Courtesy of The Archives, California Institute of Technology.

Plenty of Room at the Bottom

Richard P. Feynman

December 1959

002526

visitors since October 16, 2001

I imagine experimental physicists must often look with envy at men like Kamerlingh Onnes, who discovered a field like low temperature, which seems to be bottomless and in which one can go down and down. Such a man is then a leader and has some temporary monopoly in a scientific adventure. Percy Bridgman, in designing a way to obtain higher pressures, opened up another new field and was able to move into it and to lead us all along. The development of ever higher vacuum was a continuing development of the same kind.

I would like to describe a field, in which little has been done, but in which an enormous amount can be done in principle. This field is not quite the same as the others in that it will not tell us much of fundamental physics (in the sense of, ``What are the strange particles?'') but it is more like solid-state physics in the sense that it might tell us much of great interest about the strange phenomena that occur in complex situations. Furthermore, a point that is most important is that it would have an enormous number of technical applications.

What I want to talk about is the problem of manipulating and controlling things on a small scale.

As soon as I mention this, people tell me about miniaturization, and how far it has progressed today. They tell me about electric motors that are the size of the nail on your small finger. And there is a device on the market, they tell me, by which you can write the Lord's Prayer on the head of a pin. But that's nothing; that's the most primitive, halting step in the direction I intend to discuss. It is a staggeringly small world that is below. In the year 2000, when they look back at this age, they will wonder why it was not until the year 1960 that anybody began seriously to move in this direction.

Why cannot we write the entire 24 volumes of the Encyclopedia Brittanica on the head of a pin?

Let's see what would be involved. The head of a pin is a sixteenth of an inch across. If you magnify it by 25,000 diameters, the area of the head of the pin is then equal to the area of all the pages of the Encyclopaedia Brittanica. Therefore, all it is necessary to do is to reduce in size all the writing in the Encyclopaedia by 25,000 times. Is that possible? The resolving power of the eye is about 1/120 of an inch---that is roughly the diameter of one of the little dots on the fine half-tone reproductions in the Encyclopaedia. This, when you demagnify it by 25,000 times, is still 80 angstroms in diameter---32 atoms across, in an ordinary metal. In other words, one of those dots still would contain in its area 1,000 atoms. So, each dot can easily be adjusted in size as required by the photoengraving, and there is no question that there is enough room on the head of a pin to put all of the Encyclopaedia Brittanica.

Furthermore, it can be read if it is so written. Let's imagine that it is written in raised letters of metal; that is, where the black is in the Encyclopedia, we have raised letters of metal that are actually 1/25,000 of their ordinary size. How would we read it?

If we had something written in such a way, we could read it using techniques in common use today. (They will undoubtedly find a better way when we do actually have it written, but to make my point conservatively I shall just take techniques we know today.) We would press the metal into a plastic material and make a mold of it, then peel the plastic off very carefully, evaporate silica into the plastic to get a very thin film, then shadow it by evaporating gold at an angle against the silica so that all the little letters will appear clearly, dissolve the plastic away from the silica film, and then look through it with an electron microscope!

There is no question that if the thing were reduced by 25,000 times in the form of raised letters on the pin, it would be easy for us to read it today. Furthermore; there is no question that we would find it easy to make copies of the master; we would just need to press the same metal plate again into plastic and we would have another copy.

How do we write small?

The next question is: How do we *write* it? We have no standard technique to do this now. But let me argue that it is not as difficult as it first appears to be. We can reverse the lenses of the electron microscope in order to demagnify as well as magnify. A source of ions, sent through the microscope lenses in reverse, could be focused to a very small spot. We could write with that spot like we write in a TV cathode ray oscilloscope, by going across in lines, and having an adjustment which determines the amount of material which is going to be deposited as we scan in lines.

This method might be very slow because of space charge limitations. There will be more rapid methods. We could first make, perhaps by some photo process, a screen which has holes in it in the form of the letters. Then we would strike an arc behind the holes and draw metallic ions through the holes; then we could again use our system of lenses and make a small image in the form of ions, which would deposit the metal on the pin.

A simpler way might be this (though I am not sure it would work): We take light and, through an optical microscope running backwards, we focus it onto a very small photoelectric screen. Then electrons come away from the screen where the light is shining. These electrons are focused down in size by the electron microscope lenses to impinge directly upon the surface of the metal. Will such a beam etch away the metal if it is run long enough? I don't know. If it doesn't work for a metal surface, it must be possible to find some surface with which to coat the original pin so that, where the electrons bombard, a change is made which we could recognize later.

There is no intensity problem in these devices---not what you are used to in magnification, where you have to take a few electrons and spread them over a bigger and bigger screen; it is just the opposite. The light which we get from a page is concentrated onto a very small area so it is very intense. The few electrons which come from the photoelectric screen are demagnified down to a very tiny area so that, again, they are very intense. I don't know why this hasn't been done yet!

That's the Encyclopaedia Brittanica on the head of a pin, but let's consider all the books in the world. The Library of Congress has approximately 9 million volumes; the British Museum Library has 5 million volumes; there are also 5 million volumes in the National Library in France. Undoubtedly there are duplications, so let us say that there are some 24 million volumes of interest in the world.

What would happen if I print all this down at the scale we have been discussing? How much space would it take? It would take, of course, the area of about a million pinheads because, instead of there being just the 24 volumes of the Encyclopaedia, there are 24 million volumes. The million pinheads can be put in a square of a thousand pins on a side, or an area of about 3 square yards. That is to say, the silica replica with the paper-thin backing of plastic, with which we have made the copies, with all this information, is on an area of approximately the size of 35 pages of the Encyclopaedia. That is about half as many pages as there are in this magazine. All of the information which all of mankind has every recorded in books can be carried around in a pamphlet in your hand---and not written in code, but a simple reproduction of the original pictures, engravings, and everything else on a small scale without loss of resolution.

What would our librarian at Caltech say, as she runs all over from one building to another, if I tell her that, ten years from now, all of the information that she is struggling to keep track of--- 120,000 volumes, stacked from the floor to the ceiling, drawers full of cards, storage rooms full of the older books---can be kept on just one library card! When the University of Brazil, for example, finds that their library is burned, we can send them a copy of every book in our library by striking off a copy from the master plate in a few hours and mailing it in an envelope no bigger or heavier than any other ordinary air mail letter.

Now, the name of this talk is ``There is *Plenty* of Room at the Bottom"---not just ``There is Room at the Bottom." What I have demonstrated is that there *is* room---that you can decrease the size of things in a practical way. I now want to show that there is *plenty* of room. I will not now discuss how we are going to do it, but only what is possible in principle---in other words, what is possible according to the laws of physics. I am not inventing anti-gravity, which is possible someday only if the laws are not what we think. I am telling you what could be done if the laws *are* what we think; we are not doing it simply because we haven't yet gotten around to it.

Information on a small scale

Suppose that, instead of trying to reproduce the pictures and all the information directly in its present form, we write only the information content in a code of dots and dashes, or something like that, to represent the various letters. Each letter represents six or seven ``bits" of information; that is, you need only about six or seven dots or dashes for each letter. Now, instead of writing everything, as I did before, on the *surface* of the head of a pin, I am going to use the interior of the material as well.

Let us represent a dot by a small spot of one metal, the next dash, by an adjacent spot of another metal, and so on. Suppose, to be conservative, that a bit of information is going to require a little cube of atoms 5 times 5 times 5---that is 125 atoms. Perhaps we need a hundred and some odd atoms to make sure that the information is not lost through diffusion, or through some other process.

I have estimated how many letters there are in the Encyclopaedia, and I have assumed that each of my 24 million books is as big as an Encyclopaedia volume, and have calculated, then, how many bits of information there are (10^15). For each bit I allow 100 atoms. And it turns out that all of the information that man has carefully accumulated in all the books in the world can be written in this form in a cube of material one two-hundredth of an inch wide--- which is the barest piece of dust that can be made out by the human eye. So there is *plenty* of room at the bottom! Don't tell me about microfilm!

This fact---that enormous amounts of information can be carried in an exceedingly small space---is, of course, well known to the biologists, and resolves the mystery which existed before we understood all this clearly, of how it could be that, in the tiniest cell, all of the information for the organization of a complex creature such as ourselves can be stored. All this information---whether we have brown eyes, or whether we think at all, or that in the embryo the jawbone should first develop with a little hole in the side so that later a nerve can grow through it---all this information is contained in a very tiny fraction of the cell in the form of long-chain DNA molecules in which approximately 50 atoms are used for one bit of information about the cell.

Better electron microscopes

If I have written in a code, with 5 times 5 times 5 atoms to a bit, the question is: How could I read it today? The electron microscope is not quite good enough, with the greatest care and effort, it can only resolve about 10 angstroms. I would like to try and impress upon you while I am talking about all of these things on a small scale, the importance of improving the electron microscope by a hundred times. It is not impossible; it is not against the laws of diffraction of the electron. The wave length of the electron in such a microscope is only 1/20 of an angstrom. So it should be possible to see the individual atoms. What good would it be to see individual atoms distinctly?

We have friends in other fields---in biology, for instance. We physicists often look at them and say, ``You know the reason you fellows are making so little progress?'' (Actually I don't know any field where they are making more rapid progress than they are in biology today.) ``You should use more mathematics, like we do.'' They could answer us---but they're polite, so I'll answer for them: ``What *you* should do in order for *us* to make more rapid progress is to make the electron microscope 100 times better.''

What are the most central and fundamental problems of biology today? They are questions like: What is the sequence of bases in the DNA? What happens when you have a mutation? How is the base order in the DNA connected to the order of amino acids in the protein? What is the structure of the RNA; is it single-chain or double-chain, and how is it related in its order of bases to the DNA? What is the organization of the microsomes? How are proteins synthesized? Where does the RNA go? How does it sit? Where do the proteins sit? Where do the amino acids go in? In photosynthesis, where is the chlorophyll; how is it arranged; where are the carotenoids involved in this thing? What is the system of the conversion of light into chemical energy?

It is very easy to answer many of these fundamental biological questions; you just *look at the thing!* You will see the order of bases in the chain; you will see the structure of the microsome. Unfortunately, the present microscope sees at a scale which is just a bit too crude. Make the microscope one hundred times more powerful, and many problems of biology would be made very much easier. I exaggerate, of course, but the biologists would surely be very thankful to you---and they would prefer that to the criticism that they should use more mathematics.

The theory of chemical processes today is based on theoretical physics. In this sense, physics supplies the foundation of chemistry. But chemistry also has analysis. If you have a strange substance and you want to know what it is, you go through a long and complicated process of chemical analysis. You can analyze almost anything today, so I am a little late with my idea. But if the physicists wanted to, they could also dig under the chemists in the problem of chemical analysis. It would be very easy to make an analysis of any complicated chemical substance; all one would have to do would be to look at it and see where the atoms are. The only trouble is that the electron microscope is one hundred times too poor. (Later, I would like to ask the question: Can the physicists do something about the third problem of chemistry---namely, synthesis? Is there a *physical* way to synthesize any chemical substance?

The reason the electron microscope is so poor is that the f- value of the lenses is only 1 part to 1,000; you don't have a big enough numerical aperture. And I know that there are theorems which prove that it is impossible, with axially symmetrical stationary field lenses, to produce an f-value any bigger than so and so; and therefore the resolving power at the present time is at its theoretical maximum. But in every theorem there are assumptions. Why must the field be symmetrical? I put this out as a challenge: Is there no way to make the electron microscope more powerful?

The marvelous biological system

The biological example of writing information on a small scale has inspired me to think of something that should be possible. Biology is not simply writing information; it is *doing something* about it. A biological system can be exceedingly small. Many of the cells are very tiny, but they are very active; they manufacture various substances; they walk around; they wiggle; and they do all kinds of marvelous things---all on a very small scale. Also, they store information. Consider the possibility that we too can make a thing very small which does what we want---that we can manufacture an object that maneuvers at that level!

There may even be an economic point to this business of making things very small. Let me remind you of some of the problems of computing machines. In computers we have to store an enormous amount of information. The kind of writing that I was mentioning before, in which I had everything down as a distribution of metal, is permanent. Much more interesting to a computer is a way of writing, erasing, and writing something else. (This is usually because we don't want to waste the material on which we have just written. Yet if we could write it in a very small space, it wouldn't make any difference; it could just be thrown away after it was read. It doesn't cost very much for the material).

Miniaturizing the computer

I don't know how to do this on a small scale in a practical way, but I do know that computing machines are very large; they fill rooms. Why can't we make them very small, make them of little wires, little elements---and by little, I mean *little*. For instance, the wires should be 10 or 100 atoms in diameter, and the circuits should be a few thousand angstroms across. Everybody who has analyzed the logical theory of computers has come to the conclusion that the possibilities of computers are very interesting---if they could be made to be more complicated by several orders of magnitude. If they had millions of times as many elements, they could make judgments. They would have time to calculate what is the best way to make the calculation that they are about to make. They could select the method of analysis which, from their experience, is better than the one that we would give to them. And in many other ways, they would have new qualitative features.

If I look at your face I immediately recognize that I have seen it before. (Actually, my friends will say I have chosen an unfortunate example here for the subject of this illustration. At least I recognize that it is a *man* and not an *apple*.) Yet there is no machine which, with that speed, can take a picture of a face and say even that it is a man; and much less that it is the same man that you showed it before---unless it is exactly the same picture. If the face is changed; if I am closer to the face; if I am further from the face; if the light changes---I recognize it anyway. Now, this little computer I carry in my head is easily able to do that. The computers that we build are not able to do that. The number of elements in this bone box of mine are enormously greater than the number of elements in our ``wonderful'' computers. But our mechanical computers are too big; the elements in this box are microscopic. I want to make some that are *sub*microscopic.

If we wanted to make a computer that had all these marvelous extra qualitative abilities, we would have to make it, perhaps, the size of the Pentagon. This has several disadvantages. First, it requires too much material; there may not be enough germanium in the world for all the transistors which would have to be put into this enormous thing. There is also the problem of heat generation and power consumption; TVA would be needed to run the computer. But an even more practical difficulty is that the computer would be limited to a certain speed. Because of its large size, there is finite time required to get the information from one place to another. The information cannot go any faster than the speed of light---so, ultimately, when our computers get faster and faster and more and more elaborate, we will have to make them smaller and smaller.

But there is plenty of room to make them smaller. There is nothing that I can see in the physical laws that says the computer elements cannot be made enormously smaller than they are now. In fact, there may be certain advantages.

Miniaturization by evaporation

How can we make such a device? What kind of manufacturing processes would we use? One possibility we might consider, since we have talked about writing by putting atoms down in a certain arrangement, would be to evaporate the material, then evaporate the insulator next to it. Then, for the next layer, evaporate another position of a wire, another insulator, and so on. So, you simply evaporate until you have a block of stuff which has the elements--- coils and condensers, transistors and so on---of exceedingly fine dimensions.

But I would like to discuss, just for amusement, that there are other possibilities. Why can't we manufacture these small computers somewhat like we manufacture the big ones? Why can't we drill holes, cut things, solder things, stamp things out, mold different shapes all at an infinitesimal level? What are the limitations as to how small a thing has to be before you can no longer mold it? How many times when you are working on something frustratingly tiny like your wife's wrist watch, have you said to yourself, ``If I could only train an ant to do this!'' What I would like to suggest is the possibility of training an ant to train a mite to do this. What are the possibilities of small but movable machines? They may or may not be useful, but they surely would be fun to make.

Consider any machine---for example, an automobile---and ask about the problems of making an infinitesimal machine like it. Suppose, in the particular design of the automobile, we need a certain precision of the parts; we need an accuracy, let's suppose, of 4/10,000 of an inch. If things are more inaccurate than that in the shape of the cylinder and so on, it isn't going to work very well. If I make the thing too small, I have to worry about the size of the atoms; I can't make a circle of ``balls'' so to speak, if the circle is too small. So, if I make the error, corresponding to 4/10,000 of an inch, correspond to an error of 10 atoms, it turns out that I can reduce the dimensions of an automobile 4,000 times, approximately---so that it is 1 mm. across. Obviously, if you redesign the car so that it would work with a much larger tolerance, which is not at all impossible, then you could make a much smaller device.

It is interesting to consider what the problems are in such small machines. Firstly, with parts stressed to the same degree, the forces go as the area you are reducing, so that things like weight and inertia are of relatively no importance. The strength of material, in other words, is very much greater in proportion. The stresses and expansion of the flywheel from centrifugal force, for example, would be the same proportion only if the rotational speed is increased in the same proportion as we decrease the size. On the other hand, the metals that we use have a grain structure, and this would be very annoying at small scale because the material is not homogeneous. Plastics and glass and things of this amorphous nature are very much more homogeneous, and so we would have to make our machines out of such materials.

There are problems associated with the electrical part of the system---with the copper wires and the magnetic parts. The magnetic properties on a very small scale are not the same as on a large scale; there is the ``domain'' problem involved. A big magnet made of millions of domains can only be made on a small scale with one domain. The electrical equipment won't simply be scaled down; it has to be redesigned. But I can see no reason why it can't be redesigned to work again.

Problems of lubrication

Lubrication involves some interesting points. The effective viscosity of oil would be higher and higher in proportion as we went down (and if we increase the speed as much as we can). If we don't increase the speed so much, and change from oil to kerosene or some other fluid, the problem is not so bad. But actually we may not have to lubricate at all! We have a lot of extra force. Let the bearings run dry; they won't run hot because the heat escapes away from such a small device very, very rapidly.

This rapid heat loss would prevent the gasoline from exploding, so an internal combustion engine is impossible. Other chemical reactions, liberating energy when cold, can be used. Probably an external supply of electrical power would be most convenient for such small machines.

What would be the utility of such machines? Who knows? Of course, a small automobile would only be useful for the mites to drive around in, and I suppose our Christian interests don't go that far. However, we did note the possibility of the manufacture of small elements for computers in completely automatic factories, containing lathes and other machine tools at the very small level. The small lathe would not have to be exactly like our big lathe. I leave to your imagination the improvement of the design to take full advantage of the properties of things on a small scale, and in such a way that the fully automatic aspect would be easiest to manage.

A friend of mine (Albert R. Hibbs) suggests a very interesting possibility for relatively small machines. He says that, although it is a very wild idea, it would be interesting in surgery if you could swallow the surgeon. You put the mechanical surgeon inside the blood vessel and it goes into the heart and ``looks'' around. (Of course the information has to be fed out.) It finds out which valve is the faulty one and takes a little knife and slices it out. Other small machines might be permanently incorporated in the body to assist some inadequately-functioning organ.

Now comes the interesting question: How do we make such a tiny mechanism? I leave that to you. However, let me suggest one weird possibility. You know, in the atomic energy plants they have materials and machines that they can't handle directly because they have become radioactive. To unscrew nuts and put on bolts and so on, they have a set of master and slave hands, so that by operating a set of levers here, you control the ``hands'' there, and can turn them this way and that so you can handle things quite nicely.

Most of these devices are actually made rather simply, in that there is a particular cable, like a marionette string, that goes directly from the controls to the ``hands.'' But, of course, things also have been made using servo motors, so that the connection between the one thing and the other is electrical rather than mechanical. When you turn the levers, they turn a servo motor, and it changes the electrical currents in the wires, which repositions a motor at the other end.

Now, I want to build much the same device---a master-slave system which operates electrically. But I want the slaves to be made especially carefully by modern large-scale machinists so that they are one-fourth the scale of the ``hands'' that you ordinarily maneuver. So you have a scheme by which you can do things at one- quarter scale anyway---the little servo motors with little hands play with little nuts and bolts; they drill little holes; they are four times smaller. Aha! So I manufacture a quarter-size lathe; I manufacture quarter-size tools; and I make, at the one-quarter scale, still another set of hands again relatively one-quarter size! This is one-sixteenth size, from my point of view. And after I finish doing this I wire directly from my large-scale system, through transformers perhaps, to the one-sixteenth-size servo motors. Thus I can now manipulate the one-sixteenth size hands.

Well, you get the principle from there on. It is rather a difficult program, but it is a possibility. You might say that one can go much farther in one step than from one to four. Of course, this has all to be designed very carefully and it is not necessary simply to make it like hands. If you thought of it very carefully, you could probably arrive at a much better system for doing such things.

If you work through a pantograph, even today, you can get much more than a factor of four in even one step. But you can't work directly through a pantograph which makes a smaller pantograph which then makes a smaller pantograph---because of the looseness of the holes and the irregularities of construction. The end of the pantograph wiggles with a relatively greater irregularity than the irregularity with which you move your hands. In going down this scale, I would find the end of the pantograph on the end of the pantograph on the end of the pantograph shaking so badly that it wasn't doing anything sensible at all.

At each stage, it is necessary to improve the precision of the apparatus. If, for instance, having made a small lathe with a pantograph, we find its lead screw irregular---more irregular than the large-scale one---we could lap the lead screw against breakable nuts that you can reverse in the usual way back and forth until this lead screw is, at its scale, as accurate as our original lead screws, at our scale.

We can make flats by rubbing unflat surfaces in triplicates together---in three pairs---and the flats then become flatter than the thing you started with. Thus, it is not impossible to improve precision on a small scale by the correct operations. So, when we build this stuff, it is necessary at each step to improve the accuracy of the equipment by working for awhile down there, making accurate lead screws, Johansen blocks, and all the other materials which we use in accurate machine work at the higher level. We have to stop at each level and manufacture all the stuff to go to the next level---a very long and very difficult program. Perhaps you can figure a better way than that to get down to small scale more rapidly.

Yet, after all this, you have just got one little baby lathe four thousand times smaller than usual. But we were thinking of making an enormous computer, which we were going to build by drilling holes on this lathe to make little washers for the computer. How many washers can you manufacture on this one lathe?

A hundred tiny hands

When I make my first set of slave ``hands" at one-fourth scale, I am going to make ten sets. I make ten sets of ``hands," and I wire them to my original levers so they each do exactly the same thing at the same time in parallel. Now, when I am making my new devices one-quarter again as small, I let each one manufacture ten copies, so that I would have a hundred ``hands" at the 1/16th size.

Where am I going to put the million lathes that I am going to have? Why, there is nothing to it; the volume is much less than that of even one full-scale lathe. For instance, if I made a billion little lathes, each 1/4000 of the scale of a regular lathe, there are plenty of materials and space available because in the billion little ones there is less than 2 percent of the materials in one big lathe.

It doesn't cost anything for materials, you see. So I want to build a billion tiny factories, models of each other, which are manufacturing simultaneously, drilling holes, stamping parts, and so on.

As we go down in size, there are a number of interesting problems that arise. All things do not simply scale down in proportion. There is the problem that materials stick together by the molecular (Van der Waals) attractions. It would be like this: After you have made a part and you unscrew the nut from a bolt, it isn't going to fall down because the gravity isn't appreciable; it would even be hard to get it off the bolt. It would be like those old movies of a man with his hands full of molasses, trying to get rid of a glass of water. There will be several problems of this nature that we will have to be ready to design for.

Rearranging the atoms

But I am not afraid to consider the final question as to whether, ultimately---in the great future---we can arrange the atoms the way we want; the very *atoms*, all the way down! What would happen if we could arrange the atoms one by one the way we want them (within reason, of course; you can't put them so that they are chemically unstable, for example).

Up to now, we have been content to dig in the ground to find minerals. We heat them and we do things on a large scale with them, and we hope to get a pure substance with just so much impurity, and so on. But we must always accept some atomic arrangement that nature gives us. We haven't got anything, say, with a ``checkerboard" arrangement, with the impurity atoms exactly arranged 1,000 angstroms apart, or in some other particular pattern.

What could we do with layered structures with just the right layers? What would the properties of materials be if we could really arrange the atoms the way we want them? They would be very interesting to investigate theoretically. I can't see exactly what would happen, but I can hardly doubt that when we have some *control* of the arrangement of things on a small scale we will get an enormously greater range of possible properties that substances can have, and of different things that we can do.

Consider, for example, a piece of material in which we make little coils and condensers (or their solid state analogs) 1,000 or 10,000 angstroms in a circuit, one right next to the other, over a large area, with little antennas sticking out at the other end---a whole series of circuits. Is it possible, for example, to emit light from a whole set of antennas, like we emit radio waves from an organized set of antennas to beam the radio programs to Europe? The same thing would be to *beam* the light out in a definite direction with very high intensity. (Perhaps such a beam is not very useful technically or economically.)

I have thought about some of the problems of building electric circuits on a small scale, and the problem of resistance is serious. If you build a corresponding circuit on a small scale, its natural frequency goes up, since the wave length goes down as the scale; but the skin depth only decreases with the square root of the scale ratio, and so resistive problems are of increasing difficulty. Possibly we can beat resistance through the use of superconductivity if the frequency is not too high, or by other tricks.

Atoms in a small world

When we get to the very, very small world---say circuits of seven atoms---we have a lot of new things that would happen that represent completely new opportunities for design. Atoms on a small scale behave like *nothing* on a large scale, for they satisfy the laws of quantum mechanics. So, as we go down and fiddle around with the atoms down there, we are working with different laws, and we can expect to do different things. We can manufacture in different ways. We can use, not just circuits, but some system involving the quantized energy levels, or the interactions of quantized spins, etc.

Another thing we will notice is that, if we go down far enough, all of our devices can be mass produced so that they are absolutely perfect copies of one another. We cannot build two large machines so that the dimensions are exactly the same. But if your machine is only 100 atoms high, you only have to get it correct to one-half of one percent to make sure the other machine is exactly the same size---namely, 100 atoms high!

At the atomic level, we have new kinds of forces and new kinds of possibilities, new kinds of effects. The problems of manufacture and reproduction of materials will be quite different. I am, as I said, inspired by the biological phenomena in which chemical forces are used in repetitious fashion to produce all kinds of weird effects (one of which is the author).

The principles of physics, as far as I can see, do not speak against the possibility of maneuvering things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but in practice, it has not been done because we are too big.

Ultimately, we can do chemical synthesis. A chemist comes to us and says, ``Look, I want a molecule that has the atoms arranged thus and so; make me that molecule.'' The chemist does a mysterious thing when he wants to make a molecule. He sees that it has got that ring, so he mixes this and that, and he shakes it, and he fiddles around. And, at the end of a difficult process, he usually does succeed in synthesizing what he wants. By the time I get my devices working, so that we can do it by physics, he will have figured out how to synthesize absolutely anything, so that this will really be useless.

But it is interesting that it would be, in principle, possible (I think) for a physicist to synthesize any chemical substance that the chemist writes down. Give the orders and the physicist synthesizes it. How? Put the atoms down where the chemist says, and so you make the substance. The problems of chemistry and biology can be greatly helped if our ability to see what we are doing, and to do things on an atomic level, is ultimately developed---a development which I think cannot be avoided.

Now, you might say, ``Who should do this and why should they do it?'' Well, I pointed out a few of the economic applications, but I know that the reason that you would do it might be just for fun. But have some fun! Let's have a competition between laboratories. Let one laboratory make a tiny motor which it sends to another lab which sends it back with a thing that fits inside the shaft of the first motor.

High school competition

Just for the fun of it, and in order to get kids interested in this field, I would propose that someone who has some contact with the high schools think of making some kind of high school competition. After all, we haven't even started in this field, and even the kids can write smaller than has ever been written before. They could have competition in high schools. The Los Angeles high school could send a pin to the Venice high school on which it says, ``How's this?'' They get the pin back, and in the dot of the ``i'' it says, ``Not so hot.''

Perhaps this doesn't excite you to do it, and only economics will do so. Then I want to do something; but I can't do it at the present moment, because I haven't prepared the ground. It is my intention to offer a prize of $1,000 to the first guy who can take the information on the page of a book and put it on an area 1/25,000 smaller in linear scale in such manner that it can be read by an electron microscope.

And I want to offer another prize---if I can figure out how to phrase it so that I don't get into a mess of arguments about definitions---of another $1,000 to the first guy who makes an operating electric motor---a rotating electric motor which can be controlled from the outside and, not counting the lead-in wires, is only 1/64 inch cube.

I do not expect that such prizes will have to wait very long for claimants.

The experts look ahead

Cramming more components onto integrated circuits

With unit cost falling as the number of components per circuit rises, by 1975 economics may dictate squeezing as many as 65,000 components on a single silicon chip

By Gordon E. Moore

Director, Research and Development Laboratories, Fairchild Semiconductor division of Fairchild Camera and Instrument Corp.

The future of integrated electronics is the future of electronics itself. The advantages of integration will bring about a proliferation of electronics, pushing this science into many new areas.

Integrated circuits will lead to such wonders as home computers—or at least terminals connected to a central computer—automatic controls for automobiles, and personal portable communications equipment. The electronic wristwatch needs only a display to be feasible today.

But the biggest potential lies in the production of large systems. In telephone communications, integrated circuits in digital filters will separate channels on multiplex equipment. Integrated circuits will also switch telephone circuits and perform data processing.

Computers will be more powerful, and will be organized in completely different ways. For example, memories built of integrated electronics may be distributed throughout the machine instead of being concentrated in a central unit. In addition, the improved reliability made possible by integrated circuits will allow the construction of larger processing units. Machines similar to those in existence today will be built at lower costs and with faster turn-around.

Present and future

By integrated electronics, I mean all the various technologies which are referred to as microelectronics today as well as any additional ones that result in electronics functions supplied to the user as irreducible units. These technologies were first investigated in the late 1950's. The object was to miniaturize electronics equipment to include increasingly complex electronic functions in limited space with minimum weight. Several approaches evolved, including microassembly techniques for individual components, thin-film structures and semiconductor integrated circuits.

Each approach evolved rapidly and converged so that each borrowed techniques from another. Many researchers believe the way of the future to be a combination of the various approaches.

The advocates of semiconductor integrated circuitry are already using the improved characteristics of thin-film resistors by applying such films directly to an active semiconductor substrate. Those advocating a technology based upon films are developing sophisticated techniques for the attachment of active semiconductor devices to the passive film arrays.

Both approaches have worked well and are being used in equipment today.

The author

Dr. Gordon E. Moore is one of the new breed of electronic engineers, schooled in the physical sciences rather than in electronics. He earned a B.S. degree in chemistry from the University of California and a Ph.D. degree in physical chemistry from the California Institute of Technology. He was one of the founders of Fairchild Semiconductor and has been director of the research and development laboratories since 1959.

The establishment

Integrated electronics is established today. Its techniques are almost mandatory for new military systems, since the reliability, size and weight required by some of them is achievable only with integration. Such programs as Apollo, for manned moon flight, have demonstrated the reliability of integrated electronics by showing that complete circuit functions are as free from failure as the best individual transistors.

Most companies in the commercial computer field have machines in design or in early production employing integrated electronics. These machines cost less and perform better than those which use "conventional" electronics.

Instruments of various sorts, especially the rapidly increasing numbers employing digital techniques, are starting to use integration because it cuts costs of both manufacture and design.

The use of linear integrated circuitry is still restricted primarily to the military. Such integrated functions are expensive and not available in the variety required to satisfy a major fraction of linear electronics. But the first applications are beginning to appear in commercial electronics, particularly in equipment which needs low-frequency amplifiers of small size.

Reliability counts

In almost every case, integrated electronics has demonstrated high reliability. Even at the present level of production—low compared to that of discrete components—it offers reduced systems cost, and in many systems improved performance has been realized.

Integrated electronics will make electronic techniques more generally available throughout all of society, performing many functions that presently are done inadequately by other techniques or not done at all. The principal advantages will be lower costs and greatly simplified design—payoffs from a ready supply of low-cost functional packages.

For most applications, semiconductor integrated circuits will predominate. Semiconductor devices are the only reasonable candidates presently in existence for the active elements of integrated circuits. Passive semiconductor elements look attractive too, because of their potential for low cost and high reliability, but they can be used only if precision is not a prime requisite.

Silicon is likely to remain the basic material, although others will be of use in specific applications. For example, gallium arsenide will be important in integrated microwave functions. But silicon will predominate at lower frequencies because of the technology which has already evolved around it and its oxide, and because it is an abundant and relatively inexpensive starting material.

Costs and curves

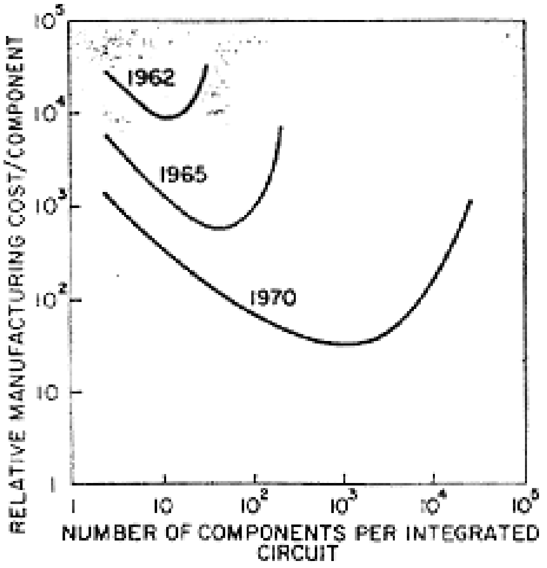
Reduced cost is one of the big attractions of integrated electronics, and the cost advantage continues to increase as the technology evolves toward the production of larger and larger circuit functions on a single semiconductor substrate. For simple circuits, the cost per component is nearly inversely proportional to the number of components, the result of the equivalent piece of semiconductor in the equivalent package containing more components. But as components are added, decreased yields more than compensate for the increased complexity, tending to raise the cost per component. Thus there is a minimum cost at any given time in the evolution of the technology. At present, it is reached when 50 components are used per circuit. But the minimum is rising rapidly while the entire cost curve is falling (see graph below). If we look ahead five years, a plot of costs suggests that the minimum cost per component might be expected in circuits with about 1,000 components per circuit (providing such circuit functions can be produced in moderate quantities.) In 1970, the manufacturing cost per component can be expected to be only a tenth of the present cost.

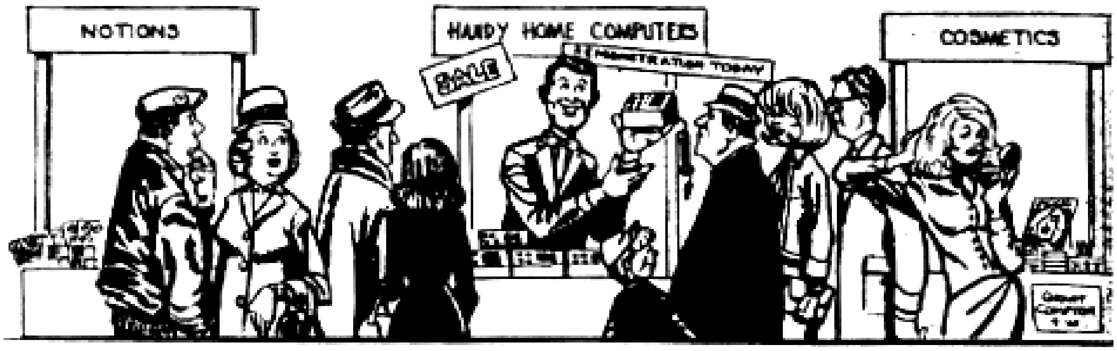
The complexity for minimum component costs has increased at a rate of roughly a factor of two per year (see graph on next page). Certainly over the short term this rate can be expected to continue, if not to increase. Over the longer term, the rate of increase is a bit more uncertain, although there is no reason to believe it will not remain nearly constant for at least 10 years. That means by 1975, the number of components per integrated circuit for minimum cost will be 65,000.

I believe that such a large circuit can be built on a single wafer.

Two-mil squares

With the dimensional tolerances already being employed in integrated circuits, isolated high-performance transistors can be built on centers two thousandths of an inch apart. Such



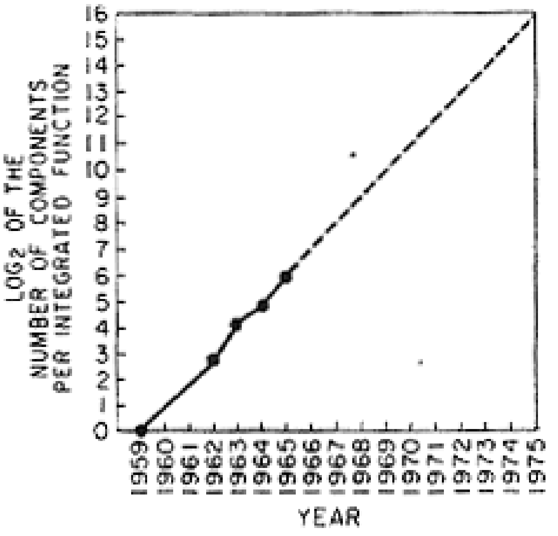


a two-mil square can also contain several kilohms of resistance or a few diodes. This allows at least 500 components per linear inch or a quarter million per square inch. Thus, 65,000 components need occupy only about one-fourth a square inch.

On the silicon wafer currently used, usually an inch or more in diameter, there is ample room for such a structure if the components can be closely packed with no space wasted for interconnection patterns. This is realistic, since efforts to achieve a level of complexity above the presently available integrated circuits are already underway using multilayer metalization patterns separated by dielectric films. Such a density of components can be achieved by present optical techniques and does not require the more exotic techniques, such as electron beam operations, which are being studied to make even smaller structures.

Increasing the yield

There is no fundamental obstacle to achieving device yields of 100%. At present, packaging costs so far exceed the cost of the semiconductor structure itself that there is no incentive to improve yields, but they can be raised as high as is economically justified. No barrier exists comparable to the thermodynamic equilibrium considerations that often limit yields in chemical reactions; it is not even necessary to do any fundamental research or to replace present processes. Only the engineering effort is needed.



In the early days of integrated circuitry, when yields were extremely low, there was such incentive. Today ordinary integrated circuits are made with yields comparable with those obtained for individual semiconductor devices. The same pattern will make larger arrays economical, if other considerations make such arrays desirable.

Heat problem

Will it be possible to remove the heat generated by tens of thousands of components in a single silicon chip?

If we could shrink the volume of a standard high-speed digital computer to that required for the components themselves, we would expect it to glow brightly with present power dissipation. But it won't happen with integrated circuits. Since integrated electronic structures are two-dimensional, they have a surface available for cooling close to each center of heat generation. In addition, power is needed primarily to drive the various lines and capacitances associated with the system. As long as a function is confined to a small area on a wafer, the amount of capacitance which must be driven is distinctly limited. In fact, shrinking dimensions on an integrated structure makes it possible to operate the structure at higher speed for the same power per unit area.

Day of reckoning

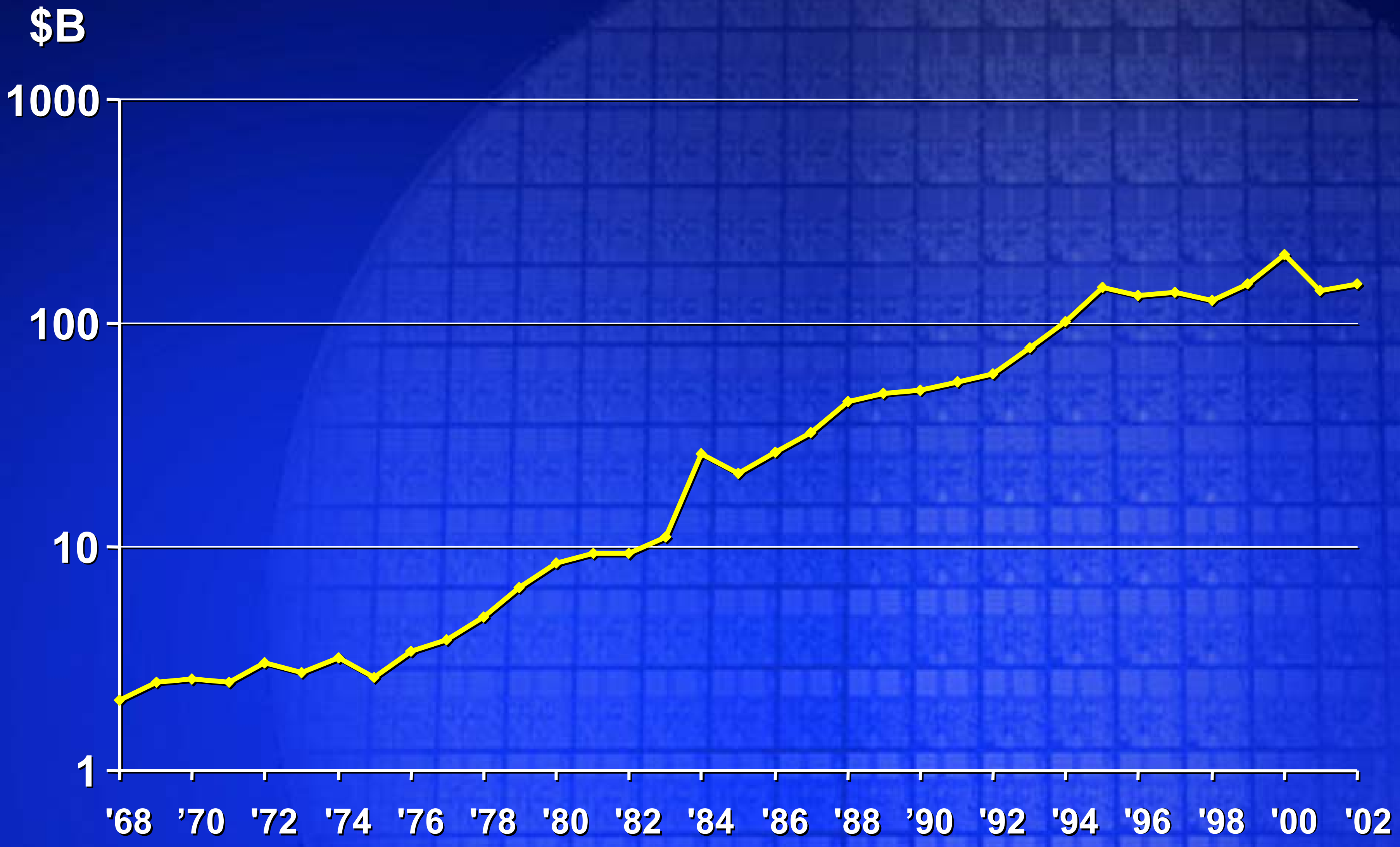
Clearly, we will be able to build such component-crammed equipment. Next, we ask under what circumstances we should do it. The total cost of making a particular system function must be minimized. To do so, we could amortize the engineering over several identical items, or evolve flexible techniques for the engineering of large functions so that no disproportionate expense need be borne by a particular array. Perhaps newly devised design automation procedures could translate from logic diagram to technological realization without any special engineering.

It may prove to be more economical to build large

NO EXPONENTIAL IS FOREVER . . .

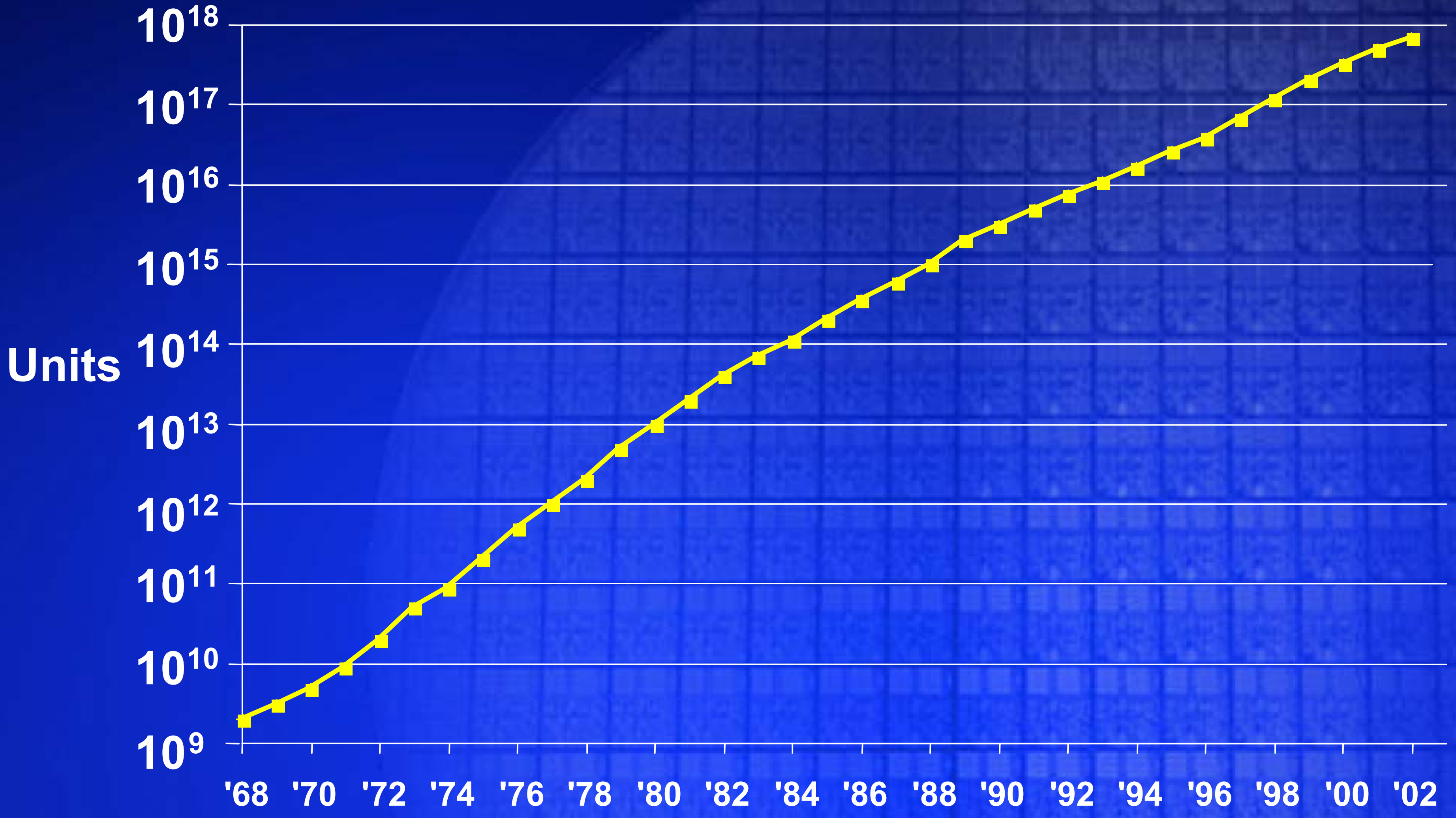
Gordon E. Moore

Worldwide Semiconductor Revenues



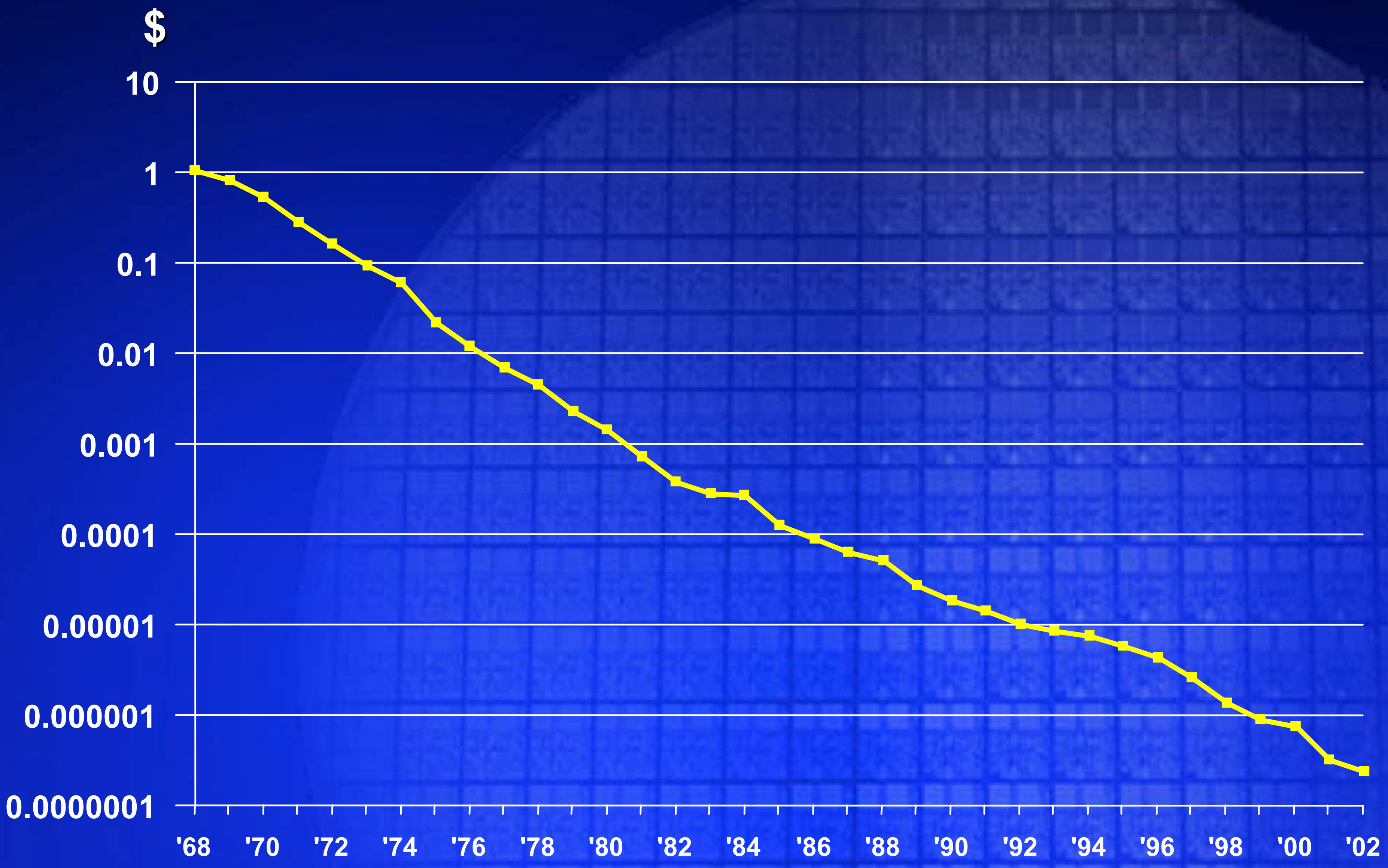
Source: Intel/WSTS,12/02

Transistors Shipped Per Year



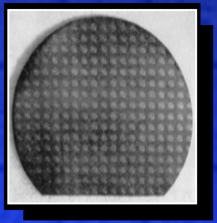
Source: Dataquest/Intel, 12/02

Average Transistor Price By Year

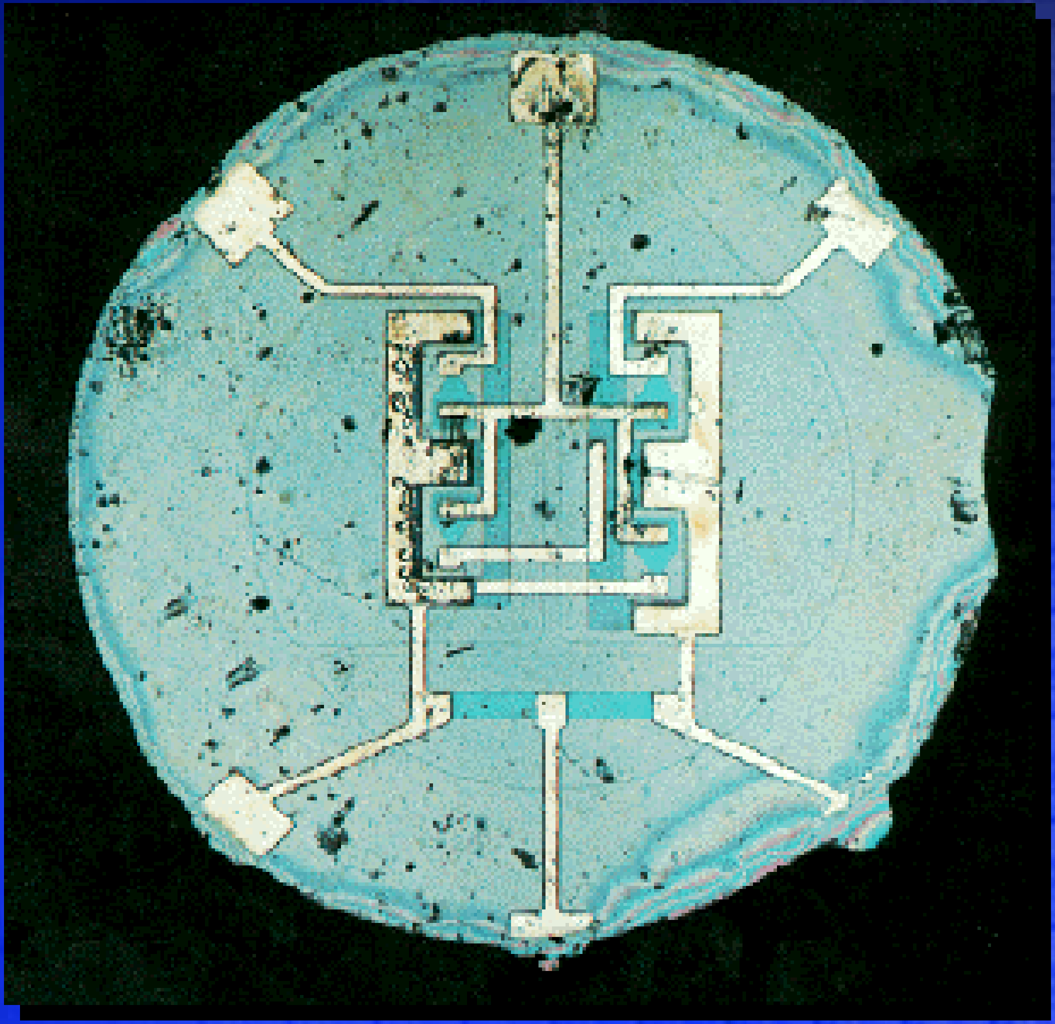


Source: Dataquest/Intel12/02

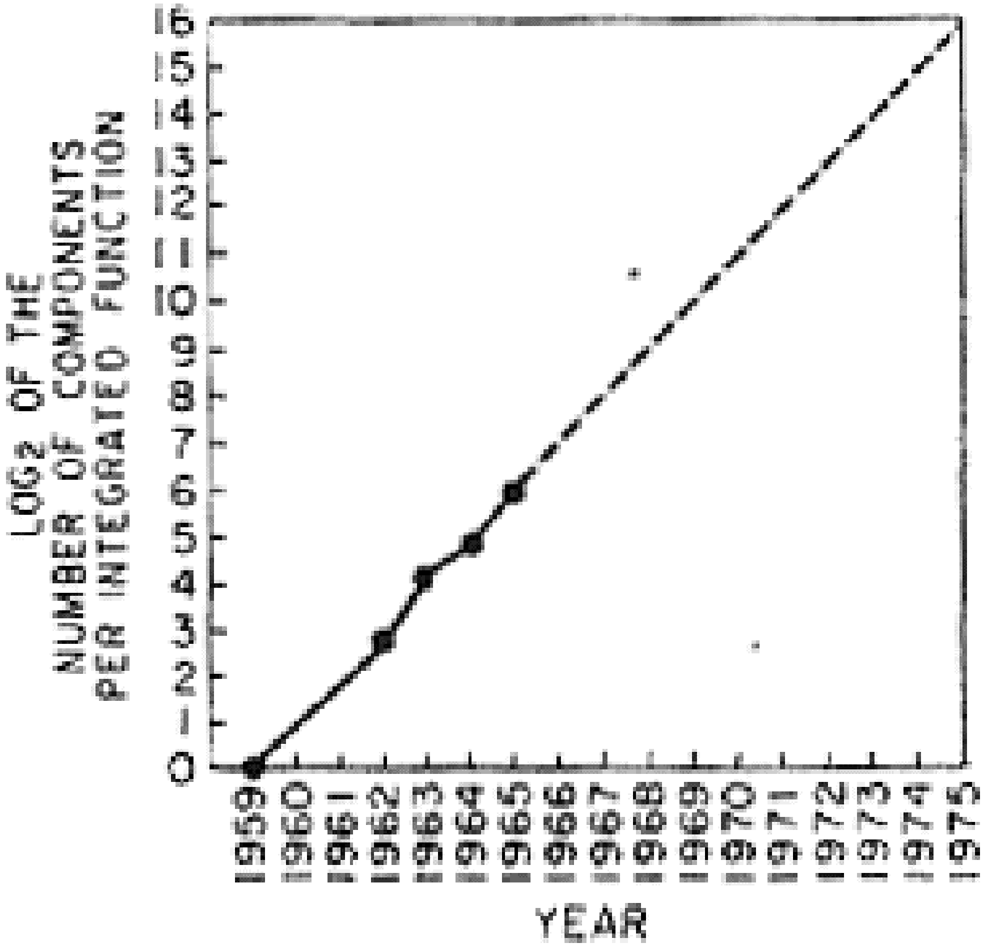
1” Wafer Of Planar Transistors, ~1959



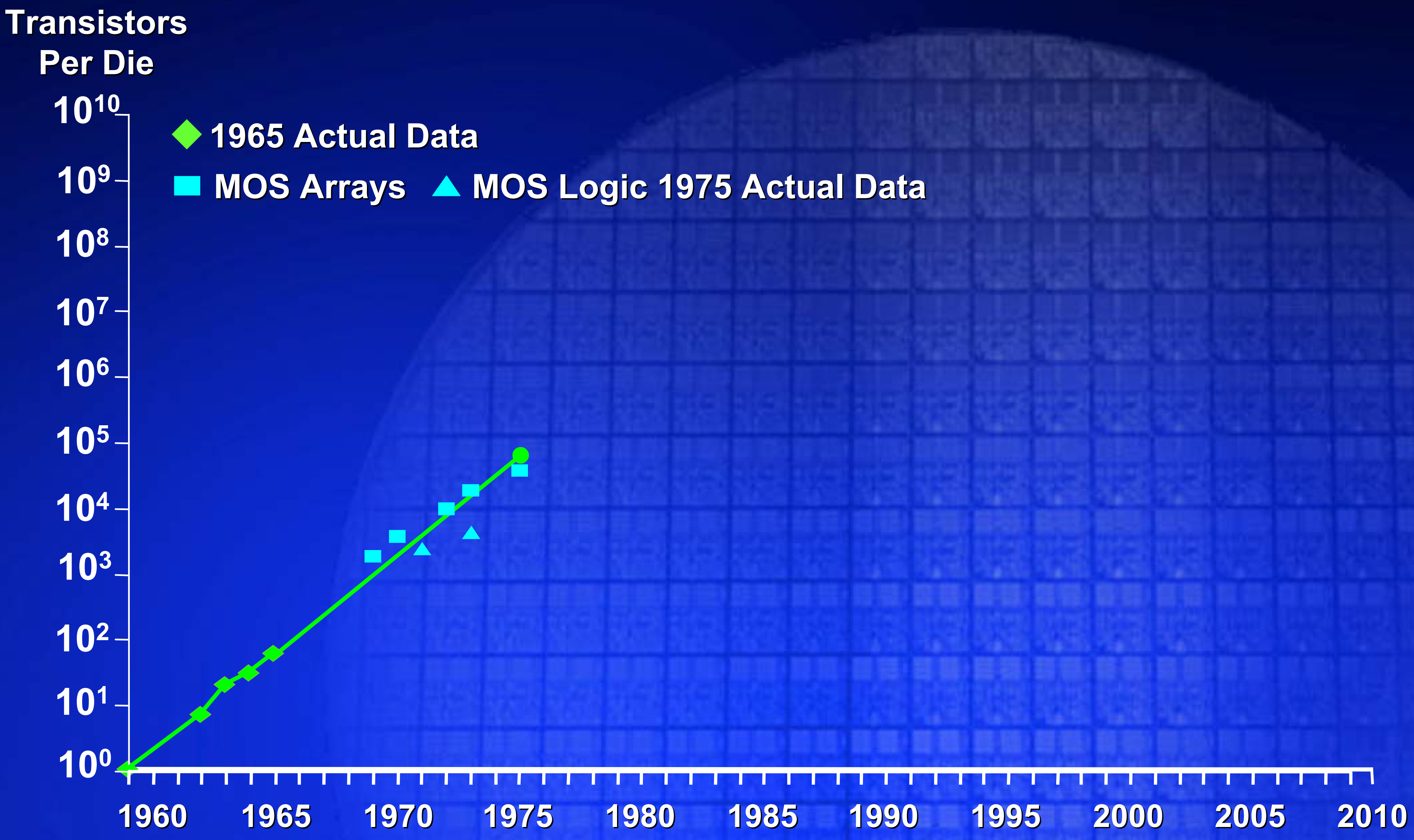
The First Planar Integrated Circuit, 1961



1965 Transistor Projection

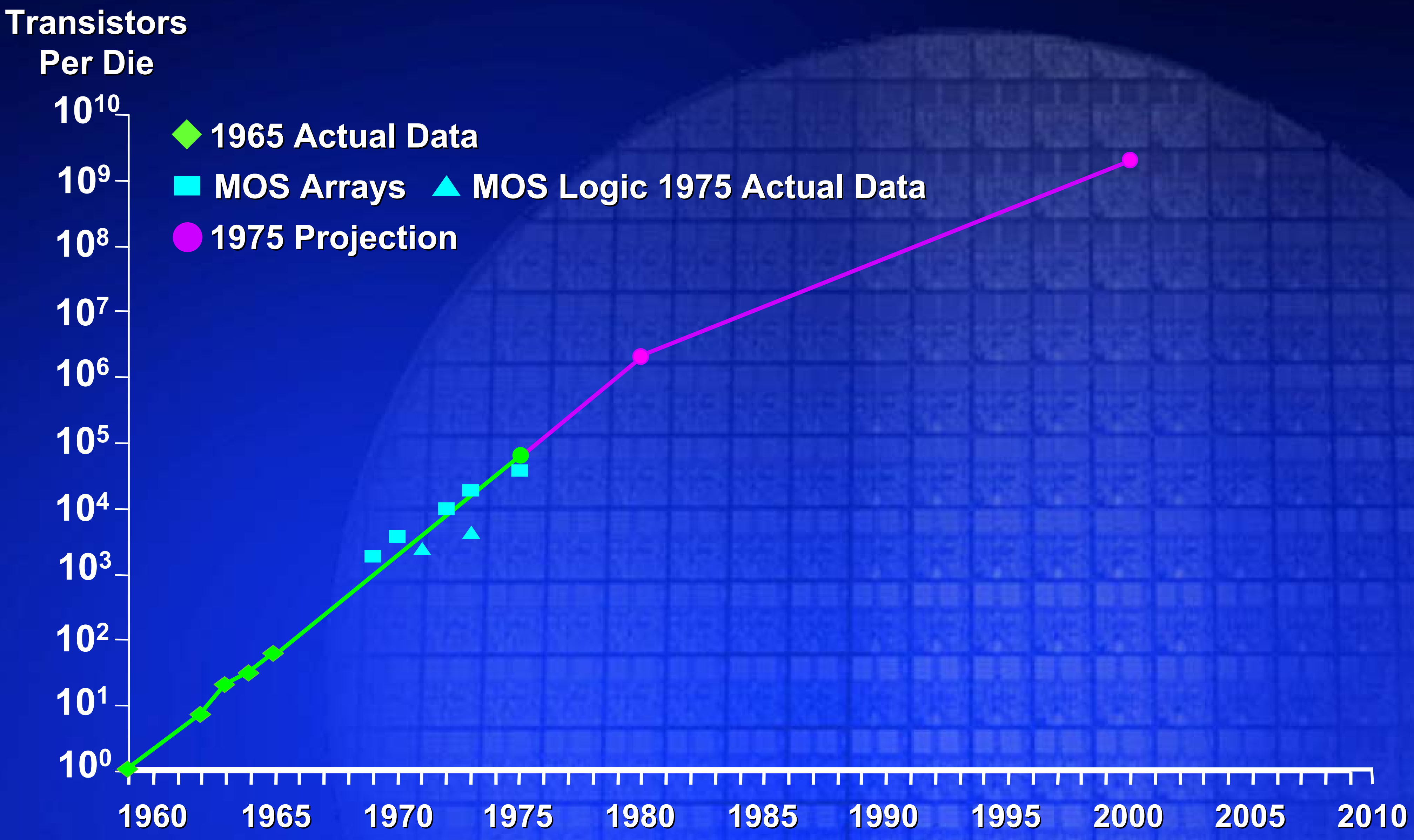


Integrated Circuit Complexity



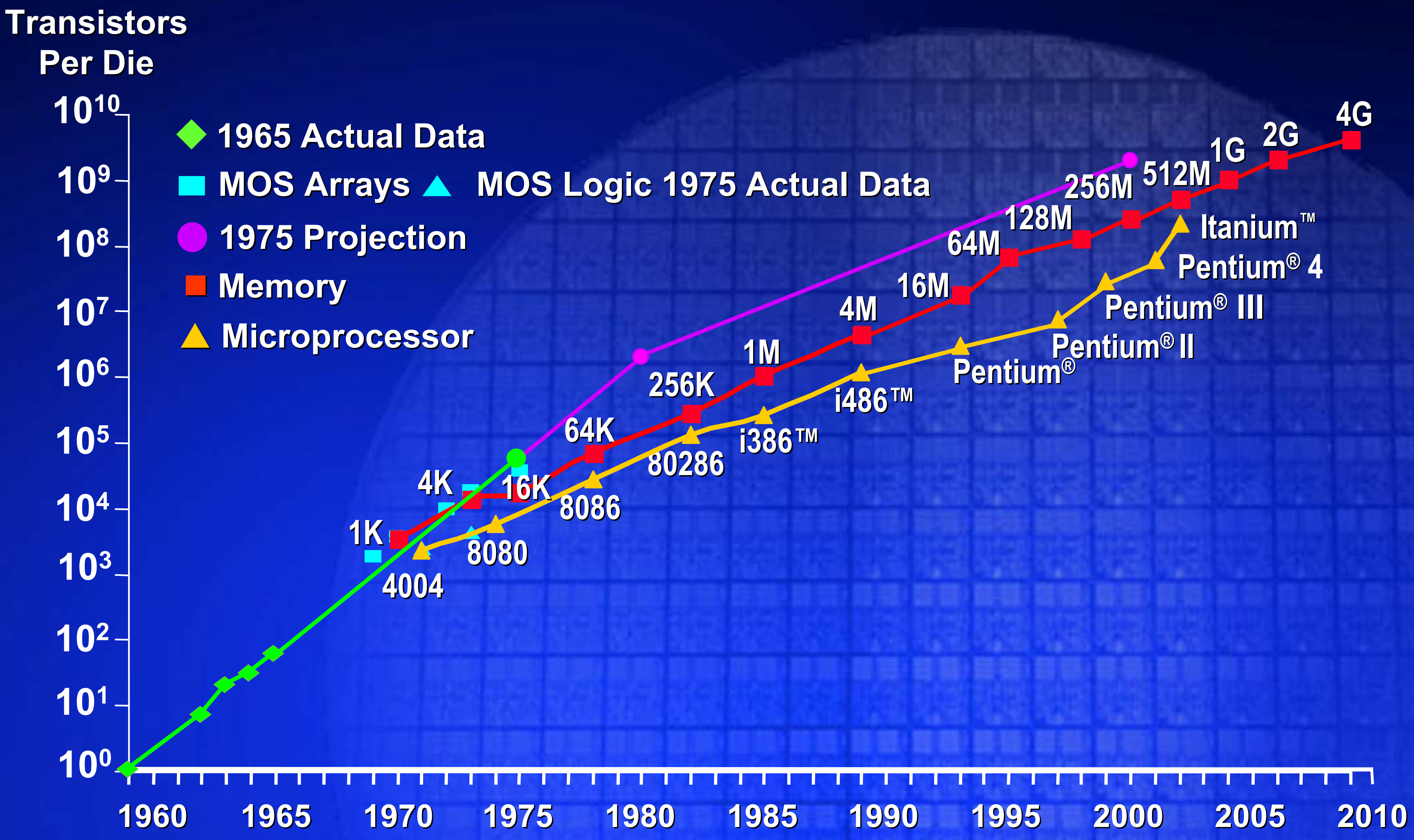
Source: Intel

Integrated Circuit Complexity



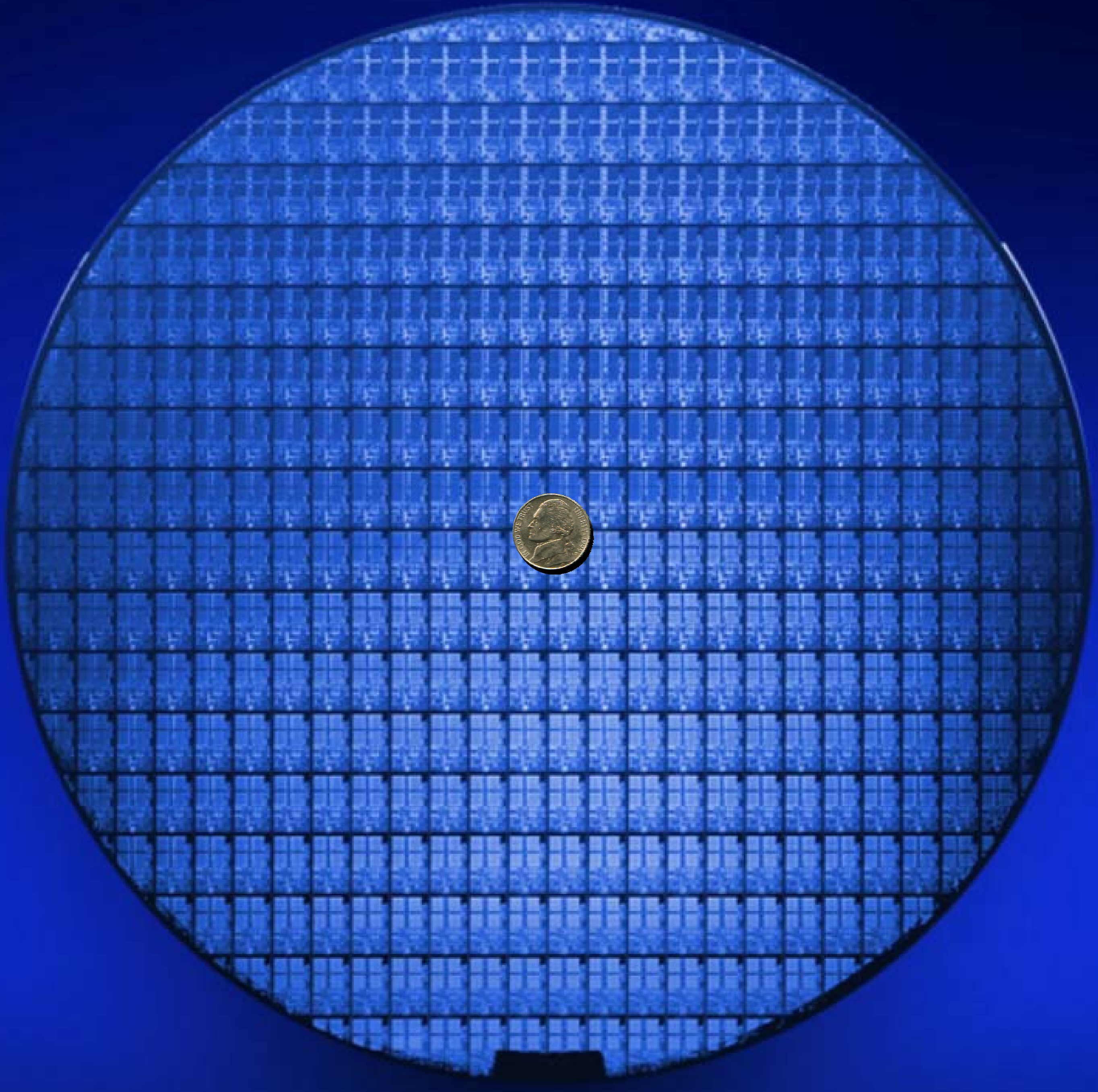
Source: Intel

Integrated Circuit Complexity

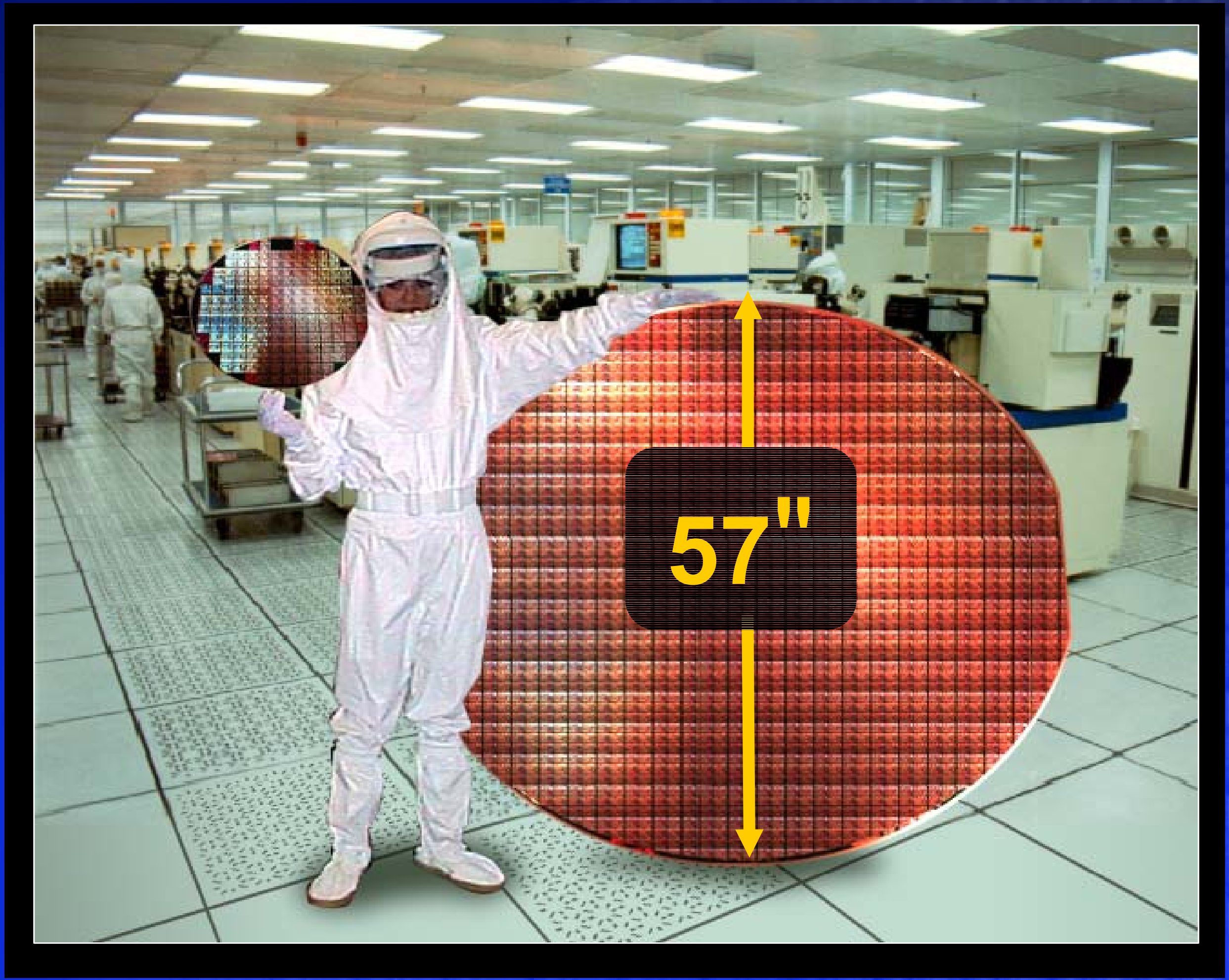


Source: Intel

300mm Wafer

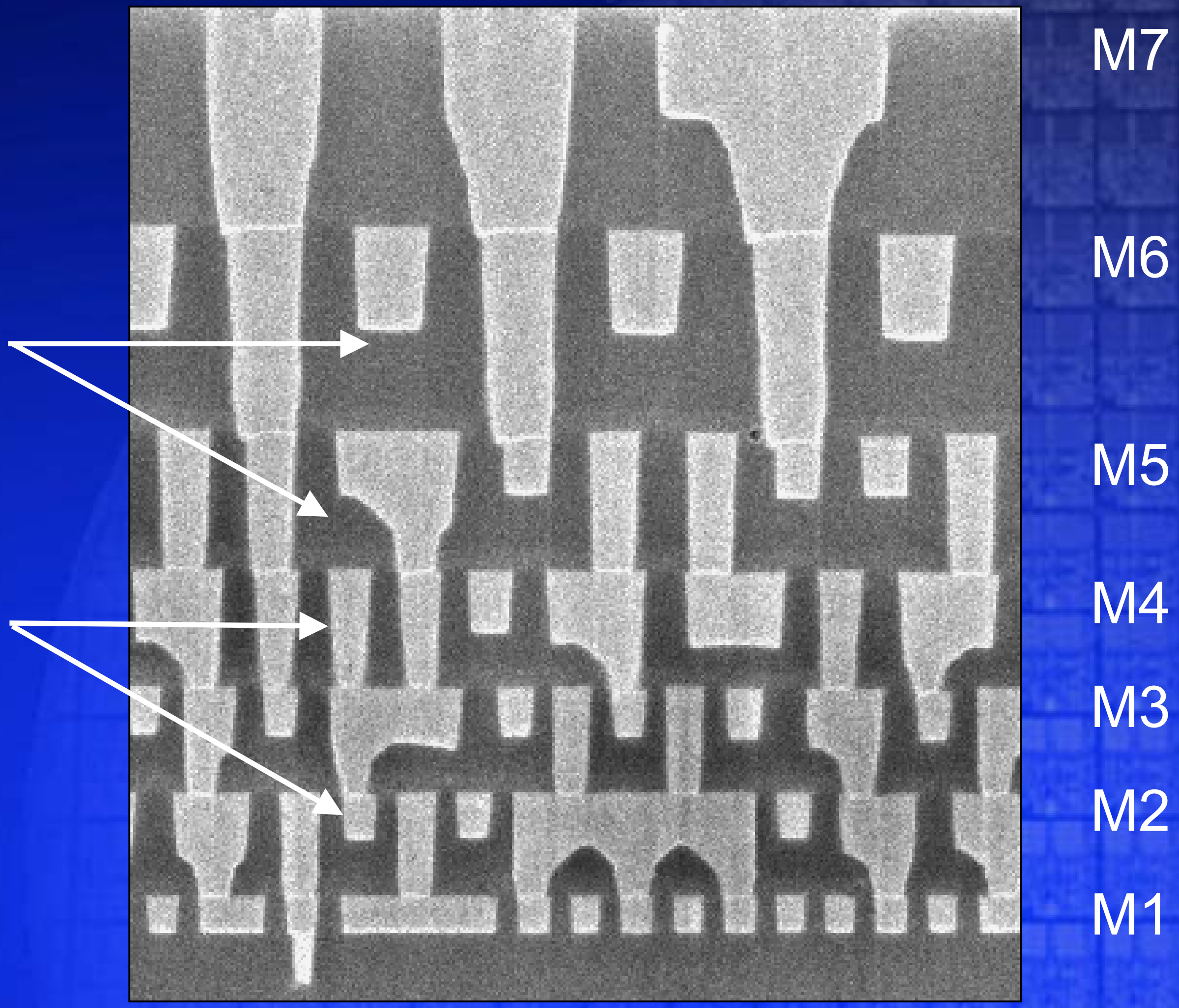


Projected 2000 Wafer, circa 1975



Moore was not always accurate

90 nm Generation Interconnects

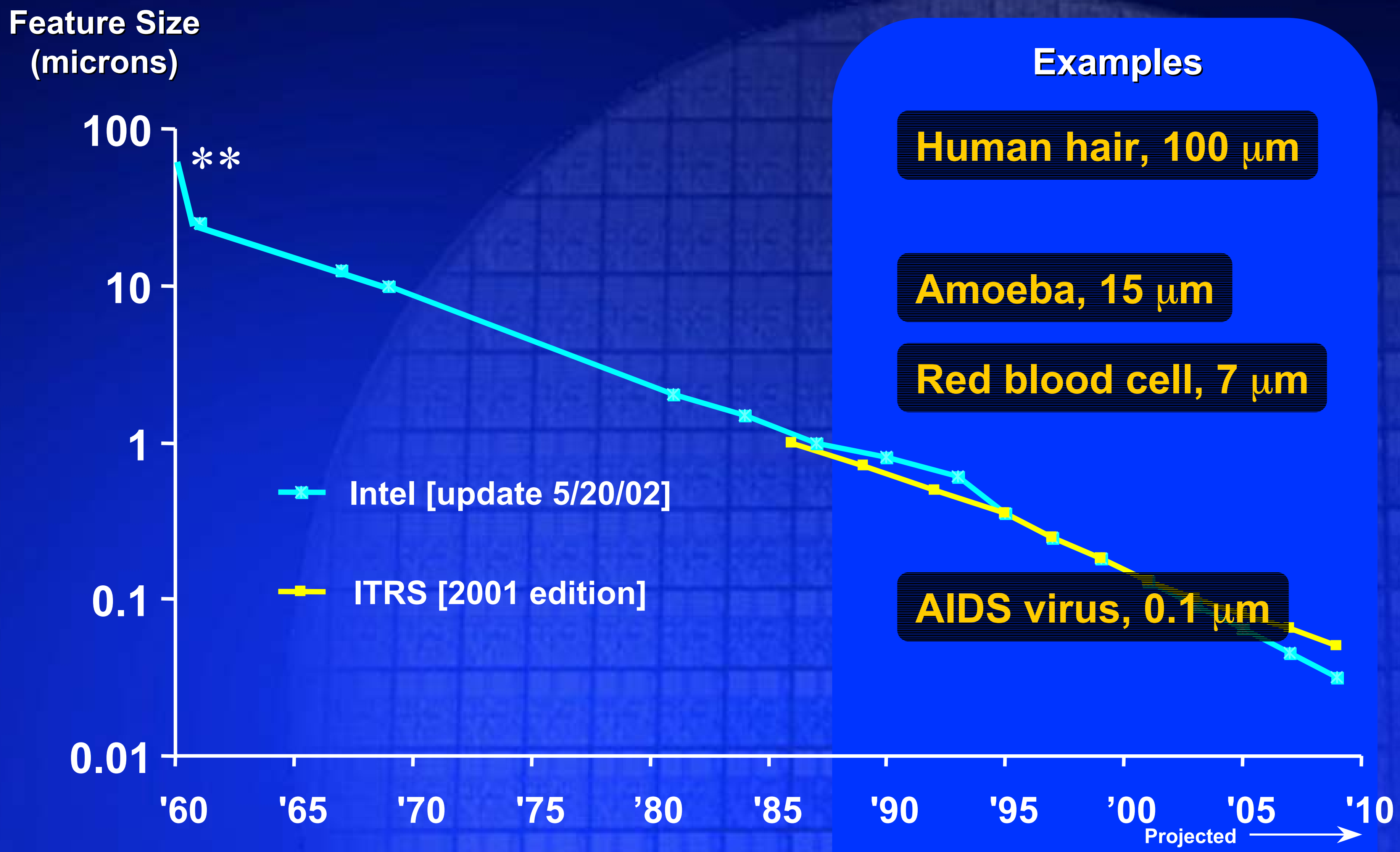


Low-k CDO Dielectric

Copper Interconnects

Combination of copper + low-k dielectric now meeting performance and manufacturing goals

Minimum Feature Size



Examples

Human hair, 100 μm

Amoeba, 15 μm

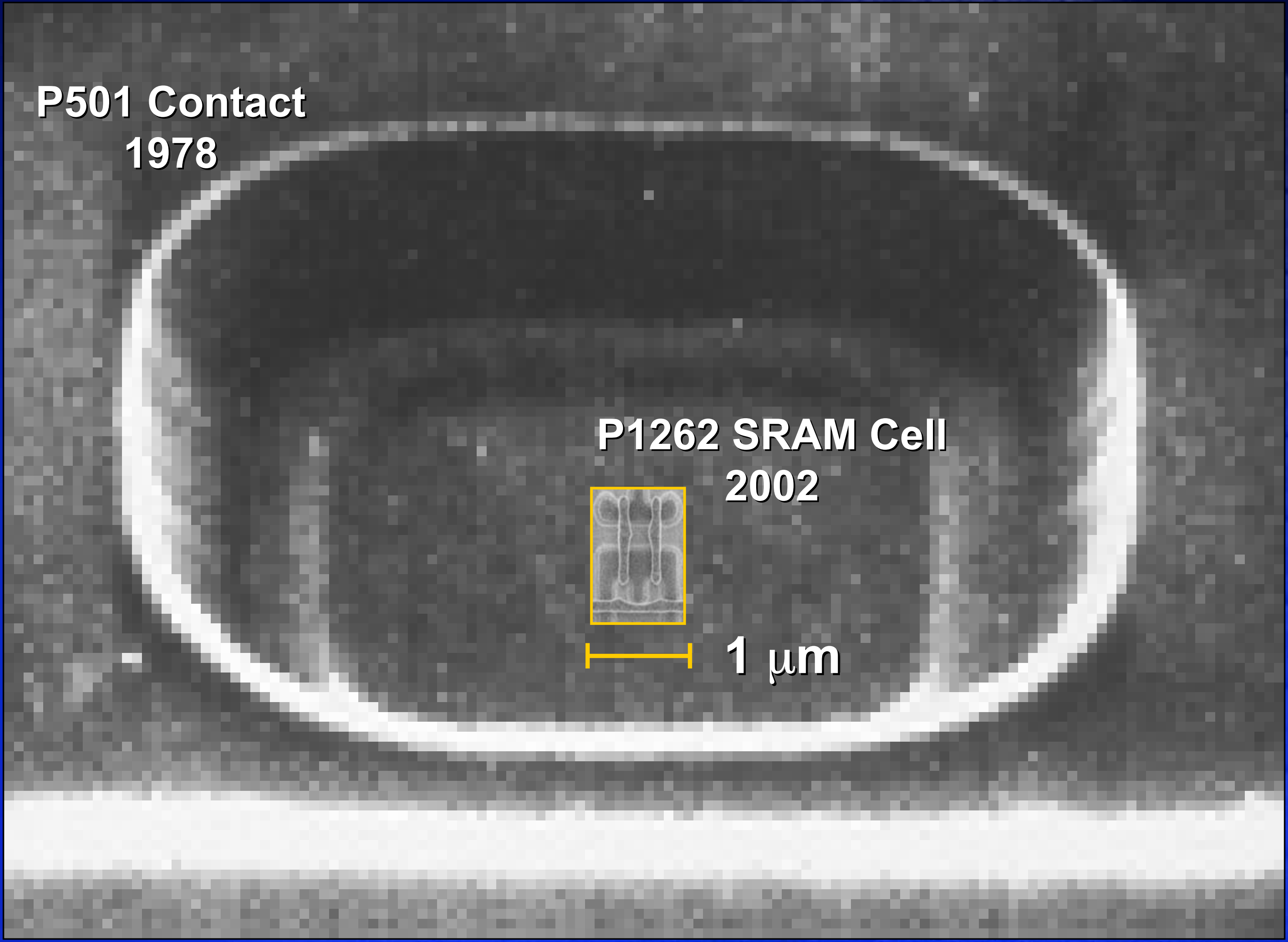
Red blood cell, 7 μm

AIDS virus, 0.1 μm

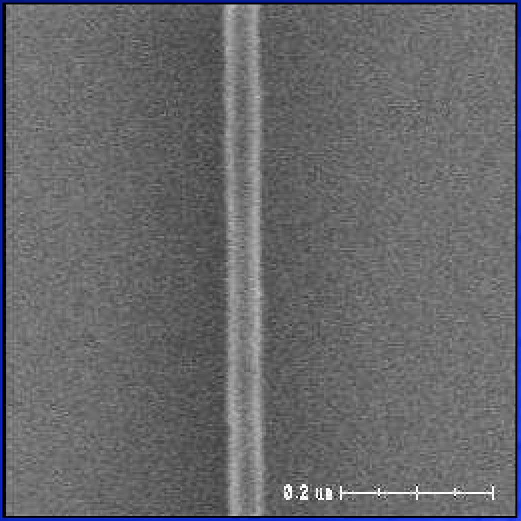
Buckyball, 0.001 μm

** Planar Transistor; remaining data points are ICs.
Source: Intel, post '96 trend data provided by SIA
International Technology Roadmap for Semiconductors (ITRS)
^ [ITRS DRAM Half-Pitch vs. Intel "Lithography"]

1 μm^2 SRAM Cell

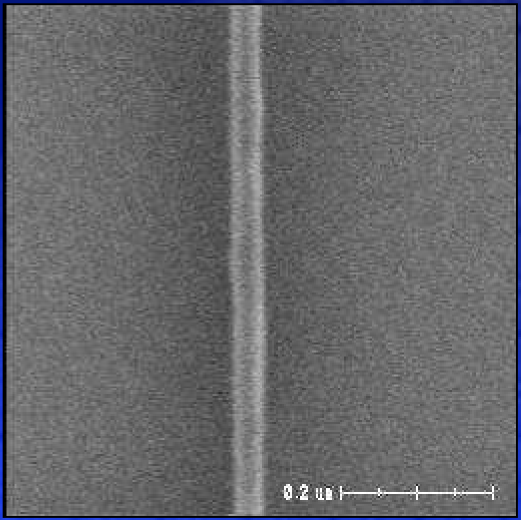


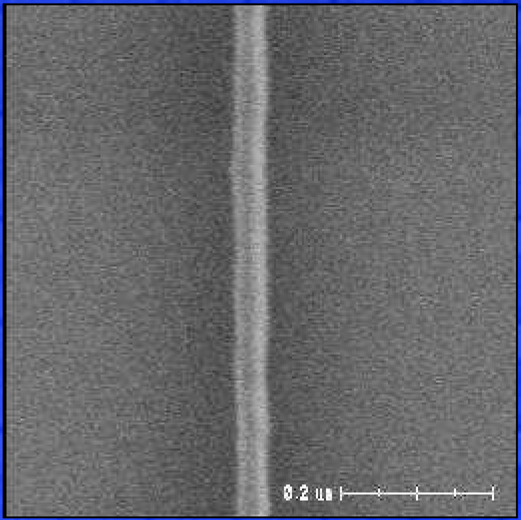
50nm Resist Lines With 193nm Light



“best focus”

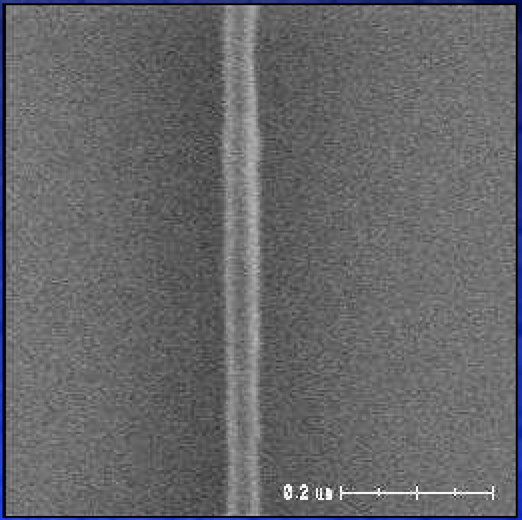
-0.2um focus

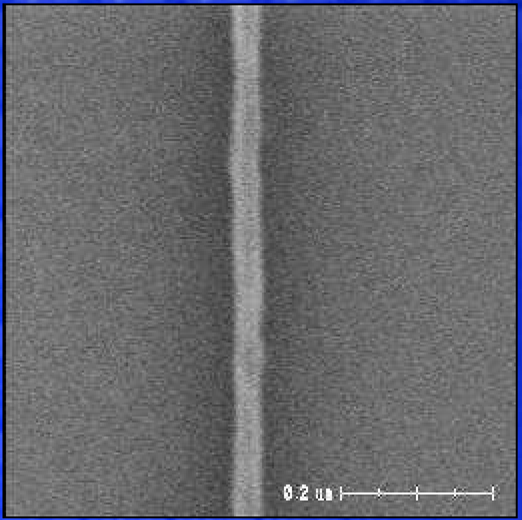




+0.2um focus

-0.3um focus



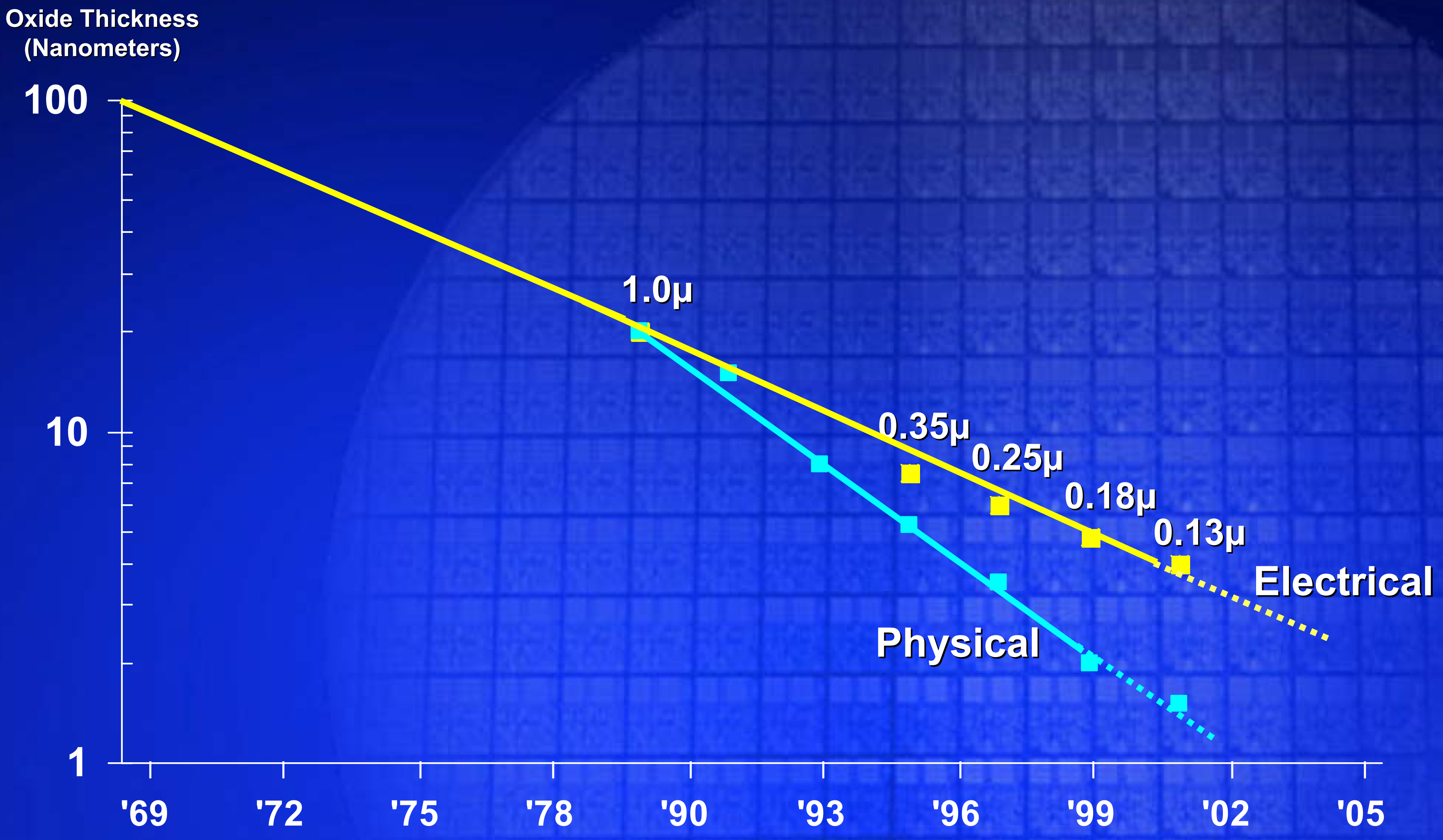


+0.3um focus

193nm Step and Scan Production Tool

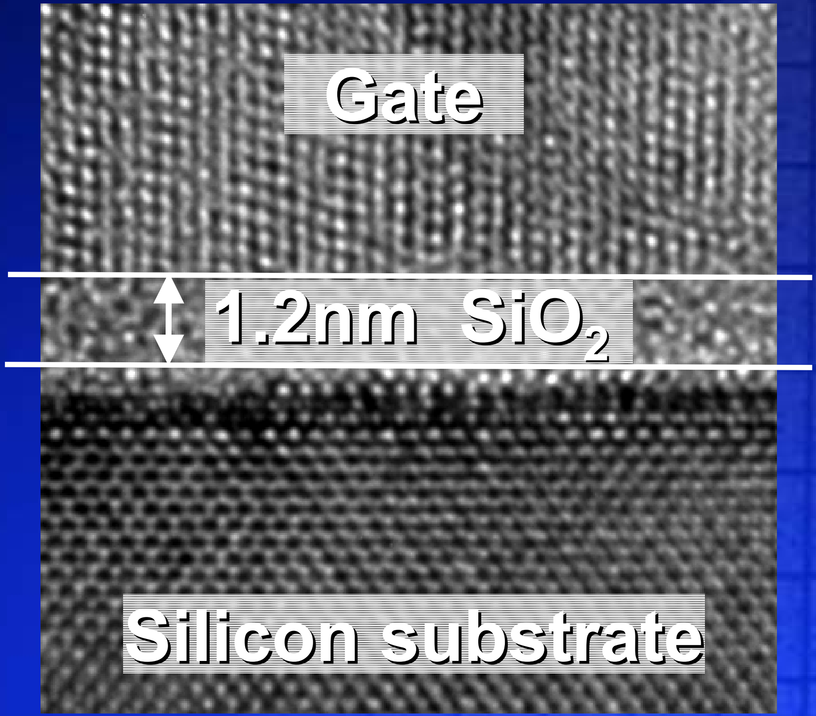
ASML
TWINSCAN
1100

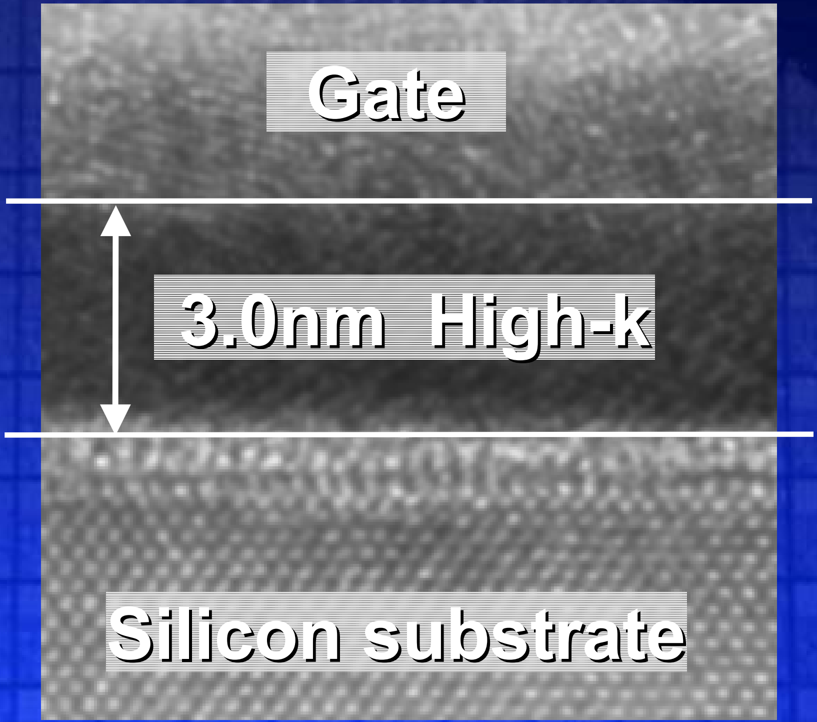
Minimum Insulator Thickness vs Time



Source: Intel

High K for Gate Dielectrics

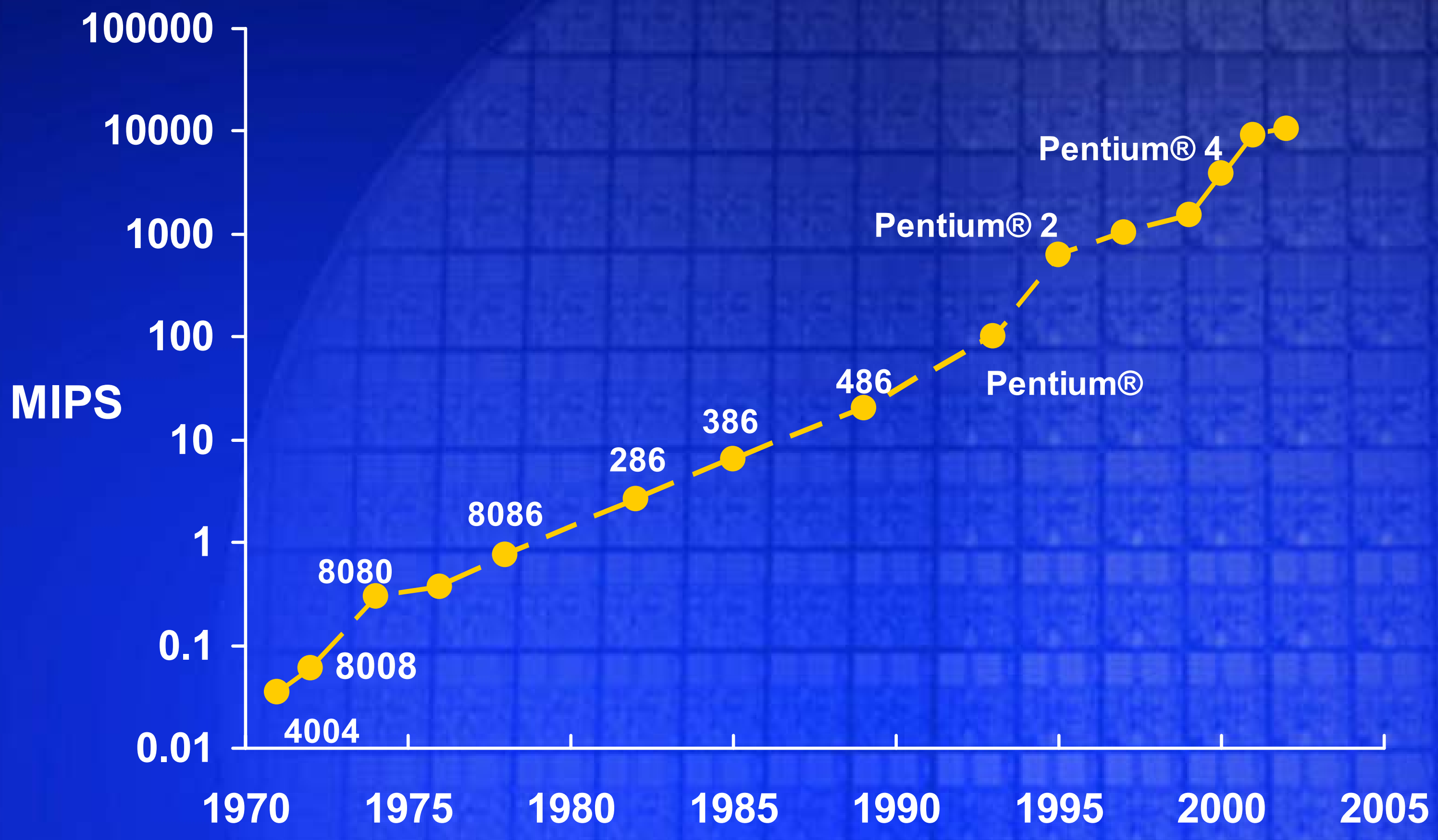




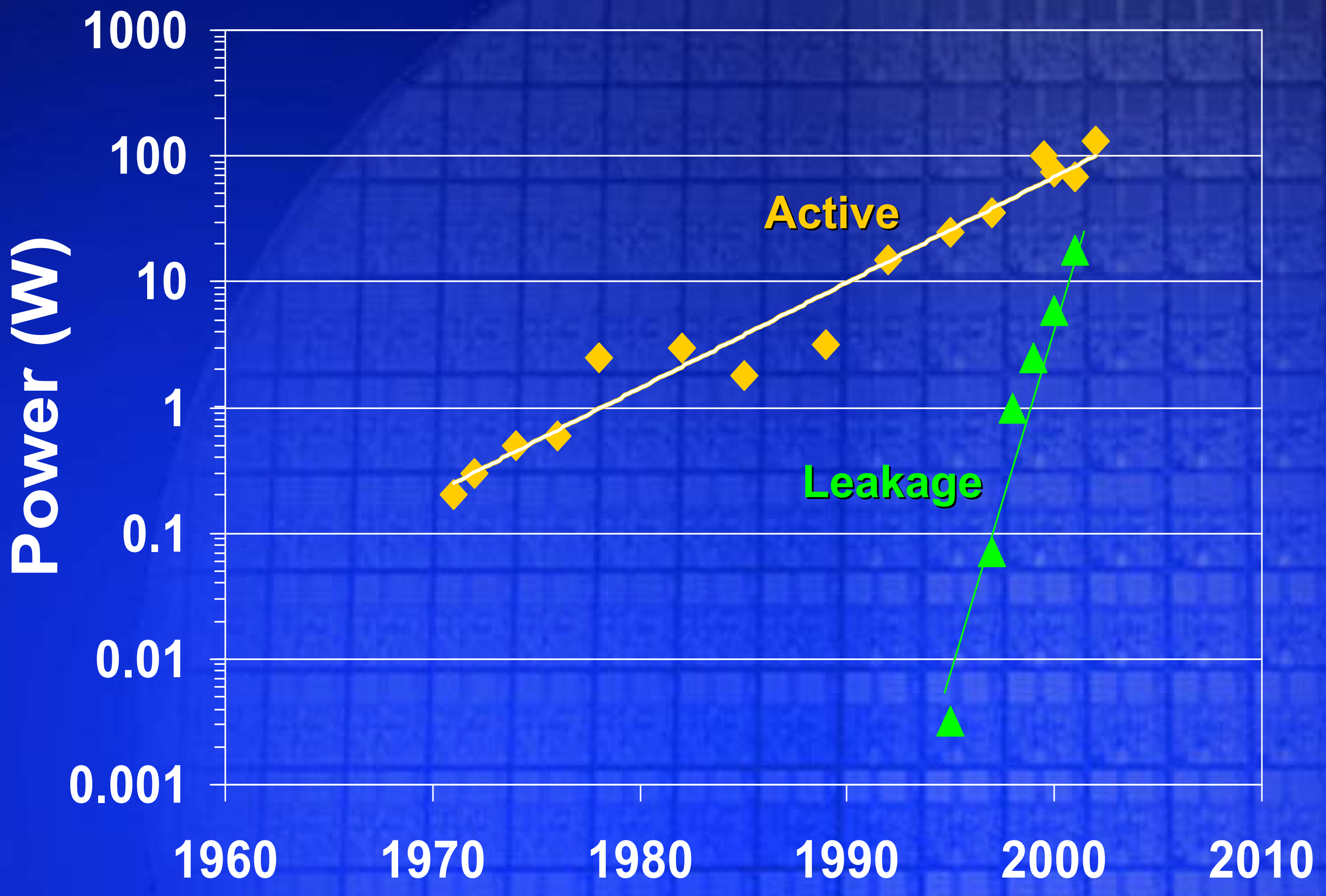
	90nm process	Experimental high-k
Capacitance	1X	1.6X
Leakage	1X	< 0.01X

Source: Intel

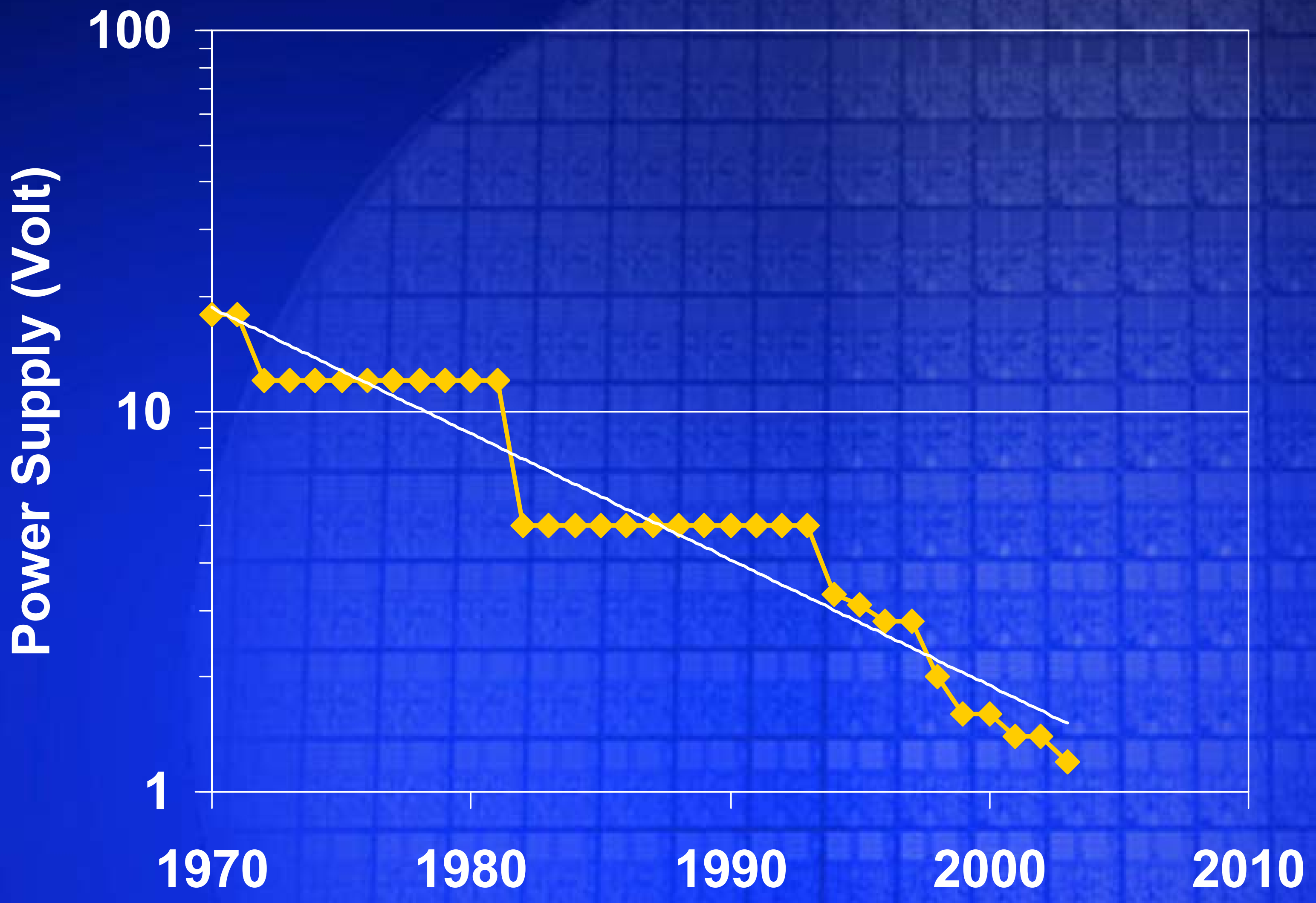
Processor Performance (MIPS)



Processor Power (Watts) - Active & Leakage



Processor Supply Voltage



New Materials and Device Structures Extending Transistor Scaling

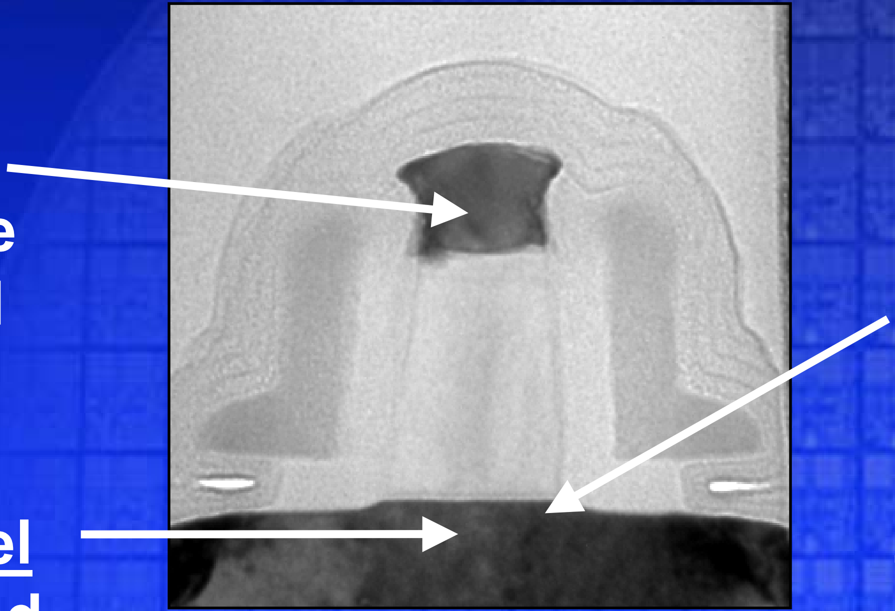
Changes Made

Gate

Silicide Added

Channel

Strained Silicon



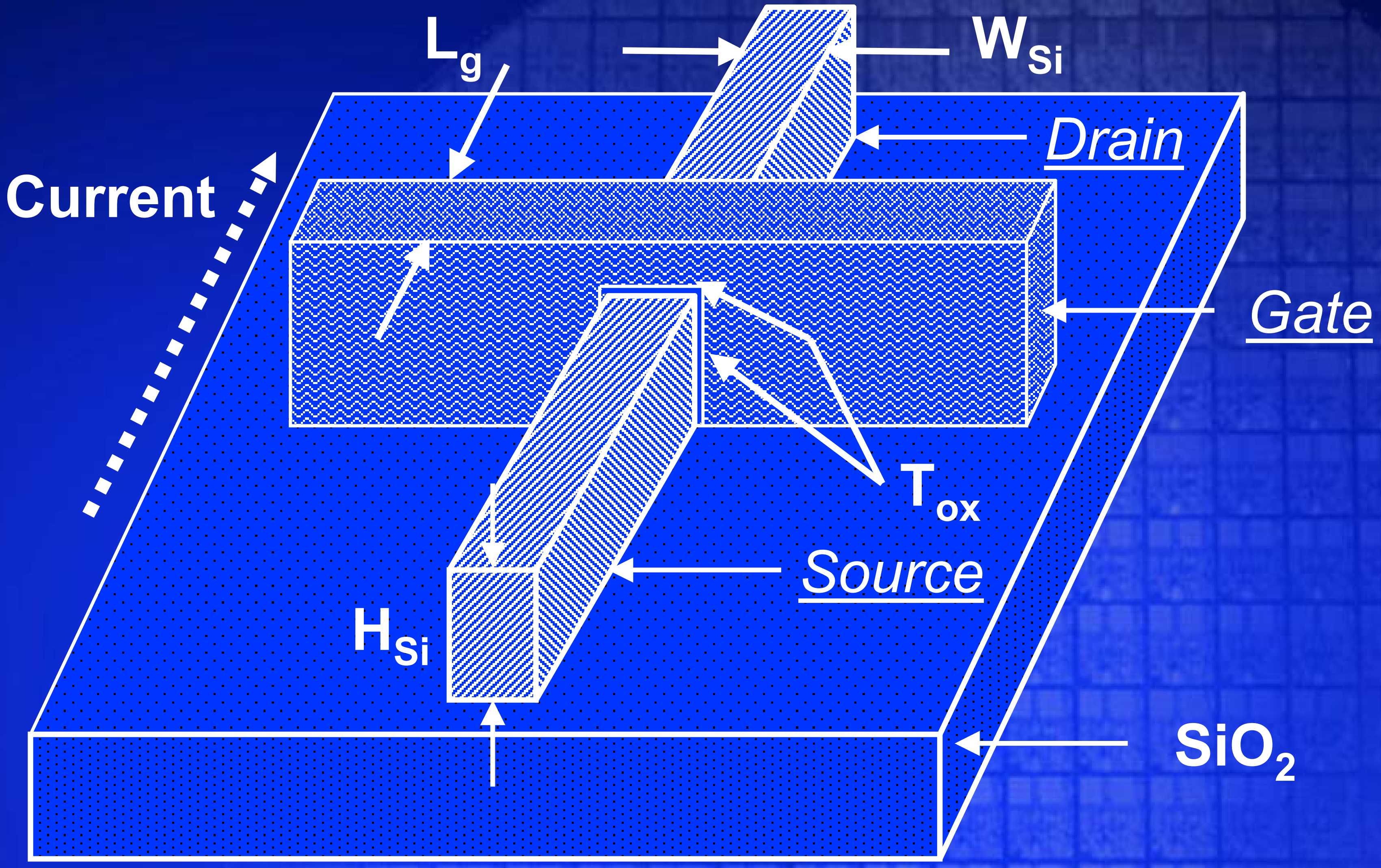
Transistor

Future Options

High-k Gate Dielectric

New Transistor Structure

Tri-Gate Transistor Structure



Technology Generations to Come

Double the Density
Reduce Line Width by 0.7x

130nm→90nm→60nm→45nm→30nm→?

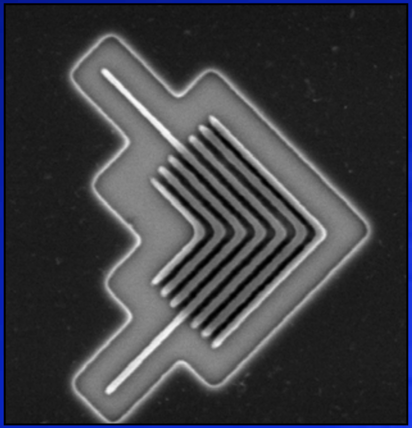
2 or 3 years between generations

∴

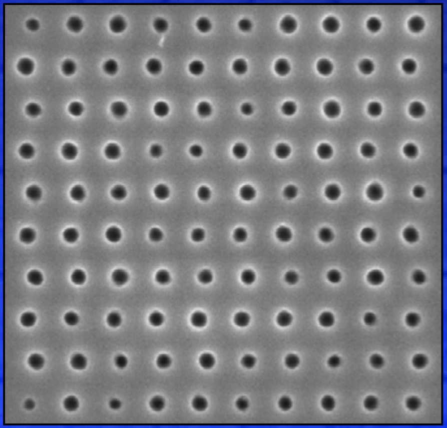
~10 ± 2 Years

EUV Printed and Etched Lines

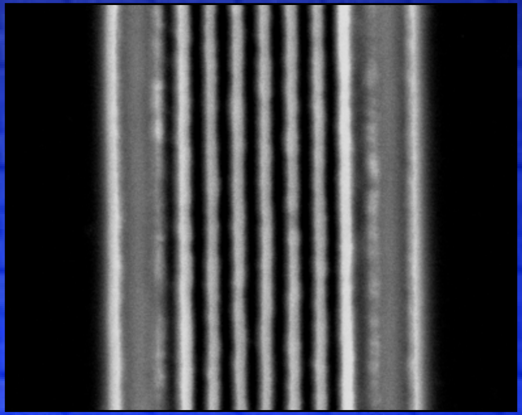
100 nm, k_1 = 0.75



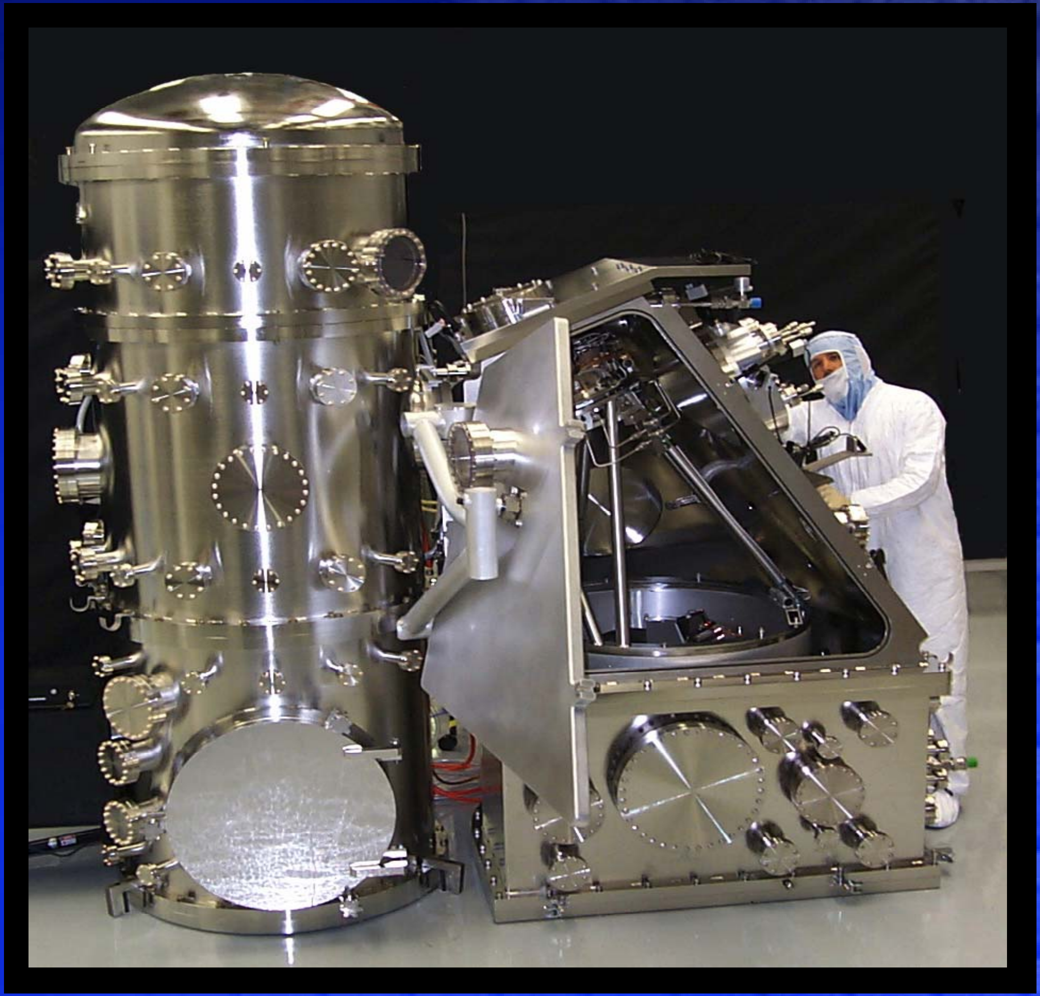
80 nm, k_1 = 0.60



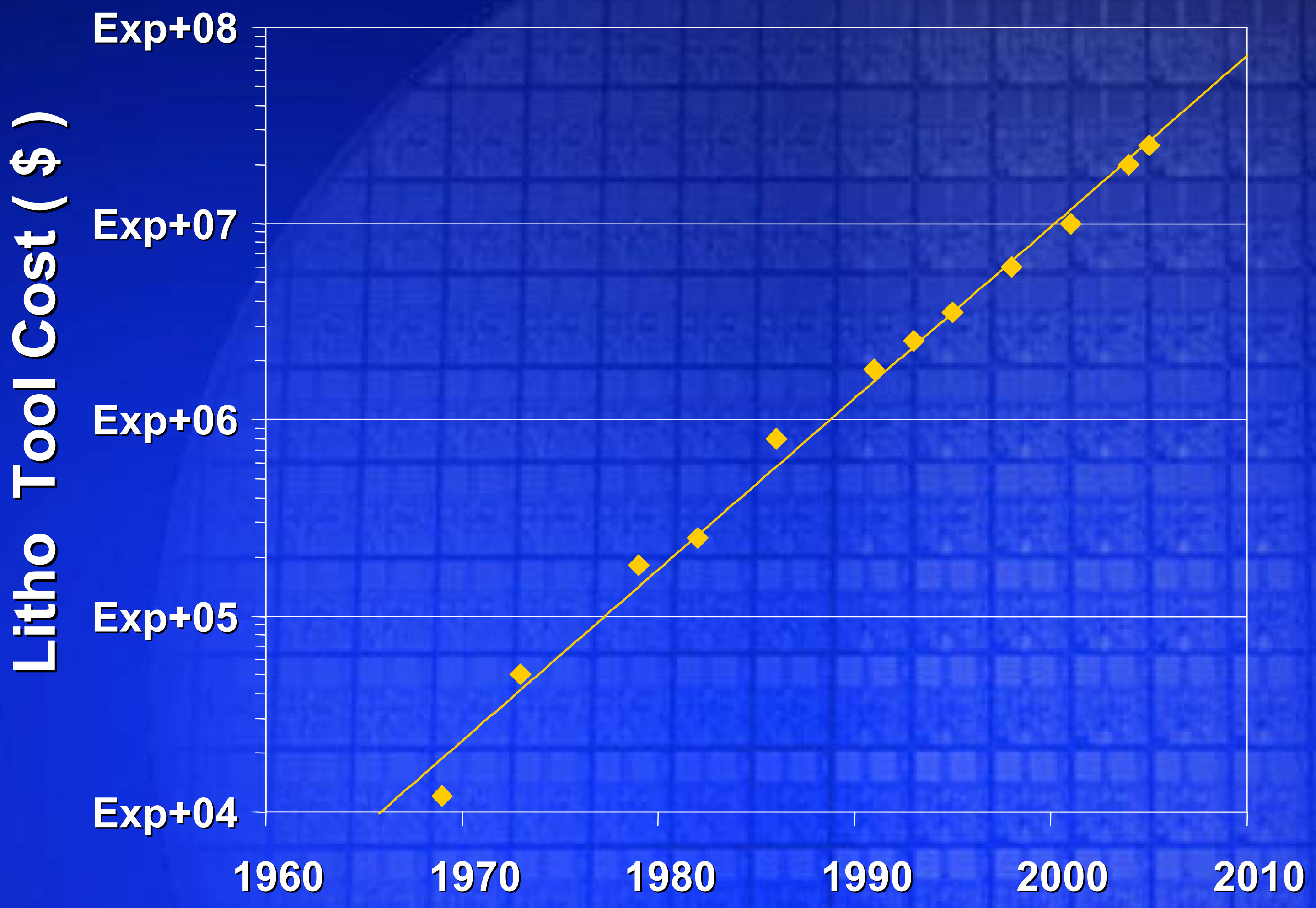
50 nm dense, k_1 = 0.37



Extreme Ultraviolet (EUV) Lithography



Lithography Tool Cost ($)



NO EXPONENTIAL IS FOREVER . . .

BUT

WE CAN DELAY “FOREVER”

2

Classical Magnitudes and Scaling Laws

2.1. Overview

Most physical magnitudes characterizing nanoscale systems differ enormously from those familiar in macroscale systems. Some of these magnitudes can, however, be estimated by applying scaling laws to the values for macroscale systems. Although later chapters seldom use this approach, it can provide orientation, preliminary estimates, and a means for testing whether answers derived by more sophisticated methods are in fact reasonable.

The first of the following sections considers the role of engineering approximations in more detail (Section 2.2); the rest present scaling relationships based on classical continuum models and discuss how those relationships break down as a consequence of atomic-scale structure, mean-free-path effects, and quantum mechanical effects. Section 2.3 discusses mechanical systems, where many scaling laws are quite accurate on the nanoscale. Section 2.4 discusses electromagnetic systems, where many scaling laws fail dramatically on the nanoscale. Section 2.5 discusses thermal systems, where scaling laws have variable accuracy. Finally, Section 2.6 briefly describes how later chapters go beyond these simple models.

2.2. Approximation and classical continuum models

When used with caution, classical continuum models of nanoscale systems can be of substantial value in design and analysis. They represent the simplest level in a hierarchy of approximations of increasing accuracy, complexity, and difficulty.

Experience teaches the value of approximation in design. A typical design process starts with the generation and preliminary evaluation of many options, then selects a few options for further elaboration and evaluation, and finally settles on a detailed specification and analysis of a single preferred design. The first steps entail little commitment to a particular approach. The ease of exploring and comparing many qualitatively distinct approaches is at a premium, and drastic approximations often suffice to screen out the worst options. Even the final design and analysis does not require an exact calculation of physical behavior: approximations and compensating safety margins suffice. Accordingly, a design process can use different approximations at different stages, moving toward greater analytical accuracy and cost.

Approximation is inescapable because the most accurate physical models are computationally intractable. In macromechanical design, engineers employ approximations based on classical mechanics, neglecting quantum mechanics, the thermal excitation of mechanical motions, and the molecular structure of matter. Since macromechanical engineering blends into nanomechanical engineering with no clear line of demarcation, the approximations of macromechanical engineering offer a point of departure for exploring the nanomechanical realm. In some circumstances, these approximations (with a few adaptations) provide an adequate basis for the design and analysis of nanoscale systems. In a broader range of circumstances, they provide an adequate basis for exploring design options and for conducting a preliminary analysis. In a yet broader range of circumstances, they provide a crude description to which one can compare more sophisticated approximations.

2.3. Scaling of classical mechanical systems

Nanomechanical systems are fundamental to molecular manufacturing and are useful in many of its products and processes. The widespread use in chemistry of molecular mechanics approximations together with the classical equations of motion (Sections 3.3, 4.2.3a) indicates the utility of describing nanoscale mechanical systems in terms of classical mechanics. This section describes scaling laws and magnitudes with the added approximation of continuous media.

2.3.1. Basic assumptions

The following discussion considers mechanical systems, neglecting fields and currents. Like later sections, it examines how different physical magnitudes depend on the size of a system (defined by a length parameter L) if all shape parameters and material properties (e.g., strengths, moduli, densities, coefficients of friction) are held constant.

A description of scaling laws must begin with choices that determine the scaling of dynamical variables. A natural choice is that of constant stress. This implies scale-independent °elastic deformation, and hence scale-independent shape; since it results in scale-independent speeds, it also implies constancy of the space-time shapes describing the trajectories of moving parts. Some exemplar calculations are provided, based on material properties like those of diamond (Table 9.1): density $\rho = 3.5 \times 10^3$ kg/m^3; °Young's modulus $E = 10^{12}$ N/m^2; and a low working stress (~0.2 times tensile strength) $\sigma = 10^{10}$ N/m^2. This choice of materials often yields large parameter values (for speeds, accelerations, etc.) relative to those characteristic of more familiar engineering materials.

2.3.2. Magnitudes and scaling

Given constancy of stress and material strength, both the strength of a structure and the force it exerts scale with its cross-sectional area

$$\textit{total strength} \propto \textit{force} \propto \textit{area} \propto L^2 \tag{2.1}$$

Nanoscale devices accordingly exert only small forces: a stress of 10^{10} N/m^2 equals 10^{-8} N/nm^2, or 10 nN/nm^2. Stiffness in °shear, like stretching stiffness, depends on both area and length

$$\textit{shear stiffness} \propto \textit{stretching stiffness} \propto \frac{\textit{area}}{\textit{length}} \propto L \tag{2.2}$$

and varies less rapidly with scale; a cubic nanometer block of $E = 10^{12}$ N/m^2 has a stretching stiffness of 1000 N/m. The bending stiffness of a rod scales in the same way

$$\textit{bending stiffness} \propto \frac{\textit{radius}^4}{\textit{length}^3} \propto L \tag{2.3}$$

Given the above scaling relationships, the magnitude of the deformation under load

$$\textit{deformation} \propto \frac{\textit{force}}{\textit{stiffness}} \propto L \tag{2.4}$$

is proportional to scale, and hence the shape of deformed structures is scale invariant.

The assumption of constant density makes mass scale with volume,

$$\textit{mass} \propto \textit{volume} \propto L^3 \tag{2.5}$$

and the mass of a cubic nanometer block of density $\rho = 3.5 \times 10^3$ kg/m^3 equals 3.5×10^{-24} kg.

The above expressions yield the scaling relationship

$$\textit{acceleration} \propto \frac{\textit{force}}{\textit{mass}} \propto L^{-1} \tag{2.6}$$

A cubic-nanometer mass subject to a net force equaling the above working stress applied to a square nanometer experiences an acceleration of $\sim 3 \times 10^{15}$ m/s^2. Accelerations in nanomechanisms commonly are large by macroscopic standards, but aside from special cases (such as transient acceleration during impact and steady acceleration in a small flywheel) they rarely approach the value just calculated. (Terrestrial gravitational accelerations and stresses usually have negligible effects on nanomechanisms.)

Modulus and density determine the acoustic speed, a scale-independent parameter [along a slim rod, the speed is $(E/\rho)^{1/2}$; in bulk material, somewhat higher]. The vibrational frequencies of a mechanical system are proportional to the acoustic transit time

$$\textit{frequency} \propto \frac{\textit{acoustic speed}}{\textit{length}} \propto L^{-1} \tag{2.7}$$

The acoustic speed in diamond is $\sim 1.75 \times 10^4$ m/s. Some vibrational modes are more conveniently described in terms of lumped parameters of stiffness and mass,

$$\textit{frequency} \propto \sqrt{\frac{\textit{stiffness}}{\textit{mass}}} \propto L^{-1} \tag{2.8}$$

but the scaling relationship is the same. The stiffness and mass associated with a cubic nanometer block yield a vibrational frequency characteristic of a stiff, nanometer-scale object: $[(1000\ \text{N/m})/(3.5 \times 10^{-24}\ \text{kg})]^{1/2} \approx 1.7 \times 10^{13}$ rad/s.

Characteristic times are inversely proportional to characteristic frequencies

$$time \propto frequency^{-1} \propto L \qquad (2.9)$$

The speed of mechanical motions is constrained by strength and density. Its scaling can be derived from the above expressions

$$speed \propto acceleration \cdot time = constant \qquad (2.10)$$

A characteristic speed (only seldom exceeded in practical mechanisms) is that at which a flywheel in the form of a slim hoop is subject to the chosen working stress as a result of its mass and centripetal acceleration. This occurs when $v = (\sigma/\rho)^{1/2} \approx 1.7 \times 10^3$ m/s (with the assumed σ and ρ). Most mechanical motions considered in this volume, however, have speeds between 0.001 and 10 m/s.

The frequencies characteristic of mechanical motions scale with transit times

$$frequency \propto \frac{speed}{length} \propto L^{-1} \qquad (2.11)$$

These frequencies scale in the same manner as vibrational frequencies, hence the assumption of constant stress leaves frequency ratios as scale invariants. At the above characteristic speed, crossing a 1 nm distance takes $\sim 6 \times 10^{-13}$ s; the large speed makes this shorter than the motion times anticipated in typical nanomechanisms. A modest 1 m/s speed, however, still yields a transit time of only 1 ns, indicating that nanomechanisms can operate at frequencies typical of modern micron-scale electronic devices.

The above expressions yield relationships for the scaling of mechanical power

$$power \propto force \cdot speed \propto L^2 \qquad (2.12)$$

and mechanical power density

$$power\ density \propto \frac{power}{volume} \propto L^{-1} \qquad (2.13)$$

A 10 nN force and a 1 nm^3 volume yield a power of 17 μW and a power density of 1.7×10^{22} W/m^3 (at a speed of 1.7×10^3 m/s) or 10 nW and 10^{19} W/m^3 (at a speed of 1 m/s). The combination of strong materials and small devices promises mechanical systems of extraordinarily high power density, even at low speeds (an example of a mechanical power density is the power transmitted by a gear divided by its volume).

Most mechanical systems use bearings to support moving parts. Macromechanical systems frequently use liquid lubricants, but (as noted by Feynman, 1961), this poses problems on a small scale. The above scaling law ordinarily holds speeds and stresses constant, but reducing the thickness of the lubricant layer increases shear rates and hence viscous shear stresses:

$$viscous\ stress\ at\ constant\ speed \propto shear\ rate \propto \frac{speed}{thickness} \propto L^{-1} \qquad (2.14)$$

In Newtonian fluids, shear stress is proportional to shear rate. Molecular simulations indicate that liquids can remain nearly Newtonian at shear rates in excess of 100 m/s across a 1 nm layer (e.g., in the calculations of Ashurst and Hoover, 1975), but they depart from bulk viscosity (or even from liquid behavior) when film thicknesses are less than 10 molecular diameters (Israelachvili, 1992; Schoen et al., 1989), owing to interface-induced alterations in liquid structure. Feynman suggested the use of low-viscosity lubricants (such as kerosene) for micromechanisms (Feynman, 1961); from the perspective of a typical nanomechanism, however, kerosene is better regarded as a collection of bulky molecular objects than as a liquid. If one nonetheless applies the classical approximation to a 1 nm film of low-viscosity fluid ($\eta = 10^{-3}$ N·s/m^2), the viscous shear stress at a speed of 1.7×10^3 m/s is 1.7×10^9 N/m^2; the shear stress at a speed of 1 m/s, 10^6 N/m^2, is still large, dissipating energy at a rate of 1 MW/m^2.

The problems of liquid lubrication motivate consideration of dry bearings (as suggested by Feynman, 1961). Assuming a constant coefficient of friction,

$$\textit{frictional force} \propto \textit{force} \propto L^2 \tag{2.15}$$

and both stresses and speeds are once again scale-independent. The frictional power,

$$\textit{frictional power} \propto \textit{force} \cdot \textit{speed} \propto L^2 \tag{2.16}$$

is proportional to the total power, implying scale-independent mechanical efficiencies. In light of engineering experience, however, the use of dry bearings would seem to present problems (as it has in silicon micromachine research). Without lubrication, efficiencies may be low, and static friction often causes jamming and vibration.

A yet more serious problem for unlubricated systems would seem to be wear. Assuming constant interfacial stresses and speeds (as implied by the above scaling relationships), the anticipated surface erosion rate is independent of scale. Assuming that wear life is determined by the time required to produce a certain fractional change in shape,

$$\textit{wear life} \propto \frac{\textit{thickness}}{\textit{erosion rate}} \propto L \tag{2.17}$$

and a centimeter-scale part having a ten-year lifetime would be expected to have a 30 s lifetime if scaled to nanometer dimensions.

Design and analysis have shown, however, that dry bearings with atomically precise surfaces need not suffer these problems. As shown in Chapters 6, 7, and 10, dynamic friction can be low, and both static friction and wear can be made negligible. The scaling laws applicable to such bearings are compatible with the constant-stress, constant-speed expressions derived previously.

2.3.3. Major corrections

The above scaling relationships treat matter as a continuum with bulk values of strength, modulus, and so forth. They readily yield results for the behavior of iron bars scaled to a length of 10^{-12} m, although such results are meaningless because a single atom of iron is over 10^{-10} m in diameter. They also neglect the influence of surfaces on mechanical properties (Section 9.4), and give (at best) crude estimates regarding small components, in which some dimensions may be only one or a few atomic diameters.

Aside from the molecular structure of matter, major corrections to the results suggested by these scaling laws include uncertainties in position and velocity resulting from statistical and quantum mechanics (examined in detail in Chapter 5). Thermal excitation superimposes random velocities on those intended by the designer. These random velocities depend on scale, such that

$$\textit{thermal speed} \propto \sqrt{\frac{\textit{thermal energy}}{\textit{mass}}} \propto L^{-3/2} \tag{2.18}$$

where the thermal energy measures the characteristic energy in a single degree of freedom, not in the object as a whole. For $\rho = 3.5 \times 10^3$ kg/m^3, the mean thermal speed of a cubic nanometer object at 300 K is ~55 m/s. Random thermal velocities (commonly occurring in vibrational modes) often exceed the velocities imposed by planned operations, and cannot be ignored in analyzing nanomechanical systems.

Quantum mechanical uncertainties in position and momentum are parallel to statistical mechanical uncertainties in their effects on nanomechanical systems. The importance of quantum mechanical effects in vibrating systems depends on the ratio of the characteristic quantum energy ($\hbar\omega$, the quantum of vibrational energy in a °harmonic oscillator of angular frequency ω) and the characteristic thermal energy (kT, the mean energy of a thermally excited harmonic oscillator at a temperature T, if $kT \gg \hbar\omega$). The ratio $\hbar\omega/kT$ varies directly with the frequency of vibration, that is, as L^{-1}. An object of cubic nanometer size with $\omega = 1.7 \times 10^{13}$ rad/s has $\hbar\omega/kT_{300} \approx 0.4$ ($T_{300} = 300$ K; $kT_{300} \approx 4.14$ maJ). The associated quantum mechanical effects (e.g., on positional uncertainty) are smaller than the classical thermal effects, but still significant (see Figure 5.2).

2.4. Scaling of classical electromagnetic systems

2.4.1. Basic assumptions

In considering the scaling of electromagnetic systems, it is convenient to assume that electrostatic field strengths (and hence electrostatic stresses) are independent of scale. With this assumption, the above constant-stress, constant-speed scaling laws for mechanical systems continue to hold for electromechanical systems, so long as magnetic forces are neglected. The onset of strong field-emission currents from conductors limits the electrostatic field strength permissible at the negative electrode of a nanoscale system; values of ~10^9 V/m can readily be tolerated (Section 11.6.2).

2.4.2. Major corrections

Chapter 11 describes several nanometer scale electromechanical systems, requiring consideration of the electrical conductivity of fine wires and of insulating layers thin enough to make tunneling a significant mechanism of electron transport. These phenomena are sometimes (within an expanding range of conditions) understood well enough to permit design calculations.

Corrections to classical continuum models are more important in electromagnetic systems than in mechanical systems: quantum effects, for example, become dominant and at small scales can render classical continuum models useless even as crude approximations. Electromagnetic systems on a nanometer scale commonly have extremely high frequencies, yielding large values of $\hbar\omega/kT_{300}$. Molecules undergoing electronic transitions typically absorb and emit light in the visible to ultraviolet range, rather than the infrared range characteristic of thermal excitation at room temperature. The mass of an electron is less than 10^{-3} that of the lightest atom, hence for comparable confining energy barriers, electron wave functions are more diffuse and permit longer-range tunneling. At high frequencies, the inertial effects of electron mass become significant, but these are neglected in the usual macroscopic expressions for electrical circuits. Accordingly, many of the following classical continuum scaling relationships fail in nanoscale systems. The assumption of scale-independent electrostatic field strengths itself fails in the opposite direction, when scaling up from the nanoscale to the macroscale: the resulting large voltages introduce additional modes of electrical breakdown. In small structures, the discrete size of the electronic charge unit, ~1.6×10^{-19} C, disrupts the smooth scaling of classical electrostatic relationships (Section 11.7.2c).

2.4.3. Magnitudes and scaling: steady-state systems

Given a scale-invariant electrostatic field strength,

$$\textit{voltage} \propto \textit{electrostatic field} \cdot \textit{length} \propto L \tag{2.19}$$

At a field strength of 10^9 V/m, a one nanometer distance yields a 1 V potential difference. A scale-invariant field strength implies a force proportional to area,

$$\textit{electrostatic force} \propto \textit{area} \cdot (\textit{electrostatic field})^2 \propto L^2 \tag{2.20}$$

and a 1 V/nm field between two charged surfaces yields an electrostatic force of ~0.0044 nN/nm^2.

Assuming a constant resistivity,

$$\textit{resistance} \propto \frac{\textit{length}}{\textit{area}} \propto L^{-1} \tag{2.21}$$

and a cubic nanometer block with the resistivity of copper would have a resistance of ~17 Ω. This yields an expression for the scaling of currents,

$$\textit{ohmic current} \propto \frac{\textit{voltage}}{\textit{resistance}} \propto L^2 \tag{2.22}$$

which leaves current density constant. In present microelectronics work, current densities in aluminum interconnections are limited to $<10^{10}$ A/m^2 or less by electromigration, which redistributes metal atoms and eventually interrupts circuit continuity (Mead and Conway, 1980). This current density equals 10 nA/nm^2 (as discussed in Section 11.6.1b, however, present electromigration limits are unlikely to apply to well-designed eutactic conductors).

For field emission into free space, current density depends on surface properties and the electrostatic field intensity, hence

$$\textit{field emission current} \propto \textit{area} \propto L^2 \tag{2.23}$$

and field emission currents scale with ohmic currents. Where surfaces are close enough together for tunneling to occur from conductor to conductor, rather than from conductor to free space, this scaling relationship breaks down.

With constant field strength, electrostatic energy scales with volume:

$$\textit{electrostatic energy} \propto \textit{volume} \cdot (\textit{electrostatic field})^2 \propto L^3 \tag{2.24}$$

A field with a strength of 10^9 V/m has an energy density of ~4.4 maJ per cubic nanometer ($\approx kT_{300}$).

Scaling of capacitance follows from the above,

$$\textit{capacitance} \propto \frac{\textit{electrostatic energy}}{(\textit{voltage})^2} \propto L \tag{2.25}$$

and is independent of assumptions regarding field strength. The calculated capacitance per square nanometer of a vacuum capacitor with parallel plates separated by 1 nm is $\sim 9 \times 10^{-21}$ F; note, however, that electron tunneling causes substantial conduction through an insulating layer this thin.

In electromechanical systems dominated by electrostatic forces,

$$\textit{electrostatic power} \propto \textit{electrostatic force} \cdot \textit{speed} \propto L^2 \tag{2.26}$$

and

$$\textit{electrostatic power density} \propto \frac{\textit{electrostatic power}}{\textit{volume}} \propto L^{-1} \tag{2.27}$$

These scaling laws are identical to those for mechanical power and power density. Like them, they suggest high power densities for small devices (see Section 11.7).

The power density caused by resistive losses scales differently, given the above current density:

$$\textit{resistive power density} \propto (\textit{current density})^2 = \textit{constant} \tag{2.28}$$

The current density needed to power an electrostatic motor, however, scales differently from that derived from a constant-field scaling analysis. In an electrostatic motor, surfaces are charged and discharged with a certain frequency, hence

$$\textit{motor current density} \propto \frac{\textit{charge}}{\textit{area}} \textit{frequency} \propto \textit{field} \cdot \textit{frequency} \propto L^{-1} \tag{2.29}$$

and the resistive power losses climb sharply with decreasing scale:

$$\textit{motor resistive power density} \propto (\textit{motor current density})^2 = L^{-2} \tag{2.30}$$

Accordingly, the efficiency of electrostatic motors decreases with decreasing scale. The absence of long conducting paths (like those in electromagnets) makes resistive losses smaller to begin with, however, and a detailed examination (Section 11.7) shows that efficiencies remain high in absolute terms for motors of submicron scale. The above relationships show that electromechanical systems cannot be scaled in the simple manner suggested for purely mechanical systems, even in the classical continuum approximation.

Electromagnets are far less attractive for nanoscale systems, since

$$\mathit{magnetic\ field} \propto \frac{\mathit{current}}{\mathit{distance}} \propto L \tag{2.31}$$

At a distance of 1 nm from a conductor carrying a current of 10 nA, the field strength is 2×10^{-6} T. The corresponding forces,

$$\mathit{magnetic\ force} \propto \mathit{area} \cdot (\mathit{magnetic\ field})^2 \propto L^4 \tag{2.32}$$

are minute in nanoscale systems: two parallel, 1 nm long segments of conductor, separated by 1 nm and carrying 10 nA, interact with a force of 2×10^{-23} N. This is 14 orders of magnitude smaller than the strength of a typical covalent bond and 11 orders of magnitude smaller than the characteristic electrostatic force just calculated. Magnetic forces between nanoscale current elements are usually negligible. Magnetic fields generated by magnetic materials, in contrast, are independent of scale: forces, energies, and so forth follow the scaling laws described for constant-field electrostatic systems. Nanoscale current elements interacting with fixed magnetic fields can produce more significant (though still small) forces: a 1 nm long segment of conductor carrying a 10 nA current experiences a force of up to 10^{-17} N when immersed in a 1 T field.

The magnetic field energy of a nanoscale current element is small:

$$\mathit{magnetic\ energy} \propto \mathit{volume} \cdot (\mathit{magnetic\ field})^2 \propto L^5 \tag{2.33}$$

The scaling of inductance can be derived from the above, but is independent of assumptions regarding the scaling of currents and magnetic field strengths:

$$\mathit{inductance} \propto \frac{\mathit{magnetic\ energy}}{(\mathit{current})^2} \propto L \tag{2.34}$$

But kinetic inductance important

The inductance per nanometer of length for a fictitious solenoid with a 1 nm^2 cross sectional area and one turn per nanometer of length would be ~10^{-15} h.

2.4.4. Magnitudes and scaling: time-varying systems

In systems with time-varying currents and fields, skin-depth effects increase resistance at high frequencies; these effects complicate scaling relationships and are ignored here. The following simplified relationships are included chiefly to illustrate trends and magnitudes that *preclude* the scaling of classical AC circuits into the nanometer size regime.

For *LR* circuits,

$$\mathit{inductive\ time\ constant} \propto \frac{\mathit{inductance}}{\mathit{resistance}} = L^2 \tag{2.35}$$

Combining the characteristic 17 Ω resistance and 10^{-15} h inductance calculated above yields an *LR* time constant of ~6×10^{-17} s. This time constant is nonphysical: it is, for example, short compared to the electron relaxation time in a typical metal at room temperature (~10^{-14} s). In reality, current decays more slowly because of electron inertia (which has effects broadly similar to those of inductance) and because of the related effect of finite electron relaxation time.

With the approximation of scale-independent resistivity,

$$\textit{capacitative time constant} \propto \textit{resistance} \cdot \textit{capacitance} = \textit{constant} \tag{2.36}$$

This implies that the time required for a capacitor to discharge through a resistor in a pure *RC* circuit is independent of scale; with the scale dependence of the *LR* time constant, however, a structure with fixed proportions can change from a nearly pure *RC* circuit (if built on a small scale) to a nearly pure *LR* circuit (if built on a large scale). The nanometer-scale *RC* time constant indicated by this expression is $(17\ \Omega) \times (9 \times 10^{-21}\ \mathrm{F}) \approx 1.5 \times 10^{-19}$ s, but this result is nonphysical because it neglects the effects of electron inertia and relaxation time.

The *LC* product defines an oscillation frequency

$$\textit{oscillation frequency} \propto \sqrt{\frac{1}{\textit{inductance} \cdot \textit{capacitance}}} \propto L^{-1} \tag{2.37}$$

The characteristic inductance and capacitance calculated above would yield an *LC* circuit with an angular frequency of $\sim 3 \times 10^{17}$ rad/s. Alternatively, in structures such as waveguides,

$$\textit{oscillation frequency} \propto \frac{\textit{wave speed}}{\textit{length}} \propto L^{-1} \tag{2.38}$$

To propagate in a hypothetical waveguide 1 nm in diameter, an electromagnetic wave would require a frequency of $\sim 9 \times 10^{17}$ rad/s or more. Even the lower of the two frequencies just mentioned corresponds to quanta with an energy of ~30 aJ, that is, to photons in the x-ray range with energies of ~200 eV. These frequencies and energies are inconsistent with physical circuits and waveguides (metals are transparent to x-rays, electrons are stripped from molecules at energies well below 200 eV, etc.). Quantum effects and electron inertia make Eq. (2.38) inapplicable in the nanometer range.

Scale also affects the quality of an oscillator:

$$Q \propto \textit{oscillation frequency} \frac{\textit{inductance}}{\textit{resistance}} \propto L \tag{2.39}$$

Since Q is a measure of the damping time relative to the oscillation time, small AC circuits will be heavily damped unless nonclassical effects intervene.

Where the following chapters consider electromagnetic systems at all, they describe systems with currents and fields that are slowly varying by the relevant standards. High-frequency quantum electronic devices, though undoubtedly of great importance, are not discussed here.

2.5. Scaling of classical thermal systems

2.5.1. Basic assumptions

The classical continuum model assumes that volumetric heat capacities and thermal conductivities are independent of scale. Since heat flows in nanomechanical systems are typically a side effect of other physical processes, no independent assumptions are made regarding their scaling laws.

2.5.2. Major corrections

Classical, diffusive models for heat flow in solids can break down in several ways. On sufficiently small scales (which can be macroscopic for crystals at low temperatures) thermal energy is transferred ballistically by °phonons for which the mean free path would, in the absence of bounding surfaces, exceed the dimensions of the structure in question. In the ballistic transport regime, interfacial properties analogous to optical reflectivity and emissivity become significant. Radiative heat flow is altered when the separation of surfaces becomes small compared to the characteristic wavelength of blackbody radiation, owing to coupling of nonradiative electromagnetic modes in the surfaces. In gases, separation of surfaces by less than a mean free path again modifies conductivity. The following assumes classical thermal diffusion, which can be a good approximation for liquids and for solids of low thermal conductivity, even on scales approaching the nanometer range.

2.5.3. Magnitudes and scaling

With a scale-independent volumetric heat capacity,

$$\textit{heat capacity} \propto \textit{volume} \propto L^3 \tag{2.40}$$

A cubic nanometer volume of a material with a (typical) volumetric heat capacity of 10^6 J/m^3·K has a heat capacity of 1 maJ/K.

Thermal conductance scales like electrical conductance, with

$$\textit{thermal conductance} \propto \frac{\textit{area}}{\textit{length}} \propto L \tag{2.41}$$

and a cubic nanometer of material with a (fairly typical) thermal conductivity of 10 W/m·K has a thermal conductance of 10^{-8} W/K.

Characteristic times for thermal equilibration follow from these relationships, yielding

$$\textit{thermal time constant} \propto \frac{\textit{heat capacity}}{\textit{thermal conductance}} \propto L^2 \tag{2.42}$$

For a cubic nanometer block separated from a heat sink by a thermal path with a conductance of 10^{-8} W/K, the calculated thermal time constant is ~10^{-13} s, which is comparable to the acoustic transit time. (In an insulator, a calculated thermal time constant approaching the acoustic transit time signals a breakdown of the diffusive model for transport of thermal energy and the need for a model accounting for ballistic transport; in the fully ballistic regime, time constants scale in proportion to L, and thermal energy moves at the speed of sound.)

The scaling relationship for frictional power dissipation, Eq. (2.16), implies a scaling law for the temperature elevation of a device in thermal contact with its environment,

$$\textit{temperature elevation} \propto \frac{\textit{frictional power}}{\textit{thermal conductance}} \propto L \tag{2.43}$$

This indicates that nanomechanical systems are more nearly isothermal than analogous systems of macroscopic size.

Table 2.1. Summary of classical continuum scaling laws.

Physical quantity	Scaling exponent	Typical magnitude	Scaling accuracy
Area	2	10^{-18} m^2	Definitional
Force, strength	2	10^{-8} N/m^2	Good
Stiffness	1	1000 N/m	Good
Deformation	1	10^{-11} m	Good
Mass	3	10^{-24} kg	Good
Acceleration	−1	10^{15} m/s^2	Good
Vibrational frequency	−1	10^{13} rad/s	Good
Stress-limited speed	0	10^{3} m/s	Good
Motion time	−1	10^{-12} to 10^{-9} s	Good
Power	2	10^{-8} W	Good
Power density	−1	10^{19} W/m^3	Good
Viscous stress	−1	10^{6} N/m^2	Moderate to poor
Frictional force	2	(see Ch. 10)	Moderate to inapplicable
Wear life	1	(see Ch. 6, 10)	Moderate to inapplicable
Thermal speed	−3/2	100 m/s	Good
Voltage	1	1 V	Good at small scales
Electrostatic force	2	10^{-12} N	Good at small scales
Resistance	−1	10 Ω	Moderate to poor
Current	2	10^{-8} A	Moderate to poor
Electrostatic energy	3	10^{-21} J	Good at small scales
Capacitance	1	10^{-20} F	Good
Magnetic field	1	10^{-6} T	Good
Magnetic force	4	10^{-23} N	Good
Inductance	1	10^{-15} h	Good
Inductive time constant	2	$<10^{-16}$ s	Bad[a]
Capacitive time constant	0	—	Moderate to poor[b]
Elect. oscill. frequency	−1	$>10^{18}$ rad/s	Bad[a]
Oscillator Q	1	—	Moderate to poor[b]
Heat capacity	3	10^{-21} J/K	Good
Thermal conductance	1	10^{-8} W/K	Good to poor
Thermal time constant	2	10^{-13} s	Good to poor

[a] Values included only to illustrate the failure of the scaling law.

[b] Values omitted; realistic geometries would require several arbitrary parameters.

2.6. Beyond classical continuum models

This chapter has described the scaling laws implied by classical continuum models for mechanical, electromagnetic, and thermal systems, together with the magnitudes they suggest for the physical parameters of nanometer scale systems. It has also considered limits to the validity of these models, imposed by statistical mechanics, quantum mechanics, the molecular structure of matter, and so forth. Different classical models fail at different length scales, with the most dramatic failures appearing in AC electrical circuits.

The following chapters go beyond classical continuum models. Chapters 3 and 4 examine models of molecular structure, dynamics, and statistical mechanics from a nanomechanical systems perspective. Chapters 5 and 6 examine the combined effects of quantum and statistical mechanics on nanomechanical systems, first analyzing positional uncertainty in systems subject to a restoring force, and then analyzing the rates of transitions, errors, and damage in systems that can settle in alternative states. Chapter 7 examines mechanisms for energy dissipation. These chapters provide a foundation for analyzing specific nanomechanical systems. Later chapters examine not only nanomechanical systems, but a few specific electrical and fluid systems; where analysis of the latter must go beyond classical continuum approximations, the needed principles are discussed in context.

2.7. Conclusions

The accuracy of classical continuum models and scaling laws to nanoscale systems depends on the physical phenomena considered. It is low for electromagnetic systems with small calculated time constants, reasonably good for thermal systems and slowly varying electromagnetic systems, and often excellent for purely mechanical systems, provided that the component dimensions substantially exceed atomic dimensions. Scaling principles indicate that mechanical components can operate at high frequencies, accelerations, and power densities. The adverse scaling of wear lifetimes suggests that bearings are a special concern. Later chapters support these expectations regarding frequency, acceleration, and power density; Chapter 10 describes suitable bearings. Table 2.1 summarizes many of the relationships discussed in this chapter.

EECS 217C 1.20

and have suggested
errata necessary. In
ions and changes as
e remains far more
e, but the additions
ew additions, such
ontinue to use this
urth-year electrical
olled are first-year
.
ica Vasileska, who
helpful in pointing
der insight into the

David K Ferry
Tempe, AZ, 2000

Chapter 1

Waves and particles

1.1 Introduction

Science has developed through a variety of investigations more or less over the time scale of human existence. On this scale, quantum mechanics is a very young field, existing essentially only since the beginning of this century. Even our understanding of classical mechanics has existed for a comparatively long period—roughly having been formalized with Newton's equations published in his *Principia Mathematica*, in April 1686. In fact, we have just celebrated more than 300 years of classical mechanics.

In contrast with this, the ideas of quantum mechanics are barely more than a century old. They had their first beginnings in the 1890s with Planck's development of a theory for black-body radiation. This form of radiation is emitted by all bodies according to their temperature. However, before Planck, there were two competing views. In one, the low-frequency view, this radiation increased as a power of the frequency, which led to a problem at very high frequencies. In the other, the high-frequency view, the radiation decreased rapidly with frequency, which led to a problem at low frequencies. Planck unified these views through the development of what is now known as the Planck black-body radiation law:

$$I(f)\,\mathrm{d}f \sim \frac{f^3}{\exp\left(\frac{hf}{k_\mathrm{B}T}\right)-1}\,\mathrm{d}f \tag{1.1}$$

where f is the frequency, T is the temperature, I is the intensity of radiation, and k_B is Boltzmann's constant (1.38×10^{-23} J K^{-1}). In order to achieve this result, Planck had to assume that matter radiated and absorbed energy in small, but non-zero quantities whose energy was defined by

$$E = hf \tag{1.2}$$

where h is now known as Planck's constant, given by 6.62×10^{-23} J s. While Planck had given us the idea of quanta of energy, he was not comfortable with this idea, but it took only a decade for Einstein's theory of the photoelectric effect (discussed later) to confirm that radiation indeed was composed of quantum particles of energy given by (1.2). Shortly after this, Bohr developed his quantum model of the atom, in which the electrons existed in discrete shells with well defined energy levels. In this way, he could explain the discrete absorption and emission lines that were seen in experimental atomic spectroscopy. While his model was developed in a somewhat *ad hoc* manner, the ideas proved correct, although the mathematical details were changed when the more formal quantum theory arrived in 1927 from Heisenberg and Schrödinger. The work of these two latter pioneers led to different, but equivalent, formulations of the quantum principles that we know to be important in modern physics. Finally, another essential concept was introduced by de Broglie. While we have assigned particle-like properties to light waves earlier, de Broglie asserted that particles, like electrons, should have wavelike properties in which their wavelength is related to their momentum by

$$\lambda = \frac{h}{p} = \frac{h}{mv}. \tag{1.3}$$

λ is now referred to as the de Broglie wavelength of the particle.

Today, there is a consensus (but not a complete agreement) as to the general understanding of the quantum principles. In essence, quantum mechanics is the mathematical description of physical systems with non-commuting operators; for example, the ordering of the operators is very important. The engineer is familiar with such an ordering dependence through the use of matrix algebra, where in general the order of two matrices is important; that is $AB \neq BA$. In quantum mechanics, the ordering of various *operators* is important, and it is these operators that do not commute. There are two additional, and quite important, postulates. These are *complementarity* and the *correspondence principle*.

Complementarity refers to the duality of waves and particles. That is, for both electrons and light waves, there is a duality between a treatment in terms of waves and a treatment in terms of particles. The wave treatment generally is described by a field theory with the corresponding operator effects introduced into the wave amplitudes. The particle is treated in a manner similar to the classical particle dynamics treatment with the appropriate operators properly introduced. In the next two sections, we will investigate two of the operator effects.

On the other hand, the *correspondence principle* relates to the limiting approach to the well known classical mechanics. It will be found that Planck's constant, h, appears in all results that truly reflect quantum mechanical behaviour. As we allow $h \to 0$, the classical results must be obtained. That is, the true quantum effects must vanish as we take this limit. Now, we really do not vary the value of such a fundamental constant, but the correspondence principle asserts that if we were to do so, the classical results would be recovered. What this means is that the quantum effects are modifications of the classical properties. These effects may be small or large, depending upon a number of factors such as time scales, size scales and energy scales. The value of Planck's constant is quite

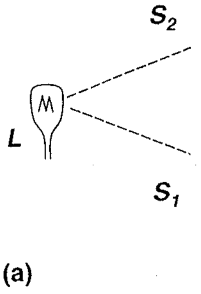


Figure 1.1. In pane
through the two slit
the right. If we bloc
through S_2 on the 's

small, 6.625 × 1
are small. For ex
metal–oxide–sem
of devices such a

Before proc
we create a sour
two slits, the wav
shown in figure 1
just a single slit, t
light waves. It is
paths so as to cr
property of the s
said that we are
properties, we tu

1.2 Light a

One of the more
of the photoelec
surface of a met
from the surface
high. The curi
only upon the w
radiation. In fa
wavelength of th
does depend up
Today, of cour

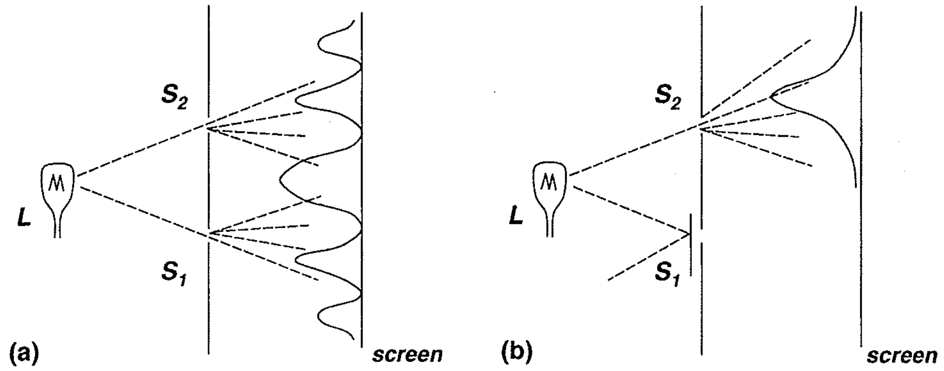


Figure 1.1. In panel (*a*), we illustrate how light coming from the source L and passing through the two slits S_1 and S_2 interferes to cause the pattern indicated on the 'screen' on the right. If we block one of the slits, say S_1, then we obtain only the light intensity passing through S_2 on the 'screen' as shown in panel (*b*).

small, 6.625×10^{-34} J s, but one should not assume that the quantum effects are small. For example, quantization is found to affect the operation of modern metal–oxide–semiconductor (MOS) transistors and to be the fundamental property of devices such as a tunnel diode.

Before proceeding, let us examine an important aspect of light as a wave. If we create a source of coherent light (a single frequency), and pass this through two slits, the wavelike property of the light will create an interference pattern, as shown in figure 1.1. Now, if we block one of the slits, so that light passes through just a single slit, this pattern disappears, and we see just the normal passage of the light waves. It is this interference between the light, passing through two different paths so as to create two different phases of the light wave, that is an essential property of the single wave. When we can see such an interference pattern, it is said that we are seeing the wavelike properties of light. To see the particle-like properties, we turn to the photoelectric effect.

1.2 Light as particles—the photoelectric effect

One of the more interesting examples of the principle of complementarity is that of the photoelectric effect. It was known that when light was shone upon the surface of a metal, or some other conducting medium, electrons could be emitted from the surface provided that the frequency of the incident light was sufficiently high. The curious effect is that the velocity of the emitted electrons depends only upon the wavelength of the incident light, and *not upon the intensity of the radiation*. In fact, the energy of the emitted particles varies inversely with the wavelength of the light waves. On the other hand, the *number* of emitted electrons does depend upon the intensity of the radiation, and not upon its wavelength. Today, of course, we do not consider this surprising at all, but this is after it has been explained in the Nobel-prize-winning work of Einstein. What Einstein concluded was that the explanation of this phenomenon required a treatment of light in terms of its 'corpuscular' nature; that is, we need to treat the light wave as a beam of particles impinging upon the surface of the metal. In fact, it is important to describe the energy of the individual light particles, which we call *photons*, using the relation (1.2) (Einstein 1905)

$$\mathcal{E} = h\nu = \hbar\omega \tag{1.2$'$}$$

where $\hbar = h/2\pi$. The photoelectric effect can be understood through consideration of figure 1.2. However, it is essential to understand that we are talking about the flow of 'particles' as directly corresponding to the wave intensity of the light wave. Where the intensity is 'high', there is a high density of photons. Conversely, where the wave amplitude is weak, there is a low density of photons.

A metal is characterized by a work function $\mathcal{E}_{\mathrm{W}}$, which is the energy required to raise an electron from the Fermi energy to the vacuum level, from which it can be emitted from the surface. Thus, in order to observe the photoelectric effect, or photoemission as it is now called, it is necessary to have the energy of the photons greater than the work function, or $\mathcal{E} > \mathcal{E}_{\mathrm{W}}$. The excess energy, that is the energy difference between that of the photon and the work function, becomes the kinetic energy of the emitted particle. Since the frequency of the photon is inversely proportional to the wavelength, the kinetic energy of the emitted particle varies inversely as the wavelength of the light. As the intensity of the light wave is increased, the number of incident photons increases, and therefore the number of emitted electrons increases. However, the momentum of each emitted electron depends upon the properties of a single photon, and therefore is independent of the intensity of the light wave.

A corollary of the acceptance of light as particles is that there is a momentum associated with each of the particles. It is well known in field theory that there is a momentum associated with the (massless) wave, which is given by $p = h\nu/c$, which leads immediately to the relationship (1.3) given earlier

$$p = \frac{h\nu}{c} = \frac{h}{\lambda}. \tag{1.3$'$}$$

Here, we have used the magnitude, rather than the vector description, of the momentum. It then follows that

$$p = \frac{h}{\lambda} = \hbar k \tag{1.4}$$

a relationship that is familiar both to those accustomed to field theory and to those familiar with solid-state theory.

It is finally clear from the interpretation of light waves as particles that there exists a relationship between the 'particle' energy and the frequency of the wave, and a connection between the momentum of the 'particle' and the wavelength

Figure 1.2. The e
energy greater thai
Fermi energy, $\mathcal{E}_{\mathrm{F}}$,

of the wave. Th
form of (1.3′) h
corresponding t
wavelength. Ho
of equations (1.
contribution wa
must possess th
able to incorpo
terms of the mo

1.3 Electro

In the previou
appropriate, an
particles, whic
when it is clear
In the correspo
the varying int
particles; the p
the wave at this
mechanics des
superposition.

stein. What Einstein
uired a treatment of
treat the light wave
metal. In fact, it is
icles, which we call

(1.2′)

ınderstood through
.erstand that we are
:o the wave intensity
ı density of photons.
density of photons.
the energy required
l, from which it can
ɔtoelectric effect, or
ıergy of the photons
y, that is the energy
ɔecomes the kinetic
photon is inversely
itted particle varies
ˈf the light wave is
ɘfore the number of
ɔh emitted electron
e is independent of

ıere is a momentum
theory that there is
iven by $p = h\nu/c$,

(1.3′)

description, of the

(1.4)

:heory and to those

particles that there
ıency of the wave,
ıd the wavelength

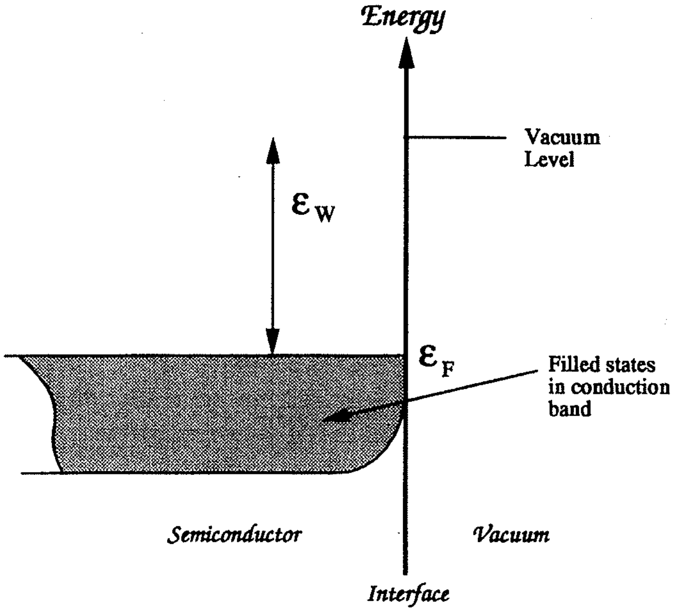


Figure 1.2. The energy bands for the surface of a metal. An incident photon with an energy greater than the work function, $\mathcal{E}_W$, can cause an electron to be raised from the Fermi energy, $\mathcal{E}_F$, to above the vacuum level, whereby it can be photoemitted.

of the wave. The two equations (1.2′) and (1.3′) give these relationships. The form of (1.3′) has usually been associated with de Broglie, and the wavelength corresponding to the particle momentum is usually described as the *de Broglie wavelength*. However, it is worth noting that de Broglie (1939) referred to the set of equations (1.2′) and (1.3′) as the Einstein relations! In fact, de Broglie's great contribution was the recognition that atoms localized in orbits about a nucleus must possess these same wavelike properties. Hence, the electron orbit must be able to incorporate an exact integer number of wavelengths, given by (1.3′) in terms of the momentum. This then leads to quantization of the energy levels.

1.3 Electrons as waves

In the previous section, we discussed how in many cases it is clearly more appropriate, and indeed necessary, to treat electromagnetic waves as the flow of particles, which in turn are termed photons. By the same token, there are times when it is clearly advantageous to describe particles, such as electrons, as waves. In the correspondence between these two viewpoints, it is important to note that the varying intensity of the wave reflects the presence of a varying number of particles; the particle density at a point x, at time t, reflects the varying intensity of the wave at this point and time. For this to be the case, it is important that quantum mechanics describe both the wave and particle pictures through the principle of superposition. That is, the amplitude of the composite wave is related to the sum of the amplitudes of the individual waves corresponding to each of the particles present. Note that it is the amplitudes, and not the intensities, that are summed, so there arises the real possibility for *interference* between the waves of individual particles. Thus, for the presence of two (non-interacting) particles at a point x, at time t, we may write the composite wave function as

$$\Psi(x,t) = \Psi_1(x,t) + \Psi_2(x,t). \tag{1.5}$$

This composite wave may be described as a *probability wave*, in that the square of the magnitude describes the probability of finding an electron at a point.

It may be noted from (1.4) that the momentum of the particles goes immediately into the so-called *wave vector* $\boldsymbol{k}$ of the wave. A special form of (1.5) is

$$\Psi(x,t) = A\mathrm{e}^{\mathrm{i}(k_1 x - \omega t)} + B\mathrm{e}^{\mathrm{i}(k_2 x - \omega t)} \tag{1.6}$$

where it has been assumed that the two components may have different momenta (but we have taken the energies equal). For the moment, the time-independent steady state will be considered, so the time-varying parts of (1.6) will be suppressed as we will talk only about steady-state results of phase interference. It is known, for example, that a time-varying magnetic field that is enclosed by a conducting loop will induce an electric field (and voltage) in the loop through Faraday's law. Can this happen for a time-independent magnetic field? The classical answer is, of course, no, and Maxwell's equations give us this answer. But do they in the quantum case where we can have the interference between the two waves corresponding to two separate electrons?

For the experiment, we consider a loop of wire. Specifically, the loop is made of Au wire deposited on a Si_3N_4 substrate. Such a loop is shown in figure 1.3, where the loop is about 820 nm in diameter, and the Au lines are 40 nm wide (Webb *et al* 1985). The loop is connected to an external circuit through Au leads (also shown), and a magnetic field is threaded through the loop.

To understand the phase interference, we proceed by assuming that the electron waves enter the ring at a point described by $\phi = -\pi$. For the moment, assume that the field induces an electric field in the ring (the time variation will in the end cancel out, and it is not the electric field *per se* that causes the effect, but this approach allows us to describe the effect). Then, for one electron passing through the upper side of the ring, the electron is accelerated by the field, as it moves *with* the field, while on the other side of the ring the electron is decelerated by the field as it moves *against* the field. The field enters through Newton's law, and

$$k = k_0 - \frac{e}{\hbar}\int E\,\mathrm{d}t. \tag{1.7}$$

If we assume that the initial wave vector is the same for both electrons, then the phase difference at the output of the ring is given by taking the difference of the integral over momentum in the top half of the ring (from an angle of π down to 0)

Figure 1.3. Transm
Au ring. The lines
Webb (1986), by pe

and the integral o

$$\Delta\phi = -\frac{e}{\hbar}$$

$$= -\frac{e}{\hbar}$$

where $\Phi_0 = h$
equations to rep
flux density. Th
phase difference
Aharonov–Bohr

In figure 1
There is a stro
ring is varied.
to magnetic fie
corresponding t
weak second h
weak non-linea
or to other phys

The cohe
observation of
are done at suc

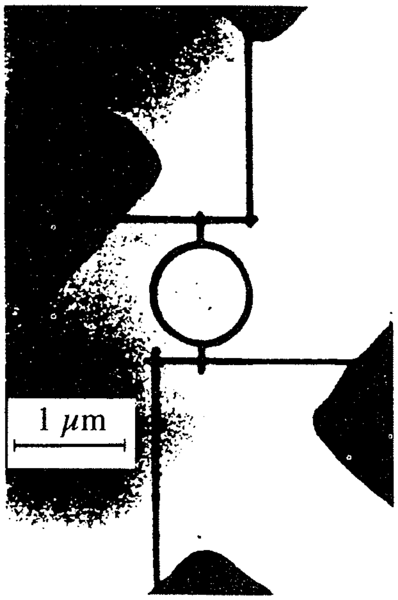


Figure 1.3. Transmission electron micrograph of a large-diameter (820 nm) polycrystalline Au ring. The lines are about 40 nm wide and about 38 nm thick. (After Washburn and Webb (1986), by permission.)

and the integral over the bottom half of the ring (from $-\pi$ up to 0):

$$\begin{aligned}\Delta\phi &= -\frac{e}{\hbar}\int \mathrm{d}t\left(\int_{\pi}^{0} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l} + \int_{-\pi}^{0} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l}\right) = -\frac{e}{\hbar}\int \mathrm{d}t\int_{0}^{2\pi} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{l}\\ &= -\frac{e}{\hbar}\int \mathrm{d}t\int \boldsymbol{\nabla}\times\boldsymbol{E}\cdot\boldsymbol{n}\,\mathrm{d}A = \frac{e}{\hbar}\int \boldsymbol{B}\cdot\boldsymbol{n}\,\mathrm{d}A = 2\pi\frac{\Phi}{\Phi_0} \qquad (1.8)\end{aligned}$$

where $\Phi_0 = h/e$ is the quantum unit of flux, and we have used Maxwell's equations to replace the electric field by the time derivative of the magnetic flux density. Thus, a *static* magnetic field coupled through the loop creates a phase difference between the waves that traverse the two paths. This effect is the Aharonov–Bohm (1959) effect.

In figure 1.4(*a*), the conductance through the ring of figure 1.3 is shown. There is a strong oscillatory behaviour as the magnetic field coupled by the ring is varied. The curve of figure 1.4(*b*) is the Fourier transform (with respect to magnetic field) of the conductance and shows a clear fundamental peak corresponding to a 'frequency' given by the periodicity of Φ_0. There is also a weak second harmonic evident in the Fourier transform, which may be due to weak non-linearities in the ring (arising from variations in thickness, width etc) or to other physical processes (some of which are understood).

The coherence of the electron waves is a clear requirement for the observation of the Aharonov–Bohm effect, and this is why the measurements are done at such low temperatures. It is important that the size of the ring be

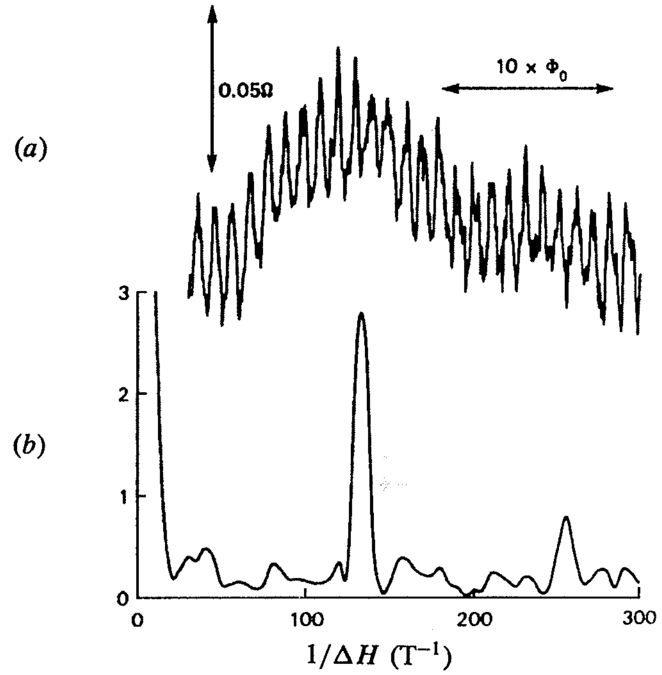


Figure 1.4. Conductance through the ring of figure 1.3. In (*a*), the conductance oscillations are shown at a temperature of 0.04 K. The Fourier transform is shown in (*b*) and gives clearly evidence of the dominant h/e period of the oscillations. (After Washburn and Webb (1986), by permission.)

smaller than some characteristic coherence length, which is termed the inelastic mean free path (where it is assumed that it is inelastic collisions between the electrons that destroy the phase coherence). Nevertheless, the understanding of this phenomenon depends upon the ability to treat the electrons as waves, and, moreover, the phenomenon is only found in a temperature regime where the phase coherence is maintained. At higher temperatures, the interactions between the electrons in the metal ring become so strong that the phase is *randomized*, and any possibility of phase interference effects is lost. Thus the quantum interference is only observable on size and energy scales (set by the coherence length and the temperature, respectively) such that the quantum interference is quite significant. As the temperature is raised, the phase is randomized by the collisions, and normal classical behaviour is recovered. This latter may be described by requiring that the two waves used above add in intensity, and not in amplitude as we have done. The addition of intensities 'throws away' the phase variables and precludes the possibility of phase interference between the two paths.

The preceding paragraphs describe how we can 'measure' the phase interference between the electron *waves* passing through two separate arms of the system. In this regard, these two arms serve as the two *slits* for the optical waves of figure 1.1. Observation of the interference phenomena shows us that the electrons

must be considered
we have a confirma
views of a coheren
used to vary the ph
behaviour of the c
phase in just one
1985). Using tech
gate voltage varies
which will lead t
according to (1.7)
the magnetic field.
has been develope
once we have disc

Which is the
wave? The answe
largely upon the
the approach that
majority of work
mechanical pictu
next chapter). Or
dealing with bo
equation is just a
formulation of th
commuting varia
the discussion of

1.4 Position

For the remainde
we can expect
particles). Do w
the wave is a co
real quantities.
the particle to b
probability of fi
of the wave. Th

is the probabilit
the wave functi

must be considered as waves, and not as particles, for this experiment. Once more, we have a confirmation of the *correspondence* between waves and particles as two views of a coherent whole. In the preceding experiment, the magnetic field was used to vary the phase in both arms of the interferometer and induce the oscillatory behaviour of the conductance on the magnetic field. It is also possible to vary the phase in just one arm of the interferometer by the use of a tuning gate (Fowler 1985). Using techniques which will be discussed in the following chapters, the gate voltage varies the propagation wave vector k in one arm of the interferometer, which will lead to additional oscillatory conductance as this voltage is tuned, according to (1.7) and (1.8), as the electric field itself is varied instead of using the magnetic field. A particularly ingenious implementation of this interferometer has been developed by Yacoby *et al* (1994), and will be discussed in later chapters once we have discussed the underlying physics.

Which is the proper interpretation to use for a general problem: particle or wave? The answer is not an easy one to give. Rather, the proper choice depends largely upon the particular quantum effect being investigated. Thus one chooses the approach that yields the answer with minimum effort. Nevertheless, the great majority of work actually has tended to treat the quantum mechanics via the wave mechanical picture, as embodied in the Schrödinger equation (discussed in the next chapter). One reason for this is the great wealth of mathematical literature dealing with boundary value problems, as the time-independent Schrödinger equation is just a typical wave equation. Most such problems actually lie in the formulation of the proper boundary conditions, and then the imposition of non-commuting variables. Before proceeding to this, however, we diverge to continue the discussion of position and momentum as variables and operators.

1.4 Position and momentum

For the remainder of this chapter, we want to concentrate on just what properties we can expect from this wave that is supposed to represent the particle (or particles). Do we represent the particle simply by the wave itself? No, because the wave is a complex quantity, while the charge and position of the particle are real quantities. Moreover, the wave is a distributed quantity, while we expect the particle to be relatively localized in space. This suggests that we relate the *probability* of finding the electron at a position x to the square of the magnitude of the wave. That is, we say that

$$|\Psi(x,t)|^2 \tag{1.9}$$

is the probability of finding an electron at point x at time t. Then, it is clear that the wave function must be normalized through

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2 \, \mathrm{d}x = 1. \tag{1.10}$$

While (1.10) extends over all space, the appropriate volume is that of the system under discussion. This leads to a slightly different normalization for the plane waves utilized in section 1.3 above. Here, we use *box normalization* (the term 'box' refers to the three-dimensional case):

$$\lim_{L\to\infty}\int_{-L/2}^{L/2}|\Psi(x,t)|^2\,\mathrm{d}x = 1. \tag{1.11}$$

This normalization keeps constant total probability and recognizes that, for a uniform probability, the amplitude must go to zero as the volume increases without limit.

There are additional constraints which we wish to place upon the wave function. The first is that the system is linear, and satisfies superposition. That is, if there are two physically realizable states, say ψ_1 and ψ_2, then the total wave function must be expressable by the linear summation of these, as

$$\Psi(x,t) = c_1\psi_1(x,t) + c_2\psi_2(x,t). \tag{1.12}$$

Here, c_1 and c_2 are arbitrary complex constants, and the summation represents a third, combination state that is physically realizable. Using (1.12) in the probability requirement places additional load on these various states. First, each ψ_i must be normalized independently. Secondly, the constants c_i must now satisfy (1.10) as

$$\begin{aligned}\int_{-\infty}^{\infty}|\Psi(x,t)|^2\,\mathrm{d}x = 1 &= |c_1|^2\int_{-\infty}^{\infty}|\psi_1(x,t)|^2\,\mathrm{d}x + |c_2|^2\int_{-\infty}^{\infty}|\psi_1(x,t)|^2\,\mathrm{d}x\\ &= |c_1|^2 + |c_2|^2.\end{aligned} \tag{1.13}$$

In order for the last equation to be correct, we must apply the third requirement of

$$\int_{-\infty}^{\infty}\psi_1^*(x,t)\psi_2(x,t)\,\mathrm{d}x = \int_{-\infty}^{\infty}\psi_2^*(x,t)\psi_1(x,t)\,\mathrm{d}x = 0 \tag{1.14}$$

which is that the individual states are *orthogonal* to one another, which must be the case for our use of the composite wave function (1.12) to find the probability.

1.4.1 Expectation of the position

With the normalizations that we have now introduced, it is clear that we are equating the square of the magnitude of the wave function with a probability density function. This allows us to compute immediately the expectation value, or average value, of the position of the particle with the normal definitions introduced in probability theory. That is, the average value of the position is given by

$$\langle x\rangle = \int_{-\infty}^{\infty}x|\Psi(x,t)|^2\,\mathrm{d}x = \int_{-\infty}^{\infty}\Psi^*(x,t)x\Psi(x,t)\,\mathrm{d}x. \tag{1.15}$$

is that of the system
ization for the plane
malization (the term

(1.11)

cognizes that, for a
ıe volume increases

lace upon the wave
superposition. That
, then the total wave
ıe, as

(1.12)

mmation represents
Jsing (1.12) in the
ıs states. First, each
c_i must now satisfy

$\int^{\infty}_{-\infty} |\psi_1(x,t)|^2\,\mathrm{d}x$

(1.13)

hird requirement of

$= 0$ (1.14)

her, which must be
ind the probability.

clear that we are
with a probability
expectation value,
normal definitions
of the position is

$t)\,\mathrm{d}x.$ (1.15)

In the last form, we have split the wave function product into its two components and placed the position *operator* between the complex conjugate of the wave function and the wave function itself. This is the standard notation, and designates that we are using the concept of an inner product of two functions to describe the average. If we use (1.10) to define the inner product of the wave function and its complex conjugate, then this may be described in the short-hand notation

$$(\Psi,\Psi) = \int_{-\infty}^{\infty} \Psi^*(x,t)\Psi(x,t)\,\mathrm{d}x = 1 \tag{1.16}$$

and

$$\langle x\rangle = (\Psi, x\Psi). \tag{1.17}$$

Before proceeding, it is worthwhile to consider an example of the expectation value of the wave function. Consider the Gaussian wave function

$$\Psi(x,t) = A\exp(-x^2/2)\mathrm{e}^{-\mathrm{i}\omega t}. \tag{1.18}$$

We first normalize this wave function as

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2\,\mathrm{d}x = A^2\int_{-\infty}^{\infty} \exp(-x^2)\,\mathrm{d}x = A^2\sqrt{\pi} = 1 \tag{1.19}$$

so that $A = \pi^{-1/4}$. Then, the expectation value of position is

$$\begin{aligned}\langle x\rangle &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \exp(-x^2/2)x\exp(-x^2/2)\,\mathrm{d}x\\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} x\mathrm{e}^{-x^2}\,\mathrm{d}x = 0.\end{aligned} \tag{1.20}$$

Our result is that the average position is at $x = 0$. On the other hand, the expectation value of x^2 is

$$\begin{aligned}\langle x^2\rangle &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} \exp(-x^2/2)x^2\exp(-x^2/2)\,\mathrm{d}x\\ &= \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} x^2\mathrm{e}^{-x^2}\,\mathrm{d}x = \frac{1}{2}.\end{aligned} \tag{1.21}$$

We say at this point that we have described the wave function corresponding to the particle in the *position representation*. That is, the wave function is a function of the position and the time, and the square of the magnitude of this function describes the probability density function for the position. The position operator itself, x, operates on the wave function to provide a new function, so the inner product of this new function with the original function gives the average value of the position. Now, if the position variable x is to be interpreted as an operator, and the wave function in the position representation is the natural function to use to describe the particle, then it may be said that the wave function $\Psi(x,t)$ has an *eigenvalue* corresponding to the operator x. This means that we can write the operation of x on $\Psi(x,t)$ as

$$x\Psi(x,t) = \underline{x}\Psi(x,t) \tag{1.22}$$

where $\underline{x}$ is the eigenvalue of x operating on $\Psi(x,t)$. It is clear that the use of (1.22) in (1.7) means that the eigenvalue $\underline{x} = \langle x \rangle$.

We may decompose the overall wave function into an expansion over a complete orthonormal set of basis functions, just like a Fourier series expansion in sines and cosines. Each member of the set has a well defined eigenvalue corresponding to an operator if the set is the proper basis set with which to describe the effect of that operator. Thus, the present use of the position representation means that our functions are the proper functions with which to describe the action of the position operator, which does no more than determine the expectation value of the position of our particle.

Consider the wave function shown in figure 1.5. Here, the real part of the wave function is plotted, as the wave function itself is in general a complex quantity. However, it is clear that the function is peaked about some point x_{peak}. While it is likely that the expectation value of the position is very near this point, this cannot be discerned exactly without actually computing the action of the position operator on this function and computing the expectation value, or inner product, directly. This circumstance arises from the fact that we are now dealing with probability functions, and the expectation value is simply the most likely position in which to find the particle. On the other hand, another quantity is evident in figure 1.5, and this is the width of the wave function, which relates to the standard deviation of the wave function. Thus, we can define

$$(\Delta x)^2 = (\Psi, (x - \langle x \rangle)^2 \Psi). \tag{1.23}$$

For our example wave function of (1.18), we see that the uncertainty may be expressed as

$$\Delta x = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} = \sqrt{\frac{1}{2} - 0} = \frac{1}{\sqrt{2}}. \tag{1.24}$$

The quantity Δx relates to the uncertainty in finding the particle at the position $\langle x \rangle$. It is clear that if we want to use a wave packet that describes the position of the particle *exactly*, then Δx must be made to go to zero. Such a function is the Dirac delta function familiar from circuit theory (the impulse function). Here, though, we use a delta function in position rather than in time; for example, we describe the wave function through

$$\Psi(x,0) = \delta(x - x_{\text{peak}}). \tag{1.25}$$

The time variable has been set to zero here for convenience, but it is easy to extend (1.25) to the time-varying case. Clearly, equation (1.25) describes the wave

Figur

function under the
We will examine
upon our knowle

1.4.2 Momentu

The wave functic
uniform quantity.
rapidly in space.
get a representati
the wave functio
as an inverse tran

The quantity $\phi($
Here, k is the spa
in (1.4). That i
which in turn is
called the *mome*
space is made us
functions. Conse
expectation valu
essence, we are

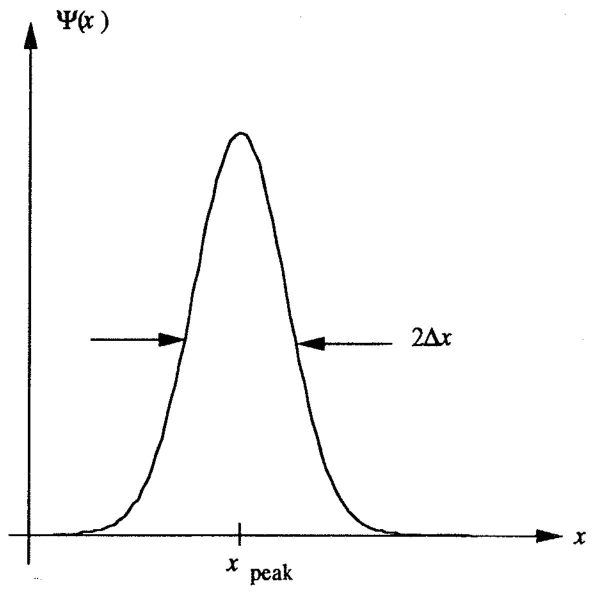


Figure 1.5. The positional variation of a typical wave function.

function under the condition that the position of the particle is known absolutely! We will examine in the following paragraphs some of the limitations this places upon our knowledge of the dynamics of the particle.

1.4.2 Momentum

The wave function shown in figure 1.5 contains variations in space, and is not a uniform quantity. In fact, if it is to describe a localized particle, it must vary quite rapidly in space. It is possible to Fourier transform this wave function in order to get a representation that describes the spatial frequencies that are involved. Then, the wave function in this figure can be written in terms of the spatial frequencies as an inverse transform:

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}kx} \, \mathrm{d}k. \tag{1.26}$$

The quantity $\phi(k)$ represents the Fourier transform of the wave function itself. Here, k is the spatial frequency. However, this k is precisely the same k as appears in (1.4). That is, the spatial frequency is described by the *wave vector* itself, which in turn is related to the momentum through (1.4). For this reason, $\phi(k)$ is called the *momentum wave function.* A description of the particle in momentum space is made using the Fourier-transformed wave functions, or momentum wave functions. Consequently, the average value of the momentum for our particle, the expectation value of the operator p, may be evaluated using these functions. In essence, we are saying that the proper basis set of functions with which to evaluate the momentum is that of the momentum wave functions. Then, it follows that

$$\langle p\rangle = \hbar(\phi, k\phi) = \hbar\int_{-\infty}^{\infty}\phi^* k\phi\,\mathrm{d}k. \tag{1.27}$$

As an example of momentum wave functions, we consider the position wave function of (1.18). We find the momentum wave function from

$$\begin{aligned}\phi(k,t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\Psi(x,t)\mathrm{e}^{-\mathrm{i}kx}\,\mathrm{d}x = \frac{1}{\sqrt{2}\pi^{3/4}}\int_{-\infty}^{\infty}\mathrm{e}^{-x^2/2-\mathrm{i}kx}\,\mathrm{d}x\\ &= \frac{1}{\sqrt{2}\pi^{3/4}}\mathrm{e}^{-k^2/2}\int_{-\infty}^{\infty}\exp\left(-\frac{(x+\mathrm{i}k)^2}{2}\right)\mathrm{d}x = \frac{1}{\pi^{1/4}}\mathrm{e}^{-k^2/2}.\end{aligned} \tag{1.28}$$

This has the same form as (1.18), so that we can immediately use (1.20) and (1.21) to infer that $\langle k\rangle = 0$ and $\langle k^2\rangle = \frac{1}{2}$.

Suppose, however, that we are using the position representation wave functions. How then are we to interpret the expectation value of the momentum? The wave functions in this representation are functions only of x and t. To evaluate the expectation value of the momentum operator, it is necessary to develop the operator corresponding to the momentum in the position representation. To do this, we use (1.27) and introduce the Fourier transforms corresponding to the functions ϕ. Then, we may write (1.27) as

$$\begin{aligned}\langle p\rangle &= \frac{\hbar}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\int_{-\infty}^{\infty}\mathrm{d}x'\,\Psi^*(x')\mathrm{e}^{\mathrm{i}kx'}k\int_{-\infty}^{\infty}\mathrm{d}x\,\Psi(x)\mathrm{e}^{-\mathrm{i}kx}\\ &= \frac{\hbar}{2\mathrm{i}\pi}\int_{-\infty}^{\infty}\mathrm{d}k\int_{-\infty}^{\infty}\mathrm{d}x'\,\Psi^*(x')\mathrm{e}^{\mathrm{i}kx'}\int_{-\infty}^{\infty}\mathrm{d}x\,\mathrm{e}^{-\mathrm{i}kx}\frac{\partial}{\partial x}\Psi(x)\\ &= -\mathrm{i}\hbar\int_{-\infty}^{\infty}\mathrm{d}x'\int_{-\infty}^{\infty}\mathrm{d}x\,\Psi^*(x')\delta(x-x')\frac{\partial}{\partial x}\Psi(x)\\ &= -\mathrm{i}\hbar\int_{-\infty}^{\infty}\mathrm{d}x\,\Psi^*(x)\frac{\partial}{\partial x}\Psi(x).\end{aligned} \tag{1.29}$$

In arriving at the final form of (1.29), an integration by parts has been done from the first line to the second (the evaluation at the limits is assumed to vanish), after replacing k by the partial derivative. The third line is achieved by recognizing the delta function:

$$\delta(x-x') = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{d}k\,\mathrm{e}^{\mathrm{i}k(x-x')}. \tag{1.30}$$

Thus, in the position representation, *the momentum operator* is given by the functional operator

$$p = -\mathrm{i}\hbar\frac{\partial}{\partial x}. \tag{1.31}$$

1.4.3 Non-comm

The description of t
a differential operat
and to the momentu

The left-hand side
bracket. However,
within the bracke
function. Thus, th
inner product, or e:

$$-(\Psi,[x,p]\Psi$$

If variables, or op
quantities cannot l
deeper meaning.
position operator a
an eigenvalue $\underline{x}$,
momentum operat
of the position rep
complex result.
simple eigenvalue
$\exp(\mathrm{i}px/\hbar)$ (whic
(1.22) with the d
form is not integ
thus the same wa
position and mor
which correspond
cannot be simulta

There is a fu
transform pair o
known, for exam
transform has u
probability of tal
since the positio
say anything ab
likely. Similarl
function, which
the position wav
the momentum i
as all values of
describe both of

1.4.3 Non-commuting operators

The description of the momentum operator in the position representation is that of a differential operator. This means that the operators corresponding to the position and to the momentum will not commute, by which we mean that

$$[x,p] = xp - px \neq 0. \tag{1.32}$$

The left-hand side of (1.32) defines a quantity that is called the *commutator bracket*. However, by itself it only has implied meaning. The terms contained within the brackets are operators and must actually operate on some wave function. Thus, the role of the commutator can be explained by considering the inner product, or expectation value. This gives

$$-(\Psi,[x,p]\Psi) = +\mathrm{i}\hbar\left\{\left(\Psi, x\frac{\partial}{\partial x}\Psi\right) - \left(\Psi, \frac{\partial}{\partial x}x\Psi\right)\right\} = -\mathrm{i}\hbar. \tag{1.33}$$

If variables, or operators, do not commute, there is an implication that these quantities cannot be measured simultaneously. Here again, there is another and deeper meaning. In the previous section, we noted that the operation of the position operator x on the wave function in the position representation produced an eigenvalue $\underline{x}$, which is actually the expectation value of the position. The momentum operator does not produce this simple result with the wave function of the position representation. Rather, the differential operator produces a more complex result. For example, if the differential operator were to produce a simple eigenvalue, then the wave function would be constrained to be of the form $\exp(\mathrm{i}px/\hbar)$ (which can be shown by assuming a simple eigenvalue form as in (1.22) with the differential operator and solving the resulting equation). This form is not integrable (it does not fit our requirements on normalization), and thus the same wave function cannot simultaneously yield eigenvalues for both position and momentum. Since the eigenvalue relates to the expectation value, which corresponds to the most likely result of an experiment, these two quantities cannot be simultaneously measured.

There is a further level of information that can be obtained from the Fourier transform pair of position and momentum wave functions. If the position is known, for example if we choose the delta function of (1.25), then the Fourier transform has unit amplitude everywhere; that is, the momentum has equal probability of taking on any value. Another way of looking at this is to say that since the position of the particle is completely determined, it is impossible to say anything about the momentum, as any value of the momentum is equally likely. Similarly, if a delta function is used to describe the momentum wave function, which implies that we know the value of the momentum exactly, then the position wave function has equal amplitude everywhere. This means that if the momentum is known, then it is impossible to say anything about the position, as all values of the latter are equally likely. As a consequence, if we want to describe both of these properties of the particle, the position wave function and its Fourier transform must be selected carefully to allow this to occur. Then there will be an uncertainty Δx in position, as indicated in figure 1.5, and there will be a corresponding uncertainty Δp in momentum.

To investigate the relationship between the two uncertainties, in position and momentum, let us choose a Gaussian wave function to describe the wave function in the position representation. Therefore, we take

$$\Psi(x) = \frac{1}{(2\pi)^{1/4}\sigma^{1/2}} \exp\left[-\frac{x^2}{4\sigma^2}\right]. \tag{1.34}$$

Here, the wave packet has been centred at $x_{\text{peak}} = 0$, and

$$\langle x \rangle = \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] x \, \mathrm{d}x = 0 \tag{1.35}$$

as expected. Similarly, the uncertainty in the position is found from (1.23) as

$$\begin{aligned} (\Delta x)^2 &= \frac{1}{(2\pi)^{1/2}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] x^2 \, \mathrm{d}x \\ &= \frac{\sigma}{(2\pi)^{1/2}} \int_{-\infty}^{\infty} \exp\left[-\frac{x^2}{2\sigma^2}\right] \mathrm{d}x = \sigma^2 \end{aligned} \tag{1.36}$$

and $\Delta x = \sigma$.

The appropriate momentum wave function can now be found by Fourier transforming this position wave function. This gives

$$\begin{aligned} \phi(k) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \Psi(x) \mathrm{e}^{-\mathrm{i}kx} \, \mathrm{d}x \\ &= \frac{1}{\sigma^{1/2}(2\pi)^{3/4}} \mathrm{e}^{-\sigma^2 k^2} \int_{-\infty}^{\infty} \exp\left[-\frac{(x - 2\mathrm{i}\sigma^2 k)^2}{4\sigma^2}\right] \mathrm{d}x \\ &= \left(\frac{2}{\pi}\right)^{1/4} \sqrt{\sigma} \mathrm{e}^{-\sigma^2 k^2}. \end{aligned} \tag{1.37}$$

We note that the momentum wave function is also centred about zero momentum. Then the uncertainty in the momentum can be found as

$$(\Delta p)^2 = \hbar^2 \sigma \sqrt{\frac{2}{\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-2\sigma^2 k^2} k^2 \, \mathrm{d}k = \frac{\hbar^2}{4\sigma^2}. \tag{1.38}$$

Hence, the uncertainty in the momentum is $\hbar/2\sigma$. We now see that the non-commuting operators x and p can be described by an uncertainty $\Delta x \Delta p = \hbar/2$. It turns out that our description in terms of the static Gaussian wave function is a *minimal-uncertainty* description, in that the product of the two uncertainties is a minimum.

The uncertainty principle describes the connection between the uncertainties in determination of the expectation values for two non-commuting operators. If we have two operators A and B, which do not commute, then the uncertainty relation states that

$$\Delta A \Delta B \geq \tfrac{1}{2}|\langle[A, B]\rangle| \tag{1.39}$$

where the angular brackets denote the expectation value, as above. It is easily confirmed that the position and momentum operators satisfy this relation.

It is important to note that the basic uncertainty relation is only really valid for non-commuting operators. It has often been asserted for variables like energy (frequency) and time, but in the non-relativistic quantum mechanics that we are investigating here, time is not a dynamic variable and has no corresponding operator. Thus, if there is any uncertainty for these latter two variables, it arises from the problems of making measurements of the energy at different times—and hence is a measurement uncertainty and not one expected from the uncertainty relation (1.39).

To understand how a *classical* measurement problem can give a result much like an uncertainty relationship, consider the simple time-varying exponential $\mathrm{e}^{-t/\tau}$. We can find the frequency content of this very simple time variation as

$$F(\omega) = \frac{1}{1 + (\omega\tau)^2}. \tag{1.40}$$

Hence, if we want to reproduce this simple exponential with our electronics, we require a bandwidth ($\Delta\omega$) that is at least of order $1/\tau$. That is, we require

$$\Delta\omega > \frac{1}{\tau} \Rightarrow \Delta E \Delta t > \hbar \tag{1.41}$$

where we have used (1.2′) to replace the angular frequency with the energy of the wave and have taken $\Delta t = \tau$. While this has significant resemblance to the quantum uncertainty principle, it is in fact a *classical* result whose only connection to quantum mechanics is through the Planck relationship. The fact that time is *not* an operator in our approach to quantum mechanics, but is simply a measure of the system progression, means that there cannot be a quantum version of (1.41).

1.4.4 Returning to temporal behaviour

While we have assumed that the momentum wave function is centred at zero momentum, this is not the general case. Suppose, we now assume that the momentum wave function is centred at a displaced value of k, given by k_0. Then, the entire position representation wave function moves with this average momentum, and shows an average velocity $v_0 = \hbar k_0/m$. We can expect that the peak of the position wave function, x_{peak}, moves, but does it move with this velocity? The position wave function is made up of a sum of a great many

Fourier components, each of which arises from a different momentum. Does this affect the uncertainty in position that characterizes the half-width of the position wave function? The answer to both of these questions is yes, but we will try to demonstrate that these are the correct answers in this section.

Our approach is based upon the definition of the Fourier inverse transform (1.26). This latter equation expresses the position wave function $\Psi(x)$ as a summation of individual Fourier components, each of whose amplitudes is given by the value of $\phi(k)$ at that particular k. From the earlier work, we can extend each of the Fourier terms into a plane wave corresponding to that value of k, by introducing the frequency term via

$$\Psi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \phi(k) \mathrm{e}^{\mathrm{i}(kx-\omega t)}\, \mathrm{d}k. \tag{1.42}$$

While the frequency term has not been shown with a variation with k, it must be recalled that each of the Fourier components may actually possess a slightly different frequency. If the main frequency corresponds to the peak of the momentum wave function, then the frequency can be expanded as

$$\omega(k) = \omega(k_0) + (k-k_0) \left.\frac{\partial \omega}{\partial k}\right|_{k=k_0} + \cdots . \tag{1.43}$$

The interpretation of the position wave function is now that it is composed of a group of closely related waves, all propagating in the same direction (we assume that $\phi(k) = 0$ for $k < 0$, but this is merely for convenience and is not critical to the overall discussion). Thus, $\Psi(x, t)$ is now defined as a *wave packet*. Equation (1.43) defines the *dispersion* across this wave packet, as it gives the gradual change in frequency for different components of the wave packet.

To understand how the dispersion affects the propagation of the wave functions, we insert (1.43) into (1.42), and define the difference variable $u = k - k_0$. Then, (1.42) becomes

$$\Psi(x,t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int_{-\infty}^{\infty} \phi(u+k_0) \mathrm{e}^{\mathrm{i}(ux-\omega' ut)}\, \mathrm{d}u \tag{1.44}$$

where ω_0 is the leading term in (1.43) and ω' is the partial derivative in the second term of (1.43). The higher-order terms of (1.43) are neglected, as the first two terms are the most significant. If u is factored out of the argument of the exponential within the integral, it is seen that the position variable varies as $x - \omega' t$. This is our guide as to how to proceed. We will reintroduce k_0 within the exponential, but multiplied by this factor, so that

$$\begin{aligned}
\Psi(x,t) &= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\mathrm{i}k_0(x-\omega' t)} \mathrm{e}^{\mathrm{i}(k_0 x - \omega_0 t)} \int_{-\infty}^{\infty} \phi(u+k_0) \mathrm{e}^{\mathrm{i}k_0(x-\omega' t)} \mathrm{e}^{\mathrm{i}u(x-\omega' t)}\, \mathrm{d}u \\
&= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\mathrm{i}(\omega_0 - \omega' k_0)t} \int_{-\infty}^{\infty} \phi(u+k_0) \mathrm{e}^{\mathrm{i}(u+k_0)(x-\omega' t)}\, \mathrm{d}u \\
&= \mathrm{e}^{-\mathrm{i}(\omega_0 - \omega' k_0)t} \Psi(x - \omega' t, 0).
\end{aligned} \tag{1.45}$$

The leading expone
phase shift has no
expectation value c
with a velocity give
derivative of the fre
describes the group
wave packet in pos

This answers the f
the peak and move
wave packet. Note
with respect to the
by k_0? The answe
value. Rather, the
motion of the wav
k_0 so that it satisfi
wave packet:

If we integrate th
recover the other
the dynamic moti

It is clear that it is
momentum of th
momentum) to th

Let us now
time variation in
packet for the m

$\Psi(x$

To proceed, we
(energy) and ave

omentum. Does this
vidth of the position
s, but we will try to
.

er inverse transform
function $\Psi(x)$ as a
amplitudes is given
vork, we can extend
ɔ that value of k, by

(1.42)

tion with k, it must
y possess a slightly
to the peak of the
ed as

. (1.43)

t it is composed of
same direction (we
venience and is not
d as a *wave packet*.
ket, as it gives the
vave packet.
ation of the wave
ence variable u =

$^{t)}\,du$ (1.44)

al derivative in the
e neglected, as the
of the argument of
n variable varies as
ntroduce k_0 within

$-\omega' t)e^{iu(x-\omega' t)}\,du$

du

(1.45)

The leading exponential provides a phase shift in the position wave function. This phase shift has no effect on the square of the magnitude, which represents the expectation value calculations. On the other hand, the entire wave function moves with a velocity given by ω'. This is not surprising. The quantity ω' is the partial derivative of the frequency with respect to the momentum wave vector, and hence describes the group velocity of the wave packet. Thus, the average velocity of the wave packet in position space is given by the group velocity

$$v_{\mathrm{g}} = \omega' = \left.\frac{\partial \omega}{\partial k}\right|_{k=k_0}. \tag{1.46}$$

This answers the first question: the peak of the position wave function remains the peak and moves with an average velocity defined as the group velocity of the wave packet. Note that this group velocity is defined by the frequency variation with respect to the wave vector. Is this related to the average momentum given by k_0? The answer again is affirmative, as we cannot let k_0 take on any arbitrary value. Rather, the peak in the momentum distribution must relate to the average motion of the wave packet in position space. Thus, we must impose a value on k_0 so that it satisfies the condition of actually being the average momentum of the wave packet:

$$v_{\mathrm{g}} = \frac{\hbar k_0}{m} = \frac{\partial \omega}{\partial k}. \tag{1.47}$$

If we integrate the last two terms of (1.47) with respect to the wave vector, we recover the other condition that ensures that our wave packet is actually describing the dynamic motion of the particles:

$$\mathcal{E} = \hbar\omega = \frac{\hbar^2 k^2}{2m} = \frac{p^2}{2m}. \tag{1.48}$$

It is clear that it is the group velocity of the wave packet that describes the average momentum of the momentum wave function and also relates the velocity (and momentum) to the energy of the particle.

Let us now turn to the question of what the wave packet looks like with the time variation included. We rewrite (1.42) to take account of the centred wave packet for the momentum representation to obtain

$$\Psi(x,t) = \sqrt{\frac{\sigma}{2\pi}}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0 x}\int_{-\infty}^{\infty} \mathrm{e}^{-\sigma^2 u^2 + \mathrm{i}ux - \mathrm{i}\omega t}\,\mathrm{d}u. \tag{1.49}$$

To proceed, we want to insert the above relationship between the frequency (energy) and average velocity:

$$\omega = \frac{\hbar k^2}{2m} = \frac{\hbar}{2m}(u+k_0)^2 = \frac{\hbar u^2}{2m} + uv_{\mathrm{g}} + \frac{\hbar k_0^2}{2m}. \tag{1.50}$$

If (1.50) is inserted into (1.49), we recognize a new form for the 'static' *effective* momentum wave function:

$$\phi(k) = \sqrt{\sigma}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0(x-v_{\mathrm{g}}t/2)} \exp\left[-u^2\left(\sigma^2 + \mathrm{i}\frac{\hbar t}{2m}\right)\right] \tag{1.51}$$

which still leads to $\langle p\rangle = 0$, and $\Delta p = \hbar/2\sigma$. We can then evaluate the position representation wave function by continuing the evaluation of (1.49) using the short-hand notation

$$\sigma' = \sqrt{\sigma^2 + \mathrm{i}\frac{\hbar t}{2m}} \tag{1.52a}$$

and

$$x' = x - v_{\mathrm{g}}t. \tag{1.52b}$$

This gives

$$\begin{aligned}\Psi(x',t) &= \sqrt{\frac{\sigma}{2\pi}}\left(\frac{2}{\pi}\right)^{1/4} \mathrm{e}^{\mathrm{i}k_0(x-v_{\mathrm{g}}t/2)} \int_{-\infty}^{\infty} \mathrm{e}^{-\sigma'^2u^2+\mathrm{i}ux'}\,\mathrm{d}u \\ &= \frac{\sqrt{\sigma}}{(2\pi)^{1/4}\sigma'}\mathrm{e}^{\mathrm{i}k_0(x-v_{\mathrm{g}}t/2)} \exp\left[-\left(\frac{x'}{2\sigma'}\right)^2\right].\end{aligned} \tag{1.53}$$

This has the exact form of the previous wave function in the position representation with one important exception. The exception is that the time variation has made this result unnormalized. If we compute the inner product now, recalling that the terms in σ' are complex, the result is

$$(\Psi,\Psi) = \frac{\sigma}{|\sigma'|} = \frac{1}{\sqrt{1+\hbar^2t^2/(4m^2\sigma^4)}} \equiv \frac{1}{S}. \tag{1.54}$$

With this normalization, it is now easy to show that the expectation value of the position is that found above:

$$\langle x\rangle = \frac{(\Psi, x\Psi)}{(\Psi,\Psi)} = v_{\mathrm{g}}t. \tag{1.55}$$

Similarly, the standard deviation in position is found to be

$$\langle(\Delta x)^2\rangle = \sigma^2S^2 = \sigma^2\left[1+\frac{\hbar^2t^2}{4m^2\sigma^4}\right]. \tag{1.56}$$

This means that the uncertainty in the two non-commuting operators x and p increases with time according to

$$\Delta x\Delta p = \frac{\hbar}{2}\sqrt{1+\frac{\hbar^2t^2}{4m^2\sigma^4}}. \tag{1.57}$$

The wave packet a
variation is a shift of
Gaussian wave pack
that the position w
propagates. This dif
minimum uncertain
At later times, the va
to become less cer
effective frequencie
for us to know as n
the actual position c

1.5 Summary

Quantum mechani
duality. The main
usually small, for v
effect can be either
of a wave packet v
mechanics, the bo
develop in the next

The result of
variables have bec
As operators, the
uncertainty relatic
nature arises from
yields eigenvalues
be simultaneously

Even if we g
correlated to a mo
of the spatial var
diffusive broaden
two operator relat

We can drav
important for the
functions in diffe
derive from a sing
must be only first
components, whi
resemblance to a
derivative). These

The wave packet actually gets wider as it propagates with time, so the time variation is a shift of the centroid plus this broadening effect. The broadening of a Gaussian wave packet is familiar in the process of diffusion, and we recognize that the position wave packet actually undergoes a diffusive broadening as it propagates. This diffusive effect accounts for the increase in the uncertainty. The minimum uncertainty arises only at the initial time when the packet was formed. At later times, the various momentum components cause the wave packet position to become less certain since different spatial variations propagate at different effective frequencies. Thus, for any times after the initial one, it is not possible for us to know as much about the wave packet and there is more uncertainty in the actual position of the particle that is represented by the wave packet.

1.5 Summary

Quantum mechanics furnishes a methodology for treating the wave–particle duality. The main importance of this treatment is for structures and times, both usually small, for which the *interference* of the waves can become important. The effect can be either the interference between two wave packets, or the interference of a wave packet with itself, such as in boundary value problems. In quantum mechanics, the boundary value problems deal with the equation that we will develop in the next chapter for the wave packet, the Schrödinger equation.

The result of dealing with the wave nature of particles is that dynamical variables have become operators which in turn operate upon the wave functions. As operators, these variables often no longer commute, and there is a basic uncertainty relation between non-commuting operators. The non-commuting nature arises from it being no longer possible to generate a wave function that yields eigenvalues for *both* of the operators, representing the fact that they cannot be simultaneously measured. It is this that introduces the uncertainty relationship.

Even if we generate a minimum-uncertainty wave packet in real space, it is correlated to a momentum space representation, which is the Fourier transform of the spatial variation. The time variation of this wave packet generates a diffusive broadening of the wave packet, which increases the uncertainty in the two operator relationships.

We can draw another set of conclusions from this behaviour that will be important for the differential equation that can be used to find the actual wave functions in different situations. The entire time variation has been found to derive from a single initial condition, which implies that the differential equation must be only first order in the time derivatives. Second, the motion has diffusive components, which suggests that the differential equation should bear a strong resemblance to a diffusion equation (which itself is only first order in the time derivative). These points will be expanded upon in the next chapter.

References

Aharonov Y and Bohm D 1959 *Phys. Rev.* **115** 485

de Broglie L 1939 *Matter and Light, The New Physics* (New York: Dover) p 267 (this is a reprint of the original translation by W H Johnston of the 1937 original *Matière et Lumière*)

Einstein A 1905 *Ann. Phys., Lpz.* **17** 132

Fowler A B 1985 *US Patent* 4550330

Landau L D and Lifshitz E M 1958 *Quantum Mechanics* (London: Pergamon)

Longair M S 1984 *Theoretical Concepts in Physics* (Cambridge: Cambridge University Press)

Washburn S and Webb R A 1986 *Adv. Phys.* **35** 375–422

Webb R A, Washburn S, Umbach C P and Laibowitz R B 1985 *Phys. Rev. Lett.* **54** 2696–99

Yacoby A, Heiblum M, Umansky V, Shtrikman H and Mahalu D 1994 *Phys. Rev. Lett.* **73** 3149–52

Problems

1. Calculate the
by the complex field

and show that its ave

2. What are th
a proton accelerate
velocities?

3. Show that
operator

in momentum space
that (1.32) is still sa

4. An electron
100 eV, is initially
time elapses before

5. Express th
packet in terms of
wave function.

6. A particle
medium, describec

What is the group

7. The long
silicon is 296 nm
irradiated with lig
the emitted electr
5 mW cm^{-2}, wh

8. For parti
300 K of electron

9. Conside
assume that the
what is their con

10. A wave

Problems

1. Calculate the energy density for the plane electromagnetic wave described by the complex field strength

$$E_\mathrm{c} = E_0 \mathrm{e}^{\mathrm{i}(\omega t - kx)}$$

and show that its average over a temporal period T is $\omega = (\varepsilon/2)|E_\mathrm{c}|^2$.

2. What are the de Broglie frequencies and wavelengths of an electron and a proton accelerated to 100 eV? What are the corresponding group and phase velocities?

3. Show that the position operator x is represented by the differential operator

$$\mathrm{i}\hbar\frac{\partial}{\partial p}$$

in momentum space, when dealing with momentum wave functions. Demonstrate that (1.32) is still satisfied when momentum wave functions are used.

4. An electron represented by a Gaussian wave packet, with average energy 100 eV, is initially prepared with $\Delta p = 0.1\langle p\rangle$ and $\Delta x = \hbar/[2(\Delta p)]$. How much time elapses before the wave packet has spread to twice the original spatial extent?

5. Express the expectation value of the kinetic energy of a Gaussian wave packet in terms of the expectation value and the uncertainty of the momentum wave function.

6. A particle is represented by a wave packet propagating in a dispersive medium, described by

$$\omega = \frac{A}{\hbar}\left\{\sqrt{1+\frac{\hbar^2 k^2}{mA}} - 1\right\}.$$

What is the group velocity as a function of k?

7. The longest wavelength that can cause the emission of electrons from silicon is 296 nm. (*a*) What is the work function of silicon? (*b*) If silicon is irradiated with light of 250 nm wavelength, what is the energy and momentum of the emitted electrons? What is their wavelength? (*c*) If the incident photon flux is 5 mW cm^{-2}, what is the photoemission current density?

8. For particles which have a thermal velocity, what is the wavelength at 300 K of electrons, helium atoms, and the α-particle (which is ionized ^{4}He)?

9. Consider that an electron is confined within a region of 10 nm. If we assume that the uncertainty principle provides a RMS value of the momentum, what is their confinement energy?

10. A wave function has been determined to be given by the spatial variation

$$\Psi(x) = \begin{cases} 2A & -a < x < 0 \\ 2A(a-x) & 0 < x < a \\ 0 & \text{elsewhere.} \end{cases}$$

Determine the value of A, the expectation value of x, x^2, p and p^2. What is the value of the uncertainty in position–momentum?

11. A wave function has been determined to be given by the spatial variation

$$\Psi(x) = \begin{cases} 2A\sin\left(\frac{\pi}{a}\right) & -a < x < a \\ 0 & \text{elsewhere.} \end{cases}$$

Determine the value of A, the expectation value of x, x^2, p and p^2. What is the value of the uncertainty in position–momentum?

Chapter 2

The Schrö

In the first chapte
mechanics arise fr
an observable leve
The most importar
and the non-comm
either in the positi
related to the prob
wave could be ex
of a diffusive nat
function evolved f
in the time derivat

It must be
quantum mechani
the new quantum
by Werner Heise
1925. In this app
This approach wa
work through in
was actually rep
was worked out
the winter vacat
was found to pro
not appreciated
two approaches
Heisenberg recei
while Schröding
in atomic physic
universally used
with a backgrou
not completely

LONG JOURNEY INTO TUNNELING

Nobel Lecture, December 12, 1973

by

Leo Esaki

IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., USA

I. Historical Background

In 1923, during the infancy of the quantum theory, de Broglie (1) introduced a new fundamental hypothesis that matter may be endowed with a dualistic nature - particles may also have the characteristics of waves. This hypothesis, in the hands of Schrodinger (2) found expression in the definite form now known as the Schrödinger wave equation, whereby an electron or a particle is assumed to be represented by a solution to this equation. The continuous nonzero nature of such solutions, even in classically forbidden regions of negative kinetic energy, implies an ability to penetrate such forbidden regions and a probability of tunneling from one classically allowed region to another. The concept of tunneling, indeed, arises from this quantum-mechanical result. The subsequent experimental manifestations of this concept can be regarded as one of the early triumphs of the quantum theory.

In 1928, theoretical physicists believed that tunneling could occur by the distortion, lowering or thinning, of a potential barrier under an externally applied high electric field. Oppenheimer (3) attributed the autoionization of excited states of atomic hydrogen to the tunnel effect: The coulombic potential well which binds an atomic electron could be distorted by a strong electric field so that the electron would see a finite potential barrier through which it could tunnel.

Fowler and Nordheim (4) explained, on the basis of electron tunneling, the main features of the phenomenon of electron emission from cold metals by high external electric fields, which had been unexplained since its observation by Lilienfeld (5) in 1922. They proposed a one-dimensional model. Metal electrons are confined by a potential wall whose height is determined by the work function ψ plus the fermi energy E_f, and the wall thickness is substantillay decreased with an externally applied high electric field, allowing electrons to tunnel through the potential wall, as shown in Fig. 1. They successfully derived the well-known Fowler-Nordheim formula for the current as a function of electric field F:

$$J = AF^2 \exp[-4(2m)^{1/2}\Phi^{3/2}/3\hbar F].$$

An application of these ideas which followed almost immediately came in the model for a decay as a tunneling process put forth by Gamow (6) and Gurney and Condon. (7) Subsequently, Rice (8) extended this theory to the description of molecular dissociation.

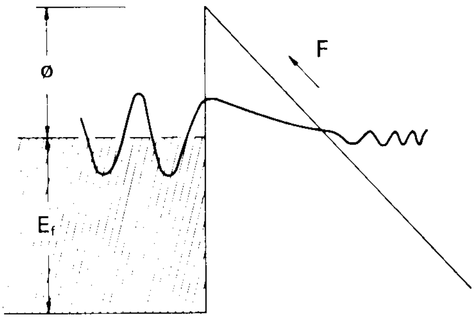


Fig. 1. Fowler-Nordheim tunneling. $J\ AF^2\ \exp\,(-4(2m)^{1\ 2}ø^{3\ 2}\ 3hF)$

The next important development was an attempt to invoke tunneling in order to understand transport properties of electrical contacts between two solid conductors. The problems of metal-to-metal and semiconductor-to-metal contacts are important technically, because they are directly related to electrical switches and rectifiers or detectors.

In 1930, Frenkel (9) proposed that the anomalous temperature independence of contact resistance between metals could be explained in terms of tunneling across a narrow vacuum separation. Holm and Meissner (10) then did careful measurements of contact resistances and showed that the magnitude and temperature independence of the resistance of insulating surface layers were in agreement with an explanation based on tunneling through a vacuum-like space. These measurements probably constitute the first correctly interpreted observations of tunneling currents in solids, (11) since the vacuum-like space was a solid insulating oxide layer.

In 1932, Wilson, (12) Frenkel and Joffe, (13) and Nordheim (14) applied quantum mechanical tunneling to the interpretation of metal-semiconductor contacts - rectifiers such as those made from selenium or cuprous oxide. From a most simplified energy diagram, shown in Fig. 2, the following well-known current-voltage relationship was derived:

$$J=J_s[\exp(eV/kT)-1]$$

Apparently, this theory was accepted for a number of years until it was finally discarded after it was realized that it predicted rectification in the wrong direction for the ordinary practical diodes. It is now clear that, in the usual circumstance, the surface barriers found by the semiconductors in contact with metals, as illustrated in Fig. 2, are much too thick to observe tunneling current. There existed a general tendency in those early days of quantum mechanics to try to explain any unusual effects in terms of tunneling. In many cases, however, conclusive experimental evidence of tunneling was lacking, primarily because of the rudimentary stage of material science.

In 1934, the development of the energy-band theory of solids prompted Zener (15) to propose interband tunneling, or internal field emission, as an explanation for dielectric breakdown. He calculated the rate of transitions

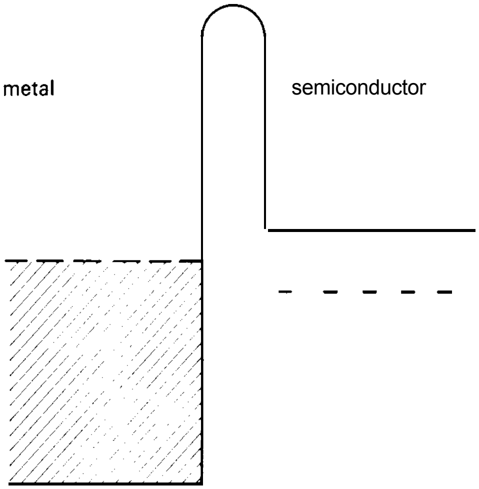


Fig. 2. Early model of metal-semiconductor rectifiers.

from a filled band to a next-higher unfilled band by the application of an electric field. In effect, he showed that an energy gap could be treated in the manner of a potential barrier. This approach was refined by Houston (16) in 1940. The Zener mechanism in dielectric breakdown, however, has never been proved to be important in reality. If a high electric field is applied to the bulk crystal of a dielectric or a semiconductor, avalanche breakdown (electron-hole pair generation) generally precedes tunneling, and thus the field never reaches a critical value for tunneling.

II. Tunnel Diode

Around 1950, the technology of Ge p-n junction diodes, being basic to transistors, was developed, and efforts were made to understand the junction properties. In explaining the reverse-bias characteristic, McAfee et al. (17) applied a modified Zener theory and asserted that low-voltage breakdown in, Ge diodes (specifically, they showed a 10-V breakdown) resulted from interband tunneling from the valence band in the p-type region to the empty conduction band in the n-type region. The work of McAfee et al. inspired a number of other investigations of breakdown in p-n junctions. Results of those later studies (18) indicated that most Ge junctions broke down by avalanche, but by that time the name "Zener diodes" had already been given to the low-breakdown Si diodes., Actually, these diodes are almost always avalanche diodes. In 1957, Chynoweth and McKay (19) examined Si junctions of low-voltage breakdown and claimed that they had finally observed tunneling.

In this circumstance, in 1956, I initiated the investigation of interband tunneling or internal field emission in semiconductor diodes primarily to scrutinize the elctronic structure of narrow (width) p-n junctions. This information, at the time, was also important from a technological point of view.

The built-in field distribution in p-n junctions is determined by the profile of impurities - donors and acceptors. If both the impurity distributions are

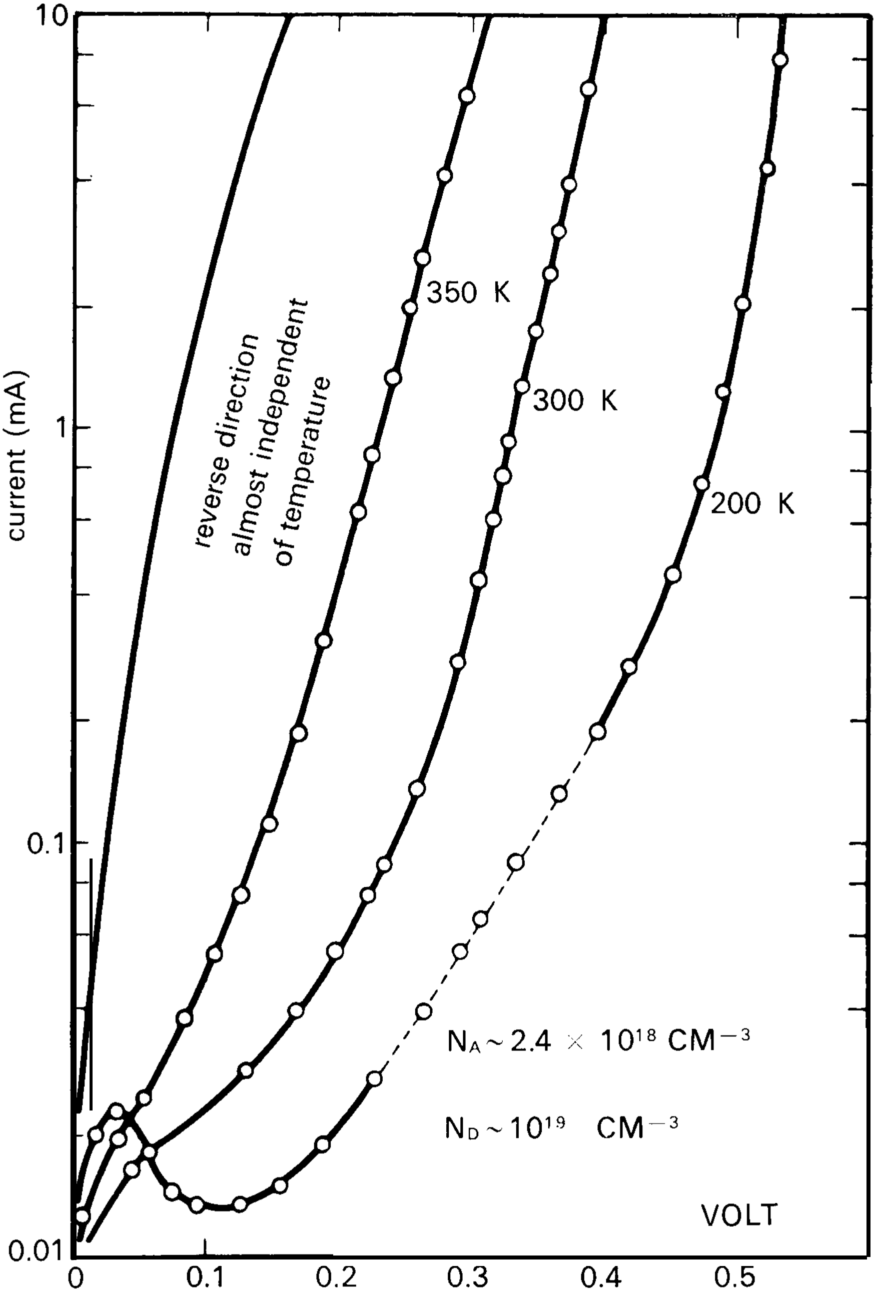


Fig. 3. Semilog plots of current-voltage characteristics in a tunnel diode, where $N_A \sim 2.4 \times 10^{18} cm^{-3}$ and $N_D \sim 10^{19} cm^{-3}$.

assumed to be step functions, the junction width W is given by W = const $[(N_A+N_D)/N_A \cdot N_D]^{1/2} \sim 1/N^{1/2}$, where N_A and N_D are the acceptor and donor concentrations and $N < N_A$ or N_D. Thus, first of all, we attempted to prepare heavily-doped Ge p-n junctions. Both the donor and acceptor concentrations are sufficiently high so that the respective sides of the junctions

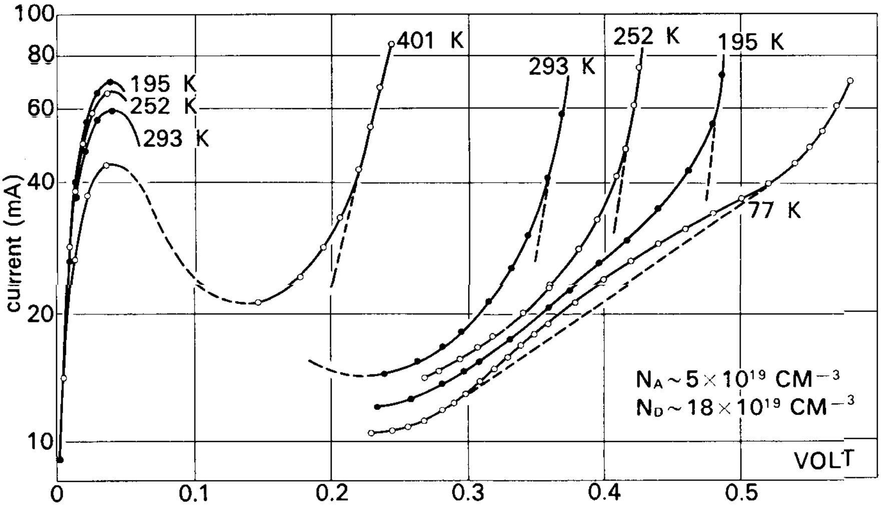


Fig. 4. Semilog plots of current-voltage characteristics in a tunnel diode, where $N_A \sim 5 \times 10^{19} cm^{-3}$ and $N_D \sim 1.8 \times 10^{19} cm^{-3}$.

are degenerate, that is, the fermi energies are located well inside the conduction or valence band.

In this study, we first obtained a backward diode which was more conductive in the reverse direction than in the forward direction. In this respect it agreed with the rectification direction predicted by the previously-mentioned old tunneling rectifier theory. The calculated junction width at zero bias was approximately 200Å, which was confirmed by capacitance measurements. In this junction, the possiblity of an avalanche was completely excluded because the breakdown occurs at much less than the threshold voltage for electron-hole pair production. The current-voltage characteristic at room temperature indicated not only that the major current-flow mechanism was convincingly tunneling in the reverse direction but also that tunneling might be responsible for current flow even in the low-voltage range of the forward direction. When the unit was cooled, we saw, for the first time, a negative resistance, appearing, as shown in Fig. 3. By further narrowing the junction width (thereby further decreasing the tunneling path), through a further increase in the doping level, the negative resistance was clearly seen at all temperatures, as shown in Fig. 4. (20)

The characteristic was analyzed in terms of interband tunneling. In the tunneling process, if it is elastic, the electron energy will be conserved. Figures 5 (a), (b), (c), and (d) show the energy diagrams of the tunnel diode at zero bias and with applied voltages, V_v, E's, and V_3, respectively. As the bias is increased up to the voltage Vi, the interband tunnel current continues to increase, as shown by an arrow in Fig. 5 (b). However, as the conduction band in the n-type side becomes uncrossed with the valence band in the p-type side, with further increase in applied voltages, as shown in Fig. 5 *(c)*, the current decreases because of the lack of allowed states of corresponding ener-

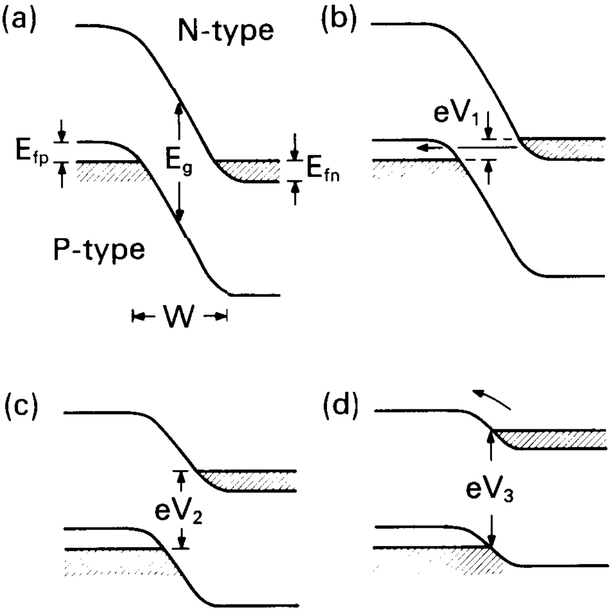


Fig. 5. Energy diagrams at varying bias-conditions in the tunnel diode.

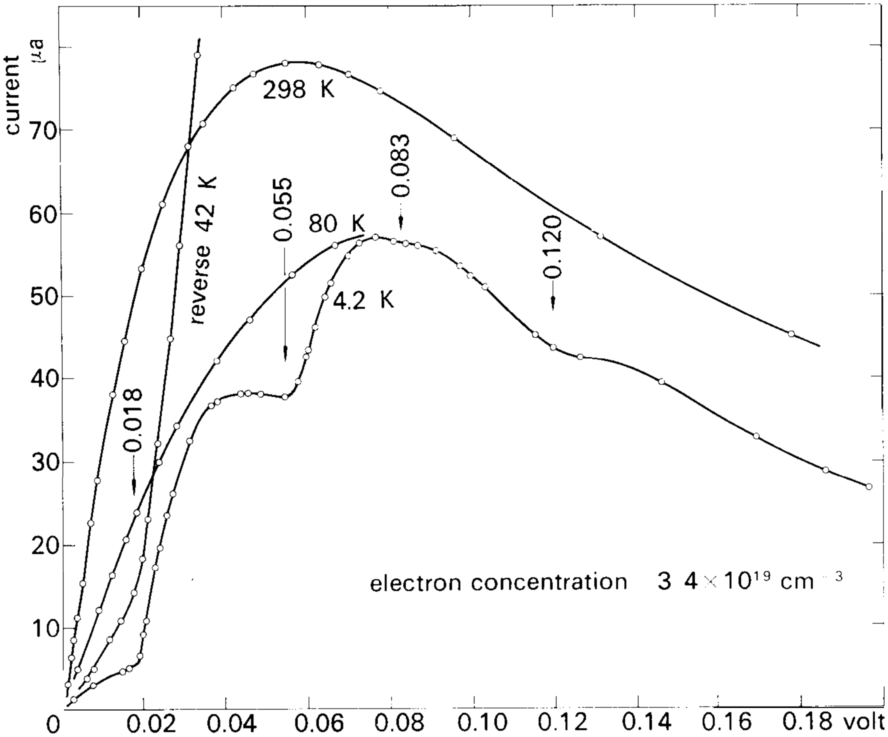


Fig. 6. Current-voltage characteristics in a Si tunnel diode at 4.2, 80 and 298 K.

gies for tunneling. When the voltage reaches V_2, or higher, the normal diffusion (or thermal) current will dominate as in the case of the usual p-n diode. Semiconductor materials other than Ge were quickly explored to obtain tunnel diodes: Si, InSb, GaAs, InAs, PbTe, GaSb, SiC, etc.

In our early study of the Si tunnel diode, (21) a surprisingly fine structure was found in the current-voltage curve at 4.2 K, indicating the existence of inelastic tunneling, as shown in Fig. 6. We were impressed with the fact that four voltages at the singularities shown in the figure agreed almost exactly with four characteristic energies due to acoustic and optical phonons, obtained from the optical absorption spectra (22) and also derived from the analysis of intrinsic recombination radiation (23) in pure silicon. The analysis of tunneling current in detail reveals not only the electronic states in the systems involved, but also the interactions of tunneling electrons with phonons, photons, plasmons, or even vibrational modes of molecular species in barriers. (24) As a result of the rich amount of information which can be obtained from a study of tunneling processes, a field called tunneling spectroscopy has emerged.

III. Negative Resistance in Metal-Oxide-Semiconductor Junctions

This talk, however, is not intended as a comprehensive review of the many theoretical and experimental investigations of tunneling, which is available elsewhere. (25) Instead, I would like to focus on only one aspect for the rest of the talk: negative resistance phenomena in semiconductors which can be observed in novel tunnel structures.

Differential negative resistance occurs only in particular circumstances, where the total number of tunneling electrons transmitted across a barrier structure per second decreases, rather than increases as in the usual case, with an increase in applied voltage. The negative resistance phenomena themselves are not only important in solid-state electronics because of possible signal amplification, but also shed light on some fundamental aspects of tunneling.

Before proceeding to the main subject, I would like to briefly outline the independent-electron theory of tunneling. (26) In tunneling, we usually deal with *a one-dimensional potential barrier V(x).* The transmission coefficient *D* for such a barrier is defined as the ratio of the intensity of the transmitted electron wave to that of the incident wave. The most common approximation for *D* is the use of the semiclassical WKB form

$$D(E_x) = \exp \left[-2/\hbar \int_{x_1}^{x_2} (2m(V-E_x))^{1/2}\, dx\right] \qquad (1)$$

where E_x is the kinetic energy in the direction normal to the barrier, and the quantities x_1 and x_2 are the classical turning points of an electron of energy E_x at the edges of the potential barrier. If the boundary regions are sharp, we first construct wave functions by matching values of functions as well as

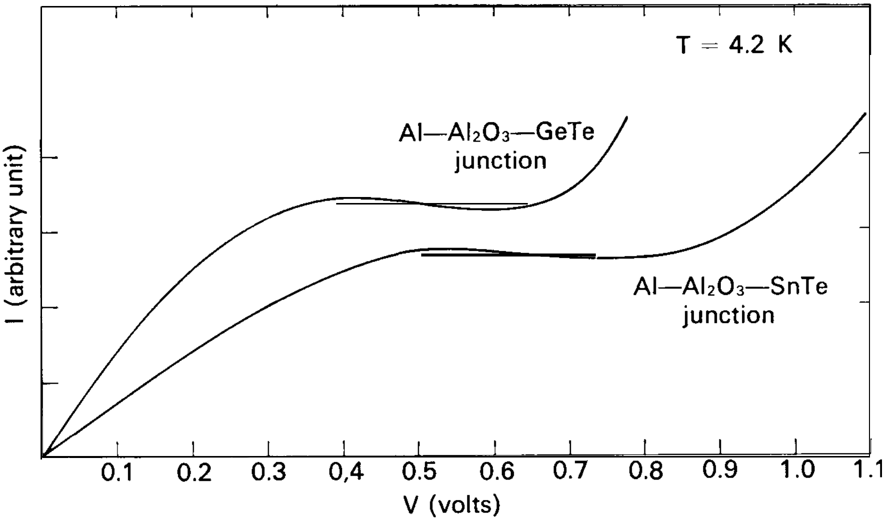


Fig. 7. Current-voltage characteristics in SnTe and GeTe tunnel junctions at 4.2 K.

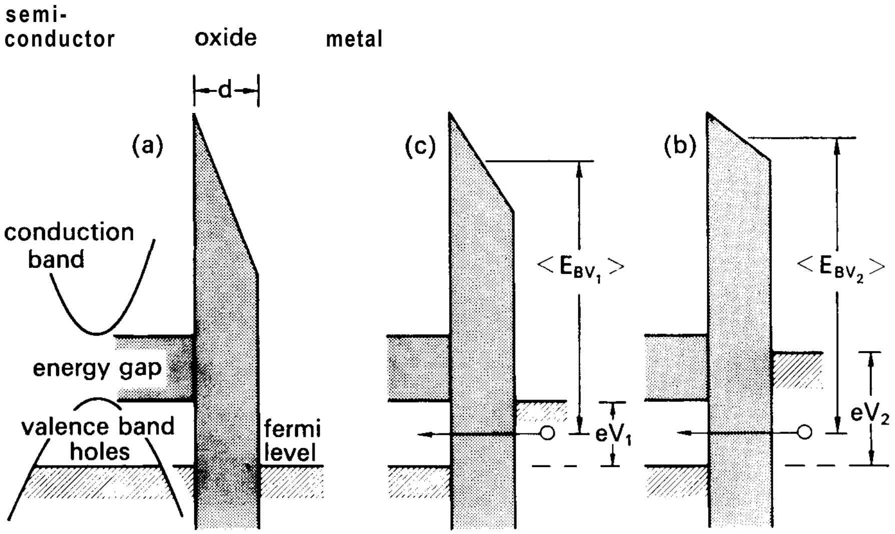
their derivatives at each boundary, then calculate the transmission coefficient D.

The tunneling expression should include two basic conservation laws: 1) Conservation of the total electron energy; and 2) Conservation of the component of the electron wave vector parallel to the plane of the junction. The velocity of an incident electron associated with a state of wave number k_x is given by 1/h $\partial E/\partial k_x$ in a one-particle approximation. Then, the tunneling current per unit area is written by

$$J = 2e/(2\pi)^3 \int\int\int D(E_x)(f(E) - f(E'))\, 1/\hbar\, \partial E/\partial k_x\, dk_x\, dk_y\, dk_z \qquad (2)$$

where f is the fermi distribution function or occupation probability, and E and E′ are the energy of the incident electron and that of the transmitted one, respectively. The front factor $2/(2\pi)^3$ comes from the fact that the volume of a state occupied by two electrons of the opposite spin is $(2\pi)^3$in the wave-vector space for a unit volume crystal.

The previously-mentioned tunnel diode is probably the first structure in which the negative resistance effect was observed. But, now, I will demonstrate that a similar characteristic can be obtained in a metal-oxide-semiconductor tunnel junction, (27) where the origin of the negative resistance is quite different from that in the tunnel diode. The semiconductors involved here (SnTe and GeTe) are rather unusual-more metallic than semiconducting; both of them are nonstoichiometric and higly p-type owing to high concentrations of Sn or Ge vacancies with typical carrier concentrations about 8 x $10^{20}$and 2 x 10^{20}c m^{-3}, respectively. The tunnel junctions were prepared by evaporating SnTe or GeTe onto an oxidized evaporated stripe of Al on quartz or sapphire substrates. In contrast to the p-n junction diodes, all ma-



Fig, 8. Energy diagrams at varying bias-conditions in Al-Al$_2$O$_3$)SnTe or-GeTe tunnel junctions.

terials involved in these junctions are polycrystalline, although the Al oxide is possibly amorphous.

Figure 7 illustrates the current-voltage curves at 4.2 K of typical SnTe and GeTe junctions and Fig. 8 shows their energy diagrams at zero bias, and at applied voltages $V_1$and $V_2$from the left to the right. As is the case in the tunnel diode, until the bias voltage is increased such that the fermi level in the metal side coincides with the top of the valence band in the semiconductor side (Fig. 8 (b)), the tunnel current continues to increase. When the bias voltage is further increased (Fig. 8 (c)), however, the total number of empty allowed states or holes in the degenerate p-type semiconductor is unchanged, whereas the tunneling barrier height is raised, for instance from $E_{BV_1}$to E_{BV_2}, resulting in a decrease in tunneling probability determined by the exponential term, $e^{-\lambda}$, where $\lambda \sim 2d\ (2mE_{BV})^{1/2}/\hbar$, and E_{BV} and d are the barrier height and width, respectively. Thus a negative resistance is exhibited in the current-voltage curve. When the bias voltage becomes higher than the level corresponding to the bottom of the conduction band in the semiconductor, a new tunneling path from the metal to the conduction band is opened and one sees the current again increasing with the voltage. The rectification direction in this junction is again backward as is the case in the tunnel diode.

We might add that, in this treatment, the tunneling exponent is assumed to be determined only by the energy difference between the bottom of the conduction band in the oxide and the metal fermi energy. This assumption should be valid because this energy difference is probably much smaller than that between the top of the valence band in the oxide and the metal fermi energy.

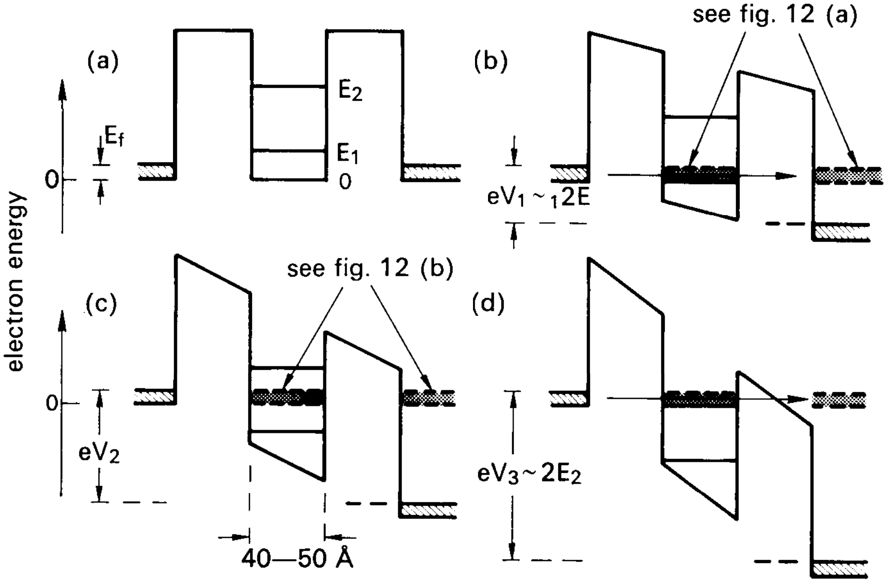


Fig. 9. Energy diagrams at varying bias-conditions in a double-barrier tunnel junction, indicating the resonant transmission in (b) and (d).

IV. Negative resistance due to Resonant Transmission

It has been known that there is a phenomenon called the resonant transmission. Historically, resonant transmission was first demonstrated in the scattering of electrons by atoms of noble gases and is known as the Ramsauer effect. In many textbooks (28) on quantum mechanics, the resonant transmission in tunneling or scattering is one of the more favored topics. In a one-dimensional double potential barrier, (29) the narrow central potential well has weakly-quantized (or quasi-stationary) bound states, of which the energies are denoted by $E_1$and E_2in Fig. 9 (a). If the energy of incident electrons coincides with these energies, the electrons may tunnel through both barriers without any attenuation. As seen in Fig. 10 (two curves at V = 0), the transmission coefficient reaches unity at the electron energy $E = E_1$or E_2. Since $E_1$is a more strongly quantized state than E_2, the resonance peak at $E_1$is much sharper than that at E_2. Although this sharpness depends upon the barrier thickness, one can achieve at some energy a resonance condition of 100% transmission, whatever thickness is given to the two barriers.

This effect is quite intriguing because the transmission coefficient (or the attenuation factor) for two barriers is usually thought of as the product of two transmission coefficients, one for each barrier, resulting in a very small value for overall transmission. The situation, however, is somewhat analogous to the Fabry-Perot type interference filter in optics. The high transmissivity arises because, for certain wavelengths, the reflected waves from inside interfere destructively with the incident waves, so that only a transmitted wave remains.

This resonating condition can be extended to a periodic barrier structure. In the Kronig-Penney model of a one-dimensional crystal which consists of a series of equally-spaced potential barriers, it is well known that allowed bands of perfect transmission are separated by forbidden bands of attenuation. These one-dimensional mathematical problems can often be elegantly treated, leading to exact analytical solutions in textbooks of quantum mechanics. Many of these problems, however, are considered to be pure mathematical fantasy, Ear from reality.

We, recently, initiated an experimental project to materialize one-dimensional potential barriers in monocrystalline semiconductors in order to observe the predicted quantum-mechanical effects. (30) We choose n-type GaAs as a host semiconductor or a matrix in which potential barriers with the height of a fraction of one electron volt are made by inserting thin layers of $Ga_{1-r\ x}Al_xAs$ or AlAs. Because of the similar properties of the chemical bond of Ga and Al together with their almost equal ion size, the introduction of AlAs into GaAs makes the least disturbance to the quality of single crystals. And yet the difference in the electronic structure between the two materials makes a sharp potential barrier inside the host semiconductor. We prepare the multi-layer structure with the technique of molecular beam epitaxy in ultra-high vacuum environment. Precise control of thickness and composition has been achieved by using a process control computer. (31)

With this facility, a double potential barrier structure has been prepared, (32) in which the barrier height and width are about 0.4 eV and a few tens of angstroms, respectively, and the width of the central well is as narrow as 40-50Å. From these data, the first two energies, $E_1$and E_2, of the weakly-quantized states in the well are estimated to be 0.08 and 0.30 eV.

We have measured the current-voltage characteristic as well as the conductance dI/dV as a function of applied voltages in this double tunnel junction. The results at 77 K are shown in Fig. 11, and they clearly indicate two singularities in each polarity and even show a negative resistance around +0.8 volt or -0.55 volt. The applied voltages at the singularities, averaged in both polarities, are roughly twice as much as the calculated bound-state energies. This general feature is not much different a 4.2 K, although no structure is seen at room temperature.

The energy diagrams at zero bias and at applied voltages V_1, $V_2$and V_3 are shown in Fig. 9. The electron densities on both the left and right GaAs sides are about $10^{18}cm^{-3}$which gives a fermi energy of 0.04 eV at zero temperature. These electrons are considered to be classical free carriers with the effective mass, m^*, of which the kinetic energy E is given by

$$E=\frac{\hbar^2}{2m^*}\ (k_x{}^2+k_y{}^2+k_z{}^2).$$

On the other hand, the electrons in the central well have the weakly-quantized levels, E_l, E_2, . . ., for motion in the x direction perpendicular to the walls with a continuum for motion in the y-z plane parallel to the walls. These

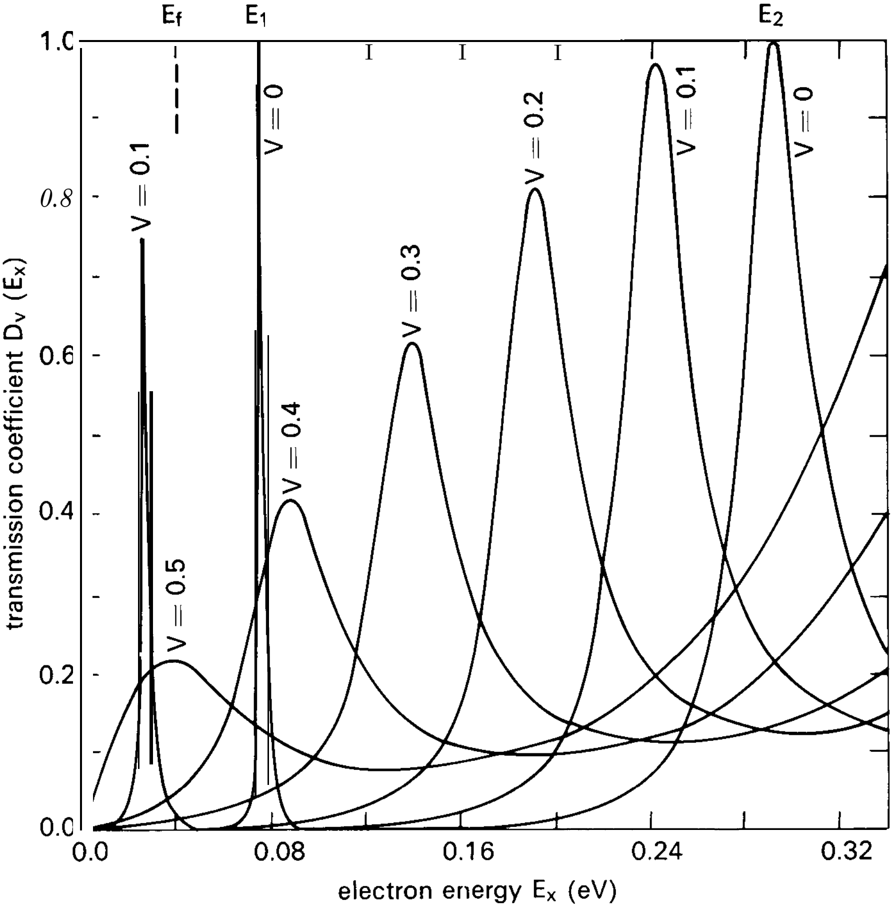


Fig. 10. Transmission coefficient versus electron energy, indicating the resonant transmission.

electrons are nearly two-dimensional, which is to say the kinetic energy E is given by

$$E = E_i + \frac{\hbar^2}{2m^*}\left(k_y^2 + k_z^2\right)$$

An approximation is made that the same electron effective mass, m*, exists throughout the structure. Then an expression for the tunneling current in this structure (33) can be derived in the framework of the previously-described tunneling formalism in Eq. 2. Using $\partial E/\partial k_x = \partial E_x/\partial k_x$, $2\pi k_t \mathrm{d}k_t = \mathrm{d}k_y \mathrm{d}k_z$ and T (temperature) = 0, the current is given by

$$J = e/2\pi^2\hbar \int_0^{E_f} D(E_x) \int_0^{(2m^*(E_f - E_x))^{1/2}\hbar} k_t dk_t dE_x$$

$$-e/2\pi^2\hbar \int_0^{E_f - eV} D(E_x) \int_0^{(2m^*(E_f - E_x - eV))^{1/2}/\hbar} k_t \, dk_t \, dE_x, \qquad (3)$$

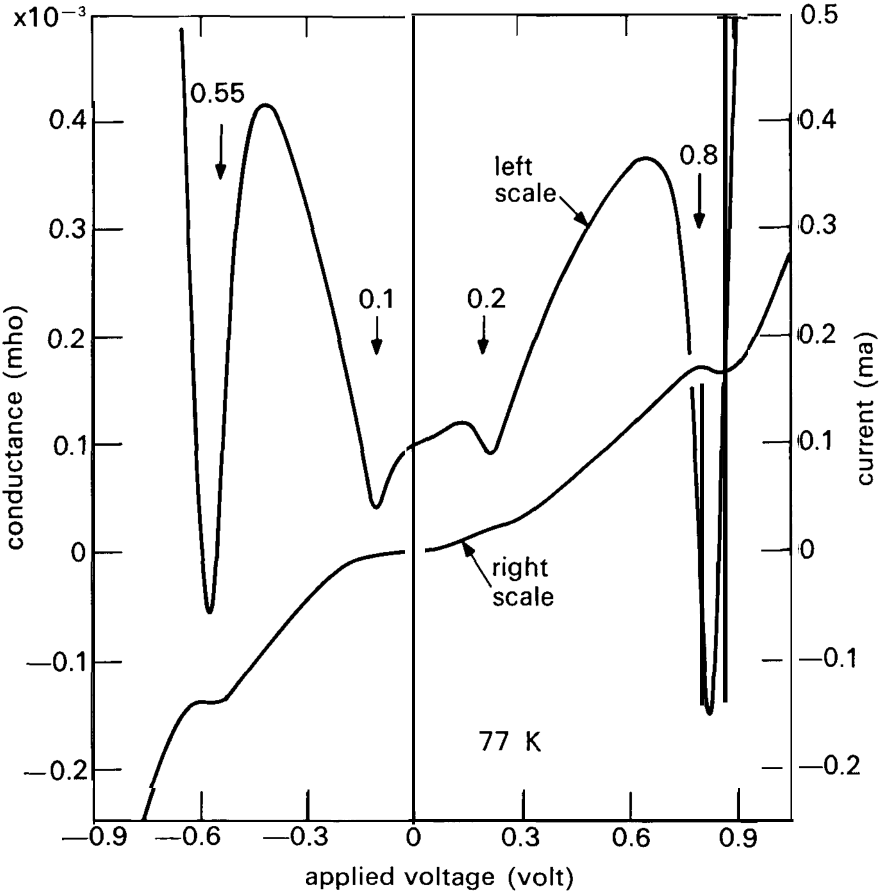


Fig. 11. Current, I, and conductance, dI/dV, versus voltage curves in a double barrier tunnel junction.

where V is the applied voltage, on which the transmission coefficient $D(E_x)$ depends. The above expression can be integrated over the transverse wave number k_t, giving

$$J = em^*/2\pi^2\hbar^3 \int_0^{E_f} D_V(E_x)(E_f - E_x)\, dE_x$$

$$- em^*/2\pi^2\hbar^3 \int_0^{E_f - eV} D_V(E_x)(E_f - E_x - eV)\, dE_x \qquad (4)$$

In both Eqs. 3 and 4, the second term is nonzero only for $eV < E_f = 0.04$ eV.

Now, the transmission coefficient D_v (E_x) can be derived for each applied voltage from wave functions which are constructed by matching their values as well as derivatives at each boundary. Figure 10 shows one example of calculated D as a function of the electron energy for applied voltages

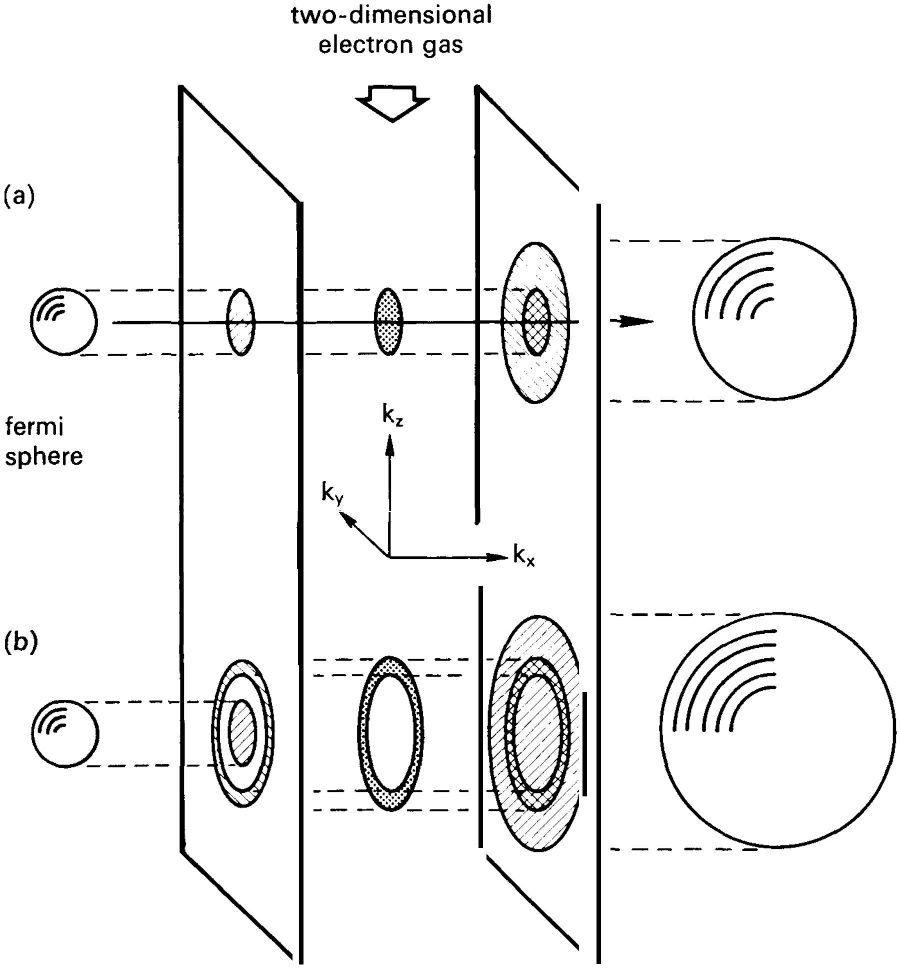


Fig. 12. Construction of shadows of energy surfaces on two k_y-$k_z$planes corresponding to two barriers.

between zero and 0.5 volt. The energy zero is taken at the bottom of the conduction band on the left as shown in Fig. 9. In this example, the well width is taken to be 45Å and the barrier height 0.4 eV at zero bias. The square shape for barriers and well is assumed for simplicity of calculation, although they are actually trapezoidal at any applied voltage.

Referring to Figs. 9 and 10, both the absolute values and the positions in energy for the maxima of the transmission coefficient decrease with increasing applied voltages, the origin of energy being the conduction band edge for the left outer GaAs layer. The current maxima occur at applied voltages such that the electron energies on the left coincide with the bound-state energies, as illustrated in Figs. 9 (b) and (d). This resonant transmission has been experimentally verified as shown in Fig. 11. The transmission coefficient itself at this resonance, however, is appreciably less than unity as indicated in

Fig. 10, primarily because of the asymmetric nature of the potential profile at applied voltages.

To gain an insight into this tunneling problem, particularly in view of the transverse wave-vector conservation (specular tunneling), a representation in the wave-vector space is useful and is shown in Fig. 12. Two k_y-k_z, planes are shown parallel to the junction plane, corresponding to the two barriers. Figures 12 (a) and (b) show two different bias-voltage conditions. First, the Fermi sphere on the left is projected on the first screen, making a circle. A similar projection, of the two-dimensional electrons in the central well which have the same total energies as electrons in the Fermi sphere on the left at the particular applied voltage, will form a circle (Fig. 12 (a)), or a ring (Fig. 12 (b)), depending upon the value of applied voltage. If the two projected patterns have no overlap, there will be no specular tunneling current. The situation on the right screen is slightly different, since an energy sphere on the right, in which electrons have the same total energies as electrons in the Fermi sphere on the left, is rather large; mereover, its size will be increased as the applied voltage increases. Thus in this case the two projected patterns always overlap. Figures 12 (a) and (b) correspond to the bias conditions in Figs. 9 (b) and (c), respectively. With an increase in applied voltage from V_1 to V_2, the current will decrease because of a disappearance of overlapping regions, thereby causing a negative resistance. Since the current density is dependent upon the half-width of the resonant peaks shown in Fig. 10, we have observed a clear negative resistance associated with the second bound-state which is not swamped by possible excess currents arising for a variety of reasons.

V. Periodic Structure -Superlattice

The natural extension of double barriers will be to construct a series of tunnel junctions by a periodic variation of alloy composition. (30) By using the same facilities for computer-controlled molecular beam epitaxy, we tried to prepare a Kronig-Penney type one-dimensional periodic structure-a man-made superlattice with a period of 100Å. (31) The materials used here are again GaAs and AlAs or $Ga_{1-x}Al_xAs$.

The composition profile of such a structure (34) has been verified by the simultaneous use of ion sputter-etching of the specimen surface and Auger electron spectroscopy and is shown in Fig. 13. The amplitudes of the Al Auger signals serve as a measure of Al concentration near the surface within a sampling depth of only 10Å or so. The damping of the oscillatory behavior evident in the experimental data is not due to thermal diffusion or other reasons but due to a surface-roughening effect or non-uniformity in the sputter-etching process. The actual profile, therefore, is believed to be one which is illustrated by the solid line in Fig. 13. This is certainly one of the highest resolution structures ever built in monocrystalline semiconductors.

It should be recognized that the period of this superlattice is ~ 100Å--still large in comparison with the crystal lattice constant. If this period ι, how-

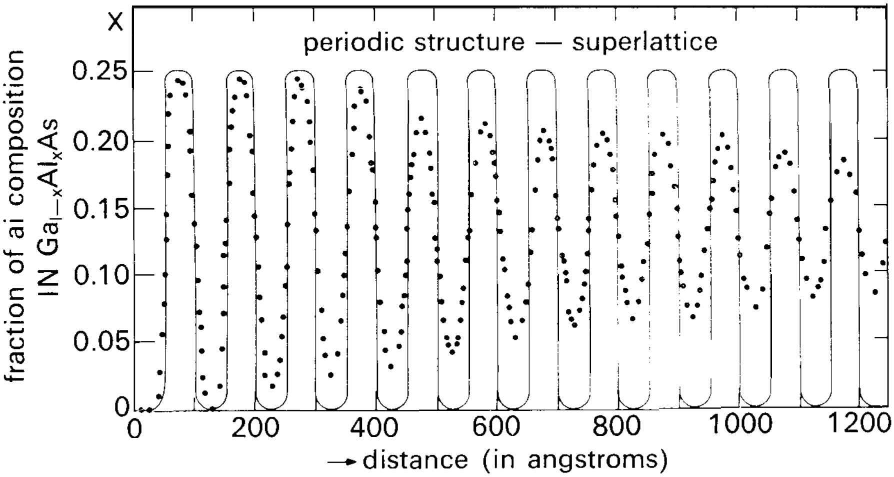


Fig. 13. Composition profile of a superlattice structure measured by a combination of ion sputter-etching and Auger electron spectroscopy.

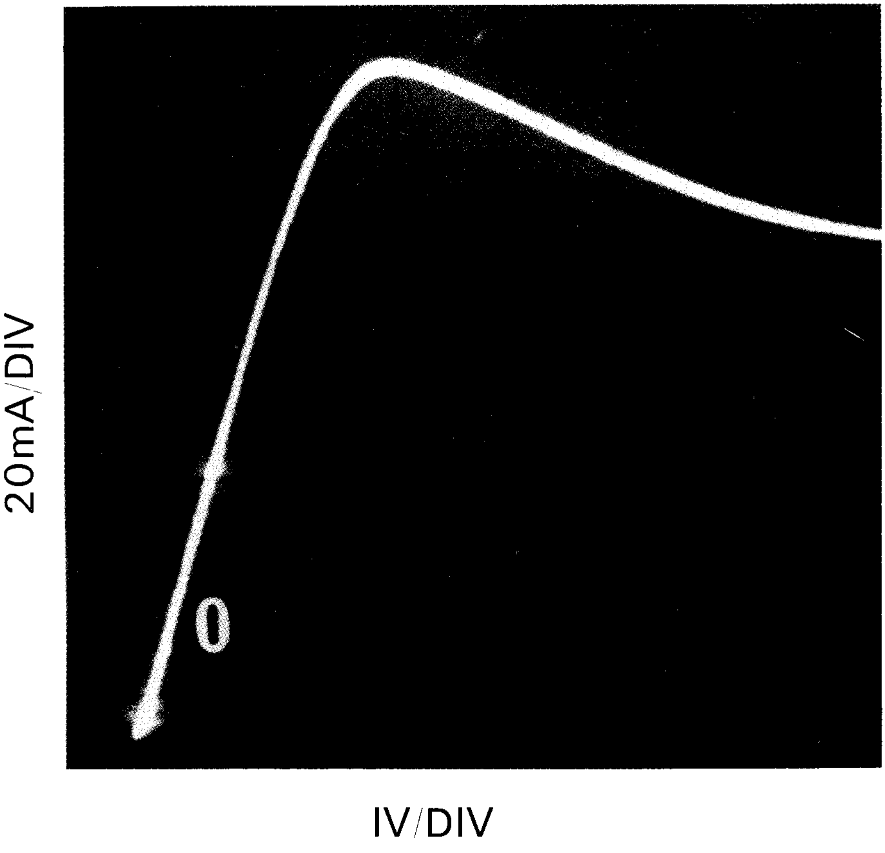


Fig. 14. Current-voltage characteristic at room temperature of a 70Å-period, GaAs-$Ga_{0,5}Al_{0,5}As$ superlattice.

ever, is still shorter than the electron mean free path, a series of narrow allowed and forbidden bands is expected, due to the subdivision of the original Brillouin zone into a series of minizones. If the electron scattering time τ, and an applied electric field F, meet a threshold condition: $eF\tau\iota/\hbar > 1$, the combined effect of the narrow energy band and the narrow wave-vector zone makes it possible for electrons to be excited beyond an inflection point in the energy-wave vector relation. This would result in a negative resistance for electrical transport in the direction of the superlattice. This can be seen in another way. The de Broglie wavelength of conduction electrons having an energy of, for instance, 0.03 eV in n-type GaAs (the effective mass ~ 0.1 m). is of the order of 200Å. Therefore, an interaction of these electron waves with the Kronig-Penney type potential with a period of 100Å can be expected, and will give rise to a nonlinear transport property.

We have begun to observe such current-voltage characteristic as shown in Fig. 14. The observed negative resistance may be interesting not only from the scientific aspect but also from a practical viewpoint because one can expect, at least theoretically, that the upper limit of operating frequencies would be higher than that for any known semiconductor devices.

VI. Conclusion

I am, of course, deeply aware of important contributions made by many colleagues and my friends throughout this long journey. The subject of Section II was carried out when I was in Japan and all the rest (35) has been performed in the United States of America. Since my journey into tunneling is still continuing, I do not come to any conclusions in this talk. However, I would like to point out that many high barriers exist in this world: Barriers between nations, races and creeds. Unfortunately, some barriers are thick and strong. But I hope, with determination, we will find a way to tunnel through these barriers easily and freely, to bring the world together so that everyone can share in the legacy of Alfred Nobel.

References

1. de Broglie, L., Nature *112, 540* (1923) ; Ann. de Physique (10 Ser.) 3, 22 (1925).
2. Schrödinger, E., Ann. d. Physik (4. Folge) 79, 361, 489 (1926).
3. Oppenheimer, J. R., Phys. Rev. 31, 66 (1928); Proc. Nat'l. Acad. Sci. U.S. 14, 363 (1928).
4. Fowler, R. H. and Nordheim, L., Proc. Roy. Soc. (London) *A 119*, 173 (1928).
5. Lilienfeld, J. E. Physik. Z. 23, 506 (1922).
6. Gamow, G., Physik, Z., 51, 204 (1928).
7. Gurney, R.W. and Condon, E. U., Nature 222,439 (1928).
8. Rice, O. K. Phys. Rev. 34, 1451 (1929).
9. Frenkel, J., Phys. Rev. 36, 1604 (1930).
10. Helm, R. and Meissner, W., Z. Physik 74, 715 (1932), 86, 787 (1933).
11. Holm, R., Electric Contacts, Springer-Verlag, New York, 1967 p.118.
12. Wilson, A. H., Proc. Roy. Soc. (London) *A 136, 487* (1932).
13. Frenkel, J. and Joffe, A., Physik. Z. Sowjetunion I, 60 (1932).
14. Nordheim, L., Z. Physik 7.5, 434 (1932).

15. Zener, C., Proc. Roy. Soc. (London) 14.5, 523 (1934).
16. Houston, W. V., Phys. Rev. 57, 184 (1940).
17. McAfee, K. B., Ryder, E. J., Shockley, W. and Sparks, M., Phys. Rev. 83, 650 (1951).
18. McKay, K. G. and McAfee, K. B. Phys. Rev. 91, 1079 (1953);
McKay, K. G., Phys. Rev. 94,877 (1954) ;
Miller, S. L., Phys. 99, 1234 (1955).
19. Chynoweth, A. G. and McKay., K. G. Phys., Rev. *106,418* (1957).
20. Esaki, L., Phys. Rev. 109, 603 (1958);
Esaki, L., Solid State Physics in Electronics and Telecommunications, Proc. of Int. Conf. Brussels, 1958 (Desirant, M. and Michels, J. L., ed.), Vol. 1, Semiconductors, Part I, Academic Press, New York, 1960, p. 514.
21. Esaki, L. and Miyahara, Y., unpublished;
Esaki, L. and Miyahara, Y., Solid-State Electron. *1,* 13 (1960) ;
Holonyak, N., Lesk, I. A., Hall, R. N., Tiemann, J. J. and Ehrenreich, H., Phys. Rev. Letters 3, 167 (1959).
22. Macfarlane, G. G., McLean, T. P., Quarrington, J. E. and Roberts, V., J. Phys. Chem. Solids 8, 388 (1959).
23. Haynes, J, R., Lax, M. and Flood, W. F., J. Phys. Chem. Solids 8, 392 (1959).
24. See, for instance, Tunneling Phenomena in Solids edited by Burstein, E. and Lundqvist, S., Plenum Press, New York, 1969.
25. Duke, C. B., Tunneling in Solids, Academic Press, New York, 1969.
26. Harrison, W. A., Phys, Rev. 123, 85 (1961).
27. Esaki, L. and Stiles, P. J., Phys. Rev. Letters 16, 1108 (1966) ; Chang, L. L., Stiles, P. J. and Esaki, L., J. Appl. Phys. 38, 4440 (1967) ; Esaki, L., J. Phys. Soc. Japan, Suppl. 21,589 (1966).
28. See, for instance, Bohm, D., Quantum Theory, Prentice-Hall, New Jersey, 1951.
29. Kane, E. O., page 9-l 1 in reference 24.
30. Esaki, L. and Tsu, R., IBM J. Res. Develop. 24, 61 (1970); Esaki, L., Chang, L. L., Howard, W. E. and Rideout, V. L., Proc. 11th Int. Conf. Phys. Semicond., Warsaw, Poland, 1972, p. 431.
31. Chang, L. L., Esaki, L., Howard, W. E. and Ludeke, R., J. Vac. Sci. Technol. 10, 11 (1973);
Chang, L. L., Esaki, L., Howard, W. E., Ludeke, R. and Schul, G., J. Vac. Sci. Technol. 10, 655 (1973).
32. Chang, L. L., Esaki, L. and Tsu, R., to be published.
33. Tsu, R. and Esaki, L., Appl. Phys. Letters 22, 562 (1973).
34. Ludeke, R., Esaki, L. and Chang, L. L., to be published.
35. Partly supported by Army Research Office, Durham, N. C.

ELECTRON TUNNELING AND SUPERCONDUCTIVITY

Nobel Lecture, December 12, 1973

by

IVAR GIÆVER

General Electric Research and Development Center, Schenectady, N.Y., USA.

In my laboratory notebook dated May 2, 1960 is the entry: "Friday, April 22, I performed the following experiment aimed at measuring the forbidden gap in a superconductor." This was obviously an extraordinary event not only because I rarely write in my notebook, but because the success of that experiment is the reason I have the great honor and pleasure of addressing you today. I shall try in this lecture, as best I can, to recollect some of the events and thoughts that led to this notebook entry, though it is difficult to describe what now appears to me as fortuitous. I hope that this personal and subjective recollection will be more interesting to you than a strictly technical lecture, particularly since there are now so many good review articles dealing with superconductive tunneling. [1,2]

A recent headline in an Oslo paper read approximately as follows: "Master in billiards and bridge, almost flunked physics - gets Nobel Prize." The paper refers to my student days in Trondheim. I have to admit that the reporting is reasonably accurate, therefore I shall not attempt a "cover up", but confess that I almost flunked in mathematics as well. In those days I was not very interested in mechanical engineering and school in general, but I did manage to graduate with an average degree in 1952. Mainly because of the housing shortage which existed in Norway, my wife and I finally decided to emigrate to Canada where I soon found employment with Canadian General Electric. A three year Company course in engineering and applied mathematics known as the A, B and C course was offered to me. I realized this time that school was for real, and since it probably would be my last chance, I really studied hard for a few years.

When I was 28 years old I found myself in Schenectady, New York where I discovered that it was possible for some people to make a good living as physicists. I had worked on various Company assignments in applied mathematics, and had developed the feeling that the mathematics was much more advanced than the actual knowledge of the physical systems that we applied it to. Thus, I thought perhaps I should learn some physics and, even though I was still an engineer, I was given the opportunity to try it at the General Electric Research Laboratory.

The assignment I was given was to work with thin films and to me films meant photography. However I was fortunate to be associated with John Fisher who obviously had other things in mind. Fisher had started out as a mechanical engineer as well, but had lately turned his atten-

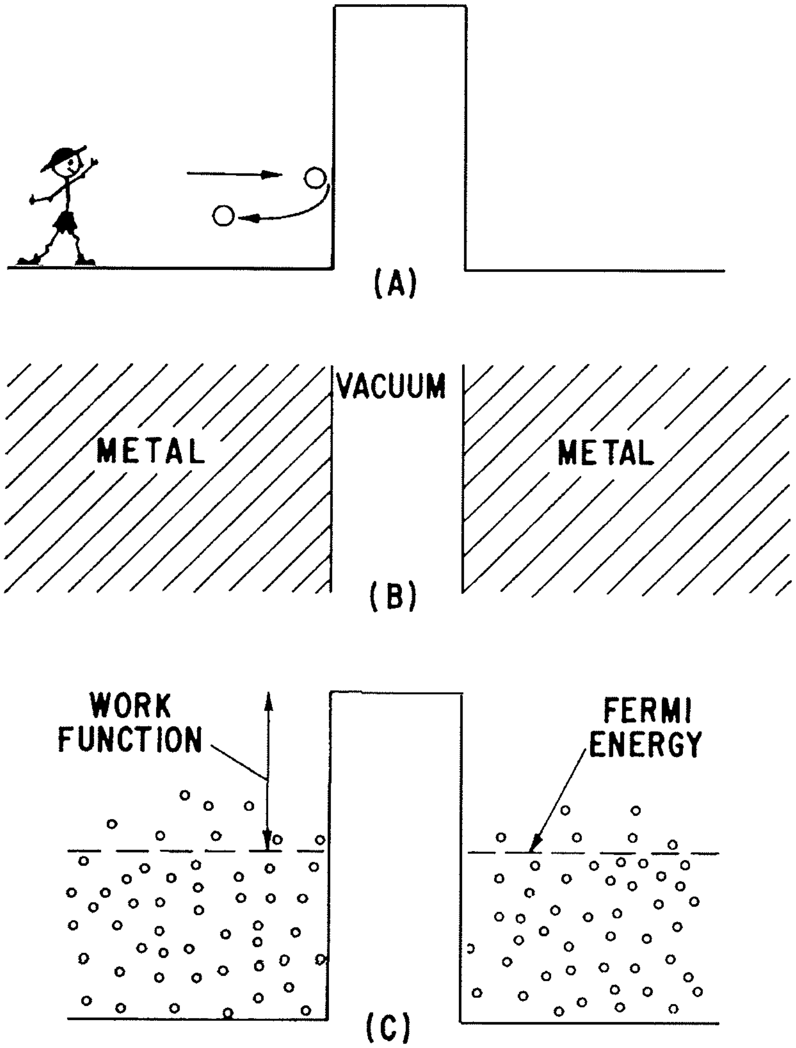


Fig. 1.
A. If a man throws a ball against a wall the ball bounces back. The laws of physics allow the ball to penetrate or tunnel through the wall but the chance is infinitesimally small because the ball is a macroscopic object. B. Two metals separated by a vacuum will approximate the above situation. The electrons in the metals are the "balls", the vacuum represents the wall. C. A pictorial energy diagram of the two metals. The electrons do not have enough energy to escape into the vacuum. The two metals can, however, exchange electrons by tunneling. If the metals are spaced close together the probability for tunneling is large because the electron is a microscopic particle.

tion towards theoretical physics. He had the notion that useful electronic devices could be made using thin film technology and before long I was working with metal films separated by thin insulating layers trying to do tunneling experiments. I have no doubt that Fisher knew about Leo Esaki's tunneling experiments at that time, but I certainly did not. The concept that a particle can go through a barrier seemed sort of strange to me, just struggling with quantum mechanics at Rensselaer Polytechnic Institute in Troy, where I took formal courses in Physics. For an engineer it sounds rather strange that if you

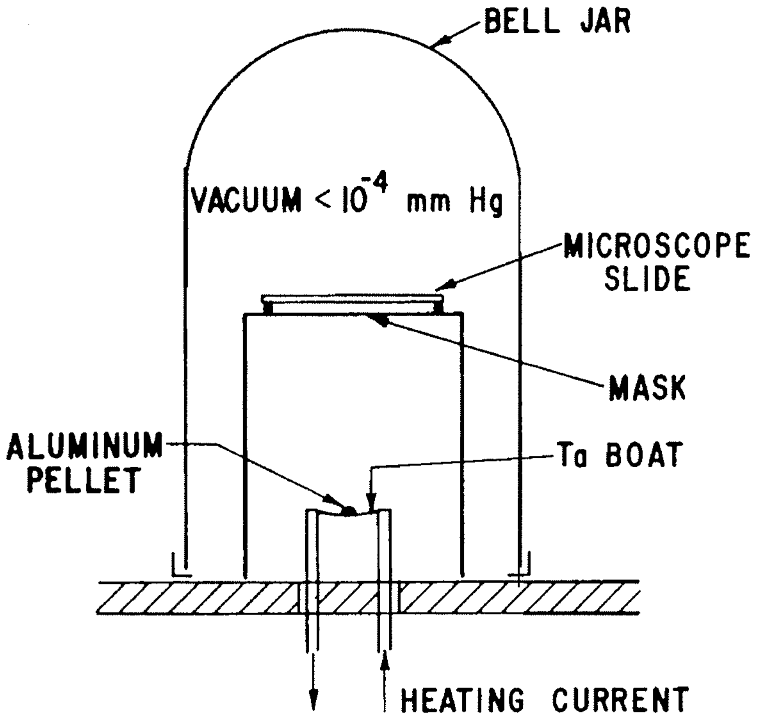


Fig. 2.
A schematic drawing of a vacuum system for depositing metal films. For example, if aluminum is heated resistively in a tantalum boat, the aluminum first melts, then boils and evaporates. The aluminum vapor will solidify on any cold substrate placed in the vapor stream. The most common substrates are ordinary microscope glass slides. Patterns can be formed on the slides by suitably shielding them with a metal mask.

throw a tennis ball against a wall enough times it will eventually go through without damaging either the wall or itself. That must be the hard way to a Nobel Prize! The trick, of course, is to use very tiny balls, and lots of them. Thus if we could place two metals very close together without making a short, the electrons in the metals can be considered as the balls and the wall is represented by the spacing between the metals. These concepts are shown in Figure 1. While classical mechanics correctly predicts the behavior of large objects such as tennis balls, to predict the behavior of small objects such as electrons we must use quantum mechanics. Physical insight relates to everyday experiences with large objects, thus we should not be too surprised that electrons sometimes behave in strange and unexpected ways.

Neither Fisher nor I had much background in experimental physics, none to be exact, and we made several false starts. To be able to measure a tunneling current the two metals must be spaced no more than about 100 Å apart, and we decided early in the game not to attempt to use air or vacuum between the two metals because of problems with vibration. After all, we both had training in mechanical engineering! We tried instead to keep the two metals apart by using a variety of thin insulators made from Langmuir films and from Formvar. Invariably, these films had pinholes and the mercury

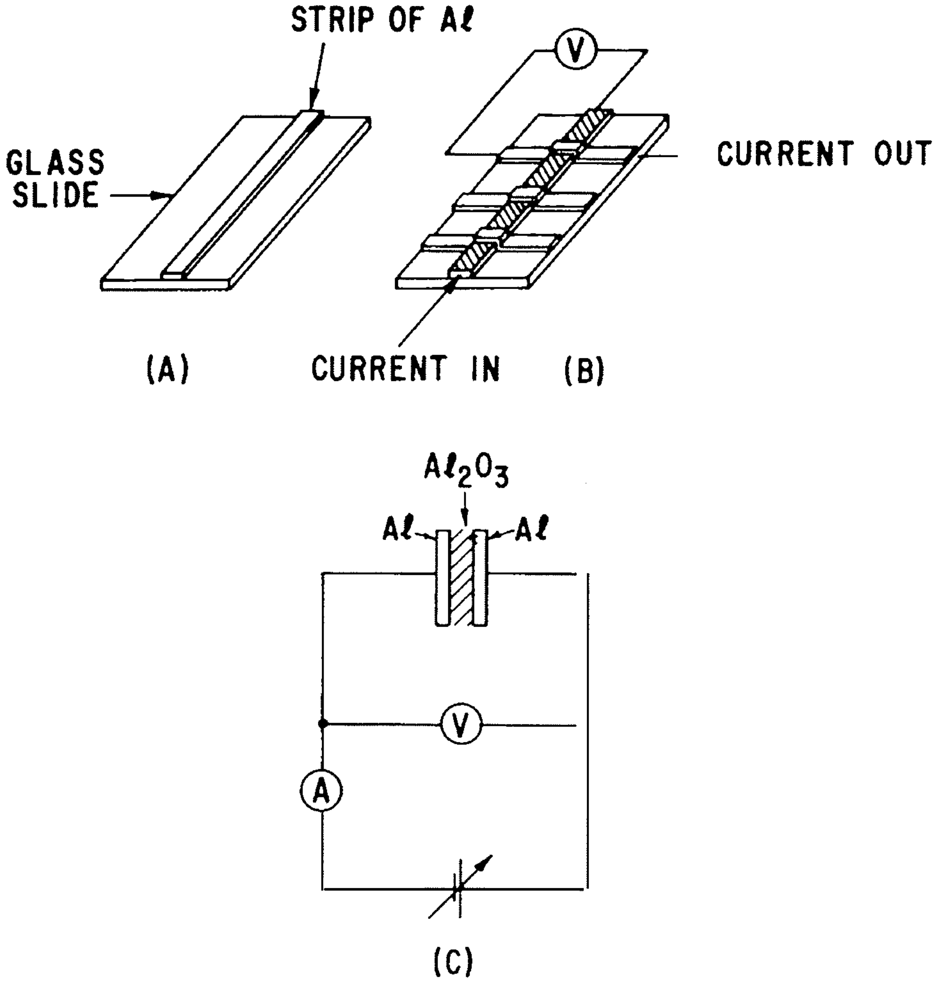


Fig. 3.
A. A microscope glass slide with a vapor deposited aluminum strip down the middle. As soon as the aluminum film is exposed to air, a protective insulating oxide forms on the surface. The thickness of the oxide depends upon such factors as time, temperature and humidity. B. After a suitable oxide has formed, cross strips of aluminum are evaporated over the first film, sandwiching the oxide between the two metal films. Current is passed along one aluminum film up through the oxide and out through the other film, while the voltage drop is monitored across the oxide. C. A schematic circuit diagram. We are measuring the current-voltage characteristics of the capacitor-like arrangement formed by the two aluminum films and the oxide. When the oxide thickness is less than 50Å or so, an appreciable dc current will flow through the oxide.

counter electrode which we used would short the films. Thus we spent some time measuring very interesting but always non-reproducible current-voltage characteristics which we referred to as miracles since each occurred only once. After a few months we hit on the correct idea: to use evaporated metal films and to separate them by a naturally grown oxide layer.

To carry out our ideas we needed an evaporator, thus I purchased my first piece of experimental equipment. While waiting for the evaporator to arrive I worried a lot-I was afraid I would get stuck in experimental physics tied

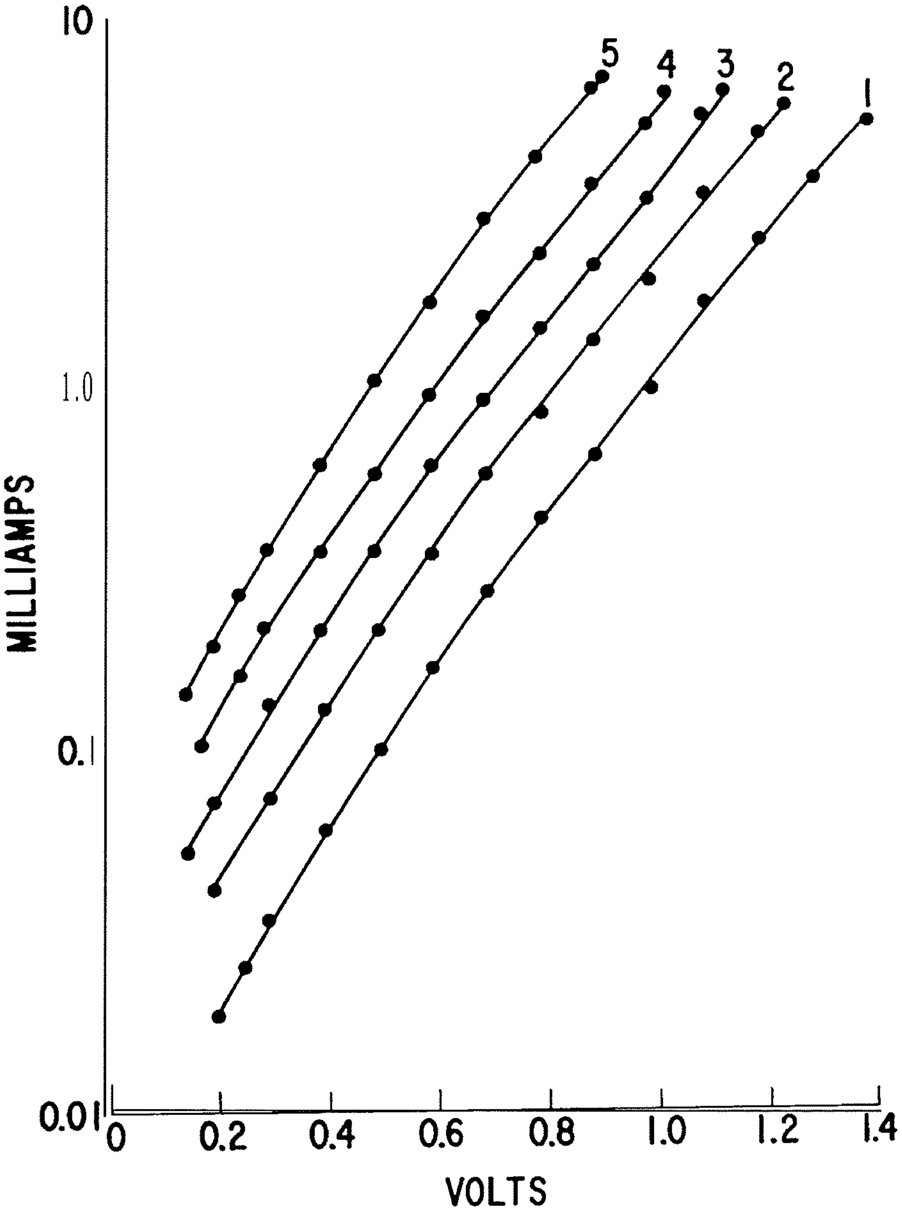


Fig. 4.
Current-voltage characteristics of five different tunnel junctions all with the same thickness, but with five different areas. The current is proportional to the area of the junction. This was one of the first clues that we were dealing with tunneling rather than shorts. In the early experiments we used a relatively thick oxide, thus very little current would flow at low voltages.

down to this expensive machine. My plans at the time were to switch into theory as soon as I had acquired enough knowledge. The premonition was correct; I did get stuck with the evaporator, not because it was expensive but because it fascinated me. Figure 2 shows a schematic diagram of an evaporator. To prepare a tunnel junction we first evaporated a strip of aluminum onto a glass slide. This film was removed from the vacuum system and heated

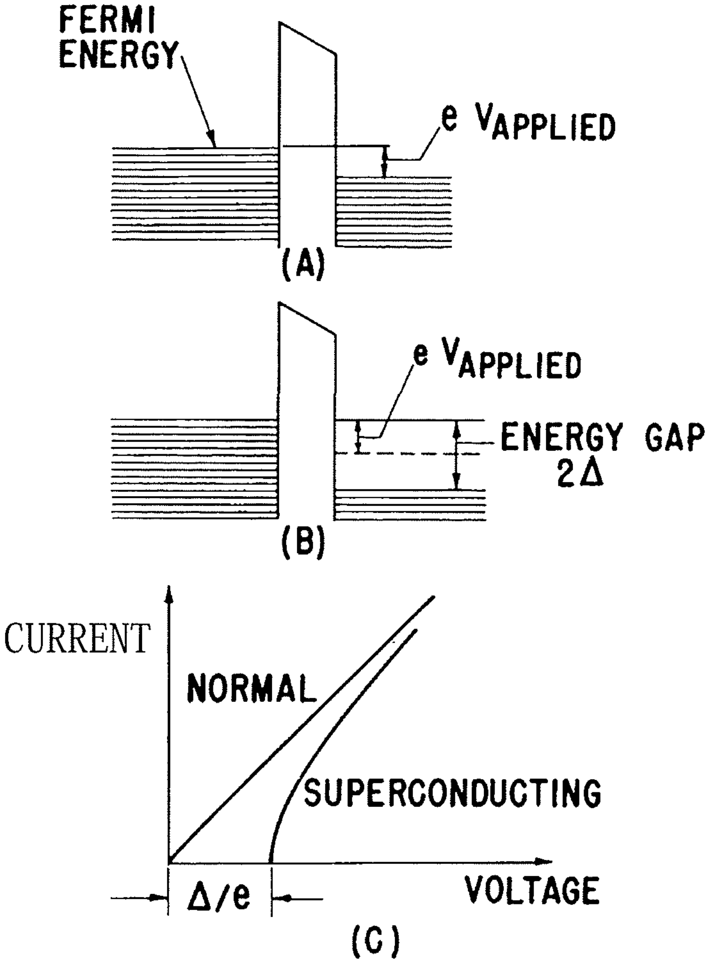


Fig. 5.
A. An energy diagram of two metals separated by a barrier. The Fermi energies in the two metals are at different levels because of the voltage difference applied between the metals. Only the left metal electrons in the energy range $e \cdot V_{\text{App}}$ can make a transition to the metal on the right, because only these electrons face empty energy states. The Pauli Principle allows only one electron in each quantum state. B. The right-hand metal is now superconducting, and an energy gap 2Δ has opened up in the electron spectrum. No single electron in a superconductor can have an energy such that it will appear inside the gap. The electrons from the metal on the left can still tunnel through the barrier, but they cannot enter into the metal on the right as long as the applied voltage is less than Δ/e, because the electrons either face a filled state or a forbidden energy range. When the applied voltage exceeds Δ/e, current will begin to flow. C. A schematic current-voltage characteristic. When both metals are in the normal state the current is simply proportional to the voltage. When one metal is superconducting the current-voltage characteristic is drastically altered. The exact shape of the curve depends on the electronic energy spectrum in the superconductor.

to oxidize the surface rapidly. Several cross strips of aluminum were then deposited over the first film making several junctions at the same time. The steps in the sample preparation are illustrated in Figure 3. This procedure solved two problems, first there were no pinholes in the oxide because it is

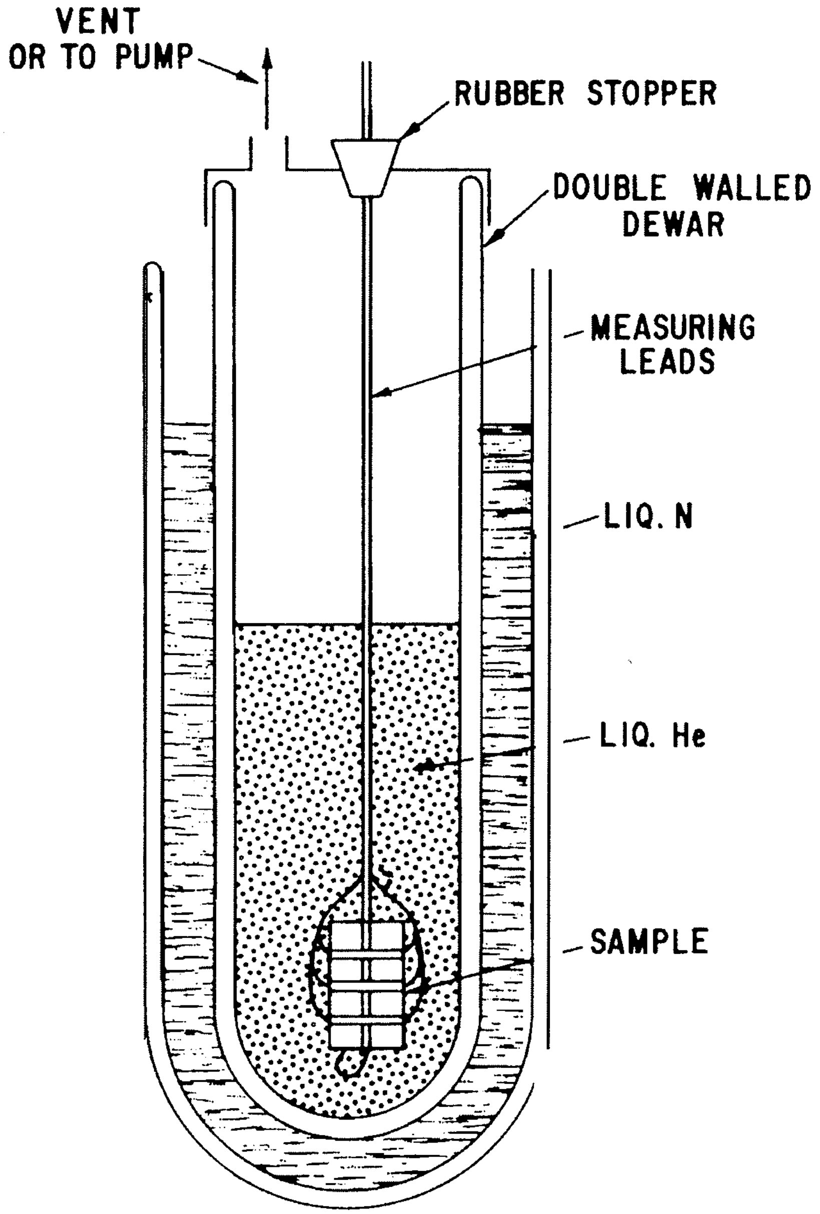


Fig. 6.
A standard experimental arrangement used for low temperature experiments. It consists of two dewars, the outer one contains liquid nitrogen, the inner one, liquid helium. Helium boils at 4.2" K at atmospheric pressure. The temperature can be lowered to about l°K by reducing the pressure. The sample simply hangs into the liquid helium supported by the measuring leads.

self-healing, and second we got rid of mechanical problems that arose with the mercury counter electrode.

By about April, 1959, we had performed several successful tunneling experiments. The current-voltage characteristics of our samples were reasonably reproducible, and conformed well to theory. A typical result is shown in Figure 4. Several checks were done, such as varying the area and the oxide thickness of the junction as well as changing the temperature. Everything looked OK, and I even gave a seminar at the Laboratory. By this time, I had solved Schrodinger's equation enough times to believe that electrons sometimes behave as waves, and I did not worry much about that part anymore.

However: there were many real physicists at the Laboratory and they properly questioned my experiment. How did I know I did not have metallic shorts? Ionic current? Semiconduction rather than tunneling3 Of course, I did not know, and even though theory and experiments agreed well, doubts about the validity were always in my mind. I spent a lot of time inventing impossible schemes such as a tunnel triode or a cold cathode, both to try to prove conclusively that I dealt with tunneling and to perhaps make my work useful. It was rather strange for me at that time to get paid for doing what I considered having fun, and my conscience bothered me. But just like quantum mechanics, you get used to it, and now I often argue the opposite point; we should pay more people to do pure research.

I continued to try out my ideas on John Fisher who was now looking into the problems of fundamental particles with his characteristic optimism and enthusiasm; in addition, I received more and more advice and guidance from Charles Bean and Walter Harrison, both physicists with the uncanny ability of making things clear as long as a piece of chalk and a blackboard were available. I continued to take formal courses at RPI, and one day in a solid state physics course taught by Professor Huntington we got to superconductivity. Well, I didn't believe that the resistance drops to exactly zero-but what really caught my attention was the mention of the energy gap in a superconductor, central to the new Bardeen-Cooper-Schrieffer theory. If the theory was any good and if my tunneling experiments were any good, it was obvious to me that by combining the two, some pretty interesting things should happen, as illustrated in Figure 5. When I got back to the GE Laboratory I tried this simple idea out on my friends, and as I remember, it did not look as good to them. The energy gap was really a many body effect and could not be interpreted literally the way I had done. But even though there was considerable skepticism, everyone urged me to go ahead and make a try. Then I realized that I did not know what the size of the gap was in units I understood-electron volts. This was easily solved by my usual method: first asking Bean and then Harrison, and, when they agreed on a few millielectron volts: I was happy because that is in a easily measured voltage range.

I had never done an experiment requiring low temperatures and liquid helium-that seemed like complicated business. However one great advantage of being associated with a large laboratory like General Electric is that there are always people around who are knowledgeable in almost any field, and better still they are willing to lend you a hand. In my case, all I had to do was go to the end of the hall where Warren DeSorbo was already doing experiments with superconductors. I no longer remember how long it took me to set up the helium dewars I borrowed, but probably no longer than a day or two. People unfamiliar with low temperature work believe that the whole field of low temperature is pretty esoteric, but all it really requires is access

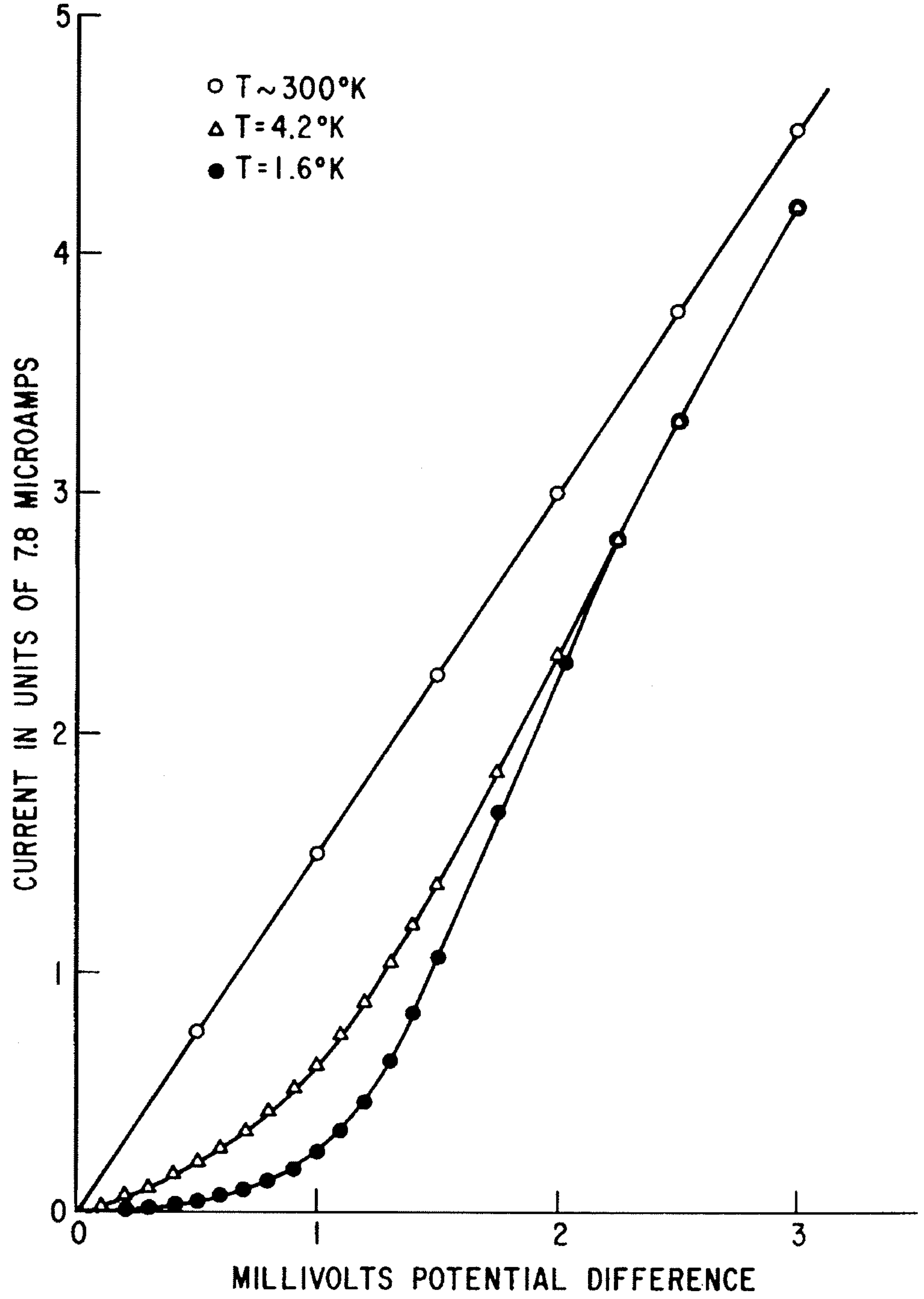


Fig. 7.
The current-voltage characteristic of an aluminum-aluminum oxide-lead sample. As soon as the lead becomes superconducting the current ceases to be proportional to the voltage. The large change between 4.2" K and 1.6" K is due to the change in the energy gap with temperature. Some current also flows at voltages less than Δ / e because of thermally excited electrons in the conductors.

to liquid helium, which was readily available at the Laboratory. The experimental setup is shown in Figure 6. Then I made my samples using the familiar aluminum-aluminum oxide, but I put lead strips on top. Both lead and alu-

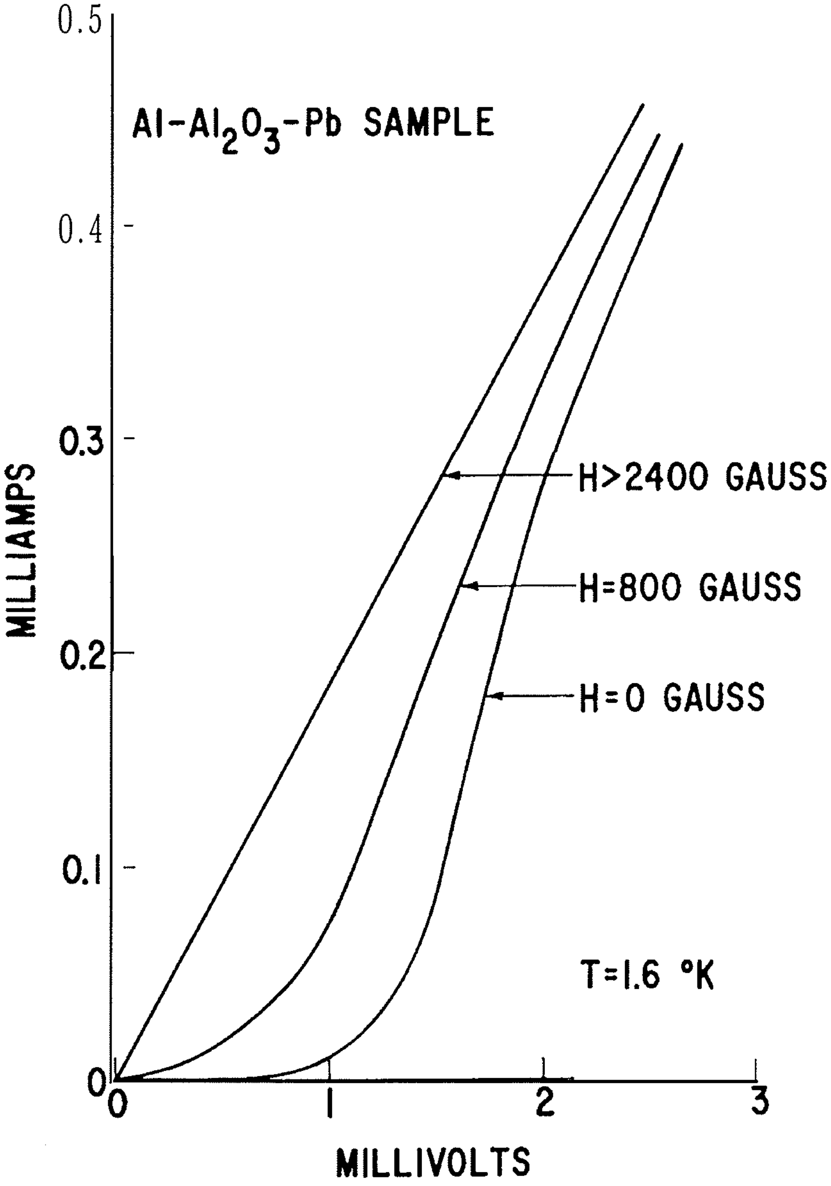


Fig. 8.
The current-voltage characteristic at 1.6" K as a function of the applied magnetic field. At 2 400 gauss the films are normal, at 0 gauss the lead film is superconducting. The reason for the change in the characteristics between 800 gauss and 0 gauss is that thin films have an energy gap that is a function of the magnetic field.

minum are superconductors, lead is superconducting at 7.2° K and thus all you need to make it superconducting is liquid helium which boils at 4.2° K. Aluminum becomes superconducting only below 1.2° K, and to reach this temperature a more complicated experimental setup is required.

The first two experiments I tried were failures because I used oxide layers which were too thick. I did not get enough current through the thick oxide to measure it reliably with the instruments I used, which were simply a standard voltmeter and a standard ammeter. It is strange to think about that

Fig. 9.
Informal discussion over a cup of coffee. From left: Ivar Giaever, Walter Harrison, Charles Bean, and John Fisher.

now, only 13 years later, when the Laboratory is full of sophisticated x-y recorders. Of course, we had plenty of oscilloscopes at that time but I was not very familiar with their use. In the third attempt rather than deliberately oxidizing the first aluminum strip, I simply exposed it to air for only a few minutes, and put it back in the evaporator to deposit the cross strips of lead. This way the oxide was no more than about 30Å thick, and I could readily measure the current-voltage characteristic with the available equipment. To me the greatest moment in an experiment is always just before I learn whether the particular idea is a good or a bad one. Thus even a failure is exciting, and most of my ideas have of course been wrong. But this time it worked! The current-voltage characteristic changed markedly when the lead changed from the normal state to the superconducting state as shown in Figure 7. That was exciting! I immediately repeated the experiment using a different sample - everything looked good! But how to make certain? It was well-known that superconductivity is destroyed by a magnetic field, but my simple setup of dewars made that experiment impossible. This time I had to go all the way across the hall where Israel Jacobs studied magnetism at low temperatures. Again I was lucky enough to go right into an experimental rig where both the temperature and the magnetic field could be controlled and I could quickly do all the proper experiments. The basic result is shown in Figure 8. Every-

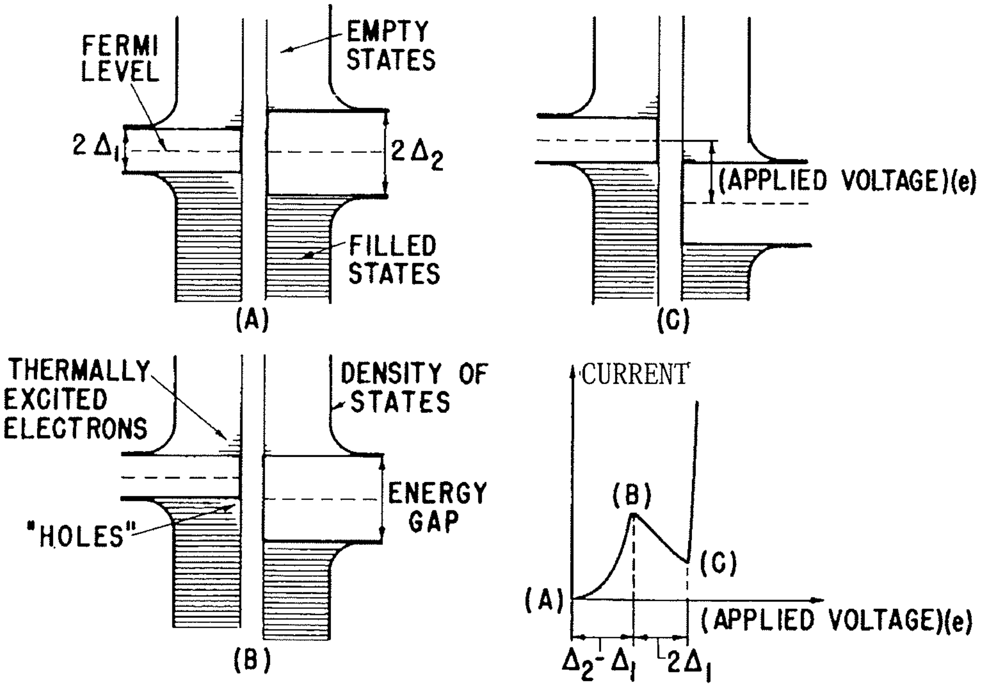


Fig. 10.
Tunneling between two superconductors with different energy gaps at a temperature larger than 0° K. A. No voltage is applied between the two conductors. B. As a voltage is applied it becomes energetically possible for more and more of the thermally excited electrons to flow from the superconductor with the smaller gap into the superconductor with the larger gap. At the voltage shown all the excited electrons can find empty states on the right. C. As the voltage is further increased, no more electrons come into play, and since the number of states the electrons can tunnel into decreases, the current will decrease as the voltage is increased. When the voltage is increased sufficiently the electrons below the gap in the superconductor on the left face empty states on the right, and a rapid increase in current will occur. D. A schematic picture of the expected current-voltage characteristic.

thing held together and the whole group, as I remember it, was very excited. In particular, I can remember Bean enthusiastically spreading the news up and down the halls in our Laboratory, and also patiently explaining to me the significance of the experiment.

I was, of course, not the first person to measure the energy gap in a superconductor, and I soon became aware of the nice experiments done by M. Tinkham and his students using infrared transmission. I can remember that I was worried that the size of the gap that I measured did not quite agree with those previous measurements. Bean set me straight with words to the effect that from then on other people would have to agree with me; my experiment would set the standard, and I felt pleased and like a physicist for the first time.

That was a very exciting time in my life; we had several great ideas to improve and extend the experiment to all sorts of materials like normal metals, magnetic materials and semiconductors. I remember many informal dis-

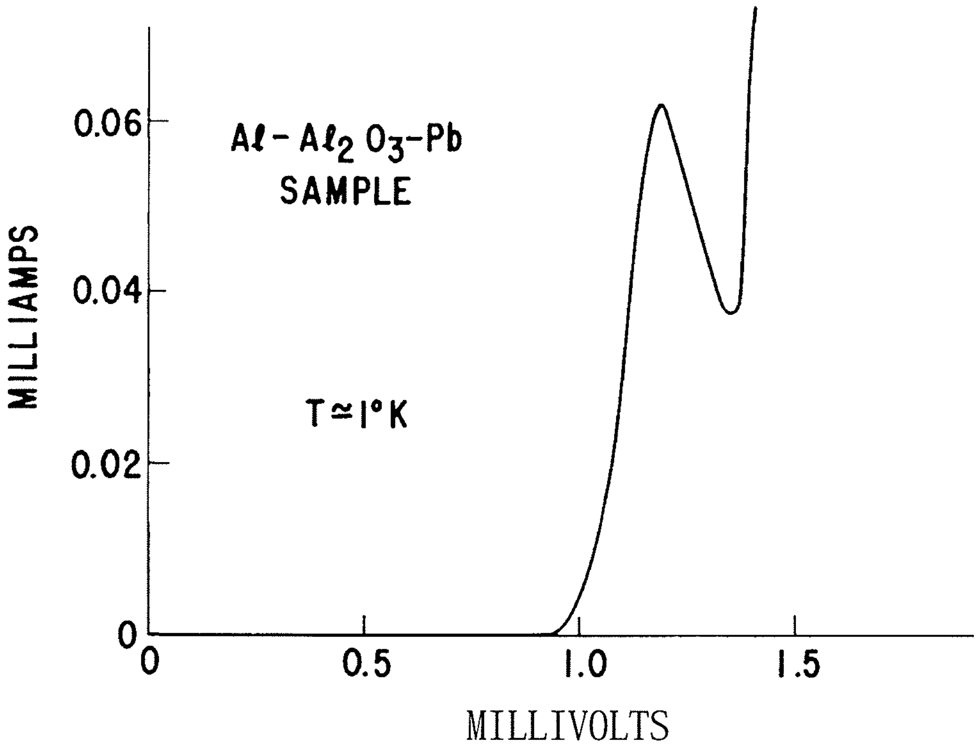


Fig. 11.
A negative resistance characteristic obtained experimentally in tunneling between two different superconductors.

cussions over coffee about what to try next and one of these sessions is in a photograph taken in 1960 which is shown in Figure 9. To be honest the picture was staged, we weren't normally so dressed up, and rarely did I find myself in charge at the blackboard! Most of the ideas we had did not work very well and Harrison soon published a theory showing that life is really complicated after all. But the superconducting experiment was charmed and always worked. It looked like the tunneling probability was directly proportional to the density of states in a superconductor. Now if this were strictly true, it did not take much imagination to realize that tunneling between two superconductors should display a negative resistance characteristic as illustrated in Figure 10. A negative resistance characteristic meant, of course, amplifiers, oscillators and other devices. But nobody around me had facilities to pump on the helium sufficiently to make aluminum become superconducting. This time I had to leave the building and reactivate an old low temperature setup in an adjacent building. Sure enough, as soon as the aluminum went superconducting a negative resistance appeared, and, indeed, the notion that the tunneling probability was directly proportional to the density of states was experimentally correct. A typical characteristic is shown in Figure 11.

Now things looked very good because all sorts of electronic devices could be made using this effect, but, of course, they would only be operative at low temperatures. We should remember that the semiconducting devices were not so advanced in 1960 and we thought that the superconducting junction

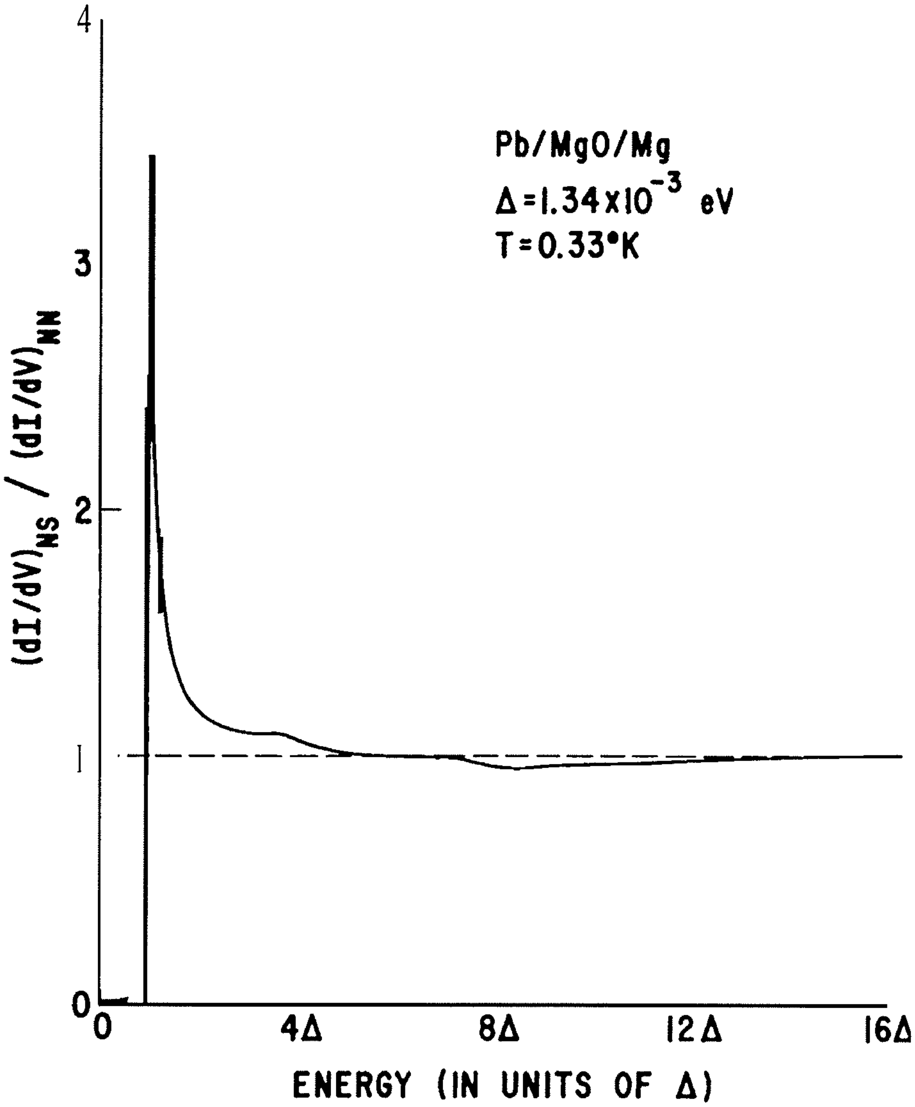


Fig. 12.
A normalized derivative of the current with respect to voltage of a lead junction at low temperature. The simple BCS-theory predicts that the derivative should approach unity asymptotically as the energy increases. Instead several wiggles are observed in the range between 4Δ and 8Δ. These wiggles are related to the phonon spectrum in lead.

would have a good chance of competing with, for example, the Esaki diode. The basic question I faced was which way to go: engineering or science? I decided that I should do the science first, and received full support from my immediate manager, Roland Schmitt.

In retrospect I realize how tempting it must have been for Schmitt to encourage other people to work in the new area, and for the much more experienced physicists around me to do so as well. Instead, at the right time, Schmitt provided me with a co-worker, Karl Megerle, who joined our Laboratory as a Research Training Fellow. Megerle and I worked well together and before long we published a paper dealing with most of the basic effects.

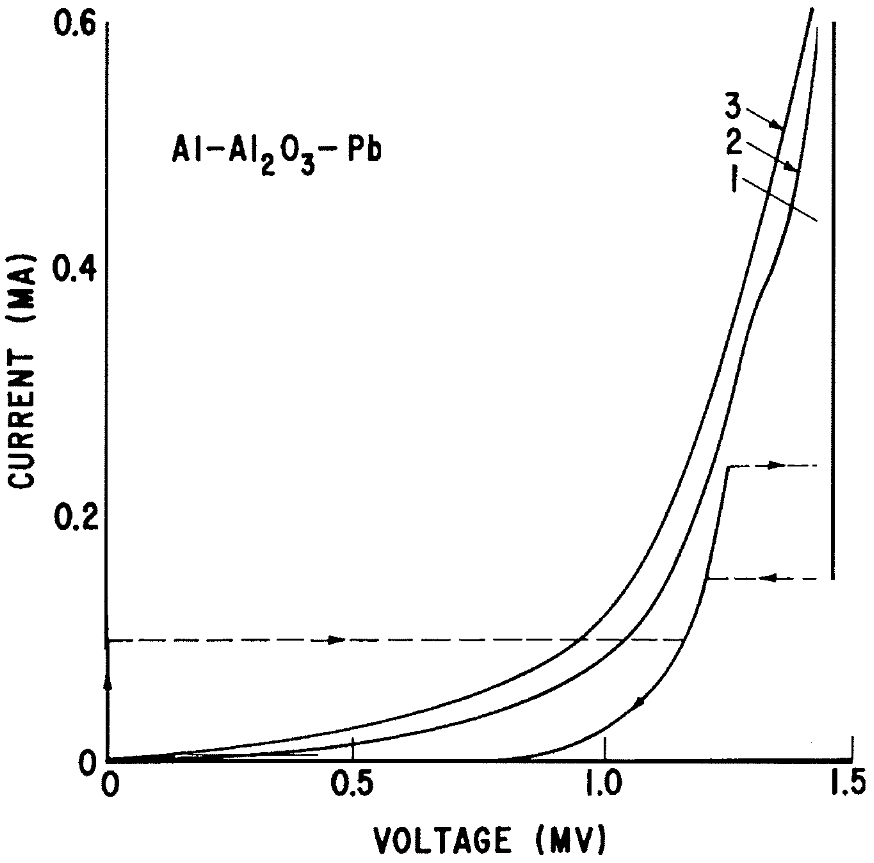


Fig. 13.
Effect of trapped magnetic field on a tunneling characteristic. Curve 1 is a virgin curve, while curve 3 is in a moderate magnetic field, and in curve 2 the magnetic field has been removed. In curve 1 we also have a small resistance-less current which we interpreted as caused by metallic shorts. In retrospect, it was actually due to the Josephson effect.

As always in physics, it is important to extend experiments to a higher energy, a greater magnetic field, or, in our case, to a lower temperature. Therefore, we joined forces with Howard Hart, who had just completed a helium 3 refrigerator that was capable of getting down to about 0.3° K. At the same time, Megerle finished a lock-in amplifier which we could use to measure directly the derivative of the current with respect to the voltage. That was really a nice looking machine with a magnet rotating past a pickup coil at eight cycles per second, but, of course, vastly inferior to the modern lock-in amplifier. We had known for some time that there were anomalies in the current-voltage characteristics of lead, and now we finally pinned them down by finding some extra wiggles in the derivative curve. This is shown in Figure 12. That made us happy because all that the tunneling experiments had done up till now was to confirm the BCS theory, and that is not what an experimentalist would really like to do. The dream is to show that a famous theory is incorrect, and now we had finally poked a hole in the theory. We speculated at the time that these wiggles were somehow associated with the phonons thought to be the cause of the attractive electron-electron interaction in a superconductor. As often happens, the theorists turned the tables on us and cleverly used these wiggles to properly extend the theory and to prove that the BCS theory indeed was correct. Professor Bardeen gave a detailed account of this in his most recent Nobel Prize lecture.

I have, so far, talked mainly about what went on at General Electric at that time; sometimes it is difficult for me to realize that Schenectady is not the center of the world. Several other people began to do tunneling work, and to mention just a few: J. M. Rowell and W. L. McMillan were really the ones who unraveled the phonon structure in a superconductor; W. J. To - masch, of course, insisted on discovering his own effect; S. Shapiro and colleagues did tunneling between two superconductors at the same time we did; and J. Bardeen, and later M. H. Cohen et al., took care of most of the theory.

Meanwhile, back at RPI, I had finished my course work and decided to do a theoretical thesis on ordered-disordered alloys with Professor Huntington because tunneling in superconductors was mainly understood. Then someone made me aware of a short paper by Brian Josephson in *Physics Letters* - what did I think? Well, I did not understand the paper, but shortly after I had the chance to meet Josephson at Cambridge and I came away impressed. One of the effects Josephson predicted was that it should be possible to pass a supercurrent with zero voltage drop through the oxide barrier when the metals on both sides were superconducting; this is now called the dc Josephson effect. We had observed this behavior many times; matter-of-fact, it is difficult not to see this current when junctions are made of tin-tin oxide-tin or lead-lead oxide-lead. The early tunnel junctions were usually made with aluminum oxide which generally is thicker and therefore thermal fluctuations suppress the dc current. In our first paper Megerle and I published a curve, which is shown in Figure 13, demonstrating such a supercurrent and also that it depended strongly on a magnetic field. However, I had a ready-made explanation for this supercurrent-it came from a metallic short or bridge. I was puzzled at the time because of the sensitivity to the magnetic field which is unexpected for a small bridge, but no one knew how a 20Å long and 20Å, wide bridge would behave anyway. If I have learned anything as a scientist it is that one should not make things complicated when a simple explanation will do. Thus all the samples we made showing the Josephson effect were discarded as having shorts. This time I was too simple-minded! Later I have been asked many times if I feel bad for missing the effect? The answer is clearly no, because to make an experimental discovery it is not enough to observe something, one must also realize the significance of the observation, and in this instance I was not even close. Even after I learned about the dc Josephson effect, I felt that it could not be distinguished from real shorts, therefore I erroneously believed that only the observation of the so-called ac effect would prove or disprove Josephson's theory.

In conclusion I hope that this rather personal account may provide some slight insight into the nature of scientific discovery. My own beliefs are that the road to a scientific discovery is seldom direct, and that it does not necessarily require great expertise. In fact, I am convinced that often a newcomer to a field has a great advantage because he is ignorant and does not know all the complicated reasons why a particular experiment should not be attempted. However, it is essential to be able to get advice and help from experts in the various sciences when you need it. For me the most important ingredients were that I was at the right place at the right time and that I found so many friends both inside and outside General Electric who unselfishly supported me.

References

1. *Tunneling Phenomena in Solids* edited by Burstein, E. and Lundquist, S. Plenum Press, New York, 1969.
2. Superconductivity edited by Parks, R. D. Marcel Dekker, Inc., New York. 1969.

Study of Superconductors by Electron Tunneling

IVAR GIAEVER AND KARL MEGERLE
General Electric Research Laboratory, Schenectady, New York

(Received January 3, 1961)

If a small potential difference is applied between two metals separated by a thin insulating film, a current will flow due to the quantum mechanical tunnel effect. For both metals in the normal state the current-voltage characteristic is linear, for one of the metals in the superconducting state the current voltage characteristic becomes nonlinear, and for both metals in the superconductive state even a negative-resistance region is obtained. From these changes in the current voltage characteristics, the change in the electron density of states when a metal goes from its normal to its superconductive state can be inferred. By using this technique we have found the energy gap in metal films 1000–3000 A thick at 1°K to be $2\epsilon_{\mathrm{Pb}}=(2.68\pm0.06)\times10^{-3}$ ev, $2\epsilon_{\mathrm{Sn}}=(1.11\pm0.03)\times10^{-3}$ ev, $2\epsilon_{\mathrm{In}}=(1.05\pm0.03)\times10^{-3}$ ev, and $2\epsilon_{\mathrm{Al}}=(0.32\pm0.03)\times10^{-3}$ ev.

The variation of the gap width with temperature is found to agree closely with the Bardeen-Cooper-Schrieffer theory. Furthermore, the energy gap in these films has been found to depend upon the applied magnetic field, decreasing with increasing field.

INTRODUCTION

THE existence of an energy gap in superconductors is well documented experimentally, and is firmly grounded in the theory of superconductivity of Bardeen, Cooper, and Schrieffer.[1] Experimental evidence for the existence of a gap and, indirectly, its width, can be obtained from measurements of specific heat, thermal conductivity, nuclear relaxation, ultrasonic attenuation, and electromagnetic absorption.[2] In general, the width of the gap is inferred from the variation of one of the above parameters and, with the exception of electromagnetic absorption, represents only an indirect measurement.

This paper describes a method for investigating the energy gap and density of electron states in superconductors by means of electron tunneling through thin insulating films. It represents an entirely new approach to the problem and results in clear, unambiguous measurements of the energy gap. Some preliminary results, employing this method, have already been published.[3–6]

The samples used in this experiment consist of a thin insulating oxide layer sandwiched between two evaporated metal films. Experimentally, the electron tunneling current through the insulating oxide layer is observed as a function of the voltage applied between the two metal films. Because of their small physical size, the samples are well suited for standard low-temperature techniques.

If a *small* potential difference is applied to the two metals in their normal, nonsuperconducting state, the tunneling current through the insulating film will vary linearly with applied voltage, as long as the density of electron states in the two metals is constant over the applied voltage range.[7] On the other hand, if the density of electron states varies rapidly in this voltage range, as it does in superconductors, the current-voltage characteristics will be nonlinear. It appears that this nonlinearity is simply correlated with the variation in the density of electron states. In particular, no electrons can flow into the energy region of the gap in superconductors.

By this method we have measured the energy gap in lead, tin, indium, and aluminum. The variation of the energy gap as a function of temperature[6] and magnetic field has also been investigated.

APPARATUS

The apparatus which is shown in Fig. 1 consists basically of a liquid helium Dewar with provisions for

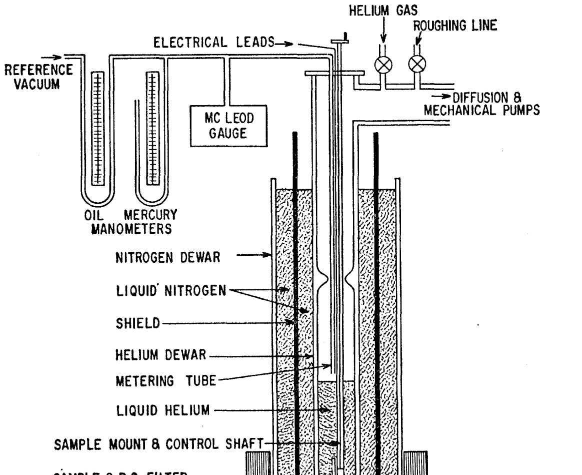


FIG. 1. A schematic drawing of the apparatus. The shield can be removed when studies are made using a magnetic field.

[1] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).

[2] M. A. Biondi, A. T. Forrester, M. P. Garfunkel, and C. B. Satterthwaite, Revs. Modern Phys. **30**, 1109 (1958).

[3] I. Giaever, Phys. Rev. Letters **5**, 147 (1960).

[4] I. Giaever, Phys. Rev. Letters **5**, 464 (1960).

[5] J. Nicol, S. Shapiro, and P. H. Smith, Phys. Rev. Letters **5**, 461 (1960).

[6] I. Giaever, Proceedings of the Seventh International Conference on Low-Temperature Physics, Toronto, 1960 (to be published).

[7] J. C. Fisher and I. Giaever, J. Appl. Phys. **32**, 172 (1961).

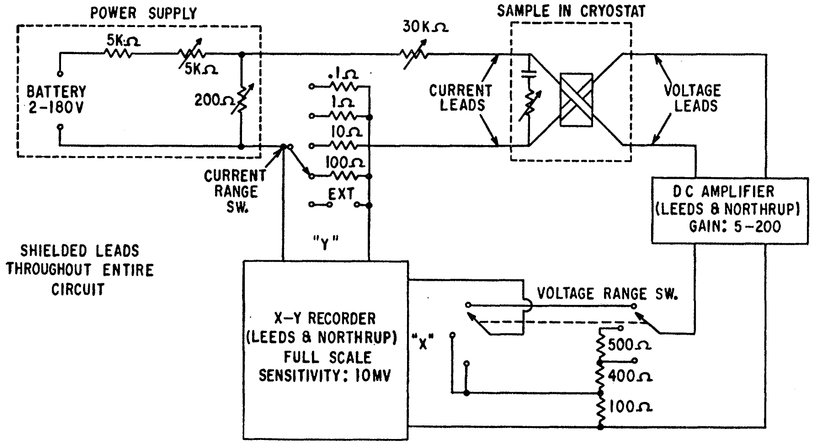


FIG. 2. Circuit diagram of the measuring circuit.

pumping on the helium, and an outer Dewar containing liquid nitrogen which acts as a radiation shield for the helium. The helium Dewar has a constriction in its diameter to minimize creep losses of the superfluid helium when the temperature is below the λ point.

Temperatures are measured by means of the helium vapor pressure. A metering tube which extends to within a few inches of the liquid helium level is connected to both oil and mercury manometers and to a McLeod gauge. The system is capable of attaining a temperature of about 0.9°K. Due to the low heat leakage, this temperature can be maintained for approximately six hours, which is adequate for making numerous measurements.

The electrical circuitry is shown in Fig. 2. To trace out the current-voltage characteristics, a Leeds and Northrup X-Y recorder and matching Leeds and Northrup dc amplifier are used in conjunction with external shunts and multipliers to extend the range of the instruments. These instruments contain chopper-stabilized amplifiers and, therefore, have virtually no drift. To accommodate various sample resistances and to obtain the necessary detail in the current-voltage curves, the current scale can be decade switched over a full-scale sensitivity from 100 ma to 1 μa and the voltage scale from 100 mv to 50 μv.

The emf source can be used as either a high- or low-impedance source, by suitable adjustment of the two variable resistors and the applied battery voltage. The high capacitance of the sample in conjunction with considerable lead inductance gives rise to very troublesome high-frequency oscillations whenever the sample is biased into its negative-resistance region. By placing an adjustable high-pass filter in parallel with the sample, the high-frequency oscillations can be eliminated or at least greatly reduced. The high-pass filter consists of a capacitor large in comparison to the sample capacitance and a variable resistor in series, and is in close proximity to the sample to minimize lead inductance. The oscillations are reduced by matching the variable resistance to the negative resistance of the sample. The variable resistor is a Bourns Trimpot which is mounted on the end of a $\frac{1}{4}\times0.008$ in. stainless steel tube. Concentric with this tube is a $\frac{1}{8}\times0.008$ in. stainless steel tube which engages the adjustment screw of the resistor and passes through an O-ring seal in the Dewar cover plate, to permit external adjustments.

The electrical connections into the Dewar, consisting of current and voltage leads, are brought out through the cover plate and are sealed in place with Apiezon wax to achieve a tight seal. In order to minimize heat leaks, four 3-mil Formex-covered copper wires are used inside the cryostat. To minimize induced noise, the entire electrical circuitry outside the Dewar is shielded. The sample and leads within the Dewar can be shielded by a copper-clad soft iron shield which sits in the liquid nitrogen, surrounding the helium Dewar. This shield is not used for measurements made with an externally applied magnetic field.

Since the voltages applied to the sample are very small, induced voltages caused by ever-present fluctuating stray fields remain a difficult problem even after careful shielding. The slow response time of the recording apparatus causes the readings to be averaged over the fluctuations resulting from induced voltages. Due to the nonlinear characteristics of our samples, this averaging tends to smooth out the current-voltage curves and results in lost detail. The difficulty is virtually eliminated by including a resistance in series with the current loop, and making this resistance as large as practical. The large resistance in series with the sample resistance acts as a voltage divider for induced noise, so that only a small amount of noise appears across the sample. This large series resistance effectively increases the emf source impedance and cannot be used when investigating the negative-resistance region of the sample. It has, however, been retained for measurements outside the negative resistance region on most samples.

The sample is mounted directly on the variable resistor which is a part of the high-pass filter. This insures mechanical rigidity and reproducible, accurate positioning for measurements involving magnetic fields.

When the sample is subjected to a magnetic field, it is aligned so that the field vector is in the plane of both films, parallel to the long dimension of the aluminum film, and normal to the long dimension of the other metal film.

SAMPLE PREPARATION

The sample consists of two metal films separated by a thin insulating layer. Aluminum/aluminum oxide/metal sandwiches are prepared by vapor-depositing aluminum on microscope glass slides in vacuum, oxidizing the aluminum, and then vapor-depositing a metal over the aluminum oxide. First the microscope slide is cut to size, $\frac{1}{2}\times 3$ in., so as to fit through the constriction in the helium Dewar. Next, indium is smeared onto the four corners of the glass slide to provide contacts between the evaporated metal strips and external leads. The glass slide with indium contacts is then washed with Alconox detergent, rinsed with distilled water and ethanol, and dried with dry nitrogen gas.

Next, the glass slide is mounted in the evaporator so that it can be positioned behind suitable masks in vacuum. The evaporations are made from tantalum strips approximately $\frac{3}{8}\times 1\frac{1}{2}\times 0.005$ in., which have previously been charged and heated in vacuum so that the charge wets the tantalum strip. The evaporations are made at a starting pressure of 5×10^{-5} mm Hg, or less.

Preparation of the metal/insulator/metal sandwich proceeds in three distinct steps, as shown in Fig. 3, during which the substrate is at room temperature.

First, a layer of aluminum is evaporated onto the glass slide between two contacts. This strip is 1 mm wide and 1000–3000 A thick. Next, the aluminum is oxidized either at atmospheric pressure or some reduced pressure. Finally, a layer of Al, Pb, In, or Sn, of dimensions similar to the aluminum strip, is evaporated over the aluminum oxide layer between the remaining two contacts.

The thickness of the Al_2O_3 insulating layer between the metal strips is subject to a number of variables. The pressure and time dependence of oxidation rate has been extensively investigated and is well documented in the literature.[8] Atmospheric humidity or residual H_2O vapor in the vacuum system also affects the oxidation rate. We have found that an increase in the amount of water effects a more rapid oxide growth rate. The oxide thickness is also contingent upon the evaporation rate and evaporation temperature of the metal layer deposited over the oxide layer. Presumably, metals which require higher temperatures for evaporation must give up more energy to the oxide film; i.e., the atoms penetrate further into the oxide, thereby effectively reducing the thickness of the oxide layer. Another parameter is oxidation temperature, elevated temperatures promoting an increase in the oxidation rate. A final variable is introduced by the evaporation rate of the aluminum layer, which influences its surface characteristics. We have noted that the oxide grows more slowly and reaches a thinner limiting value on films which were evaporated with a high deposition rate (approximately 1000 A/sec).

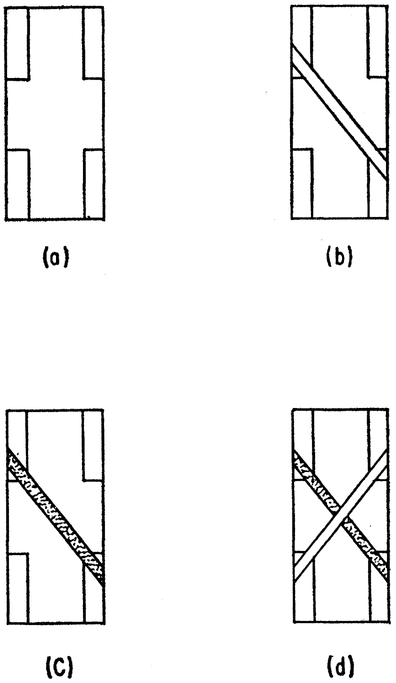


FIG. 3. Sample preparation. (a) Glass slide with indium contacts. (b) An aluminum strip has been deposited across the contacts. (c) The aluminum strip has been oxidized. (d) A lead film has been deposited across the aluminum film, forming an $Al\text{-}Al_2O_3\text{-}Pb$ sandwich.

TABLE I. Approximate relationship between oxidation time and film resistance for $Al\text{-}Al_2O_3\text{-}Pb$ sandwiches.

Time	Temperature	Pressure	Resistance (ohm/mm²)
24 hr	100°C	atmospheric	10^5-10^7
24 hr	room	atmospheric	10^3-10^5
10 min	room	atmospheric	$10-10^3$
2 min	room	atmospheric	$1-10^2$
10 min	room	200μ Hg	10^1-10^{-1}
10 min	room	50μ Hg	$10^{-2}-10^{-1}$

By controlling these parameters to some extent, the resistance of a 1-mm² junction can be made to vary between 10^{-2} and 10^7 ohm. Typical oxidation conditions for an $Al\text{-}Al_2O_3\text{-}Pb$ sandwich are given in Table I. (The resistance variations are probably due to the effect of humidity and the surface characteristics of the aluminum layer.)

It is possible to measure indirectly the thickness of the oxide layer by measuring the capacitance of the junction and then calculating the thickness.[7]

MODEL

The concept that particles can penetrate energy barriers is as old as quantum mechanics. In nuclear

[8] D. D. Eley and P. R. Wilkinson, *Structure and Properties of Thin Films* (John Wiley & Sons, Inc., New York, 1959), p. 508.

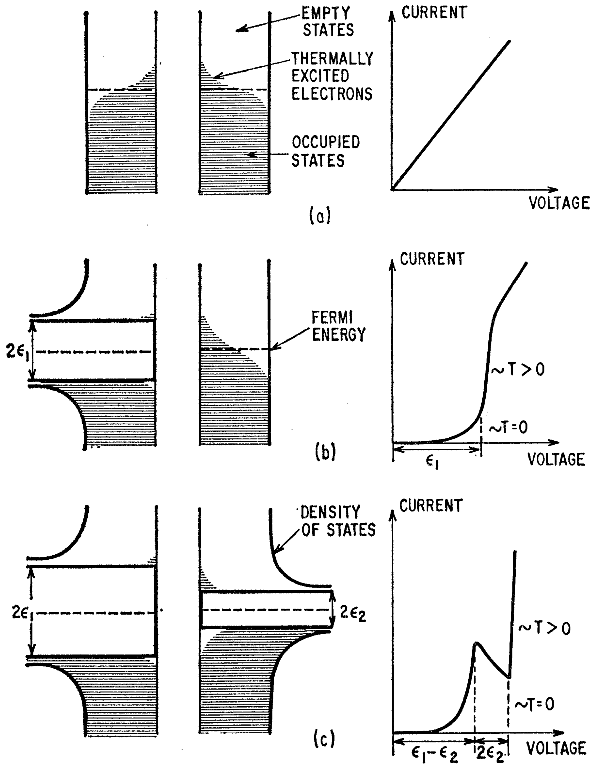


FIG. 4. Energy diagram displaying the density of states and the current-voltage characteristics for the three cases. (a) Both metals in the normal state. (b) One metal in the normal state and one in the superconducting state. (c) Both metals in the superconducting state.

physics, for example, the theory of α decay depends upon this tunnel effect. It has long been known that an electric current can flow between two metals separated by a thin insulating film because of the quantum-mechanical tunnel effect. Theoretical calculations were first made by Sommerfeld and Bethe[9] for a small potential difference applied between the two metals. These calculations were later extended by Holm.[10] An important result of these calculations is that for small voltages across the insulating film the tunnel current through the film is proportional to voltage. Holm *et al.*[11] furnished early experimental evidence for the tunneling effect, a work which was extended by Dietrich[12] and later by Fisher and Giaever.[7] The experiment of tunneling into superconductors[3] furnishes unquestionable evidence that this conduction mechanism is responsible for practically the whole current flow. In the following discussion we shall treat only the low-voltage region where the current flowing through the insulating film is proportional to the voltage across the film, provided both superconductors are in the normal state.

In Fig. 4(a) we show a simple model of two metals separated by a thin insulating film, the insulating film is pictured as a potential barrier. In Fig. 4(b) is shown the case when one of the two metals is in the superconducting state. Note how the electron density of states has changed, leaving an energy gap centered at the Fermi level as postulated by Bardeen, Cooper, and Schrieffer.[1] This particular model of a superconductor is a one-particle approximation, and it gives a surprisingly accurate picture of the experiments. In Fig. 4(c) both the metals are pictured in the superconducting state.

First we shall discuss qualitatively these three different cases, and later quantitatively calculate the current when both metals are in the normal state, and when only one of the metals is in the normal state.

The transmission coefficient of a quantum particle through a potential barrier depends exponentially upon the thickness of the barrier and upon the square root of the height of the barrier. For small voltages applied between the two metals neither the barrier thickness nor the barrier height is altered significantly. The current will then be proportional to the applied voltage, because the number of electrons which can flow increases proportionally to the voltage. The temperature effect will be very small, as the electron distribution is equal on either side of the barrier with metals in the normal state and in addition, kT is much smaller than the barrier height.

When one of the metals is in the superconducting state the situation is radically different. At absolute zero temperature, no current can flow until the applied voltage corresponds to half the energy gap. Assuming that the current is proportional to the density of states, the current will increase rapidly with voltage at first, and then will asymptotically approach the current-voltage characteristic found when both metals were in the normal state. At a temperature different from zero we will have a small current flow even at the lowest voltages. But since the two sides of the barrier now look different, the current will depend strongly upon temperature.

When both metals are in the superconducting state, the situation is again different. At absolute zero no current can flow until the applied voltage corresponds to half the sum of the two energy gaps. At a finite temperature, a current again will flow at the smallest applied voltages. The current will increase with voltage until a voltage equal to approximately half the difference of the two energy gaps is applied. When the voltage is increased further it is possible for only the same number of electrons to tunnel, but since the electrons will face a less favorable (lower) density of states, the current will actually decrease with increasing voltage. Finally, when a voltage equal to half the sum of the two gaps is applied the current will again increase rapidly with voltage and approach asymptotically the

[9] A. Sommerfeld and H. Bethe, *Handbuch der Physik*, edited by S. Flügge (Verlag Julius Springer, Berlin, 1933), Vol. 24, Part 2, p. 333.

[10] R. Holm, J. Appl. Phys. **22**, 569 (1951).

[11] R. Holm, *Electric Contacts* (Hugo Geber, Stockholm, Sweden, 1946).

[12] I. Dietrich, Z. Physik **132**, 231 (1952).

current-voltage characteristics obtained when both metals were normal.

Since we regard the distributions of holes and electrons in the metals in both the normal and superconducting state as symmetric about the Fermi level, no rectification effects are expected.

If we regard the tunneling through the insulating layer as an ordinary quantum-mechanical transition, the transition probability from an occupied state $\mathbf{k}$ on the left side of the barrier to a state $\mathbf{k}'$ on the right side can be written:

$$P_{\mathbf{k}\rightarrow\mathbf{k}'}=(2\pi/\hbar)|M|^2n'(1-f'), \qquad (4.1)$$

where n' is the density of states on the right side and f' the probability that the state $\mathbf{k}'$ is occupied. $|M|^2$ is the matrix element for the transition and we assume $|M|^2$ vanishes unless the components of $\mathbf{k}$ and $\mathbf{k}'$ transverse to the boundary are equal; that is, specular transmission and then n' is the density of states on the right for fixed wave number components parallel to the boundary. This is a convenient though not an essential assumption.

To calculate the current from left to right we sum over occupied states on the left and obtain

$$i=(4\pi e/\hbar)\textstyle\sum_{k_t}\sum_{k_x}|M|^2n'f(1-f'), \qquad (4.2)$$

where k_t is the component of wave number transverse to the barrier, k_x the component perpendicular to the barrier, e the electron charge, and f the probability that state $\mathbf{k}$ is occupied.

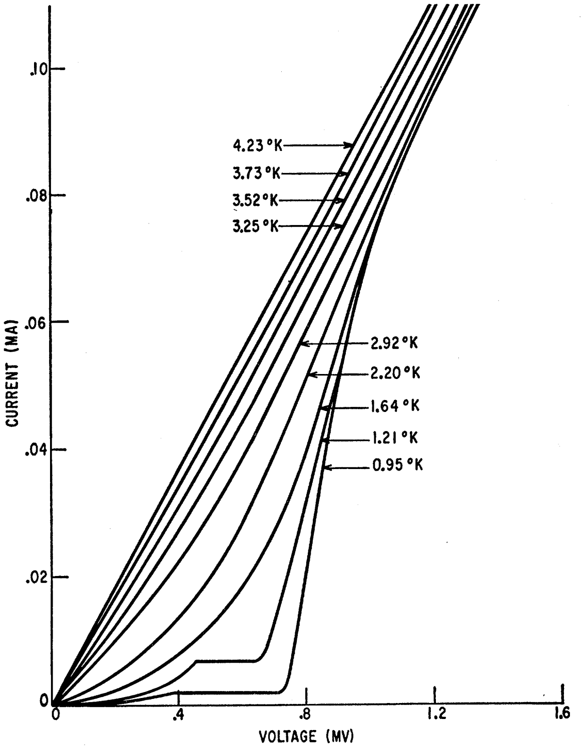


Fig. 6. Current-voltage characteristics of an Al-Al$_2$O$_3$-Sn sandwich at various temperatures.

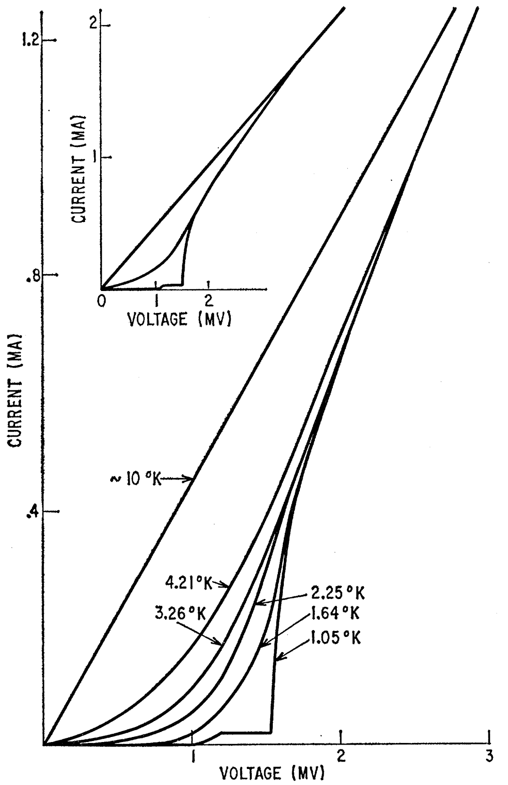


Fig. 5. Current-voltage characteristics of an Al-Al$_2$O$_3$-Pb sandwich at various temperatures.

By converting the sum over k_x to an integral over energy with fixed k_t we get

$$i=A\sum_{k_t}\int_{-\infty}^{\infty}|M|^2nn'f(1-f')dE, \qquad (4.3)$$

where A is a constant, n the density of states at the left side of the barrier (for fixed k_t), and E the energy measured from the Fermi energy.

By subtracting a similar expression for the current flowing from right to left, we get the net current flow:

$$I=A\sum_{k_t}\int_{-\infty}^{\infty}|M|^2n'n(f-f')dE. \qquad (4.4)$$

To fit the experimental results it is necessary to assume that $|M|^2\approx$constant. Bardeen,[13] using a many-particle point of view in connection with the WKB method, finds it plausible that $|M|^2$ is a constant over the energy values of interest. On assuming a constant $|M|^2$ and spherical symmetry of the dependence of energy on wave number, n' and n which are one dimensional densities of states are proportional to the total densities of states, and we may therefore sum over k_t directly and take $|M|^2$ out of the integral. We are

[13] J. Bardeen, Phys. Rev. Letters **6**, 57 (1961).

left with

$$I=A'\int_{-\infty}^{\infty} n'n\{f(E)-f(E+eV)\}dE, \qquad (4.5)$$

where A' is a constant and eV is the difference between the two Fermi levels. (V is the applied voltage.)

For the current between two normal metals we obtain at absolute zero and for small applied voltages:

$$I_{NN}=A'n'(E_F)n(E_F)eV, \qquad (4.6)$$

i.e., the current is proportional to voltage.

For a superconductor we may take the density of states from the Bardeen-Cooper-Schrieffer theory:

$$n_s=n\frac{E}{(E^2-\epsilon^2)^{\frac{1}{2}}}, \qquad (4.7)$$

where E is measured from the Fermi energy, and ϵ is half the energy gap. Thus the current between one metal in the normal state and one metal in the superconducting state can be written:

$$I_{NS}=A'n'(E_F)n(E_F)\int_{-\infty}^{\infty}\frac{|E|}{(E^2-\epsilon^2)^{\frac{1}{2}}} \times[f(E)-f(E+eV)]dE. \qquad (4.8)$$

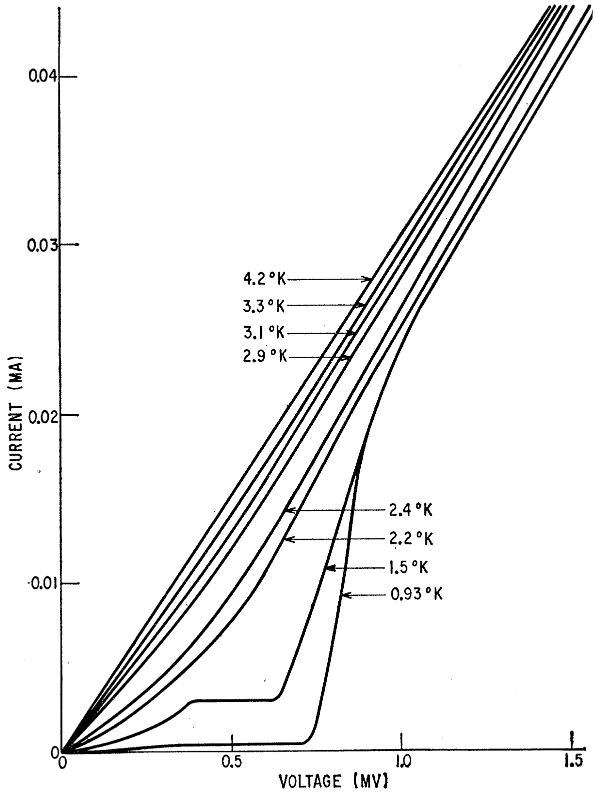


FIG. 7. Current-voltage characteristics of an Al-Al$_2$O$_3$-In sandwich at various temperatures.

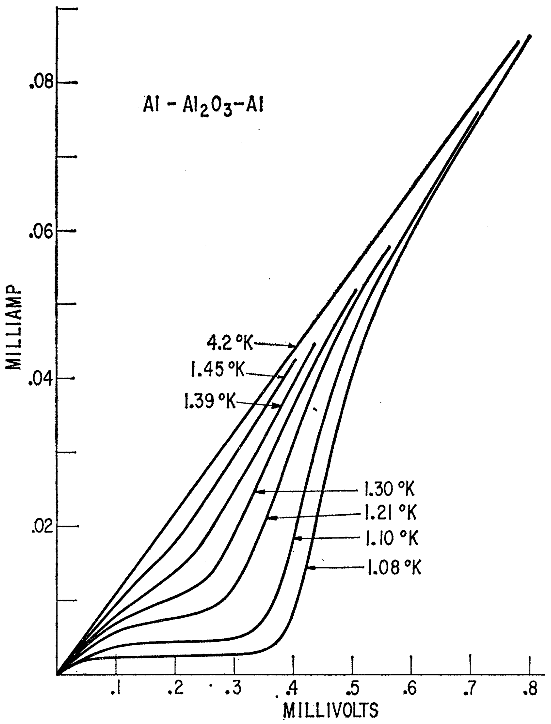


FIG. 8. Current-voltage characteristics of an Al-Al$_2$O$_3$-Al sandwich at various temperatures.

For small applied voltages such that $eV<\epsilon$ we may evaluate the above integral, as shown in Appendix I, and obtain:

$$I_{NS}=2C_{NN}\frac{\epsilon}{e}\sum_{m=1}^{\infty}(-1)^{m+1}K_1\left(m\frac{\epsilon}{kT}\right)\sinh\left(m\frac{eV}{kT}\right), \qquad (4.9)$$

where C_{NN} is the conductance when both metals are in the normal state, K_1 is the first order of the modified Bessel function of the second kind, e the electron charge, k the Boltzmann constant, T the temperature, and m an integer. Evaluation of (4.9) for special cases is given in Sec. (c) below. Calculations of the current for $eV>\epsilon$ and for tunneling between two superconductors, require more extensive computation.

Finally, it should be mentioned here that we have treated the insulating layer as if it were a vacuum. However, since the insulator has both a conduction band and a valence band, we could possibly also get a "hole" current. In this particular case this is of little importance as we are mostly interested in the current ratio I_{NS}/I_{NN} rather than the absolute values of current.

EXPERIMENTAL RESULTS

(a) Energy Gaps

We report on four different combinations of superconductors namely Al-Al$_2$O$_3$-Pb, Al-Al$_2$O$_3$-Sn, Al-Al$_2$O$_3$-In,

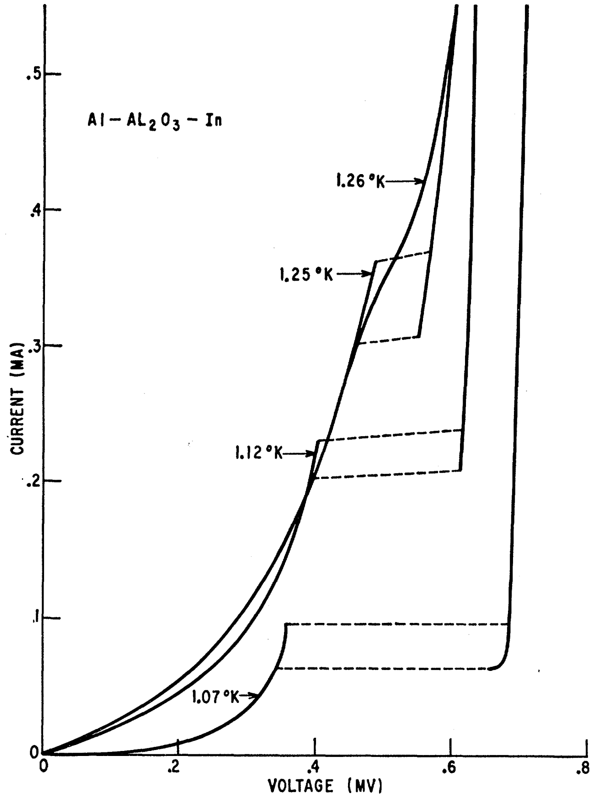


FIG. 9. Detailed current-voltage characteristics of an Al-Al_2O_3-In sandwich, showing the change of energy gap in Al as a function of temperature.

and Al-Al_2O_3-Al. The current-voltage characteristics at various temperatures for these four systems are shown in Figs. 5, 6, 7 and 8, respectively. As seen, the general behavior of the current-voltage characteristics is as predicted from the model. The negative resistance regions are not very apparent on these curves due to the current scale chosen. When the energy gaps on either side of the barrier are equal, as is the case for the Al-Al_2O_3-Al sandwich, a negative resistance region should be observable as well at sufficiently low temperatures. We did not observe this, however, due to temperature limitations in the experimental setup. Note in particular the insert on Fig. 5, showing that for larger voltages the current-voltage characteristics are independent of whether the metals are in the normal or superconducting states. This fact strongly supports the assumption that the tunnel current between superconductors is proportional to the density of states.

From these curves we find the energy gaps at approximately 1°K:

$$2\epsilon_{\mathrm{Pb}}=(2.68\pm0.06)\times10^{-3}\ \mathrm{ev}=(4.33\pm0.10)kT_c,$$

$$2\epsilon_{\mathrm{Sn}}=(1.11\pm0.03)\times10^{-3}\ \mathrm{ev}=(3.46\pm0.10)kT_c,$$

$$2\epsilon_{\mathrm{In}}=(1.05\pm0.03)\times10^{-3}\ \mathrm{ev}=(3.63\pm0.10)kT_c,$$

$$2\epsilon_{\mathrm{Al}}=(0.32\pm0.03)\times10^{-3}\ \mathrm{ev}=(3.20\pm0.30)kT_c.$$

While the energy gaps in Pb, Sn, and In will not change significantly between 1° and 0°K, this is not true for Al, because of its low transition temperature. It should be noted that the transition temperature for the aluminum films varied from sample to sample, was always greater than the bulk transition temperature, and increased with decreasing thickness of the aluminum films. The highest transition temperature observed for Al was 1.8°K. In calculating the energy gaps in terms of kT_c the bulk transition temperature has been used for all films. One reason for this choice is that the observed energy gap in the aluminum films at 1°K is approximately 0.32×10^{-3} ev, regardless of the transition temperatures observed. This experimental result may be due to the broad transition region usually observed in evaporated films.

(b) Variation of the Energy Gap with Temperature

In Fig. 9 we show detailed current-voltage characteristics for an Al-Al_2O_3-In sandwich as a function of temperature. Because the curves are traced out using a constant current source rather than a constant voltage source, the negative resistance region appears as a hysteresis loop. The width of this loop corresponds ap-

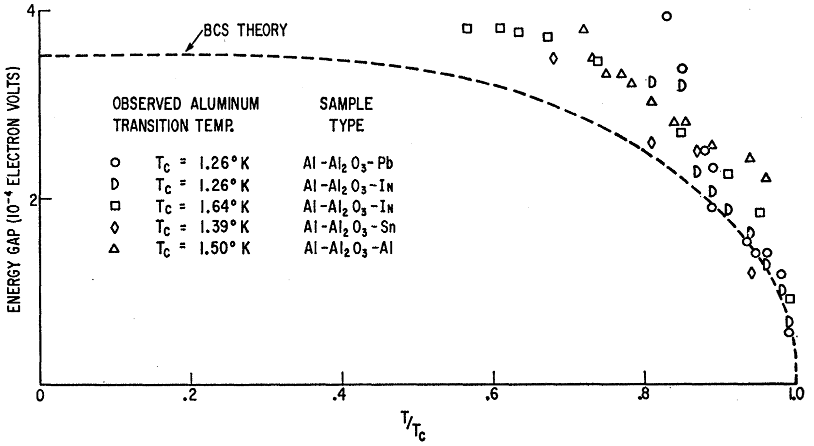


FIG. 10. The energy gap as a function of reduced temperature for several aluminum films, compared with the Bardeen-Cooper-Schrieffer theory.

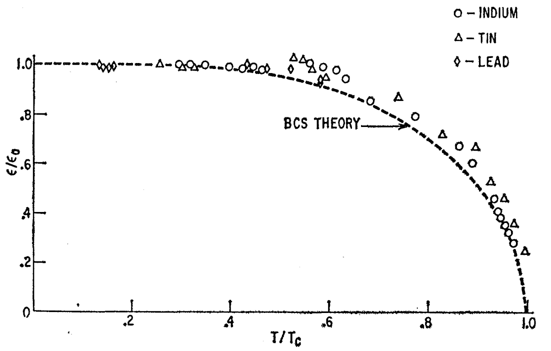


FIG. 11. The energy gap of Pb, Sn, and In films as a function of reduced temperature, compared with the Bardeen-Cooper-Schrieffer theory.

proximately to the full gap width in aluminum, and we can clearly see the variation of gap width with temperature. In Fig. 10 we have plotted the variation of the gap width as a function of reduced temperature for several different samples. For this figure we have used the observed value of the transition temperature T_c. As seen from the figure, the energy gap at $T=0$ does not appear to be very sensitive to the variations in the transition temperature actually observed for the aluminum films. One reason for this could be that the whole area of the aluminum film does not become superconducting at the same temperature, due to localized stresses or impurities. The best estimate of the energy gap for aluminum at absolute zero is

$$2\epsilon_{\rm Al}=(4.2\pm0.6)kT_c=(0.42\pm0.06)\times10^{-3}\ \text{ev},$$

where T_c is taken as the bulk value.

It is possible to observe directly the variation of the energy gap in aluminum over the entire applicable temperature range. The energy gap in indium, tin, and lead can also be observed directly in the temperature range in which aluminum is superconducting. At higher temperatures the gap in lead, tin, and indium is not directly observable; however, we are able to calculate the gap width for all temperatures. By letting $V\rightarrow 0$ in Eq. (4.9), we may write:

$$\frac{I_{NS}}{I_{NN}}=2\sum_{m=1}^{\infty}(-1)^{m+1}m\frac{\epsilon}{kT}K_1\left(m\frac{\epsilon}{kT}\right). \tag{5.1}$$

The quality I_{NS}/I_{NN} as $V\rightarrow 0$ is easily obtained from the experimental results and we may then calculate ϵ from Eq. (5.1). The results are shown in Fig. 11 and are in good agreement with the theory. It should be pointed out that the values of the energy gap, calculated in this way, are in agreement with the directly observed values in the temperature range where both of these measurements can be made. This is most gratifying since these measurements are independent of each other, one being defined at absolute zero, and the other arising solely from the temperature-dependence of the current. The calculated values of the energy gap may appear somewhat too large at low temperatures for some samples due to noise in the measuring circuit.

(c) Calculated versus Measured Current

For tunneling between a metal in the normal state and a metal in the superconducting state, we again use Eq. (4.9) and restrict the calculations to the region where $\epsilon>eV$. In Fig. 12, we compare the calculated values of current with the experimental results obtained on an Al-Al$_2$O$_3$-Pb sandwich at various temperatures. The agreement is very good using only two terms of the series in Eq. (4.9). Note in particular that for $\epsilon\gg kT$ and for large voltages such that $\sinh(eV/kT)\approx\frac{1}{2}\exp(eV/kT)$, we may write

$$\ln I_{NS}=\frac{1}{kT}eV+\alpha(\epsilon,T), \tag{5.2}$$

where α is some function of ϵ and T, independent of V. Thus we can determine the temperature directly from the slope when we plot $\ln I_{NS}$ versus V.

(d) Variation of the Energy Gap with Magnetic Field

By subjecting these samples to a magnetic field parallel to the plane of the metal films, we have found that the energy gap is a function of the applied field. In Fig. 13 we show some detailed results obtained on an Al-Al$_2$O$_3$-Pb sandwich. These results are summarized in

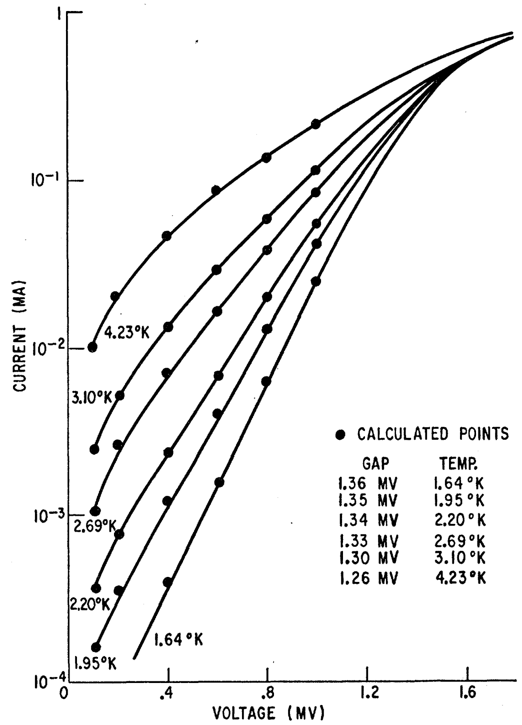


FIG. 12. Observed current-voltage characteristics for an Al-Al$_2$O$_3$-Pb sandwich at various temperatures, versus calculated values using the Bardeen-Cooper-Schrieffer density of states.

Fig. 14 where the gap width for aluminum is shown as a function of magnetic field. This curve does not agree with the observed fact that for bulk materials the transition between the normal and superconducting state is a first-order transition. A first-order transition would require a discontinuous change in gap width at the critical field. While this discrepancy may arise from the possibility that the transition is not of first order in a thin film, we believe it more likely that the surface roughness of the film will cause the magnetic field to be nonuniform. This nonuniformity will tend to smear the discontinuous change in gap that we expect at the critical field.

To make sure that the change in the current-voltage characteristics is due to a change in the energy gap, rather than being due to the aluminum film going into the intermediate state, we also investigated the effect of the magnetic field on the energy gap of lead. In Fig. 15 we plot current versus voltage for an Al-Al$_2$O$_3$-Pb sandwich at various magnetic fields. If we deal with the intermediate state in lead, then the observed current should be the sum of a current varying linearly with voltage and a current varying exponentially with voltage. This is clearly not so. On the other hand, a good fit to these curves can be obtained by using the expression derived for the tunnel current between one normal and one superconducting member with a varying gap for different field strengths. To obtain a good fit, it is necessary to use a rather large gap. This is probably due to noise in the measuring circuit or possibly a nonuniform energy gap in lead. It should be mentioned, although no detailed investigation has been made by us, that for thinner films much higher fields are needed to observe the change in the energy gap.

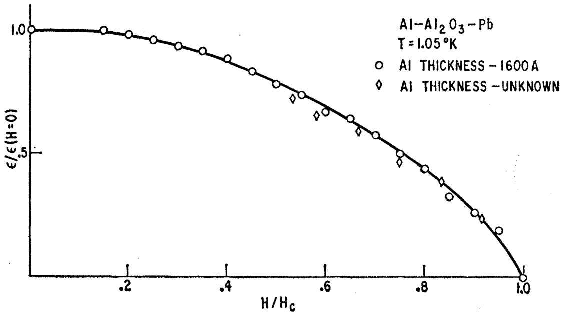


FIG. 14. Apparent variation of the energy gap in an aluminum film as a function of the applied magnetic field.

(e) Density of States

The good agreement between the experimental and calculated currents, using the density of states from the

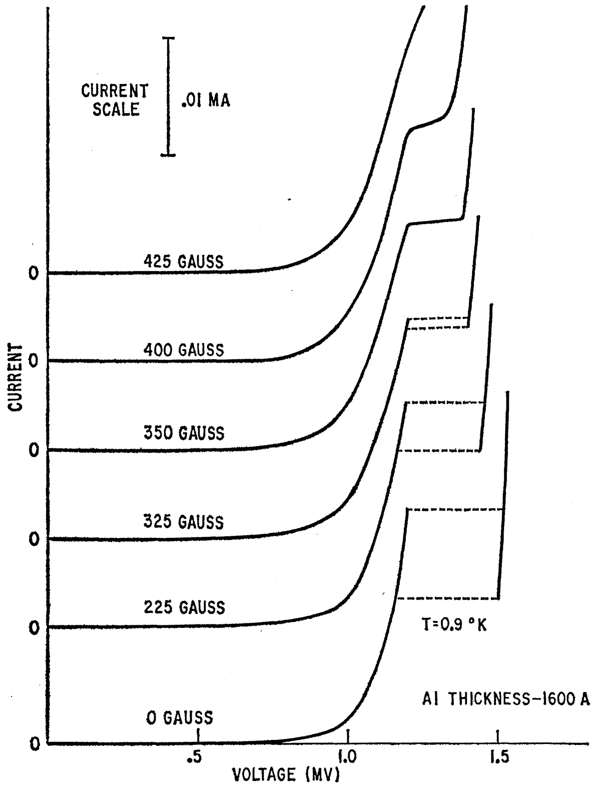


FIG. 13. Detailed current-voltage characteristics of an Al-Al$_2$O$_3$-Pb sandwich, showing a change in the energy gap of aluminum as a function of the applied magnetic field.

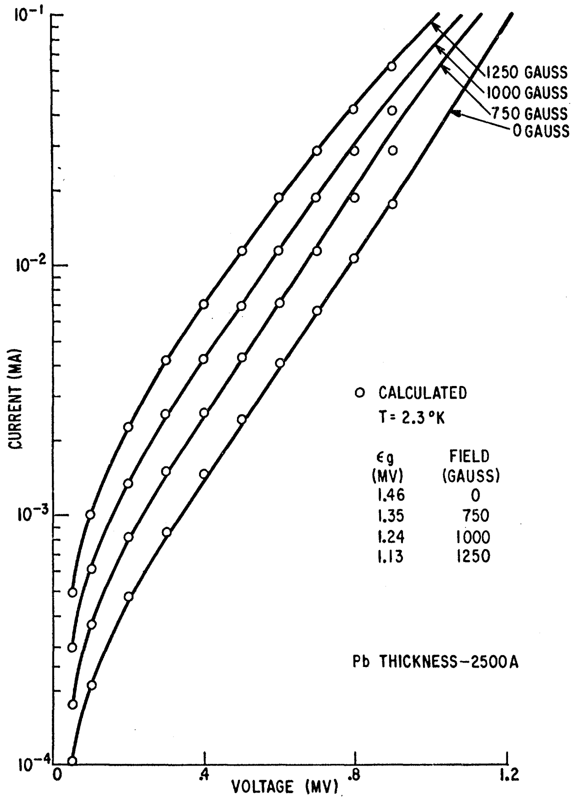


FIG. 15. The change in the current-voltage characteristics of an Al-Al$_2$O$_3$-Pb sandwich as a function of the magnetic field, demonstrating that the observed change cannot be due to the lead film being in the intermediate state.

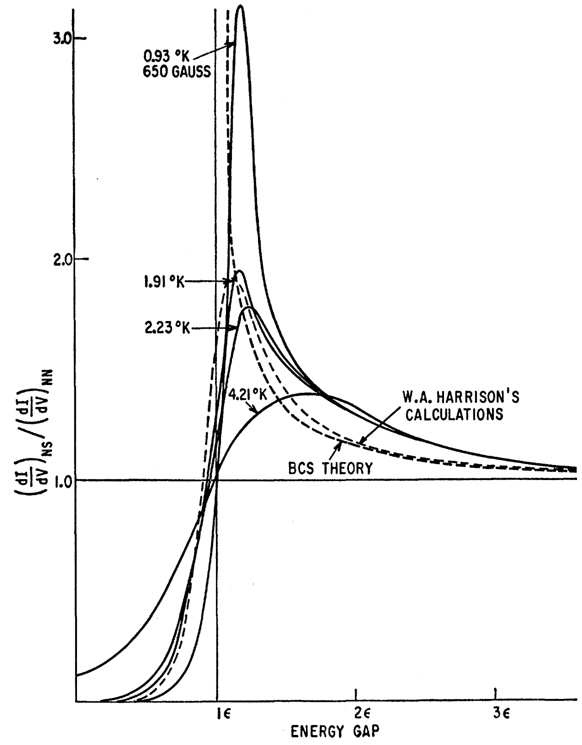


FIG. 16. The relative conductance for an Al-Al_2O_3-Pb sandwich, i.e., the conductance of the sandwich when the lead film is in the superconducting state, divided by the conductance when the lead film is in the normal state, plotted against energy, and compared with the Bardeen-Cooper-Schrieffer density of states. This density of states is used by W. Harrison in his calculations, with $\epsilon/kT=10$.

Bardeen-Cooper-Schrieffer theory, is a great triumph for this theory. In deriving Eq. (4.9), we have integrated over the density of states so that the current is relatively insensitive to small variations in the density of states. Under the assumption that the current is proportional to the density of states, we should get the relative change in the density of states directly by plotting the conductance when one of the metals is superconductive $(dI/dV)_{NS}$ divided by the conductance when both metals are normal $(dI/dV)_{NN}$ against energy. In Fig. 16 we show these results, obtained from an Al-Al_2O_3-Pb sandwich at four different temperatures. Note that at the lowest temperature we have kept the aluminum normal by applying a magnetic field and this again smears the energy gap in lead, making it difficult to assign a specific value to the gap. We see that in spite of the kT smearing, the density of states strongly resembles the theoretical density of states.

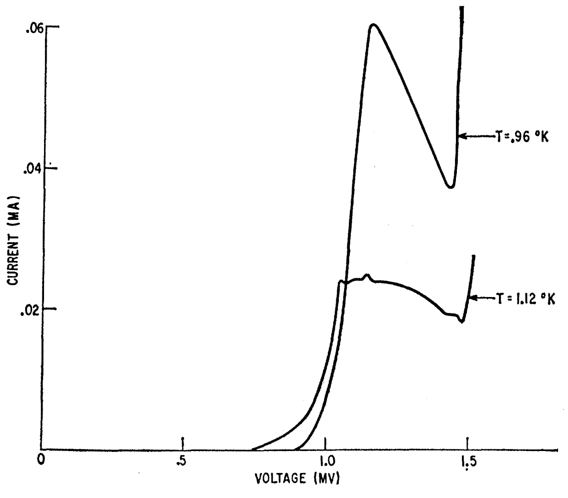


FIG. 17. The negative-resistance region traced out for two different Al-Al_2O_3-Pb sandwiches. We believe the wiggles in the lower curve are due to oscillations in the circuit.

(f) Negative-Resistance Region

In spite of the damping RC network used in parallel with the sample, we found it difficult to eliminate self-induced oscillations in the negative-resistance region. In Fig. 17 we show two attempts to trace out the negative resistance region. We believe that induced noise in the measuring circuit is the limiting factor in tracing out the negative-resistance region, as literally microvolts of induced noise will smear out the curves.

(g) Effect of Metal Bridges and Trapped Flux

In Fig. 18 we show the effect of a metal bridge short-circuiting the sample. The bridge is initially superconducting so that no voltage can be applied across the sample. Then, at a certain current density the bridge becomes normal, but now its resistance is too large to appreciably affect the tunnel current. When the voltage

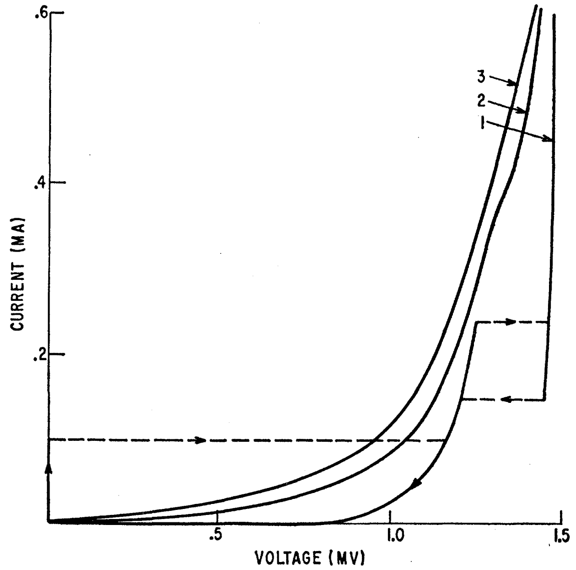


FIG. 18. The effect of trapped flux on the current-voltage characteristic of an Al-Al_2O_3-Pb sandwich. (1) The sample with no field applied; (2) the external field removed, showing the effect of the trapped flux. The figure also shows the effect of a metal bridge across the insulating film; (3) with a magnetic field applied normal to the surface of the films.

is again reduced, the bridge remains normal at a lower current density due to Joule heating.

In Fig. 18 we also show the effect of trapped flux, when the magnetic field purposely has been applied normal to the films. The trapped flux has a large effect upon the current-voltage characteristics, and this technique may possibly be helpful in studying the intermediate state.

(h) Other System

All the experiments we report on have been done by using Al_2O_3 as the insulating layer; however, the experiments may be done by using other insulating layers as well. For example, we have observed tunneling through tantalum and niobium oxides. In these experiments we used bulk specimens of tantalum and niobium; however, we did not observe any evidence for an energy gap in any of these materials. We believe the reason for this is that due to impurities, the surfaces of these materials did not become superconducting. Another superconductor used by us is lanthanum, in which we have observed evidence for an energy gap.

SUMMARY

The method of studying superconductors by electron tunneling has been very successful, and the results are in good agreement with the Bardeen-Cooper-Schrieffer theory. We have directly verified the change of energy gap with temperature. Also, we have shown that for thin films the energy gap is a function of the magnetic field.

ACKNOWLEDGMENTS

We are grateful for the generous and unselfish help received from all members of the Physical Metallurgy Section. We wish to extend special thanks to C. P. Bean, J. C. Fisher, W. A. Harrison, and R. W. Schmitt for their encouragement and interest. Our thanks also to M. V. Doyle for help in performing some of the experiments and L. B. Nesbitt for advice concerning cryogenics.

APPENDIX I

To evaluate the expression:

$$I_{NS}=A'n'(E_F)n(E_F)\int_{-\infty}^{\infty}\frac{|E|}{(E^2-\epsilon^2)^{\frac{1}{2}}}\times[f(E)-f(E+eV)]dE, \quad (A.1)$$

we introduce the conductance C_{NN} when both metals are normal, i.e.,

$$\frac{I_{NN}}{V}=C_{NN}=A'n'(E_F)n(E_F)e, \quad (A.2)$$

and split the integral into two parts:

$$I_{NS}=\frac{C_{NN}}{e}\int_{+\epsilon}^{\infty}\frac{E}{(E^2-\epsilon^2)^{\frac{1}{2}}}\{f(E)-f(E+eV)\}dE$$
$$-\frac{C_{NN}}{e}\int_{-\infty}^{-\epsilon}\frac{E}{(E^2-\epsilon^2)^{\frac{1}{2}}}\{f(E)-f(E+eV)\}dE. \quad (A.3)$$

By introducing $x+\epsilon=E$ in the first integral and $x+\epsilon=-E$ in the second integral, we get

$$I_{NS}=\frac{C_{NN}}{e}\int_{0}^{\infty}\frac{x+\epsilon}{[(x+2\epsilon)x]^{\frac{1}{2}}}[f(x+\epsilon)-f(x+\epsilon+eV)]dx$$
$$+\frac{C_{NN}}{e}\int_{\infty}^{0}\frac{x+\epsilon}{[(x+2\epsilon)x]^{\frac{1}{2}}}$$
$$\times\{f[-(x+\epsilon)]-f[eV-(x+\epsilon)]\}dx, \quad (A.4)$$

and because the Fermi function is an even function,

$$I_{NS}=\frac{C_{NN}}{e}\int_{0}^{\infty}\frac{x+\epsilon}{[(x+2\epsilon)x]^{\frac{1}{2}}}$$
$$\times[f(x+\epsilon-eV)-f(x+\epsilon+eV)]dx. \quad (A.5)$$

By expanding the Fermi function in a series valid for $\epsilon>eV$, we obtain

$$I_{NS}=2\frac{C_{NN}}{e}\sum_{m}(-1)^{m+1}e^{-m(\epsilon/kT)}\sinh(meV/kT)$$
$$\times\int_{0}^{\infty}\frac{x+\epsilon}{[(x+2\epsilon)x]^{\frac{1}{2}}}e^{-m(x/kT)}dx. \quad (A.6)$$

The last integral is of a known Laplace-integral form [A. Erdélyi, *Table of Integral Transforms* (McGraw-Hill Book Company, Inc., 1954)], and we obtain

$$I_{NS}=2C_{NN}\frac{\epsilon}{e}\sum_{m}(-1)^{m+1}K_1(m\epsilon/kT)\sinh(meV/kT), \quad (A.7)$$

where K_1 is the first-order modified Bessel function of the second kind.

APPENDIX II

Note added in proof. It is of interest to compare the values of the energy gaps obtained by using electron tunneling to previous direct measurements of the energy gap (Table II).

TABLE II.

	Energy gap in units of kT_c		
Superconductor	Tunneling measurements	Richards and Tinkham[a]	Ginsburg and Tinkham[b]
Indium	3.63±0.1	4.1±0.2	3.9±0.3
Tin	3.46±0.1	3.6±0.2	3.3±0.2
Lead	4.33±0.1	4.1±0.2	4.0±0.5

[a] P. L. Richard and M. Tinkham, Phys. Rev. **119**, 581 (1960).
[b] D. M. Ginsburg and M. Tinkham, Phys. Rev. **118**, 990 (1960).

Influence of the environment on the Coulomb blockade in submicrometer normal-metal tunnel junctions

A. N. Cleland, J. M. Schmidt, and John Clarke
Department of Physics, University of California, Berkeley, Berkeley, California 94720
and Materials Sciences Division, Lawrence Berkeley Laboratory, Berkeley, California 94720

(Received 22 August 1991)

Submicrometer normal-metal tunnel junctions were fabricated with thin-film leads of either about 2 kΩ/μm or about 30 kΩ/μm. The current-voltage (I-V) characteristics at millikelvin temperatures displayed a much sharper Coulomb blockade for the high-resistance leads than for the low-resistance leads. The zero-bias differential resistance increased as the temperature was lowered, flattening off at the lowest temperatures. A heuristic model based on the quantum Langevin equation is developed, which explains these effects qualitatively in terms of the Nyquist noise generated in the leads; in this model, the flattening of the zero-bias resistance arises from zero-point fluctuations. The data are also compared with a more accurate phase-correlation model that treats the junction and the circuit coupled to it as a single quantum circuit. This model accounts for the observed I-V characteristics quite accurately except near zero bias where it overestimates the dynamic resistance by roughly 50% at the lowest temperatures. This model, however, does not account for the flattening of the zero-bias resistance at the lowest temperatures. It is suggested that the addition of quantum fluctuations in the junction to the phase-correlation theory may account for this discrepancy.

I. INTRODUCTION

Over the past few years there has been a large theoretical and experimental effort to understand the behavior of normal-metal tunnel junctions in the limit where the junction capacitance becomes very small.[1] The inclusion of a term in the Hamiltonian which transfers single electrons introduces additional features in the I-V characteristics of these junctions. One prediction is the appearance at zero temperature of a region of voltage $-e/2C < V < e/2C$ where no tunneling occurs. At any nonzero temperature $T < e^2/2k_BC$ the tunneling rate is exponentially suppressed below that expected from the tunneling resistance. This voltage regime, called the Coulomb blockade region, appears because of the single-electron nature of tunneling: the energy of charging the capacitance C of a small junction to a charge Q is $E_Q = Q^2/2C$, whereas after an electron has tunneled, the energy is $E'_Q = (Q-e)^2/2C$, where $-e$ is the electron charge. The energy change $\Delta E = E'_Q - E_Q$ is positive for $|Q| < e/2$, and negative for $|Q| > e/2$. The additional energy required for an electron to tunnel for $|Q| < e/2$ can be supplied only by a temperature bath, so that, at low enough temperatures, tunneling is energetically unfavorable. For voltages $V = Q/C$ in the Coulomb blockade region, tunneling will therefore not occur. Using electron-beam lithographic technology, one can fabricate tunnel junctions with geometric capacitances of the order of 10^{-15} F, corresponding to the appearance of a Coulomb blockade for $T \lesssim 1$ K. A number of experiments have been performed to investigate this phenomenon,[2-6] with most of them concentrating on multiple junctions connected in series; in general, single junctions did not behave in the manner predicted, and the Coulomb blockade was visible in the current-voltage (I-V) characteristic only as an offset at very large bias currents. There were also indications from the multiple-junction experiments that the single junctions were strongly affected by the external circuit. The multiple-junction experiments of Delsing *et al.*[4,5] and Geerligs *et al.*[6] were interpreted as a single junction isolated from its external environment by the other junctions, although this picture is not clear from microscopic considerations since the multiple junctions clearly constitute a more complex system.

The experimental and theoretical work described here was intended to observe the Coulomb blockade in a single small-capacitance tunnel junction, and to provide a simple explanation for the behavior seen in these and other single-junction experiments. The basic idea behind these experiments was to attempt to isolate the small junction from the leads connected to it by means of thin-film resistors in series with the junctions. We designed these thin-film resistors to operate at high enough frequencies and to have high enough resistance that they could significantly affect the dynamic behavior of the junctions. In the following sections, we describe the design of the small junctions and the resistors connected to them. We then present measurements of the I-V characteristics and the differential resistance as functions of temperature. These data show quite clearly that, for low-resistance leads, the Coulomb blockade is very smeared out, while for higher-resistance leads the blockade is much more sharply defined. We then describe two theoretical approaches to explain these data. The first is a heuristic model based on the quantum Langevin equation and appears to explain the general features. The second, which we refer to as the phase-correlation theory,[7-9] will then be outlined and its predictions compared with the data. Finally, we discuss the possible effects of finite junction-tunnel resistance. A brief report of this work has appeared elsewhere.[10]

II. EXPERIMENTAL DESIGN

A. Thin-film resistors—*RC* transmission line model

The early theories predicting the appearance of the Coulomb blockade in a single, small tunnel junction implicitly assumed that the junction is isolated from the external environment. In any actual measurement, however, the wires attached to the junction electrodes necessarily introduce large stray capacitances, and unavoidably present a high-frequency impedance of order 10^2 to 10^3 Ω at the relevant characteristic frequencies. The presence of this large stray capacitance and small impedance is likely to affect the small junctions strongly, and most certainly has great influence on the behavior seen in the multiple- and single-junction experiments discussed above.

To minimize the effects of the stray capacitance and the low impedance leads, we designed the layout of the junctions to include thin-film resistors on the chip. In the ideal case, an infinite resistance at all relevant frequencies inserted between the small junction and the rest of the circuit isolates the junction, as it then takes an infinite time to resupply the junction with charge from the external leads. In the absence of such idealized leads, theoretical calculations indicate that a shunt resistor must be larger than the quantum of resistance $R_K/4=h/4e^2=6.45$ kΩ for the electrons to be well localized on either side of the junction;[11] the leads must therefore have a total resistance of at least 6.45 kΩ.

One can model the electrical impedance $Z(\omega)$ at frequency ω presented by a straight resistive lead typical of those used in these experiments fairly easily; a metal strip of length Λ and cross-sectional area A has a resistance R_L, a self-capacitance C_L, and a self-inductance L_L, all of which can be calculated from geometric considerations. One finds $R_L=\rho\Lambda/A$, where ρ is the resistivity, $C_L\approx(9.8f\ \text{F/mm})\Lambda$ (assuming the resistor is fabricated on a SiO$_2$ substrate),[12] and $L_L\approx(1\ \text{nH/mm})\Lambda$.[13] These distributed elements from a resistive transmission line of impedance $Z(\omega)\approx R_L$ for frequencies $\omega \ll 1/R_LC_L$ and $Z(\omega)\approx\sqrt{R_L/\omega C_L}(1-i)$ for $\omega \gg 1/R_LC_L$. Given these considerations, we see one can achieve the highest values of $Z(\omega)$ while keeping the stray capacitance at a minimum by using a high-resistivity material with as small a cross-sectional area as possible.

B. Hot-electron effect

The difficulty one encounters when reducing the cross-sectional area of a resistive lead is that hot-electron effects become very important. Electrons lose energy by emitting phonons into the environment, and, for electrons at a temperature T_e, the characteristic frequency of the emitted phonons is $\omega_{\text{ph}}=k_BT_e/\hbar$. For very low temperatures T_e, the available phonon phase space is proportional to ω_{ph}^3, the electron-phonon coupling constant is proportional to ω_{ph} (in the deformation potential approximation), and the energy of the typical emitted phonon is proportional to ω_{ph}. As a result, the electron energy emission rate is proportional to ω_{ph}^5, and the electrons and phonons fall out of thermal equilibrium at low temperatures. The electron and phonon gases are then described by different temperatures T_e and T_{ph}, respectively. If power P is dissipated by Ohmic losses when a bias current I passes through the resistive leads, the electron and phonon temperatures T_e and T_{ph} in the resistive leads satisfy[14]

$$\frac{P}{A\Lambda}=\frac{I^2\rho}{A^2}=(2\times10^9\ \text{W/m}^3\,\text{K}^5)(T_e^5-T_{\text{ph}}^5)\ . \qquad (1)$$

Clearly the resistivity ρ should be minimized and the cross-sectional area A maximized to reduce the electron heating. Unfortunately , this is in direct conflict with our need for high impedances at high frequencies.

C. Resistor design and fabrication techniques

In the final design of the resistive leads we attempted to reduce the heating while maintaining a high impedance to as high a frequency as possible. For our first resistor design, we chose an alloy of Au (25 wt. % Cu) for its reliability and ease of fabrication. A thickness of 30 nm produced a continuous film with resistivity 12 $\mu\Omega$ cm. Using Eq. (1), we find that, for this resistivity and film thickness, a 2-μm-wide resistor should not be heated appreciably by a typical bias current of a few nA (see Fig. 1). To achieve a minimum zero-frequency lead resistance of 60 kΩ [so that $R_L\approx10(R_K/4)$], we were forced to include a large number of meanders on each lead (see Fig. 2). Each meander capacitively couples across itself, and we chose the spacing and length of each meander so that this cou-

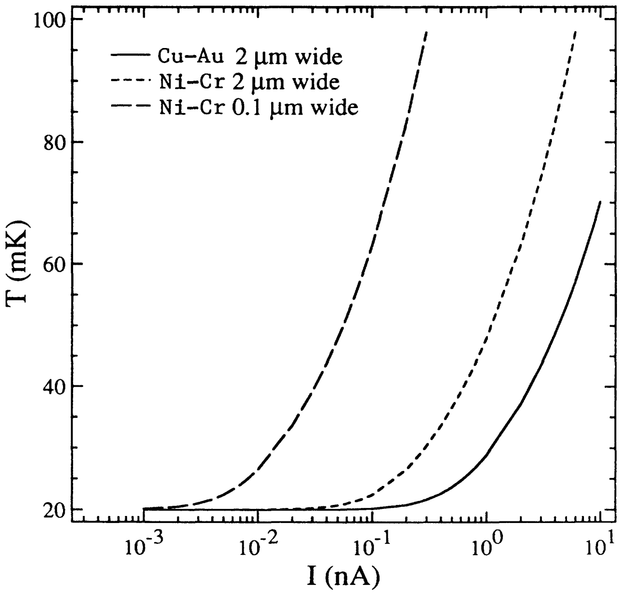


FIG. 1. Calculated dependence of electronic temperature T_e on bias current I, for 30-nm-thick and 2-μm-wide Cu-Au, for 30-nm-thick and 2-μm-wide Ni-Cr, and for 30-nm-thick and 0.1-μm-wide Ni-Cr. The calculations assume a phonon temperature $T_{\text{ph}}=20$ mK.

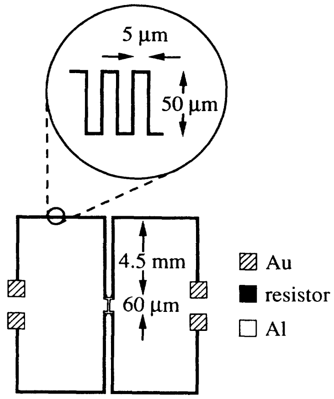


FIG. 2. Layout for the small junction measurements. The Cu-Au leads meandered after the first 4.5 mm leading out from the junction (shown in the inset); the Ni-Cr leads were straight.

pling would begin to roll off the impedance only above 10^{10} Hz. We ignored the self-capacitance of the leads. Thus, the first design involved two, 2-μm-wide lines coming straight out of each lead of the junction, extending 4.5 mm to near the edge of the chip, and then following a 2-μm-linewidth meander around the perimeter of the chip to the four 0.5 mm × 0.5 mm contact pads. We subsequently realized that the self-capacitance of the 2-μm leads was sufficient to effectively short out all but the first few millimeters of the leads above about 10^8 Hz. Thus, the meanders did not provide any significant resistance above that frequency and were replaced by straight lines to the contact pads in the second resistor design. The lines were made of 2-μm-wide Ni (30 wt. % Cr) films 25–30 nm thick with a resistivity of 120 $\mu\Omega$ cm (see Fig. 2). Although not as effective as the Cu-Au resistors for eliminating heating effects, Ni-Cr resistors of this width also should not cause a problem for low bias currents (see Fig. 1). For completeness we also show the electronic temperature versus bias current for a 0.1 μm-wide Ni-Cr resistor and note that heating can be significant.

To fabricate the samples, we used a combination of optical and scanning electron microscope (SEM) lithography. All the samples were made on Si wafers 50 mm in diameter, with a 1000-nm-thick SiO_2 insulating layer. The SEM lithography followed published recipes, with a suspended resist bridge made with a bilayer of PMMA and a copolymer, P(MMA-MAA).[15,16] The junctions were fabricated from Al with a standard angle evaporation technique.

We used evaporated Au pads as a contact surface both for the Al that made up the junctions and for external electrical lead connections. A Cr underlayer 2–3 nm thick was evaporated first to ensure adhesion of the Au layer. Both the Cu-Au and Ni-Cr resistors were deposited by evaporation; the Cu-Au was preceded by a Cr underlayer 2–3 nm thick. The evaporation of Ni-Cr from a length of wire is a technique suggested by Martinis and Kautz.[17]

To fabricate the Al tunnel junctions, we first evaporated a Cu layer, about 5 nm thick, to ensure good electrical contact between the Al and the Au contact pads at the ends of the resistive leads. The Cu film contacted the Al only at the Au contact pads and was at least 200 nm from the junction itself. We then evaporated the Al to a thickness of about 40 nm with a relatively high evaporation rate, 5–10 nm/sec, to prevent the Al from oxidizing. We oxidized the Al in about 0.5 Torr of 30 mol. % O_2+Ar for 10 min, and then deposited the second 40-nm-thick Al film. We lifted off the Al in boiling acetone. Typical junction areas were about 0.02 μm^2, and the contacts to the Au pads at the end of each of the 40-μm-long Al leads were about 10 μm^2.

D. Refrigerator design, wiring, and measurements

The sample mount consisted of a Cu plate bolted to a Cu rod which screwed into the mixing chamber of a dilution refrigerator. A rectangle of G-10 fiberglass was glued to the plate with Stycast 2850 epoxy, and a glass slide was, in turn, glued to the fiberglass with the same epoxy. The Si chip with the junction was attached to the glass slide with vacuum grease and was at least 8 mm from any metal surfaces. The Al junctions were driven into the normal state by two Nd-Fe-B permanent magnets, mounted in a steel yoke, that produced 1100 Oe at the chip (see Fig. 3). We used In to cold weld the four current and voltage leads to the Au contact pads on the chip. Each lead passed through a block of stainless-steel powder (with a grain size of about 50 μm) mixed with Stycast 2850 epoxy; this structure formed a microwave filter with at least 10 dB of attenuation above 1 GHz (see Fig. 3).[18] The leads, junction, and steel powder block were surrounded by a Cu radiation shield. Further filtering of each lead was provided at 4.2 K by filters potted in stainless-steel powder and epoxy, with a 3-dB rolloff point at 16 kHz, and at room temperature by a radiofrequency filter, with 120 dB of attenuation above 100 MHz (see Fig. 4).

Our current supply consisted of a voltage source in series with a 50-MΩ resistor, a 2-MΩ resistor to measure the current, and a low-pass RC filter with a rolloff frequency adjustable to between 3 and 100 Hz. A low-noise amplifier was connected across the voltage leads. The

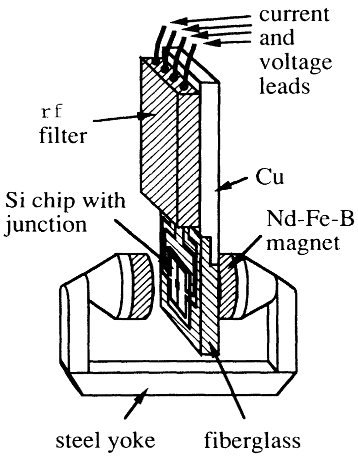


FIG. 3. Sketch of junction mount.

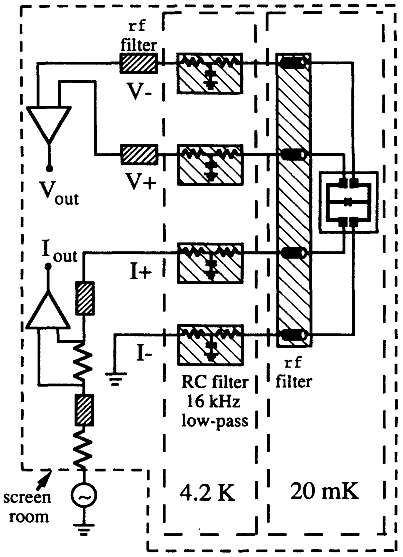


FIG. 4. Wiring schematic for dilution refrigerator.

amplifiers were battery powered and, together with the current supply, were inside the screened room. No line power was brought into the screen room during the measurements.

We recorded I-V traces on an analog X-Y plotter, and obtained I versus dV/dI traces using a lock-in technique with a very low-frequency sweep for the current (typically 1–3 mHz), and a 15–20-Hz current modulation for the lock-in measurement. The modulation amplitude was typically 1% of the full sweep amplitude: for example, for a 1-nA sweep we used a 10-pA peak-to-peak modulation current. The alternating voltage was lock-in detected, and the traces were recorded on an X-Y plotter. Reducing the modulation amplitude by a factor of 5 had no effect on the traces. We calibrated the differential resistance by comparing it with an I-V trace.

The two largest sources of spurious noise were 60-Hz pickup and vibrational noise, the latter presumably induced by wires moving relative to ground. We reduced the 60-Hz pickup to about 20 nV peak-to-peak across each pair of leads by very carefully isolating the screened room from all grounds. A μ-metal shield was placed around the lower part of the refrigerator, at the height of the sample. A second μ-metal shield could either be placed concentrically to the first shield, or placed around a platform which held the amplifiers and the current-limiting resistors forming a small magnetic and electric screened room, internal to the Cu-mesh screen room.

III. EXPERIMENTAL RESULTS

A. *I-V* characteristics and *dV/dI* measurements

We fabricated and measured nine small junctions. In Fig. 5, we show the I-V characteristics at 20 mK of two typical junctions, one with Cu-Au leads with a lead resistance of 2 kΩ/μm (junction 5 in Table I), and the other with Ni-Cr leads with a resistance of 30 kΩ/μm (junction

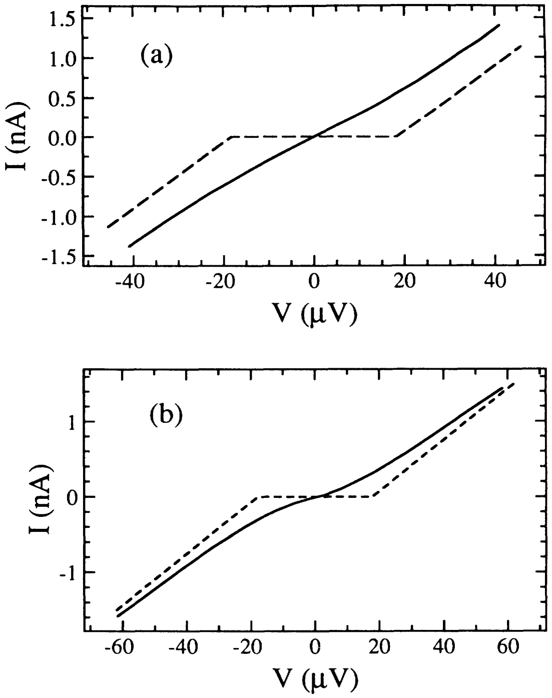


FIG. 5. I-V characteristics (solid lines) measured for two small junctions at $T=20$ mK. (a) Junction 5 with Cu-Au leads, $R_J=23$ kΩ, and $C_J=4\pm1$ fF. (b) Junction 7 with Ni-Cr leads, $R_J=29.4$ kΩ, and $C_J=5\pm1$ fF. Dotted lines show predicted voltage-biased Coulomb blockade for each junction.

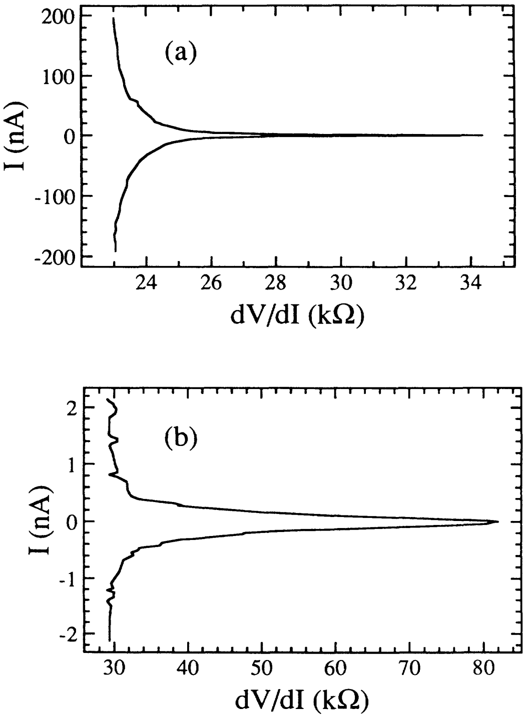


FIG. 6. Measured dV/dI at 20 mK (a) junction 5 and (b) junction 7; note the difference in current scales.

TABLE I. Summary of experimental parameters. R_J is the tunneling resistance (measured at high bias), C_J the junction capacitance inferred from the high-bias voltage offset, ZBR the zero-bias resistance at $T=20$ mK, and R_L the resistance of one lead to the junction. Lead material is given in the last column.

Junction	R_J (kΩ)	C_J (fF)	ZBR (kΩ)	R_L (kΩ)	Lead material
1	6.0	3±0.5	6.9	132	Cu-Au
2	32.2	12[a]	44	140	Cu-Au
3	11.3	3±1	14.8	162	Cu-Au
4	27	3±1	39.5	92	Cu-Au
5	23	4±1	34.3	150	Cu-Au
6	8.8	6.5±1	18.4	390	Ni-Cr
7	29.4	5±1	82	350	Ni-Cr
8	133	3±1	464	350	Ni-Cr
9	82	3.5±1	239	340	Ni-Cr

[a] Only a small I_{bias} is used, hence C is probably overestimated.

7 in Table I). The dotted lines show the voltage-biased Coulomb blockade characteristic expected at $T=0$. It is clear that the Coulomb blockade is only barely visible in junction 5, while it is quite sharply defined in junction 7. Figure 6 shows the differential resistance; it is clear that for junction 5 one must apply about 100 nA of bias current before the differential resistance approaches its asymptotic value R_J, while for junction 7 one only need apply about 1 nA. The lead resistance clearly affects the I-V characteristic profoundly. In Fig. 7 we plot the zero-bias resistance (ZBR) of seven junctions, normalized to R_J, as a function of the refrigerator temperature. The junctions with Ni-Cr leads (solid symbols) all show a significantly higher resistance than the junctions with Cu-Au leads (open symbols). We can see also that the ZBR flattens out as the temperature is reduced, at a somewhat lower temperature for the junctions with the Ni-Cr leads than for the junctions with Cu-Au leads. In Table I, we summarize the parameters of nine junctions we have studied.

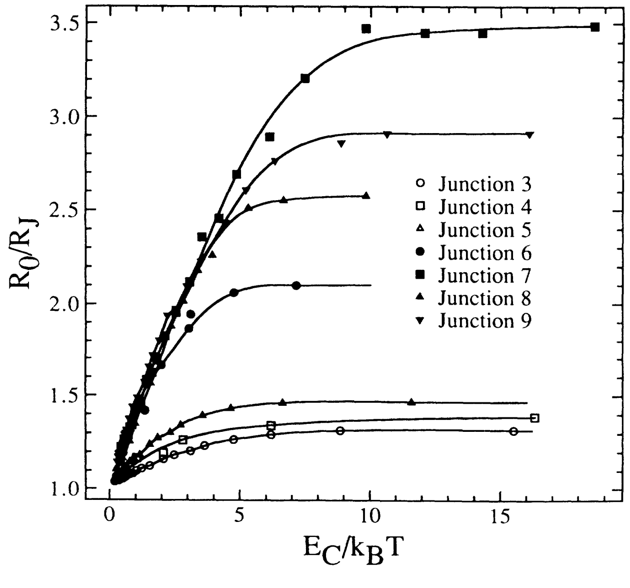


FIG. 7. Normalized zero-bias resistance vs temperature for seven small junctions. Open symbols are the Cu-Au leads, solid symbols for Ni-Cr leads. Lines are guides for the eye.

B. Diagnostic measurements

We made a number of attempts to explain the behavior of the ZBR as the temperature is lowered. We tested the possibility that the flattening was due to spurious high-frequency noise (radio- or microwave-frequency noise) by adding and removing the noise filters on the current and voltage leads at all three temperature stages (300 K, 4.2 K, and 20 mK). This had no effect on the low-temperature limit of the ZBR.

Another possibility is that the external circuit was still loading the junctions. We tested this possibility with one junction of each type by shorting out all but the first 4.5 mm of each lead with a layer of In (within the 4.5 mm section, the leads are too close together to be individually shorted with this technique). This procedure also had no effect on the ZBR, implying that the external circuit has been effectively isolated by the resistors, and that only the first few millimeters (at most) of the resistors affect the junctions.

As was mentioned earlier, there is the possibility of heating in the resistive leads, which could cause the flattening in the data. However, the design of the leads was intended to avoid problems caused by the hot-electron effect. Also it should be noted that the trend for the flattening to occur at a lower temperature for the higher resistivity Ni-Cr leads is inconsistent with this model. The fact that reducing the magnitude of the current modulation used for the differential measurements did not affect our results indicates that any heating caused by our measurement techniques should not be a factor, at least in the zero-bias limit. Finally, the magnitude of the spurious low-frequency noise also does not provide enough power to cause heating at the level needed to explain the data.

We conclude that the flattening of the ZBR at low temperature is most likely an intrinsic effect, and turn to a theoretical discussion of both this issue and the overall behavior of the I-V characteristics. We present two models which explicitly include the effects of the external electromagnetic environment and appear to reproduce the main features seen in the data. One model was presented by us previously;[10] the other has been discussed by Nazarov[7] and developed further by other authors.[8,9] We also consider quantum fluctuations in the junctions themselves as a possible mechanism for flattening the ZBR as the temperature is reduced.

IV. THEORY OF ZERO-BIAS RESISTANCE

A. Quantum Langevin equation

Consider first the circuit, shown in Fig. 8(a), consisting of a small tunnel junction with capacitance C_J and tunnel resistance R_J in series with an ideal resistor R_L and a large stray capacitance $C_{\text{stray}} \gg C_J$. Let us place a charge Q on C_J at $T=0$: What happens as the charge Q is varied? For a charge $|Q|<e/2$ an electron must gain energy in order to tunnel across the junction barrier, and tunneling therefore occurs only for $|Q|>e/2$. If an electron takes a time τ to tunnel across the junction barrier and the circuit time constant $\tau_{RC}=R_LC_J \gg \tau$, then on the scale of τ the small junction in Fig. 8(a) *acts* as if it were isolated from the external circuit. Provided $\tau_{RC} \gg \tau$, the junction should therefore allow electrons to tunnel only for $|Q|>e/2$, and for $|Q|<e/2$ no tunneling should occur. A measurement of the I-V characteristic will therefore not reveal the presence of C_{stray}; however, something is missing from this picture, given the experimental evidence presented earlier.

How difficult is it to achieve the inequality $\tau_{RC} \gg \tau$? If we take[19] $\tau \approx 10^{-15}$ sec and $C_J=10^{-15}$ F, then, even for a resistance R_L as low as 10 Ω, we find $\tau_{RC}=10^{-14}$ sec, and the inequality is satisfied. However, if we consider the Heisenberg uncertainty relation $\Delta E\, \Delta t > \hbar$, the energy corresponding to the RC time constant is equivalent to a charge of about 20 e on C_J, so that the charge on the junction is not known very precisely.

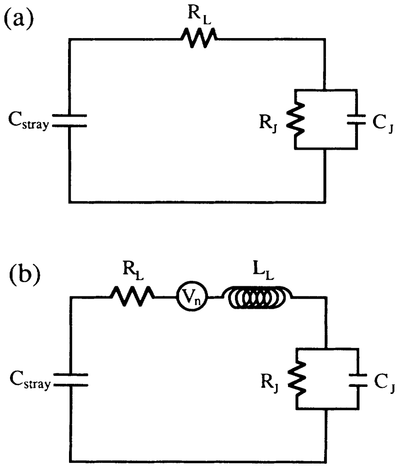


FIG. 8. (a) Circuit for small tunnel junction with resistance R_J and capacitance C_J connected through a large lead resistor R_L to low-impedance coaxial lines modeled by a large stray capacitance C_{stray}. (b) Circuit for a small tunnel junction, including lead resistance R_L and inductance L_L, and a large stray capacitance C_{stray}. Voltage noise source V_n is associated with R_L.

To complete this description, we must associate a voltage noise source V_n with R_L, as shown in Fig. 8(b), where we have also included the lead inductance L_L. The voltage source V_n drives current in the circuit loop, causing the charge Q to fluctuate over times much longer than τ. The classical equation of motion for the fluctuations $q(t)$ of the charge Q is given by

$$L_L\frac{d^2q}{dt^2}+R_L\frac{dq}{dt}+\left[\frac{1}{C_{\text{stray}}}+\frac{1}{C_J}\right]q=V_n(t)\ . \tag{2}$$

We Fourier transform Eq. (2) and solve for the spectral density of the charge fluctuations in terms of the spectral density of the voltage noise, $S_V(\omega)$:

$$S_q(\omega)=\frac{C_J^2}{(1-\omega^2/\omega_{LC}^2)^2+(\omega/\omega_{RC})^2}S_V(\omega)\ . \tag{3}$$

Here we have defined the frequencies $\omega_{RC}=1/R_LC_J$ and $\omega_{LC}=1\sqrt{L_LC_J}$, and neglected the stray capacitance with the assumption $C_{\text{stray}} \gg C_J$. To include quantum fluctuations we use the spectral distribution of voltage noise given by the full Johnson-Nyquist formula:[20]

$$S_V(\omega)=\frac{\hbar\omega R_L}{\pi}\coth(\hbar\omega/2k_BT)\ . \tag{4}$$

This expression reduces to $S_V(\omega)=2k_BTR_L/\pi$ in the limit $\hbar\omega \ll k_BT$ and to $S_V(\omega)=\hbar\omega R_L/\pi$ for $\hbar\omega \gg k_BT$. Even at $T=0$ there are significant voltage fluctuations caused by the resistor.

Using the Wiener-Khinchine theorem, we can write the instantaneous charge fluctuation $\langle q^2(t)\rangle$ in terms of the frequency-domain fluctuations,

$$\langle q^2(t)\rangle=\int_0^{\infty}S_q(\omega)d\omega\ . \tag{5}$$

We can carry out the integral in Eq. (5) analytically in two limits; for $\hbar\omega \ll k_BT$ we obtain the result expected from equipartition of energy, $\langle q^2\rangle/2C_J=k_BT/2$. For $\hbar\omega \gg k_BT$, for $\alpha=\omega_{LC}/\omega_{RC}<2$, we find

$$\frac{\langle q^2(t)\rangle}{2C_J}=\frac{\hbar\omega_{LC}}{2\pi}\frac{1}{\sqrt{4-\alpha^2}}\left[\frac{\pi}{2}-\tan^{-1}\left[\frac{\alpha^2-2}{\alpha\sqrt{4-\alpha^2}}\right]\right]\ . \tag{6}$$

For very small α, Eq. (6) reduces to $\langle q^2(t)\rangle/2C=\hbar\omega_{LC}/4$, as expected for a simple harmonic oscillator with no dissipation. For $\hbar\omega \gg k_BT$ and $\alpha>2$, we find

$$\frac{\langle q^2(t)\rangle}{2C_J}=\frac{\hbar\omega_{LC}}{4\pi}\frac{1}{\sqrt{4-\alpha^2}}\ln\left[\frac{\alpha^2-2+\alpha\sqrt{\alpha^2-4}}{\alpha^2-2-\alpha\sqrt{\alpha^2-4}}\right]\ . \tag{7}$$

In the limit of very large R_L we find that $\langle q^2\rangle/2C=(\hbar\omega_{RC}/\pi)\ln\alpha$, which varies as $1/R_L$ and has only a logarithmic dependence on L_L. In Fig. 9(a) we plot $\langle q^2\rangle/e^2$ versus R_L for $L_L=4.5$ nH and $C_J=4.5$ fF at $T=0$. We have chosen the values of C_J and L_L to be representative of the capacitances of junctions 5 and 7

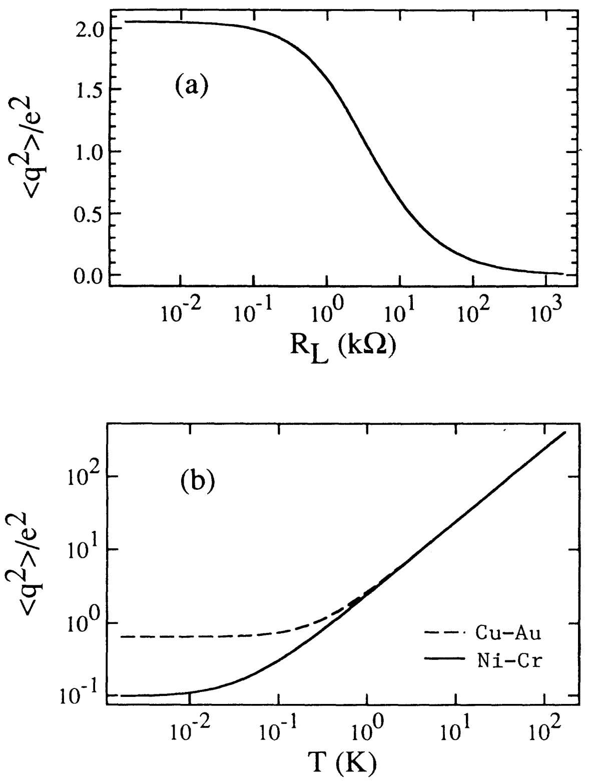


FIG. 9. (a) Dependence of $\langle q^2\rangle/e^2$ on R_L for junction capacitance $C_J=4.5$ fF, lead inductance $L_L=4.5$ nH, calculated using the quantum Langevin theory at $T=0$. (b) Plot of temperature dependence of $\langle q^2\rangle/e^2$, calculated from the quantum Langevin theory, for both Cu-Au and Ni-Cr leads; see text for circuit parameters.

and the inductance of the first 4.5 mm of the resistive leads, respectively. In Fig. 9(b) we display the temperature dependence of $\langle q^2\rangle/e^2$ for $L_L=4.5$ nH, $C_J=4.5$ fF, and $R_L=9.0$ and 130 kΩ, the latter being the resistance for the first 4.5 mm of the Cu-Au and Ni-Cr leads, respectively.

The value of $\langle q^2\rangle$ is the spread of the distribution $P(q)$, which describes the probability of having a charge fluctuation of size q on the small junction capacitor C_J; previously we assumed the probability distribution to be $P(q)=\delta(q)$. The spread in the values of q arises, in general, from both thermal and quantum fluctuations. In the limit of zero temperature, the probability distribution is the square of the wave function (i.e., the probability amplitude) of the variable q.

To accommodate the spread in the values of q in the calculation of the current-voltage characteristic, we assume that the effective tunneling rate $\langle\Gamma(Q)\rangle$ of electrons is the tunneling rate $\Gamma(Q)$ in the absence of fluctuations, convolved with the probability $P(q)$ of a given size fluctuation q. As the fluctuations are concentrated at frequencies much below the inverse electron tunneling time $1/\tau$, this should be a reasonable approximation. The resulting expression is

$$\langle\Gamma(Q)\rangle=\int_{-\infty}^{\infty}\Gamma(Q+q)P(q)dq\ . \tag{8}$$

An expression for $\Gamma(Q)$ can be derived analytically:[21]

$$\Gamma^{\pm}(Q)=-\frac{e/2\pm Q}{eR_JC_J}[\exp(\Delta E^{\pm}/k_BT)-1]^{-1}\ , \tag{9}$$

where $\Gamma^{\pm}$ is the rate for Q to go to $Q\pm e$. Assuming the probability distribution $P(q)$ to be Gaussian,

$$P(q)=\frac{1}{\sqrt{2\pi\langle q^2\rangle}}\exp\left[-\frac{q^2}{2\langle q^2\rangle}\right]\ . \tag{10}$$

We can carry out the convolution of Eq. (8) analytically at $T=0$ to obtain

$$\langle\Gamma^{\pm}\rangle=-\frac{e/2\pm Q}{2eR_JC_J}\mathrm{erfc}\left[\frac{e/2\pm Q}{\sqrt{2\langle q^2\rangle}}\right]+\frac{1}{eR_JC_J}\sqrt{\langle q^2\rangle/2\pi}\exp\left[-\frac{(e/2\pm Q)^2}{2\langle q^2\rangle}\right]\ . \tag{11}$$

In Fig. 10 we plot $\langle\Gamma(Q)\rangle$ and $\Gamma(Q)$ at $T=0$ for $\langle q^2\rangle=0.65$ and 0.098 e^2 (the values for the Cu-Au and Ni-Cr leads) to show the smearing of the Coulomb gap due to the uncertainty in the value of Q. Note that at large values of Q, the two rates are identical; the smearing only occurs for values of Q within $\sqrt{\langle q^2\rangle}$ of $\pm e/2$.

Using the above discussion, we can explain the main experimental features of the data in terms of this simple model: The smearing of the Coulomb gap at low-bias voltages is caused by fluctuations in the leads, becoming less apparent as the fluctuations are reduced in magnitude by increasing the value of the lead resistance, or by reducing the temperature. We furthermore see that the zero-temperature limit of the current-voltage characteristic can be explained by the inclusion of the zero-point quantum fluctuations in the leads. To make a more quan-

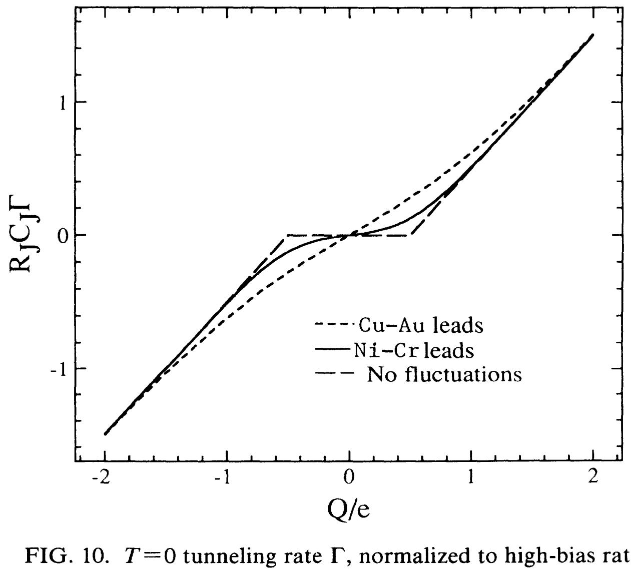


FIG. 10. $T=0$ tunneling rate Γ, normalized to high-bias rate $1/R_JC_J$ as predicted by the quantum Langevin theory vs charge Q/e. The dotted curve is for Cu-Au leads and the solid curve is for Ni-Cr leads, with circuit parameters given in text. The dashed curve is tunneling rate in the absence of fluctuations.

titative comparison between them and experiment, we have to choose parameters for our circuit model; this simple quantum Langevin approach unfortunately does not yield solutions for the more complete resistive transmission line model of the leads. To make a reasonable approximation to the actual circuit, we choose to include only the first 4.5 mm of each lead, and treat the estimated values of the resistive transmission line as lumped circuit elements choosing the total inductance to be 4.5 nH, and the resistance to be 9 kΩ for the Cu-Au leads and 130 kΩ for the Ni-Cr leads. We estimated the junction capacitance from the high-current limit of the Coulomb offset for the two junctions; according to the theory sketched out above, this procedure should be correct even in the presence of fluctuations.

One part of the experimental situation which we have thus far ignored in our use of a current bias rather than the voltage bias used in our calculations of the tunneling rates The situation is complicated by the presence of the stray capacitance in the circuit; in Fig. 11(a) we show the model circuit with a current bias element I_{bias} connected across the stray capacitance C_{stray}. The junction is modeled as a capacitance in parallel with a current source $I_J(t)$ that transfers single electrons across the junction. The junction follows a charging-discharging sequence, and on average the time between discharging events is given by $\Delta t = e/I_{\text{bias}}$. The junction voltage follows two distinct patterns, depending on whether the charging time $\tau = R_L C_J$ is much larger or much smaller than Δt (the inductance L_L is too small to have a noticeable effect). In the limit $\tau \ll \Delta t$, shown in Fig. 11(b), the voltage recovers rapidly after an electron has tunneled and spends most of its time before the next tunneling event near the average value $\langle V_J \rangle$. This voltage then satisfies the relation

$$\langle \Gamma(\langle V_J \rangle) \rangle = I_{\text{bias}}/e \ . \tag{12}$$

In other words, in this limit the junction is effectively voltage biased, and the I-V characteristic is given by Eq. (12), where $\langle \Gamma(V) \rangle$ is the result of Eq. (11).

In the opposite limit $\tau \gg \Delta t$, shown in Fig. 11(c), the junction voltage ramps up linearly and the junction is effectively current biased. One must calculate the I-V characteristic numerically by following the charging-discharging sequence in time. To calculate the I-V characteristic without using an inordinate amount of computation time, we developed an approximate method, which, for the experimental data, is estimated to have errors of at most 3%. In the approximation, we fix the time between tunneling events at $\Delta t = e/I_{\text{bias}}$, rather than allow this time to fluctuate because of the stochastic nature of electron tunneling. The current source $I_J(t)$ then provides a regular series of δ functions, with amplitude $-e$ separated in time by Δt. We calculate the voltage $V(t)$ by solving the circuit equations for Fig. 11(a) at a fixed I_{bias}, finding the mean voltage $\langle V(t) \rangle$ by solving the equation

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} \langle \Gamma[V(t)] \rangle dt = I_{\text{bias}}/e \ . \tag{13}$$

Self-consistent solutions of Eq. (13) as a function of I_{bias} yield the I-V characteristic.

In Fig. 12 we plot the comparison of the theory at $T=0$ and the experimental data. The data for the Cu-Au leads are in reasonable agreement at low bias and less

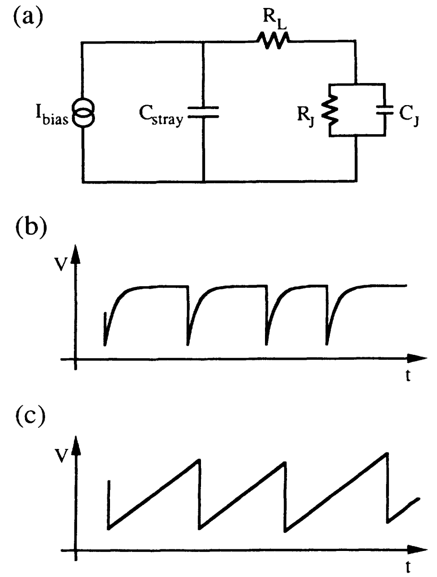


FIG. 11. (a) Model circuit for nearly current-biased measurement. Charging-discharging sequence for bias current I_{bias} with (b) $R_L C_J \ll e/I_{\text{bias}}$ and (c) $R_L C_J \gg e/I_{\text{bias}}$.

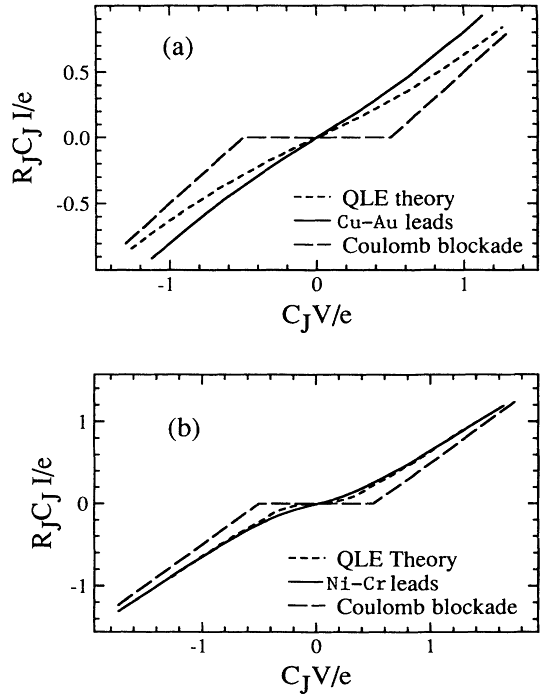


FIG. 12. Comparison of the quantum Langevin theory with experimental I-V characteristics: (a) is junction 5 with Cu-Au leads and (b) is junction 7 with Ni-Cr leads. The solid dots are quantum Langevin equation (QLE) predictions at $T=0$, and the dashed line is the predicted voltage-biased Coulomb blockade.

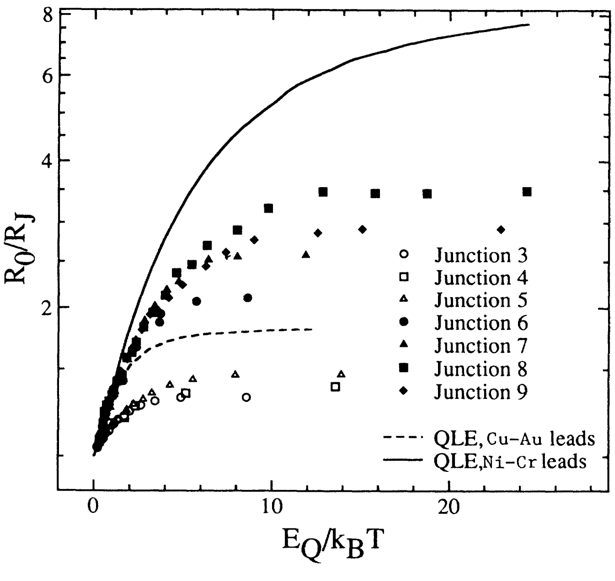


FIG. 13. Zero-bias resistance predicted by the quantum Langevin equation (lines), compared with the experimental results (symbols) for seven junctions. Open symbols and dashed line are Cu-Au leads, solid symbols and solid line are Ni-Cr leads.

good agreement at high bias, while the data for the Ni-Cr leads show reasonable agreement over the entire range of displayed bias currents. In Fig. 13 we have plotted the calculated temperature dependence of the zero-bias resistance together with the experimental results. The experiment and theory are seen to have the same approximate shape, with roughly the same rate of increase of resistance with decreasing temperature, and to flatten out at roughly the same value of resistance. Note that, by fitting the parameters, in particular, the length of lead used in the model circuit, we could obtain a much better agreement for the temperature dependence of the ZBR. However, the heuristic nature of this model is such that we wish only to show that the general behavior can be predicted; we do not claim that this model includes all relevant details. A discussion of the more rigorous theory follows.

B. Phase-correlation theory

Several authors have published different versions of an alternative theory to the quantum Langevin theory presented above.[7–9] We here summarize the version described by Devoret *et al.*[8]

The essence of the approach is to treat both the junction capacitor and its environment as a single quantum system. The tunneling in the junction is then added as a perturbative term in the Hamiltonian, coupled to the electromagnetic system. In this way the energy of the entire circuit is included when one calculates the tunneling rate of electrons across the junction; this is in contrast to the simplest theory where the external circuit is ignored and only the charging energy of the junction capacitance is considered. Taking this approach, Devoret *et al.* develop an expression for the phase-correlation function $J(t)$ given by

$$J(t)=\langle[\delta(t)-\delta(0)]\delta(0)\rangle \ , \tag{14}$$

where

$$\delta(t)=\int_{-\infty}^{t}V(t')dt' \tag{15}$$

is the phase variable across the normal junction. The phase-correlation function is given by

$$J(t)=\int_0^\infty \frac{d\omega}{\omega}\frac{\mathrm{Re}Z_t(\omega)}{(R_K/2)} \times\left[\coth\left[\frac{\beta\hbar\omega}{2}\right][\cos(\omega t)-1]-i\sin(\omega t)\right] , \tag{16}$$

where we define $\beta=1/k_BT$, the total impedance function $Z_t(\omega)$ is given by $Z_t^{-1}(\omega)=i\omega C+Z^{-1}(\omega)$, and the normal-metal resistance quantum is $R_K/2=h/2e^2=12.91\ \mathrm{k\Omega}$. Knowledge of the phase-correlation function allows one to calculate the probability $P(E)$ that an electron will lose energy E in the circuit through the expression

$$P(E)=\frac{1}{2\pi\hbar}\int_{-\infty}^{\infty}dt\,\exp[J(t)+iEt/\hbar] \ . \tag{17}$$

When the impedance of the environment approaches the infinite limit, $P(E)$ approaches a δ function, and the electron cannot exchange energy with the environment; the junction becomes isolated from its environment as we described earlier. For a finite impedance, $P(E)$ is distributed, and from this the tunneling rate Γ at a voltage $V=Q/C$ can be calculated from the expression

$$\Gamma(V)=\frac{1}{e^2R_J}\int_{-\infty}^{\infty}dE\,E\frac{1-\exp(-\beta eV)}{1-\exp(-\beta E)}P(eV-E) \tag{18}$$

for a junction with tunnel resistance R_J. Because the tunneling is included only as a perturbation, this derivation assumes that quantum fluctuations in the junction itself can be neglected, i.e., that R_J is much larger than the resistance quantum $R_K/2$; it also assumes $\mathrm{Re}[Z(\omega)]\ll R_J$. An alternative formulation of this theory appears in a paper by Ingold and Grabert.[22]

Equation (18) gives the voltage-biased tunneling rate $\Gamma(V)$; just as for the case of the quantum Langevin equation discussed in Sec. IV A, we have to account for the actual biasing circuit used in the experiment. We have chosen to use the same approximate method described in that section, using a fixed time $\Delta t=e/I_{\mathrm{bias}}$ between tunneling events. However, we replace the simple *RLC* circuit with the impedance $Z_t(\omega)$ inferred from the measured static resistance of the leads, and the numerically calculated values of the capacitance and inductance of the leads, as discussed in Sec. II A.

One would, of course, prefer to use an impedance $Z(\omega)$ obtained from a direct measurement. However the relevant frequencies involved in the phase-correlation function, Eq. (16), are of the order of $e^2/2Ch \approx 10^{10}$ Hz, a range in which such measurements are difficult. The circuit in question involves the thin-film leads, which can be accurately modeled as a resistive transmission line, connected to the low-impedance coaxial measurement leads which present a more complicated impedance. However, the impedance of the thin-film leads is large enough (of the order of a few kΩ) even at the relevant frequencies to isolate the junction from the rich environment of the coaxial leads, and we thus expect the simple impedance model to be adequate.

Figure 14 shows the calculated I-V characteristics compared with those for junctions 5 and 7, taken at $T=20$ mK. In Fig. 15 we compare the differential resistance predictions with the experiment; the temperatures for the theory are as in Fig. 14. We see that, in both cases, the high-current I-V characteristic agrees fairly well, but at lower currents the I-V shows less curvature than the theory predicts. This is borne out better in the differential resistance traces, where we find the predicted ZBR is somewhat higher than measured, and the differential resistance curve is broader in the measurement than in the theory, especially in the case of Cu-Au resistors.

A possible source of error in the calculated curves involves the choice of the parameters for the transmission line, in particular, the value of the capacitance per unit length. We calculate this capacitance numerically by allowing charge to flow onto a model section of the transmission line until the potential is uniform; the ratio of the potential to the charge gives the capacitance. This method, when applied to simple geometries, gives the correct result for the capacitance to within 20%. Variations in the value of C_L of this magnitude for the calcula-

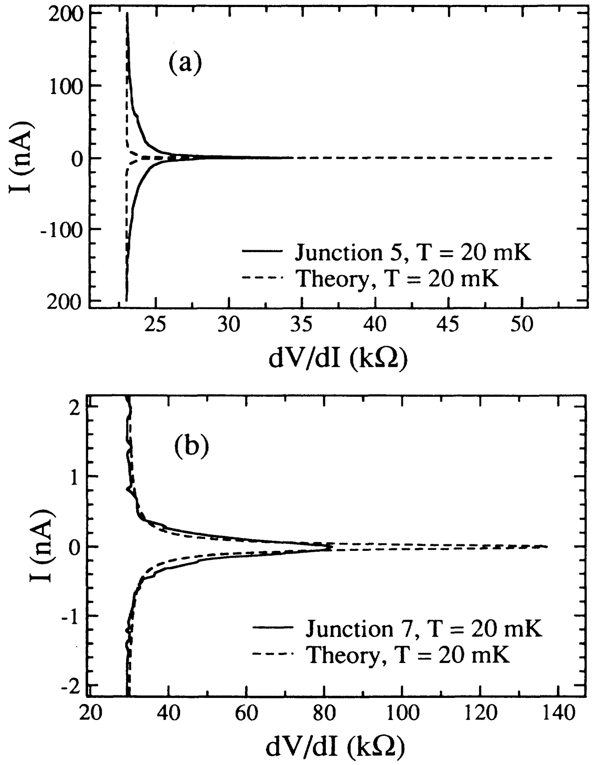


FIG. 15. Comparison of the theory of Devoret *et al.* (Ref. 20) with experimental dV/dI measurements at $T=20$ mK: (a) junction 5 and (b) junction 7.

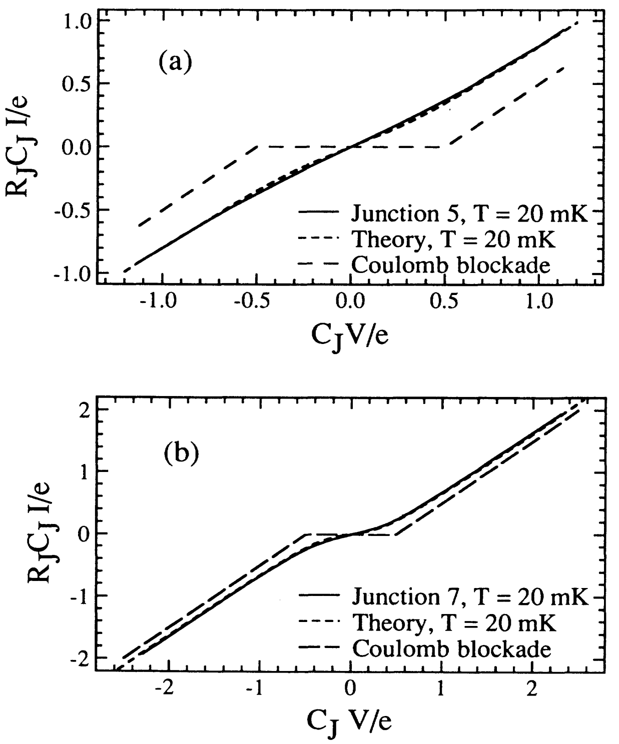


FIG. 14. Comparison of theory of Devoret *et al.* (Ref. 20) with experimental I-V characteristics; (a) is junction 5 and (b) is junction 7. Both the data, shown by solid lines, and the theory curves, shown by dashed lines, are at $T=20$ mK. The voltage-biased Coulomb blockade is also shown.

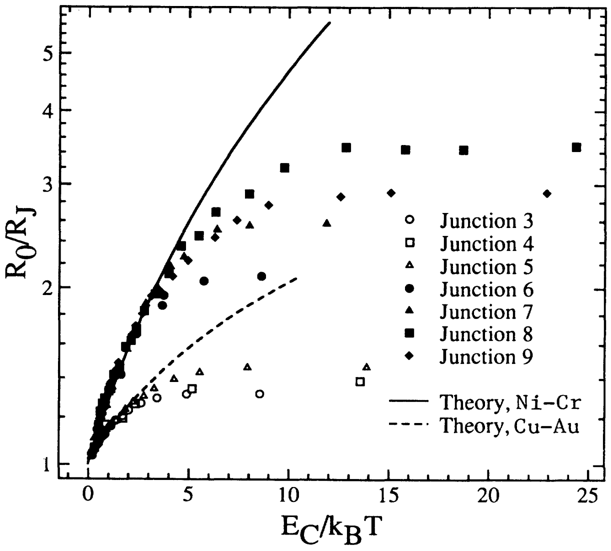


FIG. 16. Zero-bias resistance for seven junctions compared with theory of Devoret *et al.* (Ref. 20), with circuit parameters for Cu-Au leads (dashed line, open symbols) and Ni-Cr leads (solid line, solid symbols).

tions of Fig. 14 do not give noticeable differences in the curves. Larger variations in the value of C_L result in changes in the high-bias offset of these curves, with little change in the low-current bias region. The agreement between the calculated and measured curves in the high-bias region implies that our estimated value of C_L is not unreasonable.

The temperature dependence of the zero-bias resistance for both types of lead material is compared with the theory in Fig. 16; the data have been plotted as a function of $E_Q/k_BT=e^2/(2C_Jk_BT)$. A detailed comparison requires more precisely known values of the junction capacitance C_J than those given in Table I; to make the comparison of Fig. 16, we have adjusted the values of C_J slightly to give agreement at high temperatures. Values of C_J for the Cu-Au leads were adjusted upward and those for the Ni-Cr leads downward, within the estimated uncertainties from the I-V measurements. In general, the comparison at high temperatures is quite good, but the experimental results flatten off as the temperature is lowered while the theoretical predictions continue to increase. As discussed earlier, several attempts were made to explain the zero-bias behavior in terms of spurious noise and heating, none of which were shown to have any effect on the measurements. As it stands, therefore, there is no good explanation for this discrepancy.

C. Quantum fluctuations in the junctions

There is an interesting trend observed in the experimental ZBR data: the junctions with smaller R_J are seen to flatten off at smaller values of R_0/R_J, where R_0 is the zero-bias differential resistance, than those with larger R_J. This trend could be due to quantum fluctuations in the junctions themselves. This question is discussed in a paper by Brown and Šimánek,[23] who calculate the ZBR of a single small junction as a function of temperature and junction resistance. However, these authors do not include the effect of the environment, and treat the junction resistance as an Ohmic shunt. An experiment investigating the effects of finite junction resistance in devices involving multiple series-connected small junctions[6] appears to confirm the calculation. In Fig. 17 we show the predicted ZBR of a single junction versus R_J, normalized to the resistance quantum $R_K/2=h/2e^2$, at the two temperatures $E_Q/k_BT=10$ and $E_Q/k_BT=7.5$. There is clearly a strong dependence of the ZBR on tunneling resistance, and the general trends are reproduced in the experimental data obtained with Ni-Cr shunts. Thus, it seems entirely possible that a complete theory for the ZBR requires one to combine the effects of both the finite tunnel-junction resistance and the environment.

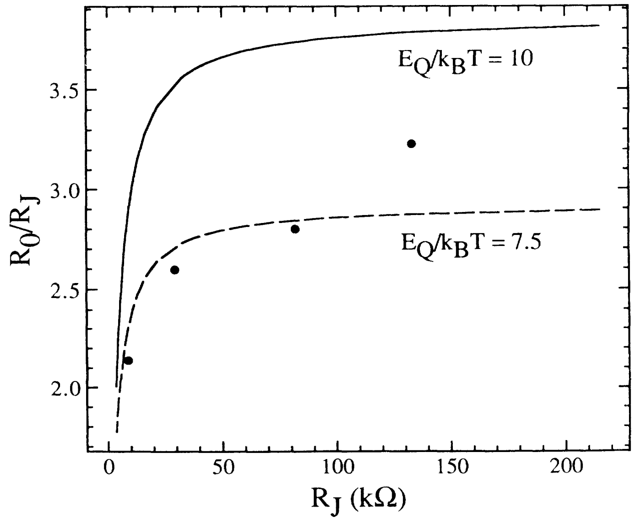


FIG. 17. Zero-bias resistance R_0/R_J calculated from theory of Brown and Šimánek for $E_Q/k_BT=10$ and 7.5. Also shown are zero-bias resistance data for the junctions with Ni-Cr leads at $E_Q/k_BT=10$.

V. SUMMARY

We have described the results of measurements on a series of single, low-capacitance tunnel junctions connected to thin-film resistors with resistances of either about 2 or 30 kΩ/μm. The I-V characteristics of the junctions are strongly influenced by the impedance presented by these leads; in particular, the Coulomb blockade is much more clearly visible for the high-resistance leads than for the low-resistance leads. The zero-bias resistance R_0 of the junctions increases as the temperature is lowered, flattening off at the lowest temperatures. We have demonstrated that this flattening could not be explained by Joule heating in the junctions and the leads, nor by the influence of extraneous noise sources.

In an attempt to obtain a heuristic understanding of the effects of the leads, we developed a simple model based on the quantum Langevin equation. In this picture, Nyquist noise in the leads produces fluctuations in the charge on the junction, smearing out the current-voltage characteristics even at $T=0$, where zero-point fluctuations remain. This model provides a qualitatively correct description of our data. In particular, the Coulomb blockade is predicted to be much less smeared out in the case of high-resistance leads than in the case of low-resistance leads, although the quantitative agreement between theory and experiment is much better for the former than for the latter. The model also predicts the flattening of R_0 as $T\rightarrow 0$, although the predicted values of R_0/R_J are too high in both cases. However, this discrepancy may well be explained by our inexact knowledge of the values of the lead resistances.

We also compared our data with the phase-correlation theory, which treats the entire circuit as a single quantum system, and includes a more accurate, transmission-line model for the leads. This model yields a better prediction of the I-V characteristics for both high- and low-resistance leads, although an examination of the differential resistance shows that the predicted zero-bias resistance at the lowest temperatures is roughly 50% higher than the data in each case. In contrast to the data and the quantum Langevin model, the phase-correlation theory predicts that R_0/R_J continues to increase as the temperature is lowered. This disagreement suggests that quantum fluctuations in the junction, which are not included in the phase-correlation theory, may indeed play a role at lower temperatures. Further theoretical work will be required to test this hypothesis quantitatively.

ACKNOWLEDGMENTS

We thank M. Devoret, U. Eckern, D. Esteve, H. Grabert, H. Mooij, Y. Nazarov, and G. Schön for helpful discussions. We thank the staff of the Microfabrication Facility of the Electrical Engineering and Computer Science Department for their expert assistance, and T. Van Duzer for the use of his scanning electron microscope. One of us (A.N.C.) acknowledges partial support by IBM. This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

[1] See, e.g., D. V. Averin and K. K. Likharev, J. Low Temp. Phys. **62**, 345 (1986).

[2] T. A. Fulton and G. J. Dolan, Phys. Rev. Lett. **59**, 109 (1987).

[3] L. S. Kuzmin, P. Delsing, T. Claeson, and K. K. Likharev, Phys. Rev. Lett. **62**, 2539 (1989).

[4] P. Delsing, K. K. Likharev, L. S. Kuzmin, and T. Claeson, Phys. Rev. Lett. **63**, 1180 (1989).

[5] P. Delsing, K. K. Likharev, L. S. Kuzmin, and T. Claeson, Phys. Rev. Lett. **63**, 1861 (1989).

[6] L. J. Geerligs, V. F. Anderegg, C. A. van der Jeugd, J. Romijn, and J. E. Mooij, Europhys. Lett. **10**, 79 (1989).

[7] Yu. V. Nazarov, Zh. Eksp. Teor. Fiz. **95**, 975 (1989).

[8] M. Devoret, D. Esteve, H. Grabert, G.-L. Ingold, H. Pothier, and C. Urbina, Phys. Rev. Lett. **64**, 1824 (1990).

[9] S. Girvin, L. I. Glazman, M. Jonson, D. R. Penn, and M. D. Stiles, Phys. Rev. Lett. **64**, 3183 (1990).

[10] A. N. Cleland, J. M. Schmidt, and J. Clarke, Phys. Rev. Lett. **64**, 1565 (1990).

[11] D. V. Averin and K. K. Likharev, J. Low Temp. Phys. **59**, 347 (1985).

[12] A. N. Cleland, Ph.D. thesis, University of California at Berkeley, 1991.

[13] J. M. Jaycox and M. B. Ketchen, IEEE Trans. Mag. **MAG-17**, 400 (1981).

[14] F. C. Wellstood, C. Urbina, and J. Clarke, Appl. Phys. Lett. **54**, 2599 (1989).

[15] G. J. Dolan, Appl. Phys. Lett. **31**, 337 (1977).

[16] R. E. Howard, E. L. Hu, L. D. Jackel, L. A. Fetter, and R. H. Bosworth, Appl. Phys. Lett. **35**, 879 (1979).

[17] J. M. Martinis and R. L. Kautz, Phys. Rev. Lett. **63**, 1507 (1989).

[18] J. M. Martinis, M. H. Devoret, and J. Clarke, Phys. Rev. B **35**, 4682 (1987).

[19] M. Büttiker and R. Landauer, IBM J. Res. Dev. **30**, 451 (1986).

[20] J. B. Johnson, Phys. Rev. **32**, 97 (1928); H. Nyquist, *ibid.* **32**, 110 (1928).

[21] U. Geigenmüller and G. Schön, Physica (Amsterdam) **152B**, 186 (1988).

[22] G.-L. Ingold and H. Grabert, Europhys. Lett. **14**, 371 (1991).

[23] R. Brown and E. Šimánek, Phys. Rev. B **34**, 2957 (1986).

Charge Fluctuations in Small-Capacitance Junctions

A. N. Cleland, J. M. Schmidt, and John Clarke

Department of Physics, University of California, Berkeley, California 94720
and Materials and Chemical Sciences Division, Lawrence Berkeley Laboratories, Berkeley, California 94720

(Received 18 December 1989)

The current-voltage characteristics of submicron normal-metal tunnel junctions at millikelvin temperatures are observed to exhibit a sharp Coulomb blockade with high-resistance thin-film leads, but to be heavily smeared for low-resistance leads. As the temperature is lowered, the zero-bias differential resistance tends asymptotically to a limit that is greater for junctions with high-resistance leads. Both observations are explained in terms of a model in which quantum fluctuations in the external circuit enhance the low-temperature tunneling rate. The predictions are in reasonable agreement with the data.

PACS numbers: 74.50.+r, 05.30.−d, 73.40.Gk

It is predicted theoretically[1-4] and well established experimentally[5-14] that submicron tunneling junctions at low temperatures T can, under appropriate conditions, exhibit charging effects due to the discreteness of the electronic charge. In particular, when the charging energy $E_C = e^2/2C$ associated with the tunneling of a single electron across a capacitance C becomes large compared with k_BT, one may observe suppression of the tunnel current I at voltages $V < e/2C$. As a result, the I-V characteristic at higher voltages is offset by the Coulomb gap, $e/2C$. However, observation of these effects depends strongly on the nature of the environment coupled to the junction. For example, Delsing *et al.*[12] varied the environment by studying two kinds of circuits. In the first, metallic leads were coupled directly to the junction while in the second, linear arrays of submicron junctions were placed in each lead to the junction under study. The Coulomb gap in the first circuit was observed to be greatly smeared out, while in the second, it was much sharper. Geerligs *et al.*[14] studied single junctions and linear arrays of junctions. They found sharp Coulomb gaps in arrays containing a minimum of two junctions, provided that the resistance of the junctions R was much greater than $\hbar/e^2$. For the single junctions the gap was extremely smeared, but at high currents the expected voltage offset was observed. Thus, although it is clear that the environment has a significant effect on the Coulomb blockade, the nature of the effect has remained unexplained until now.

In this Letter, we report measurements on small junctions with thin-film leads of high and low resistance. The Coulomb gap is well defined in the former case for junctions with $R \gg \hbar/e^2$, but very smeared out in the latter, for all values of R. In all junctions studied, the zero-bias differential resistance increases and then flattens out as the temperature is lowered. We propose a simple model in which the Nyquist noise from the external circuit produces charge fluctuations across the junction capacitance. These fluctuations, in turn, enhance the low-voltage tunneling rate, smearing the Coulomb gap. In the zero-temperature limit the noise arises from quantum fluctuations, and our model yields predictions for the I-V characteristic and the zero-bias resistance that are in reasonable agreement with our data at the lowest temperatures.

Small Al-Al-oxide-Al tunnel junctions with areas of typically 0.04 μm^2 were fabricated using standard electron-beam lithography and angled evaporations.

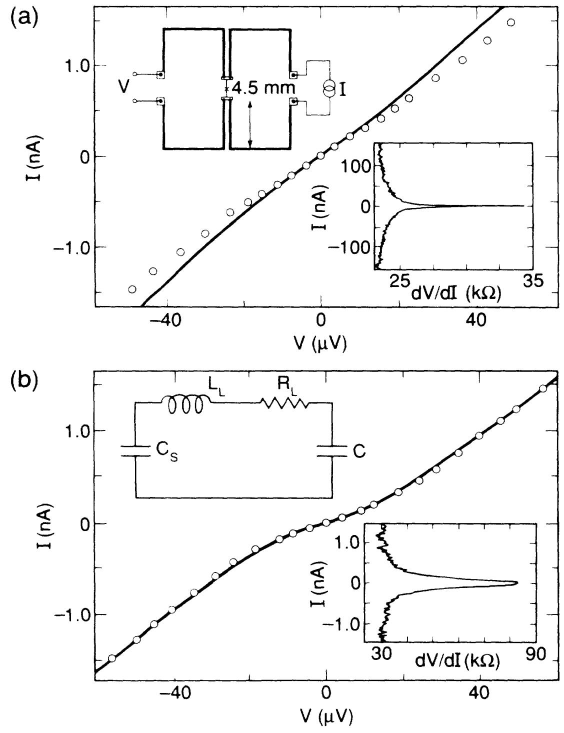


FIG. 1. I-V characteristics (solid lines) for two junctions at 20 mK with (a) $C = 4 \pm 1$ fF and $R = 23$ kΩ and (b) $C = 5 \pm 1$ fF and $R = 28$ kΩ. Dots represent predictions of theory. Inset in each figure is dV/dI vs I; note different current scales. Also inset in (a) is the configuration of junction and leads (not to scale) and in (b) its simplified representation.

Thin-film leads were connected to the junctions in the arrangement shown inset in Fig. 1; for one set of junctions the leads were made of CuAu alloy (25 wt% Cu) with a sheet resistance of about 4 Ω per square, and for the other set the leads were made of NiCr alloy (80 wt% Ni), with a sheet resistance of about 60 Ω per square. Each of these four leads was 2 μm wide and had a total length of 12 mm. The Al lead from each side of the junction to the resistive leads was 30 μm long and 0.2 μm wide. The junctions were mounted on a dilution refrigerator, and all electrical leads to the junctions were filtered above 10 kHz. The Al was driven into the normal state with an external magnetic field.

In Fig. 1 we show the *I-V* characteristics for two junctions, with low- and high-resistance leads, respectively. The tunneling resistance and capacitance of both junctions are within 25% of one another, and both characteristics were taken at the same temperature. There is a striking difference in the low-current behavior of the two junctions; the low-resistance leads clearly give rise to a very smeared Coulomb gap, while the high-resistance leads give a much sharper characteristic. This distinction is emphasized in the plots of the dynamic resistance shown as insets in Fig. 1. Similar differences in behavior were observed in all the junctions we have studied; in Fig. 2, we plot the temperature dependence of the zero-bias dynamic resistance R_0 for five junctions. We see that the low-temperature values of R_0/R are higher for junctions with high-resistance leads than for those with low-resistance leads. Furthermore, for a given lead resistance, the asymptotic value of R_0/R increases with R, in qualitative agreement with the findings of Geerligs *et al.*[14]

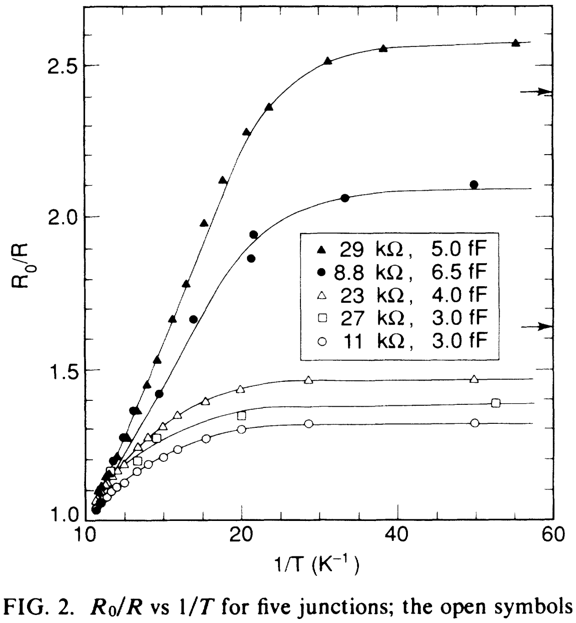


FIG. 2. R_0/R vs $1/T$ for five junctions; the open symbols are for low lead resistance and the solid symbols for high lead resistance. Arrows indicate the predicted values of R_0/R.

We have considered the possibility that the flattening out of R_0/R as the temperature is lowered is due to self-heating or to extraneous noise. Our estimates of hot-electron effects[15] in the leads and in the junction imply that heating is negligible for $V \lesssim e/2C$, even at the lowest temperatures. We have also removed the magnetic field on a low-resistance junction and measured the reduction of the superconducting energy gap due to heating at high currents. These results also imply that heating is negligible in the experiments reported here. We have tested the effects of adding and removing radio-frequency and microwave filters at room temperature, 4.2 K, and 20 mK, and found no effects on the *I-V* characteristics. These observations suggest that external noise is not responsible for the flattening of R_0/R.

To interpret our results, we consider a model circuit in which the junction is connected via leads with inductance L_L and resistance R_L to a line with a stray capacitance $C_s \gg C$. Thus, L_L and R_L represent the combined inductance and resistance of all four leads connected to the junction. The resistance R_L produces a Nyquist voltage noise V_N with a spectral density

$$S_V(\omega) = (\hbar\omega R_L/\pi)\coth(\hbar\omega/2k_BT)\,.$$

Assuming $C_s^{-1} \ll C^{-1}$ and defining $\omega_{LC}^2 \equiv 1/L_LC$, $\omega_{RC} \equiv 1/R_LC$, we write the quantum Langevin equation for the Fourier transform $q(\omega)$ of the fluctuations in charge $q(t)$ on the junction,

$$q(\omega)/C + i\omega R_L q(\omega) - \omega^2 L_L q(\omega) + V_N(\omega) = 0\,. \quad (1)$$

Solving for $q(\omega)$, we obtain the following mean-square charge fluctuation:

$$\langle q^2(t)\rangle = \int_0^\infty \frac{C^2}{(1-\omega^2/\omega_{LC}^2)^2 + (\omega/\omega_{RC})^2} S_V(\omega)d\omega\,. \quad (2)$$

We remark that, in general, the quantum Langevin equation is likely to be accurate only when any nonlinearity associated with tunneling is negligible so that the energy levels of the circuit are equally spaced, or when the damping is sufficiently large that the quantum levels are smeared out.[16] We believe the latter case to apply to the experiments described here.

In the classical limit of large T, Eq. (2) yields the result of the equipartition theory, $\langle q^2\rangle/2C = k_BT/2$. In the quantum regime of small T, for $a \equiv \omega_{LC}/\omega_{RC} = R_L(C/L_L)^{1/2} < 2$, we find

$$\frac{\langle q^2\rangle_Q}{2C} = \frac{\hbar\omega_{LC}}{4\pi}\frac{1}{(4-a^2)^{1/2}} \times \left[\frac{\pi}{2} - \tan^{-1}\left[\frac{a^2-2}{a(4-a^2)^{1/2}}\right]\right]. \quad (3)$$

In limit where R_L is very small, $a \ll 2$, Eq. (3) reduces to $\langle q^2\rangle/2C = \hbar\omega_{LC}/4$ as expected for a simple harmonic oscillator. On the other hand, in the limit $a > 2$ of interest here, we obtain

$$\frac{\langle q^2\rangle_Q}{2C}=\frac{\hbar\omega_{LC}}{2\pi}\frac{1}{(a^2-4)^{1/2}}\ln\left[\frac{a^2-2+a(a^2-4)^{1/2}}{a^2-2-a(a^2-4)^{1/2}}\right]. \tag{4}$$

In the limit where R_L is very large, $a \gg 2$, Eq. (4) becomes $\langle q^2\rangle/2C=(\hbar\omega_{RC}/\pi)\ln a$. Note that this expression depends on L_L only logarithmically.

A junction with charge Q is driven by these fluctuations to charge $Q+q$ with probability $P(q)$; $P(q)$ is Gaussian distributed with width $\langle q^2\rangle$. These fluctuations in the charge will modify the electron tunneling rate[4] $\Gamma^{\pm}(Q)$ for Q going to $Q\pm e$,

$$\Gamma^{\pm}(Q)=-\frac{e/2\pm Q}{eRC}\left[\exp\left(\frac{\Delta E^{\pm}}{k_BT}\right)-1\right]^{-1}, \tag{5}$$

where $\Delta E^{\pm}=(\pm 2eQ+e^2)/2C$ is the resultant energy change. We replace Eq. (5) with the convolution

$$\langle\Gamma^{\pm}(Q)\rangle=\int_{-\infty}^{\infty}P(q)\Gamma^{\pm}(Q+q)dq\,. \tag{6}$$

Although one can obtain expressions for the temperature-dependent tunneling rate, in this Letter we confine our attention to the limit $T=0$ where

$$\langle\Gamma^{\pm}(Q)\rangle=-\frac{e/2\pm Q}{2eRC}\operatorname{erfc}\left[\frac{e/2\pm Q}{(2\langle q^2\rangle)^{1/2}}\right]+\frac{1}{eRC}\left[\frac{\langle q^2\rangle}{2\pi}\right]^{1/2}\exp\left[-\frac{(e/2\pm Q)^2}{2\langle q^2\rangle}\right]. \tag{7}$$

Thus, even at $T=0$, for large enough fluctuations the Coulomb blockade is heavily smeared out at low Q. On the other hand, for large $Q>0$, we find the simple results $\Gamma^+=0$ and $\Gamma^-=(Q-e/2)/eRC$, at any temperature. The Coulomb blockade is still visible as a voltage shift $e/2C$ for large voltages. This result is consistent with experimental observations.[14] To enable us to make quantitative comparisons with our data, we have carried out Monte Carlo simulations of the charging sequence of a current-biased junction, using Eq. (7) for the tunneling rate. For a given bias current we compute the voltage across the junction as a function of time, and then calculate the average voltage to obtain the *I-V* characteristic.

The comparison of our data with the predictions obtained from the model circuit is not entirely straightforward. The major difficulty concerns the stray capacitance between the thin-film leads and the nearest ground plane, which is roughly 10 mm away. We estimate this distributed capacitance to be between 1 and 5 fF/(mm length of the leads). We have performed numerical calculations indicating that the direct capacitive coupling between the resistive leads on opposite sides of the junction is an order of magnitude smaller than this figure.[17] In an attempt to develop some feeling for the effect of the capacitance to ground, we performed subsidiary experiments on two junctions where all but the 4.5 mm of the leads closest to the junction were coated with indium. The presence of the indium has no observable effect on the *I-V* characteristics. Thus, we conclude that, at most, only the first 4.5 mm of each lead contribute to the loading of the junction. We estimate the inductance of each of these leads (L_L in our model) to be 5 nH, while the resistances R_L are 8 and 130 kΩ for the low- and high-resistance cases, respectively. The corresponding distributed capacitance to ground, C_L, is between 5 and 25 fF. This capacitance reduces the impedance of the leads at frequencies above $1/2\pi R_LC_L$; at $10/2\pi R_LC_L$ the resistance is about one-half of that at low frequencies. Above this frequency, which is a few GHz for the high-resistance leads, the impedance scales as $(R_L/\omega C_L)^{1/2}$ and is independent of the length of the leads. However, the high-resistance leads still present a significantly higher impedance than do the low-resistance leads. Thus, to test our model we have ignored the parasitic capacitance, and used values of L_L and R_L corresponding to the first 4.5 mm of the leads. Of course, the relevant length may be shorter, but we have no means of estimating it. We have used values of the junction capacitances C estimated from the voltage offset at high currents. The voltage offsets give capacitances somewhat larger than expected[6] from the estimated geometrical area of the junctions, but not unreasonably so; it is possible that some stray capacitance contributes to this capacitance.

The calculated *I-V* characteristics are shown as points in Figs. 1(a) and 1(b). The agreement is good for the high lead resistance, and somewhat less good for the low; however, the trends are well predicted. In Fig. 2, we show the predicted zero-temperature values of R_0/R for the high and low lead resistances. In each case, one should compare the prediction with the results from the highest junction resistance, for which the effects of dissipation in the junction should be negligible.[3] The increase in R_0/R when one increases the lead resistance for two similar junctions is clearly demonstrated by the theory. Thus, although our model calculations do not predict the observed *I-V* characteristics and zero-bias resistances precisely, given the uncertainties in the impedance loading the junction, we feel that the agreement between the predictions and the data is quite satisfactory. In particular, the results show that the flattening of R_0/R as the temperature is lowered arises from the zero-point fluctuations in the external circuit.

Our experiments support the results of Delsing *et al.*[12] and Geerligs *et al.*[14] in demonstrating the critical importance of the high-frequency properties of the circuit loading small junctions. If the external resistance is too small, our results show that noise generated in this resistance results in charge fluctuations on the junction that smear out the Coulomb gap; at low temperatures, this noise arises from quantum fluctuations. Because of the inherent stray capacitance, it is important to insert the high resistance as close as possible to the junction. We note that one could alternatively introduce high inductances in the leads close to the junctions, in which case Eq. (3) would apply. Finally, we expect the fluctuation effects described here to influence observations not only of Coulomb blockade but also of time-correlated processes involving single electrons[13] and pairs.[2]

A more detailed description of this work including calculations at nonzero temperatures will be presented elsewhere.

We thank S. Chakravarty, U. Eckern, H. Grabert, J. E. Mooij, M. Tinkham, and G. Schön for helpful discussions. A.N.C. acknowledges support from IBM. This work was supported by the Director, Office of Energy Research, Office of Basic Energy Sciences, Materials Sciences Division of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

[1] K. K. Likharev and A. B. Zorin, J. Low Temp. Phys. **59**, 347 (1985).

[2] D. V. Averin and K. K. Likharev, J. Low Temp. Phys. **62**, 345 (1986).

[3] R. Brown and E. Simanek, Phys. Rev. B **34**, 2957 (1986).

[4] U. Geigenmüller and G. Schön, Physica (Amsterdam) **152B**, 186 (1988).

[5] J. Lambe and R. C. Jaklevic, Phys. Rev. Lett. **22**, 1371 (1969).

[6] T. A. Fulton and G. J. Dolan, Phys. Rev. Lett. **59**, 109 (1987).

[7] M. Iansiti, A. T. Johnson, C. J. Lobb, and M. Tinkham, Phys. Rev. Lett. **60**, 2414 (1988).

[8] P. J. M. van Bentum, R. T. M. Smokers, and H. van Kempen, Phys. Rev. Lett. **60**, 2543 (1988).

[9] M. Iansiti, M. Tinkham, A. T. Johnson, Walter F. Smith, and C. J. Lobb, Phys. Rev. B **39**, 6465 (1989).

[10] L. S. Kuzmin, P. Delsing, T. Claeson, and K. K. Likharev, Phys. Rev. Lett. **62**, 2539 (1989).

[11] R. Wilkins, E. Ben-Jacob, and R. C. Jaklevic, Phys. Rev. Lett. **63**, 801 (1989).

[12] P. Delsing, K. K. Likharev, L. S. Kuzmin, and T. Claeson, Phys. Rev. Lett. **63**, 1180 (1989).

[13] P. Delsing, K. K. Likharev, L. S. Kuzmin, and T. Claeson, Phys. Rev. Lett. **63**, 1861 (1989).

[14] L. J. Geerligs, V. F. Anderegg, C. A. van der Jeugd, J. Romijn, and J. E. Mooij, Europhys. Lett. **10**, 79 (1989).

[15] F. C. Wellstood, C. Urbina, and J. Clarke, Appl. Phys. Lett. **54**, 2599 (1989).

[16] U. Eckern and W. Lehr, Jpn. J. Appl. Phys. **26**, 1399 (1987).

[17] A. N. Cleland and J. Clarke (unpublished).

NATO ASI Series

Advanced Science Institutes Series

A series presenting the results of activities sponsored by the NATO Science Committee, which aims at the dissemination of advanced scientific and technological knowledge, with a view to strengthening links between scientific communities.

The series is published by an international board of publishers in conjunction with the NATO Scientific Affairs Division

A	**Life Sciences**	Plenum Publishing Corporation
B	**Physics**	New York and London
C	**Mathematical and Physical Sciences**	Kluwer Academic Publishers
D	**Behavioral and Social Sciences**	Dordrecht, Boston, and London
E	**Applied Sciences**	
F	**Computer and Systems Sciences**	Springer-Verlag
G	**Ecological Sciences**	Berlin, Heidelberg, New York, London,
H	**Cell Biology**	Paris, Tokyo, Hong Kong, and Barcelona
I	**Global Environmental Change**	

Recent Volumes in this Series

Volume 287—Coherence Phenomena in Atoms and Molecules in Laser Fields
edited by André D. Bandrauk and Stephen C. Wallace

Volume 288—Intersubband Transitions in Quantum Wells
edited by Emmanuel Rosencher, Børge Vinter, and Barry Levine

Volume 289—Nuclear Shapes and Nuclear Structure at Low Excitation Energies
edited by Michel Vergnes, Jocelyne Sauvage, Paul-Henri Heenen, and Hong Tuan Doung

Volume 290—Phase Transitions in Liquid Crystals
edited by S. Martellucci and A. N. Chester

Volume 291—Proton Transfer in Hydrogen-Bonded Systems
edited by T. Bountis

Volume 292—Microscopic Simulations of Complex Hydrodynamic Phenomena
edited by Michel Mareschal and Brad L. Holian

Volume 293—Methods in Computational Molecular Physics
edited by Stephen Wilson and Geerd H. F. Diercksen

Volume 294—Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures
edited by Hermann Grabert and Michel H. Devoret

Series B: Physics

EECS 217C H.O. #1 1,30

Single Charge Tunneling

Coulomb Blockade Phenomena In Nanostructures

Edited by

Hermann Grabert

Universität Essen
Essen, Germany

and

Michel H. Devoret

Centre d'Etudes de Saclay
Gif-sur-Yvette, France

Plenum Press
New York and London
Published in cooperation with NATO Scientific Affairs Division

Chapter 1

Introduction to Single Charge Tunneling

M. H. DEVORET

Groupe Quantronique, Service de Physique de l'Etat Condensé,
Centre d'Etudes de Saclay, 91191 Gif-sur-Yvette, France

and

H. GRABERT

Fachbereich Physik, Universität–GH Essen, 4300 Essen, Germany

1. Basic ingredients of single charge tunneling phenomena

Consider a charge transport experiment in which a voltage difference is applied to two electrodes (a "source" and a "drain", see Fig. 1) separated by an insulating gap. In the middle of the gap lies a third metallic electrode, which we call an "island" since it is surrounded by an insulator. To travel from the source to the drain the electrons must go through the island. We assume that the conduction of electrons through the insulating gaps between the source and the island and between the island and the drain occurs by quantum tunneling. This process is so fast that we can consider that the electrons are traversing the insulating gaps one at a time. Nevertheless, the successive tunnel events across a particular junction are uncorrelated and constitute a Poisson process. The key point is that, during its journey from the source to the drain, the electron necessarily makes the charge of the island vary by e. This is a tiny amount of charge if we consider

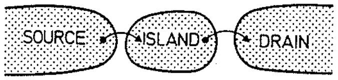


Figure 1. The quantum tunneling of electrons between a "source" and a "drain" electrode through an intermediate "island" electrode can be blocked if the electrostatic energy of a single excess electron on the island is large compared with the energy of thermal fluctuations.

Single Charge Tunneling, Edited by H. Grabert and
M.H. Devoret, Plenum Press, New York, 1992

ordinary electronic devices: each charge packet in a charge coupled device (CCD), for example, is composed of about 10^6 electrons [1]. However, if the island is small enough, the variation of the island potential due to the presence of an excess electron can be large enough to react back on the tunneling probabilities. The existence of such a feedback effect was proposed several decades ago [2]–[6].

At that time, the effect could only be observed in granular metallic materials. It was realized that the hopping of electrons from grain to grain could be inhibited at small voltages if the electrostatic energy $e^2/2C$ of a *single* excess electron on a grain of capacitance C was much greater than the electron thermal energy k_BT. The interpretation of these pioneering experiments, in which single electron effects and random media properties interplay, was complicated by the limited control over the structure of the sample. Nowadays, with modern nanofabrication techniques, it is possible to design metallic islands of known geometry separated by well controlled tunnel barriers [7]. Bias leads can impose a voltage across the whole set of barriers and the charge distribution of the islands can be acted upon by leads connected to small gate capacitors. In these nanoscale tunnel junction systems, a fully developed "Coulomb gap" arises [8, 9] which can be exploited to control a current by means of a single charge on a gate [10] and to transfer single charges from one island to another in a controlled way [11, 12]. The mechanism underlying these systems exploits the feedback effect of the Coulomb interaction energy of a charge with the other charges in the system. More generally, this feedback effect characterizes what we call *single charge tunneling* (SCT) phenomena. They can take place not only in normal metal and semiconductor junction systems in which the individual charge carriers are electrons and holes but also in superconducting systems in which the charge carriers are Cooper pairs.

The introductory remarks above indicate the basic requirements for single charge tunneling phenomena to occur in nanoscale junction systems. Leaving aside for the moment the special case of a single tunnel junction which will be discussed in the following section, these conditions are as follows. Firstly, the system must have metallic islands that are connected to other metallic regions only via tunnel barriers with a tunneling resistance R_T that exceeds the resistance quantum $R_K = h/e^2 \simeq 25.8\ \mathrm{k}\Omega$, i.e.,

$$R_T \gg R_K. \tag{1}$$

The tunneling resistance is a phenomenological quantity which is defined in the situation where a fixed voltage difference V is imposed to the two electrodes on either side of the tunnel barrier. The tunneling rate Γ of an electron through the barrier is then proportional to $V : \Gamma = V/eR_T$. The tunneling resistance can be expressed in terms of the microscopic quantity $\mathcal{T}$, which is the barrier transmission coefficient at the Fermi energy: $R_T^{-1} = 4\pi N\mathcal{T}R_K^{-1}$, where N is the number of independent electron channels through the barrier. Condition (1) is obtained by requiring that for an excess charge on the island the energy uncertainty associated with the lifetime due to tunneling $\tau_r = R_TC$ is much smaller than the Coulomb energy $E_c = e^2/2C$. Essentially, condition (1) ensures that the wave function of an excess electron or Cooper pair on an island is localized there. It is generally believed that in systems with tunneling resistances that are small on the scale provided by R_K charging effects will be suppressed since delocalized states in which electrons flow through an island without charging it are available for charge transport

[13, 14], although the exact circumstances are not precisely known at the time of this writing.

Secondly, the islands must be small enough and the temperature low enough that the energy E_c required to add a charge carrier to an island far exceeds the available energy of thermal fluctuations, i.e.,

$$E_c \gg k_BT. \tag{2}$$

In practice, only islands having capacitances not much below a fF can be reliably designed, thus imposing experiments done at a few tens of mK, now routinely attainable with a dilution refrigerator. Conditions (1) and (2) ensure that the transport of charge from island to island is governed by the Coulomb charging energy. With the use of externally applied gate voltages, the charging energy of the various islands can be sequentially lowered or increased in order to manipulate single charge carriers (see Chap. 3).

At present, two main types of systems where single charge tunneling phenomena arise are being explored. The majority of experiments carried out so far have used metallic (mostly Al) thin film systems in which the lithographically patterned islands are separated by oxide layer tunnel barriers (see Chaps. 3, 4, 7 and 8). In this case, three-dimensional electron gases confined to small regions are coupled by the tunnel effect through the oxide layer which is only about 10Å thick. The capacitances of the tunnel junctions thus make up for the most part of the capacitance of an island. These systems also allow one to explore charging effects involving Cooper pairs since the metals used to fabricate the circuits are superconductors at the temperatures required to satisfy (2). As a matter of fact, one must apply a magnetic field to keep the metals in the normal state.

Single charge tunneling effects also occur when the two-dimensional electron gas of a GaAs/AlGaAs heterostructure is confined to small islands by means of Schottky gates. In that case, covered in Chap. 5, electrons can tunnel through the depleted region between islands and the tunnel resistances of the "junctions" can be tuned by changing the confining gate voltages. Further, the islands may be reduced to quantum dots with a discrete energy spectrum for single electron wave functions. This situation yields interesting phenomena combining charging effects and resonant tunneling. It is remarkable that even for such dots containing only several tens of electrons, charging effects can usually be described through an effective island capacitance in close relationship with its geometry. The fact that the electrostatic capacitance remains a useful concept for such small islands calls for an explanation: The applicability of the notion of an island capacitance depends on the ratio between the screening length and the size of the island. This ratio, which is of order 10^{-4} in metallic systems, is still comfortably smaller than 1 in the case of quantum dots.

Apart from granular films and lithographically patterned systems covered in this book, single charge tunneling phenomena are observed in a number of other cases like small metal particles embedded in an oxide layer or disordered quantum wires. Also, one of the tunneling barriers may be formed by a scanning tunneling microscope. Detailed lists of references to these studies can be found in the review articles by Averin and Likharev [8] and Schön and Zaikin [9], and also in the Special Issue of Zeitschrift für Physik B – Condensed Matter on SCT [15].

2. Single current biased junction

In the preceding section we have stressed that the basic system in which single charge tunneling phenomena occur is a metallic island connected to electron reservoirs through at least *two* tunnel barriers. What would happen with only *one* small capacitance tunnel junction? Physicists are attracted to simple systems. Historically, it is this question which started the field a few years ago. Several new effects due to the quantization of charge were predicted to arise in an ultrasmall tunnel junction, both in the superconducting and the normal state [16]–[19]. Likharev and coworkers [17, 19] gave a major thrust to this new area of low temperature physics by making detailed predictions of Coulomb blockade phenomena in a single junction and by proposing various applications of the new effects.

The theory of Likharev and coworkers considers a tunnel junction which is biased by a current I and whose voltage V is measured by a very high impedance voltmeter. The junction is characterized by two parameters: its capacitance C and tunnel resistance R_T. The state of the junction is described by two degrees of freedom whose different nature is crucial. The first degree of freedom is the charge Q on the junction capacitance. It is a continuous variable since it describes the bodily displacement of the electron density in the electrodes with respect to the positive ionic background. In fact, Q can be an arbitrarily small fraction of the charge quantum. The second degree of freedom is the discrete number n of electrons (or Cooper pairs if the electrodes are in the superconducting state) which have passed through the tunnel barrier. The key hypothesis in the theory is that Q and n are classical variables with a well defined value at every instant of time t. Charge conservation is imposed by the relation $\dot{Q}(t) + e^*\dot{n}(t) = i(t)$, where $i(t)$ is the current flowing in the leads of the junction and e^* is the charge e (normal state) or $2e$ (superconducting state) of the carriers. Since the current bias is assumed to be ideal we have $i(t) = I$. During a tunneling event, the charge Q must thus discontinuously jump by the elementary charge e^*. The resulting change in electrostatic energy of the junction is

$$\Delta E = \frac{Q^2}{2C} - \frac{(Q-e^*)^2}{2C} = \frac{e^*(Q-e^*/2)}{C}. \tag{3}$$

At zero temperature, tunneling can only occur if ΔE is positive. This has two consequences. Firstly, the $I-V$ characteristic should have an $I=0$ branch

$$-\frac{e^*}{2C} < V < \frac{e^*}{2C} \quad \text{for} \quad I = 0 \tag{4}$$

where the particular value of V is determined by the history of the current in the junction leads: $CV = \int_{-\infty}^{t} i(t')\mathrm{d}t'$ modulo e^*. This is the Coulomb blockade for single junctions. Secondly, when a non-zero current is imposed through a junction in the normal state, the junction capacitor charge Q will increase linearly until the threshold charge $e/2$ is reached. Then, a tunneling event occurs, making Q jump to $-e/2$ and a new charging cycle starts again. This leads to single electron tunneling (SET) sawtooth oscillations of the junction voltage with the fundamental frequency

$$f_{\mathrm{SET}} = I/e. \tag{5}$$

By a similar kind of reasoning, one predicts for a superconducting junction (Josephson junction) the so-called Bloch oscillations with the frequency

$$f_{\mathrm{Bloch}} = I/2e. \tag{6}$$

The difference between the SET and Bloch oscillations is that in the normal state the charge tunnels irreversibly as Q goes beyond $e/2$ because it is accompanied by quasiparticle excitations whereas in the superconducting state the charge tunnels reversibly at $Q = e$ because Cooper pairs have no kinetic degrees of freedom.

This analysis rests on Q and n being classical variables. The classical nature of the variable n is solely determined by the properties of the junction. We can safely assume that condition (1), which is a statement about the tunnel barrier and which translates directly in terms of junction fabrication, is a sufficient condition. However, the classical nature of the variable Q depends on the junction electromagnetic environment and the original predictions concerning Coulomb blockade phenomena did not make very explicit statements about what the characteristics of this environment should be. In this theoretical void, two questions concerning the observability of Coulomb blockade and SET or Bloch oscillations arose:

Question A: The pads on the junction chip which are needed to make connections to the $I-V$ measuring apparatus have parasitic capacitances in the pF range. How should the junction environment be designed for these parasitic capacitances not to shunt the junction capacitance, which needs to be kept in the fF range to observe the charging effects?

Question B: Each mode in the environment is coupled to the charge Q and its zero point energy induces $i(t)$ and $Q(t)$ to fluctuate. How should the environment be designed for these quantum mechanical fluctuations not to affect the Coulomb blockade? In other words, how perfect does the current biasing need to be?

We will see below that these two questions are in fact closely related and that their answer can be obtained by a fully quantum mechanical analysis of the influence of the junction environment on the tunneling probability. Before presenting the results of this analysis, which is essential to the understanding of single charge tunneling phenomena, we have to discuss the various time scales of the problem, both the time scales pertaining to the junction itself and those pertaining to its environment.

The junction is characterized by three time scales. The two longer ones can be deduced from quantities we have already mentioned. The longest time scale is set by the tunneling resistance and the capacitance: $\tau_r = R_T C$. It is the reciprocal of the rate of tunnel events for a junction biased at the Coulomb voltage e/C. The intermediate time scale is the uncertainty time associated with the Coulomb energy $\tau_c = h/(e^2/C) = R_K C$. The shortest time scale is the tunneling time τ_t of the junction which is given by

$$\tau_t = \hbar \left(\frac{\partial \ln \mathcal{T}(E)}{\partial E} \right)_{E=E_F} \tag{7}$$

where, as previously, $\mathcal{T}(E)$ is the transmission probability through the tunnel barrier of an electron with energy E. This tunneling time, whose importance has been stressed by Büttiker and Landauer [20] (see also [21] and references therein), can be loosely described as the time spent by the tunneling electron under the barrier. In metallic

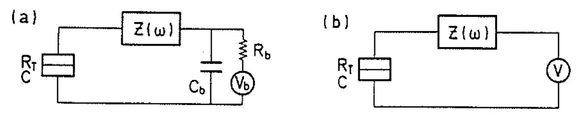


Figure 2. Lumped element model of the electromagnetic environment of a current-biased tunnel junction, which is represented by a double box symbol. The capacitance and the tunnel resistance of the junction are C and R_T, respectively. The impedance $Z(\omega)$ models the high frequency response of the environment which is dominated by the effect of the leads attached to the junction. The environmental low frequency response, which is dominated by the bias circuitry, is modelled in (a) by a resistance R_b, a capacitance C_b, and a voltage source V_b. In all practical cases $C_b \gg C$, and one can use the simplified model (b) in which the junction is biased by an effective voltage source V which is a function of the time-averaged current through the junction.

tunnel junctions it is of the order of 10^{-15}s. Here, it may be worthwhile pointing out that electron tunneling in a metallic junction is in fact a complex process, at least much more complex than what elementary textbooks might lead one to suppose: the electrons in the metallic electrodes travel as quasiparticles, i.e., bare electrons dressed by a positive cloud of charge. When a tunneling quasiparticle impinges on the insulating barrier it has to undress, leaving the positive charge cloud behind as it travels through the barrier. When this bare electron arrives in the other electrode it attracts a new cloud of positive charge and dresses again to form a quasiparticle. The characteristic time for the undressing and dressing processes is the inverse of the plasma frequency. These processes have to be taken into account in the computation of the effective tunneling time, which is the one of interest here. The tunneling rate of a quasiparticle will be quite different from the tunneling rate of a bare electron if the effective tunneling time is notably longer than the inverse of the plasma frequency [22].

Let us now discuss the time scales of the environment. We first have to indicate how one should model the junction electromagnetic environment which includes not only the $I-V$ characteristic measuring apparatus at high temperature but also the leads close to the junction. A priori, we need to consider the response of the environment up to the frequency τ_t^{-1}. Although this natural cut-off provided by the tunneling time is a frequency in the optical domain, the junction is small enough to be treated as a lumped element since its dimensions have to be of the order of 100 nm or less to ensure a capacitance in the fF range. The electromagnetic environment as seen from the location of the junction can thus be completely described in electrical engineering terms by the relationship between the complex amplitudes $v(\omega) = Q(\omega)/C$ and $i(\omega)$ at frequency ω of the voltage across the junction and the current in the first few hundred nanometers of its leads. Assuming that the environment is linear, we arrive for the electromagnetic environment as seen from the location of the junction at the general lumped element model of Fig. 2a. The bias circuitry, which includes the room temperature electronics, the filters, and the leads down to the pads on the junction chip, is modelled by a bias resistor $R_b \gg R_T$ in series with a voltage source V_b. There is also, in parallel with the resistor and the source, a capacitance C_b which models the parasitic capacitances in the bias circuitry. This three element model of the bias circuitry accounts for the low frequency response of the environment and is placed in series with a complex impedance $Z(\omega)$ which represents the impedance of the last few mm of leads on the junction chip. The impedance $Z(\omega)$ accounts for the high frequency response of the environment.

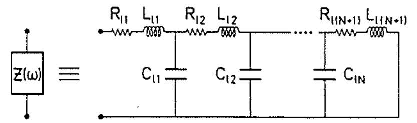


Figure 3. Resistive transmission line model for the impedance $Z(\omega)$.

This general model of the environment of the junction can be somewhat simplified, however. One has to note that the parasitic capacitance C_b is larger than the junction capacitance C by orders of magnitude (typically $C_b \simeq 10^4 C$). This means that the voltage on the capacitance C_b is essentially time independent although the current through the junction is composed of pulses corresponding to the tunnel events. One can thus replace the model of Fig. 2a by the model of Fig. 2b in which the impedance $Z(\omega)$ is simply in series with a voltage source V. Of course, V has to be determined self-consistently from the time-averaged current $I(V)$ through the junction by the relation $V = V_b - R_b I(V)$, but this is not a problem. If we know how to calculate the $I-V$ characteristic of the junction for the model of Fig. 2b, the junction voltage as a function of the bias current V_b/R_b for the model of Fig. 2a can be reconstructed.

An important remark is now in order. The value of the impedance $Z(\omega)$ at moderately high frequencies can be made large by making the leads on the junction chip very narrow and by using a resistive material like NiCr. However, at frequencies corresponding to micron wavelengths, no matter how careful one is in the fabrication of the leads, the impedance $Z(\omega)$ will be dominated by radiation phenomena and will be of the order of the impedance of free space $Z_V = (\mu_0/\epsilon_0)^{1/2} \simeq 377\,\Omega$. The modulus $|Z(\omega)|$ of the impedance is thus a decreasing function of frequency. This behavior can be crudely understood by considering the parasitic capacitance between the leads, whose shorting effect on the junction becomes more and more pronounced as the frequency gets higher.

A more precise understanding of the frequency dependence of the environment is provided by a resistive transmission line model of the function $Z(\omega)$. This model is analogous to the model described by Martinis and Kautz for experiments on the phase diffusion of a small Josephson junction [23]. The resistive transmission line can be thought of as a ladder of discrete components R_{ln}, C_{ln} and L_{ln}, as shown on Fig. 3. The total capacitance and resistance of the transmission line are $C_l = \sum_{n=1}^N C_{ln}$ and $R_l = \sum_{n=1}^N R_{ln}$, respectively, while the characteristic impedance of the line is $Z_l = (L_{ln}/C_{ln})^{1/2}$. In practice, Z_l is always a fraction of the vacuum impedance Z_V while C_l, which one tries to get as small as possible, is not much below 0.1 pF. As we mentioned above, by using very narrow leads made from NiCr, values of the order of 100 kΩ can be obtained for R_l (we refer to this as the "extreme" case). If no special effort is put in making high resistance lead resistors, typical values for R_l are in the 100 Ω–1 kΩ range (hereafter referred to as the "standard" case). A log-log plot of the function $|Z(\omega)|$ is shown schematically in Fig. 4. If one is in the standard case where the lead total resistance R_l is comparable to the line impedance Z_l, the environment behaves as a resistor Z_l. If one is in the extreme case where the lead total resistance is much higher than the line impedance, the environment behaves at low frequencies as a resistor R_l until a roll-off frequency given by $1/R_lC_l$ is reached. One then enters an RC line regime where the leads

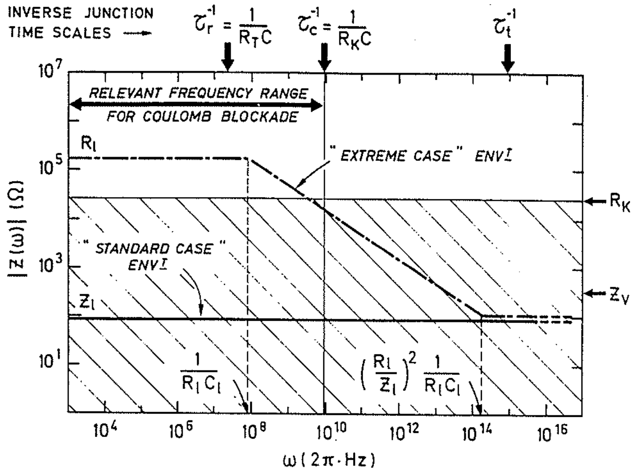


Figure 4. Schematic behavior of the modulus $|Z(\omega)|$ of the environmental impedance as a function of the frequency ω. We have shown for comparison (i) the inverse junction time scales on the frequency axis and (ii) the resistance quantum $R_K = h/e$ and the impedance of the vacuum $Z_V = \sqrt{\mu_0/\epsilon_0}$ on the resistance axis.

behave as an impedance with equal reactive and dissipative parts falling off as $\omega^{-1/2}$. Finally, at the saturation frequency $\omega_s = (R_l/Z_l)^2/(R_l C_l)$ one recovers the frequency independent behavior of the standard case. Note that the saturation frequency ω_s is independent of the length of the leads provided that $C_{ln} = C_l/N$ and $R_{ln} = R_l/N$, which is a realistic assumption. As an example, for leads with distributed resistance, capacitance, and inductance of $100\,\Omega/\mu$m, $0.5 \times 10^{-16}\,$F/μm, and $0.5 \times 10^{-12}\,$H/μm, respectively, one finds that $\omega_s/2\pi = 1.6 \times 10^{13}\,$Hz. These values correspond to NiCr resistors that are 60 nm thick and 1 μm wide.

The imperfection of the current biasing scheme, on the one hand, and the parasitic capacitances with which the environment shunts the junction, on the other hand, are thus just two aspects of the properties of the function $Z(\omega)$. Questions A and B can now be unified into a single one: What values should $|Z(\omega)|$ have at the junction characteristic frequencies in order to observe a Coulomb blockade? It is clear that in the spirit of (1) a *sufficient* condition to ensure that Q is a classical variable reads

$$|Z(\omega)| \gg R_K \quad \text{for} \quad \omega < \tau_t^{-1}, \tag{8}$$

but this is impossible to satisfy. There is a fundamental limitation since the ratio between the impedance of free space Z_V, which controls the asymptotic behavior of $Z(\omega)$ at high frequencies, and the resistance quantum R_K is equal to twice the fine structure constant 1/137. Thus one cannot avoid the problem of finding the tunneling rate as a function of V for a junction coupled to an arbitrary $Z(\omega)$. The environment needs to be treated quantum mechanically, since in most of the relevant frequency range thermal fluctuations are smaller than quantum fluctuations, i.e. $\hbar\omega \gg k_B T$. Details on the way the problem is dealt with can be found in Chap. 2. In the following we will just emphasize the important features of the theory on the effect of the electromagnetic environment [24]–[26].

The theory first assumes a clear separation of time scales

$$\tau_t \ll \tau_c \ll \tau_r. \tag{9}$$

The first inequality states that the tunneling time is negligible while the second one states the classical nature of n. The theory then considers the modes of the linear circuit formed by the environmental impedance $Z(\omega)$ in series with the junction capacitance C. Of course, for a dissipative environment, the mode frequencies form a continuous spectrum. It is assumed that before a tunnel event the environmental modes are in their equilibrium state. A tunnel event can excite them. This process is described by a function $P(E)$, which gives the probability that the tunneling electron transfers the energy E to the distribution of modes of the circuit [25]. One finds, from a quantum calculation, that $P(E)$ is a distribution function which is determined in terms of the density of environmental modes given by the real part of the total circuit impedance $Z_t(\omega) = 1/[iC\omega + Z(\omega)^{-1}]$. The probability $P(E)$ is given by

$$P(E) = \frac{1}{2\pi\hbar} \int_{-\infty}^{+\infty} \mathrm{d}t \, \exp[J(t) + iEt/\hbar] \tag{10}$$

with

$$J(t) = 2 \int_0^\infty \frac{\mathrm{d}\omega}{\omega} \frac{\mathrm{Re}[Z_t(\omega)]}{R_K} \left(\coth(\beta\hbar\omega/2)[\cos(\omega t) - 1] - i \sin(\omega t)\right) \tag{11}$$

where $\beta = 1/k_B T$ is the inverse temperature. Finally, the tunneling rate in the direction imposed by V is computed from $P(E)$ using

$$\Gamma = \frac{1}{2\tau_r E_c} \int_{-\infty}^{+\infty} \mathrm{d}E \int_{-\infty}^{+\infty} \mathrm{d}E' \; f(E)[1 - f(E')]P(E + eV - E') \tag{12}$$

which reflects the fact that only a part of the energy eV of the voltage source is used to excite the environment, the rest being used to excite one hole and one electron on either side of the barrier. At thermal energies much lower than the Coulomb energy, i.e. $\beta E_c \gg 1$, one draws the following conclusions:

i) For impedances $Z(\omega)$ such that $|Z(\omega)| \ll R_K$ for all frequencies, the function $P(E)$ is sharply peaked at $E = 0$, i.e. $P(E) \simeq \delta(E)$, and we find from (12) a straight $I - V$ characteristic with no Coulomb blockade. This result means that most tunneling transitions leave the environmental modes undisturbed except those near $\omega = 0$. The charge transferred through the junction is thus removed instantaneously by the voltage source constituted by the pads even though it is physically located a few mm away. In a way, for most tunneling events, *the environment acts as a perfect voltage source*. This is a purely quantum mechanical effect. It seems to defy locality since one would expect the charge to propagate at the speed of light from the reservoir of charge to the junction. This expectation, which seemed based on good relativistic common sense, was actually the basis for an argument in favor of the existence of Coulomb blockade phenomena for tunnel junctions in a low impedance environment [27]. There is in fact no contradiction between locality and the perfect "quantum rigidity" of the charge along the leads suggested by quasi-elastic tunneling. In this case of a low impedance environment, one calculates that the zero point motion of the environmental modes induces quantum fluctuations of the charge Q which are much larger than e. Remember that the junction and the environment behave as a whole quantum mechanically coherent unit. One can thus assume that the transferred charge is entirely provided by the zero point charge fluctuations of the environment. Crudely speaking, even though the charge is removed instantaneously by the voltage source, the source cannot tell when a tunnel event occurs because the charge pulse associated with a tunnel event is buried in the quantum fluctuations. This "charge-less" transfer of charge through the junction is analogous to the Mössbauer effect. Gamma rays can be emitted from a nucleus in a solid without exciting the phonon modes. The conservation of momentum is not violated because the recoil momentum of the nucleus is transferred to the whole crystal ("recoil-less" emission). One can think of our function $P(E)$ as being equivalent to the gamma ray energy spectrum.

ii) For ~~impedances such that $|Z(\omega)| \gg R_K$~~ for all frequencies $\omega < \tau_c^{-1}$, the tunneling electrons are well coupled to the environmental modes. One finds at zero temperature that the function $P(E)$ is sharply peaked at E_c, i.e. $P(E) \simeq \delta(E - E_c)$. Hence, like in the classical case, an electron can only tunnel when it gains at least E_c from the applied voltage, which leads to a Coulomb blockade of tunneling. The problem is that this limit is very difficult to achieve experimentally. We have represented in Fig. 4 the domain where $|Z(\omega)| < R_K$ by a shaded area. The $\omega^{-1/2}$ roll-off of the impedance must cross the impedance quantum R_K at a frequency high enough on the scale of the Coulomb frequency τ_c^{-1}. In practice this means that the on-chip lead resistors must have a saturation frequency ω_s as high as possible. This requirement is unfortunately in conflict with the requirement of no heating in the resistor and a compromise has to be found. This has been achieved by Cleland et al. [28] for normal junctions and by Kuzmin et al. [29] for superconducting junctions.

The theory we have outlined can easily be adapted to the tunneling of Cooper pairs [30]. The function $P(E)$, modified slightly to take into account the charge $2e$ of Cooper pairs, yields directly the $I - V$ characteristic if no quasiparticles are present. Schön and coworkers [31, 32] have worked out detailed predictions taking into account finite quasiparticle tunneling. As in the normal state, theory predicts that no "Cooper pair gap" exists for a single junction if the environmental impedance is less than R_K. This result explains why no Cooper pair gap was found in the Harvard group experiments described in Chap. 4. The extension of the theory to Josephson junctions is also discussed in Chap. 2.

3. Single island circuits

The preceding section showed that a current biased single tunnel junction is not, after all, a particularly simple system as far as single charge tunneling is concerned. It is certainly of interest for the foundations of the field, but it is not suited for practical applications, since the requirements for a clear-cut Coulomb blockade are so difficult to realize experimentally. We have seen that the smallness of the fine structure constant imposes the magnitude of the charge fluctuations on the junction to be much greater than e in standard cases.

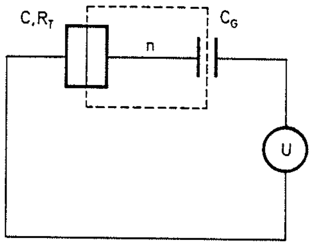


Figure 5. The single electron box, consisting of a tunnel junction in series with a capacitor. The number n of excess electrons on the island is controlled by a gate voltage U.

Another point of view can be adopted to understand why it is difficult to observe Coulomb blockade in a current biased single junction. This other point of view will make clear why islands are necessary for the occurrence of fully developed single charge tunneling phenomena. Let us compute the total *equilibrium* electrostatic energy of the circuit of Fig. 2a as a function of the number n of electrons that went through the junction. In making this calculation one assumes that the junction behaves as a perfect capacitor after the last electron has gone through it and one takes into account the work performed by the voltage source, the whole circuit having relaxed to equilibrium. One then finds

derive

$$E_{\mathrm{eq}} = -neV + \text{ terms independent of } n. \tag{13}$$

It is thus always energetically favorable for an electron to tunnel. Coulomb blockade in a current biased single junction is, at best, just a dynamical effect in which one is trying to slow down the tunneling rates as much as possible by making the environmental impedance as high as possible.

Consider now the circuit of Fig. 5 which is the simplest tunnel junction circuit containing at least one island. The island lies between a nanoscale tunnel junction with capacitance C and a gate capacitance C_G whose order of magnitude is close to that of C. This junction-capacitor combination has been nicknamed the *single electron box* [33]. The box is controlled by a voltage source U which closes the circuit. We shall restrict ourselves in the sequel to the standard case of low impedance leads. Then, as described above, the charges Q and Q_G on the junction and gate capacitances undergo large quantum fluctuations. However, these two capacitances in series couple to the leads like one capacitance $C_s = CC_G/(C + C_G)$ carrying the charge $Q_s = (C_G Q + C Q_G)/(C + C_G)$. Only this linear combination of Q and Q_G is affected by the electromagnetic environment [34]. The other linear combination, which is the island charge $Q_i = Q - Q_G$ decouples from the leads. The charge $Q_i = -ne$ is quantized in units of the elementary charge, and n, the number of "electrons in the box", is the number of excess electrons on the island. To change n an electron has to tunnel through the junction. As in (13), one can calculate the equilibrium electrostatic energy of the circuit as a function of n. It is given by

derive

$$E_{\mathrm{eq}} = \frac{(C_G U - ne)^2}{2(C + C_G)} + \text{ terms independent of } n. \tag{14}$$

There is now a big difference with the current biased junction. The equilibrium energy change

$$\Delta E_{\mathrm{eq}} = \frac{e(C_G U - ne - 1/2)}{C + C_G} \tag{15}$$

that accompanies a transition from n to $n+1$ can now be *positive*, hence ensuring an *equilibrium* Coulomb blockade. The expression (15) is in fact similar to (3), except that the junction charge is replaced by the island charge $Q_i = -ne$ which is shifted by the gate voltage. Despite the fact that ΔE_{eq} depends on the entire circuit, it can be written in terms of the average charge $\langle Q \rangle$ on the tunnel junction. Of course, $\langle Q \rangle$ depends on the applied voltage U. Using simple electrostatics one finds

$$\Delta E_{\mathrm{eq}} = \frac{e}{C}(\langle Q \rangle - Q_c), \tag{16}$$

where

$$Q_c = \frac{C}{C + C_G}\frac{e}{2} \tag{17}$$

is the so-called critical charge of the junction [11, 34] which is less than $e/2$. For the case of a low impedance environment assumed here, the tunneling rate is given by

$$\Gamma = \frac{1}{e^2 R_T}\frac{\Delta E_{\mathrm{eq}}}{1 - \exp[-\Delta E_{\mathrm{eq}}/k_B T]}, \tag{18}$$

which at zero temperature reduces to

$$\Gamma = \begin{cases} \Delta E_{\mathrm{eq}}/e^2 R_T & \text{for} \quad \Delta E_{\mathrm{eq}} > 0 \\ 0 & \text{for} \quad \Delta E_{\mathrm{eq}} < 0. \end{cases} \tag{19}$$

As soon as $\langle Q \rangle$ exceeds Q_c an electron can tunnel onto the island. According to formula (18), the tunneling rate is determined by the change (15) of the equilibrium electrostatic energy of the entire system caused by the transition. This so-called global rule rate [35, 36] is found to be very accurate when the tunneling resistance R_T satisfies (1) and the electromagnetic environment is of low impedance [34]. From (19) we see that at zero temperature a transition from n to $n+1$ only occurs for $C_G U > e(n + \frac{1}{2})$. Likewise, one finds for transitions from n to $n-1$ the condition $C_G U < e(n - \frac{1}{2})$. Hence, for sufficiently low temperatures and gate voltages U in the interval

$$e(n - \frac{1}{2}) < C_G U < e(n + \frac{1}{2}), \tag{20}$$

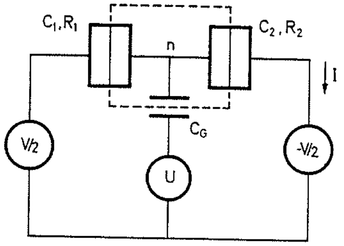


Figure 6. The SET transistor consists of two tunnel junctions in series forming an island the electrostatic potential of which is acted upon by the gate voltage U through the capacitance C_G. The transport voltage V induces a net flow of charge through the device, the value of current I being controlled by the gate voltage U.

the state with n electrons in the box is stable. By changing U electrons can thus be added one-by-one to the box. Hence, the single electron box is a simple device allowing for the manipulation of a single charge. Further details on this system are given by Lafarge et al. [33] and in Chap. 3. At the time of this writing, a box for Cooper pairs could not be operated successfully. Since for quasiparticles the threshold voltage for tunneling is always lower than that for Cooper pairs, an island between a Josephson junction and a capacitance will behave like an electron box even when the density of quasiparticles is small.

Another basic device with just one island is the double junction driven by a transport voltage V. Very often a gate capacitor with a gate voltage U is coupled to the island between the junctions. This is the single electron tunneling (SET) transistor [8] with the circuit diagram depicted in Fig. 6. The first observation of single charge tunneling in microfabricated samples by Fulton and Dolan [10] was made with this device. SET transistors are also part of more elaborate devices fabricated with oxide layer tunnel junctions (see Chap. 3), and most of the studies on charging effects in semiconductors have used this type of circuit (see Chap. 5).

In the SET transistor the island can be charged by tunneling across one junction and discharged by tunneling across the other junction, which leads to a net current through the device. If the tunneling resistances of both junctions satisfy (1), the electron transfer rates are again determined by the change of the equilibrium electrostatic energy of the circuit. In semiconductor devices with a small number of electrons in the segment between the junctions, the island may form a quantum dot with an energy level separation that exceeds $k_B T$. Then the energy difference between the Fermi level of the dot and the next available state has to be considered when calculating the energy change. This case is discussed in Chap. 5. When the discreteness of the spectrum of electronic states on the island can be disregarded, the energy change due to a transition from n to $n+1$ excess electrons as a consequence of tunneling across the first junction [cf. Fig. 3] is found to be (we drop the subscript eq)

$$\Delta E_1 = \frac{e[(C_2 + \frac{1}{2}C_G)V + C_G U + ne - \frac{e}{2}]}{C_\Sigma}, \tag{21}$$

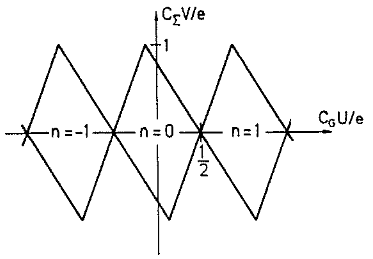


Figure 7. The stability diagram of a SET transistor with $2C_2 = 10C_G = C_1$. The transistor conducts only outside the rhombic-shaped regions. Inside these regions, there is a constant number n of electrons on the island.

where

$$C_\Sigma = C_1 + C_2 + C_G \tag{22}$$

is the capacitance of the island. In (21) the gate voltage U only appears in the combination $C_G U - ne$, which leads for all measurable quantities to a periodicity in U with period e, since the integer part of $C_G U/e$ can always be absorbed in n. In semiconductor devices this strict periodicity is usually not met because the gate voltage U influences the tunneling resistances of the junctions and because the gate capacitance is weakly dependent on U.

A straightforward analysis of the rate formula (19) shows that at zero temperature the state with n electrons on the island of the SET transistor is stable with respect to tunneling across the first and second junctions for voltages satisfying

$$e(n - \frac{1}{2}) < C_G U + (C_2 + \frac{1}{2}C_G)V < e(n + \frac{1}{2})$$

$$e(n - \frac{1}{2}) < C_G U - (C_1 + \frac{1}{2}C_G)V < e(n + \frac{1}{2}), \tag{23}$$

respectively. Hence, in the UV-plane there are rhombic-shaped regions along the U-axis within which the transistor island is charged with a fixed number of excess electrons [cf. Fig. 7]. Inside these rhombi all transitions are suppressed by a Coulomb blockade and no current flows through the device.

For example, near $U = V = 0$ the state $n = 0$ is stable. The inequalities (23) show that whenever the system leaves the stability region of $n = 0$ at a point in the UV-plane with $V \neq 0$ and a tunneling transition, say, to $n = 1$ occurs, the new state is not stable with respect to tunneling across the other junction. Hence, shortly after the first tunneling event an electron leaves the island through the other junction and the system returns to $n = 0$, where the cycle can start again. As a net effect, a current flows through the device. The second tunneling transition of this cycle occurs for $Q > Q_c$ and part of the change of electrostatic energy is left as kinetic energy of the tunneling electron. Therefore, the transistor is a dissipative element, in contrast to the single electron box discussed above which is reversible when the gate voltage is changed slowly.

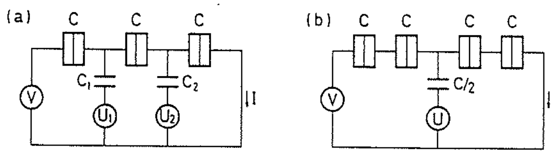


Figure 8. The circuit diagrams of the single electron pump (a) and turnstile (b). An appropriate rf modulation applied to the gate voltage(s) of these devices transfers precisely one electron through them per cycle.

For voltages V of order e/C the current I is very sensitive to U. A small change of the polarization charge $C_G U$ by a fraction of the elementary charge e can change the current from zero to values of the order of E_c/eR_T. This is why the SET transistor can be used as a highly sensitive electrometer [10, 33]. It also could serve as a low-noise amplifier of analog signals. Finally, the SET transistor is the basic active element of digital and other applications proposed for single charge tunneling (see Chap. 9). A detailed discussion of the SET transistor is given in Chap. 2. Since the transistor is a dissipative element, Cooper pairs can usually be transferred only in combined processes involving quasiparticles or environmental modes or both. The complex behavior of the superconducting device is discussed by Maassen van den Brink et al. [32].

At zero temperature and for voltages within the intervals (23), the state with n electrons on the transistor island is stable with respect to tunneling across either junction. However, in the presence of an applied voltage V the state can only be metastable. In fact, Averin and Odintsov [37] have pointed out that second order transitions always lead to a finite current in the presence of an applied voltage. In these co-tunneling events an electron tunnels onto the island while a second electron simultaneously leaves the island across the other junction. Since the charge on the island is only changed virtually, there is no Coulomb barrier for this process. The co-tunneling rate is proportional to $(R_K/R_T)^2$ and hence is a factor R_K/R_T smaller than the rate for first order processes. Accordingly, in more complicated multijunction circuits there are co-tunneling events involving N junctions with rates proportional to $(R_K/R_T)^N$. Of course, co-tunneling mainly arises when the inequality (1) is only poorly satisfied. However, since the time scale of all single charge tunneling processes is proportional to R_K/R_T, very large tunneling resistances severely reduce the speed of devices. That is why a detailed understanding of co-tunneling is of essential importance to the field. A survey of the theory is given in Chap. 6.

4. Circuits with several islands

More sophisticated multijunction circuits can be built using the single electron box or the single electron tunneling transistor as basic units. Fig. 8 shows the circuit diagram of the "pump" and "turnstile" devices fabricated recently by the Saclay and Delft groups. The pump designed by Pothier et al. [12] can be seen as two boxes connected by a tunnel junction. The boxes allow for a control of the input and output of electrons by means of the gate voltages. When the appropriate ac voltages with frequency f are applied to the gates, precisely one electron is transferred per cycle through the device, giving a current

$$I = ef. \tag{24}$$

This relation is the basis of high precision SCT current sources. The difference between (5) and (24) is that in (24) f is an externally imposed frequency. Since the pump principle employs only reversible processes, it can also be used to transfer Cooper pairs [38].

The turnstile designed by the Delft and Saclay groups [11] can be seen as two double junctions connected by a common island. The charging and discharging of this island is controlled by a gate voltage. Again, a current obeying (24) can be generated by means of an ac voltage. In a semiconductor version of the turnstile fabricated by Kouwenhoven et al. [39], the tunneling resistances are modulated via the voltages applied to the Schottky gates. Deviations from (24) mainly arise from finite temperature effects, electron heating, co-tunneling, and moving background charges. These effects must be reduced to achieve a current standard with metrological accuracy. A detailed introduction into the art of manipulating electrons one-by-one is given in Chap. 3.

Another class of multi-island circuits are one-dimensional and two-dimensional arrays (see Fig. 9). In these wonderful man-made crystals, the tolerable dispersion of "microscopic" parameters is more than compensated for by the fact that one can tune them to explore effects that would not exist in the natural world. In the one-dimensional arrays (Fig. 9a) of the Göteborg group (see Chap. 7) extended charge solitons are created by an external voltage which then makes them drift through the array. Only one soliton can exist at a time in the array and the reciprocal of the average time the soliton takes to travel along the array is given by (5). In the two-dimensional arrays (Fig. 9b) of the Delft group (see Chap. 8) many charge solitons of limited size exist at the same time and an interesting cooperative behavior similar to the Kosterlitz-Thouless transition arises. The two-dimensional arrays in the superconducting state are particularly fascinating, since, depending on the junction and island parameters, one can have either charge solitons (one extra Cooper pair on an island) or flux solitons (vortices sitting on the plaquette defined by four junctions). Under proper experimental conditions it seems possible that the motion of charge and flux vortices could be quantum mechanical. The duality relationship between the charge and flux solitons can be exploited to predict new quantum effects [40] discussed in Chap. 8.

5. Conclusions

Single charge tunneling has only recently developed into a field investigated in many laboratories world-wide, partly due to the progress in nanoscale fabrication techniques. Yet this area of research is about to leave its infancy. The main problems still to be overcome have been identified, and it might be appropriate to conclude by speculating about possible applications.

As mentioned above, the single electron tunneling transistor serves as a highly sensitive electrometer. The sensitivity of existing prototypes already exceeds those of other

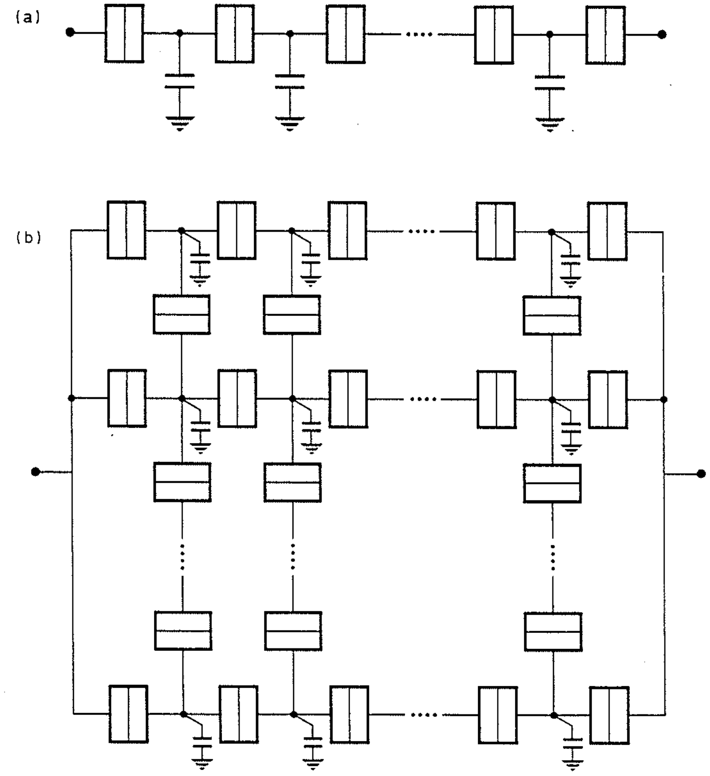


Figure 9. Arrays of tunnel junctions: (a) one-dimensional array, (b) two-dimensional array.

electrometers by 6 orders of magnitude, and the performance can certainly be improved further. The maximal sensitivity attainable is presently not known, but it should be at least $10^{-5}e/\sqrt{\text{Hz}}$. Despite this very high precision, the SET electrometer is for the measurement of electrical charges not quite as revolutionary as the SQUID was for the measurement of magnetic flux. Since there is no analogue of the superconducting flux transformer, the very small input capacitance of the single electron tunneling electrometer might limit its usefulness.

The pump and turnstile devices demonstrate that single charge tunneling can be employed to construct frequency-controlled current sources. The relative uncertainty of the current produced by existing devices is below 10^{-2}, but the error sources are being analysed and improved designs are under investigation. Whether metrological accuracy of 10^{-8} is really achievable is not known, although the prospects are rather promising. Since the ampère is presently derived from the kilogram, a closure of the quantum metrological triangle could ultimately revolutionize the metrological system, and perhaps do away with the last relic of the Bureau International des Poids et Mesures in Sèvres, i.e., the standard kilogram.

Much of the fascination of single charge tunneling derives from the idea that, in the future, a single bit in an information flow might possibly be represented by a single electron (see Chap. 9). Although ingenious designs are being proposed, the less-than-unity voltage gain of the single electron tunneling transistor remains at present a fundamental engineering problem. A complementary logic analogous to the CMOS recently proposed by Tucker [41], is an attempt to overcome this problem. It is important to note that in the single electron tunneling transistor the modulation of the flow of electrons by the gate ceases as soon as the bias voltage exceeds the Coulomb gap, whereas in the FETs used in digital circuits the modulation of the source-drain current by the gate only saturates at large bias voltages [1]. This latter feature ensures enough voltage gain to compensate for the dispersion in device parameters and makes robust integrated circuit designs possible. It may happen that the real impact of single charge tunneling phenomena on nanoelectronics will be to show how, in the next generation of FET devices, one can make Coulomb charging effects reinforce the dominant Fermi effects upon which FETs are based, instead of spoiling them. Not only is this a very valuable goal, but along the way, other striking advances in fundamental science are likely to be made, like perhaps the controlled transfer of fractional charges in semiconducting systems in the quantum Hall effect regime.

ACKNOWLEDGEMENTS. We should like to thank the participants of the NATO Advanced Study Institute on SCT in Les Houches for many inspiring discussions, in particular D. Esteve, G.-L. Ingold, P. Lafarge, H. Pothier, and C. Urbina. This work was started during an extended stay by one of us (H.G.) at the Centre d'Etudes de Saclay whose kind hospitality is gratefully acknowledged. Financial support was provided through grants from NATO, the European Community, and the Deutsche Forschungsgemeinschaft.

References

[1] D. A. Fraser, *The Physics of Semiconductor Devices*, (Clarendon, Oxford, 1986).
[2] C. J. Gorter, Physica **17**, 777 (1951).
[3] C. A. Neugebauer and M. B. Webb, J. Appl. Phys. **33**, 74 (1962).
[4] I. Giaever and H. R. Zeller, Phys. Rev. Lett. **20**, 1504 (1968).
[5] J. Lambe and R. C. Jaklevic, Phys. Rev. Lett. **22**, 1371 (1969).
[6] I. O. Kulik and R. I. Shekhter, Zh. Eksp. Teor. Fiz. **68**, 623 (1975) [Sov. Phys. JETP **41**, 308 (1975)].
[7] G. J. Dolan and J. H. Dunsmuir, Physica B **152**, 7 (1988).
[8] K. K. Likharev, IBM J. Res. Dev. **32**, 144 (1988); D. V. Averin and K. K. Likharev, in: *Quantum Effects in Small Disordered Systems*, ed. by B. L. Altshuler, P. A. Lee, and R. A. Webb (Elsevier, Amsterdam, 1991).
[9] G. Schön and A. D. Zaikin, Phys. Rep. **198**, 237 (1990).
[10] T. A. Fulton and G. J. Dolan, Phys. Rev. Lett. **59**, 109 (1987).

[11] L. J. Geerligs, V. F. Anderegg, P. A. M. Holweg, J. E. Mooij, H. Pothier, D. Esteve, C. Urbina, and M. H. Devoret, Phys. Rev. Lett. **64**, 2691 (1990).
[12] H. Pothier, P. Lafarge, P. F. Orfila, C. Urbina, D. Esteve, and M. H. Devoret, Physica B **169**, 573 (1991); Europhys. Lett. **17**, 249 (1992).
[13] K. A. Matveev, Zh. Eksp. Teor. Fiz. **99**, 1598 (1991) [Sov. Phys. JETP **72**, 892 (1991)].
[14] W. Zwerger and M. Scharpf, Z. Phys. B **85**, 421 (1991).
[15] *Special Issue on Single Charge Tunneling*, Z. Phys. B **85**, 317–468 (1991).
[16] A. Widom, G. Megaloudis, T. D. Clark, H. Prance, and R. J. Prance, J. Phys. A **15**, 3877 (1982).
[17] K. K. Likharev and A. B. Zorin, J. Low Temp. Phys. **59**, 347 (1985).
[18] E. Ben-Jacob and Y. Gefen, Phys. Lett. A **108**, 289 (1985).
[19] D. V. Averin and K. K. Likharev, J. Low Temp. Phys. **62**, 345 (1986).
[20] M. Büttiker and R. Landauer, Phys. Rev. Lett. **49**, 1739 (1982).
[21] E. H. Hauge and J. A. Støvneng, Rev. Mod. Phys. **61**, 917 (1989).
[22] B. N. J. Persson and A. Baratoff, Phys. Rev. B **38**, 9616 (1988).
[23] J. M. Martinis and R. L. Kautz, Phys. Rev. Lett. **63**, 1507 (1989).
[24] Yu. V. Nazarov, Pis'ma Zh. Eksp. Teor. Fiz. **49**, 105 (1989) [JETP Lett. **49**, 126 (1989)].
[25] M. H. Devoret, D. Esteve, H. Grabert, G.-L. Ingold, H. Pothier, and C. Urbina, Phys. Rev. Lett. **64**, 1824 (1990).
[26] S. M. Girvin, L. I. Glazman, M. Jonson, D. R. Penn, and M. D. Stiles, Phys. Rev. Lett. **64**, 3183 (1990).
[27] M. Büttiker and R. Landauer, IBM J. Res. Dev. **30**, 451 (1986).
[28] A. N. Cleland, J. M. Schmidt, and J. Clarke, Phys. Rev. Lett. **64**, 1565 (1990).
[29] L. S. Kuzmin, Yu. V. Nazarov, D. B. Haviland, P. Delsing, and T. Claeson, Phys. Rev. Lett. **67**, 1161 (1991).
[30] D. V. Averin, Yu. V. Nazarov, and A. A. Odintsov, Physica B **165&166**, 945 (1990).
[31] G. Falci, V. Bubanja, and G. Schön, Europhys. Lett. **16**, 109 (1991); Z. Phys. B **85**, 451 (1991).
[32] A. Maassen van den Brink, A. A. Odintsov, P. A. Bobbert, and G. Schön, Z. Phys. B **85**, 459 (1991).
[33] P. Lafarge, H. Pothier, E. R. Williams, D. Esteve, C. Urbina, and M. H. Devoret, Z. Phys. B **85**, 327 (1991).
[34] H. Grabert, G.-L. Ingold, M. H. Devoret, D. Esteve, H. Pothier, and C. Urbina, Z. Phys. B **84**, 143 (1991).
[35] K. K. Likharev, N. S. Bakhvalov, G. S. Kazacha, and S. I. Serdyukova, IEEE Trans. Magn. **25**, 1436 (1989).
[36] U. Geigenmüller and G. Schön, Europhys. Lett. **10**, 765 (1989).
[37] D. V. Averin and A. A. Odintsov, Phys. Lett. A **140**, 251 (1989).
[38] L. J. Geerligs, S. M. Verbrugh, P. Hadley, J. E. Mooij, H. Pothier, P. Lafarge, C. Urbina, D. Esteve, and M. H. Devoret, Z. Phys. B **85**, 349 (1991).
[39] L. P. Kouwenhoven, A. T. Johnson, N. C. van der Vaart, A. van der Enden, C. J. P. M. Harmans, and C. T. Foxon, Z. Phys. B **85**, 381 (1991).
[40] B. J. van Wees, Phys. Rev. B **44**, 2264 (1991).
[41] J. R. Tucker, to be published.

Date: Sun, 16 Mar 2003 23:47:13 -0800 (PST)
From: pburke@uci.edu
To: sldad@uci.edu, pburke@uci.edu
Subject: DDS Article request from Peter Burke

=============================
Delivery format: Web
Author: H. Grabert
Article Title: Single Charge Tunneling
Journal or Newspaper Title: Zeitschrift fur Physik B Condensed Matter
Date: 1991
Volume: 85
Issue:
Pages: 319-325
Call Number: QC 1 Z37 Sec.B
Special Instructions: rush
Recharge Number: 402558-19900
Project Code: Burke Startup
Department: EECS
Send Request to: Science Library
Name: Peter Burke
Library Card Number: 21970004137302
Email Address pburke@uci.edu
Affiliation Faculty =============================

This email message was sent from the Web form located at
http://dds.lib.uci.edu/forms/illj2.html

Z. Phys. B – Condensed Matter 85, 319–325 (1991)

Condensed Matter
Zeitschrift für Physik B
© Springer-Verlag 1991

Single charge tunneling: a brief introduction

Hermann Grabert[1,2]

[1] Fachbereich Physik, Universität-GH Essen, W-4300 Essen, Federal Republic of Germany
[2] Service de Physique de l'Etat Condensé, Centre d'Etudes de Saclay, F-91191 Gif-sur-Yvette, France

Received August 12, 1991

The field of single charge tunneling comprises of phenomena where the tunneling of a microscopic charge, usually carried by an electron or a Cooper pair, leads to macroscopically observable effects. The basic principles governing this area of research are briefly outlined and the present state of the art is discussed.

1. Introduction

The importance of Coulomb charging effects for charge transfer through small systems was first noted several decades ago [1–5]. At that time, Coulomb blockade phenomena could only be observed in granular metallic materials, in which single electron effects and random media properties interplay. Nowadays, modern lithography allows for the controlled fabrication of submicron structures, where metallic islands with capacitances C in the fF range or below are separated by tunneling barriers with resistances R_T well above the resistance quantum $R_K = h/e^2 \simeq 25.8\,\mathrm{k\Omega}$. In such systems, the charging energy $E_c = e^2/2C$ of a single excess electron on the metallic island exceeds the energy $k_B T$ of thermal fluctuations at sub Kelvin temperatures. As a consequence, a Coulomb blockade of tunneling arises [6, 7] which can be exploited to transfer single charges from one island to another in a controlled way [8, 9].

The paragraph above indicates the basic requirements for Single Charge Tunneling (SCT) phenomena to occur. Leaving aside for the moment the special case of a single tunnel junction which will be discussed in the following section, these conditions are as follows. Firstly, the system must have metallic islands that are connected to other metallic regions only via tunnel barriers with a tunneling resistance that exceeds the resistance quantum, i.e.,

$$R_T \gg R_K. \tag{1}$$

This condition ensures that the wave function of an excess electron or Cooper pair on an island is basically localized there. In systems with lower tunneling resistances, charges can be transferred through small islands without paying the charging energy as a penalty, since delocalized states with lower Coulomb energy are available for the transport. Secondly, the islands have to be small enough and the temperature has to be low enough so that the energy required to add a charge carrier to an island exceeds the mean thermal energy of the charge carriers, i.e.,

$$E_c \gg k_B T. \tag{2}$$

This ensures that the transport of charges is in fact governed by the Coulomb charging energy. With the use of externally applied voltages, the charging energy can then be influenced in order to manipulate the charge carriers.

At present, two main types of systems where SCT effects arise are being explored. Much of the work done in the last few years has used lithographically patterned tunnel junction circuits, where metallic islands (mostly made from Al) are separated by oxide layer tunnel barriers. In this case, three-dimensional electron gases confined to small regions are weakly coupled by the tunnel effect. These systems also allow one to explore charging effects involving Cooper pairs since the metals used to fabricate the circuits are superconductors. At the temperatures required to satisfy (2), one must apply a magnetic field to keep the metals in the normal state. Specific examples for such circuits are given in the articles by Haviland et al. [10], Lafarge et al. [11] and Geerligs et al. [12].

Single electron effects also arise when the two-dimensional electron gas of a GaAs/AlGaAs heterostructure is confined to small islands by means of Schottky gates. In this case the tunneling resistances of the constrictions separating the islands can be tuned by changing the gate voltages. Further, the islands may be quantum dots with a discrete energy spectrum. Such semiconductor circuits are presented in the articles by Meirav et al. [13], Kouwenhoven et al. [14, 15], and Glattli et al. [16]. A different structure where electrons tunnel vertically to the plane of the two-dimensional electron gas is discussed by

Ramdane et al. [17]. Apart from these lithographically patterned systems, single charge tunneling phenomena are observed in a number of other cases. There is a large body of work on disordered systems such as granular films [1–5], small metal particles embedded in an oxide layer [18], or disordered quantum wires [19]. Also, one of the tunneling barriers may be formed by a scanning tunneling microscope [20]. Detailed lists of references to earlier studies may be found in the review articles by Averin and Likharev [6] and Schön and Zaikin [7].

This article is not intended to review the field of single charge tunneling: rather, the main issues will be briefly discussed with particular emphasis on subjects covered in this Special Issue. A detailed introduction to this field and a survey of recent activities is given in Ref. 21. In Sect. 2, we discuss single tunnel junctions. The main predictions of the conventional theory are summarized and the disappearance of the Coulomb barrier due to the leads attached to the junction is discussed. Then, in Sect. 3 the basic components of circuits where SCT occurs are described, and possible applications are discussed. Finally, in Sect. 4, we give a short summary.

2. Single tunnel junctions

Charging effects in small capacitance tunnel junctions became a main topic of low temperature physics a few years ago when several new effects due to the quantization of the charge were predicted to arise in both superconducting [22, 23] and normal tunnel junctions [24, 25]. In particular, Likharev and coworkers have expanded the theory of Coulomb blockade phenomena and have proposed various applications of the new effects [6].

The conventional treatment of charging phenomena in ultrasmall junctions starts out from the current-biased tunnel junction [6, 7]. Charging effects result from an interplay between the continuous nature of the charge Q on the junction capacitor and the discrete nature of charge tunneling across the junction. On the one hand, the current I increases the charge Q in a continuous way, i.e., $\dot{Q}=I$, since the charge transferred from the external circuit to the capacitance C is a continuous variable. In fact, Q may be an arbitrarily small fraction of the elementary charge e, caused by a small shift of the electrons in the junction electrodes with respect to the positive ionic background. On the other hand, tunneling through the junction results in a sudden discharge by e or $2e$, depending on whether an electron or a Copper pair is tunneling. For a normal junction, the reduction of the Coulomb energy $Q^2/2C$ by a tunneling event is

$$\Delta E=\frac{Q^2}{2C}-\frac{(Q-e)^2}{2C}=\frac{e\left(Q-\frac{e}{2}\right)}{C}. \qquad (3)$$

Now, at zero temperature tunneling can only occur if ΔE is positive, which implies a Coulomb blockade of tunneling for $Q<e/2$. Hence, the current-voltage characteristic of the junction should show a Coulomb gap, i.e.,

$$I=0 \qquad \text{for } -\frac{e}{2C}<V<\frac{e}{2C}. \qquad (4)$$

Furthermore, the current source will charge the capacitor until the threshold charge $e/2$ is reached. Then, a tunneling transition occurs, leading to $Q=-e/2$, and a new charging cycle starts. This leads to Single Electron Tunneling (SET) oscillations [25] of the voltage with the fundamental frequency

$$f_{\mathrm{SET}}=I/e. \qquad (5)$$

By a similar kind of reasoning one predicts for a Josephson junction a Coulomb blockade of Cooper pair tunneling and so-called Bloch oscillations [23] with the frequency

$$f_{\mathrm{Bloch}}=I/2e. \qquad (6)$$

This analysis assumes that the tunnel junction can be considered as being independent of its electromagnetic environment, which is represented by an ideal current source. As already mentioned, a charging energy $E_c=e^2/2C$ well above k_BT is only attainable for junctions with capacitances in the fF range or below. However, such ultrasmall junctions are strongly affected by the leads attached to them [26–28]. Firstly, the leads have capacitances that always exceed the junction capacitance by several orders of magnitude. These parasitic capacitances are polarized by the average voltage across the junction and act as an effective voltage source. Secondly, the electromagnetic modes of the leads and external circuit are coupled to the electric field in the junction. As a consequence, the environmental modes influence the charge tunneling rates. Since the typical frequency of tunneling transitions is in the GHz range, the leads can usually be described in terms of a circuit model [29]. The electromagnetic environment is then characterized by the impedance $Z(\omega)$ seen from the location of the junction. These considerations lead to the more realistic junction model of Devoret et al. [27] depicted in Fig. 1.

When charge tunneling rates are calculated for the coupled system formed by the junction and its electromagnetic environment, one finds that no Coulomb blockade occurs under standard experimental conditions. Typical environmental impedances $Z(\omega)$ are of the order of the impedance of free space ($Z_V\simeq 377\ \Omega$), which is small compared to the resistance quantum $R_K\simeq 25.8\ \mathrm{k}\Omega$. The influence of the environment on electron tunneling rates in normal junctions can be described in terms of a function $P(E)$ which gives the probability that the tunneling electron transfers the energy E to the electromagnetic modes of the circuit [27]. For impedances $Z(\omega)$ with $|Z(\omega)|\ll R_K$, one finds $P(E)\simeq\delta(E)$, which means that most tunneling transitions are basically elas-

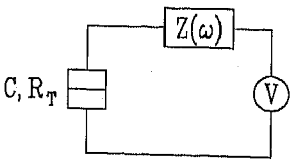


Fig. 1. A realistic model for an ultrasmall tunnel junction and its electromagnetic environment. The junction is attached to a circuit with an impedance $Z(\omega)$ and a voltage source V

tic, i.e., they do not lead to an excitation of the environment. Now, a change of the charge Q on the junction capacitor disturbs the equilibrium between the junction and its environment and thus corresponds to the excitation of electromagnetic modes. Hence, in an elastic transition the electron charge is immediately transferred to the large parasitic capacitances and no change of Q occurs. We thus see that the junction and the leads act like a system with a large capacitance, and the Coulomb blockade is therefore removed. The situation is reminiscent of the Mössbauer effect, where for recoil-less transitions which do not excite the phonon modes, the recoil momentum of the nucleus is transferred to the entire crystal. On the other hand, when the environmental impedance $Z(\omega)$ is of the order of R_K, low-frequency electromagnetic modes are more abundant and they can be more easily excited by a tunneling electron. For an impedance satisfying $|Z(\omega)| \gg R_K$ for all frequencies $\omega \lesssim E_c/\hbar$ one finds at zero temperature $P(E) \simeq \delta(E - E_c)$ [30]. Hence, an electron can only tunnel when it gains at least E_c from the applied voltage, which leads to a Coulomb blockade of tunneling.

The approach by Devoret et al. [27, 30] can equally well be applied to superconducting junctions [31, 32]. Again, the influence of the external circuit is described by the probability $P(E)$ introduced above, which for Cooper pairs is modified slightly since they carry twice the electron charge. In particular, Schön and coworkers [32, 33] have put forward detailed predictions for Josephson junctions. As for normal junctions, charging effects are found to be observable only for environmental impedances of the order of R_K. However, this limit is very hard to achieve experimentally, since a very high resistance has to be placed very close to the junction without causing substantial heating. Remarkable progress in this direction was made by Cleland et al. [34] for normal junctions and by Haviland et al. [10] for Josephson junctions.

These considerations show that single tunnel junctions are not particularly simple systems as far as SCT is concerned. They are certainly of interest for the foundations of the field, but are not suited for practical applications. Further aspects of the theory on the influence of the environment are presented in the articles by Flensberg et al. [35] and Falci et al. [32], and a survey is given by Ingold and Nazarov [36]. The main insight gained from these studies is that ultrasmall tunnel junctions with ordinary metallic leads are well described by a voltage bias except for large voltages of the order of $(E_c/e)|R_K/Z(E_c/\hbar)|$, where a crossover to predominately inelastic tunneling transitions occurs [27, 30]. The corresponding offset of the current-voltage characteristic at large voltages is the principal charging effect observable for single junctions with standard current and voltage leads [37]. Recent theoretical work on single junctions also includes a study of thermoelectric effects by Amman et al. [38]. Conductance anomalies arising from electronic excitations caused by the tunneling of localized charges are predicted in the article by Ueda and Guinea [39]. This last prediction has been questioned, since a charge disturbance due to tunneling is not necessarily localized like the core-hole in the x-ray edge problem.

3. Tunnel junction circuits

As we have seen, charging effects are usually not very important for single tunnel junctions. The leads effectively provide a voltage bias and cause large zero-point fluctuations of the charge Q on the junction capacitor that wash out single charge effects. For multijunction systems the effect of the environment on SCT was studied by Grabert et al. [40], and detailed predictions for several cases were made in subsequent articles by Grabert et al. [30], Maasen van den Brink et al. [33] and Ingold et al. [41]. Provided that the tunneling resistances of the junctions satisfy (1), pronounced charging effects arise in multijunction circuits even for a low impedance environment as a consequence of the charge quantization on the islands between the junctions.

The simplest system showing this quantization of an island charge is the Single Electron Box (SEB) studied both experimentally and theoretically by Lafarge et al. [11]. The island lies between an ultrasmall tunnel junction with capacitance C and an equally small gate capacitance C_G, and the device is controlled by a gate voltage U. The corresponding circuit diagram is shown in Fig. 2. Because of the low impedance leads from the battery to the box, the charges Q and Q_G on the junction and gate capacitors undergo large quantum fluctuations. However, since the two capacitances in series couple to the leads like one capacitance $C_\parallel = CC_G/(C + C_G)$, carrying the charge $Q_\parallel = (C_G Q + CQ_G)/(C + C_G)$, only this linear combination of Q and Q_G is affected by the electromagnetic environment [30], while the island charge

$$q = Q - Q_G = -ne \tag{7}$$

decouples from the leads. The charge q is quantized in units of the elementary charge, and is determined by the number n of electrons in the SEB. To change q an electron has to tunnel through the junction. From the Coulomb energy $q^2/2(C + C_G)$ of the island charge and the work done by the voltage source to restore equilibrium after the tunneling event, one calculates that the electrostatic energy is reduced by

$$\begin{aligned}\Delta E &= \frac{q^2}{2(C + C_G)} - \frac{(q-e)^2}{2(C + C_G)} + \frac{C_G}{(C + C_G)} eU \\ &= \frac{e\left(q + C_G U - \frac{e}{2}\right)}{C + C_G}\end{aligned} \tag{8}$$

when an electron tunnels onto the island. This is very similar to (3), except that the junction charge is replaced

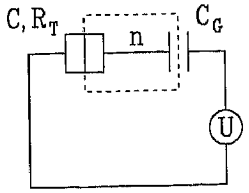


Fig. 2. The single electron box, consisting of a tunnel junction in series with a capacitor. The number n of electrons in the box is controlled by a gate voltage U

by the island charge q which is shifted by the gate voltage. Despite the fact that ΔE depends on the entire circuit, it can be written in terms of the average charge $\langle Q \rangle$ on the tunnel junction. Of course, $\langle Q \rangle$ depends on the applied voltage U. Using simple electrostatics one finds

$$\Delta E = \frac{e}{C} (\langle Q \rangle - Q_c), \tag{9}$$

where

$$Q_c = \frac{C}{C + C_G} \frac{e}{2} \tag{10}$$

is the so-called critical charge of the junction [8, 30]. As soon as $\langle Q \rangle$ exceeds Q_c an electron can tunnel onto the island. The tunneling rate is given by

$$\Gamma = \frac{1}{e^2 R_T} \frac{\Delta E}{1 - \exp[-\Delta E/k_B T]}, \tag{11}$$

which at zero temperature reduces to

$$\Gamma = \frac{1}{e^2 R_T} \Delta E \quad \text{for } T = 0 \text{ and } \Delta E > 0. \tag{12}$$

According to this simple formula the tunneling rate is determined by the change (8) of the electrostatic energy of the entire system caused by the transition. This so-called global rule rate [42, 43] is found to be very accurate when the tunneling resistance R_T satisfies (1) and the electromagnetic environment is of low impedance [30]. From (12) we see that at zero temperature a transition from n to $n+1$ only occurs for $C_G U > e(n + \frac{1}{2})$. Likewise, one finds for transitions from n to $n-1$ the condition $C_G U < e(n - \frac{1}{2})$. Hence, for sufficiently low temperatures and gate voltages U in the interval

$$e(n - \tfrac{1}{2}) < C_G U < e(n + \tfrac{1}{2}), \tag{13}$$

the state with n electrons in the box is stable. By changing U electrons can thus be added one-by-one to the box. Hence, the SEB is a simple device allowing for the manipulation of a single charge. Further details on this system are given in the articles by Lafarge et al. [11] and Esteve [44]. At the time of this writing, a box for Cooper pairs could not be operated successfully. Since for quasiparticles the threshold voltage for tunneling is always lower than that for Cooper pairs, an island between a Josephson junction and a capacitance will behave like an electron box even when the density of quasiparticles is small.

Another basic device with just one island is the double junction driven by a transport voltage V. Very often a gate capacitor with a gate voltage U is coupled to the island between the junctions. This is the Single Electron Tunneling (SET) transistor [6] with the circuit diagram depicted in Fig. 3. The first observation of SCT in microfabricated samples by Fulton and Dolan [45] was made with this device. SET transistors are also part of more elaborate devices fabricated with oxide layer tunnel junctions [8, 11, 12], and most of the studies on charging effects in semiconductors [13–16] have used this type of circuit.

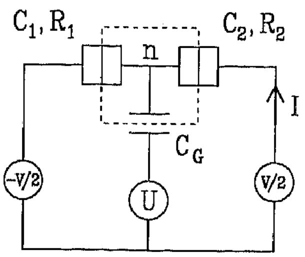


Fig. 3. The SET transistor consists of two tunnel junctions and a gate capacitor with a common electrode. The current I caused by the transport voltage V is controlled by the gate voltage U

In the SET transistor the island can be charged by tunneling across one junction and discharged by tunneling across the other junction, which leads to a net current through the device. If the tunneling resistance of both junctions satisfy (1), the electron transfer rates are again determined by the change of the electrostatic energy of the circuit. In semiconductor devices with a small equilibrium number of electrons in the segment between the junctions, the island may form a quantum dot with an energy level separation that exceeds $k_B T$. Then the energy difference between the Fermi level of the dot and the next available state has to be considered when calculating the energy change. This case is discussed in detail by van Houten et al. [46], and a quantum dot with only a few electrons is considered by Häusler et al. [47]. Experimental results on SCT through a quantum dot are presented in the articles by Meirav et al. [13] and Kouwenhoven et al. [14]. An island in the two-dimensional electron gas with a quasi continuous energy spectrum was studied by Glattli et al. [16].

When the discreteness of the spectrum of electronic states on the island can be disregarded, the energy change due to a transition from n to $n+1$ excess electrons as a consequences of tunneling across the first junction [cf. Fig. 3] is found to be

$$\Delta E_1 = \frac{e\left[\left(C_2 + \frac{1}{2} C_G\right) V + C_G U + q - \frac{e}{2}\right]}{C_\Sigma} \tag{14}$$

where

$$C_\Sigma = C_1 + C_2 + C_G \tag{15}$$

is the capacitance of the island. In (14) the gate voltage U only appears in the combination $C_G U + q = C_G U - ne$, which leads for all measurable quantities to a periodicity in $C_G U$ with period e, since the integer part of $C_G U/e$ can always be absorbed in n. In semiconductor devices this strict periodicity is usually not met because the gate voltage U weakly influences the tunneling resistances of the junctions.

A straightforward analysis of the rate formula (12) shows that at zero temperature the state with n electrons on the island of the SET transistor is stable with respect to tunneling across the first and second junctions for voltages satifying

$$\begin{aligned} &e(n - \tfrac{1}{2}) < C_G U + (C_2 + \tfrac{1}{2} C_G) V < e(n + \tfrac{1}{2}) \\ &e(n - \tfrac{1}{2}) < C_G U - (C_1 + \tfrac{1}{2} C_G) V < e(n + \tfrac{1}{2}), \end{aligned} \tag{16}$$

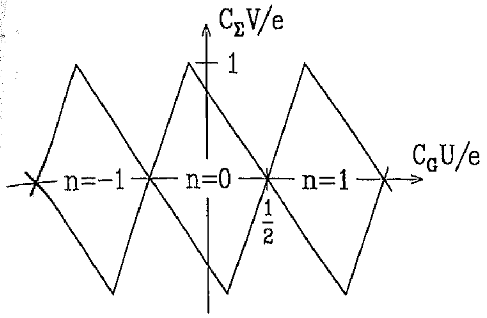


Fig. 4. The stability diagram of a SET transistor with $2C_2 = 10C_G = C_1$

respectively, Hence, in the UV-plane there are rhombic-shaped regions along the U-axis within which the transistor island is charged with a fixed number of excess electrons [cf. Fig. 4]. Inside these rhombi all transitions are suppressed by a Coulomb blockade and no current flows through the device.

For example, near $U = V = 0$ the state $n = 0$ is stable. The inequalities (16) show that whenever the system leaves the stability region of $n = 0$ at a point in the UV-plane with $V \neq 0$ and a tunneling transition, say, to $n = 1$ occurs, the new state is not stable with respect to tunneling across the other junction. Hence, shortly after the first tunneling event an electron leaves the island through the other junction and the system returns to $n = 0$, where the cycle can start again. As a net effect, a current flows through the device. The second tunneling transition of this cycle does not occur at the edge of the Coulomb blockade and part of the change of electrostatic energy is left as kinetic energy of the tunneling electron. Therefore, the transistor is a dissipative element, in contrast to the SEB discussed above which is reversible when the gate voltage is changed slowly.

For voltages V of order E_c/e the current I is very sensitive to U. A small change of the polarization charge $C_G U$ by a fraction of the elementary charge e can change the current from zero to values of the order of E_c/eR_T. This is why the SET transistor can be used as a highly sensitive electrometer [45, 11]. It also could serve as a low-noise amplifier of analog signals. Accordingly, the SET transistor is the basic active element of digital and other applications proposed for SCT [48]. A detailed discussion of the SET transistor is given by Averin and Likharev [6] and further results are presented in the article by Ingold et al. [41]. Since the transistor is a dissipative element, Cooper pairs can usually be transferred only in combined processes involving quasiparticles or environmental modes or both. The complex behavior of the superconducting device is discussed in the article by Maasen van den Brink et al. [33].

At zero temperature and for voltages within the intervals (16), the state with n electrons on the transistor island is stable with respect to tunneling across either junction. However, in the presence of an applied voltage V the state can only be metastable. In fact, Averin and Odintsov [49] have pointed out that second order transitions always lead to a finite current in the presence of an applied voltage. In these co-tunneling events an electron tunnels onto the island while a second electron simultaneously leaves the island across the other junction. Since the charge on the island is only changed virtually, there is no Coulomb barrier for this process. The co-tunneling rate is proportional to $(R_K/R_T)^2$ and hence is a factor R_K/R_Tsmaller than the rate for first order processes. Accordingly, in more complicated multijunction circuits there are co-tunneling events involving N junctions with rates proportional to $(R_K/R_T)^N$.

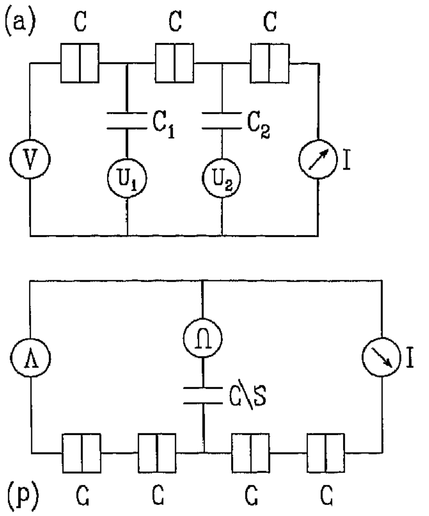


Fig. 5. The circuit diagrams of (a) the single electron pump and (b) the turnstile, which allow for a transfer of electrons one-by-one

Of course, co-tunneling mainly arises when the inequality (1) is only poorly satisfied. However, since the time scale of all SCT processes is proportional to R_K/R_T, very large tunneling resistances severely reduce the speed of devices. That is why a detailed understanding of co-tunneling is of essential importance to the field. Co-tunneling was studied experimentally by Geerligs et al. [50] in small arrays of oxide layer tunnel junctions and by Glattli et al. [16] in a semiconductor SET transistor. A survey of the theory is given by Averin and Nazarov [51]. Electron tunneling across single junctions for arbitrary tunneling resistances R_T is studied in the articles by Zwerger and Scharpf [52] and Scalia et al. [53]. This topic is still open to discussion, in fact, different conclusions on the nature of the crossover from Coulomb blocked to Ohmic conduction are reached in these articles.

More sophisticated multijunction circuits can be built using the SEB or the SET transistor as basic units. Figure 5 shows the circuit diagram of the "pump" and "turnstile" devices fabricated recently by the Saclay and Delft groups [8, 9, 12, 15]. The pump designed by Pothier et al. [9] can be seen as two SEB connected by a tunnel junction. The boxes allow for a control of the input and output of electrons by means of the gate voltages. When the appropriate ac voltages with frequency f are applied to the gates, precisely one electron is transferred per cycle through the device, giving a current

$$I = ef. \tag{17}$$

This relation is the basis of high precision SCT current sources. Since the pump principle employs only reversible processes, it can also be used to transfer Cooper pairs. Experimental results on the pump in the superconducting state are presented in the article by Geerligs et al. [12].

The turnstile designed by Geerligs et al. [8] can be seen as two double junctions connected by a common island. The charging and discharging of this island is controlled by a gate voltage. Again, a current obeying (17) can be generated by means of an ac voltage. In a semiconductor version of the turnstile fabricated by Kouwenhoven et al. [15], the tunneling resistances are modulated via the voltages applied to the Schottky gates. Deviations from (17) mainly arise from finite temperature effects, electron heating, co-tunneling, and moving background charges. These effects must be reduced to achieve a current standard with metrological accuracy. A detailed introduction into the art of manipulating electrons one-by-one is given by Esteve [44] and Urbina et al. [54].

Another class of multijunction circuits studied so far are one-dimensional arrays. Here the work by the Göteborg group recently surveyed by Delsing [55] is most remarkable. Furthermore, two-dimensional arrays show a fascinating interplay between charging effects and collective phenomena. This topic is reviewed in an article by Mooij and Schön [56]. A detailed discussion of these systems would be beyond the scope of this brief introduction to SCT.

4. Conclusion

Single charge tunneling has only recently developed into a field investigated in many laboratories world-wide, partly due to the progress in nanoscale fabrication techniques. Yet this area of research is about to leave its infancy. The main problems still to be overcome have been identified, and it might be appropriate to conclude by speculating about possible applications.

As mentioned above, the SET transistor serves as a highly sensitive electrometer. The sensitivity of existing prototypes already exceeds those of other electrometers by orders of magnitude, and the performance can certainly be improved further. The maximal sensitivity attainable is presently not known, but it should be at least $10^{-5}\, e/\sqrt{\mathrm{Hz}}$. Despite this very high precision, the SET electrometer is for the measurement of electrical charges not quite as revolutionary as the SQUID was for the measurement of magnetic flux. Since there is no analogue of the superconducting flux transformer, the very small input capacitance of the SET electrometer might limit its usefulness.

The pump and turnstile devices demonstrate that SCT can be employed to construct frequency-controlled current sources. The relative uncertainty of the current produced by existing devices is about 10^{-2}, but the error sources are being analysed and improved designs are under investigation. Whether metrological accuracy of 10^{-8} is really achievable is not known, although the prospects are rather promising. Since the ampère is presently derived from the kilogram, a closure of the quantum metrological triangle (see Fig. 6) could ultimately revolutionize the metrological system, and perhaps do away with the last relic of the Bureau International des Poids et Mesures in Sèvres.

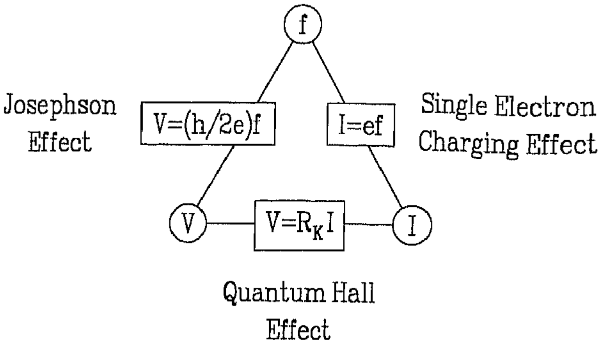


Fig. 6. The quantum metrological triangle formed by the Josephson effect, the quantum Hall effect, and the single electron charging effect

Much of the fascination of SCT derives from the fact that in the future a single bit in an information flow might possibly be represented by a single electron. Although concrete designs are being proposed, the moderate gain of the SET transistor for only a small range of input amplitudes remains a problem, and extensive research would be needed to achieve this goal. However, along the way other striking advances are likely to be made, perhaps even the controlled transfer of fractional charges in semiconductor transistors in the quantum Hall effect regime.

I should like to thank the participants of the NATO Advanced Study Institute on SCT in Les Houches for many inspiring discussions, in particular G.-L. Ingold and the members of the Quantronics Group in Saclay. This work was completed during an extended stay at the Centre d'Etudes de Saclay whose kind hospitality is gratefully acknowledged. Financial support was provided through grants from NATO, the European Community, and the Deutsche Forschungsgemeinschaft. Last but not least, special thanks are due to Springer Verlag for the considerable effort made by its staff to ensure that the original research papers collected in this Special Issue appear in accord with the traditionally fast publication scheme of Zeitschrift für Physik B – Condensed Matter.

References

1. Gorter, C.J.: Physica **17**, 777 (1951)
2. Neugebauer, C.A., Webb, M.B.: J. Appl. Phys. **33**, 74 (1962)
3. Giaever, I., Zeller, H.R.: Phys. Rev. Lett. **20**, 1504 (1968)
4. Lambe, J., Jaklevic, R.C.: Phys. Rev. Lett. **22**, 1371 (1969)
5. Kulik, I.O., Shekhter, R.I.: Zh. Eksp. Teor. Fiz. **68**, 623 (1975) [Sov. Phys. – JETP **41**, 308 (1975)]
6. For recent reviews see: Likharev, K.K.: IBM J. Res. Dev. **32**, 144 (1988); Averin, D.V., Likharev, K.K.: In: Altshuler, B.L., Lee, P.A., Webb, R.A. (eds.) *Quantum effects in small disordered systems*. Amsterdam: Elsevier 1991
7. Schön, G., Zaikin, A.D.: Phys. Rep. **198**, 237 (1990)
8. Geerligs, L.J., Anderegg, V.F., Holweg, P.A.M., Mooij, J.E., Pothier, H., Esteve, D., Urbina, C., Devoret, M.H.: Phys. Rev. Lett. **64**, 2691 (1990)
9. Pothier, H., Lafarge, P., Orfila, P.F., Urbina, C., Esteve, D., Devoret, M.H.: Physica **B169**, 573 (1991)

Haviland, D.B., Kuzmin, L.S., Delsing, P., Likharev, K.K., Claeson, T.: Z. Phys. B – Condensed Matter **85**, 339 (1991)

Lafarge, P., Pothier, H., Williams, E.R., Esteve, D., Urbina, C., Devoret, M.H.: Z. Phys. B – Condensed Matter **85**, 327 (1991)

Geerligs, L.J., Verbrugh, S.M., Hadley, P., Mooij, J.E., Pothier, H., Lafarge, P., Urbina, C., Esteve, D., Devoret, M.H.: Z. Phys. B – Condensed Matter **85**, 349 (1991)

Meirav, U., McEuen, P.L., Kastner, M.A., Foxman, E.B., Kumar, A., Wind, S.J.: Z. Phys. B – Condensed Matter **85**, 357 (1991)

Kouwenhoven, L.P., van der Vaart, N.C., Johnson, A.T., Kool, W., Williamson, J.G., Staaring, A.A.M., Foxon, C.T.: Z. Phys. B – Condensed Matter **85**, 367 (1991)

5. Kouwenhoven, L.P., Johnson, A.T., van der Vaart, N.C., van der Enden, A., Harmans, C.J.P.M., Foxon, C.T.: Z. Phys. B – Condensed Matter **85**, 381 (1991)

5. Glattli, D.C., Pasquier, C., Meirav, U., Williams, F.I.B., Jin, Y., Etienne, B.: Z. Phys. B – Condensed Matter **85**, 375 (1991)

7. Ramdane, A., Faini, G., Launois, H.: Z. Phys. B – Condensed Matter **85**, 389 (1991)

3. Ruggiero, S.T., Barner, J.B.: Z. Phys. B – Condensed Matter **85**, 333 (1991)

9. Scott-Thomas, J.H.F., Field, S.B., Kastner, M.A., Smith, H.I., Antonadis, D.A.: Phys. Rev. Lett. **62**, 583 (1989)

0. See e.g.: van Bentum, P.J.M., van Kempen, H., van de Leemput, L.E.C., Teunissen, P.A.A.: Phys. Rev. Lett. **60**, 369 (1988)

1. Grabert, H., Devoret, M.H. (eds.): Proceedings of the NATO ASI on *Single Charge Tunneling*, Les Houches, March 1991, New York: Plenum Press 1992

2. Widom, A., Megaloudis, G., Clark, T.D., Prance, H., Prance, R.J.: J. Phys. A **15**, 3877 (1982)

3. Likharev, K.K., Zorin, A.B.: J. Low Temp. Phys. **59**, 347 (1985)

4. Ben-Jacob, E., Gefen, Y.: Phys. Lett. **108**A, 289 (1985)

5. Averin, D.V., Likharev, K.K.: J. Low Temp. Phys. **62**, 345 (1986)

6. Nazarov, Yu.V.: Pis'ma Zh. Eksp. Teor. Fiz. **49**, 105 (1989) [Sov. Phys. – JETP Lett. **49**, 126 (1989)]

7. Devoret, M.H., Esteve, D., Grabert, H., Ingold, G.-L., Pothier, H., Urbina, C.: Phys. Rev. Lett. **64**, 1824 (1990)

8. Girvin, S.M., Glazman, L.I., Jonson, M., Penn, D.R., Stiles, M.D.: Phys. Rev. Lett. **64**, 3183 (1990)

9. For a specific example see e.g.: Martinis, J.M., Kautz, R.L.: Phys. Rev. Lett. **63**, 1507 (1989)

0. Grabert, H., Ingold, G.-L., Devoret, M.H., Esteve, D., Pothier, H., Urbina, C.: Z. Phys. B – Condensed Matter **84**, 143 (1991)

31. Averin, D.V., Nazarov, Yu.V., Odintsov, A.A.: Physica B**165/166**, 945 (1990)

32. Falci, G., Bubanja, V., Schön, G.: Europhys. Lett. **16**, 109 (1991) Z. Phys. B – Condensed Matter **85**, 451 (1991)

33. Maasen van den Brink, A., Odintsov, A.A., Bobbert, P.A., Schön, G.: Z. Phys. B – Condensed Matter **85**, 459 (1991)

34. Cleland, A.N., Schmidt, J.M., Clarke, J.: Phys. Rev. Lett. **64**, 1565 (1990)

35. Flensberg, K., Girvin, S.M., Jonson, M., Penn, D.R., Stiles, M.D.: Z. Phys. B – Condensed Matter **85**, 395 (1991)

36. Ingold, G.-L., Nazarov, Yu.V.: In Ref. 21

37. Geerligs, L.J., Anderegg, V.F., van der Jeugd, C.A., Romijn, J., Mooij, J.E.: Europhys. Lett. **10**, 79 (1989)

38. Amman, M., Ben-Jacob, E., Cohen, J.: Z. Phys. B – Condensed Matter **85**, 405 (1991)

39. Ueda, M., Guinea, F.: Z. Phys. B – Condensed Matter **85**, 413 (1991)

40. Grabert, H., Ingold, G.-L., Devoret, M.H., Esteve, D., Pothier, H., Urbina, C.: In: Cerdeira, H.A., Guinea Lôpez, F., Weiss, U. (eds.) Proceedings of the Adriatico Research Conference on *Quantum Fluctuations in Mesoscopic and Macroscopic Systems*, Triest, July 1990. p. 199. Singapore: World Scientific 1991

41. Ingold, G.-L., Wyrowski, P., Grabert, H.: Z. Phys. B – Condensed Matter 85, 443 (1991)

42. Likharev, K.K., Bakhvalov, N.S., Kazacha, G.S., Serdyukova, S.I.: IEEE Trans. Mag. **25**, 1436 (1989)

43. Geigenmüller, U., Schön, G.: Europhys. Lett. **10**, 765 (1989)

44. Esteve, D.: In Ref. 21

45. Fulton, T.A., Dolan, G.J.: Phys. Rev. Lett. **59**, 109 (1987)

46. van Houten, H., Beenakker, C.W.J., Staring, A.A.M.: In Ref. 21

47. Häusler, W., Kramer, B., Masek, J.: Z. Phys. B – Condensed Matter **85**, 435 (1991)

48. Averin, D.V., Likharev, K.K.: In Ref. 21

49. Averin, D.V., Odintsov, A.A.: Phys. Lett. A**140**, 251 (1989)

50. Geerligs, L.J., Averin, D.V., Mooij, J.E.: Phys. Rev. Lett. **65**, 3037 (1990)

51. Averin, D.V., Nazarov, Yu.V.: In Ref. 21

52. Zwerger, W., Scharpf, M.: Z. Phys. B – Condensed Matter **85**, 421 (1991)

53. Scalia, V., Falci, G., Fazio, R., Giaquinta, G.: Z. Phys. B – Condensed Matter **85**, 427 (1991)

54. Urbina, C., Lafarge, P., Pothier, H., Esteve, D., Devoret, M.H.: In: Koch, H. (ed.) Proceedings of SQUID '91, Berlin, June 1991, Heidelberg: Springer (to be published)

55. Delsing, P.: In Ref. 21

56. Mooij, J.E., Schön, G.: In Ref. 21

Quantized Conductance of Point Contacts in a Two-Dimensional Electron Gas

B. J. van Wees
Department of Applied Physics, Delft University of Technology, 2628 CJ Delft, The Netherlands

H. van Houten, C. W. J. Beenakker, and J. G. Williamson,
Philips Research Laboratories, 5600 JA Eindhoven, The Netherlands

L. P. Kouwenhoven and D. van der Marel
Department of Applied Physics, Delft University of Technology, 2628 CJ Delft, The Netherlands

and

C. T. Foxon
Philips Research Laboratories, Redhill, Surrey RH1 5HA, United Kingdom

(Received 31 December 1987)

Ballistic point contacts, defined in the two-dimensional electron gas of a GaAs-AlGaAs heterostructure, have been studied in zero magnetic field. The conductance changes in quantized steps of $e^2/\pi\hbar$ when the width, controlled by a gate on top of the heterojunction, is varied. Up to sixteen steps are observed when the point contact is widened from 0 to 360 nm. An explanation is proposed, which assumes quantized transverse momentum in the point-contact region.

PACS numbers: 72.20.Jv, 73.40.Cg, 73.40.Lq

As a result of the high mobility attainable in the two-dimensional electron gas (2DEG) in GaAs-AlGaAs heterostructures it is now becoming feasible to study ballistic transport in small devices.[1-6] In metals ideal tools for such studies are constrictions having a width W and length L much smaller than the mean free path l_e. These are known as Sharvin point contacts.[7] Because of the ballistic transport through these constrictions, the resistance is determined by the point-contact geometry only. Point contacts have been used extensively for the study of elastic and inelastic electron scattering. With use of biased point contacts, electrons can be injected into metals at energies above the Fermi level. This allows the study of the energy dependence of the scattering mechanisms.[8] With the use of a geometry containing two point contacts, with separation smaller than l_e, electrons injected by a point contact can be focused into the other contact, by the application of a magnetic field. This technique (transverse electron focusing) has been applied to the detailed study of Fermi surfaces.[9]

In this Letter we report the first experimental study of the resistance of ballistic point contacts in the 2DEG of high-mobility GaAs-AlGaAs heterostructures. The single-point contacts discussed in this paper are part of a double-point-contact device. The results of transverse electron focusing in these devices will be published elsewhere.[10] The point contacts are defined by electrostatic depletion of the 2DEG underneath a gate. This method, which has been used by several authors for the study of 1D conduction,[1,2] offers the possibility to control the width of the point contact by the gate voltage. Control of the width is not feasible in metal point contacts.

The classical expression for the conductance of a point contact in two dimensions (see below) is

$$G=(e^2/\pi\hbar)k_FW/\pi \tag{1}$$

in which k_F is the Fermi wave vector and W is the width of the contact. This expression is valid if $l_e \gg W$ and the Fermi wavelength $\lambda_F \ll W$. The first condition is satisfied in our devices, which have a maximum width $W_{\max} \approx 250$ nm and $l_e = 8.5$ μm. The second condition should also hold when the devices have the maximum width. We expect quantum effects to become important when the width becomes comparable to λ_F, which is 42 nm in our devices. In this way we are able to study the transition from classical to quantum ballistic transport through the point contact.

The point contacts are made on high-mobility molecular-beam-epitaxy-grown GaAs-AlGaAs heterostructures. The electron density of the material is $3.56\times10^{15}/\text{m}^2$ and the mobility 85 $\text{m}^2/\text{V s}$ (at 0.6 K). These values are obtained from the devices containing the studied point contacts. A standard Hall bar geometry is defined by wet etching. Using electron-beam lithography, a metal gate is made on top of the heterostructure, with an opening 250 nm wide (inset in Fig. 1). The point contacts are defined by the application of a negative voltage to the gate. At $V_g = -0.6$ V the electron gas underneath the gate is depleted, the conduction taking place through the point contact only. At this voltage the point contacts have their maximum width $W_{\max}$, about equal to the opening between the gates. By a further decrease of the gate voltage, the width of the point contacts can gradually be reduced, until they are fully

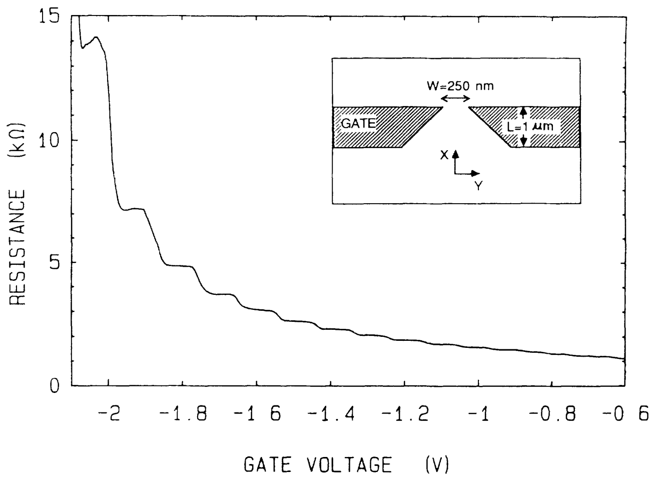


FIG. 1. Point-contact resistance as a function of gate voltage at 0.6 K. Inset: Point-contact layout.

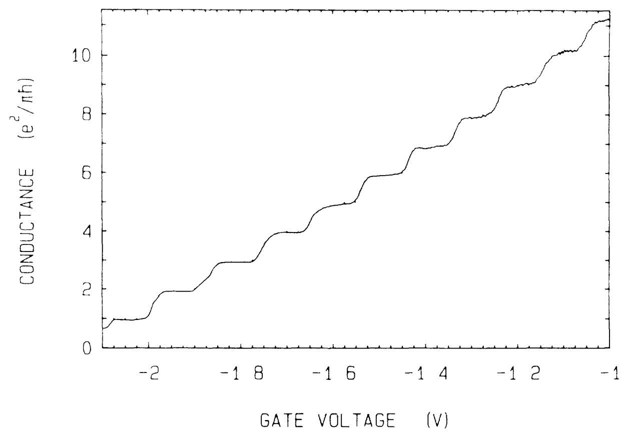


FIG. 2. Point-contact conductance as a function of gate voltage, obtained from the data of Fig. 1 after subtraction of the lead resistance. The conductance shows plateaus at multiples of $e^2/\pi\hbar$.

pinched off at $V_g = -2.2$ V.

We measured the resistance of several point contacts as a function of gate voltage. The measurements were performed in zero magnetic field, at 0.6 K. An ac lockin technique was used, with voltages across the sample kept below kT/e, to prevent electron heating. In Fig. 1 the measured resistance of a point contact as a function of gate voltage is shown. *Unexpectedly, plateaus are found in the resistance.* In total, sixteen plateaus are observed when the gate voltage is varied from -0.6 to -2.2 V. The measured resistance consists of the resistance of the point contact, which changes with gate voltage, and a constant series resistance from the 2DEG leads to the point contact. As demonstrated in Fig. 2, a plot of the conductance, calculated from the measured resistance after subtraction of a lead resistance of 400 Ω, shows clear plateaus at integer multiples of $e^2/\pi\hbar$. The above value for the lead resistance is consistent with an estimated value based on the lead geometry and the resistivity of the 2DEG. We do not know how accurate the quantization is. In this experiment the deviations from integer multiples of $e^2/\pi\hbar$ might be caused by the uncertainty in the resistance of the 2DEG leads. Inserting the point-contact resistance at $V_g = -0.6$ V (750 Ω) into Eq. (1) we find for the width $W_{\max} = 360$ nm, in reasonable agreement with the lithographically defined width between the gate electrodes.

The average conductance increases almost linearly with gate voltage. This indicates that the relation between the width and the gate voltage is also almost linear. From the maximum width $W_{\max}$ (360 nm) and the total number of observed steps (16) we estimate the increase in width between two consecutive steps to be 22 nm.

We propose an explanation of the observed quantization of the conductance, based on the assumption of quantized transverse momentum in the contact constriction. In principle this assumption requires a constriction much longer than wide, but presumably the quantization is conserved in the short and narrow constriction of the experiment. The point-contact conductance G for ballistic transport is given by[7,11]

$$G = e^2 N_0 W(\hbar/2m)\langle |k_x| \rangle. \tag{2}$$

The brackets denote an average of the longitudinal wave vector k_x over directions on the Fermi circle, $N_0 = m/\pi\hbar^2$ is the density of states in the two-dimensional electron gas, and W is the width of the constriction. The Fermi-circle average is taken over discrete transverse wave vectors $k_y = \pm n\pi/W$ $(n = 1, 2, \ldots)$, so that we can write

$$\langle |k_x| \rangle = \frac{1}{2\pi k_F} \int d^2k\, |k_x|\, \delta(k - k_F) \frac{2\pi}{W} \sum_{n=1}^{\infty} \delta\left(k_y - \frac{n\pi}{W}\right). \tag{3}$$

Carrying out the integration and substituting into Eq. (2), one obtains the result

$$G = \sum_{n=1}^{N_c} \frac{e^2}{\pi\hbar}, \tag{4}$$

where the number of channels (or one-dimensional subbands) N_c is the largest integer smaller than $k_F W/\pi$. For

$k_F W \gg 1$ this expression reduces to the classical formula [Eq. (1)]. Equation (4) tells us that G is quantized in units of $e^2/\pi\hbar$ in agreement with the experimental observation. With the increase of W by an amount of $\lambda_F/2$, an extra channel is added to the conductance. This compares well with the increase in width between two consecutive steps, determined from the experiment. Equation (4) may also be viewed as a special case of the multichannel Landauer formula,[12-14]

$$G = \frac{e^2}{\pi\hbar} \sum_{n,m=1}^{N_c} |t_{nm}|^2, \qquad (5)$$

for transmission coefficients $|t_{nm}|^2 = \delta_{nm}$ corresponding to ballistic transport with no channel mixing.

It is interesting to note that this multichannel Landauer formula has been developed to describe the idealized case of the resistance of a quantum wire, connected to massive reservoirs, in which the inelastic-scattering events are thought to take place exclusively. As discussed by Imry,[13] $|t_{nm}|^2 = \delta_{nm}$ corresponds to the case that elastic scattering is absent in the wire also. The fact that the conductance $G = N_c e^2/\pi\hbar$ of such an ideal wire is finite[15] is a consequence of the inevitable contact resistances associated with the connection to the thermalizing reservoirs. The findings described in this Letter may imply that we have realized an experimental system which closely approximates the behavior of idealized mesocopic systems.

In summary we have reported the first measurements of the conductance of single ballistic point contacts in a two-dimensional electron gas. A novel quantum effect is found: The conductance is quantized in units of $e^2/\pi\hbar$.

We would like to thank J. M. Lagemaat, C. E. Timmering, and L. W. Lander for technological support and L. J. Geerligs for assistance with the experiments. We thank the Delft Center for Submicron Technology for the facilities offered and the Stichting voor Fundamenteel Onderzoek der Materie (FOM) for financial support.

[1] T. J. Thornton, M. Pepper, H. Ahmed, D. Andrews, and G. J. Davies, Phys. Rev. Lett. **56**, 1198 (1986).

[2] H. Z. Zheng, H. P. Wei, D. C. Tsui, and G. Weimann, Phys. Rev. B **34**, 5635 (1986).

[3] K. K. Choi, D. C. Tsui, and S. C. Palmateer, Phys. Rev. B **32**, 5540 (1985).

[4] H. van Houten, C. W. J. Beenakker, B. J. van Wees, and J. E. Mooij, in Proceedings of the Seventh International Conference on the Physics of Two-Dimensional Systems, Santa Fe, 1987, Surf. Sci. (to be published).

[5] G. Timp, A. M. Chang, J. E. Cunningham, T. Y. Chang, P. Mankiewich, R. Behringer, and R. E. Howard, Phys. Rev. Lett. **58**, 2814 (1987).

[6] G. Timp, A. M. Chang, P. Mankiewich, R. Behringer, J. E. Cunningham, T. Y. Chang, and R. E. Howard, Phys. Rev. Lett. **59**, 732 (1987).

[7] Yu.V. Sharvin, Zh. Eksp. Teor. Fiz. **48**, 984 (1965) [Sov. Phys. JETP **21**, 655 (1965)].

[8] For a review, see I. K. Yanson and O. I. Shklyarevskii, Fiz. Nizk. Temp. **12**, 899 (1986) [Sov. J. Low Temp. Phys. **12**, 509 (1986)].

[9] P. C. van Son, H. van Kempen, and P. Wyder, Phys. Rev. Lett. **58**, 1567 (1987).

[10] H. van Houten, B. J. van Wees, J. E. Mooij, C. W. J. Beenakker, J. G. Williamson, and C. T. Foxon (to be published).

[11] I. B. Levinson, E. V. Sukhorukov, and A. V. Khaetskii, Pis'ma Zh. Eksp. Teor. Fiz. **45**, 384 (1987) [JETP Lett. **45**, 488 (1987)].

[12] R. Landauer, IBM J. Res. Dev. **1**, 223 (1957); R. Landauer, Phys. Lett. **85A**, 91 (1981).

[13] M. Buttiker, Y. Imry, R. Landauer, and S. Pinhas, Phys. Rev. B **31**, 6207 (1985). For a survey, see Y. Imry, in *Directions in Condensed Matter Physics,* edited by G. Grinstein and G. Mazenko (World Scientific, Singapore, 1986), Vol. 1, p. 102.

[14] D. S. Fisher and P. A. Lee, Phys. Rev. B **23**, 6851 (1981).

[15] The originai Landauer formula (Ref. 12) containing the ratio of transmission and reflection coefficients does give an infinite conductance for a perfect system. However, this formula excludes contributions from the contact resistances.

Controlled fabrication of metallic electrodes with atomic separation

A. F. Morpurgo and C. M. Marcus[a)]
Department of Physics, Stanford University, Stanford, California 94305-4060

D. B. Robinson
Department of Chemistry, Stanford University, Stanford, California 94305-5080

(Received 19 November 1998; accepted for publication 8 February 1999)

We report a technique for fabricating metallic electrodes on insulating substrates with separations on the 1 nm scale. The fabrication technique, which combines lithographic and electrochemical methods, provides atomic resolution without requiring sophisticated instrumentation. The process is simple, controllable, reversible, and robust, allowing rapid fabrication of electrode pairs with high yield. We expect the method to prove useful in interfacing molecular-scale structures to macroscopic probes and electronic devices. © *1999 American Institute of Physics.* [S0003-6951(99)04614-8]

Rapid advances in the ability to manipulate[1–3] and measure[4–7] matter at the level of single atoms and molecules suggest that future technology may allow the fabrication of electronic devices whose core consists of one or a few molecules. This possibility offers important technological advantages beyond a simple reduction in size, as single molecules can be designed and synthesized to perform a variety of specific electronic functions including molecular switches,[8] rectifiers,[9] magnetic and optically bistable systems,[10] and even molecular transistors,[11] allowing electronic functionality to be incorporated into chemical synthesis. However, what currently limits the systematic investigation of nanometer-scale electronic elements as well as their use as a viable technology (i.e., molecular electronics[12]) is the absence of a simple means of interfacing very small objects such as single molecules to macroscopic structures and devices.

At present, experiments probing the electrical properties of single atoms or molecules require either sophisticated techniques based on scanning probe microscopy, or special contacting schemes which often limit experimental flexibility. The latter is illustrated by the clever recent experiments measuring the electrical conductance of benzene–dithiol molecules using mechanical break junctions to provide two metallic contacts.[13] This approach works well but is not readily adapted to include electrostatic gates, a feature that would broaden the experimental possibilities. On the other hand, even the best conventional lithographic methods[14] cannot controllably produce electrodes separated by a few nanometers or less, which are necessary to contact most molecules of interest.

In this letter, we report a technique that readily allows the fabrication of pairs of metallic electrodes with atomic scale separation on an insulating substrate. The crucial innovation of this technique, which is based on standard lithography combined with electrochemical deposition, is active monitoring and control of the separation between electrodes *during the fabrication process*. The simplicity and robustness of the technique suggests that large-scale implementation for the purpose of nanoelectronic device fabrication should be possible.

The technique involves two main steps, as illustrated in Fig. 1. First, metallic electrodes are prepared using conventional microfabrication [Fig. 1(a)]. The separation between electrodes at this stage is not critical. In the second step, metal is electrodeposited on top of the existing pattern from an electrolyte solution [Fig. 1(b)]. This results in an increase in the size of the electrodes, and hence a decrease in their separation [Fig. 1(c)]. By measuring the electrical resistance between the two electrodes, we are able to monitor their separation once this distance becomes very small. In practice, monitoring the resistance signal allows controlled deposition with atomic-scale resolution. The process can be reversed to controllably widen gaps with similar accuracy. In fact, one can deposit until the electrodes are in contact and subsequently electrodissolve the metal to reopen the gap.

Examples of electrode pairs fabricated by this technique are shown in Fig. 2. Coarsely spaced Ti/Au (15 nm/35 nm) electrodes were patterned on a thermally oxidized silicon substate electron-beam lithography and liftoff. Initial spacings were in the range 50–400 nm. Samples were then placed in an aqueous solution consisting of 0.01 M potassium cyanaurate [$KAu(CN)_2$], and a buffer (pH 10) composed of 1 M potassium bicarbonate ($KHCO_3$) and 0.2 M potassium hydroxide. In the deposition reaction, the cyanau-

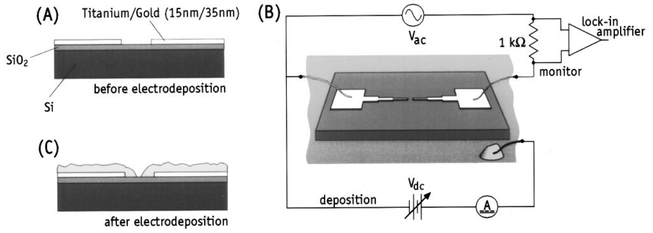


FIG. 1. Fabrication of nanoelectrodes consists of two main steps: (a) Electrodes with large separation are fabricated by conventional lithography. (b) Metal is electrodeposited onto the electrodes, reducing their separation. V_{dc} controls electrodeposition while V_{ac} is used to monitor the conductance and thus the separation between the electrodes. Reversing V_{dc} allows material to be removed rather than deposited. (c) When deposition is stopped before the electrodes touch, separations on the 1 nm scale are obtained reproducibly.

[a)] Electronic mail: cmarcus@stanford.edu

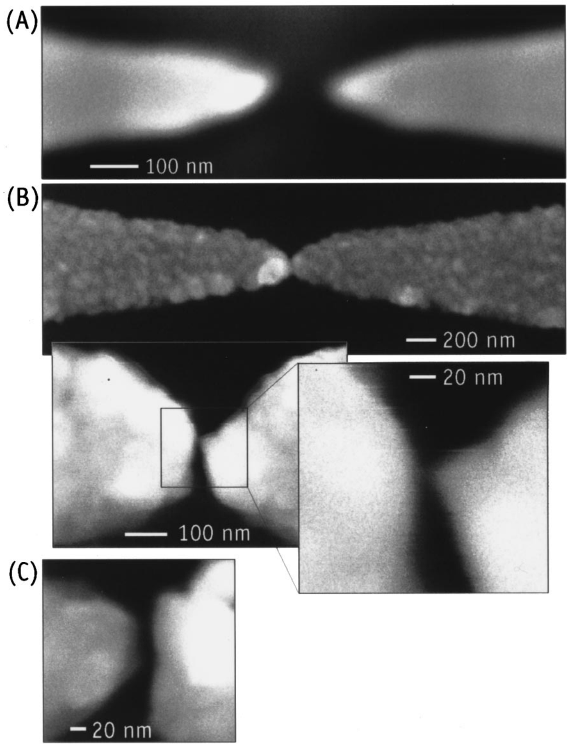


FIG. 2. SEM images before and after electrodeposition (scale bars show dimensions). (a) Electrodes before electrodeposition. (b) Electrodes after electrodeposition. The resolution of the SEM is 5 nm, not sufficient to resolve the gap. (c) Electrodes in which the gap was reopened by electrodissolution, by reversing V_{dc} following an intentional short circuiting (contacting) in a previous electrodeposition process.

rate ion accepts an electron from the electrode and liberates the cyanide ligands, leaving a neutral gold atom at the surface. A gold pellet, 2–3 mm in diameter, was immersed in the solution to act as a counterelectrode. Thin gold wires (25 μm diam, with ~3–4 mm of length in contact with the solution) were used to connect the patterned electrodes and the counterelectrode to the electrical circuit shown in Fig. 1(b). The complete circuit simultaneously serves to drive the electrodeposition process as well as monitor the interelectrode resistance.

During electrodeposition, a voltage bias of −0.5 to −0.6 V was applied to both electrodes relative to the counterelectrode, inducing a deposition current of 2–3 μA, resulting in gold plating at a lateral rate of ~1 Å/s. A number of values for the deposition current were used successfully and no effort has been made yet to optimize the process. The resistance between the two electrodes was measured by applying a 4 mV alternating current (ac) bias at 1 Hz across the electrodes and measuring the ac ''monitor'' current through a 1 kΩ series resistor using a lock-in amplifier [Fig. 1(b)].[15]

Three phases of electrodeposition corresponding to different ranges of electrode separation can be identified from the time evolution of the monitor current. In the first phase, when the electrodes are far apart, the ac monitor current (~20 nA) is small and roughly constant [Fig. 3(a)]. This current is proportional to the immersed surface area of the electrodes (dominated by the surfaces of the 25 μm gold wires) and results from the ac modulation of the direct current (dc) deposition current. The second phase is marked by a sudden increase of the monitor current [Fig. 3(a), inset]. At this point the electrodes are already very close, less than 5 nm, as shown below. The additional current observed in this phase is presumably due to direct tunneling between the contacts, enhanced by the screening effect of ions in the gap, which reduces the height of the tunnel barrier.[16] The third phase, when the contacts finally touch, is marked by a sudden jump in the monitor current, followed by its saturation at a value given by the applied voltage divided by the ~1 kΩ series resistance.

During the second phase of electrodeposition, when the electrodes are very close together but not yet touching, the monitor current is extremely sensitive to electrode distance, enabling control of the separation on an atomic scale. This is illustrated by Fig. 3(c), in which the deposition rate was reduced by a factor of 50 (by reducing the deposition current to ~50 nA) following the increase in monitor current. Using such small deposition currents allows the first atom(s) connecting the two electrodes to be resolved. These first atoms bridging the gap between the electrodes give rise to jumps in the monitor current corresponding to steps of ~2 e^2/h in the conductance [Fig. 3(c), left inset], as expected for a single gold atom,[7] which has a single electronic valence state available for conduction. Typically, only one or two steps of this magnitude are observed, followed by larger jumps presumably originating from clusters of atoms close to the contact point reassembling themselves into more energetically favorable configurations. These steps are similar to those seen in electrodeposited Cu nanowires made using a scanning tunneling microscope.[17]

The appearance of sharp steps in the monitor current associated with atomic conduction allows two important conclusions to be drawn. First, that this controlled deposition technique has atomic-scale resolution, so that it can be used to fabricate electrodes with ~1 nm separation reliably. Second, the steps unambiguously mark when the two electrodes touch; if electrodeposition is stopped at any earlier stage it is assured that the electrodes are not in direct contact.

We have fabricated many pairs of electrodes, stopping electrodeposition when the increase in the monitor current was first detected, and subsequently imaged the samples using a scanning electron microscope (SEM). Neither the SEM (Fig. 2) nor atomic force microscopy could resolve gap clearly, but placed consistent upper limits of 5 nm on the separation. Electrical resistances between such pairs of electrodes (measured using a 0.1 V bias in air after the fabrication) were between 1 and 30 GΩ, and in a few cases as low as 0.5 GΩ, whereas unplated electrodes on the same substrate had resistances above several hundred gigaohms, limited by the noise of the measurement. These values are consistent with electronic tunneling through a gap of roughly 1 nm.[18]

We emphasize that no tuning of fabrication parameters was needed to achieve the present results, demonstrating the robustness of the technique. Alternative strategies have been reported recently[19] capable of feature sizes approaching those reported here, however, the present method offers several advantages including extremely small gaps, high yield (approaching 100%) at gap sizes down to ~1 nm, relatively short fabrication time, and simple, readily available instrumentation.

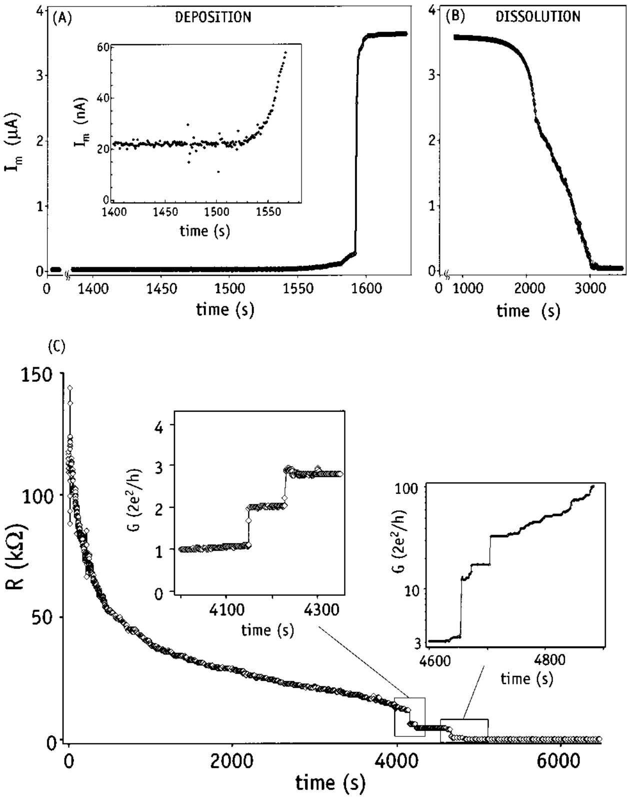


FIG. 3. Time evolution of the ac monitor current during (a) rapid electrodeposition and (b) electrodissolution. Three phases of electrodeposition can be identified. (1) In this example, for times before ~1540 s, a small ac monitor current is measured when the electrodes are well separated. (2) For times between ~1540 and 1590 s, a continuously increasing monitor current appears as the electrodes approach one another at the nanometer scale. (3) At ~1590 s, a sudden jump in the monitor current is observed as the electrodes make contact, followed by saturation. The time evolution is reversed for dissolution. (c) Time evolution of the resistance R between electrodes for slow deposition [roughly 50 times slower than in (a)]. Conductance steps close to 2 e^2/h (the expected value for Au atoms) are visible in the left inset. Following initial contact, plateau-like features and steps in the conductance on the order of a few e^2/h persist as the contact between electrodes continues to increase in size at the atomic scale (right inset).

Because this process can employ techniques and instruments that are currently in use in a variety of industries, including microelectronics manufacturers (deep-ultraviolet lithography and electroplating), it may be readily realized in an industrial setting. Note also that electronic feedback can easily be incorporated into the monitoring scheme, allowing the electrodeposition rate to be adjusted as a function of the resistance between electrodes and then stopped at a specified separation. This type of feedback control lends itself to parallel operation and provides a means of fabricating many structures at the same time.

The authors thank C. E. D. Chidsey for useful discussions and for the use of equipment in his laboratory. Research supported by the National Science Foundation PECASE program, DMR-9629180-1, and the Stanford Center for Materials Research, NSF-MRSEC.

[1] D. M. Eigler and E. K. Scweizer, Nature (London) **344**, 524 (1990).

[2] M. F. Crommie, C. P. Lutz, and D. M. Eigler, Science **262**, 218 (1993).

[3] J. K. Gimzewski, C. Joachim, R. R. Schlittler, V. Langlais, H. Tang, and I. Johannsen, Science **281**, 531 (1998).

[4] F. F. Fan and A. J. Bard, Science **267**, 871 (1995).

[5] C. Joachim, J. K. Gimzewski, R. R. Schlittler, and C. Chavy, Phys. Rev. Lett. **74**, 2102 (1995).

[6] A. Yazdani, B. A. Jones, C. P. Lutz, M. F. Crommie, and D. M. Eigler, Science **275**, 1767 (1997).

[7] E. Scheer, N. Agrait, J. C. Cuevas, A. Levy Yeyati, B. Ludoph, A. Martin-Rodero, G. R. Bollinger, J. M. v. Ruitenbeek, and C. Urbina, Nature (London) **394**, 154 (1998).

[8] M. P. O'Neil, M. P. Niemczyk, W. A. Svec, D. Gosztola, G. L. Gaines III, and M. R. Wasielewski, Science **257**, 63 (1992).

[9] A. Aviram and M. A. Ratner, Chem. Phys. Lett. **29**, 277 (1974); M. Pomerantz, A. Aviram, R. A. McCorkle, L. Li, and A. G. Schrott, Science **255**, 1115 (1992).

[10] O. Kahn and C. J. Martinez, Science **279**, 44 (1998).

[11] C. Joachim and J. K. Gimzewski, Chem. Phys. Lett. **265**, 353 (1997); S. J. Tans, A. R. M. Verschuren, and C. Dekker, Nature (London) **393**, 49 (1998).

[12] For a very recent overview see J. Gimzewski, Phys. World **11**, 29 (1998).

[13] M. A. Reed, C. Zhou, C. J. Muller, T. P. Burgin, and J. M. Tour, Science **278**, 252 (1997).

[14] See for instance L. L. Sohn, C. T. Black, M. Eriksson, M. Crommie, and H. Hess, in *Mesoscopic Electron Transport*, Nato ASI Series, edited by L. L. Sohn, L. P. Kouwenhoven, and G. Schön (Kluwer, Dordrecht, 1997), and references therein.

[15] Similar *in situ* monitoring of the electrical properties of a device to control its fabrication has been used previously, see for instance E. S. Snow and P. M. Campbell, Science **270**, 1639 (1995); Y. Nakamura, D. L. Klein, and J. S. Tsai, Appl. Phys. Lett. **68**, 275 (1996); T. Schmidt, R. Martel, R. L. Sandstrom, and P. Avouris, *ibid.* **73**, 2173 (1998).

[16] See, e.g., W. Schmickler and D. Henderson, J. Electroanal. Chem. Interfacial Electrochem. **290**, 283 (1990).

[17] C. Z. Li and N. J. Tao, Appl. Phys. Lett. **72**, 894 (1998).

[18] Y. Kuk, in *Scanning Tunneling Microscopy*, edited by J. A. Stroscio and W. J. Kaiser (Academic, San Diego, 1993), p. 281.

[19] D. L. Klein, P. L. McEuen, J. E. Bowen Katari, R. Roth, and A. P. Alivisatos, Appl. Phys. Lett. **68**, 2574 (1996); A. Bezryadin and C. Dekker, J. Vac. Sci. Technol. B **15**, 793 (1997); A. Bezryadin, C. Dekker, and G. Schmid, Appl. Phys. Lett. **71**, 1273 (1997).

Microfabrication of a mechanically controllable break junction in silicon

C. Zhou, C. J. Muller, M. R. Deshpande, J. W. Sleight, and M. A. Reed
Center for Microelectronic Materials and Structures, Yale University, P.O. Box 208284, New Haven, Connecticut 06520-8284

(Received 6 April 1995; accepted for publication 20 June 1995)

We present a detailed description of the fabrication and operation at room temperature of a novel type of tunnel displacement transducer. Instead of a feedback system it relies on a large reduction factor assuring an inherently stable device. Stability measurements in the tunnel regime infer an electrode stability within 3 pm in a 1 kHz bandwidth. In the contact regime the conductance takes on a discrete number of values when the constriction is reduced atom by atom. This reflects the conduction through discrete channels. © *1995 American Institute of Physics.*

Micromachining in silicon is an ongoing effort to provide ever smaller devices used as the active part of a sensor. Currently, it is straightforward to produce suspended beams, small springs, and vibrating or rotating structures on a chip. Engineers can make use of a number of classical transducer phenomena, such as piezoelectricity, piezoresistivity and capacitance changes to convert displacements into an electrical signal. However, the formation of smaller sensors is often obtained at the cost of precision, since the signal of the above mentioned transducer phenomena scale with size. In contrast to classical transducers, a tunnel transducer[1] (e.g., an STM) is compatible with further miniaturization and possesses an astonishing sensitivity to displacements. When a vacuum tunnel gap between two metallic electrodes is increased by 1 Å, the tunnel resistance increases approximately by an order of magnitude. This has been realized by a number of groups who have used tunnel sensors in devices.[2] The extreme sensitivity of these sensors on positional displacements however implies that the practical range of operation is limited to distances smaller than 5 Å since at larger distances the resistance becomes almost infinite and unmeasurable.

In conventional STM embodiments, one electrode is usually mounted on a flexible lever, which can be moved by an electrical signal. The tunnel gap is kept constant with the use of a feedback system, necessary since temperature fluctuations, (acoustic) vibrations or other disturbances will otherwise change the vacuum gap over distances much larger than the practical range. An accelerometer, magnetometer, and an infrared sensor have been successfully developed with these kind of tunnel sensors in feedback operation.[2] Despite these successes we have used a different approach and constructed an inherently stable tunnel sensor. When used as a displacement sensor this device can be fabricated in such a way that the electrode separation during operation remains in the practical range of about 5 Å. Due to the extreme stability of this device it can be operated without feedback; however it may also be used in a feedback loop. In this letter we present the fabrication and operation of this new type of tunnel sensor which was proposed in Ref. 3. It is inherently stable, adjustable, and compatible with silicon technology. Detailed measurements are shown, in both the contact and tunnel regimes.

The principle of operation and a schematic perspective and cross sectional view of the device are shown in Fig. 1. The starting material is a $\langle 100\rangle$ oriented 250 μm thick silicon wafer with an oxide layer of 400 nm. Standard electron-beam lithography is used to define a pattern in a PMMA bilayer used for the evaporation of an adhesion layer (10 Å Ti) and 800 Å of gold onto the oxide. The gold film has a shape as indicated in Fig. 1(a). Next a photolithographically defined thick layer of aluminum is evaporated everywhere on the oxide except over a distance u, centered around the smallest gold feature. The next step uses the gold and alumi-

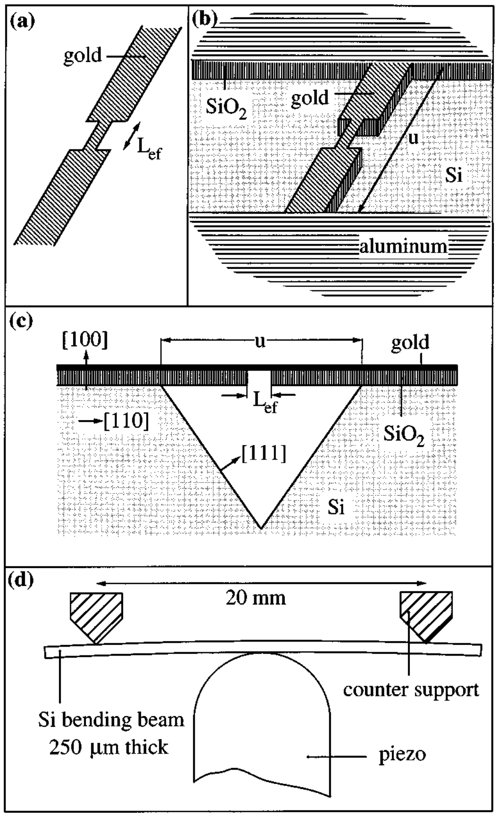


FIG. 1. (a) The gold wire defined by electron-beam lithography. The smallest width of the wire is 100 nm, L_{eff} is about 250 nm. (b) Both the aluminum and gold film are used as an etch mask to etch through the SiO$_2$ into the Si. (c) A cross section along the gold wire after the pit is etched into the silicon. Si etching is stopped at the concave corners and the intersection between the $\langle 111\rangle$ crystallographic surface and the SiO$_2$ edges. (d) The mounting configuration of the silicon bending beam in a break junction setup.

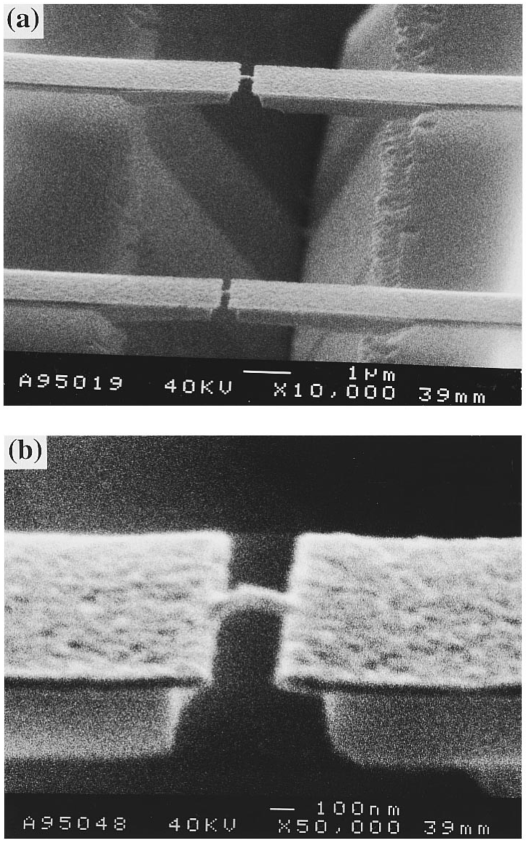


FIG. 2. (a) Two devices suspended above a triangular pit in the Si substrate before the connecting wire is broken in the break junction setup. Each device shows two SiO_2 cantilevers which are covered and bridged by the gold wire. (b) A close-up showing the connecting wire. Before operating the device in the contact or tunnel regime the small connecting wire has to be broken. Some undercut of the gold is present due to the imperfection of the reactive ion etching process.

num films as a mask to etch through the SiO_2 into the Si with a CF_4/O_2 plasma [Fig. 1(b)]. The aluminum is then removed using a standard wet etch. The last step is a wet etch of the exposed Si area using a pyrocatechol-ethylene-diamine mixture.[4] Since the two cantilevers are aligned with the $\langle 110\rangle$ direction in the substrate, a triangular pit is etched into the silicon, bounded by the SiO_2 edges and the $\langle 111\rangle$ surfaces. Rapid undercutting at the convex corners by this etchant assures that the two cantilevers are free standing after the etching process.[5] The final device consists of two small cantilever beams (2.5μm long, 4 μm wide) connected with a 100 nm wide wire over a length L_{eff} [Fig. 1(c)].

The device is mounted against two counter supports, approximately 20 mm apart, in a break junction configuration.[3] A force is exerted on the backside via the piezo element which is moved towards the device using a course adjustment screw [Fig. 1(d)]. The silicon beam is strained, resulting in an elongation of the top layer. The elongation of u is concentrated on L_{eff}, resulting in the fracture of the gold wire while the Si substrate stays intact (even though gold is more ductile than silicon). The piezo element has a maximum elongation of 5 μm and is used for fine adjustment of either atomic size contacts or vacuum barrier tunnel junctions between the fractured gold electrodes. Figure 2 shows a

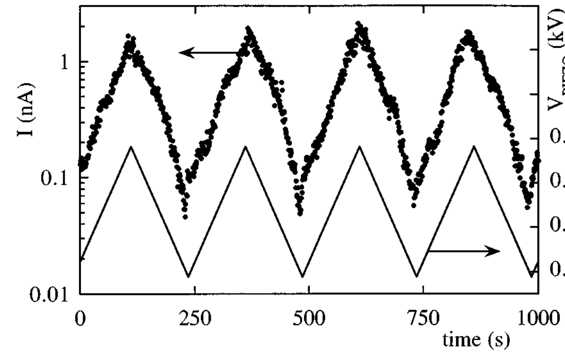


FIG. 3. The piezo voltage is changed in a triangular way (lower curve). The almost linear behavior of the tunnel current on a logarithmic scale reflects the exponential dependence on electrode separation. Note the large time scale, indicating the long term stability of the junction.

SEM photograph of a device before the bridging wire is broken. A 100 nm wire bridging the two cantilevers can be seen, and a slight undercut of the gold is visible. The etched pit into the Si [Fig. 2(a)] is bounded by a relatively rough SiO_2 edge, caused by the photolithography step. Some of the undercut below the SiO_2 layer results from this roughness and enlarges u to about 10 μm.

Experiments are performed at room temperature in a vacuum system (10^{-7}Torr) which uses an oil-free absorption/ion-pump combination in order to reduce contamination of the exposed electrodes with hydrocarbons. Figure 3 illustrates the long term stability and the exponential dependence of the tunnel current I_t on the vacuum barrier gap distance of this device. The junction is biased at 100 mV while a triangular voltage wave is applied to the piezo element (lower curve in Fig. 3). The variation in the piezo length induces a variation in the gap distance resulting in a change of the tunnel resistance (top curve in Fig. 3). The exponential dependence of I_t on the gap distance s is given by $I_t \propto \exp - \alpha\sqrt{\Phi}s$ with $\alpha = 1.025 \text{Å}^{-1}\text{eV}^{-1/2}$ and Φ is the work function of the gold electrodes. As the electrodes are displaced over about 2 Å the tunnel current changes over almost two orders of magnitude. The reason for this exceptional stability is the smallness of u which determines the reduction factor r (the ratio between the piezo elongation and the induced electrode separation). For our devices we estimate $r \simeq 5\times 10^4$.[3] From two devices we experimentally infer, from the known piezo elongation and assuming an exponential dependence of the tunnel current with $\Phi = 4$ eV, $r \simeq 10^4$. The discrepancy of a factor of five may be due to nonuniform strain near the etched pit. In the tunnel regime the current noise amplitude, which depends on the tunnel resistance, is determined at a 100 mV bias for tunnel resistances between 100 kΩ and 10 MΩ in a 1 kHz bandwidth. In this resistance range the experimental value for the current noise amplitude implies about 3 pm fluctuations in the tunnel gap distance. Although we do not know the exact origin of these fluctuations, a detailed noise analysis should include the thermal agitation of the cantilever.[6]

When the electrodes are brought close enough together, a contact is formed. Experiments performed in the contact regime are done in the following way: the contact is reduced

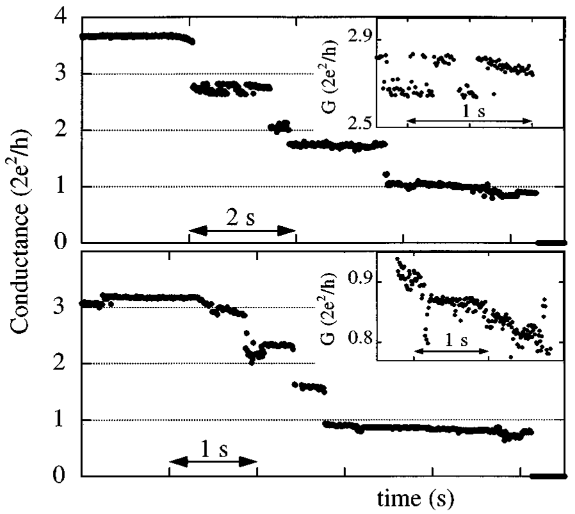


FIG. 4. Two conductance traces recorded when an atomic scale contact reduces its cross section as a function of time. Conductance plateaus are found to be near integer multiples of $2e^2/h$, reflecting the conduction through single channels. The insets show two types of intrinsic noise present in the contact regime.

in size by increasing the piezo voltage until the conductance of the contact is approximately 10 times $2e^2/h$. Then the piezo voltage is fixed, and it is found that the contact relaxes by itself, until eventually a jump to the tunnel regime takes place. Before this jump occurs, the two electrodes may be bridged by a single atom. We tentatively attribute this effect to outdiffusion of atoms, thus decreasing the constriction size. The junction is biased at 26 mV and the current is measured with a sample rate of 100 Hz. Typically the conductance decreases discontinuously as a function of time. Many conductance traces show plateaus near integer multiples of $2e^2/h$, and often the last plateau in the contact regime is near $2e^2/h$ (Fig. 4). After this smallest possible contact, the jump to the tunnel regime results in almost zero conductance (vacuum tunneling only). Upon close inspection, it is seen that the majority of the plateaus are not at exact integers. Backscattering in these metallic point contacts may be responsible for these observations.[7] The description in terms of conductance channels is still valid, although with transmission coefficients slightly different from one or zero.

Conductance noise is clearly present on the plateaus in Fig. 4. This noise is not due to external disturbances and its amplitude is much larger than the measurement accuracy. In general, two different types of noise can be present. The switching of one or a few atoms between energetically equifavorable positions in the contact region can result in closely spaced conductance levels (inset in upper panel of Fig. 4). The high kinetic energy of the atoms at room temperature can drive them between various sites, thus influencing the conductance. Another type of noise has a more random nature (inset in lower panel of Fig. 4). This may be due to small strain variations and small out-of-equilibrium displacements (small compared to the lattice constant) of a group of atoms comprising the contact.

In conclusion, we have presented a new type of displacement transducer, which is inherently stable. We have shown the operation of this device with gold electrodes as well in the contact as in the tunnel regime. The device was shown to be sensitive to positional changes of a single atom.

We would like to thank B.J. Vleeming for his assistance in building the high vacuum system. We acknowledge D.E. Prober and T.N. Todorov for stimulating discussions and the latter for providing us with Ref. 7 prior to publication. C.J.M. gratefully acknowledges the Niels Stensen Foundation for financial support. We acknowledge ARPA for financial support.

[1] M. F. Bocko, K. A. Stephenson, and R. H. Koch, Phys. Rev. Lett. **61**, 726 (1988).

[2] T. W. Kenny, S. B. Waltman, J. K. Reynolds, and W. J. Kaiser, Appl. Phys. Lett. **58**, 100 (1991); H. K. Rockstad, T. W. Kenny, J. K. Reynolds, W. J. Kaiser, and Th. B. Gabrielson, Sens. Actuators A **43**, 107 (1994), and references therein.

[3] C. J. Muller and R. de Bruyn Ouboter, J. Appl. Phys. **77**, 5231 (1995).

[4] G. Kaminsky, J. Vac. Sci. Technol. B **3**, 1015 (1985).

[5] K. E. Petersen, IEEE Trans. Electron Devices **ED-25**, 124 (1978).

[6] The resonance frequency of the cantilever is about 70 MHz. At room temperature it may be driven by k_BT resulting in a deflection of about 1.5 pm; see Th. G. Gabrielson, IEEE Trans. Electron Devices **ED-40**, 903 (1993).

[7] A. M. Bratkovsky, A. P. Sutton, and T. N. Todorov, Phys. Rev. B (to be published).

Single-Walled Carbon Nanotube Electronics

Paul L. McEuen, Michael S. Fuhrer, and Hongkun Park

***Abstract*—Single-walled carbon nanotubes (SWNTs) have emerged as a very promising new class of electronic materials. The fabrication and electronic properties of devices based on individual SWNTs are reviewed. Both metallic and semiconducting SWNTs are found to possess electrical characteristics that compare favorably to the best electronic materials available. Manufacturability issues, however, remain a major challenge.**

***Index Terms*—Field-effect transistors (FETs), interconnections, nanotechnology, nanotube.**

I. Introduction

SINGLE-WALLED carbon nanotubes (SWNTs) are nanometer-diameter cylinders consisting of a single graphene sheet wrapped up to form a tube. Since their discovery in the early 1990s [1] and [2], there has been intense activity exploring the electrical properties of these systems and their potential applications in electronics. Experiments and theory have shown that these tubes can be either metals or semiconductors, and their electrical properties can rival, or even exceed, the best metals or semiconductors known. Particularly illuminating have been electrical studies of individual nanotubes and nanotube ropes (small bundles of individual nantoubes). The first studies on metallic tubes were done in 1997 [3] and [4] and the first on semiconducting tubes in 1998 [5]. In the intervening five years, a large number of groups have constructed and measured nanotube devices, and most major universities and industrial laboratories now have at least one group studying their properties. These electrical properties are the subject of this review. The data presented here are taken entirely from work performed by the authors (in collaboration with other researchers), but they can be viewed as representative of the field.

The remarkable electrical properties of SWNTs stem from the unusual electronic structure of the two-dimensional material, graphene, from which they are constructed [6] and [7]. Graphene—a single atomic layer of graphite—consists of a 2-D honeycomb structure of sp^2 bonded carbon atoms, as seen in Fig. 1(a). Its band structure is quite unusual; it has conducting states at E_f, but only at specific points along certain directions in momentum space at the corners of the first Brillouin zone, as is seen in Fig. 1(b). It is called a zero-bandgap semiconductor since it is metallic in some directions and semiconducting in the others. In an SWNT, the momentum of the electrons moving around the circumference of the tube is quantized, reducing the available states to slices through the 2-D band structure, is illustrated in the Fig. 1(b). This quantization results in tubes that are either one-dimensional metals or semiconductors, depending on how the allowed momentum states compare to the preferred directions for conduction. Choosing the tube axis to point in one of the metallic directions results in a tube whose dispersion is a slice through the center of a cone [Fig. 1(c)]. The tube acts as a 1-D metal with a Fermi velocity $v_f = 8 \times 10^5$ m/s comparable to typical metals. If the axis is chosen differently, the allowed ks take a different conic section, such as the one shown in Fig. 1(d). The result is a 1-D semiconducting band structure, with a gap between the filled hole states and the empty electron states. The bandgap is predicted to be $E_g = 0.9$ eV/d[nm], where d is the diameter of the tube. Nanotubes can, therefor,e be either metals or semiconductors, depending on how the tube is rolled up. This remarkable theoretical prediction has been verified using a number of measurement techniques. Perhaps the most direct used scanning tunneling microscopy to image the atomic structure of a tube and then to probe its electronic structure [8] and [9].

To understand the conducting properties of nanotubes, it is useful to employ the two-terminal Landauer–Buttiker Formula, which states that, for a system with N 1-D channels in parallel: $G = (Ne^2/h)T$, where T is the transmission coefficient for electrons through the sample (see, for example, [10]). For a SWNT at low doping levels such that only one transverse subband is occupied, $N = 4$. Each channel is fourfold degenerate, due to spin degeneracy and the sublattice degeneracy of electrons in graphene. The conductance of a ballistic SWNT with perfect contacts ($T = 1$) is then $4e^2/h = 155$ μS, or about 6.5 kΩ. This is the fundamental contact resistance associated with 1-D systems that cannot be avoided. Imperfect contacts will give rise to an additional contact resistance R_c. Finally, the presence of scatters that give a mean-free path l contribute an Drude-like resistance to the tube, $R_t = (h/4e^2)(L/1)$, where L is the tube length. The total resistance is approximately the sum of these three contributions, $R = h/4e^2 + R_c + R_t$. In the sections below, we will analyze the conducting properties of metal and semiconducting nanotubes to infer the contact resistances, mean-free paths, conductivities, etc. We will concentrate almost exclusively on room temperature behavior. At low temperatures, SWNT devices exhibit a number of interesting quantum phenomena, including single-electron charging, quantum interference, Luttinger liquid behavior, and the Kondo

Manuscript received February 12, 2002; revised February 26, 2002. The work of P. L. McEuen was supported in part by the National Science Foundation (NSF) Center for Nanoscale Systems and the Microelectronics Advanced Research Corporation (MARCO) Focused Center Research Program (FCRP) on Materials, Structures, and Devices which is funded at the Massachusetts Institute of Technology, in part by MARCO under Contract 2001-MT-887 and by the Defense Advanced Research Projects Agency (DARPA) under Grant MDA972-01-1-0035.

P. L. McEuen is with the Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853 USA.

M. S. Fuhrer is with the Department of Physics, University of Maryland, College Park, MD 20742 USA.

H. Park is with the Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138 USA.

Publisher Item Identifier S 1536-125X(02)04581-7.

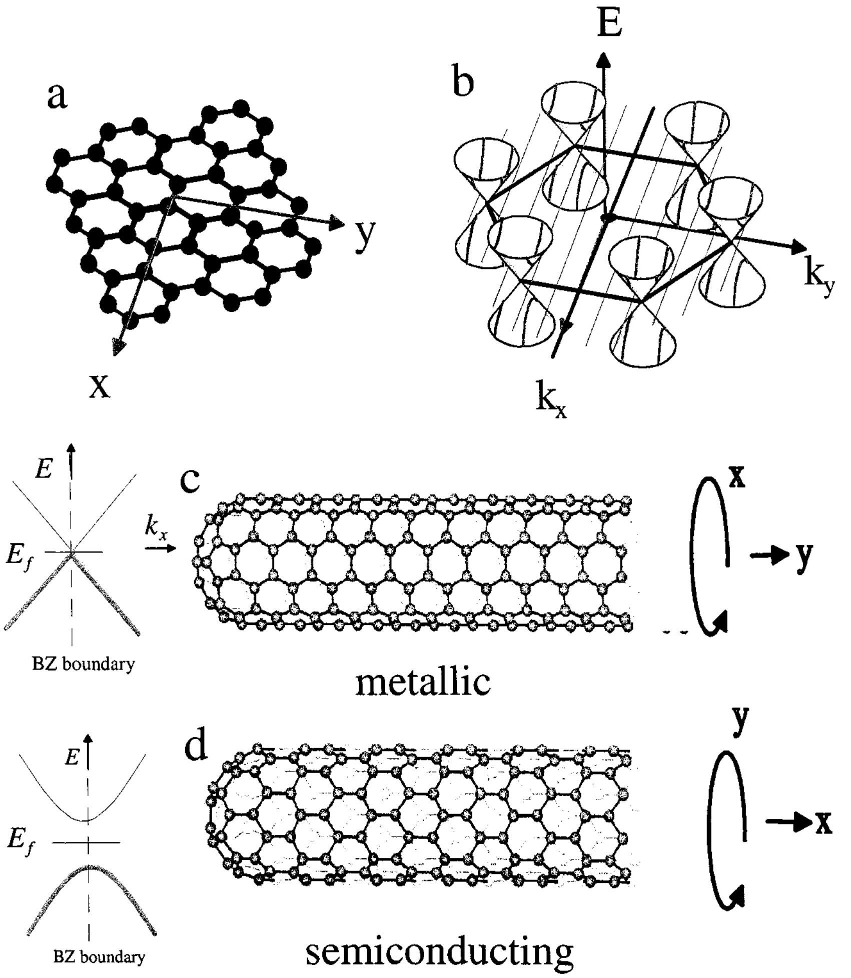


Fig. 1. (a) Lattice structure of graphene, a honeycomb lattice of carbon atoms. (b) Energy of the conducting states as a function of the electron wavevector k. There are no conducting states except along special directions where cones of states exist. (c), (d) Graphene sheets rolled into tubes. This quantizes the allowed ks around the circumferential direction, resulting in 1-D slices through the 2-D band structure in (b). Depending on the way the tube is rolled up, the result can be either (c) a metal or (d) a semiconductor.

effect, but these are not of direct relevance to most device applications. We, therefore, refer the reader to existing reviews for further discussion of these topics [11]–[13].

The critical issues with respect to device applications are twofold. The first is how reproducibly and reliably nanotube devices can be manufactured. Some current approaches to device fabrication are discussed in Section II. The second issue is how the electrical properties of SWNT devices compare to other electronic materials. These properties are described below in Sections III and IV for metallic and semiconducting tubes, respectively. These sections show that devices based on individual SWNTs have remarkable electrical characteristics, making them a very promising new class of electronic materials. The manufacturability challenges, however, are very significant. While advances are being made, controlled, reproducible device fabrication remains an unattained goal. These issues will be discussed in more detail in Section V.

II. Nanotube Growth and Device Fabrication

SWNTs are grown by combining a source of carbon with a catalytic nanostructured material such as iron or cobalt at elevated temperatures. Sources of carbon employed to date include bulk graphite, hydrocarbons, and carbon monoxide. While the details of the growth process are far from understood, the basic elements are now coming into focus. A schematic is shown in Fig. 2(a). At elevated temperatures, the catalyst has a high solubility for carbon. The carbon in the particle links up to form graphene and wraps around the catalyst to from a cylinder. Subsequent growth occurs from the continuous addition of carbon to the base of the tube at the nanoparticle/tube interface. Remarkably, tubes can grow to lengths of hundreds of microns by this process [14].

Creating the proper conditions for growth can done in a variety of ways. From the point of view of device fabrication, the techniques can be divided into categories. In the first category are tubes grown by bulk synthesis techniques that are subsequently deposited on a substrate to make devices ("deposited tubes"). The most common methods for bulk fabrication are arc synthesis [1], [2] and laser assisted growth [15], and commercial sources of SWNTs from these techniques are now available. By controlling the growth conditions, high yields of SWNTs with narrow size distributions can be obtained. Unfortunately, tubes fabricated this way are in the form of a highly tangled "felt" of tubes and bundles of tubes. For electronic devices, these tubes

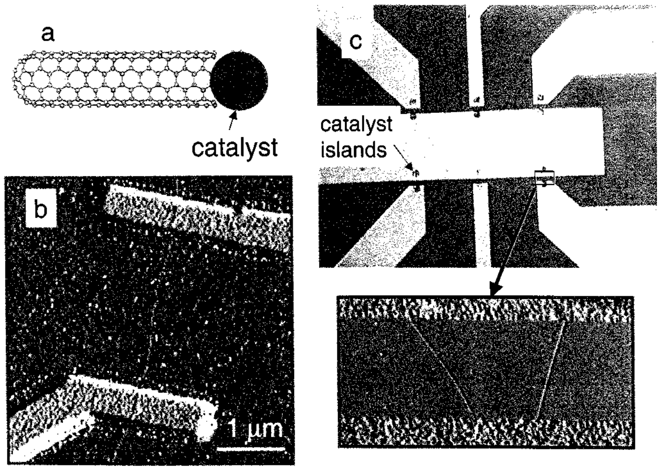


Fig. 2. (a) Schematic of a SWNT growing from a catalyst seed particle. (b) Atomic force microscope images of a single nanotube device fabricated using electron beam lithography. (c) Parallel fabrication of SWNT devices by growth from patterned catalysts and subsequent deposition of arrays of electrodes. The lower panel shows an AFM image of one pair electrodes bridged by two SWNTs.

must be separated, cut into usable sizes, and then deposited on a substrate. This is typically done by ultrasonication in an appropriate solvent to disperse and cut the SWNTs, followed by deposition onto a substrate by spinning or drying. Unfortunately, this is to date an uncontrolled process, producing tubes on the substrate of varying lengths that are often still bundled together. This processing can also induce significant numbers of defects in the tubes. However, new techniques for the wet processing, cutting, and sorting of nanotubes are under constant development [16]–[20].

An alternative approach is to grow the nanotubes directly on the wafer [21]. Currently this is done using chemical vapor deposition (CVD). The catalyst material is placed on the surface of a wafer, which is inserted in a standard furnace at 700 °C–1000 °C in a flow of a carbon source gas such as methane. The tubes grow from the catalyst seeds on the substrate. Engineering the properties of the catalyst and controlling the growth conditions control the properties of the tubes. For example, relatively monodisperse nanoparticle catalysts have been shown to yield SWNTs with a diameter closely matching that of the catalyst particle [22] and [23].

For both deposited and CVD-grown SWNTs, the tubes must be integrated with electrodes and gates on a wafer to make devices. A major challenge is the placement of the tubes relative to lithographically patterned features on the substrate. For both CVD-grown and deposited tubes, techniques have been developed that are satisfactory for research purposes, if not for mass production. Examples are shown in Fig. 2. For the device in Fig. 2(b), SWNTs were grown by CVD and located relative to alignment marks on the surface using an atomic force microscope. Polymethylmethacrylate (PMMA) resist was then spun over the tubes and an electron beam mask was designed, followed by electron beam lithography and liftoff to attach the gold leads [4]. The tubes remain bound to the substrate are unaffected by standard solvents for resist patterning. An alternate approach [21] is to pattern arrays of small catalyst islands from which SWNTs are grown. Electrode arrays are then deposited over the catalyst pads using optical or electron beam lithography. The result is pairs of electrodes with a random number of tubes connecting them, as seen in Fig. 2(c). By adjusting the parameters, a significant fraction of electrodes with only one tube bridging them can be obtained. Equivalent approaches exist to create devices for deposited tubes, with the CVD growth step replaced by a deposition step. An alternative method available for deposited tubes is to pattern the electrodes first and then deposit the tubes on top of the electrodes [3]. This avoids the high-temperature growth step, and chemical modification of the surface [24] and/or electric fields can be used to control, to some degree, the locations of the deposited tubes.

A schematic of the resulting device geometry is shown in the inset to Fig. 5. Source and drain electrodes allow the conducting properties of the nanotube to be measured, and a third gate electrode gate is used to control the carrier density on the tube. Typically, the degenerately doped Si substrate is used as the gate. Nearby metal electrodes [3], an oxidized Al electrode under the tube [25], and even an ionic solution around the tube [26] and [27] have also been employed as gates. When the conductance of the tube as the gate voltage, and hence the charge per unit length of the tube, is varied is measured, two classes of behavior are seen. For some tubes, G is relatively independent of V_g, corresponding to a metallic tube. These are discussed in Section III. For other tubes, a dramatic dependence of G on V_g is seen, indicating semiconducting tubes. These will be discussed in Section IV.

III. Electrical Properties of Metallic Tubes

Devices made from metallic SWNTs were first measured in 1997 [3] and [4], and have been extensively studied since that time. Two-terminal conductances of metallic SWNTs at room temperature can vary significantly, ranging from as small as $\sim$6-kΩ to several megaohms (MΩ). Most of this variation is due to variations in contact resistance between the electrodes and the tube. As techniques for making improved contacts have been developed, the conductances have steadily improved. The best contacts have been obtained by evaporating Au or Pt over the tube, often followed by a subsequent anneal. A number of groups have seen conductances approaching the value, $G = 4e^2/h$, predicted for a *ballistic* nanotube [28] and [29]. An example is shown in Fig. 3, where the dI/dV as a function of V_{sd} is shown for a $\sim$1-μm long SWNT. At low V_{sd}, the conductance is $\sim 2\,e^2/h$, growing to $\sim 3.4e^2/h$ at the temperature is lowered. Assuming perfect contacts, this indicates that the mean-free path is at least ~ 1 μm at room temperature and grows even larger as the device is cooled. A number of other measurements corroborate this conclusion, such as measurements of short tubes where $G = 4e^2/h$ is found [28] and [29], and scanned probe experiments that probe the local voltage drop along the length of the nanotube [30]. This mean-free path corresponds to a room temperature resistivity of $\rho \sim 10^{-6}$ cm. The conductivity of metallic nanotubes can, thus, be equal to, or even exceed, the conductivity of the best metals at room temperature.

These long scattering lengths are in striking contrast to the behavior observed in traditional metals like copper, where scattering lengths are typically on the order of tens of nanometers

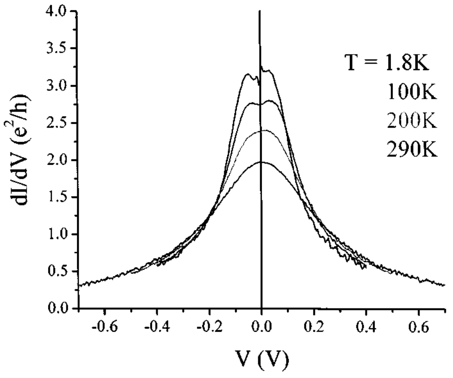


Fig. 3. Differential conductance dI/dV of a metallic SWNT as a function of V_{sd}, at different temperatures. The conductance at low V_{sd} approaches the values for a ballistic SWNT, $4e2/h$. At higher V_{sd}, the conductance drops dramatically due to optic and zone-boundary phonon scattering.

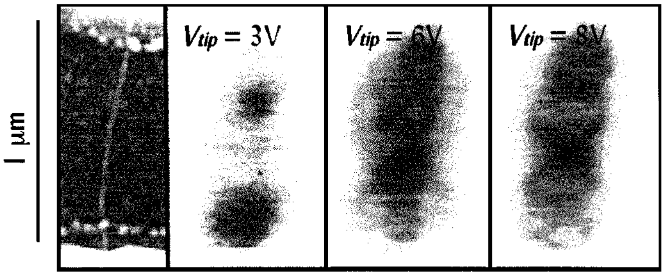


Fig. 4. Left panel: AFM image of a metallic SWNT. Other panels: Scanned gate microscopy of defects in the SWNT at different AFM tip voltages. The conductance through the SWNT is recorded as a function of the tip position. Resonant scattering at defect sites is indicated by rings of reduced conductance (dark) centered on the defects.

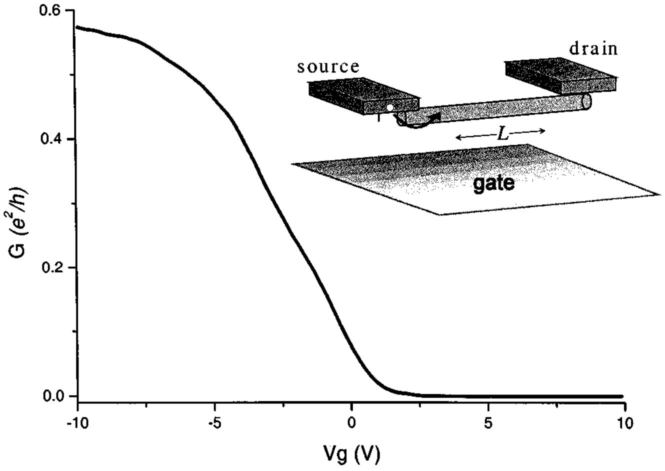


Fig. 5. Conductance G versus gate voltage V_g of a p-type semiconducting SWNT field effect transistor. The device geometry is shown schematically in the inset.

at room temperature, due to phonon scattering. The main difference is the significantly reduced phase space for scattering by acoustic phonons in a 1-D system. At room temperature, acoustic phonons have much less momentum than the electrons at the Fermi energy. In a traditional metal, phonons backscatter electrons through a series of small angle scattering events that eventually reverse the direction of an electron. This is not possible in a 1-D conductor such as a nanotube, where only forward and backward propagation is possible. Note that while the mean-free path is much larger than traditional metals, the conductivity is only comparable to slightly better. This is because the effective density of states in nanotubes is much lower than traditional metals because of the semimetallic nature of graphene.

Optic and zone-boundary phonons have the necessary momentum to backscatter electrons in nanotubes. They are too high in energy ($hf \sim 150$ meV) to be present at room temperature and low V_{sd}. At high source–drain voltages, however, electrons can emit these phonons and efficiently backscatter. This leads to a dramatic reduction of the conductance at high biases, as was first reported by Yao *et al.* [31]. This can be readily seen in the data of Fig. 3. The scattering rate grows linearly with V_{sd}, leading to a saturation of the total current through the tube. This saturation value is $\sim(4e^2/h)\ hf \sim 25\ \mu$A for small-diameter SWNT. This corresponds to a current density of $j = 2.5 \times 10^9$ A/cm^2 for a 1 nm diameter tube. This is orders of magnitude larger than current densities found in present-day interconnects. This large current density can be attributed to the strong covalent bonding of the atoms in the tube. Unlike in metals, there are no low energy defects, dislocations, etc., that can easily lead to the motion of atoms in the conductor.

In addition to phonon scattering, scattering off of static disorder (defects, etc.) is also possible in metallic tubes. A number of sources of scattering have been identified, including physical bends in the tube [32] and [33] and localized electronic states created at defects along the tube [34]. One technique that can give direct information about these scattering centers is scanned gate microscopy (SGM). In this technique, a metallized AFM tip is used as a local gate to probe the conducting properties. Fig. 4 shows a SGM image of a metallic tube [34]. The dark features in the images correspond to locations of defects, which are conjectured to be associated with a bond-rotation defect in the nanotube. Measurements show that these defects are more common in tubes grown at lower temperatures (~700 °C). With proper control of the growth parameters, however, static defects can be minimized so that they are not an important source of scattering at room temperature.

IV. Electrical Properties of Semiconducting Tubes

Semiconducting behavior in nanotubes was first reported by Tans *et al.* in 1998 [5]. Fig. 5 shows a measurement of the conductance of a semiconducting SWNT as the gate voltage applied to the conducting substrate is varied. The tube conducts at negative V_g and turns off with a positive V_g. The resistance change between the on and off state is many orders of magnitude. This device behavior is analogous to a p-type metal–oxide–semiconductor field-effect transistor (MOSFET), with the nanotube replacing Si as the semiconductor. At large positive gate voltages, n-type conductance is sometimes observed, especially in larger-diameter tubes [35] and [36]. The conductance in the n-type region is typically less than in the p-type region because of the work function of the Au electrodes. The Au Fermi level aligns with the valence band of the SWNT, making a p-type contact with a barrier for the injection of electrons.

Semiconducting nanotubes are typically p-type at $V_g = 0$ because of the contacts and also because chemical species, particularly oxygen, adsorb on the tube and act as weak p-type dopants. Experiments have shown that changing a tube's chemical environment can change this doping level—shifting the voltage at which the device turns on by a significant amount [37] and [38]. This has spurred interest in nanotubes as chemical sensors. Adsorbate doping can be a problem for reproducible device behavior, however. In air, a large hysteresis in G versus V_g is observed, with threshold voltage shifts of many volts common. In addition, the threshold voltage is very sensitive to the processing history of the device—for example, heating or exposure to UV radiation drives off oxygen [39], lowering the p-doping level of the device. Controlling adsorbate doping is an important challenge to be addressed. Recent work by the group at IBM has taken important steps in this direction [40].

Controlled chemical doping of tubes, both p- and n-type, has been accomplished in a number of ways. N-type doping was first done using alkali metals that donate electrons to the tube. This has been used to create n-type transistors [38], [41], [42], p-n junctions [43], and p-n-p devices [44]. Alkalai metals are not air-stable, however, so other techniques are under development, such as using polymers for charge-transfer doping [45]. While these techniques are progressing rapidly, we will concentrate here on tubes with no additional doping (beyond uncontrolled doping by adsorbates) and the carriers induced by the gate. For simplicity, we will further focus on the p-type conducting regime to get a sense of the basic parameters that characterize electrical transport.

In the data of Fig. 5, the conductance initially rises linearly with V_g as additional holes are added to the nanotube. At higher gate voltages, the conductance stops increasing and instead is constant. This limiting conductance is due both to the tube and to the contact resistance between the metallic electrodes and to the tube. The value of this resistance can vary by orders of magnitude from device to device, but by annealing the contacts, on-state resistances of ~20–50 kΩ can be routinely obtained. In the regime where G grows linear with V_g, the properties of the device can be described by the Drude-type relation $G = C'_g (V_g - V_{go}) \mu/L$, where C'_g is the capacitance per unit length of the tube, V_{go} is the threshold voltage, μ is the mobility. The capacitance per unit length of the tube can be estimated or obtained from other measurements [3], [4], [46]. Using this we can infer the mobility of the tube, μ. We find typical mobilities of 1000–10 000 cm^2/V·s for CVD-grown tubes, with occasional devices having mobilities as high as 20 000 cm^2/V·s. This is significantly higher than the values reported to date in deposited nanotubes [25], [40], [47], [48]. It is also higher than the mobilities in Si MOSFETs, indicating than SWNTs are a remarkably high-quality semiconducting material.

As with metallic tubes, work has also been performed to investigate the nature of the scattering sites in nanotubes. Again, scanned probe techniques has been very useful. A scanned gate microscopy measurement is shown in Fig. 6(a). The tip was biased positively, to locally deplete the carriers (holes) underneath the tip. The bright spots in the image correspond to places where the AFM tip affected the conductance of the sample, producing a map of the barriers to conduction. This data shows that the conductance is limited by a series of potential barriers that the holes see as they traverse the tube. The barriers are likely due local inhomogeneities in the surface potential from adsorbed charges, etc., at or near the tube. At higher densities, however, little effect of the tip was seen, suggesting excellent transport properties. Electric force microscopy [49] can be used to directly probe the voltage drop along the length of the channel; the result is shown if Fig. 6(b). A linear voltage drop corresponding to a resistance of ~9 kΩ/μm is observed, implying a mean-free path of ~0.7 μm, comparable to the mean-free paths in metallic tubes. This result is consistent with the maximum conductances observed for semiconducting SWNTs ($G \sim e^2/h$ for 1-μm long tubes) and the high mobilities discussed above.

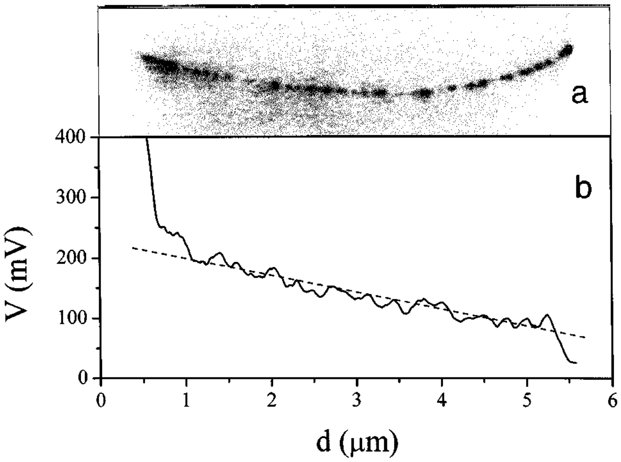


Fig. 6. (a) Scanned gate microscopy showing scattering sites in a p-type semiconducting SWNT. (b) Voltage drop along the length of the source–drain biased semiconducting SWNT, as determined by electric force microscopy. The slope of the voltage drop (dotted line) indicates a resistance per unit length of 9 kΩ/μm.

In order to maximize device performance, the tube gate capacitance C'_g should be maximized. Most experiments to date have used gate oxide thicknesses of hundreds of millimicrons. More recently, researchers have investigated a number of ways to increase the gate coupling, such as using a very thin Al oxide gate [25] or using an electrolyte solution as a gate [26] and [27]. The latter is schematically shown in Fig. 7(a), with the resulting I–V curves at different V_gs shown in Fig. 7(b). Standard FET behavior is seen; the current initially rises linearly with V_{sd} and then becomes constant in the saturation region. The nanotube exhibits excellent characteristics, with a maximum transconductance, $dI/dVg = 20$ μA/V at $V_g = -0.9$ V. Normalizing this to the device width of ~2 nm, this gives a transconductance per unit width of ~10-mS/μm. This is significantly better than current-generation MOSFETs.

The properties of semiconducting SWNTs given above are quite remarkable. Perhaps most surprising is the high mobilities obtained given the small channel width and the simplicity of the fabrication methods employed. This is largely due to the lack of surface states in these devices. As is well known from bulk semiconductors, surface states generally degrade the operating properties of the device, and controlling them is one of the key technological challenges to device miniaturization. A SWNT solves the surface state problem in an elegant fashion. First, it

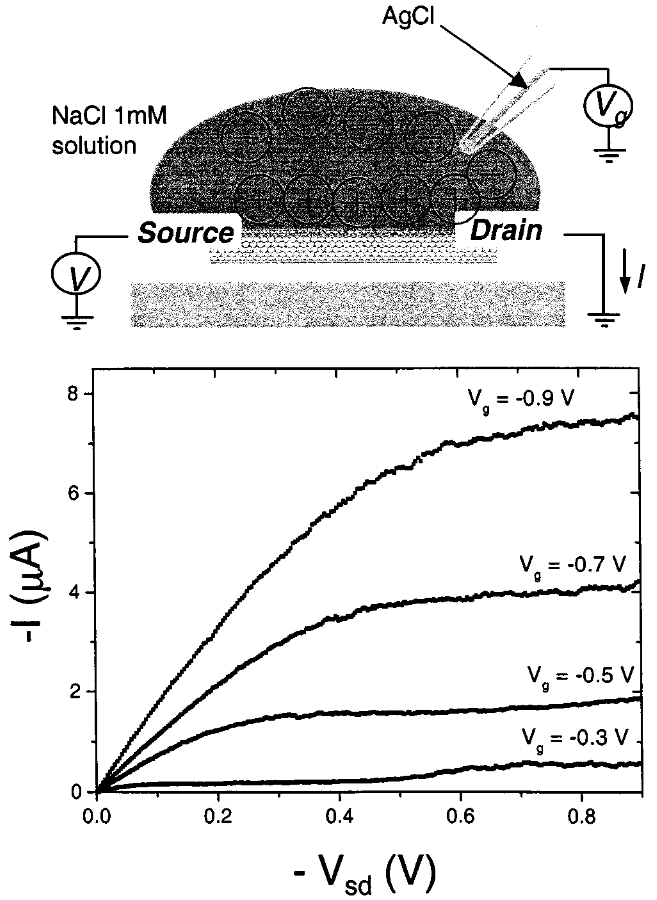


Fig. 7. I–V characteristics at different V_gs for a p-type SWNT FET utilizing an electrolyte gate. A schematic of the measurement geometry is also shown.

begins with a 2-D material with no chemically reactive dangling bonds. It then rids itself of the problem of edges by using the topological trick of rolling itself into a cylinder—which has no edges.

V. Challenges and Future Prospects

The above results show that single nanotube devices possess excellent properties. Metallic tubes have conductivities and current densities that meet or exceed the best metals, and semiconducting tubes have mobilities and transconductances that meet or exceed the best semiconductors. This clearly make them very promising candidates for electronic applications. Opportunities also exist for integrating nanotube electronics with other chemical, mechanical, or biological systems. For example, nanotube electronic devices function perfectly well under biological conditions (i.e., salty water) and have dimensions comparable to typical biomolecules (e.g., DNA, whose width is approximately 2 nm). This makes them an excellent candidate for electrical sensing of individual biomolecules. The are also a host of other device geometries beyond the simple wire and FET structures described above that are under exploration. Examples include the p-n and p-n-p devices mentioned previously [43] and [44], nanotube/nanotube junctions [50]–[52], and electromechanical devices [53] and [54].

Much more challenging is the issue of device manufacturability. Although a great deal of work has been done, the progress to date has been modest. For example, in tube synthesis, the diameter of the tubes can be controlled, but not the chirality. As a result, the tubes are a mixture of metal and semiconductors. In CVD, the general location for tube growth can be controlled by patterning the catalyst material, but the number of tubes and their orientation relative to the substrate are still not well defined. Furthermore, the high growth temperature (900 °C) for CVD tubes is incompatible with many other standard Si processes. The alternative approach, depositing tubes on a substrate after growth, avoids this high temperature issue but suffers from the chirality and positioning limitations discussed above. Furthermore, the wet processing of the tubes may degrade their electrical properties. Efforts are underway to address these issues. For example, techniques to guide tubes to desired locations during growth or deposition using electric fields [55] and/or surface modification [24] are being explored, with some success.

To date, there are no reliable, rapid, and reproducible approaches to creating complex arrays of nanotube devices. This manufacturing issue is by far the most significant impediment to using nanotubes in electronics applications. While there has been significant fanfare around "circuits" made with nanotubes, (see, e.g., the "Breakthrough of the Year" for 2001 in *Science* magazine), in reality the accomplishments to date are a far cry from anything that would impress a device engineer or circuit designer. However, there appear to be no fundamental barriers to the development of a technology. The science of nanotubes has come a long way in five years. With the involvement of the engineering community, perhaps the technology of nanotubes will see similar progress in the next five.

Acknowledgment

The authors would like thank their many collaborators at Cornell University, Ithaca, NY, Harvard University, Cambridge, MA, and at the University of California at Berkeley who participated in the work described here.

References

[1] S. Iijima and T. Ichihashi, "Single-shell carbon nanotubes if 1-Nm diameter," *Nature*, vol. 363, pp. 603–605, 1993.

[2] D. S. Bethune, C. H. Kiang, M. S. Devries, G. Gorman, R. Savoy, J. Vazquez, and R. Beyers, "Cobalt-catalyzed growth of carbon nanotubes with single-atomic-layerwalls," *Nature*, vol. 363, pp. 605–607, 1993.

[3] S. J. Tans, M. H. Devoret, H. Dai, A. Thess, R. E. Smalley, L. J. Georliga, and C. Dekker, "Individual single-wall carbon nanotubes as quantum wires," *Nature*, vol. 386, pp. 474–477, 1997.

[4] M. Bockrath, D. H. Cobden, P. L. McEuen, N. G. Chopra, A. Zettl, A. Thess, and R. E. Smalley, "Single-electron transport in ropes of carbon nanotubes," *Science*, vol. 275, pp. 1922–1925, 1997.

[5] S. J. Tans, R. M. Verschueren, and C. Dekker, "Room temperature transistor based on a single carbon nanotube," *Nature*, vol. 393, pp. 49–52, 1998.

[6] N. Hamada, S. Sawada, and A. Oshiyama, "New one-dimensional conductors—Graphitic microtubules," *Phys. Rev. Lett.*, vol. 68, pp. 1579–1581, 1992.

[7] R. Saito, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, "Electronic structure of chiral graphene tubules," *Appl. Phys. Lett.*, vol. 60, pp. 2204–2206, 1992.

[8] T. W. Odom, H. Jin-Lin, P. Kim, and C. M. Lieber, "Atomic structure and electronic properties of single-walled carbon nanotubes," *Nature*, vol. 391, pp. 62–64, 1998.

[9] J. W. G. Wildoer, L. C. Venema, A. G. Rinzler, R. E. Smalley, and C. Dekker, "Electronic structure of atomically resolved carbon nanotubes," *Nature*, vol. 391, pp. 59–62, 1998.

[10] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge, MA: Cambridge Univ. Press, 1995.

[11] C. Dekker, "Carbon nanotubes as molecular quantum wires," *Physics Today*, vol. 52, p. 22, 1999.

[12] J. Nygard, D. H. Cobden, M. Bockrath, P. L. McEuen, and P. E. Lindelof, "Electrical transport measurements on single-walled carbon nanotubes," *Appl. Phys. A, Solids Surf.*, vol. A69, pp. 297–304, 1999.

[13] Z. Yao, C. Dekker, and P. Avouris, "Electrical transport through single-wall carbon nanotubes," in *Topics in Applied Physics*, M. S. Dresselhaus, G. Dresselhaus, and P. Avouris, Eds. Berlin, Germany: Springer-Verlag, 2001, vol. 80, pp. 147–171.

[14] N. R. Franklin and H. Dai, "An enhanced CVD approach to extensive nanotube networks with directionality," *Adv. Mater.*, vol. 12, pp. 890–894, 2000.

[15] A. Thess, R. Lee, P. Nikolaev, H. Dai, P. Petit, J. Robert, X. Chunhui, L. Young Hee, K. Seong Gon, A. G. Rinzler, D. T. Colbert, G. E. Scuseria, D. Tombnek, J. E. Fischer, and R. E. Smalley, "Crystalline ropes of metallic carbon nanotubes," *Science*, vol. 273, pp. 483–487, 1996.

[16] J. Liu, A. G. Rinzler, H. Dai, J. H. Hafner, R. K. Bradley, P. J. Boul, A. Lu, T. Iverson, K. Shelimov, C. B. Huffman, F. Rodriguez-Macias, Y.-S. Shon, T. R. Lee, D. T. Colbert, and R. E. Smalley, "Fullerene pipes," *Science*, vol. 280, pp. 1253–1256, 1998.

[17] J. Chen, M. A. Hamon, H. Hu, Y. Chen, A. M. Rao, P. C. Eklund, and R. C. Haddon, "Solution properties of single-walled carbon nanotubes," *Science*, vol. 282, pp. 95–98, 1998.

[18] E. T. Mickelson, C. B. Huffman, A. G. Rinzler, R. E. Smalley, R. H. Hauge, and J. L. Margrave, "Fluorination of single-wall carbon nanotubes," *Chem. Phys. Lett.*, vol. 296, pp. 188–194, 1998.

[19] J. Chen, A. M. Rao, S. Lyuksyutov, M. E. Itkis, M. A. Hamon, H. Hu, R. W. Cohn, P. C. Eklund, D. T. Dolbert, R. E. Smalley, and R. C. Haddon, "Dissolution of full-length single-walled carbon nanotubes," *J. Phys. Chem. B*, vol. 105, pp. 2525–2528, 2001.

[20] S. Niyogi, H. Hu, M. A. Hamon, P. Bhowmik, B. Zhao, S. M. Rozenzhak, J. Chen, M. E. Itkis, M. S. Meier, and R. C. Haddon, "Chromatographic purification of soluble single-walled carbon nanotubes (s-SWNT's)," *J. Amer. Chem. Soc.*, vol. 123, pp. 733–734, 2001.

[21] J. Kong, H. T. Soh, A. Cassell, C. F. Quate, and H. Dai, "Synthesis of single single-walled carbon nanotubes on patterned silicon wafers," *Nature*, vol. 395, p. 878, 1998.

[22] Y. Li, W. Kim, Y. Zhang, M. Rolandi, D. Wang, and H. Dai, "Growth of single-walled carbon nanotubes from discrete catalytic nanoparticles of various sizes," *J. Phys.Chem. B*, vol. 105, p. 11 424, 2001.

[23] C. L. Cheung, A. Kurtz, H. Park, and C. M. Lieber, "Diameter controlled synthesis of carbon nanotubes," J. Phys. Chem. B, vol. 105, 2002, to be published.

[24] J. Liu, M. J. Casavant, M. Cox, D. A. Walters, P. Boul, L. Wei, A. J. Rimberg, K. A. Smith, D. T. Colbert, and R. E. Smalley, "Controlled deposition of individual single-walled carbon nanotubes on chemically functionalized templates," *Chem. Phys. Lett.*, vol. 303, pp. 125–129, 1999.

[25] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, "Logic circuits with carbon nanotube transistors," *Science*, vol. 294, pp. 1317–1320, 2001.

[26] M. Kruger, M. R. Buitelaar, T. Nussbaumer, C. Schonenberger, and L. Forro, "Electrochemical carbon nanotube field-effect transistor," *Appl. Phys. Lett.*, vol. 78, pp. 1291–1293, 2001.

[27] S. Rosenblatt, Y. Yaish, J. Park, J. Gore, and P. L. McEuen, "High performance electrolyte gates carbon nanotube transistors," .

[28] L. Wenjie, M. Bockrath, D. Bozovic, J. H. Hafner, M. Tinkham, and P. Hongkun, "Fabry–Perot interference in a nanotube electron waveguide," *Nature*, vol. 411, pp. 665–669, 2001.

[29] J. Kong, E. Yenilmez, T. W. Tombler, W. Kim, H. Dai, R. B. Laughlin, L. Liu, C. S. Jayanthi, and S. Y. Wu, "Quantum interference and ballistic transmission in nanotube electron waveguides," *Phys. Rev. Lett.*, vol. 87, pp. 106 801/1–106 891/4, 2001.

[30] A. Bachtold, M. S. Fuhrer, S. Plyasunov, M. Forero, E. H. Z. Anderson, Z. A. Zettl, and P. L. McEuen, "Scanned probe microscopy of electronic transport in carbon nanotubes," *Phys. Rev. Lett.*, vol. 84, pp. 6082–6085, 2000.

[31] Z. Yao, C. L. Kane, and C. Dekker, "High-field electrical transport in single-wall carbon nanotubes," *Phys. Rev. Lett.*, vol. 84, pp. 2941–2944, 2000.

[32] H. W. C. Postma, M. de Jonge, and C. Dekker, "Electrical transport through carbon nanotube junctions created by mechanical manipulation," *Phys. Rev. B, Condens. Matter*, vol. 62, pp. R10653–R10656, 2000.

[33] D. Bozovic, M. Bockrath, J. H. Hafner, C. M. Lieber, P. Hongkun, and M. Tinkham, "Electronic properties of mechanically induced kinks in single-walled carbon nanotubes," *Appl. Phys. Lett.*, vol. 78, pp. 3693–3695, 2001.

[34] M. Bockrath, L. Wenjie, D. Bozovic, J. H. Hafner, C. M. Lieber, M. Tinkham, and P. Hongkun, "Resonant electron scattering by defects in single-walled carbon nanotubes," *Science*, vol. 291, pp. 283–285, 2001.

[35] J. Park and P. L. McEuen, "Formation of a p-type quantum dot at the end of an n-type carbon nanotube," *Appl. Phys. Lett.*, vol. 79, pp. 1363–1365, 2001.

[36] A. Javey, M. Shim, and H. Dai, "Electrical properties and devices of large-diameter single-walled carbon nanotubes," *Appl. Phys. Lett.*, vol. 80, p. 1064, 2002.

[37] J. Kong, N. R. Franklin, C. Zhou, M. G. Chapline, S. Peng, K. Cho, and H. Dai, "Nanotube molecular wires as chemical sensors," *Science*, vol. 287, p. 622, 2000.

[38] V. Derycke, R. Martel, J. Appenzeller, and P. Avouris, "Carbon nanotube inter- and intramolecular logic gates," *Nano Lett.*, vol. 1, pp. 453–456, 2001.

[39] R. J. Chen, N. R. Franklin, K. Jing, C. Jien, T. W. Tombler, Z. Yuegang, and D. Hongjie, "Molecular photodesorption from single-walled carbon nanotubes," *Appl. Phys. Lett.*, vol. 79, pp. 2258–2260, 2001.

[40] R. Martel, V. Derycke, C. Lavoie, J. Appenzeller, K. K. Chan, J. Tersoff, and P. Avouris, "Ambipolar electrical transport in semiconducting single-wall carbon nanotubes," *Phys. Rev.Lett.*, vol. 87, p. 256 805, 2001.

[41] M. Bockrath, J. Hone, A. Zettl, P. L. McEuen, A. G. Rinzler, and R. E. Smalley, "Chemical doping of individual semiconducting carbon-nanotube ropes," *Phys. Rev. B, Condens. Matter*, vol. 61, pp. R10606–R10608, 2000.

[42] J. Kong, C. Zhou, E. Yenilmez, and H. Dai, "Alkaline metal-doped n-type semiconducting nanotubes as quantum dots," *Appl. Phys. Lett.*, vol. 77, pp. 3977–3979, 2000.

[43] C. Zhou, J. Kong, E. Yenilmez, and H. Dai, "Modulated chemical doping of individual carbon nanotubes," *Science*, vol. 290, pp. 1552–1555, 2000.

[44] J. Kong, J. Cao, and H. Dai, "Chemical profiling of single nanotubes: Intramolecular p–n–p junctions and on-tube single-electron transistors," *Appl. Phys. Lett.*, vol. 80, p. 73, 2002.

[45] J. Kong and H. Dai, "Full and modulated chemical gating of individual carbon nanotubes by organic amine compounds," *J. Phys. Chem. B*, vol. 105, p. 2890, 2001.

[46] D. H. Cobden, M. Bockrath, P. L. McEuen, A. G. Rinzler, and R. E. Smalley, "Spin splitting and even–odd effects in carbon nanotubes," *Phys. Rev. Lett.*, vol. 81, pp. 681–684, 1998.

[47] R. Martel, T. Schmidt, H. R. Shea, T. Hertel, and P. Avouris, "Single- and multi-wall carbon nanotube field-effect transistors," *Appl. Phys. Lett.*, vol. 73, pp. 2447–2479, 1998.

[48] P. L. McEuen, M. Bockrath, D. H. Cobden, Y.-G. Yoon, and S, G. Louie, "Disorder, pseudospins, and backscattering in carbon nanotubes," *Phys. Rev. Lett.*, vol. 83, p. 5098, 1999.

[49] A. Bachtold, M. S. Fuhrer, S. Plyasunov, M. Forero, E. H. Anderson, A. Zettl, and P. I. McEuen, "Scanned probe microscopy of electronic transport in carbon nanotubes," *Phys. Rev. Lett.*, vol. 84, pp. 6082–6085, 2000.

[50] Z. Yao, H. W. C. Postma, L. Balents, and C. Dekker, "Carbon nanotube intramolecular junctions," *Nature*, vol. 402, p. 273, 1999.

[51] J. Lefebvre, R. D. Antonov, M. Radosavljevic, J. F. Lynch, M. Llaguno, and A. T. Johnson, "Single-wall carbon nanotube based devices," *Carbon*, vol. 38, pp. 1745–1749, 2000.

[52] M. S. Fuhrer, J. Nygård, L. Shih, M. Forero, Y.-G. Yoon, M. S. C. Mazzoni, H. J. Choi, J. Ihm, S. G. Louie, Z. A. Zettl, and P. L. McEuen, "Crossed nanotube junctions," *Science*, vol. 288, pp. 494–497, 2000.

[53] T. Rueckes, K. Kim, E. Joselevich, G. Y. Tseng, C. L. Cheung, and C. M. Lieber, "Carbon nanotube-based nonvolatile random access memory for molecular computing," *Science*, vol. 289, pp. 94–97, 2000.

[54] T. W. Tombler, Z. Chongwu, L. Alxseyev, K. Jing, D. Hongjie, L. Lei, C. S. Jayanthi, T. Meijie, and W. Shi-Yu, "Reversible electromechanical characteristics of carbon nanotubes under local-probe manipulation," *Nature*, vol. 405, pp. 769–772, 2000.

[55] Y. Zhang, A. Chang, J. Cao, Q. Wang, W. Kim, Y. Li, N. Morris, E. Yenilmez, J. Kong, and H. Dai, "Electric-field-directed growth of aligned single-walled carbon nanotubes," *Appl. Phys. Lett.*, vol. 79, pp. 3155–3157, 2001.

Paul L. McEuen was born in Norman, OK, in 1963. He received the B.S. degree in engineering physics from the University of Oklahoma in 1985 and the Ph.D.degree in applied physics from Yale University, New Haven, CT, in 1991.

He was a Postdoctoral Researcher at the Massachusetts Institute of Technology, Cambridge, before joining the University of California at Berkeley in 1992, where he was an Assistant Professor and later Associate Professor of physics and a researcher at LBNL. In 2001, he joined the faculty of Cornell University, Ithaca, NY, as a Professor of physics. His research examines the science and technology of nanostructures, and has included studies of nanotubes, quantum dots, and single molecules. He also develops advanced measurement techniques to probe nanometer-scale systems.

Michael S. Fuhrer received the B.S. degree in physics from the University of Texas at Austin in 1990 and the Ph.D. degree in physics from the University of California at Berkeley in 1998.

Since 2000, he has been an Assistant Professor of physics at the University of Maryland, College Park. His research interests lie in the electronic and electromechanical properties of nanostructures, including carbon nanotubes, nanowires, and new two-dimensional nanostructures.

Hongkun Park was born in Seoul, Korea, in 1967. He received the B.S. degree in chemistry from Seoul National University, Seoul, Korea, in 1990 and the Ph.D. degree in physical chemistry from Stanford University, Stanford, CA, in 1996.

Since 1999, he has been with the Department of Chemistry and Chemical Biology at Harvard University, Cambridge, MA, as an Assistant Professor. His research interests lie in the physics and chemistry of nanostructured materials, specifically: 1) to study electrical properties of individual molecules, clusters, carbon nanotubes, and their arrays and 2) to develop synthetic schemes for transition–metal–oxide nanostructures and to study their electronic and magnetic properties.

be examined in a similar light before they can be unambiguously interpreted to constrain the absolute chronology of late Pleistocene sea level and ice volume change.

References and Notes

1. J. Imbrie *et al.*, in *Milankovitch and Climate, Part 1*, A. L. Berger, J. Imbrie, J. Hays, G. Kukla, B. Saltzman, Eds. (Reidel, Dordrecht, Netherlands, 1984), pp. 269–305.
2. B. M. Hickey, *Prog. Oceanogr.* **8**, 191 (1979).
3. W. J. Emery, K. Hamilton, *J. Geophys. Res.* **90**, 857 (1985).
4. J. J. Simpson, *Geophys. Res. Lett.* **10**, 917 (1983).
5. L. L. Ely, Y. Enzel, D. R. Cayan, *J. Clim.* **7**, 977 (1994).
6. F. G. Prahl, L. A. Muehlhausen, D. L. Zahnle, *Geochim. Cosmochim. Acta* **52**, 2303 (1988).
7. P. J. Muller, G. Kirst, G. Ruhland, I. von Storch, A. Rossell-Mele, *Geochim. Cosmochim. Acta* **62**, 1757 (1998).
8. T. D. Herbert *et al.*, *Paleoceanography* **13**, 263 (1998).
9. C. L. Reimers, R. A. Jahnke, D. C. McCorkle, *Global Biogeochem. Cycles* **6**, 199 (1992).
10. N. J. Shackleton, *Science* **289**, 1897 (2000).
11. N. J. Shackleton, A. Berger, W. R. Peltier, *Trans. R. Soc. Edinburgh Earth Sci.* **81**, 251 (1990).
12. D. G. Martinson *et al.*, *Quat. Res.* **27**, 1 (1987).
13. I. J. Winograd *et al.*, *Science* **258**, 255 (1992).
14. K. R. Ludwig *et al.*, *Science* **258**, 284 (1992).
15. Alkenone and δ^{18}O data as a function of core depth and estimated age are available at ftp://pixie.geo.brown.edu.
16. G. B. Griggs, L. D. Kulm, *J. Sediment. Petrol.* **39**, 1142 (1969).
17. D. V. Kent, N. D. Opdyke, M. Ewing, *Geol. Soc. Am. Bull.* **82**, 2741 (1971).
18. P. R. Thompson, N. J. Shackleton, *Nature* **287**, 829 (1980).
19. A. L. Sabin, N. G. Pisias, *Quat. Res.* **46**, 48 (1996).
20. J. Ortiz, A. Mix, S. Hostetler, M. Kashgarian, *Paleoceanography* **12**, 191 (1997).
21. F. G. Prahl, N. Pisias, M. A. Sparrow, A. Sabin, *Paleoceanography* **10**, 763 (1995).
22. H. Doose, F. G. Prahl, M. W. Lyle, *Paleoceanography* **12**, 615 (1997).
23. J. P. Kennett, K. Venz, *Proc. ODP Sci. Res.* **146**, 281 (1995).
24. P. G. Mortyn, R. C. Thunnell, D. M. Anderson, E. Tappa, *Paleoceanography* **11**, 415 (1996).
25. E. Bard, B. Hamelin, R. G. Fairbanks, *Nature* **346**, 456 (1990).
26. J. W. Kutzbach, H. E. Wright, *Quat. Sci. Rev.* **4**, 147 (1985).
27. S. Manabe, A. J. Broccoli, *J. Geophys. Res.* **90**, 2167 (1985).
28. W. E. Dean, J. V. Gardner, D. Z. Piper, *Geochim. Cosmochim. Acta* **61**, 4507 (1997).
29. T. D. Herbert, M. Yasuda, C. Burnett, *Proc. ODP Sci. Res.* **146**, 257 (1995).
30. C. Sancetta, M. Lyle, L. Heusser, R. Zahn, J. P. Bradbury, *Quat. Res.* **38**, 359 (1992).
31. L. V. Benson *et al.*, *Palaeogogr. Palaeoclimatol. Palaeoecol.* **78**, 241 (1990).
32. W. F. Ruddiman, A. McIntyre, *Geol. Soc. Am. Bull.* **95**, 381 (1984).
33. J. Villanueva, J. O. Grimalt, E. Cortijo, L. Vidal, L. Labeyrie, *Geochim. Cosmochim. Acta* **62**, 2421 (1998).
34. T. J. Crowley, *Paleoceanography* **9**, 1 (1994).
35. R. R. Schneider *et al.*, in *The South Atlantic: Present and Past Circulation*, G. Wefer *et al.*, Eds. (Springer, New York, 1995), pp. 527–551.
36. D. W. Lea, D. K. Pak, H. J. Spero, *Science* **289**, 1719 (2000).
37. L. E. Heusser, M. Lyle, A. Mix, *Proc. ODP Sci. Res.* **167**, 217 (2000).
38. N. G. Pisias, A. C. Mix, L. Heusser, *Quat. Sci. Rev.*, in press.
39. M. Lyle, L. Heusser, T. Herbert, A. Mix, J. Barron, in preparation.
40. Sixteen pollen units were counted by L. Heusser, with the abundance of ferns tabulated as well. *Pinus* (pine) dominates the entire record, accounting for an average of 61% of the assemblage on a fern-free basis. Other significant components include *Picea* (spruce), *Tsuga heterophylla* (western hemlock), *Abies* (fir), inaperaturate conifer types (juniper/cedar), *Sequoia* (redwood), *Quercus* (oak), *Alnus* (alder), *Compositae* (sunflower family), and family *Artemesia* (sage). We extracted statistical groupings of pollen that might reflect the changes in climate. A square root transform of pollen abundance reduced the dominance of *Pinus* and produced a more Gaussian distribution of factor scores in the resultant time series than a linear transform of pollen abundance (note that one would arrive at similar pollen factors and time series results with a linear transform as well). An analysis that excluded ferns produced three factors that account for 96% of the variance in the data. Factor 1 (42% of variance explained) contains high positive loadings of *Pinus*, *Picea*, *T. heterophylla*, and *Abies*. Negative loadings of *Pinus* and family *Artemesia* characterize factor 3 (28% of variance).
41. L. E. Heusser, N. J. Shackleton, *Science* **204**, 837 (1979).
42. L. E. Heusser, *Proc. ODP Sci. Res.* **146**, 265 (1995).
43. K. L. Rosanski, Araguas-Araguas, R. Gonfiantini, in *Climate Change in Continental Isotopic Records*, P. K. Swart, K. C. Lohmann, J. McKenzie, S. Savin, Eds. (American Geophysical Union, Washington, DC, 1993), pp. 1–36.
44. R. G. Fairbanks, R. K. Matthews, *Quat. Res.* **10**, 181 (1978).
45. D. P. Schrag, G. Hampt, D. W. Murray, *Science* **272**, 1930 (1996).
46. P. M. Grootes, in *Climate Change in Continental Isotopic Records*, P. K. Swart, K. C. Lohmann, J. McKenzie, S. Savin, Eds. (American Geophysical Union, Washington, DC, 1993), pp. 37–46.
47. I. J. Winograd, T. B. Copplen, K. R. Ludwig, J. M. Landwehr, A. C. Riggs, *Eos* **77** (suppl.), S169 (1996).
48. J. Imbrie, A. C. Mix, D. G. Martinson, *Science* **363**, 531 (1993).
49. G. M. Henderson, N. C. Slowey, *Nature* **404**, 61 (2000).
50. L. D. Stott, M. Neumann, D. Hammond, *Paleoceanography* **15**, 161 (2000).
51. J. D. Schuffert, M. Kastner, R. A. Jahnke, *Mar. Geol.* **146**, 21 (1998).
52. We thank the curators of the Ocean Drilling Program and Scripps Institution of Oceanography for making samples available. Portions of this work were supported by the NSF, JOI-USSAC, the Inter-American Institute for Global Change Research, and the U.S.-Mexico Foundation for Science.

22 January 2001; accepted 25 May 2001
Published online 7 June 2001;
10.1126/science.1059290
Include this information when citing this paper.

REPORTS

Carbon Nanotube Single-Electron Transistors at Room Temperature

Henk W. Ch. Postma, Tijs Teepen, Zhen Yao,* Milena Grifoni, Cees Dekker†

Room-temperature single-electron transistors are realized within individual metallic single-wall carbon nanotube molecules. The devices feature a short (down to ~20 nanometers) nanotube section that is created by inducing local barriers into the tube with an atomic force microscope. Coulomb charging is observed at room temperature, with an addition energy of 120 millielectron volts, which substantially exceeds the thermal energy. At low temperatures, we resolve the quantum energy levels corresponding to the small island. We observe unconventional power-law dependencies in the measured transport properties for which we suggest a resonant tunneling Luttinger-liquid mechanism.

Single-electron transistors (SETs) have been proposed as a future alternative to conventional Si electronic components (*1*). However, most SETs operate at cryogenic temperatures, which strongly limits their practical application. Some examples of SETs with room-temperature operation (RTSETs) have been realized with ultrasmall grains, but their properties are extremely hard to control (*2*–*4*). The use of conducting molecules with well-defined dimensions and properties would be a natural solution for RTSETs. We report RTSETs made within an individual metallic carbon nanotube molecule (*5*), characterizing their transport properties as a function of temperature, bias, and gate voltage and observing unexpected power-law characteristics that we describe with a Luttinger-liquid model.

SETs consist of a conducting island connected by tunnel barriers to two metallic leads (*1*). For temperatures and bias voltages that are low relative to a characteristic energy required to add an electron to the island, electrical transport through the device is blocked. Conduction

Department of Applied Physics and DIMES, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, Netherlands.

*Present address: Department of Physics, University of Texas at Austin, Austin, TX 78712, USA.
†To whom correspondence should be addressed. E-mail: dekker@mb.tn.tudelft.nl

can be restored, however, by tuning a voltage on a close-by gate, rendering this three-terminal device a transistor. Recently, we found that strong bends ("buckles") within metallic carbon nanotubes (*5*) act as nanometer-sized tunnel barriers for electron transport (*6*). This prompted us to fabricate single-electron transistors by inducing two buckles in series within an individual metallic single-wall carbon nanotube, achieved by manipulation with an atomic force microscope (AFM) (*7*) (Fig. 1). The two buckles define a 25-nm island within the nanotube. Nanotube devices have been fabricated with an island length of between 20 and 50 nm, and electrical transport has been measured through four of them. We report one representative data set obtained on the sample with a 25-nm island.

Typical RTSET transport characteristics for our nanotube devices obtained at room temperature are shown in Fig. 2A, which displays the differential conductance *dI*/*dV* versus bias voltage *V*. A voltage applied to the back gate appears to have a pronounced effect on the device conductance. A 0.2-V-wide gap is observed, which is closed upon changing the gate voltage. Upon varying the gate voltage further, the gap opens and closes in a periodic manner. At $V = 0$, this gives rise to a pattern of periodic conductance peaks (Fig. 2B). It thus appears that both bias and gate voltage can be used to modulate the conductance, and a conductance spectrum where both are varied simultaneously is shown in the inset to Fig. 2B. Diamond-shaped regions are visible where the conductance is suppressed. Traces such as those in Fig. 2A are cross sections of this conductance spectrum at fixed gate voltage, whereas those in Fig. 2B are cross sections at fixed bias voltage. Before fabrication of the two buckles, the device conductance did not change with gate voltage. It is thus evident that the modulation is due to the fabricated island.

These characteristics demonstrate Coulomb blockade (i.e., single-electron tunneling) at room temperature (*1*). Coulomb blockade as a function of bias and gate voltage occurs in diamond-shaped regions (compare with the inset to Fig. 2B). Within each diamond, the number of electrons is fixed, and electrons are added one by one to the island upon increasing the gate voltage. The height of the diamonds reads the bias voltage V^+ necessary to add an electron to the island, which defines an addition energy $E_{\mathrm{add}} = eV^+ = e^2/C + \Delta E$ (*8*), where C is the sum of all capacitances to the nanotube island and ΔE is the energy difference between consecutive quantum energy levels. We find $E_{\mathrm{add}} = 120$ meV, which is slightly larger than the largest value of 115 meV reported in previous planar RTSETs (*4*). The Coulomb blockade model describes all the basic device characteristics shown in Fig. 2, A and B. E_{add} is much larger than the thermal energy k_BT (k_B is the Boltzmann constant and T is the absolute temperature) at room temperature, which explains the room-temperature operation of our devices.

With respect to the device characteristics at low temperature, examination of the data at 30 K (Fig. 2, C to E) reveals features that were not observed at room temperature. First, the data show less scatter owing to a reduction in sample noise at low temperatures. Second, the *dI*/*dV* (V) traces show peaked features (indicated by the lines in Fig. 2C) that shift along the bias voltage axis when changing the gate voltage. We believe that these peaks are associated with energy levels of the island that become available for electronic transport, leading to an increase in current. The peaks can be followed as a function of both bias and gate voltage in the conductance spectrum (dashed lines in Fig. 2, D and E). The distance between these lines along the bias-voltage axis is equal to $2\Delta E$. We find that $\Delta E = 38$ meV. From the linear dispersion relation of nanotubes one estimates an average $\Delta E = hv_F/4L$ for a tube of finite length L, when the degeneracy between the two sets of energy levels in nanotubes has been lifted. Here v_F is the Fermi velocity and h is Planck's constant. With $v_F = 8 \times 10^5$ m/s (*9*), we obtain 34 meV for the 25-nm island, which is in good agreement with the measured value, confirming that

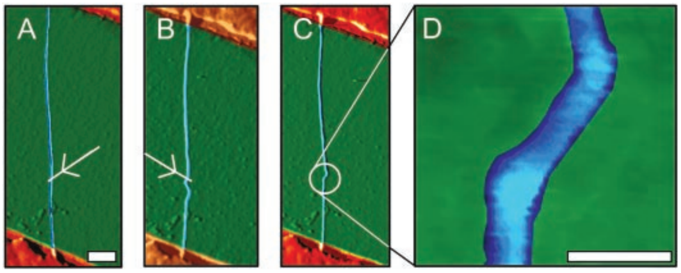


Fig. 1. Fabrication of a room-temperature single-electron transistor within an individual metallic carbon nanotube by manipulation with an AFM (*7*). (**A**) Nanotube between Au electrodes on top of a Si/SiO$_2$ substrate with a gate-independent resistance of 50 kilohm. After imaging by scanning the AFM tip over the sample in tapping mode, the tip is pressed down onto the substrate and moved along the path indicated by the arrow, thus dragging the nanotube into a new configuration. Bar, 200 nm. (**B**) Nanotube after creation of a buckle. The dragging action has resulted in a tube that is bent so strongly that it has buckled (*31*). A second dragging action is performed as indicated by the arrow. (**C**) Double-buckle nanotube device. (**D**) Enlarged image of the double-buckle device. The image shows a height increase at the buckling points, as expected (*31*). The final device resistance at room temperature is one order of magnitude larger (~0.5 megohm). The electronic transport properties of these nanotube devices are studied by application of a bias voltage *V* to the upper electrode and a measurement of the current *I* at the lower electrode. The differential conductance *dI*/*dV* is measured with a standard ac–lock-in technique with a modulation amplitude of 0.1 mV. The conducting Si substrate underneath the insulating SiO$_2$ substrate is coupled capacitively to the nanotube and acts as a back gate. Bar, 20 nm.

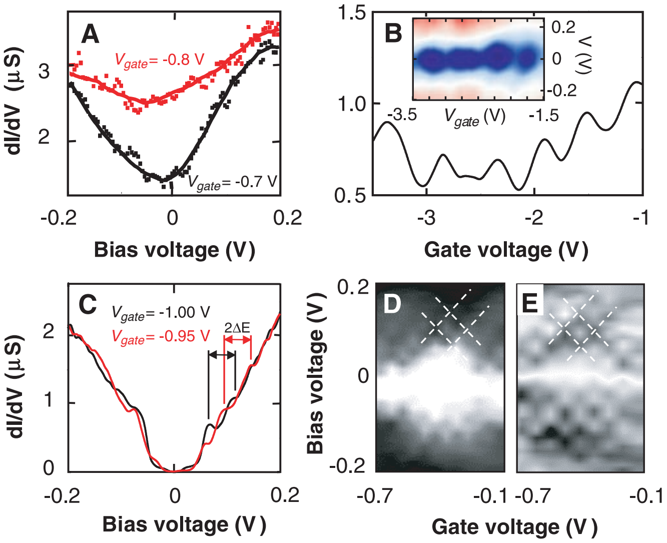


Fig. 2. Differential conductance *dI*/*dV* of the RTSET as a function of bias and gate voltage at various temperatures. (**A**) At 300 K, the differential conductance shows a thermally smeared gap around $V = 0$ with gate voltage $V_{\mathrm{gate}} = -0.7$ V (lower trace). When the gate voltage is changed to -0.8 V, the gap is closed. (**B**) Conductance oscillations as a function of gate voltage at 260 K. (Inset) *dI*/*dV* in an intensity plot. Blue represents low *dI*/*dV*, red corresponds to high *dI*/*dV*. The gap is periodically opened and closed as a function of the gate voltage, which results in diamond-shaped modulations. (**C**) *dI*/*dV* at 30 K, showing distinct peaks as indicated by the lines. The peaks in the black trace at $V_{\mathrm{gate}} = -1$ V shift up in bias voltage when V_{gate} is increased to -0.95 V (red trace). (**D**) Gray-scale image of *dI*/*dV*, where shifting peaks are indicated by the dashed lines. White represents *dI*/*dV* = 0, whereas darker shading correspond to higher values of *dI*/*dV*. (**E**) Gray-scale image of d^2I/dV^2, showing the presence of conductance peaks.

the island behaves as a well-defined quantum box for the electrons. From the addition energy we can now extract the charging energy $E_C \equiv e^2/2C$, which reads 41 meV. The fact that $\Delta E \sim E_C$ is unique to our devices, whereas in the ordinary case $E_C \gg \Delta E$. This is a direct result of the small size of these islands and the nature of the buckle junctions. In contrast to previous studies on straight undeformed nanotubes (*9–12*), the capacitances of the nanotube island to the nanotube leads (≈0.3 aF) are now a major contribution to the total capacitance *C*. Hence, whereas ΔE will increase with decreasing *L*, E_C will remain approximately constant, yielding a larger $\Delta E/E_C$ ratio.

Figure 3 shows the temperature dependence of the device conductance. A plot of a single conductance peak versus gate voltage (Fig. 3A) shows that both the conductance peak maximum G_{max} and the peak width *w* increase with increasing temperature. This is in marked contrast to the behavior expected for a conventional SET in both the classical ($k_BT > \Delta E$) and the quantum regime of Coulomb blockade ($k_BT < \Delta E$). In both of the latter cases $w \propto T$, but G_{max} = constant or $G_{max} \propto 1/T$, respectively (*8*). Our data also differ from previous results for carbon nanotube SET devices operating at low temperatures (*9–13*).

The conductance shows a power-law dependence on *T* (Fig. 4), where $G_{max}(T)$ is plotted for the peak in Fig. 3A. From 4 to 90 K, it follows a power law $G_{max} \propto T^{0.68}$, whereas at higher temperatures it increases beyond this. The inset of Fig. 4 shows that *w* follows a linear temperature dependence $w \propto T$ at low temperatures, whereas it deviates from the expected behavior at higher temperatures. We explain the high-temperature deviations from the overlap between adjacent peaks as follows. As the peak width increases with temperature, the peak tails start to overlap progressively with adjacent conductance peaks upon raising the temperature, leading to both a larger apparent peak height as well as a larger apparent width. We correct for this by integrating the conductance over gate voltage, yielding an integrated conductance G^*. We find a strong temperature dependence $G^* \propto T^{1.66}$, in excellent agreement with the expected behavior $G^* \propto T^{1+0.68}$ from $w(T)$ and $G_{max}(T)$. This power-law behavior persists well above the temperatures where both *w* and G_{max} deviate from the expected behavior, indicating that the deviations in measurements of *w* and G_{max} are indeed due to overlap of adjacent conductance peaks.

The power-law exponents observed in our experiments cannot be explained by the available models. Recent transport experiments on metallic carbon nanotubes (*6*, *14–16*) successfully used a Luttinger-liquid model (*17*, *18*), which derives from the one-dimensional electronic correlations in nanotubes. We therefore compare our experimental results with those from theoretical studies of a Luttinger island connected by tunnel barriers to two semi-infinite Luttinger liquids. Such studies have described transport in terms of sequential tunneling processes, with independent tunneling from the leads onto the island and from the island into the other lead (*19–22*). This leads to $G_{max} \propto T^{\alpha_{end}-1}$ and $w \propto T$, where $\alpha_{end} = \frac{1}{4}(\frac{1}{g} - 1)$, with the Luttinger interaction parameter *g* characterizing the electron-electron interaction strength (*17*, *18*). Experiments (*6*, *14–16*) show that *g* ranges between 0.19 and 0.26 in the case of carbon nanotubes, which leads to $G_{max} \propto T^{-0.2}$ and $G^* \propto T^{0.8}$, in contradiction with the present data.

We therefore propose another mechanism, namely, correlated sequential tunneling through the island. Here electrons tunnel coherently from the end of one nanotube lead to the end of the other nanotube lead through a quantum state in the island. In this picture, the island should be regarded as a single impurity (*23*). This is a reasonable assumption, because the thermal length $L_T \equiv v_F/k_BT$ = 70 nm at 300 K, and hence it is larger than the distance between the two barriers at all temperatures. The calculation for the conductance due to this tunneling mechanism yields $G_{max} \propto T^{\alpha_{end\text{-}end}-1}$ and $G^* \propto T^{\alpha_{end\text{-}end}}$, where $\alpha_{end\text{-}end} = 2\alpha_{end}$ (*24*). Upon identifying $\alpha_{end\text{-}end}$ with the experimental value 1.66, we obtain $g = 0.23$, in excellent agreement with earlier values for nanotubes. This shows that the proposed mechanism is the relevant transport channel for the RTSET. The surprising dominance of correlated tunneling over conventional sequential tunneling contrasts with the conventional understanding of SETs and quantum dots (*1*). Our model is further confirmed by data of the integrated differential conductance $(dI/dV)^*$ versus bias voltage at large bias ($V >$ 10 mV), which yields a power law $(dI/dV)^* \propto V^{0.87}$ (*25*). One theoretically expects that the large bias voltage $eV \gg k_BT$ will destroy the phase coherence necessary for correlated tunneling, and that conventional sequential tunneling will dominate (*24*). The expected exponent for this process is $\alpha_{end} = \alpha_{end\text{-}end}/2 = 0.83$, which is again close to the value found experimentally.

For practical applications, a figure of merit of SETs is the input equivalent charge noise q_n. Preliminary measurements on the present nanotube devices at 10 Hz and 60 K yielded $q_n \approx 2 \times 10^{-3}$ e/$\sqrt{\text{Hz}}$, which compares favorably to $q_n \approx 0.5 \times 10^{-3}$ e/$\sqrt{\text{Hz}}$ for conventional single-electron transistors that operate at mK temperatures (*26*). The prototype nanotube RTSETs presented here were obtained by manipulation with an AFM. Future use in large-scale applications will require further developments in fabrication technology such as mechanical templates or chemical methods to create short nanotubes in a parallel process. RTSETs have several advantages over room-temperature field-effect transistors using semiconducting nanotubes (*27*). Because semiconducting nanotubes are, unlike metallic tubes, intrinsically prone to disorder and unintentional doping (*28*, *29*), molec-

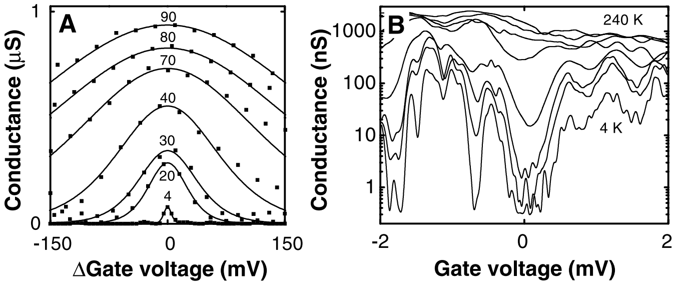


Fig. 3. (**A**) Zero-bias conductance versus gate voltage for a single conductance peak. Data are shown for *T* = 4, 20, 30, 40, 70, 80, and 90 K. The solid lines are fits to the function $G_{max}(T)/\cosh^2[2(\text{arccosh}\sqrt{2})\, V_{gate}/w(T)]$ [as derived from (*8*)]. The *T* dependence of the conductance maximum G_{max} and the full width at half maximum *w* is shown in Fig. 4. (**B**) Multiple conductance peaks on a logarithmic scale for an extended range of gate voltages at *T* = 4, 40, 60, 70, 100, 170, 200, 220, and 240 K.

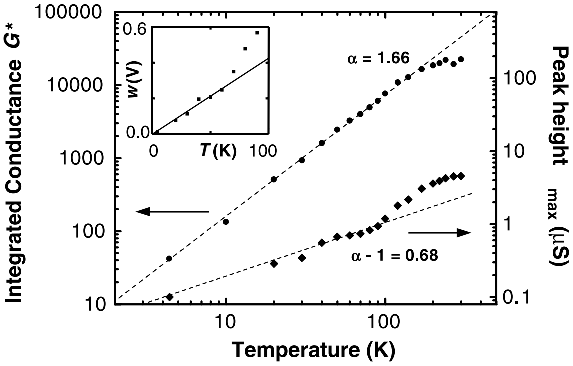


Fig. 4. Power-law temperature dependence of the conductance, demonstrating correlated sequential tunneling through the nanotube SET device. Lower data (right-hand scale) show the peak height $G_{max}(T)$ for the conductance peak in Fig. 3A, following a power-law function with exponent 0.68 (◆). The conductance integrated over the gate voltage range in Fig. 3B, $G^*(T)$ (left-hand scale), also follows a power-law function with exponent 1.66 (●). Note the double-logarithmic scales. The inset shows the peak width *w* versus *T*, which displays a linear behavior.

ular-electronics components based on metallic tubes are preferred. The present work shows that short, metallic nanotubes can be applied as RTSETs. It also serves to illustrate that the search for functional molecular devices often yields interesting fundamental science.

References and Notes

1. H. Grabert, M. Devoret, Eds., in *Single Charge Tunneling* (Plenum, New York, 1992).
2. Y. Takahashi *et al.*, *Electron. Lett.* **31**, 136 (1995).
3. K. Matsumoto *et al.*, *Appl. Phys. Lett.* **68**, 34 (1996).
4. L. Zhuang, L. Guo, S. Y. Chou, *Appl. Phys. Lett* **72**, 1205 (1998).
5. C. Dekker, *Phys. Today* **52** (5), 22 (1999).
6. H. W. Ch. Postma, M. de Jonge, Z. Yao, C. Dekker, *Phys. Rev. B* **62**, R10653 (2000).
7. H. W. Ch. Postma, A. Sellmeijer, C. Dekker, *Adv. Mater.* **17**, 1299 (2000).
8. C. W. J. Beenakker, *Phys. Rev. B* **44**, 1646 (1991).
9. S. J. Tans *et al.*, *Nature* **386**, 474 (1997).
10. M. Bockrath *et al.*, *Science* **275**, 1922 (1997).
11. D. H. Cobden, M. Bockrath, P. L. McEuen, A. G. Rinzler, R. E. Smalley, *Phys. Rev. Lett.* **81**, 681 (1998).
12. H. W. Ch. Postma, Z. Yao, C. Dekker, *J. Low Temp. Phys.* **118**, 495 (2000).
13. M. Bockrath *et al.*, *Science* **291**, 283 (2001).
14. M. Bockrath *et al.*, *Nature* **397**, 598 (1999).
15. Z. Yao, H. W. Ch. Postma, L. Balents, C. Dekker, *Nature* **402**, 273 (1999).
16. J. Nygard, D. H. Cobden, M. Bockrath, P. L. McEuen, P. E. Lindelof, *Appl. Phys. A* **69**, 297 (1999).
17. C. L. Kane, L. Balents, M. P. A. Fisher, *Phys. Rev. Lett.* **79**, 5086 (1997).
18. R. Egger, A. O. Gogolin, *Phys. Rev. Lett.* **79**, 5082 (1997).
19. C. L. Kane, M. P. A. Fisher, *Phys. Rev. Lett.* **68**, 1220 (1992).
20. A. Furusaki, *Phys. Rev. B.* **57**, 7141 (1998).
21. A. Braggio, M. Grifoni, M. Sassetti, F. Napoli, *Europhys. Lett.* **50**, 236 (2000).
22. Although the first theoretical predictions for Luttinger liquid behavior considered double-barrier devices, experimental evidence so far is scarce [see (*30*)].
23. M. P. A. Fisher, L. Balents, personal communication.
24. M. Grifoni *et al.*, unpublished data.
25. H. W. Ch. Postma, data not shown.
26. A. B. Zorin *et al.*, *Phys. Rev. B* **53**, 13682 (1996).
27. S. J. Tans, A. R. M. Verschueren, C. Dekker, *Nature* **393**, 49 (1998).
28. S. J. Tans, C. Dekker, *Nature* **404**, 834 (2000).
29. A. Bachtold *et al.*, *Phys. Rev. Lett.* **84**, 6082 (2000).
30. O. Auslaender *et al.*, *Phys. Rev. Lett.* **84**, 1764 (2000).
31. S. Iijima, C. Brabec, A. Maiti, J. Bernholc, *J. Chem. Phys.* **104**, 2089 (1996).

32. We thank R. E. Smalley and co-workers for providing the carbon nanotube material; A. Bachtold, M. P. A. Fisher, L. Balents, P. Hadley, M. Thorwart, A. Braggio, and Yu. V. Nazarov for discussions; and B. van den Enden for technical assistance. Supported by the Dutch Foundation for Fundamental Research on Matter (FOM) and the European Community SATURN project.

20 April 2001; accepted 23 May 2001

Fully Conjugated Porphyrin Tapes with Electronic Absorption Bands That Reach into Infrared

Akihiko Tsuda and Atsuhiro Osuka*

Scandium(III)-catalyzed oxidation of *meso-meso*–linked zinc(II)-porphyrin arrays (up to dodecamers) with 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (DDQ) led to efficient formation of triply *meso-meso*–, β-β–, and β-β–linked zinc(II)-oligoporphyrins with 62 to 91% yields. These fused tape-shaped porphyrin arrays display extremely red-shifted absorption bands that reflect extensively π-conjugated electronic systems and a low excitation gap. The lowest electronic absorption bands become increasingly intensified and red-shifted upon the increase in the number of porphyrins and eventually reach a peak electronic excitation for the dodecamer at ~3500 wavenumber. The one-electron oxidation potentials also decreased progressively upon the increase in the number of porphyrins. These properties in long and rigid molecular shapes suggest their potential use as molecular wires.

Discrete molecules with a very long π-system are of interest as organic conducting materials, near-infrared (near-IR) dyes, nonlinear optical materials, and molecular wires (*1*–*3*). Numerous attempts that have been made to extend the π-electronic systems have, however, encountered serious problems, such as synthetic inaccessibility, chemical instability, poor solubility, and conjugation saturation behavior that arises through the effective conjugated length (ECL) effect. The ECL defines the extent of π-conjugated systems in which the electronic delocalization is limited and at which point the optical, electrochemical, and other physical properties reach a saturation level that is common with the analogous polymer (*1*). A straightforward strategy for maximizing π-overlap may be to hold the π-systems coplanar within a tapelike framework by fusing the units edge-to-edge, to make a covalently linked long, flat array, but this goal is synthetically quite demanding. Fused π-conjugated systems are promising also in circumventing the above ECL limit, as seen for the [*n*]acene series (n = 1 to 7) (*1*, *4*), but extension to the higher conjugated systems suffers from problems of poor solubility caused by the resulting planar structures. Within a confined pigment number, charged dyes such as oxonols and cyanines can escape the ECL effect because of the absence of the bond alternation arising from effective resonance (*5*). Again, extension to the higher homologs is difficult to achieve and reveals the ECL effect (*6*).

Porphyrins are intriguing building units from which to construct large π-conjugated molecules. Two types of conjugated porphyrins have been developed, *meso*-ethyne–bridged and *meso*-butadiyne–bridged porphyrin arrays (*7*–*9*) and fused porphyrin arrays (*10*–*13*), both of which show unusual properties that result from strong π-conjugation. Here we report the synthesis of highly conjugated porphyrin arrays, in which the electronic π-conjugation over the arrays is far stronger than the π-conjugation of these precedents, as seen from extremely low-energy IR electronic excitations.

Recently, we reported the synthesis of *meso-meso*–linked porphyrin arrays of up to 128-oligomers by Ag^{I} salt–promoted coupling reaction (*14*). This extremely long, discrete, rodlike organic molecule has a molecular length of about 108 nm. These arrays adopt a nearly orthogonal conformation that tends to minimize the electronic interaction between the neighboring porphyrins (*14*, *15*). The aryl-end–capped *meso-meso*–linked Cu^{II}-diporphyrin **1** can be converted to triply linked fused diporphyrin **2** by the oxidative double-ring closure (ODRC) reaction upon treatment with (p-$BrC_6H_4)_3NSbCl_6$ in C_6F_6 (Scheme 1) (*13*). The planar structure of **2** has been revealed by x-ray analysis, and full conjugation over the two porphyrins has been demonstrated by its substantially broadened and red-shifted absorption spectrum.

We now describe a highly efficient synthetic method that allows the ODRC reaction of higher *meso-meso*–linked Zn(II)-porphyrins in good yields. The ODRC reaction was conducted simply by refluxing a toluene solution of *meso-meso*–linked Zn(II)-diporphyrin **3** in the presence of five equivalents of 2,3-dichloro-5,6-dicyano-1,4-benzoquinone (DDQ) and scandium trifluoromethanesulfonate [$Sc(OTf)_3$] for 0.5 hour, which afforded the triply linked fused diporphyrin **4** in 86% yield as a sole product. DDQ or $Sc(OTf)_3$ alone did not effect any change of **3**. Under similar conditions, the Zn(II)-porphyrin monomer **5** was also effectively coupled to give triply linked diporphyrin **4** in

Department of Chemistry, Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan.

*To whom correspondence should be addressed.

nature nanotechnology

REVIEW ARTICLE

PUBLISHED ONLINE: 29 NOVEMBER 2009 | DOI: 10.1038/NNANO.2009.355

Nanotube electronics for radiofrequency applications

Chris Rutherglen, Dheeraj Jain and Peter Burke*

Electronic devices based on carbon nanotubes are among the candidates to eventually replace silicon-based devices for logic applications. Before then, however, nanotube-based radiofrequency transistors could become competitive for high-performance analogue components such as low-noise amplifiers and power amplifiers in wireless systems. Single-walled nanotubes are well suited for use in radiofrequency transistors because they demonstrate near-ballistic electron transport and are expected to have high cut-off frequencies. To achieve the best possible performance it is necessary to use dense arrays of semiconducting nanotubes with good alignment between the nanotubes, but techniques that can economically manufacture such arrays are needed to realize this potential. Here we review progress towards nanotube electronics for radiofrequency applications in terms of device physics, circuit design and the manufacturing challenges.

The potential to exploit single-walled carbon nanotubes in advanced electronics has been a major goal in nanotechnology for over a decade[1,2]. This interest stems from the fact that carbon nanotubes offer a combination of small size, high mobility[3,4], large current density and low intrinsic capacitance: moreover, their intrinsic cut-off frequency is expected to be high. Although the long-term goal of nanotube researchers has been to replace digital CMOS devices made from silicon, and therefore to "extend Moore's law", a more realistic point of insertion into the market may be high-performance analogue radiofrequency (RF) devices, where manufacturing tolerances are relaxed and the performance metrics required for commercial systems are more suited to the materials and device properties of nanotubes. To realize this potential, it must be possible to economically manufacture dense aligned arrays of all-semiconducting nanotubes.

The use of massively parallel nanotube-based field-effect transistors (FETs) for applications such as mobile communication devices and radar is at present being investigated in both academic and industrial laboratories. So far, numerous nanotube-based FETs have been demonstrated using both single nanotubes and thin-film transistors made from mixtures of semiconducting and metallic nanotubes[5]. (The nanotubes in these devices can either be aligned or randomly oriented.) However, to achieve the highest performance, the nanotubes must be aligned at a high density (Fig. 1). Otherwise, the mobility is degraded from that of a pristine nanotube, and the fringe-field capacitance degrades the cut-off frequency by up to two orders of magnitude[6]. For this reason, the manufacturability of aligned arrays is very important, and several techniques have been investigated to solve the problems of nanotube alignment and purification: the two main techniques are 'grow in place' and 'deposition from solution'.

It has been proposed that nanotube-based FETs could, in principle, operate at frequencies well into the terahertz regime[6–11]. However, as it might not be possible to economically manufacture the perfectly dense perfectly aligned arrays containing only semiconducting nanotubes that are needed to achieve this level of performance, it is important to benchmark trade-offs that result from using less-than-perfect arrays. An intriguing aspect of nanotube-based FETs is a predicted inherent linearity[12], which is critically important for wireless communication systems. To confirm and quantify these and other device properties under realistic operating conditions, it is important to fabricate, test and demonstrate devices with high-density, aligned, all-semiconducting nanotubes in a scalable process, and to demonstrate such devices in actual working radio systems applications.

Here we review the progress so far in manufacturing, discuss the predicted and measured device properties as a function of manufacturing tolerances, and consider the implications for applications of single-walled nanotubes in analogue (as opposed to digital) RF devices and, ultimately, RF systems applications.

Grow in place by chemical vapour deposition

The most widely used method for growing single-walled nanotubes directly onto a substrate has been chemical vapour deposition (CVD). In general, a substrate holding metal catalyst particles is placed within a furnace with a flow of carbon feedstock gas and hydrogen gas at temperatures upwards of 900 °C. In such an environment, carbon nanotubes will grow from the catalyst particles with a diameter that is related to the size of these particles. To obtain aligned nanotubes during the CVD growth, multiple methods have been used to guide the alignment, such as applied electric fields[13,14], the gas flow[15–18] and interactions with the substrate. Of these, the most successful for obtaining highly dense perfectly aligned arrays of nanotubes has been surface-guided growth on single-crystal substrates such as sapphire or quartz[19–25]. Although the basic alignment mechanism remains unclear, it is assumed to involve the interactions between the nanotubes and the substrate's atomic steps, nanofacets or crystallographic lattice — or a mixture of these. Nanotube lengths of greater than 100 μm, linear densities of 10 nanotubes μm^{-1} (with peak values ~50 nanotubes μm^{-1}) and alignment within <0.01° have been achieved (Fig. 1c). Furthermore, procedures for transferring the aligned arrays to other substrates, such as $SiO_{2,}$ or flexible substrates have been developed[26]. These techniques allow heterogeneous integration of aligned single-walled nanotubes with other materials that would not otherwise survive the high temperatures involved with the CVD nanotube growth process.

Nanotubes produced by the methane CVD method typically yield a mixture of two-thirds semiconducting nanotubes and one-third metallic nanotubes. The presence of the metallic nanotubes in parallel with the semiconducting nanotubes degrades device performance, especially the on/off ratio and the output resistance. Individual nanotube-based FETs have demonstrated on/off

Integrated Nanosystems Research Facility, Departmental of Electrical Engineering and Computer Science, University of California, Irvine, California 92697, USA. *e-mail: pburke@uci.edu

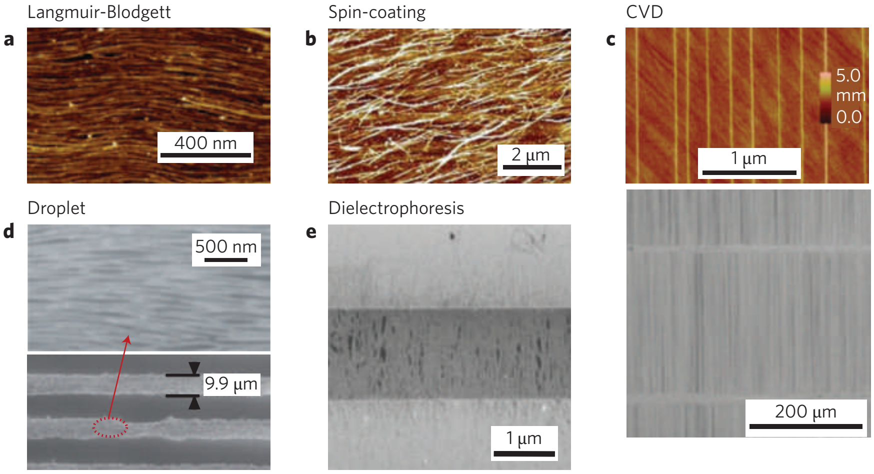


Figure 1 | Different ways to align nanotubes. To make high-frequency field-effect transistors from single-walled nanotubes (SWNTs), the nanotubes must be aligned, and they must also be long enough to span the source–drain channel. **a**, The Langmuir-Blodgett method can align SWNTs with a density of 30 nanotubes µm^{-1}, as shown in this atomic force microscopy image. Reproduced with permission from ref. 42 (© 2007 ACS). **b**, The spin-coating method is capable of aligning 10 nanotubes µm^{-2}, and an alignment of less than 10° of the radial axis, as shown in this atomic force microscopy image. Reproduced with permission from ref. 51 (© 2008 AAAS). **c**, Growing SWNTs by CVD on a single-crystal quartz substrate yields a high degree of alignment (<0.01°), as seen in the atomic force microscopy image (top). This method also produces nanotubes with lengths greater than 100 µm between the pair of catalyst lines, as shown in the scanning electron microscopy image (bottom). Reproduced with permission from ref. 29 (© 2009 ACS). **d**, The evaporating-droplet method produces densities of 10–20 nanotubes µm^{-1}, and alignment of less than 5°, as shown in these scanning electron microscopy images. Reproduced with permission from ref. 52 (© 2008 ACS). **e**, Dielectrophoresis uses the electric field to attract and align SWNTs between a pair of electrodes, as seen in this scanning electron microscopy image. Reproduced with permission from ref. 43 (© 2008 AIP).

ratios >10^6, but this ratio is much lower for combinations of metallic and semiconducting nanotubes. Although a degradation in the on/off ratio is acceptable for analogue RF applications (which relaxes the manufacturing requirements for analogue devices compared with those for digital devices), the presence of the metallic nanotubes also reduces the output resistance, which lowers the gain and frequency of operation, as discussed below. Therefore, it is necessary to devise strategies to selectively remove the metallic nanotubes while preserving (as much as possible) the semiconducting nanotubes.

Various gas-phase or plasma-etching methods have been developed to selectively remove metallic nanotubes[27,28]. Some of these methods can be incorporated into the growth process itself[29,30], and a combination of ethanol/methanol carbon feedstock mixture and copper nanoparticles as the catalyst was recently used to selectively grow >95% semiconducting nanotubes with a narrow diameter distribution and on/off ratios up to 85 (ref. 29). This selective growth is thought to be due to the OH$^-$ radical from methanol selectively etching the metallic nanotubes during the growth owing to their smaller ionization potential compared with the semiconducting variety.

Using such preferential growth, one can further enhance the on/off current ratio by post-growth removal of the metallic nanotubes. One such method[27] involves the selective etching by hydrocarbonation of metallic nanotubes with diameters between ~1.3 and 1.6 nm using a 400 °C methane plasma treatment to achieve on/off ratios of 10^4–10^5. It is found that nanotubes having a diameter smaller than this range are indiscriminately etched regardless of being metallic or semiconducting, whereas those with larger diameters are not affected at all. This general processing method has the advantage that it is scalable and compatible with other traditional semiconductor processing techniques, although some semiconducting nanotubes are also damaged in the process.

'Wet etching' of metallic nanotubes has also been demonstrated. The process originates from the selective reaction of diazonium salts with the sidewalls of the nanotubes to significantly perturb their electronic and optical properties[31–33]. On/off current ratios are found to improve to 10^4.

The electrical breakdown method is a post-growth treatment that involves selectively 'burning off' metallic nanotubes by applying a strong gate-bias to deplete or turn off the semiconducting nanotube, thus forcing the current though the metallic nanotubes[34]. By ramping up the drain–source voltage, typically to greater than 30 V, it is possible to burn off the metallic nanotubes in the presence of oxygen. This process has been shown to improve the on/off current ratio upwards of 10^5, but this improvement comes at the cost of decreasing the pre-breakdown mobility owing to the inadvertent damaging of the semiconducting nanotubes as a result of

Table 1 | Ideal parameter values for making a high-frequency field-effect transistor from single-walled nanotubes.

Property/parameter	Target value or range	Justification
Diameter	1.5–2.0 nm	Current is largest in this range[54-55].
Chirality	Semiconducting and same (*n*,*m*)	To obtain identical transport properties.
Purity	>99% semiconducting nanotubes	No metallic nanotubes for high gain and high f_{max}.
Length	>1 µm	Nanotube length must be longer than the intended channel length.
Density	>10 nanotubes µm^{-1}	Reduces the parasitic capacitance per nanotube; increases current carrying capacity; improves impedance matching.
Alignment	All parallel	Results in higher transconductance and denser nanotube packing.
Uniformity	Wafer scale	Essential for large-scale processing.

the Joule heating produced by adjacent metallic nanotubes. Such reduction in mobility has been found to result in a post-breakdown mobility of up to half its pre-breakdown value for the standard two-thirds semiconductor/one-third metallic mix with densities of ~10 nanotubes μm^{-1} (refs 35–38). As the density is further increased and the distance between nanotubes becomes smaller, one would anticipate this collateral damage to adjacent nanotubes to be even more severe. From the scalability perspective, one would face the additional challenge of applying the necessary high voltage to each device on the wafer: an alternative approach that relies on microwaves[39] or light[40] to selectively burn off the metallic nanotubes has had some limited success.

Deposition from solution

Radiofrequency FETs can also be made using the 'deposition from solution' technique. A variety of techniques have been developed to sort as-produced single-walled nanotubes: these include selective chemistry, chromatographic separation and electrophoretic separation (see ref. 41 for a review). Using these techniques, or a combination of them, in the near future it should be feasible to prepare a solution of nanotubes in which all the nanotubes have the same length and the same chirality. (The chirality of a nanotube is denoted by two integers (*n*,*m*) which define the direction in which a hypothetical sheet of graphene would be rolled up to form that nanotube, and which also determine the diameter of the nanotube and whether it is metallic or semiconducting.)

When sorting nanotubes for applications in electronics the key challenges are: the economy of the process; the ability to sort large diameter (>1.5 nm) nanotubes; and the ability to sort sufficiently long nanotubes (ideally >1 μm) so that their length is longer than the source–drain spacing. Once these challenges (which do not seem to be insurmountable) have been solved, the remaining challenges will include learning how to deposit and assemble the nanotubes into an aligned array, and understanding how residual surfactants influence the electronic properties of the array once it has been assembled. (Nanotubes tend to be insoluble, so it is usually necessary to functionalize them first to make them soluble before they can be used in 'deposition from solution' methods.) Progress in these areas is reviewed below.

In the Langmuir-Blodgett technique a solution of nanotubes is spread on top of water in a Langmuir-Blodgett trough (in much the same way that oil spreads to form a slick on water), and movable barriers in the trough are used to subject the sample to cycles of compression and retraction, which results in the formation of a self-assembled monolayer of nanotubes. The nanotubes are then transferred onto a solid substrate by successively dipping the substrate through the monolayer. This method[42] has yielded linearly aligned tubes with packing densities of more than 30 nanotubes μm^{-1} (Fig. 1a), and the process is conceivably scalable to wafer-scale processing.

Nanotubes can be aligned using a.c. electric fields and then deposited between two closely spaced electrodes using dielectrophoresis[43–46] (Fig. 1e). A disadvantage of this process is its tendency to preferentially accumulate metallic nanotubes owing to their stronger polarizability compared with semiconducting nanotubes[47–50]. The other challenges include scaling up the process for wafer-scale production and combating the tendency of the nanotubes to form bundles during deposition.

Spin coating is a simple technique that involves spinning a wafer (usually made of silicon) at high speeds, and dripping a solution of nanotubes onto it so that they are deposited in a radially aligned pattern[51]. Although on/off ratios of $>10^5$ have been achieved, the devices have a low on-state current owing to the very large sheet resistance of the nanotube film. So far the densities obtained have been ~10 nanotubes μm^{-2} with moderate alignment (within ~10° of the radial axis[51]; Fig. 1b). (For randomly aligned nanotubes, researchers tend to quote areal rather than linear densities.)

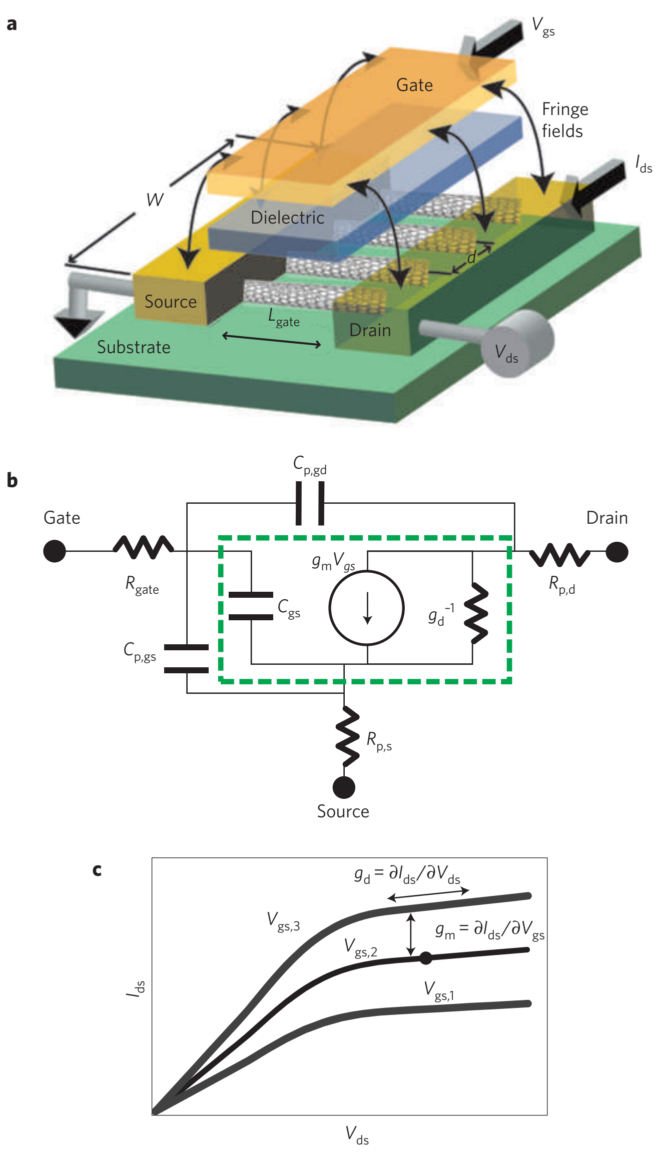


Figure 2 | The nanotube field-effect transistor. a, Schematic showing a FET in which the channel is an array of single-walled nanotubes: *W* is the gate width, L_{gate} is the gate length, *d* is the pitch (or spacing) of the nanotubes, V_{gs} and V_{ds} are the gate–source and drain–source voltages, and I_{ds} is the drain–source current. For RF-FETs, aligned arrays of nanotubes are needed to improve the impedance matching and increase the transconductance, the on-state current and the power density of the device. The fringe electrical fields from the gate to the source and drain give rise to the parasitic capacitance. **b**, A small-signal equivalent circuit for a nanotube-based FET where g_m is the transconductance, C_{gs} the intrinsic gate capacitance, and g_d the channel conductance (which can be significant if metallic nanotubes are present). These components encompass the 'intrinsic' portion of the device. The components outside the dashed line are parasitic elements: $C_{p,gs}$ and $C_{p,gd}$ are the gate–source and gate–drain parasitic capacitances, $R_{p,s}$ and $R_{p,d}$ are parasitic resistances for the source and drain, and R_{gate} is the resistance of the gate electrode. **c**, Schematic showing how the current through a nanotube transistor I_{ds} varies with the voltage across the transistor V_{ds} at three different values of the gate voltage V_{gs}. In practical applications the transistor is operated in the saturation regime at the values of V_{ds} and V_{gs} that give the optimum performance for a particular application (such as the highest gain or lowest noise). For d.c. voltages, the transconductance g_m depends on how I_{ds} changes with respect to the changes in V_{gs}, whereas the channel conductance g_d depends on how I_{ds} changes with respect to the changes in V_{ds}.

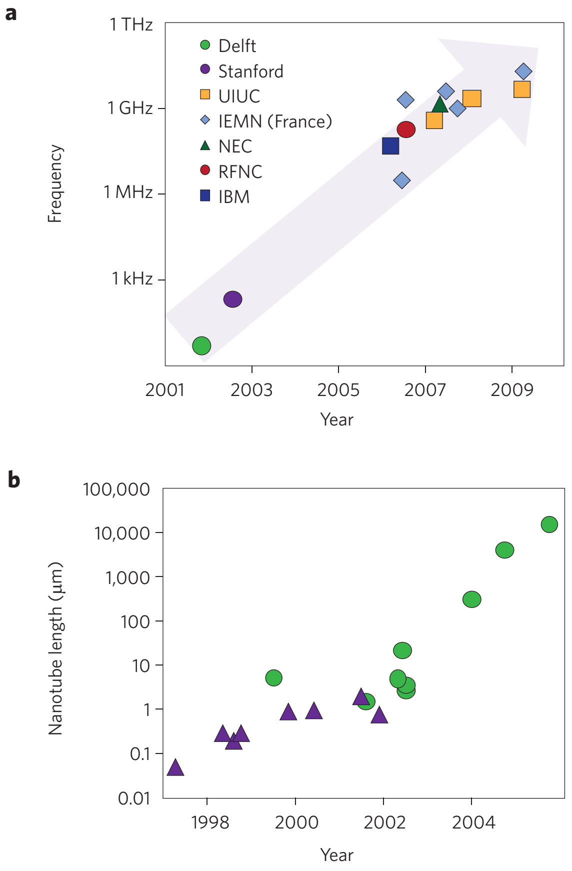


Figure 3 | Improvements over time. a, Maximum operating frequency (on a log scale) versus year for nanotube FETs. The maximum ring-oscillation frequency is plotted for the early work at Delft[69], Stanford[70] and IBM[71], and the cut-off frequency is plotted for the later work at RF Nano Corporation (RFNC)[60], NEC[72], Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN)[68,73–75,77] and the University of Illinois at Urbana-Champaign (UIUC)[35,63,76]. **b**, Length of individual single-walled nanotubes (on a log scale) produced by laser ablation (purple triangles) and chemical vapour deposition (green circles) versus year. Although nanotubes longer than ~1 cm could conceivably be produced, the chambers of scanning electron microscopes are not large enough to characterize such long nanotubes. Ropes and yarns of much longer lengths have since been made. Figure reproduced with permission from ref. 78 (© 2007 World Scientific).

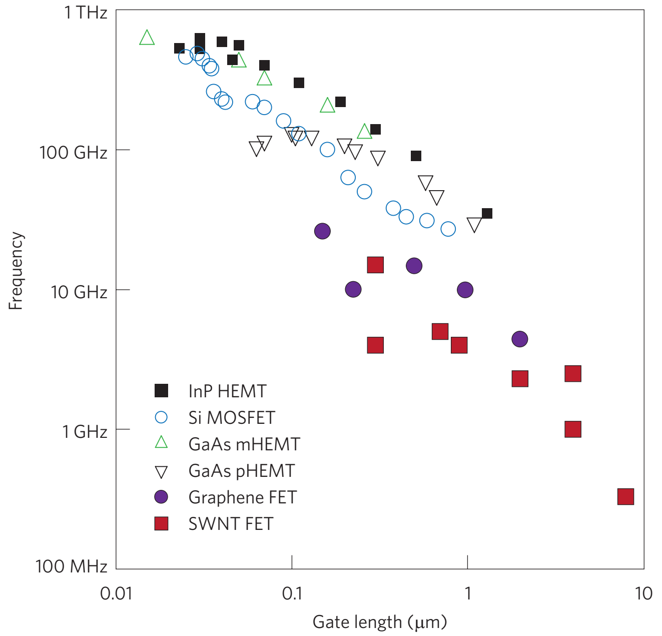


Figure 4 | Frequency performance of different materials. State-of-the-art frequency performance of traditional silicon[65,80–82] devices, III–V semiconductor devices (InP high electron mobility transistor (HEMT)[65], GaAs metamorphic-HEMT[65,83], and GaAs pseudomorphic-HEMT[65]), nanotube-based FETs[63,68,75,76] and graphene FETs[84–86,115] versus gate length. Data points for the nanotube-based FETs are the 'extrinsic' cut-off frequency. Silicon and III–V semiconductor data courtesy of Frank Schwierz.

The evaporating-droplet method has been successful in achieving self-assembled bands of high-density (~10–20 nanotubes μm^{-1}) aligned (within 5° of one another) nanotubes[52] (Fig. 1d). Similarly, using polar and nonpolar features patterned onto the substrate, linear droplet lines were formed and controlled nanotube deposition was achieved[53]. Although the process is conceivably scalable, the formation of periodic bands of aligned nanotubes could limit its utility for certain applications[52,53].

Table 1 summarizes the properties required of the final nanotube array for analogue RF electronics applications. Many of the techniques reviewed above can meet one or more of these metrics, such as diameters in the range 1.5–2 nm (required for high current[54,55]), but no single technique can meet all of them. Therefore, it is likely that some combination of the techniques will be required to meet the final requirements for practical device performance, which we discuss next.

Impact of array density on RF device performance

In the small-signal limit, the a.c. performance of RF transistors can be represented by a linear circuit model consisting of a voltage-dependent current source (the transconductance) plus associated resistances and capacitances (Fig. 2b). Such a model completely describes the input and output impedances, the voltage gain and the current gain, all of which depend on frequency.

Two different definitions of gain are widely used to characterize the frequency response of the transistor[56]: the current gain H_{21} is defined as the output current divided by the input current, and Mason's unilateral gain U is the power gain realized under conjugate impedance-matching at the input and output when the transistor is unilateralized (that is, embedded in a feedback network to isolate the output from the input) using a lossless reciprocal network[57]. For bipolar transistors in the low-frequency limit, H_{21} is better known as β, and can intuitively be considered as the current gain. For FET devices, the current gain is less intuitive, and the cut-off or transition frequency f_T — the frequency at which H_{21} falls to unity (0 dB) — is the most commonly quoted figure of merit, and is defined as such for both bipolar and FET technology. A more useful number for FETs is the maximum frequency of oscillation f_{max}, which is the frequency at which U drops to unity.

Using the effective RF circuit model shown in Fig. 2b, we can express the cut-off frequency of a nanotube FET as:

$$f_T = \frac{g_m}{2\pi} \frac{1}{(C_{gs} + C_{p,gs} + C_{p,gd})((R_{p,s} + R_{p,d})g_d + 1) + C_{p,gd}g_m(R_{p,s} + R_{p,d})} \qquad (1)$$

where g_m is the transconductance, g_d is the drain conductance, C_{gs} is the gate capacitance, $C_{p,gd}$ and $C_{p,gs}$ are the parasitic gate-drain and gate-source capacitances, and $R_{p,s}$ and $R_{p,d}$ are the parasitic series resistance for the source and drain[58]. This is sometimes referred to as the extrinsic cut-off frequency to differentiate it from the intrinsic cut-off frequency (the calculated cut-off frequency when parasitics are ignored). Sometimes, it is numerically justifiable to ignore the effects of parasitic circuit elements, but with nanotube-based FETs they are usually significant at all frequencies. Thus, the intrinsic cut-off frequency is given by:

$$f_{\mathrm{T,intrinsic}} = \frac{g_{\mathrm{m}}}{2\pi C_{\mathrm{gs}}}$$

The intrinsic cut-off frequency can be considered the ultimate frequency performance of the device when it is not slowed down by external circuit elements. As $R_{\mathrm{p,s}}$ and $R_{\mathrm{p,d}}$ are usually external metal electrodes, they can often be made smaller with modest effort. The value of g_{d} would ideally be zero, but in the presence of metallic nanotubes, it can be significant. However, the most important extrinsic element is the parasitic capacitance. For an individual nanotube-based FET, the parasitic capacitance $C_{\mathrm{p,gs}}$ is typically about two orders of magnitude larger than the intrinsic capacitance C_{gs}. (Typically, the values of both $C_{\mathrm{p,gs}}$ and $C_{\mathrm{p,gd}}$ are $\sim 10^{-16}$ F μm^{-1} of the gate width, whereas the C_{gs} of an individual nanotube is $\sim 10^{-17}$ F μm^{-1} of the nanotube length.) This reduces the cut-off frequency of individual nanotube-based FETs by about two orders of magnitude below its intrinsic limit[6,7,59–61].

To achieve the ultimate (intrinsic) limit, one must use very dense, parallel arrays of nanotubes because this increases g_{m} and C_{gs} while keeping the parasitic capacitance approximately constant. The need to use arrays to achieve the best possible performance is the most important conclusion of this Review Article.

To improve the frequency performance it is important to understand how the intrinsic cut-off frequency scales with gate length L_{gate}. First, as C_{gs} is proportional to the gate area, C_{gs} for a nanotube is proportional to L_{gate}. At present, it is not known how g_{m} for a nanotube scales with L_{gate}, so we use classical FET theory as a guide. If L_{gate} is long, the electric field E will be small (because $E = V_{\mathrm{ds}}/L_{\mathrm{gate}}$, where V_{ds} is the drain–source voltage), and the electron drift velocity will be given by $v_{\mathrm{drift}} = \mu E$, where μ is the mobility. On the other hand, if L_{gate} is short, then E will be large, and v_{drift} will saturate at a value denoted by v_{sat}. Knowing v_{drift} we can calculate the transconductance and then the cut-off frequency in these two limits by using the following expression for the drain–source current $I_{\mathrm{ds}} = v_{\mathrm{drift}} ne$, where e is the charge of an individual electron, and the charge density $n = (C_{\mathrm{gs}}/2eL_{\mathrm{gate}})(V_{\mathrm{gs}} - V_{\mathrm{T}})$, where V_{gs} is the gate–source voltage, and the threshold voltage V_{T} is related to the gate– and drain–source voltages by the expression $V_{\mathrm{ds}} = (V_{\mathrm{gs}} - V_{\mathrm{T}})$ in the current-saturation regime. For long gates and small electric fields we find the transconductance to be $\mu C_{\mathrm{gs}}(V_{\mathrm{gs}} - V_{\mathrm{T}})/L_{\mathrm{gate}}^2$; for short gates and large electric fields it is given by $v_{\mathrm{sat}}(C_{\mathrm{gs}}/L_{\mathrm{gate}})$. Consequently, the cut-off frequency can be represented by two limits:

$$f_{\mathrm{T,intrinsic}} \approx \begin{cases} \mu(V_{\mathrm{gate}} - V_{\mathrm{T}})\big/2\pi L_{\mathrm{gate}}^2 & L_{\mathrm{gate}} \text{ large} \\ \dfrac{v_{\mathrm{sat}}}{2\pi L_{\mathrm{gate}}} & L_{\mathrm{gate}} \text{ small} \end{cases}$$

The question of the definition of 'large' versus 'small' depends on the details of the velocity-field curve for carbon nanotubes, which is difficult to measure. Still, it is generally accepted that GHz frequency operation will involve going into the short-gate-length regime, so the mobility will not be the appropriate figure of merit to determine the response time of the transistor. In nanotubes the value of v_{sat} is estimated to be ≈ 1.2–2 × 10^{7} cm s^{-1} (based on carefully modelling both the d.c.[62] and RF[63] performance). Using these values, the predicted 'intrinsic' cut-off frequency will be ≈ 20–30 GHz/L_{gate} (μm) (depending on the value of v_{sat} assumed), which is comparable to the best III–V semiconductors.

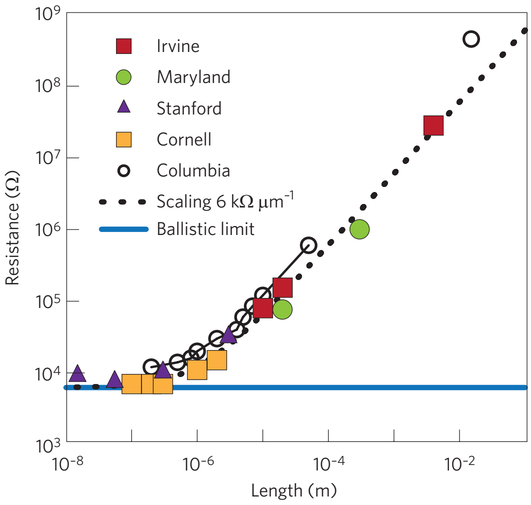


Figure 5 | Resistance performance. Resistance versus length for individual single-walled nanotubes at room temperature (except for the data point at 76 kΩ, 20 μm (left-most green circle), which was measured at 4.2 K; ref. 62) from various labs around the world. The Cornell University data[90] were taken by using an atomic force microscope to measure the voltage drop on an individual nanotube, whereas the Columbia University data[91,92] were taken with multiple contacts on an individual nanotube. The data points from University of California, Irvine[88,89], University of Maryland, College Park[4,62] and Stanford University[93–95] are for distinct nanotubes. All the data are consistent with single-walled nanotubes having a resistance of about 6 kΩ μm^{-1} (dotted line). The ballistic limit (solid blue line) is the lowest contact resistance allowed by quantum mechanics. Reproduced with permission from ref. 87 (© 2009 Wiley).

On the other hand, for long-channel devices (such as printed electronic devices with channel lengths that are longer than 10 μm), the effective mobility determines the cut-off frequency, and here individual nanotubes also have mobilities comparable to the best III–V semiconductors. So far, nanotube-array devices that realize this intrinsic limit have not yet been demonstrated, owing to limitations from parasitic capacitances (see below), but with dense enough arrays, it should be possible to approach this intrinsic speed limit.

In the extreme short-channel limit (where transport is ballistic from source to drain), it has been argued that the carrier-injection velocity into the channel strongly influences the cut-off frequency, so the mobility also becomes important in this limit[64]. Moreover, we should note that the above arguments apply mainly to 'ideal' structures where short-channel effects, parasitic effects and the overall design (for example, metal oxide field-effect transistor (MOSFET) versus high electron mobility transistor (HEMT)) are not important, so they provide only a qualitative guide in the extreme short-channel limit. (See ref. 65 for more details).

How does one construct a thin-film transistor (TFT) that achieves the intrinsic limit discussed above? In general, the best approach is to reduce the relative importance of the parasitic capacitances (which are mainly due to the fringe fields from the electrodes, and depend only mildly on the device geometry). Thus, by increasing the number of nanotubes per width, one increases the transconductance g_{m} without a significant increase in the parasitic capacitance, allowing the ultimate limit to be reached. In this context, it is important to quantify the relationship between the cut-off frequency and the intrinsic cut-off frequency as a function of nanotube array density.

In the limit of sparse nanotube arrays (that is, when the pitch (or spacing) between the nanotubes d is larger than gate–tube

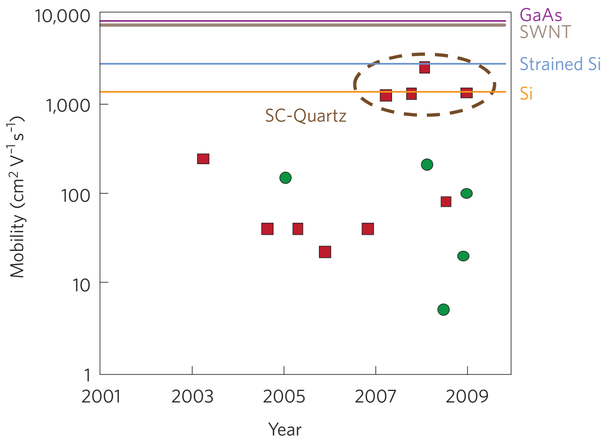


Figure 6 | Mobility performance. For long-channel devices, the mobility is important in achieving a large transconductance and a high cut-off frequency. This plot shows mobility versus year for TFTs made by two methods: devices made from single-walled nanotubes grown by CVD are shown as red squares[35,76,96-103], and devices made from nanotubes deposited from solution are shown as green circles (refs 51,52,104,105 and M. Ishida, S. Toguchi, H. Hongo and F. Nihet, unpublished observation). TFTs grown by CVD on single-crystal quartz substrates (red squares inside dashed line) have the highest mobilities. As a comparison, mobility values for n-type (undoped) silicon, strained silicon, an individual single-walled nanotube (diameter ~2 nm) and gallium arsenide are also shown.

separation), and neglecting $R_{p,s}$ and $R_{p,d}$ in equation (1), the cut-off frequency[7] in the presence of parasitic capacitances can be written as:

$$f_T = f_{T,\text{intrinsic}} \left(\frac{1}{1 + \frac{C_w}{C_{gs,1}} d} \right) \qquad (2)$$

where $C_{gs,1}$ is the nanotube–gate capacitance of an individual nanotube, and C_w is the parasitic capacitance per gate width defined as $(C_{p,gd} + C_{p,gs})/W$, where W is the gate width (see Fig. 2). Typically[6], C_w is ~10^{-16} F μm^{-1} and $C_{gs,1} \approx 10^{-17}$ F × L_{gate} (μm) so that, ideally, one wants the spacing between the nanotubes to be less than 0.1 μm (that is, a density of 10 nanotubes μm^{-1} or higher), for the cut-off frequency not to be significantly degraded by the external (parasitic) capacitance. This is achievable using some of the deposition methods described above.

Although the nanotube density is the critical parameter, g_d, $R_{p,s}$, and $R_{p,d}$ can cause further degradation in f_T as seen in equation (1). At even higher densities, screening by adjacent nanotubes will effect the values (per nanotube) of the transconductance and gate capacitance[35,66]. However, these effects cancel in the calculation of the cut-off frequency, so equation (2) is still valid in the presence of screening, but the value of $C_{gs,1}$ will be reduced compared with the sparse case.

For RF applications, power gain is the important figure of merit (rather than current gain), so f_{max} is also an important parameter. A typical approximation for f_{max} is (see ref. 58):

$$f_{max} \approx \frac{f_T}{2(g_d(R_{p,s} + R_{gate}) + 2\pi f_T C_{p,gd} R_{gate})^{\frac{1}{2}}}$$

where R_{gate} is the gate resistance. The value of f_{max} can be made as high as possible by increasing the density of the nanotubes in the array to make $C_{p,gd}$ as small as possible. However, the presence of metallic nanotubes in the array will lead to a non-zero value of g_d, which will reduce f_{max}, and this is one of the reasons for removing the metallic nanotubes. A comprehensive study of the effects of both R_{gate} and the presence of metallic nanotubes on f_{max} is an important next step in the development of RF devices[67].

Although f_T and f_{max} are generally of the same order of magnitude, either one can be higher than the other depending on the device characteristics (see, for example, Fig. 14 in ref. 65). This is especially important for nanotube transistors, where f_T can be an order of magnitude higher than f_{max} (ref. 68). Thus, both f_T and f_{max} should be compared when comparing the performance of different nanotube transistors.

Devices and measurements

Frequency performance has improved in the past few years, with individual nanotube-based FETs reaching frequencies up to 52 MHz in a multistage ring-oscillator[69–71], and arrays of nanotubes showing cut-off frequencies of up to ~10 GHz (refs 35,60,63,68,72–77; see Fig. 3a). The maximum length of nanotubes has also increased[78] (Fig. 3b). The next challenge on the road to higher frequencies is to increase the nanotube density and the percentage of semiconducting nanotubes.

The highest frequencies reported so far have been for nanotube devices made from samples with about two-thirds semiconducting nanotubes and densities of 5 nanotubes μm^{-1} grown by CVD on quartz[63,76], or from samples that are mostly (90–95%) metallic but have been deposited at higher densities with dielectrophoresis[68,75]. Both device families achieve cut-off frequencies of ~10 GHz with gate lengths ~0.3 μm, indicating that if the fraction of semiconducting nanotubes or density can be improved, the cut-off frequency can be substantially increased. This should be possible by starting with the samples of purified semiconducting nanotubes that have recently become available in a number of labs (see, for example, refs 68 and 79).

To compare nanotubes with other materials, we plot the cut-off frequency versus gate length for nanotubes, graphene and various semiconductors in Fig. 4 (see refs 63,65,68,75,76,80–86). Although it is often assumed that high-mobility materials are needed to make high-speed FETs, this relationship generally only holds true for devices with long channels, as discussed above. For example, for submicrometre gate lengths, the speed advantages of III–V semiconductors such as GaAs and InP over Si-MOSFETs[65] are mainly due to higher saturation velocities. Graphene-based FETs use two-dimensional sheets of carbon atoms as the channel material (as opposed to the one-dimensional tubes of carbon atoms used in nanotube-based FETs), and a recent report of an extrinsic cut-off frequency of ~26 GHz for a 150-nm-gate-length device is on a par with the performance of the best nanotube-based FETs if we allow for the difference in gate length[85] (Fig. 4). However, as described above, the use of denser arrays will lead to increases in the cut-off frequency for nanotube FETs.

In contrast to submicrometre devices, the effective mobility is an important figure-of-merit for TFT devices with long channel lengths. It is generally agreed[3] that electron–phonon scattering limits the peak mobility of an individual nanotube to between 6,000 and 10,000 cm^2 V^{-1} s^{-1}, with the resistance being about 6 kΩ μm^{-1} (Fig. 5 and refs 87–95). The mean free-path inferred from these measurements (at low electric fields) is ~1 μm. For arrays or thin films of nanotubes, the effective mobility is related to the nanotube density, alignment and fraction of semiconducting nanotubes[5]. It is generally believed that a thin film of nanotubes, suitably prepared, should be able to achieve an 'effective' mobility comparable to that of a single nanotube level, but this has not been demonstrated yet.

In Fig. 6, we plot the mobility versus year for nanotube films prepared by the two methods discussed earlier — grow in place with CVD, and deposition from solution — along with the mobility of a pristine nanotube and the mobilities reported for other materials (refs 35,51,52,76,96–105 and M. Ishida, S. Toguchi, H. Hongo and F. Nihet, unpublished observation). We plot mobility values computed using $\mu = (l/WC_{gs})(1/V_{ds})\partial I_{ds}/\partial V_{gs}$ from data measured typically in the linear regime (low V_{ds}). However, devices typically operate in the saturation regime (high V_{ds}), so the mobility numbers quoted in the literature (typically measured at low V_{ds}) are not always a good guide to device performance.

The mobilities of randomly aligned mats of nanotubes grown in place on silicon and those deposited from solution are comparable, with wide scatter due to differences in the nanotube density, average length and, possibly, other parameters. It is generally found that the mobility (which should be independent of gate length for single nanotubes) increases with increasing gate length, even for nanotube films of nominally the same quality. Generally speaking, we still do not have a reliable method for predicting the final device mobility based on the detailed preparation parameters. However, the mobility of nanotube arrays grown by CVD on quartz[35,76,101,102] are much higher than those deposited from solution onto other substrates.

Nanotubes deposited from solution have much higher mobilities than organic semiconductors (which typically have mobilities of ~1 cm^2 V^{-1} s^{-1}; ref. 106), and therefore they could compete with organics in applications that require only moderate mobilities such as low-cost printed electronic circuits. Although techniques for making printed circuits typically achieve resolutions (and hence gate lengths) of ~10 μm, the recent introduction of self-aligned techniques to the manufacture of printed circuits has allowed submicrometre gate lengths to be achieved, even in inkjet printed devices[107]. This approach has been shown to minimize the overlap parasitic capacitance and has made it possible to achieve a cut-off frequency 1.6 MHz from a starting material with a mobility of ~0.2 cm^2 V^{-1} s^{-1} and a gate length of 200 nm. If this new self-aligned approach to making printed electronics could be combined with the nanotube TFTs made with the solution-based approach (which have mobilities up to about 200 cm^2 V^{-1} s^{-1}; M. Ishida, S. Toguchi, H. Hongo and F. Nihet, unpublished observation), it might be possible (neglecting velocity saturation effects discussed above) to increase the cut-off frequency by a factor of 1,000 to give $f_T > 1$ GHz. Such an accomplishment would represent a great leap forward on the road to high-frequency low-cost circuit applications such as all-printed RF identification tags[108].

Demonstrations of nanotubes in RF applications

Recently, several groups have gone beyond device characterization and demonstrated applications in actual radio systems. Although these radios are not yet commercially competitive with existing systems, it is an important milestone to be able to demonstrate operating systems.

Our lab at the University of California, Irivne[109] and another lab at the University of California, Berkeley[110] have used a nanotube as the demodulator in a radio receiver, and have demonstrated a functioning radio that can pick up a signal generated in the lab by a separate generator and play music broadcast wirelessly across a room. Since the demodulation occurs owing to the nonlinearity in the source–drain current–voltage characteristics, it does not matter whether a metallic or semiconducting nanotube is used in this case. The nanotube itself simply detected an amplitude-modulated (AM) signal (replacing the diode in a classical AM radio) and, as such, does not present any particular advantage, other than small size. Moreover, the overall radio system is still large because the external components (the antenna, battery, audio amplifier and so on) are still large (Fig. 7). The UC Berkeley work adds further functionality by using the mechanical resonance frequency of the nanotube as an integrated RF filter, an elegant step towards an integrated nanoradio, but at the cost of requiring a high vacuum. Furthermore, neither of these radios were sensitive enough to receive weak radio signals from local radio stations due to lack of an RF pre-amplifier at the front end.

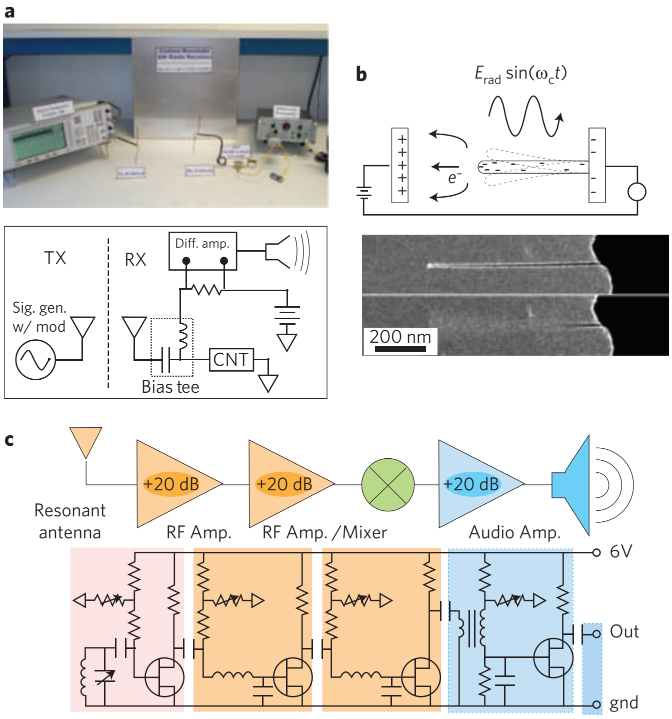


Figure 7 | Nanotubes are performing increasingly complex roles in AM radios. a, A nanotube (CNT) acts as a RF detector in an AM radio. The other components in this demonstration include a signal generator, which is used to transmit (TX) wirelessly an amplitude-modulated signal (sig. gen. w/mod) to the receiver (RX), which consists of a bias tee, a differential amplifier (diff. amp.), a speaker and battery. Reproduced with permission from ref. 109 (© 2007 ACS). **b**, A nanotube in high vacuum acts as a RF detector and an integrated RF filter in an AM radio, where an oscillating electric field (E_{rad}sin($\omega_c t$)) induces the vibration of the tube. Reproduced with permission from ref. 110 (© 2007 ACS). **c**, Nanotube-based FETs act as the RF pre-amplifier, detector (mixer) and audio-frequency amplifier, thus demonstrating a complete AM radio system. Reproduced with permission from ref. 76 (© 2008 PNAS).

A recent collaboration between the University of Illinois at Urbana-Champaign and Northrop Grumman has demonstrated the first RF amplifier based on a nanotube FET, and used it in an entire AM radio system[76]. Separate nanotube transistors also functioned as the RF detector (actually mixer) and audio amplifier. Because an RF pre-amplifier was used, the radio was able to receive weak signals from a local radio station. This demonstrates the application of nanotube electronics into a fully functional system.

Although these demonstrations show that it is possible to make nanoscale components, a true nanoradio would require all the components — including the power source (battery), antenna and the signal-processing elements — to be nanoscale. Using the RF field itself as a power source would completely obviate the need for the battery, while the use of on-chip antennas[111] or even nano-antennas[112,113] would allow for much smaller radios. More research is needed to address the trade-offs between efficiency, required external power, antenna size and heating. Based on standard CMOS technology, we have argued that a single-chip radio system (including antenna and providing space for on-board sensors) of size 100 × 100 × 1 μm is feasible, which begins to approach the size of a single living cell[114]. A true nanoradio should eventually be possible with further developments in nanotechnology.

Summary

To obtain high-performance nanotube-based RF-FETs, dense aligned arrays of all-semiconducting nanotubes are required. Progress in this direction has been rapid, and there are several potential routes towards manufacturing such materials. The advantages of high linearity predicted for one-dimensional materials, together with relaxed manufacturing tolerances, may be the defining advantage over other materials for analogue RF devices. Initial systems have been demonstrated by multiple research labs, and if the previous rate of progress is any indication, it is entirely feasible that, rather than extending Moore's law for digital electronics, the initial point of insertion of nanotube technology into commercial electronics markets will be in wireless communications systems of various kinds.

References

1. Dresselhaus, M. S., Dresselhaus, G. & Eklund, P. C. *Science of Fullerenes and Carbon Nanotubes*. (Academic Press, 1996).
2. Saito, R., Dresselhaus, G. & Dresselhaus, M. S. *Physical Properties of Carbon Nanotubes*. (Imperial College Press, 1998).
3. Zhou, X. J., Park, J. Y., Huang, S. M., Liu, J. & McEuen, P. L. Band structure, phonon scattering, and the performance limit of single-walled carbon nanotube transistors. *Phys. Rev. Lett.* **95,** 146805 (2005).
4. Durkop, T., Getty, S. A., Cobas, E. & Fuhrer, M. S. Extraordinary mobility in semiconducting carbon nanotubes. *Nano Lett.* **4,** 35–39 (2004).
5. Cao, Q. & Rogers, J. Ultrathin films of single-walled carbon nanotubes for electronics and sensors: A review of fundamental and applied aspects. *Adv. Mater.* **21,** 29–53 (2009).
6. Burke, P. J. AC performance of nanoelectronics: Towards a ballistic THz nanotube transistor. *Solid State Electron.* **40,** 1981–1986 (2004).
7. Guo, J., Hasan, S., Javey, A., Bosman, G. & Lundstrom, M. Assessment of high-frequency performance potential of carbon nanotube transistors. *IEEE Trans. Nanotech.* **4,** 715–721 (2005).
8. Alam, K. & Lake, R. Performance of 2 nm gate length carbon nanotube field-effect transistors with source/drain underlaps. *Appl. Phys. Lett.* **87,** 073104 (2005).
9. Hasan, S., Salahuddin, S., Vaidyanathan, M. & Alam, A. A. High-frequency performance projections for ballistic carbon-nanotube transistors. *IEEE Trans. Nanotech.* **5,** 14–22 (2006).
10. Castro, L. C. *et al.* Method for predicting f_T for carbon nanotube FETs. *IEEE Trans. Nanotech.* **4,** 699–704 (2005).
11. Yoon, Y., Ouyang, Y. & Guo, J. Effect of phonon scattering on intrinsic delay and cutoff frequency of carbon nanotube FETs. *IEEE Trans. Electron Dev.* **53,** 2467–2470 (2006).
12. Baumgardner, J. E. *et al.* Inherent linearity in carbon nanotube field-effect transistors. *Appl. Phys. Lett.* **91,** 052107 (2007).
13. Ural, A., Li, Y. M. & Dai, H. J. Electric-field-aligned growth of single-walled carbon nanotubes on surfaces. *Appl. Phys. Lett.* **81,** 3464–3466 (2002).
14. Joselevich, E. & Lieber, C. M. Vectorial growth of metallic and semiconducting single-wall carbon nanotubes. *Nano Lett.* **2,** 1137–1141 (2002).
15. Huang, S. M., Cai, X. Y. & Liu, J. Growth of millimeter-long and horizontally aligned single-walled carbon nanotubes on flat substrates. *J. Am. Chem. Soc.* **125,** 5636–5637 (2003).
16. Huang, S., Woodson, M., Smalley, R. & Liu, J. Growth mechanism of oriented long single walled carbon nanotubes using fast-heating chemical vapor deposition process. *Nano Lett.* **4,** 1025–1028 (2004).
17. Yu, Z., Li, S. & Burke, P. J. Synthesis of aligned arrays of millimeter long, straight single walled carbon nanotubes. *Chem. Mater.* **16,** 3414–3416 (2004).
18. Huang, L. *et al.* Cobalt ultrathin film catalyzed ethanol chemical vapor deposition of single-walled carbon nanotubes. *J. Phys. Chem. B* **110,** 11103–11109 (2006).
19. Ismach, A., Segev, L., Wachtel, E. & Joselevich, E. Atomic-step-templated formation of single wall carbon nanotube patterns. *Angew. Chem. Int. Ed.* **43,** 6140–6143 (2004).
20. Ago, H. *et al.* Aligned growth of isolated single-walled carbon nanotubes programmed by atomic arrangement of substrate surface. *Chem. Phys. Lett.* **408,** 433–438 (2005).
21. Han, S., Liu, X. L. & Zhou, C. W. Template-free directional growth of single-walled carbon nanotubes on a- and r-plane sapphire. *J. Am. Chem. Soc.* **127,** 5294–5295 (2005).
22. Kocabas, C. *et al.* Guided growth of large-scale, horizontally aligned arrays of single-walled carbon nanotubes and their use in thin-film transistors. *Small* **1,** 1110–1116 (2005).
23. Ago, H. *et al.* Competition and cooperation between lattice-oriented growth and step-templated growth of aligned carbon nanotubes on sapphire. *Appl. Phys. Lett.* **90,** 123112 (2007).
24. Ding, L., Yuan, D. N. & Liu, J. Growth of high-density parallel arrays of long single-walled carbon nanotubes on quartz substrates. *J. Am. Chem. Soc.* **130,** 5428–5429 (2008).
25. Zhou, W. W., Rutherglen, C. & Burke, P. Wafer scale synthesis of dense aligned arrays of single-walled carbon nanotubes. *Nano Research* **1,** 158–165 (2008).
26. Kang, S. J. *et al.* Printed multilayer superstructures of aligned single-walled carbon nanotubes for electronic, applications. *Nano Lett.* **7,** 3343–3348 (2007).
27. Zhang, G. *et al.* Selective etching of metallic carbon nanotubes by gas-phase reaction. *Science* **314,** 974–977 (2006).
28. Yang, C. M. *et al.* Preferential etching of metallic single-walled carbon nanotubes with small diameter by fluorine gas. *Phys. Rev. B.* **73,** 075419 (2006).
29. Ding, L. *et al.* Selective growth of well-aligned semiconducting single-walled carbon nanotubes. *Nano Lett.* **9,** 800–805 (2009).
30. Li, Y. *et al.* Preferential growth of semiconducting single-walled carbon nanotubes by a plasma enhanced CVD method. *Nano Lett.* **4,** 317–321 (2004).
31. An, L., Fu, Q., Lu, C. & Liu, J. A simple chemical route to selectively eliminate metallic carbon nanotubes in nanotube network devices. *J. Am. Chem. Soc.* **126,** 10520–10521 (2004).
32. Balasubramanian, K., Sordan, R., Burghard, M. & Kern, K. A selective electrochemical approach to carbon nanotube field-effect transistors. *Nano Lett.* **4,** 827–830 (2004).
33. Strano, M. S. *et al.* Electronic structure control of single-walled carbon nanotube functionalization. *Science* **301,** 1519–1522 (2003).
34. Collins, P. C., Arnold, M. S. & Avouris, P. Engineering carbon nanotubes and nanotube circuits using electrical breakdown. *Science* **292,** 706–709 (2001).
35. Kang, S. J. *et al.* High-performance electronics using dense, perfectly aligned arrays of single-walled carbon nanotubes. *Nature Nanotech.* **2,** 230–236 (2007).
36. Amlani, I. *et al.* in *8th IEEE Conf. Nanotech* 239–242 (IEEE, 2008).
37. Lin, A. *et al.* Threshold voltage and on-off ratio tuning for multiple-tube carbon nanotube FETs. *IEEE Trans. Nanotech.* **8,** 4–9 (2009).
38. Ryu, K. *et al.* CMOS-analogous wafer-scale nanotube-on-insulator approach for submicrometer devices and integrated circuits using aligned nanotubes. *Nano Lett.* **9,** 189–197 (2009).
39. Shim, H., Song, J., Kwak, Y., Kim, S. & Han, C. Preferential elimination of metallic single-walled carbon nanotubes using microwave irradiation. *Nanotechnology* **20,** 065707 (2009).
40. Huang, H., Maruyama, R., Noda, K., Kajiura, H. & Kadono, K. Preferential destruction of metallic single-walled carbon nanotubes by laser irradiation. *J. Phys. Chem. B* **110,** 7316–7320 (2006).
41. Hersam, M. Progress towards monodisperse single-walled carbon nanotubes. *Nature Nanotech.* **3,** 387–394 (2008).
42. Li, X. *et al.* Langmuir-Blodgett assembly of densely aligned single-walled carbon nanotubes from bulk materials. *J. Am. Chem. Soc.* **129,** 4890–4891 (2007).
43. Rutherglen, C., Jain, D. & Burke, P. RF resistance and inductance of massively parallel single walled carbon nanotubes: Direct, broadband measurements and near perfect 50 ohm impedance matching. *Appl. Phys. Lett.* **93,** 083119 (2008).
44. Krupke, R., Linden, S., Rapp, M. & Hennrich, F. Thin films of metallic carbon nanotubes prepared by dielectrophoresis. *Adv. Mater.* **18,** 1468–1468 (2006).
45. Boccaccini, A. R. *et al.* Electrophoretic deposition of carbon nanotubes. *Carbon* **44,** 3149–3160 (2006).
46. Morgan, H. & Green, N. G. *AC Electrokinetics: Colloids and Nanoparticles* (Research Studies Press, 2003).
47. Krupke, R., Hennrich, F., Lohneysen, H. & Kappes, M. M. Separation of metallic from semiconducting single-walled carbon nanotubes. *Science* **301,** 344–347 (2003).
48. Krupke, R., Hennrich, F., Kappes, M. & Lohneysen, H. Surface conductance induced dielectrophoresis of semiconducting single-walled carbon nanotubes. *Nano Lett.* **4,** 1395–1400 (2004).
49. Baik, S., Usrey, M., Rotkina, L. & Strano, M. Using the selective functionalization of metallic single-walled carbon nanotubes to control dielectrophoretic mobility. *J. Phys. Chem. B* **108,** 15560–15564 (2004).
50. Kim, Y. *et al.* Dielectrophoresis of surface conductance modulated single-walled carbon nanotubes using catanionic surfactants. *J. Phys. Chem. B* **110,** 1541–1545 (2006).
51. LeMieux, M. C. *et al.* Self-sorted, aligned nanotube networks for thin-film transistors. *Science* **321,** 101–104 (2008).
52. Engel, M. *et al.* Thin film nanotube transistors based on self-assembled, aligned, semiconducting carbon nanotube arrays. *ACS Nano* **2,** 2445–2452 (2008).
53. Sharma, R., Lee, C. Y., Choi, J. H., Chen, K. & Strano, M. S. Nanometer positioning, parallel alignment, and placement of single anisotropic nanoparticles using hydrodynamic forces in cylindrical droplets. *Nano Lett.* **7,** 2693–2700 (2007).
54. Chen, Z., Appenzeller, J., Knoch, J., Lin, Y.-M. & Avouris, P. The role of metal-nanotube contact in the performance of carbon nanotube field-effect transistors. *Nano Lett.* **5,** 1497–1502 (2005).

55. Kim, W. *et al.* Electrical contacts to carbon nanotubes down to 1 nm in diameter. *Appl. Phys. Lett.* **87,** 173101 (2005).
56. Liu, W. Fundamentals of III-V devices: HBTs, MESFETs, and HFETs/HEMTs. (Wiley, 1999).
57. Gupta, M. S. Power gain in feedback amplifiers, a classic revisited. *IEEE Trans. Microw. Theory* **40,** 864–879 (1992).
58. Schwierz, F. & Liou, J. J. *Modern Microwave Transistors: Theory, Design, and Performance.* (Wiley-Interscience, 2003).
59. Akinwande, D., Close, G. E. & Wong, H. S. P. Analysis of the frequency response of carbon nanotube transistors. *IEEE Trans. Nanotech.* **5,** 599–605 (2006).
60. Wang, D., Yu, Z., McKernan, S. & Burke, P. Ultra high frequency carbon nanotube transistor based on a single nanotube. *IEEE Trans. Nanotech.* **6,** 400–403 (2007).
61. Chaste, J. *et al.* Single carbon nanotube transistor at GHz frequency. *Nano Lett.* **8,** 525–528 (2008).
62. Chen, Y. F. & Fuhrer, M. S. Electric field-dependent charge-carrier velocity in semiconducting carbon nanotubes. *Phys. Rev. Lett.* **95,** 236803 (2005).
63. Kocabas, C. *et al.* High-frequency performance of submicrometer transistors that use aligned arrays of single-walled carbon nanotubes. *Nano Lett.* **8,** 1937–1943 (2009).
64. Lundstrom, M. Elementary scattering theory of the Si MOSFET. *IEEE Electr. Device Lett.* **18,** 361–363 (1997).
65. Schwierz, F. & Liou, J. J. RF transistors: Recent developments and roadmap toward terahertz applications. *Solid State Electron.* **51,** 1079–1091 (2007).
66. Cao, Q. *et al.* Gate capacitance coupling of singled-walled carbon nanotube thin-film transistors. *Appl. Phys. Lett.* **90,** 023516 (2007).
67. Castro, L. C. & Pulfrey, D. L. Extrapolated f_{max} for carbon nanotube field-effect transistors. *Nanotechnology* **17,** 300–304 (2006).
68. Nougaret, L. *et al.* 80 GHz field-effect transistors produced using high purity semiconducting single-walled carbon nanotubes. *Appl. Phys. Lett.* **94,** 243505 (2009).
69. Bachtold, A., Hadley, P., Nakanishi, T. & Dekker, C. Logic circuits with carbon nanotube transistors. *Science* **294,** 1317–1320 (2001).
70. Javey, A., Wang, Q., Ural, A., Li, Y. M. & Dai, H. J. Carbon nanotube transistor arrays for multistage complementary logic and ring oscillators. *Nano Lett.* **2,** 929–932 (2002).
71. Chen, Z. H. *et al.* An integrated logic circuit assembled on a single carbon nanotube. *Science* **311,** 1735–1735 (2006).
72. Narita, K., Hongo, H., Ishida, M. & Nihey, F. High-frequency performance of multiple-channel carbon nanotube transistors. *Phys. Status Solidi A* **204,** 1808–1813 (2007).
73. Bethoux, J. M. *et al.* Active properties of carbon nanotube field-effect transistors deduced from S parameters measurements. *IEEE Trans. Nanotech.* **5,** 336–342 (2006).
74. Bethoux, J. M. *et al.* An 8-GHz f_T carbon nanotube field-effect transistor for gigahertz range applications. *IEEE Electron Dev. Lett.* **27,** 681–683 (2006).
75. Louarn, A. L. *et al.* Intrinsic current gain cutoff frequency of 30 GHz with carbon nanotube transistors. *Appl. Phys. Lett.* **90,** 233108 (2007).
76. Kocabas, C. *et al.* Radio frequency analog electronics based on carbon nanotube transistors. *Proc. Natl Acad. Sci. USA* **105,** 1405–1409 (2008).
77. Chimot, N. *et al.* Gigahertz frequency flexible carbon nanotube transistors. *Appl. Phys. Lett.* **91,** 153111 (2007).
78. Burke, P. J. *Nanotubes and Nanowires* (World Scientific, 2007).
79. Rutherglen, C. *Carbon Nanotube Based Analog RF Devices* PhD thesis, Univ. California, Irvine (2009).
80. Dimitrov, V. *et al.* Small-signal performance and modeling of sub-50 nm nMOSFETs with f_T above 460 GHz. *Solid State Electron.* **52,** 899–908 (2008).
81. Stork, J. in *Proc. Symp. VLSI Tech. Dig.* 1–2 (IEEE, 2006).
82. Lee, S. *et al.* in *Electron Devices Meeting IEDM* 255–258 (IEEE, 2007).
83. Yeon, S., Park, M., Choi, J. & Seo, K. in *Electron Devices Meeting IEDM* 613–616 (IEEE, 2007).
84. Moon, J. S. *et al.* Epitaxial-graphene RF field-effect transistors on Si-face 6H-SiC substrates. *IEEE Electr. Device Lett.* **30,** 650–652 (2009).
85. Lin, Y. *et al.* Operation of graphene transistor at gigahertz frequencies. *Nano Lett.* **9,** 422–426 (2009).
86. Meric, I., Baklitskaya, P., Kim, P. & Shepard, K. RF performance of top-gated, zero-bandgap graphene field-effect transistor. *Electron Devices Meeting IEDM* 1–4 (IEEE, 2008).
87. Rutherglen, C. & Burke, P. Nanoelectromagnetics: Circuit and electromagnetic properties of carbon nanotubes. *Small* **5,** 884–906 (2009).
88. Li, S., Yu, Z., Yen, S. F., Tang, W. C. & Burke, P. J. Carbon nanotube transistor operation at 2.6 GHz. *Nano Lett.* **4,** 753–756 (2004).
89. Li, S. D., Yu, Z., Rutherglen, C. & Burke, P. J. Electrical properties of 0.4 cm long single-walled carbon nanotubes. *Nano Lett.* **4,** 2003–2007 (2004).
90. Park, J. Y. *et al.* Electron-phonon scattering in metallic single-walled carbon nanotubes. *Nano Lett.* **4,** 517–520 (2004).
91. Hong, B. H. *et al.* Quasi-continuous growth of ultralong carbon nanotube arrays. *J. Am. Chem. Soc.* **127,** 15336–15337 (2005).
92. Purewal, M. S. *et al.* Scaling of resistance and electron mean free path of single-walled carbon nanotubes. *Phys. Rev. Lett.* **98,** 186808 (2007).
93. Javey, A., Guo, J., Wang, Q., Lundstrom, M. & Dai, H. J. Ballistic carbon nanotube field-effect transistors. *Nature* **424,** 654–657 (2003).
94. Javey, A., Qi, P. F., Wang, Q. & Dai, H. J. Ten- to 50-nm-long quasi-ballistic carbon nanotube devices obtained without complex lithography. *Proc. Natl Acad. Sci. USA* **101,** 13408–13410 (2004).
95. Javey, A. *et al.* High-field quasiballistic transport in short carbon nanotubes. *Phys. Rev. Lett.* **92,** 106804 (2004).
96. Snow, E. S., Novak, J. P., Campbell, P. M. & Park, D. Random networks of carbon nanotubes as an electronic material. *Appl. Phys. Lett.* **82,** 2145–2147 (2003).
97. Zhou, Y. *et al.* P-channel, n-channel thin film transistors and p–n diodes based on single wall carbon nanotube networks. *Nano Lett.* **4,** 2031–2036 (2004).
98. Ozel, T., Gaur, A., Rogers, J. & Shim, M. Polymer electrolyte gating of carbon nanotube network transistors. *Nano Lett.* **5,** 905–911 (2005).
99. Hur, S. *et al.* Printed thin-film transistors and complementary logic gates that use polymer-coated single-walled carbon nanotube networks. *J. Appl. Phys.* **98,** 114302 (2005).
100. Cao, Q. *et al.* Medium-scale carbon nanotube thin-film integrated circuits on flexible plastic substrates. *Nature* **454,** 495–500 (2008).
101. Ishikawa, F. N. *et al.* Transparent electronics based on transfer printed aligned carbon nanotubes on rigid and flexible substrates. *ACS Nano* **3,** 73–79 (2009).
102. Kocabas, C., Kang, S. J., Ozel, T., Shim, M. & Rogers, J. A. Improved synthesis of aligned arrays of single-walled carbon nanotubes and their implementation in thin film type transistors. *J. Phys. Chem. C* **111,** 17879–17886 (2007).
103. Cao, Q., Xia, M., Shim, M. & Rogers, J. Bilayer organic-inorganic gate dielectrics for high-performance, low-voltage, single-walled carbon nanotube thin-film transistors, complementary logic gates, and p–n diodes on plastic substrates. *Adv. Funct. Mater.* **16,** 2355–2362 (2006).
104. Snow, E. S., Campbell, P. M., Ancona, M. G. & Novak, J. P. High-mobility carbon-nanotube thin-film transistors on a polymeric substrate. *Appl. Phys. Lett.* **86,** 033105 (2005).
105. Kanungo, M., Lu, H., Malliaras, G. & Blanchet, G. Suppression of metallic conductivity of single-walled carbon nanotubes by cycloaddition reactions. *Science* **323,** 234–237 (2009).
106. Bao, Z. & Locklin, J. J. *Organic Field-Effect Transistors* (CRC Press, 2007).
107. Noh, Y. Y., Zhao, N., Caironi, M. & Sirringhaus, H. Downscaling of self-aligned, all-printed polymer thin-film transistors. *Nature Nanotech.* **2,** 784–789 (2007).
108. Subramanian, V. *et al.* Progress toward development of all-printed RFID tags: Materials, processes, and devices. *Proc. IEEE* **93,** 1330–1338 (2005).
109. Rutherglen, C. & Burke, P. Carbon nanotube radio. *Nano Lett.* **7,** 3296–3299 (2007).
110. Jensen, K., Weldon, J., Garcia, H. & Zettl, A. Nanotube radio. *Nano Lett.* **7,** 3508–3511 (2007).
111. O, K. *et al.*, On-chip antennas in silicon ICs and their application. *IEEE Trans. Electron Dev.* **52,** 1312–1323 (2005).
112. Burke, P. J., Li, S. D. & Yu, Z. Quantitative theory of nanowire and nanotube antenna performance. *IEEE Trans. Nanotech.* **5,** 314–334 (2006).
113. Hanson, G. W. Fundamental transmitting properties of carbon nanotube antennas. *IEEE Trans. Antenn. Propag.* **53,** 3426–3435 (2005).
114. Burke, P. & Rutherglen, C. Towards a single-chip, implantable RFID system: Is a single-cell radio possible? *Biomed. Microdevices* doi:10.1007/s10544-008-9266-4 (2009).
115. Farmer, D. B. *et al.* Utilization of a buffered dielectric to achieve high field-effect carrier mobility in graphene transistors. *Nano Lett.* doi:10.1021/nl902788u (2009).

Acknowledgements

This work was funded by the National Science Foundation, the Army Research Office, the Office of Naval Research, Northrop Grumman and the Korean National Science Foundation (KOSEF) World Class University (WCU) programme.

Additional information

The authors declare competing financial interests: details accompany the paper at www.nature.com/naturenanotechnology.

nature nanotechnology

REVIEW ARTICLE

PUBLISHED ONLINE: 30 MAY 2010 | DOI: 10.1038/NNANO.2010.89

Graphene transistors

Frank Schwierz[1]*

Graphene has changed from being the exclusive domain of condensed-matter physicists to being explored by those in the electron-device community. In particular, graphene-based transistors have developed rapidly and are now considered an option for post-silicon electronics. However, many details about the potential performance of graphene transistors in real applications remain unclear. Here I review the properties of graphene that are relevant to electron devices, discuss the trade-offs among these properties and examine their effects on the performance of graphene transistors in both logic and radiofrequency applications. I conclude that the excellent mobility of graphene may not, as is often assumed, be its most compelling feature from a device perspective. Rather, it may be the possibility of making devices with channels that are extremely thin that will allow graphene field-effect transistors to be scaled to shorter channel lengths and higher speeds without encountering the adverse short-channel effects that restrict the performance of existing devices. Outstanding challenges for graphene transistors include opening a sizeable and well-defined bandgap in graphene, making large-area graphene transistors that operate in the current-saturation regime and fabricating graphene nanoribbons with well-defined widths and clean edges.

Every now and again, a single paper ignites a revolution in science and technology. Such a revolution was started in October 2004, when condensed-matter physicists reported that they had prepared graphene—two-dimensional sheets of carbon atoms—and observed the electric field effect in their samples[1]. It was not long before this new material attracted the attention of the electron-device community, and today a growing number of groups are successfully fabricating graphene transistors. Major chip-makers are now active in graphene research and the International Technology Roadmap for Semiconductors, the strategic planning document for the semiconductor industry, considers graphene to be among the candidate materials for post-silicon electronics[2].

Several excellent reviews on the basic science of graphene have been published in recent years[3–5]. Given the growing interest in graphene in the electron-device community, and ongoing discussions of the potential of graphene transistors, a review article focusing on graphene devices is timely. Here, from the point of view of a device engineer, I discuss the potential of graphene as a new material for electron devices, and summarize the state of the art for graphene transistors. I will focus mostly on the field-effect transistor (FET), because this is the most successful device concept in electronics and because most work on graphene devices so far has been related to FETs.

Two principal divisions of semiconductor electronics are digital logic devices and radiofrequency devices. The degree of readiness to introduce new device concepts is generally higher for radiofrequency applications, in part because the fortunes of digital logic depend almost entirely on the performance of a single type of device: the silicon metal–oxide–semiconductor FET (MOSFET). For decades, making MOSFETs smaller has been key to the progress in digital logic. This size scaling has enabled the complexity of integrated circuits to double every 18 months, leading to significant improvements in performance and decreases in price per transistor[6,7]. Today, processors containing two billion MOSFETs, many with gate lengths of just 30 nm, are in mass production (Fig. 1).

Because the fabrication of integrated circuits is highly complex, semiconductor fabrication plants are extremely expensive (at present costing several billion US dollars). Furthermore, because scaling alone has provided the needed performance improvements from one generation of integrated circuits to the next, there has been little motivation for the chip-makers to introduce devices based on a fundamentally different physics or on a material other than silicon. However, there is a consensus in the community that MOSFET scaling is approaching its limits and that, in the long run, it will be necessary to introduce new material and device concepts to ensure that performance continues to improve.

The situation is different for radiofrequency electronics. This field was dominated by defence applications until the late 1980s, and although it moved into the mainstream in the 1990s owing to advances in wireless communications, the military continued to provide generous financial support for research into new radiofrequency devices. This, together with the fact that radiofrequency circuits are much less complex than digital logic chips, has led to makers of radiofrequency chips being more open to new device concepts. An indication of this is the large variety of different transistor types and materials used in radiofrequency electronics: these include high-electron-mobility transistors (HEMTs) based on III–V semiconductors such as GaAs and InP, silicon n-channel MOSFETs, and different types of bipolar transistor[8,9].

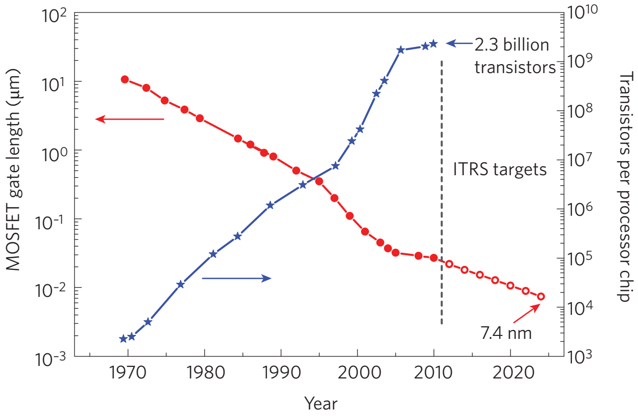


Figure 1 | Trends in digital electronics. Evolution of MOSFET gate length in production-stage integrated circuits (filled red circles) and International Technology Roadmap for Semiconductors (ITRS) targets (open red circles). As gate lengths have decreased, the number of transistors per processor chip has increased (blue stars). Maintaining these trends is a significant challenge for the semiconductor industry, which is why new materials such as graphene are being investigated.

[1]Technische Universität Ilmenau, Postfach 100565, 98694 Ilmenau, Germany. *e-mail: frank.schwierz@tu-ilmenau.de

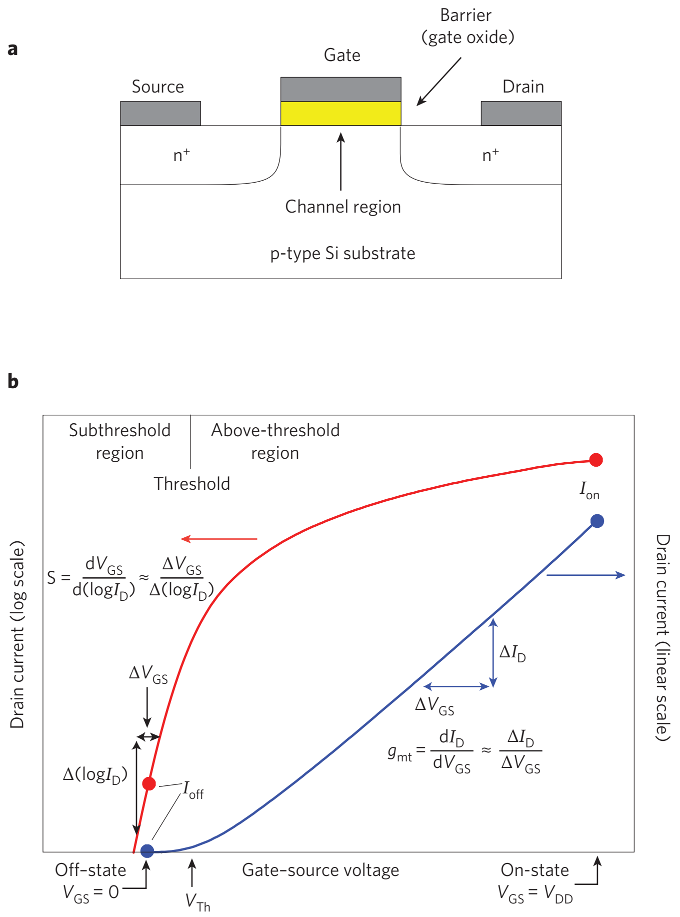


Figure 2 | Conventional FETs. a, Cross-section of an n-channel Si MOSFET. When the voltage applied between the source and gate electrodes exceeds a threshold voltage, V_{Th}, a conducting channel is formed and a drain current, I_D, flows. The length of the channel is defined by the length of the gate electrode; the thickness of the gate-controlled channel region is the depth to which the electronic properties of the semiconductor (p-doped Si in this case) are influenced by the gate. **b**, FET transfer characteristics showing I_D (on a logarithmic scale on the left and a linear scale on the right) versus the gate–source voltage, V_{GS}. The transistor is considered to be switched on when V_{GS} is equal to the maximum voltage supplied to the device, V_{DD}. The higher the slope in the subthreshold region ($V_{GS} < V_{Th}$), the better the transistor switch-on characteristics become. Above threshold, the change in I_D for a given change in V_{GS} is called the terminal transconductance, g_{mt}.

As I discuss below, graphene is potentially well suited to radiofrequency applications because of its promising carrier transport properties and its purely two-dimensional structure. This, combined with the relative openness of the radiofrequency-electronics industry to new materials, suggests that graphene might make its first appearance in radiofrequency applications rather than in logic circuits.

FET physics: what really matters

A FET consists of a gate, a channel region connecting source and drain electrodes, and a barrier separating the gate from the channel (Fig. 2a). The operation of a conventional FET relies on the control of the channel conductivity, and thus the drain current, by a voltage, V_{GS}, applied between the gate and source.

For high-speed applications, FETs should respond quickly to variations in V_{GS}; this requires short gates and fast carriers in the channel. Unfortunately, FETs with short gates frequently suffer from degraded electrostatics and other problems (collectively known as short-channel effects), such as threshold-voltage roll-off, drain-induced barrier lowering, and impaired drain-current saturation[7,10]. Scaling theory predicts that a FET with a thin barrier and a thin gate-controlled region (measured in the vertical direction in Fig. 2a) will be robust against short-channel effects down to very short gate lengths (measured in the horizontal direction in Fig. 2a)[11]. The possibility of having channels that are just one atomic layer thick is perhaps the most attractive feature of graphene for use in transistors. (Mobility, which is often considered to be graphene's most useful property for applications in nanoelectronics, is discussed later.) By comparison, the channels in III–V HEMTs are typically 10–15 nm thick, and although silicon-on-insulator MOSFETs with channel (that is, silicon body) thicknesses of less than 2 nm have been reported[12], rough interfaces caused their mobility to deteriorate. More importantly, the body of these MOSFETs showed thickness fluctuations that will lead to unacceptably large threshold-voltage variations (and similar problems are expected to occur when the thickness of the channel in a III–V HEMT is reduced to only a few nanometres). These problems occur at thicknesses that are many times greater than the thickness of graphene.

The series resistances between the channel and the source and drain terminals are also important, and their adverse impact on the FET becomes more pronounced as the gate length decreases[13]. Thus, device engineers devote considerable effort to developing transistor designs in which short-channel effects are suppressed and series resistances are minimized.

Modern digital logic is based on silicon complementary metal oxide semiconductor (CMOS) technology. CMOS logic gates consist of both n- and p-channel MOSFETs that can switch between the on-state (with a large on-current, I_{on}, and $V_{GS} = \pm V_{DD}$, where V_{DD} is the maximum voltage supplied to the device) and the off-state (with a small off-current, I_{off}, and $V_{GS} = 0$). In the terminology of digital logic, a gate is not the gate terminal of a transistor but a combination of two or more transistors that can perform a certain logic operation. The value of V_{GS} at which the FET is just on the verge of switching on is the threshold voltage, V_{Th}. Figure 2b shows the transfer characteristics of an n-channel FET indicating the on-state and the off-state. Useful measures with which to assess the switching behaviour are the subthreshold swing, S (relevant to the subthreshold region), and the terminal transconductance, g_{mt} (relevant to the above-threshold region).

In the steady state, a certain number of the MOSFETs in a CMOS logic gate are always switched off so that no current—except the small I_{off}—flows through the gate[14]. The ability of silicon MOSFETs to switch off enables silicon CMOS to offer extremely low static power dissipation (which is the reason why silicon CMOS has bested all competing logic technologies). Thus, any successor to the silicon MOSFET that is to be used in CMOS-like logic must have excellent switching capabilities, as well as an on–off ratio, I_{on}/I_{off}, of between 10^4 and 10^7 (ref. 2). In a conventional FET, this requires semiconducting channels with a sizeable bandgap, preferably 0.4 eV or more. Moreover, n- and p-channel FETs with symmetrical threshold voltages, that is, with $V_{Th,n} = -V_{Th,p}$, are needed for proper CMOS operation.

In radiofrequency applications, however, switch-off is not required *per se*. In small-signal amplifiers, for example, the transistor is operated in the on-state and small radiofrequency signals that are to be amplified are superimposed onto the d.c. gate–source voltage. To discuss the radiofrequency performance of FETs, I use the equivalent circuit from Fig. 3a and focus on the cut-off frequency, f_T, which is the frequency at which the magnitude of the small-signal current gain rolls off to unity. The cut-off frequency is the most widely used figure of merit for radiofrequency devices and is, in effect, the highest frequency at which a FET is useful in radiofrequency applications.

As can be seen from the expression for f_T given in Table 1 (refs 7,8), the cut-off frequency can be maximized by making the intrinsic transconductance, g_m, as large as possible and making the drain conductance, g_{ds}, and all the capacitances and resistances in the equivalent circuit (Fig. 3) as small as possible[7,8]. However, the values of all these quantities vary with the applied d.c. gate–source voltage, V_{GS}, and the applied d.c. drain–source voltage, V_{DS}. As shown exemplarily for a typical GaAs HEMT[15,16] (Fig. 3b,c), V_{DS} has a pronounced effect on the FET performance. For this transistor, f_T peaks around V_{DS} = 1 V, that is, deep in the region of drain-current saturation, where g_m is near its peak and g_{ds} has decreased sufficiently. For lower values of V_{DS}, the device operates in the linear regime and the cut-off frequency is low because g_m is small and g_{ds} is large.

The bottom line for radiofrequency performance is that although shorter gates, faster carriers and lower series resistances all lead to higher cut-off frequencies, saturation of the drain current is essential to reach the maximum possible operating speeds. This point is frequently missed in discussions of transistor speeds. Drain-current saturation is also necessary to maximize the intrinsic gain, $G_{int} = g_m/g_{ds}$, which has become a popular figure of merit for mixed-signal circuits.

Graphene properties relevant to transistors

Single-layer graphene is a purely two-dimensional material. Its lattice consists of regular hexagons with a carbon atom at each corner. The bond length between adjacent carbon atoms, L_b, is 1.42 Å and the lattice constant, a, is 2.46 Å (Fig. 4a). The first reports on this material appeared decades ago, even before the name graphene had been coined (see, for example, refs 17–19), but it took the pioneering 2004 paper by the Manchester group[1] to spark the present explosion of interest in the material.

At present, the most popular approaches to graphene preparation are mechanical exfoliation[1], growth on metals and subsequent graphene transfer to insulating substrates[20,21], and thermal decomposition of SiC to produce so-called epitaxial graphene on top of SiC wafers[22,23]. Exfoliation is still popular for laboratory use but it is not suited to the electronics industry, whereas the other two options both have the potential for producing wafer-scale graphene. After the graphene has been prepared, common semiconductor processing techniques (such as lithography, metallization and etching) can be applied to fabricate graphene transistors.

In this section, I discuss two important aspects of graphene: the presence (or otherwise) of a bandgap, and charge transport (mobility and high-field transport) at room temperature.

Bandgap. Large-area graphene is a semimetal with zero bandgap. Its valence and conduction bands are cone-shaped and meet at the K points of the Brillouin zone (Fig. 4b). Because the bandgap is zero, devices with channels made of large-area graphene cannot be switched off and therefore are not suitable for logic applications. However, the band structure of graphene can be modified, and it is possible to open a bandgap in three ways: by constraining large-area graphene in one dimension to form graphene nanoribbons, by biasing bilayer graphene and by applying strain to graphene. See Table 2 and refs 24–43 for more details.

It has been predicted[28] that both armchair nanoribbons and zigzag nanoribbons (the two ideal types of nanoribbon; Fig. 4a) have a bandgap that is, to a good approximation, inversely proportional to the width of the nanoribbon. The opening of a bandgap in nanoribbons has been verified experimentally for widths down to about 1 nm (refs 24–27), and theory and experiments both reveal bandgaps in excess of 200 meV for widths below 20 nm (Fig. 4c). However, it should be noted that real nanoribbons have rough edges and widths that change along their lengths. Even modest edge disorder obliterates any difference in the bandgap between nanoribbons with different edge geometries[29], and edge functionalization and doping can also affect the bandgap[44].

To open a bandgap useful for conventional field-effect devices, very narrow nanoribbons with well-defined edges are needed. This represents a serious challenge given the semiconductor processing equipment available at the moment. Recently, nanoribbons that were uniform in width and had reduced edge roughness were produced by 'unzipping' carbon nanotubes[45]. However, even a perfect nanoribbon is not perfect for electronics applications. In general, the larger the bandgap that opens in a nanoribbon, the more the

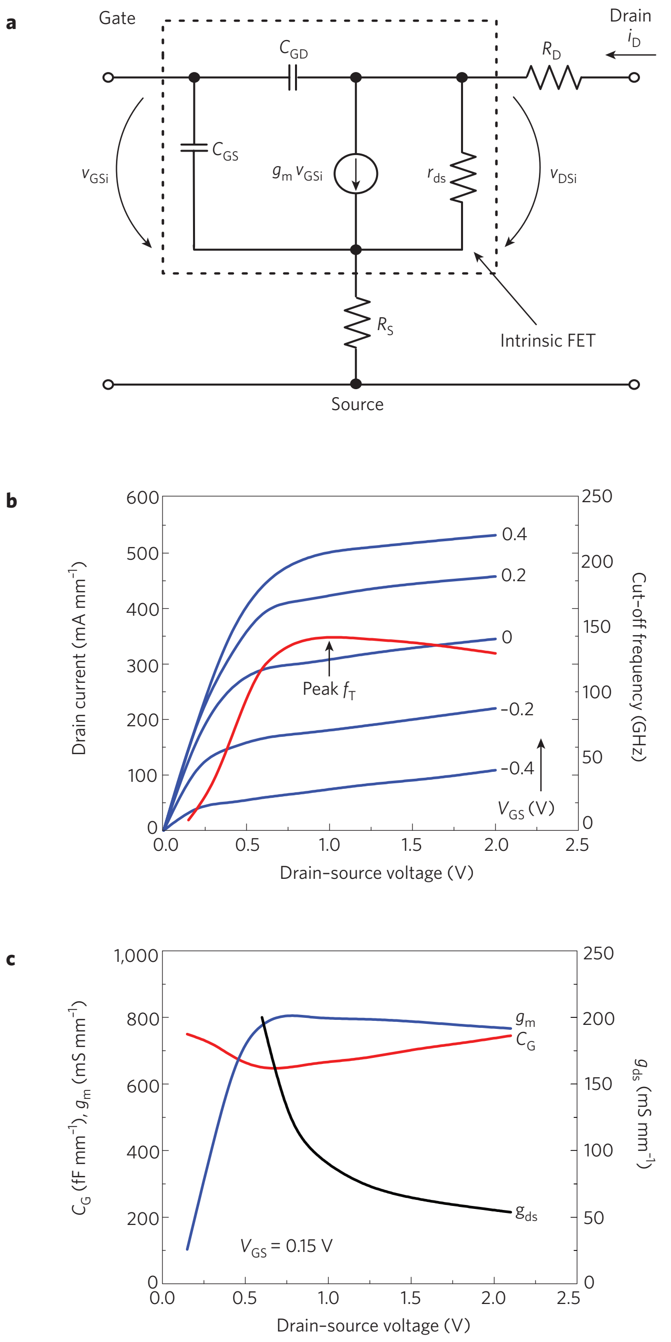


Figure 3 | FET d.c. and small-signal operation. a, Small-signal equivalent FET circuit. The intrinsic transconductance, g_m, is related to the internal small-signal gate–source and drain–source voltages, v_{GSi} and v_{DSi}, whereas the terminal transconductance, g_{mt}, is related to the applied gate–source and drain–source voltages, V_{GS} and V_{DS} (Table 1 and Fig. 2b). **b**, The drain current, I_D (blue lines), at different values of V_{GS}, and the cut-off frequency, f_T (red line), both versus V_{DS} for a radiofrequency GaAs high-electron-mobility transistor[15,16]. The cut-off frequency peaks at V_{DS} = 1 V and V_{GS} = 0.15 V. **c**, The intrinsic transconductance (blue line), the overall gate capacitance, $C_G = C_{GS} + C_{GD}$ (red line), and the drain conductance, g_{ds} ($1/r_{ds}$; black line), versus V_{DS} for the same FET.

Table 1 | Performance measures for the field-effect transistor.

Quantity	Definition
Terminal transconductance	$g_{\mathrm{mt}} = \left.\dfrac{\mathrm{d}I_{\mathrm{D}}}{\mathrm{d}V_{\mathrm{GS}}}\right\|_{V_{\mathrm{DS}}=\mathrm{const}}$
Intrinsic transconductance	$g_{\mathrm{m}} = \left.\dfrac{\mathrm{d}I_{\mathrm{D}}}{\mathrm{d}V_{\mathrm{GSi}}}\right\|_{V_{\mathrm{DSi}}=\mathrm{const}}$
Drain conductance	$g_{\mathrm{ds}} = \dfrac{1}{r_{\mathrm{ds}}} = \left.\dfrac{\mathrm{d}I_{\mathrm{D}}}{\mathrm{d}V_{\mathrm{DSi}}}\right\|_{V_{\mathrm{GSi}}=\mathrm{const}}$
Gate–source capacitance	$C_{\mathrm{GS}} = -\left.\dfrac{\mathrm{d}Q_{\mathrm{ch}}}{\mathrm{d}V_{\mathrm{GSi}}}\right\|_{V_{\mathrm{DSi}}=\mathrm{const}}$
Gate–drain capacitance	$C_{\mathrm{GD}} = -\left.\dfrac{\mathrm{d}Q_{\mathrm{ch}}}{\mathrm{d}V_{\mathrm{DSi}}}\right\|_{V_{\mathrm{GSi}}=\mathrm{const}}$
Cut-off frequency	$f_{\mathrm{T}} \approx \dfrac{g_{\mathrm{m}}}{2\pi}\,\dfrac{1}{(C_{\mathrm{GS}}+C_{\mathrm{GD}})[1+g_{\mathrm{ds}}(R_{\mathrm{S}}+R_{\mathrm{D}})]+C_{\mathrm{GD}}g_{\mathrm{m}}(R_{\mathrm{S}}+R_{\mathrm{D}})}$
Field-effect mobility	$\mu_{\mathrm{FE}} = \dfrac{L_{\mathrm{ch}}g_{\mathrm{m}}}{W_{\mathrm{ch}}C_{\mathrm{G}}V_{\mathrm{DS}}}$

V_{GS}, V_{DS}: terminal d.c. voltages; V_{GSi}, V_{DSi}: intrinsic d.c. voltages; Q_{ch}: mobile channel charge; L_{ch}, W_{ch}: channel length and width; C_{G}: gate capacitance. In the expression for μ_{FE}, C_{G} is the gate capacitance per unit area. R_{S} and R_{D} are the source and drain series resistances, respectively. Expressions for the terminal and intrinsic transconductances, drain conductance, gate–source and gate–drain capacitances, and cut-off frequency for the equivalent FET circuit shown in Fig. 3a[7,8]. The expression for the field-effect mobility in MOS channels is also shown[66].

valence and conduction bands become parabolic (rather than cone-shaped): this decreases the curvature around the K point and increases the effective mass of the charge carriers[46], which is likely to decrease the mobility.

Bilayer graphene is also gapless (Fig. 4b), and its valence and conduction bands have a parabolic shape near the K point. If an electric field is applied perpendicular to the bilayer, a bandgap opens and the bands near the K point take on the so-called Mexican-hat shape. This opening was predicted by theory[30,31] and has been verified in experiments[32,33]. Theoretical investigations have also shown that the size of the bandgap depends on the strength of the perpendicular field and can reach values of 200–250 meV for high fields ((1–3) × 10^7 V cm^{-1}; refs 30,31).

The bandgap of large-area single-layer epitaxial graphene is at present the subject of controversy[34]. Although some results suggest a zero bandgap[37,38], others report a bandgap of around 0.25 eV (refs 35,36). The transfer characteristics of epitaxial-graphene MOSFETs show no switch-off, which suggests a zero bandgap. However, a bandgap is consistently observed for epitaxial bilayer graphene[38,39].

Finally, strain has been discussed as a means of opening a bandgap in large-area graphene, and the effect of uniaxial strain on the band structure has been simulated[40,41]. At present it seems that if it is possible at all, opening a gap in this way will require a global uniaxial strain exceeding 20%, which will be difficult to achieve in practice. Moreover, little is known about the ways in which other types of strain, such as biaxial strain and local strain, influence the band structure of graphene.

Thus, although there are a number of techniques for opening a bandgap in graphene, they are all at the moment some way from being suitable for use in real-world applications.

Mobility. The most frequently stated advantage of graphene is its high carrier mobility at room temperature. Mobilities of 10,000–15,000 cm^2 V^{-1} s^{-1} are routinely measured for exfoliated graphene on SiO_2-covered silicon wafers[1,47], and upper limits of between 40,000 and 70,000 cm^2 V^{-1} s^{-1} have been suggested[47,48]. Moreover, in the absence of charged impurities and ripples, mobilities of 200,000 cm^2 V^{-1} s^{-1} have been predicted[49], and a mobility of 10^6 cm^2 V^{-1} s^{-1} was recently reported for suspended graphene[50]. For large-area graphene grown on nickel and transferred to a substrate, mobilities greater than 3,700 cm^2 V^{-1} s^{-1} have been measured[20].

Finally, for epitaxial graphene on silicon carbide, the mobility depends on whether the graphene is grown on the silicon face or the carbon face of SiC. Although graphene grown on the carbon face has higher mobility (values of ~5,000 cm^2 V^{-1} s^{-1} have been reported[23], compared with ~1,000 cm^2 V^{-1} s^{-1} for graphene grown on the silicon face[23,51]), it is easier to grow single-layer and bilayer graphene on the silicon face, which makes the silicon face of SiC more suited for electronic applications.

In early graphene MOS structures, the mobility was affected by the use of a top-gate dielectric[52,53]. However, the recent demonstration of mobilities of around 23,000 cm^2 V^{-1} s^{-1} in top-gated graphene MOS channels[54] and the observation of similar mobilities before and after top-gate formation[55] show that high-mobility graphene

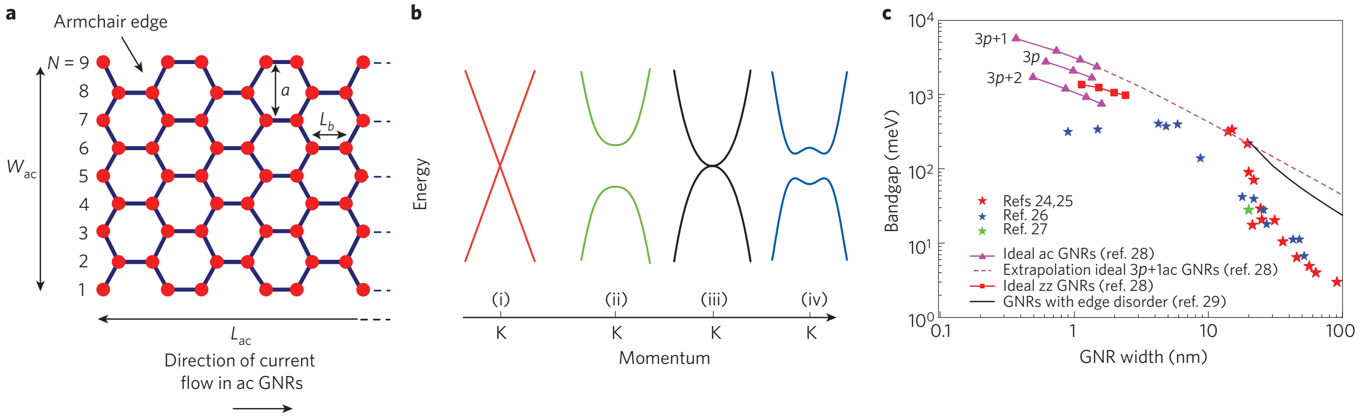


Figure 4 | Properties of graphene and graphene nanoribbons. a, Schematic of an armchair (ac) graphene nanoribbon (GNR) of length L_{ac} and width W_{ac}. The nanoribbon shown here has $N = 9$ carbon atoms along its width and thus belongs to the 3p family, where p is an integer. **b**, Band structure around the K point of (i) large-area graphene, (ii) graphene nanoribbons, (iii) unbiased bilayer graphene, and (iv) bilayer graphene with an applied perpendicular field. Large-area graphene and unbiased bilayer graphene do not have a bandgap, which makes them less useful for digital electronics. **c**, Bandgap versus nanoribbon width from experiments[24–27] and calculations[28,29]. By comparison, the bandgap of Si is above 1 eV. zz: zigzag.

MOS channels can be made with a proper choice of the gate dielectric and optimization of the deposition process.

These mobility numbers are impressive, but they require closer inspection. The high mobilities mentioned above relate to large-area graphene, which is gapless. A general trend for conventional semiconductors is that the electron mobility decreases as the bandgap increases, and a similar trend has been predicted for carbon nanotubes (CNTs)[56,57] and graphene nanoribbons[58–61] (Fig. 5a). This means that the mobility in nanoribbons with a bandgap similar to that of silicon (1.1 eV) is expected to be lower than in bulk silicon and no higher than the mobility in the silicon channel of a conventional MOS device[58]. The mobilities measured in experiments—less than 200 cm^2 V^{-1} s^{-1} for nanoribbons 1–10 nm wide[26,62] and 1,500 cm^2 V^{-1} s^{-1} for a nanoribbon 14 nm wide[45] (which is the highest mobility so far measured for a nanoribbon)—support the theoretical results (Fig. 5b). Therefore, although the high mobilities offered by graphene can increase the speed of devices, they come at the expense of making it difficult to switch devices off, thus removing one of the main advantages of the CMOS configuration—its low static power consumption.

High-field transport. In the days when FETs had gates several micrometres long, the mobility was the appropriate measure of the speed of carrier transport. Strictly speaking, however, the mobility describes carrier transport in low electric fields; the short gate lengths in modern FETs result in high fields in a sizeable portion of the channel, reducing the relevance of mobility to device performance. To illustrate this, let us consider a FET with a gate 100 nm long and a drain–source voltage of 1 V. If we assume a voltage drop of 0.3 V across the series resistances, the average field in the channel is 70 kV cm^{-1}. At such high fields, the steady-state carrier velocity saturates, and this saturation velocity becomes another important measure of carrier transport. Figure 5c shows plots of the electron velocity versus the electric field for conventional semiconductors, and simulated plots for large-area graphene[63,64] and a carbon nanotube[57]. For graphene and the nanotube, maximum carrier velocities of around 4 × 10^7 cm s^{-1} are predicted, in comparison with 2 × 10^7 cm s^{-1} for GaAs and 10^7 cm s^{-1} for silicon. Moreover, at high fields the velocity in graphene and the nanotube does not drop as drastically as in the III–V semiconductors. Unfortunately, there is at present no experimental data available on high-field transport in graphene nanoribbons and in large-area graphene. However, other measurements[65] suggest high-field carrier velocities of several 10^7 cm s^{-1} in graphene. Thus, regarding high-field transport, graphene and nanotubes seem to have a slight advantage over conventional semiconductors.

Finally, it is worth noting that reported mobilities for graphene devices need to be interpreted carefully because there are several

Table 2 | Does graphene have a bandgap?

Graphene type	Size	Bandgap	Remarks	Ref.
SL graphene on SiO$_2$	LA	No	Experiment and theory	1, 5
SL graphene on SiO$_2$	GNR	Yes	Experiment and theory; gap due to lateral confinement*	24–29
BL graphene on SiO$_2$	LA	Yes	Experiment and theory; gap due to symmetry breaking by perpendicular interlayer field	30–33
Epitaxial SL	LA	Unknown	Controversial discussion	34
		Yes	Experiment and theory, gap due to symmetry breaking	35, 36
		No	Experiment and theory	37, 38
Epitaxial BL	LA	Yes	Experiment and theory	32, 38, 39
Epitaxial SL, BL	GNR	Yes	Theory	39
Strained SL[†]	LA	Yes	Theory; gap due to level crossing	40
		No	Theory	41

SL: single-layer; BL: bilayer; LA: large-area; GNR: graphene nanoribbon. *The origin of the bandgap in nanoribbons is still under debate: in addition to pure lateral confinement[28], it has been suggested that the Coulomb blockade[42,43] or Anderson localization[29] might be responsible for the formation of the gap. [†]Theorists disagree about the existence of a bandgap for strained SL graphene.

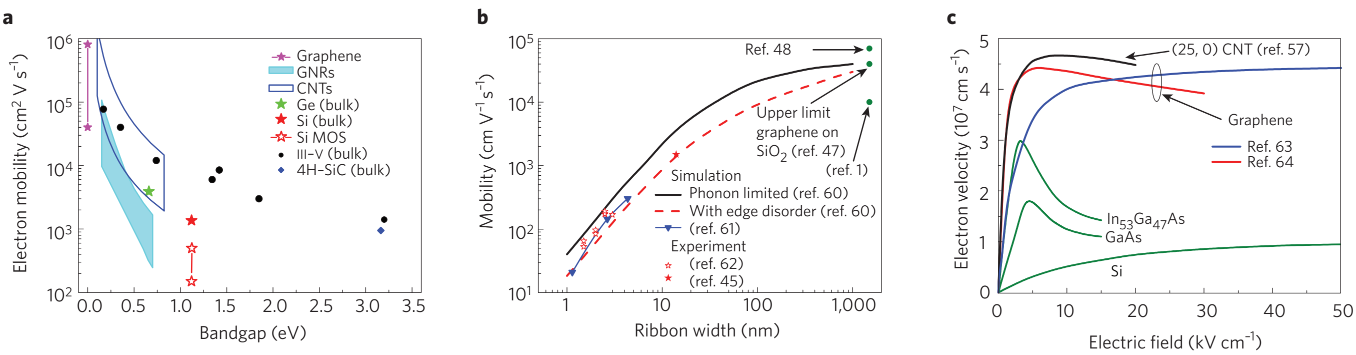


Figure 5 | Carrier transport in graphene. a, Electron mobility versus bandgap in low electric fields for different materials, as indicated (from left to right, III–V compounds are InSb, InAs, $In_{0.53}Ga_{0.47}As$, InP, GaAs, $In_{0.49}Ga_{0.51}P$, and GaN). The mobility data relates to undoped material except for the Si MOS data. Also shown are mobility data for carbon nanotubes (CNTs; simulation[56,57]), graphene nanoribbons (simulation[58,59]) and graphene (experiment and simulation[47–50]). **b**, Carrier mobility versus nanoribbon width at low electric fields from simulations[60,61] and experiments (open[62] and full[45] stars). Data for large-area graphene are also shown[1,47,48]. **c**, Electron drift velocity versus electric field for common semiconductors (Si, GaAs, $In_{0.53}Ga_{0.47}As$), a carbon nanotube (simulation[57]) and large-area graphene (simulation[63,64]).

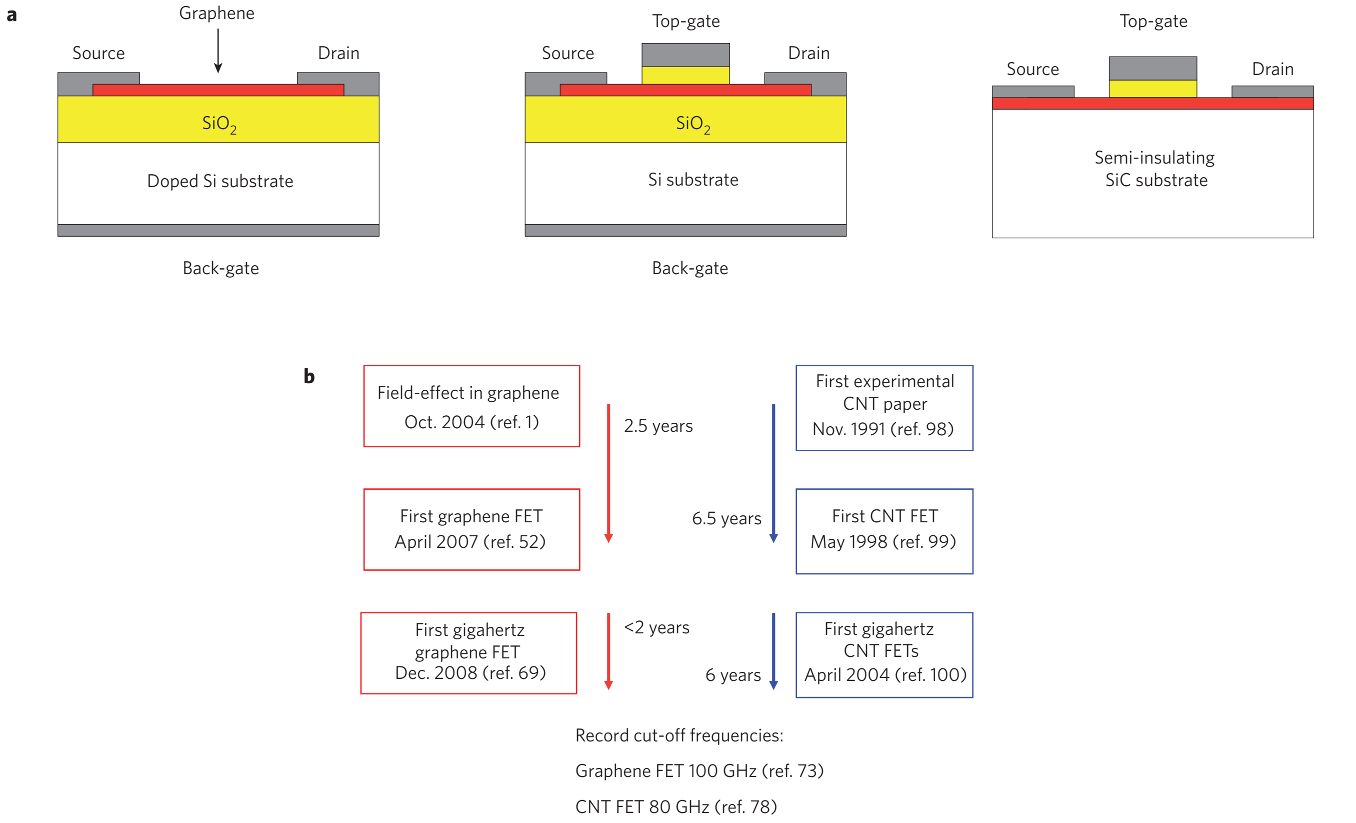

a
Graphene
Source
Drain
SiO₂
Doped Si substrate
Back-gate
Top-gate
Source
Drain
SiO₂
Si substrate
Back-gate
Top-gate
Source
Drain
Semi-insulating SiC substrate
b
Field-effect in graphene Oct. 2004 (ref. 1)
2.5 years
First graphene FET April 2007 (ref. 52)
<2 years
First gigahertz graphene FET Dec. 2008 (ref. 69)
First experimental CNT paper Nov. 1991 (ref. 98)
6.5 years
First CNT FET May 1998 (ref. 99)
6 years
First gigahertz CNT FETs April 2004 (ref. 100)
Record cut-off frequencies:
Graphene FET 100 GHz (ref. 73)
CNT FET 80 GHz (ref. 78)

Figure 6 | Structure and evolution of graphene MOSFETs. a, Schematics of different graphene MOSFET types: back-gated MOSFET (left); top-gated MOSFET with a channel of exfoliated graphene or of graphene grown on metal and transferred to a SiO_2-covered Si wafer (middle); top-gated MOSFET with an epitaxial-graphene channel (right). The channel shown in red can consist of either large-area graphene or graphene nanoribbons. **b**, Progress in graphene MOSFET development[1,52,69,73] compared with the evolution of nanotube FETs[78,98–100].

definitions for the MOSFET channel mobility and they are difficult to compare[66]. Furthermore, the techniques used to measure mobility are only vaguely described in some papers. Most frequently, the field-effect mobility, μ_{FE}, is measured (Table 1). However, the effect of the source and drain series resistances must be eliminated from the measured characteristics to determine this quantity, and it is not always clear that this has been done.

An additional complication lies in the interpretation of data from top-gated graphene MOSFETs, which involves arriving at a value for the gate capacitance, C_G. Frequently C_G is approximated by the oxide capacitance per unit area, as $C_{ox} = \varepsilon_{ox}/t_{ox}$, where ε_{ox} is the dielectric constant of the top-gate dielectric and t_{ox} is the thickness of this dielectric. However, when t_{ox} is small, the quantum capacitance, C_q, must be taken into account[67,68] because it is connected in series with C_{ox}, making the overall gate capacitance $C_G = C_{ox}C_q/(C_{ox} + C_q)$. The overall gate capacitance can be significantly smaller than C_{ox}, particularly close to the Dirac point (the point of minimum drain current), so neglecting the effect of C_q will lead to an underestimate of the field-effect mobility.

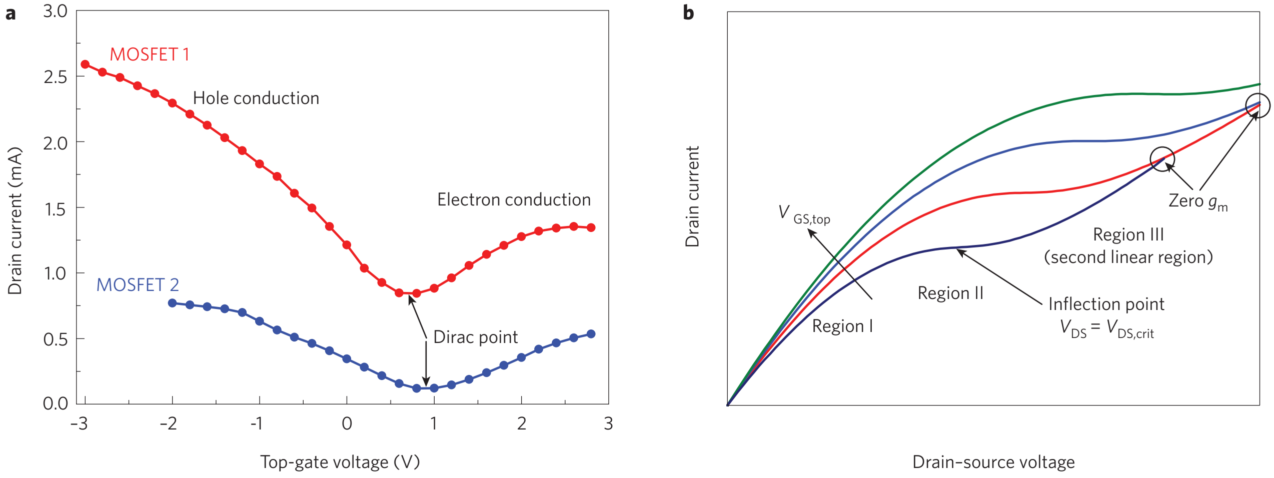


Figure 7 | Direct-current behaviour of graphene MOSFETs with a large-area-graphene channel. a, Typical transfer characteristics for two MOSFETs with large-area-graphene channels[23,71]. The on-off ratios are about 3 (MOSFET 1) and 7 (MOSFET 2), far below what is needed for applications in logic circuits. Unlike conventional Si MOSFETs, current flows for both positive and negative top-gate voltages. **b**, Qualitative shape of the output characteristics (drain current, I_D, versus drain–source voltage, V_{DS}) of a MOSFET with an n-type large-area-graphene channel, for different values of the top-gate voltage, $V_{GS,top}$. Saturation behaviour can be seen. At sufficiently large V_{DS} values, the output characteristics for different $V_{GS,top}$ values may cross[75], leading to a zero or even negative transconductance, which means that the gate has effectively lost control of the current.

State of the art of graphene transistors

A graphene MOS device was among the breakthrough results reported by the Manchester group in 2004 (ref. 1). A 300-nm SiO_2 layer underneath the graphene served as a back-gate dielectric and a doped silicon substrate acted as the back-gate (Fig. 6a). Such back-gate devices have been very useful for proof-of-concept purposes, but they suffer from unacceptably large parasitic capacitances and cannot be integrated with other components. Therefore, practical graphene transistors need a top-gate. The first graphene MOSFET with a top-gate was reported in 2007 (ref. 52), representing an important milestone, and progress has been very rapid since then (Fig. 6b). Although research into graphene is still in its infancy, graphene MOSFETs can compete with devices that have benefited from decades of research and investment.

Top-gated graphene MOSFETs have been made with exfoliated graphene[52–55,69,70], graphene grown on metals such as nickel and copper[71,72], and epitaxial graphene[23,73,74]; SiO_2, Al_2O_3, and HfO_2 have been used for the top-gate dielectric. The channels of these top-gated graphene transistors have been made using large-area graphene, which does not have a bandgap, so they have not been able to switch off.

Large-area-graphene transistors have a unique current–voltage transfer characteristic (Fig. 7a). The carrier density and the type of carrier (electrons or holes) in the channel are governed by the potential differences between the channel and the gates (top-gate and/or back-gate). Large positive gate voltages promote an electron accumulation in the channel (n-type channel), and large negative gate voltages lead to a p-type channel. This behaviour gives rise to the two branches of the transfer characteristics separated by the Dirac point (Fig. 7a). The position of the Dirac point depends on several factors: the difference between the work functions of the gate and the graphene, the type and density of the charges at the interfaces at the top and bottom of the channel (Fig. 6), and any doping of the graphene. The on–off ratios reported for MOSFET devices with large-area-graphene channels are in the range 2–20.

The output characteristics of many graphene MOSFETs either show a linear shape without any saturation[53] or only weak saturation[73,74], each of which is a disadvantage with respect to device speed. However, some graphene MOSFETs have an unusual form of saturation-like behaviour that includes a second linear region[70,71,75] (Fig. 7b). Our present understanding of the origin of this behaviour is as follows. For small values of V_{DS}, the transistor operates in the linear region and the entire channel is n-type (region I). As V_{DS} is increased, the drain current starts to saturate until the inflection point at $V_{DS} = V_{DS,crit}$ is reached (region II). At this point, the potential conditions at the drain end of the channel correspond to the Dirac point. Once V_{DS} exceeds $V_{DS,crit}$, the conduction type at the drain end of the channel switches from n-type to p-type[70,76] and the transistor enters a second linear region (region III). At sufficiently large values of V_{DS}, the output characteristics for different gate voltages may cross[75], leading to a zero or even negative transconductance—a highly undesirable situation. This peculiar behaviour is a consequence of these devices having gapless channels and does not occur in FETs with semiconducting channels.

Recently, graphene MOSFETs with gigahertz capabilities have been reported. These transistors possess large-area channels of exfoliated[53,55,69,77] and epitaxial[73,74] graphene. The fastest graphene transistor currently is a MOSFET with a 240-nm gate that has a cut-off frequency of f_T = 100 GHz (ref. 73), which is higher than those of the best silicon MOSFETs with similar gate lengths (as is the cut-off frequency of 53 GHz reported for a device with a 550-nm gate, also in ref. 73). A weak point of all radiofrequency graphene MOSFETs reported so far is the unsatisfying saturation behaviour (only weak saturation or the second linear regime), which has an adverse impact on the cut-off frequency, the intrinsic gain and other figures of merit for radiofrequency devices. However, outperforming silicon MOSFETs while operating with only weak current saturation[73] is certainly impressive.

Figure 8 shows the cut-off frequency for a variety of devices including graphene MOSFETs, nanotube FETs, and various radiofrequency FETs. For conventional radiofrequency FETs with gate lengths greater than 0.2 μm, the f_T data for each transistor type has an L^{-1} dependence, where L is the gate length. Furthermore, f_T increases with mobility[9]. Silicon MOSFETs show channel mobilities of a few 100 cm^2 V^{-1} s^{-1} compared with about 6,000 cm^2 V^{-1} s^{-1} for GaAs pHEMTs and more than 10,000 cm^2 V^{-1} s^{-1} for InP HEMTs and GaAs mHEMTs. At shorter gate lengths, however, the mobility becomes less important for transistor speed and the deleterious influence of parasitic resistances and short-channel effects increases. Both nanotube and graphene FETs are still slower than

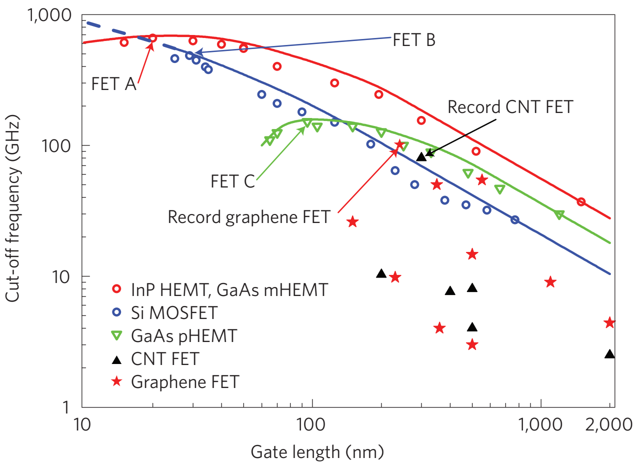


Figure 8 | Comparing cut-off frequencies for different FETs. Cut-off frequency versus gate length for graphene MOSFETs, nanotube FETs and three types of radiofrequency FET; the symbols are experimental data points and the lines are a guide to the eye for type A (InP HEMT and GaAs mHEMT), B (Si MOSFET) and C (GaAs pHEMT) devices (as indicated). The FET A with the highest cut-off frequency (660 GHz) is a GaAs metamorphic HEMT (mHEMT) with a 20-nm gate (M. Schlechtweg, personal communication). The FET B with the highest cut-off frequency (485 GHz) is a Si MOSFET with a 29-nm gate[101]. The FET C with the highest cut-off frequency (152 GHz) is a GaAs pseudomorphic HEMT (pHEMT) with a 100-nm gate[102]. The fastest nanotube device (CNT FET) has f_T = 80 GHz and L = 300 nm (ref. 78), and the fastest reported graphene MOSFET has f_T = 100 GHz and L = 240 nm (ref. 73).

the best conventional radiofrequency FETs, but they have recently overtaken the best silicon MOSFETs with gate lengths above 200 nm and are approaching the performance of GaAs pHEMTs. (See ref. 78 for details of the nanotube with the highest f_T reported so far, and ref. 79 for more information on the radiofrequency potential of nanotube FETs.)

Although the low on–off ratios demonstrated so far make use in logic devices unrealistic, transistors with large-area graphene channels are promising candidates for radiofrequency applications because radiofrequency FETs are not required to switch off and can benefit from the high mobilities offered by large-area graphene. However, the absence of drain-current saturation will limit the radiofrequency performance of graphene transistors.

One method of introducing a bandgap into graphene for logic applications is to create graphene nanoribbons. Nanoribbon MOSFETs with back-gate control and widths down to less than 5 nm have been operated as p-channel devices and had on–off ratios of up to 10^6 (refs 26,62). Such high ratios have been obtained despite simulations showing that edge disorder leads to an undesirable decrease in the on-currents and a simultaneous increase in the off-current of nanoribbon MOSFETs[80,81].This, and other evidence of a sizeable bandgap opening in narrow nanoribbons, provides proof of the suitability of nanoribbon FETs for logic applications. However, these devices had relatively thick back-gate oxides, so voltage swings of several volts were needed for switching, which is significantly more than the swings of 1 V and less needed to switch Si CMOS devices[2]. Furthermore, CMOS logic requires both n-channel and p-channel FETs with well-controlled threshold voltages, and graphene FETs with all these properties have not yet been reported.

Recently, the first graphene nanoribbon MOSFETs with top-gate control have been reported[82]. These transistors feature a thin high-dielectric-constant (high-k) top-gate dielectric (1–2 nm of HfO_2), a room-temperature on–off ratio of 70 and an outstanding transconductance of 3.2 mS μm^{-1} (which is higher than the transconductances reported for state-of-the-art silicon MOSFETs and III–V HEMTs).

Graphene bilayer MOSFETs have been investigated experimentally[83] and by device simulation[84]. Although the on–off ratios reported so far (100 at room temperature and 2,000 at low temperature[83]) are too small for logic applications, they mark a significant improvement (of about a factor of 10) over MOSFETs in which the channel is made of large-area gapless graphene.

The contact resistance between the metallic source and drain contacts and the graphene channel should be briefly mentioned. So far, the lowest reported metal–graphene contact resistances are in the range 500–1,000 Ω cm (refs 85,86), which is about ten times the contact resistance of silicon MOSFETs and III–V HEMTs[8,13]. Remarkably, in spite of the importance of the contacts (particularly for short-channel devices), only a few studies dealing with metal–graphene contacts have been published[85–87] and more work is needed to understand the contact properties.

I now return to the two-dimensional nature of graphene. According to scaling theory, as noted previously, a thin channel region allows short-channel effects to be suppressed and thus makes it feasible to scale MOSFETs to very short gate lengths. The two-dimensional nature of graphene means it offers us the thinnest possible channel, so graphene MOSFETs should be more scalable than their competitors. It should be noted, however, that scaling theory is valid only for transistors with a semiconducting channel and does not apply to graphene MOSFETs with gapless channels. Thus, the scaling theory does describe nanoribbon MOSFETs, which have a bandgap but which have significantly lower mobilities than large-area graphene, as discussed. Given that the high published values of mobility relate to gapless large-area graphene, the most attractive characteristic of graphene for use in MOSFETs, in particular those required to switch off, may be its ability to scale to shorter channels and higher speeds, rather than its mobility.

Further options for graphene devices

It has become clear that graphene devices based on the conventional MOSFET principle suffer from some fundamental problems. This has motivated researchers to explore new graphene device concepts, such as tunnel FETs and bilayer pseudospin FETs. In a tunnel FET, the band-to-band tunnelling across the source–channel junction can be controlled using the gate–source voltage. The big advantage of tunnel FETs is that their subthreshold swings are not limited to 60 mV per decade, as in conventional MOSFETs[7,10], which should lead to steeper subthreshold characteristics and better switch-off. The tunnel-FET approach has already been explored in silicon and carbon-nanotube MOSFETs[88,89]. Tunnel FETs based on nanoribbons and bilayer graphene have been investigated in simulations[84,90,91] but have not been demonstrated experimentally. In particular, the bilayer graphene tunnel FET is now considered to be a promising device for a number of reasons: narrow nanoribbons are not needed, so edge disorder will not be a problem and patterning will be relatively easy; the small bandgap opened by a vertical field applied across the two layers is sufficient to suppress band-to-band tunnelling in the off-state and thus enables effective switch-off; and the possibility of subthreshold swings below 60 mV per decade should make high on–off ratios possible[84].

The bilayer pseudospin FET consists of a vertical stack of two graphene layers separated by a thin dielectric[92]. Under certain bias conditions the tunnelling resistance between the two graphene layers becomes so small that the layers are effectively shorted, causing the FET to pass a high current, whereas under other conditions the tunnelling resistance is very large, shutting the current off. The bilayer pseudospin FET might therefore be able to deliver fast and ultralow-power logic operation.

Although graphene tunnel FETs and bilayer pseudospin FETs are both still at an embryonic stage, they have already gained considerable attention in the electron-device community and have been included in the chapter on emerging research devices in the latest edition of the ITRS[2]. It might also be possible to make interconnects from graphene, which would open the possibility of all-graphene integrated circuits in which both the active devices and the wiring were made of graphene[22]. It has been shown that graphene interconnects compete well with copper interconnects[93,94]; indeed, graphene can support current densities greater than 10^8 A cm^{-2} (which is 100 times higher than those supported by copper and is comparable with those supported by nanotubes)[95] and has a thermal conductivity of around 30–50 W cm^{-1} K^{-1} (in comparison with 4 W cm^{-1} K^{-1} for copper)[96].

Outlook

Since 2007, we have witnessed huge progress in the development of graphene transistors. Most impressive were the demonstrations of a graphene MOSFET with a cut-off frequency of 100 GHz (ref. 73), the excellent switching behaviour of nanoribbon MOSFETs[26,62], and channel mobilities exceeding 20,000 cm^2 V^{-1} s^{-1} in top-gated graphene MOSFETs[54]. However, this progress has been accompanied by the appearance of a number of problems. MOSFETs with large-area-graphene channels cannot be switched off, making them unsuitable for logic applications, and their peculiar saturation behaviour limits their radiofrequency performance. Nanoribbon graphene, which does have a bandgap and results in transistors that can be switched off, has serious fabrication issues because of the small widths required and the presence of edge disorder.

The primary challenges facing the community at present, therefore, are to create in a controlled and practical fashion a bandgap in graphene, which would allow logic transistors to switch off and radiofrequency transistors to avoid the second linear regime (Fig. 7b), and to develop other means of improving transistor saturation characteristics by, for example, developing contacts that block one kind of carrier without degrading the transistor's speed. The community may also benefit from recognizing that the motivation to use graphene in transistors in the first place stems less from ultrahigh mobilities than from graphene's ability to scale to short gate lengths and high speeds by virtue of its thinness.

This discussion of the problems of graphene MOSFETs should not lead to the conclusion that graphene is not a promising material for transistors. Rather, I have chosen a more critical view to avoid a situation that has been seen in the past, in which a new device or material concept has been prematurely declared capable of replacing the status quo. Also, I agree with David Ferry, a veteran of semiconductor device research, when he says that[97] "many such saviours have come and gone, yet the reliable silicon CMOS continues to be scaled and to reach even higher performance levels".

I conclude by noting that the first top-gated graphene transistors were reported only three years ago. Given this short history, and given that all other possible successors to conventional mainstream transistors also face serious problems, we cannot help but be impressed with the rapid development of graphene. Concepts that have been investigated for many years, such as spin transistors or molecular devices, seem to be farther from real application than does graphene, and it is not clear if they will ever reach the production stage. At the moment, it is impossible to say which, if any, of the alternative device concepts being considered will replace conventional transistors. However, the latest ITRS roadmap strongly recommends intensified research into graphene and even contains a research and development schedule for carbon-based nanoelectronics[2]. The race is still open and the prospects for graphene devices are at least as promising as those for alternative concepts.

References

1. Novoselov, K. S. *et al.* Electric field effect in atomically thin carbon films. *Science* **306,** 666–669 (2004).
2. *The International Technology Roadmap for Semiconductors* http://www.itrs.net/Links/2009ITRS/Home2009.htm (Semiconductor Industry Association, 2009).
3. Geim, A. K. & Novoselov, K. S. The rise of graphene. *Nature Mater.* **6,** 183–191 (2007).
4. Geim, A. K. Graphene: status & prospects. *Science* **324,** 1530–1534 (2009).
5. Castro Neto, A. H. *et al.* The electronic properties of graphene. *Rev. Mod. Phys.* **81,** 109–162 (2009).
6. Moore, G. E. in *Tech. Dig. ISSCC* 20–23 (IEEE, 2003).
7. Schwierz, F., Wong, H. & Liou, J. J. *Nanometer CMOS* (Pan Stanford, 2010).
8. Schwierz, F. & Liou, J. J. *Modern Microwave Transistors – Theory, Design, and Performance* (Wiley, 2003).
9. Schwierz, F. & Liou, J. J. RF transistors: recent developments and roadmap toward terahertz applications. *Solid-State Electron.* **51,** 1079–1091 (2007).
10. Taur, Y. & Ning, T. H *Fundamentals of Modern VLSI Devices* (Cambridge Univ. Press, 1998).
11. Frank, D. J., Taur, Y. & Wong, H-S. P. Generalized scale length for two-dimensional effects in MOSFETs. *IEEE Electron Dev. Lett.* **19,** 385–387 (1998).
12. Aberg, I. & Hoyt, J. L. Hole transport in ultra-thin-body MOSFETs in strained-Si directly on insulator with strained-Si thickness less than 5 nm. *IEEE Electron Dev. Lett.* **26,** 661–663 (2005).
13. Thompson, S. E. *et al.* In search of "forever", continued transistor scaling one new material at a time. *IEEE Trans. Semicond. Manuf.* **18,** 26–36 (2005).
14. Uyemura, J. P. *CMOS Logic Circuit Design* (Kluwer Academic, 1999).
15. Hughes, B. & Tasker, P. J. Bias dependence of the MODFET intrinsic model elements values at microwave frequencies. *IEEE Trans. Electron. Dev.* **36,** 2267–2273 (1989).
16. Nguyen, L. D. *et al.* in *Tech. Dig. IEDM* 176–179 (IEEE, 1988).
17. Boehm, H. P., Clauss, A., Hofmann, U. & Fischer, G. O. Dünnste Kohlenstoff-Folien. *Z. Naturforsch. B* **17,** 150–153 (1962).
18. May, J. W. Platinum surface LEED rings. *Surf. Sci.* **17,** 267–270 (1969).
19. van Bommel, A. J., Crombeen, J. E. & van Tooren, A. LEED and Auger electron observations of the SiC (0001) surface. *Surf. Sci.* **48,** 463–472 (1975).
20. Kim, K-S. *et al.* Large-scale pattern growth of graphene films for stretchable transparent electrodes. *Nature* **457,** 706–710 (2009).
21. Reina, A. *et al.* Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition. *Nano Lett.* **9,** 30–35 (2009).
22. Berger, C. *et al.* Electronic confinement and coherence in patterned epitaxial graphene. *Science* **312,** 1191–1196 (2006).
23. Kedzierski, J. *et al.* Epitaxial graphene transistors on SiC substrates. *IEEE Trans. Electron. Dev.* **55,** 2078–2085 (2008).
24. Han, M. *et al.* Energy band-gap engineering of graphene nanoribbons. *Phys. Rev. Lett.* **98,** 206805 (2007).
25. Kim, P. *et al.* in *Tech. Dig. IEDM* 241–244 (IEEE, 2009).
26. Li, X., Wang, X., Zhang, L., Lee, S. & Dai, H. Chemically derived, ultrasmooth graphene nanoribbon semiconductors. *Science* **319,** 1229–1232 (2008).
27. Chen, Z., Lin, Y-M., Rooks, M. J. & Avouris, Ph. Graphene nano-ribbon electronics. *Physica E* **40,** 228–232 (2007).
28. Yang, L. *et al.* Quasiparticle energies and band gaps in graphene nanoribbons. *Phys. Rev. Lett.* **99,** 186801 (2007).
29. Evaldsson, M., Zozoulenko, I. V., Xu, H. & Heinzel, T. Edge-disorder-induced Anderson localization and conduction gap in graphene nanoribbons. *Phys. Rev. B* **78,** 161407 (2008).
30. Castro, E. V. *et al.* Biased bilayer graphene: semiconductor with a gap tunable by the electric field effect. *Phys. Rev. Lett.* **99,** 216802 (2007).
31. Gava, P., Lazzeri, M., Saitta, A. M. & Mauri, F. *Ab initio* study of gap opening and screening effects in gated bilayer graphene. *Phys. Rev. B* **79,** 165431 (2009).
32. Ohta, T., Bostwick, A., Seyller, Th., Horn, K. & Rotenberg, E. Controlling the electronic structure of bilayer graphene. *Science* **313,** 951–954 (2006).
33. Zhang, Y. *et al.* Direct observation of a widely tunable bandgap in bilayer graphene. *Nature* **459,** 820–823 (2009).
34. Rotenberg, E. *et al.* and Zhou, S. Y. *et al.* Origin of the energy bandgap in epitaxial graphene. *Nature Mater.* **7,** 258–260 (2008).
35. Zhou, S. Y. *et al.* Substrate-induced bandgap opening in epitaxial graphene. *Nature Mater.* **6,** 770–775 (2007).
36. Kim, S., Ihm, J., Choi, H. J. & Son, Y-W. Origin of anomalous electronic structures of epitaxial graphene on silicon carbide. *Phys. Rev. Lett.* **100,** 176802 (2008).
37. Bostwick, A., Ohta, T., Seyller, Th., Horn, K. & Rotenberg, E. Quasiparticle dynamics in graphene. *Nature Phys.* **3,** 36–40 (2007).
38. Peng, X. & Ahuja, R. Symmetry breaking induced bandgap in epitaxial graphene layers on Si. *Nano Lett.* **8,** 4464–4468 (2008).
39. Sano, E. & Otsuji, T. Theoretical evaluation of channel structure in graphene field-effect transistors. *Jpn. J. Appl. Phys.* **48,** 041202 (2009).
40. Pereira, V. M., Castro Neto, A. H. & Peres, N. M. R. Tight-binding approach to uniaxial strain in graphene. *Phys. Rev. B* **80,** 045401 (2009).

41. Ni, Z. H. *et al.* Uniaxial strain on graphene: Raman spectroscopy study and band-gap opening. *ACS Nano* **2,** 2301–2305 (2008); erratum **3,** 483 (2009).
42. Sols, F., Guinea, F. & Castro Neto, A. H. Coulomb blockade in graphene nanoribbons. *Phys. Rev. Lett.* **99,** 166803 (2007).
43. Han, M. Y., Brant, J. C. & Kim, P. Electron transport in disordered graphene nanoribbons. *Phys. Rev. Lett.* **104,** 056801 (2010).
44. Cervantes-Sodi, F., Csanyi, G., Picanec, S. & Ferrari, A. C. Edge-functionalized and substitutionally doped graphene nanoribbons: electronic and spin properties. *Phys. Rev. B* **77,** 165427 (2008).
45. Jiao, J., Wang, X., Diankov, G., Wang, H. & Dai, H. Facile synthesis of high-quality graphene nanoribbons. *Nature Nanotech.* **5,** 321–325 (2010).
46. Raza, H. & Kan, E. C. Armchair graphene nanoribbons: electronic structure and electric-field modulation. *Phys. Rev. B* **77,** 245434 (2008).
47. Chen, J-H., Jang, C., Xiao, S., Ishigami, M. & Fuhrer, M. S. Intrinsic and extrinsic performance limits of graphene devices on SiO_2. *Nature Nanotech.* **3,** 206–209 (2008).
48. Chen, F., Xia, J., Ferry, D. K. & Tao, N. Dielectric screening enhanced performance in graphene FET. *Nano Lett.* **9,** 2571–2574 (2009).
49. Morozov, V. S. *et al.* Giant intrinsic carrier mobilities in graphene and its bilayer. *Phys. Rev. Lett.* **100,** 016602 (2008).
50. Geim, A. Graphene update. *Bull. Am. Phys. Soc.* **55,** abstr. J21.0004, http://meetings.aps.org/link/BAPS.2010.MAR.J21.4 (2010).
51. Emtsev, K. V. *et al.* Towards wafer-size graphene layers by atmospheric pressure graphitization of silicon carbide. *Nature Mater.* **8,** 203–207 (2009).
52. Lemme, M. C., Echtermeyer, T. J., Baus, M. & Kurz, H. A graphene field-effect device. *IEEE Electron Dev. Lett.* **28,** 282–284 (2007).
53. Lin, Y-M. *et al.* Operation of graphene transistors at gigahertz frequencies. *Nano Lett.* **9,** 422–426 (2009).
54. Liao, L. *et al.* High-κ oxide nanoribbons as gate dielectrics for high mobility top-gated graphene transistors. *Proc. Natl Acad. Sci. USA* **107,** 6711–6715 (2010).
55. Farmer, D. B. *et al.* Utilization of a buffered dielectric to achieve high field-effect carrier mobility in graphene transistors. *Nano Lett.* **9,** 4474–4478 (2009).
56. Zhou, X., Park, J-Y., Huang, S., Liu, J. & McEuen, P. L. Band structure, phonon scattering, and performance limit of single-walled carbon nanotube transistors. *Phys. Rev. Lett.* **95,** 146805 (2005).
57. Perebeinos, V., Tersoff, J. & Avouris, Ph. Electron-phonon interaction and transport in semiconducting carbon nanotubes. *Phys. Rev. Lett.* **94,** 0786802 (2005).
58. Obradovic, B. *et al.* Analysis of graphene nanoribbons as a channel material for field-effect transistors. *Appl. Phys. Lett.* **88,** 142102 (2006).
59. Fang, T., Konar, A., Xing, H. & Jena, D. Mobility in semiconducting nanoribbons: phonon, impurity, and edge roughness scattering. *Phys. Rev. B* **78,** 205403 (2008).
60. Bresciani, M., Palestri, P., Esseni, D. & Selmi, L. in *Proc. ESSDERC '09* 480–483 (IEEE, 2009).
61. Betti, A., Fiori, G., Iannaccone, G. & Mao, Y. in *Tech. Dig. IEDM 2009* 897–900 (IEEE, 2009).
62. Wang, X. *et al.* Room-temperature all-semiconducting sub-10-nm graphene nanoribbon field-effect transistors. *Phys. Rev. Lett.* **100,** 206803 (2008).
63. Akturk, A. & Goldsman, N. Electron transport and full-band electron-phonon interactions in graphene. *J. Appl. Phys.* **103,** 053702 (2008).
64. Shishir, R. S. & Ferry, D. K. Velocity saturation in intrinsic graphene. *J. Phys. Condens. Matter* **21,** 344201 (2009).
65. Barreiro, A., Lazzeri, M., Moser, J., Mauri, F. & Bachtold, A. Transport properties of graphene in the high-current limit. *Phys. Rev. Lett.* **103,** 076601 (2009).
66. Schroder, D. K. *Semiconductor Material and Device Characterization* (Wiley, 1990).
67. Fang, T., Konar, A., Xing, H. & Jena, D. Carrier statistics and quantum capacitance of graphene sheets and nanoribbons. *Appl. Phys. Lett.* **91,** 092109 (2007).
68. Chen, Z. & Appenzeller, J. in *Tech. Dig. IEDM 2008*, paper 21.1 (IEEE, 2008).
69. Meric, I., Baklitskaya, N., Kim, P. & Shepard, K. L. in *Tech. Dig. IEDM 2008*, paper 21.2 (IEEE, 2008).
70. Meric, I. *et al.* Current saturation in zero-bandgap, top-gated graphene field-effect transistors. *Nature Nanotech.* **3,** 654–659 (2008).
71. Kedzierski, J. *et al.* Graphene-on-insulator transistors made using C on Ni chemical-vapor deposition. *IEEE Electron Dev. Lett.* **30,** 745–747 (2009).
72. Li, X. *et al.* Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science* **324,** 1312–1314 (2009).
73. Lin, Y-M. *et al.* 100-GHz transistors from wafer-scale epitaxial graphene. *Science* **327,** 662 (2010).
74. Moon, J. S. *et al.* Epitaxial-graphene RF field-effect transistors on Si-face 6H-SiC substrates. *IEEE Electron Dev. Lett.* **30,** 650–652 (2009).
75. Tahy, K. *et al.* in *Proc. Dev. Res. Conf. 2009* 207–208 (IEEE, 2009).
76. Thiele, S., Schaefer, J. A. & Schwierz, F. Modeling of graphene metal–oxide–semiconductor field-effect transistors with gapless large-area graphene channels. *J. Appl. Phys.* **107,** 094505 (2010).
77. Lin, Y-M. *et al.* Dual-gate graphene FETs with f_T of 50 GHz. *IEEE Electron Dev. Lett.* **31,** 68–70 (2010).
78. Nougaret, N. *et al.* 80 GHz field-effect transistors produced using high purity semiconducting single-walled carbon nanotubes. *Appl. Phys. Lett.* **94,** 243505 (2009).
79. Rutherglen, C., Jain, D. & Burke, P. Nanotube electronics for radiofrequency applications. *Nature Nanotech.* **4,** 811–819 (2009).
80. Yoon, Y. & Guo, J. Effects of edge roughness in graphene nanoribbon transistors. *Appl. Phys. Lett.* **91,** 073103 (2007).
81. Basu, D., Gilbert, M. J., Register, L. F., Banerjee, S. K. & MacDonald, A. H. Effect of edge roughness on electronic transport in graphene nanoribbon channel metal–oxide–semiconductor field-effect transistors. *Appl. Phys. Lett.* **92,** 042114 (2008).
82. Liao, L. *et al.* Top-gated graphene nanoribbon transistors with ultrathin high-*k* dielectrics. *Nano Lett.* **10,** 1917–1921 (2010).
83. Xia, F., Farmer, D. B., Lin, Y-M. & Avouris, Ph. Graphene field-effect transistors with high on/off current ratio and large transport band gap at room temperature. *Nano Lett.* **10,** 715–718 (2010).
84. Iannaccone, G. *et al.* in *Tech. Dig. IEDM 2009* 245–248 (IEEE, 2009).
85. Nagashio, K., Nishimura, T., Kita, K. & Toriumi, A. in *Tech. Dig. IEDM 2009* 565–568 (IEEE, 2009).
86. Russo, S., Cracuin, M. F., Yamamoto, Y., Morpurgo, A. F. & Tarucha, S. Contact resistance in graphene-based devices. *Physica E* **42,** 677–679 (2010).
87. Huard, B., Stander, N., Sulpizio, J. A. & Goldhaber-Gordon, D. Evidence of the role of contacts on the observed electron-hole asymmetry in graphene. *Phys. Rev. B* **78,** 121402 (2008).
88. Boucart, K. & Ionescu, A. M. Double-gate tunnel FET with high-κ gate dielectric. *IEEE Trans. Electron. Dev.* **54,** 1725–1733 (2007).
89. Appenzeller, J., Lin, Y-M., Knoch, J. & Avouris, Ph. Band-to-band tunneling in carbon nanotube field-effect transistor. *Phys. Rev. Lett.* **93,** 196805 (2004).
90. Luisier, M. & Klimeck, G. in *Proc. Dev. Res. Conf. 2009* 201–202 (IEEE, 2009).
91. Fiori, G. & Iannaccone, G. Ultralow-voltage bilayer graphene tunnel FET. *IEEE Electron Dev. Lett.* **30,** 1096–1098 (2009).
92. Banerjee, S. K., Register, L. F., Tutuc, E., Reddy, D. & MacDonald, A. H. Bilayer pseudospin field-effect transistor (BiSFET): a proposed new logic device. *IEEE Electron Dev. Lett.* **30,** 158–160 (2009).
93. Murali, R., Brenner, K., Yang, Y., Beck, Th. & Meindl, J. D. Resistivity of graphene nanoribbon interconnects. *IEEE Electron Dev. Lett.* **30,** 611–613 (2009).
94. Awano, Y. in *Tech. Dig. IEDM 2009* 233–236 (IEEE, 2009).
95. Moser, J., Barreiro, A. & Bachtold, A. Current-induced cleaning of graphene. *Appl. Phys. Lett.* **91,** 163513 (2007).
96. Balandin, A. A. *et al.* Superior thermal conductivity of single-layer graphene. *Nano Lett.* **8,** 902–907 (2008).
97. Ferry, D. K., Gilbert, M. J. & Akis, R. Some considerations on nanowires in nanoelectronics. *IEEE Trans. Electron. Dev.* **55,** 2820–2826 (2008).
98. Iijima, S. Helical microtubules of graphitic carbon. *Nature* **354,** 56–58 (1991).
99. Tans, S. J., Verschueren, A. R. M. & Dekker, C. Room-temperature transistor based on a single carbon nanotube. *Nature* **393,** 49–52 (1998).
100. Li, S., Yu, Z., Yen, S-F., Tang, W. C. & Burke, P. J. Carbon nanotube transistor operation at 2.6 GHz. *Nano Lett.* **4,** 753–756 (2004).
101. Lee, S. *et al.* in *Tech. Dig. IEDM 2007* 255–258 (IEEE, 2007).
102. Nguyen, L. D., Tasker, P. J., Radulescu, D. C. & Eastman, L. F. Characterization of ultra-high-speed AlGaAs/InGaAs (on GaAs) MODFETs. *IEEE Trans. Electron. Dev.* **36,** 2243–2248 (1989).

Acknowledgements

This work was financially supported by the 2008–2009 Excellence Research Grant of Technische Universität Ilmenau. The author thanks A. Castro Neto, K. Novoselov and Th. Seyller for discussions. He also thanks St. Thiele for his comments and for graphene MOSFET simulations, and M. Schlechtweg for providing GaAs metamorphic-HEMT data before publication.

Additional information

The authors declare no competing financial interests.