

Semiconductors explained in 16 mins | Chris Miller

16 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=09G49L-MME](https://www.youtube.com/watch?v=09G49L-MME)

<https://www.youtube.com/watch?v=09g49L-MmeE>

SUMMARY

Chris Miller discusses the complexities and advancements in semiconductor manufacturing, emphasizing that while chips are ubiquitous, their production is highly intricate and costly. He contrasts the stagnation in nuclear weapon technology with the rapid evolution of chip technology, highlighting the critical role of companies like TSMC in the semiconductor industry.

- Semiconductor production is more complex than nuclear weapon manufacturing, despite the latter being controlled by a few governments.*
- Chips are made from silicon, but require extreme purity and sophisticated materials, with only a few companies capable of producing high-quality silicon wafers.*
- The manufacturing process is heavily automated, requiring advanced machinery that operates at the atomic level.*
- TSMC dominates the market, producing about 90% of advanced processor chips, making it one of the most critical companies globally.*
- Moore's Law continues to hold, predicting that the number of transistors on chips will double approximately every two years, driving innovation in computing power.*
- The transition to 3D transistor designs is a key trend, allowing for more transistors to be packed into smaller spaces.*
- The semiconductor supply chain is global, with materials sourced from various countries, highlighting geopolitical risks.*
- The manufacturing of cutting-edge chips requires significant investment, often exceeding \$20 billion per facility, leading to consolidation in the industry.*

When I started my research on semiconductors,

I thought that because chips were everywhere, chips were easy to make, and because nuclear bombs were only controlled by a handful of governments, they were hard to make. But what I realized is it's actually the exact opposite. If you take nuclear weapons, that technology has barely improved since the 1960s. But chips are everywhere because they're cheap and they're tiny, and making things very inexpensive and very small is extraordinarily difficult, which is why there's just a couple companies in the world that can do it. And it's done so not just for a couple years, it's done so now for over half a century.

And that's why when you compare progress in the computing industry to progress anywhere else, there's really no comparison. I'm Chris

Miller, a professor at The Fletcher School and author of, "Chip War: The Fight for the World's Most Critical Technology." - [Announcer] Semiconductors 101 - Well, a chip is a piece of silicon, often the size of your fingernail. And in it is carved thousands, or millions, in some cases billions of tiny devices called transistors, which flip circuits on or off, on and off. And when they're on, they produce a one. When they're off, they produce a zero. And all of the ones and zeros undergirding computing, undergirding data storage, all of your Instagram likes, all of your text messages, these are all just long strings of ones and zeros, which are created on the chip by these circuits flipping on and off. There are a couple different categories of chips. Some chips process data, other chips remember data, and a third category turns real world signals, like audio or pictures into ones and zeros so that they can then be processed or remembered. And so when we look at the world, we see pictures. But when a phone, for example, uses its camera to look at the world, it takes in lots of rays of light, and then has to learn how to convert those into ones and zeros that can be stored. And so there's very specific sensors for pictures, for sound, for radio waves that use semiconductors to convert these real world signals into strings in ones of zeros that can then be re-represented as pictures later on, for example, when you pull a photo up on your phone. All of this is done by different types of semiconductors. So, generally, chips have a foundation of silicon, but there are dozens of other materials that are layered on top to make the transistors at such tiny scale. So a typical advanced chip could have several dozen materials. The foundation is silicon, but there are many other chemicals involved in the process. Yeah, it's true that sand is from silicon and so are chips, but the similarities basically end there. The silicon that's used in manufacturing chips is among the most purified elements that we have. And the reason is that when you're manufacturing chips with tiny transistors, you need to place almost every atom perfectly to make those chips work.

Which means that if your silicon, or any of the other materials that you're using, has even a single atomic impurity, it can cause defects in the way your chip functions. And so the production of the silicon wafers that are used in the chip manufacturing process requires extraordinary levels of purity. There's really just four companies in the world today that are capable of producing silicon wafers at the right level of purity at the scale that's required for contemporary manufacturing. The good news is that there's silicon everywhere. It's one of the most widely-distributed elements in the Earth's crust. The hard part is really the refining and the purification of silicon to make sure there aren't any impurities that could disrupt the manufacturing process. So on top of your silicon, you could have boron, gallium, gallium arsenide, lots of different chemicals that are used, and every chip maker has its own proprietary process.

So we don't really know, inside of a typical chip, what materials are used, because chip makers usually keep it pretty secretive. That's their special sauce that lets them manufacture chips with the right level of capability. Now we're not gonna run out of silicon, nor will we run out of the other materials that are generally used in chipmaking. There are some concerns that certain materials are predominantly refined and processed

in a single country. So for some of the materials like gallium and germanium, China produces around 90% of those materials. So there's geopolitical issues that could interrupt supply, but it's not gonna be that we're running out of the capability to produce them. I visited a bunch of chipmaking facilities over the course of the research. The interesting thing though is that, when you go inside one of these massive facilities, called fabs, what you find is that there are huge machines and not much else. Because the manufacturing process has to be extraordinarily automated because humans are way too imprecise for manufacturing at nanometer scale. And so inside of a chipmaking facility, there are very few humans, and lots of big machines that, from the outside, are impressive in their size, but you can't see what's actually happening because it's happening at microscopic level. So there are a handful of companies that play a big role in the making of the machines that make chips, a couple in the United States, one in the Netherlands, and one other large one in Japan. Five companies play the dominant role in the manufacture of the machines that make chips. And in some ways, it's actually harder to make the machines that make chips than it is to make the chips themselves. Because these tools are among the most precise tools that have ever been deployed. Just to give you one example, ASML, a company based in the Netherlands, produces machines that are used in the manufacture of almost every high-end chip today. And these machines are capable of manipulating materials at basically the atomic level to produce chips with billions and billions of transistors like those that are inside of your phone or that are used for training AI systems. So there's a pretty small number of companies that make chips. And when you look at specific types of chips, you find that there's even more concentration. The biggest chip maker in the world is the Taiwan Semiconductor Manufacturing Company. When it comes to advanced processor chips, like the chips in your phone, or the chips in your computer, TSMC makes around 90% of them. So they've got an extraordinary market share, and are probably the most important semiconductor company, and arguably the most important company, in the world, because the chips that they produce, we rely on for basically everything. There's been a lot of consolidation in the chip industry over the past couple of decades, and it's been driven by economics and by technology. Today, a single cutting edge chipmaking facility can cost \$20 billion, one of the most expensive factories in all of human history. And so there's just a couple of companies that can afford to put up that sum of money on a regular basis to build more and more cutting edge facilities. And to make that work financially, you've gotta produce a ton of chips. And so there are huge benefits that accrue to the largest firms. The more chips you produce, the more your cost structure makes sense, and the better your technology gets, because you learn from every chip you manufacture, you gather data from it, and you tweak your manufacturing process to make sure you've got fewer and fewer impurities at every step. And so TSMC is both the world's largest chip maker, but it's also the world's most advanced, precisely because it gathers more data than anyone else. Because chipmaking requires ultra-purified materials and hugely complex equipment, there's not a single company that can do it on its own. Everyone requires a set of partnerships with supply chain providers to give them the materials, and the intellectual property, and the software and the tools that they need to produce advanced chips. And so if you take for example, the primary processor inside of your smartphone, it was probably made in Taiwan, but it was made in Taiwan using chipmaking tools from the Netherlands, and from the United States, and from Japan. It was produced using chemicals from Japan, and then often assembled and packaged in Malaysia

before ending up inside of your smartphone. And that's typical. A typical chip requires components and materials sourced from dozens of different companies because the process is simply too hard for any one company to do on its own. So, a nanometer is a billionth of a meter, and chips today are measured in nanometers. If you look at the chip inside of your phone, for example, and try to measure the size of the transistors, of which there will be billions on your smartphone chip, each one of these will be measured in a handful of nanometers. And so that makes them only slightly larger than atoms, smaller than any sort of living thing, far smaller than a bacteria, smaller than a mitochondria, half the size, for the most cutting edge transistors, of a coronavirus. There's basically nothing we manufacture at such tiny scale as we do with semiconductors. Every year, we make more transistors than we've made all other goods combined in all of human history. And in fact, nothing else really comes close. A typical smartphone chip could have 10 billion transistors just in the main processor chip. A big data center run by Google or Amazon Web Services would have more transistors than you could plausibly count. We know that we make more transistors than there are cells in the human body, for example. We don't even know how many we make in aggregate, because there are just so many. Moore's Law predicts that the number of transistors per chip, and as a result, the computing power per chip will double every couple of years. And that's been empirically true since the 1960s, which means that the capabilities of chips have gotten vastly better, and continue to get much, much better at a faster rate than anything else. So I like to think, for example, of airplanes to illustrate the difference. If airplanes doubled in speed every two years from the 1960s up to the present, we'd be flying faster, literally, than the speed of light. But chips have done that. Chips have increased in that capability because the scale of the transistors has shrunk to the level that today we're manufacturing them smaller than even viruses. And that has enabled the explosion of computing power, both in terms of the computing capabilities in high-powered data centers or in your phone, but also the application of computing to all sorts of devices. 'Cause today, there's computing everywhere. It's in your dishwasher, it's in your refrigerator, it's in your coffee maker, it's in your car. And it's possible to put computing everywhere because today it's so cheap, we can produce it almost for free. And that has enabled the application of chips to all sorts of different devices. To understand the change and the rate of innovation, in the 1950s, you could hold a single transistor in your hand. Today, you can hold 10 billion transistors in your hand in a chip that's the size of your fingernail. And that's not an expensive chip, that's a chip that often will just cost \$50 or so. So the rate of shrinking transistors, as well as the rate of decline in their cost, has been unparalleled in any other segment of the economy. So before transistors, computers used vacuum tubes, which are sort of light bulb like-devices that would turn on and off, on and off to produce the ones and zeros. And they were cutting edge for their time, but they had huge inefficiencies. They wasted a lot of heat, for example, they worked pretty slowly. And they also, because they created light, attracted moths, and so computers had to be regularly debugged in the early days of computing, which meant removing moths from the lights that they were attracted to. You can see why it was hard to scale that up into a 10 billion unit system. You know, I think the transistor is the key reason why we've been able to scale down. There's really nothing else, if you look all across the economy, that has shrunk in size and shrunk in cost at that level. And it's done so not just for a couple years, it's done so now for over

half a century. And that's why when you compare progress in the computing industry to progress anywhere else, there's really no comparison. Well, Moore's Law is not a law of nature, it's not a law of physics. We wish it were, because then we could rely on it to keep delivering advances far into the future. But it's really a law of economics. It says that, if you're able to find a way to shrink, shrink your transistors smaller, then you will be able to find a larger market as well. And that has incentivized huge investments in shrinking, in improving manufacturing processes, and making chemicals more purified to enable it, which has sustained this rate of advance. And if ever it turns out that the economics are on Moore's Law break down, the technology will immediately break down as well. Thankfully, the good news is that, right now, we're seeing a new wave of excitement about ways you can deploy computing, which has led to a surge of new investment into AI, but also a surge of new investment into semiconductors, because it's now clear that if we can shrink even further, we'll enable a whole new era of advances in artificial intelligence that rely on even more computing than we've been able to muster thus far.

You can define Moore's Law in a bunch of different ways. Is it based on the 2D size of the transistor, or the 3D size of the transistor? Is it based on the processing speed that comes out of it? And I think there's a lot of people in the industry that are trying to sell a certain chip with given characteristics that have an incentive to say Moore's Law, based on the other characteristics, has come to a halt. If you look at the rate of increase of machine learning semiconductors, for example, chips that are optimized for AI capabilities, they've been doubling in their capabilities every two years for the past decade or so. In other words, exactly what Gordon Moore predicted when he set out Moore's Law in 1965. And so my view is that when you zoom out and look at the rate of technological progress, there's really no slowdown that's happening. When I started my research on semiconductors, I thought that because chips were everywhere, chips were easy to make, and because nuclear bombs were only controlled by a handful of governments, they were hard to make. But what I realized is it's actually the exact opposite.

If you take nuclear weapons, that technology has barely improved since the 1960s. It's so easy to make nuclear bombs, even the North Koreans can do it. But chips are everywhere because they're cheap and they're tiny, and making things very inexpensive and very small is extraordinarily difficult, which is why there's just a couple companies in the world that can do it at the cutting edge. And the reason is that it's brutally expensive, and it requires manufacturing processes that get better, and better, and better every single year. And so if you're trying to catch up to the cutting edge in the chip industry, you're not trying to catch up to a static cutting edge, you're trying to catch up to a cutting edge that is racing forward at the rate of Moore's Law, doubling every two years. And so it's a race between companies, but it's the fastest race humans have ever undertaken, which is why it's extraordinarily difficult to reach the cutting edge. A couple years ago, it became harder to shrink transistors in two-dimensional format.

For a long time, chips were made, they were just

described as planar chips, chips in a plane, in which all the transistors were on the same level. Now we've started making transistors that have three dimensions, because we're learning to stack them on top of each other to package more of them together in a way that produces more computing power. And so one of the key trends over the next couple of years is going to be more 3D construction of groups of transistors, which will enable more of them to be crammed into a small amount of space. So the machines that make chips are extraordinarily precise in their manufacturing. For example, there are tools that can lay down thin films of material that are just a couple of atoms thick with basically perfect uniformity. And to pattern the transistors on a piece of silicon, you use a tool called a lithography tool. And today there's one company, ASML, of the Netherlands, which makes most of the world's lithography tools. And for the most advanced chips, these tools can cost \$350 million a piece for a single tool. And they cost so much because they require some of the most precise components ever used, like a mirror that's the flattest mirror humans have ever made, a laser that's the most powerful laser ever deployed in a commercial device, and a ball of tin that falls through a vacuum that is struck twice by that laser, explodes into a plasma measuring 40 times the temperature of the surface of the Sun, and this plasma emits light at just the right wavelength, 13.5 nanometers, to be bounced off the mirrors in exactly the right geometry and land on your chip to carve the transistors into the silicon. It's the most complex and expensive machine that humans have ever made, and it's required to make all of the most advanced chips. Today, there are just three companies capable of producing cutting edge processor chips, the types of chips that go in phones, or computers, or are used for AI. And it used to be a larger number of companies that could produce at the cutting edge, but it's shrunk into three, and might in the future shrink only to two for two reasons. First, the expense is extraordinary. \$20 billion per facility is a level of spending that many governments can't afford, to say nothing of companies. But second, the scale required to manufacture efficiently is vast. And that means that the benefits accrue to the largest firm. And in this case, that's TSMC, the Taiwanese firm that's at the center of the chip industry. That's why they manufacture on 90% of the most advanced chips, because they're cheaper, and they're better than their competitors when it comes to manufacturing.