Episode 69 • 1 September 2023

Jon Y (Asianometry) on Problems And Progress in Semiconductor Manufacturing

1010

ChaptersSpeed 1X

00:00 / 01:46:50

1x

Chapter 1Welcome, Jon!

00:00Welcome, Jon!

02:42The State of AI Hardware

11:22Potential Bottlenecks To Progress

36:16CPUs, GPUs, TPUs

43:26Potential Disruptions

49:12Thinking about AGI

59:26Semiconductor Supply Chains

01:11:23Chips, Taiwan, and Geopolitics

01:18:27Superconductors

01:32:10How To Learn About Chips

01:34:49Being An Online Creator

01:39:39Closing Quesitons

01:46:06Outro

00:00 / 01:46:50

0.5x5x

1x

1010

00:00 / 01:46:50

ChaptersSpeed 1X

Chapter 1Welcome, Jon!

00:00Welcome, Jon!

02:42The State of AI Hardware

11:22Potential Bottlenecks To Progress

36:16CPUs, GPUs, TPUs

43:26Potential Disruptions

49:12Thinking about AGI

59:26Semiconductor Supply Chains

01:11:23Chips, Taiwan, and Geopolitics

01:18:27Superconductors

01:32:10How To Learn About Chips

01:34:49Being An Online Creator

01:39:39Closing Quesitons

01:46:06Outro

00:00 / 01:46:50

0.5x5x

1x

← Back to all episodes Leave feedback ↗

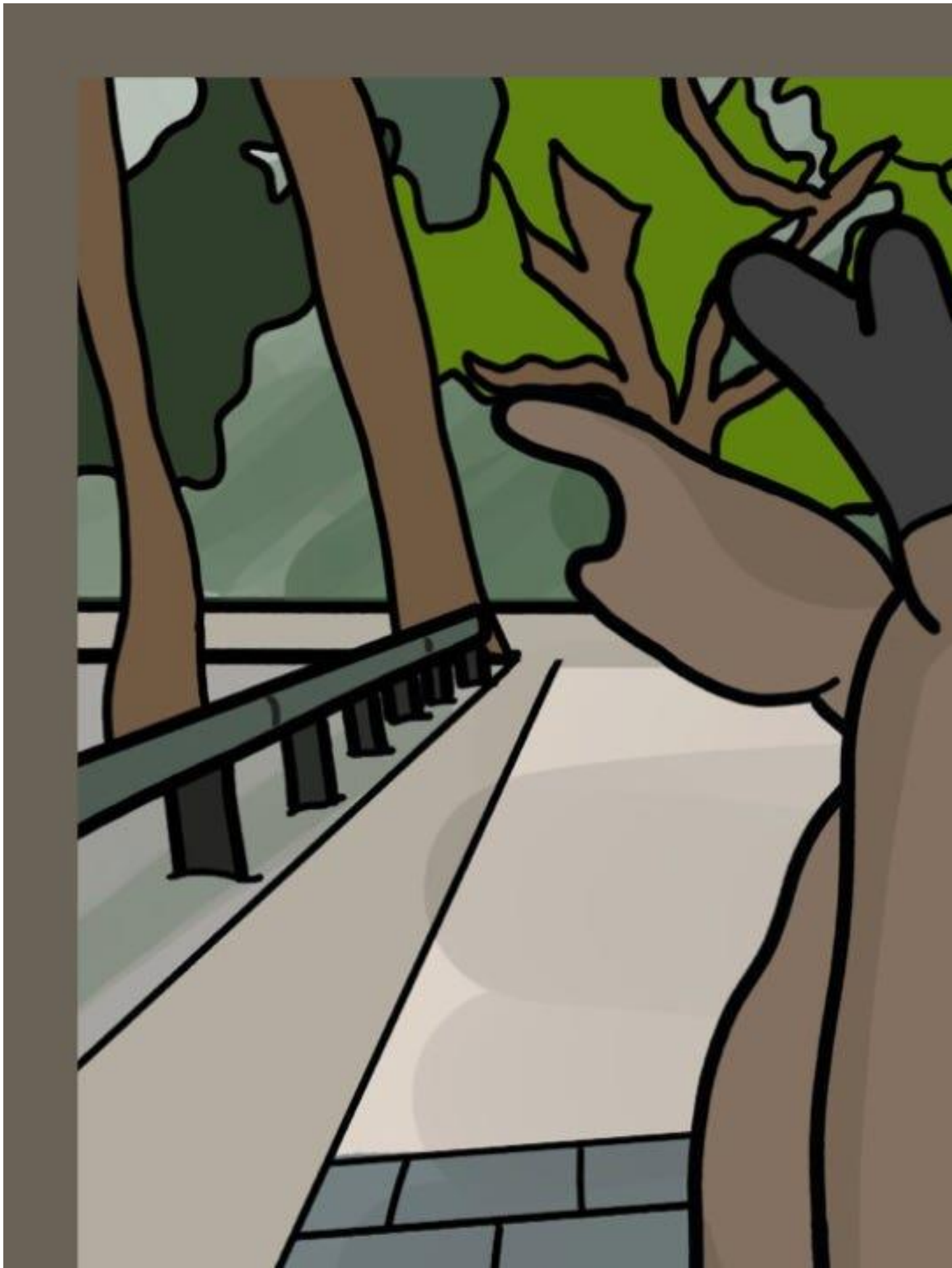
Contents

Contents

14822 words, 75 min read

Back to top ↑

Jon Y is the creator of the Asianometry YouTube channel and accompanying newsletter. He describes his channel as making "video essays on business, economics, and history. Sometimes about Asia, but not always."



In this episode we talk about:

- Compute trends driving recent progress in Artificial Intelligence;
- The semiconductor supply chain and its geopolitics;
- The buzz around LK-99 and superconductivity.

Jon's recommended reading

- The History of Semiconductor Engineering by Bo Lojek
- Tiger Technology: The Creation of a Semiconductor Industry in East Asia by John Matthews and Dong-Sung Cho
- Handbook of East Asian Entrepreneurship by Fu-Lai Tony Yu and Ho-Don Yan

More resources

- Asianometry has its own playlists, including Semiconductor "Course" and Computer History, which are great places to find relevant videos
    - In this interview we especially drew on The Coming AI Chip Boom, How Nvidia Won AI, and AI's Hardware Problem
    - We also mentioned LK-99 Wouldn't Have Changed Semiconductors Anyway, How China Got the Bomb, and India's Semiconductor Failure
- Semiconductor blogs
    - Fabricated Knowledge [blog]
    - Semi Analysis [blog]
    - The Chip Letter [blog]
    - China Talk [blog/podcast]
    - SemiWiki [open forum]
    - SemiEngineering [news site]
- Other related podcast episodes
    - Chris Miller on the History of Semiconductors, TSMC, and the CHIPS Act
    - Jaime Sevilla on Trends in Machine Learning
    - Lennart Heim on the compute governance era and what has to come after

Transcript

Note that this transcript is **machine-generated**, by a model which makes frequent mistakes and sometimes hallucinates entire sentences. Please check with the original audio before using this transcript to quote our guest.

Fin / Luca

Hi, you're listening to Hear This Idea. In this episode, we speak to Jon Y, the creator behind the Asianometry YouTube channel and its accompanying newsletter. Asianometry is, in my mind at least, one of the best resources on the intricacies of the business, history, and engineering details that shape the semiconductor industry. I've been an avid follower of John's content for years now, so it's a real treat to have this conversation with him. We talked about the many compute trends driving recent progress in artificial intelligence, and his speculations on the potential challenges and innovations we might see in the near future. We also discussed an in-depth look at the semiconductor supply chain and how it is intertwined with geopolitics, with John sharing some of his personal insights from living in Taiwan. Lastly, we talked about the buzz around LK99 and superconductivity, particularly if we were to discover such a material, what its real-world applications could be, and some of its limitations. On that last topic, I'll flag that we recorded this interview back in early August, when people were more optimistic that LK99 would replicate. Although to John's credit, I'll say that I think our conversation survives that. As John notes in the interview, his interest is much more in the nuts and bolts of how semiconductors work and the relevant players today. In contrast to some of our other recent guests, perhaps, he has thought much less explicitly about topics like artificial general intelligence and its implications. But I think his viewpoint here is really enlightening and complementary. It sheds light on the complexities behind AI scaling trends, highlights the challenges that a policymaker might face in this realm, and anchors some of our past discussions to real-world hurdles. I think I was also left with a much better sense of how companies like TSMC or Nvidia might perceive recent progress and what issues are most salient to them. So if you haven't already, I really strongly urge you to check out Asianometry videos. But without further ado, here's the episode. Cool. So welcome to the show, John. Maybe to kick things off, we often like to ask the question, what's the problem that you're currently stuck on?

Jon

I'm working right now on a question about LK99, that new room-temperature superconductor. What does that mean for the semiconductor industry? Is this something that can actually be used in semiconductors? Can it be fit in the interconnects? Can it be fit within any part of the system? I have no idea what I'm doing. I'm reading all this stuff, and I'm like… this is insane. I feel like I was walking on ice and I have now fallen into a 377 Kelvin chamber. It's freezing in here.

Fin / Luca

Excellent. I really hope this gets derailed and we get to just talk about superconductors for half an hour. But let's ask some proper questions before that. So here's just a naive question to kick things off, right? As far as I understand, in consumer hardware, like in my phone or laptop, there are chips, GPUs in some of them, CPUs. But I also understand that when it comes to the kinds of hardware used to train big AI models, they're very different and increasingly specialized. What makes that hardware special? How is it different?

I think all hardware is defined by their power requirements, what they need to do in terms of computation, and what they need to achieve for their goals. So you mentioned smartphones, which are one of the big categories in the semiconductor industry. They're special because they have these constraints that are required by power. They need to be power-efficient, not only that, but they also need to be very small. They need to be put in special packaging. For example, they need to be stacked with their memory on top of them. There's a lot that needs to be done that's specialized by the manufacturer, the fab, to ensure that the chip is suitable for the eventual end-use case, which is the phone. In certain cases, where you have these massive AI accelerators, they don't necessarily have those same restraints because you put them in a data center. They're kind of safe in their own server or "nest," I would say. Then you're looking at power, speed, and parallelization, and these constraints guide how the final chip will eventually look.

Fin / Luca

Got it. I presume there is less of a constraint on power because you don't need to rely on a single wall socket, like you would with a laptop or a battery. There's also less of a constraint on size because you just have a big building. So, unusual extents you're doing the kind of work that can be parallelized, and that changes the kind of chips you're using. Is that right?

Jon

Right, because a smartphone chip is not just a chip that does computation. These are generally called system-on-chips, so they do a whole bunch of things. The chip inside an iPhone, for example, is not just doing CPU tasks but also image processing, AI prediction, and basically all the things that a phone needs to do. That's far more than just computation. There are analog components, power-saving features, and all that needs to be integrated. They try to stuff as much of that into a single die because that's how you achieve the most scale.

Fin / Luca

Yeah, I think that leads nicely to the next question I want to ask. There seems to be a lot of attention around small nodes. In particular, when we talk about companies like TSMC or ASML, there's a lot of emphasis on these companies being able to produce equipment or manufacture at the very small end, like three nanometers versus five nanometers versus seven nanometers. Why is miniaturization useful for making chips faster and more efficient?

Jon

It's tied to the idea that when you put more transistors, when you put more devices onto an integrated circuit, it performs better. You have situations in history where you have discrete components of some prior system, and they are put together on a circuit board. Now you have the ability to compress it all into a single chip. Once you put that on a single die, you not only have the ability to add more functionality, but you also bring that into the lithography and mass production world of semiconductors. Once you can do that, it's like a printing press where you can print 100 chips in a single pass. That's part of the value of miniaturization—these chips are getting more powerful, you're putting more onto a single die, and connecting them with super-fast interconnects. When you have a billion or 10 billion of a certain thing on a single die, I think magic starts to happen.

Fin / Luca

When I read about the 5-nanometer process, and there's a new 3-nanometer chipset coming out, what do those 3 and 5 nanometers refer to?

Jon

The nanometer part is part of the popular conception. I don't think the fabs actually call it that. Apple does, which is terrible. But it's a marketing term. It used to mean something. It used to refer to the distances between certain transistors, up until a certain point in history when the industry made a fundamental change in how they built transistors and basically turned them on their side. Once that happened, all the previous metrics no longer mattered because the change was not in how many more transistors were put on a single die, but an improvement in how those transistors actually worked. Once they made that change, the industry was like, we still don't know how to convey that this is a better process. I think either TSMC or Intel did the FinFET transition first, and TSMC jumped later. The estimation that TSMC made was that they would continue decrementing the nanometer count. Back then, it was like a micron. They switched to nanometers, and that was a huge debate. Now the decrements are vague indicators of

the improvement of Moore's Law. By now, these chips are so advanced that you can no longer comprehend them with a single metric.

Fin / Luca

I think you also have a video explaining that for TSMC, there isn't just a single type of 3-nanometer fab anymore, that there are lots of sub-classifications. Can you explain what's going on there?

Jon

Nowadays, when you're talking about these super leading-edge fabs, if you imagine a TSMC process node as a manufacturing line producing chips of a certain class, you're basically working with TSMC's biggest customers like Apple, NVIDIA, and AMD to spec out what this might look like. In the end, you almost have a customization. So this new iPhone chip for the iPhone 15 is using a very specific process, customized for Apple. Eventually, after about a year of working out the kinks, they'll generalize that more to other customers. It's hard to think about which customers will use these super advanced nodes down the line because they're so expensive, different, and advanced, but that's another topic for another day.

Fin / Luca

Yeah, well, maybe to try to segue into what I'm really interested in spending a good chunk of this conversation on: using very high-end and specialized chips for the purposes of training and running very large-scale AI systems. It's obviously been a lot in the news with ChatGPT and Anthropic that large language models seem increasingly useful and that we can expect to see increasing orders of magnitude progress in their performances and capabilities. I'm curious from a hardware angle, what it would take to enable these orders of magnitudes of progress, especially given that it seems scale has really been a driver in the past decade of enabling this progress.

Jon

I've been thinking about this for a while, too, because on one side, there seems to be a capacity shortfall. They're just not delivering enough hardware, which seems to have caught everyone by surprise. I do think TSMC is very good at making more stuff if you pay them for it, so I think that'll be fine. But in terms of advancing the capabilities of hardware, you might be thinking more about things like interconnect speed, advanced packaging, and making sure that we can process more data in a certain cycle. That's part hardware and part software. A lot of people, when they complain about hardware, they

don't think they need their hardware to be better; they just need it to be cheaper. That's a manufacturing problem and a hardware problem. I don't know if I made any sense of that, but it's an interesting question: what can we do beyond just expanding capacity to make the hardware more sufficient for what AI will be in the future?

Fin / Luca

Do you think it would be possible to achieve two or three orders of magnitude improvement in performance just by scaling up the amount of training compute you're throwing at some model by 100 or 1000 times? Do you think that could be achieved just by making these manufacturing processes cheaper and scaling them up, or would that require some more innovation, like actually improving the chips themselves?

Jon

I think right now, HPC (high-performance computing) at TSMC is about 7% of their revenue, which is far short of what Apple does. Apple does around 20% of their revenue. A lot of that is iPhone. So if you think about it, there's still so much that might be done simply by TSMC ramping up capacity. Even within the company itself, they're not that big. Before we talk about changing the hardware, I think a lot of people will say, let's just let that capacity ramp up and catch up to demand.

Jon

Just to make sure I understood what you were saying there: around 7% of TSMC's capacity or revenue currently comes from AI-like chips, and 20% comes from NVIDIA.

Fin / Luca

NVIDIA.

Jon

NVIDIA, yes.

Fin / Luca

Okay. But 20%, for example, comes from iPhone chips. So if you just look at the percentages, there's easily an order of magnitude you could achieve just by having TSMC refocus its supply or refocus which customer it serves, let alone needing another TSMC or letting TSMC itself.

Jon

It's about planning, right? Right now, the situation is that they ship around 100 million iPhones in a quarter, and they all need chips. TSMC can match that capacity and bring that sort of capacity to the market. The issue with what happened with NVIDIA and the chip shortage in the automobile industry a couple of years ago was that it hit unexpectedly. It's not like an AWS server where you just ramp up more servers. There are things being made. But when you give the time to actually make those things, I think people will stop talking about a GPU shortage, which is what people are talking about right now in Silicon Valley.

Fin / Luca

Yeah. And just concretely, we're talking about chips for AI. TSMC is making a lot of them, maybe most of them. I guess the main media customer is NVIDIA, making GPUs. But what are those chips actually doing that's relevant for AI? Is it training? Is it inference? What are most of them doing?

Jon

They'll probably be doing a lot of inference, which is where you have the model, and the model is basically put into the server and generates the result. With large language models, you're generating the next token over and over again. Training is where you create the model, but inference is where you actually run it. For the most part, that's where a lot of these models will incur their cost. The training cost could still be very high, but inference will be massive, especially when you consider these AI products, which cost a lot and are being ramped across hundreds of millions, maybe even billions of people.

Fin / Luca

Okay, so you expect that most of the demand for compute will be on the inference side, especially once the models really scale and get a huge customer base?

Jon

Yeah, assuming we can find that, yeah. Right.

Jon

Right.

So two words that often get used when we talk about compute are processing and memory. Can you explain a little bit about what these are relevant for when we're talking about chips, and maybe also distinguish how important they are in the training phase of a model versus in the inference phase of a model?

Jon

That's a good question. A lot of people talk about semiconductors and what chips are, but they don't talk about the type of chips. In the 1980s and 70s, the most important chip in the world was the memory chip. These memory chips basically store information. When you talk about dynamic RAM, it takes a bit and stores it as a charge, with millions and billions of transistors and capacitors storing charges. Memory basically is the information; it stores the model and the programs. When needed, the processor, or logic chips, will pull the information out of the memory, perform operations like multiplication or addition, and send it back to memory for storage and additional processing down the line.

These are two very different chips: a logic chip and a memory chip. It's important to make clear that they are very different types of chips. They're made with the same processes and equipment, but there are subtle differences between the two. They are large categories in the semiconductor industry, each with its own market dynamics. Logic and memory are very different, but you need both to run a computer.

Fin / Luca

Yeah, I remember this detail in one of your videos where you mentioned that the energy cost for accessing data off-chip and then pulling it in to do floating-point operations is something like 200 times more costly than doing those logic operations. Similarly, something like 80% of energy usage for Google's TPUs tends to be from these electrical connections, pulling in memory rather than just the logic. So why does memory seem to be a huge bottleneck, especially for these AI applications? Why can't you just add more memory?

Jon

I think the toughest part is that when you access memory, it needs to go off the chip. There is an exception: static RAM, which sits on the chip alongside the logic. The problem with SRAM is that it has limitations because you're not using memory semiconductor manufacturing processes to make it; you're using logic memory processes. There are also size constraints and other factors.

When you move data on the chip, it's faster because you're using the shortest distances and certain types of interconnects to move the data quickly. However, when you pull data from anywhere, even on the chip, it consumes power because electrical signals encounter line resistances, and you're losing power as it travels through. Imagine this challenge on the chip, and then consider it off the chip. When you have sticks of memory stacked on top of the logic, the distances are even greater, and you have I/O issues with pulling the memory out and then into the logic parts of the chip. All of that is complicated and incurs costs.

In the context of running massive AI models with hundreds of millions of tokens and calculations that need to happen every second fast enough for the product to be viable, this becomes a significant issue.

Fin / Luca

Got it. That sounds like one major bottleneck for increasing the performance of chips specialized for AI—the bottleneck being the interconnect to pull in memory for processing. Are there any other bottlenecks that seem especially important looking to the next few years, particularly for AI?

Jon

I don't see many other bottlenecks; I think memory is the main issue. The dream would be to have enough memory to load all of these into a single chip, but that's simply not possible with the way these models are progressing. Memory is crucial to AI, perhaps even more than the logic part. The computation doesn't seem to be all that complicated, but the important part is what's stored in the model in memory. It needs to be accessed quickly, and there are restrictions on how fast that can happen. The whole race is to ensure that we don't have bottlenecks that cause serious latency in these systems, making them less competitive with whatever is out there.

Fin / Luca

Yeah, got it. You mentioned the dream of storing the entire model on a chip. Can you elaborate on that?

Jon

It would be absolutely ridiculous, right? If you could somehow have this massive chip where half of the wafer is all memory created in a specific style, it could store the entire model on the chip. Then it could be accessed using traditional on-chip interconnects

directly to the logic parts of the chip. You would gain a significant performance benefit, but you would also end up with something that might not be economical. That's why you need to scale it back down to using advanced packaging to bring that data over. It's all about accessing that stored information quickly.

Fin / Luca

I've heard that, in some ways, the actual logic operations that AI chips perform during training or inference aren't that different or complicated. Can you explain a bit more about that?

Jon

There was a paper on the Google TPU that indicated that most of it was just adding and accumulating, like multipliers and accumulators. It involves very simple operations, but there are a lot of them. If you think about it, it's somewhat similar to a Bitcoin miner, which does one repetitive thing. The TPU is a specialized chip with circuits for specific operations, primarily matrix multiplications. It involves a lot of adding and multiplying two numbers, then adding them together again, and doing that millions, even hundreds of millions of times each second.

When you consider this as a chip designer, you end up building a chip tailored to meet that need. This is different from something like a CPU or GPU, which are more generalized and have different parts dedicated to various operations. For instance, Intel would have certain operations for running specific programs, while a gaming GPU would have dedicated parts for graphics processing. Chips are built to strike a balance between being specialized for a project and being generalized for other use cases.

Fin / Luca

Like my laptop, which needs to be a jack of all trades because I might be watching videos one day and browsing the web the next.

Jon

Exactly. That's why the Apple M1 is so much faster; it's customized to handle all the tasks that macOS needs to perform.

Fin / Luca

To simplify things, when we're talking about logic and processing, we're discussing multiplying and adding, and when we're talking about memory, we're discussing remembering which matrices to use, right?

Jon

Yes, that's correct. It's about the weights.

Fin / Luca

I'm also curious about energy consumption. Is energy consumption or energy costs becoming a serious constraint? We talked about how much more energy-intensive it is to access something off the chip rather than on-chip. But where do you see this heading, especially when considering data centers, cooling, and running the chips themselves? This was clearly a big point during the crypto mining discussions, but what do you see for AI?

Jon

It's tough. For a long time, when chips got smaller, the performance per power unit improved because there was less distance between transistors. As the chips became denser, nanoscale effects began to consume energy. For example, with a TPU or NVIDIA AI accelerator, a large percentage of the power usage comes from moving data around.

Currently, energy costs are not as significant compared to the benefits gained from optimizing for more performance. I don't know where it ends up, but there could be architectural changes or the use of different materials. However, it seems challenging to find ways to cut power usage on a per-chip basis.

Fin / Luca

Do you have a sense of what fraction of costs currently come from energy versus buying the hardware upfront?

Jon

What do you mean by that? Do you mean renting hardware?

Fin / Luca

Yes, for example, if I'm OpenAI and I want people to use my ChatGPT, how much of the costs come from purchasing or renting H100s versus the energy to power them? Do you see that changing as AI scales?

Jon

I think renting AI power is likely to be more expensive. If you look at how much AWS is making, they're doing quite well. It might be more cost-effective to buy the hardware upfront and use renewable energy sources like wind or solar to run it.

Fin / Luca

I'm curious about how much of the costs come from the actual chip or hardware versus the energy to run that hardware. The reason I'm asking is that, with crypto mining, energy was a significant part of the cost, leading people to relocate to places with cooler climates and hydroelectric power. Does this apply to AI data centers and training as well?

Jon

Yes, it does. The backlash against Bitcoin mining was partly due to the anger over energy being used to create digital numbers that some believe have value. The energy usage for AI is similar; there's nothing inherent to AI that makes it use less energy. In fact, it might even use more because you're using a lot more hardware.

Consider the energy required to produce the chip itself. Manufacturing a microchip is incredibly power-intensive. I believe around 30% of the total power usage of a microchip is in its operation, while 70% is in production. TSMC is building a fab in Taiwan that uses enough power to run a small country. TSMC alone uses 7% to 8% of all of Taiwan's energy. So, when discussing power usage, I'm more concerned about how we will have enough power to make these chips.

Fin / Luca

That seems relevant for questions about how quickly we can scale up, not just the fraction of existing output used for AI applications, but the total output. If you're thinking about doubling or tenfold increasing that, you'd start consuming a significant portion of Taiwan's energy.

Jon

Yes, it's staggering how much power these things use. A single EUV machine uses the equivalent of three Walmart supercenters, and a TSMC fab might have around 80 of these machines. It adds up, and that's just one part of a very complicated supply chain, which includes clean rooms, etching, ion implantation, and lithography.

Fin / Luca

Just quickly, we were asking about the costs to buy or rent the hardware as an AI company. I realized I actually don't have a good sense. So if I'm like, you know, OpenAI or another big lab, am I more likely to be just buying the chips and running them on my own server? Or am I more likely to be going to a cloud provider like AWS and getting them to handle everything for me?

Jon

I would probably reckon that if you get to a certain size, you want to be buying the hardware yourself because otherwise you're paying… That's margin going to Andy Jassy rather than yourself, right? That's money being taken out of your pocket. I think I would definitely agree that at a certain size, you want to be making stuff. You're making your own sets of systems.

Fin / Luca

Yeah, got it. So I guess there's a difference in the company size. Maybe kind of the smaller cap AI startups, they're just going to be renting the hardware because it's much more straightforward.

Jon

Yep.

Fin / Luca

Let's maybe do a quick crash course on how we got to compute where we are today. So in particular, you mentioned CPUs and GPUs, and then I think Finn also mentioned TPUs. We're talking here about kind of an evolution of how chips, especially chips for AI, have evolved. Can you maybe very quickly talk about the relevance of having shifted from CPUs to GPUs and why this was important in enabling the explosion of neural networks?

Jon

So CPU, I think it's a very interesting question you ask, right? Because you're talking about a situation where Intel used to dominate the CPU market, right? Intel dominated the CPU market for a very long time. They had a specific type of chip that really mattered to the technology trend in vogue during the 1980s through the 1990s and early 2000s. That was probably the CPU, right? PCs and CPUs. But once Intel no longer had that, once that sort of moved away from PCs, once PCs peaked, then it moved to a certain situation, right? So that's how Intel found itself having power moved away from it, being left behind. From there, CPUs went to smartphone chips, and then you're talking about now GPUs, right? So with GPUs, what NVIDIA did was to make this transition away from graphics to generalizing to larger, more parallelized workloads, like computation workloads. That turned out to be, I mean, I don't think anyone had any idea that was going to be it. But that fit neural networks pretty well. This whole GPU thing is very recent, in my opinion. I don't want to say there was a transition from CPU to GPU because right now, if you think about it, Intel is still far larger than NVIDIA, and they still make more CPUs than NVIDIA. It's more interesting to think about that dynamic down the line a couple of years, but right now, CPUs are still very dominant in the industry, along with smartphone chips.

Fin / Luca

But dominant in the sense of compute being used across the world, as opposed to compute being used specifically for training AI models. Is that right?

Jon

Yeah, yeah, yeah. Making this stuff. Yeah.

Fin / Luca

You mentioned parallelization there. Can you briefly explain what you mean by that? And also, I guess, how is it possible to run tasks in parallel? How does that actually look like when we're talking about training or running inference on something?

Jon

Parallelization actually is a CPU thing. It started out in the CPU world. It was a situation by Intel where they were no longer able to speed up the clock speed of their CPUs. So what they did is split the CPU to create cores, right? A CPU might have a couple of cores, like 16, 20, or even maybe 50 CPU cores. But a GPU takes that core to a much larger level, thousands of cores. Each core can be thought of as an independent unit working on a certain type of software, taking in data, doing its own thing, and then

eventually bringing the result together. So parallelization is the concept of breaking a single coherent task into little bits that can be fed individually to each one of these cores, and they all work on it together to put it all together at the end. You end up doing it much faster. When Intel announced they were going to do this parallelization thing, I don't exactly remember the timeline, but I know that NVIDIA was trying to do this sort of parallelization at the same time Intel did. In the early 2000s, Intel was basically saying, "We can no longer speed up the clock speeds; we're doing this parallelization," and that was a pretty big issue for the Windows PC world because they had to rewrite their software to take advantage of more parallelization. It's very complicated. The whole chip industry makes no sense.

Fin / Luca

You mentioned software there, so I'm wondering, does this tie in with NVIDIA's CUDA software as well? I know this has gotten a lot of press attention recently, especially explaining why NVIDIA is able to stay at the frontier or lead the market here in this regard.

Jon

It's pretty interesting if you think about it because CUDA is a way to run these AI programs and everything's written on top of it. In some ways, it's not necessarily a hardware performance advantage. It's not an advantage tied to NVIDIA's hardware being simply better. It's tied to an ecosystem. It's tied to adoption. It's almost like Windows, I would say. It's a social advantage in a sense that brings benefits that other competitors struggle to break into. I hear a lot of people saying that eventually they'll break through it, but it really does feel like NVIDIA is okay for now.

Fin / Luca

I guess when we're looking at the story of computing hardware, there's this shift towards parallelization, first in CPUs, then GPUs. But then there's even more specialization in the direction of AI, right? As long as we're talking about AI. So on one hand, I have my gaming GPU that's hooked up to my desktop computer. On the other hand, I hear about, for instance, Google's TPU, which we've mentioned a couple of times. Chris, what are the differences between those two things?

Jon

A TPU, I would say, is built to run a specific workload. When you think about GPUs, in some ways, they're not all that different. The difference is sort of how their circuits are

being used. In the case of a GPU being used for graphics, they've generalized the type of programs that would be run to generate an image, to turn gaming data into a rendered image for the user. A lot of the differences are tied to software and what compromises and designs were made to the circuit to run that software at the most optimal speed.

Fin / Luca

I noticed that all of the companies making these chips that we've mentioned are enormous and decades old, and they can just throw a lot of capital at building the next generation of chips. This makes me think that it might be very difficult for new entrants to disrupt those companies. That seems relevant for thinking about who are going to be the major players in 10 years' time. So I'm curious how you think about how easy it is to disrupt these existing players like NVIDIA, Google, Intel, TSMC, etc.

Jon

This is a good question. I think it ties to the demand. In some ways, you can argue that there's almost no easier path for a startup to rise up and build an interesting niche in the industry. You have a situation where TSMC is open to taking on a customer. If you're a startup and you raise like $200 million or something, and you design a chip, you can get TSMC to use the same processes that they use to make an A100 to make your chip. In that sense, it's easier than ever. What the difficulty is is finding that use case. That requires some sort of discovery that your startup should be looking at—what is the special thing that you have that these other guys don't have and cannot have due to their existing concerns?

Fin / Luca

Can you speak maybe a bit on what's possibly on the horizon here that might be something like this? For example, I know that you've mentioned silicon photonics and startups around that. What's the relevance there, and how might that disrupt if it does end up showing a breakthrough?

Jon

Silicon photonics is a tough one because I think it's an immature technology and it requires changes to silicon to be made. The concept is that you can use this to run neural networks and have substantially less loss because you're using light rather than electrical signals. Light travels at the speed of light; electrical signals cannot travel as fast through the interconnect. They travel at a quarter of the speed of light. If you were to

somehow create a chip that sends signals around using light, a silicon chip, then you'd be able to bring a disruptively faster, better product conceptually. That's like the dream, right? You'd be able to bring that out into the market, and everyone else would be like, "Oh my God, this is amazing," and they'd all adopt it. The problem is that silicon does not emit light. That's the fundamental core problem, which causes engineering challenges. You have to have compromises in some way. Either you dope the silicon or you bring the light from off-chip. All of that adds cost, and it erodes the potential disruptive advantage that a silicon photonics chip has. When you have a chip like when TSMC is marching towards 2 nanometers and they're going to put something like 15 billion transistors on a certain chip, that's very, very difficult to compete against. You need something insanely new, some sort of groundbreaking science.

Fin / Luca

Is it possible that we will see increasing specialization even within AI? For example, I think towards the beginning of this interview, we talked about the differences between inference and training large AI models. I'm curious whether you could see there being differences in how we trade chips off to specialize in one domain versus the other, such that those end up becoming two different ecosystems or two different types of chips that companies would be buying, whereas I currently understand that with the H100, we're using it for both.

Jon

It's possible. Yeah, it's definitely possible. But that is tied to the use case. Do you have a use case that is worth giving TSMC a billion dollars to make enough chips for you to do? Even if you were to have a smaller chip, there are constraints on fixed costs that any chip startup must overcome before they can bring their chip to fruition. You need to find out your end use case in order to overcome those fixed costs. Otherwise, you have no point in creating an ASIC for that sort of inference special case that you have in mind.

Fin / Luca

Yeah, this has been really interesting. Maybe to take a bit of a sharp left turn, I want to talk a bit about artificial general intelligence. In particular, I'm curious because you've spent a bunch of time thinking about compute and scaling laws and supply chains, and I think have a very distinct perspective on all of this. I'm just curious, how seriously do you take the premise that we would develop AGI, something that could learn to accomplish any intellectual task that human beings are able to perform, say, within this century?

Jon

I haven't really paid attention to it because I think it's one of those things. I live out in Taiwan, away from Silicon Valley, so we don't really feel the hype. When you ask ordinary people on the street about AGI, they're like, "I don't care." In my opinion, I don't think about it; I don't spend my time on AGI. So when I hear other people talking about AGI, I'm like, "Why? What's so interesting?"

Fin / Luca

Do you think that applies to folks working at TSMC? Do you expect that in leadership conversations, there are people saying, "We should really take seriously this premise that in 20 years' time, things are going to get crazy; demand is going to 100x"?

Jon

No, they don't worry about that. TSMC people, what they care about is getting the chip done this year. You can leave the AGI talk and AI and all that other stuff; they leave that to big brain people who sit in rooms at some big conference and talk about it. For TSMC and other people who focus on building these chips, they need to build the chips. What TSMC is worried about is, "Can we ship N3? Can we ship N2? And can we get the capacity to do that?" They don't worry about… In a recent earnings call, they said they don't know if AI revenue is locked in. They said the same thing about Bitcoin too. When the Bitcoin miners were buying all this hardware from them, they said on the call, "We don't know if this is going to be a thing." They said the same thing about AI. I think they're skeptical, as am I. I don't know what AGI means. I think it's a moving goalpost. The hardware… In some cases, it's hard to say where some of this AI stuff is going to go in the future.

Fin / Luca

Yeah, well, maybe one way to frame this question differently that's less AGI-centric is to ask how you see compute, or say the compute used in a training run, evolving over the next couple of decades. We've currently seen this very exponential increase in how large models have been able to go. You've mentioned previously that you actually see there not being really any fundamental hardware obstacles and, in many senses, a lot of spare capacity even within TSMC to achieve, let's say, the next order of magnitude increase. I'm curious if beyond that you see it being, if the demand is there, easier to sustain the next order of magnitude increase, say even two or three, whether in this scale is all you need hypothesis, you see any fundamental roadblocks.

Jon

Yes. The only thing the semiconductor industry wants is money. So if the money is there, TSMC, Samsung, and Intel will move heaven and earth to make it happen for the next 10 to 20 years. But that's always the key issue. Moore's law is not fundamentally limited right now by physics in some ways; it's partly an economic state. Is there an end-use case that can drive the next N2, N1, or N1.7? That's what the TSMC folks are worrying about. Is Apple going to keep funding another massive push to the next node? That's the big question everyone's asking.

Fin / Luca

Do you think those folks are entertaining the possibility that maybe Apple and other companies like Apple just decide that M2 was plenty good?

Jon

Yeah, that's a big worry. That's a huge concern. If you think about it, there's one big capstone: all the money that's flowing from there. It used to be Intel. Intel used to say, "We decide, we have anointed this, and we will fund this all with this money." You have all these changes, these improvements in the semiconductor ecosystem to drive forward future changes—transitions to copper interconnects, low-k dielectric materials. I've said a bunch of terms; it doesn't matter, but it just means that all these new things are happening. And they happen because there's a big customer at the end of the line saying, "I'm willing to pay for this." So their worry, everyone's worry, is: Is Apple going to someday say, "Yeah, like you said"?

Fin / Luca

I don't need anything better than this.

Jon

Yeah, we're good.

Fin / Luca

We're too good already.

Jon

We're good. We're good, right?

Fin / Luca

And then trying to tease this out in the context of AI, I guess the question would be: Do you really need a ChatGPT-5 or -6 when -4 or -5 are maybe enough? Especially on the commercialization side, the question is less about whether AGI is possible but whether we can market it enough or raise enough revenue to get to that point because a lot of this is going to be very expensive, even if it is possible.

Jon

Now that's a question that makes a lot more sense. Yeah, I think AI has a long way to go. I think it does just on the basis that I'm looking at it right now. It's very good now, but it could be better. If you think about it that way, there is a performance need for better hardware. Now, we haven't started to see that kind of flow through to the rest of the ecosystem. Why have we not seen Nvidia step up to become one of the big customers next to Apple? That's not happened. Right? So that's one of those big concerns that maybe says to me that right now, AI as an industry for hardware—I'm not talking about the hardware side—does not seem to be as big as the iPhone yet. The iPhone is still the king; it drives everything. Next to that are GPUs, AI stuff like that. Maybe that'll change. Maybe this next node, maybe for N2, we'll see Apple and TSMC or Apple and Nvidia ramping up their newest AI accelerators next. But we don't know.

Fin / Luca

Yeah. One way those things could change, maybe wrap up quickly, and indeed the overall size of the semiconductor industry could grow quickly, is if there is some kind of feedback loop where some generative AI becomes good at innovating on the hardware. I'm curious whether that happens at all and whether that feels like a feedback loop to you. I'm also curious whether you see that becoming more of a thing in the next, let's say, five or ten years?

Jon

Maybe. I mean, maybe. A lot of the digital circuits, these are billions of digital circuits now with each particular system. AI already, in some way or form, computers already design a lot of that. I think that's something that's probably going to continue building down the future. There's no big issue. I mean, can you build that capacity? I just certainly see AI improving digital circuits down the line. Then you can talk about it deeper at the level where can you use it to improve lithography and stuff like that. A lot of that is already being done on the fab side too.

Jon

As well.

Fin / Luca

Yeah, got it. Although I presume, if you can double the performance of whatever kinds of computing you're using to improve your manufacturing process, and that gives you like a 1% or 5% kind of boost in how cheaply you're making these things, that's not like a crazy explosive feedback loop. That's just a kind of useful bonus, but it's not the main driver of your output in the next few years.

Jon

Yeah. That's like hitting your OKR for the year, right?

Fin / Luca

Yeah. Right.

Jon

That's like, yeah, good job. $5 bonus.

Fin / Luca

Maybe one last angle I want to take on this question is: Are there any specific signals that you'd be looking out for as you're kind of looking ahead that could change your mind in terms of how quickly or how slow to expect AI progress to be?

Jon

I think it'd be very interesting if we see, like I said earlier, Nvidia steps up next. It used to be Apple and Huawei, right? These two companies led TSMC's newest leading-edge node. I think TSMC preferred it that way because they do want to have customer diversification. So in this case, since Huawei is no longer available, it's been Apple alone. I think that's been a big concern for a while now. If you see one of these AI companies step up publicly in a big way to be the exclusive launch customer for N3, N2, N1 point whatever, then I think that'll be interesting. I would definitely raise my eyebrow at that.

Semiconductor Supply Chains

Fin / Luca

So turning themes again, I'm curious to talk a bit about semiconductor supply chains. A lot of your most popular videos often take the form of why country X didn't achieve Y. This really points at the difficulties of actually being able to do technological transfers, even if governments are willing to spend a lot on this. This seems to come up a lot in the semiconductor industry around lithography and Indian semiconductors. I think you also had a video on that. I'm curious if you're able to draw out any general lessons or a general theory as to why tech transfers are so hard, especially in semiconductors.

Jon

I think the hardest part about what makes the semiconductor industry hard for new entrants into the market is that, one, you need someone to teach you, and two, it's very competitive. Three, everyone's moving, right? So two and three are very important. Right now, Japan's trying to get back into leading-edge semiconductor production. They have this company called Rapidus; they're trying to do two nanometers. By the time their two nanometers ramp up in 2027, TSMC, Samsung, and Intel will have already gotten there for at least one to two years. It's trailing; it's basically old, right? If it's old, then suddenly your revenue potential declines because there's less demand for people to spend the hundreds of millions of dollars to develop a chip for that. You've sunk a lot of cash into building capacity for something that never comes. That's why a lot of these tech transfers fail—not necessarily because they didn't do it well enough, but because they couldn't get a customer. In some ways, it's trying to do tech transfer on hard mode because you decide that you're just going to find the hardest thing you can possibly do. I find that interesting. If India wants to make semiconductors and they're starting at 28 nanometers, yes, I can argue you want to build your way up and learn these things. But you've got to ask: Who's going to buy these chips that you're learning to make, which will probably be pretty bad? The general theory is that a lot of these fail simply because the product's not good, you're not learning it fast enough, and you need to work harder than everyone else. In some ways, that's very difficult.

Fin / Luca

Yeah, I guess when someone's already making it cheaper before you learn how to make it really well, then where are your incentives to get good at learning it? Or at least how do you get enough runway?

Jon

How do you catch up to someone who's running two times as fast as you?

Fin / Luca

It seems to matter as well that this industry appears to exhibit a lot of economies of scale, where just being bigger gives you an innate advantage for doing a lot of capital expenditure.

Jon

Yeah. TSMC's biggest value is its customers. The fact that you scale gives you this immense scale, and they're paying you to learn how to make their product better. I think that's incredibly valuable.

Fin / Luca

Here's a question I maybe should have asked earlier, but when we're talking about the semiconductor supply chain, what are the big components in that chain? What are the major stages from, I guess, design to I bought a laptop?

Jon

Oh, that is a big question.

Fin / Luca

You can dumb it down.

Jon

Yeah. You have the chip IP where the person makes the design. You have the chip manufacturer, right? That's the fab. They're like the restaurant that turns your recipe into food.

Fin / Luca

Then you have—this is TSMC.

Jon

TSMC, Samsung, STMicro, all these other companies. They, in turn, have their own supply chain. We'll leave that offshoot for later. After the chip comes out of the fab, it goes into packaging where you actually need to put it into the component. Then it goes to the end user, which is generally not the consumer but a supplier maker, like Apple or HP. If you're talking about supply chains within the fab side, packaging also has its own

supply chain. Design has its own supply chain. There are supply chains on top of supply chains on top of supply chains. It's ridiculous.

Fin / Luca

It was naive of me to ask about the supply chain as if it's a 1-2-3 process. But that's a useful overview. It suggests the question of which chunks in this sprawling tree of dependencies are most difficult to replicate. EUV comes to mind as an example lots of people talk about, but are there any others?

Jon

I think lithography has a really high value, but EUV itself is not a very… There are questions about whether it's economical. There are questions in and of itself whether this thing is important. Lithography itself, in general, has a lot of value because it's the printing press. It's the way that we can turn. We're not writing the Bible; we're printing the Bible, right? That's so important because that's the key to scale. That's the key to making a lot of chips. Once you make a lot of chips, you win. Get there first with the most chips. So that's the key point, in my opinion, the lithography point. That also, lithography is just because it's the biggest single point doesn't mean it's the most important point. It doesn't mean it's the majority of the point. Something like 40 to 50% of the equipment value within the system. You also have etch, ion implantation, deposition; you have all these other things that need to work together in perfect harmony at incredible speed to make it work.

Fin / Luca

Okay, got it. And then just quickly on that, when we're thinking of lithography, people often mention ASML, based in the Netherlands, as an example of a company that has more or less monopolized that part of the supply chain. Similarly, TSMC—you're raising your finger. So what am I getting wrong?

Jon

TSMC has a monopoly in EUV machines, but Canon and Nikon have a strong share in DUV machines as well. I always need to make that clarification. ASML does not have a monopoly in DUV machines.

Fin / Luca

EUV specifically, I should have said. TSMC as well is another example of a company that has an enormous share of their part of the chain, which is actually making the chips.

Jon

Yeah.

Fin / Luca

Especially on the advanced side, right? I'm asking, I guess, like as a question.

Jon

Only on the advanced side. TSMC makes the most advanced chips and therefore captures the most value. But do not equate value with volume because there are far more companies out there making far more chips, but they're low quality—or not low quality; they're low-value chips. These are the two-cent chips that go into my microphone or something.

Fin / Luca

Got it. Thanks. That's also very useful. I'm curious if there are other examples of bottlenecks or quasi-monopolies like this in other parts of the world, other than Taiwan or the Netherlands.

Jon

Japan has a lot of them. Japan makes a substantial portion of the resins, the photoresist. The photoresist is the thing that goes on top of the wafer. They make a lot of the masks. They make the mask blanks. They make a lot of the other stuff. I think you have some small European companies that help make the writers for the masks. When you make the mask, which is the template for making the chip design, you need to write that yourself. There's only two or three companies in the world that do that—one of them in Austria and the other ones in Japan. A lot of these random suppliers have emerged. This has happened over the span of 20 to 30 years when the industry exploded to its current size, and the biggest companies shrank their supply chains and basically snapped off all the competitors. Now we have the entire sub-semiconductor supply chain industry, essentially like you have a supplier, maybe an A-minus supplier, and then a bunch of C suppliers or D suppliers.

Fin / Luca

One reason to be interested in these monopolies is that from a public policy perspective, it seems that both the US and China are really interested in where there might be potential choke points when it comes to the very leading-edge types of chips. We saw that most recently with some of the legislation that the US passed. I'm curious, especially on China's side, whether there are any monopolies or so-called choke points that they might have. Maybe to give further context, we have seen recently that they applied some export controls regarding the supply of rare earth materials. But it's like a very open question, right? Of whether that's actually a choke point or whether you can find other suppliers there.

I think rare earths is probably what they're seeing as one of their most critical resources. It's not just rare earths; they also have strong monopolies, or near monopolies, in a lot of other minerals like transition metals and certain other materials. They do really well with basically any sort of refined rare element that is not economical or environmentally healthy to process; they do it, and they are the ones that can do it the best. Think about lithium and all these other processes related to batteries. I think they're really good at that, and I think that's tied to their advantages in manufacturing. Calling it a choke point is tough because, in the broadest definition of the word, I would call it a choke point. You have these supply chains that can be disrupted. But you could also argue that these are commodities that only got so cheap because someone was subsidizing them. We had the capacity to make these things before, and there's an argument to be made that maybe it's easier to do it again now. You can also argue the same thing about chips. It's really tough to compare these two, but they're clearly being used as economic weapons against one another—not necessarily outright, but they're putting gates around the doors to make sure that when the time comes, they can shut the gates closed.

Fin / Luca

I'm curious to expand a bit on what you mentioned here with economic weaponization or, more broadly, the geopolitics. You live in Taiwan yourself, and I'm really curious about the public reaction to the US's CHIPS Act and export controls, which, on the one hand, seem to target China and have a bunch of security concerns around them. On the other hand, a lot of the language is about reassuring or trying to get TSMC to produce more of its advanced nodes in the US or even help other companies such as Intel reach the leading edge. What is the overall reaction there? In particular, what is the politics in Taiwan around this?

Jon

I think it's interesting to look at it right now from a political basis. There is an acknowledgment that there are certain political lines in Taiwan-Chinese relations that

cannot be crossed. The people have an awareness of that and look nervously at situations where things happen that cause bad outcomes for everyone, and suddenly people start thinking about the unthinkable. I don't think the people here want to have conflict, but I also believe that TSMC is one of the most valuable assets—not just as a geopolitical chip, because that's definitely something we should take out of the equation. TSMC is not a geopolitical check; in any conflict, it's the first thing to get blown up. But they see it as a point of national pride. When they see discussions about re-onshoring, they think of it as the Americans not working as hard. There's a lot of concern, and I think there are cultural differences between these two that are curious. There's a lot of talk right now, particularly as they ramp up for the political election, about America stealing TSMC. For me, that's a relevant concern on Taiwan's side.

Fin / Luca

Does that get mentioned in, for instance, political debates or discussions in the run-up to the election?

Jon

Right now, I think the biggest talk will be about TSMC, but TSMC doesn't like that, so they probably would say no. The big thing is what to do about China and whether we need to be softer towards China or turn away from America a little bit. There's talk about that, but I'm not a political analyst on this. This is just my personal observation from what I see on television.

Fin / Luca

Yeah, for sure.

Jon

That's what people are worried about right now because they see these aggressions, and I think it gets a lot of people worried.

Fin / Luca

One follow-up I have specifically around fabs and the CHIPS Act is that the CHIPS Act seemed to have a lot of business incentives around subsidies or tax breaks to help make it more attractive to build fabs in the US. I'm curious to what extent Taiwan is able to compete with that or do a tit-for-tat with that. I've seen there being a new tax break around R&D expenses that seemed to be a pretty direct response to the CHIPS Act.

From a government perspective, how is that feasible both in terms of economic costs and in terms of geopolitics or relations with the US?

Jon

The number one thing that Taiwan's government can do for TSMC or all the other companies on the island is to make it easier for them to do business. That's the advantage they have because TSMC doesn't really pay that much tax anyway. The big thing that TSMC asks when they go to the government isn't for more tax breaks; they ask, "Can you get me more power? Can you get me more water? Can you help me clear this land out?" That's the sort of stuff that TSMC finds a lot easier to do in Taiwan as opposed to America. In America, they have to fight for land and other resources, whereas in Taiwan, TSMC has a lot more support and it's easier for them to get what they need. That's why they can move so fast on this, and I think that's the sort of support that a government can provide that doesn't show up on a tax break. That's what I think matters more.

Fin / Luca

One last question I have is around how some of this geopolitics affects the pace of progress or the increase in semiconductors. We've talked a lot about how complicated and international the supply chain is. One reaction would be that if governments are politicizing or even weaponizing parts of this supply chain, you would expect overall progress to slow down. But then, to counterweight this, you can also see the huge amounts of money being thrown at this, both from China and the US as well as other countries. Do you expect this recent politicization to, on net, slow down the progress in hardware or to increase it?

Jon

It's tough. One leads into the other, right? If the government starts talking up this sort of politicization, then there's going to be less money in it, right? But I would say that right now, the way the system still works—my understanding is that if you just look at ChatGPT, AI was not a big thing before November 2022. Once ChatGPT came in, all the money flowed in. And once the money flowed in, stuff started happening. So far, if there's money, then it will happen, but that money can vanish at any second. It's very precarious.

Superconductors

Fin / Luca

Okay, can we talk about superconductors?

Jon

Yes.

Fin / Luca

I mean, here's an obvious question: What exactly is a superconductor? Beyond just a thing that can conduct extremely well?

Jon

You know, that's kind of it, right? When they were doing these low-temperature tests and trying to analyze what would happen to conductivity, when an electrical current passes through an object—a wire or anything—it's going to hit interference, which causes it to slow down. But then, when you have a situation where it's very cold—at super low temperatures—certain materials can create pathways for the electrons to phase through the lattice crystal of the atoms. Essentially, you have no resistance; it's as if they have a fast pass on the freeway—they just burn right through. That's what's so cool about superconductors—literally cool! Haha. That's the holy grail because you see a lot of benefits on the power side; you're not losing power to resistance, and a lot of other things become more technically feasible, and perhaps economically feasible. Superconductivity has this mystical property that almost feels like a violation of the laws of nature.

Fin / Luca

To be clear, this isn't just an arbitrary category of materials with especially low resistance, but rather this is…

Jon

Zero resistance.

Fin / Luca

There's a kind of phase difference, and it's literally zero resistance.

Jon

It's literally zero.

Fin / Luca

Yeah. Got it.

Jon

And then…

Fin / Luca

Yeah. So my understanding is that there is a kind of frontier of pressure and temperature where typically, to make a superconductor from a given material, you'll need either an extremely low temperature or an extremely high pressure, or some combination of both.

Jon

It's tough to say, yeah.

Fin / Luca

Okay. And then I want to say that the history of pushing along this frontier is that we've figured out a way to use the same pressure but a slightly higher temperature because we've figured out some new material, and that frontier is just slowly creeping up towards ambient pressure and room temperature. Is that roughly the whole story?

Jon

Yeah. I don't think that's entirely accurate. It's tough. We don't know whether pressure is necessary; there seems to be a debate. The papers get more recent, and the more recent the papers get, the less they may mean anything. Right now, it seems like high pressure can help create superconductivity, but there are a lot of other people saying it's not. We don't know. The relevant factors for superconductivity are current density, magnetic field, and temperature. Temperature is important, but it's not the only one. They found that if you freeze something but then give it a high temperature and high magnetic field—which is exactly what happens when you start passing electricity through a wire—it will deactivate itself. Those are the three things that really matter. They found over the years that you can modify a material, like a certain element. If you create these crazy alloys—insane alloys—you can kind of work with that. That's the way we're talking about; you would lower points on some aspects to maybe get more points on another.

Fin / Luca

How long have people been trying to find new materials or new ways to modify elements to push along those different measures? How long has this been something people have been interested in?

Jon

It's been about 100 years, right? That's the crazy thing about it. No one really knows why. People don't have a good understanding of why superconductivity works, and that's why it's been so hard to find new categories of it. There's some fascination with it. I know I'm kind of geeking out about it, but it is weird. This stuff is weird.

Jon

This is a safe space for geeking out.

Fin / Luca

What then is the particular relevance of LK99? You've mentioned that there's been about a century of exploration. What is particularly special about this or could be potentially?

Jon

What these people are claiming is that this is it; this is the holy grail. After a hundred years of exploration, they've finally ended up here, essentially. You have a room-temperature superconductor that you can take out, and it will superconduct—probably with some conditions, but it will superconduct at room temperature. Suddenly, everyone is dreaming of floating hoverboards, maglev, and infinite power. The reality is much more subtle, and there's going to be a lot of industrial work that needs to be done. Almost like there have to be trade-offs made. All of that will need to be considered, but that's why everyone's gotten so excited: because superconductivity is not only a technical possibility but also partly wrapped up in a utopian dream.

Fin / Luca

Do you want to talk about those subtleties a little bit? Suppose we do get an ambient pressure, room-temperature superconductor—maybe it's LK99, maybe something else. What happens next, and what are the potential applications?

Jon

It's tough because these are compromised materials. If you think about it, I'm finishing this video on interconnects. I was thinking, okay, we can throw a superconductor into an interconnect, and suddenly we have no resistance. These chips, like a GPU, will no longer consume any power or will consume 80% less power, right? But it's actually a lot more subtle than that because the current interconnects in these machines are copper. Copper is great not only because it has low resistivity but also because it's easy to work with. We understand how it works, can predict its behavior, and can build around it. You can deposit it very evenly. That's not the case with these crazy Type II superconductors. A single-element superconductor can never be used for interconnects in an integrated circuit. You can't use superconducting copper, even if you were to freeze it. So then you need to start creating these super exotic materials.

Fin / Luca

Why can't we use these single-element superconductors?

Jon

Because once you push a current through, it generates a magnetic field, and that magnetic field essentially turns off its conductivity. It violates one of the three goals: current density, magnetic field, and temperature. Even if you have the temperature, if you exceed the boundaries for current or magnetic field, the whole thing turns off; it collapses.

Fin / Luca

Okay, makes sense. You mentioned Type II superconductors. Is there a Type I?

Jon

Type I would be like lead. Type I would be simple alloys, single elements. Mercury is a Type I superconductor because it's just mercury by itself. I've been immersed in this for the last week and a half. It's insane.

Jon

We'll also link to your videos—both the one you've already done on the history of semiconductors and the ones that might still be in the pipeline.

Fin / Luca

Yeah, I'm keen to nerd out on semiconductors. When you were doing the research for this video, did you have any conceptions that you realized were wrong? Or did you learn any details that seem very relevant for understanding how this pans out?

Yeah, it's so crazy. When I went into this LK99 thing, I thought, this is the holy grail. We have achieved eternal life, we have hoverboards now and all that. But then you learn more about how this works, and you start reading about it. Man, this stuff isn't that great actually. The key point is that a superconductor, in some ways, is like a scientific anomaly, right? And scientific anomalies are really hard to turn into commercial products. For example, in 1986, they discovered something called the 1, 2, 3 superconductor. They are called high temperature in the sense that their transition temperature, where they move into superconductivity, is higher than anything else seen before. That's good, right? But then you have the same dreams, with people saying, oh, we now have infinite power, we have floating maglev; it's going to change everything. As it turned out, they found out that these 1, 2, 3 superconductors are simply scientific curiosities. You can't use them for anything because, well, you can eventually make wires out of them, but it took years to do it. These are very strange materials. You can even think of them as more like waffles; they're like layer cakes of different materials. The superconducting happens only in specific layers, so you have to do all this crazy materials engineering to make it happen. And then you still need to do the freezing. The way people thought this would work out, the story didn't turn out that way. People thought they would be able to make these amazing superconducting wires to replace MRI machines, right? Because MRIs are the biggest commercial use for superconductors, and it's only $5 billion. The iPhone makes more in a month or two than all of the superconducting industry makes every year. But then people found out that these new superconductors from 1986 can't even be made into wire that easily. They're weird. That's the only thing I can say: they're weird. It ends up causing all these issues, and in the end, engineers were just like, you know what? We're sticking with the old stuff. It's 50 years old, but we know how to use it. That's a very clear story of how the commercialization of scientific anomalies is very difficult.

Fin / Luca

Can you maybe speak about, given all of these challenges, what industries should still be paying attention to this if it could work? You mentioned some of the challenges in using this with semiconductors. I can also imagine when it comes to decarbonizing the grid, there's a lot of emphasis on transmission lines, so being able to get energy from point A, where it's generated, to point B. That seems particularly relevant. I know fusion has been mentioned, you mentioned hoverboards, and other things as well, but what kind

of sectors, if they can make it work, would you expect to be more promising, or which sectors should be paying attention right now?

Jon

It'd be very interesting to see what these things will do to wires, right? I think it's just going to end up being wires, like power transmission. The problem is that these wires, if we need a flow, depend on how LK99 eventually works. We don't know how this thing works, right? We don't know how any room temperature superconductor will eventually work, but there are fundamental restrictions on how this thing can eventually work. What I found out was that if the current density limit is not particularly high, then if the current you need to flow through this is large, you need to make this wire really thick. It's going to be a thick wire, so then you've got materials engineering problems. How are you going to make enough LK99 to do a hundred-meter wire that's thick? It's very difficult. It might end up being yet another scientific anomaly that will win someone a Nobel Prize, which is good because it seems like the best way to get a Nobel is to do superconductor stuff.

Fin / Luca

Like we have no idea how to commercialize this stuff; there's a long way to go. Yeah, graphene just came to mind as maybe a kind of comparison. We understand its properties quite well; it has all these applications and promise for batteries and stuff. We don't have graphene batteries at scale, presumably just because it's very hard to figure out how to produce it at scale and commercialize it in different ways. Another analogy, which I'm curious if you know anything about, is semiconductors—not as shorthand for computing hardware, but actual semiconducting materials. That's a thing that people had to discover, and it turned out to be really useful. Was there a period where there were these slightly mysterious materials we didn't fully understand and therefore couldn't commercialize them, and then gradually we did come to understand them? What's the story with semiconductors?

Jon

That's a very interesting question, and I think it's actually something I find really fascinating. Within the semiconductor industry, there are two parts to this question. I'll hit them both separately. First, when they first discovered semiconductors as we know them now, there was a prior existence of materials for this, like vacuum tubes. What they did was use new materials, like silicon and germanium, to replace an existing machine, which was the tube. So, if you apply that to superconductors today, what is superconductors going to do now that already doesn't exist? It's not going to create

new magic; what it needs to do is figure out what this thing is going to do better than what already exists now. That's the big question for a lot of people. Separately, there's another part of your question where you mentioned discovery. The funny thing is that in the semiconductor industry, someone creates something, and everyone's like, that's nice, and then it just sits for like 30 years. Then they come back to it and say, we're going to use this now. This happens all the time in the packaging industry. Chiplets are sexy now, but chiplets have been around since the 1970s. They created this multi-chip module, and everyone was like, that's nice. But then no one bought it, no one cared, and it failed. Then they revitalized it like a zombie, and now it's the sexiest thing ever. That sort of stuff happens all the time. In some ways, it's ironic; it happens all the time, and it doesn't happen ever.

Fin / Luca

Yeah, interesting. So maybe there's this kind of delay. People had computers built with vacuum tubes, and then someone thinks, hey, remember those semiconducting materials we know about? Maybe we could swap that out and get a bunch more transistors on the same.

Jon

It's the magic of innovation.

Jon

Right. Indeed.

Fin / Luca

Briefly, before wrapping up this interview, I want to talk a bit about what it's like being an online content creator. You produce so many videos on so many topics at such a regular pace. I'm just wondering if you can walk us through your pipeline. Where do you get ideas? How do you learn about topics? What's it like?

Jon

Yeah, I'm very optimized now. A couple of years ago, I went to live with my grandma, and I took a sabbatical from work. I told myself I wasn't that big at that time; I think I had around 800 to 1,000 subscribers. I said, I'm just going to make one video a day, doesn't matter what. I made one video a day for 50 to 60 days, two months. That's what I did, and it could be anything. At that time, I really optimized the flow: I have a list of

topics, basically just text files. I write everything in text files, everything in Markdown. Every word, every video I do has a script, and I script every part of the work. The more you script it, the less editing you have to do afterward. I build the video and write the script at the same time, so collecting assets, making graphs, and charts all happens at the same time. By the time the script is done, I set it aside to let it mellow, and then when it's time to edit and record and create assets, that all happens in the span of three hours. That's basically the flow. It's so tightly optimized that I don't even know if I could teach it to anyone. It's so crazy. I have a job, so right now, I write scripts in the morning, go to work, go home, and then I edit and record, and then I go to bed at hopefully 11:30 to midnight. Then I wake up again.

Fin / Luca

And where do you get information from? A lot of your videos have this feeling of being both about history and very textbook-like and technical. Do you literally read textbooks on this stuff, or do you have any resources?

Jon

I do. I live next to a library, so I walk over to the library, open a real paper book, and read it. I think, you know, I hated textbooks when I was in college, but now I'm much older, and I'm reading textbooks again. I feel old, but the older the textbook, the better written I tend to find them. A lot of people say, "Hey, John, you have a textbook feeling," and that's exactly where it comes from: textbooks.

Fin / Luca

If you're making, let's say, a video on superconductors, you need to do some research. I presume you do some high-level Google research. What next? Where do you look to? How do you find the textbook to go and open up and read?

Jon

Well, they have something called the Dewey Decimal System.

Fin / Luca

Thanks.

Jon

You go in, and when I was a kid, I loved to read books. I was always searching for books in libraries. I'm very methodical; I go in and walk from left to right like a scanner. I pick the biggest textbook first, read the first three or five chapters, the introduction chapters. Then I'm like, okay, can I understand what this guy's saying? No? Next textbook. I'm very fortunate because a lot of these textbooks are open as well, so you can read them digitally and do a control F, which is very fortunate. For superconductors, I read about 30 papers that I downloaded from Google Scholar and three textbooks. I also have, yeah, so basically, I just cram that in. Luckily, it was a weekend, so I had the weekend to myself to cram through the stuff.

Fin / Luca

Do you have some elaborate note-taking system or just a little text file?

Jon

Google Docs, lots of Google Docs. I'm just rewriting, writing, writing, writing. Then I go through my notes and say, okay, I take bits; it's almost like blocks. First, you move the big blocks around, and then you start moving the little blocks within the writing to eventually end up with something that feels coherent. Then you do a final write-through to complete it.

Fin / Luca

What's the topic that you wish you knew more about? If you're looking ahead to the next few videos you're going to make, planning on doing some deep dives, what comes to mind?

Jon

I've been really interested in nuclear weapons recently. I don't know if it's because of Oppenheimer or anything, but I watched that movie, and I was really moved by it. I thought a lot about how, after the first detonation, the Soviet Union and all this set off basically an arms race by all the other countries to try to achieve nuclear weapons while, at the same time, the United States was trying hard to prevent knowledge from spreading. I'm really fascinated; I want to learn more about this because it's very difficult to actually find information on it. How did all these other countries get these bombs? I think that's something that's really fascinating.

Fin / Luca

Kind of catch up.

Jon

Yeah. I did a video about China and how China got its nuclear weapon. I'm very fortunate because I found a book at the Stanford repository that I bought. But a lot of these other countries are going to be much more challenging.

Fin / Luca

Maybe sticking on this question, for context, a lot of our listeners are early career researchers, really hunting for things they could spend, say, three to six months on. Outside of nuclear weapons, is there anything else you'd want to flag as, hey, I think this would be really cool and is currently underexplored? If you've got six months, do a deep dive on it.

Jon

I'm really fascinated right now by other parts of the semiconductor supply chain other than lithography, right? So etch, deposition, ion implantation—many of these are kind of legacy tools that need a lot more investment. One of the things I'm really fascinated by, and we'll have a video about it soon, is atomic layer deposition. That's one of the most exciting tools in the semiconductor industry right now outside of lithography. I think it and its analog, atomic layer etch, are going to be really big for semiconductors in the future, and there's going to be a lot of investment in it. If I were doing that research, that would be really good reading to learn.

Fin / Luca

If it's possible to give a quick answer, can you say what etch is?

Jon

After lithography happens, all lithography does is transfer the chip pattern from the plate to the wafer. After that, you still need to cut that design into the wafer. That's what etch does. Etch basically imprints that into the silicate or the metal layers.

Fin / Luca

All right, let's do some final questions. One question we ask everyone is whether you can recommend three resources—papers, textbooks, books—for someone listening to this who just wants to find out more about anything you've talked about?

Jon

That is a very good question. There is a great textbook by Bo Lojek called *The History of Semiconductor Engineering*. I think anyone who likes what I talk about should pick that up. There is another one called *Tiger Technology: The Creation of a Semiconductor Industry in East Asia*. I think that's a great book; it's actually publicly available, so you can just download it. Then there is another one that I read all the time: *The Handbook of East Asian Entrepreneurship*. That one is great because it tells you how all of these other companies outside of the U.S.—everyone's tired of hearing about Google. I think it's very fascinating to hear about these Chinese or East Asian companies in Taiwan, South Korea, Hong Kong, but also Malaysia and Southeast Asia. It's very fascinating.

Yeah, these are great textbook recommendations to help people get a foundational knowledge. I'm curious if there are any online resources, like blogs or Substacks, that you'd recommend for people to stay up to date with semiconductor news.

Jon

I think Doug's fabricated knowledge is really good. I also read a lot of SemiWiki; SemiWiki is pretty okay. There's SemiEngineering. A lot of this stuff doesn't move that fast, so you don't really have to worry about it. It's not like breaking news, not like exactly, you know, LK99 stuff. If it's important, it'll pop up to you eventually.

Fin / Luca

Nice. Awesome. Yeah, sorry to interrupt.

Jon

Yeah,

Fin / Luca

no worries. That's good advice. And then just lastly, we will of course link to your newsletter and your YouTube channel. But just, I guess verbally, where can people find you and your work online?

Jon

Just find me on Asianometry, youtube.com, search Asianometry. I'm the deer. You can just watch a couple of videos, and I hope you enjoy them. You can also shoot me an email at hello@asianometry.com and say hello. I try to be accessible.

Fin / Luca

What's the story behind the deer, by the way?

Jon

The deer… So one day, a long time ago, I was creating a YouTube channel for my mom to show her what I was doing in Taiwan. My mom complained to me during our last trip to Japan that I didn't tell her anything. So I created this YouTube channel to do videos of hiking and stuff I was doing in Taiwan. I picked Asianometry because I like trigonometry and I am Asian. When they asked if I needed a profile picture, I said yes. I looked into my phone, and the last picture I took—because I had just come back from vacation—was from our trip to Nara that my sister just sent me of a deer with me in the background. So if you look at that YouTube profile picture, I'm in the back. That's me.

Fin / Luca

That's so cool. I had no idea. That's a really fun little Easter egg. Well, yeah, fantastic. Thanks so much for this interview. And thanks so much, John, for coming on to the show.

Jon

No problem. Thank you so much for having me.

Jon

That was Jon Y from Asianometry on compute trends in artificial intelligence. As always, if you want to learn more, you can read the write-up at hearthisidea.com forward slash episodes forward slash Asianometry. There you'll find links to all the papers and books referenced throughout our interview, plus a whole lot more. If you enjoyed this podcast and find it valuable, one of the best ways to help us out is to write a review on whatever platform you're listening to. You can also give us a shout-out on Twitter; we're at HearThisIdea. We do have a short feedback survey, which should only take you somewhere between five to ten minutes to fill out. We read every submission, and as a thank you, you'll also get a free book from us. A big thanks, as always, to our producer, Jason, for editing these episodes, and thanks very much to you for listening.

← Back to all episodes Leave feedback ↗