Dwarkesh Podcast #81 - Dylan Patel & Jon (Asianometry) – How the Semiconductor Industry Actually Works

Published - October 2, 2024

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today, I'm chatting with Dylan Patel, who runs SemiAnalysis, and Jon, who runs the Asianometry YouTube channel.

Dylan Patel
Does he have a last name?

Jon Y
No, I do not. No, just kidding. Jon Y.

Dylan Patel
Why is it only one letter?

Dwarkesh Patel
Because Y is the best letter.

Dylan Patel
Why is your face covered?

Jon Y
Why not?

Dwarkesh Patel
No, seriously why is it covered? Because I'm afraid of looking at myself getting older and fatter over the years.

Dylan Patel
But seriously, it's for anonymity, right?

Jon Y
Anonymity, yeah.

Dwarkesh Patel
By the way, do you know what Dylan's middle name is?

Jon Y
Actually, no. I don't know.

Dylan Patel
What's my father's name?

Jon Y
I'm not going to say it, but I remember.

Dylan Patel
You could say it. It's fine.

Jon Y
Sanjay?

Dylan Patel
Yes. What's his middle name?

Jon Y
Sanjay?

Dwarkesh Patel
That's right. So I'm Dwarkesh Sanjay Patel. He's Dylan Sanjay Patel. It's like literally my white name.

Dylan Patel
It's unfortunate my parents decided between my older brother and me to give me a white name. I could have been Dwarkesh Sanjay. You know how amazing it would have been if we had the same name? Butterfly effect and all, it probably would've turned out the same way, but...

Dwarkesh Patel
Maybe it would have been even closer. We would have met each other sooner, you know? Who else would be named Dwarkesh Sanjay Patel in the world? Alright here's my first question. If you're Xi Jinping and you're scaling-pilled, what is it that you do?

Dylan Patel
Don't answer that question, Jon, that's bad for AI safety.

Jon Y
I would basically be contacting every Chinese national with family back home and saying, "I want information. I want to know your recipes. I want to know suppliers."

Dwarkesh Patel
Lab foreigners or hardware foreigners?

Jon Y
Everyone.

Dylan Patel
Honey potting OpenAI?

Jon Y
This is totally off-cycle, off the reservation, but I was doing a video about Yugoslavia's nuclear weapons program. It started with absolutely nothing. One guy from Paris showed up. He knew a little bit about making atomic nuclear weapons. He was like, "Okay, well, do I need help?" Then the state's secret police is like, "I will get you everything."

For a span of like four years, they basically drew up a list. "What do you need? What do you want? What are you going to do? What is it going to be for?" And the state police just got everything. If I were running a country and I needed to catch up on that, that's the sort of thing that I would be doing.

Dwarkesh Patel
Okay, let's talk about espionage. What is the most valuable piece, if you could have this blueprint, this one megabyte of information? Do you want it from TSMC? Do you want it from NVIDIA? Do you want it from OpenAI? What is the first thing you would try to steal?

Dylan Patel
You have to stack every layer, right? The beautiful thing about AI is that because it's growing so fast, every layer is being stressed to an incredible degree. Of course, China has been hacking ASML for over five years and ASML is kind of like, "Oh, it's fine." The Dutch government's really pissed off, but it's fine. They already have those files in my view. It's just a very difficult thing to build.

The same applies for fab recipes. They can poach Taiwanese nationals. It's not that difficult because TSMC employees do not make absurd amounts of money. You can just poach them and give them a much better life and they have. A lot of SMIC's employees are TSMC, Taiwanese nationals, especially a lot of the really good ones high up.

You go up the next layers of the stack. Of course, there are tons of model secrets. But how many of those model secrets do you not already have and just haven't deployed or implemented or organized? That's the one thing I would say. China just clearly is still not scale-pilled in my view.

Dwarkesh Patel
If you could hire these people, it would probably be worth a lot to you because you're building a fab that's worth tens of billions of dollars. This talent knows a lot. How often do they get poached? Do they get poached by foreign adversaries or do they just get poached by other companies within the same industry but in the same country? Why doesn't that drive up their wages?

Jon Y
It's because it's very compartmentalized. Back in the 2000s, before SMIC got big, it was actually much more open and more flat. After that, after Liang Mong Song and after all the Samsung issues and after SMIC's rise, you literally saw—

Dylan Patel
You should tell that story, actually, about the TSMC guy that went to Samsung and SMIC and all that. I think you should tell that story.

Jon Y
There are two stories. There's a guy who ran a semiconductor company in Taiwan called Worldwide Semiconductor. This guy, Richard Chang, was very religious. All the TSMC people are pretty religious. He particularly was very fervent and wanted to bring religion to China. So after he sold his company to TSMC—which was a huge coup for TSMC—he worked there for about eight or nine months and then went back to China. Back then, the relations between China and Taiwan were much more different. So he goes over to Shanghai and they say, "We'll give you a bunch of money." Richard Chang basically recruits a whole conga line of Taiwanese who just get on the plane and fly over. Generally that's actually true of a lot of the acceleration points within China's semiconductor industry. It's from talent flowing from Taiwan.

The second story is about Liang Mong Song. Liang Mong Song is a nut. I've not met him. I've met people who worked with him and they say he is a nut. He's probably on the spectrum. He doesn't care about people, business, or anything. He wants to take it to the limit. That's the only thing he cares about. He worked at TSMC, a literal genius with 300 patents or whatever, 285. He works his way all the way to the top tier and then one day he loses out on some power game within TSMC and gets demoted.

Dylan Patel
He was like the head of R&D or something right?

Jon Y
He was like one of the top R&D people. He was in like second or third place.

Dylan Patel
It was for the head of R&D position, basically.

Jon Y
Correct, it was for the head of R&D position. He's like, "I can't deal with this." He goes to Samsung and steals a bunch of talent from TSMC. Literally, again, it's a conga line. At some point, some of these people were getting paid more than the Samsung chairman, which is not really comparable...

Dylan Patel
Isn't the Samsung chairman usually like part of the family that owns Samsung?

Jon Y
Correctamundo.

Dylan Patel
Okay, yeah so it's kind of irrelevant.

Jon Y
So he goes over there and says, "We will make Samsung into this monster. Forget everything. Forget all the stuff you've been trying to do incrementally. Toss that out. We are going to the leading edge and that is it." They go to the leading edge. They win a big portion of Apple's business back from TSMC.

And then at TSMC, Morris Chang is like, "I'm not letting this happen." That guy is toxic to work for as well but also goddamned brilliant. He's also very good at motivating people. He sets up what is called the Nightingale Army. They split a bunch of people and they say, "You are working R&D night shift. There is no rest at the TSMC fab. As you go in, there will be a day shift going out." They called it "burning your liver." In Taiwan, they say as you get old and as you work, you're sacrificing your liver. They called it the liver buster.

They basically did this Nightingale Army for a year or two years. They finished FinFET. They basically just blow away Samsung. At the same time, they sue Liang Mong Song directly for stealing trade secrets. Samsung basically separates from Liang Mong Song and Liang Mong Song went to SMIC.

Dylan Patel
So Samsung at one point was better than TSMC. Then he goes to SMIC and SMIC caught up rapidly after.

Jon Y
Very rapid. That guy's a genius. That guy's a genius. I don't even know what to say about him. He's like 78 and he's beyond brilliant. He does not care about people.

Dwarkesh Patel
What does research to make the next process node look like? Is it just a matter of a hundred researchers going in, they do the next n+1? The next morning, the next hundred researchers go in?

Jon Y
It's experiments. They have a recipe and that's what they do. Every TSMC recipe is the culmination of long years of research. It's highly secret. The idea is that you're going to look at one particular part of it and say, "Run an experiment. Is it better? Is it not? Is it better or not?" It's a thing like that.

Dylan Patel
It's basically a multivariable problem. Every single tool sequentially you're processing the whole thing. You turn knobs up and down on every single tool. You can increase the pressure on this one specific deposition tool.

Dwarkesh Patel
What are you trying to measure? Does it increase yield?

Dylan Patel
It's yield, it's performance, it's power. It's not just better or worse. It's a multivariable search space.

Dwarkesh Patel
What do these people know such that they can do this? Is it that they understand the chemistry and physics?

Dylan Patel
It's a lot of intuition, but yeah. It's PhDs in chemistry, PhDs in physics, PhDs in electrical engineering...

Jon Y
Brilliant geniuses.

Dylan Patel
They don't even know about the end chip a lot of times. It's like, "Oh, I am an etch engineer and all I focus on is how hydrogen fluoride etches this and that's all I know. If I do it at different pressures... If I do it at different temperatures... If I do it with a slightly different recipe of chemicals... It changes everything.

Jon Y
I remember someone told me this when I was speaking. How did America lose the ability to do this sort of thing? I'm talking about etch and hydrofluoric acid and all of that. He told me basically it's very master-apprentice. You know like in Star Wars with the Sith, there's only one, right? Master apprentice, master apprentice. It used to be that there is a master, there's an apprentice, and they pass on this secret knowledge. This guy knows nothing but etch, nothing but etch.

Over time, the apprentices stopped coming. In the end, the apprentices moved to Taiwan. That's the same way it's still run. Like you have NTHU, National Tsing Hua University. There's a bunch of masters. They teach apprentices, and they just pass this secret, sacred knowledge down.

Dwarkesh Patel
Who are the most AGI-pilled people in the supply chain?

Dylan Patel
I got to have my phone call with Colette right now.

Dwarkesh Patel
Okay, go for it. Could we mention to the podcast that NVIDIA is calling Dylan to update him on the earnings call?

Dylan Patel
Well, it's not exactly that, but...

Dwarkesh Patel
Go for it, go for it...
Dylan is back from his call with Jensen Huang.

Dylan Patel
It was not with Jensen, Jesus.

Dwarkesh Patel
What did they tell you, huh? What did they tell you about next year's earnings?

Dylan Patel
No, it was just color around like Hopper, Blackwell, and margins. It's quite boring stuff for most people, I think it's interesting though.

Dwarkesh Patel
I guess we could start talking about NVIDIA. You know what, before we do...

Dylan Patel
I think we should go back to China. There's a lot of points there.

Dwarkesh Patel
Alright, we covered the chips themselves. How do they get the 10 gigawatts data center up? What else do they need?

Dylan Patel
There is a true question of how decentralized do you go versus centralized. In the US, as far as labs and such, you have OpenAI, xAI, Anthropic. Microsoft has their own effort, Anthropic has their own efforts, despite having their partner. Then you have Meta. You also have all the interesting startups doing stuff. You go down the list and there's quite a decentralization of efforts. Today in China, it is still quite decentralized. It's not like, "Alibaba, Baidu, you are the champions." You have DeepSeek doing amazing stuff and it's like, "Who the hell are you? Does the government even support you?"

If you are Xi Jinping and scale-pilled, you must now centralize the compute resources. Because you have sanctions on how many NVIDIA GPUs you can get in now. They're still north of a million a year, even post-October last year sanctions. We still have more than a million H20s, and other Hopper GPUs getting in through other means but legally the H20s. On top of that, you have your domestic chips, but that's less than a million chips.

When you look at it, it's like, "Oh, well, we're still talking about a million chips." The scale of data centers people are training on today slash over the next six months is 100,000 GPUs. OpenAI, xAI, these are quite well documented and others. But in China, they have no individual system of that scale yet. Then the question is, "How do we get there?" No company has had the centralization push to have a cluster that large and train on it yet, at least publicly and well-known. The best models seem to be from a company that has got like 10,000 GPUs or 16,000 GPUs.

It's not quite as centralized as the US companies are and the US companies are quite decentralized. If you're Xi Jinping and you're scale-pilled, do you just say, "XYZ company is now in charge and every GPU goes to one place"? Then you don't have the same issues as in the US. In the US, we have a big problem with being able to build big enough data centers, being able to build substations and transformers and all this that are large enough in a dense area. China has no issue with that at all because their supply chain adds as much power as like half of Europe every year. It's some absurd statistic.

They're building transformer substations or building new power plants constantly. They have no problem with getting power density. You go look at Bitcoin mining. Around the Three Gorges Dam, at one point at least, there was like 10 gigawatts of Bitcoin mining estimated. We're talking about gigawatt data centers coming over in 2026 or 2027 in the US. This is an absurd scale relatively. We don't have gigawatt data centers ready. China could just build it in six months, I think, around the Three Gorges Dam or many other places. They have the ability to do the substations. They have the power generation capabilities. Everything can be done like a flip of a switch, but they haven't done it yet. Then they can centralize the chips like crazy. Right now they can be like "Oh, a million chips that NVIDIA's shipping in Q3 and Q4, the H20, let's just put them all in this one data center." They just haven't had that centralization effort yet.

Jon Y
You can argue that the more you centralize it, the more you start building this monstrous thing within the industry, you start getting attention to it. Suddenly, lo and behold, you have a little bit of a little worm in there. Suddenly while you're doing your big training run, "Oh, this GPU is off. Oh, this GPU... Oh no, oh no, oh no..."

Dylan Patel
I don't know if it's like that.

Dwarkesh Patel
Is that a Chinese accent by the way?

Dylan Patel
Just to be clear, Jon is East Asian. He's Chinese.

Jon Y
I'm of East Asian descent.

Dylan Patel
Half Taiwanese, half Chinese?

Jon Y
That is right.

Dylan Patel
But I don't know if that's as simple as that because training systems are like... Is it water gated? Firewalled? What is it called? Not firewalled. I don't know. There's a word for that. Where they're...

Jon Y
Air-gapped. I think they're Chinese-walled.

Dwarkesh Patel
You're going through all the four elements. Earth, water, fire!

Dylan Patel
If you're Xi Jinping and you're scale-pilled...

Dwarkesh Patel
You got to unite the four forces.

Dylan Patel
Fuck the airbenders. Fuck the firebenders. We got the Avatar. You have to build the Avatar. Okay. I think that's possible. The question is, “Does that slow down your research?” Do you crush people like DeepSeek who are clearly not being influenced by the government? and put some idiot...

Jon Y
You put an idiot bureaucrat at the top. Suddenly, he's all thinking about these politics. He's trying to deal with all these different things. Suddenly, you have a single point of failure, and that's bad.

Dylan Patel
On the flip side, there are obviously immense gains from being centralized because of the scaling laws. The flip side is compute efficiency which is obviously going to be hurt because you can't experiment and have different people lead and try their efforts as much if you're more centralized. There is a balancing act there.

Dwarkesh Patel
That is actually really interesting, the fact that they can centralize. I didn't think about this. Even if America as a whole is getting millions of GPUs a year, the fact that any one company is only getting hundreds of thousands or fewer means that there's no one person who can do a single training run as big in America as if China as a whole decides to do one together. The 10 gigawatts you mentioned near the Three Gorges Dam, how widespread is it? Is it a state? Is it like one wire? Would you do a sort of distributed training run?

Dylan Patel
It’s not just the dam itself, but also all of the coal. There's some nuclear reactors there as well I believe. Between all of that and renewables like solar and wind, in that region there is an absurd amount of concentrated power that could be built. I'm not saying it's like one button, but it's more like, "hey within X mile radius.” That's more of the correct way to frame it. That's how the labs are also framing it in the US.

Dwarkesh Patel
If they started right now, how long would it take to build the biggest AI data center in the world?

Dylan Patel
Actually the other thing is, could we notice it? I don't think so. With the amount of factories that are being spun up—the amount of other construction, manufacturing, etc. that's being built—a gigawatt is actually like a drop in the bucket. A gigawatt is not a lot of power. 10 gigawatts is not an absurd amount of power, it's okay. Yes, it's like hundreds of thousands of homes, millions of people. But you’ve got 1.4 billion people. You’ve got most of the world's

extremely energy intensive refining and rare earth refining and all these manufacturing industries here. It would be very easy to hide it.

It would be very easy to just shut down like... I think the largest aluminum mill in the world is there and it's north of 5 gigawatts alone. Could we tell if they stopped making aluminum there and instead started making AI there? I don't know if we could tell because they could also just easily spawn like 10 other aluminum mills to make up for the production and be fine. There's many ways for them to hide compute as well.

Dwarkesh Patel
To the extent that you could just take out a five gigawatt aluminum refining center and build a giant data center there, then I guess the way to control Chinese AI has to be the chips? Just walk me through how many chips they have now. How many will they have in the future? What will that be in comparison to the US and the rest of the world?

Dylan Patel
In the world we live in, they are not restricted at all in the physical infrastructure side of things in terms of power, data centers, etc. Their supply chain is built for that. It's pretty easy to pivot that. Whereas the US adds so little power each year and Europe loses power every year. The Western industry for power is non-existent in comparison.

On the flip side, "Western" manufacturing when you include Taiwan is way, way, way larger than China's, especially on leading-edge where China theoretically has—depending on the way you look at it—either zero or a very small percentage share. There you have equipment, wafer manufacturing, and then you have advanced packaging capacity.

Where can the US control China? Advanced packaging capacity is kind of shot because the vast majority... The largest advanced packaging company in the world was Hong Kong-headquartered. They just moved to Singapore, but that's effectively in a realm where the U.S. can't sanction it. A majority of these other companies are in similar places. Advanced packaging capacity is very hard. Advanced packaging is useful for stacking memory, stacking chips on co-ops, things like that.

Then the step down is wafer fabrication. There is immense capability to restrict China there. Despite the US making some sanctions, China in the most recent quarters was like 48% of ASML's revenue and like 45% of Applied Materials'. You just go down the list. Obviously it's not being controlled that effectively. But it could be on the equipment side of things.

The chip side of things is actually being controlled quite effectively, I think. Yes, there is shipping of GPUs through Singapore and Malaysia and other countries in Asia to China. But the amount you can smuggle is quite small. Then the sanctions have limited the chip

performance to a point where it's like, “You know, this is actually kind of fair.” But there is a problem with how everything is restricted.

You want to be able to restrict China from building their own domestic chip manufacturing industry that is better than what we ship them. You want to prevent them from having chips that are better than what we have. And then you want to prevent them from having AIs that are better. That’s the ultimate goal. If you read the restrictions, it’s very clear that it's about AI. Even in 2022, which is amazing, at least the Commerce Department was kind of AI-pilled. It was like, “You want to restrict them from having AIs better than us.”

So starting on the right end, it's like, “Okay, well, if you want to restrict them from having better AIs than us. you have to restrict chips. Okay. If you want to restrict them from having chips, you have to let them have at least some level of chip that is better than what they can build internally.” But currently the restrictions are flipped the other way. They can build better chips in China than what we restrict them from in terms of chips that NVIDIA or AMD or Intel can sell to China.

So there's sort of a problem there in that the equipment that is shipped can be used to build chips that are better than what the Western companies can actually ship them.

Dwarkesh Patel
Jon, Dylan seems to think the export controls are kind of a failure. Do you agree with him?

Jon Y
That is a very interesting question because I think it's like...

Dwarkesh Patel
Why thank you.

Dylan Patel
Dwarkesh, you're so good.

Jon Y
Yeah, Dwarkesh, you're the best. Failure is a tough word to say because what are we trying to achieve?

Dylan Patel
Let’s just take lithography. If your goal is to restrict China from building chips and you just boil it down to like, “Hey, lithography is 25-30% of making a chip. Cool, let's sanction lithography. Okay, where do we draw the line? Let me figure out where the line is.” If I'm a bureaucrat or lawyer at the Commerce Department or what have you, obviously I'm going to

go talk to ASML and ASML is going to tell me, “This is the line,” because they know this and this... there's some blending over.

Jon Y
They're looking at it like, “What's going to cost us the most money?”

Dylan Patel
Then they all constantly say, “If you restrict us, then China will have their own industry.” The way I like to look at it is that chip manufacturing is like 3D chess or like a massive jigsaw puzzle. If you take away one piece, China can be like, "Oh, that's the piece. Let's put it in." Currently year by year by year, they keep updating export restrictions ever since like 2018 or 2019 when Trump started and now Biden's accelerated them.

They haven't just taken a bat to the table and broken it. It's like, “Let's take one jigsaw puzzle piece out. Walk away. Oh shit. Let's take two more out. Oh shit.” You either have to go full “bat to the fricking table/wall” or chill out and let them do whatever they want. Because the alternative is everything is focused on this thing and they make that. Then now when you take out another two pieces they can be like, “Well, I have my domestic industry for this. I can also now make a domestic industry for these.” You go deeper into the tech tree or what have you.

Jon Y
It's art in the sense that there are technologies out there that can compensate. The belief that lithography is a linchpin within the system, it's not exactly true. At some point, if you keep pulling a thread, other things will start developing to close that loop. That's why I say it's an art. I don't think you can stop the Chinese semiconductor industry from progressing. That's basically impossible.

The Chinese nation, the Chinese government, believes in the primacy of semiconductor manufacturing. They believed it for a long time, but now they really believe it.

Dylan Patel
To some extent, the sanctions have made China believe in the importance of the semiconductor industry more than anything else.

Dwarkesh Patel
So from an AI perspective, what's the point of export controls then? If they're going to be able to get these...

Dylan Patel
Well they're not centralized though. That's the big question: are they centralized? Also, I'm not sure if I really believe it but on prior podcasts, there have been people who talked about nationalization.

Dwarkesh Patel
Why are you referring to this ambiguously? "My opponent..."

Dylan Patel
No, I love Leopold, but there have been a couple where people have talked about nationalization. If you have nationalization, then all of the sudden you aggregate all the flops. There's no fucking way.

China can be centralized enough to compete with each individual US lab. They could have just as many flops in 2025 and 2026 if they decided they were scale pilled. Just from foreign chips, for an individual model.

Dwarkesh Patel
Like they can release a 1e27 model by 2026?

Dylan Patel
And a 1e28 model in the works. They totally could do this just with foreign chip supply. It's just a question of centralization. Then the question is, do you have as much innovation and compute efficiency wins when you centralize? Or do Anthropic, OpenAI, xAI, and Google develop things, and secrets shift a bit between each other, resulting in a better long-term outcome versus nationalization in the US?

China could absolutely have it in 2026-27 if they desire to, just from foreign chips. Domestic chips are the other question. You have 600,000 of the Ascend 910B, which is roughly 400 teraflops or so. If they put them all in one cluster, they could have a bigger model than any of the labs next year. I have no clue where all the Ascend 910Bs are going, but there are rumors about them being divvied up between the majors, Alibaba, ByteDance, Baidu, etc. Next year, you have more than a million.

It's possible they actually have 1e30 before the US because data center isn't as big an issue. A 10 gigawatt data center... I don't think anyone is even trying to build that today in the US, even out to 2027-28. They're focusing on linking many data centers together.

There's a possibility that come 2028-2029, China can have more flops delivered to a single model, even once the centralization question is solved. That's clearly not happening today for either party. I'd bet if AI is as important as you and I believe, they will centralize sooner than the West does. So there is a possibility.

Dwarkesh Patel
How many more wafers could they make and how many of those wafers could be dedicated to the 910B? I assume there's other things they want to do with these semiconductors.

Dylan Patel
There's two parts there. The way the US has sanctioned SMIC is really stupid. They've sanctioned a specific spot rather than the entire company. SMIC is still buying a ton of tools that can be used for their 7nm and their 5.5nm, or 6nm process, for the 910C which releases later this year. They can build as much of that as long as it's not in Shanghai. Shanghai has anywhere from 45 to 50 high-end immersion lithography tools. That's what is believed by intelligence and many other folks.

That roughly gives them as much as 60,000 wafers a month of 7 nanometer, but they also make their 14 nanometer in that fab. The belief is that they actually only have about 25,000-35,000 capacity of 7 nanometer wafers a month. Doing the math of the chip die size and all these things—Huawei also uses chiplets so they can get away with using less leading-edge wafers but then their yields are bad—you can roughly say something like 50 to 80 good chips per wafer with their bad yield.

Dwarkesh Patel
Why do they have bad yield?

Dylan Patel
Because it's hard.

Jon Y
Even if everyone knows the number, like say there's a 1000 steps. Even if you're 98-99% for each, in the end you'll still get a 40% yield.

Dylan Patel
If it's six sigma of perfection and you have your 10,000 plus steps, you end up with yield that's still dog shit by the end.

Jon Y
That is a scientific measure, dog shit percent.

Dylan Patel
Yeah, as a multiplicative effect. Yields are bad because they have hands tied behind their back. They are not getting to use EUV. On 7 nanometer Intel never used EUV, but TSMC eventually started using EUV. Initially, they used DUV.

Dwarkesh Patel
Doesn't that mean the export controls succeeded? They have bad yield because they have to use...

Dylan Patel
It's a brand new process.

Jon Y
Again, they're still determined. Success means they stop. They're not stopping.

Dylan Patel
Let's go back to the yield question. It's theoretically 60,000 wafers a month times 50-100 dies per wafer with yielded dies. Holy shit. That's millions of GPUs. Now, what are they doing with most of their wafers? They still have not become scale pilled, so they're still throwing them out. Let's make 200 million Huawei phones, right? Okay, cool. I don't care.

As the West, you don't care as much, even though Western companies will get screwed, like Qualcomm and MediaTek Taiwanese companies. Obviously there's that. The same applies to the US, but when you flip to like... Sorry, I don't fucking know what I was going to say.

Jon Y
Nailed it!

Dwarkesh Patel
We're keeping this in

Dylan Patel
That's fine, that's fine.

Dwarkesh Patel
In 2026 if they're centralized, they can have as big training runs as any one US company—

Dylan Patel
Oh, the reason why I was bringing up Shanghai. They're building 7nm capacity in Beijing. They're building 5nm capacity in Beijing, but the US government doesn't care. They're importing dozens of tools into Beijing and saying to the US government and ASML, "This is for 28nm, obviously." In the background, they're making 5nm there.

Dwarkesh Patel
Are they doing it because they believe in AI, or because they want to make Huawei phones?

Dylan Patel
Huawei was the largest TSMC customer for a few quarters before they got sanctioned. Huawei makes most of the telecom equipment in the world. Phones, modems, accelerators, networking equipment, video surveillance chips, you go through the whole gamut. A lot of that could use 7 and 5 nanometer.

Jon Y
Do you think the dominance of Huawei is actually bad for the rest of the Chinese tech industry?

Dylan Patel
Huawei is so cracked that it's hard to say that. Huawei out-competes Western firms regularly with two hands tied behind their back. What the hell are Nokia and Sony Ericsson? They're trash compared to Huawei. Huawei isn't allowed to sell to European or American companies and they don't have TSMC. Yet they still destroy them.

The new phone is as good as a year-old Qualcomm phone on a process node that's equivalent to something three or four years old. They actually out-engineered us with the worst process node. Huawei is crazy cracked.

Jon Y
Where do you think that culture comes from?

Dylan Patel
The military, because it's the PLA.

Jon Y
It's generally seen as an arm of the PLA. How do you square that with the fact that sometimes the PLA seems to mess stuff up?

Dylan Patel
Oh, like filling water in rockets?

Jon Y
I don't know if that was true. I'm not denying it.

Dylan Patel
There is that crazy conspiracy... You don't know what the hell to believe in China, especially as a non-Chinese person.

Jon Y
Nobody knows, even Chinese people don't know what's going on in China.

Dylan Patel
There's all sorts of stuff like, "Oh, they're filling water in their rockets, clearly they're incompetent." If I'm the Chinese military, I want the Western world to believe I'm completely incompetent because one day, I can just destroy everything with hypersonic missiles and drones. "No, no, we're filling water in our missiles. These are all fake. We don't actually have a hundred thousand missiles that we manufacture in a super advanced facility and Raytheon is stupid as shit because they can't make missiles nearly as fast."

That's also the flip side. How much false propaganda is there? There's a lot of, "SMIC could never, they don't have the best tools." Then it's like, "Motherfucker, they just shipped 60 million phones last year with this chip that performs only one year worse than what Qualcomm has." The proof is in the pudding. There's a lot of cope.

Jon Y
I just wonder where that culture comes from. There's something crazy about them. Everything they touch, they seem to succeed in. I wonder why.

Dylan Patel
They're making cars. I wonder what's going on there.

Jon Y
If we imagine historically... Do you think they're getting something from somewhere?

Dylan Patel
Espionage, you mean? Obviously.

Jon Y
East Germany and the Soviet industry was basically a conveyor belt of secrets coming in, and they used that to run everything. But the Soviets were never good at it. They could never mass produce it. But now you have China.

Dwarkesh Patel
How would espionage explain how they can make things with different processes?

Dylan Patel
It's not just espionage. They're just literally cracked.

Jon Y
That's why. It has to be something else.

Dylan Patel
They have the espionage without a doubt. ASML is known to have been hacked at least a few times. People have been sued who made it to China with a bunch of documents. It's not just ASML, but every company in the supply chain. Cisco code was literally in early Huawei routers. You go down the list...

But architecturally, the Ascend 910B looks nothing like a GPU or TPU. It is its own independent thing. Sure, they probably learned some things from some places, but they're good at engineering.

Jon Y
It's 9-9-6. Wherever that culture comes from, they do good. They do very good.

Dwarkesh Patel
Another thing I'm curious about is where that culture comes from, but also how it stays there. With American firms or any other firm, you can have a company that's very good, but over time it gets worse, like Intel or many others. I guess Huawei just isn't that old, but it's hard to be a big company and stay good.

Jon Y
That is true. A word I hear a lot regarding Huawei is "struggle." China has a culture where the Communist Party is big on struggle. Huawei brought that culture into their way of doing things. You said this before, right? They go crazy because they think in five years they're going to fight the United States. Everything they do, every second it's like their country depends on it.

Dylan Patel
It's the Andy Grove-ian mindset. Shout out, the based Intel of Andy Grove. Only the paranoid survive. Paranoid Western companies do well. Why did Google really screw the pooch on a lot of stuff and then resurge now? It's because they got paranoid as hell. If Huawei is just constantly paranoid about the external world and thinking, "Oh fuck, we're gonna die. They're gonna beat us. Our country depends on it. We're going to get the best people from the entire country at whatever they do..."

Jon Y
And tell them, "If you do not succeed, our country will die. Your family will be enslaved. It will be terrible."

Dylan Patel
"By the evil western pigs."

Jon Y
"Capitalist..." Or not capitalist, they don't say that anymore. It's more like, "Everyone is against China. China is being defiled. That is all on you, bro. If you can't do that..."

Dylan Patel
"If you can't get that radio to be slightly less noisy and transmit 5% more data, we are fucked."

Jon Y
"It's like the great palace fire all over again. The British are coming and they will steal all the trinkets. That's on you."

Dwarkesh Patel
Why isn't there more vertical integration in the semiconductor industry? Why is it like, "This subcomponent requires this subcomponent from this other company, which requires this subcomponent from this other company..." Why isn't more of it done in-house?

Dylan Patel
The way to look at it today is that it's super stratified. Every industry has anywhere from one to three competitors. The most competitive it gets is like 70% share, 25% share, 5% share, in any layer of manufacturing chips, anything, chemicals, different types of chips, etc. It used to be vertically integrated.

Jon Y
At the very beginning it was integrated.

Dwarkesh Patel
Why did that stop?

Jon Y
You had companies that used to do it all in one. Then suddenly a guy would be like, "I hate this. I know how to do it better." He'd spin off, do his own thing, start his own company, then go back to his old company and say, "I can sell you a product that's better." That's the beginning of what we call the semiconductor equipment industry.

Dylan Patel
In the seventies, everyone made their own equipment.

Jon Y
Sixties and seventies. All these people spin off. What happened was that the companies that accepted outside products and equipment got better stuff and did better. There were

companies that were totally vertically integrated for decades. They are still good, but nowhere near competitive.

Dwarkesh Patel
One thing I'm confused about is the actual foundries themselves, there's fewer and fewer of them every year. There's maybe more companies overall, but fewer final wafer makers. It's similar to AI foundation models where you need revenues from a previous model. You need market share to fund the next round of ever more expensive development.

Jon Y
When TSMC launched the foundry industry, there was a wave of Asian companies that funded semiconductor foundries. Malaysia with Silterra, Singapore with Chartered, one from Hong Kong...

Dylan Patel
A bunch in Japan.

Jon Y
A bunch in Japan. They all did this thing. When going to leading-edge, it got harder, which means you had to aggregate more demand from all the customers to fund the next node. Technically you're aggregating all this profit to fund the next node to the point where now there's no room in the market for an N2 or N3. You could argue that economically, N2 is a monstrosity that doesn't make sense. It should not exist without the immense concentrated spend of like five players in the market.

Dylan Patel
I'm sorry to completely derail you, but there's this video where it's like, "This is an unholy concoction of meat slurry."

Jon Y
Yes!

Dwarkesh Patel
What?

Dylan Patel
There's this video that's like, "Ham is disgusting. It's an unholy concoction of meat with no bones or collagen." I don't know. The way he was describing 2nm is like that.

Jon Y
It's like the guy who pumps his right arm so much. He's super muscular. "The human body was not meant to be so muscular!"

Dwarkesh Patel
What's the point? Why is 2 nanometer not justified?

Jon Y
I'm not saying it for N2 specifically, but N2 as a concept. The next node should technically... There will come a point where economically, the next node will not be possible at all.

Dylan Patel
Unless more technology spawned, like AI now makes one nanometer or whatever, A16, viable.

Dwarkesh Patel
Makes it viable in what sense? It makes it worth it?

Jon Y
Money. Money.

Dylan Patel
Every two years you get a shrink, like clockwork, Moore's law. Then five nanometer happened. It took three years. Holy shit. Then three nanometer happened. It took three years. Is Moore's law dead? Because TSMC didn't... and then what did Apple do? When three nanometer finally launched, Apple only moved half of the iPhone volume to three nanometer. Now they did a fourth year of five nanometer for a big chunk of iPhones. Is the mobile industry petering out?

Then you look at two nanometer and it's going to be similarly difficult for the industry to pay for this. Apple because they get to make the phone, they have so much profit they can funnel into more and more expensive chips. But finally that was really running out. How economically viable is 2nm just for one player, TSMC? Ignore Intel, ignore Samsung. Samsung is paying for it with memory, not with their actual profit. Intel is paying for it from their former CPU monopoly...

Jon Y
Private equity money, chips money, debt, and the salaries of laid-off people.

Dylan Patel
There's a strong argument that funding the next node wouldn't be economically viable anymore if it weren't for AI taking off and generating humongous demand for the most leading-edge chip.

Dwarkesh Patel
How big is the difference between 7nm to 5nm to 3nm? Is it a huge deal in terms of who can build the biggest cluster?

Dylan Patel
There's this simplistic argument that moving a process node only saves X percent in power. That has been petering out. When you moved from 90 nanometer to 80 or 70 something, you got 2x. Dennard scaling was still intact. But now when you move from 5 nanometer to 3 nanometer, you don't double density. SRAM doesn't scale at all. Logic does scale, but it's like 30%. All in all, you only save about 20% in power per transistor.

But because of data locality and movement of data, you actually get a much larger improvement in power efficiency by moving to the next node than just the individual transistors' power efficiency benefit. For example, if you're multiplying a matrix that's 8,000 by 8,000 by 8,000, you can't fit that all on one chip. But if you could fit more and more, you have to move off chip less, go to memory less, etc. The data locality helps a lot too.

AI really, really wants new process nodes because power usage is a lot less, you get higher density and higher performance. The big deal is if I have a gigawatt data center, how much more flops can I get? If I have a two gigawatt data center, how much more flops can I get? If I have a ten gigawatt data center, how much more flops can I get? You look at the scaling and everyone needs to go to the most recent process node as soon as possible.

Dwarkesh Patel
I want to ask a normie question... I won't phrase it that way.

Jon Y
Not for you nerds.

Dylan Patel
I think Jon and I could communicate to the point where you even wouldn't know what we're talking about.

Dwarkesh Patel
Suppose Taiwan is invaded or Taiwan has an earthquake. Nothing is shipped out of Taiwan from now on. What happens next? How would the rest of the world feel its impact a day in, a week, a month in, a year in?

Jon Y
It's a terrible thing to talk about. Can you just say it's all just terrible? It's not just leading-edge. People will focus on leading-edge, but there's a lot of trailing-edge stuff that people depend on every day. We all worry about AI. The reality is you're not going to get your

fridge. You're not going to get your cars. You're not going to get everything. It's terrible. Then there's the human part of it. It's all terrible. It's depressing. And I live there.

Dylan Patel
Day one, the market crashes a lot. The six or seven biggest companies, the Magnificent Seven, are like 60-75% of the S&P 500 and their entire business relies on chips. Google, Microsoft, Apple, Nvidia, Meta, they all entirely rely on AI. You would have an extremely insane tech reset.

So the market would crash. A couple weeks in, people are preparing now. People are like, "Oh shit, let's start building fabs. Fuck all the environmental stuff." War's probably happening. The supply chain is trying to figure out what the hell to do to refix it.

Six months in, the supply of chips for making new cars is gone or sequestered to make military shit. You can no longer make cars. We don't even know how to make non-semiconductor induced cars, this unholy concoction with all these chips.

Jon Y
Cars are like 40% chips now. There are chips in the tires.

Dylan Patel
There's like 2,000+ chips in every car. Every Tesla door handle has like four chips in it. It's like, "What the fuck?" Why? It's like shitty microcontrollers and stuff but there's 2000+ chips even in an internal combustion engine vehicle. Every engine has dozens and dozens of chips.
Anyways, this all shuts down. Not all of the production, there's some in Europe, some in the US, some in Japan, some in Singapore.

Jon Y
In Europe they're going to bring in a guy to work on Saturday, until four.

Dylan Patel
Yeah. TSMC always builds new fabs. They tweak production up in old fabs. New designs move to the next nodes and old stuff fills in the old nodes. Ever since TSMC has been the most important player... It's not just TSMC, there's UMC there, PSMC, and other companies. Taiwan's share of total manufacturing has grown every single process node.

In 130 nanometers, there's a lot. That's including many chips from Texas Instruments, Analog Devices, or NXP. 100% of it is manufactured in Taiwan by PSMC, TSMC, UMC or whatever. But then you step forward to 28 nanometer, 80% of the world's production is in Taiwan. Oh fuck, right? What's made on 28 nanometer today? It's tons of microcontrollers

and stuff but also every display driver I see. Cool, even if I could make my Mac chip, I can't make the chip that drives the display.

You just go down the list. That means no fridges, no automobiles, no weed whackers, because that stuff has chips. My toothbrush has Bluetooth in it. Why? I don't know. There are so many things that would just go poof. We'd have a tech reset.

Dwarkesh Patel
We were supposed to do this interview many months ago. I kept delaying because I was like, "Ah, I don't understand any of this shit." But it is a very difficult thing to understand. Whereas with AI, it's like...

Dylan Patel
You've just spent the time to —

Dwarkesh Patel
Sure but it feels like the kind of thing where you can pick up what's going on in the field in an amateur kind of way. In this field, I'm curious about how one learns the layers of the stack. It's not just papers online. You can't just look up a tutorial on how the transformer works. It's many layers of really difficult shit.

Dylan Patel
There are 18-year-olds who are cracked at AI already. There are high school dropouts who get jobs at OpenAI. This existed in the past. Pat Gelsinger, the current CEO of Intel, grew up in the Amish area of Pennsylvania. He went straight to work at Intel because he's just cracked. That's not possible in semiconductors today. You can't get a job at a tool company without at least a master's in chemistry, probably a PhD. Of the 75,000 TSMC workers, like 50,000 have a PhD or something insane. There's a next-level amount of how specialized everything's gotten.

Whereas today, you can take someone like Sholto, who started working on AI not that long ago.

Jon Y
Not to say anything bad about Sholto.

Dylan Patel
No, he's cracked. He's omega-cracked at what he does. You could drop him into another part of the AI stack. He understands it already and could probably become cracked at that too. That's not the case in semiconductors. You specialize like crazy and can't just pick it up. Sholto, what did he say... He just started—

Dwarkesh Patel
He was a consultant at McKinsey, and at night he would read papers about robotics and run experiments.

Dylan Patel
And then people noticed him and were like, "Who is this guy? I thought everyone who knew about this was at Google already. Come to Google." That can't happen in semiconductors. It's just not possible. ArXiv is a free thing. The paper publishing industry is abhorrent everywhere else. You can't just download IEEE papers or SPIE papers or from other organizations.

At least up until late 2022 or early 2023 with Google's PaLM inference paper, all the best stuff was just posted on the internet. After that, there was a little bit of clamping down by the labs, but there are also still all these companies making innovations in the public. What is state-of-the-art is public. That is not the case in semiconductors.

Jon Y
Semiconductors have been shut down since the 1970s basically. It's crazy how little information has been formally transmitted from one country to another. The last time you could really think of this was maybe the Samsung era.

Dwarkesh Patel
So then how do you guys keep up with it?

Jon Y
We don't know it. I don't personally think I know it.

Dwarkesh Patel
If you don’t know it, what are you making videos about?

Jon Y
I spoke to one guy. He’s a PhD in etch, one of the top people in etch. He’s like, “Man, you really know lithography.” I don't feel like I know lithography. But then you talk to the people who know lithography and they’re like, “You’ve done pretty good work in packaging.” Nobody knows anything.

Dwarkesh Patel
They all have Gell-Mann amnesia?

Jon Y
They’re all in this single well. They’re digging deep for what they’re getting at. They don’t know the other stuff well enough. In some ways, nobody knows the whole stack.

Dylan Patel
The stratification of just manufacturing is absurd. The tool people don't even know exactly what Intel and TSMC do in production, and vice versa, they don't know exactly how the tool is optimized like this. How many different types of tools are there? Each of those has an entire tree of all the things we've built, invented, and continue to iterate upon, and then there's the breakthrough innovation that happens every few years in it too.

Dwarkesh Patel
If that's the case and nobody knows the whole stack, how does the industry coordinate? "In two years we want to go to the next process which has gate all-around and for that we need X tools and X technologies developed..."

Jon Y
It's a fascinating social phenomenon. You can feel it. I went to Europe earlier this year. Dylan had allergies. I was talking to those people. It's like gossip. You start feeling people coalescing around something. Early on, we used to have SEMATECH where American companies came together and talked and hammered it out. But in reality it was dominated by a single company. Nowadays it's more dispersed. It's a blue moon arising kind of thing. They are going towards something. They know it. Suddenly, the whole industry suddenly is like, "This is it. Let's do it."

Dylan Patel
It's like God came and proclaimed it: "We will shrink density 2x every two years." Gordon Morris made an observation. It didn't go nowhere. It went way further than he ever expected because it's like, "There's a line of sight to get to here and here." He predicted 7-8 years out, multiple orders of magnitude of increases in transistors and it came true. But by then, the entire industry was like, "This is obviously true. This is the word of God." Every engineer in the entire industry, tens of millions of people, were driven to do this.

Now not every single engineer believed it but people were like, "Yes, to hit the next shrink, we must do this, this, this and these are the optimizations we make." You have this stratification, every single layer and abstraction layers through the entire stack. It's an unholy concoction. No one knows what's going on because there's an abstraction layer between every single layer. On this layer, the people below you and above you know what's going on. Beyond that, you can try to understand, but not really...

Dwarkesh Patel
I watched a video about IRDS or whatever, 10 or 20 years ago, where they're like, "We're going to do EUV instead of the other thing. This is the path forward." How do they do that if they don't have the whole picture of different constraints, trade-offs, and so on?

Jon Y
They kind of argue it out. They get together and they talk and argue. Basically, at some point, a guy somewhere says, "I think we can move forward with this."

Dylan Patel
Semiconductors are so siloed. The data and knowledge within each layer is: A) Not documented online at all because it's all siloed within companies. B) There's a lot of human element to it because a lot of the knowledge, as Jon was saying, is apprentice-master type of knowledge. Or it's "I've been doing this for 30 years," and there's an amazing amount of intuition on what to do just when you see something.

AI can't just learn semiconductors like that. But at the same time, there's a massive talent shortage and ability to move forward on things. Most of the equipment in semiconductor fabs runs on Windows XP. Each tool has a Windows XP server on it. All the chip design tools have CentOS version 6, which is old as hell. There are so many areas where it's so far behind. At the same time, it's so hyper-optimized that the tech stack is broken in that sense.

Jon Y
They're afraid to touch it.

Dylan Patel
Yeah, because it's an unholy amalgamation.

Jon Y
This thing should not work. It's literally a miracle.

Dylan Patel
So you have all the abstraction layers. One, there's a lot of breakthrough innovation that can happen now stretching across abstraction layers. Two, because there's so much inherent knowledge in each individual one, what if I can just experiment and test at a 1000x or 100,000x velocity?

Some examples of where this is already shown true are some of NVIDIA's AI layout tools, and Google as well, laying out the circuits within a small blob of the chip with AI. Some of these RL design things, various simulation things...

Jon Y
But is that design or is that manufacturing?

Dylan Patel
It's all design, most of it is design. Manufacturing has not really seen much of this yet, although it's starting to come in.

Jon Y
Inverse lithography, maybe.

Dylan Patel
ILT and... maybe. I don't know if that's AI. Anyway, there's a tremendous opportunity to bring breakthrough innovation simply because there are so many layers where things are unoptimized. You see all these single-digit to low double-digit advantages just from RL techniques from AlphaGo-type stuff, or not AlphaGo but 5-8 year-old RL techniques being brought in. Generative AI being brought in could really revolutionize the industry, although there's a massive data problem.

Dwarkesh Patel
Can you give the possibilities here in numbers in terms of maybe like a FLOP per dollar or whatever the relevant thing is? How much do you expect in the future to come from process node improvements? How much from just how the hardware is designed because of AI? We're talking specifically for GPUs. If you had to disaggregate future improvements.

Dylan Patel
It's important to state that semiconductor manufacturing and design is the largest search space of any problem that humans do because it is the most complicated industry that humans do. When you think about it, there are 1e10, 1e11, 100 billion transistors, on leading-edge chips. Blackwell has 220 billion transistors or something like that.

Those are just on-off switches. Think about every permutation of putting those together, contact, ground, drain source, with wires. There are 15 metal layers connecting every single transistor in every possible arrangement. This is a search space that is literally almost infinite. The search space is much larger than any other search space that humans know of.

Dwarkesh Patel
And what is the nature of the search? What are you trying to optimize over?

Dylan Patel
Useful compute. If the goal is to optimize intelligence per picojoule—and intelligence is some nebulous nature of what the model architecture is, and picojoule is a unit of energy—how do you optimize that?

There are humongous innovations possible in architecture because the vast majority of the power on an H100 does not go to compute. There are more efficient ALU (Arithmetic Logic

Unit) designs. But even then, the vast majority of the power doesn't go there. The vast majority of the power goes to moving data around. When you look at what the movement of data is, it's either networking or memory.

You have a humongous amount of movement relative to compute and a humongous amount of power consumption relative to compute. So how can you minimize that data movement and maximize the compute? There are 100x gains possible from architecture. Even if we literally stopped shrinking, we could have 100x gains from architectural advancements.

Dwarkesh Patel
Over what time period?

Dylan Patel
The question is how much can we advance the architecture. The other challenge is that the number of people designing chips has not necessarily grown in a long time. Company to company, it shifts, but within the semiconductor industry in the US—the US designs the vast majority of leading-edge chips—the number of people designing chips has not grown much.

What has happened is the output per individual has soared because of EDA (Electronic Design Assistance) tooling. Now this is all still classical tooling. There's just a little bit of AI in there yet. The question is what happens when we bring this in and how you can solve this search space somehow, with humans and AI working together to optimize this so most of the power doesn't just go to data movement and the compute is actually very small. The compute can get like 100x more efficient just with design changes, and then you could minimize that data movement massively. You can get a humongous gain in efficiency just from architecture itself.

Process node helps you innovate that there. Power delivery helps you innovate that. System design, chip-to-chip networking helps you innovate that. Memory technologies, there's so much innovation there. There are so many different vectors of innovation that people are pursuing simultaneously.

NVIDIA gen to gen to gen will do more than 2x performance per dollar. I think that's very clear. Hyperscalers are probably going to try and shoot above that, but we'll see if they can execute.

Dwarkesh Patel
There are two narratives you can tell here of how this happens. One is that these AI companies training the foundation models understand the trade-offs of. How much is the marginal increase in compute versus memory worth to them? What trade-offs do they want between different kinds of memory? They understand this, so the accelerators they build

can make these trade-offs in a way that's most optimal. They can also design the architecture of the model itself in a way that reflects the hardware trade-offs.

Another is NVIDIA. I don't know how this works. Presumably, they have some sort of know-how. They're accumulating all this knowledge about how to better design this architecture and also better search tools. Who has the better moat here? Will NVIDIA keep getting better at design, getting this 100x improvement? Or will it be OpenAI and Microsoft and Amazon and Anthropic who are designing their accelerators and will keep getting better at designing the accelerator?

Dylan Patel
There are a few vectors to go here. One you mention is important to note. Hardware has a huge influence on the model architecture that's optimal. It's not a one-way street that better chip equals... The optimal model for Google to run on TPUs, given a certain amount of dollars and compute, is different architecturally than what it is for OpenAI with NVIDIA stuff. It's absolutely different. Even down to networking decisions and data center designs that different companies make, the optimal solution—X amount of compute of TPU vs. GPU compute optimally—will diverge in what the architecture is. That's important to note.

Dwarkesh Patel
Can I ask about that real quick? Earlier we were talking about how China has the H20s or B20s, and there's much less compute per memory bandwidth than the amount of memory. Does that mean that Chinese models will actually have very different architecture and characteristics than American models in the future?

Dylan Patel
You can take this to a very large leap and say, "Oh, neuromorphic computing or whatever is the optimal path and that looks very different than what a transformer does." Or you could take it to a simple thing, the level of sparsity and coarse-grain sparsity, experts, and all this sort of stuff. You have the arrangement of what exactly the attention mechanism is, because there are a lot of tweaks. It's not just pure transformer attention.

Or how wide versus tall the model is, that's very important. D-mod versus number of layers. These are all things that would be different. I know they're different between, say, Google and OpenAI and what is optimal. But it really starts to get like, "Hey, if you were limited on a number of different things like..." China invests hugely in compute and memory. The memory cell is directly coupled or is the compute cell. These are things that China's investing hugely in. You go to conferences and there are like 20 papers from Chinese companies/universities about compute and memory.

Because the FLOP limitation is here, maybe NVIDIA pumps up the on-chip memory and changes the architecture because they still stand to benefit tens of billions of dollars by

selling chips to China. Today, it's just neutered American chips that go to China. But it'll start to diverge more and more architecturally because they'd be stupid not to make chips for China.

Huawei, obviously, has their constraints. Where are they limited on memory? Oh, they have a lot of networking capabilities and they could move to certain optical networking technologies directly onto the chip much sooner than we could. Because that is what's optimal for them within their search space of solutions, because this whole area is blocked off.

Jon Y
It's really interesting to think about how the development of Chinese AI models will differ from American AI models because of these changes or constraints.

Dylan Patel
It applies to use cases. It applies to data. American models are very focused on learning from you, being able to use you directly as a random consumer. That's not the case for Chinese models, I assume. There are probably very different use cases for them. China crushes the West at video and image recognition.

At ICML, Albert Gu of Cartesia, who invented state space models, was there. Every single Chinese person was like, "Can I take a selfie with you?" The man was harassed. In the US, you see Albert and it's awesome, he invented state space models. It's not like state space models are revered here. But that's because state space models potentially have a huge advantage in video, image, and audio, which is stuff that China does more of and is further along in and has better capabilities in.

Jon Y
Because of all the surveillance cameras there.

Dylan Patel
Yeah. That's the quiet part out loud. But there's already divergence in capabilities there. If you look at image recognition, China destroys American companies on that because of the surveillance.

You have this divergence in tech tree and people can start to design different architectures within the constraints they're given. Everyone has constraints, but the constraints different companies have are even different. Google's constraints have shown them that they built a genuinely different architecture. But now if you look at Blackwell and what's said about TPU v6, they're not exactly converging but they are getting a little bit closer in terms of how big the MatMul unit size is and some of the topology and world size of the scale-up versus scale-out network. There is some convergence slightly. I'm not saying they're similar yet,

but they're starting to. Then there are different architectures that people could go down. You see stuff from all these startups that are trying to go down different tech trees because maybe that'll work.

There's a self-fulfilling prophecy here too. All the research is in transformers that are very high arithmetic intensity because the hardware we have is very high arithmetic intensity and transformers run really well on GPUs and TPUs. You sort of have a self-fulfilling prophecy. If all of a sudden you have an architecture which is theoretically way better, but you can get only like half of the usable FLOPs out of your chip, it's worthless because even if it's a 30% compute efficiency win, it's half as fast on the chip. There are all sorts of trade-offs and self-fulfilling prophecies of what path people go down.

Dwarkesh Patel
If you were made head of compute of a new AI lab, if Ilya Sutskever's new lab SSI came to you and they're like, "Dylan, we give you $1 billion. You're our head of compute. Help us get on the map. We're going to compete with the frontier labs." What is your first step?

Dylan Patel
Okay. The constraints are that you're a US/Israeli firm because that's what SSI is. Your researchers are in the US and Israel. You probably can't build data centers in Israel because power is expensive as hell and it's probably risky. So it's still in the US most likely. Most of the researchers are here in the US, in Palo Alto or wherever.

You need a significant chunk of compute. Obviously, the whole pitch is you're going to make some research breakthrough, compute efficiency, data efficiency, or whatever it is. You're going to make some research breakthroughs but you need compute to get there. Your GPUs per researcher is your research velocity. Obviously, data centers are very tapped out. Maybe not tapped out but in terms of every new data center coming up, most of them have been sold. That's led people like Elon to go through this insane thing in Memphis. I'm just trying to square the circle.

Dwarkesh Patel
On that question, I kid you not, in my group house group chat, there have been two separate people who have been like, "I have a cluster of H100s and I have a long lease on them, but I'm trying to sell them off." Is it like a buyer's market right now? Because it does seem like people are trying to get rid of them.

Dylan Patel
For the Ilya question, a cluster of like 256 GPUs or even 4K GPUs is kind of cope. It's not enough. Yes, you're going to make compute efficiency wins, but with a billion dollars you probably just want the biggest cluster in one individual spot. Small amounts of GPUs are probably not possible to use for them, and that's what most of the sales are.

You go and look at GPU List or Vast or Foundry or a hundred different GPU resellers, the cluster sizes are small. Now, is it a buyer's market? Yeah. Last year you would buy H100s for like $4 or $3 an hour for shorter-term or mid-term deals. Right now, if you want a six-month deal, you could get it for like $2.15 or less.

The natural cost... If I have a data center and I'm paying standard data center pricing to purchase the GPUs and deploy them, it's like $1.40. Add on the debt, because I probably took debt to buy the GPUs or you have cost of equity, the cost of capital, it gets up to like $1.70 or something. You see deals that are... The good deals are like Microsoft renting from CoreWeave at like $1.90 to $2. People are getting closer and closer. But there's still a lot of profit. The natural rate even after debt and all this is $1.70. There's still a lot of profit when people are selling in the low twos.

GPU companies are deploying them, but it is a buyer's market in the sense that it's gotten a lot cheaper. Cost of compute is going to continue to tank. I don't remember the exact name of the law. It's effectively Moore's Law. Every two years, the cost of transistors halved, and yet the industry grew. Every six months or three months, the cost of intelligence... OpenAI and GPT-4 in February 2023, was roughly $120 per million tokens. Now it's like $10. The cost of intelligence is tanking, partially because of compute, partially because of the model's compute efficiency wins. That's a trend we'll see. That's gonna drive adoption as you scale up and make it cheaper and scale up and make it cheaper.

Dwarkesh Patel
Right. Anyway, if you were head of compute at SSI...

Dylan Patel
Okay, I'm head of compute at SSI. There's obviously no free data center lunch, in terms of what we see in the data. There's no free lunch if you need compute for a large cluster size, even six months out. There's some availability, but not a huge amount because of what X did.

xAI is like, "Oh shit, we're going to buy a Memphis factory, put a bunch of mobile generators usually reserved for natural disasters outside, add a Tesla battery pack, drive as much power as we can from the grid, tap the natural gas line that's going to the natural gas plant two miles away, the gigawatt natural gas plant, and just send it. Get a cluster built as fast as possible." Now you're running 100K GPUs. That costs about $4-5 billion, not $1 billion.
The scale that SSC has is much smaller. Their cluster size will be maybe 1/3 or 1/4 of that size. So now you're talking about a 25K to 32K cluster. You still don't have that. No one is willing to rent you a 32K cluster today, no matter how much money you have. Even if you had more than a billion dollars.

Now it makes the most sense to build your own cluster instead of renting, or get a very close relationship like OpenAI/Microsoft with CoreWeave or Oracle/Crusoe The next step is Bitcoin. OpenAI has a data center in Texas, or it's going to be their data center. It's kind of contracted and all that from CoreWeave. There is a 300 megawatt natural gas plant on site, powering these crypto mining data centers from a company called Core Scientific. They're just converting that. There's a lot of conversion, but the power's already there. The power infrastructure is already there. It's really about converting it, getting it ready to be water-cooled, all that sort of stuff, and converting it to a 100,000 GB200 cluster.

They have a number of those going up across the country, but that's also tapped out to some extent because NVIDIA is doing the same thing in Plano, Texas for a 32,000 GPU cluster that they're building.

Dwarkesh Patel
Is NVIDIA doing that?

Dylan Patel
Well, they're going through partners. Because this is the other interesting thing: the big tech companies can't do crazy shit like Elon did.

Dwarkesh Patel
Why?

Dylan Patel
ESG. They can't just do crazy shit like...

Dwarkesh Patel
Oh that's interesting. Do you expect Microsoft and Google and whoever to like drop their net zero commitments as the scaling picture intensifies?

Dylan Patel
Yeah. What xAI is doing isn't that polluting in the scheme of things, but it's you have 14 mobile generators and you're just burning natural gas on site on these mobile generators that sit on trucks. Then you have power directly two miles down the road. There's no way to say any of the power is green because up to two miles down the road is a natural gas plant as well.

You go to the CoreWeave thing. There's a natural gas plant literally on site from Core Scientific and all that. Then the data centers around it are horrendously inefficient. There's this metric called PUE, which is basically how much power is brought in versus how much gets delivered to the chips. The hyperscalers, because they're so efficient, their PUE is like 1.1 or lower. I.e., if you get a gigawatt in, 900 megawatts or more gets delivered to chips. It's

not wasted on cooling and all these other things. This Core Scientific one is going to be like 1.5 or 1.6. I.e., even though I have 300 megawatts of generation on site, I only deliver like 180-200 megawatts to the chips.

Dwarkesh Patel
Given how quickly solar is getting cheaper... There's also the fact that the reason solar is difficult elsewhere is you've got to power the homes at night. Here I guess it's theoretically possible to figure out only running the clusters in the day or something...

Dylan Patel
Absolutely not. That's not possible.

Dwarkesh Patel
Because it's so expensive to have these GPUs?

Dylan Patel
Yes. So when you look at the power cost of a large cluster, it's trivial to some extent. The meme that you can't build a data center in Europe or East Asia because the power is expensive, that's not really relevant. Or there's the meme that power is so cheap in China and the US that those are the only places you can build data centers. That's not really the real reason. It's the ability to generate new power for these activities. That's why it's really difficult, the economic regulation around that.

But the real thing is if you look at the cost of ownership of an H100. Let's just say you gave me a billion dollars and I already have a data center, I already have all this stuff. I'm paying regular rates for the data centers, not paying through the nose or anything. Paying regular rates for power, not paying through the nose. Power is sub 15% of the cost. It's sub 10% of the cost actually. The biggest, like 75-80% of the cost, is just the servers.

And this is on a multi-year basis, including debt financing, including cost of operation, all that. When you do a TCO (total cost of ownership), like 80% is the GPUs, 10% is the data center, 10% is the power, rough numbers. So it's kind of irrelevant how expensive the power is. You'd rather do what Taiwan does. What did they do when there were droughts? They forced people to not shower.

Jon Y
They basically reroute the power... When there was a power shortage in Taiwan, they basically rerouted power from the residential areas.

Dylan Patel
And this will happen in a capitalistic society as well, most likely because, "fuck you, you aren't going to pay X dollars per kilowatt hour. To me, the marginal cost of power is

irrelevant. Really it's all about the GPU cost and the ability to get the power. I don't want to turn it off eight hours a day.

Dwarkesh Patel
Let's zoom out a bit. Let's discuss what would maybe happen if the training regime changes and if it doesn't change. You could imagine that the training regime becomes much more parallelizable where it's about coming up with some sort of search and most of the compute for training is used to come up with synthetic data or do some kind of search. That can happen across a wide area. In that world, how fast could we scale? Let's go through the numbers year after year.

You would know more than me, but then suppose it has to be the current regime. Just explain what that would mean in terms of how distributed that would have to be and how plausible it is to get clusters of certain sizes over the next few years.

Dylan Patel
It's not too difficult for Ilya's company to get a cluster of like 32K of Blackwell next year.

Dwarkesh Patel
Forget about Ilya's company, let's talk about the clear players. Like 2025, 2026, 2027.

Dylan Patel
2025, 2026... Before I talk about the US, it's important to note that there's like a gigawatt plus of data center capacity in Malaysia next year now. That's mostly ByteDance. Power-wise there's also the humongous damming of the Nile in Ethiopia and the country uses like one-third of the power that that dam generates. There's like a ton of power there to...

Dwarkesh Patel
How much power does that dam generate?

Dylan Patel
It's like over a gigawatt. The country consumes like 400 megawatts or something trivial.

Dwarkesh Patel
Are people bidding for that power?

Dylan Patel
I think people just don't think they can build a data center in Ethiopia.

Dwarkesh Patel
Why not?

Jon Y
I don't think the dam is filled yet, is it?

Dylan Patel
No, the dam could generate that power. They just don't. There's a little bit more equipment required, but that's not too hard. Why don't they? There are true security risks. If you're China, or if you're a US lab, to build a data center with all your IP in Ethiopia... You want AGI to be in Ethiopia? You want it to be that accessible? People can't even monitor the technicians in the data center or powering the data center, all these things. There are so many things you could do... You could just destroy every GPU in a data center if you want if you just fuck with the grid, pretty easily, I think.

Dwarkesh Patel
People talk a lot about it in the Middle East.

Dylan Patel
There's a 100k GB200 cluster going up in the Middle East. The US is clearly doing stuff too. G42 is the UAE data center company, cloud company. Their CEO is a Chinese national, or not a Chinese national but there's basically Chinese allegiance. OpenAI wanted to use a data center from them but instead... The US forced Microsoft—I feel like this is what happened—to do a deal with them so that G42 has a 100K GPU cluster, but Microsoft is administering and operating it for security reasons.

There's Omniva in Kuwait, like the Kuwait super-rich guy spending five plus billion dollars on data centers. You just go down the list, all these countries. Malaysia has $10+ billion of AI data center build outs over the next couple of years. Go to every country, this stuff is happening.

But in the grand scheme of things, the vast majority of the compute is being built in the US, then China, then Malaysia, Middle East, and the rest of the world. Let's go back to your point. You have synthetic data. You have the search stuff. You have all these post-training techniques. You have all these ways to soak up flops, or you just figure out how to train across multiple data centers, which I think they have. At least, Microsoft and OpenAI have figured it out.

Dwarkesh Patel
What makes you think they figured it out?

Dylan Patel
Their actions. Microsoft has signed deals north of $10 billion with fiber companies to connect their data centers together. There are some permits already filed to show people

are digging between certain data centers. We think, with fairly high accuracy, that there are five regions that they're connecting together, which comprises many data centers.

Dwarkesh Patel
What will be the total power usage of the...

Dylan Patel
Depends on the time, but easily north of a gigawatt.

Dwarkesh Patel
Which is like close to a million GPUs.

Dylan Patel
Well, each GPU is getting higher power consumption too. The rule of thumb is that a H100 is like 700 watts, but then total power per GPU all-in is like 1200-1400 watts. But next-generation NVIDIA GPUs are like 1200 watts for the GPU. It actually ends up being like 2000 watts all in. There's a little bit of scaling of power per GPU.

You already have 100K clusters. OpenAI in Arizona, xAI in Memphis. Many others are already building 100K clusters of H100s. You have multiple, at least five, I believe GB200 100K clusters being built by Microsoft/OpenAItheir partners for them. It's potentially even more. 500K GB200s is like a gigawatt and that's online next year.

The year after that, if you aggregate all the data center sites, and how much power... You only look at net adds since 2022, instead of the total capacity at each data center, then you're still north of multi-gigawatt.

They're spending north of $10+ billion dollars on these fiber deals with a few fiber companies: Lumen, Zayo, and a couple other companies. Then they've got all these data centers where they're clearly building 100K clusters, like old crypto mining sites with CoreWeave in Texas or this Oracle/Crusoe thing in Texas. You have them in Wisconsin and Arizona and a couple other places. There's a lot of data centers being built up. You have providers like QTS and Cooper and many other providers and self-build data centers, data centers I'm building myself.

Dwarkesh Patel
Let's just give the number to like, "Okay, it's 2025 and Elon's cluster is going to be the biggest..." It doesn't matter who it is.

Dylan Patel
There's the definition game. Elon claims he has the largest cluster at 100K GPUs because they're all fully connected.

Dwarkesh Patel
Rather than who it is, I just want to know how many... I don't know if it's better to denominate in H100s...

Dylan Patel
It's 100k GPUs this year for the biggest cluster. Next year, 300-500k depending on whether it's one site or many. 300-700k I think is the upper bound of that. But it's about when they tier it on, when they can connect them, when the fiber's connected together. Let's say 300-500k but those GPUs are 2-3X faster versus the 100K cluster. So on an H100 equivalent basis, you're at a million chips next year in one cluster by the end of the year.

Well, one cluster is the wishy-washy definition. It's multisite, right? Can you do multi-site? What's the efficiency loss when you go multisite? Is it possible at all? I truly believe so. What's the efficiency loss is the question.

Dwarkesh Patel
Would it be like 20% loss, 50% loss?

Dylan Patel
Great question. This is where you need the secrets. And Anthropic's got similar plans with Amazon and you go down the list.

Dwarkesh Patel
And then the year after that? This is 2026.

Dylan Patel
In 2026 there is a single gigawatt site. And that's just part of the multiple sites for Microsoft.

Dwarkesh Patel
The Microsoft five gigawatt thing happens in 2026?

Dylan Patel
One gigawatt, one site in 2026. But then you have a number of others. You have five different locations, some with multiple sites, some with single sites. You're easily north of 2-3 gigawatts.

Then the question is, can you start using the old chips with the new chips? The flop scaling is going to continue much faster than people expect, as long as the money pours in. There's no way you can pay for the scale of clusters being planned to be built next year for OpenAI unless they raise like $50-100 billion, which I think they will raise late this year or early next year.

Jon Y
$50-100 billion? Are you kidding me?

Dylan Patel
No. Sam has a superpower. It's recruiting and raising money. That's what he's like a god at.

Dwarkesh Patel
Will chips themselves be a bottleneck to the scaling?

Dylan Patel
Not in the near term. It's more about concentration versus decentralization. The largest cluster is 100,000 GPUs. NVIDIA has manufactured close to 6 million Hoppers across last year and this year. So that's fucking tiny.

Dwarkesh Patel
But then why is Sam talking about $7 trillion to build foundries and whatever?

Dylan Patel
Draw the line. It's like a log-log line. Numbers go up, right? If you do that, you're going from 100K to 300-500K, where the equivalent is a million. You just 10X year on year. Do that again, do that again, or more. If you increase the pace...

Dwarkesh Patel
What is "do that again"? So in 2026, the number of...

Dylan Patel
If you increase the globally produced flops by like 30x year on year or 10x year on year—and the cluster size grows by 3-7x and you start getting multi-site going better and better—you can get to the point where multi-million chip clusters, even if they're regionally not connected right next to each other, are right there.

Dwarkesh Patel
And in terms of flops it would be 1e... what?

Dylan Patel
I think 1e30 is very possible in 2028 or 2029.

Dwarkesh Patel
Wow. Okay you're saying 1e30 you said by 2028-29. That is literally six orders of magnitude. That's like 100,000x more compute than GPT-4.

Dylan Patel
Yes. The other thing to say is the way you count flops on a training run is really stupid. You can't just do like active parameters x tokens x six. That's really dumb because the paradigm—as you mentioned, and you've had many great podcasts on this stuff—it's like synthetic data and RL stuff, post-training, verifying data, all these things generating and throwing it away, search, inference time compute. All these things aren't counted in the training flops.

So you can't say 1e30 is a really stupid number to say because by then the actual flops of the pre-training may be X, but the data to generate for the pre-training may be way bigger, or the search inference time may be way, way bigger.

Dwarkesh Patel
Right. Also, because you're doing adversarial synthetic data where the thing you're weakest at, you can make synthetic data for that, it might be way more sample efficient. So even though...

Dylan Patel
Pre-training flops will be irrelevant. I actually don't think pre-training flops will be 1e30. I think more reasonably, it will be like the total summation of the flops that you deliver to the model across pre-training, post-training, synthetic data for that pre-training and post-training data, as well as some of the inference time compute efficiencies. It's more like 1e30 in total.

Dwarkesh Patel
Suppose you really do get to the world where it's worth investing.... Actually, if you're doing 1e30, is that like a trillion dollar cluster, hundred billion dollar cluster?

Dylan Patel
It'll be like multi-hundred billion dollars, but I truly believe people are going to be able to use their prior generation clusters alongside their new generation clusters. Obviously it'll be smaller batch sizes or whatever, or use that to generate and verify data, all these sorts of things.

Dwarkesh Patel
And then for 1e30... Right now, I think 5% of TSMC's N5 is NVIDIA or whatever percent it is. By 2028, what percentage will it be?

Dylan Patel
Again, this is a question of how scale-pilled you are and how much money will flow into this and how you think progress works. Will models continue to get better or does the line slope over? I believe it'll continue to skyrocket in terms of capability. In that world—not of 5

nanometer, but of 2 nanometer, A16, A14, these are the nodes that'll be in that timeframe of 2028—used for AI, I could see it being like 60-80% of it. No problem.

Dwarkesh Patel
Given the fabs that are currently planned and being built, is that enough for the 1e30, or will we need more?

Dylan Patel
I think so, yeah.

Dwarkesh Patel
Okay, so then the chip goal doesn't make any sense.The chip goal stuff about how we don't have enough compute...

Dylan Patel
No, I think the plans of TSMC on two nanometer and such are quite aggressive for a reason. To be clear, Apple, which has been TSMC's largest customer, does not need how much 2nm capacity they're building. They will not need A16, they will not need A14. Apple doesn't need this shit. Although they did just hire Google's head of system design for TPU. So they are going to make an AI accelerator. But that's besides the point. Apple doesn't need this for their business. They have been 25% or so of TSMC's business for a long time. And when you zone in on just the leading-edge, they've been like more than half of the newest node or 100% of the newest node almost constantly. That paradigm goes away.

Let's say you believe in scaling and you believe the models get better, that the new models will generate amazing productivity gains for the world and so on. If you believe in that world, then TSMC needs to act accordingly and the amount of silicon that gets delivered needs to be there.

So in 2025 and 2026, TSMC is definitely there. Then on a longer timescale, the industry can be ready for it, but it's going to be a constant game of convincing them constantly that they must do this. It's not a simple game. If people work silently, it's not going to happen. They have to see the demonstrated growth over and over and over again across the industry.

Dwarkesh Patel
Who will need to see? Investors or companies?

Dylan Patel
More so, TSMC needs to see NVIDIA volumes continue to grow straight up and Google's volumes continue to grow straight up, and so on down the list. Chips in the near term, next year for example, are less of a constraint than data centers. It's likewise for 2026. The

question for 2027-28... Always when you grow super rapidly, people want to say that's the one bottleneck. Because that's the convenient thing to say.

In 2023 there was a convenient bottleneck, CoWoS (Chip on Wafer on Substrate). The picture has gotten much cloudier. Not clouder but we can see that HBM is a limiter too. CoWoS is as well, CoWoS-L especially. You have data centers, transformers, substations, power generation, batteries, UPSs, CRHs, water cooling stuff. All of this stuff is now a limitation next year and the year after. Fabs will be in 2026-27. Things will get cloudy because the moment you unlock one... Only 10% higher, the next one is the thing. Only 20% higher, the next one is the thing.

Dylan Patel
Today, data centers are like 4-5% of total US power consumption. When you think about it as a percentage of US power, that's not that much. But on the flip side you also consider that all this coal has been curtailed and all these other things. So power is not that crazy on a national basis. On a localized basis, it is, because it's about the delivery of it.

It's the same with the substation transformer supply chains. These companies have operated in an environment where the US power demand has been flat or even slightly down because of efficiency gains. There has been humongous weakening of the industry. Now all of a sudden if you tell that industry, "Your business will triple next year if you can produce more." They can only produce 50% more. Okay, fine. Year after that, now we can produce 3x as much.

You do that to the industry, the US industrial base as well as the Japanese, all across the world can get revitalized much faster than people realize. I truly believe that people can innovate when given the need to. It's one thing if it's a shitty industry where my margins are low and we're not growing really. All of a sudden it's like, "Oh, this is the sexiest time to be alive in power. We're going to do all these different plans and projects and people have all this demand. They're begging me for another percent of efficiency advantage because that gives them another percent to deliver to the chips."

You see all these things happen and innovation is unlocked. You also bring in AI tools, you bring in all these things, innovation will be unlocked. Production capacity can grow, not overnight, but it will in 6 months, 18 months, 3 year timescales. It will grow rapidly. You see the revitalization of these industries.

Getting people to understand that, getting people to believe... because you know, if we pivot to like... Yeah, I'm telling you that Sam's going to raise $50-100 billion dollars because he's telling people he's going to raise this much. He's literally having discussions with sovereigns and Saudi Arabia and the Canadian pension fund and the biggest investors in the world. Of course, Microsoft as well, but he's literally having these discussions because

they're going to drop their next model or they're going to show it off to people and raise that money. This is their plan.

Dwarkesh Patel
If these sites are already planned and...

Dylan Patel
The money's not there.

Dwarkesh Patel
So how do you plan? How do you plan a site without...

Dylan Patel
Today Microsoft is taking on immense credit risk. They've signed these deals with all these companies to do this stuff. But Microsoft doesn't have... I mean, they could pay for it. Microsoft could pay for it on the current timescale. Their CapEx is going from $50-80 billion of direct CapEx, and then another $20 billion across Oracle, CoreWeave, and then like another $10 billion across their data center partners. They can afford that for next year.

This is because Microsoft truly believes in OpenAI. They may have doubts like, "Holy shit, we're taking a lot of credit risk." Obviously, they have to message Wall Street and all these things, but that's affordable for them because they believe they're a great partner to OpenAI. They'll take on all this credit risk.

Now, obviously OpenAI has to deliver. They have to make the next model that's way better. They also have to raise the money. And I think they will. I truly believe from how amazing 4o, how small it is relative to GPT-4... The cost of it is so insanely cheap. It's much cheaper than the API prices lead you to believe. You're like, "Oh, what if you just make a big one?" It's very clear what's going to happen to me on the next jump that they can then raise this money and they can raise this capital from the world.

Jon Y
This is intense. It's very intense.

Dwarkesh Patel
Jon, actually, if he's right, or not him, but in general. If the capabilities are there, the revenue is there...

Dylan Patel
Revenue doesn't matter.

Jon Y
Revenue matters.

Dwarkesh Patel
Is there any part of that picture that still seems wrong to you in terms of displacing so much of TSMC production, wafers and power and so forth? Does any part of that seem wrong to you?

Jon Y
I can only speak to the semiconductors part, even though I'm not an expert. I think TSMC can do it. They'll do it. I just wonder though... he's right in that 2024-25 is covered. But 2026-27 is that critical point where you have to say, can the semiconductor industry and the rest of the industry be convinced that this is where the money is? That means, is there money by 2024-25?

Dwarkesh Patel
How much revenue do you think the AI industry as a whole needs by 2025 in order to keep scaling?

Dylan Patel
Doesn't matter.

Jon Y
Compared to smartphones. I know he says it doesn't matter.

Dylan Patel
I'll get to why.

Dwarkesh Patel
What are smartphones at? Like Apple's revenue is like $200 billion. So like...

Jon Y
Yeah, it needs to be another smartphone-size opportunity, right? Even the smartphone industry didn't drive this sort of growth. It's crazy. Don't you think? The only thing I can really perceive... AI Girlfriend. You know what I mean.

Dylan Patel
No, I want a real one, dammit. There's a few things. The return on invested capital for all of the big tech firms is up since 2022. Therefore, it's clear as day that investing in AI has been fruitful so far for the big tech firms just based on return on invested capital. Financially, you look at Meta's, you look at Microsoft's, you look at Amazon's, you look at Google's. The return on invested capital is up since 2022.

Dwarkesh Patel
On AI in particular?

Dylan Patel
No, just generally as a company. Now, obviously there's other factors here. Like what is Meta's ad efficiency? How much of that is AI, right?

Jon Y
That's super messy.

Dylan Patel
But here's the other thing, this is Pascal's wager, right? This is a matrix of like, do you believe in God? Yes or no. If you believe in God and God's real and you go to heaven, that's great. That's fine. Whatever. If you don't believe in God and God is real, then you're going to hell.

Dwarkesh Patel
This is the deep technical analysis you'll subscribe to SemiAnalysis for.

Jon Y
Can you imagine what happens to the stock if Satya starts talking about Pascal's wager?

Dylan Patel
But this is psychologically what's happening, right? Satya said it on his earnings call. The risk of under-investing is worse than the risk of over-investing. He has said this word for word. This is Pascal's wager.

I must believe I am AGI-pilled because if I'm not and my competitor does it, I'm absolutely fucked.

Jon Y
Other than Zuck, who seems pretty convinced...

Dylan Patel
No, Sundar said this on the earnings call. So Zuck said it. Sundar said it. Satya's actions on credit risk for Microsoft do it. He's very good at PR and messaging, so he hasn't said it so openly.

Sam believes it. Dario believes it. You look across these tech titans, they believe it. Then you look at the capital holders. The UAE believes it. Saudi believes it.

Dwarkesh Patel
How do you know the UAE and Saudi believe it?

Dylan Patel
All these major companies and capital holders also believe it because they're putting their money here.

Jon Y
But it won't last, it can't last unless there's money coming in somewhere.

Dylan Patel
Correct, correct, but then the question is... The simple truth is that GPT-4 costs like $500 million dollars to train. It has generated billions in recurring revenue. In the meantime, OpenAI raised $10 billion or $13 billion and is building a model that costs that much, effectively.

Obviously they're not making money. What happens when they do it again? They release and show GPT-5 with whatever capabilities that make everyone in the world go, "Holy fuck." Obviously the revenue takes time after you release the model to show up. You still have only a few billion dollars or $5 billion of revenue at run rate.

You just raised $50-100 billion dollars because everyone sees this like, "Holy fuck, this is going to generate tens of billions of revenue." But that tens of billions takes time to flow in. It's not an immediate click. But the time where Sam can convince, not just Sam... people's decisions to spend the money are being made then.

Therefore, you look at the data centers people are building. You don't have to spend most of the money to build the data center. Most of the money's the chips, but you're already committed to having so much data center capacity by 2027 or 2026 that you're never gonna need to build a data center again for like 3-5 years if AI is not real.

Basically, that's what all their actions are. Or I can spend over a hundred billion dollars on chips in 2026 and I can spend over a hundred billion dollars on chips in 2027. These are the actions people are taking and the lag on revenue versus when you spend the money or raise the money, there's a lag on this.

You don't necessarily need the revenue in 2025 to support this. You don't need the revenue in 2026 to support this. You need the revenue in 2025-2026 to support the $10 billion that OpenAI spent in '23, or Microsoft spent in 2023 and early 2024 to build the cluster, which model they trained in mid 2024, which they then released at the end of 2024, which then started generating revenue in 2025-2026.

Jon Y
The only thing I can say is that you look at a chart with three points on a graph: GPT-1, 2, 3, and you're like...

Dwarkesh Patel
Even that graph... The investment you have to make in GPT-4 over GPT-3 is 100X. The investment you had to make in GPT-5 over GPT-4 is 100X. Currently the ROI could be positive—this very well could be true, I think it will be true—but the revenue has to increase exponentially.

Jon Y
Of course. I agree with you. But I also agree with Dylan. It can be achieved. ROI, TSMC does this. It invests $16 billion. It expects ROI. It does that. I understand that. That's fine. Lag all that. The thing that I don't expect is that GPT-5 is not here. It's all dependent on GPT-5 being good. If GPT-5 sucks, if GPT-5 looks like it doesn't blow people's socks off, this is all void.

Dylan Patel
What kind of socks are you wearing, bro? Show them.

Jon Y
AWS. GPT-5 is not here. GPT-5 is late. We don't know.

Dylan Patel
I don't think it's late.

Jon Y
I think it's late.

Dwarkesh Patel
Okay. I want to zoom out and go back to the end of the decade picture again.

Dylan Patel
We've already lost Jon.

Jon Y
We've already accepted GPT-5 would be good? Hello?

Dwarkesh Patel
You gotta, you know?

Dylan Patel
Life is so much more fun when you just are delusionally...

Jon Y
We're just ripping bong hits, are we?

Dylan Patel
When you feel the AGI, you feel your soul.

Jon Y
This is why I don't live in San Francisco.

Dylan Patel
I have tremendous belief in the GPT-5 area.

Dwarkesh Patel
Why?

Dylan Patel
Because of what we've seen already. The public signs all show that this is very much the case. What we see beyond that is more questionable and I'm not sure because I don't know. We'll see how much they progress.

If things continue to improve, life continues to radically get reshaped for many people. Every time you increment up the intelligence, the amount of usage of it grows hugely. Every time you increment the cost down of that amount of intelligence, the amount of usage increases massively. As you continue to push that curve out, that's what really matters.

It doesn't need to be today. It doesn't need to be revenue vs. how much CapEx. In any time in the next few years, it just needs to be, did that last humongous chunk of CapEx make sense for OpenAI or whoever the leader was? How does that then flow through? Or were they able to convince enough people that they can raise this much money?

You think Elon's tapped out of his network with raising $6 billion? No. xAI is going to be able to raise $30+ billion easily. You think Sam's tapped out? You think Anthropic's tapped out? Anthropic's barely even diluted the company relatively. There's a lot of capital to be raised. Call it FOMO if you want, but during the dot-com bubble, the private industry flew through like $150 billion a year. We're nowhere close to that yet.

We're not even close to the dot-com bubble. Why would this bubble not be bigger? You go back to the prior bubbles: PC bubble, semiconductor bubble, mechatronics bubble. Throughout the US, each bubble was smaller. I don't know if you call it a bubble or not. Why wouldn't this one be bigger?

Dwarkesh Patel
How many billions of dollars a year is this bubble right now?

Dylan Patel
For private capital? It's like $55-60 billion so far for this year. It can go much higher. I think it will next year.

Dwarkesh Patel
Let me think about this.

Jon Y
You need another bong rip.

Dylan Patel
At least like finishing up and looping into the next question... Prior bubbles also didn't have the most profitable companies that humanity has ever created investing and they were debt financed. This is not debt financed yet. That's the last little point on that one. Whereas the 90s bubble was very debt financed.

Jon Y
That was disastrous for those companies.

Dylan Patel
Yeah, sure. But so much was built. You've got to blow a bubble to get real stuff to be built.

Dwarkesh Patel
It is an interesting analogy. Even though the dot-com bubble obviously burst and a lot of companies went bankrupt, they in fact did lay out the infrastructure that enabled the web and everything. You could imagine an AI... A lot of the foundation model companies or whatever, a bunch of companies will go bankrupt, but they will enable the singularity.

Jon Y
At the turn of the 1990s, there was an immense amount of money invested in things like MEMS and optical technologies because everyone expected the fiber bubble to continue. That all ended in 2003 or 2002.

Dylan Patel
It started in '94?

Jon Y
There hasn't been revitalization since. You could risk the possibility of a...

Dylan Patel
Bro, Lumen, one of the companies that's doing the fiber build out for Microsoft, its stock like fucking 4x'd last month, or this month.

Jon Y
How'd it do from 2002 to 2024?

Dylan Patel
Oh no, horrible, horrible, but we're gonna rip, baby. Rip that bong, baby!

Jon Y
You could freeze AI for another two decades.

Dylan Patel
Sure, sure, it's possible. Or people can see a badass demo from GPT-5, a slight release, they raise a fuckload of money. It could even be like a Devin-like demo, where it's like complete bullshit, but it's fine.

Jon Y
Edit that out! Edit that out!

Dylan Patel
No, it's fine. I don't really care. The capital's going to flow in. Now whether it deflates or not is an irrelevant concern in the near term because you operate in a world where it is happening. What is that Warren Buffett quote? I don't even know if it's Warren Buffett.

Jon Y
You don't know who's swimming naked until the tide goes out?

Dylan Patel
No, no, no. The one about how the market is delusional for longer than you can remain solvent, or something like that.

Jon Y
Oh, that's not Buffett.

Dylan Patel
That's not Buffett?

Jon Y
That's John Maynard Keynes.

Dylan Patel
Oh shit, that's that old? Okay. So Keynes said it. So this is the world you're operating in. It doesn't matter what exactly happens. There'll be ebbs and flows, but that's the world you're operating in.

Jon Y
I reckon that if the AI bubble pops, each one of these CEOs lose their jobs.

Dylan Patel
Sure. Or if you don't invest and you lose, it's a Pascalian wager. That's much worse. Across decades, the largest company at the end of each decade of the largest companies, that list changes a lot. And these companies are the most profitable companies ever. Are they going to let themselves lose it? Or are they going to go for it? They have one shot, one opportunity to make themselves into... the whole Eminem song.

Dwarkesh Patel
I want to hear the story of how both of you started your businesses, the thing you're doing now. Jon, how did it begin? What were you doing when you started the YouTube channel?

Dylan Patel
It's all about your textile company?

Jon Y
Oh, my God. No way.

Dylan Patel
Please, please.

Dwarkesh Patel
Wait, is he joking?

Dylan Patel
If he doesn't want to, we'll talk about it later.

Jon Y
The story's famous. I've told it a million times. Asianometry started off as a tourist channel. I moved to Taiwan for work.

Dylan Patel
DoIng what?

Jon Y
I was working in cameras.

Dylan Patel
What was the other company you started?

Jon Y
It tells too much about me. I worked in cameras and then basically I went to Japan with my mom. My mom was like, "Hey, what are you doing in Taiwan? I don't know what you're doing." I was like, "All right, mom, I will go back to Taiwan and I'll make stuff for you." I made videos. I would go to the Chiang Kai-shek Park and be like, "Hi, mom, this park was this, this."

Eventually, you run out of stuff. But then it's a pretty smooth transition from that into Chinese history, Taiwanese history. People started calling me Chinanometry. I didn't like that. So I moved to other parts of Asia.

Dwarkesh Patel
What year did people start watching your videos? Let's say like a thousand views per video or something?

Jon Y
Oh, my gosh. I started the channel in 2017 and it wasn't until 2018 or 2019 that it actually... I labored on for the first three years with no one watching. I got like 200 views and I'd be like, "Oh, this is great."

Dwarkesh Patel
Were the videos basically like the ones you have now? Sorry, backing up for the audience who might not know, I imagine basically everybody knows Asianometry, but if you don't it's the most popular channel about semiconductors, Asian business history, business history in general, geopolitics, history and so forth. Honestly, I've done research for different AI guests and different things I'm trying to understand. How does hardware work? How does AI work?

Dylan Patel
How does a zipper work? Did you watch that video? I think it was a span of three videos. It's like the Russian oil industry in the 1980s and how it funded everything and then when it collapsed, they were absolutely fucked. Then the next video was like, the zipper monopoly in Japan.

Jon Y
Not a monopoly.

Dylan Patel
The next video was about ASML.

Jon Y
Not a monopoly. Strong holding in a mid-tier size. They're like the luxury zipper makers. Asianometry is always just stuff I'm interested in. I'm interested in a whole bunch of different stuff. Then the channel, for some reason people started watching the stuff I do. I still have no idea why. To be honest, I still feel like a fraud. I sit in front of Dylan and I feel like a legit fraud, especially when he starts talking about 60,000 wafers and all that. I feel like I should know this but in the end, I just try my best to bring interesting stories out.

Dwarkesh Patel
How do you make a video every single week? These are like...

Jon Y
Two a week?

Dylan Patel
You know how long he had a full-time job?

Jon Y
Five years, six years.

Dwarkesh Patel
While doing this?

Dylan Patel
Sorry, a textile business and a full-time job. Wait, no it's a full-time job, textile business, and Asianometry for a long long time.

Jon Y
I literally just gave up the textile business this year.

Dwarkesh Patel
How are you doing research and making a video twice a week? I don't know. I do these, I'm just fucking talking. This is all I do. I do this once every like two weeks.

Dylan Patel
The difference is, Dwarkesh, you go to SF Bay Area parties constantly and Jon is locked in. He's like locked in 24/7.

Dwarkesh Patel
He's got the TSMC work ethic and I've got the Intel work ethic.

Jon Y
If I don't... I got the Huawei ethic. If I do not finish this video, my family will be pillaged.

Dylan Patel
He actually gets really stressed about it, not doing something on his schedule.

Jon Y
I do two videos per week, I write them both simultaneously.

Dwarkesh Patel
How are you scouting out future topics you want to do? You just pick up random articles, books, whatever. If you find it interesting, you make a video about it?

Jon Y
Sometimes what I'll do is Google a country, I'll Google an industry, and I'll Google what a country is exporting now and what it used to export. I compare that and I say, "That's my video." Sometimes it's also just as simple as, "I should do a video about YKK."

Dylan Patel
This zipper is nice. I should do a video about it.

Jon Y
I do. It literally is...

Dwarkesh Patel
Do you keep a list? "Here's the next one. Here's the one after that."

Jon Y
I have a long list of ideas. Sometimes it's as vague as Japanese whiskey. I have no idea what Japanese whiskey is about. I heard about it before. I watched that movie. So I was just like, "Okay, I should do a video about that."

Dwarkesh Patel
How many research topics do you have on the back burner, basically? I'm talking about something where you're reading about it constantly and then in a month or so you're like, "I'll make a video about it."

Jon Y
I just finished a video about how IBM lost the PC. Right now I'm unstressing about that. But I'll kind of move right on to... The videos do kind of lead into others. Right now this one is about how IBM lost the PC. Now what's next is how Compaq collapsed, how the wave destroyed Compaq. So I'll do that. At the same time, I'm dual lining a video about qubits. I'm dual lining a video about directed self-assembly for semiconductor manufacturing, which I'll read a lot of Dylan's work for. But then a lot of that is in the back of my head. I'm producing it as I go. Dylan, how do you work?

Dwarkesh Patel
How does one go from Reddit shitposter to running a semiconductor research and consulting firm? Let's start with the shitposting.

Dylan Patel
It's a long line. I had immigrant parents and I grew up in rural Georgia. When I was seven, I begged for an Xbox and when I was eight I got it. It was the 360. They had a manufacturing defect called the red ring of death. There are a variety of fixes that I tried like putting a wet towel around the Xbox, something called the penny trick. Those all didn't work, my Xbox still didn't work. My cousin was coming next weekend and he's like two years older than me. I look up to him. He's in between my brother and me but I'm like, "Oh, no, no, we're friends. You don't like my brother as much as you like me." My brother's more of a jock-y type. It didn't matter. He didn't really care that the Xbox was broken. He's like, "You better fix it though. Otherwise parents will be pissed."

I figured out how to fix it online. I tried a variety of fixes, ended up shorting the temperature sensor. That worked for long enough until Microsoft did the recall. But in that, I learned how to do it out of necessity on the forums. I was a nerdy kid, so I liked games, but whatever. There was no other outlet so once I was like, "Holy shit, this is Pandora's box what just got opened up," then I just shitposted on the forums constantly. I did that for many, many years and then I ended up moderating all sorts of Reddits when I was a tween and teenager. Then as soon as I started making money... I grew up in a family business, but I didn't get paid for working, of course, like yourself. But as soon as I started making money at my internships, I was like 18 or 19, I started making money. I started investing in semiconductors. I was like, "Of course, this is the shit I like."

By the way, the whole way through as technology progressed, especially mobile, it went from very shitty chips and phones to very advanced. Every generation they'd add something and I'd read every comment, I'd read every technical post about it. Also, I was interested in all the history around that technology and who's in the supply chain and just kept building and building and building. I went to college and did data science type stuff. I went to work on hurricane/earthquake/wildfire simulation and stuff for a financial company. During college I wasn't shitposting on the internet as much. I was still posting some, but I was

following the stocks and all these sorts of things, the supply chain, all the way from the tool equipment companies. The reason I liked those is because all this technology, it's made by them.

Dwarkesh Patel
Did you have friends in person who were into this shit, or was it just online?

Dylan Patel
I made friends on the internet.

Jon Y
Oh, that's dangerous.

Dylan Patel
I've only ever had like literally one bad experience and that was just because he was drugged out.

Dwarkesh Patel
One bad experience online?

Dylan Patel
Meeting someone from the internet in person. Everyone else has been genuinely... You have enough filtering before that point. Even if they're hyper mega autistic, it's cool. I am too. No, I'm just kidding. You go through the layers and you look at the economic angle, you look at the technical angle, you read a bunch of books. You can just buy engineering textbooks and read them. What's stopping you? If you bang your head against the wall, you learn it.

Dwarkesh Patel
While you were doing this, did you expect to work on this at some point or was it just pure interest?

Dylan Patel
No, it was an obsessive hobby of many years and it pivoted all around. At some point I really liked gaming. Then I moved into phones and rooting them and underclocking them, and the chips there, and screens and cameras, and then back to gaming, and then to data center stuff because that was where the most advanced stuff was happening. I liked all sorts of telecom stuff for a little bit. It bounced all around, but generally it was in computing hardware.

I did data science. I said I did AI when I interviewed but it was like bullshit multivariable regression, whatever. It was simulations of hurricanes, earthquakes, and wildfires for financial reasons. I had a job for three years after college. I was posting. I had a blog, an

anonymous blog for a long time. I'd even made some YouTube videos and stuff. Most of that stuff is scrubbed off the internet, including Internet Archive, because I asked them to remove it.

In 2020, I quiet quit my job and started shitposting more seriously on the internet. I moved out of my apartment and started traveling through the US. I went to all the national parks, in my truck/tent. I also stayed in hotels and motels for like three or four days a week. I started posting more frequently on the internet. I'd already had some small consulting arrangements in the past. But it really started to pick up in mid-2020, consulting arrangements from the internet from my persona.

Dwarkesh Patel
What kinds of people? Was it investors, hardware companies?

Dylan Patel
It was people who weren't in hardware that wanted to know about hardware. It would be some investors. Some VCs did it, some public market folks. There were times where companies would ask about three layers up in the stack. They saw me write some random posts and like “Hey, can we…” There's all sorts of random stuff. It was really small money.

Then in 2020, it really picked up and I thought, “Why don't I just arbitrarily make the price way higher?” And it worked. I made a new newsletter as well. I kept posting. The quality kept getting better because people would read it and be like, “this is fucking retarded. Here’s what’s actually right” over more than a decade.

Towards the end of 2021, I made a paid post because someone didn't pay for a report or whatever. It was about photoresist and the developments in that industry, which is the stuff you put on top of the wafer before you put in the lithography tool. It did great. I went to sleep that night and woke up the next day with 40 paid subscriptions. I thought, "What? Okay, let's keep going." I started posting more paid content, partially free, partially paid. I did all sorts of stuff covering advanced packaging, chips, data center stuff, and AI chips. It was all sorts of stuff I was interested in and thought was interesting.

I always bridged economically—because I've read all the company's earnings since I was 18 and I'm 28 now—all the way through to the technical stuff that I could. In 2022, I also started going to every conference I could. I go to about 40 conferences a year, not trade show type conferences, but technical ones: chip architecture, photoresist, AI and NeurIPS, ICML, and so on.

Dwarkesh Patel
How many conferences do you go to a year?

Dylan Patel
Like 40.

Dwarkesh Patel
So you basically live at conferences?

Dylan Patel
Yeah. I've been a digital nomad since 2020. I've basically stopped and I moved to SF now, kind of not really.

Jon Y
You can't say that the California government...

Dylan Patel
I don't live in SF, come on... But I basically do now.

Dwarkesh Patel
California Internal Revenue Service.

Dylan Patel
Do not joke about those guys. Seriously, don't joke about this.

Jon Y
They're going to send you a clip of this podcast and be like, "40% please."

Dylan Patel
I am in San Francisco sub four months a year contiguously, exactly 100 days or whatever it is, 179 days. Let's go, right? Over the full course of the year...but no, I go to every conference, make connections with all these very technical things like international electron device manufacturing, lithography and advanced patterning, very large scale integration. You have circuits conferences. You just go to every single layer of the stack. It's so siloed, there's tens of millions of people that work in this industry.

But you go to every single one. You try and understand the presentations. You do the required reading. You look at the economics of it. You are just curious and want to learn. You can start to build up more and more. The content got better and what I followed got better. Then I started hiring people in mid-2022, as well. I got people in different layers of the stack.

Now today, almost every hyperscaler is a customer, not for the newsletter but for the data we sell. Many major semiconductor companies, many investors, all these people are customers of the data and stuff we sell. The company has people all the way from

ex-CYMER, ex-ASML, all the way to ex-Microsoft and an AI company. Through the stratification, now there's 14 people here in the company and all across the US, Japan, Taiwan, Singapore, France. It's all over the world and across many ranges of... You have ex-hedge funds as well. You kind of have this amalgamation of tech and finance expertise. We just do the best work there, I think.

Jon Y
Are you still talking about a monstrosity?

Dylan Patel
An unholy concoction. We have data analysis, consulting, etc. for anyone who really wants to get deeper into this. We can talk about people building big data centers, but how many chips are being made in every quarter of what kind for each company? What are the subcomponents of these chips? What are the subcomponents of the servers? We try to track all of that. We follow every server manufacturer, every component manufacturer, every cable manufacturer, all the way down the stack, tool manufacturer. We know how much is being sold, where and how, and project out all the way out to like, "Hey, where's every single data center? What is the pace that it's being built out?" This is the sort of data we want to have and sell. The validation is that hyperscalers purchase it and they like it a lot. AI companies do and semiconductor companies do. How it got there to where it is, is just like, "Just try and do the best and try to be the best."

Dwarkesh Patel
If you were an entrepreneur who's like, "I want to get involved in the hardware chain somewhere..." If you could start a business today, somewhere in the stack, what would you pick?

Dylan Patel
Jon, tell him about your textile business.

Jon Y
I'd work in memory, something in memory. The concept is that they have to hold immense amounts of memory. I think memory already is tapped technologically. HBM exists because of limitations in DRAM. I think it's fundamentally... We've forgotten it because it is a commodity, but we shouldn't. I think breaking memory would change the world in that scenario.

Dylan Patel
The context here is that Moore's Law was predicted in 1965. Intel was founded in 1968 and released their first memory chips in 1969 and 1970. So a lot of Moore's Law was about memory. The memory industry followed Moore's law up until 2012 when it stopped. Tt became very incremental gains since then. Whereas logic has continued and people are

like, "Oh, it's dying. It's slowing down." At least there's still a little bit of coming. It's still more than 10-15% a year CAGR of growth and density and cost improvement. Memory is literally since 2012, really bad.

When you think about the cost of memory... It's been considered a commodity. But memory integration with accelerators, this is something... I don't know if you can be an entrepreneur here though. That's the real challenge. Because you have to manufacture at some really absurdly large scale or design something in an industry that does not allow you to make custom memory devices.

Jon Y
Or use materials that don't work that way.

Dylan Patel
So there's a lot of work there. I don't necessarily agree with you. But I do agree it's one of the most important things for people to invest in.

It's really about where you are good at. Where can you vibe? Where can you enjoy your work and be productive in society? There are a thousand different layers of the abstraction stack. Where can you make it more efficient? Where can you utilize AI to build better and make everything more efficient in the world and produce more bounty and iterate feedback loop? There's more opportunity today than any other time in human history, in my view. Just go out there and try. What engages you? Because if you're interested in it, you'll work harder.

If you have a passion for copper wires... I promise to God, if you make the best copper wires, you'll make a shitload of money. If you have a passion for B2B SaaS, I promise to God, you'll make fuck loads of money. I don't like B2B SaaS, but whatever you have a passion for. Just work your ass off, try and innovate, bring AI into it. Try and use AI to make yourself more efficient and make everything more efficient. I promise, you will be successful.

That's really the view. It's not necessarily that there's one specific spot because every layer of the supply chain has... You go to the conferences. You go to talk to the experts there. It's like, "Dude, this is the stuff that's breaking and we could innovate in this way. These five extraction layers, we could innovate this way." Yeah, do it. There's so many layers where we're not at the Pareto optimal. There's so much more to go in terms of innovation and inefficiency.

Dwarkesh Patel
That's a great place to close. Dylan, Jon, thank you so much for coming on the podcast. I'll just give people the reminder, Dylan Patel, semianalysis.com. That's where you can find all the technical breakdowns that we've been discussing today. Asianometry YouTube channel,

everybody will already be aware of Asianometry, but anyways, thanks so much for doing this. It was a lot of fun.

Dylan Patel
Thank you.

Jon Y
Yeah, thank you.