Why Hardware-Software Co-Design Is AI's Real 100x: Dylan Patel of SemiAnalysis

69 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=F6D_AIY8QYU
https://www.youtube.com/watch?v=f6D_aiy8qyU

SUMMARY

Dylan Patel discusses the evolution and current state of the semiconductor industry, highlighting the dynamic interplay between technology, economics, and market demand. He emphasizes the importance of co-optimization in hardware and software, the competitive landscape among AI model providers, and the ongoing compute crunch driven by skyrocketing demand for AI capabilities.

- *Semi Analysis has grown significantly, focusing on semiconductor research and consulting, with a rumored revenue of over $100 million.*
- *The company thrives on a mix of technologists and finance experts, fostering debates on technology versus cost.*
- *Patel's background includes early experiences with hardware and a deep understanding of the semiconductor supply chain, leading to the founding of Semi Analysis.*
- *The AI landscape is rapidly evolving, with inference performance benchmarks needing to adapt to constant model updates and optimizations.*
- *There is a notable disparity in the value of compute resources, with companies like Google and Amazon having different efficiencies and pricing structures for their AI workloads.*
- *The Neocloud opportunity exists due to hyperscalers' inefficiencies in AI-focused compute, allowing smaller players to thrive.*
- *Jensen Huang of Nvidia is strategically investing in diverse AI labs to prevent a monopolistic environment dominated by hyperscalers.*
- *The compute crunch is expected to persist as demand for AI capabilities continues to outpace supply, despite ongoing data center buildouts.*

I think it's really fun inside of semi analysis because we have 90 people and like a big chunk of them are technologists engineers across the whole supply chain. Um, and then a big chunk is people who are formerly at hedge funds. And you see these arguments like people are like, "Oh, well that doesn't matter." And it's like, then someone's like, "Well, but cost." And then someone the engineers like, "No, no, no, but this technology is the coolest." And you see this you see this organically like fight it out. Um, and and we're pretty informal and you know, given the fact that I was a for moderator, you can imagine what the the enjoying it. You don't wrestle with a pig because a pig enjoys it, right?

Exactly. We're here in the semi analysis office with Dylan Patel. You know, I'm Sean from Sequoia. My partner Sonia Huang. It's pretty insane what you've done. Semi semis 5 years ago were not very sexy in the west. They were sexy in the east but uh people here in the west had kind of forgotten about them. You did not forget about them though. You went very long. You created probably the premier research company in the space that's been

educating the world and you know the state of the art from very technical details to supply chain you know to the bigger picture. Um there's rumors that semi analysis recently passed 100 million of revenue. I don't know how accurate those are. Whatever the numbers are you guys are crushing.

>> It's it's as accurate as the information is. Yeah. Cool. You know, you never you never know. Um there's also rumors that you might start a venture fund like you know I I hear all the time in the ecosystem people wanting you know affiliation with semi analysis. You you've built this trusted brand and so whatever you do it's working. This clearly like just the beginning of the journey for you. Congratulations all of that. But how did this happen? Like how did you first question is like what is the background? How did you kind of get to where you are now? Well, well, when I was a young boy in the, you know, coming out of the womb. No. So, so, okay. So, I grew up in like a small business. My parents had a motel. We lived in the motel. We laid our gas station. So, you know, uh I was selling. You know, I joke a lot of times the first neural network I trained was uh racially and and visually profiling people based on when they enter the gas station, which cigarette to uh pick. Right. Basically, you know, the cigarettes were all extrudeed across the top and I was too short to actually like, you know, reach them and technically it wasn't legal to sell cigarettes at that age, but whatever. I I had to move the step stool over to the right area.

>> I started working my first job was before it was legal, too. So, but it's good experience. >> Well, I didn't get paid, right? It's a family business. Same. >> Same. Um, but yeah, we had our motel and then across the street was our gas station. So, you know, sometimes, you know, you know, someone would walk in and so like if a old white lady with curly hair walked in, I'd move the ladder or the step stool over to where the camels are. And if you know and you know different different age, demographic, profession, you know, race, etc. I would move the step stool over and I joke this is the first neural network I trained because if I waited for them to tell me I'd have to like move it over and then I'd step up versus like just being ready. Um so you know menthols versus, you know, 100 slims and all these things, you know, I joke that's the first neural network I trained. But I grew up in family businesses. Um lived in a motel and um it all really goes back to when I was like, you know, it was my 8th birthday.

Um, my birthday's in May. Um, and it was April when the Xbox 360 was announced. Um, for my birthday, I didn't ask for the Xbox or I didn't ask for a birthday gift. My parents asked what I wanted. I asked for it for Christmas. Uh, we celebrated Christmas, but there was no way, at least at the time, I thought there was no way they would ask would give me the Xbox 360 for Christmas and so I got it for I asked for my birthday for tab for Christmas. Anyways, Christmas comes around, I get it. Um, you know, fast forward a couple months, my cousin who lives in Alabama, they also lived in a motel, was going to come over for spring break, um, for his spring break and we were going to just hang out at my house and he's in between me and my older brother in age, brother's a bit more jockey. Um, so he didn't really care too much about the Xbox. He played sometimes, but he didn't really care. Um, but my cousin, you know, I wanted to think I'm him to think I was cool, right? You know, so I bragged many times on the phone. I was like, "Yeah, I got an Xbox." And then the Xbox broke. There was something there's a hardware defect called the red ring of death. Um, but long story short, I had to open it up and, you know, short the temperature sensor and it fixed it.

Um, but I there was many other tricks I tried first and none of them worked. Um, and so that's sort of how I like got into hardware. I was like open Pandora's box. By the time I was 12, I was like on these forums a lot reading uh posting a lot and this is around the time when Reddit ate all other forms and so I became a moderator of you know Android and Apple and Google as well as like hardware and was watch you know looking at Intel, Nvidia and AMD and all these other forms right? I was build a PC. All these forms I was watching, reading, posting a lot, but some of them I was moderating a lot. Um, and so, you know, smartphones, watching smartphones develop from like very simple to speed racing to being technologically more advanced than PCs, um, in many ways architecturally and and same with like, you know, all the in GPUs like just tracking and watching at that, reading every comment. Um, always having the economic tinge because I grew up in a small business. So, I was always looking at the economics, right? There was a time where all the like I'd say neck beards on the internet loved AMD GPUs and like I personally had bought an AMD GPU too because price performance but then when it came down to like what's technically better I'd always be like no no no Nvidia is better because they use a smaller chip to get you know better performance at better power efficiencies and their margins better and and so like I would always like talk about how Nvidia's margins were better than than AMD's in the GPU landscape and so it's like very fun.

>> And you were 12 at the time. I started moderating when I was 12, but this is all through my teenage tween age and high school years, right? >> Do you have any other weird hobbies or was it just semis? >> I played a ton of Starcraft. At one point, I was grandmaster on the North American ladder. Starcraft 2. >> Very serious. >> So, you've gotten just obsessively good at multiple things. >> Yeah. I mean, it's it's it's obsession is good.

>> How were your grades? >> Um, they were decent. Um, I would say like I had mostly A's, but they're classes that I like were thought were really boring or, you know, I just didn't enjoy. Um, like Spanish I got like not the greatest grades. Um, you know, but but it was like I speak fluent Spanish by the way, so it's really dumb. But like it's just sort of >> maybe that's why you didn't get a good grade.

>> I didn't learn Spanish till later to be fair. But yeah, so sort of my grades were fine, right? Like I mean I like they were fine enough for Asian parents. >> I was better than most of school but you know it wasn't like you know tryh hard maxing for like you know all A's. >> Okay. So you're very much a student of the internet then this is how you how you develop this expertise. At what point do you decide to start semi analysis and what's been the biggest surprise since starting the company?

>> Yeah so I went to school I got a few degrees in stuff that wasn't related to semiconductors. Um was a quant for two years at a small quant risk firm. Um and then basically, you know, there's a culmination of events that happened, right? One was that my um you know, sort of like I got screwed out of a bonus. I I'd made my company many millions of revenue of risk-free revenue because I exploited like a risk, you know, thing in the market. Um you know, I think well over 10 million and they then someone else took credit for my work and all this sort of stuff. But eventually I did get rightsized. But you know, I lost a social contract with the company I was working with. Um add some you know, my grand my grandparents grew up in my house with us, right? are in the motel with us. Uh they lived with us and so you know very close with them and my grandmother got dementia and she forgot who I was and she she fell down some stairs and had like a tragic accident and passed

away. So all of that happened in early 2020. Um additionally there were some like you know girl things and so you know there's a few things that happened that made me like kind of very sad. Um and and and so all of those things sort of culminated. Then co happened and my brother's like dude just just come stay with me. He lived in Nashville so I came and stayed with him in Nashville. We were like, "Oh, lockdowns will be a few weeks. You can stay with me while they happen and then you can go back home and you know, whatever." Famous last words.

Lockdowns lasted much longer. But, you know, living with my brother for a few months, you know, was like sort of like, okay, didn't know what I was doing. I was now at my brother's home. Um, everything was his rules. You know, sort of like, you know, him and him and his fiance at the time, now wife, you know, were like there. And so, like, I basically had to tiptoe around, but I didn't care about my job. And so, I was like posting even more than normal. I'd always been posting a lot on the internet. I'd always been trading stocks a lot, but like I made a lot of money shorting COVID and long in COVID and like all this stuff. Semiconductor shortages happened around then too. And anyways, I was like very much obsessed with posting and and things like that.

And eventually um around that time someone I got into an argument with someone on the internet and they doxed me, right? They they publicly revealed my identity for my anonymous account. And at the time I was like, "Oh no, I scared. I stopped posting for like three weeks and I was like, what am I doing? Why do I care?" So then I just started posting under I had like had like blogs and stuff as well. I made a real blog, semi- analysis, and on my 24th birthday, I posted um you know, two blogs and and then from there it just like it was not a newsletter, but I got so much traction because now instead of posting on an anonymous name, it was a real name and I put a lot more effort into those two posts than I usually did. Instead of like posting on the internet, it was like real effort into the blog. Um you can actually go back and read those if you want. They're they're not that great, but you know, they were they were good for the time. They were the best stuff you could find on the internet about semis. Um, and and I just kept posting, posting, posting. I started getting a lot of consulting business.

You know, 2020, I also sort of I was again crashing out. Didn't know what I wanted to do. So, I uh packed everything up or sort of I I I took my truck, I bought a tent that fits on the back of the tent, truck, um, bought a air mattress, whatever, and would like and drove around all these national parks all around America. And so, like two or three or four days of the week, I'd stay in a random motel where I negotiated the price to be like $30 a night for a room.

And I was work on something else stuff. And then the weekends I'd read books and oftentimes read textbooks um while in some random national park or hiking and listen to audiobooks um about semiconductors about AI about all the things that I cared a lot about and got way more educated over these six months where I'm just like going to every national park. Um and the whole time I was I was alone the whole time I was posting blogs. Um everyone was like DD what the are you doing >> pre-larink or the very early days of Starling?

>> Pre-star link pre-tar link. Um yeah, so it was like very much like what are you doing? Um I travel around Latam again like for for a year initially with my friend and then with my ex you know for you know about a

year and then I go then 22 23 24 end of 21 22 23 and 24 I'm completely I'm still completely homeless since mid2020 right um but I'm traveling around to every conference in the world.

I go to 40 plus conferences a year no matter where in the supply chain it is. I'm like oh that looks interesting. I guess I'll go to that. And I'm like I went to one conference like wow this is amazing. you get to talk to the experts and they just like they they're they they're going to talk to you because and then you're so excited and in the case of semiconductors everyone's a boomer so it's like it's great to like you know they're like they don't see young people who are like excited about it so they're really happy to tell stuff and so you just have to ask on this was there like a part of the supply chain or one of these conferences that you know particularly changed your view of the semi-world or that you felt then or feel now is particularly underrated. I think I think the trade shows like r and conferences range really widely. Um obviously some of the you know the ones I have the most fun at you know include NERPS. Why why is that? Because it's 20,000 AI researchers and they're generally in my distribution of age range. So it's like a lot of fun but they're also like leading AI researchers and it's a lot of fun and lot you learn a lot. Um there's also a lot of parties and then it ranges all the way to like you know there's random chemical conference in Japan where it's 300 Japanese dudes. It's like 20 guys from ASML, 20 guys from TSMC, 20 guys from Intel, and those are the only people who speak English. Uh, everyone else speaks only Japanese, and you're like, h, I guess they're still pretty interesting and fun. I think I think like one thing that I have like a skill set of is like I'm able to bond with anyone regardless of their background and like who they are. I'm able to talk to them, find something interesting to talk about.

Oftentimes, it's the tech stuff, but, you know, it's it's and so I think like the most interesting conferences are oftentimes like, you know, the really big ones because that's where the biggest stuff is happening. Um but I think the niches that are really really exciting is like you know SPIE um so there's IE which is international electrical engineering something um and there's SPI which is another ecosystem. SPIE conferences are super super deep in details. Every single one that I went to, especially like SPI advanced lithography or SPI photo mask, I went to them the first time I didn't even understand 90% of what I heard. And then I read red, red, red, I had made some context, of course, and then next time I went, I understood like half of what I went to. Third time I went, I understood like 75% of what I went to. Even now, I went and I was like, I still don't understand everything that's going on.

Whereas like you go to like Nurips, you know, a couple times you can understand, okay, what's neurosymbolic reasoning? Okay, what's this? What's that? like you can you can kind of get a mapping of what everything is pretty quickly but some parts of the supply chain are so arcane and so deep and so technical. It takes a lot of times for you to even understand what's happening in and you know on everything right um for every research paper doesn't necessarily mean you didn't you know you go to a conference for a few reasons right you understand the research you understand but like it's all the research that's being published but what you really care about is understanding how does that research intersection intersect with technology also how does that research differ from what's there today and none of these research papers tell you what's happening today but then you just ask people and you you build contacts and you learn and then you like learn about the supply chain and oh this company supplies this company even though it's not publicly stated anywhere or like you know you learn that the the this chemical is like cost about this much and a tool uses about this much and you >> hear you hear the horror stories of like this chemical had a shortage and it totally threw off this part of the

supply chain and then it turns out there's only three companies in the world that make that chemical and it's like >> my favorite one is I learned uh a Japanese guy at that specific Japanese uh conference that I went to where no almost no one spoke English in very broken English he told me about how uh his father worked in this in in in this industry in the 1980s that the the only factory in the world that built this chemical uh burned down and that caused memory prices to like double or triple and I was like wow not too different from today >> not not not at all >> crazy >> um >> inference going to be the biggest market on earth biggest market beyond earth agree or disagree >> um I mean obviously use of tokens is going to be the biggest market um and the value that's created from tokens is going to be the biggest market but I think tokconomics sort of the use of tokens adoption of AI sort of is the most important thing that's happening and inference whether it's open models or closed models will be like one of the biggest markets in the world much bigger than oil I think much bigger than like you know many other parts like inference of AI will be you know many percentage points of the GDP yeah right >> what you've done with inference X I think is you know industry standard maybe say a word on why you started it what it does and you know what do people misunderstand about uh performance benchmarking on inference >> yeah So, so to zoom back, right, like semi analysis, uh, we do a lot of stuff that's like, you know, a lot of it is like research for institutional clients and and our subscription versus products, but a lot of it is also like, hey, you know, this would just be cool to figure out. Let's figure out how to figure it out and just post it publicly.

And that gets, you know, more and more scale. And so we've done this with a lot of GPU benchmarking and testing and training performance and inference performance, but you know, ultimately we saw like inference benchmarking was like point in time. you know, you test it and you take some time, you release it and it's like slow and arcane and out outdated because models change all the time. Every I feel like every week there's a new model whether it's a Chinese model or you know today mythos 5, Fable dropped and new models are coming out all the time. Um on the software layer uh PyTorch, VLM, SG lang um new drivers, new new something drops, you know, in fact the update cycle for most of these libraries is twice a week.

So you basically have the software updating all the time and therefore performance changing. Um you know new inference optimizations are coming out and those get updated and and so I feel like it's a relentless breakthrough after breakthrough after breakthrough that keeps driving efficiency and cost down which is why we've seen you know model cost drop for equivalent quality by like 60x a year. It's incredible. Um but to stay on top of that you can't have point in time benchmarking. You need to have benchmarks be living and breathing i.e. you know constantly running on the latest hardware on the latest models. And so we embarked on a project and we got a lot of buyin from the ecosystem. This was only possible because we had you know enough aura with some of the ecosystem where we're able to get coreweave and cruso and nebus and Oracle and Microsoft and Amazon and Google and OpenAI to contribute to us um compute and then we were able to work with SG Lang and VLM and now Radix Arc and InRact uh which are the private companies who are sort of leading those efforts um the open source efforts um to collaborate with us. We're able to get Nvidia and AMD and Google and Amazon now because we're adding TPUs and trrenium uh to collaborate. Now we've got all these people collaborating. We've got over $50 million of hardware uh donated to us. Um once we launch TPUs and trainum it actually should be over $und00 million of hardware. Um you know maybe about like 15 different chip types all running these benchmarks every single day on all the latest model, right? the best model from Moonshot, the best model from Alibaba, the best model from um there's about five different Chinese models, the best open

source models, the best Chinese labs there. We run benchmarks on their models every day and then also the best US open source models um GPTOSS, Neotron, etc. So we're running these benchmarks every day um in an automated fashion and they run on these these servers that are dedicated to us for inference benchmarking and we sweep across so many different configurations and optimization types and then what it creates is and all the results are public and all the configurations are public. So now we have the paralo optimal curve because a lot of you know times when people are comparing inference performance they're like taking a suboptimal curve or point for someone else and comparing it to their optimal one. And it's like, well, yeah, I can make I can I can stick, you know, if I drove a Porsche versus like some some race car driver, obviously I'd drive it slower. The same thing with inference benchmarking. And so what we did is we created open- source uh basically containers for the optimal points across every uh point on the interactivity, i.e. how fast is it responding to me versus you know batch size, i.e. how many users am I simultaneously serving curve? And so now anyone who wants the optimal point can just go to inference X download it and run that as the optimal point and they can check every day if they want or they can even autod download the most optimal point for that model and and their inference performance will be near peak.

Um >> is that curve like the most important curve in your opinion? The throughput interactivity curve is the most important one. >> Yeah, I think I think um most things in hardware infrastructure uh model application layer everything is downstream of that curve, right? Is it is it something that needs to be super super fast, super low latency? Um, and I don't really care about the cost, so I make batch size very low and I use techniques like speculative decoding or multi-token prediction heavily and and there's so many, you know, possible techniques there. Or is it something where actually I'm batch processing a ton of documents and I don't really care about all these things. I don't use these techniques that actually are worse on cost efficiency but help you with speed for an individual user because I just want to pack a bunch of users. I don't care if the document takes all night to process, right? Um, and right now the way we treat AI infrastructures, it's like one-sizefits-all. But over time, we're going to get to the point where, you know, there's stuff where you you have batch workloads or, you know, you need instant response and there's there's the whole curve that's going to matter for uh users. And so we see this with entropic, right? Cloud code fast mode cost way more than regular mode.

Um, and same with open eyes priority Q thing. Um, >> sorry, dumb question. How does cost factor into the chart? So if if I let's say imaginary example, I have 100 I have a batch size of 100, >> okay? >> And I can do 10 tokens per second per user. So in total I'm doing a thousand tokens per second uh off of that one piece of compute. That's one side of the curve. Super slow, 10 tokens per second.

Um you know, other side is I have uh uh 500 tokens per second, but I only have one user. And so maybe 250 tokens per second, one user. And then there's points on the middle that are more fraal optimal, right? the average person actually wants like 50 or 100 tokens a second and maybe you know the this the the number of users I can batch together. So the curve is okay a thousand tokens total uh per second or 250 tokens total per second depending on how many users I batch and there's a curve in the middle and so ultimately some workloads will actually want the 4x cost decrease because the same unit of hardware can do a th000 versus 250 and some users I'll pay 4x more because I don't care about the price I care about time because the person using the tokens is expensive or the feedback loop that I have here is expense is is expensive. If you had to guess, you choose the time frame 10 years or 15 years. What percent of inference compete do you think will happen in

space?

>> Can be 0% 50% >> Sean 99% like >> this is a tough one. Um you choose the time frame like 10 whatever time frame and you're >> so I think I think the non- consensus or at least against SpaceX thing, you know, I love SpaceX by the way and I totally would buy the IPO if I could buy stocks. Um >> not investment. >> Not investment advice. Thank you. Thank you. not invested ice um from either um I don't think that space data centers will really matter in the next um you know 3 to 5 years um with that said I think in you know 20 years I think the vast majority of compute will be going in space um and so the real real factor there is sort of you know what's the cost it's the time frame it's the cost of building power on terrestrial land and how much power you going to be able to do on terrestrial land and I think obviously my views of where inference you know you know how many gigawatts or terowatts are devoted to inference is it's a crazy curve for me personally >> what's your forecast how many gigawatts or >> um yeah I think I think by you know 2030 just open anthropic we'll have over 100 gigawatts combined um and then you'll add you know meta and Google and you know so on and so on so forth it's it's a humongous amount of compute that will be dedicated to inference um and by like 2040 it'll be terowatts right um the the curve of like productivity that we're going to get and so you know inference deployments is going to be huge And so if you look at like 2040, I think like you know probably more than half of the incremental compute will be going in space. But if you look at 2030, I think it's sub 1%.

>> Do you think intelligence per watt has been increasing? Uh and then it seems like there's still a giant gap between where we are intelligence per watt versus like human biology and so like if we are do you think we are to close that gap? And if so, where is that game going to come from? >> Yeah, I think I think it often depends on what you're doing too, right? Like a TI84 is way more intelligence per watt in terms of doing math than us and it's like 30 years old, right? Obviously this is like a dumb dumb you know sort of >> general intelligence.

>> Yeah. But general intelligence wise um so one of the things inference X does is we also measure the power and cost of all of these this hardware. And so we offer not just you know throughput versus interactivity we offer cost versus interactivity. We offer power versus interactivity. And so as far as has you know intelligence per watt been increasing? Um I mentioned you know it's been a 60x cost decrease for same benchmark level. Um we've also seen the same on on uh intelligence per watt. Um it's not been it's not been exactly 60x.

It's been closer to like 40x. Uh some of the efficiencies are nonp power ways, but there's been a humongous improvement in in intelligence per watt on an annual basis at least so far this year, last year, year before, year before. And I expect that to continue as far as where we are from the human brain. We're we're many orders of magnitude away. Thankfully, doesn't really matter. We can devote a lot of power to computers. Much easier to power computers than human brains. Like you know, we have sickness, disease, and like food preferences. sleep.

>> Uh yeah, exactly. >> Let me just ask one more question on the like on the general theme in my opinion in terms of like you know intelligence per watt or intelligence per per dollar like any any of these metrics. I think there's kind of three levels of input. You can get hardware improvements that are where the hardware is more efficient. You can get lowlevel systems optimizations like kernel level you know improvements matri multiplication libraries you know things like that or you can get like highlevel like model level algorithmic

improvements you know at the highest level. It see like to me it seems like in the last three years most of the gains have come from hardware level and you know and some from the model level like do you think that that is what do you agree with that do you think that's what look like in the future do like do you think there's a bunch of juice to squeeze in a say like kernel level like >> yeah Sean I completely disagree with you by the way great great that's why I'm asking this question >> um okay so I I think you know one way is to look at as these three different layers Um and in that sense like okay from hopper to blackwell which is all we've had over the last three years roughly 30x improvement on deepseeek on the most optimized deployment which is you know you can see on inference there's about a 30x improvement but you know over the last three years um we've had way more improvement intelligence per watt a lot of that coming from the model layer right if you look back three years it's GPD4 now it's like you know you know maybe like Quen one of the smaller Quen models that's like you know 27B parameters total and like 2 billion active is like way better. Um, and so you've got this huge improvement on model layer, you've got this pretty sizable improvement on hardware, but it's that co-design layer and I think that's that's what's important, right?

If you look at the architecture of, you know, any of these models, but deepseek is the most famous one at least, uh, that's public and people have seen. >> Yeah, Deepseek got huge efficiency gains from like co-op optimization or kernel level optimizing memory. >> Yes, I I think it's it's it's like kernels of course, but it's actually you build the hardware architecture for the chip. So if you look at the shapes of all the experts in in DeepSeek, uh V3, they were all optimized for Hopper. And if you look at for V4, they're optimized for Blackwell and Huawei's chip. And what's interesting is despite the fact that TPUs are objectively an amazing chip, you know, and and they run all of deep mind and they do all the training uh for anthropic as well on the pre-training side at least. TPUs suck at running deepseeek, but they are really really great at running other kinds of models that don't run well on NVIDIA.

there is some level of such deep optimization that has been done um whether it be shapes uh network IO uh patterns you know how you do the collectives how you do um things around you know the the arithmetic intensity of the attention mechanism all these different things are co-optimized between the model and the and the and the hardware and the infrasoftware in between and it's it's hard to say you can disentangle the games >> do you think that like my understanding is that like China has done this a lot better than the west the last few years like in the deep sea was one of the first models to really like do this.

>> I don't necessarily think so. think it's more so that the west doesn't tell people what they do right like open eye didn't tell people that you know GP40 was uh how sparse it was what the shape size was all these things but GP40 is roughly the same size slightly smaller than deepseeek v3 and 40 came out you know a little bit earlier right if I recall correctly >> so is your is your view that like all three of these things have been happening simultaneously at like roughly the same rate and the most the biggest gains are when you just co-optimize I would say I would say there's been more gains on the model layer than on that co-op than than on the sort of software infrastructure layer and the hardware layer. Um, but there's been innovations on every layer and and and really the biggest gain and the beauty of the best labs is when they co-optimize all three, you know, and and and that's what like you know when Enthropic is is, you know, even though they used many different kinds of hardware, they don't really inference too much on TPUs. They mostly train on TPUs. um and and they inference a lot on cranium and GPUs and GPU is more a jack of all trades but they've optimized their hardware

they're optimized their model they've optimized everything so they can do that whereas open AI they're you know prior models were optimized for hopper more now they're more optimized for blackwell and you know you you step forward through time these these um these labs and and and the same with Google right they've they've they've optimized you know Gemini 2 was really optimized for the TPU uh v uh v6e or tp Gemini 3 was and then Gemini uh you know the next Gemini that's coming out is really optimized for TPV7.

Um and so sort of like a lot of these things are being co-optimized and actually when you pull that model and put it run it on the old hardware it's really not that great. Um and so I think a lot of this co-optimization is is the most important thing. It's called software hardware co-design and that's what's like really exciting about like you know sort of what what you know I think my day-to-day is like you know great you get to look at one layer there's all these innovations happening here there's all these innovations happening on every layer. The real breakthrough innovation is when you leaprog a few layers, you co-optimize and co-design them, and now all of a sudden you've you've taken what could have been a 2x here, 2x here, 2x here, and instead of being multiplicative to 8x, it's actually 100x because you've optimized across all three layers. And so that's what's really exciting about sort of like what you see at the labs, which you see at like a company like Nvidia who's not co-optimizing on the model layer per per se, but a little bit from the model layer all the way downstream to, you know, silicon. Or you look at a company like TSMC, they're co-optimizing not just, you know, fabrication, but all the way from the components and the consumables and the tools all the way upstream to what the designs, their chips are, the customers are telling them is this co-optimization across many layers of the abstraction stack.

>> There will always be bottlenecks somewhere in that optimization though that are like lagging behind and then need to get pulled forward, >> you know, >> and band-aids to toact. If you had to predict like what are at any level of the stack, it can be literally anywhere. What are some of the bottlenecks you're most like you're kind of tracking most acutely the next year? And not necessarily in the supply chain, not in like scale, but in terms of the actual um and it can it can be in the supply chain too, but just like you know, is it memory improvements? Is it is it that like just like scaling? So memory memory is memory is an easy one that everyone's talked about, but I'm not going to talk about from a supply chain angle. I'm talking about from a technology angle, right? Memory um capacity and bandwidth have been improving very slowly. The NAN cell was invented like 25 years ago. The DM cell was invented like 40 years ago and there's been no major breakthrough in in cell like you know how what a NAND cell is. Obviously NAND is like a very simple gate or DAM cell. There there is stuff that could come down the pipeline that could be hugely innovative. But even over the last, you know, five years, all we've really done is make the HBM, you know, more stacks, faster, but actually there's like new innovations coming in the next few years where instead of, you know, stacking the HPM separately from the chip, you stack the memory directly on the chip and that makes your bandwidth explode. Um, and so there's interesting companies in that space and interesting PC's that companies are trying to do there. I think like memory bandwidth is one of the biggest. Another one is um for the history of like silicon basically for the last two decades at least you know how many watts a chip is can be easily predicted just by looking at it for for a data center or desktop chip it it peaks up at one watt per millimeter squared and so if a chip is 100 millime squared generally the power consumption is around 100 or a little bit less um and if you look at the newest Nvidia silicon the newest TPU silicon it's still on that range of one watt per millimeter squared so you know chips are now getting to you know, 1400 watts.

Next generation is 2,000 watts for Nvidia. Um, with Reuben and such. Uh, and and you move forward to Reuben Ultra, it's going to be like 4,000 watts or something like that. But really, there's increasing the amount of silicon. What's exciting is we're now finally doing things and and it's in development right now where you actually can pump the amount of power into the silicon uh to be way more than one watt per millimeter squared. And now that all of a sudden means you need less silicon.

Obviously, it's running at higher power. It's less efficient in some cases, but you reduce the amount of silicon and you're able to like >> like over thermal issues, >> thermal issues. Um there's uh interference of like electrical interference issues. There's all sorts of different issues uh that crop up and that's why it's a hard engineering problem. That's why we've stuck at about one. But what's exciting is the world is trying to change these things. I think interesting like in a different part of the supply chain it's sort of like you know people people will talk about like energy is hard and you know we have energy bottlenecks and it's like yeah but there's actually like very simple solutions you know one could think of right um take the millions of diesel engines for trucks that the US has the capacity to make um you can very trivially convert them to be using for gas uh in the assembly line and then stick them up to a electrical motor like back driving it so the electrical motor generates electricity rather than the electrical motor causing the the rotation of the wheel, for example, but doing it the opposite direction. And now you've generated electricity by pumping gas into something that us can make millions of. Um, and then, okay, well, that sounds like a pain in the ass to uh service, right? Because now you have to have hundreds of these on a data center site. Well, actually, you can just pull people out of car mechanic shops and have them run around and repair truck engines. Actually, it's actually pretty trivial to not I don't want to say it's trivial, I couldn't do it. Um >> I think you're making a really good point which is that like because the west wasn't really thinking about semic even hardware more broadly the last 20 30 years we didn't have like much innovation we'd have the best minds like thinking about how do you improve these >> why why would you why would you want to go work in hardware when you can uh make ads to ads >> yeah exactly >> um okay I'm dying to ask Nvidia versus TPU what are your thoughts >> um I think I think like everyone wants to pick one or the other for this, but it's really like a function of like look, you know, you look two years from now, Google's going to make 10 plus million TPUs and through their supply chain and Nvidia is going to make, you know, many more million tens of millions of GPUs and both are going to be 100 plus billion dollar, you know, well, Google's going to be 100 plus billion dollars, you know, of TPU created a year and and Nvidia will be, you know, 500 plus or, you know, whatever. I'm not making a specific estimate.

>> This is not revenue forecast. This is just a thought experiment. >> Yeah. Or research. >> You've been media trans. >> Absolutely. you know, getting ready for the SpaceX idea. >> Um, are you guys big in SpaceX? Okay, so that makes sense. Um, >> we're very lucky to be very large investors. >> Awesome. Awesome. Um, so I would say um the the case of sort of like Google TPUs versus uh Nvidia GPUs, they both have like points that are really like in their favor, right? You know, Nvidia will be like, "Oh, well, we have switches and we're general purpose." And and TPUs will be like, "Well, we're more optimized. actually more energy efficient and our network is actually more um optimized for certain types of network architectures. And so you have like these counterpoints that both would really uh get into and you know I could with a straight face argue with you like that GPUs are way better than TPUs or TPUs are way better than GPUs but it comes down to hardware software codeesign. So actually the way OpenAI's models are headed, it would be a terrible decision for them to use TPUs potentially. And the way that Enthropic and Google's uh models are headed, it's actually a terrible

decision potentially for them to train with GPUs. I mean, it'd be fun to what's the what's the fundamental difference there?

>> There's various things, right? Like the size of the matrix multiply unit is different as a as a very simple thing. And therefore, the shape of the matrix multiply you do, the attention mechanism you use, uh the way that attention mechanism is structured, the way the experts are structured. So you think so open AI and anthropic are converging the very different model architectures. I think they're I think they have quite different model architectures. In fact, um, you know, open eyes are much more sparse, um, and that has benefits. And then anthropics are, you know, they're still sparse, but more dense in general, and and that has different benefits. And there's many other things, right? The network topology, right? Nvidia, all of their chips are connected to switches, NVLink switches. For Google, they have no switch. Um, but what they've done is they've been able to, you know, Nvidia, the NVLink can only connect 72 GPUs. for Google, their ICI can connect 8,000 chips at super high bandwidth, but you have to pass through other chips to get there because there's no switch. And so there's like there's trade-offs there.

There's positives and negatives and that influences the model architecture. It's not necessarily that you should uh you know claim one is better than the other because at the end of the day, how do you say that this is better than that when you can't measure them in isolation because it also extends up to the model layer, right? Um >> but I remember for a long time thinking you know one the programmability of Nvidia and just CUDA as as such a big moat. It seems to me that narrative has kind of changed at least in my mind for the last three six months like model companies no longer care about if we have to write custom kernels for you know this other chip so be it. We'll work with four or five chips if we have to. Um Claude and Codeex are actually quite good at doing a lot of that optimization work. And so it seems like some of the and then and then it's you know it's not like there's 10,000 model companies that are each you know each need programmability. There's on the order of tens maybe model companies and so it seems to me that like if you the fundamental premise of like tens of thousands of big customers that need CUDA compatibility like it seems that kind of thesis is is changing in the last >> Yeah. I mean I mean certainly the CUDA mode and software remote is at least partially uh disentangled because you know models are just great at coding and all software gets commoditized in that case. I do think there is some level of like open source and you know what people call the CUDA mode is not actually anything to do with CUDA but it's like the fact that DeepSeek Kimmy and and Zippui and and Alibaba and Tens all these all these companies Xiaomi had an awesome model recently their models are co-designed for GPUs and therefore if I want to run them on TPUs actually in some cases they don't run really well on TPUs now Google just has to create their own open source model ecosystem or open source models themselves so they have the Gemma models and and so you end up with like well that's not really CUDA as a moat it's that the downstream product is more optimized for Nvidia and in these cases these companies are just open sourcing them or like Neotron is just open sourcing it and then the users of it for example the open you know the inference uh API providers the RL companies that are trying to take open models and customize them for company's business use cases all these different companies are downstream of the fact that like okay well I guess I need to use Nvidia because the ecosystem uses Nvidia even though I don't partic particularly care about writing CUDA kernels because the models are great at that, but it's like the shape of like well this expert the the demod is this and you know the hidden dimension blah blah blah is this right and and so therefore it's better to run on Nvidia GPUs than it is on TPUs and vice versa right if Google were to actually open source really good models you know this would be the

same thing right people would take their models and they'd be like oh wow these don't run that well on Nvidia GPUs um I should actually just rent TPUs or buy TPUs and do it on there for small teams you're going to want to use all the open source software like VLM MSG laying um pietorch all that stuff but the big labs they don't necessarily need to use all that right open I forked PyTorch long ago and you know anthropic and all these other people don't necessarily rely heavily on the open- source implementation of you know these things they forked things or built it on their own already and so they don't need to rely on the open source and therefore now it's more like you know I'll choose the best hardware and I'll co-design my model and infrastructure software through and through for that hardware uh that is the best and most costefficient >> and you know I'll have AI help me write all that software.

>> What do you think of Cerebrus? >> I think Cerebrus is a really innovative company. Um I I think in in some spots of the market they're really really good. Um very fast inference. I think that's a big market. Uh we use fast mode almost exclusively at semi analysis. Um >> by the way I love how disciplined you've been about accounting for I don't know if that was one exhibit you did or if you do it consistently but accounting for the dollar spent and the ROI on each task.

>> Awesome analysis. >> Yeah. Yeah. We we we uh we do it pretty diligently and so thank you. That was the dark GDP article that we wrote. Um and so and and also like track everyone's token spend by day and if someone's like spiked up I'm like what did you do? It's like okay thank you for telling me that that seems worth it. Cool. On with my day. I think fast mode is obviously worth a lot for high-end tasks, right? I could just see so many different use cases where you know super fast tokens are worth it. I can also see the flip side where there's a lot of use cases where super fast tokens aren't needed and and therefore uh the market won't pay for them and they'll use GPUs and TPUs instead. I think the big risk for Cerebrus is I mostly think the best models are the ones that you want to use fast mode on and small models you necessarily might not use fast mode on.

I could see that being wrong with you know financial markets maybe or something like that like a Jane Street high frequency trading or something like that um or medium frequency trading. Um but ultimately you know running really large models at really long context is very difficult on SRAMM based chips like Cerebras like Grock and so now it all of a sudden is like you know what happens then if like the models get too big right if open's model is not you know on the order of uh you know hundreds of billions parameters or you know low trillion parameters but it's actually 10 plus trillion parameters now all of a sudden I don't think that that will fit on cerebrus right and then if that doesn't with a long context length right if you have a million context length now that makes it really difficult to justify you know, and and as all so far we've seen the bulk of revenue and usage at the labs be on their best model. Even when the model price has gone up, we've seen that. Um there's some data that shows that even though Fable just released today, they've had incredible amounts of people switch to Fable and Mythos, sort of that next tier model, even though it's way more expensive. And so um >> is that and that's volume by dollars totally. But was that volume by tokens?

>> Well, I guess who cares about volume by tokens? It's about the dollars. >> Fair enough. >> Right. If I don't care that there's, you know, uh, you know, I don't know, 200,000 Mini Coopers or Toyota Camry sold if if, uh, you know, I don't know, Ford50s are 5x ASP and they sell only half as much. >> Okay. Right. >> And then and

and therefore the most lucrative market is pickup trucks in America. Right. Mostly being facicious, but like >> I do think this is one of the things that you've done so well and differentiates you from almost everyone else is that you you care so much about the economics in addition to the technology. And I think very few people bridged those two thing things well. And so >> I think I think it's really fun inside of semi analysis because we have 90 people and like a big chunk of them are technologist engineers across the whole supply chain. Um and then a big chunk is people who are formerly at hedge funds and you see these arguments like people are like oh well that doesn't matter and it's like then someone's like well but cost and then someone the engineers like no no but this technology is the coolest. You see this you see this organically like fight it out. Um and and were pretty informal and you know given the fact that I was a for moderator is you can imagine what the the enjoying it >> you don't wrestle with a pig because a pig enjoys it.

>> Exactly. Just on this topic before going to the next question. Are there like trigger topics in semis for you? You know like if someone's like which is like such a meme you think this person must be a like if you know if it's like oh you like memory is the bottleneck. I mean it's true but like um I think I think moreover the one that really gets me is people are like AI has no ROI >> infuriates me right like there's like what's the ROI or like denying model progress right there's these people that are like models aren't getting better they're not reasoning they can't think they're going to deadend and plateau and it's like bro the line has been up and to the right in terms of capabilities this entire time and they're like look this benchmark didn't improve that's cuz it said 90% look at the new benchmark you saturated now they're skyrocketing, right? It's like I think that's more so the issue and challenge. Like I think semis are really complex and I don't fault people for um lacking like understanding of it.

Like I learn stuff every day about the semiconductor supply chain from people and I've been studying it for you know arguably 18 years since I started moderating the forums when I was 12 right like you know arguably been studying it for that long but even then like and it's like live breathed and that's all I care about but there's so many layers of the abstraction stack it's like like I learned about a new chemical that does like a hundred million dollars of sales like yesterday and I'm like whoa didn't know this one existed and what process it did and it's like but it's like you know you learn about things all the It's like okay hundred billion dollar sales in a you know couple hundred billion dollar industry is whatever but like you know it's like >> but it's essential >> it's essential and it's like actually every chip requires it. It's like wow I guess there are a thousand process steps and you know it's like oh yeah you like semiconductors name every process step.

It's like no come on. What what I think is the most funny is when people have all the facts in front of them and then they get the conclusion completely wrong. Um and that's >> that happens in our job all the time too. >> Yeah. >> Yeah. I mean, I can't I I get I I think my attitude is not to be mad that you do that. It's to do it as fast as possible. >> I think the industry because it's so it's just like AI is the most important thing in the world right now and there's so many near-term bottlenecks. We talk a lot about the near-term. Are there longer term things that you're really excited about? Like say on a 10-year time frame? We talked about orbital data centers, but like like siliconics, you think they're underrated or overrated on a 10-year time frame? Are there other things that on a 10-year time frame?

Yeah, I mean I think on SP I think space is like super crazy awesome in the 10-year time frame that I'm you know for space data centers and all these sort of mining asteroids and all these things which is you know super excited about the vision of SpaceX right um again not investment advice before you hop in um I think I think on the semiconductor side tremendous market movements and tremendous like things can happen just when like things happen one year later or sooner and so that's all like technology that like you know in terms of like co-ackage optics like well like everyone knows it's going to happen by the end of the decade the the debate is like 27 7 28 29 2030 but some point along there it's going to happen. I think the more interesting thing is like there's companies like um I did you guys invest in Navian Ral's company?

>> We did. >> Okay. Yeah. So I think like he's trying to innovate on like the silicon layer on the software abstraction layer and the model layer simultaneously and he fully understands that it's not a like a you know we're going to do this in a few years. >> It's not a two-year time frame. >> Yeah. It's not a few year time frame. It's a long-term bet. Um, and like stuff like that is like, okay, we're going to bring like potentially like analog compute with energy based models and like all this crazy all at once.

It's like that's exciting. Probably won't work, but you know, that's exciting and I I like really look forward to >> definitely won't work quickly. >> Yeah, definitely won't work quickly is what I should say. I believe in Deaveen and like, you know, I I I met him very, you know, I think he's one of the first people I met in the industry um, funnily enough, like in 2020 or 2021. Um, actually 2020. Yeah. It says something about him. I think he's someone in my experience. He's always trying to >> I baited him on the internet. I baited him on the internet. That's >> He's always trying to help the younger generation. He's trying to identify talent. And >> he was also so ahead of his time with Mosaic. I remember getting pitched.

>> No, it was 2019. I was still I was still anonymous then actually. I I baited him on the internet and he started replying and then I just took it to DMs and then took it to a call and like that was the first person who's like really important that I talked to in the entire semiconductor industry. funny. >> Um, but yeah, sorry to interrupt. >> That's funny. What do you think is the end state of the ecosystem? Like do you think every lab, every hyperscaler just has its own chips? Like train seems like it's now working, right? So do you think we end up with every lab, every hyperscaler has it own chips at least for inference and then maybe for training you go to Nvidia or whoever or what do you think is the end state?

>> I think everyone will try and stop trying. I think ultimately um you know supply chains matter. what technology you can bring in matters and more and more as the industry gets bigger supply chain diversification happens. Um you know right now everyone's chip more or less looks the same. It's a big logic compute die in the center and there's some HBM on the right and left and on the top and bottom top side is networking and then the bottom side is PCIe and other IO. Um and that is the exact same structure for tranium TPU Nvidia chips. Um and most of the startups um not Grock and Fris are doing weird but that's cool you know um I think like as you step forward we're going to get more bifurcation of hardware architecture and model architecture and therefore people are going to co-optimize them and you know some of them will end up in local minimas right you know as we're you know if this is like gradation gradient descent like people are like trying to go to the most optimized solution some people will race to a local minima and then the question is like how do you leap how do

you scoot back over to like the absolute minima and some to some extent like a general more Nvidia will always be more general purpose than anyone else's chip in general um at least on a parallel AI compute basis because they have so many customers who care about different things who will always give them feedback in the design you know the minima will always be better than them but is that minima a local minima like is is the TPU or tranium or grock or cerebras or whoever's design optimized awesomely for here but in the end state actually you got to go over here and so they're the wrong >> um and Maybe they make a great time, they're great for a little bit of time, but then they end up being wrong. It's like that's the real question. Um, and so I think I think there will be a big market for general purpose AI compute.

Um, because you talk to people at labs, they don't even know what architecture they're going to be doing in a year. Like, right, like they literally don't know what architecture they're going to be doing in a year. They have bets. They have many research bets and and that's this exciting thing, but they don't know where where it's going. generally they like know what hardware they have and they're trying to co-optimize but ultimately like if a new breakthrough happens on model architecture it's like just replace the tension mechanism with something else right who knows or you know all of a sudden you know something happens the best hardware will change and therefore like are people going to make fiveyear investments on hardware solely on you know an an asich that is more specialized or are they going to do so they're going to have some bucket of more general purpose compute and so you see this with like Google's paying $11 an hour per GPU to XAI for G for GPUs, right? Like that's insane, right? It's a very high amount of uh obviously compute is limited and and so on and so forth, but it's like very like insane, but at the same, you know, despite the fact that they have TPUs and so there's like some questions there like why do they do that? Um Google actually has three different design programs for TPUs.

They're making a TPU with Broadcom. That's a different architecture than the TPU with MediaTek. That's a different TPU than the architecture that is, you know, I won't disclose, you know, by research. Um but, you know, they're they're making different architectures. It's not just like, oh, they're making TPUs with a couple vendors. It's the same architecture. It's different architectures. And the third one is a very different architecture from the first two. And so, I think people recognize that the local minima can happen. And therefore, um, I think everyone will have their own ASIC program. I think everyone will deploy billions of dollars of their own AS6, tens of billions of dollars. In the case of Google, hundreds of billions of dollars a year of their own AS6. But ultimately, they're also going to have workloads that don't use TPUs, right?

Some of the Google bets that are not Gemini Deepbind actually primarily use GPUs. They don't use TPUs. Um, some of them also primarily use TPUs, right? It's a bit of a broad thing, but like, you know, maybe for drug discovery or for Whimo, you might not want to use TPUs. I won't say which one it is, but like, you know, there's there's there's there's different architecture bets and different paths for AI. AI for science may have different algorithmic patterns than than general intelligence AGI models. Um, and so I think we'll see we'll see diversity continue to proliferate. Yeah. and and and because the market has gotten so big, niches will be carved out and so that's makes it possible for companies to have their niche and actually make money even if the majority of the pie goes to Nvidia and TPU and tranium.

>> Yeah. >> Okay. Love that. Can we talk about the data center buildout? Like one, it seems like I mean by all

accounts if you look at the charts like dollars per compute hour, we are in the middle of a crazy compute crunch. Um and it seems like it's both a demand and supply side crunch, right? demand for long agents skyrocketing, supply, all these data center buildouts are delayed. Um, do you think this we're in a compute crunch for the foreseeable future or do you think it alleviates at some point?

>> Yes, every quarter we're deploying vastly more compute than the prior quarter and there's more data centers built than the prior quarter. Um, this year there's going to be 20 gigawatts uh even accounting for the delays and next year there's going to be more than 30 gigawatts accounting for the delays. Um, of course delays happen on everything, right? Anything hardware can have a delay. That's that's just the reality of life. Are we gonna have a compute crunch for the rest of our lives? It depends on what happens with models. But like the TAM for Mythos, you know, Mythos 5, Fable 5 is not just like 2x that of Opus, right? The model is so much better and it can do so many more tasks that the Tamford is way larger than that. And yet compute in the world did not double in the last, you know, six months, right? From, you know, Opus or maybe like seven or eight months since Opus 45 launched to now. huge you know 46 47 48 were improvements but fable and methos were like a huge step function improvement the world's compute did not double in that or or quadruple or whatever in that same time frame but the demand for useful tasks that can be done by AI the number of useful tasks and the value of them that can be done by AI has and so now the question is what happens well obviously anthropic in Q2 is profitable their net income profitable um excluding stockbased compensation um And and I think by Q3 they may even be profitable including stockbased compensation. That's like how profitable they're getting. And their margins on a on a on an Opus token, at least Opus 48 token is like north of 80% for the API price. They've got a lot of deals where their total corporate gross margins gets clawed down a little bit uh because of like how they do bedrock deals and vertex deals and things like that. But ultimately their their per token margin is so high. Well, then if you don't have the cap, they have the capability to pay ultimately every GPU they buy at above market rate. You know, they also bought GPUs at above market rate from SpaceX, which is below the rate of Google, but that's because they signed earlier. Um, you know, it's it's something that, you know, other companies, maybe a ventureback company or company that's not really got positive uh margins can't necessarily do, right? What is the cost benefit ratios like every GPU I rent because I'm out of compute capacity I can immediately turn around and sell tokens on it or every TPU or every tranium I can immediately sell tokens on it at a positive margin and if I'm running 75% gross margin and I double the cost of the compute it's fine I'm still running 50% gross margin and spinning up more compute nodes is not really necessarily a human requiring task for them if they're renting them and so ultimately it's like well my NOI still goes up right and and so I'm going to rent GPUs at whatever price at some level whatever price I want to pay I can pay.

>> I have almost the reverse question of like at some point does this compute build out go bump at night? Earlier today I think there was a tweet like Cuso publicly said one of their customers had asked to halt construction on one of their data center buildouts. Like it seems like everybody in the ecosystem is so levered right now to like we got to build, we got to go build, we got to build. High leverage high growth to me is like makes me very very nervous as investor. Like >> wait hold on. High leverage high growth means small amount of equity has huge upside. You're not a debt investor.

>> You're a credit you're an equity investor, right? >> Let's go. >> Um, >> look, you got you got to go to the school of private equity. Levered buyouts only. >> I actually come from the school of private equity. >> Oh,

awesome. >> She forgot the school. It's been a VC for too long. >> Yeah. >> No, I just do revenue multiples. No, but are you do you see any signs of that? Are you worried about that?

>> I I I see what you mean. Right. And that sort of goes back to the model point, right? Obviously if the models expanding the total economic valuable like work sort of the dark GDP uh report that we did and the you mentioned earlier um if the work that these models can do does not expand faster than the compute capacity then that tide turns right and over the last six months that tide has been you know very much levered in this direction of um you know the models can do more work or can is exp or expanding their TAM of work they can do faster than the compute is increasing And so prices go up. It's very possible that all of a sudden model progress stops.

You talk to anyone at Enthropic or OpenAI, maybe they're drinking the Kool-Aid, but you talk to basically all of them, they're like, "No, no, no, no. Model progress still go up." Um, and so, you know, ultimately, you know, current methods could stall somewhere. I'm not sure where that would be. It seems like we have line of sight to model improvement, rapid model improvement. And in fact, models are improving faster than they were six months ago or a year ago because there's I wouldn't call it recursive self-improvement, but basically the engineer the models are helping write all the info and and launch the next model sooner and sooner and sooner. So you've got this like pseudo recursive self-improvement loop going and so the models are getting better and better and better faster. Um and so but ultimately, you know, capital is a big problem which is why Google raised capital. You know, they they've got an ungodly amount of SpaceX, right?

They own like 5% of the company. >> I think a little more, but yeah. Yeah, maybe. >> I think at one point they had like 10%. >> Larry Page invested a billion dollars at a $10 billion valuation, got 10% of the company, it got diluted, like all this. But that was one of the greatest investments of all time. Good job, Larry. The guy. >> So, they know they have like a hundred billion dollars in the bank that they can sell in, you know, nine months or whatever from the lockup.

>> And they have all the gross profit they do, and yet they still modeled that. and they were like we need to raise capital and so they did an offering and it's like that's insane. So that tells you how much they think they need to spend. But capital is like really, you know, you know, Meta's do Meta did announce that they're going to do a raise. Stock tanked. People don't like it, but you know, that's all these companies are going to raise capital, whether it be debt or equity. At some point, money spiggots will have to, you know, slow down. But right now, every GPU that Amazon adds, they're making higher revenue or every TPU or tranium, you know, whoever anyone adds is is making is making gross profit. I do a little bit of a tea up on this to turn into a question for you. But like >> as we talk about this for me, the thing that's going my through my head that's that is almost an alternative hypothesis for like the Crusoe example. I'm going use an analogy in oil like in oil Saudi Arabia has way lower cost per barrel to produce oil than a lot of other countries. There's also like the purity of the oil. A lot of you know Saudi has generally like very low contaminants in their oil which makes refining easier all of this. The question for me is like when you look at for every gigawatt that's being put in the ground if call it the 20 gigawatts coming online today like how much like how much homogeneity do you see in those gigawatts? Is it something like and I don't you can tell me whatever metric you think is right but like are Google's gigawatts two

times more valuable than say most Neoclouds because they have optical switches and they have like they've been doing it for a long time and like they know how to do power smoothing because I think this could be the alternative hypothesis that some of the people that are it's like the people that are good at at building data centers they they should just do it to the max because there's so much demand and there's so much better than it, but then maybe we're starting to see the early signs of the people that are like not as good at it kind of getting hit a little. So I like I don't know the reality here. I'm just curious how you think about this.

>> So so far um there there are metrics for this, right? So uh tranium sells at sub10 billion per gawatt rental rate uh to anthropic and to open aai. GPUs at least before the craziness of the last six months usually went around 12 to$13 billion per gigawatt. So the rental rate and this is from a neocloud versus Amazon even and now when Amazon sells GPUs they'd also be 13 or so >> and my understanding of that also is that those number like Amazon subsidized that a little bit so that it's like I actually think the numbers were even like I think the disparity was even more >> it's less than 10. It's less than 10 but there's like some weird basically >> and like look I my understanding obvious like anthropic played a big role in making tranium useful in terms of you know writing all the libraries etc and and so like I >> everything I hear is that tranium's really freaking good hardware and it's getting way like way better and obviously anthropic now using it a lot so hopefully we would see that price go up you know like per the the deal they did was actually like there was a floor mechanism them and like it if it didn't do well it would be like cheaper and then to the point where it's cancelceable and you know if it if it did really well the price is kind of higher um but effectively um less than 10 right is is where tranium shakes out at whereas GPUs I mean this the SpaceX deal again was like 25 or something crazy billion dollars per gigawatt or $25 million per megawatt right a year rental rate with Google I was like that's a crazy divergence now obviously if if if Amazon was selling tranium today it' probably be more expensive than 10 because the comput shortages, but you you do see this already in the sense of uh with data centers oftent times a rental price of a data center if you're doing collocation, right? Not compute in there, but just power. Here's the data center. Um you you price it generally on a uh dollars per kilowatt per month. And so they used to be $60 per kilowatt hour per month, and now you see things transacting at anywhere from like 120 to 160. Um but different quality data centers, this actually you've I've seen data centers go as high as 200. um when the customer is not such a great credit rating and then the data center is a pretty good one. And I've seen stuff go as low as 100 still or in like India go like as low as 80 because the grid's not reliable, the internet connection's not great and it's a pretty mid data center but at least it's a data center. Um and so you you see this huge discrepancy there already.

Um in the case of like data center construction, usually the pitfalls they just fail. There's a lot of people who fail, you know, claim they're g they're like they're like four guys they they're like, "Yeah, we here I bought some turbines. I put the money down for them. I'm gonna build a data center." And then they get delayed, delayed, delayed, and fail. Um, so you have to like probability, weight, time, weight, time lag, the teams that suck versus don't.

Um, and and sort of, you know, our data center model does that. We kind of track every data center uh and try and do this for every single one based on, you know, equipment that they're using and all these things. One of the things you mentioned about Google is you know in a gigawatt data center they'll actually put like 1.5 gigawatts of hardware and because they have such understanding all the way from workload to u you know

they're able to slosh the power around and so instead of you know constantly you know a gigawatt of compute which typically runs at like 60 or 70% utilization in terms of power consumption not utilization of the hardware someone's always renting it um they're now running it at like you know you know that 60 to 70% means it's at a gigawatt and you're using the full gigawatt um you see people doing deals with including Google with utilities where they're like, "Oh, well, I know this grid can sustainably take a gigawatt, but you know, except for three days of the year, you can actually do two gigawatts, so give me two gigawatts and then just tell me to turn off." And so they'll do that. And so these sorts of tricks and then you need to have supreme management of workload, backup power, all these things, um, generators on site to figure out how to actually keep it 2 gigawatts sustainably. When people do this, they're able to charge more. Whether it be I'm actually selling two gigawatts despite only having one gigawatt because those three deers deal days I'm be able to deal with via battery, gas, etc. or I figured out how to build power on site. Now I have a gigawatt where no one else does and so I'm able to do it quickly. Um it's not necessarily transacting for a higher price. It's that I'm selling more gigawatts. And sometimes there are levers where you're selling more gigawatts is where where each gigawatt is selling at a different price. Um I think it's more on the data center and energy layer. It's more about just having it versus not and then that being delayed or not. It's more binary. But on the compute side, I do think there's a lot more um interesting work there.

Right? A gigawatt given to Enthropic is objectively worth more revenue than a gigawatt given to OpenAI. And it seems that both of them could sell every gigawatt that they have right now. Uh given rate limit problems and token max limit and all these sorts of things at OpenAI and anthropic. Uh especially since Codex 5.5 came out, it's much better. And then likewise, if you gave a gigawatt to SpaceX, you know, they turn >> my my guess, like my suspicion is that they're they probably make better use of the, you know, hardware than most people. Um, just like I think people underestimate how much networking experience they have from Starlink in particular and also how much just like power management experience they have via from Tesla.

>> Yeah. people like Brett Mayo are like incredible like >> pretty good. >> Yeah. >> And so I I think that like for me that's actually I think probably the thing that might I don't actually know the answer but I think that might be missing from the analysis a lot of people are doing. >> I think I think it's also the fact that when Coreweave builds a gigawatt even though their GPU compute is objectively better than Amazon or Google or Microsoft's in terms of performance.

We've tested the performance and reliability. Um, the problem is Google sells it six months before they have it up and they need to turn around and take that paper that they signed to get debt uh with that credit backing and then turn around so they can actually pay for the PO that they've already issued you know for the order that they've already issued. Whereas SpaceX was like no no no this is running now buy it right and it's it's a big discrepancy when you have a balance sheet to do that versus not and that also helps your revenue per megawatt like be much higher.

>> Why does the Neocloud opportunity even exist? Because if you had asked me five years ago, I would have said the hyperscalers are going to own this. And you know, you mentioned just now core weight has better performance than than the hyperscalers. Like what why does this opportunity exist maybe at the macro level

and then in the execution level? >> Yeah. So in 2023 I wrote a report that had uh Amazon really hate me. Um it was called Amazon cloud crisis. So I talked about how Amazon was the best cloud because they had their nitro nicks which offered like tenant isolation. all the hypervisor ran on the nick and then you could sell all the cores and they had you know custom SSDs that they made and they'd buy the raw NAND and they'd have lower cost because they'd buy the raw nand and build their own SSDs um and you know they had their custom graviton CPUs and that drove down cost for for per core and so they had all these things that enabled them to sell more cores have better security good networking for but this was all for the traditional CPU better storage for the traditional you know cloud world but in the AI cloud a lot of this stuff hurt performance right these nitro necks bad for performance, still are worse performance. Although they've caught up a lot because they've had a couple iterations to like, you know, improve them, but they're still worse for performance. Um, a lot of the security stuff doesn't matter because it's not like I'm time splicing users or splicing a socket into many users, right? It's like no one buy rents a single GPU and an 8GPU server. No one rents a single GPU in a 72GPU rack. They rent the whole rack and in fact, they rent many of the racks. And so, and then and then there's no like, oh, I rent for six hours and I give it back. it's everyone has these long-term contracts.

So, the mechanics of the GPU rental market meant that a lot of the expertise of the hyperscalers fell away. Um, and a lot of the expertise that they did have were actually some of them were detrimental, right? Network performance for Google and Amazon. It was they had custom networks that were better for traditional CPU and for the stuff that they were doing, but actually worked for AI. Um and then in other cases it's like well you know Microsoft would save money by building their own data centers but their data center teams are actually were not actually that great and so when it came time to run you know when it was predictable building it was like fine when it came time to like actually double your forecast for the year it's like they fell on their face and they had to go get a bunch of NeoCloud capacity. I think so performance I think you know I think time to market's another one right ne you know these massive organizations no one's getting rich from building this data center faster right but you look at Crusoe for example Chase and and and and all the other people at the team you know I was going to name some people at the team I'd rather not you know these all these people are getting rich if they deliver these this compute faster they're they're you know they're lever they're hyperlevered equity owners >> hey look they're also all coming from Bitcoin And and you know they you're not supposed to say that.

>> Uh I mean a lot of the data center like their main data center guy came from Microsoft. >> I don't know. I'm just I'm just teasing. But it's uh you know it's like you learn you learn a lot when you're in a very high fluctuation you know market. >> What do you think was Jensen playing for each chess? >> Jensen absolutely hates a world where all the hyperscalers have all the power. There's a reason he's like blowing money on like random AI labs that like I don't even know if like it makes sense to but like you know he's blowing money and pumping them up and going to you know everyone around the world and saying you should invest in this company because he wants to create a multipolar world.

That's why he loves Chinese labs because he wants to create a multipolar world. A world where open anthropic and Google models are the only models is one in which he's screwed. Yep. >> Right. Um a world in which you know the hyperscalers are the only ones building compute is one he's screwed in. Yeah. >> And so, you know, of course he needs to point the allocation gun at NeoClouds, help back stop their clusters, do anything and

everything because while today a GPU sold to Cruso and a GPU sold to um Coree and a GPU sold to Google and Amazon are all the same price for him, five years from now Cruso and Coree existing means Google TPU will be weaker and means Amazon tranium will be weaker and more inference being done with you know non-clos model labs is is better for firm. So I think you know the neocloud ecosystem is you know it's these people that are wild west these neolabs as well a lot of them have investments from Nvidia it's the wild west some will fail many will fail but you know some will emerge as really great teams whether it be you know oddly cruso who's a bunch of crypto guys who then started building data centers and doing flared gas stuff or you know corweave who initially was a bunch of New York hedge fun guys >> they were also and then they were doing >> and crypto guys but then they they they like built you know there were a lot of people who didn't bubble up like them started around the same time just failed Right. And so I think you know um >> I gotta say both those teams are phenomenal. They deserve a lot of credit and it's like that's your point. But >> yeah, I mean my point is like he he you know you throw it's like you throw a bunch of like bait into the water and the best fish will figure out and survive, right? Um and and sort of the same way with the Neoclouds and and and he hopes the Neols as well. We'll see if any of the Neolabs really bubble up, but like you know Thinking Machines has a few hundred million dollars of ARR, right? That's pretty impressive even though they've had you know in the media it's like oh they've lost all this talent. It's like, well, but Tinker is doing a few hundred million of ARR.

Like, that's pretty impressive for out of the gate a product that's less than six months old or whatever. Um, and and you, you know, we hope the same happens to other Neolabs and and so um you know, he wants a multipolar world. >> Truly, congratulations on the success. Thank you. Just the last thing I'll say is I've seen a little bit of this. I think the public, they can probably tell from listening to you how hard you work, but like it's clear you've just been working your ass off for more than a decade and it, you know, led to the last few years of being in the right place, right time. But like it's unbelievable what you've accomplished and I know it's just the beginning. So, >> thank you so much.

>> Thank you for doing this. >> Awesome.