

More good News for Intel, AI Model Trends, Future of Datacenter

43 MIN · YOUTUBE · [HTTPS://WWW.YOUTUBE.COM/WATCH?V=XHaIAvNLEyK](https://www.youtube.com/watch?v=XHaIAvNLEyK)

<https://www.youtube.com/watch?v=XHaIAvNLEyK>

SUMMARY

The discussion revolves around the current state of the semiconductor market, particularly focusing on Intel's recent partnerships and the evolving dynamics of AI and data centers. The hosts express concern over the valuation discrepancies in the semiconductor sector, where established companies are undervalued compared to newer entrants. They also highlight the strategic implications of AI advancements, particularly with Anthropic's Claude, and the necessity for data centers to optimize for power and compute efficiency.

- Intel's partnership with Tesla and its implications for the semiconductor industry are discussed, though details remain vague.*
- Concerns are raised about the market's valuation discrepancies, with established companies trading at low multiples compared to less proven firms.*
- The conversation highlights the importance of advanced packaging and the potential for Intel's technology to gain traction.*
- AI advancements, particularly Anthropic's Claude, are reshaping the software landscape, potentially usurping traditional productivity tools like Microsoft Word.*
- The hosts note the increasing complexity of data center design, emphasizing co-optimization of power and cooling systems alongside compute resources.*
- There is a growing trend of companies measuring employee AI tool usage, but concerns exist over the effectiveness of token usage as a productivity metric.*
- Future costs per transistor in the semiconductor industry are projected to rise significantly, but the value captured from AI may offset these increased costs.*

It is a little bit concerning to me that there are a lot of really what we would consider high quality names, companies that you just know we're going to be around in 10 years trading at massive discounts to companies who may not be around in 10 years. And it is cognitive dissonance and I I'm I'm not sure what to make of it. It's just a it's a weird time in the market. Hello everyone. Welcome to another episode of the circuit. I am Ben Bajarín.

Greetings Claus. I'm Jake Goldberg. All right, so we have on the docket da da da another announcement from Intel. This uh actually two two announcements from Intel to be honest with you uh in the same week. So, a couple a couple of things on they they made an announcement with Tesla. There was a nice picture of Lip-Bu Tan uh in front of the Intel sign for Fab 52 in Chandler, Arizona where Ben also took a selfie when he toured Fab 52. Uh standing there shaking the hands of Elon Musk because sparse on details, but somehow in some way Intel is a partner with it with with Terafab. Now, everybody was like, "Oh, this is how he does it." cuz we were

all like, "How in the world are you going to do what you want to do in in CapEx and uh you know, gigawatts of compute and foundry when you have no ASML machines, you've got no process technology. You were clearly going to use somebody else's stuff. It's just a matter of who.

And so now there's more clarity that Intel is involved. We we know Samsung's involved, too. So, both these people are involved to to varying degrees. But but my initial take and I'm sort of curious if you land on something is is sparing details and we have no idea what the economics are. I felt it was a vote of confidence for Intel IP. And again, I'm not going to say it's just manufacturing capabilities, whether this was based on uh on things they can do with with wafer production or advanced packaging, although I think it might lean toward more toward advanced packaging. Um a a a big customer a notable person, celebrity force of the industry and Elon Musk is going to use some stuff from Intel under one roof at Terafab.

So, I thought that was you know, a a vote of confidence. I I'll I'll give it that. It was a vote of confidence, but beyond that I don't know really what to make of it. I I feel like my feelings about this Terafab announcement are the shrug emoji. Like, "I don't know." Um cuz they were just very light on detail. Uh what exactly are they planning? Uh you know, In- Intel right now it has its own capacity shortages. They're going to work through those hopefully sooner than TSMC, but like >> Yeah. they don't exactly have a lot of EUV systems to spare.

Um I have a theory on where they're going to get their lithography >> I want to from I want to hear it. I don't I I it's it's a theory. Let's We'll we'll >> Boo. Yeah, it's it's too crazy. So, I'm going to hold off for now. Um but yeah, let me let me do some more. I'm I'm I'm working away at that, but I can't get confirmation. So, I'm going to hold off on that theory.

Okay. But it's something. Like you're right. It's it is like a a a big potential customer like a massive potential customer when you add up x.ai and Tesla and their silicon needs and E- Elon is uh he runs a lot of companies that are going to need a lot of chips. So, I'll just leave it at that. And Intel is going to be in that mix and that's that's good for Intel.

Probably. >> Yeah. Well, I mean, so there there was two other things, too. I think there was a little bit of a confusion like, "Oh, he's going to use Intel fabs." And you know, the premise of Terafab is everything under one roof, which I understand, right? When you talk to anybody who manufactures product to TSMC they essentially say the same thing, which is that I like that it's all under under one roof, which which is the reason why I'm I'm somewhat optimistic on TM- TSMC coop for optical is that if you're doing silicon photonics or if you're doing co-packaged optics or need these things, packaging in the same place and doing your manufacturing from your silicon wafers make make sense. And so I I get that, right? I like I I get the idea.

But what is going under his roof? Like it's highly unlikely Intel moves to your point, they're capacity short constrained uh manufacturing, so wafer equipment material into Terafab. But is it like he's going to build on their process? Is it that he's going to use some of their like It's very vague. It's very vague. It's something, you know? Yeah. I I want to I want to be positive about it. I do think it is important to remember not even a year ago Elon made a similar announcement with Samsung.

Yes. About using their their fab in in Texas. Right. And that doesn't seem to be playing out, which is in a very weird way actually good for Samsung cuz But it is like it's Elon. It's it is uh an ex- massive wildcard, a massive X Factor. Right? x.ai, get it? That's a little pun there. But the reaction the market's reaction was positive, right? Despite, you know, the news. They were like, "Hey, look." And and I think they landed similarly like that I did. Like this is you know, more technology validation for Intel, which again, to be honest with you, yes, we would love a wafer scale customer to Intel. Yes, we would love some big name to be using their packaging publicly. The reality of you know, this being a baby step that there's validation that something Intel has, process, package, etc.

Um is good. So, I I think that was positive. Again, we'd all love to know more, but it was directionally good. Agreed. And if if you look at what what X and Elon did with Samsung as as much as volumes are probably not what we thought they would be that announcement was a major like turning point for Samsung. Right. Right? Now, maybe maybe just optically, maybe substantively they were going to I'm sure they were going to get there anyway, but like that announcement was the start of like a real resurgence in their fortunes.

Like they they got Not long after that they got their process working again. They got HBM qualified. Like obviously that's all, you know, mostly on Samsung, but like I think it's safe to say that Elon had done some diligence on their process beforehand. And that served as a pretty important public validation of Samsung. And I think this is equally true for Intel. Yeah. Fingers crossed. Yeah. Agreed. And And you know, we have honestly, I I feel like they've they've been getting momentum as a whole. I mean, I do think, you know, again, the stock's up. I forget. I can look briefly for the week. Uh 19% actually. They closed at 62.38 today.

So, that's great. That's amongst the highest it almost been in some time actually to be honest with you. Ew. Yep. I mean, like literally since 2020. It's the highest highest they've been. Um I you know, and okay, so let's let's jump to Google. I Okay, and I'm just going to make a very brief analogy and in no way, shape or form am I saying these two things are similar. But like Anthropic today made another announcement that all of a sudden uh you can now work with Claude in Word.

Like it it's insane. They're they're usurping Microsoft's productivity suite. But they're just launching. Like they're just announcing after announcing after announcing. And to some degree, I feel like Intel is doing this in foundry. Like they actually have a whole series of things that are compounding upon each other. The Google announcement, yes, Google's using Xeon. We should we should all know, right? That it's not a shock when a hyperscaler works with Intel to do a somewhat bind up variant of their CPU.

Like every- Amazon does that. Like everybody does this with Intel. Um They do co-design what they call the IPU, which is basically a DPU if you follow this show and you kind of hear all of our love for a DPU. Uh their IPU is a DPU. And I think there's more to this announcement that led on. I I think there might be some packaging involved in this, maybe a little bit more customization that's done for Google on the on the IPU cuz again, IPU's are CPUs. They're not random things. They're CPUs architected to do network security, noisy neighbor, Kubernetes, multi-management, tenant management, all those things. Um But but I like that Intel sort of just keeps dropping announcements is my point. Like Yeah.

Anthropic like we just keep shipping. Intel keeps making announcements on things that are beneficial for for foundry or or the health of the business. Yeah, I mean, I I did some work on that exact subject this week just going through their announcements over the last like 3 weeks. Cuz you have you have Terafab. You have Google. You have their new chairman, which I think we talked about. That's a major positive sign. We talked about last week they got out of that deal in Ireland. Yep. Uh right? I mean, it's just been And then and then they have uh In addition to that, there are a lot of press reports about other customers kicking the tires. Like I have a whole list of who their announced customers are. And it's it's a surprisingly long list. And now, admittedly, it's mostly CPU custom CPU DPU customers. It's not foundry customers.

>> Yeah. Which is what we all want. Right? But then the you know, there are there are a half dozen or more companies where there have been credible press reports saying these guys are evaluating Yes. Intel foundry. Right. >> point that you have to think there's like some substance there. Like there's at least momentum. Yeah, I I agree and I and I do think it's it's it's not coincidence that you actually have seen and this was I don't know, maybe a quarter or so ago, folks at Broadcom, folks at Google, folks at others with in the job title say um understanding of Intel's EMIB.

And you keep hearing that EMIB is is becoming such a key part of of what's what's getting them in the door. And and this was their thesis a long time ago. Like I think I talked about it, especially when I toured the advanced packaging facility. Like their belief was people are coming to us for packaging cuz we've got some really unique differentiators there and that's opening the door to other conversations. Now now again, right? Zisner on the call was like, yeah, packaging could do a billion, like sure, we'd love that. But but but packaging is not wafers. So but but great, right? You're you're growing that that business. But if that gets you, in a customer conversation for an ASIC or custom CPU and and that's where they're at. And so I make that anthropic point to say I think they have a lot more interesting foundry announcements coming through the rest of the year while they're absolutely like TSMC not going to just name drop customers cuz they're not Hock Tan apparently who can just name drop customers.

Um I I think it will become clear who their customers are very quickly for those of us who will be vested to figure that out, but I think they've got the second half of this year will be very interesting for unpacking foundry and and momentum for for Intel foundry. Yep, I agree. I agree with that. All right, good stuff all right, good stuff on that. Um Jay has a thesis we're all going to hear for the first time. Ben hasn't heard it, so he's going to launch into his thesis.

All right, so my thesis isn't so much of thesis, it's just a reality that I've been struggling with a lot lately. Which is if you look at the you look at a lot of the stocks in in semis right now, there is this weird mix of valuation where you have some very prominent companies who we all know are doing incredibly well trading at relative to their history low multiples.

Right, relative to the S&P low multiples. And then you have a bunch of other companies whose prospects are further out, less proven and their multiples are through the roof. Right? Really crazy ones. And I'm I'm I'm deliberately avoiding names cuz I don't want to give any I'm not giving investment advice. Yeah. But it is this

weird disconnect in the market that I'm I'm struggling to wrap my head around. And I think a lot of it just comes down to this expectations game. Right? Where we all kind of have a good sense of where the sort of the biggest chip companies, especially in the compute side of things, we have an idea pretty good idea of what they're going to do this year.

What while as like some of the other ones, the networking names the optical components and optical systems that's like they they seem to have capacity. Right? TSMC is sold out. A lot of other companies aren't sold out yet. Uh and this is this demand is sort of saturating has saturated compute and now demand is spreading out into other areas creating a lot of expectations. And it is it is very confusing, I would say, for investors. Certainly confusing for me.

Yeah, so I know the names James Jay is talking about are relative to optical names. So do your own assumptions there. I I think what's come down to, to be honest with you, is like so so I agree with you on expectations. I think what is interesting is when you unpack, I think some of the market narratives there's this like thesis similar to use the name as to who owns the scarcity.

And then who has to some degree some monopoly on that scarcity. Now rightly or wrongly there are some names in optical that have believed to be like the only ones who can do X, Y, and Z and that's not true. But the reality is that's where some folks have landed. And there's some of that emotional sentiment, you know, that you've talked about right before. There's some of that Yes, there's a lot of retail people pumping these up. I've even seen the bigger we won't to get name them uh influential substacks, if you will, of the last few weeks do some things where they're shorting or longing like some of these stocks. So okay. Um But but I think it I think there's such a desire to kind of make a pick on a name who essentially's got call it like the the edge on something that's scarce.

And with an optical this is an interesting debate, right? Because this is very difficult. And we are not sure which direction it's going to go yet. And someone might win and be a very big winner, but we don't know who that's going to be in. So I think there's those dynamics. But then there's also to your point about like the Nvidias and the Broadcoms and some of the other of the world, which is like I mean, do you not think like are you are you discounting really the trajectory that they are when they're actually the the gatekeepers for the success of those optical names, for example? Um They're the ones that will define that path and to some degree pick those winners.

And I would argue they also have a fair element of control of scarcity given their ability to secure the supply chain. So I don't know. I mean, I'm with you. It's it's weird. I mean, I watch these stocks regularly and I see a couple of these names and it's like 400% in 6 months and you're like Okay. Okay. You certainly aren't making that much money in the next 2 years, man. Yeah, it's it's it's very I mean, especially we're also getting to the point where we're plumbing the very dark corners of the industry.

Um where, you know, a lot of the substacks I'm on and Twitter and Discord servers I I lurk in we're having debates about like really really violent debates about some very obscure names that no one would have talked

about 2 3 years ago. Right? Mostly in optical and test and optical test, but other areas, too. And yep. I And it's it's just it is I I'm I'm somewhat cautious on parts of the sector.

And it is a little bit concerning to me that there are some some like a lot of really what we would consider high quality names, companies that you just know are going to be around in 10 years trading at massive discounts to companies who may not be around in 10 years. Yeah. And it is cognitive dissonance and I I'm I'm not sure what to make of it. It's just a it's a weird time in the market. Yeah.

I I think that's true and just being, you know, sort of aware of that I think is important as you, you know, factor through this and as folks do do some research. Um But, you know, like I said, at the same time it's like the echo chamber that is Twitter, which and I just can't emphasize this enough, has become a fairly vocal place in which some momentum does go to things for that is just the reality of it. Um It it just moves really quick, you know, in some of these things. Like emotion and sentiment can move quickly. And it's it's just it's just it's just interesting.

Um so, you know, be aware of that obviously I think is important. And stay on top of the the ever-changing dynamics of But but what's interesting, I will say like I I I don't know if you've seen it, but I have had a lot of conversations with investors and I think I talked I think I had said this a while back. Like it's actually interesting to me that so many, you know, people on the equity side are actually bringing in technical people to be a part of their investment team.

And what that has led them to do is to and so I use this like, you know, obviously there's big picture and then, you know, for a long time I was saying like, hey, let's get in the weeds in semis. And then when we start talking about like testing and substrates and all these things that we're talking about tungsten like we talked about, you're talking about the soil now. And you're having people go like the weeds aren't enough anymore. We need to get into the minerals in the and the soil of the semiconductor industry and the supply chain.

And that just opens up a whole another bottle of you know, can of worms. And so I just find it fascinating that that's where now these conversations have led is like, no, no, no, I don't I don't want to talk about manufacturing anymore. I don't want to talk about wafer equipment. I want to talk about minerals. I want to talk about, you know, testing. I want to talk about how we do like I'm going to find every bottleneck I can.

In the soil. Yeah, I love that metaphor. Although I do I do want to say I need I owe an apology to the tungsten community. Tungsten is very important. I mentioned last week there was a supply that I didn't really know what it was used for. Now I know because you've all educated me and I'm sorry. Tongue twister. >> J has J I I mean I didn't know either when we talked about this. It's it's fascinatingly important, especially to memory.

>> But but we're at the point now where we're not talking about chips, we're talking about Exactly. >> elements. Not not even like how many wafers or who makes the wafers. It's like no no no no. I want to talk about the the all of the nitty-gritty of uh of into the soils and the mineral composite of the soils that makes the semi nectar and any I anyway. Everybody's going deeper. That's my high-level point and I think it's great.

I also think in that depth becomes a whole lot more unanswered questions. Like we don't know you know, just I'm a gardener to use the garden like you don't always know what's going to sprout from the mixture in your soil. >> [laughter] >> Oh, that was good. Um All right. So, future of the data center I wrote a couple of pieces in the past couple of weeks one on 800 VDC and uh one this week on liquid cooling and it was sort of like I said my future of data center series and the and the more I unpack this there's sort of two conclusions I think are very very interesting.

Um the first is that the reason in which you might move to 800 VDC and then again start to use more liquid cooling across the board is is yes, there is a economic savings, right? To the to the to the data center. You you could walk away from this saying I just want to do that so I can save money on my gigawatts or milliwatts of compute. That's that's one angle and and that does exist. But what this really does and where everybody sorts of lands and where they get excited is that that savings in energy you can actually convert to more compute. So essentially there's a variable model of right how much I how many more computer racks I can get if I go to 800 VDC or if how much I can get if I go to liquid cooling and and it's the capacity unlock for more compute that everybody that kind of overshadows any economic savings benefit. And and I thought that was interesting cuz where that then took my brain was okay, so then let's just say that this entire thing, right? 800 VDC plus the infrastructure that you decide you decide and then the liquid cooling that sits alongside of that all unlocks more your ability to put more compute under a thermal and a wattage uh budget that you have for your data center. So, that then means this becomes a TCO conversation. If it becomes a TCO conversation to the degree that we've talked about right in video being a completely co-optimized rack then the future of the data center is actually a co-optimized thing. So, like you are designing this in cohort with the infrastructure decisions that you're making with your power and your thermal budgets and the infrastructure that goes to that 800 VDC and how you deploy liquid cooling tightly integrated together in order to just squeeze tokens per dollar per watt out of that out of that data center. And so, I I say that not to say like you know, it it I mean well there is a degree of this that's being designed with Nvidia to to be honest with you. So, like there are data centers that will just be plug and play 800 VDC with liquid cooling with Nvidia racks designed for maximum TD as a complete co-design.

But that the hyperscalers I think are to to a degree I hadn't appreciated verticalizing even more their work with the liquid cooling ecosystem and the power ecosystem deploying 800 V D in order to again squeeze every bit of value of that megawatt. It's not just cobbled together from merchants anymore is my point. Like this is becoming an extreme co-design the entire data center footprint, not just your infrastructure. I I agree with that. I I the one place I would quibble is you said now now that the only thing matters is TCO and performance per watt. And I I would say it always has been that way, right? We've we've been for at least 10 probably 15 years power has been the sort of key bottleneck in data centers. Going back to pre-AI cloud just sort of cloud buildout.

And what's what's happened with AI is just supercharged all that. There's so much money being poured into it that it's driving everything acceler- it's accelerated that trend extremely. And we're in a place now where necessity is the mother of invention. Like you you're everyone's limited don't have enough tokens which because they don't have enough compute because they don't have enough power. And so you're turning all those all the knobs you can now extreme co-optimization you're turning you're optimizing every part of the data center that you can. So, it makes a lot of sense that you're going to work more closely with all those companies.

And I think what's what's really interesting is especially around something like 800 volt data centers is that that's been a topic for forever for 20 years people have been talking about that. And it would have happened eventually but now because there is so much demand and the money to supply it that's providing a forcing function that means it's going to happen in two years and not 10. Yeah. Yeah. And you know, and I it reminds me a lot of like the US energy market where same thing we it data centers are driving demand for power and so we're starting to see a lot of changes to the grid that people have talked about for even longer.

Uh and we're finally now having the demand and the capital to to make some of those changes and try out new new systems and bring new things on on grid. And so, like you know, if if and if and when if or when the AI build slows down at the very least when it all comes crashing down we'll still have better a better better power grid. Hopefully cheaper. For those of us in California. >> [laughter] >> Um you know, I I I agree with you. I I and so again, I'm not going to disagree with your TCO point. Yes, that's true. I I would just say that the data centers of past um well well firstly not not as complicated when it came down to thinking them like a system like of a like a system of a fully functioning system.

>> Right. And and two, I think you could sort of just do and I again, I have read I have a friend at Schneider. Uh I have a friend at Eaton. I have a friend at Rosen Den. So, like we've had these conversations about they have they have co-designed for a long time. But I would say that there was still a degree of you could just kind of pick your random merchant for a standardized power system.

To I my point is I think where we're going is that like these data centers become and and and who they pick, whether that's Schneider, Eaton, uh Vertiv, etc. is actually a designed in partner. Meaning that that's a long-term conversation, not just pick my price, pick my vendor, build my data center. And and to me that's a fundamental shift of how these things get architected which is which is just fascinating. Because again, right? I hate to use this analogy and I'm sorry if I did it before, but like the way you think about a system of compute or even just an SOC, right? An SOC becoming a system of transistors working together each doing different things, CPU, GPU, NPU like the data center sort of just becoming like that system. It is the system's optimization working together as a compute unit which gets you your efficiency. And and that I just think that's a fascinating change at a at a at a footprint level.

Build this giant factory that is solely designed to produce the best tokens per watt and you have systematically designed that thing. Like with with a lot of intention. You know, purpose-built for that and I think that's that just to me it has a lot of implications about how I think about the future of the data center. Yeah. I I I got to give a shout out to Doug O'Laughlin at Fabricated Knowledge. We wrote it when he was at Fabricated Knowledge and now he's at SemiAnalysis.

But he he called this two years ago where he said what we're seeing in designing data centers is just and it was very much led by Nvidia then is Jensen is now he's he's run out of all the optimizations he can get on the chip and he's optimized the server and the rack and now the next thing is he's going to optimize the whole data center. Um and it's a natural outgrowth of that. Yeah, and and to be honest with you, so dang, I'm sorry to the anybody who tracks this with Nvidia, but they have a they they have a piece of software that does a digital twin

of data center Nvidia where you can sort of build and configure these things and then see how network traffic goes.

Which on that aside, I I hear a fairly popular like video game out there is one where you like can start designing data centers. It's like it's like Farming Simulator or or like the one where you were an air pay pilot like having to schedule routes. Like you're building data centers, moving traffic and you can watch the show. Like people are just sitting there like enjoying Yeah, it's even it's even more than that. It's it's you're you're you're cabling. You're you're plugging in all the cable.

Right? Which I think if you build a data center that's like the most tedious part of it and now it's like a super popular video game. Oh man. But but Nvidia has the software that does like this digital twin simulator where you can just kind of put all these pieces together and basically see what you're going to get in TCO. And so anyway, I I just I think this is fascinating. Like like again, how much it's moved from the compute layer to the physical foundation of how you build data centers now in such an extreme co-optimized way to use you know, Jensen's nomenclature. Um it's just interesting. With again, right? The secondary point to this, which is like data centers built today are outdated next year.

And so then you're sort of like all these configurations I'm thinking about today for my next X number of gigawatts, when I go build greenfield builds next year, I'm talking about very different things. And the dynamicness of this I think is is extremely interesting. Oh, yeah. And it I know people who are in those positions planning those things out and like they have you know, they they have to deal with 5-year roadmaps for not just for computer anymore, but for all the things we just talked about.

Right? And it's that's that's like that's a big task, right? Especially cuz some of these things are sort of at the edge of techno- technological possibility. You're on that frontier and you're like, well, if I want my data center that I'm going to I'm not going to start building it for 3 years, I and when it is done completing a year after that, I need to make sure it's at the cutting edge. Like there's a huge amount of technical risk in that. I mean and and just execution risk, like is the power going to be ready? Is the interconnect going to be ready?

Is is this fancy thing I'm co-designing with my power systems provider, is that going to be ready? Right? On top of all the normal sort of things like is whatever flavor of HBM is coming out, is my chip going to be ready? That's a That's a that's a big task. There are a lot of people deeply deeply involved in it right now. It's pretty fascinating. Yeah. So I did want to go back to this Claude announcement just because yes, I get that they're shipping quickly. Yes, there's Mythos and we talked a little bit about that last week, but the you know, if you didn't see it, they basically have a connector that within Word, a Word document, you can use Claude. Now we've known they had this in Excel.

And I say this because they are usurping, to use like one of those old medieval terms, they're becoming the great usurper of productivity from in this case Microsoft, embedding themselves as that functional operating system layer of productivity that you rely on, and disaggregating others to just be the software platform. They're the more valuable layer that sits on top of that. And I just the one the pace that they move, two, the fact Microsoft

or others who should be doing these things are not doing these things. And three, that it's becoming fairly clear, like the more you get invested in Claude and yes, they don't have as much compute and there's this, you know, maybe we talk about it too, this rumor about them building their own ASIC, yada yada yada, designing their own ASIC to be built by someone else.

Um If they do this, like they really put themselves in sort of a position to again usurp software's value and capture it themselves. And I I just think this has so many strategic implications to the software world. Again, not that it goes away, but that the value capture shifts. Cuz again, right? It's not like you're not going to stop using Word. You're going to use it. You're going to pay that fee. You are however going to use all your your value, what you walk away with being like that was useful, I got this done, is Claude, not Word.

So one, full full credit to Anthropic. They're on fire. Cuz they had it I mean they they unveiled Mythos, their all-time planet killer model. Um Planet killer. Just which I which it like I'm so torn cuz I'm like, that sounds really scary. I really want to try it. Um so they have that, they have they have a Word connector. They have a bunch of other things like uh Hermes and there's a multi-agent tool that they have now that does I mean they're they're basically going to displace open Claude.

Um so they're on they're on fire. Full credit to them. I I there's two things I take away. I agree with you that what's taking place now in AI and its impact on the software world isn't we're not going to do away with all software, but we're going to see massive shifts in market share and where in or share of wallet and where where people are spending their money. I think that's that's very clear. And it's very clear that it's going to happen. It's very unclear what the end result will be.

Um but certainly at this moment in time Anthropic is in the lead and seems to be accelerating. The flip side of that though is another complaint that I've dealing with a lot of things I I read is the extent to which uh over the last, call it 2 weeks, 10 days, core core Claude products like Opus have have been giving noticeably poorer results. Yeah, absolutely. Right? I've seen it firsthand. I went and talked to a bunch of people. Same. They're all seeing it.

And it's still pretty good, but it is not like amazing the way it was 2 months ago. And and and I think that's interesting because it reminds me so much of Telco. I was going to say this. You took it from me. You st- took it from me. It's it's Telco, right? You get on you get on the first 3G network or the first 4G network, the first month it's incredible data rates. Just wow, this is so fast. And then everybody else gets on it too, and it's like, oh, this is just kind of average. Um cuz it's >> get get away.

I I have I have I have to steal your momentum by going even deeper. My thought was the internet. Remember when like you couldn't get on the internet in the middle of the day? You go to a website and it was super slow, but then at night you could use the internet. >> [laughter] >> It's like so like it's to me like now I'm like, oh, I'll use Claude at night cuz it's you can use GPT during the day or chat as they in Open AI like to call it cuz they got capacity.

Claude's got no capacity. Everybody wakes up by 8:00 a.m. on the West Coast time, people are pounding Claude all day, and you can't use it. So like I got to use Claude at night for my deeper, you know, agentic stuff because it's just like the internet. Couldn't browse in the middle of the day, but I could browse the internet at night when it when everybody was offline cuz it was much faster. We we clog it we clog in the pipes. We clog in the compute.

People doing their freaking agentic things, man. So I think for for us here, what where that matters, like sort of bringing it back to semis, one of the things that I I saw in the the press materials they put out for Mythos, they put out a comparison of different uh token usage per Anthropic model. Okay. >> And the extent to which Mythos is can can do much more more complicated tasks for far fewer tokens.

Okay. Right? Which right? They're they're like, call it double the performance for fewer tokens than even Opus can do. And so to me I think that's an important reminder that there is this race that's going to take place between we don't have enough compute. We just all the things we just talked about for the last 40 minutes about everything people are doing to optimize to get more to squeeze more compute out of it. That's going to take place, but at the same time all the software companies are going to do find every trick they can. They're going to do their part to make token usage more efficient. So you need fewer tokens. So you need less compute.

Now it doesn't mean we're going to stop buying chips, but like those those two things are happening at the same time. And for the moment what we're doing is we're we're coming out with things like Mythos, which are more efficient on a per token basis, and we're just going to do more and more compute with that. That's the tradeoff we're giving away. We're not we're not going to lower the prices. We're just going to do more and more compute.

Mhm. All right? And those things are going to keep going on. They're going to continue to co-evolve. I mean we I I I so I heard like this is going to be one of them situations where token prices are astronomical, too, despite what you just said. Like it's going to be expensive at le- at least right? A- and that most people probably won't see it in its full, right? They'll distill that down to whatever you get in a version of Opus or something like that. So it's still be capable, but but again, ver- very different in its ability to just do bolt bolt chat or tasks one shot at if you will, right? From its from its reasoning capabilities. Um but we want like we want token costs to go down. Like that's the thing I keep looking at these and I'm like, look, a model comes out, it's high per, you know, X number per million tokens. It's competitive. You're you're going to have premium tokens always. I believe this like hands down. Somebody wants to pay X amount to get this thing done in an hour versus wait 2 days, they'll do that. But that might not be everybody.

But this this like we we right now, I mean I don't you've tried this too, but I've talked to friends who've been doing this, like they go out, they don't just do what they're doing in Claude or GPT. They use the API, right? To build some software. They have no idea how much it's going to cost. They just task it to go do it. They come back. They got a bill for \$250. And you're like, it's whether or not, right? It actually got to the outcome that you want. So like we need people should not think about their token budgets. That's a question in enterprises circling this week, the these last few weeks.

Um but we need those costs to come down. That'll help it diffuse cuz right now it is still very expensive to serve at scale um your enterprise. And to me that's an issue for adoption. In some ways that's no different than cloud, right? Nobody you know, that's a perennial complaint. You don't know what your cloud computing bill is going to be in any given month. That's true. But this could get it's more expensive right now is my point. Like I think it'd be nice if it was at least somewhat I mean it's not going to be predictable, but it's not astronomical.

versus like Yeah, I mean I I think in some senses like Claude in particular I know I can track the number of tokens I use. There's a tool to do that. Yeah. Um but I do have that problem where I'll like spend all day and I'll I'll like try to like I This matters cuz Opus wasn't working that well this week. I had it build something for me. It wasn't quite right. I had to iterate through it several times.

And then I I I hit my limit without completing it. And there is that sort of ongoing frustration. But at the same time you have lots of news this week about token usage per employee. Right? Meta put out some Anthropic put out some numbers about Meta. Yeah. >> And Meta employees are using on average 750 million tokens a month, I think. >> Yeah. Yes. >> Which like to me is like sounds like rookie I know. Right? And And you have lots of people saying, "Oh, you know, I We're There are lots of companies now who are beginning to measure their employees usage of AI tools because it's seen as such a productivity enhancer.

And I I sus- strongly suspect that most of what the most of those companies are doing are at the moment they're they're measuring uh token usage. Oh, look, employee number seven is using a lot of tokens. They must be really good at AI. And that that That's not going to last cuz cuz I can get I can use a lot of I can use a lot of tokens on really stupid things and not be >> Yes. I agree. That that was my concern with this where I was like, "Just because you use a lot of tokens doesn't say anything about your productivity."

You could be You could be using those and and not coming to any decent outcome cuz you're not doing It's not a It's not an efficient way to use tokens, right? The best people will get more done with less tokens. Like that will be the skill that you have in your ability to either preempt or prompt or work with AI as a collaborator to get more done with less tokens. You be be more efficient with your token usage. So, I agree with you.

It was like a little bit of a silly and disingenuous metric today. Um but you know, it's Yeah, like I said. Anyway, but but you're right. Compute-wise up and to the right still for freaking ever. To the moon. Not capex to the moon. Remember rises to the moon. GPUs it to the moon. Data center costs to the moon. Um On that, before we go, I'm just going to tease what I'm publishing next week. So, I had not seen an updated model on cost per transistor in some time. There's generally only a handful of people who did this and several of them are retired.

Um so, I'm working on one. And let me just say when I say cost per transistor has gone up, it has gone up a lot. Like a lot a lot. Like way more than we anticipated a couple years ago when we were just like, "Hey, the angstrom era, chipllets, like they're going to cause like it that all them models are gone. Like However, the point that I'm going to flesh out in this in this piece is that it has fortuitously arrived at the right time in which there's so much value to AI that you can absorb those costs because of your ability to capture that value on these chips. If this was five years ago and I told you the number of cost per billion trans- you'd have been like, "Everybody's

screwed. Nobody's going to pay for that. There's not enough value in getting out of that compute." But today, they're very lucky that this all worked out. So, all of that to again say something we've been fond of saying, this is not your grandfather's semiconductor industry. This is something different in almost all of these capacities. So, I'll share more about that next week, but it is quite the jump in cost per billion transistors.

Uh so, let that marinate. >> [snorts] >> Okay, then. Okay, then. All right, thanks for listening everybody. We will talk to you next week. Tell your friends. Leave a review. Tell your agents to tune in to us.