Huang Isn't Planning to Wake Up a Loser

MBI DEEP DIVES
APR 18, 2026 • PAID

14 3 Share

You may have already listened to Jensen Huang on the recent episode of Dwarke podcast. If you haven't, I highly recommend you listen to the entire episode and just the bits and pieces clips. This wasn't the kind of episode that you want an A summarize either.

What made the episode so enjoyable to listen to is I thought Dwarkesh did an ex job in eliciting Huang's opinions on a range of topics that are hotly debated. Fra speaking, I came to the conversation with a slight bias to Huang's point of view selling chips to China, but came away thinking Huang's arguments weren't conv enough.

Here on MBI Deep Dives, I am more interested in micro economics concerns su competitive dynamics, moats etc as I consider the broader macro, geopolitical concerns are largely beyond my paygrade even though they can be quite impactf times. There were plenty of materials in the episode that were very much in my micro economics interest even though most people are mostly talking about the around selling chips to China.

Speaking of competitive dynamics, the below excerpt on Huang's opinions about Trainium and TPUs are particularly worth highlighting:

> I would welcome Trainium to demonstrate their 40% that they claim all the ti would love to hear them demonstrate the cost advantage of TPUs. It makes n sense in my mind. It makes absolutely zero sense. On first principles, it make sense.
>
> So I think the reason why we're so successful is simply because our TCO is so Secondly, **you say 60% of our customers are the top five, but most of that bus**

> **is external**. For example, most of Nvidia in AWS is for external customers, not internal use. Most of our customers at Azure, obviously all of our customers are external. All of our customers at OCI are external, not internal use. **The reason they favor us is because our reach is so great**. We can bring them all of the great customers in the world. They're all built on Nvidia. And the reason why all these companies are built on Nvidia is because our reach and our versatility is so great.
>
> So **I think the flywheel is really install base, the programmability of our architecture, the richness of our ecosystem**, and the fact that there's so many companies in the world. There's tens of thousands of them now. If you were one of those AI startups, what architecture would you choose? You would choose an architecture that's most abundant. We're the most abundant in the world. You choose the one that has the largest installed base. We're the largest install base, you'd choose the one that has a rich ecosystem.
>
> So that's the flywheel. That's the reason why, between the combination of: one perf per dollar is so great that they have the lowest cost tokens. Second, our perf per watt is the highest in the world. So if one of these companies, if our partners, has a one gigawatt data center, that one gigawatt data center better deliver the maximum amount of revenues and number of tokens, which directly translates to revenue. You want it to generate as many tokens as possible, maximize the revenues for that data center. We are the highest tokens per watt architecture in the world. Last, **your goal is to rent the infrastructure, we have the most customers in the world**. So that's the reason why the flywheel works.

Jensen correctly notes that Amazon (AWS), Microsoft (Azure), and Google (GCP) are buying mountains of GPUs primarily because their **external cloud tenants demand them**. Hyperscalers, simplistically speaking, are increasingly acting as capital-intensive GPU distributors for Nvidia. It may take a while, but I still wonder unless we get past this compute constrained environment, the true competitive dynamics between ASICs and Nvidia GPUs may be hard to ascertain. In the current environment, you will probably sign a contract whoever can offer you compute. As a result, Trainium, TPUs, and GPUs all can coexist and thrive in this environment. I am noticing a growing

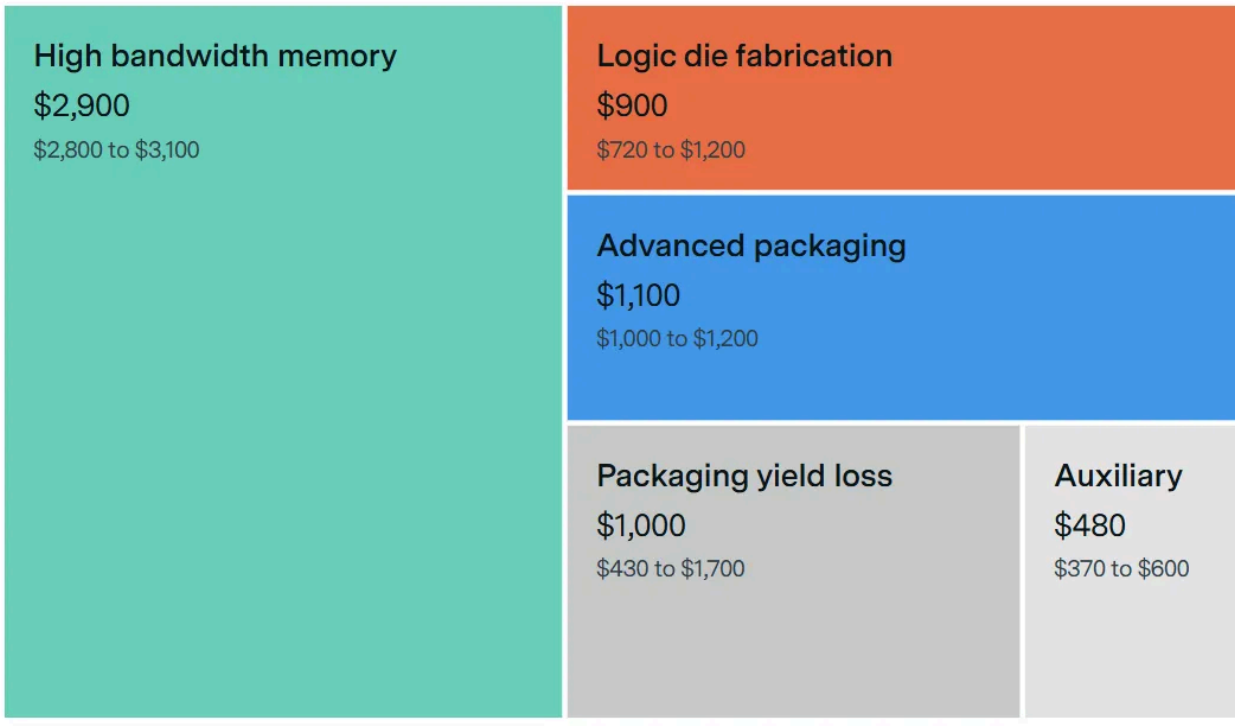
consensus among investors that "we may never have enough compute" which im worrying about competitive dynamics in a relatively compute abundance scenari largely irrelevant. I, however, lean towards believing in the power of capitalism a there is hardly any shortage that capitalism has failed to eliminate over time. As result, while it's excruciatingly challenging to pinpoint a timeline **when** we will r such relative compute abundance scenario, I suspect we will eventually have to d with such question. If you value these businesses assuming such question will ne arrive but it does arrive five years from now, that has much more valuation implications than the most investors may imagine.

I continue to think that in a relative compute abundance scenario, TPU is the mo protected precisely because Google has enormous captive internal workloads: Se query serving, YouTube ranking, ad targeting, Gemini across every surface, Wayı and Workspace AI features. Even if external TPU demand goes to zero, internal workload volume alone justifies continued investment and tape-outs. They're obviously not immune, just less exposed than Trainium and Nvidia's GPUs. My g is DeepMind would be very happy to get higher compute allocation if Google Clo faces any difficulty for selling compute to external customers at attractive ROIC.

Huang also made an interesting point about ASICs vs GPU margins:

> "don't forget, even in ASICs margins are really quite high. Nvidia's margin is l let's say. But ASIC margins are 65%. What are you really saving?
>
> You've got to pay somebody. I think the ASIC margins are incredibly good, fro what I can tell. They believe it too. They're quite proud of their incredible ASI margins."

I found this argument to be a bit misleading. Epoch AI [suggests](#) that Blackwell c sell for $30k to $40k whereas the bill of materials (BOM) costs only ~$5.7k to $7.3 which implies an eye-watering 82% margin. They, however, did point out that "*Si most Blackwell revenue comes from servers and rack-scale systems, which may carry [low margins](#), NVIDIA's realized margins on Blackwell sales may be lower than these chip-lev estimates.*"



Source: Epoch AI

When a hyperscaler pays Nvidia $30k+ for a Blackwell unit, Nvidia's margin is captured on the entire bill of materials plus the software stack. Broadcom's or Marvell's "65% ASIC margin" is on a much narrower slice. The hyperscaler pays TSMC directly for wafers and pays SK Hynix, Samsung, or Micron directly for H which together are the majority of the BOM on a current-generation accelerator. Broadcom's margin sits on top of its contribution, not on top of the whole chip. S "65% vs 70%" makes it sound like near-parity, but the 65% is applied to perhaps a quarter of the system cost while the 70% is applied to all of it.

A hypothetical chip with ~$6k of real BOM cost sold by Nvidia at $30k captures S of margin per unit. The same chip produced through a design-services relationsh with Broadcom might carry ~$8k of Broadcom margin plus the same ~$6k BOM, landing the hyperscaler at $14-15k all-in. Even if Broadcom's percentage margin slice is as high as Nvidia's, the dollar savings per unit are enormous because Broadcom's slice is so much smaller.

Custom silicon programs do, however, carry real costs that doesn’t get surfaced i simple framing. There is internal engineering headcount, inference and training framework support, and the opportunity cost of engineers who could be working

revenue-generating products instead. These are fixed and amortized, but they ar zero, and for a program that fails to reach sufficient volume, the amortized per-u cost can be meaningful vs just merely comparing the BOM. This is the actual rea most companies do not build custom silicon because you need Google-scale or A scale internal demand **AND** a top-tier engineering team to make the amortizatio math work. For the handful of operators where the math does work, the savings a not 65% vs 70%. They are likely closer to single-digit-thousands vs tens-of-thous per chip, and I'm sure Huang knows this.

However, there are different layers to this debate. Gavin Baker pointed out that v model portability used to make investors think model developers might gain leve over the chip suppliers over time, that may be changing with Blackwell and even so with Rubin. From Baker's post (emphasis mine):

> As system level architectures diverge (torus vs. switched scale-up topologies, memory hierarchies, networking primitives), **true portability is eroding.** The and **Mi325 had roughly the same scale-up domain size as Hopper while Blackwell's scale-up domain is 9x larger than the Mi355 scale-up domain**, et
>
> Many frontier models are now being explicitly co-designed for inference on s hardware like GB300 racks. Codex on Cerebras is another example. **Those mo run less efficiently on other systems and the performance differentials will widen.** A model that runs well on Google's torus topology will run less efficie Nvidia's switched scale-up topology and vice versa - the data traffic is fundamentally different as a byproduct of the models being parallelized acros different topologies.
>
> Google's internal teams - and increasingly the Anthropic teams as they becom most important customer of almost every cloud - have the luxury of operating across the stack (models, chips, networking) - but that is not the case for the r the market and other prospective users. Anthropic is the exception, not the ru **wit, Anthropic and Google allegedly have a mutual understanding where Anthropic can hire the TPU engineers they need every year to ensure that tl**

> **can continue to get the most out of the TPU.**
>
> Given the overwhelming importance of cost per token to the economics of the models will be run where they run best. Most extremely large MoE models wi best on GB300s given the importance of having a switched scale-up network l NVLink for MoE inference. **When training was the dominant cost for labs ar power was broadly available, labs were optimizing to minimize capex dollar Model portability was a way to create leverage over suppliers. I think that dr lot of the focus on portability.**
>
> Today, inference costs as measured by tokens per watt per dollar are everythir **Inference is way more important than training costs (inference is effectively part of training via RL). Labs are therefore now optimizing for inference. Th means increasing co-design and higher go-forward switching costs for indiv models between systems. I do think this explains why Anthropic and Nvidia together**: Anthropic needed Blackwells and Rubins to inference at least *some their models economically. And Mythos might just end up being released coincident with the availability of Rubins for inference.
>
> TLDR: **as labs shift their focus from training to inference, the costs of porta and the upside of co-design to maximize tokens per watt per dollar both rise Portability is likely to begin decreasing as a result.**

Reading Baker's post made me think that the hyperscalers will have to pay the "N tax" whether they like it or not; as long as the hyperscaler customers see **materia** benefit for using Nvidia GPUs over hyperscaler ASICs, hyperscalers hands will b And if hyperscalers appear unwilling or hesitant to pay such exorbitant taxes, Nv will be happy to allocate more of their chips towards neoclouds, and many custo especially AI startups will likely follow. Always remember, "[You're not talking to somebody who woke up a loser]."

*In addition to "Daily Dose" (yes, **DAILY**) like this, MBI Deep Dives publishes one Deep on a publicly listed company every month. You can find all the 67 Deep Dives here.*

Current Portfolio:

Please note that these are **NOT** my recommendation to buy/sell these securities, just disclosure from my end so that you can assess potential biases that I may hav because of my own personal portfolio holdings. Always consider my write-up my personal investing journal and never forget my objectives, risk tolerance, and constraints may have no resemblance to yours.

My current portfolio is disclosed below: