December 25, 2024

Nvidia’s Christmas Present: GB300 & B300 – Reasoning Inference, Amazon, Memory, Supply Chain // Blackwell Delays, Microsoft Orders, GB300 BOM, Nvidia Gross Margin, ConnectX-8, VRMs, Micron, Samsung, SK Hynix, Wistron, FII Foxconn, Aspeed, Axiado

8 minutes

7 comments

By Dylan Patel, Myron Xie and Daniel Nishball

Merry Christmas has come thanks to Santa Huang. Despite Nvidia’s Blackwell GPU’s having multiple delays, discussed here, and numerous times through the Accelerator Model due to silicon, packaging, and backplane issues, that hasn’t stopped Nvidia from continuing their relentless march.

Aug 04, 2024

Nvidia’s Blackwell Reworked – Shipment Delays & GB200A Reworked Platforms

Dylan Patel, Wega Chu, Daniel Nishball, Myron Xie, Chaolien Tseng

They are bringing to market a brand-new GPU only 6 months after GB200 & B200, titled GB300 & B300. While on the surface it sounds incremental, there’s a lot more than meets the eye.

The changes are especially important because they include a huge boost to reasoning model inference and training performance. There is a special Christmas present from Nvidia to all the hyperscalers, especially Amazon, certain players in the supply chain, memory vendors, and their investors. The entire supply chain is reorganizing and shifting with the move to B300, bringing many winners presents, but also some losers get coal.

B300 & GB300 – Not Just An Incremental Upgrade

The B300 GPU is a brand-new tape out on the TSMC 4NP process node, IE it is a tweaked design, for the compute die. This enables the GPU to deliver **50% higher FLOPS** versus the B200 on the product level. Some of this performance gain will come from 200W additional power with TDP going to 1.4KW and 1.2KW for the GB300 and B300 HGX respectively (compared to 1.2KW and 1KW for GB200 and B200).

The rest of the performance increase will come from architectural enhancements and system level enhancements such as power sloshing between CPU & GPU. Power sloshing is when the CPU and GPU dynamically reallocate power between the CPU and GPU

In addition to more FLOPS, the memory is upgraded to 12-Hi HBM3E from 8-Hi growing the HBM capacity per GPU to 288GB. However, the pin speed will remain the same so memory bandwidth is still 8TB/s per GPU. Note Samsung is receiving coal from Santa, because they have no shot at getting into the GB200 or GB300 for at least another 9 months.

Furthermore, Nvidia, because they are in the Christmas spirit, the pricing of it is quite interesting. This shifts the margins of Blackwell, but more on pricing and margins later. Most important to discuss first is the performance changes.

Built For Reasoning Model Inference

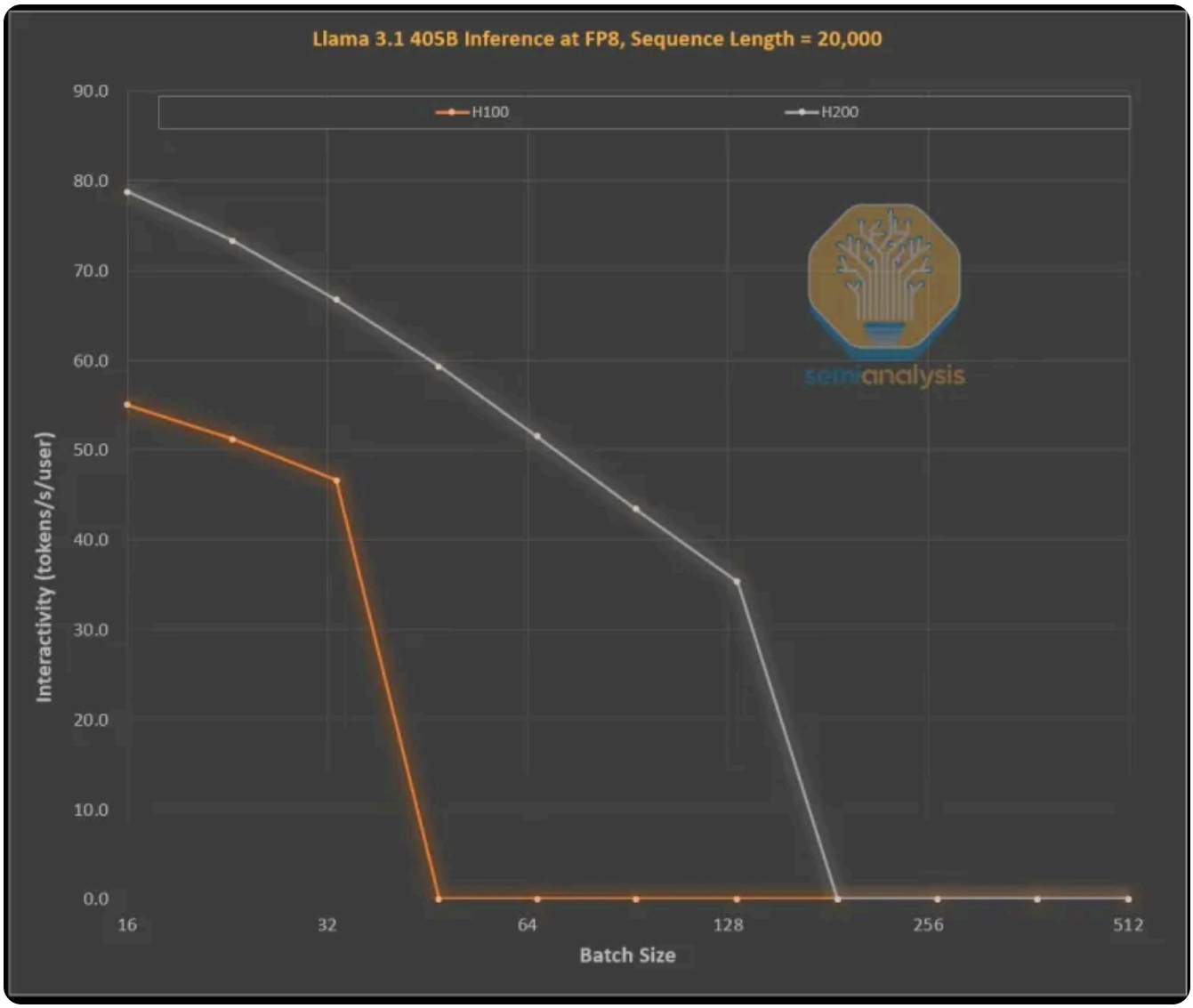
The improvements to memory are key for OpenAI O3 style LLM Reasoning training and inference due to long sequence lengths growing KVCache, limiting critical batch sizes and latency. We explained this in our Scaling Laws defense piece where we discussed reasoning model training, synthetic data, inference, and much more.

Dec 11, 2024

Scaling Laws – O1 Pro Architecture, Reasoning Training Infrastructure, Orion and Claude 3.5 Opus "Failures"

Dylan Patel, Daniel Nishball, AJ Kourabi, Reyk Knuhtsen

The chart below shows improvements to tokenomics through Nvidia's current generations of GPUs running on 1k input tokens, 19k output tokens, which is similar to a chain of thought in OpenAI's o1 and o3 models. This demonstrative roofline simulation is run on LLAMA 405B at FP8 as it is the best public model we can simulate, with H100 and H200 GPUs, the GPUs we have access to.



Source: SemiAnalysis

When going from H100 to H200, which is purely only an upgrade with more, faster memory, there are two effects.

1. 43% higher interactivity generally across all comparable batch sizes due to more memory bandwidth (H200 @ 4.8TB/s vs H100 @ 3.35TB/s).
2. ~3x reduction in cost due to H200 running higher batch size then H100, enabling generation of 3x as many tokens per second. This difference is primarily because of KVCache limiting total batch size.

The dynamic of more memory capacity offering a seemingly disproportional benefit on to are massive. The performance and economic difference for the

operator between the two GPUs is much larger than what the paper specs suggest:

1. Reasoning models can be a poor user experience due to significant waiting time between requests and responses. If you can offer significantly faster reasoning time, this will increase the user's propensity to use and pay for them.
2. A 3x difference in cost is massive. Hardware delivering 3x with a mid-generation memory upgrade is frankly insane, way faster than Moore's law, Huang's Law, or any other pace of hardware improvement we've seen.
3. We have observed that the most capable and differentiated models are able to charge a significant premium over even slightly less capable models. Gross margins on frontier models are north of 70%, but on trailing models with open source competition, margins are below 20%. Reasoning models don't have to be 1 chain of thought. Search exists and can be scaled up to improve performance as it has in O1 Pro and O3. This enables smarter models that can solve more problems and generate significantly more revenue per GPU.

Nvidia's not the only one that can increase memory capacity of course. ASICs can do this and in fact AMD may be well positioned due to their higher memory capacity versus Nvidia generally with MI300X's 192GB, MI325X 256GB, and MI350X 288GB... Well except Santa Huang has a Red-Nosed Reindeer called NVLink.

Apr 10, 2024

Nvidia Blackwell Perf TCO Analysis – B100 vs B200 vs GB200 NVL72

Dylan Patel, Daniel Nishball

When we step forward to GB200 NVL72 and GB300 NVL72, the performance and cost for Nvidia based systems improve massively. The key point for using NVL72 in inference is because it enables 72 GPUs to work on the same problem, sharing their memory, at extremely low latency. No other accelerator in the world has all-to-all switched connectivity. No other accelerator in the world can do all reduce through a switch.

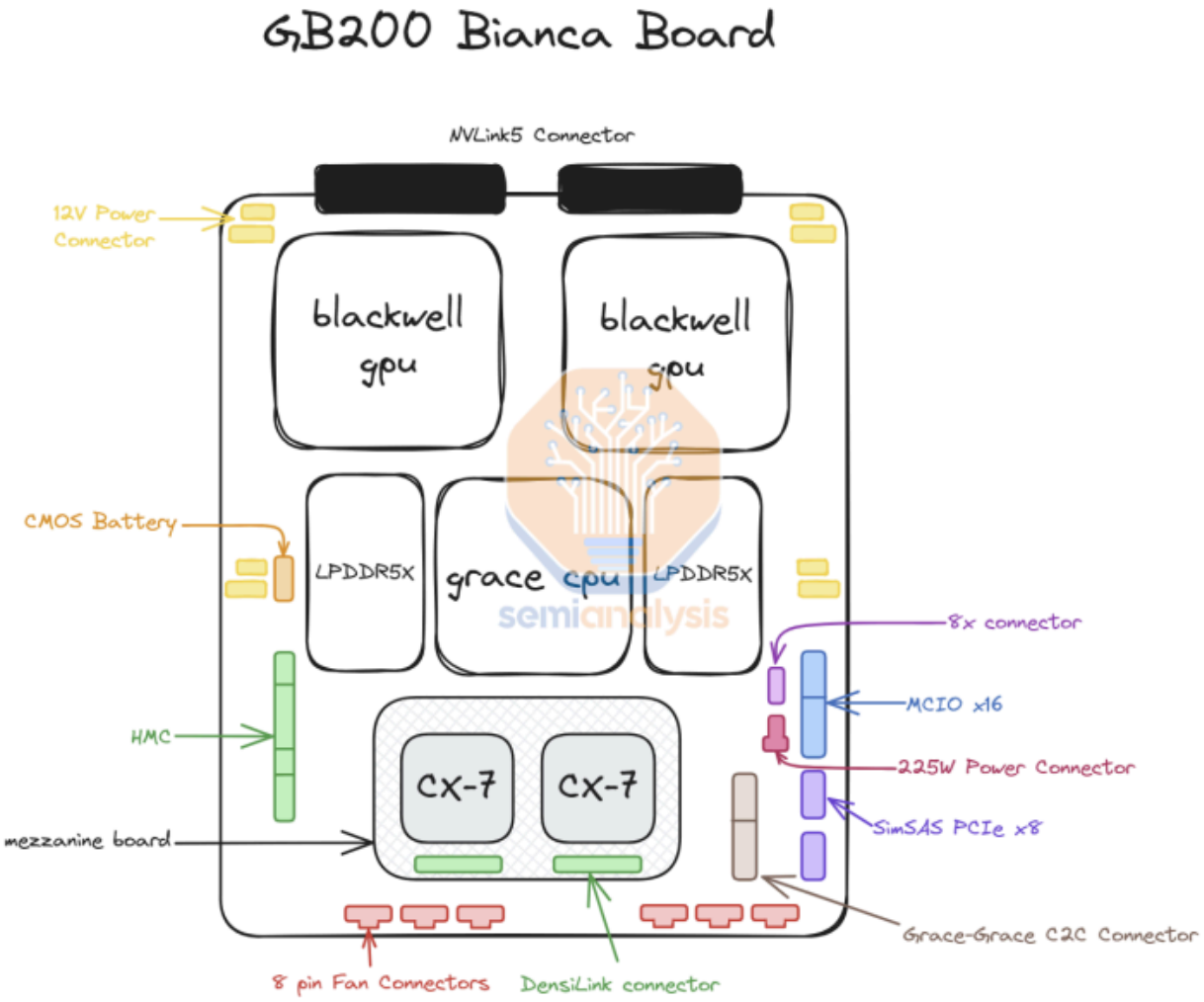
Nvidia's GB200 NVL72 and GB300 NVL72 is incredibly important to enabling a number of key capabilities.

1. Much higher interactivity enabling lower latency per chain of thought.
2. 72 GPUs to spread KVCache over to enable much longer chains of thought (increased intelligence).
3. Much better batch size scaling versus the typical 8 GPU servers, enabling much lower cost.
4. Many more samples to search with working on the same problem to improve accuracy and ultimately model performance.

As such, the tokenomics with NVL72 are more than 10x better, especially on long reasoning chains. KVCache eating up memory is a killer for economics, but NVL72 is the only way to scale reasoning lengths to 100k+ tokens at high batches.

Blackwell Supply Chain Reworked for GB300

With GB300, the supply chain and content that Nvidia supplies drastically changes. For the GB200 Nvidia provides the whole Bianca board (including the Blackwell GPU, Grace CPU, 512GB of LPDDR5X, VRM content all integrated onto one PCB) as well as the switch tray and copper backplane.

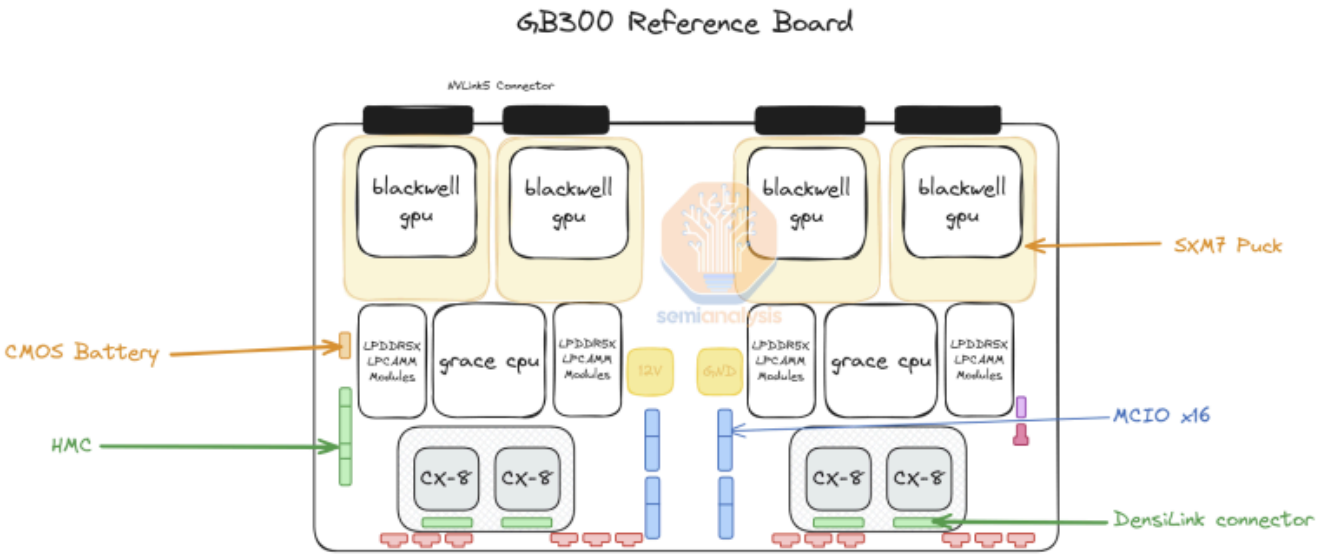


Source: SemiAnalysis

For GB300, instead of supplying the whole Bianca board, Nvidia will only supply the B300 on an "SXM Puck" module, the Grace CPU on a BGA package, and the HMC which will be from US based startup Axiado instead of Aspeed for GB200.

End customers will now directly procure the remaining components on the compute board and the second tier of memory will be LPCAMM modules instead of soldered-on LPDDR5X. Micron will be the main supplier of these modules.

The switch tray and copper backplane stays the same with Nvidia supplying these components entirely.



Source: SemiAnalysis

The shift to the SXM Puck opens up opportunities for more OEMs and ODMs to participate in the compute tray. Where previously only Wistron and FII could manufacture the Bianca compute board, now more OEMs and ODMs can. Wistron is the biggest loser in the form of ODMs as they lose share of the Bianca board. For FII, share loss at the Bianca board level is offset by the fact that they are the exclusive manufacturer of the SXM Puck and socket that the SXM Puck sits on. Nvidia is attempting to bring other suppliers for both the Puck and socket, but they have not placed any other orders yet.

Another major shift is with the VRM content. While there is some VRM content on the SXM Puck, much of the on board VRM content will be procured by Hyperscalers/OEMs directly from VRM suppliers. On October 25th for Core Research subscribers, we sent a note on how B300 was reshaping the supply chain specifically around Voltage Regulator Modules ("VRM"). We specifically called out how Monolithic Power Systems would lose market share due to the shift in business model and which new entrants were gaining market share. In the month following our note to clients, the MPWR fell over 37% due to the market's realization of the facts in our leading research.

Nvidia also offers the 800G ConnectX-8 NIC on the GB300 platform, offering twice the scale out bandwidth on InfiniBand and Ethernet. Nvidia cancelled

ConnectX-8 for GB200 a while ago due to time to market complexities and foregoing enabling PCIe Gen 6 on the Bianca board.

ConnectX-8 offers a huge improvement versus ConnectX-7. Not only does it have 2x bandwidth, but it also has 48 PCIe lanes instead of 32 PCIe lanes, enabling unique architectures such as the air cooled MGX B300A. Furthermore, ConnectX-8 is SpectrumX capable whereas on the prior 400G generation, SpectrumX required much less efficient Bluefield 3 DPUs.

Hyperscaler Impacts with GB300

The hyperscaler impacts from delayed GB200 and GB300 mean that many orders starting in Q3 shift over to Nvidia's new more expensive GPU. As of last week, all hyperscalers have decided to go forward with GB300. Partially this is due to the increased performance of GB300 due to higher FLOPS and more Memory, but also a portion of this is due to the ability to have control of their destiny.

Due to time to market challenges and significant changes in rack, cooling, and power delivery/density, hyperscalers were not allowed to change the GB200 much at the server level. This resulted in Meta abandoning all hope of being able to multi-source NICs from Broadcom and Nvidia in favor of relying solely on Nvidia. In other cases, such as Google, they abandoned their in-house NIC in favor of only going with Nvidia.

This is like nails on a chalkboard for the multi-thousand people organizations at hyperscalers who are used to cost optimizing everything from CPUs to networking down to screws and sheet metal.

The most egregious example was Amazon, who choose a very sub-optimal configurations that had worse TCO versus the reference design. Amazon specifically has not been able to deploy NVL72 racks like Meta, Google, Microsoft, Oracle, X.AI, and Coreweave due to the use of PCIe switches and

less efficient 200G Elastic Fabric Adaptor NICs needing to be air cooled. Amazon due to their internal NICs had to use NVL36 which also costs more per GPU due to higher backplane and switch content. All in all, Amazon's configuration was sub-optimal, due to their constraints around customization.

Now with GB300, hyperscalers are able to customize the main board, cooling, and much more. This enables Amazon to build their own custom mainboard which is watercooled with previously air-cooled components integrated such as the Astera Labs PCIe Switches. Watercooling more components alongside finally getting to HVM on the K2V6 400G NIC in Q3 25 means that Amazon can move back to NVL72 architecture and greatly improve their TCO.

There is one big downside though, which is that hyperscalers have to design, verify, and validate a ton more. This is easily the most complicated platform hyperscalers have ever had to design (save for Google's TPU systems). Certain hyperscalers will be able to design this quickly, but others with slower teams are behind. Generally, despite market cancellation reports, we see Microsoft as one of the slowest to deploy GB300 due to design speed, with them still buying some GB200 in Q4.

The total price the customer pays differs a lot as components get pulled out of Nvidia's margin stacking, onto ODMs. ODM's revenue is impacted, and most importantly, Nvidia's gross margin shifts through the year as well. Below we will show these impacts.

Nvidia Margin Impact

The additional memory capacity comes with upgrading from SK Hynix and Micron 8-Hi HBM3E stacks to 12-Hi HBM3E stacks. This upgrade drives the BOM cost increase for Nvidia at the chip level by around $2,500 higher (both by more capacity, an additional premium on $/GB due to higher stack count as well as the cost associated from packaging yield loss). This illustrates the trend of HBM being the single biggest gainer of share of BOM which will only

continue as reasoning model architectures demand more memory capacity and bandwidth.

Overall, the Nvidia ASP of the GB300 is ~$4,000 higher than the GB200 equivalent. As mentioned, the big change in BOM is a ~$2.5k increase for HBM. This would mean an incremental margin stack of less than 40% on face compared to mid to low 70% margin for the GB200 as a whole.

However, this is offset for the GB300 due to the content change as discussed above with Nvidia offering less content as a whole and making hyperscalers procure themselves

Nvidia is no longer providing the 512GB of LPDDR5X per Grace CPU which helps them claw back a lot of the additional HBM cost. Then the PCB savings are the next most significant. Net of all this the BOM cost for Nvidia goes up just over $1,000 on an ASP increase of $4,000. The incremental gross margin for GB300 vs GB200 is 73% meaning the product is margin neutral assuming constant yields.

While this may seem anti-climactic, note that HBM refresh cycles have generally been margin dilutive (such as the case of the H200, MI325X) and this notable as a trend break. Also, as various engineering issues are getting resolved, so yields will improves and margins will actually improve through the year after the gnarly ramp of Blackwell brings them down. Santa Huang even has a present for you Wall Street!

Component	Delta: GB300 vs GB200	Delta (per GPU)
Total compute tray content	**$78,161**	
Switch trays	**$0**	
Copper backplane	**$0**	
Total BOM cost to Nvidia	$78,161	$1,086
Nvidia Markup	$209,839	$2,914
ASP to end customer (NVL rack cost)	**$288,000**	**$4,000**
Other Nvidia add on items	**$32,400**	**$450**
Non-Nvidia BOM	**$106,049**	**$1,473**
ODM Assembly fee	**$18,001**	**$250**
Total GB200/300 NVL72 system cost	**$444,450**	**$6,173**
Chip level BOM		
B200/B300 chip level:	[redacted]	
Active Silicon Cost		
Memory Costs		
Packaging Costs		
Assembly & Test Costs		
Yield loss cost adders		
Total B200 chip level BOM cost	**$2,494**	**$2,494**
Grace CPU chip level cost	**$0**	**$0**
Compute tray content from Nvidia		
B200 GPUs	[redacted]	
Grace CPU		
VRM		
LPDDR5X (512GB)		
PCB		
BMC/HMC		

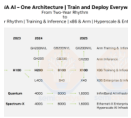
Source: SemiAnalysis

Next Article

2025 AI Diffusion Export Controls – Microsoft Regulatory Capture, Oracle Tears, Impacts Quantified, Model Restrictions // Malaysia's Stranded Capacity, Western Hyperscale Strength, AI Accelerator Restrictions, India Brazil Middle East Concerns, Sovereign AI Killed, VEU Restrictions and Implications, Compute Budget, Open-Source Creating Lower Bound, Model-Weights Controlled

Keep Reading

Nvidia Hacked – A National Security Disaster



Nvidia's Plans To Crush Competition – B100, "X100", H200, 224G SerDes, OCS, CPO, PCIe 7.0, HBM3E // Roadmap, Supply, Anti-competitive: AMD, Broadcom, Google, Amazon, and Microsoft Have Their Work Cutout For Them

Nvidia's Ramp – Volume, ASP, Cloud Pricing, Margins, EPS, Cashflow, China, Competition // Unraveling the Mixed Signals in the Supply Chain from AI Head-Fakes

Comments

fredhstein@gmail.com
December 25, 2024

Great article. The mention of Micron, raises the question of whether use of their HBM is major catalyst for MU. And, follow up, has the market ignored this

catalyst in the recent decline in MU.

Reply

Eric Liu
December 27, 2024

It would seem the GB300 degree of improvement for reasoning inference scaling would shift the balance away from ASIC's from the hyperscalers and OPEN AI towards GPU's. What do you expect that ratio to be going forward?

Reply

Dylan Patel Author
December 30, 2024

See accelerator model
https://semianalysis.com/accelerator-industry-model/

Reply

dantrump@gmail.com
December 30, 2024

50% more flops with minor architectural update? B300 sounds like they are using 3x ~800mm2 dies and sticking with their standard HBM3e 8hi. Either that or you are referencing "effective" flops? and they are getting a big bump in some workloads going to 12hi?

I just don't think you can get 50% more flops on the same process node when you are already maxing out the reticle size and only have a "minor" update, and power shifting that is giving on you 200 more watts.

very cool they are able to piece together so much silicon.

Reply

Dylan Patel Author
December 30, 2024

No, still 2 reticles

Reply

~~john benson~~
January 2, 2025

Does higher customization mean server manufacturing share could shift back from the likes of Quanta / Foxconn on the GB200 NVL systems (which they effectively dominate) to the likes of Super Micro and others? Or will these also be mass produced as L10 NVL systems like the GB 200s?

Reply

Kyle Kochanski
January 5, 2025

Micron’s LPDDR5X seems to be in high demand these days. Samsung selected them for the S25 phone, now also for these modules in GB300 separate from the 12-hi HBM3E. Seems another tailwind for them if on-device memory keeps increasing?

Reply