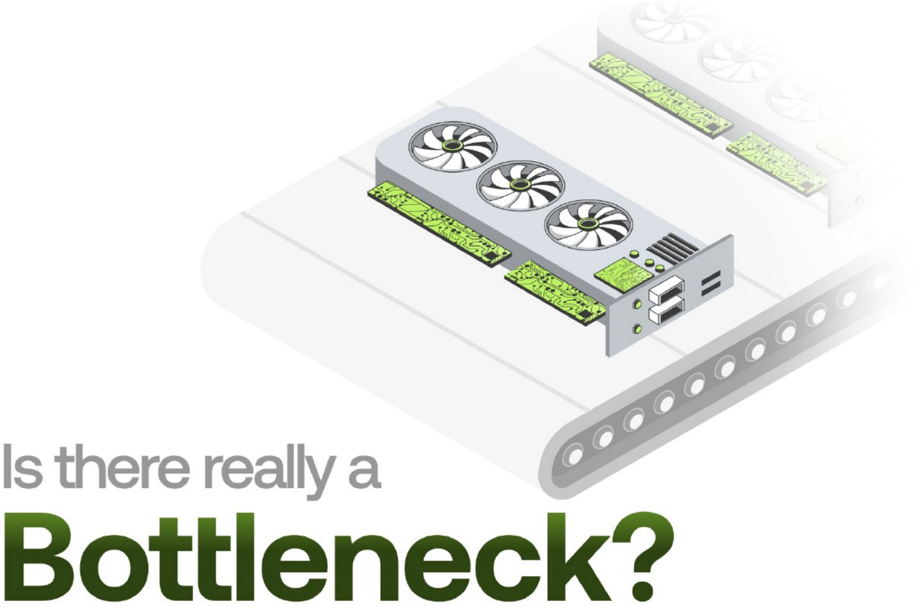
Nvidia H100 GPUs: Supply and Demand

July 2023 · Updated: September 2023



▼ Table of Contents

This post is an exploration of the supply and demand of GPUs, particularly Nvidia H100s. We're also releasing a song and music video on the same day as this post.

This post went mega viral. It was on the frontpage of HN, techmeme, many email newsletters, got tweets from Andrej Karpathy and others, comments from Mustafa (who will have $1B of GPUs online soon) from Inflection and Emad from Stability, the song was mentioned in the NY Times, and various asset managers and AI founders reached out. If you haven't read it yet, I hope you enjoy!

Introduction

As of August 2023, it seems AI might be bottlenecked by the supply of GPUs.

“One reason the AI boom is being underestimated is the GPU/TPU shortage. This shortage is causing all kinds of limits on product rollouts and model training but these are not visible. Instead all we see is Nvidia spiking in price. Things will accelerate once supply meets demand.”

— Adam D'Angelo, CEO of Quora, Poe.com, former Facebook CTO

These Are The CEOs And Companies That Are Most Important to GPU Supply and Demand - And To AI. Larger version

Is There Really A Bottleneck?

Elon Musk says that “GPUs are at this point considerably harder to get than drugs.”[1]

Sam Altman says that OpenAI is GPU-limited and it's delaying their short term plans (fine-tuning, dedicated capacity, 32k context windows, multimodality).[2]

Capacity of large scale H100 clusters at small and large cloud providers is running out.[3]

“Rn everybody wishes Nvidia could produce more A/H100”[4]

— Message from an exec at a cloud provider

“We’re so short on GPUs the less people use our products the better”

“We’d love it if they use it less because we don’t have enough GPUs”

Sam Altman, CEO at OpenAI[5]

It’s a good soundbite to remind the world how much users love your product, but it’s also true that OpenAI needs more GPUs.

For Azure/Microsoft:

They are rate limiting employees on GPUs internally. They have to queue up like it was a university mainframe in the 1970s. I think OpenAI is sucking up all of it right now.

The Coreweave deal is all about pasting on their GPU infrastructure.

— Anonymous

In short: Yes, there’s a supply shortage of H100 GPUs. I’m told that for companies seeking 100s or 1000s of H100s, Azure and GCP are effectively out of capacity, and AWS is close to being out.[6]

This “out of capacity” is based on the allocations that Nvidia gave them.

What do we want to know about the bottleneck?

What’s causing it (how much demand, how much supply)

How long will it last

What’s going to help resolve it

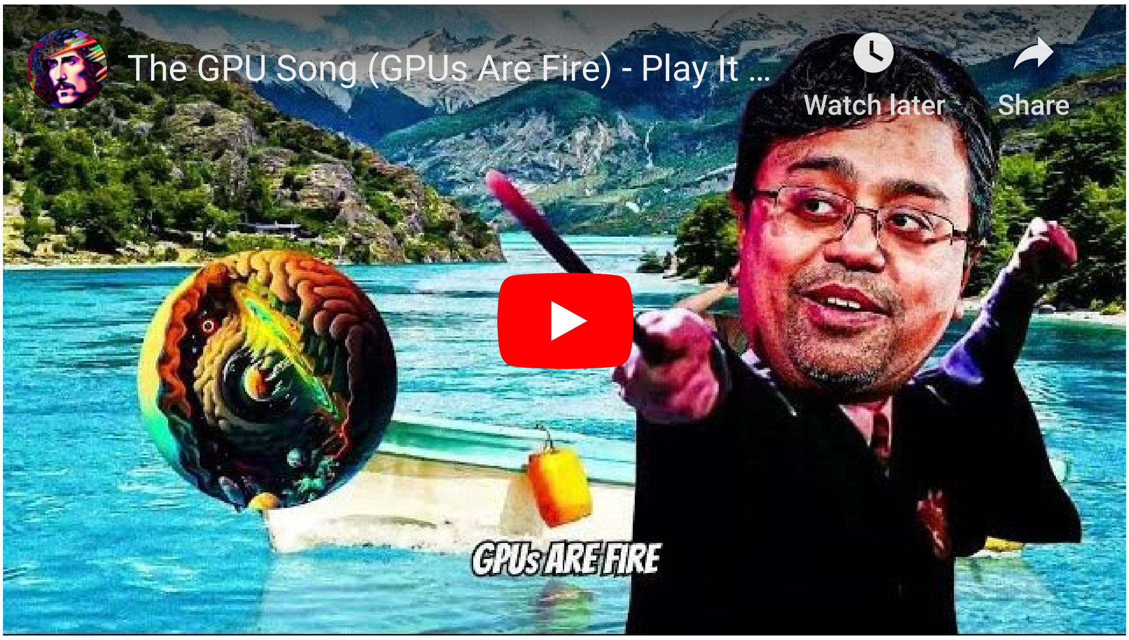
Table Of Contents

The GPU Song

Uh… We're also releasing a song on the same day as we're releasing this post.

It's fire.

If you haven't heard The GPU Song yet, do yourself a favor and play it.



i just watched the video. very funny. nice work.

– Mustafa Suleyman, CEO at Inflection AI

It's on Spotify, Apple Music and YouTube.

See more info on the song here.

Demand For H100 GPUs

What's causing the bottleneck - Demand

Specifically, what do people want to buy that they can't?

How many of those GPUs do they need?

Why can't they use a different GPU?

What are the different product names?

Where do companies buy them and how much do they cost?

Who Needs H100s?

"It seems like everyone and their dog is buying GPUs at this point"[7]

– Elon

Who Needs/Has 1,000+ H100 Or A100s

Startups training LLMs

OpenAI (through Azure), Anthropic, Inflection (through Azure[8] and CoreWeave[9]), Mistral AI

CSPs (Cloud Service Providers)

The big 3: Azure, GCP, AWS

The other public cloud: Oracle

Larger private clouds like CoreWeave, Lambda

Other large companies

Tesla[7] [10]

Who Needs/Has 100+ H100 Or A100s

Startups doing significant fine-tuning large open source models.

What Are Most Of The High End GPUs Being Used For?

For companies using private clouds (CoreWeave, Lambda), of companies with hundreds or thousands of H100s, it's almost all LLMs, and some diffusion model work. Some of it is fine-tuning of existing models, but mostly it's new startups that you may not yet know about that are building new models from scratch. They're doing $10mm-50mm contracts done over 3 years, with a few hundred to a few thousand GPUs.

For companies using on-demand H100s with a handful of GPUs, it's still probably >50% LLM related usage.

Private clouds are now starting to see inbound demand from enterprises who would normally be going with their default big cloud provider, but everyone is out.

Are The Big AI Labs More Constrained On Inference Or Training?

Depends on how much product traction they have! Sam Altman says OpenAI would rather have more inference capacity if forced to choose, but OpenAI is still constrained on both.[11]

Which GPUs Do People Need?

Mostly H100s. Why? It's the fastest both for inference and training for LLMs. (The H100 is often also the best price-performance ratio for inference, too)

Specifically: 8-GPU HGX H100 SXM servers.

My analysis is it's cheaper to run for the same work as well. The V100 a great deal if you could find them used, which you can't

– Anonymous

honestly not sure about [it being the best price-performance ratio]? price/performance for training looks about the same for A100 as for H100. for inference, we find that A10Gs are more than enough and much cheaper.

– Private cloud exec

this [A10G's being more than enough] was true for a while. but in the world of falcon 40b and llama2 70b, which we're seeing a lot of usage for, it's not true anymore. we need A100s for these

2xA100s to be exact. so the interconnect speed matters for inference.

– (Different) Private cloud exec

What's The Most Common Need From LLM Startups?

For training LLMs: H100s with 3.2Tb/s InfiniBand.

What Do Companies Want For LLM Training And Inference?

For training they tend to want H100s, for inference it's much more about performance per dollar.

It's still a performance per dollar question with H100s vs A100s, but H100s are generally favored as they can scale better with higher numbers of GPUs and give faster training times, and speed / compressing time to launch or train or improve models is critical for startups.

"For multi-node training, all of them are asking for A100 or H100 with InfiniBand networking. Only non A/H100 request we see are for inference where workloads are single GPU or single node"

– Private cloud exec

What Is Important For LLM Training?

Memory bandwidth

FLOPS (tensor cores or equivalent matrix multiplication units)

Caches and cache latencies

Additional features like FP8 compute

Compute performance (related to number of cuda cores)

Interconnect speed (eg InfiniBand)

The H100 is preferred over A100 partly because of things like lower cache latencies and FP8 compute.

H100 is preferred because it is up to 3x more efficient, but the costs are only (1.5 - 2x). Combined with the overall system cost, H100 yields much more performance per dollar (if you look at system performance, probably 4-5x more performance per dollar).

— Deep learning researcher

What Are The Other Costs Of Training And Running LLMs?

GPUs are the most expensive individual component, but there are other costs.

System RAM and NVMe SSDs are expensive.

InfiniBand networking is costly.

10-15% of total cost for running a cluster might go to power and hosting (electricity, cost of the datacenter building, cost of the land, staff) - roughly split between the two, can be 5-8% for power and 5-10% for other elements of hosting cost (land, building, staff).

It’s mostly networking and reliable datacenters. AWS is difficult to work with because of network limitations and unreliable hardware

— Deep learning researcher

What About GPUDirect?

GPUDirect is not a critical requirement, but can be helpful.

I would not say it is supercritical, but it makes a difference in performance. I guess it depends on where your bottleneck is. For some architectures / software implementations, the bottleneck is not necessarily networking, but if it is GPUDirect can make a difference of 10-20%, and that are some pretty significant numbers for expensive training runs.

That being said, GPUDirect RDMA is now so ubiquitous that it goes almost without saying that it is supported. I think support is less strong for non-InfiniBand networking, but most GPU clusters optimized for neural network training have Infiniband networks / cards. A bigger factor for performance might be NVLink, since this is rarer than Infiniband, but it is also only critical if you have particular parallelization strategies.

So features like strong networking and GPUDirect allows you to be lazy and you can guarantee that naive software is better out of the box. But it is not a strict requirement if you care about cost or using infrastructure that you already have.

– Deep learning researcher

What Stops LLM Companies From Using AMD GPUs?

Theoretically a company can buy a bunch of AMD GPUs, but it just takes time to get everything to work. That dev time (even if just 2 months) might mean being later to market than a competitor. So CUDA is NVIDIA's moat right now.

– Private cloud exec

I suspect 2 months is off by an order of magnitude, it's probably not a meaningful difference, see https://www.mosaicml.com/blog/amd-mi250

– ML Engineer

Who is going to take the risk of deplying 10,000 AMD GPUs or 10,000 random startup silicon chips? That's almost a $300 million investment.

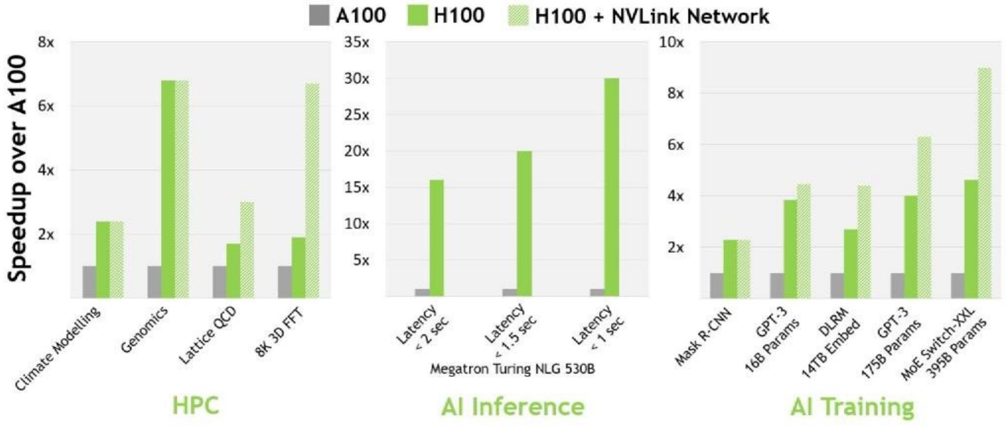
– Private cloud exec

MosaicML/MI250 - Has anyone asked AMD about availability? It doesn't seem like AMD built many beyond what they needed for Frontier, and now TSMC CoWoS capacity is sucked up by Nvidia. MI250 may be a viable alternative but unavailable.

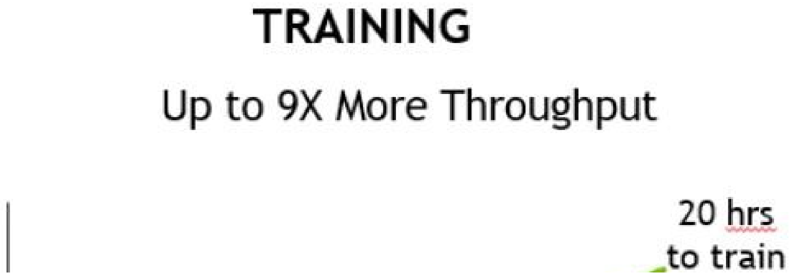
– Retired semiconductor industry professional

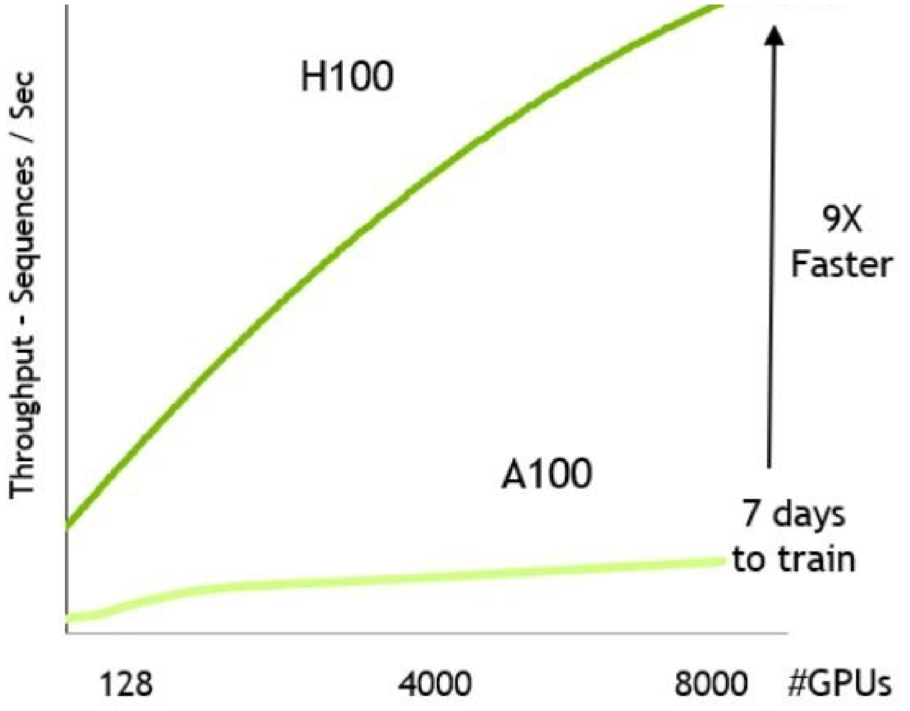
H100 Vs A100: How Much Faster Are H100s Than A100s?

About 3.5x faster for 16-bit inference[12] and about 2.3x faster for 16-bit training.[13]



A100 vs H100 Speed

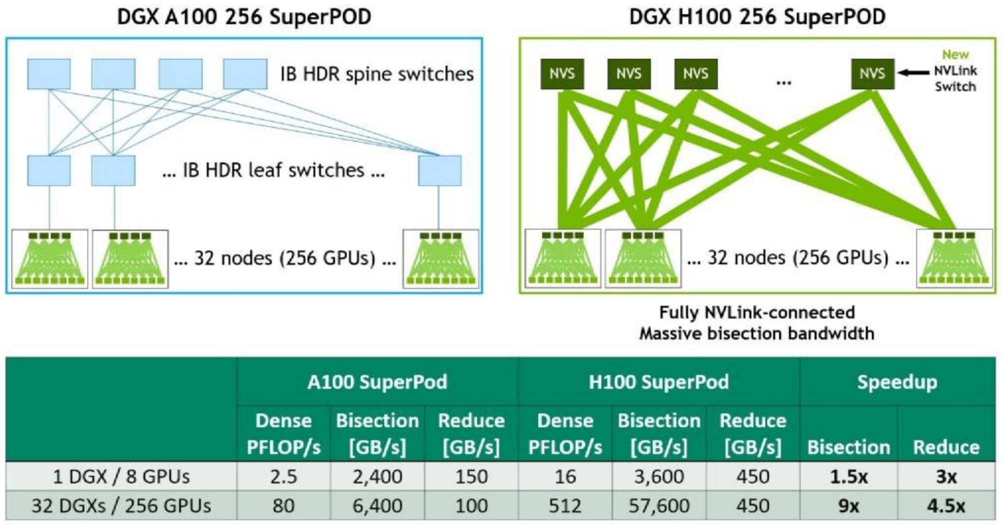




Mixture of Experts (395B) Training vs A100

Projected performance subject to change
Training Mixture of Experts (MoE) Transformer Switch-XXL variant with 395B parameters on 1T token dataset

H100 Training MoE



	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Bisection	Reduce
1 DGX / 8 GPUs	2.5	2,400	150	16	3,600	450	**1.5x**	**3x**
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	**9x**	**4.5x**

H100 Speedup At Scale

Here’s some more reading for you: 1 2 3.

Is Everyone Going To Want To Upgrade From A100s To H100s?

Mostly people will want to buy H100s and use them for training and inference and switch their A100s to be used primarily for inference. But, some people might be hesitant to switch due to cost, capacity, the risk of using new hardware and setting it up, and their existing software being already optimized for A100s.

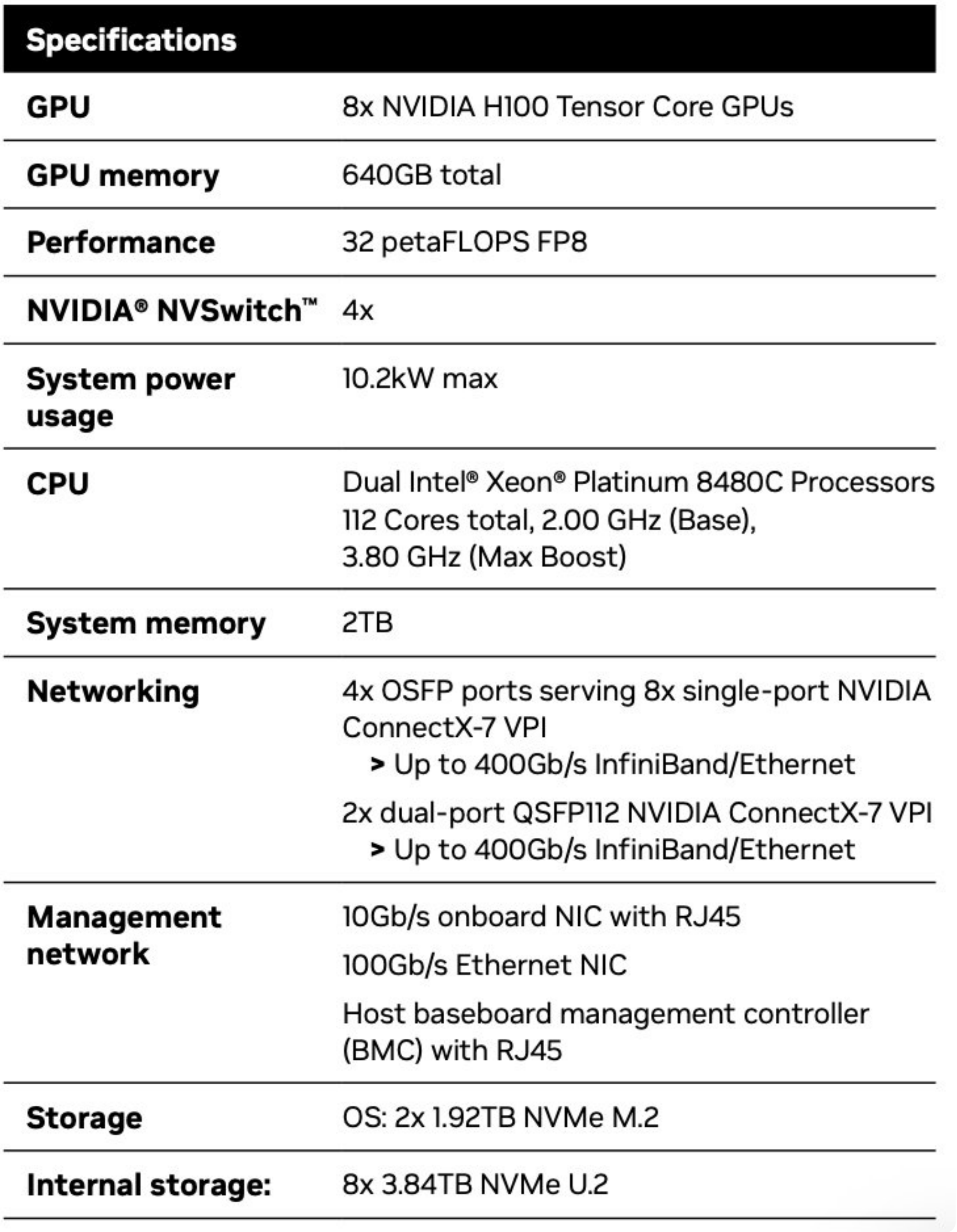
Yes, A100s will become today’s V100s in a few years. I don’t know of anyone

training LLMs on V100s right now because of performance constraints. But they are still used in inference and other workloads. Similarly, A100 pricing will come down as more AI companies shift workloads to H100s, but there will always be demand, especially for inference.
– Private cloud exec
think it's also plausible some of the startups that raised huge rounds end up folding and then there's a lot of A100s coming back on the market
– (Different) Private cloud exec
Over time people will move and the A100s will be more used for inference.
What about V100s? Higher VRAM cards are better for large models, so cutting edge groups much prefer H100s or A100s.
The main reason for not using V100 is the lack of brainfloat16 (bfloat16, BF16) data type. Without that, its very difficult to train models easily. The poor performance of OPT and BLOOM can be mostly attributed to not having this data type (OPT was trained in float16, BLOOM's prototyping was mostly done in fp16, which did not yield data was generalized to the training run which was done in bf16)
— Deep learning researcher
What's The Difference Between H100s, GH200s, DGX GH200s, HGX H100s, And DGX H100s?
H100 = 1x H100 GPU
HGX H100 = the Nvidia server reference platform that OEMs use to build 4-GPU or 8-GPU servers. Built by third-party OEMs like Supermicro.
DGX H100 = the Nvidia official H100 server with 8x H100s.[14] Nvidia is the sole vendor.
GH200 = 1x H100 GPU plus 1x Grace CPU.[15]
DGX GH200 = 256x GH200s,[16] available toward the end of 2023.[17] Likely only offered by Nvidia.
There's also MGX which is aimed at large cloud companies.

Which Of Those Will Be Most Popular?

Most companies will buy 8-GPU HGX H100s,[18] rather than DGX H100s or 4-GPU HGX H100 servers.
How Much Do These GPUs Cost?
1x DGX H100 (SXM) with 8x H100 GPUs is $460k including the required support. $100k of the $460k is required support. The specs are below. Startups

can get the Inception discount which is about $50k off, and can be used on up to 8x DGX H100 boxes for a total of 64 H100s.



Specifications	
GPU	8x NVIDIA H100 Tensor Core GPUs
GPU memory	640GB total
Performance	32 petaFLOPS FP8
NVIDIA® NVSwitch™	4x
System power usage	10.2kW max
CPU	Dual Intel® Xeon® Platinum 8480C Processors 112 Cores total, 2.00 GHz (Base), 3.80 GHz (Max Boost)
System memory	2TB
Networking	4x OSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA ConnectX-7 VPI > Up to 400Gb/s InfiniBand/Ethernet
Management network	10Gb/s onboard NIC with RJ45 100Gb/s Ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	OS: 2x 1.92TB NVMe M.2
Internal storage:	8x 3.84TB NVMe U.2

DGX H100 Specs

1x HGX H100 (SXM) with 8x H100 GPUs is between $300k-380k, depending on the specs (networking, storage, ram, CPUs) and the margins of whoever is selling it and the level of support. The higher end of that range, $360k-380k including support, is what you might expect for identical specs to a DGX H100.

1x HGX H100 (PCIe) with 8x H100 GPUs is approx $300k including support, depending on specs.

PCIe cards are around $30k-32k market prices.

SXM cards aren’t really sold as single cards, so it’s tough to give pricing there. Generally only sold as 4-GPU and 8-GPU servers.

Around 70-80% of the demand is for SXM H100s, the rest is for PCIe H100s.

And the SXM portion of the demand is trending upwards, because PCIe cards were the only ones available for the first few months. Given most companies buy 8-GPU HGX H100s (SXM), the approximate spend is $360k-380k per 8 H100s, including other server components.

The DGX GH200 (which as a reminder, contains 256x GH200s, and each GH200 contains 1x H100 GPU and 1x Grace CPU) might cost in the range of $15mm-25mm - though this is a guess, not based on a pricing sheet.[19]

How Many GPUs Are Needed?

GPT-4 was likely trained on somewhere between 10,000 to 25,000 A100s.[20]

Meta has about 21,000 A100s, Tesla has about 7,000 A100s, and Stability AI has about 5,000 A100s.[21]

Falcon-40B was trained on 384 A100s.[22]

Inflection used 3,500 H100s for their GPT-3.5 equivalent model.[23]

We have 22k operational by December btw. and way more than 3.5k operational today.

– Mustafa Suleyman, CEO at Inflection AI

GPT-5 might need 30k-50k H100s according to Elon. Morgan Stanley said in Feb 2023 that GPT-5 would use 25,000 GPUs, but they also said it was already being trained as of Feb 2023 and Sam Altman said in May 2023 that it's not yet being trained, so MS's info may be outdated.

GCP has approx 25k H100s. Azure probably has 10k-40k H100s. Should be similar for Oracle. Most of Azure's capacity is going to OpenAI.

CoreWeave is in the ballpark of 35k-40k H100s - not live, but based on bookings.

How Many H100s Are Most Startups Ordering?

For LLMs: For fine tuning, dozens or low hundreds. For training, thousands.

How Many H100s Might Companies Be Wanting?

OpenAI might want 50k. Inflection wants 22k.[24] Meta maybe 25k (I'm told actually Meta wants 100k or more). Big clouds might want 30k each (Azure, Google Cloud, AWS, plus Oracle). Lambda and CoreWeave and the other private clouds might want 100k total. Anthropic, Helsing, Mistral, Character, might want 10k each. Total ballparks and guessing, and some of that is double counting both the cloud and the end customer who will rent from the cloud. But that gets to about 432k H100s. At approx $35k a piece, that's about $15b worth of GPUs. That also excludes Chinese companies like ByteDance (TikTok), Baidu, and Tencent who will want a lot of H800s.

There are also financial companies each doing deployments starting with hundreds of A100s or H100s and going to thousands of A/H100s: names like Jane Street, JP Morgan, Two Sigma, Citadel.

How does that compare to Nvidia's data center revenue?

Feb-April 2023 was $4.28b data center revenue.[25] May-July 2023 might be around $8b data center revenue, assuming most of the higher guidance for that quarter is due to gain in data center revenue rather than other segments.

So might take a while for the supply shortage to go away. But also all my ballparks could be wildly overstated, and many of these companies aren't going to go right out and buy the H100s today, they'll upgrade over time. Plus, Nvidia is aggressively ramping production capacity.

Seems possible. 400k H100s doesn't sound out of reach, especially given how everyone is doing a massive 4 or 5-figure H100 deployment right now.

– Private cloud exec

Summary: H100 Demand

The main things to keep in mind as you go onto the next section are that most of the big CSPs (Azure, AWS, GCP, and also Oracle) and private clouds (CoreWeave, Lambda, and various others) want more H100s than they can get access to. Most of the big AI product companies want more H100s than they can get access to, as well. Generally they want 8-GPU HGX H100 boxes with SXM cards, which cost approx $300k-400k per 8-GPU server depending on specs and support. There may be a few hundred thousand H100 GPUs worth of excess demand ($15b+ of GPUs). With a limited supply, Nvidia could purely raise prices to find a clearing price, and are doing that to some extent. But it's important to know that ultimately H100 allocation comes down to who Nvidia prefers to give that allocation to.

Supply Of H100 GPUs

What's causing the bottleneck - Supply

What are the bottlenecks on the production side?

Which components?

Who produces them?

Who Makes The H100s?

TSMC.

Can Nvidia Use Other Chip Fabs For H100 Production?

Not really, at least not yet. They've worked with Samsung in the past. But on the H100s and other 5nm GPUs they only use TSMC. Implication is that

Samsung can’t yet meet their needs for cutting edge GPUs. They might work with Intel in the future, and Samsung again on cutting edge, but neither of those will be happening in the short term in a way that’d help the H100 supply crunch.

How Do The Different TSMC Nodes Relate?

TSMC 5nm family:

N5[26]

4N either fits here as an enhanced version of N5, or below N5P

N5P

4N either fits here as an enhanced version of N5P, or below N5 as an enhanced version of N5

N4

N4P

Which TSMC Node Is The H100 Made On?

TSMC 4N. This is a special node for Nvidia, it’s in the 5nm family and is enhanced 5nm though rather than truly 4nm.

Who Else Uses That Node?

It was Apple, but they’ve moved primarily to N3 and have reserved most of the N3 capacity. Qualcomm and AMD are the other big N5-family customers.

Which TSMC Node Does The A100 Use?

N7[27]

How Long In Advance Is Fab Capacity Normally Reserved?

Not sure though maybe 12+ months.

that applies to TSM and their big customers They sort of plan it out together

Which is why TSM/NVDA may have underestimated what they need

– Anonymous

How Long Does Production Take (Production, Packaging, Testing)?

6 months from production on a H100 starting to that H100 being ready to be sold to a customer (est from a conversation, would like to get a confirmation)

Where Are The Bottlenecks?

Wafer starts are not the bottleneck at TSMC. Mentioned earlier CoWoS (3D stacking) packaging is the gate at TSMC.

– Retired semiconductor industry professional

H100 Memory

What Impacts Memory Bandwidth On GPUs?

Memory type, memory bus width, and memory clock speed.

It’s mostly HBM. Manufacturing it is a nightmare. Supply is also mostly limited because HBM is so difficult to produce. Once you have HBM the design follows intuitively

— Deep learning researcher

What Memory Is Used On The H100s?

On the H100 SXM, it’s HBM3.[28] On the H100 PCIe, it’s actually HBM2e.[29]

Who Makes The Memory On The H100s?

The bus width and clock speed are designed by Nvidia as part of the GPU architecture.

For the HBM3 memory itself, I think Nvidia uses either all or mostly SK Hynix. Not sure if Nvidia uses any from Samsung for the H100s and I believe it’s nothing from Micron for the H100s.

In terms of HBM3 generally, SK Hynix makes the most, then Samsung not that far behind, then Micron far behind. Seems like SK Hynix is ramped up but Nvidia still wants them to make more, and Samsung and Micron haven’t successfully ramped up production yet.

What Else Is Used When Making GPUs?

Note that some of these pieces are significantly more bottlenecked than others.

Metal Elements: These are essential in the production of GPUs. They include:

Copper: Used in the creation of electrical connections due to its high conductivity.

Tantalum: Often used in capacitors due to its ability to hold a high electrical charge.

Gold: Used in high-quality plating and connectors due to its resistance to corrosion.

Aluminum: Frequently used in the heatsink to help dissipate heat.

Nickel: Often used in the coating of connectors for its corrosion resistance.

Tin: Used in soldering components together.

Indium: Used in thermal interface materials for its good thermal conductivity.

Palladium: Used in certain types of capacitors and semiconductor devices.

Silicon (Metalloid): This is the primary material used in the creation of semiconductor devices.

Rare Earth Elements: These are used in various parts of the GPU for their unique properties.

Other Metals and Chemicals: These are used in various stages of production, from creating the silicon wafers to the final assembly of the GPU.

Substrates: These are the material on which the GPU components are mounted.

Package Materials: These are used to house and protect the GPU chip.

Solder Balls and Bonding Wires: These are used to connect the GPU chip to the substrate and other components.

Passive Components: These include capacitors and resistors, which are essential for the operation of the GPU.

Printed Circuit Board (PCB): This is the board on which all the components of the GPU are mounted. It provides the electrical connections between the components.

Thermal Compounds: These are used to improve heat conduction between the chip and the heatsink.

Semiconductor Manufacturing Equipment: This includes photolithography machines, etching equipment, ion implantation equipment, etc.

Clean Room Facilities: These are necessary for the production of GPUs to prevent contamination of the silicon wafers and other components.

Testing and Quality Control Equipment: These are used to ensure that the GPUs meet the required performance and reliability standards.

Software and Firmware: These are essential for controlling the operation of the GPU and for interfacing with the rest of the computer system.

Packaging and Shipping Materials: These are necessary for delivering the final product to customers in good condition.

Software Tools: Software tools for Computer-Aided Design (CAD) and simulations are crucial in designing the structure and testing functionality of the GPU.

Energy Consumption: A significant amount of electricity is required in the manufacturing process of GPU chips due to the usage of high-precision machinery.

Waste Management: The production of GPUs results in waste which has to be properly managed and disposed of, as many of the materials used can be harmful to the environment.

Test capacity: Custom/specialty test equipment that verifies functionality and performance.

Chip packaging: Assembling the silicon wafer into a component package that can be utilized in a larger system.

Outlook And Predictions

What Is Nvidia Saying?

Nvidia has disclosed that they have more supply in the second half of the year, but beyond that they haven't said much more, and nothing quantitative.

"We are working on both supply today for this quarter, but we have also procured a substantial amount of supply for the second half"

"We believe that the supply that we will have for the second half of the year will be substantially larger than h1"

– Nvidia CFO Colette Kress during the earnings call for Feb-April 2023

What'll Happen Next?

I think it's possible we have a self-reinforcing cycle right now where scarcity causes GPU capacity to be perceived as a moat, which causes more GPU-hoarding, which exacerbates scarcity.

– Private cloud exec

When Will There Be A H100 Successor?

Probably won't be announced until late 2024 (mid 2024 to early 2025), based on historical Nvidia time between architectures.

The H100 will be the top of the line Nvidia GPU until then. (The GH200 and DGX GH200 don't count, they're not pure GPUs, they all use H100s as their GPU)

Will There Be Higher VRAM H100s?

Maybe liquid cooled 120GB H100s.

When Will The Shortage End?

One group I talked with mentioned they are effectively sold out until the end of 2023.

Sourcing H100s

Who Sells H100s?

OEMs like Dell, HPE, Lenovo, Supermicro and Quanta sell H100s and HGX H100s.[30]

And when you need InfiniBand, you'll need to speak directly to Mellanox at Nvidia.[31]

So GPU clouds like CoreWeave and Lambda buy from OEMs and then rent to startups.

Hyperscalers (Azure, GCP, AWS, Oracle) work more directly with Nvidia but they are generally also working with the OEMs as well.

And even for DGX you'll still buy through an OEM. You can talk to Nvidia, but you'll buy through an OEM. You won't do a purchase order directly to Nvidia.

How Are The Lead Times?

Lead times on 8-GPU HGX servers are terrible, lead times on 4-GPU HGX servers are good. Everyone wants the 8-GPU servers!

If A Startup Places An Order Today, When Would They Have SSH Access?

It'd be a staggered deployment. Say it was a 5,000 GPU order. They might get access to 2,000 or 4,000 in 4-5 months and then the remaining by around 6 months total.

Do Startups Buy From OEMs And Resellers?

Not really. Startups will generally go to big clouds like Oracle to rent access, or to private clouds like Lambda and CoreWeave, or to providers that work with OEMs and data centers like FluidStack.

When Do Startups Build Their Own Datacenter Vs Doing Colocation?

For building a datacenter, the considerations are the time to build the datacenter, whether you have the people and experience in hardware, and that it's capex expensive.

Much easier to rent & colo servers. If you want to build your own DC, you literally have to run a dark fiber line out to your location to connect to the internet - $10k per km. Most of this infra was already built & paid for during dot-com boom. Now you can just rent it, quite cheap

– Private cloud exec

The spectrum from rent to own is: on-demand cloud (pure rental using cloud services), reserved cloud, colo (buy the servers, work with a provider to host and manage the servers), self-hosting (buy and host the servers yourself).

Most startups needing large H100 quantities will do either reserved cloud or colo.

How Do The Big Clouds Compare?

The sentiment is that Oracle infrastructure is less reliable than the big 3 clouds. In exchange, Oracle gives more tech support help and time.

100%. a big feeder of unhappy customers lol

– Private cloud exec

i think [oracle has] better networking though

– (Different) Private cloud exec

Generally startups will pick whoever offers the best blend of support, price, and capacity.

The main big differences at the large clouds are:

Networking (AWS and Google Cloud have been slower to adopt InfiniBand because they have their own approaches, though most startups looking for large A100/H100 clusters are seeking InfiniBand)

Availability (Azure's H100s are mostly going to OpenAI. GCP is struggling to get H100s.)

Nvidia seems to tend to give better allocations to clouds that aren't building competing machine learning chips. (This is all speculation, not hard facts.) All of the big 3 clouds are working on machine learning chips, but the Nvidia-alternative offerings from AWS and Google are already available and taking dollars that might've gone to Nvidia.

also speculation but i agree that nvidia likes oracle for this reason

– Private cloud exec

Some big clouds have better pricing than others. As one private cloud exec noted, "a100s are much more expensive on aws/azure than gcp for instance."

oracle told me they have "10s of thousands of H100s" coming online later this year. they boasted about their special relationship with nvidia.

but… when it came to pricing, they were way higher than anyone else. they didn't give me H100 pricing but for A100 80gb they quoted me close to $4/hour, which is nearly 2x more than gcp's quote for the same hw and same commit.

– Anonymous

The smaller clouds are better for pricing, except in some instances where the one of the big clouds does a weird deal in exchange for equity.

It might be something like: Oracle & Azure > GCP & AWS in terms of Nvidia relationship. But that's speculation.

Oracle was the first to launch A100s, and they worked with Nvidia to host an NVIDIA-based cluster. Nvidia is also a customer of Azure.

Which Big Cloud Has The Best Networking?

Azure, CoreWeave and Lambda all use InfiniBand. Oracle has good networking, it is 3200 Gbps, but it's ethernet rather than InfiniBand, which may be around 15-20% slower than IB for use cases like high-parameter count LLM training. AWS and GCP's networking isn't as good.

Which Big Clouds Do Enterprises Use?

In one private datapoint of about 15 enterprises, all 15 were either AWS, GCP or Azure, zero Oracle.

Most enterprises will stick with their existing cloud. Desperate startups will go wherever the supply is.

How About DGX Cloud, Who Is Nvidia Working With For That?

"NVIDIA is partnering with leading cloud service providers to host DGX Cloud infrastructure, starting with Oracle Cloud Infrastructure (OCI)" - you deal with Nvidia sales but you rent it through an existing cloud provider (first launching with Oracle, then Azure, then Google Cloud, not launching with AWS)[32] [33]

Jensen said on the last earnings call: "The ideal mix is something like 10% Nvidia DGX Cloud and 90% the CSPs clouds"

When Did The Big Clouds Launch Their H100 Previews?

CoreWeave was first.[34] Nvidia gave them an earlier allocation, presumably to help strengthen competition (and because Nvidia is an investor) amongst large clouds.

Azure on March 13 announced that H100s were available for preview.[35]

Oracle on March 21 announced that H100s were available in limited availability.[36]

Lambda Labs on March 21 announced that H100s would be added in early April.[37]

AWS on March 21 announced that H100s would be available for preview starting in a few weeks.[38]

Google Cloud on May 10 announced the start of a private preview for H100s.[39]

Which Companies Use Which Clouds?

OpenAI: Azure.

Inflection: Azure and CoreWeave.

Anthropic: AWS and Google Cloud.

Cohere: AWS and Google Cloud.

Hugging Face: AWS.

Stability AI: CoreWeave and AWS.

Character.ai: Google Cloud.

X.ai: Oracle.

Nvidia: Azure.[35]

How Can A Company Or Cloud Service Provider Get More GPUs?

The ultimate bottleneck is getting allocation from Nvidia.

How Do Nvidia Allocations Work?

They have an allocation they give per customer. But for example, Azure saying "hey we would like 10,000 H100s all to be used by Inflection" is different from

Azure saying “hey we would like 10,000 H100s for Azure’s cloud” - Nvidia cares about who the end customer is, and so clouds might be able to get an extra allocation for a specific end customer if Nvidia is excited about the end customer. Nvidia also wants to know who that end customer is, as much as possible. And they prefer customers with nice brand names or startups with strong pedigrees.

Yes, this seems to be the case. NVIDIA likes to guarantee GPU access to rising AI companies (many of which they have a close relationship with). See Inflection — an AI company they invested in — testing a huge H100 cluster on CoreWeave, which they also invested in

– Private cloud exec

If a cloud brings Nvidia an end customer and says they’re ready to purchase xxxx H100s, if Nvidia is excited about that end customer they’ll generally give an allocation, which effectively boosts the total capacity allocated by Nvidia to that cloud - because it won’t count against the original allocation that Nvidia gave to that cloud.

It’s a unique situation in that Nvidia is giving large allocations to private clouds: CoreWeave has more H100s than GCP.

Nvidia would prefer not to give large allocations to companies that are attempting to compete directly with them (AWS Inferentia and Tranium, Google TPUs, Azure Project Athena).

But ultimately, if you put the purchase order and money in front of Nvidia, committing to a bigger deal and more money up front and show that you have a low risk profile, then you’ll get a larger allocation than others get.

Closing Thoughts

For now, we are GPU-limited. Even if we are at the “end of the era where it’s going to be these giant models” as Sam Altman has said.

It’s both bubble-ish and not-bubble-ish depending on where you look. Some companies like OpenAI have products like ChatGPT with intense product-market-fit, and can’t get enough GPUs. Other companies are buying or reserving GPU capacity so they’ll have access in the future, or to train LLMs that are much less likely to have product-market-fit.

Nvidia is the green king of the castle right now.

Tracing The Journey Of GPU Supply And Demand

The LLM product with the strongest product-market fit is ChatGPT. Here’s the story of GPU demand with respect to ChatGPT:

Users love ChatGPT. It’s probably making $500mm++ annual recurring revenue.

ChatGPT runs on the GPT-4 and GPT-3.5 APIs.

The GPT-4 and GPT-3.5 APIs need GPUs to run. Lots of them. And OpenAI wants to release more features for ChatGPT and their APIs, but they can’t, because they don’t have access to enough GPUs.

They buy lots of Nvidia GPUs through Microsoft/Azure. Specifically the GPU they want most is the Nvidia H100 GPU.

To make H100 SXM GPUs, Nvidia uses TSMC for fabrication and uses TSMC’s CoWoS packaging tech and uses HBM3 primarily from SK Hynix.

OpenAI isn’t the only company that wants GPUs (but they are the company with the strongest product-market-fit that wants GPUs). Other companies are also wanting to train large AI models. Some of these use cases will make sense, but some are more hype driven and unlikely to get product-market-fit. This is pushing up demand. Also, some companies are concerned about not being able to access GPUs in the future so they’re placing their orders now even when they don’t need them yet. So there’s a bit of “expectations of supply shortages create even more supply shortages” going on.

The other major contributor to GPU demand is from companies that want to create new LLMs. Here’s the story of GPU demand with respect to companies wanting to build new LLMs:

A company executive or founder knows there’s big opportunities in the AI space. Maybe they’re an enterprise that wants to train an LLM on their own data and use it externally or sell access, or maybe they’re a startup that wants to build an LLM and sell access.

They know they need GPUs to train large models.

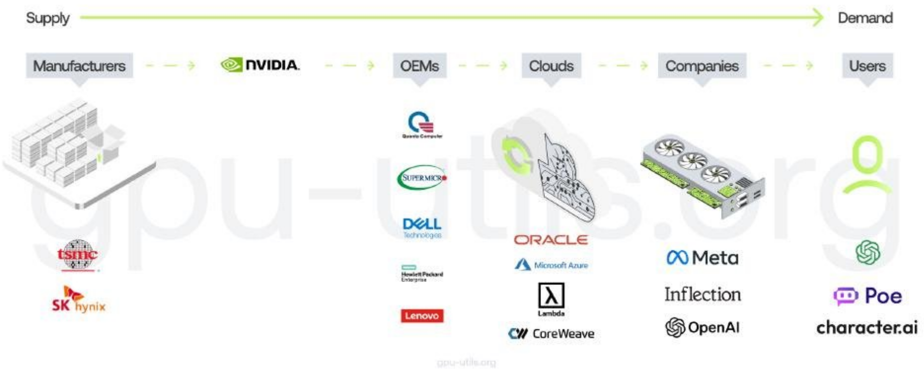
They talk with some set of people from the big clouds (Azure, Google Cloud, AWS) to try and get many H100s.

They find out that they can’t get a big allocation from the big clouds, and that some of the big clouds don’t have good networking setups. So they go and talk with other providers like CoreWeave, Oracle, Lambda, FluidStack. If they want to buy the GPUs themselves and own them, maybe they also talk with OEMs and Nvidia.

Eventually, they acquire a lot of GPUs.

Now, they try and get product-market-fit.

In case it’s not obvious, this pathway isn’t as good - remember that OpenAI got product-market-fit on much smaller models and then scaled them up. But, now to get product-market-fit you have to be better than OpenAI's models for your users' use-cases, so to start you will need more GPUs than OpenAI started with. Expect H100 shortages for multi-hundred or multi-thousand deployments through the end of 2023 at least. At the end of 2023 the picture will be clearer, but for now it looks like the shortages may persist through some of 2024 as well.



The Journey of GPU Supply and Demand. Larger version

Getting In Touch

Author: Clay Pascal. Questions and notes can be sent in via email.

New posts: get notified about new posts via email.

Helping: see here.

The Natural Next Question - What About Nvidia Alternatives?

The natural next question is "ok, what about the competition and alternatives?" I'm exploring hardware alternatives as well as software approaches. Submit things I should explore as alternatives to this form. For example, TPUs, Inferentia, LLM ASICs and others on the hardware side, and Mojo, Triton and others on the software side, and what it looks like to use AMD hardware and software. I'm exploring everything, though focusing on things that are usable today. If you're a freelancer and want to help get Llama 2 running on different hardware, email me. So far we've gotten it running on AMD, Gaudi, in progress for TPUs and Inferentia, and have people from AWS Silicon, Rain, Groq, Cerebras and other companies who've offered to help.

Acknowledgements

This article contains a decent amount of proprietary and previously unpublished information. When you see people wondering about GPU production capacity, please point them in the direction of this post.

Thanks to a handful of execs and founders at private GPU cloud companies, a few AI founders, an ML engineer, a deep learning researcher, a few other

industry experts, and some non-industry readers, for providing helpful comments. Thanks to Hamid for illustrations.

https://www.youtube.com/watch?v=nxbZVH9kLao&t=35s ↩

https://humanloop.com/blog/openai-plans ↩

Comment from a custom LLMs-for-enterprises startup founder ↩

Message from an exec at a cloud provider ↩

https://www.youtube.com/watch?v=TO0J2Yw7usM ↩

Conversations with execs at cloud companies and GPU providers ↩

https://www.tomshardware.com/news/more-details-about-elon-musk-ai-project-emerge ↩ ↩

https://azure.microsoft.com/en-us/blog/azure-previews-powerful-and-scalable-virtual-machine-series-to-accelerate-generative-ai/ ↩

https://inflection.ai/nvidia-coreweave-mlperf ↩

Tesla Q1 2023 (covers Jan 1 2023 to Mar 31 2023) earnings call ↩

https://llm-utils.org/OpenAI+Interviews/Sam+Altman+interviewed+by+Patrick+Collison+-+Transcript+(May+9%2C+2023) ↩

https://timdettmers.com/2023/01/30/which-gpu-for-deep-learning/ ↩

https://www.mosaicml.com/blog/mpt-30b ↩

https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx ↩

https://resources.nvidia.com/en-us-grace-cpu/grace-hopper-superchip ↩

https://resources.nvidia.com/en-us-dgx-gh200/nvidia-dgx-gh200-datasheet-web-us ↩

https://llm-utils.org/DGX+GH200+Stats+and+Release+Date ↩

A comment from an exec at a cloud company ↩

A guesstimate ballpark from an exec at a cloud company ↩

https://www.fierceelectronics.com/sensors/chatgpt-runs-10k-nvidia-training-gpus-potential-thousands-more ↩

https://www.stateof.ai/compute ↩

https://huggingface.co/tiiuae/falcon-40b ↩

https://inflection.ai/nvidia-coreweave-mlperf ↩

https://inflection.ai/inflection-ai-announces-1-3-billion-of-funding ↩

https://nvidianews.nvidia.com/news/nvidia-announces-financial-results-for-first-quarter-fiscal-2024 ↩

https://fuse.wikichip.org/news/6439/tsmc-extends-its-5nm-family-with-a-new-enhanced-performance-n4p-node/, https://pr.tsmc.com/english/news/2874 ↩

https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/ ↩

https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/ ↩

https://www.nvidia.com/content/dam/en-zz/Solutions/gtcs22/data-center/h100/PB-11133-001_v01.pdf ↩

https://www.nvidia.com/en-us/data-center/products/certified-systems/ ↩

https://llm-utils.org/Building+your+own+GPU+cluster ↩

https://nvidianews.nvidia.com/news/nvidia-launches-dgx-cloud-giving-every-enterprise-instant-access-to-ai-supercomputer-from-a-browser ↩

https://www.reuters.com/technology/amazons-cloud-unit-is-considering-amds-new-ai-chips-2023-06-14/ ↩

https://www.businesswire.com/news/home/20230321005245/en/CoreWeave-Announces-NovelAI-as-Among-the-First-to-Have-NVIDIA-HGX-H100-GPUs-Online ↩

https://azure.microsoft.com/en-us/blog/azure-previews-powerful-and-scalable-virtual-machine-series-to-accelerate-generative-ai/ ↩ ↩

https://blogs.oracle.com/cloud-infrastructure/post/limited-availability-oci-compute-nvidia-h100 ↩

https://lambdalabs.com/blog/lambda-cloud-adding-nvidia-h100-tensor-core-gpus-in-early-april ↩

https://nvidianews.nvidia.com/news/aws-and-nvidia-collaborate-on-next-generation-infrastructure-for-training-large-machine-learning-models-and-building-generative-ai-applications ↩

https://cloud.google.com/blog/products/compute/introducing-a3-supercomputers-with-nvidia-h100-gpus ↩

←→ Llama 2 Prompt Template
July 2023

The GPU Song (GPUs Are Fire) - Lyrics and Video →←
July 2023