AI Compute Workloads Shift: Training vs. Inference and the Impact on Chip Leaders

Training vs. Inference: Diverging Compute Demands in Conversational AI

This article was written by ChatGPT's new Research offering. It took 6 minutes of compute and researched 42 articles. My intent was to both shed some light on the topic, and on ChatGPT's new Research capabilities. Please excuse the formatting.

AI training and inference place very different demands on hardware. **Training** a large language model (LLM) means running billions of examples through the network with high numerical precision, often distributed across many GPUs for days or weeks

petewarden.com
. The goal is maximizing throughput – crunching as many matrix multiplications as possible – so enormous compute and memory bandwidth are needed. In contrast, **inference** – using the trained model to respond to user prompts – involves serving one input (or a batch of inputs) at a time under tight latency constraints. Inference can leverage lower precision arithmetic (int8 or FP8) and model optimizations like quantization to cut compute per query. But while each inference step is lighter than training, the *scale* of queries can be massive once deployed to millions of users
epoch.ai
.

Reasoning models in conversational AI further blur these lines. Techniques like chain-of-thought prompting or OpenAI's new "GPT-o1" approach use *more* compute during inference to boost reasoning accuracy

blog.heim.xyz
. Essentially, the model performs additional internal inference steps (for example, iteratively refining an answer or consulting tools), trading extra compute for better results
blog.heim.xyz

epoch.ai
. This means advanced chatbots may perform multiple forward passes or search steps for a single user query. The takeaway: training is a huge one-time expense requiring maximal horsepower, whereas inference is a continuous service workload that must be optimized per query – especially as reasoning-heavy use cases increase inference computation per prompt.

Key differences in compute needs:

- *Scale & Duration*: Training runs involve *trillions* of operations over long durations; inference handles fewer operations per query but runs continuously at scale.

  epoch.ai

- *Precision & Batch Size*: Training uses higher precision (FP16/FP8) and large batches to maximize throughput; inference uses low precision (int8) and processes smaller batches or single requests for low latency.
- *Compute Patterns*: Training benefits from dense GPU clusters with fast interconnects (for parallelism), whereas inference often can be served on smaller systems or even edge devices if the model fits, focusing on response time.
- *Workload Flexibility*: If more compute is available, one can train a bigger model or, alternatively, spend it on more thorough inference. Notably, a smaller model with more inference compute (e.g. iterative reasoning) can sometimes match a larger model's quality

  epoch.ai

  . This "training-inference tradeoff" means AI practitioners must balance investing compute in making models smarter versus serving them smarter.

The Shift to Inference: From One-Time Training to Everyday AI Services

With several GPT-scale models now trained to impressive capabilities, the industry's focus is shifting from building models to *deploying* them. In fact, some analysts suggest the "training era is over" and 2024 will be the **"year of LLM inference"**

foundationcapital.com
. The rationale is simple: once powerful models exist, the next challenge is running them in real products and services. The compute cost of inference is quickly overtaking the cost of training in many cases. OpenAI, for example, is estimated to spend more **per week** on ChatGPT's inference than it cost to train the model in the first place
foundationcapital.com
. Serving millions of user queries day in and day out adds up – Sam Altman revealed ChatGPT processes around 100 billion tokens per day, roughly 36 trillion per year
epoch.ai
. By one analysis, that puts OpenAI's annual inference compute on par with its training compute usage
epoch.ai
.

This is a seismic shift. Historically, AI budgets were skewed toward training huge models (an expensive R&D investment by a few tech giants

uncoveralpha.com
), while inference deployment was limited. Now, as LLM-powered products proliferate, **inference workloads are exploding**. Industry watchers predict that in the next couple of years, the compute capacity dedicated to inference will be a multiple of that for training – estimates range from **3–4× up to 10× more** hardware needed for inference globally
nextplatform.com
. Enterprises rolling out AI features will contribute to this surge, moving beyond just the "five frontier model labs" to thousands of organizations running AI inference.

Crucially, inference also needs to get *cheaper* to enable this broad adoption

nextplatform.com
. It's not viable for a chatbot or AI assistant to require dozens of top-end GPUs per user session. Early GPT-4 deployments reportedly needed 8–16 GPUs to serve a single query within 200ms latency; with larger models, that might rise to 32 GPUs per query
nextplatform.com
. This ratio is improving (the largest models might train on 24,000 GPUs but serve on 32 GPUs)
nextplatform.com
, yet the overall cost pressure is immense. If millions of GPUs worldwide are used for training today, **tens of millions** could be needed for inference in coming years
nextplatform.com
. One analysis concludes that even if inference hardware becomes 10× cheaper, the sheer volume deployed could equal training hardware spend – but with far tighter margins
nextplatform.com
. In other words, inference might generate as much revenue as training, but it will be a more cost-sensitive, high-volume business
nextplatform.com
.

All this sets the stage for a new phase of competition. When AI was all about headline-grabbing model training runs, NVIDIA reigned supreme. Now that *serving* those models efficiently is the bottleneck, a wider cast of players (cloud providers, startups, alternative chipmakers) see an opening

nextplatform.com

nextplatform.com
. The inference boom is a double-edged sword: it promises more ubiquitous AI services (and thus greater demand for chips), but it also pressures companies to **optimize cost per inference** aggressively. The industry consensus is clear – AI inference must get dramatically cheaper and more efficient for generative AI to achieve wide adoption
nextplatform.com

nextplatform.com
.

“Nobody Needs an H100 Anymore”? The Push for Cheaper Inference

With this shift, some in the AI community provocatively claim that *“nobody needs an NVIDIA H100 anymore”* for most work – the idea being that once a model is trained, you can deploy it on far cheaper hardware. There’s a kernel of truth here for many use cases. Inference doesn’t require the full prowess of a $30,000 high-end GPU designed for giant multi-node training jobs

latent.space

. Companies are discovering they can save dramatically by using more *cost-efficient* accelerators for serving models. Even NVIDIA acknowledges this: the company now offers the **L40S**, a data center GPU optimized for inference, which delivers roughly one-third the performance of an H100 but at about one-fifth the price
latent.space
. The L40S skips costly training-centric features like NVLink interconnects, providing 48GB of VRAM and ample tensor cores for inference tasks – a clear sign of NVIDIA segmenting its lineup to address mass-market inference needs
latent.space
.

Comparison of NVIDIA’s Hopper H100 vs. Ada Lovelace L40S GPUs. The L40S offers about one-third the BF16 compute throughput and less memory bandwidth of an H100, but at a fraction of the cost – making it attractive for inference deployments where multi-GPU training connectivity (NVLink/InfiniBand) isn’t required
latent.space
.

Beyond NVIDIA’s own lineup, **alternative chips are stepping up for inference.** Startup and academic communities have shown that even older or less specialized hardware can handle surprising loads. For instance, Huawei’s Ascend 910C AI processor (launched in 2019) can achieve about **60% of an H100’s inference performance**

tomshardware.com
despite being several generations behind in process technology. It’s not viable for cutting-edge training, but for serving a model it can suffice, demonstrating how inference workloads are more forgiving of non-NVIDIA hardware
tomshardware.com
. Likewise, when running smaller open-source models (say 7B–13B parameters), many practitioners get by with consumer-grade GPUs like NVIDIA RTX cards or older A100s, avoiding the premium of H100. In fact, a glut of GPUs from the crypto-mining world has flooded the market and is being **repurposed for AI inference**, albeit mostly for modest model sizes due to memory and bandwidth constraints
latent.space

latent.space

.

AMD and Intel are also keen to exploit this trend. **AMD’s Instinct MI300X** accelerators, for example, come with 192GB of HBM memory – allowing a large LLM to reside on a single card – and they’ve shown strong single-node performance per dollar in both training and inference

latent.space
. A recent analysis found AMD’s latest GPU can be a *drop-in replacement* for H100 in many inference scenarios, offering more memory and competitive throughput at lower cost
latent.space

latent.space
. The catch has been software maturity (drivers and libraries), but for the common AI models (like Meta’s Llama 2 family), these alternatives are increasingly “good enough” once compatibility issues are sorted
latent.space
. Intel, too, has entered the fray with its Gaudi2 and upcoming Gaudi3 AI accelerators focused on cost-efficient throughput; one published test showed Gaudi2 training a vision model on par with A100s at lower cost
medium.com
(though Intel’s traction in the market remains limited).

So, does *nobody* need an H100 anymore? That’s an overstatement – the **frontier model developers still can’t get enough of them**. As of late 2024, demand for H100s was so extreme that secondary rental prices hit $8 per hour, and big cloud providers were effectively sold out for months

latent.space

latent.space
. Meta, for one, reportedly deployed over **100,000 H100 GPUs** for training its next-gen Llama 3 and Llama 4 models
tomshardware.com
. For cutting-edge R&D on 100B+ parameter models, nothing yet beats NVIDIA’s top-tier silicon and its software ecosystem. However, the *median* AI inference workload is indeed gravitating away from such ultra-high-end chips. Startups and enterprises running GPT-sized models for their own applications are finding they can use cheaper NVIDIA cards (like older A100s or the L40 series) or rival accelerators, and still meet their service requirements
news.ycombinator.com
. The market is stratifying: H100s (and soon NVIDIA’s even mightier Blackwell GPUs) will power the **frontier training and the largest inference clusters**, while a long tail of inference jobs runs on a mix of less costly devices. In short, not everybody needs an H100 – but those who do *really* do, and they’ll pay for the privilege.

NVIDIA’s Blackwell Generation: Upping the Ante for Training *and* Inference

Far from ceding to this inference commoditization, NVIDIA is doubling down on innovation at both ends of the spectrum. The upcoming **Blackwell** GPU series (expected in 2025) is poised to deliver major leaps in performance – not just for training, but explicitly for inference efficiency at scale. Jensen Huang, NVIDIA’s CEO, has hinted that Blackwell could be the company’s *“most successful product in the history of the industry”* and noted early demand has it **sold out for the next year** before launch

latent.space
. One headline spec: a Blackwell-based system called the **GB200 NVL72** (combining 72 Blackwell GPUs with NVIDIA’s Grace CPUs in a unified architecture) is claimed to provide **up to 30× the LLM inference throughput of an equivalent H100 setup**
nvidianews.nvidia.com
. NVIDIA says this rack-scale Blackwell platform can slash inference cost and energy per query by 25× relative to today’s state-of-the-art
nvidianews.nvidia.com
– a clear acknowledgment that the future is about serving models efficiently. It’s designed to handle models up to 10 *trillion* parameters in real-time, underscoring how Blackwell targets the next chapter of generative AI deployment
nvidianews.nvidia.com
.

In essence, NVIDIA is trying to stay ahead of the “nobody needs H100” narrative by making sure tomorrow’s flagship is indispensable. Blackwell GPUs pack 208 billion transistors each

nvidianews.nvidia.com
, and NVIDIA is leveraging advanced packaging (like multi-die “chiplets” on a shared module) to boost performance per node. For example, Blackwell’s multi-GPU modules use 10 TB/s chip-to-chip links to function as one giant GPU with 30 TB of memory
nvidianews.nvidia.com

nvidianews.nvidia.com
– a dream for both massive model training and ultra-fast inference on gigantic models. These engineering feats aim to keep hyperscalers locked into NVIDIA for the highest-end needs.

At the same time, NVIDIA is expanding its product stack downward. The company’s latest announcements include not only the 8-GPU HGX boards for Blackwell (for large-scale training clusters), but also inference-focused cards and cloud services. The fact that NVIDIA itself markets the L40S as an H100 alternative for inference

latent.space
shows a pragmatic strategy: capture the inference market with tailored offerings before competitors do. NVIDIA’s **software moat** also remains a huge advantage – its TensorRT and AI

frameworks are deeply integrated into enterprise AI stacks, making GPU deployment turnkey in many cases. Even if a competitor's chip matches performance, customers often stick with NVIDIA to avoid porting code or retraining talent
news.ycombinator.com

petewarden.com
. In short, NVIDIA is preparing to service *both* extremes: the highest-end training and inference with Blackwell (at premium pricing), and mainstream inference with more cost-effective GPU options. This two-pronged approach is how NVIDIA plans to retain as much share as possible in a world where inference workloads multiply and diversify.

Cloud Providers' ASICs: Google and Amazon Bet on In-House Silicon

The major cloud providers are pursuing a different path to cope with the inference surge: build their own chips. **Google** was first, developing its Tensor Processing Units (TPUs) to power internal AI workloads and Google Cloud offerings. Now on its fifth generation, Google's TPUs have evolved into a dual strategy: a max-performance chip for training and a cost-optimized chip for inference. The recently announced **TPU v5e** (the "e" for efficiency) is explicitly aimed at high-volume inference and smaller-scale training. It has lower FLOPS and memory bandwidth than an H100

semianalysis.com

semianalysis.com
, but Google's goal isn't raw peak – it's **better performance-per-dollar at scale**. Thanks to vertical integration, Google can operate TPUv5e at lower power and cost (no NVIDIA markup), achieving excellent total cost of ownership for serving models up to ~200B parameters
semianalysis.com
. Early figures from Google show that Cloud TPU v5e can offer a *massive cost advantage* on many workloads – in one case, generating 1,000 AI images on a TPU v5e pod cost only $0.10, a fraction of the cost on GPUs
cloud.google.com
. Google has also improved the software tooling (like PyTorch XLA) to make porting models to TPU easier, boasting that for most LLMs, using TPUv5e "is as easy as a GPU" and can even yield higher utilization with less effort
semianalysis.com
.

Amazon Web Services is similarly all-in on custom silicon. AWS's **Inferentia** chips (for inference) and **Trainium** chips (for training) are built by its Annapurna Labs team to reduce dependence on NVIDIA in the cloud. The latest **Inferentia2** (launched 2023) delivers impressive results on NLP tasks. Amazon reports that, compared to an NVIDIA A10G GPU instance, Inferentia2 offers **2.6× higher throughput** and **8× lower latency** for popular transformer models, at **70% lower cost** per inference versus GPU instances huggingface.co

huggingface.co
. In practical terms, an Inf2 instance can serve a BERT or GPT model several times faster and cheaper than a comparable GPU server – a huge draw for cost-conscious customers. AWS has successfully attracted users like **Anthropic**, an AI startup rivaling OpenAI, which agreed to **train and deploy its future models on AWS's Trainium and Inferentia chips** as part of a strategic partnership
press.aboutamazon.com
. This kind of deal not only validates AWS's silicon but also ensures high-profile workloads commit to AWS infrastructure. Inferentia2 is designed so that even very large models (like GPT-3 175B) can be sharded across multiple chips – an Inf2 48xlarge instance has 12 Inferentia2 chips with 384 GB total memory, enough to host a model of that size in memory
huggingface.co

huggingface.co
. The emphasis is on scalability and integration: AWS offers a full software stack (Neuron SDK, deep framework integration) to make using these chips as painless as possible for developers, reducing the inertia to switch off GPUs
ankursnewsletter.com

huggingface.co
.

Other cloud giants are not sitting idle either. **Microsoft**, heavily invested in OpenAI, is reportedly developing an AI chip (codenamed Athena) to deploy OpenAI's models more cost-effectively on Azure – though details are scarce, and for now Azure continues to buy tens of thousands of NVIDIA GPUs for its AI supercomputer. **Meta** has designed an internal accelerator as well (the MTIA for inference), aiming to handle recommendation models and LLM inference in-house. While Meta's first attempt was reportedly shelved for being underpowered, they are likely to return with a more competitive chip around 2025. For Meta, which uses AI across Facebook, Instagram, Ads, and more, shaving inference costs at its immense scale can save billions. Indeed, Meta has already adopted AMD's MI300X GPUs for some inference/training workloads in its datacenters

crn.com
, showing it will mix and match technologies to optimize performance per dollar.

The **performance of these in-house ASICs vs. NVIDIA GPUs** has reached a tipping point where they are "good enough" for many inference tasks. Google's latest **TPU v5p** (the performance-optimized sibling to v5e) is reportedly **faster than NVIDIA's H100** on certain workloads

news.ycombinator.com

, indicating Google can now outgun NVIDIA in specific scenarios (especially when Google codes the software to take full advantage). Meanwhile, AWS's chips focus on efficiency: AWS claims many customers achieved **50–70% cost savings** using Inferentia1 and 2 versus GPUs huggingface.co

. The gap in absolute performance is narrowing, and any remaining shortfall is often made up by the pricing advantage. However, these cloud chips largely remain available only within their respective clouds (AWS Trainium/Inferentia on AWS, TPU on Google Cloud). They give the cloud providers a margin edge and a way to lure customers, rather than directly threatening NVIDIA's merchant business in the open market. Still, the broader implication is clear – for the growing inferencing market, **NVIDIA no longer has a monopoly** on high-performance silicon uncoveralpha.com

. Inference is viewed as a more **open playing field**: one expert quipped that the only reason competitors can beat NVIDIA in inference is NVIDIA's hefty 70% margins – implying if NVIDIA simply cut prices, much of the advantage of custom chips would evaporate news.ycombinator.com

. In response, NVIDIA could indeed adjust pricing or offer cloud-friendly models (like lower-cost licenses for its software on cloud ASICs, or special pricing on older GPU generations for inference pools). The competition between cloud ASICs and NVIDIA's GPUs will likely drive down inference costs further, benefitting users in the end.

Outlook: Inference Growth, Market Forecasts, and Financial Implications

Over the next 1–3 years, industry insiders widely agree that **inference workloads will see explosive growth** relative to training. By 2025, we can expect a vast expansion of AI inference capacity across cloud and enterprise data centers. One forecast pegs the global "AI inference server" market at ~$13.2B in 2024, tripling to nearly $50B by 2033

wicz.com

, but even that may underestimate the near-term craze of deploying generative AI into every app, call center, and productivity tool. It's conceivable that **inference compute demand could outstrip training by a factor of 5–10×** in aggregate hardware usage nextplatform.com

, given the millions of potential end-users for AI services petewarden.com

. As Pete Warden noted, even if a single training run is enormously expensive, there are *billions* of daily queries potentially to run – eventually the cycles spent answering questions will surpass those spent learning in the first place petewarden.com

. We're already seeing hints of this: for one major LLM, roughly equal compute has been invested in training the model and running it for users so far epoch.ai

, and usage is still climbing.

That said, *training* isn't slowing down either. Every year brings new state-of-the-art models that often require 2× or 4× more compute than the previous generation. Competitive pressure in AI (among a handful of leaders like OpenAI, Google, Meta, Anthropic) ensures that money will be poured into ever-larger training runs to maintain an edge

news.ycombinator.com
. For example, if GPT-5 or a successor to PaLM or Llama is in the works, it might consume tens of thousands of the latest GPUs for months – each such project is a bonanza for NVIDIA (and possibly AMD) sales. We may also see a proliferation of specialized models (multimodal AI, domain-specific LLMs, etc.) which require separate training efforts. So in terms of revenue, the **training hardware market will continue growing**, but it's concentrated in a few big buyers and remains NVIDIA's fortress (with some cracks now accessible to AMD). NVIDIA's launch of Blackwell in 2025, and AMD's launch of its MI300/MI350 series, are timed to serve this ongoing training arms race. AMD's CEO Lisa Su recently raised the company's internal estimate for the AI accelerator market to **$500B by 2028**
www2.deloitte.com
, reflecting the sum of both training and inference opportunities. By 2025, Deloitte projects that "GenAI chips" (GPUs, NPUs, TPUs, etc. for AI) will exceed **$150B in annual revenue**
www2.deloitte.com
– a huge jump from only ~$50B in 2023. This implies an immense uptick in volume, much of it likely attributable to inference deployment at scale.

For NVIDIA, the financial narrative might shift from puregrowth to mix and margins. 2023 saw NVIDIA's data center revenue skyrocket – the company's data center sales roughly doubled year-over-year, on track to ~$40 billion in the second half of 2024 alone, fueled by a backlog of H100 orders for training clusters

crn.com
. Going forward, Nvidia will still sell out its flagship GPUs (especially with Blackwell's debut), but a larger portion of units sold may be the lower-cost cards for inference or previous-gen chips repurposed for serving. This could moderate the average selling price. On the other hand, **unit volumes** could be much higher. If the world truly needs "hundreds of millions of GPUs" for inference as Next Platform mused
nextplatform.com
(even if many of those are specialized ASICs, a good portion will still be NVIDIA-based), NVIDIA could see continued revenue growth by sheer volume – albeit with somewhat lower margins per unit. The company's strategy to offer DGX Cloud (renting GPU capacity directly) might also play a role, converting some one-time hardware sales into ongoing service revenue as inference demand grows. Investors will be watching how well NVIDIA can fend off competitive pricing pressure in inference. If it maintains a strong enough ecosystem, it may not need to slash margins much; but if cloud providers start achieving too much price/performance advantage with their own silicon, NVIDIA might choose to adjust pricing or offer "inferencing licenses" for its GPUs at discount to keep customers in-house
news.ycombinator.com
.

AMD stands to be a wild card beneficiary of the inference trend. After years of minimal share in AI accelerators, AMD's Instinct line is finally gaining traction. In 2024, AMD exceeded **$5 billion in AI GPU revenue** for the first time

crn.com

, thanks to MI300 series sales to major cloud players. Lisa Su is openly bullish, expecting AMD's AI chip business to scale to "tens of billions" annually in the next few years
crn.com

crn.com

. Much of that growth will come if AMD can position its GPUs as a cost-effective choice for inference farms and fine-tuning tasks. The company has already scored wins: AMD revealed that **Meta is deploying Instinct GPUs at scale for both inference and training** – a breakthrough reference customer
crn.com
. If Meta's usage expands (and if other hyperscalers like Microsoft or Oracle also adopt AMD for certain workloads), AMD could steadily chip away at NVIDIA's dominance. Financially, even a 10–15% share of the AI accelerator market by 2025 would translate into a multi-billion jump in AMD's top line. AMD is also leveraging its CPU presence: its upcoming EPYC processors integrate AI engines that could handle lightweight inference on the CPU side, potentially appealing for edge or less demanding scenarios (though NVIDIA and Intel are doing similar with their CPUs). The big question mark for AMD is software – a "large software gap" still exists in comparison to NVIDIA's CUDA ecosystem
latent.space
. If AMD can close that gap (through tools like ROCm, or by leaning into open AI frameworks that abstract the hardware), it can truly capitalize on the inference boom. If not, some potential customers may stick with NVIDIA despite higher cost, simply for ease of development.

For the **cloud providers**, the surge in inference is both a challenge and an opportunity. On one hand, they must invest heavily to build out capacity – both purchasing GPUs (still largely from NVIDIA) and developing/deploying their own ASICs. It's notable that **capital expenditures at the big cloud companies have spiked** due to AI; Meta, for instance, significantly raised its 2023 CapEx guidance largely to expand AI infrastructure (training and inference both). In the short term, this is a cost, but the strategic play is to **own the customer's AI workload** and reap the rewards over time. By offering AI inference as a service, cloud providers will generate substantial recurring revenue. Microsoft, Google, and Amazon all have introduced managed AI inference services (from OpenAI's API on Azure to Bedrock on AWS to Vertex AI on GCP). The profitability of these services hinges on how efficiently the cloud can run the models – hence the incentive to use their in-house chips where possible. If AWS's Inferentia can serve a model at half the cost of using an NVIDIA GPU, AWS can either pocket the margin or undercut competitors' pricing to win business. Cloud providers are also likely to bundle AI inference with other cloud services, making it sticky.

One interesting dynamic is that **cloud providers are simultaneously NVIDIA's biggest customers and emerging competitors**. AWS, Google, and Azure collectively buy tens of billions of dollars worth of NVIDIA GPUs, which bolsters NVIDIA's financials. But each is also trying to reduce that reliance by shifting workloads to their own silicon. The financial impact on NVIDIA will depend on how successful they are. For example, if Google manages to run 50% of its AI inference on TPUs by 2025 (hypothetically), that's thousands fewer GPUs NVIDIA sells – but NVIDIA might still sell out to other customers. The cloud titans have the deep pockets to pursue custom chips, and their efforts will likely **compress the margins of AI compute** over time (good for customers, tougher for chip vendors).

In summary, we are entering a phase where **AI inference is the main growth driver** of compute demand. Training isn't going away – in fact, it may keep scaling in step with model ambitions – but the *financial center of gravity* in AI is tilting toward deployment. We can expect overall spending on AI infrastructure to keep climbing steeply, with more of that pie allocated to inference. NVIDIA sits in a strong position but will have to navigate a more crowded competitive landscape. AMD has a window to substantially grow its AI revenue if it executes well in this inference-focused era. And the cloud providers, by wielding custom silicon and massive investments, will seek to capture value themselves rather than simply passing checks to chip vendors.

From an investor's vantage, one implication is that NVIDIA's current sky-high margins may normalize a bit as focus shifts to the "brutal low-margin business" of inference volume

news.ycombinator.com

. NVIDIA could end up analogous to an "Apple of AI" – dominating the high end (training, premium systems) and still selling a lot of units for inference, but facing competition in the commodity tiers

news.ycombinator.com

. Meanwhile, companies enabling cheaper inference (whether AMD, cloud ASIC divisions, or startups like Groq and Cerebras pivoting to cloud inference

nextplatform.com

nextplatform.com

) could see new revenue streams. The next 1–3 years will likely redefine market share in accelerated computing more than any time in the last decade.

One thing is certain: **AI workload demands will continue rising**. The exact split of dollars between training and inference may settle around parity or tip in inference's favor, but both are growing in absolute terms. As Karl Freund and other industry analysts have pointed out, the real winners will be those who can supply this appetite in the most cost-effective way. That means higher-density chips, better software optimization, and innovative architectures purpose-built for inference are all on the horizon. It's a classic tech cycle – after a burst of innovation (in model design) comes the burst in deployment and optimization. For NVIDIA, AMD, and the cloud giants, the race is on to provide the **pick-and-shovel infrastructure** of the AI gold rush, from training the motherlode models to serving billions of AI queries efficiently. The balance of power in AI hardware is poised to shift, driven by the simple fact that answering questions for millions of users is now a bigger business than just teaching AI to answer them in the first place.

Sources:

epoch.ai

foundationcapital.com

nextplatform.com

nextplatform.com

news.ycombinator.com

latent.space

latent.space

tomshardware.com

latent.space

latent.space