Rethinking AI Demand Part 1: AI Data Centers Are Experiencing a Surge of Training Demand – What Happens When the Surge is Over?

Published on Alvarez & Marsal | Management Consulting | Professional Services (https://www.alvarezandmarsal.com)

February 25, 2025

BACKGROUND

Larger Checks with Fewer Certainties

Hyperscalers and AI players are spending billions of dollars on AI infrastructure and data centers, with Amazon, Google, MSFT and Meta spending $217 billion in 2024 alone[1] (and committing to more than $320 billion in 2025),[2,3,4,5] and OpenAI and Softbank have announced plans for investments of $500 billion over the next four years.[6]

Most of this will be spent on IT hardware, but a meaningful portion is allocated to data center builds. More than half of these mega-campus builds will be outsourced to third-party operators to expedite deployment. Currently, a handful of third-party operators provide the lion's share of new hyperscale colocation capacity, but a new class of crypto-mining-turned-AI-data-center operators is also emerging.

Third-party operators are increasingly turning to outside partnerships for coinvestor funding to build these $1 billion+ mega-campuses. In the past, investors found these partnerships attractive for several reasons: Capex was not expended speculatively (build funding was committed only after contracts were in hand), pricing was tied to build costs, and rents were seen as a sticky and stable revenue stream on a fully utilized asset with creditworthy tenants.

The current data center builds are much more focused on training, which is a fundamentally different workload type from what has been built in the past. Previously, the focus had been on hyperscale cloud workloads, and there is palpable market anxiety about the durability of AI training deployments given the amount of speculative investment poured into the space. Even before the recent DeepSeek news, we've seen a stream of headlines questioning the ROI on training investments dating as far back as the Sequoia Capital article[7] in 2023.

A&M's Telecommunications & Digital Infrastructure (TDI) group has deep experience within the hyperscale data center space. In this paper, we seek to answer three fundamental questions that are critical for data center operators and investors: 1) How will training needs evolve over time from a space and power consumption perspective? 2) What happens to AI training data centers once the training boom is over? 3) What is the impact of these dynamics to data center investments?

INSIGHTS

Insight One: While Some Need for Ongoing Training Will Persist, Training Needs in Deployed Facilities Are Likely to Decline In Five

to Ten Years

We know that training demand is going to be front-loaded into the next two to five years, to be followed by deployment and commercialization. *"[Currently] 80 percent of that demand curve is driven by the need for these large-scale training facilities [while] only 20 percent of that is driven by inference and Gen AI. … [By] the time we get to the 2030/32 horizon, we'll see that ratio reversed,"* a recent lead for AI capacity planning at a "Big 3" hyperscaler[8] tells us. This inversion of training to inference ratio for new workloads (absorption) is widely accepted across the industry, but what does it mean for deployed training facilities? Our analysis leads us to conclude that training needs in deployed facilities are also likely to decline to some extent, driven by three key factors: 1) Training is a cost, not a direct revenue driver. 2) For any given model generation, training is frontloaded, with the majority of follow on workloads stemming from inference 3) Training will get more efficient over time.

1. Training is a cost, not a direct revenue driver. For any training model, the monetization strategy involves a shift from training to inferencing to drive end user adoption and revenue generation. While hyperscalers are currently in a spending frenzy deploying AI training Capex, they are also cognizant of the difficulty and risk involved in estimating ROI on this. The former Big 3 lead of AI capacity planning offered the following analogy: *"When we initially discovered electrification, we couldn't understand all the use cases that we could monetize around [it]; we're in [a similar moment right now, where] we know that we can monetize a lot of use cases once we get to AGI/ASI… but we don't know what those exact use cases are, or when they're coming. [This strategy is now pursued by large hyperscalers] … with deep pockets as they are going for a moonshot goal … but it's still a risky proposition for a public company."* They were careful to draw the distinction between measuring the ROI for capex expended on cloud and inference workloads, which can be directly tied to revenue, from those of training: *"When you're doing capacity planning and ROI planning around your own services like enterprise cloud [or around] Gen AI and inference, that's a very direct activity we can attach ROI to. When we're looking at ROI for training, there is no pragmatic model that we can apply against this. ... The ROI on something like a Stargate is only known 10, 15 years down the road,"* they explained.

2. The second important distinction to draw between the two workload types is that training is largely frontloaded, whereas cloud/inference workloads are recurrent and ramping over time. Each generational model is trained over a training run and then shipped off to be deployed via inferencing, a recurring/reoccuring activity. Yes, future iterations of models will be continuously trained and refined, and the need for additional training will likely persist into the longer term for newer models and new use cases, but the amount of compute necessary to refine models needs to approach a fractional share of total compute to generate a positive ROI and justify the long-term business case for any given use case. Furthermore, the majority of AI spend are currently going toward IT hardware (mostly NVIDIA GPUs) with an incredibly short lifespan of under three years[9] when used for training. This means that training requires a constant stream of reinvestment, making it more capitally intensive to do so on an ongoing basis.

3. Finally, training is bound to become more efficient over time. Even before the DeepSeek news, capacity planners and industry experts have long known that training models become more efficient over time. Epoch AI[10] expects training runs to increase in efficiency by a factor of ~24 times by 2030, driven by hardware efficiency, longer training runs and power efficiency gains. The question, of course, is how much greater the compute needs for future training demand will be versus today's models. According to Epoch AI (writing in summer of 2024, prior to the latest discussions on further training efficiency gains), future training models could be 5,000 times more computationally intensive than today's models, but 24 times as efficient. With this math, we would need approximately 200 times more total power to run a future training model than we run today's models (the power difference between ~30 MWs per training run today, versus 6 GWs per training run in 2030). It is important to note that for these predictions, from either a compute or efficiency perspective, the cone of uncertainty is very large, leaving even the best data scientists and hyperscale capacity planners in a wait-and-see mode when asking the "500 billion dollar question" of how much training capacity will be necessary in the medium to longer term, and for how long.

For now, the broad consensus is that even in deployed facilities, training requirements are likely to persist for some time, but then eventually decrease. *"So there's always going to be demand for training,"* a director of AI and HPC at another hyperscaler (non-Big 3) says, *"but at some point, you are not going to need that entire hundred megawatts forever…you'll only need maybe 60 or 70 out of the hundred, but it will take a while [five to ten years] to get to that point."* A third hyperscaler market participant offered the following view after bouncing our line of questioning off their data center infrastructure team: *"Seems like you are asking the billion dollar question. … Some data centers might find a niche in supporting specific types of AI workloads for ongoing training …[but*

other] data centers built specifically for the training boom might become obsolete and either be mothballed [or] redeveloped."

Insight Two: With So Many "Known Unknowns," Today's Training Facilities Need to Accommodate for Infrastructure Flexibility Because Their Underlying Workload Type May Change Over Time

With this much uncertainty around future training needs and the nature of facility usage, a key feature highlighted in our research was facility flexibility: "*So I think what the key feature ["Big-3" Hyperscaler] is building for in designing their new data centers is to really be agnostic to the type of innovation that will transpire. There are a lot of known unknowns around the space and site planning*," says recent lead of AI capacity planning for a Big 3 hyperscaler. A vice president of data center development at one of the top third-party operators put it another way: *"Part of the reason why hyperscalers have greenlit these hundreds of billions of dollars of training centers is because they see multiple use cases for them down the line. These aren't one-trick ponies. These are 20-plus year centers that can be useful for training, inferencing or cloud."* Along the same lines, our third hyperscaler market participant (Big 3) emphasized the importance of "*design[ing] data centers with flexibility allowing for potential repurposing or adaptation"* and *"collaborat[ing] with AI companies and cloud providers to understand their long-term needs and plans"* to maintain a competitive advantage for a third-party operator.

At the same time, not all facilities can be as easily repurposed for inferencing; rural facilities far from population centers, for example, do not meet many of the requirements for inferencing deployments. While training benefits most from scale and ultrahigh-density compute, and is not sensitive to latency, inference is more latency sensitive and needs to be proximal to end users to drive positive experience and further adoption (similarly to cloud). With training, hyperscalers are trying to optimize for markets with available power at scale, speed to market, and lowest TCO (total-cost-of-ownership). With inference, while TCO is still important, connectivity and latency become determining table stakes. In light of the uncertainty around long-term training requirements, many hyperscalers are preferring to deploy in lower latency locations for training, even though that is not a direct technical requirement for the workload currently, because of easier repurposing capabilities down the line, if required.

Insight Three: Despite Uncertainty, Data Center Development Revenue Durability Is Likely to Persist, Especially for Facilities in Established Markets With Appropriate Specs

Despite the many uncertainties around the demand arc for training deployments, there appears to be a broad consensus around the revenue durability for most mega-campus developments. As long as we accept the argument that today's training facilities can be repurposed for tomorrow's mixed-use needs, third-party facility durability/customer stickiness largely holds for the same reasons it's always held: because migrating workloads is costly and almost never worth the effort for hyperscalers, and because need for power and compute in core and edge markets only grows. A vice president of data center development at one of the leading North American operators makes the following case for customer stickiness: "*1) Hyperscalers use [data centers] until the point of diminishing returns, and that's historically always been like 20-plus years out. 2) Even if there is a slowdown [in demand], the needs for centralized digital infrastructures for [the hyperscalers] will continue to increase. 3) Even if there is a step function technological advancement … the best facilities in the industry, like what [my company] builds, what my competitors build, can also be upgraded [to stay] relevant based on technological changes."*

To hedge against the risks of stranded facilities in rural locations after a transition from training to inference, many hyperscalers are choosing to deploy in safer low-cost/relatively low latency markets, with partners known for their ability to execute and deliver future-proof facilities. The recent lead of AI capacity planning for a Big 3 hyperscaler explains in her own words: *"Those large leases with the very reputable colocation providers will absolutely have revenue durability. These will be very sticky because they've done the fit-out to accommodate for the lease, they bring their own dielectric liquid cooling system, they've got the right air systems in place, and they've got the right portfolio level partnerships. As the cloud is going to grow in tandem with the inference in Gen AI [as training needs wane], anyone who has large-scale leasable space under 10, preferably under three milliseconds, from a population center [is well positioned]. If these data centers were just for training, I wouldn't say there was a lot of durability."* Another hyperscaler AI capacity expert reaches a similar conclusion: *"You will have a stable, stable rent from your client base, especially for the larger reputable companies."* He further elaborates on the dual use appeal of population center deployments: *"The AI companies that are deploying AI-only workloads, they still prefer to have these data centers in highly populated areas … because they utilize a two-pronged approach where half their workload is running AI training and the other half is inferencing, or they may have an entire site that's built for training [that] completely converts over to inferencing over time."*

Insight Four: We Need to Consider Workload Churn When Forecasting Training Capacity Over Time

The consensus around facility flexibility and possibility of inference/cloud repurposing has some profound implications for AI

workload demand forecasting — it implies the possibility of training workload churn; the tens of GWs of training power accruing in top U.S. markets in the next two to five years may always stay live, but may not always be entirely earmarked for training. In the pre-AI hyperscale cloud paradigm, the industry has always treated market absorption on a gross basis as entirely accretive to TAM (total addressable market): New demand for cloud went into new facilities; existing demand stuck around. If inferencing and cloud demand may be replacing training demand to some extent in existing facilities, it needs to be factored into capacity planning in a way that isn't being thought through yet by operators and investors, as far as we can tell.

Further, in rural markets where the business case for deployment rests entirely on TCO, and connectivity is poor/latency is high, the potential for churn is higher still. Our Big-3 AI capacity planning lead, who expressed generally bearish views on rural deployments, envisions certain scenarios where there is stranded power that may need to be downsized or sold off: *"I think we potentially could end up with some stranded power. ... You can either downsize your PPA, or there's a transfer clause. … Now I have a powered land asset that I can turn around and sell to someone who may want to be in that rural area, so it's still a valuable asset; it just may not be used for training forever."* A more optimistic outlook expressed by other hyperscalers willing to take the bet on rural deployments follows one of the following arguments: a) Some degree of recurring training compute will be served by these facilities in perpetuity, due to highly favorable TCO conditions, b) An emerging labor force ecosystem will spring up in rural locations where latency insensitive workloads of all types from different hyperscalers converge into a rural ecosystem due to talent availability and favorable TCO conditions, and c) Virtual compute and/or ongoing GPUaaS will be supplied to enterprise/SMB users for their long-term latency-insensitive AI training/inference needs.

Regardless of whether training deployments will be repurposed for other inference needs, or other types of workloads, it's clear we should stop viewing all forecasted future absorption as net-new incremental demand. Many of today's capacity forecasts are oversimplifying the picture when they paint the installed capacity of training MWs as entirely permanent, or attempt to arrive at training demand by assuming all future chip shipments contribute to net-new demand (even as we know GPU lifespans for training are very short). As training MWs sunset and are replaced with inference and cloud/AI hybrid MWs, some of the forecasted absorption will replace workload churn rather than add to the workload base on a net-new basis, while other deployments may be downsized altogether in the medium to longer term.

CONCLUSION

Reflecting on these insights in light of the rapidly shifting AI landscape, we leave you with three practical considerations:

1. **Consider your market:** Some markets lend themselves easily to revenue durability cases beyond the AI training case, while others rely on a series of looser suppositions to make the long-term durability argument hold.

2. **Rethink your demand forecast:** There is a lot of uncertainty around the scale, need and pace of future training deployments versus efficiency gains. Capacity forecasts that paint training installed capacity MWs as permanent without accounting for workload churn and evolution are oversimplifying the picture.

3. **When it comes to discrete data center deployments, the bubble narrative is overhyped**: Prudent investments into deployments with a high level of execution confidence and market durability will often still make sense. Desirable facilities provide core AI infrastructure with flexibility in mind at increased scale. Even though future requirements remain hazy, as long as we believe that overall compute demand will grow over time, best-in-class facilities can stay relevant.

Stay tuned for additional white papers in this series in the coming months.

Sources:

[1] S&P Capital IQ Pro, Financial Data Pull for FY2024 (Derived From 10K Reporting), https://www.spglobal.com/market-intelligence/en/solutions/products/sp-capital-iq-pro.

[2] Dan Swinhoe, "Amazon 2025 Capex to Reach $100bn, AWS Revenue Hit $100bn in 2024," DataCenter Dynamics, February 7,

2025. https://www.datacenterdynamics.com/en/news/amazon-2025-capex-to-reach-100bn-aws-revenue-hit-100bn-in-2024/.

[3] Georgia Butler, "Google Expects 2025 Capex to Surge to $75bn on AI Data Center Buildout," DataCenter Dynamics, February 5, 2025, https://www.datacenterdynamics.com/en/news/google-expects-2025-capex-to-surge-to-75bn-on-ai-data-center-buildout.

[4] Matthew Gooding, "Microsoft to Spend $80bn on AI Data Centers in 2025," DataCenter Dynamics, January 6, 2025, https://www.datacenterdynamics.com/en/news/microsoft-ai-data-center-80-billion/.

[5] Sebastian Moss, "Meta Plans $60-65bn Capex on AI Data Center Boom, Will Bring ~1GW of Compute Online This Year," DataCenter Dynamics, January 24, 2025, https://www.datacenterdynamics.com/en/news/meta-plans-60-65bn-capex-on-ai-data-center-boom-will-bring-1gw-of-compute-online-this-year/.

[6] "Announcing The Stargate Project, OpenAI, January 21, 2025, https://openai.com/index/announcing-the-stargate-project/.

[7] David Cahn, "AI's $200B Question," Sequoia Capital, September 20, 2023, https://www.sequoiacap.com/article/follow-the-gpus-perspective/.

[8] Amazon Web Services, Microsoft Azure and Google Cloud Platform offer the most extensive cloud infrastructure services worldwide. For this white paper, Alvarez & Marsal's TDI team interviewed more than five hyperscale data center market participants across the AI capacity planning and third-party operator landscape.

[9] Anton Shilov, "Datacenter GPU Service Life Can Be Surprisingly Short — Only One to Three Years Is Expected According to Unnamed Google Architect," Tom's Hardware, October 24, 2024, https://www.tomshardware.com/pc-components/gpus/datacenter-gpu-service-life-can-be-surprisingly-short-only-one-to-three-years-is-expected-according-to-unnamed-google-architect.

[10] Jaime Sevilla et al., "Can AI Scaling Continue Through 2030?" Epoch AI, August 20, 2024, https://epoch.ai/blog/can-ai-scaling-continue-through-2030.

Source URL:https://www.alvarezandmarsal.com/insights/rethinking-ai-demand-part-1-ai-data-centers-are-experiencing-surge-training-demand-what

Authors:
Asya Walters
Andy Walker
Moe Kelley