DaVinci

D|A|DAVIDSON

Compute

INDUSTRY UPDATE

Institutional Equity Research
February 12, 2026

Price (2/12/26)
Industry: TECHNOLOGY

Breaking Down the Gigawatt-Gigacycle

Alexander Platt
(503) 603-3045
AJPlatt@dadco.com

Alongside our sector launch of the semiconductor space, we are republishing our first trend within our 6-for-26 trends report, that focuses primarily around where we see the biggest opportunities in the compute ecosystem for the year. While our original report primarily provides insights within the model-layer, we isolate the non-model takeaways for investors to get a clear view on where we're focusing in the AI data center buildout for 2026.

Our high-level view as follows:

2026 will be a year of constraints in the gigawatt-gigacycle. At 100MW campuses, the story could be reduced to "can you get the chips", however, at gigawatt scale, the constraint stack widens across several bottlenecks including CPUs, co-packaged optics, advanced packaging, high bandwidth memory, NAND flash, behind-the-meter power, and more. The reality is that clearing any one bottleneck will not be enough to deliver compute at scale, and the slowest components will increasingly impact timelines for these campuses, especially as we start to see more GW campuses being planned and constructed.

On a more granular level:

Co-packaged optics will be the defining architectural transition in AI networking this year. NVIDIA's Q3450 and Spectrum-X 6800 ship in volume on TSMC's COUPE platform, and while the ~65% power reduction per link matters, we believe the real story to focus on in scale-up networking. To that end, CPO breaks the two-meter copper reach constraint on NVLink domains, enabling optical fabrics spanning hundreds of meters and dramatically larger coherent GPU topologies. The pluggable market's own 800G to 1.6T transition and the approaching limits of electrical SerDes scaling creates compounding scarcity and architectural forcing functions that we anticipate will accelerate adoption.

Advanced packaging, not wafer fabrication, is the binding production constraint for frontier AI accelerators. TSMC's CoWoS capacity has consistently lagged demand despite repeated doublings since 2023, with yield mathematics compounding multiplicatively as packages grow larger and integrate more dies. Announced timelines for custom hyperscaler/frontier lab silicon should be interpreted as more aspirational targets conditional on CoWoS allocation, with us noting in a couple of our initiations that the companies with secured packaging partnerships will separate from those constrained by bonding tool throughput in Taichung.

The memory hierarchy is shifting rapidly as KV cache becomes of greater importance alongside model weights as the primary memory consumer. High bandwidth memory (HBM) remains a binding constraint on system availability independent of GPU die supply, but for frontier models serving longer contexts (i.e. Claude Opus 4.6 1M token context window), cache memory is growing to the point where we believe it could exceed weight storage. This structurally shifts traditional assumptions about memory and most prominently pulls NAND flash into the infrastructure bottleneck narrative as KV cache offloading and video generation workloads stress storage tiers previously treated more as commodity inputs.

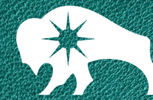
Energy is the upstream constraint governing all others, and behind-the-meter generation is the only path to gigawatt-scale deployment on 2026 timelines. Grid interconnection faces multi-year queue delays and transmission infrastructure not designed for concentrated gigawatt loads, driving adoption of on-site gas turbines that deliver hundreds of megawatts in 18-24 months. Many of the hyperscalers and neoclouds we speak to all stress power availability, and it's our belief that the AI infrastructure leaders of the next few years will be partly determined by who will be able to secure power at the gigawatt-scale.

DaVinci Overview

D.A. Davidson's DaVinci initiative focuses our technical-oriented research, data-driven insights, and prescient think pieces under one unified framework. We note that for our DaVinci coverage of deep tech businesses, we employ an early-stage venture approach focusing on technical foundations, disruptive potential, and long-term strategic value, rather than near-term financial and valuation metrics given the unique growth trajectories of pre-inflection markets.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

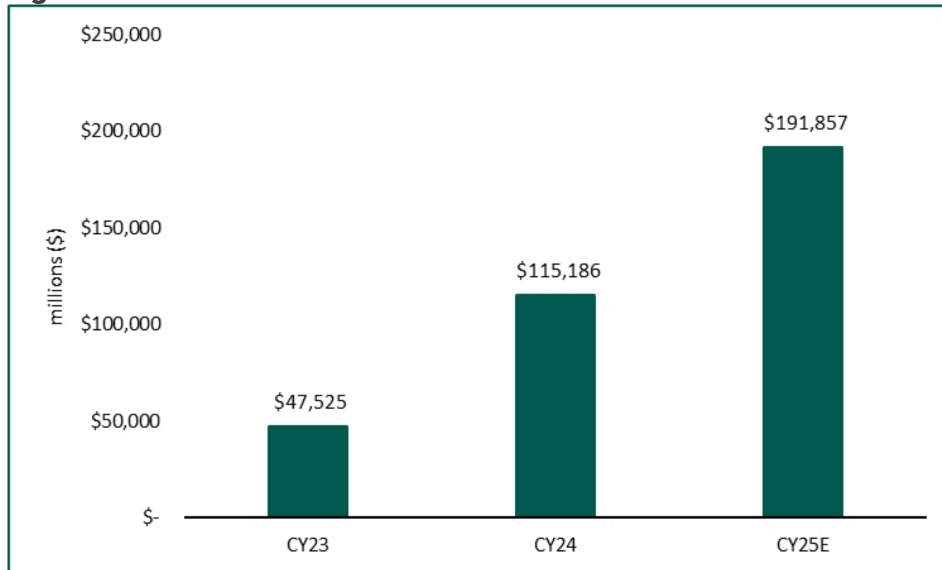
Please refer to pages 15 - 16 of this report for detailed disclosure and certification information.



The Gigawatt Giga-Cycle

The investment narrative for AI infrastructure from 2023 through early 2025 was organized around a single constraint: GPUs. Demand for AI compute exceeded the supply of NVIDIA GPUs, and this imbalance determined pricing power, capital allocation, and equity valuations across the sector. The framework was simple and largely correct. NVIDIA's data center revenue grew from $47.5B in CY2023 to our estimate of over $191B in CY2025 because it controlled the binding constraint on AI deployment. On the flip-side of things, adjacent suppliers benefited in proportion to their proximity to that constraint. However, this year will not be the same, and the single-bottleneck model is no longer adequate. The AI infrastructure buildout has reached a scale where constraints propagate through the entire system, binding at various points depending on the specific deployment configuration and timeline. GPUs remain important, and we don't expect them to lose importance anytime soon, but they are no longer the sole determinant of who can deploy compute at scale or how fast. The binding constraint has shifted in any given quarter or year across advanced packaging capacity, HBM memory allocation, optical transceiver availability, power delivery infrastructure, or cooling system lead times. And the reality is that it's often several of these at once.

Figure 1: NVIDIA Data Center Revenue CY23-CY25E



Source: Company reports, D.A. Davidson & Co.

The shift we're witnessing from a single-bottleneck regime to a multi-bottleneck regime has a very specific cause, which is that the infrastructure being deployed has crossed a threshold of system complexity where no single component's supply can be expanded independently. A 100k GPU training cluster is not 100k independent GPUs but rather it's an integrated system where the GPUs must communicate over a fabric that requires specific networking components, draw power from delivery systems rated for specific loads, dissipate heat through cooling infrastructure with specific capacity, and be packaged using processes with specific throughput. Expanding GPU supply without proportionally expanding each of these adjacent capacities does not yield additional deployable compute, it just yields chips that are waiting for the rest of the system to catch up. This type of interdependence creates what we'd categorize as constraint propagation. When any single bottleneck is resolved, the binding constraint just shifts to the next limiting component, and that component's suppliers experience the pricing power and demand surge that previously accrued to the resolved bottleneck. The total system capacity is determined by the minimum of all component capacities, following a Liebig's law dynamic familiar from agricultural and biological systems which states that growth is dictated not by total resources available, but by the scarcest resource. Which means that investment in expanding any single component beyond the capacity of other components yields no incremental system output until those other components are also expanded. This creates significant implications on the duration of the actual AI buildout, as under a single-bottleneck model, the buildout ends when supply of the bottleneck component catches demand. But under a multi-bottleneck model, the buildout continues as long as any component remains supply-constrained because resolving each constraint merely reveals the next one. Making the total duration of the "cycle" not the time to resolve the primary bottleneck but the sum of resolution times for the sequence of bottlenecks, adjusted for whatever parallelism is achievable in capacity expansion.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

Current evidence would suggest that there's substantial seriality in this sequence. For example, power infrastructure cannot be parallelized with data center construction because the power systems must be designed for specific load profiles that depend on the compute configuration. Each dependency introduces sequencing constraints that extend the total timeline beyond what a parallel-expansion model would predict. To further expand on this point, the multi-bottleneck regime produces mini-cycles within the larger buildout as different constraints bind and release. When HBM is the binding constraint, HBM suppliers capture incremental margin while GPU suppliers face inventory accumulation. When the constraint shifts to advanced packaging, the margin capture shifts accordingly. These mini-cycles create rotational dynamics within the AI infrastructure sector that are invisible to investors using aggregate "AI demand" as their primary analytical lens. To this point, we're already seeing this rotational pattern emerge in data from last year. The first half of the year saw elevated pricing power for HBM suppliers as memory allocation constrained system shipments. The second half saw that pricing power moderate as HBM capacity expanded, while optical component suppliers experienced tightening as 800G transceiver demand exceeded production capacity. Regardless of the trend, neither shift reflected a change in aggregate AI demand, which continued to grow throughout the year. Both of which reflected the internal dynamics of constraint propagation through a complex system.

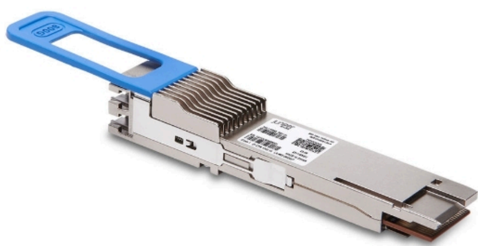
Optical Networking and Co-Packaged Optics

The networking fabric connecting GPUs within and across data centers has emerged as a binding constraint on compute cluster scale. This specific constraint operates at multiple levels being (1) the bandwidth available per link (2) the power consumed per bit transmitted (3) the physical density of connections achievable in a given form factor and (4) the reach over which high-bandwidth connections can be maintained. And each level presents their own distinct technical challenges, and the solutions we'll see being deployed this year represents a fundamental architectural transition from pluggable optical transceivers to co-packaged optics.

The Transition from Pluggables to CPO

The current paradigm for data center optical connectivity relies on pluggable transceivers which are modular units that plug into cages on the front or back panel of switches and servers, converting electrical signals to optical signals for transmission over fiber. A pluggable transceiver contains an optical engine (the components that perform electro-optical conversion), a digital signal processor (DSP) that conditions the electrical signal before conversion, and supporting circuitry for power management and thermal control. The DSP is the critical component for understanding why this architecture faces scaling limits. An electrical signal traveling from a switch ASIC or GPU to a front-panel transceiver cage must traverse 15 to 30 centimeters of copper trace on a printed circuit board. Over this distance, the signal degrades substantially due to insertion loss, crosstalk, and impedance discontinuities. The DSP's function is to recover this degraded signal through equalization, retiming, and error correction before the optical engine converts it to light. This recovery process is computationally intensive and power-hungry.

Figure 2: Breakdown of a Pluggable Transceiver



Source: Juniper Networks

In a typical 800G transceiver, the DSP accounts for approximately 50% of total module power consumption and 20-30% of the bill of materials cost. An 800G DR4 transceiver consumes roughly 16-17W, of which 6-8W is attributable to the DSP. At cluster scale, this power consumption becomes substantial. A 200k GPU cluster on a three-layer InfiniBand network requires tens of thousands of transceivers, consuming on the order of 17MW in transceiver power alone. The DSP portion of this load, approximately 8-9MW, performs no function other than compensating for the signal degradation introduced by the physical distance between the switch ASIC and the transceiver. Co-packaged optics eliminates this inefficiency by placing the optical engine on the same package substrate as the switch ASIC or GPU. The electrical signal path shrinks from tens of centimeters to tens of millimeters, reducing signal degradation to the point where DSP-based recovery is unnecessary. The optical engine can be driven directly by shorter-reach SerDes from the host ASIC, consuming substantially less power per bit.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

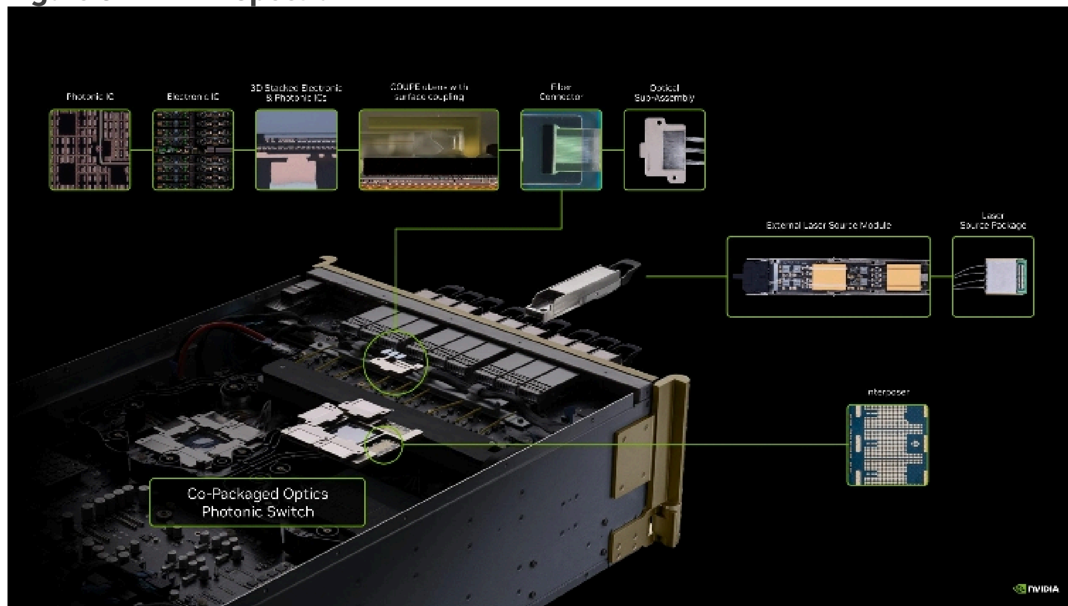
And what we'd point out is that the power savings are significant. Meta's testing of Broadcom's Bailly 51.2T CPO switch, published at ECOC 2025, demonstrated that optical engine plus external laser source power consumption is approximately 5.4W per 800G of bandwidth, compared to approximately 15W for an equivalent 800G pluggable transceiver with DSP. This represents a 65% reduction in power per bit at the optical interface. Additionally, NVIDIA's CPO implementations show similar results, with estimates of 4-5W per 800G for the optical engine and external laser source combined, representing a 70-75% reduction versus DSP-based pluggables. The total cluster-level impact is more modest because networking represents only a fraction of total cluster power. For a GB300 NVL72 cluster on a three-layer network, switching from DSP transceivers to CPO reduces total networking power by approximately 23% but reduces total cluster power by only 2-3%. The value proposition for CPO in scale-out networking (GPU-to-switch connectivity) is therefore meaningful but not transformational.

What we'd argue is that the more compelling application is scale-up networking (though scale-out networking happens first), where CPO enables capabilities that pluggable transceivers cannot achieve at any power budget. Scale-up networks connect GPUs within a coherent domain where they can share memory and coordinate at fine granularity. Current scale-up implementations, such as Nvidia's NVLink, use copper interconnects that provide high bandwidth (7.2 Tbit/s per GPU in NVLink 5.0) but are limited to approximately two meters of reach. This reach constraint limits scale-up domain size to one or two racks, which in turn limits the number of GPUs that can be interconnected in an all-to-all topology. CPO enables optical scale-up links that maintain NVLink-class bandwidth over distances of tens or hundreds of meters. This reach extension allows scale-up domains to span multiple racks or even multiple buildings, dramatically increasing the number of GPUs that can participate in a single coherent training run. The performance implications are substantial as collective communication operations that currently require traversing the slower scale-out network could instead execute over the faster scale-up fabric, reducing synchronization overhead and improving training efficiency.

Why 2026 is a Big Year for CPO

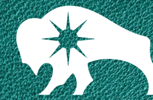
CPO has been discussed as an imminent transition for over a decade. The reason it matters specifically in 2026 is the convergence of three factors (1) products shipping in volume (2) a maturing supply chain centered on TSMC's COUPE platform and (3) accumulating reliability data that addresses customer concerns about field serviceability. NVIDIA announced two CPO-enabled scale-out switches at GTC 2025 with the Quantum X800-Q3450 for InfiniBand and the Spectrum-X 6800 for Ethernet. These are not prototypes or limited-availability products but are intended for volume deployment in production data centers. The Q3450 uses 72 optical engines at 1.6 Tbit/s each, providing 144 ports of 800G connectivity. The Spectrum 6800 offers 512 ports of 800G in its high-radix configuration. Both products integrate optical engines on the switch package substrate, eliminating front-panel transceivers for back-end network connectivity.

Figure 3: NVIDIA Spectrum X



Source: NVIDIA

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



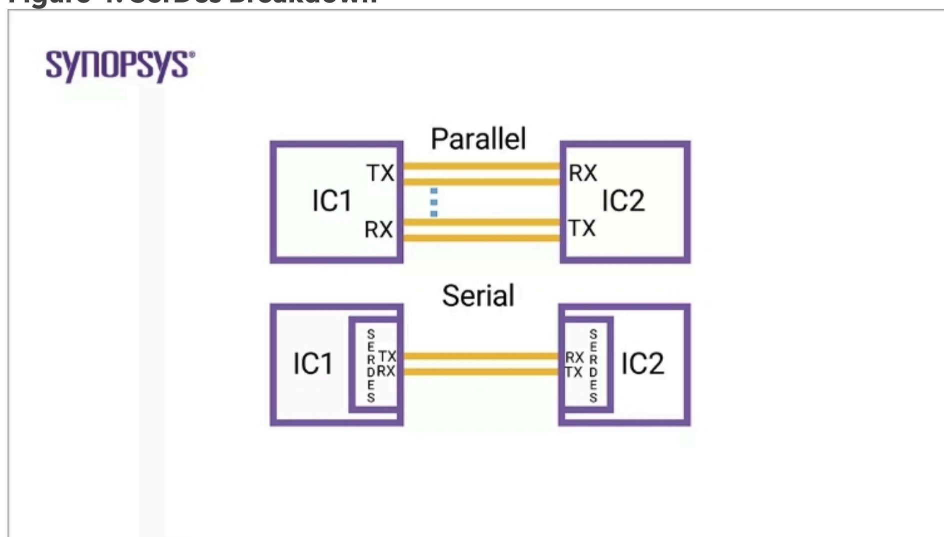
The manufacturing foundation for these products is TSMC's Compact Universal Photonic Engine (COUPE) platform. COUPE provides an integrated solution for CPO fabrication including the electrical integrated circuit (EIC) containing drivers and transimpedance amplifiers is manufactured on TSMC's N7 node, the photonic integrated circuit (PIC) containing modulators and photodetectors is manufactured on TSMC's SOI N65 node, and the two are bonded using TSMC's SoIC process, which provides a bumpless interface with minimal parasitic capacitance. The integration here is actually quite critical for performance, as the parasitic capacitance introduced by bump-based bonding limits achievable bandwidth per lane while SoIC-based bonding enables scaling to 100G per lane and beyond. Furthermore, TSMC's entry into CPO manufacturing is significant because it brings the company's established strengths in advanced logic and packaging to a domain previously served by smaller foundries with more limited capacity. NVIDIA, Broadcom, and Ayar Labs have all adopted COUPE for their CPO roadmaps, consolidating the supply chain around a single integration platform which reduces supply chain risk for customers while increasing TSMC's leverage over CPO pricing and allocation.

Reliability data is accumulating that addresses the primary customer concern about CPO which is the inability to field-replace optical components. In a pluggable architecture, a failed transceiver can be swapped by a technician in minutes, however, in a CPO architecture, a failed optical engine could render the entire switch unusable. Meta's ECOC 2025 data provides some reassurance to these problems as they stated that across 15 million 400G port-device-hours of testing (approximately 15 CPO switches operating for 11 months), the observed mean time between failure for CPO was 2.6 million device-hours, compared to 0.5-1 million device-hours for pluggable 2xFR4 transceivers. This means CPO appears to be more reliable than pluggables, not less, likely because it eliminates the mechanical connectors and contamination-prone interfaces that cause many pluggable failures. This data is helpful but not yet sufficient for broad adoption though. Fifteen switches over eleven months in a lab environment is a small sample relative to production deployments of thousands of switches in variable data center conditions. The 2026 CPO deployments from NVIDIA and Broadcom will serve partly as supply chain pipe-cleaners, generating the field reliability data that larger customers require before committing to CPO at scale. The production deployments will likely begin in scale-out networking, where the blast radius of failures is smaller, before extending to scale-up networking where reliability requirements are more stringent.

SerDes Scaling Limits

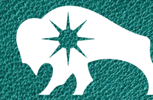
The technical pressure toward CPO is intensified by the approaching limits of electrical SerDes scaling. SerDes (serializer/deserializer) circuits convert parallel data within a chip to high-speed serial data for transmission over copper traces or cables. The bandwidth of an off-chip electrical interface is the product of the number of SerDes lanes and the data rate per lane. Increasing bandwidth therefore requires either more lanes (consuming more chip area and package pins) or faster per-lane data rates (requiring more sophisticated and power-hungry circuitry). Per-lane data rates have scaled from 25G in 2015 to 112G in 2022 to 224G in 2024-2025. NVIDIA's Blackwell architecture ships with 224G SerDes enabling NVLink 5.0's 900 GB/s bidirectional bandwidth per GPU. Broadcom has sampled 224G SerDes in its optical DSPs. This generation of SerDes represents the current state of the art for production deployment. The path to 448G SerDes is less clear though the fundamental challenge is that signal attenuation in copper traces increases with frequency. For example, a 224G signal operating at 112 Gbaud (using PAM4 modulation, which encodes two bits per symbol) experiences acceptable attenuation over short distances but requires extensive equalization for longer reaches, while a 448G signal at 224 Gbaud would experience substantially higher attenuation, potentially requiring either higher-order modulation (PAM6 or PAM8, which degrades signal-to-noise ratio) or dramatic shortening of the copper path.

Figure 4: SerDes Breakdown



Source: Synopsys

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



NVIDIA's approach for Rubin uses bidirectional SerDes to achieve 448G per physical channel. So rather than doubling the symbol rate, bidirectional SerDes transmits and receives simultaneously on the same conductor pair, effectively doubling the data rate per wire without increasing frequency. This approach works for scale-up interconnects within a rack but does not extend to longer-reach applications where signal propagation delays make bidirectional operation impractical. True 448G unidirectional SerDes, which would be required for longer-reach copper interconnects, remains a research challenge, with industry consensus that achieving 448G over meaningful distances will require either a breakthrough in equalization techniques, a transition to PAM8 modulation with associated SNR penalties, or abandonment of electrical transmission in favor of optics. The latter option is precisely what CPO provides. This SerDes scaling limit creates a strategic forcing function as companies that solve the interconnect bandwidth problem will be able to build larger and faster AI clusters while those that do not will face architectural ceilings on cluster scale. NVIDIA's continued investment in copper-based scale-up (NVLink over copper, bidirectional SerDes) reflects confidence that copper can be extended for at least one more generation, while their parallel investment in CPO for scale-out reflects recognition that optics will eventually be required across the full interconnect hierarchy.

The 800G to 1.6T Transceiver Transition

Independent of the CPO transition, the pluggable transceiver market is undergoing its own generational shift from 800G to 1.6T modules. This transition creates supply chain stress that compounds the broader networking bottleneck. An 800G transceiver typically uses four lanes at 200G each (DR4 or FR4 configuration) while a 1.6T transceiver uses eight lanes at 200G (DR8 or FR8) or four lanes at 400G (DR4-400G). The eight-lane approach is simpler technically but requires twice as many optical components per transceiver and the four-lane approach requires doubling the per-lane data rate, which demands more sophisticated modulators, drivers, and DSPs.

Figure 5: Comparison Between 800G Transceiver and 1.6T Transceiver

Source: AscentOptics

This merely continues to make the case that the supply chain is constrained at multiple points. For example, EML (electroabsorption modulated laser) production, which provides the light sources for DR4 and DR8 transceivers, is concentrated in a small number of facilities with limited expansion capacity. DSP silicon for 1.6T transceivers requires advanced process nodes (7nm or below) and competes for wafer capacity with other high-demand products. Testing and qualification throughput is rate-limited by the availability of specialized equipment that must itself be manufactured and deployed. All these constraints translate directly into cluster deployment timelines so a hyperscaler planning a new training cluster must secure not only GPU allocation but also transceiver allocation, and transceiver lead times for 1.6T modules currently extend to multiple quarters. And the reality is that the companies that locked in transceiver supply early in 2025 are likely the ones that will be deploying clusters in 2026, while those that didn't will have to wait.

Another thing that's key to mention is that the transceiver bottleneck also influences the economic case for CPO. If pluggable transceivers were abundant and cheap, the incremental complexity of CPO integration would be harder to justify. In a world where pluggables are scarce and expensive, CPO becomes more attractive even if its technical advantages were marginal. So we'd argue that the current supply situation therefore accelerates CPO adoption beyond what a pure technology comparison would predict.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

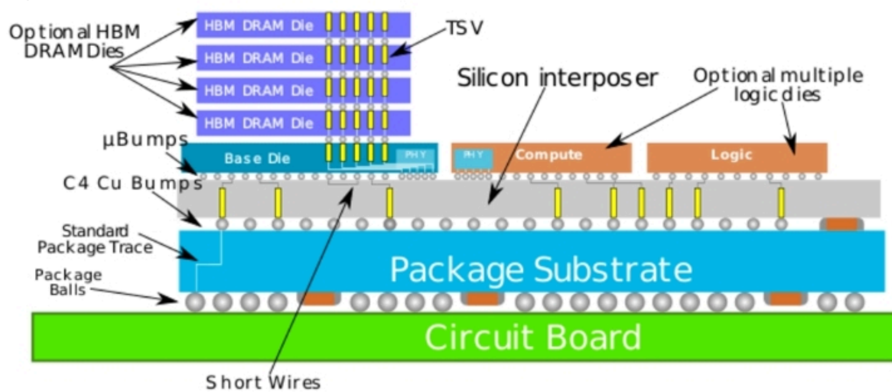
CoWoS and Advanced Packaging

The discourse around AI chip supply has focused predominantly on wafer fabrication capacity, so "how many wafers can TSMC produce at the N4 or N5 node", "how NVIDIA's allocation competes with Apple and AMD", or "does capacity expansion at Arizona or Kumamoto relieve shortages?" We'd argue that for the most part this framing is rather incomplete, as for most advanced AI accelerators, the binding constraint is not wafer fabrication itself (though this is still important) but advanced packaging.

CoWoS as a System-Level Checkpoint

Modern AI accelerators are not monolithic chips but multi-die systems integrated on a shared substrate. NVIDIA's Blackwell GB200 combines two GPU dies with eight HBM3e memory stacks on a single package. AMD's MI300X integrates multiple compute chiplets with HBM stacks on a unified base die. Google's TPU v5p and Amazon's Trainium 2 follow similar multi-die architectures. The technology enabling this integration is advanced packaging, and more specifically, TSMC's CoWoS (Chip on Wafer on Substrate), which is the dominant platform for AI accelerator packaging. CoWoS creates a silicon interposer that serves as a high-bandwidth interconnect layer between dies. The GPU or accelerator dies are placed on the interposer along with HBM memory stacks, with the interposer providing the dense wiring (thousands of interconnects per millimeter) required for high-bandwidth memory access. The interposer-plus-dies assembly is then bonded to an organic package substrate that provides power delivery and external I/O, resulting in a system-in-package that achieves memory bandwidths impossible with conventional packaging.

Figure 6: Breakdown of CoWoS Architecture



Source: GitHub

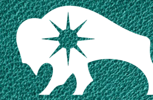
The constraint arises because CoWoS capacity scales differently than wafer fabrication capacity. Wafer fabs can increase output by adding tools and running additional shifts, and while the capital intensity is high, the expansion physics are well understood. CoWoS capacity on the other hand is constrained by the availability of specialized equipment (bonding tools, testing systems), the yield of the multi-die assembly process, and the physical throughput of operations that cannot be parallelized in the same way as wafer processing. While TSMC's CoWoS capacity has expanded substantially since 2023, expansion has consistently lagged demand regardless. The company has repeatedly doubled CoWoS capacity over the past couple of years, yet customers continue to report allocation constraints. To this point, the gap between wafer output and packaging throughput is visible in inventory dynamics, as GPU dies accumulate waiting for packaging slots while finished package units remain supply-constrained.

The Substrate and Interposer Scaling Challenge

CoWoS packaging faces technical challenges that intensify as AI accelerators grow larger. Two issues are particularly relevant, with the first being interposer size limits and the second being yield degradation from multi-die integration. Silicon interposers are fabricated using semiconductor lithography, which imposes a reticle size limit of approximately 26mm by 33mm (858 square millimeters) per exposure. Interposers larger than this limit must be created by stitching multiple exposures together, a process that introduces alignment challenges and potential defect sites at stitch boundaries. For instance, the Blackwell GB200 requires an interposer substantially larger than a single reticle, necessitating multi-exposure stitching with associated yield implications.

As interposers grow larger, mechanical stress becomes increasingly problematic. The interposer, dies, and substrate have different coefficients of thermal expansion meaning as that package heats and cools during operation and testing, these materials expand and contract at different rates. Larger packages experience greater absolute dimensional change, increasing the risk of warpage, delamination, and interconnect failure. Managing this stress requires careful co-design of materials, die placement, and thermal management, adding engineering complexity that does not scale linearly with package size. The yield mathematics of multi-die packaging compound these challenges. A CoWoS package containing *n* known-good dies has a yield ceiling equal to the product of individual die yields raised to the *nth* power, multiplied by the packaging process yield. If individual GPU dies have 95% yield and the packaging process has 90% yield, a two-die package like Blackwell has a theoretical yield ceiling of approximately 81% (0.95 × 0.95 × 0.90). Adding HBM stacks, each with their own yield, further reduces the ceiling.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



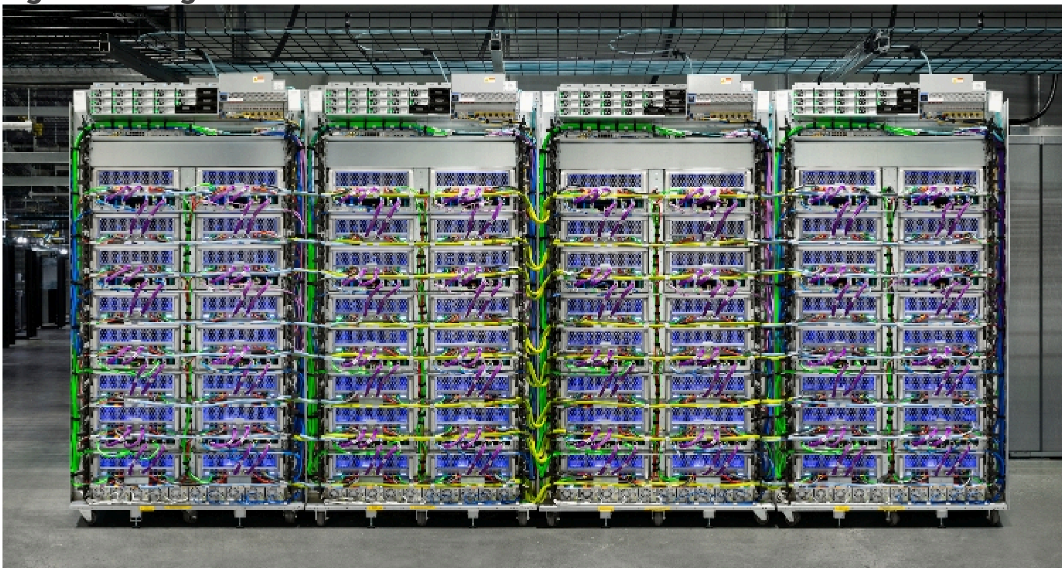
In practice, sophisticated binning and redundancy schemes recover some yield loss, but the fundamental dynamic remains which is that packages containing more dies have lower yield than packages containing fewer dies, all else equal. This yield penalty acts as a tax on the multi-die architectures that enable the highest-performance AI accelerators, making them disproportionately expensive relative to their component costs. The CPO transition discussed in the previous section intensifies these packaging challenges. Co-packaged optics requires placing optical engines on the package substrate alongside the switch ASIC or accelerator die. NVIDIA's Spectrum-X Photonics switch package for example measures 110mm by 110mm to accommodate 36 optical engines surrounding the switch ASIC, compared to Blackwell's 70mm by 76mm package. This larger substrate must maintain signal integrity for high-speed electrical connections to each optical engine while managing the thermal load of both the switch ASIC and the optical components. The engineering complexity is substantially higher than conventional packaging, and the yield implications of bonding 36 additional known-good optical engines onto each package are significant.

Implications for Custom Silicon Timelines

The advanced packaging constraint has direct implications for the timeline of custom AI silicon from hyperscalers. Including but not limited to Google's TPU, Amazon's Trainium, Microsoft's Maia, and Meta's MTIA all require advanced packaging for their highest-performance configurations. These chips compete with NVIDIA and AMD for the same TSMC CoWoS capacity. The allocation dynamics favor high-volume, high-margin customers. NVIDIA, as TSMC's largest CoWoS customer by revenue, receives priority allocation. Hyperscalers developing custom silicon face a structural disadvantage: their volumes typically are lower (many of them barring Google with the TPU are developing their chips for internal uses only), their design iterations are less frequent (reducing learning curve benefits), and their packaging requirements are often more demanding (custom configurations optimized for specific workloads rather than general-purpose designs).

Because of this, custom silicon programs have a higher risk of slipping their announced timelines. A hyperscaler announcing a new AI accelerator for deployment in 2026 is making an implicit assumption about CoWoS allocation that may not hold if NVIDIA's demand increases or if TSMC's capacity expansion encounters delays. The history of custom silicon announcements includes numerous examples of products that were technically ready but supply-constrained due to packaging limitations. This dynamic creates an information asymmetry that sophisticated investors can exploit. Announced timelines for custom silicon should be interpreted as aspirational targets conditional on packaging availability, not firm commitments. The companies with the strongest packaging partnerships (long-term agreements, prepaid capacity, co-investment in expansion) will meet their timelines; those without such partnerships will experience delays that may not be disclosed until they affect reported metrics.

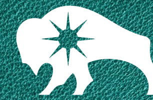
Figure 7: Google TPU v5e



Source: Google

The packaging constraint also influences architectural decisions. Some hyperscalers have opted for designs that use less advanced packaging (standard flip-chip rather than CoWoS) to avoid the capacity bottleneck, accepting lower memory bandwidth in exchange for supply certainty. Others have invested in alternative packaging approaches (Intel's EMIB, proprietary solutions) to reduce dependence on TSMC. These strategic responses are visible in product specifications and partnership announcements for those who know where to look. For 2026, we expect the packaging constraint to remain binding for the highest-performance AI accelerators. TSMC's capacity expansion will absorb much of the incremental demand from Blackwell ramp and next-generation products from AMD and hyperscalers, leaving limited slack for unexpected demand increases. The companies with secured allocation will execute on their roadmaps; those without will find their ambitions constrained by the physical throughput of bonding tools in Taichung.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



High Bandwidth Memory & NAND Flash

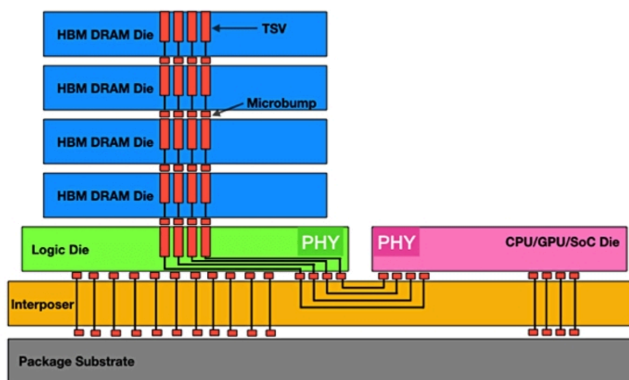
The memory system of an AI accelerator determines what workloads it can execute efficiently. Training and inference place different demands on memory capacity and bandwidth, and as models have grown larger and inference patterns have shifted toward longer contexts and reasoning-intensive workloads, the memory hierarchy has emerged as a binding constraint on AI deployment. This constraint operates at two levels (1) the HBM attached directly to accelerators and (2) the storage systems used for model weights, checkpoints, and intermediate state.

HBM as an Inference Bottleneck

High Bandwidth Memory (HBM) provides the memory capacity and bandwidth for modern AI accelerators. An NVIDIA H100 SXM for instance contains 80GB of HBM3 with approximately 3.35 TB/s of memory bandwidth while a GB200 increases this to 192GB of HBM3e with approximately 8 TB/s of bandwidth per GPU. Simply put, these specifications determine the size of models that can be served from a single accelerator and the throughput achievable for a given batch size. And the relationship between memory bandwidth and inference throughput is direct for large language models, as each token generated requires reading the model weights from memory, performing matrix multiplications, and writing intermediate activations. For a model with p parameters at b bytes per parameter, generating a single token requires approximately $p*b$ bytes from memory at minimum. The achievable tokens per second is therefore bounded above by memory bandwidth divided by model size, before accounting for compute utilization and other overheads.

For a 70B parameter model stored in FP16 (2 bytes per parameter), the minimum memory read per token is 140GB. On an H100 with 3.35 TB/s bandwidth, this implies a theoretical maximum of approximately 24 tokens per second per GPU for single-batch inference, assuming perfect memory bandwidth utilization. Practical throughput is lower due to memory access patterns, activation storage, and KV cache overhead. Larger models face proportionally tighter constraints like if we were to take a 405B parameter model like Llama 3.1 405B which requires distributing inference across multiple GPUs not because any single GPU lacks sufficient compute but because no single GPU has sufficient memory capacity or bandwidth. The transition from Hopper to Blackwell relaxes this constraint meaningfully as the combination of 2.4x higher memory capacity (192GB versus 80GB) and 2.4x higher memory bandwidth (8 TB/s versus 3.35 TB/s) enables serving larger models on fewer GPUs and achieving higher throughput for memory-bound workloads. For inference providers, this translates directly to cost per token because fewer GPUs per model instance means lower capital and operating costs per unit of inference output.

Figure 8: High Bandwidth Memory Architecture



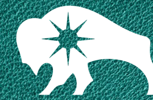
Source: Semiconductor Engineering

Another point we feel important to make is that this constraint will reassert itself at the frontier, as models designed for Blackwell-class memory systems will be larger than models designed on previous generations of chips, essentially absorbing the capacity and bandwidth gains. The memory bandwidth requirement scales with model size, and model size scales with available memory, creating a co-evolution dynamic where hardware improvements enable larger models rather than making existing models cheaper to serve. The memory constraint does not disappear; it migrates to a new equilibrium.

Yield and Capacity Dynamics

HBM production is primarily concentrated among three suppliers: SK Hynix, Micron, and Samsung. SK Hynix has maintained a consistent lead in both capacity and yield, supplying the majority of HBM for NVIDIA's data center GPUs. Samsung has faced yield challenges that have limited its qualification for high-volume NVIDIA orders, while Micron entered the HBM3e market later but has achieved qualification for select NVIDIA products and supplies HBM for other accelerators including AMD's MI300 series.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



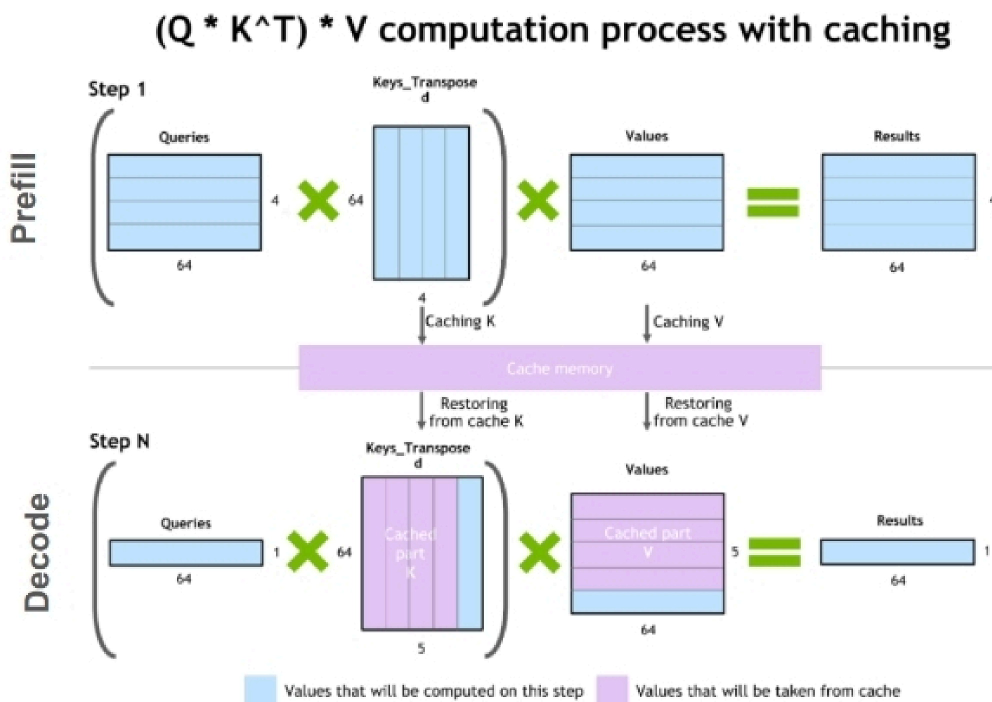
The yield dynamics of HBM are particularly challenging because HBM stacks multiple DRAM dies vertically, connected by through-silicon vias (TSVs). An HBM3e stack typically contains 8 or 12 DRAM dies, while the stack yield is the product of individual die yields multiplied by the stacking and bonding yield. If individual DRAM dies have 95% yield and the stacking process has 90% yield, an 8-high stack has a theoretical yield of approximately 66% (0.95^8 × 0.90). Higher stacks, required for higher capacity per package, face steeper yield penalties. The yield differential between suppliers translates into capacity and allocation constraints. SK Hynix's yield advantage means it can produce more functional HBM stacks per wafer start, making it the preferred supplier for volume applications. Samsung's yield challenges have resulted in limited allocation from NVIDIA despite Samsung's substantial DRAM manufacturing capacity. While Micron's position is intermediate, its yields have improved sufficiently for qualification but not for displacing SK Hynix as the primary supplier.

For AI system availability, HBM allocation can be the binding constraint independent of GPU die supply. A GB200 requires eight HBM3e stacks, so if HBM supply is constrained while GPU dies are available, the result is unshippable systems despite adequate GPU production. This dynamic has been visible in multiple quarters where Nvidia's reported data center revenue was limited by memory availability rather than GPU production. The geographic concentration of HBM production also introduces additional risk given all three major HBM suppliers have their primary production facilities in East Asia, with SK Hynix and Samsung concentrated in South Korea and Micron's HBM production primarily in Taiwan and Japan. This concentration creates supply chain vulnerability to regional disruption, a factor that hyperscalers increasingly weigh in their infrastructure planning.

KV Cache as a First-Class Problem

Beyond model weights, inference workloads must store key-value (KV) cache which is the accumulated context from previous tokens in a sequence that enables the model to attend to earlier parts of the conversation. KV cache size scales linearly with sequence length and is multiplicative across attention layers. For a transformer with *l* layers, *h* attention heads, *d* dimensions per head, and sequence length *s*, the KV cache requires storage of $2 \times l \times h \times d \times s$ values, with the factor of 2 accounting for both keys and values. For a model like Llama 3.1 70B with 80 layers, 64 heads, and 128 dimensions per head serving a 128K context window in FP16, the KV cache per sequence is approximately 2 × 80 × 64 × 128 × 128,000 × 2 bytes, or roughly 167GB. This exceeds the HBM capacity of a single H100 for the KV cache alone, before accounting for model weights or activations. Serving long-context workloads therefore requires either distributing the KV cache across multiple GPUs, compressing the cache through quantization or pruning, or offloading portions of the cache to slower storage tiers.

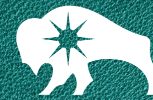
Figure 9: Explanation of KV Caching in LLMs



Source: Medium

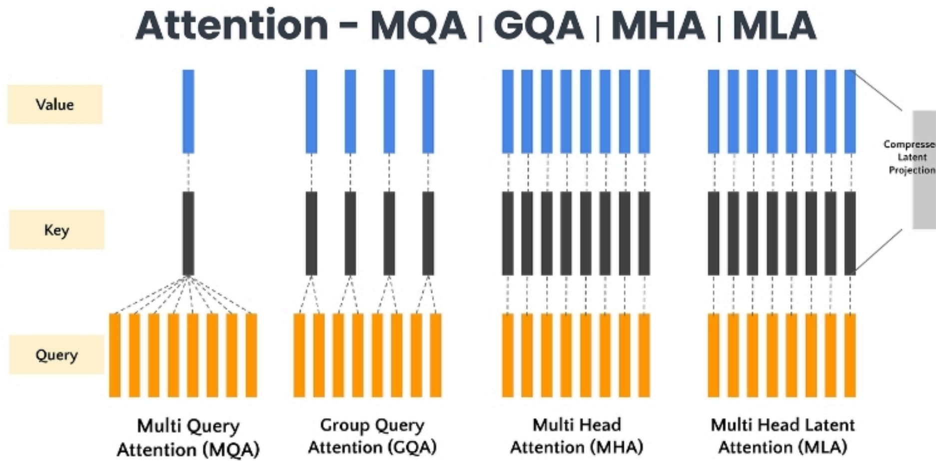
The growth of context windows from 4K tokens in early GPT-3 deployments to 128K or 1M tokens in current frontier models has transformed KV cache from an incidental overhead to a primary memory consumer. Inference providers report that for long-context workloads, KV cache memory consumption often exceeds model weight storage, inverting the traditional assumption that weights dominate memory usage. This inversion changes the optimization target as memory efficiency techniques must address cache management, not just weight compression.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



Several architectural responses have emerged. Multi-query attention (MQA) and grouped-query attention (GQA) reduce KV cache size by sharing key-value projections across multiple query heads. DeepSeek's multi-head latent attention (MLA) compresses the KV cache through learned projections. Sliding window attention limits the context each token attends to, bounding cache growth at the cost of reduced long-range dependency modeling. Each approach trades some capability for memory efficiency, and the optimal trade-off depends on the workload distribution. For inference infrastructure, KV cache dynamics affect provisioning and pricing. A server optimized for short-context, high-throughput workloads (many concurrent users with brief interactions) has different memory requirements than one optimized for long-context, low-throughput workloads (few concurrent users with extended reasoning). Inference providers increasingly differentiate pricing based on context length, reflecting the real resource cost difference rather than treating all tokens as equivalent.

Figure 10: Multi-Query Attention (MQA) vs. Grouped-Query Attention (GQA)



Source: Medium

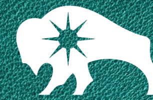
Storage Entering the Narrative

The constraints discussed above concern HBM, the fastest tier of the memory hierarchy. But AI workloads increasingly stress lower tiers as well, pulling NAND flash storage into the infrastructure bottleneck narrative. KV cache offloading represents one driver of storage demand. When KV cache exceeds HBM capacity, portions can be offloaded to NVMe SSDs and retrieved when needed. The latency penalty is substantial (microseconds for NVMe versus nanoseconds for HBM) but acceptable for workloads where the alternative is failing to serve the request at all. Offloading requires high-bandwidth, low-latency storage with sufficient write endurance to handle the continuous churn of cache eviction and retrieval.

Video generation represents a second driver. Generating video requires producing sequences of frames, each represented as a spatial grid of tokens that must be stored and processed in temporal order. A single minute of generated video at 24 frames per second with 1080p resolution can require storing and manipulating hundreds of gigabytes of intermediate activations. Training video generation models requires even larger storage for checkpoints, gradient accumulations, and dataset shards. The storage requirements scale with video length, resolution, and frame rate, creating demand growth that substantially exceeds the growth in text and image workloads. Retrieval-augmented generation (RAG) represents a third driver. RAG systems supplement model generation with retrieved passages from external knowledge bases, which must be stored, indexed, and accessed with low latency. Enterprise RAG deployments commonly involve knowledge bases of hundreds of gigabytes to terabytes, stored on SSDs for fast retrieval. The growth of RAG as a deployment pattern creates storage demand proportional to the knowledge base sizes organizations choose to index.

The flash endurance question becomes relevant under these access patterns. SSDs have finite write endurance, typically specified in drive writes per day (DWPD) over a warranty period. KV cache offloading, checkpoint storage, and video generation all involve sustained high-write workloads that stress endurance limits. Enterprise-grade SSDs with high DWPD ratings command significant price premiums over consumer-grade drives, and inference providers must factor endurance-limited drive replacement into their operating cost models. For 2026, we expect storage to transition from a commodity input to a considered constraint for specific workloads. The overall NAND market is not supply-constrained in the way that HBM is; flash memory is a mature commodity with multiple suppliers and adequate capacity. But the specific categories of storage optimized for AI workloads (high-endurance NVMe, low-latency enterprise SSDs, high-capacity drives for checkpoint storage) may face tighter supply as demand from AI deployments grows faster than these segments' traditional markets anticipated.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.


D|A|DAVIDSON

Energy and Power Generation

The constraints we've discussed thus far operate within the data center with packaging, memory, and networking determining what systems can be built and how they perform. However, energy operates upstream of all of these, as a data center cannot deploy compute it cannot power, and the availability of power at the scale required for gigawatt or even just multi-hundred megawatt clusters has become a binding constraint on where and how fast clusters can come online.

The Arithmetic of Gigawatt Scale

The power requirements for frontier AI clusters have grown by roughly an order of magnitude since 2022. A DGX A100 system (eight A100 GPUs) consumes approximately 6.5kW. A DGX H100 system (eight H100 GPUs) consumes approximately 10.2kW. A GB200 NVL72 rack (72 Blackwell GPUs in a liquid-cooled, high-density configuration) consumes approximately 120kW. The per-GPU power has increased, and the packaging density that enables high-bandwidth interconnects has concentrated that power into smaller physical footprints. A 100k GPU training cluster built on GB200 NVL72 racks requires approximately 1,400 racks, consuming roughly 168MW of IT load before accounting for cooling and facility overhead. Applying a power usage effectiveness (PUE) ratio of 1.2 (achievable with liquid cooling) yields total facility power of approximately 200MW. A cluster targeting 500k Blackwell-class GPUs requires on the order of one gigawatt of facility power. These figures assume current-generation hardware. The trend in accelerator power consumption continues upward. NVIDIA's has already announced that Rubin will have higher per-GPU power draw to support increased compute density. So each hardware generation increases the power required per unit of deployed compute, even as efficiency (FLOPS per watt) improves. The efficiency gains reduce the power required per unit of useful work, but the total work demanded grows faster than efficiency improves, resulting in net power growth.

Cooling is inseparable from power at these densities. A 120kW rack dissipates 120kW of heat in a footprint of approximately 5 square meters. Air cooling cannot remove heat at this density; the thermal gradient between chip surface and ambient air is insufficient to drive adequate convective heat transfer regardless of airflow volume. Liquid cooling is required, either through direct-to-chip cold plates (as in the GB200 NVL72) or through immersion in dielectric fluid. Liquid cooling reduces PUE by eliminating the energy spent moving large volumes of air, but it introduces its own infrastructure requirements: coolant distribution systems, heat exchangers, and often cooling towers or dry coolers sized for the full thermal load. These systems require space, capital, and lead time to deploy. A data center designed for air-cooled servers cannot be trivially retrofitted for liquid cooling at GB200 densities; the mechanical and plumbing infrastructure must be purpose-built.

The gap between announced cluster capacity and actual power availability reflects these compounding requirements. A hyperscaler can announce a 500MW AI cluster, but delivering that power to racks requires: securing a power purchase agreement or utility interconnection, building or upgrading substation capacity, installing switchgear and power distribution within the facility, deploying cooling infrastructure sized for the thermal load, and commissioning the integrated system. Each step has its own lead time, and the total timeline is the critical path through these dependencies.

Grid Interconnection vs. Behind-the-Meter

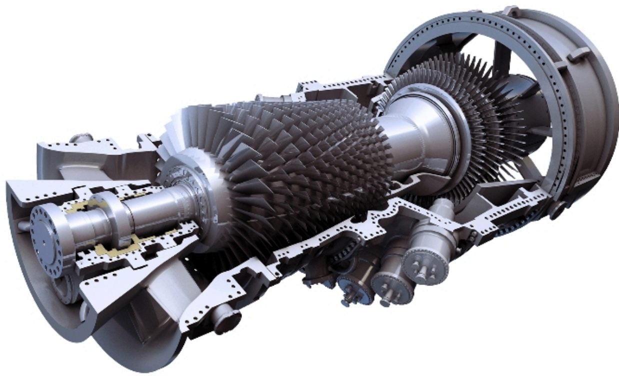
The traditional model for data center power involves grid interconnection: the facility connects to the local utility's transmission or distribution system and draws power as needed. This model faces two challenges at AI cluster scale: interconnection queue delays and transmission capacity constraints. Interconnection queues in major U.S. markets have extended to multi-year timelines. A large load seeking to connect to the PJM Interconnection (serving the mid-Atlantic and Midwest) may wait three to five years between application and energization. The queue backlog reflects both the volume of applications (driven heavily by renewable energy projects and data centers) and the studies required to assess grid impact. Each large interconnection requires analysis of how the new load affects power flows, voltage stability, and fault currents throughout the surrounding network. These studies cannot be parallelized indefinitely, and the engineering resources to conduct them are finite.

Transmission capacity constraints compound the queue delays. Even after completing the interconnection process, a data center may face curtailment if the local transmission network cannot deliver the contracted power during peak periods. Building new transmission lines requires rights-of-way, environmental review, and construction timelines measured in years to decades. The grid infrastructure in most regions was not designed for concentrated gigawatt-scale loads appearing in locations chosen for land cost and fiber connectivity rather than proximity to generation. These constraints have driven interest in behind-the-meter power: generation assets located at the data center site, connected directly to the facility load without passing through the utility grid. Behind-the-meter generation avoids interconnection queues (the data center connects to the grid at a smaller capacity for backup and supplemental power) and avoids transmission constraints (power is generated where it is consumed).

Gas turbines represent the most mature behind-the-meter option for AI datacenters. A utility-scale gas turbine installation can deliver hundreds of megawatts with a construction timeline of 18 to 24 months, substantially faster than grid interconnection in congested markets. The economics depend on natural gas prices and carbon emission costs, but in many jurisdictions, gas generation remains cost-competitive with grid power while offering faster deployment and greater supply certainty. Several announced AI clusters are being designed around behind-the-meter gas generation. The power architecture involves on-site turbines providing base load, grid interconnection providing backup and supplemental capacity, and potentially battery storage for load smoothing and peak shaving. This hybrid approach allows clusters to energize on the turbine timeline rather than the grid interconnection timeline, accelerating deployment by years in some cases.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

Figure 11: GE Vernova 9F Gas Turbines



Source: GE Vernova

Fuel cells represent an emerging alternative, particularly for facilities seeking lower carbon intensity than gas turbines while maintaining behind-the-meter deployment speed. Solid oxide fuel cells and molten carbonate fuel cells can achieve electrical efficiencies above 60%, higher than simple-cycle gas turbines, and can operate on natural gas or hydrogen. The installed base and manufacturing capacity for utility-scale fuel cells remains smaller than for gas turbines, limiting near-term deployment at gigawatt scale, but several hyperscalers have announced fuel cell installations for AI data centers. The nuclear option frequently appears in discussions of AI data center power but faces timeline constraints that exclude it from 2026 deployment relevance. Small modular reactors (SMRs) are not yet licensed for commercial deployment in the United States; the NRC licensing process requires years of review before construction can begin. Even after licensing, construction and commissioning timelines for nuclear facilities are measured in years to decades. Announcements of nuclear-powered AI data centers should be understood as statements about the 2030s, not the 2020s.

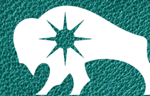
Figure 12: Bloom Energy Fuel Cells

Source: Bloom Energy

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

Implications for Cluster Geography and Timelines

Power availability has become a primary driver of site selection for frontier AI clusters, often superseding traditional factors like fiber connectivity, labor markets, and tax incentives. The question is not where a hyperscaler would like to build but where sufficient power can be delivered on the required timeline. This dynamic explains the geographic clustering of announced AI infrastructure projects. Texas, particularly the ERCOT grid region, offers faster interconnection timelines than PJM or California ISO due to its different regulatory structure and available transmission capacity. The announced Stargate project in Abilene, Texas reflects this calculus: the site offers access to power that would be unavailable on the same timeline in more traditional data center markets.


D|A|DAVIDSON

Similarly, locations with existing heavy industrial infrastructure offer advantages. Sites with retired or underutilized power plants may have transmission interconnections already in place, dramatically reducing the timeline to energize a new load. Sites adjacent to large generation facilities (hydroelectric dams, nuclear plants, industrial cogeneration) may have access to power that is not available to the broader grid. The Fairwater facility in Atlanta represents this pattern: leveraging existing infrastructure to accelerate deployment. International locations are increasingly competitive. Regions with surplus generation capacity, streamlined permitting, or state-owned utilities willing to prioritize AI infrastructure can offer timelines unavailable in the United States. The Middle East, Scandinavia, and parts of Asia have attracted AI infrastructure investment partly on the basis of power availability.

For investors, power constraints create both risk and opportunity. Announced cluster deployments should be evaluated against realistic power availability timelines, not aspirational commissioning dates. A cluster announcement without a credible power strategy (identified site, secured generation or interconnection, plausible construction timeline) is an intention, not a plan. The companies that have locked in power agreements and begun infrastructure construction will deploy on schedule; those still searching for sites or waiting in interconnection queues will experience delays that may not be disclosed until they affect financial guidance. The power constraint also creates long-duration competitive advantages. A hyperscaler that secures a gigawatt of power capacity with a long-term agreement has an asset that cannot be quickly replicated by competitors. Power agreements and site infrastructure represent committed capital that raises barriers to entry, unlike software or model weights that can be copied or approximated. The AI infrastructure leaders of the late 2020s will be partly determined by who secured power in 2024 and 2025.

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.

Copyright D.A. Davidson & Co., 2026. All rights reserved.

Potential Risks

Required Disclosures

D.A. Davidson & Co, or any of its affiliates, does or seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.

D.A. Davidson & Co. is a full service investment firm that provides both brokerage and investment banking services. Alexander Platt, the research analyst principally responsible for the preparation of this report has received and is eligible to receive compensation, including bonus compensation, based on D.A. Davidson's overall operating revenues, including revenues generated by its investment banking and institutional equities activities. D.A. Davidson & Co.'s analysts, however, are not directly compensated for involvement in specific investment banking transactions.
I, Alexander Platt, attest that (i) all the views expressed in this research report accurately reflect my personal views about the common stock of the subject company, and (ii) no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this report.

Rating Information
D.A. Davidson & Co.'s Institutional Research Rating Scale Definitions (maintained since October 10, 2017); information regarding our previous definitions is available upon request:
BUY: Expected to produce a total return of over 15% on a risk adjusted basis over the next 12-18 months
NEUTRAL: Expected to produce a total return of -15% to +15% on a risk adjusted basis over the next 12-18 months
UNDERPERFORM: Expected to lose value of over 15% on a risk adjusted basis over the next 12-18 months

Rating Distribution (as of 12/31/25)	Coverage Universe Distribution			Investment Banking Distribution		
	IR	WMR	Combined	IR	WMR	Combined
BUY (Buy)	60%	85%	63%	8%	0%	8%
NEUTRAL (Hold)	40%	13%	36%	4%	0%	3%
UNDERPERFORM (Sell)	0%	2%	1%	0%	0%	0%

IR denotes Institutional Research; WMR denotes Wealth Management Research whose rating scale is Buy/Add, Neutral, Sell/Reduce. Investment Banking Distribution denotes companies from whom D.A. Davidson & Co. has received compensation in the last 12 months. Best-of-Breed: Expected to outperform on a risk adjusted basis over a five-year time horizon.

Target prices are our Institutional Research Department's evaluation of price potential over the next 12 months, based upon our assessment of future earnings and cash flow, comparable company valuations, growth prospects and other financial criteria. Certain risks may impede achievement of these price targets including, but not limited to, broader market and macroeconomic fluctuations and unforeseen changes in the subject company's fundamentals or business trends.
While the Best-of-Breed designation does not contain a separate rating and/or price target from that of the standard ratings system referenced above, the expectation is that the security, based on the 12 criteria utilized in assessing the "Best-of-Breed" designation, will outperform over a five-year time horizon, not the standard 12-18 month time horizon.

For a copy of the most recent reports containing all required disclosure information for covered companies referenced in this report, please contact your D.A. Davidson & Co. representative or call 1-800-755-7848.

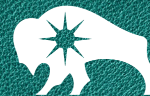
Other Disclosures

Information contained herein has been obtained by sources we consider reliable, but is not guaranteed and we are not soliciting any action based upon it. Any opinions expressed are based on our interpretation of data available to us at the time of the original publication of the report. These opinions are subject to change at any time without notice. Investors must bear in mind that inherent in investments are the risks of fluctuating prices and the uncertainties of dividends, rates of return and yield. Investors should also remember that past performance is not necessarily an indicator of future performance and D.A. Davidson & Co. makes no guarantee, express or implied, as to future performance. Investors should note this report was prepared by D.A. Davidson & Co.'s Institutional Research Department for distribution to D.A. Davidson & Co.'s institutional investor clients and assumes a certain level of investment sophistication on the part of the recipient. Readers, who are not institutional investors or other market professionals, should seek the advice of their individual investment advisor for an explanation of this report's contents, and should always seek such advisor's advice before making any investment decisions. Consensus estimates are obtained from Capital IQ. Further information and elaboration will be furnished upon request.

Other Companies Mentioned in this Report

Company Name	Ticker	Rating	Price
Apple Inc.	AAPL	NEUTRAL	$275.50
Advanced Micro Devices, Inc.	AMD	NEUTRAL	213.58
Amazon.com, Inc.	AMZN	NEUTRAL	$204.08
CoreWeave, Inc.	CRWV	BUY	$95.15
Alphabet Inc.	GOOGL	NEUTRAL	$310.96

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.



Company Name	Ticker	Rating	Price
Meta Platforms, Inc.	META	BUY	$668.69
Microsoft Corporation	MSFT	BUY	$404.37
Nebius Group N.V.	NBIS	BUY	$88.61
NVIDIA Corporation	NVDA	BUY	$190.05
Oracle Corporation	ORCL	BUY	$157.16

This report is intended for AJPlatt@dadco.com. Unauthorized distribution prohibited.