TechArena®

The Integrated Machine

How rack-scale integration is rewriting the requirements for AI infrastructure

Executive Editor: Allyson Klein
Writer: Rachel Horton
Copy editor: Amy Reifenrath
Reporting by TechArena Editorial Team:
Cinthya Anand, Rachel Horton, Deanna Oothoudt, Amy Reifenrath

September 2026

Executive Summary

As AI infrastructure continues to grow in complexity, the way it's measured has simplified to the cost and quality of the tokens it produces. Over the summer, TechArena's editorial team sat down with more than 20 companies building different components of the AI infrastructure stack to talk about the challenges they face and how to make the cost of a token predictable. A pattern surfaced in our discussions: Each layer that solves its own bottleneck hands a harder problem to the layer above, from compute to memory to the network to the interconnect, until the constraint turns physical and the integrated rack becomes the unit of design, with power, cooling, and firmware as partners to the silicon. Operators who run their infrastructure as one integrated machine can forecast their own cost per token with a confidence the stitched-together alternative cannot match. Every seam between vendors and layers hides variance someone has to absorb. Integration removes the uncertainty operators create for themselves, and silicon roadmaps, model efficiency, and demand will keep moving on schedules nobody controls, which is exactly why no operator can afford to stack self-made uncertainty on top. In this era of AI infrastructure, integration has become the quiet requirement underlying all the others.

Opening: The Unit of Value

Every currency needs a mint, and every mint has one job: strike the same coin, at the same weight, at a cost the treasury can predict. The AI economy is building its mints now. Its coin is the token, struck each time a model answers a question, writes a line of code, or completes a step in an agent's chain of work. And the mints striking it hold thousands of parts from dozens of vendors, similar to a rocket or a jetliner, with a key difference: This machine has to hit its numbers every hour, for years, at a cost promised in advance.

The AI buildout is well underway. By Dario Amodei's math, the AI industry will stand up 10 to 15 gigawatts of computing capacity this year alone at a cost near $10 billion per gigawatt. Meta is planning a single site, Hyperion, with plans to scale to 5 gigawatts over several years, on a footprint that would cover a significant part of Manhattan. Sundar Pichai said in Alphabet's Q3 2025 earnings call that Google was "processing over 1.3 quadrillion monthly tokens, more than 20x growth in a year." Commitments of this size get made years before the first token ships, so every buyer in the market is underwriting a forecast. An operator that cannot forecast its own cost per token hands that uncertainty to whoever finances the buildout, and capital charges for uncertainty.

Neocloud providers already operate by this rule: CoreWeave reported its average customer contract stretching from four years to five as its backlog grew, and long commitments at known economics are what make a gigawatt bankable. Gigawatts and dollars measure the size of the machine. Its nature is harder to hold in one view. An AI data center brings together silicon, memory, storage, network fabric, optics, power trains, and cooling loops, thousands of parts from dozens of vendors that all have to work in concert. The mint's discipline of same

coin, same weight, same cost holds only when every one of those parts does its job. So, every AI budget conversation eventually lands on the same two numbers: what a token costs to produce and whether its quality holds. The question buyers now put to every vendor, every operator, and every architecture has become: Who can make the price and quality of a token predictable over time?

The industry sells itself in superlatives: the biggest cluster, the fastest chip, the largest model. Next to those, predictability sounds like a modest goal. It is the harder one to deliver.

Lynn Comp, Intel's head of global sales and go-to-market, and vice president of its AI Center of Excellence, has watched enterprises cross from experimentation into production, and she hears the question in its rawest form.

"No CFO is happy when they are told 'I don't know how much budget I will need for AIOps at scale, nor can I guarantee I can hit that budget' by their sysadmins and IT architects," she said.

A proof of concept can tolerate surprise. A production line cannot.

The people who buy the most compute frame the goal the same way. Satya Nadella told investors on Microsoft's Q1 FY2026 earnings call that the company's fleet objective is "maximizing tokens per dollar per watt," a metric that compresses the whole infrastructure stack into one ratio. Sam Altman put the demand side on the record in early 2025: "The cost to use a given level of AI falls about 10x every 12 months, and lower prices lead to much more use." Those two statements together explain the paradox this report unpacks. Token prices keep falling, and budgets still surprise their owners, since demand, context lengths, and workload mix all grow faster than any one component improves. Falling unit prices make predictability harder to deliver, and the operators who can promise it hold the scarcest product in the market.

Diane Bryant has watched the industry rebuild itself around this kind of prize before. Her career runs from Intel, where she served as CIO and led the data center group, through Google Cloud and board seats including Broadcom. She describes the current buildout as the unwinding of the standard data center she helped create. The industry spent 20 years converging on 19-inch racks, common heights, Ethernet, PCIe, DDR, and one processor architecture. By 2016, she noted, x86 held 99.5% of server CPUs. Homogeneity kept operations simple and markets competitive, and enterprises demanded it, accepting the efficiency any general-purpose design gives up on a specific workload.

AI broke the bargain. Data center demand now concentrates in a handful of companies rather than thousands of enterprises, and those companies own their software stacks, so no processor architecture holds them.

“With scale, these companies can afford to customize their infrastructure, tuned to run a particular workload,” she said. Her math on why is stark. Working from Anthropic’s public statements on revenue and capacity, she calculates that the company generates “$30B in revenue per 1GW of data center power.” An operator who tunes infrastructure to extract more compute from that gigawatt lifts revenue and cuts operating cost at once.

“The ROIC is massive,” she said. Custom silicon, custom racks, custom networks, and custom orchestration all follow from that arithmetic.

Bryant rejects the alarm that usually attaches to this story.

“It used to be Intel that had the power-on infrastructure. It’s shifted to the hyperscalers and AI model players. It’s naive to think this is new,” she said.

“It used to be Intel that had the power-on infrastructure. It’s shifted to the hyperscalers and AI model players. It’s naive to think this is new.”

Diane Bryant
Former Intel executive

Power over infrastructure has moved before. What matters for everyone downstream is what the new holders of it now require. The operators one tier below the giants feel those requirements without hyperscale budgets to absorb mistakes. David Driggers, CEO and founder of Cirrascale Cloud Services, lives on the difference between a well-matched deployment and a wasteful one.

“Inference is forever, as we like to put it, and those costs compound,” he said.

Training is an event with an end date. Serving tokens is a permanent operating condition, and every inefficiency baked into the infrastructure repeats with every request, forever.

So the stakes are set. Tokens are the unit of value, predictability is the prize, and the customers with the most buying power are rebuilding everything in pursuit of both. What follows is a walk up the stack in the order the industry built it, one silo at a time. Every layer that solves its own bottleneck hands a harder one to the layer above. The constraint starts at the chip.

Part I. The Silo World: Components in Order

1. Compute: The Driver of AI Infrastructure Performance

A token's cost starts in silicon, and the numbers on accelerator spec sheets no longer predict it. Delivered compute now depends on process node, advanced packaging, memory bandwidth, and interconnect reach, and the gap between peak performance on a slide and sustained throughput in a rack has become the first place token economics go wrong. The interviews for this report converge on a replacement metric from three separate directions: useful work per watt.

Eddie Ramirez, vice president of marketing for Arm's infrastructure business, starts from the constraint that reframes everything else.

"AI infrastructure is increasingly constrained by fixed power, cooling, and rack density rather than demand for compute," he said.

Once the power envelope is fixed, efficiency stops being a virtue and becomes the whole game.

"Efficiency therefore isn't simply about lowering TCO, but it is about how much useful AI work can be delivered within a fixed power envelope," Ramirez added. On that accounting, a chip that wastes watts costs its owner twice, once in the power bill and again in the tokens those watts never produced.

The merchant silicon leaders have adopted the same arithmetic. At GTC 2026, NVIDIA CEO Jensen Huang put it in revenue terms: "AI factory revenues are equal to tokens per watt, so with power constraints every unused watt is revenue lost." He named the output in the same breath, declaring that "tokens are the new commodity," and committed the company to a new architecture every year to keep chasing the ratio. When the company selling the most accelerators in the world describes its own product in tokens per watt and revs it annually, the metric has finished its migration from engineering detail to industry scoreboard.

Qualcomm's entry into data center inference tests a specific bet: that two decades of thermally constrained phone design taught the company the right instincts for a power-constrained rack. "Designing for efficiency first enables us to optimize the entire rack around cost per token and power efficiency rather than chasing benchmark peaks," said Tony Pialis, executive vice president and general manager of data center. The phrase to notice is "the entire rack." Even the silicon vendors have stopped talking about chips in isolation, a shift this report returns to at its fulcrum. Manuel Botija, vice president of product at Axelera AI, applies the harshest version of the discipline. His chips go where the resources run out.

"Data center architectures are built for abundant power, abundant budget, and abundant time. None of that is available at the edge," he said.

Axelera's answer is digital in-memory computing, which "performs that math directly inside the memory cell, so the data barely has to move at all." Scarcity forced a design discipline, and Axelera now scales that discipline upward, from vision workloads to multiuser generative AI on one architecture.

"A unified architecture that handles both on the same silicon keeps cost per token stable as workloads evolve from vision to generative AI, which is the discipline that will define the platforms operators can build a business on two years from now," he said. His premise, that moving data now costs more than computing with it, is the deeper truth the next section takes up.

Buyers, meanwhile, have learned that no single accelerator wins everywhere.

"No single chip serves a 1B-parameter model and a 400B-parameter model economically, so our customers get the accelerator that fits their workload instead of bending their workload to fit whatever one vendor is selling," Driggers said. "In training, you can run a thousand GPUs and barely notice a network hiccup. In inference, that same hiccup is a failed request in front of a customer."

Matching workload to silicon is itself a lever on cost per token, and it only works when the rest of the stack can feed whatever chip wins the assignment.

One more gap defines this layer: the distance between the performance a buyer purchases and the performance a rack delivers. Peak throughput on a launch slide assumes a fed accelerator, and every section that follows in this report describes a way accelerators go hungry. Supply adds its own tax, with packaging capacity, HBM allocation, and lead times shaping the performance actually available to buy in any given quarter.

That is where the trouble begins. Push the compute ceiling higher and the chip starves. The constraint moves off the die and into memory.

2. Systemic Performance: Storage and Memory

The memory wall stopped being a research topic and became a line item.

"Model sizes are growing far faster than memory bandwidth and capacity, creating what the industry increasingly describes as the memory wall," Pialis said. He carries the point into the power budget. Memory architecture matters, he said, because "data movement increasingly consumes more energy than computation itself."

Inference made the wall taller. Randy Kreiser, field CTO at Graid Technology, draws the distinction that defines the new constraint.

"Training is limited by how fast you can process data; inference is limited by how much conversational state you can keep available," he said. Every active request carries a KV cache,

the model's working memory of a conversation, and that cache grows with context length and concurrent users rather than model size. Agentic workloads compound it, since a long-running agent accumulates state across every step it takes. "The constraint is not the GPU; it is the memory available to hold and reuse state," Kreiser said. HBM cannot rescue the economics on its own. It arrives soldered to an accelerator, and buying memory by buying compute you do not need is the waste Lynn Comp pointed to in the Opening — the one CFOs refuse to fund.

Storage vendors read the same shift as a change in their job description.

"The shift is from 'store it cheaply and reliably' to 'keep the GPU from ever waiting,'" said Jeniece Wnorowski, director of content strategy and industry expert programs at Solidigm. "Every idle GPU second caused by a stalled data loader, slow checkpointing operation, storage-network congestion, or memory bottleneck directly increases training costs and degrades economics."

The company's high-density NVMe now plays a role no storage roadmap predicted a decade ago.

"SSDs increasingly act as a memory-expansion tier, storing model weights or KV cache data that can be swapped in and out of GPU memory on demand," she said. Capacity still matters. It just stopped being the headline.

The economics of failure here are unforgiving, said Andy Pernsteiner, field CTO at VAST Data.

"Every cycle spent waiting on I/O is a cycle that's gone for good. You can't bank idle GPU time and use it later," he said. "For model builders, GPUs are the most expensive resource in their pipelines. For nearly everyone else, GPUs are the scarcest resource they have, outside of skilled staff."

Agentic workloads deepen the dependency. An agent's retrieval step rarely ends at a vector search; it usually needs the structured or unstructured source data the vector points back to. That puts the data platform in the serving path of every request.

Turning storage into a memory tier only works if the tier is fast enough, and Kreiser's benchmark numbers show how narrow the margin is. In a controlled vLLM and LMCache test on a 235 billion parameter mixture-of-experts model across four NVIDIA H200 GPUs, conventional Linux software RAID made things worse, stretching time to first token from 29.4 seconds with no

offload to 36.6. Graid's GPU-accelerated approach, which runs RAID's parallel parity math on a sliver of an installed GPU, cut it to 9.0 seconds, 3.26 times faster than no offload at all.

"The business case is simple: Fetch must beat recompute. If it does not, the offload tier works against you," Kreiser said. "Adding protection to an insufficiently fast storage path can make inference worse, not better."

Graid organizes its agentic storage portfolio by deployment scale, server to rack to platform, aligns the platform tier with NVIDIA's STX reference architecture, and plans native execution on BlueField-4 DPUs in the second half of 2026. He expects buyers to grade storage on new metrics.

"The business case is simple: Fetch must beat recompute. If it does not, the offload tier works against you."

Randy Kreiser
Graid Technology

"Over the next two years, storage will increasingly be evaluated in inference outcomes rather than raw capacity: cost per million tokens served, cache-hit rate, and time to first token," Kreiser said. "Terabytes remain necessary, but they stop being the headline metric."

The pattern of the silo walk is now visible. Memory and storage achieved their assignment, keeping the accelerator fed, and the solution multiplied the traffic between nodes. A fed node is fast. A thousand fed nodes have to talk to each other, and the constraint jumps to the fabric.

3. Systemic Performance: Network

AI clusters fail as teams, and the network is where the teamwork happens. Training and large-scale inference run collective operations that force thousands of accelerators to exchange results and wait for the slowest participant, so one congested link or one mistuned switch taxes every GPU in the job. Tail latency, an afterthought in traditional networking, becomes the governing statistic. Operators respond the expensive way, overprovisioning bandwidth they will rarely use to insure against the percentile that ruins the job. The relevant number is job completion time, and it belongs to the fabric.

The technology contest at this layer has settled into a familiar shape. InfiniBand held the early AI clusters on the strength of its latency discipline, and purpose-built scale-up fabrics hold the frontier. Ethernet absorbs capability from both, and its ecosystem and economics keep widening its share of new deployments. Buyers push the market toward open standards for the same reason they resist single-vendor racks: A fabric commitment outlasts several generations of the chips it connects.

Brandon Draeger, chief marketing officer at Cornelis Networks, makes the case for purpose-built over general-purpose.

“The network is not simply moving bits; it is determining how efficiently the entire machine works,” he said.

Draeger puts a number on what a congested fabric costs: A sustained 1% utilization improvement across 10,000 GPUs, at an assumed cost of $4 per GPU-hour, works out to roughly $3.5 million in annualized compute capacity.

“At scale, a small fabric inefficiency becomes a very large infrastructure bill,” he said.

Marc Austin, CEO and co-founder of Hedgehog, built a company on the observation that most operators cannot spend what hyperscalers spend to get job completion time right.

“The network defends it or destroys it,” he said. Hedgehog packages the hyperscaler playbook, open networking and cloud-style abstractions, into software an ordinary operator can run.

“It turns the network from a monthslong engineering project into a product you deploy,” he said. “Any one misconfigured switch quietly taxes every job on the cluster.”

Aanchal Sharma, senior director of product management at Astera Labs, prices the same failure in the report’s currency.

“Put the fastest chip in the world behind a slow, high-latency fabric, and you’ve built an expensive space heater, with GPUs sitting idle waiting on data instead of generating tokens,” she said. Astera’s connectivity silicon spans the three directions the industry now scales in, up within the rack, out across the cluster, and across sites, and Sharma’s point holds at every one of them.

“Put the fastest chip in the world behind a slow, high-latency fabric, and you’ve built an expensive space heater, with GPUs sitting idle waiting on data instead of generating tokens.”

Aanchal Sharma
Astera Labs

“Faster chips buy more compute. They don’t buy less waiting,” she said.

“Tokens per watt and tokens per dollar get decided in the fabric long before anyone reads a chip’s spec sheet,” she added.

The fabric is also where the industry’s standards politics now play out. Bryant catalogs the fragmentation: The single Ethernet world

has split into scale-up variants, with UALink, SUE, and ESUN advancing alongside NVLink. Sharma makes the case for keeping those interfaces open.

“An open fabric means an operator never has to bet the entire rack on one company’s roadmap,” she said.

The network carries one more assignment in the integrated machine, and it is the one the arc of this report depends on: composability. Infrastructure that gets pooled and reassembled around a workload, rather than fixed at the moment of purchase, needs a fabric that can redraw the machine’s boundaries in software. Austin sees a solved problem waiting to be borrowed.

“Composability is fundamentally a network abstraction problem, and hyperscalers already showed the answer: open networking plus VPC abstractions,” he said.

Operators learned in the cloud era to treat compute as fungible. AI infrastructure brings the same lesson one layer down, with the fabric as the instrument.

Every fix at this layer raises the load on the physical medium underneath it. Faster fabrics push more bits through copper that must carry them farther, at a higher density, and on a shrinking power budget. Copper is running out of room, and the constraint drops into the interconnect itself.

4. The Interconnect: Optical and the Limit of Copper

As data rates climb, electrical signaling loses reach, bandwidth density, and power efficiency at the same time, and the industry’s answer is to move light closer to the silicon until it arrives inside the package. Co-packaged optics is the structural version of that answer, and it changes what the interconnect is: a component of the chip rather than a cable between boxes.

Vishal Chandrasekar, director of product management at Ayar Labs, states the goal in system terms.

“They are trying to connect thousands of accelerators so they can operate as a single unified system, with the bandwidth and latency needed to support increasingly large AI models,” he said of the operators driving demand. Copper forces a choice between bandwidth and distance at the moment AI needs both.

“CPO removes that tradeoff by using light to extend high-bandwidth, low-latency connectivity across tens of meters,” he said, citing up to 10 times higher bandwidth, 10 times lower latency, and three to five times better power efficiency than copper and pluggable alternatives. “A successful demonstration matters, but customers also need confidence in reliability, supply, packaging, fiber attachment, thermal performance, and production yield.”

Under the optics sits a materials problem, and Robert Blum, senior vice president of sales and marketing at Lightwave Logic, works at the layer where it gets solved. The materials that carried the industry to 200 gigabits per lane, improved III-V compounds and silicon photonics, are reaching their ceiling.

“New materials are required for the next modulator generation where 400 Gbps speeds are needed,” Blum said. His company’s electro-optic polymers compete for that generation.

“EO polymers have really improved in performance and reliability, thanks in part to the lessons learned from the OLED industry, and are now ready for deployment,” he said.

“EO polymers have really improved in performance and reliability, thanks in part to the lessons learned from the OLED industry, and are now ready for deployment.”

Robert Blum
Lightwave Logic

Foundries favor the polymers, Blum said, since they integrate into standard silicon photonics processes more easily than lithium niobate.

Tying this back to the token, moving a bit costs energy, and the interconnect moves more bits than any other layer. Optics attacks the joules per bit directly, and every picojoule saved in transit returns to the power budget as compute.

Co-packaging also rewrites the manufacturing contract between the optics and the silicon they serve, and Blum’s description of the change is the integration thesis in miniature. Optical assemblies for pluggable transceivers tolerate standard solder reflow and wire bonding.

“For CPO, you need to integrate much more tightly with the switch ASIC, CPU, or GPU,” Blum said, which pulls in hybrid bonding, higher processing temperatures, and more complex assemblies. “All this requires much closer collaboration with system integrators.”

A component that once shipped in a box now gets engineered alongside the chip it feeds.

“Optics tends to be more complicated than copper,” he said, with high-fiber-count detachable connectors still a bottleneck and optical engine form factors not yet standardized.

What remains is the least glamorous part of any technology transition.

“The real race is about ramping production capacity and getting the 400G ecosystem in place,” Blum said. Lightwave Logic has five Fortune Global 500 customers engaged in prototype testing and targets high-volume production in 2027, with the remaining gates spread across foundry

process maturity, back-end qualification, and an ecosystem of DSPs, SerDes, connectors, and laser sources that all have to arrive together.

That phrase, arrive together, is the hinge of this report. Optics solves the last movement problem inside and between racks, and in doing so, it removes the excuse every silo had for engineering alone. Once the parts can move data as one machine, someone has to build them as one machine. The problem stops being electrical. It becomes physical.

Part II. The Turn

5. Rack-Scale: The Unit of Compute Becomes the Building Block

Every voice in Part I described a different layer and ended in the same place.

“AI infrastructure is increasingly a systems problem rather than a chip problem,” Qualcomm EVP Pialis said.

The unit of compute is no longer the chip or the server. It is the rack, designed, sold, and bought as a single product, with the row not far behind.

“AI infrastructure performs best when power, cooling, controls, and software are designed as one integrated technology system instead of assembled piece by piece,” said Steven Carlini, chief advocate for AI and data centers at Schneider Electric.

Operators using the catalog model, picking servers from one vendor, power from a second, cooling from a third, and management software from a fourth, could once assume the interfaces between those purchases were forgiving. At AI densities, no interface is forgiving. A megawatt rack punishes every assumption its designers did not share.

Adam Morton, CTO of data center infrastructure at Flex, makes the same case from the manufacturing side. “Designed as a system from the outset, [integrated AI racks] are generally more efficient, cost-effective, and scalable than racks that come together as a collection of components,” he said.

“Customized data centers have been the industry standard for decades. That era is coming to an end,” Morton said. “‘Snowflake’ projects with bespoke designs, supply chains, permitting paths, and commissioning plans

> “‘Snowflake’ projects with bespoke designs, supply chains, permitting paths, and commissioning plans don’t scale, and certainly not at an accelerated pace. … The future AI factory will be designed once and repeated many times.”
>
> **Adam Morton**
> Flex

don't scale, and certainly not at an accelerated pace. … The future AI factory will be designed once and repeated many times."

Diane Bryant tracks the same turn from the buyer's side.

"Racks are all now custom built to the dimensions that optimize power and cooling for the custom xPUs," she said. The 19-inch rack survived every previous platform shift in computing. It did not survive this one. When the companies with the deepest pockets abandon the most durable standard in the data center to win thermal and power headroom, the message to the rest of the market is unambiguous.

The merchant vendors heard it. AMD CEO Lisa Su highlighted the company's Helios platform during an earnings call with investors, describing it as an integrated rack-scale solution featuring Instinct MI400-series GPUs, Venice EPYC CPUs, and Pensando NICs. Built on Meta's double-wide Open Rack Wide standard, Helios was designed and "optimized for the performance, power, cooling, and serviceability required for the next generation of AI infrastructure," Su said. She reported "a lot of interest in the full rack-scale solution."

NVIDIA sells its flagship as a rack. The neoclouds pour the same logic into concrete: Chase Lochmiller, CEO of Crusoe, calls the data center "the new unit of compute." Each phrasing moves the boundary of the product outward, from the chip to the rack to the building.

Moving the boundary moves the accountability with it, and the industry has not finished deciding who holds it. A rack that spans one vendor's silicon, a second vendor's power train, a third's cooling loop, and an integrator's assembly has to answer an old question at a new scale: When the machine underdelivers, whose machine is it? Our interviews suggest the market is answering with engineering rather than contracts. Thermal, power, and management vendors describe co-design relationships that start at the silicon roadmap, years before a purchase order. The alternative is discovering incompatibility at commissioning, when every idle day burns the most expensive depreciation schedule in the industry.

Component vendors now design for that boundary or design themselves out of the market. Randy Kreiser of Graid sees it from the storage layer.

"Once the rack is the unit of deployment, storage cannot be treated as a component to integrate afterward," he said. Graid organizes its roadmap by deployment scale rather than by SKU for exactly that reason, and some version of that reorganization appears in nearly every interview in this report.

Rack-scale architecture also delivers the promise the cloud made and AI briefly broke: composability. A machine designed as one system can be pooled, partitioned, and reassembled around a workload in software rather than fixed at the moment of purchase. The fabric abstractions that Hedgehog's Marc Austin described in the network section are the mechanism. The rack designed as a product is the precondition.

“The technology is converging faster than most organizations are,” said Lakecia Gunter, former global CTO and corporate board director. “Companies cannot operate integrated AI infrastructure through disconnected technology, facilities, finance, security, and sustainability teams. The operating model must evolve with the architecture.”

Treating the rack as one product also knocks down a wall inside the operator’s own organization. The teams who run servers and the teams who run power and cooling grew up in different professions, with different tools, different budget lines, and different definitions of an emergency. A rack designed as one machine makes them tenders of one converged system. Facilities engineers now read GPU telemetry, and IT architects sit in utility interconnection meetings. The hardware merged first; the people are catching up. The chip and the facility used to be separate conversations. From here on, they are the same one, and the rest of this report walks the layers those merged teams share. The first is the one gating the entire industry: power.

Part III. The Integrated Machine: Systems

6. Power as a Design Partner

Power used to enter the conversation after the IT was specified. It now opens the conversation, and often ends it. “Today, speed is driven by time to power,” Carlini said, and the sentence explains more of the current market than any chip roadmap. Utility interconnection queues, substation lead times, and grid capacity now gate AI deployment harder than silicon supply. Satya Nadella made the same point from the buyer’s chair.

“You may actually have a bunch of chips sitting in inventory that I can’t plug in. In fact, that is my problem today,” he said. The scarcest input to an AI factory today is the energized shell around the accelerators.

Operators are answering the queue with every tool that shortens it. Some buy their way into existing capacity, and some build behind the meter. The energy-first developers invert the old site-selection logic, putting the data center where the power is and running the fiber to it. Crusoe built its business on that inversion. Whichever way the shell gets energized, the arithmetic that follows is unforgiving: An unpowered rack has no cost per token, since it has no tokens, and a partially powered facility pays full capital cost for partial revenue.

The premium on an energized megawatt now shows up on income statements. Nebius founder and CEO Arkady Volozh told analysts in August 2026 that the company’s pipeline makes it one of just a few players able to build more than a gigawatt of new capacity a year, and the unit economics reported from that call, tens of millions of dollars in annual revenue per megawatt, make each stranded watt a measurable loss. Carlini frames the resulting agenda in exactly those terms.

"The next challenge is not only securing more power; it is reducing the time it takes to bring that power into service and making every available kilowatt work harder to generate tokens and intelligence," he said.

Scarcity of that order forces discipline on every watt that does arrive, and the discipline starts inside the rack. Steve Thorne, chief commercial officer at CelLink, makes power delivery a peer of the components it feeds.

"Designing at rack scale forces power delivery to be considered as a first-class constraint alongside compute and cooling, rather than an afterthought bolted on at the end," he said. "At each step down in voltage, current rises sharply and I^2R losses compound accordingly."

"The next challenge is not only securing more power; it is reducing the time it takes to bring that power into service and making every available kilowatt work harder to generate tokens and intelligence."

Steven Carlini
Schneider Electric

CelLink's flat, flexible harnesses attack those losses and reclaim tray space, and Thorne's roadmap points at the same convergence Carlini described, power delivery "co-designed and co-assembled with liquid cooling cold plates as a single unified subsystem."

Flex's Adam Morton ties the shift to a specific electrical transition. "Within 800 VDC environments, electrical, mechanical, and thermal management decisions are increasingly interdependent, even as power, cooling, and IT equipment gets disaggregated to accommodate more computing capacity in the rack," he said. "A co-designed approach to the rack, power infrastructure, and cooling system maximizes performance."

AI also changed the shape of demand, not just its size. Brandon Smith, vice president of global sales and product management at ZincFive, describes a load profile no facility engineer trained for.

"An AI workload can swing from idle to full draw and back in milliseconds, then repeat it thousands of times an hour," he said. Batteries designed to sit idle until a blackout solve a problem different from the one Smith describes.

"Rather than sitting idle until the grid fails, the system absorbs the spikes and releases energy as the workload calls for it, shaping load in real time before it travels through the facility and out to the grid. Power moves from passive backup to active stabilization," Smith said. ZincFive's case for nickel-zinc chemistry rests on that repetitive, high-rate duty cycle.

"A site that smooths its own demand internally reads as a better neighbor, and that can shape how much capacity gets allocated and how quickly a project breaks ground," he said. In a market

gated by time to power, good citizenship at the meter converts into schedule, and schedule converts into tokens.

Every watt the power team wins arrives in the rack with an obligation attached: It all comes back out as heat.

7. Cooling as a Design Partner

Every watt that enters a rack leaves it as heat, and operators now plan for heat with the same care they give power.

“Cooling is now a foundational design decision for AI infrastructure,” said Rich Whitmore, president and CEO of Motivair by Schneider Electric.

“As rack power increases, the challenge shifts from simply removing heat to delivering reliable, efficient, and scalable thermal management across an entire facility.”

The industry crossed the density line where air alone stops working, and the liquid systems replacing it cannot be bolted on after the fact. Scale changed the job, too. Whitmore sets the bar for production readiness in operational terms.

“Production environments require more than excellent thermal performance. They require repeatability, uptime, ease of service, and seamless integration with the rest of the infrastructure,” he said.

Paul Quigley, chief strategic relations officer at AIRSYS, connects cooling to the constraint running through this whole report: the power budget. His company has spent three decades in precision air cooling and now builds hybrid systems that add liquid where density demands it, which gives him an unusual vantage on how the two share the load.

“Power has become a primary constraint, so every infrastructure decision comes back to how much provisioned capacity can ultimately be allocated to compute,” he said. “The first question should not always be, ‘How do I get more power?’ It should also be, ‘Am I making the best use of the power already provisioned?’”

“The first question should not always be, ‘How do I get more power?’ It should also be, ‘Am I making the best use of the power already provisioned?’”

Paul Quigley
AIRSYS

AIRSYS proposes measuring exactly that with a planning metric it calls Power Compute Effectiveness, which evaluates how provisioned electrical capacity gets structurally allocated at design time, a complement to PUE's operational lens. Quigley expects liquid to carry a growing share of the thermal load and air to remain necessary for portions of the IT load, which makes the design question one of proportion.

Ryan Brown, director of data center product management at Phononic, pushes the thermal case down to the component.

"As AI infrastructure becomes more tightly integrated, performance is increasingly determined by the weakest thermal link in the system," he said.

The weakest link has moved past the GPU under its cold plate. Optical transceivers, memory, and NICs now throttle systems whose headline silicon is perfectly cooled, and Phononic's solid-state thermoelectric coolers target those hotspots directly. Precision at the component pays off at the facility.

"By actively and dynamically controlling temperature at the component level, operators can maintain tighter thermal margins on the devices that matter most, even as rack densities continue to climb," Brown said.

Tighter margins let operators run warmer coolant and cut the cooling headroom they overprovision out of fear.

"Thermal intelligence turns cooling from a supporting utility into an operational tool that helps improve reliability, efficiency, and asset management," Brown added. A cooling system that senses and reports becomes part of the machine's telemetry, the subject of the next section. One more wall between operations and IT comes down.

"Thermal intelligence turns cooling from a supporting utility into an operational tool that helps improve reliability, efficiency, and asset management."

Ryan Brown
Phononic

Simon Jesenko, CEO and CFO of Iceotope, closes the spectrum at the whole board. The company's precision liquid cooling seals the entire server and delivers dielectric fluid exactly where heat appears.

"Cooling the whole server with dielectric fluid allows the operator to cool all heat-generating components within the server: GPUs, CPUs, memory, networking, and PSUs," he said. His

efficiency case rests on first principles. Every heat exchange between fluids gives up energy, so, in his words, “the most efficient system is one which has fewer (or no) heat exchangers.” He also treats the heat itself as inventory rather than exhaust.

“Captured heat can be reused in other deployments. For example, a precision liquid-cooled data center in a hotel basement could use the excess heat to heat the hotel pool,” Jesenko said.

“Captured heat can be reused in other deployments. For example, a precision liquid-cooled data center in a hotel basement could use the excess heat to heat the hotel pool.”

Simon Jesenko
Iceotope

Two pressures from outside the machine now shape every choice inside it. Buyers apply a sustainability lens with real procurement weight. They ask about water consumption, energy reuse, and the carbon attached to each token, and liquid systems that run warmer coolant, reject heat efficiently, or hand it to someone who can use it perform better on all three counts than air. The installed base pushes back from the other side. Most of the world’s data center square footage was built for air, so operators choose between retrofitting live facilities toward liquid and reserving the densest AI for greenfield builds.

A rack full of power and coolant is still a body without a nervous system. Someone has to make the machine knowable.

8. The Control Plane: Firmware, Telemetry, Software

Integration remains an aspiration until something operates it, and the layer that does the operating spent decades beneath anyone’s notice. Firmware booted the server, reported its health, and stayed out of sight. AI ended that obscurity, said Colin Brix, vice president of marketing at AMI, a Lattice Company. Operators now demand granular telemetry on GPU performance, power draw, thermal behavior, and interconnect health, feeding orchestration systems that act on it automatically.

“Firmware is no longer just reporting health status. It’s becoming the trusted source of telemetry that drives automated decisions around workload placement, power balancing, cooling optimization, and predictive maintenance,” Brix said. “In short, firmware is evolving from a management layer into a data and control layer for the AI factory.”

The stakes of getting that layer wrong scale with the cluster. In a traditional environment, a misconfigured node inconveniences a workload. In an AI factory, Brix said, “a node that is misconfigured, running an inconsistent firmware level, or failing attestation checks can prevent thousands of GPUs from operating at full efficiency.” The boot-to-workload chain, invisible when it works, becomes the foundation of the fleet’s economics. Secure provisioning, firmware integrity validation, and hardware attestation moved from compliance checkboxes to operating requirements.

Automation is only as good as the telemetry beneath it.

“At AI factory scale, telemetry must be accurate, consistent, and synchronized,” Brix noted, since operators now correlate events across thousands of systems at once, and decisions about workload placement, power allocation, and thermal management are only as good as the data feeding them. Heterogeneity works against all three properties. A fleet whose compute, power, cooling, and networking each expose different interfaces and telemetry formats makes automation fragile and lets operational overhead grow with every added vendor.

At fleet scale, the case for unification stops being aesthetic. AMI’s MegaRAC OneTree pulls management of compute, power, cooling, and networking into one open codebase across different silicon, so a patch propagates once instead of dozens of times and every subsystem speaks the same telemetry language. The endgame is a fleet that runs itself.

“The reality is that nobody manually operates a 50,000-node AI factory,” Brix said. “The telemetry and automation architecture must be designed so the fleet effectively manages itself, with humans focusing on policy and optimization rather than individual device administration.”

Predictable cost per token, Brix said, “requires three things: trusted data, automated optimization, and strong security.” The security leg reaches all the way down to power-on.

“The control plane must be built on a common hardware root of trust that validates systems from initial power-on through workload execution,” he said, so an operator knows every node in the fleet runs authorized firmware, trusted software, and verified hardware before it earns a workload.

“Nobody manually operates a 50,000-node AI factory. The telemetry and automation architecture must be designed so the fleet effectively manages itself, with humans focusing on policy and optimization.”

Colin Brix
AMI, a Lattice Company

The market has started pricing this layer accordingly. Lattice Semiconductor closed its acquisition of AMI on July 27, 2026, in a deal trade press valued at $1.65 billion. A sum that size for a firmware company measures what control of the control plane is now worth. Eddie Ramirez of Arm sees the same promotion happening in silicon, where “the CPU becomes the control plane for the entire rack, coordinating data movement, scheduling work, feeding accelerators efficiently, and ensuring entire system resources are fully utilized.”

A fleet an operator can see is a fleet a customer, a regulator, or a board can ask hard questions about, and those questions are the final layer of the machine.

“Boards should not view AI infrastructure as a technical procurement decision,” Gunter said. “It is a capital allocation, resilience, and competitive-positioning decision that will shape how quickly the company can turn AI ambition into enterprise value.”

9. Privacy and Policy: The Outermost Layer of Trust

The last requirement is the one no rack diagram shows. A token’s price means nothing if its buyer cannot trust where it was made, what data fed it, and who can be held accountable for it. Lynn Comp, whose vantage at Intel spans the enterprises crossing into production, watches trust move from the legal department to the architecture review.

“I’m seeing harder questions being asked about private and hybrid AI as the true costs of frontier models hosted in hyperscaler data centers become more obvious,” she said. Cost and control turn out to be the same question, and buyers have begun answering it with deployment decisions rather than policy memos.

The economics run deeper than compliance, since the data that privacy rules protect is the same data that makes AI worth buying.

“The AI is generic until it applies your business information and context, but it is difficult to see ROI from AI that spends the majority of tokens on the data input processes rather than getting insights from the underlying data,” Comp said. An enterprise’s proprietary data is its entire differentiation in AI. Where that data can safely go becomes an infrastructure requirement rather than a legal afterthought.

Diane Bryant maps the same boundary from the CIO’s chair she once occupied at Intel. In her assessment, an enterprise CIO today should move everything possible to the cloud, since no internal operation competes with hyperscale efficiency, and keep on premises only the workloads with compliance or privacy concerns. The exceptions define the territory. Sovereignty rules, data residency requirements, and sector regulations now shape where tokens get produced as surely as grid capacity does, and they are the reason private and hybrid AI keeps its seat at a table the economics would otherwise clear.

Purpose-built machinery for those exceptions is arriving. Confidential computing has moved from research program to procurement checkbox. Buyers ask whether inference can run without exposing plaintext the model should never see. The hardware root of trust Brix described in the control plane section turns out to be the foundation for every trust claim above it: The same attestation that tells an operator a node runs authorized firmware tells a regulator, a customer, or a court where a token came from and what touched it along the way. Integration pays a dividend here too. A machine designed as one system can answer for itself as one system. Agentic AI raises the stakes on all of it, and Comp's recommendations are design constraints rather than aspirations.

"Infrastructure needs to have very tight design constraints for a given AI-based agent, avoiding the ability of an agent to exercise functionality beyond its intended purpose," she said. "Any decision that would be difficult to defend in a current regulatory framework needs to have a human in the loop."

Dana Bos, TechArena co-founding advisor and founder and principal of Bos Solutions, sees that gap as the one most AI stack diagrams don't show.

"Every AI stack diagram stops at the application layer, but the layer that actually determines ROI sits above it: the humans deciding whether to believe the output. Skip that layer and you've built a very expensive system nobody uses correctly," she said.

Trust, in Comp's telling, is an engineering deliverable with an audit trail, and it pays like one. A token a customer can trust, produced where the rules require, from data that never left its owner's control, commands a premium over an identical token that cannot show its provenance. Trust is the only layer of the machine the buyer experiences directly, and the whole integrated machine underneath exists to earn it.

This report has traced a single constraint from the chip through memory, fabric, and light, into the physical realities of the rack, and up through power, heat, and telemetry to the question of trust. One question remains, the one this report opened with.

Close: The Requirement

Who can make the price and quality of a token predictable over time?

The operators who run their infrastructure as one integrated machine, and the vendors who design for that machine, hold the pricing power. The ones still stitching a machine together from catalog parts inherit every seam as a variance they cannot forecast. Theirs are mints that strike a different coin every day. The market pays for the mint that strikes the same one.

The evidence assembled here points one direction. Satya Nadella compresses the hyperscaler's entire strategy into tokens per dollar per watt. Steven Carlini's power, cooling, controls, and software perform best "designed as one integrated technology system instead of assembled

piece by piece." Aanchal Sharma's tokens get decided in the fabric, Randy Kreiser's in the cache tier, Brandon Smith's in the milliseconds between power spikes, Ryan Brown's at the hottest component on the board.

"Thinking of agent design like building an appliance helps keep the architecture clean, which then has the byproduct of being more deterministic and predictable," Comp noted.

"Thinking of agent design like building an appliance helps keep the architecture clean, which then has the byproduct of being more deterministic and predictable."

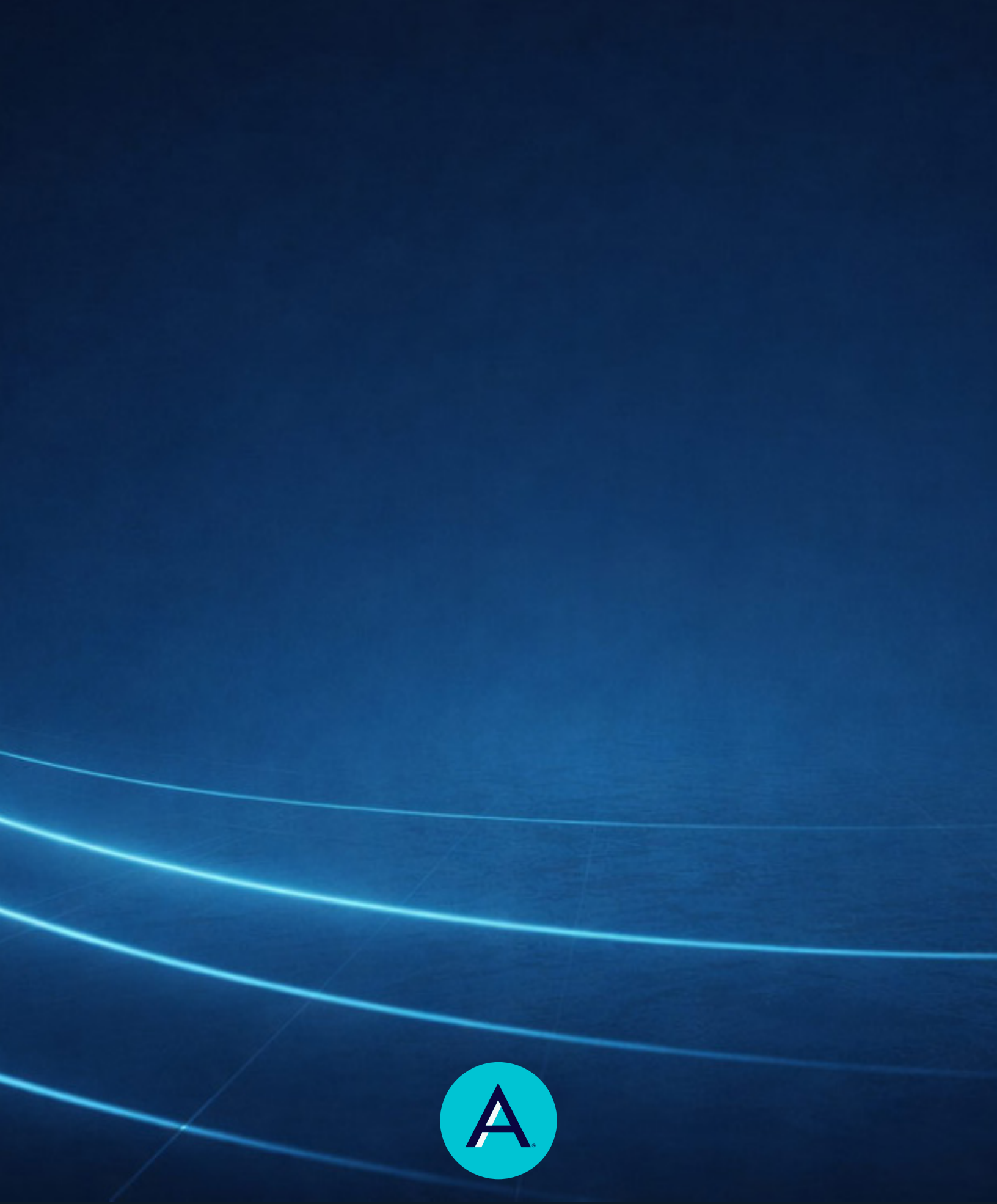
Lynn Comp
Intel

An appliance is the humblest thing engineering produces, and the most trusted. It does what it says, every time, at a cost its owner can state. The AI industry, for all its scale and speed, is working its way toward the appliance's virtue.

The constraint will keep moving. The voices in this report point to where it travels next: trust and energy, the two inputs no engineering team can manufacture alone. The next year offers clean markers for anyone tracking the turn. Watch whether the scale-up fabric standards consolidate or keep splintering, and whether co-packaged optics crosses from prototype to volume on the timelines vendors like Lightwave Logic have staked. Watch how fast the firmware layer consolidates now that Lattice has priced it, and whether time to power shortens or keeps stretching. Each is a proxy for the same underlying question: how quickly the industry finishes becoming one machine.

What will not change is the shape of the winning response, which every section here has traced. Watch the constraint, follow it across the old silo walls, and design the next layer with the last one instead of after it.

Everything this report has argued comes down to a single payoff: the determinacy dividend. It accrues to the operators who treat integration as the requirement underneath all the others, who co-design power with silicon and cooling with both, who can see every node and trust what they see. Those operators will quote a customer a price per token two years out and hit it, and they will collect the capital, the customers, and the trust that follow. In this era of AI infrastructure, integration is the requirement. Everything else is a line item.



TechArena 2026