RCR Tech
NEWS. INSIGHT. IMPACT.

MARKET PULSE REPORT

THE AI INFRASTRUCTURE PLATFORM SHIFT IS HERE

BY SEAN KINNEY, PRINCIPAL ANALYST, RCRTECH

Contents

Introduction

Artificial intelligence (AI) is entering a new phase. What began as an era defined by model breakthroughs and hyperscale experimentation has become an exercise in systems engineering. The limiting factors aren't algorithms or access to GPUs. The limiting factors are organizational, operational, physical and economic. Power, networking, data readiness, governance and geography are shaping what AI can realistically deliver and at what cost.

This report starts from a simple premise: AI advantage is not determined by who builds the most intelligence. Advantage is determined by who can operate — who can deliver — intelligence reliably, efficiently and at scale. As AI moves from pilots into production, the infrastructure required to support it begins to resemble an industrial system rather than a software stack. That shift has profound implications for enterprises, cloud providers, telecommunications operators, data center developers and policymakers alike.

This conversation, encapsulated in RCRTech's recent AI Infrastructure Week virtual event, is often framed in binaries like cloud versus edge, hyperscalers versus enterprises and centralization versus distribution. In practice, these distinctions are dissolving. The real challenge is continuity across layers of compute, networks and energy systems, and the ability to orchestrate workloads across environments constrained by latency, cost, regulation and sustainability.

This report is intended as a strategy guide for navigating that reality. Drawing on interviews with infrastructure vendors, operators, analysts and technology leaders, it examines how AI is being operationalized, scaled and constrained in the real world. It highlights where assumptions inherited from cloud computing no longer hold, where new architectural patterns are emerging and where the next bottlenecks are already forming.

Top *RCR* takeaways

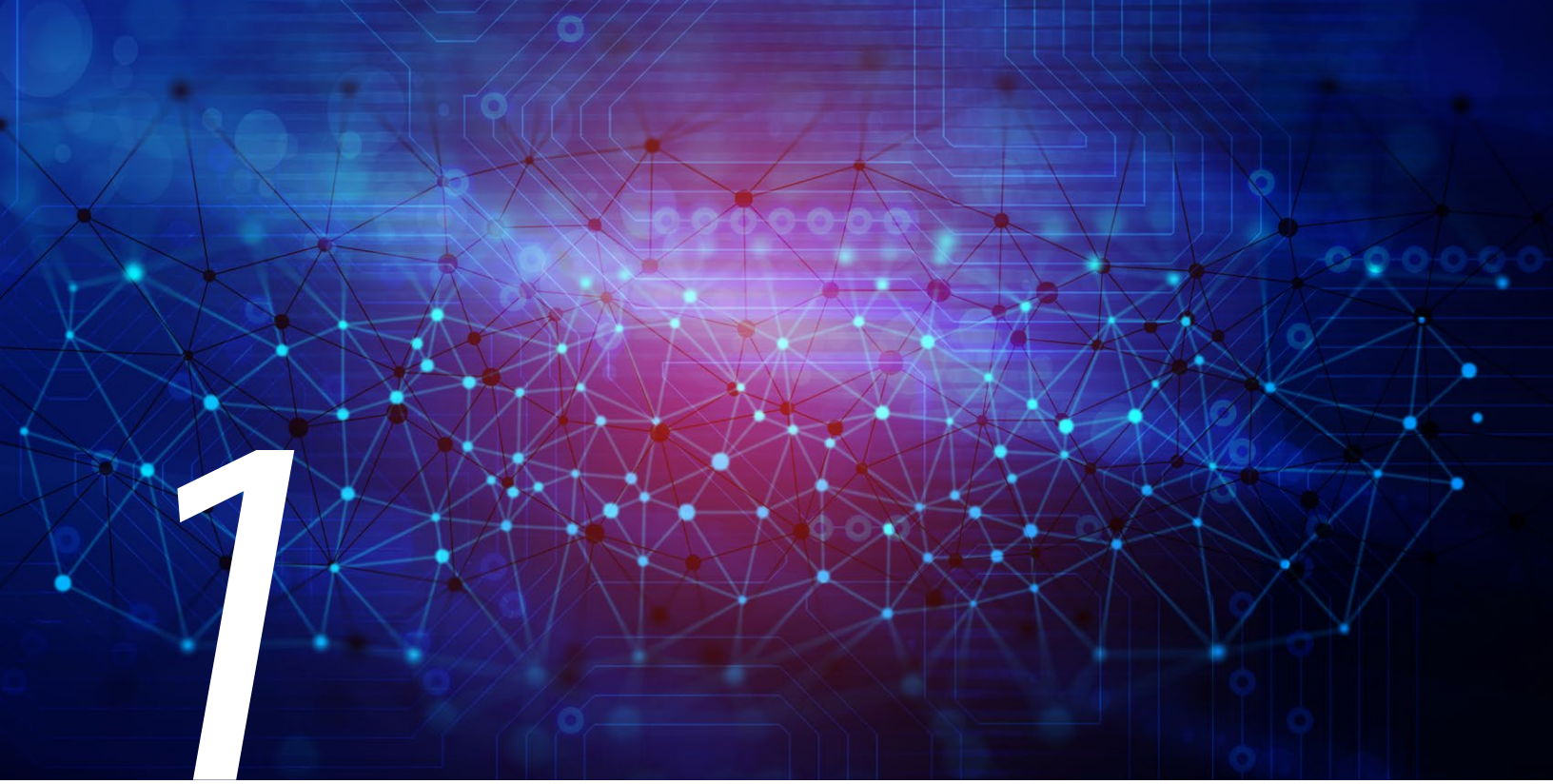
AI infrastructure has two distinct optimization problems. Infrastructure to build AI prioritizes scale and concentration. Infrastructure to consume AI prioritizes efficiency, distribution and proximity. Approaching both from the same perspective leads to overbuild and systemic fragility.

Operational readiness now matters more than model capability. Enterprise AI adoption is hindered by data readiness, workflow integration, networking and governance. AI success depends on production discipline.

Networking is a strategic layer in AI systems. As AI workloads grow more distributed and latency-sensitive, networks shift from passive transport to active enablers of utilization and economics. Observability, testing and automation directly affect AI ROI.

Energy and cooling are redefining where AI can scale. Power availability and thermal management are primary concerns. They shape AI geography, pace of deployment and sustainability.

The advantage is moving from accumulation to orchestration. Competitive differentiation is about more than owning compute; orchestrating compute, data and automation across environments is paramount. Systems designed for adaptability will outperform those optimized for scale.



1

The AI platform shift – speed, scale and the infrastructure of innovation

AI is no longer a future capability to be explored at the margins of the enterprise. It is becoming the defining platform shift of this decade and is reshaping how organizations operate, compete and create value across industries.

In an interview during RCRTech's recent AI Infrastructure Week virtual event (available on demand), Google Cloud AI Practice Leader Guruaj Bhat said, "We are frankly living through the biggest platform shift since the internet." The implication, he said, extends beyond the adoption of new technologies into a broader rethinking of business models, operating structures and the relationship between people and increasingly capable AI systems.

That reframing is essential, because the current phase of AI adoption is not being constrained by model capability. It is being constrained by organizational readiness, infrastructure maturity and the ability to move fast without creating long-term fragility. Enterprises today face a familiar tension. They are under pressure to move quickly with AI to capture near-term value, while also building systems that can scale, integrate and be governed over time. Bhat described this as an "innovation paradox" that Google Cloud sees repeatedly across customer engagements.

"In the short term you cannot afford analysis paralysis," he said. Enterprises need to identify "high-impact, low-risk pilots to solve a specific business problem or friction point." These early initiatives "will prove value quickly, generate excitement and build momentum." But speed without structure carries its own risks. "If you do only pilots," Bhat warned, organizations can end up with "a scattered landscape of solutions" that becomes difficult to scale, secure or govern. His prescription is to operate on two parallel tracks: "While you build these short-term initiatives, you focus simultaneously on building a very unified AI platform with the relevant security and governance controls."

Bhat compared the approach to building a city. You can open a shop quickly, but you still need to construct the roads, utilities and zoning that will eventually support skyscrapers. "Don't wait for the foundation to be perfect to start building," he advised — but do not ignore the foundation either. This dual-focus imperative, for speed and structure alongside experimentation and platformization, defines the core challenge of the AI era. It also sets the stage for a series of paradoxes now reshaping the AI infrastructure landscape.

The paradoxes defining the AI moment

As Vish Nandlall, an independent technology analyst, observed in recent writing, much of the public conversation around AI remains fixated on surface-level features like chatbots. "While most are focused on chatbot features, the real story is happening in the trenches of infrastructure, economics, and policy." The signals, he argued, "are clear, but they're not what you think."

He framed cost, underdog, measurement and ambition paradoxes. One at a time: training costs continue to rise while the cost to use AI through inference is falling rapidly. This suggests that efficiency, rather than raw compute power, is a competitive moat — over time there may be more value in delivering and diffusing intelligence than in strictly manufacturing it. On the competitive front, hyperscalers draw the most attention while Oracle has quietly become the single largest lessor of data center capacity in the United States. The takeaway is that control of physical infrastructure matters as much as algorithmic leadership. In a period of perceived rapid advancement, the tools used to measure progress can't keep pace. If AI improves faster than our ability to gauge its improvement, this complicates our understanding of differentiated capability. Finally, no amount of business- or state-level ambition can overcome structural bottlenecks, particularly access to space and energy. Put another way, ambition is outpacing institutional capacity.

Taken together, these paradoxes point to a common conclusion: AI has moved beyond its "magic" phase and into the brutal economics of an industrial-scale utility. The competitive game is no longer just about having the best model. It is about securing the cheapest power, the most resilient logistics, the most efficient delivery mechanisms, and the most durable partnerships.

As Nandlall wrote: "We've moved past the 'magic' of AI and into the brutal economics of an industrial-scale utility. The game is no longer just about having the best algorithm, but about having the cheapest power, the smartest logistics and the savviest partnerships."

Does AI need its own Manhattan Project?

Nandlall's industrial framing leads to a broader geopolitical and economic question explicitly posed by Futurum Group Chief and Senior Fellow David Nicholson. "The question we should be asking isn't whether to involve the government in AI," he wrote, "but whether we're prepared to treat it like the next Manhattan Project, because our competitors already are."

Nicholson pointed to historical precedent, including the Manhattan Project but also the interstate highway system, DARPA's pioneering work on the internet and the Apollo program. All were driven by a combination of public investment, private ingenuity and a shared strategic objective. The goal here wasn't to optimize short-term return on capital, but to secure long-term prosperity, security and technological leadership.

Nicholson argued that AI now belongs in that category. It "supports the general welfare" and "provides for the common defense" as surely as roads, bridges or military systems. In an increasingly competitive global landscape where nation-states are aligning against one another in the pursuit of AI leadership, the pressing question is how to act in a way that sustains leadership while advancing AI for broad societal benefit.

This report begins from the premise that AI is not just a software or cloud story. It is decidedly an infrastructure story that spans data centers, networks, power grids, silicon supply chains, governance frameworks and public-private collaboration. The chapters that follow examine how this infrastructure is being built, where it is under strain and how the balance is shifting from centralized creation toward distributed consumption. The platform shift Bhat set up is already underway. The challenge is ensuring the foundations being laid today can support the scale, complexity and responsibility that AI will demand tomorrow. And inertia is not an option.

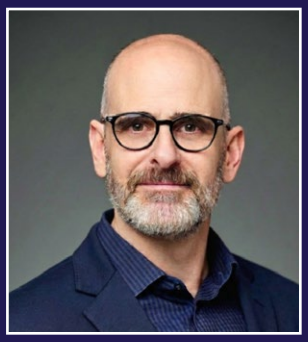
2

From raw compute to manufacturing intelligence – operationalizing AI infrastructure

The last 12 to 18 months of AI infrastructure headlines have largely been written by hyperscalers. That attention is warranted, but it can obscure a harder question for enterprise leaders: what does it actually take to operationalize AI as a repeatable production capability that delivers measurable value rather than a series of impressive pilots. The shift underway is from buying raw compute to manufacturing intelligence, where workflows, data readiness, operating models and networking determine how successfully AI scales.

"Do I actually have the data to make this business case fly, and is it something that scale and is it something that is secure and compliant?"

- Shawn Rosemarin, Vice President, R&D and Customer Engineering, Pure Storage



In a panel discussion during AI Infrastructure Week, Pure Storage's Shawn Rosemarin agreed that hyperscalers are taking the lion's share of headlines, but said that doesn't mean enterprises are behind. It means enterprises need to be deliberate in making infrastructure, people and process decisions as they develop AI strategies. Thomas Nadeau of the Open Compute Project framed what many CIOs are doing right now as an evaluation process with a familiar dilemma. "There's an evaluation happening at this point in the enterprise space where people are trying to figure out where to invest." The core question echoes every major platform shift: do you build or do you buy. "It hearkens back to the same questions we have with every technological change — do I need to build all of this myself or do I need to invest in purchasing a service?"

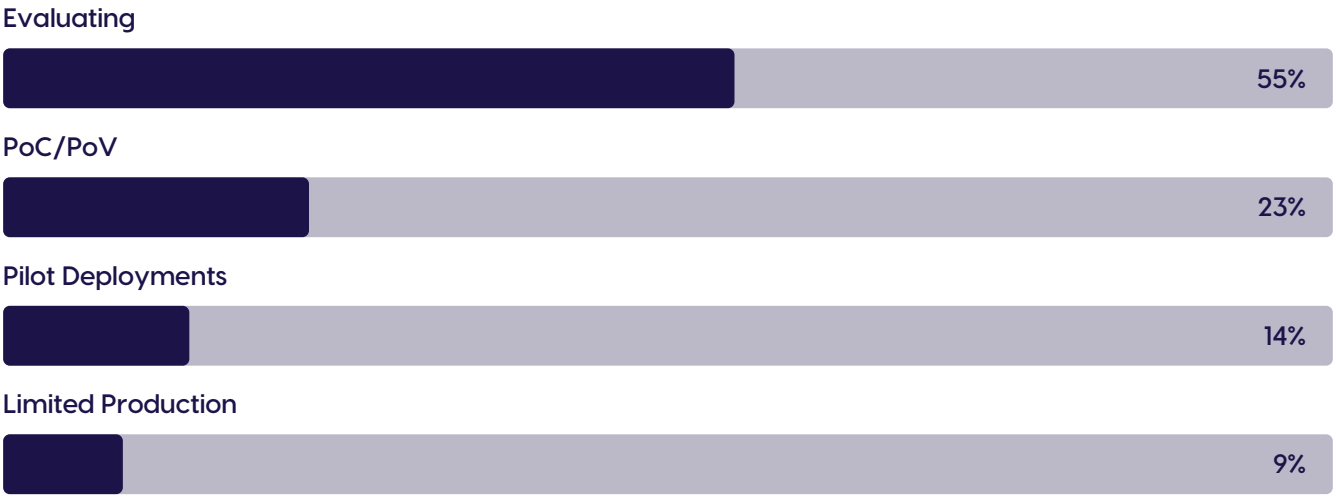
That decision is complicated by the fact that the ecosystem is diversifying fast. Ted Weatherford, vice president of business development for Xsight Labs, described AI as it is today as optimized for flexibility because few buyers can confidently predict what the stable application patterns will be. "In the end, once we know what the killer apps are, things will harden, they'll specialize." Until then, "You're going to see things stay flexible for a while… people don't know what they are gonna do exactly, they need flexibility." Rosemarin reinforced the same direction of travel from brute force toward smarter architectures. "We've solved all AI problems with brute force and solve all problems with hardware…But we are seeing not only models get more efficient…We're also seeing new architectures, we're seeing edge analytics, we're seeing DPUs, we're seeing advances in Ethernet fabric that are all driving a new paradigm to how we architect to these solutions."

In enterprise contexts, though, the bottleneck is often not a GPU count. It is whether the organization can build a credible business case and execute it safely. Rosemarin put the ROI question plainly: "What's the economic story behind this? Can I get to an ROI that makes sense?" Then he went to the root cause. "Data is what feeds all of this, and ultimately is my data ready?"

He also highlighted a less discussed mismatch between how enterprise data has been captured and what machines need to learn reliably. "What many organizations are finding…is that while we've digitized the majority of our data…all that data was essentially recorded by humans…Frankly, machines don't think like humans." As a result, AI quality issues often trace back to the fidelity and structure of the underlying data. "Machines want high fidelity, raw data."

This is where the "AI factory" concept becomes less about a room full of accelerators and more about operational discipline. When asked how enterprise IT leaders should navigate the uncertainty, Rosemarin argued for first principles thinking and for acknowledgement of the limits of technology alone. "AI is not solved by technology. AI is solved actually by the modernization of people and process." Nadeau added a similar caution against overpromising or framing AI as a replacement project. "That's the folly I think…[AI is] an augmentative experience. And if it's augmented in the right way, it actually is additive."

Where are you today in your Enterprise AI Journey?



Responses to an audience question on organizational AI maturity underscored the point. Most participants are not yet operating at scale. A poll showed 55% evaluating, 23% in a proof of concept phase, 14% in pilot deployments and 9% in limited production. The market is moving, but it is still early.

An increasingly diverse AI ecosystem underscores the importance of networking

If AI factories are meant to turn infrastructure into production capability, networking is one of the key conversion mechanisms. Weatherford described DPUs as a foundational building block that unifies security, storage, virtualization and networking, often over Ethernet. "It's just a basic building block for the whole story." He also emphasized software-defined networking as a vital flexibility layer while applications and architectures are still evolving.

Mike Bushong of Nokia pushed this further in an interview. He drove home the point that AI infrastructure is not a set of components, but a system that must behave coherently under load. Compute resources "at some point they've got to talk to one another. So the network really is the connective tissue that brings it all together." He warned against reducing networking to "bits and bytes" because decisions now carry workforce, budget, supply chain and vendor management implications. "There are broader dynamics."

On whether networking innovation is keeping pace with rapid advancements in AI capabilities, his answer was that it largely is, but the pacing factor is shifting. "Where's the long pole in the tent?" In his view, it is no longer networking. It is power, density and grid realities. Even so, the architecture choices driven by training versus inference are reshaping the network's role. On inference specifically, he expects "older generation GPUs become viable" and that "latency becomes really, really important." The consequence is distribution. "You're not going to see everything centralized." This in turn creates a scale up versus scale out fork that touches everything from silicon to topology to operations.

Operations is where the narrative sharpens. As systems scale, failures become routine rather than exceptional. A former colleague's line stuck with Bushong: "Even if you're operating at six nines, at sufficient scale, things

are breaking every couple of hours." That reality shifts the emphasis from building a fabric to running a fabric. The critical capability is diagnosis, remediation and confidence at scale. "This whole operational piece is really challenging."

That is also why observability has moved from a best practice to an enabling requirement for AI infrastructure. Bushong described observability as an attempt to see into "systems of systems" where not everything is explicitly instrumented and where tight tolerances matter. "In an AI world, where lossless is king, your tolerances are dialed in way tighter than they otherwise would've been." At scale, debugging is not "a needle in a haystack," it becomes "a needle in a stack of other needles."

He also offered a grounded view on cost structure. Many benchmark networking at roughly 10% of total data center cost, but the more interesting question is the operational and software cost needed to run and monetize AI services. "I would just encourage people to think it's not only about the bits and bytes." In many cases, the differentiator is not hardware procurement but software development, integration and the talent required to operate complex stacks. "It's really about software development cost."

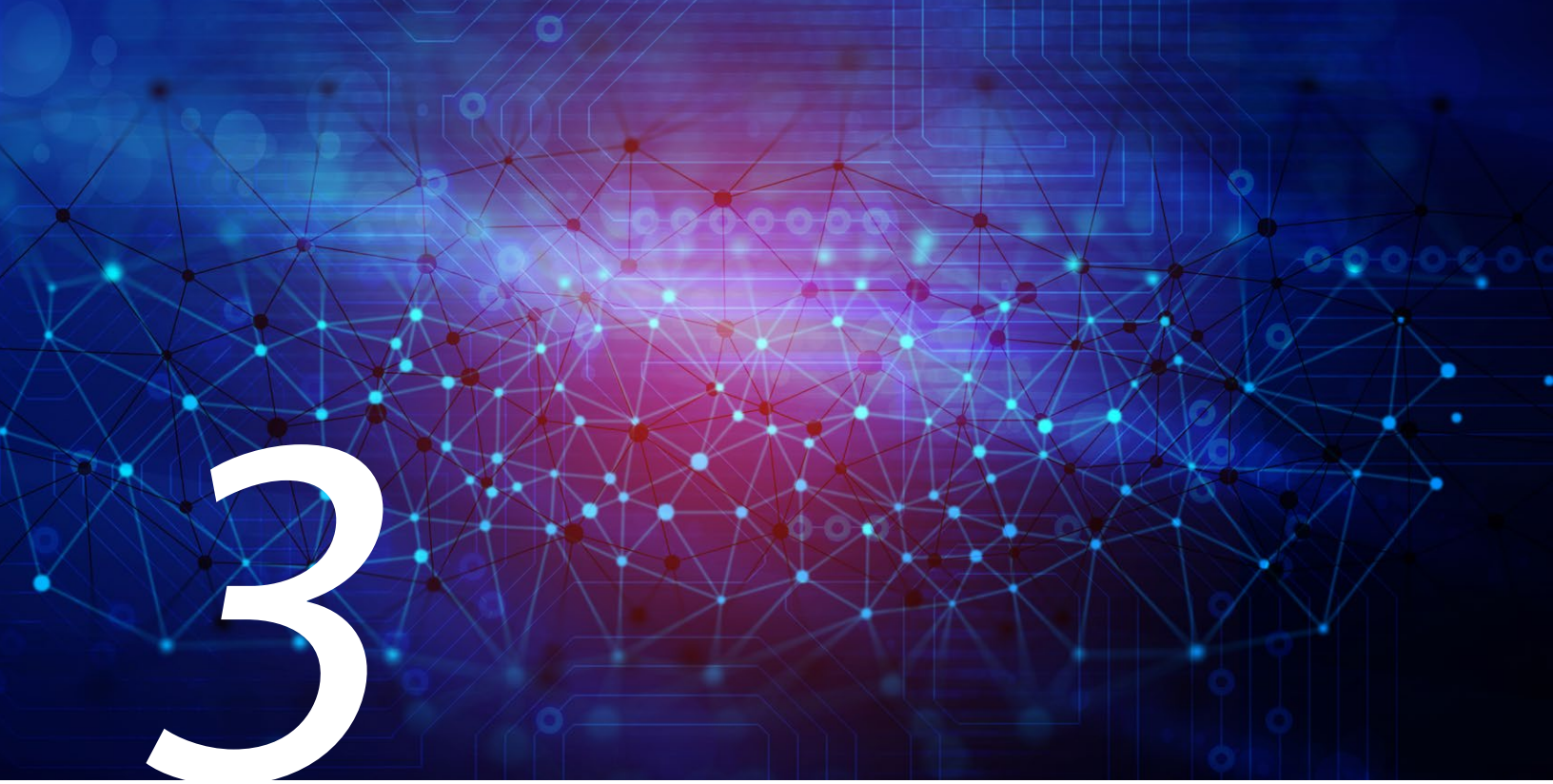
Agentic AI – making complexity consumable

If networking is the connective tissue, automation is the muscle that turns connectivity and compute into a dependable production line. Anil Kollipara of Spirent, part of Keysight, framed the operational problem through the evolution from 3G to 5G. Disaggregation and cloud native design increased flexibility but shifted responsibility. With 3G and 4G, "The responsibility to test and put the network together and make sure everything is working laid with the vendors." With 5G and disaggregation, the ownership of system integration now falls to service providers.

Kollipara's central point was that complexity has grown beyond what manual processes can sustainably manage. "The networks today are 150x more complex than legacy networks. And the only way to address or manage this operational complexity is through continuous testing and…total automation." Even with automation pipelines, failure handling often remains slow because the workflow is human intensive. "The process of getting to the bottom of it…is a very painstaking and tedious process." He described a chain of artifact collection, diagnosis, reproductions, root cause analysis and remediation planning that can take weeks.

This is where the chapter's arc reaches its destination. Agentic AI is positioned as the mechanism that can absorb workflow complexity and make AI infrastructure and AI-enabled networks usable at scale. Kollipara distinguished the key shift with agentic AI as execution. "Agentic AI is specifically executing a complete workflow." It may include LLM reasoning, but it also orchestrates data fetching, processing and sequencing across tools and domains. "That's where agentic AI is different." He offered a concrete example from a Spirent customer case where a ticket opened Oct. 7 took seven weeks to reach root cause analysis with a remediation plan on Nov. 18 through the normal escalation path. With agentic AI, root cause analysis was done in two minutes and produced the same answer as the final remediation plan.

Taken together, these perspectives show what it means to operationalize AI. The AI factory is not just a pile of GPUs. It is a readiness discipline grounded in data, people and process. Networking is not a cost line item, it is connective tissue that determines how systems behave under scale and strain. Agentic AI then becomes the control layer that turns operational complexity into something consumable, repeatable and fast enough to keep up with the pace of change. That is the shift from raw compute to manufacturing intelligence.



3

Scaling AI – clusters, networks and the multi-site reality

As AI enters sustained production, the realities of massive scale come into sharp focus. Beyond standing up ever more GPU capacity, scaling AI is about orchestrating dense compute clusters, managing power consumption, redesigning networks for new traffic patterns and extending those architectures across multiple sites. The work ahead is technological, operational and economic all at once and none of it is finished.

Raymond Chik of the IEEE AI Hardware and Infrastructure Working Group tracked the evolution of power consumption which he said "has always been a concern" for large-scale computing. Specifically what's changing is where that power is consumed. "Most of the power that is consumed in AI compute actually is burned through data movement." For LLMs in particular, the cost of moving bits between compute, memory and nodes increasingly impacts systemic efficiency.

That observation reframes how clusters must be designed. When optimizing AI systems, "We all have to think about how do we efficiently minimize data transfer between all these nodes." The challenge spans from chip-level design to full cluster architecture and it cuts across both scale up and scale out models.

Clayton Wagar of Nokia emphasized that AI clusters behave differently from traditional cloud infrastructure. "It's a coherent process that requires equipment to be working in tandem which is why interconnect — the network — is so important." In classical cloud environments, power consumption can be diffused through load balancing. In AI training and inference, the network becomes "a fundamental part of how AI works."

That reality changes the economics of power allocation. As Wagar noted, "By saving the power in the network, or by being as efficient as possible in the network, then you have the ability to deploy more of the GPUs…and ostensibly that's where the revenue is." Power efficiency in networking is not a secondary optimization; it directly determines how much productive compute can be deployed.

Looking ahead, both speakers pointed to optical technologies as a necessary step in sustaining scale. Wagar described an industry-wide densification trend. "We have this densification; racks that are consuming…hundreds of thousands of watts of power." Copper interconnects can only go so far, literally and figuratively. "Electrical to optical transition over the next two to five years" is underway for both scale up and scale out networks. "Everyone of those will be impacted by optical technologies in order to get to the power consumption that we know we need to get to."

Chik outlined the tradeoffs in more detail. Copper remains cheaper and is being pushed further with active electrical cables but, "There's a physical limit." Long term, "No doubt it's optical." He pointed to co-packaged optics and integrated photonics as major industry drivers, while also noting the reliability and serviceability challenges that must be solved before mass deployment.

Wagar was clear that this transition will not be incremental. "The change in the construction of a network…will change more radically over the next five years than it has over the past 20." Vendors must adapt their engineering disciplines and operators must learn to manage new tradeoffs between flexibility, integration and reliability. Scaling across sites, from single clusters to distributed systems

As clusters grow, a second challenge emerges: scale across. "We can't make a single building large enough or maybe don't want to," Wagar said. Multi-site architectures allow operators to capture power availability, connectivity or geographic optionality. The technical problem then becomes extending high bandwidth, low latency connectivity across buildings and campuses. "The key element really is providing that bandwidth between buildings." This is not traditional data center interconnect for peering; it is closer to a LAN extension operating at massive scale.

Chik reinforced that hardware is only part of the efficiency story. "The whole AI deployment is not just hardware… It also has to do with the software…the data science, the model efficiency." He pointed to increasingly efficient models as proof that architectural choices at the application and model level matter as much as silicon.

At the lowest level, even high bandwidth memory is becoming a bottleneck. For large models, "All these parameters have to sit somewhere and now they're sitting in HBM." Streaming them through GPUs for every token creates significant energy overhead. New approaches such as 3D stacking and memory-centric architectures aim to reduce this traffic by bringing compute and memory closer together. Each of these approaches still relies on robust networking. As Wagar put it, "Every one of the things that we've mentioned…all of them require robust networking."

"One thing is clear…AI traffic is fundamentally different from convention network traffic."

- Stephen Douglas, Head of Market Strategy, Spirent, now part of Keysight Technologies

AI traffic breaks old assumptions

If cluster design defines what is possible, network testing defines what is sustainable. Training and inference generate massive, bursty East-West flows that are highly parallel and bidirectional. They are both latency sensitive and congestion sensitive. Inference introduces additional complexity. It requires high connection rates and concurrency from large numbers of devices and applications. Latency varies with compute placement and network paths. Uplink traffic often exceeds downlink traffic due to multimodal data upload.

As Douglas warned, "Networks for AI must be designed and tested to deliver deterministic performance, scalability and dynamic workload orchestration far beyond the conventional sort of IP-based traffic we've been used to dealing with."

This applies inside data centers and across wireline and wireless networks. Without proper testing, AI traffic can degrade existing services. "It's fair to say, connectivity is becoming the lynchpin for AI scaling." Douglas argued that testing is now inseparable from scaling. In data center fabrics, "Testing actually the interconnect fabric…is absolutely vital to keep GPUs fed rather than sitting idle." Metrics like job completion time expose the real business impact because overall progress is gated by the slowest node in a synchronized workload. Encryption performance also matters as East-West traffic is secured.

Historically, testing required real GPUs and full scale clusters which was "incredibly costly," he said. Emulation changes the equation. "Up until this point, the only way to test out the buildout of these huge, mega GPU clusters has been to use real GPUs." Emulating GPU workloads and traffic patterns enables repeatable and cost-efficient validation. Douglas outlined three pillars of modern testing. Digital twins replicate networks and workloads. Intelligent automation enables continuous and active testing. AI test systems help scale test creation, execution and remediation. The ROI is threefold. "Being able to actually confidently and quickly validate…so you can monetize it is absolutely critical." Emulation reduces cost. Automation and reduced lab overhead improve energy efficiency by avoiding unnecessary GPU usage.

Neoclouds, hyperscalers and the economic layer of scale

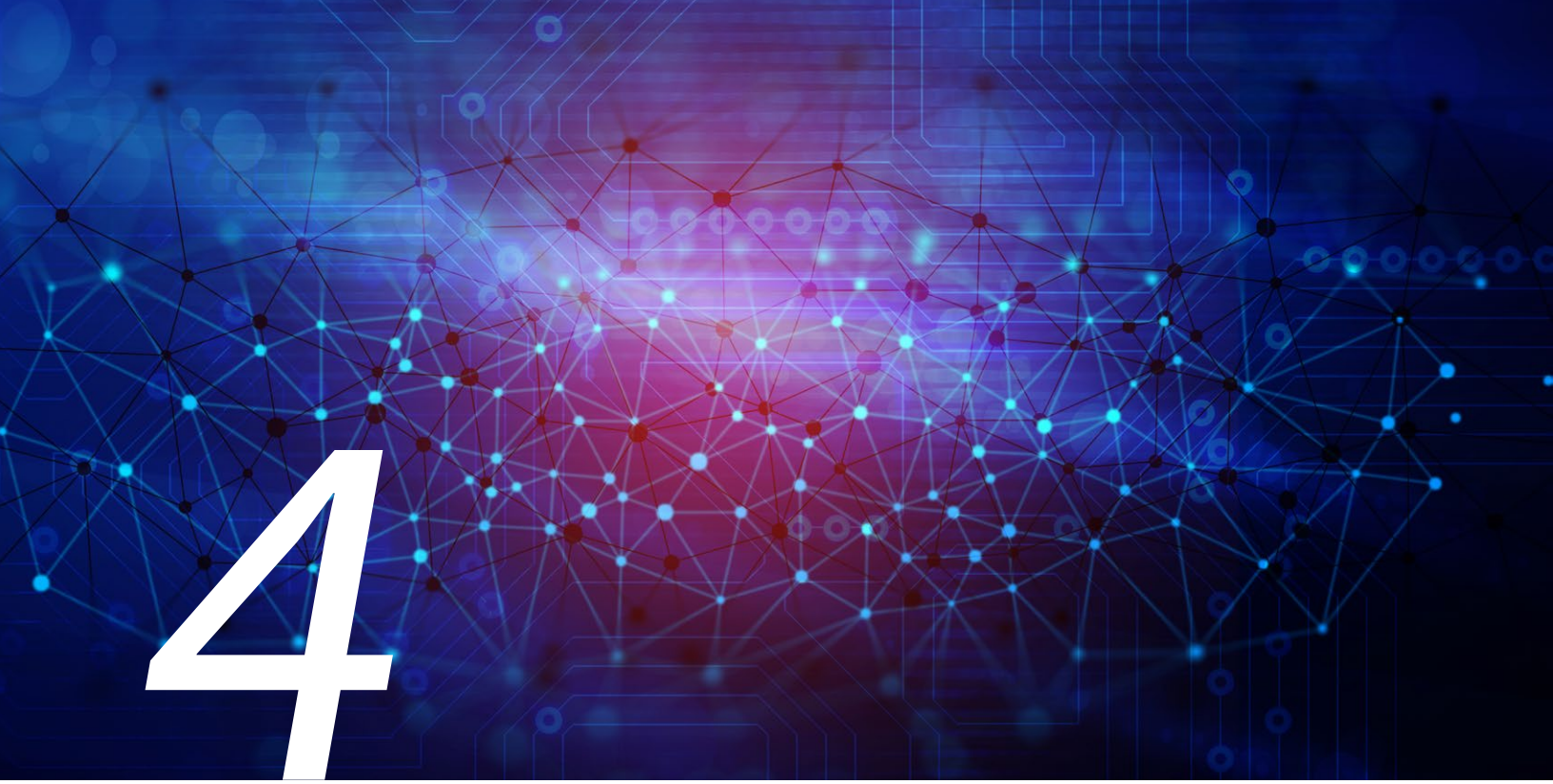
At the economic layer, scaling AI plays out differently across infrastructure providers. Reece Hayden of ABI Research described neoclouds as focused on "providing GPUs in the most effective way in purpose-built

architectures." Unlike hyperscalers, neoclouds primarily deliver infrastructure rather than managed services or broad ecosystems on top of infrastructure.

The differences matter. Neoclouds offer predictable pricing and faster access to leading edge hardware. Hyperscalers offer global scale and deep integration across enterprise workloads. "What's really driving them at the moment is the big issue in the AI market and that's cost." For training in particular, neoclouds can be cheaper. Sovereignty and power availability also factor into regional decisions. Still, neoclouds face structural challenges. Software lock-in and ecosystem breadth favor hyperscalers. "I think [neoclouds] will capture meaningful share of AI greenfield workloads…I don't think it's going to be comparable to the hyperscalers."

Scaling AI remains a work in progress. Power efficiency, optical interconnects, multi-site architectures, new traffic patterns and economic tradeoffs all intersect. None of them can be solved in isolation. The path forward requires continued innovation in hardware, disciplined operations and rigorous testing. As AI traffic grows and spreads across networks and sites, the ability to validate performance, cost and energy efficiency becomes a strategic capability.

As AI clusters scale across sites and networks, they're increasingly bounded by very real physical limitations around energy availability and thermal limits. At this point, scaling AI stops being purely a question of architecture and becomes a question of power (of course), cooling and sustainability.



4

Energy, sustainability and the physical limits of AI

Driving AI scale is an exercise in managing hardware and software; however, how physical constraints are managed matters today and carries long-term implications. Compute can be virtualized and workloads abstracted, but power, water, land and materials cannot. The next phase of AI infrastructure growth will be shaped less by algorithmic ambition than by the availability and governance of real-world resources.

This framing is central to recent research from Opna, which positions AI infrastructure as a climate and energy systems challenge as much as a digital one. As Opna Founder and CEO Shilpika Gautam argues, "The expansion of AI data center infrastructure is one of the largest, if not the largest and fastest, capital investments and infrastructure rollouts of our time." Without deliberate intervention, that expansion risks intensifying pressure on energy grids, water systems and local communities, while locking in decades of embodied carbon through construction materials and site selection.

At the same time, Opna's work highlights a structural shift underway in how AI is deployed. Training large foundation models favors centralized hyperscale facilities built for peak capacity. Inference, by contrast, can be modular, distributed and aligned with actual demand. "If training demands centralization, inference opens the door to distribution," the paper notes. "Modularity and right-sizing is what can allow inference to meet the physical world on its own terms, relevant for local context, adaptable, and sustainable."

This distinction is critical because energy availability is poised to bind AI growth which reiterates the earlier points made on the key role of networking. International Energy Agency estimates cited by Opna project that global data center electricity demand could more than double by 2030. In many established data center markets, grid interconnects and generation capacity are already limiting new deployments. Without changes in architecture and operating models, AI risks becoming a destabilizing force for power systems rather than a driver of modernization.

Opna's analysis reframes data centers as potential assets rather than liabilities within energy systems. Rather than behaving as inflexible loads, modular inference-first facilities can function as anchor customers for renewable generation, participate in demand response, and support local grid balancing. Smaller footprints also make it more feasible to reuse waste heat and integrate with district energy systems. In Opna's words, this allows AI infrastructure to become "a lever for accelerating, rather than obstructing, the path to net zero."

These sustainability considerations extend beyond operations. A significant share of a data center's emissions footprint is embedded in construction materials such as steel and concrete. Bespoke hyperscale builds slow the adoption of low-carbon alternatives and increase the risk of stranded assets as cooling and compute technologies evolve. Modular designs enable repeatability, prefabrication, and aggregation of demand for lower-carbon materials, aligning infrastructure economics with climate goals.

The liquid cooling imperative

If energy availability defines the boundary conditions for AI infrastructure, cooling is where those limits are most immediately felt. Power consumed by AI workloads ultimately becomes heat, and the ability to remove that heat efficiently is now inseparable from the ability to scale compute itself.

Joe Capes, CEO of LiquidStack, captured the urgency succinctly. "The three most pressing pain points are power, power and power." As AI factories push toward ever-higher rack densities, traditional air-cooled designs are reaching practical limits. GPUs and CPUs designed for AI training and inference generate thermal loads that air systems cannot dissipate without excessive energy overhead.

"Liquid cooling is a means of removing heat or providing thermal management for a variety of different components within data centers," Capes explained. While air cooling has historically dominated, liquid cooling has matured through years of R&D and earlier adoption in high performance computing and crypto mining. With AI, that transition is accelerating. "With the advent of AI scale up and much higher power from GPUs and CPUs that are being used for this type of compute, we're seeing the adoption of direct-to-chip liquid cooling."

Direct-to-chip approaches use cold plates mounted directly on processors, circulating coolant through precision loops. While the concept dates back decades, the scale is new. AI workloads are driving rack power densities far beyond historical norms, placing pressure not only on cooling systems but across the entire supporting ecosystem. Capes pointed to constraints in "switching gear, transformers, back up generators, low and medium voltage switch gears, chillers and coolant distribution units which are the heart of a direct-to-chipliquid cooling deployment." In many cases, demand now exceeds supply.

Despite these realities, adoption remains uneven. Some operators view liquid cooling as complex or costly. Capes acknowledged the learning curve. "A lot of the operators are deploying direct-to-chipliquid cooling for the first time or for the first time at scale, so there is a very steep learning curve." At the same time, the efficiency gains are significant. "Using liquid cooling can be 40 to 50% more efficient than air cooling."

Looking forward, Capes expects rapid evolution in cooling technology. "The next three to four years is going to see" widespread adoption of single-phase direct-to-chip systems, followed by a transition toward two-phase technologies as power densities rise further. "The reality is…we're going to see one to two megawatt IT racks in the future and in order to support those types of power densities and heat loads, we will need to move to two-phase."

That transition introduces a final and often overlooked dimension. Advanced cooling relies on specialized refrigerants, and Capes emphasized the importance of developing solutions that are "appropriate for use cases and friendly for the environment." In other words, solving AI's thermal challenge is not only an engineering problem. It is a sustainability problem.

Taken together, the Opna and LiquidStack perspectives underscore a shared conclusion. AI infrastructure has reached a point where physical limits, environmental impact, and system-level design can no longer be treated as secondary concerns. Energy efficiency, cooling innovation, and sustainability are now core enablers of scale. The next phase of AI growth will not be defined solely by faster chips or larger clusters, but by whether the industry can align compute ambition with the realities of power, heat, and the planet.

As energy, cooling, and environmental constraints tighten, they are not just limiting how large AI infrastructure can grow, but also where it can be placed and how it must operate. Those pressures are accelerating a shift away from monolithic, centralized deployments toward architectures that distribute inference across cloud, regional, and edge environments—setting the stage for the next infrastructure frontier.

5

Cloud to edge continuity – distributed inference and the next infrastructure frontier

For the past few years, the AI infrastructure narrative has been dominated by the race to build intelligence through investment in hyperscale data centers and frontier training clusters with unprecedented GPU density and power draw. That build phase is real, it's ongoing, and it's strategically critical; but it now represents only one side of the infrastructure equation. As AI systems move from experimentation into daily operational use, a different set of constraints is coming to the fore.

The infrastructure required to consume AI — that is to deliver inference reliably, securely and economically at scale — is becoming the next frontier. The pendulum is swinging from centralized build-out toward distributed delivery. This shift does not pit cloud against edge. Instead, it elevates continuity across a cloud-

to-edge continuum, designed around how inference flows and where value is realized. In other words, training infrastructure optimizes for scale and concentration; inference infrastructure optimizes for distribution and proximity to action.

The rise of edge AI and the age of inference

In discussing last-mile AI infrastructure, Apple, Meta and Tesla veteran Nikhil Tyagi challenged the tendency to frame edge inference as a location (latency) problem, instead positioning it as a latency and orchestration problem. "It's not a single point…it's more about designing how the inference will flow across the continuum." Latency, cost, governance and policy considerations are all shaping that flow. What once looked like a simple cloud-to-device model is now a more complex topology, including cloud to regional edge, edge to on-prem when available, and onward to increasingly capable devices. "Connectivity in that spectrum becomes fairly important," Tyagi said, raising practical questions about Wi-Fi versus 5G, indoor versus outdoor mobility, and seamless handoff across environments.

That framing aligns with how hyperscalers themselves are classifying AI workloads. Examining workload definitions, Majed Al Amine, head of distributed computing and artificial intelligence for Google Cloud in EMEA, distilled a complex taxonomy into two core dimensions. The first is proximity to delivered value — "sizing of the activity itself meaning is it something that is requiring huge infrastructure for training models…or basically closer to the user where we call it inferencing. That's one dimension: how close is it to delivering value to the end user." The second is where a workload sits in a dynamic model lifecycle. "We, as specialists, look at the AI models as either training vs. inferencing or from a learning perspective: pre-training models to fine-tuning and then delivering it." Together, these dimensions clearly separate infrastructure optimized for training from infrastructure optimized for inference consumption. The convergence of demand-side, supply-side and economic pressures makes distributed inference unavoidable.

Colocation providers are already building to serve that second category. Colovore's Tomek Mackowiak, vice president of product and business development, framed an infrastructure profile tuned explicitly for enterprise inference. "We typically concern ourselves with inference workloads…for the enterprise." Rack densities ranging from 20 kW to 120 kW reflect rising model complexity, as well as the diversity of models in production. "The variability is very high but the one common thread across them all is that you really need high-density infrastructure to deliver it." Enterprises increasingly rely on workload routers to direct inference requests across a heterogeneous mix of models and platforms, and Mackowiak emphasized that "all of it kind of goes beyond the scope of the regular air-cooled data center."

From the telecom perspective, this is not the first time distributed compute has been positioned as transformative. Stewart Dudding of Vodafone Business International was clear that "distributed edge compute isn't new. It's just that up to now AI hasn't been the use case for it." What is new is the interaction model. He pointed to a steady rise in voice streaming as users engage agents conversationally — "that's just going to become the norm" — along with generative gaming and public safety applications. Importantly, Vodafone is also applying AI inwardly. "Edge AI for us brings the kind of self-healing nature of our investments in our core infrastructure to the edge as well." Observability and inference are moving closer to where customer experience is shaped.

At the same time, Dudding urged restraint. Operators cannot afford to repeat earlier MEC cycles by overcapitalizing too early. The challenge is identifying where generative AI and edge inference genuinely justify incremental infrastructure, without starving core CDN and connectivity investments.

From MEC to CDN 2.O but this time with AI

When experts were polled as to whether this moment simply replays earlier edge narratives, Al Amine reframed the question around feasibility. Latency, he argued, is now being positioned more pragmatically, particularly for consumer applications. For enterprises, sovereignty and security loom larger; they need the ability to use AI "without exposing information." But the decisive factor is supply-side scalability. "I think the bigger element that's pushing edge AI is not the use case itself but the feasibility of service providers to deliver that use case." Power and cooling constraints in centralized facilities limit how far traditional GPU-centric scaling can go. "By pushing it to the edge, you're able to use a more ASIC-based approach…for specific applications at the edge that are more power efficient, they're smaller footprint, they can be hosted in locations a big data center would not be able to accommodate."

To that delivery point, Dudding said collaboration would be key as Vodafone and other service providers with an edge footprint figure out the business model. "Someone's got to pay for it. There's got to be a customer at the end of this that's willing to pay more for a faster response." Vodafone, he said, caches content at the edge already, so why can't they cache inference capabilities and compute at that same edge? Dudding referred to it as CDN 2.0 and added, "The architecture feels fairly clean-cut to me."

Scanning what he sees gaining market momentum in this quarter and beyond, Mackowiak concurred with the CDN 2.0 assessment. "The folks that are responsible for the CDNs in the US are very much the next wave of enterprise infrastructure procurement." He recalled an initial AI training wave, followed by a shift toward inference that continues to evolve into something more workable and usable, and now a content delivery and networking wave. "It's going to hit…[this] year." Legacy, but perhaps relevant again, CDN providers "are really footing the bill because they're going to be first in line…to reap the benefits of the infrastructure." Foundation model developers like OpenAI and Anthropic, he said, have a core competency in developing AI tools, not "getting the services to the people that use them."

Regional data centers and infrastructure decentralization

Decentralization was a central theme in a CXO discussion with Scott Willis and Brad Alexander, the CEO and CTO respectively, of regional data center and services provider DartPoints. Willis explained that DartPoints' investment thesis has historically focused on enabling data and compute ecosystems outside Tier 1 markets, and that AI has accelerated that strategy. Historically, training clusters concentrate in major hubs. Alexander was clear that this will continue. "The foundational training will continue to occur…in these large markets." What changes is where inference lands. "What we'll see is we'll continue to see decentralization into markets outside our Tier 1 markets… where inference becomes a huge impact." Manufacturing floors, operating rooms and research universities all demand low-latency, local inference coupled with dense connectivity.

That demand is reshaping DartPoints' approach to data center design itself. Where 5–8 kW per cabinet was once standard, "now 30 kW becomes an entry point," with facilities increasingly built or retrofitted for liquid cooling at 120 kW per cabinet, Alexander explained. But even in Tier 2 and Tier 3 markets, limits remain. "We'll be limited not by space and power, but we'll be limited purely by data gravity and the amount of customers that can come in." Success, Willis argued, depends on creating sustainable regional ecosystems that serve enterprises, HPC users, AI inference customers and hyperscalers seeking regional reach.

KPMG's Philip Wong tracked the financial and regulatory implications of this decentralization. He described how as enterprises move from experimenting with AI to scaling it, the trend is "from large computation to closer to the edge in order to reduce latency, drive speed." Smaller, task-specific models, including those trained solely on

enterprise data, enable inference on increasingly lightweight infrastructure, including devices themselves. "All of these changes and advances…is going to drive a more distributed use of digital infrastructure."

But distribution introduces cost-discipline challenges. On the supply side, data center economics are constrained by land, permitting, power sourcing and supply chain factors. On the demand side, enterprises face rising operational costs as token volumes grow with agentic systems and reasoning workloads. "From the enterprise perspective, a lot of the cost is around managing computation and managing storage cost." Wong warned that without strong governance, "You can quickly get out of control if you don't really step back and think about the cost, think about where the ROI is." The trajectory mirrors early cloud adoption in that experimentation is cheap, but scaled consumption is not.

Taken together, these perspectives point to a structural transition. Infrastructure to build AI remains centralized, capital-intensive and hyperscale. Infrastructure to consume AI is increasingly distributed, latency-sensitive, tailored to particular use cases and governance-driven. The pendulum is swinging not because the cloud is failing, but because inference demand is exploding, and because feasibility, economics and customer experience now dictate where intelligence must live. The next phase of AI infrastructure, again, isn't a false cloud/edge binary; the key is creating a continuum that effectively stitches the two together.

As data gravity diffuses from Tier 1 markets to every corner of the globe, and as the focus shifts from making AI work at all to making AI work for enterprises and individuals, stakeholders should ask themselves three key questions to inform their technology strategies:

1. Where does inference have to run for high-value use cases?
2. Which parts of the AI stack are becoming cost centers versus differentiators?
3. Which partners (hyperscalers, telcos, colo providers, etc…) have advantages that match relevant consumption patterns?

The next phase of AI infrastructure will be won by those who align intelligence delivery with real-world constraints and outcomes. That alignment depends on infrastructure strategy, ecosystem partnerships and long-term execution.

In the AI era, testing turns a risk into an advantage

A Q&A with VIAVI Solutions CTO Sameh Yamany

Q: Considering this coming shift from AI compute for training to AI compute for inference, what do you see communications service providers (CSPs) getting right, and where do you think they should put more emphasis and capital?

A: CSPs increasingly recognize that AI has moved beyond experimental lab workloads and is rapidly becoming a real-time service layer. They are already investing in distributed edge infrastructure to enable low-latency inference, and they've begun modernizing transport networks for higher throughput and reduced jitter while evaluating GPU pooling and cloud-native designs. The next step where greater focus and capital are needed is building a network-aware AI architecture that delivers deterministic latency across metro and edge domains, supported by AI-driven testing and assurance tightly integrated into orchestration workflows.

Q: Further on the training/inference shift, how does this change the way we should think about network performance, observability and assurance for AI-driven use cases?

A: Training workloads primarily tax bandwidth, while inference demands determinism, latency stability and operational resilience. As a result, networks must shift from chasing peak throughput to guaranteeing predictable latency, real-time microburst detection and validated lossless Ethernet performance under congestion. Operators will require AI-workload-aware telemetry, end-to-end GPU-to-fabric visibility and cross-layer correlation spanning compute, network and application domains. Assurance must evolve into a continuous, automated AI-driven system — effectively 'AI testing AI.' Because inference workloads compress tolerance margins, even millisecond-level deviations can degrade service quality, making proactive validation essential instead of reactive troubleshooting.

Q: Whether it's a CSP or another large, distributed enterprise, how do you assess whether they're actually ready to run AI in production, as opposed to simply experimenting with it?

A: From my experience with hyperscalers, governments and telecom operators, true production-readiness for AI rests on five pillars:

1. Infrastructure readiness: a validated fabric that performs at scale, proven multi-vendor interoperability and deterministic latency under real-world load.
2. Observability maturity: end-to-end, cross-layer visibility from optical transport through applications; AI workload fingerprinting; and readiness for closed-loop automation.
3. Operational integration: AI embedded in IT and NOC workflows, with AIOps delivering decision automation, not just dashboards.
4. Security and quantum-safe posture: encrypted data in motion, a defined PQC transition roadmap and validated key-lifecycle and rekey strategies.

"In AI-native networks, testing is no longer an event as it becomes an always-on discipline embedded directly into orchestration.."

- Sameh Yamany, Chief Technology Officer, VIAVI Solutions

5. Energy and sustainability alignment: workload-aware power optimization and continuous monitoring of fabric efficiency.

If AI remains confined to a sandbox cluster with manual intervention, it is still experimentation. When it becomes validated, observable, automated and resilient under production traffic, it is truly operationalized.

Q: All of these AI infrastructure systems are complex and getting more complex. From your vantage point, how does continuous testing and validation become a mechanism that turns that complexity from a risk into an operational advantage?

A: AI infrastructure introduces multi-layer fabrics, accelerated compute, disaggregated architectures and software-defined control planes, capabilities that amplify complexity, and without validation, turn it into operational risk. Continuous testing flips that equation, enabling early congestion detection, automated root-cause isolation, correlation between model drift and network behavior, faster and safer software rollouts, and confidence across multi-vendor environments at scale. By shifting testing both left into design and right into production telemetry, operators move from firefighting to continuous optimization.

Q: Looking ahead, as AI infrastructure becomes more distributed, energy constraints continue (and become more acute), and software-based orchestration rises as a differentiator, what capabilities will separate infrastructure leaders from those who struggle to operationalize AI at scale?

A: As AI infrastructure becomes increasingly distributed, energy-constrained and software-defined, the organizations that lead will excel in five areas: engineering deterministic fabrics that guarantee low-latency inference across metro and edge; deploying AI-native observability that delivers real-time, cross-layer predictive insight; orchestrating compute with energy intelligence based on power, cooling and network efficiency; leveraging continuous validation through digital twins and emulation to de-risk deployments; and building quantum-safe, secure-by-design networks resilient to future cryptographic threats.

Conclusion

The AI platform shift outlined at the beginning of this report is happening and it's not slowing down. The associated infrastructure buildout is an exercise in managing constraints but, again, it's not slowing down. While users have many technology and strategy options in front of them, what is not optional is inaction. There will be winners and losers, and companies that are left behind may face existential questions. Early movers will carve out durable competitive advantage.

We return to Guruaj Bhat of Google Cloud: "Start now and stay flexible…The companies that will lead their industries in five years are the ones who will get started quickly, who will ground their models with data [and] upskill their workforce." The cost of waiting, he concluded, may be higher than the cost of experimentation.

"The cost of not acting will be a significant challenge for these companies if they do not accelerate the innovation with AI right here, right now."

- Guruaj Bhat, AI Practice Leader, Google Cloud

Acknowledgements

AMD

AMD is the high performance and adaptive computing leader, powering the products and services that help solve the world's most important challenges. Our technologies advance the future of the data center, embedded, gaming and PC markets. Founded in 1969 as a Silicon Valley start-up, the AMD journey began with dozens of employees who were passionate about creating leading-edge semiconductor products. AMD has grown into a global company setting the standard for modern computing, with many important industry firsts and major technological achievements along the way.

Learn more

Dell Technologies

Dell Technologies offers a portfolio of use-cases with Dell AI for Telecom, combining AI Factories, edge computing, and 5G solutions to transform telecom networks. With open, scalable architectures and strong industry partnerships, Dell empowers operators to optimize networks, enhance efficiency, monetize in the enterprise and deliver smarter, more connected experiences.

Learn more

Google Cloud

At Google Cloud, we're partnering with communication service providers around the world to deliver the AI-Driven Telecom. By harnessing intelligent data, advancing cloud-native networks, and unlocking new monetization models, we are enabling the future of telecom. We are proud of our growing ecosystem that helps drive these innovations forward.

Learn more

Pure Storage

Pure Storage uncomplicates data storage, forever. Pure delivers a cloud experience that empowers every organization to get the most from their data while reducing the complexity and expense of managing the infrastructure behind

it. Pure's commitment to providing true storage as-a-service gives customers the agility to meet changing data needs at speed and scale, whether they are deploying traditional workloads, modern applications, containers, or more. Pure believes it can make a significant impact in reducing data center emissions worldwide through its environmental sustainability efforts, including designing products and solutions that enable customers to reduce their carbon and energy footprint. And with a certified customer satisfaction score in the top one percent of B2B companies, Pure's ever-expanding list of customers are among the happiest in the world.

Learn more

Spirent

Spirent Communications, now part of Keysight (NYSE: KEYS), is a leading global provider of automated test and assurance solutions for networks and positioning. The company provides innovative products, services and managed solutions that address the test, assurance, and automation challenges of a new generation of technologies, including 5G, AI, cloud, autonomous vehicles and beyond. From the lab to the real world, Spirent helps companies deliver on their promise to their customers of a new generation of connected devices and technologies.

Learn more

VIAVI

From testing, assuring, and securing the largest communications networks around the globe, to the coatings and filters that make your car's spatial sensing possible, our technologies have a diverse impact on the world. Discover how we enable new possibilities that touch every area of life.

Learn more

Xsight

Founded in 2017, Xsight Labs is a fabless semiconductor company re-architecting the foundation of cloud infrastructure. Our breakthrough technology delivers exponential bandwidth growth while reducing power consumption and total cost of ownership, enabling next-generation, end-to-end data center connectivity. Our mission is to redefine data center scale and efficiency through innovative architectures that power the connected world of tomorrow.

Learn more

MARKET PULSE REPORT

The AI infrastructure platform shift

RCR Tech
NEWS. INSIGHT. IMPACT.