RCR Wireless News
INTELLIGENCE ON ALL THINGS WIRELESS

THE AI POWER PLAY

Data center infrastructure is being reinvented for the intelligence era

By Sean Kinney

In partnership with

SUPERMICRO

TELCO AI FORUM

AI INFRASTRUCTURE FORUM

CONTENTS

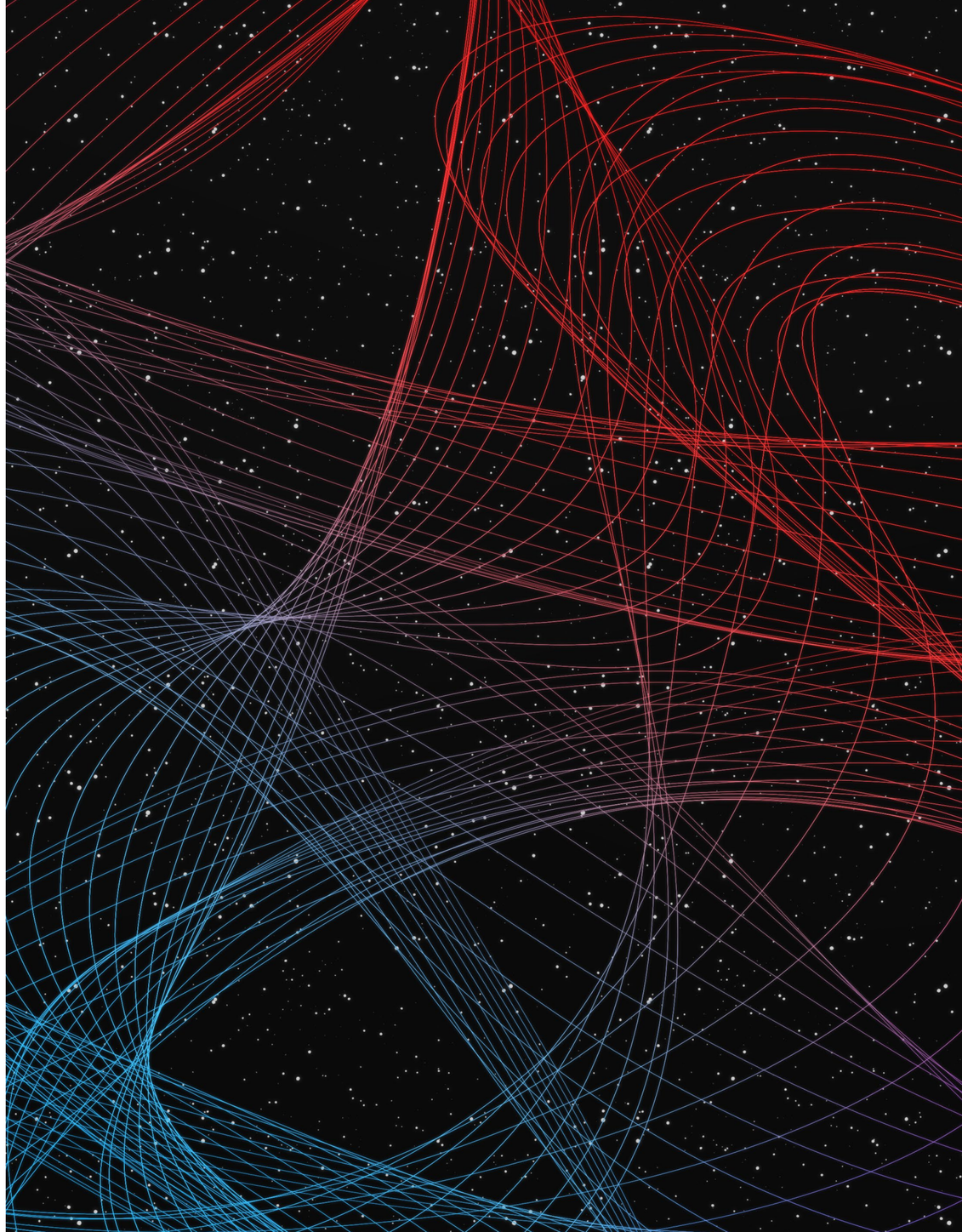
Introduction

Artificial intelligence (AI) is reshaping the economics, purpose, and scale of digital infrastructure. But as trillions of dollars flow into new data center capacity and compute platforms, access to power has emerged as a clear constraint. And it's not a simple constraint with a simple solution; rather, it's one variable in a multi-dimensional system that governs how fast AI can scale and deliver value.

This report looks at how power has become the bottleneck and blueprint for AI data center evolution. It explores the difference between building frontier-scale infrastructure for training general-purpose models and deploying energy-optimized systems to serve inference at the edge. And it highlights how solving for scale means solving for alignment of servers and megawatts, but also alignment of delivery timelines, ecosystem coordination, and trust.

Key takeaways

- **Power is a strategic constraint:** Adding power capacity isn't a magic wand for driving AI infrastructure scale. To realize the promise of an AI-enabled economic revolution, we must acknowledge power as a limiting variable embedded in geographic, political, regulatory, and technology cycles.
- **Training and inference require different energy strategies:** Hyperscale AI factories demand dense, power-hungry clusters. But delivering AI to enterprises and edges, in contrast, requires distributed infrastructure designed to optimize performance per watt across varied environments.
- **Coordination is as critical as compute:** Solving for power at scale means aligning stakeholders across OEMs, operators, regulators, utilities, and more. AI infrastructure is being industrialized within a globally interdependent system.



Situation—from AI boom to power bottleneck

Dell Technologies Engineering Technologist Tim Shedd, who has worked in cooling, power, and thermal management for compute infrastructure for more than two decades, reflected on the evolving power demands of data centers. “Power has always been a challenge,” he said, “but there was always some seemingly infinite government fund to provide that.” During the Obama administration, he said, “I observed a real concentration of understanding…the implication of the growth of power in the data center.” Now, he says, the impact is more diffuse, shaped by both enterprise and hyperscale trends.

The perception of AI as the enabler of the next wave of economic and industrial revolution has prompted long-term, multi-trillion-dollar investments in the modernization of existing data centers and the construction of new ones. In January, AFCOM released its annual State of the Data Center report. Jumping straight to the conclusion: “Every data center is becoming an AI data center.” Rack density has more than doubled since 2021, with nearly 80% of operators expecting density to rise even further due to AI and high-performance workloads. That shift is prompting widespread adoption of liquid cooling, advanced airflow optimization, and AI-based environmental sensors.

This wave of infrastructure investment has also placed the spotlight squarely on power from generation and sourcing to delivery and intelligent management. Without power, there is no AI.

Big picture, AI-based power demand is reshaping energy strategies. Data center electricity consumption is expected to double in the coming years, driving a shift toward ever more efficient hardware all the way down to the silicon level, a laser focus on power and thermal optimization strategies, and even recursively using AI to model and manage how data centers can continue delivering compute within constrained power envelopes.

Since late 2021, more than 100 megawatts of new data center construction has been added each month. The US colocation market alone has more than doubled in size over the past four years. Vacancy rates are at record lows, and rents are steadily increasing. Meanwhile, 55% of operators cite solar as the most viable renewable energy source, with growing interest in nuclear power as well.

Globally, data centers consumed an estimated 240–340 TWh of electricity in 2022—about 1% to 1.3% of total global electricity use, according to the International Energy Agency.

For years, that footprint grew at a modest pace due to hyperscale efficiencies. But now, both the IEA and IDC project a dramatic shift: global data center electricity consumption could double between 2022 and 2026. IDC forecasts that AI data center capacity alone will expand at a 40.5% CAGR through 2027, with total global electricity consumption by data centers projected to hit 857 TWh by 2028.

The perception of AI as the engine of the next economic revolution has triggered a supercycle in data center investment: not only to modernize existing facilities, but to construct entirely new infrastructure purpose-built for AI. According to AFCOM, 80% of data center operators are planning major capacity increases to support AI workloads; this represents an alignment of capital, engineering, and urgency that rivals any infrastructure wave of the past century.

But with this boom comes a bottleneck. AI is radically altering the energy equation that underpins digital infrastructure. Power demand is expected to double within just a few years, and rack density has already done so. Data centers now face a new design imperative: how to deliver more and more performance within constrained energy and thermal envelopes.

Historically, the industry operated on an inefficient paradigm: for every watt of computing power, approximately one watt was spent on cooling. “It was almost one-to-one…that was not sustainable,” Shedd said. The rise of cloud and high-performance computing prompted major gains in power efficiency—but those gains had limits. “As long as we got better at the cooling, we had more power for the compute…I think what’s happened now is we’re getting to where we’ve squeezed that orange pretty hard.”

Complication — whether for training or inference, AI diffusion requires more power

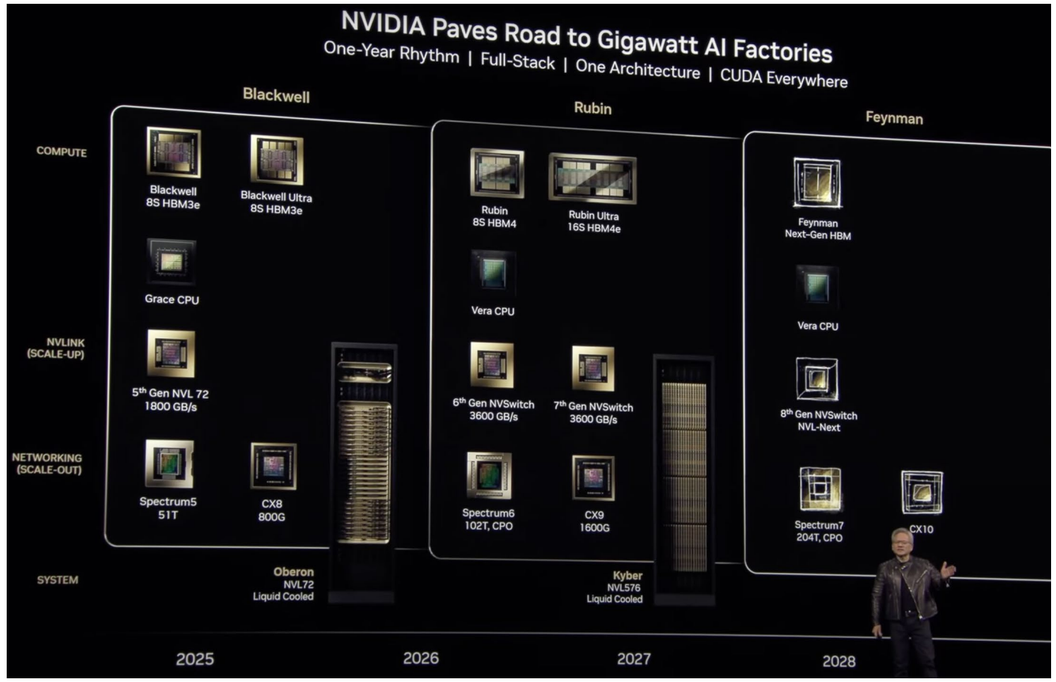
AI is necessarily redrawing the blueprint for data center infrastructure. Power densities are surging with no signs of slowing. During his GTC keynote, NVIDIA CEO Jensen Huang projected a staggering 600 kW per rack power draw for the company's upcoming Rubin Ultra NVL576 platform, a leap well beyond today's rack averages. These numbers grab attention, but they don't tell the whole story.

The Rubin-class platforms are optimized for training frontier-scale large multimodal models. These are general-purpose intelligence engines that will later be tuned and optimized for more domain-specific use cases. And herein lies a kind of circular logic at the heart of AI's infrastructure challenge: to create general models, we need unprecedented scale; to apply them meaningfully, we need customized deployments that are often at the edge, or on-premises, or tightly integrated with enterprise workflows. Either way, more infrastructure is needed. Either way, more power is consumed.

"I've said before," Huang said, "that I expect data center buildout to reach a trillion dollars, and I'm fairly certain we're going to reach that very soon…We're building AI factories and AI infrastructure. It's going to take years of planning. This isn't like buying a laptop. This isn't discretionary spend. This is spend that we have to plan…you need the land, the power, the engineering teams. You've got to lay it out two, three years in advance." He then offered the crux of the issue: "Everything is related ultimately to energy."

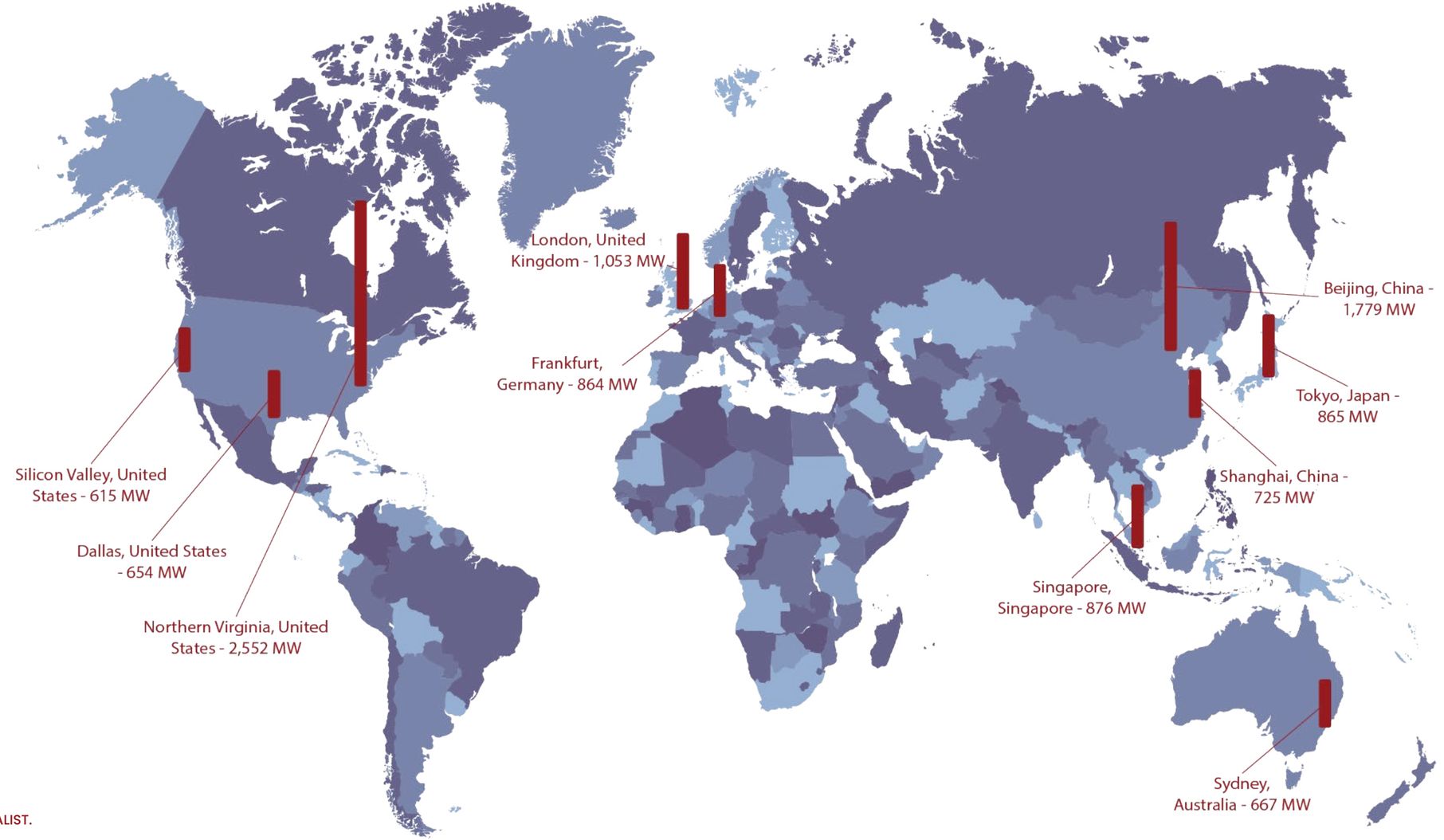
But if Huang articulates the frontier model while acknowledging the long tail of enterprise-specific adoption, it's Dell's Shedd who articulates its translation into the broader market. "It really is customer specific," he said of rack power densities. "We have to enable them. But when we talk about deployment—the scale of the deployment, even the scale of the power consumption—most of it is not going to be 100 MW-plus data centers." Instead, Shedd explained, the real complexity lies in supporting dense compute wherever the data is generated which often means colocation, edge, and hybrid environments. "We're aware of much higher densities coming. We do not currently see a limiting factor. But we also need to prepare to enable our customers to deploy where the data is as effectively and efficiently as possible."

IMAGE COURTESY OF NVIDIA.



This is the paradox of progress in AI infrastructure. The larger the vision, the more granular the execution needs to be, and the faster the technology accelerates, the more time and coordination it demands. Regardless, as AI workloads proliferate and energy remains a primary constraint, the goal (the solution to scale) is in solving for megawatts and milliwatts to power both hyperscale clusters and lightweight edge deployments.

The top 10 global data center markets by megawatts



DATA COURTESY OF VISUAL CAPITALIST.

Question — power is the problem, but what are we really solving for

At a high level, everyone agrees: a primary constraint to scaling AI infrastructure is power. But solving for it is not as simple as adding megawatts. It means solving for delivery timelines, distribution, field services, location, precision, thermals, uptime, and workload placement. In practice, the constraint expresses itself differently depending on whether you're designing a national energy strategy, deploying inference at the edge, or training frontier models.

As Shedd put it, "I've got one hammer…maybe two. That's power and cooling…Every watt into the building has to leave as heat. There's no way around that. That's a challenge." Even small improvements can have large ripple effects. "The larger that number [power budget], the larger the lever we have to provide value to our customers." For Shedd, Dell's eRDHx system is one such lever. "This is one of our answers to allow for extremely efficient removal of 100% of the heat that's generated in the rack."

Peter Downey, senior vice president with Worley, echoes this sentiment from the perspective of industrial engineering and procurement. "A big part of what we do is managing cost. When we shop for data center capacity, the cost of electricity is hugely variable from market to market." He explained that while Worley operates in 48 countries, the core opportunity lies in optimizing heat, cooling, and energy generation via simulation: "We have, I would say, it's a mandate…to make our customers much more energy efficient."

Downey points to two compounding imperatives: engage the ecosystem early, and shift the efficiency conversation from uptime to consumption. "Compute providers that don't maintain 100% uptime but have seriously optimized energy consumption" are becoming more attractive. The idea of 500 kW racks sounds scary, but "in our world, in all the countries we operate in, these are going to be small installations…We're not developing foundation models." The goal now isn't minimizing data; instead it's better managing data "so we can feed our models." That includes techniques like workload scheduling. "As somebody who's paying for all of this in data centers and using it, I can tell you that would sway our decision-making in a big way."

Mike McDonald, vice president of product at Fluidstack sees the question as one of enabling reliable, performant infrastructure

PETER DOWNEY OF WORLEY, RIGHT, AND MIKE MCDONALD OF FLUIDSTACK, SECOND FROM RIGHT, DISCUSS THEIR OWN AI JOURNEYS DURING DELL TECHNOLOGIES WORLD. IMAGE COURTESY OF DELL TECHNOLOGIES.

at scale without compromising sustainability. “Deploying large-scale H100 clusters was not boring,” he said, recalling “hand-racking” GPUs for some of the world’s largest labs. Beyond delivering pure performance, the trickier challenges are reliability and resiliency in hyperscale environments. “There was a lot of infrastructure that was sort of just under the surface that enabled years two and beyond.”

That’s where AI might help itself. McDonald sees potential in agentic site reliability engineers who can assist humans in debugging and managing hardware. “We haven’t quite given [agents] access to servers just yet but I think we’re very close to giving them access to parts of production.” As inference becomes more dominant, “cheap, abundant tokens will start powering…the enterprise.” That shift introduces yet another energy profile defining smaller clusters, but more of them. “We’ve been focused increasingly on how do we get the scale necessary to provide that compute efficiently and in a climate-aligned way.”

That means siting facilities not just where customers are, but where the power is clean and abundant. Fluidstack has partnered with Borealis Data Center to deploy exascale clusters across Iceland and Europe, leveraging hydro and geothermal energy in cold climates. In France, the company is investing €10 billion into a 1-gigawatt facility powered by nuclear energy. “This €10 billion agreement with Fluidstack embodies my ambition,” said French President Emmanuel Macron. “We must not slow down because the world is accelerating and the battle for innovation is happening now.”

The point is that solving the power question clearly means acknowledging that it’s the overarching constraint, but it’s a variable constraint that has to be considered as part of a larger, more complex system of capital, governments, operators, and vendors.

A DELL FACILITY THAT ASSEMBLES SERVER RACKS FEATURING NVIDIA BLACKWELL GPUS. IMAGE COURTESY OF NVIDIA.

Answer — solving for power is really about solving for alignment

If the current supercycle of AI infrastructure deployment is all about manufacturing intelligence, it's perhaps appropriate to apply a management framework developed for the manufacturing sector to understand how key AI players can deliver the compute capacity needed to translate AI infrastructure investments into broad economic gains.

In his 1984 book The Goal, Eliyahu M. Glodratt introduced the Theory of Constraints framework described by the eponymous Theory of Constraints Institute as built around the idea that "every system has a limiting factor or constraint. Focusing improvement efforts to better utilize this constraint is normally the fastest and most effective way to improve profitability." The methodology suggests that by identifying and leveraging a particular constraint, "Organizations can achieve their financial goals while delivering on-time-in-full…to customers, avoiding stock-outs in the supply chain, reducing lead time," and simultaneously gain "better control over operations, less inventory, reduced conflicts between team member[s] and drastically reduced firefighting." And, generally, once one constraint is addressed, another will emerge; rinse and repeat.

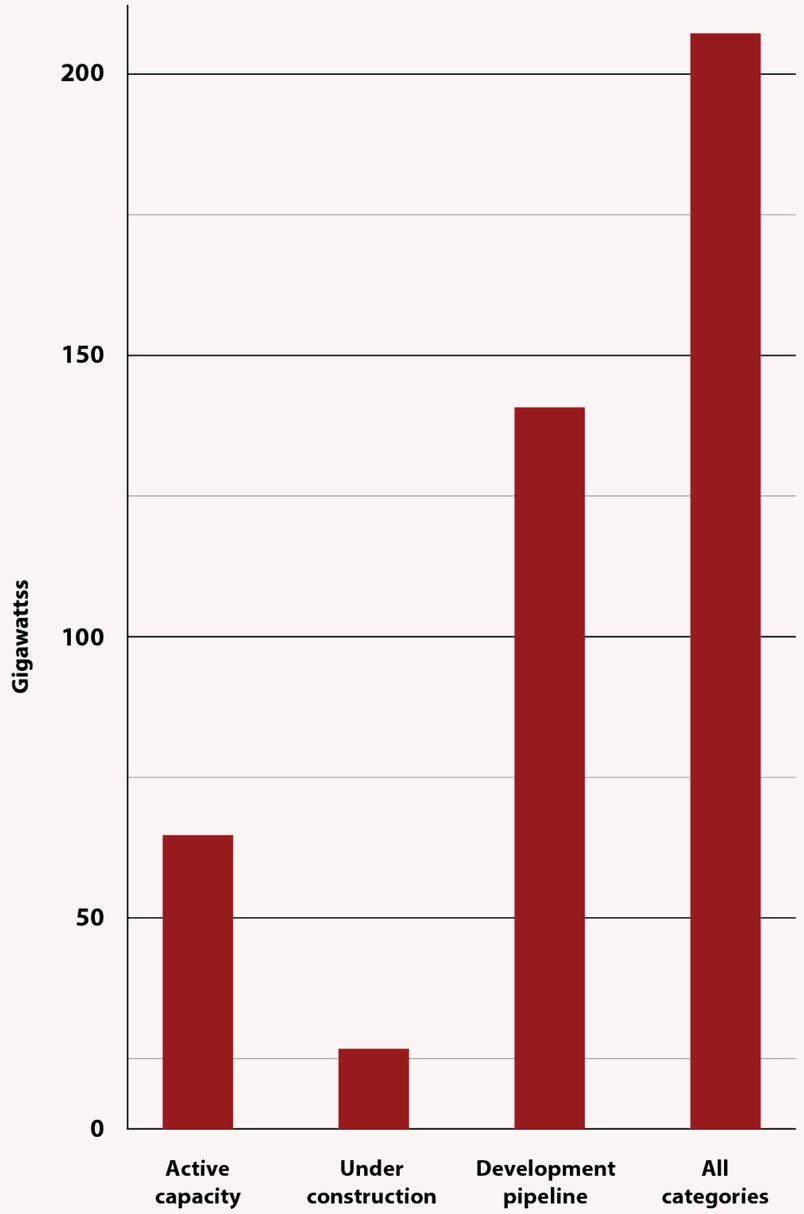
So what is the primary constraint faced by AI infrastructure providers? In a new report, Infrastructure Masons estimates that there's currently 55 Gigawatts of "active data center power capacity worldwide," 15 Gigawatts more "under construction and at least 135 [Gigawatts] in the development pipeline." Looking broadly at digital infrastructure, the report authors write that, "Access to power is the top industry challenge. AI amplifies this challenge." And it's not a static target. "Today, 90% of power growth is for AI model training. The use of these trained AI models, called inference, is forecast to become 90% of AI workloads by 2030…This will transform how digital infrastructure is used and spur data center development at the near and far edge for lightningfast interactions with people and machines in every community around the world."

The Theory of Constraints dictates "five focusing steps" meant to establish a "process of ongoing improvement." Those steps delineated in the graphic on page 10.

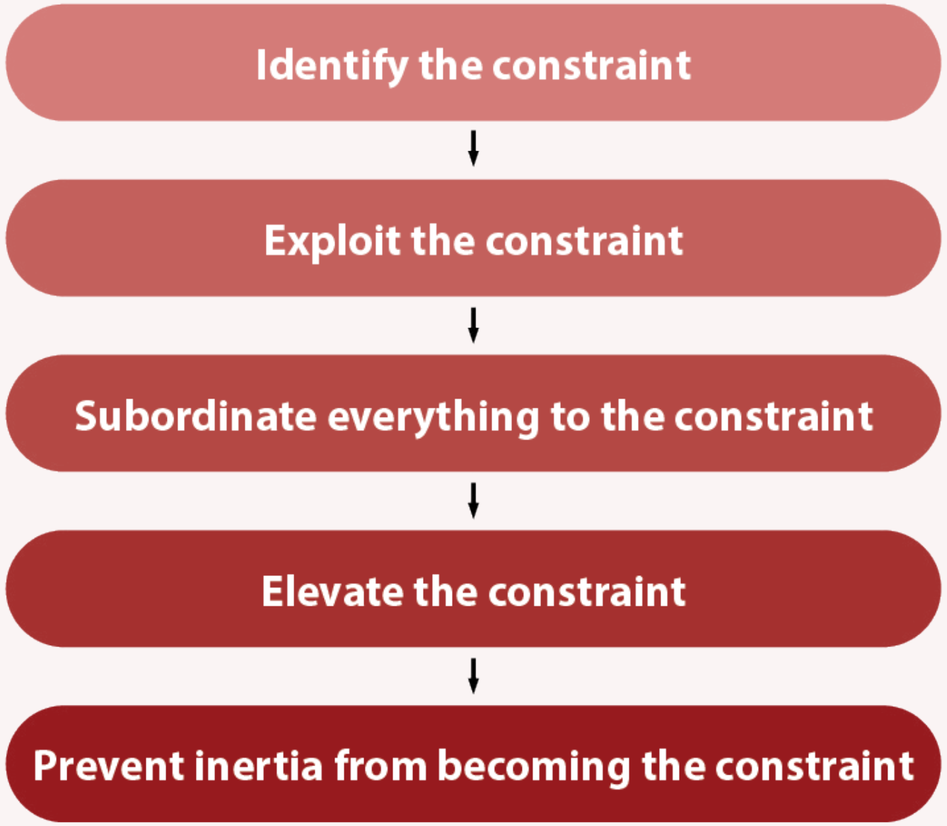
In the context of AI infrastructure, the constraint is power. Exploiting the constraint means squeezing as much value as possible from the

DATA COURTESY OF AFCOM.

Powering present and future data centers



The five focusing steps



existing power supply—design within existing power envelopes, optimize high-revenue workloads based on power availability, and scheduling workloads to maximize compute per Watt. Subordinating everything else to the constraint means aligning construction timelines with power delivery schedules, matching rack deployments and GPU deliveries with powered sites, and shaping sales and customer on-boarding around power limitations. Elevating the constraint means engaging earlier with utility providers, investing in on-site power generation and power purchase agreements (PPAs), lobbying for priority access, and partnering with data center providers who already have power. And avoiding inertia becoming the primary constraint means hammering on the power piece until a new constraint emerges. Then rinse and repeat this process of ongoing improvement.

Beyond these tech-founded strategic pieces that can power the AI era, there's also a bit of a softer skill that came up in numerous conversations with compute builders and buyers — getting better and faster at project management in a complex, multi-stakeholder, global environment while the clock is ticking.

Whether it's one rack, 100 racks, or 1,000 racks, Shedd said, the key is "working together…You cannot do this without excellent deployment and field services. If you can't deploy the hardware at scale quickly, you don't have a business and our customers cannot make money off the hardware they've purchased… These GPUs are so expensive. If they wait… that's money lost." While power is certainly a major hurdle, in the current world of building and operationalizing AI, "As far as where the primary constraint is, it can shift daily."

Solving for scale also means solving for precision. Fluidstack Vice President of Product Mike McDonald elaborated on this point. There are two big issues, he said. "One is just the sheer number of things that need to be right. And there is one right answer. When you cable a super computer, there is one way to do it… The second [issue] is just reliable operation… The laws of physics are harsh, so you're constantly dealing with failure." Infrastructure deployment at scale is less about perfection and more about systemic alignment.

This is why partnership emerges as the real enabler of constraint-based innovation. Dell's CTO of Global Industries David Holmes put it this way: "Partnership is absolutely key to success. So whether that's working with regulators, policymakers, utility companies, technology providers, [or] data center operators…the partnership between those different entities to ensure the…delivery of technology and infrastructure is absolutely critical."

In the end, solving for power is not just about electricity. It's about coordination and trust in pursuit of a common goal. That's what it takes to industrialize intelligence. As it advances, AI won't be trained in theoretical infinity. It will be built, deployed, and sustained within the constraints we solve for today.

THE AI POWER PLAY

Data center infrastructure is being reinvented for the intelligence era

In partnership with

Would you like to be involved with out upcoming reports? Contact the team here.

Find out more about our events