



Insight Paper • September 2026

INSIDE THE MINDS OF
AI INFRASTRUCTURE BUILDERS:
ONE MIND AT A TIME

The AI Infrastructure Imperative



Insights from Griffin Lacey

Solutions Architect,
NVIDIA



As artificial intelligence transitions from research investment to operational infrastructure, the organizations that understand the architectural shift will be better positioned to adapt as capabilities evolve.

Griffin Lacey, Solutions Architect at NVIDIA, offers a grounded, forward-looking account of what that shift requires and why flexibility across the full technology stack matters.

A Rate of Progress That Demands a Structural Response

The pace at which AI capability is advancing has moved well beyond what most enterprise planning cycles were designed to accommodate. Lacey describes the current environment well: “Every day there is a new foundation model released. Every week the capabilities of these models are something that was not even imaginable the week before.”

For technology leaders, the implication is that infrastructure must be able to absorb rapid change. Rather than optimizing around a single generation of hardware or one fixed model architecture, Lacey’s perspective points toward platforms and ecosystems designed to adapt as new capabilities, algorithms, and workloads emerge.

Lacey situates the NVIDIA full-stack philosophy as follows: AI is not a hardware problem, a software problem, or a systems problem in isolation; performance depends on how well those layers work together. “Our latest GPU architectures, like NVIDIA Rubin, are carefully co-designed with our NVIDIA CUDA® platform and accelerated libraries so that the software can take full advantage of the hardware,” he explains. The value of the platform, in this account, comes from that coordination across the stack.

The Shift to Inference: New Design Criteria, Broader Access

For much of modern AI’s recent development, infrastructure design was organized around training. The primary question was throughput: how quickly a large model could be trained on a given cluster. That framing shaped the way infrastructure was built and optimized.

That emphasis is now shifting. “When I started my career in AI, the sentiment was that everyone needs to train their own model. Now there is a real shift, we can use foundation models that other people have trained, and either use them off the shelf or fine-tune them to a specific application” says Lacey. The consequence is a reordering of infrastructure priorities. Training emphasizes throughput. Inference brings renewed attention to latency and the speed at which users receive a response.

We think of AI as a full-stack problem – not just hardware, not just software, not just systems, but everything working together in unison.

Griffin Lacey

Solutions Architect, NVIDIA Canada



With AI inference, there is a renewed emphasis on latency. When you talk to an AI, you want to feel like you're talking to a human, so you expect an answer very quickly.

That requirement for ultra-low latency is a very different type of consideration for AI infrastructure design.

Griffin Lacey

For organizations building AI applications, foundation models can reduce the need to train every model from the ground up.

The infrastructure question therefore changes as well: not only how to train efficiently, but how to serve AI reliably and responsively for real users.

Customer-facing AI applications require response times that feel immediate. When inference requests must travel long distances to reach a data center, geography becomes part of the infrastructure design problem. Lacey explains: "Having geographically distributed data centers is now a key requirement if you want to drive low latency across your entire global community of users." His point is straightforward: as inference becomes more important, distributed infrastructure becomes an increasingly important consideration for serving users with consistently low latency.

The Next Capability Threshold

Among the near-term developments Lacey regards as most consequential is the expansion of context windows to a scale that changes what AI systems can reason about. Current limitations on context length constrain AI usefulness in ways that practitioners encounter routinely: a model that loses track of prior conversation turns, or that can only engage with a fragment of a codebase rather than its full structure, is working with an incomplete view of the problem.

Lacey points to next-generation hardware as an important enabler of much larger context. He describes the practical implications with precision: "Imagine the ability to have incredibly complicated multi-agent conversations with many tool calls, ingesting large documents, all without ever losing memory. Or, instead of putting code snippets into a prompt, you can put an entire codebase into a prompt and have the AI reason against it."

The significance is not simply a longer prompt. Larger context can allow AI systems to retain more of the information needed to reason across complex conversations, documents, software environments, and multi-step workflows. That opens the door to more capable forms of AI-assisted analysis, software development, and decision support.

Disaggregated Inference: Better Economics Through Architectural Discipline

AI inference can involve workloads with very different infrastructure demands. Lacey distinguishes between parts of the problem that are compute-bound and parts that are memory-bound. His perspective is that treating those workloads differently creates an opportunity to use infrastructure more efficiently.

“With disaggregated architectures we can allocate infrastructure that is compute-heavy to the compute-bound parts of the problem, and infrastructure with more memory performance to the memory-bound parts. Overall, this drives maximum utilization of the infrastructure.”

For Lacey, the economic benefit follows from that better utilization: it can ultimately help drive down cost per token. As AI systems move into wider production use, improving utilization while reducing the cost of inference becomes an important part of making those deployments more efficient at scale.

Disaggregated architectures are one of the best examples of extreme hardware-software co-design, allocating the right infrastructure to the right part of the problem, driving maximum utilization, and ultimately driving down cost per token.

Griffin Lacey

Networking, compute, and storage all need to be designed, operated, and optimized together. If any one of those elements is out of sync with the others, that will quickly become the bottleneck. This is not a series of independent pieces.

This is a total full-stack AI factory problem.

Griffin Lacey

The AI Factory: Unified Design as a System Requirement

One of the clearest changes in Lacey's description of next-generation AI infrastructure is the shift away from thinking about independent servers and instead, about the complete AI factory. It is an environment in which the infrastructure acts and behaves as one large compute engine.

"Partners need to stop thinking about independent servers and start thinking about the complete AI factory design, which acts and behaves like one large compute engine," he states. The distinction is structural. Networking, compute, storage, power, and cooling cannot be treated as isolated decisions when overall performance depends on how they work together. As Lacey notes, if one element is out of sync with the others, it can quickly become the bottleneck.

Compute densification also changes the physical infrastructure requirements. Lacey points to direct liquid cooling as increasingly necessary where traditional air cooling is no longer adequate, and he identifies 800-volt DC power distribution and energy storage for workload volatility as examples of the new design considerations operators may need to address. These shifts challenge traditional assumptions about what a data center looks like and how it operates.

The broader implication is a need for deeper coordination across the organizations and specialists involved in building AI infrastructure. Competency in AI factory design increasingly means understanding not only individual components, but also the dependencies between them.

Digital Twins: Reducing Risk Before Construction Begins

The complexity of AI factory design spans physical layout, cooling, thermal behavior, power distribution, workload dynamics, and the coordination of multiple vendors. The digital twins can help teams test those interactions before they commit to the physical build.

By creating a simulated representation of the planned facility, operators can evaluate design choices and explore how changes in layout, cooling, thermal distribution, and power may affect the infrastructure.

Lacey describes the value: “These are all things that are incredibly expensive to tinker with in the real world. If you can build that digital twin and do this in the simulated world, it allows you to anticipate these problems ahead of time, and often save not just time, but cost.”

Digital twins can also support collaboration among the independent vendors that need to make the AI factory work as a system. Teams can simulate different scenarios and workload conditions in a digital environment before committing those decisions to the physical world.

He also points out why this area is especially relevant to NVIDIA’s capabilities: digital twins combine high-performance computing, physics, and graphics. In his view, that combination makes simulation a powerful tool for anticipating infrastructure challenges that are difficult and expensive to discover after construction.

Algorithmic Progress and the Need for an Adaptable Platform

A consistent feature of modern AI history is that major capability shifts have often been triggered by algorithmic breakthroughs and made practical through the availability of data and accelerated computing. Lacey, who comes from a foundation in AI research, points to AlexNet and the Transformer architecture as pivotal examples. He also emphasizes that AlexNet’s impact depended not on algorithmic progress alone, but on the combination of new ideas, large data sets, and GPU computing that made those ideas trainable at useful speed.

“I believe that looking forward, the future will be defined by the next algorithmic improvements, the next transformer architecture,” Lacey says. He points to neural networks’ long-standing foundations in backpropagation and stochastic gradient descent, and the fundamental building blocks of AI may have the largest impact on what becomes possible next.

That uncertainty makes adaptability especially important. Lacey notes that a new paper or architectural idea can challenge existing assumptions quickly. He describes the NVIDIA CUDA platform as a mechanism for responding to those shifts by accelerating new building blocks across widely used frameworks and libraries. In this framing, platform value comes not only from performance today, but from the ability to adapt as AI itself changes.

Digital twins can help those data center operators, and the independent vendors that all need to work together, simulate different scenarios before they have to commit to building in the physical world. This helps reduce costs and anticipate challenges that traditional AI infrastructure never had to address.

Griffin Lacey

I'm really an optimist in the sense that I believe AI can be used for good. All these applications are not only important to push the science forward, but they are also incredibly important for the good of humanity. And we're seeing a lot of that evidence already.

Griffin Lacey

AI's Contribution to Scientific Progress

Lacey believes in AI's contribution to science and its potential to improve how complex problems are understood and solved. He points to advances in genome sequencing, drug discovery, personalized healthcare, large-scale data analysis, climate and weather prediction, satellite imagery, and capital markets as examples of the breadth of applications already emerging.

He concludes, "I really believe that the capabilities of AI as it relates to fundamental science are completely unbounded." Rather than predicting exactly where the next breakthrough will come from, Lacey emphasizes the range of fields in which AI is already advancing research and the possibility that future achievements may arrive faster than expected.



HOME OF AI.

Contact us at: 5c.ai

Follow us on:



5C.