The next *AI bottleneck*: from memory constraints to infrastructure limits

AI's next bottleneck is no longer just chips or memory, but the infrastructure needed to connect, power, and sustain AI at scale.

Insights

The evolution of artificial intelligence over the past decade has been defined by a sequence of shifting bottlenecks. In the early phase of scaling, progress was primarily constrained by access to computing power, especially GPUs, as companies raced to train larger models.

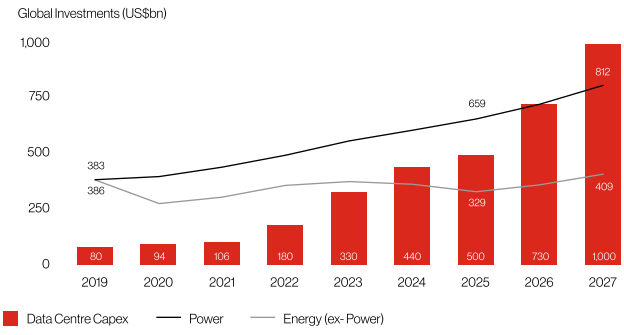
As GPU supply gradually improved, the bottleneck moved toward memory, because larger models need to access and process enormous amounts of data quickly. Today, however, the industry appears to be entering another transition. The next major limitation is no longer confined to the chip itself but increasingly resides at the system and infrastructure level.

This new bottleneck is best understood as the convergence of three interrelated constraints: **data movement, power delivery, and power flexibility.**

First, AI systems must move data quickly between chips, servers, and data centers. Second, dense AI racks require power to be delivered safely and efficiently. Third, electricity systems must become more flexible as real-time AI usage creates more variable demand.

This is why AI capex is expanding beyond semiconductors into energy and infrastructure.

Figure 1 – AI Capex Is Expanding Beyond Compute Into Energy and Infrastructure



Source: IEA, Morgan Stanley Research estimates

From chip optimization to AI factory optimization

A useful way to frame this transition is that AI is moving from a **chip-centric scaling problem to an AI factory optimization problem**. In this context, an AI factory refers to the full industrial system required to produce AI outputs: GPUs, CPUs, memory, networking, cooling, power conversion, and the data-center infrastructure around them.

This shift is likely to become more visible with next-generation architectures such as NVIDIA's Rubin platform, where the design emphasis appears to be expanding beyond the GPU alone toward full rack-level and system-level efficiency.

Using the Rubin architecture, we can illustrate this broader system approach. Rather than focusing only on a faster GPU, it brings together multiple layers of compute, networking, switching, and power management. This matters because the economics of AI are increasingly determined by **cost per token at the rack and cluster level**, rather than by peak chip performance in isolation.

As clusters become larger and more heterogeneous, the relevant question is no longer simply how many calculations a chip can perform, but how efficiently the full system works together.

This is an important conceptual change for investors. AI scaling now depends not only on chip performance but also on how effectively those chips are integrated into larger systems. Even where semiconductor performance continues to improve, scaling benefits may diminish if supporting infrastructure fails to keep pace.

Data movement is emerging as the first system bottleneck

At the current stage of AI development, even as memory remains important, the efficiency of large-scale systems is increasingly governed by the ability to move data across chips, racks, and data centers.

Frontier model training already requires thousands of accelerators operating in parallel, and the next generation of AI clusters is expected to become even more tightly interconnected. In such environments, interconnect bandwidth and latency become critical determinants of utilization.

This is why optical connectivity is becoming increasingly relevant. Today, many AI systems still rely heavily on copper connections to move data over short distances. As clusters grow, however, copper can become heavier, harder to cool, and less efficient.

Optical links, including emerging approaches such as co-packaged optics, may help move data faster and more efficiently across larger AI systems. The industry debate therefore appears to be shifting from 'when will optical scale-up be needed?' to 'when will adoption become economically and operationally compelling?'.

The implication is that future computational efficiency gains may come less from improvements within the GPU itself and more from the network that **connects those GPUs.** This shifts the bottleneck from processing power to communication capability.

Improving AI performance requires increasing utilization. Even if compute and memory are technically available, insufficient data movement can leave expensive hardware underutilized and reduce returns on capital deployed.

From an investment perspective, the companies that would benefit from this will include the broader interconnect ecosystem, such as optical components, packaging, testing infrastructure, and related enabling technologies.

The deeper constraint lies in power delivery

While data movement is the first system bottleneck, power may prove even more binding.

As AI infrastructure scales, data centres are becoming a major source of incremental electricity demand, driven by both training and, increasingly, inference workloads. This is unfolding against the backdrop of energy systems that in many regions have faced years of underinvestment in generation, grid infrastructure, and transmission capacity.

The result is that AI scaling is no longer constrained solely by how many chips can be manufactured. It is increasingly dependent on whether sufficient power can be delivered reliably and continuously to run those chips at high utilization.

Importantly, the power challenge is not just about the total megawatt availability. We need to consider how electricity is **converted, distributed, and managed** within the AI factory itself. As rack densities rise sharply, traditional power architectures become increasingly inefficient.

This points to a next-order bottleneck, **power delivery architecture**—the ability to move power into ever-denser compute environments efficiently, safely, and economically.

In this context, higher-voltage direct-current designs, such as 800V HVDC, are gaining traction. These systems are designed to deliver power into dense computing environments with lower losses and greater efficiency. NVIDIA's push in this direction underscores a broader transition already underway, enabling significantly higher rack power densities.

This has significant investment implications. The transition toward more demanding AI power architectures should benefit the power ecosystem that includes power semiconductors, converters, electrical equipment suppliers, power management systems, and selected cabling and infrastructure providers.

In particular, it creates a more constructive backdrop for areas such as silicon carbide (SiC) and gallium nitride (GaN) based power solutions, advanced power conversion platforms, and AI-specific rack power systems.

Inference changes the nature of power demand

Importantly, the character of AI power demand is also evolving. The next phase of AI deployment is likely to be shaped not just by large, scheduled training workloads, but by a growing layer of real-time inference demand.

Training is episodic and concentrated; inference is more continuous, more user-driven, and more vulnerable to spikes and short-term volatility. As AI becomes embedded in search, enterprise applications, digital advertising, automation, and agentic workflows, the demand profile becomes more dynamic and less predictable.

That distinction matters because it changes the infrastructure problem. The challenge is no longer simply securing enough steady-state generation to support training clusters. It is increasingly about designing a system that can respond to sudden and variable load fluctuations without excessive overbuilding of fixed capacity.

In other words, the power bottleneck becomes not only one of supply, but also one of **responsiveness and flexibility**.

This marks a fundamental shift in how AI bottlenecks should be conceptualized. Semiconductor bottlenecks could largely be addressed within the digital stack. Energy, by contrast, introduces harder physical-world limits.

Data centers require uninterrupted, high-quality power, yet electricity systems must increasingly contend with intermittency, transmission bottlenecks, and balancing challenges. AI therefore becomes constrained by the weakest link in the broader energy value chain.

Why energy storage systems matter more than before

This is where energy storage systems (ESS) move closer to the centre of the AI investment case. If inference demand creates short bursts of electricity usage, storage can help absorb those spikes, provide peak shaving, improve power quality, and reduce the need to build permanent capacity for occasional demand peaks.

Storage is therefore no longer merely a supplementary clean-energy technology; it is increasingly a strategic component of AI power architecture.

A useful way to think about storage is as a form of **power inventory**. It allows excess or lower-cost electricity to be stored and deployed when demand surges, pricing rises, or system flexibility becomes more valuable. This changes the economics of AI infrastructure.

Instead of building fixed generation and grid capacity to satisfy the most extreme hours of demand, operators can increasingly use storage to smooth load volatility, improve equipment utilization, and defer expensive system upgrades.

The importance of storage therefore goes well beyond simple levelized cost comparisons. Its real value lies in **infrastructure deferral, deployment optionality, and capital flexibility**.

Conventional generation assets are capital intensive, slow to permit, and difficult to resize once built. Storage, by contrast, can often be deployed incrementally and scaled in line with realized demand. In a market where AI demand growth remains powerful but uncertain in shape and timing, that flexibility is particularly valuable.

The story, therefore, is not that one bottleneck has disappeared and another has replaced it, nor is it that AI is riddled with bottlenecks. Rather, AI scaling is a layered infrastructure problem. Compute and memory remain essential, but their value increasingly depends on the infrastructure that connects them, powers them, cools them, and keeps them operating at high utilization.

A broader AI capex cycle is emerging

Once we recognise the multiple layers contributing to AI development, we can see how AI capex is likely to broaden meaningfully beyond semiconductors.

The first leg of the AI cycle was dominated by accelerators and memory. The next leg is increasingly likely to be shaped by expenditure on networking, optical interconnect, power conversion, electrical architecture, thermal management, storage, and grid-facing infrastructure.

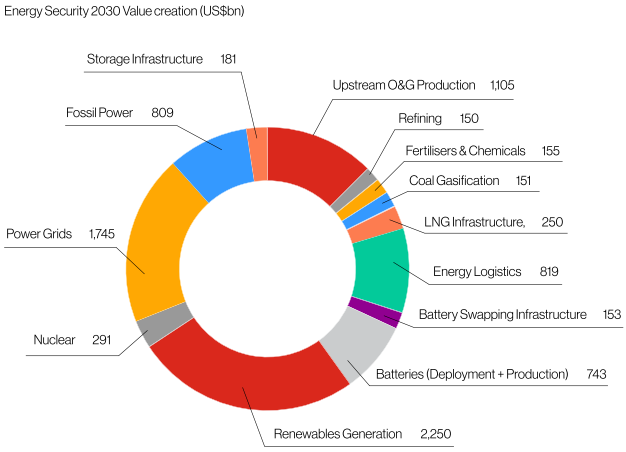
This does not mean semiconductors become less important. Compute, memory, and networking remain central to AI scaling. But the marginal bottleneck—and therefore an increasing share of marginal investment—appears to be shifting toward the physical systems that allow those chips to operate efficiently at scale.

AI is no longer simply a chip story; it is a systems engineering story and, increasingly, a power systems story.

This also has geographic implications. Regions with abundant, reliable, and competitively priced electricity are likely to attract a disproportionate share of AI-related capital.

Over time, data-centre siting decisions may be influenced not only by fibre connectivity and tax incentives, but also by grid readiness, storage economics, renewable integration, and the ability to deliver resilient high-quality power into dense compute environments.

Figure 2 – AI scaling depends on the entire energy value chain, not a single constraint



Source: Morgan Stanley Research Estimates

Investment implications

For investors, the key takeaway is that the next AI bottleneck represents a structural transition from **in-chip constraints to system-wide constraints**. The relevant beneficiary set therefore broadens beyond the traditional AI winners.

The most attractive opportunities may lie where AI-related demand is visible, capacity is constrained, and revenue conversion is not yet fully reflected in expectations.

While GPUs, memory, and advanced packaging remain central, increasing attention should also be paid to:

- **optical and interconnect enablers**, that improve data movement efficiency,
- **power delivery and conversion players**, especially those exposed to higher-voltage AI data-center architectures,
- **energy storage and flexible infrastructure providers**, that help manage more variable inference demand and load management,
- **thermal management, electrical equipment, and integration specialists**, that support higher rack densities.

In that sense, the next phase of AI capex may increasingly flow into the intersection of digital and physical infrastructure.

However, investors should recognize that part of this thesis is already well priced. Many companies exposed to optics, power management, data-center electrification, and cooling have benefited from elevated expectations around AI infrastructure capex.

The key question is therefore not whether the theme is real, but how quickly revenues materialize and whether current valuations already discount that growth.

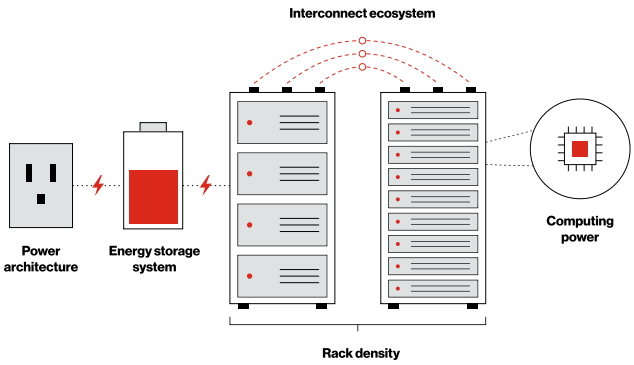
In 2026, revenue contribution may still be relatively limited for some emerging technologies, particularly next-generation interconnects and new power delivery architectures. It will be important to monitor production ramps, qualification cycles, order conversion, and the translation of technology roadmaps into revenue.

That said, signals from NVIDIA and its supplier ecosystem remain encouraging. The key investment distinction is between companies merely exposed to a well-understood theme and companies where the path from roadmap to revenue remains underappreciated.

This also has geographic implications. Regions with abundant, reliable, and competitively priced electricity are likely to attract a disproportionate share of AI-related capital.

Over time, data-centre siting decisions may be influenced not only by fibre connectivity and tax incentives, but also by grid readiness, storage economics, renewable integration, and the ability to deliver resilient high-quality power into dense compute environments.

Figure 3 – Beyond the chip: the ecosystem powering AI



Conclusion

The next AI bottleneck is no longer just about faster chips or larger memory. It is increasingly about whether data can move fast enough across the system, whether power can be delivered efficiently into dense AI racks, and whether electricity systems can respond flexibly to a new generation of volatile inference demand.

What is emerging is a broader convergence between digital infrastructure and physical infrastructure. AI scaling is no longer an isolated technological challenge; it is embedded within the limits of networking, power architecture, and grid flexibility.

The future trajectory of AI will therefore depend not only on continued advances in semiconductors, but also on the expansion and modernization of the infrastructure that connects, powers, and sustains those semiconductors over time.

In this new paradigm, the limiting factor is not simply how powerful the machine is, but how effectively the real-world infrastructure connects, powers and sustains that machine.