Goldman Sachs Global Institute

Tracking Trillions: The Assumptions Shaping the Scale of the AI Build-Out

APRIL 2026

George Lee
Co-Head, Goldman Sachs Global Institute

Lucas Greenbaum
Vice President, Goldman Sachs Global Institute

Executive Summary

The AI CapEx debate is usually framed as a demand-side question—will adoption justify the spend?—but the size of the investment itself is not a single, fixed number. It is highly sensitive to a small set of assumptions about how the infrastructure itself is built and renewed.

Four assumptions are most impactful in determining the scale of the build-out:

1. The economic useful life of AI silicon, where small shifts in replacement cadence move cumulative spend by hundreds of billions
2. The cost and complexity of next-generation data centers, which are rising as AI workloads push power density higher and system integration deeper
3. The chip and architecture mix, whose impact depends on whether compute demand is elastic (reshaping margins) or inelastic (reshaping totals)
4. Elongation from power, labor, and equipment bottlenecks, which in stress scenarios can feed back into demand-side doubt

Several widely discussed dynamics matter for returns, volatility, and value distribution across the ecosystem but do not materially change the aggregate scale of capital required.

Current estimates of the ultimate scale of the AI build-out—regardless of the demand side—are far more conditional than they appear. For investors and operators, critical questions remain: What fundamental assumptions do we have about the future, and how resilient to changes in those assumptions are our plans?

This analysis is a scenario-based framework intended to explore how different infrastructure assumptions may affect aggregate capital requirements, not a forecast of future spending.

Framing the Question

A single AI query feels weightless—a question typed, an answer returned, no moving parts in sight. But the progress of AI rests on a deeply physical edifice: millions of processors, hundreds of thousands of kilometers of cabling, industrial cooling systems, and power demands that rival those of midsize countries. Better understanding of the complexity of that physical infrastructure—and the assumptions upon which its build-out rests—should inform how we think about the scale, durability, and risks of today's AI capital expenditure boom.

The scale of these expenditures is enormous. Estimates of $4 trillion to $8 trillion of total capital investment over the next five years have featured prominently in recent market commentary. That capital is used to buy new chips, build new data centers, and construct new power, all in an effort assemble sufficient computing infrastructure to meet the moment. Debates about whether this figure is "too high" are usually framed around a demand-side question: Will AI adoption and monetization justify the spend?

The scale of required investment for the AI build-out is itself more uncertain than commonly assumed. Estimates rest on a number of assumptions that, if changed, can significantly increase or decrease the amount of capital required.

But there is an equally important supply-side unknown. **The scale of required investment for the AI build-out is itself more uncertain than commonly assumed. Estimates rest on a number of assumptions that, if changed, can significantly increase or decrease the amount of capital required.**

Not all assumptions matter equally in this equation. A small number of assumptions determines how much capital must ultimately be deployed to build AI infrastructure, while other assumptions—despite commanding significant attention—primarily influence timing, monetization, or the distribution of returns.

The most critical assumptions for the level of capital expenditure required for the AI build-out include the following:

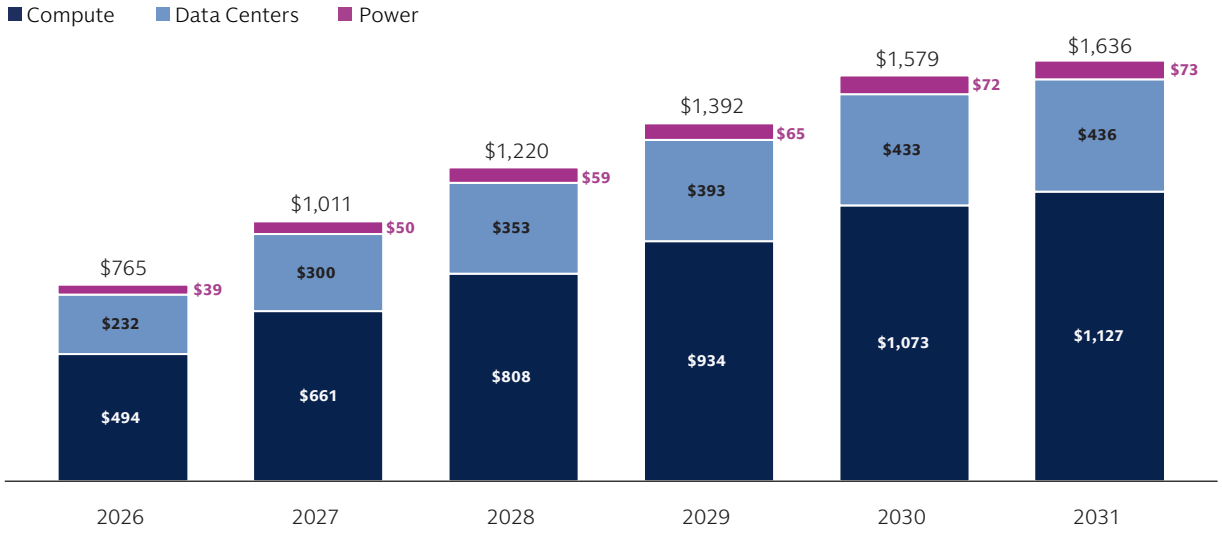
- The economic useful life of AI chips
- The cost and complexity of building next-generation data centers
- The way chip architectural choices translate into system-level costs
- The elongation of the build-out due to physical and institutional bottlenecks

Much of the broader debate focuses on dynamics that matter for returns but do not materially alter the amount of capital that must be deployed. This analysis examines these assumptions and suggests a framework for understanding which changes would push the headline capital expenditure figures higher or lower than current estimates.

Baseline Estimates

Baseline aggregate AI CapEx estimates (bn)

~$7.6tr of capital between 2026 and 2031 across compute, data centers, and power



Source: Goldman Sachs Global Institute, Goldman Sachs Global Investment Research NVIDIA projections (as of March 3, 2026)
Note: Forecasts and expectations are based on material assumptions subject to change. Assumes NVIDIA accounts for 75% of total compute spend in each period. Assumes 5% YoY compute growth past the projection period (2031). Uses VR200 (Rubin) chip as baseline spec ($80.5K per GPU [incl. node costs] and 3,000 W per package) across all years. Assumes 1.2 PUE, $15mn per MW for data centers, and $2,500 per kW for new power. Assumes 15% of required data center space is brownfield (i.e., excluded from calculation) in 2026, growing to 30% in 2031. Totals may not sum due to rounding.

We begin with a baseline model that projects the total scale of AI infrastructure investment implied by today's chip sales estimates. We anchor this baseline to NVIDIA's forward data center revenue Wall Street estimates as a proxy for prevailing expectations around XPU (GPU and other accelerators) deployment and then infer the associated requirements for data centers, power, and supporting infrastructure. This approach does not attempt to forecast AI adoption or end-market demand; rather, it provides a consistent reference point against which we can test how different supply-side assumptions expand or contract the overall scale of investment.

The baseline model implies $765 billion in annual AI CapEx in 2026, growing to $1.6 trillion in annual CapEx in 2031.

These figures include a variety of components necessary for the AI build-out. The core unit of AI infrastructure is the accelerator—a processor purpose-built for the parallel computation that AI workloads demand. Today's leading systems, such as NVIDIA's GB300 NVL72, pack 72 of these processors into a single rack, connected by high-speed backplanes and linked across facilities by hundreds of thousands of kilometers of cabling. These systems generate enormous heat, requiring industrial-scale liquid cooling. And all of it sits within data centers equipped with dedicated power delivery, redundancy systems, and grid or behind-the-meter generation. Together these layers account for baseline estimates that anticipate roughly $7.6 trillion of cumulative CapEx between 2026 and 2031. The key question is, how might changes in the useful life of silicon, the cost and complexity of data centers, the mix of chip architecture, or the pace at which physical bottlenecks persist push that figure materially higher or lower?

The key question is, how might changes in the useful life of silicon, the cost and complexity of data centers, the mix of chip architecture, or the pace at which physical bottlenecks persist push that figure materially higher or lower?

Assumption 1
The Economic Useful Life of AI Chips

AI accelerators (GPUs, ASICs, etc.) are the engines of AI infrastructure, and large-scale data centers house hundreds of thousands of these chips. These devices have a useful life—typically estimated at four to six years—bounded by physical degradation on one side and economic obsolescence on the other, as each new generation delivers step-change improvements in performance. Useful life of silicon chips is the single most influential variable in determining the scale of cumulative AI infrastructure investment.

Unlike other major components of the stack—data center buildings, which are typically depreciated over roughly 20 years, or power infrastructure, which often spans 25 years or more—AI silicon turns over on much shorter cycles. This fact, paired with its high cost per unit, is what makes the silicon replacement cadence so consequential.

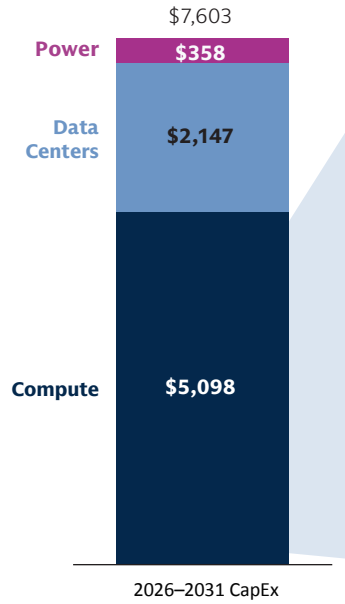
Uncertainty around AI silicon's useful life reflects a core tension: Rapid improvements in performance per dollar between generations of AI silicon push companies to replace hardware quickly, while the growing range of AI tasks means older chips can still deliver value for longer. This tension is sharpened by NVIDIA's unprecedented annual release cadence for GPU architectures, with each generation delivering step-function leaps in capability rather than incremental improvements. Many analysts believe that this mismatch between the annual release schedule and the quantum advances of each new generation makes the prevailing accounting treatment of four-to-six-year depreciation schedules less reflective of the value of the underlying assets.

Because silicon accounts for a large share of AI infrastructure CapEx, small changes in assumed useful life have outsize effects on cumulative spend. Extending average economic life from four years to six years materially reduces the number of replacement cycles over a given horizon—while shortening it has the opposite effect. At scale, these differences translate into substantial changes in aggregate capital requirements—and, critically, into the level of annual depreciation borne by the ecosystem—even as spending on buildings and power infrastructure remains largely unchanged.

Sensitizing the useful life of silicon

Impact on annual compute depreciation from altering silicon useful life from 3 years to 7 years

Baseline 2026–2031 Est. CapEx (bn)



Implied Annual Compute Depreciation by Silicon Useful Life

		Silicon Useful Life and Implied Annual Depreciation (bn)				
Year	Compute CapEx (bn)	3 Years	4 Years	5 Years	6 Years	7 Years
2026	$494	$165	$124	$99	$82	$71
2027	$661	$385	$289	$231	$196	$165
2028	$808	$655	$491	$393	$327	$281
2029	$934	$801	$724	$580	$483	$414
2030	$1,073	$939	$869	$794	$662	$567
2031	$1,127	$1,045	$986	$921	$850	$728
2026–2031 Total	**$5,098**	**$3,989**	**$3,482**	**$3,017**	**$2,597**	**$2,226**

Source: Goldman Sachs Global Institute, Goldman Sachs Global Investment Research NVIDIA projections (as of March 3, 2026)
Note: Forecasts and expectations are based on material assumptions subject to change. Assumes straight-line depreciation and no terminal value for GPUs. Totals may not sum due to rounding.

To illustrate: A single accelerator purchased at $50,000 and depreciated over five years carries $10,000 per year in depreciation expense. However, if that chip becomes operationally obsolete or uneconomic to run before the depreciation schedule expires—because a new generation delivers dramatically better performance per dollar—the operator is still carrying the cost of an asset that no longer drives the economic value it once did. Multiply that dynamic across hundreds of thousands of devices, and the risk becomes a threat to the fundamental economics of the AI ecosystem. Accounting statements may reflect orderly depreciation, but operational obsolescence can impose a very different economic reality—and those shifts can arrive abruptly.

But one dynamic that could extend useful lives—and lend support to the prevailing depreciation treatment—is the emergence of a tiered deployment model for AI silicon. Beyond the demand for leading-edge training lies many less performance-sensitive workloads that may be well suited to trailing-edge silicon and benefit from the depreciated cost of such devices—such as certain inference scenarios, edge computing, deployment in emerging markets, and synthetic data generation. Today, the rental price of trailing-edge NVIDIA devices such as A100s and H100s remains high enough to suggest useful lives of five to six-plus years. That could be a consequence of the extreme capacity constraints model providers are operating under today, or it could be a signal about the sustained value of silicon in the AI era—and thus the appropriateness of its prevailing depreciation timelines.

Buildings and power systems are long-lived assets, while AI silicon turns over far more quickly. As a result, assumptions about accelerator replacement cycles can plausibly shift multiyear infrastructure investment totals by hundreds of billions of dollars.

Assumption 2
The Cost and Complexity of Building Next-Generation Data Centers

AI accelerators run inside data center facilities composed of physical components including power distribution systems, cooling infrastructure, and high-speed networking. As AI workloads push power density higher and system integration deeper, the cost to construct a data center in the AI era has risen meaningfully relative to during prior generations of cloud infrastructure.

Several forces are driving this increase. Compared to prior generations, today's AI data centers operate at significantly higher rack densities, requiring advanced cooling solutions, tighter power delivery tolerances, and greater redundancy. As a result, compute, memory, networking, cooling, and power systems are now codesigned rather than layered independently, shrinking failure domains and increasing the consequences of localized outages. Data centers are therefore now increasingly built with tightly coupled, system-like designs.

Evolution in data center specifications

Rapidly increasing scale, complexity, and density

	Silicon (Generation)	Rack Scale	Power per Rack	Prototypical Scale	Thermals
1 Cloud Data Center[1]	x86/ARM	highly variable # of CPUs	5–15 kW	**10s of MWs**	air
2 Retrofit AI Data Center	GPU (Hopper)	8 GPUs	~40 kW	**10s of MWs** (10s of thousands of GPUs)	air
3 Transitional AI Data Center	GPU (Blackwell)	144 GPUs	~130–200 kW	**100s of MWs** (100s of thousands of GPUs)	liquid/air
4 AI "Factory" of the Future[2]	GPU (Rubin/Feynman)	576 GPUs	500+ kW	**>1 GW** (millions of GPUs)	liquid only

Source: Goldman Sachs Global Institute, NVIDIA GTC presentations (March 2025 & 2026) and public disclosures
Note: All data center specifications are prototypical of the approximate scale and composition. Utilizing NVIDIA silicon road map as representative baseline figures.
[1] Cloud data centers come in all shapes and sizes; for simplicity's sake, a "typical" cloud data center is represented.
[2] A representative model of a "scaled-up" AI data center of the future.

Cloud data centers from the 2010s were built to last 15 to 20 years. However, the rate of progress in AI system design suggests that tomorrow's AI data centers may face a very different trajectory, with future requirements bearing little resemblance not only to traditional cloud data centers, but even to the AI-optimized facilities that have been constructed in the last two years. These design shifts translate directly into higher capital costs per megawatt. Whereas traditional hyperscale cloud facilities might have been constructed at roughly $10 million per MW, in current market commentary, next-generation AI data centers increasingly fall in the $15 million to 20 million per MW range, with further upside risk as density and redundancy requirements rise.

Sensitizing data center cost
Impact on data center CapEx from altering data center cost per megawatt (MW) from $11mn/MW to $19mn/MW

	Cost per Megawatt (MW) for New Data Center Construction (bn)				
	$11mn/MW	**$13mn/MW**	**$15mn/MW (baseline)**	**$17mn/MW**	**$19mn/MW**
2026	$170	$201	$232	$263	$294
2027	$220	$260	$300	$340	$380
2028	$259	$306	$353	$400	$447
2029	$288	$340	$392	$445	$497
2030	$318	$376	$433	$491	$549
2031	$320	$378	$436	$494	$553
Total CapEx	**$1,574**	**$1,861**	**$2,147**	**$2,433**	**$2,720**

Source: Goldman Sachs Global Institute, Goldman Sachs Global Investment Research NVIDIA projections (as of March 3, 2026)
Note: Forecasts and expectations are based on material assumptions subject to change. Totals may not sum due to rounding.

Because data centers are built at massive scale, even modest changes in cost per megawatt compound quickly, making $/MW a key driver of total infrastructure investment. But cost per megawatt is only part of the equation. The rate of architectural change also makes it harder to underwrite assets that have historically been quite durable. Today, developers are realizing that data centers designed less than two years ago (i.e., "transitional AI data centers") may be insufficiently provisioned for the next generation of cutting-edge AI chips, given their significant power and cooling demands. When the design requirements of a facility may shift materially within a few years of commissioning—and when ambitious new concepts like novel cooling architectures or entirely new power delivery paradigms could shift the data center model altogether—the long-lived nature of these assets becomes as much a risk as an advantage.

Assumption 3
AI Chip Architectural Choices

AI workloads run on specialized accelerator chips (GPUs, ASICs, etc.), and the mix of architectures used to deliver that compute represents another key lever on total infrastructure cost. Today, most AI compute is delivered through NVIDIA GPUs. But an increasing share may shift toward custom silicon such as application-specific integrated circuits (ASICs), which are designed to perform specific tasks due to cost, availability, and competitive incentives. Different chip architectures deliver AI workloads at different costs. ASICs trade flexibility for efficiency, often delivering lower costs per unit of useful compute and, in some cases, better power efficiency. Today there is active debate over the percentage of workloads that will shift from GPUs to ASICs. Whether shifts toward these architectures reduce overall AI infrastructure spending depends on one factor: elasticity of demand—i.e., when compute gets cheaper, do buyers spend less—or do they use more of it?

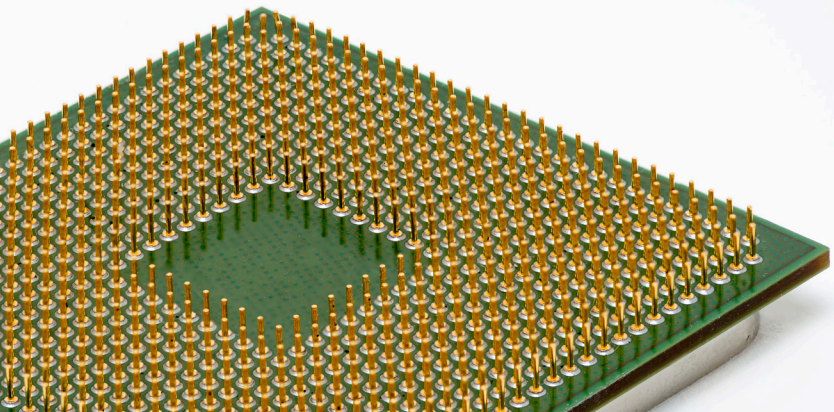
Whether shifts toward these architectures reduce overall AI infrastructure spending depends on one factor: elasticity of demand—i.e., when compute gets cheaper, do buyers spend less—or do they use more of it?

If organizations are building toward a relatively fixed compute requirement—training and serving a defined set of models—then less expensive silicon translates directly into lower capital requirements. In this inelastic-demand scenario, chip choice becomes a meaningful lever on the scale of total investment.

But if demand is instead elastic, lower costs would unlock more usage. Cheaper compute enables larger models, longer training runs, and broader AI deployment, while the overall infrastructure footprint may be broadly similar. In this case, chip mix reshapes *who* earns the margins rather than *how much* is spent, shifting value away from merchant silicon providers and toward hyperscalers, integrators, and end users. To put a sharper point on this: NVIDIA earns gross margins of roughly 75% on its data center GPUs, far above those of alternative silicon providers. At the scale we are discussing, that differential is significant—and it provides a natural economic motivation for buyers to consider alternative architectures.

For illustrative purposes, the baseline used in this analysis aligns more closely with the elastic case. Under that assumption, chip and architecture mix primarily reshape the composition of spend rather than the total amount of infrastructure built.

Over time, the balance could shift. As AI workloads become more inference-heavy and margin-attuned, or as returns to incremental compute diminish, turning to less expensive silicon architectures could begin to constrain total spending rather than expand usage. That regime is plausible—but it does not yet define the current phase of the AI build-out.



Assumption 4
Elongation of the Build-Out Due to Bottlenecks

Elongation refers to the widening gap between capital deployments and new compute capacity coming online. Elongation becomes a greater concern due to power interconnection queues, permitting, shortages of specialized labor, and long lead times for critical equipment—transformers, switchgear, turbines, and cooling systems—all of which can lengthen the time gap between capital commitment and data center operations.

Elongation does not alter the per-unit cost of delivering AI infrastructure. It also does not change the price of silicon, the cost per megawatt of a data center, or the efficiency of a given chip architecture. Rather, its effect on the scale of investment operates through a different channel—stretching timelines, increasing coordination complexity, and, in stress scenarios, eroding the confidence required to sustain capital deployment at current rates.

In the base case, bottlenecks slow deployments without reducing the total amount of infrastructure ultimately built. Projects may slip, phases can extend, and capital sometimes is duplicated through work-arounds, with behind-the-meter generation being the most visible example. While the result is a build-out that is less efficient and more drawn out than road maps imply, it is not materially smaller in aggregate. Under these conditions, elongation is primarily a timing and volatility problem.

The more consequential risk arises if bottlenecks prove severe or persistent enough to shift the narrative around the build-out itself. When enough projects stall simultaneously, attention tends to migrate from supply-side mechanics to demand-side questions—whether end-market revenue will materialize on a timeline that justifies the capital at risk. At that point, elongation begins to function as a feedback loop, one in which supply-side friction introduces demand-side doubt, potentially leading to deferred or downsized investment plans.

The current environment sits closer to the base case than the stress case, though the buffer is not wide. At the scale of capital being committed, even modest delays in execution invite real scrutiny around the demand assumptions used to underwrite these investments. Elongation is a negative factor for nearly every participant in the ecosystem. Credit providers face longer duration risk, offtakers absorb exposure through take-or-pay contracts signed against uncertain timelines, and companies relying on public market fundraising must sustain investor confidence through extended periods of capital deployment without commensurate returns.

Important Dynamics That Do Not Materially Alter the Scale of Investment

The scale of AI infrastructure investment is most determined by assumptions around silicon useful life, data center cost and complexity, and composition and timing of the build-out. Beyond these critical factors, there is enormous movement across the ecosystem—extreme shifts in pricing, supply chain dynamics, and fundamental rewriting of long-held assumptions. But not all of the factors that create headlines move the needle on the overall scale of spend over the medium term. Though they are critical for understanding returns, volatility, and value distribution—especially within individual subsectors—they do not materially alter the overall scale of infrastructure investment.

These are some of the factors:

- Training vs. inference mix
- Per chip memory growth & memory pricing volatility
- Behind-the-meter vs. grid power sourcing

There is enormous movement across the ecosystem—extreme shifts in pricing, supply chain dynamics, and fundamental rewriting of long-held assumptions. But not all of the factors that create headlines move the needle on the overall scale of spend over the medium term.

The balance between training and inference primarily affects the timing of economic realization rather than the scale of infrastructure investment. A faster transition to inference-heavy workloads accelerates revenue generation by converting a fixed capital base into usage, improving utilization and near-term returns. By contrast, a prolonged training-dominant phase extends the ROI timeline as CapEx and R&D continue to be deployed in advance of broad monetization. This distinction has limited inherent impact on the magnitude of AI infrastructure spend. Instead, changes in training vs. inference mix alter how quickly that infrastructure begins to pay for itself.

Memory per accelerator continues to rise, but this trend is largely priced into current infrastructure estimates. Today's accelerator road maps already assume significantly higher memory density per chip, reflecting longer context windows, more stateful inference, and increasingly agentic workloads. When we consider sensitivities around this trend—whether memory per chip evolves broadly in line with today's expectations or ends up 25%+ higher or lower—the impact on our aggregate ~$7.6 trillion infrastructure estimate is modest. This is because rising memory intensity is largely baked into prevailing system designs and pricing, therefore shifting the *composition* of spend within the silicon stack more than it expands the overall envelope of investment. Near-term volatility in memory pricing, even when dramatic, should be understood primarily as a function of supply-demand imbalance at unprecedented and highly lumpy volumes, rather than evidence of a permanent departure from historical cyclicality—informative for margin distribution and vendor positioning but with less impact on the long-run scale of AI infrastructure spend. As with prior chokepoints, we would expect capacity expansion and yield improvements to normalize pricing over time. Memory is unlikely to be the last component to experience this kind of volatility. The "buy out the store before the storm" dynamics that define parts of the AI supply chain suggest that similar episodes of intense, short-term pricing pressure are likely to recur across other critical components such as interconnect, optics, storage, and packaging.

Behind-the-meter power does increase absolute spending on power infrastructure relative to a grid-connected world, as it replaces shared generation and transmission assets with bespoke, project-specific solutions. On a project-by-project basis, this represents a real cost difference: Captive generation typically requires higher upfront capital and results in lower average utilization than grid-scale assets. However, power remains a relatively small share of total AI infrastructure investment when compared with silicon, data-center construction, and supporting systems. As a result, even meaningful shifts in power sourcing strategy—such as widespread adoption of behind-the-meter generation—are unlikely to materially change the aggregate ~$7.6 trillion estimates for ecosystem-wide AI spend. This dynamic has important implications for the power sector itself, where suppliers may enjoy greater pricing power than historical precedent would suggest. The more significant effects are felt in deployment timing, coordination efficiency, and volatility, rather than in the long-run magnitude of capital deployed.

What Actually Moves the Number

Debates about the scale of AI infrastructure investment are often framed as a referendum on demand: whether AI adoption, monetization, or productivity gains will ultimately justify trillions of dollars of capital investments. But the amount of capital required to support today's AI ambitions is not a single, fixed number—it is highly sensitive to a small set of structural assumptions about how the infrastructure itself is built and renewed.

The implication is not that current estimates are obviously too high or too low, but that they are far more conditional than they appear and so may shift over time. As assumptions around technology progress and system design and market demand behavior shift, estimates of required capital will move with them. For investors and operators, critical questions remain: What fundamental assumptions do we have about the future, and how resilient to changes in those assumptions are our plans?

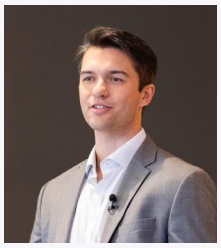
Innovation may be the most important wild card. Current assumptions are largely based on current technologies. But a truly discontinuous innovation—for example, one that materially reduces the compute complexity of both training and inference—could further shift the investment landscape. The DeepSeek moment in January 2025 offered a glimpse of how markets might react to such a development. While events that followed suggest it was not, in the end, the kind of paradigm shift that would fundamentally alter the trajectory of infrastructure investment, the episode was a reminder that this risk exists and that investment frameworks for the AI build-out need to be attuned to such factors.

There is a certain circularity hinted at in this analysis. Much of it has focused on how difficult it will be to deploy trillions of dollars of capital against the physical, institutional, and economic constraints we have described. But if the ecosystem does manage to conquer those constraints—if the infrastructure is built, the bottlenecks are cleared, and the cost of compute continues to fall—then the history of technology suggests that the result may not be surplus capacity but rather a new wave of demand and use cases that could not have existed at higher price points. The success of the build-out for today's AI ambitions may be what ensures that it is not enough for tomorrow's technological opportunities.

Authors

GEORGE LEE

Co-Head
Goldman Sachs Global Institute



LUCAS GREENBAUM

Vice President
Goldman Sachs Global Institute

Disclaimer