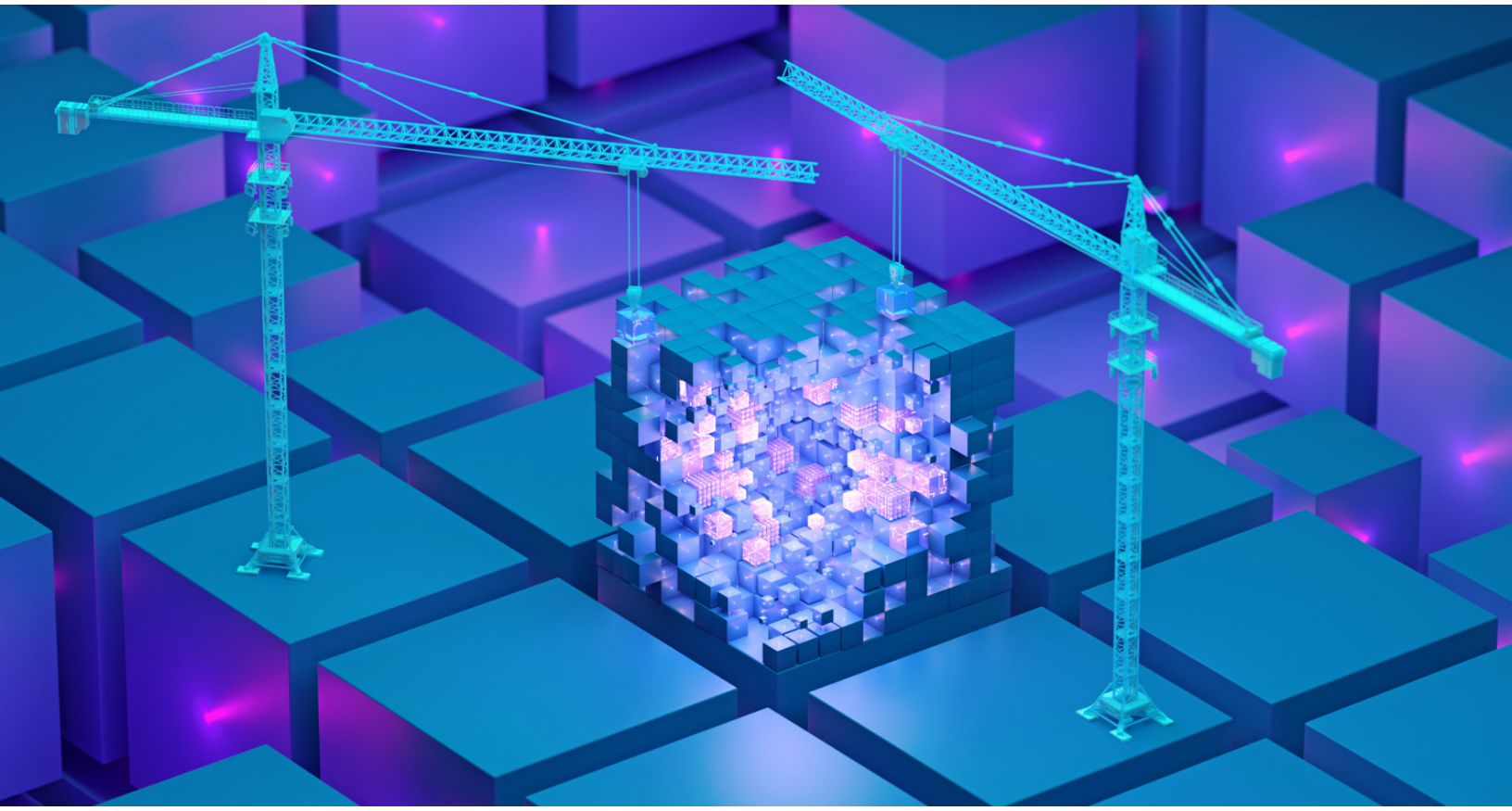
McKinsey Quarterly

Technology, Media & Telecommunications Practice

The cost of compute: A $7 trillion race to scale data centers

AI is fueling high demand for compute power, spurring companies to invest billions of dollars in infrastructure. But with future demand uncertain, investors will need to make calculated decisions.

This article is a collaborative effort by Jesse Noffsinger, Mark Patel, and Pankaj Sachdeva, with Arjita Bhan, Haley Chang, and Maria Goodpaster, representing views from McKinsey's Technology, Media & Telecommunications Practice.



April 2025

Amid the AI boom, compute power is emerging as one of this decade's most critical resources. In data centers across the globe, millions of servers run 24/7 to process the foundation models and machine learning applications that underpin AI. The hardware, processors, memory, storage, and energy needed to operate these data centers are collectively known as compute power—and there is an unquenchable need for more.

Our research shows that by 2030, data centers are projected to require $6.7 trillion worldwide to keep pace with the demand for compute power. Data centers equipped to handle AI processing loads are projected to require $5.2 trillion in capital expenditures, while those powering traditional IT applications are projected to require $1.5 trillion in capital expenditures (see sidebar "What about non-AI workloads?"). Overall, that's nearly *$7 trillion* in capital outlays needed by 2030—a staggering number by any measure.[1]

What about non-AI workloads?

While AI workloads dominate the conversation, non-AI processing loads remain a significant portion of data center activity. These include traditional enterprise IT tasks such as web hosting, enterprise resource planning systems, email, and file storage. Non-AI loads are less compute-intensive and can operate efficiently on central processing units rather than the specialized graphics processing units or AI accelerators that AI workloads require. They also tend to have more predictable usage patterns and lower power densities, which allow for less demanding cooling and energy requirements. As a result, data centers focused on non-AI processing typically have different infrastructure needs, capital intensity, and operational considerations compared with those optimized for AI.

To meet this demand, companies across the compute power value chain must strike a balance between deploying capital quickly and doing so prudently. To improve the odds that their data center investments will provide strong returns, companies can tackle projects in stages, assessing ROI at each step. Still, a lack of clarity about future demand makes precise investment calculations difficult.

The compute power value chain is complex—from the real estate developers that build data centers to the utilities that power them, to the semiconductor firms that produce chips to the cloud service hyperscalers that host trillions of terabytes of data. Leaders across this value chain know that they must invest in compute power to accelerate AI growth. But their challenge

[1] McKinsey Data Center Demand Model, McKinsey Data Center Capex TAM Model, and expert interviews.

is formidable: deciding how much capital to allocate to which projects, all while remaining uncertain of how AI's future growth and development will impact compute power demand. Will hyperscalers continue shouldering the cost burden, or will enterprises, governments, and financial institutions step in with new financing models? Will demand for data centers rise amid a continued surge in AI usage, or will it fall as technological advances make AI less compute-heavy?

One thing is certain: The stakes are high. Overinvesting in data center infrastructure risks stranding assets, while underinvesting means falling behind. This article, based on McKinsey research and analysis, provides companies across the compute power value chain with an overview of the investment landscape for the next five years. Despite the rigor behind these forecasts, we acknowledge that AI is a radically evolving space. Our analysis is built on thoroughly researched hypotheses, but there are critical uncertainties that cannot yet be quantified.

Predicting the compute power demand curve

To decide how much to invest in compute power, companies should first accurately forecast future demand—a challenging task given that the AI sector is shifting so rapidly. Our research shows that global demand for data center capacity could almost triple by 2030, with about 70 percent of that demand coming from AI workloads (Exhibit 1). However, this projection hinges on two key uncertainties:

— *AI use cases*. The value in AI lies at the application layer—how enterprises turn AI into real business impact. If companies fail to create meaningful value from AI, demand for compute power could fall short of expectations. Conversely, transformative AI applications could fuel even greater demand than current projections suggest.

— *Rapid innovation cycles and disruptions*. Continuous advancements in AI technologies, such as processors, large language model (LLM) architectures, and power consumption, could significantly enhance efficiency. For instance, in February 2025, Chinese LLM player DeepSeek reported that its V3 model achieved substantial improvements in training and reasoning efficiency, notably reducing training costs by approximately 18 times and inferencing costs by about 36 times, compared with GPT-4o.[2] However, preliminary analysis suggests that these types of efficiency gains will likely be offset by increased experimentation and training across the broader AI market. As a result, efficiency gains may not substantially impact overall compute power demand over the long term.[3]

AI demand alone will require $5.2 trillion in investment

We calculate that companies across the compute power value chain will need to invest *$5.2 trillion into data centers by 2030* to meet worldwide demand for AI alone. We based this figure on extensive analysis and key assumptions, including a projected 156 gigawatts (GW) of AI-

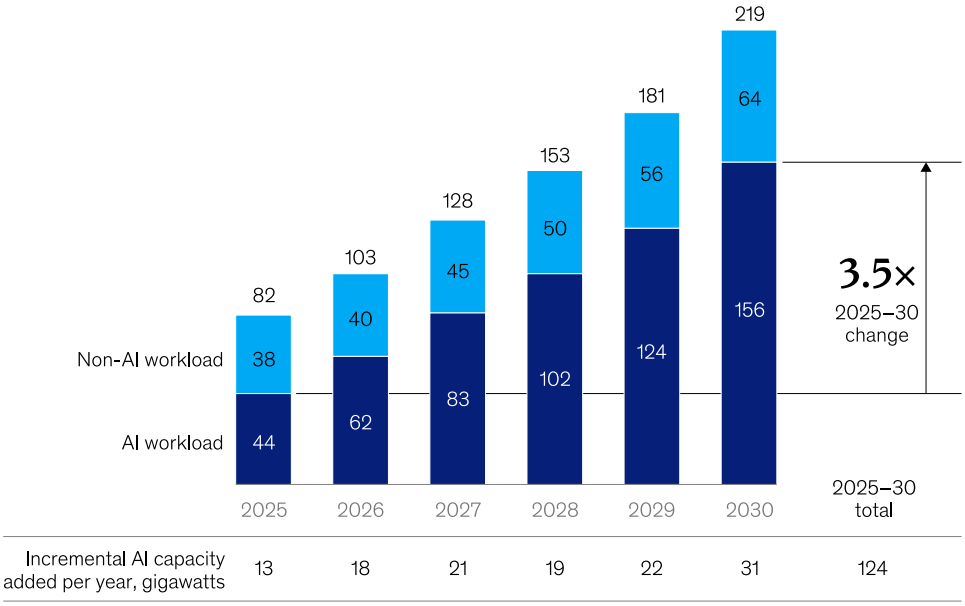
[2] Manish Singh, "DeepSeek 'punctures' AI leaders' spending plans, and what analysts are saying," TechCrunch, January 27, 2025; Wayne Williams, "OpenAI spent $80M to $100M training GPT-4; Chinese firm claims it trained its rival AI model for $3 million using just 2,000 GPUs," TechRadar, December 2, 2024; "Independent analysis of AI models and API providers," Artificial Analysis, January 27, 2025.

[3] This aligns with the concept of Jevons Paradox, which posits that improvements in efficiency can lead to increased overall demand. In the context of AI, more efficient and accessible computing resources could spur greater adoption and utilization, potentially offsetting the anticipated reductions in compute demand.

Exhibit 1

Both AI and non-AI workloads will be key drivers of global data center capacity demand growth through 2030.

Estimated global data center capacity demand, 'continued momentum' scenario, gigawatts



Note: Figures may not sum to totals, because of rounding.
Source: McKinsey Data Center Demand Model; Gartner reports; IDC reports; Nvidia capital markets reports

McKinsey & Company

related data center capacity demand by 2030, with 125 incremental GW added between 2025 and 2030. This $5.2 trillion figure reflects the sheer scale of investment required to meet the growing demand for AI compute power—a significant capital commitment that underscores the magnitude of the challenge ahead (see sidebar "The scale of investment").

Amid the uncertainty about future needs for compute power, we created three investment scenarios ranging from constrained to accelerated demand (Exhibit 2). In the first of our three scenarios, growth accelerates significantly and 205 incremental GW of AI-related data center capacity is added between 2025 and 2030. This would require an estimated $7.9 trillion in capital expenditures. The second scenario is the one we use in this article: Demand grows, but not as much as in the first scenario, and the expected capital expenditure is $5.2 trillion. In our third scenario, in which demand is more constrained, with 78 incremental GW added in the next five years, the total capital expenditure is $3.7 trillion (see sidebar "Methodology ").

The scale of investment

To put the trillion-dollar size of investment needed by 2030 into perspective, consider these unrelated statistics that illustrate the sheer scale of capital needed:

— *Labor*. $500 billion in labor costs is roughly equivalent to 12 billion labor hours (six million people working full time for an entire year).[1]

— *Fiber*. $150 billion worth of fiber is equivalent to installing three million miles of fiber-optic cables—enough to circle the Earth 120 times.[2]

— *Power generation*. $300 billion worth of power generation is equivalent to adding 150 to 200 gigawatts of gas, which would be enough to power 150 million homes for a year—more than the total number of households in the United States.[3]

[1] Estimated conservatively, using a high-end hourly wage of $40 (for roles ranging from construction workers to data center technicians), assuming a standard 40-hour workweek and 52 weeks per year. "Occupational employment and wage statistics survey by occupation–May 2024," US Bureau of Labor Statistics news release, April 2, 2025.
[2] Estimated based on an average of $50,000 as the per-mile aerial installation cost of fiber-optic cables. Jonathan Kim, "Fiber optic network construction: Process and build costs," Dgtl Infra, January 15, 2024.
[3] Estimated assuming full utilization and average US household electricity consumption of 10,500 kilowatt-hours per year. Based on data from the US Energy Information Administration.

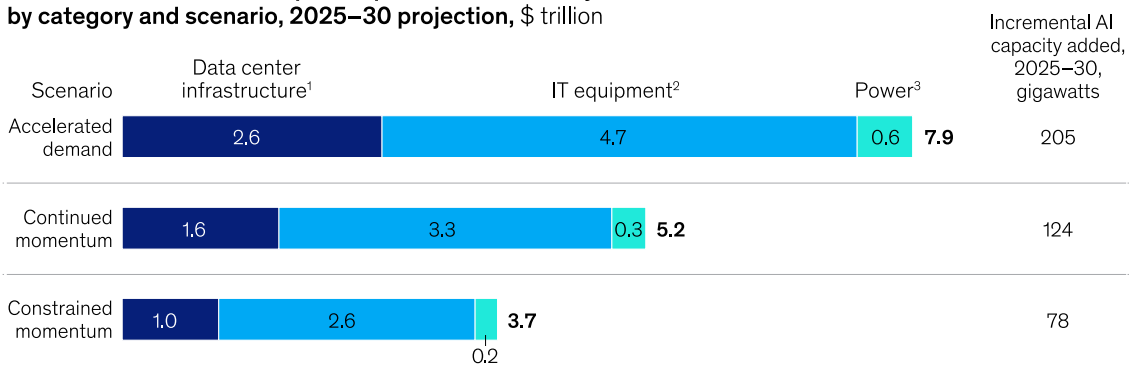
Methodology

Capital expenditure estimates in this article are derived from McKinsey's proprietary data center demand model, which projects data center capacity under multiple scenarios shaped by factors such as semiconductor supply constraints, enterprise AI adoption, efficiency improvements, and regulatory challenges. Investment requirements were calculated by translating demand projections for gigawatt capacity into capital expenditures across major cost categories, including power (for example, generation, transmission), data center infrastructure (for example, electrical, mechanical, site, shell), and IT equipment (for example, AI accelerators, networking, storage).

Exhibit 2

Capital investments to support AI-related data center capacity demand could range from about $3 trillion to $8 trillion by 2030.

Global data center total capital expenditures driven by AI, by category and scenario, 2025–30 projection, $ trillion



Note: Figures may not sum to totals, because of rounding.
[1]Excludes IT services and software (eg, operating system, data center infrastructure management), since they require relatively low capex compared with other components.
[2]Includes server, storage, and network infrastructure. IT capex also accounts for replacing AI accelerators every 4 years.
[3]Assumes $2.2 billion–$3.2 billion/gigawatt (including power generation and transmission cost) to account for a range of power generation scenarios (eg, fully powered by gas, a combination of gas power and storage, and solar) and regional cost differences. Distribution cost is neglected, as most AI centers are expected to be >50 megawatt scale and connected to a transmission grid.
Source: McKinsey Data Center Capex TAM Model; McKinsey Data Center Demand Model

McKinsey & Company

In any scenario, these are staggering investment numbers. They are fueled by several factors:

— *Mass adoption of gen AI*. The foundation models that underpin gen AI require significant compute power resources to train and operate. Both training and inference workloads are contributing to infrastructure growth, with inference expected to become the dominant workload by 2030.

— *Enterprise integration*. Deploying AI-powered applications across industries—from automotive to financial services—demands massive cloud computing power. As use cases grow, AI applications will grow more sophisticated, integrating specialized foundation models tailored to specific domains.

— *Competitive infrastructure race*. Hyperscalers and enterprises are racing to build proprietary AI capacity to gain competitive advantage, which is fueling the construction of more and more data centers. These “builders” (as further described below) hope to gain competitive advantage by achieving scale, optimizing across data center tech stacks, and ultimately driving down the cost of compute.

— *Geopolitical priorities*. Governments are investing heavily in AI infrastructure to enhance security, economic leadership, and technological independence.

Where is the investment going?

To qualify our $5.2 trillion investment forecast for AI infrastructure, it's important to note that our analysis likely undercounts the total capital investment needed, as our estimate quantifies capital investment for only three out of five compute power investor archetypes—builders, energizers, and technology developers and designers—that directly finance the infrastructure and foundational technologies necessary for AI growth (see sidebar "Five types of data center investors"). Approximately 15 percent ($0.8 trillion) of investment will flow to builders for land, materials, and site development. Another 25 percent ($1.3 trillion) will be allocated to energizers for power generation and transmission, cooling, and electrical equipment. The largest share of investment, 60 percent ($3.1 trillion), will go to technology developers and designers, which produce chips and computing hardware for data centers. The other two investor archetypes, operators, such as hyperscalers and colocation providers, and AI architects, which build AI models and applications, also invest in compute power, particularly in areas such as AI-driven automation and data center software. But quantifying their compute power investment is challenging because it overlaps with their broader R&D spending.

Despite these projected capital requirements, our research shows that current investment levels lag demand. In dozens of client interviews, we found that CEOs are hesitant to invest in compute power capacity at maximum levels because they have limited visibility into future demand. Uncertainty about whether AI adoption will continue its rapid ascent and the fact that infrastructure projects have long lead times make it difficult for companies to make informed investment decisions. Many companies are unsure whether large capital expenditures on AI infrastructure today will produce measurable ROI in the future. So how can business leaders move forward confidently with their investments? As a first step, they can determine where their organizations fall within the compute power ecosystem.

Five archetypes of AI infrastructure investors

Who are the investors behind the multitrillion-dollar race to fund AI compute power? We have identified five key investor archetypes, each navigating distinct challenges and opportunities, and detailed how much they could spend in the next five years.

1. Builders

— *Who they are*: real estate developers, design firms, and construction companies expanding data center capacity

Five types of data center investors

As AI drives a surge in compute power demand, five types[1] of organizations are leading the massive capital investments required to scale data centers:

— *Builders*: real estate developers, design firms, and construction companies that expand and upgrade data centers, such as Turner Construction and AECOM

— *Energizers*: companies that supply the electricity and cooling systems essential for data center operations, including utilities like Duke Energy and Entergy and infrastructure and equipment providers like Schneider Electric and Vertiv

— *Technology developers and designers*: semiconductor companies that develop the chips powering AI workloads, such as NVIDIA and Intel, and computing hardware suppliers such as Foxconn and Flex

— *Operators*: cloud providers and co-location firms that own and run large-scale data centers, such as Amazon Web Services, Google Cloud, and Equinix

— *AI architects*: companies developing AI models and infrastructure, including OpenAI and Anthropic

[1]The companies listed as examples in each archetype category may also span adjacent categories. For instance, hyperscalers in the operators category (Amazon Web Services, Google Cloud) are developing specialized computing hardware and investing in both their own and third-party LLM products—activities that align with the roles of technology developers and designers and AI architects, respectively.

— *AI workload capital expenditure*: $800 billion

— *Non-AI workload capital expenditure*: $100 billion

— *Key investments*: land and material acquisition, skilled labor, site development

Opportunities. Builders that optimize site selection can secure prime locations, reduce construction timelines, and integrate operational feedback early, ensuring faster deployment and higher data center efficiency.

Challenges. Labor shortages could impact technician and construction worker availability, while location constraints could limit site selection options. Meanwhile, increased rack power density could create space and cooling challenges.

Solutions. Forward-thinking builders can find solutions to core challenges, adding certainty to their investment decisions. For example, some are solving the labor shortage issue by adopting modular designs that streamline the construction process, such as off-site construction of large components that can be assembled on-site.

2. Energizers

— *Who they are*: utilities, energy providers, cooling/electrical equipment manufacturers, and telecom operators building the power and connectivity infrastructure for AI data centers[4]

— *AI workload capital expenditure*: $1.3 trillion

— *Non-AI workload capital expenditure*: $200 billion

— *Key investments*: power generation (plants, transmission lines), cooling solutions (air cooling, direct-to-chip liquid cooling, immersion cooling), electrical infrastructure (transformers, generators), network connectivity (fiber, cable)

Opportunities. Energizers that scale power infrastructure and innovate in sustainable energy solutions will be best positioned to benefit from hyperscalers' growing energy demands.

Challenges. Powering data centers could stall due to existing grid weaknesses and solving heat management challenges from rising processor densities remains an obstacle. Energizers also face clean-energy transition requirements and lengthy grid connection approval processes.

Solutions. With over $1 trillion in investment at stake, energizers are finding ways to deliver reliable power while driving ROI. They are making substantial investments in emerging power-generation technologies—including nuclear, geothermal, carbon capture and storage, and long-duration energy storage. They are also doubling down on efforts to bring as much capacity online as quickly as possible across both renewable sources and traditional energy infrastructure, such as gas and fossil fuels. What is changing now is the sheer scale of that demand, which brings a new urgency to build power capacity at unprecedented speed. As demand—especially for clean energy—surges, power generation is expected to grow rapidly, with renewables projected to account for approximately 45 to 50 percent of the energy mix by 2030, up from about a third today.[5]

3. Technology developers and designers

— *Who they are*: semiconductor firms and IT suppliers producing chips and computing hardware for data centers

— *AI workload capital expenditure*: $3.1 trillion

— *Non-AI workload capital expenditure*: $1.1 trillion

[4]For more on how utility and energy providers are investing in AI infrastructure, see "How data centers and the energy sector can sate AI's hunger for power," McKinsey, September 17, 2024; for more on how telecom operators are investing in AI infrastructure, see "AI infrastructure: A new growth avenue for telco operators," McKinsey, February 28, 2025.
[5]"Global Energy Perspective," McKinsey, September 17, 2024.

— *Key investments*: GPUs, CPUs, memory, servers, and rack hardware

Opportunities. Technology developers and designers that invest in scalable, future-ready technologies supported by clear demand visibility could gain a competitive edge in AI computing.

Challenges. A small number of semiconductor firms control the market supply, stifling competition. Capacity building remains insufficient to meet current demand, while at the same time, shifts in AI model training methods and workloads make it difficult to predict future demand for specific chips.

Solutions. Technology developers and designers have the most to gain in the compute power race because they are the ones providing the processors and hardware that do the actual computing. Demand for their products is currently high, but their investment needs are also the greatest—more than $3 trillion over the next five years. A small number of semiconductor firms have a disproportionate influence on industry supply, making them potential chokepoints in compute power growth. Technology developers and designers can mitigate this risk by expanding fabrication capacity and diversifying supply chains to prevent bottlenecks.

4. Operators

— *Who they are*: hyperscalers, colocation providers, GPU-as-a-service platforms, and enterprises optimizing their computing resources by improving server utilization and efficiency

— *AI workload capital expenditure*: not included in this analysis

— *Non-AI workload capital expenditure*: not included in this analysis

— *Key investments*: data center software, AI-driven automation, custom silicon

Opportunities. Operators that scale efficiently while balancing ROI, performance, and energy use can drive long-term industry leadership.

Challenges. Immature AI-hosted applications can obscure long-term ROI calculations. Inefficiencies in data center operations are driving up costs, but uncertainty in AI demand continues to disrupt long-term infrastructure planning and procurement decisions.

Solutions. While data centers today operate at high-efficiency levels, the rapid pace of AI innovation will require operators to optimize both energy consumption and workload management. Some operators are improving energy efficiency in their data centers by investing in more effective cooling solutions and increasing rack stackability to reduce space requirements without sacrificing processing power, for example. Others are investing in AI model development itself to create architectures that need less compute power to be trained and operated.

5. AI architects

— *Who they are*: AI model developers, foundation model providers, and enterprises building proprietary AI capabilities

— *AI workload capital expenditure*: not included in this analysis

— *Non-AI workload capital expenditure*: not included in this analysis

— *Key investments*: model training and inference infrastructure, algorithm research

Opportunities. AI architects that develop architectures that balance performance with lower compute requirements will lead the next wave of AI adoption. Enterprises investing in proprietary AI capabilities can gain competitiveness by developing specialized models tailored to their needs.

Challenges. AI governance issues, including bias, security, and regulation, add complexity and can slow development. Meanwhile, inference poses a major unpredictable cost component, and enterprises are facing difficulties demonstrating clear ROI from AI investments.

Solutions. The escalating computational demands of large-scale AI models are driving up the costs to train them, particularly regarding inference, or the process where trained AI models apply their learned knowledge to new, unseen data to make predictions or decisions. Models with advanced reasoning capabilities, such as OpenAI's o1, require significantly higher inference costs. For example, it costs six times more for inference on OpenAI's o1 compared with the company's nonreasoning GPT-4o. To bring down inference costs, leading AI companies are optimizing their model architectures by using techniques like sparse activations and distillation. These solutions reduce the computational power needed when an AI model generates a response, making operations more efficient.

Critical considerations for AI infrastructure growth

As companies plan their AI infrastructure investments, they will have to navigate a wide range of potential outcomes. In a constrained-demand scenario, AI-related data center capacity could require $3.7 trillion in capital expenditures—limited by supply chain constraints, technological disruptions, and geopolitical uncertainty. These barriers are mitigated, however, in an accelerated-demand scenario, leading to investments as high as $7.9 trillion. Staying on top of the evolving landscape is critical to making informed, strategic investment decisions. Some of the uncertainties investors must consider include:

— *Technological disruptions*. Breakthroughs in model architectures, including efficiency gains in compute utilization, could reduce expected hardware and energy demand.

— *Supply chain constraints*. Labor shortages, supply chain bottlenecks, and regulatory hurdles could delay grid connections, chip availability, and data center expansion—slowing overall AI adoption and innovation. To address supply chain bottlenecks for critical chips, semiconductor companies are investing significant capital to construct new fabrication facilities, but this construction could stall due to regulatory constraints and long lead times from upstream equipment suppliers.

— *Geopolitical tensions*. Fluctuating tariffs and technology export controls could introduce uncertainty in compute power demand, potentially impacting infrastructure investments and AI growth.

The race for competitive advantage

The winners of the AI-driven computing era will be the companies that anticipate compute power demand and invest accordingly. Companies across the compute power value chain that proactively secure critical resources—land, materials, energy capacity, and computing power—could gain a significant competitive edge. To invest with confidence, they can take a three-pronged approach.

First, investors will need to understand demand projections amid uncertainty. Companies should assess AI computing needs early, anticipate potential shifts in demand, and design scalable investment strategies that can adapt as AI models and use cases evolve. Second, investors should find ways to innovate on compute efficiency. To do so, they can prioritize investments in cost- and energy-efficient computing technologies, optimizing performance while managing power consumption and infrastructure costs. Third, they can build supply-side resilience to sustain AI infrastructure growth without overextending capital. This will require investors to secure critical inputs such as energy and chips, optimize site selection, and build flexibility into their supply chains.

Striking the right balance between growth and capital efficiency will be critical. Investing strategically is not just a race to scale data infrastructure—it's a race to shape the future of AI itself.

We are celebrating the 60th birthday of the McKinsey Quarterly *with a yearlong campaign featuring four issues on major themes related to the future of business and society, as well as related interactives, collections from the magazine's archives, and more. This article will appear in the fourth themed issue, which will launch in July. Sign up for the* McKinsey Quarterly *alert list to be notified as soon as other new* Quarterly *articles are published.*

Jesse Noffsinger is a partner in McKinsey's Seattle office, where **Maria Goodpaster** is an associate partner; **Mark Patel** is a senior partner in the Bay Area office, where **Haley Chang** is a consultant; **Pankaj Sachdeva** is a senior partner in the Philadelphia office; and Arjita Bhan is a knowledge expert in the Boston office.

The authors wish to thank Andrea Boza Zanatta, Jason Amri, Rishi Gupta, Senem Bilir, and Shraddha Kumar for their contributions to this article.

This article was edited by Kristi Essick, an executive editor in the Bay Area office.