Financing the AI Buildout

Stijn Van Nieuwerburgh

Draft: March 20, 2026

I. Introduction

Artificial intelligence is often described as an intangible revolution. Public debate focuses on models, algorithms, productivity gains, and the automation of cognitive tasks. Yet the near-term economic footprint of AI is strikingly physical. Training and deploying modern AI systems requires a rapidly expanding capital stock of data centers, power infrastructure, cooling systems, network equipment, and specialized chips. The current phase of the AI boom is therefore not only a software story. It is also an infrastructure buildout.

This physical dimension matters because AI workloads are changing the economics of digital infrastructure. Compared with earlier generations of cloud computing, frontier AI systems require far greater power densities, more advanced cooling technologies, and tighter integration between hardware and facilities. As a result, many existing data centers are ill-suited for state-of-the-art workloads. AI is not simply increasing demand for data centers in the aggregate; it is changing what kinds of facilities are valuable and making access to large amounts of reliable power a central constraint.

A first claim of this paper is that AI is driving a wave of physical capital formation that is unusual in both scale and composition. The relevant investment spans not only buildings but also substations, transmission upgrades, cooling systems, and large volumes of specialized IT equipment. In aggregate, the resulting buildout resembles earlier infrastructure booms—such as railroads, electrification, and telecommunications—more than a conventional increase in technology-sector capital expenditures.

A second claim is that the ownership and financing of this buildout are changing in important ways. Although hyperscalers remain central actors, a substantial share of the required capital is increasingly supplied by external investors, including real estate developers, infrastructure funds, private credit providers, and structured finance vehicles. In many cases, the user of compute is not the owner of the underlying assets, and the economic risk is distributed across a layered set of claims held by different investors.

A third claim is that this financial architecture reshapes how risk is allocated and perceived. Off-balance-sheet structures, long-term leases, project finance, and asset-backed lending allow firms to scale infrastructure while preserving balance sheet flexibility and high equity valuations. At the same time, these arrangements

can increase asset-level leverage, introduce contractual complexity, and make the ultimate distribution of risk less transparent. Historical experience with large infrastructure booms as well as with prior episodes of complex structured finance suggests that such dynamics can become financially consequential when expectations about demand or financing conditions change.

This paper connects to several strands of the existing literature. First, it relates to work on general-purpose technologies and large-scale infrastructure investment, which emphasizes the role of complementary physical capital in enabling productivity gains (Jovanovic and Rousseau, 2005; Fernald, 1999). Second, it contributes to the literature on corporate finance and securitization, which studies how financial structures can expand investment capacity while reshaping incentives and the allocation of risk (DeMarzo, 2005; Coval, Jurek and Stafford, 2009). Third, it connects to recent work on technological innovation and wealth inequality, which highlights how financial claims on new technologies can both redistribute and concentrate economic gains (Kogan, Papanikolaou and Stoffman, 2020). Finally, it relates to a growing literature on intangible capital (Haskel and Westlake, 2017; Eberly, 2019; Eisfeldt, Schubert and Zhang, 2023) and the data economy (Farboodi and Veldkamp, 2026; Chung and Veldkamp, 2024; Jones and Tonetti, 2020), which emphasizes the role of data as a key input in modern production and the interaction between intangible assets and the tangible capital required to process and scale them. The AI infrastructure buildout provides a novel setting in which these forces interact at an unprecedented scale.

The paper proceeds as follows. Section II describes the scale, physical constraints, and economic characteristics of AI infrastructure. Section III analyzes how this infrastructure is owned and financed, emphasizing the role of external capital and the structure of the capital stack. Section IV examines how these financing arrangements redistribute risk, using recent transactions as illustrative examples. Section V discusses the main channels through which financial and technological risks may materialize, and relates them to historical infrastructure cycles. Section VI concludes with directions for future research in this interesting yet young topic.

II. AI as a Physical Capital Boom

The rapid progress in artificial intelligence is often described as a software revolution. In economic terms, however, the current phase of AI development is better understood as a large-scale investment cycle in physical capital. Training and deploying modern AI systems requires vast quantities of specialized hardware, electricity, and purpose-built facilities. As a result, the expansion of AI is tightly linked to the construction of a new layer of industrial infrastructure.

At the center of this buildout are data centers designed for high-performance computing. While data centers have long been part of the digital economy, the requirements of AI workloads differ sharply from those of earlier generations of cloud computing. Traditional facilities were optimized for storage, web hosting,

and enterprise applications, with relatively modest power densities. By contrast, modern AI clusters concentrate large numbers of graphics processing units (GPUs) and high-speed interconnects in a single location, placing far greater demands on power delivery, cooling, and physical space.

The intensity of these requirements reflects the rapid evolution of chip design. In earlier generations of data centers, a rack of CPU-based servers might consume on the order of 5 to 10 kilowatts (kW) and could be cooled with conventional airflow systems. By contrast, a modern AI rack populated with high-end accelerators can draw 80 to 120 kW or more. For example, a single rack of 72 next-generation GPUs—such as NVIDIA's Blackwell-class systems—can exceed 100 kW once associated networking and power overhead are included. At these densities, traditional air cooling becomes insufficient. Facilities increasingly rely on direct-to-chip liquid cooling or closed-loop refrigerant systems to remove heat efficiently and maintain performance.

These rack-level differences scale up dramatically at the campus level. A 200 megawatt (MW) data center can be thought of as the equivalent of roughly two thousand such high-density AI racks operating simultaneously, drawing as much electricity as 150,000 to 200,000 U.S. homes. Unlike residential demand, however, this load is highly concentrated and nearly continuous, with utilization rates close to full capacity around the clock. Delivering this power reliably requires dedicated substations, high-voltage transmission connections, and substantial investment in grid interconnection, making electricity and cooling infrastructure central constraints on the expansion of AI compute.

These changes are reflected in the scale and cost of new facilities. A contemporary hyperscale campus designed for AI training may require on the order of 200 megawatts (MW) of capacity. At current construction costs of roughly $11 million per MW, the data center facility itself represents an investment of about $2.2 billion. Supporting this load often requires substantial incremental spending on power infrastructure–such as substations, transmission upgrades, and grid interconnection,–adding on the order of $0.4 billion.

The largest component of the investment, however, lies inside the building. Outfitting a 200 MW campus with AI hardware can require approximately $5.6 billion in IT equipment, assuming that roughly 70 percent of capacity is dedicated to GPU-based training workloads at a cost of about $40 million per MW. Within this category, the majority of spending reflects GPU servers (including accelerator-attached memory), with additional expenditures on networking fabric, storage systems, control nodes, and rack-level integration.

Taken together, these figures imply a total capital cost of roughly $8.2 billion for a single 200 MW AI training campus, of which about one-third is associated with real estate and power infrastructure and two-thirds with compute hardware. This composition is a departure from earlier generations of data centers, where the physical facility accounted for a larger share of total cost. In the AI era, the economic center of gravity shifts toward the compute layer, even as the feasibility of that compute remains tightly constrained by physical infrastructure. At the same

time, the new scale of these data center assets have drawn increasing attention from institutional investors, as data centers emerge as a new asset class in real estate and infrastructure portfolios (Green Street, 2026).

Aggregating across projects, the scale of the investment cycle becomes substantial. Recent work has begun to construct more comprehensive measures of data center investment using project-level data; Brandsaas et al. (2025), for example, use detailed information on planned and ongoing developments to nowcast aggregate investment in real time. Based on announced and planned developments, U.S. data center capacity could increase by as much as 200 gigawatts (GW) over the 2026–2032 period, compared with a current installed base of about 50 GW and an additional 36 GW under construction.[1]

Using the earlier benchmark of approximately $8.2 billion for a 200 MW campus, a 200 GW buildout implies a total investment on the order of $8.2 trillion in data centers, power infrastructure, and IT equipment. These estimates are broadly consistent with industry projections that place the global scale of AI-related data center investment in the trillions of dollars (McKinsey & Company, 2025; Morgan Stanley Research, 2025; Moody's Ratings, 2026*b*). Spread over eight years, and assuming nominal GDP growth of 5 percent, this $8.2 trillion corresponds to annual investment of roughly 2.8 percent of GDP. Already today, the scale of the AI buildout is clearly visible in aggregate activity. In the fourth quarter of 2025, investment associated with AI infrastructure accounted for essentially all of the observed growth in U.S. GDP, highlighting the extent to which the current expansion is being driven by a single, capital-intensive sector.

Historical comparisons help place this magnitude in context. Railroad investment in the United States between 1865 and 1890 averaged approximately 2.4 percent of GDP, electrification between 1905 and 1925 about 1.1 percent, the construction of the interstate highway system between 1956 and 1973 about 1.6 percent, and the telecom and fiber expansion between 1996 and 2003 about 0.8 percent. The projected AI buildout between 2025 and 2032, at roughly 2.8 percent of GDP, would exceed these historical episodes.

The AI infrastructure boom shares key features with these earlier episodes: a general-purpose technology induces complementary investment in a large-scale physical network, whose returns depend on uncertain future demand and whose financing requirements are large.

The physical nature of this investment also reshapes the geography of AI. A useful distinction is between data centers used for model training and those used for inference. Training large models involves running massive, batch-style computations that are relatively insensitive to latency and can therefore be located far from end users. By contrast, inference—serving model outputs in real time—often requires low latency and proximity to population centers. Examples include agentic AI systems responding to user queries, streaming video and recommendation engines,

[1]Estimates from the National Laboratory of the Rockies, which aggregates commercial data sources (such as Baxtel) with infrastructure datasets on transmission and fiber networks to map current, under-construction, and planned data center capacity in the United States.

or even routine interactions such as checking a weather app. These applications benefit from fast response times and are typically deployed closer to users within established network hubs.

This distinction helps explain why the primary constraint for large AI training facilities is increasingly access to reliable, scalable, and inexpensive power. Securing grid capacity, transmission access, and permitting approvals can take several years, making electricity infrastructure, rather than land or connectivity, the binding constraint in many locations. As a result, new development is spreading toward regions that offer a favorable combination of energy availability, regulatory approval, and expansion potential, even as established hubs continue to play an important role for latency-sensitive inference workloads. Gargano and Giacoletti (2025) document that state and local subsidies, often structured as sales tax exemptions on data center equipment, are an additional, meaningful determinant of data center location choice.

In summary, the combination of scale, capital intensity, and physical constraints distinguishes the current phase of AI from earlier digital expansions. The deployment of compute is no longer limited primarily by software or demand, but by the available supply of capital, power, and specialized infrastructure such as data centers and GPUs. As a result, the key economic questions concern how this infrastructure is financed, who ultimately owns it, and where the risks associated with the AI buildout reside.

III. Ownership and Financing of AI Infrastructure

The scale of the AI infrastructure buildout raises a natural question: who is financing and owning the underlying assets? While hyperscalers and large technology firms are the primary drivers of demand, they are not, in general, the sole providers of capital. The magnitude of required investment, on the order of trillions of dollars, exceeds what even the largest firms can comfortably fund on balance sheet without affecting leverage, valuation, or strategic flexibility.

Historically, major cloud providers tended to own and operate much of their data center infrastructure directly. This model aligned control over physical assets with control over compute workloads and allowed firms to internalize both the costs and benefits of expansion. In the current cycle, however, the scale and speed of investment have led to a more mixed model, in which hyperscalers increasingly combine owned capacity with leased facilities, joint ventures, and partnerships with specialized third-party developers. Figure 1 illustrates the fast growth of capital expenditures on the five main hyperscalers' balance sheets.

A central feature enabling this shift from owning to renting is the credit quality of the primary tenants. Hyperscalers such as Amazon, Microsoft, Meta, Google, and Oracle carry investment-grade credit ratings, which are the gold standard in real estate and infrastructure leasing. Long-term contracts with such tenants provide predictable cash flows and make projects attractive to lenders and institutional

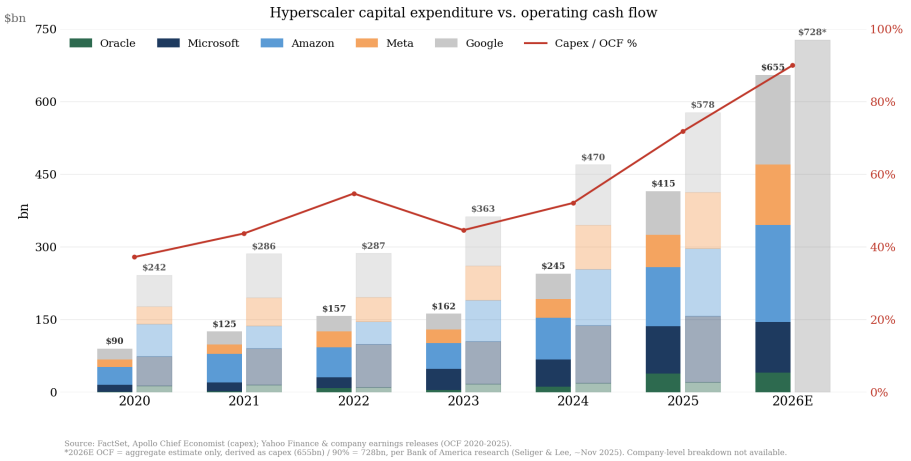

Figure 1. Hyperscaler Capital Expenditures, Operating Cash Flow, and Capex Intensity

Notes: The figure plots capital expenditures and operating cash flow, in dollars, for the five largest hyperscalers using the left axis. The line, plotted against the right axis, shows the ratio of capital expenditures to operating cash flow. The figure illustrates the growing capital intensity of the AI buildout and the extent to which capital expenditures are absorbing an increasing share of internally generated funds.

investors. By contrast, many of the firms driving demand for frontier AI, such as OpenAI and Anthropic, do not have a public credit rating and would not, on their own, be able to support large-scale infrastructure financing. In practice, these firms access compute capacity through agreements with hyperscalers, effectively renting their balance sheet strength and creditworthiness.

This shift has opened the door to a broader set of capital providers. Data center developers, real estate investment trusts (REITs), and infrastructure funds play a growing role in building and owning facilities, often entering into long-term lease agreements with hyperscalers or other large tenants. Data centers have recently become a core asset class for institutional real estate investors (Green Street, 2026). These arrangements with third-party providers allow technology firms to secure access to capacity without bearing the full upfront capital cost, while providing investors with exposure to long-duration, contract-backed cash flows.

Debt financing has expanded alongside this equity participation. Traditional bank lending remains part of the capital structure, but a substantial share of funding now comes from private credit markets and structured finance vehicles. Private credit funds, in particular, have emerged as major providers of capital, attracted by the scale of the opportunity and the perceived stability of contracted revenues. In addition, some projects are financed through securitized structures, in which claims on data center cash flows are packaged and sold to investors.

A further extension of this model is the emergence of financing structures

backed directly by IT hardware. In addition to shifting data center ownership off balance sheet, hyperscalers and AI firms are beginning to explore ways to finance GPUs and related equipment through asset-backed or lease-based structures.[2] These arrangements resemble established forms of equipment finance, such as aircraft leasing, in which long-lived assets generate predictable cash flows that support external funding. However, the analogy is imperfect. Compared with aircraft or other traditional collateral, AI hardware has a shorter economic life and faces substantial technological obsolescence risk, as it depends more directly on the evolution of model architectures and demand for compute. As a result, financing structures backed by GPUs may carry greater residual value risk, even as they further expand the scope for distributing AI infrastructure exposure across a broader set of investors.

As a result, a significant portion of the AI infrastructure buildout—spanning both physical facilities and, increasingly, the underlying compute hardware—is financed externally rather than directly by the firms that ultimately use the compute. Estimates by Morgan Stanley Research suggest that more than half of the roughly $2.9 trillion in investment required to meet hyperscalers' additional compute needs over 2025–2028 will be funded by outside capital (Morgan Stanley Research, 2025).

Across the full investment, Morgan Stanley projects an approximate 60–40 split between equity and debt, with hyperscalers providing a substantial share of the equity capital alongside third-party private equity investors. Within the debt component, private credit is expected to account for the majority (about $800 billion, or roughly 70 percent), with the remainder coming from corporate debt (approximately $200 billion, 17 percent) and structured finance (about $150 billion, 13 percent).

The aggregate 60–40 equity-debt split, however, understates the degree of leverage at the asset level. Much of the internal equity provided by hyperscalers is used to finance IT equipment, while a large share of external debt is tied to the construction of data centers and associated power infrastructure. As a result, leverage on these physical assets is likely to be significantly higher, often in the range of 70 to 80 percent, consistent with standard project finance and real estate lending practices. In this sense, the shift toward external, asset-backed financing may increase effective leverage relative to a model in which infrastructure is financed directly on hyperscalers' balance sheets through unsecured corporate debt.

The resulting structure has several important economic implications. First, ownership of AI infrastructure is increasingly dispersed across a broad set of investors, allowing the sector to tap into deeper pools of capital and potentially improving diversification of the funding base. Second, this expansion is accompanied by a growing complexity in the number and structure of claims on underlying cash flows, as equity holders, private credit funds, banks, and structured finance vehicles each hold distinct contractual positions. Third, the allocation of financing across

[2]Recent credit ratings agency research points to growing investor demand for AI-linked collateral, including both infrastructure and equipment-backed exposures (Moody's Ratings, 2025).

asset classes, where IT equipment is largely financed with internal equity while real estate and power infrastructure are financed with external asset-backed debt, can lead to higher effective leverage at the asset level than suggested by hyperscaler balance-sheet leverage measures.

Taken together, these developments separate the users of compute from the owners of physical assets and from the ultimate bearers of financial risk. Hyperscalers and AI firms contract for capacity through leases and service agreements, while a diverse set of investors holds long-duration claims backed by infrastructure, real estate, and IT hardware assets. Understanding how this separation reshapes the distribution of risk across investors and institutions is the focus of the next section.

IV. Financial Architecture and Risk

The financial structure that has emerged around AI infrastructure separates the users of compute from the owners of physical assets and from the ultimate bearers of financial risk. Hyperscalers and AI firms contract for capacity, while a wide range of investors, through equity stakes, private credit, and structured finance vehicles, hold claims on the underlying cash flows. This separation enables the rapid scaling of infrastructure by drawing on a broad pool of capital. At the same time, it reshapes how risk is distributed, priced, and monitored within the financial system.

Importantly, this redistribution of risk does not imply a reduction in aggregate risk. Instead, it changes who holds the risk and how transparent those exposures are. The resulting system bears similarities to other forms of securitized and infrastructure finance, in which long-duration assets are funded through layered claims held by diverse investors. Understanding these mechanisms is essential for assessing both the financial capacity for a further AI buildout and its potential financial vulnerabilities.

A useful illustration of these dynamics is provided by the recently announced Hyperion data center financing. The project involves the development of approximately 2.0 GW of capacity, with a total investment of around $30 billion, making it one of the largest single data center developments to date. Meta was initially the sole owner of the project but subsequently sold an 80 percent equity stake to the private credit firm Blue Owl for approximately $2.5 billion. The resulting joint venture, named Beignet, then raised $27 billion in external debt in October 2025. The debt was rated A+ by Standard & Poor's, one notch below Meta's own corporate credit rating, making it the largest individual investment-grade corporate debt issuance in U.S. history.

Importantly, the capital structure implies a leverage ratio of roughly 90 percent debt, computed as $27 billion in debt out of $30 billion in asset value, far above what would typically be observed on the balance sheet of an investment-grade corporate bond issuer.[3]

[3]Indeed, at the end of 2025, Meta's own on-balance sheet leverage ratio was 25% when using

At the same time, the structure exhibits a high degree of financial engineering. The project is organized through a network of bankruptcy-remote special purpose vehicles (SPVs), illustrated in Figure 2, which issue debt backed by contractual claims on future lease payments. This structure increases the opacity of the system by making it more difficult to trace ultimate exposures. The lease contract itself reflect a tension between the needs of capital providers and tenants. While lenders require long-duration, stable cash flows to support large-scale financing, tenants value flexibility in a rapidly evolving technological environment. In this case, the solution takes the form of a sequence of shorter-term leases, supplemented by a residual value guarantee that replicates most aspects of a longer-term commitment.

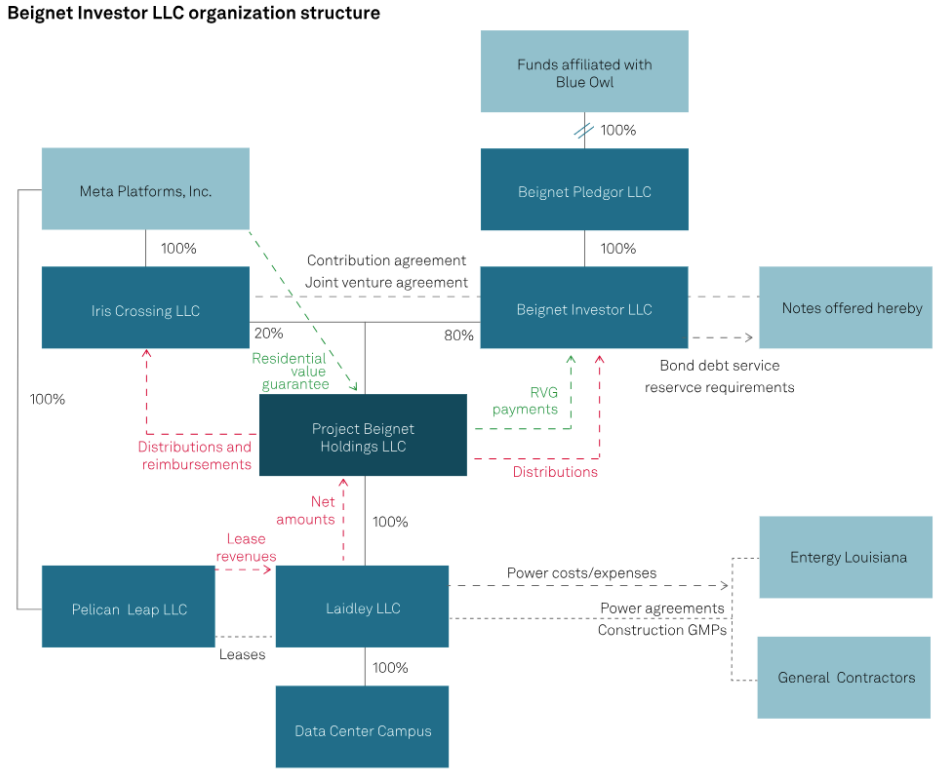

Figure 2. Structure of the Hyperion (Beignet) Data Center Financing

Notes: The figure illustrates the ownership and financing structure of the Hyperion data center project.

Specifically, Meta is the sole tenant of the Hyperion data center and has committed to five successive four-year leases, beginning in 2029–when the facility is

the book value of equity and 4% when using the market value of equity.

expected to become operational–and extending through 2049, when the Beignet bond matures. At each renewal date, Meta retains the option to terminate the lease for one or more portions of the campus. If it exercises this option, the underlying asset is offered for sale, and Meta must cover any shortfall between the realized sale value and a contractually specified guaranteed minimum value. This residual value guarantee (RVG) protects lenders by ensuring that the debt can be repaid even if the tenant exits. The RVG naturally has counterparty risk.

The arrangement reflects a trade-off between flexibility and financing. Shorter-term leases provide the tenant with optionality in a rapidly evolving technological environment, while the RVG restores the long-duration, predictable cash flows required by creditors.

At the same time, the accounting treatment of these commitments creates a form of regulatory arbitrage. Under current GAAP rules, Meta recognizes neither future lease obligations nor the RVG on its balance sheet, even though one of these two liabilities is certain to materialize. As a result, Meta's reported leverage understates the economic commitments embedded in the lease contract. This understatement will only be corrected at the time that either the lease is renewed or the RVG option is exercised. Moody's Ratings (2026*a*) finds that hyperscalers have about $970 billion in combined lease commitments, $660 billion of which are not reflected on balance sheet.

This off-balance sheet debt arrangement is not without cost. Beignet's debt was issued at a yield of 6.58 percent, at least 100 basis points higher than what Meta would likely have paid to issue an equivalent amount of unsecured corporate debt. Over the life of the bond, this difference translates in over $5 billion of additional interest expense. This higher cost of debt is ultimately passed through to Meta in the form of higher lease payments.

The benefit to Meta lies in preserving an asset-light balance sheet. By financing the project off balance sheet, the firm avoids adding substantial capital expenditures and debt that could put pressure on its credit rating and, more importantly, on its equity valuation. Meta is valued more like a high-growth software company than a capital-intensive infrastructure provider, with correspondingly higher valuation multiples. Maintaining that valuation generates gains that far outweigh the higher cost of debt associated with external, asset-backed financing.

The Hyperion-Meta transaction stands as a landmark deal that may well serve as a template for the next generation of data center finance. The transaction highlights how the separation of ownership, control, and risk-bearing can give rise to forms of regulatory and financial arbitrage. By locating assets and liabilities within off-balance sheet entities, firms can obtain large amounts of secured financing while preserving balance sheet flexibility and credit ratings. At the same time, the resulting structure increases leverage at the asset level and distributes exposure across a wide range of investors, raising questions about how risks are measured and monitored within the system.

V. Risks in the AI Infrastructure Buildout

The rapid expansion of AI infrastructure is supported by a financial structure that redistributes risk across a wide set of investors and institutions. While this structure facilitates investment at scale, it also introduces new forms of risk that are shaped by the technological, contractual, and financial characteristics of the sector.

A first source of risk arises from concentration and counterparty exposure. Many large data center projects are effectively single-tenant facilities, with hyperscalers serving as the primary or sole occupants. While these tenants currently have strong credit profiles, their financial positions are tied to the evolution of the AI ecosystem, including demand for AI services and the profitability of downstream applications. In addition, hyperscalers are themselves exposed to a relatively small number of large AI model developers and enterprise clients. This creates the possibility of correlated risks, in which a downturn in AI demand propagates through tenants to infrastructure investors. Given that hyperscalers frequently own part of their data center footprint, a contraction in compute demand may prompt a reallocation of activity from leased facilities to owned capacity.

A second risk relates to technological obsolescence. The rapid pace of innovation in AI hardware and system design implies that data centers built for current generations of chips may become less competitive over time. Increases in power density, cooling requirements, and interconnect technologies can render existing facilities costly to retrofit or upgrade. The threat is not limited to incremental hardware upgrades. More disruptive shifts, such as quantum computing for select computational workloads or edge computing that moves processing onto phones, vehicles, factories, and other local devices, could structurally weaken demand for large centralized data center campuses. As a result, assets that are financed over long horizons may face shorter effective economic lives than anticipated, creating a mismatch between asset durability and financial structure.

A third set of risks stems from power and infrastructure constraints. The viability of large AI facilities depends critically on access to reliable and scalable electricity, as well as on permitting and transmission infrastructure. Delays in grid interconnection, rising electricity costs, or regulatory opposition at the local level can affect both the timing and profitability of projects. These constraints introduce an additional layer of uncertainty that is largely outside the control of developers and tenants.

A fourth source of risk arises from potential bottlenecks in the supply chain for AI hardware. The expansion of compute capacity depends critically on the availability of advanced semiconductors and related components, including GPUs, high-bandwidth memory, and specialized networking equipment. These, in turn, rely on a highly concentrated and technologically complex global production chain, spanning chip design, fabrication, and the manufacture of advanced lithography equipment. Capacity constraints at any point in this supply chain can slow the pace of deployment. Geopolitical risks further amplify these concerns, as key segments of production are geographically concentrated, for example, the chip manufacturer

TSMC in Taiwan.

These supply-side constraints raise the possibility that the projected expansion of tens of gigawatts of AI infrastructure may not fully materialize. While such an outcome could increase the value of existing facilities by limiting future competition, it would pose risks for investors whose returns depend on the timely deployment and utilization of new capacity. In particular, debt holders in projects collateralized by data center leasing revenues may face shortfalls if the required hardware cannot be delivered at scale, causing actual compute capacity to fall below contracted or underwritten levels.

Financial structure introduces a fifth dimension of risk. As discussed in the previous section, the separation of ownership and use is often accompanied by high levels of asset-level leverage and by complex contractual arrangements. While these features allow more capital to flow into the sector, they can also amplify losses in adverse scenarios. These dynamics echo a broader literature on securitization, which emphasizes both its ability to expand credit supply and distribute risk, as well as the incentive and information problems it can create. DeMarzo (2005) shows how securitization can improve risk sharing but may weaken monitoring incentives, while Coval, Jurek and Stafford (2009) highlight how structured finance can obscure tail risks and amplify losses when underlying assumptions fail. Lenders who originate construction loans or mortgages for data centers with the intention of selling the loans to structured finance vehicles may lower underwriting discipline. For early evidence in this direction,Gete and Mercader (2026) argue that an originate-to-distribute model is emerging in datacenter mortgage lending. The combination of high leverage, long-duration debt, and uncertain future demand creates the potential for valuation adjustments that are transmitted across multiple layers of investors.

More broadly, the interaction of these risks may give rise to systemic effects. The AI ecosystem is characterized by increasing circularity, with interlocking strategic relationships along the value chain linking chip designers, cloud providers, AI model developers, and data center operators. These interdependencies create channels through which shocks to one segment, for example, a decline in AI demand or financial stress among model developers, can propagate across the entire system.

Recent asset pricing evidence is consistent with increasing systematic risk. The estimated stock market beta of publicly-traded data center REITs has risen from roughly 0.5 in earlier periods to around 1 in recent years, as shown in the black line in Figure 3, indicating that these assets now co-move closely with the broader equity market. This shift suggests that data centers no longer behave as relatively defensive infrastructure investments, but instead are increasingly exposed to the same macroeconomic and technological risks that affect large technology firms. Given the central role of these firms in equity markets, the risks associated with AI infrastructure may therefore be more systemically important than what traditional real estate or infrastructure investors may be accustomed to.

Data center stocks also increasingly comove with the non-AI component of the overall stock market. This undermines the case for data centers as a good hedge for

AI-related risk. Over the past three years, the exposure of data centers to non-AI stocks has become as high as its exposure to AI stocks.

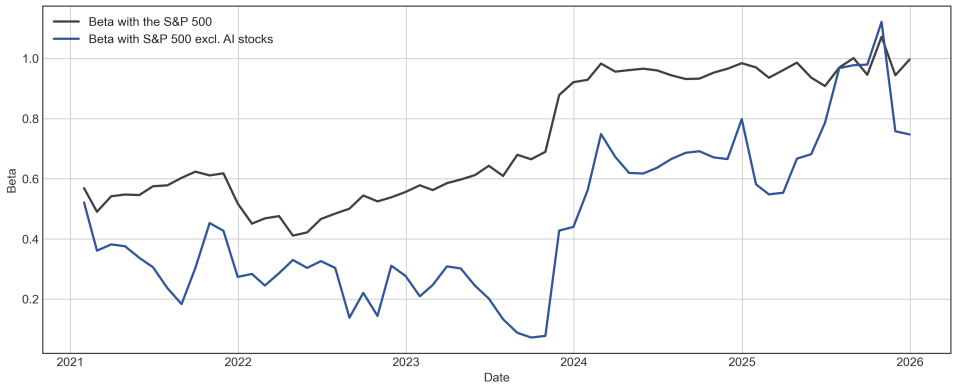

Figure 3. Rolling Betas of Datacenter Returns with the Stock Market

Notes: The black line plots the 36-month rolling-window beta of data center returns with the S&P 500 value-weighted excess return. The blue line plots the 36-month rolling-window beta of data center returns with the component of the S&P 500 that excludes all AI-related stocks. The dependent variable is the value-weighted return on data center REITS, from the FTSE Nareit U.S. Real Estate Index Series, minus the Treasury bill rate. The independent variables are the return on the S&P 500 minus the T-bill rate, the 10-year U.S. Treasury total return minus the T-bill rate, and the size, value, profitability, and investment factors of the Fama-French five-factor model from Kenneth French data library. In the second specification, underlying the blue line, we replace the excess return on the S&P with the excess return on the S&P500 excluding AI stocks and the excess return on the AI stocks in the S&P500. AI stocks are defined as the constituents of the Global X's AIQ ETF on March 18, 2026. The sample is monthly from January 2018 until December 2025.

None of these risks imply that adverse outcomes are imminent. A more optimistic view emphasizes the strength of underlying demand for AI compute, the continued scaling of model capabilities, and the strategic importance of AI infrastructure to both private firms and governments. Under this scenario, high utilization rates, sustained growth in AI applications, and ongoing technological improvements support strong and stable cash flows for infrastructure assets. In addition, the involvement of well-capitalized hyperscalers and institutional investors may enhance the resilience of the system.

At the same time, the scale and structure of the current buildout make it important to understand how risks could materialize under less favorable conditions. Historical experience with large infrastructure booms, from railroads to telecom, shows that periods of rapid, capital-intensive expansion accompanied by growing leverage are often followed by financial stress when demand falls short of expectations or financing conditions tighten. The interaction of technological uncertainty, financial leverage, and complex contractual arrangements in the AI sector suggests that similar dynamics could emerge. Even if such outcomes do not materialize,

careful consideration of their probabilities is essential for the adequate pricing of risk and for evaluating the long-term sustainability of the AI infrastructure boom.

VI. Policy and Research Questions

The rise of AI infrastructure raises questions that cut across industrial organization, corporate finance, accounting, and financial stability. A central lesson of this paper is that the physical capital of AI cannot be understood by looking only at the reported capital expenditures of large technology firms. The relevant investment is distributed across buildings, power systems, cooling equipment, chips, leases, guarantees, and project-level financing vehicles. The first task for researchers and policymakers is therefore conceptual as much as empirical: to define what constitutes AI infrastructure exposure and to identify where, in economic rather than purely legal terms, that exposure resides.

A natural starting point is measurement and disclosure. Existing data sources provide only partial views of the buildout, and no single framework captures the full capital stack linking compute users to the physical assets and financial claims that support them. At the same time, current accounting practices were not designed for long-duration infrastructure financed through short-term leases, contingent guarantees, and off-balance-sheet entities. As a result, economically meaningful commitments may appear only gradually in financial statements. Improving transparency—through better mapping of exposures and enhanced disclosure of contractual features such as lease optionality and residual value guarantees—would help investors and regulators assess where risks are accumulating.

A second set of questions concerns valuation and capital structure. The assets underlying AI infrastructure combine long-lived physical systems with rapidly evolving hardware, making their economic life and recovery value difficult to assess. At the same time, firms have strong incentives to separate infrastructure-intensive investments from their balance sheets in order to preserve high equity valuations. This raises broader questions about how leverage should be measured, how collateral should be valued, and whether standard performance metrics adequately capture the true capital intensity of AI-related activities.

A related set of questions concerns the distribution of gains from AI and the role of financial markets in shaping them. Kogan, Papanikolaou and Stoffman (2020) argue that investors can partially hedge the risk of technological disruption by holding claims on innovative firms, but that such hedging is inherently incomplete because innovators must retain equity stakes for incentive and control reasons. As a result, periods of rapid innovation tend to be associated with rising wealth inequality. The current AI buildout raises a similar issue. To the extent that households can gain exposure to AI through public equities, infrastructure funds, or credit instruments linked to data centers and compute, broader participation in the financing of AI may provide a partial hedge against the labor-displacement risks of AI. At the same time, if access to these investments is uneven, or if the most valuable claims

remain concentrated among a small set of firms and insiders, the expansion of AI infrastructure could reinforce existing patterns of wealth inequality.

Finally, the scale and financial structure of the AI buildout raise questions about financial stability. It would be premature to conclude that the sector poses systemic risk in the sense associated with past credit booms. Yet historical experience with large infrastructure expansions suggests that periods of rapid, capital-intensive growth can become financially consequential when expectations shift. The relevant question is therefore not whether risks are already systemic, but under what conditions they could become so, and how they would propagate across a system characterized by high leverage, complex contractual arrangements, and dispersed ownership.

These considerations suggest that AI infrastructure should be viewed not as a narrow industry topic, but as a case study in how modern economies finance general-purpose technologies. The current buildout combines large-scale physical investment with increasingly sophisticated financial structures, linking technological progress to the organization of capital markets. Because these institutions and contracts are still evolving, the present moment offers a valuable opportunity for research before their long-run consequences are fully realized.

REFERENCES

Brandsaas, Eirik Eylands, Daniel Garcia, Robert Kurtzman, Joseph Nichols, and Adelia Zytek. 2025. "Estimating Aggregate Data Center Investment with Project-level Data." Board of Governors of the Federal Reserve System 2025-109.

Chung, Hee Kwon, and Laura Veldkamp. 2024. "Data and the Aggregate Economy." *Journal of Economic Literature*.

Coval, Joshua, Jakub Jurek, and Erik Stafford. 2009. "The Economics of Structured Finance." *Journal of Economic Perspectives*, 23(1): 3–25.

DeMarzo, Peter M. 2005. "The Pooling and Tranching of Securities: A Model of Informed Intermediation." *Journal of Finance*, 60(1): 1–35.

Eberly, Janice C. 2019. "The Rise of Intangible Capital." *Journal of Economic Perspectives*, 33(3): 3–26.

Eisfeldt, Andrea L., Gregor Schubert, and Miao Ben Zhang. 2023. "Generative AI and Firm Values." National Bureau of Economic Research Working Paper 31222.

Farboodi, Maryam, and Laura Veldkamp. 2026. "A Model of the Data Economy." *Review of Economic Studies*.

Fernald, John G. 1999. "Roads to Prosperity? Assessing the Link between Public Capital and Productivity." *American Economic Review*, 89(3): 619–638.

Gargano, Antonio, and Marco Giacoletti. 2025. "Subsidizing the Cloud: U.S. State Incentives to Data Centers." Working paper / SSRN 5881105.

Gete, Pedro, and Amparo Mercader. 2026. "Datacenter Mortgages and Originate-to-Distribute." Working paper.

Green Street. 2026. "Data Center Outlook." Sector outlook report.

Haskel, Jonathan, and Stian Westlake. 2017. *Capitalism without Capital: The Rise of the Intangible Economy.* Princeton University Press.

Jones, Charles I., and Christopher Tonetti. 2020. "Nonrivalry and the Economics of Data." *American Economic Review*, 110(9): 2819–2858.

Jovanovic, Boyan, and Peter L. Rousseau. 2005. "General Purpose Technologies." *Handbook of Economic Growth*, 1: 1181–1224.

Kogan, Leonid, Dimitris Papanikolaou, and Noah Stoffman. 2020. "Technological Innovation, Resource Allocation, and Growth." *Quarterly Journal of Economics*, 135(2): 665–712.

McKinsey & Company. 2025. "The Cost of Compute: A $7 Trillion Race to Scale Data Centers." Published April 28, 2025.

Moody's Ratings. 2025. "Corporate ABS – U.S.: 2026 Outlook — AI and Used Asset Demand to Aid ABS Quality and Collateral Values." Ratings outlook.

Moody's Ratings. 2026*a*. "Accounting – U.S.: Hyperscalers' Reported AI-related Lease Commitments May Understate Economic Risk." Accounting commentary.

Moody's Ratings. 2026*b*. "Data Centers – Global: 2026 Outlook — Capacity Growth Remains Robust as Tenants Prioritize Speed to Market." Ratings outlook.

Morgan Stanley Research. 2025. "AI Is Now a Macro Variable." *https://www.morganstanley.com/insights/articles/ai-market-trends-institute-2026*, Accessed 2026.