Brookings Papers
ON ECONOMIC ACTIVITY

BPEA Conference Draft, September 24-25, 2026

Financing the AI Buildout

Stijn Van Nieuwerburgh (Columbia Business School)

STIJN VAN NIEUWERBURGH
Columbia Business School

Financing the AI Buildout

Draft: September 4, 2026

ABSTRACT Artificial intelligence is driving a large expansion in data center capacity, power infrastructure, and specialized computing equipment. We estimate that a 1 GW AI campus costs about $41 billion. A 183 GW U.S. buildout completed by 2032, together with investment in projects completed later, would require annual investment averaging roughly 3.6 percent of GDP over 2025–2032. As hyperscaler capital expenditures outgrow internal cash flow, financing is shifting toward leases, joint ventures, project debt, private credit, securitization, and special-purpose vehicles. These structures expand funding capacity but also raise asset-level leverage, obscure contingent obligations, and expose investors to tenant concentration, technological obsolescence, execution delays, and uncertain residual values.

I. Introduction

Artificial intelligence is often described as an intangible revolution. The near-term economic footprint of AI, however, is strikingly physical. Training and deploying frontier models requires large quantities of specialized chips, electricity, cooling, network equipment, and purpose-built data center capacity. The current AI boom is therefore not only a software story; it is also a large-scale infrastructure investment cycle.

This physical dimension matters because AI workloads are changing both the scale and composition of digital capital. Modern AI campuses operate at far higher power densities than traditional data centers and require tighter integration among computing hardware, buildings, cooling systems, and power infrastructure. Many existing facilities cannot accommodate these demands without extensive retrofits. Access to reliable power, transmission capacity, and suitable sites has consequently become as important to the expansion of AI as advances in models and chips.

The required investment is enormous. A representative 200 MW AI training campus costs roughly $8.2 billion, about two-thirds of which reflects IT equipment and one-third the real estate and associated power infrastructure. We estimate that 183 GW is completed by 2032 with a further 118 GW completed after 2032. This pipeline will trigger about $10.3 trillion dollars of investment over 2025–2032. At an average of 3.63 percent of GDP per year, the projected buildout would be larger relative to the economy than the major U.S. canal, railroad, electrification, highway, and telecommunications investment booms. It would double the electricity consumption of the entire U.S. residential sector.

Financing this investment increasingly requires more than the internal cash flows of the firms demanding the compute. Aggregate capital expenditures by Oracle, Microsoft, Amazon, Meta, and Alphabet rose from about $97 billion in 2020 to more than $400 billion in 2025 and are projected to exceed $800 billion in 2026, surpassing their combined operating cash flow for the first time. Hyperscalers therefore increasingly combine direct ownership with leases, joint ventures, project finance, private credit, and structured-finance vehicles.

This shift expands financing capacity by separating the users of compute from the owners of the underlying assets. Data center developers, infrastructure funds, REITs, and private equity investors provide equity capital, while banks, private credit funds, and securitization vehicles supply debt. The same logic is beginning to extend from buildings and power infrastructure to GPUs and other IT equipment. Investment-grade hyperscaler tenants make these structures financeable by supporting long-duration contractual cash flows, while frontier AI firms effectively access infrastructure through the hyperscalers' balance sheets and creditworthiness.

These arrangements transform rather than eliminate risk. Moving assets into separately financed vehicles can raise leverage on the underlying infrastructure even when reported hyperscaler leverage remains low. Long-duration debt is then supported by cash flows and collateral values that depend on uncertain AI demand, rapid technological change, timely access to power and hardware, and the continued credit quality of a small number of tenants. The resulting capital structure can therefore make exposures more layered, correlated, and difficult to observe.

The paper combines three elements. First, it quantifies the physical and financial scale of the AI infrastructure buildout and compares it with earlier U.S. infrastructure booms. Second, it describes the emerging capital stack, including the growing roles of third-party ownership, leases, project finance, private credit, and structured-finance vehicles, and uses Meta's Hyperion transaction to illustrate these arrangements in detail. Third, it analyzes how this financial architecture allocates and potentially amplifies technological, operating, and demand risk, with particular attention to asset-level leverage, contingent obligations, and the visibility of exposures outside hyperscaler balance sheets.

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 complementary physical capital required for new technologies to generate productivity gains (Jovanovic and Rousseau, 2005; Fernald, 1999; Acemoglu, 2025; Baily et al., 2025). Recent work on the AI investment race similarly highlights how competition for future rents can generate exceptionally large investment responses under substantial uncertainty (Rungcharoenkitkul, 2026). Second, the paper 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 et al., 2009). In this respect, it is also related to recent analysis of the financing structures and potential vulnerabilities surrounding data center development (Arun, 2025; Gete and Mercader, 2026). Finally, the paper connects to the literature on intangible capital (Haskel and Westlake, 2017; Crouzet et al., 2022; Eisfeldt et al., 2023) and the data economy (Farboodi and Veldkamp, 2026; Chung and Veldkamp, 2024; Jones and Tonetti, 2020), which emphasizes the role of data as an input into 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 technological, physical, and financial forces interact at unusually large scale.

The paper proceeds as follows. Section II describes the scale, physical requirements, and economic characteristics of AI infrastructure. Section III examines how this infrastructure is owned and financed, with particular emphasis on external capital and the evolving capital stack. Section IV illustrates this financial architecture through the Hyperion transaction and shows how project-level structures allocate ownership, leverage, and risk. Section V analyzes the main technological, operating, and financial vulnerabilities and relates them to earlier infrastructure cycles. Section VI concludes by outlining the principal policy and research questions raised by the AI buildout.

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 the AI buildout are data centers designed for high-performance computing. Unlike earlier facilities optimized for storage, web hosting, and enterprise applications, modern AI campuses concentrate large numbers of GPUs and high-speed interconnects in a single location. This sharply raises rack-level power density and requires more demanding electrical and cooling systems. Whereas traditional CPU-based racks often consumed 5–10 kilowatts (kW), current rack-scale AI systems can draw well above 100 kW and increasingly rely on direct-to-chip liquid cooling.

These demands become substantial at the campus level. A facility drawing 200 megawatts (MW) of total power supports approximately 166.7 MW of critical IT capacity at a power usage effectiveness ratio of 1.2. At roughly 142 kW per NVIDIA GB300 NVL72 rack, this corresponds to approximately 1,170 high-density AI racks operating simultaneously. A continuous 200 MW load would consume about as much electricity annually as 170,000 average U.S. households. This highly concentrated, nearly continuous demand requires dedicated substations, high-voltage connections, and significant grid-interconnection investment.

The associated capital requirements are correspondingly large. We estimate approximately $2.2 billion for the data center facility and $0.4 billion for incremental power infrastructure. The installed IT system costs approximately $5.6 billion, including the integrated GPU racks (at $4 million per rack), external networking fabric, shared storage, and system integration. Appendix A provides the underlying assumptions and cost breakdown.

Taken together, these estimates imply an initial capital cost of approximately $8.2 billion for a 200 MW AI training campus. About one-third reflects the facility and power infrastructure, while roughly two-thirds consists of compute hardware and related IT systems. Thus, the economic center of gravity has shifted toward the compute layer, even as the feasibility of that investment remains tightly constrained by physical infrastructure.

Aggregating across projects, the scale of the data center buildout becomes enormous. Recent work has begun to construct comprehensive measures of data center investment using project-level information; Brandsaas et al. (2025), for example, use data on planned and ongoing developments to nowcast aggregate investment. Using Cleanview project-level data from July 2026, we estimate approximately 57 GW of operating U.S. data center capacity and a planned development pipeline of approximately 509 GW. The latter is not a completion forecast: projects may be delayed, reduced in scale, or never completed.[1]

To translate the announced pipeline into a capacity scenario, we allocate planned capacity among completion by 2032, completion after 2032, and non-realization, with lower realization rates for more distant and less fully documented projects. Our central scenario implies that 182.8 GW of additional U.S. capacity becomes operational during 2025–32, including 19.0 GW associated with projects already operational in 2025 or 2026. A further 117.2 GW is completed after 2032, while 226.9 GW is never completed. This is a scenario rather than a forecast and is particularly sensitive to large, multiphase projects and projects without reported completion dates.

To estimate annual investment, we apply the benchmark of approximately $8.2 billion per 200 MW, allow costs to grow by 4 percent annually, and allocate project costs across the completion year and the preceding three years. The resulting $10.3 trillion represents investment incurred during 2025–32, including pre-completion spending on projects that become operational after 2032. Assuming nominal GDP also grows by 4 percent annually, investment averages 3.63 percent of GDP over 2025–32. Table 1 places this estimate in historical perspective: it exceeds the corresponding investment shares associated with the canal, railroad, electrification,

[1] Appendix B describes the treatment of canceled projects, missing capacity and completion years, and the realization assumptions applied to planned projects.

highway, and telecommunications and fiber booms. Other sources likewise project a multi-trillion-dollar AI infrastructure buildout (McKinsey & Company, 2025; Morgan Stanley Research, 2025; Moody's Ratings, 2026b).

Table 1. Selected U.S. Capital-Expenditure Booms

Episode	Period	Size (% of GDP)
Canals	1836–1841	0.66
Railroads	1870–1890	2.24
Electrification	1905–1925	0.50
Highways	1956–1973	1.13
Telecom and fiber	1996–2003	1.10
AI buildout	2025–2032	3.63

Notes: Size is average annual capital expenditure as a percentage of GDP over the indicated period. Appendix C provides sources and further details.

The AI infrastructure boom also 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 substantial.

The boom is already visible in aggregate activity. In 2025 and the first quarter of 2026, AI infrastructure investment contributed meaningfully to U.S. GDP growth (Soto et al., 2026), illustrating the scale the expansion has already reached.

The physical nature of this investment also reshapes the geography of AI. Training large models involves massive, batch-style computations that are relatively insensitive to latency and can therefore be located far from end users. Inference workloads, by contrast, often require faster response times and greater proximity to population centers and established network hubs.

For large AI training facilities, the primary constraint is increasingly access to reliable, scalable, and inexpensive power. Grid capacity, transmission access, and permitting can take years to secure, pushing development toward regions with available energy, favorable regulation, and room to expand. Established network hubs nevertheless remain important for latency-sensitive inference. State and local subsidies, often provided through sales-tax exemptions on data center equipment, also influence location choice (Gargano and Giacoletti, 2025).

The scale and physical constraints of the AI buildout raise a distinct set of economic questions: how this infrastructure is financed, who ultimately owns it, and where the associated risks reside.

III. Ownership and Financing of AI Infrastructure

The scale of the AI infrastructure buildout raises a basic financing question: who supplies the capital and who ultimately owns the underlying assets? Hyperscalers and large technology firms remain the principal sources of demand, but they increasingly rely on a mix of internal funds, leases, joint ventures, and third-party capital. As investment needs have accelerated, financing the buildout entirely on balance sheet has become more costly in terms of leverage, liquidity, and strategic flexibility.

Figure 1 shows how quickly this financing constraint is tightening. Aggregate capital expenditures by Oracle, Microsoft, Amazon, Meta, and Alphabet rose from roughly $97 billion in 2020 to more than $400 billion in 2025 and are projected to exceed $800 billion in 2026. On current estimates, capex will surpass the five firms' combined operating cash flow for the first time in 2026. The figure therefore marks an important transition: internal cash generation remains substantial, but it is no longer sufficient to finance the projected pace of investment without greater reliance on external capital or financing structures that shift assets and obligations away from the operating companies' balance sheets.

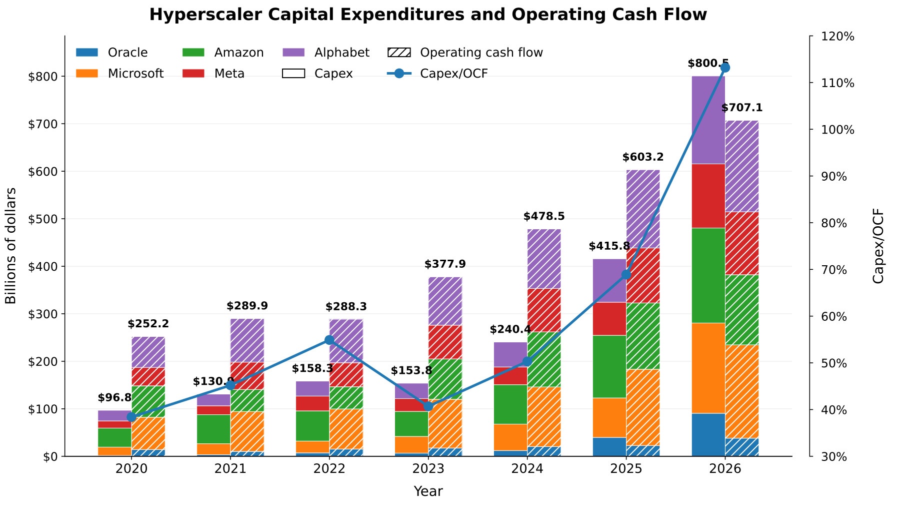

Figure 1. Hyperscaler Capital Expenditures, Operating Cash Flow, and Capex Intensity

Notes: The figure plots capital expenditures and operating cash flow, in billions of dollars, for Oracle, Microsoft, Amazon, Meta, and Alphabet using the left axis. The line, plotted against the right axis, shows capital intensity, measured as the ratio of aggregate capital expenditures to aggregate operating cash flow. Historical values are constructed from company 10-K and 10-Q SEC filings and measure cash capital expenditures and GAAP operating cash flow. Microsoft and Oracle are converted from fiscal-year to calendar-year values using quarterly and year-to-date filings. The 2026 values combine reported results with the latest company guidance and Wall Street estimates. Hyperscaler capex includes non-AI related capex but excludes capex for data centers leased by the hyperscalers.

The ability to attract outside capital rests heavily on the credit quality of the firms contracting for capacity. Amazon, Microsoft, Meta, Alphabet, and Oracle are investment-grade borrowers, and long-term agreements with these tenants can generate predictable cash flows that support real estate, infrastructure, and project-level financing. Frontier AI firms such as OpenAI and Anthropic generally lack comparable standalone credit profiles. They therefore obtain compute largely through contracts with hyperscalers, effectively relying on the hyperscalers' balance sheets and creditworthiness.

Historically, hyperscalers financed much of their data center investment through retained earnings and unsecured corporate debt, owned the facilities directly, and bore the associated construction, utilization, and residual-value risk. That model remains important, especially for IT equipment, but the marginal financing model is shifting. Specialized data center developers, REITs, infrastructure funds, and private equity investors increasingly build or acquire facilities and lease capacity to hyperscalers and other large users. These arrangements allow technology firms to secure capacity without supplying all of the upfront equity or issuing the associated debt directly, while giving outside investors exposure to long-duration, contract-backed cash flows. Data centers have consequently become a major institutional real estate and infrastructure asset class (Green Street, 2026).

Debt financing has expanded alongside this third-party equity. Bank loans, comprised of mortgage loans backed by datacenter assets and syndicated lines of credit, remain important, but private credit and structured finance now provide a growing share of project funding, usually backed by long-term contracted revenues. Financing is also extending beyond buildings and power infrastructure to the underlying IT equipment. GPUs and related hardware can be financed through leases or asset-backed structures, further broadening the pool of external capital available to AI firms and hyperscalers. Yet this collateral differs from conventional real estate or aircraft: its economic life is shorter and rapid technological change creates substantial obsolescence risk, making its resale value less certain (Moody's Ratings, 2025). More broadly, these arrangements allow the sector to draw on investors with different risk appetites, liabilities, and investment horizons, increasing the amount of capital available while reducing the immediate funding burden on hyperscaler balance sheets.

Economically, these structures substitute project-level debt and third-party equity for hyperscaler equity and corporate debt. Yet hyperscalers often remain central to the financing through leases, capacity commitments, guarantees, and their underlying credit quality. The shift therefore broadens funding capacity while relocating ownership and leverage, rather than fully transferring the underlying economic risk.

The scale of external financing is already substantial. Morgan Stanley estimates that more than half of the roughly \$2.9 trillion required to meet hyperscalers' incremental compute needs over 2025–2028 will come from outside capital (Morgan Stanley Research, 2025). Across the full investment, it projects an approximate 60–40 split between equity and debt. Within the debt component, private credit accounts for the largest share—about \$800 billion—followed by corporate debt of roughly \$200 billion and structured finance of approximately \$150 billion.

The aggregate 60–40 equity-debt split, however, can understate leverage at the asset level. Hyperscalers often direct a large share of their internal capital toward IT equipment, while data center shells, power infrastructure, and related real estate are financed through project-level debt, leases, and other asset-backed structures. Moving these assets into separately financed vehicles can therefore raise loan-to-value and debt-to-equity ratios on the underlying infrastructure, even if the hyperscalers' own reported leverage ratios remain relatively low. The shift toward external financing thus expands funding capacity but may also increase leverage within the broader AI infrastructure system and make that leverage less visible in operating-company balance sheets.

The resulting capital structure separates the users of compute from the owners of the underlying assets and, increasingly, from the investors who ultimately bear the financial risk. Hyperscalers and AI firms secure capacity through leases, service agreements, and joint ventures, while developers, infrastructure funds, private credit lenders, and structured-finance investors hold distinct claims on data centers, power assets, and IT equipment. This separation draws deeper pools of capital into the sector, enabling faster scaling, but it also creates a more layered and potentially opaque allocation of risk. The next section examines where those risks reside and how they may propagate across firms and financial institutions.

IV. Financial Architecture and Risk

A useful illustration of this financial architecture is the Hyperion data center financing. The project involves approximately 2.0 GW of capacity and a total investment of about $30 billion, making it one of the largest single data center developments to date.[2] Meta initially owned the project outright but subsequently sold an 80 percent equity stake to Blue Owl for approximately $2.5 billion. The resulting joint venture, Beignet, then raised $27 billion of external debt in October 2025. The debt was rated A+ by Standard & Poor's, one notch below Meta's own corporate credit rating, and represented the largest individual investment-grade corporate debt issuance in U.S. history.

The transaction is highly levered at the project level. Beignet financed roughly $27 billion of the $30 billion asset value with debt, implying a debt-to-asset ratio of about 90 percent. The debt service coverage ratio is only 1.12×. This contrast illustrates how off-balance-sheet project financing can support much higher leverage on the underlying infrastructure than would typically appear on the balance sheet of an investment-grade technology company.[3]

The financing is organized through a network of bankruptcy-remote special-purpose vehicles, shown in Figure 2, that issue debt backed by contractual claims

[2]This corresponds to about $15 million per MW, broadly consistent with our estimate of $11 million per MW for the facility plus additional power-infrastructure costs. The figure excludes IT equipment, which Meta will finance separately as the tenant.

[3]By comparison, Meta's own year-end 2025 leverage ratio was approximately 25 percent using book equity and just 4 percent using market equity. Meta's operating cash flow of $130 billion is at least 18× its 2026 debt service.

on future lease payments. This structure isolates the project from Meta's broader corporate balance sheet and enhances creditor protection, but it also makes ultimate exposures more difficult to trace. The central contractual challenge is to reconcile lenders' demand for long-duration, predictable cash flows with Meta's desire for flexibility in a rapidly changing technological environment.

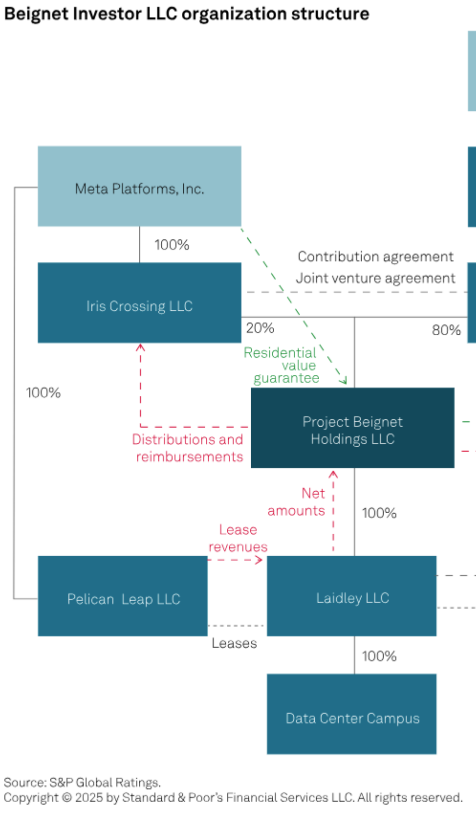

Figure 2. Structure of the Hyperion (Beignet) Data Center Financing

Notes: The figure illustrates the ownership and financing structure of the Hyperion data center project. Laidley LLC owns the datacenter campus, which is funded by its subsidiary Project Beignet Holdings LLC. The latter is owned in Joint Venture for 20% by Meta, through Iris Crossing LLC, and by Blue Owl, through Beignet Investor LLC. Laidley enters in a Power Purchase Agreement with Entergy Louisiana, which will develop three gas-fired generating units. Sources: Standard & Poor's pre-sale report, Entergy 2025 Form 10-K.

Meta addresses this tension through a sequence of five four-year leases beginning in 2029 and extending through the bond's 2049 maturity. At each renewal date, Meta may terminate the lease for some or all of the Hyperion campus. If it does so, the affected assets are offered for sale, and Meta must cover any shortfall between the sale proceeds and a contractually specified minimum value. This residual value guarantee gives Meta operational flexibility while preserving much of the long-duration credit support that lenders would otherwise obtain from a conventional long-term lease.

The accounting treatment of these commitments creates a further distinction between reported and economic leverage. Under current GAAP rules, Meta does not recognize the future lease obligations nor the residual value guarantee on its balance sheet before the relevant lease or guarantee becomes effective, even though one of the two commitments will ultimately materialize for sure. Meta's reported leverage therefore understates the economic obligations embedded in the contract. More broadly, Moody's estimates that hyperscalers have roughly $970 billion of lease commitments, of which about $660 billion relates to future leases not yet reflected on balance sheet (Moody's Ratings, 2026a).[4]

This off-balance-sheet structure is not costless. Beignet's debt was issued at a yield of 6.58%, at least 100 basis points above the rate Meta likely would have paid on comparable unsecured corporate debt at the time of issuance. Over the life of the financing, that spread implies more than $5 billion of additional interest expense for Beignet, a cost which is ultimately reflected in Meta's lease payments. The benefit is that Meta preserves balance-sheet capacity and avoids reporting the project's full capital intensity and leverage directly at the corporate level.

The economic rationale for Meta is therefore not cheaper financing, but preservation of an asset-light corporate profile. By placing the project in a separately financed vehicle, Meta avoids adding the full amount of the associated assets and debt to its own balance sheet and limits the effect on reported capital intensity, leverage, and credit metrics. For a firm valued partly on the basis of high growth and relatively low capital intensity, those benefits may outweigh the higher borrowing cost embedded in the lease.

Beignet bondholders receive a higher interest rate because they do not have a direct claim on Meta's balance sheet. Repayment instead depends on cash flowing through the project structure and on Meta's lease and residual-value guarantees, creating additional structural, contractual, liquidity, and enforcement risk. The bonds' longer duration and specialized private-placement structure also contribute to the interest premium.

The Hyperion transaction may therefore provide a template for future data center financings.[5] It combines investment-grade tenant credit, high project-level leverage, bankruptcy-remote entities, lease optionality, and a residual value guarantee to attract large amounts of external capital while preserving the tenant's balance-sheet flexibility. At the same time, the structure illustrates how financial engineering can make leverage and contingent obligations less visible and distribute exposure

[4] A recent analysis by the Wall Street Journal tallies $2.4 trillion in off-balance sheet liabilities for the four large hyperscalers (Microsoft, Alphabet, Amazon, and Meta): $904 billion in future lease commitments and $1.52 trillion in future purchase commitments, mostly of chips. This compares to $604 billion in on-balance sheet obligations: $356 billion in long-term debt and $248 billion in current lease obligations. Only one-fifth of the $3 trillion in datacenter-related obligations are currently recognized on hyperscalers' balance sheets.

[5] Indeed, the 960MW Sopaipilla datacenter project in El Paso, TX, a joint venture of Meta and Blackrock, follows a nearly identical structure. Sopaipilla investor issued $12.3 billion in A+-rated bonds in July 2026.

across investors whose claims depend on the same underlying tenant and asset values.

Although the Hyperion financing is contractually complex and contain numerous contingencies, investors at least benefit from a public credit rating and accompanying analysis by a rating agency. However, publicly rated debt represents only a relatively small share of datacenter financing; much larger volumes are provided through syndicated bank loans and private-credit facilities, for which little information is publicly available.

V. Risks in the AI Infrastructure Buildout

The central issue is not simply that AI infrastructure is exposed to technological and operating risks, but that the sector's financing structure can transmit and amplify those risks. Long-duration debt and contractual claims are being written against assets whose utilization, technological relevance, and residual value remain unusually uncertain. The severity of any adverse shock will therefore depend not only on the underlying economics of AI demand, but also on where leverage resides and how losses are allocated across tenants, asset owners, and creditors.

A first vulnerability is concentration in tenants and end demand. Many large AI campuses are effectively single-tenant facilities whose cash flows depend on a hyperscaler's continued willingness to lease and use the capacity. Although these tenants currently have strong credit profiles, their infrastructure commitments ultimately depend on the profitability of AI services and on demand from a relatively small set of model developers and enterprise customers. If compute demand disappoints, hyperscalers may also protect utilization of their owned facilities by shifting workloads away from leased capacity, concentrating the adjustment on third-party assets.

A second vulnerability is technological obsolescence. AI hardware, cooling systems, and facility design are evolving rapidly, so campuses built around current chip generations may become less competitive well before their physical structures wear out. Rising rack densities, new interconnect standards, and changing cooling requirements can require costly retrofits, while a shift toward more efficient large language models, edge computing, or other decentralized forms of inference could reduce demand for large centralized campuses. This creates a mismatch between the long maturities of infrastructure financing and the potentially much shorter economic lives of the underlying assets.

A third vulnerability lies in execution. Large AI campuses require timely access to electricity, transmission, permits, and specialized hardware, yet each of these inputs is capacity-constrained and often outside the control of developers. Grid interconnection delays, rising power costs, local opposition, or shortages of GPUs, high-bandwidth memory, and networking equipment can postpone completion and reduce utilization. Because advanced semiconductor production is concentrated in a small number of firms and locations, geopolitical disruptions could amplify these bottlenecks. For investors, the resulting risk is not only that project costs escalate, but that revenue begins later or never reaches the levels assumed at underwriting.

Under the central assumptions in Appendix D, the 2025–2032 buildout requires approximately \$3.7 trillion in annual revenue by 2032 to earn a 10% unlevered return at a 50% operating cash-flow margin. This is equivalent to roughly \$5.5 per installed GB300 GPU-hour at full utilization, or about \$6.9 per billed GPU-hour at 80 percent utilization, placing the required price in the broad range of current frontier-GPU rental prices. The underwriting risk is therefore whether demand and pricing power remain strong enough to sustain such prices once 180 GW of compute capacity has been added.

These underlying risks become more consequential because they are financed with high asset-level leverage and layered contractual claims. Debt service is fixed, while project revenues depend on utilization, lease renewals, and uncertain residual values. A relatively modest deterioration in demand, delays, or asset values can therefore produce much larger losses for equity holders and potentially impair creditors. The use of SPVs, guarantees, securitization, and originate-to-distribute lending can further weaken monitoring and make it harder to identify where exposures ultimately reside (DeMarzo, 2005; Coval et al., 2009; Gete and Mercader, 2026).

The risks are also linked across the AI value chain. Chip designers, cloud providers, model developers, data center operators, and infrastructure investors increasingly depend on one another through long-term contracts, strategic investments, and financing arrangements. A shock to one segment—such as weaker demand from model developers or financial stress at a major tenant—can therefore propagate to cloud revenues, lease payments, asset values, and creditor recoveries. This circularity makes exposures more correlated than they may appear when each transaction is viewed in isolation.[6]

Recent asset-pricing evidence is consistent with this greater exposure to common risk. Figure 3 shows that the market beta of publicly traded data center REITs has risen from roughly 0.5 to around 1, indicating that their returns now move much more closely with the broader equity market. Their exposure to the non-AI component of the S&P 500 has also increased and is now comparable to their exposure to AI stocks. Data centers therefore appear less like defensive infrastructure assets or a hedge against AI-related risk and more like part of the same systematic risk complex.

[6] The \$500 billion credit facility stood up by Apollo, BlackRock, Blackstone, Brookfield, Goldman Sachs, and KKR in August 2026 to finance NVIDIA chips is a good example. NVIDIA will provide a 25% residual value guarantee, exposing it to as much as \$125 billion of backstops. It illustrates the circularity created when a dominant equipment supplier helps finance customers' purchases of infrastructure built around its own products. Google and Broadcom entered into a similar arrangement to help finance acquisition of its TPU chips.

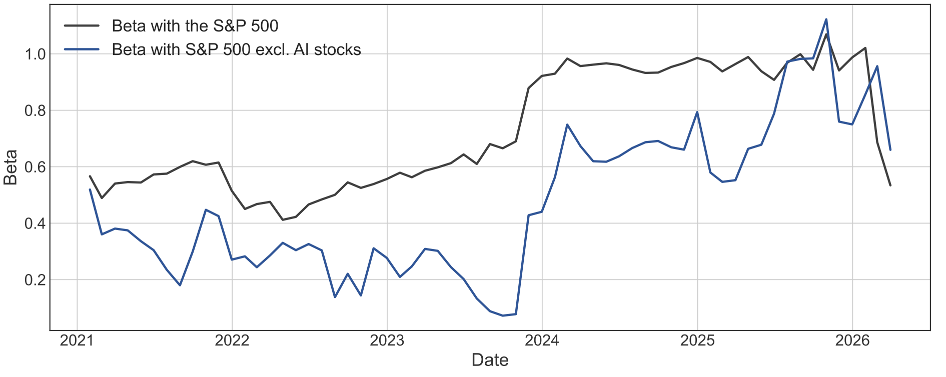

Figure 3. Rolling Betas of Data Center Returns with the Stock Market

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 March 2026.

These developments do not imply that financial distress is imminent. Strong growth in AI applications, high utilization, and continued improvements in model capability could support the projected infrastructure and generate stable cash flows. But the combination of uncertain demand, rapid technological change, execution bottlenecks, and high leverage creates meaningful downside risk if expectations are revised. Historical infrastructure booms—from railroads to telecommunications—show that large, capital-intensive expansions become financially consequential when demand falls short, financing conditions tighten, or there is a mismatch between the adoption of a new technology and the maturity of the debt issued to finance it. The relevant question is therefore not whether AI infrastructure is already systemically risky, but under what conditions project-level losses could become correlated and propagate across firms and financial institutions.

VI. Policy and Research Questions

The preceding analysis raises a set of policy and research questions that extend beyond the data center sector itself. The central challenge is that AI infrastructure combines large-scale physical investment with contractual and financial arrangements that are still evolving and are only partially visible in conventional corporate accounts. The most immediate priorities are therefore to improve measurement, clarify disclosure, understand asset valuation and capital-structure choices, and identify the conditions under which these exposures could become financially consequential.

A first challenge is measurement. The physical capital supporting AI is distributed across data center buildings, power systems, networking infrastructure, and specialized chips, while the associated financial exposure is distributed across corporate balance sheets, leases, joint ventures, guarantees, project debt, and special-purpose vehicles. Reported hyperscaler capital expenditures therefore capture only part of the buildout. A central task for researchers and policymakers is to construct measures that identify not only who legally owns each asset, but also who bears the underlying demand, refinancing, and residual-value risk. Ongoing work by Van Nieuwerburgh and Sun (2026) as well as private data providers such as Atrium.ai is trying to triangulate multiple data sources to begin to address this measurement gap.

A second set of questions concerns disclosure and accounting. Conventional measures of corporate leverage may be incomplete when economically important obligations migrate to separately financed vehicles or arise through future leases and contingent guarantees. More informative disclosure would map project-level debt, equity interests in joint ventures, future lease commitments, residual-value guarantees, construction guarantees, tenant concentration, renewal options, and the assumed recovery values of specialized assets. The objective need not be to eliminate off-balance-sheet financing, which can serve legitimate economic purposes, but to make the resulting commitments and risk transfers more transparent.

A third issue is valuation and capital-structure design. AI infrastructure combines long-lived buildings and power systems with hardware whose economic life may be much shorter and more uncertain. This raises basic questions about depreciation, collateral values, lease duration, refinancing risk, and the appropriate maturity of debt. It also creates incentives to separate infrastructure-intensive assets from high-growth operating companies in order to preserve reported capital intensity and valuation multiples. Research is needed to determine whether current accounting and performance measures adequately capture the economic cost of the capital employed.

A fourth set of questions concerns financial stability. It would be premature to conclude that AI infrastructure already poses systemic risk comparable to earlier credit booms. The more relevant questions are how much exposure is held by banks, insurers, private credit funds, securitization vehicles, and other institutional investors; how concentrated those claims are in the same tenants and how sensitive to AI demand forecasts; and how losses would propagate under a joint decline in utilization, collateral values, and refinancing capacity. Off-balance-sheet structures matter in this context not because they necessarily increase aggregate risk, but because they may make correlated exposures harder to observe before a downturn.

More broadly, the AI buildout offers an unusual opportunity to study how modern economies finance a general-purpose technology in real time. The institutions, contracts, accounting practices, and investor base supporting this infrastructure are still being formed. This creates a valuable research setting for understanding how financing capacity expands, how ownership and risk become separated, and when financial innovation improves risk sharing rather than obscuring leverage and tail exposure. The most important policy contribution at this stage may therefore be to improve measurement and transparency while the capital structure of the industry is still evolving.

Conflict of Interest Disclosure: The author has no conflicts of interest to disclose.

References

Acemoglu, D. (2025). The simple macroeconomics of AI. *Economic Policy 40*(121), 13–58.

Arun, A. (2025, November). Bubble or nothing: Data center project finance. Working paper, Center for Public Enterprise.

Baily, M. N., D. M. Byrne, A. Kane, and P. Soto (2025). Generative AI at the crossroads: Light bulb, dynamo, or microscope? *arXiv preprint arXiv:2505.14588*.

Brandsaas, E. E., D. Garcia, R. Kurtzman, J. Nichols, and A. Zytek (2025, December). Estimating aggregate data center investment with project-level data. Technical Report 2025-109, Board of Governors of the Federal Reserve System.

CBRE (2026). Global data center trends 2026. Technical report, CBRE. Accessed July 2026.

Chung, H. K. and L. Veldkamp (2024). Data and the aggregate economy. *Journal of Economic Literature*.

Couper, E. A., J. P. Hejkal, and A. L. Wolman (2003). Boom and bust in telecommunications. *Federal Reserve Bank of Richmond Economic Quarterly 89*(4), 1–24.

Coval, J., J. Jurek, and E. Stafford (2009). The economics of structured finance. *Journal of Economic Perspectives 23*(1), 3–25.

Cranmer, H. J. (1960). Canal investment, 1815–1860. In W. N. Parker (Ed.), *Trends in the American Economy in the Nineteenth Century*, Volume 24 of *Studies in Income and Wealth*, pp. 547–570. Princeton, NJ: Princeton University Press.

Crouzet, N., J. C. Eberly, A. L. Eisfeldt, and D. Papanikolaou (2022). The economics of intangible capital. *Journal of Economic Perspectives 36*(3), 29–52.

DeMarzo, P. M. (2005). The pooling and tranching of securities: A model of informed intermediation. *Journal of Finance 60*(1), 1–35.

Eisfeldt, A. L., G. Schubert, and M. B. Zhang (2023). Generative ai and firm values. Working Paper 31222, National Bureau of Economic Research.

Farboodi, M. and L. Veldkamp (2026). A model of the data economy. *Review of Economic Studies*.

Federal Highway Administration (2014). Total disbursements for highways, by function, 1945–2014. Highway Statistics 2014, Table DISB-C. Accessed July 2026.

Fernald, J. G. (1999). Roads to prosperity? assessing the link between public capital and productivity. *American Economic Review 89*(3), 619–638.

Gallman, R. E. (1966). Gross national product in the united states, 1834–1909. In D. S. Brady (Ed.), *Output, Employment, and Productivity in the United States after 1800*, pp. 3–76. New York: National Bureau of Economic Research.

Gargano, A. and M. Giacoletti (2025). Subsidizing the cloud: U.s. state incentives to data centers. Working paper / SSRN 5881105.

Gete, P. and A. Mercader (2026). Datacenter mortgages and originate-to-distribute. Working paper.

Green Street (2026, January). Data center outlook. Sector outlook report.

Haskel, J. and S. Westlake (2017). *Capitalism without Capital: The Rise of the Intangible Economy*. Princeton University Press.

JLL (2026, January). 2026 global data center outlook. Technical report, Jones Lang LaSalle. Accessed July 2026.

Johnston, L. D. and S. H. Williamson (2026). What was the u.s. gdp then? MeasuringWorth. Accessed July 2026.

Jones, C. I. and C. Tonetti (2020). Nonrivalry and the economics of data. *American Economic Review 110*(9), 2819–2858.

Jovanovic, B. and P. L. Rousseau (2005). General purpose technologies. *Handbook of Economic Growth 1*, 1181–1224.

McKinsey & Company (2025, April). The cost of compute: A $7 trillion race to scale data centers. Published April 28, 2025.

Moody's Ratings (2025, December). Corporate abs – u.s.: 2026 outlook — ai and used asset demand to aid abs quality and collateral values. Ratings outlook.

Moody's Ratings (2026a, February). Accounting – u.s.: Hyperscalers' reported ai-related lease commitments may understate economic risk. Accounting commentary.

Moody's Ratings (2026b, January). 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.

NVIDIA (2026, May). Nvidia nvl72 ai factory: Enterprise reference architecture. Technical report, NVIDIA. Last updated May 18, 2026; accessed July 2026.

Rhode, P. W. (2002). Gallman's annual output series for the united states, 1834–1909. Working Paper 8860, National Bureau of Economic Research.

Rungcharoenkitkul, P. (2026, July). The ai investment race. BIS Working Paper 1367, Bank for International Settlements.

Soto, P. E., M. Thieu, and J. S. Allen (2026, July). The AI buildout and the economy: Publicly available data to assess AI's impact. FEDS Notes, Board of Governors of the Federal Reserve System. July 17, 2026.

Turner & Townsend (2025, November). Data centre construction cost index 2025–2026. Technical report, Turner & Townsend. Accessed July 2026.

Ulmer, M. J. (1960). *Capital in Transportation, Communications, and Public Utilities: Its Formation and Financing*. Princeton, NJ: Princeton University Press.

Van Nieuwerburgh, S. and Z. Sun (2026, November). Clouds of credit: Mapping the financial web behind ai compute. Working paper, Columbia Business School.

A. Representative AI Campus Cost

We estimate the initial capital cost of a representative AI training campus with 200 MW of total facility power. Assuming a power usage effectiveness (PUE) ratio of 1.2, the facility supports approximately 166.7 MW of critical IT capacity. This PUE ratio is consistent with Cleanview data on facilities that report both measures. The remaining approximately 33.3 MW powers cooling, electrical conversion and distribution, lighting, and other non-IT facility operations.

We assume that the data center facility costs $11 million per MW of total power capacity, implying a construction cost of $2.2 billion. This benchmark is close to the $11.3 million-per-MW global construction cost forecast by JLL (2026). It is also consistent with evidence from Turner & Townsend (2025) that liquid-cooled U.S. data centers cost 7–10 percent more than otherwise comparable air-cooled facilities, and with CBRE (2026), who emphasize continuing cost pressure from high-density cooling requirements, power-equipment constraints, and specialized labor.

We separately allow $0.4 billion for substations, transmission upgrades, and grid interconnection, equivalent to $2 million per MW of total facility capacity. Because these costs vary widely across sites and may be borne partly by utilities, we interpret this amount as an allowance for incremental campus-related power infrastructure rather than a comprehensive measure of all induced grid investment.

We estimate IT costs using NVIDIA's GB300 NVL72 rack-scale system. NVIDIA reports power requirements of approximately 142 kW per rack (NVIDIA, 2026), implying that the campus can accommodate approximately 1,170 high-density racks.[7] We assume a cost of $4.0 million per integrated GB300 NVL72 system, or $4.68 billion in total. This price already includes the GPUs, attached high-bandwidth memory, Grace CPUs, NVLink switching, power systems, and the physical rack. We add 10 percent of the rack cost for the external network fabric and fiber, 5 percent for shared storage, and 5 percent for cluster-level installation,

[7]We interpret this rack count as a cluster-equivalent capacity measure. That is, each GB300 NVL72 rack represents not only the rack-scale compute system itself, but also its pro rata share of the external network, storage, and integration layer required to operate the cluster. Under this convention, the 166.7 MW critical IT load is allocated across fully configured AI-compute capacity, rather than only the GPU rack nameplate draw. The network, storage, and integration allowances below are therefore cost markups applied to the same cluster-equivalent unit of capacity, not additional power loads layered on top of the 166.7 MW IT budget.

integration, and commissioning. The resulting installed IT cost is approximately \$4.8 million per rack, or \$5.62 billion in total, equivalent to about \$34 million per MW of critical IT capacity.

Table 2. Illustrative Cost of a 200 MW AI Training Campus

Component	Assumption	Cost ($ billions)	Share (%)
Data center facility	200 MW × $11m	2.200	26.8
Incremental power infrastructure	Campus-level allowance	0.400	4.9
IT equipment	1,170 racks × $4.8m	5.616	68.4
Integrated GB300 NVL72 systems	1,170 racks × $4.0m	4.680	57.0
External network fabric and fiber	10% of rack-system cost	0.468	5.7
Shared data storage	5% of rack-system cost	0.234	2.8
Integration and commissioning	5% of rack-system cost	0.234	2.8
Total cost 200 MW data center		8.216	100.0

Notes: The campus has 200 MW of total facility power, equivalent to 166.7 MW of critical IT capacity at a PUE of 1.2. The rack count is interpreted as a cluster-equivalent measure: the 142 kW per rack denominator allocates the critical IT load across fully configured compute capacity, including each rack's pro rata share of external networking and storage power. The network, storage, and integration rows are cost allowances rather than additional power loads. Indented rows sum to total IT equipment. Appendix A provides details.

The resulting initial capital cost is approximately \$8.2 billion. The data center facility and incremental power infrastructure account for about 32 percent of the total, while the installed IT system accounts for about 68 percent.

B. Data Centers Capacity Count

Data Center Projects. We measure U.S. data center capacity using project-level records from Cleanview, accessed on July 23, 2026. The database contains 2,972 projects and reports a project identifier, project name, developer, capacity (MW), completion year, county, state, land area, building area, investment, latitude, longitude, and links to the underlying sources. Cleanview classifies each project as *operational*, *planned*, or *canceled*. We exclude the 39 canceled projects, leaving 2,933 operational or planned projects. Completion year denotes the reported operating year for operational projects and the expected completion year for planned projects.

Imputing Missing Capacity. Capacity is reported for 2,501 of the 2,933 operational or planned projects and is missing for 432. We impute the missing values using a model estimated on all operational and planned projects with positive reported capacity. The dependent variable is the natural logarithm of capacity in MW. The covariates are $\log(1+\text{land acres}/100)$, $\log(1+\text{building square feet}/100{,}000)$, and log(1 + investment/\$1 billion), together with a separate missing-value indicator for each covariate. The model includes completion-year fixed effects, with 2024 as the omitted category. Operational and planned projects without a reported completion year receive separate missing-year effects.

We also include regularized county and developer effects. Developer names are normalized to combine common aliases, such as Amazon and AWS, Google and Alphabet, and Meta and Facebook. A developer receives a separate effect only if the estimation sample contains at least 20 of its projects with reported capacity; all remaining developers form the omitted *Other* category. This rule produces 17 named-developer groups. County and developer effects are shrunk toward zero using a ridge penalty selected by five-fold cross-validation; the selected penalty is $\lambda = 1$. The regression has 2,501 observations and an R^2 of 0.846 in logs.

Exponentiating a fitted log value estimates a conditional geometric mean rather than expected capacity in levels. We therefore apply Duan's smearing correction, which yields an estimated smearing factor of 1.1513. To limit extrapolation, imputed values are bounded between the first and ninety-ninth percentiles of reported capacity, 1 and 2,500 MW. This bound affects 10 predictions. We impute capacity for 61 operational projects and 371 planned projects.

Capacity Estimates. Table 3 reports project counts and capacity by completion year before applying any realization adjustments. The database identifies 971 operational projects completed by the end of 2024 or operating without a reported completion year. Together, these projects represent 39.0 GW, of which 1.6 GW is imputed. The 2025 category contains 170 projects and 12.7 GW. Seven of these projects, representing 0.6 GW, remain classified as planned despite a reported 2025 completion year; we treat them as operational in the realization scenario below. The 2026 category contains 80 operational projects representing 6.3 GW and 205 planned projects representing 27.5 GW. In total, 566 GW worth of data center projects are reported in the database, a huge change from the existing stock of 39 GW at the end of 2024.

Capacity Operational by 2032. We next estimate the additional U.S. data center capacity that becomes operational between 2025 and 2032. This quantity differs from the announced pipeline because planned projects may be reduced in scale, abandoned, delayed beyond 2032, or completed as announced. We therefore interpret the calculation as a scenario for realized capacity, not as a forecast that every project is built on its reported schedule. Projects without a reported completion date are treated as relatively early-stage developments: some are completed by 2032, some after 2032, and some never reach completion.

Capacity associated with operational projects is included without adjustment, as is all capacity with a reported 2025 completion year. For each remaining planned category, we assign separate shares to completion by 2032, completion after 2032,

Table 3. Operational and Planned U.S. Data Center Capacity by Completion Year

	Observed capacity		Imputed capacity		Total	
Completion year	Projects	GW	Projects	GW	Projects	GW
Before 2025	915	37.5	56	1.6	971	39.0
2025	167	12.5	3	0.1	170	12.7
2026	272	31.8	13	1.9	285	33.8
2027	330	38.8	25	4.9	355	43.7
2028	264	36.5	21	5.5	285	42.0
2029	91	13.2	6	1.3	97	14.6
2030	47	7.8	7	2.3	54	10.2
2031	20	1.6	5	2.2	25	3.8
After 2031	14	2.7	15	9.9	29	12.6
Planned, no date	381	243.6	281	110.0	662	353.6
Total	**2,501**	**426.1**	**432**	**139.9**	**2,933**	**566.0**

Notes: The table includes operational and planned projects and excludes canceled projects. "Observed" denotes capacity reported by Cleanview; "imputed" denotes capacity estimated using the model described in the text. "Before 2025" includes operational projects without a reported year, and "After 2031" includes projects dated 2032 or later. No realization adjustments are applied.

and non-realization. The share completed by 2032 declines with the reported completion horizon and is lower when capacity must be imputed, reflecting the greater uncertainty surrounding projects for which no capacity has been disclosed. The residual "Never" category captures cancellation and downsizing; the realization shares should therefore be interpreted as capacity-weighted outcomes rather than project-level cancellation probabilities.

Table 4. Planned-Project Realization Assumptions and U.S. Data Center Capacity Outcomes

	Observed capacity		Imputed capacity		Capacity outcome (GW)		
Completion	By 2032	After 2032	By 2032	After 2032	By 2032	After 2032	Never
2025	–	–	–	–	12.7	0.0	0.0
2026	90%	0%	80%	0%	30.8	0.0	2.9
2027	80%	5%	65%	10%	34.2	2.4	7.1
2028	70%	10%	55%	15%	28.6	4.5	8.9
2029	60%	15%	45%	20%	8.5	2.3	3.8
2030	50%	20%	35%	25%	4.7	2.1	3.3
2031	40%	25%	25%	30%	1.2	1.1	1.6
After 2031	30%	30%	15%	35%	2.3	4.3	6.0
Planned, no date	20%	30%	10%	25%	59.7	100.6	193.3
Total					**182.7**	**117.2**	**226.9**

Notes: Rates apply only to planned projects; operational capacity and all 2025 capacity are included in full. "Never" is the residual capacity not completed. Columns may not sum exactly because of rounding.

Table 4 divides the 526.9 GW gross post-2024 pipeline into three outcomes. Our central scenario has 182.8 GW becoming operational by the end of 2032, 117.2 GW completed after 2032, and 226.9 GW never completed. The 182.7 GW buildout consists of 19.0 GW associated with projects that are operational in 2025 or 2026, 141.7 GW of planned capacity for which MW are observed, and 22.1 GW of planned capacity for which MW are imputed. Projects completed by the end of 2024 are part of the pre-2025 installed base and are excluded from the buildout calculations.

Planned projects without a reported completion date receive separate treatment because the database provides no direct basis for assigning them to a particular year. For reported capacity, we assume that 20 percent is completed by 2032, 30 percent is completed after 2032, and 50 percent is never realized. For imputed capacity, the corresponding shares are 10 percent, 25 percent, and 65 percent. Consequently, of the 353.6 GW associated with 662 planned projects without a completion date, 59.7 GW is completed by 2032, 100.6 GW is completed after 2032, and 193.3 GW is never completed. Capacity is imputed for 281 of these projects. Because this category accounts for most of the gross pipeline, its ultimate realization remains the largest source of uncertainty in the central scenario.

Finally, Cleanview's capacity field does not consistently distinguish total facility power from critical IT load. We therefore refer to these quantities throughout the paper as reported total project power capacity.

Annual Data Center Investment, 2025–32. We translate the capacity scenario into annual investment in data centers, power infrastructure, and IT equipment. We assume a total cost of \$41 million per MW in 2025, increasing by 4 percent annually. Because project expenditures precede completion, we allocate 10 percent of the cost of capacity completed in year t to year $t-3$, 25 percent to $t-2$, 40 percent to $t-1$, and 25 percent to t. Annual investment therefore includes spending on projects completed in that year and in the following three years.

Let K_t denote GW of capacity completed in year t, and let $c_t = \$41$ million $\times$ 1.04^{t-2025} denote the cost per MW in the year in which expenditure occurs. Annual investment is

$$I_t = 1{,}000 \times c_t \times (0.25K_t + 0.40K_{t+1} + 0.25K_{t+2} + 0.10K_{t+3})\,. \qquad (1)$$

We distribute the 59.7 GW of no-date capacity completed by 2032 evenly over 2028–32. Capacity completed after 2032 is assigned to 2033–35, with 50 percent completed in 2033, 30 percent in 2034, and 20 percent in 2035. These projects contribute to investment through 2032 only to the extent that their pre-completion expenditures fall within the 2025–2032 measurement window. Capacity that is never completed generates no investment.

Table 5 reports annual investment and its ratio to nominal GDP. Nominal GDP is \$30.762 trillion in 2025 and is assumed to grow by 4 percent annually, the same rate as cost per MW. The scenario implies \$10.3 trillion of investment during 2025–32, equivalent to 3.63 percent of cumulative nominal GDP. The arithmetic average of the eight annual investment-to-GDP ratios is also 3.63 percent.

Table 5. Scenario-Implied U.S. Data Center Investment, 2025–32

Year	Investment (\$ billions)	Nominal GDP (\$ billions)	Investment/GDP (percent)
2025	1,153	30,762	3.7
2026	1,432	31,992	4.5
2027	1,399	33,272	4.2
2028	1,098	34,603	3.2
2029	791	35,987	2.2
2030	940	37,427	2.5
2031	1,408	38,924	3.6
2032	2,058	40,481	5.1
2025–32	**10,279**	**283,448**	**3.63**

Notes: Investment uses the capacity outcomes in Table 4, a 2025 cost of \$41 million per MW, 4 percent annual cost and GDP growth, and the expenditure profile in equation (1). The final ratio is cumulative investment divided by cumulative GDP. The investment numbers include the expenditures incurred in or before 2032 on data centers completed after 2032.

C. Historical Investment Booms

This appendix documents the construction of the historical comparisons in Table 1. For each episode, we measure average annual capital expenditure as a share of contemporaneous GDP over the principal buildout period. Because the underlying historical sources differ in scope and accounting conventions, the estimates should be interpreted as approximate measures of the scale of each investment boom rather than as perfectly harmonized national-accounting series.

Canal investment is based on the annual construction-expenditure estimates assembled by Cranmer (1960). During the principal canal boom of 1836–1841, annual investment rose from \$4.4 million to a peak of \$14.3 million in 1840 before declining to \$11.7 million in 1841. Dividing these expenditures by contemporaneous nominal GDP from Johnston and Williamson (2026), canal investment averaged approximately 0.66 percent of GDP over 1836–1841 and peaked at 0.90 percent in 1840.

Railroad investment is based on Gallman's revised annual series for railroad construction and his annual national-product estimates, both measured in constant 1860 dollars and reported by Rhode (2002). Gallman (1966) constructed the revised investment series using improved estimates of construction costs and annual information on railroad mileage built. Over 1870–1890, railroad construction investment averaged 2.24 percent of GDP; the annual share exceeded 4 percent in

1870 and 1881. This measure covers railroad construction and therefore excludes at least some investment in locomotives and rolling stock.

Electrification investment is based on the annual gross capital expenditure series for electric light and power assembled by Ulmer (1960). The measure includes private and municipally owned generation, transmission, and distribution plant and equipment, but excludes land and complementary investment in electrical machinery by electricity users. For years before 1923, we allocate Ulmer's municipal-investment totals across years in proportion to private investment, consistent with his construction of the annual series. Dividing current-dollar investment by annual nominal GDP from Johnston and Williamson (2026), we find that electric-utility investment averaged 0.50 percent of GDP during 1905–1925 and peaked at 0.96 percent in 1924.

Highway investment is based on the Federal Highway Administration's annual series for capital outlays by federal, state, and local governments (Federal Highway Administration, 2014). Capital outlays include land acquisition, engineering, construction and reconstruction, rehabilitation, and related highway improvements. Dividing these expenditures by annual nominal GDP, we find that highway capital investment averaged 1.13 percent of GDP during 1956–1973 and peaked at 1.32 percent in 1958.

Telecommunications investment is measured as private fixed investment in communications equipment and communications structures in the BEA national accounts. This captures both fiber and other network structures and the switches, routers, and related communications equipment purchased across the economy. Dividing these expenditures by nominal GDP, we find that investment averaged 1.10 percent of GDP during the 1996–2003 telecom boom and peaked at 1.39 percent in 2000. These data are consistent with Couper et al. (2003), who document the sharp rise and subsequent collapse in communications investment surrounding the expansion of long-haul fiber networks.

The annual investment to GDP ratio for the AI buildout was discussed in Appendix B and averages 3.63% over the period 2025–2032.

Historical comparisons based on gross investment require caution because the assets created have different economic lives. Canals, railroads, and fiber optic cables often provided services for decades, whereas GPUs may become economically obsolete within three to six years. The AI buildout therefore generates less net capital formation per dollar of gross investment than many earlier infrastructure booms, although buildings, power connections, and other components of data centers remain long-lived.

At the same time, our estimates capture only the initial construction and equipping of new capacity. They do not include subsequent hardware refreshes. Capacity commissioned in 2025–26 may require substantial replacement of GPUs and related IT equipment by around 2030, even though the underlying buildings and power infrastructure remain usable. Such refresh cycles would increase cumulative gross investment. Our estimates incorporate neither component-specific depreciation nor future replacement investment.

D. Revenue Required to Support the Investment

This appendix considers the revenues required for the data center investment described in Appendix B to earn a given unlevered return. The exercise is not a forecast of AI revenues. Rather, it calculates the mature annual revenue required to recover the investment, replace depreciating capital, and compensate investors for the opportunity cost of capital.

We use the capacity outcomes reported in Table 4. We exclude capacity that became operational before 2025. We assign the 2.3 GW reported in the "after 2031" category to 2032. In addition, Table 4 assumes that 59.7 GW of projects without a reported completion date are completed by 2032. We allocate this capacity equally across 2029–2032.

For each completion cohort, Q_t measures the total cost of the capacity completed in year t:

$$Q_t = 1{,}000 \times c_t \times K_t. \tag{2}$$

Because expenditures occur before capacity begins generating revenue, investors must also be compensated for capital committed during construction. Under the expenditure profile in equation (1), one dollar of project cost has a value at commissioning, using a annual cost of capital of r, of

$$\phi = 0.10(1 + r)^3 + 0.25(1 + r)^2 + 0.40(1 + r) + 0.25 \tag{3}$$

The capital base at the date cohort t becomes operational is therefore

$$B_t = \phi Q_t. \tag{4}$$

The assets embodied in a data center have substantially different economic lives. We assumed in Table 2 that IT equipment accounts for 68 percent of total cost. Based on current industry assessment and accounting practices, we assume IT has a six-year economic life. Buildings, power infrastructure, cooling equipment, and other non-IT assets account for the remaining 32 percent and have a 20-year economic life. For an asset with economic life N and required unlevered return r, define the annual capital-recovery factor as

$$a(r, N) = \frac{r(1 + r)^N}{(1 + r)^N - 1}. \tag{5}$$

This is the constant annual cash flow, as a fraction of initial capital, whose present value over N years equals one dollar.

The mature annual operating cash flow required from cohort t is

$$CF_t(r) = B_t\,[0.68a(r,6) + 0.32a(r,20)]\,. \tag{6}$$

At a $r = 10\%$ required unlevered return, the six-year and 20-year capital-recovery factors are 22.96% and 11.75%, respectively. The weighted annual capital charge is therefore 19.37% of the capital base.

Let μ denote the operating cash-flow margin, defined as operating cash flow divided by revenue. Mature annual revenue required from cohort t is

$$Rev_t(r,\mu) = \frac{CF_t(r)}{\mu}. \tag{7}$$

At a 10 percent required return and a $\mu = 50\%$ margin, mature annual revenue equals 38.74% of the capital base.

Table 6 reports the calculation by completion cohort. The 182.7 GW of capacity completed between 2025 and 2032 has an initial cost of approximately \$8.54 trillion. Including the return on capital committed during construction raises the capital base at commissioning to approximately \$9.61 trillion. At a 10 percent required unlevered return, this capacity must generate \$1.86 trillion in annual operating cash flow once all cohorts are operational and mature. At a 50 percent operating cash-flow margin, this corresponds to mature *annual* revenue of approximately \$3.725 trillion by 2032. This represents 9.2% of 2032 GDP of \$40.481 trillion.

Table 6. Capital and Required Revenue by Completion Cohort

Year	GW	Cost	Capital	Req. Profit ($ billions)	Req. Rev.
2025	12.7	520.7	586.3	113.6	227.2
2026	30.8	1,313.3	1,478.8	286.5	572.9
2027	34.2	1,516.6	1,707.7	330.8	661.6
2028	28.6	1,319.0	1,485.2	287.7	575.4
2029	23.4	1,123.6	1,265.1	245.1	490.2
2030	19.6	978.9	1,102.3	213.5	427.1
2031	16.1	836.5	941.9	182.5	364.9
2032	17.2	929.3	1,046.4	202.7	405.4
Total	182.700	8,538.0	9,613.8	1,862.4	3,724.8

Notes: Capacity is based on the outcomes in Table 4. The 2.3 GW in the "after 2031" category is assigned to 2032, and the 59.7 GW of undated capacity assumed to be completed by 2032 is allocated equally across 2029–2032. Cost is $Q_t = 1{,}000c_tK_t$. Capital is $B_t = 1.126Q_t$. Required operating cash flow assumes a 10 percent unlevered return, a six-year life for the 68 percent IT share, and a 20-year life for the remaining assets. Required revenue assumes a 50 percent operating cash-flow margin. Figures may not sum exactly because of rounding.

Table 7 shows how the aggregate mature revenue requirement varies with the required unlevered return and the operating cash-flow margin. At the central assumptions of a 10 percent return and a 50 percent margin, required annual revenue is \$3.72 trillion. The estimate ranges from \$2.29 trillion under a 5 percent required return and a 67 percent margin to \$6.72 trillion under a 15 percent required return and a 33 percent margin.

Table 7. Required Mature Annual Revenue

	Operating cash-flow margin		
Required return	33%	50%	67%
5%	4,651	3,070	2,291
10%	5,644	3,725	2,780
15%	6,724	4,438	3,312

Notes: Entries are billions of dollars of mature annual revenue required from all 2025–32 completion cohorts. Calculations use the \$9.614 trillion capital base in Table 6, including the fixed 1.126 construction-period adjustment. IT equipment is 68 percent of cost and has a six-year economic life; other assets have a 20-year economic life. Each entry equals the weighted annual capital charge divided by the indicated operating cash-flow margin.

Expressed in market units, the central revenue requirement is also within the range of current GPU-rental economics. The \$3.725 trillion mature annual revenue requirement in Table 6, spread over 182.7 GW of total facility capacity, corresponds to approximately \$20.4 million per MW-year of total facility power, or \$1.70 million per MW-month. At a PUE of 1.2, the same revenue corresponds to approximately \$24.5 million per MW-year of critical IT load, or \$2.79 per kWh of GPU power consumed. Using the representative GB300 NVL72 configuration in Appendix A, the 182.7 GW buildout supports roughly 1.07 million racks, or 77 million GPUs, implying required revenue of about \$5.5 per installed GPU-hour at full utilization. This figure is broadly consistent with current market prices for high-end NVIDIA GPU capacity, which commonly range from roughly \$6 to more than \$10 per GPU-hour for on-demand or short-term rental, with lower effective prices available under long-term reservations or spot capacity. At utilization rates below 100 percent, the required billed price rises mechanically: for example, 80 percent utilization implies a required price of about \$6.9 per billed GPU-hour, while 70 percent utilization implies about \$7.9 per billed GPU-hour. The revenue requirement therefore does not appear implausible relative to current pricing for frontier compute. The central economic question is instead whether demand is deep enough to absorb roughly 180 GW of new capacity at high utilization, and whether providers retain sufficient pricing power for GPU-hour prices to remain near current levels as capacity expands and competition among hyperscalers, cloud providers, and AI model developers intensifies.

For context, given current estimates of annual combined revenues of OpenAI and Anthropic of around \$100 billion, revenues would need to grow at roughly 80% per year until 2032 in order to reach the required \$3.725 trillion revenue.

If the economic life of IT equipment is shorter, the revenue requirement rises substantially. For example, if the IT share of capital has a three-year rather than six-year economic life, while all other assumptions remain unchanged, the weighted annual capital charge increases from 19.37 percent to 31.10 percent of the capital base. At a 10 percent required return and a 50 percent operating cash-flow margin, mature annual required revenue rises from $3.7 trillion to about $6.0 trillion, or from 9.2 percent to 14.8 percent of 2032 GDP. In GPU-hour terms, the required price rises from about $5.5 to $8.8 per installed GPU-hour at full utilization.