FIDE PARTNERS

AI bubble from a digital infrastructure perspective

Whitepaper

February 2026

Authors:
Tom Allegaert, Alejandro Cardenas, Nicolas Betancur, Natalia Serrano, Daniel Mansi, Gabriela Villegas

FIDE PARTNERS

AI bubble from a digital infrastructure perspective

Tel: +44 2038 187213
info@fidepartners.com
www.fidepartners.com

February 2026

Whitepaper

Inside

Page 04-05
Circular financing and artificial demand

Page 06-10
Valuation

Page 11-15
Overinvestment

Page 16-24
ROI and Capex commitments

Page 23-25
Monetisation models

AI bubble in the context of digital infrastructure dynamics

We are witnessing an unprecedented capital deployment cycle, where hundreds of billions of dollars are being invested in AI compute, energy, and networking capacity. The core tension lies between massive infrastructure buildout, driven by enterprise commitments to spend 10-40% of budgets on AI, and the current revenue reality of roughly USD 20 billion across AI-native companies worldwide.

However, this apparent mismatch may reflect measurement limitations rather than bubble dynamics. The critical infrastructure challenge ahead involves energy constraints potentially outpacing data centre construction, with blackout risks posing systemic threats. The key question is not whether demand exists, but whether the infrastructure stack can scale sustainably to meet the requirements of embodied AI and AI-driven solutions that are not yet priced into current projections.

This document aims to connect the main drivers of the alleged AI-bubble with the dynamics of the digital infrastructure industry and present our view on whether, from this perspective, it may represent a bubble in the making.

What has been happening so far?

November 2022 – Q1 2023

AI bubble fears began shortly after the release of ChatGPT in November 2022, as AI-linked stocks started to rise based solely on headlines with no commercial data to support expectations.

H2 2023

Bubble fears appear again due to a funding boom where venture capitalists and public markets began to pour billions into gen-AI start ups + an initial supply shortage of GPUs

Throughout 2024

Fears began to deepen as AI start-ups began to struggle to monetise AI, with most pilots delivering no or low ROI.

The costs of GPUs and digital infrastructure to keep pace with demand and new models were increasing at a rate higher than expected.

Nvidia becomes the most valuable company in the world, raising concerns about overvaluation

Initial policy tailwinds (EU AI Act, U.S. CHIPS Act)

Q1 2025

DeepSeek's new model crashes the market due to the quality of the model, using far fewer resources. This contradicted the notion of exponential funding, compute and power for AI deployments

August – November 2025

Periods of peak bubble fears in August and October, followed by early correction signals such as Palantir and Meta.

Increased debt issuance and multiple calls for shorting positions are adding pressure to an already stirred market.

Conflicting views on ROI and the adoption of AI services, with some mentioning that 95% of projects yield zero ROI, while others report 75% positive outcomes, along with some initial concerns about government intervention from regulators across multiple jurisdictions, have increased the noise around the topic.

The debate over whether Artificial Intelligence (AI) constitutes a bubble is an ongoing discussion that emerged prominently in 2023 following the launch of ChatGPT. The explosion of Artificial Intelligence (AI), particularly generative AI (gen-AI), has triggered a wave of market capitalisation and enthusiasm that draws direct parallels with the tech bubble of the late 1990s. The central question facing financial analysis is whether this rapid escalation represents a fundamental and sustainable technological transformation, or merely a wave of speculative euphoria driven by hype.

Although AI's valuation has evolved to unprecedented highs, it remains below the price multiples observed at the onset of previous bubbles in terms of the price multiple relative to the bubble start. In fact, as of the cut-off date for this document[1], no single KPI or conclusive indicator confirms that AI represents a financial bubble. Meanwhile, a surge in Google searches for 'AI Bubble' since August, along with the thousands of articles and comments on the topic, signals an unprecedented level of widespread public concern.

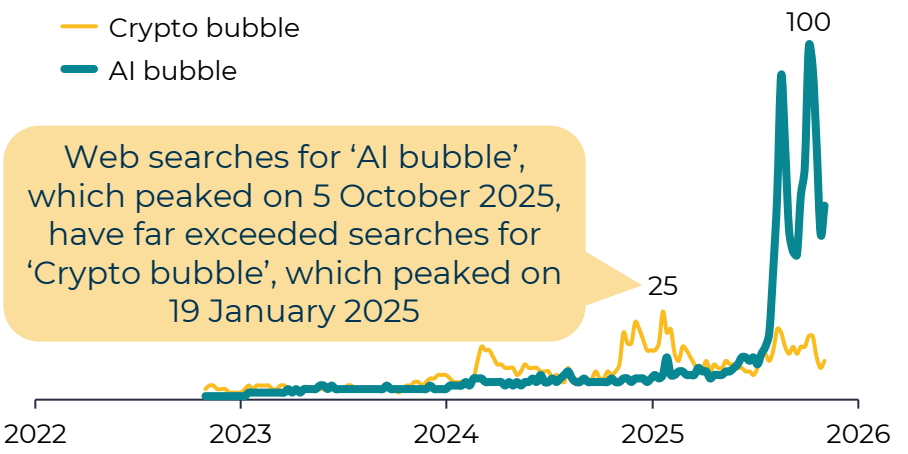
Beyond this debate, it is also important to recognise that technology has become an essential component of modern economies. Its exclusion from regional growth strategies would mean ignoring not only its positive contributions but also the expectations of a prosperous future associated with it.

A closer examination of the underlying components of the entire value chain, particularly digital infrastructure, reveals key insights that either support or contradict concerns about the AI bubble. This report will examine main points around the AI bubble debate (economic strain, investment dynamics, valuations, ROI, hardware requirements, circular investment flows, and monetisation models) and link each of them with the economics, drivers, and characteristics of the digital infrastructure space. By weighing both supporting and opposing arguments, we expect to provide a clear view on **whether AI is a bubble from a digital infrastructure angle.**

In light of these dynamics, Fide Partners supports investors and operators in identifying real value across the digital infrastructure landscape and navigating the strategic implications of AI with clarity and confidence.

Note 1: *The cut-off date is November 2025. Given the pace of developments and AI-related announcements, some aspects discussed may be outdated by the time this document is read.*

Exhibit 1.1: Web searches for 'Crypto bubble' and 'AI bubble' relative to the 'AI bubble' peak



Source: *Google Trends, 2025*

The investment paradox of AI: Circular flows, inflated valuations, and real fundamentals

AI circularity reveals a risk of interdependence: if one fails, could all fail?

Why it might be a bubble

The recent boom of circular agreements between OpenAI and the largest chip providers (NVIDIA and AMD), as well as with the infrastructure principal players (Oracle, Microsoft, CoreWeave), has turned on the alarms on possible speculative excess in the AI market. As an example, NVIDIA has announced that it will invest USD 100 billion in OpenAI, while the latter will purchase more chips from NVIDIA. Simultaneously, AMD publicly disclosed a strategic multi-year partnership with OpenAI that includes issuing warrants allowing OpenAI to acquire up to 10% of AMD's equity, alongside multi-billion-dollar purchases of AMD's AI accelerators.

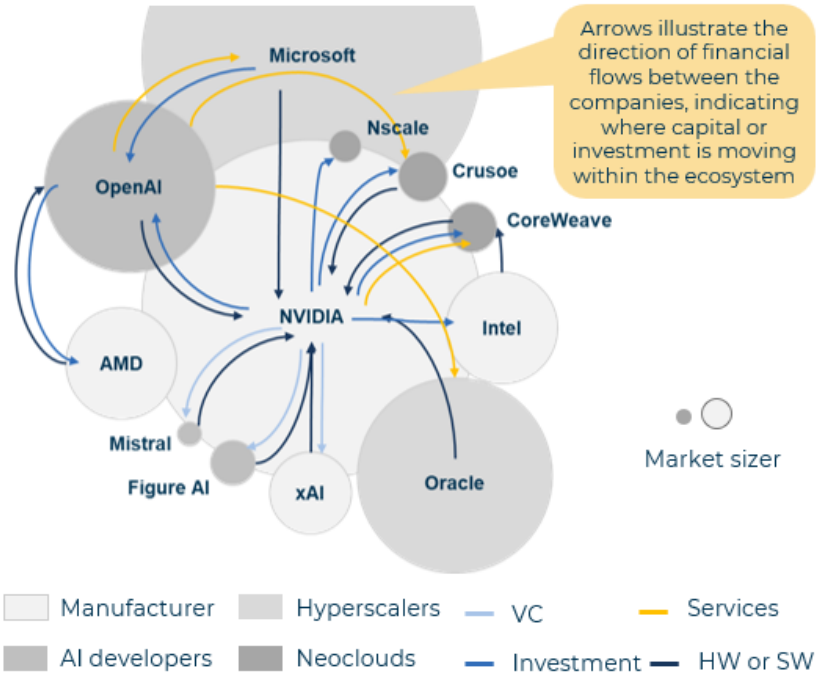
This pattern is also evident in the partnership between OpenAI, Thrive Capital and Thrive Holdings, where Thrive Capital, which invests in OpenAI, has launched Thrive Holdings and granted OpenAI equity for supplying models and technical expertise to its portfolio companies, forming a circular flow in which OpenAI powers the value creation that underpins Thrive's returns. These examples suggest that chip providers are implicitly 'subsidising' their own buyers and have led some analysts to identify similarities with the dot-com bubble.

Similar to the above, current capital investments are clustering around a limited number of AI model developers and infrastructure providers that rely heavily on hyperscale cloud operators. This has created a self-reinforcing ecosystem where hyperscalers fund start-ups, start-ups depend on their infrastructure, and valuations rise simultaneously on both sides. Such interdependence magnifies vulnerability. Should hyperscaler investment slow or compute access tighten, valuations across the ecosystem could adjust sharply.

Neocloud providers, such as CoreWeave, Crusoe, and Nscale, play a pivotal intermediary role within the AI investment loop, positioned between GPU manufacturers and hyperscalers. The next diagram illustrates how these operators both acquire GPUs from NVIDIA and receive indirect capital or offtake support from it, forming a reinforcing cycle of funding and hardware dependency. This structure accelerates GPU deployment but also concentrates financial exposure, as many neoclouds' margins and liquidity remain tied to NVIDIA's pricing and product cadence.

Moreover, the financial structure is worrisome: Microsoft reported a USD 3.1 billion hit to its net income in Q3 2025 from its equity-method investment in OpenAI. Given Microsoft's ~27% stake,

Exhibit 2.1: AI ecosystem financial flow



Source: *Bloomberg, Data Center Dynamics*

this implies an estimated loss for OpenAI of around USD 11.5 billion in that quarter, though the figure remains a derived estimate rather than an officially disclosed amount, according to disclosures within Microsoft's earnings filings. Additionally, OpenAI has projected a cumulative cash burn of up to USD 115 billion through 2029. These figures underline that the funding model relies heavily on new capital commitments and circular investment (rather than on robust FCF generated by the business itself), which is a hallmark of earlier speculative bubbles.

Beyond the record-breaking capital expenditure, the market's current valuations further illustrate the speculative tension surrounding AI. OpenAI, for instance, is now valued at nearly USD 500 billion, up from USD 157 billion a year ago, despite projecting multi-billion-dollar annual losses through 2028 and not expecting to become profitable until 2030. This reinforces the disconnect between valuation and near-term fundamentals, particularly as its CEO, Sam Altman, has stated that profitability is not a priority at this stage while investment requirements continue to scale toward the USD 1 trillion mark.

Meanwhile, the IMF and the Bank of England have both warned that risk assets, particularly those linked to AI, are "well above fundamentals" and show "stretched valuations." According to Morgan Stanley, just five AI-driven companies accounted for roughly 75% of the S&P 500's gains, raising concerns about a potential "Cisco moment"; a concentrated collapse if demand fails to meet projections. These dynamics echo the speculative excesses of past bubbles.

FCF: Free cash flows

The investment paradox of AI: Circular flows, inflated valuations, and real fundamentals

AI circularity reveals a risk of interdependence: if one fails, could all fail?

Why it might not be a bubble

The funding dynamics surrounding AI differ from the excesses of the late 1990s. The key point is the strong cash position of the leading players within this circular ecosystem; unlike most startups, companies such as Nvidia and AMD generate substantial cash flows from diversified revenue streams, limiting the risk of default. As economist Noah Smith[1] argues, this resembles legitimate commercial financing rather than fictitious revenue recognition. For example, Microsoft's free cash flow recently grew by around 33 % year-over-year as of Q3 2025, according to its latest earnings release, reported at the end of October 2025 (though distinct from OpenAI's numbers). Nevertheless, a certain degree of systemic risk persists, albeit to a reduced extent than in previous market cycles.

The initial market sentiment follows this belief. Goldman Sachs' global equity strategist, Peter Oppenheimer, stressed that "*this is not 1999: valuations are becoming extended, though still below levels seen in past speculative bubbles*". Unlike the dot-com era, today's tech giants are essentially funding their AI-related capital expenditure through internal cash generation.

Technology firms now enjoy more substantial margins, healthier balance sheets, and more sustainable cash flows than during the dot-com period, suggesting that today's market is underpinned by genuine financial fundamentals rather than speculative capital. For instance, Microsoft reported a year-over-year increase of approximately 33% in free cash flow in Q3 2025, reaching USD 29.8 billion according to its October 2025 earnings release, while maintaining an operating margin above 42%. These figures contrast sharply with the late-1990s tech sector, where median free-cash-flow margins among large-cap firms were close to zero, and fast-growing companies had negative cash generation.

Additionally, long-term demand projections remain robust. UBS forecasts that global income generated by AI could reach approximately USD 2.6 trillion by 2030 (a CAGR of roughly 41%), broadly consistent with McKinsey's estimate that generative AI could create between USD 2.6 and 4.4 trillion in annual economic value. Analysts from Barclays and UBS argue that most AI projects are built upon tangible use cases and sustainable growth paths, interpreting recent market corrections as consolidation phases and contending that current investment in AI represents productive capital directed toward infrastructure and innovation.

UBS: Union Bank of Switzerland | Note 1: Former Bloomberg columnist and PhD in economics from the University of Michigan

At the same time, OpenAI's own projections illustrate the scale of expected monetisation, as according to The Information and Reuters, the company anticipates a cumulative cash burn of around USD 115 billion through 2029, while targeting close to USD 200 billion in annual revenue by 2030. These figures, although highly optimistic, highlight the magnitude of revenue expectations underpinning current investment flows.

The digital-infrastructure layer remains notably resilient. Data centre operators rely on long-term contracts with hyperscale and enterprise clients, generating predictable recurring cash flows supported by barriers such as power capacity and connectivity. In diversified tech firms, data-centre and cloud infrastructure represent a smaller but structurally more stable revenue stream; for instance, Microsoft's Intelligent Cloud accounts for roughly 40% of total revenue. Meanwhile, according to CBRE, global data-centre vacancy fell to 6.6% in Q1 2025, down 2.1 points year-over-year, signalling robust demand beyond AI. These fundamentals provide a natural safety buffer even if AI-related investment slows.

Finally, even the sharp rise in share prices among leading AI firms appears to be grounded in operational performance. Demand for AI chips continues to surge, driving record earnings for suppliers like Nvidia. As Goldman Sachs' Oppenheimer notes, while valuations are high, they are supported by credible future earnings potential. Current price-to-earnings multiples remain below dot-com levels, and the sector's expansion is being financed through solid cash flows, mitigating the characteristics of a classic speculative bubble.

In summary, there is a compelling case for identifying warning signs of a possible bubble in AI, driven by the circular investment structure, significant losses at OpenAI, and the gap between future expectations and current free cash flows. On the other hand, there are significant factors that differentiate this cycle from past bubbles: more resilient business models, a broader demand base beyond AI, and stronger cash flows underpinning major infrastructure firms.

From our perspective, the greatest risk is that if AI adoption slows down, the infrastructure investments will still proceed, generating capacity oversupply and forcing valuation corrections (even if a sudden collapse is less likely). We therefore believe that market participants should monitor closely free cash flow metrics, infrastructure utilisation rates, and the quality of long-term contracts in the data centre and cloud segments.

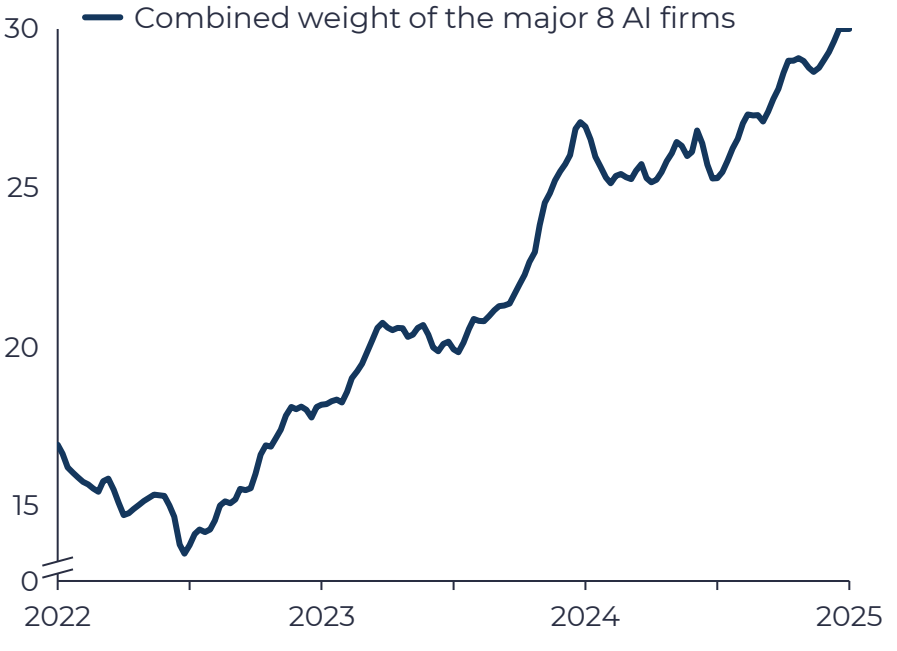
AI equities are driving disproportionate gains, echoing past speculative phases

Why overvaluation could indicate an AI bubble

Since late 2022, after the release of ChatGPT, AI has become the dominant theme in equity markets, directing an outsized share of capital into a small cluster of AI-linked stocks. Circular investment patterns have amplified the shift, pushing valuations higher by turning internally generated demand into what can appear to be genuine commercial traction.

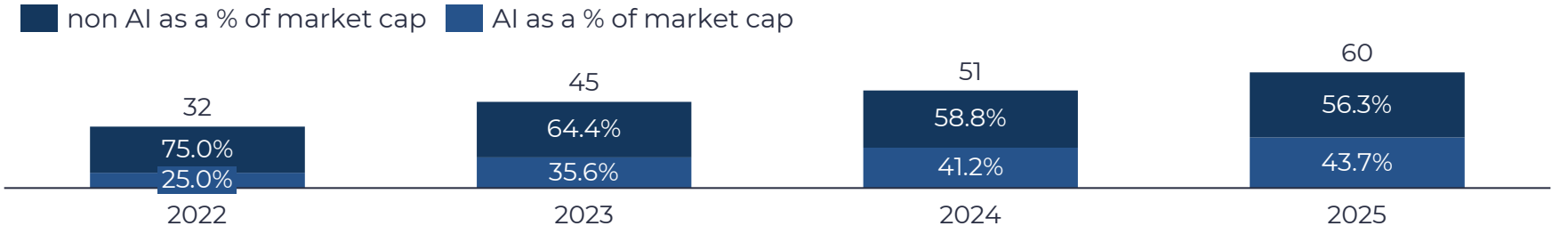
This concentration is evident in major equity indices. JP Morgan's AI Index, which comprises 38 AI-related names across technology, consumer discretionary, and real estate sectors, represent 44% of the S&P 500's total market capitalisation s of November 2025, up from 26% in 2022. Within this group, eight firms, Nvidia, Microsoft, Amazon, Alphabet, Meta, Broadcom, Oracle, and Tesla, account for almost one-third of the index, roughly double their share three years earlier. The scale of this concentration echoes the late 1990s, when the largest technology companies accounted for approximately 25-27% of the S&P 500 before the dot-com bubble burst.

Exhibit 3.1: Major 8 AI players as a percentage of the S&P500 (%)



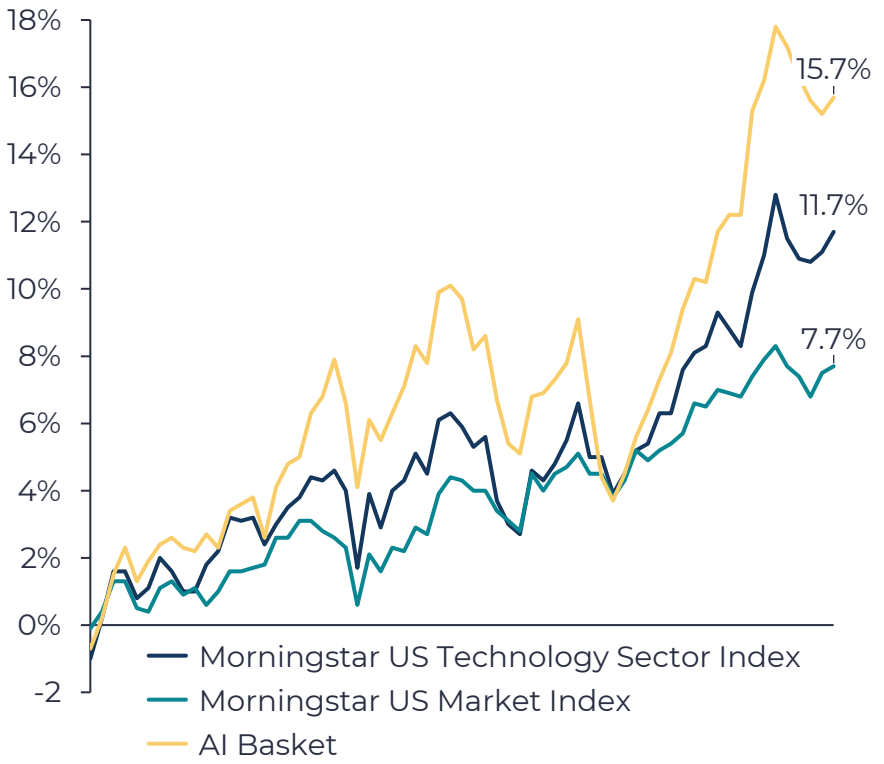
Source: *LSEG Datastream, Reuters*

Exhibit 3.2: AI players as a percentage of the S&P500 (USD trillion, %)



Source: *JPMorgan, S&P Global*

Exhibit 3.3: AI basket outperformance in Q3 of 2025 (%)



Source: *Morningstar*

Rising concentration in market capitalisation has been accompanied by a widening divergence in returns between AI-linked stocks and the rest of the market. Over the same period, a Morningstar index tracking 38 leading AI firms, including hyperscalers, chipmakers, and software providers, rose 27% in the second quarter of 2025 and a further 16% in the third. During these same two quarters, the broader US market advanced by only 7.7% and the wider technology index by 11.7%, underscoring that most of the market's recent gains have been driven by companies exposed to AI.

Within the basket, the most significant gains came from hardware and infrastructure providers such as Corning, Teradyne, and Arista Networks. These companies supply essential components for digital infrastructure, including optical fibre, semiconductor testing equipment, and network switches, all of which are critical to AI data centre build-outs. Their valuations have been fuelled by expectations of sustained AI data centre expansion. This pattern of concentrated outperformance shows that investor enthusiasm and capital inflows are clustering around a narrow group of firms, amplifying the imbalance between AI-related firms and the broader market.

AI valuations are expanding faster than earnings, signalling speculative momentum

Why overvaluation could indicate an AI bubble

Exhibit 3.4: AI-linked companies' P/E ratios

Company	P/E ratio	12-month forward P/E ratio
Nvidia	56	28
Microsoft	36	27
Alphabet	29	25
Meta	28	20
Broadcom	88	37
Telsa	295	185
Oracle	55	28
Palantir	444	173
Amazon	35	30
AMD	129	38

Source: *LSEG datastream, Finviz*

Another growing concern is that AI-linked equities are trading increasingly above their intrinsic value, suggesting that market enthusiasm may be outpacing fundamentals. Investors are assigning exceptionally high valuations to companies associated with artificial intelligence. Nvidia, for example, is now valued at more than USD 4.7 trillion, Microsoft is close to USD 3.8 trillion, and Alphabet is close to USD 3.5 trillion, reflecting expectations that future growth will justify current stock prices.

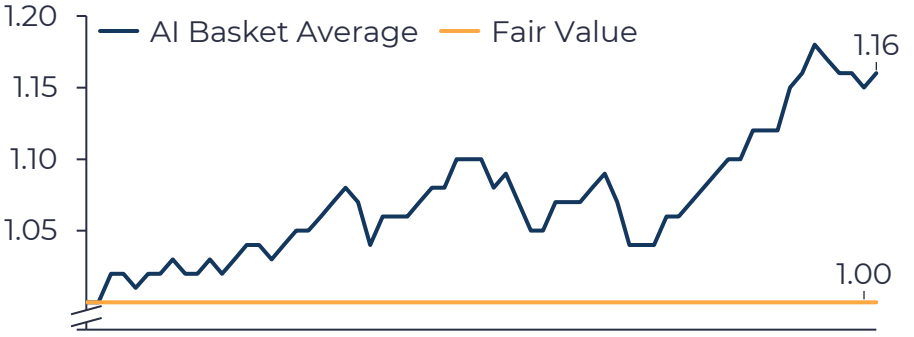
One measure of this is to look at relative valuation multiples across the major AI players. One particular measure that is useful for determining whether value has departed from intrinsic value is examining the price–earnings ratio, which measures how much a dollar of earnings is priced by the firm's valuation.

The median P/E ratio for AI-linked firms stands at 55X, surpassing the US technology industry's three-year average of 43X and significantly exceeding the S&P 500's long-term average of between 18X and 25X. This difference, resulting in an average valuation of more than 2x that of the rest of the S&P 500, rests on the belief that these firms will sustain extraordinary growth.

Forward valuations underscore investor optimism even more clearly. The median 12-month forward price-to-earnings ratio for leading AI firms, including Nvidia, Microsoft, and Amazon, is around 29x, compared with 19.7x for the S&P 500 excluding the Magnificent Seven. Investors are effectively paying almost twice as much for every expected dollar of future earnings.

This creates a fragile setup where even modest earnings downgrades or slower AI adoption could trigger sharp corrections. The extremes within this group highlight how stretched valuations have become. Palantir, whose profitability remains limited and reliant on government and enterprise contracts, trades at a price-to-earnings ratio of 444X and a forward ratio of 173X. Tesla trades at 295X earnings and 185X forward P/E, levels typically seen only during early-stage hypergrowth. Such valuations reflect strong confidence in long-term AI growth but also reveal how dependent the sector has become on sustained optimism and liquidity. When sentiment shifts, corrections can be swift, as seen when Palantir's stock fell 11% within days amid a broader sell-off driven by concerns over excessive AI spending across global markets.

Exhibit 3.5: 2025 Q3 AI basket to fair value estimate



Source: *Morningstar*

Data from Morningstar further supports this observation. The average AI stock now trades at roughly a 16% premium to its estimated fair value, up from near parity just one quarter earlier. This widening premium indicates that share prices are rising faster than revisions in earnings forecasts, signalling that markets are increasingly rewarding narrative over near-term cash flow generation. In other words, the belief in AI's transformative potential has already been monetised into present valuations.

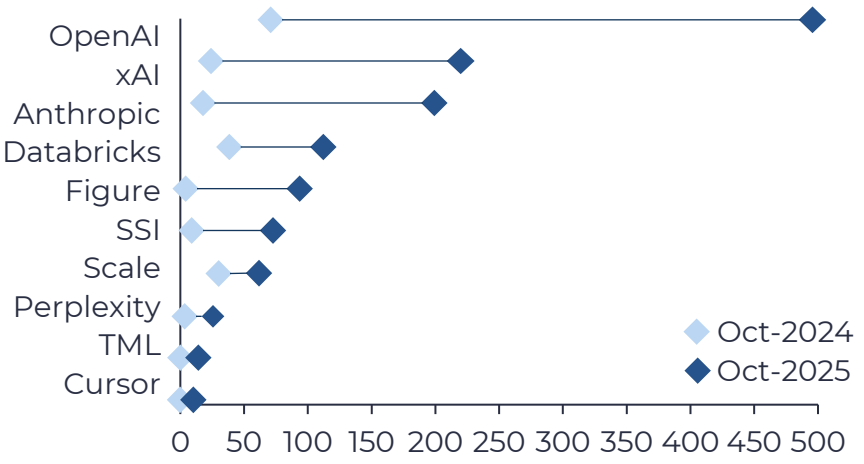
Overall, these valuation metrics show that the AI sector is priced for huge future earnings growth. With trailing and forward multiples far above historical averages and some firms trading at levels disconnected from their current earnings base, the market appears increasingly dependent on continued optimism and liquidity. Such a combination of concentration, elevated expectations, and limited earnings headroom has often preceded corrections in past market cycles, most notably during Japan's late-1980s asset bubble, when its equity market reached around 42% of global equity capitalisation before collapsing.

Private capital and valuations are expanding faster than realised fundamentals

Why overvaluation could indicate an AI bubble

A defining feature of this environment is the widening gap between private and public valuation frameworks. Publicly listed firms are priced on measurable indicators such as free cash flow, margins, and return on capital. In contrast, private AI companies are often valued primarily based on their potential for revenue growth and scale. Profits are treated as future outcomes rather than immediate requirements. This creates a structural dislocation between markets valuing realised earnings and those that value promise, a dynamic that has historically signalled fragility when sustained for long periods.

Exhibit 3.6: AI start-up valuations (USD billion)



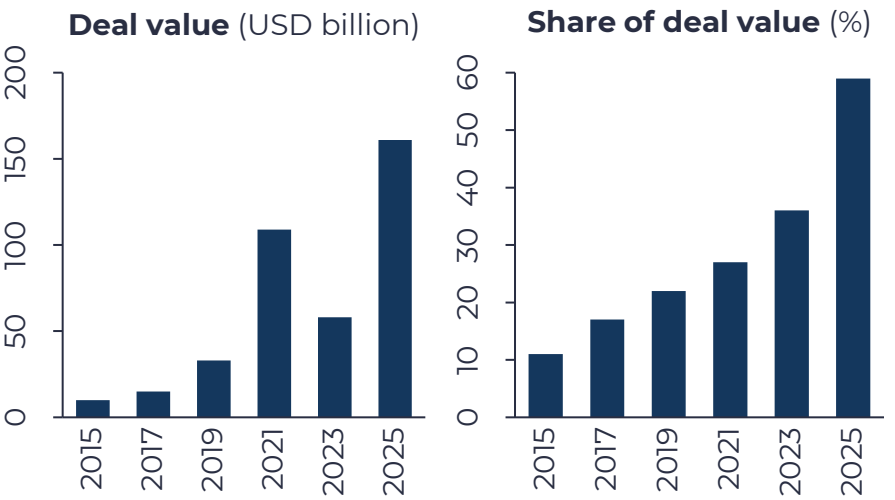
Source: *Financial Times, PitchBook*

The economic logic underpinning many current valuations appears fragile. While leading private firms have reported rapid revenue growth, many continue to post substantial losses; for instance, OpenAI is estimated to have lost around USD11 billion in 2025. Valuations are based on the expectation that future generations of models, expanding use cases, and eventual breakthroughs in artificial general intelligence will unlock vast new markets. Across the sector, infrastructure and research costs are rising faster than revenues, leaving most business models dependent on continued investor appetite to fund high expenditure in exchange for potential long-term dominance. This has driven the collective valuation of the ten most prominent AI start-ups to nearly USD 1 trillion, led by OpenAI, xAI, and Anthropic.

Many of these valuations have expanded without proven monetisation pathways. Some early-stage firms with only a few million dollars in annual revenue are now valued at over USD 500 million, implying multiples exceeding 100X. This escalation appears to be driven less by fundamentals than by massive capital inflows seeking to chase perceived future leaders.

The result is a momentum-driven dynamic reminiscent of past speculative phases, where narrative power and scarcity of opportunity inflated valuations.

Exhibit 3.7: AI venture capital inflows in the US



Source: *Financial Times, PitchBook*

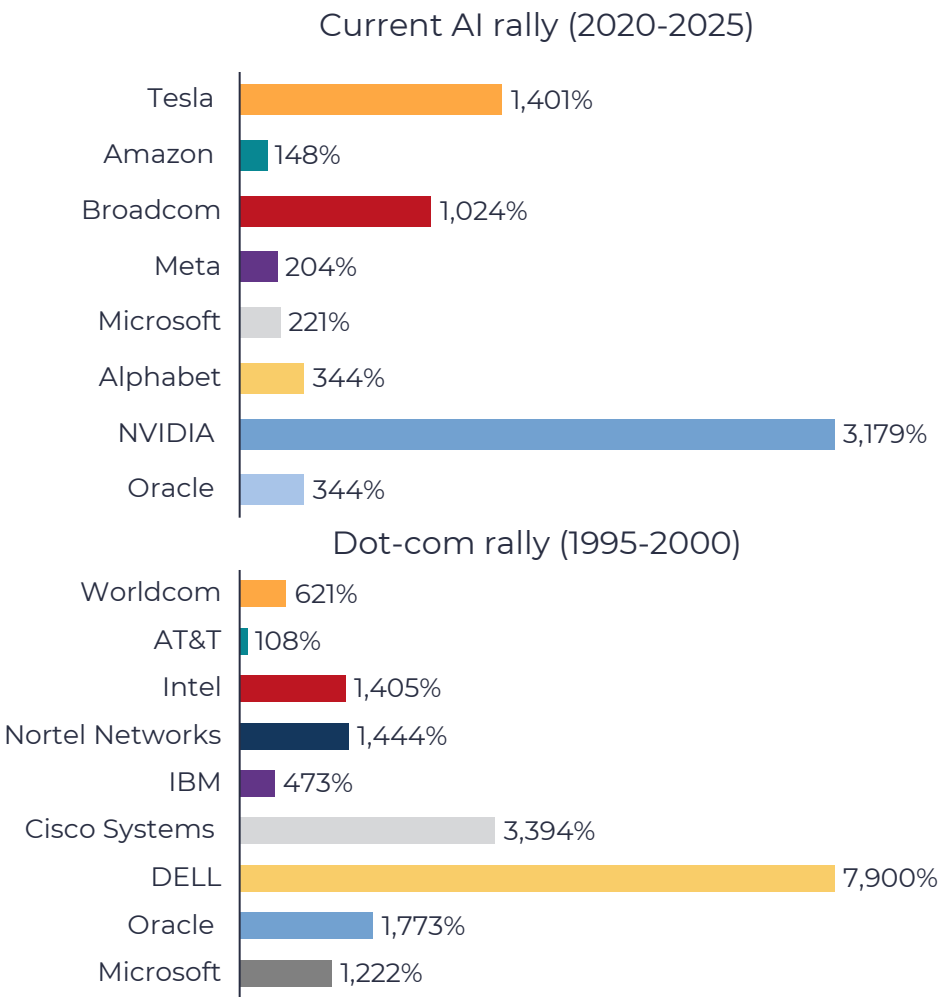
The scale of capital now flowing into private AI ventures raises concerns about broader systemic risks. Venture and private equity funding exceeded USD 235 billion in 2025, more than 10X the inflation-adjusted peak of the dot-com era. This surge has been driven not only by traditional venture capital but also by sovereign wealth funds, infrastructure investors, and crossover institutions seeking exposure to the next technological frontier. The rapid pace of capital deployment may also carry implications. In such a competitive environment, investors can be incentivised to accelerate processes, which in some instances may result in weaker due diligence and increased willingness to back pre-revenue companies on the assumption that early scale and first-mover advantage will secure future market share.

The magnitude of current AI funding invites comparison with past speculative phases. At the height of the dot-com boom, venture investment in internet start-ups reached around USD 10 billion, or USD 20 billion in today's terms. In 2021, software-as-a-service (SaaS) firms attracted USD 135 billion. By contrast, AI-focused investment now exceeds USD 200 billion annually, the largest and fastest inflow in venture history. This expansion reflects a broad conviction that artificial intelligence will create multi-trillion-dollar markets spanning automation, drug discovery, and industrial optimisation. Yet as in previous cycles, investment appears to be outpacing the sector's ability to convert capital into sustainable revenue. History suggests that when capital growth decouples from commercial reality, consolidation and correction often follow.

Current AI rally shows strength but lacks the speculative extremes of the dot-com era

Why current AI valuations do not yet signal a bubble

Exhibit 3.8: Current AI stock rally vs dot com rally (% gain)



Source: *LSEG datastream, CEPR*

While AI-linked equities have delivered strong returns, these remain moderate compared with the extremes of the late 1990s. During the dot-com boom, several individual stocks appreciated by more than 1000% and up to 8000% before the correction. By contrast, many of the leading AI companies have grown at a pace only slightly faster than that, suggesting that the market is expanding rapidly but still within historical bounds.

The composition of the rally also differs from earlier speculative phases. In the 1990s, gains were spread across hundreds of untested firms with limited revenue. Today, most of the increase in value is concentrated in large, profitable companies such as Nvidia, Microsoft, Alphabet, and Meta, which already generate strong cash flows and hold durable competitive positions. Between 2020 and mid-2025, Nvidia's share price rose by over 3000%, while Microsoft, Meta, and Amazon gained between 140-400%. Although valuations are high, they are underpinned by measurable profitability and ongoing investment in data centres, semiconductors, and software infrastructure, and are also far below the speculative rallies seen during the dot-com bubble.

Exhibit 3.9: Price-Implied growth rate of AI-linked companies (%)

Company	Previous 5 years	Analyst expectations	Implied growth rate
Nvidia	84	29	21
Microsoft	13	12	13
Alphabet	25	14	9
Meta	25	10	8
Broadcom	32	18	30
Telsa	49	-2	53
Oracle	2	14	21
Palantir	74	38	61
Amazon	38	15	13

Source: *LSEG datastream, CEPR*

AI stocks exhibit a wide dispersion in implied growth rates, which helps illustrate how investors distinguish between mature platforms and companies positioned for faster expansion. Firms such as Microsoft and Alphabet sit at the more conservative end, with implied growth rates of 14% and 9% that are broadly consistent with their historical performance and the steady trajectory of their core businesses. These expectations align with fundamentals and do not indicate overvaluation.

Palantir provides an example of how higher implied growth does not automatically suggest excessive pricing. Its implied growth rate of 74% is clearly ambitious when compared with analyst expectations of 38%. Still, it can be seen as a reflection of the company's significant exposure to high-growth markets for analytics and artificial intelligence software. In this context, the valuation appears less like a mispricing and more like an acknowledgement of its potential to scale rapidly within emerging artificial intelligence segments. However, this naturally comes with a higher level of risk given the growth that investors are assuming.

This variation in growth rates suggests that investors are differentiating based on business models, growth potential, and sector exposure, rather than applying uniform optimism across the sector. Companies like Microsoft and Alphabet are priced for stable growth, while firms like Palantir show higher expectations, grounded in their position in transformative industries.

Taken together, these patterns suggest that the market is confident but not irrational. The revaluation of AI equities reflects credible long-term growth expectations, not speculative exuberance.

Valuation resets in AI would ripple unevenly across digital infrastructure

Implications for digital infrastructure if **AI valuations** are bubble-like

Hyperscaler data centres

If AI equity valuations fall meaningfully, hyperscalers are likely to shift focus from aggressive expansion to capital efficiency and clearer monetisation. In a high-valuation environment, markets tend to reward scale and early positioning; however, once sentiment shifts, investors place greater emphasis on tangible returns. Capital expenditure aimed at speculative or long-dated AI growth may therefore face greater scrutiny, particularly where revenue models remain uncertain, encouraging hyperscalers to absorb adjustment internally through optimisation rather than continued outward expansion.

This does not imply that existing infrastructure will become irrelevant. Current data centre assets are likely to continue supporting essential workloads, such as productivity software, cloud computing, storage, and advertising, which remain central to hyperscaler economics. Artificial intelligence should continue to enhance these functions through improvements in automation, relevance and efficiency, even if expectations for frontier models become more restrained. Some reprioritisation is therefore possible, with new builds in emerging regions or facilities designed mainly for training workloads being paused or adjusted. Hyperscalers may place greater emphasis on improving existing facilities by increasing GPU utilisation, managing energy efficiency and extracting higher returns from their current footprint, reducing reliance on external capacity and, in turn, transmitting adjustment unevenly across the broader infrastructure ecosystem rather than through balance sheet stress at the hyperscaler level.

Neocloud providers

Neocloud providers like CoreWeave, Lambda Labs, Voltage Park and Crusoe Cloud have grown rapidly by supplying high-performance GPU infrastructure tailored to AI workloads. Their value lies in speed, specialisation, and flexibility, serving both AI startups and overflow from hyperscalers. However, this model relies heavily on two fragile pillars, external capital and concentrated tenant demand.

While many are revenue-generating and in some cases profitable, they do not generate enough free cash flow to fund internal infrastructure growth. Expansion has depended on frequent equity and debt raises to secure GPUs, prepay leases and lock in power, with much of this committed ahead of revenue and reliant on continued investor enthusiasm. If AI valuations fall, fundraising becomes slower, more expensive and more dilutive, while fixed obligations on GPUs, leases and power create immediate pressure. At the same time, demand could weaken, exposing the financial rigidity of commitments incurred ahead of stabilised utilisation. Hyperscaler partners such as Microsoft and Google, which have helped validate and scale neocloud providers, may shift from external expansion to internal optimisation, reducing resale activity, delaying precommits and shrinking workload volumes. A correction in venture funding would further limit customers' ability to pay for compute or renew contracts, exposing neocloud providers to falling utilisation across both anchor and peripheral demand.

The risk is particularly acute for neocloud providers moving into data centre ownership. Although this shift offers greater control and the prospect of improved long-term margins, it also increases direct asset exposure. If demand softens, providers could be left with underutilised infrastructure and rising overhead costs, without the financial resilience required to absorb the impact.

Diversified data centre providers

A correction in AI-related valuations is likely to produce targeted rather than systemic impacts for diversified data centre operators. While many have made significant investments to support GPU workloads, including liquid-cooled campuses and hyperscale builds, these facilities typically represent only part of a broader and more diversified footprint.

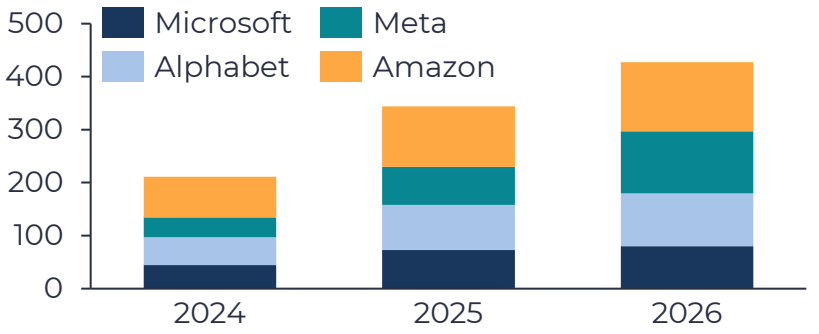
The AI-focused segment may still face pressure. Lease signings could slow, pricing may soften, and backfill conditions could become more difficult, particularly at sites leased to newer cloud providers with higher funding and demand risks. Some projects initiated ahead of confirmed tenancy or based on optimistic growth assumptions may require revisions to timelines or underwriting assumptions, resulting in lower expected IRRs even where long-term utilisation remains intact. These pressures are partly offset by more stable asset classes. Colocation campuses serving enterprise IT, interconnection-rich sites supporting cloud and network exchanges, and regional data centres enabling hybrid or legacy workloads remain underpinned by resilient enterprise demand and ongoing digital transformation. These segments tend to have longer contract terms, more diverse tenant bases, and lower sensitivity to cyclical shifts.

As a result, while AI demand has influenced recent development strategy, it does not determine the financial health of diversified providers. Cash flows from non-AI workloads continue to anchor platform performance, allowing operators to absorb delays or repricing in AI-facing assets without materially undermining overall stability.

The accelerating wave of AI capital: spending growth outpacing historical cycles

Why overinvestment could indicate an AI bubble

Exhibit 4.1: Capex spend from major tech firms (USD billion)

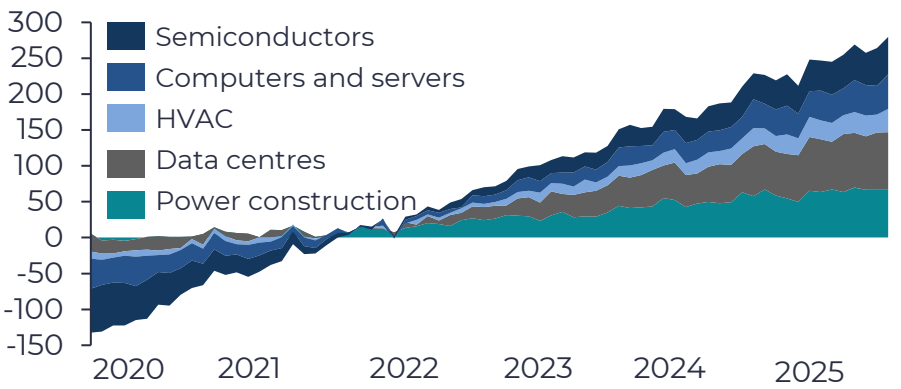


Source: *Company reports, Bloomberg*

When valuations are built on strong expectations, investment behaviour can begin to mirror that confidence, drawing ever more capital into the infrastructure behind AI growth. This has raised concerns of overinvestment, particularly as a small number of huge firms are driving spending. AI-related capex by Amazon, Meta, Alphabet and Microsoft is expected to reach ~USD344 billion in 2025, up from USD200 billion in 2024 and 2.75X the 2022 level.

The concentration of this spending amplifies the risk; of the roughly USD 430 billion in total global AI capex projected for 2025, the top four hyperscalers, as seen above, account for 80%, the largest eight for 86%, and the largest eleven for more than 90%. This level of concentration means that the trajectory of AI investment is now shaped by the decisions of a handful of firms, placing them at the centre of mounting fears that the current surge may carry the hallmarks of a speculative bubble.

Exhibit 4.2: Change in AI hardware spending in the US from 2022 (USD billion)



Source: Bureau of Economic Analysis, Goldman Sachs GIR

As shown in Exhibit 4.2, much of the investment in AI has been concentrated across semiconductors, servers, and data centre infrastructure. Since 2022, the most significant growth in US AI infrastructure has been driven by data centre construction (projected to increase by USD 80 billion), power infrastructure build-outs (up by USD 55 billion), and semiconductor manufacturing facilities (up by USD 45 billion).

Accounting for a USD 219 billion increase in annualised hardware spending, semiconductor and server investment has accelerated as hyperscalers expand training and inference capacity. At the same time, data centre construction and power infrastructure have grown rapidly to accommodate increasing compute density and energy demand. HVAC investment has also risen to address the cooling requirements associated with AI workloads.

AI investment in the US now accounts for roughly 1.5% of GDP, exceeding the 1% spent on telecom during the 5G build-out in 2019–2020 and approaching the 1.1–1.2% seen at the peak of the fibre and telecom expansions in the dot-com era. While still below historic infrastructure surges such as the US railway boom of the late 1800s or the electrification wave of the 1920s, it remains unprecedented in recent decades. High levels of capital formation indicate rapid technological progress; however, history shows that large build-outs can outpace actual demand, leading to financial strain once expectations are adjusted. In both the dot-com and electrification cycles, similar investment spikes led to overcapacity and eventual corrections. The current AI wave may reflect comparable dynamics, where optimism drives spending faster than proven economic returns.

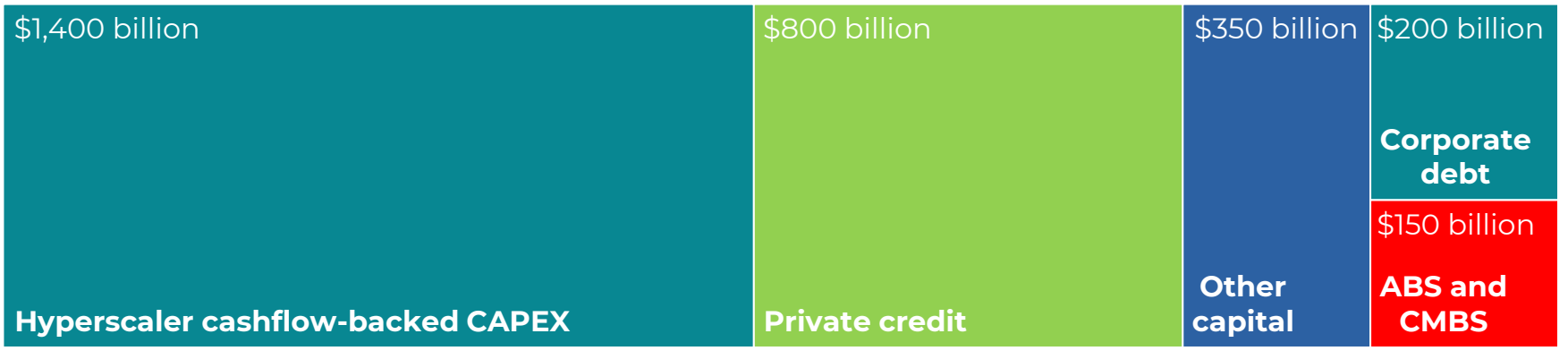
Forecasts for future capital needs vary, but year-to-date capex for the four largest hyperscalers has already exceeded earlier 2025 projections of USD 300 billion by around USD 44 billion. Nevertheless, the expected investment remains substantial. Some estimates suggest cumulative AI infrastructure spending could reach USD 3 trillion by 2028 and USD 5.2 trillion by 2030, spanning hyperscale data centres, semiconductor fabrication, power distribution, cooling, and networking. At this pace, annual investment could peak at the equivalent of around 4% of US GDP by 2030.

Most of the recent surge in AI spending has been funded primarily through hyperscaler cash flows; however, rising capital demands are shifting attention toward how this build-out will be financed. The pressure became clearer when Oracle issued USD 18 billion of bonds in September 2025 to support data centre expansion tied to its commitments to OpenAI. In the seven weeks that followed, Meta issued USD 30 billion in bonds and secured USD 27 billion in private debt for its Hyperion facility, while Alphabet raised USD 25 billion for AI infrastructure. In total, USD 120 billion of AI-related debt was raised over a short period. As investment needs continue to rise, internal financing will reach its limits, and the next phase of AI development will increasingly rely on external borrowing, introducing a new layer of financial risk.

Funding the AI boom: rising credit exposure and refinancing risks

Why overinvestment could indicate an AI bubble

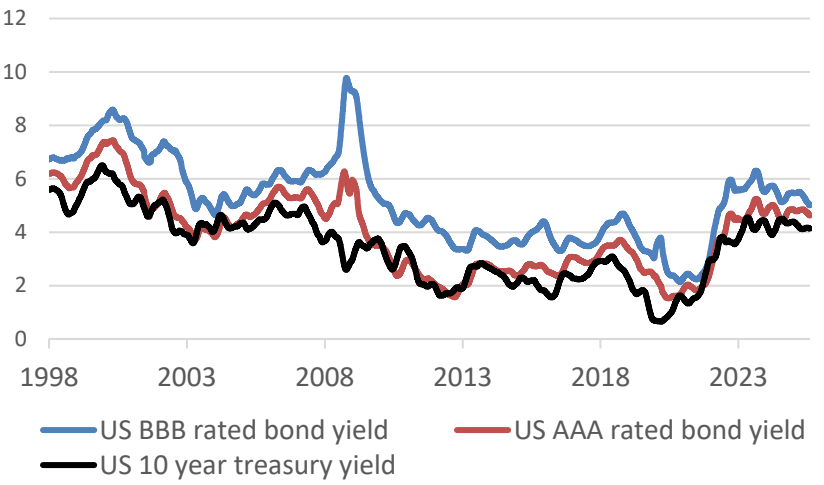
Exhibit 4.3: Expected financing of future AI investments



Source: *Morgan Stanley Research, Bank of England*

Bain and Co. estimates that USD 2 trillion in annual revenue is needed to fund the computing power required to meet anticipated AI demand by 2030. However, even with AI-related savings, the world is still USD 800 billion short of meeting demand, creating a need for external funding to bridge this shortfall. Morgan Stanley estimates that of the almost USD 3 trillion set to be spent on AI between 2025 and 2028, USD 1.4 trillion will be funded by hyperscaler cash flow, with the remainder distributed among external sources such as private credit, corporate debt, ABS, and other forms of capital, such as private equity.

Exhibit 4.4: Yield on the US 10-year treasury, corporate BBB, and corporate AAA bonds (%)



Source: *Federal Reserve Bank of St. Louis*

A growing reliance on leverage introduces a dual financing risk for the AI infrastructure build-out. As shown in **Exhibit 4.4**, borrowing costs have eased from recent peaks but remain materially above pandemic-era lows, with both AAA- and BBB-rated corporate bond yields still significantly higher than the levels at which much AI-related debt was raised between 2019 and 2022. As this legacy debt approaches refinancing, higher rates are likely to compress returns and reduce financial flexibility. At the same time, credit spreads over the 10-year US Treasury remain near historic lows, suggesting that new AI-related borrowing is still being priced with limited risk premium despite increasing market volatility.

While this supports continued capital deployment in the near term, it also heightens sensitivity to any repricing of risk, particularly for more leveraged developers and expansion-led platforms. Together, elevated base rates and tight credit spreads increase the exposure of AI infrastructure investment to shifts in funding conditions.

At the same time, the financing model is evolving. The growing use of special-purpose vehicles and private credit is moving leverage away from corporate balance sheets and into less transparent structures. Meta's Hyperion Partners partnership with Blue Owl is a clear example, with roughly USD 27 billion channelled through an SPV that carries much of the borrowing off balance sheet, even though Meta remains the developer and long-term tenant. These arrangements support rapid capital deployment but reduce visibility over system-wide indebtedness and introduce more complex refinancing chains.

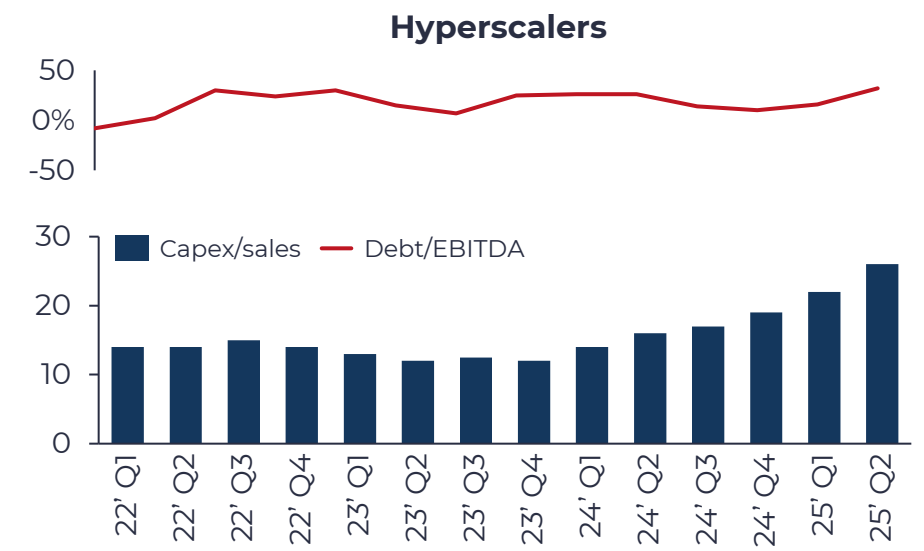
As a result, banks and lenders are increasingly seeking to reduce direct exposure to the AI investment cycle even as funding volumes continue to grow. Large debt raises by hyperscalers and data-centre developers are now accompanied by greater use of hedging, credit derivatives, and risk-transfer mechanisms, signalling caution around the scale and duration of AI-related investments. Rising default protection costs on issuers such as Oracle and Microsoft reflect this shift, even among highly rated borrowers.

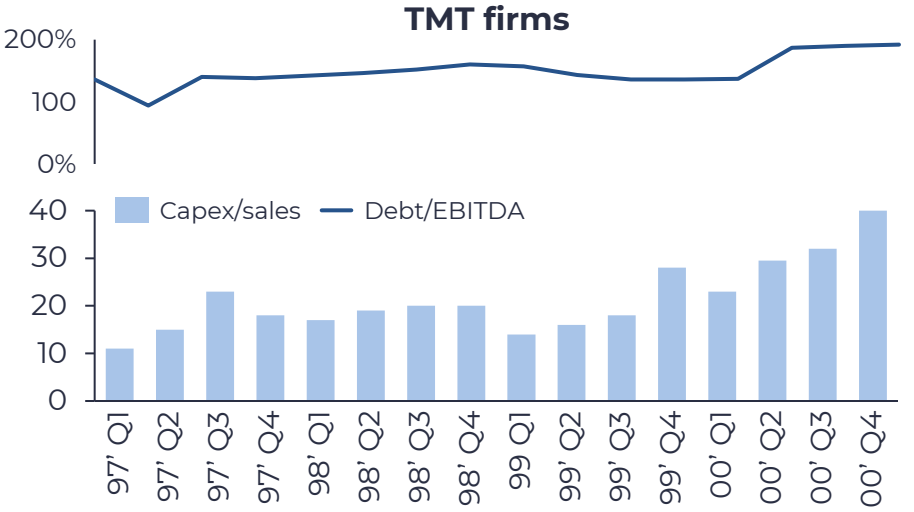
This evolution is beginning to attract regulatory attention. The Bank of England and the IMF have noted that aspects of private-credit lending increasingly resemble pre-crisis patterns, particularly the use of layered and opaque structures that go beyond traditional bank balance sheets. While not signalling immediate stress, this reinforces concerns that risk transmission in the AI financing ecosystem may become harder to monitor and more sensitive to changes in funding conditions as the cycle matures.

Earnings, not leverage, are powering the AI Buildout

Why AI investment does not yet signal a bubble

Exhibit 4.6: Relative leverage of hyperscalers vs TMT firms during the dot-com bubble (%)



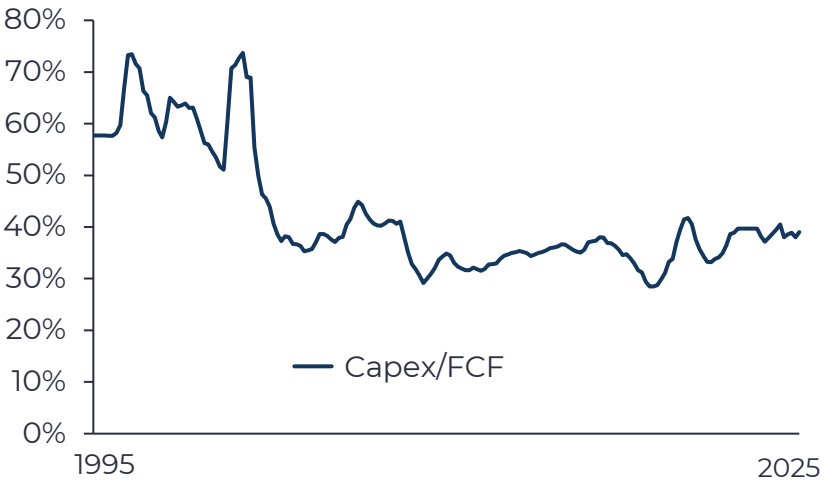


Source: *Barclays, Financial Times*

Fears about the scale of current AI investment are understandable when looking at absolute spending levels. Still, the firms driving this cycle differ materially from those that dominated earlier technology booms. Exhibit 4.6 shows that capex to sales for hyperscalers is now in the high-20s %, a noticeable increase but still meaningfully below the 30–40% levels reached by TMT firms during the dot-com bubble. The resemblance is therefore more visual than substantive. Today's hyperscalers operate from a far stronger financial base, supported by far greater revenue scale and healthier margins.

Despite rising capex intensity, hyperscaler leverage remains substantially lower than during the dot-com era. At the peak of the bubble, TMT companies carried debt-to-EBITDA ratios of roughly 190%, whereas current hyperscalers remain in the 30–40% range. Even with recent increases in debt issuance, balance sheets are supported by diversified revenue streams and robust cash flow generation, making current levels of indebtedness significantly more manageable.

Exhibit 4.7: US technology capex as a proportion to FCF (%)



Source: *LSEG Datastream, Goldman Sachs GIR*

Furthermore, Exhibit 4.7 reinforces the structural differences between today's major technology spenders and the firms that drove earlier investment cycles. At the height of the dot-com period, capex consumed close to 70% of free cash flow, leaving companies with limited capacity to absorb operational pressures or shifts in funding conditions. The equivalent figure today is closer to 38%, underscoring the extent to which current investment is supported by stronger and more predictable cash generation.

Firms with higher and more stable free cash flow are correspondingly better positioned to sustain elevated levels of investment without generating financial strain. They are less reliant on external financing, more resilient to changes in credit markets and better able to absorb fluctuations in demand.

Taken together, these factors materially weaken the case for a bubble-like risk profile. In earlier technology cycles, high capex intensity was paired with aggressive borrowing and limited internal cash generation, leaving firms exposed when revenue expectations faltered. The current environment is structurally different.

Hyperscalers are increasing investment, but they are not overextending with leverage; debt levels remain well within sustainable ranges and are supported by substantial free cash flow and diversified revenue bases. This ensures that rising capex is being financed from a position of financial strength rather than dependency on external credit. As a result, the risks associated with this investment cycle are far lower than those seen in past booms, and concerns about systemic overleveraging or a repeat of bubble-era vulnerabilities appear overstated.

Tangible assets are anchoring the AI expansion

Why AI investment does not yet signal a bubble

AI infrastructure investment and resilience

Much of today's investment in artificial intelligence is focused on digital infrastructure, including data centres, semiconductor manufacturing, power systems, and advanced computing equipment, rather than software-based AI services. These tangible, income-generating assets form the foundation of the current AI investment cycle, contrasting with earlier technology booms in which capacity was built ahead of demand. Fibre networks, for example, were already core to telecom operations before the dot-com overbuild, but later investments led to large volumes of unused capacity.

Today's digital infrastructure, however, is tied to established demand across various commercial services. AI-oriented data centres, semiconductor plants, and power systems are integral to the broader digital economy, supporting workloads in cloud computing, enterprise software, and content delivery. This connection to consistent demand gives these assets enduring value, limiting potential financial loss even if AI profitability expectations soften. Additionally, capacity adjustments are manageable for major operators and can be integrated into regular renewal cycles, thereby reinforcing the long-term resilience of these assets.

Supply constraints and market dynamics

The resilience described above is reinforced by the fact that today's AI-related infrastructure cycle operates within much tighter physical and structural constraints than past technology buildouts. Land, grid access, and power capacity are increasingly scarce in most mature data centre markets, while high-performance chips and supporting equipment remain in short supply. These bottlenecks naturally pace deployment and make sustained overbuilding difficult, even in the face of strong capital inflows.

This creates a fundamentally different dynamic from the late 1990s fibre boom, when falling technology costs and regulatory liberalisation enabled virtually unrestricted network expansion. That environment produced large surpluses of unused capacity, often referred to as the "dark fibre moment." In contrast, the current development of AI data centres is constrained by protracted permitting processes, environmental scrutiny, and extended construction timelines, all of which limit the short-term elasticity of supply.

Vacancy rates across major markets illustrate this tightness: in the US, availability fell to a record low of roughly 1.6% in 2025, while in Europe it declined below 10%, despite a 20% year-on-year increase in capacity.

These figures underscore how persistent supply constraints are preventing the kind of speculative overbuild that characterised previous cycles. Even with advances in model efficiency, similar to the reaction sparked by DeepSeek, which appeared capable of delivering ChatGPT-level performance at a fraction of the compute cost, the overall balance between supply and demand would likely remain tight. Forecasts for computing needs have expanded far beyond existing infrastructure capacity, meaning that even significant efficiency improvements would slow the pace of new construction rather than create large-scale oversupply.

The market's reaction to DeepSeek illustrates this dynamic. When the model was first announced, investors briefly sold off AI and semiconductor stocks amid speculation that demand for compute could fall sharply. However, within days, those stocks rebounded, led by Nvidia, as the market recognised that structural demand for compute, power, and data centre capacity remained exceptionally strong. This episode highlights that while efficiency gains may temporarily affect sentiment, they do little to change the underlying trajectory of infrastructure demand.

Stability of AI infrastructure in a slower-growth environment

As such, Data centre cash flows are unlikely to decline sharply even if expectations for AI workloads moderate. Most facilities are supported by long-term leases or hyperscaler contracts tied to cloud and enterprise services, ensuring stable revenue streams. Given the combination of physical constraints and contracted demand, the risk of a systemic overinvestment correction in digital infrastructure is much lower than in earlier speculative cycles.

Should profitability expectations for AI weaken, the adjustment would likely occur gradually through slower expansion and softer valuations rather than a sharp fall in income. Most infrastructure assets are supported by long-term leases and service contracts, ensuring stable cash generation and preserving financial strength. Even if short-term returns moderate, these assets would continue to provide essential digital capacity and generate predictable revenue, resulting in slower growth and greater capital discipline rather than systemic correction. This durability of income underpins the view that today's AI infrastructure expansion, while substantial, is not exhibiting the characteristics of a speculative bubble.

Credit exposure increases, but structural demand remains intact

Implications for digital infrastructure if **AI investment** is bubble-like

The acceleration of investment in artificial intelligence infrastructure is reshaping the digital ecosystem, creating both opportunity and exposure across every layer of the value chain. While the largest technology and cloud providers remain fundamentally resilient, with diversified revenue bases and strong balance sheets, the second and third tiers of the market are far more exposed to a potential slowdown or recalibration in AI demand. The effects of any adjustment are therefore likely to be uneven, transmitting stress first through the most leveraged and least diversified parts of the ecosystem.

Developer-led hyperscale campuses and external financing

The most significant near-term vulnerability lies in developer-led hyperscale campuses that rely heavily on external financing. Unlike hyperscalers, which can fund expansion from retained earnings and recurring cash flow, developers depend on debt or institutional capital to finance construction. Their model focuses on delivering capacity for third-party tenants, requiring significant upfront expenditure while revenues materialise only once space is leased. This reliance on external funding increases exposure to interest rate volatility, cost inflation, and shifts in credit conditions.

Many of these projects are financed through private credit or structured debt facilities and depend on rapid absorption by large cloud tenants to remain viable. If credit conditions tighten or utilisation falls short of underwriting assumptions, developers could face higher refinancing costs, covenant pressure, or valuation adjustments. Exposure is amplified by high fixed costs, rising construction expenses, and constrained grid access, all of which reduce flexibility even if the underlying demand for compute remains intact. In practice, this means that financial stress is likely to emerge through compressed returns, refinancing risk, or extended stabilisation periods rather than a sharp decline in underlying utilisation.

Mature markets such as Northern Virginia, Dublin, and parts of continental Europe continue to experience grid congestion and rising energy prices, which place pressure on margins. However, these effects occur against a backdrop of structurally constrained supply and record-low vacancy rates. Should AI-related demand moderate, operators may rebalance development toward lower-cost or less constrained markets rather than abandon existing campuses, with utilisation more likely to adjust through timing and pricing than through a sharp contraction in activity.

Diversified infrastructure operators

In contrast to these more vulnerable segments, diversified infrastructure operators have a more substantial capacity to withstand cyclical fluctuations. Established data centre real estate investment trusts with large tenant bases and recurring colocation revenues retain stable baseline utilisation even if AI-related leasing moderates. Their exposure to non-AI workloads, including cloud, enterprise software, and digital content delivery, provides a steady source of cash flow that supports debt service and valuations even in the event of a contraction in expected AI-linked compute demand. Network and interconnection assets also remain supported by persistent structural drivers such as video streaming, remote working, and cloud migration, which sustain traffic growth independently of AI demand. As a result, any adjustment is more likely to be reflected in slower growth or modest return compression rather than in material declines in utilisation or cash flow.

Hyperscaler campuses

At the system-wide level, the most integrated and capital-strong entities are likely to remain insulated from short-term volatility. Core hyperscalers such as Microsoft, Amazon, and Alphabet possess the liquidity and internal demand to repurpose compute capacity across cloud, consumer, and enterprise services. Their ability to re-optimise workloads, defer specific capital projects, and redirect investment allows them to avoid the liquidity constraints that affect developer-led operators. The energy and utility infrastructure that underpins this entire ecosystem is also positioned to retain long-term relevance. The ongoing electrification of transport, the expansion of renewable generation, and the digitalisation of industry all ensure that grid upgrades, transmission expansion, and battery storage investments will continue to attract capital even if AI-specific demand slows.

Uneven repricing and structural stability

Taken together, these dynamics suggest that an AI investment correction would not trigger a systemic collapse in digital infrastructure but rather an uneven repricing across segments. Stress would be concentrated in leveraged developers, power-constrained regions, and hardware suppliers tied to short-order cycles, while diversified operators and capital-rich platforms would maintain relative stability.

Such a recalibration would accelerate consolidation, improve capital discipline, and sharpen the distinction between speculative expansion and sustainable, revenue-backed infrastructure.

Concentrated gains and long payback cycles raise questions about the sustainability of AI

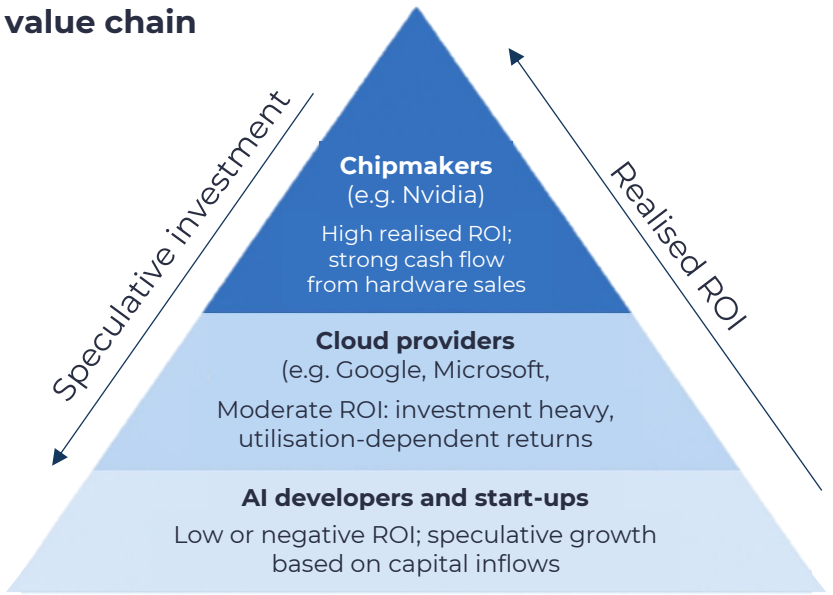
Why ROI could indicate an AI bubble

Fears of overspending largely stem from uncertainty over whether these technologies will generate returns that justify the scale of investment, especially in light of the rapid acceleration in spending.

Narrow ROI concentration

Financial returns from artificial intelligence remain concentrated among a small group of dominant firms, suggesting that much of the sector has yet to convert substantial capital spending into broad profitability. Most realised gains continue to flow to companies at the core of the AI value chain, with Nvidia capturing a large share of industry profit growth and Microsoft and Alphabet beginning to see meaningful earnings uplift through cloud and productivity services. Returns beyond this group are far less consistent. Meta has improved advertising and operational efficiency, albeit modestly relative to the scale of its infrastructure investment. Oracle has experienced rapid growth in cloud and AI-related revenues; however, these gains have not yet translated into stronger cash generation. In fiscal year 2025, it reported a negative free cash flow of approximately USD 390 million following substantial spending on data centre and AI capacity.

Exhibit 5.1: ROI concentration across the AI value chain



Deferred payback cycle

The long and uncertain payback period adds further risk. The largest technology firms are investing at record levels, yet revenue growth is slower. Meta's costs rose by 35% last year compared with an 18% rise in revenue, and Microsoft's depreciation and amortisation expenses increased from roughly 11-17% of revenue. Deutsche Bank expects that many large AI infrastructure projects will not reach breakeven until 2028 or 2029. Such delays are typical in capital-intensive industries, but they become problematic when expectations of return outpace actual cash generation.

Substantial gains are required to achieve even a modest ROI

The issue of delayed payback is further compounded by the large amount of revenue needed to achieve even a modest ROI from AI investments. J.P. Morgan estimates that USD 650 billion in annual revenue is required in perpetuity to deliver just a 10% ROI on AI infrastructure investments. This substantial target underscores the immense scale of financial commitment required to justify modest returns in a sector where growth remains speculative.

The capital-intensive nature of AI projects only furthers this challenge. Infrastructure such as data centres and high-performance computing systems requires significant upfront investment, with returns spread over a long period. As payback cycles extend, the pressure mounts to sustain massive annual revenue generation.

Model convergence and competitive pressure

The rapid proliferation of frontier AI models is creating a structural ROI problem as the same demand pool is now being chased by more capital-intensive competitors, compressing the economic return available to each. Alphabet's Gemini 3 has overtaken ChatGPT-5 on key benchmarks, and Anthropic's Opus 4.5 is outperforming OpenAI in several enterprise-relevant tests. Gemini 3's monthly downloads are also now approaching those of ChatGPT, and monthly users have reached 650 million, narrowing the gap with ChatGPT's 856 million. As performance converges across models, competition is likely to shift toward distribution and pricing, which may compress margins and put downward pressure on ROI, while rising spend and duplicated investment could further weaken returns over time

Historical precedent and case for caution

The telecommunications expansion of the late 1990s offers a useful comparison. At that time, network operators assumed that bandwidth demand would grow exponentially and invested far ahead of actual usage. When demand failed to keep pace, excess capacity and weak utilisation followed, limiting returns for years.

The AI sector shows similar characteristics today. Capital spending is expanding rapidly; profits are concentrated and returns depend on future utilisation. Although AI is a far more transformative technology, the investment cycle may be running ahead of realised productivity. This imbalance between financial momentum and tangible output has often been followed by periods of consolidation, after which steadier growth resumes.

Early profitability is concentrated but operationally grounded

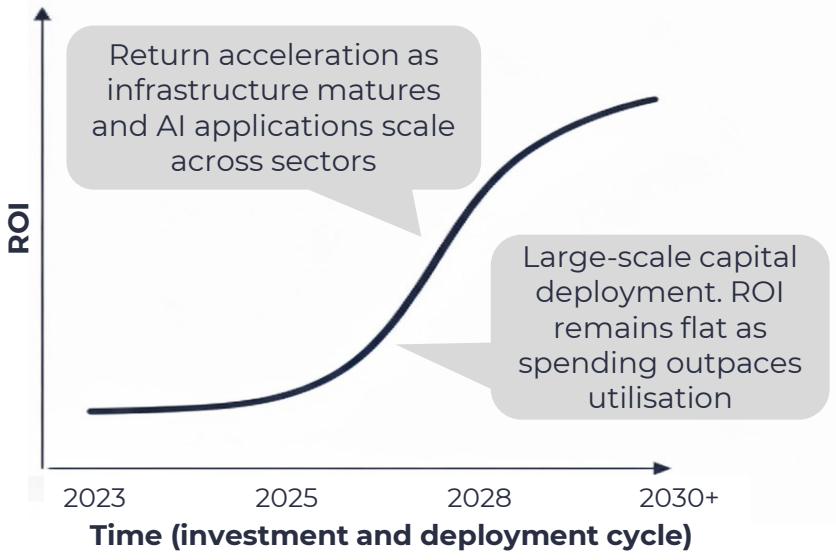
Why current AI returns do not yet signal a bubble

Concentrated but legitimate early-stage profitability

Financial returns from AI investment are concentrated among a few leading firms; however, this profitability is becoming increasingly tangible and measurable. For example, Azure's AI revenue grew by 39% from the previous year based on Q3 of 2025 earnings reports, Alphabet's AI-supported advertising income rose by 15%, and Nvidia's net income increased by 22%, with operating margins surpassing 50%. Collectively, these companies generated over USD 110 billion in quarterly free cash flow and maintained steady double-digit earnings growth.

Much of this growth is driven by the adoption of AI infrastructure, including cloud services and AI hardware like Nvidia's GPUs, both of which are foundational for AI scalability. While a portion of this growth is driven by multi-year enterprise contracts that provide predictable demand, some AI applications, particularly in newer, innovative sectors, remain speculative. However, the early profitability of infrastructure providers does not necessarily indicate overvaluation. Instead, it shows that returns are first captured by firms that develop foundational technologies, and as adoption deepens, the broader ecosystem benefits.

Exhibit 5.2: Illustrative ROI trajectory across the AI investment cycle



Improving efficiency and structural ROI timing

Operational efficiency across the sector is improving quickly, supporting the view that current investment is delivering genuine productivity gains. Deutsche Bank data show that GPU throughput per watt has increased three to four times since 2022, while energy efficiency per inference task has more than doubled since 2023. Utilisation within large data centres is steadily rising as workloads become better optimised and trained models are reused across different applications.

These improvements suggest that the economics of AI production are strengthening, not deteriorating. Earlier inefficiencies reflected a typical adjustment period during which capacity expanded faster than demand. The delay between capital spending and return, often cited as a speculative risk, can also be interpreted as a structural feature of infrastructure rollouts. The current lag in returns, therefore, may reflect sequencing and maturation rather than financial overreach.

ROI from AI Adoption: evidence against an AI bubble

A recent Wharton report reveals that 74% of enterprises are already seeing positive ROI from AI investments, particularly in sectors like Tech, Telecom, and Banking/Finance. These industries are leveraging AI to drive measurable business value, including improved customer service, enhanced data analysis, and operational efficiency. Additionally, 80% of enterprises expect continued ROI over the next two to three years, indicating that AI adoption is delivering long-term benefits rather than speculative returns.

This growing demand for AI solutions, combined with tangible productivity gains and improved decision-making, shows that AI investments are grounded in real business value. The consistent and sustained ROI supports the view that AI is moving beyond pilot projects, signalling strategic growth driven by real-world demand, rather than an overinflated bubble.

Rational infrastructure buildout phase

Taken together, these developments indicate that the AI industry is undergoing a rational buildout rather than an unsustainable speculative surge. High capital intensity does not automatically imply overvaluation. In transformative industries, investment almost always precedes measurable output, as seen in the early years of cloud computing and broadband networks. Those sectors initially appeared unprofitable before utilisation and scale efficiencies produced substantial returns.

What may appear to be speculative behaviour can therefore be read as rational preparation for expected demand. Firms are investing heavily to secure a competitive advantage, backed by growing cash flows, stronger efficiency metrics, and robust balance sheets. The AI sector may therefore be in an early phase of a long-term growth cycle in which the heavy spending of today lays the groundwork for productivity and profitability in the years ahead.

The return divide of the AI buildout: Uneven gains but enduring fundamentals

Implications for Digital Infrastructure if AI ROI Is Bubble-Like

Implications for digital infrastructure

If returns on artificial intelligence remain sluggish or concentrated among a few dominant firms, the effects on digital infrastructure would be mixed. Underlying demand for capacity would persist, but growth would become more selective and geographically uneven. Concentrated profitability among large technology groups would reinforce the resilience of the biggest operators but slow expansion across the broader market. Over time, investment flows would shift from speculative projects to mature platforms with predictable income and strong counterparties.

Hyperscale data centres

Hyperscale data centres would remain the most resilient part of the ecosystem. The companies capturing the most AI returns, such as Microsoft, Google, Amazon and Nvidia, are also the leading developers, owners and tenants of extensive facilities. Even if AI profitability slows, these firms would continue to generate steady cash flows from cloud computing and enterprise software that sustain utilisation.

A slower ROI cycle would temper expansion rather than income. The same concentration that limits broader profit distribution would strengthen credit quality, ensuring that the most extensive facilities remain well-financed and fully occupied.

Colocation and enterprise facilities

Colocation and enterprise facilities would be more sensitive to a slowdown in AI-related investment, as their client base includes smaller technology firms and corporate users testing AI solutions. If access to venture capital weakens, demand for incremental space could plateau, and renewal pricing could come under pressure.

At the same time, this could trigger substitution effects. When confidence declines, enterprises often lease space instead of building new facilities. Colocation providers with diversified clients and strong balance sheets would be best positioned to capture this shift. A similar adjustment would occur in regional and edge data centres, where expansion would slow but remain supported by diverse workloads such as content delivery, automation and the Internet of Things.

Edge and regional data centres

Edge and regional data centres would experience a slower buildout if AI activity remains concentrated in centralised hyperscale clusters. The expected growth of distributed inference and low-latency computing could arrive more gradually than anticipated, moderating utilisation growth in smaller facilities. However, these sites serve a diverse range of applications, including content delivery, gaming, industrial automation and the Internet of Things, which continue to expand independently of AI profitability.

Their shorter contract durations and higher yields allow operators to adjust pricing and scale flexibly. A period of sluggish AI ROI could therefore produce a more disciplined expansion cycle, focused on proven workloads rather than speculative demand forecasts. The sector would grow at a steadier pace but would not lose structural importance.

Fibre networks and power infrastructure

Fibre and network infrastructure would face little direct impact from slower AI profitability. Revenues are tied to long-term contracts for capacity and interconnection, based on throughput and latency rather than technology cycles. Even if AI traffic growth eases, baseline demand from streaming, enterprise connectivity and cloud services would continue to rise, supported by financially strong hyperscalers.

Energy infrastructure would remain similarly stable. A slowdown in AI-related expansion might delay new power projects, but existing purchase agreements and grid connections would keep income secure. In some markets, reduced growth could ease the strain on power systems and support the integration of renewables.

Sector-wide outcomes

A slower or more concentrated pattern of AI returns would prompt adjustment rather than contraction across digital infrastructure. Hyperscale operators would consolidate their position, maintain stable cash flows and capture greater market share as smaller developers retrench. Colocation and edge providers would experience moderate volatility but remain essential to enterprise and distributed workloads, supported by ongoing digitalisation trends. Fibre and power networks would continue to provide a steady income underpinned by long-term contracts and baseline data demand.

Overall, the concentration of realised returns among leading technology firms would slow the pace of new investment but reinforce the sector's financial stability. Rather than a systemic downturn, the outcome would be one of capital rotation toward mature, income-generating assets and stronger discipline in project selection. This shift would primarily affect growth and capital allocation rather than the credit quality or income stability of core digital infrastructure assets.

The physical limits of the AI rise: Infrastructure strain and asset depreciation

Hardware assumptions may be quietly shaping the AI bubble debate

One of the most critical yet under-examined dimensions in the discussion as to whether AI constitutes a speculative bubble is the underlying hardware infrastructure. Recently, capital expenditure (capex) related to chips, GPUs, and accelerators, and especially their depreciation cycle, has sparked the debate, given its criticality for current valuations of chip manufacturers and underlying digital infrastructure assets.

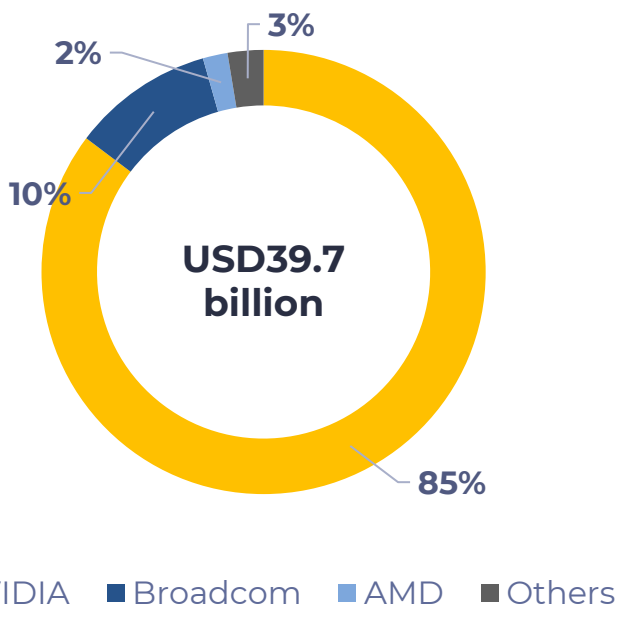
Why might it be a bubble?

AI is both capital and energy-intensive, resulting in high entry barriers and financial risks associated with the effective life cycle of essential hardware. Meta's Llama 3 LLM, for instance, was trained using approximately 24,000 NVIDIA H100 chips, which cost over USD 720 million. Such costs make it nearly impossible for new LLM entrants lacking significant financial capacity to compete at scale. Accelerated servers dedicated to AI training and specialised workloads are projected to account for roughly half of total data-centre capex by 2029, highlighting the importance of hardware investment in the current AI cycle.

Nvidia's market dominance and the hardware power structure

Nvidia holds dominance in the AI chip market for data centres, particularly in AI accelerators used for training large foundation models. IDC estimates that as of 2025 Q2, Nvidia captured 85.2% of the AI accelerator market share, followed by Broadcom (10.3%) and Marvell (2.1%).

Exhibit 6.1: AI accelerator vendor share, Q2 2025



Source: *IDC CY1Q25 Semiconductor Application Forecast, 2025*

Furthermore, the market exhibits structural stratification. At the same time, Intel and AMD can compete in more general-purpose or inference-oriented accelerators. Nvidia, however, retains near-total control of the premium training segment, supported by its proprietary hardware architectures (Hopper and Blackwell) and interconnect systems. Competitors such as AMD and Intel are advancing, but they have yet to erode Nvidia's lead significantly. A more credible long-term threat may come from efficiency-focused entrants such as Qualcomm, whose AI200 and AI250 chips (expected 2026–27) target the inference segment with high-memory, low-power designs.

Beneath this dominance lies Nvidia's actual competitive shield: its CUDA ecosystem. CUDA, a parallel computing platform developed over the past two decades, provides a complete software stack (compiler, driver, runtime and toolkit) that has become indispensable for developers seeking to accelerate complex AI applications. The deep learning ecosystem has effectively standardised around CUDA, creating a powerful network effect that makes switching costs prohibitively high. Competing hardware vendors face not only the technological challenge of producing equivalent silicon but also the immense difficulty of fostering a software community capable of optimising workloads at a comparable scale. Nvidia's vertical integration of hardware and software thus functions as a barrier to entry, reinforcing both market power and the perception of speculative dependence on a single supplier.

However, the competition is no longer just aiming to create a better chip; they are working to de-standardise CUDA through parallel ecosystems, open-source toolkits, and alternative custom silicon. For the next 24 to 36 months, CUDA's moat is likely secure, but its long-term viability is under the most intense pressure it has ever faced.

Hardware lifecycle and depreciation risk

A central piece of the "bubble" argument rests on the rapid cadence of investment in AI infrastructure and the possibility that the depreciation assumptions embedded in corporate and investor models are overly optimistic or mis-specified. The very short lifecycle often attributed to AI-specific hardware is a warning flag, considering that by reducing the assumed years of useful life, chip manufacturers may inflate future sales expectations. Industry estimates place the practical lifespan of data-centre GPUs at only one to three years, particularly under high-utilisation training workloads.

The physical limits of the AI rise: Infrastructure strain and asset depreciation

Hardware assumptions may be quietly shaping the AI bubble debate

Two interrelated factors constrain this cycle: physical stress and technological obsolescence. GPUs running AI workloads typically operate at sustained utilisation rates of 60–70%, generating significant thermal and electrical strain. Google's hardware architects estimate that such operating conditions limit physical lifespan to roughly one or two years, with three years as an upper bound. Even if the chips remain functional, the rapid pace of innovation renders older generations economically obsolete. For example, Nvidia's GB200 (Blackwell) offers four to five times faster inference than the H100, making older hardware non-competitive for latency-critical training tasks.

This short lifecycle aligns closely with Nvidia's own support policies. Data-centre GPU drivers in the vGPU C-Series receive formal support for roughly one year, with major releases every six months, while long-term support branches (LTSB) extend coverage to a maximum of three years. The implication is clear: three years represents the upper limit of Nvidia's guaranteed operational reliability window. Many firms, however, still depreciate AI hardware over five- or six-year horizons, creating an accounting mismatch between reported and economic depreciation. This mismatch, similar to those observed in previous speculative episodes, such as the dot-com bubble of the 2000s, smooths costs over a more extended period while concealing the actual rate of obsolescence.

Algorithmic efficiency and the cost-disruption paradox

A further emerging challenge comes from algorithmic efficiency, which, in theory, could undermine demand for high-end hardware. The case of China's DeepSeek, backed by the hedge fund High-Flyer, is illustrative.

DeepSeek has demonstrated that large language models (LLMs) with GPT-4-level performance can be trained for as little as USD 6 million (barely a fraction of the USD 100 million reportedly required for GPT-4 in 2023). This efficiency stems from architectures such as *Mixture-of-Experts* (MoE), which activate only a subset of model parameters per token, drastically reducing GPU operations. Combined with techniques like context caching and quantisation (INT8/INT4), such methods can cut inference costs by 75-90%.

At first glance, this appears to threaten Nvidia's dominance and to suggest a looming collapse in hardware demand. In reality, it may have the opposite effect. By reducing entry barriers and training costs, algorithmic efficiency democratises AI development, enabling a broader base of firms and research institutions to train and deploy their own models.

Aggregate hardware demand thus remains strong, even if distributed across more actors. Moreover, Nvidia has strategically embraced this shift: its NeMo Automodel toolkit integrates MoE optimisation directly into the CUDA ecosystem, ensuring that even more efficient models remain dependent on Nvidia's software infrastructure. Paradoxically, the very innovations that appear to erode the need for high-end chips may reinforce Nvidia's centrality, transforming what could have been a deflationary risk into a driver of long-term platform entrenchment.

Taken together, these factors sustain the argument that valuations may be inflated by short-term optimism around hardware turnover and market dominance. Any revision of lifecycle assumptions or a sudden loss of confidence in the durability of Nvidia's lead could trigger a rapid re-rating of AI infrastructure assets, exposing speculative excess.

Why might it not be a bubble?

There are, however, several arguments suggesting that the hardware-life challenge does not necessarily mean we are in a classic speculative bubble or that the risk is more nuanced.

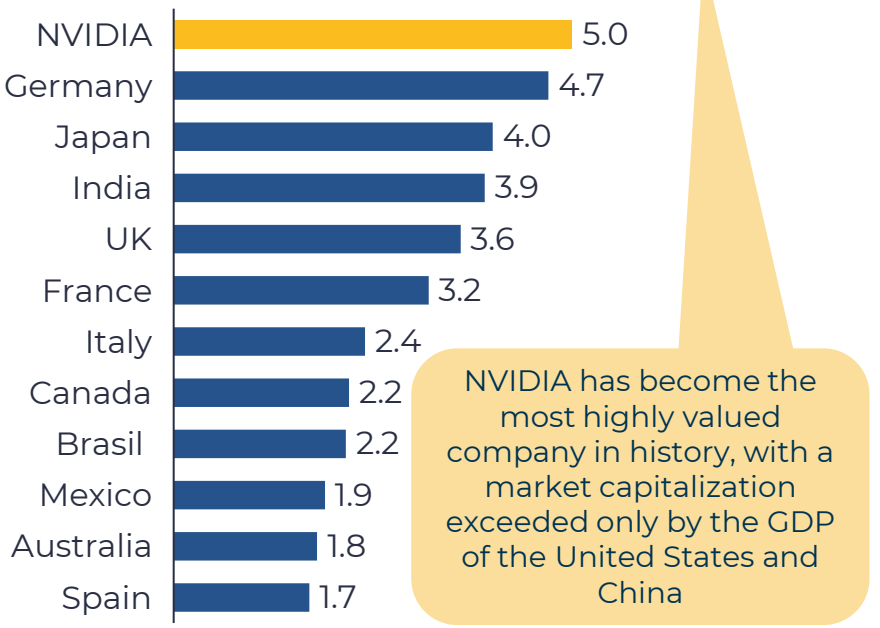
Reutilisation and secondary market resilience

First, GPUs enjoy a "second life" that extends their economic usefulness well beyond the initial training phase. Once superseded by newer generations, top-tier chips (such as H100s or A100s) are redeployed for lower-latency inference tasks, where performance thresholds are less demanding, but volumes are high. This strategic redeployment (shifting from training to inference) significantly prolongs the economic life of the asset. According to Morgan Stanley, the inference business is "astonishingly profitable": an A100 purchased in 2021 for elite training may still be profitably employed in 2024 for high-throughput inference, and again in 2026 for batch inference or less time-sensitive workloads.

The physical limits of the AI rise: Infrastructure strain and asset depreciation

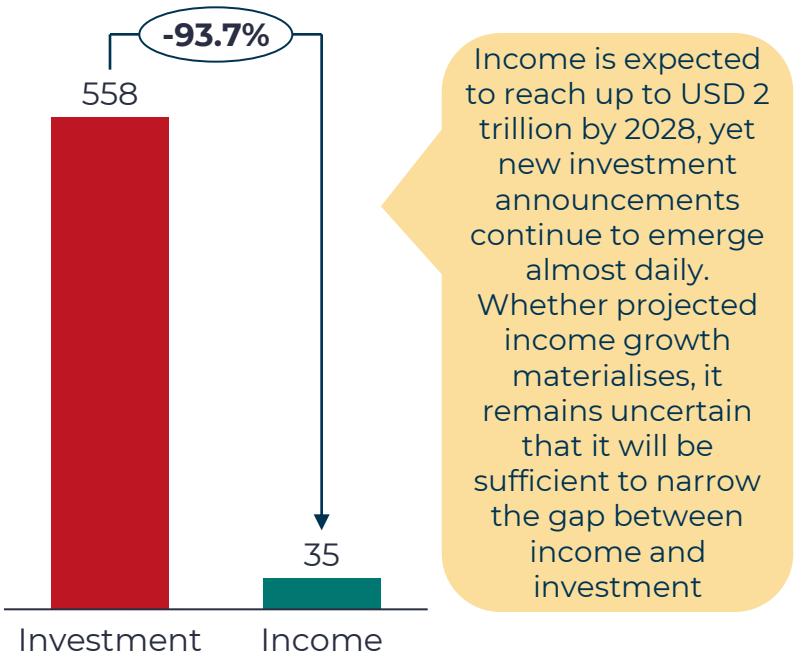
Scale does not always mean speculation

Exhibit 6.2: NVIDIA's valuation vs. selected countries' GDP as of October 2025 (USD trillion)



Source: *Reuters, World Bank, October 2025*

Exhibit 6.3: AI investment vs. AI income as of August 2025 (USD billion)



Source: *The Information, August 2025*

The robustness of the secondary market supports this view. Unlike the collapse of residual values in crypto-mining equipment, the resale market for previous-generation GPUs (V100s, A100s) remains active. Specialist resellers test and certify refurbished units for enterprise-grade performance, offering warranties that are comparable to those of new equipment. Major cloud providers (including AWS, Azure, Google Cloud, CoreWeave and Lambda Labs) continue to rent out such GPUs alongside the latest H100s, confirming that depreciation in the training domain does not equate to full economic obsolescence. Nvidia's continued software support and backward compatibility through CUDA further sustain this residual value, reinforcing the hardware's longevity as a productive asset.

Capex commitment and the physical infrastructure of AI

The scale and nature of current AI investment suggest a transformation of industrial, not speculative, character. Hyperscalers, such as Meta, Alphabet, and Microsoft, have all reported sustained and accelerating capex growth dedicated to AI infrastructure. Meta, for instance, raised its 2025 capex guidance to USD 71 billion and expects a "notably higher" figure in 2026, potentially exceeding USD 100 billion, compared with USD 39.2 billion in 2024. Alphabet's projections for 2026 have similarly risen to USD 92 billion. Nvidia's CEO Jensen Huang estimates total AI infrastructure spending of between USD 3-4 trillion over the next five years, with annual outlays potentially surpassing USD 1 trillion by 2029.

Crucially, this investment extends far beyond chips themselves. Supporting components (power supply, cooling systems, and electrical infrastructure) represent immense complementary capex. AI-ready data centre racks now consume 60 kW or more per rack (often >100 kW), compared with fewer than 10 kW for traditional deployments. Power infrastructure alone adds approximately USD 4.5 million per MW of capacity, on top of USD 25 million per MW for the servers themselves. The physical, energy-intensive nature of these facilities underlines that this is not a speculative boom but a transformation with tangible utility.

Long-term financial commitments further reinforce this point. OCI, for example, has reported record AI demand, underpinned by a five-year, USD 300 billion infrastructure partnership with OpenAI. This single contract caused Oracle's RPO to surge 359% YoY to USD 455 billion, much of it tied directly to AI compute. These legally binding contracts ensure sustained demand for state-of-the-art GPUs and related infrastructure for years ahead.

Maturity of market adjustment

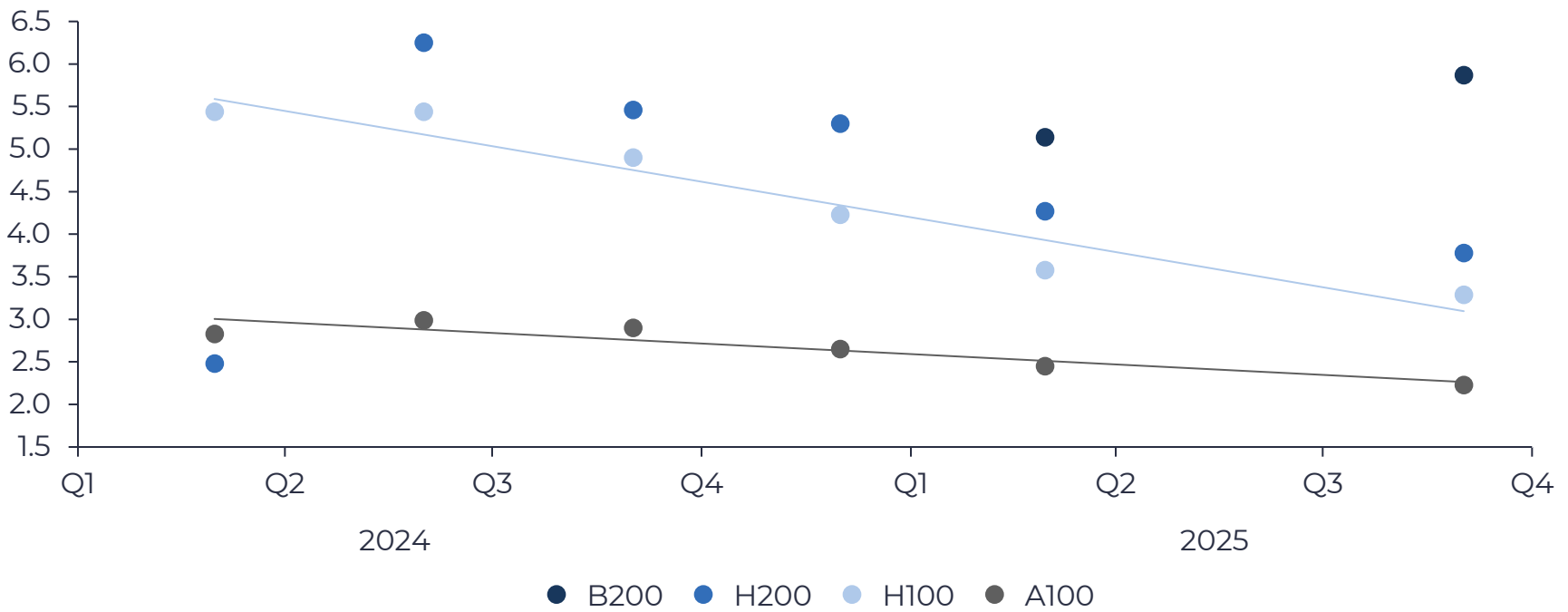
Finally, while depreciation cycles are short and costly, market participants appear aware of these dynamics. Disclosure practices, investor scrutiny, and accounting adjustments are evolving in response. Hyperscalers' ability to internalise costs through diversified business models and repurpose hardware across multiple use cases mitigates the systemic risk of overinvestment. As awareness of lifecycle economics spreads, valuations are likely to adjust in a more measured way, reflecting a maturing phase of capital formation rather than a speculative bubble.

ASIC: Application-Specific Integrated Circuit | TCO: Total cost of ownership | OCI: Oracle Cloud Infrastructure | YoY: Year-on-year | RPO: Remaining Performance Obligations

The physical limits of the AI rise: Infrastructure strain and asset depreciation

Market maturity behind the AI capex cycle

Exhibit 6.4: Nvidia GPU rental pricing (USD/hour)



Source: *Financial Times*

Insight on GPUaaS capex compression and market maturity

Recent Fide Partners' analysis of GPU-as-a-Service (GPUaaS) providers illustrates how market adjustment mechanisms are already materialising. Across NVIDIA's latest models (A100 to B200), average rental rates have declined from around USD 5.5 per hour to below USD 3 per hour between 2024 and 2025, even as compute capacity per chip continues to expand. This signals the beginning of a deflationary trend in compute pricing, one that mirrors earlier waves of commoditisation observed in cloud infrastructure and fibre connectivity.

The implications are significant. As GPUs become more powerful and abundant, revenue per teraflop continues to decline, compressing margins across the ecosystem and shifting competition away from hardware availability toward operational efficiency. Neocloud operators such as Crusoe and CoreWeave are being pushed to scale volumes at unprecedented rates, leveraging proximity to data centres and energy efficiency as the new differentiators in a market where hardware parity is rapidly emerging.

Unit-economics modelling for H100-based deployments suggests a capex outlay of approximately USD 35,000 per GPU and a payback horizon of 2.3–2.5 years under a 30% annual price-erosion assumption. Beyond this window, profitability erodes quickly as newer GPU generations displace earlier units, compressing both asset value and resale potential. For smaller GPUaaS providers, the resulting capital intensity and depreciation pressure limit reinvestment capacity and increase dependency on financing cycles tightly linked to NVIDIA's product refresh cadence.

From a strategic standpoint, these dynamics reinforce the notion that the AI infrastructure market is entering a phase of rationalisation rather than speculative contraction. The bubble narrative gives way to a more structural adjustment, one where efficiency, access to low-cost energy, and supply-chain coordination determine long-term competitiveness. In this context, NVIDIA's dominance extends beyond chip performance to ecosystem control: the company effectively sets the pace of reinvestment for an entire downstream industry now racing to maintain equilibrium between utilisation, obsolescence, and capital recovery.

Ultimately, the ongoing compression in GPU rental economics highlights the fragility of the current capex cycle. While aggregate demand for compute remains strong, returns are increasingly constrained by physical depreciation and accelerating technological obsolescence, suggesting that sustained profitability will depend on transforming AI infrastructure from a speculative asset class into a stable, utility-like service layer.

ASIC: Application-Specific Integrated Circuit | TCO: Total cost of ownership

Monetisation on the back end: Justifying the investment in AI infrastructure

Will monetisation catch up with investment?

The unprecedented levels of investment in AI infrastructure raise some crucial questions: what is the monetisation model on the back end that will justify such outlays? How much longer can market players sustain high capex without visible profit? Is the benefit effectively a tacit or embedded one (for example, increased cloud consumption, greater lock-in, higher customer lifetime value)?

Why it might be a bubble

If the monetisation models remain implicit rather than explicit (i.e., if customers do not pay directly for AI services or if revenue growth is driven purely by ecosystem effects), then valuations may be acutely vulnerable. The path to viability for the pure foundation-model operators, such as OpenAI, depends on achieving scale, attaining broader applicability, and shifting to enterprise subscription models. Analysts from UBS expect AI spending to exceed USD 2.6 trillion by 2030, underscoring the vast upside potential. Yet, expectations must be converted into predictable revenue streams; otherwise, investment becomes speculative. Recent forecast revisions suggest that some headline projections may be overly optimistic. While specific analyses estimate up to USD 2.6 trillion in monetisable AI value by 2030, more conservative views place cumulative revenues below USD 2 trillion, compared to nearly USD 4 trillion in infrastructure spending. This imbalance implies that a significant share of today's investment may be chasing demand that has yet to materialise.

Moreover, enterprise software marketplace sales via hyperscaler clouds are expected to grow from USD 30 billion in 2024 to USD 163 billion by 2030, representing a significant leap but still modest compared to the underlying infrastructure investment. However, it is essential to note that only a fraction of this marketplace growth corresponds to AI-driven enterprise solutions. Most SaaS products distributed through hyperscaler marketplaces are still non-AI or only partially AI-enhanced, meaning that the revenue base directly attributable to AI remains smaller than the headline figures suggest. This gap reinforces the risk that infrastructure investment may be outpacing the monetisable portion of AI-native workloads.

Emerging evidence already points to pressure in the underlying economics. If the monetisation model depends on perpetual reinvestment, rapid hardware turnover, or fast model churn, any slowdown in demand or prolonged inference cycles could quickly undermine viability. Sustaining today's infrastructure build-out would require approximately USD 2 trillion in annual AI revenues by 2030, a target that is increasingly difficult to achieve given current adoption and refresh dynamics. Deloitte further notes a widening "AI ROI gap", characterised by rising spend but persistently elusive returns across enterprises. The limited scrutiny applied to back-end revenue assumptions is therefore an additional red flag.

Should a valuation correction or abrupt funding slowdown occur, the impact would propagate unevenly across the AI value chain. The first point of exposure lies with chip manufacturers and GPU vendors (e.g., NVIDIA), who are financing large-scale capacity expansions based on exceptionally strong demand expectations. If revenues fall short or capital inflows weaken, hardware suppliers would face disproportionate downside risks, including valuation impairments, margin pressure, and balance-sheet strain that could curtail supply or trigger rapid markdowns.

Hyperscalers (e.g., Microsoft, Google, AWS) represent the second channel of exposure. Their risk stems from both direct capex commitments and indirect exposure through model partnerships and supply contracts. Although their diversified revenue bases and substantial cash flows provide resilience, weaker AI monetisation could extend capex cycles, compress returns, and force a re-prioritisation of investment pipelines, slowing expected ROI.

The third and most vulnerable segment comprises companies that rely almost entirely on the digital infrastructure stack, including many SaaS and AI-enabled application providers. Their dependence on affordable GPU access, hyperscaler distribution channels, and investor willingness to tolerate long monetisation timelines leaves them highly sensitive to shifts in sentiment or contracting terms. If either tightens, many AI-enhanced product roll-outs may fail to convert into sustainable revenue—an outcome already hinted at by recent sell-offs in social and ad-tech firms following weaker-than-expected AI adoption signals.

Why it might not be a bubble

On the other hand, there are strong arguments that suggest the monetisation model of AI infrastructure may become sustainable, and that current investment is strategic rather than purely speculative.

Monetisation on the back end: Justifying the investment in AI infrastructure

Will monetisation catch up with investment?

First, hyperscalers' AI investments are tightly integrated into broader cloud ecosystems. AI services act as catalysts for consumption of high-margin cloud offerings (compute, storage, security, databases), enhancing customer lifetime value and increasing platform lock-in. The model is not simply "sell AI" but "use AI to sell more cloud". Hyperscalers are uniquely positioned to monetise AI in this way, and their scale and cash-flow strength mitigate many of the risks associated with smaller pure-AI plays.

Second, the pure foundation-model operators are increasingly moving toward enterprise subscription and licensing frameworks, shifting away from volatile pay-per-token API models to longer-term contracts with higher predictability. Anthropic is one of the clearest examples of this shift; the company has launched multi-tier enterprise subscriptions and model-use licences that bundle fixed monthly fees, dedicated capacity, and usage guarantees, effectively smoothing revenue volatility and aligning monetisation with corporate procurement cycles. This approach marks a deliberate move toward stable recurring income rather than transactional API variability.

Third, multiple industry analyses suggest that the AI infrastructure build-out is not merely a short-term cycle but rather part of a structural transformation of computing. For example, the migration from traditional IT to cloud and AI is only in its early innings, with a large legacy market still to be transitioned. A figure that supports this is the actual percentage of enterprises that use AI in the US: even though approximately 78% of surveyed companies claim they are using AI, only 9.7% are effectively utilising it in 2025. If adoption maintains a slow pace, then the monetisation path will have time to mature, and the business models may evolve without collapsing under the weight of upfront investment.

Finally, some revenue realisation is already evident. Enterprises are beginning to commit to multi-year cloud and AI marketplace agreements, and agentic AI demand is growing at a double-digit annual rate, suggesting that the "back-end" monetisation may indeed be materialising rather than remaining hypothetical. The AI boom is a long marathon, not a sprint.

Monetisation on the back end: Justifying the investment in AI infrastructure

AI infrastructure monetisation models on the back end

Model	Revenue sources	Cost structure	Key risks	Indicative sustainability
Hyperscaler-integrated model (Microsoft Azure, AWS, Google Cloud)	• Indirect monetisation via increased cloud consumption (compute, storage, networking) • Higher customer lifetime value and platform lock-in • Premium enterprise AI services and co-developed solutions • Marketplace commissions and third-party AI integrations	• Extremely high capex in data centres, GPUs and networking • Energy costs and PPA exposure • R&D on AI orchestration layers and model integration	• Revenue dilution if AI does not materially increase cloud usage • Over-capacity risk if demand slows • Hardware obsolescence (shorter asset life)	 **High (structurally sustainable)** due to diversified cash flows and ecosystem synergies
Foundation-model pure-play (OpenAI, Anthropic, Mistral, xAI)	• API usage fees (pay-per-token) • Enterprise subscriptions and licensing • Model fine-tuning and hosted solutions • Partnerships with hyperscalers for infrastructure or distributionText	• Extremely high training costs (compute, data, engineering talent) • Large upfront investment before monetisation • Limited scale economies at early stage	• Monetisation lag vs. capex cycle • Dependency on few large customers (e.g. hyperscalers) • Competitive erosion as open-source models improve	 **Medium (conditional)** requires scale, brand and differentiated data to reach break-even
Vertical-embedded AI model (AI integrated in sector-specific applications: finance, healthcare, industry)	• AI-enhanced SaaS subscriptions • Productivity tools embedding AI features (Copilot-type models) • Value capture through efficiency gains or pricing premiums	• Lower direct capex • Higher marginal R&D per use case • Integration and compliance costs	• Monetisation may be offset by price compression or substitution • Dependency on underlying model providers	 **High to medium** depends on ability to capture tangible productivity improvements

The unfinished story of the AI boom: bubble or transformation?

So, is AI a bubble from a digital infrastructure angle?

Taken together, the elements analysed illustrate the dual nature of the ongoing debate. While speculative patterns and extreme valuations evoke historical parallels with past bubbles, the structural, financial, and operational foundations of today's AI sector appear stronger and more resilient. The evidence suggests that, rather than a speculative hype, the AI boom may reflect a genuine, albeit uneven, process of technological transformation with profound macroeconomic implications.

The market may envisage different potential outcomes. In the most pessimistic case, AI investment could prove overextended, echoing past episodes of capital misallocation and triggering a sharp correction in equity markets. A second possibility is a protracted geopolitical race, in which the United States and China double down on AI as a strategic priority, fuelling state-backed spending and fiscal expansion. At the opposite end, AI could ultimately deliver on its transformative promise, driving a new wave of productivity and innovation, albeit with disruptive effects on labour markets and policy frameworks.

Tom Allegaert
Managing Director - Americas

Alejandro Cárdenas
Director - UK

Nicolas Betancur
Senior Consultant

Natalia Serrano
Bogota office

Gabriela Villegas
Bogota office

Daniel Mansi
London office

Our offices:

London:
125 Kingsway London
England WC2B 6NH
6th floor, Office 107
United Kingdom

Madrid:
C/Don Ramón de la Cruz, 6, 1°
28001 - Madrid
Spain

Bogotá:
Carrera 11A #98-50
Ofc. 704, Edificio Punto99
110221, Bogota
Colombia

Boston:
50 Milk Street,
Planta 15, C.P. 02109.
Boston, MA

Mexico

Telephone:
+34 910 244 113

Mail:
info@fidepartners.com

Web:
https://fidepartners.com