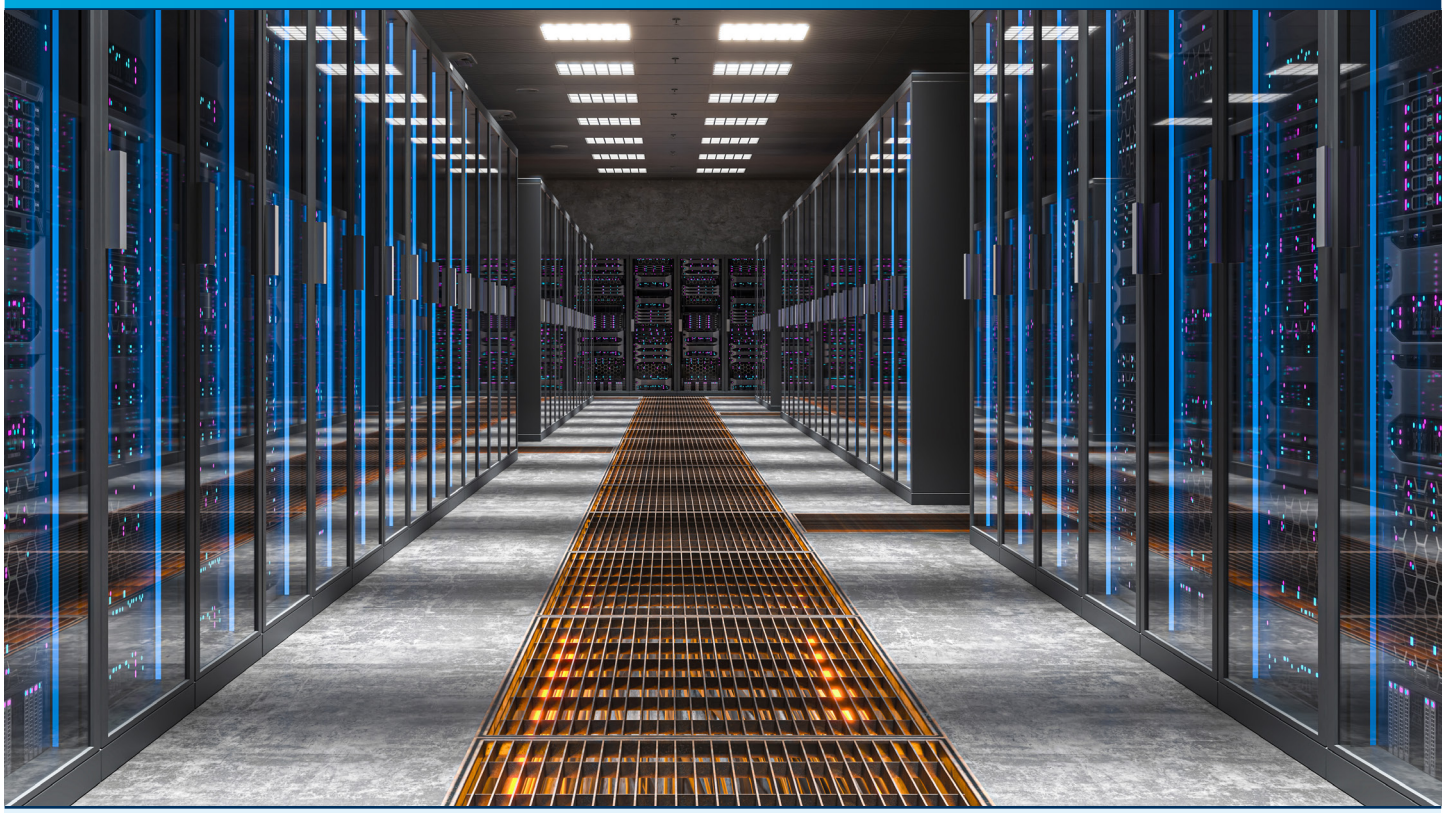
Issued: May 2026

Ashmore

The AI Supercycle and Bottlenecks: A Multi-year Tailwind for EM

By Gustavo Medeiros and Ben Underhill



As demand for artificial intelligence (AI) compute scales rapidly, concerns that hyperscalers may be 'overbuilding' have receded. Anthropic and OpenAI have taken revenue into the tens of billions far faster than any company in history. Broader adoption of agentic AI is now expected to drive the next leg of exponential demand growth. NVIDIA's CEO Jensen Huang has recently argued agentic AI could soon drive a 10x increase in compute demand relative to large language models (LLMs).

So, rather than fretting that hallucinations might make the technology unusable, the market is now laser focused on supply bottlenecks in the race to build enough compute to meet exploding demand. So far this year, the companies positioned within these bottlenecks have driven most equity market gains, none more so Korean memory chip producers.

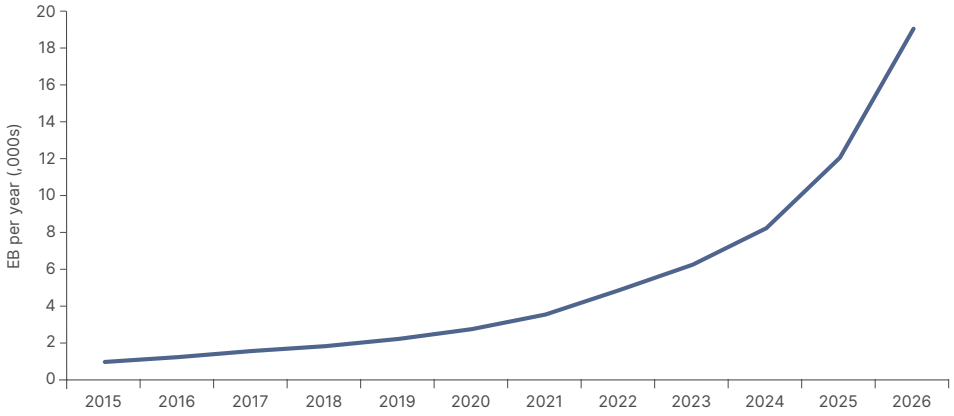
The question increasingly coming into focus is whether supply bottlenecks around energy and chips will soon force a slowdown in the investment cycle. In this month's Emerging View we chart these bottlenecks and explore how they might affect the trajectory of capex, and the implications for emerging market assets.

1. The Semiconductor Supercycle: A key tailwind for EM

A massive, global investment cycle is well and truly underway. After the Strait of Hormuz situation normalises, there will likely be substantial increases in investment across defence, energy security, and supply chain resilience. But the backbone of the investment cycle of our time is based on a transformational technological breakthrough – truly powerful AI. Over the last 15-years there has already been a very sharp increase in computing data consumption. This has been facilitated by data centres built by hyperscalers. But since OpenAI launched ChatGPT in 2022, the graph has gone vertical. More data centres, and more investment, must follow.

Since ChatGPT, the demand for compute has gone vertical.

Fig 1: **Global IP traffic in exabytes per year**



Source: Cisco VNI, Statista, IDC

AI is the largest capex cycle since 19th century railways.

The AI capex cycle is already the largest investment cycle in history, except for railways in the late-19th century. Shares of semiconductors are dominating market caps of global indices. The IPOs of Space X, Anthropic, and Open AI this year are likely to be historic – among the largest ever. The combined market capitalisation of these three could soon approach USD 4trn, sucking more liquidity toward the AI sector.

Seasoned investors sense euphoria and are cautious about jumping on the bandwagon. However, there is still a large opportunity afoot. The massive capex from hyperscalers is leading to one of the largest transfers of capital from developed markets (DM) to EM companies in history. The closest comparison would be the commodity-driven 2002-2008 EM bull market cycle, but the scale is bigger. The cycle is also likely to be longer than the one experienced post China's World Trade Organisation (WTO) accession, in our view. This is due to both supply and demand factors. Supply constraints will elongate project timelines while ongoing innovation will keep spurring on new investment.

2. The AI capex landscape

Today, there are about 100 gigawatts (GW) of data centres globally.[1] Projections from banks and sell-side research predict incremental data centre demand of 100-200GW by 2030, mostly driven by AI. McKinsey estimates a range of 125GW to 205GW, a framework which we will use for our analysis.

Meeting this demand will require an exorbitant amount of investment. The cost of building a frontier level 1GW AI data centre in America is now in the ballpark of USD 50bn,[2] according to industry leaders.

[1] See – https://www.jll.com/en-us/insights/market-outlook/data-center-outlook
[2] Midpoint of estimates, in line with Nvidia CEO comments.

Ashmore

The AI capex landscape

Estimates suggest that frontier AI data centres now cost around USD 50bn per GW.

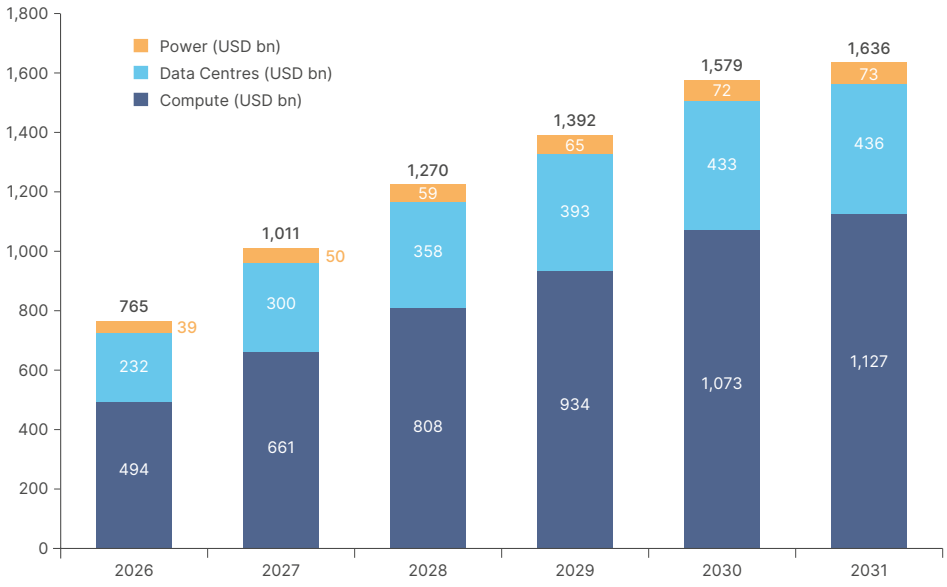
Goldman Sachs sees USD 6trn of cumulative capex by 2030.

Back-of-the-envelope framework for AI data centre cost

- **USD 15bn/GW** for the data centre shell, power delivery and site infrastructure.
- **USD 5bn/GW** for server-floor fit-out, including liquid cooling, CDUs and high-density electrical systems.
- **USD 30bn/GW** for GPUs, CPUs, networking, memory and other compute hardware.
  - Within this, memory could represent roughly **half of the chip bill**, equivalent to around **30% of total AI data centre capex** in 2026, consistent with SemiAnalysis estimates.
- **USD 50bn/GW** in total, with a broad sensitivity range of **±20%**.

This is already much higher than just two years ago. The evident constraints across AI data centre inputs suggest these costs are likely to keep rising. This implies that data centre capex will have to be at least USD 5trn over the next five years for demand assumptions to be met. Goldman Sachs model USD 6trn out to 2030 and USD 7.6trn out to 2031 as per Fig 2. Other experts give more aggressive forecasts. Nvidia gave a forecast in its Q1 earnings call of an annual run rate of USD 4trn by 2030.

Fig 2: **AI Capex estimates by 2030** (Goldman Sachs) **– USD billion**



Source: Goldman Sachs Global Institute, Goldman Sachs Global Investment Research NVIDIA projections (as at 3 March 2026).

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

Global AI capex remains US driven, led by the 'Big Five' hyperscalers.[3] Big Five capex is projected to reach c. USD 725bn this year, of which around 75% (circa. USD 540bn) is AI-related. Roughly two-thirds of the spend is currently marked for US projects, with the remainder split across Europe, Asia and increasingly the Gulf Cooperation Council (GCC).

Beyond the Big Five, the 'neoclouds' (CoreWeave, Nebius, and others) are expected to spend close to USD 85bn collectively on AI infrastructure in 2026, with CoreWeave alone guiding to over USD 30bn and Nebius USD 16-20bn. Neoclouds are US-centric (around 85%) but several have meaningful European footprints.

Chinese hyperscalers – Alibaba, Tencent, ByteDance, Baidu – are restricted by chip access, but currently still invest another USD 50-60bn on AI infrastructure annually. Other sovereign AI commitments add roughly USD 80-100bn per year, dominated so far by the GCC (USD 1trn+ committed over the decade between UAE, Saudi, and Qatar). Last month, the

[3] Google, Amazon, Microsoft, Meta and Oracle.

The AI capex landscape

UAE announced it would reduce the number of public servants by 40%, in a push to replace bureaucratic roles with AI. While extreme, this is a strong leading indicator for the paths other countries may take. The UK's NHS contract with Palantir is another example.

Put together, global AI capex in 2026 is likely to reach around USD 780-820bn, of which the US accounts for roughly 60%. Although much of this flows back to Nvidia, it is still the cleanest DM-to-EM wealth transfer in modern history. The hyperscalers do not have a Western supplier of any scale for the silicon, the memory, or the assembled systems. They must buy from EM.

3. Picks-and-shovels: From DM to EM

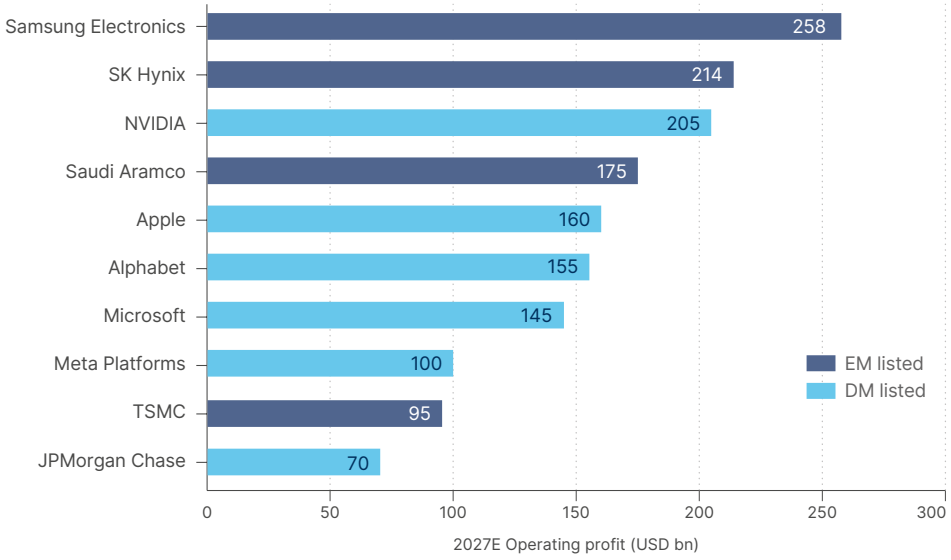
The first-order beneficiaries from the capex boom are clearly the prime contractors.

- Taiwan Semiconductor Manufacturing Corporation (TSMC) delivered FY25 revenue of USD 122bn (+35% yoy) and net income of USD 55bn (+40% yoy), with 2026 capex guided up 25% to USD 52-56bn.
- In Korea, Nomura sees Samsung 2027 operating profit at KRW 322trn (c. USD 258bn) – which would make Samsung the most profitable listed company in the world, with SK Hynix not far behind at KRW 267trn (~USD 214bn).
- Memory operating margins of 73-80% comfortably exceed Nvidia (65%), Apple (35%) and Google (32%).

AI capex is the cleanest DM-to-EM wealth transfer in history.

The result is that, based on 2027 forecasts, four of the world's 10 most profitable companies are EM-listed: Samsung #1, Hynix #2, Aramco #4, TSMC #9. That is a reconfiguration of where global profit pools sit, as per Fig 3. It has happened in just three years.

Fig 3: **World's most profitable companies in 2027(E)**



The world' most profitable companies in 2027 will sit in EM.

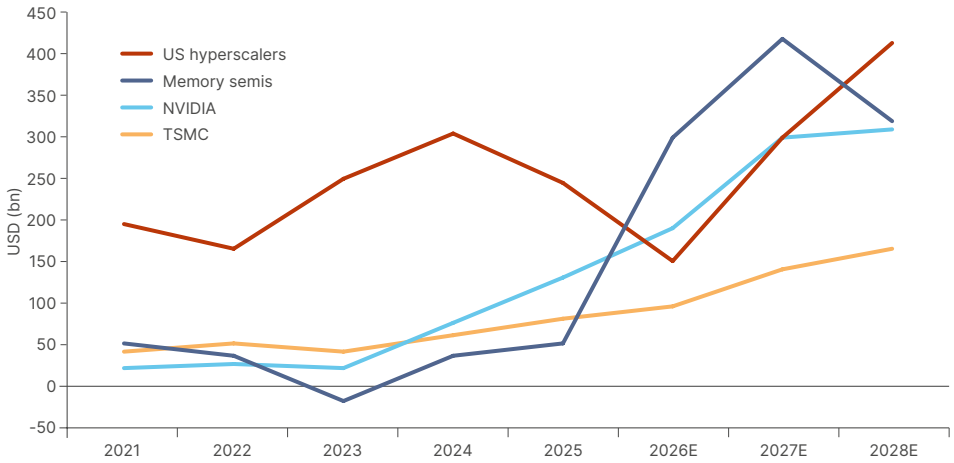
Source: Nomura (Samsung & SK Hynix 2027E via BigGo February 2026); company filings, consensus 2027E for other names; Fortune Global 500 for Aramco. Four of the world's top ten profit pools sit in EM.

Three of the top 10 most profitable companies are the largest investors in the AI race. (Alphabet, Microsoft, and Meta), which alongside Amazon, and Oracle makes up the five hyperscalers. The enormous capex of Alphabet, Microsoft and Meta means their free cash flow will be much leaner in both 2026 and 2027. Instead, memory companies are the ones enjoying the supernova profits. As memory emerges as a key AI bottleneck, memory companies could end up surpassing other parts of the AI supply chain in free cash flow earned.

Picks-and-shovels: From DM to EM

Memory cash flow may overtake the hyperscalers next year.

Fig 4: **Memory free cash flow (mostly Korea) to surpass hyperscaler free cash flow in 2026 and 2027**



Source: Bloomberg, Ashmore. Data as at May 2026.

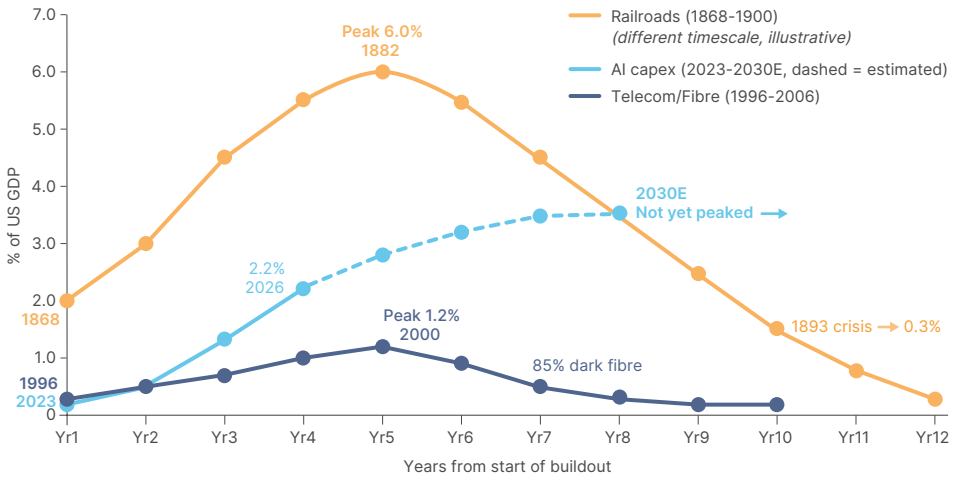
a. Comparisons to previous capex buildouts

Both Goldman Sachs and Morgan Stanley expect global AI capex to top USD 1.0trn in 2027, close to 1% of global GDP. In the US, estimated AI spending of USD 700bn in 2026 would breach 2% of GDP. A meaningful percentage of this spending reflects imports from EM, so is not GDP accretive. Nevertheless, AI capex remains the engine behind US GDP and corporate earnings growth.

Investment cycles of this magnitude are rare. The two capex cycles most often compared with AI infrastructure are railroads in the late 20th century and fibreoptics during the dotcom boom of the late 1990s and early 2000s. In both cases, capex peaked as ballooning leverage and overbuild ultimately caused the profitability outlook to collapse.

Unlike fibre and rail, the AI capex boom is not debt-fuelled.

Fig 5: **Infrastructure capex booms: AI (2023-?) vs fibreoptics (1996-2006) vs railways (1868-1900)**



Source: Ashmore. Data as at May 2026.

"This time may be different" is of the most dangerous phrases in investing. But there are key caveats around the 2000 comparison. While the telecom buildout was highly levered, feeding a big boom and a big bust, AI infrastructure spending remains buttressed by enormous hyperscaler cashflows. Leverage is beginning to rise, particularly for pure-play cloud providers like Oracle. However, the net debt/EBITDA ratios across the hyperscalers remain low or still negative. Although not every project will succeed, this basis makes a classic leverage-induced bust less likely.

Picks-and-shovels: From DM to EM

One company alone can manufacture frontier AI chips: TSMC.

Perhaps the key reason, in our view, that AI is unlikely to resemble prior capex driven bubbles is the tightness of the bottleneck around leading edge chip production. This all comes back to the fact that there currently is only one company capable of manufacturing the most advanced AI chips at scale: TSMC. Unless the current demand curve for compute slows down meaningfully, it is unlikely TSMC will or indeed can expand fast enough to overrun demand. This does not mean we will not see a bubble form in financial markets. But it does make it less likely that this will lead to overcapacity in the real economy, analogous to the dotcom bubble or the railroad bubble.

4. Supply vs demand: A rolling question

In late 2025, concerns grew that AI and data centre investment was overzealous. Meta CEO Mark Zuckerberg's comment that the risk of being left behind far outweighs the risk of "misspending a couple of hundred billion dollars" captured the zeitgeist. But in early 2026, the release of Claude Opus 4.6 flipped the narrative. Models were suddenly good enough to shift future demand curves sharply higher.

Particularly for AI sceptics, this was a watershed moment. The ease with which even laypeople could now 'vibe code' apps triggered a broad software sell-off, as moats looked suddenly narrow. Meanwhile, Anthropic's revenue scaling from USD 1bn to USD 40bn annual recurring revenue (ARR) in just over a year demonstrated the monetisation potential that had previously been in doubt.

Narrative shifts will continue as the technology evolves. One key variable is efficiency. Future models may compound efficiency gains atop hardware improvements, leading to significantly more output per data centre. However, to what extent efficiency gains can reduce bottlenecks is not modellable. History shows that efficiency gains in technology tend to raise demand, not lower it. This rule is known as Jevons' Paradox, first observed when greater steam train efficiency led to greater coal demand by enabling larger rail networks. Each train journey used less coal, but the price of coal went up.

Jevons Paradox = efficiency gains raise demand.

Jevons' Paradox suggests compute demand risks are skewed to the upside. The leading AI CEOs: Demis Hassabis (Google DeepMind), Sam Altman (OpenAI), Dario Amodei (Anthropic), and Elon Musk (SpaceX) all flag a structural compute shortage ahead. Their bias is obvious, but so is their informational edge. So far, investors would have done well to believe them.

5. AI Investment has vastly exceeded expectations. Can this continue?

"USD 1trn per year of total annual AI investment by 2027 seems outrageous... dramatic... one of the very largest capital buildouts ever," wrote German AI researcher and investor Leopold Aschenbrenner in 2024. What seemed an outrageous forecast then for a famous AI bull is now the base case. Goldman Sachs and Morgan Stanley both see roughly USD 1trn of AI capex commitments in 2027 from the hyperscalers alone – equivalent to around 20GW of capacity at USD 50bn/GW, likely online by 2028-29.

The question is where we go from there. AI leaders speak as though no amount of capital is too much to pave the road that will fast-track society into the future. Sam Altman wants OpenAI alone scaling at 1GW per week by 2030. But ultimately investment cycles must generate an acceptable return, or they end.

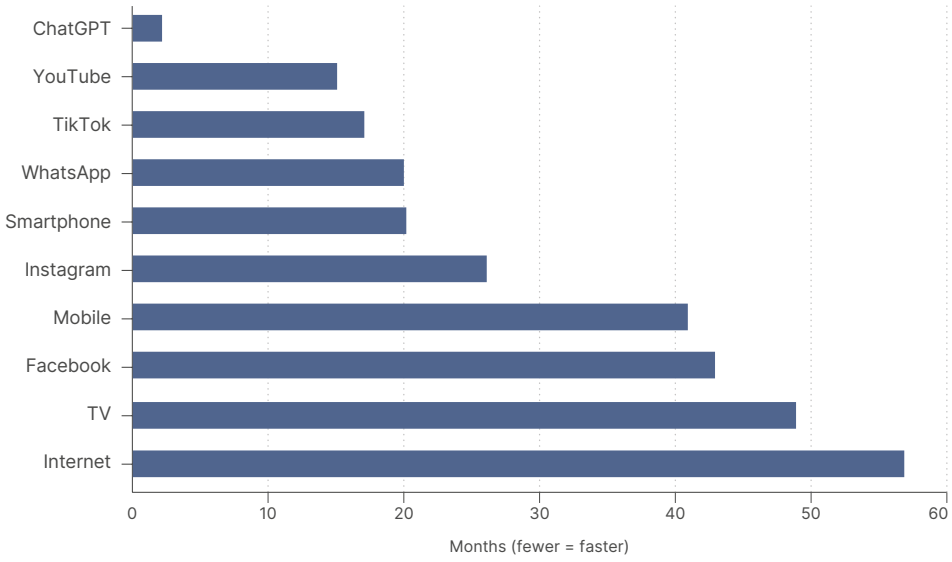
For now, adoption and revenue are still moving fast enough to justify further investment. LLMs have been the fastest major technology adoption cycle on record by far, even adjusting for today's larger population base. Across all models, cumulative AI revenue

AI Investment has vastly exceeded expectations. Can this continue?

stands at between USD 80bn-100bn today, from zero just three years ago. This revenue trajectory has no precedent in the history of capitalism. AI also enhances hyperscaler core businesses: ad revenue at Meta and Google, conversion rates at Amazon.

AI is the fastest technology adoption ever recorded.

Fig 6: **Population-adjusted months to 100 million users**



Source: Ashmore. Data as at May 2026.

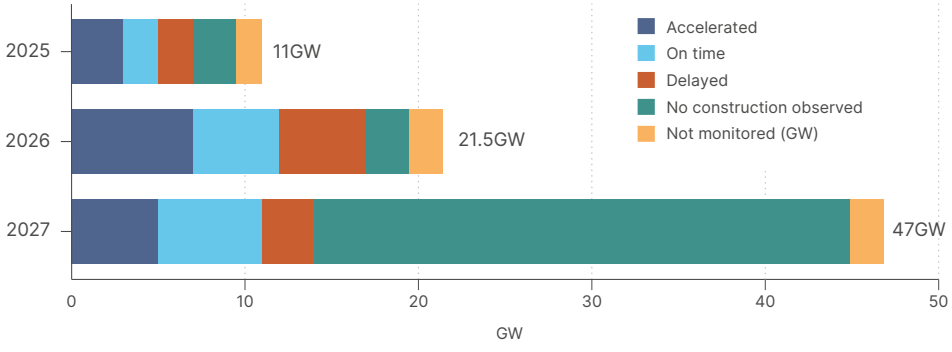
However, the scale of investment necessitates that revenue continues expanding at a rapid clip. If not, capex is likely to be rationalised. As a rough roadmap, using a simplified blended asset life of around eight years, and a mid-to-high single-digit cost of capital, a USD 6trn installed base would likely need to generate roughly USD 900bn-1.2trn of annual operating cash flow by 2030 to be economically justified.

Time will tell whether this will happen. But for now, capital does not look like the binding constraint. Physical supply does. In our view, tight bottlenecks in semiconductors and energy will make sustained scaling of annual investment beyond 2026 vdifficult. Economic incentives to clear the bottlenecks are enormous, and margins across the AI infrastructure stack are high and expanding. But the complexity of the products in shortage, particularly chips, means production lead times are long.

Supply constraints are already extending project timelines. Satellite imagery shows only half of the 23GW of US data centre capacity scheduled for this year is on time. In 2027, less than a quarter of projected capacity is on track. With 12- to 18-month construction timelines the norm, many 2027 deliveries now look unlikely.

Only half of 2025 data centre capacity is on schedule.

Fig 7: **US data centre capacity (GW), by status and completion year**



Source: Syntax Vulcan Platform, IRR Energy. Data as at April 2026.

AI Investment has vastly exceeded expectations. Can this continue?

Constrained supply against strong demand are currently elongating the investment cycle, rather than derailing it. But more delays are already becoming cancellations. Trellis tracks a rise from two cancelled data centre projects in 2023 to six in 2024 and 25 in 2025.[4] Most 2025 cancellations were projects announced in 2022-23, a potential leading indicator for the larger 2024-25 cohort currently in 'delayed' status.

6. Physical constraints: Chips (global constraint)

Frontier AI capacity depends on a narrow global semiconductor supply chain – bottlenecked not just at TSMC but at multiple upstream points, particularly ASML's Extreme Ultraviolet (EUV) lithography machines. Since the chain ultimately funnels through TSMC's "Chip on Wafer on Substrate" (CoWoS) packaging, any upstream node can become the binding constraint at any time, making long-term chip supply growth hard to forecast.

Chip supply has historically scaled to cyclical consumer electronics demand, producing notorious booms and busts – a history that instils caution in producers. TSMC, SK Hynix and Samsung are run by managers bruised by past overinvestment and loathe to make the same mistakes again. But the AI era is shifting the nature of semiconductor cyclicality. Leading-edge chip demand is now driven overwhelmingly by Nvidia and AI accelerators, not consumers buying iPhones. Agentic AI sharpens the picture further. Agents can now be trained to specific requirements, scaled exponentially, and run compute-intensive tasks requiring constant high-bandwidth memory (HBM) around the clock. Supply caution against vertical demand growth leads many analysts to conclude we may be entering a multi-year chip deficit.

Each new generation of AI chips are harder to build.

Dylan Patel of SemiAnalysis estimates 1GW of Nvidia Vera Rubin data centre capacity requires roughly 55k 3nm wafers for GPU logic, 6k 5nm wafers for central processing units (CPU) and supporting logic, and 170k dynamic random-access memory (DRAM) wafers across the memory stack, including HBM. We use this framework to illustrate how many GW of current frontier infrastructure chip supply *could* facilitate. Obviously, not all data centres from now on will use 100% Vera Rubin – far from it. Nevertheless, Vera Rubin is the best expression of the kind of architecture that increasingly defines the constraint.

a. TSMC: Leading-edge logic

NVIDIA, AMD and most custom AI application-specific integrated circuit (ASIC) designers rely entirely on TSMC for leading-edge manufacturing and chip-on-wafer-on-substrate (CoWoS) packaging. TSMC's 3nm monthly capacity is reportedly rising from ~150k to ~180k wafers by end-2026 – a 40% year-on-year increase – softening the leading-edge bottleneck for now, even as capacity remains fully allocated.

AI's share of TSMC's 3nm mix is also expanding from ~5% in 2025 to ~36% in 2026, or roughly 792k chips this year. Using Patel's rule of thumb, that enables approximately 14.4GW of Vera Rubin. As it turns out, this is a plausible estimate for what might come online globally in 2026, in the same ballpark as various forecasts.

Note that only a small fraction of new chips will refresh legacy facilities, since most older data centres lack the cooling and power infrastructure for newer Nvidia architectures. But as the installed base ages, replacement rates will rise, tightening supply available for new builds.

Three key variables determine how much new capacity leading-edge logic can enable: TSMC's 2nm/3nm supply growth, AI's allocated share, and the replacement rate of the existing AI compute base. Using these three inputs, a crude benchmark for annual data centre enablement can be expressed as:

Gross GW enabled = (TSMC 3nm/2nm wafer supply x AI allocation share) / 55k

Net new GW enabled = Gross GW enabled x (1 – replacement drag)

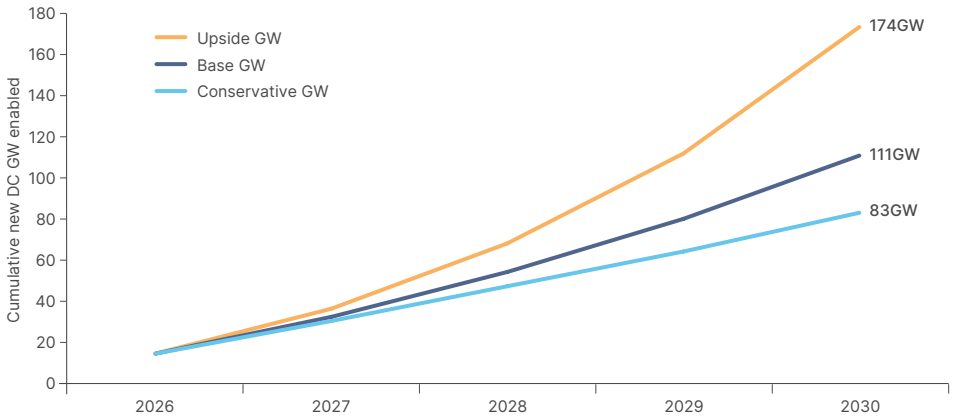
[4] See – https://trellis.net/article/data-centers-ground-zero-ai-backlash-what-must-change/

Physical Constraints: Chips (Global Constraint)

Even unprecedented semiconductor supply growth may fall short of demand.

Scenarios with conservative, base and upside cases, varying wafer supply combined annual growth rate (CAGR) (10%, 15%, 25%) AI allocation by 2030: 45%, 55%, 70%, and the replacement drag from refreshing the existing compute base by 2030: 10%, 20%, 30%, produce annual data centre capacity that may struggle to meet even the lower end of McKinsey's demand growth projection (125GW).

Fig 8: **Illustrative net new data centre capacity enabled by leading-edge logic supply**



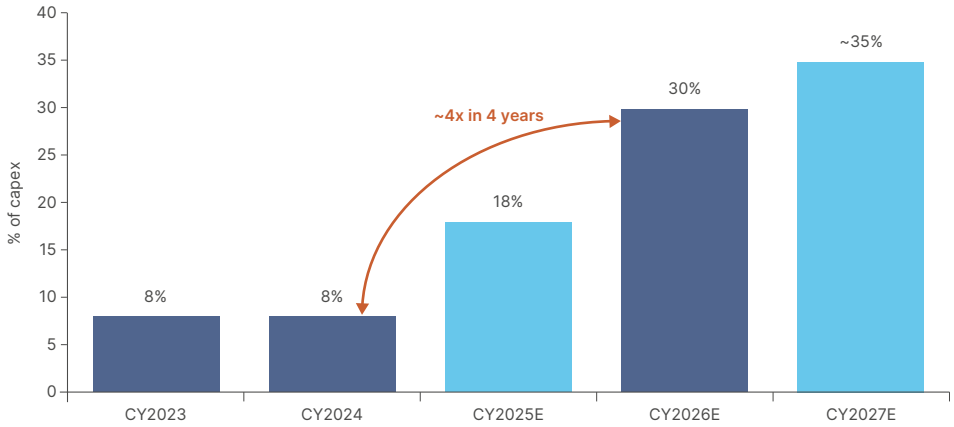
Source: Ashmore estimates. Uses 55k 3nm/2nm wafers per GW of Rubin-class capacity. AI allocation and replacement drag scale linearly from anchors of 36%/0%.

b. Memory

As AI workloads shift further towards inference and agentic use cases, demand is rising across the full memory stack: HBM, DRAM and NAND. Supply is tight across all three. This explains why memory has moved from a relatively small percentage of hyperscalers' capex, to a major one as per Fig 9.

Memory chips: from 8% to 35% of AI capex in four years.

Fig 9: **Memory as a share of capex**



Source: SemiAnalysis via Tom's Hardware.
https://www.tomshardware.com/tech-industry/memory-will-consume-30-percent-of-hyperscaler-spending-this-year

The more specialised product – frontier HBM stacks – are likely to remain the binding constraint. HBM is made from DRAM, but it is not directly fungible with it. Conventional DRAM is closer to a commodity. In periods of shortage, higher prices draw in supply and destroy demand elsewhere. AI customers can therefore obtain more conventional DRAM by outbidding phones, PCs and standard servers. HBM is different in that it clears through allocation. It must be manufactured, stacked, bonded, tested and qualified for specific accelerator platforms. It is to commodity DRAM what jet fuel is to crude oil: a specialised derivative that can absorb upstream capacity even as the downstream bottleneck remains tight.

Only SK Hynix, Samsung and Micron are delivering leading-edge HBM at scale. Together, they form a supply-disciplined oligopoly, with their philosophy shaped by repeated cycles

Physical Constraints: Chips (Global Constraint)

of overcapacity and margin destruction. SK Hynix leads with 57% of HBM market share in Q3 2025, followed by Samsung at 22% and Micron at 21%. China is trying to catch up but remains structurally behind. ChangXin Memory Technologies (CXMT) is not expected to launch HBM3E until 2027, by which point the leading producers will be making HBM4.[5]

Scaling HBM chip production is far harder than DRAM.

Global HBM capacity is sold out through 2026 and heavily pre-committed through 2027. SK Hynix forecasts ~30% annual demand growth through 2030. Supply is expanding, but not at that pace. SK Hynix's HBM-allocated wafers may grow 2.5-3x by 2030 against demand growing nearly 4x on its own guidance. Its M15X fab ramps to 60k wafers/month by mid-2027; the larger Yongin cluster adds ~350k wafers/month at full output between 2028-30.

Samsung targets a ~50% capacity increase in 2026; Micron has announced high-volume HBM4 production for Vera Rubin. Both companies may gain market share from SK Hynix, but blended supply growth across all three producers is unlikely to keep pace. This will force further HBM cannibalisation of standard DRAM and therefore also higher DDR5 prices until demand destruction bites in PCs, smartphones and traditional servers. This points to higher-for-longer memory pricing, structural rent capture by the HBM oligopoly and continued allocation-based market clearing until supply, qualification or memory efficiency decisively shifts.

c. ASML: Extreme Ultraviolet Machines

Every advanced chip funnels through two chokepoints.

ASML is the sole global supplier of EUV lithography tools – the machines required to manufacture the most advanced logic and memory chips and widely regarded as the most complex machinery ever built. ASML is probably not the acute bottleneck today, but if AI demand compounds anywhere near frontier labs' ambitions, the constraint could shift upstream to ASML by 2028-29.

d. Chip on Wafer on Substrate (CoWoS)

CoWoS is the final assembly step connecting GPU/ASIC dies with HBM on a silicon interposer, enabling the memory bandwidth required for training and inference. TSMC holds a near-monopoly on this process. CoWoS was a visible bottleneck in 2023-24 but TSMC has since expanded capacity from ~35k wafers/month in late 2024 to a planned 120-140k by end-2026. Even so, TrendForce reports TSMC's CoWoS lines remain fully booked for the next two years across NVIDIA, Google, Amazon and others.[6]

e. The Life of a Chip – crucial variable

Due to compute demand, older GPUs are rising in price, not falling.

Nvidia releases new GPU architectures annually with step-change performance gains, but intense demand against constrained cutting-edge supply creates strong incentives to extend the life of older chips. Newer chips dominate training, but inference runs well on legacy hardware, particularly for simpler tasks.

As a result, prices of legacy GPUs like Nvidia H-100 (2024) are now rising, not falling as dated hardware should. Anecdotally, there are numerous three to four-year-old Nvidia Hopper (2022) clusters signing multi-year extensions. This indicates the useful life of chips may well be closer to seven to eight years, improving the economics of the overall capex materially. A longer lifespan also means a lower replacement rate, which means more capacity free to bring on new GWs of capacity.

[5] See – https://www.scmp.com/business/china-business/article/3354223/why-chinese-dram-maker-cxmts-ipo-attracting-so-much-attention
[6] See – https://www.trendforce.com/news/2025/12/08/news-tsmcs-cowos-l-s-reportedly-fully-booked-osat-partners-step-up-with-ases-cowop-in-focus/

7. The EM semiconductor supply chain

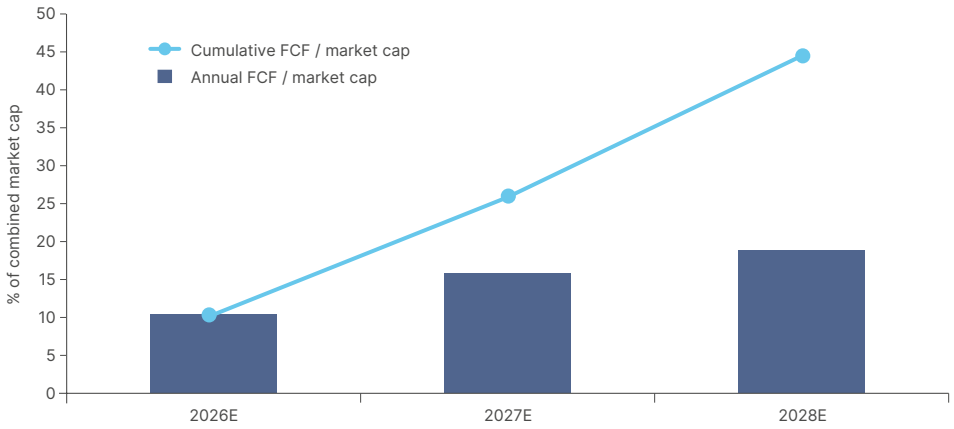
Asia dominates semiconductor supply chains, not the West.

The constraints in TSMC and memory are not new and have already been reflected in asset prices and company results. Seasoned investors would point out that cyclical businesses are usually bought at low price-to-book (PB) and high price-to-earnings (PE). Today, the opposite is true: these companies trade at low PE, given supernova earnings, but high PB.

The challenge with this valuation analysis is that book value is a poor anchor when demand for the product has shifted structurally. On a discounted cash flow basis, SK Hynix and Samsung look much less expensive than headline PB multiples imply. Consensus free cash flow over the next three years represents a large share of current market capitalisation before assigning any terminal value to structurally higher HBM demand, as per Fig 10. Unless new entrants take market share and close the supply-demand gap, which is unrealistic for 2026-27, prices are likely to remain elevated.

Korean memory chip producers are still cheap in cash flow terms.

Fig 10: **SK Hynix and Samsung FCF/market cap (2024-2028)**



Source: MarketScreener consensus FCF estimates and current share counts/prices.
Market cap calculated using 26 May 2026 prices.

As more value accrues to semiconductor bottleneck companies in EM, supply chains across Asia benefit disproportionately. Hon Hai Precision Industry, Wistron and Quanta Computer assemble the L10 racks in Taiwan, with manufacturing footprints across China, Mexico and Vietnam. Delta Electronics dominates power supplies, thermal management and liquid cooling – the underrated, but mission-critical kit that turns silicon into a functioning AI data centre, with operations spanning Taiwan, Thailand and Vietnam. MediaTek and other ASIC partners design the custom chips. Substrate, printed circuit boards, copper clad laminates and connectors come from Taiwan and Korean specialists. Outsourced semiconductor assembly and testing is concentrated with ASE Technology Holding in Taiwan, plus the buildout in Malaysia, the Philippines and Vietnam. The Western contribution centres on the GPU/ASIC design, including Nvidia, AMD, and in-house silicon teams and the data centre shell and connectivity – a minority share of the bill.

8. Physical (US) constraints: Energy

The second constraint on AI expansion is energy availability. This bottleneck is particularly acute in the US, where power availability is already delaying data centre projects. The constraint is considered severe enough over the longer term that Elon Musk and others have floated siting data centres in orbit for the energy and cooling advantages.

Bloom Energy survey data show hyperscalers and colocation providers consistently expect power up to two years earlier than utilities and IPPs believe they can deliver it. PJM Interconnection, the largest US grid operator and the system covering Virginia, the world's largest data centre market, is the clearest example. The average time from interconnection

Physical (US) constraints: Energy

application to commercial operation has risen from under two years in 2008 to more than eight years in 2025. PJM connected only 3.0GW of new generation across all uses in 2025, against more than 30GW of incremental data centre demand projected through 2030.

The lack of available power in PJM is pushing more data centre development towards Texas and Georgia, where near-term prospects appear better. Texas, however, is also tight. ERCOT peak demand can reach around 85GW, against roughly 89GW of effective evening peak supply (ERCOT CDR), leaving little spare margin. ERCOT is adding capacity quickly, but not all capacity is equally useful for new data centre connections. Recent additions have run at 12 to 15GW per year of nameplate, but around 77% of 2024 to 2025 additions were solar and storage, while net new gas was only around 1.5GW. This matters because grid space is allocated against effective supply at system stress, not average annual generation. Peak stress tends to fall on summer evenings, when solar output is weak: ERCOT credits its combined wind and solar fleet with effective load carrying capability of just 6.2GW at the evening peak, against 73GW of nameplate capacity.

Surplus US energy supply across peak hours is very thin.

Assumptions for US grid capacity available for data centres by 2030

- Rather than model each grid region from the bottom up, we anchor the grid contribution to external authorities, since even the most rigorous studies decline to give a single point estimate. The US Department of Energy's July 2025 Resource Adequacy Report finds that only around 22GW of new firm, dispatchable capacity is likely to be added nationally by 2030 (DOE, Evaluating the Reliability and Security of the United States Electric Grid, July 2025). This is a deliberately strict measure: it counts only firm dispatchable generation and has been criticised for excluding battery storage and assuming few additions after 2026. RAND's parallel 2025 study puts front-of-the-meter net available additions at 33GW on a fuller effective-capacity basis (Arciniegas Rueda et al., Assessing the United States' Additional Power Capacity by 2030, RAND, 2025). We take the midpoint of these two anchors, around 28GW, as the realistic addition of effective grid capacity by 2030.

Forecasts for grid capacity growth vary widely.

- Data centres are the single largest driver of incremental US electricity demand and, critically, the most able to capture scarce new capacity: developers can pay premium rates and move faster than diffuse residential electrification or industrial load in the inter-connection queue, and they actively target the regions and substations where headroom exists. We therefore assume data centres capture most of these effective additions, around 20 GW, with the balance absorbed by electrification and industrial growth.
- This as the conservative end of the range. The same RAND study reaches a headline 82GW of net available additions, but only by adding 49GW of behind-the-meter capacity that we do not credit for data centre purposes. That 49GW is overwhelmingly customer-sited residential solar and storage that reduces household peak demand rather than generation deliverable to data centres..

Behind the Meter (BTM) Solutions

Today there are effectively no large, fully off-grid AI data centres operating at scale in the US. But the severity of the grid constraint has driven most US data centre projects toward generating power "behind the meter" to bypass the grid partially or entirely. The most viable source of BTM power is natural gas, which is abundant and cheap in the US.

Over the longer term, BTM power has the potential to meaningfully ease the bottleneck. But for now it shifts the bottleneck rather than solving it, because the power equipment itself is supply constrained. The clearest example is gas turbines. GE Vernova, Siemens Energy and Mitsubishi, who together control roughly 90% of the global market, are booked through 2028 to 2030, and new orders placed today may not be delivered for three to four years. Because the data-centre-relevant share of these order books is only around a fifth to a quarter of the total, and because the fast-deploy aeroderivative units that developers

Physical (US) constraints: Energy

New gas turbines are sold out three to four years.

most want are physically small, the gigawattage that can actually reach US data centres by 2030 is limited even as order books look enormous. We assume new-build gas turbines, aeroderivative and large-frame combined, add around 13GW cumulatively to US data centre capacity by 2030.

With new turbines backlogged, developers are also turning to faster or smaller alternatives: refurbished and repurposed units, reciprocating gas engines, and portable generators. Reciprocating engines in particular are moving from backup duty to primary power, with Wärtsilä alone having booked over 1.6GW of US data centre engine capacity. These help individual projects move sooner but are smaller in scale and less efficient.

Solid oxide fuel cells (Bloom Energy) are one way around the gas turbine bottleneck. The technology converts natural gas or hydrogen directly to electricity through electrochemical oxidation rather than combustion spinning a turbine. The same technology was originally developed to convert Martian atmospheric gases into oxygen for propulsion and life support. Lead times from order to delivery are short, around six months. But Bloom cannot fill the gap alone in the coming years. Production is scaling quickly, but from a base of around 1GW per year, and is not expected to exceed 3GW per year by 2030.

BTM power doesn't solve the bottleneck, it shifts it.

Nuclear reactors, including small modular reactors (SMRs), are another potential long-term solution but are unlikely to move the needle before 2030. An illustrative base case for the possible availability of onsite power by 2030 is shown in Fig 11, based on company reporting and synthesised industry expert commentary. We assume annual capacity scales linearly between 2026 and 2030.

Fig 11: **BTM solutions total availability by 2030 estimates** (optimistic base case)

BTM pathway	2026 (GW/yr)	2030 (GW/yr)	Cumulative by 2030 (GW)	Key constraint / driver
Gas turbines (new-build: aero + frame)	1.5	3.5	13.0	Aero (LM2500/LM6000) for fast bridge/islanded BTM; heavy-duty frames in combined cycle for large co-located builds (GEV JVs with Chevron, NRG, NextEra). All three OEMs booked through 2028-30. DC share of orders ~20-25% blended.
Refurb/salvaged/relocated turbines			5.0	One-off, not sustained flow. Repurposed aero/turbofan cores; finite airframe pool.
Gas turbines (reciprocating)	0.5	1.2	4.0	12-18 month lead. 5-20MW units. Caterpillar, Wartsila (1.6GW+ booked), MAN ES, INNIO. Moving from backup to primary.
Fuel cells (Bloom)	2.0	2.5	9.0	Fremont ~1GW/yr to 2GW/yr end-2026. 3rd plant unannounced caps >3GW/yr. Oracle/Brookfield partly international; US-DC share ~60%.
Fuel cells (non-Bloom)	0.0	0.2	0.5	Not competitive at DC scale.
Solar + storage (8hr+)	0.2	0.7	2.0	8hr+ duration not at scale. Form Energy 2028+. SW/TX only.
Diesel gensets (as primary)	0.2	0.3	1.0	Emissions/permitting limits. Stopgap while waiting for gas.
Total US BTM (base case)	**4.4**	**8.4**	**34.5**	

Source: Bloom Energy Data Center Survey, November 2025.

Fig 12 outlines estimations for US data centre demand growth from two sources against what is likely buildable based on the supply estimates above. Energy capacity will not come online in a straight line as the graph depicts. Nevertheless, the trajectory between supply and demand widens significantly from 2027. Jigar Shah, ex Director of Loans in the US department of Energy, says only 50GW extra capacity can come online over next five years,[7] which is close to our combined estimate of potential incremental power supply for data centres from the grid (20GW) and BTM (34.5 GW).

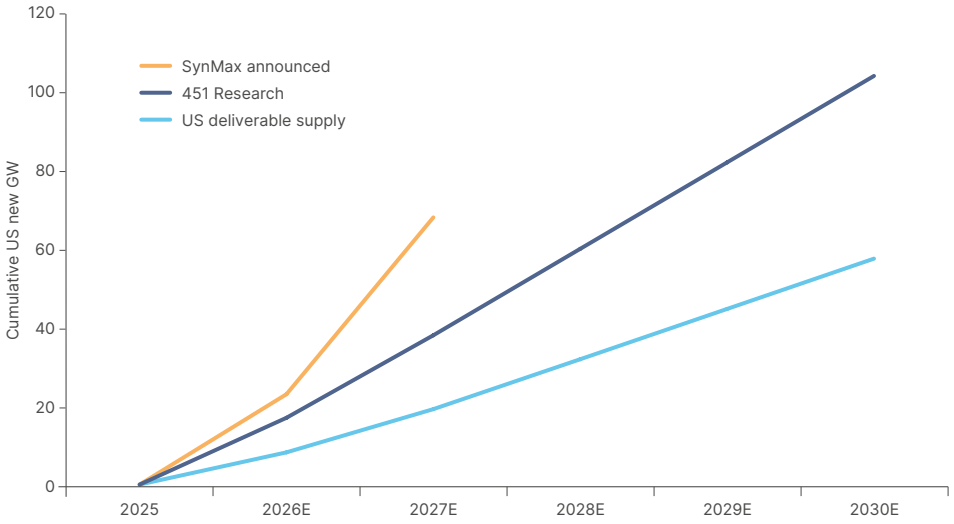
[7] Prof G transcript: https://pod.wave.co/podcast/prof-g-markets-2a0afe72-164a-4443-9f5a-558ae6db44f4/data-center-debate-are-energy-bills-about-to-explode

Ashmore

Physical (US) constraints: Energy

The energy supply-demand gap may widens sharply from 2027.

Fig 12: **US data centre demand against energy capacity**



Ashmore, 451 Research, Synmax. Data as at April 2026.

Power supply growth trends look weak...

55GW of incremental data centre supply in the US by 2030 undershoots most demand forecasts. 451 Research forecast 100 GW demand growth in the US between 2025 and 2030. Synmax show that nearly 80GW of projects were scheduled for completion between 2025 and 2027. It is currently difficult to see how many of these will come online on schedule, given the energy bottleneck.

9. Energy bottleneck release valves

Significant innovation needs to happen to avoid the energy constraint. History suggests it is unwise to bet against innovation. Rapidly expanding power generation or availability in the US for data centres is both a regulatory and engineering problem, but it is not unsolvable. Two key 'release valves' that could help ease the US energy bottleneck, in our view:

a. BTM scales at unprecedented rates

...but bet against innovation at your peril.

Fuel cell production can scale quicker than expected. The other potential solution involves gas turbines and rotational engines expanding supply at an unprecedented rate. The best and most efficient turbines made by companies such as GE Vernova are constrained and backlogged – but backlogs could of course clear faster than expected. Other non-optimal turbine solutions could also scale at an unprecedented pace. Solar powered battery technology is improving and could soon be another solution that can scale quickly as a supplement to grid or gas-based BTM power, particularly in the US solar belt.

BTM scaling can also take place outside of data centres, through residential/industrial grid clients own on-site solar and battery power. This can, in theory, free up extra firm capacity for the data centre usage.

b. US grid optimisation

Only 100-200 hours a year actually stress the US grid.

Power grids are planned around maximum load. Power outages or load shedding are not acceptable in America, so reserve margins are built around the tightest hours of the year. But in practice, there are only around 100-200 hours a year when grid load is close to maximum levels. These are typically summer evenings, when people return home from work and blast their air conditioning. Hence, the problem is not simply one of total energy availability, but of peak management.

Energy bottleneck release valves

The practical solution may be a hybrid model. Data centres that cannot get full grid access or build 100% BTM power could rely on the grid for most of the time and then use batteries, onsite generation and workload shifting during peak demand hours. Nvidia's Jensen Huang is an advocate for this model, publicly raising the possible solution numerous times. Market design is already heading in this direction, with Texas recently giving ERCOT the authority to disconnect large loads such as data centres during grid emergencies.

Further complications

Even where power is theoretically available, the equipment required to deliver it to the data centres is in severe shortage. The lead time for generation step-up transformers has tripled since 2022, to more than 150 weeks. Power transformers have followed a similar trajectory. Of all 47 categories in the US producer price index (PPI), transformers and power regulators have seen the second highest level of inflation since 2020, at roughly 80%. New transformer factories are being built in the US, but meaningful output is two to three years away. As of today, there is no software solution, no workaround, and no substitute for a physical transformer. A USD 100bn data centre campus can sit idle waiting on a USD 40m transformer order.

Other constraints include permitting, with data centres now more unpopular in some US states than the Iran War. The possibility of a Democrat-dominated Congress next year raises the risk of stricter regulation around data centre construction. With each generation of chips becoming increasingly powerful, cooling infrastructure is another key bottleneck. Despite Elon Musk's best efforts, humanoid robots cannot build data centres yet – so skilled construction labour is another.

Constraints vs Capex

Power supply may cap US data centre supply growth...

If innovation around energy generation capacity and unprecedented chip supply growth do not materialise the question becomes: how much capex can the physical economy absorb? A simple way to frame this is to convert buildable GW into deployable capex. As we have suggested, if the US only adds around 55GW of AI-suitable, effective data centre power by 2030, that implies around USD 2.75trn of US AI infrastructure capex can be deployed, at our USD 50bn per GW anchor. That is still enormous, but it means US capex commitments are unlikely to keep growing at the recent pace, given USD 1trn is now expected to be announced over 2026 and 2027.

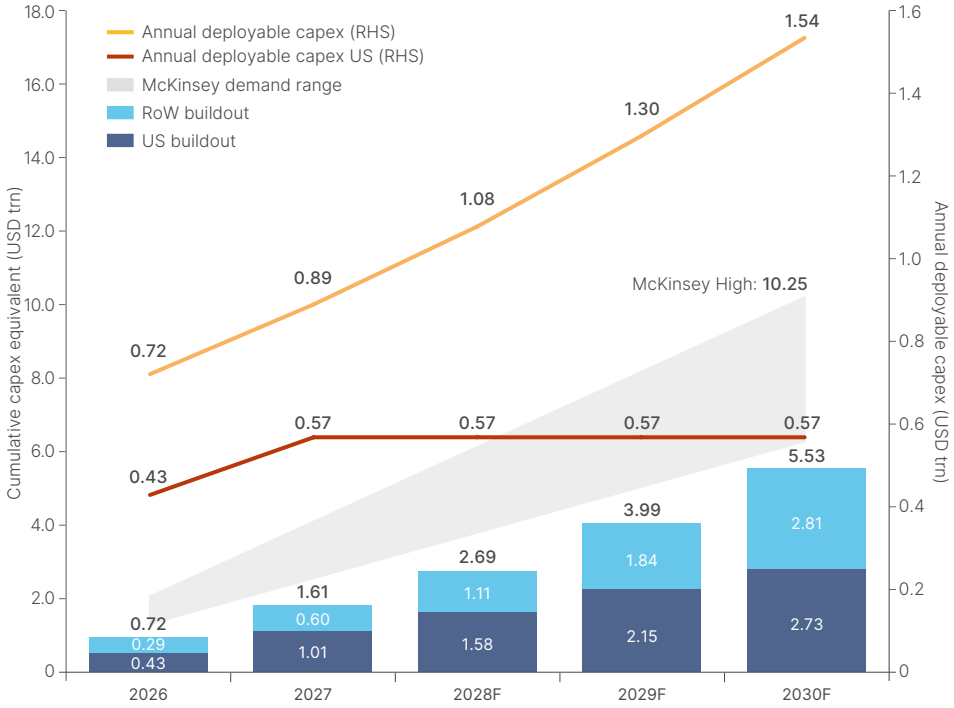
If US power is the constraint, power-rich regions such as Latin America, the GCC, Morocco, Scandinavia, and parts of Southeast Asia (i.e., Malaysia), can absorb some of the slack. We think more capex from hyperscalers and others will indeed flow to these regions. However, this doesn't solve the semiconductor constraint, where our base case of 111GW AI-enabling production through to 2030 undershoots the low end of McKinsey demand estimates (125GW). To put this into real world perspective, TSMC CEO said last November that TSMC advanced node-capacity falls 'about three times short' of current AI demand.

If physical constraints bind, the key question is whether projects are delayed, moved offshore, repriced or cancelled. The market's comfort with the constraint thesis rests on the assumption that most delayed projects are deferred rather than abandoned. A 12-month delay is a roll; a 36-month delay starts to look like a cancellation. Net present value erodes as GPU generations move on, competitors capture the revenue opportunity and CFOs reprioritise spending.

Energy bottleneck release valves

...but chips are a global constraint.

Fig 13: **Global and US commitment of data centres against power and semiconductor ceilings**



Source: Ashmore, 2026.

US 'annual deployable' capex in 2028/2029/2030 is a yearly average of the difference between what we model power enabling by 2030 and forecasts of US AI capex announcements across 2026 and 2027.

Our base case forecast for TSMC leading edge node growth available for new AI data centres leans optimistic. But to reiterate, CoWos packaging or any point upstream of it may bind chip supply growth before TSMC 3/2nm production does. We also acknowledge that our anchor of USD 50bn per GW of data centre is unlikely to be accurate over the longer term. If price inflation in semiconductors continues to surge, costs could run meaningfully higher per GW by 2030. How much costs can rise before data centre projects become uneconomical is not clear. It depends in large part on the profitability of AI. The magnitude of the many variables are evident. Fig 13 should therefore be seen as illustrative – what the trajectory of spending may look like against current costs and current bottlenecks.

10. Macro implications: Three paths for the AI capex cycle

Scenario 1: Boom continues

Boom: constraints ease, the cycle accelerates.

Constraints ease faster than expected. BTM power scales, grid optimisation works, CoWoS expands, HBM supply improves and non-US regions absorb more of the US power bottleneck. AI infrastructure investment remains well above consensus, and the cycle continues to accelerate. Demand for chips, memory, copper, power equipment, gas turbines, transformers and construction labour remains strong. EM continues to benefit, led by TSMC, Korean memory, Asian hardware, copper, energy and power-rich regions such as the GCC. Margins of bottleneck supplier may compress due to greater supply but profits likely to remain very high. The USD impact is mixed: strong US nominal growth supports the dollar, but a wider US external deficit and stronger EM terms of trade lean against it.

Macro implications: Three paths for the AI capex cycle

Base case: constraints slow US, EM captures the rents.

Scenario 2: Bottleneck-constrained cycle – base case

Demand remains strong, but physical constraints bite. US power availability and global chip supply prevent the buildout from scaling at the pace implied by announcements. Some projects move offshore, some are deferred, and a larger share of spend shifts toward chips, powered shells and refits rather than speculative greenfield campuses. Capex growth slows but does not collapse. Fewer GW are added than demand would justify, while bottleneck price inflation keeps nominal spend elevated. High prices persist across bottleneck suppliers.

This is the best relative setup for EM, in our view. The US bears more of the capex-growth slowdown, while EM captures the bottleneck rents. Korean memory, TSMC, Taiwanese ODMs, power equipment, copper and selected energy exporters remain supported. The USD should weaken at the margin if US growth slows while EM external balances improve. EM equities outperform, but the trade becomes selective: own the constraints, not generic AI beta.

Bear case: demand, not supply, is what disappoints.

Scenario 3: Demand-led rationalisation – bear case

The bear case is not that supply constraints bite, but that end demand disappoints. AI revenues fail to scale fast enough, commoditisation of LLMs (perhaps accelerated by Chinese models) competition compresses token pricing, regulation raises the cost of compute, or investors lose confidence in returns. Capex then slows because expected returns fall. Delays become cancellations, order books normalise and pricing power across bottleneck suppliers fades. North Asian semiconductors, Taiwanese hardware, copper, energy and power equipment de-rate. DM may outperform defensively, particularly if lower inflation allows rate cuts and supports duration. The USD would probably strengthen initially on risk aversion, even if US growth also slows.

Own the constraints, not generic AI beta.

Bottom line

In our view, the AI capex buildout is a long-term trend which is here to stay. This is likely the transformational technology of our lifetime. Unlike the fibreoptic buildout during the DOTCOM era, utilisation rates of GPUs and data centres are running at close to 100%, with demand considerably higher. However, physical constraints will continue to be a headwind for AI infrastructure until innovative solutions can be devised to circumvent them. This does not mean that the AI cycle will collapse. Instead, projects shift right, data centre projects redirect to more power rich regions and scarce suppliers keep pricing power.

If demand continues to exceed what chips, memory, power and grid infrastructure can physically deliver, the investment implication is to own the constraints. This is now the consensus view. But what is less talked about is the multi-year tailwind this backdrop can provide for EM assets, even if headline US AI capex growth begins to flatten. This is a strong argument for owning North Asian semiconductors, selected Asian hardware supply chains, power equipment, copper and power-rich EM regions.

Ashmore

Head office

Ashmore Investment Management Limited, 16 Palace Street, London, SW1E 5JD T: +44 (0)20 3077 6000

Local offices

Bogota
T: +57 601 589 9000

Dublin
T: +353 1588 1300

Jakarta
T: +6221 2953 9000

Lima
T: +511 391 0396

Mexico
T: +52 55 4065 8898

Mumbai
T: +9122 6269 0000

New York
T: +1 212 661 0061

Riyadh
T: +966 11 483 9100

Singapore
T: +65 6580 8288

Tokyo
T: +81 03 4571 1601

Qatar

Fund prices
www.ashmoregroup.com
Bloomberg
FT.com
Reuters
S&P
Lipper

www.ashmoregroup.com 𝕏 @AshmoreEM