

The AI Infrastructure Constraint Map

A practical map of the bottlenecks shaping the next phase of AI capital spending.

M3I Research by Grish Ram · m3iresearch.com · July 2026 edition

Forecast record: 6 confirmed · 1 missed · 2 open — every call timestamped publicly at m3iresearch.com/audit-trail

Published for informational and educational purposes only. Not investment advice, an offer to buy or sell securities, or a recommendation tailored to any investor. Figures are drawn from public reporting as of July 2026 and may be revised.

How to read this map

The AI buildout is a capital cycle. At any moment, one layer of the stack tends to act as the **binding constraint** — the layer where demand outruns supply, lead times stretch, and pricing power concentrates. As capacity is added, the constraint migrates to the next layer. This map works through the stack layer by layer, and for each one asks three questions:

1. The constraint	Where is supply tight, and why is it structural rather than cyclical?
2. The evidence	Which observable indicators confirm the constraint is binding?
3. The challenge	What evidence would indicate the constraint is easing — and the thesis weakening?

Everything here is framed to be falsifiable. The point is not certainty about the future — it is a structured way to notice, earlier than the headlines, when the facts change.

Layer 1 — Compute — GPUs, accelerators and custom silicon

Merchant accelerators remain the largest single line of AI capex, but the structural story of 2026 is **second-sourcing**: hyperscalers building custom silicon programs to diversify beyond a single vendor. Marvell's custom-silicon franchise spans multiple hyperscaler design programs, with Google reported in 2026 as a third TPU design partner alongside Broadcom. Apple committed more than **\$30 billion** to Broadcom in July 2026 for US-made custom components under its American Manufacturing Program.

Policy is now a first-order variable: US export controls cap advanced-chip volumes to China (with tariffs layered on re-exports), and Nvidia's share of the Chinese AI-chip market roughly halved to ~50% by early 2026. Compute supply is increasingly shaped by regulation as much as by fabs.

What to monitor

Hyperscaler capex guidance; custom-silicon design wins and tape-outs; export-control changes; accelerator lead times; the ratio of AI revenue growth to AI capex growth.

Evidence that would challenge the thesis

A sustained hyperscaler capex deceleration; custom-silicon programs cancelled or consolidated; export-control relaxation that reopens supply faster than demand absorbs it.

Layer 2 — Memory — HBM and DRAM (the current binding constraint)

Memory is where the constraint bound in 2026. In the June–July reporting round: Micron reported quarterly revenue of ~\$41.5B, up 346% year-on-year, described HBM as **fully booked**, guided that the shortage persists **well beyond 2027**, and disclosed **~\$100B in cumulative minimum revenue commitments** across 16 strategic customer agreements. Samsung followed with the largest quarter in its history — roughly \$58B of operating profit — on DRAM prices up ~44% and NAND up ~53% quarter-on-quarter, HBM-led.

The same data cuts both ways: contract pricing rising 40–60% in a single quarter confirms the constraint **and** marks a later phase of the memory cycle. Multi-year contracted demand distinguishes this cycle from prior commodity-DRAM cycles; stretched pricing is where the cycle risk now concentrates.

What to monitor

DRAM/HBM contract price trajectory vs spot; memory-maker gross-margin trajectory; the contracted share of revenue; Chinese memory-capacity ramp and yield progress.

Evidence that would challenge the thesis

Contract prices peaking or spot inverting below contract; hyperscalers renegotiating committed volumes; Chinese suppliers scaling qualified HBM/DRAM capacity faster than AI demand absorbs it.

Layer 3 — Interconnect — networking and storage

Every added accelerator multiplies east-west traffic: networking scales super-linearly with cluster size, spanning switching silicon, optics and interconnect fabrics. Custom-silicon programs also pull networking attach with them.

Storage has quietly tightened alongside memory — Micron's data-centre SSD revenue more than doubled sequentially in its June quarter — as inference workloads pull petabyte-scale fast storage closer to compute.

What to monitor

Networking attach rates per cluster; optics lead times; data-centre SSD pricing and volumes; the share of cluster capex going to interconnect vs compute.

Evidence that would challenge the thesis

Cluster architectures that economize on interconnect (fewer, larger nodes); optics/switching capacity additions outpacing cluster growth; storage pricing rolling over.

Layer 4 — Physical plant — power, cooling and the grid

The layer most likely to inherit the constraint. Gas-turbine manufacturing slots are reported sold out **through 2030**; grid interconnection queues stretch for years in the major data-centre corridors; and high-density racks are forcing the transition from air to liquid cooling.

Unlike chips, power infrastructure cannot be de-bottlenecked in one product cycle — the capital assets take half a decade to permit and build. Energy-security politics currently act as a tailwind for domestic buildout. This layer is **power-constrained rather than oil-constrained**: its exposure to the current geopolitical conflict is limited, which is also why it screens as the least war-exposed segment of the stack.

What to monitor

Turbine and transformer lead times; interconnection-queue depth; utility capex plans; data-centre power-purchase agreements; liquid-cooling attach rates.

Evidence that would challenge the thesis

Interconnection reform or grid capacity landing faster than expected; data-centre efficiency gains (power per token falling) outpacing workload growth; permitting acceleration.

The watchlist — categories to monitor

Categories (with examples drawn from M3I's published research) rather than recommendations. The examples are companies M3I has covered publicly; inclusion is not an endorsement.

Layer	Category	Coverage examples (published research)
Compute	Merchant accelerators; custom-silicon designers	NVDA; MRVL, AVGO (custom programs)
Memory	HBM/DRAM manufacturers	MU; Samsung (005930.KS)
Interconnect	Networking silicon; optics; DC storage	MRVL; AVGO
Foundry	Leading-edge manufacturing	TSM; Samsung Foundry
Physical plant	Electrical infrastructure; turbines; grid equipment	IESC (data-centre electrical)

The macro and geopolitical risk overlay

The constraint thesis lives inside a macro regime. M3I tracks five flashpoints as live 0–100 tension scores, weighted by market impact (m3iresearch.com/predictions):

Flashpoint	Weight	Primary market channel
US–Iran / Hormuz	28%	Oil, tankers, safe-havens

China–Taiwan	26%	Semiconductors, TSMC, shipping lanes
Russia–Ukraine / NATO	22%	European gas, grain, defence
US–China trade & tech	14%	Chip export controls, tariffs, rare earths
Red Sea / Houthis	10%	Container freight, insurance, Suez routing

Two overlay principles matter for the AI stack specifically. First, **Taiwan concentration**: leading-edge manufacturing remains the single largest structural tail risk to the entire map. Second, **the conflict channel is indirect**: the AI/power buildout is domestically financed and power-constrained rather than oil-constrained — its principal sensitivities are positioning and hyperscaler capex guidance, not the Strait of Hormuz.

The invalidation checklist

The indicators that would most directly challenge the overall constraint thesis. These are tracked in M3I's models and revisited in the weekly briefing:

- 1. Memory contract prices roll over** — DRAM/HBM contract prices peaking, or spot inverting below contract (the memory-cycle top signal).
- 2. Hyperscaler capex guides cut** — two or more major platforms lowering AI infrastructure guidance in the same season.
- 3. Capex-to-revenue divergence widens** — AI revenue growth decelerating while infrastructure spend accelerates.
- 4. Chinese memory scales** — qualified HBM/DRAM capacity from Chinese suppliers ramping faster than demand absorbs it.
- 5. Breadth narrows at highs** — index concentration rising while participation thins (tracked live in the M3I AI Bubble composite).
- 6. A geopolitical trigger fires** — any Danger-Window trigger (strike on US personnel, Gulf civilian-infrastructure attack, direct Israel–Iran escalation) re-weights the wider-war tail and the risk overlay.

About M3I Research

M3I Research, by Grish Ram, publishes independent research on AI infrastructure through a macro-risk lens: the constraint, the regime it lives in, and a publicly audited scoreboard of every forecast — currently 6 confirmed, 1 missed, 2 open, with the misses and methodology revisions kept on the record.

Weekly briefing: m3iresearch.substack.com · Live models: m3iresearch.com · Record: m3iresearch.com/audit-trail

M3I Research is published for informational and educational purposes only. It does not constitute investment advice, an offer to buy or sell securities, or a recommendation tailored to any individual investor. Forecasts and models are uncertain and may be incorrect. Company figures are drawn from public reporting as of July 2026 and may be revised. All investing involves risk, including loss of principal.