INDUSTRY NOTE · SEMICONDUCTORS & AI INFRASTRUCTURE

The Physical Limits of Compute

Structural Capital Allocation in the 2026–2027 AI-Infrastructure Cycle

AUTHOR
Aydin Ali

PUBLISHED
July 2026

COVERAGE
HBM · Foundry · Power

CORE THESIS

The market is pricing this cycle through the accelerator designer. I think the higher-margin, longer-duration opportunity sits one layer down — in the physical infrastructure that gates how much compute can actually get built: high-bandwidth memory, advanced packaging capacity, and the power to run it.

CONTENTS

$418.6B
Projected 2026 DRAM revenue, nearly tripling on HBM demand alone.

1st
Layer of the AI stack I think is mispriced: physical infrastructure, not accelerator design.

4 risks
Argued against my own thesis in Section 04, including one that already happened.

Independent research note by a rising junior at Northern Virginia Community College, applying as a transfer student to UVA's McIntire School of Commerce. Documents the reasoning behind personal capital allocation inside AEA Capital Research, tracked publicly at aea-capital-research.haiden1690.workers.dev under a written Investment Policy Statement. Not a FINRA/SEC-defined research report, not produced by a broker-dealer or registered investment adviser, and not investment advice or a solicitation. Sourcing is documented in the footnotes on each page.

anymore. Hyperscalers are guiding to a combined $635–725 billion of 2026 capital expenditure, up roughly 77% year over year, and the industry is now bottlenecked below the chip design layer — in high-bandwidth memory supply, advanced packaging capacity, and the power required to run what gets built.[1]

DRAM revenue is on pace to nearly triple in 2026 to roughly $418.6 billion, almost entirely on HBM demand, while the broader memory segment grows an estimated 250% year over year.[2] Leading-edge foundry capacity remains the primary chokepoint for every accelerator design, regardless of who designed it.[3] And inside the neocloud sector I track directly, the framing has already shifted: the AI infrastructure race increasingly "looks less like a scramble for Nvidia GPUs and more like a fight over electricity."[4]

I think the highest-margin trade in this cycle is owning the names that gate what the accelerator designer cannot build without — not the designer itself, where I think crowding is worst and the multiple already assumes the best case. This note lays out why, where I think the thesis breaks, and how it changes what I actually hold.

02 THE SECULAR TAILWIND & THE ACADEMIC LENS

Computing is going through its third architectural shift in fifty years — mainframe to client-server, client-server to cloud, and now general-purpose to accelerated. The distinction matters more than the label. A CPU is built for flexibility: it can run anything, adequately. An accelerator is built for one operation — parallel matrix multiplication — at a throughput a CPU cannot approach. Hyperscalers aren't buying accelerators because they're fashionable; the unit economics of running a large model on general-purpose silicon stopped making sense years ago, and the gap has widened every quarter since.

That shift shows up first in the income statement, but it's only fully legible once you trace it through the balance sheet and cash flow statement together — which is where my Managerial Accounting coursework earns its keep. Hyperscaler capex doesn't hit the P&L as an expense; it capitalizes, then depreciates over a useful life every one of these companies has been quietly shortening as GPU generations turn over faster than the original schedules assumed. Watch the depreciation assumption, not just the capex headline: when a company extends useful life to smooth EPS, that is a real signal about how confident management actually is in the hardware's durability.

> *"A low trailing P/E on a memory name can be a value trap if pricing is about to roll over — or a genuine bargain if the up-cycle has further to run. The multiple alone doesn't tell you which."*

Inventory velocity tells the other half of the story. A memory maker sitting on rising inventory with slowing turns is signaling oversupply, regardless of what management says on the call. A memory maker with accelerating turns into a supply-constrained market is signaling real pricing power — not a qualitative read, but a ratio I can pull directly from a 10-Q, and the same lens I'd apply to any inventory-heavy business, hype cycle or not.

[INSERT CHART 1: AEA Custom Graphic — Hyperscaler CapEx vs. HBM Pricing Premium, 2023–2026]

03 THE BOTTLENECK ANALYSIS

High-Bandwidth Memory

growing an estimated 250% year over year.[2] That capacity does not come online quickly — new fabrication lines take eighteen to twenty-four months even once capital is committed today, and that lag is the entire trade. Memory pricing has historically been the most cyclical, least defensible line item in semiconductors; I think the market is still underwriting the old cyclicality here instead of the new structural floor. Pricing power that clean in a historically commoditized product does not last forever, but it does not unwind in a quarter either.

Foundry & Advanced Packaging

Every high-end AI accelerator, regardless of who designs it, needs advanced packaging — increasingly Chip-on-Wafer-on-Substrate (CoWoS) integration — to physically combine the logic die with HBM stacks. Leading-edge packaging and foundry capacity remains the primary bottleneck for the entire AI supply chain: design wins mean little without wafer and packaging allocation, and that capacity runs overwhelmingly through **$TSM**, with Samsung and Intel Foundry a distant second and third.[3] **$TSM** is on my watchlist, not yet a position — I think its premium multiple already prices in most of this. But the structural point stands regardless of what I own: a company that controls the binding constraint in a supply chain carries a different risk-reward than one that controls a component with real substitutes, and I don't think the market prices that distinction correctly yet.

[INSERT CHART 2: AEA Custom Graphic — Leading-Edge Packaging Capacity Expansion vs. AI Accelerator Shipment Forecasts]

Power & Cooling

This constraint shows up last in a model and first in reality. Management teams broadly expect data-center-driven demand to meaningfully tighten U.S. power markets by late 2027 or early 2028, and inside the neocloud names I track, power availability — not chip supply — is already becoming the binding constraint on new capacity.[4] Two positions give me direct exposure to the response: **$VST** and **$TLN** are merchant power producers that have converted that scarcity into long-duration, contracted cash flow — Vistra has signed 20-year nuclear power purchase agreements covering over 2,600 MW of zero-carbon generation and committed roughly $4 billion to acquire additional gas and nuclear capacity specifically for data-center demand, while Talen has agreed to supply 1,920 MW of carbon-free power from its Susquehanna nuclear plant to AWS through 2042.[5] **$VRT**, a third position, sits one layer removed as a supplier of the power and liquid-cooling infrastructure that rack densities at modern accelerator clusters now require as a baseline, not a niche spec.[6] This is the part of the thesis that reads least like a semiconductor note and most like a utilities note — which is exactly why I think it is underpriced by investors who entered this trade through the chip layer and never had a reason to look at merchant power economics.

[INSERT CHART 3: AEA Custom Graphic — U.S. Data Center Power Demand Growth vs. Grid Interconnection Queue Times]

04 RED TEAM: STEELMANNING AGAINST MY OWN THESIS

Risk 1 — Scaling laws plateau. The entire physical-bottleneck thesis assumes demand keeps outrunning supply. If the marginal performance gain from larger models keeps shrinking — and there is real research suggesting exactly that — hyperscalers stop chasing parameter count and capex growth decelerates faster than any physical constraint resolves. In that world the bottleneck stops being a scarcity story and be-

stops. If future AI value creation shifts from training frontier models to cheaply serving already-good-enough ones, the build-out this note underwrites gets smaller than the market currently assumes.

Risk 3 — Disintermediation is already live, not hypothetical. Meta's entry into GPU-capacity rental ("Meta Compute") triggered a real, dated selloff in July 2026 — two neocloud names I track directly fell roughly 15% on the news.[7] That is not a scenario in a model; it already happened, and it is the clearest live evidence that hyperscaler in-house build-out can compress the economics of the very infrastructure layer I am arguing is structurally advantaged.

Risk 4 — Taiwan concentration is binary, not diversifiable. Leading-edge fabrication and packaging capacity outside Taiwan remains years behind. Any disruption to that capacity does not just hurt the packaging trade — it removes the physical substrate the entire thesis depends on, chip designers included. I hold that risk explicitly, not as a footnote.

None of these four risks are fully hedged in my own positioning today. I think pretending otherwise would undercut the entire premise of this note.

concentration cap for exactly this reason: this thesis is high-conviction, not certain, and sizing has to reflect the difference.

THE MANDATE, IN PRACTICE

It is also why my current positioning leans into the physical layer — memory (**$MU $SNDK**), power infrastructure (**$VST $TLN $VRT**) — rather than concentrating further in accelerator design alone, where I think crowding is worst and the multiple already assumes the best case. The bottleneck thesis is testable, dated, and published under my own name specifically so it can be checked against what actually happens over the next eighteen months, not just cited when it's convenient.

I would rather be visibly, checkably wrong about a specific claim than vaguely right about a trend everyone already agrees with. That is the discipline this entire portfolio is built on, and it is the standard I am holding this note to.

1. Hyperscaler 2026 capex guidance aggregated from company reporting; see AEA Capital Research, Portfolio Sector Overview (July 2026), Mega-Cap Tech Platforms and AI Infrastructure & Neoclouds sections.

2. DRAM/memory 2026 revenue growth figures per WSTS-based industry estimates, as cited in AEA Capital Research, Portfolio Sector Overview (July 2026), Semiconductors & Chips section.

3. Leading-edge foundry/packaging bottleneck framing per AEA Capital Research, Portfolio Sector Overview (July 2026), Semiconductors & Chips, Key Trends & Drivers.

4. Quotation and power-availability framing per AEA Capital Research, Portfolio Sector Overview (July 2026), AI Infrastructure & Neoclouds section.

5. Vistra and Talen PPA terms per AEA Capital Research, Portfolio Sector Overview (July 2026), Power & Utilities section, sourced from company disclosures.

6. Liquid-cooling baseline shift per AEA Capital Research, Portfolio Sector Overview (July 2026), AI Infrastructure & Neoclouds section.

7. July 2026 neocloud selloff on Meta Compute news per AEA Capital Research, Portfolio Sector Overview (July 2026), AI Infrastructure & Neoclouds, Key Trends & Drivers.

Methodology note. This note synthesizes AEA Capital Research's own previously-published Sector Comps Analysis and Portfolio Sector Overview, both built from public company filings and third-party financial-data aggregators, cross-referenced against the author's own live portfolio positioning as disclosed on Holdings and Watchlist. All ticker references reflect real positions or watchlist names as of publication.