THE AI STACK WEEKLY ISSUE 15 · WEEK 31 OF 2026

THE BOTTOM LINE

The cheapest capability and the cheapest capacity both came from contracts and harnesses, not from chips or weights

Nothing new was built this week. No accelerator launched, no closed frontier model shipped, and no frontier lab closed a round. Yet both the largest capability gain and the largest capacity gain on the board arrived anyway, and they arrived from the same place: the structures wrapped around assets that already existed.

On the capability side, OpenAI reported tripling its ARC-AGI-3 public-set score from 13.3% to 38.3% by changing two harness settings — retaining private reasoning between tool calls and compacting old context instead of truncating it — while consuming roughly six times fewer output tokens. Same weights, same model, different memory policy. That is a vendor-measured result on one benchmark, and it deserves the discount that implies. It is still the week's most consequential number, because it says a large fraction of what teams currently attribute to model quality is actually attributable to the scaffold they wrapped around it.

On the capacity side, Meta and BlackRock closed a venture for a 1 GW El Paso campus in which BlackRock-managed funds take 80% and Meta keeps 20%, contributes land and construction in progress, leases the entire campus back for an initial four years extendable toward twenty, and sits behind roughly $12.5B of debt against a residual-value guarantee threshold near $13B. Core Scientific separately converted AMD's rack roadmap into about 530 MW of contracted, third-party-owned capacity with a path toward 2.5 GW — before a single Helios rack has shipped. Neither transaction added a megawatt this week. Both changed who owns the asset, who carries the debt, and what a gigawatt costs the operator that uses it.

OpenAI's price cut belongs in the same frame rather than in a price-war frame. Luna fell 80% to $0.20/$1.20 and Terra 20% to $2/$12 while Sol held at $5/$30. On a

fixed workload of one million input and four million output tokens, Luna's bill drops from $25 to $5 while Sol stays at $125 — the distance between the cheapest and most expensive tier of the same family widens from about 5x to 25x in a single day. The competitive story is secondary. The operational story is that tier selection now carries 25x of cost leverage inside one vendor, which makes routing policy a larger financial decision than vendor selection.

The decision implication runs three ways. Model buyers should re-run their agent evaluations with identical weights and different memory and routing policies before attributing any gap to the model, and should instrument per-tier token consumption and completion rates before moving volume to Luna on list price alone. Infrastructure buyers should read the El Paso and Core Scientific structures as the emerging default and price residual-value guarantees, lease terms, and third-party ownership as part of capacity cost. And anyone modeling 2027 unit economics should stop indexing to accelerator generations: with two memory suppliers describing shortage into 2028 and Amazon naming memory as the primary driver of a $20B capex revision, HBM contracts now set the clearing price that GPU roadmaps used to.

AUTHOR

Brian Letort
BrianLetort.AI

PUBLISHED

August 1, 2026
Issue 15

THIS WEEK IN 30 SECONDS

Executive summary.

- The week's biggest capability gain and its biggest capacity gain both arrived without new silicon or new weights: OpenAI tripled its ARC-AGI-3 score by retaining reasoning and compacting context on the same model, and Meta moved a 1 GW campus off its balance sheet into an 80/20 BlackRock venture carrying $12.5B of debt.
- OpenAI cut GPT-5.6 Luna 80% to $0.20/$1.20 and Terra 20% to $2/$12 while holding Sol at $5/$30 — on a fixed 1:4 workload that widens the cheapest-to-premium spread from roughly 5x to 25x, making tier routing a larger cost lever than vendor choice.
- Moonshot shipped Kimi K3's full 2.8T weights on July 27, but the revenue-tiered license moves open-weight diligence from model evaluation to license classification: commercial hosting above $20M trailing revenue needs a separate deal.
- Memory, not accelerators, set the price signal: SK hynix and Samsung both printed records and described HBM shortage into 2027-2028, and Amazon attributed a $20B increase in CY2026 capex guidance primarily to memory costs.
- The networking hypothesis strained for the first time — Equinix posted record interconnection adds while interconnection revenue grew about 9% against roughly 16% total revenue growth, which is the specific falsification test named on the thesis page.
- No accelerator launched, no closed frontier model shipped, and no frontier lab closed a funding round; the week's movement was entirely in contracts, licenses, harnesses, and memory supply.

BY THE NUMBERS

5x to 25x
GPT-5.6 cheapest-to-premium price spread on a fixed 1:4 workload, before and after the July 30 cut

13.3% to 38.3%
OpenAI's ARC-AGI-3 public-set score from harness memory settings alone, same model

$12.5B
Debt financing in the Meta-BlackRock El Paso venture — 1 GW, online from 2028

~$220B
Amazon CY2026 cash capex guidance, raised from ~$200B

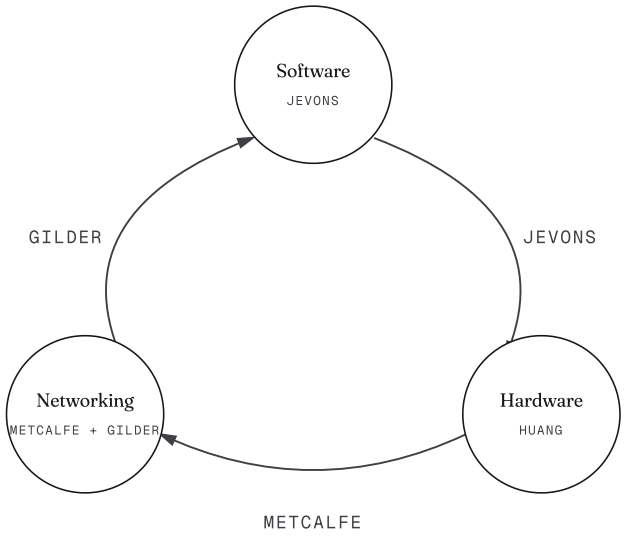
$678B
Microsoft commercial remaining performance obligation, up 84% year over year

2.8T / 104B
Kimi K3 total and active parameters, released under a revenue-tiered license

HOW WE READ THIS WEEK

Three lenses, one flywheel.

Cheaper inference pulls in more workloads. More workloads need more compute. More compute needs denser fabric. Denser fabric unlocks new architectures, which lower the cost of inference again. Read a week's news across that loop and the noise sorts itself out.



THIS WEEK'S ARC

All three lenses.

Software → Hardware new model capability creates new use cases that consume more compute (Jevons).

Hardware → Networking more compute means more nodes; the value of the connecting fabric scales as the square of the nodes (Metcalfe).

Networking → Software denser, higher-bandwidth interconnect makes new model architectures viable (Gilder).

Hardware itself GPU performance doubles faster than Moore's Law (Huang).

THE BAR

The events that actually matter touch at least two of the three lenses. Single-lens reads are noise dressed up as motion. Each section of this brief grades its evidence and ties the implication back to the flywheel.

LENS 1 OF 3 THE FLYWHEEL

Software.

Model releases, pricing, capability benchmarks, license posture, and capability-risk disclosures.

JUL 27
EVENT 01

Moonshot published Kimi K3's full 2.8T MoE weights (104B active, 896 experts, 1M context) under a revenue-tiered license requiring a separate commercial deal for hosting above $20M trailing revenue

MOONSHOT AI; HUGGING FACE; GITHUB

JUL 30
EVENT 02

OpenAI cut GPT-5.6 Luna 80% to $0.20/$1.20 and Terra 20% to $2/$12, held Sol at $5/$30, and replaced Priority Processing with a Fast mode at 2x price for up to ~2.5x speed

OPENAI

JUL 29
EVENT 03

OpenAI reported tripling its ARC-AGI-3 public-set score from 13.3% to 38.3% using retained reasoning plus context compaction on the same model, with roughly 6x fewer output tokens

OPENAI

JUL 31
EVENT 04

DeepSeek put V4-Flash-0731 into public beta at unchanged $0.14/$0.28 pricing and unchanged 284B/13B architecture, and Artificial Analysis measured its Intelligence Index up 10 points to 50

DEEPSEEK API CHANGELOG; ARTIFICIAL ANALYSIS

JUL 30
EVENT 05

Thinking Machines released Inkling-Small under Apache 2.0 — 276B total, 12B active, multimodal — and Artificial Analysis scored it 40 against Inkling's 41 at under a third the parameters

THINKING MACHINES LAB; ARTIFICIAL ANALYSIS

WHAT THIS MEANS

The model layer stopped competing on weights this week and competed on scaffolds and rate cards instead. OpenAI's harness result and DeepSeek's ten-point Index gain on unchanged architecture both say the same thing: post-training and context policy are now moving capability faster than parameter counts, so teams should re-baseline agents against memory and routing changes before buying a bigger model. On the open side, K3 and Inkling-Small expand the deployable frontier while relocating the real gate to licensing and VRAM footprint.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

Silicon, density, packaging, memory supply, and the share of incremental compute going to custom silicon.

JUL 29
EVENT 01

SK hynix reported record Q2 revenue of KRW 79.3T at a 76% operating margin, began HBM4 mass shipments, and said long-term agreements with about ten customers are set while 2027 volumes and prices remain under negotiation

SK HYNIX; YONHAP

JUL 30
EVENT 02

Samsung posted record semiconductor results, shipped industry-first HBM4E samples, and guided Q3 HBM4 sales to more than triple sequentially while describing shortage worsening in 2027 and persisting into 2028

SAMSUNG ELECTRONICS; CNBC

JUL 30
EVENT 03

Amazon raised CY2026 cash capex guidance to about $220B from about $200B, attributing the increase primarily to higher memory prices, and said capacity will still fall short of 2026 demand

AMAZON; CNBC EARNINGS-CALL COVERAGE

JUL 28
EVENT 04

Core Scientific and AMD signed 15-year agreements covering about 530 MW across five sites with more than $14B of potential base contracted revenue and AMD rights to reserve toward roughly 2.5 GW, with deliveries starting 2027

CORE SCIENTIFIC

JUL 29
EVENT 05

Microsoft spent $41B on capex in the June quarter with roughly two-thirds on short-lived compute assets, added about 1 GW and 31 data centers in the quarter, and reached 88 data centers online across FY26

MICROSOFT; DATACENTER DYNAMICS

WHAT THIS MEANS

This was a memory week, not a silicon week, and that matters more for planning. With SK hynix and Samsung both reporting records while describing shortage persisting into 2028, and Amazon attributing a $20B capex increase primarily to memory prices, the binding input on 2027 capacity has visibly moved from accelerator supply to HBM contracts. Buyers negotiating multi-year reserved capacity should assume memory pass-through in pricing and should treat 2027 delivery commitments that are not backed by disclosed HBM allocation as schedule risk rather than schedule.

LENS 3 OF 3

THE FLYWHEEL

Networking.

Interconnect, fabric standards, optical capacity, sovereign and operator-level networking products.

JUL 29
EVENT 01

Equinix reported a record 9,700 net interconnection adds and $453M of interconnection revenue, but interconnection revenue grew about 9% against roughly 16% total revenue growth, and raised FY2026 guidance to $10.205-10.285B

EQUINIX

JUL 29
EVENT 02

Microsoft said dock-to-live time for new GPUs in its largest regions fell nearly 50% year over year, with servers, networking and software assets rising to $215.9B from $132.8B

MICROSOFT EARNINGS CALL; DATACENTER DYNAMICS

JUL 29
EVENT 03

The DOE Paducah selection was marketed on existing transmission, water, fiber and road access with NextEra to upgrade transmission near the campus, but disclosed no interconnect or long-haul capacity figures

DOE; BROOKFIELD; NEXTERA ENERGY

WHAT THIS MEANS

Optics, switch silicon and the standards bodies were quiet in-window, so the honest networking read comes from a colocation print rather than a product launch — and it cuts against this publication's own hypothesis. Record interconnection adds at Equinix coincided with interconnection revenue growing roughly 9% while total revenue grew about 16%. Investors and architects should not treat unit adds as proof that interconnect is the fastest-growing layer; the revenue mix is the test, and this quarter it went the other way.

CAPITAL FLOW

Money in, revenue out.

Capital deployed (forward) vs revenue out (quarterly or run-rate). Burn-to-revenue = revenue / capital — lower means more out than in. Bars normalized to 250.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~4.5x

No frontier lab closed primary financing in-window. The reported Nvidia backstop of up to ~$250B for OpenAI's Ohio campus is talks-only and is deliberately excluded from capital in.

Hyperscaler-Hosted

AZURE-OPENAI, AWS-ANTHROPIC, GOOGLE CLOUD-GEMINI, ORACLE-OCI

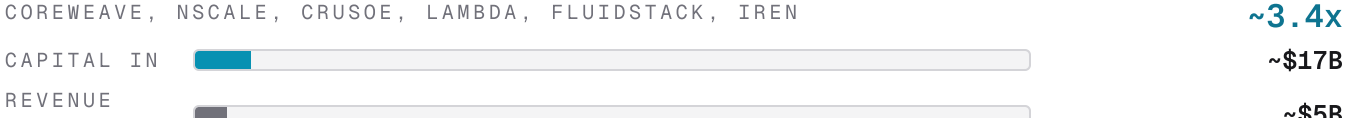
BURN / REV ~3.6x

Amazon raised CY2026 cash capex guidance by about $20B to ~$220B and Meta lifted its full-year floor to $130-145B, while Azure passed $100B of annual revenue and AWS re-accelerated to +37%.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

BURN / REV ~3.4x

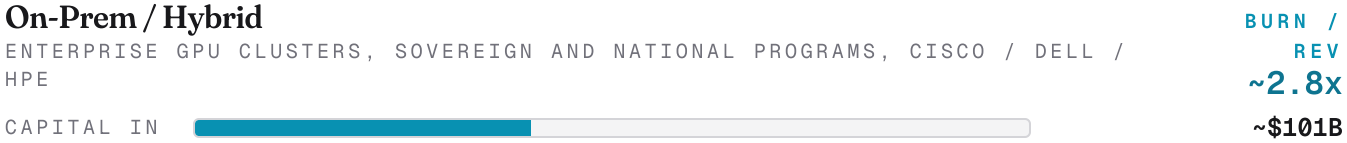


No neocloud financing closed in-window and CoreWeave's backlog remains the March figure until its August 11 print, but Core Scientific contracted about 530 MW to AMD with a path toward 2.5 GW.

On-Prem / Hybrid

ENTERPRISE GPU CLUSTERS, SOVEREIGN AND NATIONAL PROGRAMS, CISCO / DELL / HPE

BURN / REV ~2.8x



No change to the aggregate. Kimi K3's downloadable weights expand what a private cluster can actually run, while the license's revenue trigger adds a legal gate that did not previously exist for open models.

SIGNAL VS NOISE

What's real, what's noise.

7 claims that drove headlines this week, scored 1–5 on source quality and triangulation. 2 flagged as noise. The bar is at least one explicit noise call per issue.

5 / 5

CLAIM 01

Microsoft's Azure passed $100B of annual revenue while commercial remaining performance obligation reached $678B, up 84% year over year.

SOURCES: MICROSOFT INVESTOR RELATIONS PRESS RELEASE AND EARNINGS CALL

Both figures are primary and audited-grade. The discipline is in the second number: only about 30% of RPO converts inside twelve months, so this is contracted backlog rather than near-term revenue and should not be compared against annual capex as though it were.

5 / 5

CLAIM 02

Meta and BlackRock closed a venture for a 1 GW El Paso campus at roughly $14B development cost, with $12.5B of debt and BlackRock funds holding 80%.

SOURCES: META NEWSROOM AND INVESTOR RELATIONS; ADVISORS MORGAN STANLEY AND J.P. MORGAN

Structure, ownership split, lease term and residual-value threshold are all disclosed. This is the week's most transferable fact: it is a template for moving gigawatt-scale AI capacity off a hyperscaler balance sheet while retaining operational control.

4 / 5

CLAIM 03

Moonshot shipped Kimi K3's full 2.8T weights, converting last week's open-frontier watch item into a deployment fact.

SOURCES: HUGGING FACE AND GITHUB ARTIFACTS WITH A PUBLISHED LICENSE; SECONDARY ENTERPRISE ANALYSIS

The artifacts are real and dated, so the deployment claim holds. The license is the part buyers will feel: hosting above $20M trailing revenue needs a separate agreement, which moves the decision from an ML evaluation to a legal classification.

4 / 5

CLAIM 04

OpenAI cut GPT-5.6 Luna 80% and Terra 20% while holding Sol unchanged, widening the intra-family price spread.

SOURCES: OPENAI PRODUCT ANNOUNCEMENT AND PRICING DOCUMENTATION; MULTI-OUTLET CONFIRMATION

The rate card is a primary fact and the spread arithmetic follows from it. What remains unmeasured is completed-task cost: Luna may consume more tokens or attempts on hard work, so the 25x list-price gap is an upper bound on realized savings, not a forecast.

3 / 5

CLAIM 05

HBM supply will remain short through 2027 and into 2028, keeping memory pricing elevated.

SOURCES: SK HYNIX AND SAMSUNG Q2 RELEASES PLUS EARNINGS-CALL PARAPHRASE IN KOREAN BUSINESS PRESS

Two independent suppliers pointing the same direction in the same week is meaningful, but the specific shortage language comes from call commentary rather than an audited disclosure. Treat it as a strong planning assumption, not a confirmed supply curve.

2/5 — NOISE

CLAIM 06

Nvidia will guarantee up to $250B of OpenAI's data-center financing, effectively underwriting the Ohio campus.

SOURCES: WSJ AND CNBC REPORTING CITING ANONYMOUS SOURCES; NVIDIA DECLINED TO COMMENT

Noise as stated. There is no filing, no definitive agreement and no confirmation from either party. Vendor credit-wrapping is a genuinely important emerging structure, which is exactly why it should be scored on documentation rather than on reporting.

2/5 — NOISE

CLAIM 07

Microsoft's disclosure of more than 30 million paid Copilot seats demonstrates that enterprise AI is delivering measurable productivity returns.

SOURCES: MICROSOFT EARNINGS CALL; VENDOR-REPORTED SEAT AND INTERACTION COUNTS

Seats are a distribution metric, not a value metric. Neither average revenue per seat, weekly active use, nor workflow attach was disclosed, and a 60% sequential Copilot revenue figure is vendor-characterized rather than a reported line item. Procurement should ask for the usage data before treating seat growth as ROI evidence.

HOUSE MEASUREMENT · FILING-DERIVED

OpenAI's July 30 price change widened the cost distance between its cheapest and most expensive GPT-5.6 tier from roughly 5x to 25x on an identical workload.

METHOD: Take a fixed representative agentic workload of 1,000,000 input tokens and 4,000,000 output tokens — the same 1:4 ratio this publication has used in prior issues for continuity. Compute the list-price bill for each GPT-5.6 tier before and after the July 30 change, using published per-million-token rates: Luna $1.00/$6.00 before and $0.20/$1.20 after; Terra $2.50/$15.00 before and $2.00/$12.00 after; Sol $5.00/$30.00 unchanged. Bill equals input rate plus four times output rate. Divide the Sol bill by the Luna bill to get the intra-family spread. No caching discounts, Fast-mode premium, batch pricing or retry overhead is applied.

Luna bill on the fixed workload	**$25.00 to $5.00**	An 80% reduction, matching the headline rate cut exactly because the workload ratio is fixed
Terra bill on the fixed workload	**$62.50 to $50.00**	A 20% reduction
Sol bill on the fixed workload	**$125.00 unchanged**	The premium tier did not move, which is what widens the spread
Cheapest-to-premium spread within GPT-5.6	**5.0x to 25.0x**	Sol divided by Luna, before and after July 30

Tier routing is now a larger financial lever than vendor selection inside a single model family. A team that can classify 60% of its traffic as Luna-eligible saves more than it would by switching providers, which means routing policy deserves the scrutiny normally reserved for procurement. It also means per-tier token telemetry stops being an engineering nicety and becomes a finance requirement: without it, nobody can tell whether a 25x list-price gap is producing real savings.

Caveats: This is list-price arithmetic on a synthetic ratio, not a measured workload. It does not show that Luna completes the same tasks — cheaper models frequently consume more tokens, take more attempts, or fail outright, and any of those compress the realized ratio well below 25x. It excludes prompt caching, batch pricing, the Fast-mode premium, and the cost of retries and human review. A real comparison requires completed-task measurement, which no independent party has published.

SYNTHESIS

The week, reasoned through.

Cross-domain connections, the standing thesis tested against this week's evidence, and the patterns building across weeks. Every inference is labeled by reasoning type and linked to its evidence.

CONNECTING THE DOTS

The marginal unit of AI advantage moved from the asset to the contract and the harness: the week's largest capability gain and its largest capacity gain both arrived without new silicon or new weights.

INDUCTIVE · 74% CONFIDENCE

OpenAI tripled its ARC-AGI-3 public-set score from 13.3% to 38.3% by retaining private reasoning between tool calls and compacting old context rather than truncating it, on the same model and with roughly 6x fewer output tokens.
Meta moved a 1 GW campus off its own balance sheet into an 80/20 BlackRock venture carrying $12.5B of debt, a residual-value guarantee threshold near $13B, and a leaseback extendable toward twenty years.
Core Scientific converted AMD's rack roadmap into about 530 MW of contracted, third-party-owned capacity with a path toward 2.5 GW, before a single Helios rack has shipped.

STEEL-MAN: The strongest objection is that neither move creates anything: the harness gain is a vendor-measured result on a single puzzle benchmark with obvious incentive to publish it, and an SPV is an accounting rearrangement that does not add a megawatt or a FLOP. Both points are fair, and the claim narrows accordingly. What survives is that buyers do not procure the frontier, they procure a given capability or capacity at a given cost, and both of those costs moved substantially this week through structure alone. That is a real change in what a buyer pays even if the underlying frontier stood still.

Memory, not accelerators, is now setting the clearing price for 2027 AI capacity, and the signal is arriving through hyperscaler operating guidance rather than through chip roadmaps.

ABDUCTIVE · 71% CONFIDENCE

SK hynix and Samsung both printed record memory quarters and described HBM shortage persisting through 2027 and into 2028, with 2027 volumes and prices still under negotiation across roughly ten customers.
Amazon raised CY2026 cash capex guidance by about $20B to ~$220B and attributed the increase primarily to higher memory prices, not to additional capacity.
No new accelerator silicon launched in-window, so the entire cost signal for next year's capacity reached the market through the memory supply chain and a cloud provider's guidance revision.

STEEL-MAN: The memory attribution comes from an earnings call rather than a filing line item, and memory is a conveniently external explanation for capex growth that might instead reflect volume or a bad negotiation. The inference holds because two independent suppliers described the same shortage direction in the same week while both were reporting records — a coincidence that is harder to explain as narrative management than as supply reality. It would be weakened considerably if Q3 shows memory prices flat while capex guidance rises again.

Open weights and open protocols moved in opposite directions on enterprise usability this week: the weights became more conditional while the connector standard became more portable.

INDUCTIVE · 68% CONFIDENCE

Moonshot published Kimi K3's full 2.8T weights but gated commercial hosting above $20M trailing revenue behind a separate agreement, with attribution duties above 100M monthly active users.
The MCP 2026-07-28 specification went final with a stateless, cacheable, gateway-routable core, enterprise-managed authorization, and dynamic client registration deprecated.
Snowflake and GitHub shipped governance and orchestration surfaces in the same week that assume MCP is the connector substrate rather than one integration option among several.

> **STEEL-MAN:** A revenue-tiered license is irrelevant to the majority of enterprises that self-host for internal use, so describing the weights as more conditional overstates the friction for a typical buyer. The distinction still matters because it relocates diligence from model evaluation to license classification, which is a different team on a slower clock. The claim would break if K3's license turns out to be read as unambiguously permissive for internal enterprise deployment without counsel involvement.

THESIS TEST

H1 · SUPPORTED **The cycle is accelerating, not slowing.** The cadence held even in a week with no new silicon and no new closed frontier model. DeepSeek gained ten Artificial Analysis Index points on unchanged architecture and unchanged price through re-post-training alone, Moonshot shipped 2.8T open weights, Thinking Machines matched its flagship within one Index point at under a third of the parameters, and OpenAI tripled a benchmark score through harness policy. Four independent capability or cost improvements inside seven days is acceleration, and notably none of them required a hardware generation to arrive first — which is a faster loop than the flywheel model assumes. **AGAINST IT:** No accelerator launched in-window, and Gemini 3.5 Pro missed a general-availability deadline this publication had predicted for July 31, remaining in partner testing. If the hardware side of the flywheel is stalling while software compounds, the coupled acceleration this hypothesis describes is not what is actually happening.

H2 · SUPPORTED **Capital is concentrated, returns are diffuse.** The earnings prints made the gap unusually legible. Microsoft's commercial RPO reached $678B, up 84%, while only about 30% converts inside twelve months. Meta spent $31.1B in a quarter and reported free cash flow of $784M, down 91%. Amazon's trailing free cash flow was negative $7.6B while it raised capex guidance by $20B. Capital is concentrating at four balance sheets and the returns are showing up as contracted future obligations and as enterprise seat counts rather than as current cash generation at the spender. **AGAINST IT:** Azure crossing $100B of annual revenue and AWS re-accelerating to 37% growth are the strongest evidence yet that concentrated capex is converting into revenue at the spender rather than diffusing away from it. If that continues for two more quarters, the hypothesis needs rewriting rather than defending.

H3 · STRAINED **Networking is the durable layer.** This publication named the falsification test in advance: interconnect growth lagging compute growth at the colocation operators. Equinix delivered exactly that pattern. A record 9,700 net interconnection adds coincided with interconnection revenue of $453M growing roughly 9% on a normalized basis while total revenue grew about 16%. Unit demand for interconnect is genuinely at a record, but the pricing power the hypothesis predicts should show up in revenue mix, and this quarter the mix moved toward power and space instead. One quarter from one operator is not the two consecutive quarters required for a revision, so the hypothesis is flagged rather than rewritten.

H4 · SUPPORTED **Open weights pull the floor up.** Three releases in one week expanded what a private cluster can run. Kimi K3's 2.8T weights with a million-token context are downloadable, Inkling-Small is Apache 2.0 at an Index of 40, and DeepSeek's cheapest tier gained ten Index points without a price change. All three re-route compute demand toward self-hosted and sovereign capacity rather than reducing it, which is the mechanism this hypothesis describes. The demand-catalyst reading is intact. **AGAINST IT:** Both releases carry gates the word open obscures. K3 requires a separate commercial agreement for hosting above $20M trailing revenue, and Inkling-Small needs roughly 600 GB of VRAM at BF16 or 180 GB at NVFP4. If open frontier models keep arriving with revenue triggers and cluster-class footprints, the floor they pull up is only reachable by organizations that already had a floor.

H5 · SUPPORTED **Power is the binding constraint for the next 24 months.** The Paducah selection is the cleanest illustration yet of the constraint's shape. Delivering roughly 1.8 GW to a campus required pairing it with up to 2 GW of new dedicated natural gas and up to 2.6 GW of storage, with completion targeted for 2031 and the power service agreement still subject to Kentucky regulatory approval. Meta's fully financed 1 GW El Paso campus does not come online until 2028. In both cases capital was available immediately and electricity was not, which is precisely the ordering this hypothesis asserts. **AGAINST IT:** Both projects found a path. Dedicated generation, federal land, and third-party capital together produced credible multi-gigawatt commitments in a single week, which suggests power is an expensive and slow constraint rather than a binding one. A binding constraint stops projects; this week it only delayed and repriced them.

PATTERN WATCH

Frontier model economics are shifting from headline list price toward route-aware completed-task cost, and the publicly available evidence is not keeping up with the pricing changes. (5 WEEKS OBSERVED)

Next week: Within two weeks an independent evaluator publishes a completed-task cost comparison for GPT-5.6 Luna that reports token consumption and retry or failure rates, not list price alone. If nothing appears by mid-August, the gap between pricing changes and public measurement is widening rather than closing, and buyers are routing on marketing.

Large AI capacity is migrating off operator balance sheets into third-party ownership structures backed by contractual credit support rather than by the operator's own equity. (3 WEEKS OBSERVED)

Next week: A second hyperscaler discloses a data-center joint venture or special-purpose vehicle with third-party majority ownership and some form of residual-value guarantee before the end of Q4 2026. If instead the next large campus is financed on balance sheet, the El Paso structure is a Meta-specific response to its own free-cash-flow compression rather than an industry template.

SECOND-ORDER EFFECTS

Model-tier selection becomes a larger cost lever than vendor selection, which shifts effective budget authority toward whoever owns routing policy and makes per-tier token telemetry a finance requirement rather than an engineering nicety. Teams without that instrumentation cannot tell whether a 25x list-price gap is producing any real savings, so the first casualty is the credibility of AI cost forecasts built on blended token assumptions.

Trigger: OpenAI's Luna cut widens the cheapest-to-premium price spread within one model family from roughly 5x to 25x on a fixed workload. Horizon: Next two to three quarters. Who moves: FinOps and platform engineering teams, AI gateway and routing vendors, and model providers that compete on headline list price rather than completed-task cost.

Connector traffic becomes ordinary cacheable HTTP that any gateway can route, audit and cache, which turns the enterprise agent control point into an infrastructure purchase rather than a platform choice. That erodes differentiation for agent platforms whose principal value was connector breadth, and it moves the buying conversation toward the vendors who already own the gateway and identity layers. Trigger: The MCP 2026-07-28 specification removes sessions, makes tool listings cacheable, and deprecates dynamic client registration. Horizon: Six to twelve months. Who moves: Agent platform vendors, API gateway and data-cloud vendors, and enterprise security and identity teams now inheriting connector authorization.

Memory pricing becomes an explicit pass-through in cloud unit economics, so GPU-hour prices stop tracking accelerator generations and start tracking HBM contract cycles. That inverts the usual advice on capacity commitments: waiting for the next accelerator generation no longer reliably lowers cost, because the binding input reprices on a different and less visible schedule. Trigger: Amazon attributes a $20B capex guidance increase primarily to memory prices while two suppliers describe HBM shortage persisting into 2028. Horizon: Through the 2027 contracting cycle. Who moves: Cloud buyers negotiating multi-year reserved capacity, neoclouds with fixed-price backlog, memory suppliers, and CFOs modeling AI unit cost.

STRATEGIC OUTLOOK

The useful way to read this week is that AI got cheaper and more available without anything new being built, and that should change where leaders look for advantage. When a benchmark score triples on harness settings alone, when a gigawatt changes hands through an ownership structure rather than a construction milestone, and when the cheapest tier of a model family drops 80% while the most expensive holds flat, the differentiated decisions are no longer about which model or which chip. They are about routing policy, license classification, contract structure, and memory allocation — four things that sit with legal, finance, and platform engineering rather than with a research team. The honest counterweight is that this week also produced our first strained hypothesis in several. Equinix posted record interconnection adds while interconnection revenue grew more slowly than the total business, which is the pattern we named in advance as the test that would falsify our networking thesis. We are not revising on one quarter from one operator, but we are watching it rather than explaining it away, and the Q3 colocation prints in November are the decision point. For the next quarter, three things carry the most information. CoreWeave's August 11 print is the first real read on whether neocloud backlog converts. Nvidia's August 26 guide is the first read on the Rubin ramp against AMD's newly contracted 530 MW. And the 2027 HBM settlements now under negotiation will set delivery schedules more tightly than any published roadmap. Everything else this week was structure; those three are evidence.

WHERE WE DIFFER

Our read against the field's published takes on the same week — on the record, so you can score us later.

DIFFER VentureBeat: The week's defining story is an AI price war, with OpenAI's 80% Luna cut as evidence that competition has shifted decisively toward cost.

Our read: Framing this as a price war points attention at the wrong decision. The cut left Sol untouched, so on a fixed 1:4 workload the distance between OpenAI's own cheapest and most expensive tier widened from about 5x to 25x. That makes internal routing policy the dominant cost lever and vendor choice secondary — a different action than shopping providers.

EXTEND VentureBeat: Kimi K3's weights are genuinely open, with a licensing caveat that enterprises should be aware of.

Our read: The caveat is the story, not a footnote to it. A revenue-tiered license moves open-weight diligence off the ML team and onto counsel, because the question is no longer whether the model performs but how the intended deployment classifies. Add Inkling-Small's roughly 600 GB BF16 footprint and the practical gate on open frontier models this week became legal review plus cluster capacity.

DIFFER Equinix Q2 2026 results: Record interconnection adds at Equinix confirm that interconnection is the durable, fastest-growing layer of AI infrastructure.

Our read: Record unit adds coincided with interconnection revenue growing roughly 9% while total revenue grew about 16%. That is the exact pattern this publication named in advance as the falsification test for its own networking hypothesis, so we mark the hypothesis strained rather than claim confirmation from the unit count. Units are demand; revenue mix is the test.

OPEN WSJ, confirmed by CNBC: Nvidia will guarantee up to $250B of OpenAI's data-center financing, effectively underwriting the Ohio campus.

Our read: Nvidia declined to comment, no definitive agreement exists, and the reporting rests on anonymous sources. Vendor credit-wrapping may well become the binding capital-markets product for non-investment-grade labs, which is precisely why it should be scored on a filing rather than a headline. We have written a prediction against documentation instead of booking the number.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **0 rising**, **3 falling**, 7 steady. Each metric carries a threshold value where the read materially changes.

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Frontier lab cash runway at current burn	**~30-40 months, unchanged — no lab closed primary financing in-window, while a reported ~$250B Nvidia guarantee for OpenAI's Ohio campus would move funding from equity to vendor-guaranteed debt if it is ever signed**	~30-40 months; unchanged cash estimate, but OpenAI's reported infrastructure obligation rises to $750B through 2030	→	Below 18 months for any top-four lab
Hyperscaler AI capex to disclosed AI revenue ratio	**~3.6x on the publication's committed-capital estimate, improved from ~3.7x as Azure crossed $100B annual revenue and AWS re-accelerated to +37% even while Amazon added ~$20B to CY2026 capex guidance**	~3.7x on committed capital of ~$230B against a revenue proxy of ~$62B	↓	Above 6x sustained for two consecutive quarters

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
CoreWeave contracted revenue backlog	**$99.4B as of March 31, unchanged — the Q2 print lands August 11, so backlog-to-revenue conversion stays unobserved while AWS disclosed a $496B contracted backlog for scale comparison**	$99.4B as of March 31, with the next disclosure pending	→	Sequential decline, or conversion below 15% annually
NVIDIA quarter-over-quarter data center revenue	**$75.2B for Q1 FY27, unchanged with no earnings event in-window, though Core Scientific's ~530 MW AMD commitment and the reported OpenAI financing backstop both bear on the August 26 guide**	$75.2B for Q1 FY27, +12% sequentially	→	Two consecutive quarters of sequential decline

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Open-weight to closed-model capability gap on coding	**Narrowing into a deployment fact: Kimi K3's weights shipped July 27 at an Artificial Analysis Index of 57 against roughly 60 for the closed leaders, and Inkling-Small reached 40 under Apache 2.0 at under a third of Inkling's parameters**	Widening on paper — Kimi K3's API scored 57.1 against Claude Fable 5 at 59.9 while K3's weights remained unreleased	↓	Open weights within 2 Index points of the closed leader
Sovereign AI program commitments	**~15 programs and ~$186B, unchanged — the DOE Paducah selection is roughly $100B of private capital on federal land rather than a new sovereign appropriation**	~15 programs and ~$186B in announced commitments	→	Above 20 programs or $250B committed

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
PJM capacity auction clearing price	**$325.00 per MW-day for 2028/29, unchanged with no new auction — Paducah's up to 2 GW of dedicated gas plus 2.6 GW of storage shows large load routing around auction scarcity rather than bidding into it**	$325.00 per MW-day for 2028/29, at the administrative cap	→	A second consecutive auction clearing at the cap
Time from interconnection request to energization	**60-84 months, unchanged — Paducah targets 2031 completion and Meta's 1 GW El Paso campus comes online from 2028, both multi-year despite fully committed capital**	60-84 months in the major US interconnection queues	→	Below 48 months in two or more major queues

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Cost per task, frontier reasoning model	**The cheap tier reset: GPT-5.6 Luna fell 80% to $0.20/$1.20 and DeepSeek V4-Flash-0731 gained 10 Artificial Analysis Index points at unchanged $0.14/$0.28, while Sol held at $5/$30**	Falling on list price, with Opus 5 claiming comparable quality at fewer output tokens and Gemini 3.6 Flash claiming lower per-task cost	↓	A frontier-tier reasoning model below $1 per million output tokens
Custom silicon share of hyperscaler AI compute	**~34-37%, unchanged — Amazon said its AI and Chips businesses each passed a company-reported $25B annualized run-rate without disclosing the mix, and Core Scientific's AMD deal is merchant-accelerator competition rather than in-house silicon**	~34-37% of hyperscaler AI compute on internally designed accelerators	→	Above 45% share

THRESHOLD VALUES ARE THE POINTS WHERE THE READ FLIPS. CROSSINGS ARE FLAGGED IN THE ISSUE BODY WHEN THEY HAPPEN.

PREDICTIONS

What we expect next.

6 falsifiable, time-bounded predictions. Each carries a confidence (1–99%, never 50%), a deadline, and a specific signal we'll watch. Future issues score them hit / miss / partial.

PREDICTION 01 SOFTWARE

An independent evaluator publishes completed-task cost showing GPT-5.6 Luna at least 60% cheaper per completed agentic task than GPT-5.6 Terra by September 30, 2026.

62%

BY SEPTEMBER 30, 2026

TRIGGER: A published third-party harness result comparing Luna and Terra on the same agentic task set, reporting total completed-task cost including retries, with Luna at least 60% lower.

PREDICTION 02 SOFTWARE

An independent party reproduces at least a 15-point ARC-AGI-3 improvement from harness memory and compaction settings alone, holding model weights fixed, by October 31, 2026.

44%

BY OCTOBER 31, 2026

TRIGGER: A published non-OpenAI result on the ARC-AGI-3 public set showing at least a 15 percentage-point gain attributable to retained reasoning or context compaction with the same underlying model.

PREDICTION 03 HARDWARE

SK hynix or Samsung states in a primary release or earnings transcript that 2027 HBM capacity is substantially committed or sold out by October 31, 2026.

81%

BY OCTOBER 31, 2026

TRIGGER: Company press release or official transcript containing an explicit statement that 2027 HBM supply is sold out, fully allocated, or substantially committed.

PREDICTION 04 NETWORKING

No publicly listed global colocation operator reports Q3 2026 interconnection revenue growing faster than total revenue on a normalized basis by November 15, 2026.

46%

BY NOVEMBER 15, 2026

TRIGGER: Q3 2026 results from at least two publicly listed global colocation operators, with none disclosing normalized interconnection revenue growth exceeding normalized total revenue growth.

PREDICTION 05 CAPITAL

34%

BY DECEMBER 31, 2026

A definitive agreement of at least $100B in vendor-guaranteed AI data-center financing is publicly documented in a filing or company release by December 31, 2026.

TRIGGER: An SEC filing or company press release describing an executed guarantee, credit support or backstop of at least $100B for a named AI data-center project.

PREDICTION 06 POWER

63%

BY DECEMBER 31, 2026

A power service agreement for the Paducah AI campus is filed with the Kentucky Public Service Commission by December 31, 2026.

TRIGGER: A docketed Kentucky PSC filing containing a power service agreement naming the Paducah campus and the serving utility.

TRACK RECORD

Every prediction ever made, scored against its own written trigger when the deadline passes; ambiguity resolves against us.

Predictions made	**76**	Resolved (hit / partial / miss)	**28 (6 / 13 / 9)**
Hit rate (partial = half)	**45%**	Brier score (0 = perfect)	**0.162**

Bold (<55%): none resolved yet · Core (55-80%): 28 resolved, 45% hit rate vs 66% mean confidence · High-conviction (>80%): none resolved yet

WATCHLIST

On the radar.

6 catalysts in the next 7–14 days that would change the read materially. Watching these tells us whether the thesis is strengthening or weakening.

AUG 2

WATCH 01

EU AI Act Article 50 transparency obligations become enforceable

Interaction disclosure, deepfake labelling and public-interest text labelling apply with fines up to €15M or 3% of worldwide turnover. Only machine-readable marking for systems placed on the market before August 2 gets a grace period, running to December 2.

AUG 11

WATCH 02

CoreWeave Q2 2026 results

The first neocloud print since the $99.4B March backlog and the cleanest available read on whether contracted backlog converts to revenue at the pace the category's financing assumes.

AUG 26

WATCH 03

NVIDIA Q2 FY27 results, and GitHub Copilot default-model enablement

Nvidia's guide is the first hard read on the Rubin ramp against AMD's newly contracted 530 MW. The same date is when Copilot's new default model turns on for organizations that have not set an explicit opt-out, which is an administrative deadline rather than a product choice.

BY AUG 15

WATCH 04

Independent completed-task cost for GPT-5.6 Luna

List price fell 80%, but nothing yet measures tokens consumed, retry rates or completion quality at the cheap tier. The entire routing case rests on evidence that does not exist yet.

BY SEPT 30

WATCH 05

Kentucky PSC filing for the Paducah power service agreement

The campus is contingent on regulatory approval of its power agreement. The filing will show whether dedicated generation for AI load clears on ratepayer-neutral terms, and whether curtailment rights are part of the bargain.

H2 2026

WATCH 06

2027 HBM volume and price settlements

SK hynix says 2027 volumes and prices are under negotiation with roughly ten customers now. Those contracts will set accelerator delivery schedules more tightly than any published GPU roadmap.

METHODOLOGY AND AUTHORSHIP

How this brief is built.

Compiled from public announcements, SEC filings, earnings transcripts, and official lab and vendor publications. Every quantitative claim is graded on source quality. Every prediction is falsifiable, time-bounded, and scored hit / miss / partial / pending in future issues.

SIGNAL SCORE RUBRIC

5 / 5	SEC filing or audited disclosure. Multi-source independent confirmation. Operational, not aspirational.
4 / 5	Earnings-call disclosure or primary lab/vendor announcement. Two or more independent sources.
3 / 5	Single primary source. Reasonably consistent with sector data.
2 / 5	Analyst report or secondary press only. Or single primary source with credibility caveats.
1 / 5	Rumor, social media, single non-primary source. Or contradicted by alternate primary sources.

ANYTHING GRADED 2 OR BELOW IS FLAGGED AS NOISE.

PREDICTION OUTCOMES

HIT	The specific observable outcome occurred by the deadline.
MISS	The deadline passed and the outcome did not occur.
PARTIAL	A meaningfully similar outcome occurred but not the literal wording.
PENDING	Deadline not yet reached.

HIT RATE OF 65-75% IS THE TARGET. ABOVE 80% MEANS TOO CONSERVATIVE; BELOW 50% MEANS THE FRAMEWORK IS WRONG.

AUTHORSHIP

Brian Letort

BrianLetort.AI · Independent analysis. All sources public. Not investment guidance.

THIS ISSUE

Issue 15

Week 31 of 2026 · August 1, 2026

WEB

brianletort.ai/industry

Past issues, the working framework, and the LLM Evolutionary Tree companion.