THE AI STACK WEEKLY ISSUE 13 · WEEK 29 OF 2026

THE BOTTOM LINE

The open frontier arrived priced like the closed one: Kimi K3 hit #4 on the independent index and #1 on a flagship arena at $3/$15 — while TSMC, PJM, and a first-ever state moratorium confirmed the constraints have moved fully downstream of silicon.

Moonshot AI's Kimi K3 (July 16) is the first model slated for open weights to reach the closed-frontier tier on independent evaluations — 57.1 on the Artificial Analysis Intelligence Index (#4 of 189, within three points of Claude Fable 5 and GPT-5.6 Sol) and #1 on LMArena's Frontend Code Arena at 1,679, ahead of both closed flagships. But it launched hosted-only at $0.30/$3/$15 per MTok — roughly 3x its predecessor and at parity with Claude Sonnet 5 — with weights and license promised by July 27. Add DeepSeek's first-of-its-kind peak-hour 2x surge pricing and the same-week Apache-2.0 release of Thinking Machines' 975B Inkling (weights actually live on day one), and the 'Chinese models are 10x cheaper' assumption in procurement models is over: open-weight labs now price like the frontier vendors they have become, and the differentiation is shifting to who actually ships weights. Meanwhile the physical layer told the other half of the story. TSMC posted a record $40.2B quarter with HPC at 66% of revenue, raised 2026 capex to $60-64B, added $100B to Arizona — and said advanced packaging capacity is 'so tight that now it's limiting my customers' growth.' PJM's 2028/29 capacity auction cleared at the FERC cap for the third straight year while coming up 6,831 MW short, and New York signed the first statewide data-center permit moratorium a day later. Compute itself started trading across former battle lines: Anthropic in reported talks to lease up to $10B of capacity from Meta, while GPU-backed credit institutionalized ($775M Nebius facility at SOFR+250, a

Fitch-rated $2.6B CoreWeave DDTL, the first inference-ASIC-collateralized loan). Boards should re-price single-vendor model dependency against a credible self-host option arriving July 27; investors should note credit and equity now tell opposite stories about the same GPU assets; architects and operators should treat packaging, power, and market structure — not chips or capital — as the binding constraints for 2027 planning.

AUTHOR

Brian Letort
BrianLetort.AI

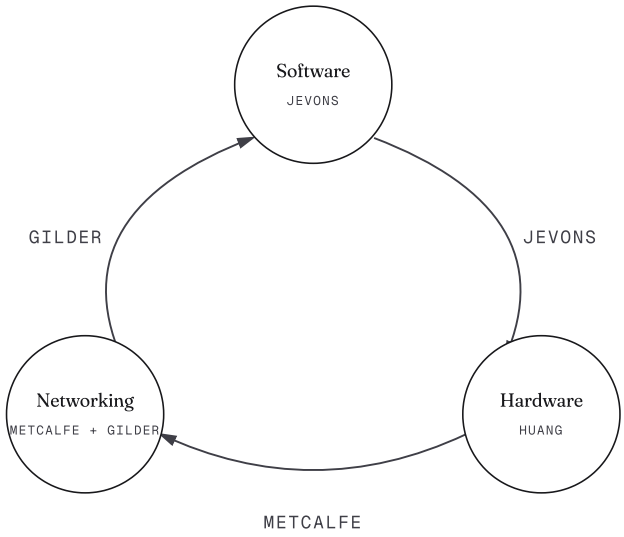
PUBLISHED

July 18, 2026
Issue 13

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 13 EVENT 01

OpenAI's GPT-5.6 family (Sol, Terra, Luna) reached GA on Amazon Bedrock via the bedrock-mantle Responses API at OpenAI first-party rates counting toward AWS commitments — with 272K context, 90% prompt-cache discounts, and Sol limited to two US East regions

AWS WHAT'S NEW; AWS BEDROCK MODEL CARDS

JUL 15 EVENT 02

Thinking Machines Lab released Inkling — a 975B-parameter (41B active) Apache 2.0 multimodal MoE with 1M context, the largest US-origin open-weights model — with weights live on Hugging Face day one, BF16 plus calibrated NVFP4 checkpoints, and day-0 vLLM/SGLang/llama.cpp support

HUGGING FACE BLOG (CO-PUBLISHED); LATENT SPACE

JUL 15 EVENT 03

OpenAI disclosed GPT-Red, an internal-only self-play RL red-teaming model that beat human red-teamers 84% vs 13% on indirect prompt-injection scenarios and hardened GPT-5.6 Sol to a 0.05% direct-injection failure rate — and said it will not be released due to offensive capability

OPENAI; THE NEW STACK; HELP NET SECURITY

JUL 16 EVENT 04

Moonshot AI launched Kimi K3 — a vendor-reported 2.8T-parameter MoE (16 of 896 experts active) with 1M context, native text/image/video input, and Kimi Delta Attention — hosted-only at $0.30/$3/$15 per MTok, with open weights promised by July 27

MOONSHOT/KIMI BLOG; SIMON WILLISON; MARKTECHPOST

JUL 16 EVENT 05

Kimi K3 debuted at 57.1 on the Artificial Analysis Intelligence Index (#4 of 189, vs Fable 5's 59.9 and GPT-5.6 Sol's 58.9) and took #1 on LMArena's Frontend Code Arena at 1,679 — the first open-weight-lab model to top an LMArena flagship board

ARTIFICIAL ANALYSIS; ARENA.AI LEADERBOARD

WHAT THIS MEANS

The Jevons arc ran hot this week: near-frontier capability got dramatically cheaper to acquire (K3 independently measured at $0.94 per Index task; Inkling free to self-host under Apache 2.0), and both flagship open releases ship 4-bit-native targeting very large self-hosted footprints — Moonshot recommends 64+ accelerator supernodes, Inkling needs ~600GB in NVFP4 — pulling inference demand toward private clusters and their memory and interconnect fabrics. Architects should hold procurement decisions on the K3 tier until the July 27 weights-and-license drop resolves, and see the Model Pulse for the full architecture read on why routing stability and prompt-cache hit rates are the new cost levers.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

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

JUL 15
EVENT 01

ASML Q2: EUR 9.3B net sales above the high end of guidance, system sales split 51% logic / 49% memory, ~65 Low-NA EUV shipments planned for 2026 (+45% YoY EUV system sales), 2027 Low-NA capacity 'close to fully covered' with +30% expansion planned

ASML Q2 2026 PRESS RELEASE AND INVESTOR CALL

JUL 15
EVENT 02

Intel Foundry became the first to ship high-volume logic patterned with High-NA EUV — Panther Lake layers on Intel 18A, dual-qualified in Oregon at yields matching the standard NXE platform; TSMC has said it won't run High-NA in production before ~2029

ASML PRESS RELEASE (JOINT WITH INTEL FOUNDRY)

JUL 15
EVENT 03

NVIDIA's Huang, in Tokyo, denied the Rubin delay reports: 'Vera Rubin is already in production. Giant amounts of production incoming' — but named no customer-delivery date; Japan's FRONTia AI factory (27,500 Rubin GPUs) was announced the next day

TOM'S HARDWARE; MULTIPLE OUTLETS

JUL 16
EVENT 04

TSMC Q2: record $40.2B revenue (+33.7% YoY), HPC at 66% of revenue, 2nm at 3% of wafer revenue in its first material quarter, FY26 capex raised to $60-64B, $100B added to Arizona — and CEO C.C. Wei: advanced packaging capacity is 'so tight that now it's limiting my customers' growth'

TSMC 2Q26 MANAGEMENT REPORT (SEC 6-K); EARNINGS CALL

JUL 15
EVENT 05

The Information: Google reportedly pitching TPUs to NVIDIA-centric neoclouds — offering to financially backstop TPU data centers and rent capacity back; Nebius, Lambda, and CoreWeave publicly demurred

THE INFORMATION VIA WINBUZZER (SECONDARY)

WHAT THIS MEANS

Both in-window prints confirmed the same structure: demand visibility is lengthening (TSMC guiding to slightly above 40% full-year growth; ASML nearly fully booked on 2027 EUV) while the binding constraint migrated downstream to advanced packaging — the CEO of the world's foundry said so on the record. Operators should treat Wei's packaging warning as the arbiter of the Rubin rack-delivery dispute, and watch Intel's High-NA production first as the first credible fork in leading-edge lithography cadence in a decade; if it converts to 14A customer wins, the Huang's-law slope gets a second supplier.

LENS 3 OF 3 THE FLYWHEEL

Networking.

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

JUL 14
EVENT 01

Tower Semiconductor announced a ~$3B (net of $1B Japan government grants) dual-track 300mm silicon photonics and SiGe capacity expansion in Japan with METI support — targeting $3.6B revenue in 2028 on optical interconnect demand

TOWER SEMICONDUCTOR PRESS RELEASE

JUL 15
EVENT 02

3M and Microsoft announced a strategic partnership making Azure the first announced hyperscale cloud to deploy 3M Expanded Beam Optical fiber connectivity across AI data centers — attacking contamination-sensitive connector install and service time

MICROSOFT SOURCE NEWSROOM; 3M INVESTOR RELEASE

JUL 16
EVENT 03

Japan's METI-backed FRONTia project launched national AI infrastructure — a 140 MW Vera Rubin AI factory (27,500 Rubin GPUs, 13,750 Vera CPUs) scaled with Spectrum-X Ethernet fabric, with ¥1T (~$6.2B) of government support over five years

NVIDIA NEWSROOM; JAPAN TIMES; NHK

JUL 16
EVENT 04

Chelsio launched its seventh-generation AI Interconnect Platform with native 400Gb Ethernet and unified RDMA (RoCEv2 + iWARP) — widening the pool of standards-based NIC endpoints that can terminate open Ethernet AI fabrics

CHELSIO COMMUNICATIONS PRESS RELEASE

WHAT THIS MEANS

The center of gravity this week was optics manufacturing capacity and deployment velocity rather than protocol wars: a $3B state-co-funded foundry bet on photonic ICs (Gilder — bandwidth supply industrializing ahead of demand) and a hyperscaler attacking the unglamorous fiber-connector bottleneck that throttles how fast bandwidth physically installs. A sovereign program standardizing on a single vendor's Ethernet fabric shows national fabrics becoming Metcalfe machines — every domestic lab and manufacturer that plugs in multiplies shared-fabric value, and deepens the lock-in debate at state scale. Watch the colo interconnect Q2 prints starting Jul 29 for the revenue confirmation.

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 210.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~1.3x

No new primary capital closed in-window; the action moved to compute sourcing — Anthropic in reported talks to lease up to ~$10B of capacity from Meta over two years (figures explicitly disputed as 'speculative' by one source), on top of its earlier $45B/3yr Colossus 1 access commitment.

Hyperscaler-Hosted

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

BURN / REV ~3.4x

Meta expanded Hyperion (Louisiana) to 5 GW and >$50B — the largest disclosed single-site AI investment — while Google was confirmed as the customer behind the 2.7 GW 'Project Tembo' campus near Cheyenne, Wyoming (scalable to 10 GW); Meta simultaneously negotiated to sell capacity to Anthropic.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

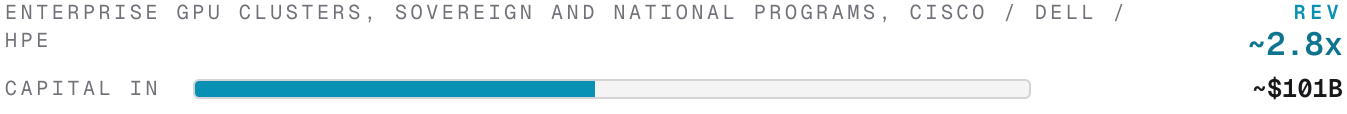
BURN / REV ~3.4x

The debt market institutionalized around AI compute in one week: Nebius closed a $775M GPU/contract-backed facility at SOFR+250 (SEC 6-K filed), Fitch rated CoreWeave's proposed $2.6B contract-backed DDTL at BB+, and General Compute secured the first inference-ASIC-collateralized facility (up to $400M against SambaNova chips) — while neocloud equity kept de-rating on Meta's market entry.

On-Prem / Hybrid

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

BURN / REV ~2.8x



Japan launched FRONTia, the first national physical-AI infrastructure program — ¥1T (~$6.2B) of government support over five years for a 140 MW / 27,500-Rubin-GPU AI factory built by a 48-company consortium — while New York moved the other direction with the first statewide moratorium on discretionary environmental permits for data centers ≥50 MW.

SIGNAL VS NOISE

What's real, what's noise.

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

5 / 5

CLAIM 01

PJM's 2028/29 capacity auction cleared at the $325/MW-day FERC cap for the third straight year — and still came up 6,831 MW short of the reliability requirement, with only 525 MW of new generation clearing.

SOURCES: PJM 2028/2029 BRA RESULTS REPORT (PRIMARY); UTILITY DIVE; MODO ENERGY

Real, and the headline understates it: the price fell 2.5% only because the cap itself fell, while PJM's own uncapped simulation cleared 71% higher ($554.72 RTO / $776.69 ComEd) — the gap between capped and simulated price has widened three auctions running. Anyone budgeting AI load in the largest US grid should watch PJM's FERC filings this month for the backstop auction and data-center connect-and-manage framework, not the headline price.

4 / 5

CLAIM 02

TSMC will spend another $100B in Arizona (total $265B) after a record Q2 — revenue $40.2B (+33.7% YoY), net income +77% YoY, FY26 capex raised to $60-64B.

SOURCES: TSMC SEC 6-K AND EARNINGS CALL (Q2 RESULTS GRADE 5); NYT, REUTERS (ARIZONA COMMITMENT)

The financial capacity is audited and extraordinary; the commitment is real but the timeline is elastic — CEO C.C. Wei explicitly tied fab timing to 'how market conditions develop,' and the announcement's political utility is part of its purpose. Treat $265B as a ceiling with option value, not a schedule; groundbreaking dates and tool orders are the confirmation to watch.

4 / 5

CLAIM 03

Kimi K3 reached the closed-frontier tier on independent evaluations — 57.1 on the Artificial Analysis Index (#4 of 189) and #1 on LMArena's Frontend Code Arena — as an open-weights model.

SOURCES: ARTIFICIAL ANALYSIS (INDEPENDENT EVAL); ARENA.AI LEADERBOARD; MOONSHOT LAUNCH BLOG

The placements are genuinely independent and real; the 'open' framing is promissory — K3 launched hosted-only, the license text is unpublished, the HF repo 404'd, and Artificial Analysis classifies it proprietary until the July 27 weight drop lands. Buyers should also note effective cost exceeds sticker: always-on max-effort thinking consumed 130M tokens on the Index eval versus a 63M peer average. Hold the procurement conclusion two weeks.

3 / 5

CLAIM 04

New York became the first state to impose a data-center moratorium, halting AI buildout in the state.

SOURCES: EO 62 TEXT (GRADE 5 FOR THE EVENT); DATA CENTER KNOWLEDGE; LAW-FIRM CLIENT ALERTS

Directionally real, rhetorically inflated: the order pauses only discretionary state environmental permits for facilities ≥50 MW, exempting completed applications, ministerial permits, and local approvals. New York is a minor AI-capacity market — the significance is precedential, landing one day after PJM confirmed the shortage. Copycat orders in Virginia, Texas, or Georgia would change the read from symbolic to structural.

2/5 — NOISE

CLAIM 05

Meta and Anthropic have a $10B, two-year compute leasing deal.

SOURCES: NYT (THREE UNNAMED SOURCES); CNBC; CNN — CNN'S SOURCE CALLS THE FIGURES 'SPECULATIVE'

Hype as priced: the talks are real (multiple outlets independently confirm talks), but the tradable 'fact' the market moved on — $10B — is explicitly disputed, terms are in flux, and no term sheet exists. A buyer-proposed monthly-payment lease with walk-away rights is an overflow-capacity option, not a strategic commitment. The direction it signals (labs renting rivals' GPUs, Meta seeking compute revenue) is real; the number is not yet.

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 'Chinese models are 10x cheaper' assumption in enterprise procurement models ended this week — Chinese open-weight labs now price like the frontier vendors they have become, and the differentiation is shifting from price to who actually ships weights.

DEDUCTIVE · 77% CONFIDENCE

Moonshot priced Kimi K3 at $0.30/$3/$15 per MTok — roughly 3x its K2.6 flagship tier and at parity with Claude Sonnet 5's $3/$15 — the same day independent evaluations placed it #4 of 189 on the AA Intelligence Index.
DeepSeek introduced the first time-of-day surge pricing from a major model API (2x listed rates during Beijing peak hours, effective Jul 24 alongside its legacy-alias retirement) — importing electricity-market demand management into inference pricing.
The counter-move came from a US lab: Thinking Machines shipped Inkling under clean Apache 2.0 with weights live on day one — while K3's weights remained promissory (license unpublished, HF repo 404, AA classifying it proprietary until the Jul 27 drop).

Compute is becoming a traded commodity across former battle lines — the market structure is shifting from vertical exclusivity to a lease-and-credit market for GPU capacity, with credit and equity now telling opposite stories about the same assets.

ABDUCTIVE · 68% CONFIDENCE

Anthropic entered reported talks to lease up to ~$10B of capacity from Meta — a top-2 lab renting a rival's fleet — days after Meta doubled Hyperion to >$50B, meaning surplus hyperscaler capex is being productized as a fifth cloud.
The debt market institutionalized in the same week: Nebius closed an SEC-filed $775M GPU/contract-backed facility at SOFR+250, Fitch rated CoreWeave's $2.6B take-or-pay DDTL at BB+, and Upper90 wrote the first inference-ASIC-collateralized loan.
Meanwhile neocloud equity kept de-rating on exactly this development (CoreWeave down ~35% since the Meta Compute report) — lenders are underwriting contracted compute cash flows at institutional spreads while equity reprices the same assets' concentration risk downward.

The binding constraints on AI deployment have fully migrated downstream of silicon and capital — in one week, the world's foundry said packaging caps its customers' growth, the largest US grid cleared short at an administratively capped price, and a state added the first policy-layer pause — while capital itself flowed unimpeded.

DEDUCTIVE · 81% CONFIDENCE

TSMC posted a record quarter, raised capex to $60-64B, and added $100B to Arizona — capital is abundant — while CEO C.C. Wei stated advanced packaging capacity is 'so tight that now it's limiting my customers' growth.'
PJM's 2028/29 auction cleared at the FERC cap for the third straight year, 6,831 MW short of the reliability requirement, with PJM's own uncapped simulation 71% higher — administrative price suppression masking a worsening shortage.
New York signed EO 62 the next day — the first statewide data-center permit moratorium — converting local opposition into a governor-signed template exactly as the grid data confirmed the scarcity that fuels it.

THESIS TEST

H1 · SUPPORTED **The cycle is accelerating, not slowing.** Two trillion-parameter-class open MoEs landed in a single week (K3 at a vendor-reported 2.8T, Inkling at 975B), and an open-weights-slated model reached within ~3 points of the closed frontier on the independent index just five weeks after GPT-5.6's government preview began — a gap-closing cadence that took DeepSeek V4 months at the start of the year. On the hardware clock, TSMC's 2nm hit its first material revenue quarter (3% of wafer revenue) while FY26 growth guidance was raised to 'slightly above 40%', and Intel shipped the first high-volume High-NA EUV logic. Both the capability clock and the silicon clock ticked faster this week.

H2 · SUPPORTED **Capital is concentrated, returns are diffuse.** The realized profits again landed at the physical layer while the model layer spent: TSMC booked a record $40.2B quarter with net income up 77% YoY and ASML beat the high end of guidance with 2027 EUV nearly fully booked — audited, banked returns — while frontier labs cut effective prices, rented rivals' compute, and Meta committed >$50B to a single site with no revenue attached. The concentration side sharpened too: two single-company sites (Hyperion, Tembo) now each approach the aggregate size of Japan's entire national AI program, and the clearest new revenue stream in the category was lease payments and debt coupons on contracted GPU capacity, not model margins.

H3 · SUPPORTED **Networking is the durable layer.** The supply side voted with capital this week: Tower Semiconductor committed ~$3B (state-co-funded) to silicon photonics capacity explicitly for optical interconnect demand, Microsoft became the first hyperscaler to deploy 3M's expanded-beam optical connectivity across AI data centers (attacking deployment velocity, not just link speed), and Japan's sovereign program standardized on a single-vendor Ethernet fabric — a state-scale Metcalfe machine. The revenue-side test (interconnect growth outpacing compute growth at colo operators) comes with the Q2 prints starting Jul 29 and is now prediction p64.

H4 · SUPPORTED **Open weights pull the floor up.** The demand-catalyst mechanism operated on cue, with a new wrinkle: Inkling shipped Apache 2.0 weights at 975B scale with day-0 serving support and a calibrated NVFP4 checkpoint that fits ~600GB — explicitly positioned as a fine-tuning base for private deployment — and Moonshot's K3 launch materials recommend 64+ accelerator supernodes for self-hosting, a direct on-prem hardware pull at frontier scale. The wrinkle is pricing: the open-weight floor is rising in capability while its API price floor rises too (K3 at closed parity, DeepSeek surge pricing), which strengthens rather than weakens the self-host re-routing thesis — the cheaper path to open-model capability is increasingly your own cluster.

H5 · SUPPORTED **Power is the binding constraint for the next 24 months.** The strongest single-week confirmation since the hypothesis was authored: PJM cleared at the administrative cap for the third consecutive year while coming up 6,831 MW short — with the operator's own uncapped simulation 71% above the clearing price and only 525 MW of new generation clearing — and New York layered the first statewide permit moratorium on top of queue physics the very next day. The constraint now has three stacked layers: physical (generation shortfall), administrative (price caps masking scarcity), and political (state-level pauses). Japan's FRONTia sizing its national AI factory at 140 MW — a rounding error against a single Meta campus — shows even sovereign programs designing around power availability.

PATTERN WATCH

Model-API pricing is converging from both directions onto a $1-3 input / $6-15 output frontier band — closed labs cutting down into it, Chinese open-weight labs raising up into it. (3 WEEKS OBSERVED)

Next week: DeepSeek V4's official GA pricing (due within days) lands at or above current V4-Pro rates rather than resetting the old ultra-cheap floor. If V4 GA undercuts to pre-K3 Chinese pricing levels, the convergence read fails and the price war resumes.

Government action is a standing gate on frontier-model availability — and the gate now includes labs gating themselves. (5 WEEKS OBSERVED)

Next week: The GPT-Red technical preprint ships within two weeks but the model does not — and no API access materializes. If OpenAI releases GPT-Red weights or endpoints in any form, the self-gating leg of this pattern breaks.

Physical-input scarcity (memory, power, now packaging) keeps marking itself to market with progressively harder money and harder words. (5 WEEKS OBSERVED)

Next week: SK hynix (Jul 29) and Samsung (Jul 30) Q2 prints disclose HBM substantially sold out or committed into 2027 with capex raises attached. Prints showing softening HBM commitments or flat memory capex would be the first hard counter-evidence in five weeks.

SECOND-ORDER EFFECTS

A fifth cloud priced against neoclouds emerges from surplus hyperscaler capex, splitting the neocloud category by funding cost: operators with bankable take-or-pay contracts refinance into cheap secured debt while the rest face equity markets that just repriced their concentration risk. Expect consolidation — the weaker half of the neocloud category becomes acquisition inventory for hyperscalers and infrastructure funds within two or three quarters. Trigger: Meta doubles Hyperion to >$50B while simultaneously negotiating to lease capacity to Anthropic, as GPU-backed credit institutionalizes at rated spreads. Horizon: Q4 2026 - H1 2027. Who moves: Neocloud equity holders and lenders, frontier labs negotiating compute, infrastructure funds, and hyperscaler corp-dev teams.

Data-center opposition movements in states with real AI load now have a governor-signed template, converting diffuse local resistance into replicable state policy — which raises the siting-risk premium on any campus without secured permits, accelerates the behind-the-meter generation and scale-across fabric workarounds, and pulls forward the value of already-permitted powered land as a distinct asset class. Trigger: New York signs the first statewide data-center permit moratorium one day after PJM clears at the cap with a 6,831 MW shortfall. Horizon: 2027. Who moves: Data-center developers without generation balance sheets, state energy regulators, utilities, and holders of permitted powered-land inventory.

If the weights and a commercial-friendly license land, data-residency-constrained enterprises and sovereign programs get their first credible exit from frontier API dependency — which re-routes inference spend from closed-lab APIs toward private accelerator clusters, HBM, and interconnect, and forces closed labs to defend on harness, tooling, and injection-hardening (exactly where OpenAI positioned GPT-Red this week) rather than raw capability. Trigger: Kimi K3 reaches the closed-frontier tier on independent evaluations with open weights promised July 27 and self-hosting guidance targeting 64+ accelerator supernodes. Horizon: 30-90 days. Who moves: Closed-lab API revenue, sovereign AI programs, enterprise infrastructure teams, and the accelerator/HBM supply chain.

STRATEGIC OUTLOOK

This week moved the 12-month posture on three fronts. First, procurement: the open-vs-closed decision is no longer about a capability gap (~3 Index points) or a price gap (K3 at closed parity) — it is about deployment control, and July 27 is the date that tells you whether frontier-class self-hosting is real; budget the evaluation now, because if the weights land clean, every closed-lab contract renewal after August negotiates against a credible internal alternative. Second, capital structure: compute is becoming a two-sided market — leased across rival lines, financed with rated secured debt — so treat GPU capacity commitments the way treasurers treat interest-rate exposure: as a portfolio of owned, leased, and optioned positions rather than a single vendor bet. Third, the constraint stack: with packaging capped at the foundry, power capped at the auction, and permits now pausable by executive order, the scarce inputs for 2027 are all downstream of money — which means the durable advantages are secured packaging allocation, energized land, and permitted sites, and the capital markets just started pricing all three accordingly.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **2 rising**, **1 falling**, 7 steady. Each metric carries a threshold value where the read materially changes.

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Frontier lab cash position (avg months runway, top 3)	**~34-37 mo; no new primary capital — the movement was compute-sourcing (Anthropic-Meta lease talks, reported ~$10B/2yr, disputed)**	~34-37 mo; Anthropic at implied $1.2T on secondaries (broker-reported), IPO calendars unchanged	→	<18 mo triggers re-rating risk
Hyperscaler capex / AI revenue ratio (top 4 weighted)	**~5.0-5.5; Meta Hyperion doubled to >$50B/5 GW, Google behind 2.7 GW Wyoming 'Project Tembo' — numerator still hardening ahead of the Jul 22-30 earnings wave**	~5.0-5.3; Meta targets 14 GW of compute in 2027 (2x 2026) on $125-145B capex guidance	↑	>6.0 invites investor pushback at next earnings
CoreWeave revenue backlog	**$99.4B as of Mar 31 (+284% YoY), restated in Fitch's Jul 16 DDTL note; equity down ~35% since the Meta Compute report**	~$100B reported; Helios Phase I (133 MW) delivered on schedule — backlog now converting to lease revenue	→	Conversion velocity matters more than gross figure

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
NVIDIA Q-over-Q data center revenue	**$75.2B Q1 FY27 (unchanged); Huang in Tokyo: 'Vera Rubin is already in production' — delay narrative rebutted, but no customer-delivery date named**	$75.2B Q1 FY27; Q2 guide $91B (reports Aug 26); SemiAnalysis sees H2 ~20% above consensus despite Kyber dispute	→	Q2 FY27 guide $91B implies further +21% QoQ
Open vs closed gap on coding (SWE-Bench / agentic)	**~3 pts on the AA Intelligence Index — Kimi K3 at 57.1 vs Fable 5's 59.9 — with K3 weights pending Jul 27; K3 already #1 on LMArena Frontend Code Arena**	Effectively closed on cost-quality: GLM 5.2 statistically tied with Opus 4.8 at $1.28 vs $1.94/task (Databricks); Hy3 adds Apache-2.0 agentic-search lead	↓	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**~15 / ~$186B; Japan FRONTia adds ¥1T (~$6.2B/5yr) — the first national program explicitly scoped to physical AI and robotics**	~14 / ~$180B+ (flat; Meta Alberta and MARA Texas are corporate capital on power-rich land, not sovereign programs)	↑	—

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
PJM 2026/27 capacity auction price ($/MW-day)	**$325.00 — the 2028/29 BRA cleared at the (lowered) FERC cap Jul 14, 6,831 MW short of the reliability requirement; PJM's uncapped simulation: $554.72 RTO / $776.69 ComEd**	$329.17; 2028/29 BRA bids closed Jul 7 — results post Jul 14 after 4 p.m. ET (prediction p54 resolves)	→	11x in 24 months — power is the new binding constraint
Time-to-power, busiest US markets (months)	**60-84 unchanged; New York adds the first statewide ≥50 MW discretionary-permit moratorium — state policy now stacks on top of queue physics**	60-84; hyperscalers routing around queues — Meta fully funds its own generation in Alberta because the grid cannot host multiple large loads	→	—
Cost-per-task, frontier reasoning model	**~$0.06-$0.12 effective unchanged at the floor; new AA Index task costs: DeepSeek V4 Pro $0.04, GLM-5.2 $0.32, K3 $0.94, Sol $1.04 — and DeepSeek adds the first peak-hour surge pricing**	~$0.06-$0.12 effective; GPT-5.6 GA tiering (Sol $5/$30 / Terra $2.50/$15 / Luna $1/$6), Grok 4.5 at $2/$6 — and harness choice swings per-task cost 2x+	→	DeepSeek's 2x Beijing-peak-hours surcharge (from Jul 24) imports electricity-market demand management into model APIs

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Custom silicon share of incremental AI compute	**~34-37% unchanged; Google reportedly pitching TPUs to NVIDIA-centric neoclouds with financial backstops — custom silicon escalating from internal cost play to merchant distribution fight**	~34-37%; Meta's Iris enters production in September, Broadcom-Apple extended through 2031 (8-K), AWS raising Trainium 3 orders 20-30%	→	>35% materially compresses merchant GPU pricing

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

PREDICTIONS

What we expect next.

5 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

Moonshot publishes Kimi K3 open weights on Hugging Face with a license permitting commercial self-hosting by August 10, 2026 (vendor-committed 'by July 27', with slippage buffer).

72%

BY AUGUST 10, 2026

TRIGGER: Kimi K3 weights live on Hugging Face with published license text; Artificial Analysis reclassifies K3 from proprietary to open-weights.

PREDICTION 02 SOFTWARE

DeepSeek retires its legacy deepseek-chat and deepseek-reasoner aliases on July 24 as scheduled and ships DeepSeek V4 to official GA by July 31, 2026, with peak-hour surge pricing in effect.

76%

BY JULY 31, 2026

TRIGGER: DeepSeek API docs showing V4 GA model IDs and the alias-retirement notice executed; surge pricing live in the rate card.

PREDICTION 03 HARDWARE

SK hynix's July 29 Q2 earnings call discloses that 2027 HBM capacity is substantially sold out or committed under long-term agreements, extending the memory-scarcity trade into a second year.

68%

BY JULY 29, 2026

TRIGGER: SK hynix Q2 2026 earnings call commentary on 2027 HBM capacity commitments and capex.

PREDICTION 04 NETWORKING

The largest colocation operators' Q2 prints (starting July 29) show interconnect/fabric revenue growth again outpacing overall revenue growth, sustaining the Metcalfe read for a second consecutive quarter.

71%

BY AUGUST 15, 2026

TRIGGER: Q2 2026 colo earnings disclosures: interconnection and fabric revenue growth rates vs total revenue growth.

PREDICTION 05 POWER

At least one additional US state announces a moratorium, discretionary-permit pause, or equivalent statewide restriction on large data-center development by October 31, 2026, following New York's EO 62.

57%

BY OCTOBER 31, 2026

TRIGGER: A governor's executive order or enacted state legislation pausing or restricting ≥50 MW-class data-center permitting in a second state.

WATCHLIST

On the radar.

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

JUL 20-24
WATCH 01

OpenAI's GPT-Red technical preprint

Promised 'later this week' — enables third-party scrutiny of the 84%-vs-13% red-teaming and 0.05% injection-failure claims that are currently vendor-reported only. If it substantiates, injection-hardening becomes a hard procurement criterion industry-wide.

JUL 22-23
WATCH 02

Q2 earnings triple-header: ServiceNow and Alphabet (Jul 22), then SAP, Intel, and AMD's Advancing AI event (Jul 23)

First reads on agentic-AI monetization (ServiceNow), cloud AI revenue (Alphabet), the closed-gateway Joule strategy (SAP), 18A economics post-High-NA (Intel), and MI455X/Helios rack shipment clarity (AMD).

JUL 24
WATCH 03

Two hard deadlines: M365 Copilot's OpenAI-subprocessor setting auto-enables, and DeepSeek retires its legacy API aliases

M365 tenants that don't opt out are auto-enrolled in a changed data-processing chain (silence is consent); DeepSeek's alias retirement plus surge pricing is a forced-migration event for one of the most-called model APIs.

JUL 27
WATCH 04

Kimi K3 open-weights drop, license text, and technical report

The single highest-information event in view: if the weights and a commercial-friendly license land, frontier-class capability becomes self-hostable for the first time and prediction p61 resolves early; if they slip or the license restricts, the 'open frontier' narrative resets.

JUL 29-30
WATCH 05

Memory and hyperscaler prints: SK hynix Q2 (Jul 29), Microsoft FY26 Q4 (Jul 29), Samsung divisional results (Jul 30), colo interconnect Q2 prints begin (Jul 29)

The supply-side read on 2027 HBM commitments (p63), whether hyperscaler capex guides absorb this week's Hyperion/Tembo adds, and the interconnect-revenue test of the networking hypothesis (p64).

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 13

Week 29 of 2026 · July 18, 2026

WEB

brianletort.ai/industry

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