THE AI STACK WEEKLY ISSUE 09 · WEEK 25 OF 2026

THE BOTTOM LINE

Power policy and open weights became the week's leverage while the closed frontier stalled.

W25 moved at the two ends of the stack that the AI factory actually depends on — the grid and the open model layer — while the closed frontier marked time. The single most consequential event was regulatory: on Jun 18 FERC issued six simultaneous Section 206 show-cause orders telling PJM, MISO, SPP, CAISO, ISO-NE, and NYISO their tariffs 'appear unjust and unreasonable' for large data-center loads, forcing co-location/behind-the-meter rules and cost-shift transparency within 60 days. Developers kept routing around the interconnection queue entirely (Cummins–Circe's 2GW behind-the-meter gas) while hyperscalers added grid-connected capacity explicitly paying 100% of interconnection cost (Amazon's $10B Missouri campus, Google's $1.5B Alabama expansion). At the model layer the story was openness and stasis: Z.ai shipped GLM-5.2 under a genuine MIT license, which independent testing says beats GPT-5.5 on several long-horizon coding benchmarks at roughly one-sixth the cost; MiniMax-M3's sparse-attention weights matured; and Artificial Analysis rebased its Intelligence Index to v4.1 around agentic tasks — all while Anthropic's Claude Fable 5 stayed government-suspended the entire week (Opus 4.8 remained the working leader) and Gemini 3.5 Pro slipped to July. Hardware and networking were quiet (Supermicro opened Vera Rubin NVL72 order availability; Coherent expanded indium-phosphide laser capacity). Net/net: boards should treat power policy as a board-level siting and cost risk now being repriced by FERC; investors should watch the open-weight cost curve collapsing frontier-adjacent capability toward commodity inference; architects should pilot MIT-licensed GLM-5.2 for sovereign and self-host coding while keeping multi-model fallbacks given the closed frontier's demonstrated fragility; operators should lock generation and long-lead transformers before GPUs.

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

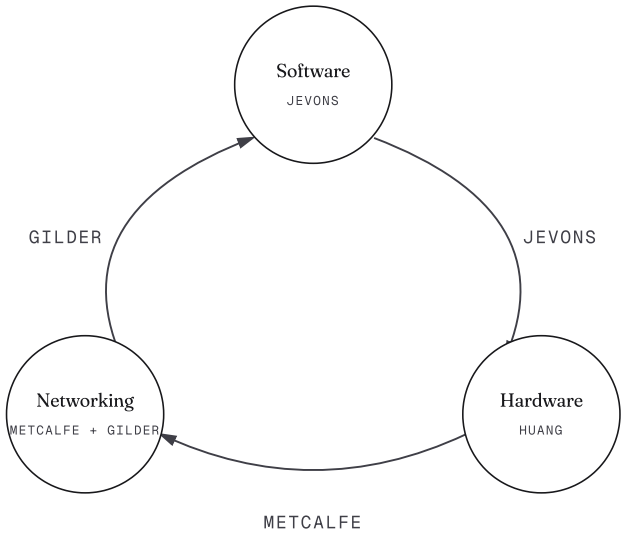
June 20, 2026

Issue 09

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.

JUN 16
EVENT 01

Z.ai released GLM-5.2 under an MIT license (~744B / ~40B-active sparse-attention MoE, 1M context); independent testing reports it beats GPT-5.5 on several long-horizon coding benchmarks at ~1/6 the cost, and Artificial Analysis cites it as the leading open-weight model

Z.AI BLOG; VENTUREBEAT; HUGGING FACE

JUN 15
EVENT 02

Artificial Analysis rebased its Intelligence Index to v4.1, re-weighting the industry's headline benchmark around agentic tasks (GDPval-AA v2, Terminal-Bench, banking agents); scores are not back-comparable to v4.0

ARTIFICIAL ANALYSIS

JUN 15-21
EVENT 03

Anthropic's Claude Fable 5 and Mythos 5 remained government-suspended for the full week with no restoration; Opus 4.8 stayed the working closed-frontier leader as queries continued to fall back to it

ANTHROPIC

JUN 18
EVENT 04

Gemini 3.5 Pro slipped its GA from June to July: still a limited Vertex preview with no public model card, pricing, or independent benchmark vs Fable 5/Opus 4.8

BUSINESS INSIDER; GOOGLE

WHAT THIS MEANS

The closed frontier paused — no GA from OpenAI, Google, or a restored Anthropic — and the momentum shifted to open weights and to how 'best' is measured. Architects should pilot MIT-licensed GLM-5.2 for self-host/sovereign coding now and re-baseline on the agentic AA v4.1 index; see The Model Pulse for the full architecture read.

LENS 2 OF 3

THE FLYWHEEL

Hardware.

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

JUN 17
EVENT 01

Supermicro opened customer engagement / order availability for Vera Rubin NVL72 and HGX Rubin NVL8 reference designs (1,152-GPU scalable unit, 331TB HBM4), though deployments remain H2 2026 at GA — order availability, not shipments

THE DATA CENTER ENGINEER (SUPERMICRO DCBBS)

JUN 16
EVENT 02

HPE Discover expanded Private Cloud AI with NVIDIA Vera-CPU servers (ProLiant DL394) and the NVIDIA Agent Toolkit — on-prem agentic-AI enablement rather than new silicon or shipments

SILICONANGLE

JUN 18
EVENT 03

A J.P. Morgan note pegged the 2026 custom-ASIC market at ~$60-70B with Broadcom at 80-85% and Marvell 10-12% share, reinforcing the ASIC-overtakes-merchant-GPU-units thesis for 2027

J.P. MORGAN NOTE (FINVAULTA SUMMARY)

WHAT THIS MEANS

Hardware was a follow-through week: the marquee catalysts (Micron earnings, SK hynix HBM4 dynamics, NVIDIA ISC, Groq's raise) all landed Jun 22-24, just outside the window. In-window the signal was OEM/system enablement (Supermicro order availability, HPE Vera CPU) and the custom-silicon thesis hardening. Operators should treat Vera Rubin as orderable but not shippable until H2; investors should keep diversifying toward the Broadcom/Marvell ASIC duopoly.

LENS 3 OF 3 THE FLYWHEEL

Networking.

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

JUN 16
EVENT 01

Coherent broke ground on a Sherman, TX indium-phosphide fab expansion and signed a $50M CHIPS LOI to roughly quadruple InP wafer output — the lasers feeding 800G-to-1.6T transceivers and NVIDIA CPO external laser sources

COHERENT; NVIDIA BLOG; THE REGISTER

JUN 17
EVENT 02

HPE Juniper began shipping the QFX5250-64OE-L, an industry-first 1.6T, fully liquid-cooled 102.4T switch (64x 1.6T ports, ORV3-compliant) built on Broadcom-class silicon

JUNIPER COMMUNITY BLOG

JUN 17
EVENT 03

Marvell published a plasmonics optical-roadmap (with Polariton/ETH Zurich) citing a 1.1THz modulator lab record and 400G/lane SiPho devices — a forward R&D signal, not a commercial design win

MARVELL BLOG

WHAT THIS MEANS

Networking was quiet between the March (OFC) and June-1 (earnings) news clusters: the week's signal was supply-chain capacity (Coherent's InP expansion addressing the laser chokepoint) plus a liquid-cooled 1.6T system, not a new chip design win. Critically, no second non-Broadcom vendor (Marvell or Credo) cited a fresh 1.6T/CPO production win — architects pressing for a second silicon source still have to wait.

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 187.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~1.3x

No new frontier-lab round closed; the action moved to open weights (GLM-5.2 MIT) and to Fable 5 staying government-suspended, not to lab balance sheets.

Hyperscaler-Hosted

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

BURN / REV ~0.3x

Amazon ($10B Missouri) and Google ($1.5B Alabama) added grid-connected capacity, both explicitly paying 100% of interconnection cost amid FERC cost-shift scrutiny.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

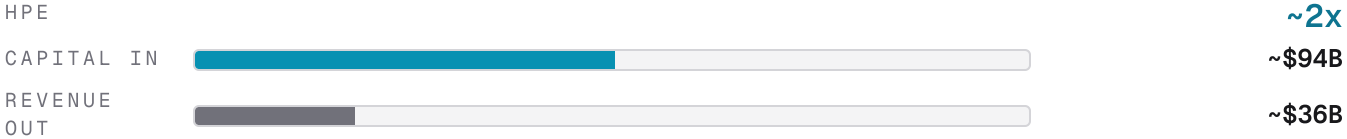
BURN / REV ~3x

CoreWeave's backlog ticked to ~$100B with a Cantor estimate of ~$131B by end-Q2; Meta–Crusoe's reported ~1.6GW deal stayed unconfirmed.

On-Prem / Hybrid

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

BURN / REV ~2x



FERC's six show-cause orders put national interconnection rules in play; Cummins–Circe committed 2GW of behind-the-meter gas, and TensorX/Solstice lined up up-to-$1B of EU sovereign-AI financing.

SIGNAL VS NOISE

What's real, what's noise.

5 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

FERC issued six simultaneous Section 206 show-cause orders rewriting large-load/data-center interconnection across all US RTOs/ISOs.

SOURCES: FERC PRESS RELEASE AND ORDERS; DAY PITNEY AND AKIN/JDSUPRA LEGAL ANALYSES. PRIMARY REGULATORY ACTION, MULTI-SOURCE.

SIGNAL. The rules for how AI campuses connect and who pays are being rewritten nationally on a 60-day clock. This is a board-level siting and cost variable — pro-speed medium-term, but it injects near-term regulatory uncertainty into every US data-center plan.

4/5

CLAIM 02

Z.ai's GLM-5.2 shipped MIT open weights that beat GPT-5.5 on several long-horizon coding benchmarks at ~1/6 the cost and lead the open-weight field.

SOURCES: Z.AI RELEASE AND HUGGING FACE CARD (PRIMARY); VENTUREBEAT INDEPENDENT TESTING; ARTIFICIAL ANALYSIS PLACEMENT. CAVEAT: SWE-BENCH/AIDER NUMBERS NOT YET FULLY SETTLED.

SIGNAL. A genuinely permissive (MIT, no regional limits) frontier-adjacent open model is now self-hostable at a fraction of closed-API cost. Architects should pilot it for sovereign/self-host coding; it pressures closed-flagship pricing.

4/5

CLAIM 03

Anthropic's Claude Fable 5 and Mythos 5 remained globally suspended for the entire week following the Jun 12 government order.

SOURCES: ANTHROPIC NEWS PAGE (PRIMARY) PLUS SECONDARY TRACKERS; NO RESTORATION DATE ANNOUNCED.

SIGNAL. The sovereign/regulatory single-point-of-failure on a top-tier closed model is not a one-day outage — it persisted a full week. Keep Opus 4.8/Sonnet fallbacks wired and do not single-source the frontier for production-critical paths.

2/5 — NOISE

CLAIM 04

GPT-5.6 (OpenAI) and Grok V9 (xAI) are launching imminently and will reset the frontier.

SOURCES: BACKEND CODENAMES, TRANSIENT ROUTING STRINGS, AND PREDICTION-MARKET ODDS ONLY; NO SYSTEM CARD OR API CONTRACT. GPT-5.5 REMAINS THE DOCUMENTED FLAGSHIP; 'GROK V9' DOES NOT EXIST.

NOISE. Rumor with no primary artifact. Keep current baselines and do not delay procurement on an unreleased SKU; any GPT-5.6 was tracking to land after this window.

2/5 — NOISE

CLAIM 05

Meta has signed a ~1.6GW compute deal with Crusoe.

SOURCES: BLOOMBERG-SOURCED REPORTING (JUN 18); BOTH META AND CRUSOE DECLINED TO COMMENT, TERMS UNDISCLOSED.

NOISE / unconfirmed. Treat as reported, not closed. The structural pattern (Meta adding external behind-the-meter compute) is real, but the specific deal is not yet verifiable for a procurement or investment decision.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **2 rising**, **2 falling**, 6 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 (flat; no new in-window round)**	~34-37 mo (xAI now public via SpaceX IPO)	→	<18 mo triggers re-rating risk
Hyperscaler capex / AI revenue ratio (top 4 weighted)	**~5.0-5.3 (Amazon $10B + Google $1.5B added; top-4 guides flat)**	~5.0-5.3 (Oracle FY27 ~$70B guide; top-4 flat)	→	>6.0 invites investor pushback at next earnings
CoreWeave revenue backlog	**~$100B (Jun 15); Cantor estimate ~$131B by end-Q2**	$99.4B (next print Q2, early Aug)	↑	Conversion velocity matters more than gross figure
NVIDIA Q-over-Q data center revenue	**$75.2B (Q1 FY27); Q2 guide $91B, reports Aug 26**	$75.2B (Q1 FY27); Q2 guide $91B, reports Aug 26	→	Q2 FY27 guide $91B implies further +21% QoQ
Open vs closed gap on coding (SWE-Bench / agentic)	**Open closing fast: GLM-5.2 (MIT) reportedly beats GPT-5.5 on long-horizon coding at ~1/6 cost; AA rebased to v4.1**	Closed lead widened on paper (Fable 5 80.3% vendor scaffold) vs open ~58.6%; contested	↓	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**~14 / ~$180B+ (flat; TensorX/Solstice up-to-$1B EU facility is capacity, not drawn)**	~14 / ~$180B+ (NAVER+NVIDIA 1GW added)	→	—

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
PJM 2026/27 capacity auction price ($/MW-day)	**$329.17; 2028/29 BRA results expected ~July 7 (pending)**	$329.17 (2028/29 BRA clears Jun 30-Jul 7)	→	11x in 24 months — power is the new binding constraint
Time-to-power, busiest US markets (months)	**60-84; FERC Jun 18 show-cause may compress large-load study timelines**	60-84 (new PJM); 2027/28 auction cleared short of target	→	—
Cost-per-task, frontier reasoning model	**~$0.10-$0.15 (effective); open weights (GLM-5.2) push commodity capability to ~1/6 frontier cost**	~$0.10-$0.15 (effective); Fable 5 sets a 2x-Opus top tier ($10/$50 per MTok)	↓	GLM-5.2 API ~$1.40/$4.40 per MTok; Grok 4.3 on Bedrock $1.25/$2.50
Custom silicon share of incremental AI compute	**~33-36%; J.P. Morgan pegs 2026 custom-ASIC TAM ~$60-70B (Broadcom 80-85%)**	~33-36%; d-Matrix Corsair inference ASIC in volume production	↑	>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.

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 MIT- or Apache-licensed open-weight model (e.g., GLM-5.2) enters the overall top 5 of the Artificial Analysis Intelligence Index v4.1 — not just the open-weight subset — by August 31, 2026.

58%

BY AUGUST 31, 2026

TRIGGER: Artificial Analysis Intelligence Index v4.1 leaderboard update.

PREDICTION 02 HARDWARE

By August 31, 2026, all three HBM makers (SK hynix, Samsung, Micron) confirm HBM fully allocated for 2026 and/or 2027 price increases.

75%

BY AUGUST 31, 2026

TRIGGER: Earnings calls or supply-chain reporting (TrendForce, Bloomberg/Reuters) from the three memory vendors.

PREDICTION 03 NETWORKING

A second non-Broadcom vendor (Marvell or Credo) cites a 1.6T or co-packaged-optics production design win by September 30, 2026.

55%

BY SEPTEMBER 30, 2026

TRIGGER: Earnings call, product release, or customer design-win disclosure.

PREDICTION 04 POWER

At least one RTO/ISO files a large-load interconnection compliance proposal answering FERC's Jun 18 show-cause orders by the August 17, 2026 deadline.

80%

BY AUGUST 17, 2026

TRIGGER: FERC docket filings from PJM, MISO, SPP, CAISO, ISO-NE, or NYISO.

PREDICTION 05 POWER

A named hyperscaler announces a >1GW behind-the-meter or off-grid generation deal for AI data centers by August 31, 2026.

68%

BY AUGUST 31, 2026

TRIGGER: Hyperscaler energy/data-center announcement; utility or developer disclosure.

PREDICTION 06 CAPITAL

At least one major closed lab cuts flagship API prices or ships a cheaper tier by August 31, 2026, in response to open-weight cost pressure.

55%

BY AUGUST 31, 2026

TRIGGER: Lab pricing page or API changelog (OpenAI, Anthropic, Google).

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.

JUN 22-JUL 7

WATCH 01

PJM 2028/29 base residual capacity auction results

After two near-cap clears and a 2027/28 reliability shortfall, the 2028/29 results (collar near $325/MW-day) are the single most important power-cost signal of the quarter and will reprice US data-center economics.

BY AUG 17

WATCH 02

RTO/ISO responses to FERC's Section 206 show-cause orders

FERC gave six grid operators 60 days to fix large-load interconnection rules. The compliance filings will define how AI campuses connect and who pays — a board-level siting variable for every US deployment.

JUN 22-24 (W26 CATALYSTS)

WATCH 03

Project Kilby, Micron/SK hynix HBM4, Cerebras/Groq

A cluster of marquee items landed just after this window: Microsoft-Chevron's 2.67GW off-grid gas deal, Micron's record HBM-allocated quarter, SK hynix's HBM4-vs-DDR5 pivot, and Cerebras/Groq milestones. They will anchor next week's read.

JULY

WATCH 04

Gemini 3.5 Pro GA and Fable 5 restoration

Both slipped: Pro moved to July and Fable 5 stayed suspended. Their resolution will decide whether the closed frontier re-takes the lead from the surging open-weight field.

JUN-AUG

WATCH 05

Open-weight enterprise adoption (GLM-5.2)

Downloads, integrations, and independent SWE-Bench replication for GLM-5.2 and MiniMax-M3 will show whether MIT/permissive open weights become production substrate and force closed-flagship price cuts.

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 09

Week 25 of 2026 · June 20, 2026

WEB

brianletort.ai/industry

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