THE AI STACK WEEKLY ISSUE 07 · WEEK 23 OF 2026

THE BOTTOM LINE

The AI factory became a power-and-fabric problem, not a model-release problem.

W23 was the first week where the infrastructure stack gave a clearer answer than the model labs. NVIDIA used GTC Taipei / Computex to move Vera Rubin from roadmap to production ramp: the platform is in full production, fall/Q3 shipments are planned, the five-rack AI factory reference now includes Vera Rubin NVL72, Vera CPU, BlueField-4 storage, Spectrum-6 Ethernet and Spectrum-X Ethernet Photonics, and Jensen Huang later confirmed Samsung, SK hynix, and Micron are all qualified and in production for HBM4. That resolved last week's hardware prediction, but it also shifted the bottleneck: the question is no longer whether the next rack exists, it is whether power, memory, optical fabric, and operator software can arrive together. On software, the closed frontier was quiet — Gemini 3.5 Pro still had not GA'd by the end of the window — while open weights widened in the efficient-agent layer: JetBrains Mellum2, NVIDIA Cosmos 3, and Holo3.1 all targeted deployable sub-agents, physical-AI reasoning, or local computer-use rather than a monolithic chatbot benchmark. On applications, Microsoft Scout, Salesforce Coworker, ServiceNow Otto, Wordsmith, and Stilta all pointed at the same control-plane fight: governed agents with identities, permissions, and workflow authority. Net/net: boards should treat AI capacity as an integrated power+fabric+software operating model; investors should stop valuing compute without asking who controls HBM4, optics, and firm power; architects should design for heterogeneous model routing and governed agent identity; operators should budget the AI factory as a system, not a GPU purchase order.

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

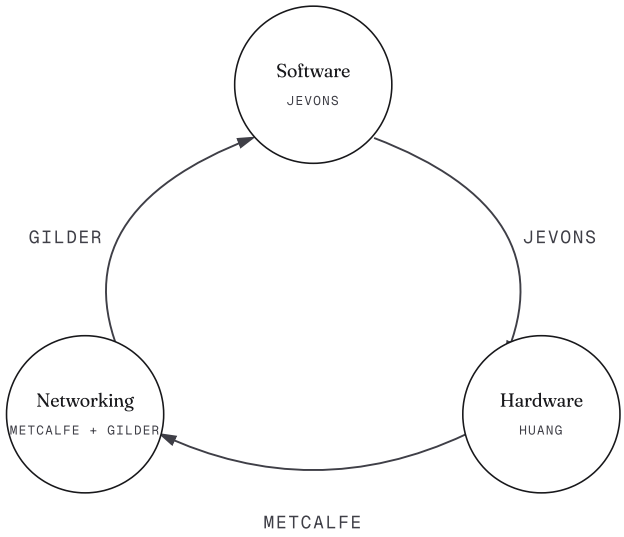
June 6, 2026

Issue 07

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 1
EVENT 01

JetBrains released Mellum2, an Apache-2.0 12B sparse MoE with 2.5B active parameters per token, positioned for low-latency routing, RAG, summarization, validation, sub-agents, and private text/code deployments

HUGGING FACE JETBRAINS MELLUM2 LAUNCH

JUN 1
EVENT 02

NVIDIA released Cosmos 3 on Hugging Face as an open omni-model for physical-AI reasoning and action, with Nano 16B and Super 64B variants plus Diffusers integration and synthetic-data workflows

HUGGING FACE NVIDIA COSMOS 3 LAUNCH

JUN 2
EVENT 03

H Company released Holo3.1 for local computer-use agents, adding 0.8B / 4B / 9B / 35B-A3B sizes plus FP8, Q4 GGUF and NVFP4 checkpoints for private deployment

HUGGING FACE HOLO3.1 LAUNCH

WHAT THIS MEANS

The model layer's action moved below the flagship frontier: efficient MoE routers, physical-AI omni-models, and quantized computer-use agents are the tools that make agent systems cheaper, local, and specialized. Architects should route cheap sub-agent work to open/local models and reserve Opus/GPT/Gemini-class spend for high-risk reasoning, because the software flywheel is now about orchestration economics as much as raw intelligence.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

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

JUN 1
EVENT 01

NVIDIA announced Vera Rubin is in full production, with a five-rack platform spanning Vera Rubin NVL72, Vera CPU, BlueField-4 STX storage, Spectrum-6 SPX Ethernet, and partner manufacturing across 350+ factories and 30 countries

NVIDIA NEWSROOM, GTC TAIPEI

JUN 1
EVENT 02

GTC Taipei positioned DSX OS as the lifecycle, health, resiliency, and multi-tenant operating layer for AI factories, shifting attention from rack shipment to fleet operations

DATA CENTER KNOWLEDGE GTC TAIPEI COVERAGE

JUN 5
EVENT 03

Jensen Huang confirmed Samsung, SK hynix, and Micron are all qualified and in production for Vera Rubin HBM4, resolving the near-term supplier uncertainty around the Q3/fall ramp

TECHTIMES SUMMARY OF REUTERS/BLOOMBERG REMARKS

WHAT THIS MEANS

The hardware read changed from 'will Rubin be on schedule?' to 'can the whole AI factory be delivered as a coordinated system?' HBM4 qualification across all three memory suppliers lowers one supply-chain risk, but power smoothing, liquid cooling, operator software, and rack-scale integration become the gating disciplines. Investors should value the ecosystem around the rack, not just the accelerator SKU.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

JUN 1
EVENT 01

NVIDIA said Spectrum-X Ethernet Photonics, a CPO-based switch platform with 200Gb/s SerDes, is now in production as part of the Vera Rubin AI factory fabric

NVIDIA NEWSROOM

JUN 3
EVENT 02

Marvell framed CPO and 1.6T optical DSPs as the next AI connectivity bottleneck, citing a CPO switch design, 100T Ethernet switch work, and NVIDIA partnership around optics, photonics, and NVLink Fusion

DATACENTERNEWS ASIA

JUN 6
EVENT 03

Broadcom reported AI semiconductor revenue up 143% YoY, with networking nearly 40% of AI revenue and demand for XPUs plus networking described as insatiable

SDXCENTRAL BROADCOM Q2 FY2026 EARNINGS COVERAGE

WHAT THIS MEANS

Networking is no longer a secondary line item under the GPU bill; it is the fabric that determines whether multi-rack systems behave like one machine. The week put CPO, 1.6T/3.2T optics, and AI Ethernet economics into the same frame as HBM4. Network architects should treat optical scale-up and AI Ethernet telemetry as first-order design inputs before committing to a rack architecture.

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 181.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~1.3x

No new mega-round closed; the action moved from lab balance sheets to the infrastructure stack those labs consume.

Hyperscaler-Hosted

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

BURN / REV ~0.3x

Google's power-first Texas campus made energy procurement the visible hyperscaler battleground.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

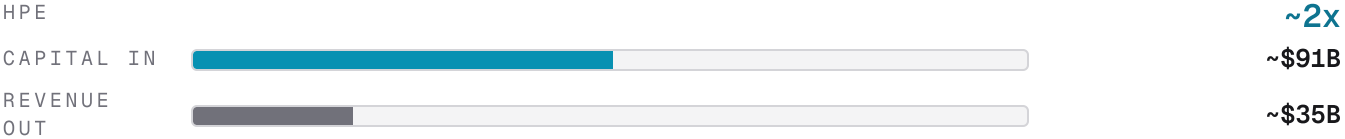
BURN / REV ~3x

No new W23 financing reset; prior IREN/Microsoft-style deals remain the relevant neocloud proof point.

On-Prem / Hybrid

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

BURN / REV ~2x



Sovereign AI infrastructure moved from compute ambition to power-first site selection and public-consent risk.

SIGNAL VS NOISE

What's real, what's noise.

4 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

Vera Rubin is in full production and NVIDIA named a fall/Q3 shipment path for the next AI-factory platform.

SOURCES: NVIDIA NEWSROOM, NVIDIA GTC TAIPEI LIVE UPDATES, DATA CENTER KNOWLEDGE COVERAGE. CAVEAT: VENDOR ANNOUNCEMENT, NOT CUSTOMER ACCEPTANCE DATA.

SIGNAL. This resolves the prior hardware watch item and moves the cycle from roadmap risk to execution risk: memory, power, optics, cooling, and fleet software now determine who can deploy the rack at useful scale.

4/5

CLAIM 02

All three HBM4 suppliers are qualified and in production for Vera Rubin.

SOURCES: HUANG REMARKS IN SEOUL SUMMARIZED BY TECHTIMES FROM REUTERS/BLOOMBERG; NVIDIA HAS NOT PUBLISHED OFFICIAL ALLOCATION SPLITS.

SIGNAL with allocation caveat. Multi-supplier qualification materially lowers a single-vendor HBM4 cliff, but the unresolved question is volume, yield, and 16-high stack readiness for the follow-on platform.

2/5 — NOISE

CLAIM 03

Gemini 3.5 Pro has launched and already displaced Opus 4.8 on public benchmarks.

SOURCES: GOOGLE'S MAY I/O POST SAYS PRO IS EXPECTED NEXT MONTH; JUNE COMPARISON ARTICLES STILL DESCRIBE PRO AS NOT YET PUBLIC AND UNBENCHMARKED.

NOISE for this window. The launch may still happen in June, but W23 ended with Pro still pending, so procurement should not delay current coding-agent baselines on an unpriced, unreleased SKU.

4/5

CLAIM 04

Enterprise application vendors are converging on governed autonomous agents with identity, permissions, and workflow authority.

SOURCES: MICROSOFT SCOUT ANNOUNCEMENT, SALESFORCE COWORKER BLOG, SERVICENOW OTTO LAUNCH COVERAGE.

SIGNAL. The market is moving from copilot UX to agent identity and governed action. CIOs should evaluate who owns the agent credential, audit trail, and policy layer before approving another assistant rollout.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **2 rising**, **0 falling**, 8 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)	**~33-36 mo**	~33-36 mo	→	<18 mo triggers re-rating risk
Hyperscaler capex / AI revenue ratio (top 4 weighted)	**~5.0-5.2**	~5.0-5.2	→	>6.0 invites investor pushback at next earnings
CoreWeave revenue backlog	**$99.4B**	$99.4B	→	Conversion velocity matters more than gross figure
NVIDIA Q-over-Q data center revenue	**$75.2B (Q1 FY27); Rubin production ramp confirmed**	$75.2B (Q1 FY27)	↑	Q2 FY27 guide $91B implies further +21% QoQ
Open vs closed gap on SWE-Bench Pro (coding)	**Closed +~19pp (no new Pro challenger yet)**	Closed +~19pp (audit caveat)	→	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**~13 / ~$160B+; power-first gating rising**	~13 / ~$160B+	→	—
PJM 2026/27 capacity auction price ($/MW-day)	**$329.17**	$329.17	→	11x in 24 months — power is the new binding constraint
Time-to-power, busiest US markets (months)	**60-84 (new PJM); power-first campuses rising**	60-84 (new PJM); 36-48 (existing PJM queue)	→	—

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Cost-per-task, frontier reasoning model	**~$0.10-$0.15 (effective; unchanged)**	~$0.10-$0.15 (effective)	→	Opus 4.8 fast mode dropped ~3x but no verifiable per-task reading in-window
Custom silicon share of incremental AI compute	**~33-36%; Broadcom AI revenue +143% YoY**	~33-36%	↑	>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

Gemini 3.5 Pro reaches public GA by June 30, 2026, but does not exceed Claude Opus 4.8 on SWE-Bench Pro in its first independent Artificial Analysis run.

TRIGGER: Google AI Studio / Gemini API changelog plus Artificial Analysis leaderboard update.

60%
BY JUNE 30, 2026

PREDICTION 02 HARDWARE

At least one major OEM announces customer shipment or formal order availability for Vera Rubin NVL72-class systems before September 30, 2026.

TRIGGER: Dell, HPE, Lenovo, Supermicro, or NVIDIA customer-shipment announcement.

70%
BY SEPTEMBER 30, 2026

PREDICTION 03 HARDWARE

Before August 31, 2026, at least one memory supplier or supply-chain analyst reports HBM4 allocation tightness despite three-supplier qualification.

TRIGGER: SK hynix, Samsung, Micron, TrendForce, or Bloomberg/Reuters supply-chain reporting.

65%
BY AUGUST 31, 2026

PREDICTION 04 NETWORKING

Broadcom, Marvell, or NVIDIA announces a new CPO/1.6T production design win or revenue guide uplift tied to AI networking before August 31, 2026.

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

65%
BY AUGUST 31, 2026

PREDICTION 05 POWER

A hyperscaler announces another >500MW power-first AI campus or behind-the-meter generation deal by September 30, 2026.

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

60%
BY SEPTEMBER 30, 2026

WATCHLIST

On the radar.

4 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 7-30

WATCH 01

Gemini 3.5 Pro GA and first independent benchmark pass

Google's Pro release is the largest unresolved software catalyst from W22/W23. If it ships below Opus 4.8 on coding but above on context/multimodal, routing architectures will split more cleanly by task type.

JUN-AUG

WATCH 02

HBM4 allocation and Vera Rubin first customer shipment evidence

Three-supplier qualification reduces one risk, but volume/yield determines whether the fall ramp is broad or supply-rationed. Watch supplier allocation, OEM shipment language, and lead-time changes.

JUN-AUG

WATCH 03

CPO and 1.6T optics revenue conversion

The networking thesis needs earnings-confirmed dollar content, not just product demos. Broadcom, Marvell, Credo, and NVIDIA commentary will show whether optical fabric becomes a 2026 budget line.

JUN-SEP

WATCH 04

Power-first campus replication

Google/Intersect's model could become the hyperscaler template. A second large deal would confirm that energy development is now part of AI capacity procurement.

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 07

Week 23 of 2026 · June 6, 2026

WEB

brianletort.ai/industry

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