THE AI STACK WEEKLY ISSUE 04 · WEEK 20 OF 2026

THE BOTTOM LINE

Frontier text took a breath. The constraint stack moved to supply, regulation, and pricing structure.

The model layer took a breath this week. No new GPT-class or Claude refresh between W19 and Google I/O (May 19-20), no headline frontier-text GA — but the specialist canopy widened sharply: Perceptron Mk1 priced video and embodied reasoning 80-90% below Gemini Flash Lite on May 12; NVIDIA's open-weights SANA-WM put a minute-scale 720p world model on a single RTX 5090 on May 15; OpenBMB's MiniCPM-V 4.6 collapsed multimodal SLMs to 1.3B running on-device across iOS, Android, and HarmonyOS on May 11.

Pricing structure became the dominant decision input. Anthropic emailed Max-20x subscribers a June 15 policy that moves Claude Agent SDK, `claude -p`, GitHub Actions, and third-party agents off subscription rate limits onto separate $20-$200/mo metered credit at API list prices — 12x-175x effective price increase per workload, ending Claude Code subscription arbitrage. OpenAI and Microsoft formalized a $38B cumulative cap on Microsoft's revenue-share through 2030 — the first hard ceiling on any hyperscaler's AI revenue-capture. xAI shipped Grok Build CLI and a SuperGrok Heavy tier at $300/mo. The unit of competition shifted from model intelligence to agent-runtime monetization.

Demand pulled forward into prints. Cisco Q3 FY26 raised FY26 AI infrastructure orders guide from $5B to $9B (Q3 AI orders $1.9B vs $600M YoY; ~$300M from non-hyperscaler neocloud / sovereign / enterprise buyers). Applied Materials Q2 FY26 printed a record $7.91B with record 50% non-GAAP gross margin and guided calendar-2026 semicap growth above 30%. Anthropic in talks at over $900B post-money on $30B April ARR (Bloomberg May 12), target close end-May. OpenAI launched a $4B+ Deployment Company with TPG, Advent, Bain, and Brookfield.

The supply chain wedged itself into the read. Samsung-union talks collapsed on May 13, locking in an 18-day general walkout May 21-June 7 across 50,000+ workers — Samsung is the sole HBM4 mass-shipper for NVIDIA's Vera Rubin platform, holds ~40% global DRAM share, and a single April 23 half-day rally already cut daily memory output 18.4%. TSMC's Taiwan Tech Symposium guided AI wafer demand 11x 2022-2026 and CoWoS capacity CAGR above 80% through 2027 — capacity is coming, but not before Q3.

The so-what for the four audiences: boards should treat federal procurement (Canada-TELUS BC sovereign AI factory May 11; HMRC-Quantexa £175M sovereign data and AI May 14) and capability gating (UK AISI's 4.7-month doubling rate on autonomous cyber) as durable vendor-risk dimensions. Investors should treat the Anthropic round close (not the headline valuation) as the binary read of the week and watch for Samsung-walkout impact on NVIDIA Q1 FY27 (May 20) and H2 hyperscaler capex. Architects and operators running Claude Code at scale need to model true API burn before the June 15 cutover this week, lock HBM allocation and Q3 GPU delivery slots, and add at least one specialist physical-AI model (Perceptron Mk1, MolmoAct 2) to 2H roadmaps before procurement budgets close.

AUTHOR

Brian Letort
BrianLetort.AI

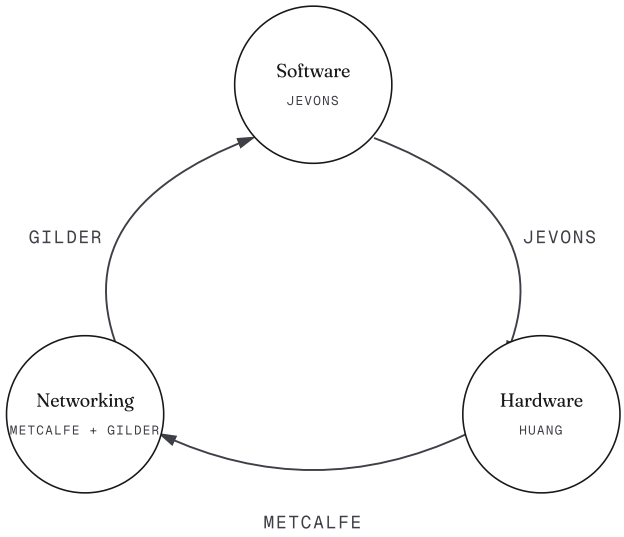
PUBLISHED

May 17, 2026
Issue 04

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.

MAY 13
EVENT 01

Anthropic emails Max-20x subscribers a June 15 policy: Claude Agent SDK, `claude -p`, GitHub Actions, and third-party agents (OpenClaw, T3 Code, Conductor, Zed, Jean) move off subscription rate limits onto separate $20-$200/mo metered credit at API list prices — community math shows 12x-175x effective price increase per workload, ending Claude Code subscription arbitrage

SUPPORT.CLAUDE.COM, VENTUREBEAT, XDA-DEVELOPERS, GATE NEWS

MAY 13
EVENT 02

UK AISI publishes paired evaluations: GPT-5.5 hits 71.4% expert-task pass on AISI's cyber suite and becomes the second model after Claude Mythos Preview to solve the 32-step 'Last Ones' corporate-network attack end-to-end; companion post finds autonomous cyber task length is doubling every 4.7 months, aligned with METR's 4.2-month SWE figure

AISI.GOV.UK, THE REGISTER, CYBERSCOOP, IT PRO, METACURITY, THE DECODER

MAY 12
EVENT 03

Google's Android Show: I/O Edition unveils 'Gemini Intelligence' (built on Gemini 3.1) and Googlebook AI laptops, pushing multi-step agentic Gemini — Chrome auto-browse, cross-app task automation, Rambler dictation, vibe-coded widgets — into the Android OS layer ahead of I/O 2026 (May 19-20)

BLOG.GOOGLE, TECHCRUNCH, ENGADGET, 9TO5GOOGLE, CNBC

MAY 14
EVENT 04

OpenAI Codex changelog ships Codex remote access (Mac-to-mobile project handoff) plus Codex access tokens for trusted non-interactive ChatGPT Enterprise workspaces — Hooks now GA — formalizing a workspace-permissioned CI/CD automation rail distinct from raw API keys

DEVELOPERS.OPENAI.COM/CODEX/CHANGELOG, OPENAI HELP CENTER

WHAT THIS MEANS

W20 traded model novelty for structural posture. AISI's paired disclosures confirm autonomous offensive-cyber capability is now doubling every ~4.7 months (matched to METR's 4.2-month SWE figure) — CISOs and boards should treat that doubling rate as the new 2026 baseline for cyber-risk planning. Anthropic's $200 Agent SDK credit ends subscription-arbitraged Claude Code economics — architects running Claude Code at scale must model true API-rate burn before the June 15 cutover this week or shift workloads. Gemini Intelligence pushing the frontier-model surface into the Android OS layer and OpenAI Codex workspace-permissioned automation rails mean distribution and pricing structure are now the dominant procurement variables, not benchmark scores. See the Model Pulse for the three new tree rows (Perceptron Mk1, SANA-WM, MiniCPM-V 4.6) that widened the specialist canopy this week.

LENS 2 OF 3

THE FLYWHEEL

Hardware.

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

MAY 13
EVENT 01

Government-mediated Samsung-union talks collapse after a 17-hour final round, locking in an 18-day general walkout May 21-June 7 covering 50,000+ workers at the only company currently mass-shipping HBM4 for NVIDIA's Vera Rubin (Samsung holds ~40% global DRAM share; a single April 23 half-day rally already cut daily memory output 18.4%)

KOREA HERALD, SEOUL ECONOMIC DAILY, TECHTIMES

MAY 14
EVENT 02

TSMC at Taiwan Tech Symposium guides AI wafer demand up 11x 2022-2026, CoWoS capacity CAGR >80% through 2027, 14-reticle CoWoS with 20 HBM stacks in 2028 / >14 reticles with 24 HBM stacks in 2029, SoIC Chiayi line scaling 5x to 50K wpm by 2027, and the world's first 200Gbps Micro Ring Modulator using COUPE silicon photonics entering production this year

TRENDFORCE, REUTERS, COMMERCIAL TIMES, DIGITIMES

MAY 12
EVENT 03

TrendForce reports MediaTek adopting a dual advanced-packaging strategy (Intel EMIB + TSMC CoWoS) for AI ASICs, with Google's TPU v8e expected to use Intel EMIB; EMIB-M is scaling toward 6-12x effective reticle size vs CoWoS-S at 3.3x, opening a credible second source for hyperscaler custom-silicon supply outside the CoWoS bottleneck

TRENDFORCE, DIGITIMES

MAY 11
EVENT 04

Applied Materials and TSMC announce a partnership at AMAT's $5B EPIC Center in Silicon Valley to co-develop materials, equipment, and process integration for next-gen 3D transistor and interconnect structures, with TSMC named a founding partner getting early access to AMAT's roadmap

APPLIED MATERIALS IR / GLOBENEWSWIRE PRESS RELEASE

WHAT THIS MEANS

Architects and procurement leads should lock HBM allocation and triage Q3 2026 GPU delivery slots this week rather than waiting for the May 20 NVIDIA print: the Samsung walkout (May 21-June 7) threatens the sole HBM4 mass-shipper for Vera Rubin, while TSMC's 14-reticle CoWoS and 24-stack HBM trajectory (2028-29) sets the density envelope to plan 2028+ platforms against. MediaTek's validation of Intel EMIB as a real CoWoS alternative for AI ASICs (with TPU v8e named) should reopen the ~33-36% custom-silicon-share assumption upward — when packaging supply diversifies, hyperscaler ASIC volume scales faster, and the hardware-to-software flywheel turns again because new packaging unlocks architectures (24-stack HBM, EMIB-bridged ASICs) that weren't economic last quarter. Investors should model HBM supply continuity — not GPU pricing — as the binding swing variable on the May 20 NVIDIA print and H2 hyperscaler capex.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

MAY 12
EVENT 01

3M, AMD, Arista, Cisco, Meta, Oracle (co-chair), Molex, Amphenol, TE Connectivity, Senko, Sumitomo, Source Photonics, Accelink, Aperion, Nexthop-AI, viaPhoton, and Xscape Photonics launch the Expanded Beam Optical (EBO) MSA at ebomsa.org to standardize open, contamination-resilient connector specs for hyperscale AI fabrics

3M, PR NEWSWIRE, CONVERGE DIGEST, SDXCENTRAL, RCR TECH

MAY 11
EVENT 02

TELUS and the Government of Canada unveil a three-site British Columbia Sovereign AI Factory cluster (Kamloops, Vancouver Mount Pleasant, 150 W Georgia) scaling to 60,000+ NVIDIA GPUs and 150 MW by 2032, networked with NVIDIA Quantum InfiniBand and Spectrum-X Ethernet under the federal Enabling Large-Scale Sovereign AI Data Centres program

TELUS, CNW NEWSWIRE, CBC, GLOBE AND MAIL

MAY 12
EVENT 03

Lumen launches NorthLine, a ~2,000-mile Seattle-Minneapolis long-haul fiber route (100G/400G wavelengths now, architected for future 800G and 1.6T) — its first major long-haul build in decades — targeting AI east-west DCI traffic across emerging northern data-center corridors

LUMEN, BUSINESSWIRE, FIERCE NETWORK, TELECOMPAPER

MAY 14
EVENT 04

Equinix announces global expansion of Fabric Geo Zones across five continents (preview live in AU/BR/CA/JP/CH/UK/US; EU in June), positioning network-layer multicloud sovereignty enforcement that blocks rerouting across non-compliant jurisdictions on its 77-metro software-defined fabric

EQUINIX NEWSROOM, PR NEWSWIRE

WHAT THIS MEANS

W20's pattern is network-internal compounding — AI fabric matured at the connector layer (17-vendor EBO MSA pulling AMD, Cisco, Meta, and Oracle into a shared open spec), the inter-DC fiber map (Lumen NorthLine, first major US long-haul build in decades), and sovereignty enforced at the wire (Equinix Geo Zones; TELUS' federal-backed InfiniBand + Spectrum-X cluster). Architects and procurement leads writing 2027 AI RFPs should treat 800G as capacity-bound through mid-2027 (per Applied Optoelectronics Q1 commentary), start EBO connector qualification and sovereignty-aware routing requirements now, and stop treating both as application-overlay problems. Investors should watch interconnect-category pricing power compound — Equinix Geo Zones is a wire-level enforcement product, not a feature.

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 58.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV **0.9**

Anthropic in talks to raise at least $30B at a >$900B valuation (Bloomberg May 12), target close end-May, on April annualized run-rate revenue of ~$30B — boards on Anthropic-dependent contracts should underwrite ~30 months counterparty runway conditional on close; investors should treat the close (not the headline valuation) as the binary read of the week.

Hyperscaler-Hosted

MICROSOFT AZURE, GOOGLE CLOUD, AWS, META, ORACLE

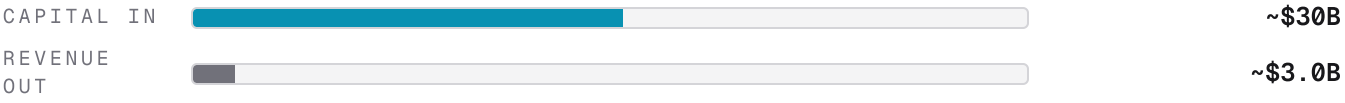
BURN / REV **0.26**

OpenAI and Microsoft formalized a $38B cumulative cap on Microsoft's revenue-share payments through 2030 on May 12 — the first hard ceiling on any hyperscaler's AI revenue-capture; architects building on Azure-OpenAI should plan for Microsoft pushing native MAI-line and direct API products into more enterprise workloads through 2027, and operators should expect Azure-OpenAI commercial terms to harden, not loosen, post-cap.

Neoclouds

COREWEAVE, NSCALE, IREN, CRUSOE, LAMBDA, APPLIED DIGITAL, NEBIUS

BURN / REV **0.10**

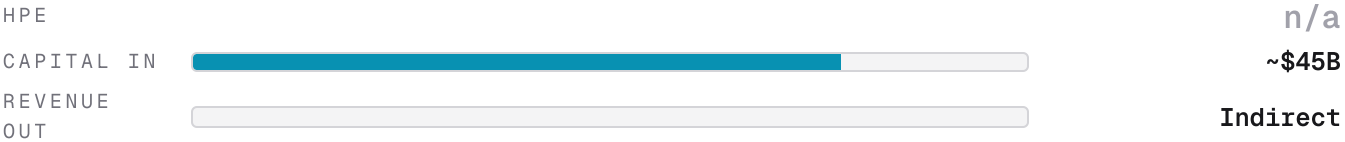


Nscale closed $790M committed senior debt (+$790M accordion) from five Nordic banks for the Narvik, Norway expansion on May 11 — operators evaluating non-CoreWeave neocloud capacity now have a fresh credit comp; investors should treat this as the template for second-tier neocloud refinancings, not a one-off, and watch the accordion draw cadence as a real-time demand signal.

On-Prem / Hybrid

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

BURN / REV **n/a**



Cisco Q3 FY26 raised FY26 AI infrastructure orders guide from $5B to $9B and AI revenue guide from $3B to $4B on May 13, with Q3 AI orders of $1.9B vs $600M YoY (~$300M from non-hyperscaler neocloud / sovereign / enterprise) — architects should re-baseline enterprise AI fabric demand from a 2027-2028 thesis to a real 2026-print risk; operators should expect lead-time pressure on Silicon One, Nexus 9000, and Hypershield SKUs.

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

Cisco raised FY26 AI infrastructure orders guide from $5B to $9B and AI revenue guide from $3B to $4B (Q3 FY26 release May 13, 2026); Q3 AI orders alone were $1.9B vs $600M YoY, with ~$300M from non-hyperscaler neocloud / sovereign / enterprise buyers against a $3B pipeline.

SOURCES: CISCO PR NEWSWIRE (PRIMARY PRESS RELEASE), CNBC, FT.COM DATA FEED

Real and material — enterprise / on-prem AI fabric demand pulled forward into 2026 prints; operators on legacy refresh budgets should re-baseline networking lead-times this week, not next quarter. What would change the read: Q4 FY26 AI orders failing to deliver the implied ~$3.7B step needed to hit the $9B annual guide.

5 / 5

CLAIM 02

Anthropic in talks to raise at least $30B at a valuation exceeding $900B, target close end-May 2026, on April annualized run-rate revenue of ~$30B (doubled from $14B in February); previous Feb 2026 round was $30B at $350B.

SOURCES: BLOOMBERG (PRIMARY MAY 12), YAHOO FINANCE / INVESTING.COM / TECHZINE / MONEYCONTROL (ECHO)

Real but unsigned — term sheet not finalized, so investors should treat 'closes' vs 'collapses' as the binary read; boards on Anthropic-dependent vendor contracts can underwrite ~30 months counterparty runway conditional on close. What would change the read: lead investor list leak, a haircut to <$700B, or a slip past the end-May target.

4 / 5

CLAIM 03

OpenAI and Microsoft formalized a $38B cumulative cap on Microsoft revenue-share payments through 2030 (May 12); follows the late-2025 contract amendment that allowed OpenAI to add Anthropic-style multi-cloud commitments (SpaceX, Akamai, Amazon).

SOURCES: THE HINDU / REUTERS-DERIVED, RESULTSENSE SECONDARY; NO MSFT 8-K INSIDE-WINDOW

Real and structural — first hard ceiling on any hyperscaler's AI revenue-share capture. Architects building on Azure-OpenAI should plan for Microsoft pushing MAI-line and native direct-API products into more enterprise workloads through 2027; operators should expect Azure-OpenAI commercial terms to harden post-cap. What would change the read: a counter-disclosure from MSFT IR walking back the cap framing or a different framing in the next MSFT 10-Q.

4 / 5

CLAIM 04

Samsung Electronics and union talks collapsed May 13 after 17 hours; an 18-day general walkout May 21-June 7 covering 50,000+ workers is locked in at the only company currently mass-shipping HBM4 for NVIDIA Vera Rubin; Samsung holds ~40% global DRAM share and a single April 23 half-day rally cut daily memory output 18.4%.

SOURCES: KOREA HERALD, SEOUL ECONOMIC DAILY, TECHTIMES (MAY 16 SYNTHESIS)

Real and consequential. The Vera Rubin H2 ramp leans on Samsung HBM4 — an 18-day walkout almost certainly compresses Q2-Q3 HBM4 output by single-digit-billion-dollar revenue impact and risks NVIDIA's May 20 print color. Operators planning H2 capacity should triage HBM allocation this week; investors should size the supply-chain wedge against Goldman's $80B Q1 forecast.

2 / 5 — NOISE

CLAIM 05

Alpha Compute closed a $32.2M two-year lease with a 'leading frontier AI lab' for 504 NVIDIA B200 GPUs at a Canadian data center (May 12), generating ~$16.1M annualized recurring revenue.

SOURCES: MARKETS INSIDER / BUSINESS INSIDER (PR-LED, SINGLE PRIMARY, COUNTERPARTY UNNAMED)

Mostly noise — small ARR contract dressed up with 'frontier AI lab' framing; counterparty undisclosed, deal size irrelevant at category level. Operators should not treat this as a comp for capacity pricing or deployment density; investors should discount until the lab is named and the deal is corroborated against the lab's public compute roadmap.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **1 rising**, **0 falling**, 9 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)	**~30 mo**	~30 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	**$62.3B (Q4 FY26)**	$62.3B (Q4 FY26)	→	Q1 FY27 print Tuesday May 20 — Goldman forecasts ~$70B+ DC implied
Open vs closed gap on SWE-Bench Pro (coding)	**Closed +19pp**	Closed +6 to +20pp	→	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**12 / ~$80B+**	10 / ~$80B+	↑	—
PJM 2026/27 capacity auction price ($/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); 36-48 (existing PJM queue)**	60-84 (new PJM); 36-48 (existing PJM queue)	→	—
Cost-per-task, frontier reasoning model	**~$0.10-$0.15 (effective, with hidden reasoning tokens)**	~$0.10-$0.15 (effective, with hidden reasoning tokens)	→	Anthropic Agent SDK June 15 cutover will reset effective Claude Code economics
Custom silicon share of incremental AI compute	**~33-36% (estimated)**	~33-36% (estimated)	→	>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 HARDWARE

55%

BY JULY 31, 2026

Samsung HBM4 shipments to NVIDIA Vera Rubin drop by more than 25% in May vs April due to the May 21-June 7 walkout, with at least one downstream hyperscaler publicly delaying a Q3 capacity ramp.

TRIGGER: TrendForce / SemiAnalysis May-June supply data; NVIDIA Q1 FY27 commentary May 20; hyperscaler Q2 capex / supply commentary.

PREDICTION 02 SOFTWARE

65%

BY AUGUST 15, 2026

Anthropic discloses or third-party developer surveys document a greater than 20% drop in active Claude Code third-party agent users within 30 days of the June 15 pricing cutover.

TRIGGER: Anthropic earnings / blog commentary; GitHub stars / npm downloads on Claude Agent SDK; third-party agent vendors (OpenClaw, T3 Code, Conductor, Zed) public statements.

PREDICTION 03 SOFTWARE

60%

BY SEPTEMBER 30, 2026

At least one Fortune 500 enterprise discloses a production deployment greater than $10M annualized of a specialist video / embodied multimodal model (Perceptron Mk1, MolmoAct 2 successors, or analogous) as a primary tier rather than a general LMM.

TRIGGER: Vendor case studies; enterprise architecture announcements; quarterly earnings color from Perceptron, Ai2 partners.

PREDICTION 04 CAPITAL

70%

BY AUGUST 20, 2026

Cisco's Q4 FY26 AI orders confirm the $9B annual run rate (Q4 AI orders at least $2.5B), validating the enterprise AI fabric demand-pull narrative.

TRIGGER: Cisco Q4 FY26 earnings release (early August).

PREDICTION 05 NETWORKING

60%

BY AUGUST 12, 2026

Expanded Beam Optical MSA publishes a v1.0 spec within 90 days of launch (May 12), with at least one in-production deployment announced by a hyperscaler member (AMD, Cisco, Meta, Oracle).

TRIGGER: ebomsa.org publications; member vendor blogs; OCP Future Technologies Symposium filings.

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.

MAY 19-20

WATCH 01

Google I/O 2026 — Gemini 3.2 Flash / Pro confirmation + fabric session

Gemini 3.2 Flash leaked May 5 at ~$0.25 / $2.00 per Mtok with ~92% of GPT-5.5 capability; if confirmed at I/O, becomes the new price anchor for the cheap-frontier tier and forces OpenAI / Anthropic price moves within 60 days. Watch the fabric session for any second-hyperscaler MRC adoption (tests p13).

MAY 20

WATCH 02

NVIDIA Q1 FY27 earnings

First disclosure on Vera Rubin ramp velocity, Samsung HBM4 supply commentary post-walkout context, and warrant-stake terms for CoreWeave / IREN / Corning equity investments. Goldman's $80B total / ~$70B+ DC implied bar sets the test. A beat resets H2 hyperscaler capex tail; a miss flattens the doubling slope; circular-flow framing stress-tests on the call (tests p16).

MAY 21 - JUNE 7

WATCH 03

Samsung HBM4 walkout impact on daily output

18-day general walkout locked in across 50,000+ workers at the sole HBM4 mass-shipper for NVIDIA Vera Rubin. A single half-day rally April 23 cut daily memory output 18.4%; an 18-day stoppage at the same scale puts single-digit-billion of Q2-Q3 HBM4 revenue at risk and threatens Vera Rubin H2 ramp commitments. Tests p17.

MAY 19-22

WATCH 04

Microsoft Build 2026

First major Microsoft venue since the OpenAI revenue-share cap formalized. Watch for MAI-line product announcements that signal Microsoft pushing direct-API and native models into enterprise workloads; Azure-OpenAI commercial terms color; any new Bedrock-like distribution moves through OCI / Oracle. Conditions architects' Azure-OpenAI procurement read through 2027.

BY END-MAY

WATCH 05

Anthropic $900B round close or pricing slip

Bloomberg-confirmed target end-May close on $30B / >$900B round. The price ($900B vs OpenAI's $852B) and the lead investor's identity re-rate the entire frontier-lab valuation curve. A slip past end-May, a pricing haircut below $700B, or a walk-away from the price would be the loudest re-rating signal in a year. Tests p14.

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 04

Week 20 of 2026 · May 17, 2026

WEB

brianletort.ai/industry

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