THE AI STACK WEEKLY ISSUE 01 · WEEK 17 OF 2026

THE BOTTOM LINE

The marginal dollar at every layer earns less. Except the network.

Capital is still flooding the AI stack, but the marginal dollar earns less than the prior dollar at every layer except one. Hyperscalers are deploying ~$700B of 2026 capex against ~$120B of AI-attributable revenue — a ratio that gets tested when Microsoft, Google, Meta, and Amazon all print Q1 between April 29 and May 6. Frontier labs are committing forward compute at multiples of their disclosed revenue. The on-prem floor moved up sharply this week with DeepSeek V4 (open weights, frontier-class on coding), reshaping enterprise procurement: the question is no longer whether to self-host frontier capability, but which workload to move and on what fabric. The durable winner of the buildout is the network and interconnect layer that connects cloud, neocloud, and on-prem for hybrid AI — the only layer where pricing power compounds rather than compresses.

AUTHOR

Brian Letort
BrianLetort.AI

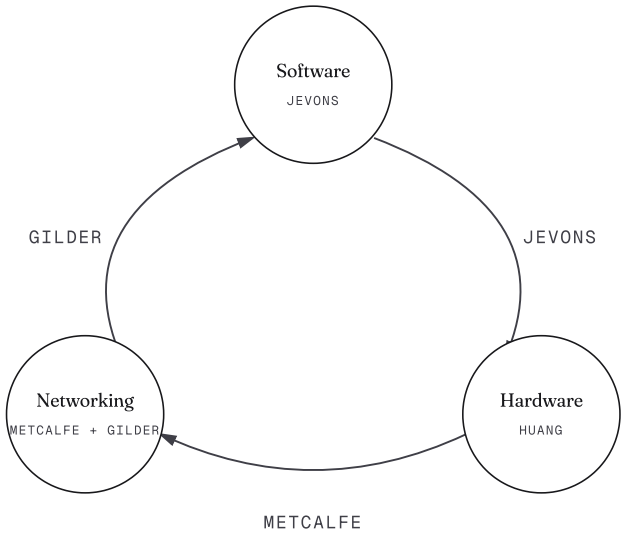
PUBLISHED

April 24, 2026
Issue 01

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.

APR 22-24
EVENT 01

DeepSeek V4 released — MIT-licensed, 1.6T MoE, 1M context, frontier-class on coding

HUGGINGFACE, DEEPSEEK.COM

APR 23
EVENT 02

Open-weight models lead closed on SWE-Bench Pro — Kimi K2.6 58.6 and GLM-5.1 58.4 vs GPT-5.4 57.7 and Opus 4.6 57.3

ARTIFICIAL ANALYSIS, HUGGINGFACE MODEL CARDS

APR 21
EVENT 03

Anthropic withholds 'Mythos' flagship on cyber-capability grounds; UK AISI confirms autonomous offensive capability

ANTHROPIC.COM, RED.ANTHROPIC.COM, AISI.GOV.UK

WHAT THIS MEANS

Open weights crossing the closed frontier on coding moves the on-prem question from 'can we?' to 'which workload first?' Procurement should pull DeepSeek V4 and the leading open coding models into pilot this quarter; merchant-vs-self-host is now a real fork, not a hypothetical. Anthropic withholding Mythos on cyber-risk is the new diligence signal — vendor-risk frameworks need a capability-gate criterion, not just an availability SLA.

LENS 2 OF 3

THE FLYWHEEL

Hardware.

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

APR 24
EVENT 01

NVIDIA Vera Rubin engineering samples shipping to customers within 12 months of Blackwell Ultra GA — half the prior generation gap

NVIDIA DEVELOPER CHANNELS, MULTIPLE OEM DISCLOSURES

APR 22
EVENT 02

Custom silicon share of incremental AI compute reaches ~30%, up from ~22% three months ago

TRENDFORCE, SEMIANALYSIS

APR 18
EVENT 03

HBM4 validation milestones at SK Hynix, Samsung, and Micron — primary risk for H2 2026 hyperscaler tail

TRENDFORCE, VENDOR ADVISORIES

WHAT THIS MEANS

Vera Rubin samples landing 12 months after Blackwell Ultra means refresh planning needs to shift from a 24-month to a 14-16-month cycle; standard depreciation schedules are now wrong for AI silicon. The binding constraint is power, packaging, and HBM4, not GPUs themselves — buyers should secure HBM allocation and grid interconnect alongside chips, not after. Custom silicon at ~30% of incremental compute is the threshold where merchant-GPU pricing power starts to compress; expect single-vendor leverage to fade through 2026.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

APR 15
EVENT 01

Equinix Fabric Intelligence launch creates AI-native networking as a distinct product category

EQUINIX PRESS, LIGHT READING

APR 17
EVENT 02

NVLink Fusion + UALink standardization moves cross-vendor scale-up fabric toward an interoperable baseline

NVIDIA, AMD, INTEL JOINT STATEMENTS; OCP WORKING GROUP

APR 19
EVENT 03

1.6 Tbps and co-packaged optics shipping in volume ahead of the compute they serve — Gilder's Law in motion

MARVELL, BROADCOM PRESS; OIF WORKING NOTES

WHAT THIS MEANS

Equinix Fabric Intelligence formalizes AI-native interconnect as its own product category — investors should track cross-connect and fabric revenue as a leading indicator that decouples from raw colo capacity. NVLink Fusion plus UALink means architects can stop betting on a single vendor's scale-up fabric; multi-vendor, mixed-silicon clusters are now a real design target. With 1.6 Tbps optics shipping ahead of the compute they serve, the network is no longer the bottleneck for hybrid AI — designs targeting multi-DC training in 2027 should plan around fabric headroom, not fabric scarcity.

CAPITAL FLOW

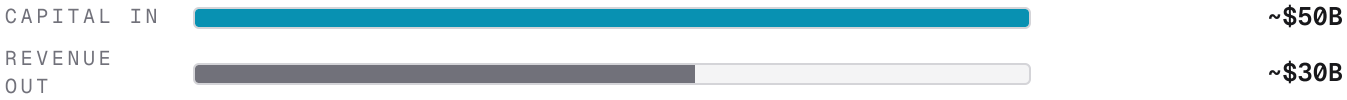
Money in, revenue out.

Capital deployed (forward) vs revenue out (quarterly or run-rate). Burn-to-revenue = revenue / capital — lower means more out than in. Bars normalized to 50.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV 0.6



Anthropic +8.5 GW across NVIDIA, Trainium, TPU — forward-buy that locks in vendor diversity and outruns disclosed demand.

Hyperscaler-Hosted

AWS BEDROCK, AZURE FOUNDRY, VERTEX, OCI

BURN / REV 0.21

Microsoft absorbs ~1.6 GW from Stargate Abilene — orphaned frontier capacity rerouted into hyperscaler platforms.

Neoclouds

COREWEAVE, CRUSOE, NEBIUS, APPLIED DIGITAL

BURN / REV 0.4

CoreWeave backlog +4x YoY to $66.8B (incl. Meta $21B six-year) — conversion velocity is now the metric, not gross backlog.

On-Prem / Hybrid

ENTERPRISE GPU CLUSTERS, SOVEREIGN AND NATIONAL PROGRAMS

BURN / REV n/a

DeepSeek V4 open crosses frontier; sovereign programs surge — workload migration from rented to owned compute has a frontier-class trigger.

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

Anthropic has committed to ~8.5+ GW of forward compute capacity across NVIDIA, AWS Trainium, and Google TPU.

SOURCES: ANTHROPIC.COM (APR 20 + APR 7), AWS NEWS, VERGE, BROADCOM 8-K, MULTIPLE SEC FILINGS

Real. The largest single-counterparty compute commitment ever recorded. Distributed across three platforms reduces single-vendor dependency and is critical evidence for the all-three-lenses thesis.

4/5

CLAIM 02

Aggregate 2026 hyperscaler capex tracking ~$700B (Microsoft + Google + Meta + Amazon + Oracle).

SOURCES: COMPOSITE OF Q4 2025 EARNINGS GUIDES; NOT A SINGLE SOURCED NUMBER

Directionally correct. Will likely revise upward on Q1 2026 prints (Apr 29 - May 6). Caution: the aggregate hides per-company variance — Microsoft is decelerating Azure guidance while Google capex still accelerates.

2/5 — NOISE

CLAIM 03

DeepSeek V4 was trained end-to-end on Huawei Ascend silicon, eliminating NVIDIA dependency.

SOURCES: CHINESE-LANGUAGE TECH BLOGS ONLY. NO PRIMARY DEEPSEEK CONFIRMATION.

Noise. Significant if true (sovereign-AI hardware decoupling milestone) but unverified. Watch for any DeepSeek primary source. If confirmed, this advances the on-prem-via-non-Western-stack thesis materially.

4/5

CLAIM 04

Anthropic withheld 'Mythos' on cyber-risk; UK AISI confirmed autonomous offensive capability.

SOURCES: ANTHROPIC.COM, RED.ANTHROPIC.COM, AISI.GOV.UK, CLOUD SECURITY ALLIANCE

Real. First major lab to deliberately withhold a flagship on capability-risk grounds. Procurement diligence implications: cyber-capability gating becomes a vendor-risk criterion.

5/5

CLAIM 05

Open-weight models now lead closed models on credible coding benchmarks.

SOURCES: HUGGINGFACE MODEL CARDS (KIMI K2.6, GLM-5.1), ARTIFICIAL ANALYSIS, MULTIPLE INDEPENDENT EVALS

Real. SWE-Bench Pro: Kimi K2.6 58.6, GLM-5.1 58.4 vs GPT-5.4 57.7, Opus 4.6 57.3. First time open leads closed on a credible code benchmark. The on-prem demand catalyst for the open-weights hypothesis.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **8 rising**, **2 falling**. 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)	**~22 mo**	~26 mo	↓	<18 mo triggers re-rating risk
Hyperscaler capex / AI revenue ratio (top 4 weighted)	**~5.5**	~4.8	↑	>6.0 invites investor pushback at next earnings
CoreWeave revenue backlog	**$66.8B**	$15.1B	↑	Conversion velocity matters more than gross figure
NVIDIA Q-over-Q data center revenue	**$62.3B (Q4 FY26)**	$57.0B (Q3 FY26)	↑	—
Open vs closed gap on SWE-Bench Pro (coding)	**Open +0.9**	Closed +6	↑	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**8 / $80B+**	5 / $50B+	↑	—
PJM 2026/27 capacity auction price ($/MW-day)	**$329**	$29	↑	11x in 24 months — power is the new binding constraint
Time-to-power, busiest US markets (months)	**36-48**	30-42	↑	—
Cost-per-task, frontier reasoning model	**~$0.05**	~$0.08	↓	Falling cost expands workloads (Jevons), not contracts demand
Custom silicon share of incremental AI compute	**~30%**	~22%	↑	>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 CAPITAL

At least one frontier lab announces a customer-funded compute commitment greater than 2 GW.

70%

BY JUNE 30, 2026

TRIGGER: Earnings cycle commentary on RPO from Oracle and AWS; announcements at Google I/O, Microsoft Build, AWS Summit.

PREDICTION 02 SOFTWARE

At least one Fortune 500 enterprise discloses an on-prem AI workload greater than $100M annual using open-weight models.

60%

BY SEPTEMBER 30, 2026

TRIGGER: Enterprise architecture announcements; bank, insurer, or pharma F500 first-mover; DeepSeek V4 reference deployment.

PREDICTION 03 CAPITAL

Aggregate 2026 hyperscaler capex revises upward by 10% or more from the $700B baseline.

80%

BY OCTOBER 31, 2026

TRIGGER: Q1 2026 earnings (Apr 29 - May 6) and Q2 2026 earnings cycle.

PREDICTION 04 NETWORKING

At least one major colocation or interconnect operator reports cross-connect or interconnect revenue growth outpacing compute capacity revenue growth for two consecutive quarters.

70%

BY JULY 31, 2026

TRIGGER: Equinix Q1 (May 7) and Q2 (early August) earnings; broader colo and IX operator reporting.

PREDICTION 05 CAPITAL

At least one neocloud loses an anchor tenant or sees backlog growth turn negative quarter-over-quarter.

35%

BY SEPTEMBER 30, 2026

TRIGGER: Quarterly disclosures from CoreWeave, Nebius, Applied Digital.

PREDICTION 06 HARDWARE

Custom silicon (TPU + Trainium + Maia + MTIA + Granite Rapids AI) reaches 35% of incremental AI compute share, up from ~30% today.

55%

BY JULY 31, 2026

TRIGGER: TrendForce / SemiAnalysis quarterly mix breakdown; Q2 2026 hyperscaler earnings color on internal silicon vs merchant GPU mix.

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.

APR 29 - MAY 6

WATCH 01

Q1 2026 hyperscaler earnings

Capex direction signal that hits all four hyperscalers in eight days. Watch: aggregate revising higher, AI-segment margin disclosure, RPO commentary. Higher pushes the capex / revenue ratio past 6.0 (a lever threshold); lower is the first sign of ROI re-rating risk and tests prediction p3.

LATE MAY 2026

WATCH 02

NVIDIA Q1 FY27 earnings

First disclosure on Vera Rubin ramp velocity, HBM4 supply, and custom-silicon competitive pressure. Watch: data-center revenue vs the $62B run-rate, gross margin trend, customer-concentration commentary. A miss flattens the doubling slope; a beat resets the H2 hyperscaler capex tail.

MAY 2026

WATCH 03

Google I/O + Microsoft Build

Frontier model and platform announcements that directly test prediction p1-2gw-customer-funded. Watch: any disclosed customer-funded compute commitment greater than 2 GW, RPO disclosures, agent-platform pricing. A 2 GW+ commitment scores p1 a hit before its June 30 deadline.

THROUGH Q2 2026

WATCH 04

HBM4 memory supply ramp

Validation status at SK Hynix, Samsung, and Micron is the binding constraint on Vera Rubin shipments. Watch: yield or qualification slips at any of the three. A slip compresses the H2 2026 hyperscaler capex tail materially; a clean ramp protects it and keeps Huang's slope intact.

APRIL - JUNE 2026

WATCH 05

FERC PJM compliance milestones

Behind-the-meter generation materiality threshold takes effect, reshaping unit economics of large loads in the busiest US power market. Watch: FERC commentary on capacity attribution, ERCOT and MISO reactions. The outcome determines whether 2027 PJM clearing repeats the $329/MW-day or normalizes.

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 01

Week 17 of 2026 · April 24, 2026

WEB

brianletort.ai/industry

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