THE AI STACK WEEKLY ISSUE 06 · WEEK 22 OF 2026

THE BOTTOM LINE

A quiet release week resolved last week's two biggest open questions — and moved the contested ground to optics and the agent control plane.

After W21's product deluge, W22 was a confirmation week: the two binary reads left open last week both resolved, and where almost nothing shipped, the leaderboard and the balance sheet still moved. On software, Claude Opus 4.8 (May 28) retook #1 on the Artificial Analysis Intelligence Index (61.4 vs GPT-5.5's 60.2) with a >10-point SWE-Bench Pro lead — at flat list pricing — while Gemini 3.5 Pro slipped to June, leaving Anthropic the unchallenged closed frontier. On capital, the W21 watch item resolved cleanly: Anthropic's round closed May 28 as a $65B Series H at a $965B post-money, the largest private AI raise on record (though ~$15B is previously-committed hyperscaler capital, not fresh cash). On hardware, the other open risk closed too — Samsung's union ratified its wage deal on May 27 (73.7% approval), formally averting the strike that threatened HBM4 packaging, so the binding constraint stays supply (CoWoS + HBM4 + surging DRAM pricing), not demand. The genuinely new front was networking: four independent vendors (GlobalFoundries, Wiwynn, Credo, Edgecore) pushed co-packaged optics and all-photonic fabric from demos toward deployable product in-window, with one analyst house putting a ~$154B TAM on optical interconnect — a textbook Gilder arc where per-GPU bandwidth, and the dollar content that carries it, scales faster than compute. So-what for the four audiences: boards should treat the HBM4 supply scare as closed and the custom-silicon counter-flywheel (ASIC shipments +44.6% YoY vs +16.1% for merchant GPUs) as the durable margin question; investors should read the Anthropic close as the answered binary and discount the louder unverified headlines (a '$45B Anthropic-pays-xAI-through-2029' filing the principal himself recut to a ~180-day lease; a secondary-sourced '$1T OpenAI September IPO'); architects should re-baseline coding-agent evals on Opus 4.8 but pin effort levels first; operators should put agent

governance and the data-cloud control plane (Snowflake's Natoma/MCP buy, Salesforce's $1.2B Agentforce ARR) on this quarter's procurement roadmap.

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

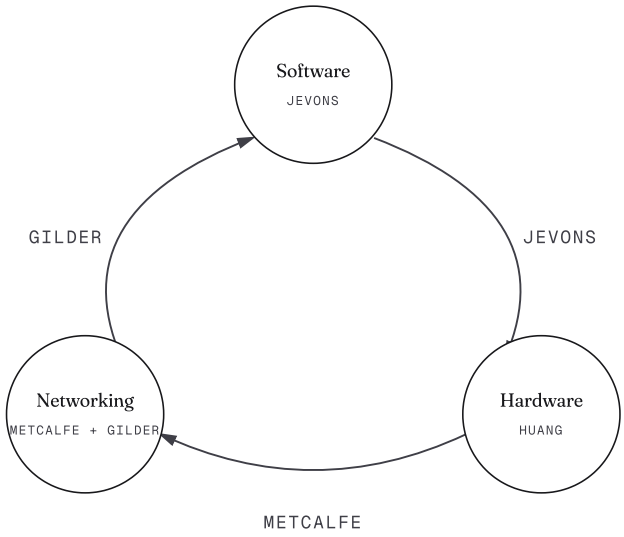
May 30, 2026

Issue 06

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 28
EVENT 01

Anthropic released Claude Opus 4.8 (claude-opus-4-8) at flat pricing ($5/$25 per Mtok, fast mode ~3x cheaper) plus a Dynamic Workflows research preview; retakes the Artificial Analysis Intelligence Index lead at 61.4 (vs GPT-5.5 60.2) and SWE-Bench Pro 69.2%, though it loses Terminal-Bench 2.1 to GPT-5.5

ANTHROPIC.COM/NEWS/CLAUDE-OPUS-4-8, ARTIFICIALANALYSIS.AI, TECHCRUNCH

MAY 28
EVENT 02

Liquid AI shipped open-weight LFM2.5-8B-A1B, an on-device MoE (8.3B total / 1.5B active, 128K context) decoding ~253 tok/s on a laptop with day-one llama.cpp / MLX / vLLM / SGLang support

LIQUID.AI/BLOG/LFM2-5-8B-A1B, MARKTECHPOST

MAY 28
EVENT 03

StepFun released open-weight Step 3.7 Flash (198B sparse MoE, ~11B active, 256K context) under Apache 2.0 at ~400 tok/s, claiming SWE-PRO 56.3 and 98%+ on tau-squared-bench, positioned for agentic / coding / search workloads

STEPFUN HUGGINGFACE MODEL CARD, DIGG

MAY 27
EVENT 04

A Financial Times report showed the open-source 'Heretic' abliteration tool (~17.8K stars) stripping safety guardrails from open-weight models such as Llama 3.3 and Gemma 4 within minutes, spotlighting the open-weight safety floor ahead of the EU AI Office's Aug 2 Article 92 evaluation powers

FINANCIAL TIMES (VIA SINGULARITY.KIWI), LAWFARE

WHAT THIS MEANS

Two arcs run in parallel: (1) a software→hardware Jevons signal — the closed frontier pushed the ceiling (Opus 4.8) while a same-day open-weight efficiency wave (Liquid on-device 128K at 253 tok/s, StepFun 400 tok/s under Apache 2.0) pushed cheap-and-good inference outward, so architects should re-baseline cost-per-task now rather than wait for price cuts; (2) capability-risk procurement hardens as trivially abliterated open weights lower the deployed safety floor just as the EU AI Office's Article 92 access powers arrive Aug 2 — boards should treat open-weight governance as a vendor-risk dimension, not a compliance afterthought.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

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

MAY 27 EVENT 01

Samsung's union ratified its 2026 wage-and-bonus deal with 73.7% approval on 95.5% turnout, formally averting the strike that had threatened the testing/packaging lines feeding HBM4 supply for next-gen AI accelerators

BLOOMBERG, KOREA HERALD, UNION JOINT-BARGAINING-TEAM ANNOUNCEMENT

MAY 27 EVENT 02

TrendForce's in-window DRAM Market Bulletin reported seller-driven contract price hikes, said HBM prices will surge once a temporary margin lag clears, and warned fab-expansion lead times will keep memory supply bottlenecked 'for years,' with HBM4 the gating 2026 constraint

TRENDFORCE DRAM MARKET BULLETIN (RP260527ZM)

MAY 26 EVENT 03

Alchip's CEO said custom-ASIC revenue growth will outpace the GPU market, aligning with TrendForce data showing cloud custom-silicon shipments tracking +44.6% in 2026 vs +16.1% for merchant GPUs — the first AI-era year custom chips meaningfully outgrow GPUs, with ASICs ~27.8% of AI-server shipments

ALCHIP CEO REMARKS + TRENDFORCE, VIA TECHTIMES

MAY 21 EVENT 04

AMD began the production ramp of its 6th-gen EPYC 'Venice' (Zen 6, up to 256 cores, 1.6 TB/s memory bandwidth) — the industry's first HPC product on TSMC's 2nm node and built on SoIC-X / CoWoS-L packaging, feeding the MI450X-based Helios rack into the Computex run-up

AMD IR PRESS RELEASE, TOM'S HARDWARE, WCCFTECH

WHAT THIS MEANS

A constraint-and-cadence week, not a launch week: the binding limits on the hardware flywheel are packaging (CoWoS) and memory (HBM4 plus surging DRAM/HBM pricing), and the Samsung labor resolution defused the single biggest near-term HBM4 supply shock just before the June 1 GTC Taipei keynote. Underneath, the custom-silicon counter-flywheel is real (ASIC shipments +44.6% vs +16.1% GPUs) and AMD's 2nm/SoIC-X Venice ramp shows rivals fighting for the same TSMC node-and-packaging stack that gates every accelerator from wafer to rack — investors with concentrated merchant-GPU exposure should diversify into ASIC co-design and advanced packaging.

LENS 3 OF 3 THE FLYWHEEL

Networking.

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

MAY 25
EVENT 01

GlobalFoundries launched its SCALE co-packaged silicon-photonics module, claiming the first platform to comply with and exceed the Optical Compute Interconnect (OCI) MSA, with 50G/100G micro-ring modulators and demonstrated 8-lambda/16-lambda bidirectional DWDM for optical scale-up interconnect

GLOBALFOUNDRIES RELEASE VIA ET DATACENTERS

MAY 26
EVENT 02

Wiwynn unveiled a CPO-based optical scale-up rack design (Ayar Labs TeraPHY engines, ELSFP SuperNova remote light sources, GUC 2.5D packaging) ahead of Computex 2026, framing co-packaged optics as the path to move AI scale-up fabric beyond copper limits

WIWYNN PRESS RELEASE

MAY 27
EVENT 03

Credo closed its acquisition of DustPhotonics, adding silicon-photonics PIC technology to build a vertically integrated interconnect stack (SerDes + DSP + SiPho + system integration) spanning 800G, 1.6T and 3.2T near-packaged and co-packaged optics for scale-out and scale-up AI networks

CREDO VIA BUSINESS WIRE / MORNINGSTAR

MAY 28
EVENT 04

Edgecore (Accton) and NTT announced 'Edgecore Open Fabric: Built for IOWN,' an all-photonics fabric anchored by an IOWN DCI Gateway (LCoS optical switch doing DWDM/ROADM wavelength routing with zero O-E-O conversion) over a 102.4 Tbps Broadcom-silicon, SONiC-compatible spine for standards-based multi-site optical interconnect

EDGECORE/ACCTON VIA BUSINESS WIRE & TELECOMPAPER

WHAT THIS MEANS

W22 was an optical-interconnect week: four independent vendors pushed CPO and all-photonics from component demos toward deployable fabric, and one analyst house put a ~$154B TAM on it — a textbook Gilder arc where per-GPU bandwidth (and the optical dollar content that carries it) scales faster than compute itself, so value capture migrates into the fabric/optics layer that gates whether GPUs can act as one machine. As OCI-MSA and IOWN-APN standardization let more nodes interconnect coherently across racks and sites, the Metcalfe payoff becomes the explicit design target — networking architects should treat optical scale-up, not just scale-out, as a near-term procurement line.

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 180.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV ~1.3x

Anthropic closed a $65B Series H on May 28 at a $965B post-money — the largest private AI raise on record

Hyperscaler-Hosted

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

BURN / REV ~0.3x

Anthropic's round confirmed ~$15B of previously-committed hyperscaler capital, deepening lab-cloud entanglement

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

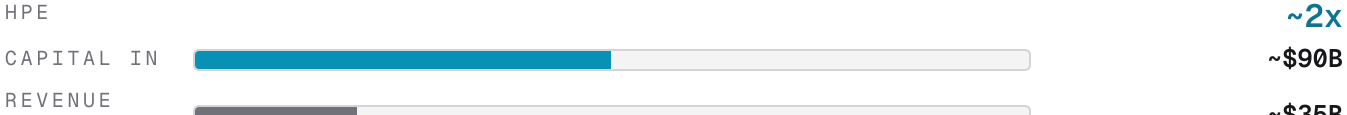
BURN / REV ~3x

IREN signed a ~$9.7B multi-year Microsoft AI-cloud deal (May 29) plus a ~$5.8B Dell GPU purchase

On-Prem / Hybrid

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

BURN / REV ~2x



SoftBank committed up to EUR 75B to build 5GW of AI data centers in France (May 30, Choose France summit)

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.

4/5

CLAIM 01

Anthropic closed a $65B Series H on May 28 at a $965B post-money valuation, becoming the world's most valuable private AI company.

SOURCES: COMPANY PRESS RELEASE (MAY 28) + TECHCRUNCH + BLOOMBERG (WHICH THE PRIOR WEEK FLAGGED A '>$30B / >$900B' ROUND CLOSING 'AS SOON AS NEXT WEEK' — IT CLOSED LARGER). CAVEAT: PRIVATE ROUND, NOT SEC-AUDITED; ~$15B IS PREVIOUSLY-COMMITTED HYPERSCALER CAPITAL, SO FRESH CASH IS MATERIALLY BELOW THE $65B HEADLINE.

Real and historic — the largest private AI raise on record and a clean answer to the W21 watch item (yes, it closed in-window). It resets the private-market ceiling ahead of an IPO race and locks strategic silicon/cloud partners onto the cap table. What would change the read: an eventual S-1 separating fresh cash from in-kind compute commitments.

2/5 — NOISE

CLAIM 02

Anthropic has committed ~$45B to xAI/SpaceX for compute 'through 2029.'

SOURCES: SPACEX S-1 (FILED MAY 20, SEC) STATES $1.25B/MONTH THROUGH MAY 2029 (~$45B CUMULATIVE); BUT THE PRINCIPAL PUBLICLY RECAST IT ON MAY 28 AS A '180-DAY LEASE WITH 90-DAY MUTUAL CANCELLATION,' IMPLYING ~$7.5B. CNBC/IBD FLAGGED THE FILING-VS-STATEMENT DIVERGENCE.

NOISE / overstated. The 'through-2029 / ~$45B' figure is a filing ceiling, not a binding commitment — either-side 90-day termination means only the monthly run-rate is reliable, and the principal himself walked back the duration. It distorts perceived cross-lab compute dependence. What would change the read: an amended S-1 or counterparty confirmation of a minimum term.

2/5 — NOISE

CLAIM 03

OpenAI filed to IPO at ~$1T with a September 2026 listing.

SOURCES: REUTERS/AXIOS/FORTUNE (CONFIDENTIAL DRAFT S-1 REPORTED ~MAY 22). CAVEAT: A CONFIDENTIAL DRS IS NOT VISIBLE ON EDGAR; THE VALUATION AND TIMING ARE REPORTER ESTIMATES, AND THE CFO HAD PREVIOUSLY SIGNALED A 2027 PREFERENCE.

Mostly hype on specifics. The confidential filing is a real signaling step, but '$1T' and 'September' are secondary-sourced extrapolations that conflict with prior company guidance. It matters as the anchor for the entire AI-IPO calendar. What would change the read: a public S-1 on EDGAR.

3/5

CLAIM 04

A new benchmark (DeepSWE) shows SWE-Bench Pro is contaminated, reordering the coding leaderboard.

SOURCES: VENTUREBEAT + DATACURVE (RELEASED MAY 25; DATASET, HARNESS, AND TRAJECTORIES PUBLISHED ON GITHUB). CAVEAT: SINGLE PRIMARY SOURCE WITH COMMERCIAL INTEREST IN A COMPETING BENCHMARK; VERDICTS ARE LLM-JUDGED ON MODEST SAMPLES.

Credibly real as a methodological finding — the .git-history exploit (cited at 18-25% of certain passes) is reproducible from the published artifacts — but self-interested and not yet independently replicated. It implies the open-vs-closed coding gap is partly an artifact. What would change the read: third-party replication of the contamination rate.

4 / 5

CLAIM 05

SoftBank committed up to EUR 75B to build 5GW of AI data-center capacity in France (Choose France, May 30).

SOURCES: SOFTBANK PRESS RELEASE (MAY 30) + SUMMIT COVERAGE. CAVEAT: PHASED AND BACK-LOADED (PHASE 1 ~EUR 45B / 3.1GW BY 2031); AN ANNOUNCED COMMITMENT, NOT YET FINANCED OR UNDER CONSTRUCTION.

Real and the largest single European AI-infrastructure pledge to date, validating the sovereign-AI capital thesis — but the capacity is years out and privately funded rather than state capex, so near-term grid/revenue impact is limited. What would change the read: financing close, groundbreaking, or grid-interconnection milestones.

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**	~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	**$75.2B (Q1 FY27)**	$75.2B (Q1 FY27)	→	Q2 FY27 guide $91B implies further +21% QoQ
Open vs closed gap on SWE-Bench Pro (coding)	**Closed +~19pp (audit caveat)**	Closed +19pp	→	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**~13 / ~$160B+**	12 / ~$80B+	↑	—
PJM 2026/27 capacity auction price ($/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)**	~$0.10-$0.15 (effective, with hidden reasoning tokens)	→	Opus 4.8 fast mode dropped ~3x but no verifiable per-task reading in-window
Custom silicon share of incremental AI compute	**~33-36%**	~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 CAPITAL

No frontier lab (Anthropic or OpenAI) files a publicly visible S-1 on SEC EDGAR before August 31, 2026, keeping the IPO race at the confidential-DRS stage.

65%

BY AUGUST 31, 2026

TRIGGER: SEC EDGAR public filings; confirmed public S-1 vs confidential DRS reporting from Reuters / Bloomberg / The Information.

PREDICTION 02 SOFTWARE

Gemini 3.5 Pro reaches general availability by June 30, 2026 and scores AA Intelligence Index >= 61, contesting Claude Opus 4.8's fresh lead.

60%

BY JUNE 30, 2026

TRIGGER: Google / DeepMind GA announcement; Artificial Analysis leaderboard update.

PREDICTION 03 HARDWARE

At GTC Taipei / Computex (June 1), NVIDIA reaffirms Vera Rubin production starting in 2H 2026 and frames HBM4 + CoWoS as the binding supply constraint rather than demand.

75%

BY JUNE 7, 2026

TRIGGER: NVIDIA GTC Taipei keynote; press coverage; investor notes.

PREDICTION 04 NETWORKING

At least two of (Credo, Marvell, Broadcom) cite co-packaged-optics or 1.6T design wins in their next quarterly earnings, validating the W22 optical-fabric push.

65%

BY AUGUST 31, 2026

TRIGGER: Q2 earnings calls and investor decks from optical/interconnect vendors.

PREDICTION 05 POWER

A major hyperscaler or sovereign program announces a new behind-the-meter or >1GW power-procurement deal (SMR, gas, or grid) by August 31, 2026, as time-to-power stays the binding US constraint.

60%

BY AUGUST 31, 2026

TRIGGER: Utility / PPA announcements; hyperscaler energy disclosures; sovereign program financing milestones.

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 1
WATCH 01

NVIDIA GTC Taipei / Computex keynote

Vera Rubin and Feynman cadence plus HBM4/CoWoS supply color will set the hardware-flywheel baseline for 2H 2026.

JUN 2026
WATCH 02

Gemini 3.5 Pro general availability

Its arrival would directly contest Opus 4.8's fresh AA Index lead and reset the closed-frontier scorecard.

JUN 1
WATCH 03

PJM 2026/27 capacity prices take effect

The $329.17/MW-day cap-bound rate begins billing June 1, hard-coding power as the binding cost constraint for the year.

JUN 2026
WATCH 04

Frontier-lab S-1 / IPO signaling

Whether any public S-1 appears on EDGAR will separate the real IPO calendar from the secondary-sourced '$1T September' headline.

JUN 2026
WATCH 05

Computex optical / CPO follow-through

Post-show, watch whether the W22 co-packaged-optics demos convert into named design wins and shipping fabric products.

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 06

Week 22 of 2026 · May 30, 2026

WEB

brianletort.ai/industry

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