THE AI STACK WEEKLY ISSUE 05 · WEEK 21 OF 2026

THE BOTTOM LINE

The bull thesis got every confirmation it needed. The strategic battleground moved to runtimes.

W21 was the loudest week of Q2 across every lens at once. NVIDIA Q1 FY27 (Tuesday May 20) printed $81.6B revenue (+85% YoY) with Data Center at $75.2B (+92% YoY, +21% QoQ from $62.3B) and a Q2 guide of $91B with zero China DC compute assumed; total supply commitments stepped to $145B; $80B additional buyback authorized and the dividend lifted 25x. Huang said NVIDIA 'will be supply-constrained through the entire life of Vera Rubin' (Q3 2026 production) — the bear thesis that hyperscaler capex isn't translating to chip revenue is gone, and the binding constraint moved firmly to HBM4 and CoWoS packaging.

The frontier-lab valuation curve reset 15x in 14 months. Bloomberg (Thursday May 22) reported Anthropic's $30B-plus round closing 'as soon as next week' at a $900-plus-billion valuation, vaulting past OpenAI's $852B March mark, with Sequoia, Dragoneer, Altimeter, and Greenoaks each writing ~$2B. NVIDIA's Q1 transcript explicitly named Anthropic alongside the hyperscalers as a Blackwell deployment customer, confirming the equity bid matches compute-side demand.

Google I/O 2026 (May 19-22) compressed a normal three-month product cycle into one keynote — Gemini 3.5 Flash (AA Intelligence Index 55.3, Terminal-Bench 76.2%, 4x faster than incumbent frontier at ~280 tok/sec), Gemini Omni Flash (native text-and-image-and-audio-and-video into grounded video out, shipping immediately in Gemini app / Flow / YouTube Shorts), Antigravity 2.0 standalone agent IDE with managed Linux sandboxes, and TPU 8t/8i dual-chip 8th gen. Combined with OpenAI Codex 'Goal mode' GA + macOS Appshots (May 21), Anthropic's self-hosted Managed Agent sandboxes + MCP tunnels (May 19), and xAI Grok Build (W20), the strategic battleground shifted decisively from model intelligence to agent-runtime monetization and ecosystem lock-in.

The specialist canopy kept widening at a fraction of frontier cost. Cohere Command A+ shipped the first-ever Apache 2.0 frontier-adjacent MoE (218B/25B-active, runs on 2x H100, τ²-Bench Telecom jumped 37%→85% generation-over-generation). Microsoft Research Fara1.5-27B (open weights, fine-tuned on Qwen3.5) hit 72% Online-Mind2Web — beating OpenAI Operator (58.3%) and Gemini Computer Use (57.3%) — collapsing the cost structure for browser-agent fleets. Alibaba Qwen3.7-Max (May 20) became the first Chinese model in the AA Intelligence Index top 5 (56.6, ahead of Gemini 3.5 Flash) with a demonstrated 35-hour autonomous tool-use run.

The one bullet dodged: Samsung-union talks reached a tentative wage agreement late Wednesday May 20 — about an hour before the planned 18-day general strike — granting a 10.5% performance bonus pending ratification (vote May 27-28). The W20 downside scenario (Vera Rubin HBM4 supply at acute risk) is largely off the table conditional on the vote passing. So-what for the four audiences: boards should retire 'AI capex peak' theses and treat federal-procurement + capability-gating as durable vendor-risk dimensions; investors should treat the Anthropic close (not the headline valuation) as the binary read of the week and price NVIDIA on supply-not-demand; architects and operators should pull agent-runtime selection into procurement bake-offs now — model swaps inside someone else's harness are constrained.

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

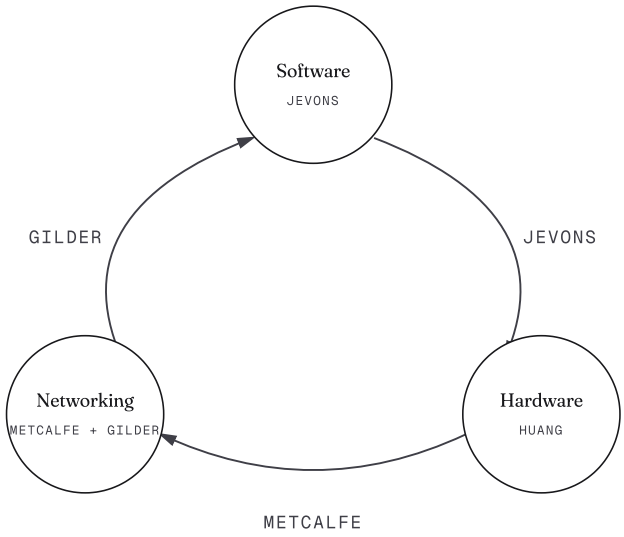
May 23, 2026

Issue 05

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 19
EVENT 01

Google I/O 2026: Gemini 3.5 Flash GA (AA Intelligence Index 55.3, Terminal-Bench 2.1 76.2%, ~280 tok/sec — beats prior-gen Gemini 3.1 Pro on most agent benchmarks at 4x speed; list price $1.50/$9 per Mtok), Gemini Omni Flash (native multimodal create-and-edit with grounded video out, immediately in Gemini app + Flow + YouTube Shorts), Antigravity 2.0 standalone agent IDE with managed Linux sandboxes, TPU 8t/8i dual-chip 8th gen, and Managed Agents API

BLOG.GOOGLE/INNOVATION-AND-AI/TECHNOLOGY/DEVELOPERS-TOOLS/GOOGLE-IO-2026-DEVELOPER-HIGHLIGHTS, DEEPMIND.GOOGLE MODEL CARD, ARTIFICIALANALYSIS.AI, SIMONWILLISON.NET

MAY 20
EVENT 02

Alibaba Cloud Summit ships Qwen3.7-Max (AA Intelligence Index 56.6 — first Chinese model in the global top 5, ahead of Gemini 3.5 Flash; 1M context, demonstrated 35-hour autonomous 1,158-tool-call run), Qwen3.7-Plus-Preview (multimodal sibling), homegrown Zhenwu M890 AI chip (3x predecessor), and a rebuilt full-stack agentic cloud platform

ALIBABACLOUD.COM/BLOG/ALIBABA-UNVEILS-NEW-AI-CHIP-FLAGSHIP-MODEL-AND-REBUILT-CLOUD-STACK-AI-FOR-AGENTIC-ERA, MARKTECHPOST.COM, ARTIFICIAL ANALYSIS

MAY 20
EVENT 03

Cohere ships Command A+ under Apache 2.0 — first-ever fully-permissive license from Cohere on a 218B/25B-active MoE that runs on 2x H100s; τ^2-Bench Telecom jumped 37% (Command A Reasoning) to 85%, Terminal-Bench Hard 3% to 25%, AA-Omniscience Non-Hallucination #1 at 86%; 48 languages with citation grounding for sovereign / on-prem RAG deployments

COHERE.COM/BLOG/COMMAND-A-PLUS, VENTUREBEAT.COM, ARTIFICIALANALYSIS.AI, MARKTECHPOST

MAY 22
EVENT 04

Microsoft Research Fara1.5 family (4B / 9B / 27B, open weights, fine-tuned on Qwen3.5) — 27B variant hits 72% Online-Mind2Web, beating OpenAI Operator (58.3%) and Gemini 2.5 Computer Use (57.3%); released with MagenticLite sandboxed browser plus FaraGen1.5 synthetic-data pipeline; collapses per-task cost for browser-agent fleets

MICROSOFT.COM/EN-US/RESEARCH/ARTICLES/FARA1-5-COMPUTER-USE-AGENT, DECRYPT.CO, AZURE AI FOUNDRY

MAY 22

EVENT 05

Anthropic publishes Project Glasswing initial update — Claude Security beta (Opus 4.7-powered) patched 2,100 vulnerabilities in three weeks; Anthropic reiterates Mythos will NOT release publicly until 'far stronger safeguards' exist despite UK AISI showing Mythos Preview at 73% on expert CTFs and first to autonomously complete the 32-step 'Last Ones' corporate intrusion (3 of 10 runs)

ANTHROPIC.COM/RESEARCH/GLASSWING-INITIAL-UPDATE, AISI.GOV.UK EVALUATION-OF-CLAUDE-MYTHOS-PREVIEWS-CYBER-CAPABILITIES

WHAT THIS MEANS

Two simultaneous arcs: (1) speed-intelligence Pareto frontier reset — Gemini 3.5 Flash now beats prior-gen Pro at 4x throughput, so architects should re-baseline cost-per-task this quarter rather than waiting for Pro-tier price drops; (2) agent-platform lock-in becomes the procurement decision — every flagship this week (Antigravity 2.0, Codex Goal mode, Claude Managed Agents sandboxes, Grok Build, Command A+) is engineered for sustained agentic loops, with the harness (AGENTS.md, SKILL.md, sandbox runtime) now the multi-year commit rather than the model. Boards should treat federal procurement and capability-gating (Anthropic's continued Mythos non-release; METR's first Frontier Risk Report May 19) as durable vendor-risk dimensions, not capability-fade speculation.

LENS 2 OF 3 THE FLYWHEEL

Hardware.

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

MAY 20
EVENT 01

NVIDIA Q1 FY27: revenue $81.6B (+85% YoY), Data Center $75.2B (+92% YoY, +21% QoQ), Q2 guided to $91B (~$4B above Street, zero China DC compute assumed); supply commitments stepped to $145B; $80B buyback authorized; dividend raised 25x; Huang says Vera Rubin (NVL144, 3.6 EFLOPS FP4, Q3 2026 production) 'off to a tremendous start' and NVIDIA 'will be supply-constrained through the entire life of Vera Rubin'

NVIDIA IR PRESS RELEASE, SEC 8-K Q1FY27PR, CFO COMMENTARY, CNBC, REUTERS, MOTLEY FOOL TRANSCRIPT

MAY 20
EVENT 02

Samsung Electronics and the National Samsung Electronics Union reach a tentative wage agreement at ~22:30 KST — about an hour before the planned 18-day general strike (May 21 to June 7, ~48,000 workers, ~40% global DRAM share at risk) — granting a 10.5% special semiconductor performance bonus with no cap; outcome conditional on ratification vote May 27-28; Yonhap industry estimates the strike would have removed 3-4% of global DRAM and 2-3% of NAND supply and directly threatened HBM4 ramp for Vera Rubin

YONHAP, KOREA JOONGANG DAILY, AJU PRESS, SEOUL ECONOMIC DAILY, REUTERS VIA INVESTING.COM

MAY 21
EVENT 03

AMD announces >$10B in Taiwan ecosystem investments tied to Helios rack-scale platform (Instinct MI450X + 6th-gen EPYC 'Venice' CPUs + Pensando 'Vulcano' NICs) on track for multi-gigawatt deployments H2 2026; introduces EFB-based 2.5D packaging for Venice; confirms ODM build-out with Sanmina, Wiwynn, Wistron, Inventec — Helios positioned as the only credible second source to Vera Rubin in the same shipment window

AMD IR PRESS RELEASE VIA MARKETSCREENER, AMD ON X (MAY 21), WCCFTECH, TECHNOSPORTS

MAY 19
EVENT 04

Intel 'Crescent Island' inference accelerator PCB leak: single Xe3P die paired with 20 LPDDR5X packages for 160GB at ~0.94 TB/s on a 640-bit bus, explicitly skipping HBM to sidestep the HBM4 supply crunch; H2 2026 customer sampling, volume early 2027 — first mainstream data-center AI accelerator designed around LPDDR5X as a supply hedge against the SK Hynix / Samsung HBM4 duopoly NVIDIA confirmed on Q1 call

WCCFTECH, TWEAKTOWN, LAVX NEWS, ORIGINAL PCB LEAK BY YUUKI_ANS / @M0_R4_H4

WHAT THIS MEANS

Architects and operators should now treat HBM4 + CoWoS allocation, not announced TFLOPS, as the binding spec for fleet density through 2027: NVIDIA's $145B supply commitments, $91B Q2 guide, and Huang's 'supply-constrained through the entire life of Vera Rubin' push the bottleneck firmly into back-end packaging, and the Samsung tentative agreement keeps that bottleneck one rejected ballot from re-tightening on May 27-28. For boards and investors, Huang's $200B Vera CPU TAM plus an $80B buyback and 25x dividend hike is the Huang flywheel re-leveraging from product to balance sheet — scarcity, not speed, is now the moat heading into AMD Helios' H2 2026 ramp and Intel Crescent Island's HBM-bypassing sampling. Custom-silicon share trajectory remains intact (TrendForce reaffirms ASIC +44.6% CAGR vs GPU +16.1%).

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

MAY 20
EVENT 01

NVIDIA Q1 FY27 segment-reporting split: Data Center Networking revenue $14.8B (+199% YoY, +35% QoQ) — InfiniBand, Spectrum-X, NVLink combined; multi-year silicon-photonics agreements with Coherent, Corning, and Lumentum reaffirmed alongside the Marvell NVLink Fusion partnership

DATA CENTER KNOWLEDGE, NVIDIA INVESTOR MATERIALS, NVIDIA CFO COMMENTARY

MAY 18
EVENT 02

EBO MSA membership expands from 17 to 23 vendors with HPE (marquee server / switch signature), Bellwether, JPC Connectivity, Mixx Technologies, TIME Interconnect, and TFC Communication joining the AMD / Cisco / Meta / Oracle-led expanded-beam optical connector standard

EBOMSA.ORG, SDXCENTRAL, CONVERGE DIGEST, 3M NEWSROOM

MAY 21
EVENT 03

Lightmatter unveils Guide DR liquid-cooled in-chassis Laser NIC in OCP NIC 3.0 form factor at InterConnect 2026 — up to 51.2 Tbps per module and 204.8 Tbps per 1RU switch tray; relocates laser sources from faceplate to chassis to break the CPO / NPO scale-up faceplate bottleneck; samples Q4 2026

LIGHTMATTER PRESS RELEASE, CONVERGE DIGEST, THE AI JOURNAL

MAY 19
EVENT 04

Google I/O keynote concentrated entirely on Gemini 3.5 Flash / Omni Flash and Antigravity 2.0 — no MRC adoption confirmation, no Jupiter / Apollo update, no proprietary fabric alternative disclosed, leaving W19 prediction p13 (second-hyperscaler MRC at >50K-GPU scale) untested at Google's biggest annual stage

GOOGLE DEVELOPERS BLOG, GOOGLE I/O 2026 KEYNOTE, BLOG.GOOGLE

WHAT THIS MEANS

Architects and capex planners should now treat optical interconnect as a peer procurement category to GPUs rather than a GPU attachment: NVIDIA's $14.8B networking quarter (+199% YoY) plus HPE's signature on the EBO MSA put scale-across optics on the same line item as accelerators, while Google's fabric silence at I/O leaves the open MRC + EBO + UEC stack as the de facto hyperscaler standard by default. The Gilder networking-to-software arc is now legible in dollars — bandwidth spend is compounding faster than compute inside the AI factory, and 2026-2027 designs should budget optics as a separately-clearable, multi-vendor stream rather than a NIC-and-cable line buried under GPU.

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 60.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV **0.9**

Anthropic's $30B-plus round at $900B-plus valuation moved from 'in talks' (Bloomberg May 12) to 'closing as soon as next week' (Bloomberg May 22) — Sequoia, Dragoneer, Altimeter, Greenoaks each writing ~$2B; vaulting past OpenAI's $852B March mark. Boards holding 'AI capex peak' theses should retire them this week; investors should treat the close (not the headline valuation) as the binary read.

Hyperscaler-Hosted

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

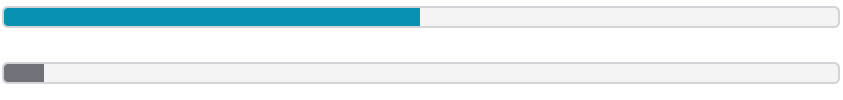
BURN / REV **0.33**

NVIDIA Q1 FY27 (May 20) printed Data Center revenue $75.2B (+21% QoQ from Q4 FY26's $62.3B, +92% YoY) — investors should retire the 'hyperscaler capex isn't translating to chip revenue' bear case; it is, demonstrably, and the Q2 guide of $91B with zero China assumed says next quarter accelerates. Operators planning 2H26 should escalate HBM4 supply assumptions immediately.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

BURN / REV **0.10**

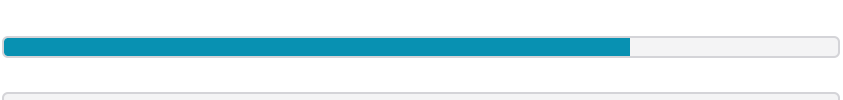


Quiet week for fresh neocloud closures, but NVIDIA Q1 FY27 transcript reframes the category: NVIDIA's $145B supply commitments + in-quarter buybacks while remaining a guaranteed buyer of CoreWeave's unsold capacity (per Reuters reporting on the $6.3B arrangement) means architects must now price the 'circular flow' risk into neocloud GPU contracts — vendor concentration is no longer just a procurement question, it's a counterparty question.

On-Prem / Hybrid

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

BURN / REV **n/a**



Cisco's W20 print (Q3 FY26 raised FY26 AI orders guide $5B to $9B) anchors W21 on-prem demand; NVIDIA Q1 FY27 networking +199% YoY confirms the same demand wave from the chip side. PJM accelerated its proposed backstop reliability auction to September (from March, May 19) for data-center load — architects should pull network / optical line items forward in 2H26 build plans and pre-commit power before pre-committing GPU SKUs.

SIGNAL VS NOISE

What's real, what's noise.

6 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

NVIDIA Q1 FY27 (May 20) Data Center revenue $75.2B (+21% QoQ from $62.3B; +92% YoY); Q2 FY27 guide $91B plus-or-minus 2% with zero China DC compute assumed; total supply commitments + prepaids $145B; $80B additional buyback authorized; quarterly dividend lifted 25x ($0.01 to $0.25).

SOURCES: NVIDIA INVESTOR RELATIONS, SEC 8-K (Q1FY27PR), MOTLEY FOOL TRANSCRIPT, MARKETSCREENER CFO COMMENTARY

Real and decisive. Strongest possible primary-source signal that hyperscaler + AI-cloud + sovereign capex is still translating into chip revenue, and supply not demand is the 2H26 binding constraint. Would change the read only if Q2 prints below ~$86B Street consensus or if a clear China DC compute upside emerges and isn't ramped (signaling demand-side pressure).

4/5

CLAIM 02

Anthropic to close $30B-plus round at $900B-plus valuation as soon as the week of May 25, vaulting past OpenAI's $852B March mark; Sequoia / Dragoneer / Altimeter / Greenoaks each ~$2B; Founders Fund + General Catalyst participating.

SOURCES: BLOOMBERG (MASCARENHAS / TORRENCE / GHAFFARY MAY 22), MONEYCONTROL / ECONOTIMES SYNDICATION

Real but not yet closed in W21. Sourcing is Bloomberg primary plus syndication; round reported by people with direct knowledge but commitments 'still being finalized.' Would change the read if the close slips into mid-June (signal that bond-market term-sheet pressure is biting) or if Microsoft / Salesforce join as strategic allocators (would change the antitrust risk profile).

4/5

CLAIM 03

Samsung Electronics and union reached a tentative wage agreement May 20 ~22:30 KST — about an hour before the planned 18-day strike (May 21 to June 7, 48K workers, ~40% global DRAM share at risk); 10.5% special semiconductor performance bonus with no cap; outcome conditional on ratification vote May 27-28; strike would have removed 3-4% global DRAM and 2-3% NAND supply and directly threatened HBM4 ramp for Vera Rubin.

SOURCES: YONHAP, KOREA JOONGANG DAILY, AJU PRESS, SEOUL ECONOMIC DAILY, REUTERS VIA INVESTING.COM

Real and conditional. Tentative agreement is primary-sourced; ratification vote is binary. Operators planning H2 capacity should hold HBM allocation triage until the vote outcome is reported May 28-29. A ratification rejection would reactivate the W20 supply-risk thesis on three days' notice.

4/5

CLAIM 04

Cohere ships Command A+ under Apache 2.0 — first-ever fully-permissive license from Cohere on a 218B/25B-active MoE that runs on 2x H100s; τ²-Bench Telecom 37% to 85% generation-over-generation; AA-Omniscience Non-Hallucination #1 at 86%; 48 languages with citation grounding.

SOURCES: COHERE.COM/BLOG/COMMAND-A-PLUS, VENTUREBEAT.COM, ARTIFICIALANALYSIS.AI, MARKTECHPOST

Real and structural. First Apache 2.0 frontier-adjacent MoE from a Western lab — sovereign-AI and data-residency procurement teams now have an on-prem option that previously required a closed API or a license-restricted open-weights model. Would change the read only if Cohere reverses the license posture on a subsequent model release, or if a major hyperscaler ships an equally permissive frontier MoE first.

2/5 – NOISE

CLAIM 05

Custom-silicon ASIC share of AI server shipments will reach ~40% by 2030, with hyperscaler in-house silicon (TPU, Trainium, MTIA, Maia) collapsing NVIDIA inference market share to 20-30% by 2028.

SOURCES: TRENDFORCE MARCH 2026, TOM'S HARDWARE MAY 2026, BLOOMBERG INTELLIGENCE (VIA INTROL); NEW STREET MODEL OUTPUT REPRODUCED IN DERIVATIVE COVERAGE

Plausible directionally; the +44.6% YoY ASIC growth vs +16.1% GPU is well-sourced. But the 20-30% NVIDIA inference share by 2028 is a New Street model output — vendor-friendly framing for ASIC co-designers. Thesis requires hyperscalers actually shipping at scale; historically those programs slip. Would change the read if AWS, Google, or Meta publish per-rack TCO numbers proving 65% TCO inference advantage in production.

1/5 – NOISE

CLAIM 06

Microsoft Build 2026 (May 19-22) delivered new MAI / capex / RPO color this week.

SOURCES: INHERITED FROM W20 WATCHLIST FRAMING; NOT CORROBORATED BY ANY IN-WINDOW PRIMARY SOURCE

Hype / misattribution. Microsoft Build 2026 is scheduled June 2-3, 2026 (per Microsoft Foundry blog session schedule), NOT May 19-22 — that window is Google I/O, not Build. The W21 Microsoft news was Surface for Business hardware launch (May 19) and Microsoft India data-center commentary, not capex / RPO disclosures. Watch W23 (June 2-3) for the actual Build content.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **3 rising**, **0 falling**, 7 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	**$75.2B (Q1 FY27)**	$62.3B (Q4 FY26)	↑	+21% QoQ; Q2 FY27 guide $91B implies further +21% QoQ
Open vs closed gap on SWE-Bench Pro (coding)	**Closed +19pp**	Closed +19pp	→	Sustained open lead reshapes enterprise procurement
Sovereign AI commitments (count / aggregate $)	**12 / ~$80B+**	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, with hidden reasoning tokens)**	~$0.10-$0.15 (effective, with hidden reasoning tokens)	→	Gemini 3.5 Flash GA at I/O priced 3x prior Flash list but 'half per-task'
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 HARDWARE

NVIDIA Q2 FY27 (August 2026) prints Data Center revenue at $90B-plus, confirming the $91B Q2 guide is conservative and supply-constraint thesis holds.

70%
BY AUGUST 31, 2026

TRIGGER: NVIDIA Q2 FY27 earnings release; HBM4 supply commentary post-Samsung ratification vote; hyperscaler Q2 capex prints.

PREDICTION 02 CAPITAL

Anthropic round closes at a final post-money valuation between $900B and $1.1T with at least three of (Sequoia, Dragoneer, Altimeter, Greenoaks) named as lead investors publicly.

80%
BY JUNE 13, 2026

TRIGGER: Anthropic primary disclosure; SEC Form D if applicable; lead-investor naming via Bloomberg / Reuters / The Information.

PREDICTION 03 SOFTWARE

At least one Fortune 500 enterprise discloses standardization on a single agent-platform runtime (Claude Code / Codex / Antigravity 2.0 / Grok Build) for greater than $50M annualized commitment.

60%
BY AUGUST 31, 2026

TRIGGER: Enterprise architecture announcements; vendor earnings color from OpenAI partnership disclosures, Anthropic, Google Cloud, xAI; F500 case studies.

PREDICTION 04 NETWORKING

At least three of (Marvell, Broadcom, NVIDIA networking, Arista) post Q2 networking-segment revenue growth greater than 40% YoY, validating EBO MSA momentum into Q2 prints.

70%
BY AUGUST 31, 2026

TRIGGER: Q2 earnings prints from networking-focused vendors; NVIDIA Q1 already at +199% YoY sets the bar.

PREDICTION 05 SOFTWARE

Gemini 3.5 Pro launches by July 15, 2026, scoring AA Intelligence Index greater than or equal to 60, taking the closed-frontier intelligence lead from GPT-5.5 (currently 60.2).

60%
BY JULY 15, 2026

TRIGGER: Google blog / DeepMind announcement; Artificial Analysis leaderboard update; vendor benchmark releases.

WATCHLIST

On the radar.

6 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 25 - 30

WATCH 01

Anthropic round close — final terms, lead investor, full investor list

Bloomberg May 22 reported close 'as soon as next week' at $30B-plus / $900B-plus. Tests prediction p23; reset frontier-lab valuation curve. A slip past end-May or a haircut below $700B would be the loudest re-rating signal in a year.

MAY 27 - 28

WATCH 02

Samsung-union ratification vote on tentative HBM4 agreement

Binary outcome. Approval keeps the 18-day walkout off the table and HBM4 supply for Vera Rubin intact; rejection re-activates the W20 supply-risk thesis on three days' notice and tests prediction p17 the other way.

JUNE 2 - 3

WATCH 03

Microsoft Build 2026 (San Francisco) — MAI-line announcements + Foundry updates

Microsoft notably quiet through I/O week. Build is where MAI-line frontier models, Copilot agent stack moves, and Foundry consolidation are likely to land. Procurement teams running Azure / Foundry should hold model-selection decisions until after Build.

JUNE 1 - 5

WATCH 04

Computex Taipei 2026 + NVIDIA GTC Taipei (Huang keynote June 1)

Computex + NVIDIA GTC Taipei deliver the H2 2026 silicon roadmap (Vera Rubin partner ecosystem, Helios shipping milestones, Crescent Island OEM commitments). First Computex with Saturday-cadence reporting — fully inside W22 window.

JUNE 9 - 13

WATCH 05

Apple WWDC 2026 — on-device model strategy, Apple Intelligence v2

Apple absent from the model-launch cycle through Q2. WWDC is the only major Apple model-disclosure window in 1H. Mobile / edge architects building on-device AI should expect significant Apple Intelligence repositioning and possibly partnered closed-model integration changes.

MAY 26 - JULY 15

WATCH 06

Gemini 3.5 Pro launch — frontier intelligence rematch

Google promised Pro 'next month' (June 2026). This is the model that determines whether Google retakes the AA Intelligence Index lead from GPT-5.5 (currently 60.2) or stays on the second podium tier. Tests prediction p26. Buyers planning Pro-tier procurement should hold pending Pro pricing and benchmarks.

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 05

Week 21 of 2026 · May 23, 2026

WEB

brianletort.ai/industry

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