THE AI STACK WEEKLY ISSUE 18 · WEEK 34 OF 2026

THE BOTTOM LINE

The duration-mismatch thesis just moved into the GPU vendor's own 8-K — NVIDIA is now the residual-value underwriter of a 20-year OpenAI lease, and Marvell is teaching Google how to buy silicon on the revenue that has not yet arrived

Last week this publication argued that the AI infrastructure trade was systematically financing itself on a longer duration than the contracts underneath it, and that the mismatch was structural rather than specific to any one operator. This week that thesis moved into a filing. NVIDIA's Form 8-K dated August 17, 2026 discloses residual value guaranties supporting SB Energy's PORTS-Pike technology campus in Pike County, Ohio, at approximately 4.25 gigawatts of IT load, with an OpenAI affiliate as the tenant on 20-year leases. NVIDIA can optionally support an additional roughly 3.8 GW at the same site. The aggregate payment obligation is cumulatively capped at $105B for the initial commitment, disclosed under Item 2.03 as an off-balance-sheet arrangement, subject to ready-for-service conditions expected beginning 2028, and triggered by tenant insolvency or default under the lease. In a companion transaction NVIDIA committed $1.5B of equity investment into SB Energy, with the site contracted exclusively for NVIDIA compute for 20 years. This publication's filing-derived house measurement is that $105B / 4.25 IT-GW is approximately $24.7B of residual-value credit support per IT-GW of initial commitment — reproducible from the 8-K text alone, and material to how the category should now be modelled.

The first thing to say about the number is what it is not. It is not $105B of cash out the door. It is a cumulative cap on a contingent obligation that turns on a tenant default across a 20-year lease that has not yet commenced service. NVIDIA's July press narrative talked about ~$250B; the filing is $105B, conditional, and phased. Every revenue-opportunity figure the vendor puts alongside this — the ~1.5 million GPUs per generation, the ~$150-200B revenue opportunity language on the NVIDIA blog — is vendor-stated and belongs behind that label. What the filing does say precisely is who is now on the hook if the tenant fails to pay: not the developer, not the neocloud, not the bond insurance market, but the accelerator vendor itself. Once the residual-value assumption on 20-year land-power-shell sits inside a public GPU vendor's off-balance-sheet disclosure, the industry has stopped debating whether GPUs are financeable long assets and started underwriting them like them. Two competing narratives are running through coverage this week — 'NVIDIA becomes a financier' and 'a $105B guarantee' — and both are correct in outline and misleading in emphasis. The tighter reading is that NVIDIA has taken the residual-value bet the neoclouds had been carrying and rehypothecated it onto its own balance sheet's contingent line, which is a structural change in how AI capital gets underwritten and a rating question that will not be resolved by the next earnings call.

The same week produced the mirror image on the silicon side. On August 19, Marvell filed an 8-K disclosing a warrant issued to Google covering up to 58,970,907 shares of Marvell common stock at an exercise price of $206.58 per share — roughly $12.2B if fully exercised and approximately 7% of the company. The commercial agreement underneath is dated July 29 and covers custom silicon around the TPU ecosystem. 1,360,867 shares vest quarterly through the first year; the remaining shares vest in 240 equal tranches, one tranche released for each $500M of Custom Products revenue Marvell attributes to Google under the agreement between Q3 FY2027 and FY2033. Google has no purchase obligation. Multiply the tranche count by the revenue trigger and the implied ceiling is $120B of Custom Products revenue over roughly six fiscal years for Google's silicon relationship with Marvell alone — a ceiling the market has already begun quoting as a headline number even though the vesting is optional and back-loaded. What matters for this issue is the shape of the instrument. Marvell has effectively sold Google the option to acquire ~7% of the company at $206.58 against an August 19 close of $237.27 — modestly in-the-money at disclosure, with nearly all shares still unvested and funded by revenue that has not yet arrived and that only appears if Google actually chooses to source at that scale. The silicon vendor is not underwriting demand in the same way NVIDIA is at PORTS-Pike, but it is issuing equity as a function of a customer's future purchasing decision — which is a different and quieter form of the same residual-value logic.

Around those two structural filings, the rest of the week's capital moved in ways that either confirmed the pattern or made the promised timeline slower. Nebius priced $5B of convertibles on August 19 — $3B of 0.50% notes due 2030 and $2B of 4.50% notes due 2034, upsized from $4.5B, with roughly $4.94B of net proceeds — which is a second public neocloud rolling long paper against shorter customer contracts inside the same quarter that CoreWeave did it. Cerebras announced the CS-4 SUPERNOVA on August 18 with three WSE-3 Turbo processors at 750 vendor-claimed PFLOPS and first shipments at the end of Q3 2026, which is a genuine architectural step and a claim measured on Cerebras's own harness. Etched raised $700M at a $21B valuation, roughly double its June mark, after Jane Street led the round following a rack deployment — an eye-catching valuation move on thin corroboration and best treated as noise-adjacent. Pennsylvania Executive Order 2026-05 dated August 18 bound GRID requirements to data centers above 25 MW, removed those DCs from Permit Fast Track, and conditioned the state tax exemption on GRID compliance while requiring local approvals before DEP permits — the first state-level rule to explicitly rope hyperscale into a slower permitting lane. Read together, the week did four things at once: filed the duration mismatch into a public balance sheet's contingent line, invented a silicon-issuance vehicle indexed to a customer's future purchasing, added a $5B convertible print to the neocloud tenor stack, and lengthened the regulatory path for Pennsylvania sites inside PJM.

The practical instruction is narrower and sharper than 'watch NVIDIA.' If you underwrite AI infrastructure, the residual-value line is no longer a private credit assumption inside a neocloud model. It is now a public off-balance-sheet disclosure at a $4T market-cap company, and the rating implication compounds every time another operator's lease falls under the same guaranty envelope — so read the 10-Q exhibits when they arrive alongside the quarter ended July 26 print. If you buy silicon at scale, the Marvell warrant is the template you will be offered next; require the vesting mechanics in the term sheet before the industry standardizes on a version that only vests when convenient to the vendor. If you plan capacity in PJM territory, PA EO 2026-05 has just added a permitting stage to every site above 25 MW, and this compounds with an interconnect queue already at 60-84 months rather than shortens it. And if you architect agents, the layer with the shortest tenor just moved: Anthropic's Skills API, Files API and computer use went GA on August 20, Salesforce, UiPath and Google converged on identity-inheritance in the same week, and OpenAI cut Sol pricing for a dated three-month window while publishing the reversal date. Each of those is a full home in another publication this week — the important thing to see at the weekly level is that all three are now compressing capital durations and agent extensibility durations against each other faster than any board paper cycles.

AUTHOR

Brian Letort

BrianLetort.AI

PUBLISHED

August 22, 2026

Issue 18

THIS WEEK IN 30 SECONDS

Executive summary.

- Last week's house-original thesis just moved into a filing: NVIDIA's own 8-K discloses a residual-value guaranty on a 20-year OpenAI lease at PORTS-Pike, Ohio, cumulatively capped at $105B on roughly 4.25 IT-GW of initial commitment — the GPU vendor is now the underwriter of land-power-shell, not just a supplier into it.
- Marvell sold Google a warrant on up to ~58.97M shares at $206.58 (~$12.2B fully exercised) that vests only as Custom Products revenue arrives — 240 tranches of $500M each for a $120B ceiling, and Google has no purchase obligation. The residual-value bet is no longer just neocloud credit; it is how silicon vendors underwrite demand.
- Anthropic promoted Skills API, Files API, and computer use to GA on August 20 with HIPAA/BAA availability, 1TB/org file capacity, and progressive-disclosure loading — the first agent extensibility surface that treats skills as versioned artifacts under inherited identity. Full technique read is in the Agent Techniques Weekly.
- Salesforce, UiPath, and Google converged on identity-inheritance for enterprise agents inside the same week (Headless 360 MCP surfaces, Maestro Flow, Antigravity bundled into Gemini Enterprise). Service-account patterns are now the exception rather than the pattern. See the Application Layer.
- OpenAI cut GPT-5.6 Sol pricing by more than 20% on August 21 for a dated three-month promotional window and published the reversal date at launch — the frontier is now pricing on scheduled windows rather than steady state. Full architecture read is in the Model Pulse.
- Pennsylvania's August 18 executive order binds GRID requirements to DCs above 25 MW, removes them from Permit Fast Track, and conditions the tax exemption on GRID — regulatory friction lengthening the interconnect queue rather than shortening it, and the first state-level rule to explicitly rope hyperscale into a slower permitting lane.

BY THE NUMBERS

$105B / 4.25 GW
Cumulative cap and initial IT-GW covered by NVIDIA's residual-value guaranty for the SB Energy PORTS-Pike lease

$120B / 240 tranches
Implied ceiling and structure of Marvell's warrant to Google, one $500M tranche of Custom Products revenue at a time

$4.94B net
Nebius upsized convertible notes offering priced August 19 ($3.0B 0.50% due 2030 + $2.0B 4.50% due 2034)

>20%
OpenAI's dated GPT-5.6 Sol price cut, three-month promotional window with published reversal

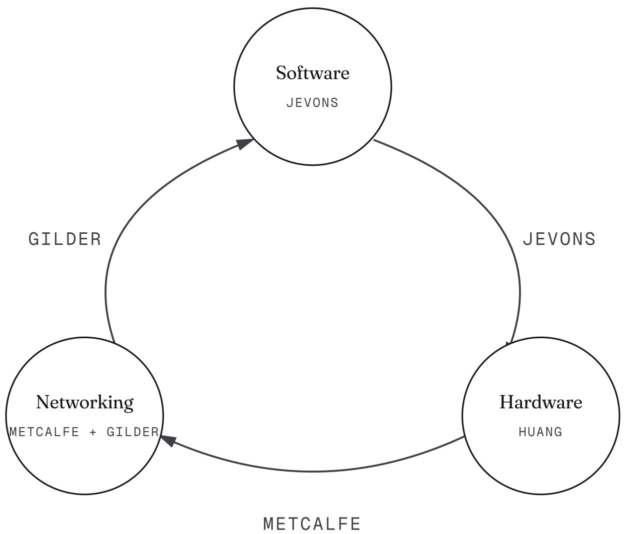
5x / 1TB
Anthropic Files API rate-limit multiple and per-organization capacity at GA (August 20)

>25 MW
Pennsylvania Executive Order 2026-05 GRID threshold binding on data centers effective August 18

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.

AUG 18
EVENT 01

NVIDIA published TensorRT Model Connect as a public preview under Apache 2.0 — a Hugging Face to TensorRT path with no ONNX intermediate and aarch64 wheels, and stated on the release page that the tool was built using Codex agents

NVIDIA TENSORRT MODEL CONNECT ON GITHUB

AUG 18
EVENT 02

OpenAI announced ChatGPT advertising expanding to 31 EU countries with rollout beginning August 24 and separately launched ChatGPT for Teens with new safeguards — consumer surfaces monetizing and segmenting on the same day

OPENAI, COVERAGE VIA TECHCRUNCH AND REUTERS

AUG 19
EVENT 03

Ornith AI released Ornith-1.5 as an MIT-licensed open-weights family — a 397B/35B-active mixture-of-experts flagship with 35B dense and 9B distilled companions, positioned against DeepSeek V4 Pro and GLM-5.3 on vendor-run coding and terminal benchmarks

RUNTIMEWIRE; INDEPENDENT TRACKERS UNRANKED AT PUBLISH

AUG 21
EVENT 04

DeepSeek launched V4-Flash-Vision-Exp as an experimental API-only multimodal model with no downloadable weights and no committed graduation date — a second week of served-versus-downloadable divergence at the same vendor

DEEPSEEK PLATFORM ANNOUNCEMENT

AUG 21
EVENT 05

OpenAI cut GPT-5.6 Sol token pricing by more than 20% for a dated three-month promotional window, publishing the reversal date at launch and preserving the same price on Sol Max — see the Model Pulse for the architecture read

OPENAI GPT-5.6 ANNOUNCEMENT

WHAT THIS MEANS

The software layer's week was a dated cut and a widening definition of what 'open' means, both of which push the same procurement conclusion: agent contracts should be short, indexed, and written with the assumption that both the price and the benchmarked artifact can change without notice. OpenAI publishing a reversal date at launch is now the second frontier vendor to price on a schedule rather than a spot rate in the last six weeks, which turns inference cost from a market to a calendar. DeepSeek serving a vision endpoint whose weights are not obtainable, in the same week Ornith released weights already benchmarked against DeepSeek on the vendor's own harness, means the industry's headline measure of the open-versus-closed gap keeps drifting further from measuring substitutability. Architects should treat inference contracts and skill artifacts as separate durations — one repricing in months, the other in weeks — and stop letting either one anchor the other. The Model Pulse carries the full architecture read on Ornith, the Sol price cut, and DeepSeek-V4-Flash-Vision-Exp.

LENS 2 OF 3

THE FLYWHEEL

Hardware.

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

AUG 17
EVENT 01

NVIDIA filed Form 8-K disclosing residual value guaranties supporting SB Energy's PORTS-Pike, Ohio campus for a 20-year OpenAI-affiliate lease at approximately 4.25 IT-GW, cumulatively capped at $105B for the initial commitment and subject to ready-for-service beginning 2028 (see Big Story for full treatment)

NVIDIA FORM 8-K, AUGUST 17, 2026

AUG 18
EVENT 02

Cerebras announced the CS-4 SUPERNOVA — three WSE-3 Turbo processors delivering a vendor-claimed 750 PFLOPS with first shipments targeted at the end of Q3 2026

CEREBRAS SYSTEMS

AUG 18
EVENT 03

Qualcomm and Modular unveiled Dragonfly AI200 at ModCon with Modular's Mojo 1.0 shipping under Apache 2.0 — no TOPS figure was published for AI200 at launch and the software stack is where the news lives

QUALCOMM AND MODULAR MODCON 2026

AUG 18
EVENT 04

Etched raised $700M at a roughly $21B valuation — approximately doubling its June $10B mark 26 days later — with Jane Street leading after deploying a rack for evaluation, third-party benchmark comparability still limited

TECHCRUNCH

AUG 21
EVENT 05

Anthropic hired former Google TPU founder Amir Salek to lead in-house silicon work, per Bloomberg reporting relayed through secondary coverage — a spend-side commitment against custom accelerators rather than a primary raise

BLOOMBERG VIA SECONDARY COVERAGE

WHAT THIS MEANS

The hardware lens produced two structural filings and one noise call in the same seven days. NVIDIA's disclosure changes how the industry should model residual-value risk: what used to be a private credit assumption inside a neocloud DSCR test is now an off-balance-sheet contingent obligation on a mega-cap public semiconductor company, and it will move rating agency posture on both sides of that trade. Cerebras and Qualcomm/Modular each shipped a real architectural step, but the numbers around them are vendor harnesses without independent third-party reproduction — and Etched raising at $21B on a rack demo is the week's reminder that private-mark inflation on custom accelerators is running ahead of any measured procurement case. The honest read on custom silicon this week is that the demand side (Marvell/Google warrant, Anthropic hiring Salek, Microsoft Maia 300 talks in market) is moving faster than the measured performance side, and the risk is a category priced on optionality rather than on shipped throughput per dollar.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

AUG 19
EVENT 01

Marvell filed an 8-K disclosing a warrant issued to Google covering up to 58,970,907 shares at $206.58 (~$12.2B if fully exercised, ~7% of the company), with 1,360,867 vesting quarterly year one and the remainder in 240 tranches of one per $500M of Custom Products revenue from Q3 FY2027–FY2033 — an implied $120B ceiling, no purchase obligation (see Big Story)

MARVELL FORM 8-K, AUGUST 19, 2026

AUG 19-21
EVENT 02

Hot Interconnects 2026 confirmed NVIDIA co-packaged-optics Ethernet switches shipping in production, with vendors and independent commentators putting optical scale-out at approximately 10% of total compute power in advanced deployments

CONVERGE DIGEST COVERAGE OF HOT INTERCONNECTS 2026

AUG 19-21
EVENT 03

Broadcom detailed Ethernet scale-up, scale-out and scale-across positioning at Hot Interconnects with new switch and optics roadmaps, positioning against NVIDIA CPO on a merchant fabric

BROADCOM HOT INTERCONNECTS SESSIONS, CONVERGE DIGEST

AUG 19
EVENT 04

Ciena launched Nitro 2004 targeting approximately 4 meter copper reach at roughly 1.5 pJ/bit for in-rack and near-rack AI interconnects — the strongest continuing case for copper's role as the near-rack default before optics

CIENA PRESS AND HOT INTERCONNECTS SESSIONS

AUG 18
EVENT 05

Pennsylvania Executive Order 2026-05 imposed binding GRID requirements on data centers above 25 MW, removed qualifying DCs from Permit Fast Track and conditioned the tax exemption on GRID — regulatory friction added to a PJM interconnect queue already at 60-84 months

PENNSYLVANIA GOVERNOR'S OFFICE EXECUTIVE ORDER 2026-05

WHAT THIS MEANS

Networking produced this week's second structural filing and its most consequential policy move on the same three-day window. The Marvell warrant is a different form of the residual-value logic that appeared in NVIDIA's 8-K: NVIDIA is guaranteeing land-power-shell against a tenant default, and Marvell is issuing equity against a customer's future purchasing decision, but both are the industry moving underwriting risk onto public balance sheets that used to sit inside private neocloud credit assumptions. On the physical layer, Hot Interconnects 2026 confirmed that CPO is now a shipping technology from one vendor rather than a coalition-standard technology from many, and the OCP CPO specification's stated Q4 2026 submission window remains the schedule input that determines whether a second source exists for 2027-28 refresh cycles. Pennsylvania's EO 2026-05 is the reminder that fabric planning still sits inside a physical constraint stack, and that constraint just got heavier at the largest US market for hyperscale demand. Architects should now separate three fabric decisions explicitly rather than roll them together: single-vendor CPO now, multi-vendor CPO later, and the near-rack copper layer where Ciena's Nitro 2004 keeps the pluggable window open through the transition.

CAPITAL FLOW

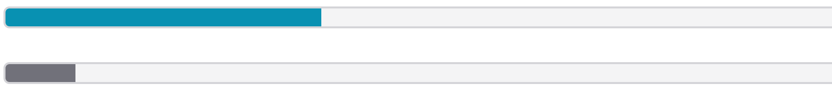
Money in, revenue out.

Capital deployed (forward) vs revenue out (quarterly or run-rate). Burn-to-revenue = revenue / capital — lower means more out than in. Bars normalized to 250.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV **~4.5x**



CAPITAL IN — ~$95B

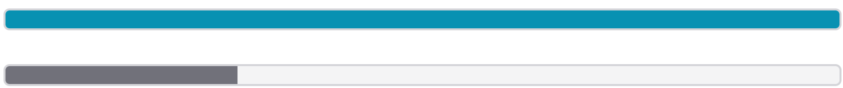
REVENUE OUT — ~$21B

No lab closed primary financing in-window. Anthropic hired former Google TPU founder Amir Salek to lead in-house silicon work — a spend-side commitment against custom accelerators rather than a raise. OpenAI's Sol price cut and EU ads rollout are product-and-monetization moves that do not change the lab's aggregate capital or disclosed revenue in-window.

Hyperscaler-Hosted

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

BURN / REV **~3.6x**



CAPITAL IN — ~$250B

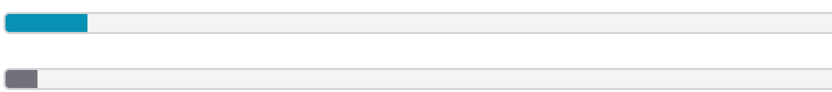
REVENUE OUT — ~$70B

No hyperscaler reported or revised capex guidance in-window. The category's motion was structural rather than budgetary: NVIDIA's residual-value guaranty on the SB Energy PORTS-Pike site is a contingent obligation on the accelerator vendor tied to an OpenAI affiliate's 20-year lease, and Marvell's warrant to Google indexes silicon vendor equity to custom-product revenue. Google also bundled Antigravity into Gemini Enterprise on August 20 — see the Application Layer.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

BURN / REV **~3.1x**



CAPITAL IN — ~$24.6B

REVENUE OUT — ~$8B

Nebius priced $5B of convertible notes on August 19 — $3B of 0.50% notes due 2030 and $2B of 4.50% notes due 2034, upsized from $4.5B, with roughly $4.94B in net proceeds — the largest AI-infrastructure convertible print of the year. No neocloud reported earnings in-window.

On-Prem / Hybrid

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

BURN / REV **~2.7x**

CAPITAL IN — ~$103B

REVENUE OUT — ~$38B

Pennsylvania Executive Order 2026-05 on August 18 bound GRID requirements to data centers above 25 MW, removed qualifying DCs from Permit Fast Track, and conditioned the state tax exemption on GRID compliance — regulatory friction added to the largest US hyperscale market's interconnect queue. Cerebras announced CS-4 SUPERNOVA with vendor-claimed 750 PFLOPS and first shipments at end of Q3 2026.

SIGNAL VS NOISE

What's real, what's noise.

7 claims that drove headlines this week, scored 1–5 on source quality and triangulation. 3 flagged as noise. The bar is at least one explicit noise call per issue.

5 / 5

CLAIM 01

NVIDIA's Form 8-K discloses residual value guaranties supporting SB Energy's PORTS-Pike, Ohio campus at approximately 4.25 IT-GW for a 20-year OpenAI-affiliate lease, cumulatively capped at $105B for the initial commitment, disclosed under Item 2.03 as an off-balance-sheet arrangement and triggered by tenant insolvency or default.

SOURCES: NVIDIA FORM 8-K FILED AUGUST 17, 2026; NVIDIA NEWSROOM ANNOUNCEMENT; NVIDIA CORPORATE BLOG. EXHIBITS PENDING WITH THE 10-Q FOR THE QUARTER ENDED JULY 26, 2026.

Structural change in how AI capital gets underwritten, disclosed in a primary filing. The residual-value assumption that used to live inside neocloud credit models is now on a mega-cap public semiconductor company's contingent-obligation line, and the rating implication compounds every time another site falls under the same guaranty envelope. Model the 10-Q exhibits when they arrive. The $105B figure is a cumulative cap on a contingent obligation, not cash out the door, and every revenue-opportunity figure NVIDIA has attached to the transaction is vendor-stated — treat the label accordingly.

---

5 / 5

CLAIM 02

Marvell issued Google a warrant on up to 58,970,907 shares at $206.58 (~$12.2B if fully exercised, ~7% of the company) vesting one $500M-revenue tranche at a time over 240 tranches from Q3 FY2027 through FY2033, implying a $120B ceiling of Google Custom Products revenue with no purchase obligation.

SOURCES: MARVELL FORM 8-K FILED AUGUST 19, 2026; COMMERCIAL AGREEMENT DATED JULY 29 COVERING CUSTOM SILICON AROUND THE TPU ECOSYSTEM.

The other structural filing of the week and a different form of the residual-value logic. Silicon vendor equity is now indexed to a hyperscaler's future purchasing decision at 240 discrete revenue triggers, funded by revenue that has not yet arrived. Enterprises buying at scale should require warrant-style vesting mechanics in their term sheets before the industry standardizes on a version that only vests when convenient to the vendor. The $120B ceiling is the multiplication of tranche count and trigger; do not carry it as a revenue forecast.

---

4 / 5

CLAIM 03

Nebius priced $5B of upsized convertible notes on August 19 — $3B of 0.50% notes due 2030 and $2B of 4.50% notes due 2034, with roughly $4.94B in net proceeds — the largest AI-infrastructure convertible print of the year.

SOURCES: NEBIUS GROUP FORM 6-K FILED AUGUST 19, 2026.

Public disclosure at scale by an audited issuer, so directionally reliable. It continues the pattern this publication named in W33 — long paper against shorter contracts — with a second neocloud rolling the same tenor stack inside the same quarter as CoreWeave. Rates term structure (0.50% near, 4.50% far) is the market rewarding the growth story on the near tranche and demanding a step-up on the further one, which is a rational and unremarkable answer to a still-forming private-mark environment. The number itself is signal; the structural implication for GPU residual value beyond 2030 remains a hypothesis rather than a settled question.

---

3 / 5

CLAIM 04

Cerebras CS-4 SUPERNOVA delivers 750 PFLOPS using three WSE-3 Turbo processors, with first shipments at the end of Q3 2026 — a step change in wafer-scale performance for large model training and inference.

SOURCES: CEREBRAS SYSTEMS PRODUCT ANNOUNCEMENT DATED AUGUST 18, 2026. NUMBER IS A VENDOR CLAIM MEASURED ON CEREBRAS'S OWN HARNESS; NO INDEPENDENT THIRD-PARTY REPRODUCTION AVAILABLE AT PUBLISH.

Architectural progress is real and the shipping timeline is credible. The PFLOPS number is a vendor claim without independent verification, and Cerebras's methodology on prior generations has diverged from Artificial Analysis measurement on comparable workloads. Treat the announcement as directionally reliable for planning and as vendor-stated for procurement. On-prem operators evaluating wafer-scale should schedule a measured comparison against a Blackwell-Ultra reference workload rather than rely on the harness delta.

---

2 / 5 — NOISE

CLAIM 05

Etched's transformer-ASIC valuation doubled to $21B in 26 days on the basis of a Jane Street rack deployment — a signal that dedicated inference silicon is being repriced by the market against merchant GPUs.

SOURCES: TECHCRUNCH REPORTING ON THE ROUND DATED AUGUST 18, 2026, CITING ANONYMOUS SOURCES ON THE PRICING AND JANE STREET'S PARTICIPATION.

Noise-adjacent. Private-mark inflation of 100% in a month is not evidence of category revaluation; it is evidence of a specific investor's willingness to lead at that mark. Jane Street's rack deployment is a datapoint, not a benchmark, and dedicated-inference ASIC economics still depend on measured throughput per dollar on a stable workload that has not been published. The category may well reprice but nothing in this transaction supports the claim as it stands.

---

2 / 5 — NOISE

CLAIM 06

HBM/DRAM Korean press reports on August 20 indicate a further tightening at Samsung and SK Hynix that will require accelerator vendors to de-spec memory configurations across the next generation, with material shipment implications for 2027.

SOURCES: GRADE-3 KOREAN PRESS AGGREGATION DATED AUGUST 20, 2026, UNSOURCED BACK TO A MANUFACTURER DISCLOSURE.

Directionally consistent with what Micron said publicly in W33 and with the pattern this publication has been tracking, but this specific report is aggregation without a manufacturer citation and should not be carried as a fifth-grade signal. The underlying constraint is real; the specific de-spec framing is a rumour continuation without primary confirmation. Wait for a manufacturer disclosure or an accelerator-vendor product page rather than move a share estimate on a Korean press item.

---

1 / 5 — NOISE

CLAIM 07

SB Energy is preparing a $50B initial public offering on the strength of its PORTS-Pike hyperscale contracts, per unnamed sources cited across secondary coverage of the NVIDIA guaranty announcement.

SOURCES: SECONDARY COVERAGE TYING THE NVIDIA GUARANTY ANNOUNCEMENT TO A RUMOURED SB ENERGY IPO VALUATION; NO PRIMARY SOURCE, NO FILING, NO COMPANY CONFIRMATION.

Lowest-confidence claim of the week and the one most likely to end up in a slide. Every part of the number is inference: the valuation, the timing, and even that a listing is being prepared. Anchoring a $50B private mark to the same-day announcement of a contingent guaranty conflates two very different balance-sheet events, and no company involved has confirmed any listing preparation. Do not quote the number in any comparison table.

SYNTHESIS

The week, reasoned through.

Cross-domain connections, the standing thesis tested against this week's evidence, and the patterns building across weeks. Every inference is labeled by reasoning type and linked to its evidence.

CONNECTING THE DOTS

The residual-value bet on AI infrastructure has moved from private neocloud credit assumptions into two different public-issuer disclosures in the same week — NVIDIA guaranteeing land-power-shell on a tenant default and Marvell indexing equity to a hyperscaler's future silicon purchasing — which means the underwriting question is no longer whether GPUs are a financeable long asset, but who publicly carries the residual-value bet once they are.

ABDUCTIVE · 78% CONFIDENCE

NVIDIA's Form 8-K dated August 17 discloses residual-value guaranties supporting SB Energy PORTS-Pike at approximately 4.25 IT-GW for a 20-year OpenAI-affiliate lease, cumulatively capped at $105B, Item 2.03 off-balance-sheet, triggered by tenant insolvency or default.

Marvell's Form 8-K dated August 19 discloses a warrant to Google covering up to 58,970,907 shares at $206.58 with the remainder vesting one $500M-revenue tranche at a time across 240 tranches from Q3 FY2027 through FY2033, implying a $120B ceiling with no purchase obligation.

Nebius priced $5B of upsized convertibles on August 19, rolling long paper (0.50% 2030 + 4.50% 2034) against still-shorter customer contracts — a second neocloud in the same quarter as CoreWeave doing the same duration trade.

The 8-K exhibits from NVIDIA are still to come with the 10-Q for the quarter ended July 26, which means the auditability of the residual-value line grows this quarter regardless of whether NVIDIA reports a revenue surprise.

> **STEEL-MAN:** The strongest counter is that this reading conflates three genuinely different instruments — a contingent obligation, a warrant, and a convertible — into a single narrative when they answer different capital questions. NVIDIA guaranteeing a tenant default is not the same instrument as Marvell issuing Google an equity option indexed to future purchasing, and Nebius issuing convertibles is normal growth-cap-stack behaviour rather than novel duration engineering. There is real risk of pattern-matching separate filings into a shared story. The claim still stands in a narrower form: what all three do share is that the residual value of AI infrastructure — either the site behind the accelerator, the accelerator vendor's equity, or the neocloud debt — is being underwritten longer than the contracts underneath it, and that is a common structural feature rather than a coincidence.

The framework's hypothesis that Metcalfe's law and the power constraint reinforce each other predicted the fabric layer's pricing behaviour, and this week's PA EO 2026-05 confirms the mechanism from the regulatory side by lengthening the interconnect queue for the largest US hyperscale market — which means the two-vendor CPO decision (single-source now versus multi-vendor Q4 2026 spec later) becomes a scheduling question against a queue that is measurably getting longer rather than shorter.

DEDUCTIVE · 66% CONFIDENCE

The working framework's Metcalfe law predicts that interconnect value scales as the square of connected nodes, and hypothesis 5 holds power as the binding constraint on deployment — together these predict that campuses fragment under power limits and drive optics demand, which W33's Cisco call commentary directly supported.

Pennsylvania Executive Order 2026-05 on August 18 binds GRID requirements to DCs above 25 MW, removes qualifying DCs from Permit Fast Track and conditions the tax exemption on GRID compliance — the first state-level rule to explicitly rope hyperscale into a slower permitting lane in the largest US market.

Hot Interconnects 2026 confirmed NVIDIA CPO Ethernet switches shipping in production with vendors and independent commentators putting optical scale-out at approximately 10% of total compute power, and Ciena launched Nitro 2004 at approximately 4m copper reach at 1.5 pJ/bit as the near-rack default before optics.

The OCP Open Silicon Photonics workstream's stated Q4 2026 specification submission target remains the fork in the road for whether a multi-vendor path exists for 2027-28 fabric refresh cycles.

> **STEEL-MAN:** The strongest counter is that PA EO 2026-05 is one state's regulatory posture rather than a national trend, and that fabric-refresh scheduling decisions run off national accelerator supply and hyperscaler capex plans that PJM-region permitting does not gate. If operators respond by routing new sites out of PJM into ERCOT, MISO, or Nordic markets — as several have already signalled they can — the queue-length effect is a Pennsylvania problem rather than a national fabric-choice problem. The deduction still holds because it does not depend on PA being unique; it depends on state-level friction lengthening some part of the effective queue somewhere, and PA EO 2026-05 is direct evidence of that mechanism arriving in the market that matters most for hyperscale demand. If a comparable rule appears in a second major-market state within the quarter, the claim gets stronger; if PA reverses, it gets weaker.

Anthropic's Skills API and computer use GA on August 20 and Salesforce, UiPath, and Google's identity-inheritance releases in the same week are converging on the same architectural pattern — versioned skill artifacts running in a hosted sandbox under the caller's identity — without any coordinating body having asked for it, which suggests the industry has decided identity inheritance is the enterprise-viable default and that service-account patterns have become an exception carrying a written justification rather than a routine choice.

INDUCTIVE · 72% CONFIDENCE

Anthropic Skills API, Files API and computer use graduated to GA on August 20 with progressive-disclosure loading, HIPAA/BAA availability, 5x file rate limits and 1TB per organization — skills executing in Anthropic's sandbox with outbound calls under the caller's identity.

Salesforce expanded Headless 360 on August 19 with MCP servers, skills and identity/permission inheritance across Data Cloud, Agentforce, MuleSoft and Tableau — agents calling Salesforce data under caller entitlements rather than service credentials.

UiPath launched Maestro Flow on August 19 for identity-scoped agent orchestration, and Google bundled Antigravity into Gemini Enterprise on August 20 for identity-inherited agent workflows on the Google identity plane.

Alipay's agentic commerce release on August 17 uses MCP as commerce infrastructure with per-user consent scoping rather than aggregate merchant credentials — the same architectural choice on a different vendor's substrate.

> **STEEL-MAN:** The strongest counter is that four announcements in a single week are a coincidence of release calendars rather than a convergent architectural decision, and that identity inheritance is what these vendors have been building for six to twelve months regardless of what any of them shipped this week. The 'convergence' language may also be reading a shared pattern into products that solve the identity problem in materially different ways (Salesforce's is a data platform choice, Anthropic's is a lab-side sandbox, Google's is a hyperscaler identity plane). The claim survives because the useful thing to say about the pattern is not who moved first but that the enterprise architecture question — service account or inherited identity — now has one visible answer across the three surfaces most enterprises will encounter, and buyers who assume the older default in a term sheet drafted today are the exception rather than the pattern. See the Agent Techniques Weekly for the full technique treatment and the Application Layer for the platform-side read.

THESIS TEST

H1 · SUPPORTED **The cycle is accelerating, not slowing.** The frontier lab release cadence continues to compress: OpenAI cut GPT-5.6 Sol pricing more than 20% inside a dated three-month window, Anthropic promoted Skills API and computer use to GA, DeepSeek launched V4-Flash-Vision-Exp as an experimental API model, and Ornith AI shipped a 397B/35B-active MoE flagship with 35B and 9B companions in one release. Agent tooling GA'd inside the same seven-day window as consumer-tier product surfaces monetized in 31 EU countries. The framework's law that pricing on a schedule rather than a spot rate is now the observable behaviour of the top tier is holding. **AGAINST IT:** The main honest counter is that this week's biggest lab-side events were about extensibility and pricing rather than capability advances. Anthropic did not ship a new base model; OpenAI did not raise Sol's ceiling capability; Ornith's benchmarks are vendor-run. Release cadence accelerating while pretraining cadence does not is the weaker form of the hypothesis, and this week's evidence sits on that side of the line.

H2 · SUPPORTED **Capital is concentrated, returns are diffuse.** The two structural filings of the week are direct evidence: NVIDIA carrying a $105B cumulative cap of contingent obligation for a single OpenAI-affiliate site and Marvell issuing Google an equity option indexed to $120B implied Custom Products revenue. Concentration is compounding. Nebius adding $5B of convertible paper against still-shorter contracts continues the same shape. Returns remain diffuse on the enterprise side — the Application Layer documents Salesforce, UiPath, Google and Alipay all shipping identity-inheritance features in the same week, spreading agent value across horizontal and vertical platforms rather than concentrating it back to the lab. **AGAINST IT:** The honest counter is that both filings are contingent instruments, not booked capital. NVIDIA's $105B is a cap on a payment obligation triggered by tenant default with obligations only beginning 2028; Marvell's warrant only vests as Google chooses to purchase. Neither counts as capex concentration in the near-term reporting sense. The hypothesis still holds because the underwriting concentration — who has to carry the residual value if things go wrong — is now on public balance sheets in a way it was not a month ago.

H3 · SUPPORTED **Networking is the durable layer.** Hot Interconnects 2026 confirmed NVIDIA CPO Ethernet switches shipping in production, Broadcom staked scale-up/out/across positioning on merchant fabric, and Ciena launched Nitro 2004 for the near-rack copper window. Marvell's warrant to Google indexes the silicon vendor's equity value to hyperscaler purchasing behaviour on custom silicon, which is the closest thing to a valuation-fed pricing-power confirmation the networking layer has produced. Optical scale-out at approximately 10% of compute power in advanced deployments is the durable-layer share showing up as a fabric-power datapoint rather than an ARR one. **AGAINST IT:** Single-vendor CPO shipping while OCP's multi-vendor specification is still on a Q4 2026 target and ASE says the test ecosystem is not ready means the durable pricing power may capture to one vertically integrated silicon vendor rather than accrue to networking suppliers as a class. If the merchant fabric layer loses share to NVIDIA CPO across the next refresh cycle, the durability applies to a smaller category than the hypothesis assumes.

H4 · STRAINED **Open weights pull the floor up.** Ornith-1.5 shipped as MIT-licensed at 397B/35B-active with vendor-run benchmarks against DeepSeek V4 Pro and GLM-5.3, and no independent tracker had ranked it at publish. DeepSeek in parallel launched V4-Flash-Vision-Exp as API-only with no downloadable weights. The mechanism the hypothesis depends on — open weights re-routing compute demand to on-prem and sovereign deployment — requires that the weights be obtainable in the benchmarked form, and this week produced one open release that has not been externally benchmarked and one served build that is not downloadable at all.

H5 · SUPPORTED **Power is the binding constraint for the next 24 months.** NVIDIA guaranteeing a 20-year land-power-shell arrangement at 4.25 IT-GW in Ohio is the strongest single week of evidence for power (and grid interconnect) as the binding physical constraint the framework has produced. The deal exists because power at that scale is the scarce resource the tenant is contracting for, and NVIDIA writing a residual-value guaranty against tenant default is the accelerator vendor absorbing what would otherwise be an unfinanceable siting risk. Pennsylvania Executive Order 2026-05 in the same week added a permitting-side friction to sites above 25 MW in the largest US hyperscale market — a state directly acting on the constraint. The honest ambiguity is whether NVIDIA's absorption of the residual-value bet should be read as power constraint (evidence for the hypothesis) or as capital reroute (evidence for hypothesis 2 instead). This publication reads it as power-first because the guaranty is contingent on the physical site rather than on the accelerator. **AGAINST IT:** The strongest counter is that NVIDIA writing a residual-value guaranty on a 20-year lease could be read as demand routing capital around a power constraint that stops binding once the site is in service in 2028 — in which case this week's evidence is capital solving power on the ground rather than power binding capital. Under that reading the hypothesis is supported for the near term and structurally weakening for the medium term. The verdict remains supported because the guaranty exists at all only because power at scale is scarce, but the counter is honest and worth naming.

PATTERN WATCH

Silicon vendors are moving residual-value and demand-underwriting risk from private customer credit assumptions onto their own public disclosure lines — first as commercial narrative, then as filed contingent obligations and warrants. (2 WEEKS OBSERVED)

Next week: At least one additional silicon vendor or major hyperscaler counterparty discloses a residual-value guaranty, warrant, or take-or-pay-linked equity instrument tied to future purchasing behaviour in an SEC or equivalent filing by October 31, 2026. Falsified if the next eight weeks produce no such filing and no vendor commentary confirming the trend at a scheduled earnings event.

Enterprise agent surfaces are converging on identity inheritance across the agent boundary as the default authentication pattern, deprecating service-account-based orchestration as an exception rather than a routine choice. (2 WEEKS OBSERVED)

Next week: At least one major enterprise vendor publishes a deprecation notice or migration path for service-account-based agent orchestration by December 31, 2026, or two additional platforms ship inherited-identity MCP surfaces in Q4. Falsified if two consecutive quarterly release cycles produce new agent surfaces built primarily on service-principal patterns without an inherited-identity alternative.

SECOND-ORDER EFFECTS

The rating agencies now have to model contingent-obligation exposure on the GPU vendor whose credit was previously priced almost entirely off product margins. Even without a call event, the disclosure changes the ratings math on future issuance and on any hyperscaler counterparty that would sit under the same guaranty envelope — which will show up as tighter documentation on the next comparable lease before it shows up as a downgrade. Trigger: NVIDIA disclosed a $105B cumulatively capped residual-value guaranty as an Item 2.03 off-balance-sheet arrangement with exhibits pending in the 10-Q for the quarter ended July 26, 2026. Horizon: 3-9 months. Who moves: NVIDIA rating agencies, SB Energy investors, hyperscaler tenants negotiating similar leases, insurance underwriters writing residual-value cover, and neocloud borrowers whose next facility now has a public comparable for the guaranty market.

The warrant becomes a template. Broadcom, Astera Labs, Alchip and every custom-silicon vendor selling into a hyperscaler will be offered — or asked to accept — a similar instrument on the next major supply agreement. Enterprises buying at scale will either standardize the vesting mechanics in their term sheets before the industry converges on a vendor-friendly version, or inherit whatever the first two or three major deals set as the norm. Trigger: Marvell's warrant to Google indexes silicon vendor equity to a customer's future purchasing decision with 240 discrete revenue triggers and no purchase obligation. Horizon: 6-18 months. Who moves: Custom-silicon vendors, hyperscaler procurement organizations, buy-side analysts pricing merchant-versus-custom share, and legal/tax teams designing the next generation of long-term supply contracts.

Site selection for above-25-MW campuses shifts materially against Pennsylvania and toward ERCOT, MISO and Nordic markets. The demand does not disappear; it routes around the state, which lengthens effective queue times in less-friction markets by adding demand and puts fabric planning under further pressure to design for multi-site training runs that cross ISO boundaries. Trigger: Pennsylvania Executive Order 2026-05 binds GRID requirements to DCs above 25 MW and conditions the state tax exemption on GRID compliance, adding a permitting-side stage to the largest US hyperscale market's site pipeline. Horizon: 6-24 months. Who moves: PJM-region developers, hyperscaler site-selection teams, ERCOT and MISO grid operators anticipating displaced demand, and network operators being asked to design more inter-DC fabric than the current physical layer supports.

STRATEGIC OUTLOOK

The twelve-month posture this week argues for is to treat the residual-value bet on AI infrastructure as a public disclosure item rather than a private credit assumption, and to write it into vendor term sheets before the market standardizes on the version that is most convenient to the issuer. NVIDIA has moved the residual-value line inside its own contingent-obligation footnote, Marvell has moved it into a warrant vesting mechanism, and Nebius has extended the neocloud tenor stack another $5B against still-shorter contracts — three different vehicles, one shared structural feature. For a board setting AI infrastructure strategy into 2027, the useful diagnostic is not capacity or capability but disclosure surface: which counterparties in your stack have written the residual-value bet down publicly, in what form, on what schedule of exhibits. If yours has not, it is either not thinking about it or it is not telling you. On the shorter-tenor side, the agent-tooling layer just re-based on Anthropic's Skills API and computer use GA — a decision that will look either premature or obviously right depending on whether OpenAI ships a comparable versioned-skill surface inside the quarter — and the OpenAI Sol promotional window has published its own reversal. Contracts in that layer should be dated, indexed, and written with the assumption that both the price and the benchmarked artifact can change without notice. Pennsylvania's EO 2026-05 is a reminder that the physical layer still owns the schedule, and the last twelve months of net site displacement out of PJM will be measurable well before any of the near-tenor layers rebase.

WHERE WE DIFFER

Our read against the field's published takes on the same week — on the record, so you can score us later.

DIFFER General coverage across financial and technology press, August 18-21: NVIDIA has effectively pledged $105B to guarantee the SB Energy PORTS-Pike campus, becoming the financier behind OpenAI's next 20-year home.

Our read: The framing collapses two important distinctions. The $105B is a cumulative cap on a contingent obligation triggered by tenant default across a 20-year lease with obligations subject to ready-for-service starting 2028 — it is not a $105B pledge and it is not cash out the door. And 'becoming the financier' overstates the shift: the developer is SB Energy, the tenant is an OpenAI affiliate, and NVIDIA is the residual-value underwriter on the lease alongside a $1.5B equity investment in SB Energy. The structural point is real and larger than the framing: the accelerator vendor now sits inside the contingent-obligation footnote for land-power-shell, which is a rating question. But the number and the identity of the underwriter both need the primary-filing precision.

DIFFER Aggregated secondary coverage referencing the July press narrative: The July NVIDIA-OpenAI announcement was worth $250B; the filing confirms the deal at that scale.

Our read: The July $250B figure was a press narrative that the filed 8-K does not confirm. The filed number is a $105B cumulative cap on a contingent obligation for the initial commitment, with an optional additional ~3.8 GW capacity NVIDIA could support but has not committed to. Anyone quoting $250B this week is quoting a July press figure that the primary filing revised downward and conditioned, and the revision matters because the ratings implication runs off the filed number rather than the press one.

OPEN Cerebras marketing materials and press aggregation: Cerebras CS-4 at 750 vendor-claimed PFLOPS is roughly a 30x lift over Blackwell-Ultra rack performance on comparable workloads.

Our read: The 30x framing is a vendor comparison against a specific reference configuration and specific workload, not an independently reproduced measurement. Cerebras's architectural step is genuine and the shipping timeline is credible; the specific multiple over Blackwell-Ultra depends on assumptions about model, precision, batch size, and utilization that are not visible from the marketing page. We are not calling this either way until independent measurement lands. On-prem operators evaluating wafer-scale should schedule a measured comparison rather than accept the vendor multiple.

DIFFER Aggregated financial-press coverage of the August 19 8-K: Marvell's warrant to Google is a $120B silicon supply agreement.

Our read: The $120B figure is the implied ceiling of the vesting mechanism (240 tranches × $500M revenue trigger), not a booked supply agreement and not a Google purchase commitment. Google has no obligation to source at that scale, and the tranches only vest as attributed Custom Products revenue actually arrives across Q3 FY2027 through FY2033. Calling this a $120B supply deal reads the maximum of a forward option as an executed contract. The right framing is that Marvell issued Google an optionality vehicle indexed to a customer's future purchasing — a real instrument, worth watching, but not a booked commitment.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **0 rising**, **1 falling**, 9 steady. Each metric carries a threshold value where the read materially changes.

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Frontier lab cash runway at current burn	~30-40 months, unchanged — no lab closed primary financing in-window. Anthropic's hire of former Google TPU founder Amir Salek for in-house silicon is a spend-side commitment that turns forward cash needs upward without disclosing a delta; OpenAI's Sol promotional cut and EU ads rollout are demand-side moves that do not change disclosed runway	~30-40 months, unchanged — no lab closed primary financing in-window. OpenAI's ~$7B transaction was an employee tender at a flat $852B mark, which supplies liquidity to shareholders rather than capital to the business	→	Below 18 months for any top-four lab

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Hyperscaler AI capex to disclosed AI revenue ratio	**~3.6x, unchanged — no hyperscaler reported or revised guidance in-window. NVIDIA's residual-value guaranty on the SB Energy PORTS-Pike site is a contingent obligation at the accelerator vendor, not hyperscaler capex, and does not enter the numerator until it is called; Google's Marvell warrant is silicon vendor equity indexed to future purchasing rather than a capex disclosure**	~3.6x, unchanged — no hyperscaler reported or revised guidance in-window; the last moves were the late-July earnings cycle and nothing this week touched either side of the ratio	→	Above 6x sustained for two consecutive quarters

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
CoreWeave contracted revenue backlog	**$104.2B as of June 30, unchanged with no CoreWeave reporting event in-window — the next scheduled print is the Q3 filing. Nebius's $5B convertible print on August 19 is the category's parallel move and confirms the tenor pattern that made backlog conversion the binding lever, without changing the CoreWeave number**	$104.2B as of June 30, up 4.8% sequentially from $99.4B and 246% year over year, excluding more than $25B of commitments added in early Q3 — but conversion crossed the threshold: FY2026 revenue guidance of $12.4-13.2B implies 11.9-12.7% annual conversion, below the 15% floor	→	Sequential decline, or conversion below 15% annually
NVIDIA quarter-over-quarter data center revenue	**$75.2B for Q1 FY27, unchanged with no earnings event in-window — the Q2 print and guide land August 26, and this week's 8-K contingent-obligation disclosure raises a separate question about how the residual-value line is footnoted in the accompanying 10-Q exhibits**	$75.2B for Q1 FY27, unchanged with no earnings event in-window — the Q2 print and guide land August 26, and reports of reduced Rubin Ultra memory configurations make delivered content per GPU the variable to watch alongside unit volume	→	Two consecutive quarters of sequential decline

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Open-weight to closed-model capability gap on coding	**Still narrowed on paper and widened in practice. Ornith-1.5 arrived MIT-licensed at 397B/35B-active with vendor-run coding and terminal benchmarks against DeepSeek V4 Pro and GLM-5.3, but with no independent tracker ranking — a third consecutive week where a claimed open-weight advance did not clear the substitutability test**	Narrowed on paper and widened in practice. DeepSeek's 0813 build reports DeepSWE at 62.7 against the April build's 12.8, and GLM-5.3 claims the best open-source Terminal Bench 3.0 result — but neither model's weights were obtainable in the benchmarked form this week, so the gap that closed is between closed models and models you cannot yet download	→	Open weights within 2 Index points of the closed leader

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Sovereign AI program commitments	**~15 programs and ~$186B, unchanged — no new national program announced in-window. Pennsylvania's Executive Order 2026-05 is a state-level regulatory posture rather than a compute commitment; the White House NSSTS action on August 18 continues to circulate in secondary coverage without a primary text supporting the 'data centers removed from CET list' framing**	~15 programs and ~$186B, unchanged — no new national program announced in-window. The week's state-level activity was regulatory rather than capital: Colorado opened ADMT and chatbot rulemaking and Taiwan's digital ministry confirmed AI agents were used in attacks on government sites	→	Above 20 programs or $250B committed
PJM capacity auction clearing price	**$325.00 per MW-day for 2028/29, unchanged with no auction and no in-window filings. Pennsylvania EO 2026-05 does not change the clearing price but does add a permitting-side stage that compounds the queue length behind the same price**	$325.00 per MW-day for 2028/29, unchanged with no auction and no in-window filings	→	A second consecutive auction clearing at the cap

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Time from interconnection request to energization	**60-84 months, held flat with a lengthening bias. PA EO 2026-05 adds GRID compliance as a binding stage for DCs above 25 MW, removes qualifying DCs from Permit Fast Track and conditions the tax exemption on GRID — regulatory friction lengthening the effective queue for the largest US hyperscale market rather than shortening it**	60-84 months, unchanged — no new interconnection data in-window. Cisco's call commentary that campuses are fragmenting under power limits, driving scale-across optics demand, is indirect evidence the queue is shaping architecture rather than shortening	→	Below 48 months in two or more major queues

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Cost per task, frontier reasoning model	**Falling further at the top of the market on a scheduled cut: OpenAI reduced GPT-5.6 Sol pricing by more than 20% on August 21 for a dated three-month promotional window with the reversal date published at launch. The efficiency component of the lever is unchanged; the rate-card component now has three published increases within six months**	Falling sharply at the top of the market for the first time since April, on turn efficiency rather than rate cards: Artificial Analysis measured Grok 4.6 resolving long-horizon agentic tasks in ~53 turns and ~0.5B input tokens against ~103 turns and ~2.0B for Claude Opus 5, at $2 / $6 against Sol's $5 / $30 — but two published increases land within five months	↓	A frontier-tier reasoning model below $1 per million output tokens

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Custom silicon share of hyperscaler AI compute	**~34-37%, held flat with a directional-up note. Marvell's warrant to Google indexes ~7% of the silicon vendor's equity to Custom Products revenue on a $120B implied ceiling and no purchase obligation — a structural argument for share moving upward on Google's roadmap, but the vesting mechanism is forward and not yet a booked shipment**	~34-37%, unchanged — reports of Microsoft seeking 300,000-plus Maia 300 units from TSMC would move this materially if confirmed, but the company disputed the reported figures and capacity talks are not booked wafers	→	Above 45% share

THRESHOLD VALUES ARE THE POINTS WHERE THE READ FLIPS. CROSSINGS ARE FLAGGED IN THE ISSUE BODY WHEN THEY HAPPEN.

PREDICTIONS

What we expect next.

7 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

72%

BY OCTOBER 31, 2026

NVIDIA files exhibits with the 10-Q for the quarter ended July 26, 2026 that translate the SB Energy PORTS-Pike residual-value guaranty into a per-quarter contingent-obligation disclosure and identify the OpenAI affiliate as tenant, by October 31, 2026.

TRIGGER: NVIDIA's next Form 10-Q filed on EDGAR containing exhibit or footnote text that specifies the guaranty structure disclosed in the August 17 8-K, including the aggregate cap, ready-for-service conditions, and identification of the tenant.

PREDICTION 02 HARDWARE

34%

BY NOVEMBER 30, 2026

Marvell discloses the first attributable Custom Products revenue tranche under the Google warrant agreement in a filed earnings release or subsequent 10-Q, by November 30, 2026.

TRIGGER: A Marvell Form 10-Q or earnings release identifying a $500M-plus Custom Products revenue amount and confirming a warrant tranche has vested under the Google agreement disclosed August 19, 2026.

PREDICTION 03 SOFTWARE

43%

BY SEPTEMBER 15, 2026

Z.ai publishes GLM-5.3 weights to Hugging Face by September 15, 2026, closing the two-week window promised at the model's August 14 announcement.

TRIGGER: A Hugging Face repository under Z.ai containing model weights for GLM-5.3 distinct from GLM-5.2, dated on or before September 15, 2026.

PREDICTION 04 SOFTWARE

58%

BY OCTOBER 31, 2026

Anthropic or a named enterprise customer publishes an independently attributed Skills API production case study with a stated manual baseline and a measured post-Skills outcome (time, cost, or completion), by October 31, 2026.

TRIGGER: A published case study, press release or customer-authored blog post that identifies a named enterprise customer, cites a specific SOP or workflow, includes a numerical baseline for the pre-Skills-API state, and includes a measured post-implementation result attributable to Skills API rather than to a general workflow redesign.

PREDICTION 05 POWER

39%

BY DECEMBER 31, 2026

A second US state with major hyperscale exposure (Virginia, Texas, Georgia, Oregon, or Illinois) publishes a permitting or interconnection rule that adds a binding compliance stage above a stated MW threshold, comparable in kind to Pennsylvania Executive Order 2026-05, by December 31, 2026.

TRIGGER: An executive order, PUC ruling, or legislature-passed and governor-signed statute in one of the named states that binds an additional permitting, grid-compliance, or interconnection-review stage on data centers above a numeric MW threshold, published on or before December 31, 2026.

PREDICTION 06 HARDWARE

22%

BY JUNE 30, 2027

Anthropic publicly confirms a tape-out or first-silicon milestone on the in-house accelerator program being led by Amir Salek, by June 30, 2027.

TRIGGER: An Anthropic release, blog post, executive statement at a scheduled event, or major-media report attributed to Anthropic that identifies a tape-out, first-silicon, or physical prototype milestone on the in-house accelerator program.

PREDICTION 07 NETWORKING

47%

BY MARCH 31, 2027

A second major networking or accelerator vendor (Broadcom, Marvell, Cisco, or Arista) publicly discloses production shipments of co-packaged-optics Ethernet or scale-up switches to a named hyperscaler by March 31, 2027.

TRIGGER: A press release, earnings statement, or product-page update from Broadcom, Marvell, Cisco, or Arista identifying a co-packaged-optics switch product shipping in volume to a named hyperscaler customer, distinct from NVIDIA's Spectrum-6 CPO family.

TRACK RECORD

Every prediction ever made, scored against its own written trigger when the deadline passes; ambiguity resolves against us.

Predictions made	94	Resolved (hit / partial / miss)	29 (7 / 13 / 9)
Hit rate (partial = half)	47%	Brier score (0 = perfect)	0.159

Bold (<55%): none resolved yet · Core (55-80%): 29 resolved, 47% hit rate vs 66% mean confidence · High-conviction (>80%): none resolved yet · 6 overdue, unresolved (visible until adjudicated)

WATCHLIST

On the radar.

7 catalysts in the next 7–14 days that would change the read materially. Watching these tells us whether the thesis is strengthening or weakening.

AUG 24
WATCH 01

ChatGPT advertising rollout across 31 EU countries begins

The first live deployment of ChatGPT ads in the EU market — a direct test of whether consumer-tier AI monetization can be built on advertising rather than subscription in a jurisdiction with distinct data-privacy expectations. Watch complaint volume, national regulator response, and any adjacent product surfaces (Teens, Autopilot) that reprice on the same schedule.

AUG 26
WATCH 02

NVIDIA Q2 FY2027 earnings and 10-Q exhibits

The single largest scheduled datapoint of the quarter, and this time the residual-value guaranty disclosure changes what the print measures. Watch for 10-Q exhibits translating the SB Energy PORTS-Pike guaranty into per-quarter contingent-obligation disclosure, guidance commentary on Rubin Ultra memory configurations, and any change in framing on the ~$150-200B revenue-opportunity language attached to the transaction.

SEP 1
WATCH 03

OpenAI Daybreak hardware security key requirement takes effect

Physical-authentication gate on frontier model access lands. Whether Anthropic or Google adopt equivalent controls within the quarter determines whether physical authentication becomes a norm for dual-use agent access or stays an OpenAI-specific control. The Skills API GA raises the stakes because more capability is now behind the identity gate.

SEP 15
WATCH 04

GLM-5.3 weights promised window closes

Z.ai's August 14 announcement promised weights within two weeks of the release. September 15 is the outside edge of that promise plus a two-week buffer. If GLM-5.3 weights are still not on Hugging Face by that date, the pattern of announced-but-unavailable open-weight releases moves from a curiosity to a category-level trend, and the substitutability question this publication has been holding open should be reopened.

SEP 30
WATCH 05

DeepSeek V4-Pro-0813 weights deadline (Prediction p83)

The scoring window for the outstanding prediction from W33 closes. If the 0813 build weights are on Hugging Face by September 30, the served-versus-downloadable divergence remains a timing artifact; if not, DeepSeek's release cadence has structurally changed toward API-only endpoints and the substitutability lever this publication tracks needs to be revised.

NOV 21
WATCH 06

Dated Sol reversal exposes November business cases

The three-month scheduled step-up on the more-than-20% cut lands. Any November-priced agent business case must be modelled against the reversal, and watch whether a competing frontier vendor uses the moment to publish a comparable dated cut. Full architecture read lives in the Model Pulse.

Q4 2026
WATCH 07

OCP Open Silicon Photonics first specification submission

The stated target from the nineteen-company coalition launched at OCP APAC. Meeting it puts a multi-vendor CPO path on the table for 2027-28 refresh cycles; missing it effectively hands the next fabric refresh generation to single-vendor co-packaged optics by default. This is the fork in the road that the Big Story identifies as needing an explicit written decision by any operator refreshing fabric inside eighteen months.

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 18

Week 34 of 2026 · August 22, 2026

WEB

brianletort.ai/industry

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