THE AI STACK WEEKLY ISSUE 20 · WEEK 36 OF 2026

THE BOTTOM LINE

AI's scarce asset is now a governed path to service, not an announced model or megawatt

Astra's benchmark split and this week's infrastructure contracts point to the same operating reality: the asset being bought is a controlled path from intent to completed work, not a standalone model, chip, cable, or megawatt. ARC Prize's matched-effort comparison associates a 35.9-point runtime spread with the harness; Model Pulse carries the Standard and Provider Adapter scores. Infrastructure buyers likewise committed to dated future power, fiber, and deployment paths rather than operating supply.

Boards should require one schedule-and-control ledger across the stack: model plus harness and memory policy, power and fiber service dates, funding milestones, authority scopes, evidence custody, and rollback. Do not count options as contracted capacity, commitments as funded cash, previews as production services, or provider-adapter scores as model-only capability.

Vertiv's UtilityInnovation Group structure makes the underwriting visible: 44.2% of maximum consideration remains contingent on undisclosed post-close EBITDA targets. The arithmetic is house-original; the view that this transfers delivery risk to sellers is an interpretation because seller obligations and payout curves are undisclosed.

AUTHOR

Brian Letort
BrianLetort.AI

PUBLISHED

September 5, 2026
Issue 20

THIS WEEK IN 30 SECONDS

Executive summary.

- ARC Prize's matched-effort Astra comparison shows a 35.9-point runtime-associated spread; version the model, memory policy, context compression, and harness together.
- The week's infrastructure pattern is capital committed to dated future service, not operating capacity: a power-purchase agreement starts in 2028, a 150-terabit-per-second cable targets Q4 2028, and delayed-draw funding releases against milestones.
- Vertiv put 44.2% of UtilityInnovation Group's maximum $2.60B consideration behind EBITDA delivery; interpreting that as seller risk transfer is a house reading, not a disclosed fact.
- Broadcom's $16.7B AI-semiconductor quarter proves supplier monetization, not utilization or application returns; custom accelerators and networking remain combined.
- PJM's comment window closed while Texas staff proposed $50,000/MW security above 75 MW, keeping power obligations inside operating contracts.

BY THE NUMBERS

35.9 points
Astra matched-effort harness-associated spread

44.2%
Vertiv-UIG maximum price that is contingent

$16.7B
Broadcom Q3 AI-semiconductor revenue

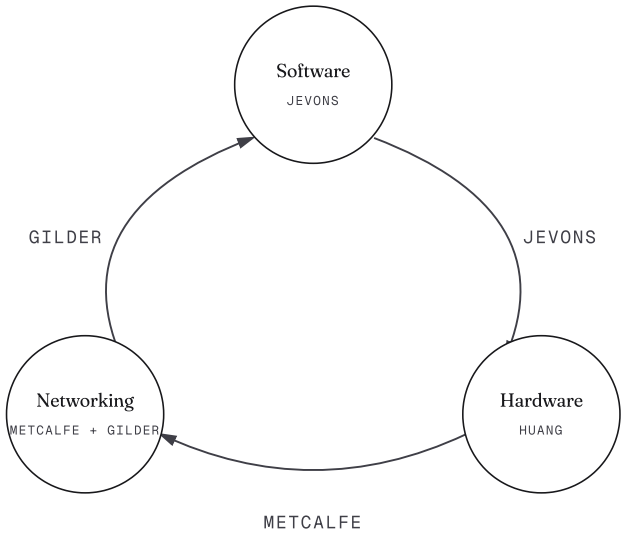
396 MW
Firm geothermal capacity under Google-Fervo power-purchase agreement

$3.05B
Nscale delayed-draw commitments

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.

SEP 3
EVENT 01

OpenAI begins phased GPT-6 Astra rollout with 1M context and $10/$50 per million-token Standard API pricing

OPENAI, ARC PRIZE, ARTIFICIAL ANALYSIS

SEP 1
EVENT 02

Claude Fable 5.1 becomes generally available across GitHub Copilot; enterprise administrators must enable it

GITHUB

SEP 3
EVENT 03

Gemini 3.8 Flash enters GitHub Copilot under introductory provider pricing through year-end

GITHUB

SEP 1
EVENT 04

NVIDIA releases Apache-2.0 Muse-Glimmer-30B NVFP4 (4-bit floating-point) checkpoint, reducing storage from 60 GB to 24.7 GB

NVIDIA ON HUGGING FACE

WHAT THIS MEANS

Architects should stop procuring model names as if the surrounding runtime were interchangeable. Astra's neutral-versus-provider harness spread, GitHub's administrator gates, and year-end Gemini pricing all make deployment policy part of measured capability and cost. See The Model Pulse for the full model and tree read.

LENS 2 OF 3

THE FLYWHEEL

Hardware.

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

AUG 31
EVENT 01

AMD, Cisco, and HUMAIN put an MI355X and 800G Ethernet cluster into production; live scale and utilization undisclosed

AMD, CISCO

AUG 31
EVENT 02

MediaTek joins NVIDIA NVLink Fusion for custom cloud accelerators (XPUs) without a named customer, tape-out, or shipment date

NVIDIA

SEP 1
EVENT 03

Cerebras contracts phased 50 MW, 80 MW, and 165 MW Finnish campus capacity under seven-year service orders

CEREBRAS ANNOUNCEMENT MIRROR, CONVERGE DIGEST

SEP 2
EVENT 04

Broadcom reports AI-semiconductor revenue at 56.4% of Q3 company revenue, without splitting accelerators from networking

BROADCOM EARNINGS RELEASE

WHAT THIS MEANS

Infrastructure buyers should separate production systems, contracted phases, and architecture announcements. The HUMAIN cluster is live but unquantified; Cerebras has only its first Finnish phase under construction; MediaTek-NVLink has no delivery milestone. Broadcom's filed unaudited quarterly disclosure is the hard signal, but its combined custom-accelerator-and-networking line cannot resolve which layer holds margin.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

AUG 31
EVENT 01

AWS-Azure managed private multicloud interconnect enters preview in four paired regions

AWS

SEP 1
EVENT 02

DE-CIX deploys dual-homed Azure ExpressRoute Metro across five global metros

DE-CIX

SEP 2
EVENT 03

Equinix announces pre-beta Fabric One for intent-driven routing, encryption, resilience, and failover

EQUINIX

SEP 3
EVENT 04

Firmus commits approximately $300M for up to 150 terabits per second on APX East for 25 years, targeting Q4 2028 service

FIRMUS AND SUBCO

WHAT THIS MEANS

Network architects should design delegated changes around simulation, approval, evidence, and rollback, not just easier invocation. AWS-Azure and Fabric One make intent-driven multicloud connectivity concrete, while the Firmus contract shows long-duration queue hedging. The durable-layer thesis remains strained until revenue growth proves pricing power.

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 1.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, XAI

BURN / REV
n/a

CAPITAL IN		Unknown
REVENUE OUT		Unknown

No reproducible W36 constituent ledger supports a category point estimate; no in-window financing or annual revenue disclosure establishes one.

Hyperscaler-Hosted

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

BURN / REV
n/a

CAPITAL IN		Unknown
REVENUE OUT		Unknown

No reproducible W36 constituent ledger supports a category point estimate; Google's Fervo agreement has no disclosed consideration.

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, FLUIDSTACK, IREN

BURN / REV
n/a

CAPITAL IN		Unknown
REVENUE OUT		Unknown

Nscale added up to $3.05B of delayed-draw commitments, but the unreproduced carried base prevents an honest category total or direction.

On-Prem / Hybrid

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

BURN / REV
n/a

CAPITAL IN		Unknown
REVENUE OUT		Unknown

No reproducible W36 constituent ledger supports a category point estimate; the atNorth deal changes ownership rather than measuring new category capital.

SIGNAL VS NOISE

What's real, what's noise.

5 claims that drove headlines this week, scored 1–5 on source quality and triangulation. 2 flagged as noise. The bar is at least one explicit noise call per issue.

4/5

CLAIM 01

Broadcom reported Q3 FY2026 revenue of $29.6B and AI-semiconductor revenue of $16.7B, 56.4% of company revenue.

SOURCES: BROADCOM FILED UNAUDITED QUARTERLY EARNINGS RELEASE, SEPTEMBER 2, 2026.

The reported line supports supplier monetization. It does not separate custom accelerators from networking or show customer utilization and application returns.

5/5

CLAIM 02

Google Energy signed a 396 MW, 15-year Fervo power-purchase agreement in four 99 MW tranches beginning in Q3 2028.

SOURCES: FERVO FILING MIRROR AND UTILITY DIVE, SEPTEMBER 1, 2026.

Count 396 MW as firm future capacity. Exclude conditional expansion until Google accepts it and a definitive agreement is executed; do not treat either amount as energized today.

4/5

CLAIM 03

Astra scored 99.9% on ARC-AGI-3 and therefore nearly saturates general intelligence.

SOURCES: OPENAI AND ARC PRIZE, SEPTEMBER 3, 2026.

The number is real but the conclusion is not. ARC Prize's Standard harness landed far below the headline adapter score; the best-observed adapter result used different reasoning effort, provider-private state, and context compression. Model Pulse carries the matched-effort scores. ARC says the closed-ended benchmark is not proof of general intelligence.

2/5 – NOISE

CLAIM 04

DeepSeek will deploy at least 160,000 Huawei Ascend accelerators at a roughly 1 GW Inner Mongolia inference site.

SOURCES: BLOOMBERG AND THE DECODER, SEPTEMBER 4, 2026; NO PRIMARY CONFIRMATION.

Potentially material domestic-inference substitution, but no purchase contract, delivered quantity, rack design, or schedule is public. Do not put it into capacity forecasts yet.

1/5 – NOISE

CLAIM 05

Fabric One has already made the network an autonomous production control plane for enterprise AI.

SOURCES: EQUINIX ANNOUNCEMENT AND LAUNCH COVERAGE, SEPTEMBER 2-4, 2026.

Architecture direction, not operating proof. The product is pre-beta with no price, service level, provisioning benchmark, rollback evidence, or production customer result.

HOUSE MEASUREMENT · FILING-DERIVED

Vertiv made 44.2% of UtilityInnovation Group's maximum $2.60 billion consideration performance-contingent, putting 79.3 cents of earnout behind every upfront dollar.

METHOD: House computation from Vertiv's September 2, 2026 acquisition disclosure. The primary release states: (1) approximately $1.45 billion cash consideration at closing; (2) up to $1.15 billion additional cash consideration tied to 12- and 24-month EBITDA targets; and (3) the upfront price is approximately 13 times expected 2027 EBITDA. Arithmetic, all in US dollars: maximum consideration = $1.45B + $1.15B = $2.60B; contingent share of maximum = $1.15B ÷ $2.60B = 44.23%; upfront share = $1.45B ÷ $2.60B = 55.77%; contingent dollars per upfront dollar = $1.15B ÷ $1.45B = $0.793; current implied expected 2027 EBITDA = $1.45B ÷ 13 = approximately $111.54M; maximum earnout ÷ current implied expected EBITDA = $1.15B ÷ $111.54M = 10.31x. The last ratio is a scale comparison, not the deal's full-price EBITDA multiple, because paying the earnout requires undisclosed higher EBITDA targets.

Performance-contingent share of maximum consideration	**44.2% ($1.15B of $2.60B)**	$1.15B maximum earnout ÷ ($1.45B upfront + $1.15B maximum earnout); this is the selected house measurement
Earnout dollars per upfront dollar	**$0.793 per $1.00 upfront**	$1.15B maximum earnout ÷ $1.45B cash at close; the structure places nearly eighty cents of additional price behind each upfront dollar
Upfront share of maximum consideration	**55.8% ($1.45B of $2.60B)**	$1.45B cash at close ÷ $2.60B maximum; closing cash is still the majority of potential consideration
Implied expected 2027 EBITDA	**Approximately $111.5M**	$1.45B upfront price ÷ the disclosed approximately 13x expected 2027 EBITDA multiple; approximate because both inputs are rounded
Maximum earnout relative to current implied EBITDA	**10.31x current implied expected 2027 EBITDA**	$1.15B ÷ $111.5M; this sizes the contingent pool only and is not an acquisition multiple because the earnout depends on higher undisclosed EBITDA targets

The disclosed structure makes a material share of potential consideration contingent rather than fixed. The house interpretation is that this transfers some delivery risk to sellers, but the disclosure does not quantify that transfer. Boards should compare the contingent share of maximum consideration and the earnout-to-upfront ratio before comparing headline transaction values.

Caveats: Seller risk transfer is an interpretation, not a disclosed fact: Vertiv does not publish seller obligations, the EBITDA thresholds, their weighting, payout curve, revenue concentration, or baseline. The approximate inputs limit precision, the $2.60B maximum is not the expected purchase price, and the 10.31x scale comparison is not a transaction multiple. Closing remains expected in Q4 2026.

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

Enterprise AI is re-layering around a delegated-control boundary: policy, evidence custody, approval, and rollback are becoming separately owned infrastructure around model execution.

ABDUCTIVE · 76% CONFIDENCE

Astra's launch joined persistent model state with action-level classifiers and automatic review, making runtime control part of the released capability rather than an external compliance attachment.
Anthropic's Enterprise Frontier Safeguards separates detector operation from data custody: monitoring runs across sessions, while logs, keys, flagged-event review, and human access stay in the customer's cloud account.
GitHub now lets a Copilot approval satisfy branch-protection rules, but scopes the authority by repository and path and dismisses the approval after a new commit, converting AI judgment into revocable policy state.
Equinix plans to translate portal, application programming interface, agent, and natural-language intent into routing, encryption, resilience, and failover, moving delegated control from software changes into network state.

STEEL-MAN: Three of the four controls are new, previewed, or rolling out later: Anthropic's Enterprise Frontier Safeguards starts in phases, Copilot approval is public preview, and Fabric One is not yet in beta. These could remain vendor-specific features rather than a durable independent layer. The strongest part of the claim is narrower: across code, model traffic, and networking, vendors are converging on explicit scopes, retained evidence, and revocation. Falsified if production releases fold those controls back into opaque model behavior or customers cannot export and audit the resulting policy state by Q2 2027.

An accumulating procurement pattern is visible across AI infrastructure: buyers are committing capital to dated future-service paths in power, fiber, and deployment before those assets operate.

INDUCTIVE · 74% CONFIDENCE

Google's 15-year Fervo power-purchase agreement contracts 396 MW in four tranches beginning in Q3 2028 and includes approximately 600 MW of conditional expansion, reserving a future delivery path rather than consuming present generation.
Firmus committed approximately $300 million for up to 150 terabits per second over 25 years on an Australia-US cable targeted for service in Q4 2028, paying for route position before the system is built.
Nscale's $3.05 billion facilities are delayed-draw commitments, meaning funds become available against future milestones rather than all at closing, for two deployments spanning compute, network, cooling, and site work.
Vertiv and Flex agreed to acquire power-control and conversion capabilities that influence whether a permitted site can move from interconnect application to energized rack.

STEEL-MAN: Power agreements, subsea capacity, delayed-draw loans, and acquisitions are different instruments with different risks; grouping them can obscure ordinary long-lead infrastructure planning. None proves that queue positions are independently tradable or earn excess returns. The inference is a continuing procurement pattern, not a new asset class or measured directional shift. Falsified if comparable power, fiber, and compute capacity becomes repeatedly available on short notice without precommitment or scarcity-linked terms by mid-2027.

Agent procurement complexity is accumulating in bundles, credits, workflow entitlements, retention routes, and runtime choices, making utilization forecasting a core buying competency.

ABDUCTIVE · 67% CONFIDENCE

Salesforce embedded Agentforce, collaboration, analytics, security, and support into three seat editions while coupling access rights to different consumption pools.

Astra launched as a premium application-programming-interface tier while Gemini 3.8 Flash entered Copilot under temporary introductory provider pricing, widening the model-price range inside comparable agent surfaces.

GitHub administrators can choose whether Fable 5.1 is available across nine surfaces, showing that enterprise entitlement and retention policy can matter as much as public model availability.

AWS's migration pattern deliberately separates tools, loop, runtime, memory, and observability so model choice can change without rebuilding every operating layer.

> **STEEL-MAN:** Salesforce is one vendor, its credits are not directly convertible to model tokens, and introductory Gemini pricing may expire without changing Copilot seat economics. Customers can still capture value primarily from superior capability on difficult tasks. This evidence supports procurement complexity, not margin migration. Falsified if major agent suites converge on transparent per-outcome prices and publish stable task-level usage denominators by Q2 2027.

THESIS TEST

H1 · STRAINED **The cycle is accelerating, not slowing.** W36 shows rapid software packaging: Astra arrived with persistent memory, asynchronous clarification, and action monitoring, while Anthropic, GitHub, and AWS announced adjacent distribution and control changes in the same week. It does not measure a shorter prior-to-current doubling interval across software, hardware, and networking, and therefore cannot support the hypothesis's flywheel-cadence test. **AGAINST IT:** W36 had no production-volume next-generation silicon release: MediaTek-NVLink is a collaboration, Tensordyne is modeled pre-silicon, and Cerebras's 165 MW is a phased campus target with only the first phase under construction. The acceleration is clearest in software distribution and control-layer iteration, not uniformly across the hardware lens; two quarters of stalled production would still trigger the framework's refutation test.

H2 · STRAINED **Capital is concentrated, returns are diffuse.** Large infrastructure commitments show capital concentration, including Nscale's delayed-draw facilities and the atNorth, UtilityInnovation Group, and EPC Power transactions. Salesforce packaging, Wonderful's valuation, and Atira's reported deployments do not measure diffuse returns or margins. The week supports the concentration half of the hypothesis but leaves the return-distribution half untested. **AGAINST IT:** Broadcom captured $16.7 billion of quarterly AI-semiconductor revenue, 56.4% of company revenue, showing returns can concentrate with an upstream custom-silicon and networking supplier rather than diffuse downstream. Wonderful and Atira disclose funding, customer counts, and selected outcomes but not audited margins or retention, so application-layer return diffusion remains more plausible than measured.

H3 · STRAINED **Networking is the durable layer.** W36 strengthened networking's strategic role: AWS-Azure private multicloud interconnect entered preview, DE-CIX expanded dual-homed ExpressRoute Metro, Equinix announced intent-driven Fabric One, and Firmus committed approximately $300 million to 25 years of APX East capacity. These events show orchestration, route diversity, and long-duration capacity value. They do not satisfy the framework's economic test because none discloses cross-connect or fabric revenue growth relative to compute revenue. **AGAINST IT:** Fabric One is pre-beta, the AWS-Azure preview currently reports up to 1 Gbps, and APX East targets Q4 2028. Meanwhile Broadcom's AI-semiconductor line combines networking with custom accelerators, preventing a clean margin comparison. The architectural evidence is favorable, but the week's public data cannot show networking pricing power holding longer than compute pricing power.

H4 · UNTESTED **Open weights pull the floor up.** NVIDIA's Apache-2.0 low-precision Muse-Glimmer checkpoint reduces deployment footprint while preserving published capabilities. But it is an optimization of an existing W33 model, not a new open-weight frontier release, and W36's dominant application and agent launches were closed or metered. One optimized checkpoint cannot test whether open weights narrowed the closed-model capability gap or rerouted meaningful new enterprise compute demand this week. **AGAINST IT:** Astra's closed launch produced the week's strongest benchmark movement, and Salesforce, GitHub, and Anthropic all reinforced governed hosted access. The open floor may still be rising over a multiweek horizon, especially after W35's model wave, but W36 adds no comparable fresh benchmark set and should not be scored as support from absence of refutation.

H5 · STRAINED **Power is the binding constraint for the next 24 months.** Capital and policy converged on energization: Vertiv and Flex agreed to power-control and conversion acquisitions; Google contracted 396 MW of geothermal with delivery beginning in 2028; Texas staff retained $50,000 per MW security for large-load interconnection; PJM's comment window closed on a bring-capacity-or-curtail proposal; and the Department of Energy temporarily authorized customer standby generation during emergency conditions. This supports power as a major site-timing constraint, but does not establish it as the unique or dominant constraint across the industry. **AGAINST IT:** The Nscale facilities fund GPUs, networking, storage, cooling, and site work, not power alone, while Cerebras's Finnish build and the HUMAIN production cluster show projects can still secure viable sites. The evidence supports power as a site-timing constraint, not the only industry constraint; accelerator supply, financing, permits, and customer readiness still determine how fast announced capacity becomes revenue.

PATTERN WATCH

Agent deployment is shifting from capability access to explicit containment, delegated authority, and evidence custody. (3 WEEKS OBSERVED)

Next week: Before 2026-12-31, at least one major enterprise agent platform will publish an exportable authorization and evidence schema covering tool scope, approvals, retained traces, and revocation. Falsified if the next two major agent releases provide capability GA without new authority or audit controls.

Power constraints are migrating from site-selection assumptions into enforceable contracts, financial security, curtailment priority, and emergency dispatch. (4 WEEKS OBSERVED)

Next week: By 2026-10-31, either FERC will act on PJM ER26-3515 or another US state or grid operator will publish a large-load rule that assigns measurable security, curtailment, or capacity obligations. Falsified if PJM's proposal is withdrawn and no comparable rule advances during that period.

AI infrastructure buyers are reserving future schedule position through guarantees, capacity contracts, and milestone-timed capital rather than waiting for operating supply. (3 WEEKS OBSERVED)

Next week: Before year-end 2026, another AI infrastructure contract above $500 million will disclose a future service date plus an option, earnout, guarantee, or delayed-draw mechanism that allocates schedule risk. Falsified if the next three comparable commitments are fully funded against currently operating capacity without milestone conditions.

SECOND-ORDER EFFECTS

Enterprise evaluations will have to version the model, harness, memory policy, tool surface, and retry budget as one tested system. Model-only scorecards will become procurement-incomplete, and portability reviews will ask which gains disappear when provider-private state is removed. ARC Prize, not this publication, established the core harness finding. Trigger: ARC Prize's matched maximum-reasoning comparison measured Astra at 62.7% on the Standard harness and 98.6% on the Provider Adapter, a 35.9-point spread associated with runtime state and context compression. Horizon: Q4 2026. Who moves: Enterprise AI architecture teams, model-evaluation vendors, frontier API providers, procurement and model-risk functions.

Security and audit teams will define a new separation-of-duties rule: an AI may execute or approve, but the same provider-controlled evidence path cannot be the sole basis for both. Expect independent trace retention, human escalation thresholds, and policy-engine attestations in regulated software delivery. Trigger: GitHub allowed Copilot approvals to count toward protected-branch requirements while Anthropic placed long-window monitoring evidence in customer-controlled accounts. Horizon: Two procurement cycles. Who moves: Software engineering leaders, internal audit, regulated-industry security teams, GitHub administrators, agent-platform vendors.

Data-center design authority will move earlier toward firms that can model utility, onsite generation, storage, power conversion, and rack loads as one permitted system. Cooling and compute vendors without an upstream power-control partner will face acquisition, partnership, or specification risk before equipment selection. Trigger: Vertiv and Flex committed up to $7.0 billion to grid-to-chip controls and power conversion while large-load rules attached security and curtailment obligations to interconnection. Horizon: 2027 planning cycle. Who moves: Data-center developers, electrical equipment vendors, utilities, engineering firms, hyperscalers, neoclouds, and infrastructure investors.

STRATEGIC OUTLOOK

W36 favors control-plane and schedule discipline over headline capacity. Treat Astra as a model-plus-runtime release: reproduce its benchmark advantage under the memory, context-compression, tool, and audit constraints you can govern. For agent procurement, forecast entitlements and credits against completed workflow outcomes rather than seats or tokens, and require exportable authorization and evidence records before allowing AI approvals to satisfy human control gates. For infrastructure, separate operating megawatts from phased, optioned, and contracted future capacity; track dated service paths, then assign delivery risk through earnouts, delayed draws, guarantees, or tranche milestones. Power is a major site-timing constraint, not a proven unique bottleneck. Networking's role is strengthening architecturally, yet the thesis needs revenue evidence before claiming durable pricing power.

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 OpenAI and ARC Prize, Sep 3, 2026: GPT-6 Astra's 99.9% ARC-AGI-3 result is near-saturation evidence of broadly human-level agentic intelligence.

Our read: The 99.9% adapter result is best-observed and descriptive, but it uses a different reasoning effort from the matched Standard-harness maximum. Model Pulse carries the matched-effort Standard versus Provider Adapter scores. Astra plus its runtime nearly saturates a closed-ended benchmark; model-only saturation and general intelligence do not follow.

OPEN Broadcom earnings release; RCR Tech: Broadcom's 56.4% AI revenue share proves custom-silicon returns have already concentrated upstream.

Our read: The reported revenue is material, but it comes from a filed unaudited quarterly disclosure, combines custom accelerators and networking, and omits customer utilization, component margins, and downstream returns. It supports shipment monetization, not where durable return pools settle.

DIFFER Vertiv and Flex announcements, Sep 2-3, 2026: Vertiv and Flex spent $7.0B to buy one integrated grid-to-chip control plane.

Our read: The headline sum collapses different assets. Vertiv buys site architecture and microgrid orchestration with a material contingent share of maximum price (see house measurement); Flex buys conversion hardware and grid-forming controls at 5.5x forecast 2026 revenue. Both are pending acquisitions, not operating capacity.

EXTEND Equinix and SiliconANGLE, Sep 2-4, 2026: Equinix Fabric One has turned the network into the autonomous control plane for distributed AI.

Our read: Intent-driven connectivity is the direction, but Fabric One is pre-beta. The control-plane product is not natural-language provisioning alone; it is policy simulation, scoped approval, retained intent history, rollback, and production service evidence that have not yet published.

EARLY WARNING PANEL

The levers we monitor.

10 metrics, current vs prior period. **0 rising**, **0 falling**, 10 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	**Unknown — W36 has no reproducible cash-and-burn input table**	Prior estimate withheld for the same reason	→	Below 18 months for any disclosed-burn lab
Hyperscaler AI capex to disclosed AI revenue ratio	**Unknown — no reproducible top-four capex and AI-revenue input range**	Prior point estimate withheld because AI revenue is not segment-audited	→	Above 6x sustained for two consecutive quarters
CoreWeave contracted revenue backlog	**$104.2B as of June 30, unchanged — no CoreWeave filing in-window**	$104.2B as of June 30, unchanged — no CoreWeave filing in-window; Vera Rubin production deep-dive reinforces operational moat ahead of Q3 print	→	Sequential decline, or conversion below 15% annually
NVIDIA quarter-over-quarter data center revenue	**$89.0B for Q2 FY27, held pending Q3; Broadcom separately reported AI-semiconductor revenue (see byTheNumbers)**	$89.0B for Q2 FY27 (+117% YoY), up sequentially from $75.2B Q1; Vera Rubin production began in August; ~20% datacenter mix guided for Q3 (vendor-stated)	→	Two consecutive quarters of sequential decline

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Open-weight to closed-model capability gap on coding	**Untested this week — Muse-Glimmer NVFP4 cuts checkpoint storage (see software events), but no new independent open-frontier benchmark set**	Narrowed on vendor-reported rows but substitutability improved: IBM Granite 4.2 30B Apache 2.0 with vendor-reported 57% SWE-Bench Verified; GLM-5.3-Flash MIT weights with vendor-reported 84.3% Terminal-Bench 2.1 — independent tracker confirmation still pending	→	Open weights within 2 Index points of the closed leader
Sovereign AI program commitments	**Unknown — W36 has no reproducible program ledger**	Prior aggregate withheld for the same reason	→	Above 20 programs or $250B committed
PJM capacity auction clearing price	**$325.00 per MW-day for 2028/29, unchanged — ER26-3515 comment deadline passed Sep 3**	$325.00 per MW-day for 2028/29, unchanged	→	A second consecutive auction clearing at the cap
Time from interconnection request to energization	**Unknown — no reproducible multi-queue duration table in W36**	Prior range withheld for the same reason	→	Below 48 months in two or more major queues

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Cost per task, frontier reasoning model	**No commensurate W36 update; ARC Prize suite totals are not a market floor or median**	W35 named-basis floor: GLM-5.3-Flash vendor-reported $0.045 per task on Artificial Analysis Index 57; no same-basis median available	→	A frontier-tier reasoning model below $1 per million output tokens
Custom silicon share of hyperscaler AI compute	**Unknown — Broadcom disclosed combined custom-accelerator and networking revenue but no compute-share or component split**	Unknown — Hot Chips disclosed inference ASIC roadmaps (Google TPU 8i, Jalapeño, Maia 200, MTIA 400) but no in-window hyperscaler compute-mix filing supports a booked share estimate; disclosure cadence ≠ shipped mix	→	Above 45% share with audited hyperscaler mix disclosure

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 SOFTWARE

32%

BY DECEMBER 15, 2026

ARC Prize or another provider-neutral evaluator publishes an Astra run without provider-private reasoning state that closes at least half of the 35.9-point matched-effort Standard-to-Provider-Adapter gap by December 15, 2026.

TRIGGER: Public ARC-AGI-3 result with a reproducible visible-memory and context-compression configuration, no provider-private reasoning state, and Astra at 80.7% or higher.

PREDICTION 02 POWER

41%

BY JUNE 30, 2027

Google accepts at least 500 MW of Fervo's conditional expansion and the parties execute a definitive agreement by June 30, 2027.

TRIGGER: Fervo filing or Google announcement stating that at least 500 MW of the approximately 600 MW expansion option has become firm under a definitive agreement.

PREDICTION 03 CAPITAL

68%

BY DECEMBER 31, 2026

A second AI infrastructure contract above $500M discloses both a future service date and an option, delayed-draw, earnout, or guarantee allocating schedule risk by December 31, 2026.

TRIGGER: Primary filing or announcement with transaction value above $500M, named service date, and explicit contingent or milestone mechanism.

PREDICTION 04 NETWORKING

57%

BY DECEMBER 31, 2026

Equinix publishes Fabric One beta documentation that exposes approval, rollback, or auditable intent-history controls before December 31, 2026.

TRIGGER: Public Fabric One beta documentation naming at least one of policy simulation, approval workflow, rollback, or exportable intent history.

PREDICTION 05 HARDWARE

24%

BY DECEMBER 31, 2026

Broadcom discloses separate quarterly revenue figures for custom AI accelerators and AI networking by December 31, 2026.

TRIGGER: Hit only if a Broadcom earnings release, 10-Q, or call transcript reports distinct dollar revenue for both custom AI accelerators and AI networking; a combined AI-semiconductor line is a miss.

TRACK RECORD

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

Predictions made	**104**	Resolved (hit / partial / miss)	**57 (23 / 15 / 19)**
Hit rate (partial = half)	**54%**	Brier score (0 = perfect)	**0.200**

Bold (<55%): 1 resolved, 100% hit rate vs 43% mean confidence · Core (55-80%): 55 resolved, 52% hit rate vs 66% mean confidence · High-conviction (>80%): 1 resolved, 100% hit rate vs 84% mean confidence

WATCHLIST

On the radar.

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

SEP 8

WATCH 01

DOE emergency-order expiry

Order 202-26-43 expires after temporarily authorizing customer standby generation as a last-resort grid resource.

SEP 9

WATCH 02

GLM-5.3-Flash promotion ends

Reprice W35 open-weight agent pilots on steady-state list rates rather than the launch discount.

SEP 11

WATCH 03

Texas large-load rule vote

Tests whether the 75 MW threshold and $50,000/MW security move from staff recommendation into final policy.

LATER IN 2026

WATCH 04

Fabric One beta documentation

Look for policy simulation, approval, rollback, service levels, and exportable intent evidence before production use.

Q4 2026

WATCH 05

Vertiv-UIG and Flex-EPC closings

Regulatory clearance and disclosed integration terms will test whether grid-to-chip consolidation converts from agreement to operating capability.

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 20

Week 36 of 2026 · September 5, 2026

WEB

brianletort.ai/industry

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