THE AI STACK WEEKLY ISSUE 22 · WEEK 38 OF 2026

THE BOTTOM LINE

The agent bill moved from tokens to the whole system, just as power and financing became the binding constraints

The week's original connection is not that voice models improved; it is that the conversational layer, reasoning layer, tools, power envelope, and financing structure can now be priced separately. Google released Gemini 3.8 Live at $0.005 per minute of audio input and $0.018 per minute of output while allowing tool calls and deeper reasoning to continue behind an uninterrupted conversation. NVIDIA then reported a measured Blackwell deployment where factory-level power management raised cluster token throughput from about four million to five million tokens per second inside the same power budget. At the same time, CoreWeave launched a $3 billion convertible offering and loans tied to an Oracle-leased AI campus were reported at 89 to 91 cents on the dollar. Buyers should therefore model cost per completed workflow across model, harness, tools, recovery, and infrastructure rather than treating the token rate as the cost of the agent.

AUTHOR

Brian Letort
BrianLetort.AI

PUBLISHED

September 19, 2026
Issue 22

THIS WEEK IN 30 SECONDS

Executive summary.

- Voice agents became execution systems: Google now keeps dialogue running while tools and deeper reasoning work in parallel.
- Infrastructure efficiency moved from component benchmarks to tokens per megawatt, with one measured Blackwell deployment lifting throughput 24% inside the same power budget.
- Capital markets supplied the counter-signal: CoreWeave sought $3B of convertible debt while debt tied to an Oracle-leased campus traded below par.
- The operating decision is to price agent work end to end, including the conversational layer, reasoning model, tools, recovery, and power.

BY THE NUMBERS

$0.005/min
Gemini 3.8 Live audio input

$0.018/min
Gemini 3.8 Live audio output

+24%
Measured cluster token throughput

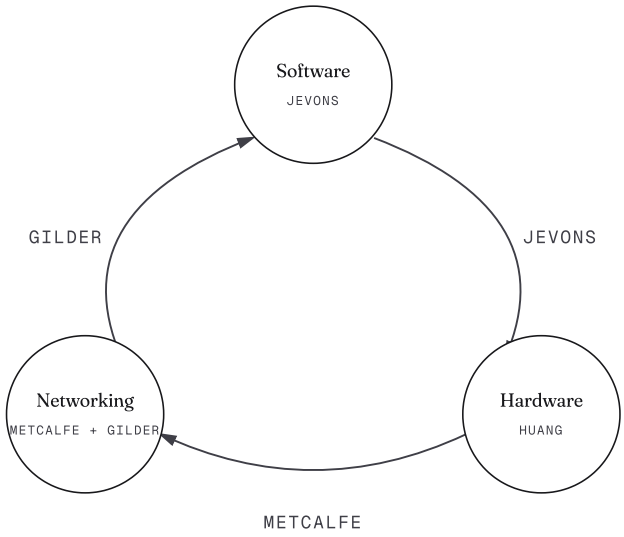
$3.0B
CoreWeave convertible offering

89–91¢
Reported price of Project Jupiter loans

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 14
EVENT 01

GitHub adds efficiency, balance, and intelligence tiers to automatic model routing

GITHUB CHANGELOG

SEP 15
EVENT 02

Google releases Gemini 3.8 Live and Live Extended Thinking through the Live API

GOOGLE

SEP 15
EVENT 03

Salesforce introduces Koa, a CRM reasoning model post-trained from NVIDIA Nemotron 3 Super

SALESFORCE, NVIDIA

SEP 18
EVENT 04

GitHub agents gain local Dev Containers and direct pull-request creation from agent sessions

GITHUB CHANGELOG

WHAT THIS MEANS

Architects should specify routing policy and the complete metering chain, not a favorite model. Automatic model selection now exposes an explicit cost-quality-latency policy, while live voice can keep a conversation open as tools and reasoning continue; see The Model Pulse for the model and pricing detail.

LENS 2 OF 3

THE FLYWHEEL

Hardware.

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

SEP 15
EVENT 01

NVIDIA reports Vera Rubin in production and publishes agentic throughput-per-megawatt claims

NVIDIA

SEP 15
EVENT 02

Lambda measures 24% more cluster token throughput under the same Blackwell power budget

NVIDIA, LAMBDA

SEP 17
EVENT 03

AMD argues for movable workloads across CPUs, accelerators, edge devices, and open interconnects

AMD

WHAT THIS MEANS

Operators should make tokens per megawatt and performance under a site power cap acceptance metrics for new infrastructure. The measured Lambda result is stronger evidence than NVIDIA's forward Rubin multipliers, but both point to system-level power management becoming as important as accelerator peak throughput.

LENS 3 OF 3

THE FLYWHEEL

Networking.

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

SEP 15
EVENT 01

NVIDIA frames NVLink, Spectrum-X, ConnectX, BlueField, storage, and cooling as one power-managed factory

NVIDIA

SEP 15
EVENT 02

Cisco highlights cross-domain network, security, and observability context for agentic operations

CISCO

SEP 17
EVENT 03

AMD makes open interconnects and workload portability central to its AI infrastructure posture

AMD

WHAT THIS MEANS

Network buyers should test the fabric as part of a power-capped system rather than procure switching on headline bandwidth alone. The week's common design point is cross-layer control: compute, memory, fabric, cooling, and observability must expose enough telemetry to optimize one completed-work metric.

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 18.0 $B; On-Prem revenue is indirect.

Frontier Labs

OPENAI, ANTHROPIC, GOOGLE DEEPMIND, DEEPSEEK

BURN / REV
Unknown

CAPITAL IN 

No disclosed primary financing in the observation window

REVENUE OUT

Undisclosed

Flat on disclosed financing; Anthropic's possible model release and IPO timing remain reported deliberations, not transactions

Hyperscaler-Hosted

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

BURN / REV
Unknown

CAPITAL IN 

No new category-wide financing disclosed

REVENUE OUT

No new AI-segment revenue disclosure

Down in credit quality for one project: $18B of Oracle-leased campus loans were reported at 89-91 cents

Neoclouds

COREWEAVE, NSCALE, CRUSOE, LAMBDA, IREN, ZANKORE, NEXTDC

BURN / REV
Unknown; financing need is disclosed, operating cash conversion is not

CAPITAL IN 

$3.0B convertible notes, with a $500M buyer option

REVENUE OUT

No new revenue disclosure

Up sharply as CoreWeave returned to debt and opened an at-the-market equity program

On-Prem / Hybrid

ENTERPRISE GPU CLUSTERS, SOVEREIGN AND NATIONAL PROGRAMS, OPEN-WEIGHT AND ON-DEVICE DEPLOYMENT

BURN / REV

Not applicable

No comparable disclosed program in the window

CAPITAL IN

REVENUE OUT

Indirect

Flat on disclosed capital; up in architectural emphasis as vendors stressed portability and private deployment

SIGNAL VS NOISE

What's real, what's noise.

4 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

Lambda raised cluster token throughput 24% inside the same power budget by running 19 nodes where 16 full-power nodes were normally allocated.

SOURCES: NVIDIA REPORT OF LAMBDA MEASUREMENT

This is the strongest operating evidence in the window because it names the before and after configuration. Buyers should request the same power-capped test on their workload.

3/5

CLAIM 02

Google's Gemini 3.8 Live can continue a conversation while tools and deeper reasoning run in the background.

SOURCES: GOOGLE LAUNCH POSTS

The capability is shipped through the Live API, but outcome quality and interruption behavior still need workload-specific testing.

2/5 — NOISE

CLAIM 03

Vera Rubin delivers up to 30 times more agentic throughput per megawatt than GB300 NVL72.

SOURCES: NVIDIA VENDOR BENCHMARK USING AGENTX

The number is vendor-reported and workload-specific. Treat it as a test target, not a capacity-plan input, until independently reproduced.

2/5 — NOISE

CLAIM 04

Koa produces three times fewer errors than leading models on CRM actions.

SOURCES: SALESFORCE CRM BENCHMARK

The benchmark, dataset, and comparison set are vendor-controlled. Pilot availability is real; the multiple is not yet procurement-grade evidence.

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

AI procurement is moving from model selection to system-yield contracting.

ABDUCTIVE · 76% CONFIDENCE

Google split live-agent audio into input and output meters while background reasoning and tools continue.
GitHub exposed cost, quality, and latency as selectable routing policy.
NVIDIA and Lambda measured completed token throughput under a fixed power budget.

STEEL-MAN: These are vendor-defined meters and one customer measurement, not a standardized contracting unit. The claim is bounded to procurement direction, not current market practice.

Infrastructure credit risk is becoming an operational architecture input rather than a finance-only concern.

INDUCTIVE · 69% CONFIDENCE

CoreWeave sought $3 billion of convertible debt plus an equity-sale facility to fund operations.
Loans tied to an Oracle-leased campus were reported below par amid project-specific concerns.
Hardware vendors simultaneously shifted their headline metric to useful work inside a fixed megawatt.

STEEL-MAN: Both financings may be issuer- or project-specific and demand remains strong. The inference is about diligence scope, not a sector-wide credit event.

THESIS TEST

H1 · SUPPORTED **Software demand pulls hardware and network investment forward.** Parallel voice reasoning increases the number of simultaneous model, tool, and memory paths, while infrastructure vendors answer with system-level throughput optimization. **AGAINST IT:** No buyer disclosed incremental capacity ordered specifically for Gemini Live.

H2 · SUPPORTED **Hardware efficiency expands economically viable AI demand.** Lambda's measured 24% throughput gain inside the same power budget directly lowers the infrastructure consumed per token. **AGAINST IT:** One Blackwell deployment does not establish portability across models or facilities.

H3 · SUPPORTED **Networking becomes a first-order limiter as AI systems scale.** The week's vendor architectures treat fabric, memory, storage, and cooling as one controlled system rather than separable components. **AGAINST IT:** No in-window outage or benchmark isolated networking as the binding bottleneck.

H4 · STRAINED **Capital follows visible utilization and contracted demand.** CoreWeave accessed large financing, but below-par project debt shows that contracted demand no longer neutralizes construction, environmental, and concentration risk.

H5 · UNTESTED **Open ecosystems gain when switching costs become material.** AMD emphasized open interconnects, but the window supplied no comparable buyer migration or audited share shift.

PATTERN WATCH

Agent cost decomposes into independently managed system meters (2 WEEKS OBSERVED)

Next week: A major provider will publish a worked multi-meter agent cost example before year end.

Power-capped throughput replaces peak component performance (2 WEEKS OBSERVED)

Next week: At least one infrastructure launch next week will headline useful work per watt or megawatt rather than peak FLOPS.

SECOND-ORDER EFFECTS

FinOps must attribute one user interaction across several concurrent meters and failure domains. Trigger: Voice agents continue talking while background tools and reasoning execute. Horizon: Next two quarters. Who moves: CIOs, application architects, and AI platform teams.

Workload portability and staged capacity commitments become credit-risk mitigations, not only architecture preferences. Trigger: AI campus debt trades below par while new capacity still seeks billions in financing. Horizon: Next 12 months. Who moves: Cloud buyers, lenders, neoclouds, and infrastructure vendors.

STRATEGIC OUTLOOK

Over the next 12 months, the winning capital posture is optionality with measurement: buy capacity in stages, require power-capped workload tests, preserve model and fabric portability where practical, and price agents as complete workflows rather than token endpoints. The risk is no longer only overpaying for a model that becomes obsolete. It is locking a workflow to a conversational meter, a reasoning meter, a tool chain, and an infrastructure financing structure that can each move independently.

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 — no lab disclosed cash, burn, or financing in the window**	Unknown — no constituent disclosed cash, burn, or financing in the window	→	Below 18 months for any disclosed-burn lab
Hyperscaler AI capex to disclosed AI revenue ratio	**Unknown — no hyperscaler disclosed AI-segment revenue against capital expenditure**	Unknown as a top-four ratio — one constituent disclosed $28.5B of quarterly capital expenditure and roughly $5B of negative free cash flow, with no AI-segment revenue line	→	Above 6x sustained for two consecutive quarters
CoreWeave contracted revenue backlog	**$104.2B as of June 30; unchanged pending the next filing**	$104.2B as of June 30, unchanged — no CoreWeave filing landed in the window	→	Sequential decline, or conversion below 15% annually

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
NVIDIA quarter-over-quarter data center revenue	**$89.0B for Q2 FY27; unchanged pending the next NVIDIA print**	$89.0B for Q2 FY27, unchanged — no NVIDIA print in the window	→	Two consecutive quarters of sequential decline
Open-weight to closed-model capability gap on coding	**Not comparable — Gemini Live launched without an open-weight peer on the same voice benchmark**	Not measurable this week — the reference index changed basis twice in four days, breaking comparability with the prior reading	→	Open weights within 2 Index points of the closed leader
Sovereign AI program commitments	**Unknown — no qualifying government-funded national compute commitment in the window**	Unknown — no government-funded national compute programme in the window meets the ledger method	→	Above 20 programs or $250B committed
PJM capacity auction clearing price	**$325.00 per MW-day for 2028/29; unchanged**	$325.00 per MW-day for 2028/29, unchanged — no auction occurred in the window	→	An auction clearing below the cap, or a FERC-approved increase in the cap itself

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Time from interconnection request to energization	**Unknown — no comparable queue-duration update was published**	Unknown as a queue duration — the window's only hard figure is 850 MW delivered within one quarter by a single operator, which measures delivery rather than energisation or queue time	→	Below 48 months in two or more major queues
Cost per task, frontier reasoning model	**Voice layer now $0.005/min input plus $0.018/min output before reasoning and tool charges**	$3.26 is the floor of the two leading models and $7.63 the other, per index task on the new v4.3 basis; no frontier-tier median is computable because the basis changed twice inside the window, and neither figure is commensurate with the prior reading	→	A frontier-tier reasoning model below $1 per million output tokens

METRIC	CURRENT	PRIOR	DIR	THRESHOLD / NOTE
Custom silicon share of hyperscaler AI compute	**Unknown — no audited hyperscaler compute-mix disclosure**	Unknown — a multi-generation custom inference agreement was signed but discloses a warrant-vesting ceiling rather than units, share, or a service date	→	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

At least one major voice-agent provider publishes an end-to-end worked cost example that includes voice, reasoning, and tool execution by December 31, 2026.

43%

BY DECEMBER 31, 2026

TRIGGER: Hit only if a first-party pricing or documentation page shows all three cost components in one worked workflow.

PREDICTION 02 HARDWARE

A major server, accelerator, or cloud vendor publishes a customer procurement template using tokens per megawatt as an acceptance metric by December 31, 2026.

31%

BY DECEMBER 31, 2026

TRIGGER: Hit only with a public first-party RFP, reference architecture, or acceptance guide naming tokens per megawatt.

PREDICTION 03 CAPITAL

CoreWeave completes at least $3 billion of the convertible note offering announced September 17 by November 30, 2026.

84%

BY NOVEMBER 30, 2026

TRIGGER: Hit only if a CoreWeave filing states gross proceeds of at least $3 billion from the announced notes.

PREDICTION 04 SOFTWARE

Salesforce makes Koa generally available in at least one US region by March 31, 2027.

67%

BY MARCH 31, 2027

TRIGGER: Hit only if Salesforce documentation marks Koa generally available, not pilot or beta, in a named US region.

PREDICTION 05 NETWORKING

A major AI networking vendor adds workload-level token-throughput correlation to a generally available fabric telemetry product by March 31, 2027.

37%

BY MARCH 31, 2027

TRIGGER: Hit only if public product documentation joins network telemetry to model token throughput for a named workload.

TRACK RECORD

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

Predictions made	**116**	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.

4 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 21-25

WATCH 01

Independent Gemini 3.8 Live latency and interruption tests

The shipped rate card is useful only when paired with measured turn latency, tool-call delay, and interruption recovery.

SEP 21-30

WATCH 02

CoreWeave convertible pricing and closing

Coupon, conversion premium, hedging cost, and final proceeds will show what capital markets charge for the next unit of neocloud capacity.

SEP 21-OCT 2

WATCH 03

Project Jupiter credit and permitting response

Any lender, county, or Oracle disclosure could distinguish temporary syndication pressure from a material construction risk.

OCT 2026

WATCH 04

Koa pilot evidence and open-beta timing

Named pilots need workload definitions and error baselines before the vendor's three-times-fewer-errors claim can inform procurement.

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 22

Week 38 of 2026 · September 19, 2026

WEB

brianletort.ai/industry

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