

HUMAN INTEREST · #GRID

Sunday, August 30, 2026

# Orchestrated job curtailment unlocks 100 GW of grid capacity

PRIMARY SOURCE   Volts · volts.wtf · published Aug 26, 2026 · 68 minute podcast

THE ARGUMENT

Emerald Conductor software ramps, pauses, and relocates GPU jobs to meet over 200 grid events while protecting token throughput, trading curtailment for faster, larger interconnections.

CONTEXT — WHERE THIS SITS

Emerald AI CEO Varun Sivaram tells Volts its Emerald Conductor software just moved from kilowatt demos into live multi-megawatt deployment. PJM’s market monitor calls such flexibility ‘a regulatory fiction.’ ERCOT counters with a binding ‘provisional controllable load resource’ tariff and a \$2,000 per MWh dispatch price.

TOP LEARNINGS

- 01 Utilities grant larger, faster interconnections when a load commits to curtail, because they then build no new capacity.
- 02 Compute flexing adds no cost by hour 24, so on-site batteries only need to cover short events.
- 03 Massively synchronized frontier training runs resist curtailment, so the flexible pool is fine-tuning and batch inference jobs.

BY THE NUMBERS

|                 |                                                                                                         |
|-----------------|---------------------------------------------------------------------------------------------------------|
| 100 GW          | New AI load Duke’s Nicholas Institute says the existing grid can absorb without new generating capacity |
| 0.5%            | Share of uptime a new load must curtail to unlock that capacity                                         |
| 200 grid events | Events met across published demos while protecting token throughput of critical jobs                    |
| \$2,000 per MWh | Willingness-to-curtail price at which ERCOT dispatches a ‘provisional controllable load resource’       |

TRIPWIRES

- 01 Watch the Vera Rubin Test Center’s 96 MW telemetry. If Emerald AI protects critical token throughput through repeated grid events, flexibility scales beyond controlled pilots.
- 02 Watch utility approvals granting 75 MW instead of 50 MW. If approvals avoid upgrades through enforceable curtailment, flexible loads earn larger and faster interconnections.

EVIDENCE

- [1] rtinsider.com — <https://www.rtinsider.com/125894-duke-university-study-benefits-data-center-flexibility/>
- [2] keentelengineering.com — <https://www.keentelengineering.com/ercot-pclr-batch-zero-guide>
- [3] utilitydive.com — <https://www.utilitydive.com/news/data-centers-flexibility-utilities-speed-to-power/822588/>
- [4] texasenergyandpower.com — <https://www.texasenergyandpower.com/p/how-load-flexibility-could-unlock>

LISTEN TO THE PIECE   <https://www.volts.wtf/p/making-data-centers-flexible-so-they>

RED TEAM

COMPLICATES

The piece says 0.5% curtailment can add roughly 100 GW without new generation. Duke models 76–126 GW across 22 balancing authorities, while ERCOT yields only 10–15 GW and competing load growth reduces headroom.  
texasenergyandpower.com

COMPLICATES

The piece says NVIDIA, EPRI, and National Grid demos met over 200 events while protecting critical token throughput. Public records describe small staged or simulated pilots, without a field-verified sample from full-scale production clusters.  
utilitydive.com

WHAT HELD

ERCOT’s dispatchable PCLR pathway for large flexible loads survived scrutiny.

# Agentic workloads move the CUDA moat into KV-cache software

PRIMARY SOURCE SemiAnalysis · newsletter.semianalysis.com · published Aug 24, 2026 · 45 minute read

THE ARGUMENT

SemiAnalysis replays 393 anonymized Claude Code traces across 1,000+ chips, showing Nvidia’s lead lives in KV-cache routing, offload, and tokenization, where AMD’s open-source stack trails.

CONTEXT — WHERE THIS SITS

SemiAnalysis’s AgentX 1.0 is a \$3M open-source agentic coding benchmark at 1M context, with tier-1 lab capacity teams already consuming its results API. Its contrarian move is benchmarking the whole serving stack, not kernels. AMD’s strength shows only in ATOM, which no major lab runs in production.

- TOP LEARNINGS
- 01

Agentic inference is a systems problem, so cache routing, offload, and tokenization decide perf/\$ more than decode kernels.
- 02

Under agentic load, HBM capacity decides cache hit rates, so B300 squeezes out throughput that B200 cannot.
- 03

At 150 tok/s/user on GLM 5.3, Nvidia stays cheaper per token even if AMD chips were free.

BY THE NUMBERS

|      |                                                                                                                              |
|------|------------------------------------------------------------------------------------------------------------------------------|
| 5x   | Nvidia cost-efficiency lead over AMD on GLM 5.3 at 150 tok/s/user, even with AMD hardware priced free                        |
| 20x  | Nvidia’s SGLang performance advantage over AMD on Qwen3.5 397B at 90 tok/s/user                                              |
| 95%+ | Prefix-cache hit rate vLLM reported after selective retention, with fourteen concurrent requests at contexts up to 1M tokens |
| 70+  | Upstream optimization PRs across vLLM, SGLang, TensorRT-LLM, ATOM, and Mooncake that use AgentX as north star                |

- TRIPWIRES
- 01

Watch production trace distributions across major DeepSeek V4 deployments. If medians converge near 88k and p95 reaches 404k, fixed-length benchmarks lose production relevance.
- 02

Watch vLLM’s AMD DCP and PCP support matrix. If optimized AMD backends post competitive results, context parallelism stops extending the CUDA moat.

EVIDENCE

[1] vllm.ai — <https://vllm.ai/blog>

[2] signal65.com — <https://signal65.com/research/ai/amd-instinct-mi355x-vs-nvidia-b200-inference-evaluation/>

[3] inferencex.semianalysis.com — <https://inferencex.semianalysis.com/compare/deepseek-r1-b200-vs-mi355x>

[4] vllm.ai — <https://vllm.ai/events/vllm-conference/2026>

READ THE PIECE <https://newsletter.semianalysis.com/p/agentx-inferencexv3-does-cuda-moat>

# Junior hiring collapse makes small AI unemployment politically explosive

PRIMARY SOURCE ChinaTalk · chinatalk.media · published Aug 24, 2026 · 90 minute podcast

THE ARGUMENT

Leicht argues even a 1-point unemployment uptick, if concentrated in junior white-collar roles that agents can replace, forces real national policy after the 2026 midterms.

CONTEXT — WHERE THIS SITS

ChinaTalk hands the floor to Carnegie fellow Anton Leicht, an ex-German policy hand. He maps AI’s road to ballot salience ahead of the midterms and a predicted 2027 policy window. Contrarian positions include corporate income tax over token taxes, data center backlash as own goal, and the NVIDIA open-weights letter versus the national security state.

- TOP LEARNINGS
- 01 If agents replace junior white-collar work, a coordination failure lets every firm skip entry-level hiring, destroying the five-to-eight-year talent layer needed to orchestrate agents.
  - 02 Token taxes punish the adopters you want and spare legacy firms, so corporate income tax remains the least distorting revenue option.
  - 03 If buildout shifts to allies like Australia under favorable access deals, the US sheds backlash risk, shares capital costs, and gets better sites.

BY THE NUMBERS

|                  |                                                                                                                         |
|------------------|-------------------------------------------------------------------------------------------------------------------------|
| 1 percent        | The transition-phase unemployment uptick Leicht expects, small enough that one targeted hit creates front-page politics |
| 10 or 20 percent | Doomer headline-unemployment scenario, above anything a functioning Western democracy has survived                      |
| \$100 million    | Extra per-site spend on quieting and amenities Schneider says the industry skips, calling backlash an own goal          |
| \$20,000         | Annual per-household payment, floated for ten years, to buy off host-town resistance                                    |

- TRIPWIRES
- 01 Watch 2026 AP VoteCast polling on AI voting priorities. If AI ranks below economic issues, the 2027 must-act window weakens.
  - 02 Watch Chinese frontier-chip output and domestic DUV shipments. If both scale despite controls, a narrow pause becomes asymmetric.

EVIDENCE

[1] americanedgeproject.org — <https://americanedgeproject.org/new-poll-more-than-9-in-10-likely-midterm-voters-prioritize-economic-is-sues-and-regulating-tech-and-ai-is-not-a-top-priority/>

[2] chinatalk.media — <https://www.chinatalk.media/p/how-ai-becomes-a-political-crisis>

[3] smarterx.ai — <https://smarterx.ai/smarterxblog/young-americans-ai-job-fears>

[4] brookings.edu — <https://www.brookings.edu/articles/competing-ai-strategies-for-the-us-and-china/>

LISTEN TO THE PIECE <https://www.chinatalk.media/p/how-ai-becomes-a-political-crisis>

RED TEAM

CONTRADICTS

The piece expects quiet 2026 midterms, followed by a 2027 window when post-midterm polling forces both parties to act. August 2026 election data showed AI becoming a major campaign issue before the midterms, reversing the proposed sequence.

[washingtonpost.com](https://www.washingtonpost.com)

COMPLICATES

The piece says a narrow frontier-development pause would disarm the U.S. semiconductor edge while China indigenizes DUV and frontier chips. Brookings finds export controls and limited capital still constrain Chinese compute, while the United States retains clear frontier-model and semiconductor leads.

[brookings.edu](https://www.brookings.edu)

WHAT HELD

The token-tax incidence claim survived because volume-based charges burden heavy adopters more than low-use legacy firms.

# Adoption outruns institutions, opening the world’s largest arbitrage

PRIMARY SOURCE   Aatish Nayak   · x.com   · 9 minute read

THE ARGUMENT

Nayak argues labs will “tokenmaxx” individual productivity while institutions lag adoption, and maps seven moves for app companies to convert that gap into durable moats.

CONTEXT — WHERE THIS SITS

This answers the “app layer is dead” narrative riding record lab revenue. Nayak concedes labs become the biggest companies ever yet says value accrues above them, as cloud did not stop Stripe or Salesforce. He cites Harvey’s F500 law-firm orchestration, Factory’s software-factory bet, and Profound’s neutral awareness layer.

- TOP LEARNINGS
- 01

Factory electricity predates its productivity statistics by decades, and institutional change still lags AI adoption today.
- 02

Labs will “tokenmaxx” individual productivity because it diffuses easiest, and that caps gains since firms exceed the sum of their people.
- 03

Once token spend becomes a 2027 P&L line item, companies already pricing on outcomes like resolved claims will own the economics.

BY THE NUMBERS

|                       |                                                                                                                  |
|-----------------------|------------------------------------------------------------------------------------------------------------------|
| 1880s-1920s           | Electricity’s factory wiring versus productivity statistics, the piece’s evidence that institutions lag adoption |
| 2027                  | Year token and AI spend becomes a P&L line item C-suites must justify                                            |
| order of magnitude    | Per-cycle compression of adoption speed while institutional change keeps lagging                                 |
| thousand agent swarms | Armadin’s vulnerability-hunting scale no service firm could replicate, backing outcome pricing                   |

RED TEAM

COMPLICATES

The piece says fragmenting model ecosystems will divorce continual learning from models, letting application companies own durable memory. A 2026 technical analysis finds continual updates alter the representations that stored memories reference, making durable memory harder to separate from models.

baseten.co

COMPLICATES

The piece says 2027 will make token spend a P&L line item and position outcome-priced application companies to own the economics. A June 2026 finance article reported token spending already reaching P&Ls, weakening 2027 as a distinct budgeting inflection.

investmentnews.com

WHAT HELD

The claim that powerful labs can coexist with major value for application companies survived the available counter-evidence.

- TRIPWIRES
- 01

Watch enterprise filings for model-switching retention and migration costs. If retention survives model changes without migration losses, workflow gravity becomes a durable moat.
- 02

Watch 2027 filings for token budgets and outcome-priced revenue. If outcome-priced revenue expands margins, application companies can own the economics.

EVIDENCE

[1]

analysisdata.co.id — <https://analysisdata.co.id/index.php/JEBI/article/view/263>

[2]

baseten.co — <https://www.baseten.co/research/continual-learning/>

[3]

larridin.com — <https://larridin.com/blog/ai-budget-2027>

[4]

bis.org — <https://www.bis.org/publ/work1325.pdf?>

READ THE PIECE   <https://x.com/nayakkayak/status/2092276375935963537>

# Write endurance decides which AI bytes belong on NAND

PRIMARY SOURCE   BEP Research   ·   bepresearch.substack.com   ·   published Aug 29, 2026   ·   36 minute read

THE ARGUMENT

Endurance arithmetic converts SanDisk’s 420-exabyte KV forecast into a DWPDP budget where write amplification and reuse economics decide which bytes earn flash placement.

CONTEXT — WHERE THIS SITS

A third of SanDisk’s 2030 data-center NAND forecast is KV cache, a write stream on media priced per byte but consumed by writes. The piece rates HBF a weights tier while NVIDIA’s Dynamo and CMX route KV bytes over fabric. It retracts its July 2027 HBF date to SanDisk’s late-2028 roadmap.

- TOP LEARNINGS
- 01

NAND is priced by the byte but consumed by the write, so capacity forecasts become endurance forecasts once write amplification enters.
- 02

Persisted KV earns flash placement only when expected reuse makes reload plus admission-write cost cheaper than recompute.
- 03

NVIDIA’s placement software owns the endurance budget and collects repeatedly, while the flash vendor collects once at part sale.

BY THE NUMBERS

|         |                                                                                                                   |
|---------|-------------------------------------------------------------------------------------------------------------------|
| 420 EB  | KV cache slice of SanDisk’s 1.2 ZB 2030 data-center NAND forecast, the write-heavy third the deck prices as bytes |
| 320 KiB | KV state written per token at the piece’s reference attention configuration, the input to every endurance figure  |
| 42 EB   | media writes per day per 0.1 physical DWPDP once scaled to the 420 EB tier                                        |
| 8x      | DeepSeek’s KV cache reduction per million tokens from R1 to V4 in about a year, the compression bear              |

- TRIPWIRES
- 01

Watch SanDisk’s published 2030 workload table. If it confirms 420 EB of KV cache, the forecast attribution holds.
- 02

Watch production KV offload disclosures with bytes per token and WAF. If measured wear stays within rated DWPDP, persisted KV belongs on NAND.

EVIDENCE

[1]

investor.sandisk.com — <https://investor.sandisk.com/events/event-details/2026-sandisk-investor-day>

[2]

bepresearch.substack.com — <https://bepresearch.substack.com/p/the-write-problem-which-ai-bytes>

[3]

moomoo.com — <http://www.moomoo.com/community/feed/2026-sandisk-investor-day-117080241668101>

[4]

news.ainvest.com — <https://news.ainvest.com/deep-topic/topic/sandisk-investor-day-sparks-storage-sector-rally-2608/>

READ THE PIECE   <https://bepresearch.substack.com/p/the-write-problem-which-ai-bytes>

RED TEAM

COMPLICATES

The piece treats SanDisk’s 420 EB KV cache slice as a primary 2030 forecast. Counterpoint describes persistent KV as an aggressive scenario and does not verify that specific breakdown.

[counterpointresearch.com](https://counterpointresearch.com)

COMPLICATES

The piece uses 320 KiB per token to derive 0.984 physical DWPDP at 3.0 WAF. A Llama-3-70B example reaches about 2.5 MB per token, making that endurance result model-specific.

WHAT HELD

The claim that persisted KV faces admission, reuse, reload, and endurance gates survived.

# China’s NdFeB magnet chokehold, not software, gates humanoid robotics

PRIMARY SOURCE · CrossCurrents · crosscurrents28.substack.com · published Feb 26, 2026 · 41 minute read

THE ARGUMENT

Three price shocks show non-Chinese magnet supply barely moves while mass humanoid deployment needs 186x today’s NdFeB output.

CONTEXT — WHERE THIS SITS

This piece rejects the software-first humanoid view and extends the author’s China debt thesis, framing embodied AI as China’s “Second Lever” growth engine. It argues Morgan Stanley and Adamas models omit robotics from demand, so analysts understate magnet repricing. It calls a Western ‘Brain’ versus Chinese ‘Body’ division a strategic suicide pact.

- TOP LEARNINGS
- 01

Under mass humanoid deployment, NdFeB output must expand 186x, two orders of magnitude beyond battery inputs.
- 02

Past NdPr price spikes barely moved non-Chinese supply because new capacity needs 15 to 20 years.
- 03

When a bottleneck input clears by allocation, robot manufacturing clusters around its source, so learning compounds in China.

BY THE NUMBERS

|             |                                                                                                  |
|-------------|--------------------------------------------------------------------------------------------------|
| 186x        | Required NdFeB output expansion under mass humanoid deployment, versus 14x for lithium           |
| >90%        | China’s control of global magnet manufacturing and the chemical midstream behind it              |
| ~95%        | Share of humanoid build cost in actuator-heavy limb and joint modules, about \$52.3k per unit    |
| 15–20 years | Lead time from geological discovery to commercial oxide production, the core inelasticity driver |

- TRIPWIRES
- 01

Watch Morgan Stanley’s next Humanoid 100 BOM. If actuation approaches 95%, scaling collapses into a materials question.
- 02

Watch IEA data on China’s NdPr separation share. If China’s share falls below 70%, its chemical midstream stops gating Western humanoid scale.

EVIDENCE

[1]

blog.robozaps.com — <https://blog.robozaps.com/b/economics-of-humanoid-robot-production>

[2]

standardbots.com — <https://standardbots.com/blog/tesla-robot>

[3]

reddit.com — [https://www.reddit.com/r/robotics/comments/1k2ecr7/what\\_are\\_your\\_thoughts\\_on\\_figure\\_ai/](https://www.reddit.com/r/robotics/comments/1k2ecr7/what_are_your_thoughts_on_figure_ai/)

[4]

x.com — [https://x.com/lukas\\_m\\_ziegler/status/2051307452541382998](https://x.com/lukas_m_ziegler/status/2051307452541382998)

READ THE PIECE <https://crosscurrents28.substack.com/p/ndfeb-magnets-the-material-bottleneck>

RED TEAM

CONTRADICTS

Actuator-heavy limb and joint modules carry \$52.3k, or 95% of humanoid build cost, so scaling becomes a materials question. A McKinsey-cited analysis puts actuation at 40–60% of a \$32k–\$35k bill, leaving several major cost centers.

COMPLICATES

NdFeB’s roughly 400 kJ/m³ energy density leaves no commercially viable alternative and forces motors using weaker magnets to become 4x larger. Commercial SmCo reaches 16–33 MGOe, roughly one-half to two-thirds of top Nd-FeB grades, weakening a universal 4x motor-size rule.

WHAT HELD

China’s roughly 90% control of separation and magnet production survived the evidence review.