SECOND ORDER

AI, Capital & Work · A Second Order research briefing

ISSUE 07 · JULY 20, 2026

How Long Can an AI Chip Earn?

Separating prompt processing from answer generation can give older GPUs another job later in life. The reviewed loans still depend on the original customer's payments.

One claim, followed through vendor materials, benchmark results, and company filings.

IN THIS BRIEFING

How Long Can an AI Chip Earn?

THE THROUGH-LINE

An AI model does two different jobs when it answers a question. It first processes the prompt and builds a working memory, then produces the answer one small unit of text at a time. The first job wants a burst of computation; the second often waits on memory and communication. Running them on different machines can give an older GPU a useful role alongside a newer or specialized processor, provided the working memory moves between them quickly enough.

WHY NOW

NVIDIA and Sarvam AI reported in February 2026 that separation and related tuning lifted completed-token throughput on one H100 system from a 1.00 baseline to 2.00. A June 2026 research prototype paired Tenstorrent hardware with NVIDIA A100s. AWS found in July 2026 that the split worked best for long prompts and heavy simultaneous use.

WHAT YOU'LL LEARN

1. Current revenue from several GPU generations
2. Prompt processing and answer generation on different hardware
3. H100 retention and B200 replacement at a power-bound site
4. Equipment-life, customer-term and debt-maturity timelines
5. Equipment-life revisions and reported depreciation
6. Contract cash flow, guarantees and equipment controls
7. Three events would show post-contract financing works

IF YOU REMEMBER ONE LINE: Our base case assigns no earnings to years 6 to 10 without post-contract economics for a named fleet.

THE CLAIM

Older GPUs can get another job later in life; the reviewed loans still depend on the original customer's payments

> *Separating prompt processing from answer generation can give older GPUs productive work. In a February 2026 NVIDIA and Sarvam AI demonstration, separation and related tuning helped lift completed-token throughput from a 1.00 baseline to 2.00 on one H100 system. The reviewed loans take the GPUs as collateral but repay from original customer payments.*
>
> Second Order · Issue 07 claim

1.5x

the throughput step from the separation itself in the NVIDIA and Sarvam AI H100 demonstration; 2.00 with related tuning, from a 1.00 baseline.

$0.126

per million added tokens, the modeled requirement for B200 replacement at $32,000 per nameplate kW, three years and 70% utilization.

2 to 6 yr

disclosed GPU debt maturities in the broader financing sample. No reviewed maturity reaches the proposed ten-year horizon.

EXHIBIT 1

The split works on H100s; an old-vintage fleet remains unobserved

Disaggregated-inference evidence by maturity stage: who runs prefill, who runs decode, and what links them

Scheduler and router allocate requests across both pools

EVIDENCE	PREFILL	DECODE	TRANSFER & SYSTEM LAYER	WHAT IT PROVES
NVIDIA + Sarvam AI	H100	H100	SGLang router, 1P+1D KV transfer	Kernels, scheduling, and the split lifted one H100 case to 2.0x; the split alone about 1.5x
AWS + Cerebras	AWS Trainium	Cerebras CS-3	Amazon EFA; Bedrock, forthcoming	Commercial design assigns phases to different accelerator types
NVIDIA Rubin + Groq 3 LPX	Rubin GPU	Groq 3 LPU	Fabric logic, BlueField-4, liquid-cooled rack	Current product pairs GPUs, LPUs, network, and rack
HMA-Serve (preprint)	Tenstorrent Blackhole	A100 80GB (old)	RDMA, cross-vendor KV-format conversion	Prototype puts an A100 80GB on decode; higher goodput per dollar
Life-extension claim	Hopper or Ampere	Cerebras or Groq LPU	Required, not quantified	Proposed 10 to 15 year vintage extension; no operator result

Note: NVIDIA and Sarvam use H100 for both phases; the split contributes roughly 1.5 times for the tested model, workload and latency target. AWS-Cerebras and NVIDIA Rubin-Groq use current-generation components. HMA-Serve assigns an A100 80GB to decode. No case reports post-year-five operator economics.

Source: NVIDIA, Cerebras, AWS, HMA-Serve and Gavin Baker.

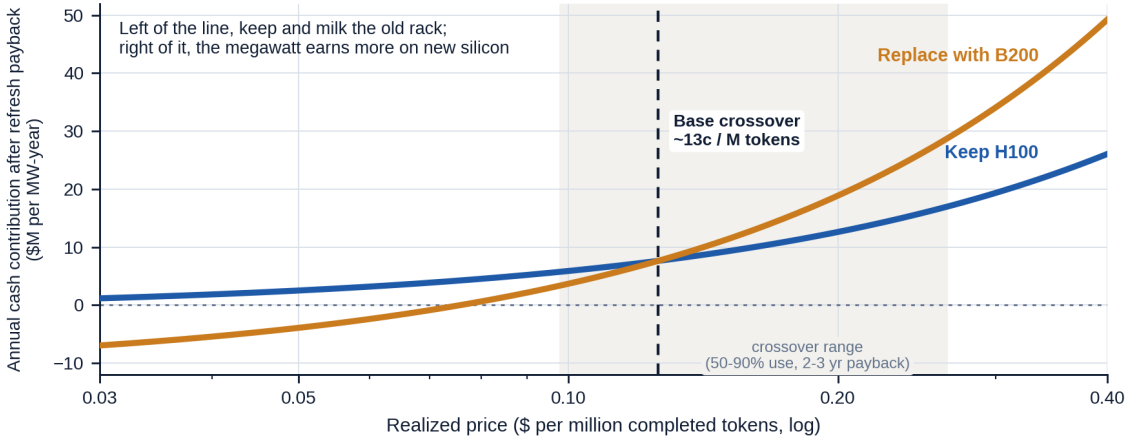
SO WHAT

- Producing an AI answer has two phases. The system first processes the prompt and builds a working memory, then generates the answer one token at a time. Separating those phases can match each job to different hardware.
- NVIDIA and Sarvam AI lifted a completed-token throughput index from 1.00 to 2.00 with separation and related tuning on one H100 system; the split itself accounted for roughly a 1.5 times step.
- HMA-Serve paired Tenstorrent hardware with A100s and reported up to 3.2 times more useful throughput. AWS found that the gain depended on long prompts, heavy use and fast movement of the working memory.

EXHIBIT 2

A profitable old rack can still lose a scarce megawatt

Annual cash contribution after refresh payback: keep an H100 rack vs. replace with B200, by realized price ($M per MW-year)



Note: Simple-payback sensitivity with no discounting. The keep floor uses delivered power of 4 to 12 cents per kWh plus 6 cents of avoidable operating cost. The replacement band assumes $32,000 per nameplate kW, a 2 to 3 year payback and 60 to 80 percent utilization. The calculation omits complementary hardware, networking, integration, downtime and H100 sale proceeds.
Source: MLCommons Inference v5.0 Llama2-70B Server (DGX H100, B200), NVIDIA max input power; realized-price anchors DeepSeek and Silicon Data token index; Second Order scenario.

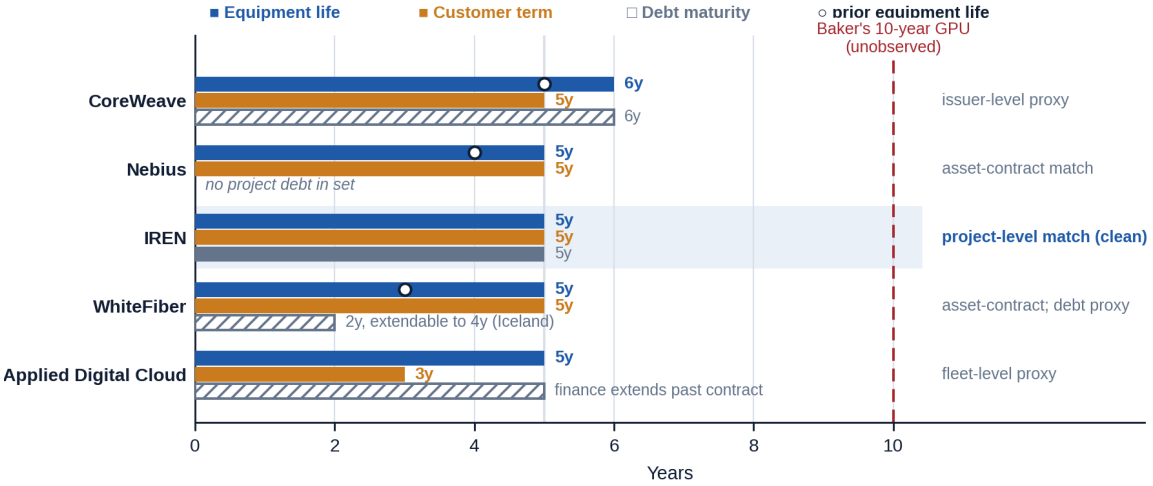
SO WHAT

- At a site with spare power and rack space, the hurdle for an owned H100 is its avoidable operating cost. At a full site, keeping it also means giving up the output of a newer server.
- MLCommons v5.0 results and NVIDIA specifications put a complete B200 server at about 2.3 times the completed tokens per maximum input kW of an H100 server. Spreading a purchase priced at $32,000 per nameplate kW across three years of added output at 70% utilization, our model requires about $0.126 per million added tokens.
- DeepSeek's $0.28 first-party V4 Flash output price supplies one customer-price anchor. The comparison excludes integration work, downtime, financing cost and the H100's sale value, so it does not establish payback.

EXHIBIT 3

No disclosed GPU-linked maturity in this source set reaches year ten

Equipment life, customer commitment, and disclosed debt maturity, by operator (years)



hatched debt bars: approximate or redacted; solid: disclosed

Note: Five operators selected for disclosure quality. CoreWeave is an issuer-level proxy; WhiteFiber's GPU loan applies to a different fleet from the Paris contract; Applied Digital Cloud is a fleet-level proxy. IREN is the only clean project-level asset-contract-debt match. Hollow circles mark prior equipment lives.

Source: CoreWeave, Nebius, IREN, WhiteFiber, and Applied Digital Cloud filings. Second Order calculations.

SO WHAT

- Operators in our five-company sample estimate equipment lives of five to six years and disclose customer terms of three to five years. The broader set of public GPU debt maturities runs from two to six years.
- IREN is the only same-project match across equipment life, customer term and debt: its five-year Microsoft project financing ends no later than the last Microsoft service payment.
- No reviewed maturity reaches the proposed ten-year life-extension horizon.

WHAT IT MEANS

Takeaways

1. Current revenue from several GPU generations

CoreWeave reported in May 2026 that Ampere, Hopper and Blackwell capacity were all earning revenue, and Applied Digital Cloud reported all 6,144 deployed GPUs earning revenue at year-end 2025. The financing question begins after the first customer contract.

2. Equipment-life revisions and reported depreciation

CoreWeave, Nebius and WhiteFiber lengthened equipment-life estimates to five or six years. These management estimates determine the timing of reported expense; the disclosures provide no sale-price forecast or lender judgment.

3. Contract cash flow, guarantees and equipment controls

DDTL 4.0's payment order directs receipts to operating costs, interest and principal before the owner can receive cash. Both agreements reduce a contractual GPU value evenly over six years, but that convention supplies no appraisal or expected recovery price.

4. What changes if older GPUs develop a real second market?

New GPUs would still enter the best-powered sites, while older high-memory fleets could move to cheaper electricity, less urgent work or facilities that cannot support the newest systems.

WATCHLIST

Three events would show post-contract financing works

An operator publishes vintage economics: economics for a named GPU vintage after year five, including its assigned job, utilization, realized revenue and margin, power and hosting costs.

A third party buys a fleet: and discloses its age, unit count and price.

A lender finances the next workload: advancing money after the original contract ends, relying on later cash flow or recoverable sale value.

What we would need disclosed: the spending on new accelerators, networks or site work required to keep the fleet useful, plus the customer's term and renewal, the loan's repayment schedule and maturity, and the hardware's eventual sale or recovery value.

SECOND-ORDER EFFECT · WHAT CHANGES IF OLDER GPUS DEVELOP A REAL SECOND MARKET? Financing that transition would require prices for used fleets, independent inspection and valuation, specialist owners willing to take customer turnover risk, and loans for the network, processor or site work needed to redeploy the machines.