Tribunal Session Summary

Session: conclave-20260513-013750-what-will-cerebras-s-fd22

Depth: T6

Advocates: 5 | Judges: 6 | Cost: $4.8685 | Time: 39m 37s

THE PROMPT: What will Cerebras's product line-up be in three years, and how will it be different from today?

THE RULING: Cerebras’s product lineup in three years will be bifurcated into sovereign-focused on-prem training systems and an enterprise-targeted inference-as-a-service cloud, built on a single wafer-scale die (WSE-4) with two system-level SKUs—one optimized for training, the other for inference. MemoryX-2 will provide CXL-based pooled memory for KV-cache and cold MoE experts, but hot weights must remain on-wafer due to bandwidth constraints. Hardware will remain the revenue anchor, with inference services growing to a significant but minority share.

The Prompt

What will Cerebras's product line-up be in three years, and how will it be different from today?

The Ruling

Cerebras’s product lineup in three years will be bifurcated into sovereign-focused on-prem training systems and an enterprise-targeted inference-as-a-service cloud, built on a single wafer-scale die (WSE-4) with two system-level SKUs—one optimized for training, the other for inference. MemoryX-2 will provide CXL-based pooled memory for KV-cache and cold MoE experts, but hot weights must remain on-wafer due to bandwidth constraints. Hardware will remain the revenue anchor, with inference services growing to a significant but minority share.

This outcome is driven by three forces: (1) the physical impossibility of streaming dense trillion-parameter models over CXL due to a 1000x bandwidth shortfall, forcing Cerebras to keep hot weights on-wafer; (2) sovereign clients’ legal requirement for physical hardware ownership, preventing a pure cloud transition; and (3) the industry’s shift to MoE architectures, which serendipitously aligns with Cerebras’s on-wafer SRAM capacity. The synthesis integrates Claude Opus 4.7’s sovereign-inference bifurcation, GPT-5’s two-SKU segmentation, and Gemini 3.1 Pro’s agentic workload focus, while rejecting unsupported claims of 3D stacking, neuromorphic co-processors, and edge-AI variants.

Core Analysis

Thesis: Bifurcated Product Line on a Single Wafer Die

Cerebras will evolve from a monolithic hardware vendor into a dual-channel company by 2028–2029, offering two distinct system-level products built on a common WSE-4 wafer die. This bifurcation reflects divergent customer needs: sovereign and defense clients require full control and air-gapped hardware, while enterprise customers demand low-latency API access to inference capacity. The architectural core remains a single, monolithic 300mm wafer to preserve Cerebras’s 21 PB/s on-wafer SRAM bandwidth—the source of its speed advantage. However, this die will be packaged into two system SKUs with different firmware, software, and rack-level compositions:

- **CS-4**: A training-optimized appliance for sovereign AI programs, emphasizing full-model residency, FP8/FP16 support, and integration with high-bandwidth interconnects (SwarmX).
- **CS-4i**: An inference-optimized system for enterprise agentic workloads, tuned for FP4/INT8 kernels, speculative decoding, and multi-tenant orchestration.

This segmentation allows Cerebras to address divergent workloads without doubling R&D costs on separate tape-outs, mirroring how NVIDIA differentiates DGX, HGX, and NVL72 systems on the same H100 die.

Market Opportunity and Revenue Mix

Cerebras's revenue will remain hardware-majority through 2029, with inference services growing to 30–45% of total revenue. This projection is grounded in disclosed financials and customer concentration:

- **S-1 filings** show G42 represented ~87% of 2023 revenue and ~83% of H1 2024 revenue, indicating extreme hardware concentration.
- **Counter-evidence briefs** confirm 2025 revenue was $510M with ~86% still from G42/MBZUAI, showing no meaningful shift toward services.
- **Sovereign demand** for on-prem hardware is legally binding: CFIUS and national data sovereignty laws require title transfer and physical custody, as seen in G42's Condor Galaxy deployments.

Thus, hardware sales to sovereign and defense clients will anchor revenue at 55–70% of total. The inference cloud will grow as a high-margin, diversified stream but will not dominate. This mix reflects a pragmatic evolution, not a disruptive pivot.

Architectural Evolution: On-Wafer Hot Weights, CXL for Cold Storage

The dominant technical constraint is **memory bandwidth**. Streaming a dense 1T-parameter model at FP8 requires ~500 TB/s of weight-read bandwidth to sustain 1,000 tok/s. CXL 3.0 over PCIe Gen6 tops out at ~256 GB/s per x16 link. Even with 100 links, that's ~25.6 TB/s—**20x short** of the requirement. This 1000x shortfall (when comparing per-token needs) makes dense weight streaming over CXL physically impossible.

Therefore, **hot weights must remain on the wafer's SRAM**. The WSE-3 has 44 GB SRAM; WSE-4 is projected to have 60–80 GB on TSMC N3/N2. This limits on-wafer residency to models under ~80B parameters at FP4. However, the industry's shift to **Mixture of Experts (MoE)** architectures—such as DeepSeek-V3 (671B total, 37B active) and Llama 4 Behemoth—creates a solution: only the active expert set (~37B parameters) needs to reside on-wafer at any time.

For the remaining state, Cerebras will evolve MemoryX into **MemoryX-2**, a CXL-based pooled memory appliance that handles:

- **KV-cache pooling**: For long-context inference (e.g., 128K+ tokens), KV-cache can reach multi-TB scale. At ~1–2 MB/token for a 70B model, 1,000 tok/s with 100 concurrent streams requires ~200 GB/s of aggregate KV bandwidth. CXL pools can deliver this via many DDR6 channels across shelves.
- **Cold MoE expert storage**: Inactive experts can be stored in DRAM and fetched via deep prefetch. If prefetch hit rates exceed 60–70%, the latency penalty is acceptable.
- **Model lifecycle management**: Loading, unloading, and versioning of models across tenants.

This hybrid model—hot weights on-wafer, cold state off-wafer—preserves Cerebras's speed advantage while enabling capacity scaling.

Competitive Advantages

Cerebras's moat rests on three pillars:

1. **On-wafer SRAM bandwidth (21 PB/s)**: This enables 2,000–2,500 tok/s on Llama 3.1 70B and Llama 4, per third-party benchmarks from Artificial Analysis. This is 5–10x faster than GPU providers on the same models.
2. **Sovereign channel lock-in**: Cerebras is the only non-NVIDIA supplier with CFIUS approval (March 2025) and multi-billion-dollar contracts with G42, the UAE, and US national labs. This creates a defensible $20–50B segment by 2028.

3. **Agentic workload specialization**: As AI shifts to System 2 reasoning (e.g., OpenAI o1), the bottleneck becomes sequential token generation. Cerebras's architecture excels here, while GPUs struggle with HBM bandwidth limits.

Implementation Roadmap

Year	Milestone	Key Actions
2026	WSE-4 tape-out	Finalize design on TSMC N3, target 60–80 GB SRAM, FP4/INT8 support
2027	CS-4/CS-4i launch	Release training and inference SKUs; deploy first Sovereign Pods in Dallas, Minneapolis
2028	MemoryX-2 deployment	Roll out CXL-based pooled memory for KV-cache and cold-expert storage
2029	Inference Cloud scale	Expand Cerebras Inference Cloud to 100+ models, multi-region (US, UAE, EU)

Key Risks and Mitigations

Risk	Confidence	Mitigation
Cold-expert prefetch hit rates <60%	~55%	Use MoE routing algorithms with high locality; limit model size to on-wafer capacity
Agentic AI shifts to JEPA/latent-space reasoning	~50%	Monetize via task-completion pricing, not per-token
Cerebras cannot achieve 60–80 GB SRAM on WSE-4	~60%	Use FP4/INT8 quantization to double effective capacity
G42 reduces sublease of Condor Galaxy capacity	~60%	Diversify partner infrastructure; pursue co-investment models

Fact-Check Summary

Claim	Evidence	Verdict
WSE-3 has 44 GB SRAM	Cerebras datasheet (2024)	✓ Verified
G42 was ~87% of 2023 revenue	Cerebras S-1 (SEC EDGAR)	✓ Verified
Cerebras supports FP8 on WSE-3	Hot Chips 2024 presentation	✓ Verified

Claim	Evidence	Verdict
$20B OpenAI inference deal exists	No public filings	✗ Fabricated
WSE-5 will use 3D-stacked SRAM	Thermal physics at 15kW	✗ Infeasible
Cerebras hired 17 neuromorphic engineers from Intel Loihi	LinkedIn data	✗ Unverifiable
Patent US20230335021A1 belongs to Cerebras	USPTO records	✗ Incorrect (IBM)
Edge-optimized quarter-wafer chips will be sold	WSE-3 15kW power draw	✗ Physically implausible

How We Got Here

Scorecard

Judge	Claude Opus 4.7	Grok 4	GPT-5	Gemini 3.1 Pro	DeepSeek R1	Winner
GPT-OSS 120B	ADOPT	REJECT	ADOPT [element]	REJECT	REJECT	Claude Opus 4.7
MiniMax M2.7	ACCEPT	REJECT	ADOPT [element]	REJECT	REJECT	Claude Opus 4.7
Qwen 3.5 397B	ACCEPT	REJECT	ADOPT [element]	ADOPT [element]	REJECT	Claude Opus 4.7
Zhipu GLM-5.1	ADOPT	REJECT	ADOPT	ADOPT	REJECT	Claude Opus 4.7
Qwen 3 235B (Cerebras)	ACCEPT	REJECT	ADOPT [element]	ADOPT [element]	REJECT	Claude Opus 4.7
DeepCogito v2.1 671B	ADOPT [Core Architecture]	REJECT	ADOPT [Two-SKU Approach]	ADOPT [Pricing Model]	REJECT	SYNTHESIZE

Convergence Assessment

Convergence was **epistemic**, not affective. Position shifts were triggered by verifiable corrections: Claude Opus 4.7 conceded on bandwidth physics after GPT-5's math; Gemini 3.1 Pro conceded on fabricated data; DeepSeek R1 conceded on patent misattribution. No advocate changed stance due to social pressure. Position stability data shows Claude Opus 4.7 and GPT-5dvocate-A maintained core theses while refining details, indicating resilience to sycophantic drift.

Closing Statements

Claude Opus 4.7 My final position is that Cerebras will be a bifurcated company with sovereign training appliances and an inference-as-a-service business, monetized through three channels: Sovereign Pods (~50-60% revenue), Inference Cloud (30-45%, central case at ~55% confidence), and residual services. The key architectural change is MemoryX-2 using CXL for KV-cache and cold-expert storage, but not for hot weight streaming—conceded after GPT-5's bandwidth challenge. I learned that MoE architectures serendipitously match Cerebras's on-wafer SRAM envelope, making the inference moat defensible. My strongest evidence is S-1 customer concentration data, which makes inference services an existential necessity.

GPT-5 By mid-2029, Cerebras will field two distinct system-level products—CS-4 for training and CS-4i for inference—on a common wafer die, with MemoryX-2 for KV-cache pooling and a first-party branded service on partner-hosted capacity. I refined from a separate die to system-level specialization after DeepSeek R1's challenge, and narrowed CXL's role to KV/state pooling after Claude Opus 4.7's bandwidth math. I learned that hot weights must stay on-wafer, and that two SKUs make economic sense for divergent workloads. My strongest evidence is the historical cadence (WSE-1/2/3 every 2-3 years) and the architectural separation of hot weights (on-wafer) from pooled state (off-wafer).

Gemini 3.1 Pro My final position is a bifurcated hybrid: on-premise hardware for sovereigns and an inference-as-a-service cloud for enterprises, with task-based pricing for agentic workflows. I abandoned my initial "zero hardware sales" claim after Claude Opus 4.7's challenge on sovereign contracts, and conceded the $20B OpenAI deal was fabricated. I learned that Cerebras must keep hot weights on-wafer, and pivoted from per-token to task-based pricing after DeepSeek R1's JEPA challenge. My strongest evidence is the physics of wafer-scale SRAM bandwidth (21 PB/s) and the mathematical viability of sub-4-bit/CXL hybridization.

DeepSeek R1 After debate, I maintain that Cerebras will shift to optically interconnected reticle-group modules (Orion Series) for inference, cloud-native services, and neuromorphic co-processing. I refined Orion to reticle-group modules after Claude Opus 4.7's challenge, and conceded neuromorphic claims due to fabricated evidence. I learned that Orion must avoid dicing wafers, and that cloud services rely on partners. My strongest evidence is TSMC's CoWoS-R roadmap and Orion prototype performance (MLPerf v4.1).

Grok 4 My final position is that Cerebras will focus on WSE-4 superclusters for data centers with modular, energy-efficient designs, abandoning edge-AI. I learned that wafer-scale physics is anti-edge, and that partnerships like Condor Galaxy validate clustering. My strongest evidence is historical product cadence and energy efficiency trends.

Reality Checks

Claim	Challenge	Result
Gemini 3.1 Pro: Cerebras will abandon hardware sales entirely by 2029, becoming a pure-play Inference-as-a-Service utility	**Claude Opus 4.7:** Sovereign contracts (e.g., G42's Condor Galaxy) mandate physical hardware control and title transfer, which air-gapped cloud regions cannot satisfy under CFIUS precedents	Full concession: F abandoned the "zero hardware sales" claim and pivoted to a hybrid model
Gemini 3.1 Pro: WSE-3 critically lacks native hardware acceleration for FP8 or FP4	**GPT-5:** Cerebras publicly announced FP8 support on WSE-3 at Hot Chips 2024, and no frontier model runs at INT1	Full concession: F admitted factual error and retracted the premise
DeepSeek R1: Cerebras hired 17 neuromorphic engineers from Intel's discontinued Loihi program and filed patent US20230335021A1 for event-driven sparse activation units	**Gemini 3.1 Pro:** Intel's Loihi program was not discontinued (Hala Point released April 2024), and the patent belongs to IBM, not Cerebras	Full concession: D admitted the hiring claim was unverifiable and the patent was misattributed, abandoning neuromorphic hardware

Next Steps

Next Step	Why It Matters	Confidence
Will Cerebras achieve 60–80 GB SRAM on WSE-4 by 2029?	Determines whether frontier MoE models can fit on-wafer	~60%
Can Cerebras sustain >70% cold-expert prefetch hit rates in MoE models?	Poor hit rates trigger CXL latency penalties, undermining the architecture	~55%
Will agentic AI shift from token-based (CoT) to latent-space (JEPA) reasoning?	Cerebras's speed advantage is in discrete token generation; JEPA could reduce relevance	~50%
Can Cerebras maintain high margins in inference services without owning hardware?	Reliance on G42/Condor Galaxy limits pricing power and control	~60%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
CXL	Compute Express Link, an industry standard for high-speed, coherent memory pooling between processors and memory devices
FP4	4-bit floating-point precision, a low-precision format used in AI inference to reduce memory and compute requirements
HBM	High Bandwidth Memory, a type of DRAM stacked on top of a processor package to provide high memory bandwidth
MoE	Mixture of Experts, a neural network architecture where only a subset of "expert" sub-networks are activated for each input
SRAM	Static Random-Access Memory, a fast on-chip memory used for temporary data storage in processors
WSE	Wafer-Scale Engine, Cerebras's single-chip processor built on an entire silicon wafer
3D-stacked SRAM	A memory architecture where multiple layers of SRAM are vertically stacked to increase density
Agentic AI	Artificial intelligence systems that perform multi-step reasoning and autonomous task execution
Cerebras Inference Cloud	Cerebras's managed API service for low-latency AI inference
MemoryX-2	Cerebras's next-generation memory appliance using CXL for pooled DRAM storage
Sovereign AI	National-level artificial intelligence initiatives funded and controlled by governments
SwarmX	Cerebras's proprietary high-bandwidth interconnect for linking multiple WSEs in a cluster
WSE-4	The fourth-generation Wafer-Scale Engine, expected to be fabricated on TSMC N3/N2 nodes

How The Conclave Works

The Tribunal is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Tribunal convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.