廣發証券(香港)經紀有限公司
GF SECURITIES (HONG KONG) BROKERAGE LIMITED

Overseas
Technology

On-chip Memory: Growing but Niche

Henry Huang
SFC CE No. BYA749
henryhuang@gfgroup.com.hk

Jeff Pu, CFA
SFC CE No. BNO719
jeffpu@gfgroup.com.hk

What's new

We view on-chip memory as a niche AI inference trend. Cerebras (CRBS US) and Groq appear the most topical ones. We believe the trend will benefit EDA, SRAM IP and tier-1 foundry vendors while take a neutral stance towards Cerebras.

Comments

On-chip memory growing in AI inference but niche: We believe memory wall is accelerating adoption of on-chip memory in AI inference, as data movement increasingly constrains performance. AI inference is becoming more bandwidth-bound and latency-sensitive, driving disaggregated prefill/decode architectures, while on-chip SRAM offers ~100x higher bandwidth than DRAM to improve token rate, adopted by Cerebras and Groq. Although AI inference accelerator TAM is projected to grow from $122bn in 2025 to $736bn by 2029, ~57% CAGR, we expect premium level of high token rate services to remain niche due to low model size capacity, low throughput, and context limits, with the SAM of $7bn / 20bn / 40bn in 2026–2028.

Cerebras WSE with leading token rate but in niche market: We believe Cerebras differentiates through its wafer-scale on-chip memory architecture for memory-bound AI inference. Cerebras believes that WSE integrates 44GB SRAM, delivering 21PB/s on-chip and 214PB/s fabric bandwidth, enabling ~6x higher token rate and ~5x faster response vs. Groq LPU, while it is not apple-to-apple. However, we expect WSE to remain niche due to limited SRAM scaling, low off-chip I/O bandwidth, reliance on external memory, and higher TCO. While Cerebras leads in token rate, we see competitors such as SambaNova, Tenstorrent, and Google MPU in early stage, while the next to watch will be any M&As from tier-1 GPU/ASIC makers.

View on Cerebras, taking a neutral stance: First of all, we view Cerebras' MRA with OpenAI improving revenue visibility, supported by over $20bn backlog and take-or-pay commitment for 750MW capacity in 2026–2028, with an additional 1.25GW option, and backed by $1bn working capital loan and warrant incentives. We expect the agreement to accelerate Cerebras's transition from fabless toward cloud services, which pressures profitability. Additionally, we believe the multi-year partnership with AWS should add recurring revenue and reduce customer concentration risk. Based on the partnership with OpenAI/AWS and our analysis on TAM, we expect its revenue to be $1.2bn / $3.2bn / $5.5bn level in 2026/2027E/2028E, suggesting 10x 2028E P/S at current level. That said, the valuation doesn't seem particularly cheap, while the upside risks would be the breakthrough in scaling.

Supply chain beneficiaries: We believe the on-chip memory or wafer-scale architectures will drive demand for more sophisticated SRAM planning and EDA simulation, with memory compilers becoming a critical layer. We expect EDA and SRAM IP vendors to benefit, including Synopsys (SNPS.US) and Cadence (CDNS.US). For foundry and backend, we believe the design of SRAM, which has been difficult for die shrink, will benefit foundry makers such as TSMC (2330 TT Buy) and Samsung, and the wafer-scale architecture will also require SoIC WoW. Following current Groq's LP30/LP35 foundry by Samsung, TSMC's mgmt indicated the next-gen chip will be made by TSMC.

Risks

1) AI demand slowing down; 2) Geo-political risks; 3) Competition.

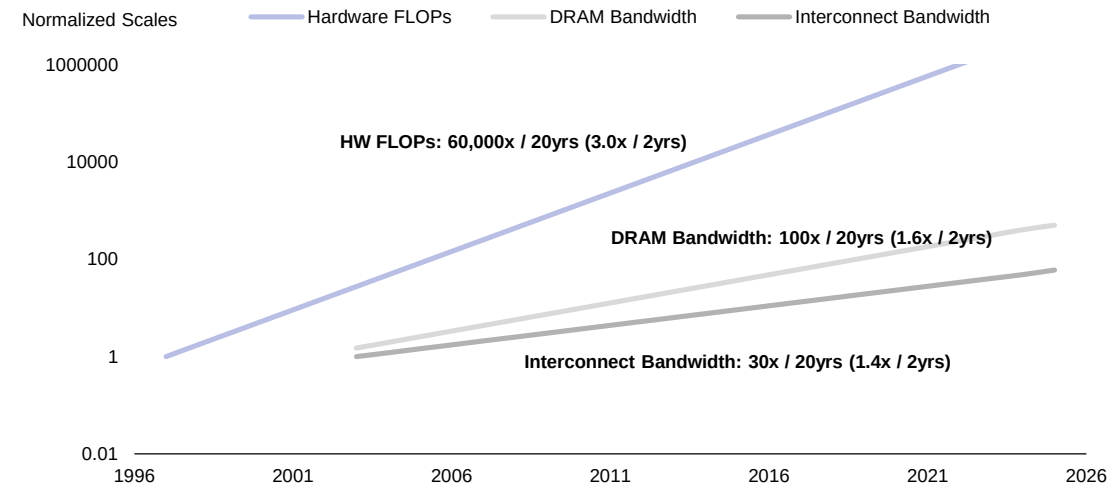
Memory wall drives on-chip memory adoption

AI chips are constrained by memory wall, where compute performance has significantly outpaced data movement, causing accelerators to spend more time waiting for data than performing actual computation. During the training phase, chip architecture prioritized compute intensity, with relatively low sensitivity to latency, making HBM's large capacity characteristic well suited for GPU workloads. As workloads shift toward inference, performance becomes increasingly bandwidth-intensive and latency-sensitive, driving the increasing use of disaggregated inference architectures to optimize data access and response time.

In disaggregated inference setup, prefill and decode workloads are routed to separate nodes. Prefill remains compute-intensive, while decode is increasingly bandwidth-bound; separating the two improves both TTFT and token generation consistency. As decode phase is sensitive to data access latency, on-chip memory becomes more important. Technology wise, DRAM latency is typically ~100ns vs. ~1ns of SRAM, or GPUs with ~100x more internal bandwidth than external DRAM/HBM. By keeping model weights resident directly in SRAM on chips, chips can bypass cache prefetching and memory wait states, reducing latency at the physical level. This has led vendors to adopt more SRAM as embedded primary on-chip memory to enhance bandwidth, such as Groq LPU and Cerebras WSE.

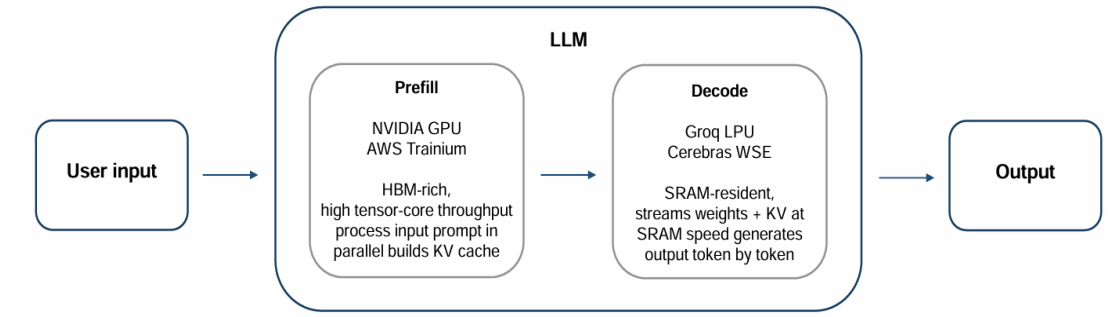
NVIDIA combines Vera Rubin NVL72 for prefill and LPX for decode delivers up to 35x higher TPS per megawatt at 400 TPS per user vs. GB200 NVL72, effectively creating a new premium performance tier for AI inference, and unlocking a new category of AI experiences on Pareto frontier (Throughput vs. Interactivity). AWS is also developing heterogeneous inference by combining Trainium and Cerebras WSE to accelerate memory-bound inference and improve token generation efficiency.

Figure 1: Scaling of peak hardware FLOPs, memory and interconnect bandwidth



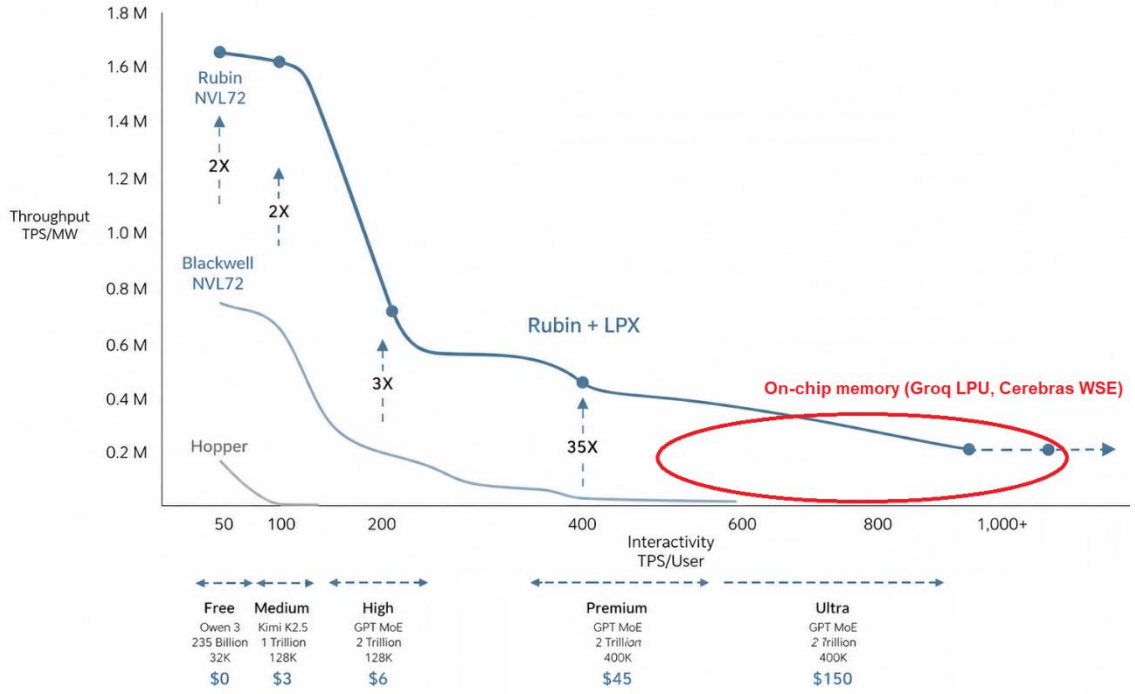
Sources: TrendForce, GF Securities (Hong Kong) Brokerage.

Figure 2: Disaggregated inference



Sources: GF Securities (Hong Kong) Brokerage.

Figure 3: NVIDIA unlocks new category of AI experience with Rubin and Groq LPX



Sources: NVIDIA, GF Securities (Hong Kong) Brokerage.

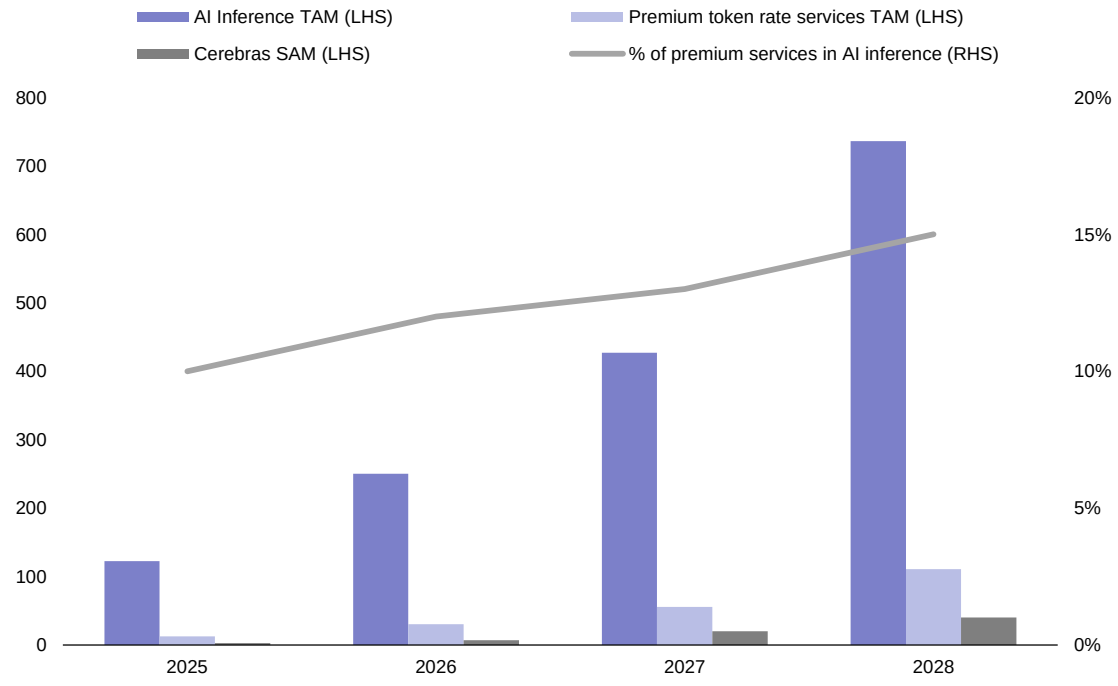
AI inference accelerating, while the position remain niche

Based on our analysis, we expect AI inference TAM to reach $736bn by 2029 from $122bn in 2025, with ~57% CAGR. Inference is expected to account for over 70% of total AI accelerator market in 2028, driven by rising usage, reasoning workloads, and token consumption. However, we expect premium level with high token rate services of on-chip memory application would stay a niche market, due to low throughput, low model size capacity, and low context absorption under agentic AI trend.

NVIDIA's CEO Jensen Huang mentioned in the recent earnings call and GTC, "The use case for LPX is not broad, intended for somebody who has a fairly large a portfolio of different types of token services. And for the high token rate, maybe these services are quite premium and the number of customers is not significant. I expect that LPX and other SRAM-based decode-focus token gen will always be a niche product for some time to come." And Jensen also gave the market share projection of these premium token rate services, "Whether it's 20% or 10%, just depends on where we are in the development of AI, I think today is a lot less than 20%. Someday, these premium tokens could be 20%."

We now forecast premium token rate services TAM to be $30bn / 56bn / 111bn in 2026-2028. Excluding NVIDIA, we estimate Cerebras SAM to be $7bn / 20bn / 40bn in 2026-2028.

Figure 4: AI inference TAM, premium token rate services TAM, Cerebras SAM forecast (USD bn)



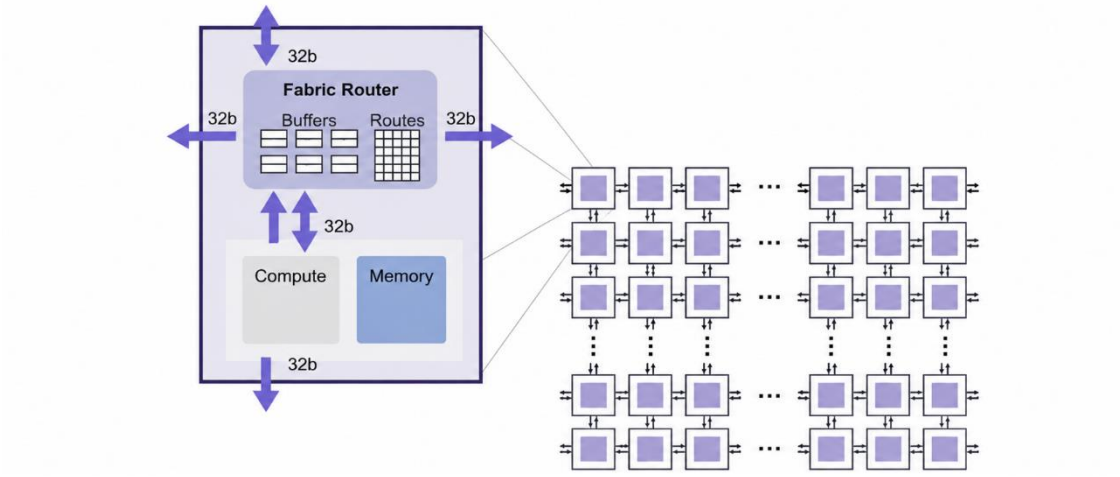
Sources: GF Securities (Hong Kong) Brokerage.

Cerebras WSE with extremely high on-chip bandwidth and token rate

WSE-3 integrates 84 interconnected “dies” onto a single silicon wafer spanning 46,225 mm², with TSMC’s 5nm process and incorporating 44GB on-chip SRAM, 4 trillion transistors and 900k independent ML-optimized cores. Its employs fully distributed on-chip SRAM to provide ultra-high aggregate memory bandwidth directly to compute units. Each WSE core is only ~0.05 mm², with approximately half allocated to 48 kB SRAM and the remainder to compute logic, and 2D mesh with embedded five-port router in every core enabling scaling, which eliminates memory wall constraints. Die-to-die connectivity relies on short metal wires (<1 mm) with redundancy and defect correction mechanisms. Compared to conventional off-chip SerDes, WSE’s on-chip fabric achieves ~7x higher bandwidth density and ~66x better power efficiency. Thus, WSE achieves extremely high on-chip memory bandwidth of 21PB/s and on-chip fabric bandwidth of 214PB/s.

WSE-3 equips 44GB of on-chip SRAM vs. Groq LPU of 500MB, NVIDIA B200 of 126MB, and Google TPU v8i of 384MB, with on-chip memory bandwidth of 21PB/s vs. Groq LPU of 150TB/s, B200 inter-chip of 8TB/s, and TPU v8i inter-chip of 19.2TB/s, making its leading token rate under AI inference, ~2.5x token rate of B200.

Figure 5: Cerebras WSE’s fully distributed on-chip SRAM architecture



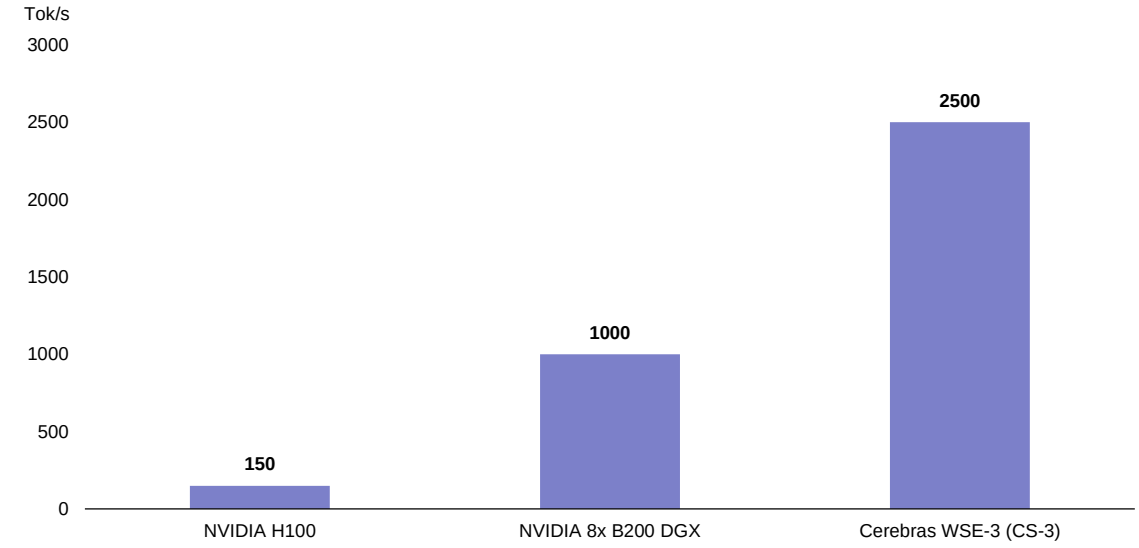
Sources: Cerebras, GF Securities (Hong Kong) Brokerage.

Figure 6: Cerebras WSE-3 spec

	WSE-3 Spec
Process node	TSMC 5nm
Chip area	46,225 mm²
Transistors	4 trillion
AI cores	900,000
On-chip SRAM	44GB
On-chip memory bandwidth	21PB / s
On-chip fabric bandwidth	214PB / s
Power	~23kW

Sources: Cerebras, GF Securities (Hong Kong) Brokerage.

Figure 7: Token rate per user comparison (Llama 4 Maverick benchmark)



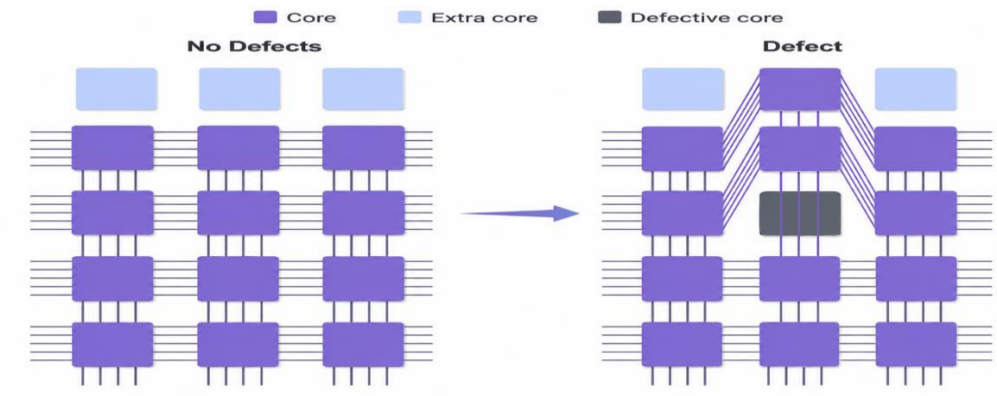
Sources: Cerebras, GF Securities (Hong Kong) Brokerage.

Failure routing architecture and cooling system form Cerebras's edge

Since each WSE core is ~0.05 mm², a defect in a core would disable ~1/900,000 of the whole chip. Cerebras developed a routing architecture enabling reconfigure connections between cores. When a defect is detected, the system can automatically reroute through redundant communication paths and leverages adjacent cores to preserve overall computational capability. In addition, each wafer batch uses customized upper metal layer mask, with tailored wiring layouts to bypass defective tiles. Thus, wafer-level yield remains high, with nearly 100% of TSMC wafer output meeting production server assembly requirements.

On the other hand, Cerebras employs a customized liquid-cooling stack co-designed with the wafer architecture to address the ~23kW thermal load in a single wafer. To support this design, Cerebras partnered with LiquidStack to develop a L2L single-phase CDU optimized for the CS-3's flow and pressure requirements. The cooling solution adopts a four-layer structure consisting of the cold plate, wafer, compliant connector, and PCB, with the cooling manifold integrated on the back side of the cold plate.

Figure 8: Cerebras Failure routing architecture



Sources: Cerebras, GF Securities (Hong Kong) Brokerage.

Model size capacity limits and low I/O bandwidth constrain Cerebras WSE in niche market

The biggest constraint for WSE is model size capacity. WSE-1 on TSMC 16nm equips 18GB of SRAM, and WSE-2 with 7nm jumped to 40GB, but only increased to 44GB in WSE-3. Across a full node transition, SRAM capacity only increased by 10%, while transistor grew by ~50%. Looking ahead, we expect it's hard for Cerebras on SRAM scaling, as the SRAM occupies 50% of chip area but density stops rising beyond TSMC N5 node. With limited SRAM capacity, the key disadvantage of Cerebras is model size capacity constraint in terms of total model parameters and KV cache.

In agentic AI era, model parameters are getting larger (Deepseek V4 with 1.6T total parameters). When the model is too big to fit in on-chip SRAM, Cerebras disaggregates weights from compute such that weights live in external memory server, MemoryX, which composes of DRAM and flash, and stream through the machine layer by layer to CS-3. Though on-chip bandwidth of WSE is extremely high, off-chip bandwidth is only 150GB/s, or 0.17GB/s per mm of edge, due to the trade-off inherent in wafer-scale design, where uniform reticle replication enables seamless on-wafer scaling but constrains shoreline density and external I/O expansion. Despite Cerebras's ongoing development of photonic interconnect wafer hybrid bonded onto the WSE to improve scale-out bandwidth, current I/O constraint remains a significant limitation for large model deployment.

In terms of KV cache, current mainstream LLM context is up to ~1M, which corresponds ~100GB KV cache. Due to limitation of only 44GB on-chip SRAM, each CS server is equipped with 6TB of DDR5 RDIMM for KV cache offload, with a dual socket AMD CPU. DDR5 offers lower bandwidth, and CPU mediation introduces additional latency, thus low off-chip bandwidth ultimately restricts scaling for large model deployment. Although KV cache compression techniques in the future might alleviate Cerebras's long-context serving problem, slow I/O remains bottleneck as KV cache transfers still take several milliseconds across on- and off-chip communication, affecting TTFT and utilization efficiency.

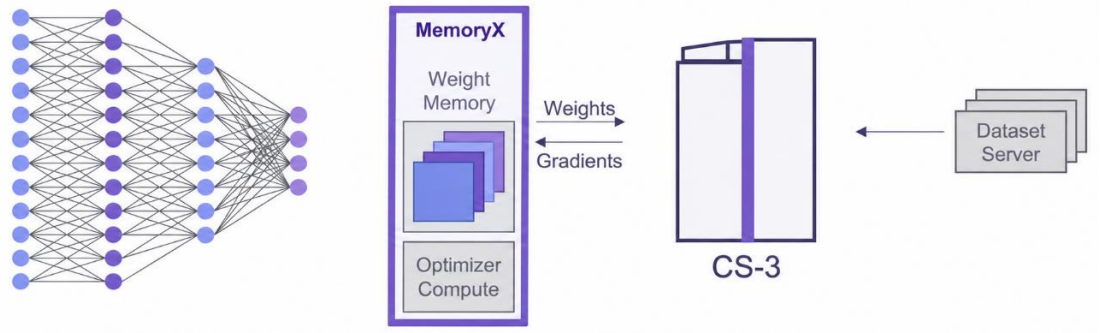
At the same time, limited SRAM capacity leads to low throughput per chip or per watt, and Cerebras didn't mention about any high-concurrency aggregate throughput information which shows it could win on tokens per second per dollar at scale. All in all, we expect Cerebras WSE remains in niche market due to these constraints in the near future.

Figure 9: SRAM scaling by node

Node	HD SRAM Cell (μm²)	Density (MB/mm²)
N7	0.027	25
N5	0.021	32
N3B	0.020	33
N3E	0.021	32
N2	0.021	32
A16	0.021	32

Sources: Semianalysis, TSMC, GF Securities (Hong Kong) Brokerage.

Figure 10: Cerebras's weight streaming architecture



Sources: Cerebras, GF Securities (Hong Kong) Brokerage.

Cerebras WSE delivers faster inference than Groq LPU, but with higher TCO

Groq's LPU features a silicon area of 725 mm², 512 MB of on-chip SRAM, and up to 150 TB/s of on-chip memory bandwidth. Given the finite capacity of on-chip SRAM, larger models are scaled across multiple interconnected LPUs through chip-to-chip (C2C) communication architecture. Each LPU is equipped with 96 C2C links operating at 112 GB/s, with aggregate scale-up bandwidth of 640 TB/s in LPX system configured with 256 LPUs. LPUs rely on extensive C2C communication for scaling, they incur higher latency and power overhead relative to Cerebras's on-chip interconnect architecture, resulting in LPUs scale-up bandwidth below that of Cerebras's fully integrated on-chip fabric.

Token rate wise, according to Artificial Analysis benchmark, Cerebras delivers token rate of ~6x vs. Groq LPU on the same models, such as GPT OSS 120B, Llama 4 Maverick, and Llama 3.3 70B, while it is not apple-to-apple. At the same time, Cerebras delivers end-to-end response time of ~5x faster than Groq LPU on identical models.

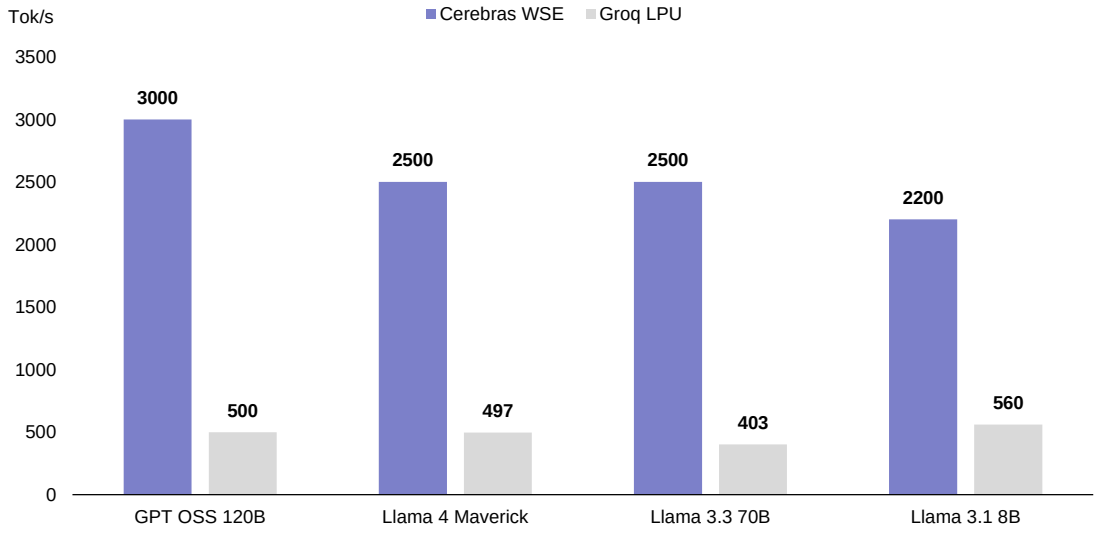
However, TCO of Cerebras is higher than Groq. Input / Output cost per million tokens on GPT OSS 120B of Cerebras is $0.35 / $0.75 vs. Groq of $0.15 / $0.6, and on Llama 3.1 8B of Cerebras is $0.1 / $0.1 vs. Groq of $0.05 / $0.08. As the power is finite, though Cerebras's token rate is much higher than Groq, cost per token or token per watt determine whether it can scale and remain margin at the same time. We expect Cerebras's performance advantage may outweigh the cost difference for small models, but higher TCO could become a limiting factor as model sizes scale, potentially slowing adoption and constraining share expansion.

Figure 11: Cerebras WSE and Groq LPU spec

	Cerevras WSE	Groq LPU
Process node	TSMC 5nm	Samsung 4nm
Chip area	46,225 mm²	725 mm²
On-chip SRAM	44GB	512 MB
On-chip memory bandwidth	21PB / s	150 TB/s
On-chip fabric bandwidth / Scale-up bandwidth	214PB / s	640 TB/s

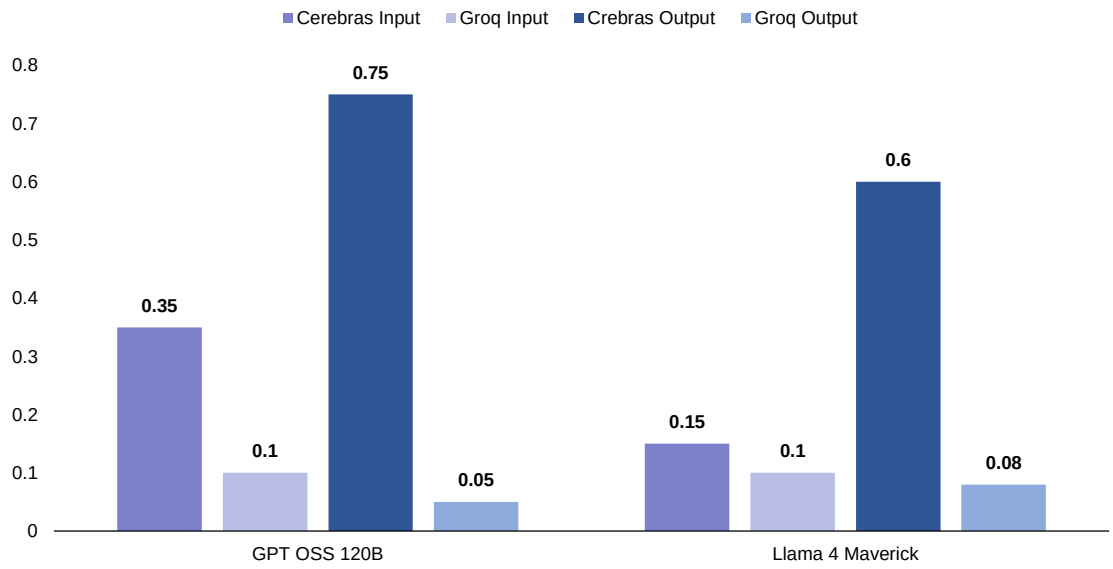
Sources: Cerebras, NVIDIA, GF Securities (Hong Kong) Brokerage.

Figure 12: Cerebras WSE and Groq LPU token rate



Sources: Artificial Analysis, GF Securities (Hong Kong) Brokerage.

Figure 13: Cerebras WSE and Groq LPU TCO ($ per million tokens)



Sources: Cerebras, GF Securities (Hong Kong) Brokerage.

Other on-chip memory firms soaring but in early stage

Beyond Cerebras and Groq, several AI vendors are also pursuing on-chip memory to mitigate the memory wall, including SambaNova, Tenstorrent, and Rebellions. However, most other competitors adopt a more balanced approach by maintaining external HBM/DRAM and scaling through conventional off-chip interconnect. SambaNova SN50 RDU equips 520MB SRAM, 64GB HBM, and 1.5TB DRAM, combining 16 RDU to form a SambaRack, then interconnecting several racks to execute inference, where lots of off-chip I/O and scale-out would make the latency compared with Cerebras on-chip data movement. In 2025, valuation of SambaNova is only ~$5bn, with expected ARR of ~$100mn.

Tenstorrent Galaxy Rack comprises of 32 Blackhole chips, in total 6.2GB SRAM with bandwidth of 2.9PB/s, 1TB DRAM with 16TB/s, and 56 800G Ethernet ports for 11.2 GB/s of scale-out bandwidth. Tenstorrent claimed that they perform faster than Cerebras and Groq on DeepSeek-R1-0528 671B model. Recently, Intel and Qualcomm are intended to acquire Tenstorrent. In 2025, valuation of Tenstorrent is only ~$3.2bn, with expected ARR of ~$360mn. Although current commercialization remains largely strategic partnerships and joint development projects, such as US$50mn automotive chip agreement with Korea's Hyundai and the 2nm R&D collaboration with Japan's LSTC, with up to 30% of revenue from IP licensing and silicon-as-a-service, we believe the company could emerge as a potential competitor in high token rate inference market in the future if acquired by tier-1 GPU/ASIC makers.

On the other hand, Google is also cooperating with Marvell to co-develop Memory Processing Unit (MPU) to work alongside with TPUs. MPU aims to offload in-memory computing tasks to alleviate TPU bottlenecks in memory bandwidth, thereby improving system efficiency in high-concurrency inference scenarios. MPU design will be finalized by 2027 at the earliest.

MRA with OpenAI secures revenue visibility, with potential margin compression

Cerebras reported a backlog o$24.6bn, mostly belonging to OpenAI under a take- or-pay agreement. Under the MRA, OpenAI committed to procure 750MW o AI inference capacity in three years, with 250MW deployed annually across 2026–2028. The initial 250MW deployment in 2026 will be delivered via the cloud, with OpenAI retaining extension options for years our and five, while deployments in 2027 –2028 can either remain delivering by cloud or convert into direct hardware purchases. In addition, OpenAI holds options for an extra 1.25GW ocapacity, bringing total potential deployment to 2GW. OpenAI also provided Cerebras with a $1bn working capital loan through a secured promissory note carrying a 6% annual interest rate.

At the same time, Cerebras issued warrants to OpenAI, granting 33,445,026 Class N common shares at an exercise price oUS$0.00001 per share, could be viewed as for free. Within the warrants, 4,459,337 shares vested immediately upon receipt o working capital loan, while 5,574,171 shares are tied to either Cerebras reaching a $40bn market capitalization or OpenAI achieving specified payment milestones under MRA. The remaining 23,411,518 shares are linked to committed and optional capacity deployment. Cerebras assessed that the working capital loan tranche, the market capitalization/payment milestone tranche, and the committed capacity tranche are probable ovesting, whereas the additional capacity tranche is currently not considered probable.

Corporate gross margin declines from 42.3% in 2024 to 39.0% in 2025, due to higher data center costs related to cloud services. Cloud services gross margin declined signiicantly to ~30% in 2025 from over 60% in 2024, reflecting higher data center costs, lower initial use and pass-through expenses tied to capacity expansion. We expect Cerebras's MRA with OpenAI could transform its business model from abless to asset-heavy cloud services, leading to margin compression in the near term.

AWS partnership encouraging, also reducing customer concentration risk

Beyond OpenAI, we view bullish on Cerebras' strategic collaboration with AWS, which is a multi-year cooperation with recurring payment. Under the term sheet, AWS is expected to become theirst hyperscaler to deploy Cerebras systems in its own data centers and jointly develop an inerence architecture integrating AWS Trainium3 and Cerebras CS-3. At the same time, Cerebras issued warrants to AWS, granting 2,696,678 Class N common shares at an exercise price o$100 per share, ~$270mn in total. We believe this helps reduce reliance on customers such as MBZUAI and G42 and improves the durability and quality olong-term growth.

Valuation assessment for Cerebras

Based on the partnership with OpenAI/AWS, our analysis on TAM and the assumptions of more competitors, we expect its revenue to be $1.2bn / $3.2bn / $5.5bn level in 2026E/2027E/2028E, suggesting 10x 2028E P/S at current level. That said, the valuation doesn't seem particularly cheap, while the upside risks would be the breakthrough in scaling.

Figure 14: Cerebras's revenue analysis

(USD bn)	2025	2026	2027	2028
Total revenue forecast				
AI accelerator TAM (ex. NVIDIA)	45	81	190	295
% of AI inference	60%	70%	80%	90%
AI inference TAM (ex. NVIDIA)	27	56	152	266
% of premium token rate services	10%	12%	13%	15%
SAM	3	7	20	40
Market share	20%	18%	16%	14%
Total Revenue	**0.5**	**1.2**	**3.2**	**5.5**
OpenAI revenue forecast				
OpenAI revenue		1.0	2.3	3.6
% of RPO		5%	10%	17%
OpenAI revenue contribution		80%	71%	65%

Sources: GF Securities (Hong Kong) Brokerage.

Key beneficiaries of on-chip memory trend

We believe the adoption of on-chip memory and wafer-scale architectures will require increasingly sophisticated SRAM planning and greater reliance on EDA simulation tools. In particular, memory compilers are expected to become a critical layer in enabling efficient SRAM generation and integration under this architecture trend. As a result, we expect design tool vendors and SRAM IP providers to benefit, including Synopsys (SNPS.US) and Cadence (CDNS.US).

For foundry and backend, we believe the design of SRAM, which has been difficult for die shrink, will benefit foundry makers such as TSMC (2330 TT Buy) and Samsung, and the wafer-scale architecture will also require SoIC WoW. Following current Groq's LP30/LP35 foundry by Samsung, TSMC's mgmt indicated the next-gen chip will be made by TSMC.

Risks

1) AI demand slowing down; 2) Geo-political risks; 3) Competition.

Rating definitions Benchmark: Hang Seng Index (Hong Kong)

Company ratings	**Buy**	Stock expected to outperform benchmark by more than 10%
	Hold	Expected stock relative performance ranges between -10% and 10%
	Underperform	Stock expected to underperform benchmark by more than 10%

Sector ratings	**Positive**	Sector expected to outperform benchmark by more than 10%
	Neutral	Expected sector relative performance ranges between -10% and 10%
	Cautious	Sector expected to underperform benchmark by more than 10%

Hong Kong	**Company**	GF Securities (Hong Kong) Brokerage Limited
	Address	27/F, GF Tower, 81 Lockhart Road, Wan Chai, Hong Kong
	Telephone	(852) 37191111
	Email	evanlee@gfgroup.com.hk

Disclaimer

Disclosure of Interests

1) The research analyst(s) and his/her associate has not served as an officer of the company(ies) mentioned in this report, and does not have any financial interests in the company(ies) mentioned in this report.

2) GF Securities (Hong Kong) Brokerage and/or its affiliates does not have any financial interests and has not invested in interests aggregate to an amount equal to or more than (i) 1% of the market capitalization; or (ii) 1% of the issued share capital, or issued units, in the company(ies) mentioned in this report.

3) GF Securities (Hong Kong) Brokerage and/or its affiliates does not have any market making activities, and has not employed any individual(s) serving as officer(s) of the company(ies) mentioned in this report.

4) GF Securities (Hong Kong) Brokerage and/or its affiliates does not have any investment banking relationship in Hong Kong with the company(ies) mentioned in this report in the past 12 months.