Cerebras's Next Generation CS-4: Fast Just Got Faster

Cerebras's Next Generation CS-4: Fast Just Got Faster Double the Performance with Double the Power

AUG 19 https://newsletter.semianalysis.com/p/cerebrass-next-generation-cs-4-fast

View in browser

Cerebras's Next Generation CS-4: Fast Just Got Faster

Double the Performance, Double the Power, Double the Fun

MYRON XIE, BRYAN SHAN, WEGA CHU, JORDAN NANOS, AND DYLAN PATEL

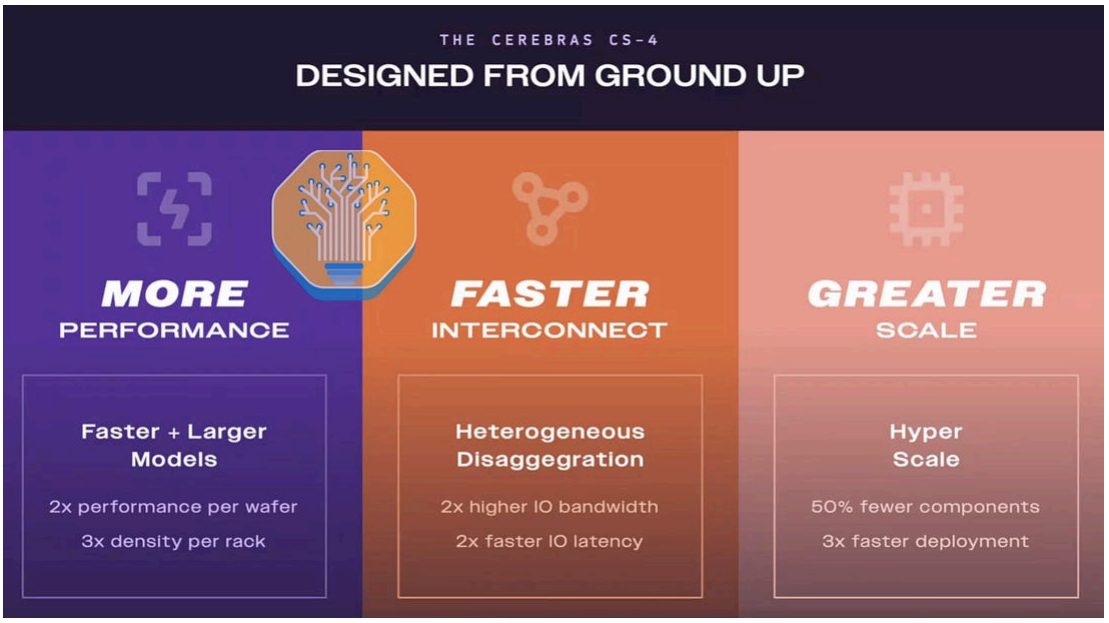
AUG 18 • PAID

READ IN APP

Cerebras revealed CS-4 this week, with more details to come at Hot Chips. CS-4 is their fourth-generation rack built around the same third generation 5nm wafer-scale engine: WSE-3. CS-4 doubles the performance of CS-3 through increased power consumption and clock frequency per wafer, and better rack-scale density.

This all translate into CS-4 being able to double the tokens/s/user per wafer from CS-3, and at around the same cost as the previous generation. This is a no-brainer for customers who can enjoy double the token revenue with the same hardware spend. It’s not only the tokens that are getting faster, but so is time time market: the rack architecture itself is redesigned to be more modular, allowing improved manufacturability and deployment times. Last but not least, a new I/O module will enable open, heterogeneous and disaggregated inference architectures going forward. These disaggregated inference setups will go a long way to help overcome the memory capacity constraints of the CS-4 by pairing it with HBM-based systems.

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis

Same Wafer, Double the Clock

Cerebras Specs		
	CS-3	CS-4
Dense FP16 Compute	15.6 PFLOPs	30.3 PFLOPs
Sparse FP16 Compute	125 PFLOPs	250 PFLOPs
Memory Capacity	44 GB	44 GB
Memory Bandwidth	21.6 PB/s	43.2 PB/s
On-chip fabric Bandwidth	26.7 PB/s	53.4PB/s
I/O Bandwidth	1.2 Tbit/s	2.4 Tbit/s

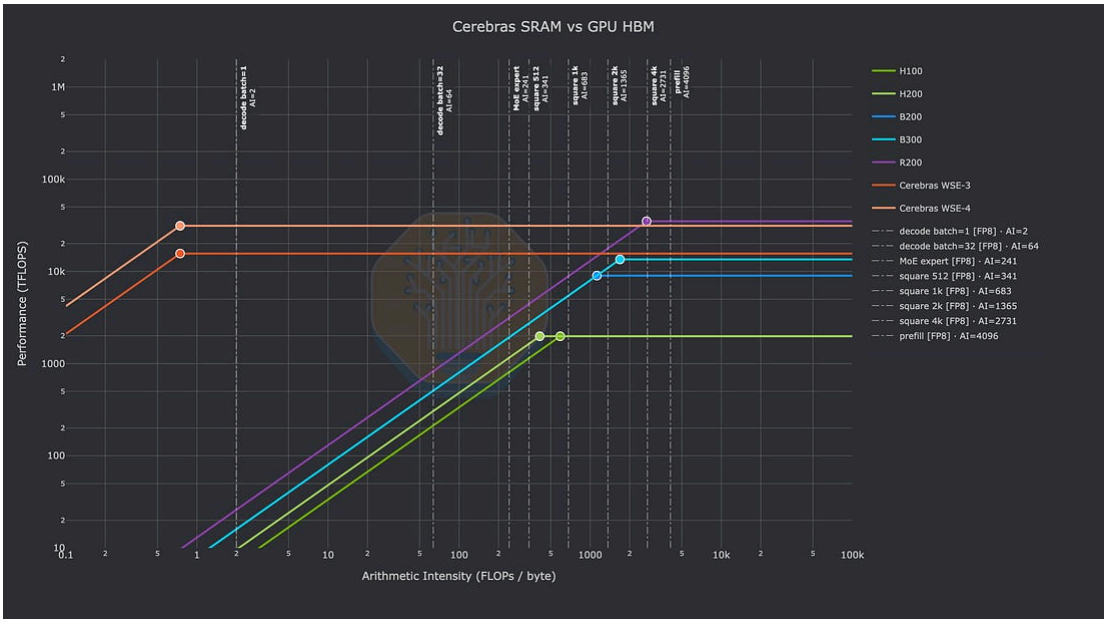
Source: SemiAnalysis

The CS-4 uses the same 5nm WSE-3 as the CS-3, but Cerebras is extracting double the performance by doubling clock speeds. This comes from feeding dramatically more power to the wafer, and is enabled by the CS-4's improvement in power delivery and cooling technology. While staying on the same 5nm silicon sounds underwhelming, Cerebras can still double the metric that matters most: memory bandwidth. The doubling in memory bandwidth should translate into a near doubling of tokens/sec/user all else equal, and this is what customers want from Cerebras. Clock speed doubling also drives double the peak theoretical FLOPs and the WSE's parallel off-wafer I/O, allowing the CS-4 to upgrade to 2.4Tb/s of off-wafer I/O from 1.2Tb/s with CS-3. However, what remains the same is 44GB of SRAM capacity per

QQ:1413221435

wafer, as this is determined by the number of SRAM bit cells available on each wafer. so we'll have to wait for the next generation silicon before we can see any improvement here. This is the main drawback of re-using the same WSE-3 as the low memory capacity per wafer is one of the key tradeoffs inherent with Cerebras's architecture.

As we described in our previous article on Cerebras, the wafer has an incredibly unique architecture due to the use of SRAM that makes it well suited for running kernels with low Arithmetic Intensity, such as low-batch size decode.

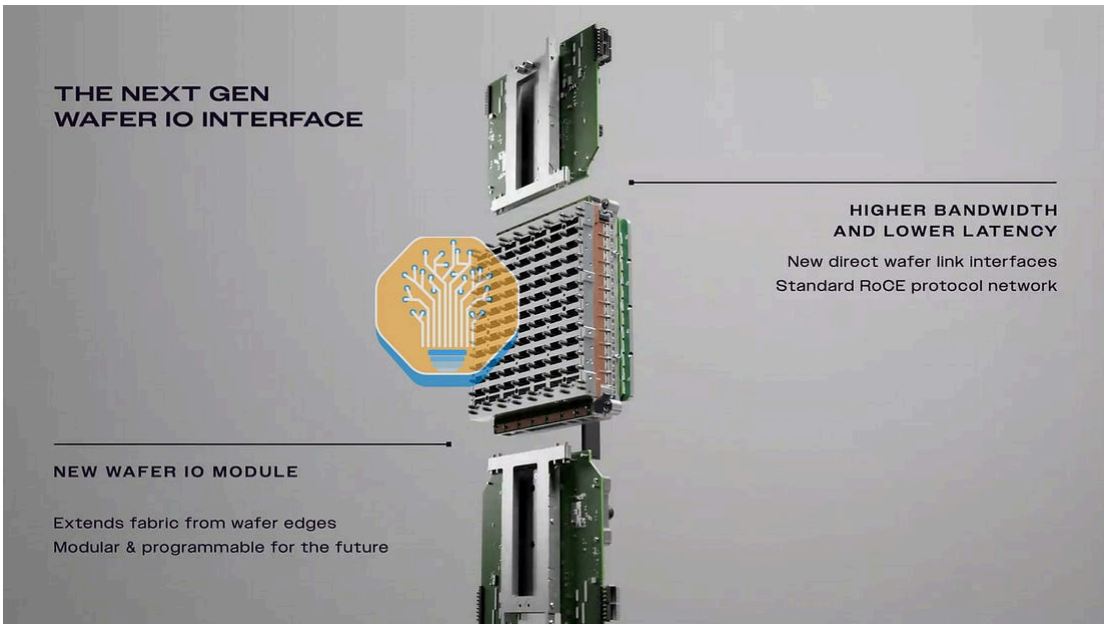


Source: SemiAnalysis

Network Improvements

Beyond the doubling of off-wafer I/O bandwidth, there are further improvements made on off-wafer communications coming from a new Wafer I/O interface, which is an upgraded FPGA card that is used as a NIC to convert Cerebras proprietary I/O to standard ethernet. From the image below, we can see that there are 2 I/O modules coming off the north and south of the wafer.

QQ:1413221435
QQ:1413221435



Source: Cerebras

QQ:1413221435

This I/O module is field-upgradeable, so Cerebras can move to new networking standards without redesigning the chassis. This seems like a small change but there are big implications. This makes it easier for the CS-4 to interface with other systems for disaggregated inference setups. Pairing the CS-4 with HBM-based XPUs in a disaggregated attention feed-forward network setup is one way to overcome the CS-4's low memory capacity, in the same way that Nvidia is positioning the Groq LPUs. We write more about this below. This seems to be designed especially for AWS in mind, who would like to have its EFA NICs on CS-4 to interface with Trainium servers for disaggregated inference.

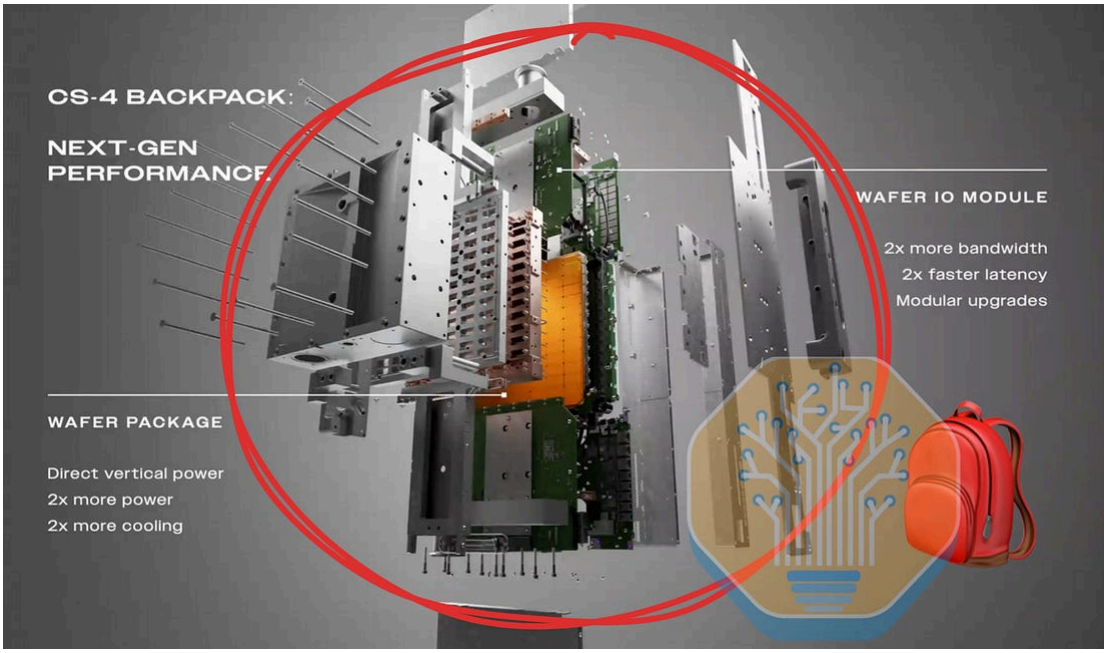
In addition, latency through the 2 layer fat-tree network (using Arista ethernet switches) is reduced to 3 microseconds from 5 microseconds for CS-3 through a new low latency package processing pipeline. Now, direct wafer to wafer links are also possible rather than going through the switched network which further reduces latency to 2 microseconds. The direct wafer paths are also configurable so this means that the FPGA has switching capability to route data through various wafers.

This is a real improvement, but with many Cerebras competitors now quoting all in switch latencies in nanoseconds, "ultrafast" networking is relative and we view it as a modest improvement. We believe this 3µs and bandwidth limitations continues to be a bottleneck that prevents parallelism setups such as EP and ETP where expert layers span multiple wafers. Token dispatch and combine from router to expert is latency sensitive, and the expert imbalance problem coupled with an extra network hop makes pipeline parallelism the only viable solution.

QQ:1413221435

The Backpack Rack

At the system level, the headline change in CS-4 is physical. Cerebras has split the rack into a front half dedicated to power delivery and a rear half dedicated to compute, packaged as modular, pluggable “backpacks.” Each backpack houses a single wafer-scale engine, and a CS-4 rack holds three of them, up from two wafers per rack in CS-3. The cooling infrastructure such as the pump and the heat exchangers are also removed from the rack, as data centers nowadays are built to support fully liquid cooled racks.

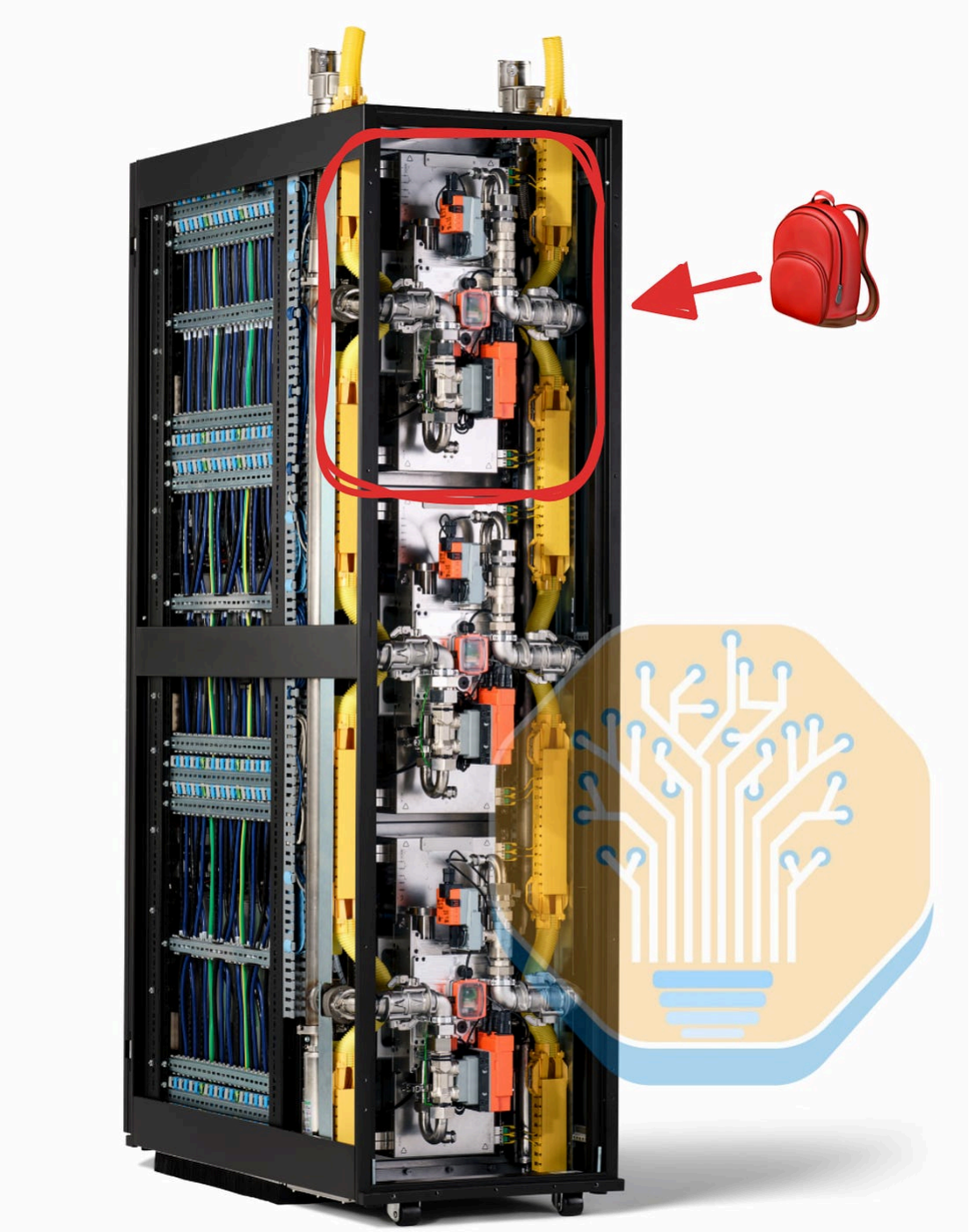


Source: Cerebras

The backpack is a vertical enclosure that is split into power, cooling, I/O, and WSE-3 modules independently, which makes the whole system meaningfully simpler to manufacture than CS-3. The WSE-3 wafer sits vertically with the power side facing the front of the rack. The power modules will deliver power via the front side of the wafer, which is the same as CS-3. The cooling of the wafer will be approached from the rear side of the wafer. The I/O modules are attached to the top and bottom edges of the wafer forming a rectangular surface with the wafer.

The backpack design allows a smoother deployment process. Customers can set up the rack with the power modules before simply socketing in the the wafer backpack onto the rack on site. Given the upgrade to three wafer engines per rack running at higher clock speed. One CS-4 rack lands at 125-135kW TDP, which is up around or just short of double the 23kW power draw of a single CS-3. Overall, this means performance/W has at best a slight improvement over CS-3.

QQ:1413221435
QQ:1413221435



Source: Cerebras

The big improvement gen-on-gen should be cost. While more cooling and power alone would bring up the BOM, the CS-4's reduction in components and simpler assembly should be a significant offset to these cost items, and we think the effective BOM per wafer could end up similar to the CS-3. To customers, this means getting nearly double the interactivity and token revenue but at similar TCO which is a very attractive proposition.

Cerebras's favorite number for CS-4 is 43 PB/s of total on-chip memory bandwidth, which the company markets as roughly 2,000x more memory bandwidth than

QQ:1413221435

Nvidia’s Rubin. That’s the number that will lead most coverage of this launch, since on-wafer SRAM bandwidth scales up with the increase in power.

The result is that in spite of 2,000x more memory bandwidth, Cerebras claims a more reasonable interactivity improvement of up to 30x when compared to GPUs, which it’s branding as a new “ultrafast” performance tier. We believe CS-4 will hit near 4,000 tok/sec/user on frontier models, while CS-3 hits 2,000 tok/sec/user. Meanwhile we expect Blackwell GPUs will continue to top out at a theoretical 200 tok/sec/user (which no one actually runs at), and a more realistic 100 tok/sec/user for reasonable amounts of concurrency. That looks like 20-40x more interactivity to us, so why not “up to” 40x faster? Seems fair enough.

Parallelism Strategies

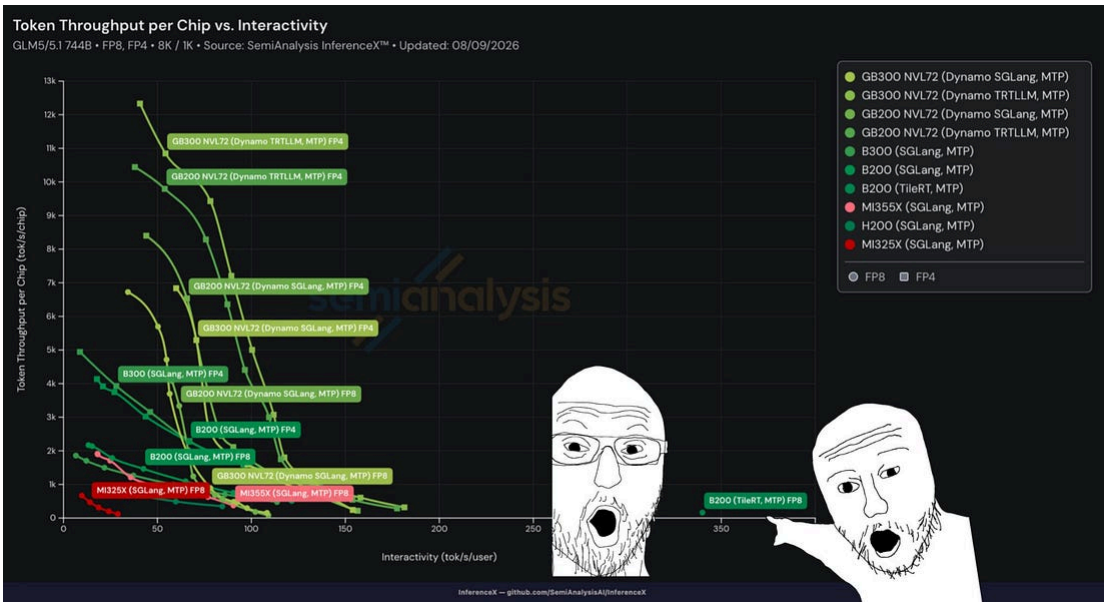
QQ:1413221435

Because a WSE does not have enough SRAM on wafer to hold an entire model’s weights, Cerebras continues to focus their efforts on pipeline parallel inference with this system. On CS-4, every MoE expert for a given model will sit on a single wafer, interleaved. Pipeline parallelism by default is different than GPUs, where tensor and expert parallelism are most common to get big models to fit in the available HBM. Cerebras has always maintained that using the GPU’s HBM to store weights is slower, more power hungry, and more expensive than doing everything on one wafer.

Of course, when comparing a cluster of WSEs to a cluster of GPUs on performance, power consumption, and cost, it really depends what parallelism strategy you choose. GPUs have a very wide range of configuration options (from high throughput/low interactivity configs to high interactivity/low throughput) while the wafer’s range is more modest. Only high interactivity/low throughput is considered.

In order to compare WSEs directly to GPUs, we are particularly interested in NVIDIA’s release of TileRT, which brings high throughput/low interactivity configs to GPU clusters. We discussed some of this in our TileRT article last week

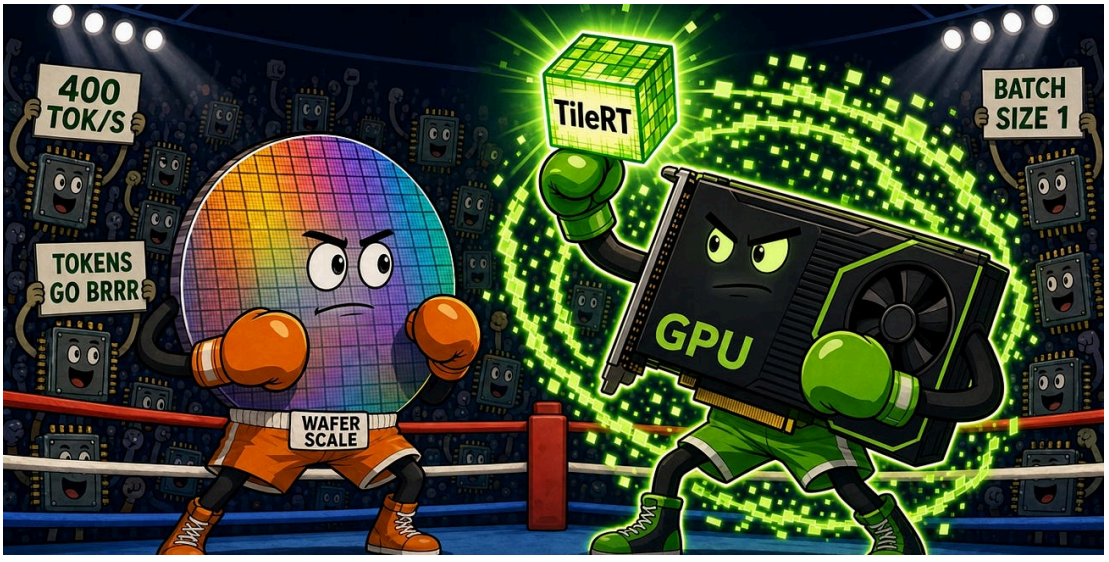
QQ:1413221435



Source: InferenceX

Ultra-High Interactivity on NVIDIA GPUs? - TileRT InferenceX

BRYAN SHAN, DANIEL NISHBALL, AND 4 OTHERS · AUG 10



Premium-priced “fast modes” are proving that users will pay more for lower latency and faster tokens, potentially yielding higher gross margins. Frontier AI labs such as OpenAI are therefore evaluating purpose-built inference systems, including Cerebras and NVIDIA Groq LPUs

Read full story

Because the wafer itself is 44GB, Cerebras hasn’t yet needed to shard individual experts in a MoE model across multiple wafers, even for frontier models that they run

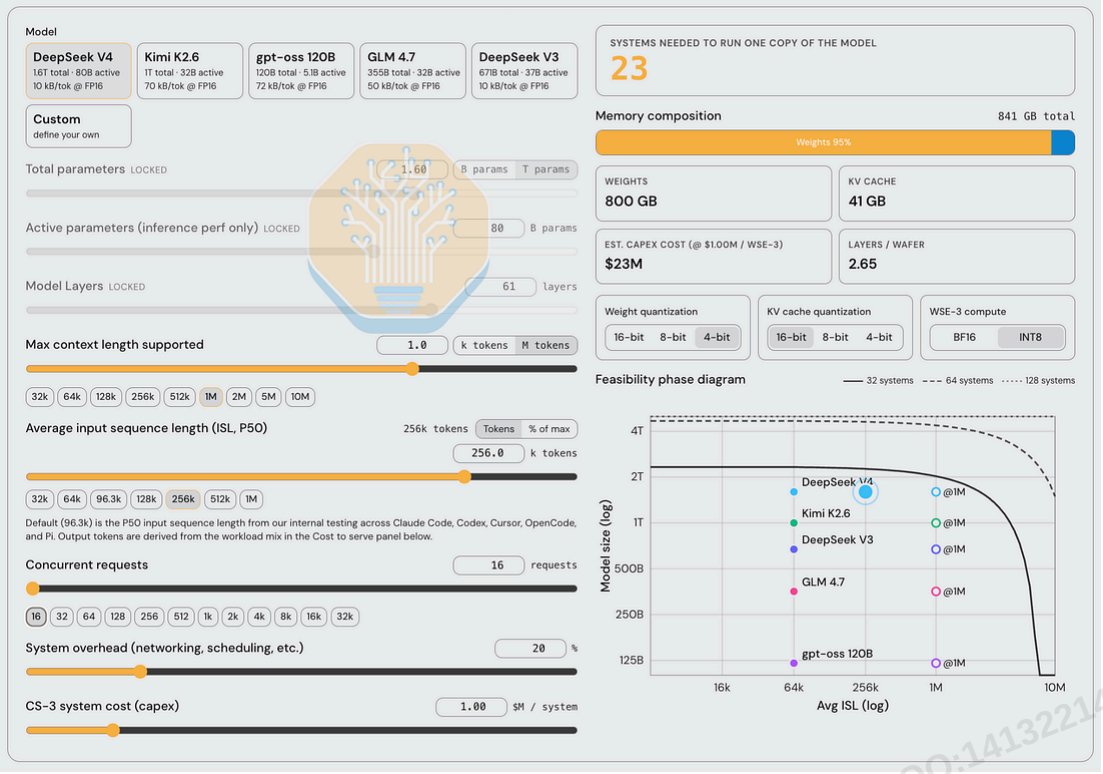
QQ:1413221435

today such as GPT 5.6 Sol. However, due to the demands of long context inference, we expect Cerebras customers such as OpenAI to try and save on the cost of keeping KV Cache on-wafer for long context workloads, and run 5.6 Sol at 256k context window, rather than the full 1M context.

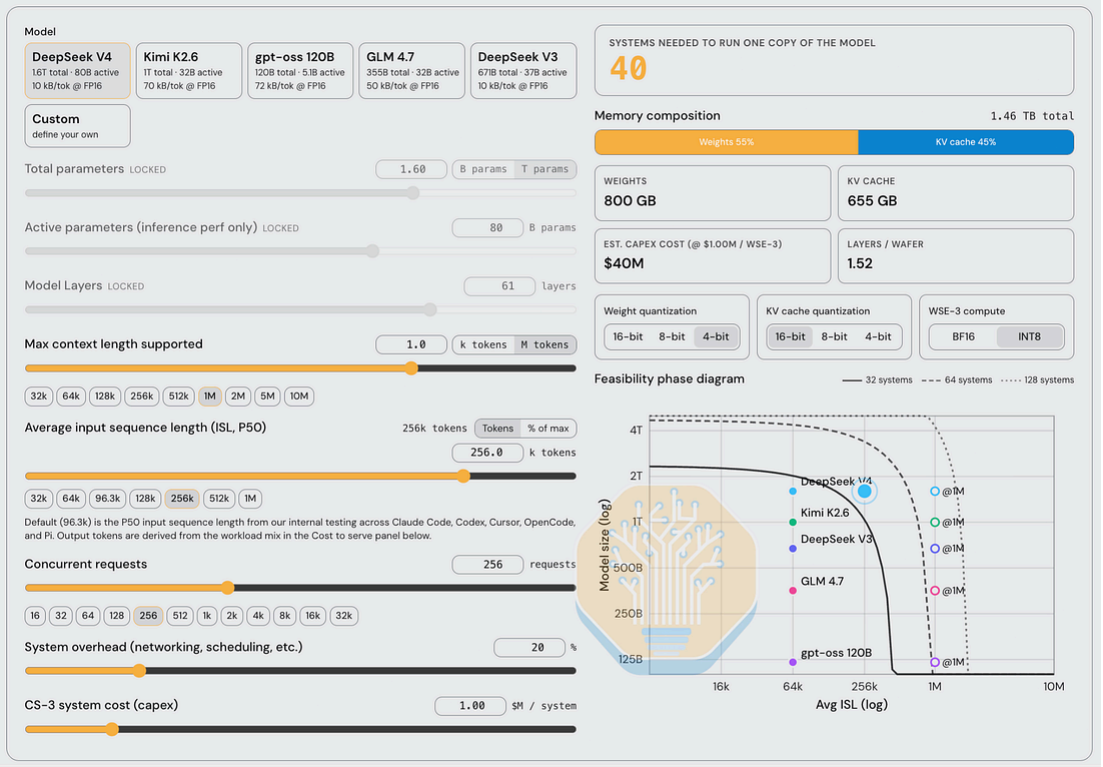
As we described in our Cerebras IPO article, the cost of supporting long context inference is massive. Most people still understand that the amount of memory required to get one forward pass from a model is proportional to the total amount of parameters in the model. However, it still seems to be a well kept secret that the amount of memory required to hold KV Cache grows proportionally to the number of concurrent users and the average/max size of those users requests (which is dictated by the context window of the model). Running a model with a large context window, and supporting many concurrent users, requires lots of memory capacity.

When we run a simple analysis of what it would take to run a large model (say, the 1.6T parameter DeepSeek V4 Pro), we find that the minimum number of Cerebras WSE's needed to run this model at 1M ctx is around 20 systems, and at a reasonable concurrency of 256 requests, its around 40 systems. That's over $20M of CAPEX and 1MW of power consumption before you can get a forward pass on a frontier model.

This analysis is available on our public tokenomics website:



QQ:1413221435



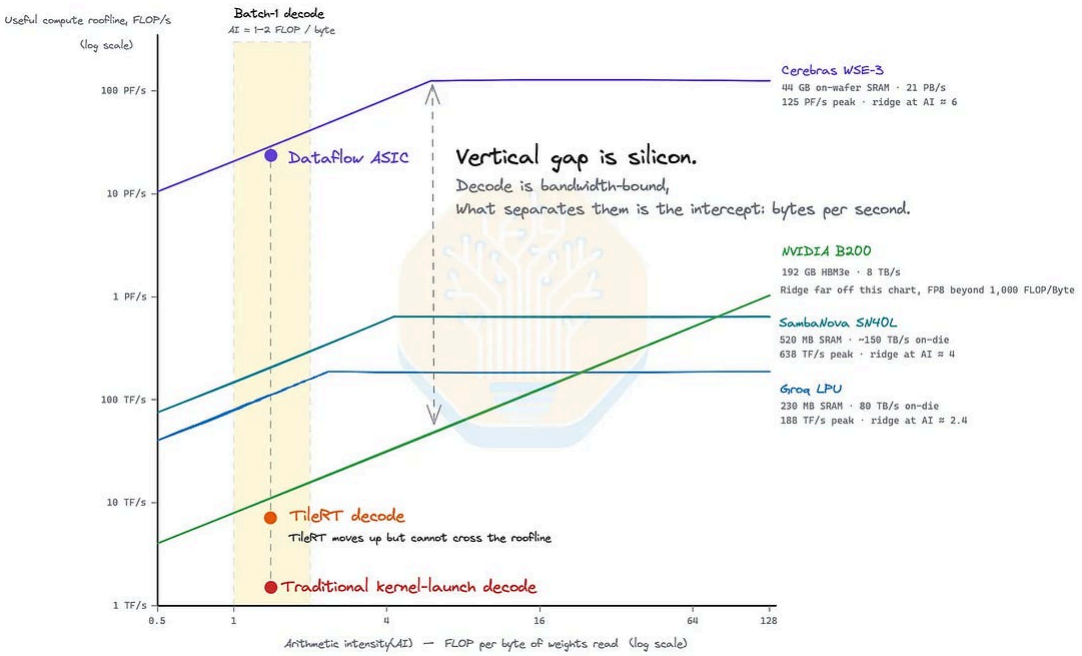
Source: https://www.tokenomics.info/cerebras

Of course, after this up front investment comes some pretty remarkable aggregate performance metrics. Lets dig into that.

Betting on Disaggregation

Cerebras is positioning CS-4 as being built for open, heterogenous, disaggregated inference from the ground up. They are currently working with AMD and AWS Trainium as partners, but have more coming. In all these configurations, Cerebras will serve as the decode chip, since its rooflines aren’t optimal for compute-bound prefill. The company believes in all disagg setups, and claims support for attention feedforward disaggregation (AFD) in addition to the traditional prefill decode disaggregation (PDD).

QQ:1413221435
QQ:1413221435



Source: SemiAnalysis

QQ:1413221435

Software approaches trying to replicate dataflow execution on GPUs can approach the HBM roofline, but cannot compare to SRAM-only architectures on bandwidth. However, all heterogeneous disagg setups are double-edged, since the ratio of Prefill to Decode resources in your cluster is fixed the day that the hardware PO is signed. Meanwhile a fleet of GPUs or TPUs can be dynamically allocated into different ratios as workload profiles from users shift over time. And in the real world, workloads do shift over time. We saw reasoning models drive decode costs up as models thought longer, and then agentic cache hits drive prefill costs down while decode costs remained the same. One P:D ratio to rule them all is unlikely to be perfectly optimal for the 5+ year lifespan of these systems.

To get a better understanding of these dynamics, and the thought process that infrastructure teams are using to design high performance inference clusters, read our TileRT article.



Ultra-High Interactivity on NVIDIA GPUs? - TileRT InferenceX

BRYAN SHAN, DANIEL NISHBALL, AND 4 OTHERS · AUG 10

Read full story

Roadmap: Nexus and CS-5

Cerebras has already co-designed the next-generation wafer-scale engine alongside a new rack-scale platform “Nexus” meaning the CS-4 rack chassis will likely carry forward into future chip generations without major mechanical redesign. The

QQ:1413221435

company’s stated roadmap commitment is roughly 2x faster performance every year, with a specific target of 20x throughput improvement by 2027.

On the reliability side, Cerebras is addressing the operational difficulty of swapping wafer trays in the field, and continues investment in error recovery, tray-level redundancy, and yield harvesting across cores and channels on the chip itself.

Overall, we are impressed by this announcement. Cerebras has made improvements on performance, manufacturability, and deployability as the company continues to scale its own production and deployments. While Cerebras’s weakness in the form of low memory capacity was not addressed directly this generation, greater support for disaggregation addresses this indirectly. Double the bandwidth can double customers’ token revenue and result in a 2x perf/TCO improvement. However, Cerebras is still a very expensive solution that is betting on customers’ willingness to pay significantly more for fast tokens. We hope that Cerebras can iterate further to deliver higher concurrency, so that Cerebras can drive costs down further and bring these ultrafast tokens to the mass market.

QQ:1413221435

Like Comment Restack

© 2026 Dylan Patel
548 Market Street PMB 72296, San Francisco, CA 94104
Unsubscribe

Start writing

QQ:1413221435