Wafer-Scale AI:
GPU Impossible Performance

Sean Lie, Co-founder and CTO, Cerebras Systems

Hot Chips 2024

Cerebras Systems

Founded in 2016

400 Employees

Offices

Silicon Valley | San Diego | Toronto | Bangalore

Customers

North America | Asia | Europe

ALEPH ALPHA

Argonne NATIONAL LABORATORY

core42

DEPARTMENT OF DEFENSE UNITED STATES OF AMERICA

GSK

lrz

MAYO CLINIC

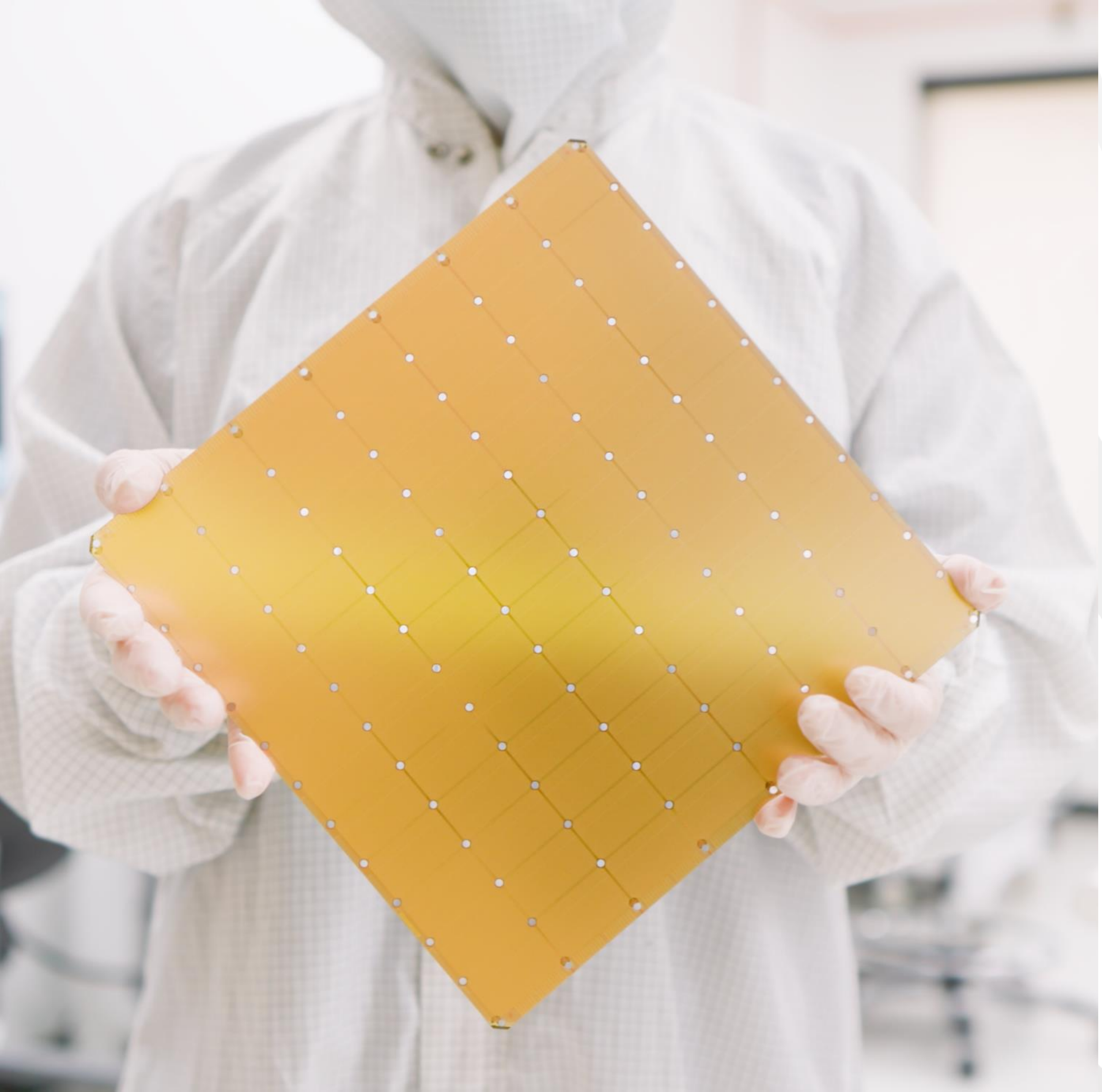
NNSA National Nuclear Security Administration

NETL NATIONAL ENERGY TECHNOLOGY LABORATORY

novo nordisk®

Sandia National Laboratories

TotalEnergies



Cerebras Wafer-Scale Engine

The largest chip ever produced

46,225 mm^2 silicon

4 trillion transistors

900,000 AI cores

125 Petaflops of AI compute

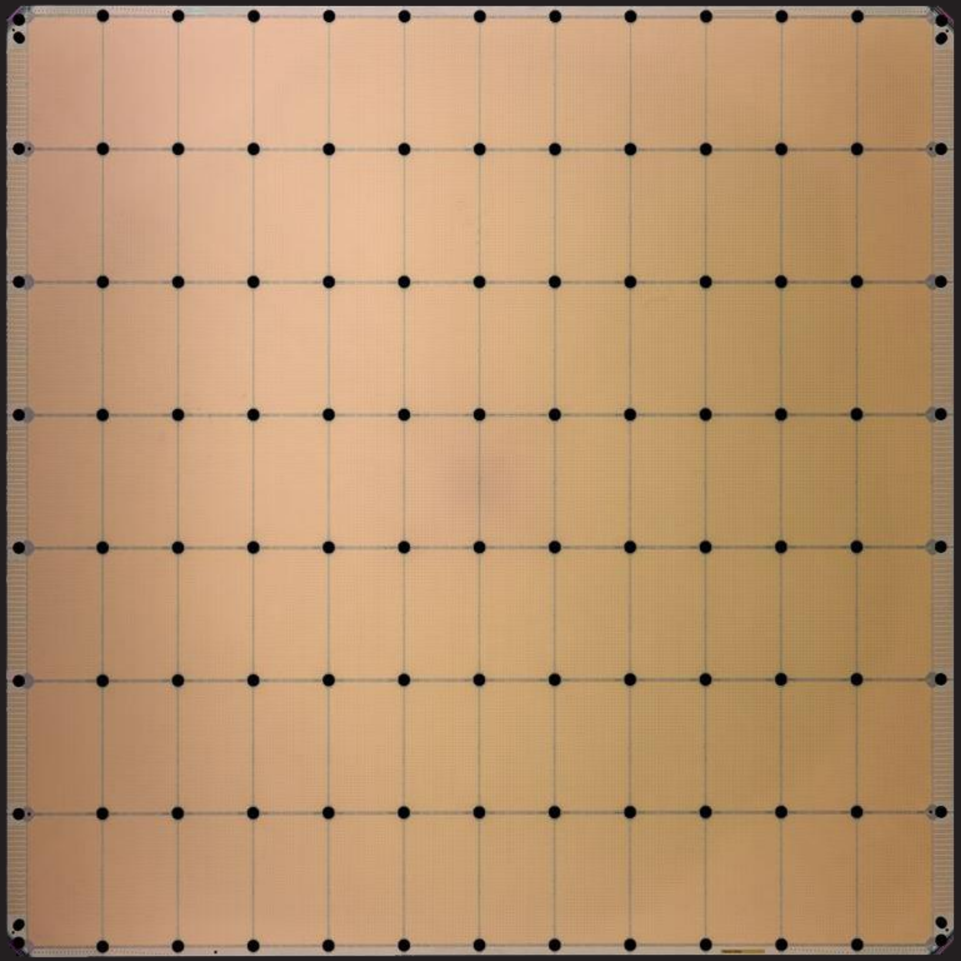
44 Gigabytes of on-chip memory

21 PByte/s memory bandwidth

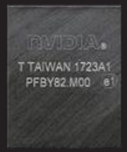
214 Pbit/s fabric bandwidth

5nm TSMC process

Cerebras Wafer-Scale Engine Versus the H100



Cerebras WSE-3
4 Trillion Transistors
46,225 mm^2 Silicon



Largest GPU
80 Billion Transistors
814 mm^2 Silicon

Cerebras CS-3

Condor Galaxy 1 - 4 Exaflops
Santa Clara, California

Condor Galaxy 2 - 4 Exaflops

Stockton, California

Condor Galaxy 3-5 - 20 Exaflops

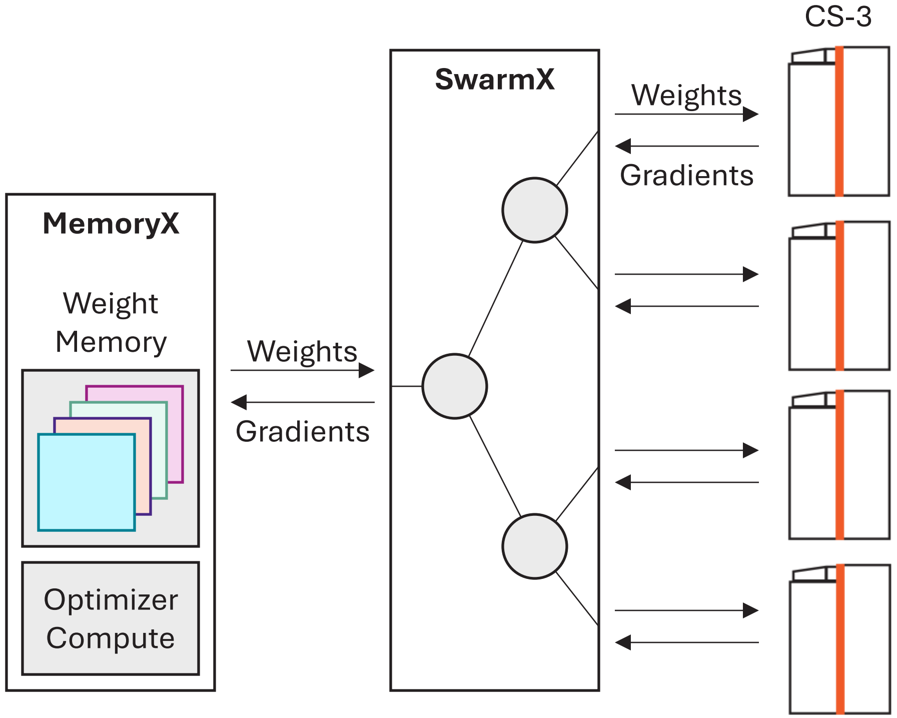
Dallas, Texas

Condor Galaxy 6-9 - 32 Exaflops
Minneapolis, MN

Designed end-to-end for large-scale **Training**

Co-designed cluster architecture to scale using data-parallelism only

- WSE large enough to run even the largest models on a single chip
- Avoid hybrid model parallelism complexity
- MemoryX store streams weights to CS-3s
- SwarmX fabric performs broadcast/reduce
- Multi-system scaling with the same execution model as single system



The only architecture with

Exaflop-scale training performance

But programs like a single device

Training SOTA large models everyday

Sample of open-source models trained on Cerebras

From multi-lingual LLMs to healthcare chatbots to code models

BTLM-3B-8K

3B PARAMETERS • 8K CONTEXT

7B Performance in a 3B Model

Open Source. Trained on Cerebras

CrystalCoder

7B PARAMETERS • 1.3T TOKENS

Coding + English. The most open source & reproducible model in the world.

Open Source. Trained on Cerebras

cerebras

Jais

13B & 30B

State of the art Arabic + English models

Open Weights. Trained on Cerebras

Med42

FINED-TUNED LLAMA2-70B

Medical Q&A LLM Scores 72% on USMLE

Trained on Cerebras

gigaGPT

GPT-3 in 565 LINES OF CODE

Cerebras implementation of nanoGPT

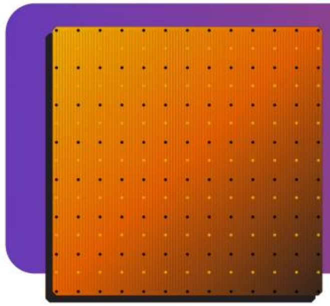
Open Source. Trained on Cerebras

SlimPajama

627BTOKEN DATASET

Extensively deduplicated dataset with twice the perf/token

Open Source



Cerebras-GPT

111M–13B PARAMETERS

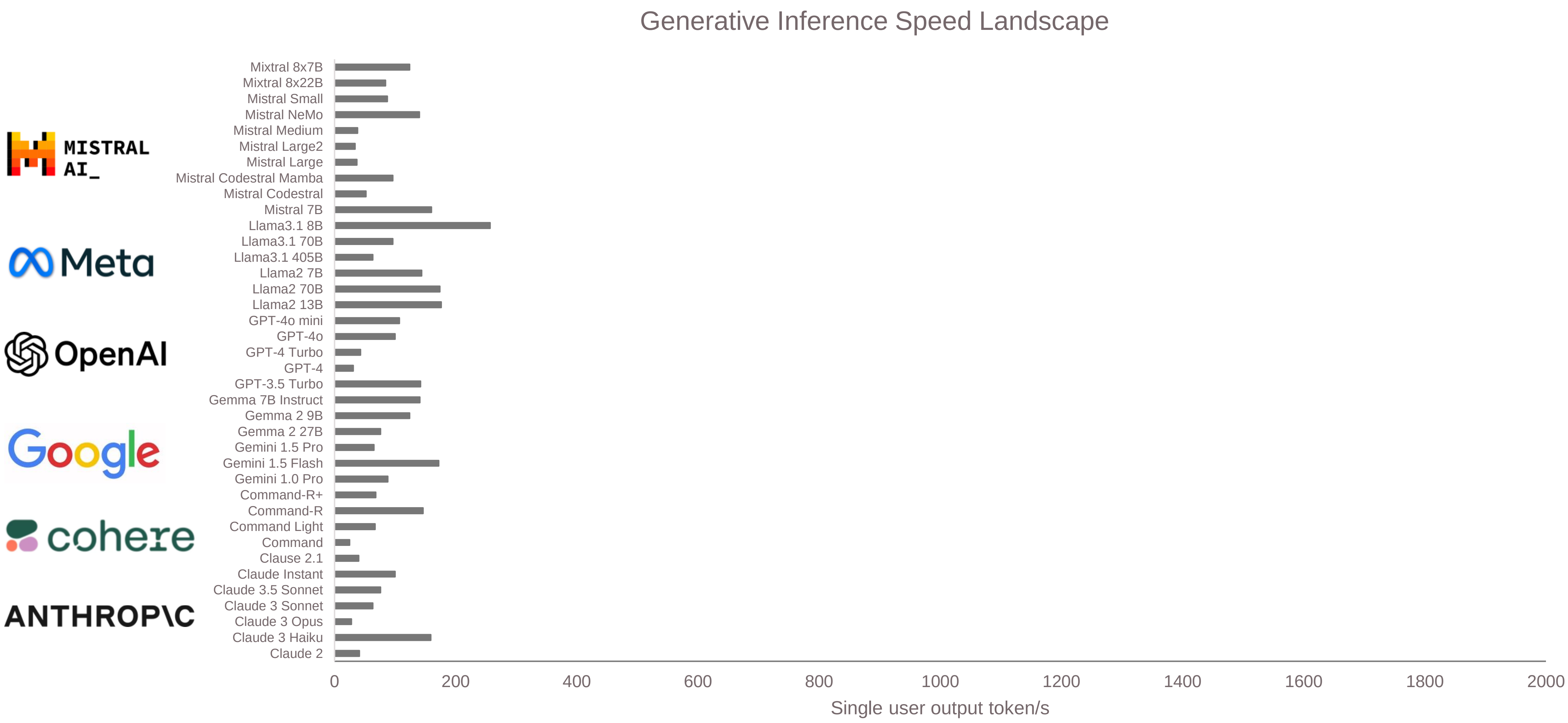
First family of GPT models released under Apache 2.0

Open Source. Trained on Cerebras

But also designed for **inference...**

The Generative Inference Problem

Generative inference today is ***really slow***

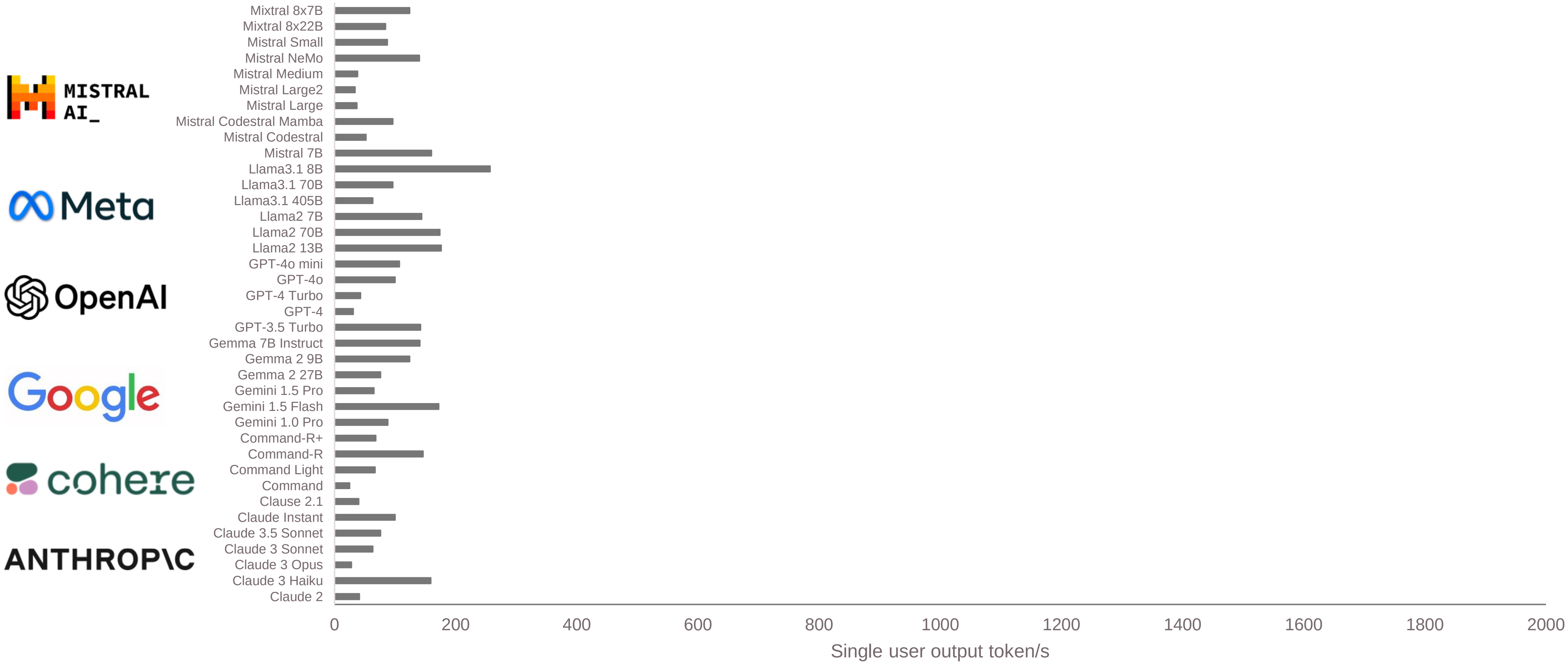


Source: Artificial Analysis https://artificialanalysis.ai/

Generative inference today is ***really slow***

Different ML model architectures
Different hardware: H100, MI300, TPU, ...
Similar performance. Why?

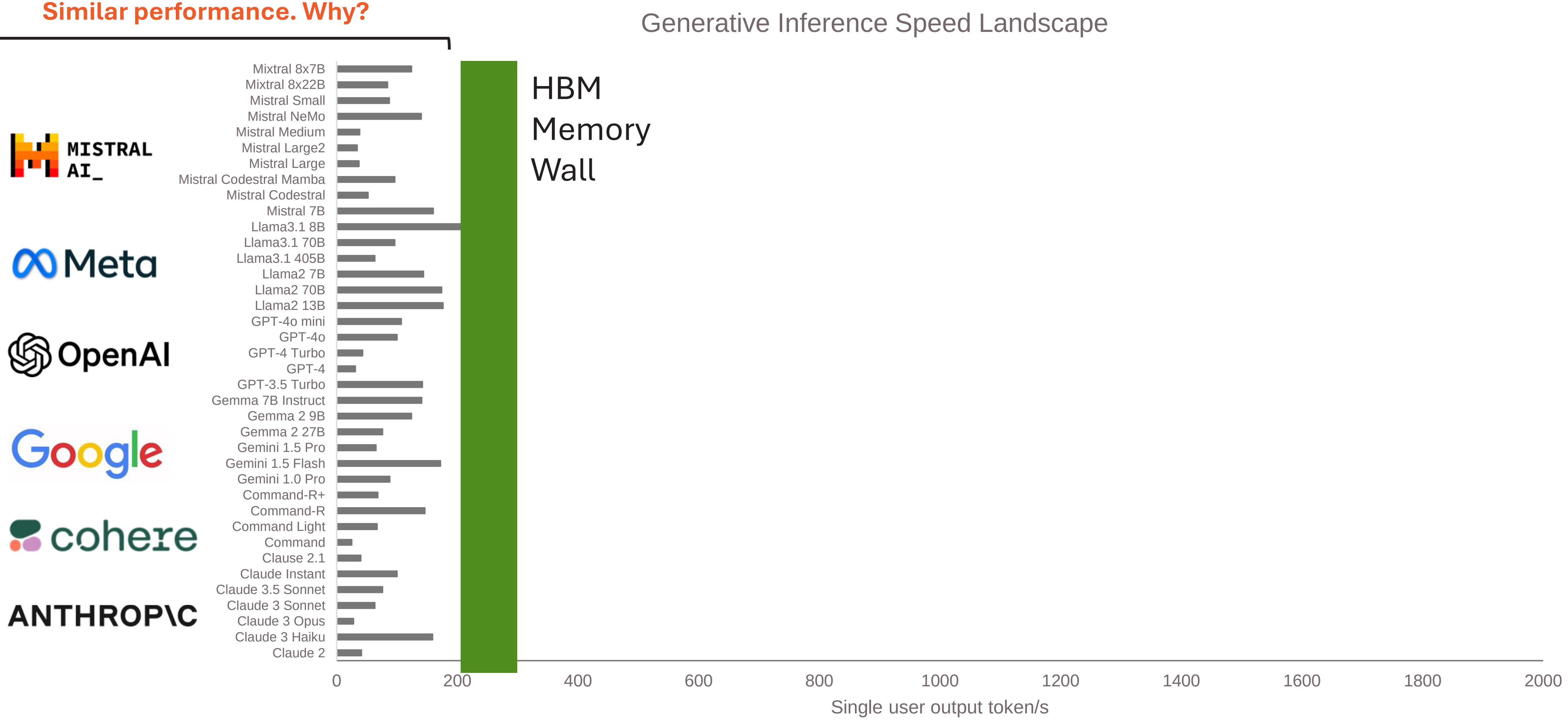
Generative Inference Speed Landscape



Source: Artificial Analysis https://artificialanalysis.ai/

Generative inference today is ***really slow***

Different ML model architectures
Different hardware: H100, MI300, TPU, ...
Similar performance. Why?

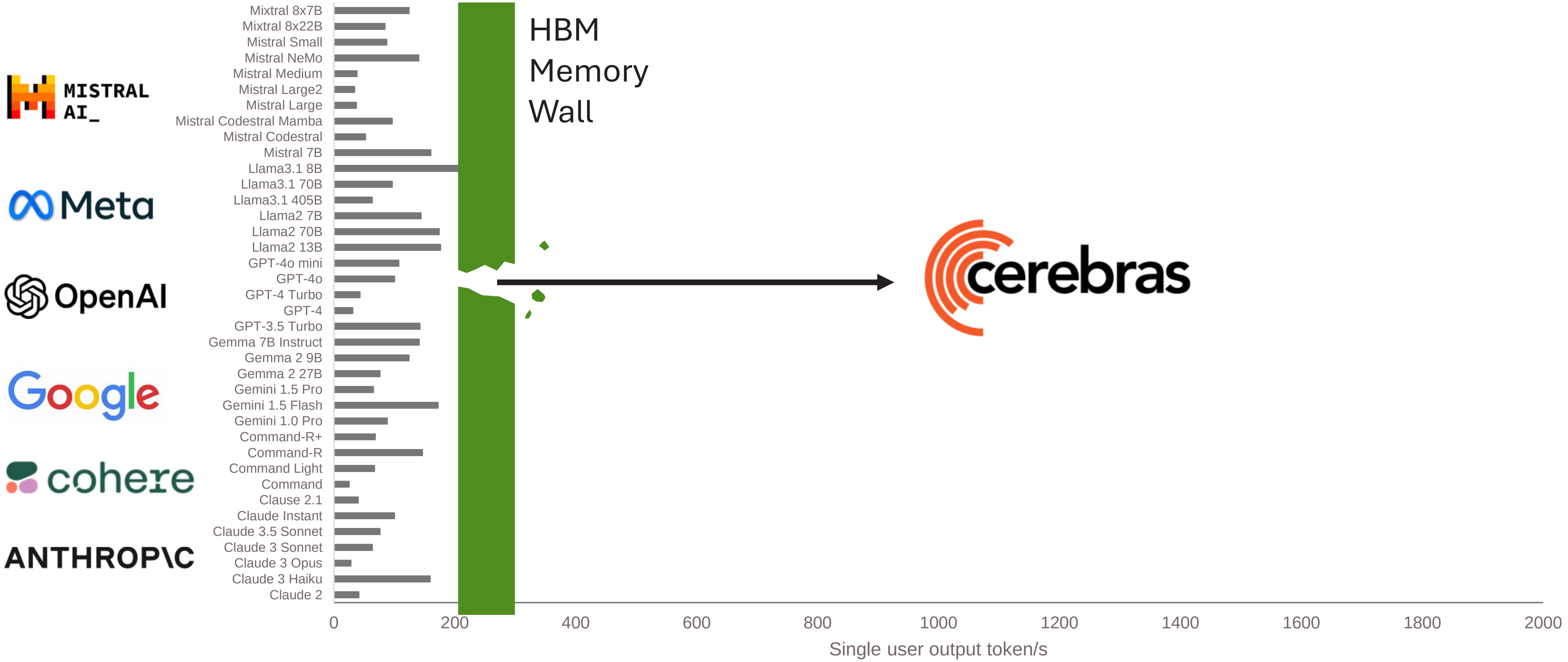


Source: Artificial Analysis https://artificialanalysis.ai/

Generative inference today is ***really slow***

Different ML model architectures
Different hardware: H100, MI300, TPU, ...
Similar performance. Why?

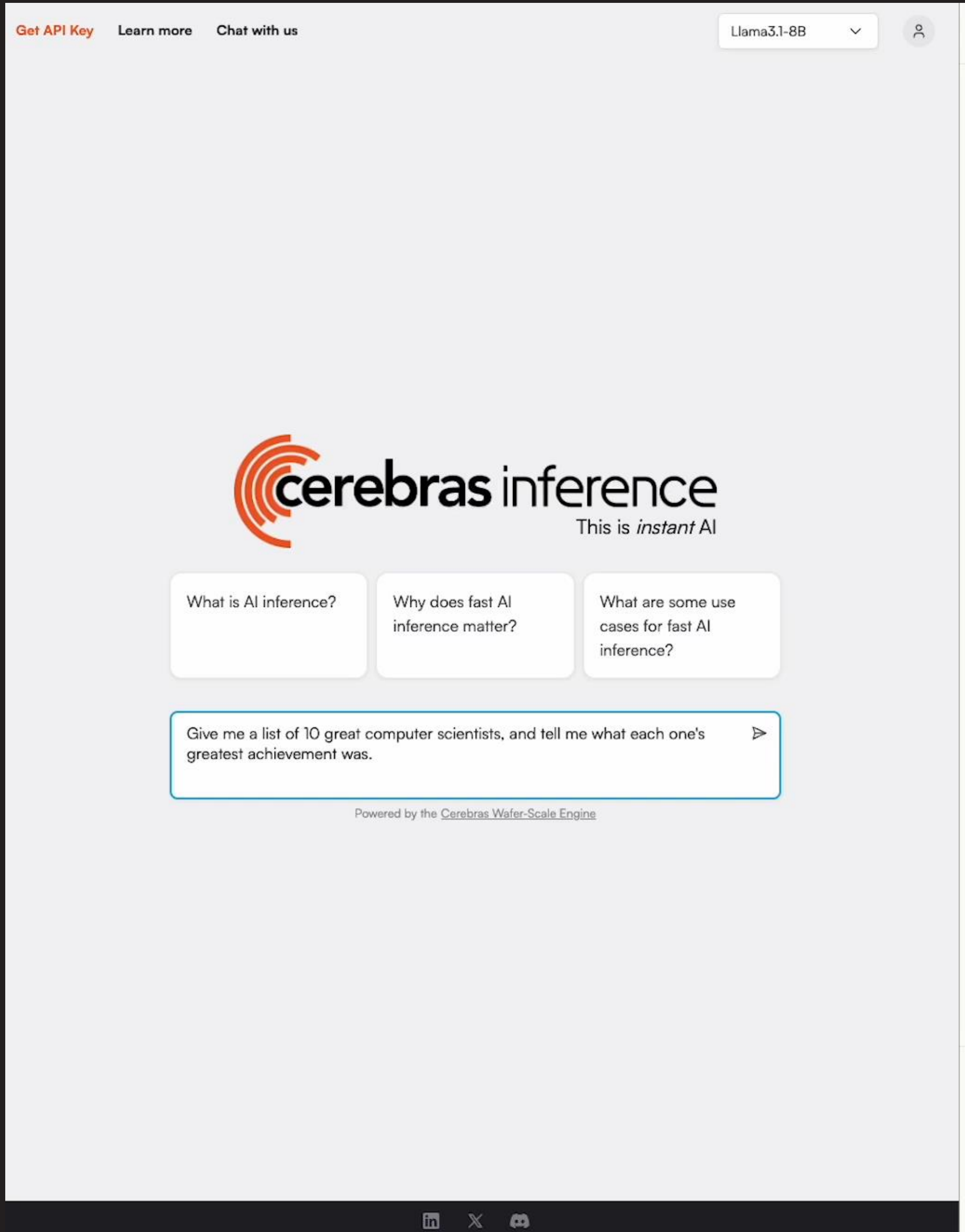
Generative Inference Speed Landscape



Source: Artificial Analysis https://artificialanalysis.ai/

Llama3.1 8B

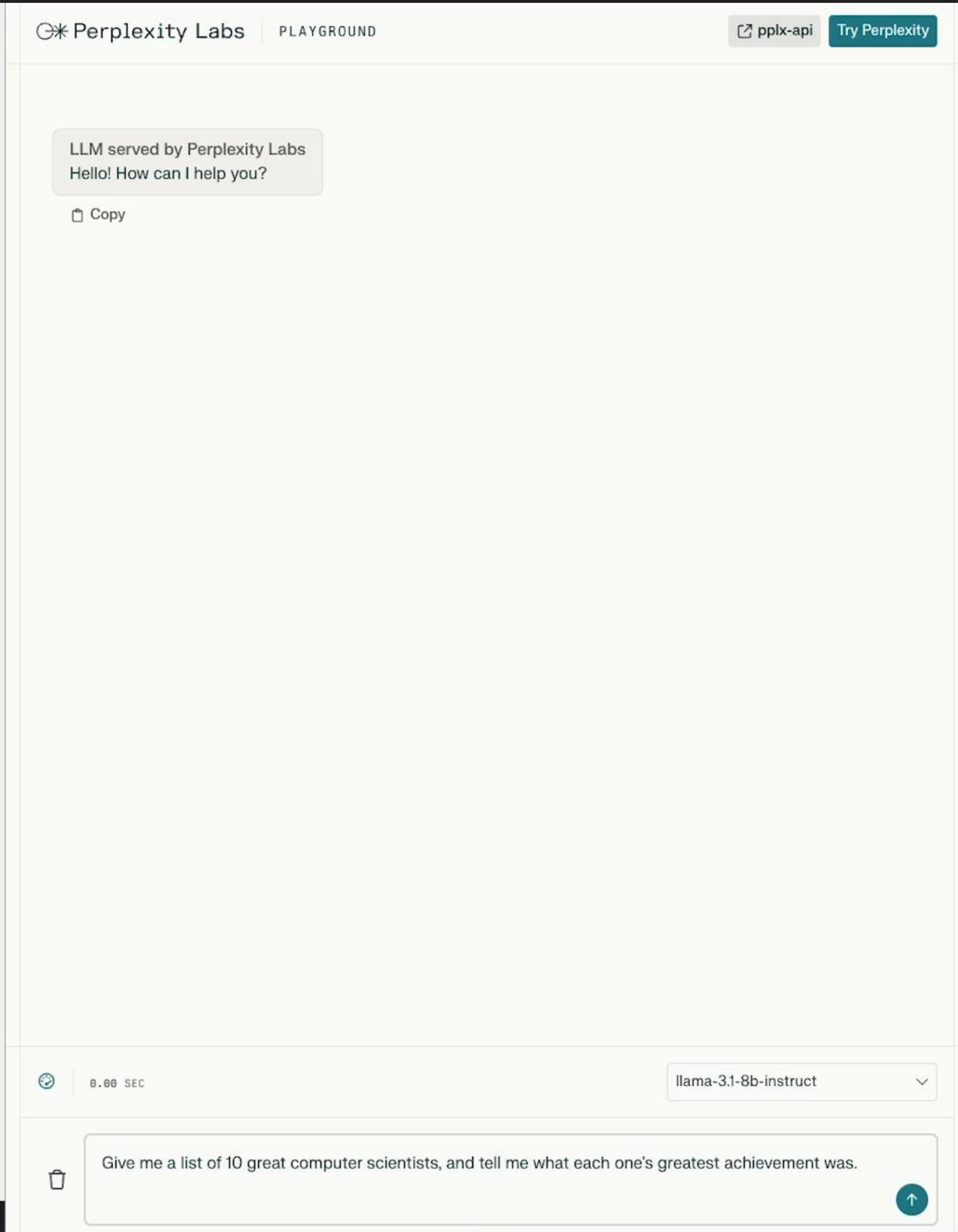
Cerebras CS-3



Give me a list of 10 great computer scientists, and tell me what each one's greatest achievement was.

Powered by the Cerebras Wafer-Scale Engine

DGX-H100



LLM served by Perplexity Labs
Hello! How can I help you?

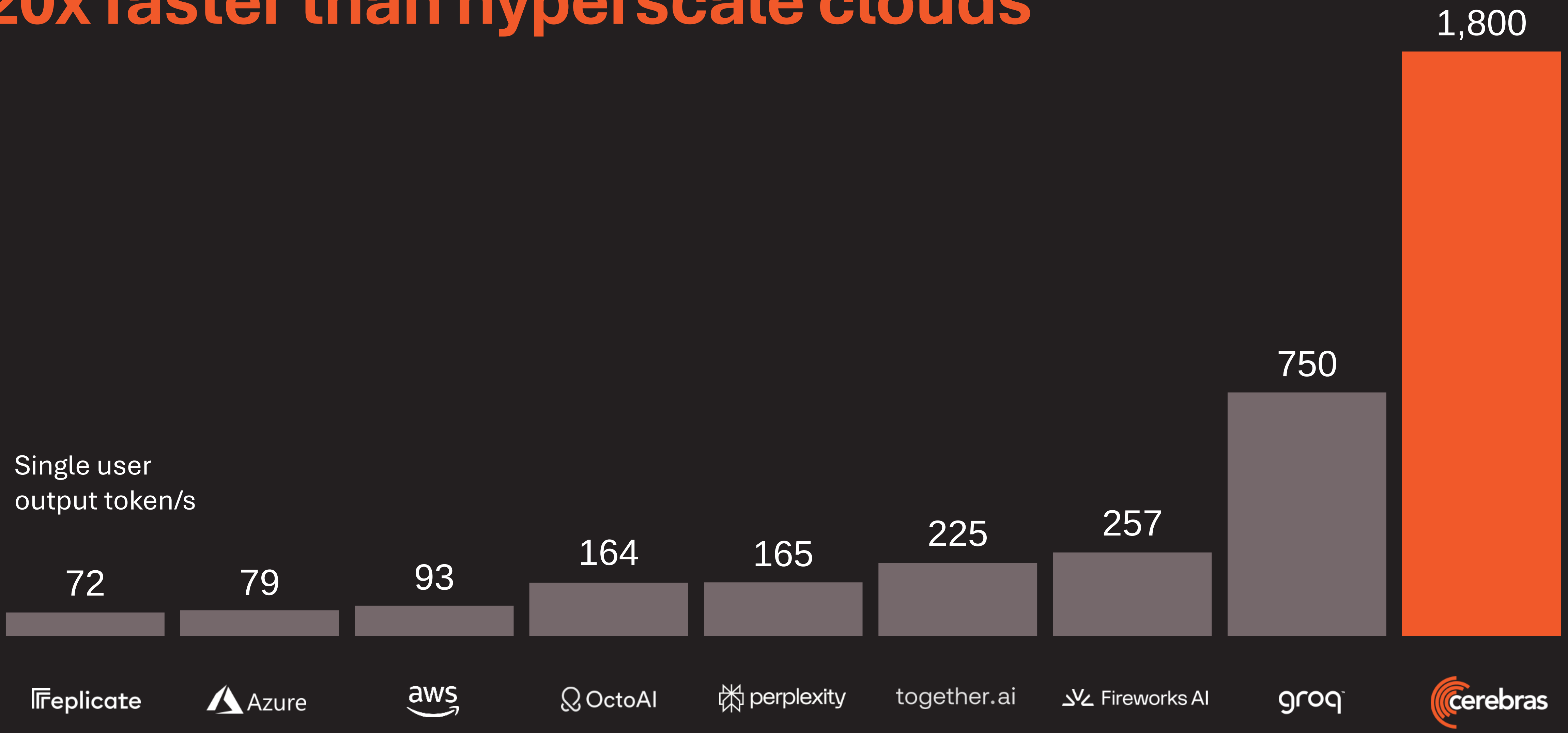
0.00 SEC

llama-3.1-8b-instruct

Give me a list of 10 great computer scientists, and tell me what each one's greatest achievement was.

The fastest inference on the planet on Llama3.1-8B

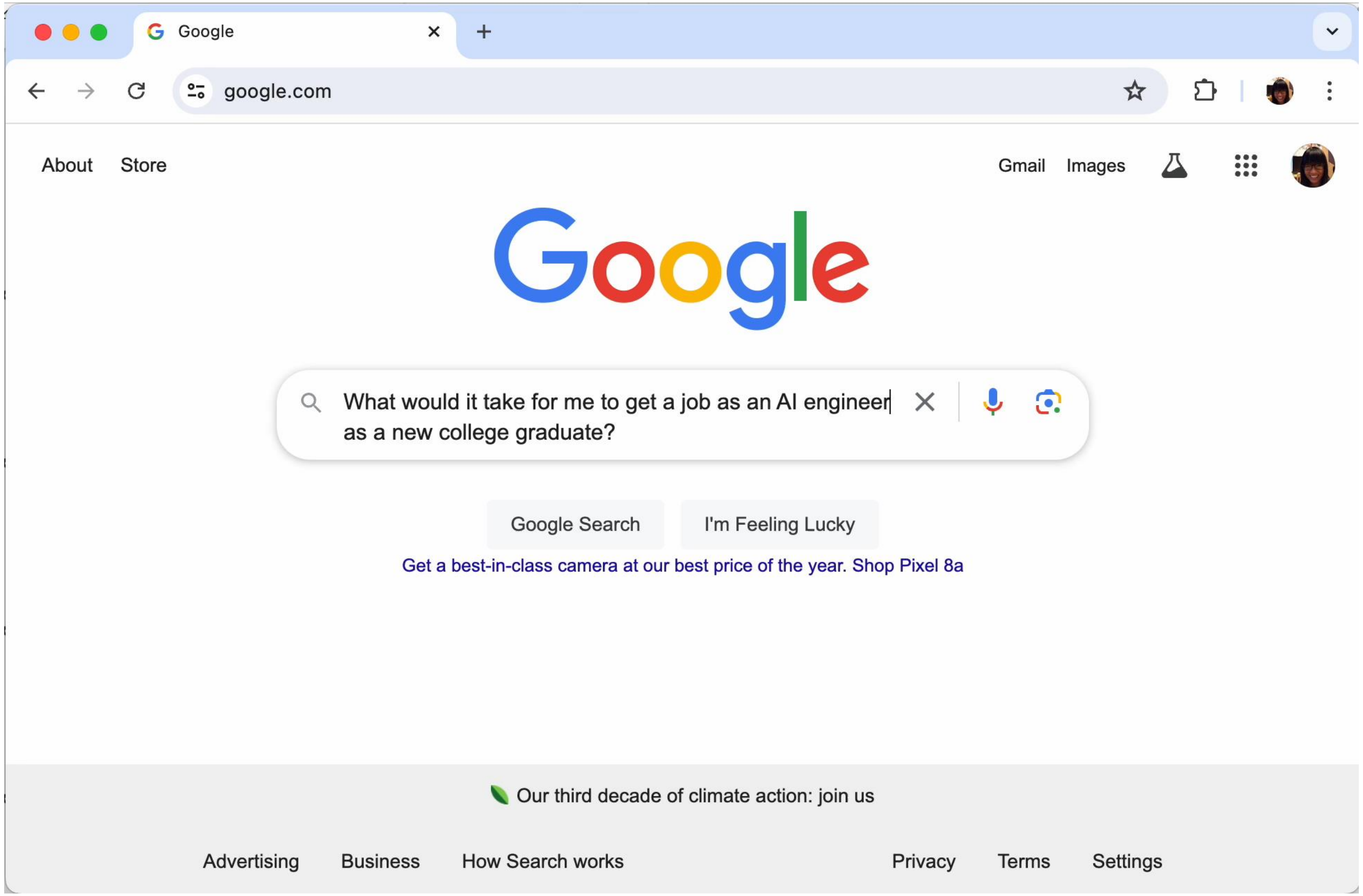
20x faster than hyperscale clouds



Source [Artificial Analysis](), Llama3 & 3.1 Results.

What **20x speed** enables...

Today GenAI applications are **promising but...**



What would it take for me to get a job as an AI engineer as a new college graduate?

Google Search
I'm Feeling Lucky
Get a best-in-class camera at our best price of the year. Shop Pixel 8a
Our third decade of climate action: join us
Advertising
Business
How Search works
Privacy
Terms
Settings

They are slow

Limited user engagement

Still primitive

20x speed eliminates the wait
Unlocks AI agent and copilot capabilities

20x Speed enables

- **20x more** interactive response for higher user engagement
- **20x more** model calls for chain-of-thought reasoning and more accurate responses

Industry trying to move to *agentic* frameworks that need many LLM calls

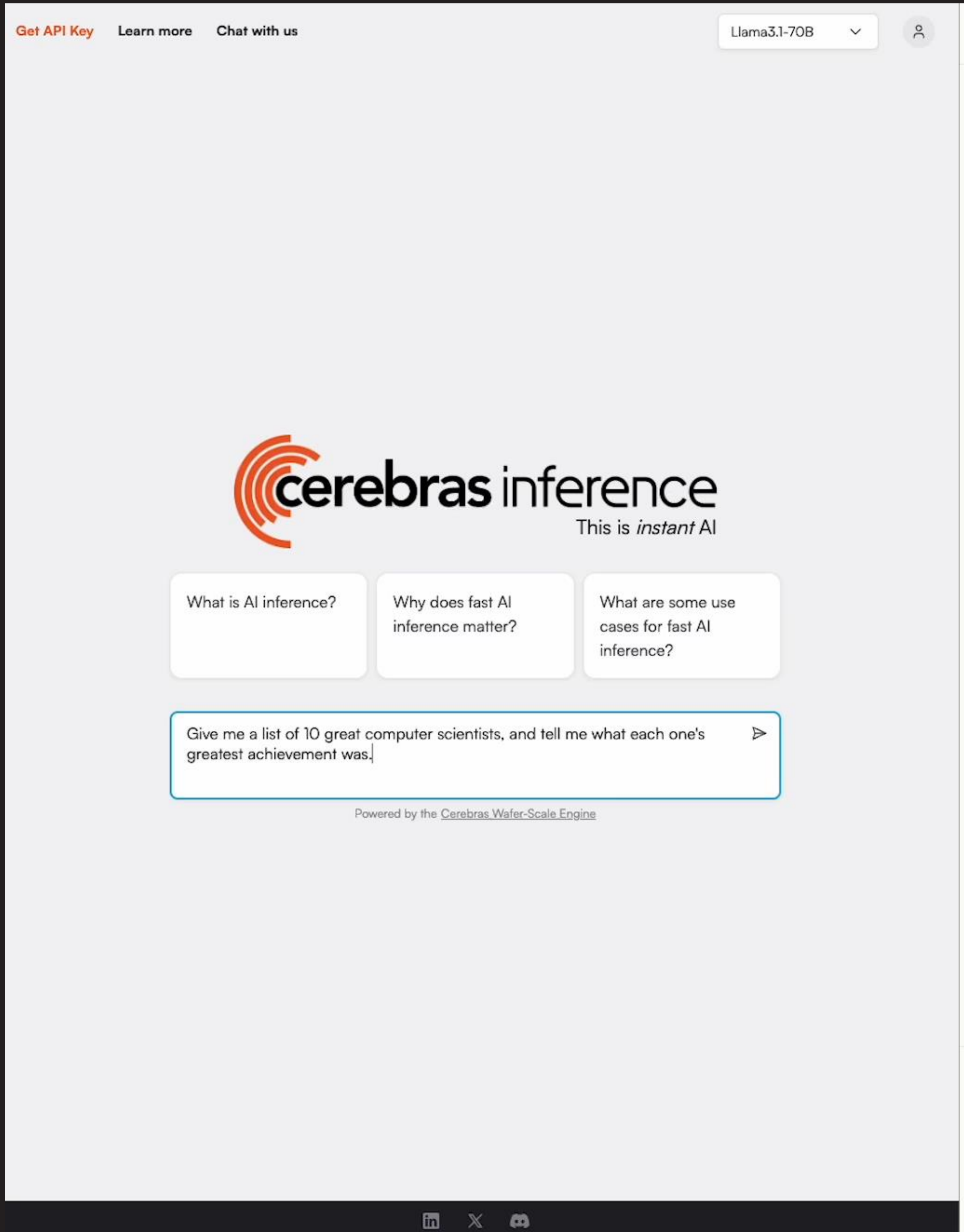
- **GraphRAG (MS Research) :** 5 calls per page, 100s per user request
- **GIST (DeepMind) :** 5 calls per page, several calls per user request
- **ReAct :** 2 calls per turn, 8 calls per user request

This will lead to the most powerful, most sophisticated, most engaging applications

The need for speed is even more evident with larger models

Llama3.1
70B

Cerebras CS-3



Llama3.1-70B

What is AI inference?

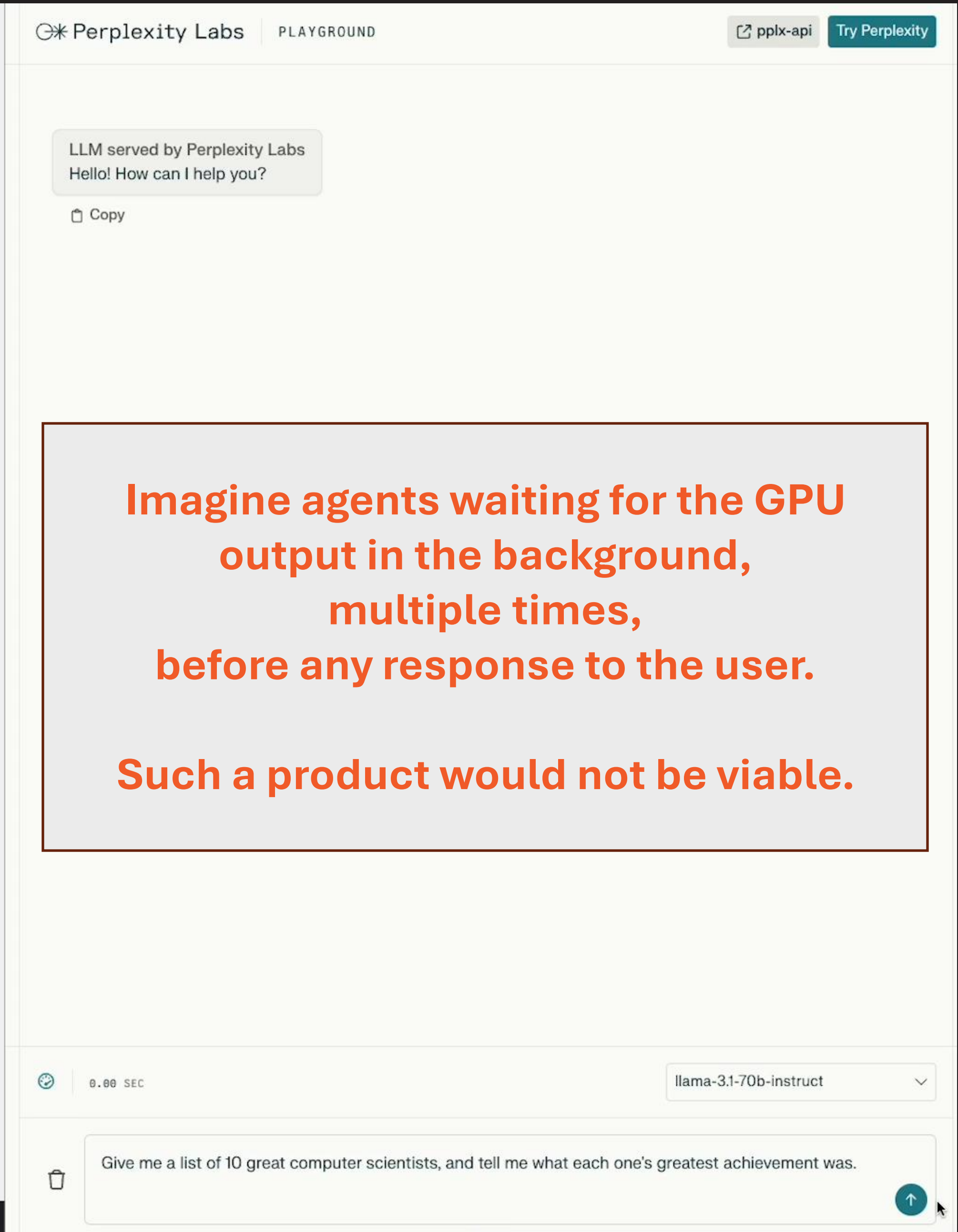
Why does fast AI inference matter?

What are some use cases for fast AI inference?

Give me a list of 10 great computer scientists, and tell me what each one's greatest achievement was.

Powered by the Cerebras Wafer-Scale Engine

DGX-H100



LLM served by Perplexity Labs
Hello! How can I help you?

0.00 SEC

llama-3.1-70b-instruct

Give me a list of 10 great computer scientists, and tell me what each one's greatest achievement was.

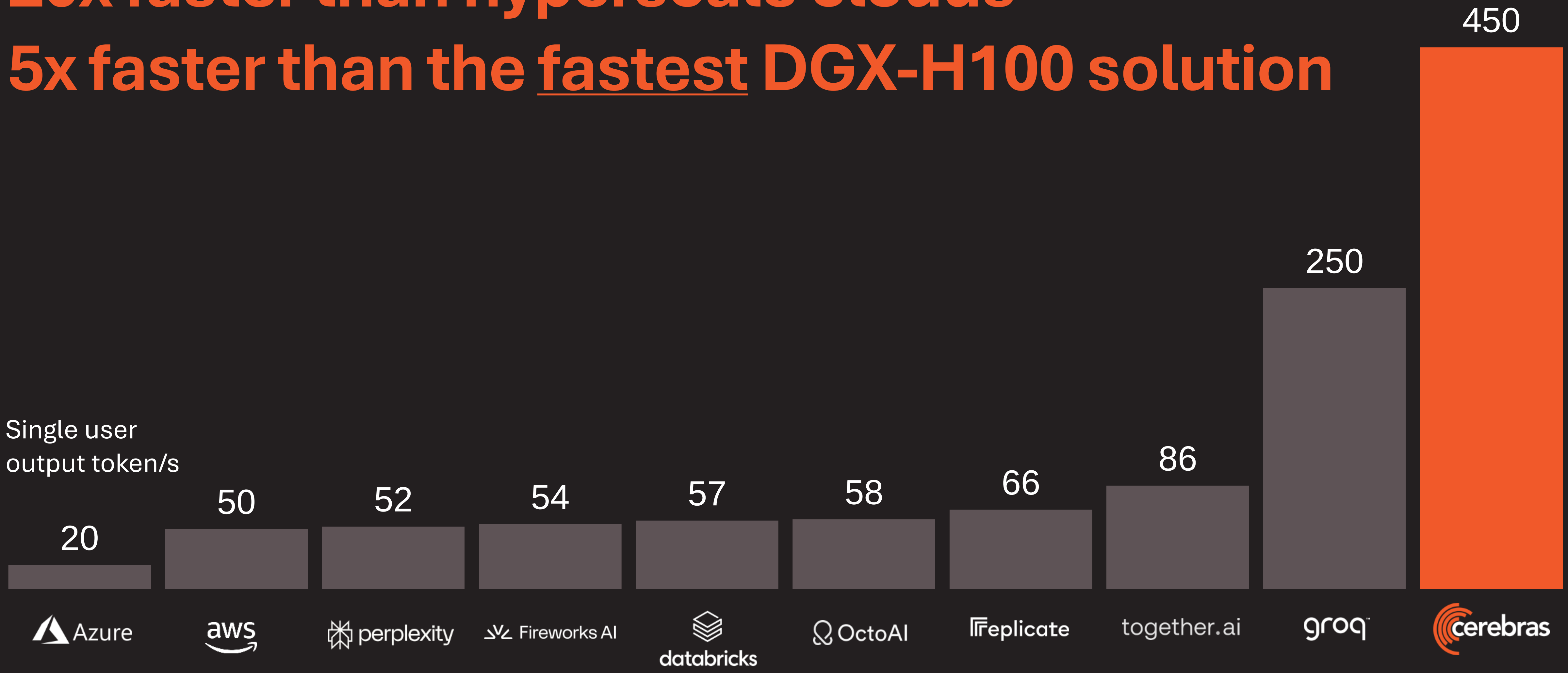
Imagine agents waiting for the GPU output in the background, multiple times, before any response to the user.

Such a product would not be viable.

The fastest inference on the planet on Llama3.1-70B

20x faster than hyperscale clouds

5x faster than the fastest DGX-H100 solution



Source Artificial Analysis, Llama3 & 3.1 Results.

The GPU Impossible

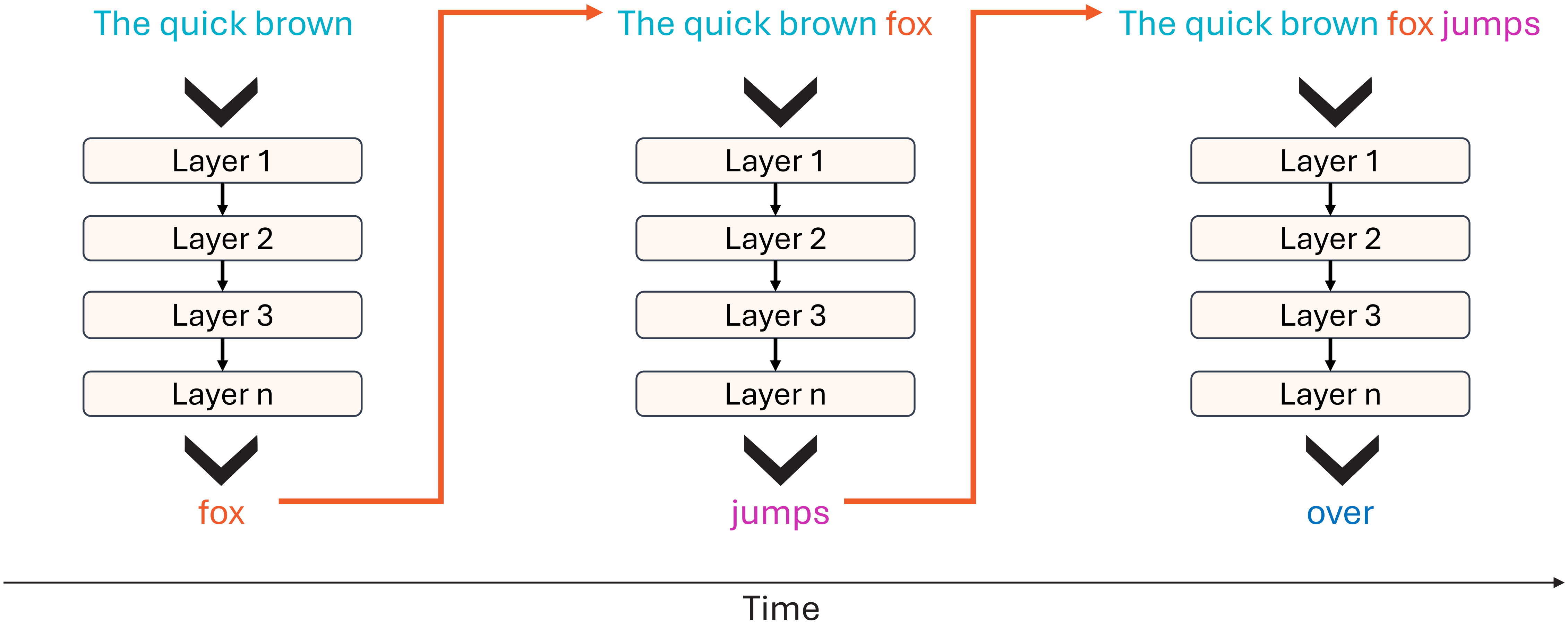
5x faster than the fastest GPU solution.

No number of GPUs can do this.

Memory Bandwidth

GenAI inference is a **memory bandwidth** problem

- Generating 1000 tokens takes 1000 serial passes through the model
- Each pass requires reading all model parameters from memory
- Low memory bandwidth is the bottleneck for generation performance



Why Wafer Scale Matters

	Cerebras WSE-3	Nvidia H100	Cerebras Advantage
Chip size	46,225 mm^2	814 mm^2	**57x**
Cores	900,000	16,896 FP32 + 528 Tensor	**52x**
On-chip memory	44 Gigabytes	0.05 Gigabytes	**880x**
Memory bandwidth	21 Petabytes/s	0.003 Petabytes/s	**7,000x**

Completely removes the memory bandwidth bottleneck

WSE-3 Core

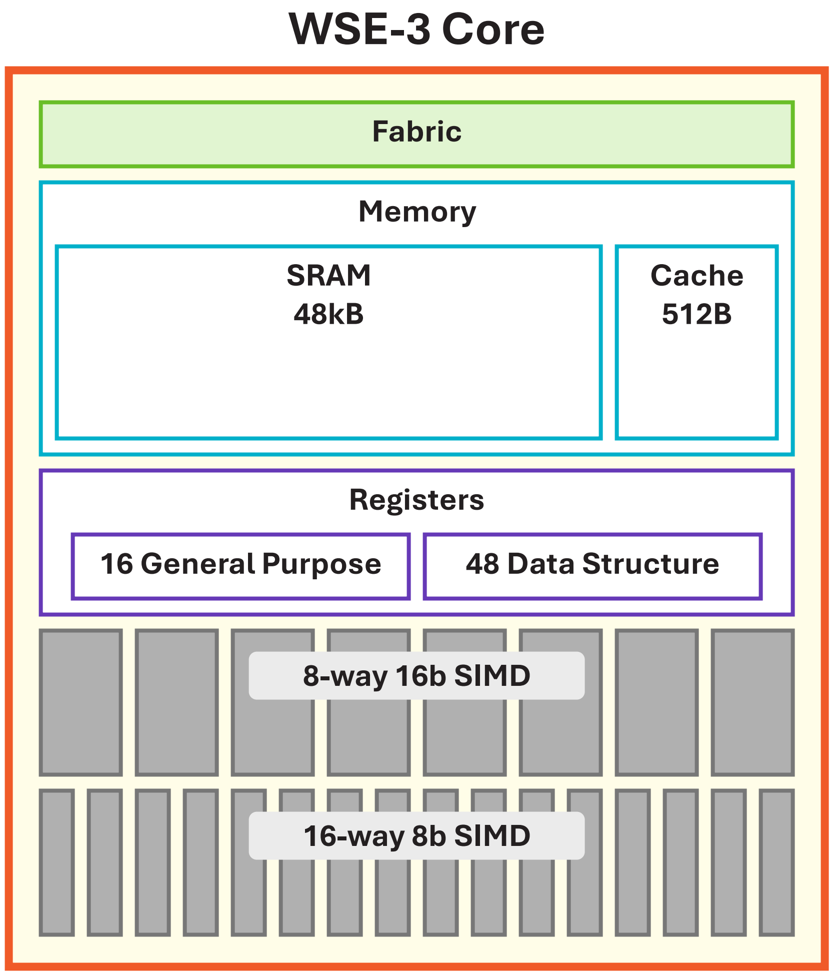
Tightly coupled compute and memory

High performance for AI compute

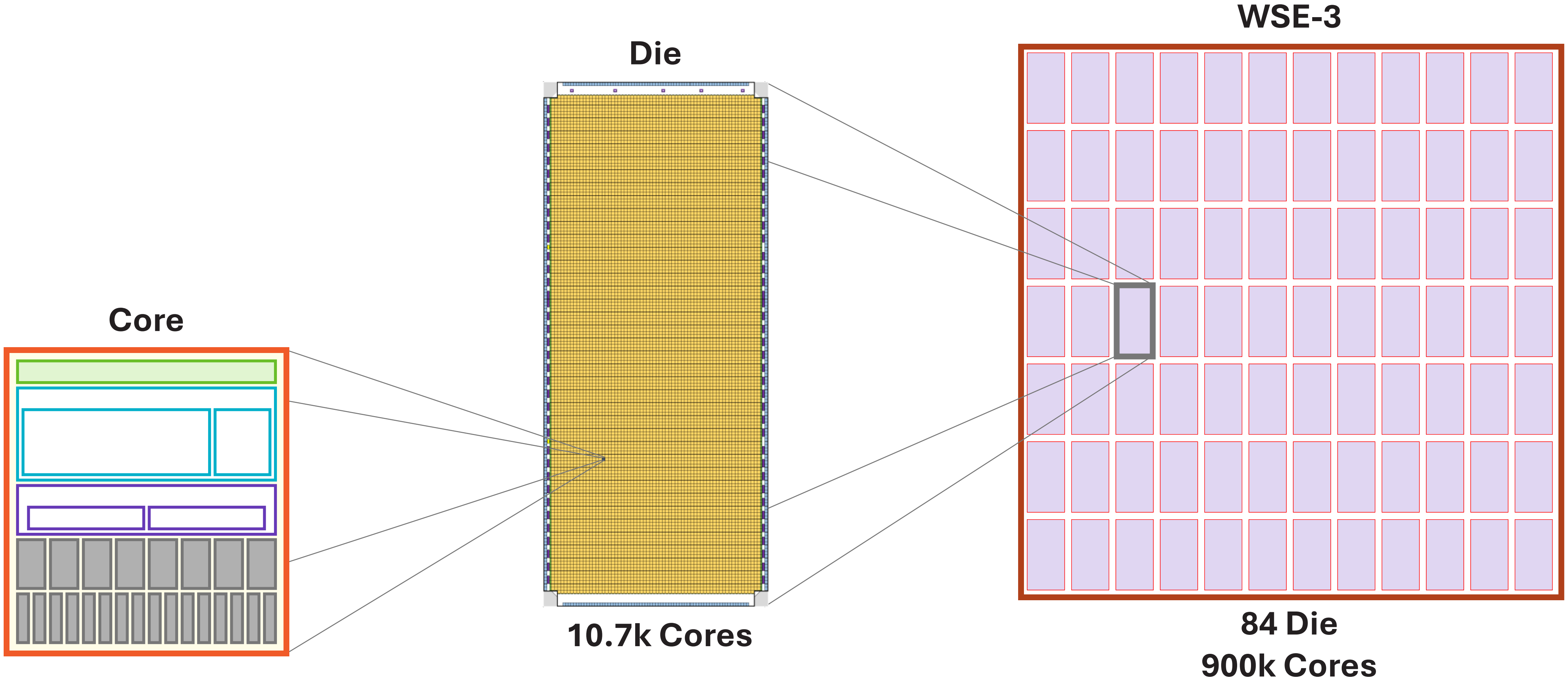
- High performance tensor operations
  - 8-way SIMD for 16b data (FP/BF16)
  - 16-way SIMD for 8b data (Fixed/INT8)
- Fast non-linear functions instructions
- Fine-grained dataflow scheduling
- Native unstructured sparsity acceleration

High bandwidth memory and cache

- 48kB SRAM per core
- 512B local cache per core
- Full bandwidth for full SIMD performance



From Small Core to Massive Wafer



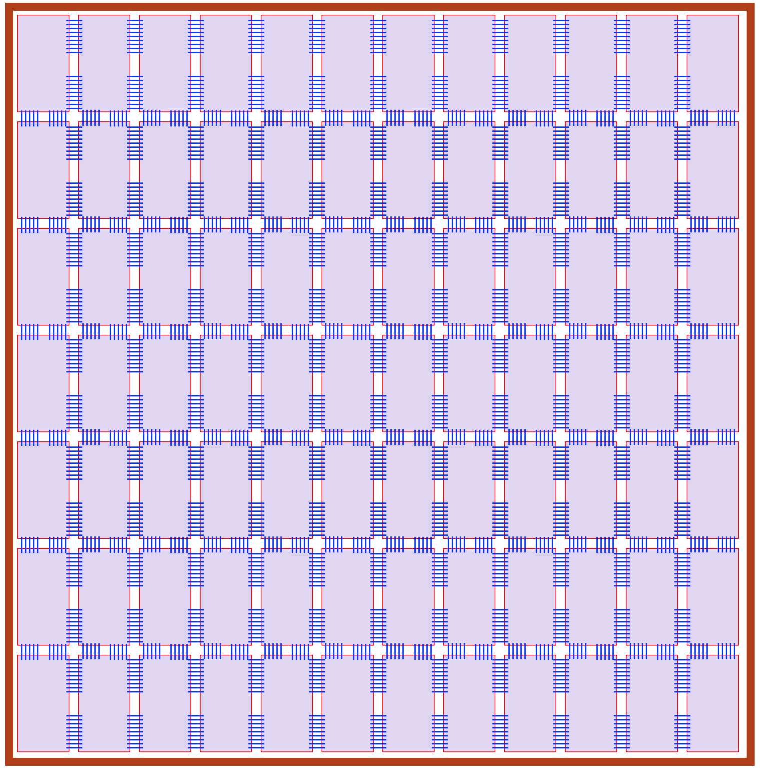
WSE-3 Interconnect

Connection between dies crossing reticle boundaries

- Invented process in first generation WSE
- Extended to 5nm in collaboration with TSMC

Co-designed with die-level fabric and system software

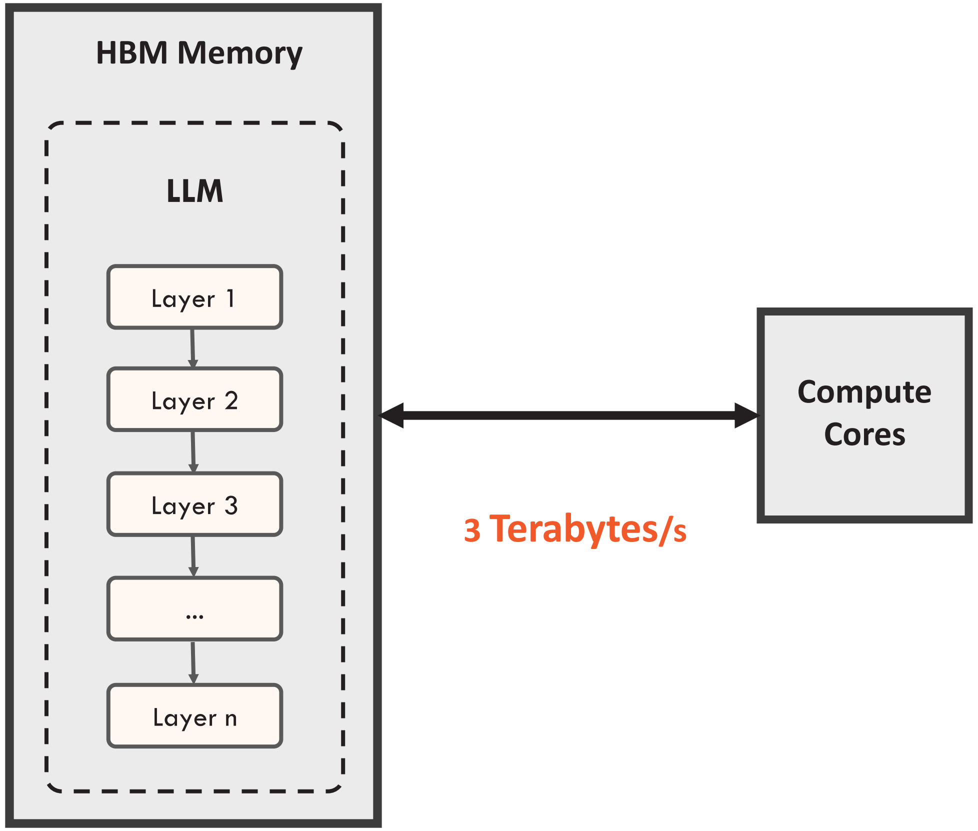
- Each die has 2D mesh fabric connecting all cores
- Extend across die boundaries at full performance
- Uniform fabric at die level and wafer level
- Built-in redundancy to route around failures
- Software always sees a fully uniform 2D mesh



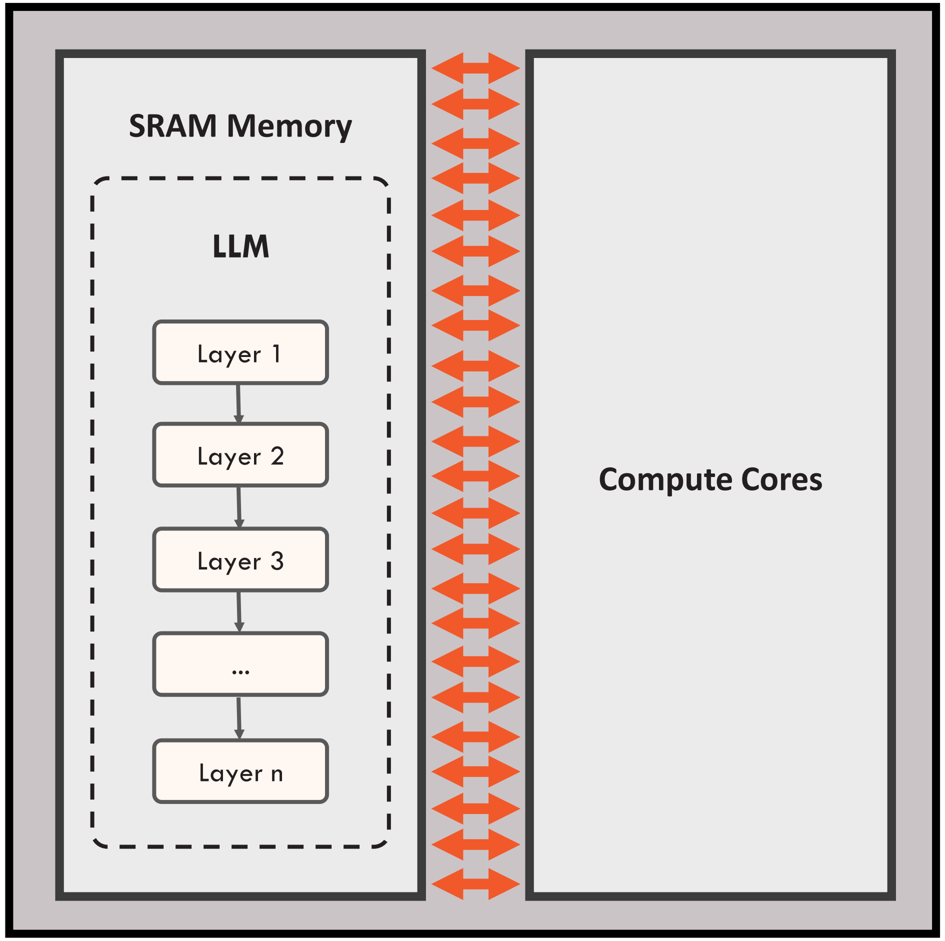
Wafer Scale, SRAM based, Compute-in-Memory

7,000x more memory bandwidth than the GPU

Nvidia H100

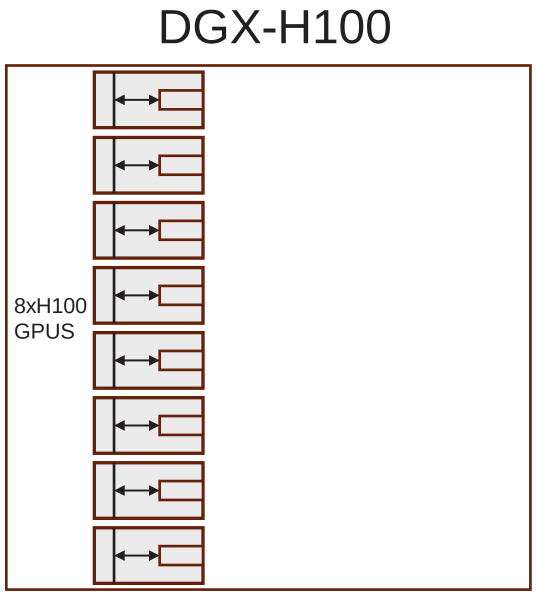


Cerebras Wafer Scale Engine 3



21 Petabytes/s

Multi-GPU memory bandwidth aggregation



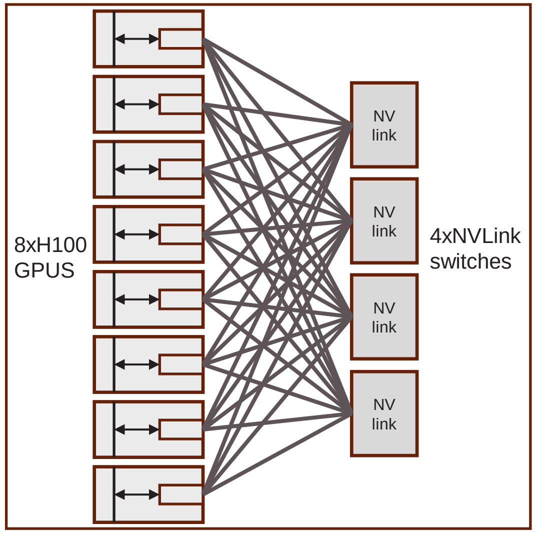
	Each H100	8xH100
Mem Bandwidth	3.35 TB/s	26.8 TB/s

To achieve higher aggregate memory bandwidth…

- Integrate 8 GPUs into a single server
- Accessed with tensor parallel execution
- 8x higher aggregate memory bandwidth

Multi-GPU memory bandwidth aggregation

DGX-H100



	Each H100	8xH100
Mem Bandwidth	3.35 TB/s	26.8 TB/s
IO Bandwidth	9000GB/s 36x 100Gb/s serial	7.2 TB/s 288x 100Gb/s serial
Power	36W (5.0 pJ/bit*2)	288W IO + 300 Switch 588W Total

*GPU estimate based on 5nm 100G serdes power with Nvidia H100 NVLink bandwidth

To achieve higher aggregate memory bandwidth…

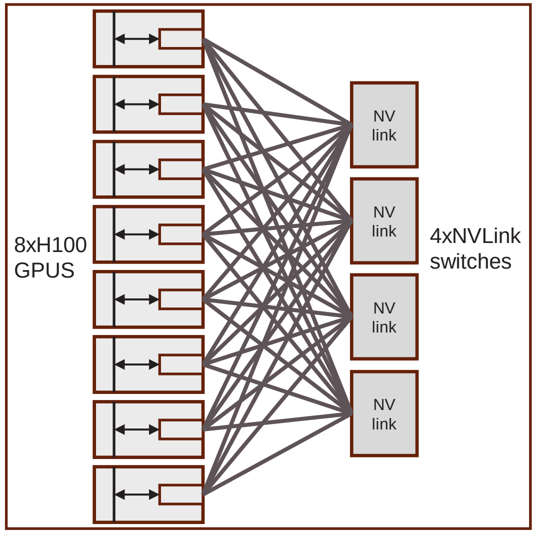
- Integrate 8 GPUs into a single server
- Accessed with tensor parallel execution
- 8x higher aggregate memory bandwidth

But it comes at a cost

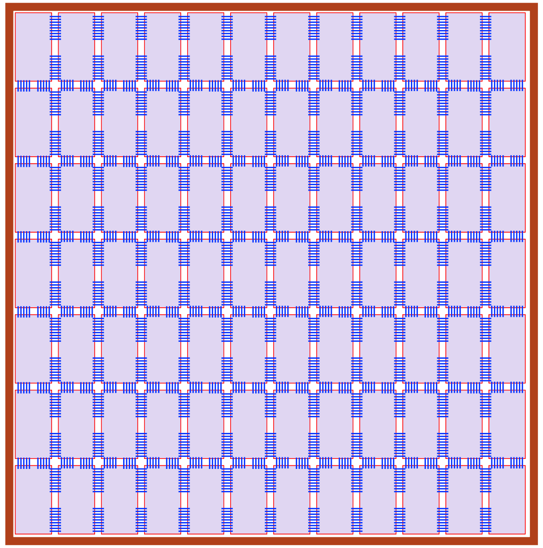
- Hundreds of high speed serial links
- Multiple interconnect switch chips
- High cost
- High power – 0.5 kW

Multi-GPU vs. Wafer-Scale Integration

DGX-H100



Wafer Scale Engine

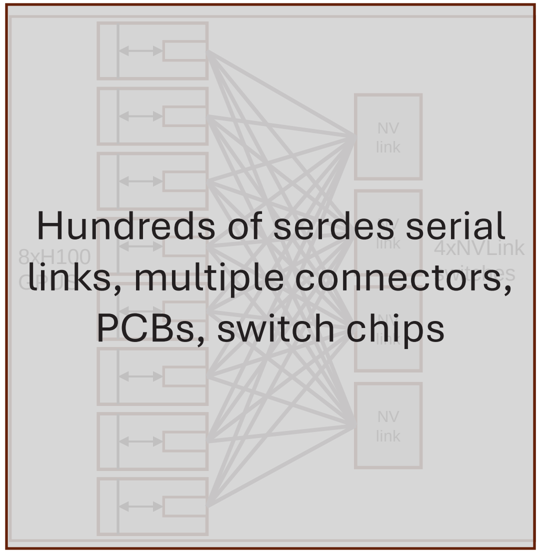


	Each H100	8xH100		Each Die	84xDie
Mem Bandwidth	3.35 TB/s	26.8 TB/s	**800x more Mem bandwidth**	255 TB/s	21 PB/s
IO Bandwidth	9000GB/s 36x 100Gb/s serial	7.2 TB/s 288x 100Gb/s serial	**33x more Inter-Die IO**	2880GB/s 480x 24Gb/s parallel	242TB/s 40320x 24Gb/s parallel
Power	36W (5.0 pJ/bit*2)	288W IO + 300 Switch 588W Total	**6x Lower Power**	1.1W (0.05 pJ/bit*2)	97W

*GPU estimate based on 5nm 100G serdes power with Nvidia H100 NVLink bandwidth

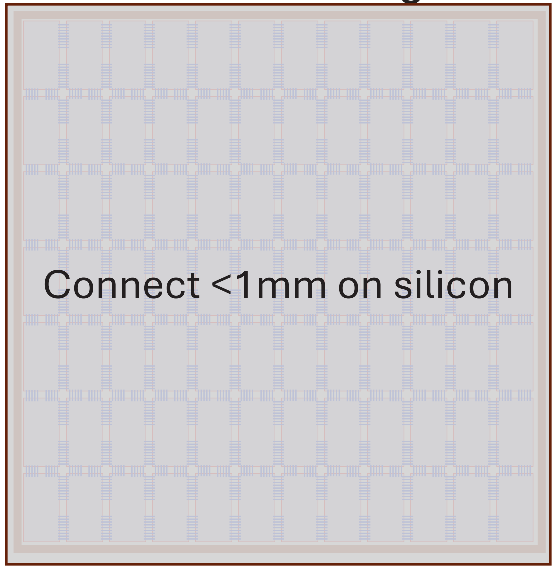
Multi-GPU vs. Wafer-Scale Integration

DGX-H100



Hundreds of serdes serial links, multiple connectors, PCBs, switch chips

Wafer Scale Engine



Connect <1mm on silicon

	Each H100	8xH100		Each Die	84xDie
Mem Bandwidth	3.35 TB/s	26.8 TB/s	**800x more Mem bandwidth**	255 TB/s	21 PB/s
IO Bandwidth	9000GB/s 36x 100Gb/s serial	7.2 TB/s 288x 100Gb/s serial	**33x more Inter-Die IO**	2880GB/s 480x 24Gb/s parallel	242TB/s 40320x 24Gb/s parallel
Power	36W (5.0 pJ/bit*2)	288W IO + 300 Switch 588W Total	**6x Lower Power**	1.1W (0.05 pJ/bit*2)	97W

*GPU estimate based on 5nm 100G serdes power with Nvidia H100 NVLink bandwidth

Multi-GPU memory bandwidth does not scale well

Higher memory bandwidth on paper only

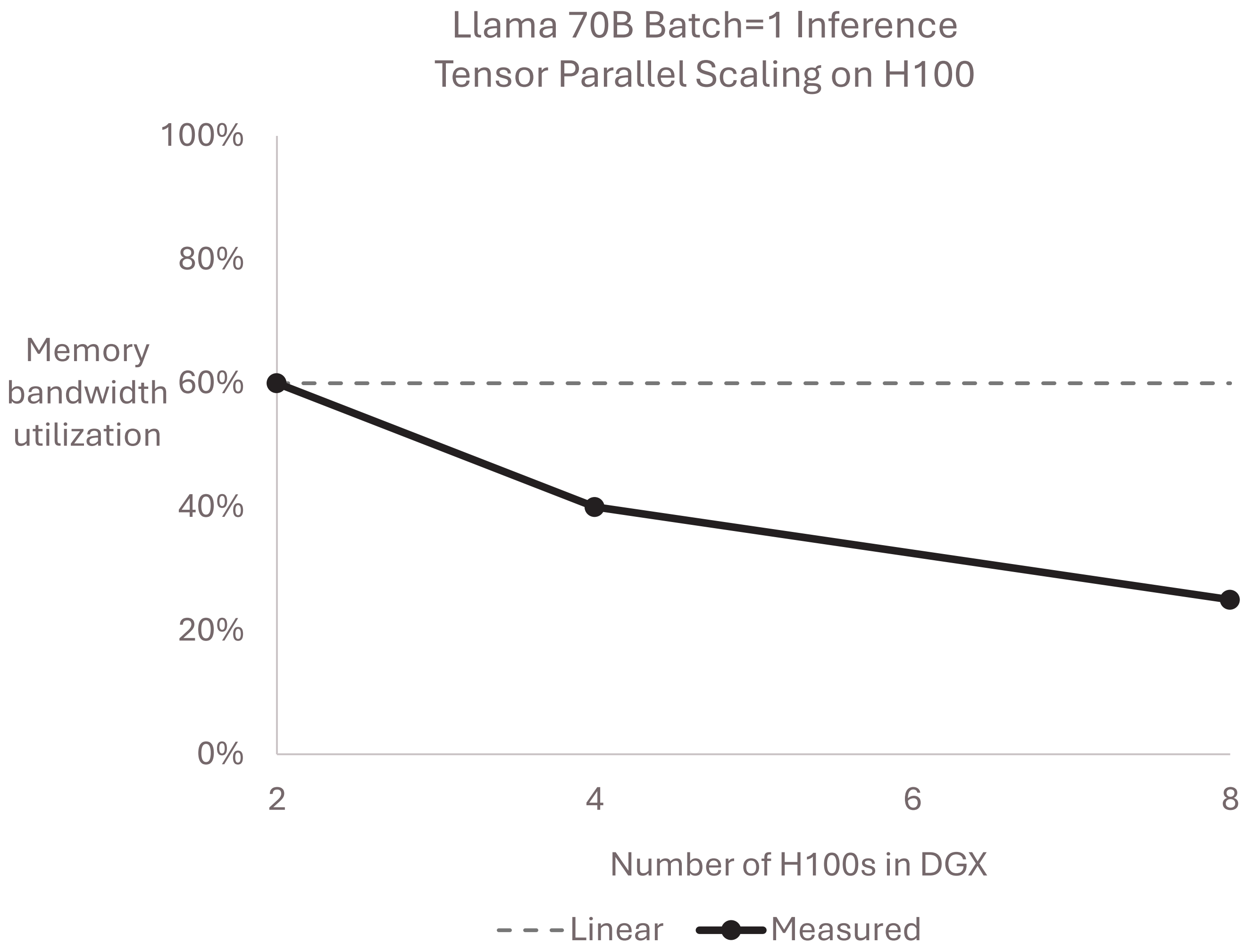
- In reality, tensor parallel scales poorly due to interconnect bandwidth and latency overhead
- Measured memory bandwidth utilization drops from 60% on 2 H100s to 25% on 8 H100s
- That's even on the highest performance interconnect within a DGX

Multi-GPU architecture has poor scaling

- 4x theoretical higher memory bandwidth
- 1.7x performance only
- 42% scaling efficiency

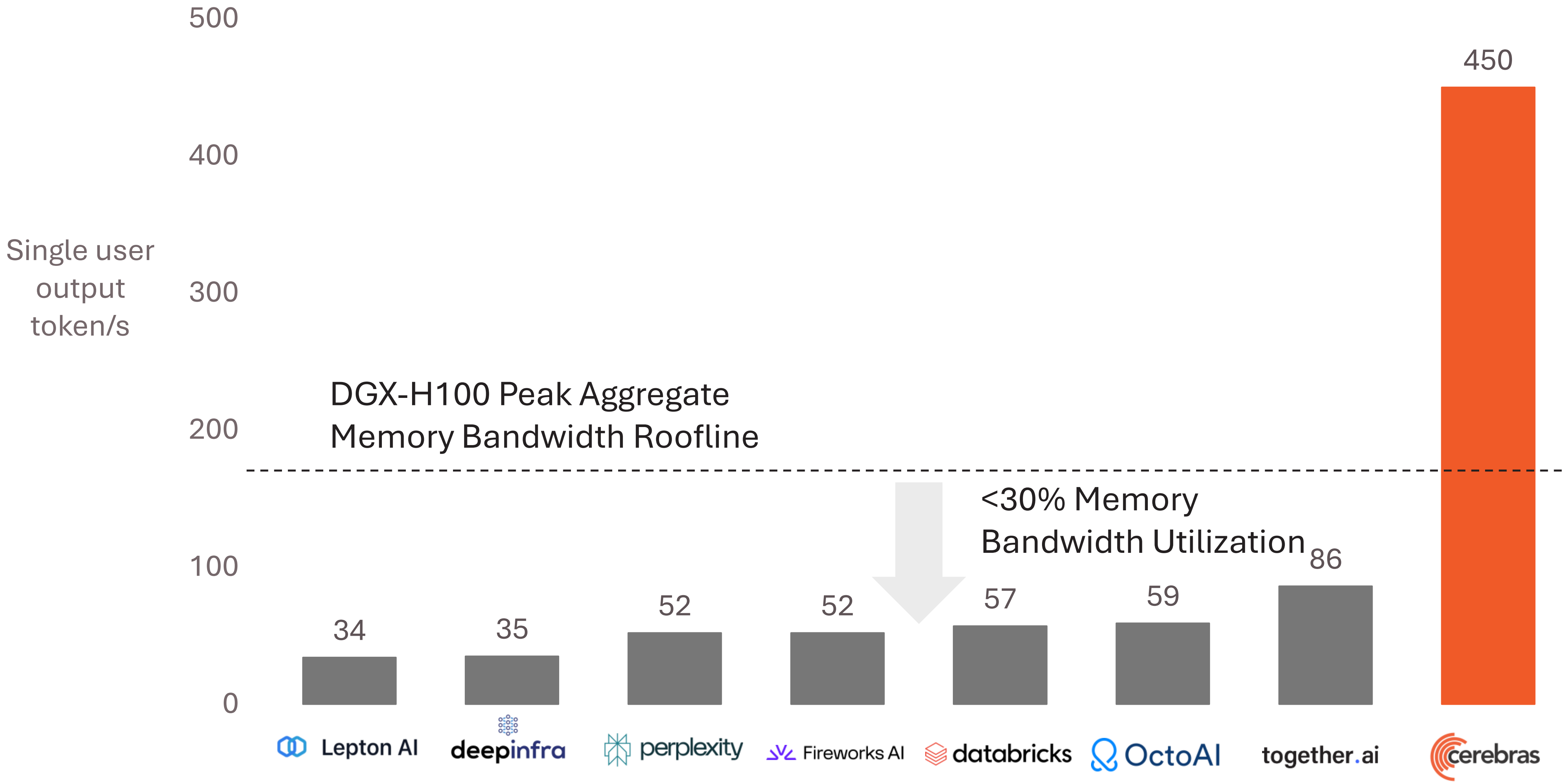
This is why GPU cannot scale inference generation performance beyond DGX

Lower IO limits performance scaling



Source: Databricks inference performance results
https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices

Real world low memory bandwidth utilization on Llama3.1-70B



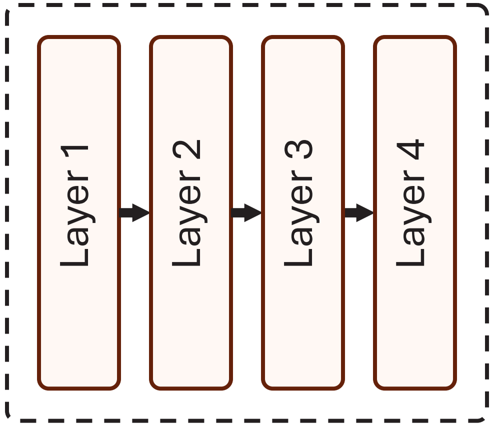
Source: Artificial Analysis https://artificialanalysis.ai/

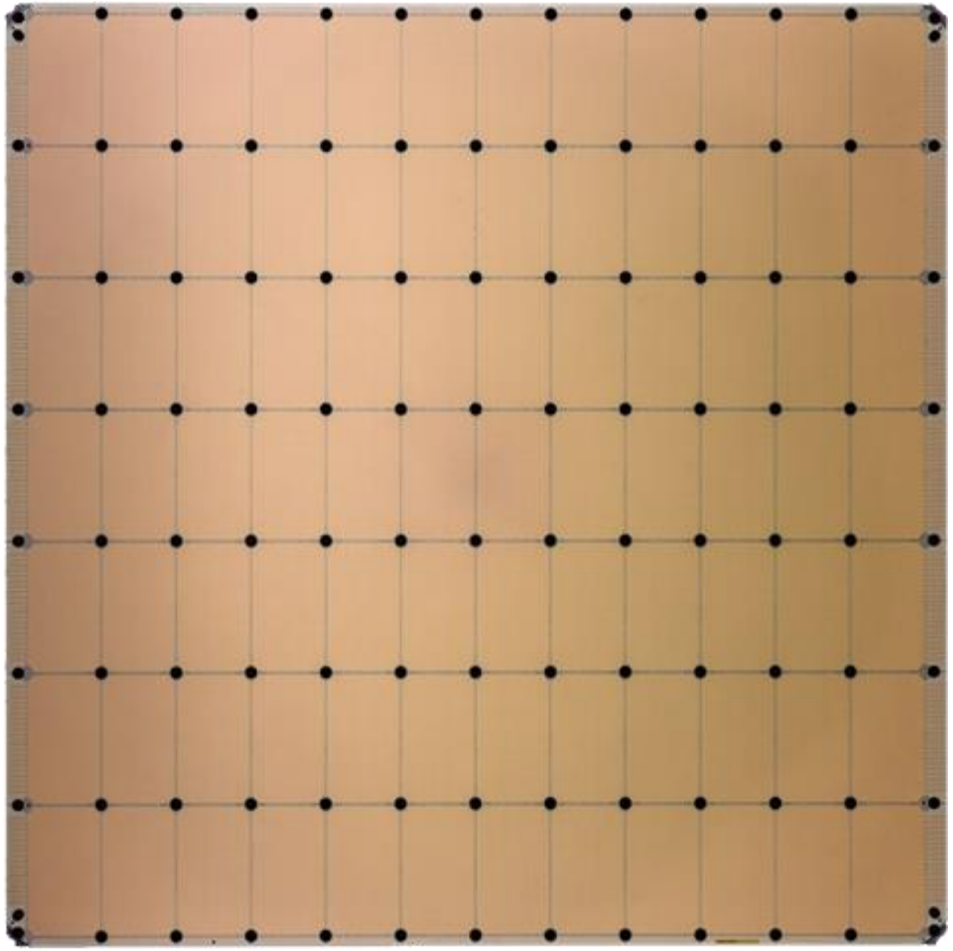
Lowest Latency Generation

Pipeline execution on a single chip

Massive memory bandwidth enables opposite execution model

- GPU: multiple chips to run a single layer
- Cerebras: fraction of a chip to run a single layer





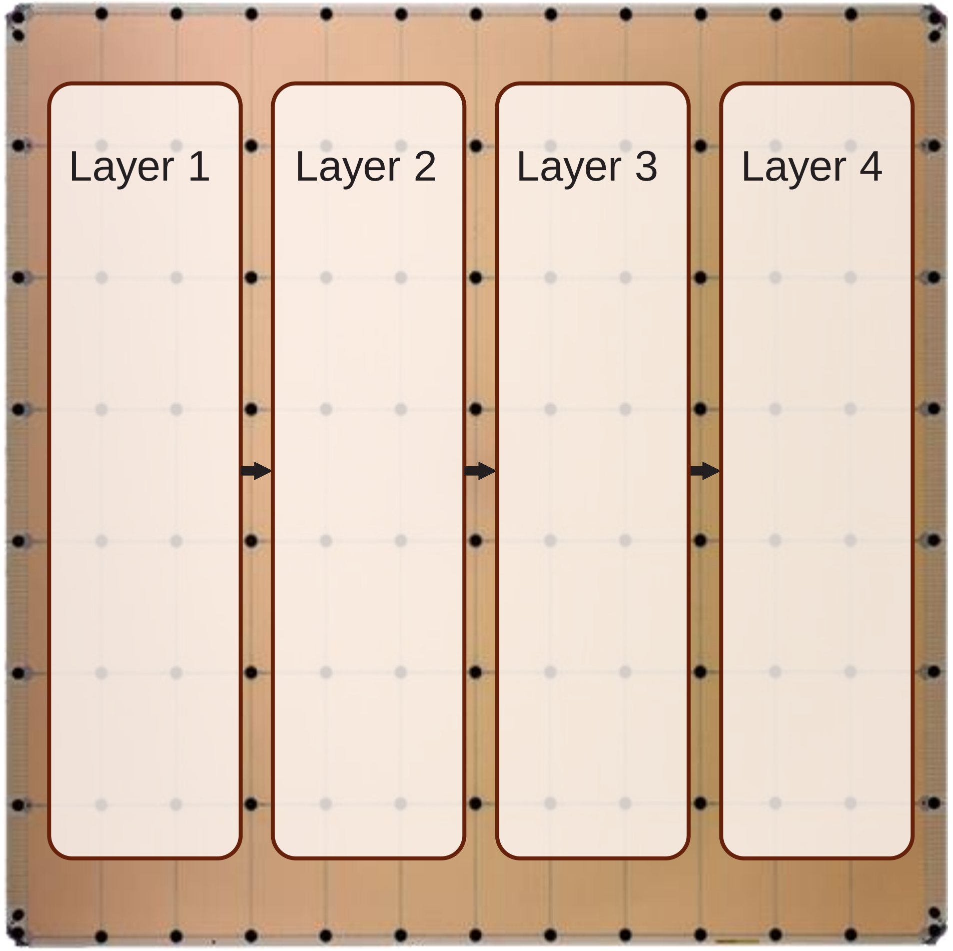
Pipeline execution on a single chip

Massive memory bandwidth enables opposite execution model

- GPU: multiple chips to run a single layer
- Cerebras: fraction of a chip to run a single layer

Pipeline mapping of model

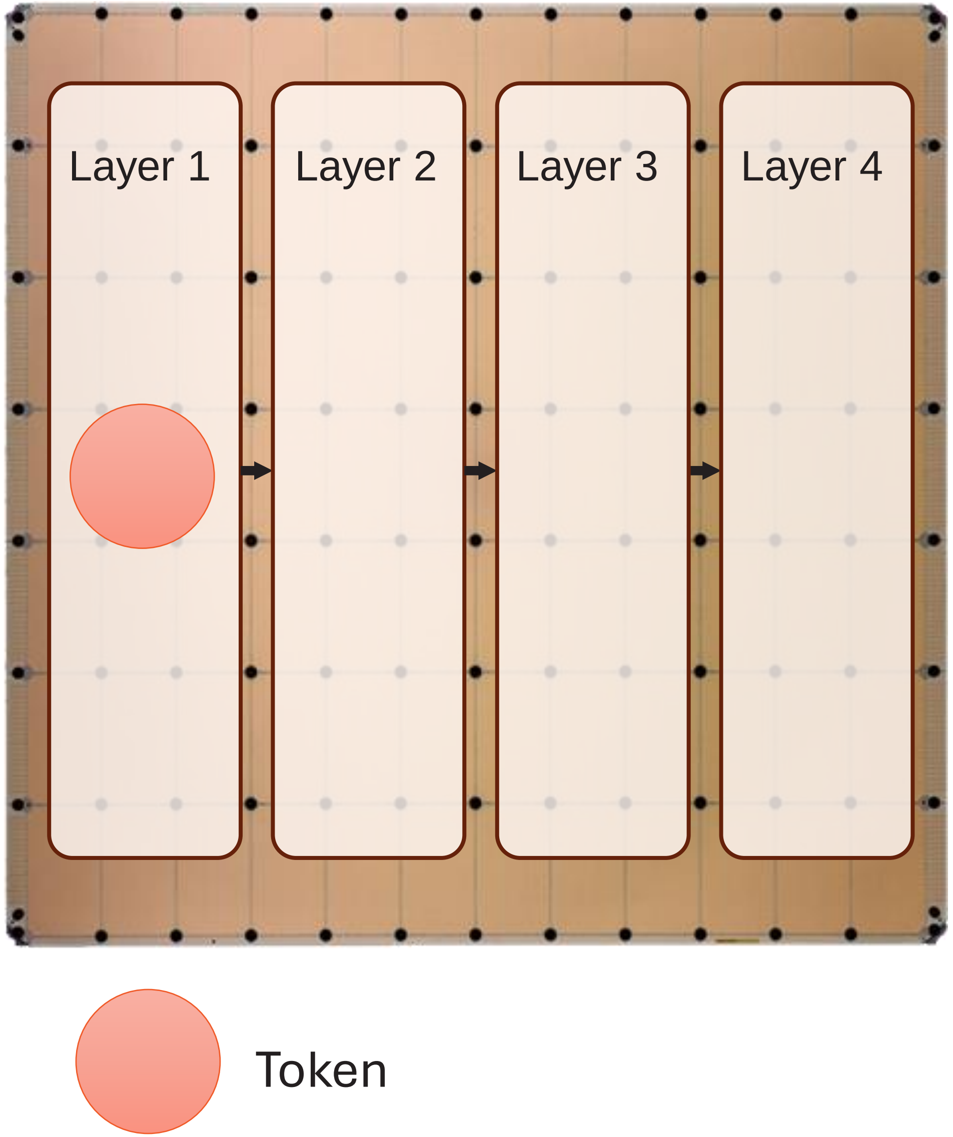
- Model layers are mapped to wafer regions
- Size of wafer region determined based on memory and compute requirements
- Model weights and KV cache stored locally in the region memory close to the compute



Pipeline execution

A low latency pipeline

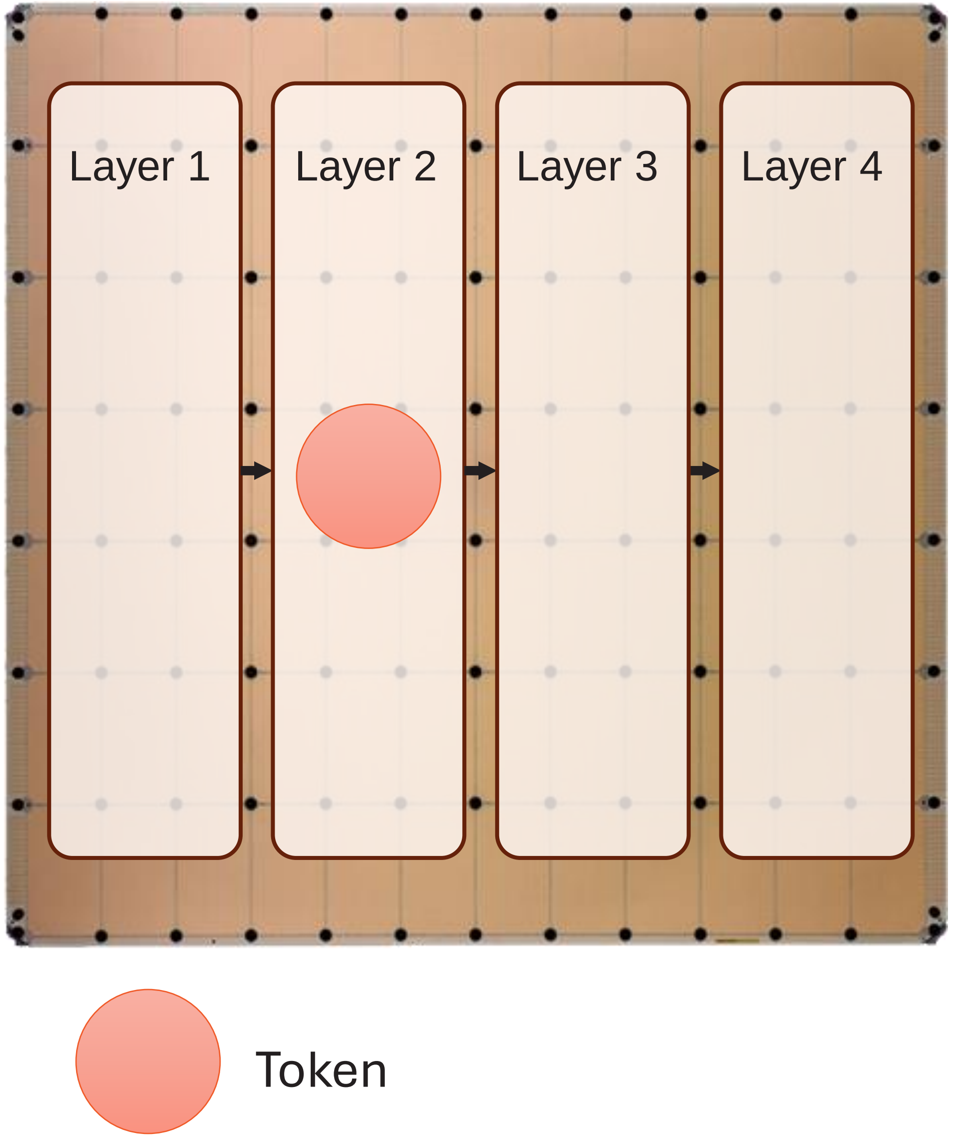
- Each wafer region processes 1 token
  - Enough memory bandwidth to run local batch size 1
  - Memory to feed compute datapath at full speed
  - Enabling optimized performance for matrix*vector
  - Local fabric interconnect to maintain low latency



Pipeline execution

A low latency pipeline

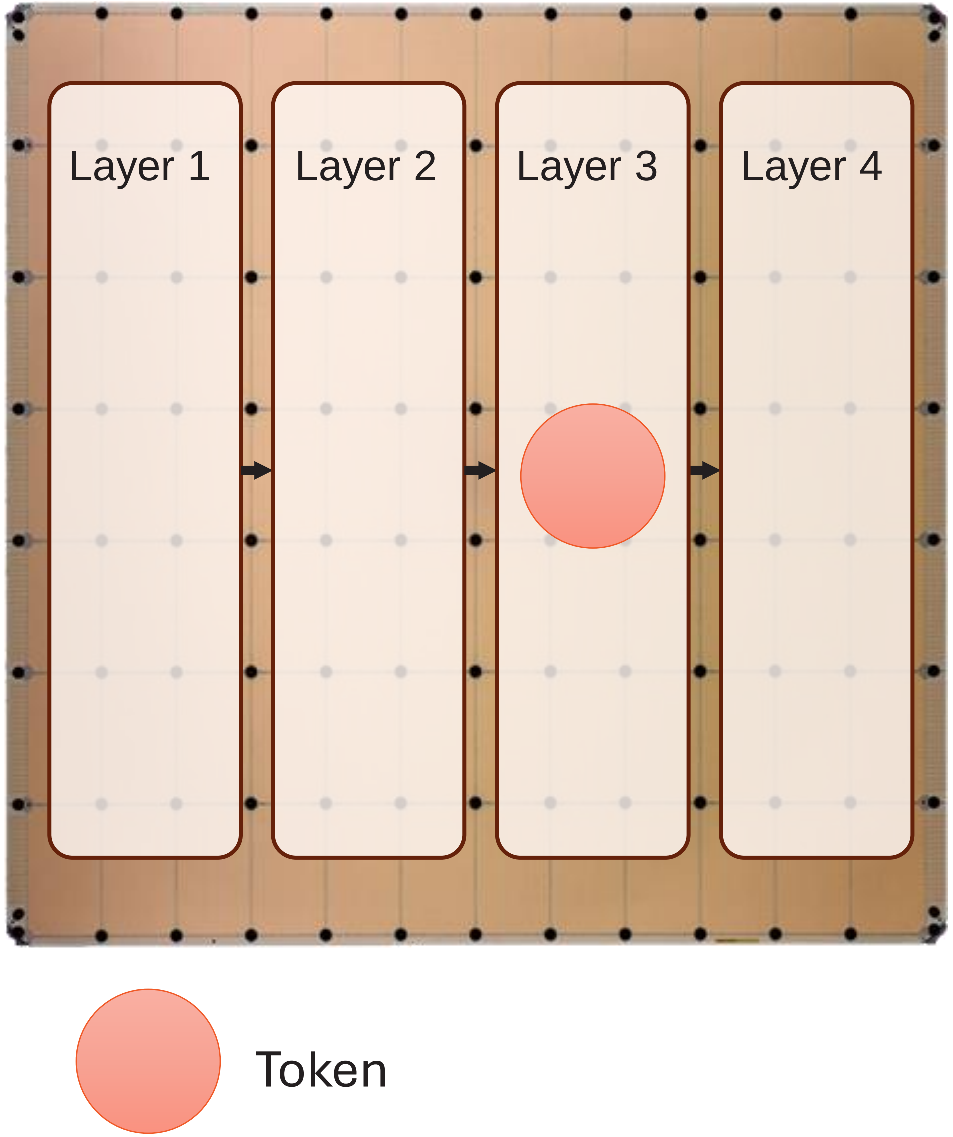
- Each wafer region processes 1 token
  - Enough memory bandwidth to run local batch size 1
  - Memory to feed compute datapath at full speed
  - Enabling optimized performance for matrix*vector
  - Local fabric interconnect to maintain low latency
- Regions are placed to reduce latency
  - The next region is physically adjacent
  - Virtually no latency between pipe stages
  - Possible because it’s all done on-chip



Pipeline execution

A low latency pipeline

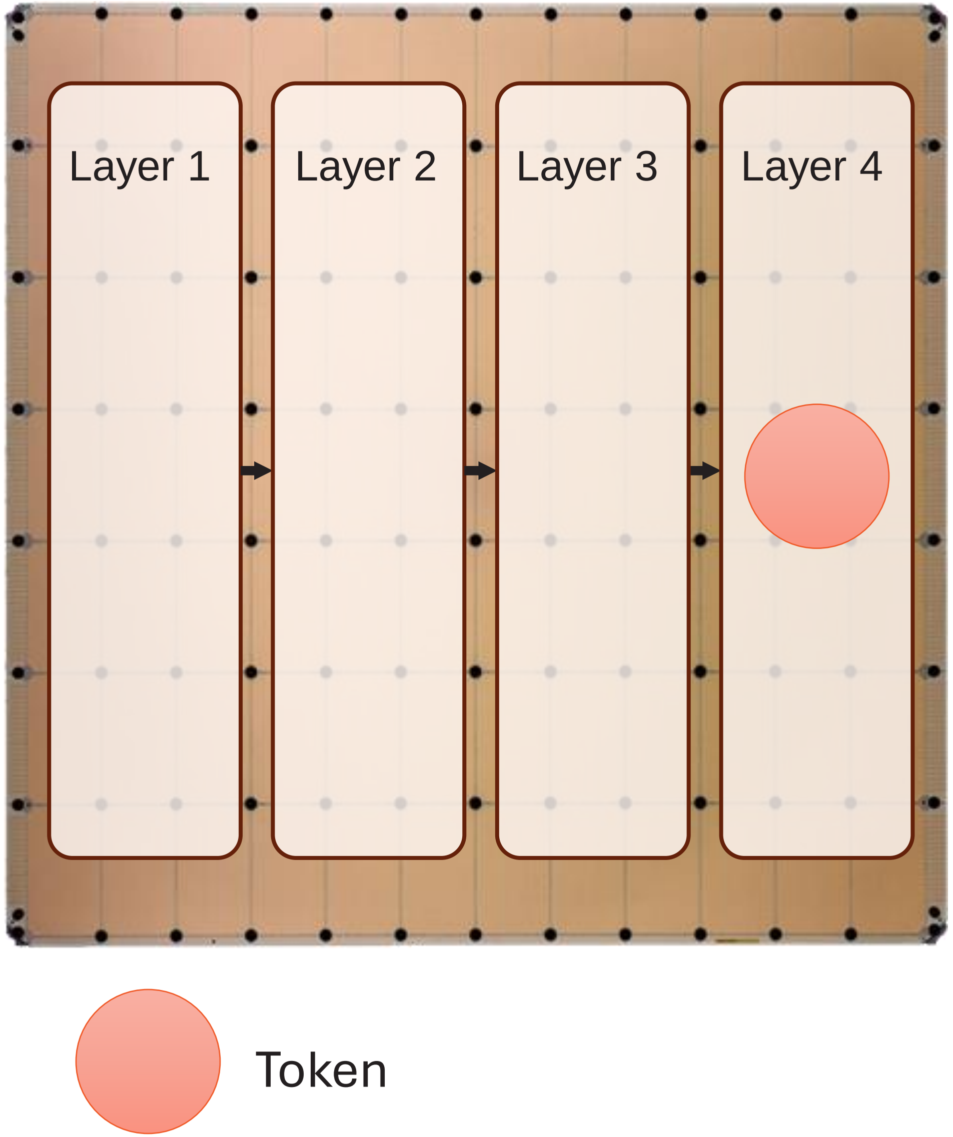
- Each wafer region processes 1 token
  - Enough memory bandwidth to run local batch size 1
  - Memory to feed compute datapath at full speed
  - Enabling optimized performance for matrix*vector
  - Local fabric interconnect to maintain low latency
- Regions are placed to reduce latency
  - The next region is physically adjacent
  - Virtually no latency between pipe stages
  - Possible because it's all done on-chip



Pipeline execution

A low latency pipeline

- Each wafer region processes 1 token
  - Enough memory bandwidth to run local batch size 1
  - Memory to feed compute datapath at full speed
  - Enabling optimized performance for matrix*vector
  - Local fabric interconnect to maintain low latency
- Regions are placed to reduce latency
  - The next region is physically adjacent
  - Virtually no latency between pipe stages
  - Possible because it’s all done on-chip

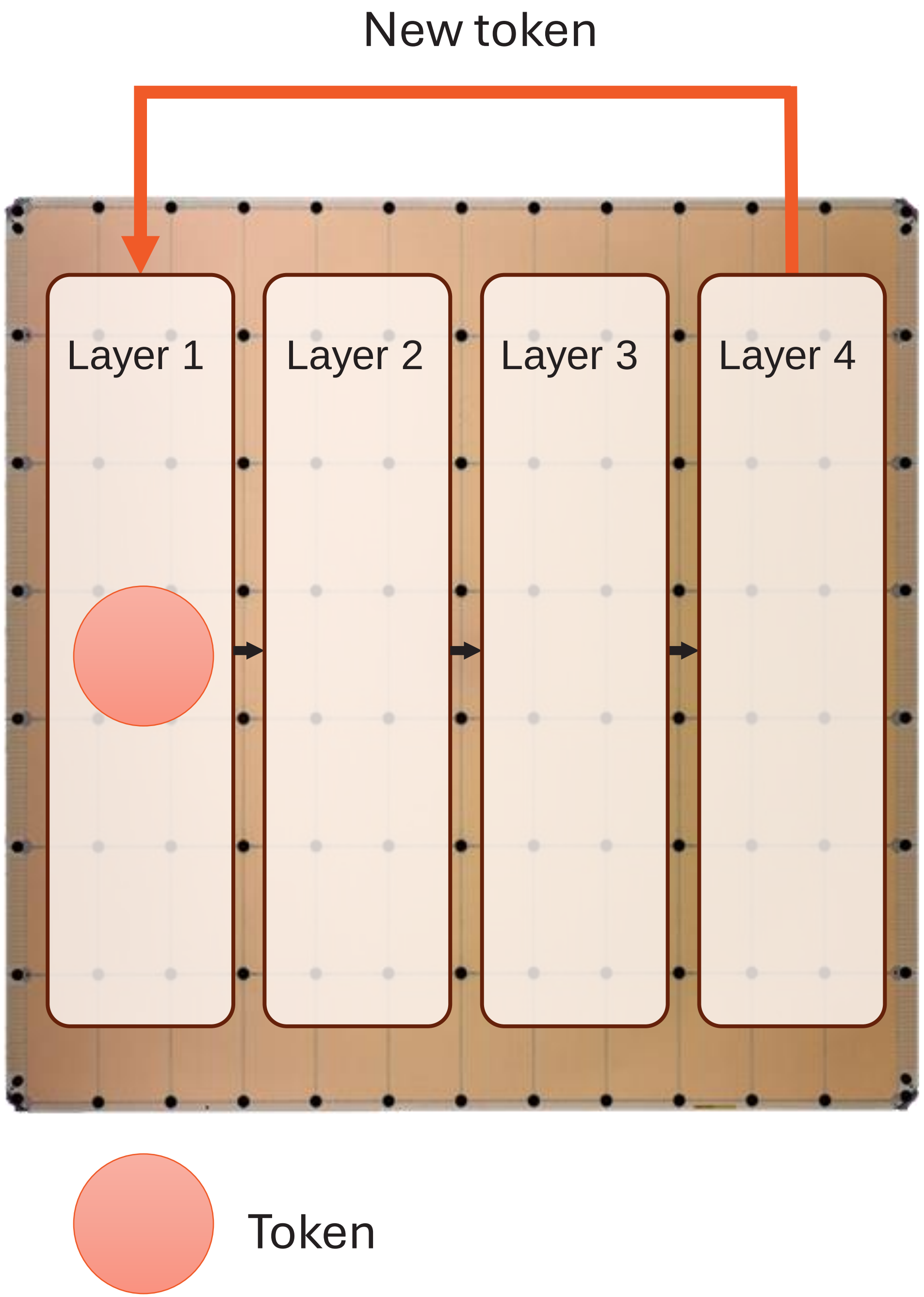


Pipeline execution

A low latency pipeline

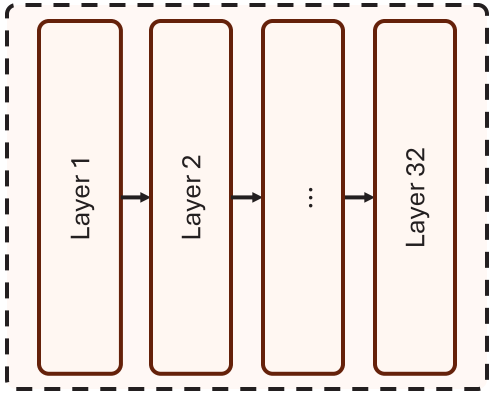
- Each wafer region processes 1 token
  - Enough memory bandwidth to run local batch size 1
  - Memory to feed compute datapath at full speed
  - Enabling optimized performance for matrix*vector
  - Local fabric interconnect to maintain low latency
- Regions are placed to reduce latency
  - The next region is physically adjacent
  - Virtually no latency between pipe stages
  - Possible because it’s all done on-chip
- Last region output cycled back to generate next token

Pipelined execution enables super-fast token generation

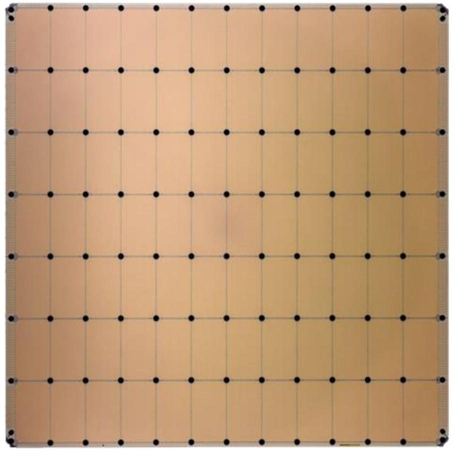


Scalability

Large models fit on our large on-chip memory



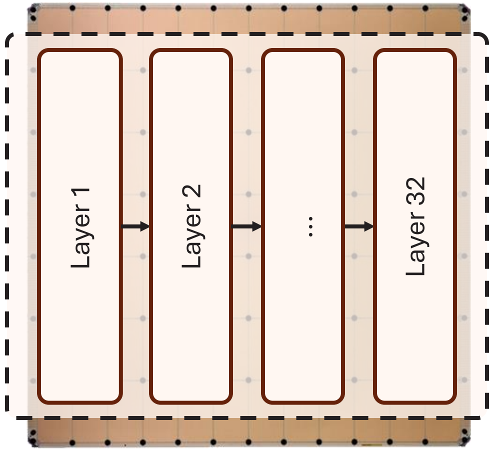
Llama3.1-8B
8 billion parameters
16 GB of memory (FP16)



WSE-3
44GB of SRAM

Large models fit on our large on-chip memory

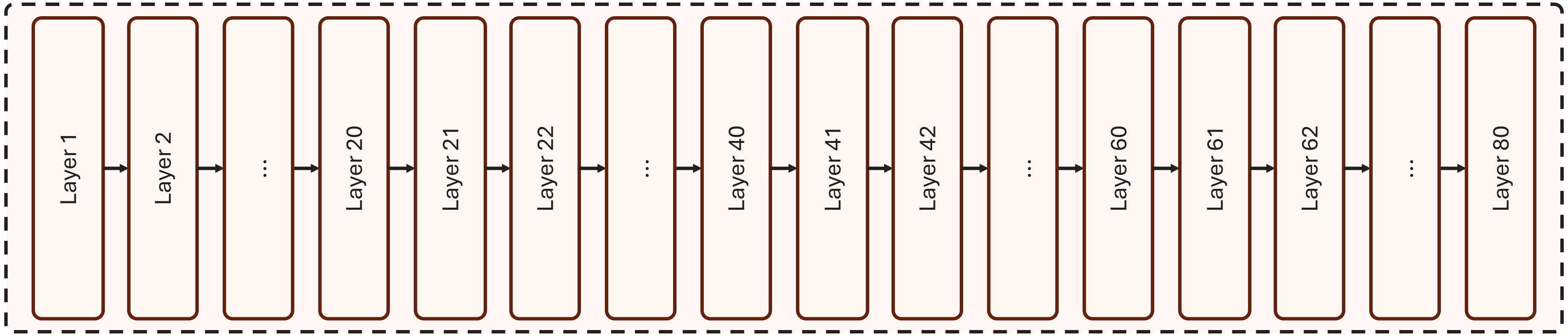
- Models that fit entirely on the wafer are mapped directly



Llama3.1-8B
8 billion parameters
16 GB of memory (FP16)

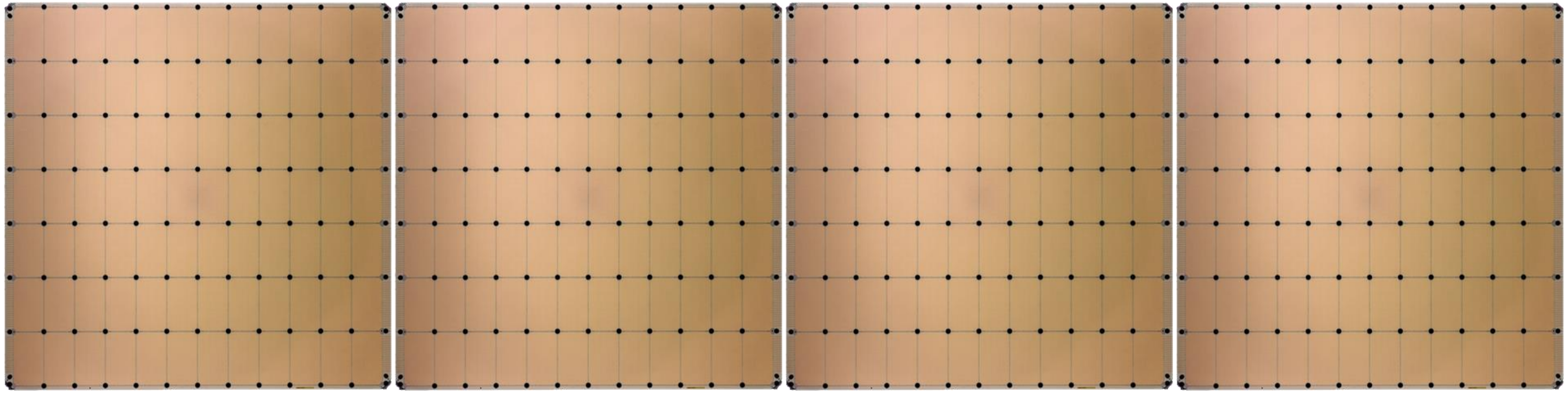
WSE-3
44GB of SRAM

Naturally scales to multiple wafers



Layer 1 → Layer 2 → … → Layer 20 → Layer 21 → Layer 22 → … → Layer 40 → Layer 41 → Layer 42 → … → Layer 60 → Layer 61 → Layer 62 → … → Layer 80

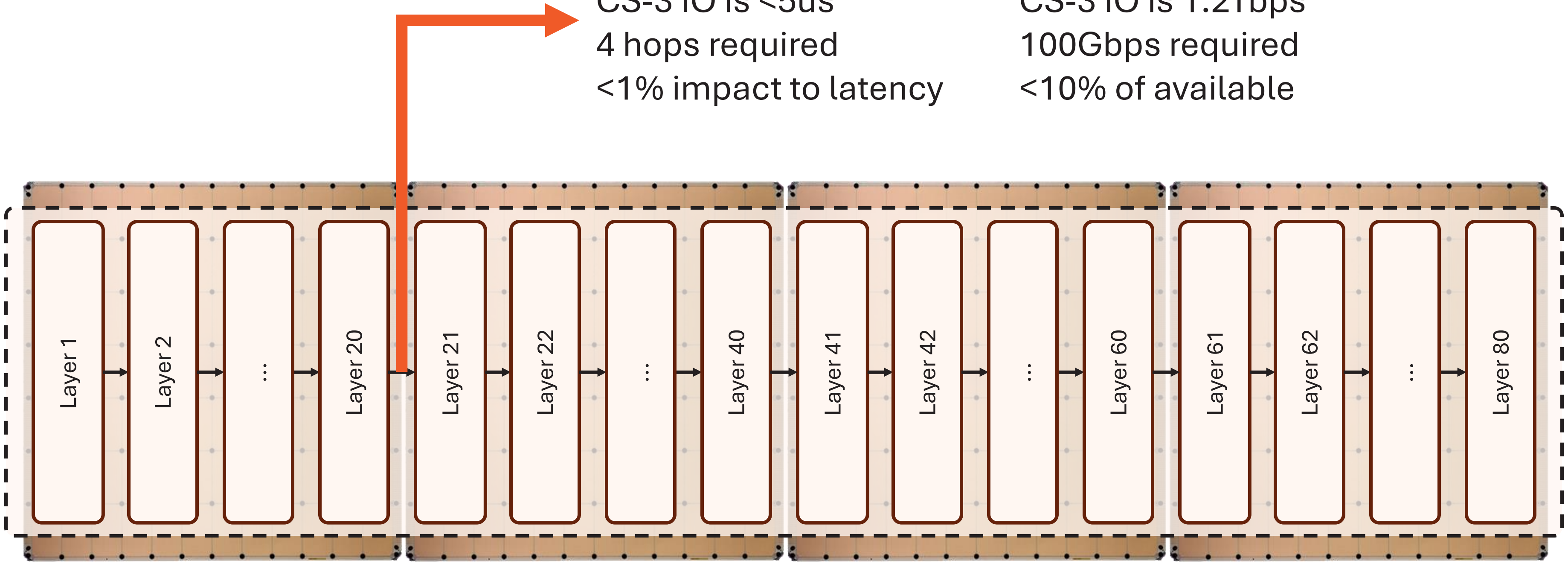
Llama3.1-70B
70 billion parameters
140 GB of memory (FP16)



4x WSE-3
176GB of SRAM

Naturally scales to multiple wafers

- Models that require more memory are mapped to multiple wafers
- Keep almost all high communication on wafer, using internal high bandwidth fabric
- Transfer only activations between wafers, requiring relatively lower bandwidth
- Using CS-3 low latency RDMA over Ethernet interconnect between systems



Latency:
CS-3 IO is <5us
4 hops required
<1% impact to latency

Bandwidth:
CS-3 IO is 1.2Tbps
100Gbps required
<10% of available

Llama3.1-70B
70 billion parameters
140 GB of memory (FP16)

4x WSE-3
176GB of SRAM

High Throughput

The nasty GPU latency vs. throughput tradeoff

GPUs are designed for high throughput

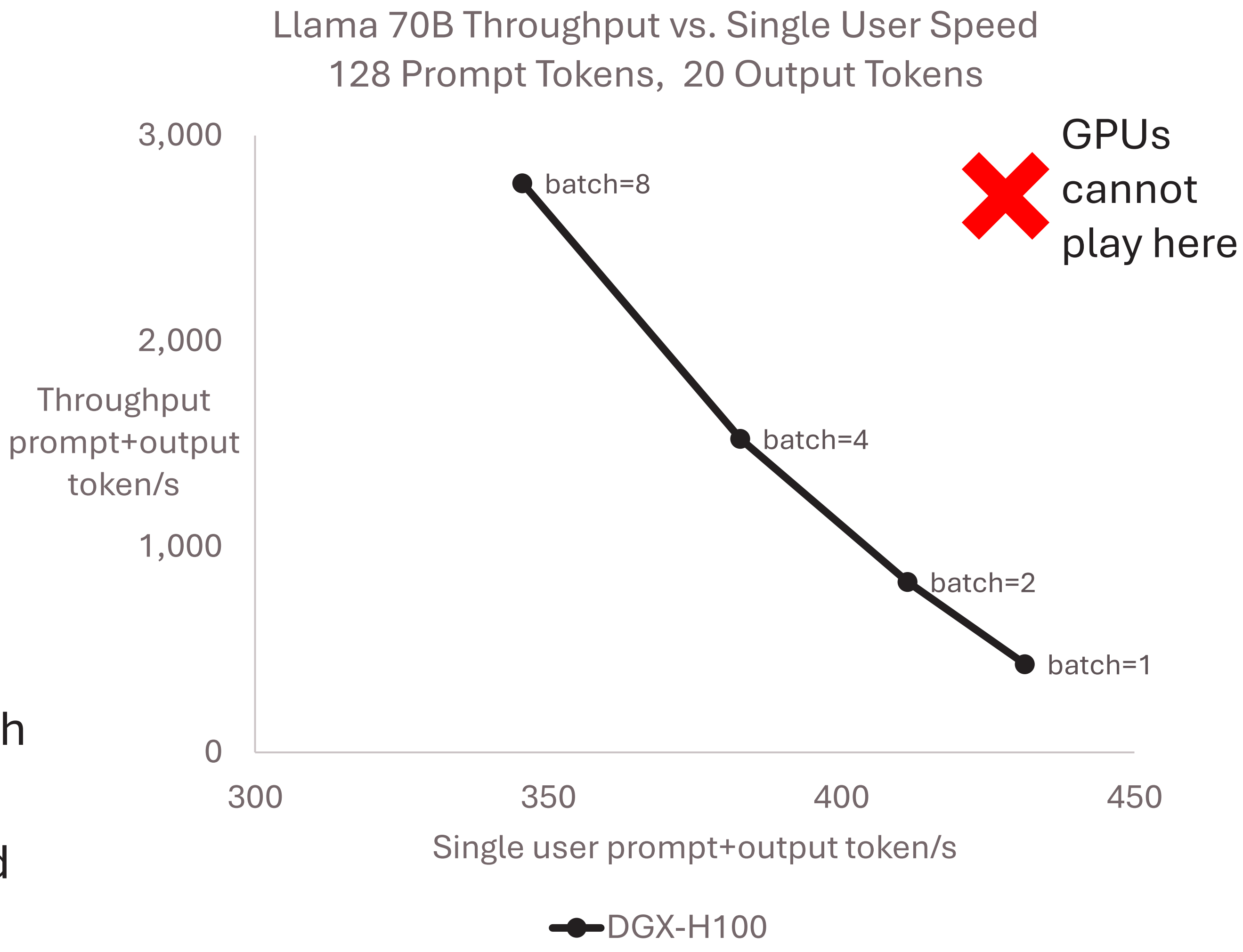
- High throughput requires high batch size
- Batch size = concurrent users
- But high batch size results in worse latency

Latency is critical in generative inference

- Latency = single user speed
- Better latency enables better user experience
- Better latency enables agentic workflows

The GPU latency vs. throughput tradeoff

- Single user speed low to start, even at low batch
- To achieve higher throughput requires higher batch, resulting in even lower single user speed



Llama 70B Throughput vs. Single User Speed
128 Prompt Tokens, 20 Output Tokens

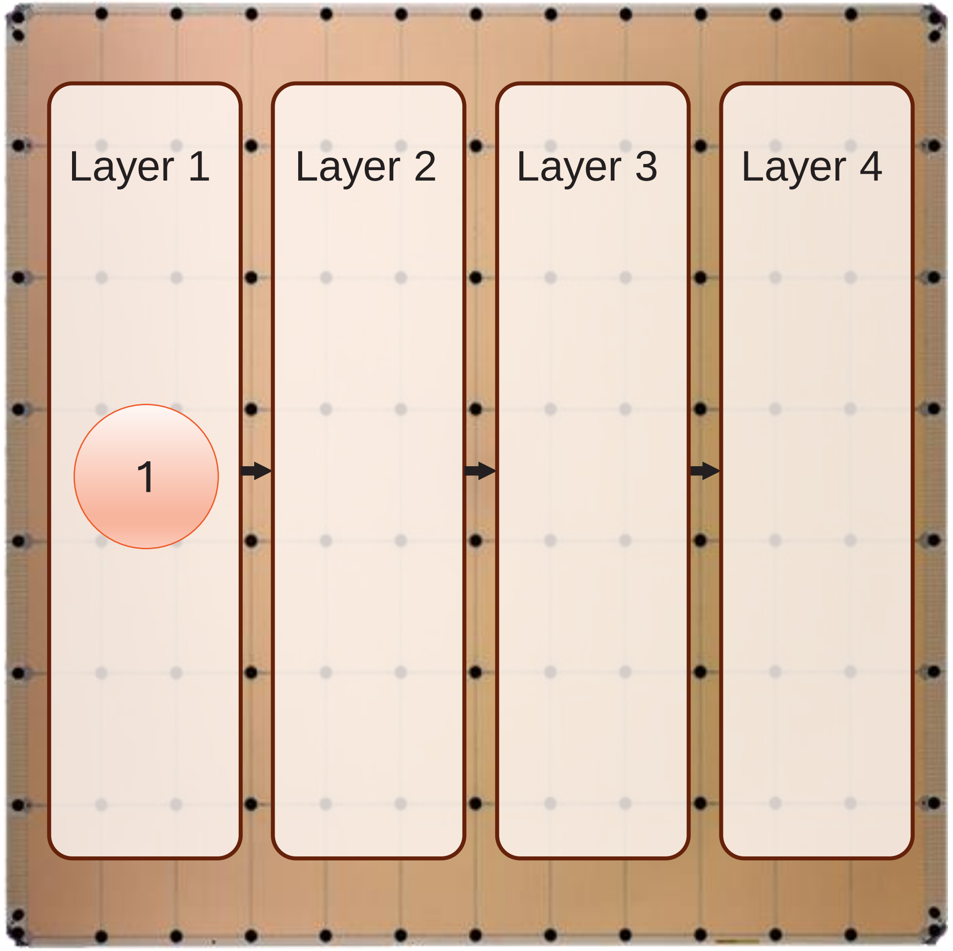
GPUs cannot play here

Source: Nvidia published Nemo results
https://docs.nvidia.com/nemo-framework/user-guide/latest/performance/llama.html

Cerebras enables **low latency and high throughput**

More than enough memory bandwidth for single user enables high multi-user throughput

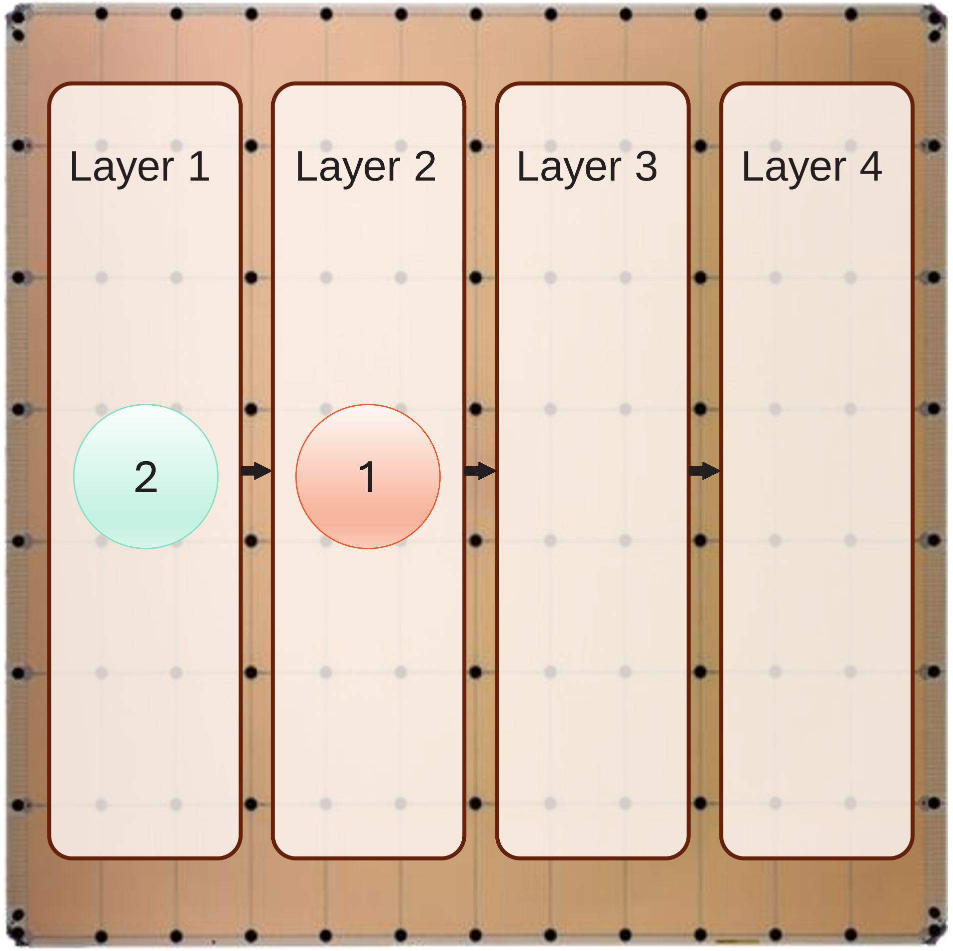
- Single user is using only a fraction of the total memory bandwidth
- We can use the extra bandwidth to support multiple users



Cerebras enables **low latency and high throughput**

More than enough memory bandwidth for single user enables high multi-user throughput

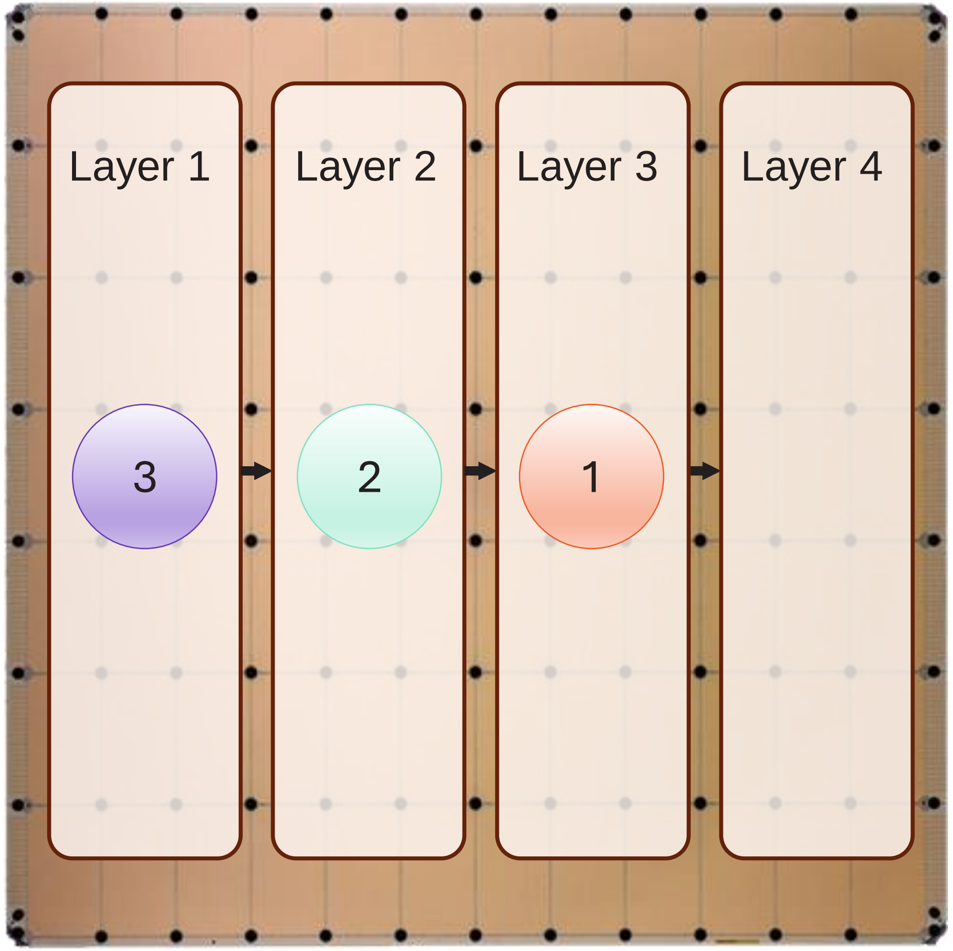
- Single user is using only a fraction of the total memory bandwidth
- We can use the extra bandwidth to support multiple users
- Additional users all run in parallel
- Each accessing the model simultaneously



Cerebras enables **low latency and high throughput**

More than enough memory bandwidth for single user enables high multi-user throughput

- Single user is using only a fraction of the total memory bandwidth
- We can use the extra bandwidth to support multiple users
- Additional users all run in parallel
- Each accessing the model simultaneously
- Every user gets full performance
- All pipe stages to run at the same time

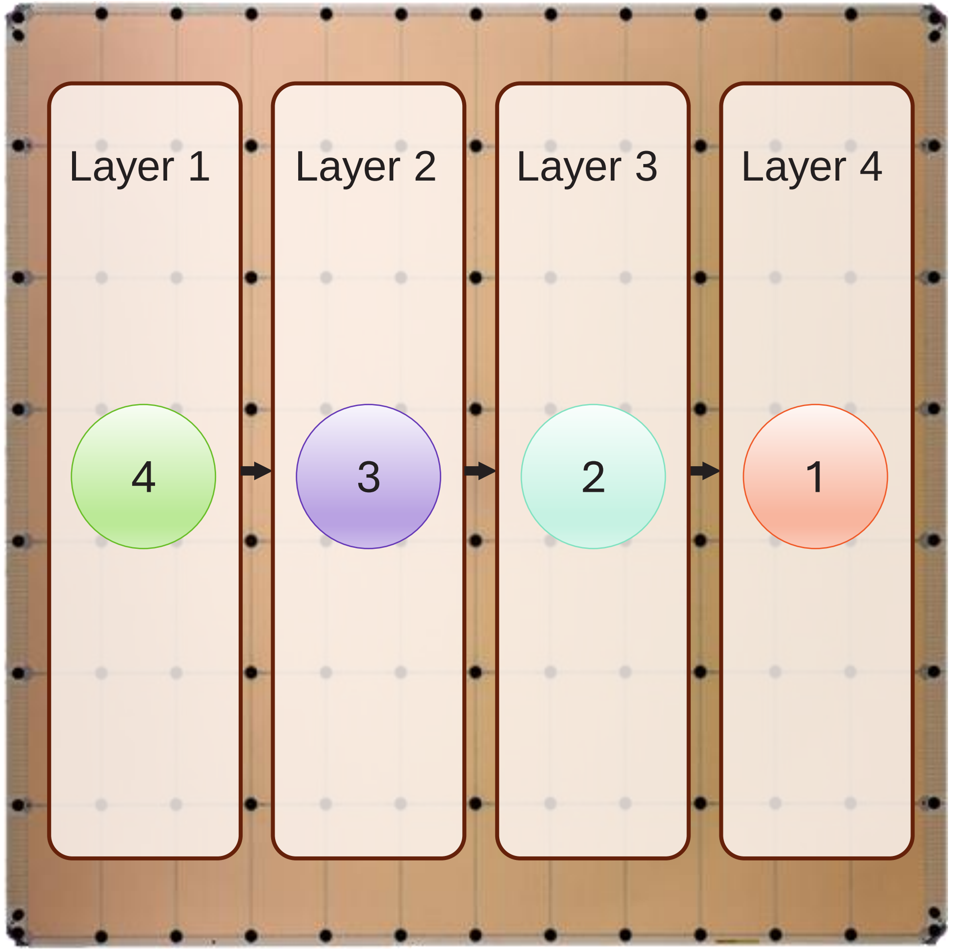


Cerebras enables **low latency and high throughput**

More than enough memory bandwidth for single user enables high multi-user throughput

- Single user is using only a fraction of the total memory bandwidth
- We can use the extra bandwidth to support multiple users
- Additional users all run in parallel
- Each accessing the model simultaneously
- Every user gets full performance
- All pipe stages to run at the same time

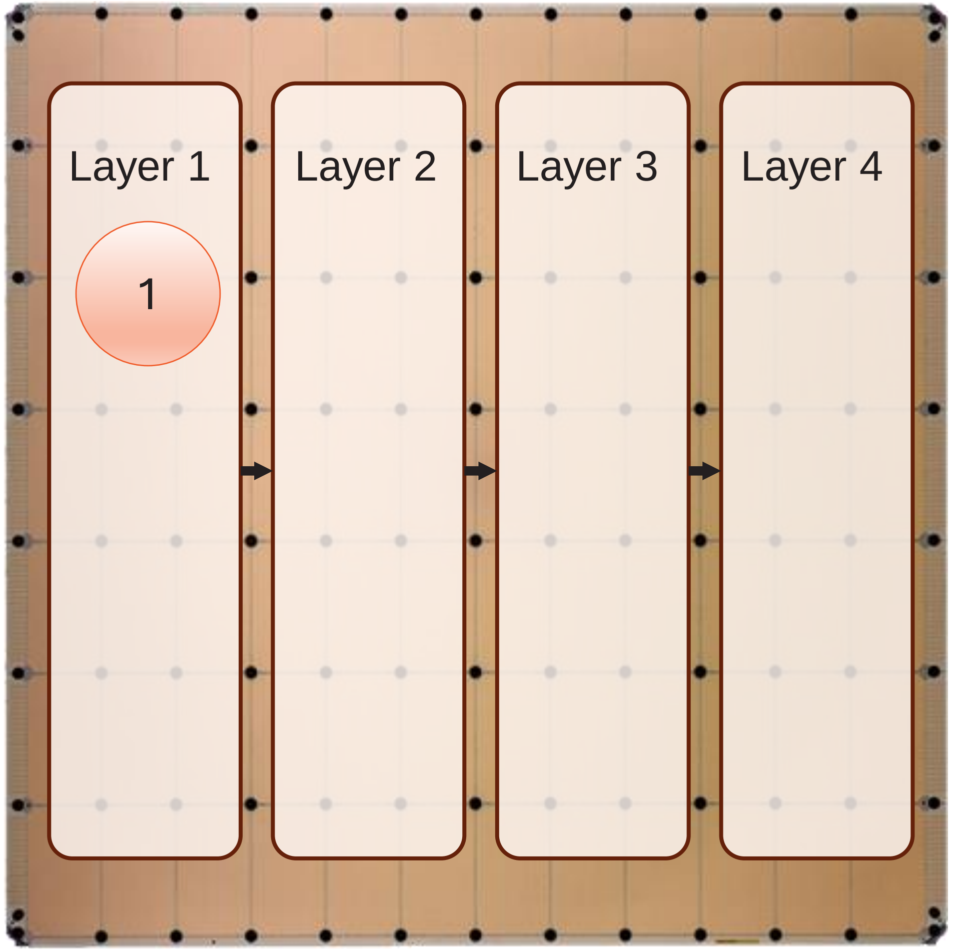
Full pipeline model parallelism on single chip



High performance prompt processing

Prompt tokens can be processed in parallel

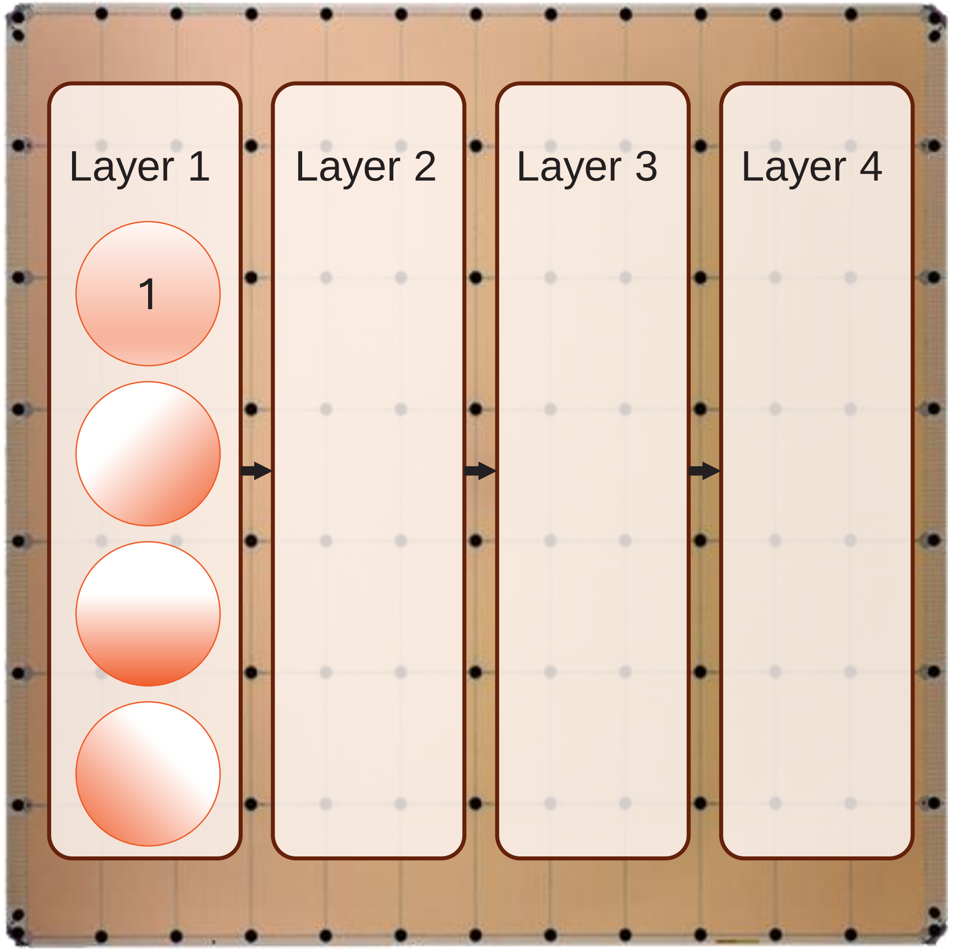
- Prompt is known upfront so no need to wait for full pipeline execution for next token



High performance prompt processing

Prompt tokens can be processed in parallel

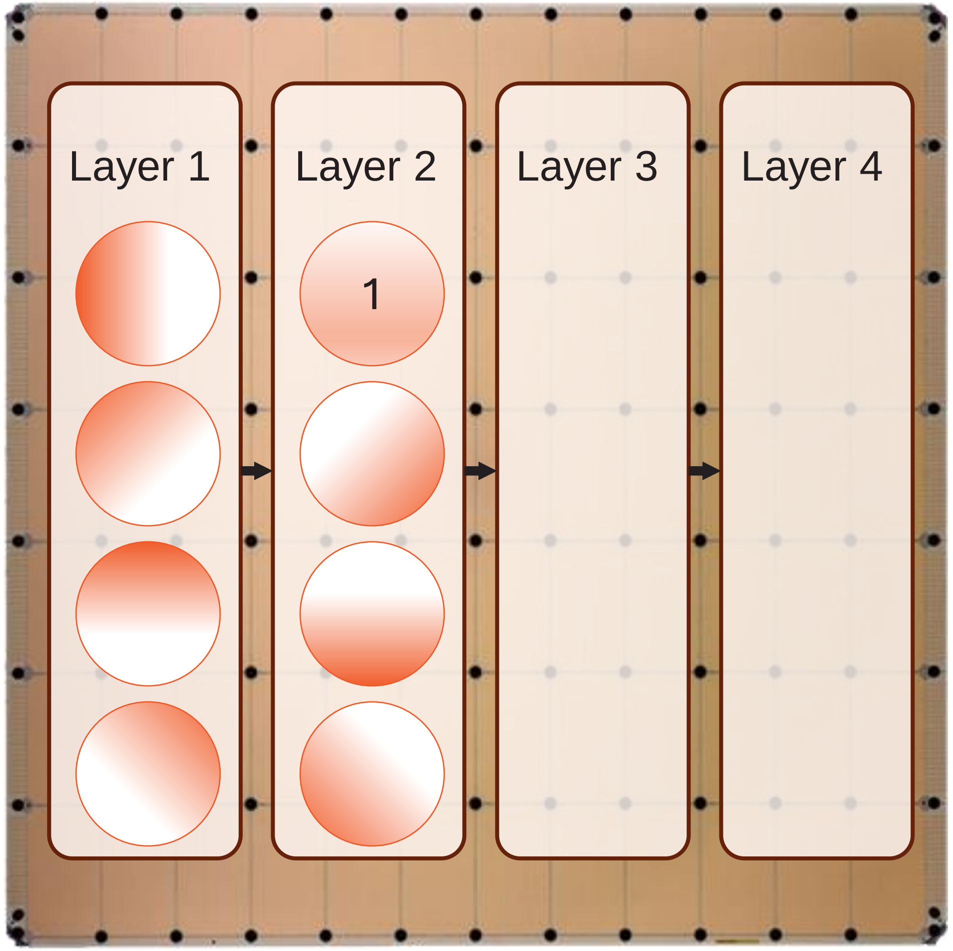
- Prompt is known upfront so no need to wait for full pipeline execution for next token
- Run multiple single user prompt tokens in the same pipe stages



High performance prompt processing

Prompt tokens can be processed in parallel

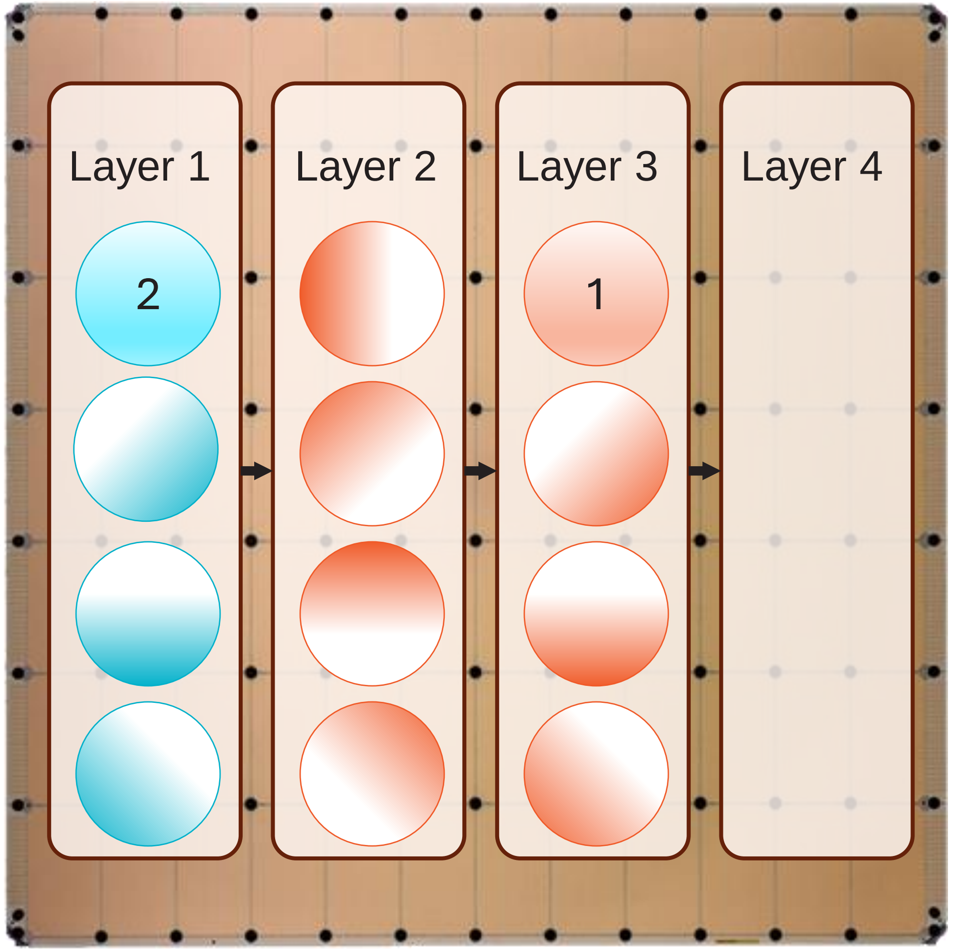
- Prompt is known upfront so no need to wait for full pipeline execution for next token
- Run multiple single user prompt tokens in the same pipe stages
- Run multiple single user prompt tokens in multiple pipe stages



High performance prompt processing

Prompt tokens can be processed in parallel

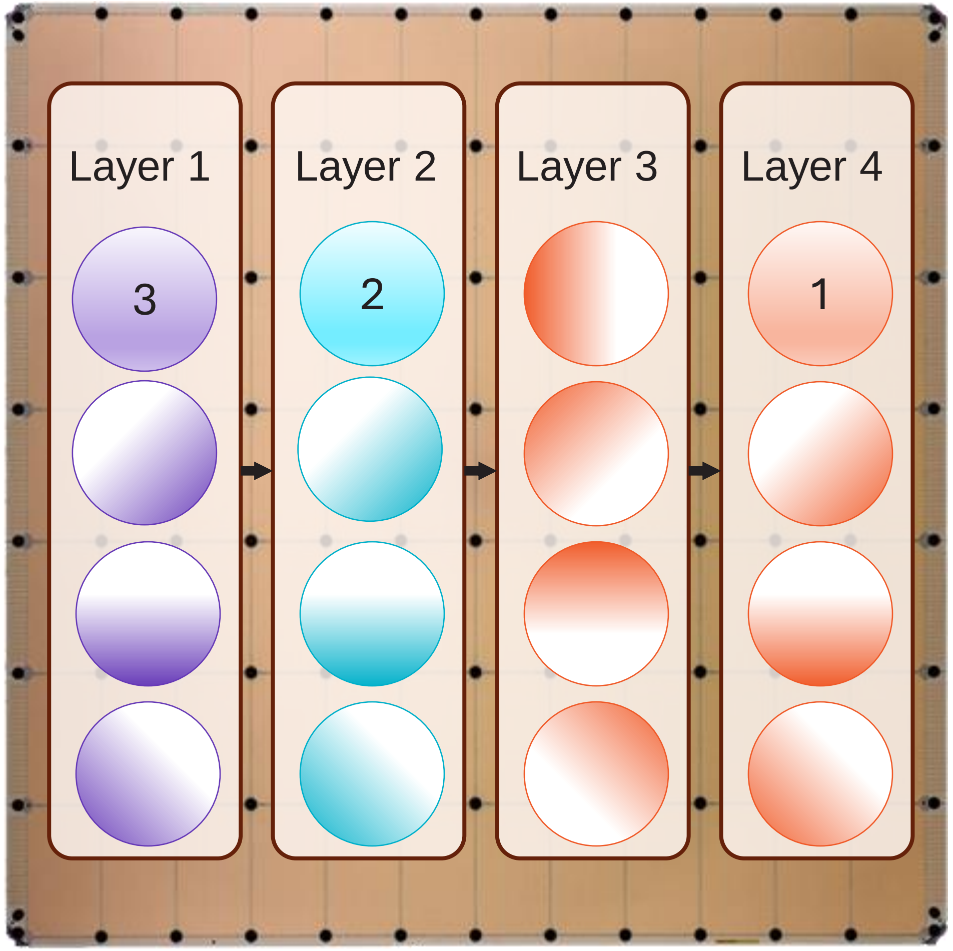
- Prompt is known upfront so no need to wait for full pipeline execution for next token
- Run multiple single user prompt tokens in the same pipe stages
- Run multiple single user prompt tokens in multiple pipe stages
- Can achieve higher single user prompt speed using empty pipe stages when fewer users



High performance prompt processing

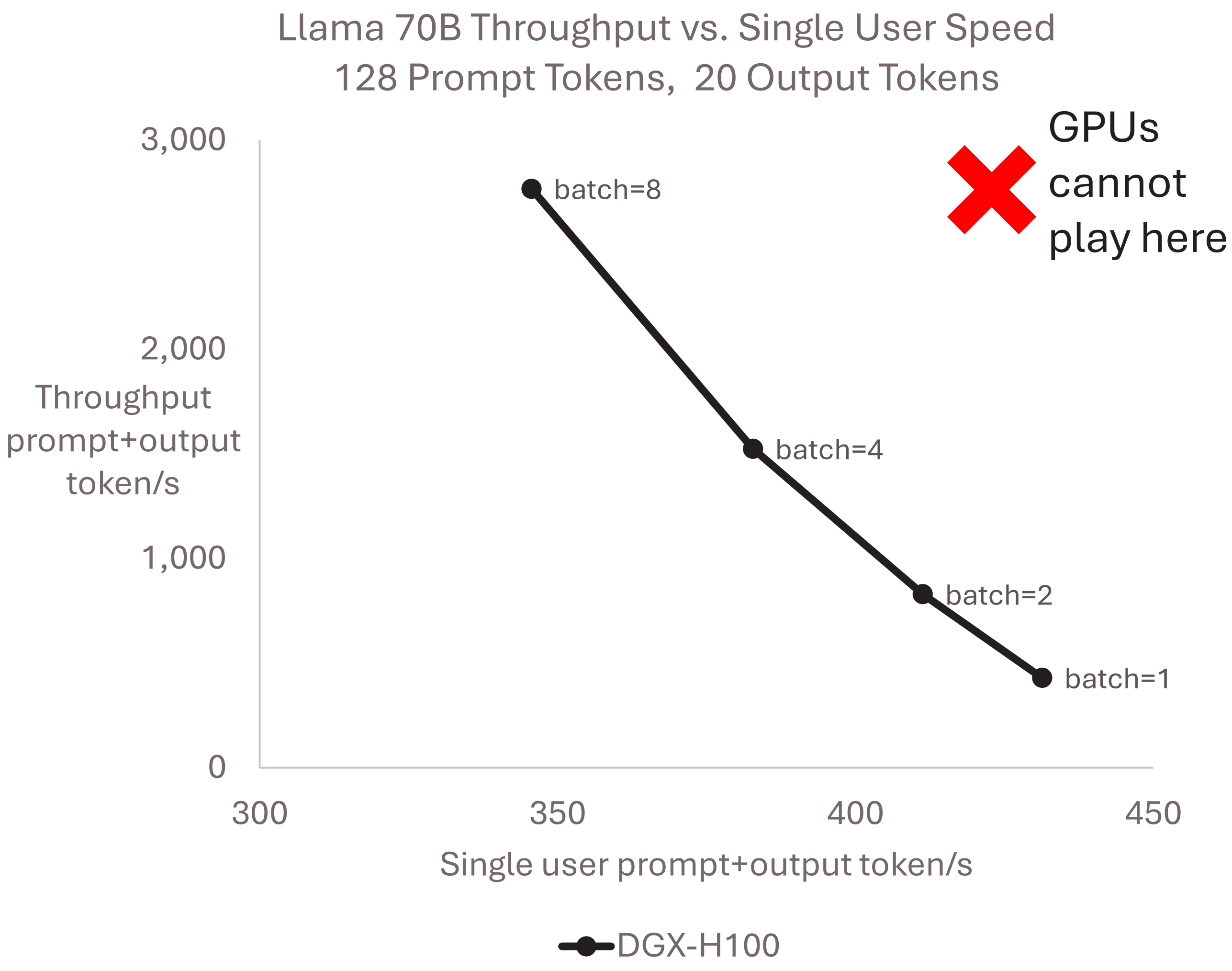
Prompt tokens can be processed in parallel

- Prompt is known upfront so no need to wait for full pipeline execution for next token
- Run multiple single user prompt tokens in the same pipe stages
- Run multiple single user prompt tokens in multiple pipe stages
- Can achieve higher single user prompt speed using empty pipe stages when fewer users



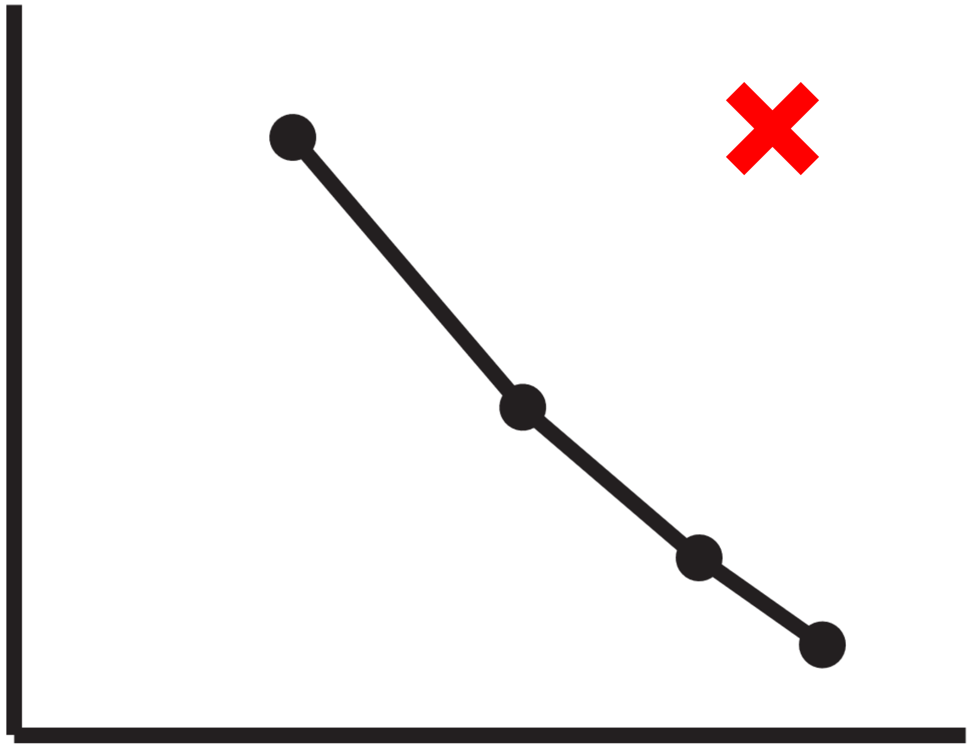
Flexible faster prompt processing driving maximum throughput

The latency vs. throughput tradeoff



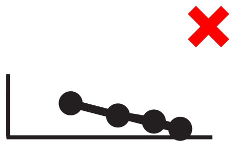
Where is Cerebras on this graph?

Zoom out 10x because Cerebras is so much faster...



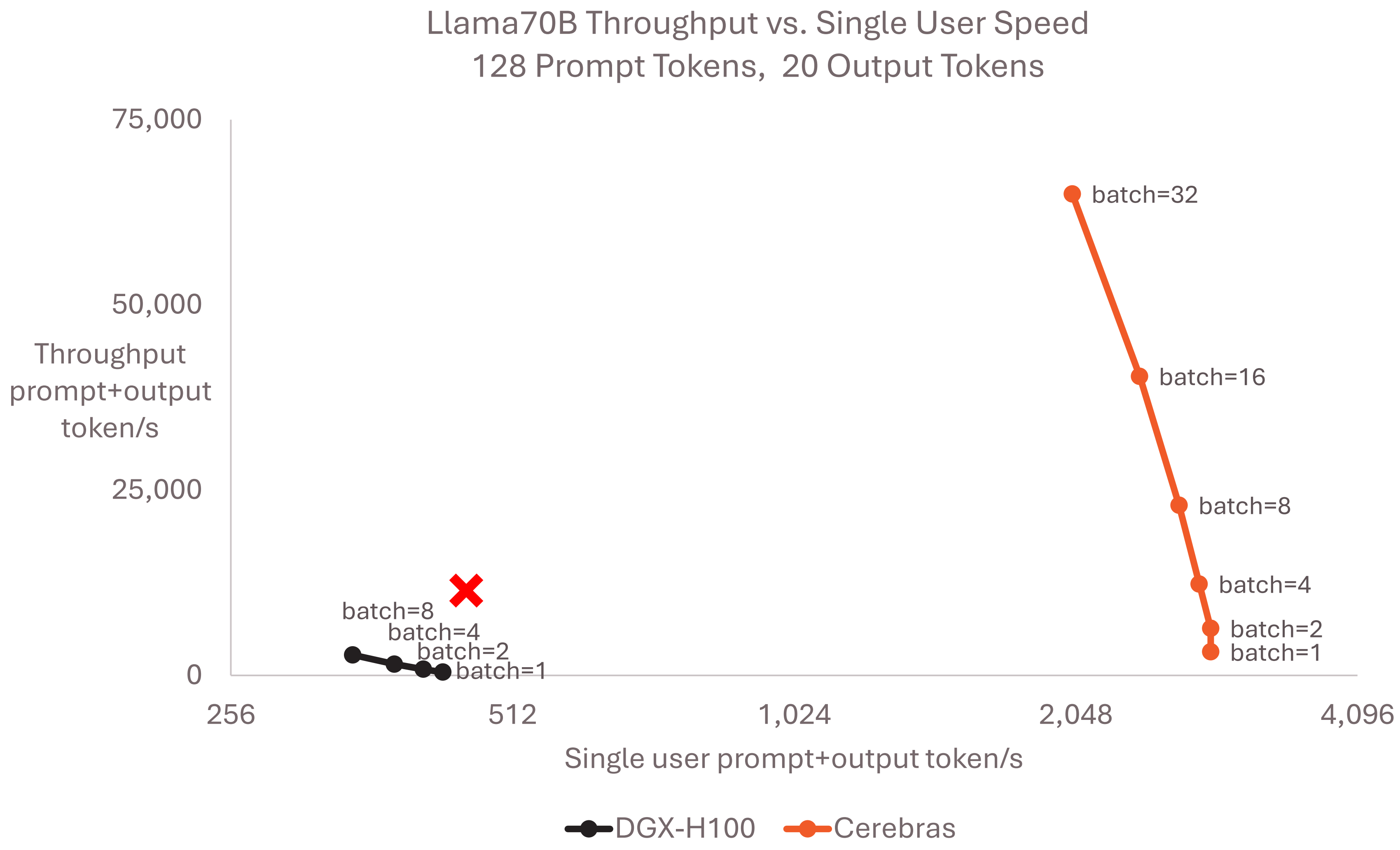
Where is Cerebras on this graph?

Zoom out 10x because Cerebras is so much faster...



Where is Cerebras on this graph?

Wafer-scale architecture enables highest single user speed **and** high throughput



With these throughput techniques, we expect

20-40x higher throughput

5-20x single user speed

lower cost per token

Source: Nvidia published Nemo results
https://docs.nvidia.com/nemo-framework/user-guide/latest/performance/llama.html

And this is just the beginning...

We are already working on many techniques to improve further:

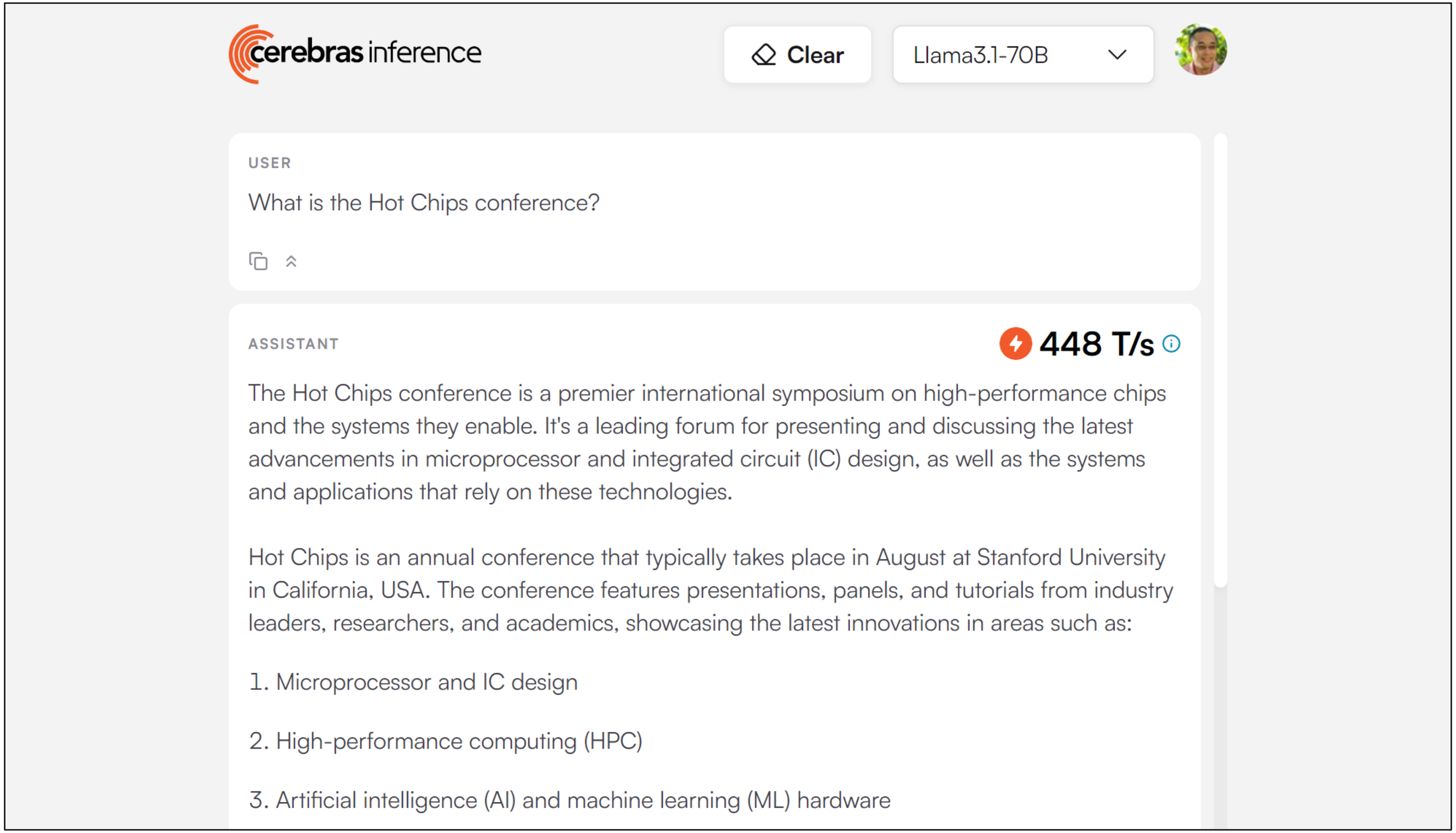
Technique	Speed	Footprint & Throughput
Speculative decoding	✓	
KV cache optimizations		✓
Quantization	✓	✓
Sparsity	✓	✓
More to come...	✓	✓

We are continually improving performance and supporting larger models everyday

Cerebras Inference Service

Cerebras Inference Service Launching Today!

Go to **inference.cerebras.ai** to try it out!



USER

What is the Hot Chips conference?

ASSISTANT

448 T/s

The Hot Chips conference is a premier international symposium on high-performance chips and the systems they enable. It's a leading forum for presenting and discussing the latest advancements in microprocessor and integrated circuit (IC) design, as well as the systems and applications that rely on these technologies.

Hot Chips is an annual conference that typically takes place in August at Stanford University in California, USA. The conference features presentations, panels, and tutorials from industry leaders, researchers, and academics, showcasing the latest innovations in areas such as:

1. Microprocessor and IC design
2. High-performance computing (HPC)
3. Artificial intelligence (AI) and machine learning (ML) hardware

Cerebras Inference Service Launching Today!

Llama3.1-8B

1,800 tokens/s

Free tier
30 requests/m
1M daily token limit

Paid tier
10¢ per M tokens

Llama3.1-70B

450 tokens/s

Free tier
30 requests/m
1M daily token limit

Paid tier
60¢ per M tokens

Coming Soon

Llama-405B
Mistral Large 2
Cohere Command R
Whisper
Custom fine-tunes

inference.cerebras.ai

to experience the GPU impossible

Thank you