Microsoft

Compute Scaling for AI

Fan Yang

Systems Research Group

Microsoft Research Aisa

In collaboration with







Imperial College London

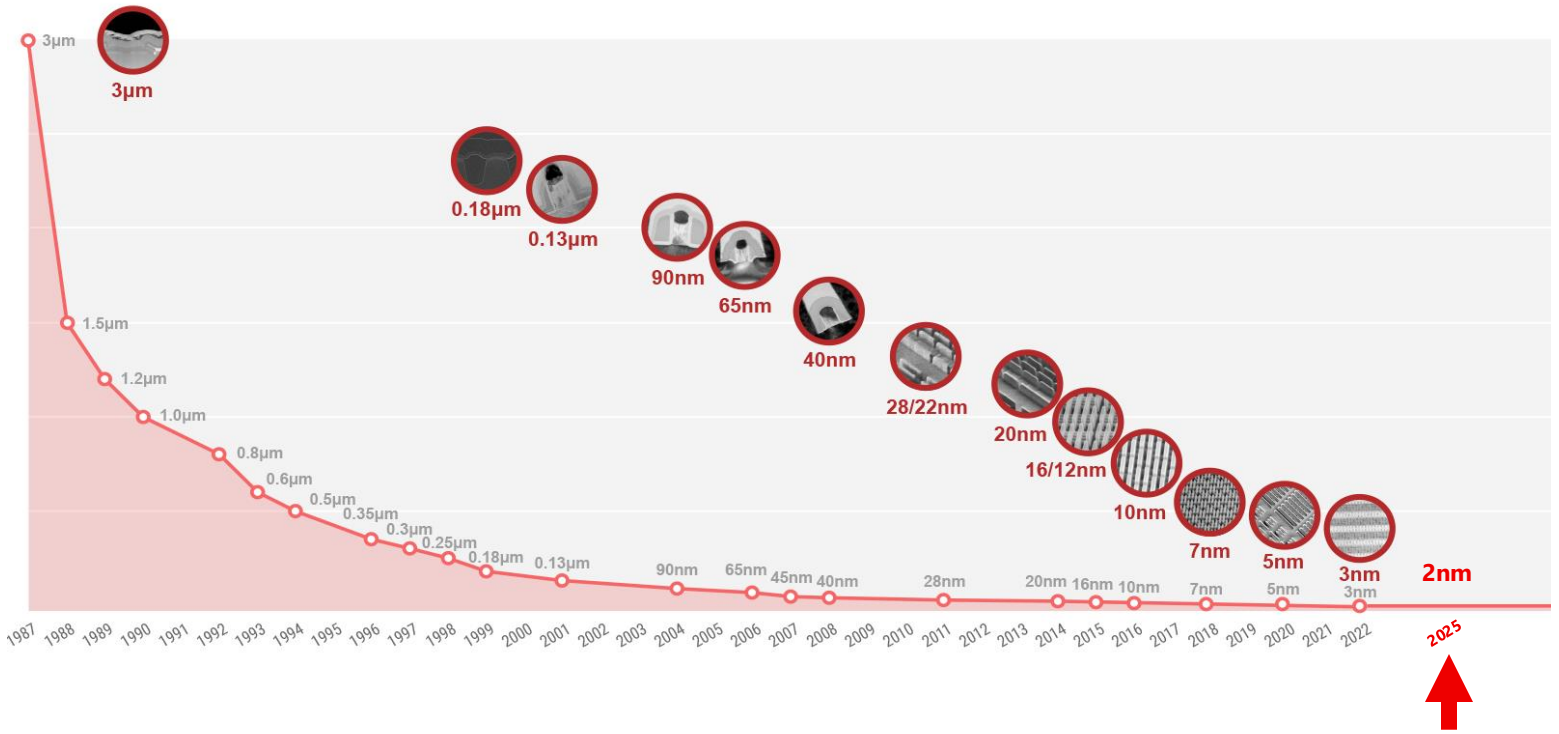
The Key to AI Success: *Scaling Laws*

Four dimensions

- Model – size
- Data – volume
- Time – test-time scaling
- Compute – foundation

Compute Scaling Faces Fundamental Challenges

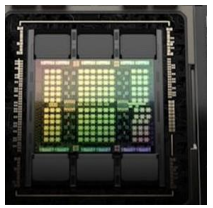
- Semiconductor nodes approaching physical limits



Source: TSMC

An Arms Race for Larger, More Complicated AI Chips

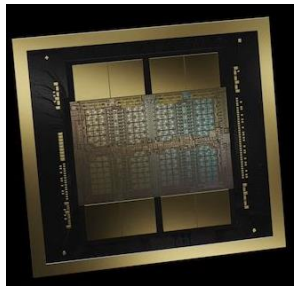
NVIDIA H100



Single-Die
814 mm^2

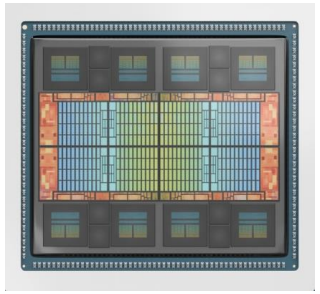
Larger chips

NVIDIA B200



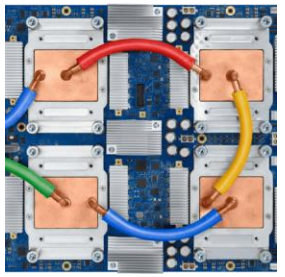
Multi-Die, 1600 mm^2

AMD MI300

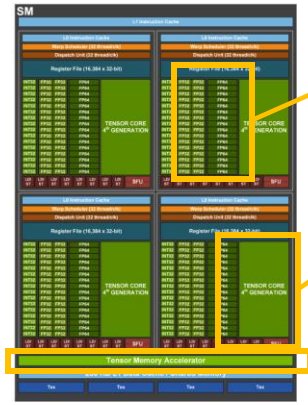


Multi-Die, 1017 mm^2

Google TPU v3



Multi-Chip, 4 x 700 mm^2



GPU architecture

CUDA core: general-purpose computation

TensorCore: matrix multiplication /w wider tile shape
(MMA → WMMA → WGMMA)

TMA: async memory load & store

More complex hardware

Problem – Design & Manufacturing Difficulties

- Higher chances of design flaws and/or manufacture defects

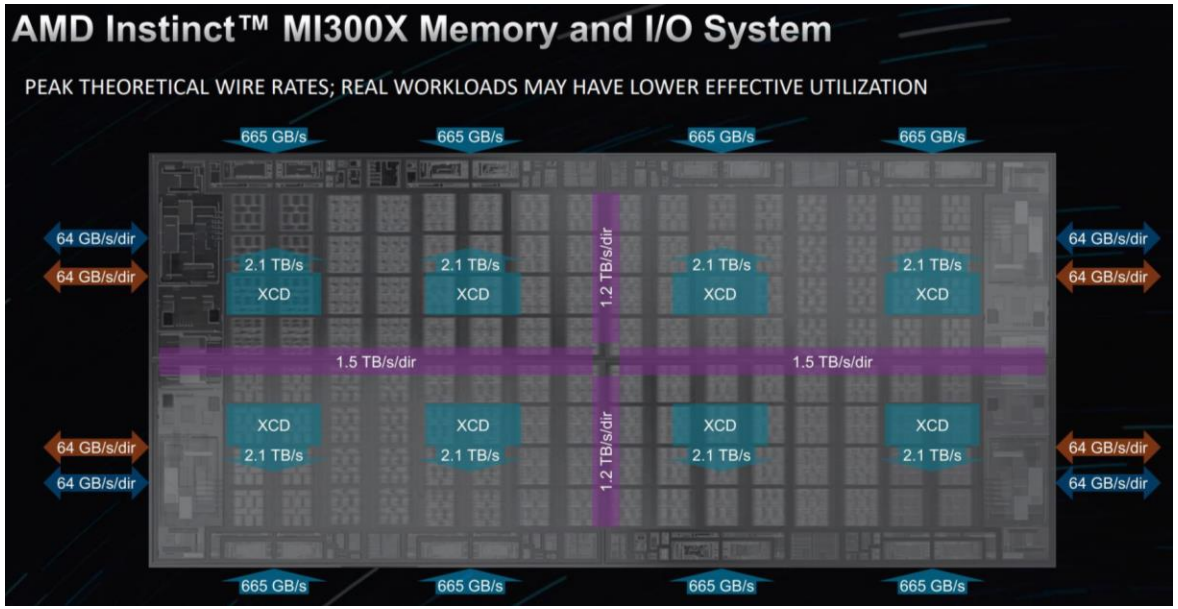
Mon 5 Aug 2024 // 13:23 UTC

UPDATED Nvidia is understood to be delaying shipments of its Blackwell GPUs until the first quarter of 2025, and it appears the problems may be due to the complexity of the chip-on-wafer-on-substrate (CoWoS) packaging tech that TSMC is using to manufacture the next-gen hardware.

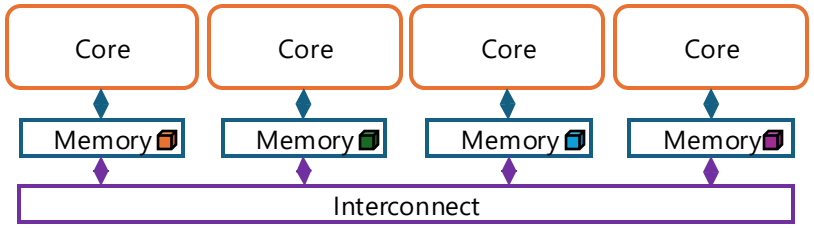
2025 Jan 13 (Reuters) - Nvidia's (NVDA.O) top customers are delaying orders of the AI chip leader's latest 'Blackwell' racks due to overheating issues, the Information reported on Monday.

Problem – Addressing the Uniqueness of Large Chips

- Example: harder to hide non-uniformity in larger chips



Source: AMD



AI workloads were predominantly optimized for chips with shared memory architectures

A Path Forward Driven by Simplicity

- A codesign to *simplify* both AI models and AI chips
  - BitNet – a simpler 1-bit LLM (Bitnet.cpp)
  - LUT tensor core – a simplified tensor core for further compute scaling (ISCA'25)

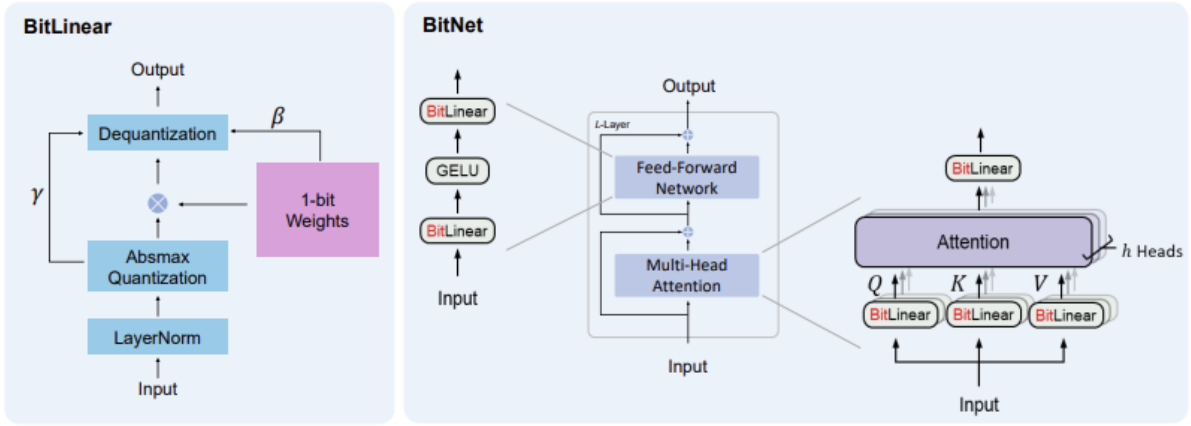
- A stack *modeling* the key properties of modern AI chips
  - WaferLLM – a new system stack for wafer-scale chips (OSDI'25)

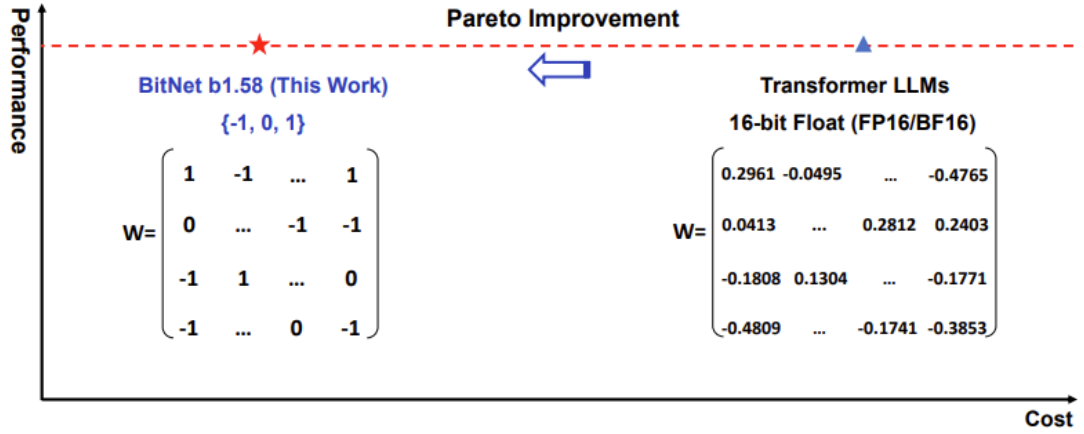
A Path Forward Driven by Simplicity

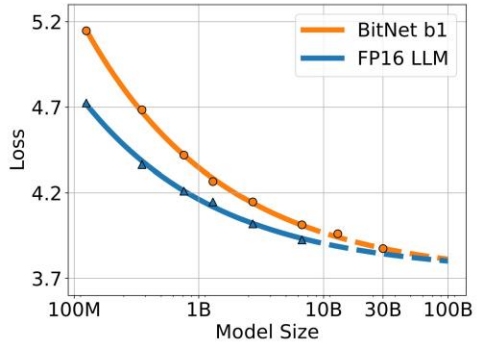
- A codesign to *simplify* both AI models and AI chips
  - **BitNet – a simpler 1-bit LLM** (Bitnet.cpp)
  - LUT tensor core – a simplified tensor core for further compute scaling (ISCA'25)

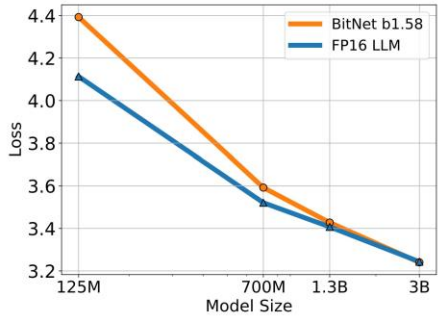
- A stack *modeling* the key properties of modern AI chips
  - WaferLLM – a new system stack for wafer-scale chips (OSDI'25)

The Era of 1-bit LLMs









Credit: BitNet team

1-bit LLMs match full-precision LLMs when the model scale is large enough

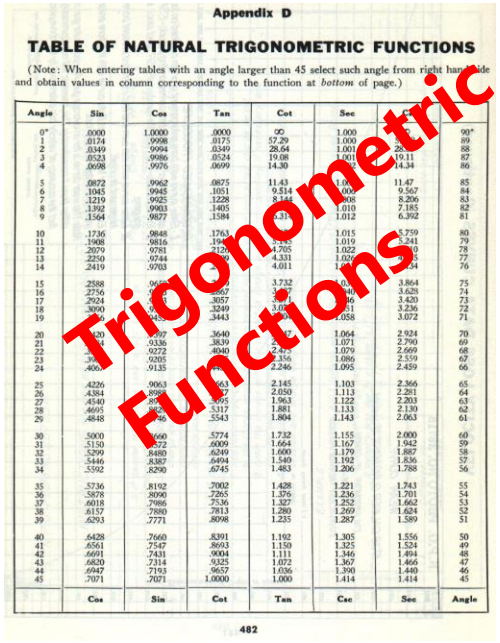
A Path Forward Driven by Simplicity

- A codesign to *simplify* both AI models and AI chips
  - BitNet – a simpler 1-bit LLM (Bitnet.cpp)
  - **LUT tensor core – a simplified tensor core for further compute scaling** (ISCA'25)

- A stack *modeling* the key properties of modern AI chips
  - WaferLLM – a new system stack for wafer-scale chips (OSDI'25)

LUT TensorCore: Compute Scaling by Simplicity

- 1-bit LLMs: replacing ALU w/ LUT (Lookup Table)



Appendix D

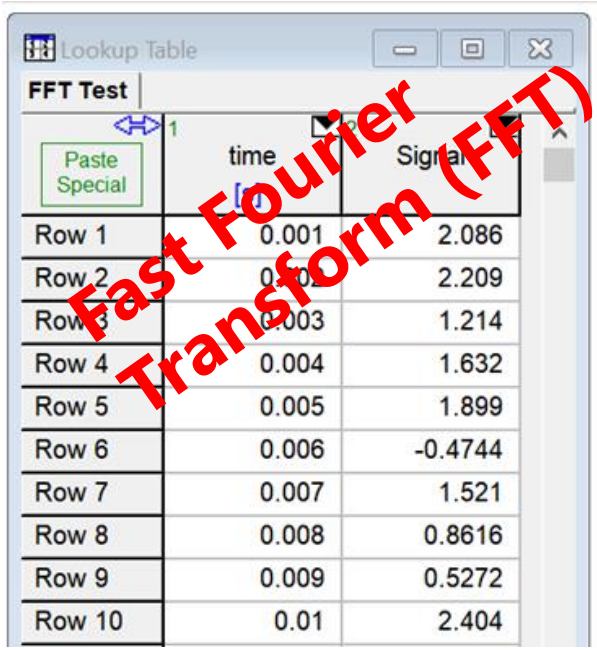
TABLE OF NATURAL TRIGONOMETRIC FUNCTIONS

(Note: When entering tables with an angle larger than 45 select such angle from right hand side and obtain values in column corresponding to the function at *bottom* of page.)

Angle	Sin	Cos	Tan	Cot	Sec	Csc	
0°	.0000	1.0000	.0000	∞	1.000	∞	90°
1	.0174	.9998	.0175	57.29	1.000	57.30	89
2	.0349	.9994	.0349	28.64	1.001	28.65	88
3	.0523	.9986	.0524	19.08	1.001	19.11	87
4	.0698	.9976	.0699	14.30	1.002	14.34	86
5	.0872	.9962	.0875	11.43	1.004	11.47	85
6	.1045	.9945	.1051	9.514	1.006	9.567	84
7	.1219	.9925	.1228	8.144	1.008	8.206	83
8	.1392	.9903	.1405	7.115	1.010	7.185	82
9	.1564	.9877	.1584	6.314	1.012	6.392	81
10	.1736	.9848	.1763	5.671	1.015	5.759	80
11	.1908	.9816	.1944	5.145	1.019	5.241	79
12	.2079	.9781	.2126	4.705	1.022	4.810	78
13	.2250	.9744	.2309	4.331	1.026	4.445	77
14	.2419	.9703	.2493	4.011	1.031	4.134	76
15	.2588	.9659	.2679	3.732	1.035	3.864	75
16	.2756	.9613	.2867	3.487	1.040	3.628	74
17	.2924	.9563	.3057	3.271	1.046	3.420	73
18	.3090	.9511	.3249	3.078	1.051	3.236	72
19	.3256	.9455	.3443	2.904	1.058	3.072	71
20	.3420	.9397	.3640	2.747	1.064	2.924	70
21	.3584	.9336	.3839	2.605	1.071	2.790	69
22	.3746	.9272	.4040	2.475	1.079	2.669	68
23	.3907	.9205	.4245	2.356	1.086	2.559	67
24	.4067	.9135	.4452	2.246	1.095	2.459	66
25	.4226	.9063	.4663	2.145	1.103	2.366	65
26	.4384	.8988	.4877	2.050	1.113	2.281	64
27	.4540	.8910	.5095	1.963	1.122	2.203	63
28	.4695	.8829	.5317	1.881	1.133	2.130	62
29	.4848	.8746	.5543	1.804	1.143	2.063	61
30	.5000	.8660	.5774	1.732	1.155	2.000	60
31	.5150	.8572	.6009	1.664	1.167	1.942	59
32	.5299	.8480	.6249	1.600	1.179	1.887	58
33	.5446	.8387	.6494	1.540	1.192	1.836	57
34	.5592	.8290	.6745	1.483	1.206	1.788	56
35	.5736	.8192	.7002	1.428	1.221	1.743	55
36	.5878	.8090	.7265	1.376	1.236	1.701	54
37	.6018	.7986	.7536	1.327	1.252	1.662	53
38	.6157	.7880	.7813	1.280	1.269	1.624	52
39	.6293	.7771	.8098	1.235	1.287	1.589	51
40	.6428	.7660	.8391	1.192	1.305	1.556	50
41	.6561	.7547	.8693	1.150	1.325	1.524	49
42	.6691	.7431	.9004	1.111	1.346	1.494	48
43	.6820	.7314	.9325	1.072	1.367	1.466	47
44	.6947	.7193	.9657	1.036	1.390	1.440	46
45	.7071	.7071	1.0000	1.000	1.414	1.414	45
	Cos	Sin	Cot	Tan	Csc	Sec	Angle

482

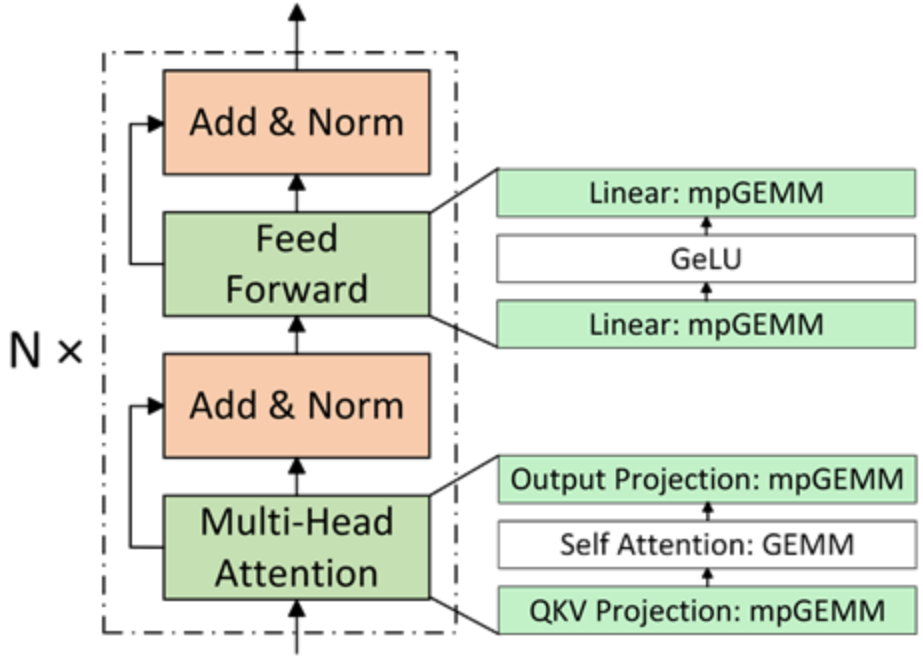
Trigonometric Functions



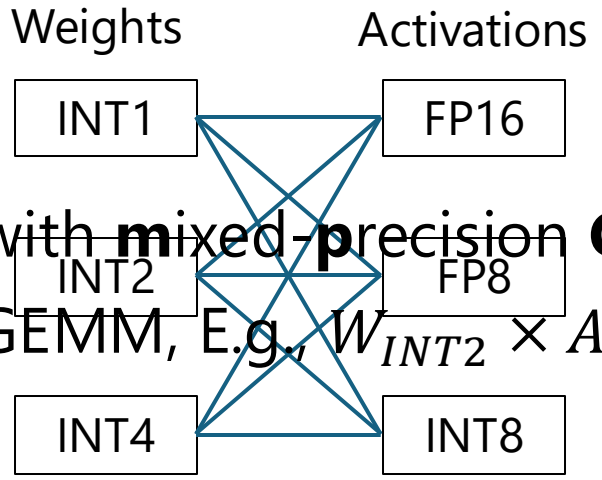
Fast Fourier Transform (FFT)

When compute couldn't scale, scientists transform them into tables/codebooks

The Inconvenience – Not Entirely 1-Bit



A transformer layer



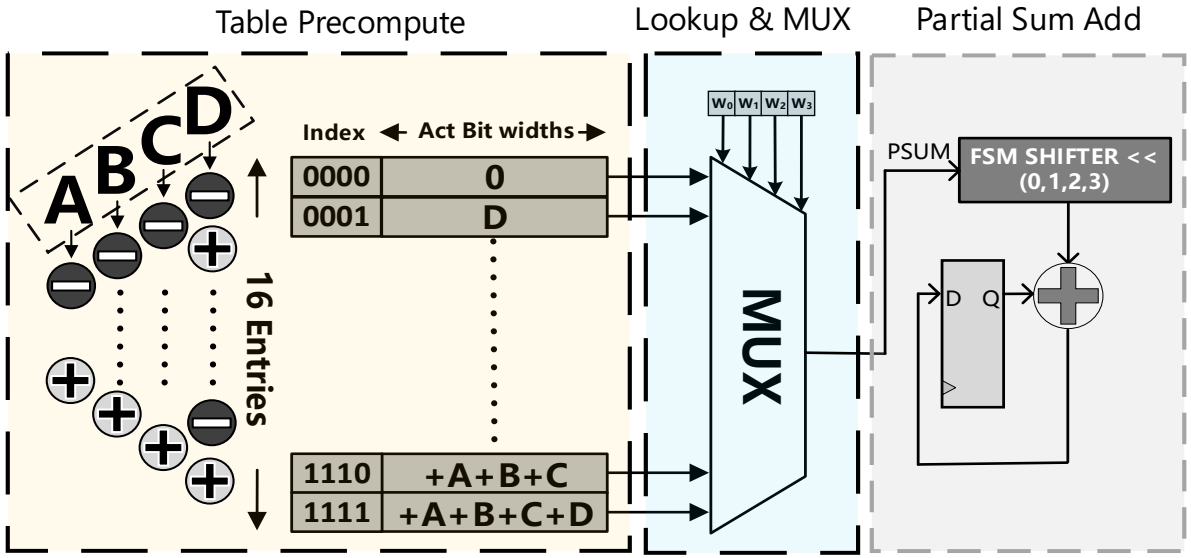
LLMs with **m**ixed-**p**recision **GEMM**

(mpGEMM, E.g., $W_{INT2} \times A_{FP16}$)

Different Precision Combination

The Inconvenience – Table Overhead

- Table size still non-negligible : e.g., $W_{INT2} \times A_{FP16}$
- Table precompute overhead



LUT-based dot product

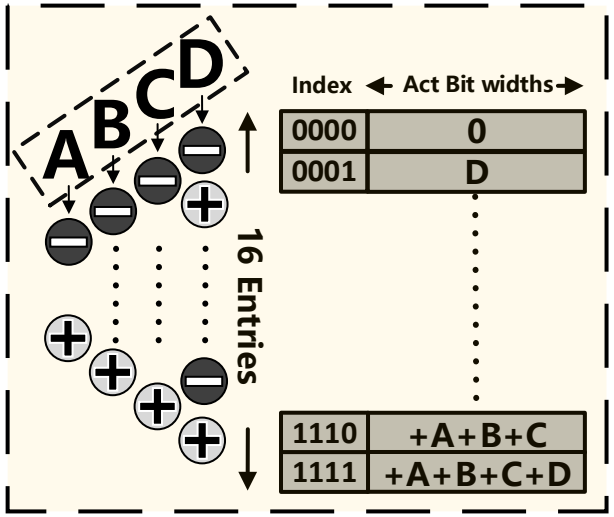
A Software-Hardware Codesign

Large Table Size

	000	001	010	011	100	101	110	111
000	000000	000000	000000	000000	000000	000000	000000	000000
001	000000	000001	000010	000011	000100	000101	000110	000111
010	000000	000010	000100	000110	001000	001010	001100	001110
011	000000	000011	000110	001001	001100	001111	010010	010101
100	000000	000100	001000	001100	010000	010100	011000	011100
101	000000	000101	001010	001111	010100	011001	011110	100011
110	000000	000110	001100	010010	011000	011110	100100	101010
111	000000	000111	001110	010101	011100	100011	101010	110001

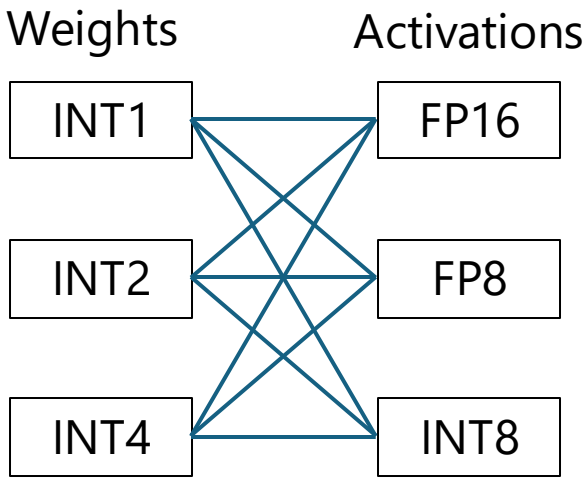
Table Symmetrization

Table Precompute Overhead



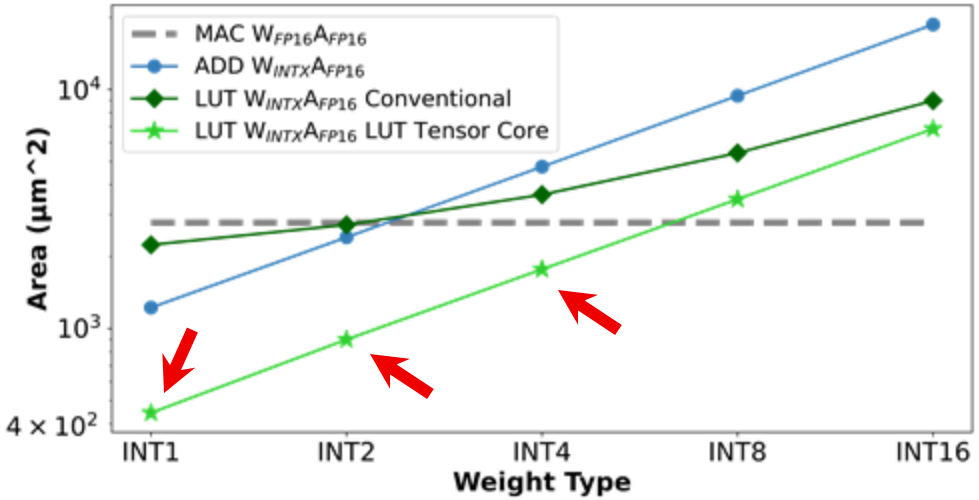
Dedup & Fusion (Ladder, OSDI'24)

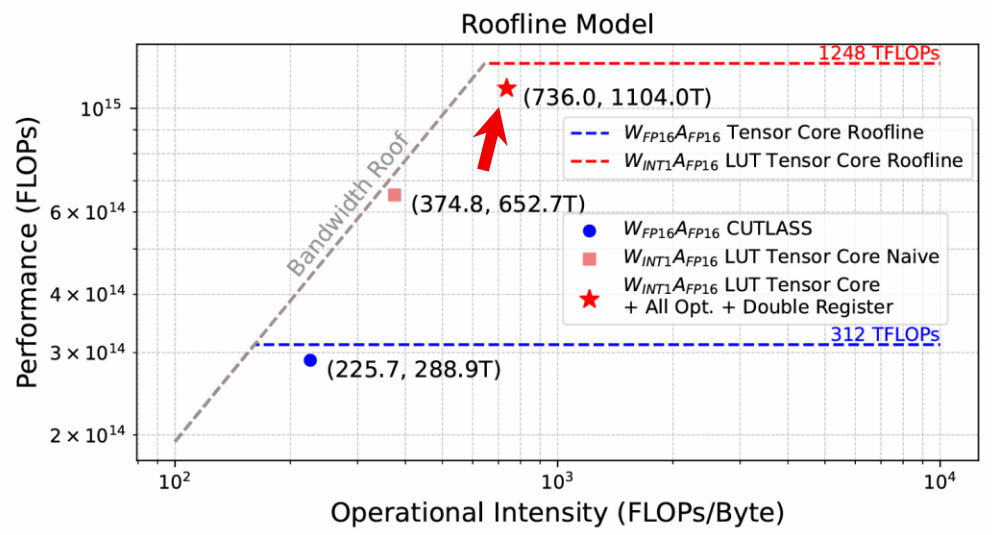
Different Precision Combination



Bit-Serial Circuit

LUT Tensor Core Scales Better at Lower Bits





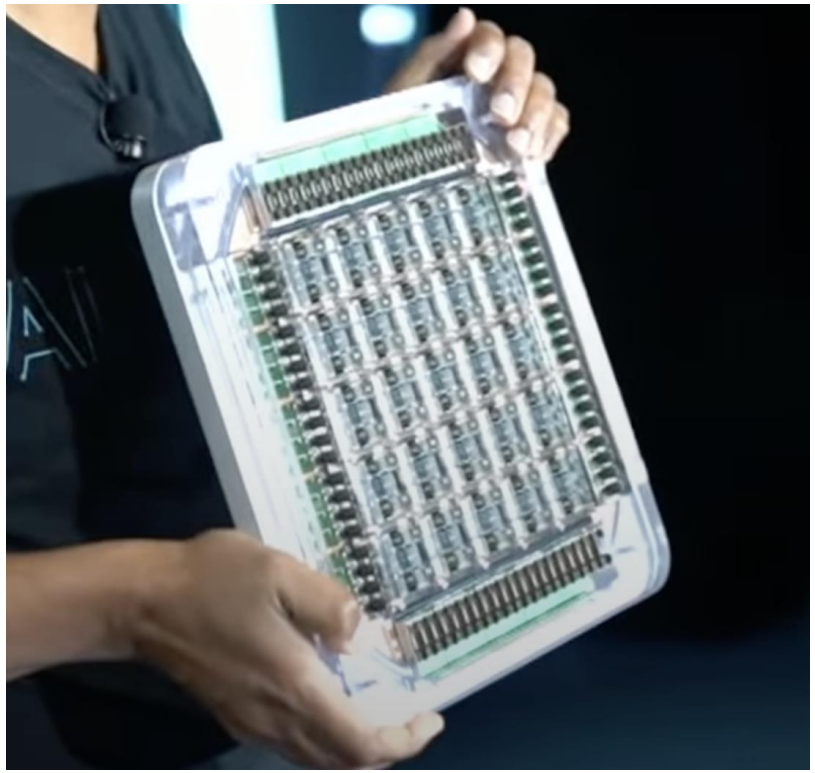
A Path Forward Driven by Simplicity

- A codesign to *simplify* both AI models and AI chips
  - BitNet – a simpler 1-bit LLM (Bitnet.cpp)
  - LUT tensor core – a simplified tensor core for further compute scaling (ISCA'25)

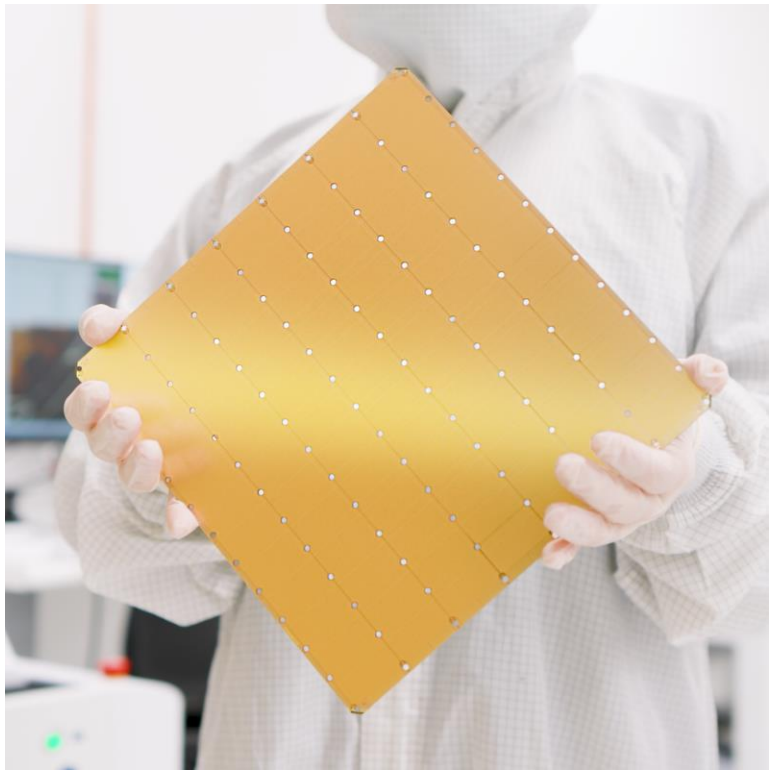
- A stack *modeling* the key properties of modern AI chips
  - **WaferLLM – a new system stack for wafer-scale chips** (OSDI'25)

A System Stack Modeling Modern AI Chips

Case study – Wafer-scale chips



Tesla Dojo



Cerebras WSE-3

NVIDIA B200

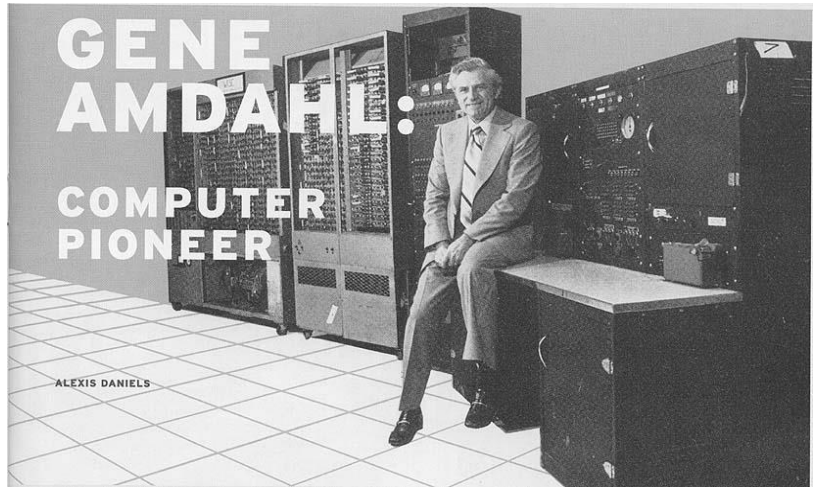
Credit: Cerebras and Tesla

Wafer-Scale is Not a New Dream

- Gene Amdahl shared a similar observation [1]
  - The pioneer of mainframe machines
  - The author of Amdahl's Law

- Amdahl co-founded Trilogy Systems
  - Attempted to design the first wafer-scale chips
  - The biggest investment ($200M) in Silicon Valley in the 1980s

- Trilogy Systems failed due to
  - **Low yields at wafer-scale**
  - **Weak market demand**



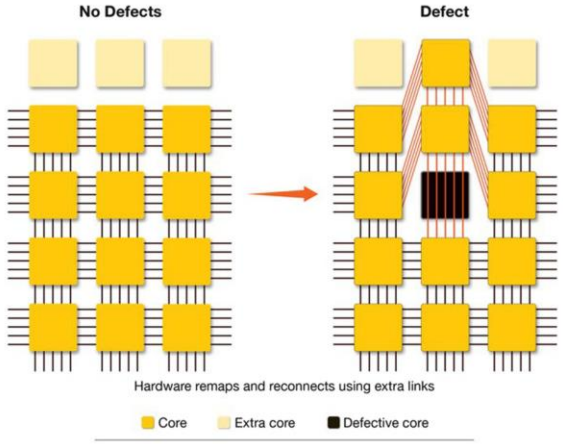
[1] https://spectrum.ieee.org/whats-better-than-40-gpubased-servers-a-server-with-40-gpus

It is about Time (40 Years Later)

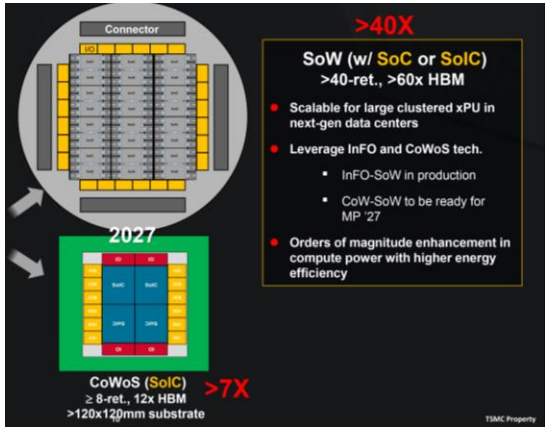
- AI chases extreme efficiency/performance
- Manufacturing improvement
  - Bypass the defective cores with redundant wires

  High yield: 93% core active (WSE-3) vs ***92%*** (NV H100)

- A wave of wafer-scale computers is coming
  - **>40X** compute and bandwidth expected **by 2027** [2]
  - Advanced packaging (CoWoS), 3DIC (TSMC SoIC)



Example of hardware remapping [1]



TSMC Roadmap – System-on-Wafer[2]

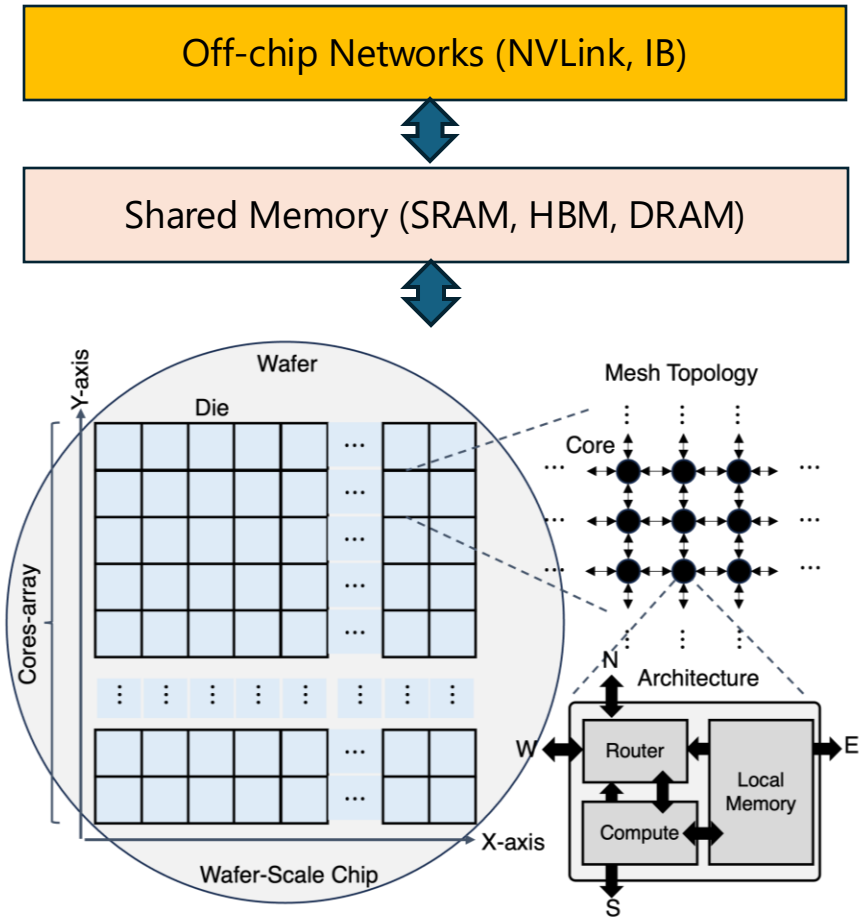
[1] https://www.cerebras.ai/blog/100x-defect-tolerance-how-cerebras-solved-the-yield-problem
[2] https://www.tomshardware.com/tech-industry/tsmc-to-go-3d-with-wafer-sized-processors-cow-sow-system-on-wafer-technology-allows-3d-stacking-for-the-worlds-largest-chips

Wafer-Scale Integration – Better Compute Scaling

	System-on-Die	System-on-Wafer	
Area	Typically 858 mm^2	Typically 73062 mm^2	
#Transistors (TSMC n3)	1 trillion	Up to 91 trillion	
Interconnect	PCB/RDL/SUB/WoW	Wafer	
Die-to-die efficiency	**~10s pJ/bit**	**~0.1s pJ/bit**	~100x
Die-to-die bandwidth	**~ 1-10s TB/s**	**~ 10 - 100s TB/s**	~100x
Memory Bandwidth	**10s TB/s** (crossbar)	**10s PB/s** (aggregated via mesh)	~1,000x
Off-chip memory	10s - 100s GB HBM	10s TB DRAM via Ethernet **10s TB HBM/DRAM via TSMC SoW in 2027**	

- **Emerging wafer-scale systems**: Cerebras, Tesla Dojo, NVIDIA and more reported by TSMC
- **Growing adoption**: Perplexity, Mixtral, Meta AI, G42, …

Systems Ready for Wafer-Scale Chips?



Extensive research on scaling LLM with off-chip networks

- **Topology**: Clos, 3D-Torus
- **System**: Megatron-LM, PyTorch, JAX, TensorFlow, nnScaler
- **Multi-dimensional Parallelism**: TP, PP, DP, EP
- **Communication Operator**: Ring allreduce, All-to-All for MoE

Extensive research on LLM with on-chip shared memory

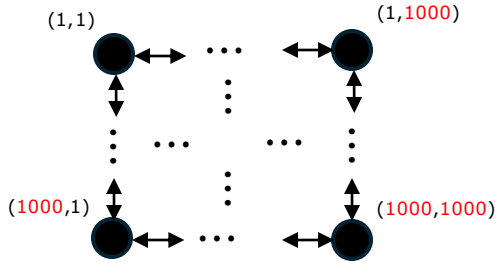
- **Operator**: FlashAttention, MLA, PageAttention,
- **Compiler**: Ladder [OSDI'24], and T10 [SOSP'24]

Wafer-scale AI software remains largely unexplored

- Existing NoC research targets CPUs and small scale (up to 100s)
- Suffer severe communication bottlenecks
- …

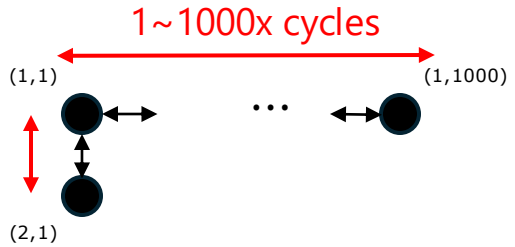
PLMR – A Simplicity-Driven Model for Wafer-Scale Chips

1. Million-scale Parallelism (PLMR)



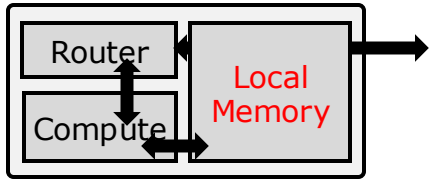
From hundreds of parallelism in a crossbar **to** millions of parallelism in a mesh

2. Highly non-uniform access Latency (PLMR)



From shared memory and small NUMA to large-scale non-uniform memory

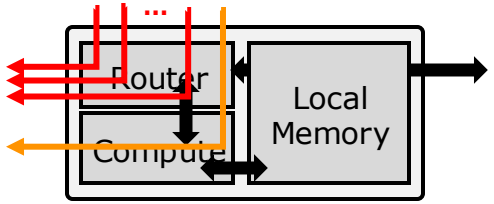
3. Constrained local Memory (PLMR)



100s KB – 1s MB

From coarse-grained tile pipeline to fine-grained tile pipeline

4. Constrained Routing resources (PLMR)



Only support 10s routing entries

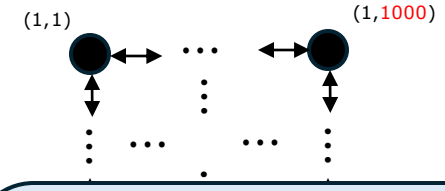
From centralized routing to decentralised NoC routing

Routing on NoC

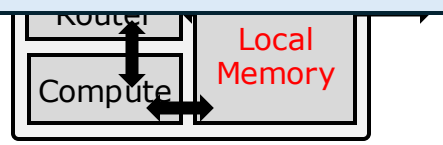
Routing on Compute Engine

PLMR – A Simplicity-Driven Model for Wafer-Scale Chips

1. Million-scale Parallelism (PLMR)



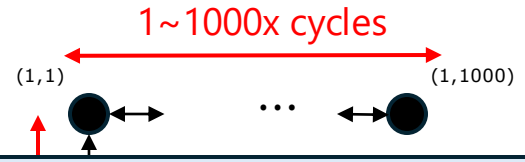
From hundreds of parallelism in a crossbar **to** millions of



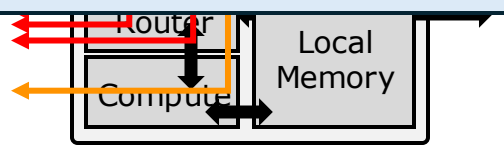
100s KB – 1s MB

From coarse-grained tile pipeline to fine-grained tile pipeline

2. Highly non-uniform access Latency (PLMR)



From shared memory and small NUMA to large-scale non-uniform



Only support 10s routing entries

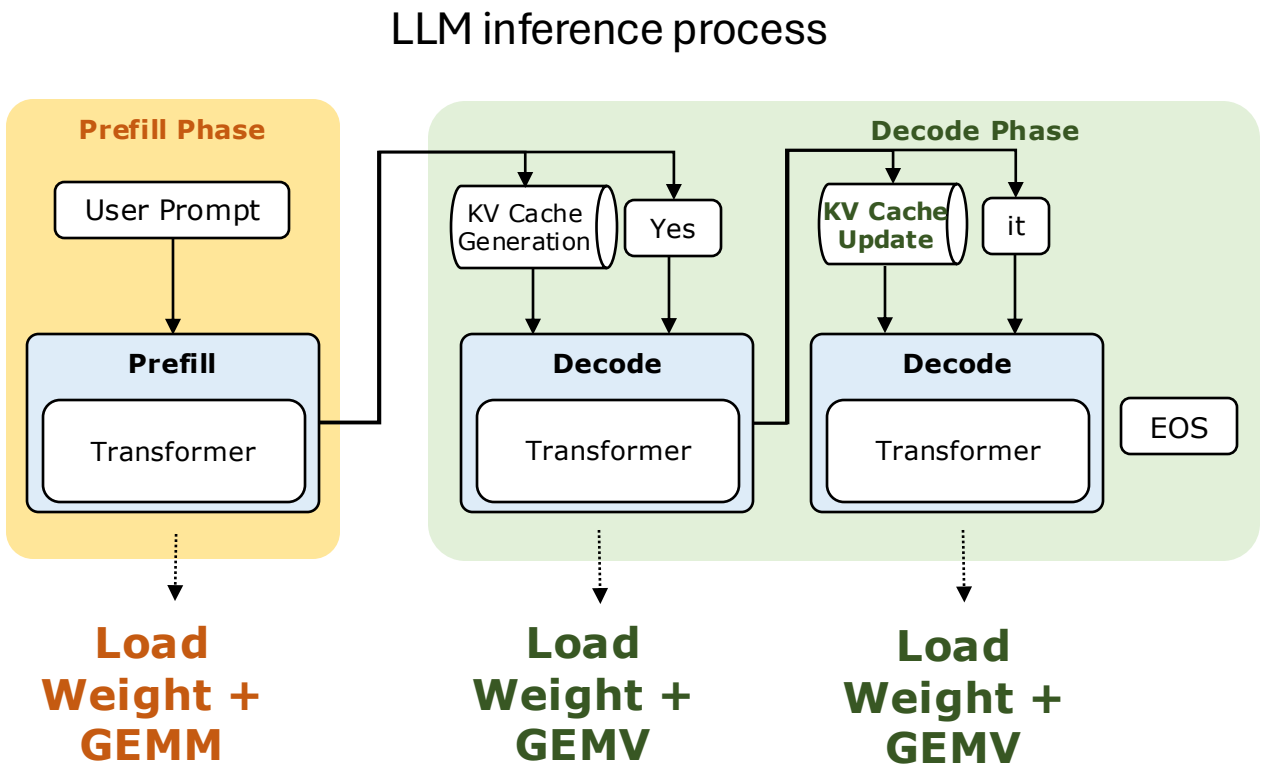
From centralized routing to decentralised NoC routing

Routing on NoC

Routing on Compute Engine

PLMR model - The **key shift** from ***shared-memory architectures*** to ***on-chip large-scale, distributed memory systems***

LLM Inference on Wafer-Scale Chips



PLMR Compliant

WaferLLM: World-First Wafer-Scale LLM Inference System

Goals

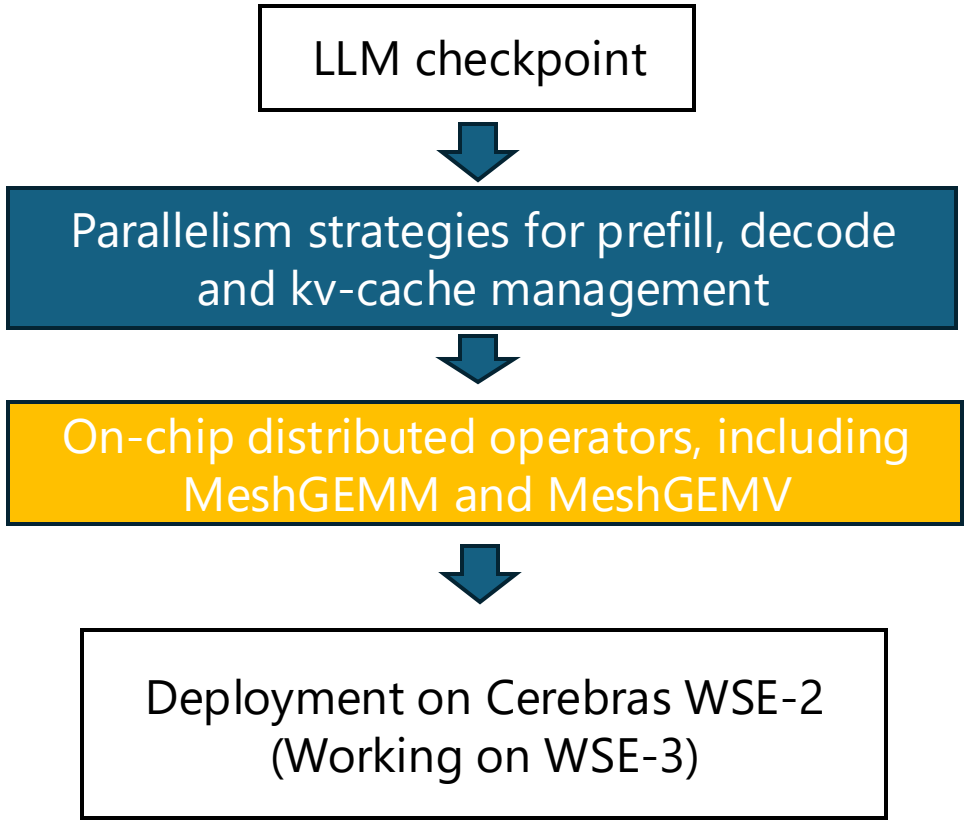
- **Design the entire stack guided by PLMR**
- Generalise across hardware backends

Contributions

- New prefill parallelism strategies
- New decode parallelism strategies
- New KV-cache algorithm – Shift-based update
- New GEMM algorithm - MeshGEMM
- New GEMV algorithm - MeshGEMV

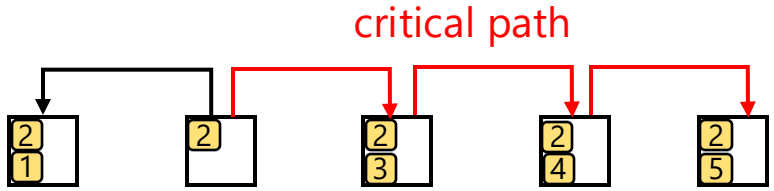
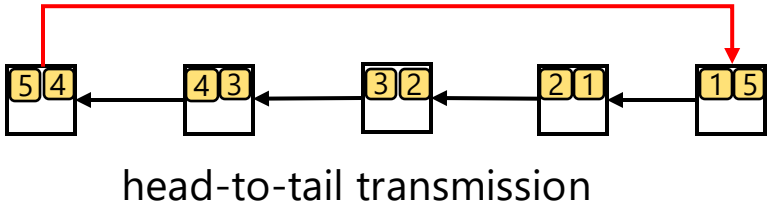
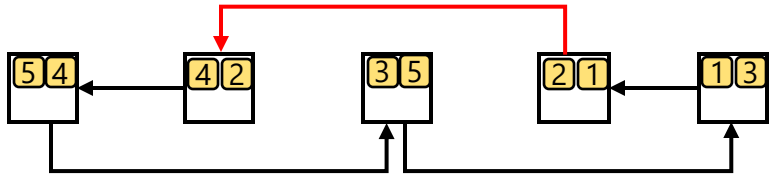
First LLM inference system to reach 2700 token/s per request

Current WaferLLM Architecture



PLMR Compliant MeshGEMM

Prefill is bottlenecked by GEMM, which requires each submatrix to traverse all row (or column) cores, constrained by properties L and R.

		Latency (L)	Routing (R)
SUMMA (SOTA in All-to-All / Clos)		$O(N)$	$O(N)$
Cannon (SOTA in 2D Torus)		$O(N)$	$O(1)$
MeshGEMM (Ours)		$O(1)$	$O(1)$

Evaluation results:

1. **1.3–2×** faster GEMM with matrix sizes from 2K to 8K
2. Reduces communication overhead by **2–5×**.

Key techniques: **Cyclic Shifting** and **Interleave**

Comparison with SOTA LLM Inference Systems

We compare **WaferLLM on real Cerebras WSE-2 chip (TSMC 7nm)** with **SGLang/vLLM on NVLink/IB-connected A100 GPU (TSMC 7nm)** in **performance** and **energy efficiency**

Decode (4K in, 4K out, BSZ=1)	LlaMA3-8B		
SGLang (A100) Token/s per request	1 GPU	8 GPUs	2x8 GPUs
	78	260	164
WaferLLM (WSE-2) Tokens/s per request		**2700**	
A100/WSE-2 Energy Ratio	0.92	2.22	7.02

- **6-20x** faster than SoTA off-chip solutions on LLM model size range from 8B to 70B
- **2-2.5x** more energy efficiency than GPU interconnect – currently the only one on the market beyond NVLink
- Outperforms best-case off-chip scaling in both speed and efficiency

Summary

- To further scale compute for AI
  - Going after simplicity that enables true scalability
  - A new AI stack following the PLMR model to address the system challenges