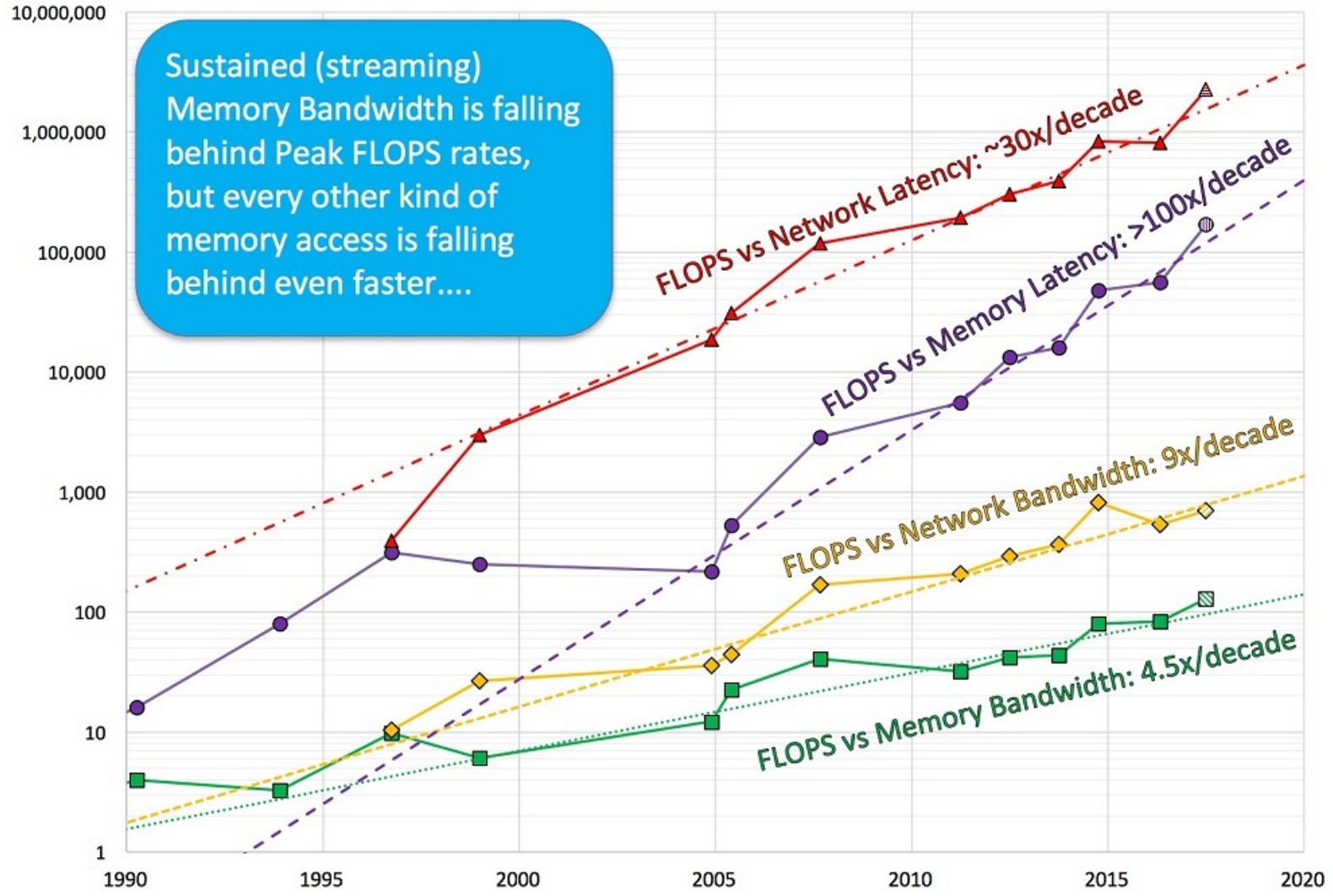
Wafer-Scale Processors for HPC

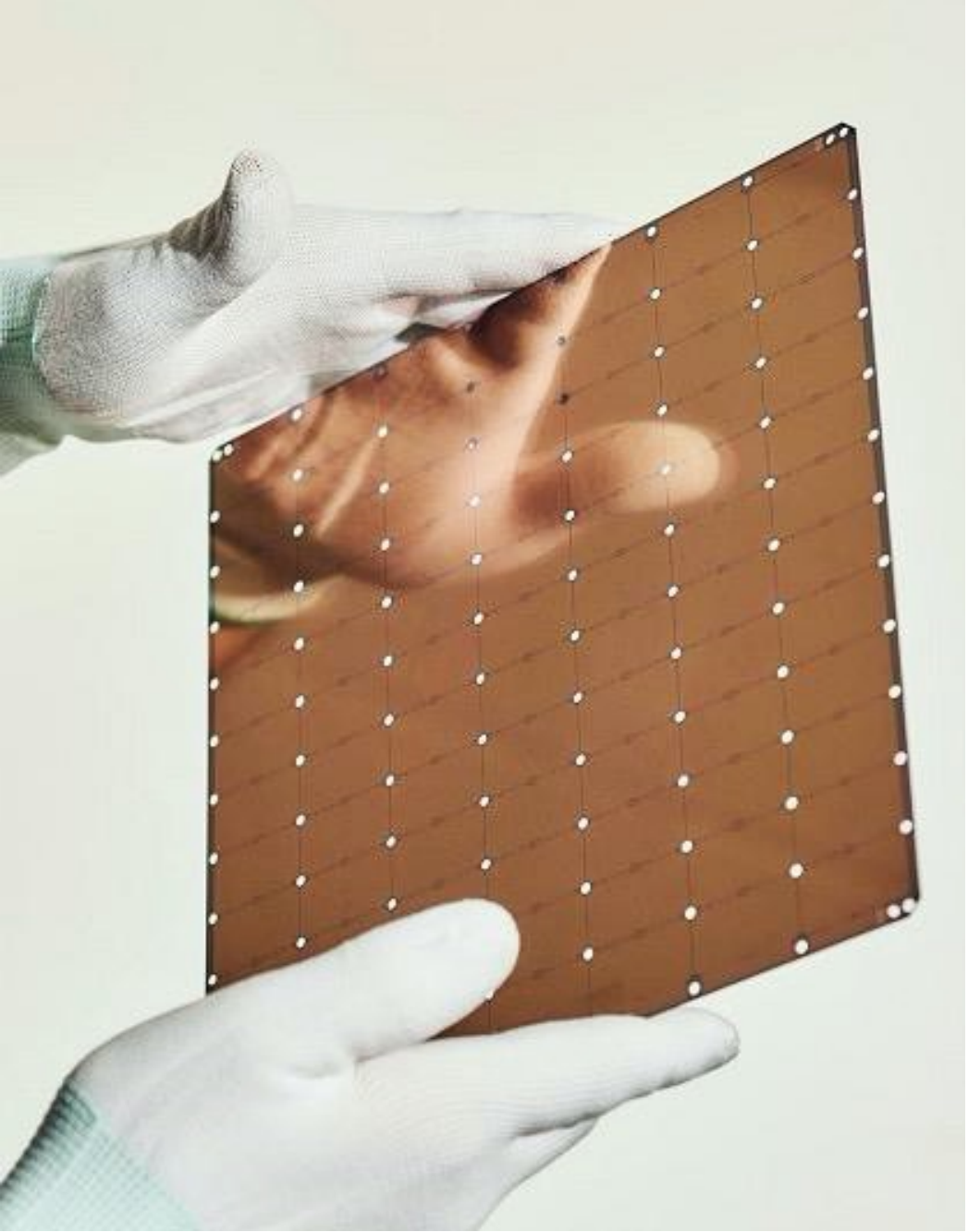
Rob Schreiber
ModSim, October 2021

The Memory & Interconnect “Walls”



A plausible solution:

The HPC-optimized Wafer-Scale Processor



Cerebras Wafer Scale Engine (WSE-2)

The Most Powerful Processor for AI

850,000 AI-optimized cores

46,225 mm² silicon

2.6 trillion transistors

40 Gigabytes of On-chip Memory

20 PByte/s memory bandwidth

220 Pbit/s fabric bandwidth

7nm process technology

Cluster-scale acceleration on a single chip

cerebras

	Cerebras WSE-2	A100	Cerebras Advantage
Chip size	46,225 mm^2	826 mm^2	56 X
Cores	850,000	6912 + 432	123X
On-chip memory	40 Gigabytes	40 Megabytes	1,000 X
Memory bandwidth	20 Petabytes/sec	1555 Gigabytes/sec	12,733 X
Fabric bandwidth	220 Petabits/sec	600 Gigabytes/sec	45,833 X

Argonne National Lab is speeding up cancer research.

GlaxoSmithKline is exploring more ideas in less time.

"The Cerebras system will be critical in the development of next generation ML to uncover next set of more viable drug targets. The incredible power of the Cerebras architecture allows us to explore these new frontiers and decode the language of the cell."

– Kim Branson, Head of AI R&D, GlaxoSmithKline

AstraZeneca is training AI in 2 days instead of weeks.

What (else) can it do?

Fast Stencil-Code Computation on a Wafer-Scale Processor

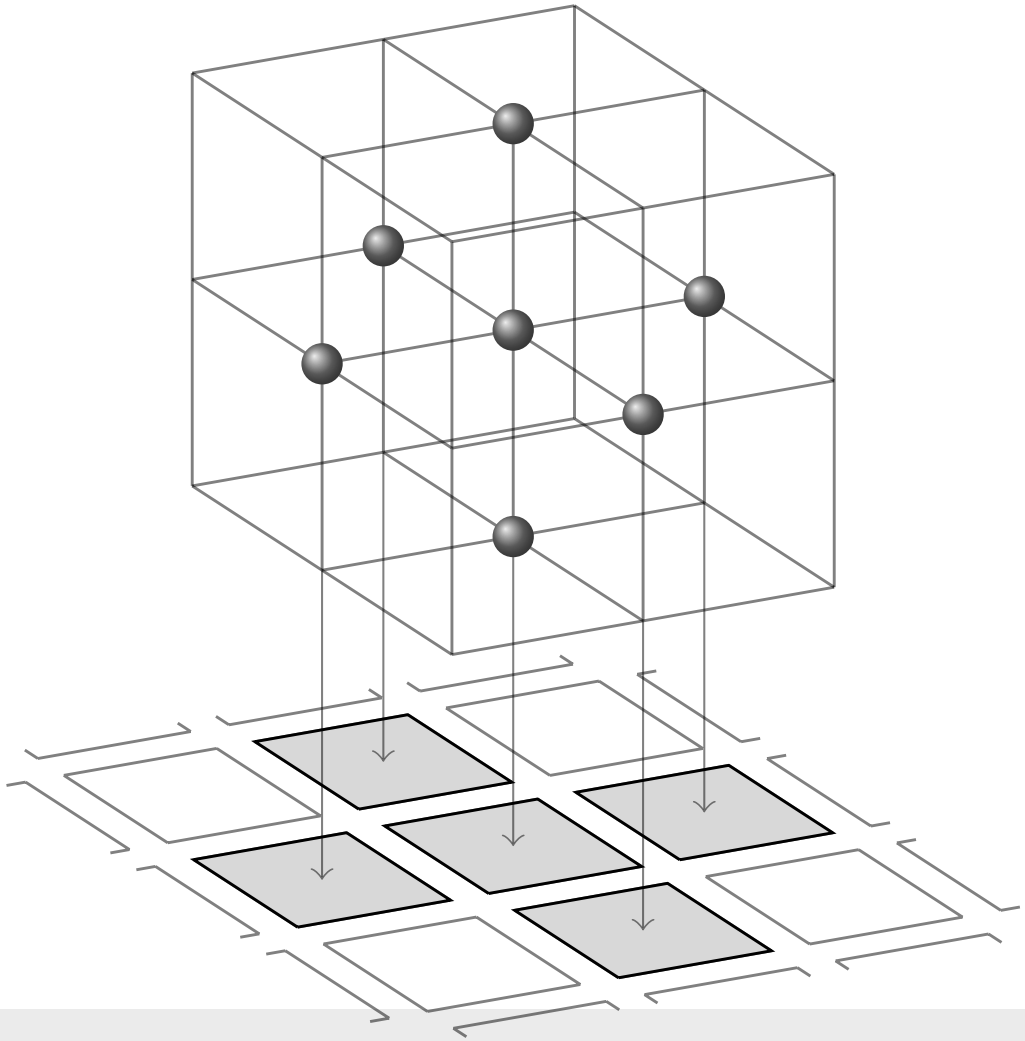
Kamil Rocki*, Dirk Van Essendelft†, Ilya Sharapov*, Robert Schreiber*, Michael Morrison*,
Vladimir Kibardin*, Andrey Portnoy*, Jean Francois Dietiker†‡, Madhava Syamlal† and Michael James*

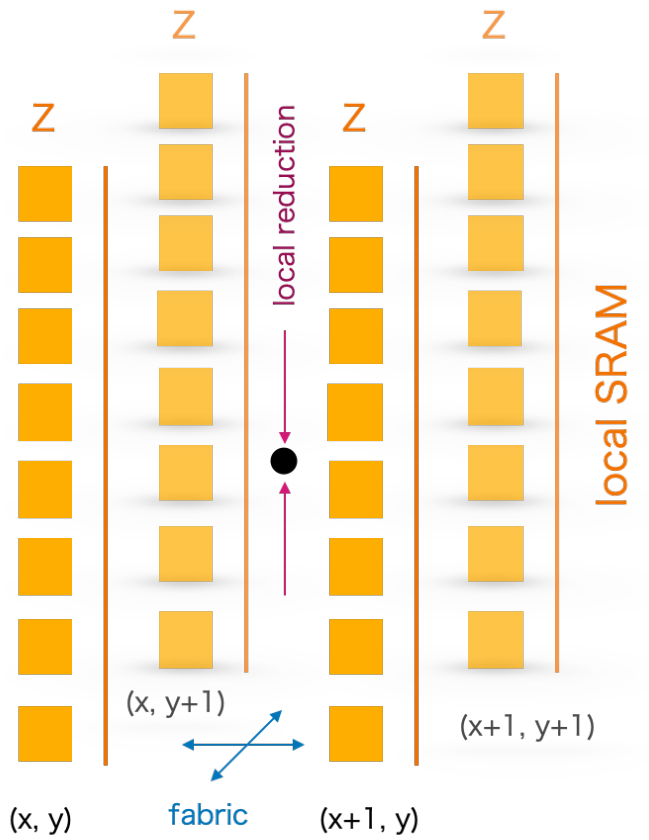
* Cerebras Systems Inc., Los Altos, California, USA
Email: {kamil,michael}@cerebras.net

† National Energy Technology Laboratory, Morgantown, West Virginia, USA
Email: dirk.vanessendelft@netl.doe.gov

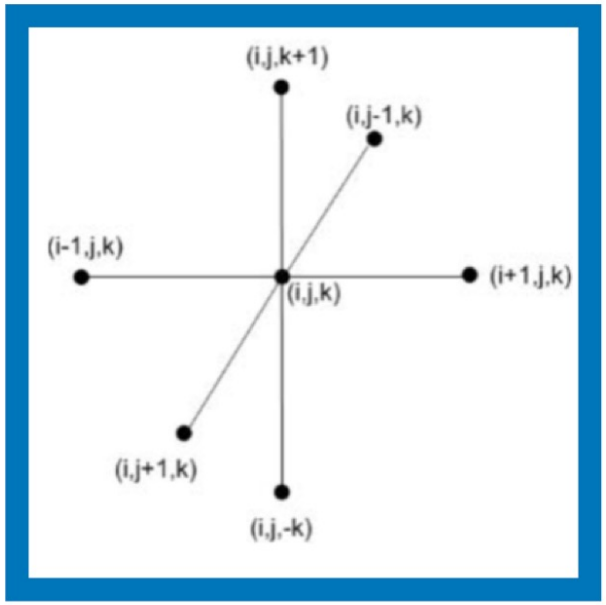
‡ Leidos Research Support Team, Pittsburgh, Pennsylvania, USA
Email: jean.dietiker@netl.doe.gov

On the CS-1: 3D mesh > 2D machine

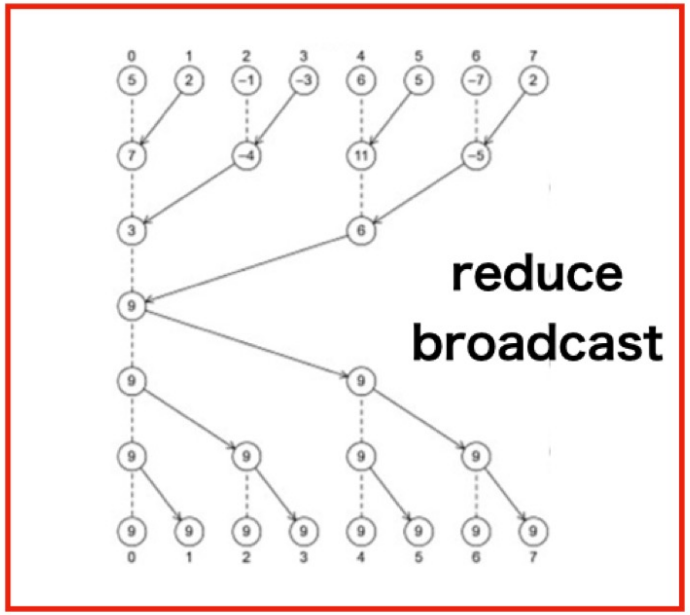




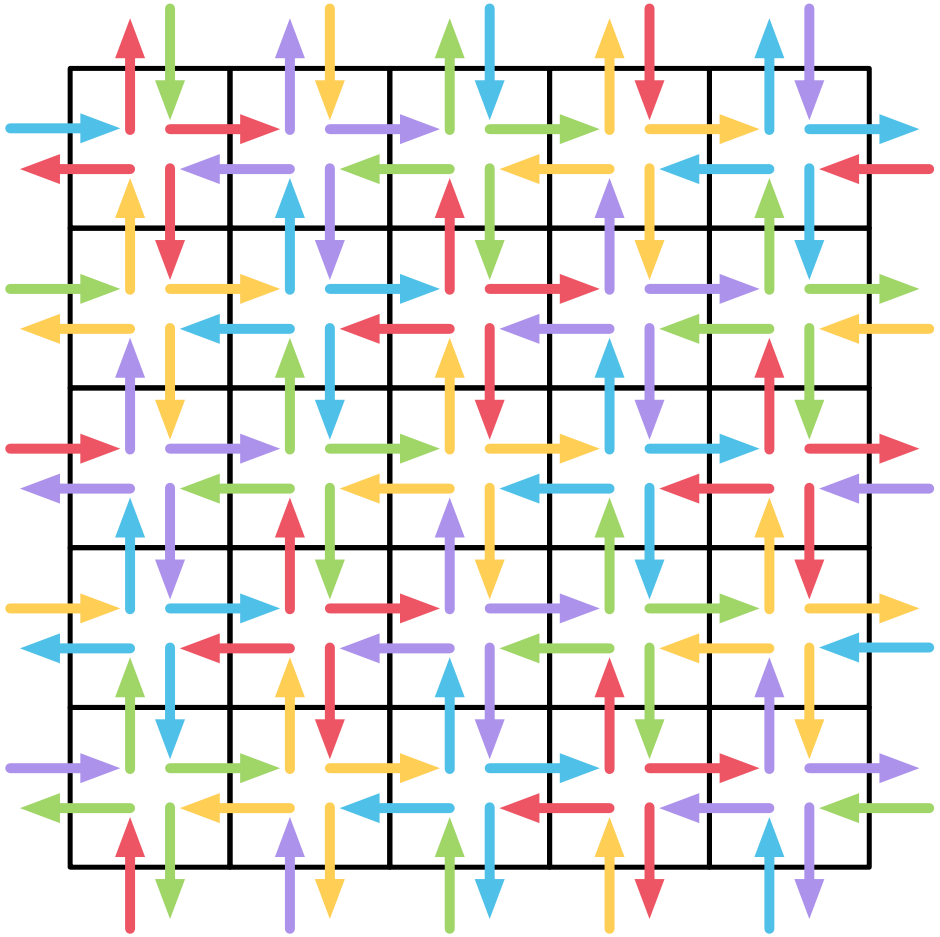
Sparse Linear Solver Building Blocks



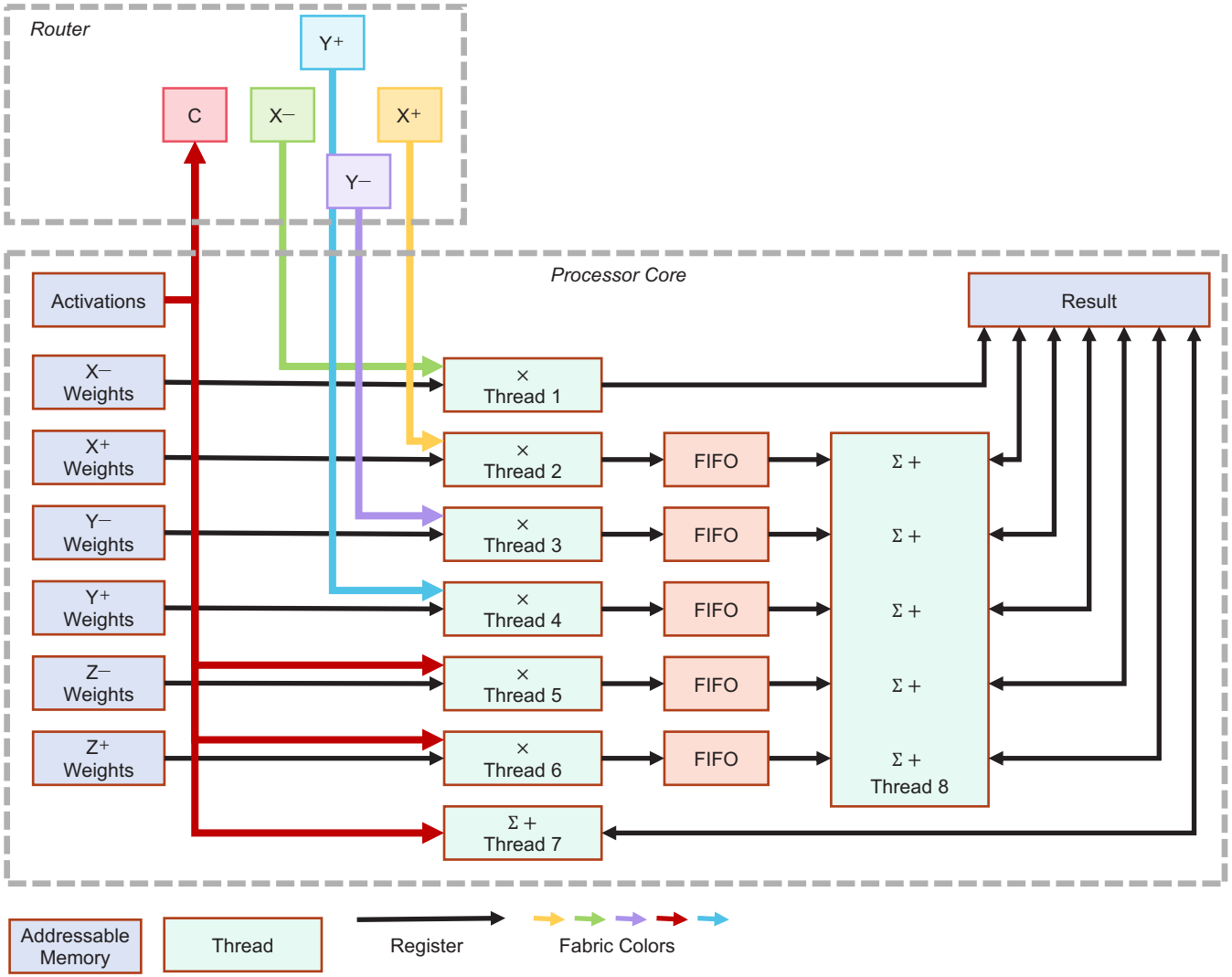
$$\begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \leftarrow \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} + \alpha \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$



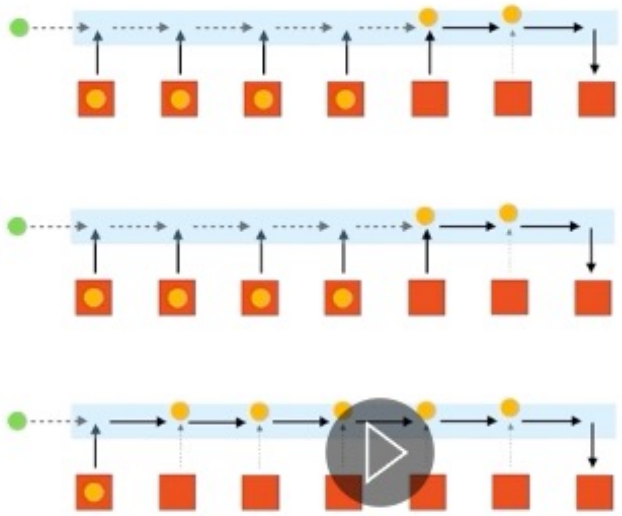
Data communication, sparse Ax

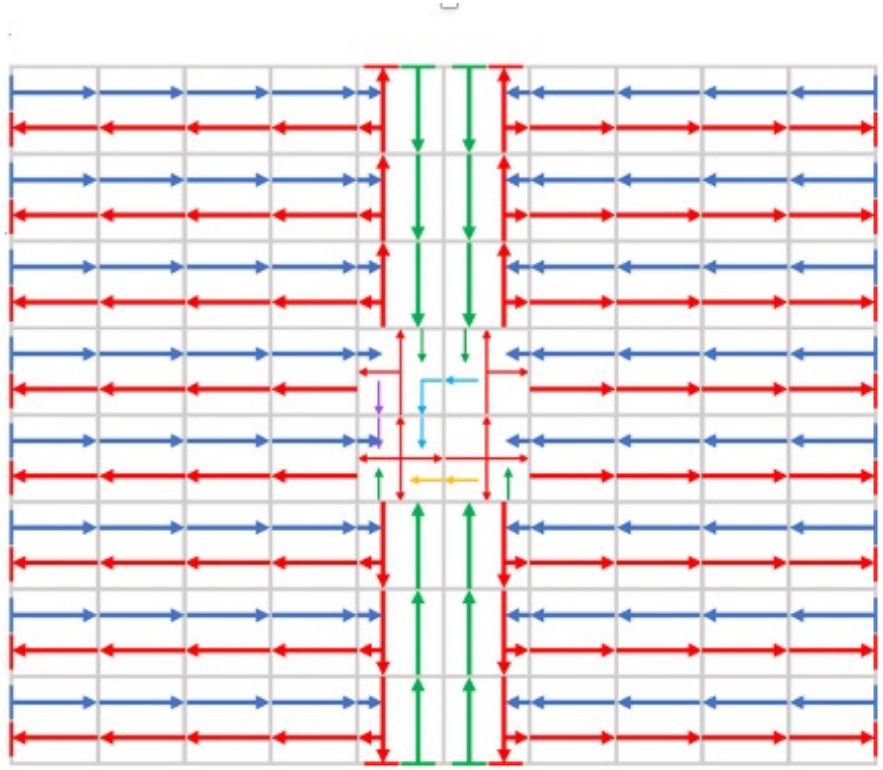


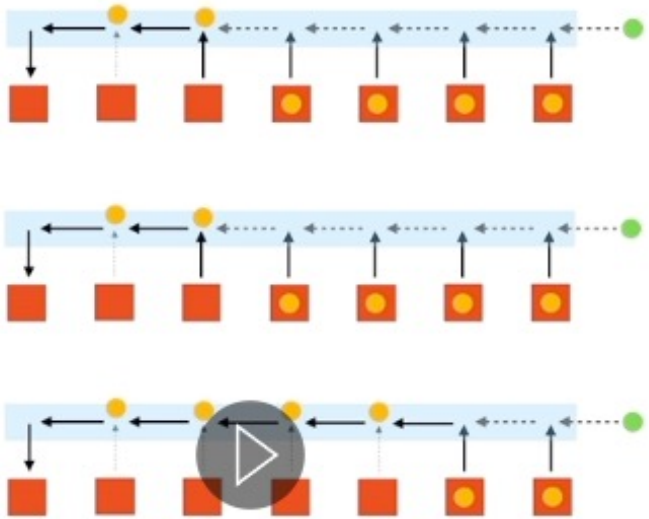
Code, sparse Ax



Allreduce for 360,000 processor dot product

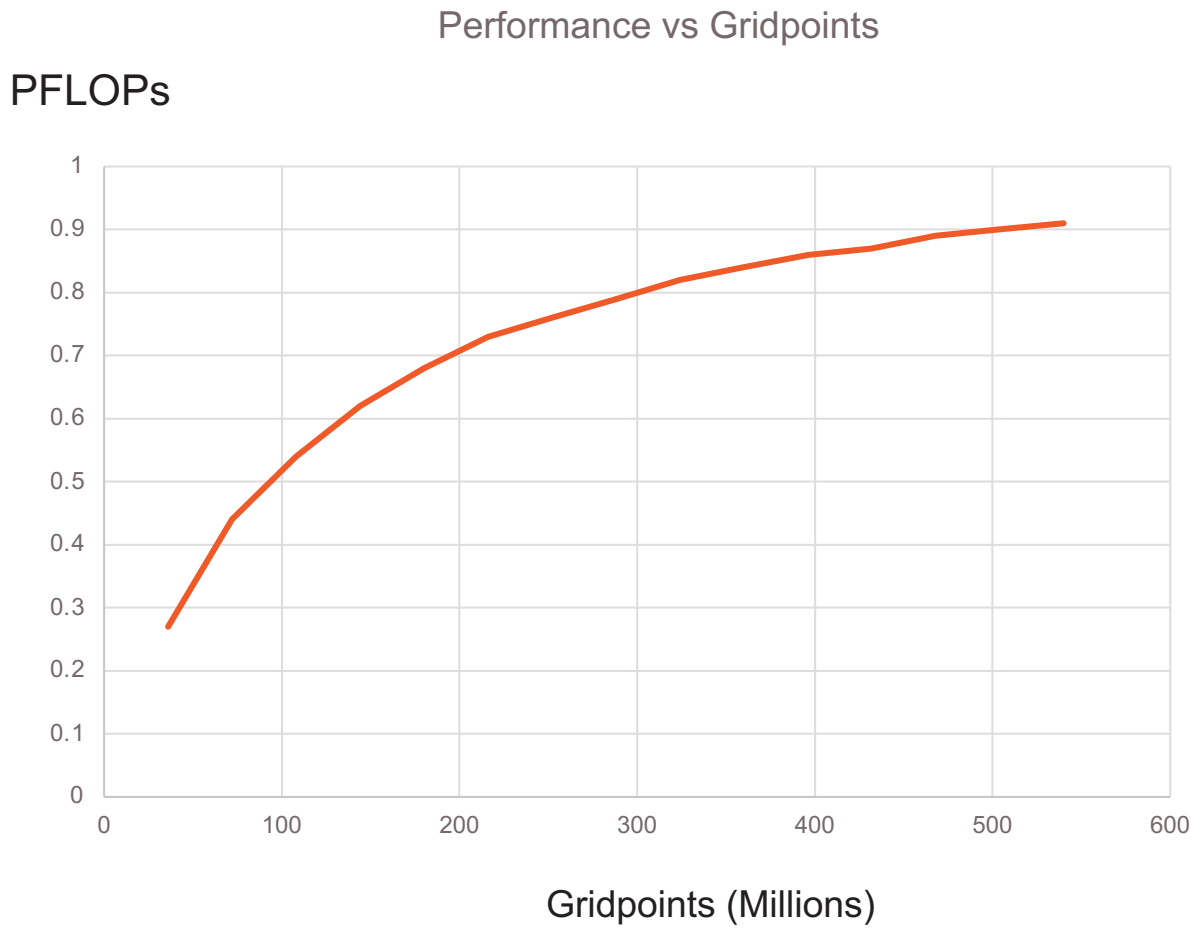






BiCGstab Performance

- Stream data to neighboring PEs
- No memory bandwidth limits
- 1.3 microsecond allreduce on 350,000 PEs
- 16-bit (32-bit add in dot products)
- Measured 0.86 PFLOP/s in lab on CS-1
  - 28 microseconds per iteration for a 600 × 595 × 1536 mesh on 602 × 595 compute fabric
  - Compared to 6 milliseconds on NETL's Joule supercomputer.
- 0.86 Pflops measured out of 2.43 = 35% of peak.



November 2020 HPCG Results

New HPCG results announced at SC20

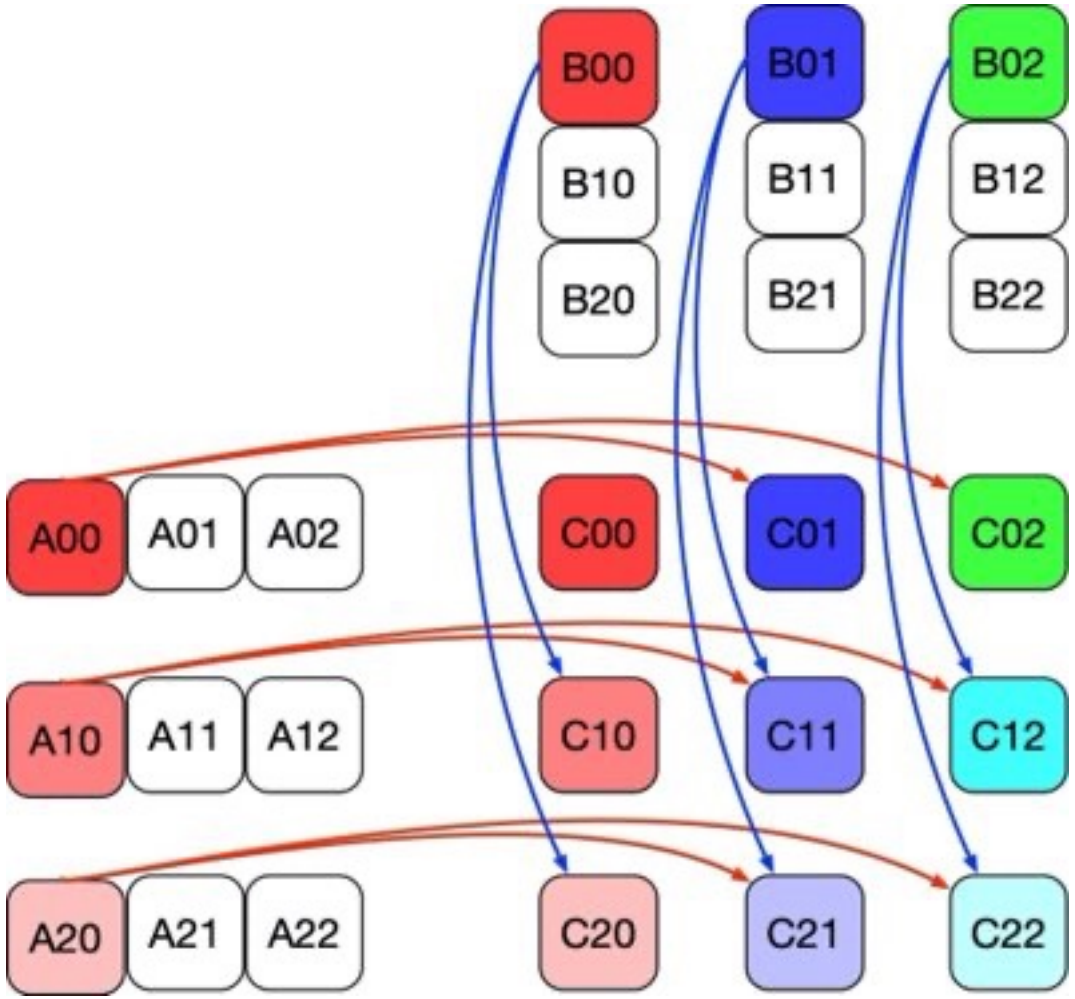
Rank	Site	Computer	Cores	HPL Rmax (Pflop/s)	TOP500 Rank	HPCG (Pflop/s)	Fraction of Peak
1	RIKEN Center for Computational Science **Japan**	**Supercomputer Fugaku** — A64FX 48C 2.2GHz, Tofu Interconnect D	7,630,848	442.01	1	16.00	3.0%
2	DOE/SC/ORNL **USA**	**Summit** — IBM POWER9 22C 3.07GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta V100	2,414,592	148.60	2	2.926	1.5%
3	DOE/NNSA/LLNL **USA**	**Sierra** — IBM POWER9 22C 3.1GHz, Dual-rail Mellanox EDR Infiniband, NVIDIA Volta V100	1,572,480	94.64	3	1.796	1.4%
4	NVIDIA Corporation **USA**	**Selene** — AMD EPYC 7742 64C 2.25GHz, Mellanox HDR Infiniband, NVIDIA Tesla A100 40GB	555,520	63.46	5	1.623	2.0%
5	Forschungszentrum Juelich (FZJ) **Germany**	**JUWELS Booster Module** — AMD EPYC 7402 24C 2.8GHz, Mellanox HDR InfiniBand/ParTec ParaStation ClusterSuite, NVIDIA Ampere A100	449,280	44.12	7	1.275	1.8%

Random small messages

- Small (single word) messages are the limiter for CPU/GPU
- Bisection bandwidth on the wafer
- ***ISA-based communication*** is key
- Over 200B cross-machine single word communications per second on 256K processors (512 X 512 fabric)
- Over 300 B on CS-2, larger fabric
- Further results coming

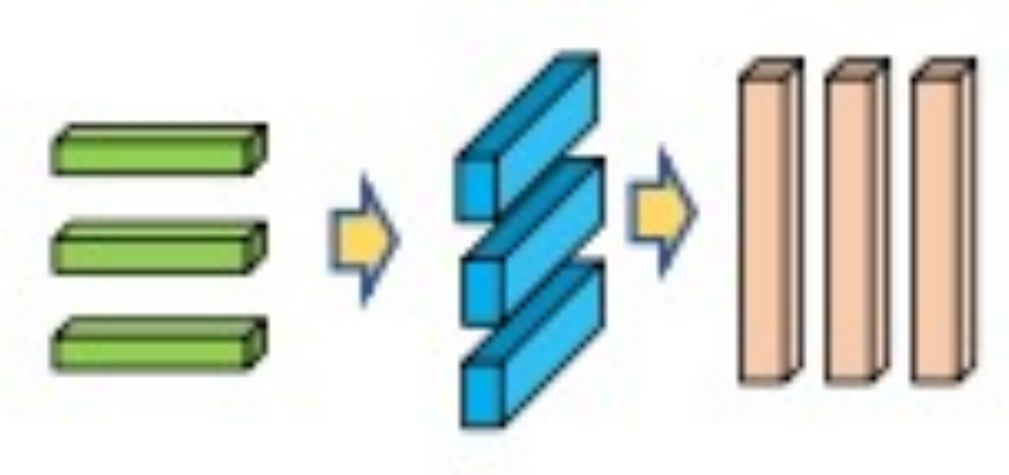
GEMM

- Single GEMM, matrices on wafer.
  N = 6,000 -- 30,000
- Use outer product formulation
- Fast broadcasts of rows and columns
- > 90 percent of peak for largest matrices.



FFT

- 454 MFLOPS per PE
- 3D --> Transposes
- Bisection bandwidth 1.74 TB/s in each direction
- Lower bound on time from BB: 4.8 msec
- Actual transposes: 8.4 msec
- Comparable to world's fastest results



Life on a wafer

+ Memory bandwidth and latency

+ Communication bandwidth and latency

+ Compute density, efficiency

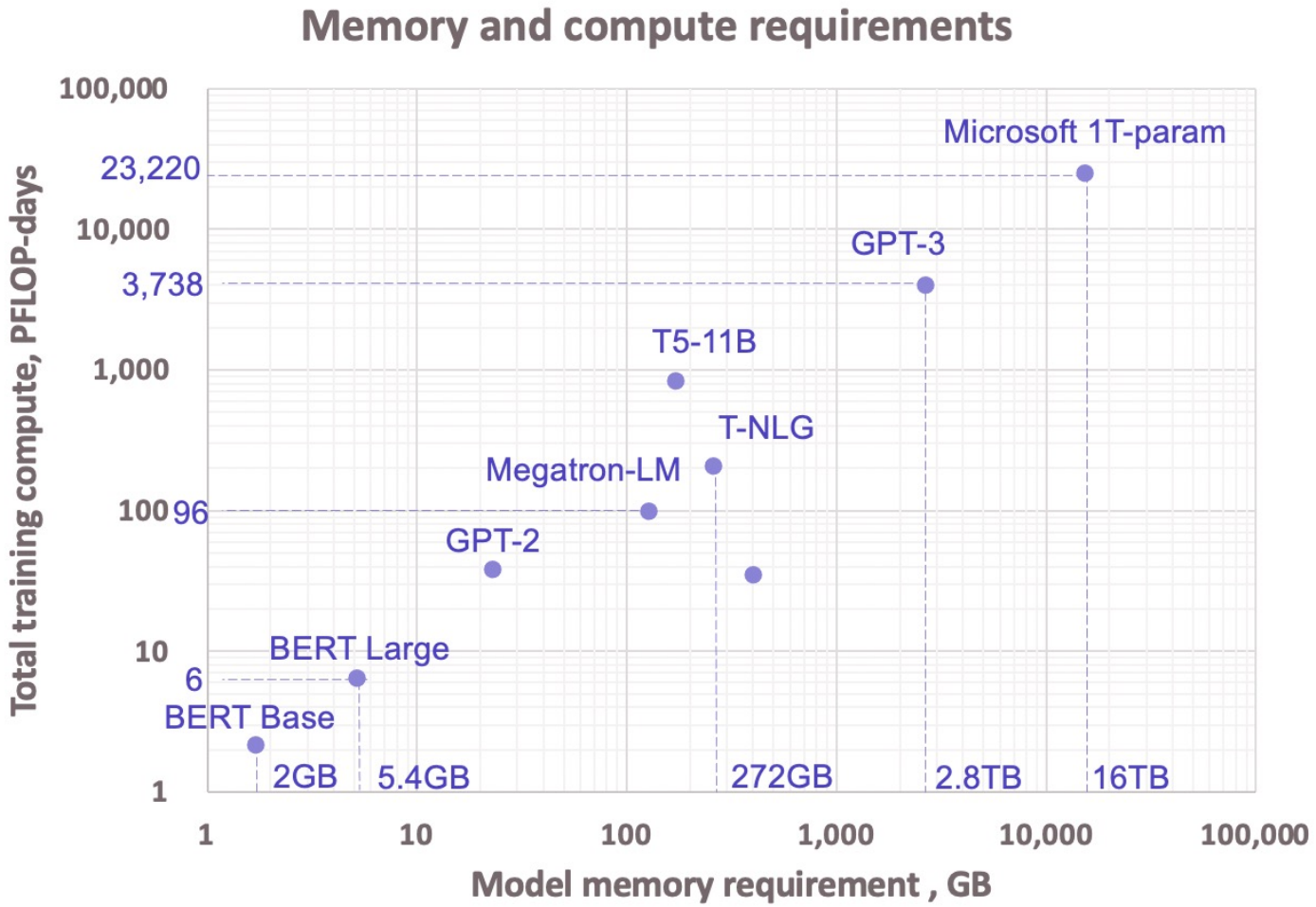
- Small memory

- bisection bandwidth $O(\sqrt{P})$

Wafer-scale: Not Just for AI

Scale out and scale up for DL

Modern models are needing more and more compute



Estimated time-to-train:

- NVIDIA Megatron-LM:
  trained on **512 V100** (32 DGX-2H)
  for **about 10 days**
- OpenAI GPT-3:
  trained on **1024 V100** (64 DGX-2H)
  for **about 116 days**

1 PFLOP-day is about
1 x DGX-2H or 1 x DGX-A100
busy for a day

Training Giant Models

Scale out to enormous scale has limits

- memory capacity per node insufficient for the model
- batch size

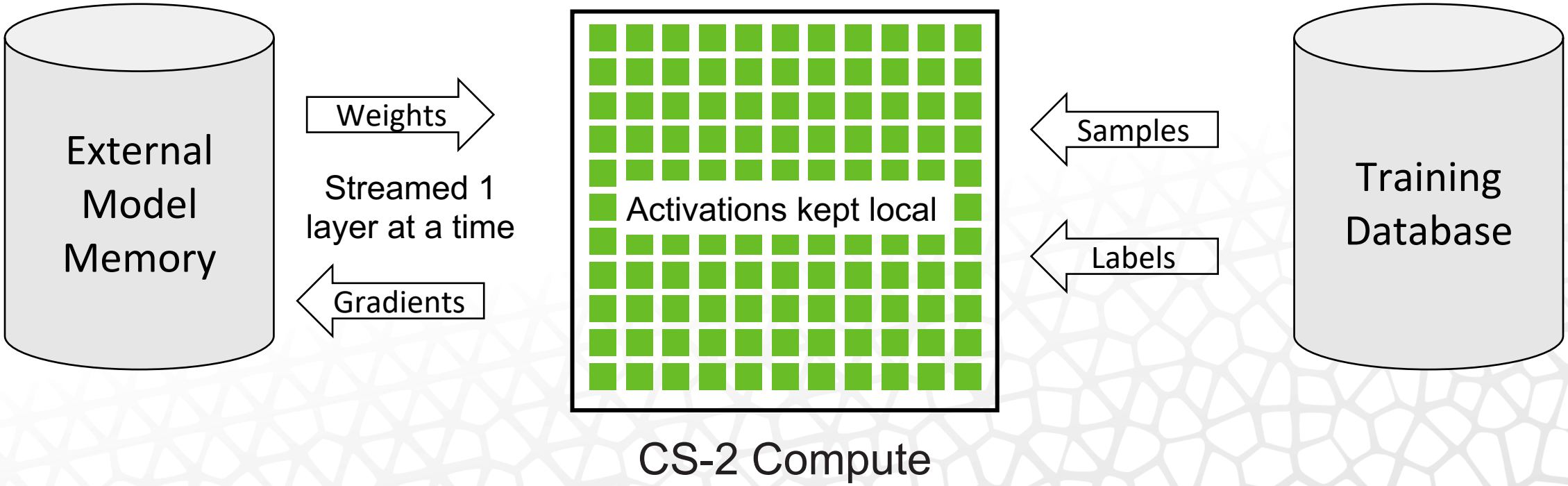
Scale both – a smaller ensemble of wafer-scale (scaled up) compute nodes

- hundreds of nodes > exaflops
- manageable replication, smaller batches, more efficient training
- less pressure on training data server

Weight streaming

- Do not store the whole model on the compute units
- Stream a layer at a time
- Decouple compute speed from memory capacity

Disaggregate Compute and Parameter Storage



Scale model size and training speed independently