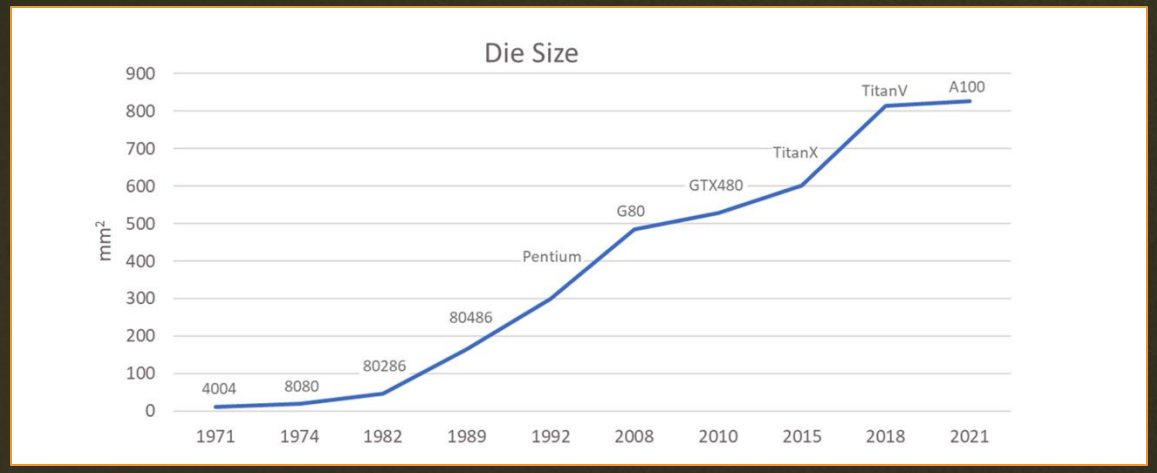
Scientific computing on the largest chip ever built

Cerebras in HPC

Motivation for Cerebras

	Manufacturing Technology node (μm)	Clock frequency	Op/Sec[1]	Die size (mm)^2	Transistor count	Architecture factor	Launch Date
Intel 4004	10	740 kHz	1176	12	2250	7×10^{-7}	1971
Nvidia A100	0.007	1.4 GHz	19.5×10^{15}	826	54 billion	2.6×10^{-4}	2021
Ratio	1429	1892	1.66×10^{13}	68.9	24 million	365	

[1]For comparison, Op is defined as a 32-bit BCD addition for the 4004 and a 32-bit integer add for the A100.



- Over the last 50 years…
- Die size growth approaches an asymptote
- Geometry shrink slows down
- Growth in # of transistors is slowing

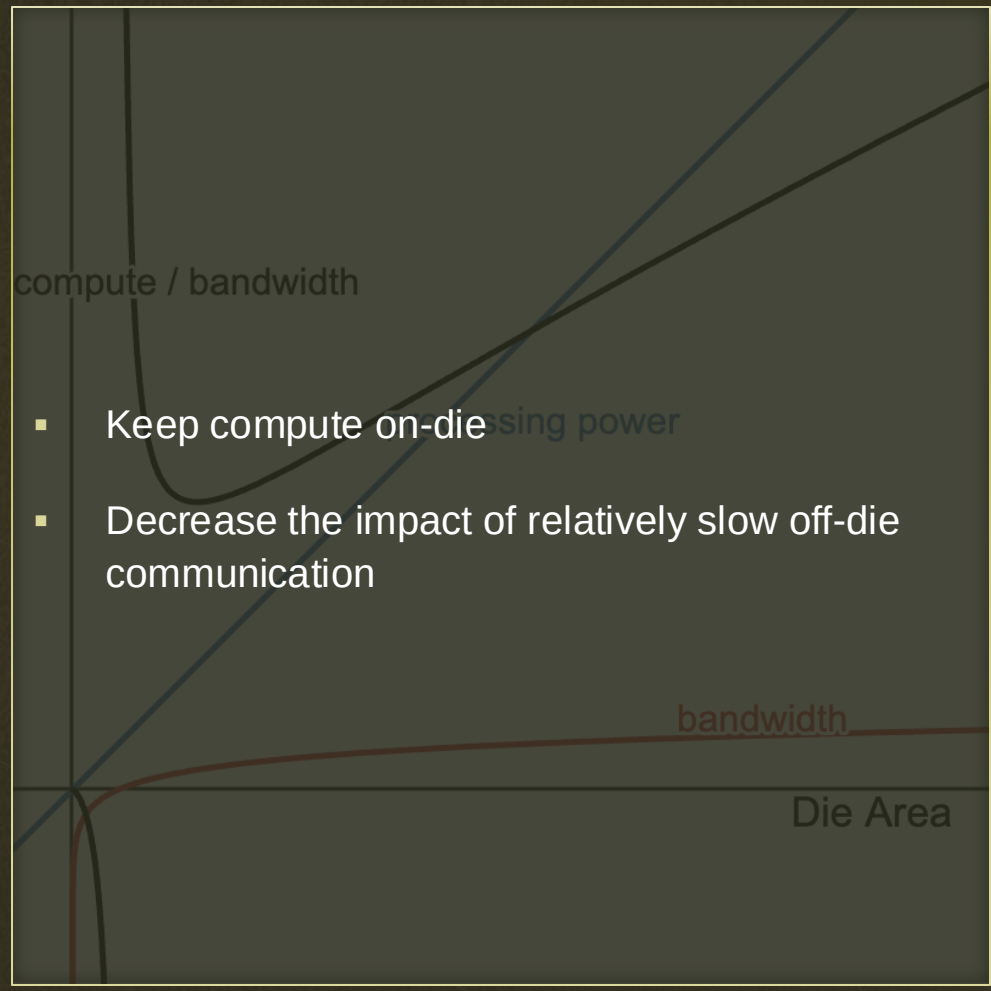
(Lauterbach, 2021)

Emerging Workloads – Neural Networks

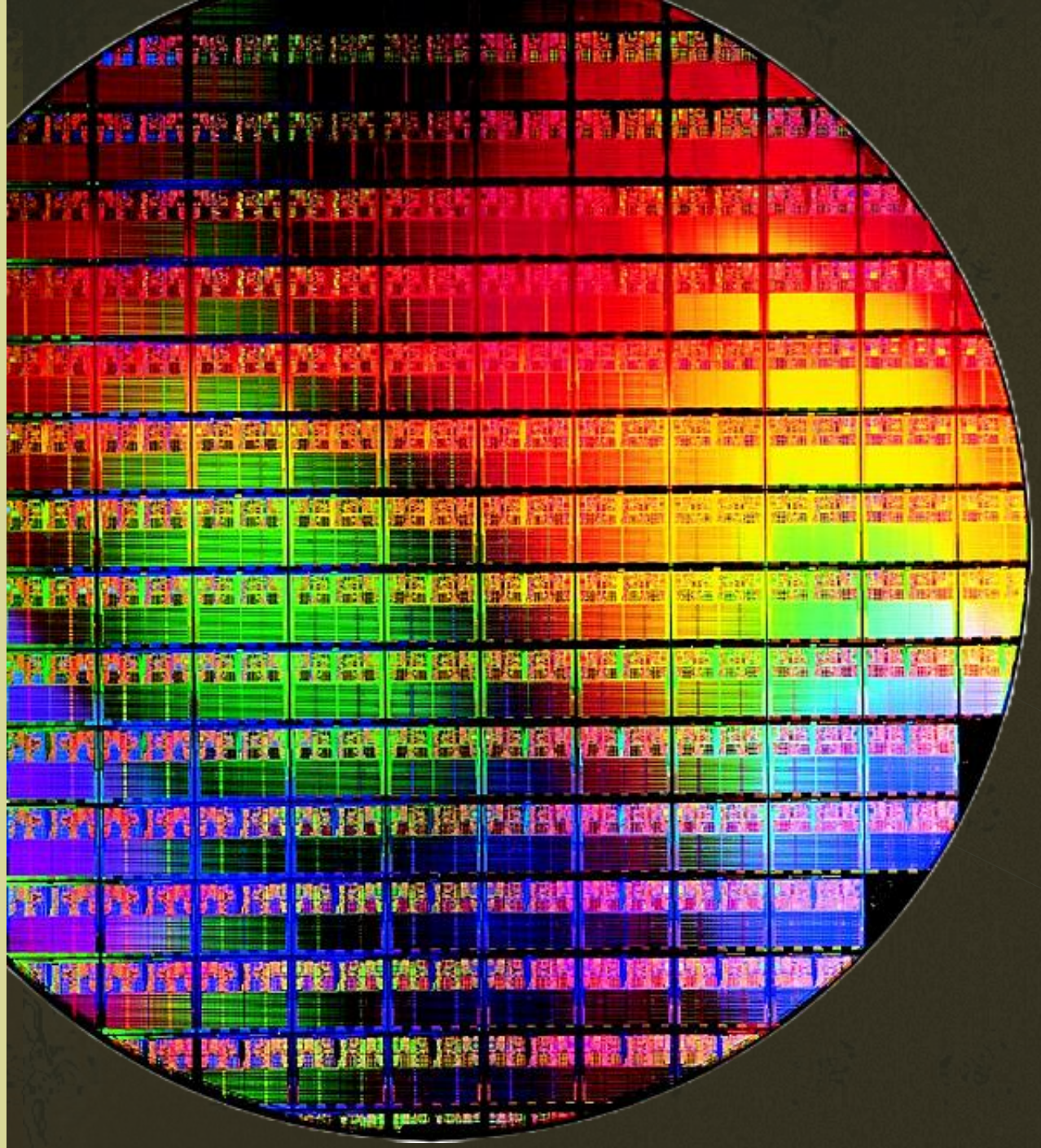
- performance bound
- abundant parallelism
- drive the demand for more transistors

One large die? Many smaller ones?

- Off-die bandwidth proportional to the log of the die area (Rent's rule)
- Processing power grows linearly with die area
- Ratio of compute to off-die bandwidth increases with die area.



- Keep compute on-die
- Decrease the impact of relatively slow off-die communication

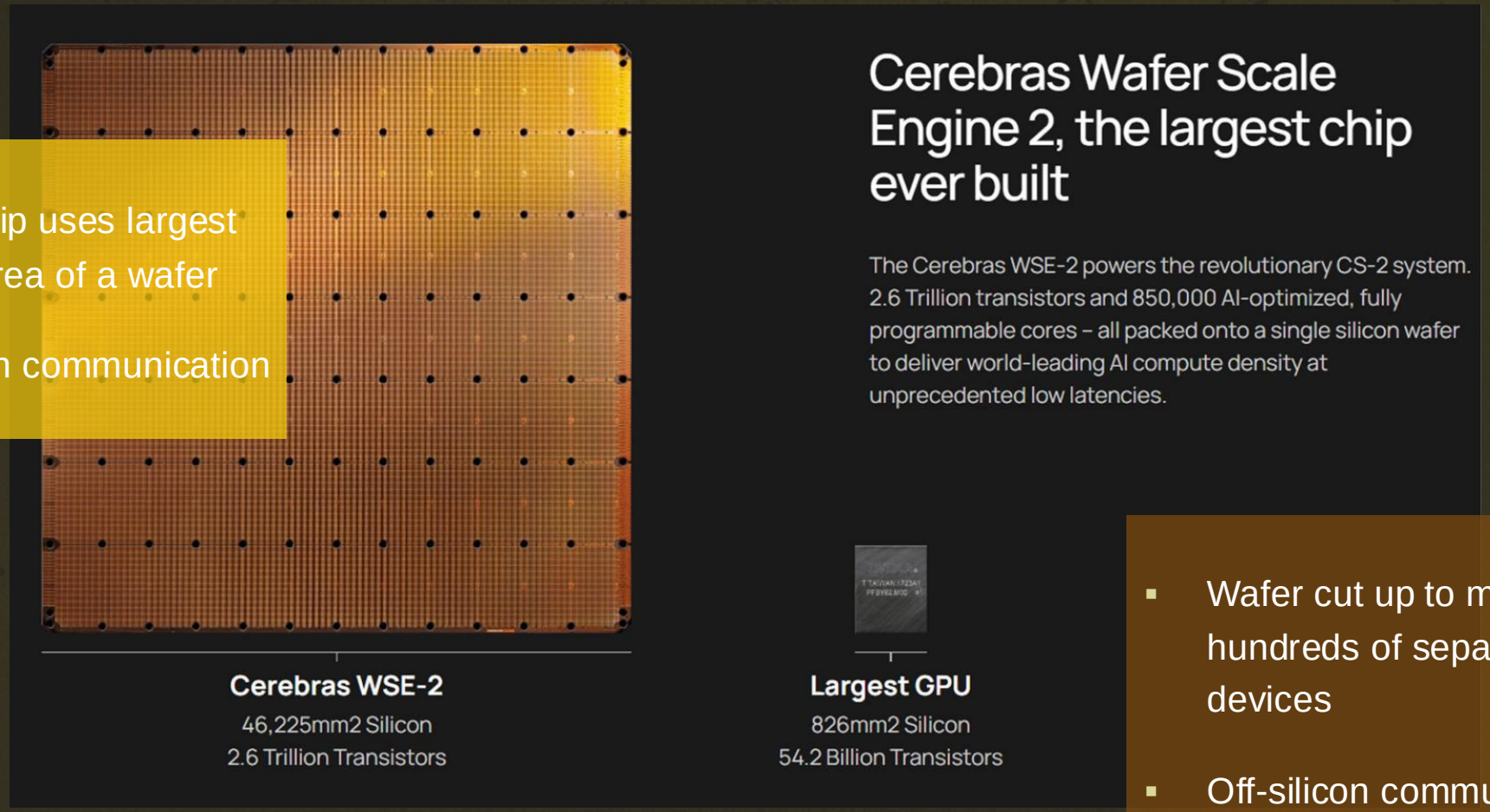


The Beginnings of Wafer Scale

- Use an entire wafer to make a single chip (Wafer Scale Integration)
- Trilogy 1980s attempt
  - Addressing the wafer yield problem
  - triple-modular latency: logic gate and flip-flop were triplicated
  - binary two-out-of-three voting at each triplication

The Wafer Scale Engine (WSE-2)

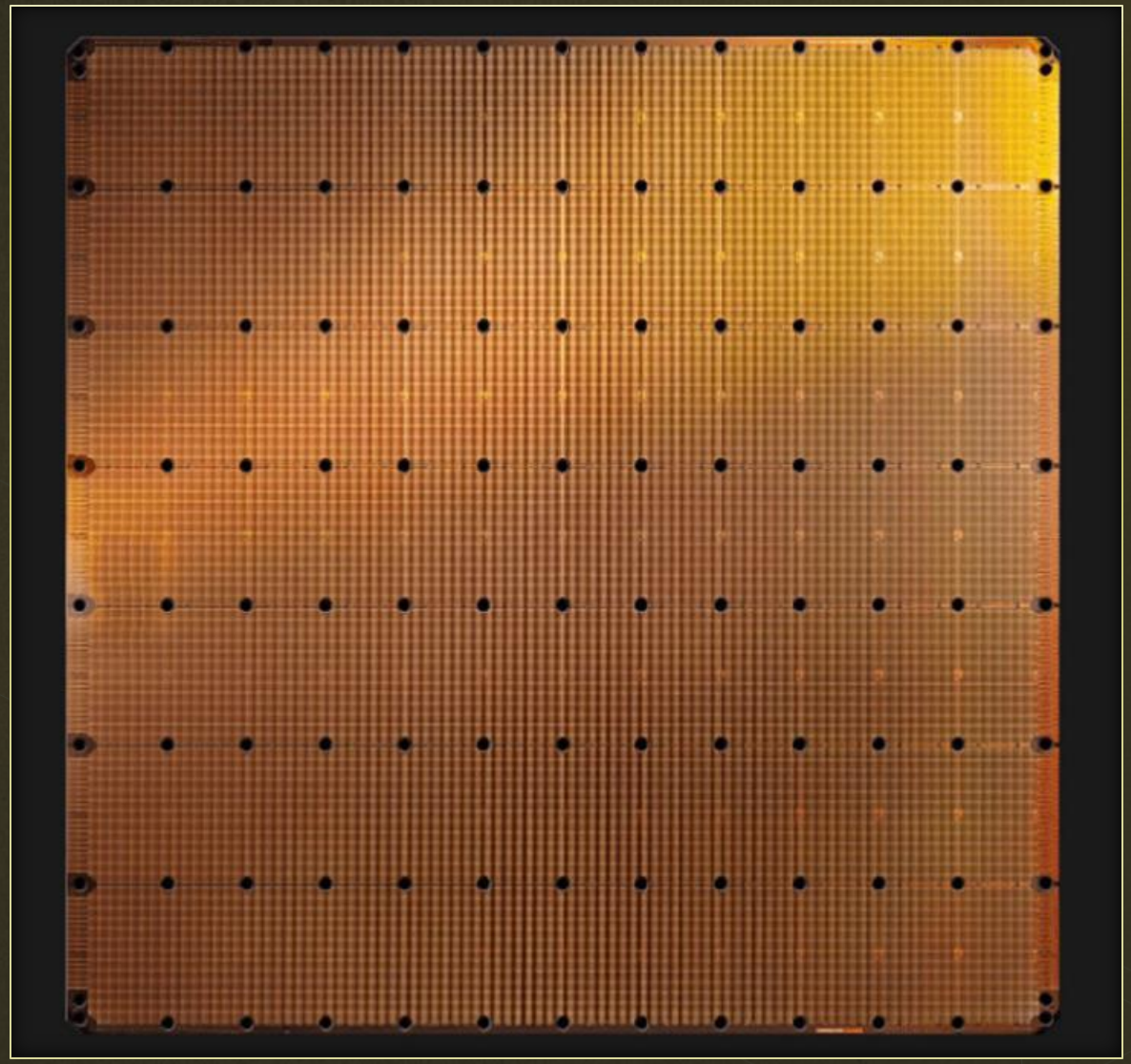
- Single chip uses largest square area of a wafer
- On-silicon communication



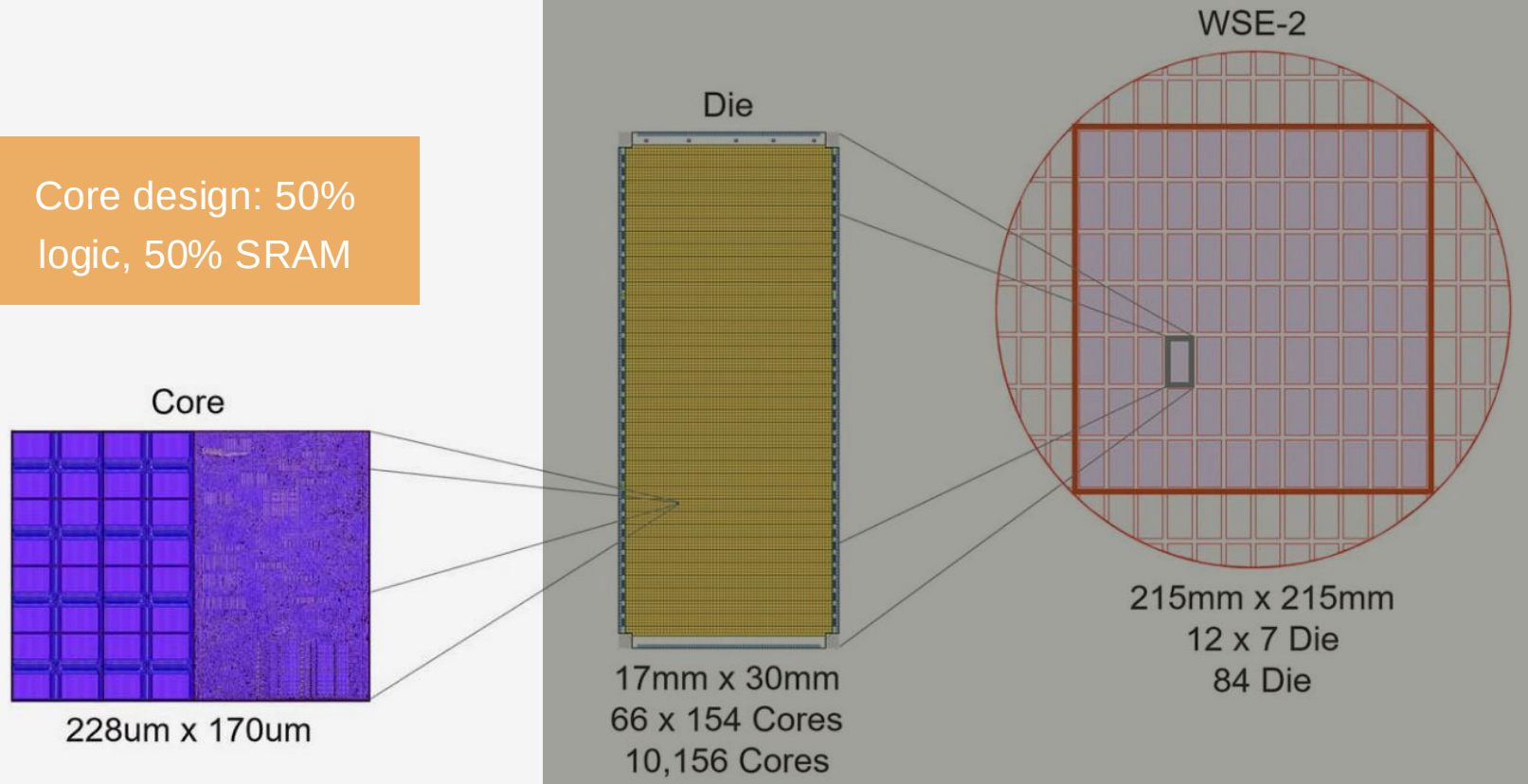
(Lauterbach, 2021)

- Wafer cut up to make hundreds of separate devices
- Off-silicon communication

- grid of processing elements (PEs)
- 850,000 PEs in total



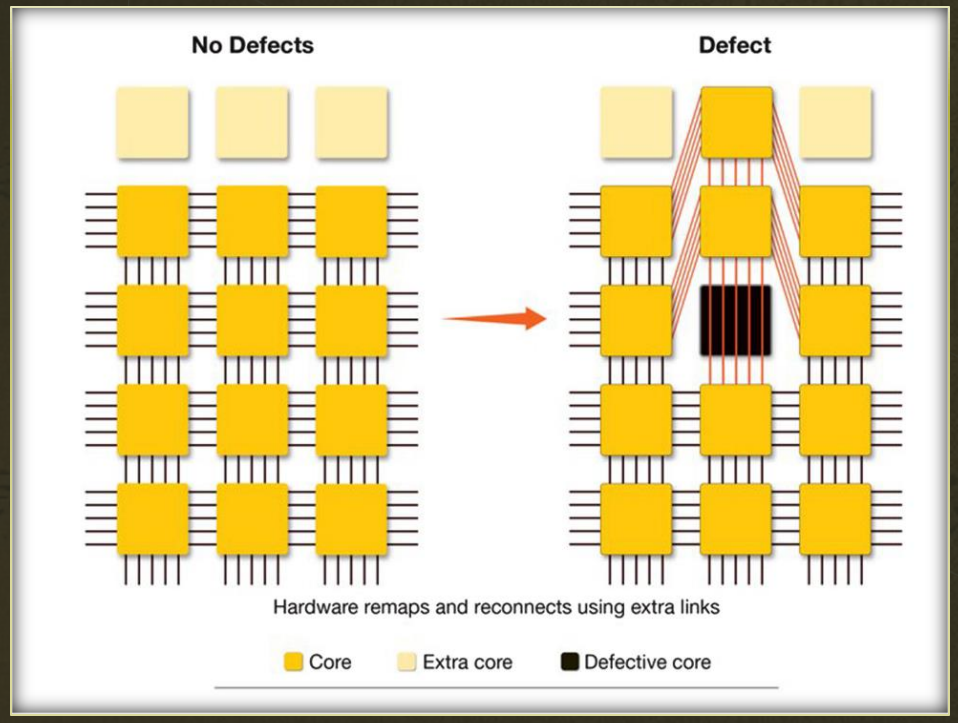
Core design: 50% logic, 50% SRAM



(Lie, 2023)

Architecture Details

The Wafer Yield Problem

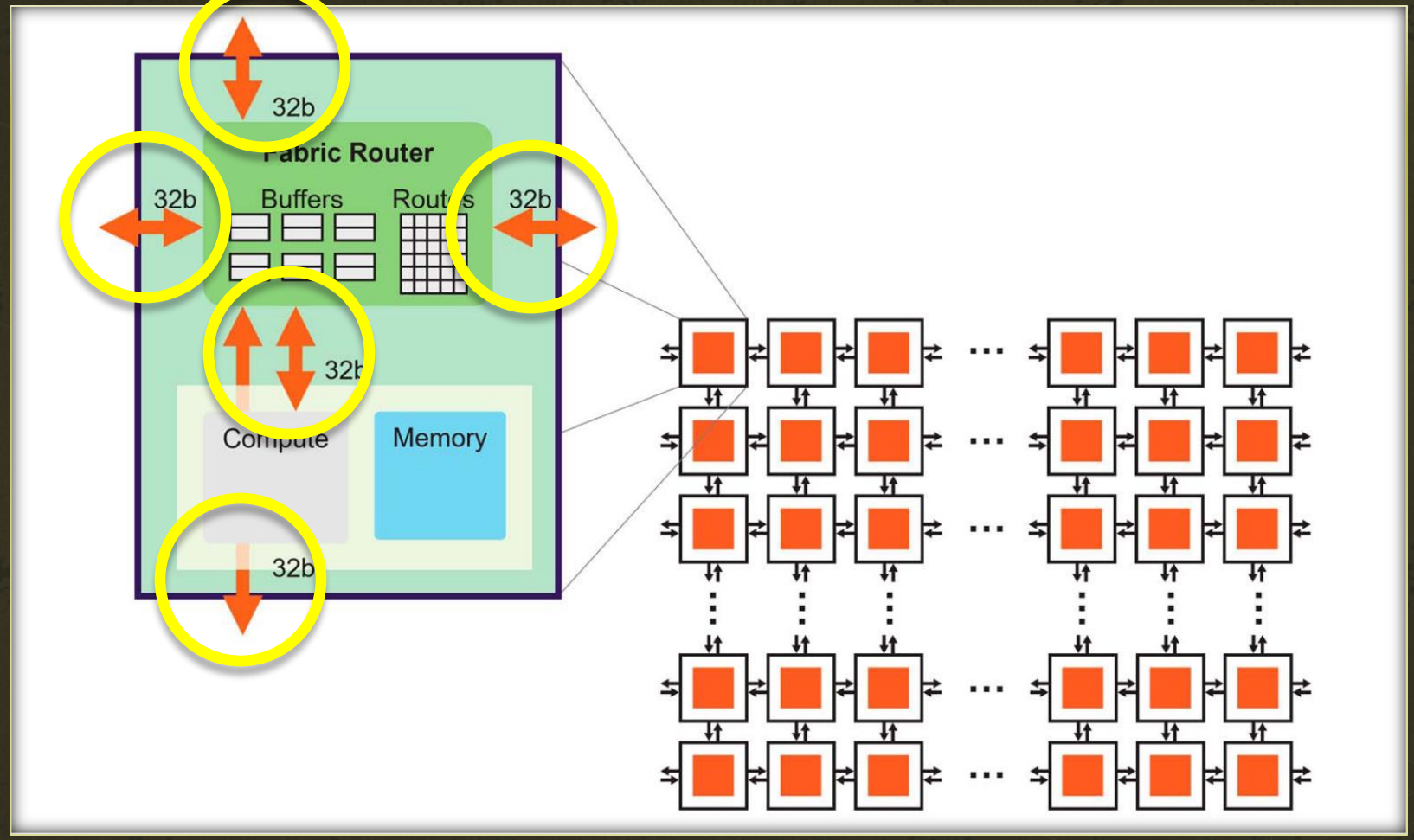


(Lauterbach, 2021)

- Trilogy: triple-modular redundancy
- Cerebras: Homogenous array of processing elements (PEs)
- Approximately 1% held in reserve to “repair” defective PEs

The WSE-2 Core

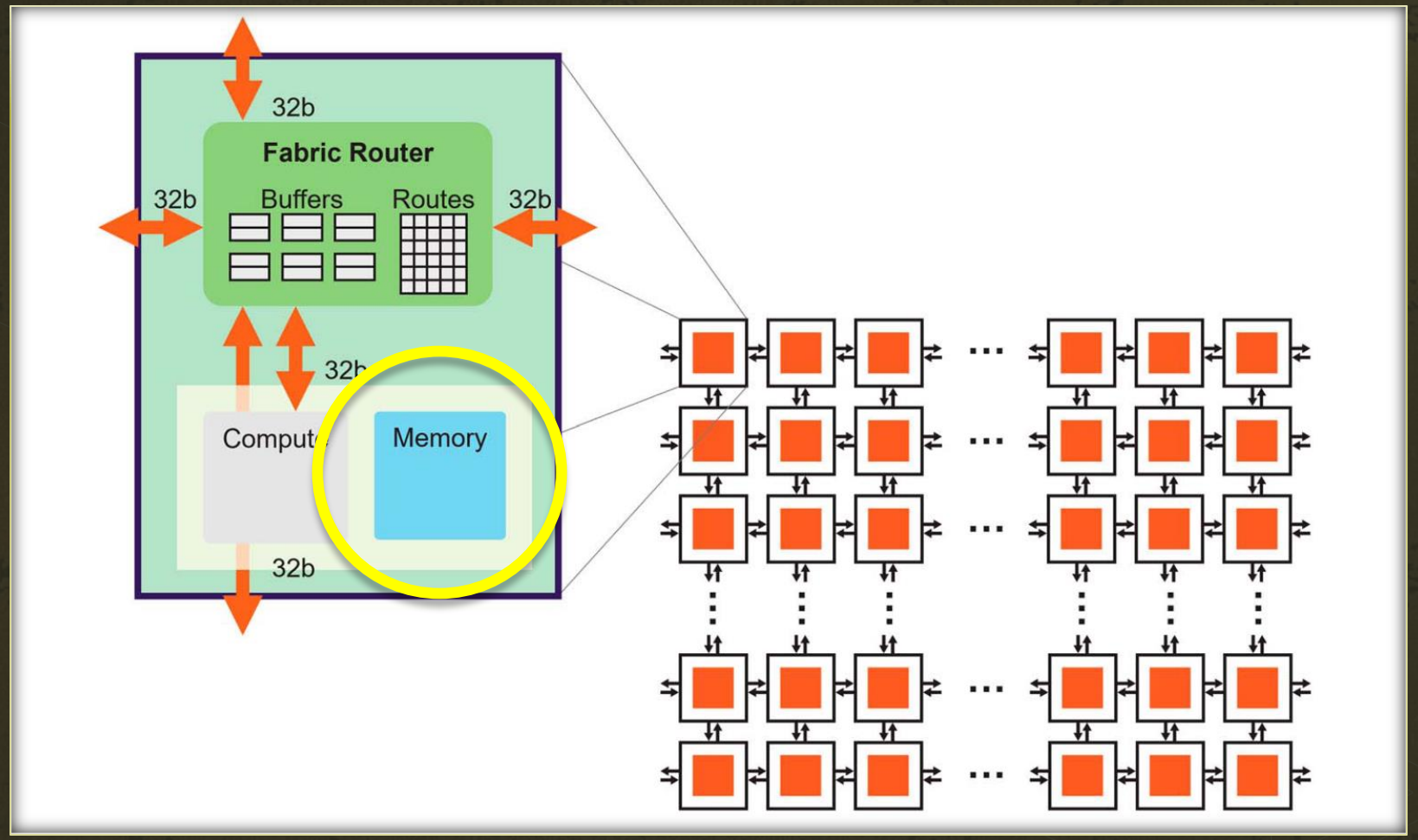
- Bidirectional interfaces
- Data packet: 16 bits data, 16 bits control info
- Fabric extended across die boundaries



(Lie, 2023)

The WSE-2 Core

- 48kB SRAM
- Per cycle: two 64-bit reads and one 64-bit write

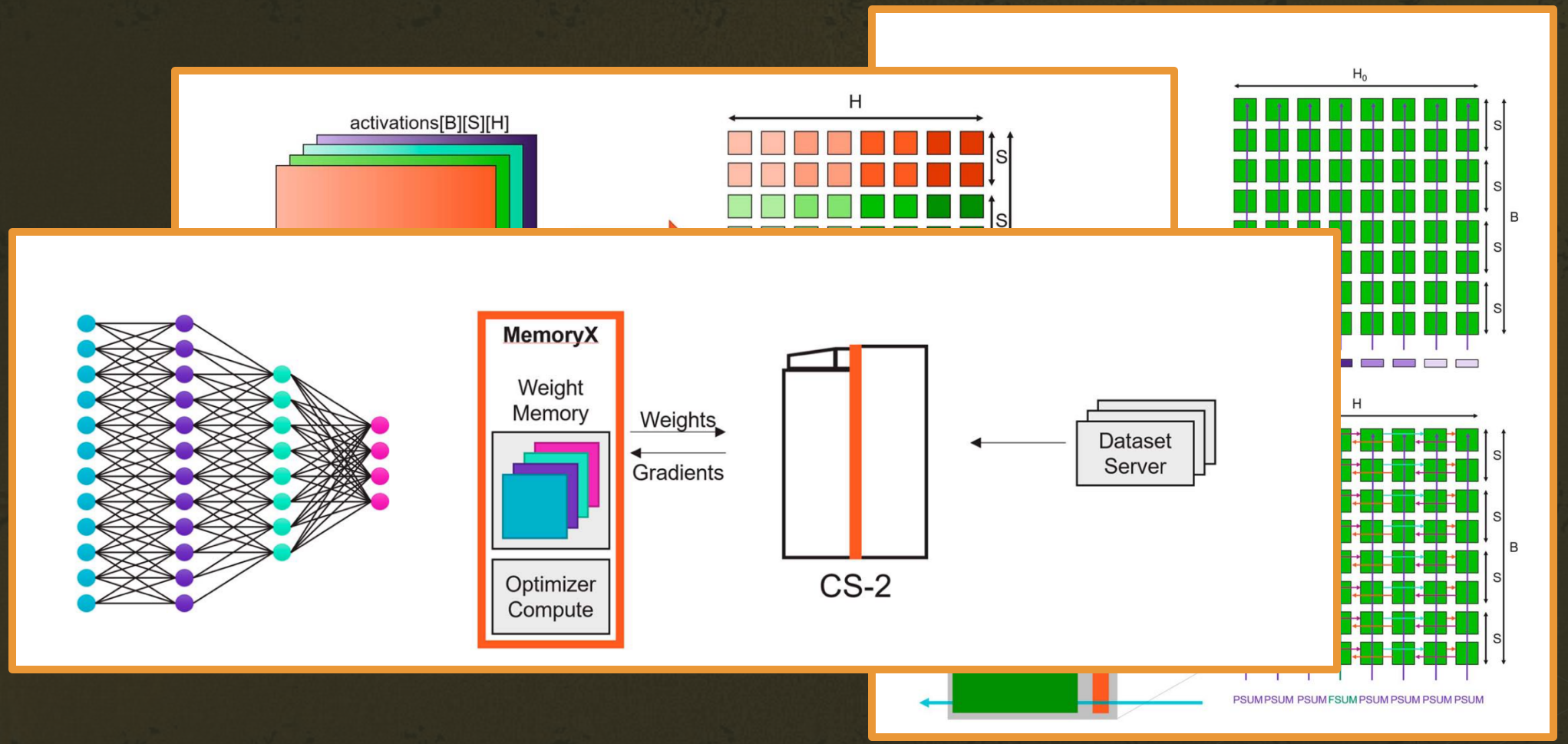


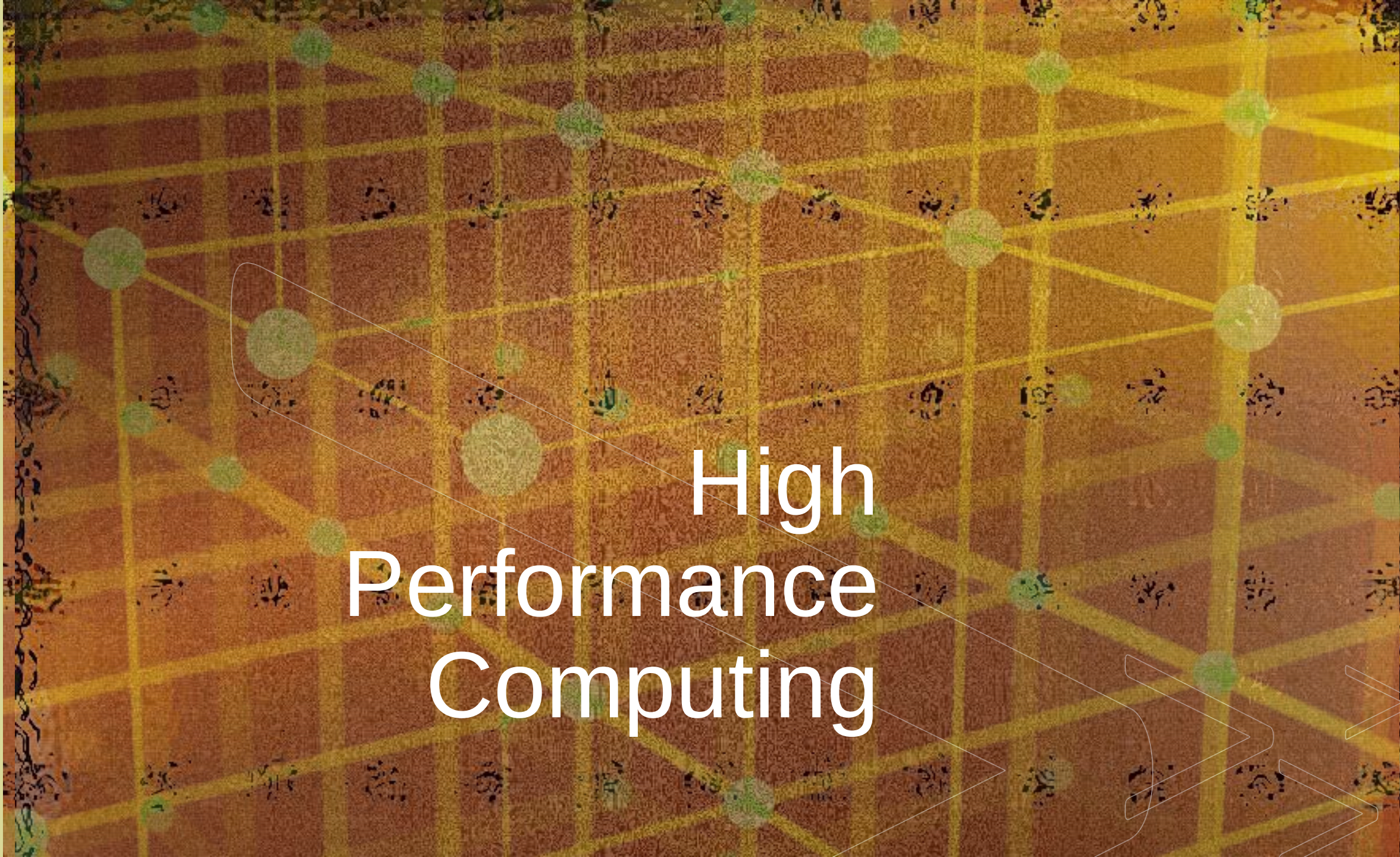
(Lie, 2023)

Fine-grained Dataflow

- Dataflow: computation is triggered by data arrival, and an instruction is executed when all inputs have arrived
- Traditional von Neumann architecture: instructions executed in an order specified by control flow

The Intended Purpose…




High
Performance
Computing

Why?

- Its important to support scientific applications even as AI drives the hardware industry
- Porting irregular applications to dataflow architectures is a new and interesting problem

Computational Molecular Dynamics

- Resolving atomic vibrations at a tiny timestep (10^{-15} sec)
- Simulating long time scales to observe physical phenomena
  - for example, on the order of 100 microseconds
- Month-long exascale runs can at most only simulate a few microseconds

Strong Scaling

- Keep problem size constant, increase the number of processors, and achieve proportional speedup
- Obstacles include:
  - Kernel launch overhead
  - MPI communication costs
- CPU/GPU machines cannot achieve the required performance

MD on the WSE-2

v1 [physics.comp-ph] 13 May 2024

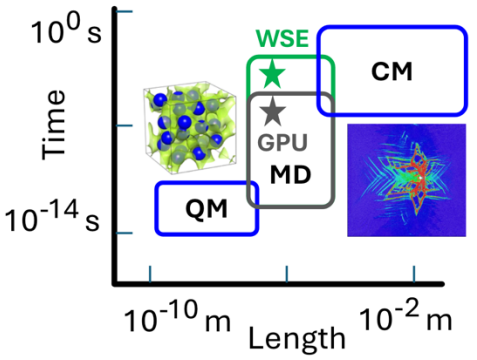
Breaking the Molecular Dynamics Timescale Barrier Using a Wafer-Scale System

Kylee Santos*, Stan Moore†, Tomas Oppelstrup‡, Amirali Sharifian*, Ilya Sharapov*, Aidan Thompson†, Delyan Z Kalchev*, Danny Perez§, Robert Schreiber*, Scott Pakin§, Edgar A. Leon‡, James H Laros III†, Michael James*, and Sivasankaran Rajamanickam†
*Cerebras Systems, Sunnyvale, CA
†Sandia National Laboratories, Albuquerque, NM
‡Lawrence Livermore National Laboratory, Livermore, CA
§Los Alamos National Laboratory, Los Alamos, NM

***Abstract*—Molecular dynamics (MD) simulations have transformed our understanding of the nanoscale, driving breakthroughs in materials science, computational chemistry, and several other fields, including biophysics and drug design. Even on exascale supercomputers, however, runtimes are excessive for systems and timescales of scientific interest. Here, we demonstrate strong scaling of MD simulations on the Cerebras Wafer-Scale Engine. By dedicating a processor core for each simulated atom, we demonstrate a 179-fold improvement in timesteps per second versus the Frontier GPU-based Exascale platform, along with a large improvement in timesteps per unit energy. Reducing every year of runtime to two days unlocks currently inaccessible timescales of slow microstructure transformation processes that are critical for understanding material behavior and function.**

Our dataflow algorithm runs Embedded Atom Method (EAM) simulations at rates over 270,000 timesteps per second for problems with up to 800k atoms. This demonstrated performance is unprecedented for general-purpose processing cores.

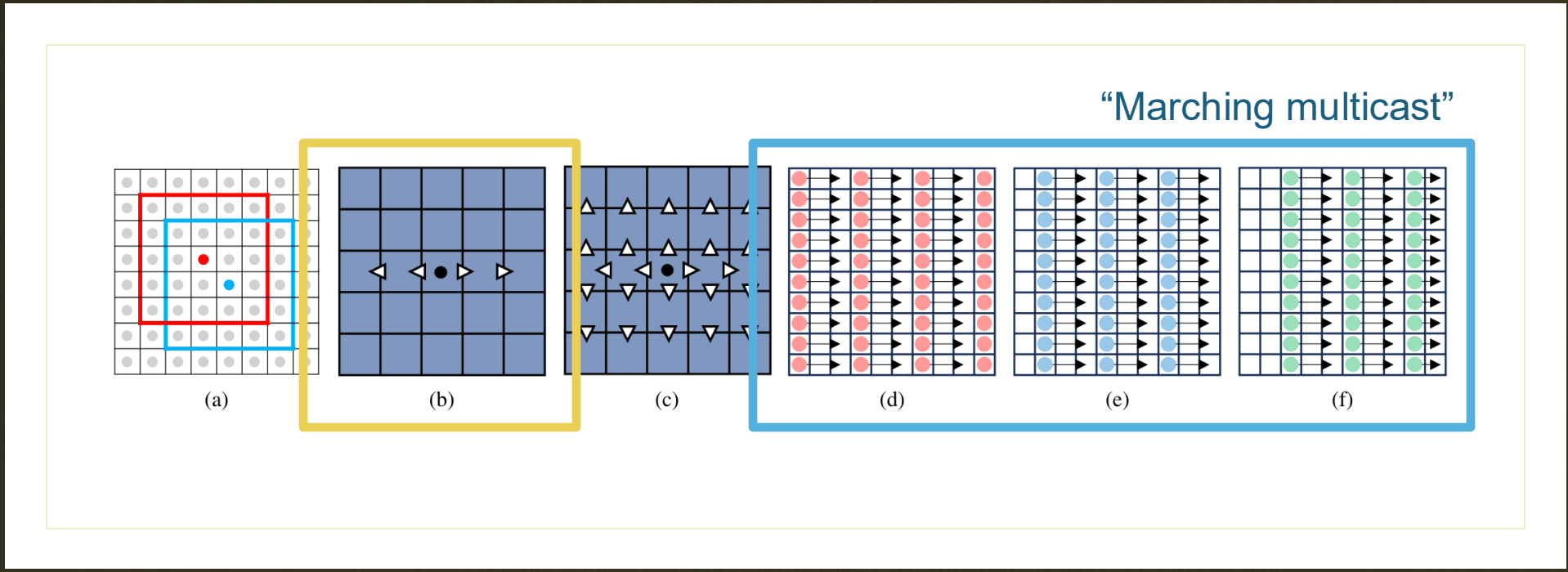
***Index Terms*—wafer-scale engine, molecular dynamics, materials, EAM, strong scaling**



(Santos et al., 2024)

Mapping Atoms to PEs

- Atom-based MD simulation
- Modeling Tungsten
- Map one atom per WSE-2 core
- Mapping is *locality preserving*

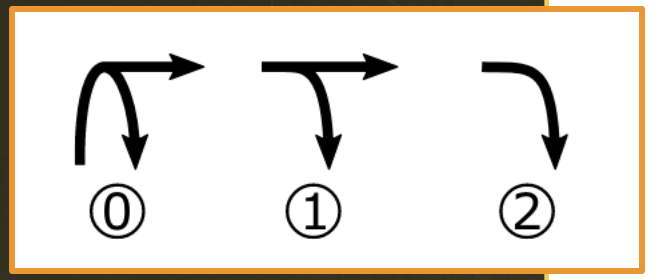


(Santos et al., 2024)

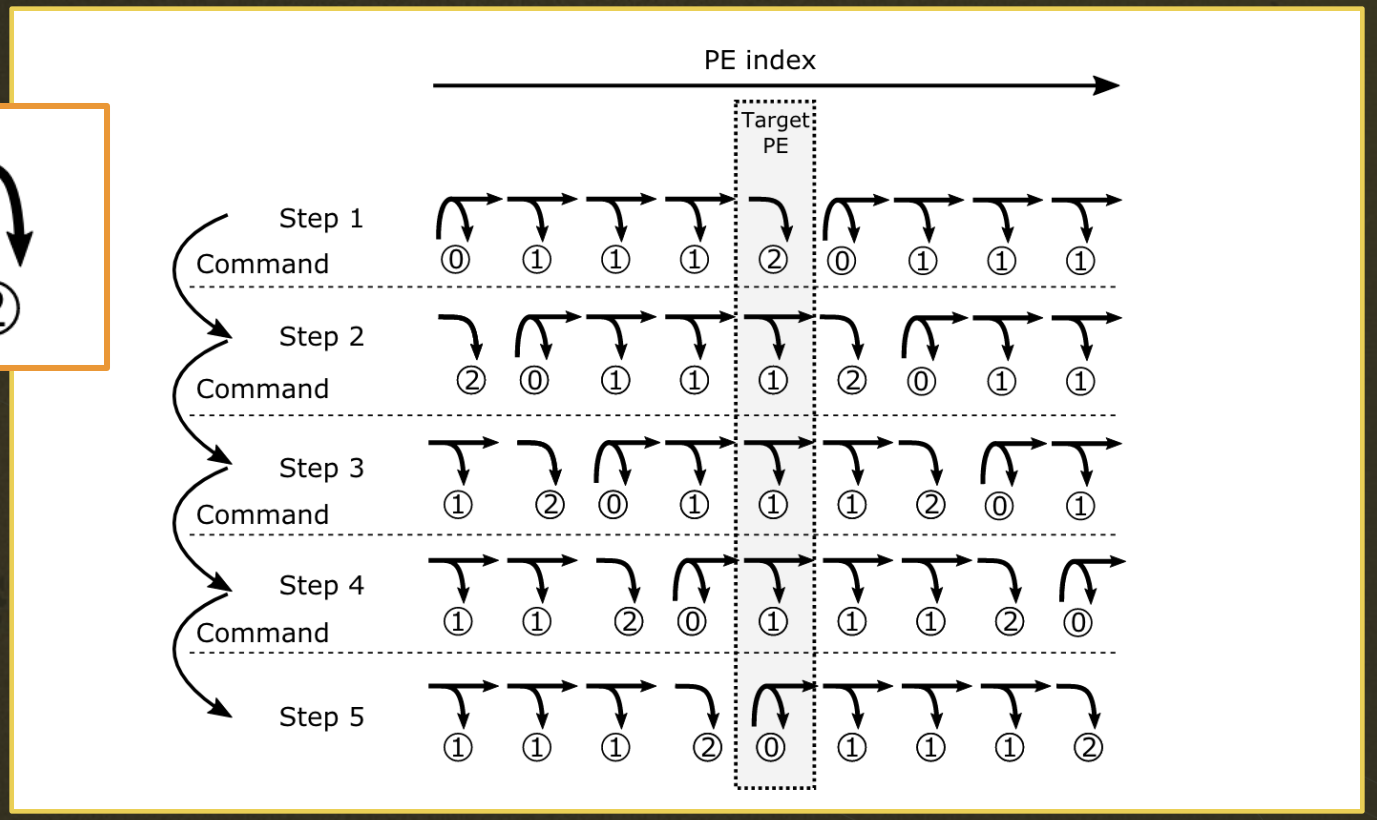
Candidate Exchange

- At every time step, particles exchange data with their neighborhood in all-to-all communication
- Marching multicast phases are used to prevent link contention

Router Message Configurations



Possible router configurations



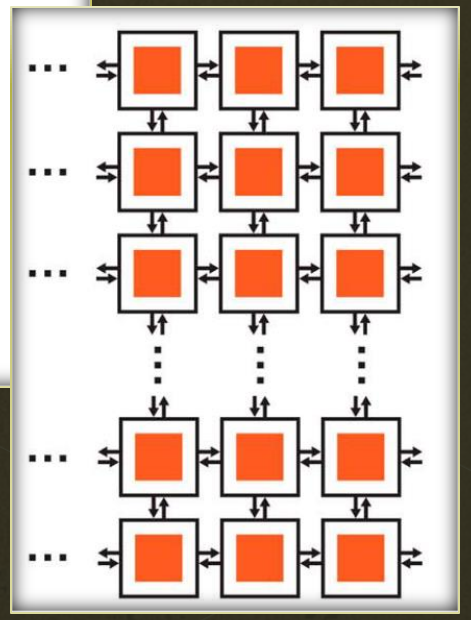
Marching multicast horizontal phase

(Jacquelin et al., 2022)

Tungsten Implementation

```
118         /* Process horizontal transfer */
119         parallel {
120             {
121                 ∀p lr[] ← payload[p];
122                 ∀s lr[] ← control(mcast_ctrl[s]);
123             }
124             {
125                 ∀p rl[] ← payload[p];
126                 ∀s rl[] ← control(mcast_ctrl[s]);
127             }
128             ∀l ∀p row.half.left[l][p]  ← lr[];
129             ∀r ∀p row.half.right[r][p] ← rl[];
130         }
```

https://github.com/CerebrasResearch/Cerebras-Trilabs/



Performance

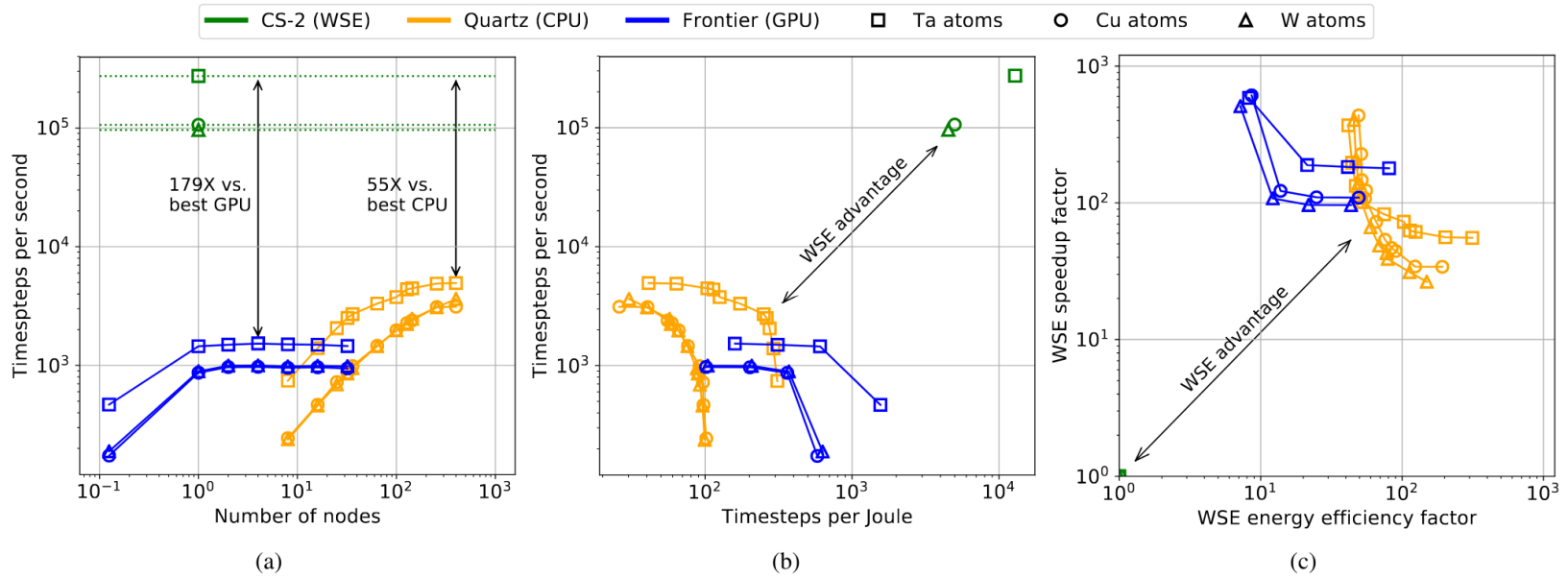


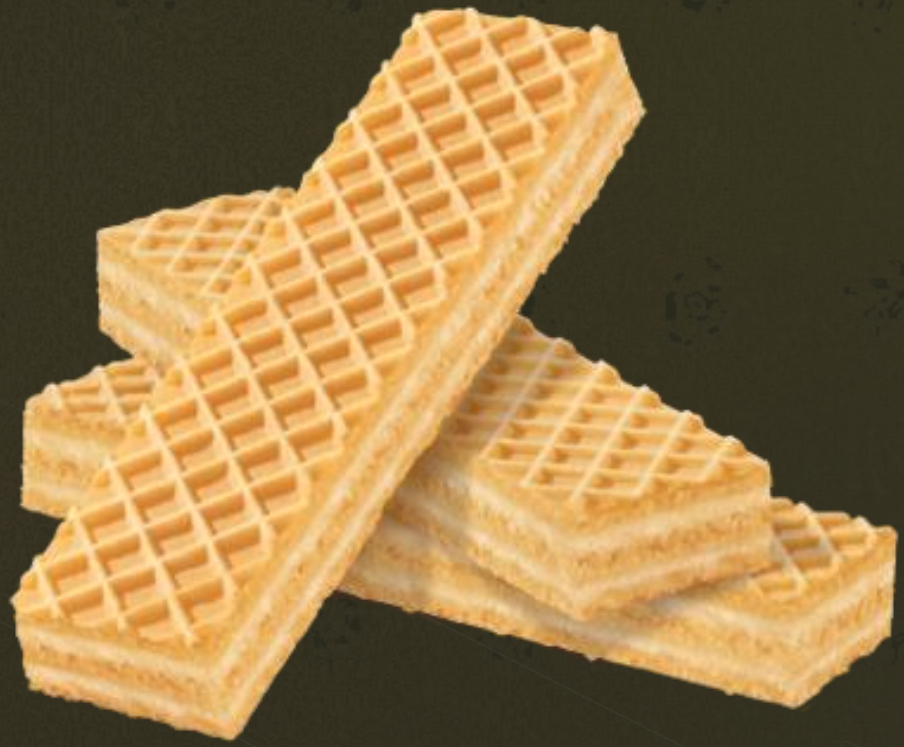
Fig. 7. Measured performance and energy efficiency of single WSE compared to multi-node GPU and CPU systems for Ta, Cu, and W EAM benchmark simulations with 801,792 atoms. **(a)** For Ta, WSE (green square) achieved 179x and 55x speedup compared to the maximum simulation rates on GPU (blue squares) and CPU (orange squares) systems, respectively; **(b)** WSE also demonstrated one to two orders of magnitude improvement in energy efficiency over both CPU and GPU systems; **(c)** Relative energy efficiency and performance of CPU and GPU systems compared to WSE, showing Pareto front dominance of WSE on both metrics.

(Jacquelin et al., 2022)

Conclusion

- Summary
  - Cerebras WSE-2 Dataflow architecture
  - MD simulation and candidate exchange algorithm
- My interests
  - Intersection of architecture and HPC
  - Task-based programming models for dataflow

Questions?



Citations and Further Reading

- Brown, N., Echols, B., Zarins, J., & Grosser, T. (2022). *TensorFlow as a DSL for stencil-based computation on the Cerebras Wafer Scale Engine* (arXiv:2210.04795). arXiv. https://doi.org/10.48550/arXiv.2210.04795
- Jacquelin, M., Araya-Polo, M., & Meng, J. (2022). *Massively scalable stencil algorithm* (arXiv:2204.03775). arXiv. http://arxiv.org/abs/2204.03775
- Lauterbach, G. (2021). The Path to Successful Wafer-Scale Integration: The Cerebras Story. *IEEE Micro*, *41*(6), 52–57. IEEE Micro. https://doi.org/10.1109/MM.2021.3112025
- Lie, S. (2023). Cerebras Architecture Deep Dive: First Look Inside the Hardware/Software Co-Design for Deep Learning. *IEEE Micro*, *43*(3), 18–30. https://doi.org/10.1109/mm.2023.3256384
- Orenes-Vera, M., Sharapov, I., Schreiber, R., Jacquelin, M., Vandermersch, P., & Chetlur, S. (2023). Wafer-Scale Fast Fourier Transforms. *Proceedings of the 37th International Conference on Supercomputing*, 180–191. https://doi.org/10.1145/3577193.3593708
- Rocki, K., Van Essendelft, D., Sharapov, I., Schreiber, R., Morrison, M., Kibardin, V., Portnoy, A., Dietiker, J. F., Syamlal, M., & James, M. (2020). *Fast Stencil-Code Computation on a Wafer-Scale Processor* (arXiv:2010.03660). arXiv. https://doi.org/10.48550/arXiv.2010.03660
- Santos, K., Moore, S., Oppelstrup, T., Sharifian, A., Sharapov, I., Thompson, A., Kalchev, D. Z., Perez, D., Schreiber, R., Pakin, S., Leon, E. A., Laros III, J. H., James, M., & Rajamanickam, S. (2024). *Breaking the Molecular Dynamics Timescale Barrier Using a Wafer-Scale System* (arXiv:2405.07898). arXiv. http://arxiv.org/abs/2405.07898
- Tramm, J., Allen, B., Yoshii, K., Siegel, A., & Wilson, L. (2023). *Efficient Algorithms for Monte Carlo Particle Transport on AI Accelerator Hardware* (arXiv:2311.01739). arXiv. https://doi.org/10.48550/arXiv.2311.01739
- Woo, M., Jordan, T., Schreiber, R., Sharapov, I., Muhammad, S., Koneru, A., James, M., & Van Essendelft, D. (2022). *Disruptive Changes in Field Equation Modeling: A Simple Interface for Wafer Scale Engines* (arXiv:2209.13768). arXiv. https://doi.org/10.48550/arXiv.2209.13768