# Towards Acceleration of Sparse Matrix-Vector Multiplication on the Cerebras CS-2

Saho Orihara, Koki Sakai and Takaaki Miyajima
Graduate school of Science and Technology, Meiji University, Japan, e-mail:ce255032@meji.ac.jp

明治大学 MEIJI UNIVERSITY

## Overview

Sparse matrix–vector multiplication (SpMV) is a fundamental computational kernel in scientific computing, but its scalability on conventional multi-GPU systems is often constrained by communication overhead. The Cerebras CS-2 features a large two-dimensional mesh of processing elements (PEs) with high-speed neighbor-to-neighbor communication, making it well-suited for communication-intensive workloads. In this study, we implement a two-dimensional block-partitioned SpMV on the CS-2 and evaluate its performance using the Poisson3Da matrix on PE configurations ranging from 30×30 to 60×60. Strong-scaling results show that execution time decreases as the number of PEs increases, while parallel efficiency declines. A heatmap of computation, communication, and reduction-cycle counts on each PE reveals that computation time strongly correlates with the distribution of nonzero elements in the matrix. This imbalance causes communication and reduction to start asynchronously across PEs. However, the overall execution time is dominated by the PE with the latest computation completion, indicating that performance is limited primarily by load imbalance induced by the matrix structure rather than by communication cost.

## Wafer Scale Engine 2 (WSE-2)

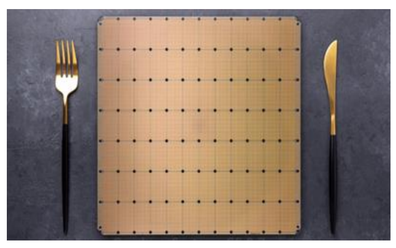
### The entire 300 mm wafer is used as one chip

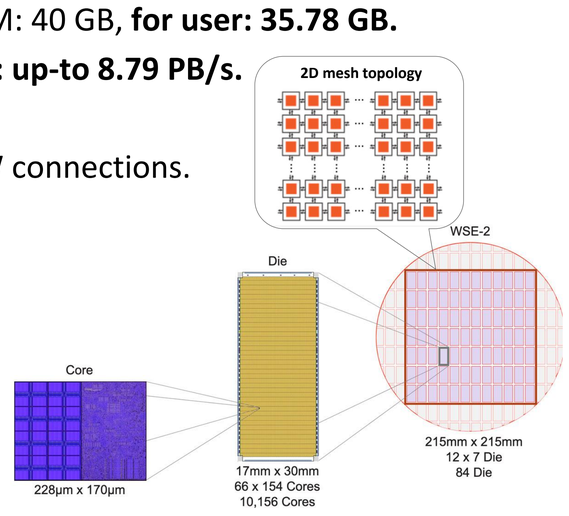
- 84 dies are interconnected via a parallel interface called Cross Scribe Line Connections.
- Process rule: TSMC 7 nm, chip area: 46225 mm^2.
- 853104 processing elements (PEs) are arranged in a 2D mesh topology.
- Each PE consists of memory, a compute element, and a router.

Cerebras CS-2

### Specifications

- Total number of PEs: 853104 (792x1078), **for user: 745500 (750x994).**
- Total number amount of on-chip SRAM: 40 GB, **for user: 35.78 GB.**
- Memory bandwidth: 20 PB/s, **for user: up-to 8.79 PB/s.**
- Inter-PE interconnect: 220 Pb/s.
- PE connectivity: bidirectional N/S/E/W connections.
- Running at 850MHz.


WSE-2 Chip

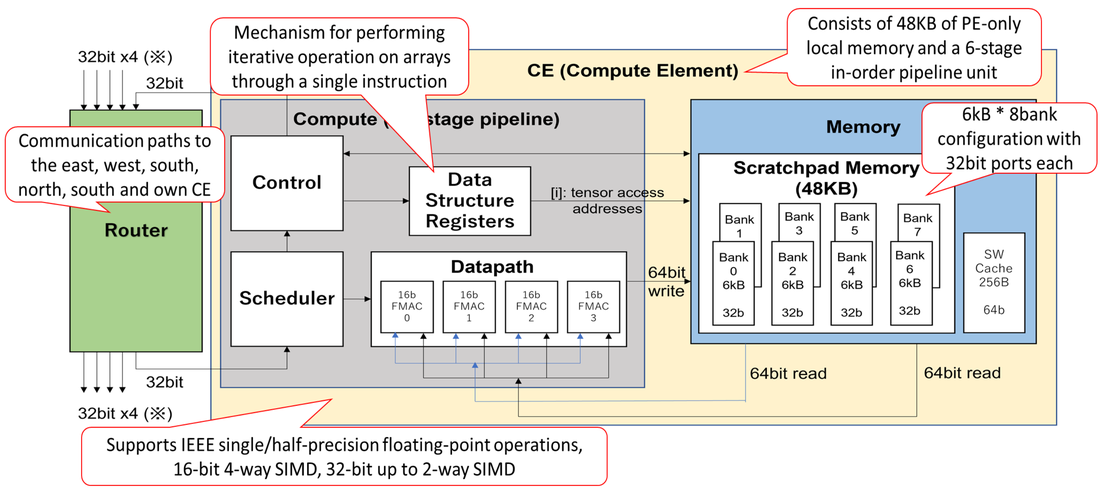


Physical structure of WSE-2

## Processing Element (PE)

### PE comes with scratchpad memory, computing element (CE) and router

- CE and memory are connected via two 64-bit read ports and one write port.
- There is no dedicated hardware unit for matrix operations.



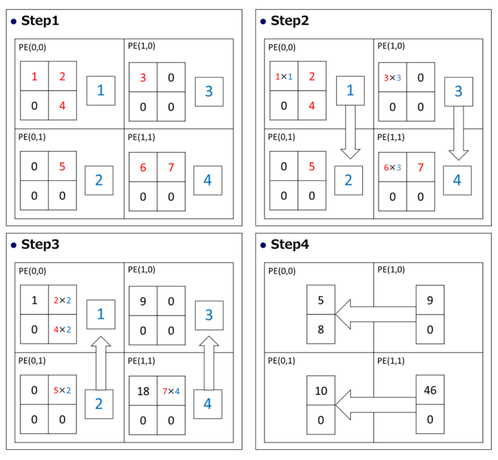
Structure of the CE

## Implementation SpMV in Cerebras CSL

### Algorithm overview

- Our 2-D SpMV on CS-2 comprised local computation, Y-axis broadcast, and reduction.

1. The matrix is 2D-partitioned and mapped onto the PE grid.
2. The x vector is sequentially distributed, starting from the top-row PEs.
3. Each PE performs local SpMV.
4. Results are aggregated leftward to generate the output vector.



Computation Algorithm

## Evaluation method

### Experimental Setup

- The Poisson3Da matrix is used as the evaluation target.
- Poisson3Da is a 13514×13514 sparse matrix with 352762 non-zero values.
- The ratio of non-zero elements to all matrix entries is approximately 0.19%, indicating high sparsity.
- The matrix has a regular structure, with an almost constant number of non-zero elements per row.
- This matrix is derived from the discretization of a three-dimensional Poisson problem and is included in the SuiteSparse Matrix Collection.
- Measurement environment: Simulator in Cerebras SDK version: 1.3.0

## Evaluation on the simulator

### Strong-Scaling Configuration

- In this strong-scaling experiment, the problem size (Poisson3Da) is fixed, while the PE grid size is increased from 30×30 to 60×60 in steps of 5.
- Speedup and parallel efficiency are computed using the 30×30 configuration as the baseline.

### Result

- Out implementation achieves a 2.17-fold speedup when expanding the used PEs from 30×30 to 60×60.
- Performance increases with the number of PEs, reaching up to 5.44 GFLOP/s.
- Parallel efficiency decreases gradually as the number of PEs increases.

| PE grid | PEs | Cycles | Time [ms] | Speedup | Parallel efficiency | GFLOP/s |
|---|---|---|---|---|---|---|
| 30 × 30 | 900 | 238854 | 0.28 | 1.00 | 1.00 | 2.51 |
| 35 × 35 | 1225 | 204946 | 0.24 | 1.17 | 0.86 | 2.93 |
| 40 × 40 | 1600 | 181624 | 0.21 | 1.32 | 0.74 | 3.30 |
| 45 × 45 | 2025 | 154926 | 0.18 | 1.54 | 0.69 | 3.87 |
| 50 × 50 | 2500 | 140927 | 0.17 | 1.69 | 0.61 | 4.26 |
| 55 × 55 | 3025 | 122123 | 0.14 | 1.96 | 0.58 | 4.91 |
| 60 × 60 | 3600 | 110133 | 0.13 | 2.17 | 0.54 | 5.44 |

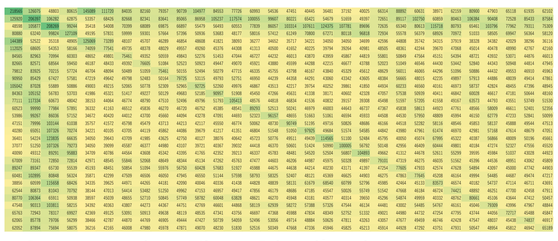
Strong Scaling of SpMV on multiple PEs

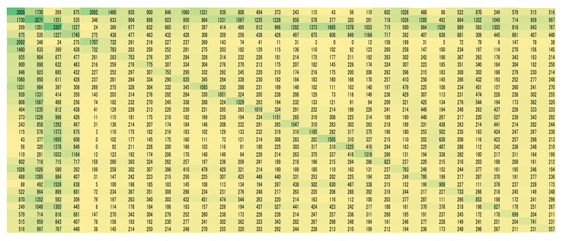
### Runtime breakdown visualization using heatmap

- Computation, Y-axis broadcast, and reduction times are visualized as heatmaps mapped to PE locations on a 30×30 grid.
- Yellow indicates fewer cycles, while green indicates larger cycle counts.

### Local computation part

- Computation time varies significantly across PEs due to the matrix nonzero distribution, and the slowest PE ultimately limits entire processing time.


Computation Time Distribution Across PEs

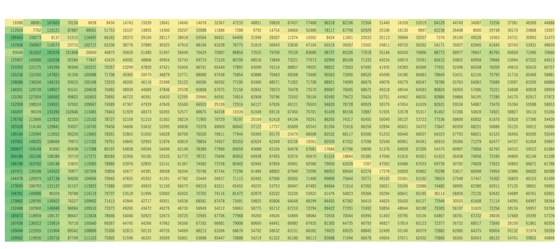

Nonzero Distribution of the Poisson3Da Matrix

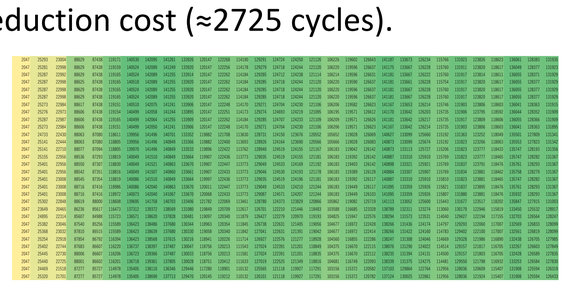
### Y-axis broadcast part

- Vector communication itself is constant (≈260 cycles), but delayed communication start on slower PEs dominates column-wise communication completion
- As a result, the observed waiting time originates from computation imbalance rather than communication overhead.

### Reduction part

- The large reduction time mainly reflects waiting due to computation imbalance, rather than the intrinsic reduction cost (≈2725 cycles).


Communication Time Distribution Across PEs


Reduction Time Distribution Across PEs

## Future Work

Our results show that computation imbalance leads to increased waiting time during communication and reduction.As future work, we plan to balance these workloads to shorten the overall execution time.

## Acknowledgement and references

[1] Ryunosuke Matsuzaki, Daichi Mukunoki, and Takaaki Miyajima. Performance evaluation and modelling of single-precision matrix multiplication on cerebras cs-2. In Proceedings of the SC '24 Workshops of the International Conference on High Performance Computing, Network, Storage, and Analysis, SC-W '24, page 727–731. IEEE Press, 2025.
[2]Leon Fukuoka and Takaaki Miyajima.
Performance evaluation of inter-processing-element communication on Cerebras CS-2.
IPSJ SIG Technical Report, 2024-HPC-197, Information Processing Society of Japan, 2024.
This work was supported by Japan Science and Technology Agency (JST) as part of Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE), Grant Number JPMJAP2341. This work was supported by JSPS KAKENHI Grant Number 24K14972.