Cornell Bowers CIS
Computer Science

Sparse Matrix-Vector Multiplication on Cerebras WSE-3

Zenchang Sun*, Arman Margarian*, Jefferson Zhou*, Giulia Guidi
* Equal contribution · Cornell University

1. Population Genomics Becomes SpMV

Population-genomics analyses repeatedly aggregate mutation effects across hundreds of thousands of haplotypes. A Genotype Representation Graph (GRG) stores shared ancestry compactly; we materialize its implied 5.28B-nonzero genotype matrix as a baseline for SpMV. CPUs and GPUs face irregular off-chip memory traffic, while WSE-3 keeps sparse tiles in local SRAM and combines results on-wafer:

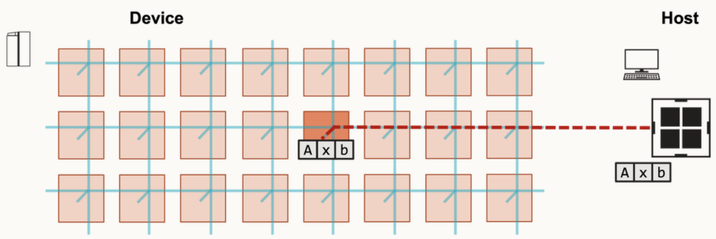
$$y = Gx$$

G: haplotype × mutation incidence · **x**: mutation weights · **y**: haplotype totals.

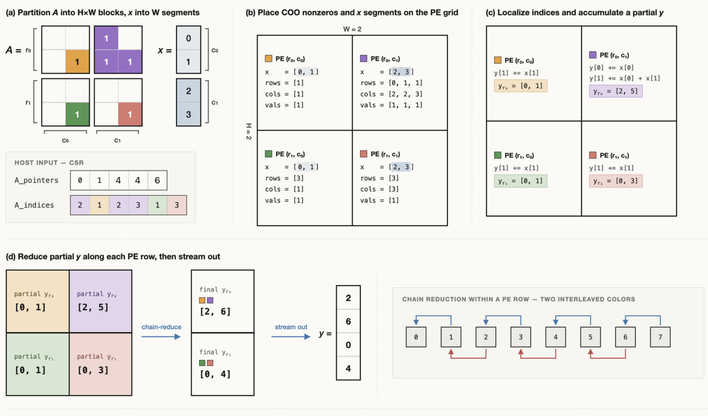
200,000	2,502,559	>98.9%
haplotypes	mutations	sparse

2. WSE-3 Architecture

A 755 × 1,170 subgrid stores one sparse tile per PE in 48 KiB of local SRAM. Single-cycle neighbor links keep computation and communication on wafer.



3. Partition, Compute, and Reduction



On-wafer kernel. Each PE computes a partial y; row-wise partials reduce eastward.

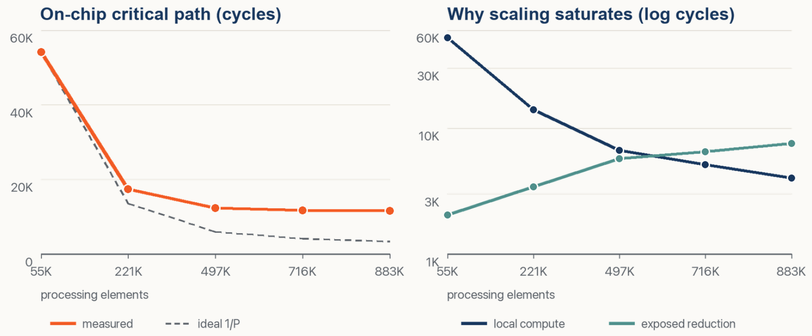
This work: measures binary SpMV on WSE-3 across 17 configurations and identifies its one-axis reduction limit.

5.28B-nonzero case. Three passes sum to 138,773 on-chip cycles (≈163 µs at 850 MHz*; host work excluded), or 32.3 Tnnz/s (64,700 effective GFLOP/s).

Tnnz/s = trillion nonzeros/s. Effective FLOP rate counts 2 operations per nonzero; the binary kernel omits the multiply.

4. Strong Scaling

16× more PEs, 4.68× speedup.

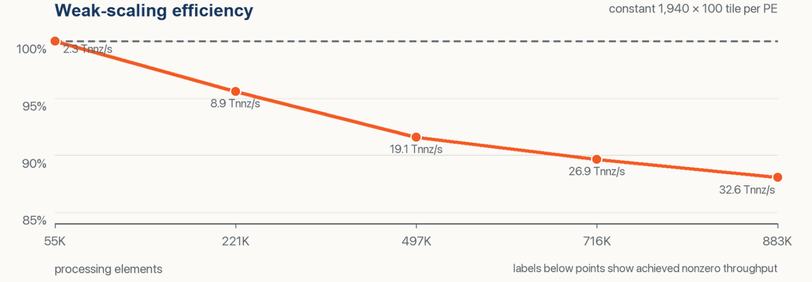


More PEs cut local work only 13× (versus 16× ideal) as the busiest PE becomes more imbalanced. The longer eastbound reduction then dominates at large grids.

Fixed 145.1M-nonzero matrix: runtime falls from **54,114 to 11,561 cycles**; the measured reduction remainder rises from 2,048 to 7,555 cycles.

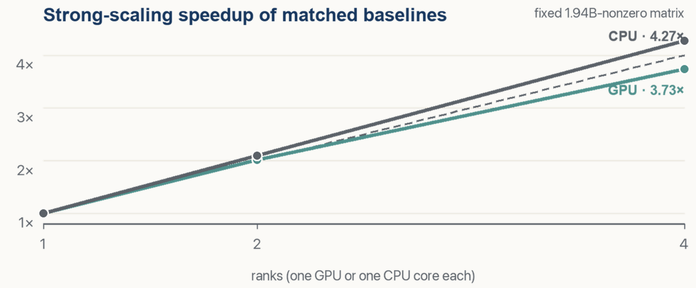
5. Weak Scaling

32.6 Tnnz/s at 883,350 PEs.



With fixed work per PE, local time stays near 43K cycles. Reduction grows from **1,927 to 7,600 cycles**, lowering efficiency from 100% to 88%.

6. CPU & GPU Baseline Scaling

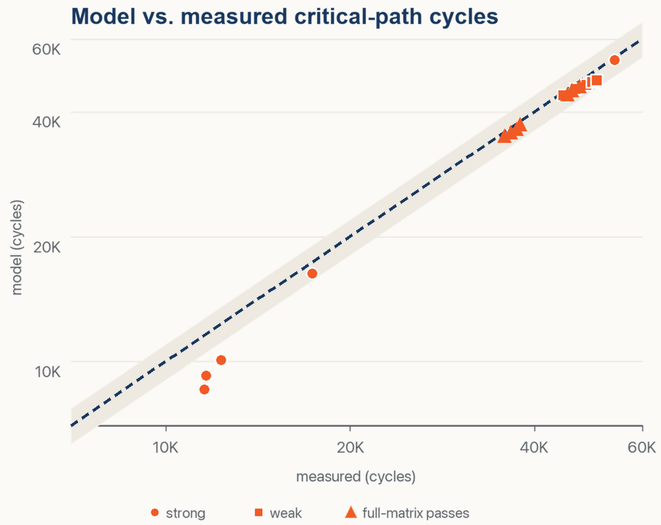


At four ranks, GPU and CPU baselines achieve **3.73×** and **4.27×** strong-scaling speedup (CPU superlinear from cache residency), with **≥93%** weak-scaling efficiency. SciPy CSR (CPU); cuSPARSE CSR (GPU).

7. Performance Model

Local cycles fit the busiest PE. The measured remainder, max(total) − max(local), is compared with a published WSE-2 chain model; overlap is not measured separately.

$$T_{model} = 46 + 18 \cdot max_nnz + rows_per_PE + 6(W-1)$$



The local term has zero in-sample residual across 17 runs. The measured remainder grows ~10 cycles per added PE column versus 6 in the borrowed model; total error reaches **26%** at full width.

8. Platform Comparison

Four-A100 comparison. On the same 1.94B-nonzero matrix, WSE-3 reports **50,648 on-chip cycles (≈59.6 µs at 850 MHz*)** versus **5.46 ms** on four NVIDIA A100s—about **92× less reported time**.

Reduction included · WSE-3 excludes matrix loading; CPU/GPU times exclude host↔device transfers

PLATFORM	REPORTED TIME	TNNZ/S
WSE-3 · 883K PEs	**0.0596 ms**	**32.6**
4× NVIDIA A100	5.46 ms	0.355
EPYC 7763 · 4 cores	434 ms	0.00447
EPYC 7763 · 64 cores	93.8 ms	0.0207

Illustrative peak-power proxy, not measured energy. Published maxima (23 kW for CS-3; assuming 400 W per A100) give P_max·T ≈ 1.37 versus 8.74 J per SpMV, a ~6.4× proxy ratio; GPU host power is excluded.

9. Conclusion

Biobank scale: 5.28B nonzeros across three passes in 138,773 summed on-chip cycles (~163 µs at 850 MHz).

Local compute fits the 17 measured runs with zero in-sample residual; load imbalance also limits scaling.

Reduction dominates at large grids: ~10 cycles per added PE column observed versus 6 modeled.

With transfers excluded on both, WSE-3 reports ~92× less time than four A100s.

Acknowledgments. Max Zhao and Luk Burchard, Cerebras Systems; Yifan Li and Qingyao Sun for GRG expertise; Pittsburgh Supercomputing Center for WSE-3 access. CPU/GPU baselines used NERSC Perlmutter. ChatGPT and Claude assisted with text and figure generation; authors verified all content.

* WSE-3 time and throughput estimates convert on-chip cycle counts at 850 MHz (Cerebras Systems, personal communication).

References.
[1] D. DeHaas et al., Nat. Comput. Sci., vol. 5, pp. 112–124, 2025.
[2] F. Dobrosavljević et al., "Sparse Matrix Multiplication on Cerebras WSE-2," 2023.
[3] P. Luczynski et al., HPDC '24, pp. 334–347, 2024.