Sparse Matrix-Vector Multiplication on Cerebras WSE-3

Zenchang Sun*
zs434@cornell.edu
Cornell University
Ithaca, New York, USA

Arman Margarian*
am2738@cornell.edu
Cornell University
Ithaca, New York, USA

Jefferson Zhou*
jz2336@cornell.edu
Cornell University
Ithaca, New York, USA

Giulia Guidi†
gg434@cornell.edu
Cornell University
Ithaca, New York, USA

1 Introduction

Biobank-scale genomic studies now span hundreds of thousands of genomes, and their computational core is sparse matrix-vector multiplication (SpMV), $y = Ax$, bounded by data movement rather than arithmetic. The Cerebras Wafer-Scale Engine 3 (WSE-3) addresses this in hardware: ~893,000 processing elements (PEs) on a 2D on-chip fabric, each with 48 KiB of local memory and nearest-neighbor links, hold the working set on chip. Our tile-wise Cerebras Software Language (CSL) kernel partitions A with a deterministic host-side balancer and sums per-PE partials eastward along each PE row. On the full wafer it sustains 64,700 effective GFLOP/s on a 5.28 B-nonzero genotype matrix. Local compute scales to 883k PEs and a closed-form model fits it exactly. The reduction phase is the dominant limit on total scaling.

2 Methods

Workload. A Genotype Representation Graph (GRG) [1] compresses population genotype data into a DAG whose internal nodes capture ancestry shared among haplotypes. The genotype matrix $G \in \{0,1\}^{n\times m}$, for n haplotypes and m mutations, follows from the GRG adjacency A_{G} by reachability. GRG-native analyses apply repeated sparse products against A_{G}, never forming G. We instead materialize G to isolate the kernel, evaluating $y = Ax$ with $A = G$, x weighting each mutation, and y accumulating each haplotype's total. Our simulated GRG instance has 200,000 haplotypes, 2,502,559 mutations, and 5.28 B nonzeros (98.9% sparse).

Partitioning and per-PE kernel (Figure 1). G is tiled across a $W \times H$ PE subgrid following the tile–broadcast–reduce architecture of Dobrosavljević et al. [2], with x sliced along the column axis. Their kernels degrade past 95% sparsity; ours is, to our knowledge, the first Cerebras wafer-scale SpMV specialized to the binary, >98%-sparse regime. Each PE stores its block as nonzero coordinates with 32-bit indices local to its own slice of x, and accumulates each fetched entry into the partial y.

Load balancing. Runtime is set by the fullest PE, so we minimize its nonzero count. Genotype matrices concentrate nonzeros in high-frequency mutation columns, so a naive partition overloads a few PE-columns. We sort columns by nonzero count and place them greedily onto the least-loaded PE-column, a longest-processing-time schedule run host-side.

Reduction. A row's result sums the W partial y vectors in that row, formed with the contention-optimal chain reduce of Luczynski et al. [3]. Unlike their standalone collective, ours is fused into the SpMV and pipelined with local compute. We therefore report T_{reduce} as $\max(T_{\mathrm{total}}) - \max(T_{\mathrm{local}})$, the non-overlapped remainder.

3 Performance Model

We model the critical path as the local phase plus that non-overlapped overhead:

$$T_{\mathrm{total}} = \underbrace{\alpha + \beta \cdot \texttt{max_pe_nnz}}_{T_{\mathrm{local}}} + \underbrace{\texttt{rows_per_PE} + 6\,(W-1)}_{T_{\mathrm{reduce}}} \quad (1)$$

The local coefficients come from least squares of local cycles against per-PE nonzero count; because the phase incurs no branch misprediction, cache miss, or off-chip traffic, residuals are effectively zero. The reduction term comes from the $T_{\mathrm{chain}} = B + (2T_R + 2)(P-1)$ formula of [3], with their fitted WSE-2 ramp latency $T_R = 2$ and P the PEs per row.

4 Results and Conclusions

Setup. We report the critical-path PE's on-chip timestamp count, sweeping grid size as a fraction ρ of a $755 \times 1{,}170$ subgrid of the $762 \times 1{,}172$ fabric. Strong scaling fixes a $32{,}000 \times 400{,}000$ matrix and grows the grid; weak scaling fixes a $1{,}940 \times 100$ tile per PE and grows the matrix with it, both on synthetic matrices matching the biobank distribution. Biobank-scale runs split G into k column blocks. k=3 is the smallest k that fits on-chip and host memory.

Scaling. Under strong scaling, a 16× grid increase yields only 4.7× speedup: local compute falls 13× against an ideal 16× (load imbalance), while the reduction rises 3.7× as the chain lengthens, crossing near ρ=0.8. Weak scaling isolates the effect: local time stays flat while the reduction grows 3.9× for essentially the entire runtime increase, dropping efficiency to 88%. Both reflect a reduction that parallelizes along only one grid axis.

CPU and GPU baselines. On the identical $117{,}000 \times 1{,}464{,}700$ matrix, WSE-3 completes in 59.6 μs versus 5.46 ms on four A100 GPUs and 93.8 ms on a 64-core AMD EPYC 7763: gaps of 92× and 1,574×. Both sweeps, repeated with SciPy and cuSPARSE CSR, a 1-D column partition, and MPI allreduce, scale well: 3.73× (4.27×) strong-scaling speedup at P=4 on GPUs (cores), the latter superlinear from cache residency, and weak-scaling efficiency above 0.93.

*All three authors contributed equally to this research.
†Advisor.

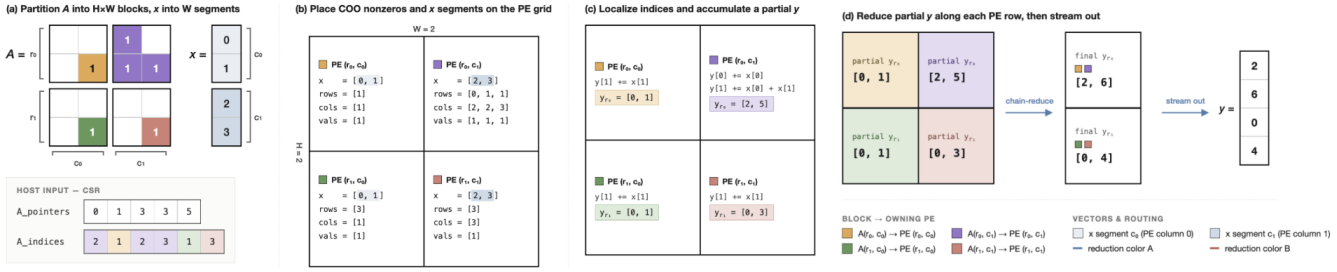


Figure 1: Tile-wise partitioning and eastward reduction. (a–c) G is split across a $W \times H$ PE subgrid, with x sliced along the column axis so each PE holds only the entries of x its block needs, stored as nonzero coordinates with indices local to that slice. (d) The W partial y vectors in a row are summed eastward: each PE accumulates its west neighbor's partial and forwards the running sum, so the easternmost PE ends with the finished row result.

Table 1: Every measured WSE-3 configuration. The local term $46 + 18 \cdot$ `max_pe_nnz` reproduces measured local cycles exactly (0.0% error) in all 17 runs, so only total-runtime error is shown, taken against the reduction cost published by Luczynski et al. [3]. That error is negative throughout and grows monotonically with ρ as the reduction comes to dominate. T_{reduce} is the non-overlapped remainder $T_{\text{total}} - T_{\text{local}}$. The total column is the critical-path PE's count, local is the maximum over all PEs, and reduce their difference; wall-clock times are those counts at the 850 MHz clock based on discussions with Cerebras Systems. Bold rows sum their constituent passes over the full 5.28 B-nonzero matrix and so carry no independent model error.

	Fabric		Problem			Measured cycles				
Configuration	Grid $W \times H$	PEs	Matrix $M \times K$	total_nnz	max_pe_nnz	local	reduce	total	Time (μs)	Err. (%)
Strong scaling — fixed matrix, growing grid										
strong ρ=0.25	189×292	55,188	32k×400k	145.1 M	2,890	52,066	2,048	54,114	63.7	−1.5
strong ρ=0.5	378×585	221,130	32k×400k	145.1 M	773	13,960	3,410	17,370	20.4	−6.3
strong ρ=0.75	566×878	496,948	32k×400k	145.1 M	366	6,634	5,692	12,326	14.5	−18.4
strong ρ=0.9	680×1053	716,040	32k×400k	145.1 M	282	5,122	6,515	11,637	13.7	−20.7
strong ρ=1	755×1170	883,350	32k×400k	145.1 M	220	4,006	7,555	11,561	13.6	−26.0
Weak scaling — fixed 1,940 × 100 tile per PE										
weak ρ=0.25	189×292	55,188	29.2k×366.7k	121.4 M	2,368	42,670	1,927	44,597	52.5	−1.6
weak ρ=0.5	378×585	221,130	58.5k×733.3k	485.9 M	2,389	43,048	3,600	46,648	54.9	−2.7
weak ρ=0.75	566×878	496,948	87.8k×1.10M	1.09 B	2,383	42,940	5,752	48,692	57.3	−4.6
weak ρ=0.9	680×1053	716,040	105.3k×1.32M	1.58 B	2,387	43,012	6,734	49,746	58.5	−5.1
weak ρ=1	755×1170	883,350	117k×1.46M	1.94 B	2,389	43,048	7,600	50,648	59.6	−5.9
Biobank scale, 200,000 × 2,502,559, split into k column blocks										
pass 0	755×1170	883,350	200k×834.2k	1.81 B	2,294	41,338	6,069	47,407	55.8	−2.9
pass 1	755×1170	883,350	200k×834.2k	1.76 B	2,236	40,294	5,814	46,108	54.2	−2.4
pass 2	755×1170	883,350	200k×834.2k	1.71 B	2,181	39,304	5,954	45,258	53.2	−2.8
k=3 total	755×1170	883,350	200k×2.50M	5.28 B	2,294	120,936	17,837	138,773	163.3	—
pass 0	755×1170	883,350	200k×625.6k	1.38 B	1,806	32,554	5,382	37,936	44.6	−1.8
pass 1	755×1170	883,350	200k×625.6k	1.34 B	1,755	31,636	5,766	37,402	44.0	−2.9
pass 2	755×1170	883,350	200k×625.6k	1.30 B	1,715	30,916	5,740	36,656	43.1	−2.9
pass 3	755×1170	883,350	200k×625.6k	1.26 B	1,677	30,232	5,503	35,735	42.0	−2.3
k=4 total	755×1170	883,350	200k×2.50M	5.28 B	1,806	125,338	22,391	147,729	173.8	—

Power envelopes (23 kW for CS-3, 400 W per A100) imply 1.37 versus 8.74 J/SpMV, ~6.4× higher energy efficiency; power was not measured. All times exclude host staging and transfers.

Biobank scale. At k=3 the complete matrix takes 138,773 summed on-chip kernel cycles (Table 1), achieving 64,700 effective GFLOP/s (2·nnz/T, the conventional SpMV binary; A is binary, so hardware issues one add per nonzero). k=4 drops to 60,800 effective GFLOP/s because each pass pays a full chain traversal and finer blocks raise the summed per-pass maximum.

Model accuracy (Table 1). We measure $\alpha = 46$ cycles and $\beta = 18$ cycles per nonzero, fitting local cycles exactly in all 17 runs. The chain term underpredicts the measured reduction by a near-constant 40% — ~10 cycles per hop against the 6 predicted — so total error grows with the reduction's share, to −26% at ρ=1.

Acknowledgments

We thank our collaborators Max Zhao and Luk Burchard at Cerebras Systems for mentorship. We thank Yifan Li and Qingyao Sun for GRG domain expertise.

References

[1] D. DeHaas, Z. Pan, and X. Wei, "Enabling efficient analysis of biobank-scale data with genotype representation graphs," *Nature Computational Science*, vol. 5, pp. 112–124, 2025.

[2] F. Dobrosavljević, E. Derouiche, A. Ivanov, L. Gianinazzi, and T. Hoefler, "Sparse matrix multiplication on Cerebras WSE-2: Evaluating SpMM algorithms in spatial computing," ETH Zürich and NTNU Trondheim, Tech. Rep., 2023.

[3] P. Luczynski, L. Gianinazzi, P. Iff, L. Wilson, D. De Sensi, and T. Hoefler, "Near-optimal wafer-scale reduce," in *Proc. 33rd Int. Symp. High-Performance Parallel and Distributed Computing (HPDC)*, Pisa, Italy, Jun. 2024, pp. 334–347.