Exploring Tree-Based Approaches for Collective Communications on Cerebras WSE-3

Saverio Pasqualoni
KAUST & Sapienza University of Rome

Marco Canini (advisor)
KAUST

Daniele De Sensi (advisor)
Sapienza University of Rome

Abstract

Fixed-pattern collectives on the Cerebras WSE-3 minimize communication rounds, although on a wafer a round costs a few cycles while distance grows with scale. We design and implement tree-based 1D Reduce schedules that minimize communication distance, outperforming fixed-pattern state-of-the-art algorithms by up to 1.63×.

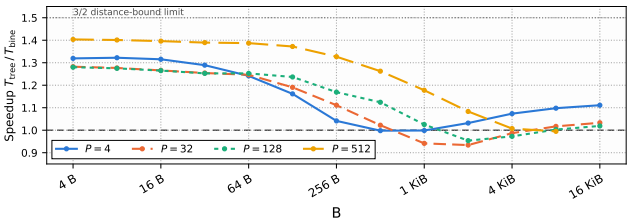


Figure 1: Speedup of the best Bine implementation over the binomial tree. Bine wins by 1.28–1.40× for $B \lesssim 64$ wavelets, the gain growing with P toward the $3/2$ distance-bound limit; as contention dominates, both converge within 3%.

1 Introduction

The Cerebras WSE-3 is a wafer-scale processor: 900,000 processing elements (PEs), organized in a 2D mesh, exchange 32-bit messages (*wavelets*) with their neighbors in a single clock cycle. Collective operations are a key performance factor, and the vendor collectives leave headroom. Existing patterns [1] minimize communication rounds — the right trade on clusters, but on the wafer a round costs a handful of cycles while distance grows with the number of PEs, making locality a first-order design concern.

We develop tree-style schedules minimizing communication distance for 1D Reduce, the building block of 2D collectives and AllReduce (broadcast is near-optimal through hardware multicast [1]). We (i) port Bine trees [2], the locality-optimized binomial trees we recently introduced, to the WSE-3; (ii) implement balanced binary-tree reductions; (iii) introduce unbalanced binary trees; (iv) introduce two-phase variants with tree-based outer phases, retuned analytically; and (v) pair every schedule with a cost model and transport-level optimizations.

2 Cost Model

We adopt the model of [1]: $T = \max\left(C, \frac{E}{N} + L\right) + (2T_R + 1)\, D$, where C is per-PE contention, E total hops over N links, L the longest path, D the depth, and T_R the PE–router ramp latency. The model is a lower bound: transport decisions alone span more than 2×.

3 Bine Trees

Bine (binomial negabinary) trees [2] rebuild binomial trees through negabinary rank encodings, preserving the $\log_2 P$ step count (D) and contention (C) while shrinking energy to $E \approx \frac{B}{3} P \log_2 P$ and maximum distance to $L = \frac{2(P-1)}{3}$, for vectors of B wavelets. Performance depends critically on the transport: for a 4-byte Reduce on 512 PEs, three implementations of the same schedule span 2.2× — a CE relay, each intermediate PE pulling the vector down the ramp and re-sending it (4071 cycles); static routing, wavelets never entering the compute element (2078 cycles); and one synchronous task per rank (1841 cycles). The best variant outperforms the binomial tree by 1.28–1.40× for B up to 64, approaching the $3/2$ distance-bound limit (Figure 1).

4 Binary Trees

Binary trees recursively bisect the rank range, each subroot at the midpoint of its range: distances halve with depth over $\log_2 P - 1$ levels. Since a PE receives one message per round, siblings serialize at the parent's port and the schedule depth doubles to $D = 2(\log_2 P - 1)$; with $E \approx \frac{B}{2} P(\log_2 P - 1)$ and $L = P/2$, the model follows. The unbalanced binary tree cuts each range at one third rather than the median, reducing total distance by 25% at the cost of a deeper schedule. Unfused, a node sends only once its vector is complete; fusing each receive with the onward send, children ordered far-side last, is worth 2.1× (balanced) and 3.1× (unbalanced).

5 Two-Phase Trees

Two-Phase [1] combines the tree's low depth with the chain's low contention: groups of S consecutive PEs chain-reduce to representatives, which chain-reduce among themselves; $S = \sqrt{P}$ balances the two depths. We replace the outer chain with binomial and Bine trees, trading $P/S - 1$ rounds for $\log_2(P/S)$. Retuning yields $S^\star = \frac{1 + cB/(2T_R+1)}{\ln 2}$, with $c = 1/2$ (binomial) or $1/3$ (Bine) — independent of P and linear in the vector length.

6 Evaluation

Times are data-transmission only, from the official cycle-accurate simulator, measured at the root; Auto-Gen [1] is excluded by kind, generating collective code off-chip. Figure 3 maps the per-configuration winner. In latency-bound regions with $P \geq 128$, our trees lead the fixed-pattern state of the art: the unbalanced binary tree by up to 1.63× at 512 PEs, Bine by 1.12× at 128, owing to lower communication distances (Figure 2). At fewer PEs, fixed per-round overheads let Star [1] lead; such configurations matter little in practice: applications engage most of a row [1, 3]. Head to head, the two-phase Bine improves standard two-phase on long vectors by up to 1.12×, although the chain hides the win against SotA by sitting on the

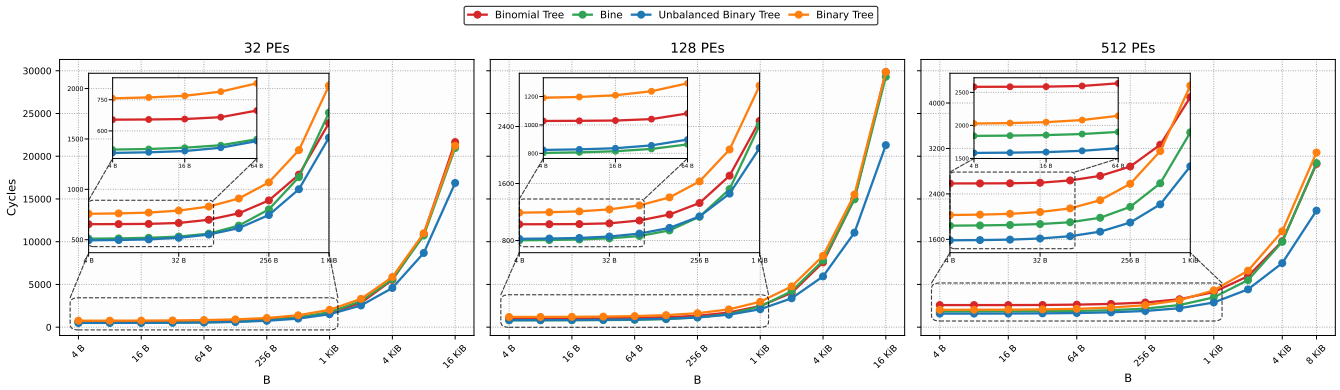


Figure 2: Cycle counts of the best implementation of each tree schedule at 32, 128, and 512 PEs. Bine outruns the binomial tree at 77–92% of sizes, the margin growing with P; the unbalanced binary tree beats the balanced one at every size and scale.

port bound $C = B$. Across the grid, measured per-wavelet costs match the derived schedule depths for every pattern at every scale.

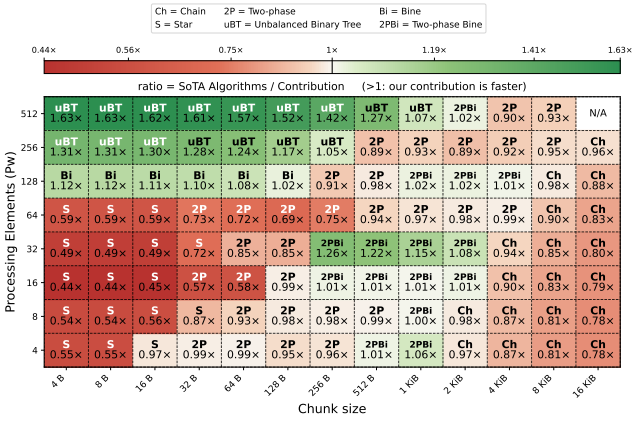


Figure 3: Best fixed pattern [1] against the best of our contributions per configuration (ratio > 1: ours faster). Our trees take the latency-bound region; star and chain hold the smallest problems and the longest vectors.

7 Conclusion and Future Work

On the WSE a round costs five cycles and a hop costs one: trading rounds for distance wins the latency-bound region, up to 1.63× over the fixed-pattern state of the art. Half the result is the implementation, worth 2–3× on the trees that rely on it most. We plan to expand validation to physical WSE-3 systems, extend the schedules to 2D collectives and AllReduce, and measure end-to-end gains in applications built on row reductions.

References

[1] P. Luczynski, L. Gianinazzi, P. Iff, L. Wilson, D. De Sensi, and T. Hoefler. Near-Optimal Wafer-Scale Reduce. In *HPDC '24*. https://doi.org/10.1145/3625549.3658693

[2] D. De Sensi, S. Pasqualoni, L. Piarulli, T. Bonato, S. Ba, M. Turisini, J. Domke, and T. Hoefler. Bine Trees: Enhancing Collective Operations by Optimizing Communication Locality. In *SC '25*. https://doi.org/10.1145/3712285.3759835

[3] C. He, Y. Huang, P. Mu, Z. Miao, J. Xue, L. Ma, F. Yang, and L. Mai. WaferLLM: Large Language Model Inference at Wafer Scale. In *OSDI '25*. https://doi.org/10.48550/arXiv.2502.04563