Short Hops, Big Gains: Beyond Asymptotic Costs in Collective Communication on Dragonfly+

Chiara Di Stefano, Daniele De Sensi
distefano.2061657@studenti.uniroma1.it, desensi@di.uniroma1.it
Sapienza University of Rome

Motivation

THEORY

Traditional cost models evaluate collective operations based on asymptotic complexity

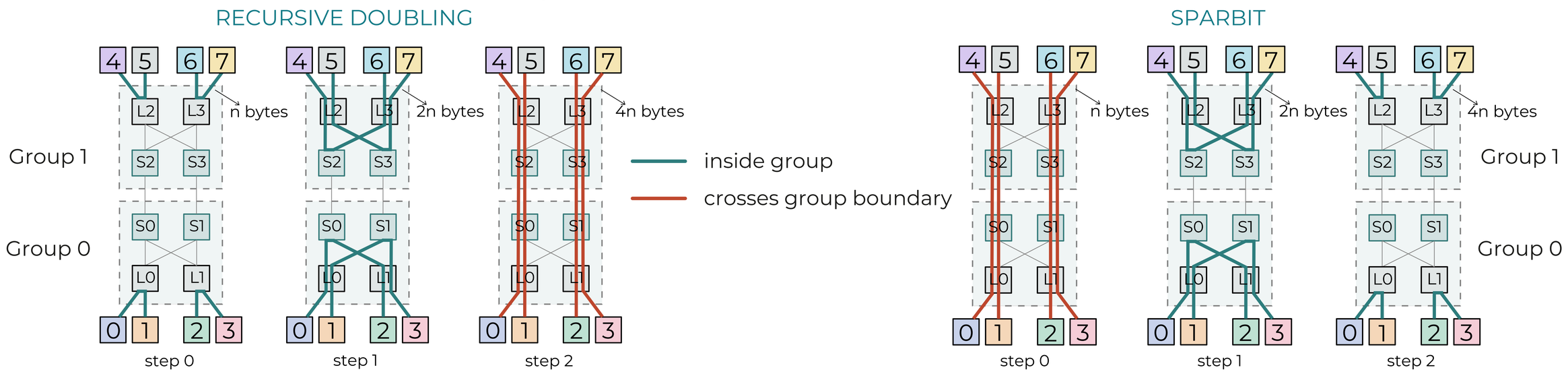
PRACTICE

Different algorithms with identical theoretical cost have different performance

SAME THEORETICAL COST

$$\log(p)\cdot\alpha + \frac{(p-1)}{p}\cdot n\cdot\beta$$

Bine, Bruck, Recursive Doubling, Sparbit

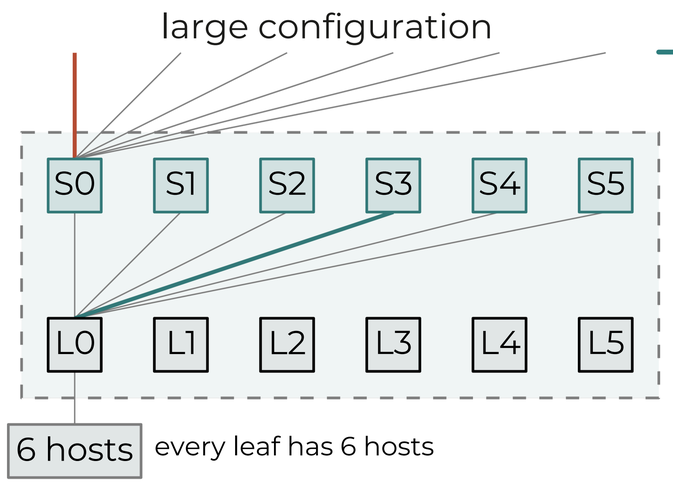


Recursive doubling and Sparbit AllGather exchange the same distances and the same message sizes, in opposite order
The first sends its ***largest message*** across a group boundary, Sparbit sends its ***smallest*** one

Methodology

SIMULATION:

ht-sim (extended for Dragonfly+), tracking port usage
400 Gbps bandwidth, 20 ns link latency, RoCE
AllGather, 1 MiB output vector, 1024 ranks
Minimal routing used, adaptive routing showed no significant difference

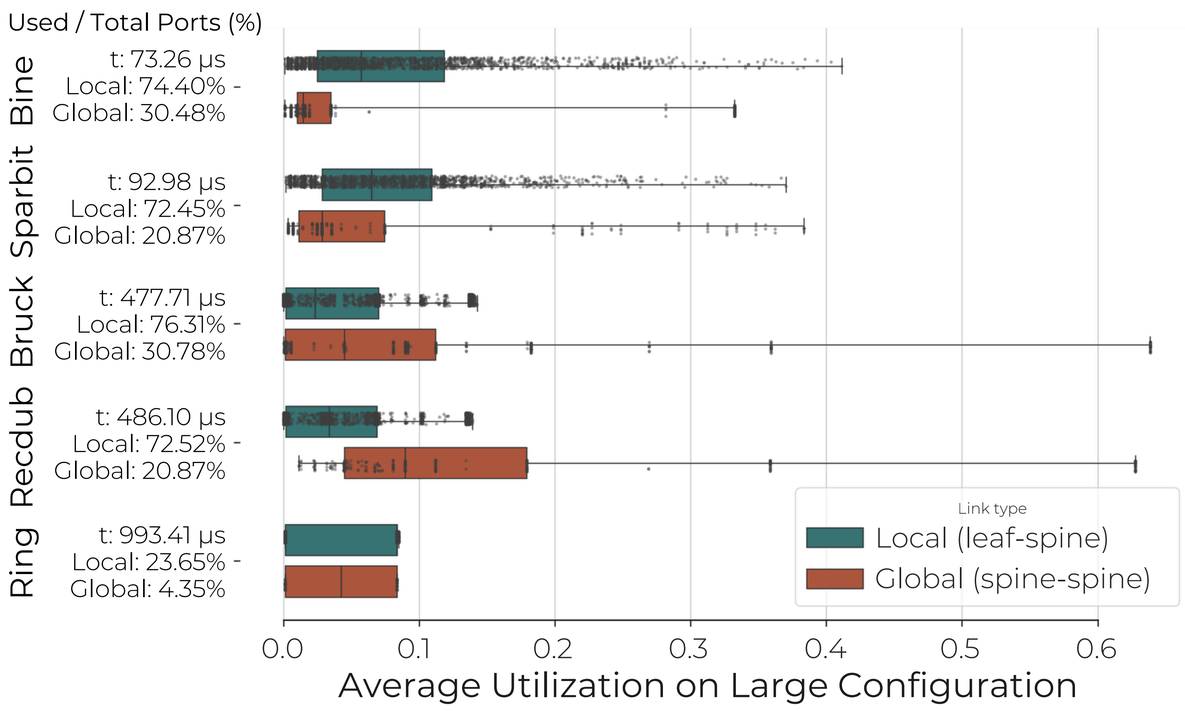


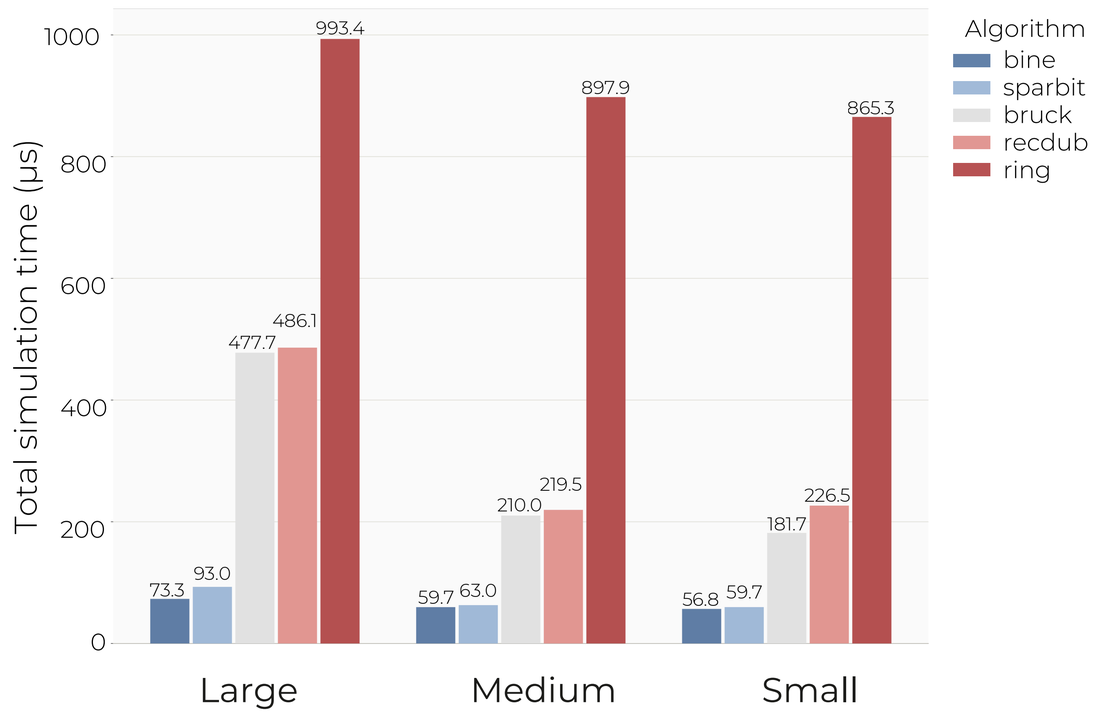
TOPOLOGY: Dragonfly+

	nodes	groups	radix	link between each pair of groups
Large	1332	37	12	1
Medium	1100	11	20	10
Small	1183	7	26	26

Results

Same cost model for the first four algorithms, 6.6× apart in *completion time* based on how traffic is distributed across local and global links





Conclusion and Future Work

Need to evaluate collective algorithms on the traffic they induce, not only on asymptotic cost

Extend our analysis to:

- other topologies
- different routing algorithm
- other collective operations

References

- Thakur, R., Rabenseifner, R., & Gropp, W. (2005). Optimization of collective communication operations in MPICH. IJHPCA, 19(1), 49-66.
- Loch, W. J., & Koslovski, G. P. (2021). Sparbit: A new logarithmic-cost and data locality-aware MPI Allgather algorithm.
- De Sensi, D., Pasqualoni, S., Piarulli, L., Bonato, T., Ba, S., Turisini, M., Domke, J., & Hoefler, T. (2025). Bine Trees: Enhancing Collective Operations by Optimizing Communication Locality. Proc. SC '25.

hlc-lab.github.io

Abstract