Completion-Driven, Staging-Free AlltoAllv with Local-NIC Balancing

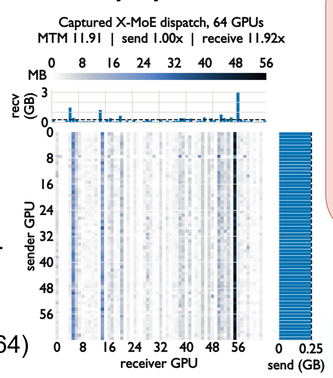
Emir Gencer[1], Abhinav Bhatele[1], Sajal Dash[2]
[1]Department of Computer Science, University of Maryland [2] Oak Ridge National Laboratory

Abstract

Mixture-of-Experts (MoE) training workloads use imbalanced AlltoAllv collectives at bandwidth-bound message sizes. On Frontier[1], one-sided AlltoAll can beat RCCL[6] and Cray-MPICH, but only with the right schedule and only as per-peer fetches shrink[2]. Other approaches[3,4] balance load across a node's NICs by explicitly copying data between GPUs, which costs GPU staging buffers and overlaps that copy with network traffic only coarsely. We introduce CURLEW, a completion-driven, GPU-staging-free RDMA AlltoAllv with dynamic local-NIC balancing. CURLEW achieves near-flat bandwidth across GPU counts and message sizes on balanced patterns, is 1.9x faster than previous work[3] at max-to-mean skew above 8, and reduces end-to-end training time of a 201B-parameter model by 22% at a capacity factor of 1.25.

AlltoAllv in MoE Training

- Two AlltoAllv per layer, per forward and backward pass:
  - Dispatch token activations to expert ranks, combine back.
- A learned router decides the traffic pattern every layer.
  - Skew is worst early in training.
- In Dispatch sends are balanced, in Combine receives are.
- Capacity factor caps the skew by dropping tokens.
  - Recent models do not drop tokens[5].
- Skew ranges from 1.2-11.9 Max-To-Mean (MTM) ratio.
- Message size config dependent.
  - 3 MiB per peer in XMoE[7] (EP=64)

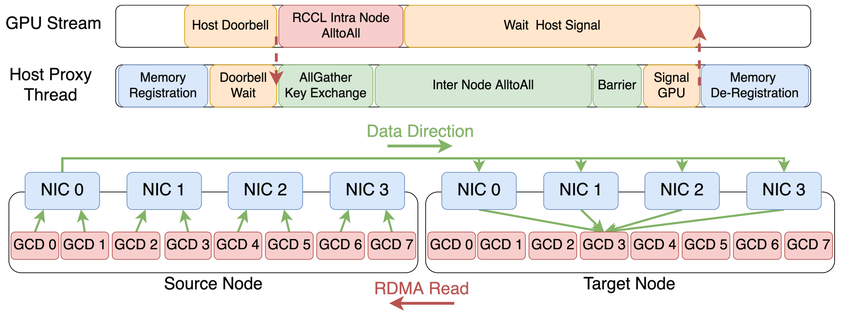


Previous Work

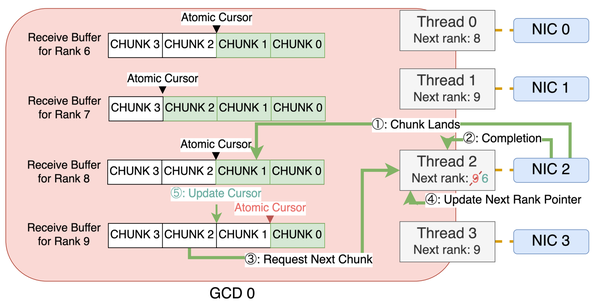
- MPI_Get based AlltoAll is **schedule dependent** and performs better as **per-peer message sizes get smaller** as total receive buffer stays the same[2].
  - Same bytes in smaller Gets achieve better bandwidth.
- Hierarchical approaches for balancing skew across local Network Interface Cards (NIC) [3,4].
  - GPU **staging buffers** sized O(message size).
  - Explicit software copy with **coarse grained overlap**.
- In a Frontier[1] node 8 GCDs share 4 NICs.
  - Harder to explicitly balance work across NICs compared to one NIC per GCD.
  - DMA requests to far GCDs are implicitly routed over xGMI.

CURLEW Design

- For dispatch, sends are balanced, receives are imbalanced.
  - For dispatch use RDMA read, for combine use RDMA write.
  - In both cases, stripe only local (imbalanced side) NICs.



1. Fill the pipeline so that there are 128 chunks in-flight, 32 per NIC.
2. As completions return, threads pull the next chunk from the peer their cursor points to.

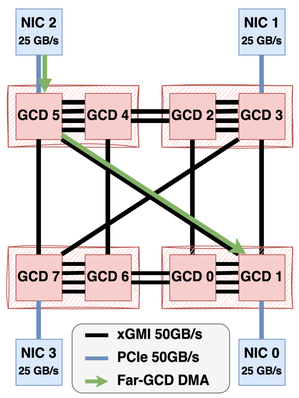


- Far GCDs are reached over xGMI by the hardware.
- No staging buffer, no software copy.
- The rail carrying a chunk is chosen at completion time, dynamically load balancing NICs.
- We use 16KiB chunks.

Experimental Setup

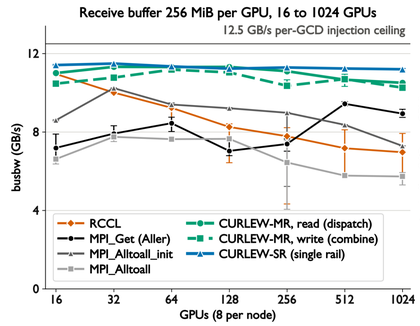
- Experiments conducted on Frontier
- Benchmarking balanced AlltoAll:
  - CURLEW, RCCL, MPI{AlltoAll, Alltoall_init, Get}
  - Receive buffer at 256MiB, on 16-1024 GCDs.
- End-to-end training of Xmoe-Medium and Large[7]
  - 55B and 201B respectively.
  - Both with EP = 64, capacity factor = 1.25
- Benchmarking high-skew AlltoAllv:
  - CURLEW, RCCL, SABRE[3]
  - Schedules captured from training a 10B model at EP=64, no token dropping.
  - Captured first 500 iterations, every 25 iteration.

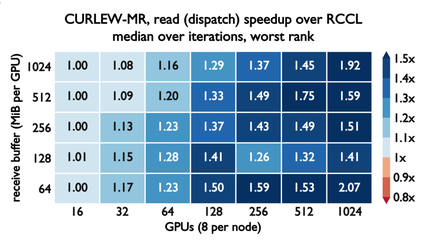
A Frontier Node



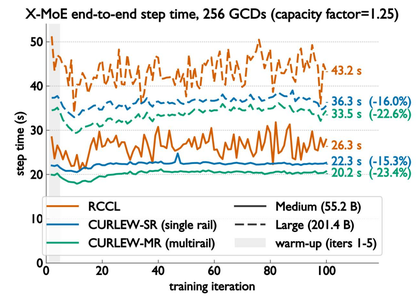
This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Advanced Scientific Computing Research programs in the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

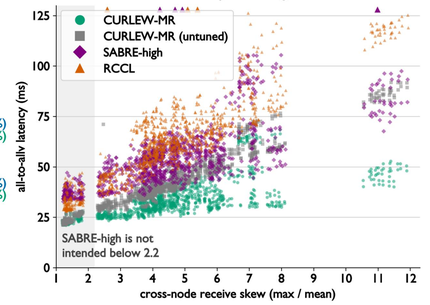
Results





- Both multi rail and single rail CURLEW achieve 90% of peak bandwidth and show near-flat scaling on balanced AlltoAll.
- MPI_Get will converge to CURLEW once per-peer message size becomes less than the chunk size.





- **Left:** CURLEW reduces end to end training time for XMoE-Medium and Large by 23% and 22% respectively.
- **Right:** CURLEW is 1.9x faster than SABRE and 2.5x faster than RCCL on skewed workloads.

Conclusion and Future Work

- This work identifies the gaps in existing work in large-message AlltoAll and imbalanced AlltoAllv scenarios.
- We introduce CURLEW, a completion-driven, staging buffer free AlltoAllv with dynamic local NIC balancing.
- We show CURLEW can beat vendor libraries and previous work on balanced AlltoAll, highly skewed AlltoAllv, and end-to-end training performance.

References

[1] O. R. N. L. (ORNL), "Frontier (hpe cray ex) exascale supercomputer at the oak ridge leadership computing facility," 05 2022. [Online]. Available: https://www.osti.gov/biblio/3023606
[2] J. B. White, "Large-message all-to-all communication at frontier scale," in Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, ser. SC Workshops '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 461–467. [Online]. Available: https://doi.org/10.1145/3731599.3767389
[3] C. Wei and A. Bhatele, "Skew-aware adaptive all-to-ally algorithms for dynamic deep learning workloads," in Proceedings of the 40th ACM International Conference on Supercomputing, ser. ICS '26. New York, NY, USA: Association for Computing Machinery, 2026, p. 450–462. [Online]. Available: https://doi.org/10.1145/3797905.3800541
[4] C.-C. Chen, J. Yao, L. Xu, H. Subramoni, and D. K. Panda, "Unified designs of multi-rail-aware mpi allreduce and alltoall operations across diverse gpu and interconnect systems," in 2025 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2025, pp. 938–949.
[5] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan et al., "Deepseek-v3 technical report," arXiv preprint arXiv:2412.19437, 2024.
[6] "Rccl," 2025. [Online]. Available: https://github.com/ROCm/rccl
[7] Y. Yuan, A. Gupta, J. Li, S. Dash, F. Wang, and M. Zhang, "X-moe: Enabling scalable training for emerging mixture-of- experts architectures on hpc platforms," 2025. [Online]. Available: https://arxiv.org/abs/2508.13337