HiBD
High-Performance
Big Data

MVAPICH
MPI, PGAS and Hybrid MPI+PGAS Library

HiDL
High-Performance
Deep Learning

Benefits of Streaming Aggregation with SHARPv2 in MVAPICH2

9th MVAPICH User Group (MUG) Conference '21

Bharath Ramesh

The Ohio State University
ramesh.113@osu.edu

Follow us on

https://twitter.com/mvapich

Presentation Outline

- Introduction/Background
- Performance Evaluation for MPI_Allreduce and MPI_Reduce
- Summary

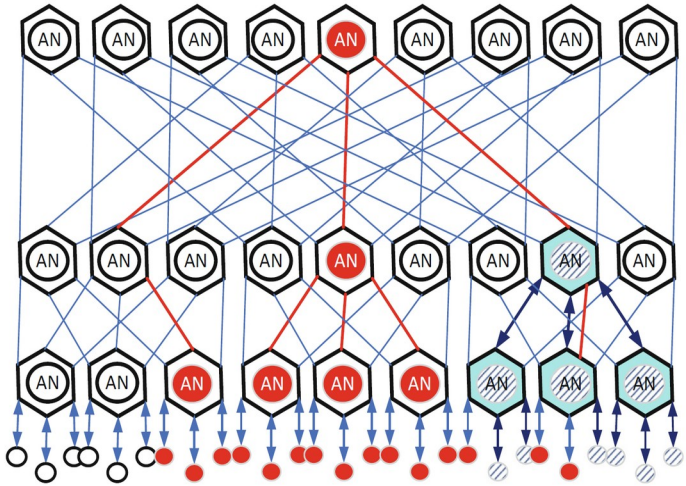
Considerations for Accelerating HPC Applications

- MPI collectives using aggregation (Allreduce and Reduce) significant to application run time
- Ideal set of goals
  - Overlap computation and communication
  - Maximally utilize CPU resources
  - Scale-out and Scale-up efficiency
  - Ideally no changes to application code for performance
  - Utilize high levels of parallelism
- Co-design software and hardware elements for best results

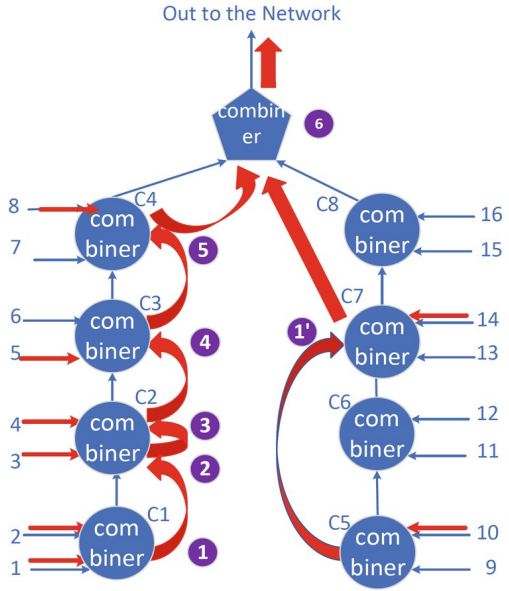
What is SHARP?

- Scalable Hierarchical Aggregation and Reduction Protocol
- Advantages
  - Progress, offload computation and communication
  - Focuses on low latency for small messages, maximal bandwidth utilization for large messages
  - Hierarchical Aggregation in a logical tree (LLT) providing low latency for small messages
  - Streaming aggregation with pipelined ring-based algorithms for large messages
- In-Network computing
- Focus on creating groups, which can simultaneously execute operations

SHARP Reduction trees and Streaming Aggregation



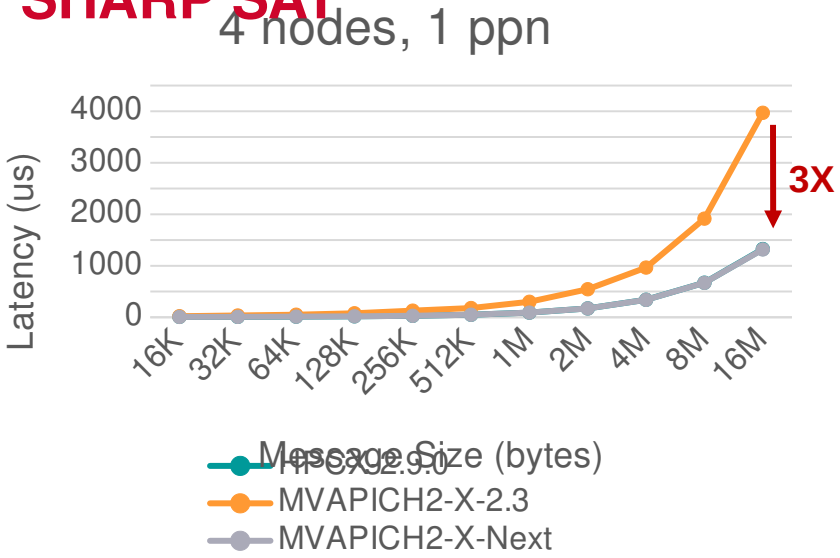
Aggregation Tree

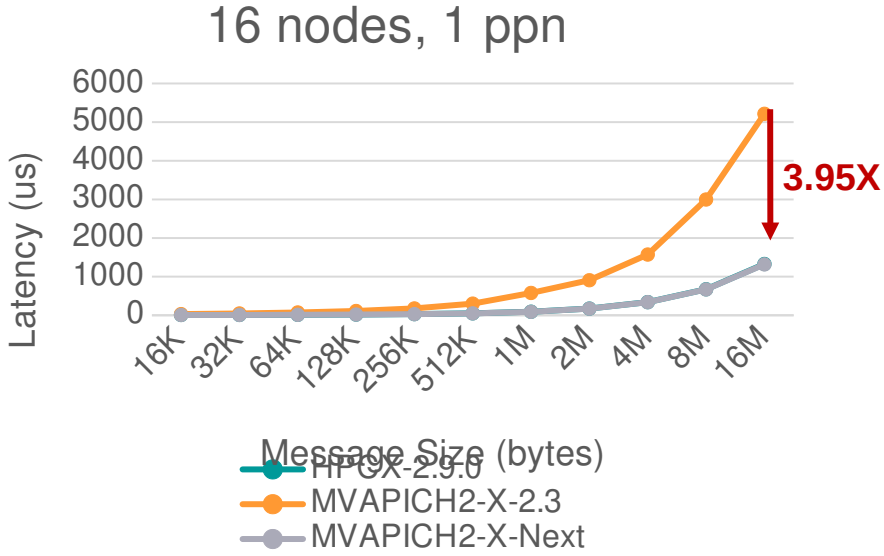


Switch-level reduction (radix 16)

Images taken from "Graham, Richard et al. Scalable Hierarchical Aggregation and Reduction Protocol (SHARP)™ Streaming-Aggregation Hardware Design and Evaluation. DOI : 10.1007/978-3-030-50743-5_3

Optimized MPI_Allreduce Performance with MVAPICH2-X + SHARP SAT

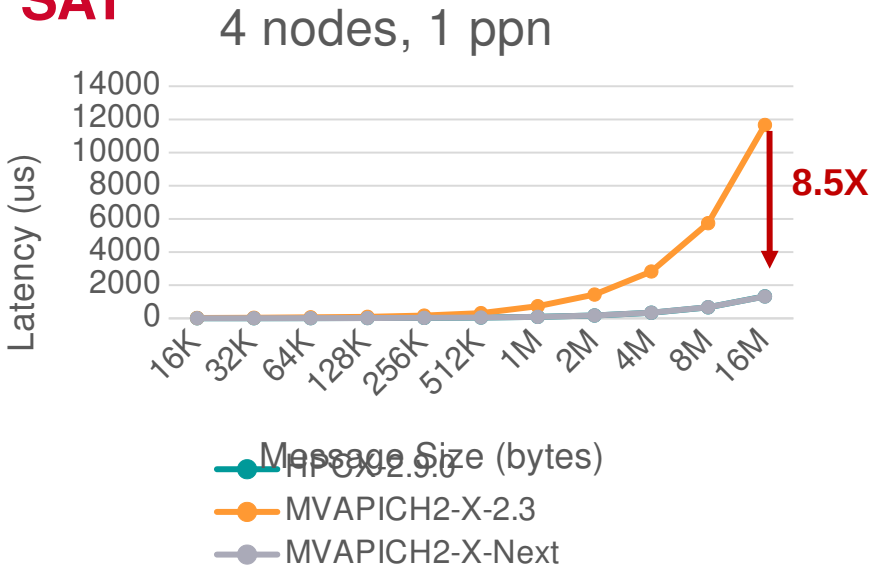


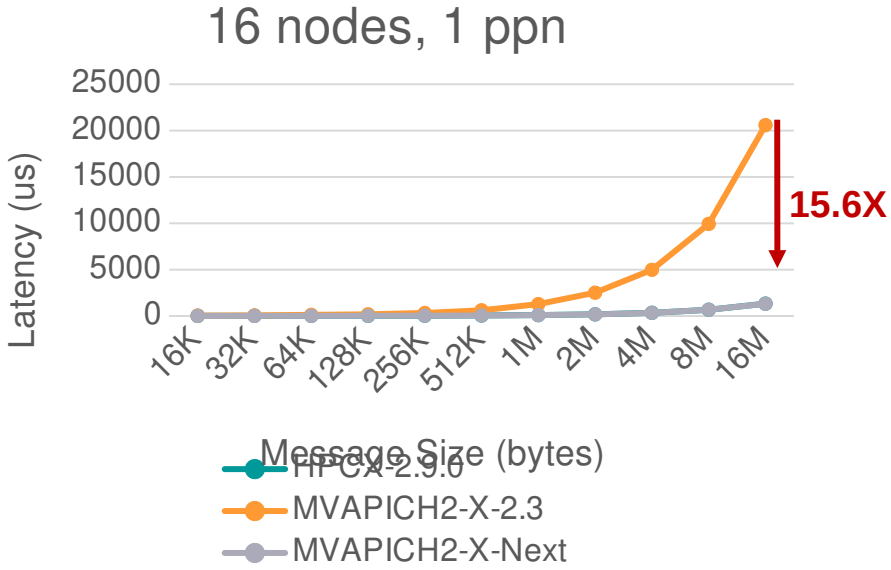


- SHARP provides flat scaling, even for large messages
- Up to 3.95X benefits over MVAPICH2-X-2.3 using SAT + optimized designs

Platform: Intel(R) Xeon(R) Gold 6138 nodes equipped with a dual-socket CPU and InfiniBand HDR-200 Interconnect

Optimized MPI_Reduce Performance with MVAPICH2-X + SHARP SAT





- Comparing max latency for MPI_Reduce as the root is the bottleneck
- Up to 15.6X benefits over MVAPICH2-X-2.3 using SAT + optimized designs

Platform: Intel(R) Xeon(R) Gold 6138 nodes equipped with a dual-socket CPU and InfiniBand HDR-200 Interconnect

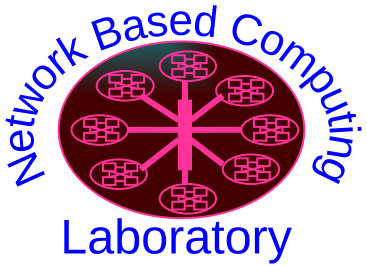
Conclusion and Future work

- Conclusion
  - Pure software-based schemes limiting as message size and scale increases
  - SHARP highly effective with good scalability and low latency
    - Flat scaling up to a fixed node count, even with streaming aggregation
    - Close to point-to-point latency

- Future work
  - Comprehensive evaluation with benchmarks and applications at large scales
  - Scaling studies with larger number of processes per node
  - Optimize non-blocking collectives with streaming aggregation

Thank You!

ramesh.113@osu.edu



Follow us on

https://twitter.com/mvapich

Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

The High-Performance MPI/PGAS Project

http://mvapich.cse.ohio-state.edu/

The High-Performance Big Data Project

http://hibd.cse.ohio-state.edu/

The High-Performance Deep Learning Project

http://hidl.cse.ohio-state.edu/