SmartNICs SUMMIT

April 26-28, 2022
DoubleTree by Hilton San Jose
SmartNICsSummit.com

Accelerating HPC Applications with SmartNICs

Donglai Dai

Chief Engineer

contactus@x-scalesolutions.com

X-ScaleSolutions

Outline

- Motivation
- Basic Idea for MVAPICH2-DPU Library Design
- Main Features of MVAPICH2-DPU Library
- Performance Benefits for Benchmarks and Applications
- Conclusion

Requirements for Next-Generation Communication Libraries

- SmartNICs have the potential to take over a wide range of overhead tasks in a variety of applications from the host CPUs in systems
- Message Passing Interface (MPI) libraries are widely used for parallel and distributed HPC and AI applications in HPC/data centers and clouds
- Requirements for a high-performance and scalable MPI library:
    - Low latency communication
    - High bandwidth communication
    - Minimum contention for host CPU resources to progress non-blocking collectives
    - High overlap of computation with communication
- CPU based non-blocking communication progress can lead to sub-par performance as the main application has less CPU resources for useful application-level computation

Can MPI Functions be Offloaded?

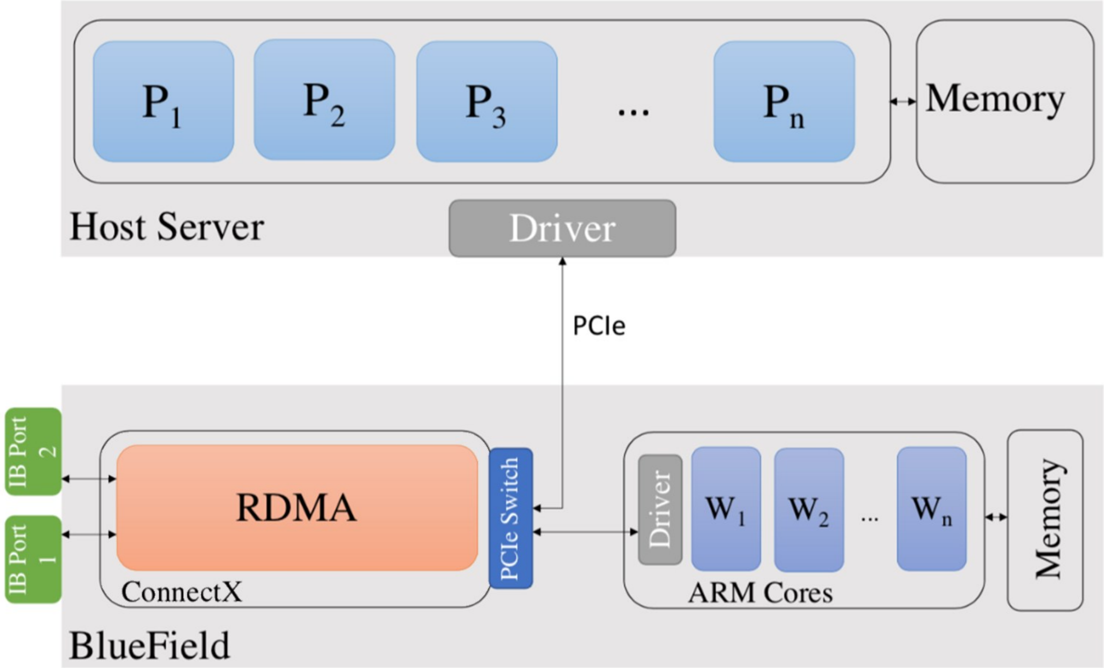
- The area of network offloading of MPI primitives is still nascent
- State-of-the-art BlueField DPUs bring more compute power into the network
- Exploit additional compute capabilities of modern BlueField DPUs into existing MPI middleware to extract
  - Peak pure communication performance
  - Overlap of communication and computation

Outline

- Motivation
- Basic Idea for MVAPICH2-DPU Library Design
- Main Features of MVAPICH2-DPU Library
- Performance Benefits for Benchmarks and Applications
- Conclusion

Overview of BlueField-2 DPU

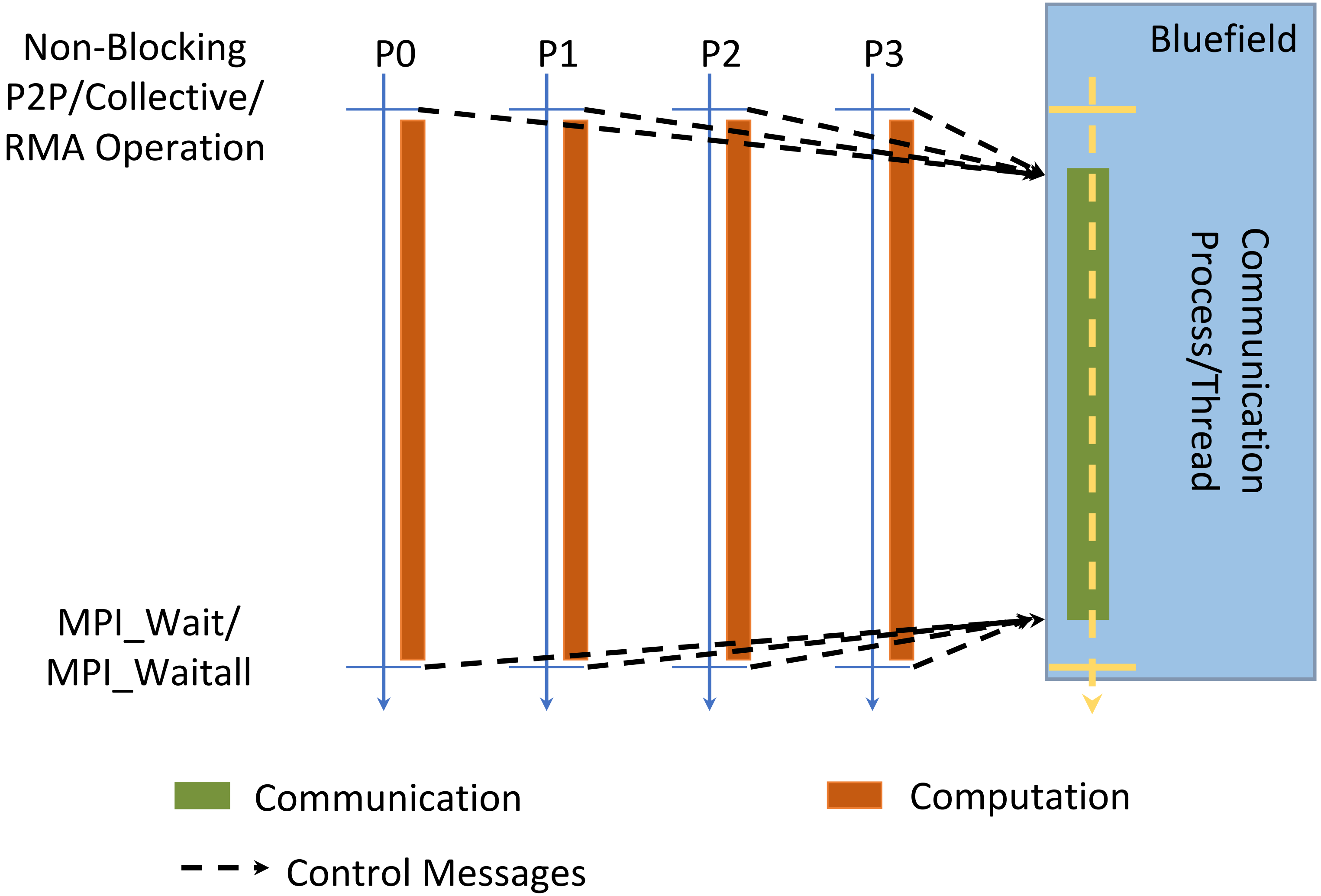
- ConnectX-6 network adapter with 200Gbps InfiniBand
- System-on-chip containing eight 64-bit ARMv8 A72 cores with 2.7 GHz each
- 16GB of memory for the ARM cores



MVAPICH2-DPU MPI library is designed to take advantage of DPUs and accelerate scientific applications

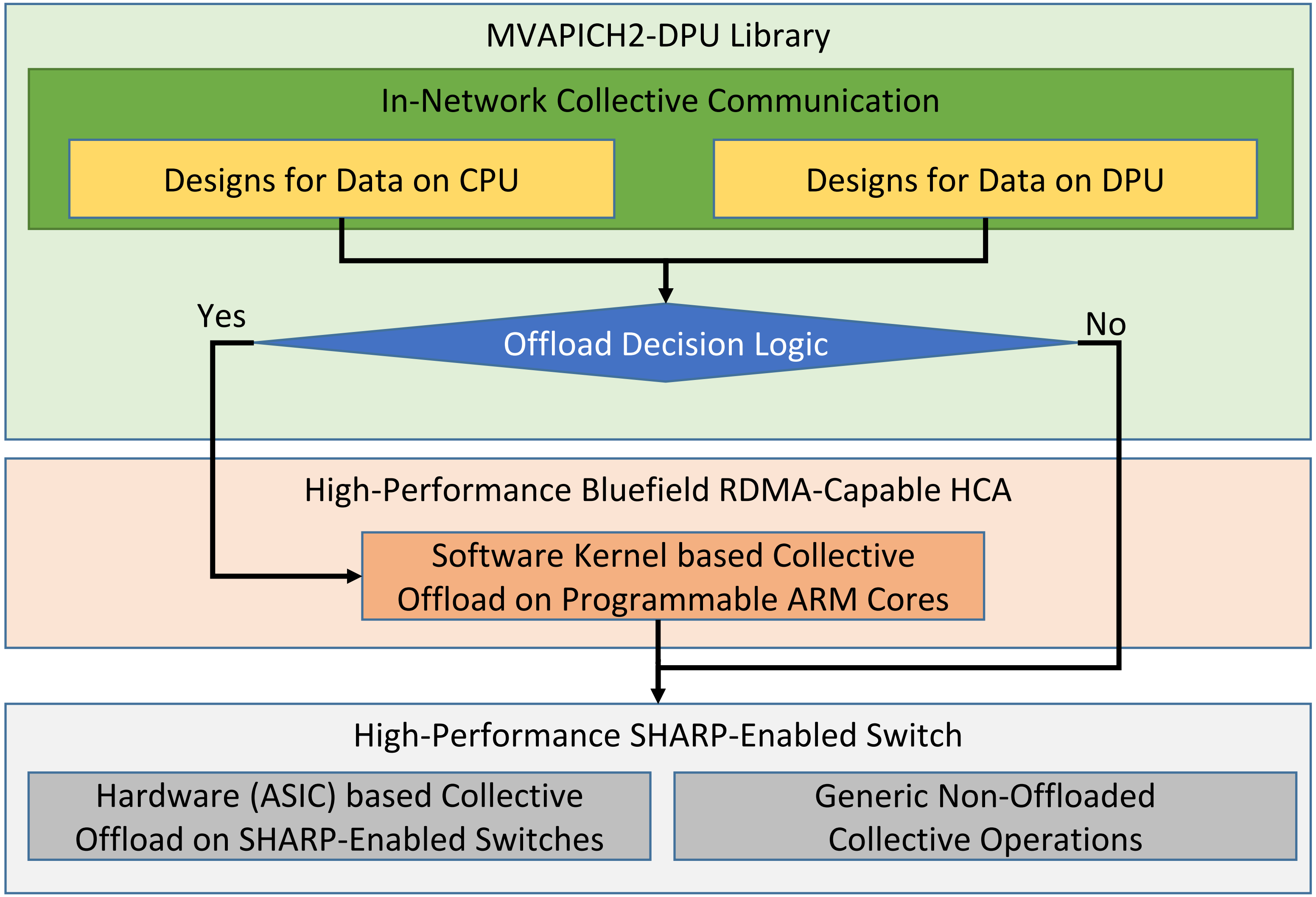
Basic Idea for MPI offloading to DPU

- Use of generic and optimized asynchronous progress threads on ARM cores for
  - Point-to-point
  - Collectives
  - RMA operations



High Level Design for MPI offloading to DPU

- Better support for critical collective communication operations
  - Enable offloading to the Bluefield ARM SoC
  - Performance enhancing algorithm selection based on the communication characteristics of application



MVAPICH2-DPU Library
In-Network Collective Communication
Designs for Data on CPU
Designs for Data on DPU
Yes
Offload Decision Logic
No
High-Performance Bluefield RDMA-Capable HCA
Software Kernel based Collective Offload on Programmable ARM Cores
High-Performance SHARP-Enabled Switch
Hardware (ASIC) based Collective Offload on SHARP-Enabled Switches
Generic Non-Offloaded Collective Operations

Outline

- Motivation
- Basic Idea for MVAPICH2-DPU Library Design
- Main Features of MVAPICH2-DPU Library
- Performance Benefits for Benchmarks and Applications
- Conclusion

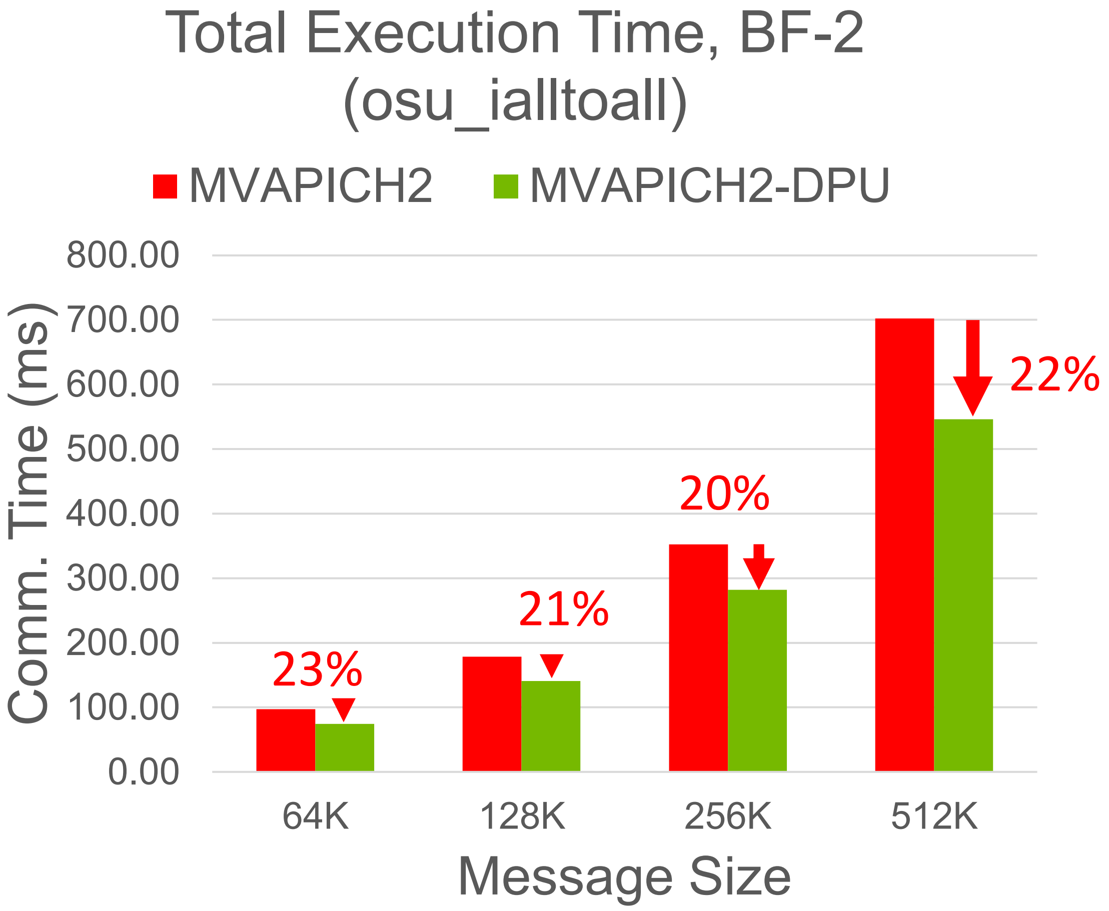
MVAPICH2-DPU Library 2022.02 Release

- Implemented by X-ScaleSolutions
- Based on MVAPICH2 2.3.6, compliant to MPI 3.1 standard
- Supports all features available with the MVAPICH2 2.3.6 release (http://mvapich.cse.ohio-state.edu)
- Novel framework to offload non-blocking collectives to DPU
- Offloads non-blocking collectives (MPI_Ialltoall, MPI_Iallgather, MPI_Ibcast, etc) to DPU
- Up to 100% overlap of computation with non-blocking collective
- Accelerates scientific applications using non-blocking collectives

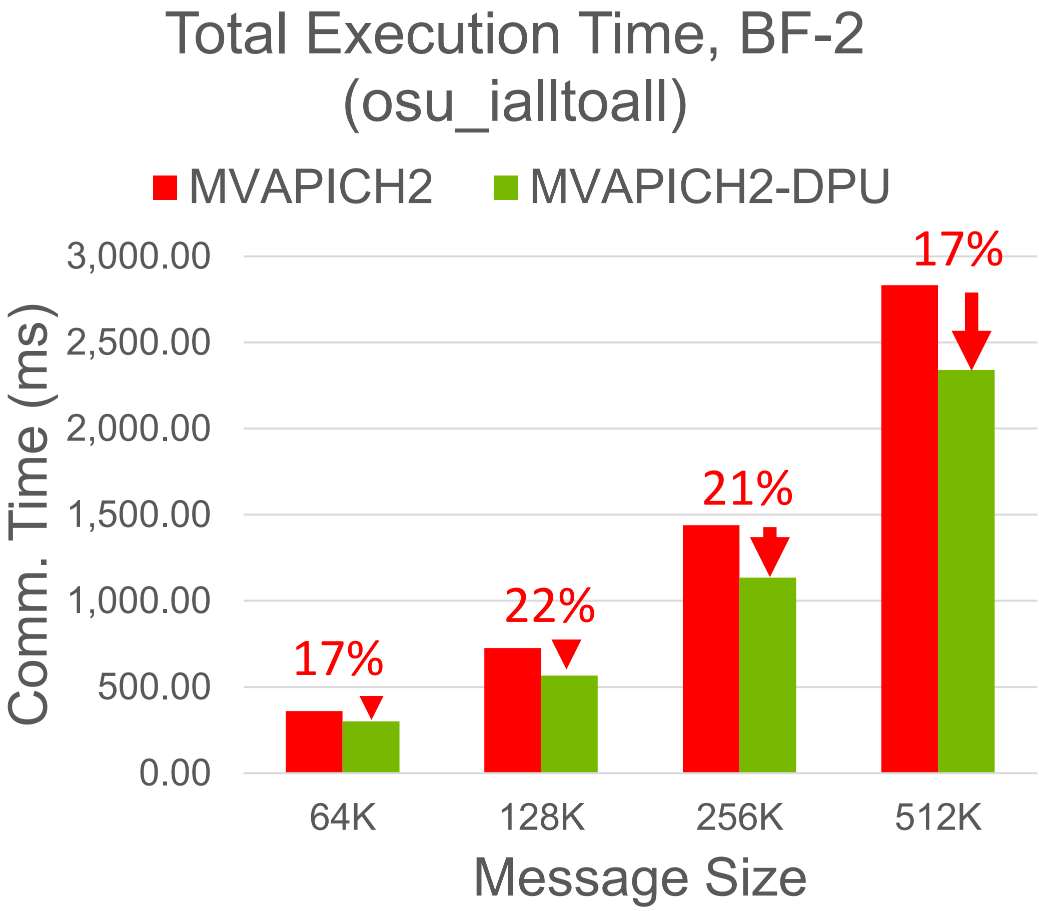
Outline

- Motivation
- Basic Idea for MVAPICH2-DPU Library Design
- Main Features of MVAPICH2-DPU Library
- Performance Benefits for Benchmarks and Applications
- Conclusion

Total Execution Time with osu_Ialltoall (32 nodes)

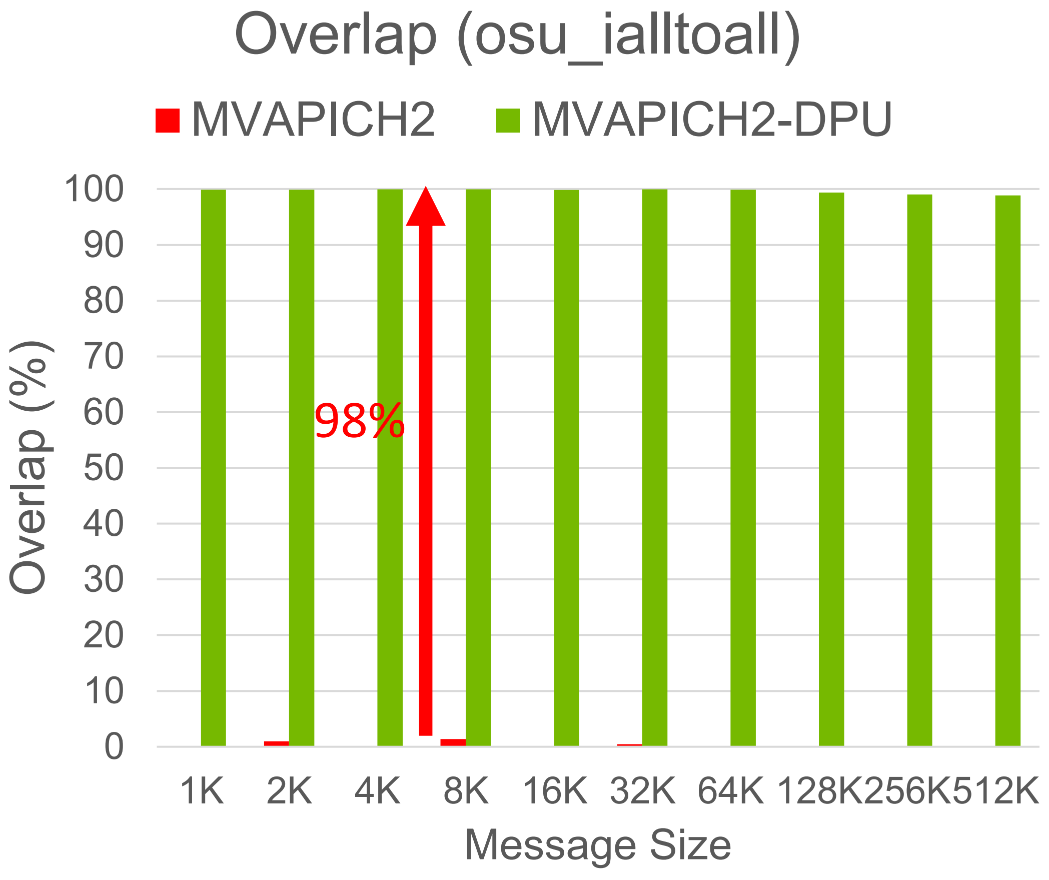


32 Nodes, 16 PPN

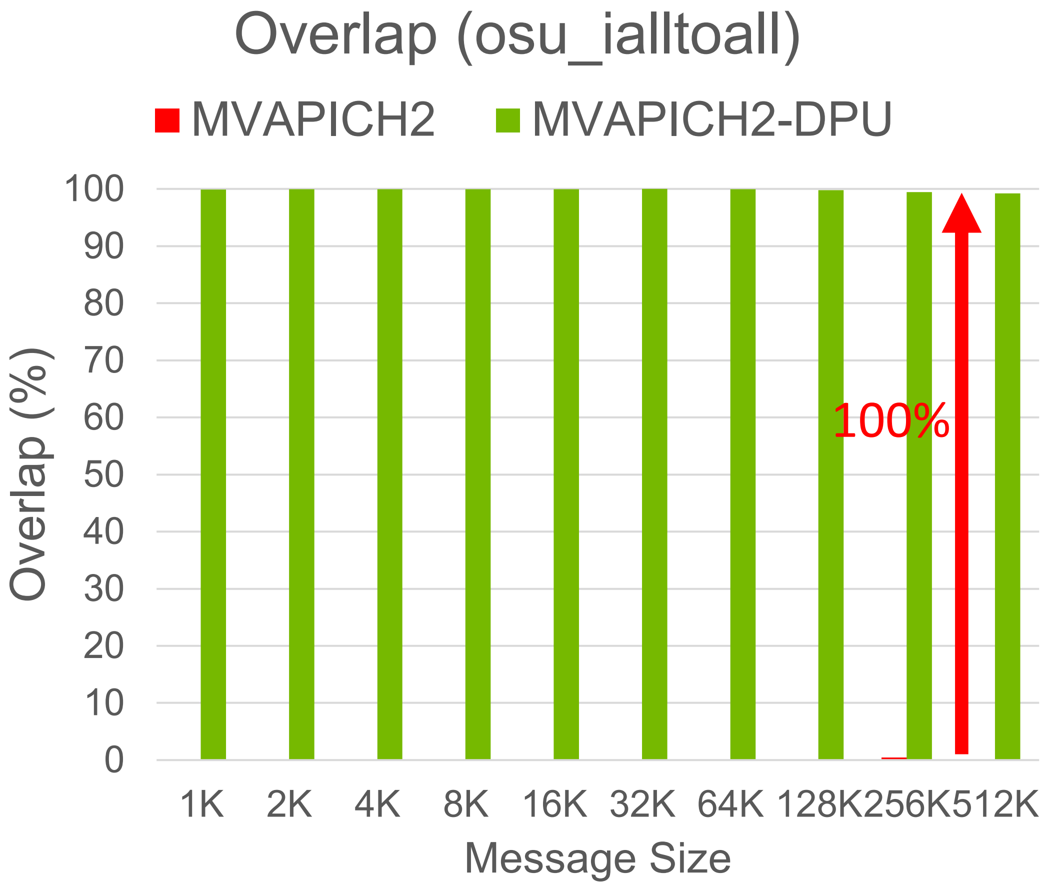


32 Nodes, 32 PPN

Overlap Between Computation & Communication with osu_Ialltoall (32 nodes)



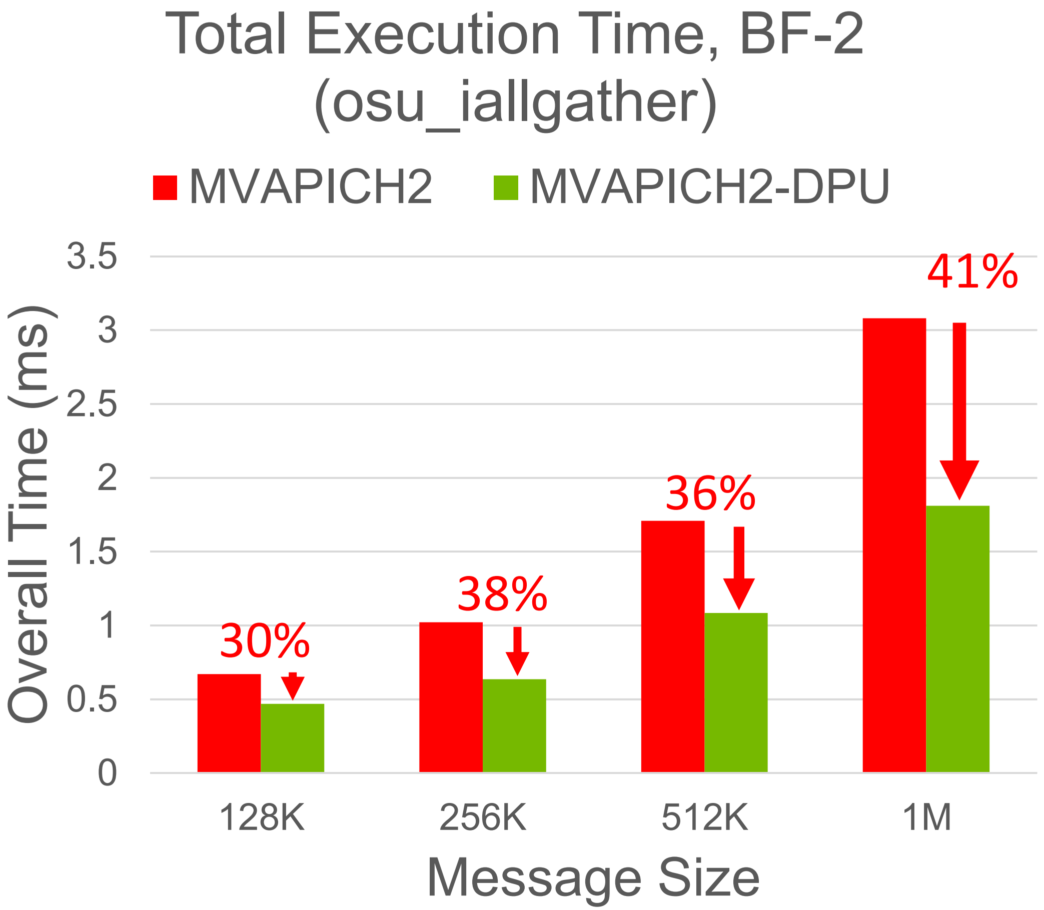
32 Nodes, 16 PPN



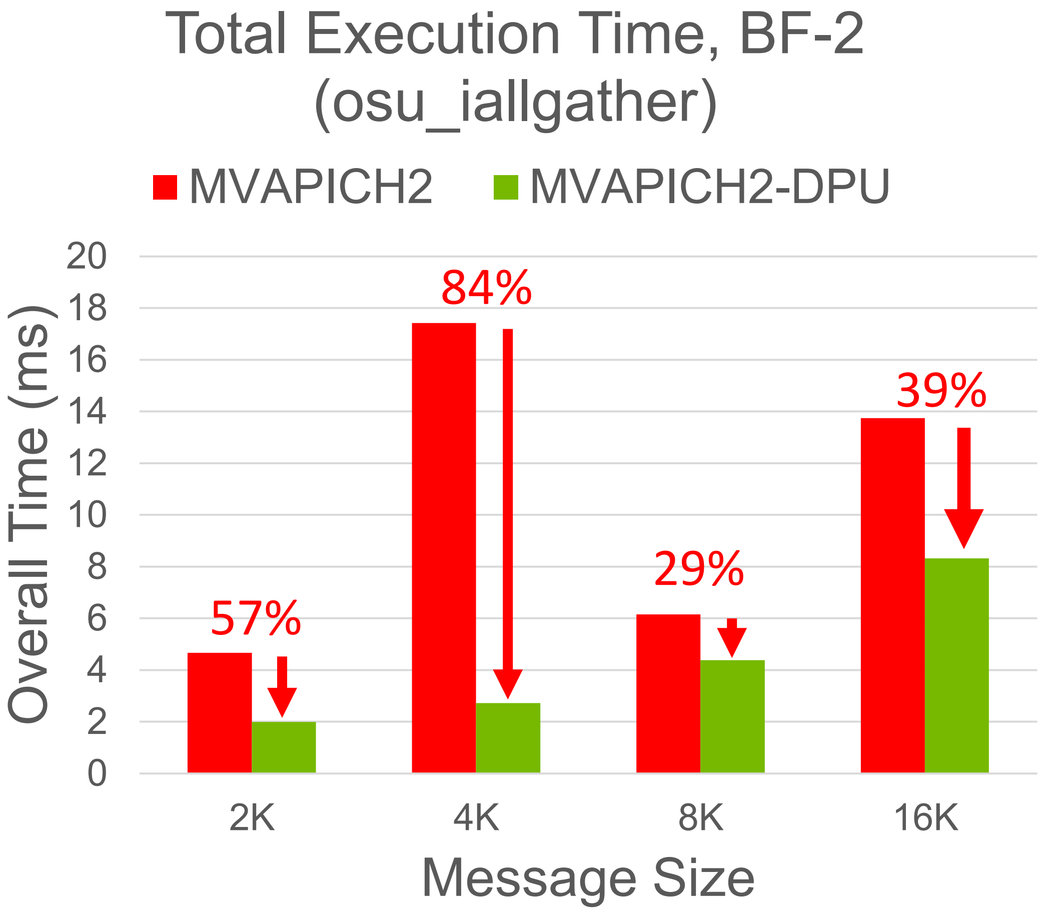
32 Nodes, 32 PPN

Delivers peak overlap

Total Execution Time with osu_Iallgather (16 nodes)

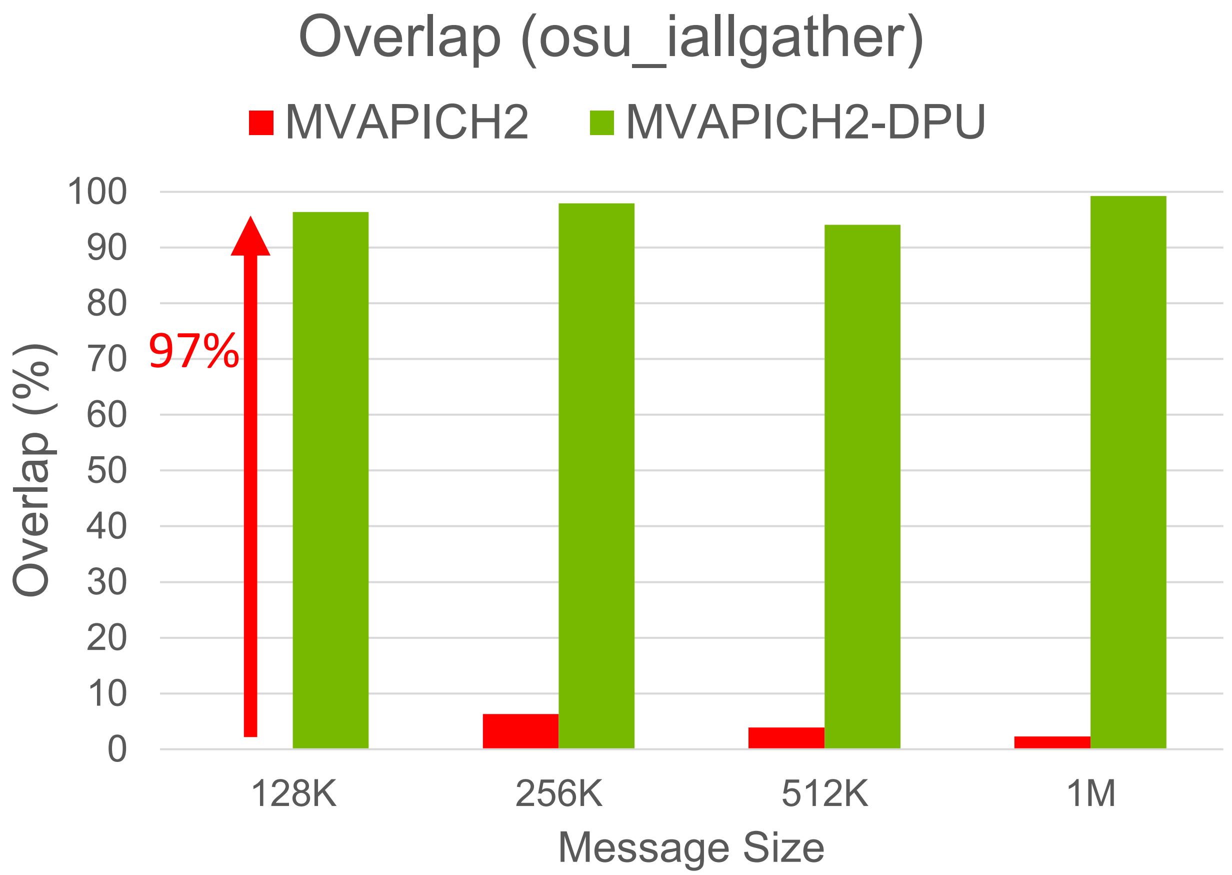


16 Nodes, 1 PPN



16 Nodes, 32 PPN

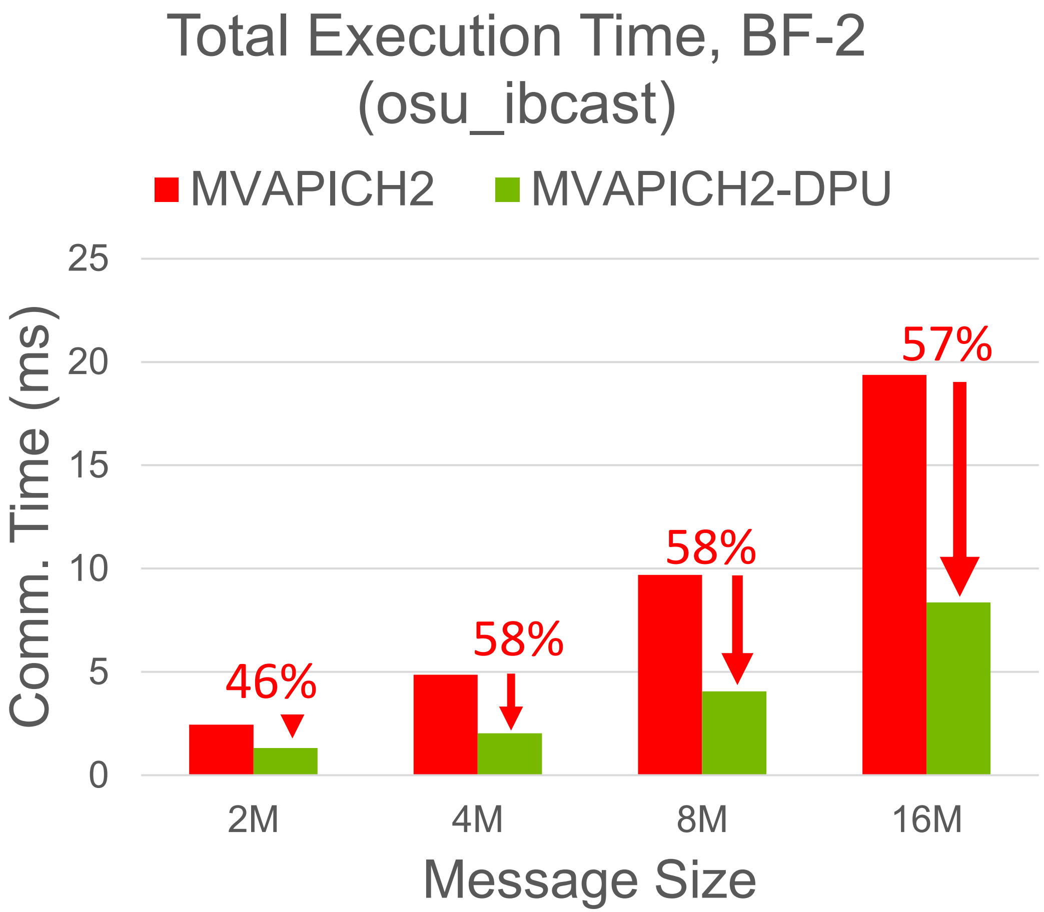
Overlap Between Computation & Communication with osu_Iallgather (16 nodes)



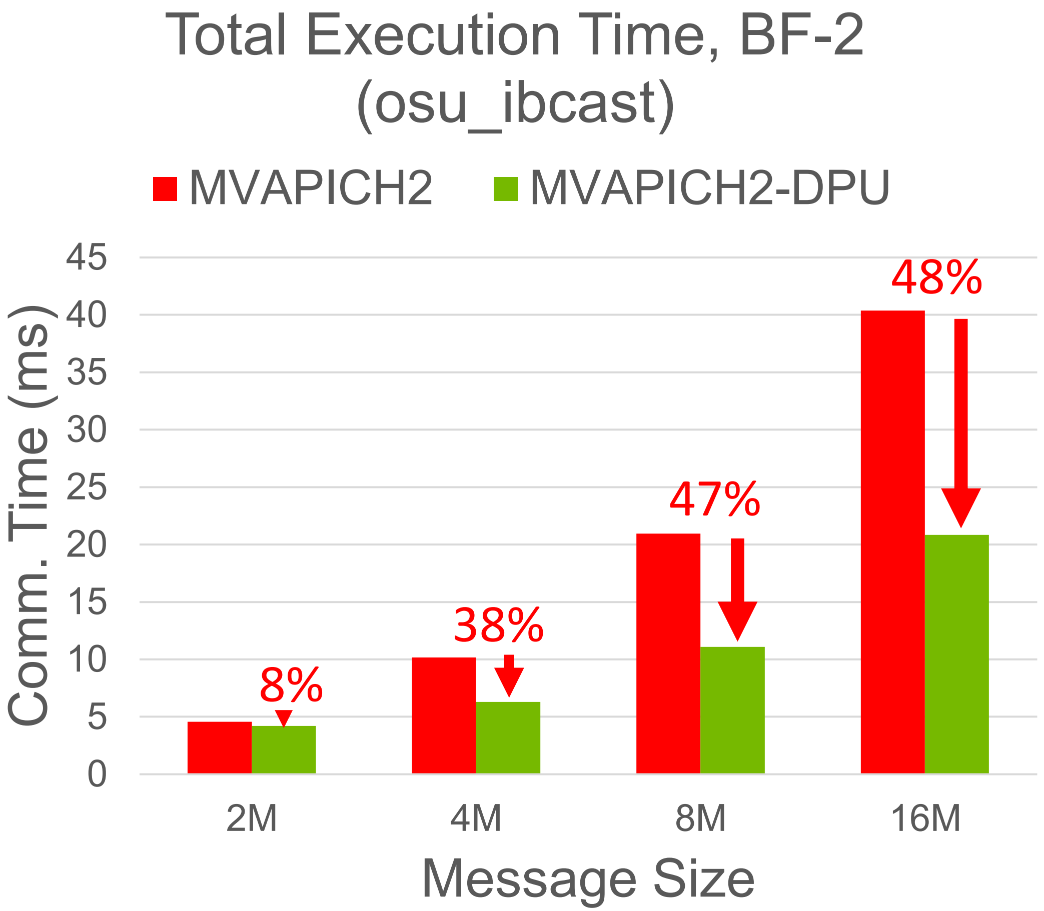
16 Nodes, 1 PPN

Delivers peak overlap

Total Execution Time with osu_Ibcast (32 nodes)

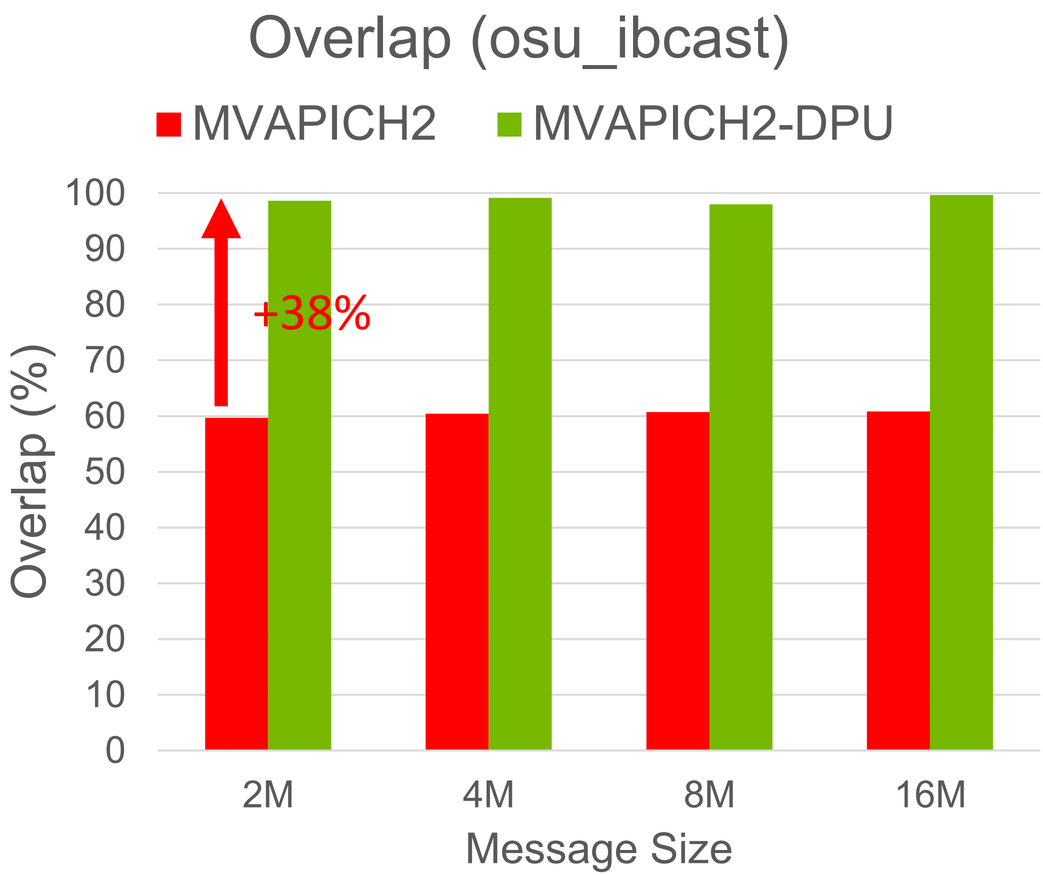


32 Nodes, 1 PPN

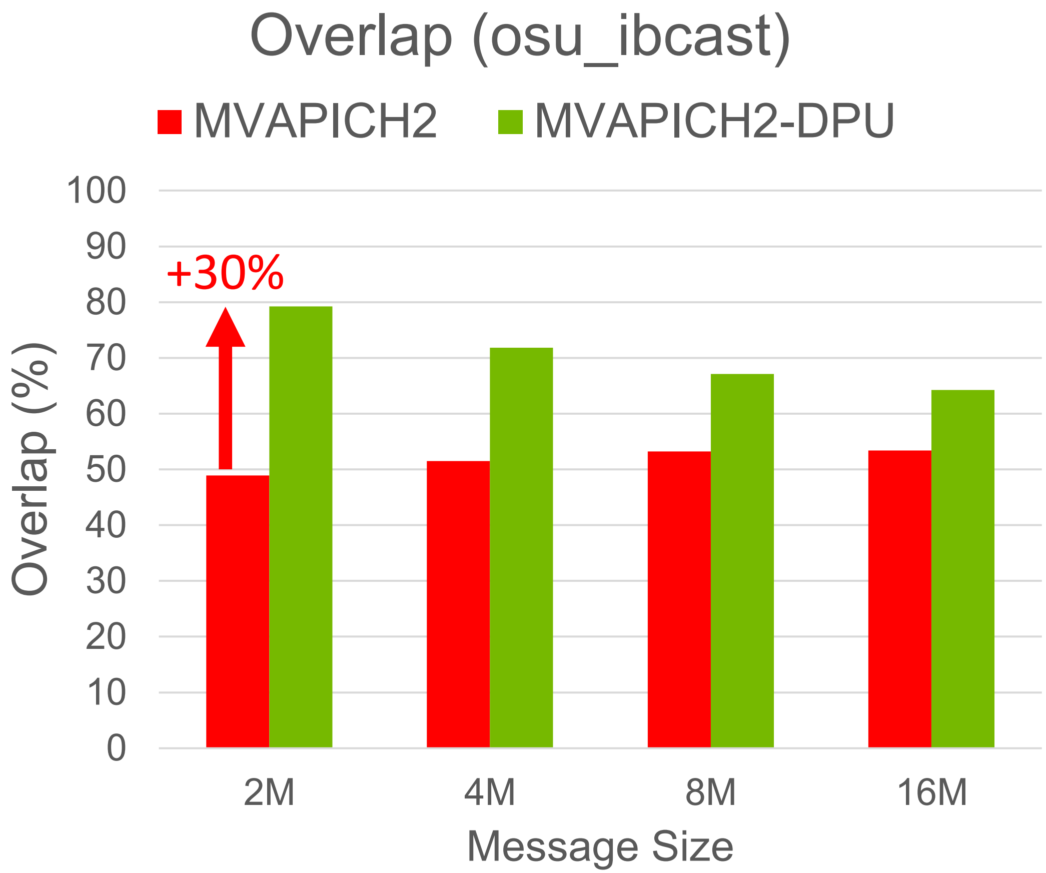


32 Nodes, 16 PPN

Overlap Between Computation & Communication with osu_Ibcast (32 nodes)



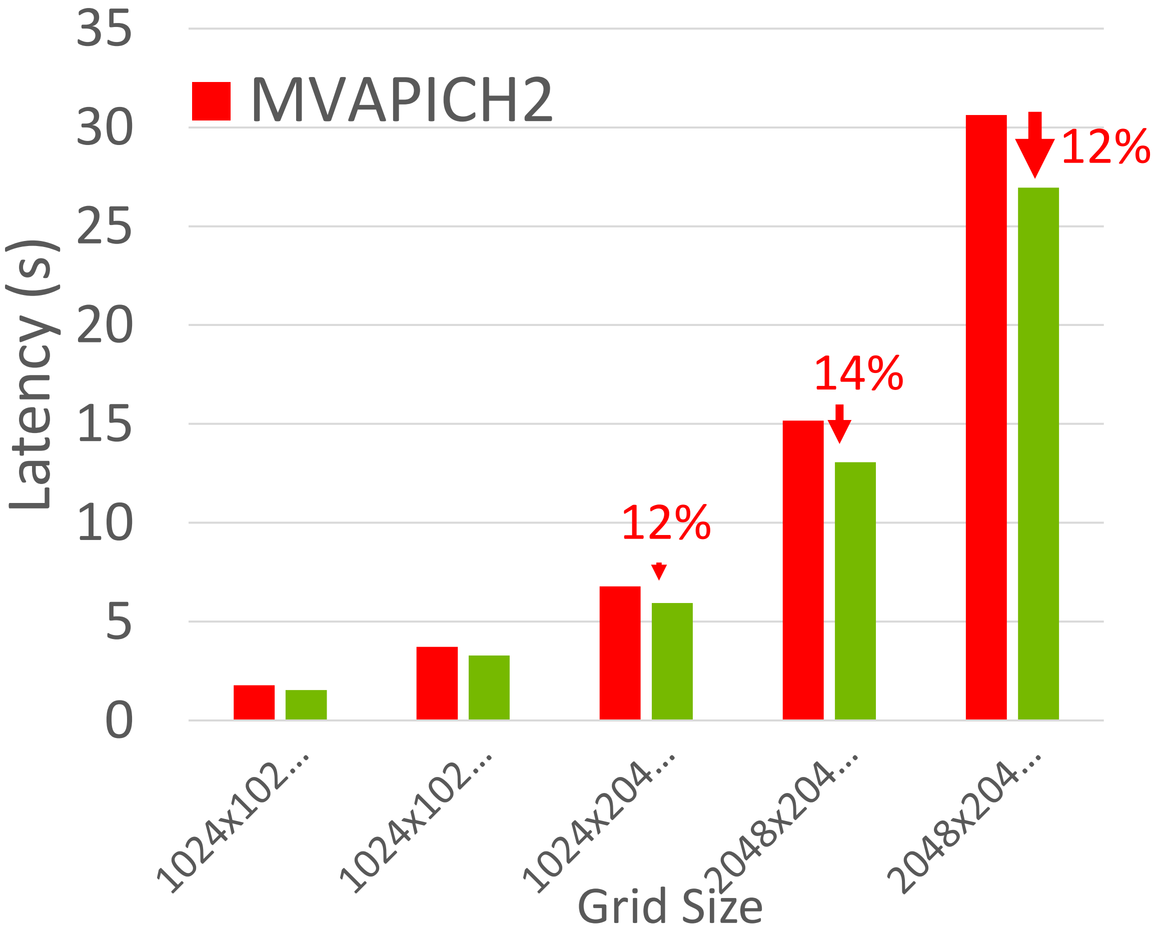
32 Nodes, 1 PPN



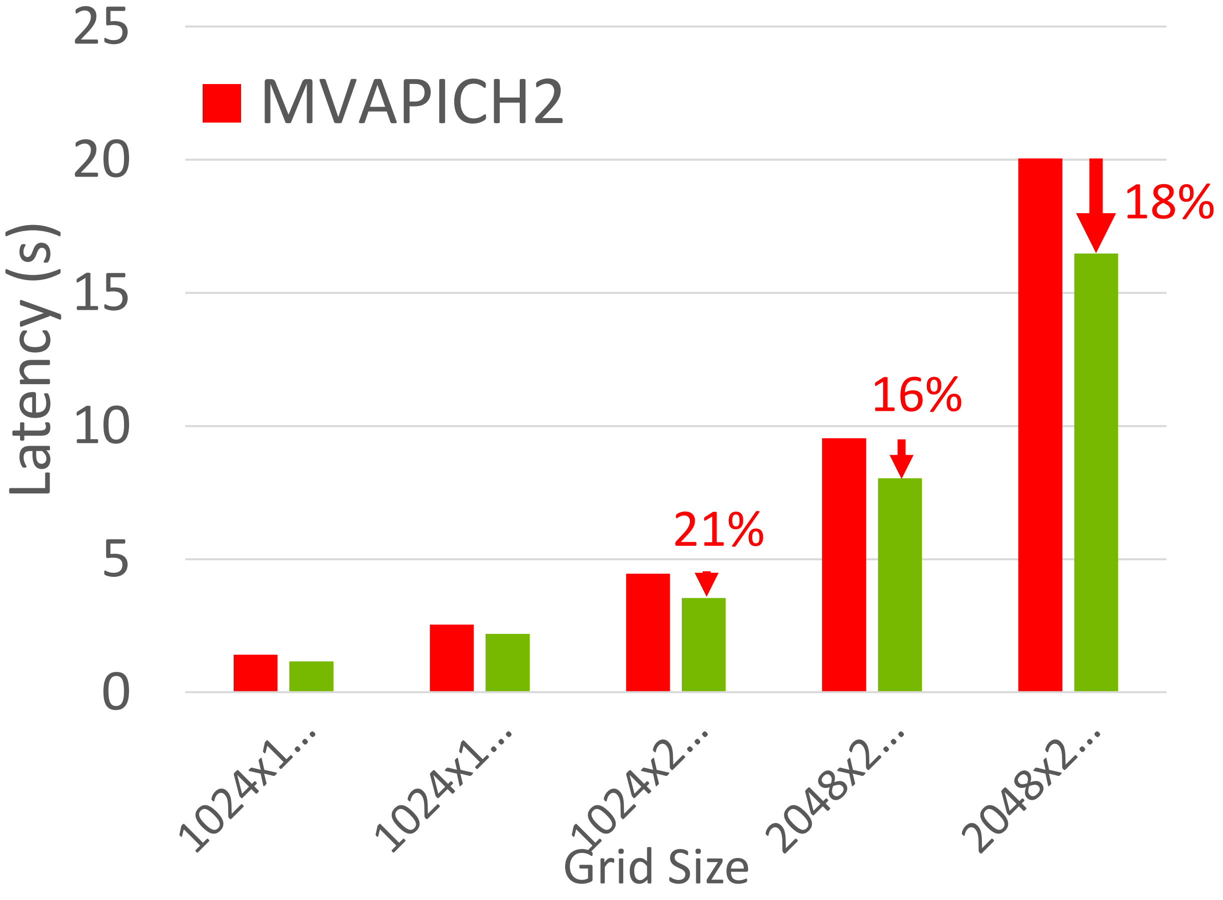
32 Nodes, 16 PPN

Delivers peak overlap

P3DFFT Application Execution Time (32 nodes)



32 Nodes, 16 PPN



32 Nodes, 32 PPN

Benefits in application-level execution time

Outline

- Motivation
- Basic Idea for MVAPICH2-DPU Library Design
- Main Features of MVAPICH2-DPU Library
- Performance Benefits for Benchmarks and Applications
- Conclusion

Conclusion

- Efficient MVAPICH2-DPU MPI library utilizes the BlueField DPU to progress MPI non-blocking collective operations
- Provides up to 100% overlap of communication and computation for non-blocking Alltoall, Allgather, Bcast, etc
- Reduces the total execution time of P3DFFT application up to 21% on 1,024 processes
- Work in progress for MVAPICH2-DPU library to efficiently offload more types of non-blocking collective operations to DPUs

Exhibition and Live Demo

- If you are interested in knowing more details, please come and visit our exhibit booth #8 next door
- Live demo on MVAPICH2-DPU library at our booth
  - 6-7 pm, today
  - 1-2 pm, tomorrow

Thank You!

Donglai Dai

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/