Parallel Programing Languages and Accelerations

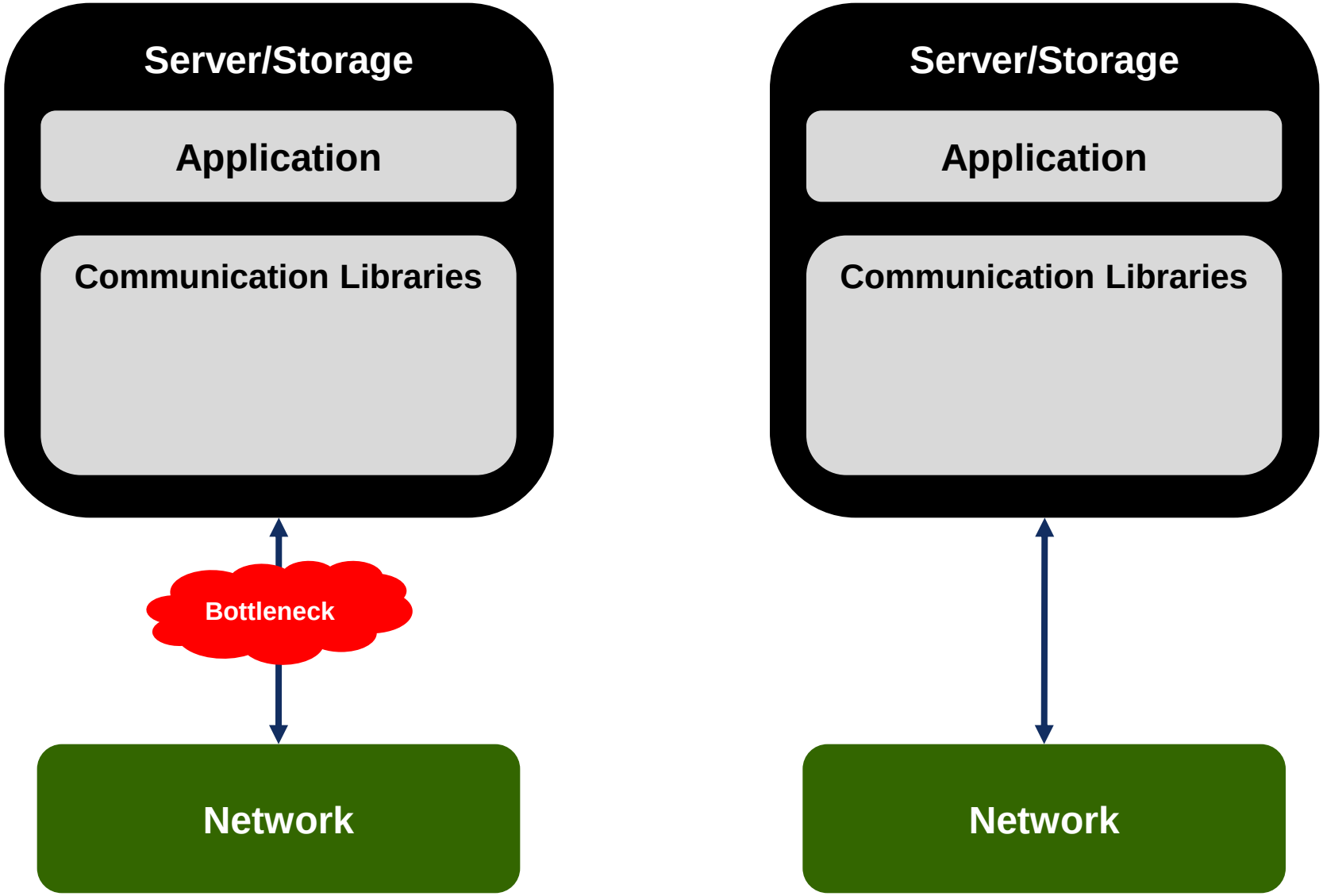
HPC@mellanox.com

Mellanox ScalableHPC

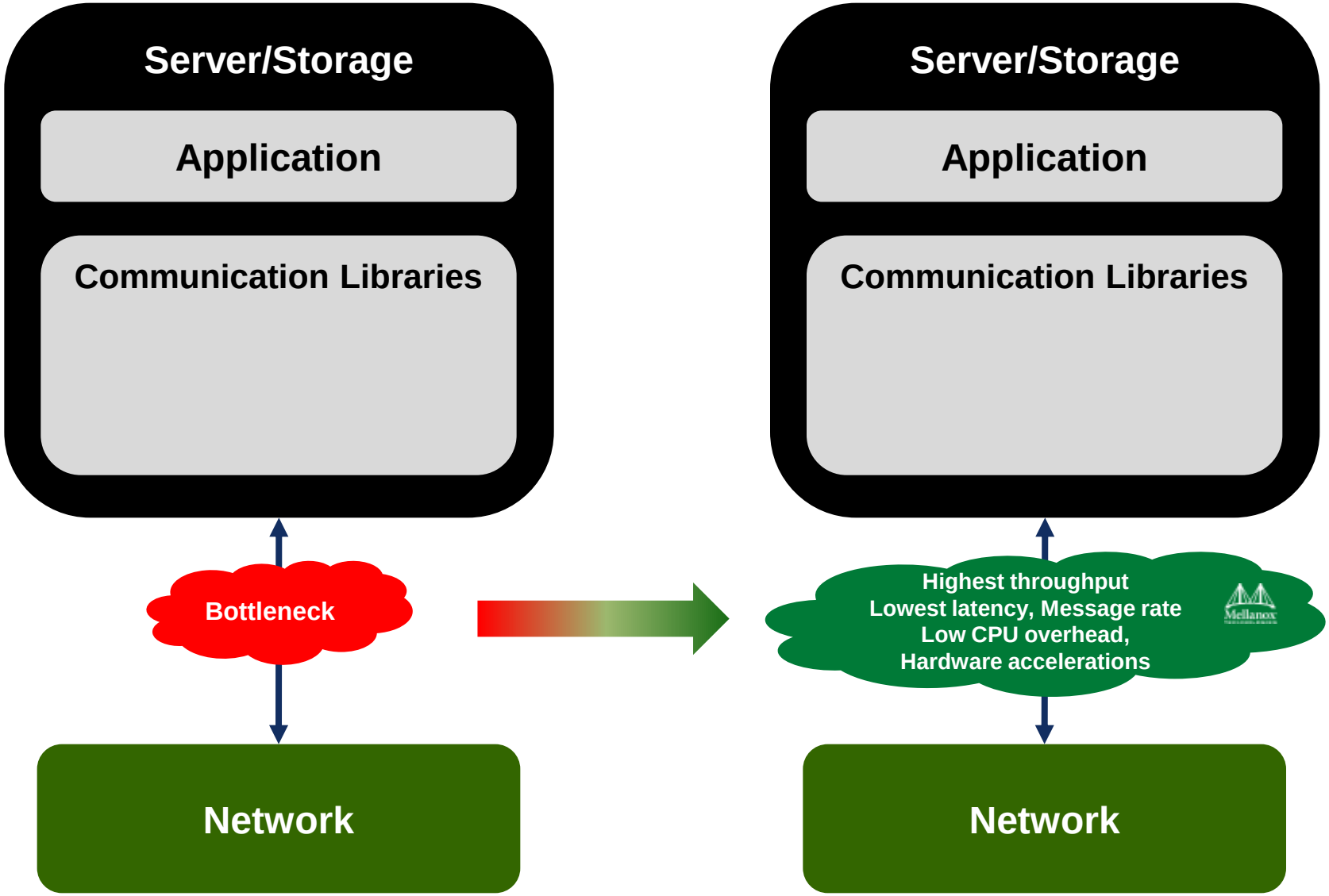
Mellanox ScalableHPC

- Offer high performing and scalable parallel programming libraries for HPC
- Support a comprehensive set of MPIs and PGAS languages
  - Integration of Mellanox acceleration technology into broad list of languages
  - Provide our own language library package when there is no open source alternative
- Integrates Mellanox acceleration components into MPIs/PGAS languages
  - MXM – MellanoX Messaging Accelerator
  - FCA – Mellanox Fabric Collective Accelerator

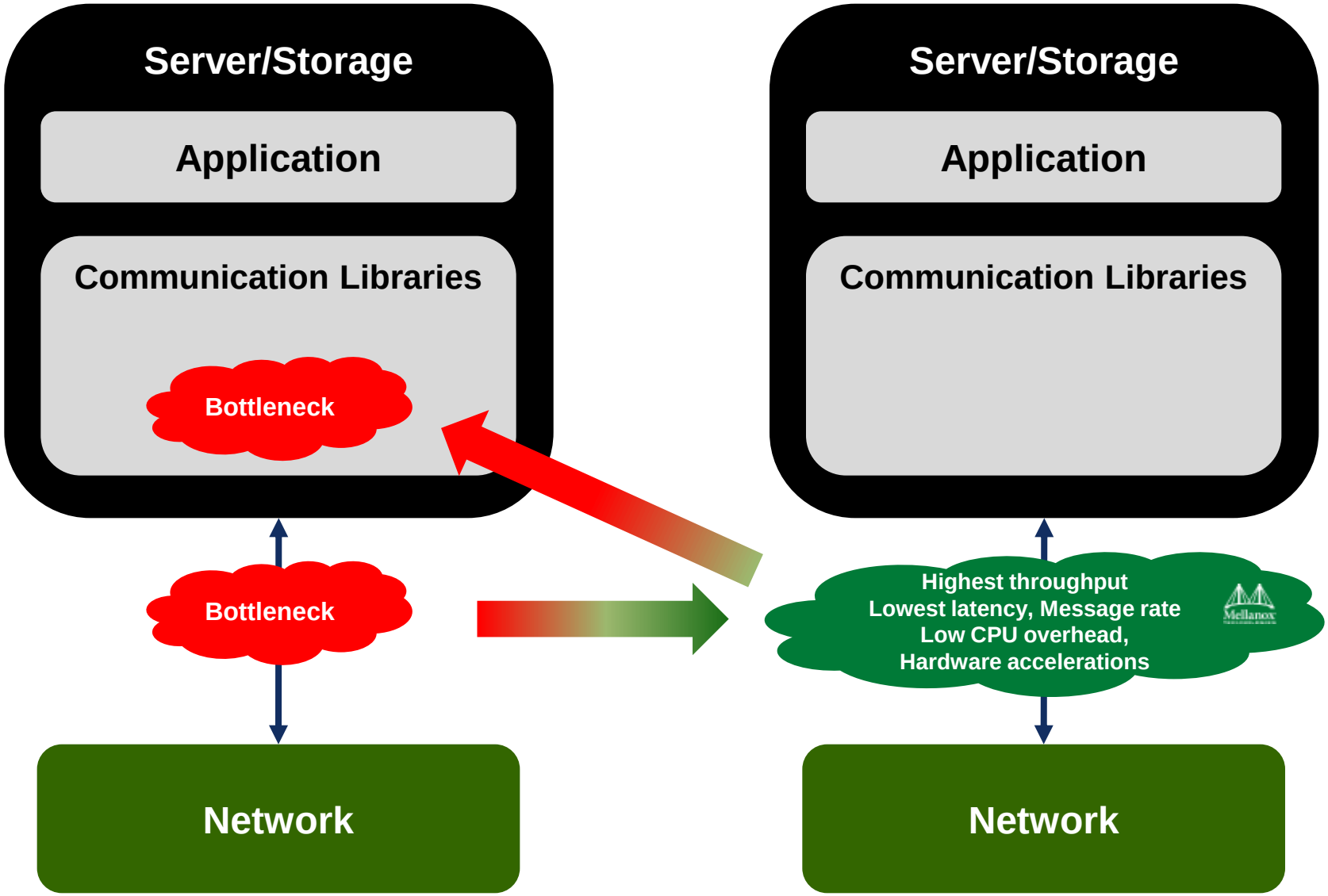
The I/O Bottleneck Paradigm



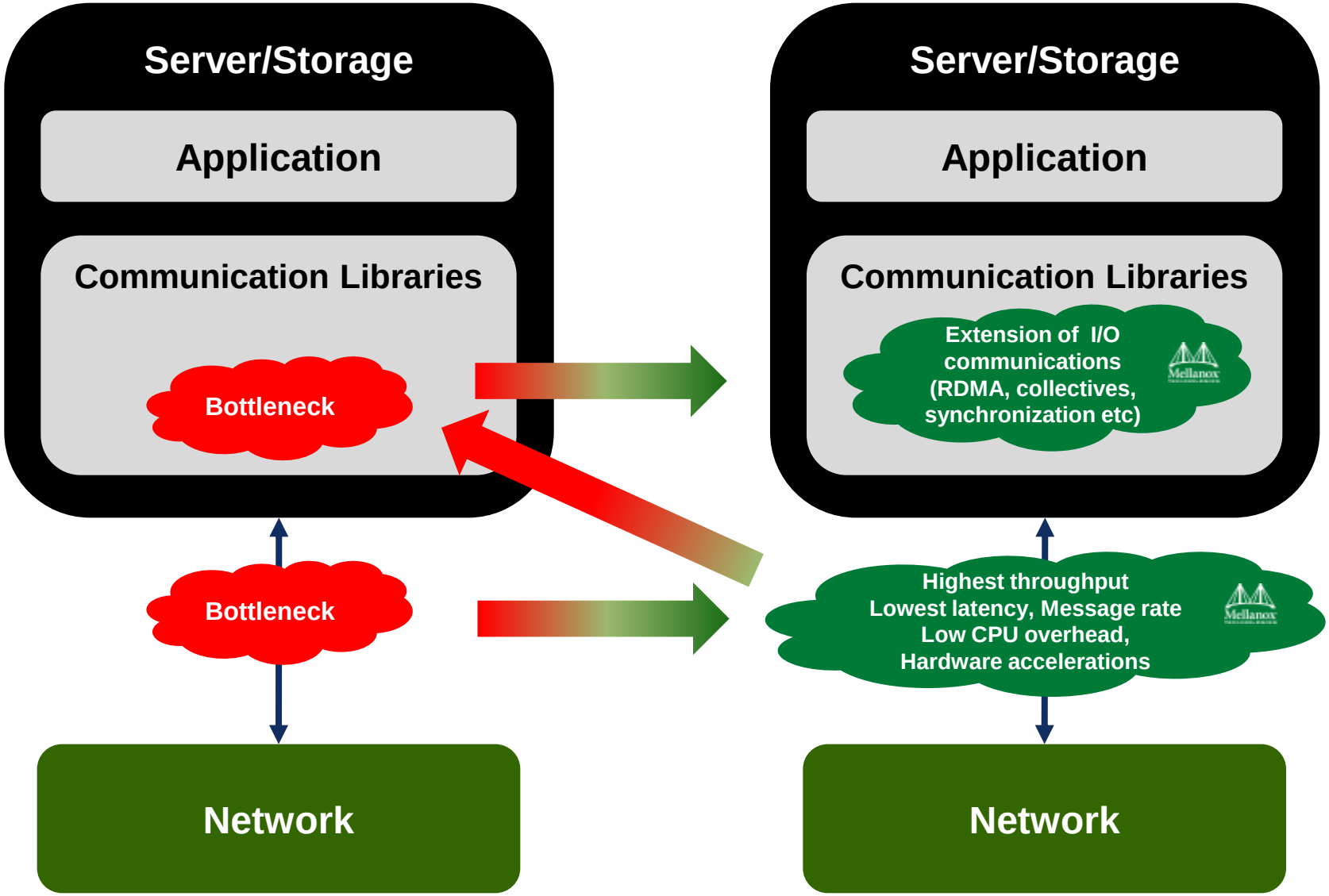
The I/O Bottleneck Paradigm



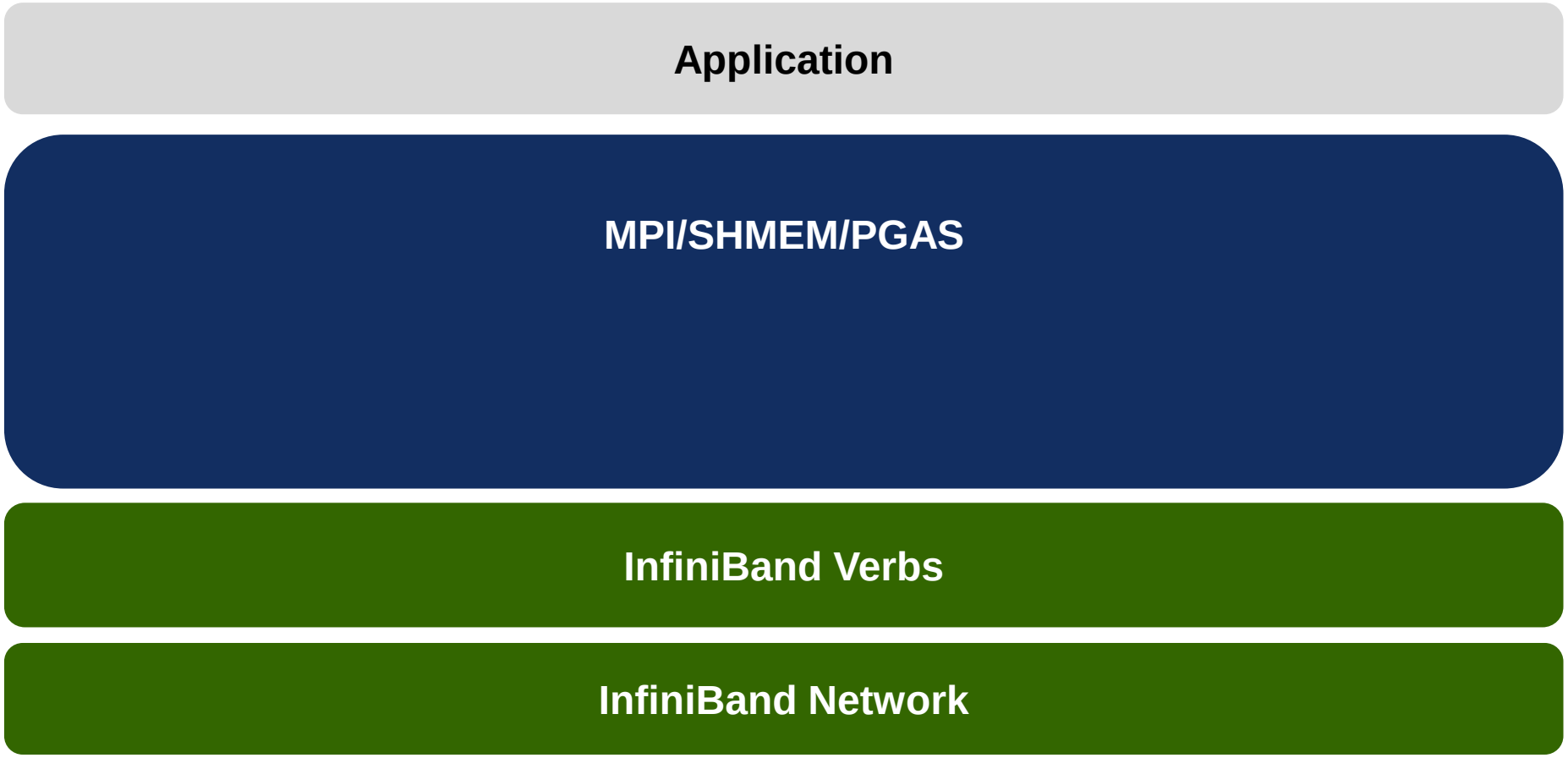
The I/O Bottleneck Paradigm – Scaling Issues



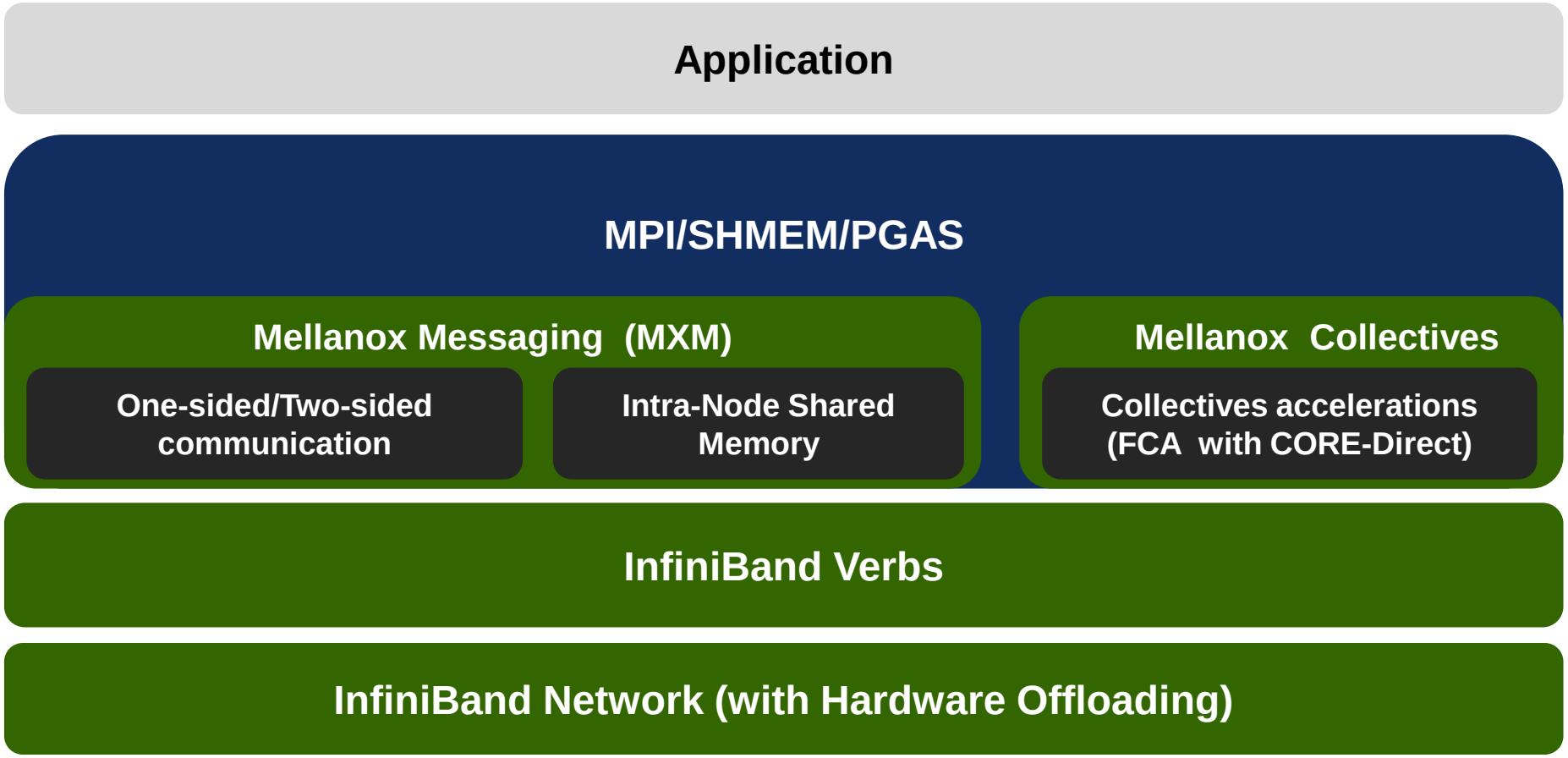
The I/O Bottleneck Paradigm – Co-Design Architecture



MPI/SHMEM/PGAS Architecture



Mellanox ScalableHPC Architecture



Mellanox Fabric Collective Accelerations (FCA)

- High performing and scalable accelerations for collective operations
  - Topology aware collectives take advantage of optimized message coalescing
  - Makes use of powerful multicast capabilities in network for one-to-many communications
  - Run collectives on separate service level so no interference with other communications
  - Utilizes Mellanox CoreDirect collective hardware offload to minimize system noise

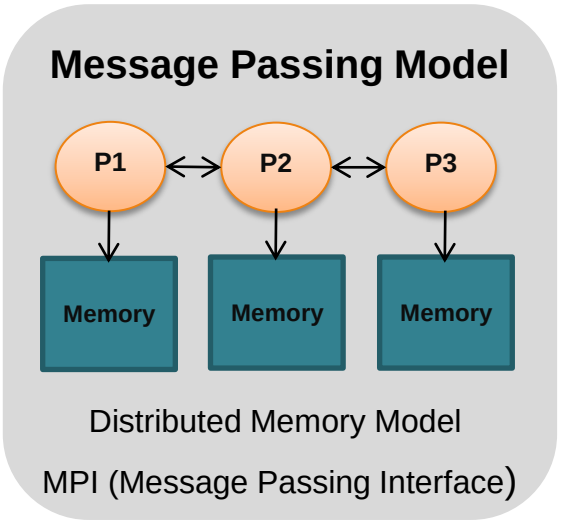
MellanoX Messaging (MXM)

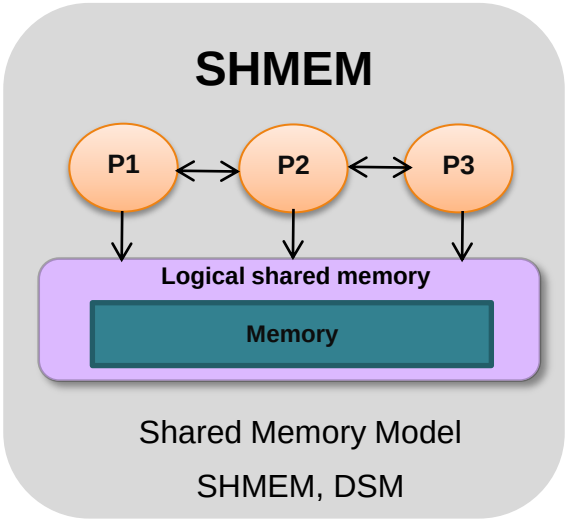
- High performance and scalability for send/receive (or put/get) messages
  - Proper management of HCA resources and memory structures
  - Optimized intra-node communication
  - Hybrid transport technology for large scale deployments
  - Efficient memory registration
  - Connection management
  - Receive Side tag matching
  - Fully utilizes hardware offloads and capabilities
  - Incorporated in MLNX_OFED-1.5.3-300 and later
    - Also provided as a stand-along package

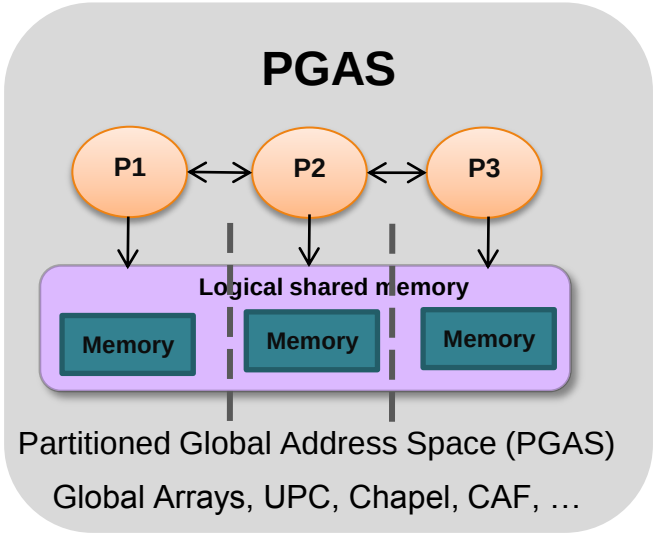
ScalableHPC Communication Libraries

HPC Parallel Programming Models

- MPI - Message Passing Interface
  - Based on Send/Receive and collectives communication semantics
- SHMEM - Shared Memory
  - Provides logically shared memory model and one-way put/get communications
- PGAS - Partitioned Global Address Space
  - Message passing abstracted into a partitioned global address space
  - UPC (Unified Parallel C) is one example of a PGAS language





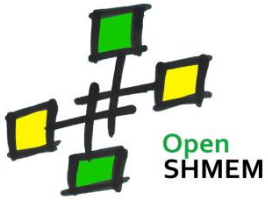


SHMEM Details

- **SH**ared **MEM**ory library
  - Library of functions somewhat similar to MPI (e.g. *shmem_get())*
  - *….but SHMEM supports one-sided communication (puts/gets vs. MPI's send/receive)*

- SHMEM and PGAS both allow for a unique combination of using a 'Distributed Memory Model' (like MPI), and a 'Shared Memorv Model' (like SMP machines)

- Cray first introduced SHMEM in 1993



- OpenSHMEM consortium formed to consolidate the various SHMEM versions into a widely accepted standard

- Mellanox ScalableSHMEM based on OpenSHMEM-1.0 specification with FCA/MXM integration

UPC Details

- UPC, or ‘Unified Parallel C’ is another PGAS language
- Higher level abstraction than MPI or SHMEM
- Allows programmers to directly represent and manipulate distributed data structures
- Commercial compilers are available for Cray, SGI and HP machines
- Open source compiler from LBNL/UCB (Berkeley UPC) available on InfiniBand
- Mellanox ScalableUPC based on BUPC with FCA/MXM integration

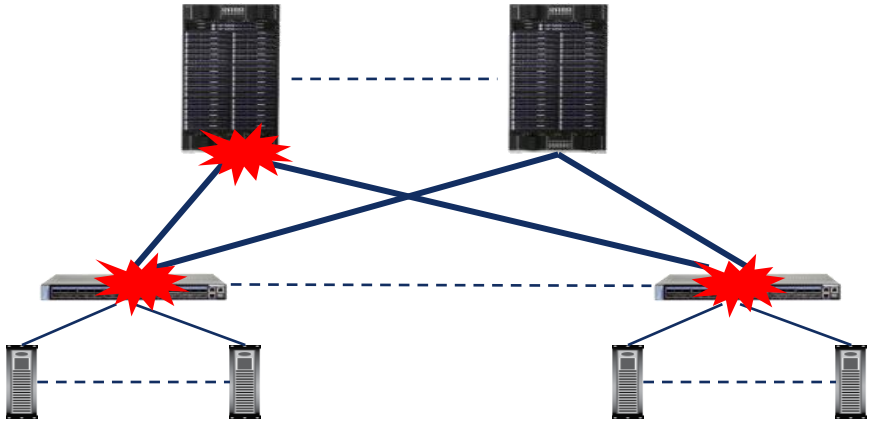
FCA Details

What are Collective Operations?

- Collective Operations are Group Communications involving all processes in job
- Synchronous operations
  - By nature consume many ‘Wait’ cycles on large clusters
- Popular examples
  - Barrier
  - Reduce
  - Allreduce
  - Gather
  - Allgather
  - Bcast

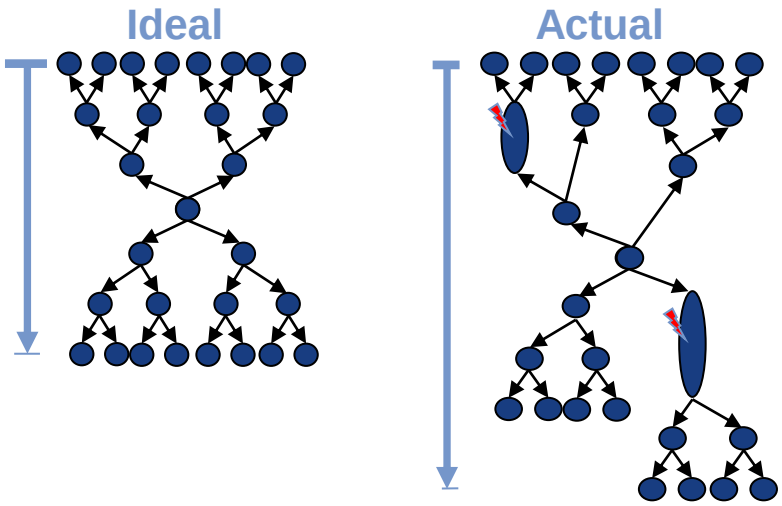
Collective Operation Challenges at Large Scale

- Collective algorithms are not topology aware and can be inefficient

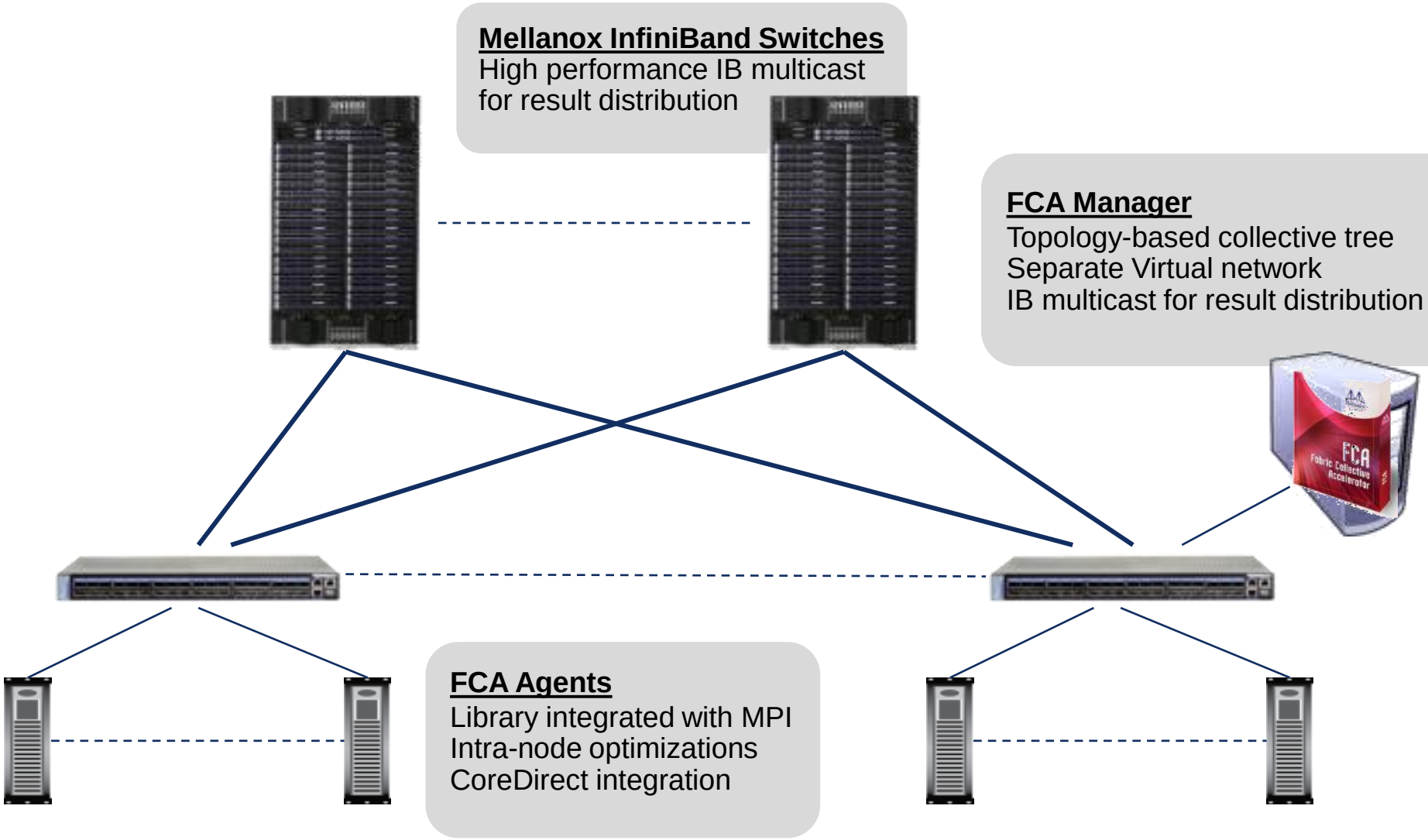


- Congestion due to many-to-many communications

- Slow nodes and OS jitter affect scalability and increase variability

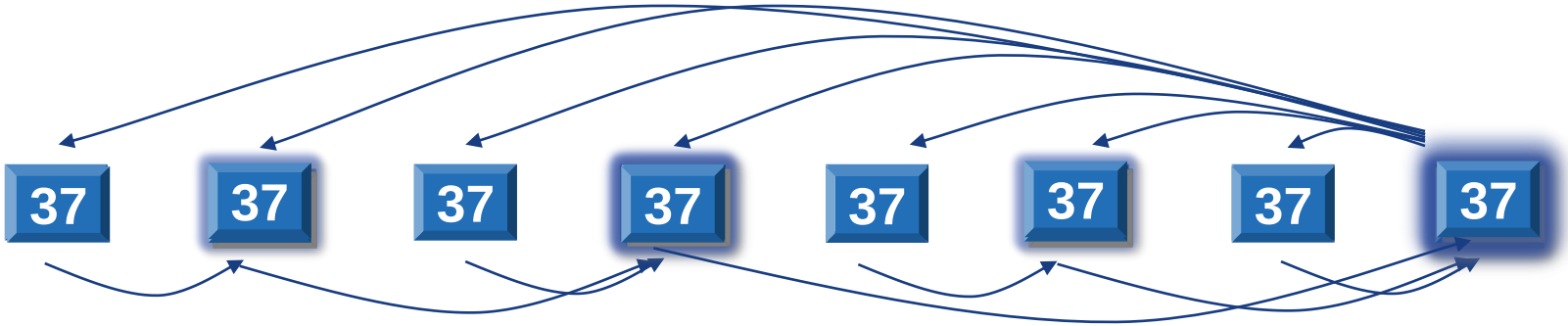


Mellanox Fabric Collectives Accelerations (FCA)



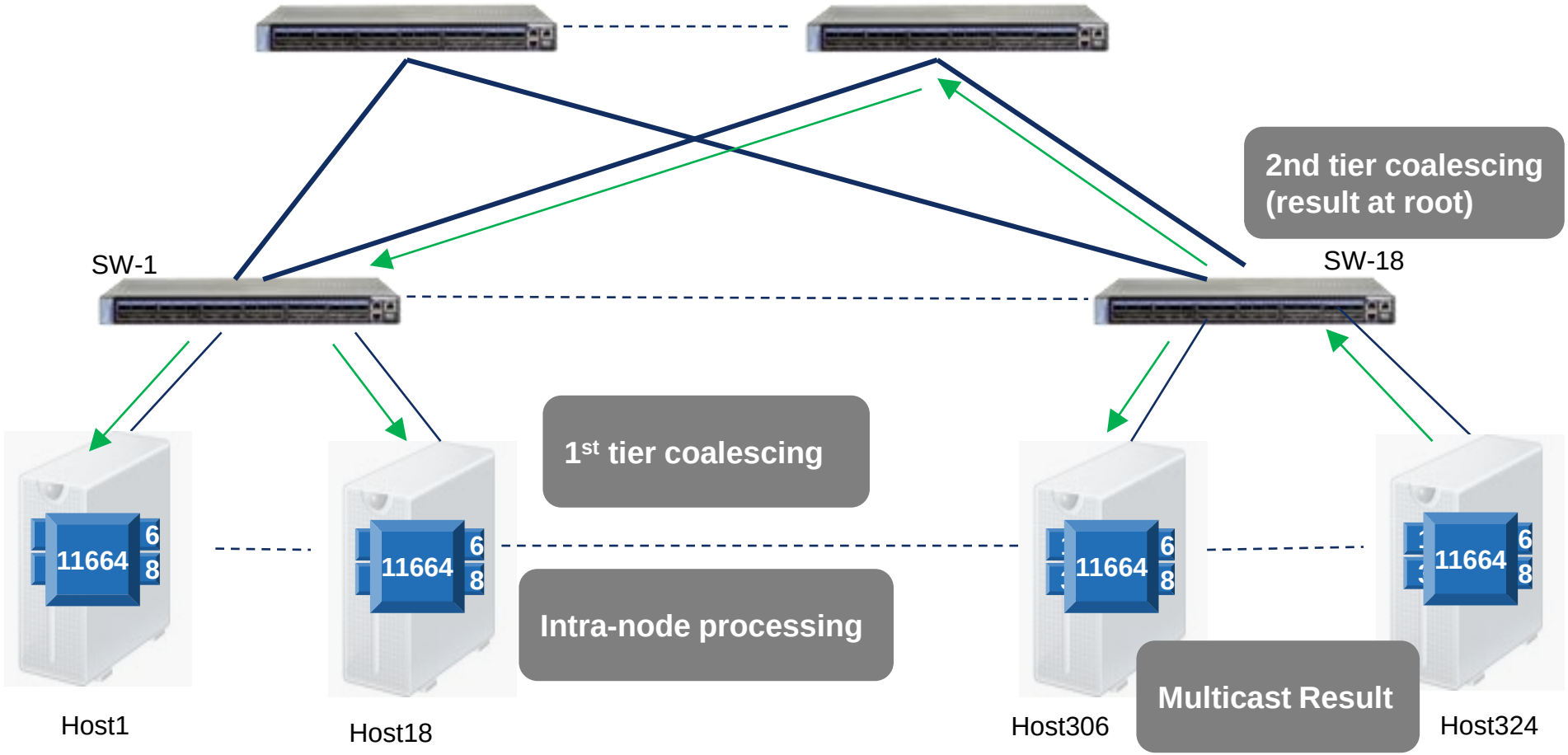
Collective Example – Allreduce using Recursive Doubling

- Collective Operations are Group Communications involving all processes in job



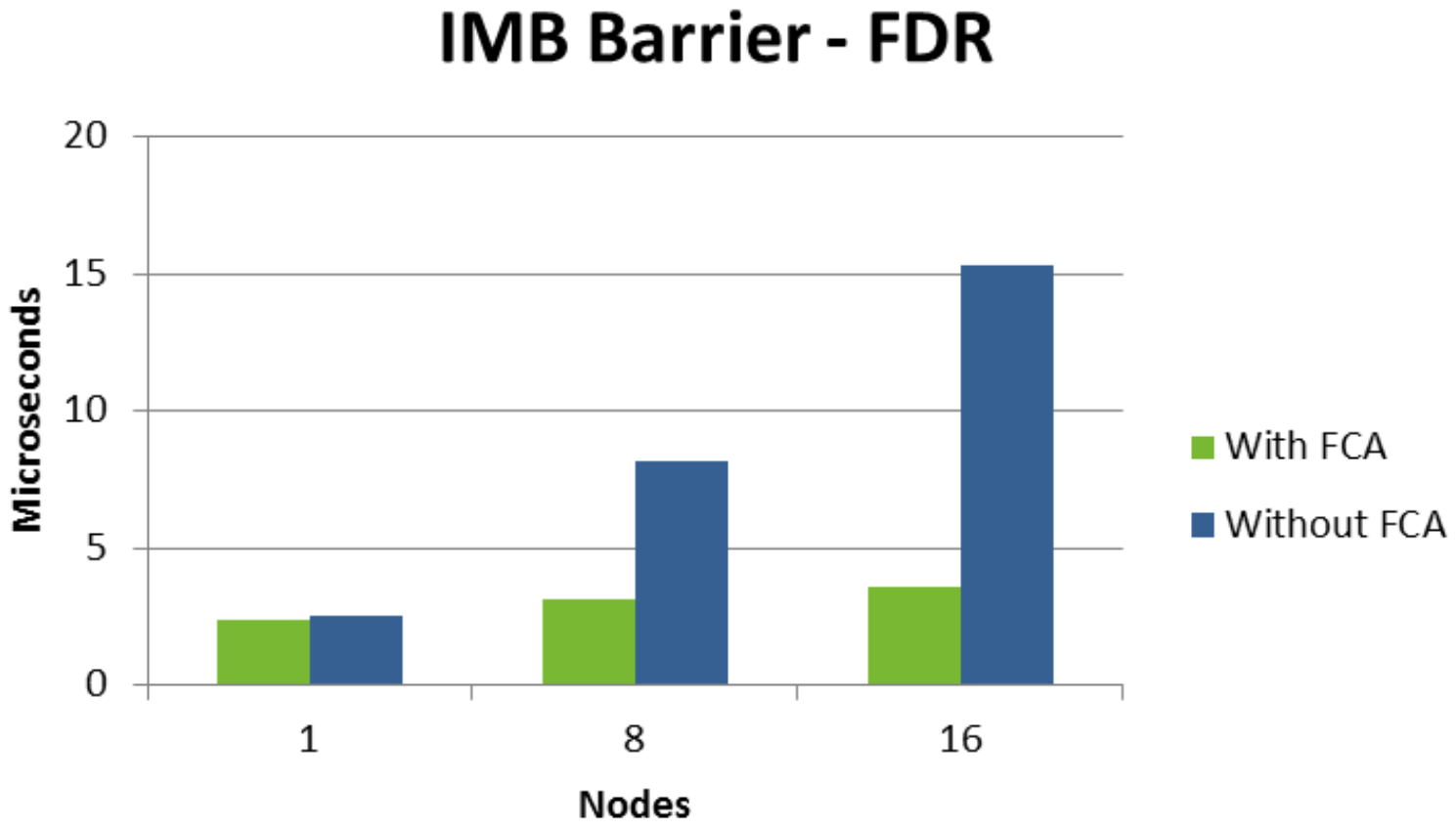
- A 4000 process Allreduce using recursive doubling is 12 stages

Scalable Collectives with FCA

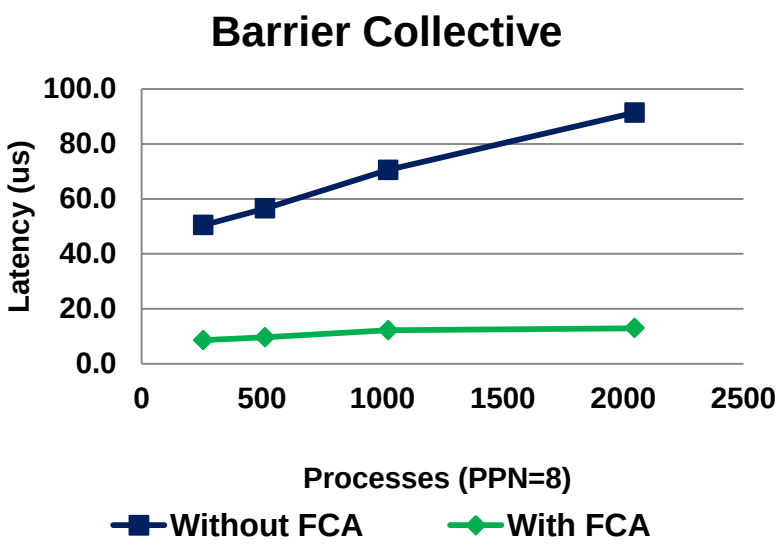


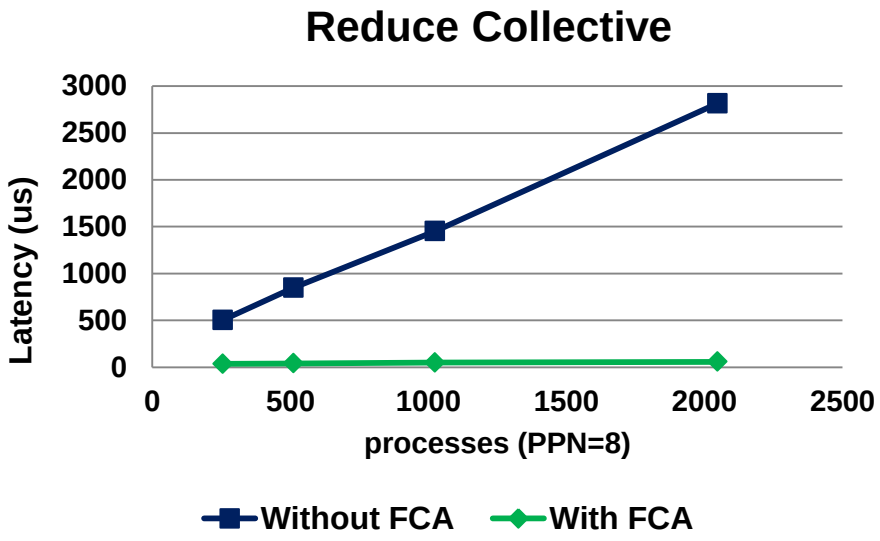
Performance Results

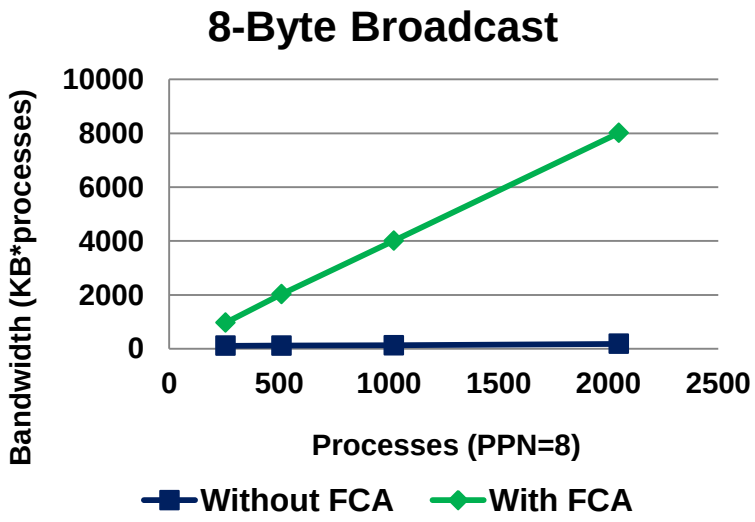
FCA collective performance with OpenMPI



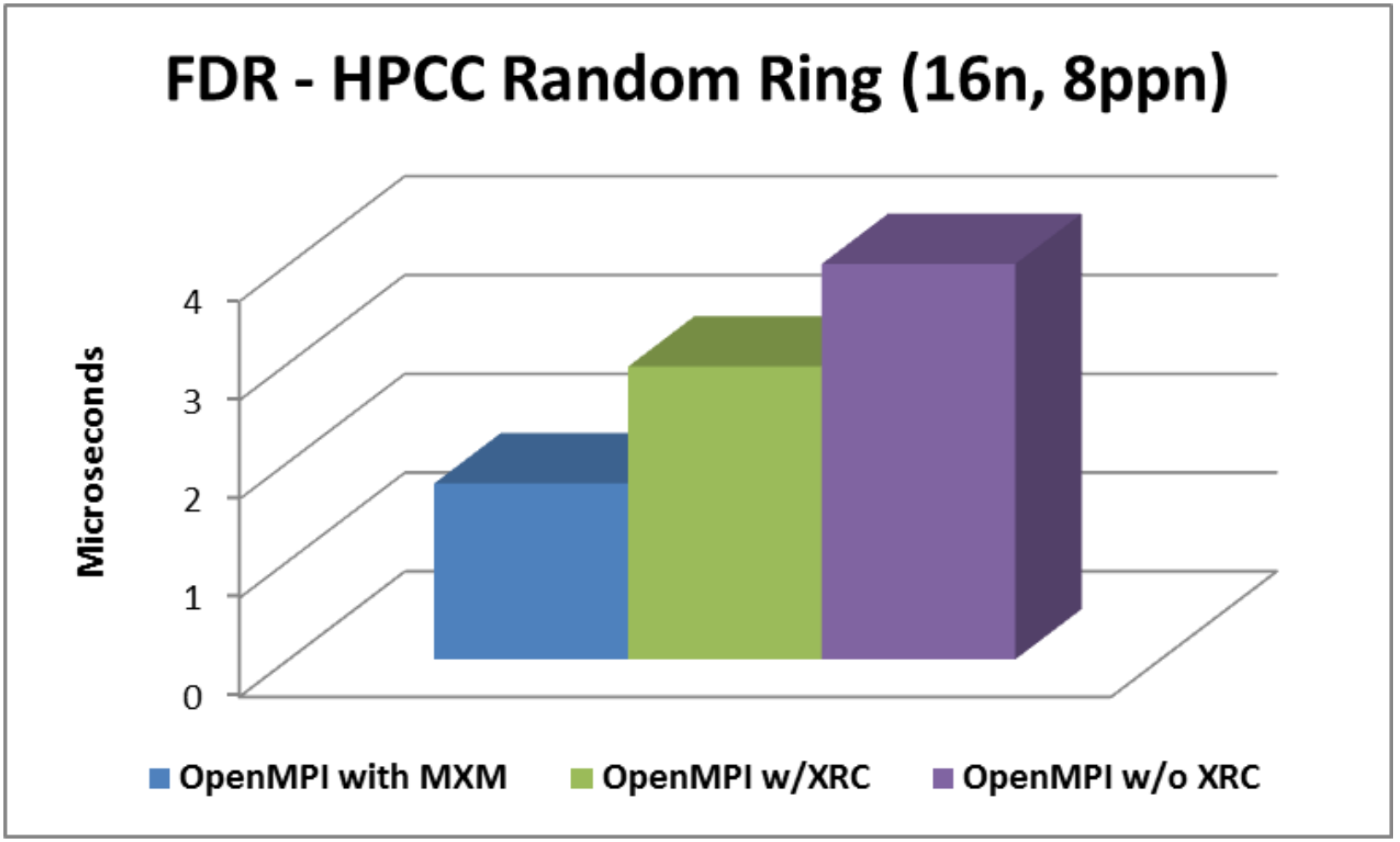
FCA collective scalability for SHMEM







Mellanox MXM – HPCC Random Ring Latency



Thank You

HPC@mellanox.com