The InfiniBand Advantage

Joshua S. Ladd, PhD

Staff Engineer, HPC Software R&D Group

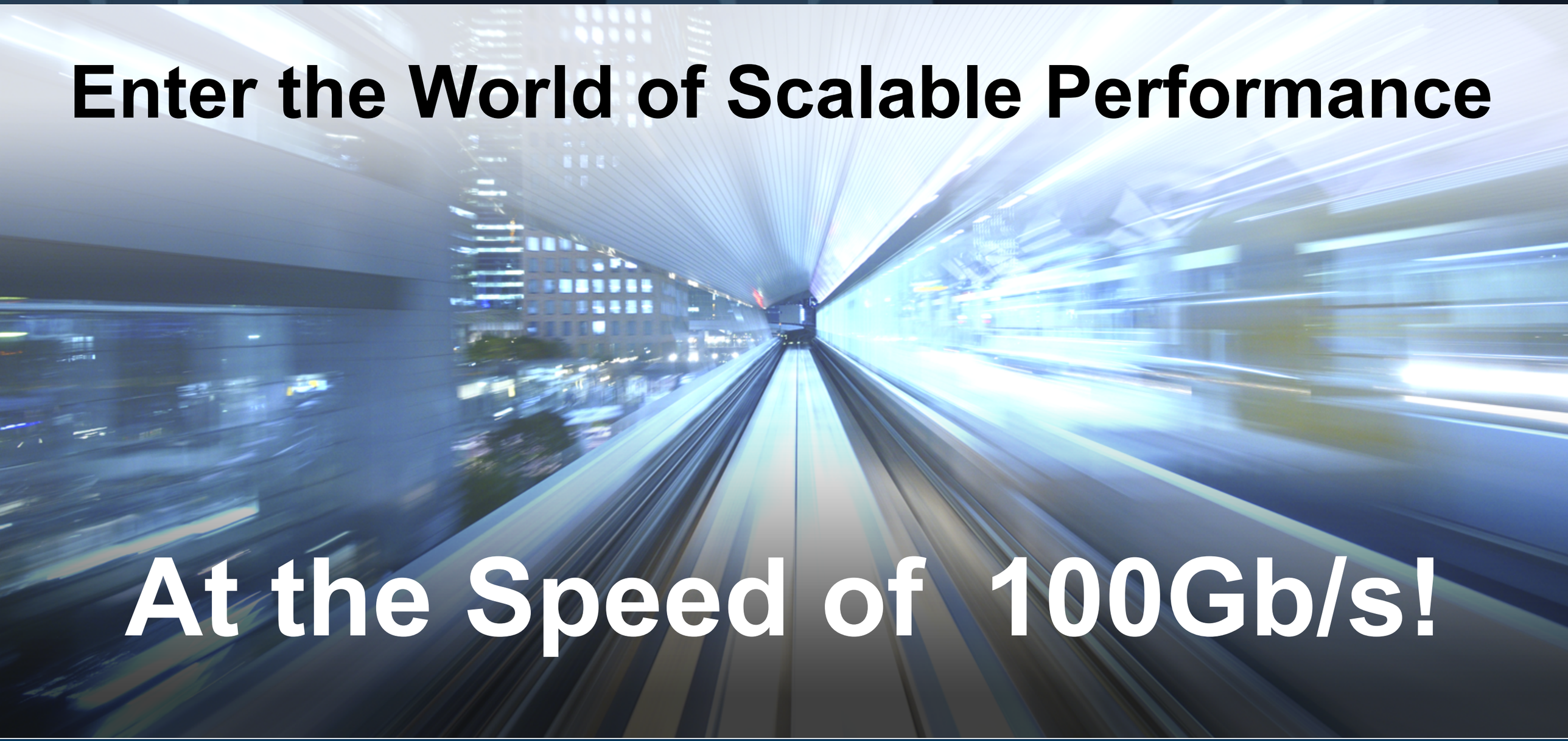
OpenSHMEM 2015, Annapolis, Maryland

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Exascale-Class Computer Platforms – Communication Challenges

Challenge	Solution focus
Very large functional unit count ~10M	Scalable communication capabilities: point-to-point & collectives Scalable Network: Adaptive routing
Large on-”node” functional unit count ~500	Scalable HCA architecture
Deeper memory hierarchies	Cache aware network access
Smaller amounts of memory per functional unit	Low latency, high b/w capabilities
May have functional unit heterogeneity	Support for data heterogeneity
Component failures part of “normal” operation	Resilient and redundant stack
Data movement is expensive	Optimize data movement
Independent remote progress	Independent hardware progress
Power costs	Power aware hardware



Enter the World of Scalable Performance

At the Speed of 100Gb/s!

The Future is Here

Entering the Era of 100Gb/s

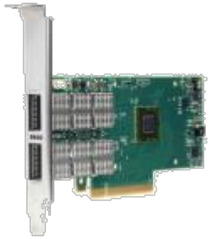
Adapters

ConnectX®-4

100Gb/s Adapter, 0.7us latency

150 million messages per second

(10 / 25 / 40 / 50 / 56 / 100Gb/s)



Switch

SwitchIB™

36 EDR (100Gb/s) Ports, <90ns Latency

Throughput of 7.2Tb/s

Interconnect

LinkX™

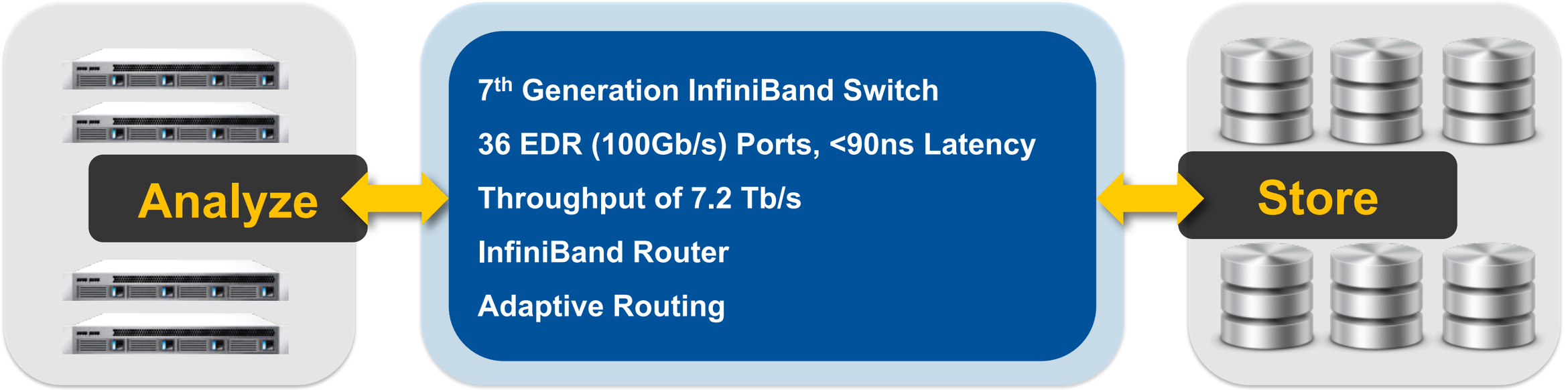
Copper (Passive, Active)

Optical Cables (VCSEL)

Silicon Photonics

Enter the World of Scalable Performance – 100Gb/s Switch

Switch-IB: Highest Performance Switch in the Market



SwitchIB™

Enter the World of Scalable Performance – 100Gb/s Adapter

ConnectX-4: Highest Performance Adapter in the Market

InfiniBand: SDR / DDR / QDR / FDR / EDR

Ethernet: 10 / 25 / 40 / 50 / 56 / 100GbE

100Gb/s, <0.7us latency

150 million messages per second

OpenPOWER CAPI technology

CORE-Direct technology

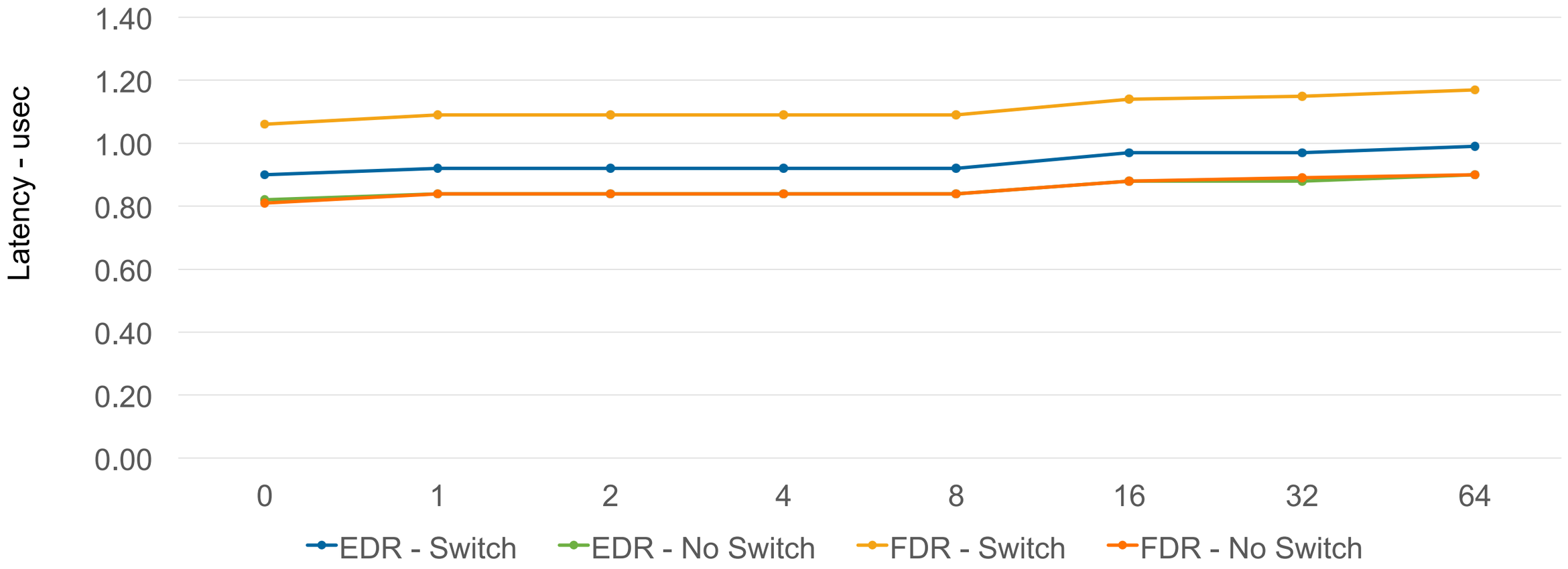
GPUDirect RDMA

Dynamically Connected Transport (DCT)

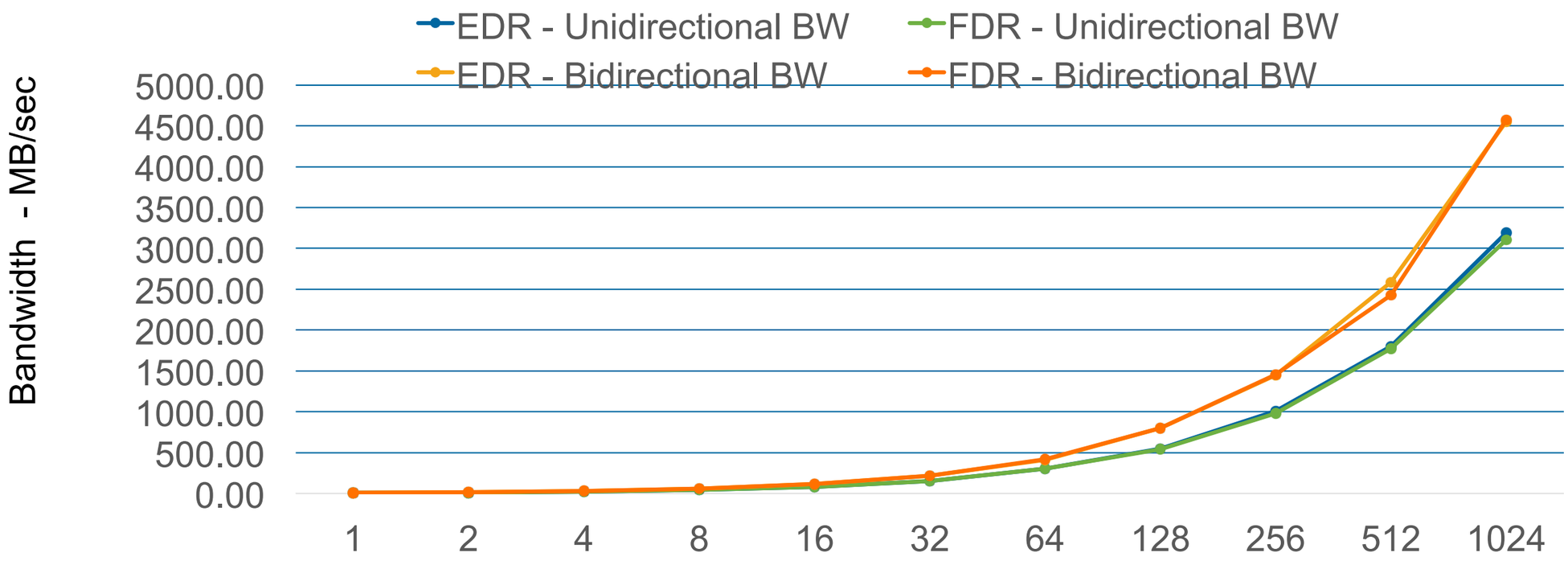
Ethernet offloads (HDS, RSS, TSS, LRO, LSOv2)

Point-to-Point Data

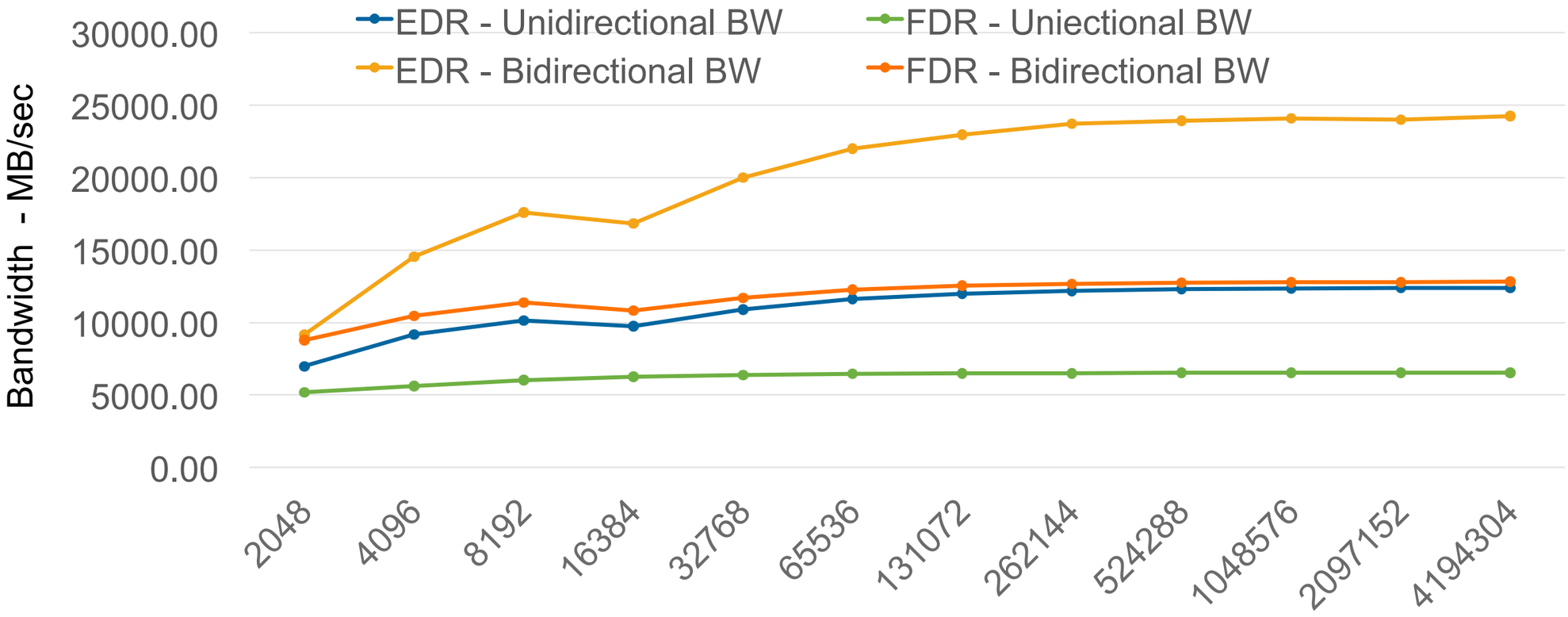
MPI Latency – OSU Latency test



MPI Bandwidth – OSU Bandwidth test

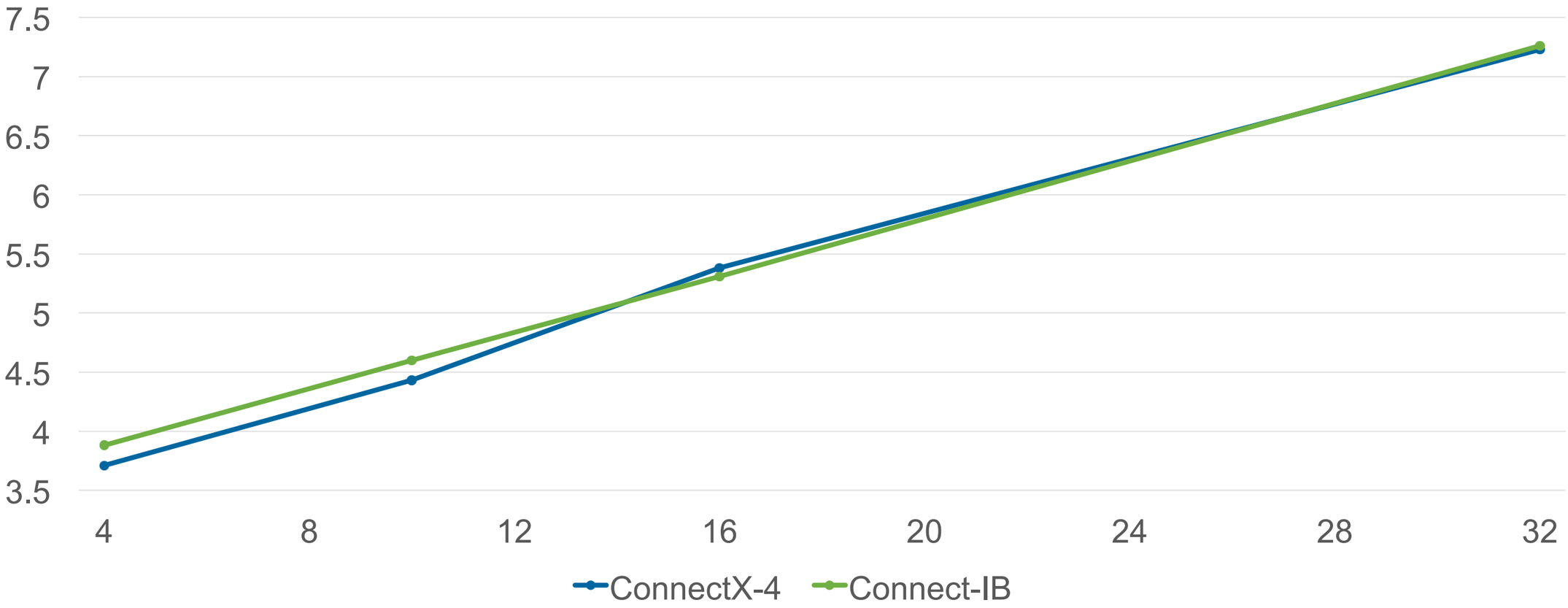


MPI Bandwidth – OSU Bandwidth test

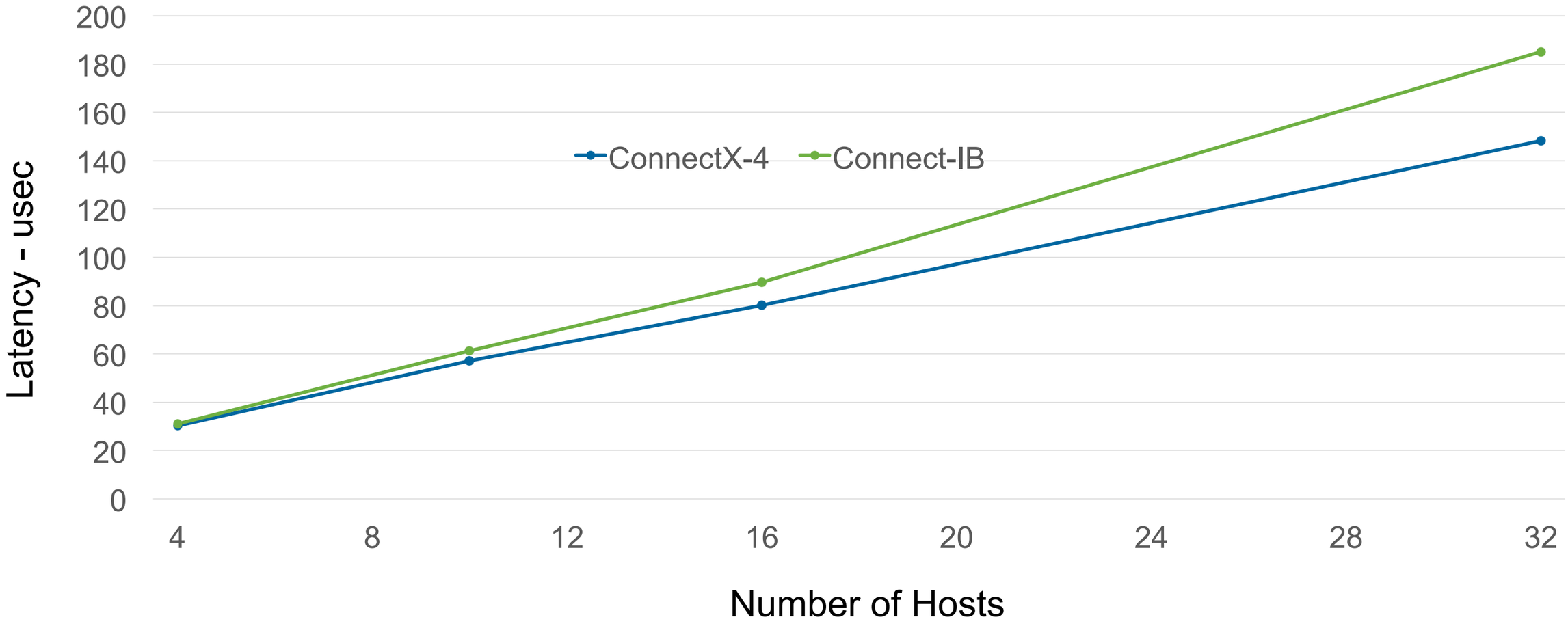


Collective Communication

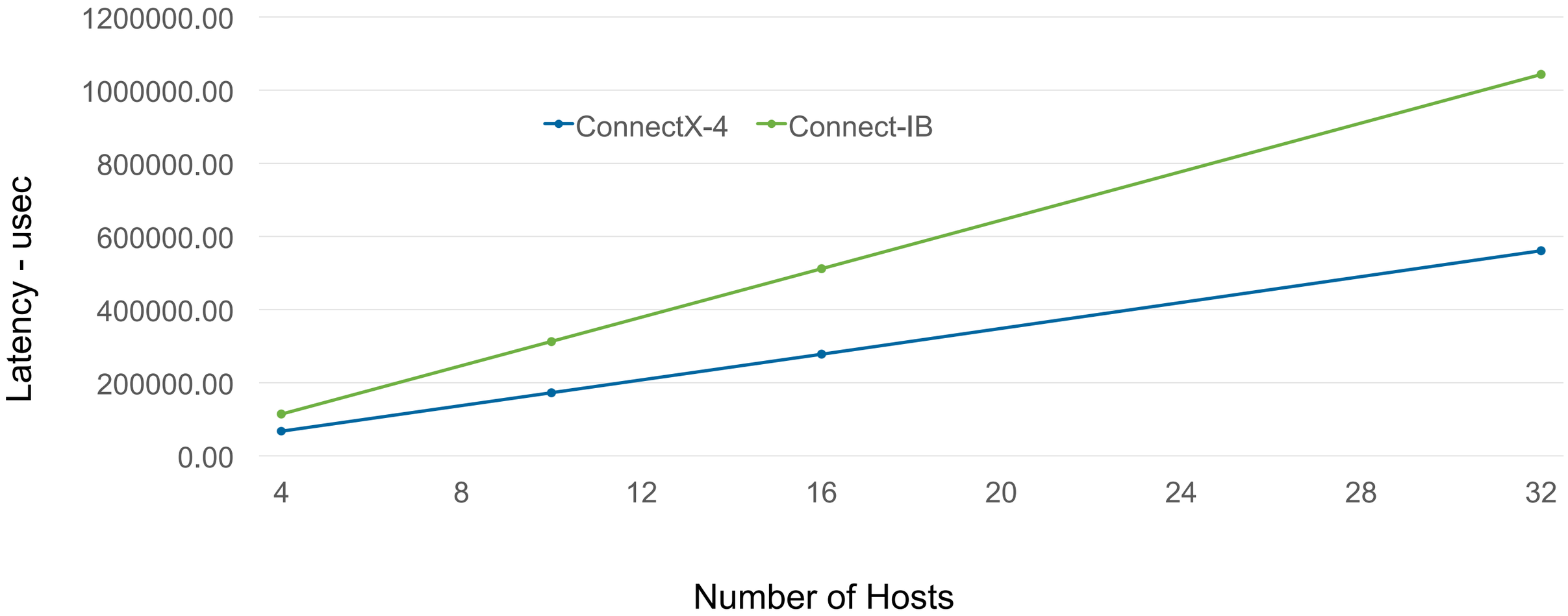
Barrier



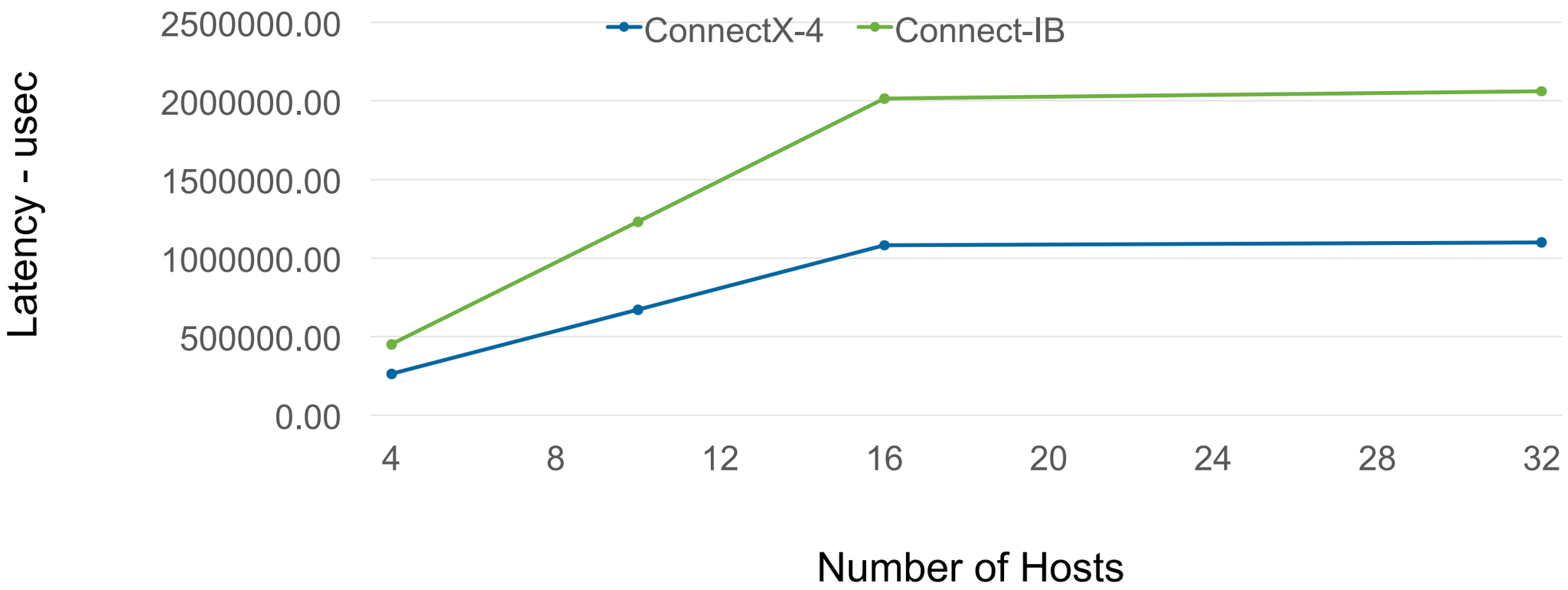
All-to-All - 8 bytes



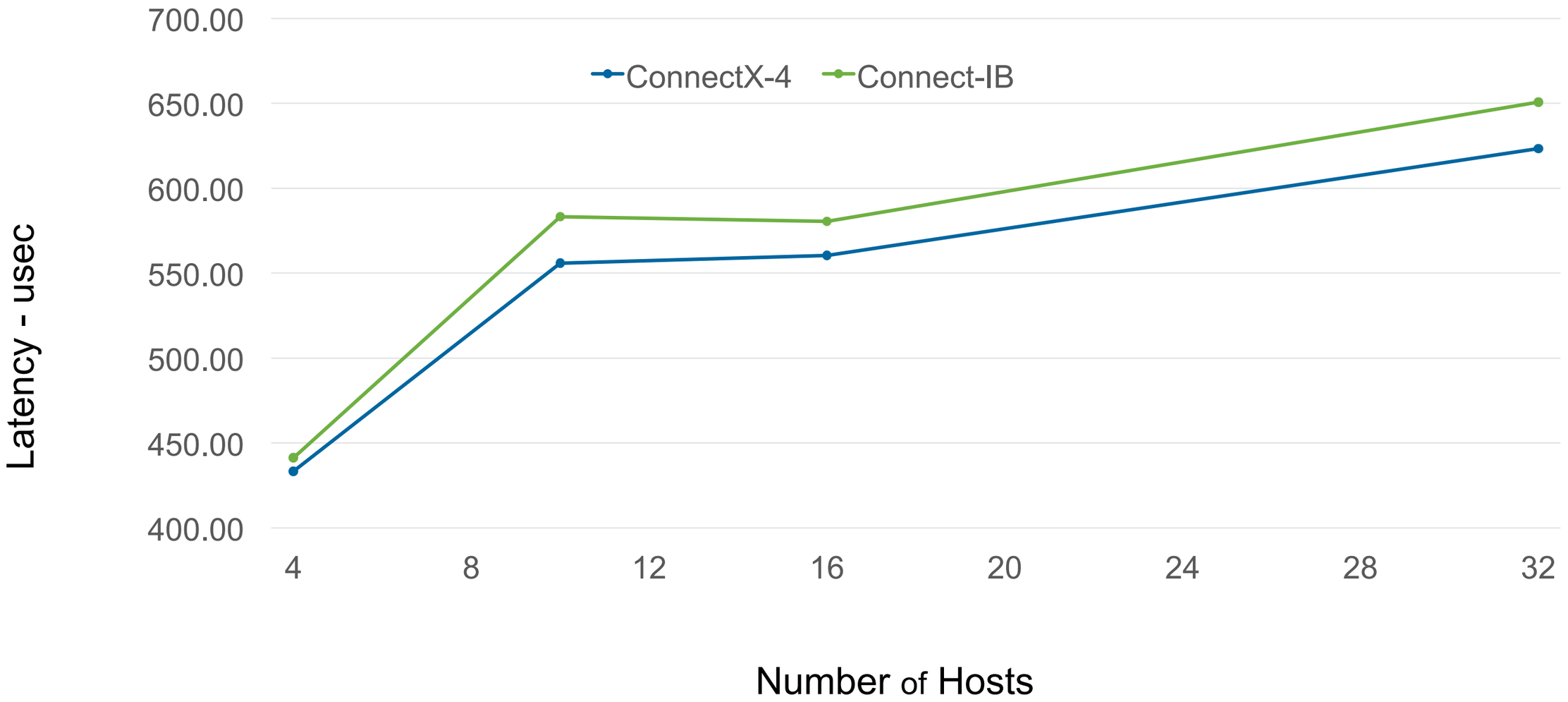
All-to-All – 256 Kbytes



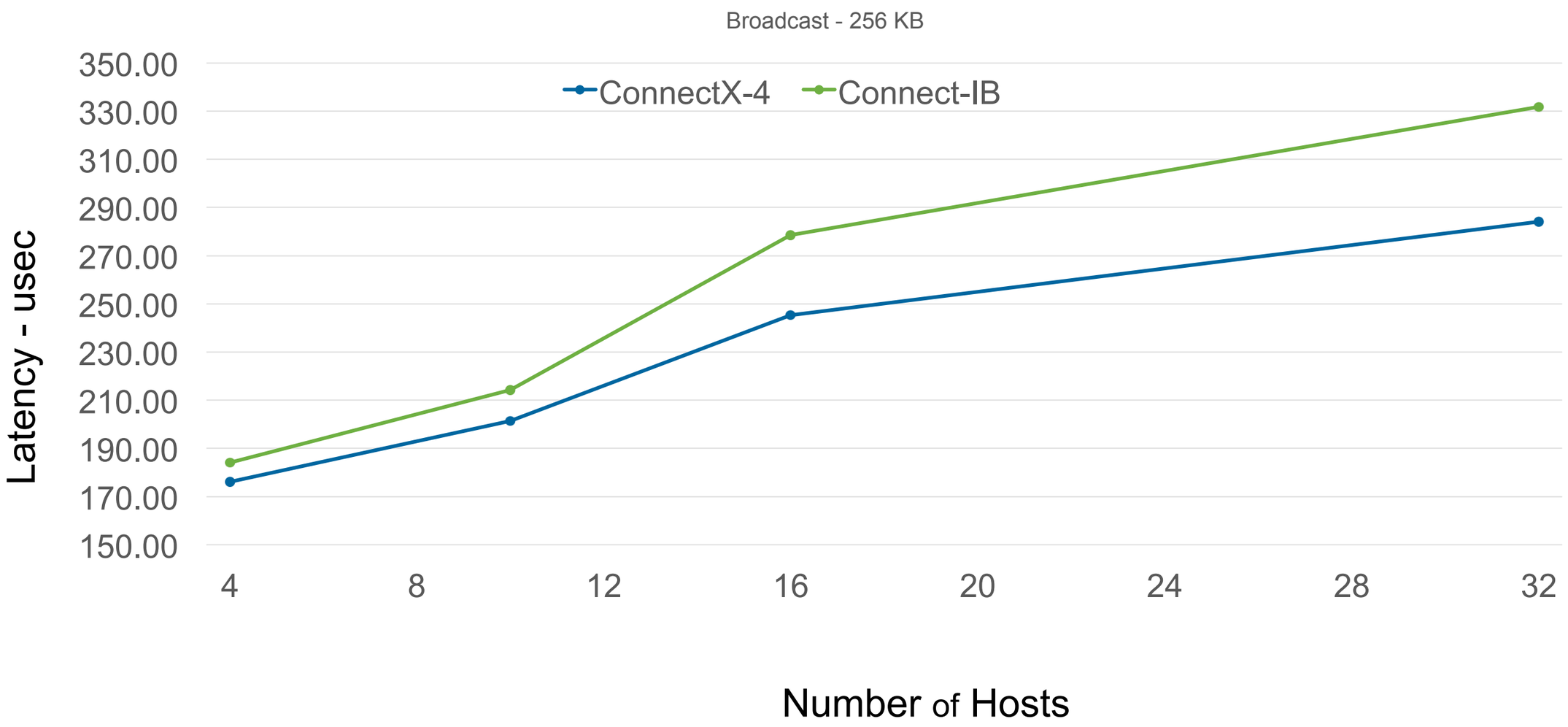
All-to-All – 1 Mbyte



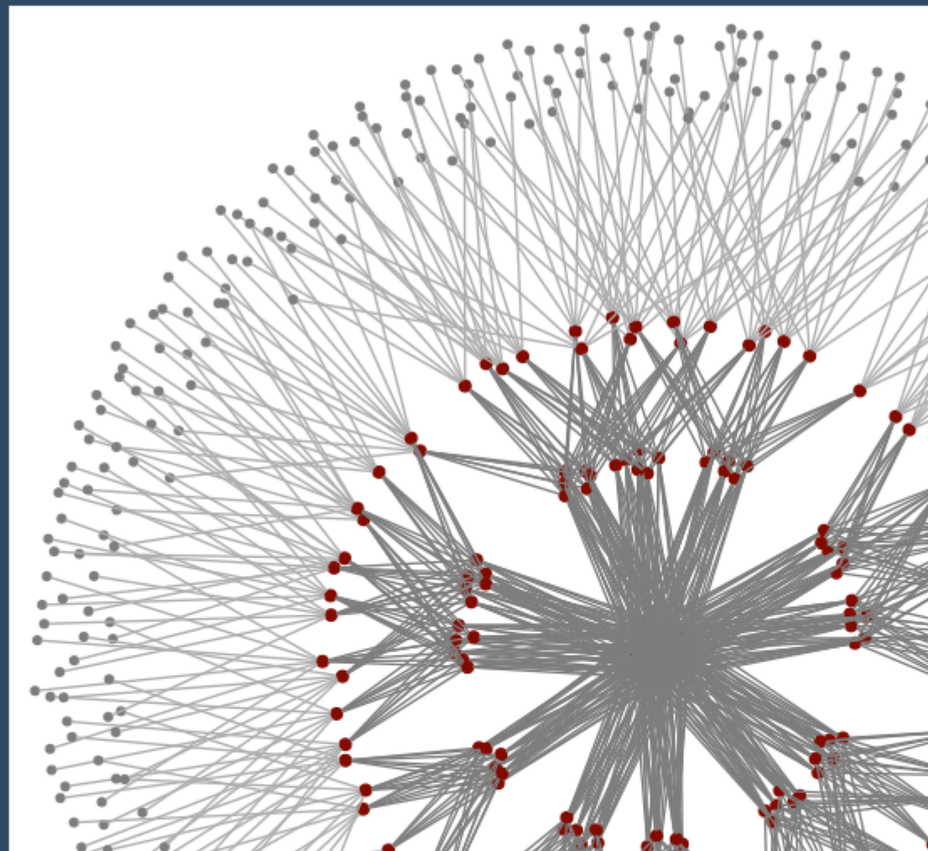
Allreduce – 256K Bytes



Broadcast – 256K Bytes

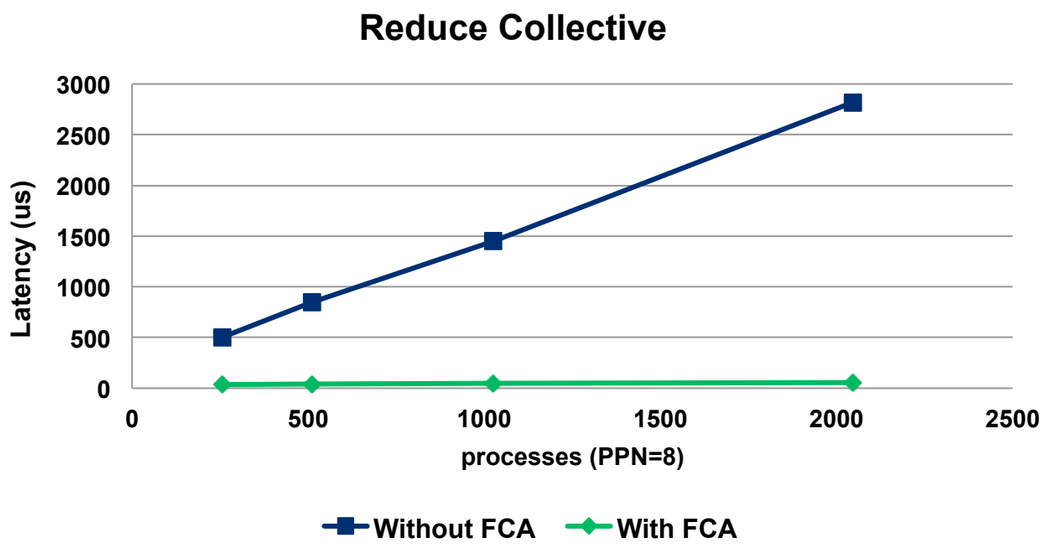


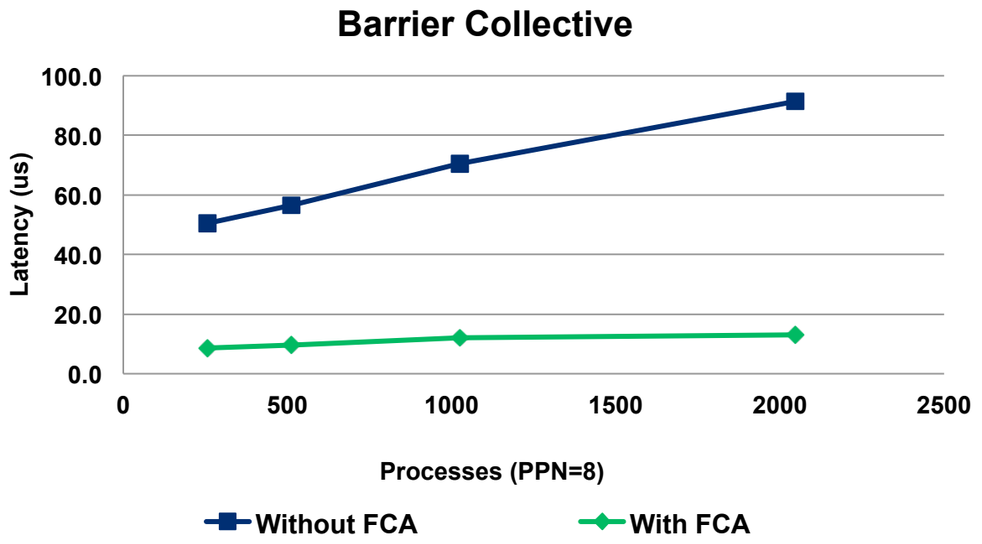
Scalability



FCA - Collective Operations

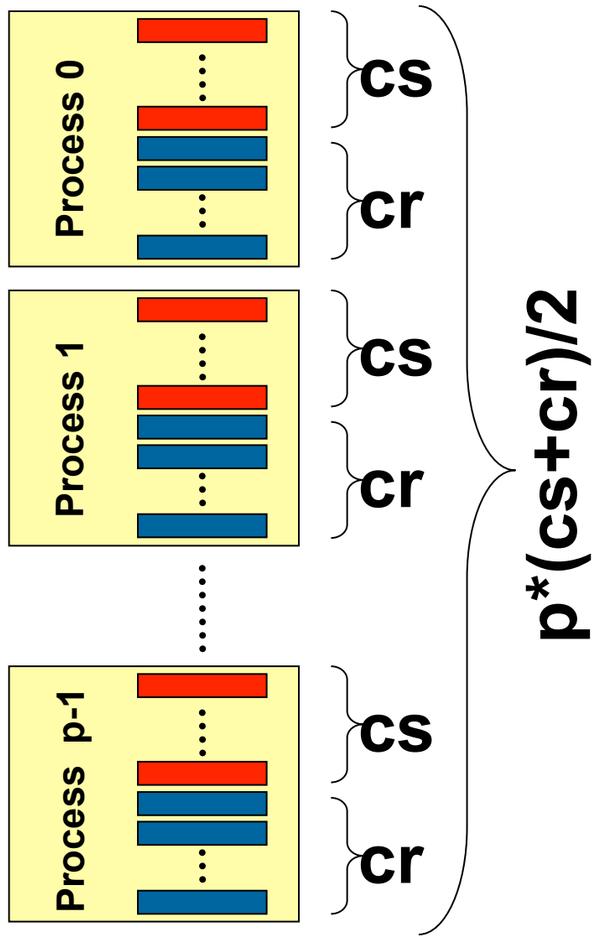
- Topology Aware
- Hardware Multicast
- Offload
- Scalable algorithms





The Dynamically Connected Transport Model

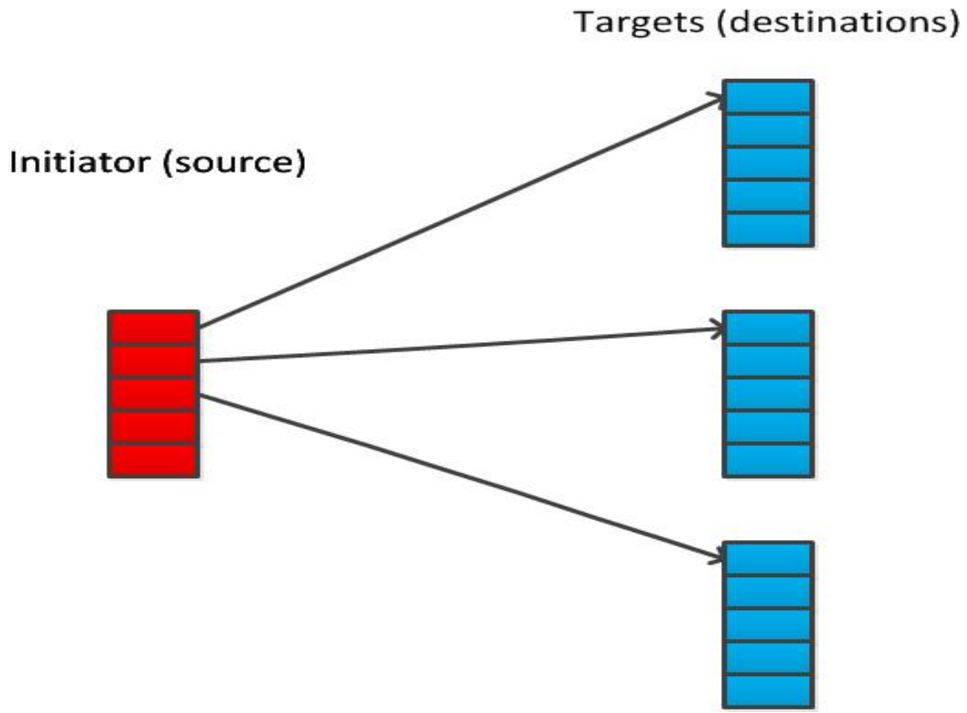
- Dynamic Connectivity
- Each DC Initiator can be used to reach any remote DC Target
- No resources' sharing between processes
  - process controls how many (and can adapt to load)
  - process controls usage model (e.g. SQ allocation policy)
  - no inter-process dependencies
- Resource footprint
  - Function of node and HCA capability
  - Independent of system size
- Fast Communication Setup Time



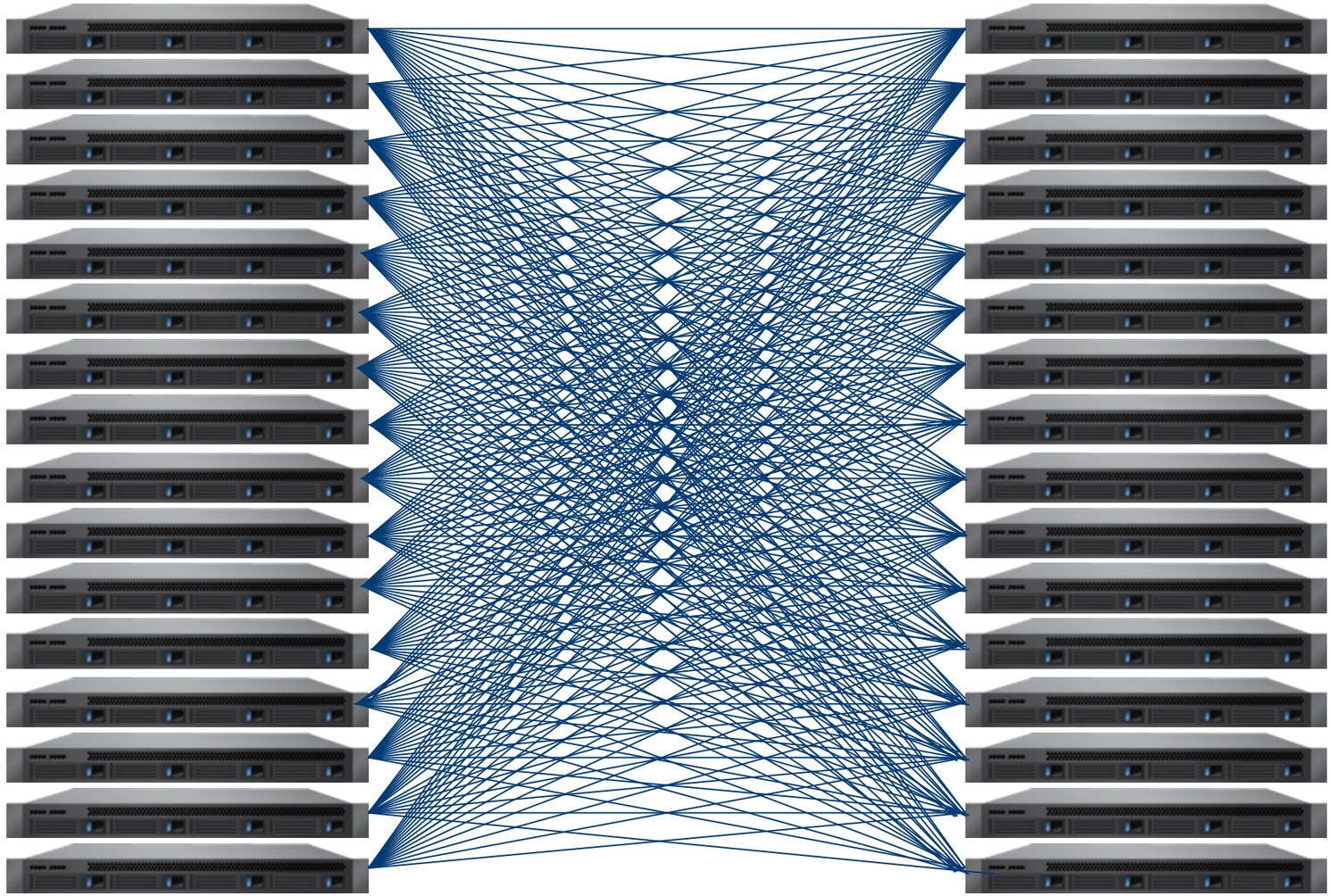
cs – concurrency of the sender
cr=concurrency of the responder

Dynamically Connected Transport

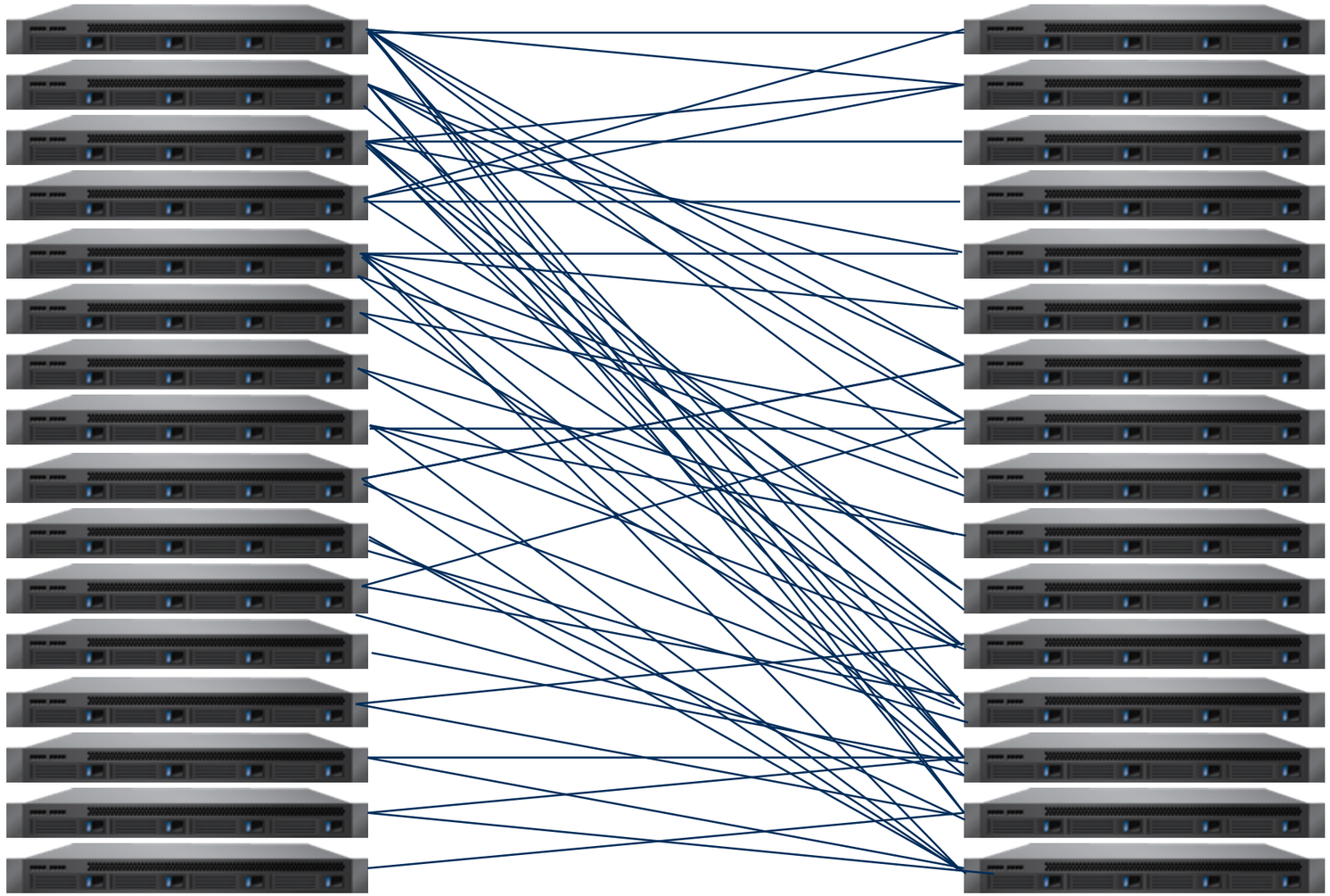
- Key objects
  - DC Initiator: Initiates data transfer
  - DC Target: Handles incoming data



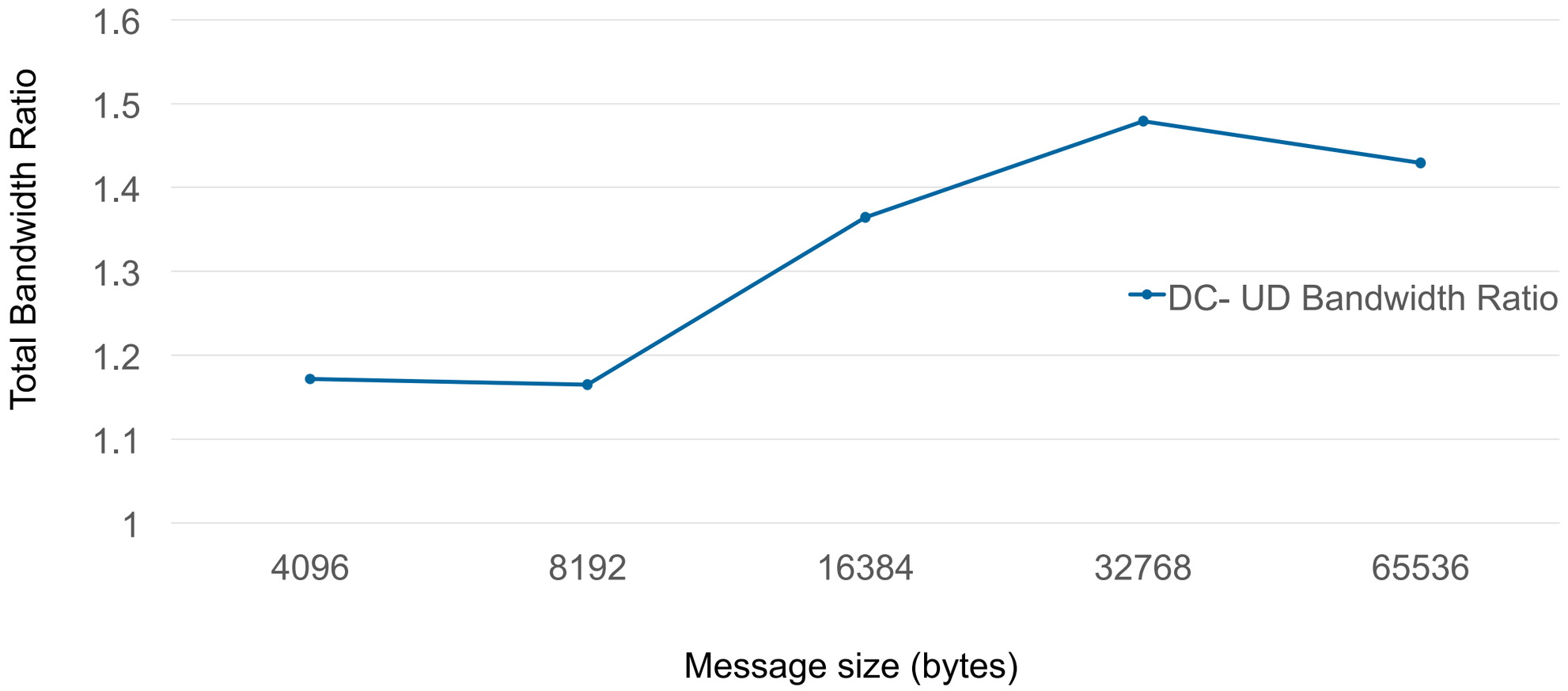
Reliable Connection Transport Mode



Dynamically Connected Transport Mode



All-To-All Performance



Scalability Under Load

Adaptive Routing

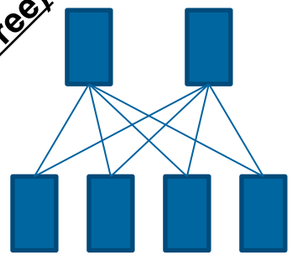
- Purpose
  - Improved Network utilization: choose alternate routes on congestion
  - Network resilience: Alternative routes on failure
- Supported Hardware
  - SwitchX-2
  - Switch-IB: Adaptive routing notification added

Mellanox Adaptive Routing – Hardware Support

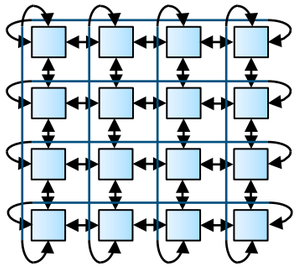
- Mellanox hardware is NOT topology specific
  - SDN concept separates the configuration plane from the data plane
  - Every feature is software controlled
  - Fat-Tree, Dragonfly and Dragonfly+ are fully supported
  - New hardware features introduced to support Dragonfly and Dragonfly+

Topologies

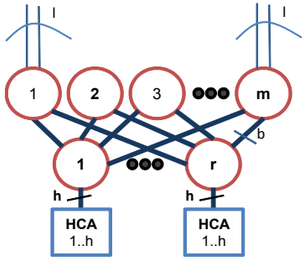
CLOS (Fat Tree)



Mesh / 3D Torus



Dragonfly+



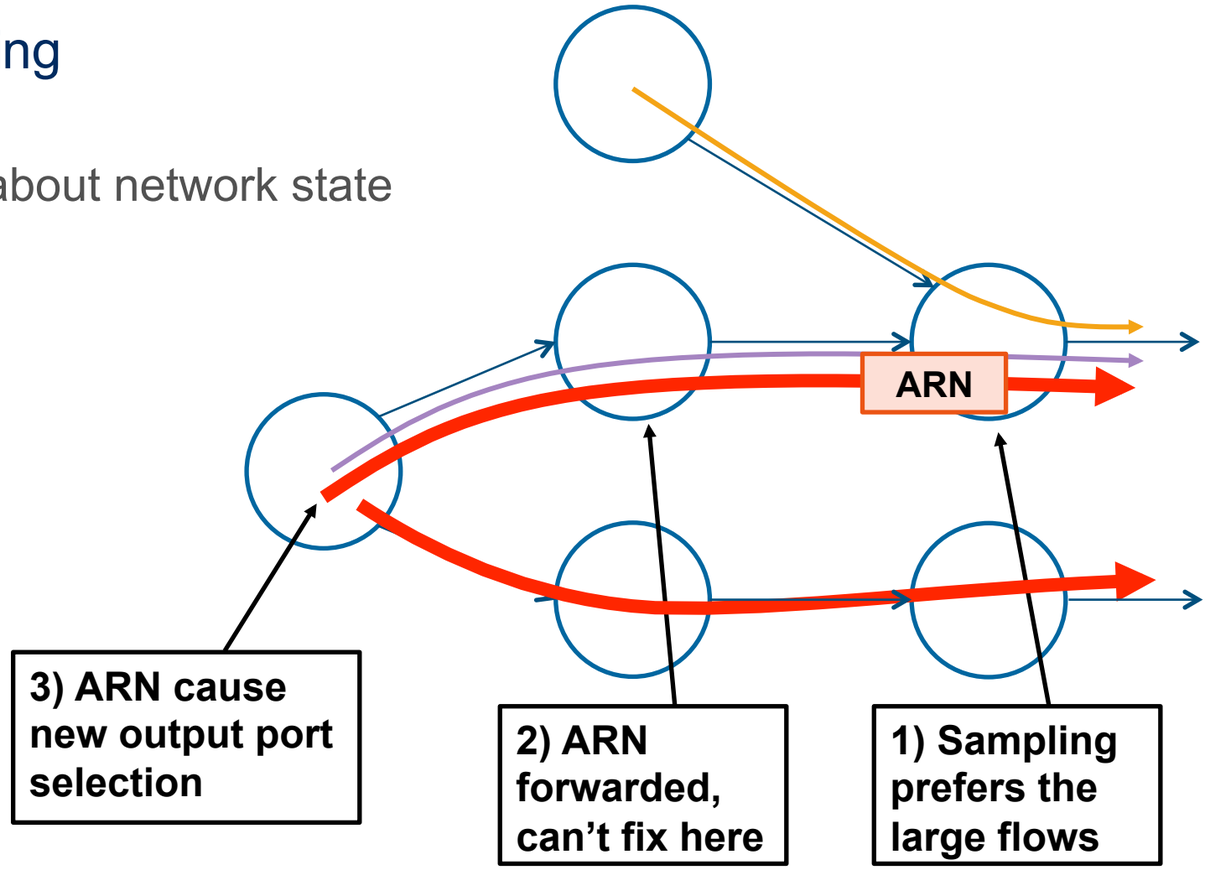
Is the Packet Allowed to Adapt?

- For every incoming packet the adaptive routing process has two main stages
  - Route Change Decision (to adapt or not to adapt)
  - New output port selection
- AR Modes
  - Static – traffic is always bound to a specific port
  - Time-Bound – traffic is bound to the last port used if not more than Tb [sec] passed since that event
  - Free – traffic may select a new out port freely

- Packets are classified to be either Legacy, Restricted or Unrestricted
- Destinations are classified to be either Legacy, Restricted, Timely-Restricted or Unrestricted
- A matrix maps possible combinations of packet and destination based classification to AR modes

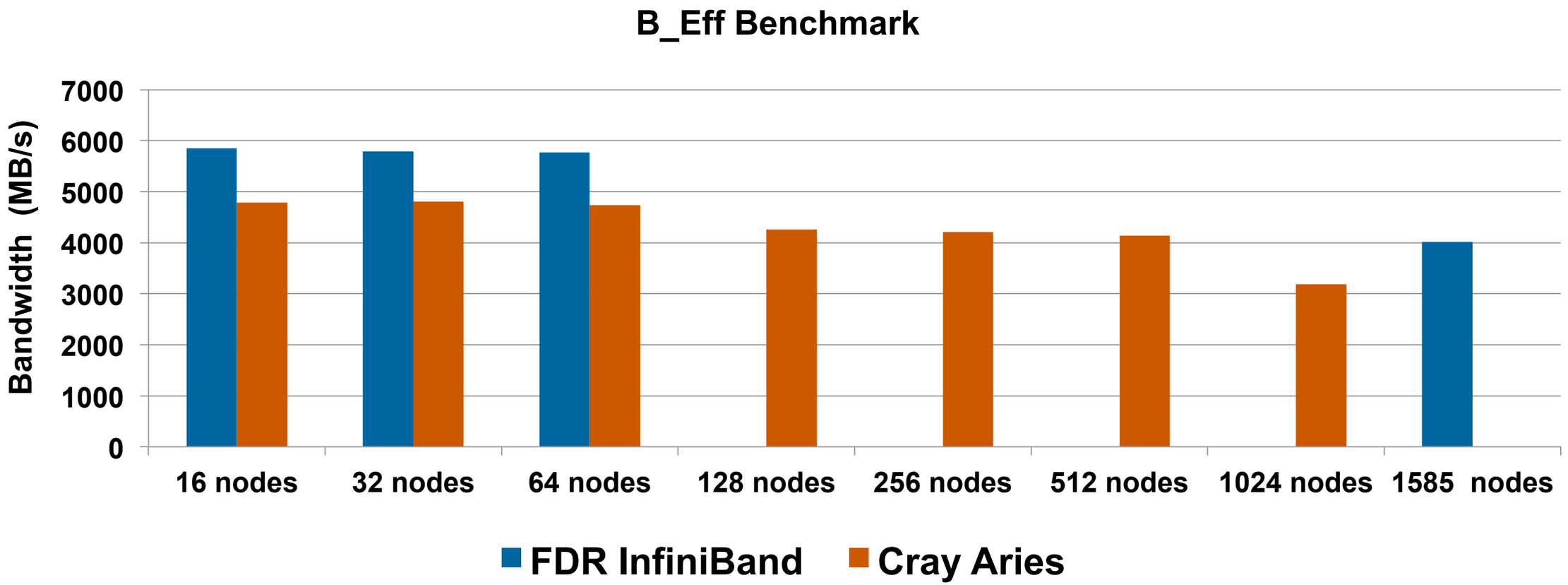
Mellanox Adaptive Routing Notification (ARN)

- The “reaction” time is critical to Adaptive Routing
  - Traffic modes change fast
  - A “better” AR decision requires some knowledge about network state
- Internal switch to switch communications
- Faster convergence after routing changes
- Fast notification to decision point
- Fully configurable (topology agnostic)



Faster Routing Modifications, Resilient Network

InfiniBand Adaptive Routing



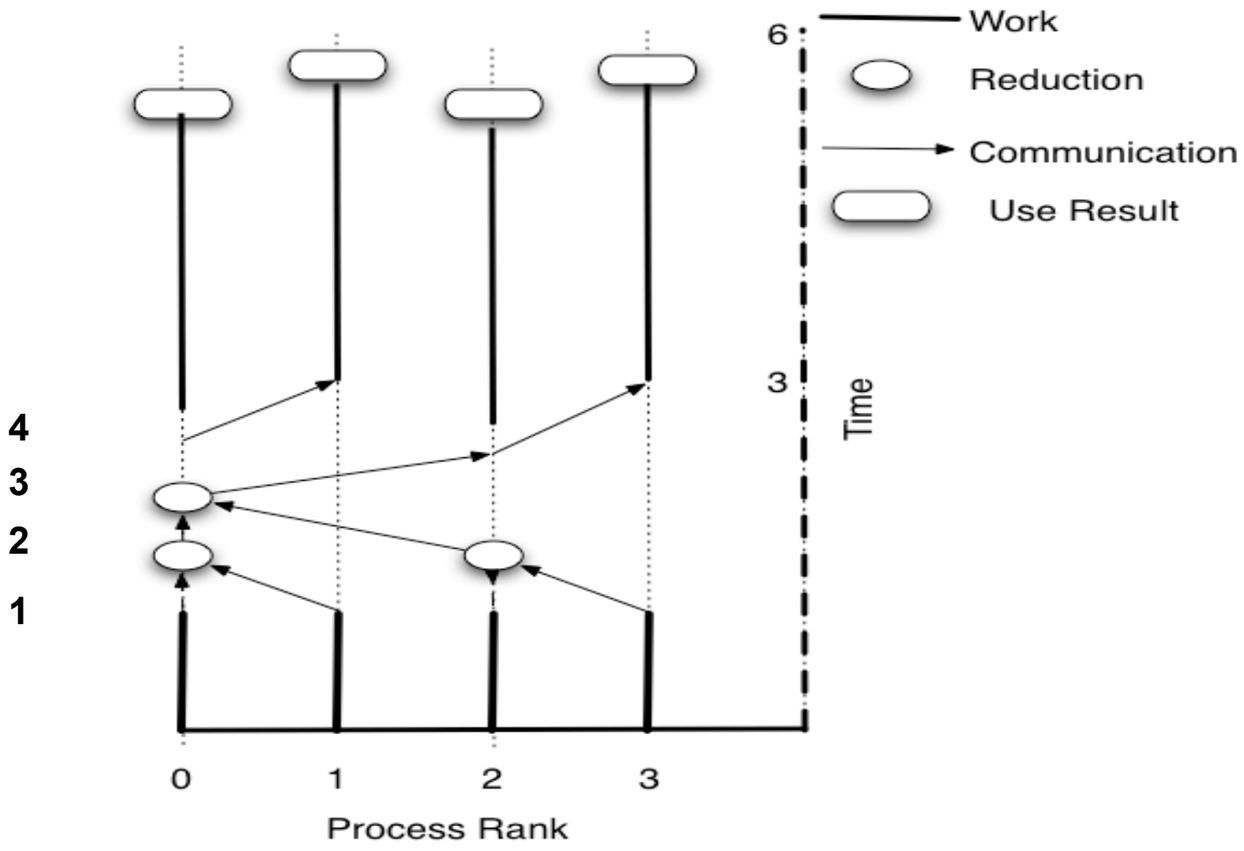
Higher Performance and Better Network Utilization

Network Offload

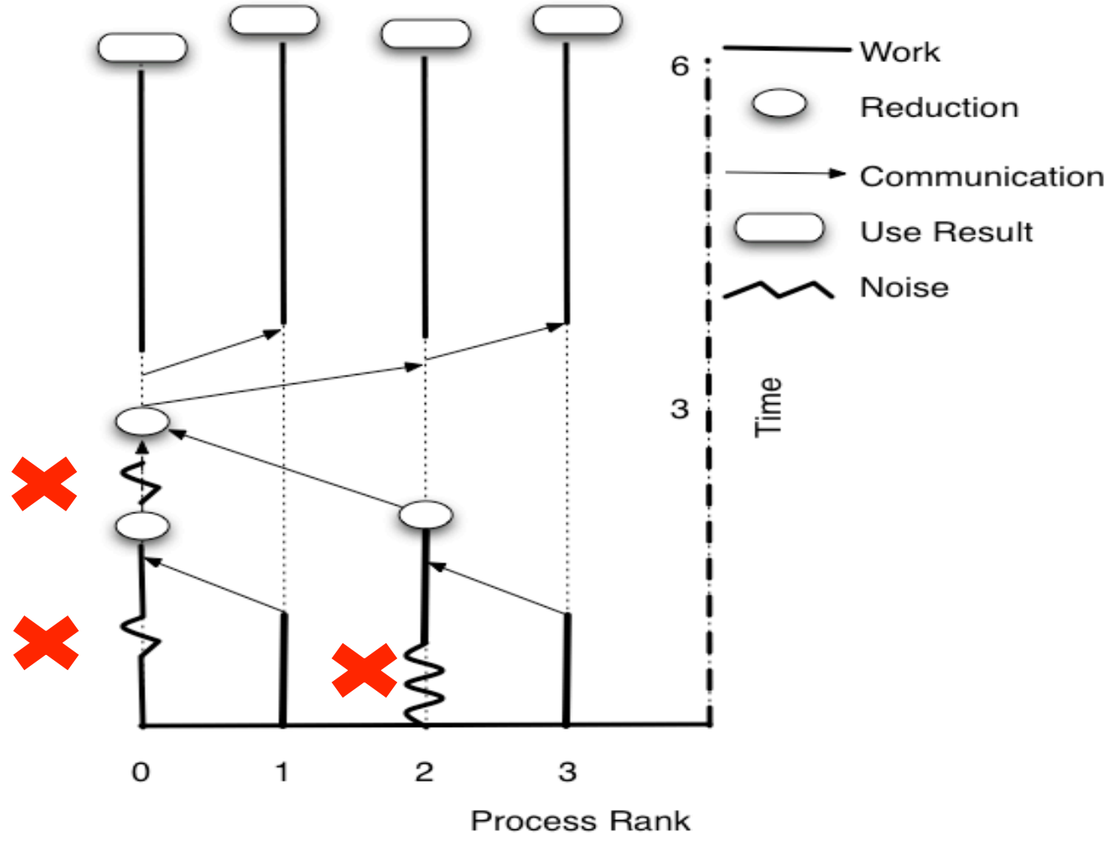
Cross Channel Synchronization

Scalability of Collective Operations

Ideal Algorithm

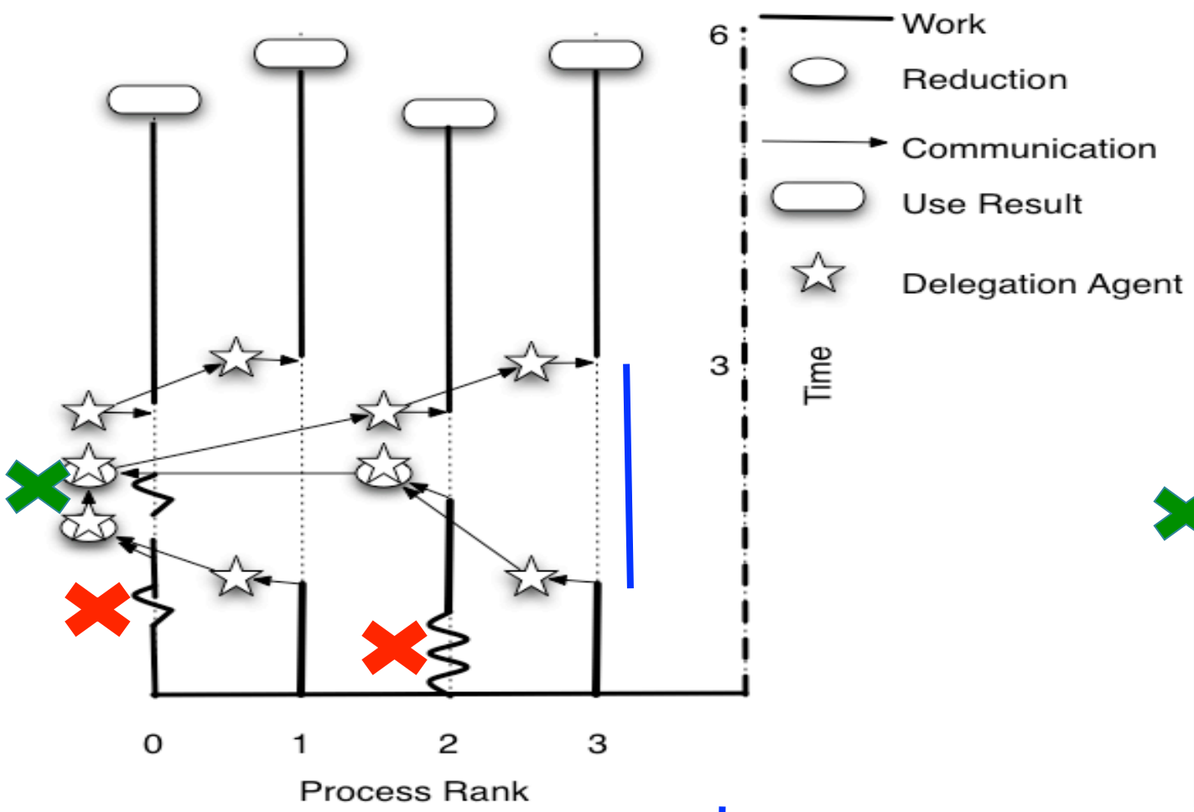


Impact of System Noise

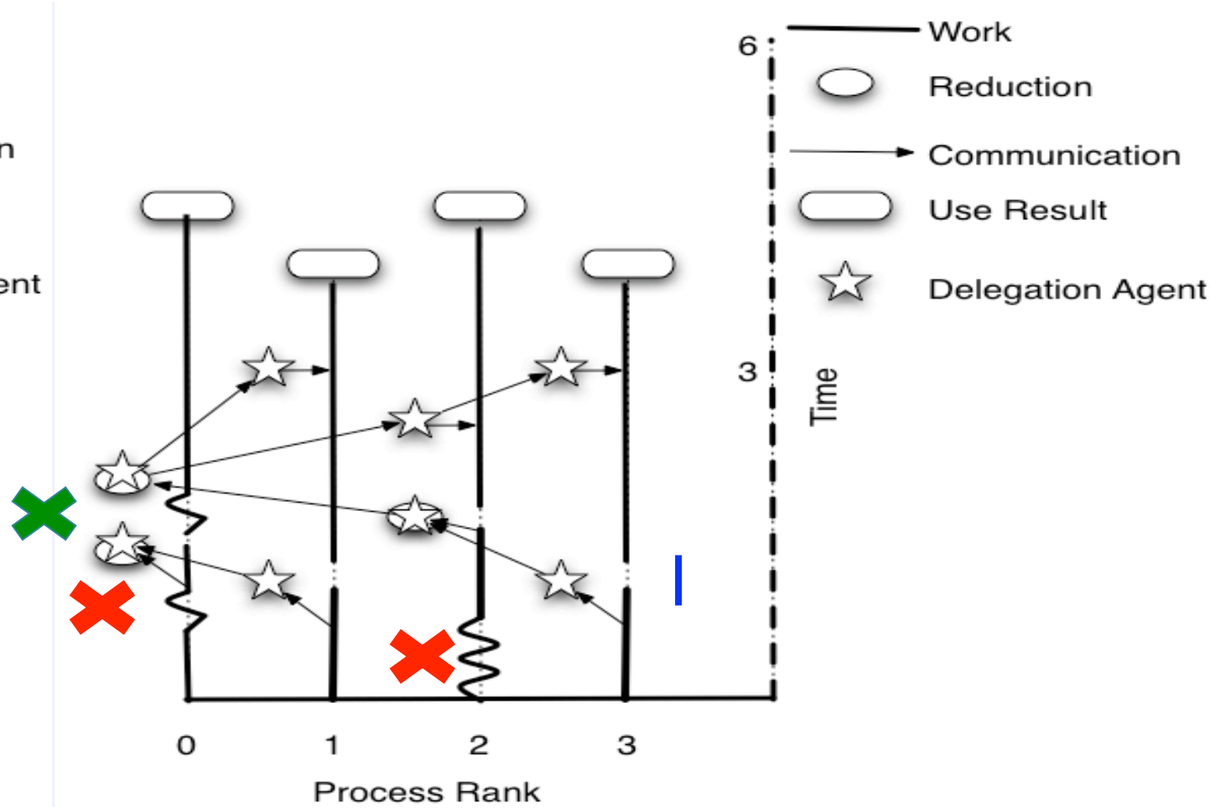


Scalability of Collective Operations - II

Offloaded Algorithm



Nonblocking Algorithm

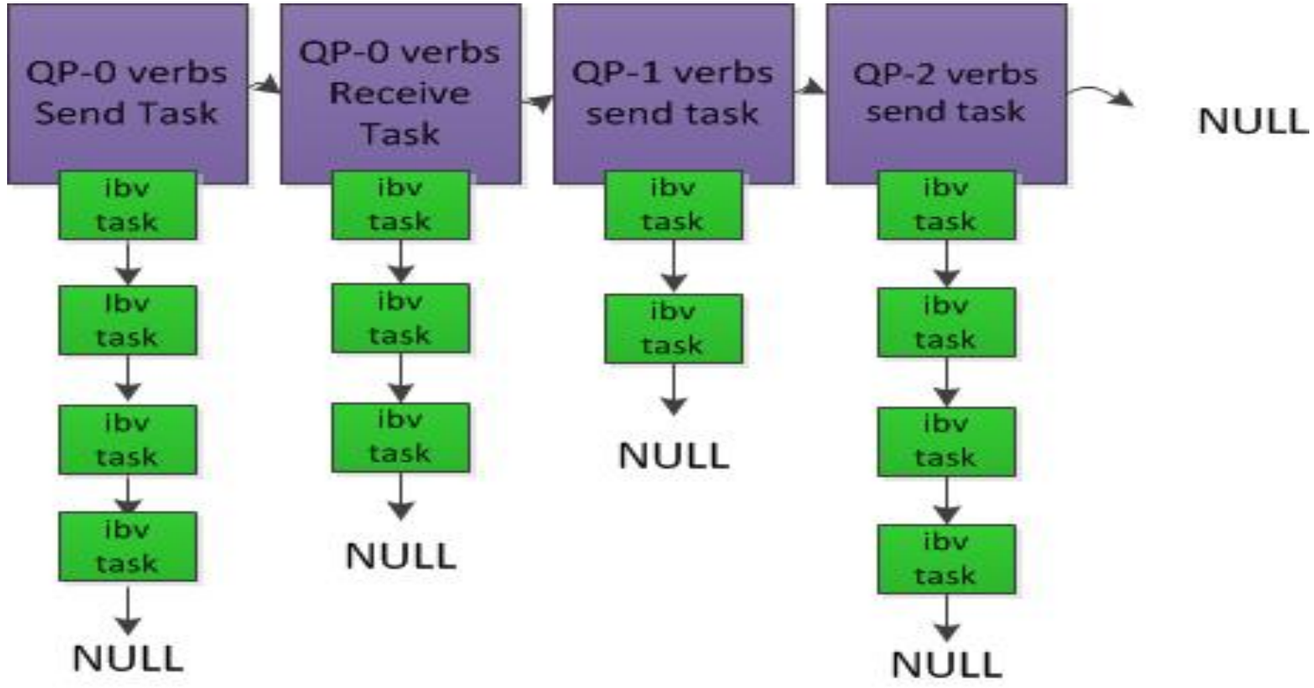


- Communication processing

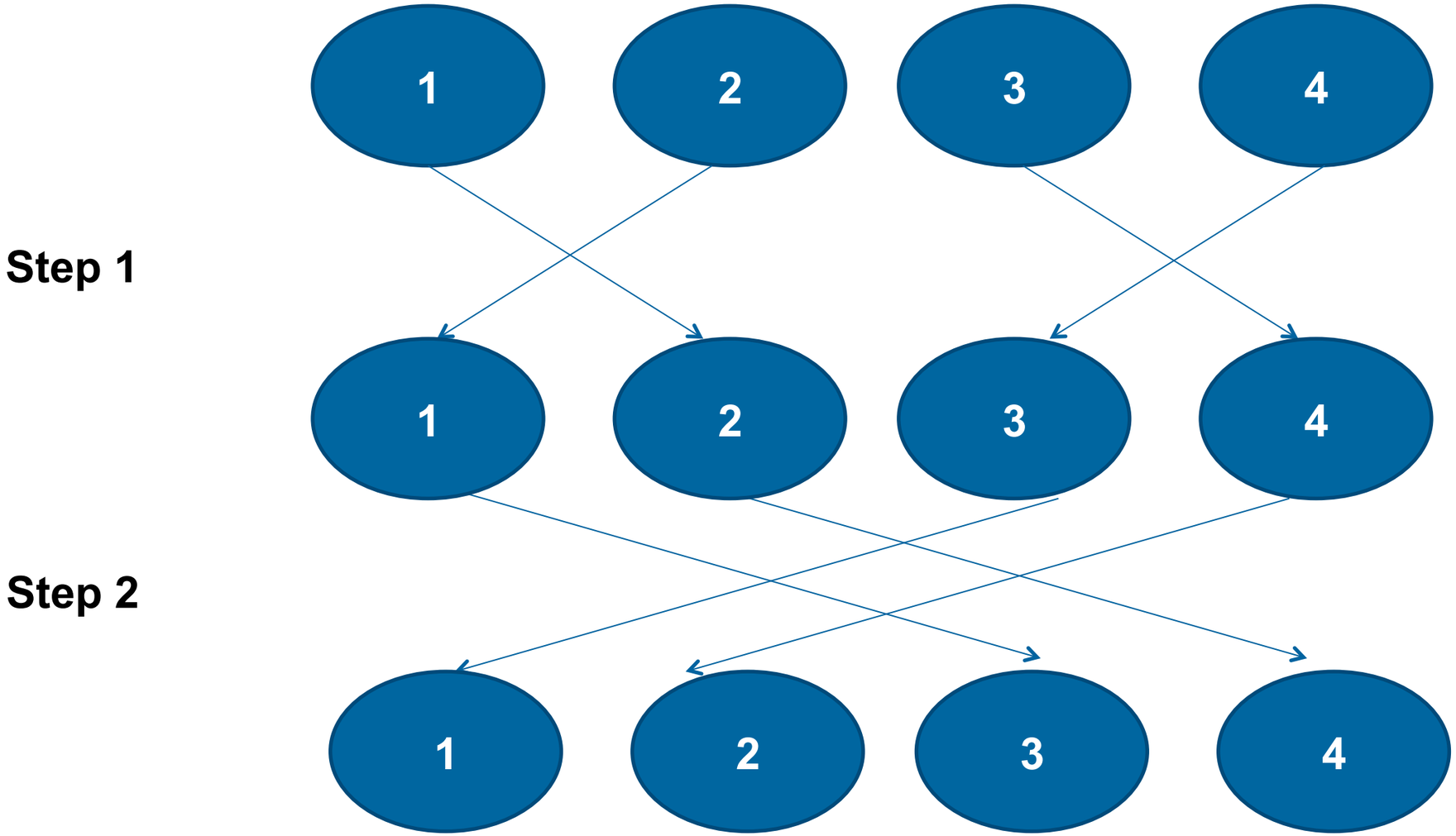
Cross Channel Synchronization (aka CORE-Direct)

- Scalable collective communication
- Asynchronous communication
- Manage communication by communication resources
- Avoid system noise

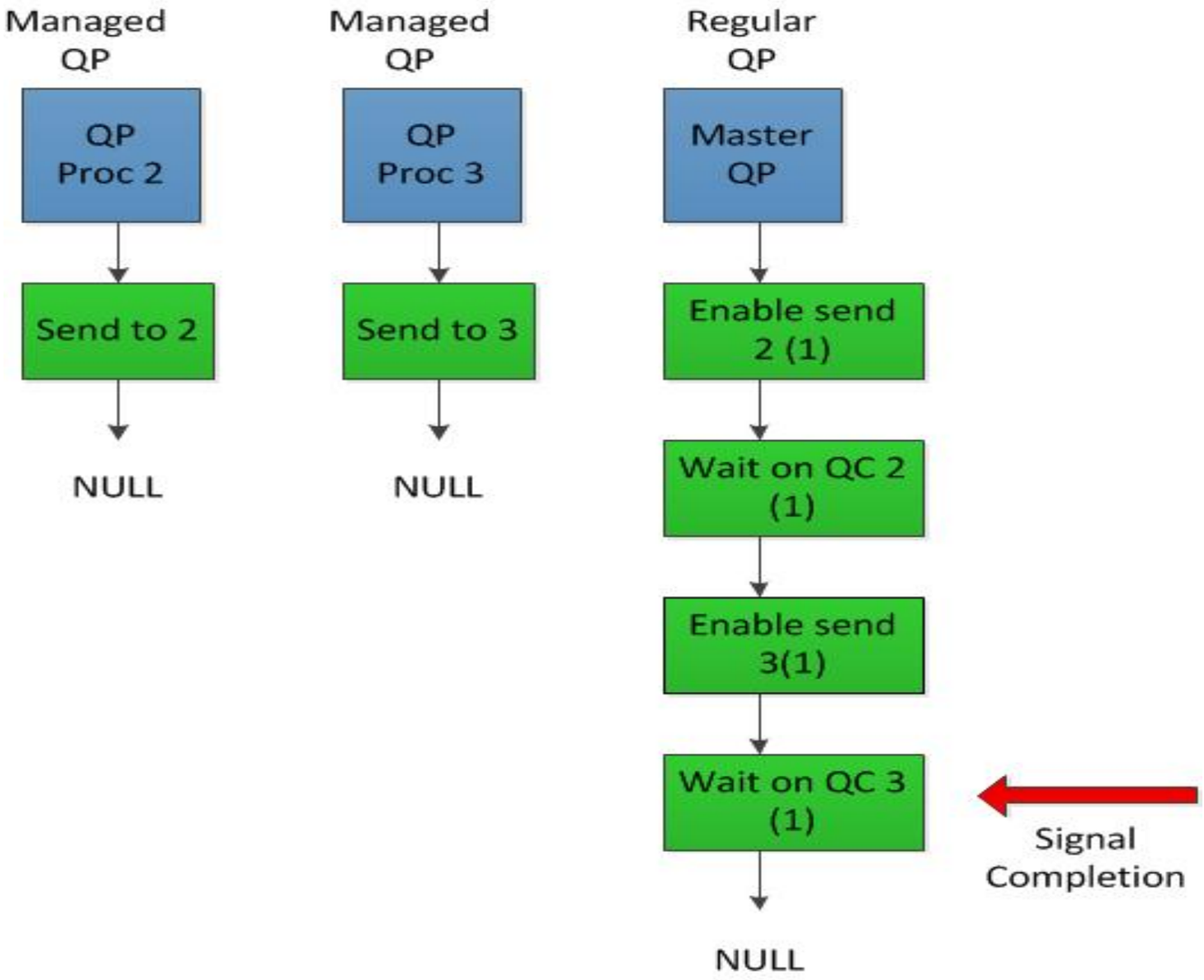
- Task list
- Target QP for task
- Operation
  - Send
  - Wait for completions
  - Enable
  - Calculate



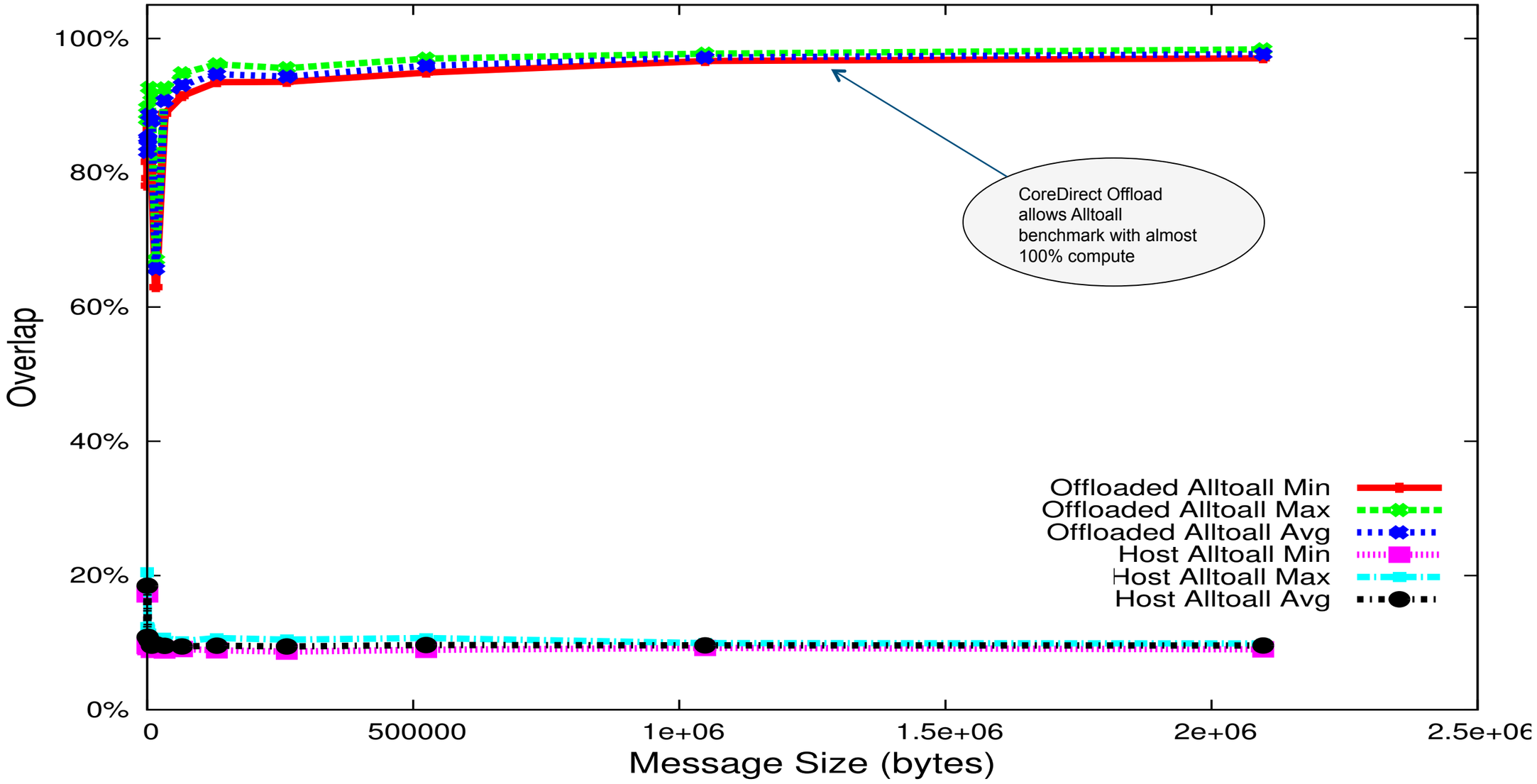
Example – Four Process Recursive Doubling



Four Process Barrier Example – Using Managed Queues – Rank 0



Nonblocking Alltoall (Overlap-Wait) Benchmark



Non-Contiguous Data

Optimizing Non Contiguous Memory Transfers

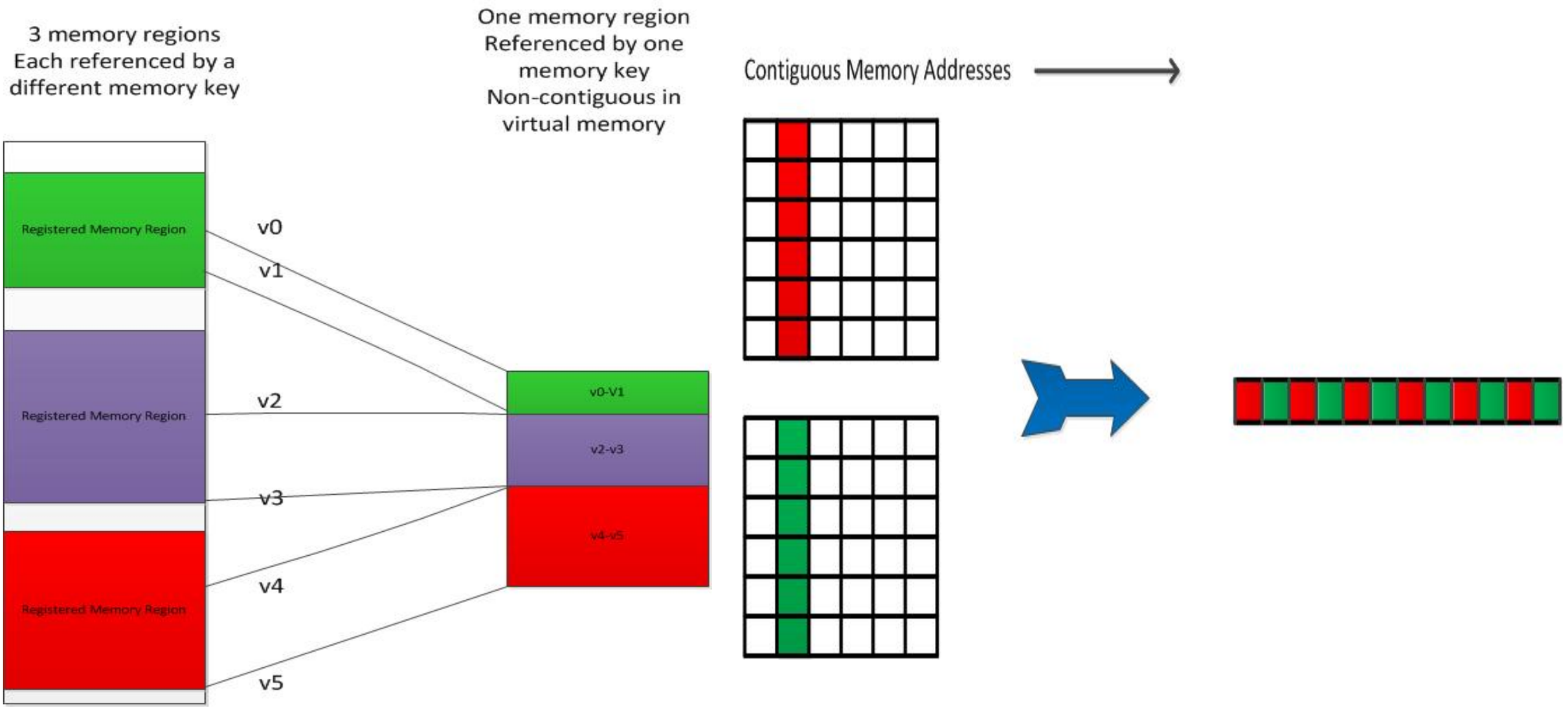
- Support combining contiguous registered memory regions into a single memory region. H/W treats them as a single contiguous region (and handles the non-contiguous regions)

- For a given memory region, supports non-contiguous access to memory, using a regular structure representation – base pointer, element length, stride, repeat count.
  - Can combine these from multiple different memory keys

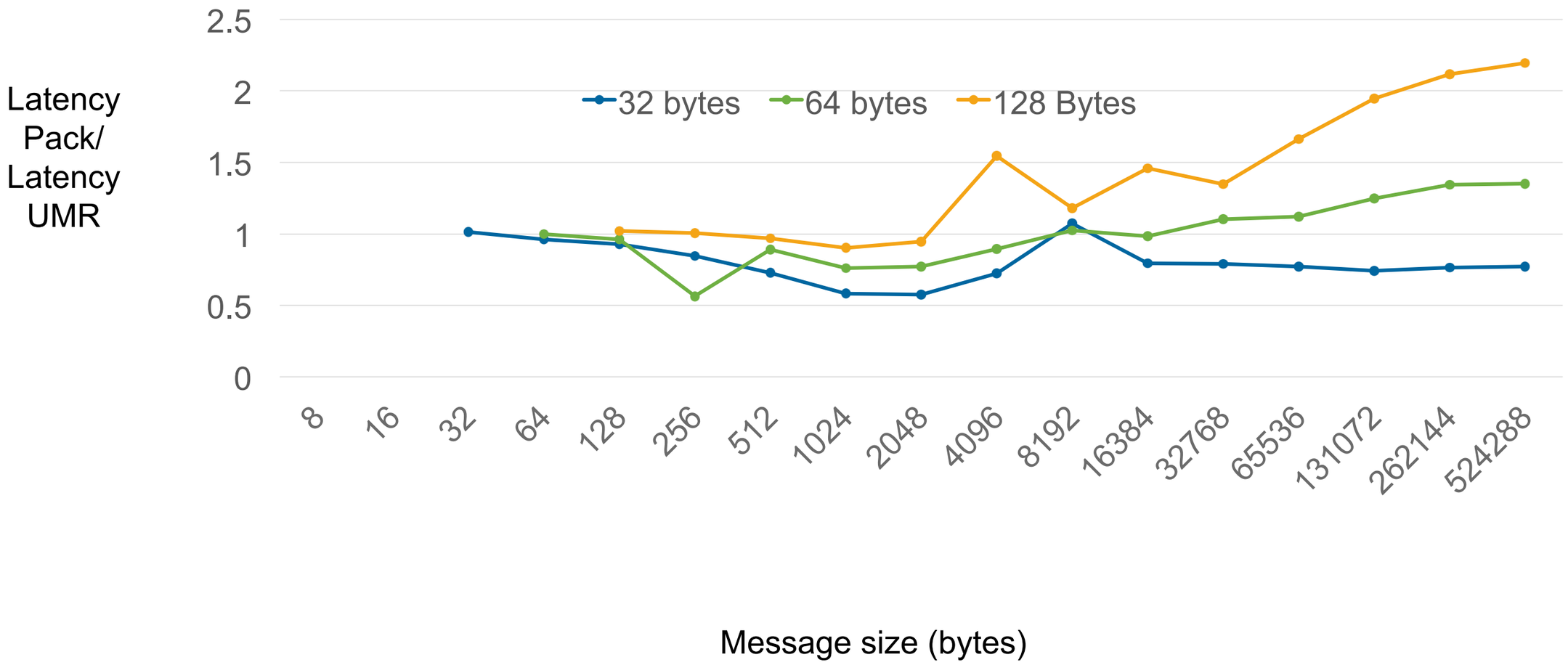
- Memory descriptors are created by posting WQE's to fill in the memory key

- Supports local and remote non-contiguous memory access
  - Eliminates the need for some memory copies

Optimizing Non Contiguous Memory Transfers



Hardware Gather/Scatter Capabilities – Regular Structure – Ping-Pong latency



New Effort – Application Optimization

- Starting up effort to work on improving application performance
  - In house application domain experts
  - In house performance optimization experts
  - Looking for interested partners

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™