Scalable HPC Communication Capabilities

Rich Graham

HPC 2014, Cetraro, Italy

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

HPC
The Challenges

Exascale-Class Computer Platforms – Communication Challenges

Challenge	Solution focus
Very large functional unit count ~10M	Scalable communication capabilities: point-to-point & collectives
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

Challenges of Scalable HPC Communication Libraries

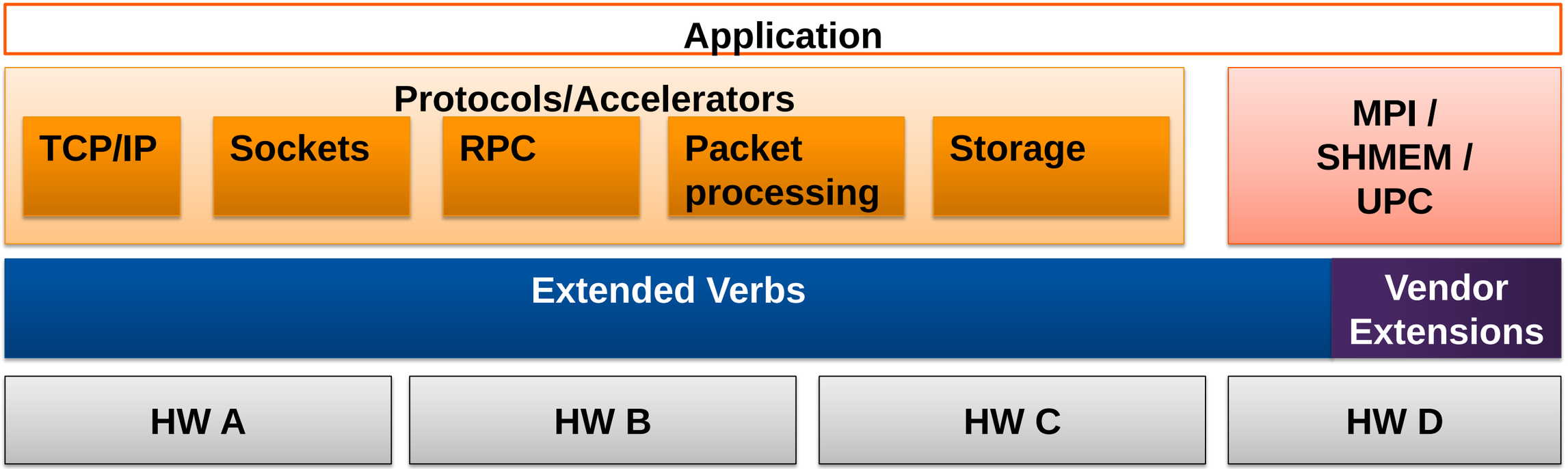
- Truly asynchronous communication capabilities
  - Two-sided point-to-point
  - One-sided point-to-point
  - Collective Communication
- Truly one-sided remote progress
- System noise mitigation
- Topology awareness

Standardize your Interconnect

Standard !

- Standardized wire protocol
  - Mix and match of vertical and horizontal components
  - Ecosystem – build together
- Open-source, extensible interfaces
  - Extend and optimize per applications needs

Extensible Open Source Framework



Speed, Speed, Speed

Technology Roadmap – Always One-Generation Ahead

Mellanox → 20Gbs → 40Gbs → 56Gbs → 100Gbs → 200Gbs

Terascale

3rd

TOP500 2003
Virginia Tech (Apple)

Petascale

1st

“Roadrunner”
Mellanox Connected

Exascale

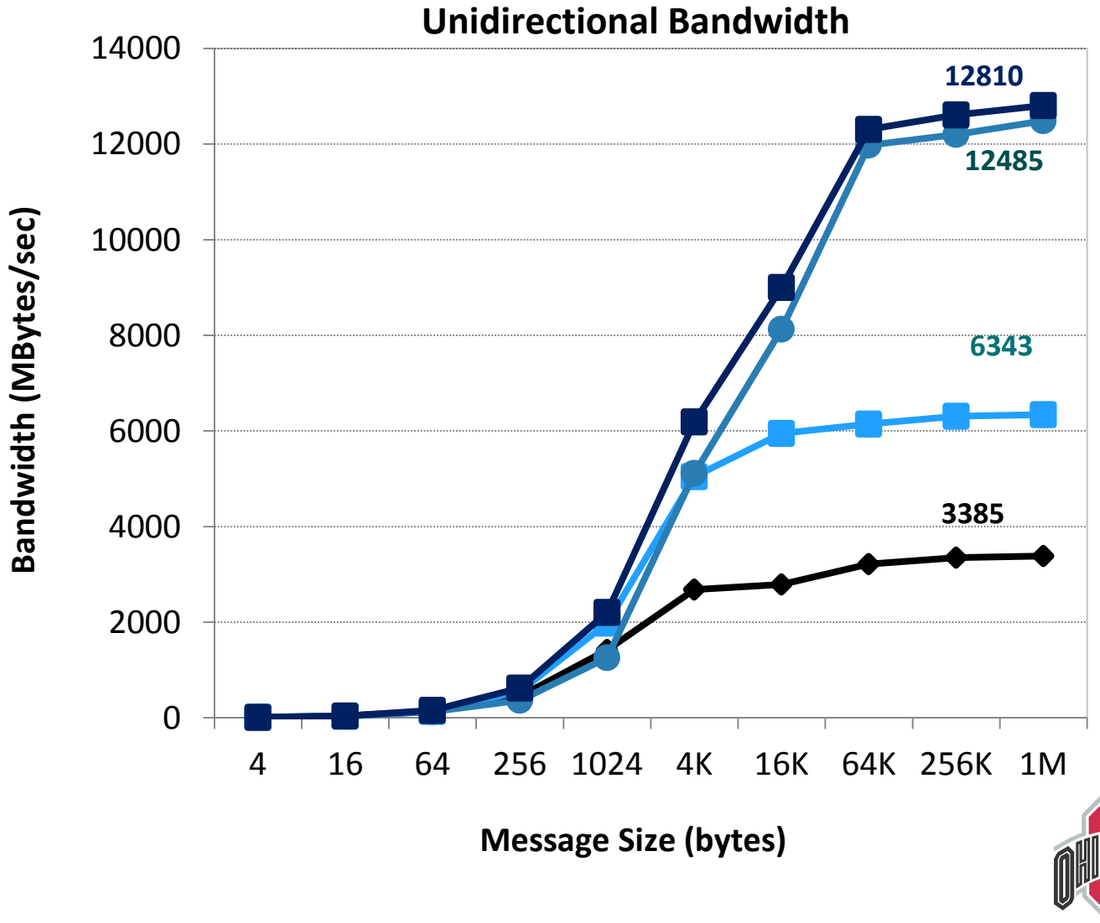
2000 2005 2010 2015 2020

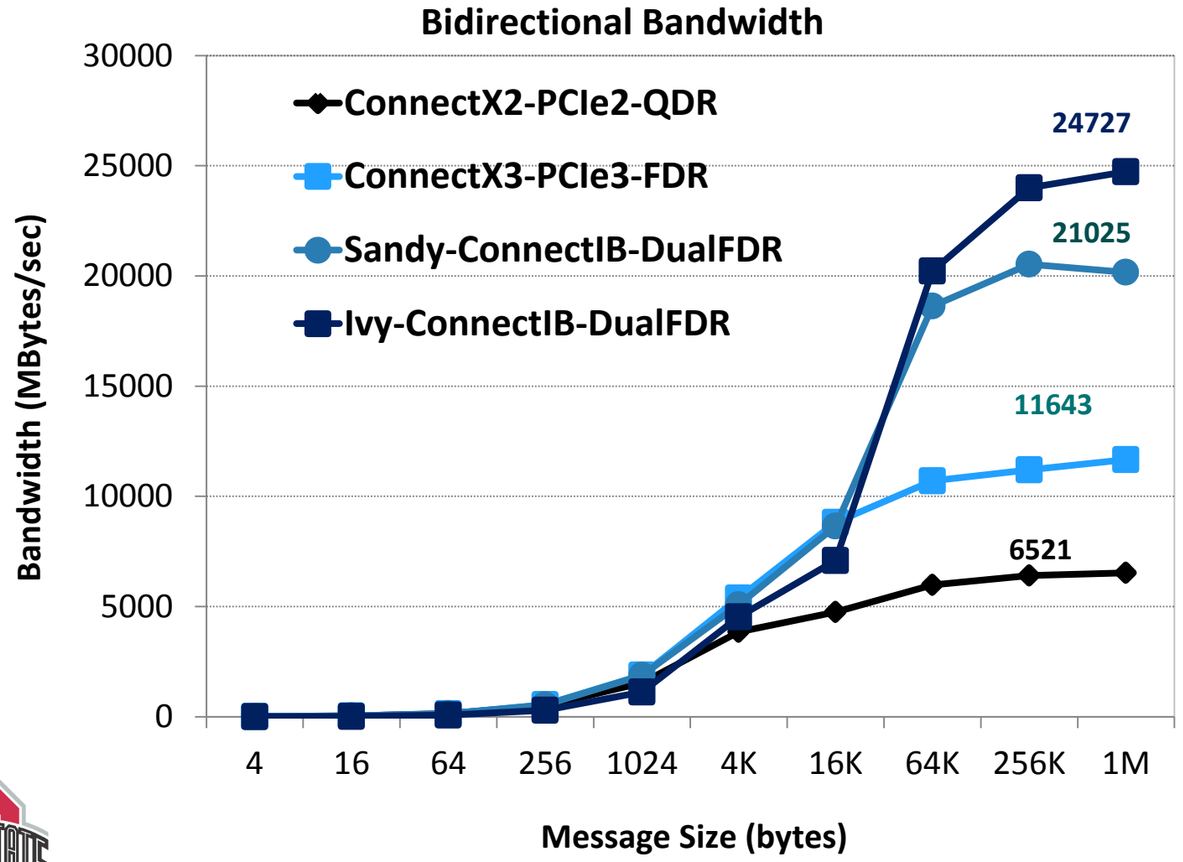
Mellanox Connect-IB The World's Fastest Adapter

- The 7th generation of Mellanox interconnect adapters
- World's first 100Gb/s interconnect adapter (dual-port FDR 56Gb/s InfiniBand)
- Delivers 137 million messages per second – 4X higher than competition
- Support the new innovative InfiniBand scalable transport – Dynamically Connected

Connect-IB Provides Highest Interconnect Throughput

Higher is Better

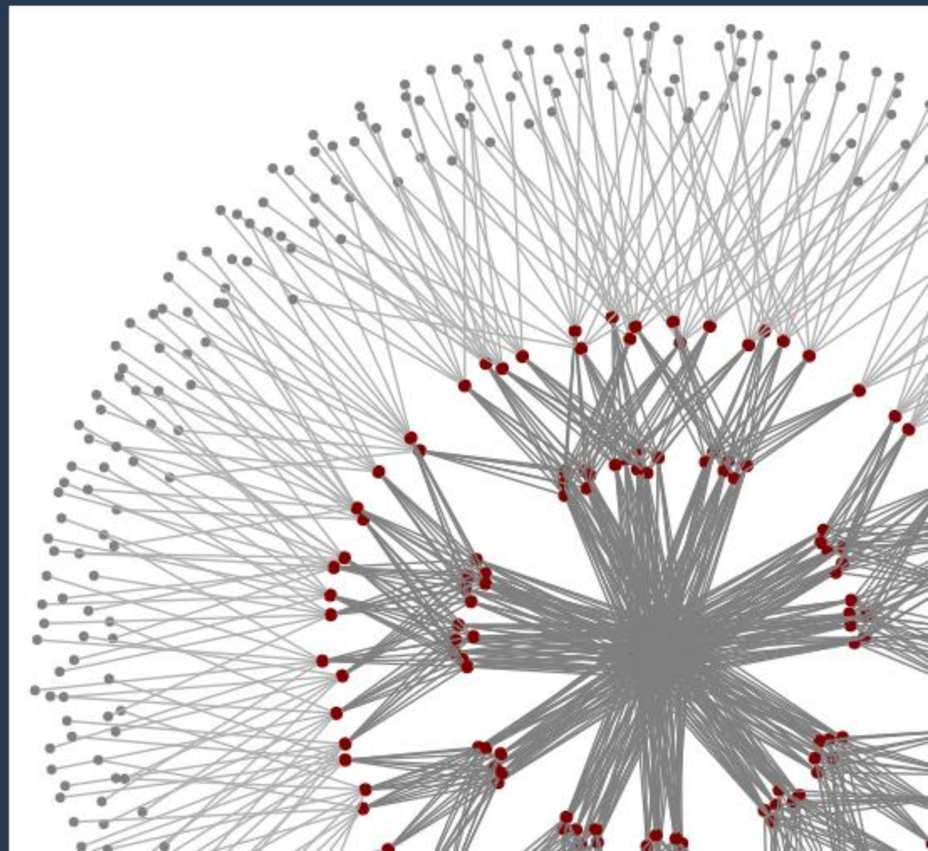




Source: Prof. DK Panda

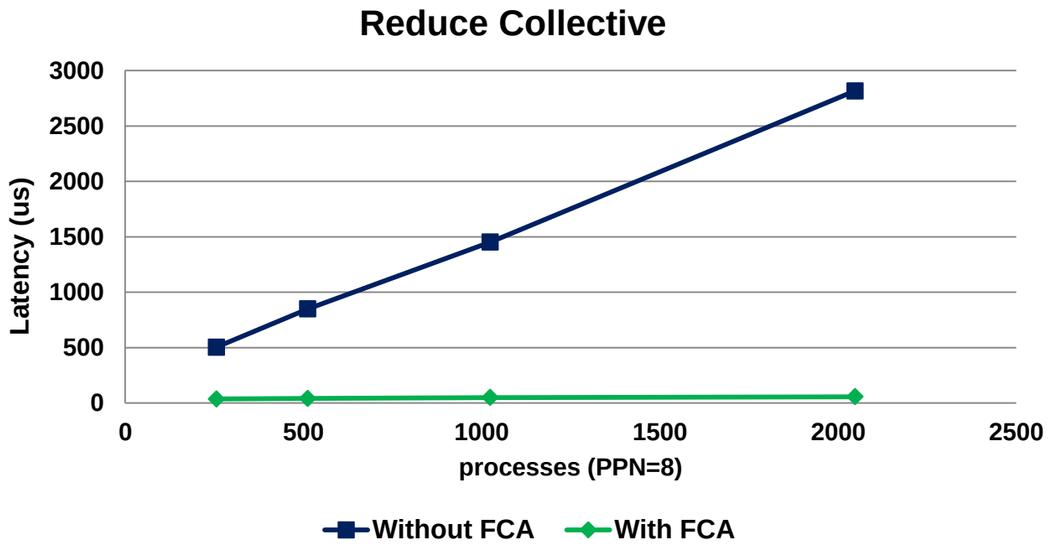
Gain Your Performance Leadership With Connect-IB Adapters

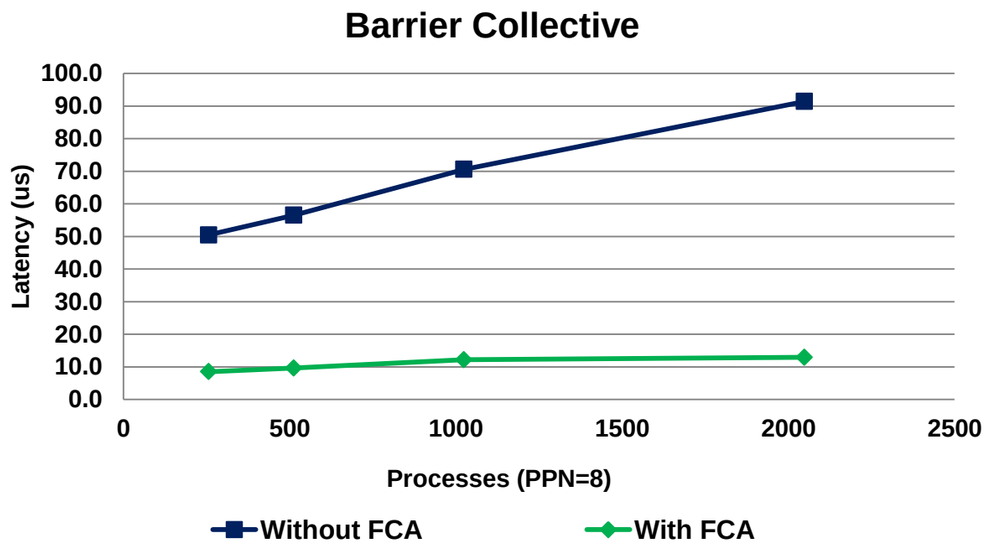
Scalability



Collective Operations

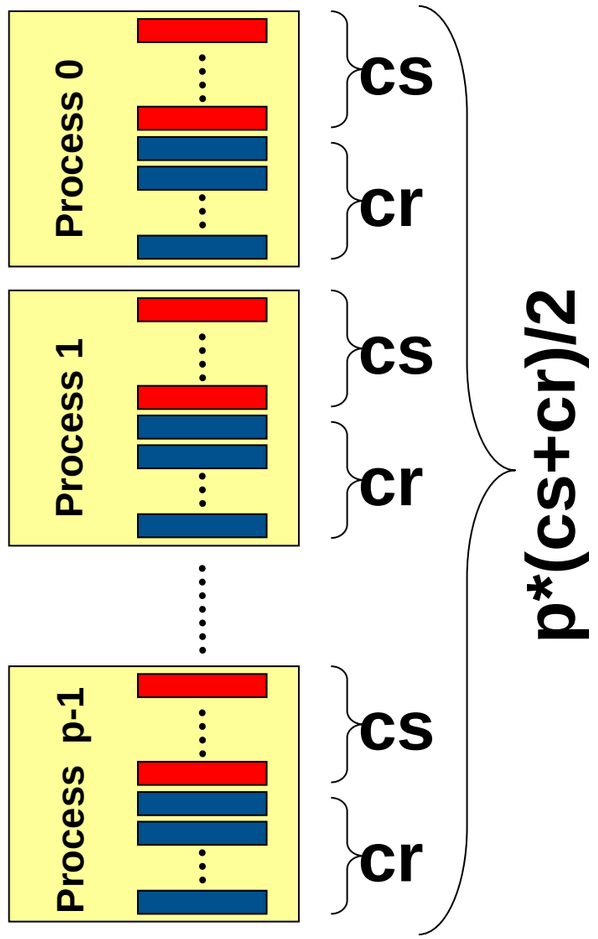
- Topology Aware
- Hardware Multicast
- Separate Virtual Fabric
- Offload
- Scalable algorithms





The DC Model

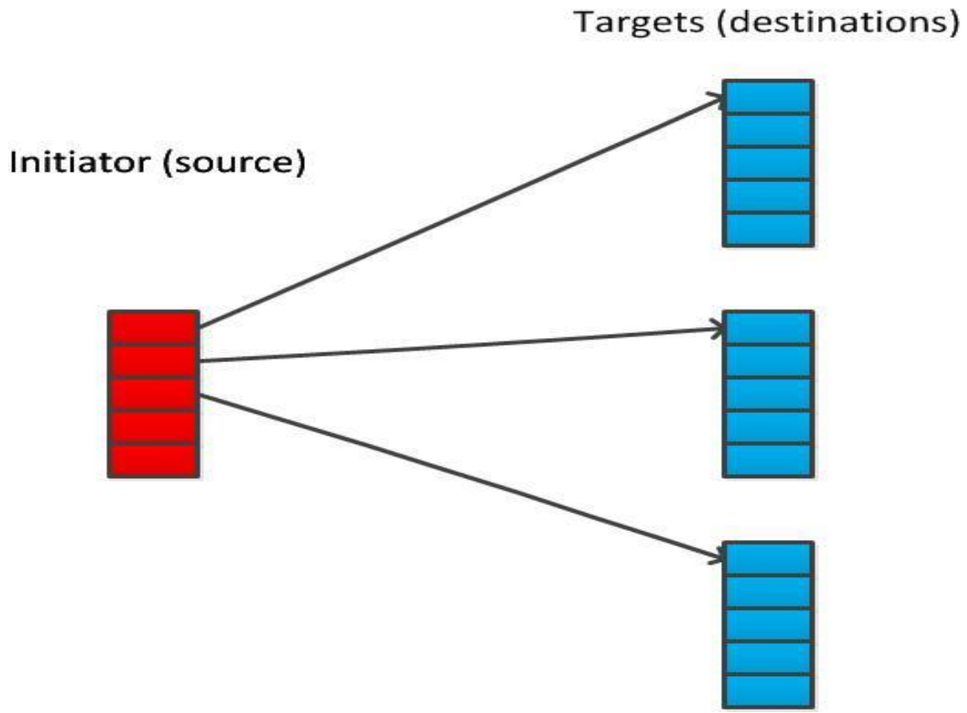
- Dynamic Connectivity
- Each DC Initiator can be used to reach any remote DC Target
- No resources' sharing between processes
  - process controls how many (and can adapt to load)
  - process controls usage model (e.g. SQ allocation policy)
  - no inter-process dependencies
- Resource footprint
  -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



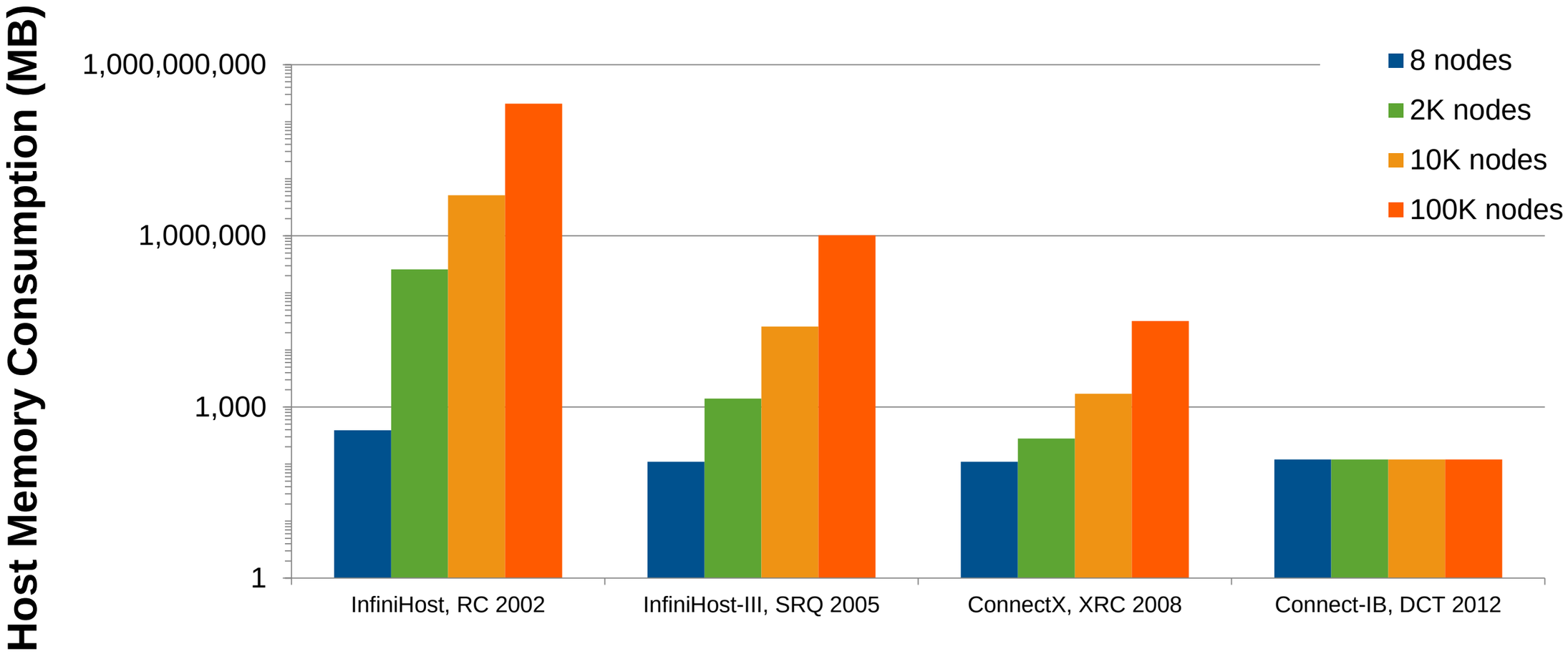
cs – concurrency of the sender
cr=concurrency of the responder

Dynamically Connected Transport

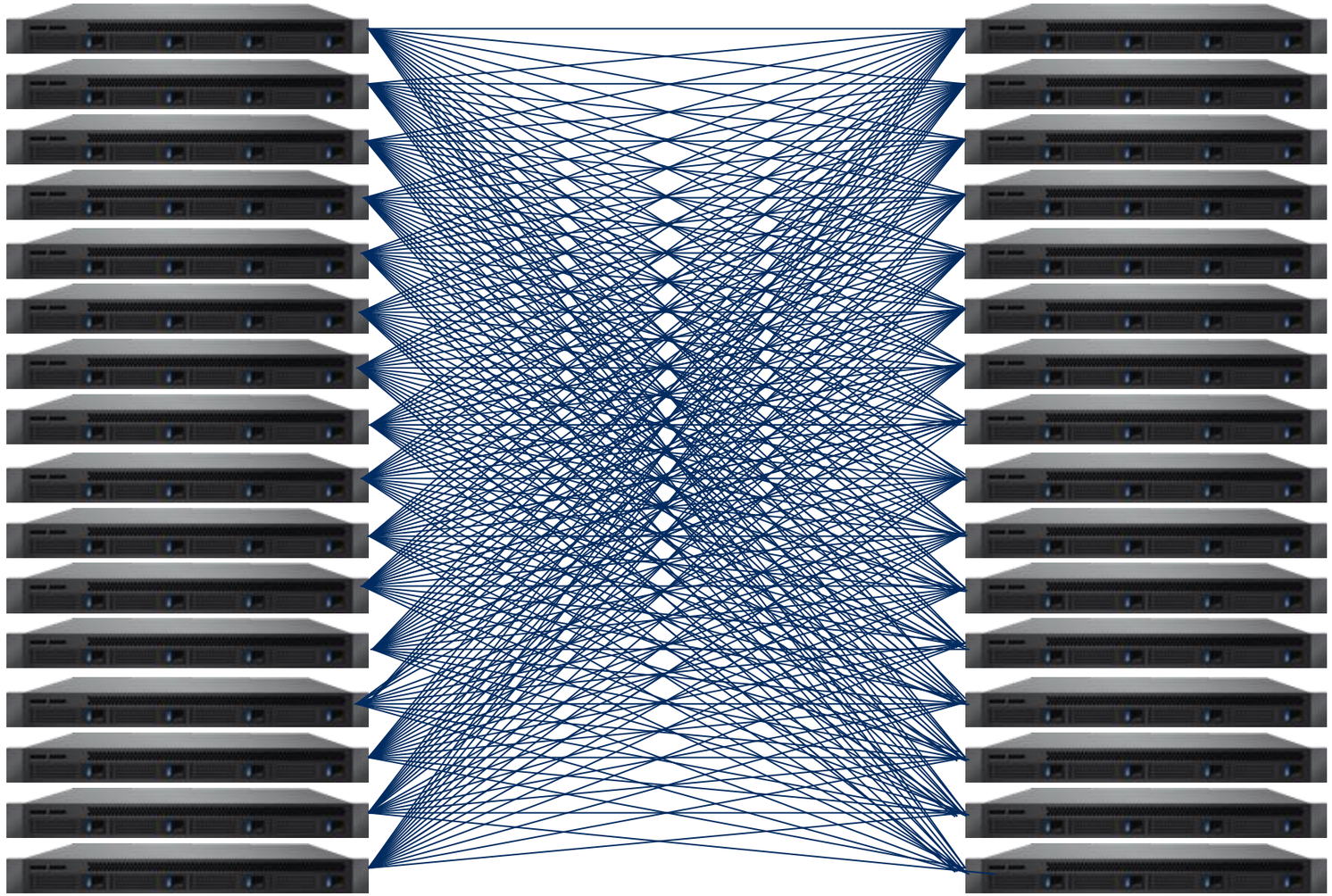
- Key objects
  - DC Initiator: Initiates data transfer
  - DC Target: Handles incoming data



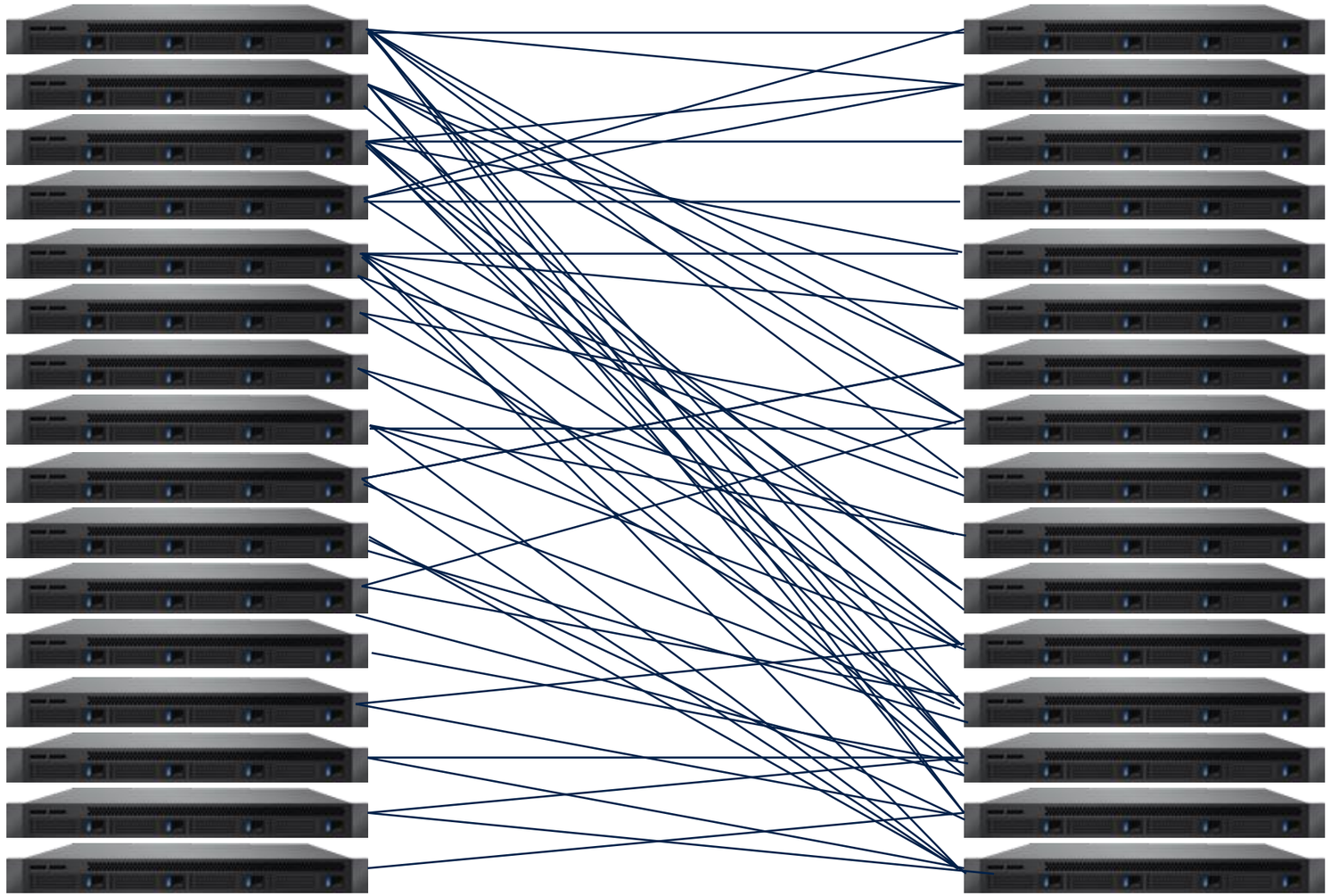
Dynamically Connected Transport – Exascale Scalability



Reliable Connection Transport Mode

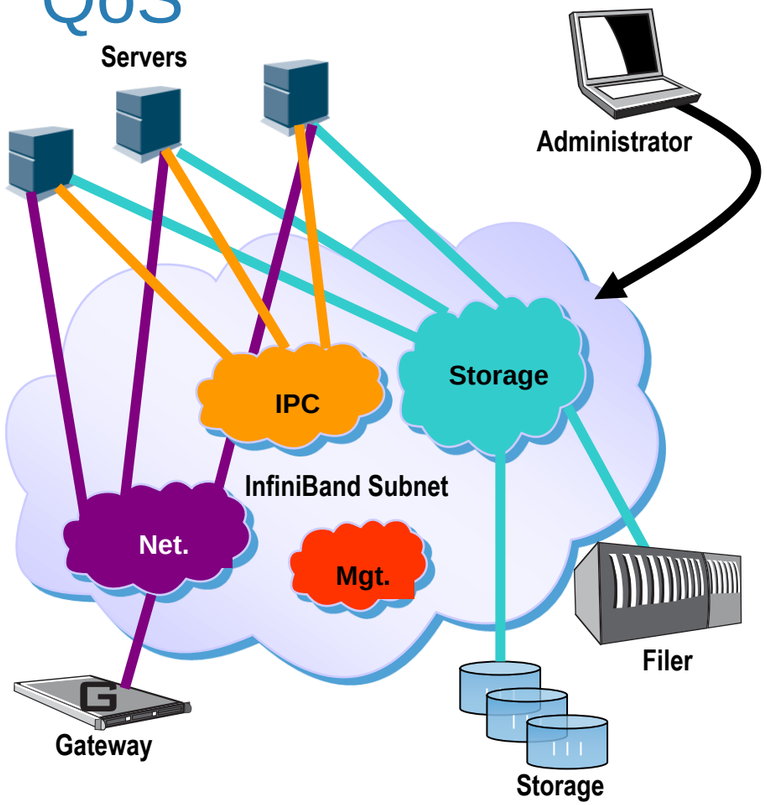


Dynamically Connected Transport Mode



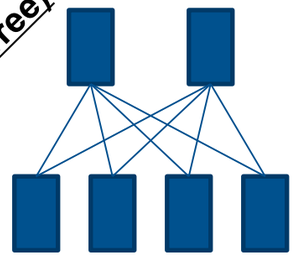
More…

QoS

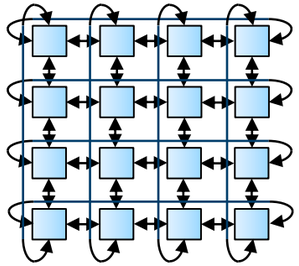


Topologies

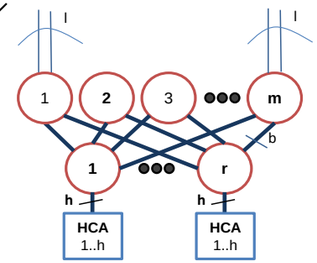
CLOS (Fat Tree)



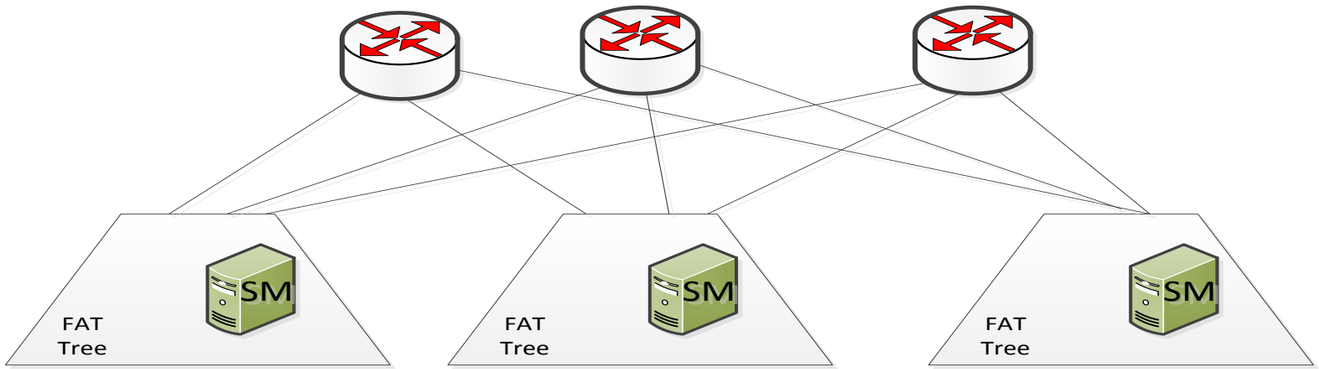
Mesh / 3D Torus



Dragonfly+



Routers

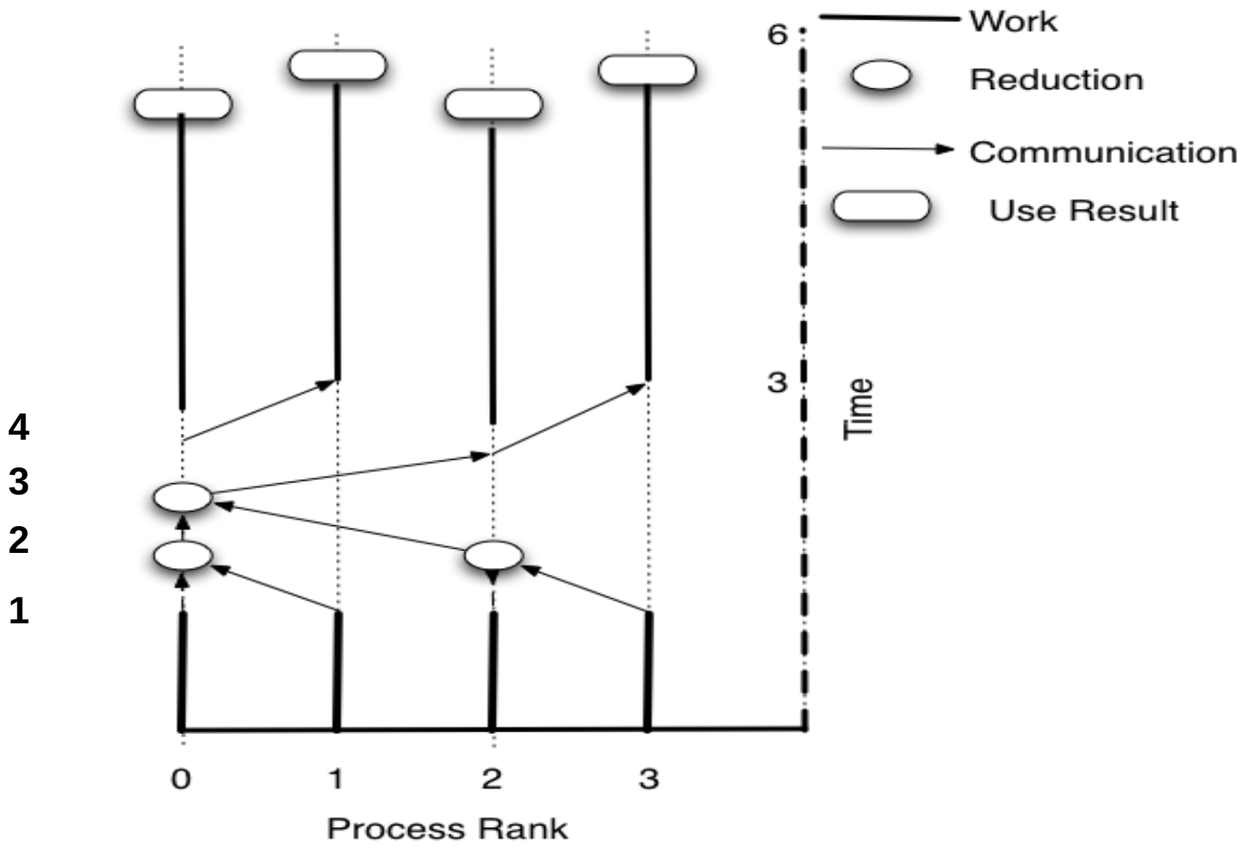


Distance

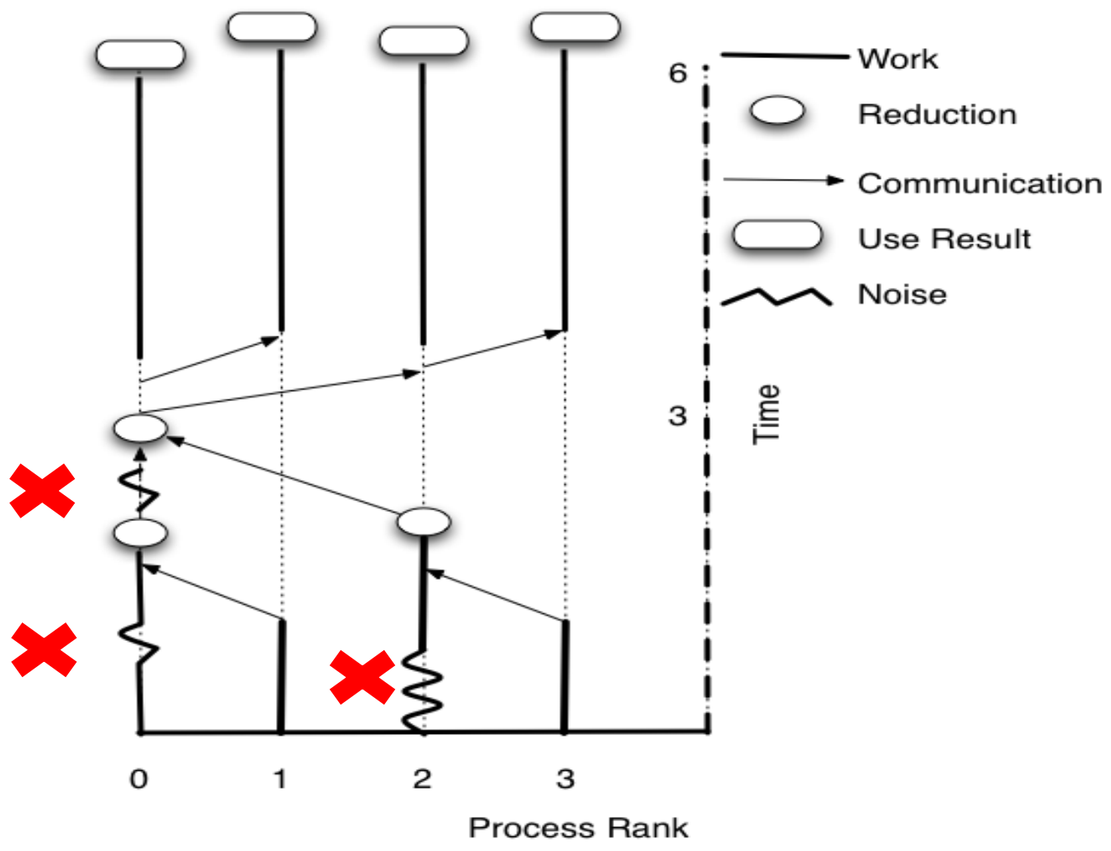
Network Offload

Scalability of Collective Operations

Ideal Algorithm

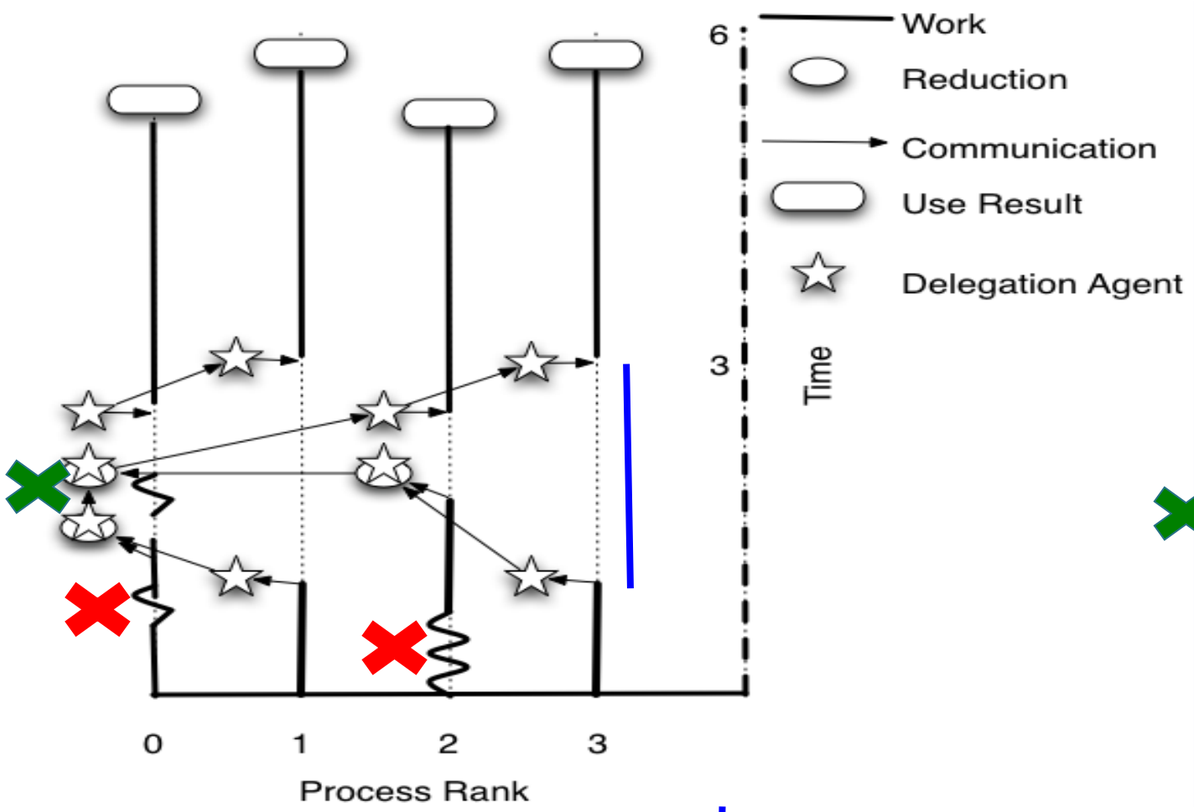


Impact of System Noise

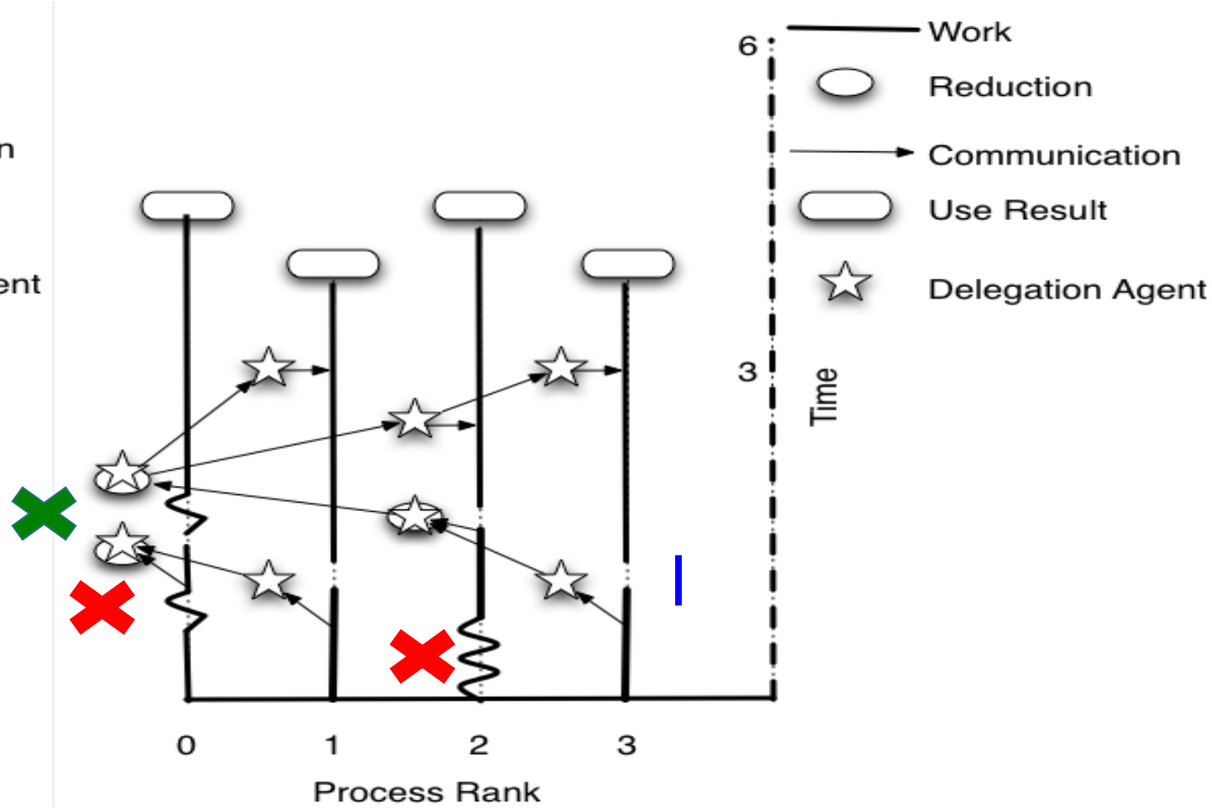


Scalability of Collective Operations - II

Offloaded Algorithm



Nonblocking Algorithm

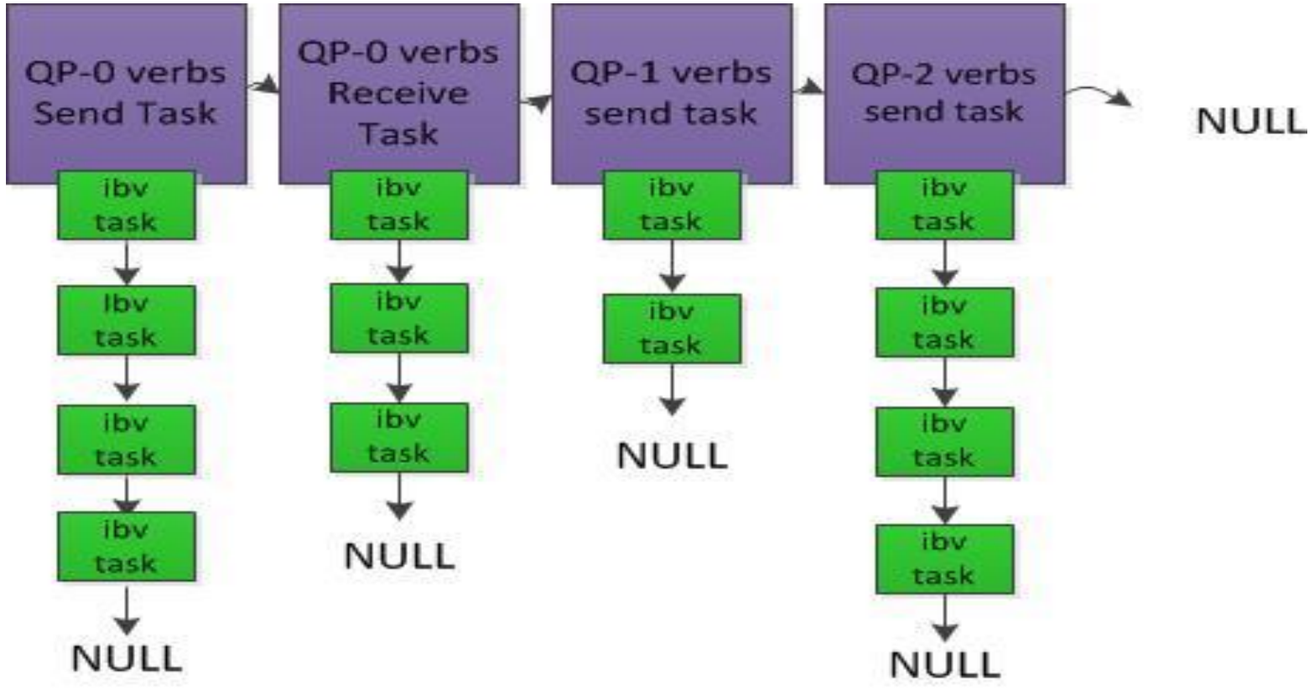


- Communication processing

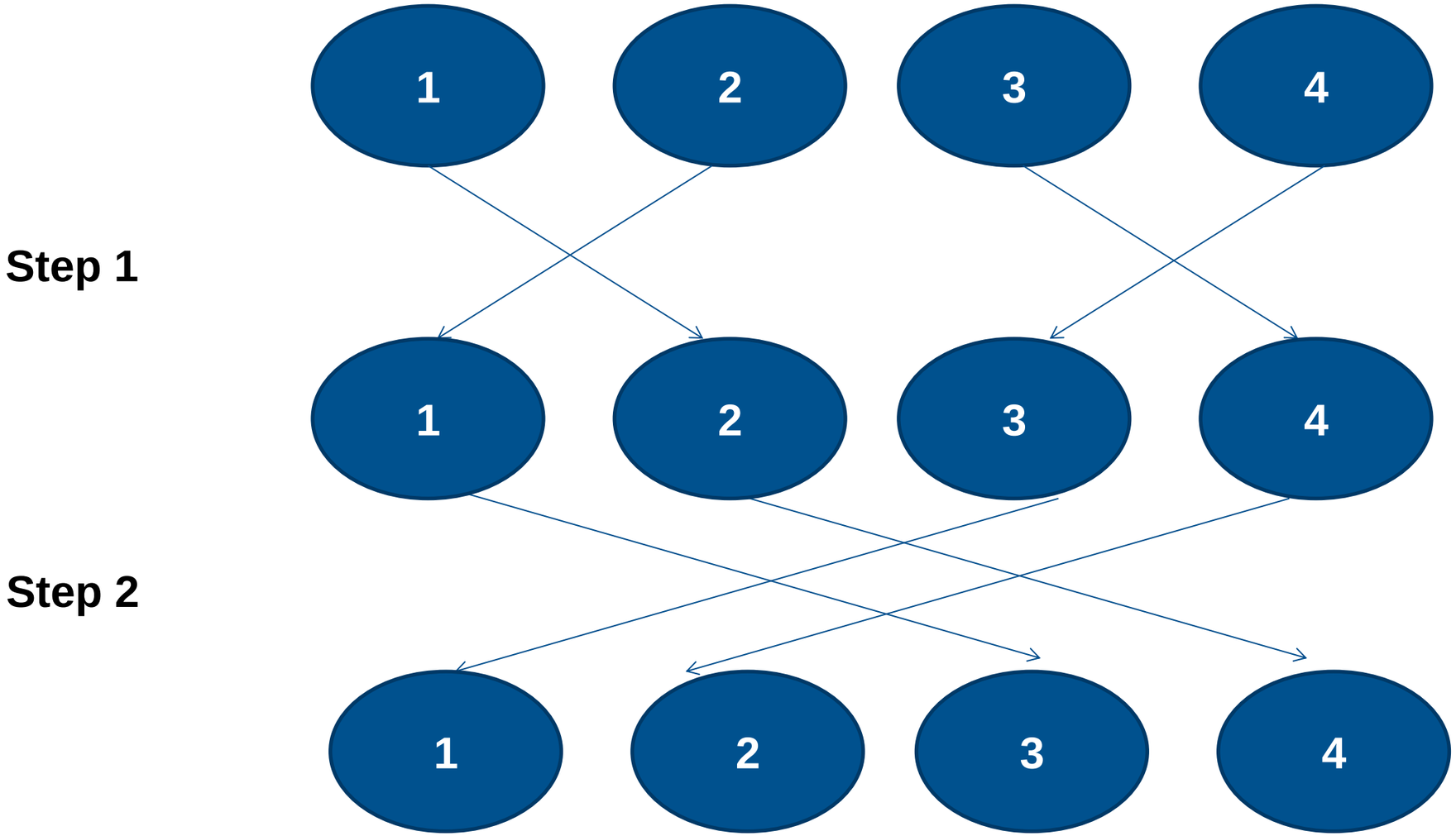
CORE-Direct

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

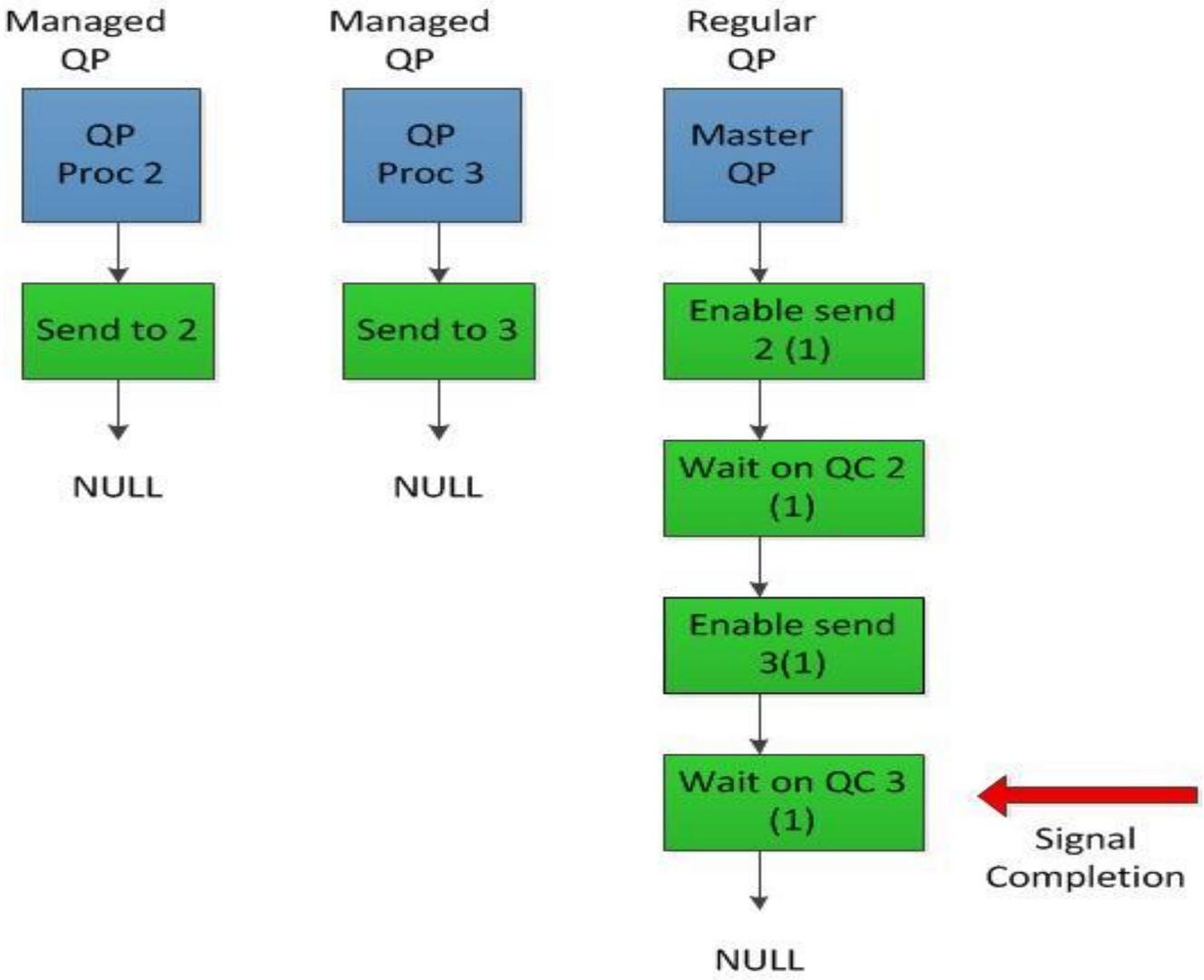
- Task list
- Target QP for task
- Operation
  - Send
  - Wait for completions
  - Enable
  - Calculate



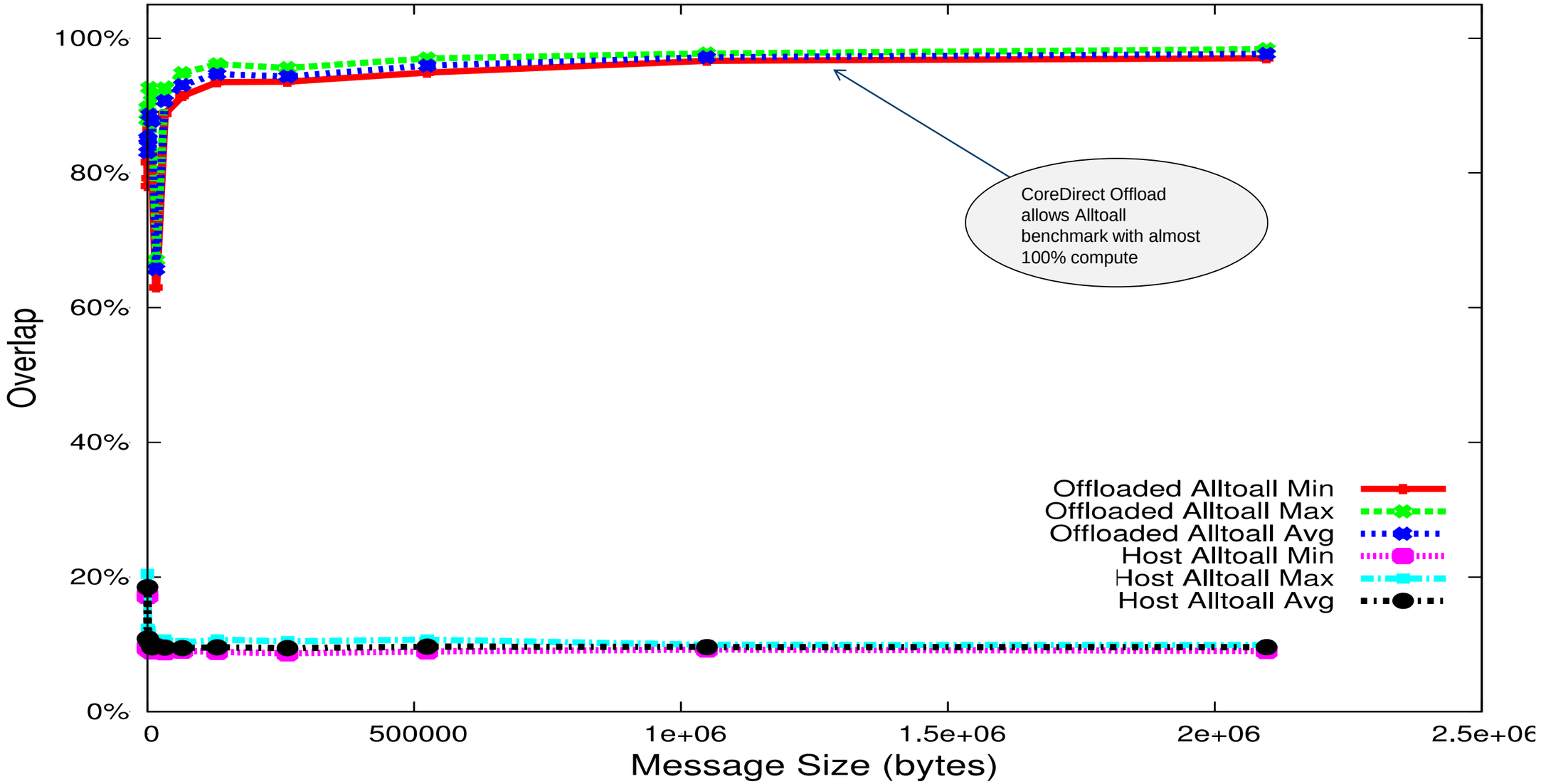
Example – Four Process Recursive Doubling



Four Process Barrier Example – Using Managed Queues – Rank 0



Nonblocking Alltoall (Overlap-Wait) Benchmark



Optimizing Non Contiguous Memory Transfers

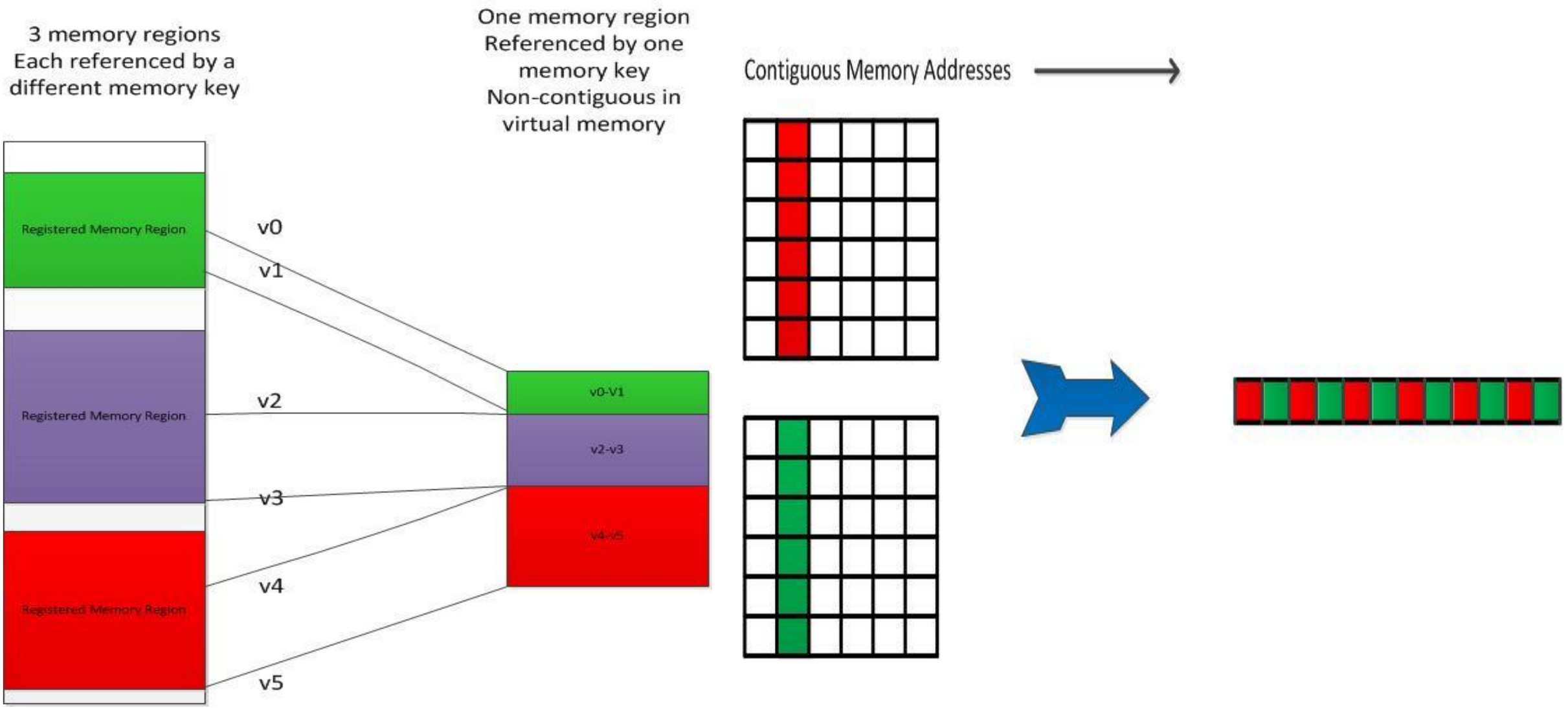
- Support combining contiguous registered memory regions into a single memory region. H/W treats them as a single contiguous region (and handles the non-contiguous regions)

- 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

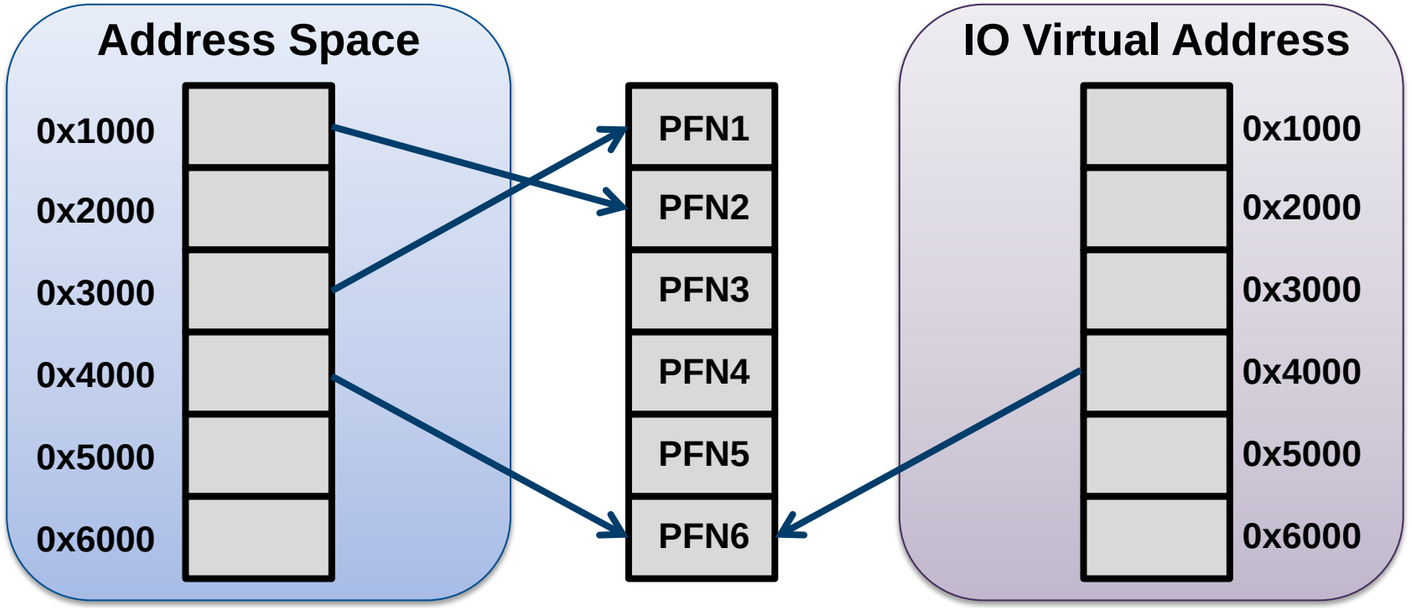
- Supports local and remote non-contiguous memory access
  - Eliminates the need for some memory copies

Optimizing Non Contiguous Memory Transfers



On Demand Paging

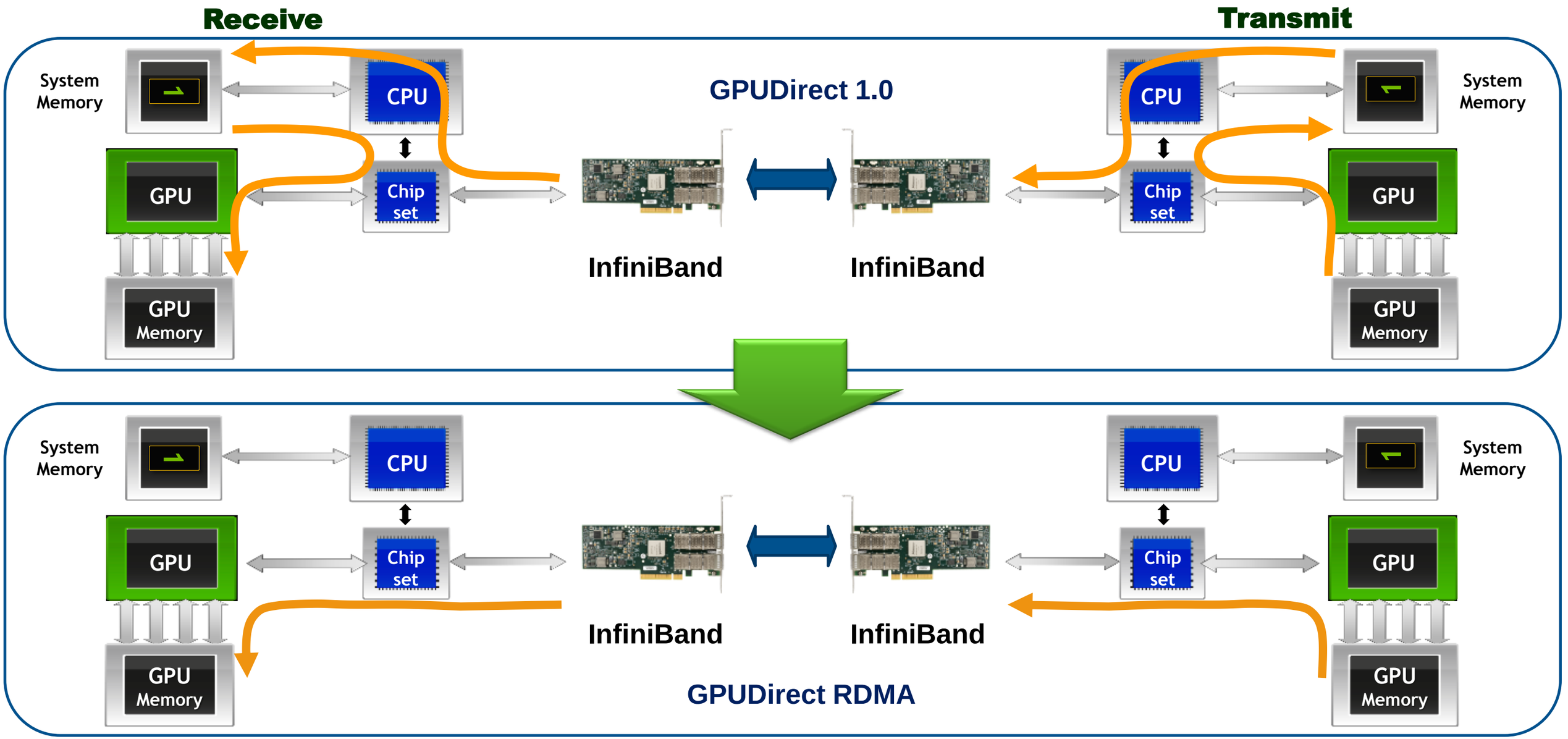
- No memory pinning, no memory registration, no registration caches!
- Advantages
  - Greatly simplified programming
  - Unlimited MR sizes
  - Physical memory optimized to hold current working set



ODP promise:
IO virtual address mapping == Process virtual address mapping

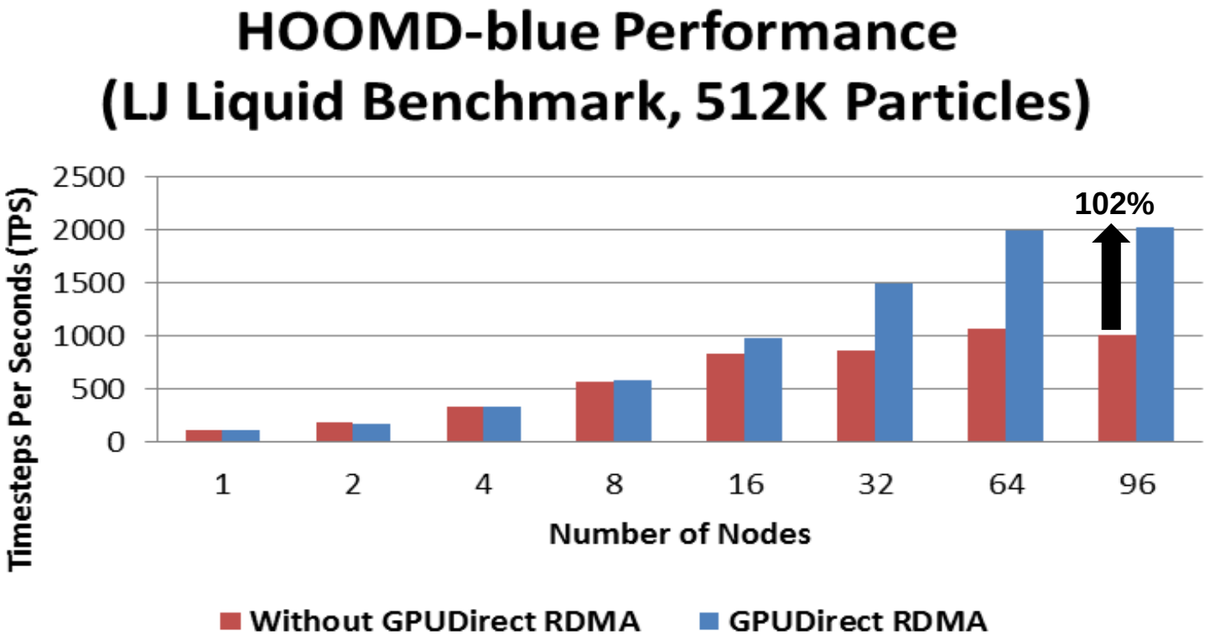
Connecting Compute Elements

GPUDirect RDMA



Mellanox PeerDirect™ with NVIDIA GPUDirect RDMA

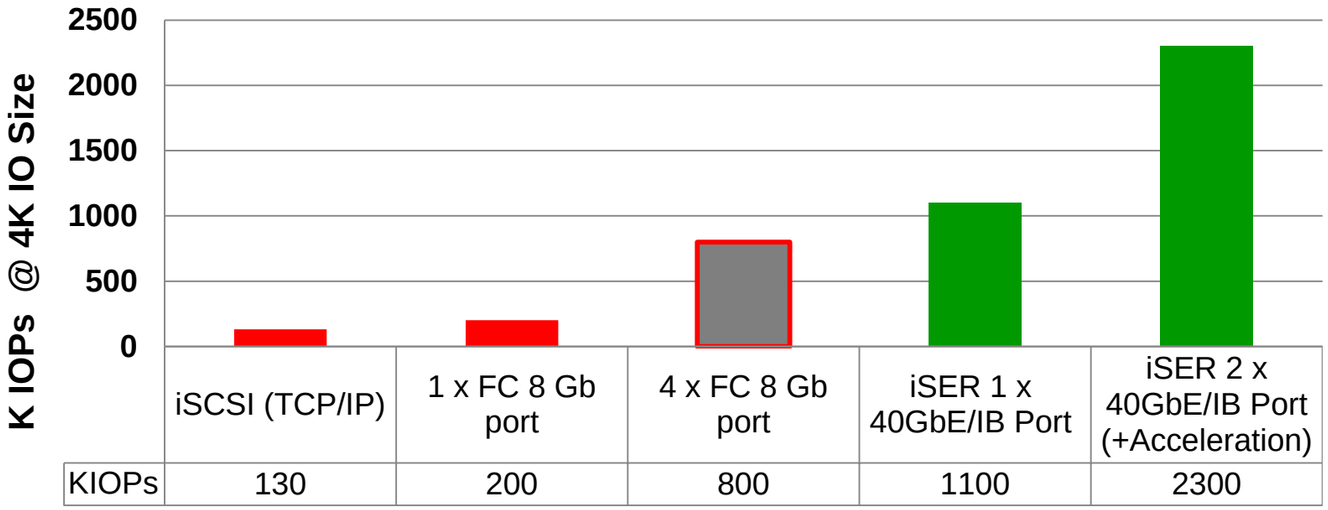
- HOOMD-blue is a general-purpose Molecular Dynamics simulation code accelerated on GPUs
- GPUDirect RDMA allows direct peer to peer GPU communications over InfiniBand
  - Unlocks performance between GPU and InfiniBand
  - This provides a significant decrease in GPU-GPU communication latency
  - Provides complete CPU offload from all GPU communications across the network
- Demonstrated up to 102% performance improvement with large number of particles

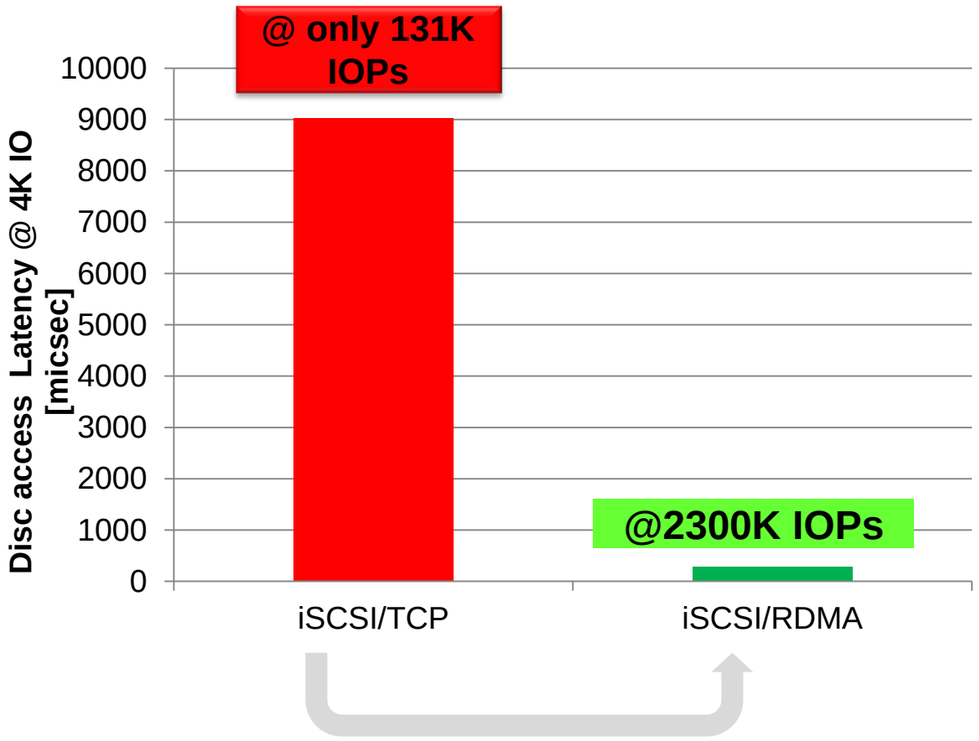


Storage

InfiniBand RDMA Storage – Get the Best Performance!

- Transport protocol implemented in hardware
- Zero copy using RDMA

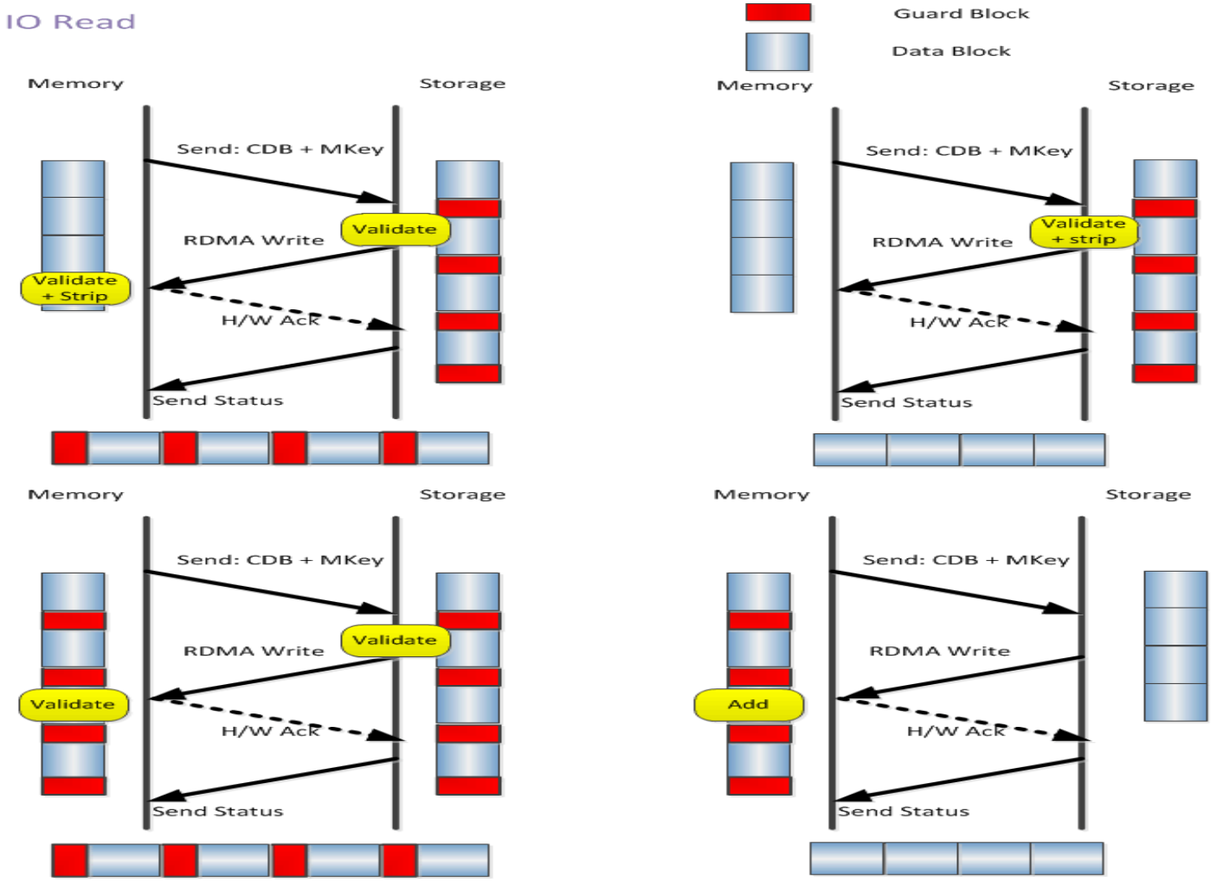


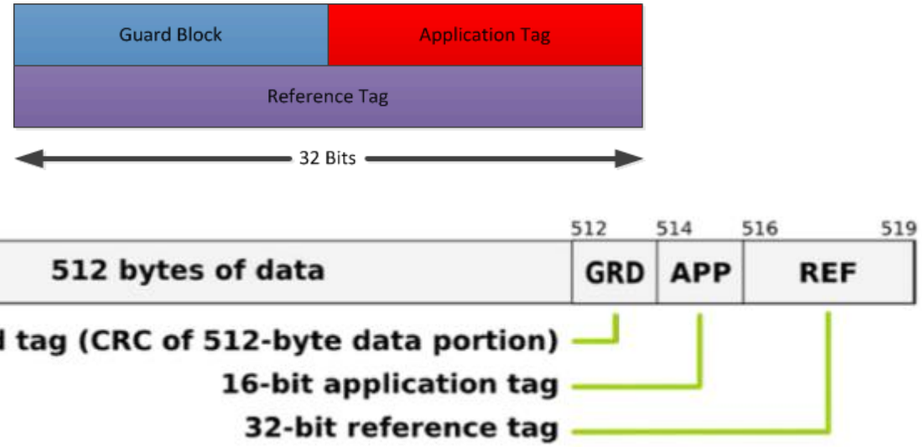


5-10% the latency under 20x the workload

Data Protection Offloaded

- Used to provide data block integrity check capabilities (CRC) for block storage (SCSI)
- Proposed by the T10 committee
- DIF extends the support to main memory



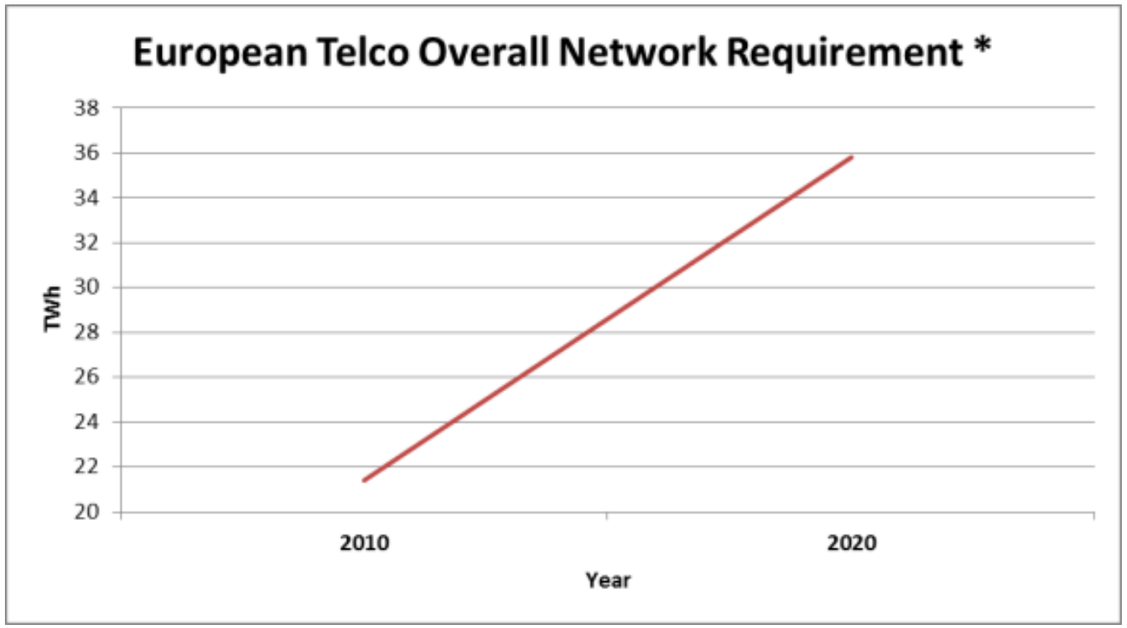


Power

Motivation For Power Aware Design

- Today networks work at maximum capacity with almost constant dissipation
- Low power silicon devices may not suffice to meet future requirements for low-energy networks
- The electricity consumption of datacenters is a significant contributor to the total cost of operation
  - Cooling costs scale as 1.3X the total energy consumption

- Lower power consumption lowers OPEX
  - **Annual OPEX for 1KW is ~$1000**



*** According to Global e-Sustainability Initiative (GeSI) if green network technologies (GNTs) are not adopted**

Increasing Energy Efficiency with Mellanox Switches

ASIC level

Voltage/Frequency scaling
Dynamic array power-save

Link level

Width/Speed reduction
Energy Efficient Ethernet

Aggregate Power Aware API
FW / MLNX-OS™ / UFM™

Switch Level

Port/Module/System shutdown

System Level

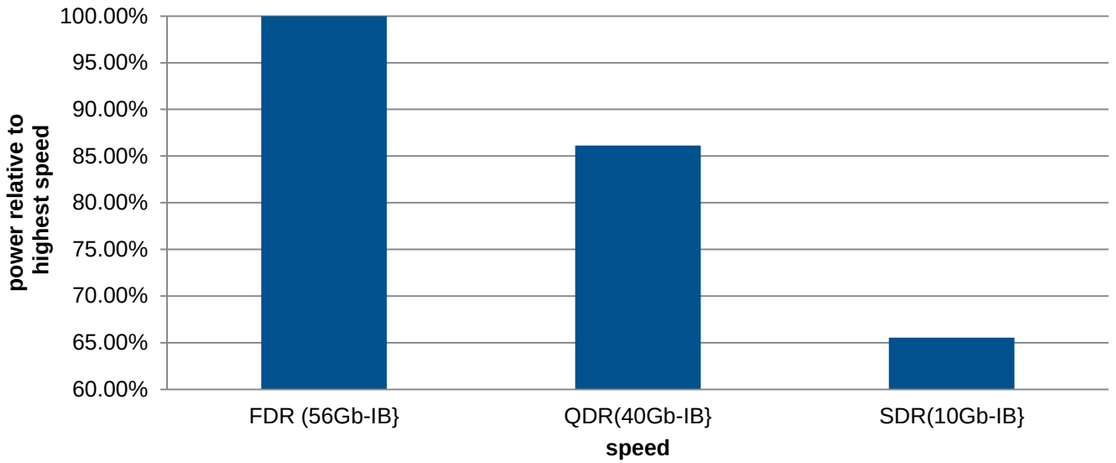
Energy aware cooling control

ECO net

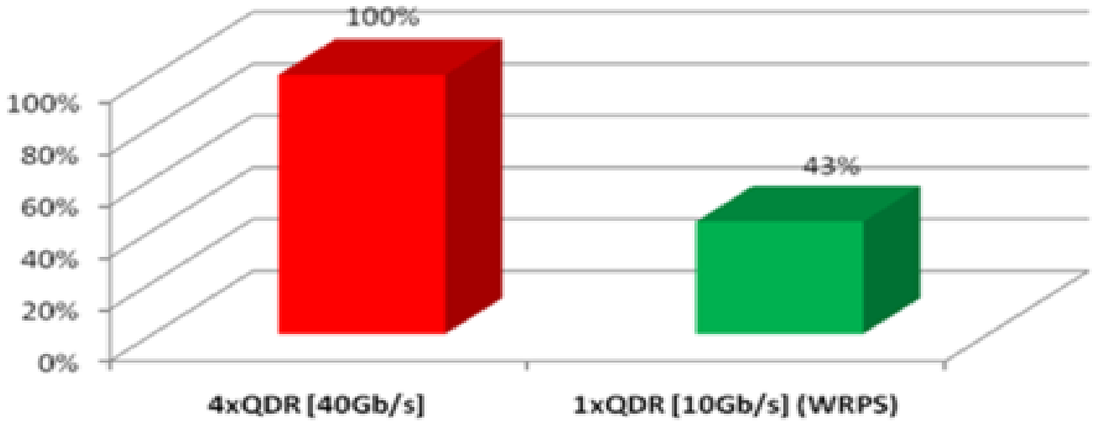
Low Energy COnsumption NETworks

Prototype Results - Power Scaling @ Link Level

- Speed Reduction Power Save [SRPS]

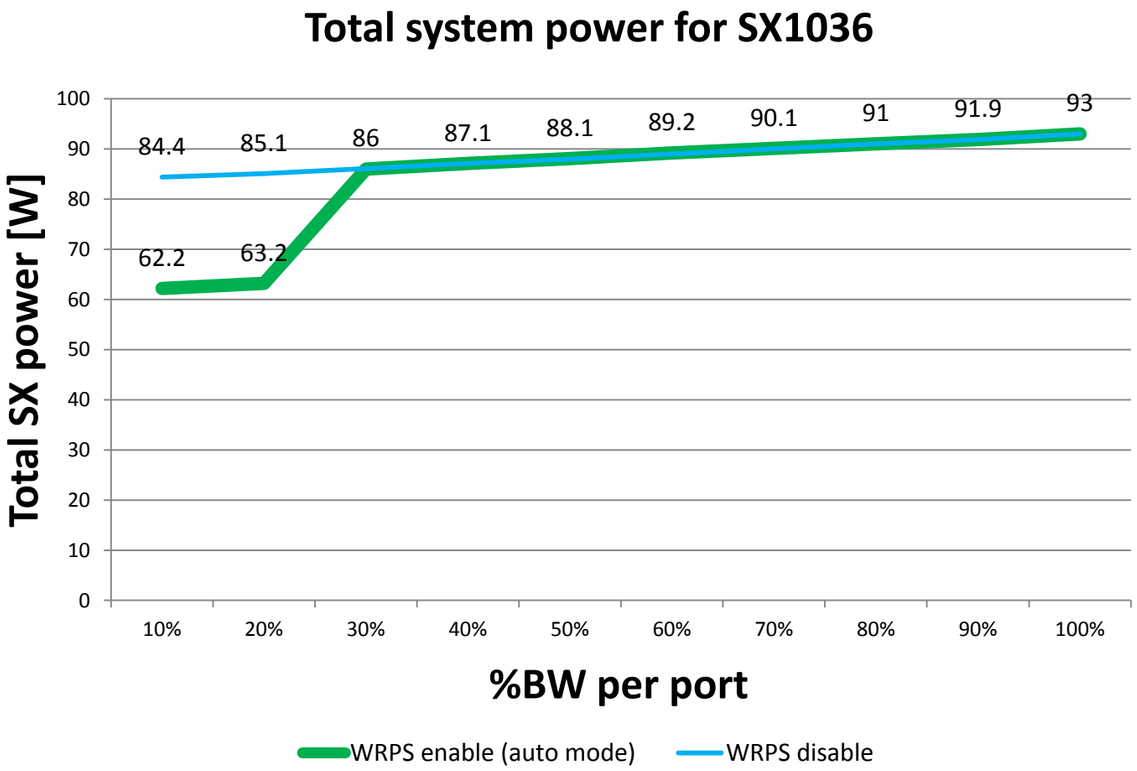


- Width Reduction Power Save [WRPS]

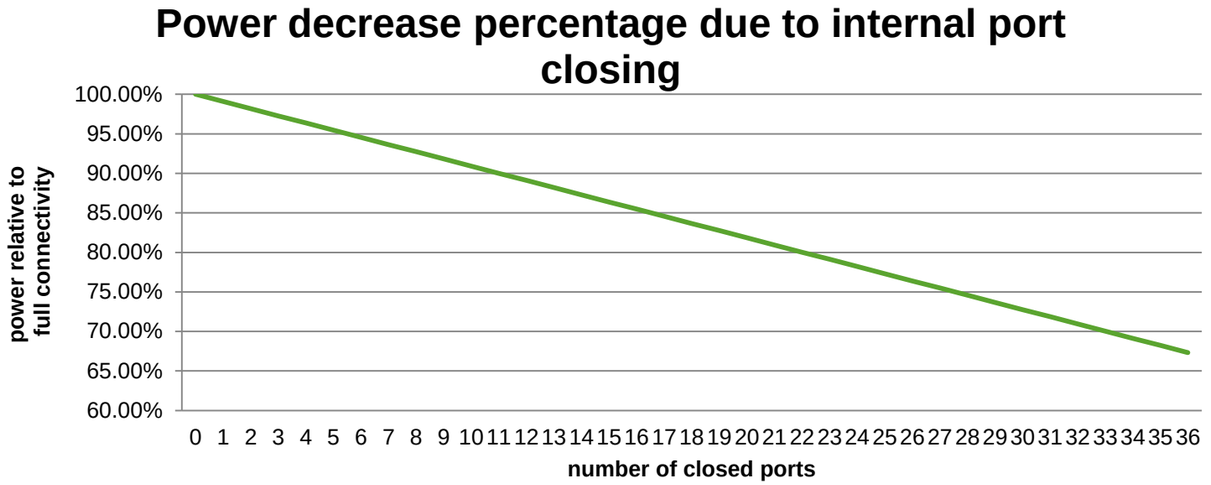


Prototype Results - Power Scaling @ System Level

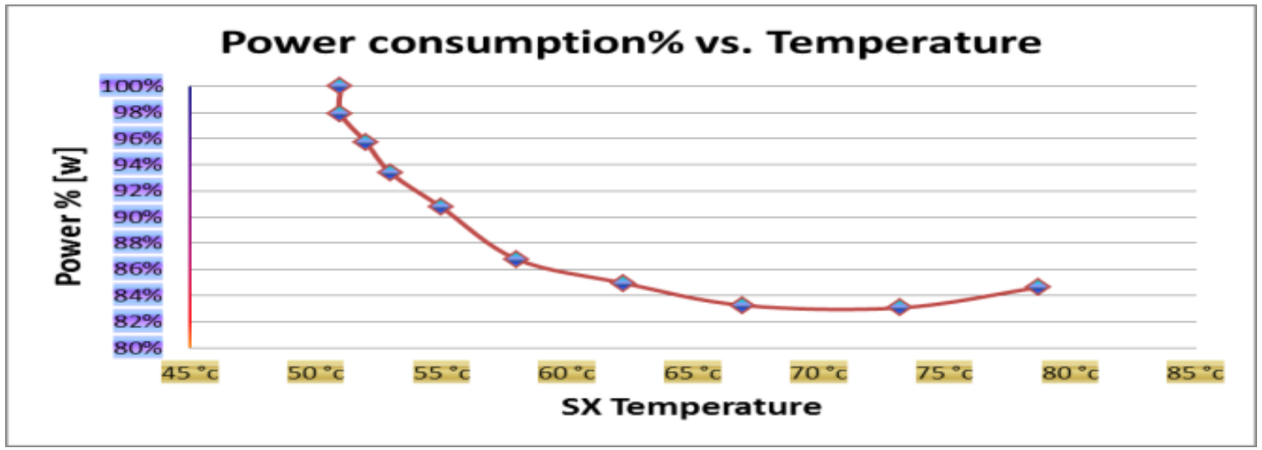
- Load based power scaling



- Internal port shutdown (Director switch) - ~1%/port



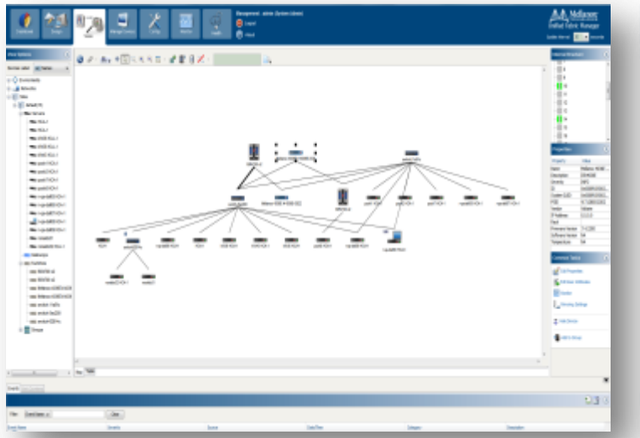
- Fan system analysis for improved power algorithm



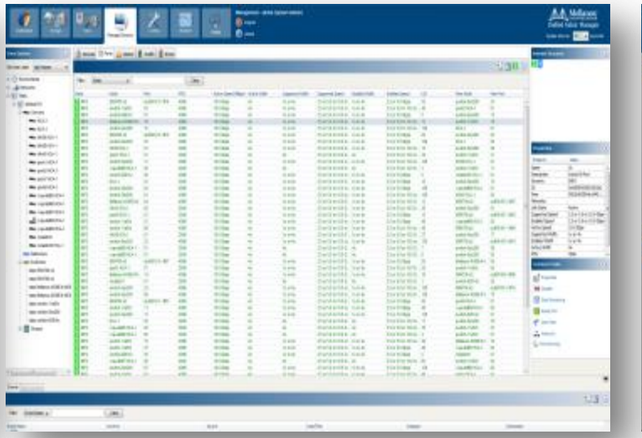
Monitoring and Diagnostics

Unified Fabric Manager

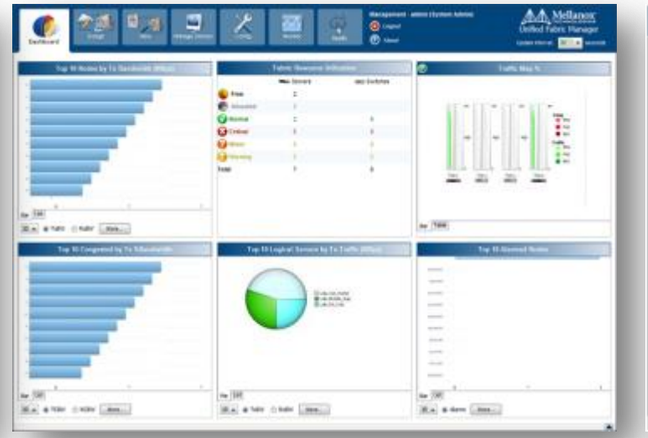
Automatic Discovery



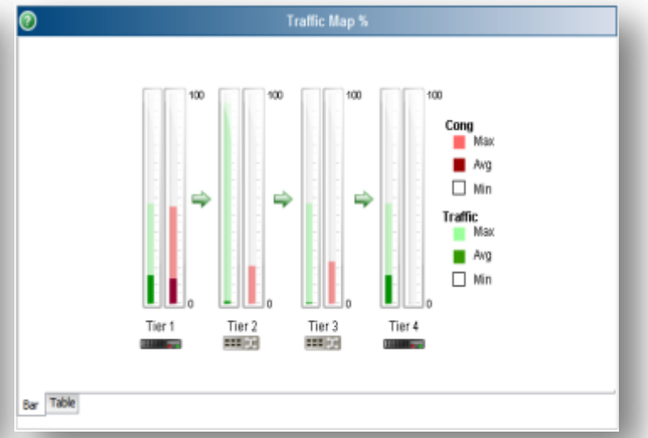
Central Device Mgmt



Fabric Dashboard

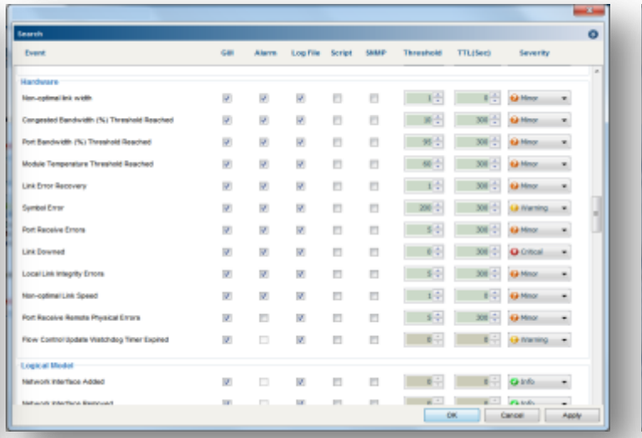


Congestion Analysis

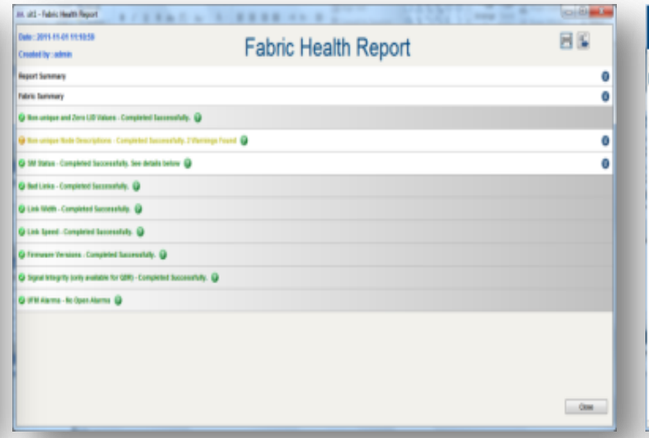


Health & Perf Monitoring

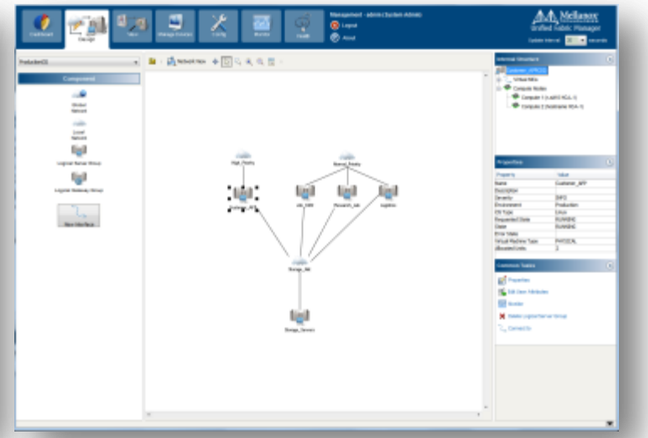
Advanced Alerting



Fabric Health Reports



Service Oriented Provisioning



ExaScale !

Mellanox InfiniBand Connected Petascale Systems

Connecting Half of the World's Petascale Systems

Mellanox Connected Petascale System Examples

The Only Provider of End-to-End 40/56Gb/s Solutions

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

ICs

Adapter Cards

Switches/Gateways

Host/Fabric Software

Metro / WAN

Cables/Modules

From Data Center to Metro and WAN

X86, ARM and Power based Compute and Storage Platforms

The Interconnect Provider For 10Gb/s and Beyond

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™