New Accelerations for Parallel Programming

Todd Wilde – Director of Technical Computing and HPC

HPC@mellanox.com

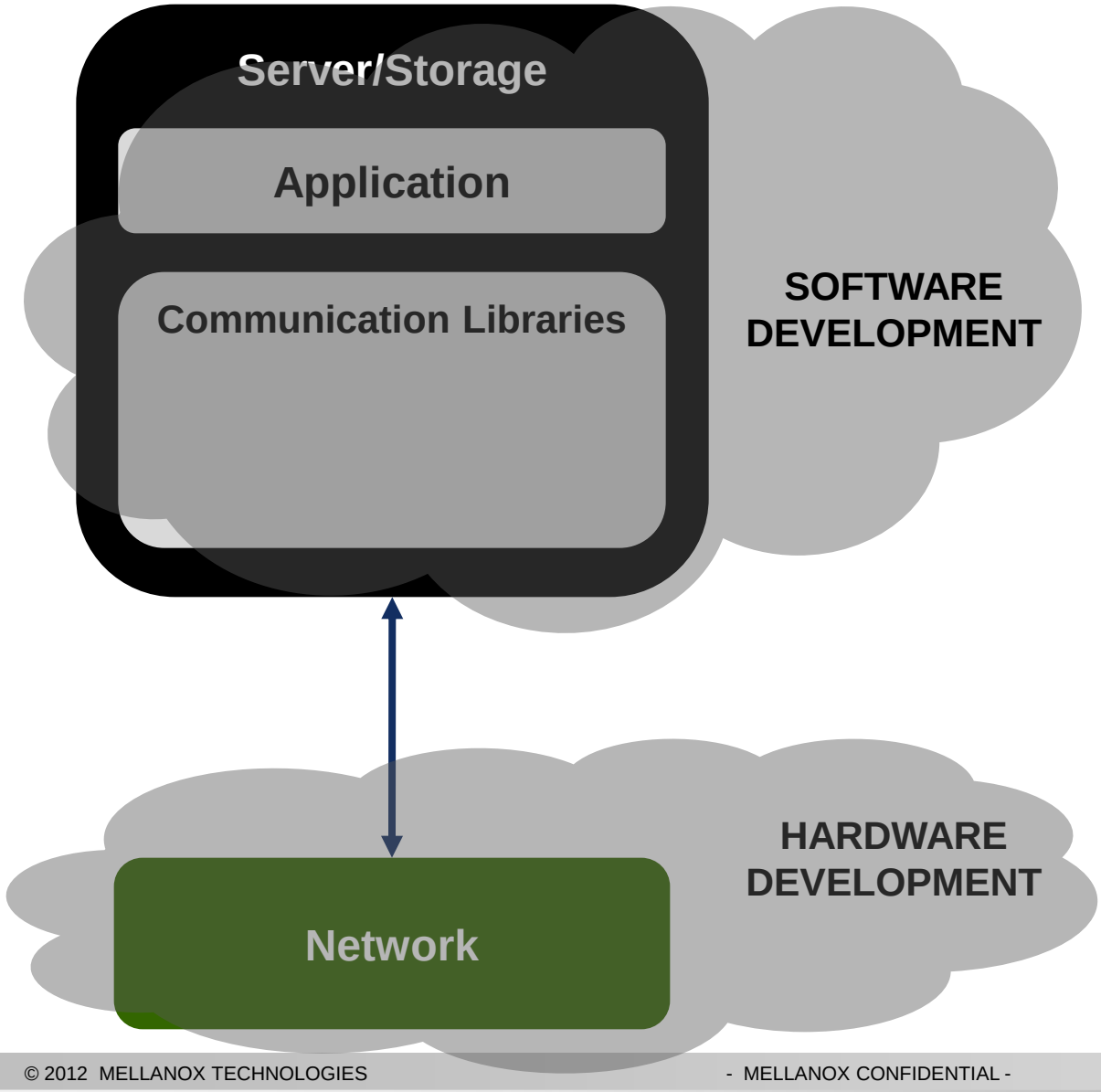
Mellanox ScalableHPC

- Offer high performing and scalable parallel programming libraries for high performance computing

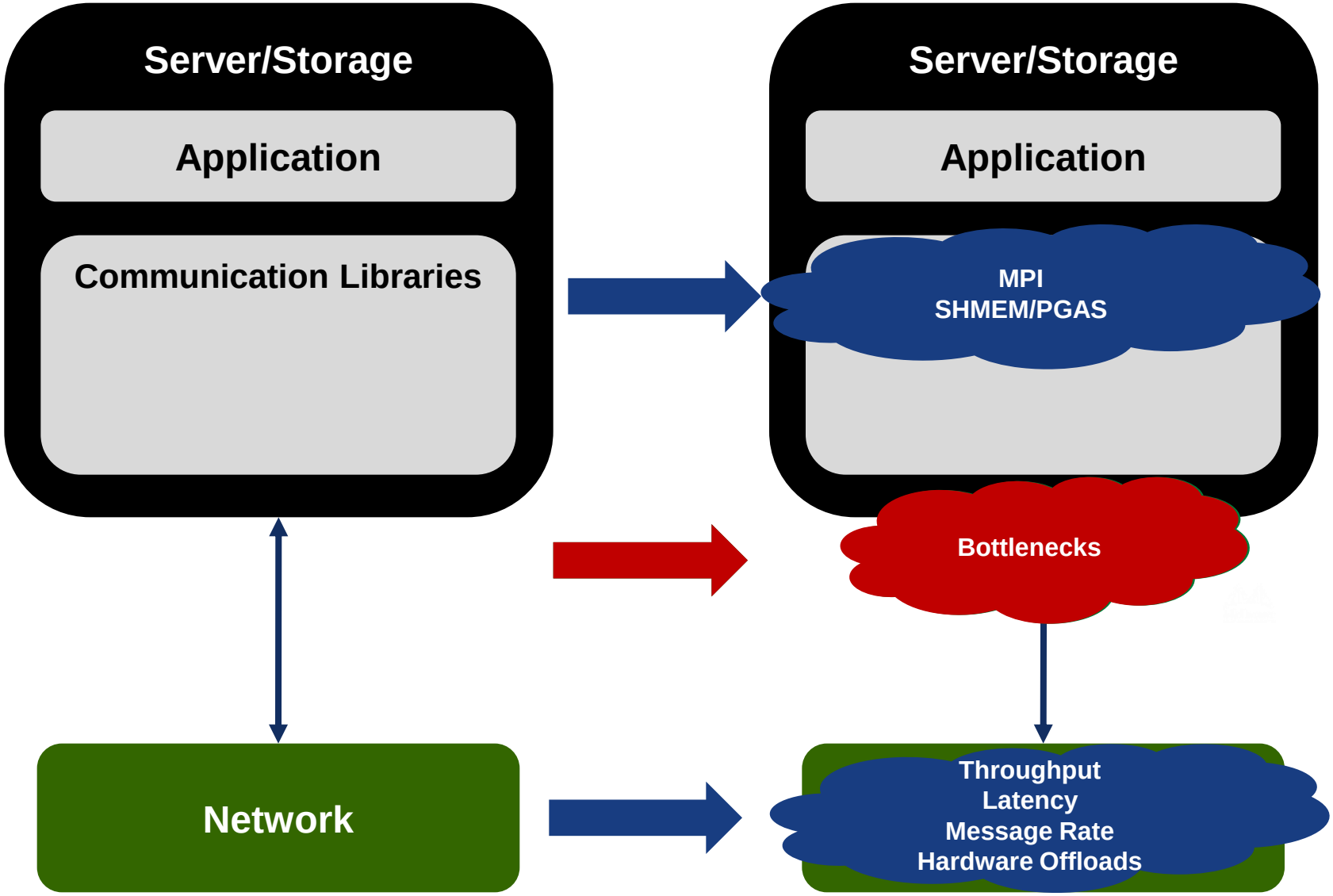
- Support a comprehensive set of MPIs and PGAS languages
  - Integration of acceleration technology into broad list of languages and implementations
  - Provide Mellanox library packages when there is no open source alternative

- Integrates Mellanox acceleration components into MPIs/PGAS languages
  - MXM – MellanoX Messaging Accelerator
  - FCA – Mellanox Fabric Collective Accelerator

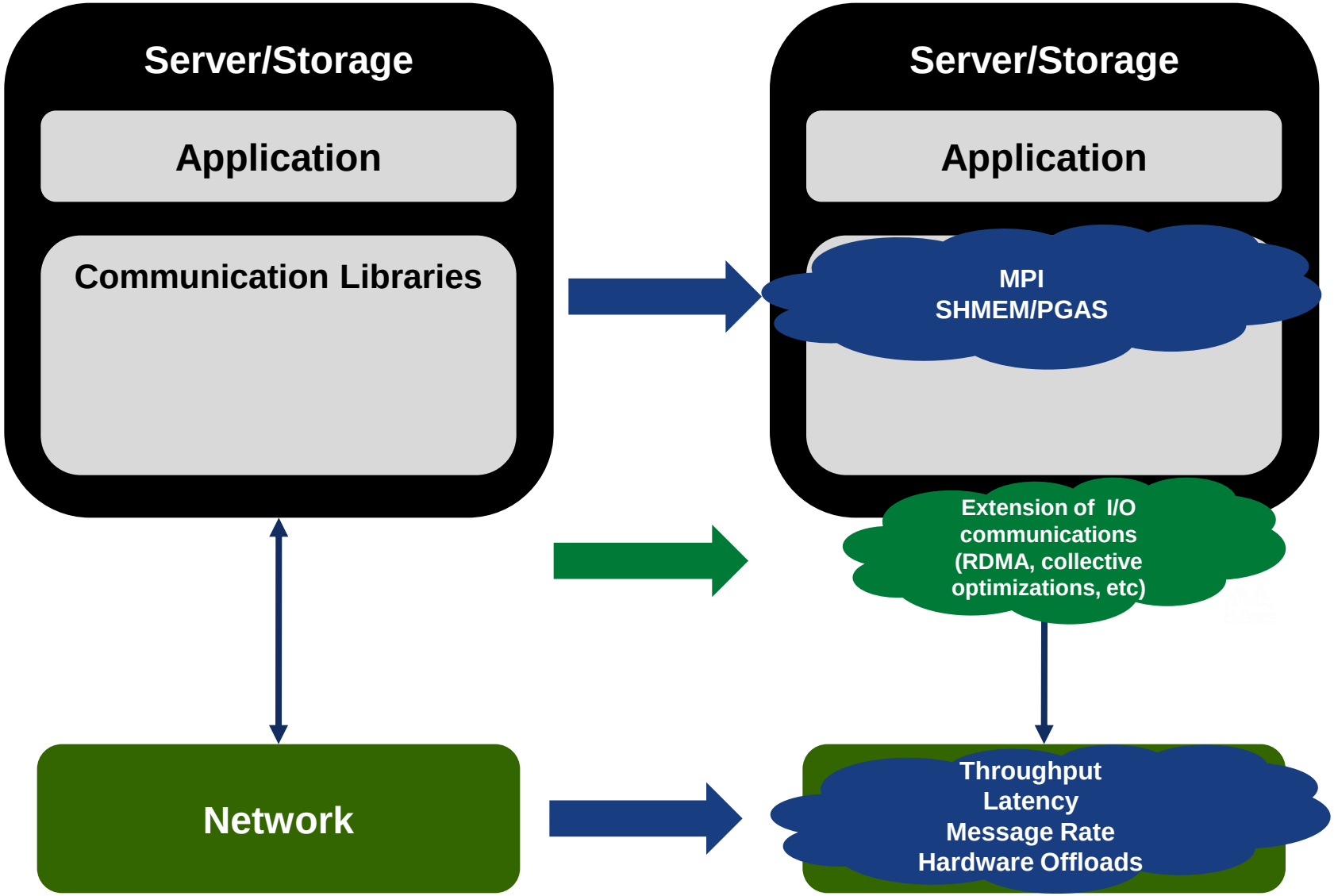
Co-Design Strategy



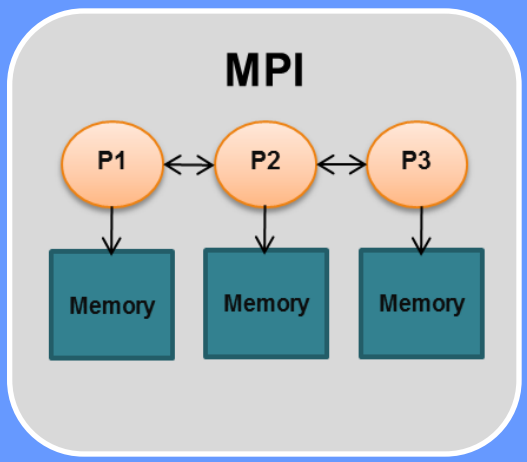
Co-Design Strategy

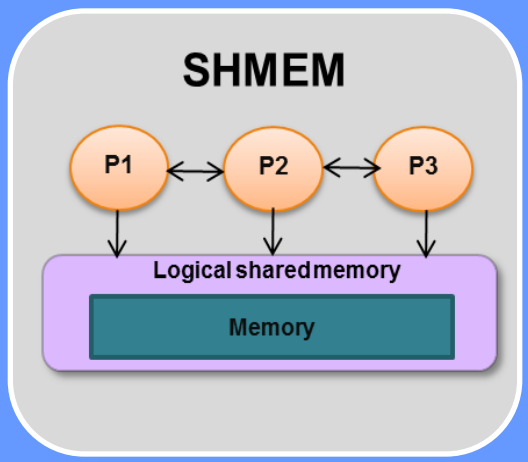


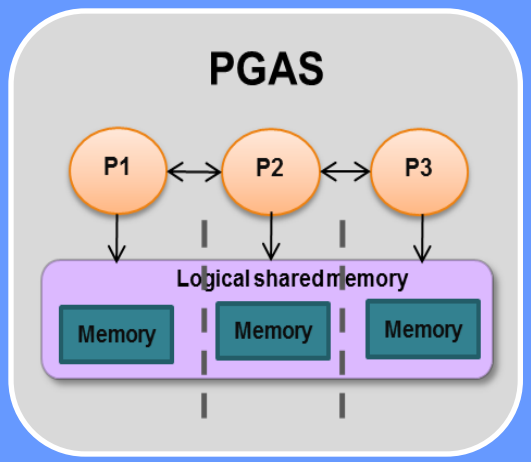
Co-Design Strategy



Mellanox ScalableHPC Accelerate Parallel Applications







MXM

- Reliable Messaging Optimized for Mellanox HCA
- Hybrid Transport Mechanism
- Efficient Memory Registration
- Receive Side Tag Matching

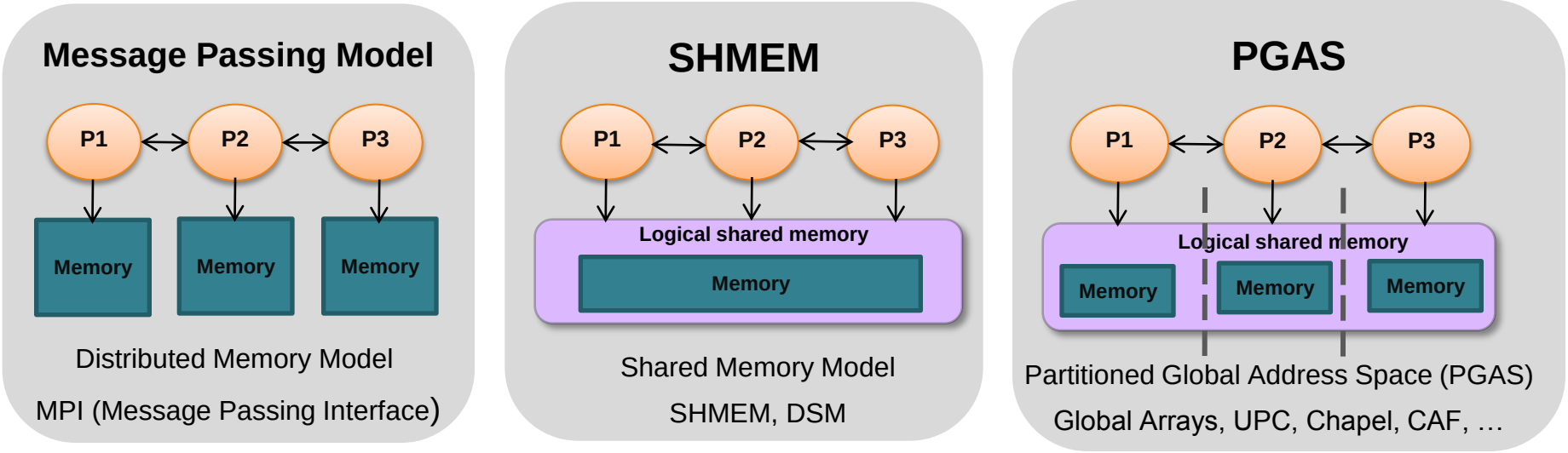
FCA

- Topology Aware Collective Optimization
- Hardware Multicast
- Separate Virtual Fabric for Collectives
- CoreDIrect Hardware Offload

InfiniBand Verbs API

HPC Parallel Programming Models over InfiniBand

- 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) and Chapel are popular examples



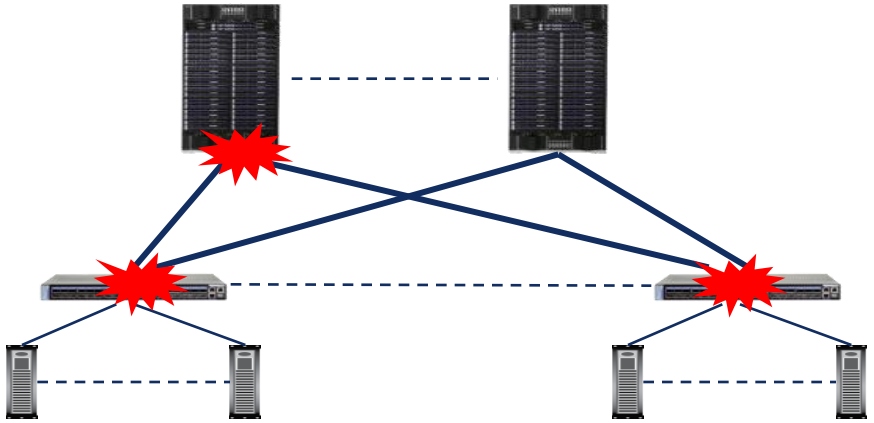
Fabric Collective Accelerations

What are Collective Operations?

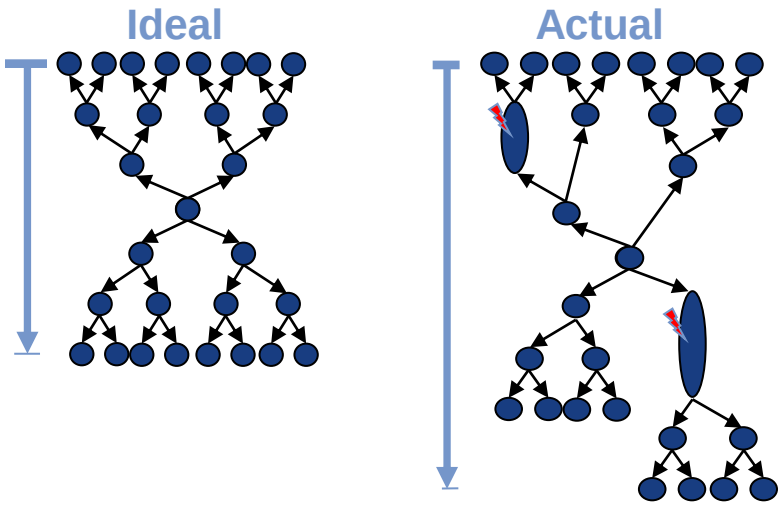
- Collective Operations are Group Communications involving all processes in job
- Synchronous operations
  - By nature consume many ‘Wait’ cycles on large clusters
- Popular examples
  - Barrier
  - Reduce
  - Allreduce
  - Gather
  - Allgather
  - Bcast

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)

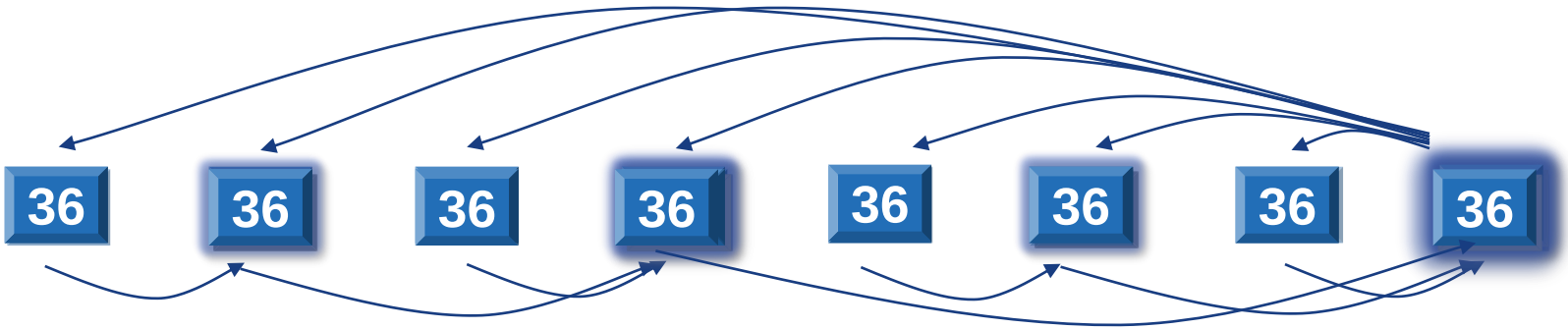
Mellanox InfiniBand Switches
High performance IB multicast
for result distribution

FCA Manager
Topology-based collective tree
Separate Virtual network
IB multicast for result distribution

FCA Agents
Library integrated with MPI
Intra-node optimizations
CoreDirect integration

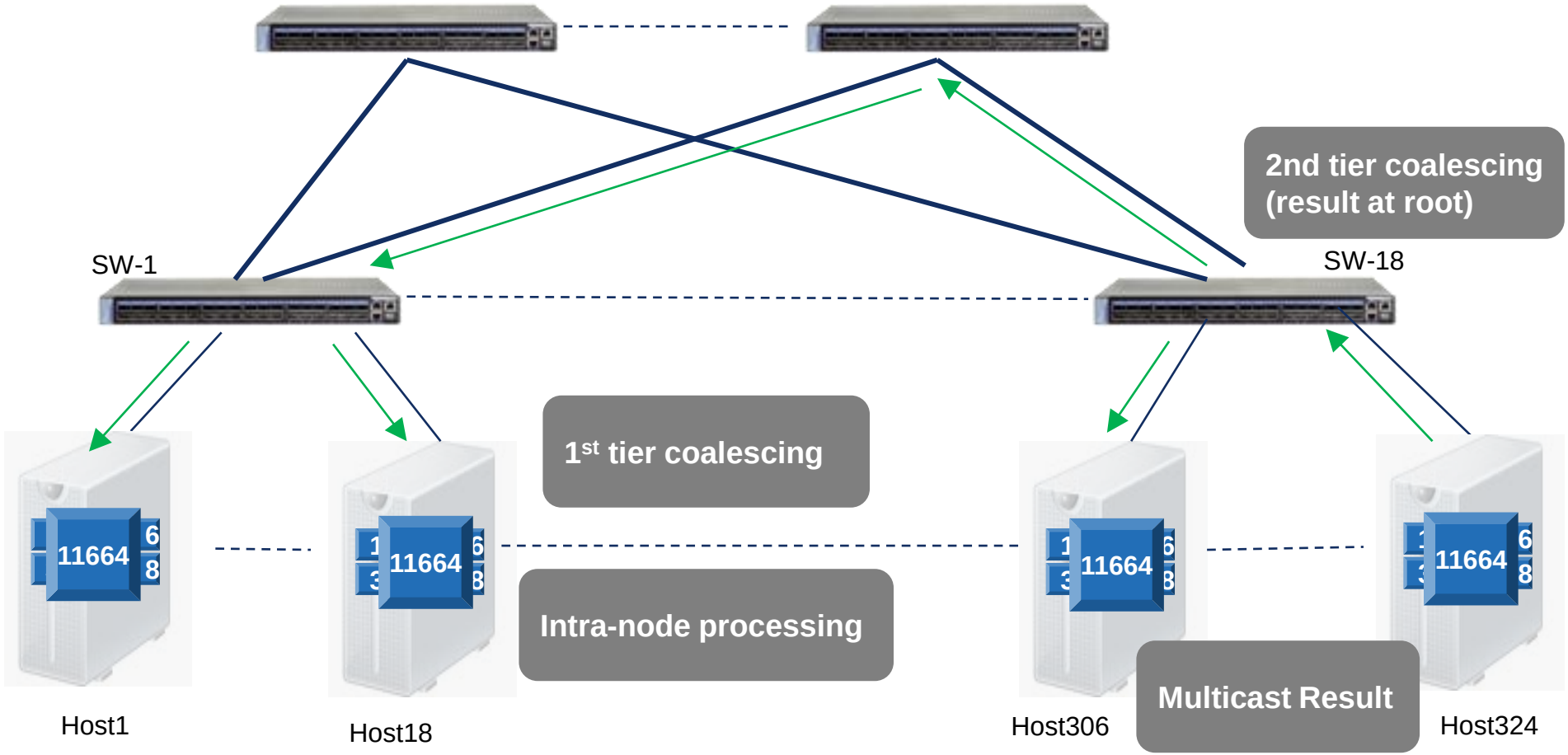
Collective Example – Allreduce using Recursive Doubling

- Collective Operations are Group Communications involving all processes in job



- A 4000 process Allreduce using recursive doubling is 12 stages

Scalable Collectives with FCA



Thank You

HPC@mellanox.com