OPENFABRICS ALLIANCE

MPI Collectives Offloads

Author: Gilad Shainer, shainer@mellanox.com
Date: April 2011

MPI Collective Operations

- MPI provides messaging interface for parallel computing
  - Communications options include send/receive and collectives
  - Used by applications processes for communications
    - One-to-one (one process to another), many-to-one, one-to-many
- Collectives communications
  - Have a crucial impact on the application's scalability and performance
  - Communications used for one-to-many or many-to-one
    - Used for sending around initial input data
    - Reductions for consolidating data from multiple sources
    - Barriers for global synchronization
- Collectives operations
  - Must be executed as fast as possible
  - Each local node delay will impact the entire cluster performance
  - Consume high percentage of CPU cycles
- Offloading MPI collectives to the network
  - Overlapping and minimizing jitter

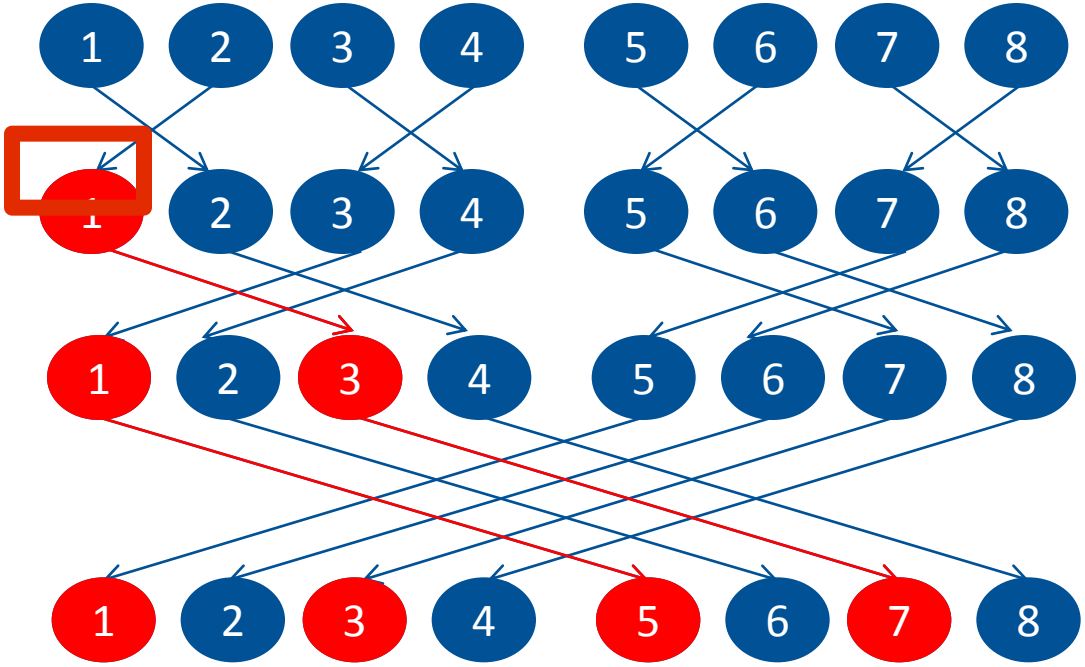
MPI Collectives Offloads

- Offloading means: move collectives outside of the CPU – have the network to manage these operations

- Offloading collectives to the network
  - NOT because InfiniBand verbs provide limited performance
    - InfiniBand verbs can provide great performance
  - NOT because vendors needs it for performance increase
    - Vendor solution can work great with CPU-based collectives

- Offloading collectives to the network
  - Because of overlapping (MPI-3) and Jitter elimination

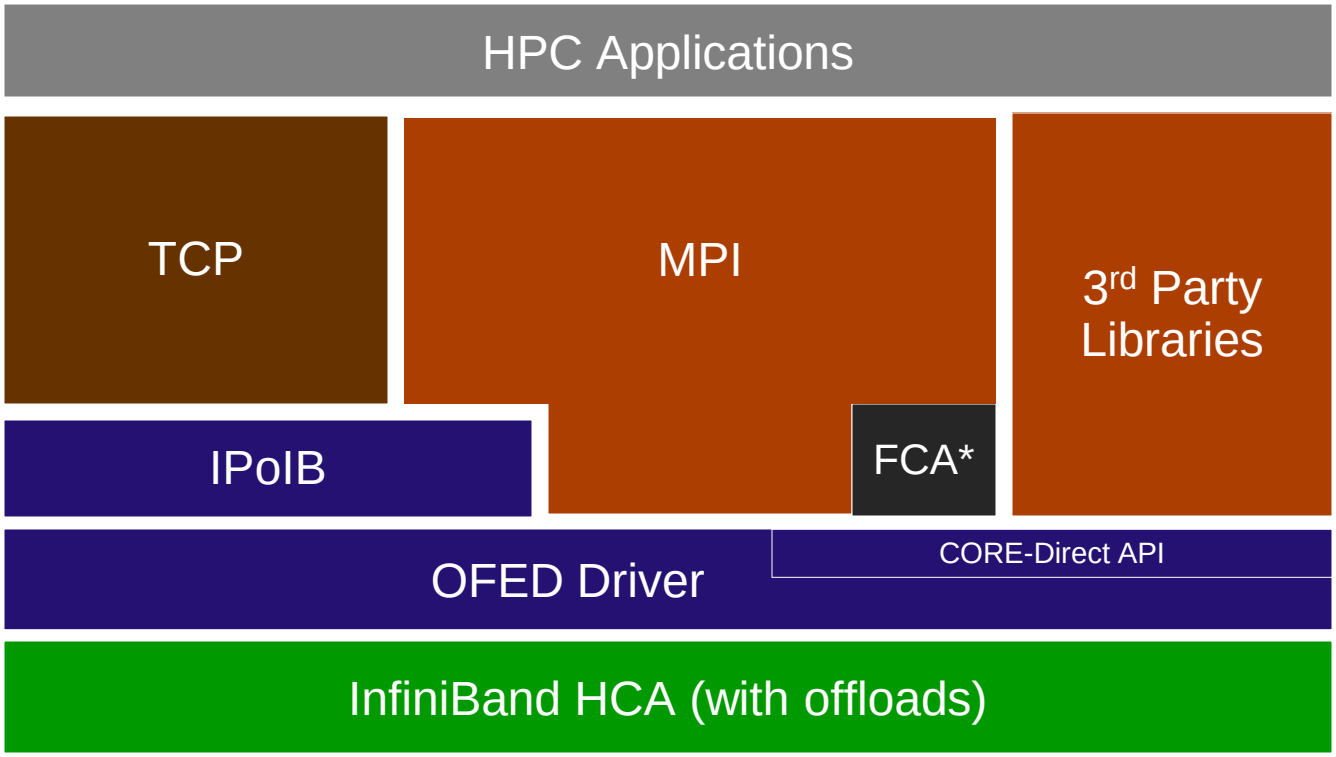
Collective Algorithms (Recursive Doubling)



Efficient Execution of Collectives Operation

	CPU Executes Collectives	Interconnect Executes Collectives
Fast propagation throughout the system	Fast	Fastest
Negative effect of a single node on the entire system (system noise/jitter)	Maximizing the effect	Minimizing the effect
Reducing CPU overhead and maximizing CPU availability for the application	Maximum CPU overhead	Minimum CPU overhead, allowing overlap between computations and communications

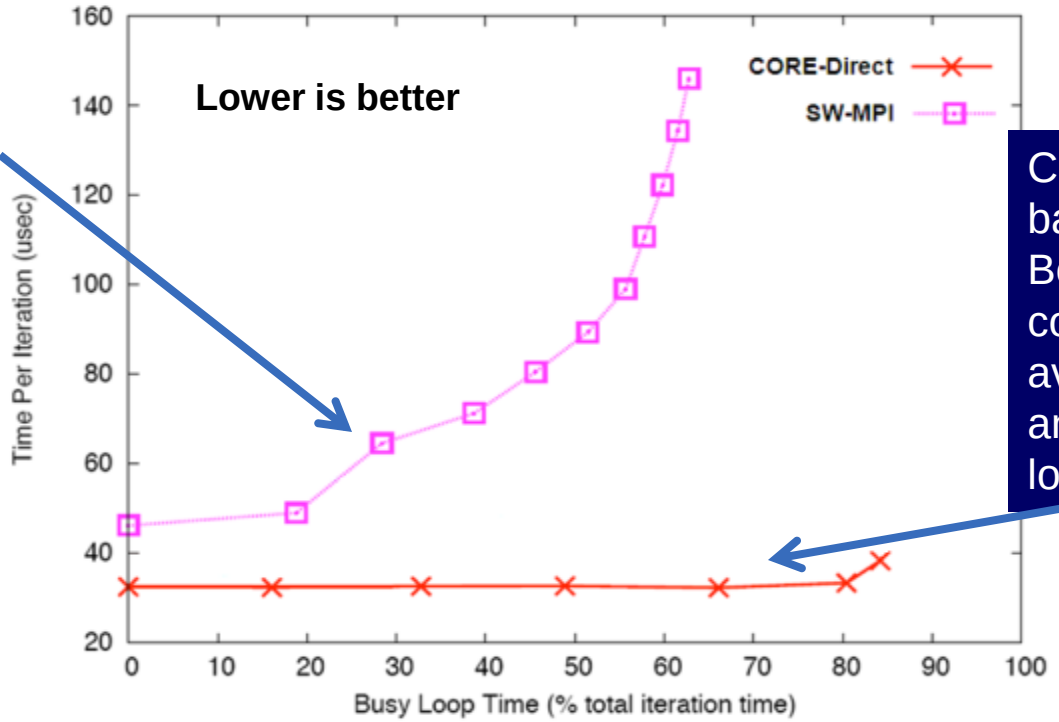
Software Stack



Collectives Offloads - Non Blocking Collectives

- Presents the overlapping benefit of collective offloads
  - Non-blocking collective implementation - non-blocking barrier
    - Initiate non-blocking MPI barrier
    - CPU to performance application calculations
    - Wait for non-blocking barrier to complete

Software MPI: Losing performance beyond 20% CPU computation availability



Collectives Offload based MPI: Beyond 80% CPU computation availability without any performance loss!

Thank You