



# Advances in High Performance Computing

**Brian Forbes – Sr. System Engineer**



- Performance : 100Gb/s network injection rate, under 1us end-end latency
- Scalability : Flat, Fat and Fast. Powering 8 of the TOP20 Petascale Supercomputers
- Advanced Features : Collective Communication Offloads, Flexible Topologies
- Big Data : Boost Hadoop Performance with Direct InfiniBand RDMA Access

## Comprehensive End-to-End 10/40/56Gb/s Ethernet and 56Gb/s InfiniBand Portfolio



**Scalability, Reliability, Power, Performance**

# FDR INFINIBAND TECHNOLOGY

THE NEXT GENERATION OF  
HIGH-PERFORMANCE SCALABLE CONNECTIVITY

## FDR InfiniBand New Features and Capabilities

### Performance / Scalability

- >12GB/s bandwidth, <0.7usec latency
- PCI Express 3.0
- InfiniBand Routing and IB-Ethernet Bridging

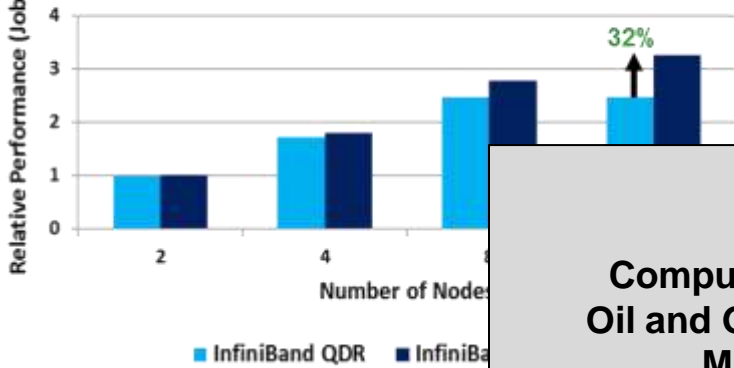
### Reliability / Efficiency

- Link bit encoding – 64/66
- Forward Error Correction
- Lower power consumption

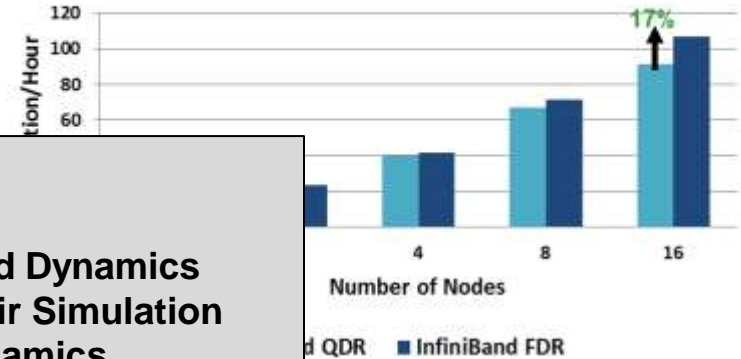
# FDR Application Benchmarks



### ECLIPSE 2012 Performance (FOURMILL, Platform MPI)



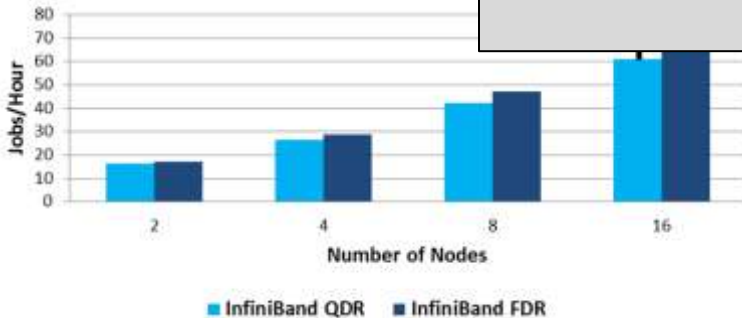
### WRF Benchmark (conus12km)



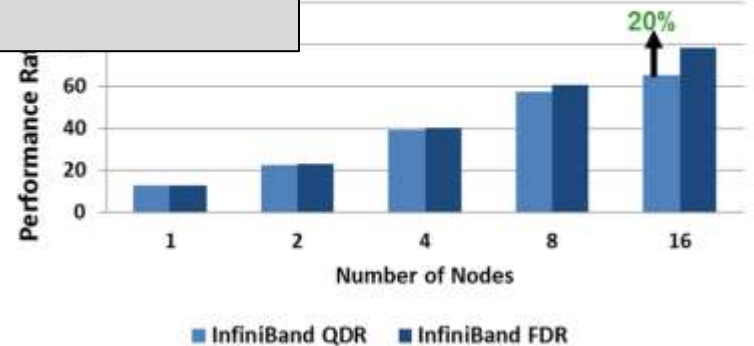
Computational Fluid Dynamics  
Oil and Gas Reservoir Simulation  
Molecular Dynamics  
Weather and Earth Sciences

Up to 32% ROI on equipment and operating costs

### CP2K Benchmark (H2O-128, Intel MPI)



### MPS Benchmark (Rhodopsin Protein)

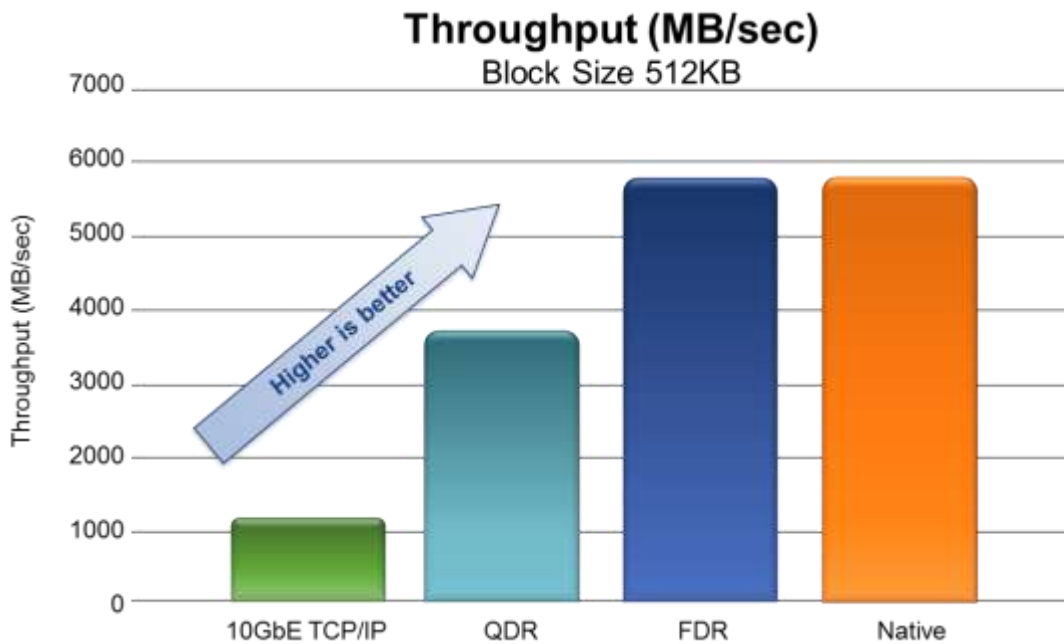


# FDR InfiniBand Meets the Needs of Changing Storage World

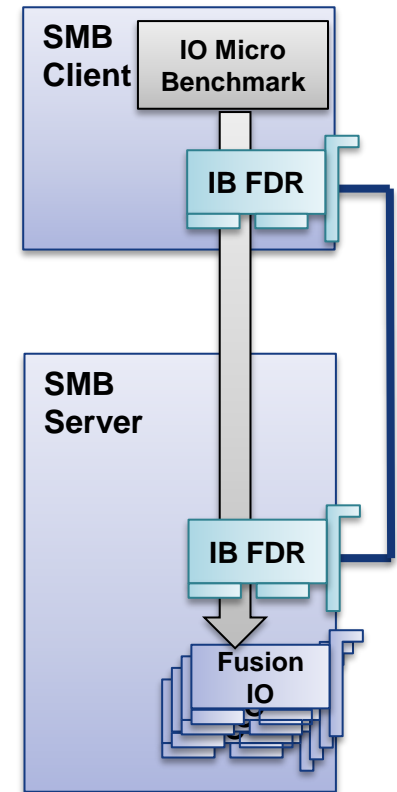


SSDs, the storage hierarchy, In-Memory Computing.....

Remote I/O access needs to be equal to local I/O access

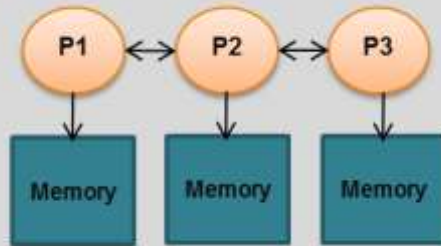


**SMB 3.0 + RDMA  
(InfiniBand FDR)**

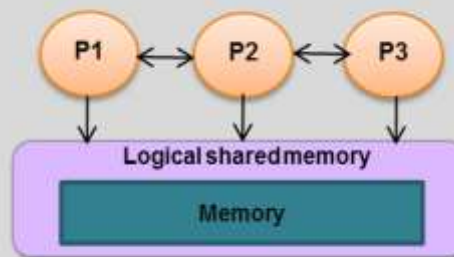


**Native Throughput Performance over InfiniBand FDR**

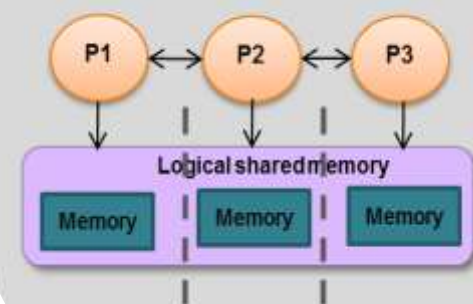
## MPI



## SHMEM



## PGAS



## MXM

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

## FCA

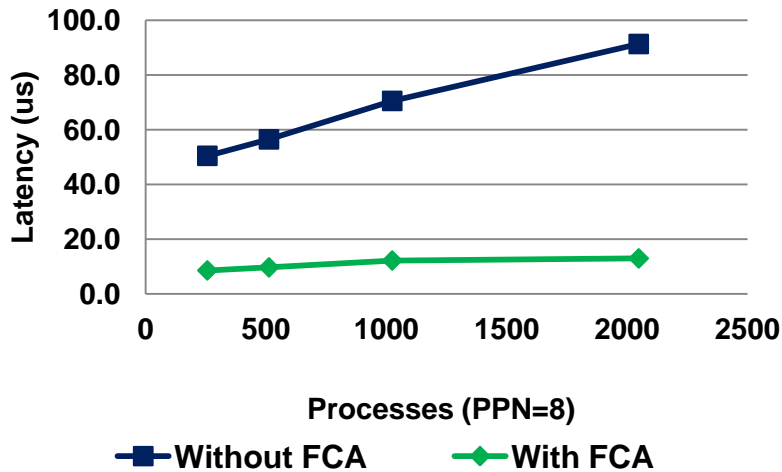
- Topology Aware Collective Optimization
- Hardware Multicast
- Separate Virtual Fabric for Collectives
- CoreDirect Hardware Offload

## InfiniBand Verbs API

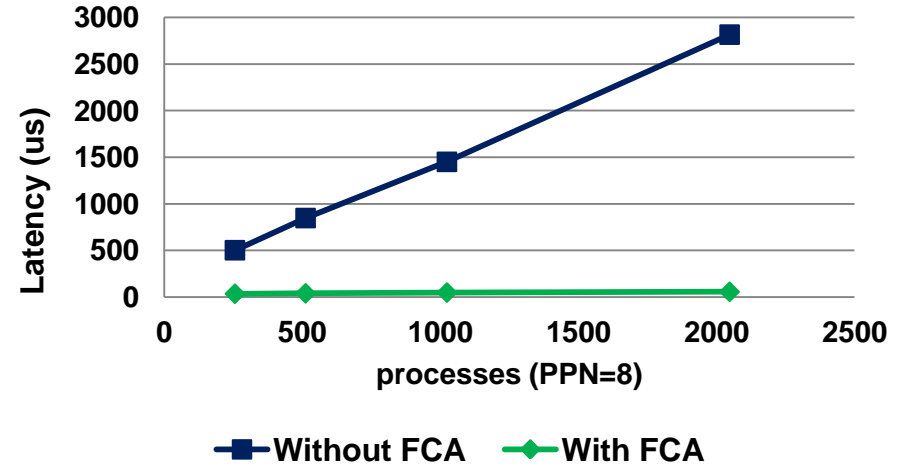
# Fabric Collective Accelerations Provide Linear Scalability



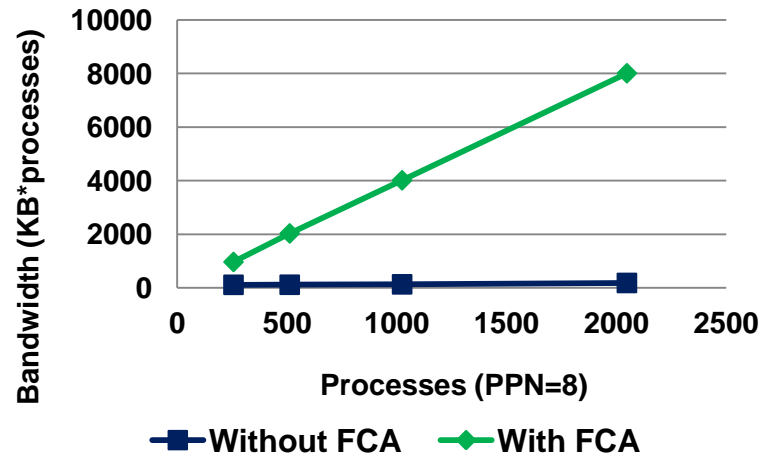
## Barrier Collective



## Reduce Collective



## 8-Byte Broadcast



# Double Hadoop Performance with UDA

## Terasort Benchmark\*

(20GB file size, 16GB Data per node, 8 Mappers, 4 Reducers, 4 Disks)



Disk Writes

40%

Disk Reads

15%

CPU Utilization

2.5X

\*TeraSort is a popular benchmark used to measure the performance of Hadoop cluster

# ~2X Faster Job Completion!



## InfiniHost

World's first  
InfiniBand HCA

10Gb/s InfiniBand  
PCI-X host interface  
1 million msg/sec



2002

## InfiniHost III

World's first PCIe  
InfiniBand HCA

20Gb/s InfiniBand  
PCIe 1.0  
2 million msg/sec



2005

## ConnectX (1,2,3)

World's first  
Virtual Protocol  
Interconnect (VPI)  
Adapter

40/56Gb/s InfiniBand  
PCIe 2.0, 3.0 x8  
33 million msg/sec



2008-11

## Connect-IB

The Exascale  
Foundation



# Announcing Connect-IB: The Exascale Foundation



- A new interconnect architecture for compute intensive applications
- World's fastest server and storage interconnect solution providing 100Gb/s injection bandwidth
- Enables unlimited clustering scalability with new Dynamically Connected Transport service
- Accelerates compute-intensive and parallel-intensive applications with over 130 million msg/sec
- Optimized for multi-tenant environments of 100s of Virtual Machines per server

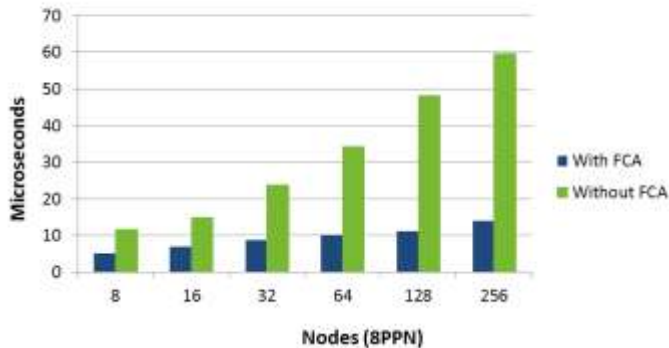
**Enter the World of Scalable Performance**

**Connect IB™**



## ScalableHPC

IMB Barrier - FDR



## Ultimate Scalability with Connect-IB

- 100Gb/s throughput to network
- Over 130-million messages/second
- Dynamically Connected Transport service for unlimited inter-node scaling

Connect **IB**™

## Highest Performing Interconnect

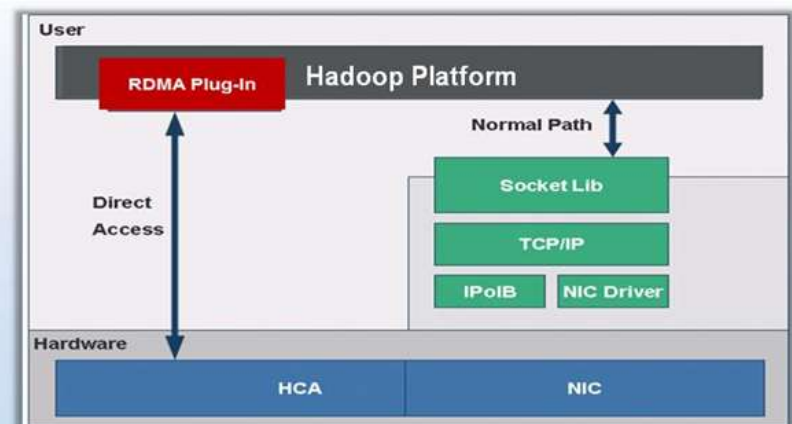


Switch **X**™

- <0.7usec latency
- 56Gb/s throughput
- Higher scalability
- Maximum Reliability



## Accelerating Big Data



# Thank You

[HPC@mellanox.com](mailto:HPC@mellanox.com)

PAVING THE ROAD  
TO **EXASCALE**

ADVANCING NETWORK PERFORMANCE,  
EFFICIENCY, AND SCALABILITY.