Datacenter Fabric Workshop
MPICH2 Windows

MPICH2

MPI-2 over InfiniBand Interconnect

Gil Bloch

Mellanox Technologies Inc.

gil@mellanox.co.il

August 22, 2005

Agenda

- MPI Overview
- MPICH2
- Current Status
- Future Work

Datacenter Fabric Workshop
MPICH2 Windows

MPI Overview

Introduction to MPI and the HPC market

August 22, 2005

Message Passing Interface

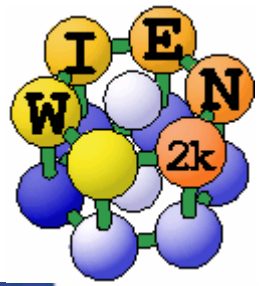
- *de facto* standard for parallel programming
  - Message Passing Library
    - Portable
    - High performance
  - Point-to-point & collective communication
- MPI-2 - important features
  - One-sided communication
  - Data types
  - MPI-I/O

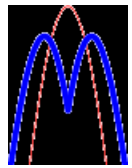
MPI in the Industry

- Aerospace / Defense
- Appliances
- Automotive
- Biomedical & Healthcare
- Buildings
- Chemical Process
- Consumer Packaged Goods
- Data Centers
- Electronic Cooling
- Environmental
- Food & Beverage
- Fuel Cells
- Glass
- HVAC&R
- Marine & Off-shore
- Mixing
- Nonwoven Materials
- Nuclear Power
- Oil & Gas
- Polymer Processing
- Power Generation
- Pumps
- Protein
- Semiconductors/MEMS
- Sport & Athletic Equipment
- Steel
- Turbomachinery

Cluster Ready Applications
Accelerating Market Growth

Cluster-Ready Application Providers





Microsoft targets HPC

Microsoft Windows Server 2003 Compute Cluster Edition

"If you look at the classic adoption curve of new technologies [an S-curve], I would say we think high-performance clusters as an approach are hitting the knee of the curve there"

"Windows Server 2003 HPC Edition, will integrate Microsoft's Windows Server with other software that is considered standard in HPC, including a cluster manager, a scheduler and an implementation of the Message Passing Interface protocol "

Dennis Oldroyd, director of Microsoft's Windows Server Group

Datacenter Fabric Workshop
MPICH2 Windows

MPICH2

MPI-2 Implementation over InfiniBand

August 22, 2005

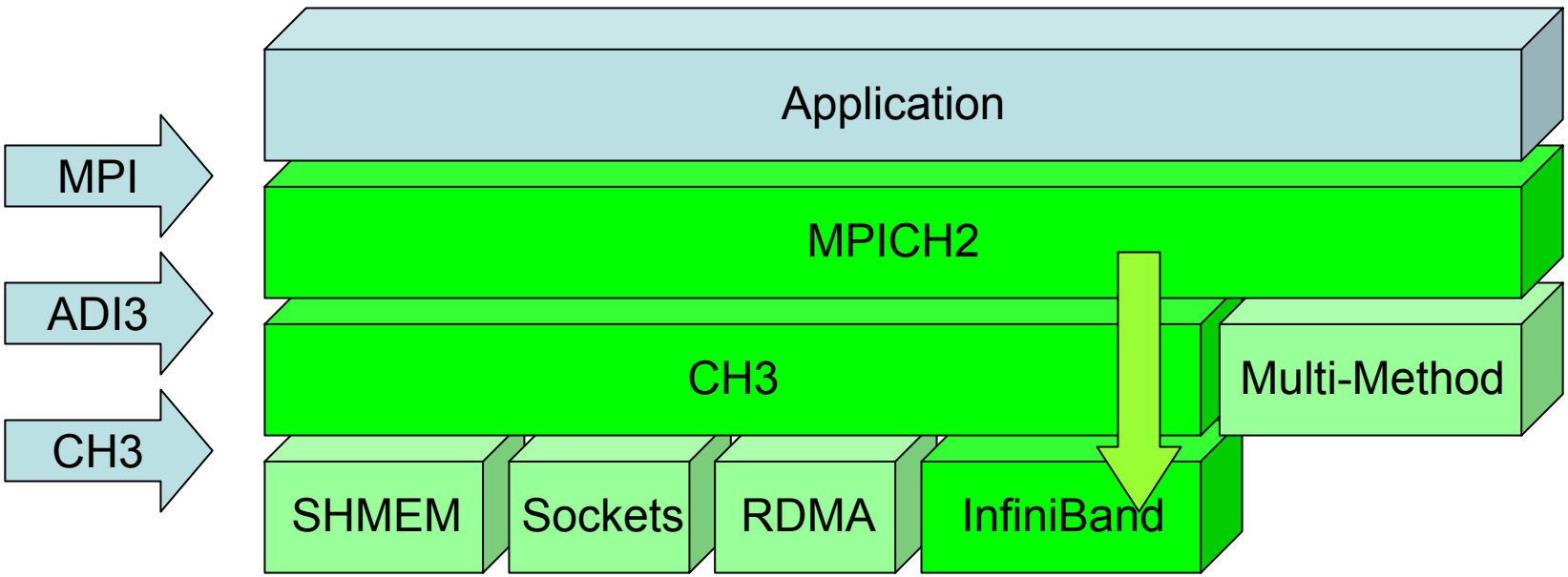
MPICH2 Overview

- Successor of MPICH
  - One of the most popular open source MPI implementations
- Portable, high performance
- Supports both MPI-1 and MPI-2 standards
- Totally new design
  - *Cleaner*
  - *More flexible*
  - *Faster*
  - *Better internal instrumentation*

MPICH2 over InfiniBand

- Implemented by Argonne National Lab
  - Channel Interface (CH3)
- Enhanced by Mellanox MPI team
  - CH3 InfiniBand on top of the access layer
  - For better performance and scalability
- Two sided point-to-point communication using RDMA
- Runtime tunable parameters

MPICH2 Implementation



Enhancements

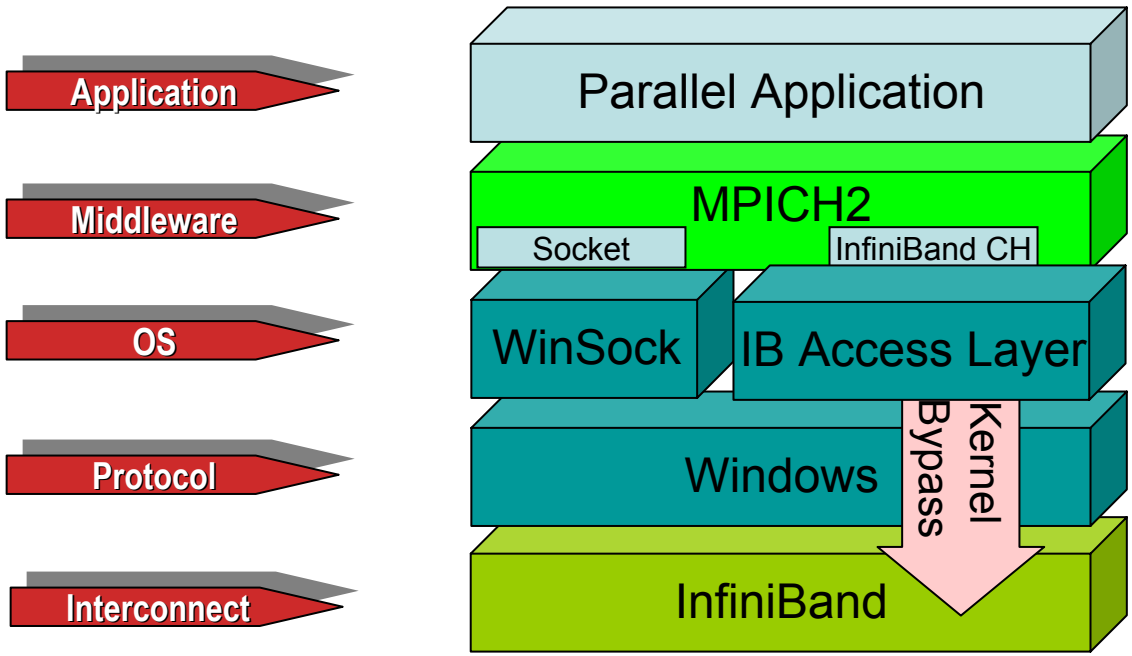
Benefit	Implementation
Reduced Latency	Eager RDMA protocol
Overlapping Computation & networking	Zero copy Rendezvous protocol using RDMA
Memory Scalability	Large message segmentation

MPICH2 cont.

- Collectives
  - On top of point-to-point communication
- One Sided communication (get / put / accumulate)
  - On top of two-sided communication

MPI over IB Channel

MPICH2 over native InfiniBand



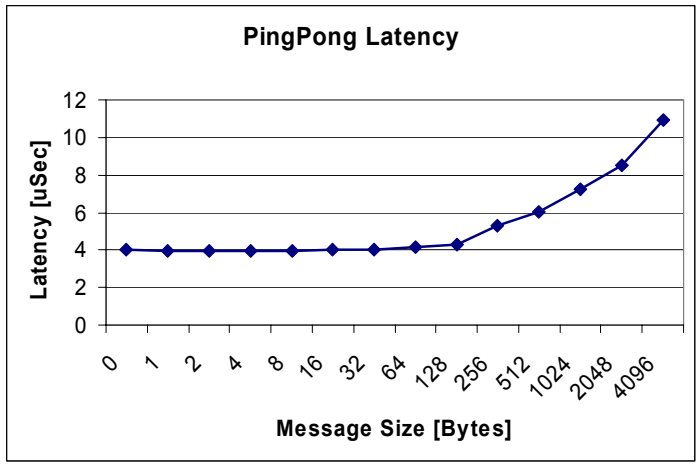
Pallas Benchmark

Performance

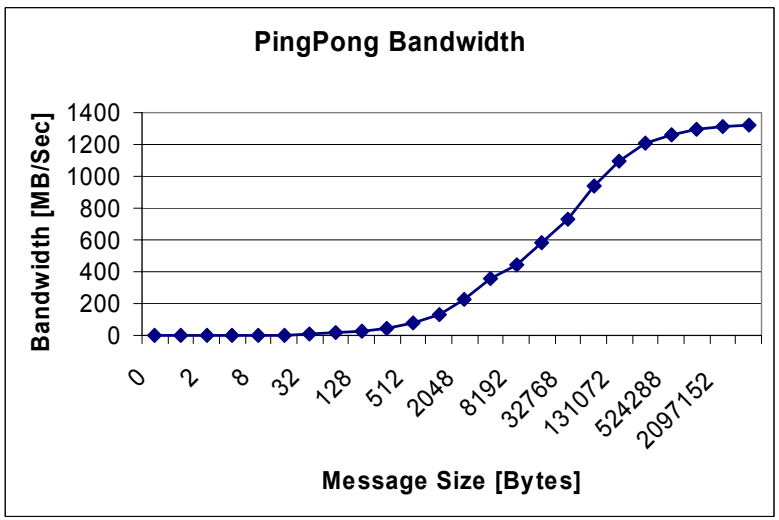
4 uSec — **Ping latency**

Performance Results

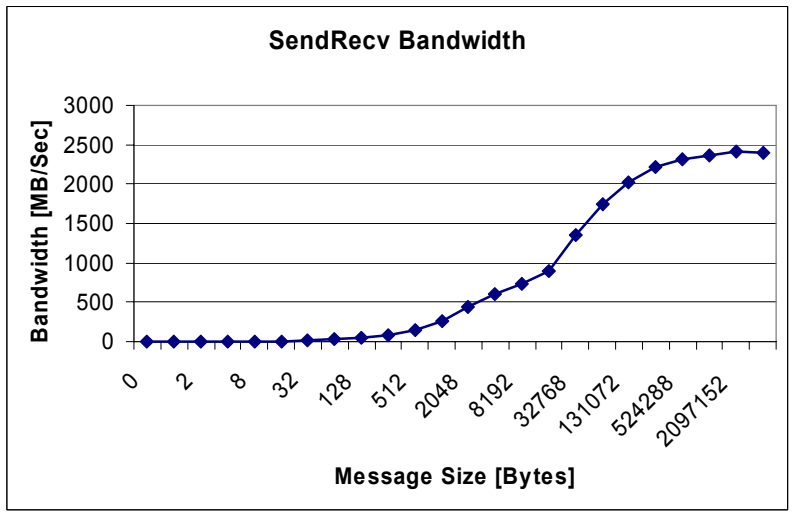
Sub 4us latency



>1.3GB Ping BW



>2.4GB Ping SendRecv



Future Work

- Parameter Tuning
- Collectives Optimizations
  - Hardware multicast
  - Atomic operations
  - Collectives over RDMA
- One-Sided Communication
  - RDMA
- Shared Memory
- Registration Cache
- Stabilization and testing

Resources

- OpenIB WiKi
  - https://openib.org/tiki/tiki-index.php?page=OpenIB+Windows
- Openib-windows mailing list
  - http://openib.org/mailman/listinfo/openib-windows
- Sign up to contribute
  - http://windows.openib.org/openib/contribute.aspx

Q & A