Designing Next Generation Clusters with InfiniBand and 10GE/iWARP: Opportunities and Challenges

Invited Talk at Cluster '08

by

Dhabaleswar K. (DK) Panda

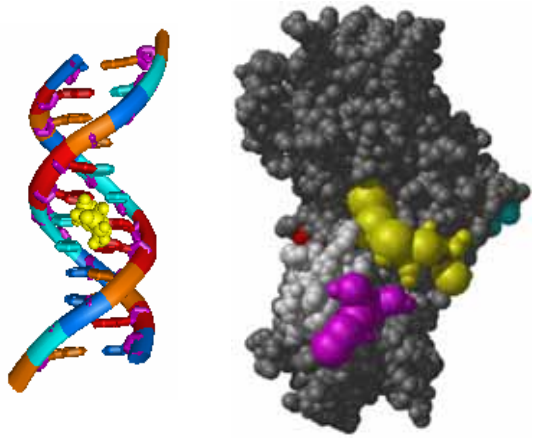
The Ohio State University

E-mail: panda@cse.ohio-state.edu

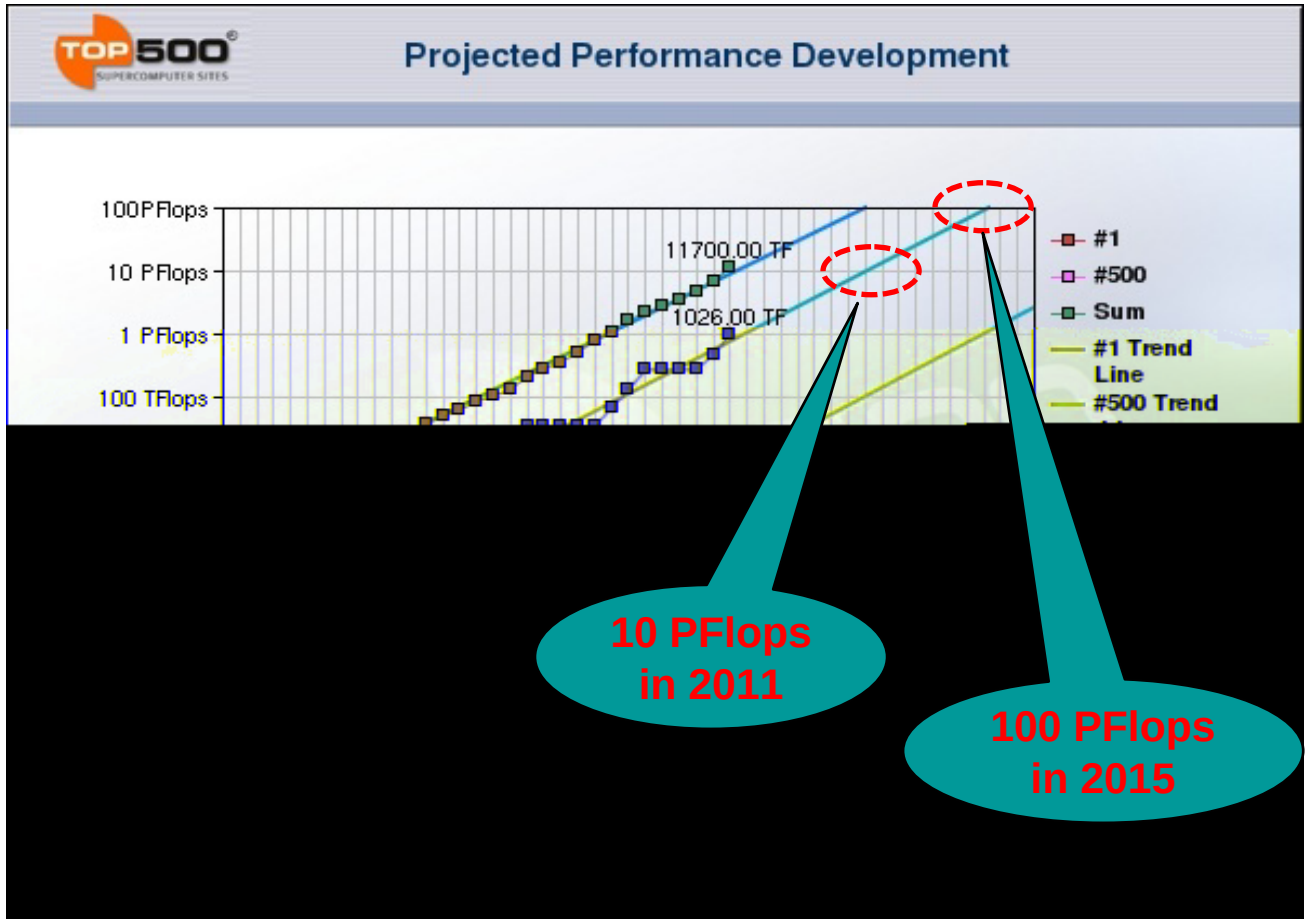
http://www.cse.ohio-state.edu/~panda

Current and Next Generation Applications and Computing Systems

- Big demand for
  - High Performance Computing (HPC)
  - File-systems, multimedia, database, visualization
  - Internet data-centers
- Processor performance continues to grow
  - Chip density doubling every 18 months
  - Multi-core chips are emerging
- Commodity networking also continues to grow
  - Increase in speed and features
  - Affordable pricing
- Clusters are increasingly becoming popular to design next generation computing systems
  - Scalability, Modularity and Upgradeability with compute and network technologies



PetaFlop to ExaFlop Computing



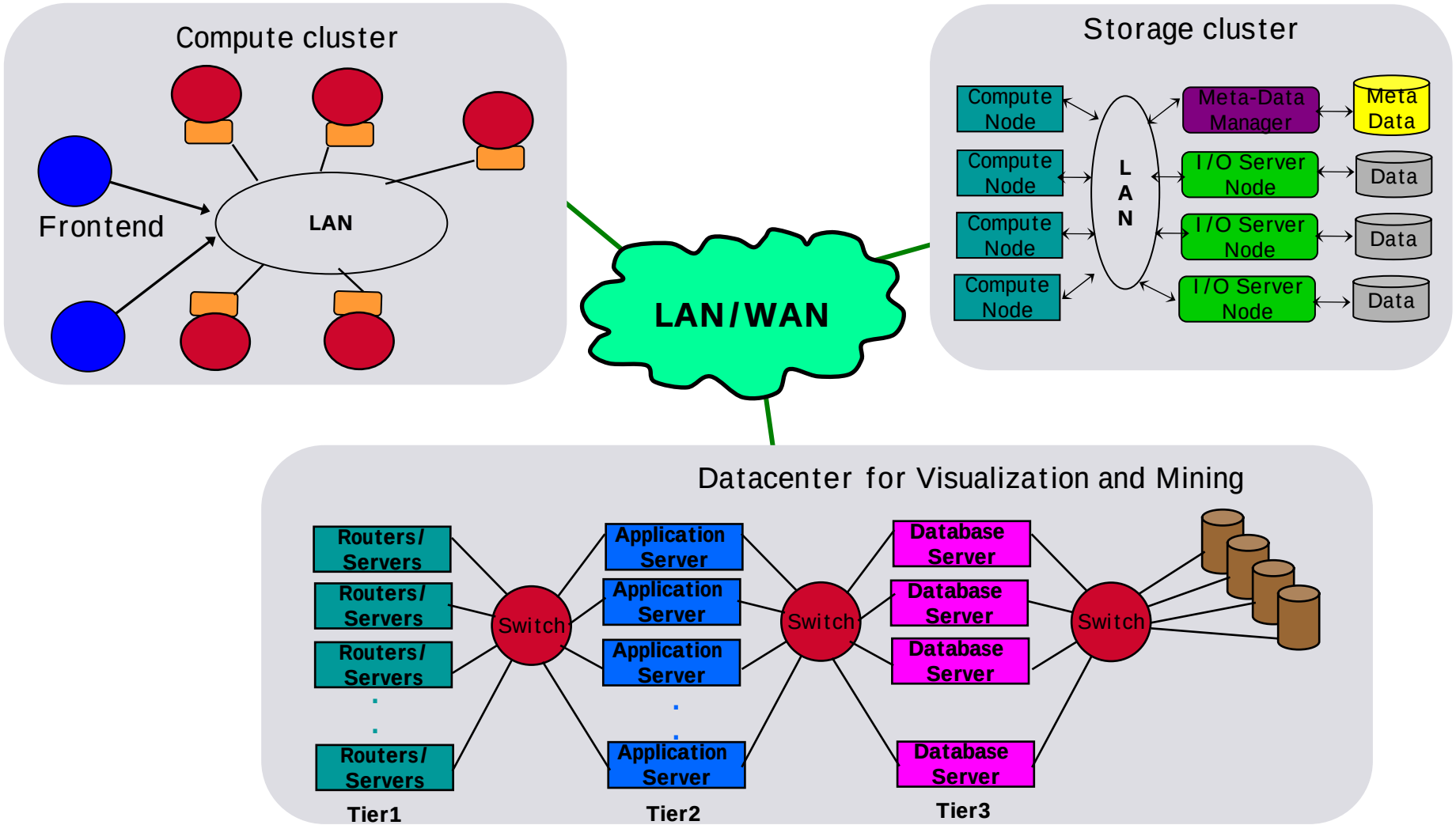
Expected to have an ExaFlop system in 2018-2019 !

Trends for Computing Clusters in the Top 500 List

- Top 500 list of Supercomputers (www.top500.org)

June 2001: 33/500 (6.6%)	June 2005: 304/500 (60.8%)
Nov 2001: 43/500 (8.6%)	Nov 2005: 360/500 (72.0%)
June 2002: 80/500 (16%)	June 2006: 364/500 (72.8%)
Nov 2002: 93/500 (18.6%)	Nov 2006: 361/500 (72.2%)
June 2003: 149/500 (29.8%)	June 2007: 373/500 (74.6%)
Nov 2003: 208/500 (41.6%)	Nov 2007: 406/500 (81.2%)
June 2004: 291/500 (58.2%)	June 2008: 400/500 (80.0%)
Nov 2004: 294/500 (58.8%)	Nov 2008: To be Announced

Integrated Environment with Multiple Clusters



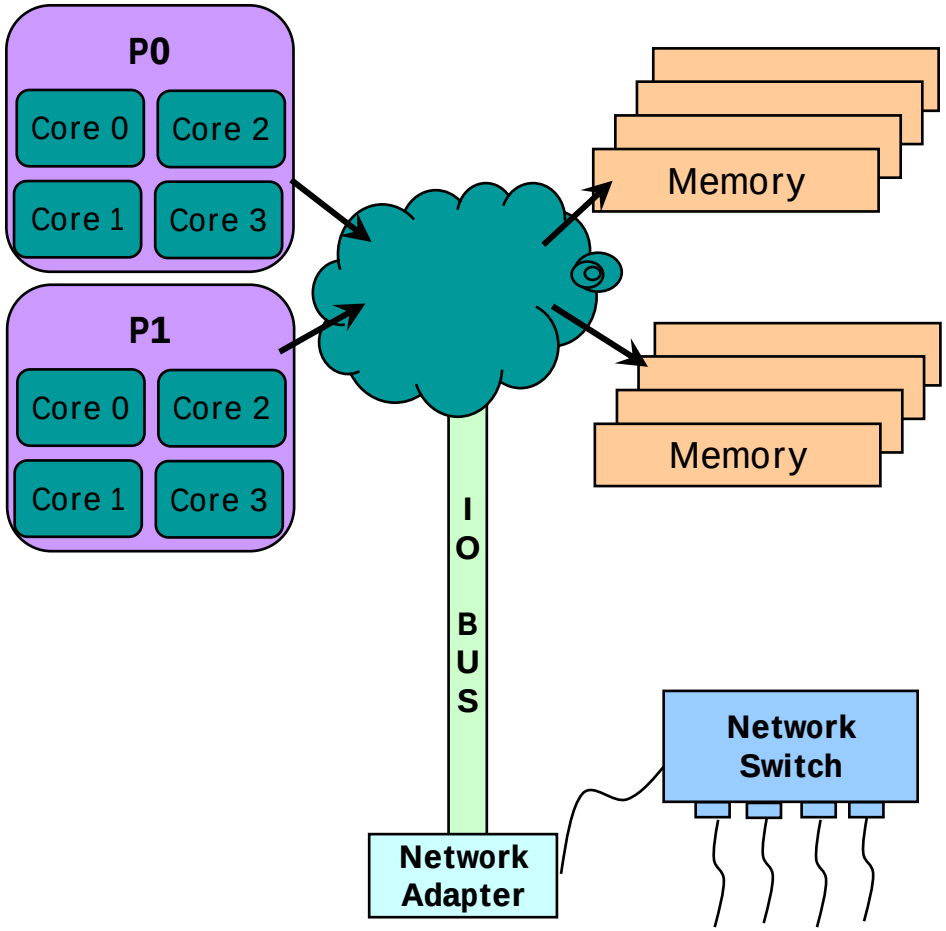
Networking and I/O Requirements

- Good Systems Area Network with excellent performance (low latency and high bandwidth) for inter-processor communication (IPC) and I/O
- Good Storage Area Networks high performance I/O
- Good WAN connectivity in addition to intra-cluster SAN/LAN connectivity
- Quality of Service (QoS) for interactive applications
- RAS (Reliability, Availability, and Serviceability)
- With low cost

Presentation Overview

- Trends in Networking Technologies
- Overview of InfiniBand and Its Features
- Overview of 10GigE/iWARP and Its Features
- Performance on Current Clusters and Trends
- Challenges
- Conclusions and Q&A

Major Components in Computing Systems



- Hardware Components
  - Processing Core and Memory sub-system
  - I/O Bus (PCI-X, PCIe, HT)
  - Network Adapter (InfiniBand, 10GE)
  - Network Switch (InfiniBand, 10GE)
- Software Components
  - Communication software

Processing Units

- Multi-processor systems have existed for many years
- Multi-core processors have also started coming into the market
- Quad-core processors are considered “commodity”
- Many-core processors upcoming
  - Intel planning to release an 80-core processor by 2011

Growth in Commodity Network Technology

Representative commodity networks; their entries into the market

Ethernet (1979 -)	10 Mbit/sec
Fast Ethernet (1993 -)	100 Mbit/sec
Gigabit Ethernet (1995 -)	1000 Mbit /sec
ATM (1995 -)	155/622/1024 Mbit/sec
Myrinet (1993 -)	1 Gbit/sec
Fibre Channel (1994 -)	1 Gbit/sec
InfiniBand (2001 -)	2 Gbit/sec (1X SDR)
10-Gigabit Ethernet (2001 -)	10 Gbit/sec
InfiniBand (2003 -)	8 Gbit/sec (4X SDR)
InfiniBand (2005 -)	16 Gbit/sec (4X DDR)
	24 Gbit/sec (12X SDR)
InfiniBand (2007 -)	32 Gbit/sec (4X QDR)

16 times in the last 7 years

Trends in I/O Interfaces with Servers

- Network performance depends on
  - Networking technology (adapter + switch)
  - Network interface (last mile bottleneck)

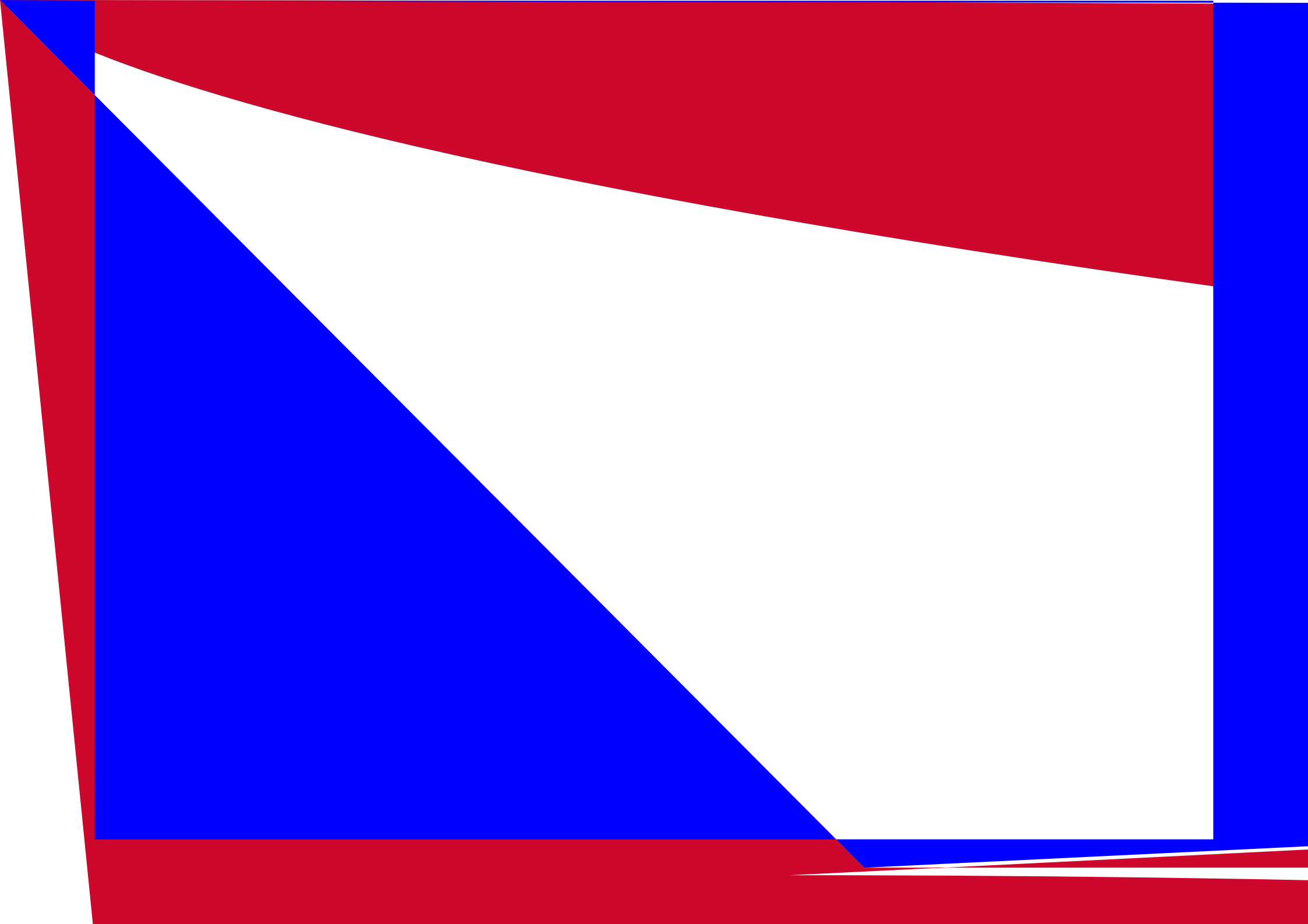
PCI	1990	33MHz/32bit: 1.05Gbps (shared bidirectional)
PCI-X	1998 (v1.0) 2003 (v2.0)	133MHz/64bit: 8.5Gbps (shared bidirectional) 266-533MHz/64bit: 17Gbps (shared bidirectional)
HyperTransport (HT) by AMD	2001 (v1.0), 2004 (v2.0) 2006 (v3.0), 2008 (v3.1)	102.4Gbps (v1.0), 179.2Gbps (v2.0) 332.8Gbps (v3.0), 409.6Gbps (v3.1)
PCI-Express (PCIe) by Intel	2003 (Gen1) 2007 (Gen2)	Gen1: 4X (8Gbps), 8X (16Gbps), 16X (32Gbps) Gen2: 4X (16Gbps), 8X (32Gbps), 16X (64Gbps)
PCIe Gen3		

Limitations of Traditional Host-based Protocols

- Ex: TCP/IP, UDP/IP
- Generic architecture for all network interfaces
- Host-handles almost all aspects of communication
  - Data buffering (copies on sender and receiver)
  - Data integrity (checksum)
  - Routing aspects (IP routing)
- Signaling between different layers
  - Hardware interrupt whenever a packet arrives or is sent
  - Software signals between different layers to handle protocol processing in different priority levels

Previous High Performance Network Stacks

- Virtual Interface Architecture
  - Standardized by Intel, Compaq, Microsoft
- Fast Messages (FM)
  - Developed by UIUC
- Myricom GM
  - Proprietary protocol stack from Myricom
- These network stacks set the trend for high-performance communication requirements
  - Hardware offloaded protocol stack
  - Support for fast and secure user-level access to the protocol stack



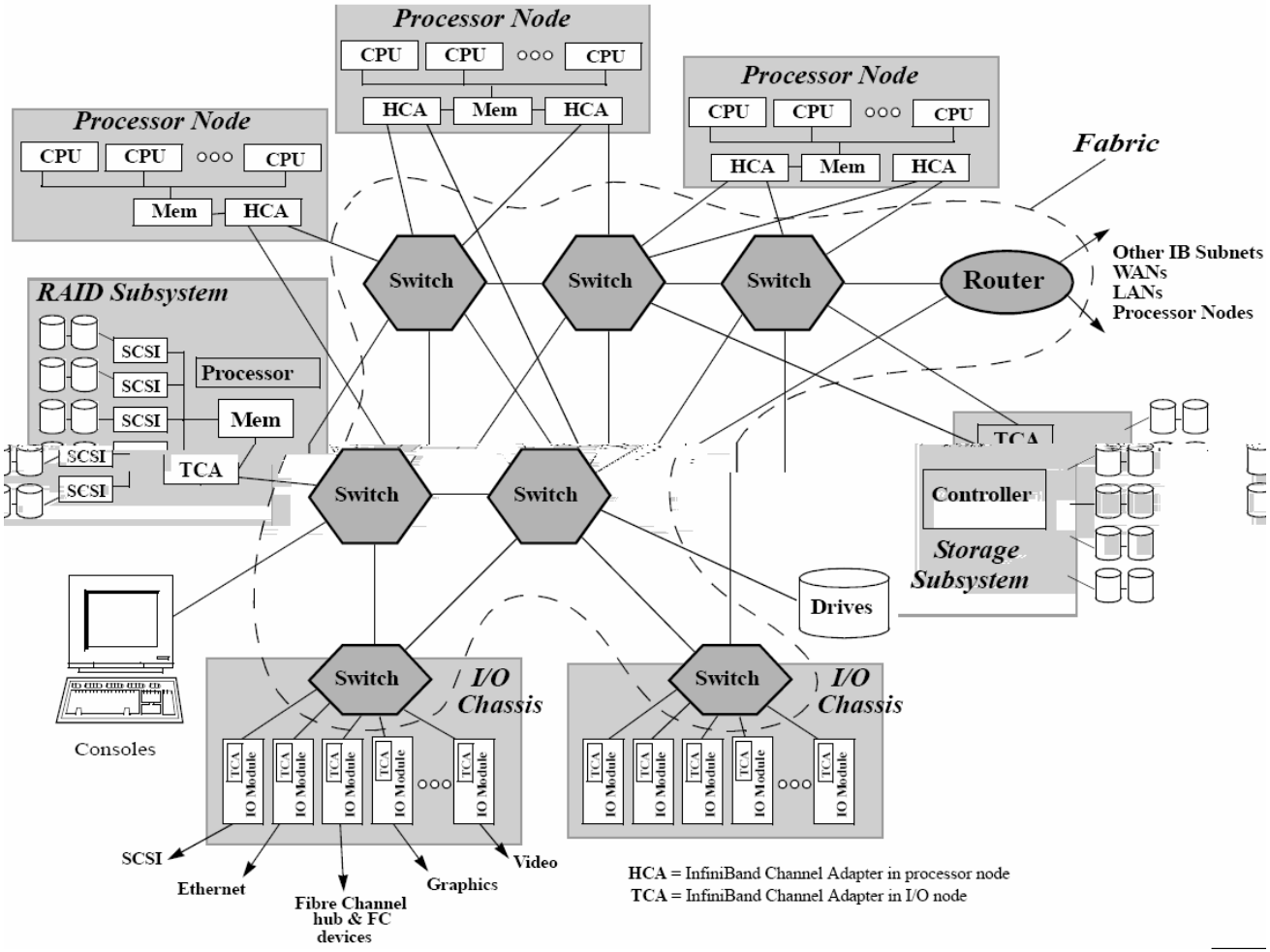
10-Gigabit Ethernet Consortium

- 10GE Alliance formed by several industry leaders to take the Ethernet family to the next speed step
- Goal: To achieve a scalable and high performance communication architecture while maintaining backward compatibility with Ethernet
- http://www.ethernetalliance.org
- Upcoming 40-Gbps (Servers) and 100-Gbps Ethernet (Backbones, Switches, Routers): IEEE 802.3 WG
- Energy-efficient and power-conscious protocols
  - On-the-fly link speed reduction for under-utilized links

Presentation Overview

- Trends in Networking Technologies
- Overview of InfiniBand and Its Features
- Overview of 10GigE/iWARP and Its Features
- Performance on Current Clusters and Trends
- Challenges
- Conclusions and Q&A

A Typical IB Network



Three primary components

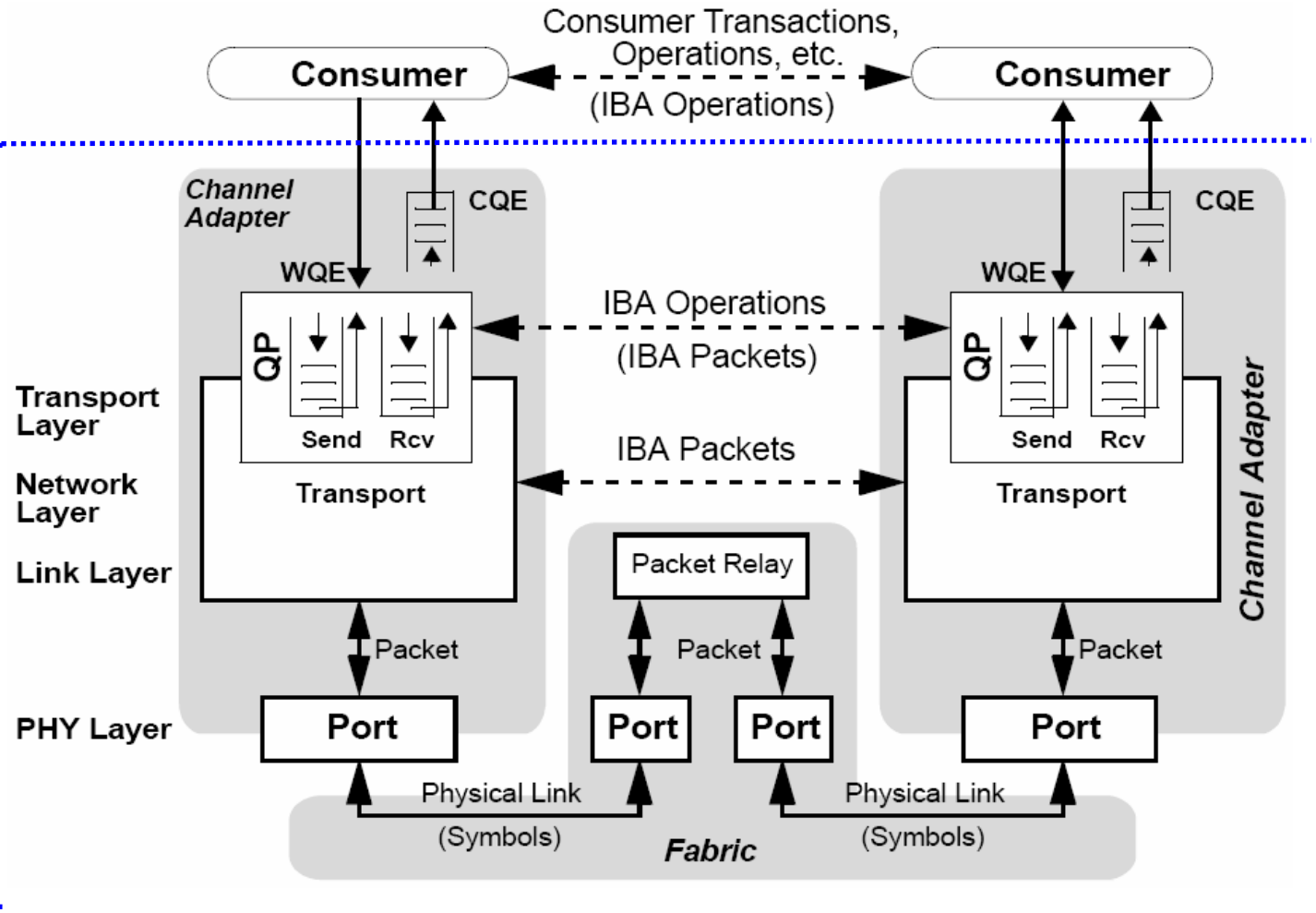
Channel Adapters

Switches/Routers

Links and connectors

Hardware Protocol Offload

Complete Hardware Implementations Exist



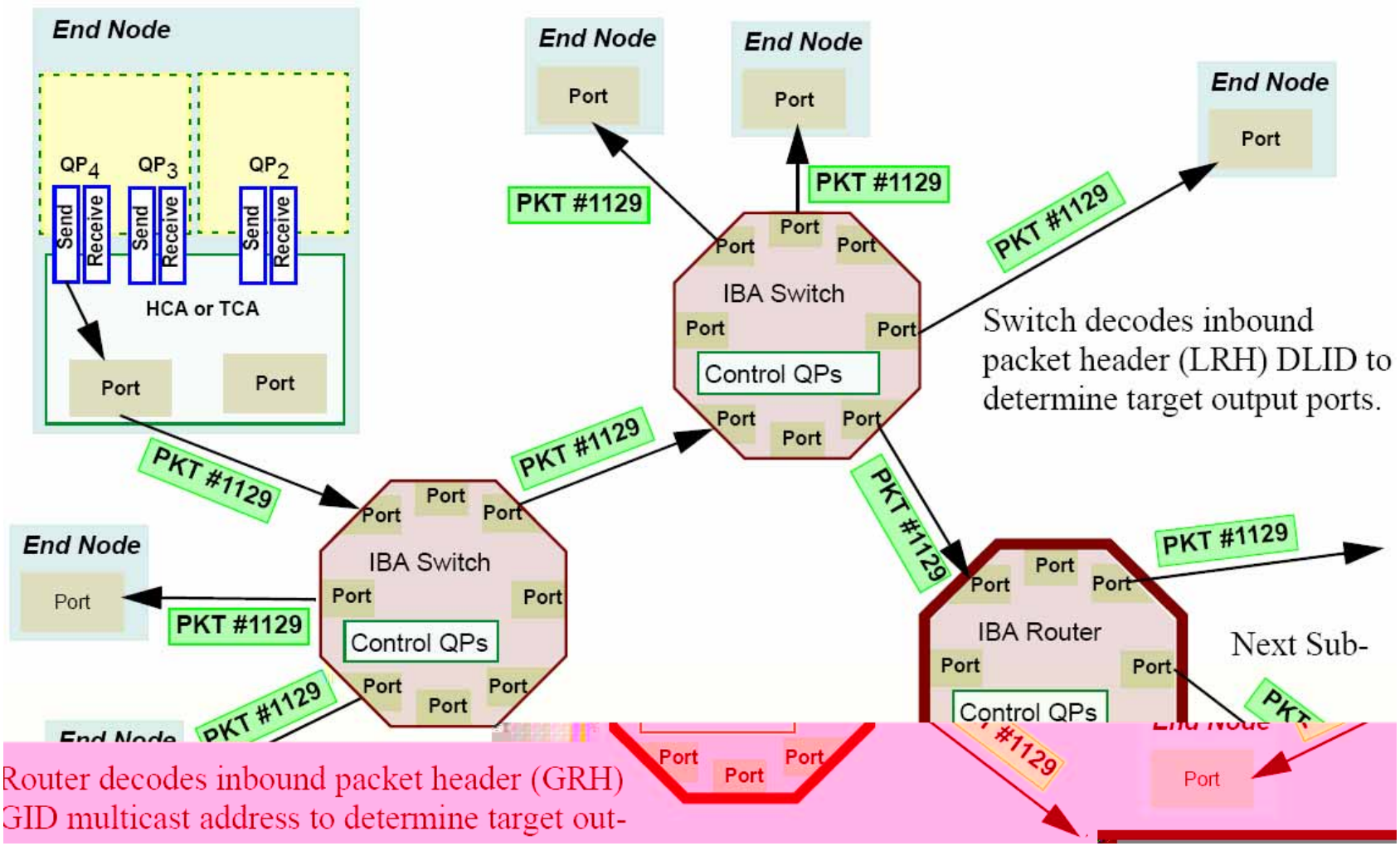
Basic IB Capabilities at Each Protocol Layer

- Link Layer
  - CRC-based data integrity, Buffering and Flow-control, Virtual Lanes, Service Levels and QoS, Switching and Multicast, WAN capabilities
- Network Layer
  - Routing and Flow Labels
- Transport Layer
  - Reliable Connection, Unreliable Datagram, Reliable Datagram and Unreliable Connection
  - Shared Receive Queued and Extended Reliable Connections

Communication and Management Semantics

- Two forms of communication semantics
  - Channel semantics (Send/Recv)
  - Memory semantics (RDMA, Atomic operations)
- Management model
  - A detailed management model complete with managers, agents, messages and protocols
- Verbs Interface
  - A low-level programming interface for performing communication as well as management

IB Multicast Example



Service Levels and QoS

- Service Level (SL):
  - Packets may operate at one of 16 different SLs
  - Meaning not defined by IB
- SL to Virtual Lane (VL) mapping:
  - SL determines which VL on the next link is to be used
  - Each port (switches, routers, end nodes) has a SL to VL mapping table configured by the subnet management
  - Similar to DiffServ model in Internet
- Allows sharing of different traffic (communication and storage), even from different applications, on the same physical links with different QoS

IB WAN Capability

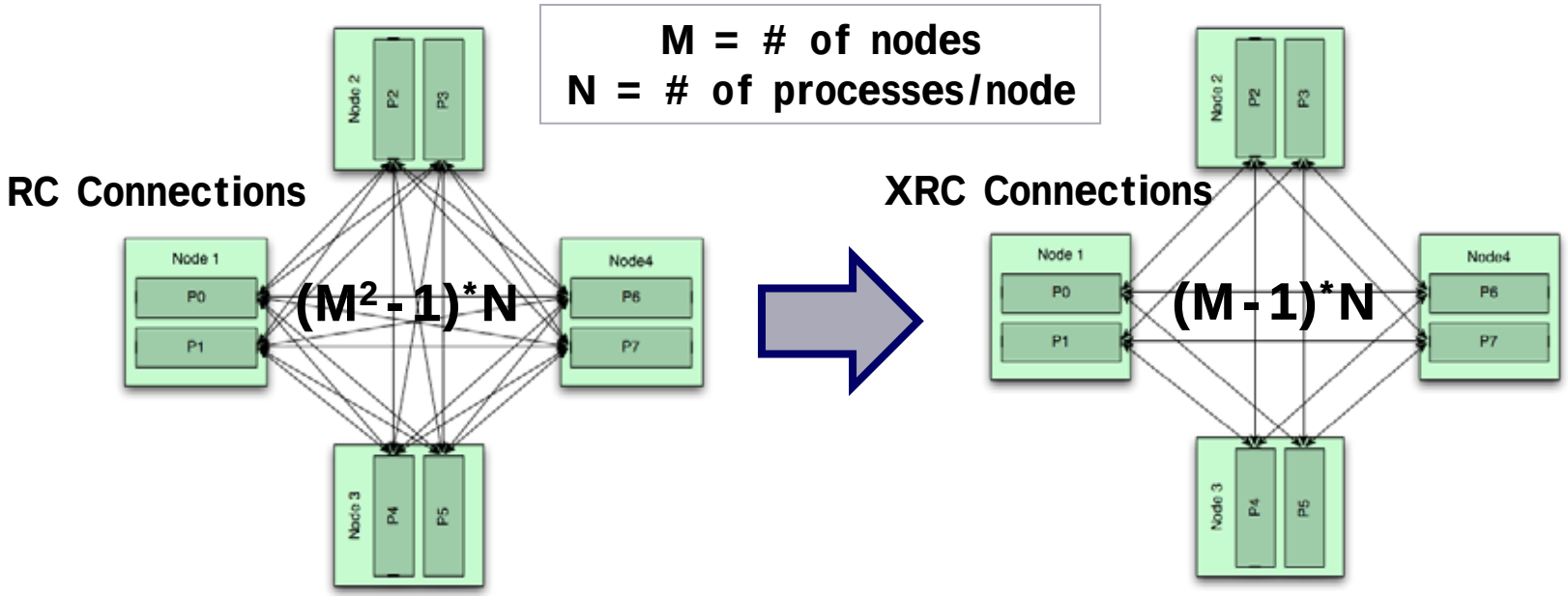
- Getting increased attention for:
  - Remote Storage, Remote Visualization
  - Cluster Aggregation (Cluster-of-clusters)
- IB-Optical switches by multiple vendors
  - Obsidian Research Corporation: www.obsidianresearch.com
  - Network Equipment Technology (NET): www.net.com
  - Layer-1 changes from copper to optical; everything else stays the same
    - Low-latency copper-optical-copper conversion

IB Transport Services

Service Type	Connection Oriented	Acknowledged	Transport
Reliable Connection	yes	Yes	IBA
Unreliable Connection	yes	no	IBA
Reliable Datagram	no	Yes	IBA
Unreliable Datagram	no	no	IBA
RAW Datagram	no	no	Raw

A new transport eXtended Reliable Connection (XRC) is introduced recently

eXtended Reliable Connection (XRC)



- New IB Transport added: eXtended Reliable Connection
  - Allows connections between nodes instead of processes

Communication in the Channel Semantics (Send-Receive Model)

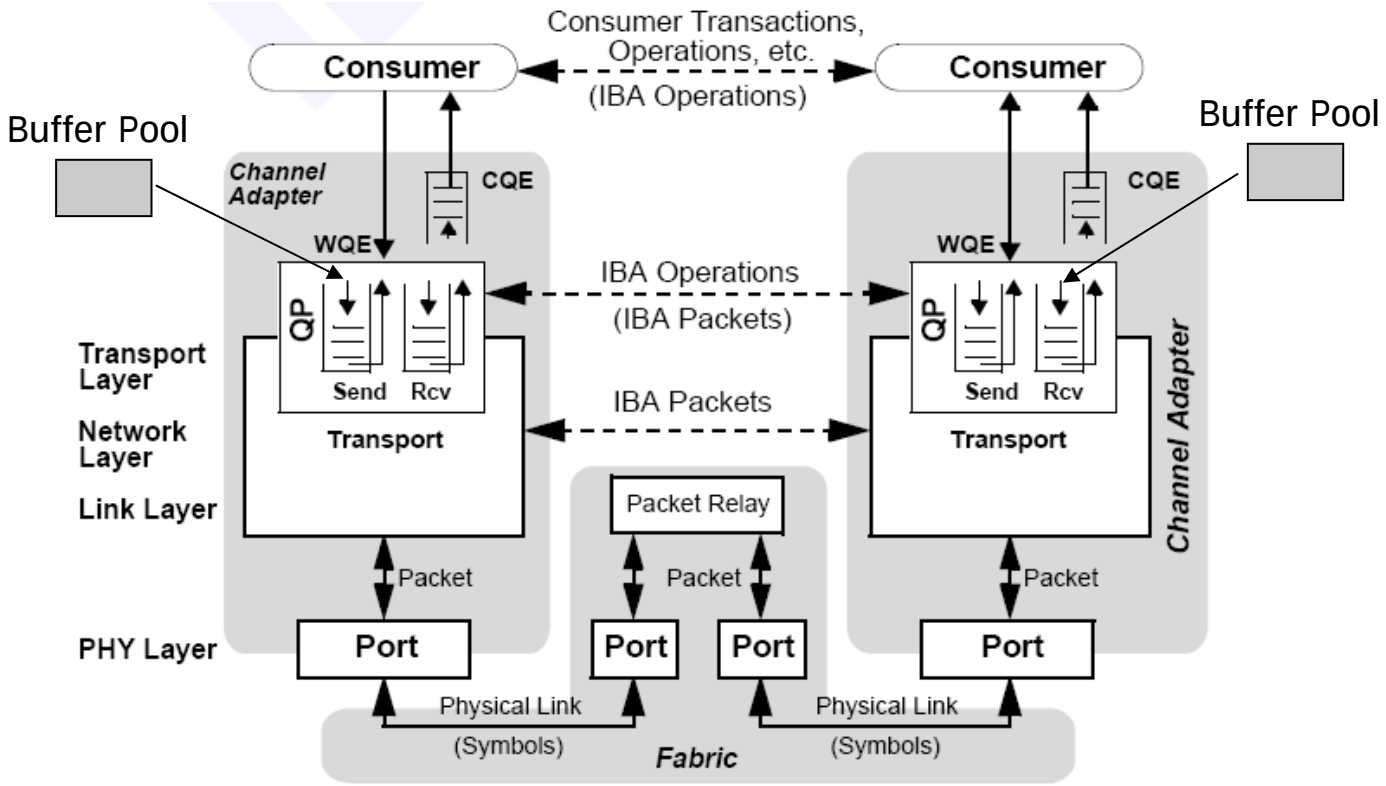
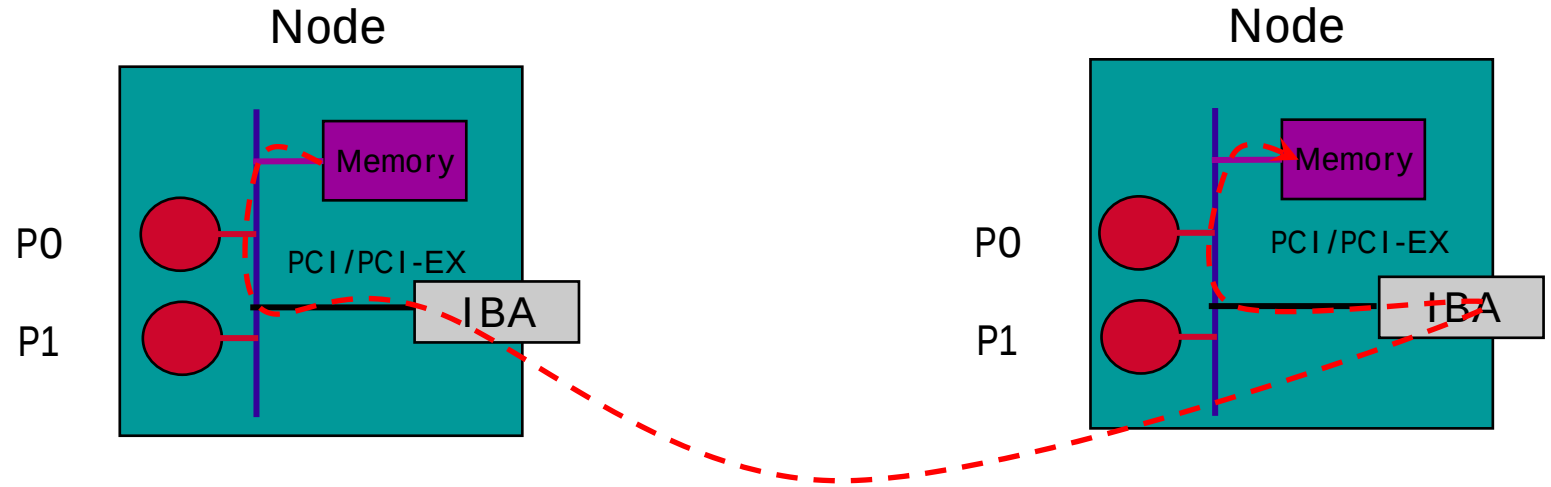


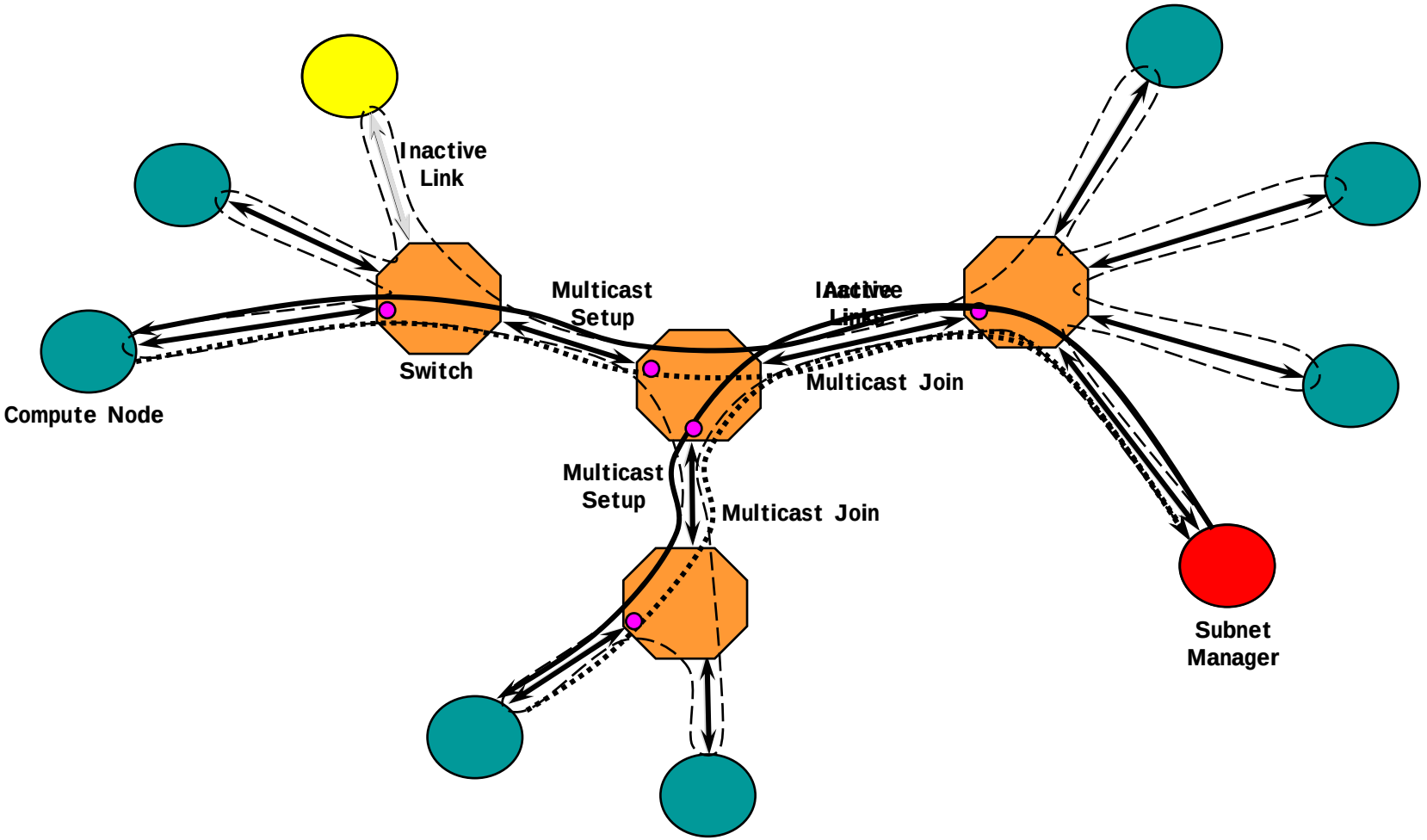
Figure 13 IBA Communication Stack

Communication in the Memory Semantics (RDMA Model)



- No involvement by the CPU at the receiver (RDMA Write/Put)
- No involvement by the CPU at the sender (RDMA Read/get)
- 1-2 µs latency (for short data)
- 1.5 – 2.6 GBps bandwidth (for large data)
- 3-5 µs for atomic operation

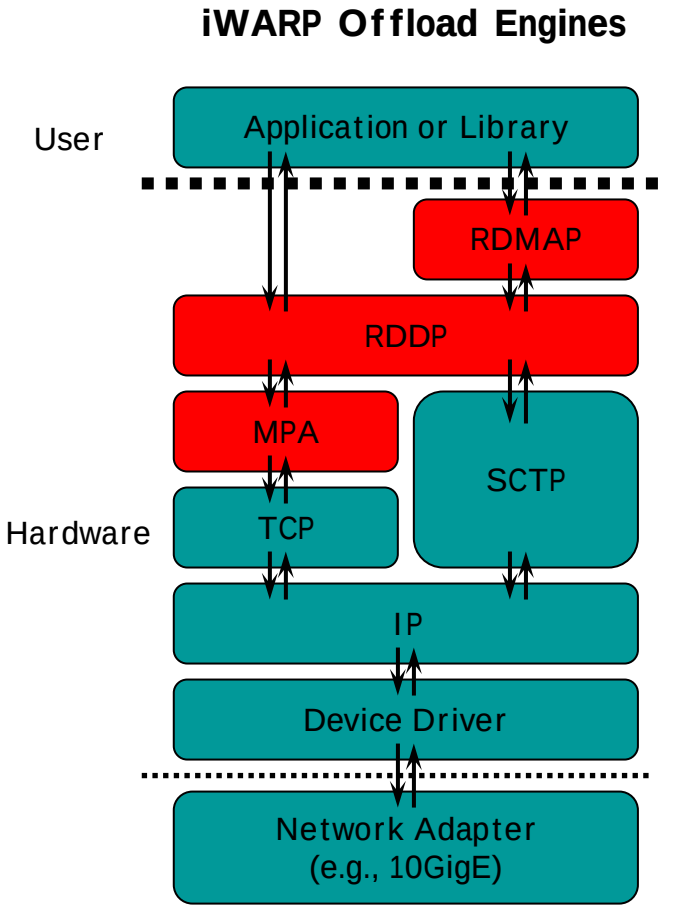
Subnet Manager



Presentation Overview

- Trends in Networking Technologies
- Overview of InfiniBand and Its Features
- Overview of 10GigE/iWARP and Its Features
- Performance on Current Clusters and Trends
- Challenges
- Conclusions and Q&A

iWARP Architecture and Components



(Courtesy iWARP Specification)

- RDMA Protocol (RDMAP)
  - Feature-rich interface
  - Security Management
- Remote Direct Data Placement (RDDP)
  - Data Placement and Delivery
  - Multi Stream Semantics
  - Connection Management

Basic iWARP Capabilities

- Supports most of the communication features supported by IB (with minor differences)
  - Hardware acceleration, RDMA, Multicast, QoS
- Lacks some features
  - E.g., Atomic operations
- … but supports some other features
  - Out-of-Order data placement (useful for iSCSI semantics)
  - Fine-grained data rate control (very useful for long-haul networks)
  - Fixed bandwidth QoS

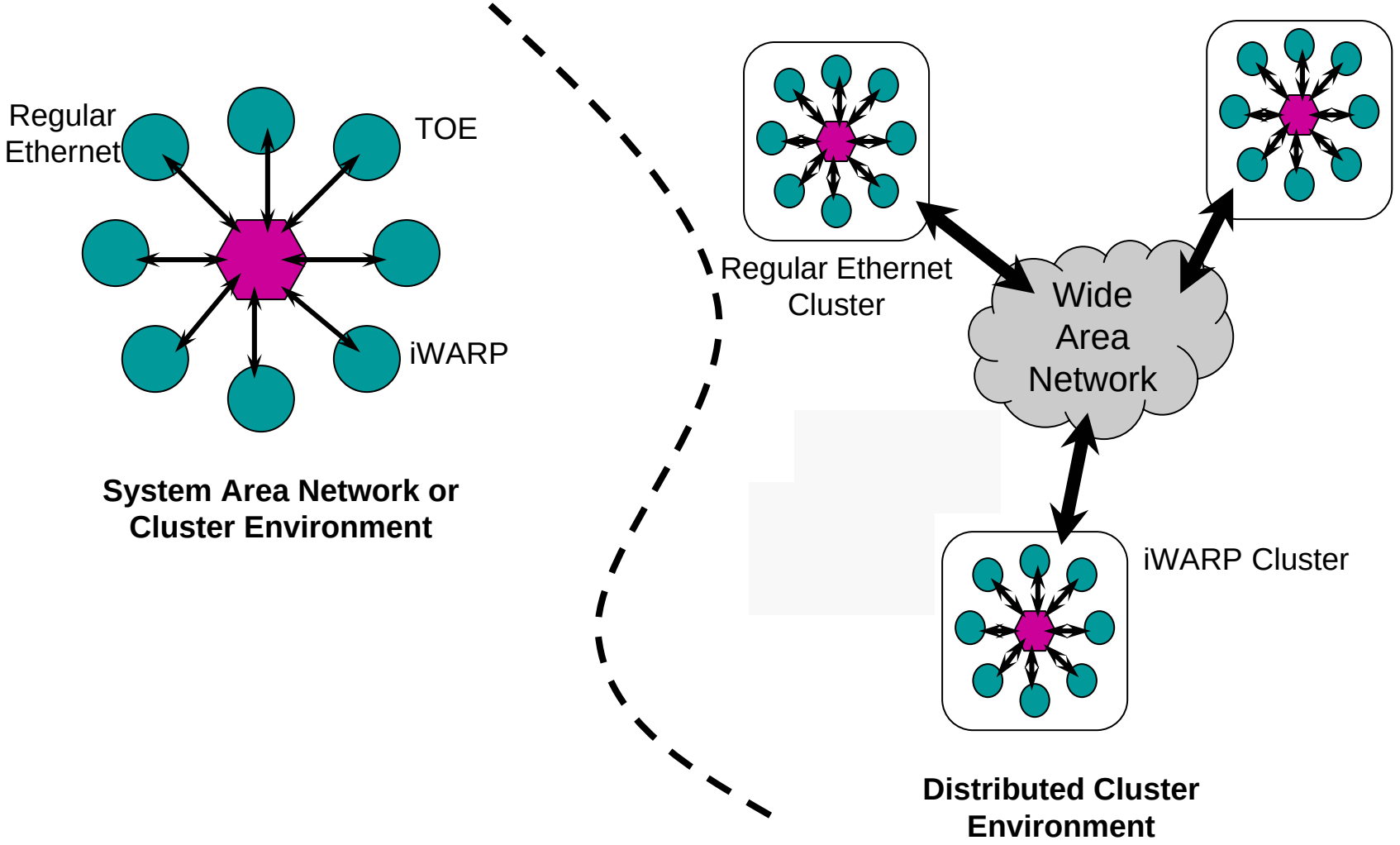
IB and 10GE: Commonalities and Differences

	IB	iWARP/10GE
Hardware Acceleration	Supported	Supported (for TOE and iWARP)
RDMA	Supported	Supported (for iWARP)
Atomic Operations	Supported	Not supported
Multicast	Supported	Supported
Data Placement	Ordered	Out-of-order (for iWARP)
Data Rate-control	Static and Coarse-grained	Dynamic and Fine-grained (for TOE and iWARP)
QoS	Prioritization	Prioritization and Fixed Bandwidth QoS

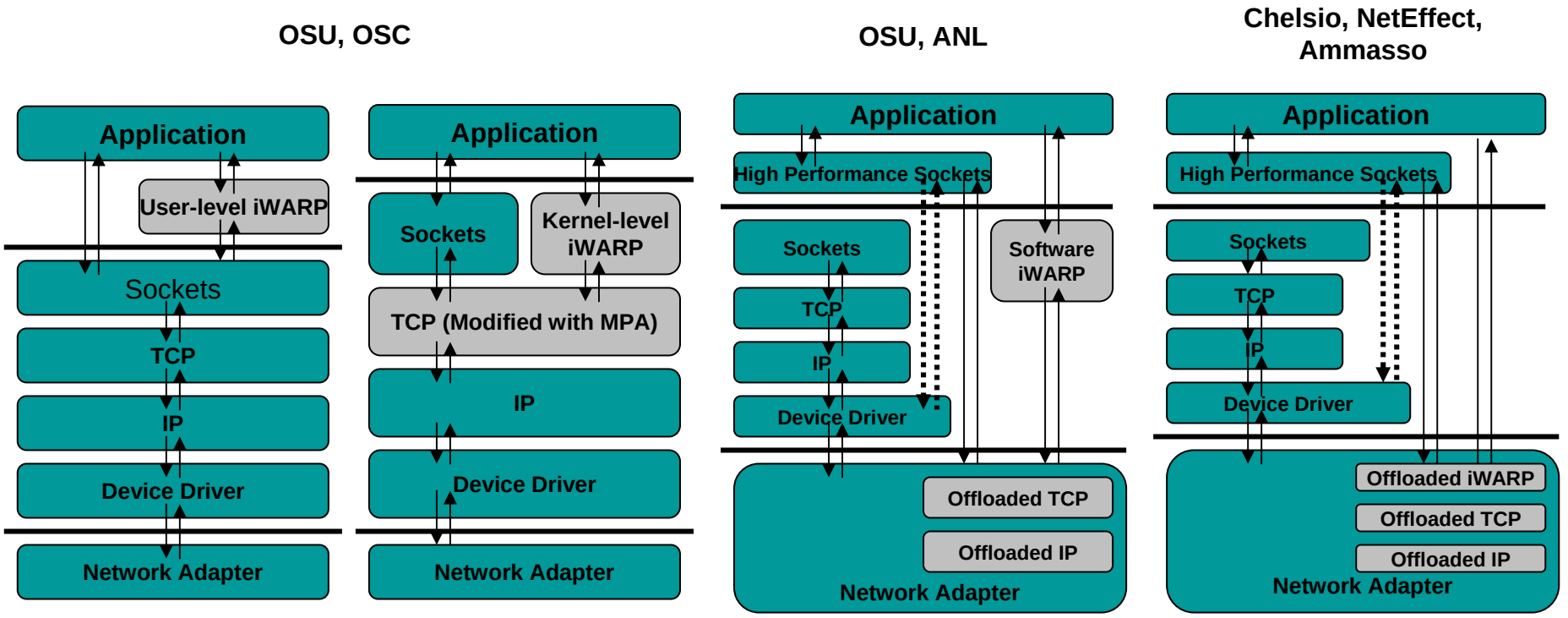
Prioritization vs. Fixed Bandwidth QoS

- Can allow for simple prioritization:
  - E.g., connection 1 performs better than connection 2
  - 8 classes provided (a connection can be in any class)
    - Similar to SLs in InfiniBand
  - Two priority classes for high-priority traffic
    - E.g., management traffic or your favorite application
- Or can allow for specific bandwidth requests:
  - E.g., can request for 3.62 Gbps bandwidth
  - Packet pacing and stalls used to achieve this
- Query functionality to find out "remaining bandwidth"

Current Usage of Ethernet



Different iWARP Implementations



Regular Ethernet Adapters

TCP Offload Engines

iWARP compliant Adapters

Presentation Overview

- Trends in Networking Technologies
- Overview of InfiniBand and Its Features
- Overview of 10GigE/iWARP and Its Features
- Performance on Current Clusters and Trends
- Challenges
- Conclusions and Q&A

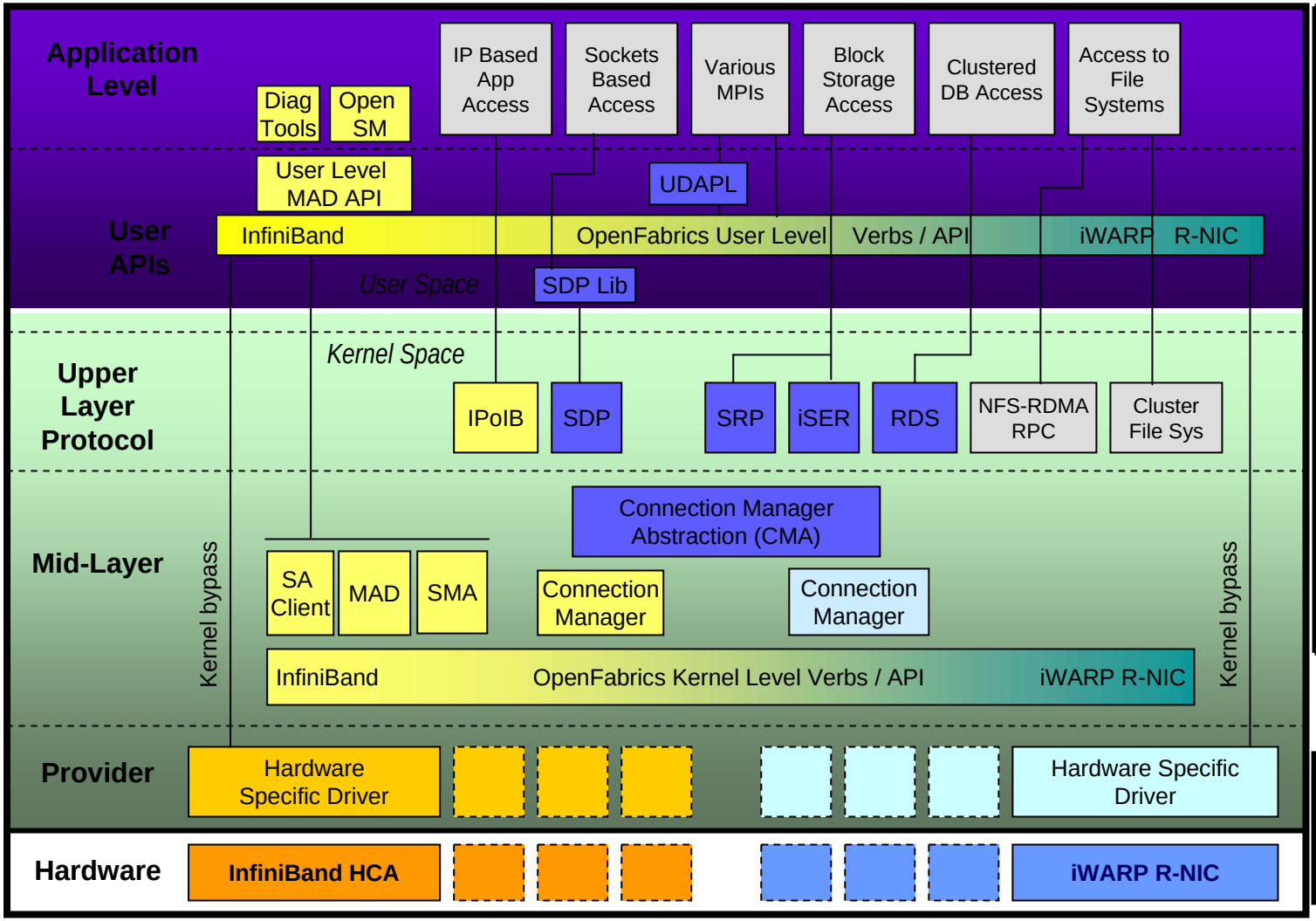
Performance on Current Clusters and Trends

- OpenFabrics Stack
- Sample Performance Numbers
  - MPI
  - File Systems (Lustre and NFS/RDMA)
  - Datacenters
- IB and 10GigE Installations and Trends

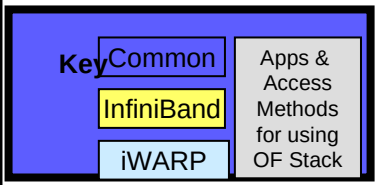
OpenFabrics

- www.openfabrics.org
- Open source organization (formerly OpenIB)
- Incorporates both IB and iWARP in a unified manner
- Focusing on effort for Open Source IBA and iWARP support for Linux and Windows
- Design of complete software stack with `best of breed' components
  - Gen1
  - Gen2 (current focus)
- Users can download the entire stack and run
  - Latest release is OFED 1.3.1
  - OFED 1.4 is being worked out

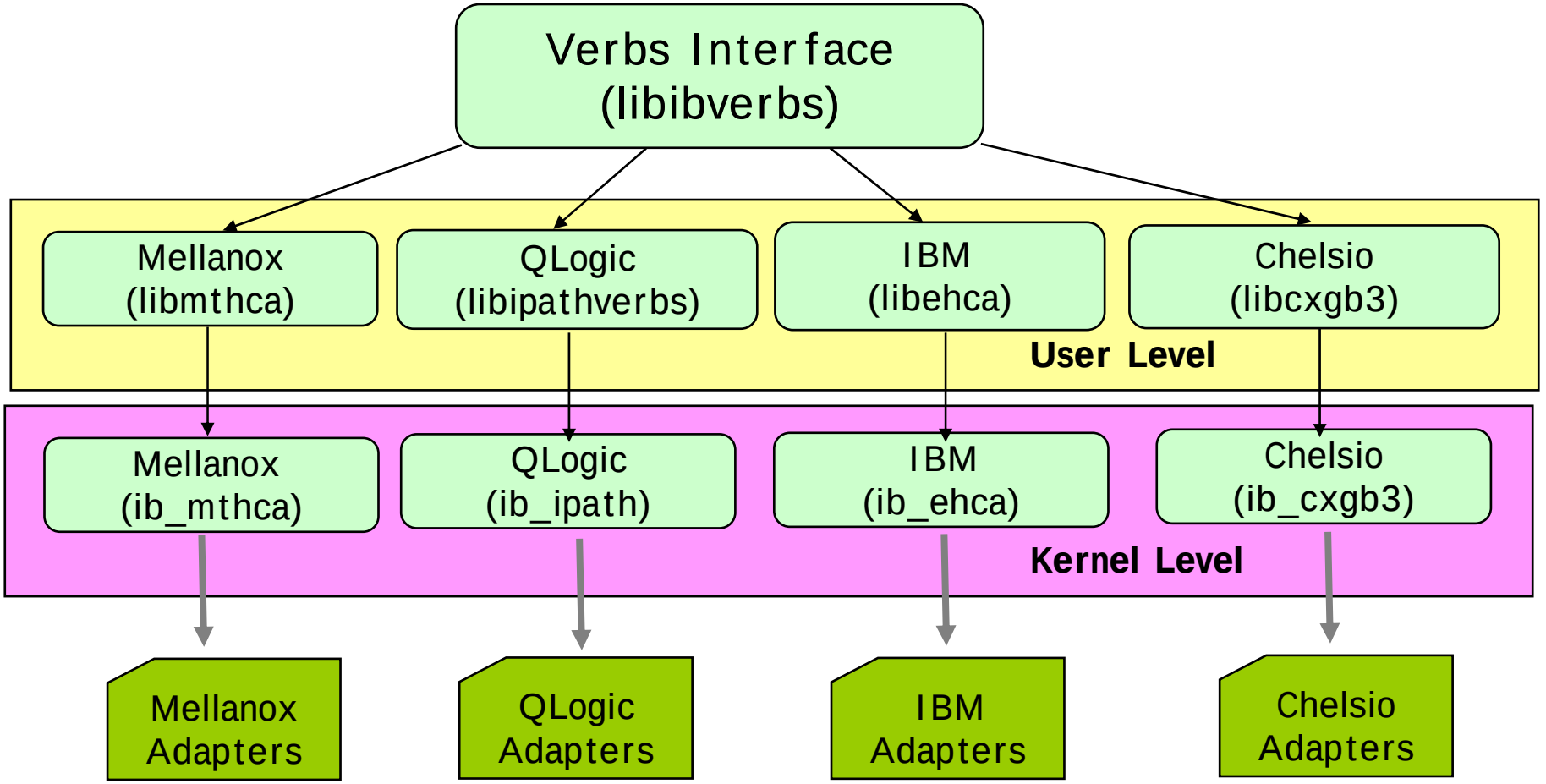
OpenFabrics Software Stack



SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC



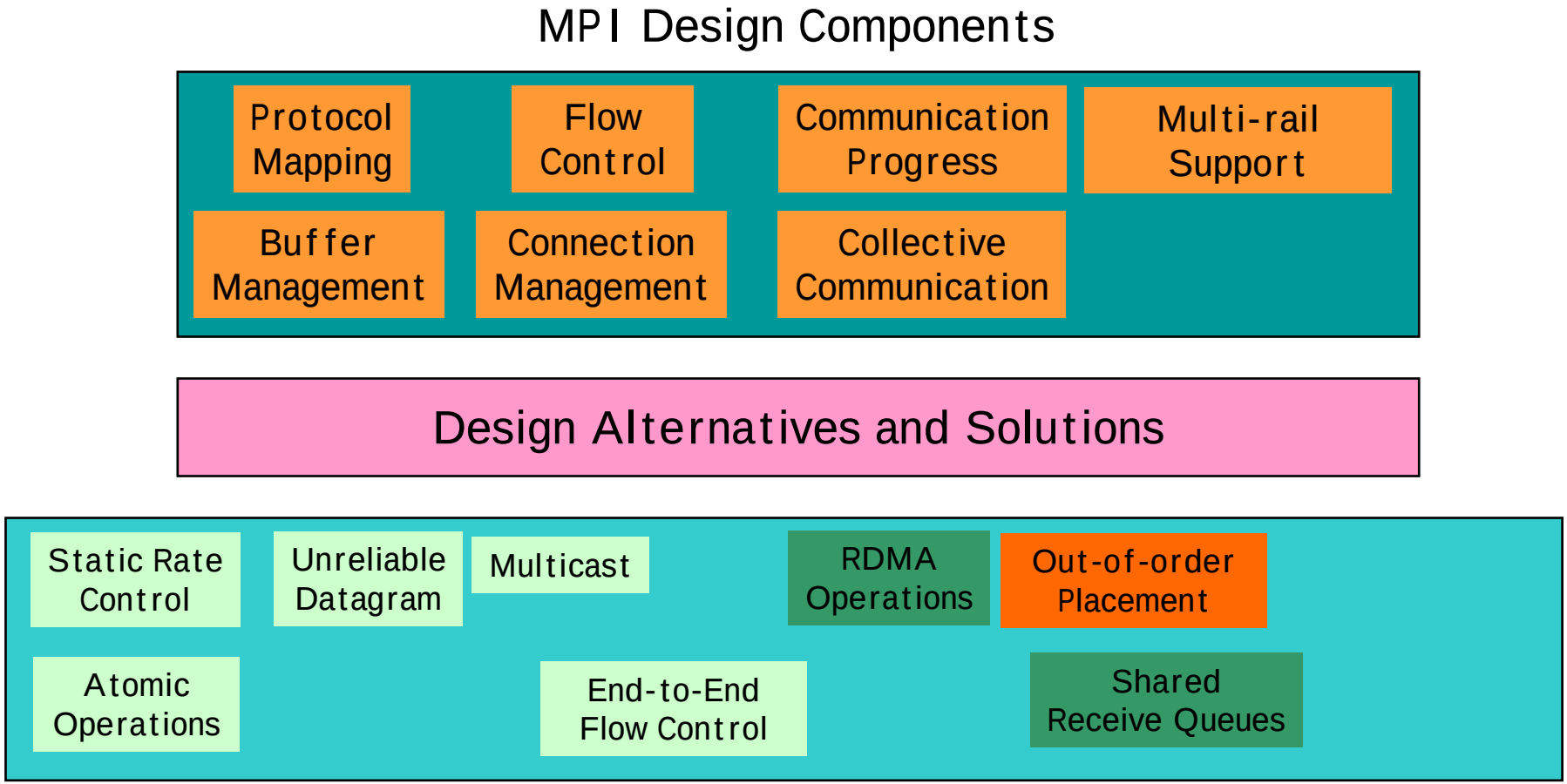
OpenFabrics Stack with Unified Verbs Interface



Performance on Current Clusters and Trends

- OpenFabrics Stack
- Sample Performance Numbers
  - MPI
  - File Systems (Lustre and NFS/RDMA)
  - Datacenters
- IB and 10GigE Installations and Trends

Designing MPI Using IB/iWARP Features



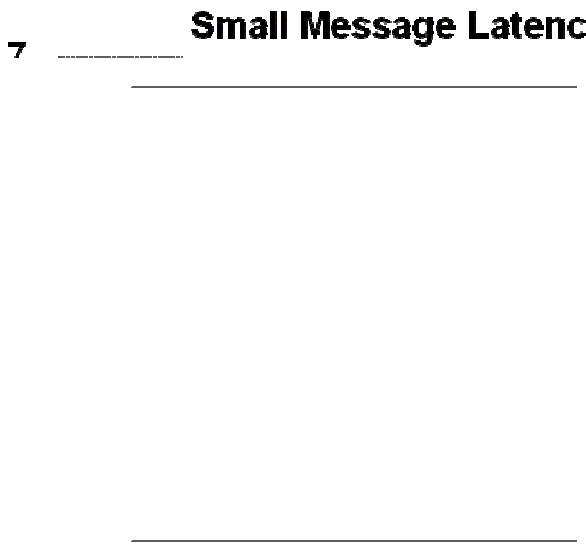
MVAPICH/MVAPICH2 Software

- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 760 organizations in 42 countries
  - More than 23,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 4th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
  - http://mvapich.cse.ohio-state.edu/

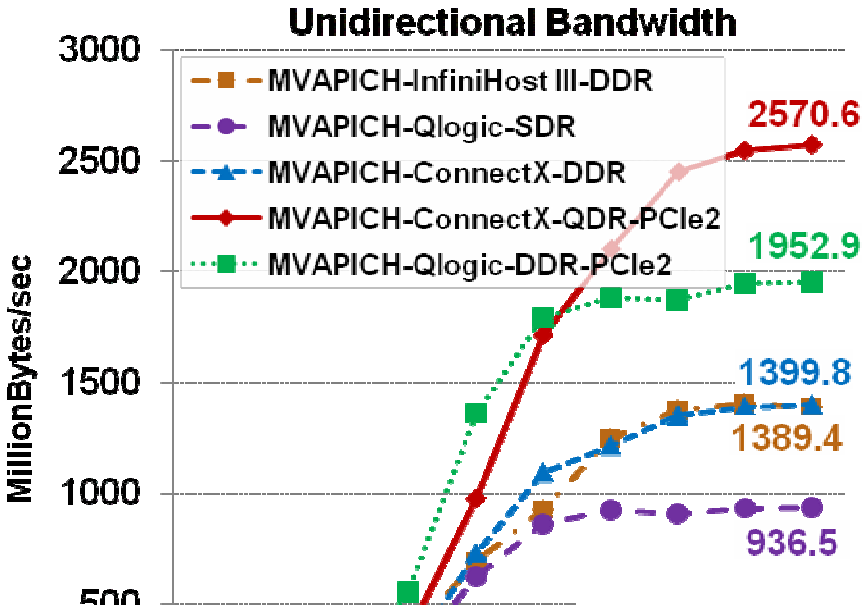
MVAPICH/MVAPICH2 Software

- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 765 organizations in 42 countries
  - More than 23,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 4th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
  - http://mvapich.cse.ohio-state.edu/

One-way Latency: MPI over IB

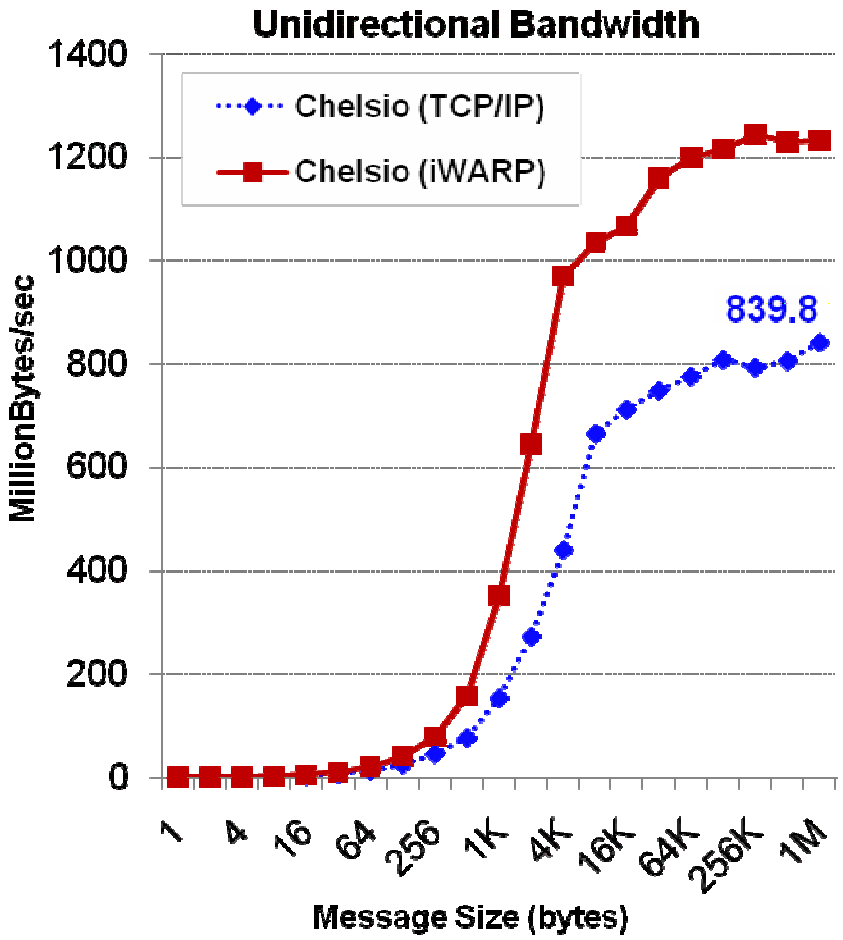


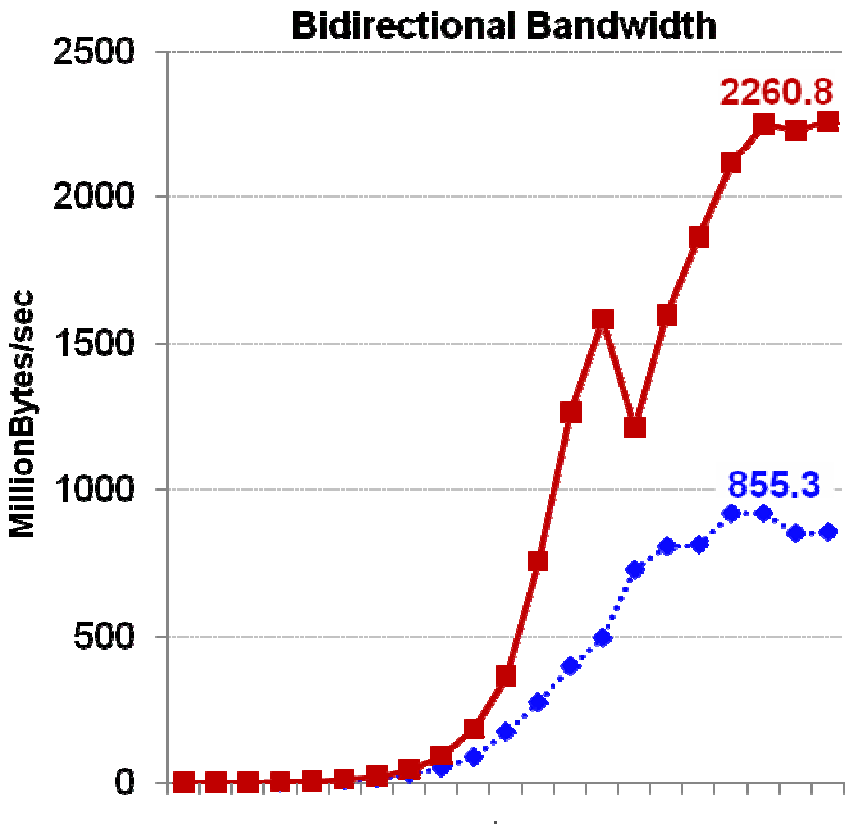
Bandwidth: MPI over IB




One-way Latency
35
30
Chelsio (TCP/IP)

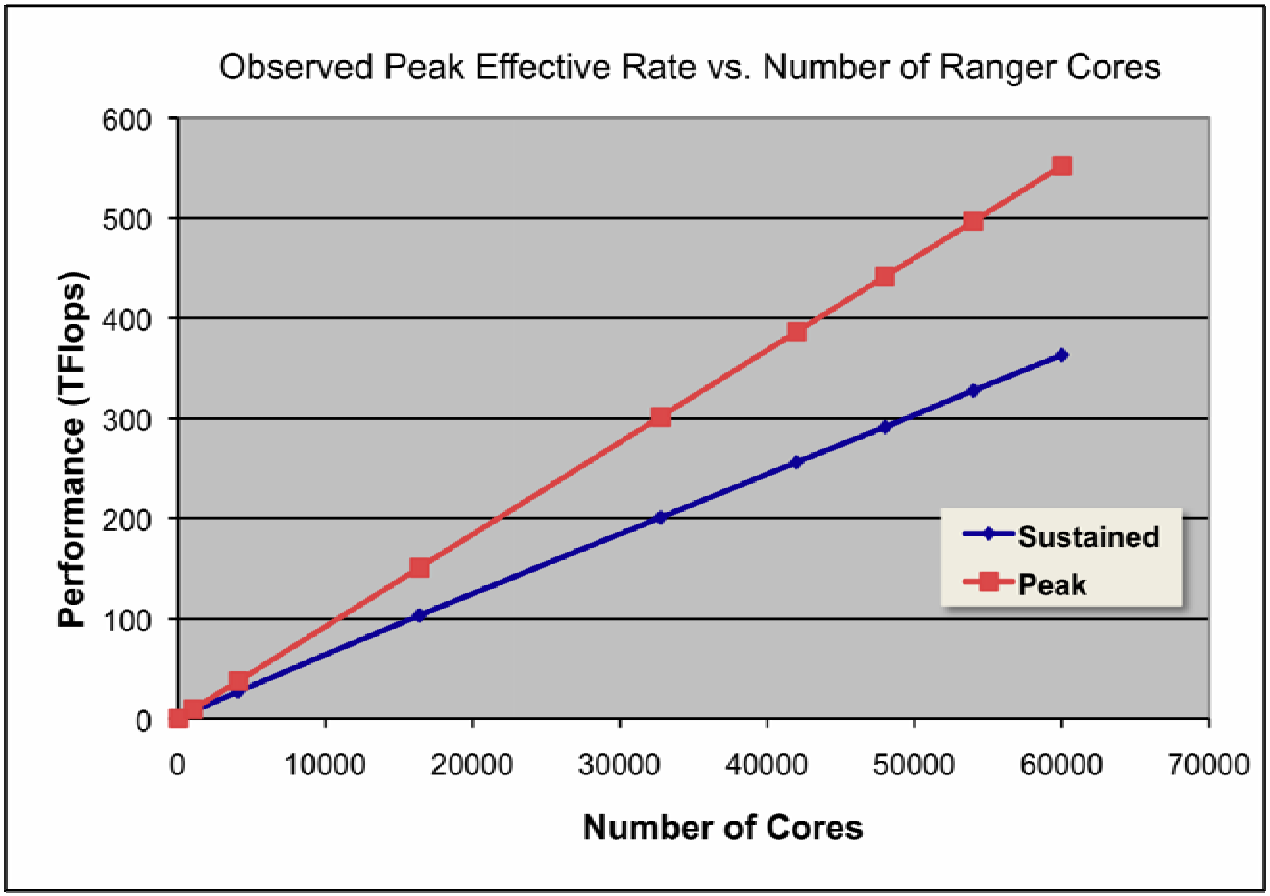
Bandwidth: MPI over iWARP





Performance of HPC Applications on TACC Ranger using MVAPICH + IB

- Rob Farber's facial recognition application was run up to 60K cores using MVAPICH
- Ranges from 84% of peak at low end to 65% of peak at high end



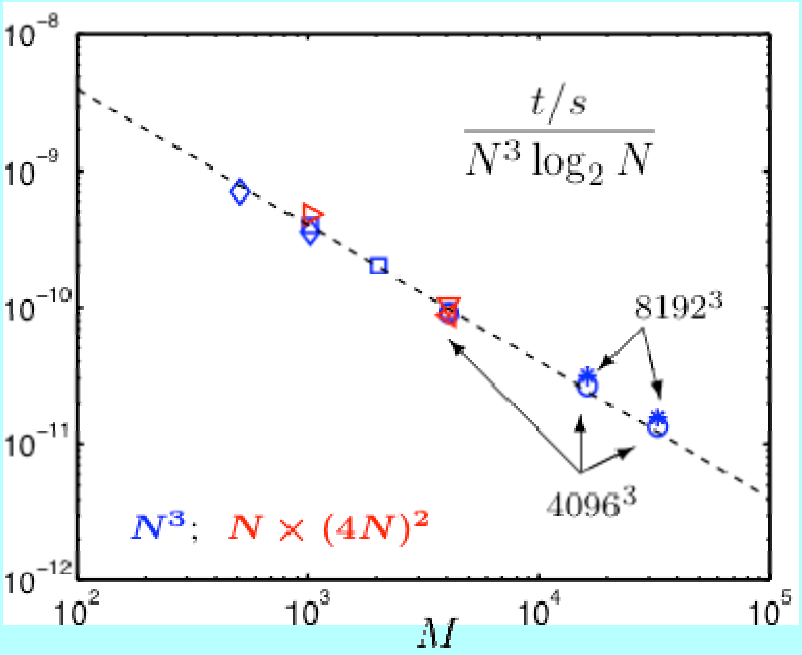
http://www.tacc.utexas.edu/research/users/features/index.php?m_b_c=farber

Performance of HPC Applications on TACC Ranger: DNS/Turbulence

- 3D FFT flop count $\propto N^3 \log_2 N$
- Perfect scaling:

$$\frac{t/s}{N^3 \log_2 N} \propto M^{-1}$$

- **Strong scaling**: $> 98\%$ at both 4096^3 and 8192^3 from $M = 16K$ to $32K$
- **Weak scaling**: $\sim 80\%$ from $(N, M) = (2048, 2048)$ to $(8192, 32768)$

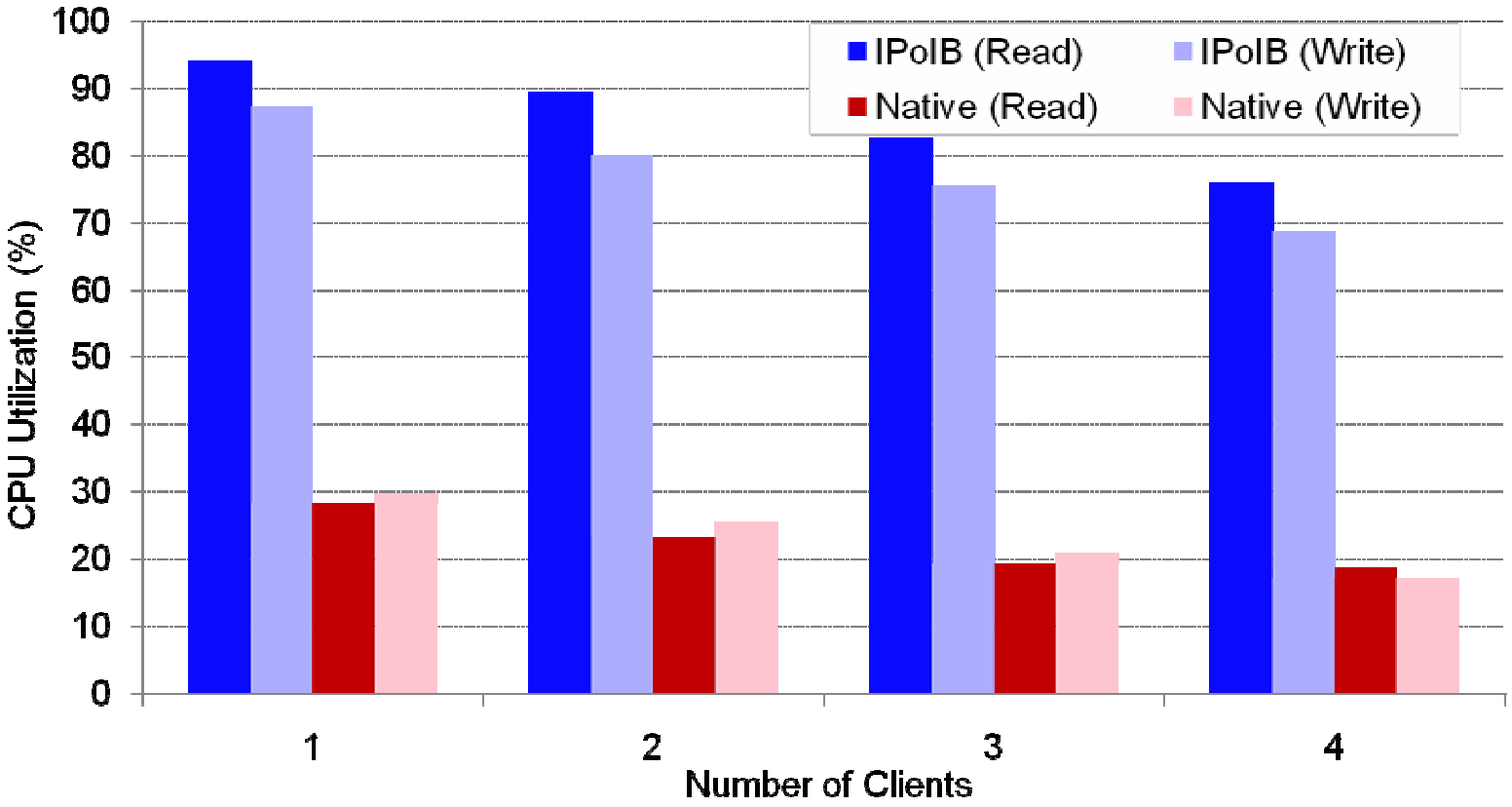


- Best timings for small M_1: row communicator within node (16 cores) or within socket (4 cores)

Courtesy: P.K. Yeung, Diego Donzis, TG 2008

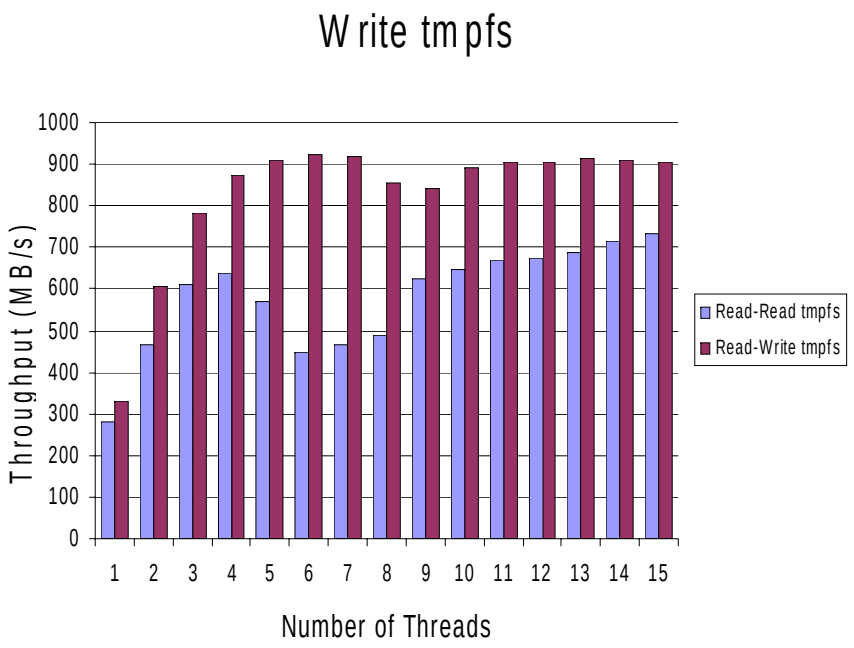
Lustre Performance

CPU Utilization



- 4 OSS nodes, IOzone record size 1MB
- Offers potential for greater scalability

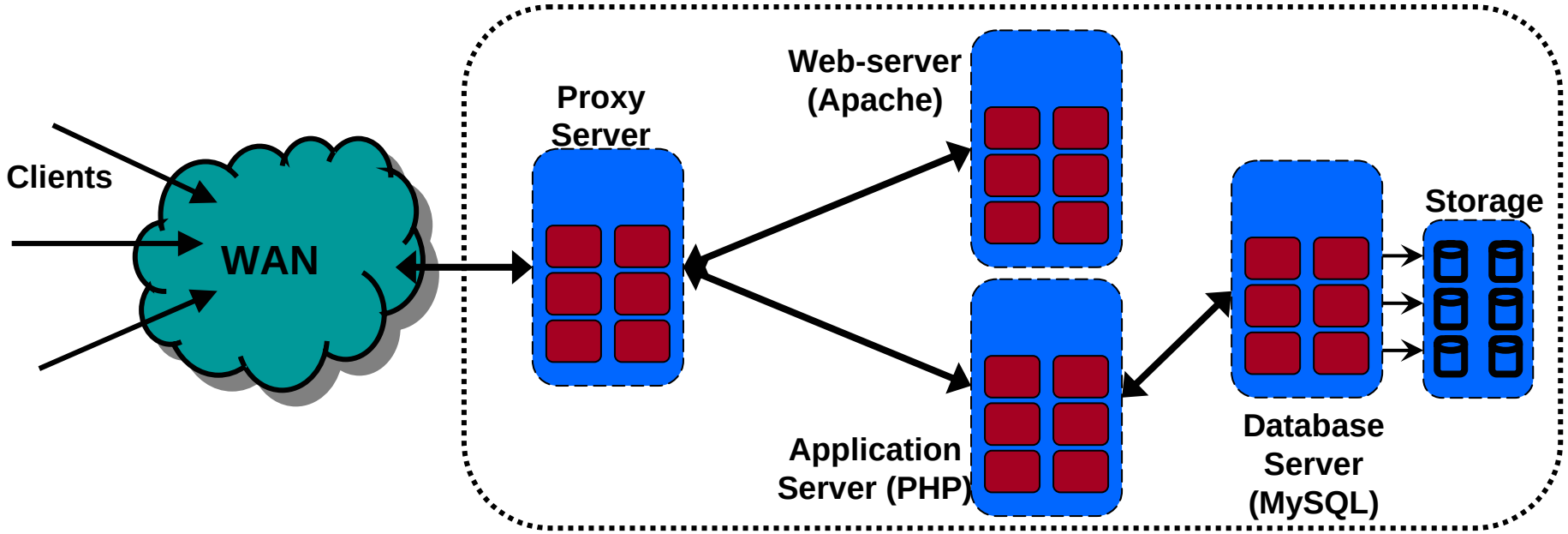
NFS/RDMA Performance



- IOzone Read Bandwidth up to 913 MB/s (Sun x2200's with x8 PCIe)
- Read-Write design by OSU, available with the latest OpenSolaris
- NFS/RDMA will also be added in OFED 1.4

R. Noronha, L. Chai, T. Talpey and D. K. Panda, "Designing NFS With RDMA For Security, Performance and Scalability", ICPP '07

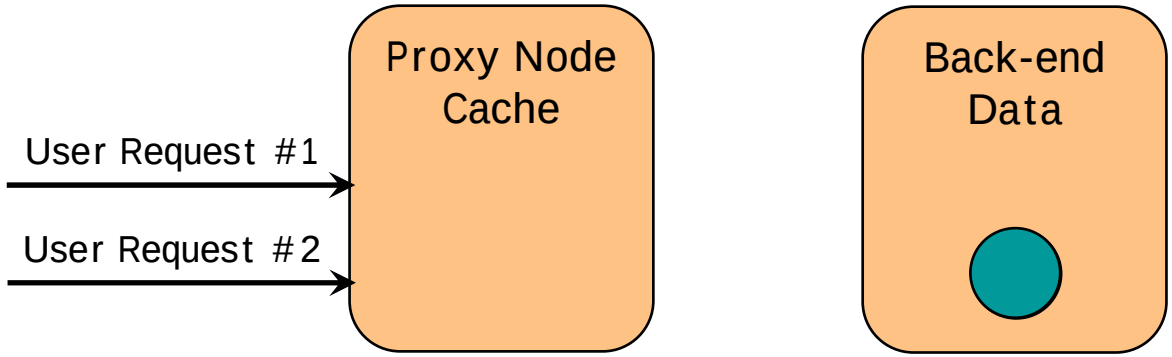
Typical Multi-Tier Datacenter



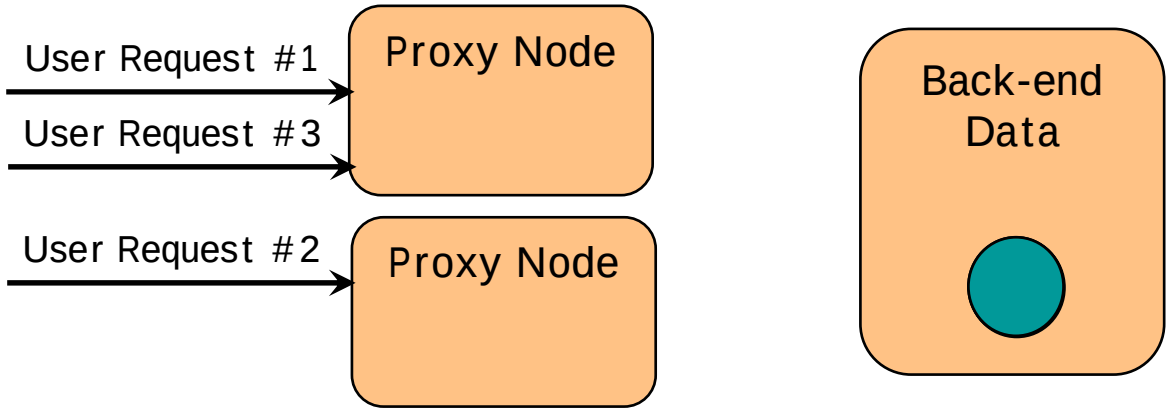
- Requests are received from clients over the WAN
- Proxy nodes perform caching, load balancing, resource monitoring, etc.
  - If not cached, request forwarded to the next tier → Application Server
- Can RDMA mechanism (one-sided communication) help?

Cache Coherency and Consistency with Dynamic Data

Example of Strong Cache Coherency: Never Send Stale Data



Example of Strong Cache Consistency:
Always Follow Increasing Time Line of Events



Strong Cache Coherency with RDMA

Performance on Current Clusters and Trends

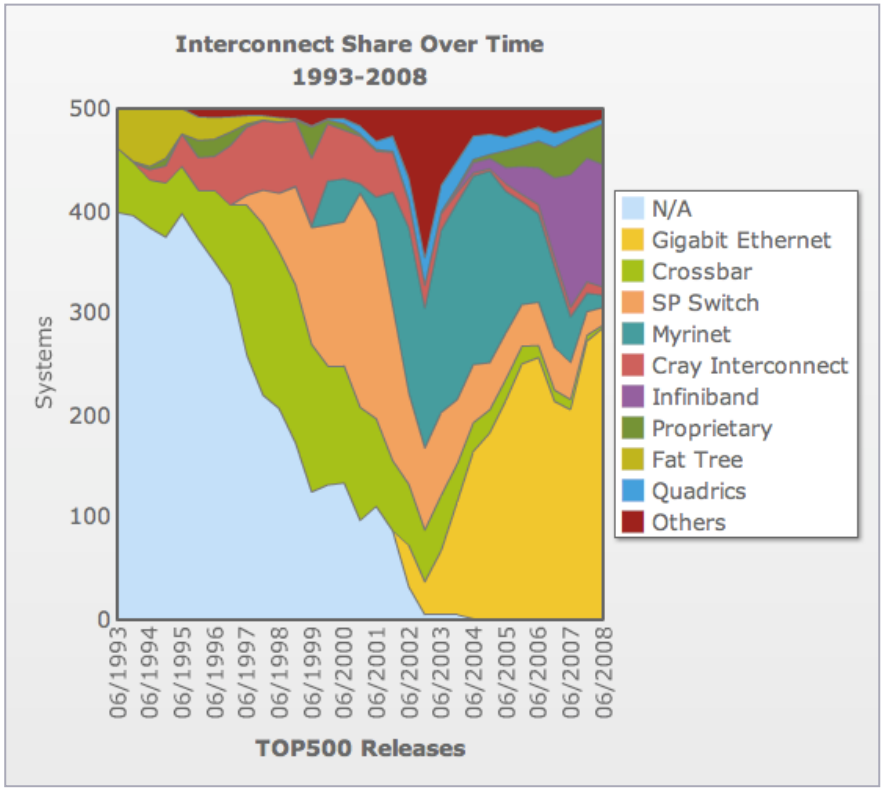
- OpenFabrics Stack
- Sample Performance Numbers
  - MPI
  - File Systems (Lustre and NFS/RDMA)
  - Datacenters
- IB and 10GigE Installations and Trends

IB Installations

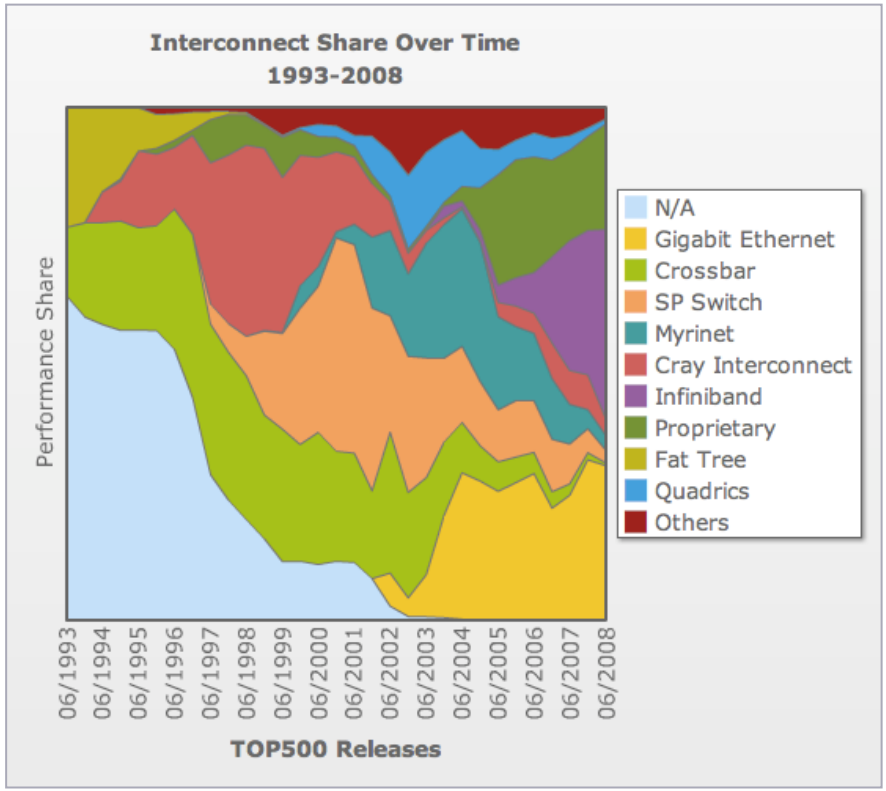
- 121 IB clusters (24.2%) in June '08 TOP500 list (www.top500.org)
- 12 IB clusters in TOP25
  - 122,400-cores (RoadRunner) at LANL (1st)
  - 62,976-cores (Ranger) at TACC (4th)
  - 14,336-cores at New Mexico (7th)
  - 14,384-cores at Tata CRL, India (8th)
  - 10,240-cores at TEP, France (10th)
  - 13,728-cores in Sweden (11th)
  - 8,320-cores in UK (18th)
  - 6,720-cores in Germany (19th)
  - 10,000-cores at CCS, Tsukuba, Japan (20th)
  - 9,600-cores at NCSA (23rd)
  - 12,344-cores at Tokyo Inst. of Technology (24th)
  - 13,824-cores at NASA/Columbia (25th)
- More are getting installed ….

InfiniBand in the Top500

Systems



Performance

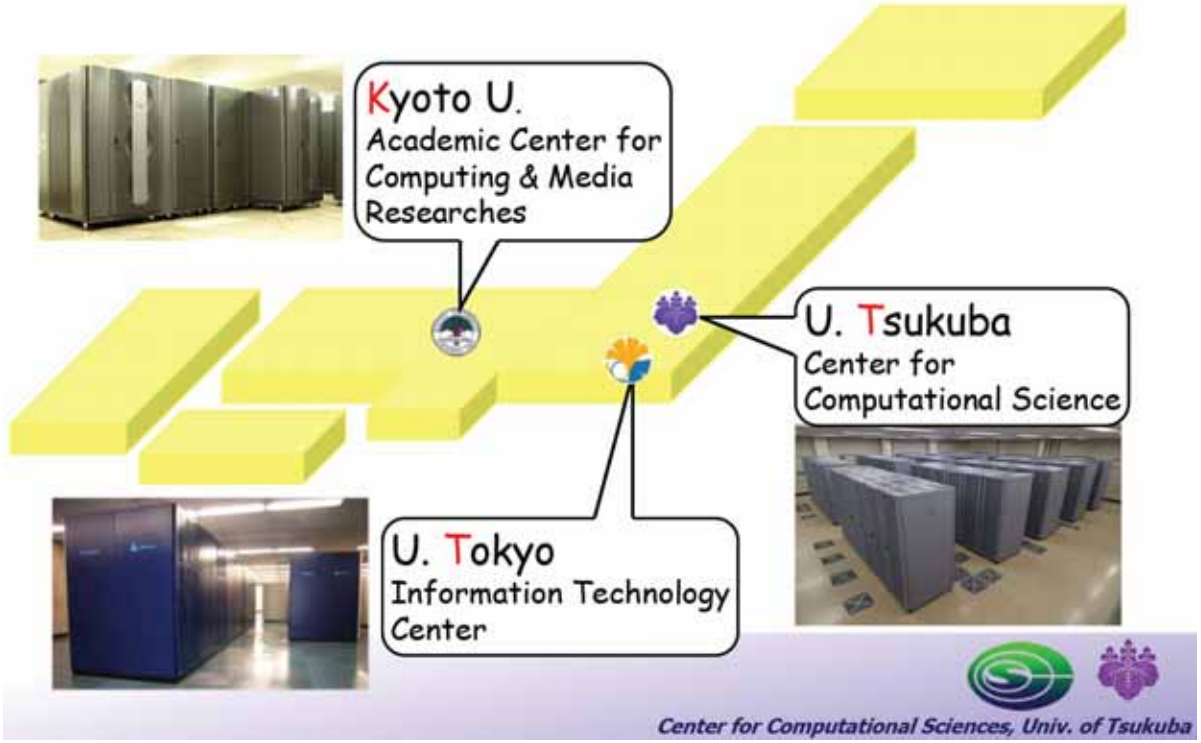


Percentage share of InfiniBand is steadily increasing

10GE Installations

- Several Enterprise Computing Domains
  - Enterprise Datacenters (HP, Intel)
  - Animation firms (e.g., Universal Studios created “The Hulk” and many new movies using 10GE)
- Scientific Computing Installations
  - 640-core installation in University of Heidelberg, Germany
  - 512-core installation at Sandia National Laboratory (SNL) with Chelsio/iWARP and Woven Systems switch
  - 256-core installation at Ohio Supercomputer Center (OSC) with Ammasso/iWARP
  - 256-core installation at Argonne National Lab with Myri-10G
- Integrated Systems
  - BG/P uses 10GE for I/O (ranks 3, 6, 9, 13, 37 in the Top 50)

Dual IB/10GE Systems

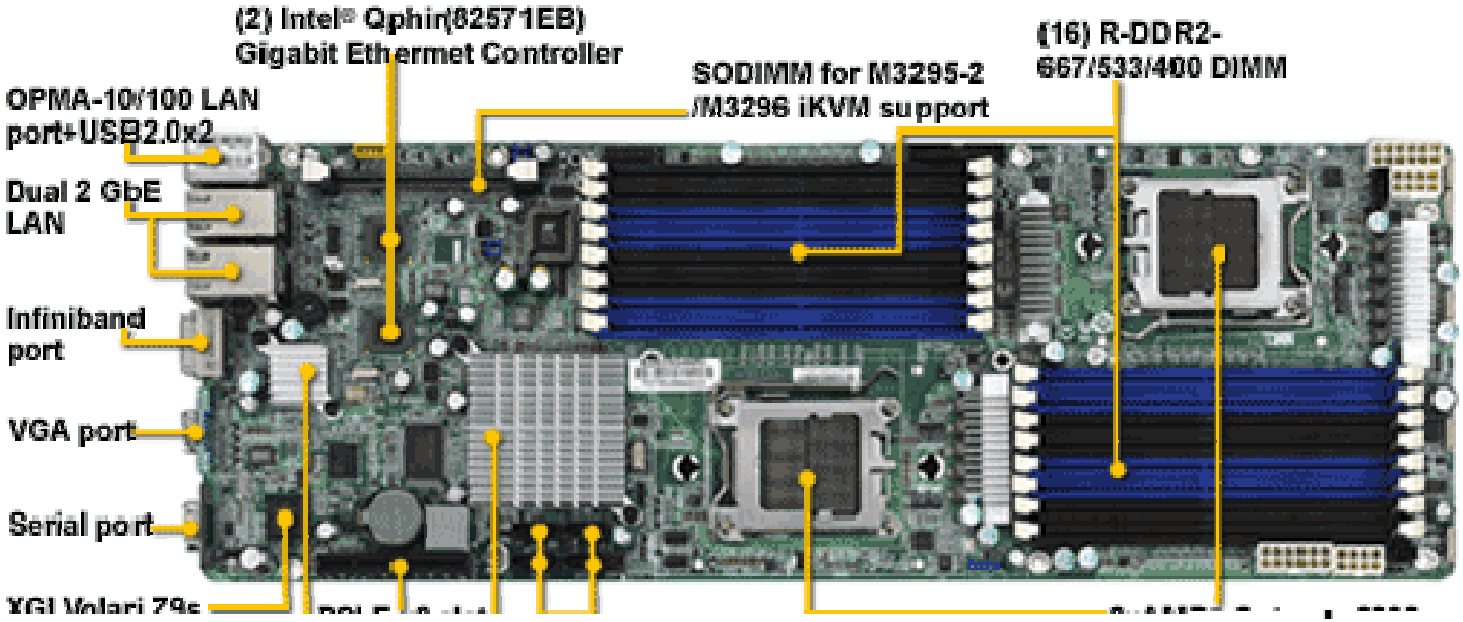


- Such systems are being integrated
- E.g., the T2K-Tsukuba system (95 TFlop System)
- Systems at three sites (Tsukuba, Tokyo, Kyoto)

(Courtesy Taisuke Boku, University of Tsukuba)

- Internal connectivity: Quad-rail IB ConnectX network
- External connectivity: 10GE

[BTO only]


(2) Intel® Qphin(82571EB)
Gigabit Ethernet Controller
SODIMM for M3295-2
/M3296 iKVM support
(16) R-DDR2-
667/533/400 DIMM
OPMA-10/100 LAN
port+USB2.0x2
Dual 2 GbE
LAN
Infiniband
port
VGA port
Serial port

Presentation Overview

- Trends in Networking Technologies
- Overview of InfiniBand and Its Features
- Overview of 10GigE/iWARP and Its Features
- Performance on Current Clusters and Trends
- Challenges
- Conclusions and Q&A

Major Challenges in Designing Exascale Clusters

- Scalability
  - Systems with 100-250K cores are being deployed
  - Exascale systems will have hundreds of millions of cores
  - Requires networking and I/O technologies for integrating such large-scale systems
- Programmability and Ease of Use (The "Wizard Gap")
  - Not everyone is a parallel programming expert
  - … and yet, everyone wants the best performance
- Performance
  - How to extract performance as system size increases
  - Small performance problems can get brutal at scale !
- Testing and Debugging
  - Need for scalable tools
- Fault Tolerance, Reliability and Maintainability
  - Failure detection and transparency
  - Ease of administration

Failures on Large-Scale Systems

System #	CPUs	MTBF/I
ASCI Q	8,192	6.5 hrs
ASCI White	8.192	40 hrs
PSC Lemieux	3,016	9.7 hrs
Google	15,000	20 reboots/day

- Results collected by [1]
- MTBF will be 1.25 hrs with current trend on Petascale machine [2]
- TACC Ranger (0.5 Petaflop) has 60K cores
- Larger configurations (100K cores) are on the horizon

[1] Chung-hsing Hsu and Wu-chun Feng, A Power-Aware Run-Time System for High-Performance Computing, Proceedings of ICS'05, Nov., 2005

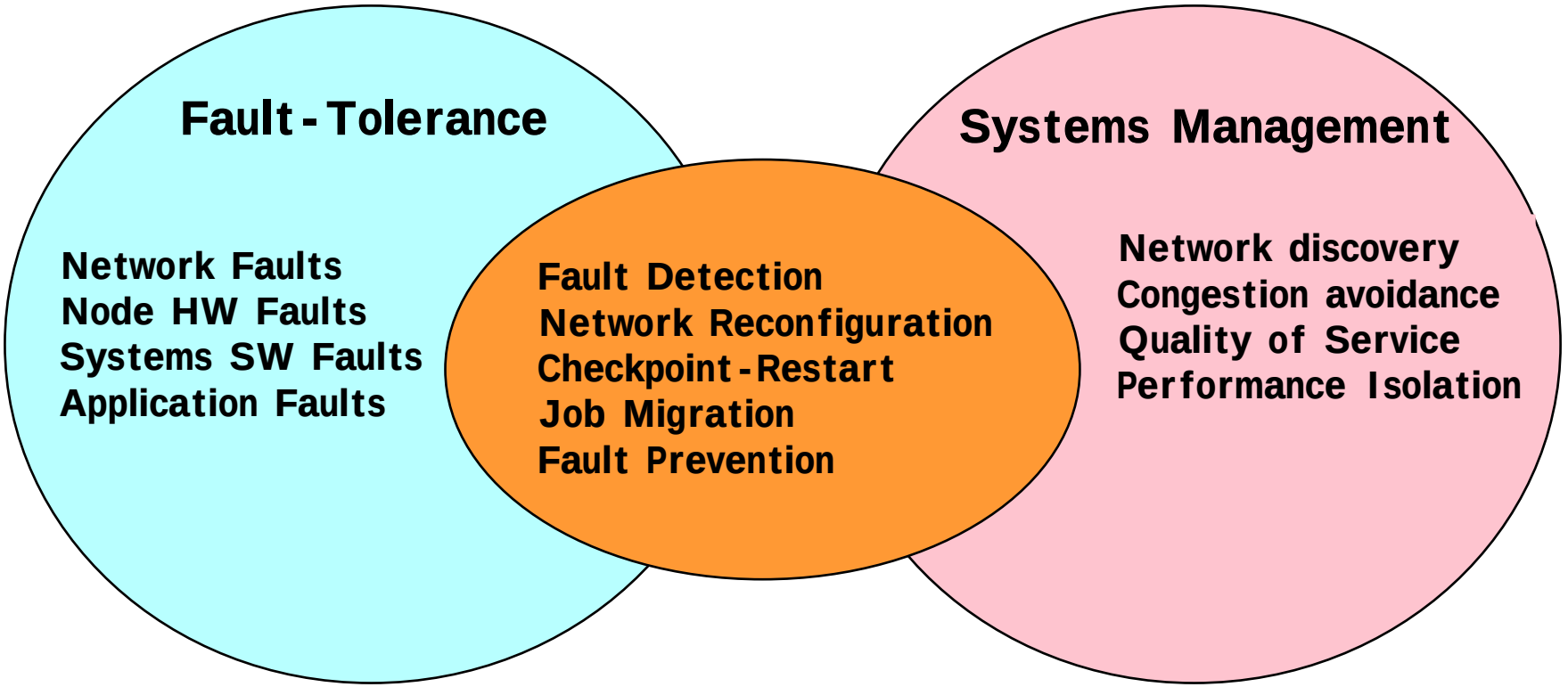
[2] I. Philp. Software Failures and the Road to a Petaflop Machine. In *HPCRI: 1st Workshop on High Performance Computing Reliability Issues, in Proceedings of the 11th International Symposium on High Performance Computer Architecture (HPCA-11)*. IEEE Computer Society, 2005.

Cost of Downtime

*Service	Cost of One Hour Downtime
Brokerage Operations	$6,450,000
Credit Card Authorization	$2,600,000
eBay	$225,000
Amazon	$180,000
Catalog Sales Center	$90,000

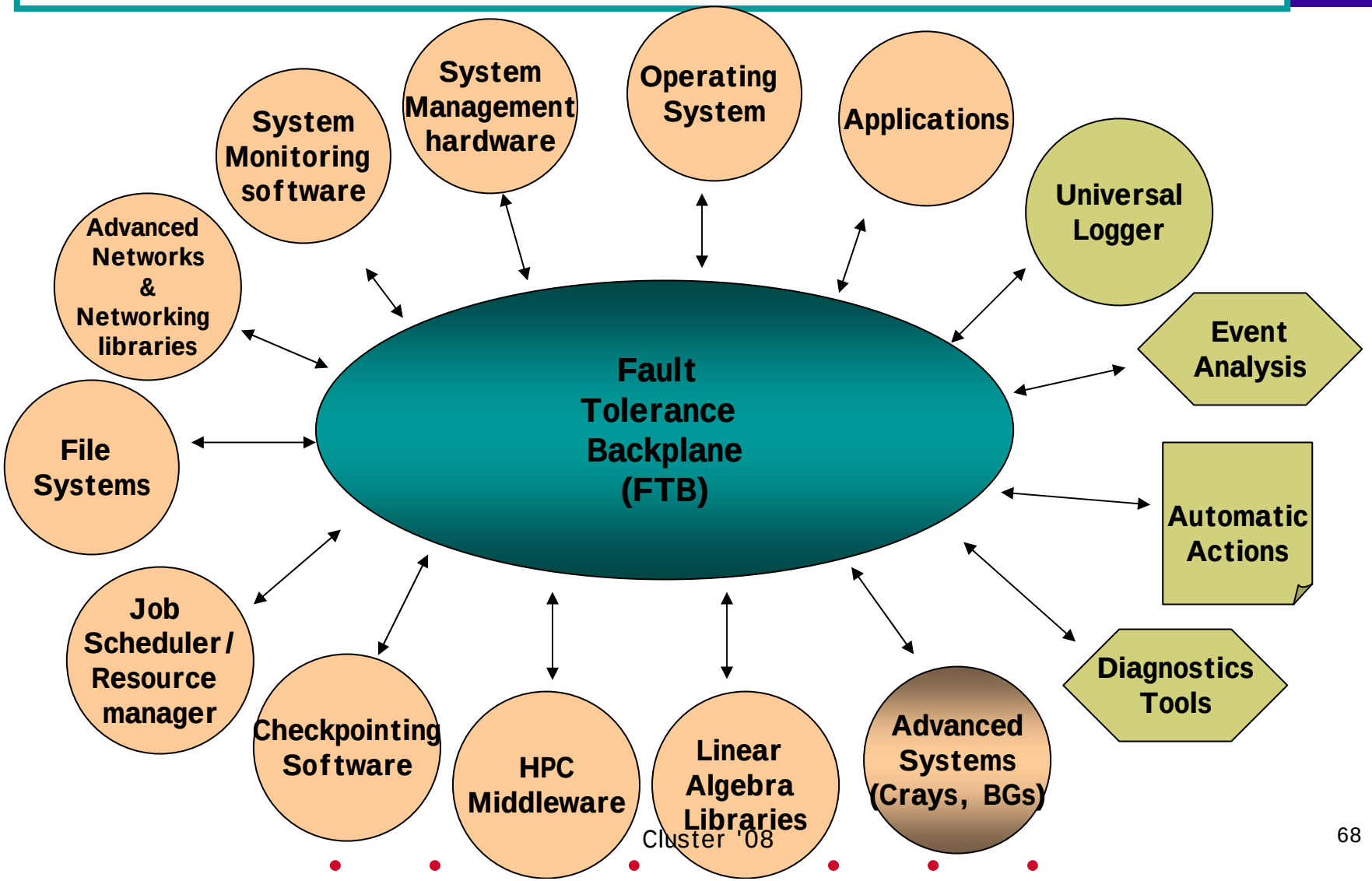
**"A Power-aware Run-Time System for High-Performance Computing", Chung-hsing Hsu and Wu-chun Feng, IEEE International Supercomputing Conference (SC), 2005*

Fault-Tolerance and Systems Management



InfiniBand provides a lot of underlying support to achieve these.

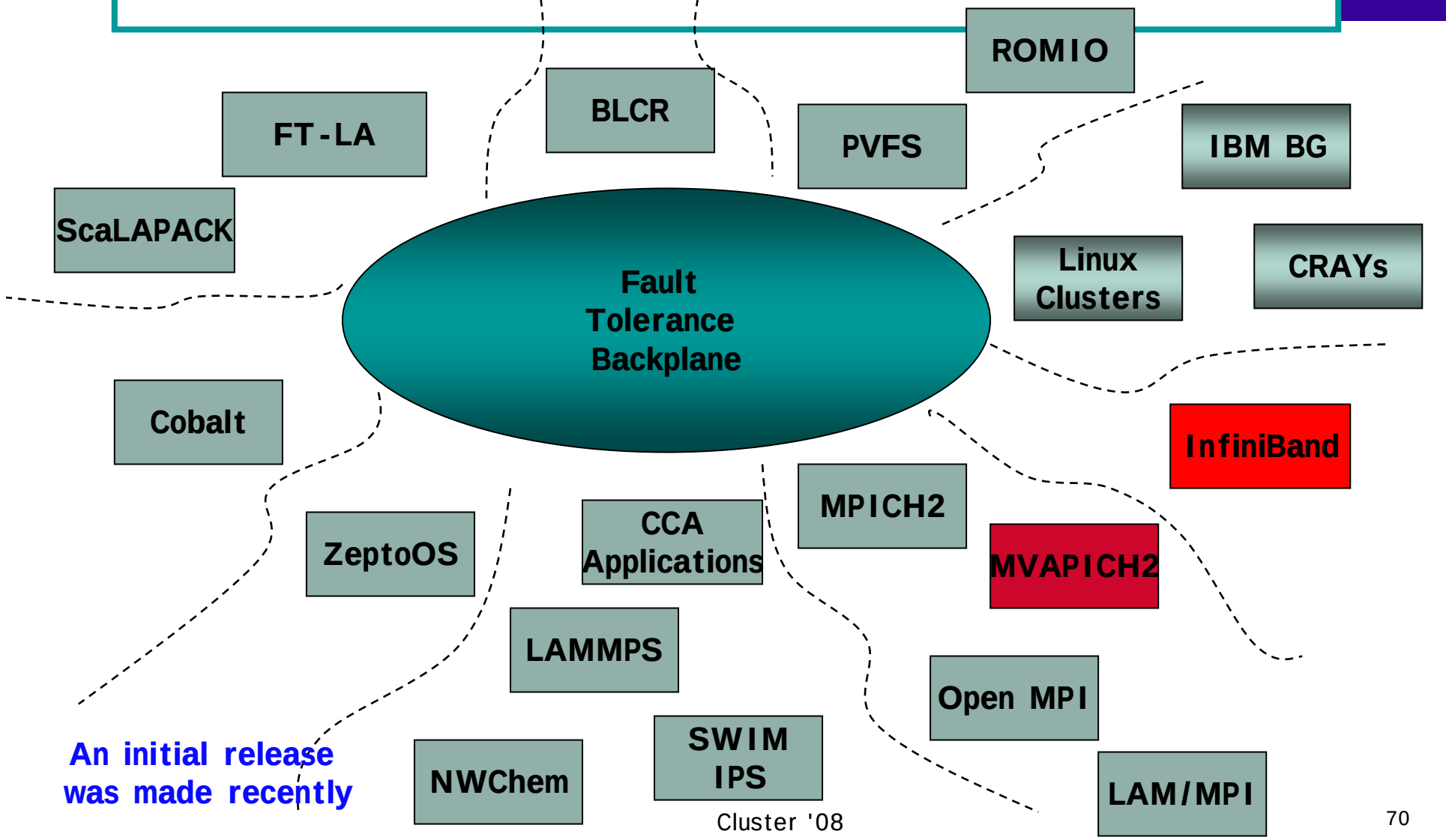
Coordinated Infrastructure for Fault Tolerant Systems (CIFTS) Framework



CIFTS team

- Argonne National Laboratory
  - Pete Beckman, Rinku Gupta, Ewing Lusk, Rob Ross, Rajeev Thakur
- Indiana University
  - Andrew Lumsdaine
- Lawrence Berkeley National Laboratory
  - Paul Hargrove
- Oak Ridge National Laboratory
  - Al Geist, David Bernholdt, Pratul Agarwal, Scott Hampton, Byung-Hoon Park, Aniruddha Shet
- Ohio State University
  - D.K. Panda
- University of Tennessee, Knoxville
  - Jack Dongarra

FTB-enabled Software



Concluding Remarks

- Presented network architectures & trends in Clusters
- Presented background and details of IB and 10GE/iWARP architectures
- Discussed sample performance numbers in designing various high-end systems with IB and 10GE
- Outlined challenges for designing next generation Exascale clusters
- Golden Era of Cluster Computing
- IB and 10GE are emerging as new architectures leading to a new generation of networked computing systems, opening many research issues needing novel solutions

Funding Acknowledgments

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by

Personnel Acknowledgments

Current Students

- L. Chai (Ph.D.)
- T. Gangadharappa (M. S.)
- K. Gopalakrishnan (M. S.)
- M. Koop (Ph.D.)
- P. Lai (Ph. D.)
- G. Marsh (Ph. D.)
- X. Ouyang (Ph.D.)
- G. Santhanaraman (Ph.D.)
- J. Sridhar (M. S.)
- H. Subramoni (M. S.)

Current Programmer

- J. Perkins

Past Students

- P. Balaji (Ph.D.)
- D. Buntinas (Ph.D.)
- S. Bhagvat (M.S.)
- B. Chandrasekharan (M.S.)
- W. Jiang (M.S.)
- W. Huang (Ph.D.)
- S. Kini (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- J. Liu (Ph.D.)
- A. Mamidala (Ph.D.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- S. Sur (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- R. Noronha (Ph.D.)
- S. Sur (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Web Pointers

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu/

E-mail: panda@cse.ohio-state.edu