NETWORK-BASED COMPUTING LABORATORY

Head-to-TOE Evaluation of High Performance Sockets over Protocol Offload Engines

P. Balaji[¥] W. Feng[α] Q. Gao[¥]

R. Noronha[¥] W. Yu[¥] D. K. Panda [¥]

[¥]Network Based Computing Lab,
Ohio State University

[α]Advanced Computing Lab,
Los Alamos National Lab

OHIO STATE

Ethernet Trends

- Ethernet is the most widely used network architecture today
- Traditionally Ethernet has been notorious for performance issues
  - Near an order-of-magnitude performance gap compared to InfiniBand, Myrinet
    - Cost conscious architecture
    - Relied on host-based TCP/IP for network and transport layer support
    - Compatibility with existing infrastructure (switch buffering, MTU)
  - Used by 42.4% of the Top500 supercomputers
  - Key: Extremely high performance per unit cost
    - GigE can give about 900Mbps (performance) / 0$ (cost)
- 10-Gigabit Ethernet (10GigE) recently introduced
  - 10-fold (theoretical) increase in performance while retaining existing features
  - Can 10GigE bridge the performance between Ethernet and InfiniBand/Myrinet?

InfiniBand, Myrinet and 10GigE: Brief Overview

- InfiniBand (IBA)
  - Industry Standard Network Architecture
  - Supports 10Gbps and higher network bandwidths
  - Offloaded Protocol Stack
  - Rich feature set (one-sided, zero-copy communication, multicast, etc.)
- Myrinet
  - Proprietary network by Myricom
  - Supports up to 4Gbps with dual ports (10G adapter announced !)
  - Offloaded Protocol Stack and rich feature set like IBA
- 10-Gigabit Ethernet (10GigE)
  - The next step for the Ethernet family
  - Supports up to 10Gbps link bandwidth
  - Offloaded Protocol Stack
  - Promises a richer feature set too with the upcoming iWARP stack

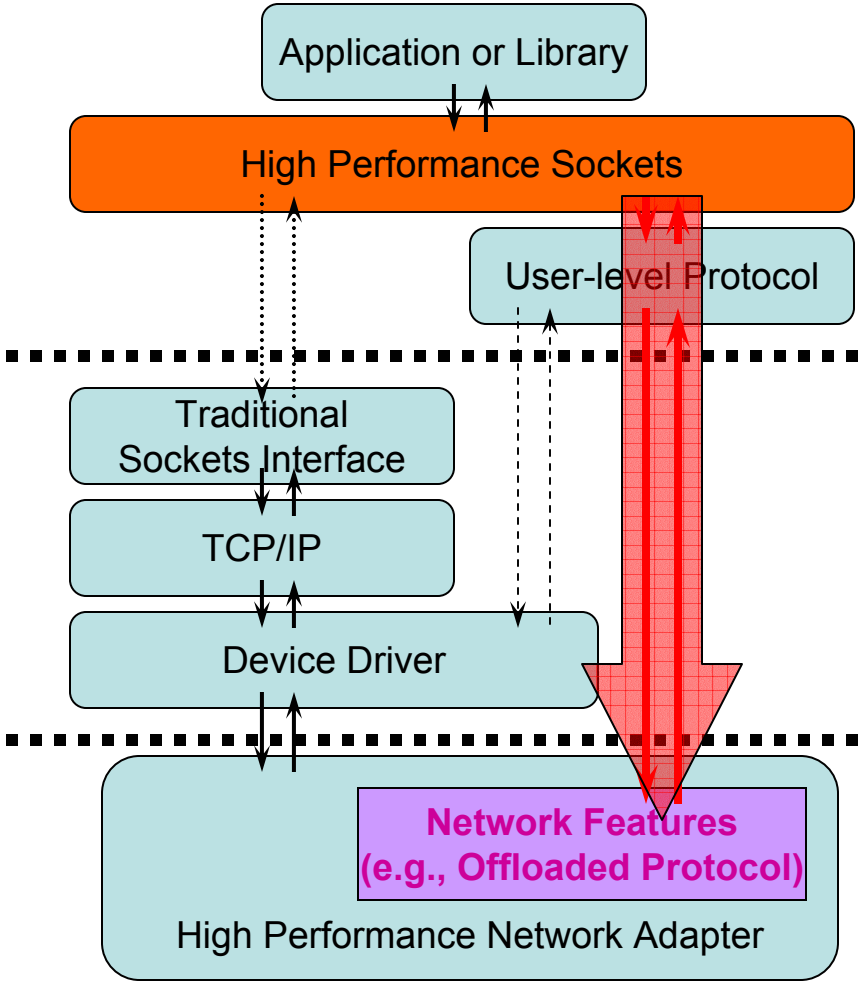
Characterizing the Performance Gap

- Each High Performance Interconnect has its own interface
  - Characterizing the performance gap is no longer straight forward
- Portability in Application Development
  - Portability across various networks is a must
  - Message Passing Interface (MPI)
    - De facto standard for Scientific Applications
  - Sockets Interface
    - Legacy Scientific Applications
    - Grid-based or Heterogeneous computing applications
    - File and Storage Systems
    - Other Commercial Applications
- Sockets and MPI are the right choices to characterize the GAP
  - In this paper we concentrate only on the Sockets interface

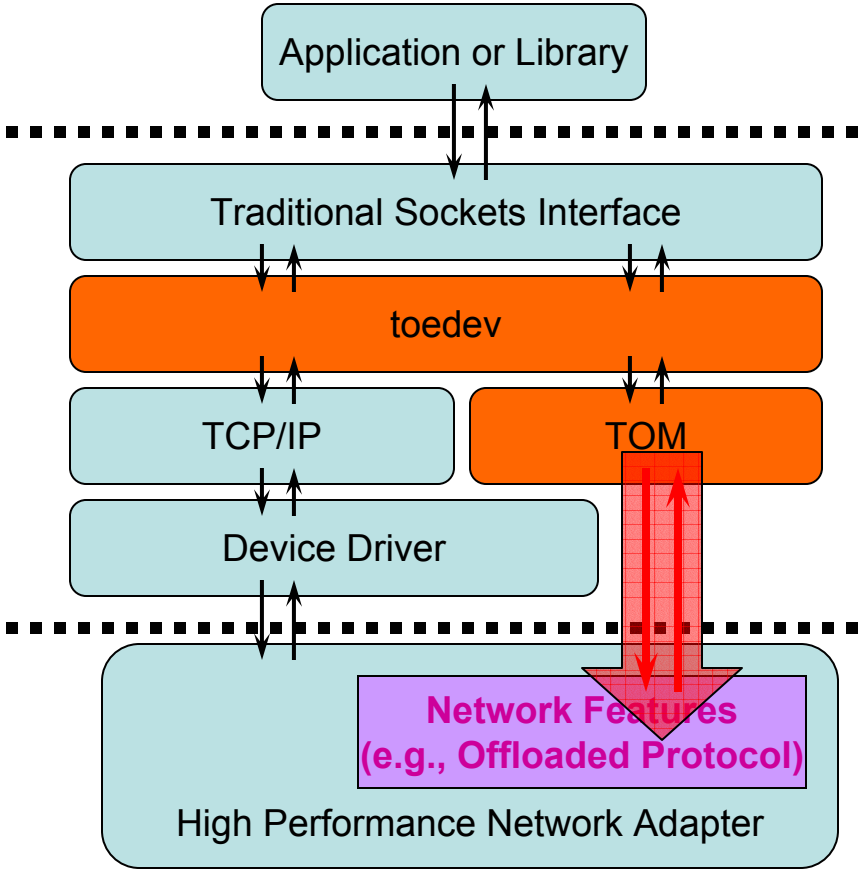
Presentation Overview

- Introduction and Motivation
- **Protocol Offload Engines**
- Experimental Evaluation
- Conclusions and Future Work

Interfacing with Protocol Offload Engines



High Performance Sockets



TCP Stack Override

High Performance Sockets

- Pseudo Sockets-like Interface
  - Smooth transition for existing sockets applications
    - Existing applications do not have to rewritten or recompiled !
  - Improved Performance by using the Offloaded Protocol Stack on networks
- High Performance Sockets exist for many networks
  - Initial implementations on VIA *[shah98:canpc, kim00:cluster, balaji02:hpdc]*
  - Follow-up implementations on Myrinet and Gigabit Ethernet *[balaji02:cluster]*
  - Sockets Direct Protocol is an industry standard specification for IBA
    - Implementations by Voltaire, Mellanox and OSU exist *[balaji03:ispass, mellanox05:hoti]*

TCP Stack Override

- Similar to the High Performance Sockets approach, but…
  - Overrides the TCP layer instead of the Sockets layer
- Advantages compared to High Performance Sockets
  - Sockets features do not have to be duplicated
    - E.g., Buffer management, Memory registration
  - Features implemented in the Berkeley sockets implementation can be used
- Disadvantages compared to High Performance Sockets
  - A kernel patch is required
  - Some TCP functionality has to be duplicated
- This approach is used by Chelsio in their 10GigE adapters

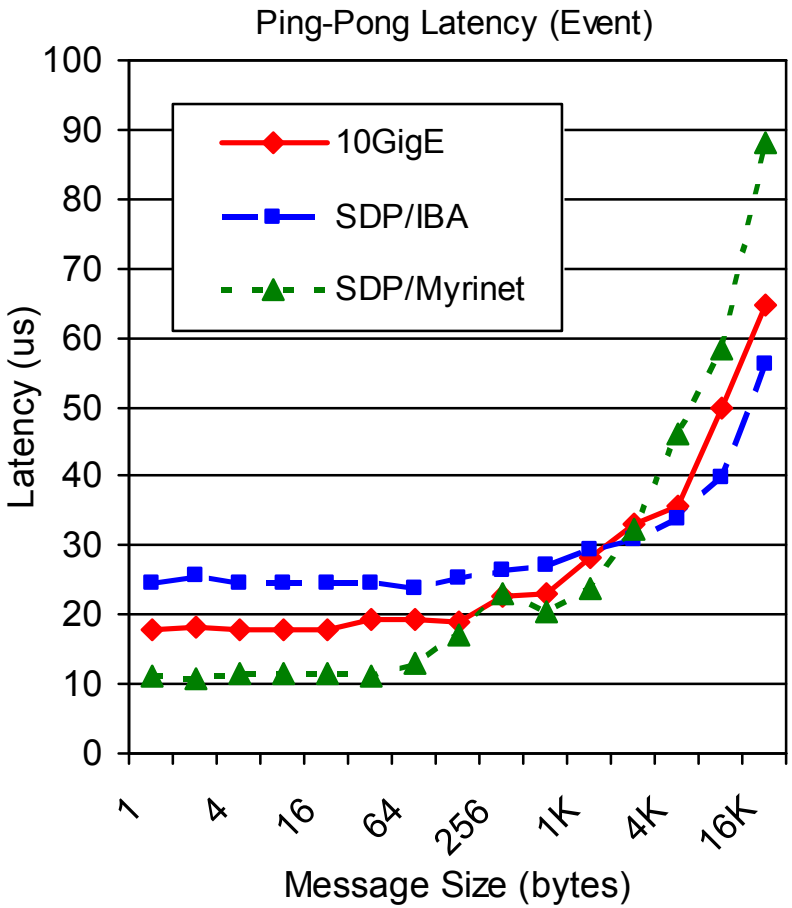
Presentation Overview

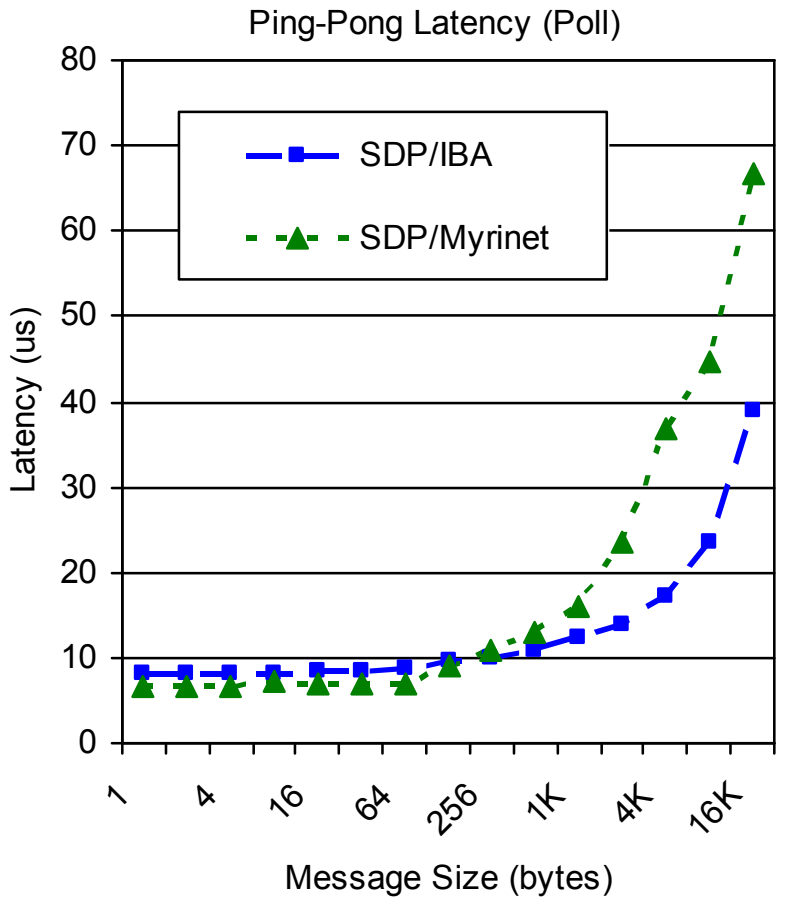
- Introduction and Motivation
- High Performance Sockets over Protocol Offload Engines
- **Experimental Evaluation**
- Conclusions and Future Work

Experimental Test-bed and Evaluation

- 4 node cluster: Dual Xeon 3.0GHz; SuperMicro SUPER X5DL8-GG nodes
- 512 KB L2 cache; 2GB of 266MHz DDR SDRAM memory; PCI-X 64-bit 133MHz
- InfiniBand
  - Mellanox MT23108 Dual Port 4x HCAs **(10Gbps link bandwidth)**; MT43132 24-port switch
  - Voltaire IBHost-3.0.0-16 stack
- Myrinet
  - Myrinet-2000 dual port adapters **(4Gbps link bandwidth)**
  - SDP/Myrinet v1.7.9 over GM v2.1.9
- 10GigE
  - Chelsio T110 adapters **(10Gbps link bandwidth)**; Foundry 16-port SuperX switch
  - Driver v1.2.0 for the adapters; Firmware v2.2.0 for the switch
- Experimental Results:
  - Micro-benchmarks (latency, bandwidth, bi-dir bandwidth, multi-stream, hot-spot, fan-tests)
  - Application-level Evaluation

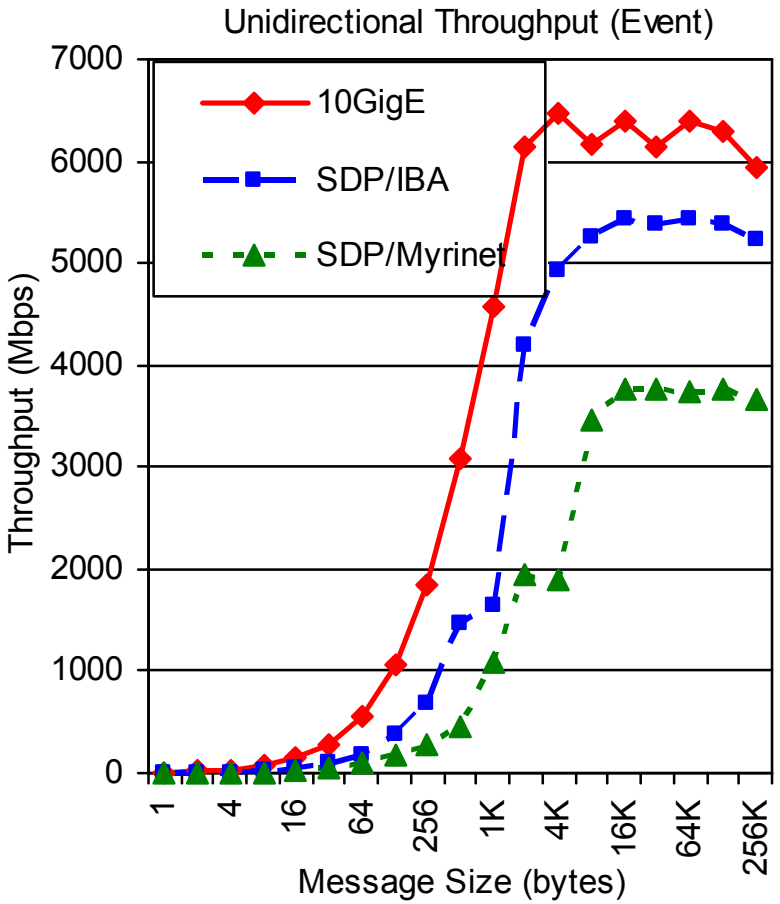
Ping-Pong Latency Measurements

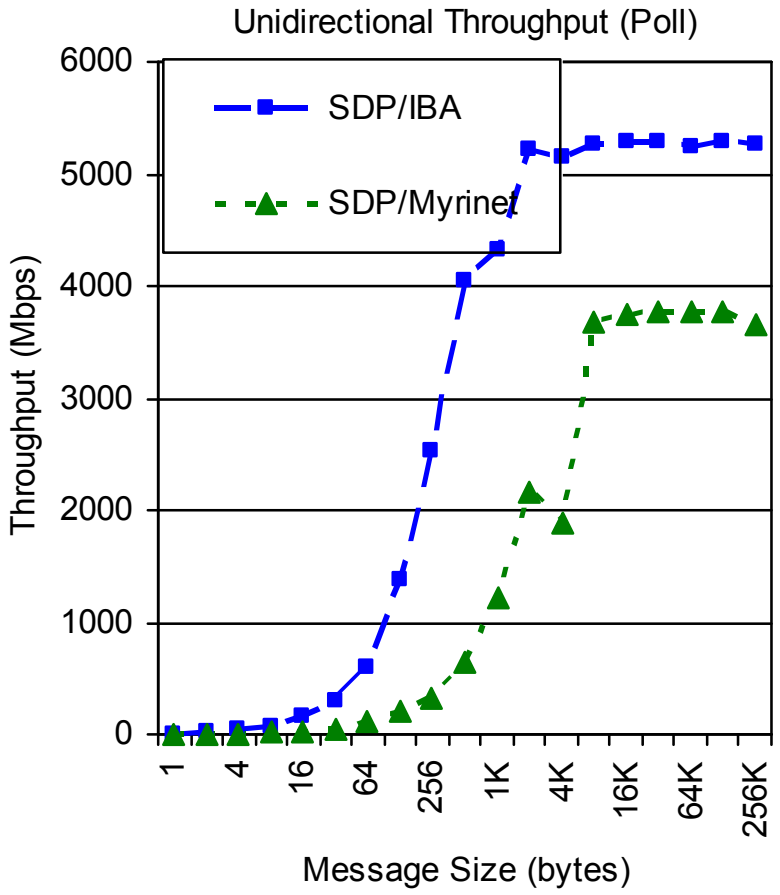




- SDP/Myrinet achieves the best small message latency at 11.3us
  - 10GigE and IBA achieve latencies of 17.7us and 24.4us respectively
- As message size increases, IBA performs the best → Myrinet cards are 4Gbps links right now !

Unidirectional Throughput Measurements



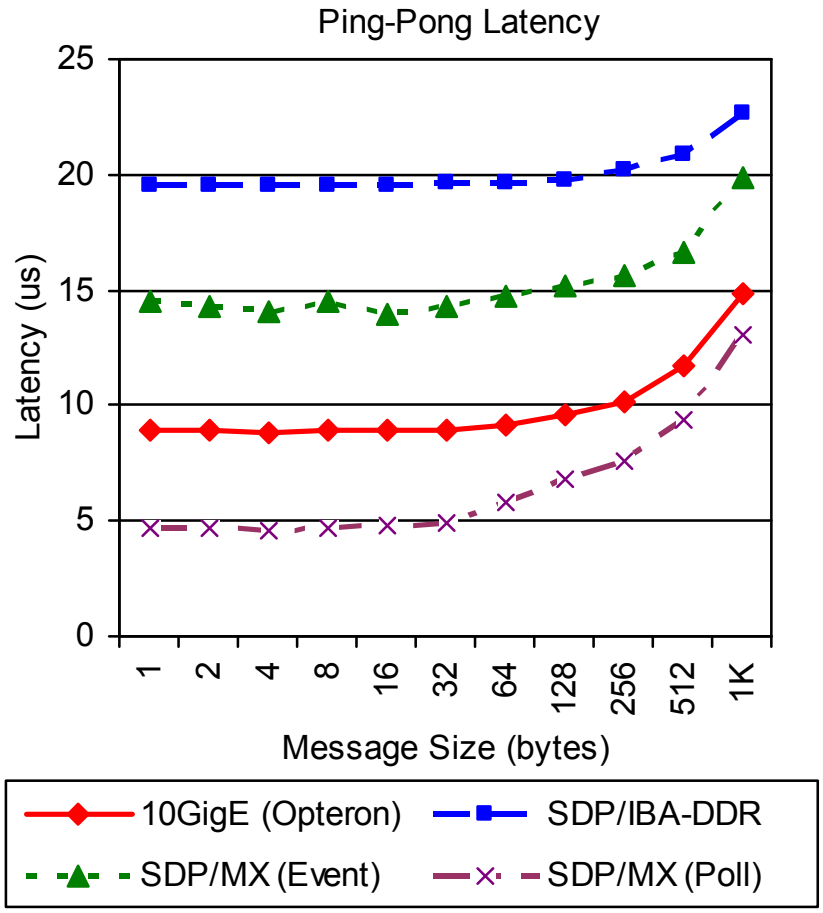


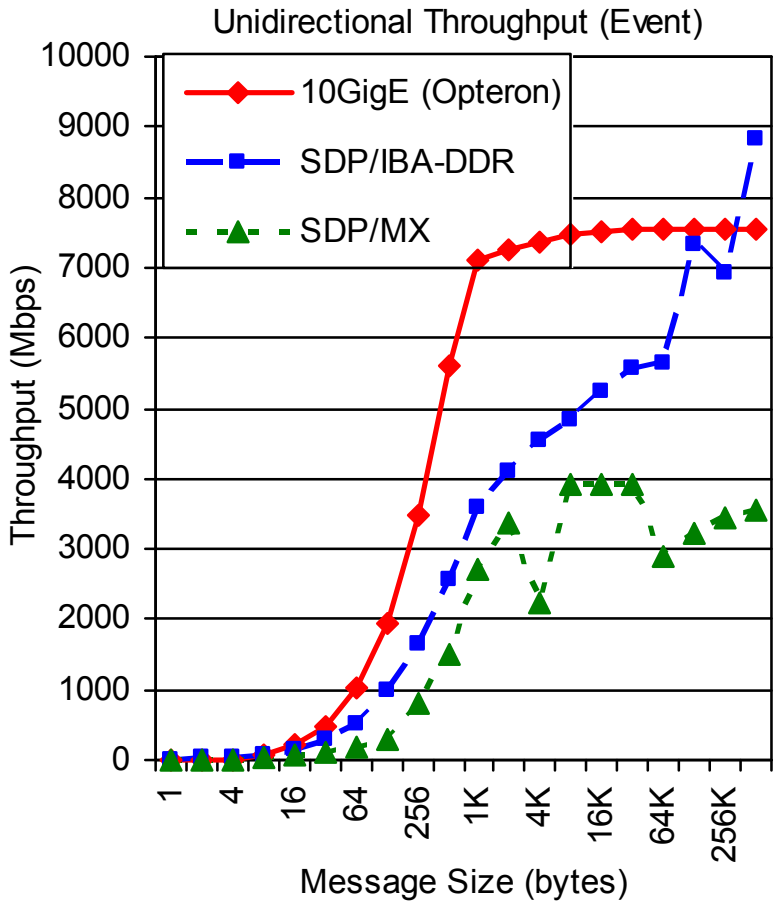
- 10GigE achieves the highest throughput at 6.4Gbps
  - IBA and Myrinet achieve about 5.3Gbps and 3.8Gbps → Myrinet is only a 4Gbps link

Snapshot Results ("Apples-to-Oranges" comparison)

10GigE: Opteron 2.2GHz IBA: Xeon 3.6GHz Myrinet: Xeon 3.0GHz

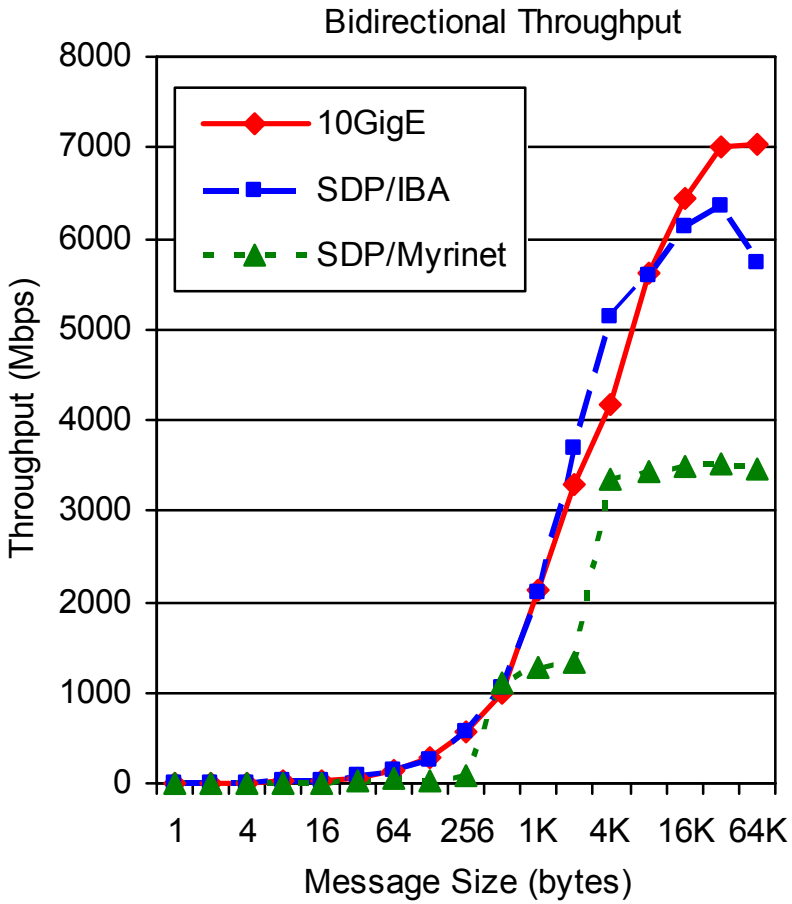
Provide a reference point ! Only valid for THIS slide !

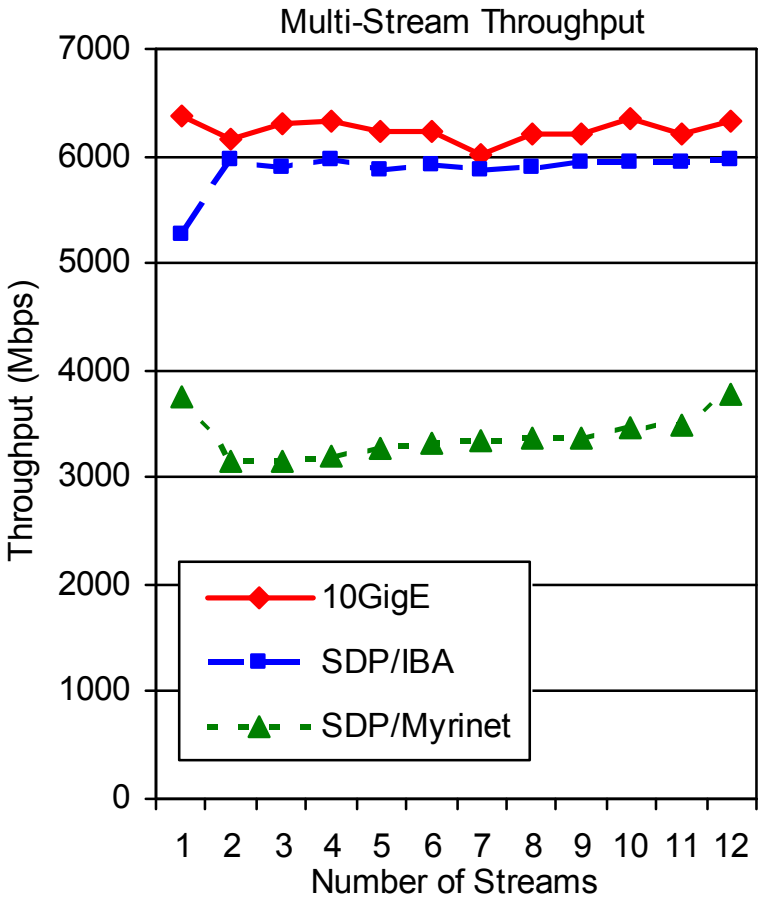




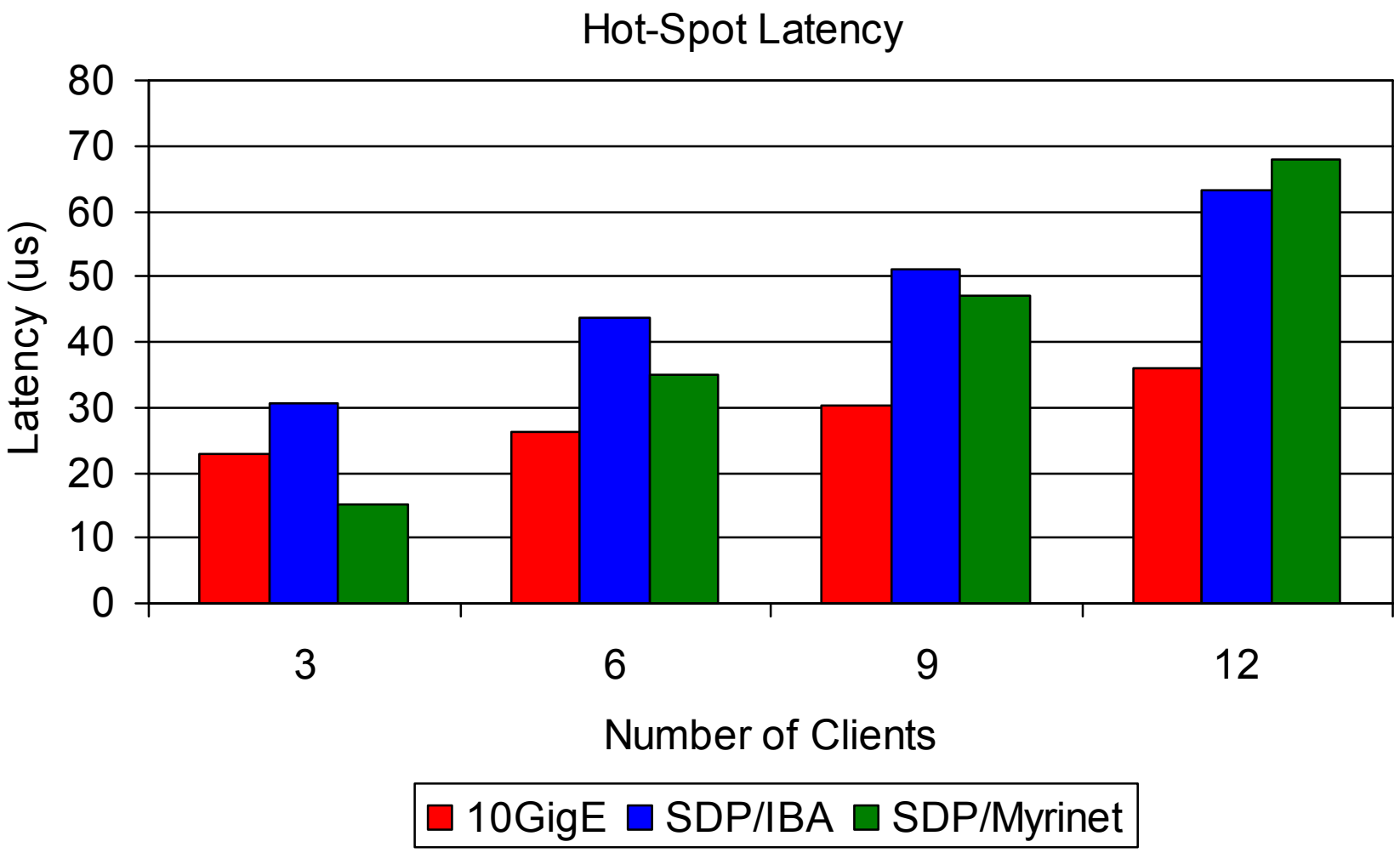
- SDP/Myrinet with MX allows polling and achieves about 4.6us latency (event-based is better for SDP/GM ~ 11.3us)
- 10GigE achieves the lowest event-based latency of 8.9us on Opteron systems
- IBA achieves a 9Gbps throughput with their DDR cards (link speed of 20Gbps)

Bidirectional and Multi-Stream Throughput



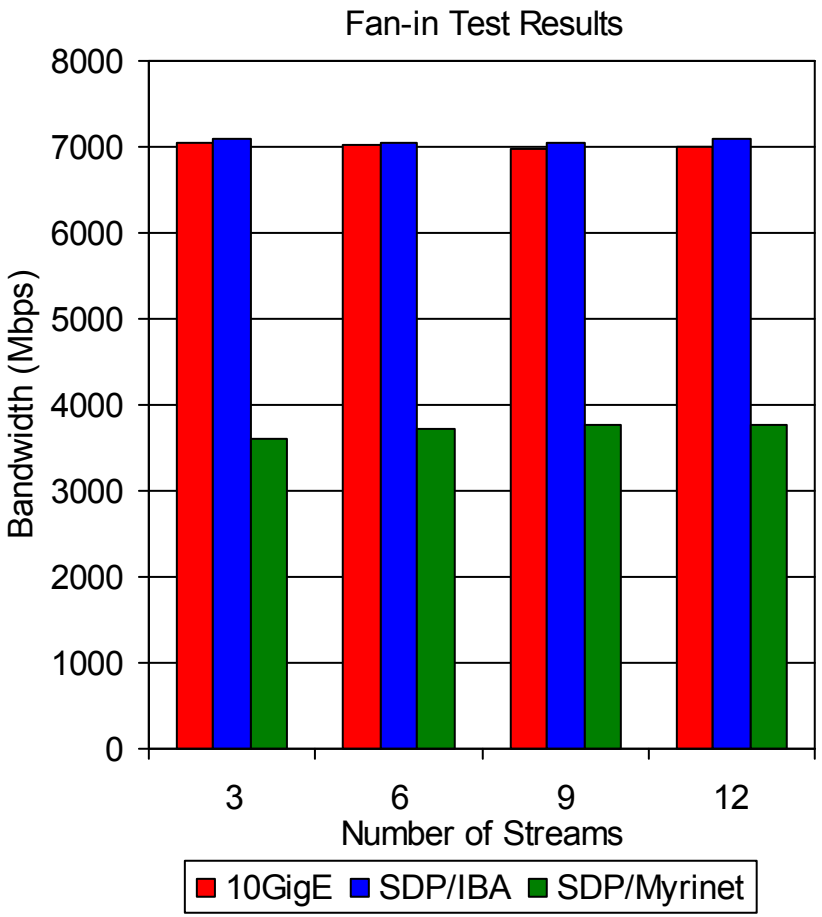


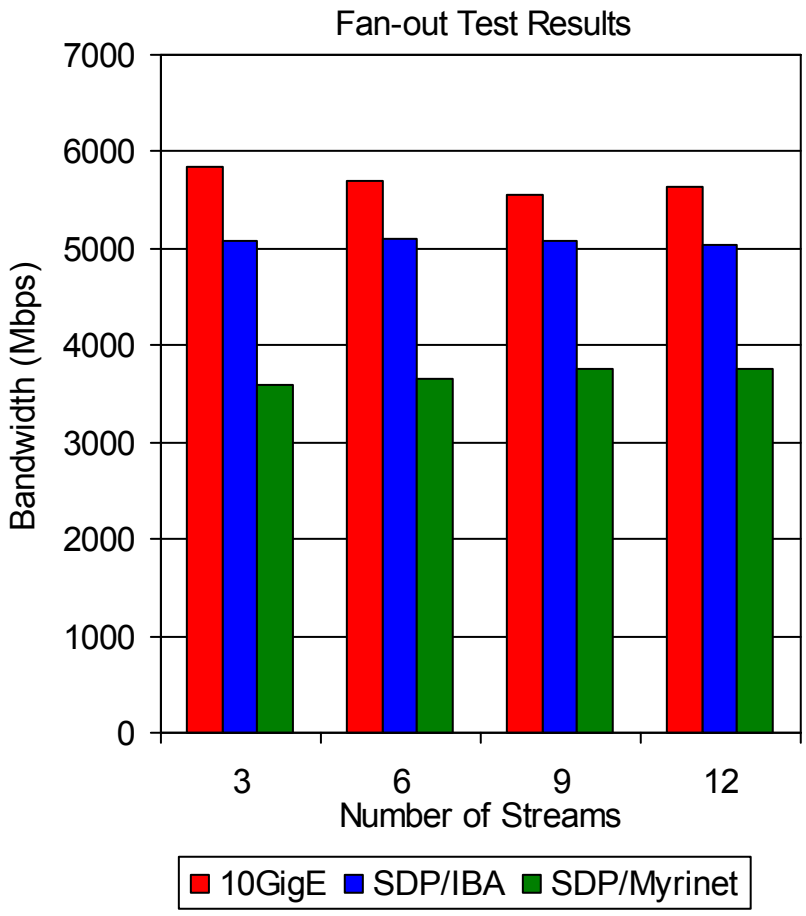
Hot-Spot Latency



- 10GigE and IBA demonstrate similar scalability with increasing number of clients
- Myrinet's performance deteriorates faster than the other two

Fan-in and Fan-out Throughput Test





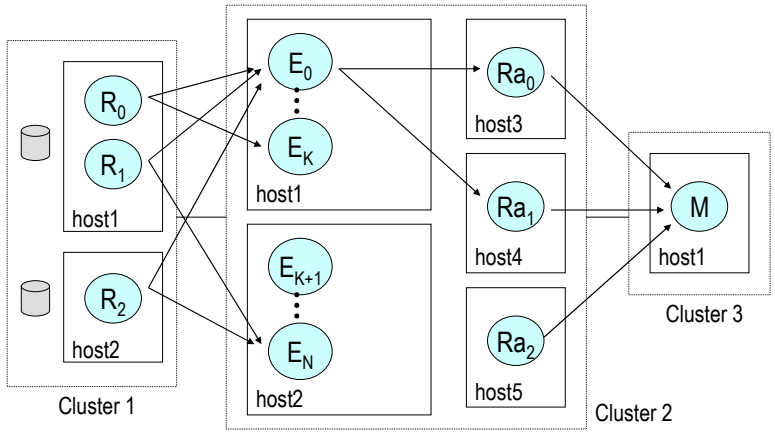
- 10GigE and IBA achieve a similar performance for the fan-in test
- 10GigE performs slightly better for the fan-out test

Data-Cutter Run-time Library

(Software Support for Data Driven Applications)

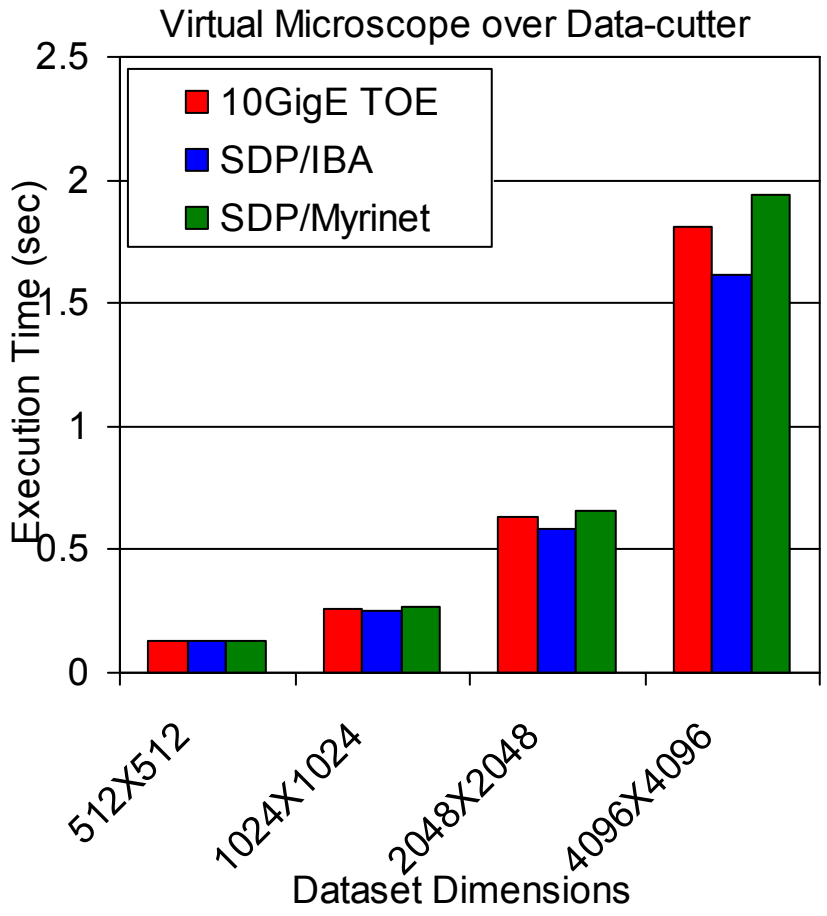
- Designed by Univ. of Maryland
- Component framework
- User-defined pipeline of components
  - Each component is a filter
  - The collection of components is a filter group
  - Replicated filters as well as filter groups
  - Illusion of a single stream in the filter group
  - Stream based communication
- Flow control between components
  - Unit of Work (UOW) based flow control
  - Each UOW contains about 16 to 64KB of data
- Several applications supported
  - Virtual Microscope
  - ISO Surface Oil Reservoir Simulator

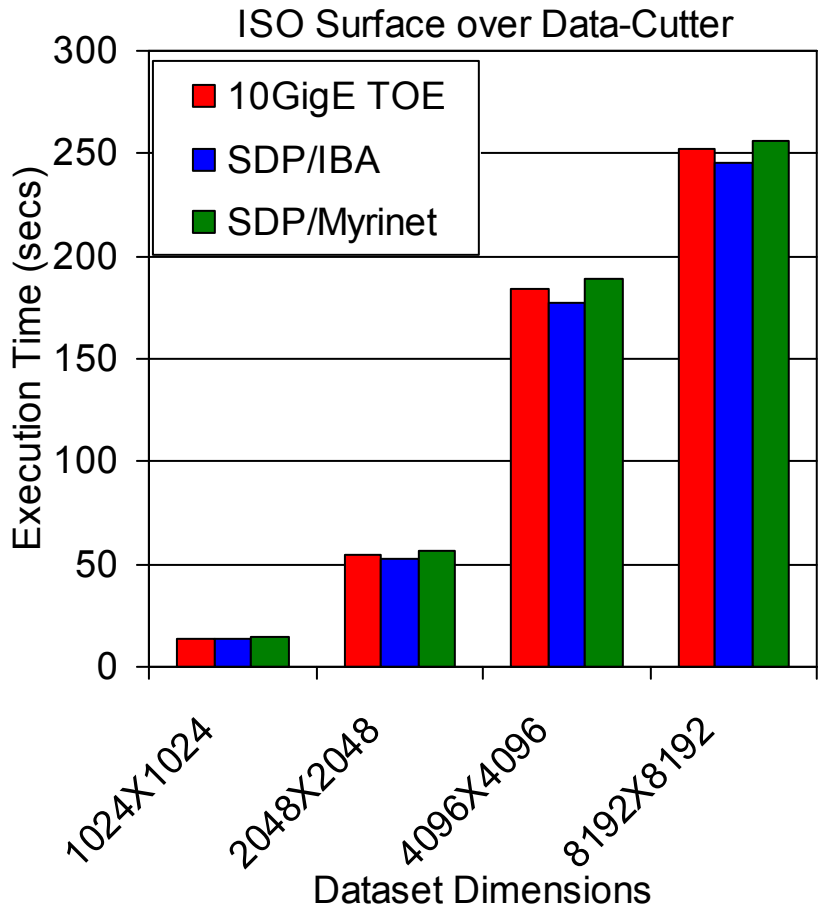
Combined Data/Task Parallelism



http://www.datacutter.org

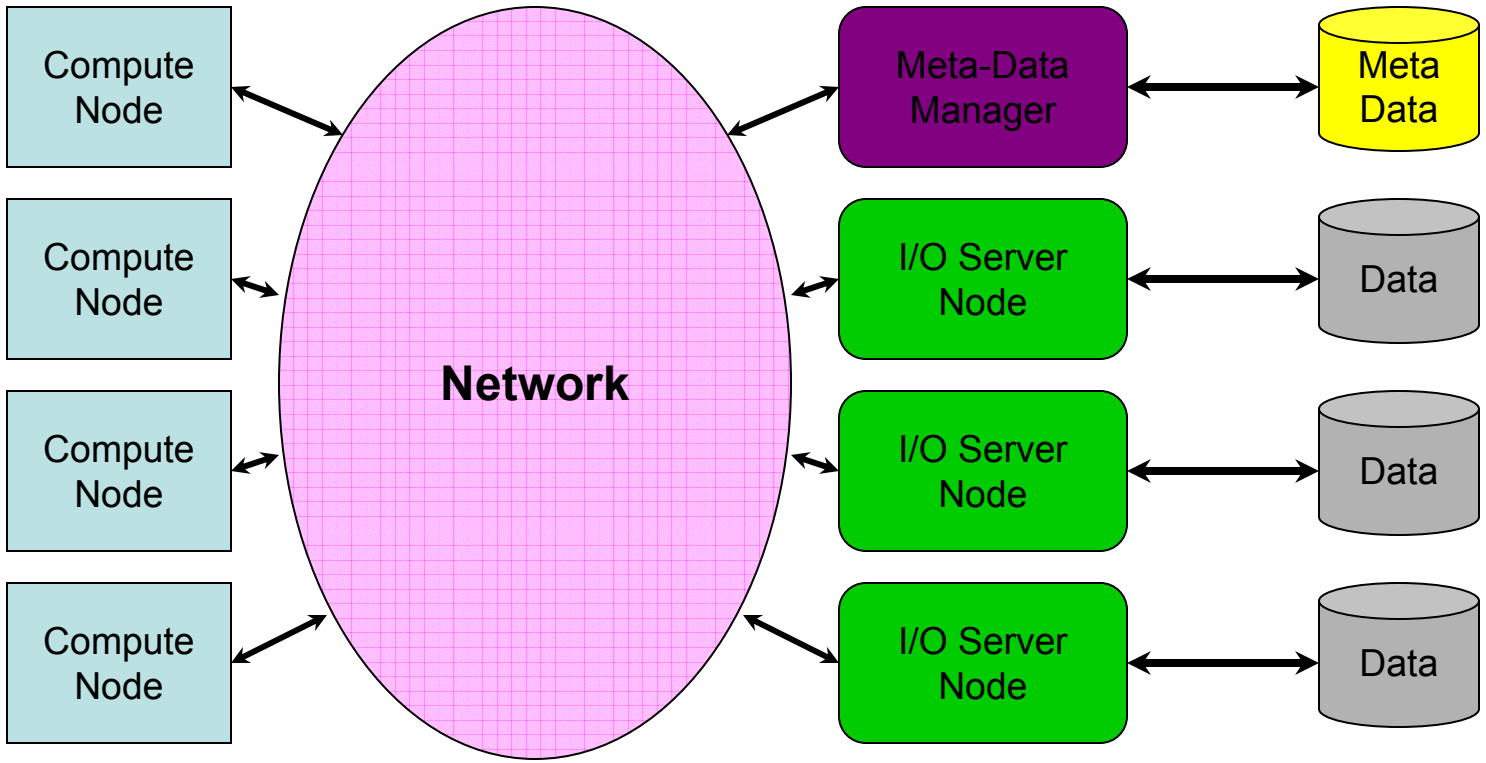
Data-Cutter Performance Evaluation





- InfiniBand performs the best for both the data-cutter applications (especially Virtual Microscope)
- The filter-based approach makes the environment medium message latency sensitive

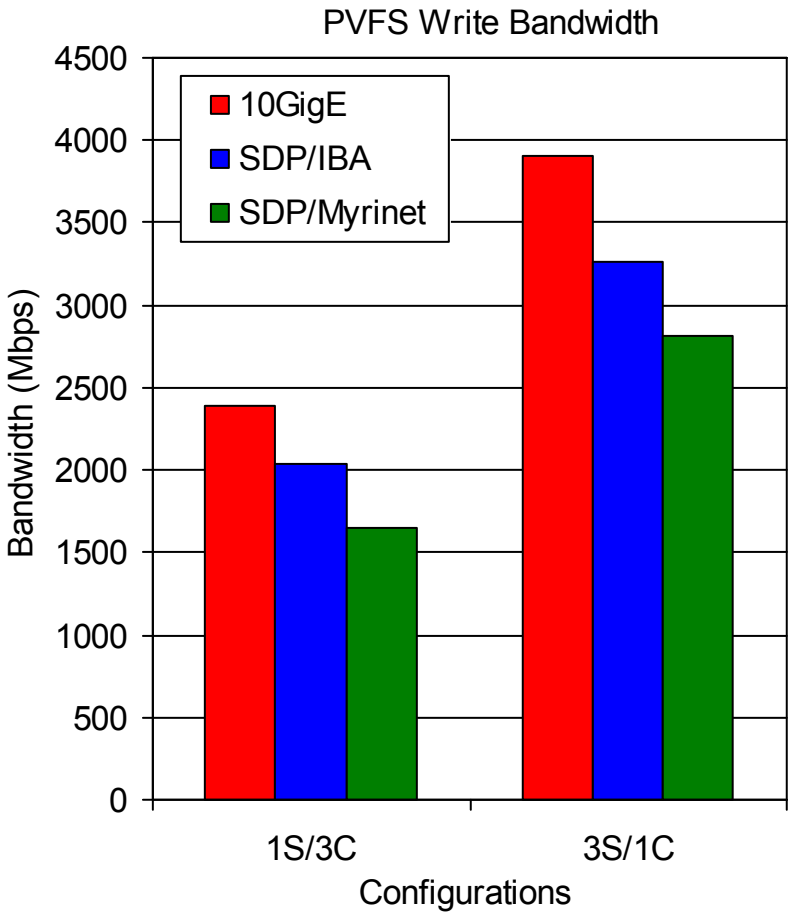
Parallel Virtual File System (PVFS)

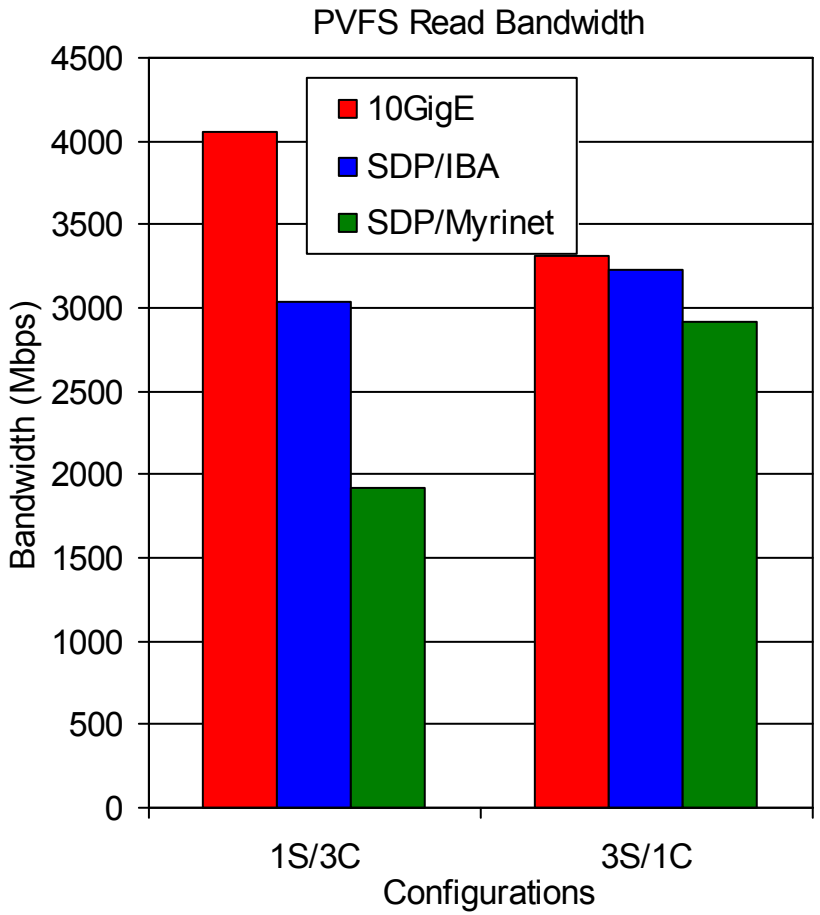


Designed by ANL and Clemson University

- Relies on Striping of data across different nodes
- Tries to aggregate I/O bandwidth from multiple nodes
- Utilizes the local file system on the I/O Server nodes

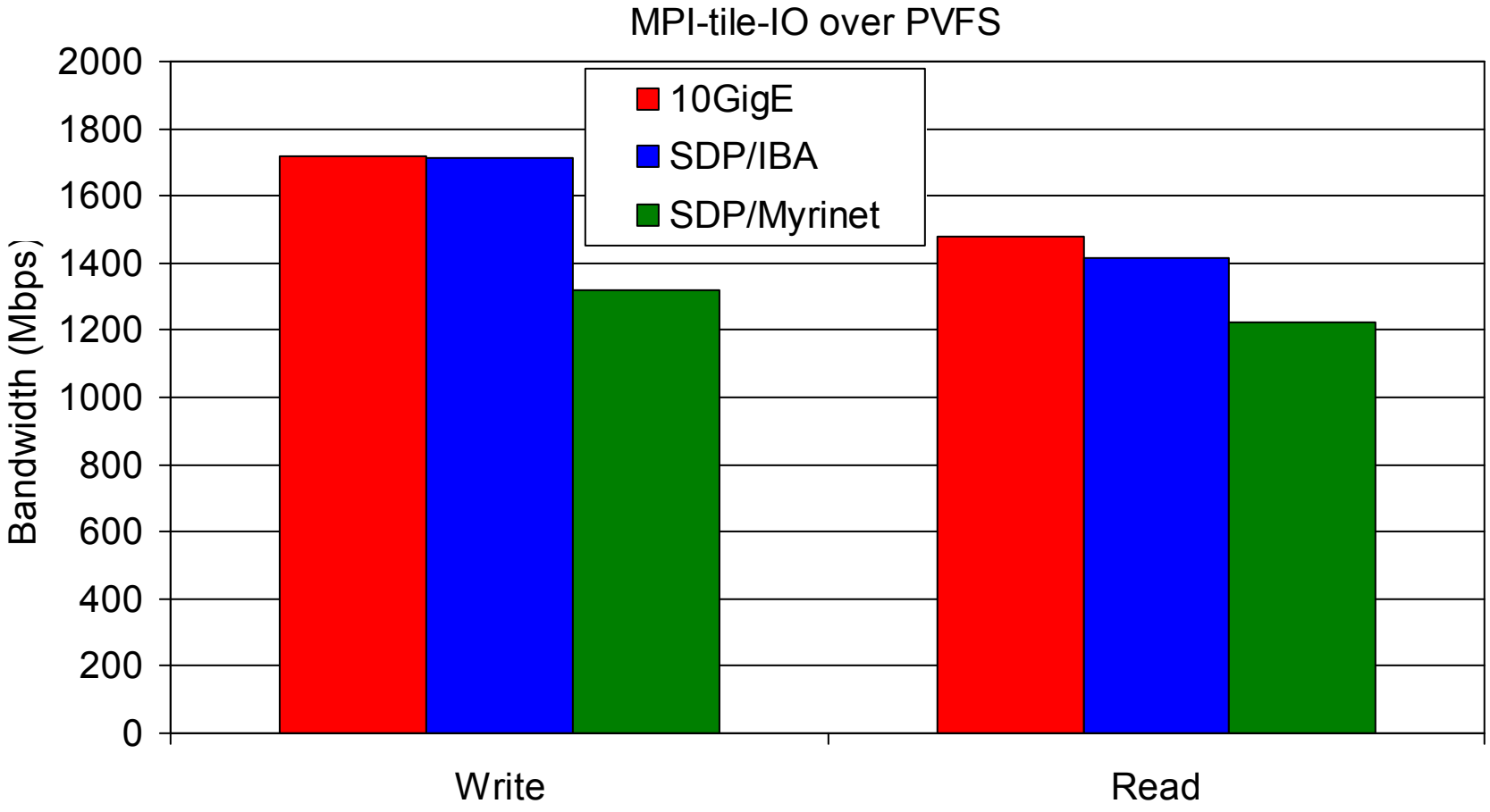
PVFS Contiguous I/O





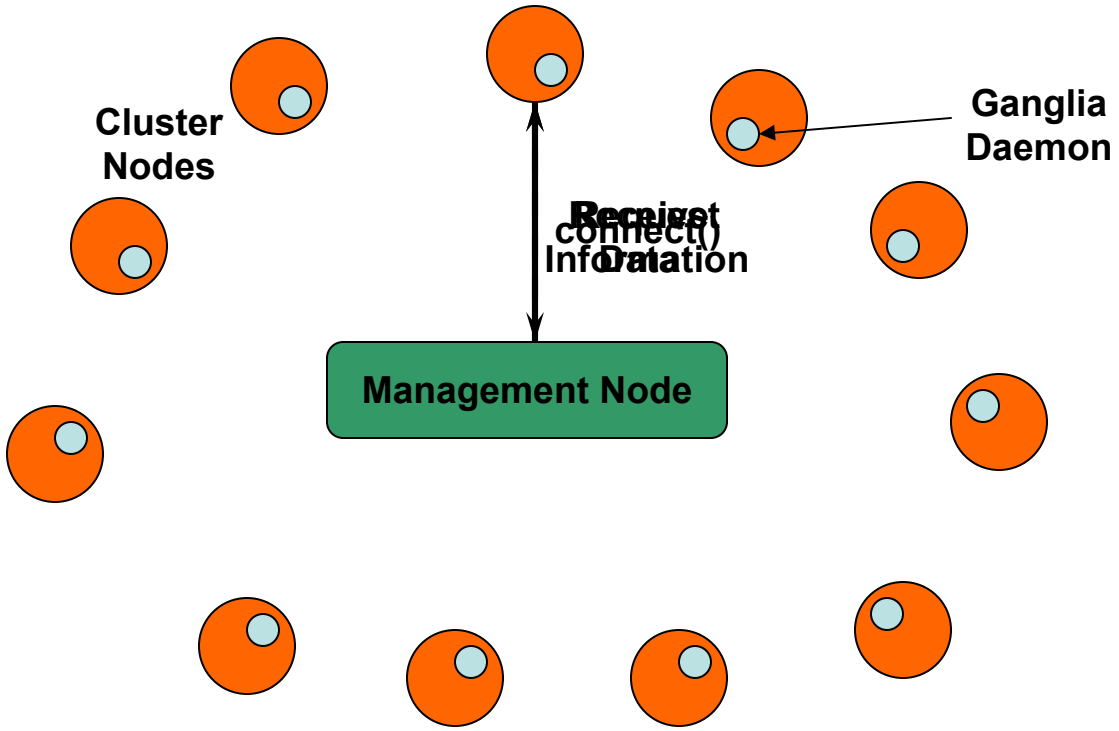
- Performance trends are similar to the throughput test
  - Experiment is throughput intensive

MPI-Tile I/O (PVFS Non-contiguous I/O)



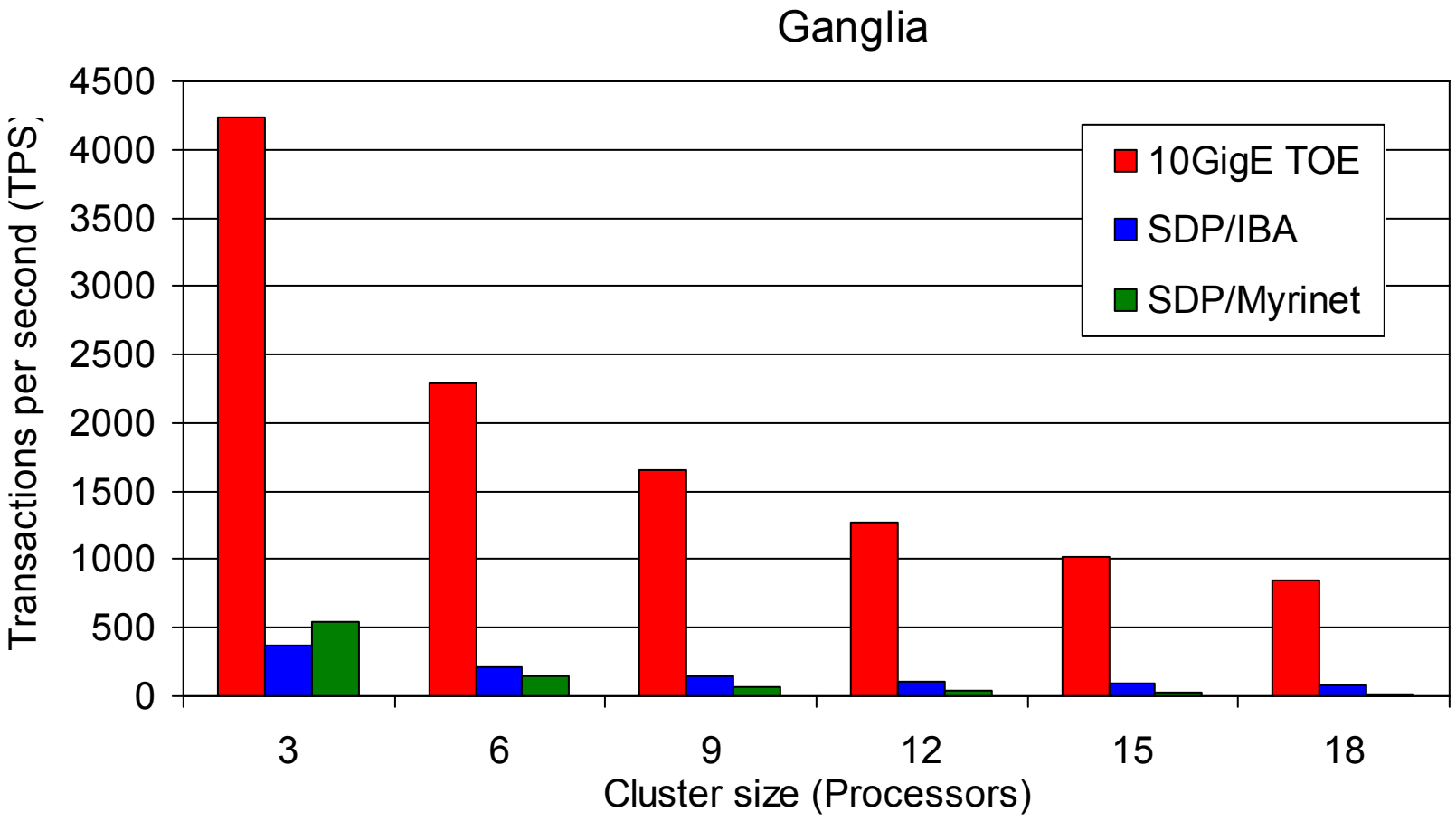
- 10GigE and IBA perform quite equally
- Myrinet is very close behind inspite of being only a 4Gbps network

Ganglia Cluster Management Infrastructure



- Developed by UC Berkeley
- For each transaction, one connection and one medium sized message (~6KB) transfer is required

Ganglia Performance Evaluation



- Performance is dominated by connection time
  - IBA, Myrinet take about a millisecond to establish connections while 10GigE takes about 60us
  - Optimizations for SDP/IBA and SDP/Myrinet such as connection caching are possible (being implemented)

Presentation Overview

- Introduction and Motivation
- High Performance Sockets over Protocol Offload Engines
- Experimental Evaluation
- **Conclusions and Future Work**

Concluding Remarks

- Ethernet has traditionally been notorious for performance reasons
  - Close to an order-of-magnitude performance gap compared to IBA/Myrinet
- 10GigE: Recently introduced as a successor in the Ethernet family
  - Can 10GigE help bridge the gap for Ethernet with IBA/Myrinet?
- We showed comparisons between Ethernet, IBA and Myrinet
  - Sockets Interface was used for the comparison
  - Micro-benchmark as well as application-level evaluations have been shown
- 10GigE performs quite comparably with the IBA and Myrinet
  - Better in some cases and worse in others; but around the same ballpark
  - Quite ubiquitous in Grid and WAN environments
  - Comparable performance in SAN environments

Continuing and Future Work

- Sockets is only one end of the comparison chart
  - Other middleware are quite widely used too (e.g., MPI)
  - IBA/Myrinet might have an advantage due to RDMA/multicast capabilities
- Network interfaces and software stacks change
  - Myrinet coming out with 10Gig adapters
  - 10GigE might release a PCI-Express based card
  - IBA has a zero-copy sockets interface for improved performance
  - IBM's 12x InfiniBand adapters increase the performance of IBA by 3 fold
  - *These results keep changing with time; more snapshots needed for fairness*
- Multi-NIC comparison for Sockets/MPI
  - Awful lot of work at the host
  - Scalability might be bound by the host
- iWARP compatibility and features for 10GigE TOE adapters

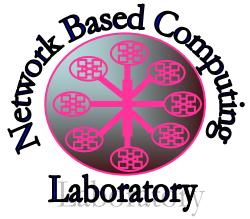
Acknowledgements

Our research is supported by the following organizations

- Funding support by

- Equipment donations by

Web Pointers



NBCL

LANL

Websites:

http://nowlab.cse.ohio-state.edu

http://public.lanl.gov/radiant/index.html

Emails:

balaji@cse.ohio-state.edu

feng@lanl.gov

panda@cse.ohio-state.edu