NETWORK-BASED COMPUTING LABORATORY

Sockets vs. RDMA Interface over 10-Gigabit Networks: An In-depth Analysis of the Memory Traffic Bottleneck

Pavan Balaji[ф] Hemal V. Shah[¥] D. K. Panda[ф]

[ф]Network Based Computing Lab
Computer Science and Engineering
Ohio State University

[¥]Embedded IA Division
Intel Corporation
Austin, Texas

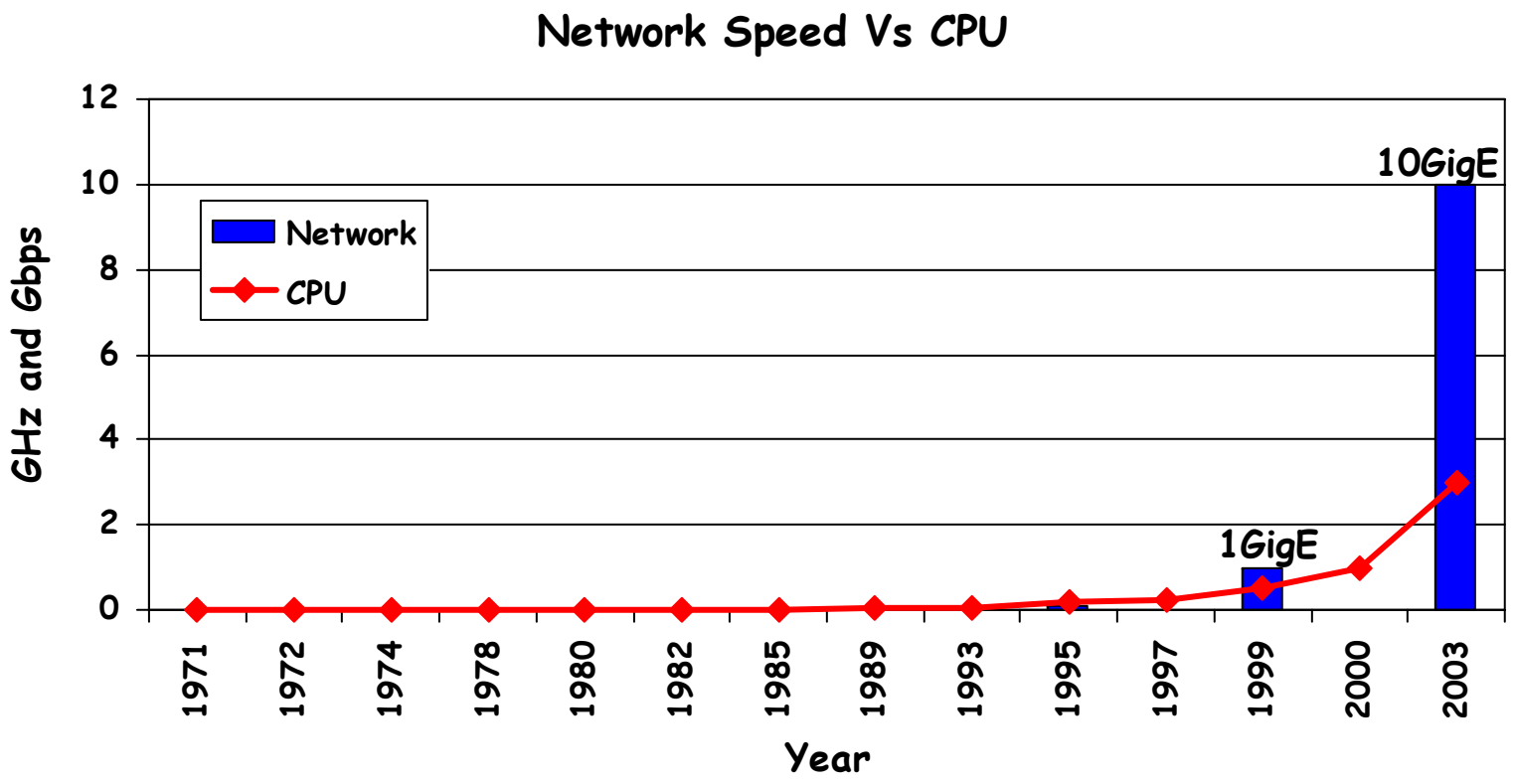
OHIO STATE

Introduction and Motivation

- Advent of High Performance Networks
  - Ex: InfiniBand, 10-Gigabit Ethernet, Myrinet, etc.
  - High Performance Protocols: VAPI / IBAL, GM
  - Good to build new applications
  - Not so beneficial for existing applications
    - Built around portability: Should run on all platforms
    - TCP/IP based sockets: A popular choice
    - Several ***GENERIC*** optimizations proposed and implemented for TCP/IP
      - Jacobson Optimization: Integrated Checksum-Copy [Jacob89]
      - Header Prediction for Single Stream data transfer

[Jacob89]: "An analysis of TCP Processing Overhead", D. Clark, V. Jacobson, J. Romkey and H. Salwen. IEEE Communications

Generic Optimizations Insufficient!



http://www.intel.com/research/silicon/mooreslaw.htm

- Processor Speed DOES NOT scale with Network Speeds
- Protocol processing too expensive for current day systems

Network Specific Optimizations

- Sockets can utilize some network features
  - Hardware support for protocol processing
  - Interrupt Coalescing (can be considered generic)
  - Checksum Offload (TCP stack has to modified)
  - Insufficient!
- Network Specific Optimizations
  - High Performance Sockets [shah99, balaji02]
  - TCP Offload Engines (TOE)

[shah99]: "High Performance Sockets and RPC over Virtual Interface (VI) Architecture", H. Shah, C. Pu, R. S. Madukkarumukumana, In CANPC '99

[balaji02]: "Impact of High Performance Sockets on Data Intensive Applications", P. Balaji, J. Wu, T. Kurc, U. Catalyurek, D. K. Panda, J. Saltz, In HPDC '03

Memory Traffic Bottleneck

- Offloaded Transport Layers provide some performance gains
  - Protocol processing is offloaded; lesser host CPU overhead
  - Better network performance for slower hosts
  - Quite effective for 1-2 Gigabit networks
  - Effective for faster (10-Gigabit) networks in some scenarios
- Memory Traffic Constraints
  - Offloaded Transport Layers rely on the sockets interface
  - Sockets API forces memory access operations in several scenarios
    - Transactional protocols such as RPC, File I/O, etc.
  - For 10-Gigabit networks memory access operations can limit network performance !

10-Gigabit Networks

- 10-Gigabit Ethernet
  - Recently released as a successor in the Ethernet family
  - Some adapters support TCP/IP checksum and Segmentation offload
- InfiniBand
  - Open Industry Standard
  - Interconnect for connecting compute and I/O nodes
  - Provides High Performance
    - Offloaded Transport Layer; Zero-Copy data-transfer
    - Provides one-sided communication (RDMA, Remote Atomics)
  - Becoming increasingly popular
  - An example RDMA capable 10-Gigabit network

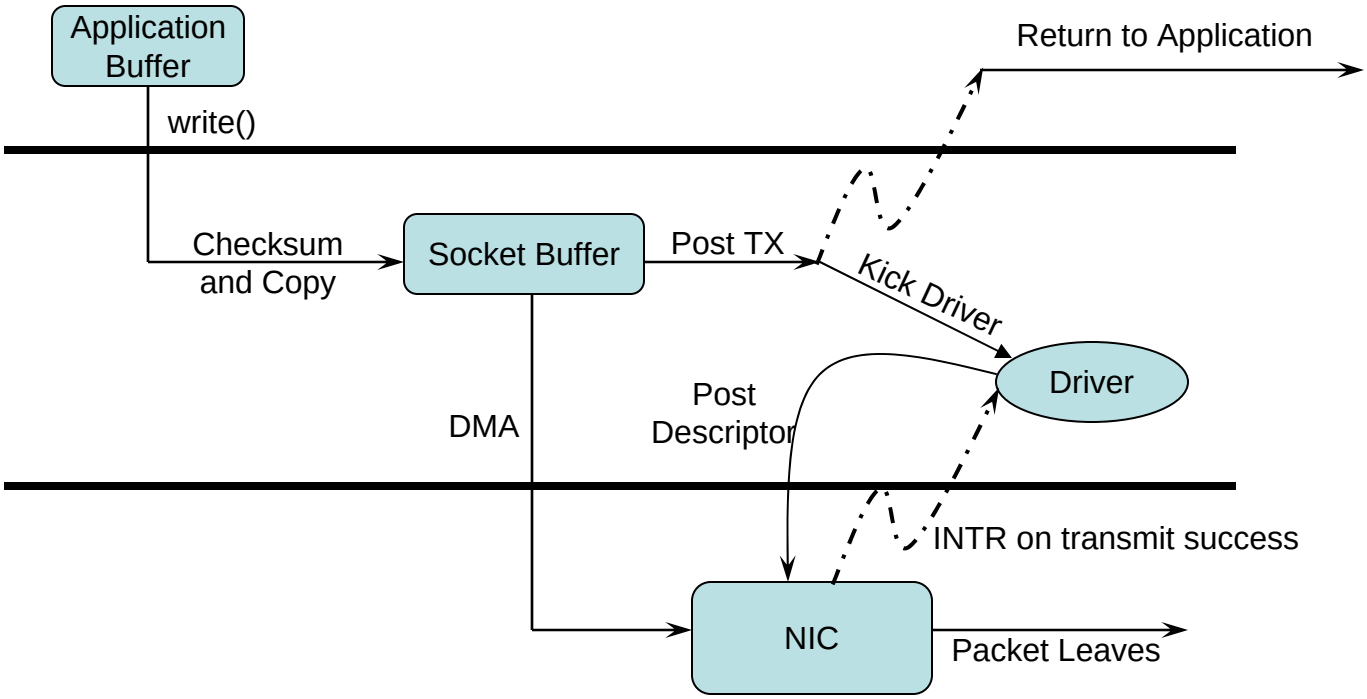
Objective

- New standards proposed for RDMA over IP
  - Utilizes an offloaded TCP/IP stack on the network adapter
  - Supports additional logic for zero-copy data transfer to the application
  - Compatible with existing Layer 3 and 4 switches
- What's the impact of an RDMA interface over TCP/IP?
  - Implications on CPU Utilization
  - Implications on Memory Traffic
  - Is it beneficial?
- We analyze these issues using InfiniBand's RDMA capabilities!

Presentation Outline

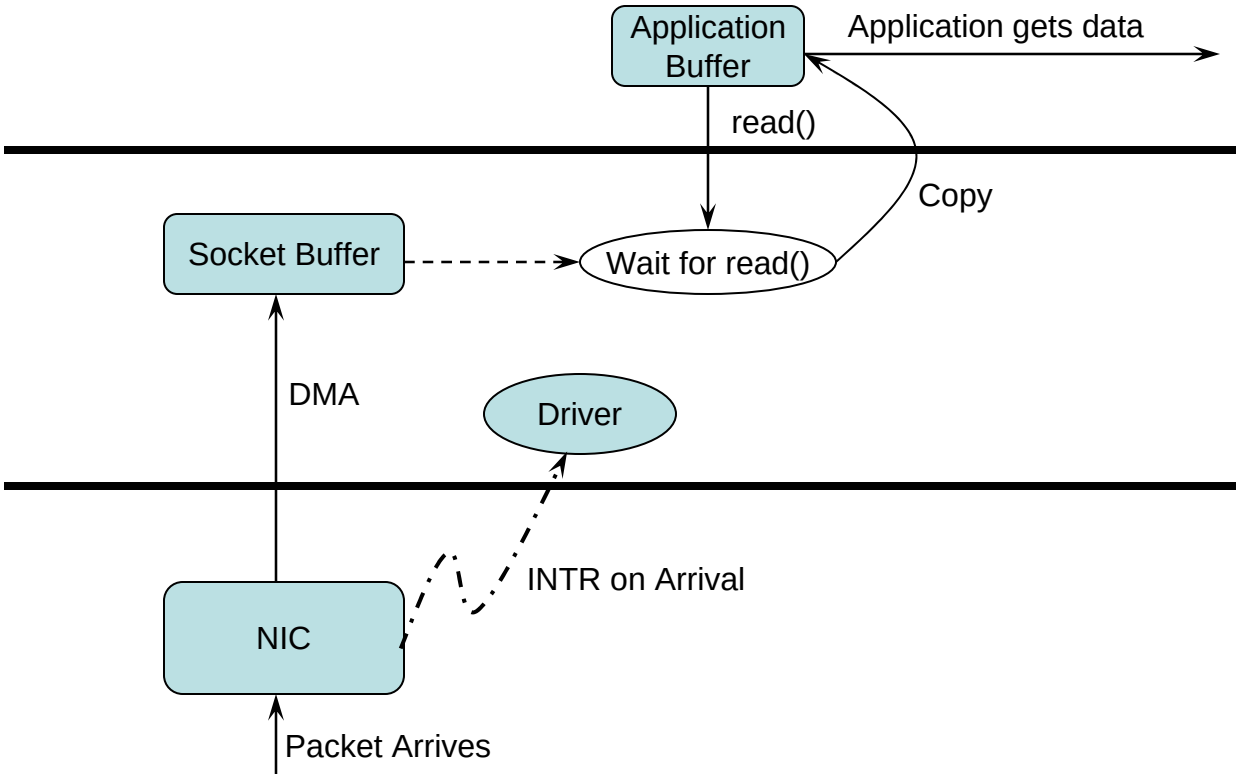
- Introduction and Motivation
- **TCP/IP Control Path and Memory Traffic**
- 10-Gigabit network performance for TCP/IP
- 10-Gigabit network performance for RDMA
- Memory Traffic Analysis for 10-Gigabit networks
- Conclusions and Future Work

TCP/IP Control Path (Sender Side)



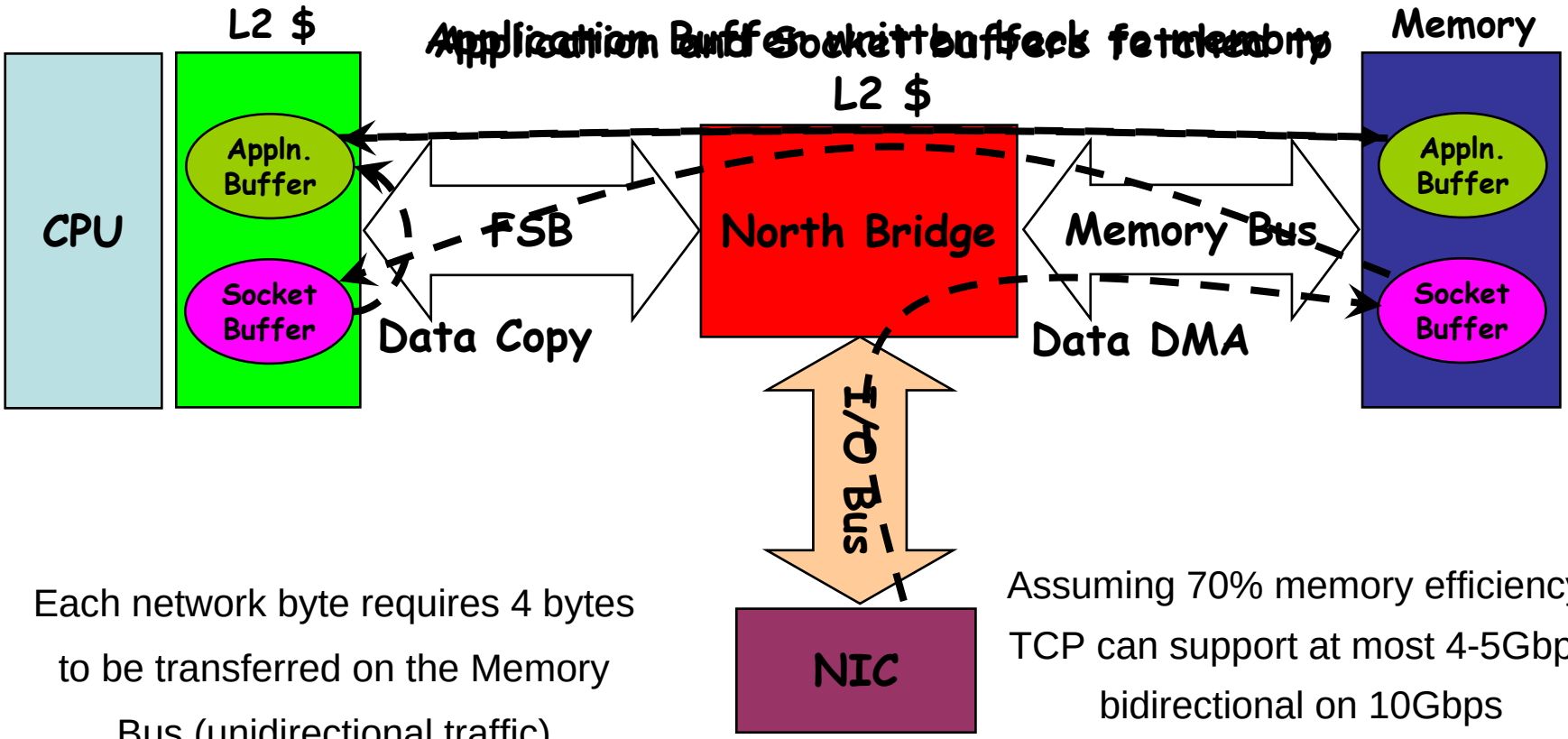
- Checksum, Copy and DMA are the data touching portions in TCP/IP
- Offloaded protocol stacks avoid checksum at the host; copy and DMA are still present

TCP/IP Control Path (Receiver Side)



- Data might need to be buffered on the receiver side
- Pick-and-Post techniques force a memory copy on the receiver side

Memory Bus Traffic for TCP



Each network byte requires 4 bytes to be transferred on the Memory Bus (unidirectional traffic)

Assuming 70% memory efficiency, TCP can support at most 4-5Gbps bidirectional on 10Gbps (400MHz/64bit FSB)

Network to Memory Traffic Ratio

	Application Buffer Fits in Cache	Application Buffer Doesn't fit in Cache
Transmit (Worst Case)	1-4	2-4
Transmit (Best Case)	**1**	**2-4**
Receive (Worst Case)	2-4	4
Receive (Best Case)	**2**	**4**

This table shows the minimum memory traffic associated with network data

In reality socket buffer cache misses, control messages and noise traffic may cause these to be higher

Details of other cases present in the paper

Presentation Outline

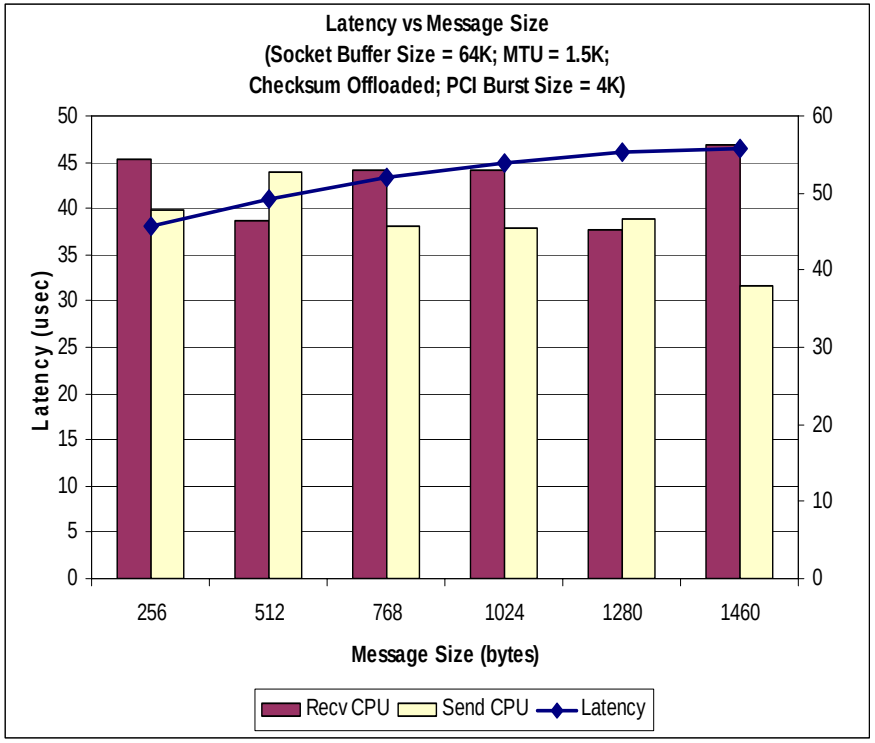
- Introduction and Motivation
- TCP/IP Control Path and Memory Traffic
- **10-Gigabit network performance for TCP/IP**
- 10-Gigabit network performance for RDMA
- Memory Traffic Analysis for 10-Gigabit networks
- Conclusions and Future Work

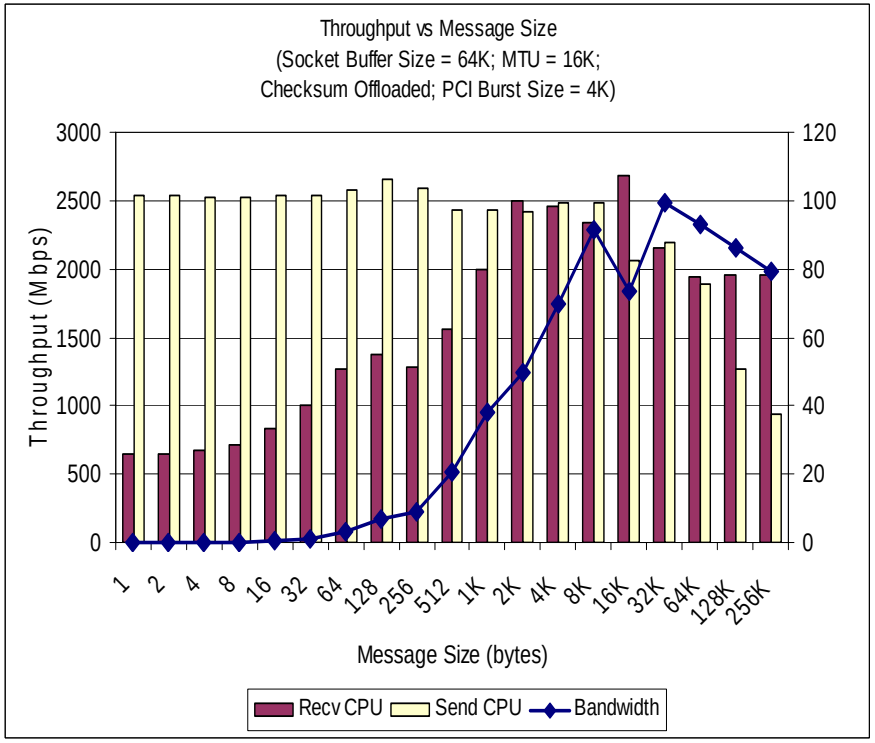
Experimental Test-bed (10-Gig Ethernet)

- Two Dell2600 Xeon 2.4GHz 2-way SMP node
- 1GB main memory (333MHz, DDR)
- Intel E7501 Chipset
- 32K L1, 512K L2, 400MHz/64bit FSB
- PCI-X 133MHz/64bit I/O bus
- Intel 10GbE/Pro 10-Gigabit Ethernet adapters

- 8 P4 2.0 GHz nodes (IBM xSeries 305; 8673-12X)
- Intel Pro/1000 MT Server Gig-E adapters
- 256K main memory

10-Gigabit Ethernet: Latency and Bandwidth

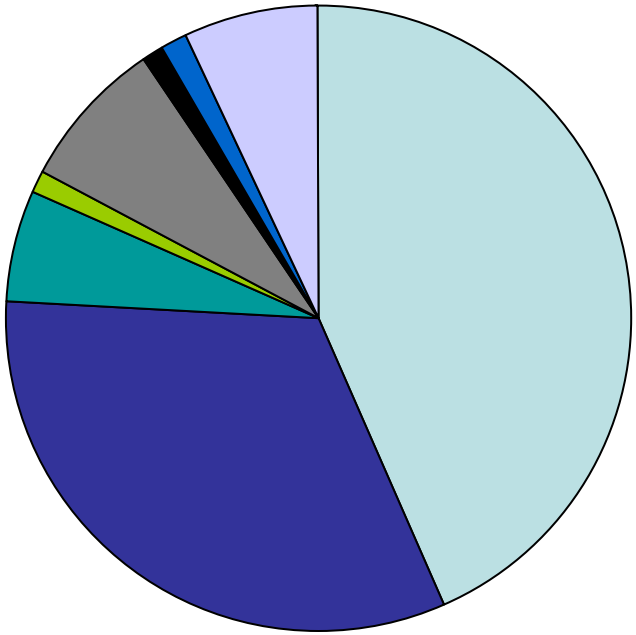




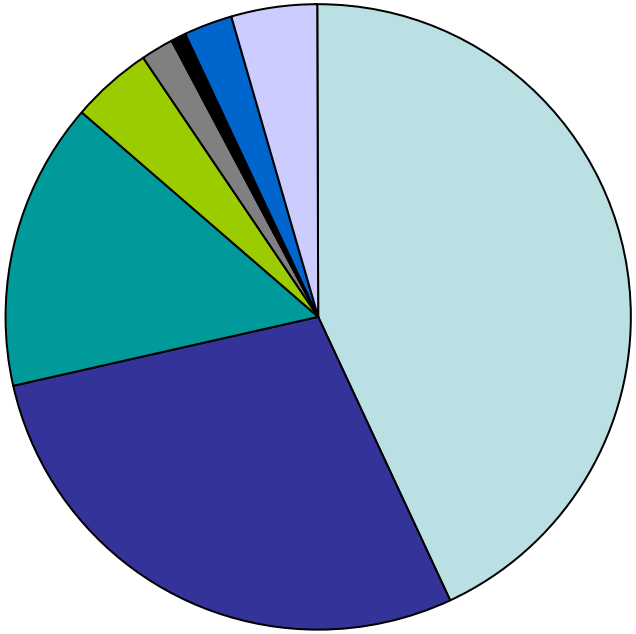
- TCP/IP achieves a latency of 37us (Win Server 2003) – 20us on Linux
  - About 50% CPU utilization on both platforms
- Peak Throughput of about 2500Mbps; 80-100% CPU Utilization
- Application buffer is always in Cache !!

TCP Stack Pareto Analysis (64 byte)

Sender



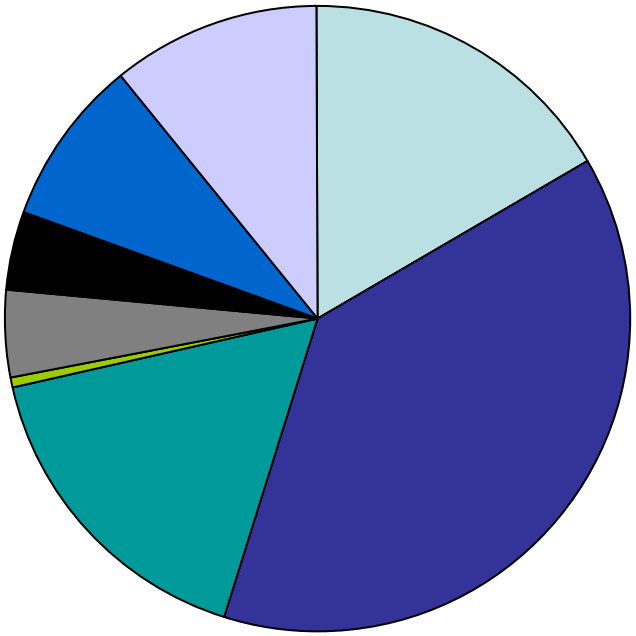
Receiver

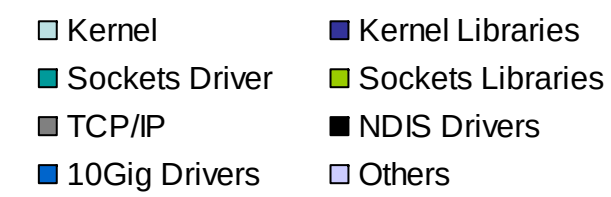


Kernel, Kernel Libraries and TCP/IP contribute to the Offloadable TCP/IP stack

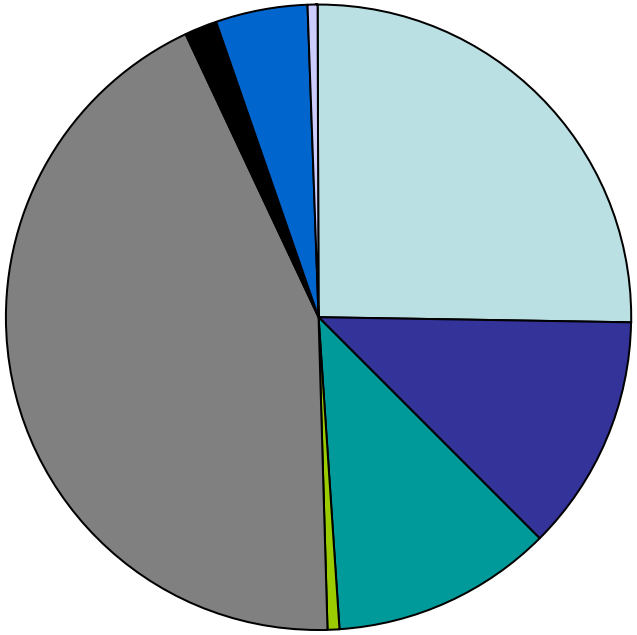
TCP Stack Pareto Analysis (16K byte)

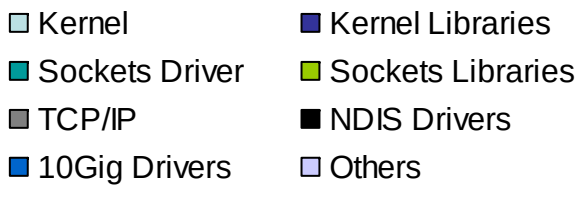
Sender





Receiver

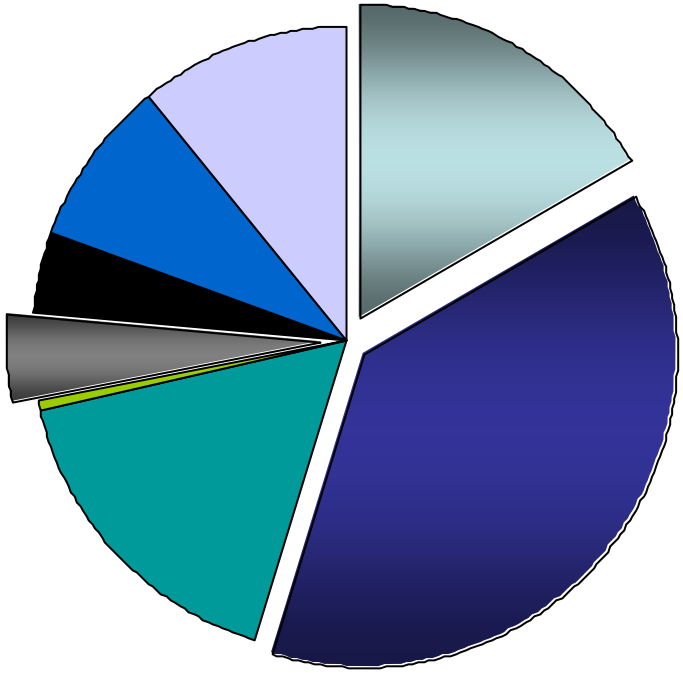




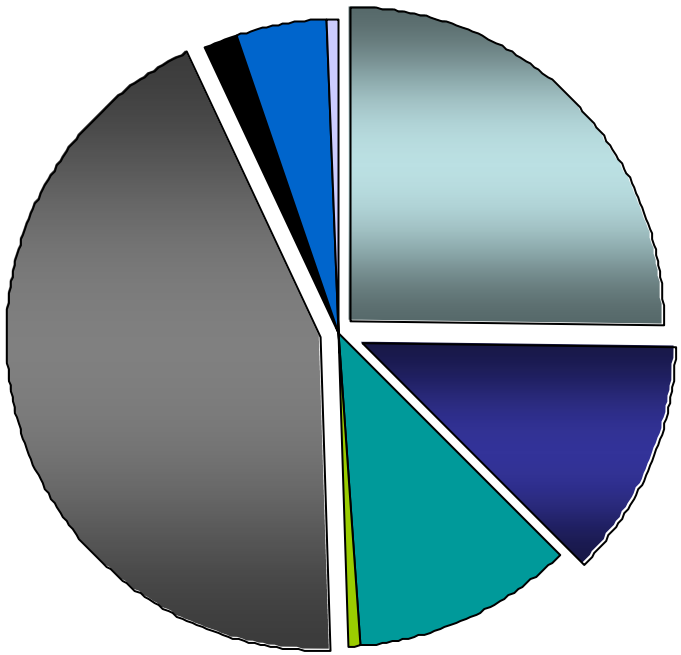
- TCP and other protocol overhead takes up most of the CPU
- Offload is beneficial when buffers fit into cache

TCP Stack Pareto Analysis (16K byte)

Sender



Receiver

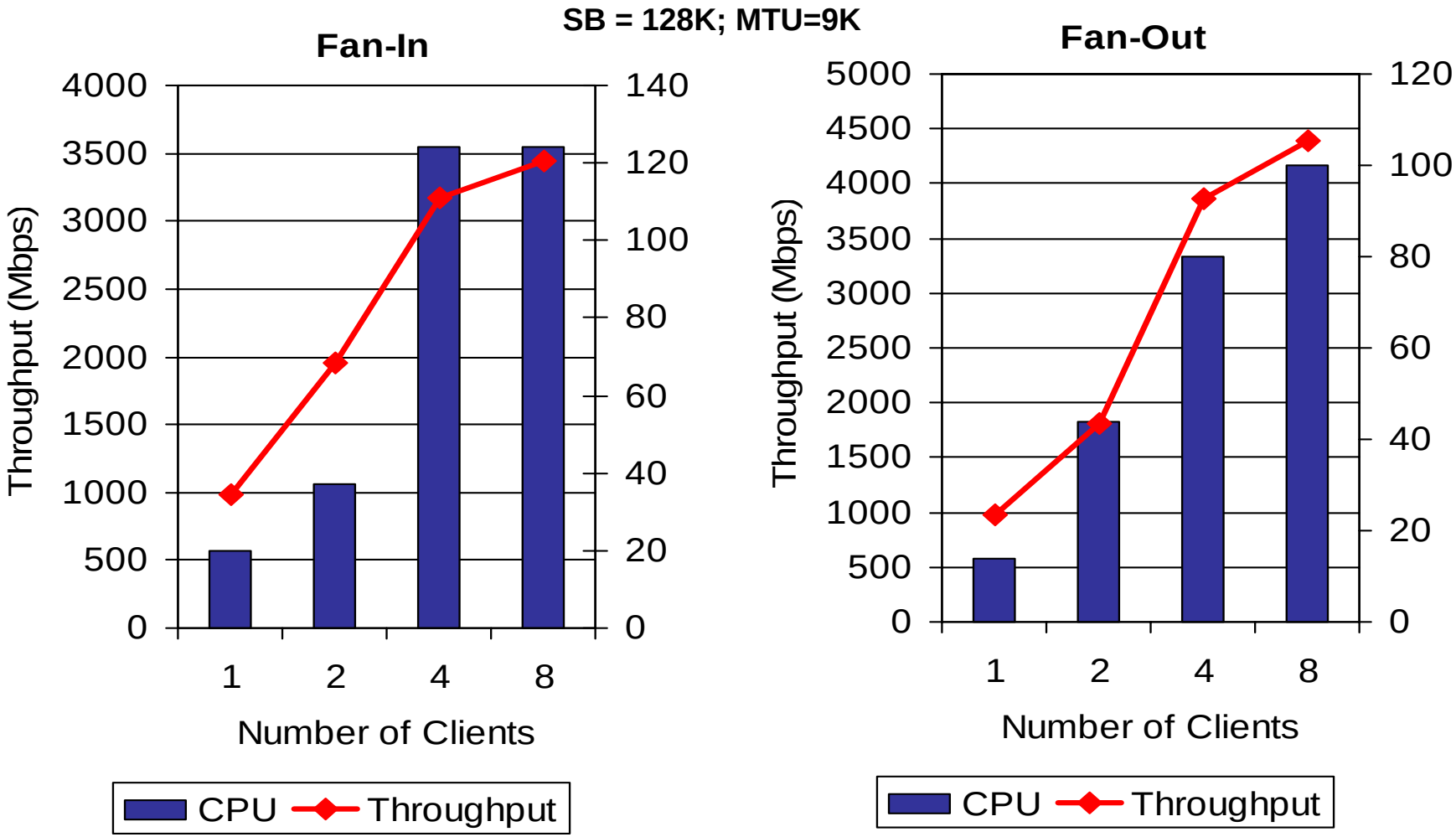


Kernel
Kernel Libraries
Sockets Driver
Sockets Libraries
TCP/IP
NDIS Drivers
10Gig Drivers
Others

Kernel
Kernel Libraries
Sockets Driver
Sockets Libraries
TCP/IP
NDIS Drivers
10Gig Drivers
Others

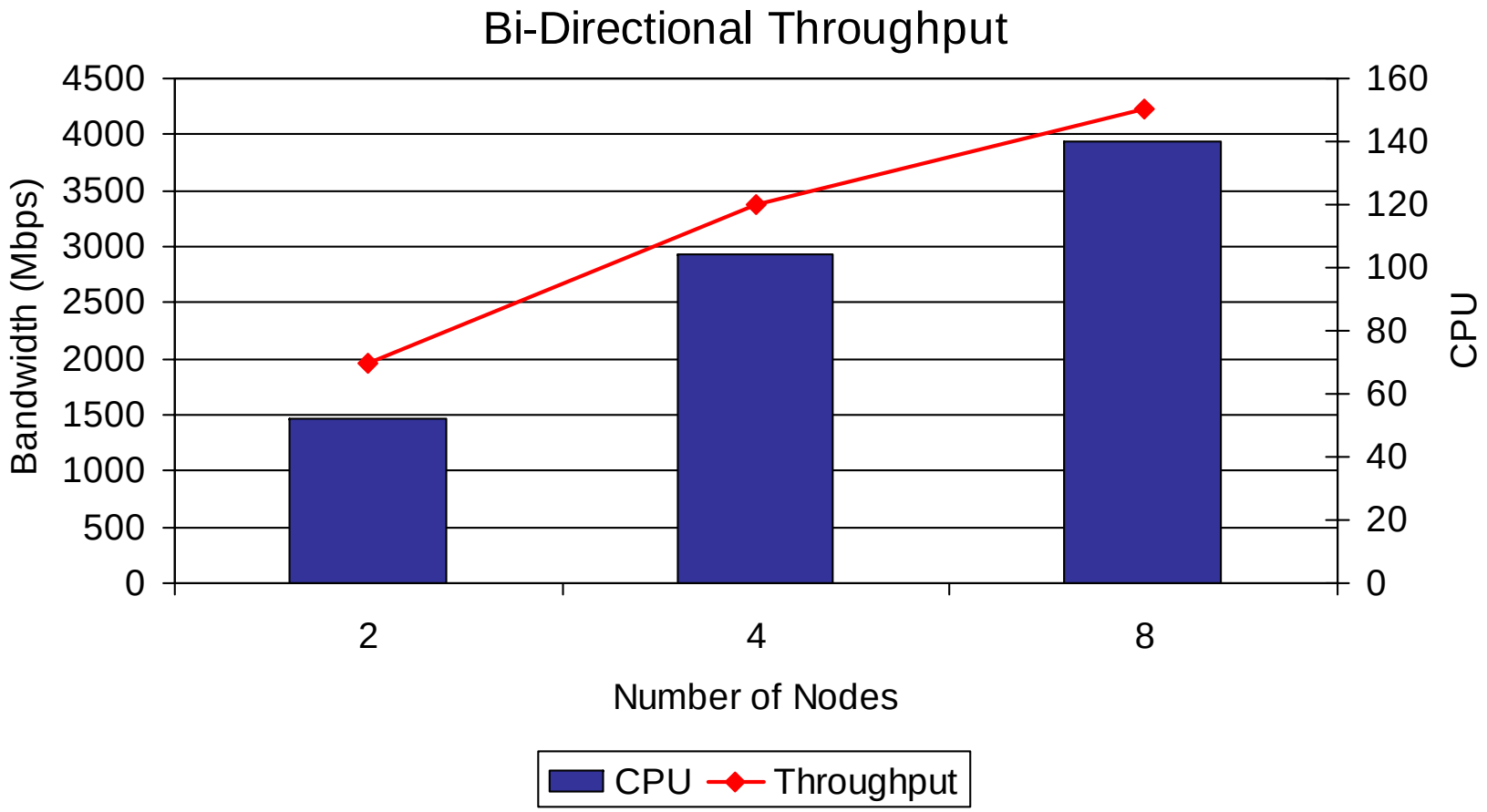
- TCP and other protocol overhead takes up most of the CPU
- Offload is beneficial when buffers fit into cache

Throughput (Fan-in/Fan-out)



Peak throughput of 3500Mbps for Fan-In and 4200Mbps for Fan-out

Bi-Directional Throughput



- Not the traditional Bi-directional Bandwidth test
- Fan-in with half the nodes and Fan-out with the other half

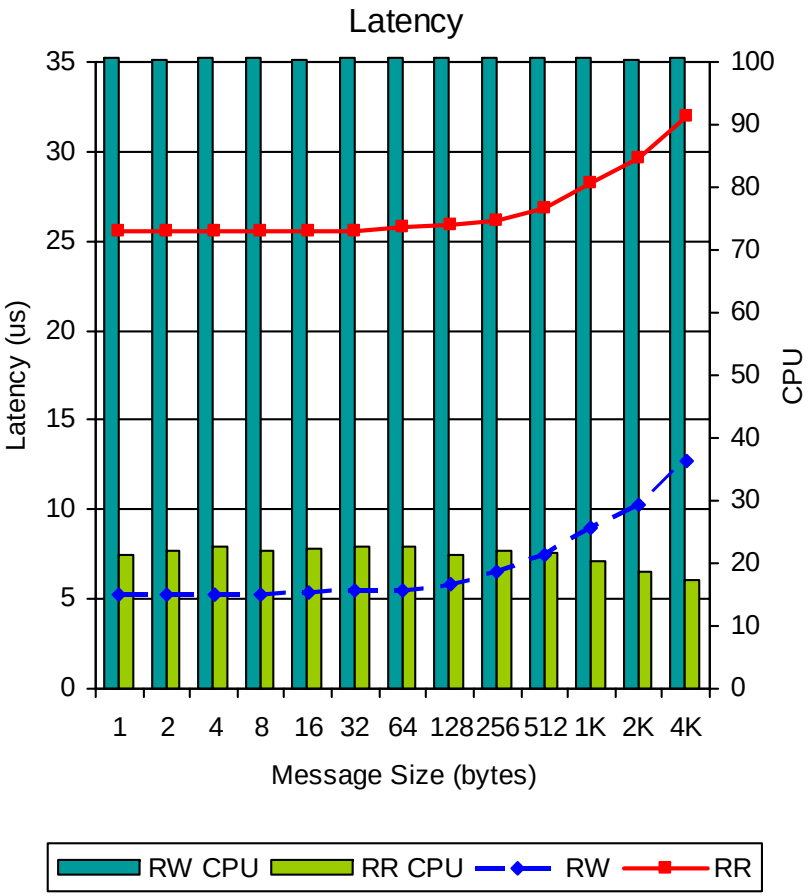
Presentation Outline

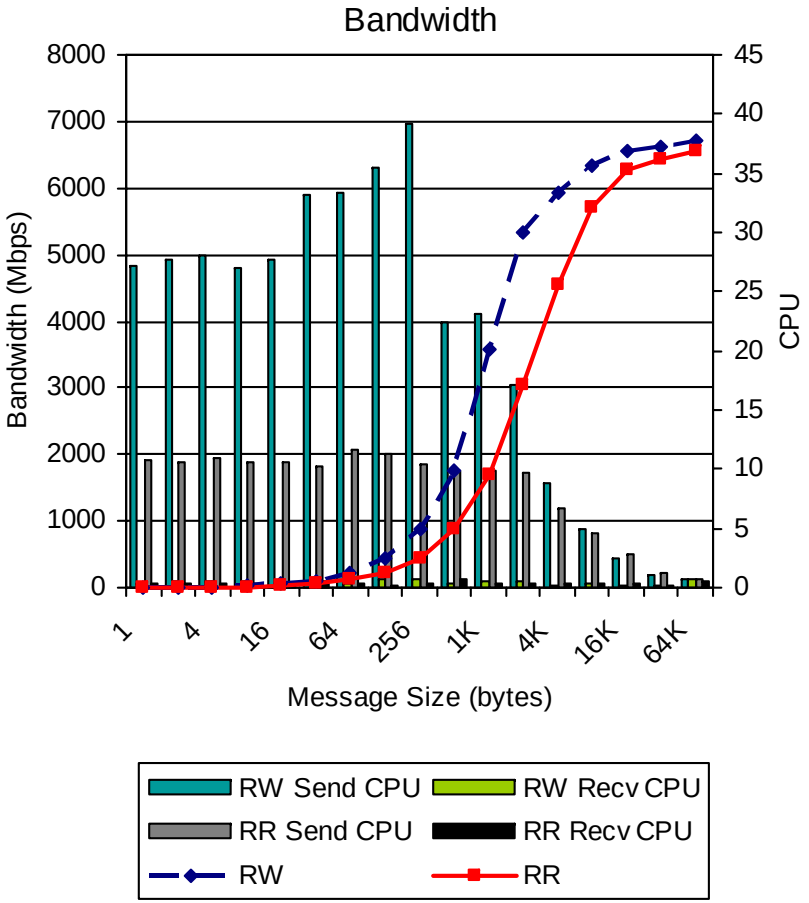
- Introduction and Motivation
- TCP/IP Control Path and Memory Traffic
- 10-Gigabit network performance for TCP/IP
- **10-Gigabit network performance for RDMA**
- Memory Traffic Analysis for 10-Gigabit networks
- Conclusions and Future Work

Experimental Test-bed (InfiniBand)

- 8 SuperMicro SUPER P4DL6 nodes
  - Xeon 2.4GHz 2-way SMP nodes
  - 512MB main memory (DDR)
  - PCI-X 133MHZ/64bit I/O bus
- Mellanox InfiniHost MT23108 DualPort 4x HCA
  - InfiniHost SDK version 0.2.0
  - HCA firmware version 1.17
- Mellanox InfiniScale MT43132 8-port switch (4x)
- Linux kernel version 2.4.7-10smp

InfiniBand RDMA: Latency and Bandwidth



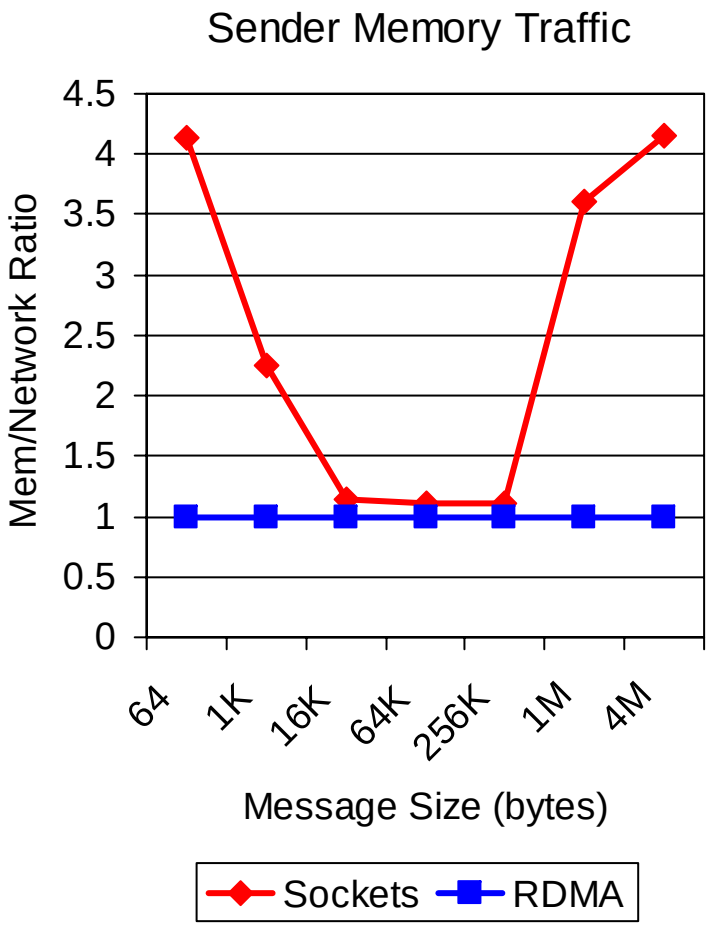


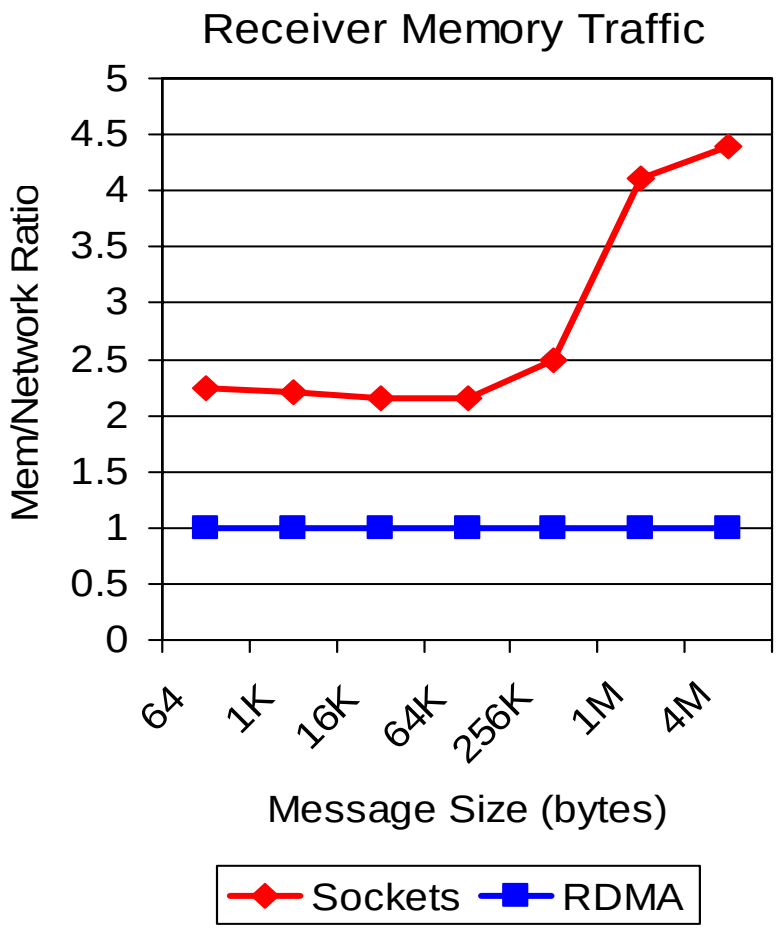
- Performance improvement due to hardware support and zero-copy data transfer
- Near zero CPU Utilization at the data sink for large messages
- Performance limited by PCI-X I/O bus

Presentation Outline

- Introduction and Motivation
- TCP/IP Control Path and Memory Traffic
- 10-Gigabit network performance for TCP/IP
- 10-Gigabit network performance for RDMA
- **Memory Traffic Analysis for 10-Gigabit networks**
- Conclusions and Future Work

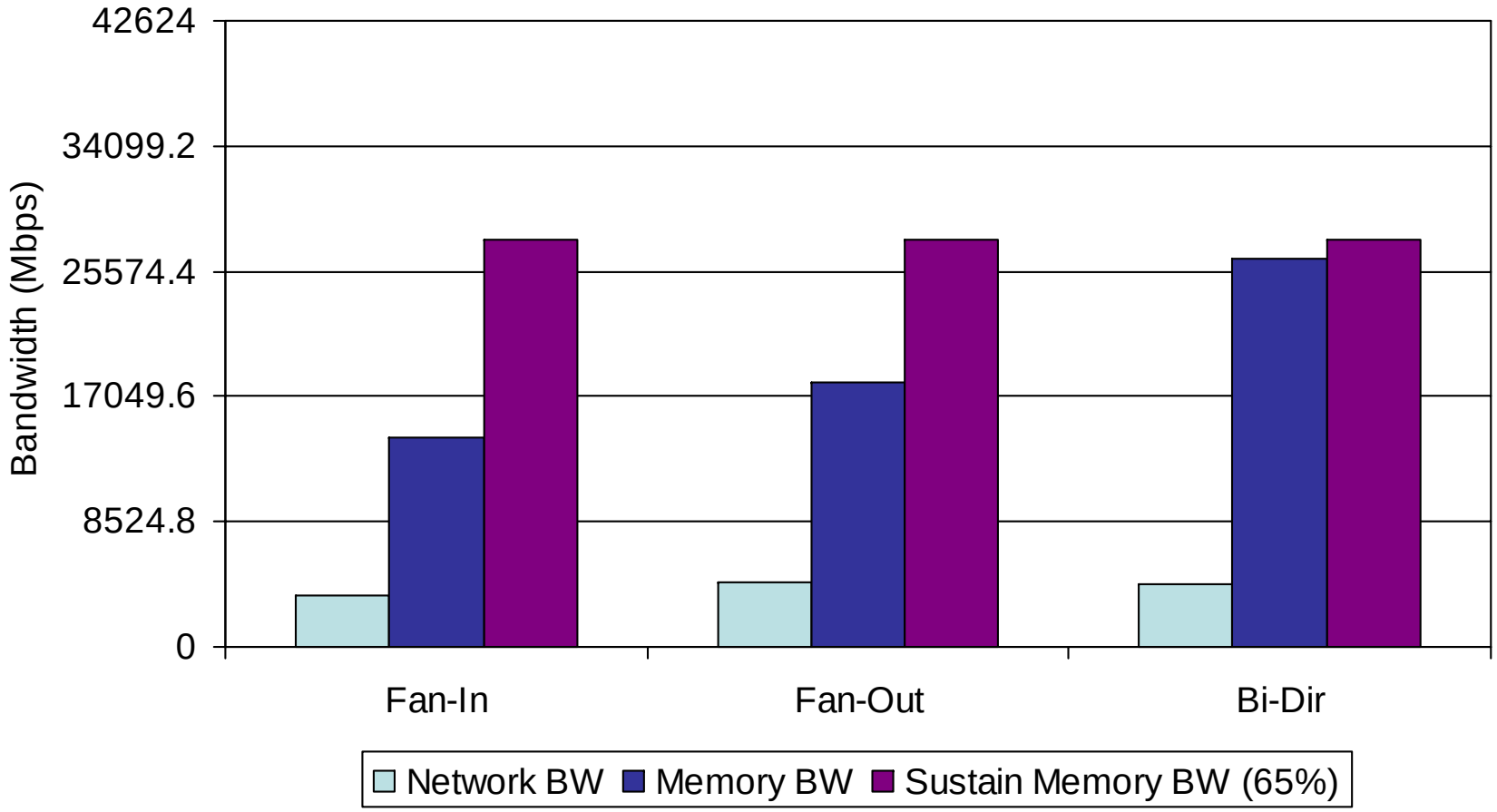
Throughput test: Memory Traffic





- Sockets can force up to 4 times more memory traffic compared to the network traffic
- RDMA allows has a ratio of 1 !!

Multi-Stream Tests: Memory Traffic



- Memory Traffic is significantly higher than the network traffic
- Comes to within 5% of the practically attainable peak memory bandwidth

Presentation Outline

- Introduction and Motivation
- TCP/IP Control Path and Memory Traffic
- 10-Gigabit network performance for TCP/IP
- 10-Gigabit network performance for RDMA
- Memory Traffic Analysis for 10-Gigabit networks
- **Conclusions and Future Work**

Conclusions

- TCP/IP performance on High Performance Networks
  - High Performance Sockets
  - TCP Offload Engines
- 10-Gigabit Networks
  - A new dimension of complexity – *Memory Traffic*
- Sockets API can require significant memory traffic
  - Up to 4 times more than the network traffic
  - Allows saturation on less than 35% of the network bandwidth
  - Shows potential benefits of providing RDMA over IP
  - Significant benefits in performance, CPU and memory traffic

Future Work

- Memory Traffic Analysis for 64-bit systems
- Potential of the L3-Cache available in some systems
- Evaluation of various applications
  - Transactional (SpecWeb)
  - Streaming (Multimedia Services)

Thank You!

For more information, please visit the

Home Page

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

Network Based Computing Laboratory,

The Ohio State University

Backup Slides