Boosting ProLiant Network Performance

David Koenen, Chuck Hudson
ProLiant Networking
Industry Standard Server Group

HP WORLD 2003
Solutions and Technology Conference & Expo

Before we begin..

- The system performance is limited by the slowest element in the link.
- Network file transfer performance is limited by the local file transfer performance.
- To keep focus on network performance, only socket to socket performance data is presented. No disk/file subsystem was involved.
- The Ethernet protocol limits the max throughput to:

$$\text{Max Thruput} = \frac{\textit{Avg Data Payload Size}}{\textit{Hdr Size} + \textit{Avg Data Payload Size}}$$

- All performance data given in this presentation were measured using NetIQ's™ Chariot "FileSendLong" script.

Top Performance Tips

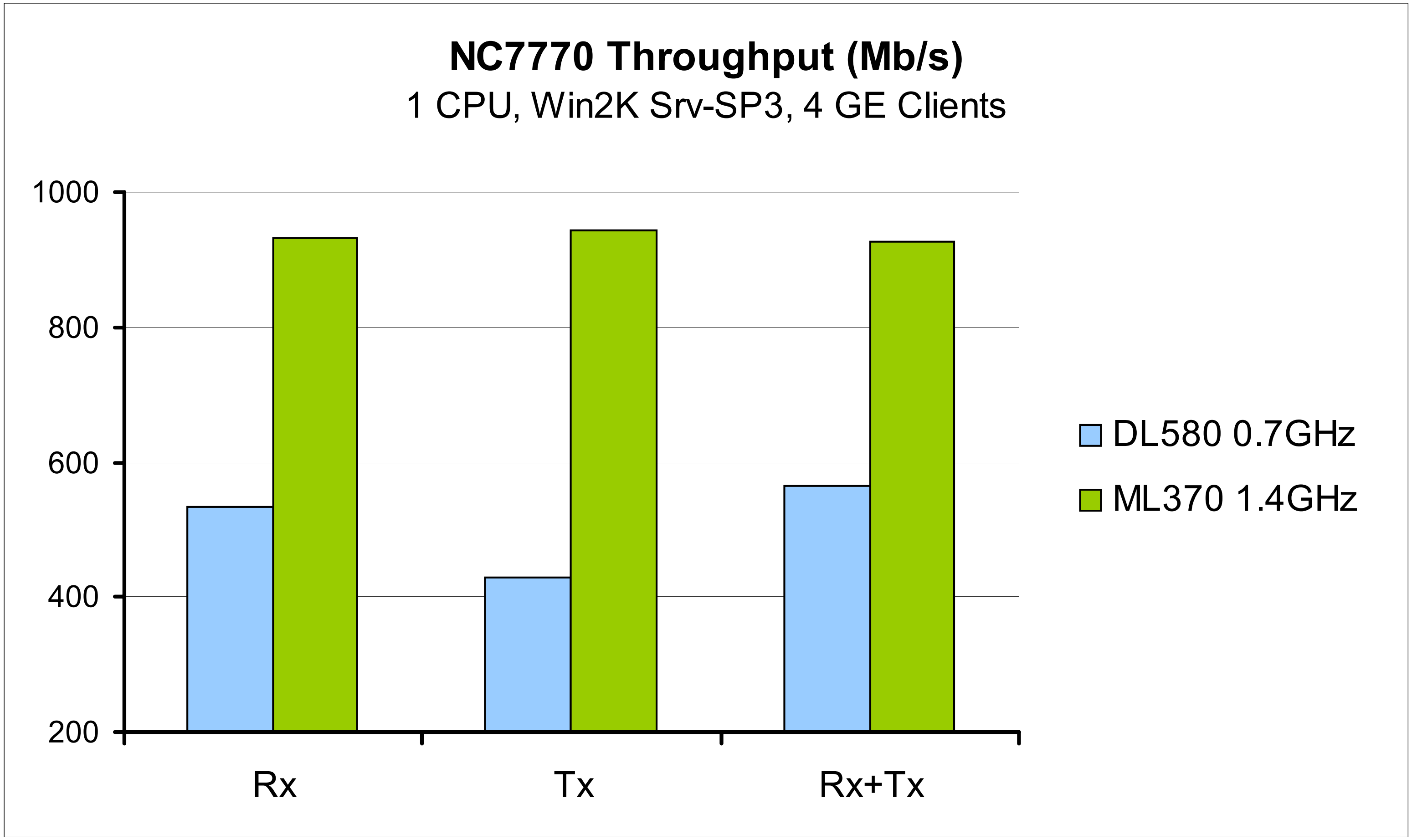
- Today:
  - Upgrade to faster hardware: CPU, NIC, & PCI-Bus
  - Use multiple, simultaneous transfers
  - Enable Jumbo Frames between endpoints
  - Buy/write applications with Asynchronous I/O support
  - Use NIC and O/S with Large Send Offload support
  - Use multiple NICs: multi-subnets or Link-Aggregation (team/trunk)
- Future:
  - Receive Side Scaling (RSS)
  - TCP Offload Engines (TOE)
  - Remote Direct Memory Access Procotol (RDMAP)

Faster Hardware

Faster Hardware Increases Server Network Performance

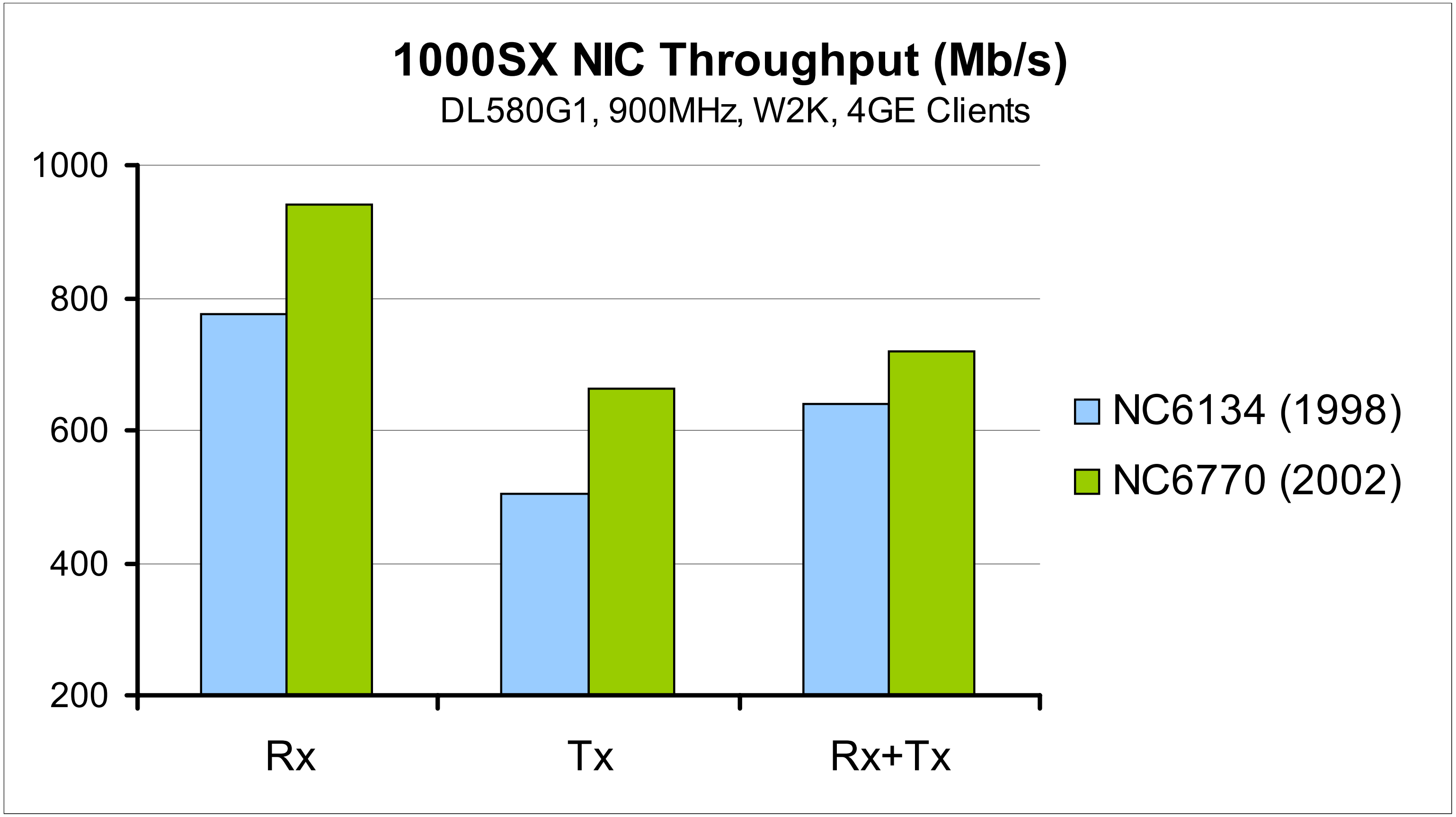
1. Upgrade to faster CPU speeds to help reduce TCP/IP processing time.
2. Update Gigabit NICs: larger buffers, task offloading and PCI-X Bus speed and protocol efficiencies.
3. Use servers with faster PCI or PCI-X support for multiple Gigabit NICs or storage devices.

Faster CPU

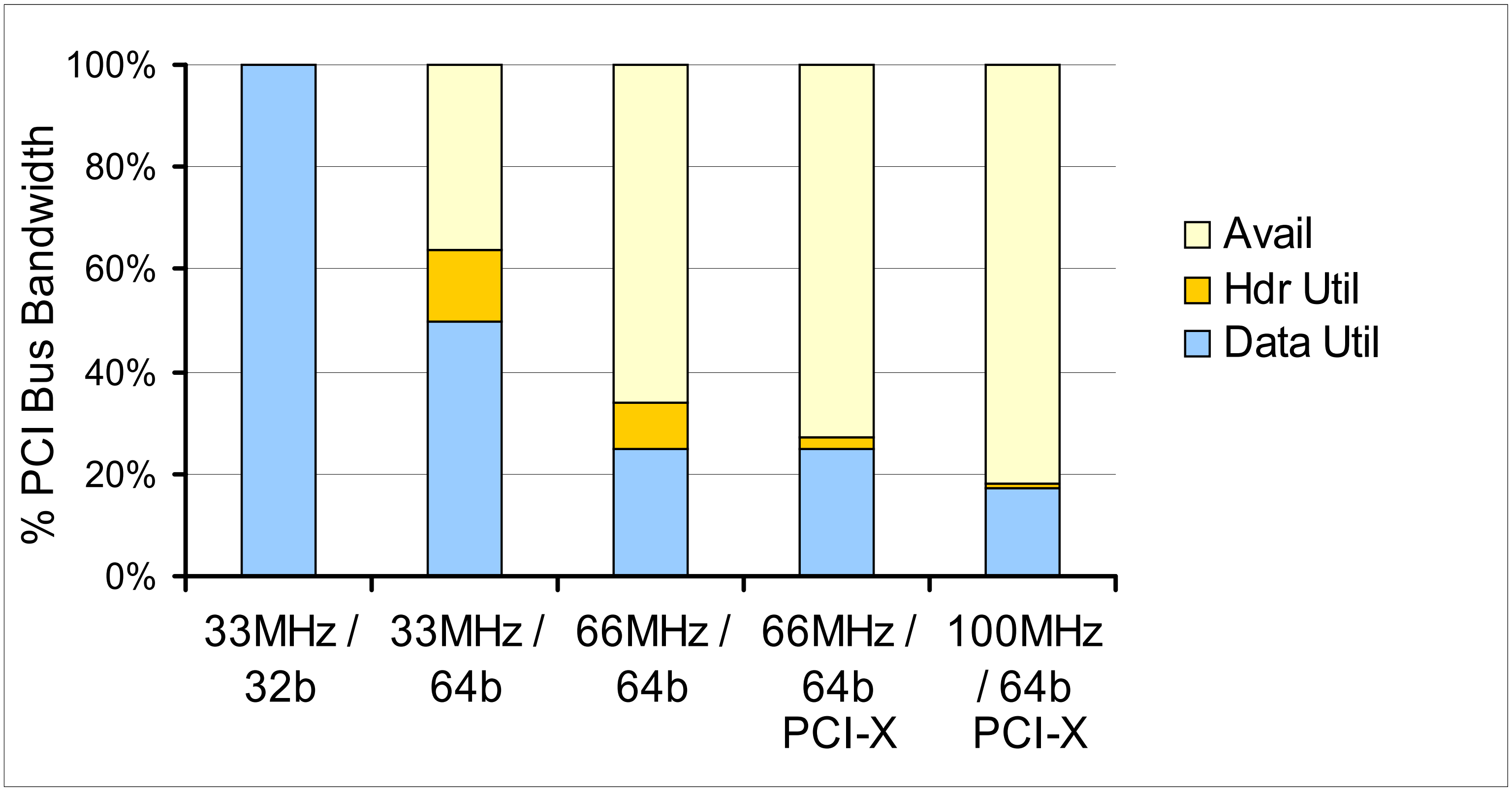


NC7770 Throughput (Mb/s)
1 CPU, Win2K Srv-SP3, 4 GE Clients
1000
800
600
400
200
Rx
Tx
Rx+Tx
DL580 0.7GHz
ML370 1.4GHz

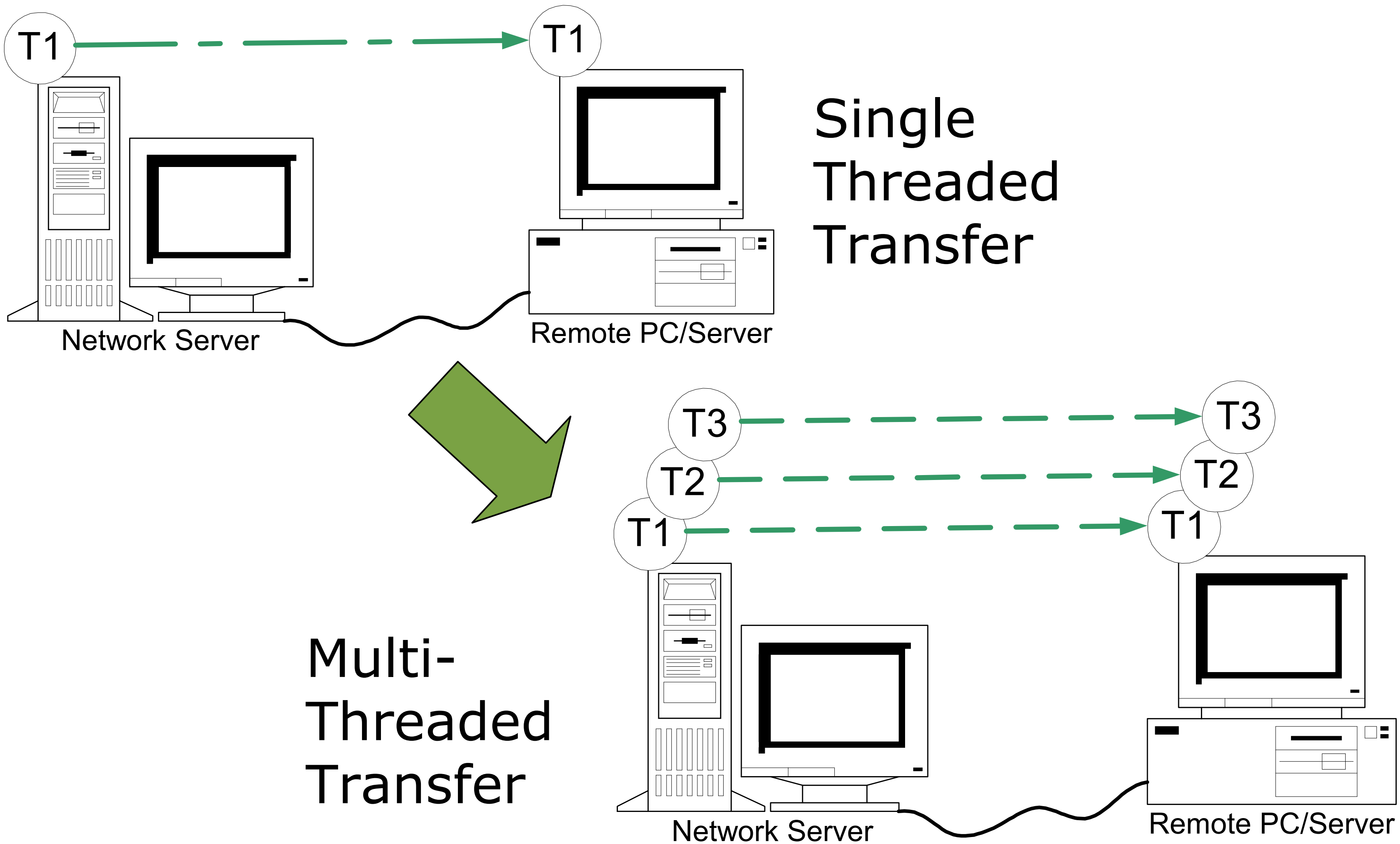
Newer NIC Technology



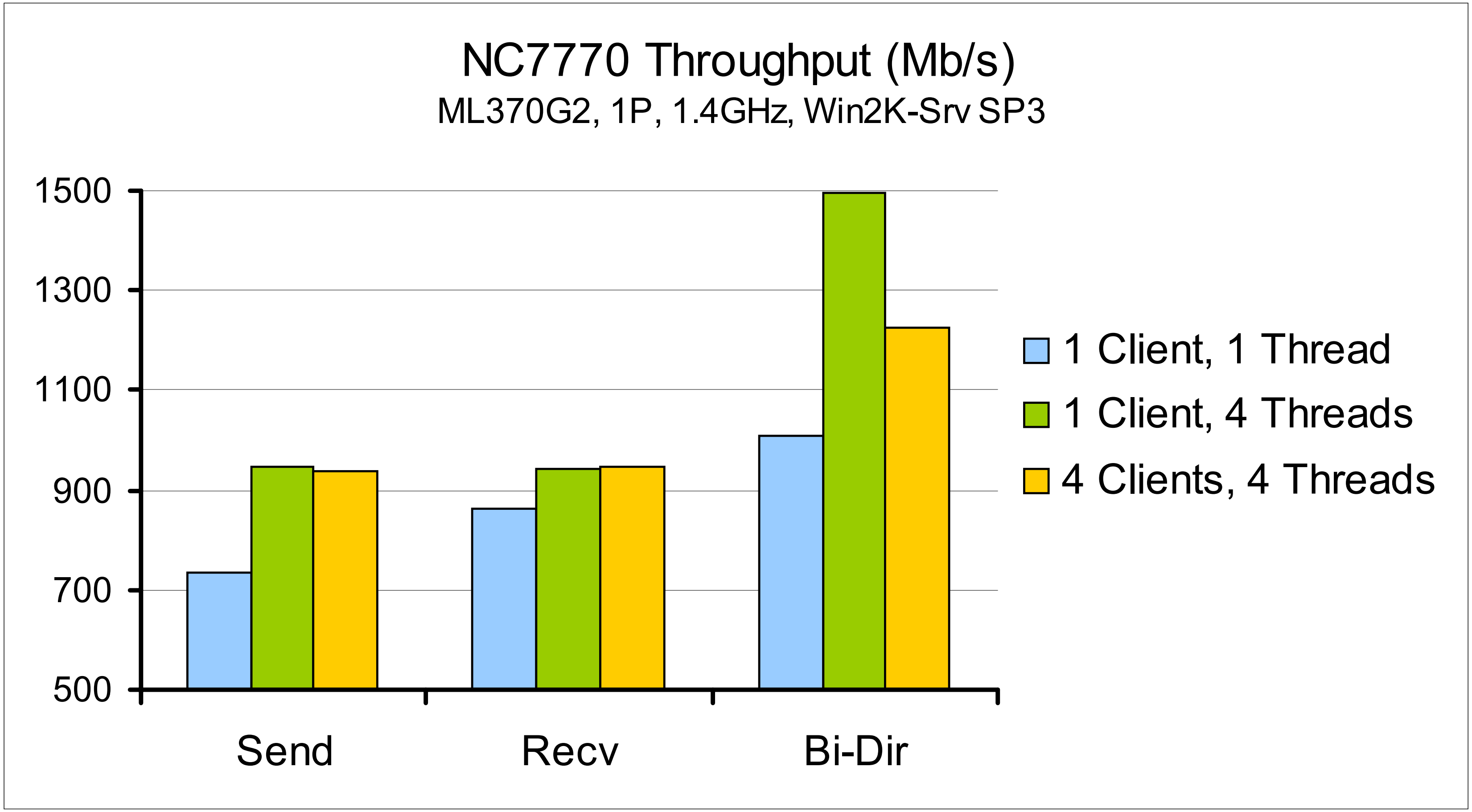
Gb NIC PCI Utilization



Multiple, Simultaneous Transfer Threads



Multi-Simultaneous Sessions



Jumbo Frame

- standard Ethernet

64 - 1518 Bytes

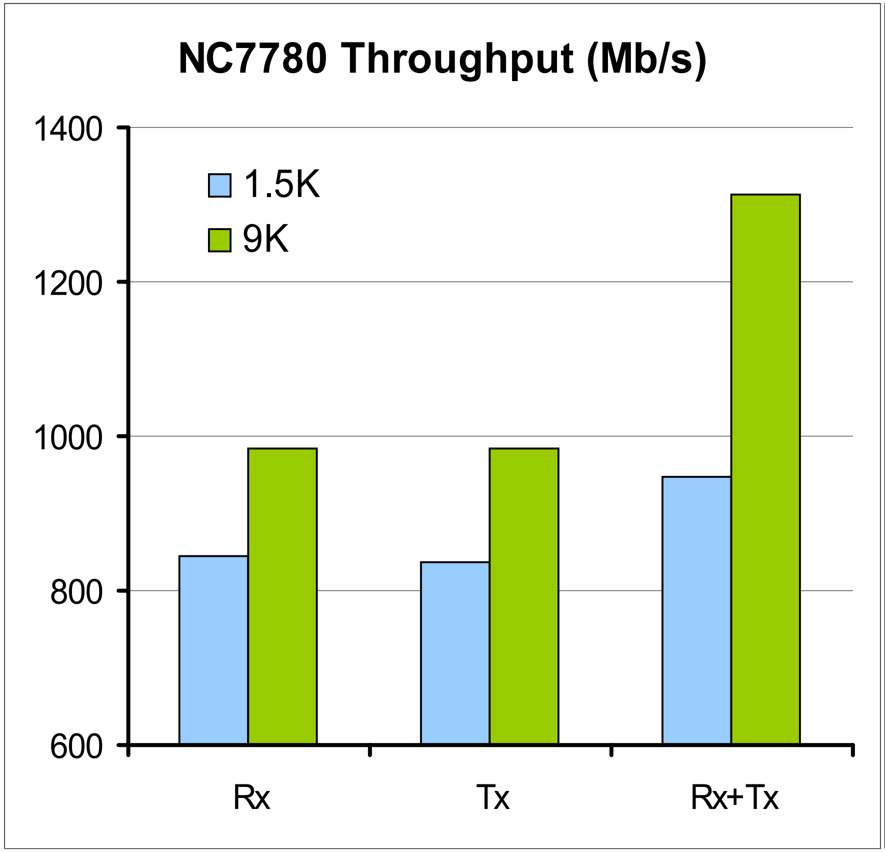
8 Bytes	6 Bytes	6 Bytes	2 Bytes	46 - 1500 Bytes	4 Bytes
Preamble /SFD	Destination Address	Source Address	Type/ Length Field	Data and Pad	Frame Check Sequence

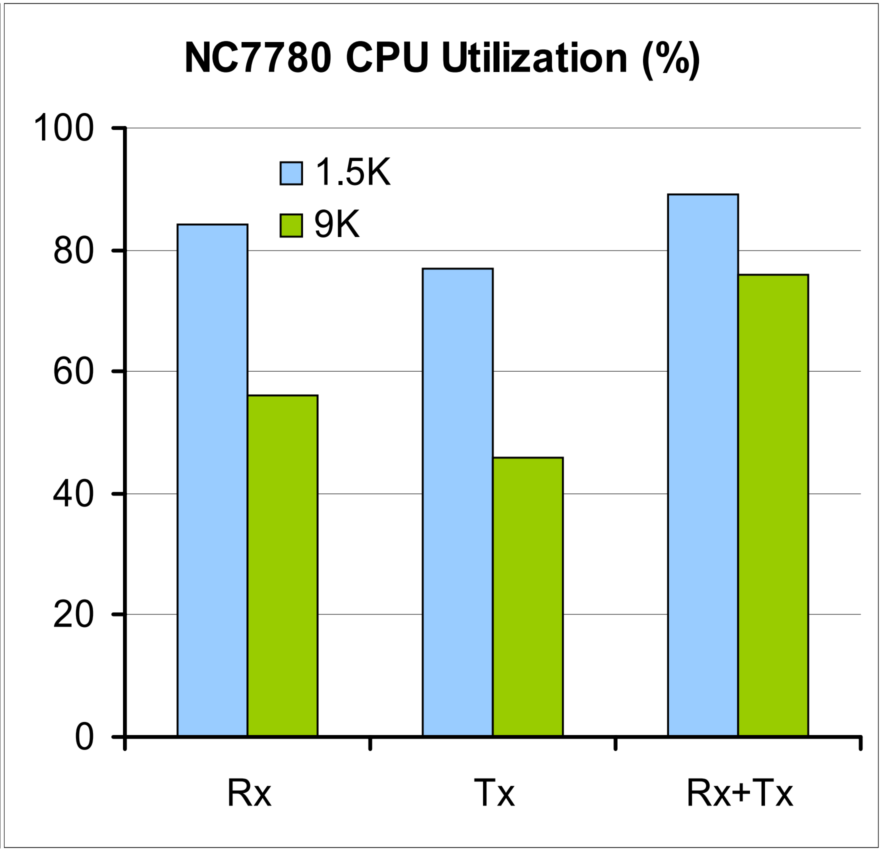
- jumbo frames

64 - 9018 Bytes

8 Bytes	6 Bytes	6 Bytes	2 Bytes	46 - 9000 Bytes	4 Bytes
Preamble /SFD	Destination Address	Source Address	Type/ Length Field	Data and Pad	Frame Check Sequence

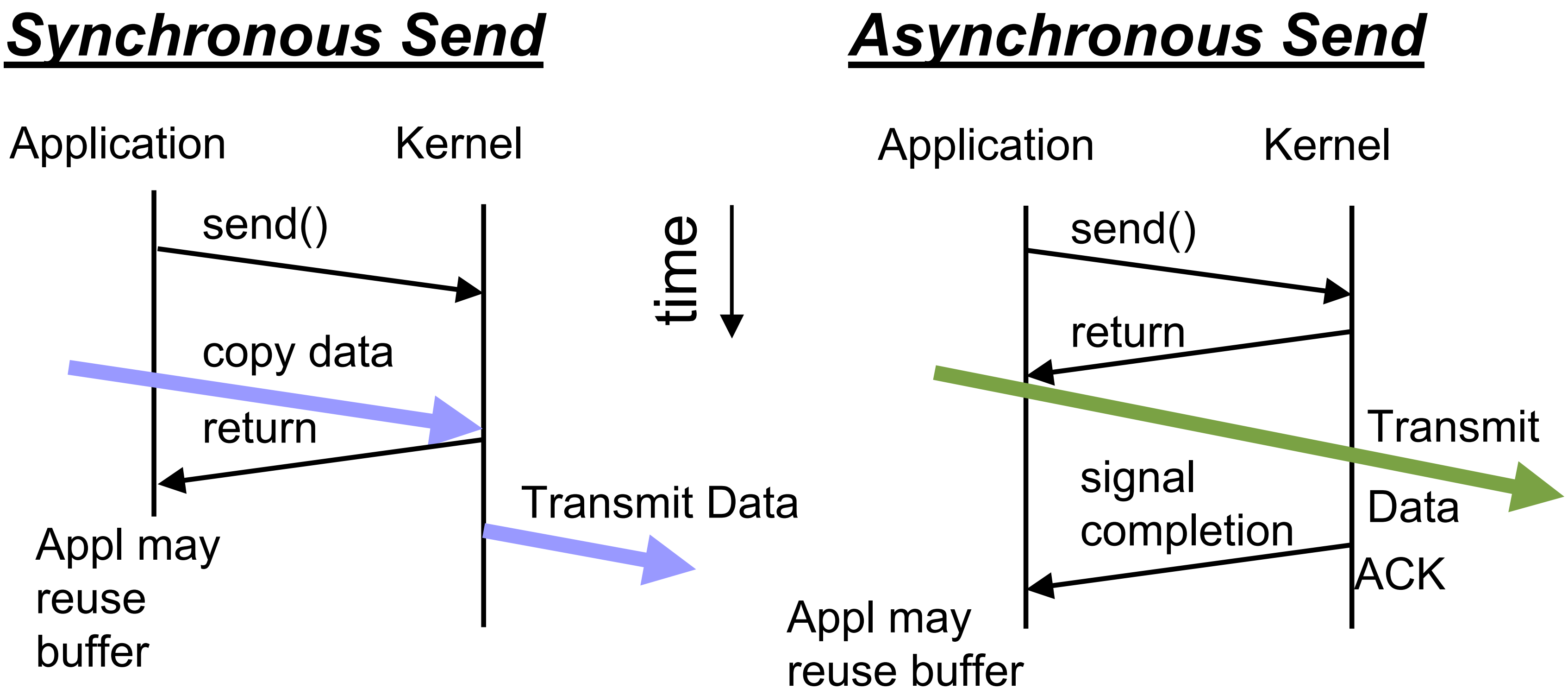
Jumbo Frames (cont.)





DL380G3, 1P-2.4GHz, 1 Client, 4 Threads

Asynchronous I/O



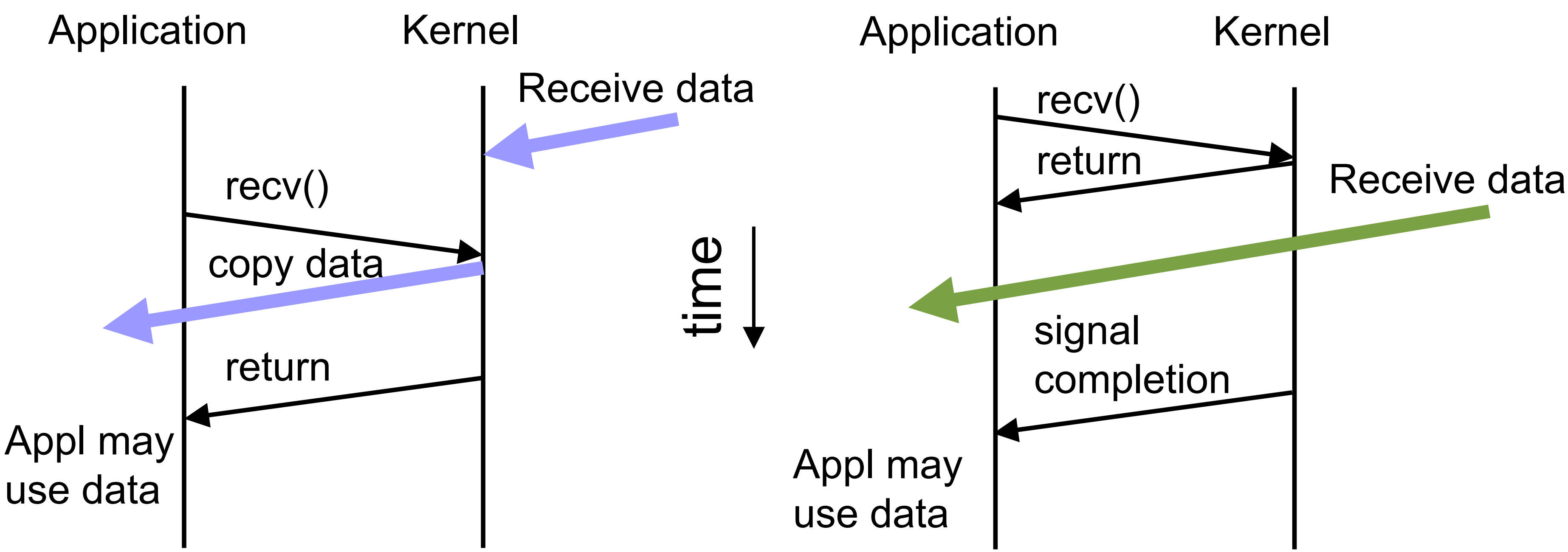
- Simpler Programming
- Appl to Kernel Copy
- Wastes CPU cycles

- Wait for Signal Cmplt before re-use
- Data DMA'd from Appl Space
- More CPU cycles for Appl

Asynchronous I/O (cont.)

Synchronous Receive

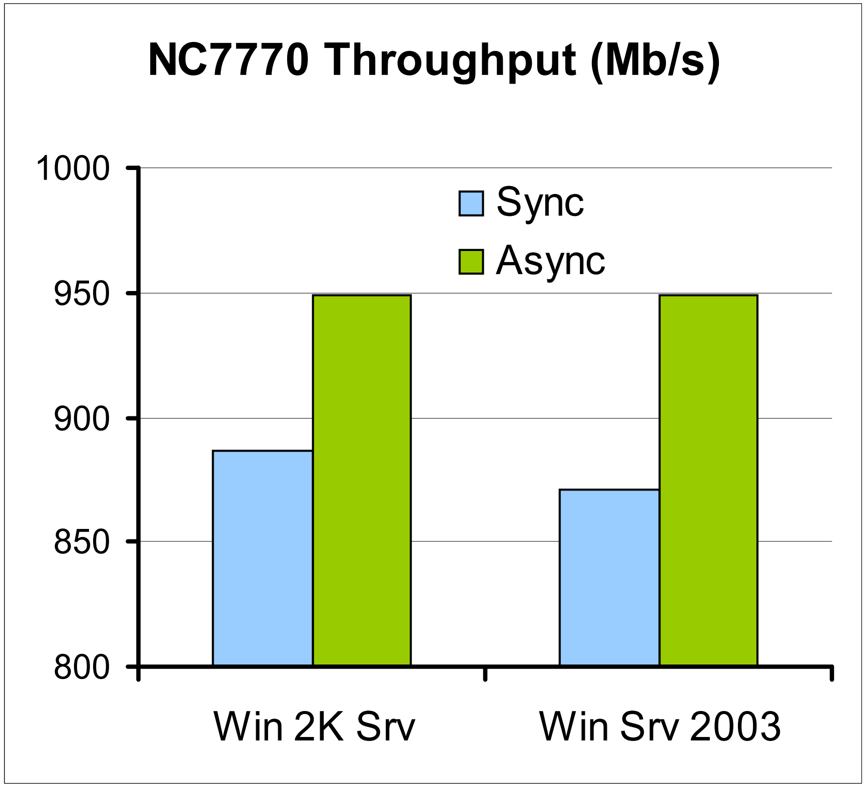
Asynchronous Receive

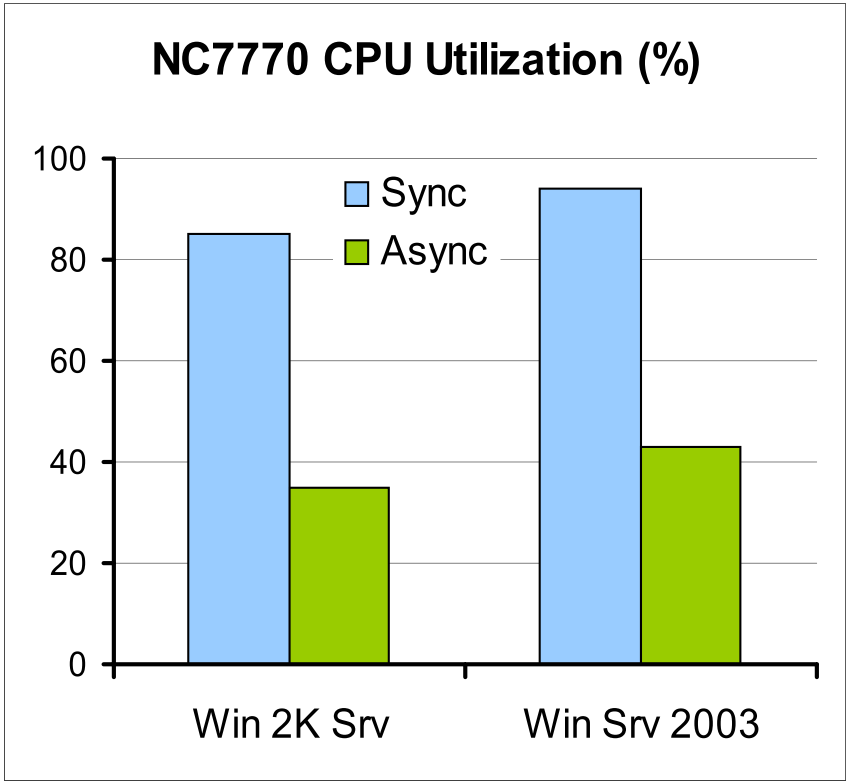


- Simple programming model
- Appl to Kernel Copy
- Wastes CPU cycles

- Appl must avoid race conditions
- Data DMA'd to Appl Space
- More CPU cycles for Appl

Asynchronous I/O (cont.)





ML370G2, 2P-1.4GHz, Send to 8 GE Clients

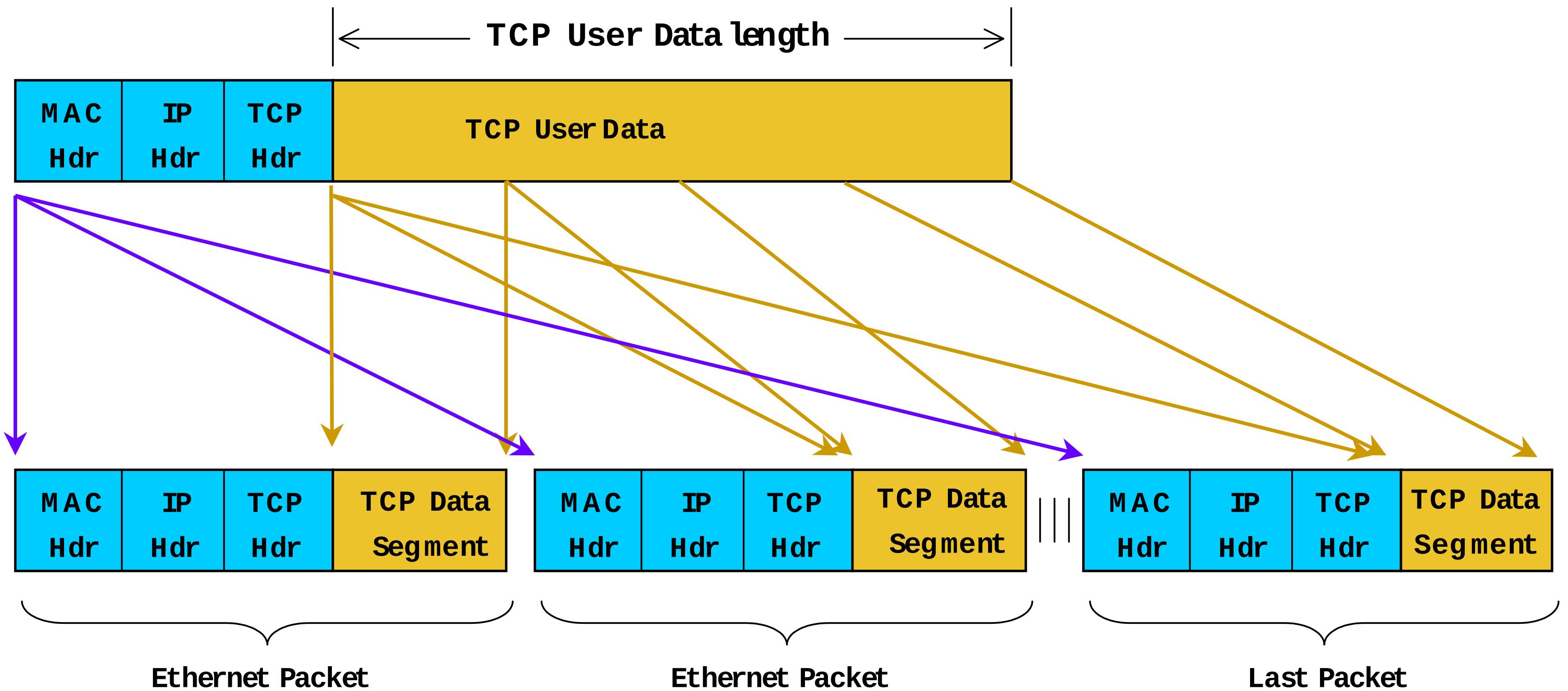
Asynch I/O Under Linux

```
sendfile(out_fd, in_fd, offset, count)
```

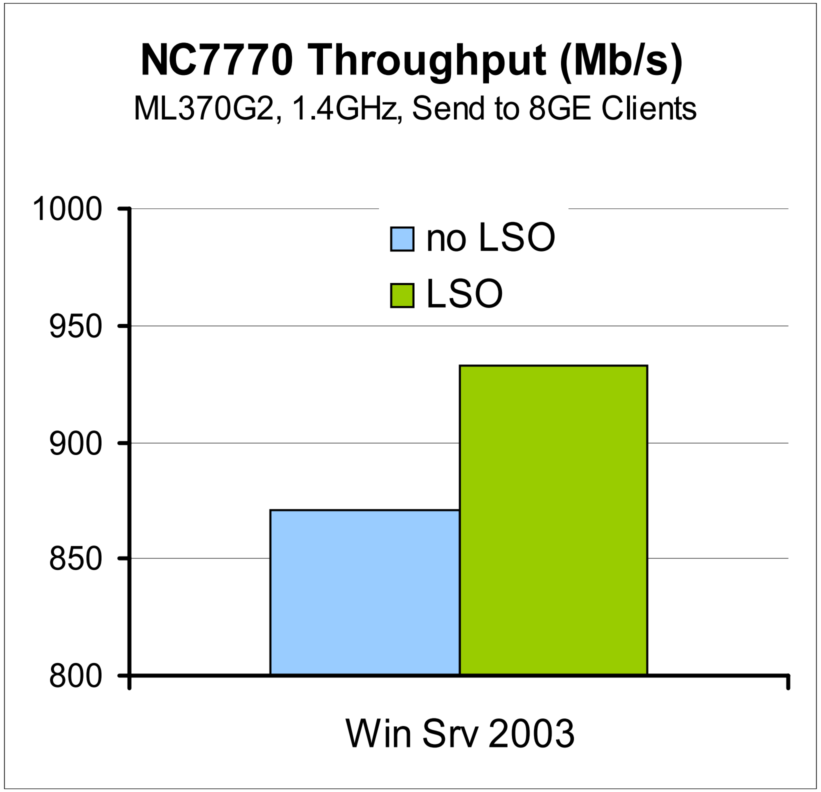
- copies contents of *in_fd* starting at *offset* for *count* bytes to *out_fd*
- uses kernel buffers directly; avoids two data copies between kernel and user space
- avoids system call overhead; one call does the work of multiple reads and writes

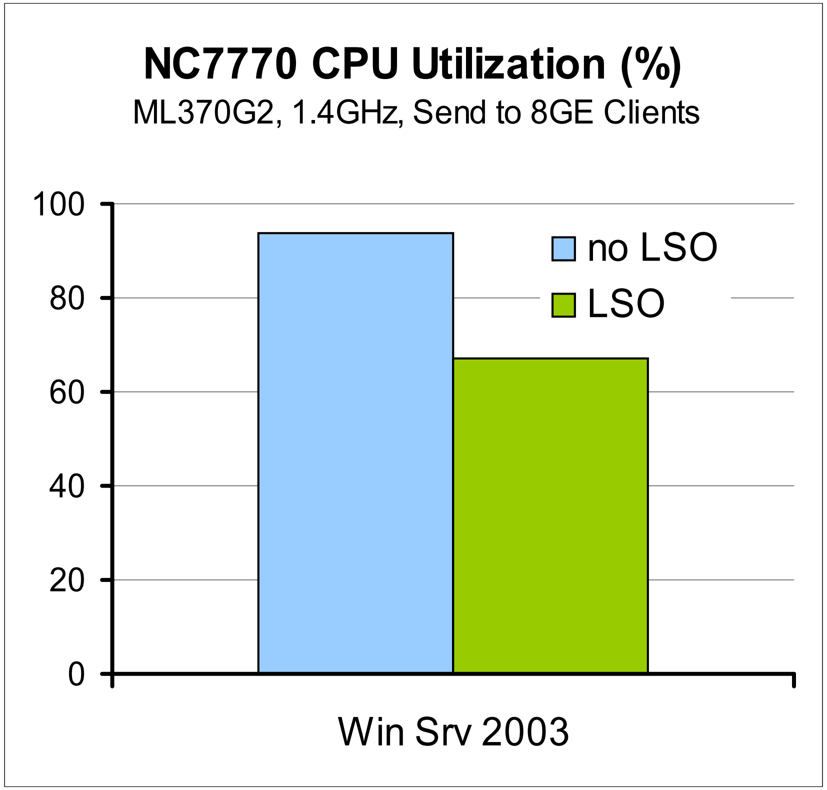
Large Send Offload

Also known as TCP Segmentation Offload



Large Send Offload (cont.)



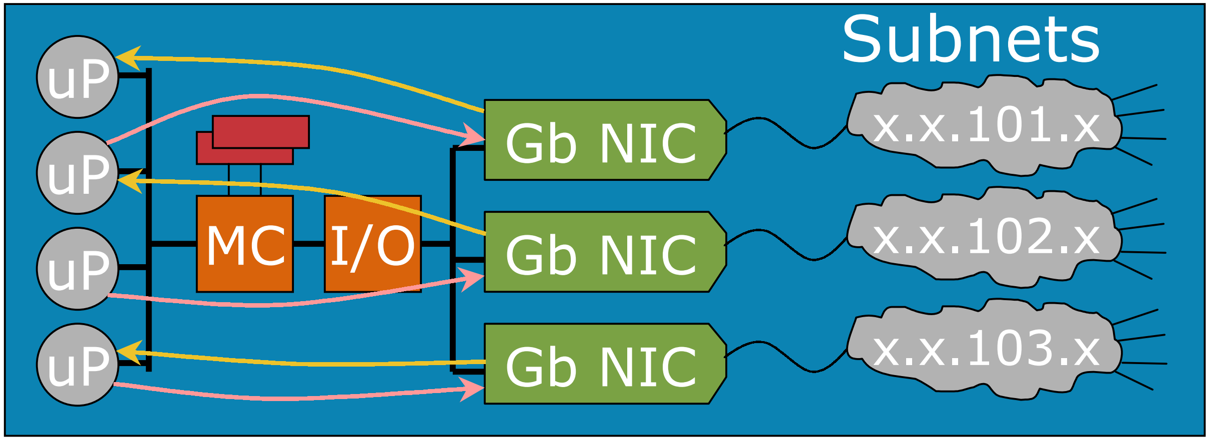


- Standard and enabled by default on Win Srv 2003
- Optionally supported on Win2K Srv in SP4

Multiple NIC Subnets

Separate Subnet per NIC

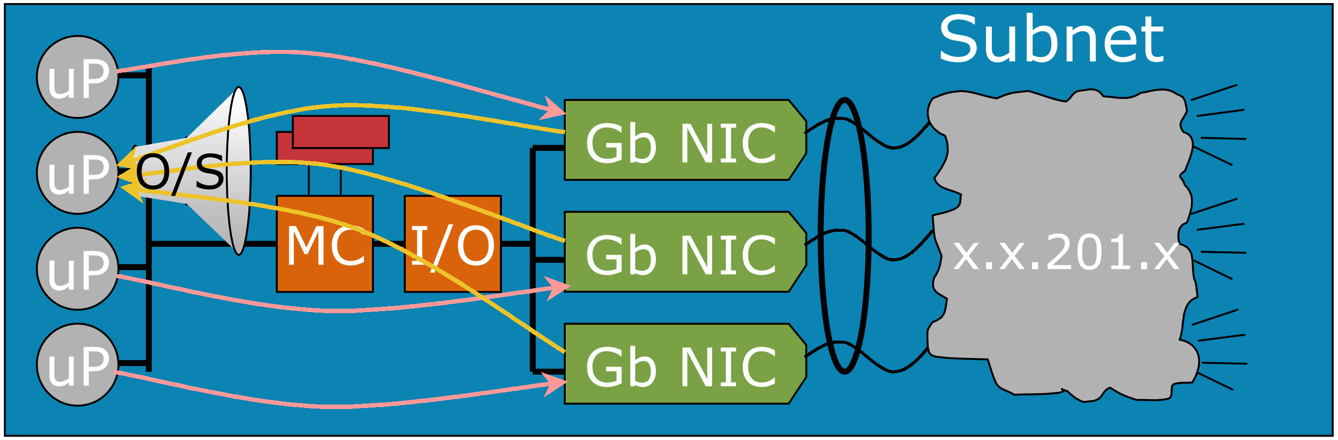
- Achieves Optimal Server Network Performance
- Harder to Admin Multi-Subnets, DHCP, and Masks
- Need One 1.2+GHz CPU per Gb NIC for Wire Speed
- Additional CPUs needed for Appl or File System
- Will work Point to Point, but needs 1+ thread per NIC



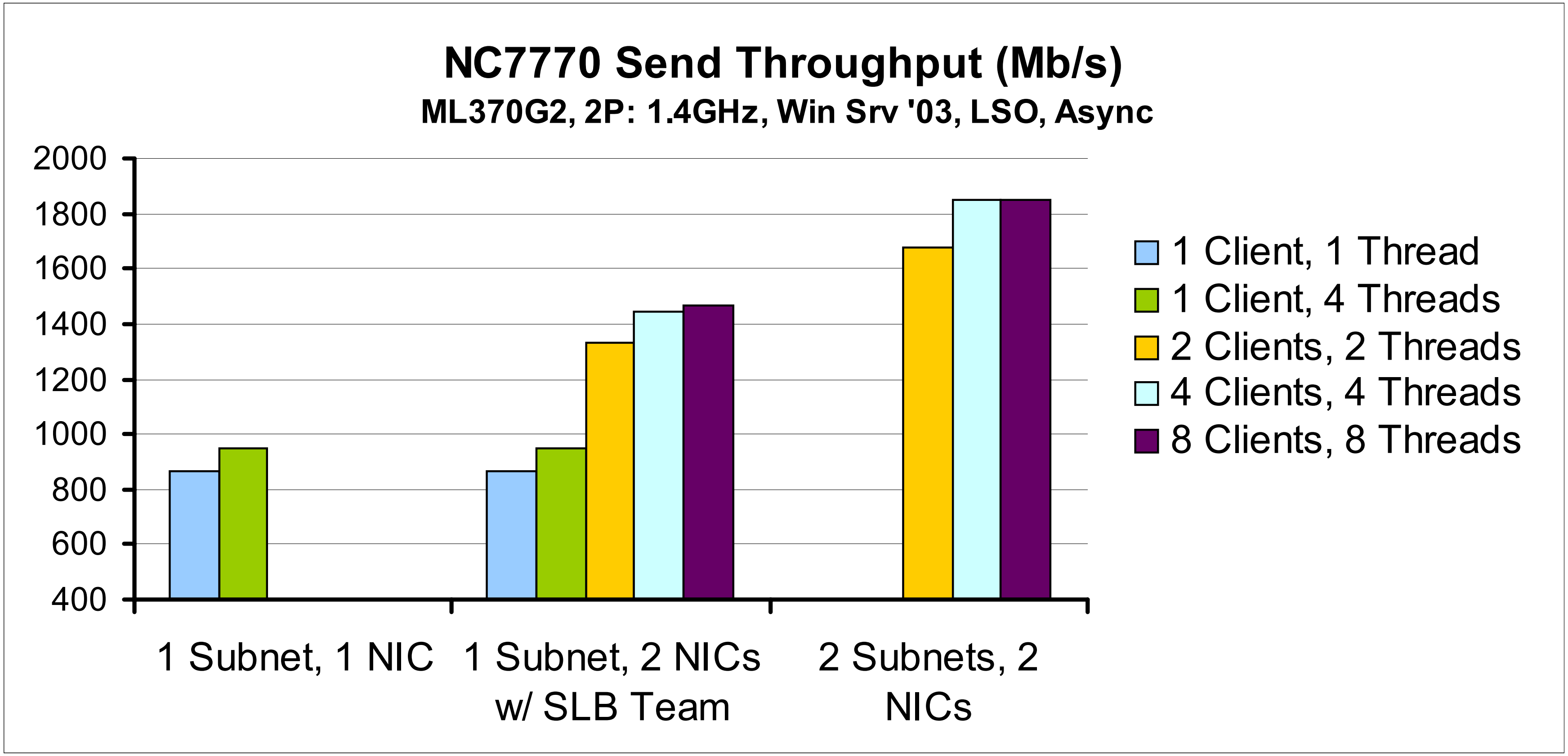
Multiple NICs Teamed

Link Aggregation (Alias : Team or Trunk)

- Increases Bandwidth to Single Subnet
- Easier to Manage Single IP and Subnet
- Need One 1.2+GHz CPU per Gb NIC for Wire Speed
- Additional CPUs needed for Application or File System
- Point to Point transfer gets no benefit from Team /Trunk



Multi-NICs Team vs. Multi-Subnets



Multi-NIC Team (cont..)

Note: *802.1ad Link Aggregations port selection algorithm requires multiple source or destination addresses to balance load.*

- Our Switch assisted Load-Balancing (SLB) and Transmit Load Balancing (TLB) supports 802.1ad Link Aggregation based on MAC or IP address.
- For optimal performance, NIC team must have multiple clients with evenly distributed MAC or IP address.
- Single Server to Single Client trunk will not benefit from today's Link Aggregation Algorithms. *(eg. Network backup of Server to Single IP Dest [Tape/Disk])*

Tips for Future

- Today:
  - Upgrade to faster hardware: CPU, NIC, & PCI Bus
  - Use multiple, simultaneous transfers
  - Enable Jumbo Frames between endpoints
  - Buy/write applications with Asynchronous I/O support
  - Upgrade NIC and O/S with Large Send Offload support
  - Multiple NICs: multi-subnets or Link-Aggregation (team/trunk)

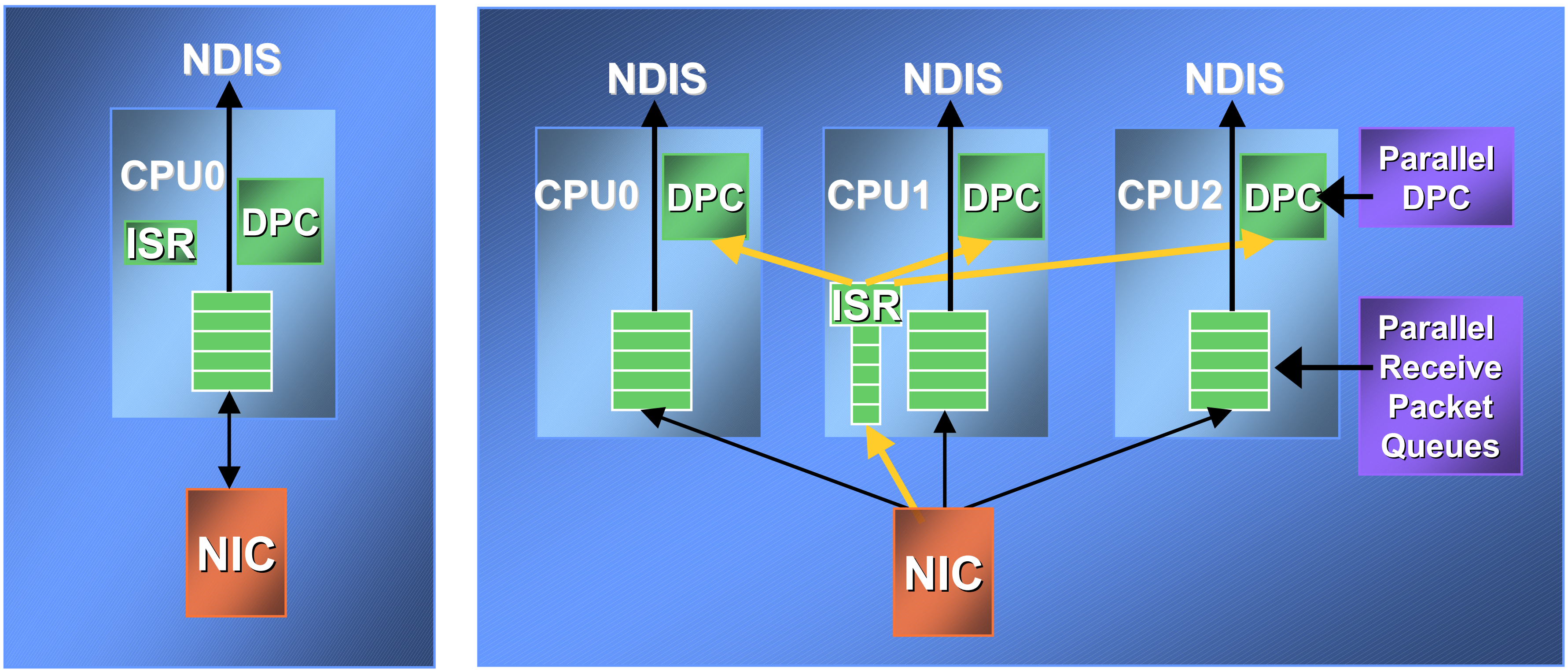
- Future:
  - Receive Side Scaling (RSS)
  - TCP Offload Engines (TOE)
  - Remote Direct Memory Access Protocol (RDMAP)

Future Performance Enhancers

Receive Side Scaling (RSS)

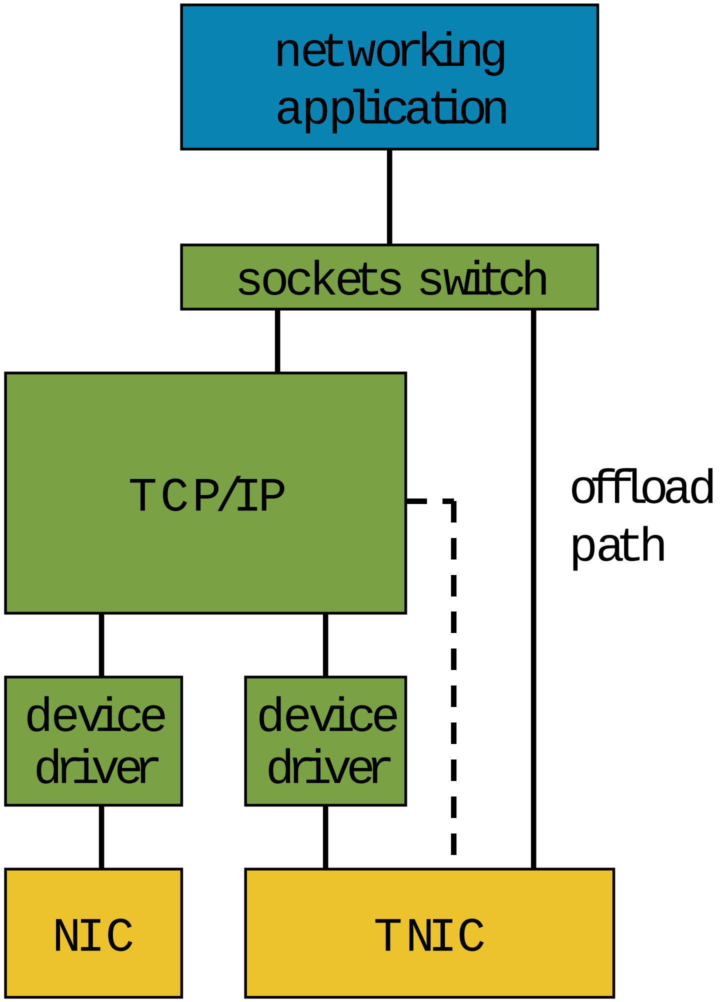
- Existing network stack limits receive processing to 1 CPU, Restricts network receive performance to that of single CPU can manage (per NIC)
- ACK (TX) processing tied to RX processor for CPU cache affinity
- Needed method to load balance RX processing across available CPUs while maintaining process affinity.
- Microsoft has developed RSS to parallelize Receive processing.
- Will require new hardware hooks in NIC to assist in load distribution.

Receive Side Scaling



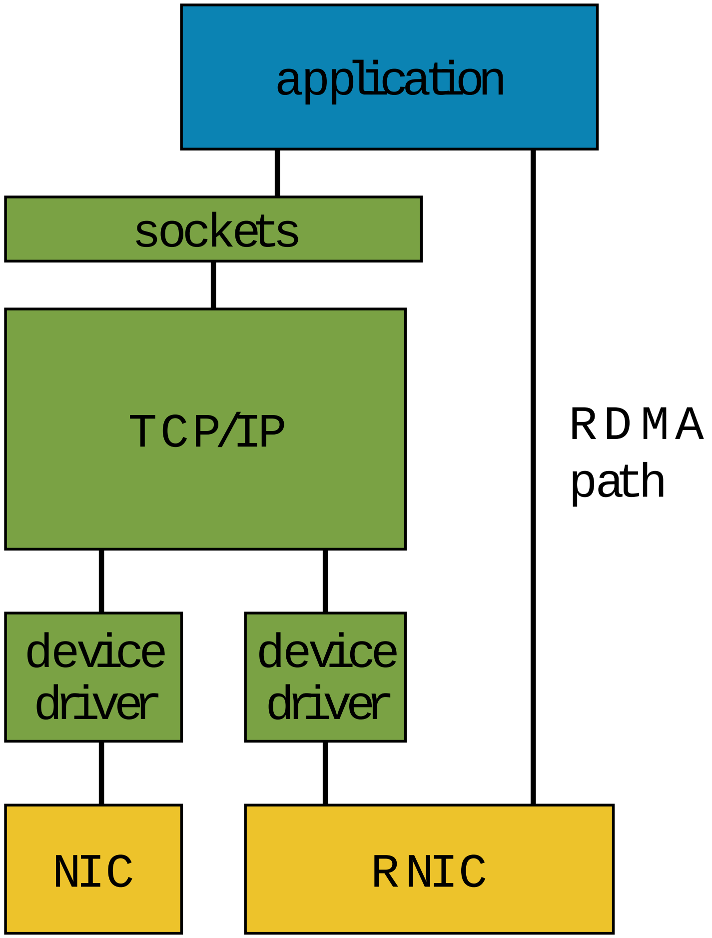
TCP/IP Offload Engines (TOE)

- TCP/IP processing moved from kernel to TOE NIC
- TCP connections may be established in TNIC or in kernel
- Reduces CPU utilization for segmentation and reassembly
- Reduces interrupts and context switches
- Allows for zero-copy receives to kernel buffers
- Works best with async IO

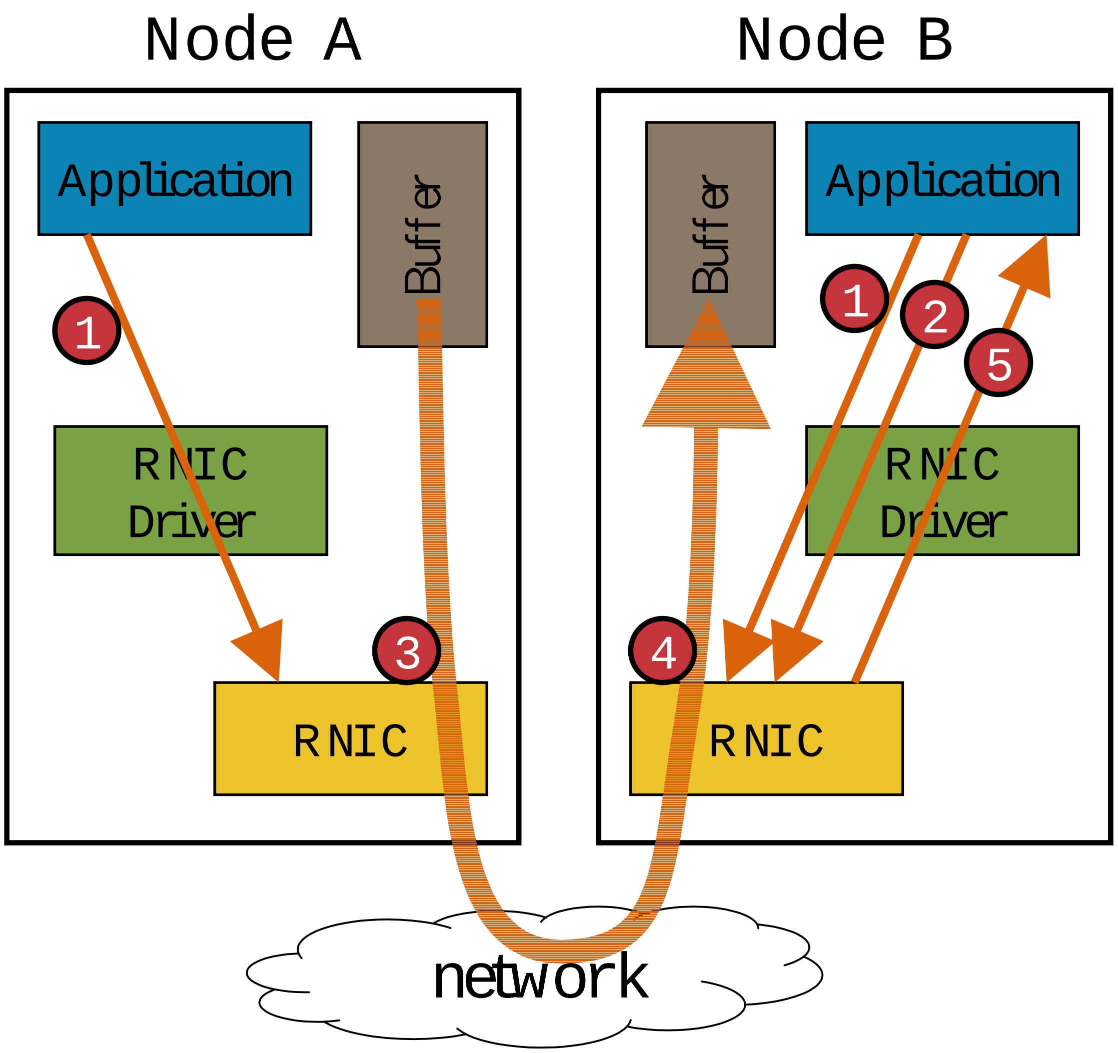


Remote Direct Memory Access (RDMA)

- Provides direct communication between user-space buffers in separate servers.
- Bypasses the kernel
  - avoids protocol processing
  - avoids context switches
  - avoids interrupt processing
  - yet, preserves kernel protections
- Improves both
  - throughput scaling
  - message latency
- Provides the performance needed by networking, IPC, and storage

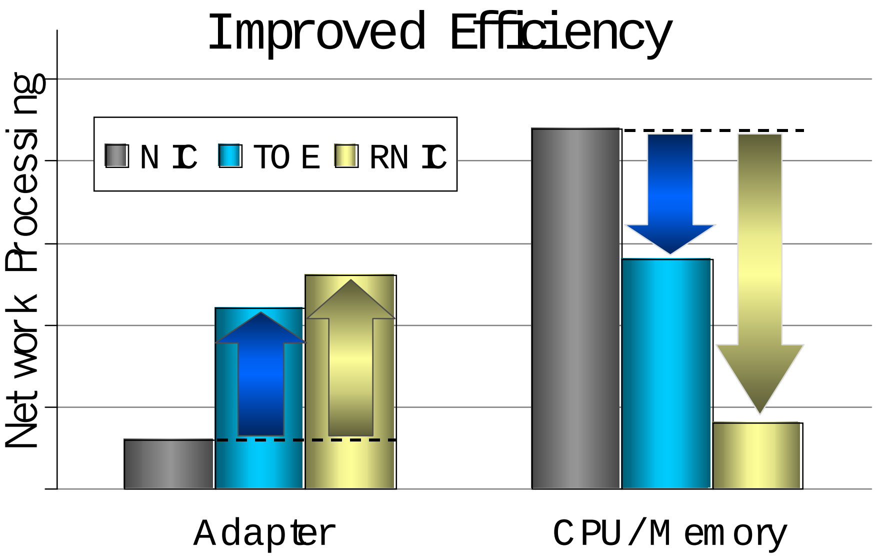


RDMA read operation



1. Both nodes have suitable memory regions registered
2. Node B initiates RDMA Read
3. RNIC in Node A sends data
4. RNIC in Node B places data in buffer as directed
5. RNIC in Node B complete read (w/o kernel intervention)

RDMA NICs Provide Better Performance



Networking Benchmarks	BW Mbps	CPU Util %	Perf. Index
1Gb/s Enet	1000	60%	17
TOE	1000	40%	25
1Gb/s RDMA	1250	15%	74
10Gb/s RDMA	8500	15%	567

4X
30X

Note: Based on internal HP projections

RDMA enabled NICs (RNICs)

- More efficient network communications
- TOE moves TCP/IP work from the CPU
- RDMA reduces the communication work

CPU/memory freed up for applications

- Zero-copy RDMA protocol conserves valuable memory bandwidth
- Much lower CPU utilization
- Per message communication overhead

Improved application performance

- Opportunity for increased application throughput or server consolidation
- Improved scalability for streaming applications or large data files

Boosting ProLiant Networking Performance

- Today:
  - Upgrade to faster hardware: CPU, NIC, & PCI Bus
  - Use multiple, simultaneous transfers
  - Enable Jumbo Packets between endpoints
  - Buy/write applications with Asynchronous I/O support
  - Upgrade NIC and O/S with Large Send Offload support
  - Multiple NICs: multi-subnets or Link-Aggregation (Team)

- Future:
  - Receive Side Scaling (RSS)
  - TCP Offload Engines (TOE)
  - Remote Direct Memory Access Procotol (RDMAP)

Questions?

HP WORLD 2003
Solutions and Technology Conference & Expo

Interex, Encompass and HP bring you a powerful new HP World.

interex
shared knowledge • shared power

encompass
AN HP USER GROUP

hp
invent