# ProLiant Networking Performance and Tuning for Real-World Scenarios

**Alan Banks**
ProLiant Networking
Industry Standard Servers

# System and Network Performance Considerations

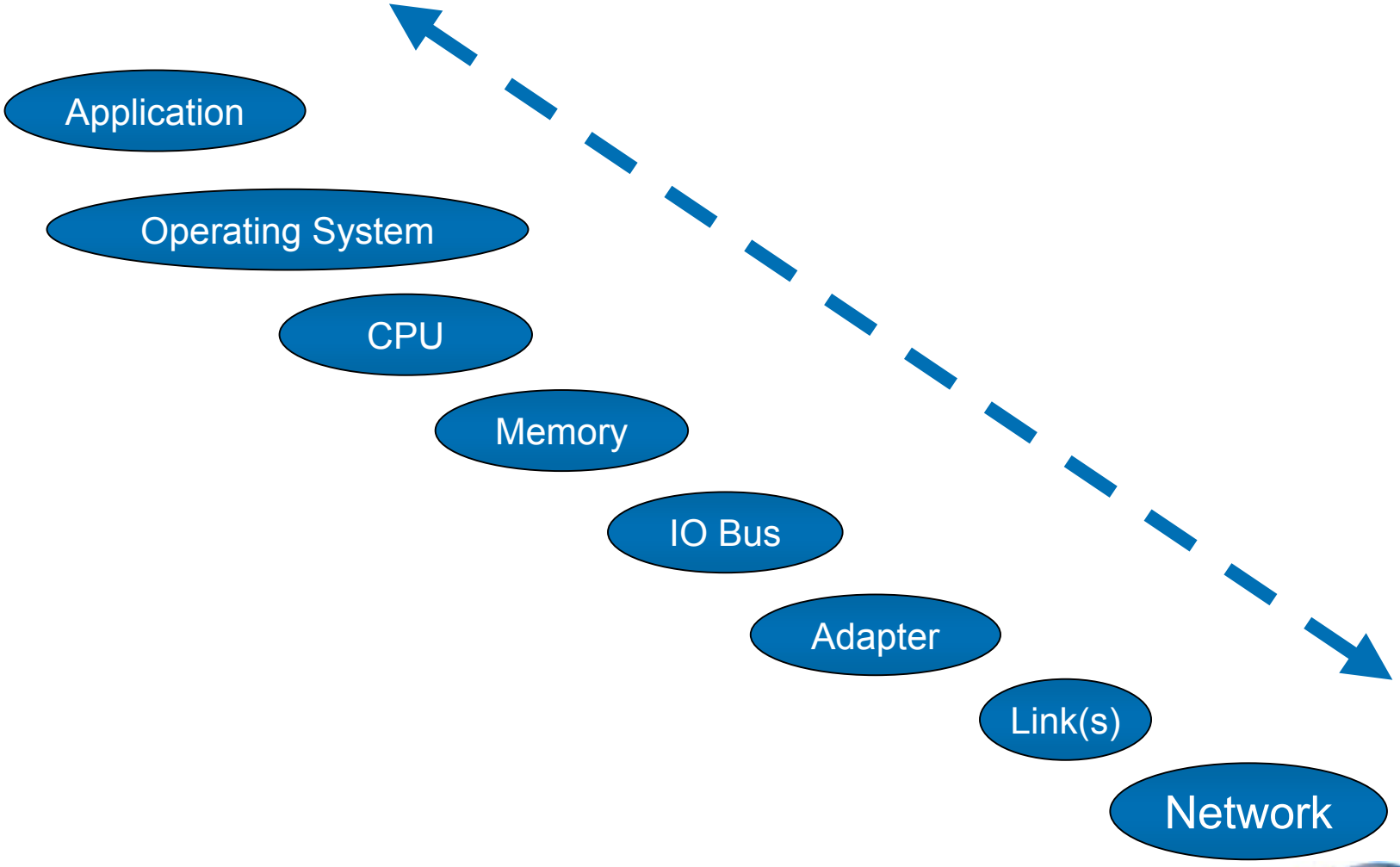
# Before We Begin…..

- 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:

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

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

# Network Performance
# Finding the Weakest Link



# Network Performance Which Definition?

- Response Time
  - "Time between the start and completion of an event"
  - End users are aware
  - Server to server applications are aware (latency)
- Examples
  - Web page load time
  - Database record locks
  - Memory access

- Throughput
  - "Total amount of work accomplished in a given amount time"
  - End users are aware
  - Sys Admins are aware
- Examples
  - Backups
  - TPC Benchmarks
  - File access

# Network Latency - (Response Time) Application to Application

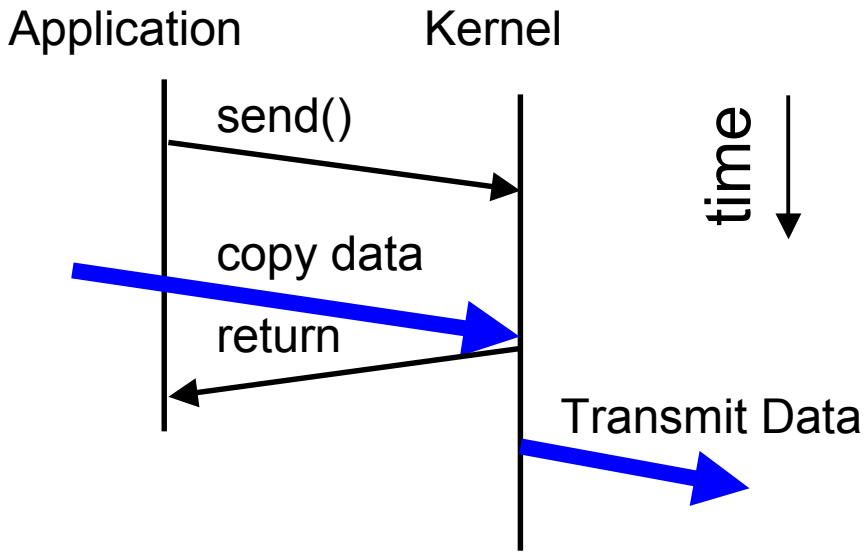

- One Way Trip Latency

Sender CPU Overhead Time
Network Adapter Overhead Time
Media Bandwidth Time
Network Processor Overhead Time
(Iterate Media and Network for multiple hops)
Media Bandwidth Time
Receiver Adapter Overhead Time
Receiver CPU Overhead Time

- Round Trip = Multiply by two ….

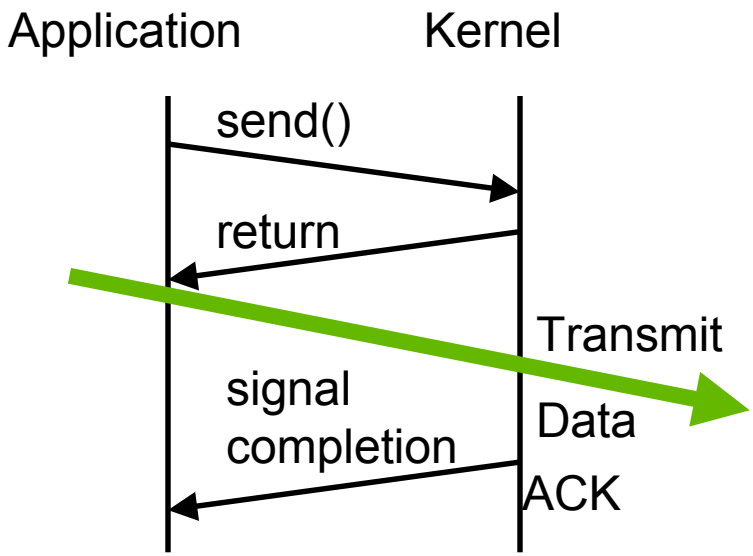
# Improving Network Performance - Asynchronous I/O

***Synchronous Send***



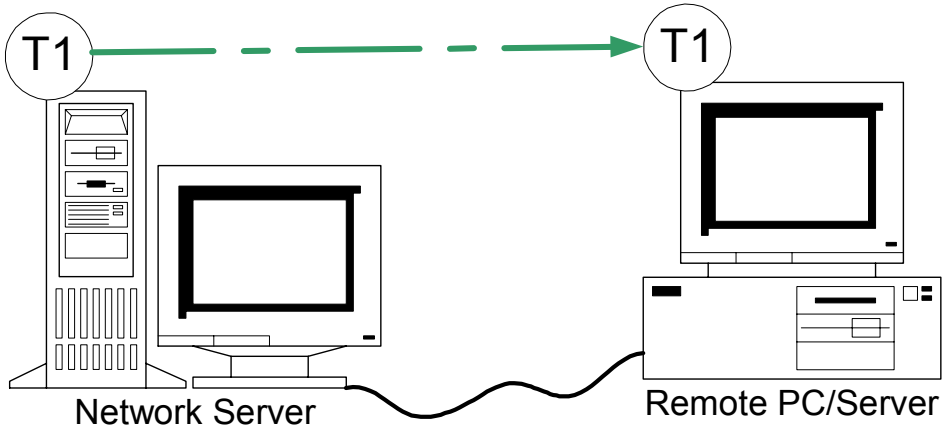

- Simpler Programming
- Appl to Kernel Copy
- Takes more CPU cycles

***Asynchronous Send***

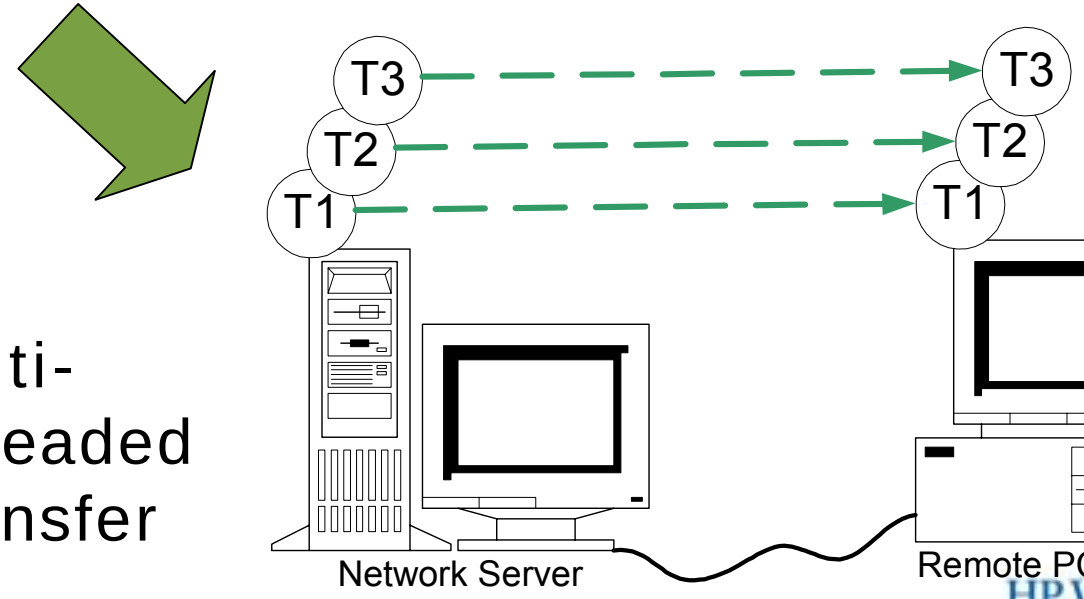



- Wait for signal completion before re-use
- Data DMA'd from Appl Space

# Improving Network Performance - Multiple, Simultaneous Transfer Threads




Single Threaded Transfer




Multi-Threaded Transfer

# Improving Network Performance - Registry Settings

- Increase TCP Window Size for large transfers. Set hidden registry parameter *TcpWindowSize to 32K - 63K*

  *HKEY_LOCAL_MACHINE/System/CurrentControlSet/Services/Tcpip/Parameters /*

  *Documentation on TcpWindowSize can be found here:*
  *http://www.microsoft.com/resources/documentation/WindowsServ/2003/all/deployguide/en-us/Default.asp?url=/resources/documentation/WindowsServ/2003/all/deployguide/en-us/242669.asp*

- NIC drivers set Interrupt Coalescence to reduce the number of interrupts per packet, at the cost of latency. To reduce latency, decrease coalescence buffer:
  - Intel based NICs (NC61xx, NC71xx)
    - TX: *NumCoalesceBuffer*
  - Broadcom based NICs (NC67xx, NC77xx)
    - TX: *TxMaxCoalescedFrames*
    - TX: *TxCoalescingTicks*
    - RX: *RxMaxCoalescedFrames*
    - RX: *RxCoalescingTicks*

# Improving Performance - Faster Hardware

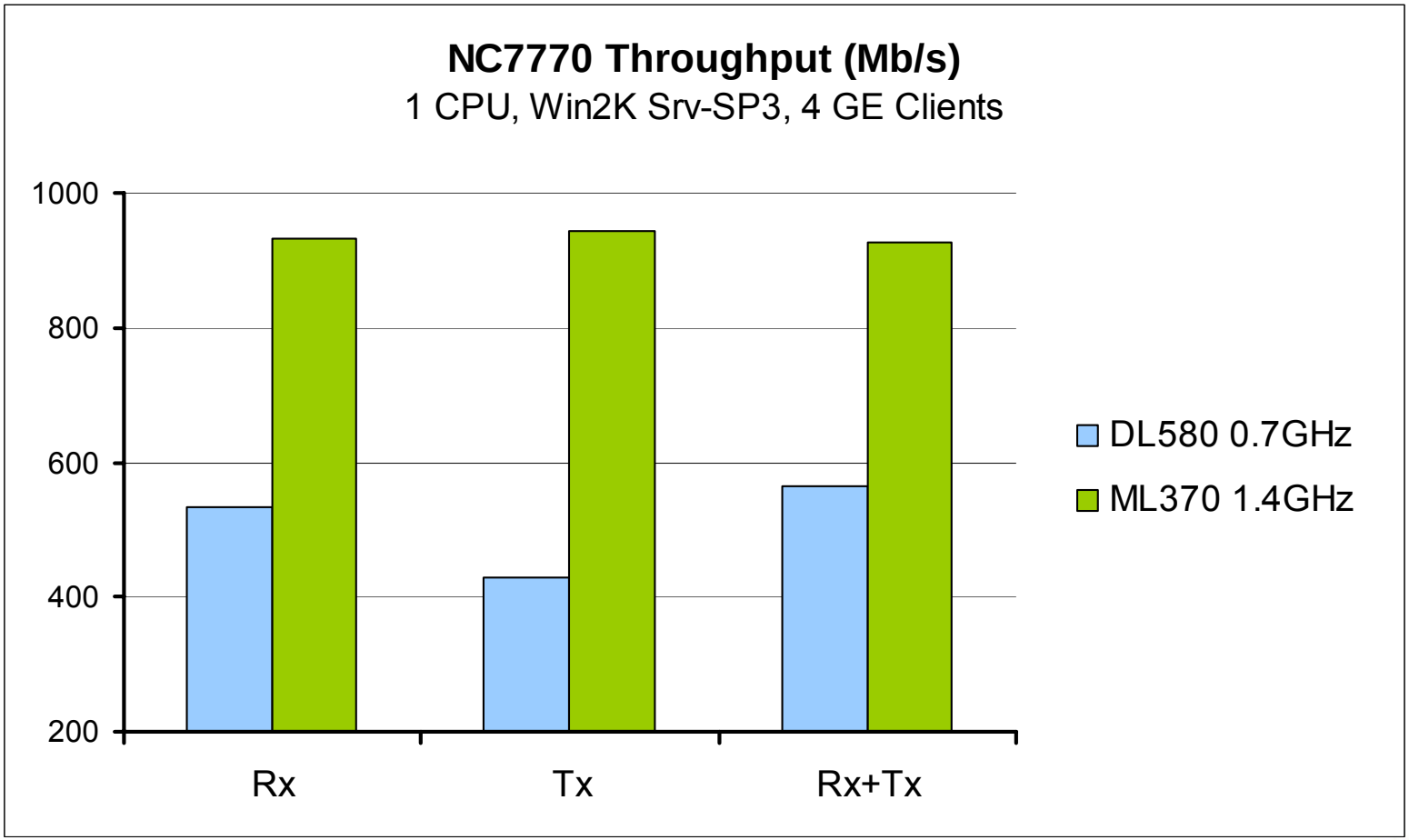
*Faster Hardware Increases Server Network Performance*

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.
3. Use servers with faster PCI, PCI-X, or PCI-Express support for multiple Gigabit NICs or storage devices.

# Improving Performance - Processor Recommendations

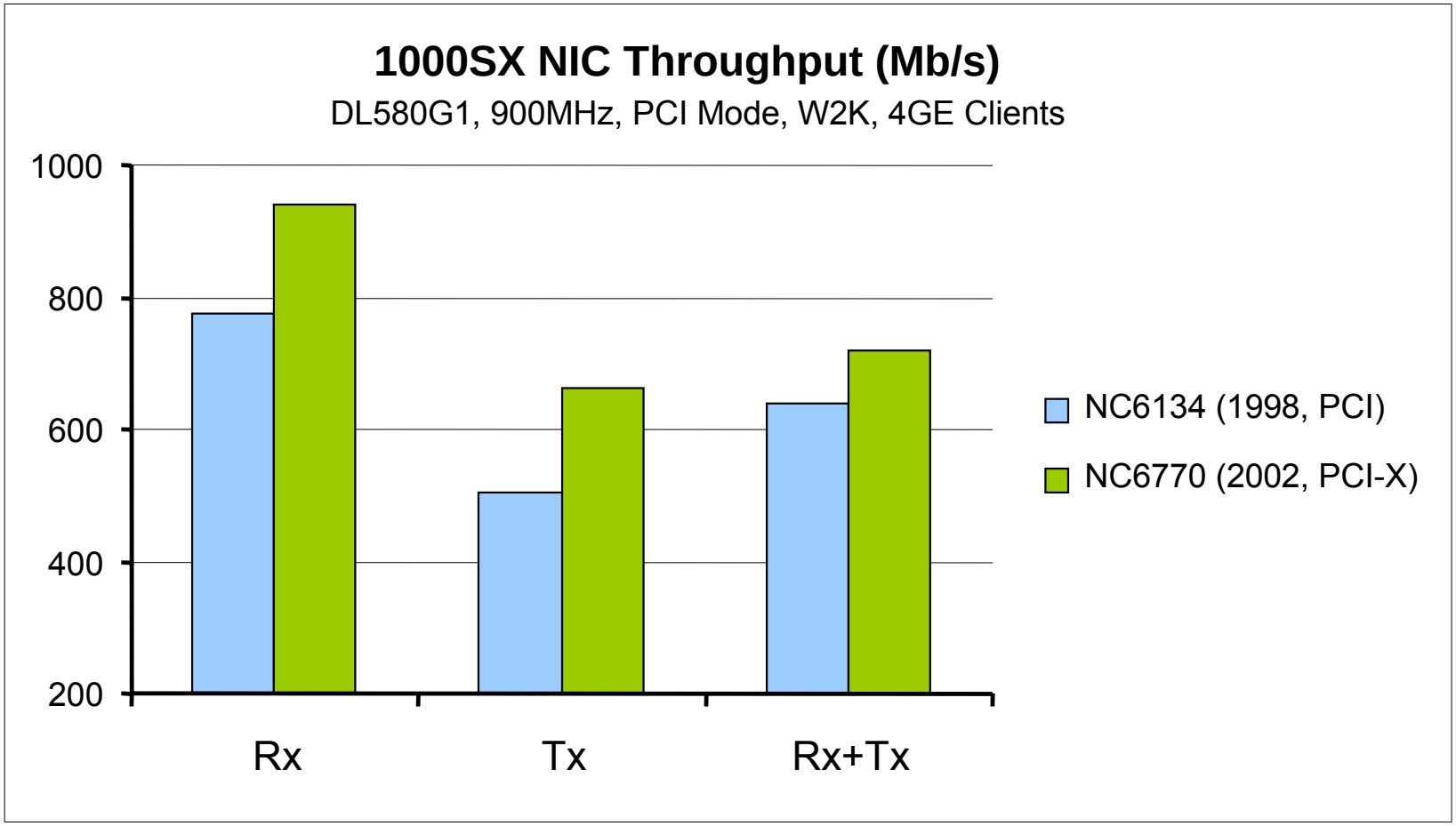
- Minimum Processor Recommendations to Achieve Maximum Throughput with a 1Gbps NIC
  - At least one 1.4GHz (or 2x 700MHz for Xmit) Physical CPU Per Gb NIC and at least one more for the Application or File System.
    - Note: Hyper Threaded CPU = 1.15x Performance of non-HT CPU at same Freq.

# CPU Throughput Implications – Faster CPU



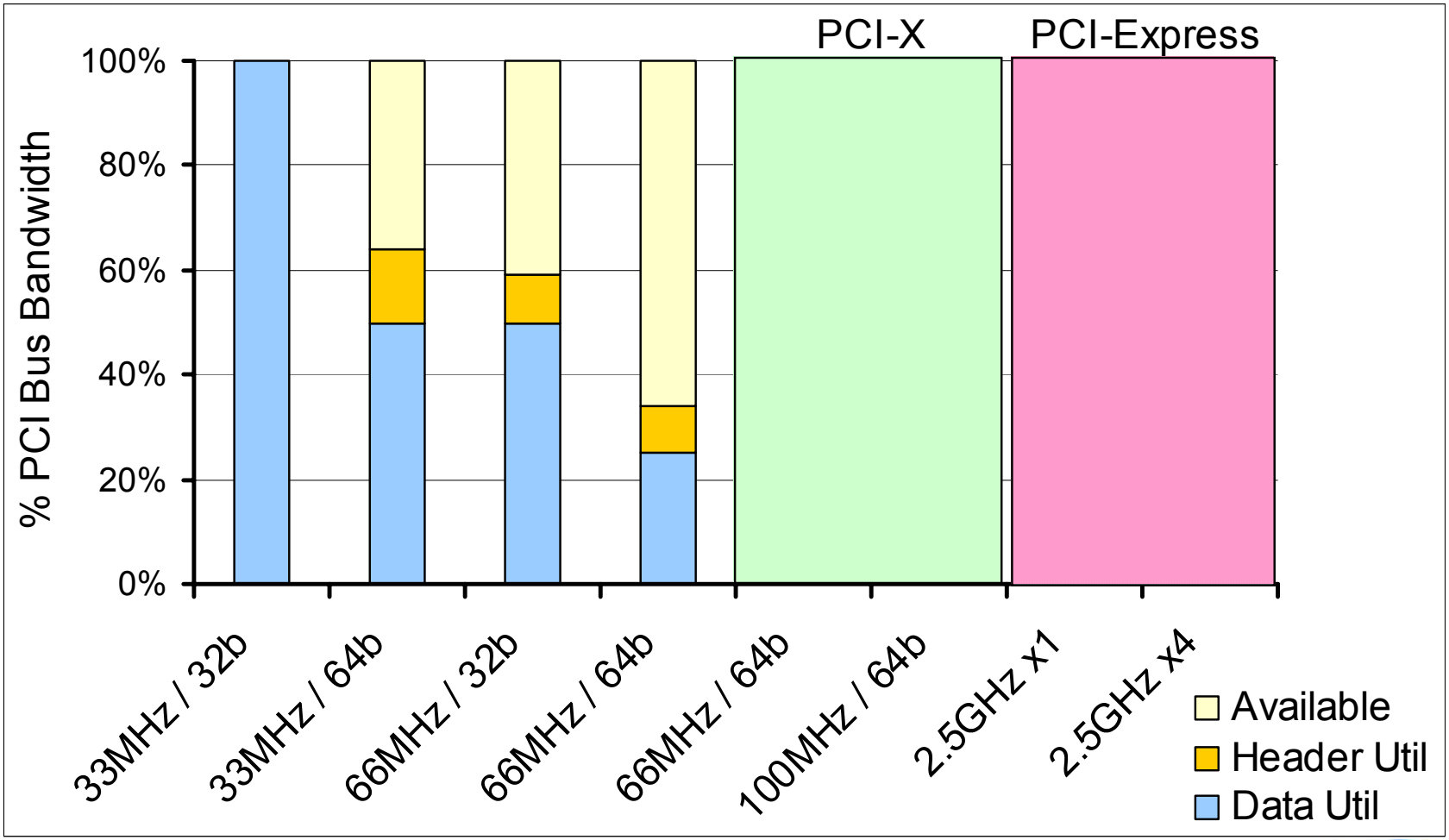

- NetIQ's™ Chariot "FileSendLong" script.

# Newer NIC Technology




- NetIQ's™ Chariot "FileSendLong" script.

# 1Gb NIC PCI Utilization




(950Mb/s, 1500B Frame Size, One Direction)

- NetIQ's™ Chariot "FileSendLong" script.

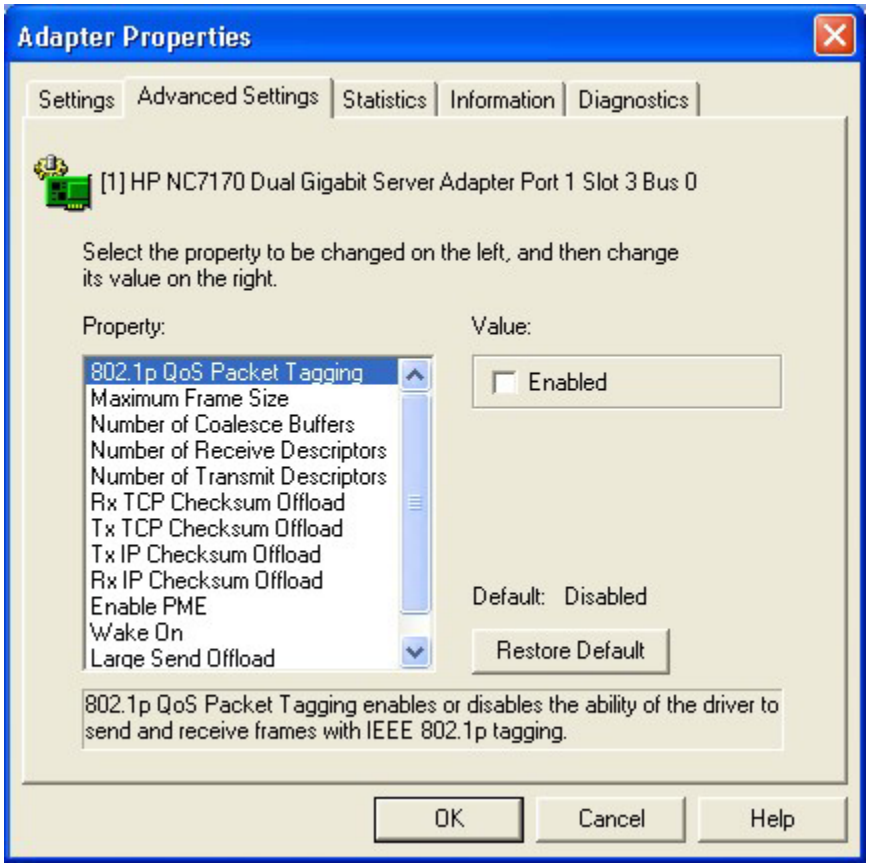
# Remember

You are only as fast as the slowest link of your network

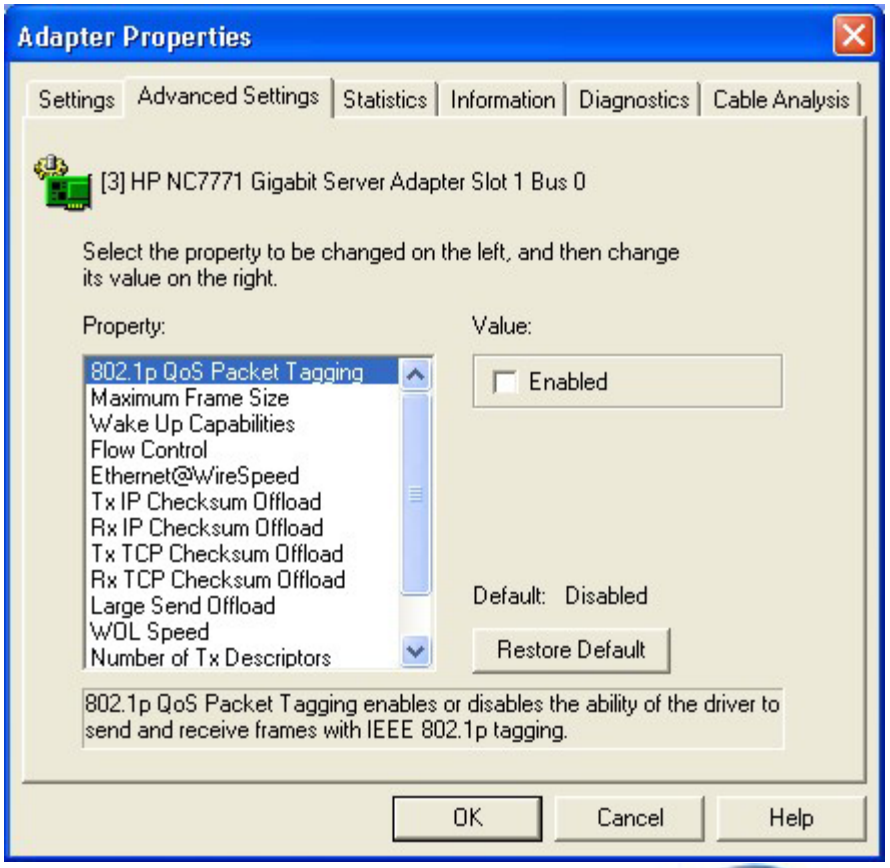
# NIC Tunables

# ProLiant NIC Tunables

## Intel Based Adapters

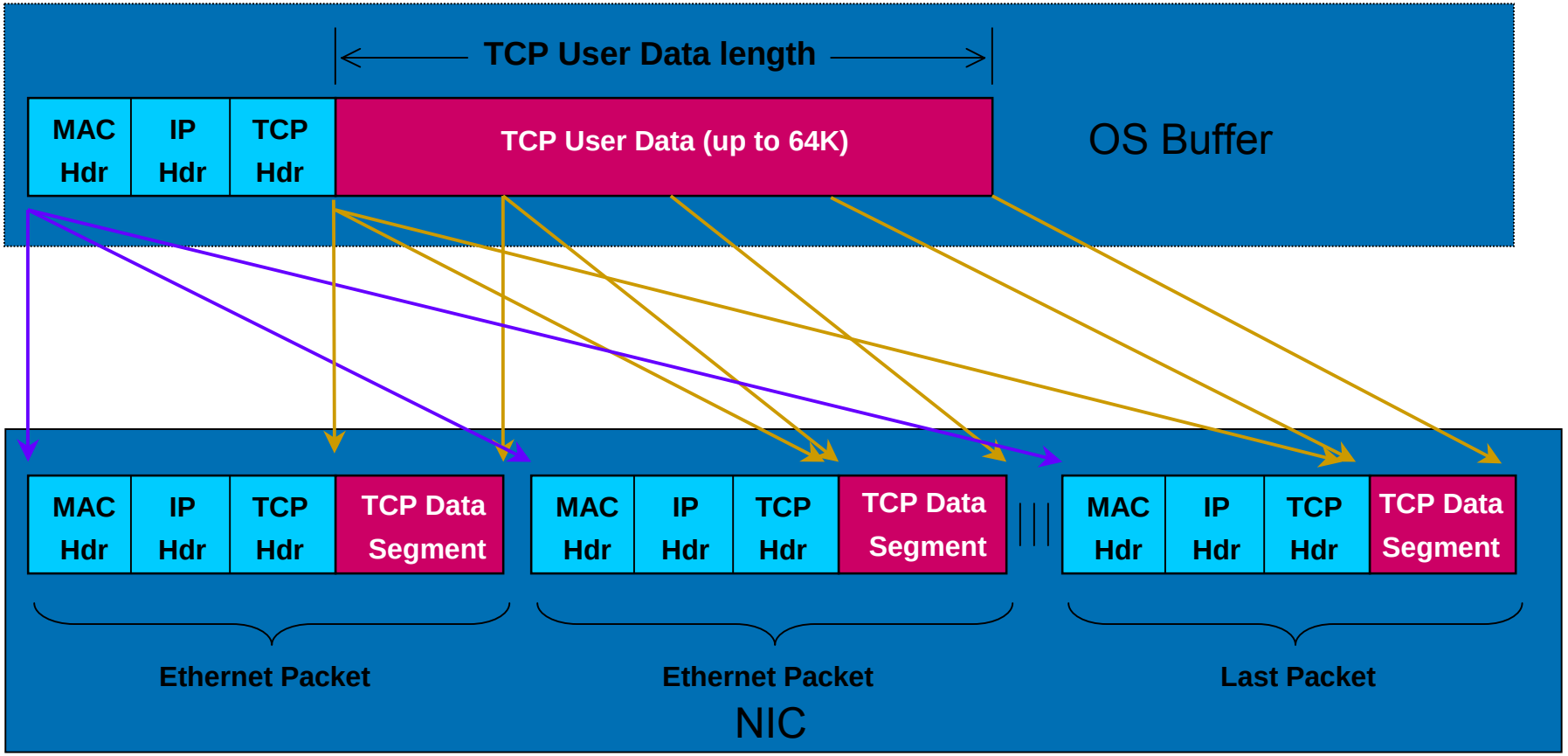



## Broadcom Based Adapters

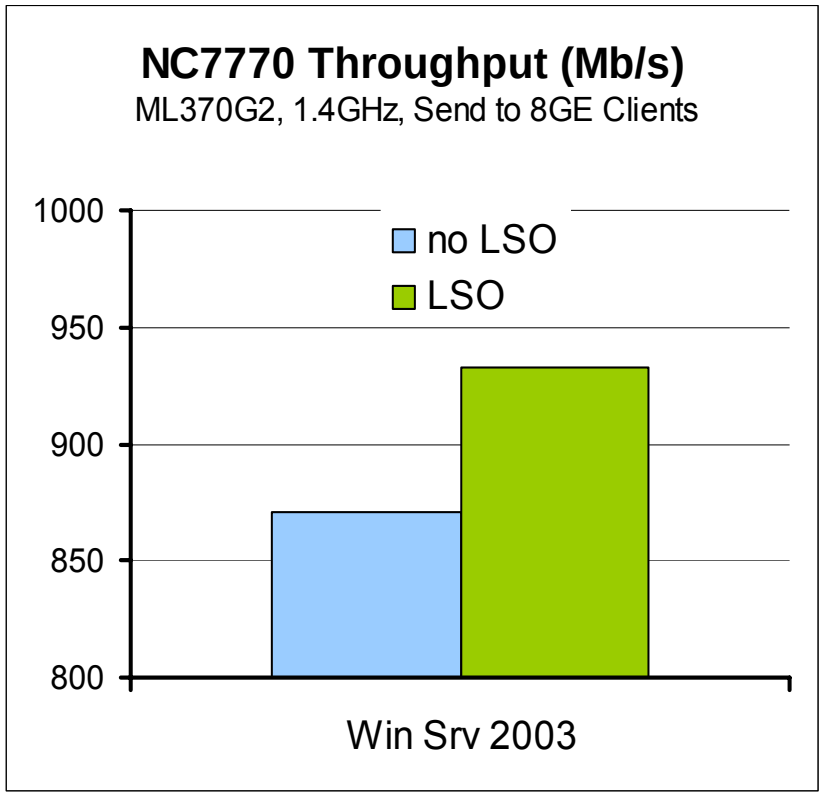


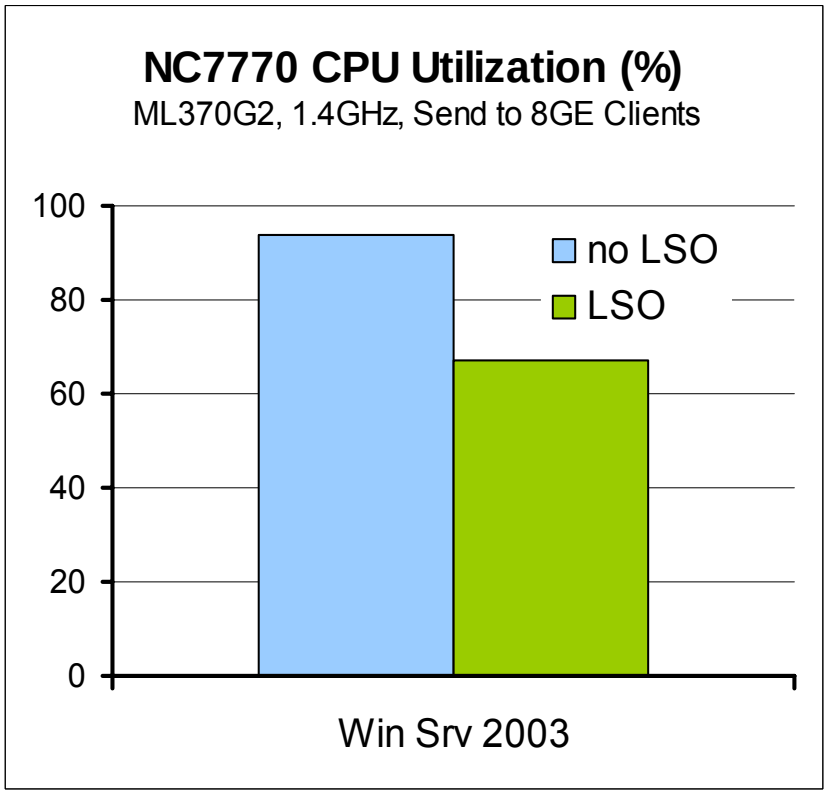
# Large Send Offload

*Also known as TCP Segmentation Offload*



# Large Send Offload




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

- NetIQ's™ Chariot "FileSendLong" script.

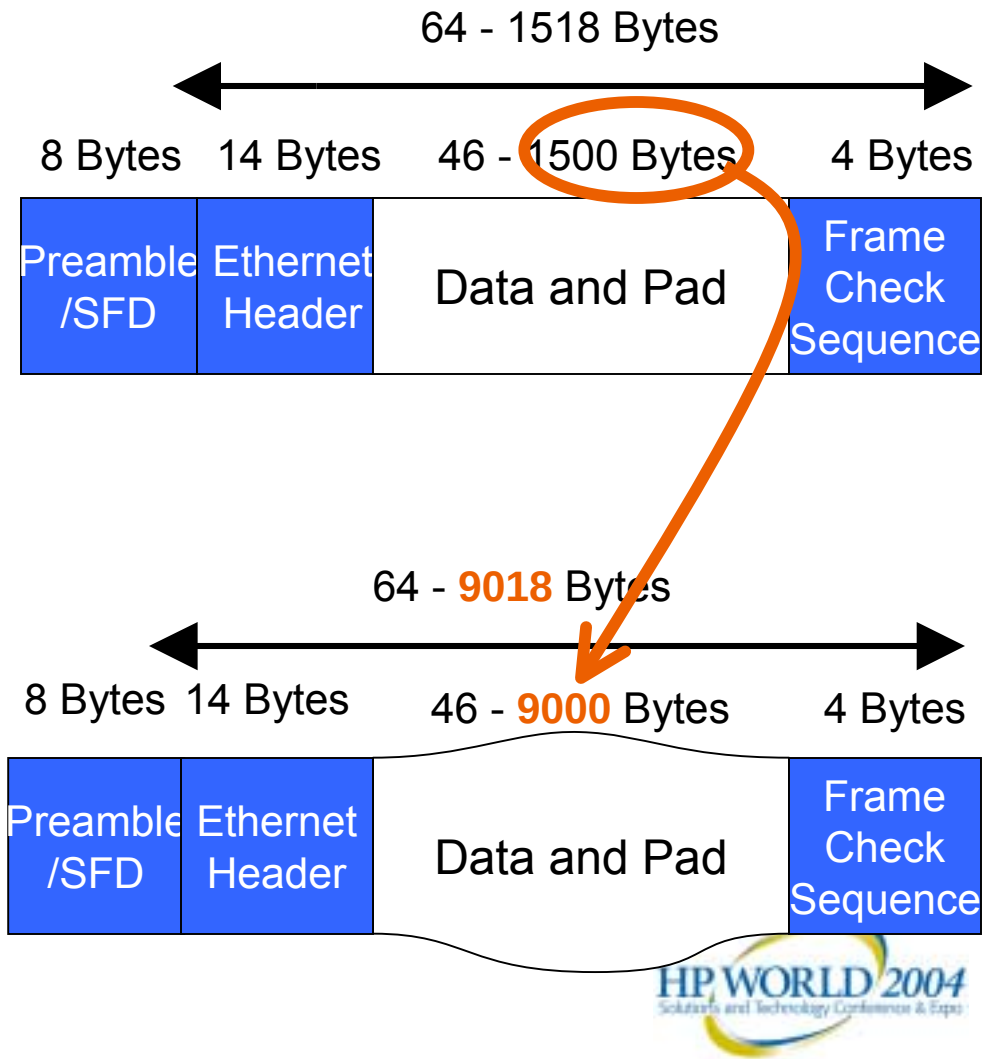
# Checksum Offload

- On the send side, the checksum offload enabled network adapter calculates any needed checksums. The Microsoft transport calculates the TCP pseudo header checksum and place this value in the checksum field, so that the network adapter can calculate the correct TCP checksum without touching the IP header.
- On the receive side, the network adapter will fill in a NDIS_TCP_IP_CHECKSUM_PACKET_INFO structure and set the appropriate bits. If for any reason the network adapter cannot perform the checksum, it will not set any bits and indicate the packet. TCP/IP will then look at the packet and calculate the appropriate checksum itself.

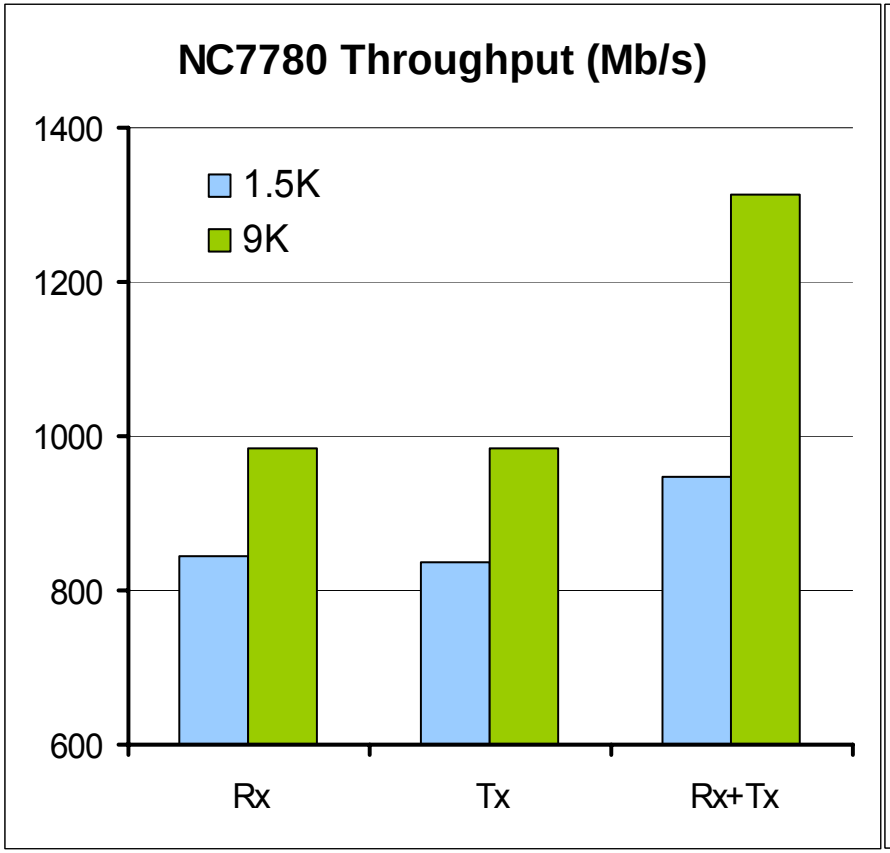
Improving Throughput Scaling

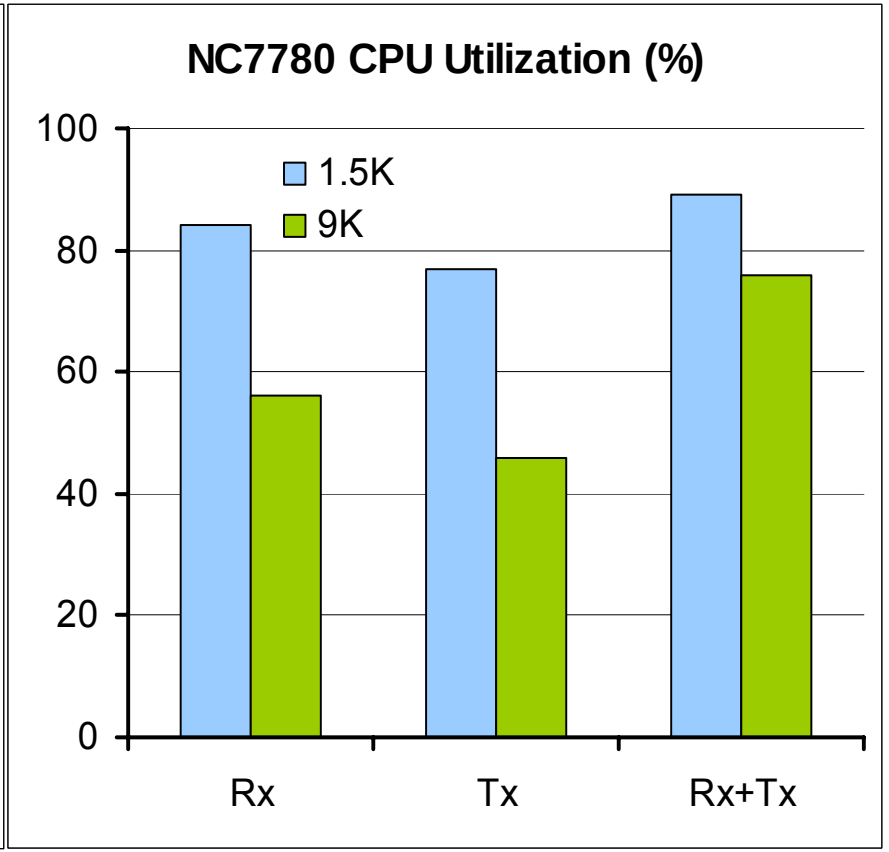
# Jumbo Frames

- Ethernet frame size is increased.
- Reduces the amount of segmentation and reassembly overhead.
- Requires all points on the network to support jumbo frames
- Informal standard.



# Jumbo Frames




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

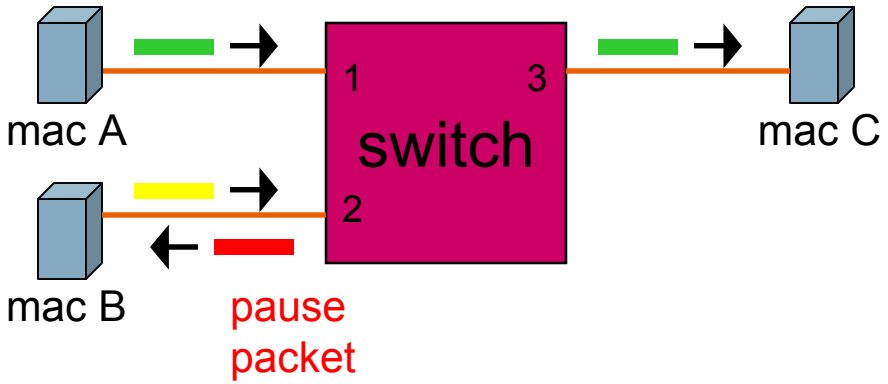
- NetIQ's™ Chariot "FileSendLong" script.

# Transmit and Receive Descriptors

- The default configuration works well in most situations
- If you run out of Tx/Rx descriptors you must wait for one to free up before additional networking transactions can take place
- Increase when using larger block sizes
- Increase for high load situations

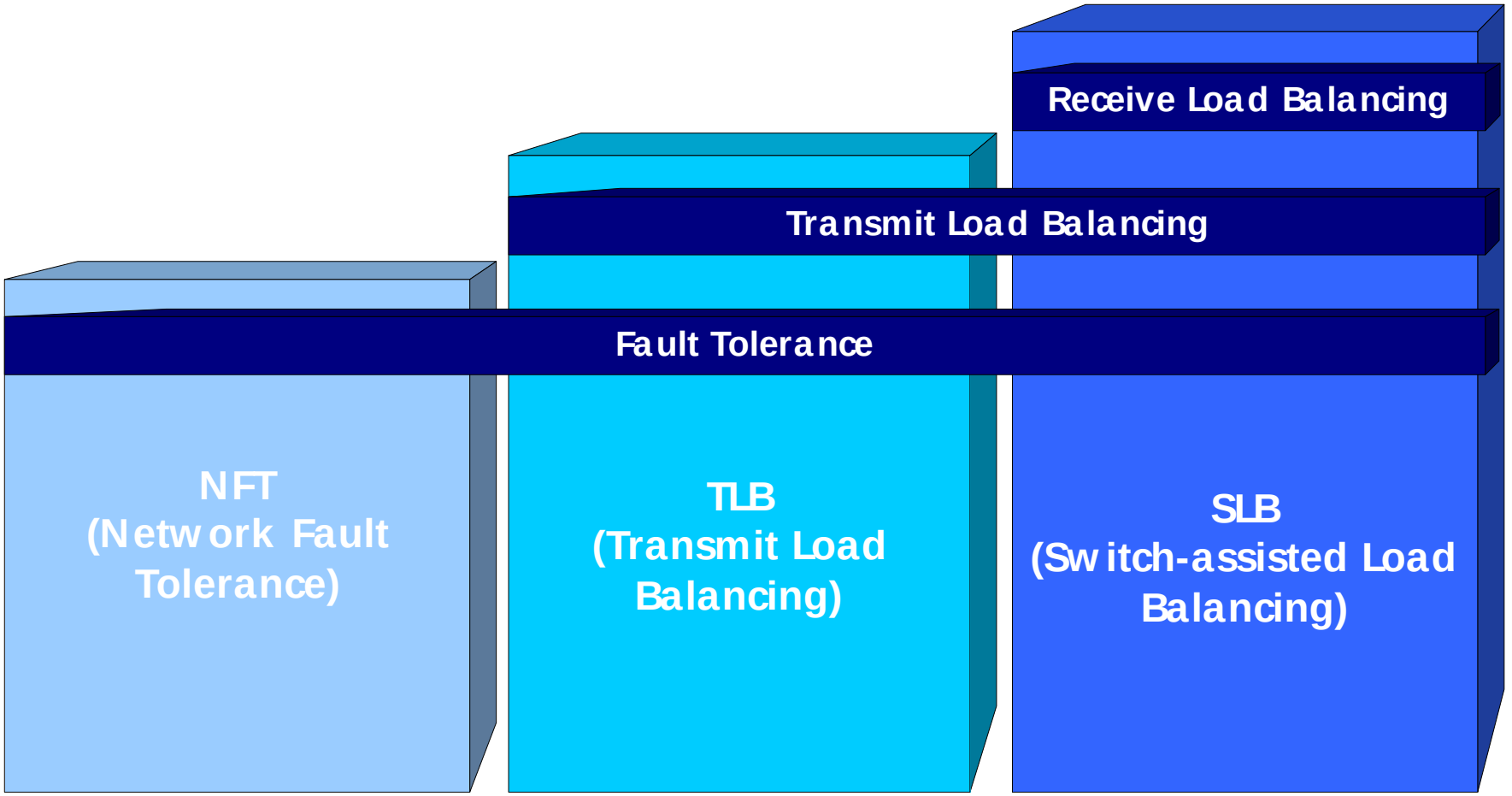
# Ethernet features:
## flow control (802.3x)

- if a device (switch) is receiving more data than it can buffer it
  - drops packets
  - or requests flow control
- IEEE 802.3x defines flow control
  - sends ‘pause packets’ to stop flow
  - prevents dropped packets
- must be supported by both end node (s) and switch

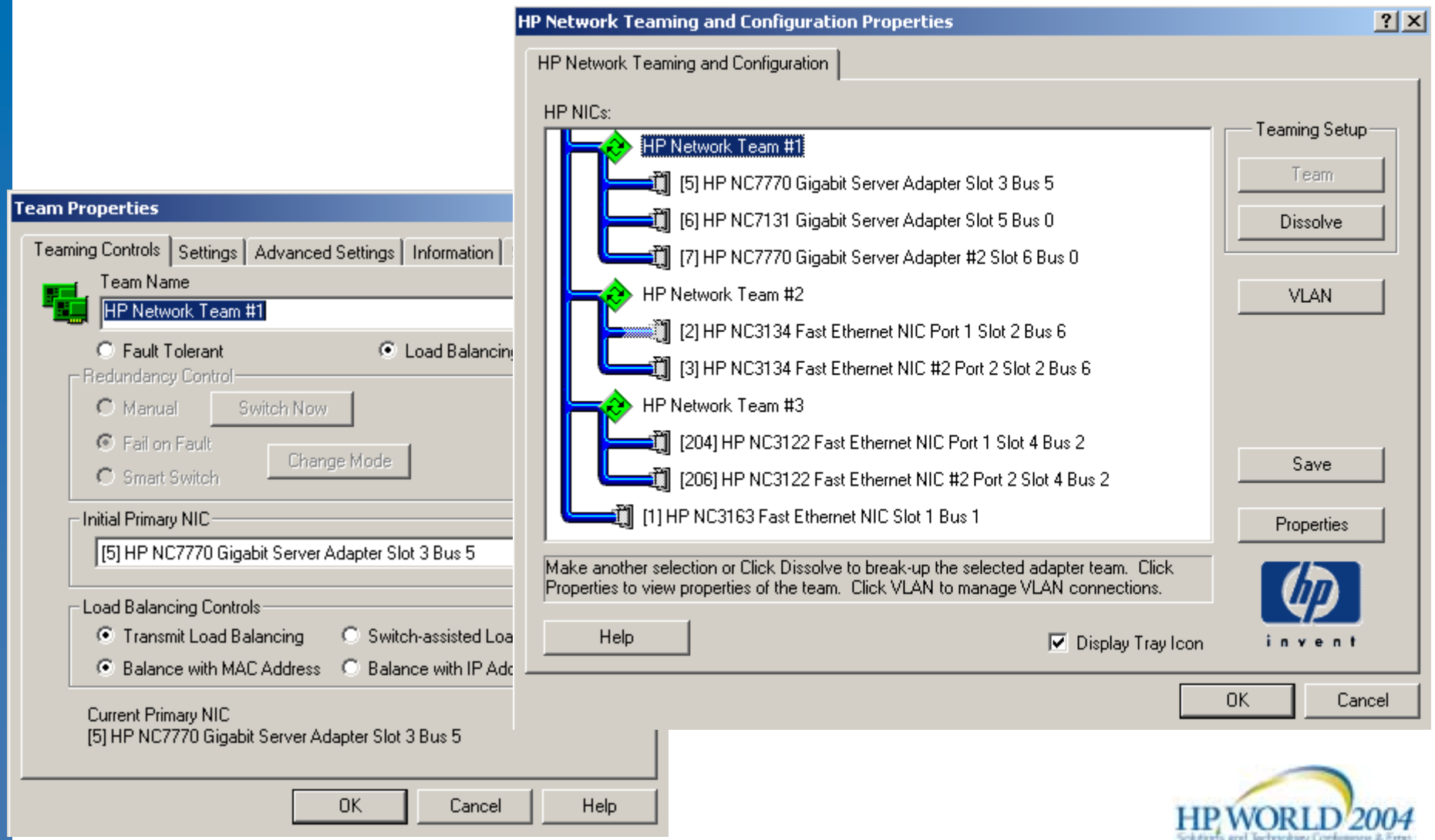


# Teaming

# Network Adapter Teaming: Three Modes

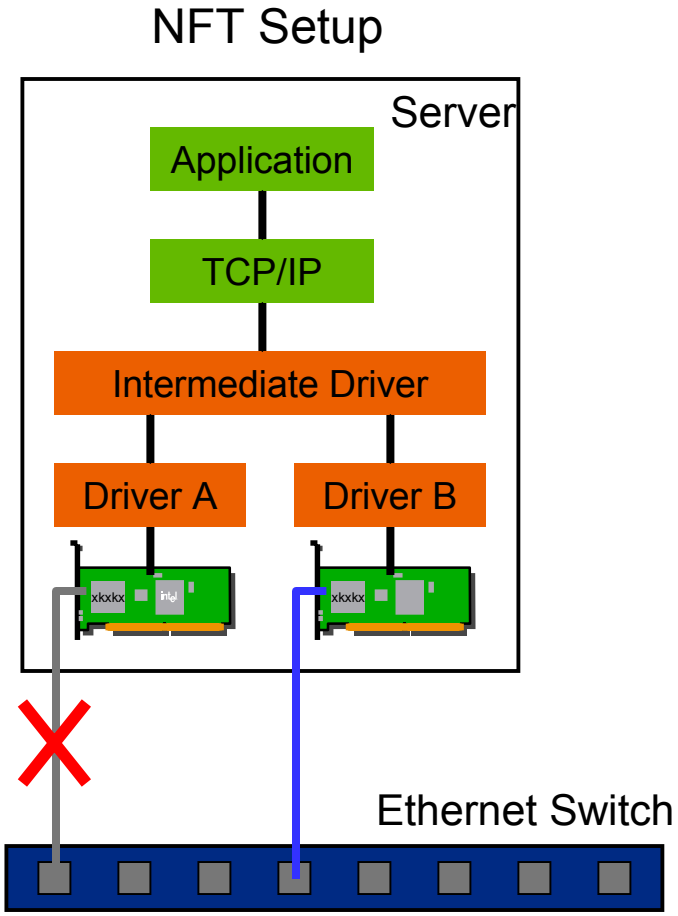


# Teaming:
# Network Configuration Utility



# Teaming Benefits

- two or more physical adapters are configured as one logical adapter
- provides support for fail-over
- provides increased throughput
- multiple algorithms available
  - fail-over
  - transmit load-balancing
  - IEEE 802.3ad standard
- reference:
  http://h18004.www1.hp.com/products/servers/networking/whitepapers.html



# NFT: Network Fault Tolerance

- Simple redundancy
- Two to eight adapters in a fault tolerant team
- One adapter is defined as the primary
  - Only the primary adapter transmits and receives
  - May have a preferred primary
  - All others are secondary and in standby mode
- Any speed, any media
- Team can be split across switches
  - But they must be on the same broadcast domain
- Can be configured to carry multiple VLANs

# TLB: Transmit Load Balancing

- Two to eight adapters in a team
- Primary adapter transmits and receives
  - All other adapters (secondaries) transmit only
- Must support a common speed
- Team can be split across switches
  - But they must be on the same Layer 2 network

Example:

TLB Team of 4 Fast Ethernet NICs

Transmit Throughput = 400Mb

Receive Throughput = 100Mb

# SLB: Switch-assisted Load Balancing

- Two to eight adapters in a team
- All adapters transmit & receive
  - Same speed, same media
- Must be used with an intelligent switch that supports this type of teaming (802.3ad, EtherChannel, etc.)
- All ports must be part of the same switch port trunk group

Example:

SLB Team of 4 Fast Ethernet NICs

Transmit Throughput = 400Mb

# Teaming Mode Comparison Table

| | NFT | TLB | SLB |
|---|---|---|---|
| Number of NICs supported per Team | 2-8 | 2-8 | 2-8 |
| Supports network adapter Fault Tolerance | X | X | X |
| Supports Transmit Load Balancing | | X | X |
| Supports Receive Load Balancing | | | X |
| Requires a switch that supports the feature | | | X |
| Can connect a single Team to more than one switch for switch fault tolerance | X | X | Depends on Switch |
| All NICs must support a common speed | | X | X |
| Can Team NICs of different media | X | X | X |

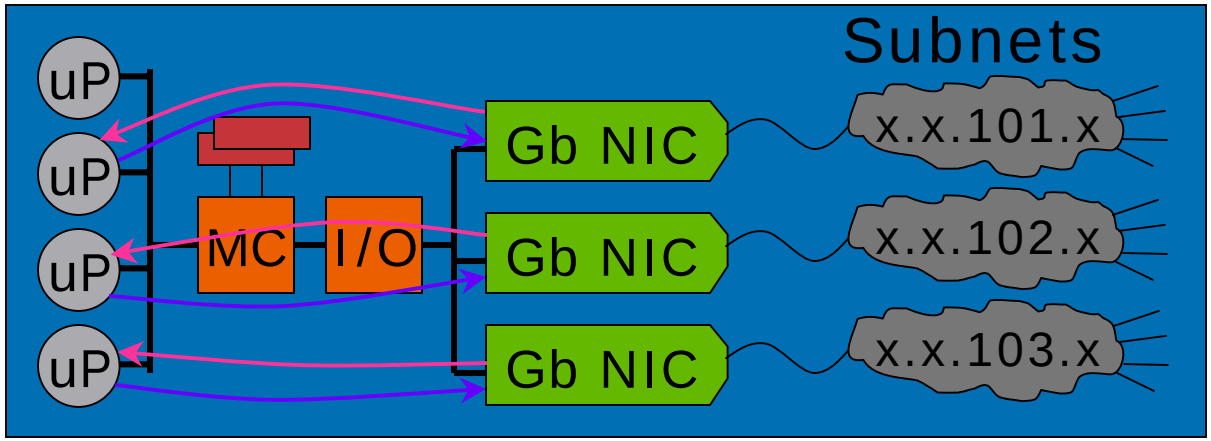
# Teaming Mode Comparison Table Cont.

| | NFT | TLB | SLB |
|---|---|---|---|
| Maximum theoretical transmit/receive throughput (in Mbps) with maximum number of 100 Mbps adapters | 100/ 100 | 800/ 100 | 800/ 800 |
| Maximum theoretical transmit/receive throughput (in Mbps) with maximum number of 1000 Mbps adapters | 1000/1 000 | 8000/1 000 | 8000/ 8000 |
| Load Balances TCP/IP Traffic | X | X | X |
| Load Balances non-IP Traffic | | | X |
| All NICs transmit using same MAC address | | | X |
| All NICs use the same IP address on the network | X | X | X |

# Multiple NIC Subnets

*Separate Subnet per NIC*

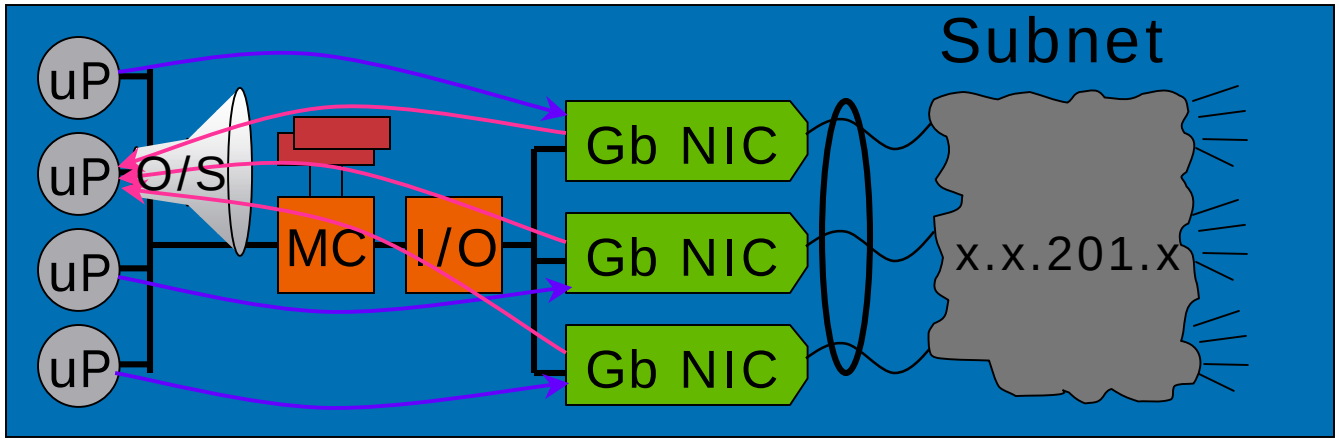
- Achieves Optimal Server Network Performance
- Harder to Admin Multi-Subnets, DHCP, and Masks
- Need One 1.4+GHz CPU per Gb NIC for Wire Speed
- Additional CPUs needed for Appl or File System
- Will work Point to Point, but needs 1+ thread per NIC



# Multiple NICs Teamed

*Link Aggregation (Alias : Team or Trunk )*

- Increases Bandwidth to Single Subnet
- Easier to Manage Single IP and Subnet
- Need One 1.4+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

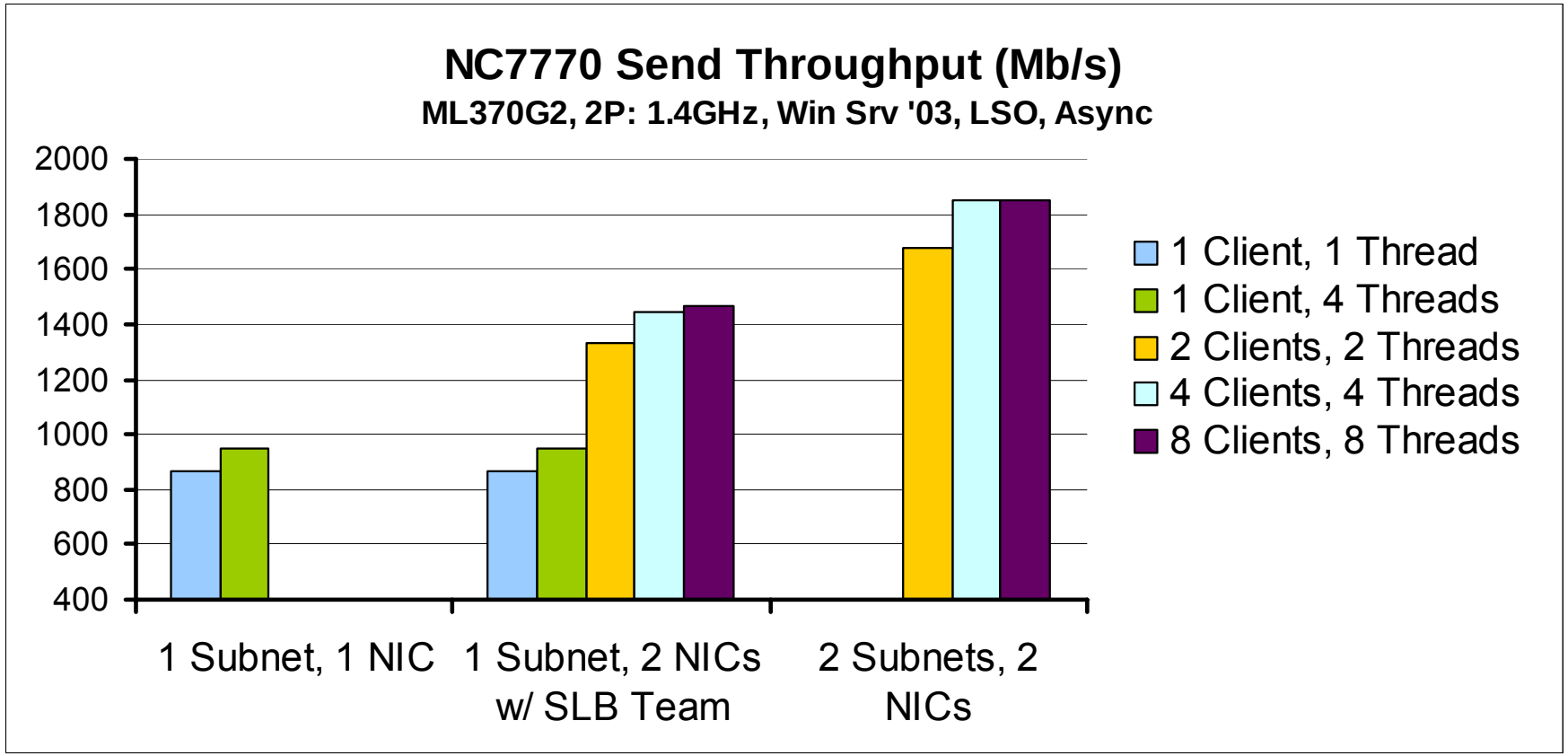


# Multi-NIC Team (cont..)

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 802.3ad 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] )*

# Multi-NICs
# Team vs. Multi-Subnets



# Future Improvements

# 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

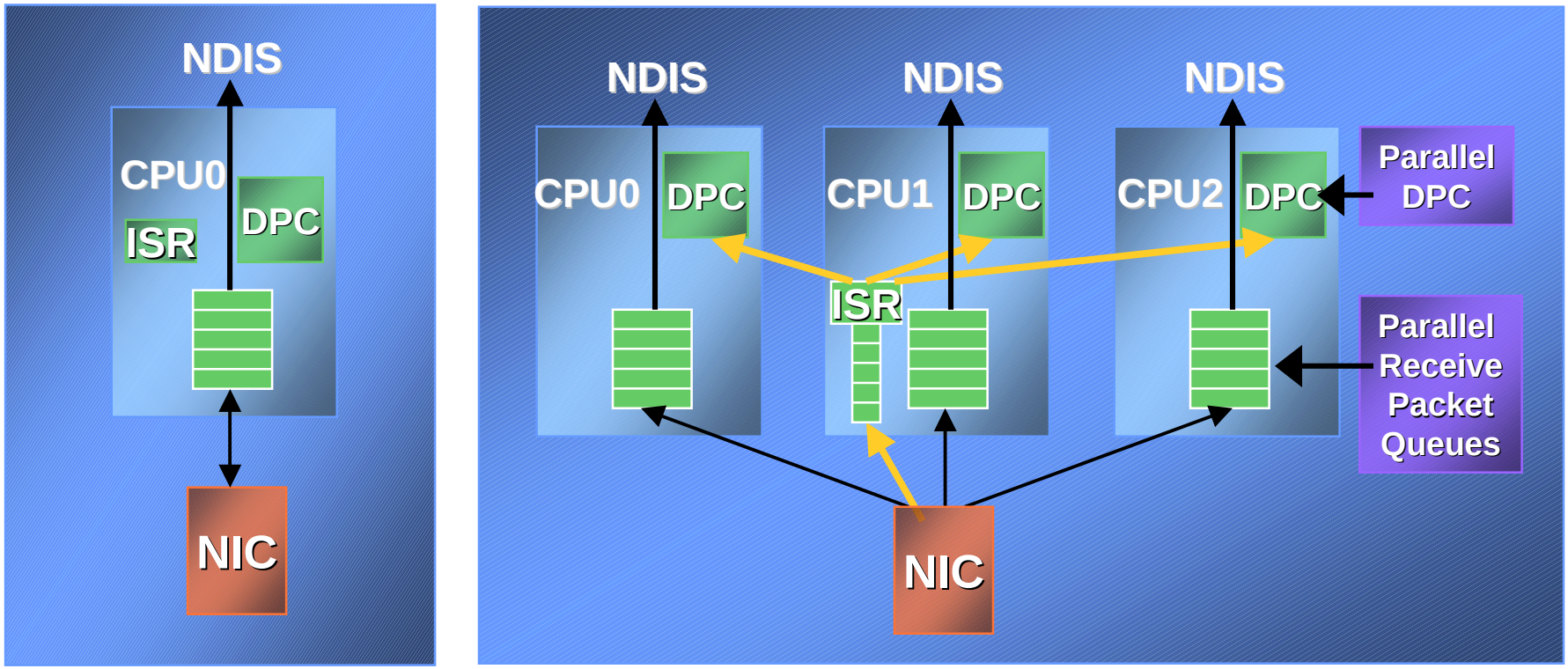
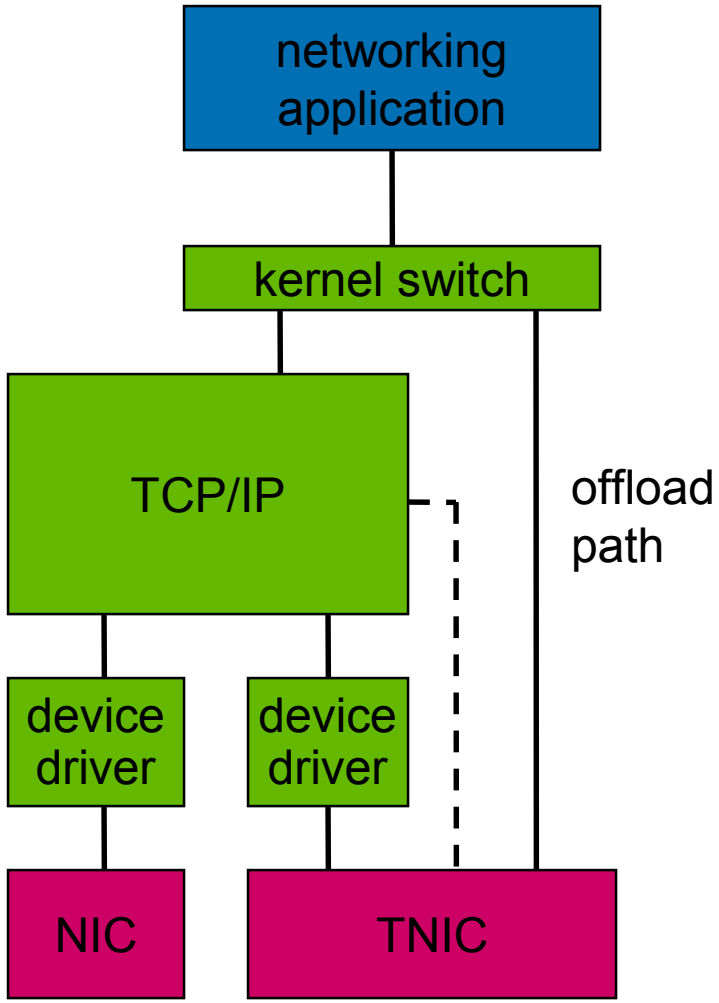



Figure courtesy of Microsoft, Copyright © 2003 Microsoft Corp.

Improving Throughput Scaling

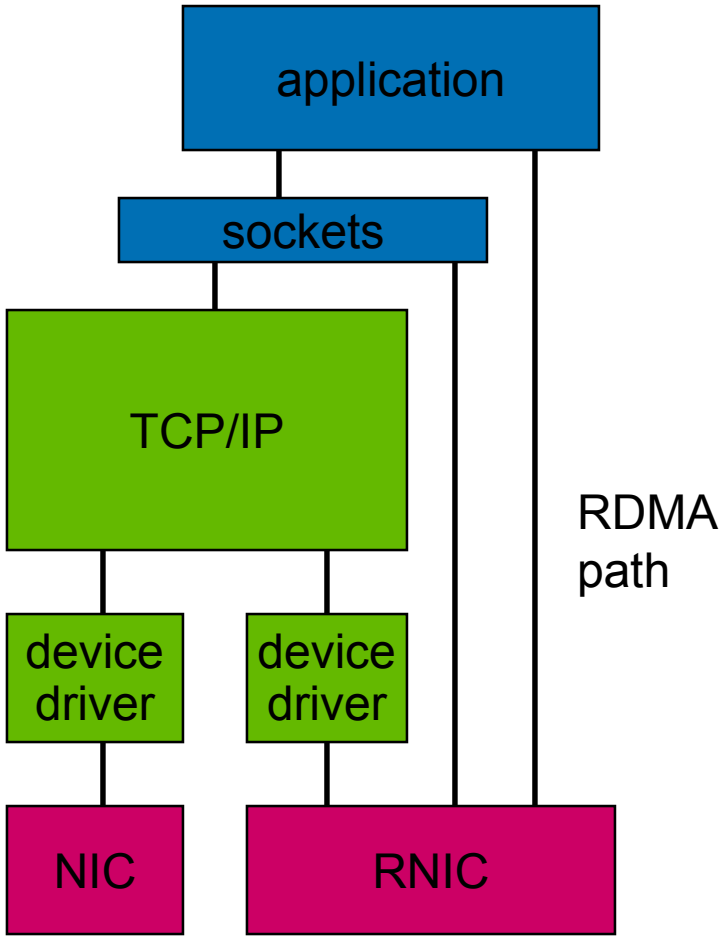
# TCP/IP Offload Engines (TOE)

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



# Remote DMA NIC (RNIC)

- Provides direct communication between application buffers in separate servers.
- Bypasses the OS 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



# Top Performance Tips

- fatter pipe
- multiple pipes (teaming)
- greater processing power
- reduce interrupts (coalescence)
- checksum offloads (TCP, IP, UDP)
- large send offload (LSO)
- use asynchronous sockets
- receive side scaling
- TCP/IP offload (TOE)
- RDMA offload (RNIC)

Now

Future

# Whitepapers and References

- HP Network Adapter Teaming Technology
  - http://h18000.www1.hp.com/products/servers/networking/whitepapers.html
- Network-Level Redundancy/Resilience For High-Availability Campus LANs
  - http://www.cisco.com/warp/public/779/largeent/learn/technologies/campuslan.pdf
- Network Adapter and Teaming Drivers
  - http://h18007.www1.hp.com/support/files/networking/nics/index.html

# HP WORLD 2004

Solutions and Technology Conference & Expo

Co-produced by:

interex

shared knowledge • shared power

encompass

AN HP USER GROUP

RECOMMENDED TRAINING VENUE FOR THE

HP Certified Professional