HP WORLD 2004
Solutions and Technology Conference & Expo

Ethernet's Extreme Makeover: Multifunction Networking Adapters

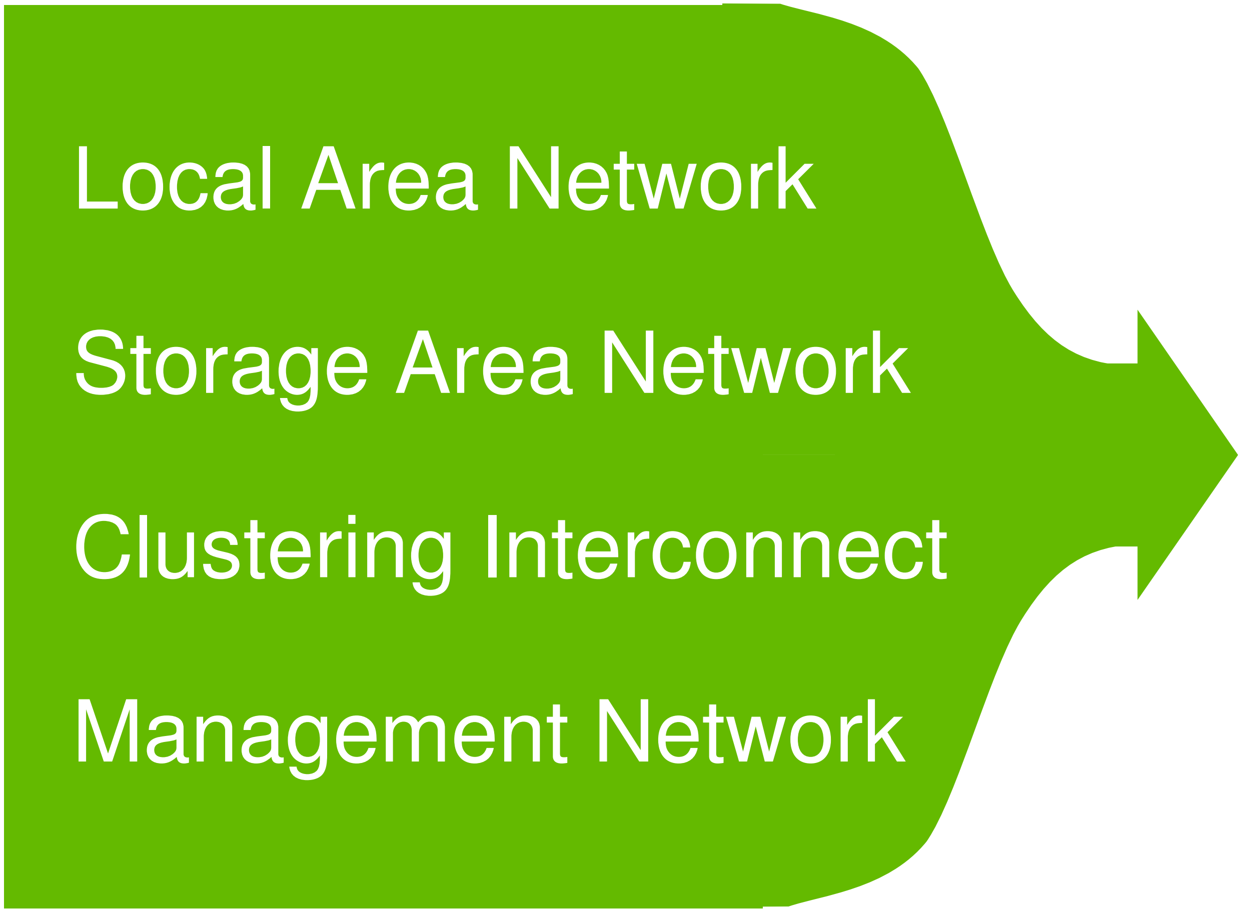
Chuck Hudson

Manager, ProLiant Networking Technology
Hewlett-Packard

Outline

- Introduction
- Multifunction Networking Technology
- Performance
- Application Programming Interfaces
- Matching Applications and Offload Technology
- Technology Availability

What is multifunction networking?



Local Area Network

Storage Area Network

Clustering Interconnect

Management Network

Multifunction Network

Why use a multifunction networking adapter?

- Flexibility
  - Pick and chose the capabilities you need
  - Each interface can be used for any one or more roles
- Better Networking
  - Adapters include the latest performance features
- IO Consolidation
  - Multi-use interfaces allow for fewer cables
- Enable Server Virtualization
  - iSCSI boot
  - Flexible IO connectivity

Multifunction Networking Technology

RSS

iSCSI TOE

iSER LS
O

RDMA

Etherne
t

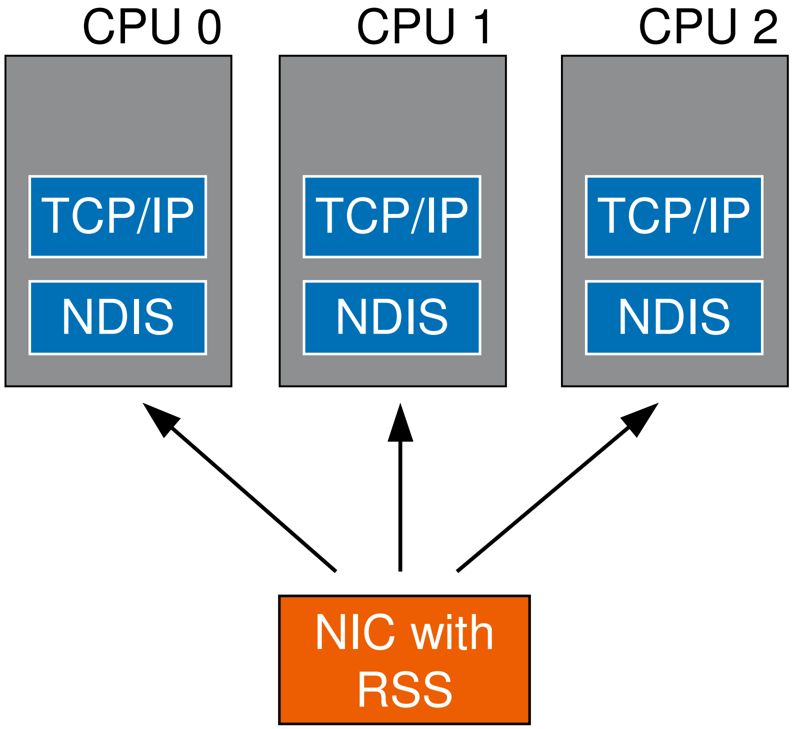
iLO

Ethernet Today

- 1Gb Ethernet (10/100/1000BASE-T)
- TCP/IP Checksum Offload
- Jumbo Frames
  - Uses larger frames (typically 9K bytes)
  - Reduces TCP/IP segmentation/reassembly and interrupts
- Large Send Offload / Segmentation Offload
  - Works for transmits only
  - Similar benefits to jumbo frames
- Asynchronous Sockets
  - Avoids extra data copies (transmits)

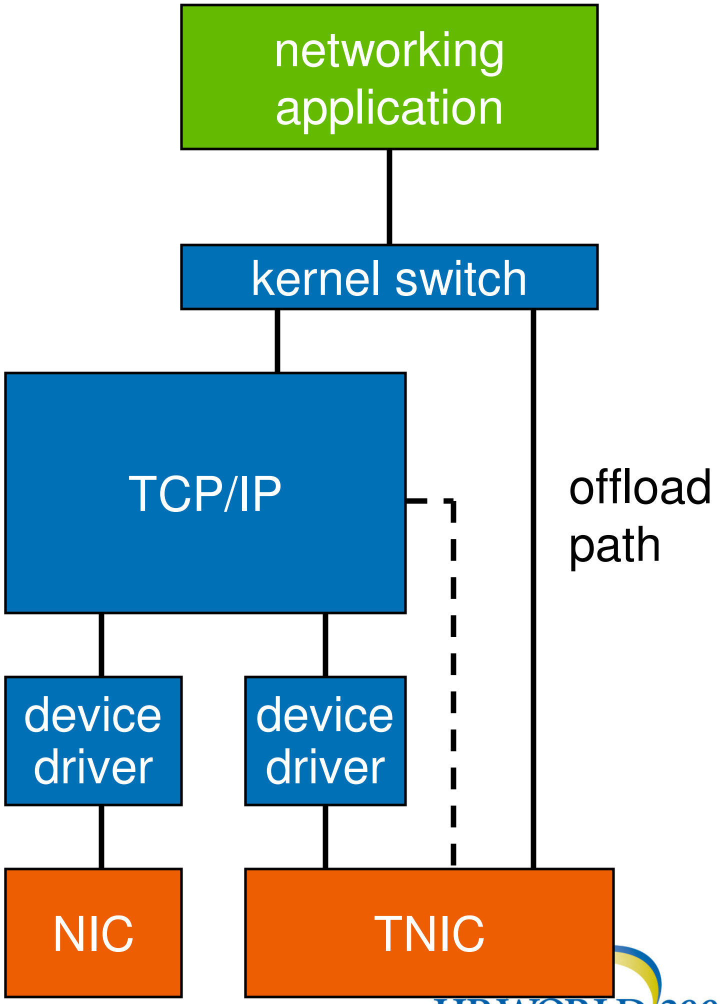
Receive Side Scaling (RSS)

- Spreads incoming connections across the CPUs within a server.
- Overcomes the single CPU bottleneck.
- Works well in applications with lots of short-lived connections (where TOE doesn't work well).
- Supported on Windows 2003 with Scalable Networking Pack (Beta in 2H2004).



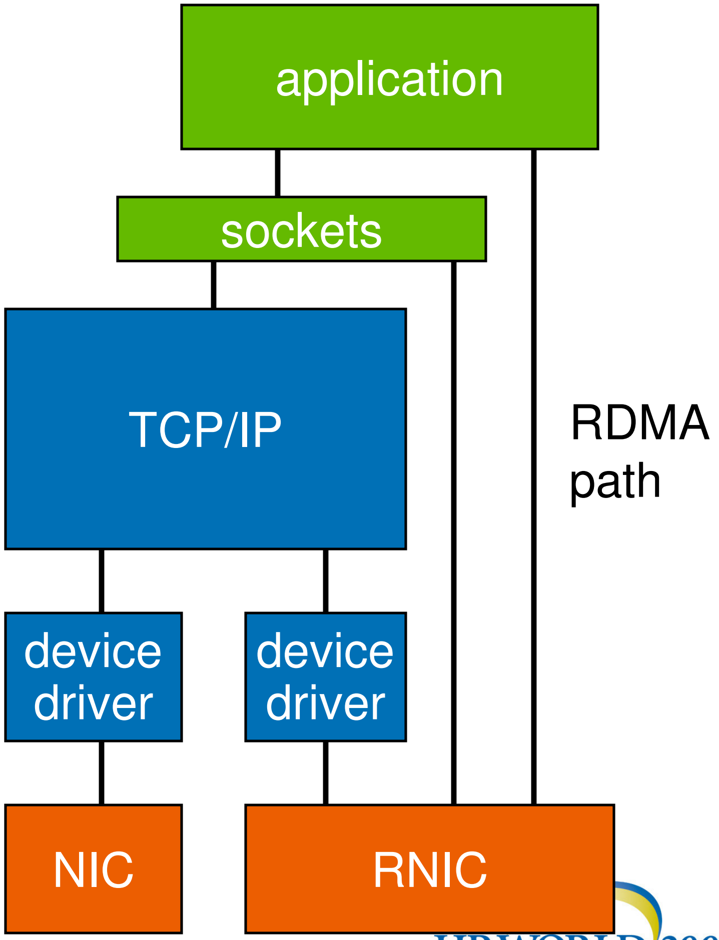
TCP/IP Offload Engines (TOE)

- TCP/IP processing moved from the host CPU to TOE NIC (TNIC)
- Improves performance
  - Reduces CPU utilization for segmentation and reassembly
  - Reduces interrupts and context switches
  - Allows for zero-copy receives to kernel memory buffers
- Supported on Windows 2003 with Scalable Networking Pack (Beta in 2H2004).

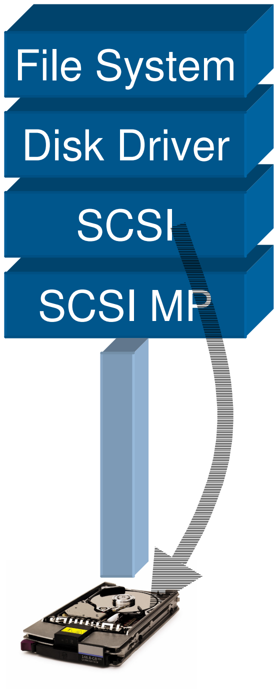


Remote DMA NIC (RNIC)

- 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



iSCSI standard for SCSI over TCP/IP



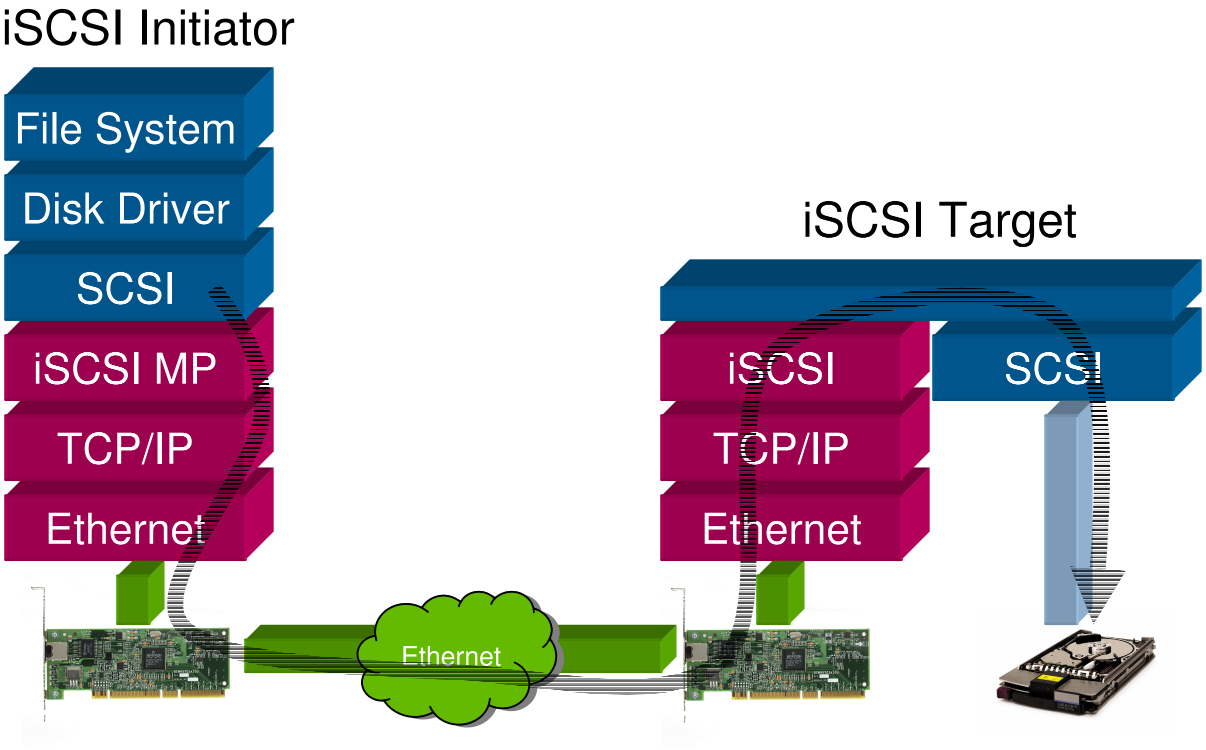
File System

Disk Driver

SCSI

SCSI MP

direct attached block storage using SCSI/SAS



iSCSI Initiator

File System

Disk Driver

SCSI

iSCSI MP

TCP/IP

Ethernet

iSCSI Target

iSCSI

TCP/IP

Ethernet

SCSI

remote block storage using iSCSI

iSCSI & iSER

- iSCSI Standardized by the IETF
  - By the Internet Engineering Task Force (IETF) IP storage working group (RFC 3720)
  - http://www.ietf.org/html.charters/ips-charter.html
- iSCSI Microsoft software initiator available today
  - http://www.microsoft.com/windows/storage/iscsi.mspx
- Linux iSCSI initiator available
  - ‘UNH’ and ‘Cisco’ iSCSI sourceforge projects
- iSCSI Extensions for RDMA (iSER)
  - Specified by the RDMA consortium
  - iSER provides better performance by allowing the storage device to control the data transfers

Multifunction Network Adapters…

Local Area Network

Clustering Interconnect

Storage Area Network

TOE iSCSI

RDMA

…include protocol offload technology

- One card
- One cable
- Multiple functions

Performance

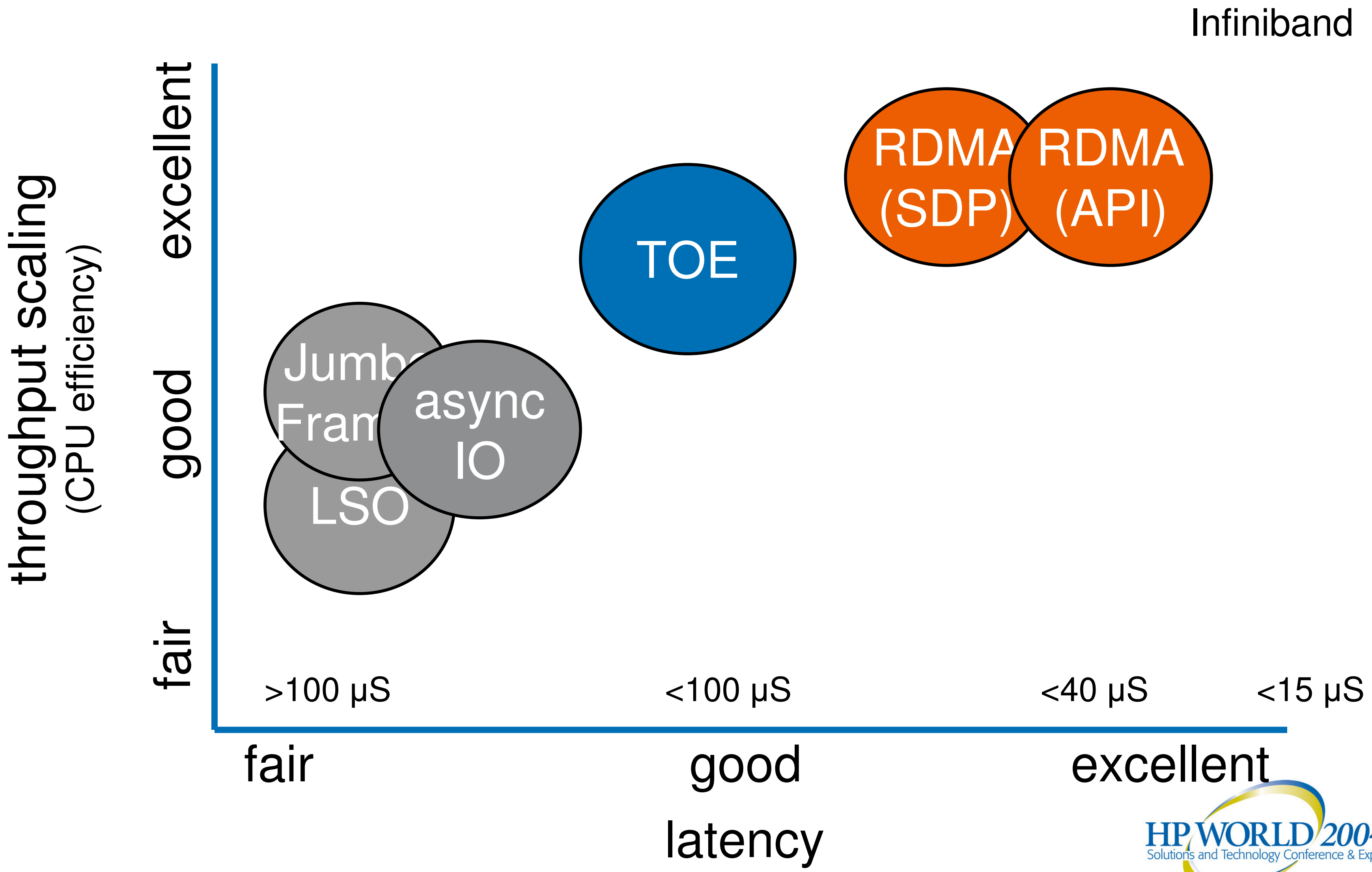
Performance Dimensions

- Throughput
- CPU Utilization

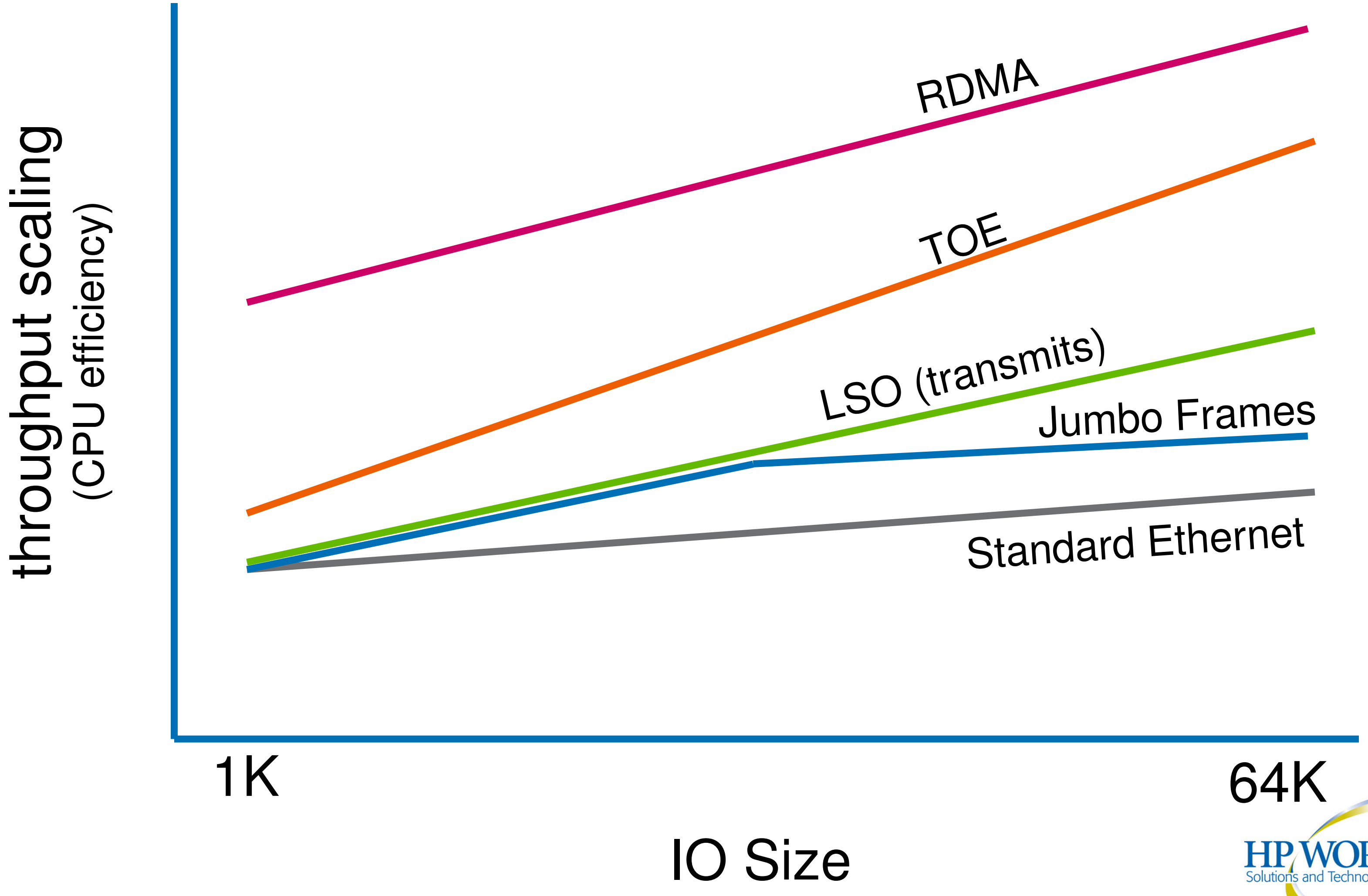
CPU Efficiency

- Latency
- Memory Bus Usage (data copies)
- IOs per second (IOPS)
- IO Size
- Connection Lifetime
- Transmit vs. Receive
- Cost

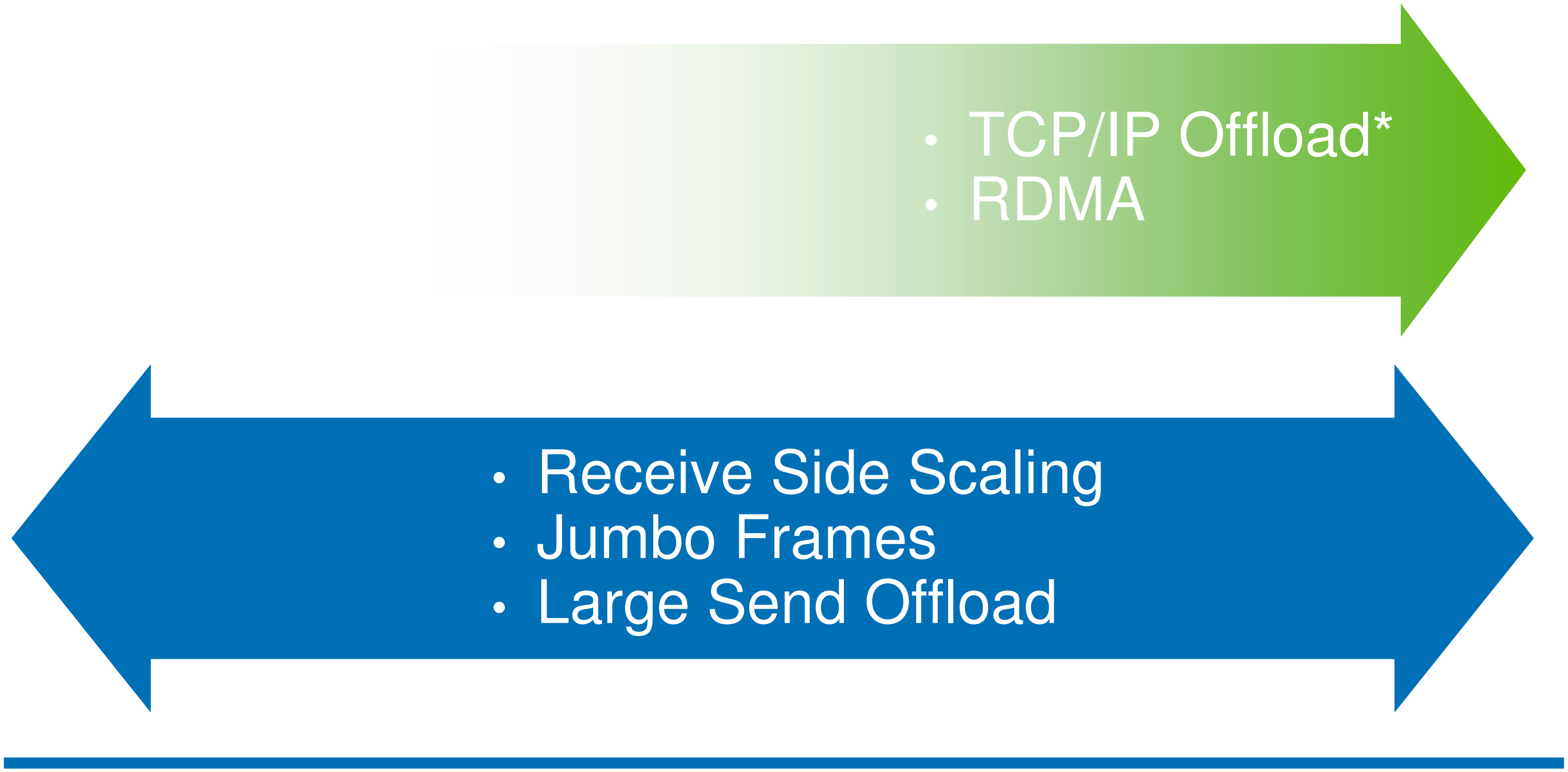
Throughput and latency



Throughput and IO Size



Connection Duration



- TCP/IP Offload*
- RDMA

- Receive Side Scaling
- Jumbo Frames
- Large Send Offload

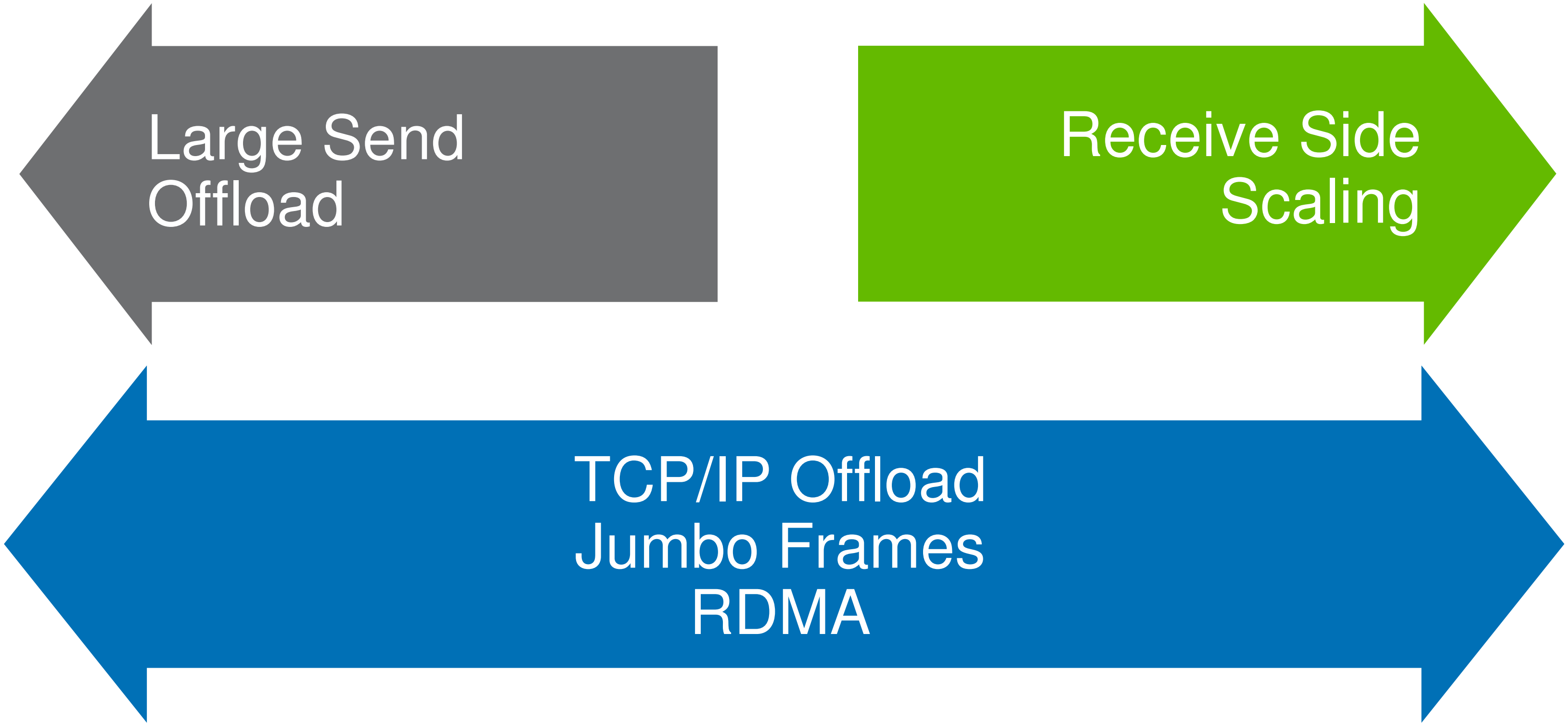
Short-lived Connections
(e.g., HTTP 1.0)

Long-Lived Connections
(e.g., Backup)

Connection Duration

* host-based connection setup

Transmit / Receive



Large Send Offload

Receive Side Scaling

TCP/IP Offload
Jumbo Frames
RDMA

Improves Transmit Performance

Improves Receive Perfomance

Transmit vs. Receive

* host-based connection setup

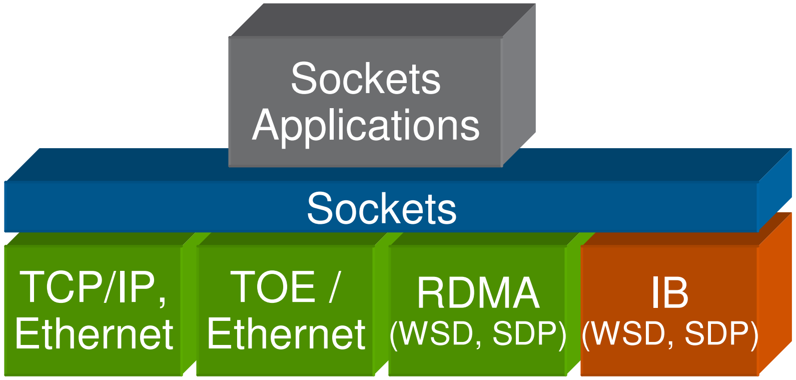
Performance Benefits Summary

Technique	Benefit	Where?
Jumbo Frames	Reduces CPU utilization (segmentation and reassembly) and interrupts for large transfers.	Requires equipment that supports jumbo frames all through the network.
Large Send Offload	Reduces CPU utilization (segmentation and reassembly) and interrupts for large transmits.	Only helps transmits.
Receive Side Scaling	Distributes connections and receive processing across CPUs. Improves scaling, but not efficiency.	Works well for short-lived connections where other techniques will not work well.
TOE host-based connections	Reduces CPU utilization and interrupts for large transfers. Zero copy on transmits (receives with pre-posted buffers).	Needs long-lived connections.
RDMA via Sockets	TOE benefits plus zero copy on receives. Reduced latency.	Long-lived connections. Passing information from point to point.
RDMA via APIs	Benefits as above PLUS ability to post once for many reads. Best low-latency messaging.	Long-lived connections Multiple nodes (single posted buffer can be read by many).

Applications & Programming Interfaces

Multifunction Networking builds on Standard Interfaces

Networking:



Sockets Applications

Sockets

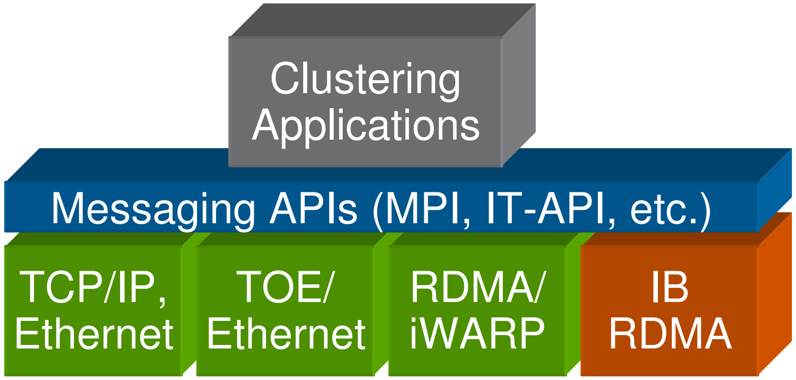
TCP/IP, Ethernet

TOE / Ethernet

RDMA (WSD, SDP)

IB (WSD, SDP)

Cluster:



Clustering Applications

Messaging APIs (MPI, IT-API, etc.)

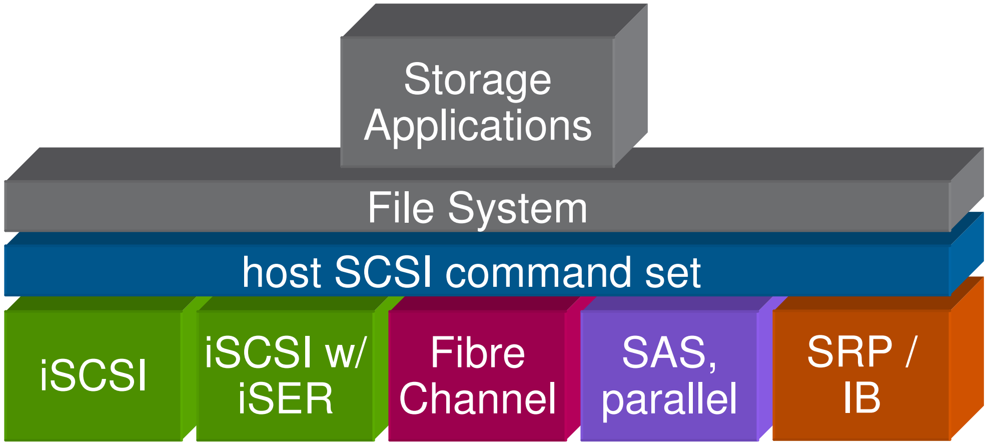
TCP/IP, Ethernet

TOE/ Ethernet

RDMA/ iWARP

IB RDMA

Storage:



Storage Applications

File System

host SCSI command set

iSCSI

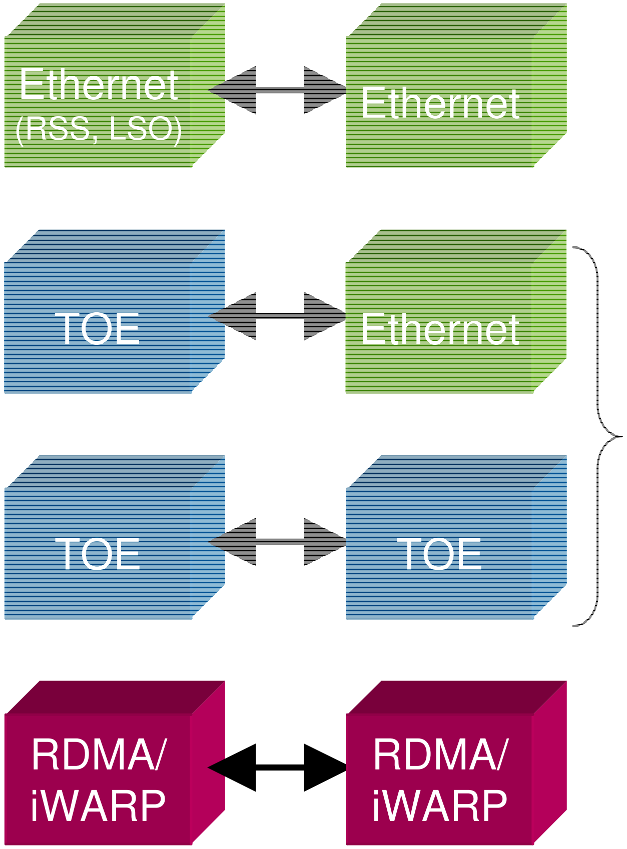
iSCSI w/ iSER

Fibre Channel

SAS, parallel

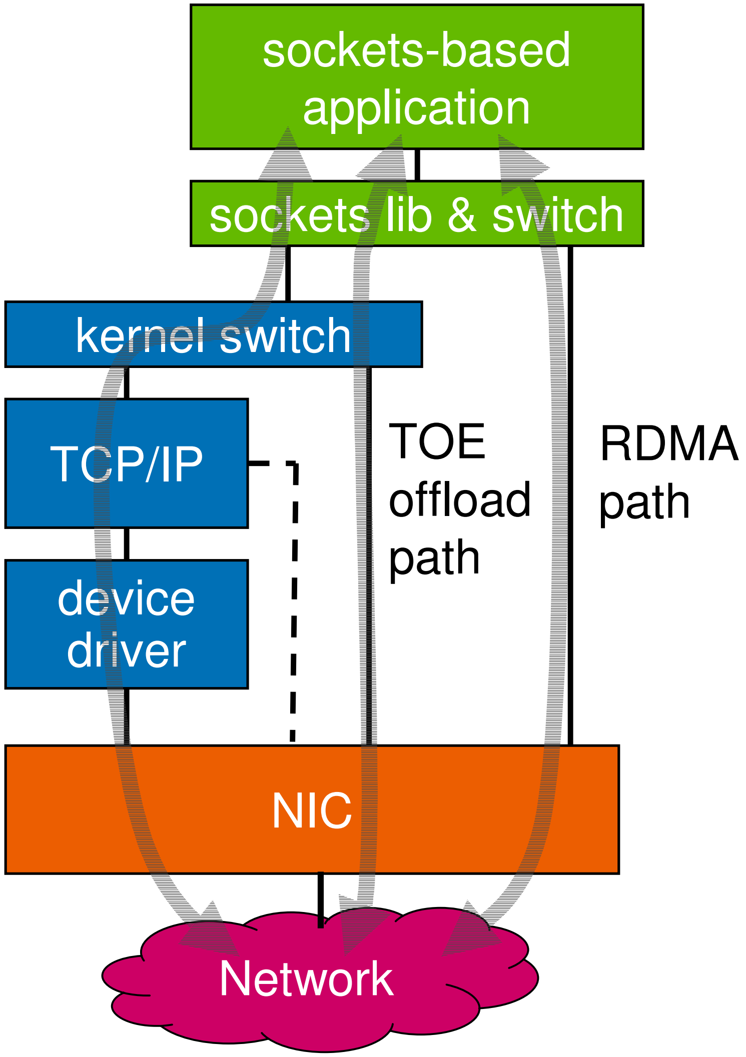
SRP / IB

Socket Applications use of offload technologies



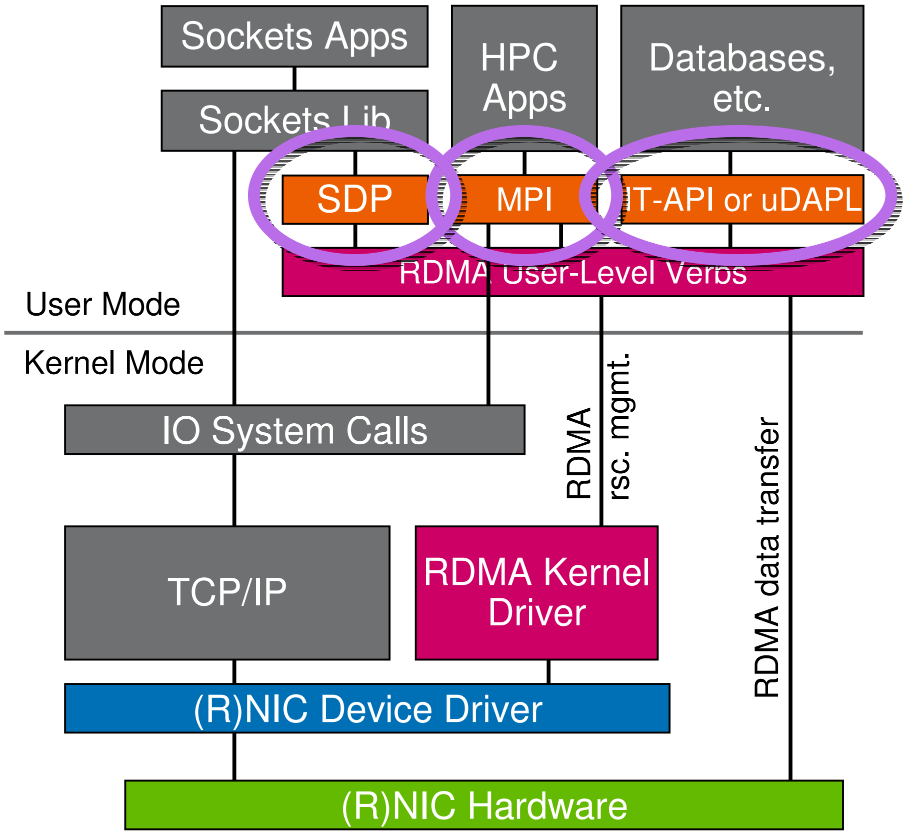
Server	"Client"	Best for this Connection Type	Application Examples
Ethernet (RSS, LSO)	Ethernet	Short-lived connections	Web servers (Server to web client)
TOE	Ethernet	Long-lived connections. Larger message sizes.	NAS, backup, iSCSI
TOE	TOE		
RDMA/ iWARP	RDMA/ iWARP	Long-lived connections. Any message Size	Databases, HPC, Grid, MS Exchange, NAS, backup, iSCSI/iSER

Let the OS choose the best transport for sockets applications



- If both sides have matching RDMA
  - the OS will connect with RDMA (SDP or WSD).
- If not,
  - Long-lived TCP/IP connections will be offloaded
  - Short-lived can make use of existing partial offload
    - TCP/IP Checksums
    - Large Send Offload
    - Receive Side Scaling
    - Jumbo Frames

Clustering APIs



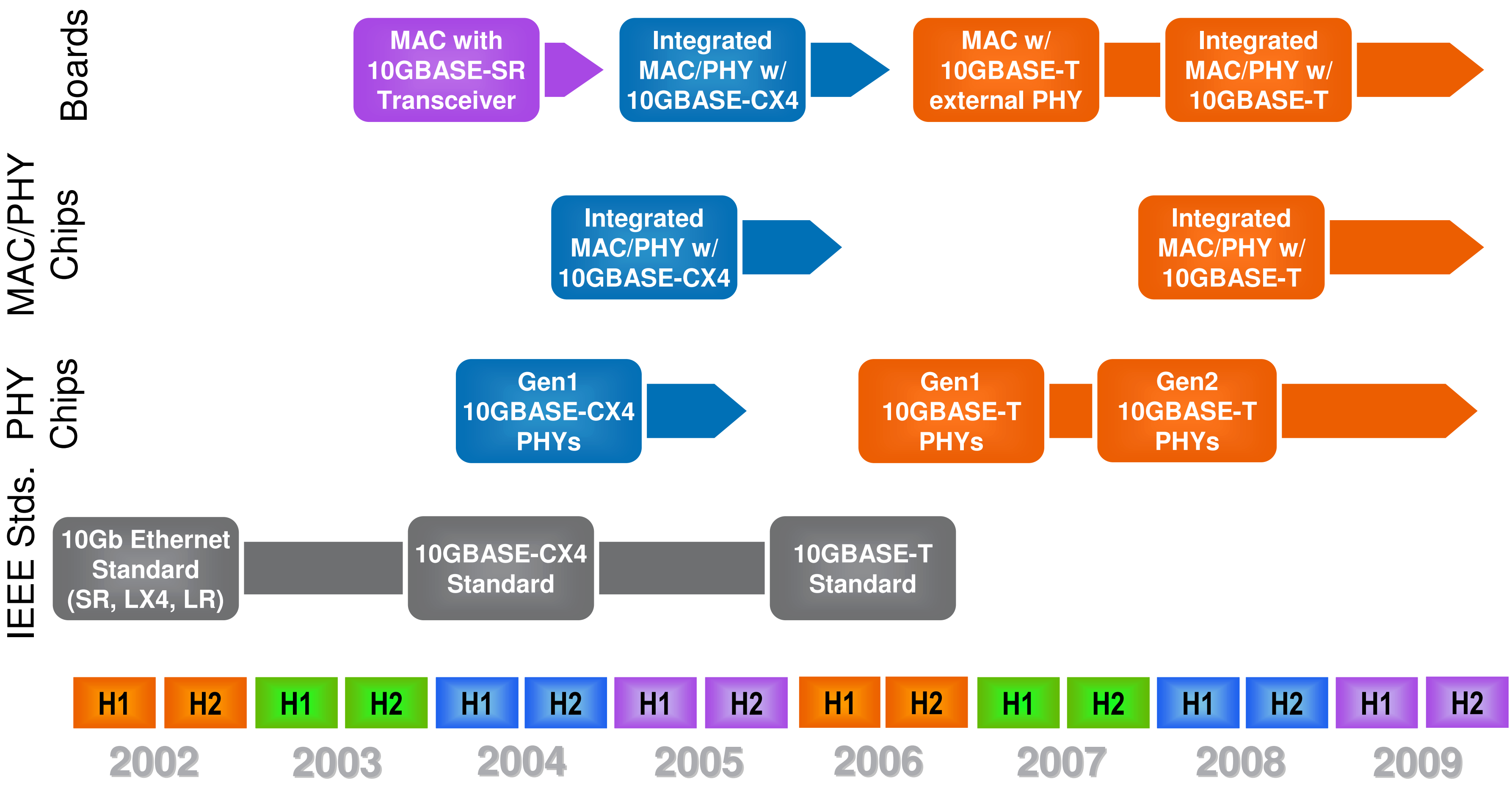
- Message Passing Interface (MPI)
  - Used by most HPC applications
  - Runs on Ethernet, RDMA over TCP/IP, and other RDMA media (Infiniband).
- IT-API
  - Open Group’s Transport API for RDMA over TCP/IP/Ethernet
  - uDAPL is Infiniband version
  - Used by applications that want to manipulate buffers and queue pairs directly.
- Sockets Direct Protocol
  - Allows sockets transfers over the RDMA protocol.
  - Usable by any sockets application
  - Requires SDP on both sides of the wire.

Multifunction Networking Technology Timeline

10Gb Ethernet Adoption

- Current 10Gb Ethernet adapters are expensive
- CX4 will lower the cost a bit
- Lower prices require higher volumes
- Volumes will be limited until customers can
  - ‘future-proof’ (using RJ45 connector)
  - at a reasonable cost (10x perf for 3x cost)
  - blade servers may be exception
- 10GBASE-T is not easy
  - Pressing against the cable theoretical limits
  - Affordable and low-power interface may be a while

10Gb Ethernet Timeline



Boards

MAC with 10GBASE-SR Transceiver

Integrated MAC/PHY w/ 10GBASE-CX4

MAC w/ 10GBASE-T external PHY

Integrated MAC/PHY w/ 10GBASE-T

MAC/PHY Chips

Integrated MAC/PHY w/ 10GBASE-CX4

Integrated MAC/PHY w/ 10GBASE-T

PHY Chips

Gen1 10GBASE-CX4 PHYs

Gen1 10GBASE-T PHYs

Gen2 10GBASE-T PHYs

IEEE Stds.

10Gb Ethernet Standard (SR, LX4, LR)

10GBASE-CX4 Standard

10GBASE-T Standard

H1 H2 H1 H2 H1 H2 H1 H2 H1 H2 H1 H2 H1 H2 H1 H2

2002 2003 2004 2005 2006 2007 2008 2009

Ethernet or Infiniband?

Fabric	Strengths	Weaknesses	Outlook
Ethernet	Ubiquitous; Standard Affordable adapters and switches. Minimal training costs. Extends beyond the datacenter. Mature foundation	~~CPU utilization~~ ~~Memory bandwidth utilization~~ ~~Latency~~ 10Gb cost	Already the default choice for most applications. Overcoming its weaknesses (with TOE, RDMA, iSCSI, and CX4)
Infiniband	Lowest latency. Most affordable 10Gb link (today).	Expensive adapters, switches. Unique fabric and fabric management. Bridges needed to go to storage and Ethernet fabrics.	Good in niche applications where performance advantages outweigh disadvantages.

Why use offload with 1Gb NICs?

- Some applications now need more than 1Gb
- Multiple 1Gb ports are more affordable than 10Gb
- 1Gb Ethernet interfaces need offload to scale
  - but must retain cost advantages
- Many applications need reduced latency
  - Scale out applications (e.g., HPC)
  - Controlled latency across data center

Affordable offload

- TOE
  - Parts now becoming available without external RAM
  - Microsoft has announced timeline for OS support
- RDMA
  - Standards completed in 2003
  - Some vendors have announced components & boards
- iSCSI
  - Software initiators available for Windows/Linux
  - Release of iSCSI targets slowed by availability of affordable TOE

Conclusion

Multifunction Networking

Affordable TOE

Affordable iSCSI Initiators

Low-latency RDMA

Scalability with RSS

Realistic Multifunction Network

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