Chelsio Communications

The State-of-the-Art of TOE Technology

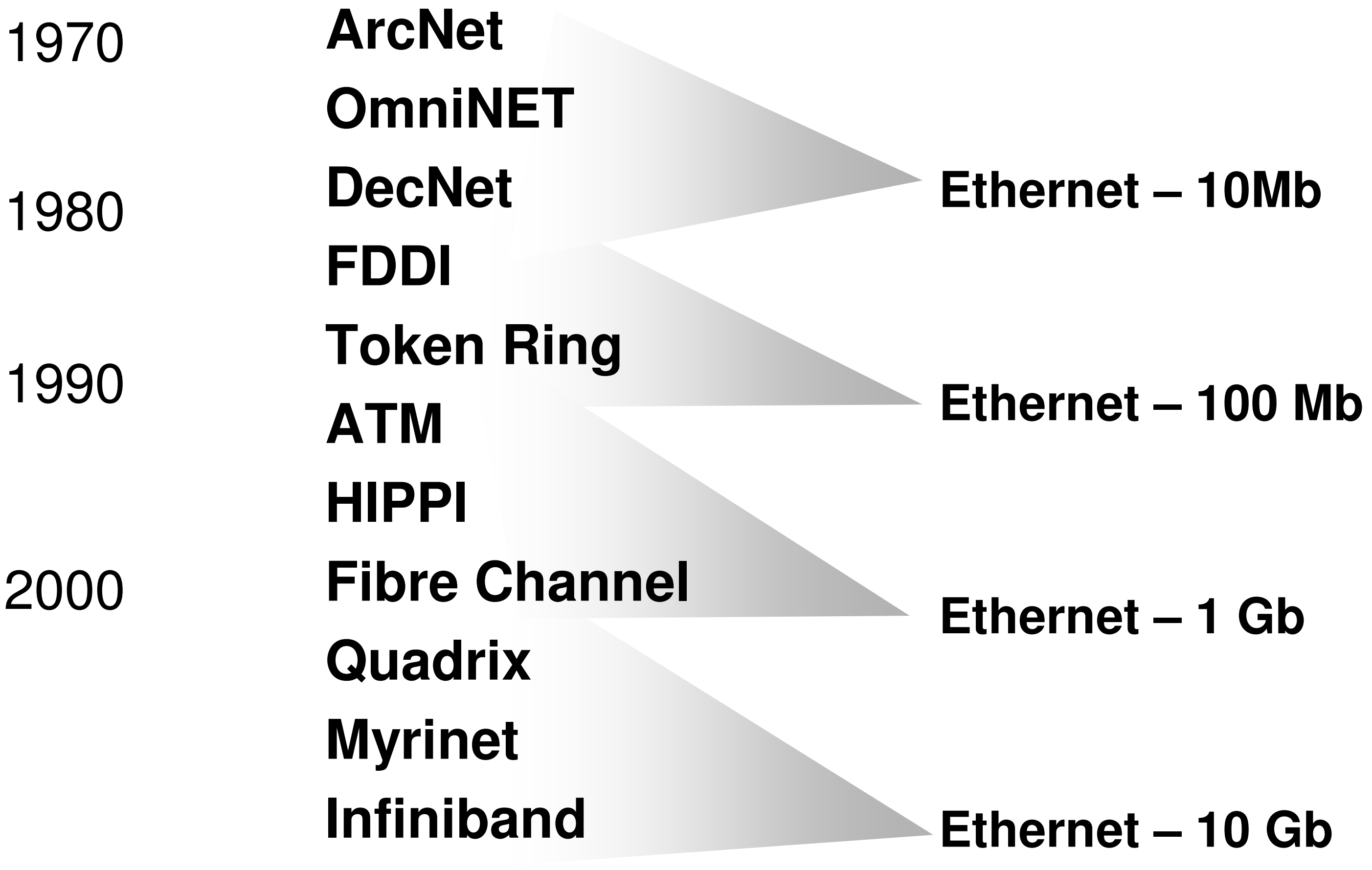
Michael Chen, PhD

PFLDnet2006 Presentation, Feb. 2006

Agenda

- Technology Trend
- 10GbE TOE Architecture
- TOE Support of LFN
- 10GbE TOE Performance
- Network Convergence and ULP Acceleration

Ethernet’s History of Absorbing Proprietary Networking Technologies



1970

1980

1990

2000

ArcNet

OmniNET

DecNet

FDDI

Token Ring

ATM

HIPPI

Fibre Channel

Quadrix

Myrinet

Infiniband

Ethernet – 10Mb

Ethernet – 100 Mb

Ethernet – 1 Gb

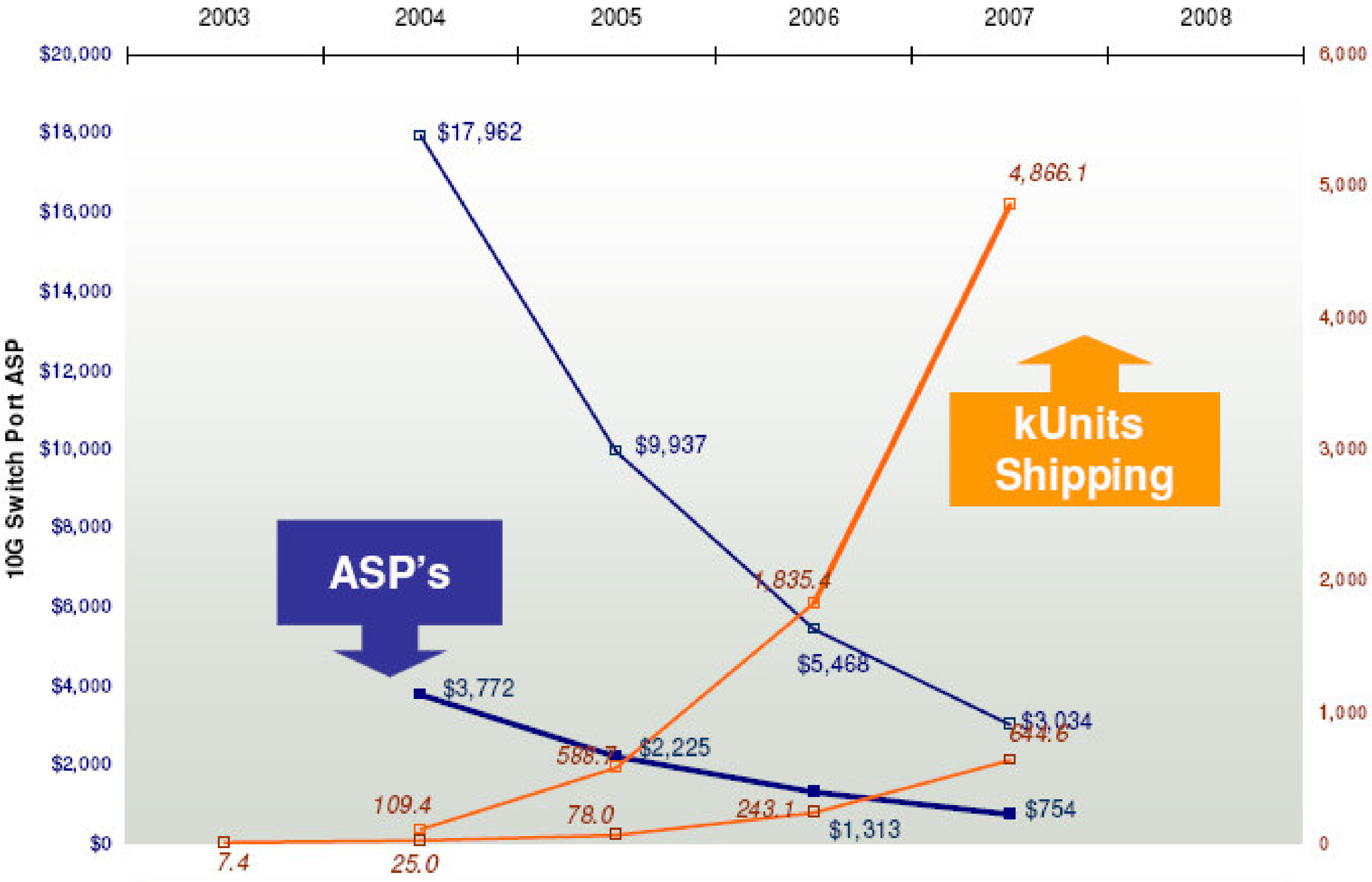
Ethernet – 10 Gb

10G Ready for Prime Time

Criteria	*Market Drivers / Enablers*
Units growth	▪ 3x volume growth in the 10GbE NIC market in 2006 (synergy report) ▪ 3x volume growth in 10GbE switch market in 2006 (Dell'Oro) ▪ iSCSI market growing at 50% per quarter ▪ 55% of top 500 HPC installations are Gb Ethernet
Infrastructure	▪ High density 10GbE optical switches available now ▪ High density 10GbE CX4 switches available now ▪ Low latency switch chips available from at least 4 vendors ▪ UTP PHY chips available from at least 3 vendors
Prices	▪ XFP over 12 months has dropped by more than 100% ▪ CX4 switch port at $700/port list price now ▪ CX4 adapter pricing dropping past the knee ▪ 10GbE HBA pricing has been halving every 12 months
Standards	▪ 10G CX4 (copper media) introduced and shipping ▪ 10G-baseT silicon expected by year-end

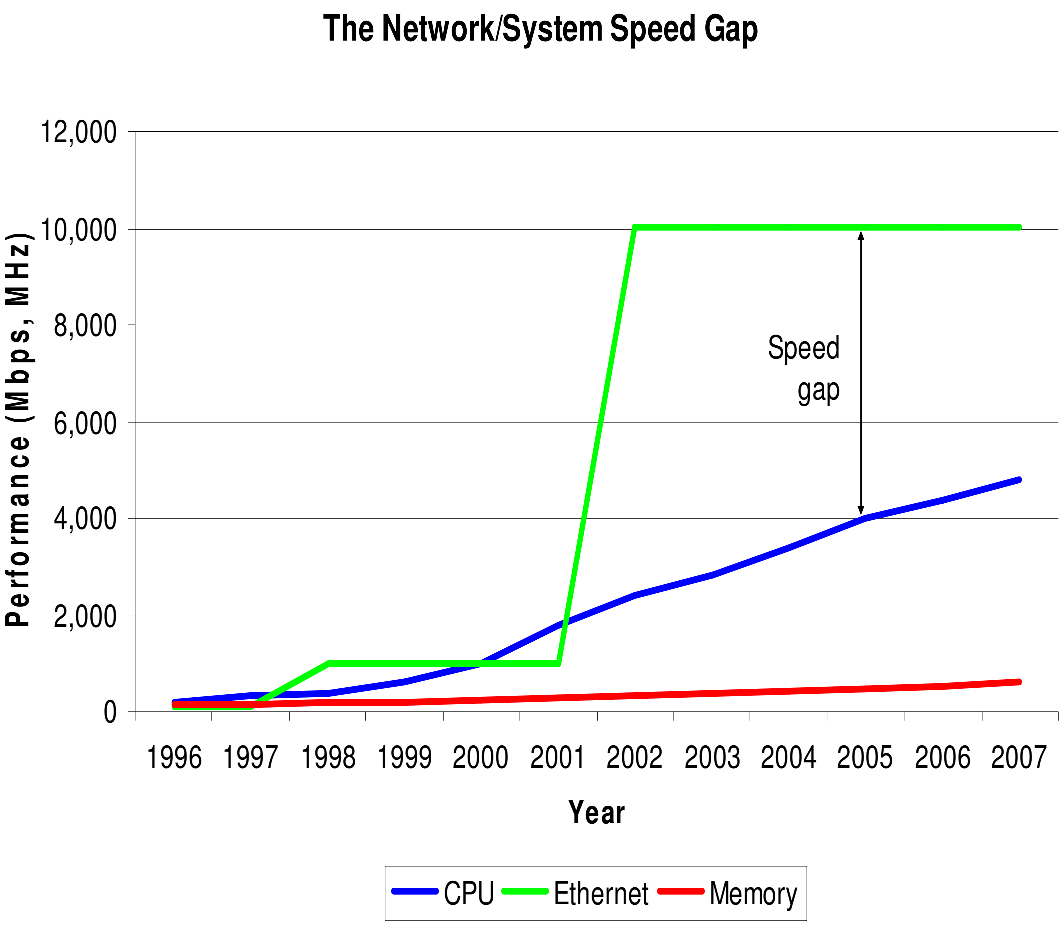
10GbE is Beating Forecast

Chelsio Communications
Accelerate



With CX4, HBA prices are at 2007 levels

The Speed Gap



- **Rule-of-thumb: 1GHz of CPU needed to process 1Gbps data rates**
- **At 10Gbps, today's highest performance CPUs lag by 2.5x**
- **In 2006, 10GbE full-duplex (20Gbps) will further widen the gap**
- **Memory speeds lag even further behind and will become main obstacle in the future**
- **The SOLUTION is Protocol Offload**

Ethernet Popular HPC Deployments

2003 Top Supercomputer Cluster Interconnects

	Clusters	%	Servers	%	Avg Server/Cluster
Ethernet	88	55%	15,112	53%	172
Myrinet	57	35%	8,890	31%	156
InfiniBand	3	2%	1,484	5%	495
Quadrics	9	6%	2,608	9%	290
SCI	4	2%	310	1%	78
Total	161	100%	28,404	100%	176

Source Top500.org Nov 2003 (161 new clustered systems were added to the list)

Ethernet is the *dominant*
High Performance Cluster Interconnect
Today!

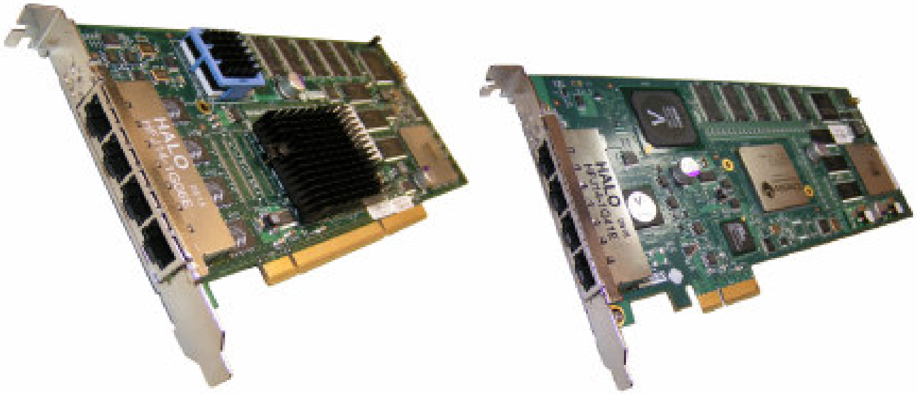
Chelsio Product Family

Chelsio
Communications
Accelerate

N210: 10GbE Server Adapters

T210: 10GbE Protocol Engines – Fiber & Copper
Server Adapter + TCP + iSCSI + RDMA

T204: 4-port 1GbE Protocol Engines



Protocol software and drivers

10GbE PHY Technologies

- **Fiber**

– 10GBase-SR	85 m	shipping
– 10GBase-LR	10,000 m	shipping

- **Copper**

– 10GBase-CX4	15 m	shipping
– 10GBase-T	55-100 m	

- **Backplane**

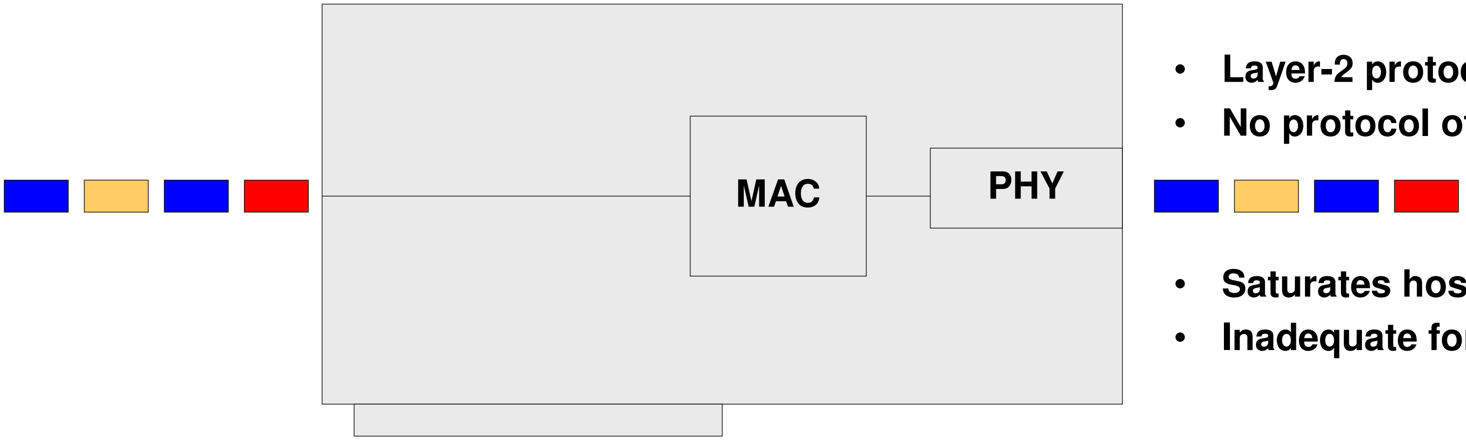
– 10GBase-KX4	0.5-1 m	
– 10GBase-KR	0.5-1 m	

Chelsio Communications

TOE Architecture

Alternative 10G Solutions

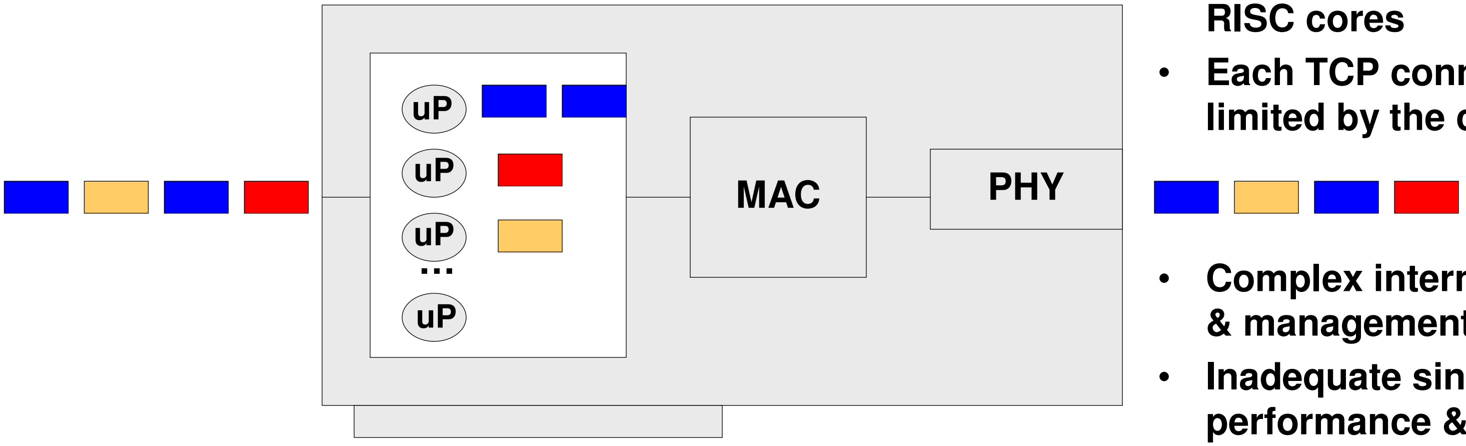
Basic 10G NIC



- Layer-2 protocols only
- No protocol offload intelligence

- Saturates host CPU
- Inadequate for high-performance

Multi-RISC based Architecture



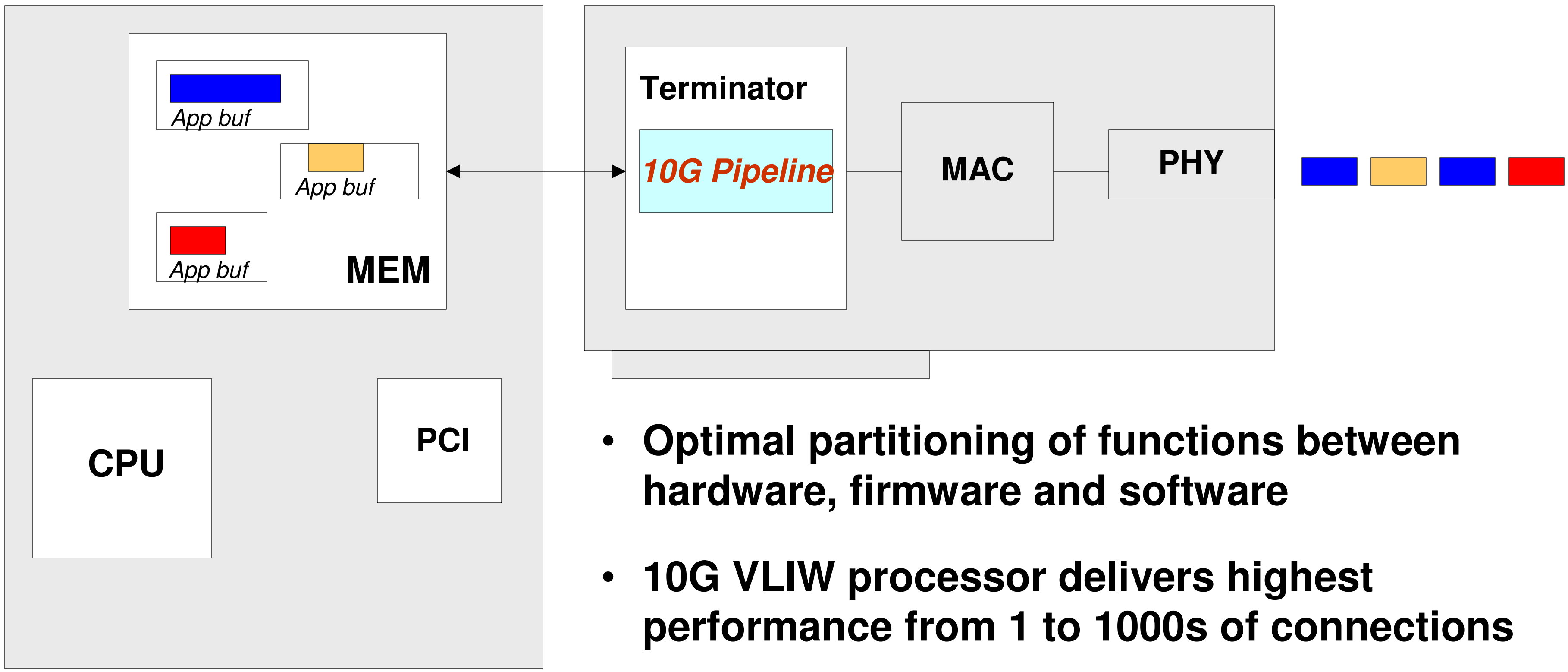
- Offload engine consists of multiple RISC cores
- Each TCP connection’s bandwidth limited by the core frequency

- Complex internal software for control & management of multiple cores
- Inadequate single channel performance & scalability

Chelsio's Unique Architecture

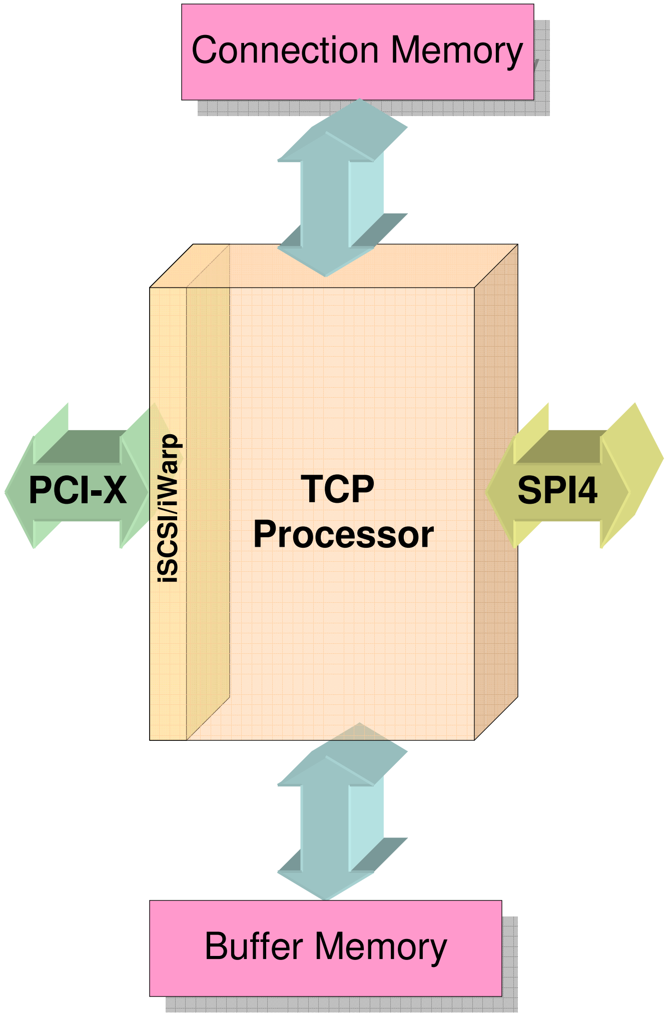
Host

Chelsio Terminator Architecture



- Optimal partitioning of functions between hardware, firmware and software
- 10G VLIW processor delivers highest performance from 1 to 1000s of connections
- Pipelined architecture uses cut-through processing for low latency
- Direct data placement into application buffers eliminates copy overhead

“Terminator” Processor ASIC



- Cut-through, wire-speed architecture
- Scalable from 10G to 1G line speeds
- TCP, iSCSI, RDMA, DDP acceleration
- 400+ configuration registers
- Programmable TCP rules per connection

> L4	*iSCSI*	*iWARP*	*DDP*		
L4	UDP/TCP Checksum Offload	Full TCP Protocol Conformance	Per-Connection, per-Server Control		
		Congestion Avoidance and Control	High Performance Extensions	Increased Initial Window Size	
		TCP Traffic Filtering	SYN Cookies	Wildcard Sockets	
IP	IP Checksu\m Offload	Route Table Cache	QoS, DiffServ, DSCP	ARP Cache	Path MTU Discovery
MAC	Multiple MAC Address per Port	Jumbo Frames	VLAN Tagging *802.1q*	Prioritization *802.1p*	Ethernet-II and 802.3 Frames
Ether	1 Gbps, 10 Gbps Ethernet	Copper / LR and SR Fiber			

Accelerated

Offloaded

10GE TCP Processing

- **Current generation NPU not good match for 10GE TCP processing**

- **TCP at 10Gb characteristics and requirements**
  - Stateful protocol -> efficient RMW
  - Large number of connections -> scalable architecture
  - Jumbled byte stream -> intelligent memory system
  - < 10us latency requirement -> cut-through processing
  - Still an evolving protocol -> programmability

10GE VLIW TCP Processor Innovation

- **The TCP protocol is stateful**
  - At 10Gbps, 1500 byte packets are 1us apart
  - TCP state has poor cache locality
  - Wire speed needs to be attainable for 1 connection
  - Wire speed needs to be attainable for 1000s connections
  - Requires
    - An efficient pipelined pre-fetch of the TCP state
    - Single engine that can process 10Gbps traffic

10GE VLIW TCP Processor Innovations

- **The TCP protocol provides a FIFO stream abstraction to the end points**
  - Packets can partially overlap and/or arrive out of order
  - Latency requirement rules out store-and-forward
  - Requires
    - Specialized memory subsystem to unravel the packet jumble @10Gbps speeds

10GE VLIW TCP Processor Innovations

- **There are stringent end-to-end latency requirements (< 10us) in addition to high BW requirement (10Gbps)**
  - This requires cut-through processing
  - Cut-through processing refers to the packet arriving on one terminal, being processed and the being forwarded out the other terminal without ever being stored in off-chip memory
- **Measured end-to-end latency < 10us with L2 switch**
  - DMA engine used interrupts but could push number lower by using polling mode

Optimized Architecture

Chelsio Communications
Accelerate

	10G VLIW	Multi-RISC
Scalability	Unlimited	Limited by # of CPUs
Firmware Complexity	Low	Typically 1+ year firmware debug
Cache Capacity	Unlimited	Limits maximum # of accelerated connections
Performance Profile	Linear, uniform bandwidth per connection	Falls off once IPC becomes significant
Roadmap	Low-risk upgrade path	Complex firmware more difficult to scale

Chelsio Communications

TOE Support of LFN

Congestion Control in LFN

- **RFC 3649: Highspeed TCP**
- **In Congestion Avoidance,**
  - for each ACK, increase the window by
    - $w = w + a(w) / w$

    Note: in standard TCP

    a(w)=1 // when w is maintained by #segs

    a(w) = mss * mss // when w is maintained by bytes
  - For each congestion event, decrease the window by
    - $w = (1-b(w)) * w$, // where $0 < b(w) <= 0.5$

Congestion Control in LFN

- **Table-driven Implementation**
  - For each ACK received, using current ***w*** as index to lookup a table for *a(w)*
  - For each congestion event, using current ***w*** as index to lookup a table for *b(w)*
  - The lookup table is SW configurable which provides the max flexibility for various LFN environments.

AB-Table

W	***A(w)***	***B(w)***
< 38	**1**	**0.5**
38 (56k)	**1**	**0.5**
118 (172k)	**2**	**0.44**
221 (322k)	**4**	**0.41**
…	**…**	**…**
5610 (11M)	**21**	**0.24**
…	**…**	**…**
83000 (120M)	**70**	**0.09**
…	**…**	**…**

Traffic Pacing and Shaping

- Researches [Hiraki-SC04, etc] indicated the importance of pacing TCP streams across LFN to reduce the traffic burstness

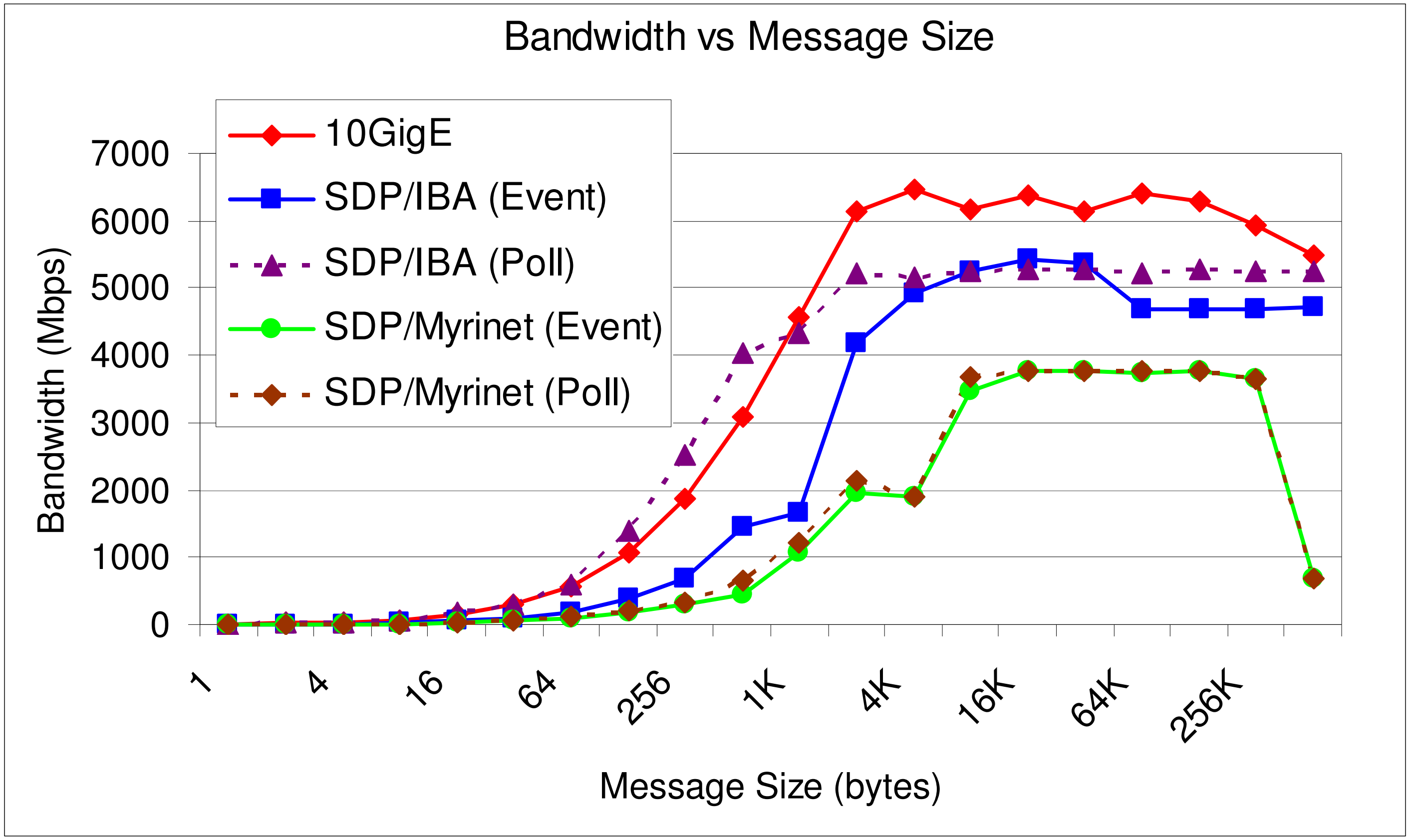
- SW traffic pacing/shaping at 10Gb rate is CPU intensive

- TOE enables HW traffic pacing at TCP level

Chelsio Communications

TOE Performance

OSU/LANL HPC Benchmarks 10GbE TOE vs IB & Myrinet

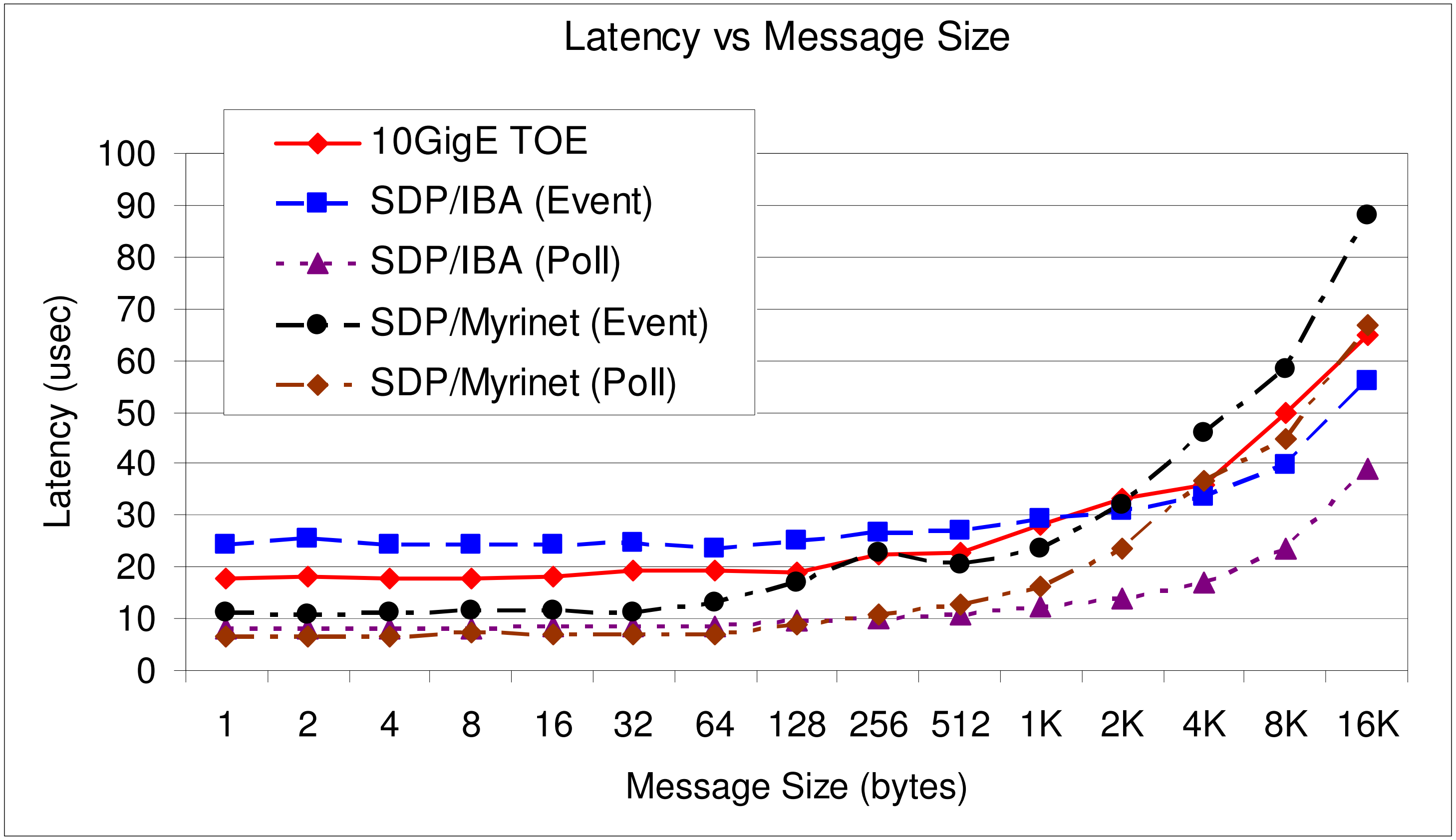


<u>Source</u>: ***Head-to-TOE Evaluation of High-Performance Sockets over Protocol Offload Engines*** **by DK Panda et al**

<u>Test configuration</u>: 4-node cluster connected through 10GbE switch running single connection

<u>System configuration</u>: Dual 32-bit Intel Xeon 3.0GHz processors running Red Hat 9.0 Linux kernel 2.4.25smp

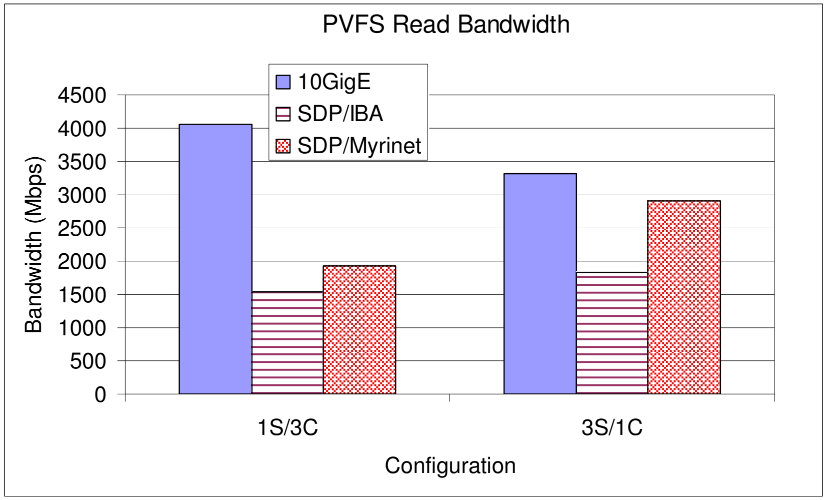
OSU/LANL Benchmarks
10GbE TOE vs IB & Myrinet

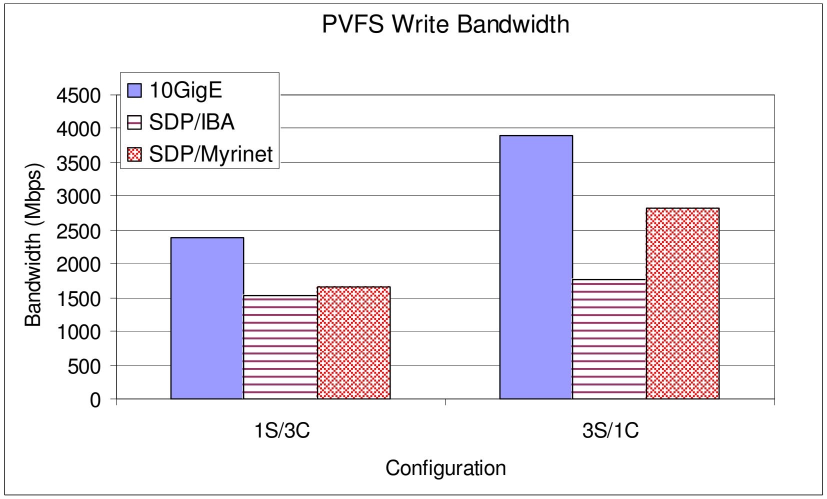


Source: ***Head-to-TOE Evaluation of High-Performance Sockets over Protocol Offload Engines*** **by DK Panda et al**
Test configuration: **4-node cluster connected through 10GbE switch running single connection**
System configuration: **Dual 32-bit Intel Xeon 3.0GHz processors running Red Hat 9.0 Linux kernel 2.4.25smp**

OSU/LANL Benchmarks
10GbE TOE vs IB & Myrinet

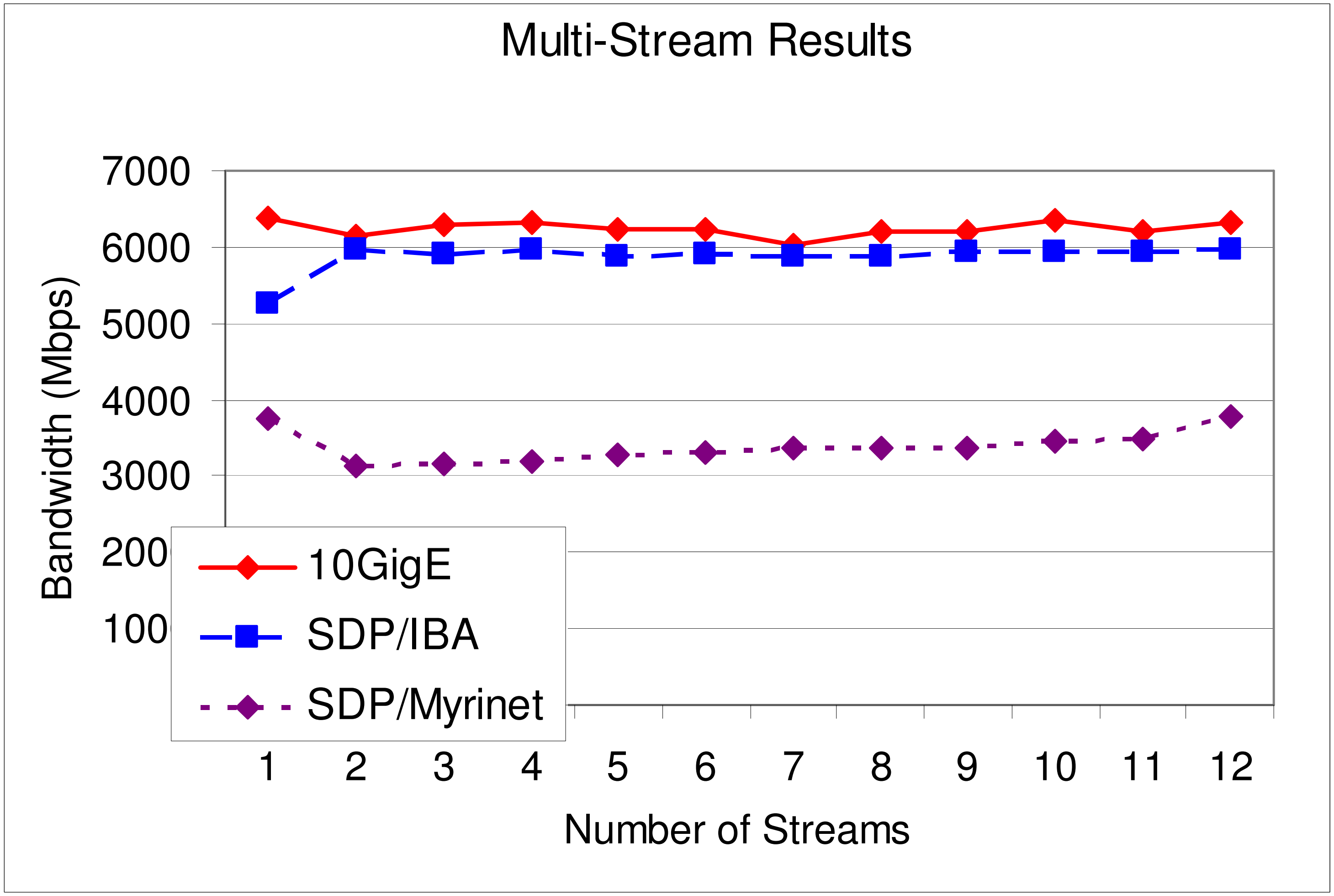
- Parallel Virtual File System (PVFS) – concurrent read/write performance





Source: ***Head-to-TOE Evaluation of High-Performance Sockets over Protocol Offload Engines*** **by DK Panda et al**
Test configuration: 4-node cluster connected through 10GbE switch running single connection
System configuration: Dual 32-bit Intel Xeon 3.0GHz processors running Red Hat 9.0 Linux kernel 2.4.25smp

OSU/LANL Benchmarks
10GbE TOE vs IB & Myrinet



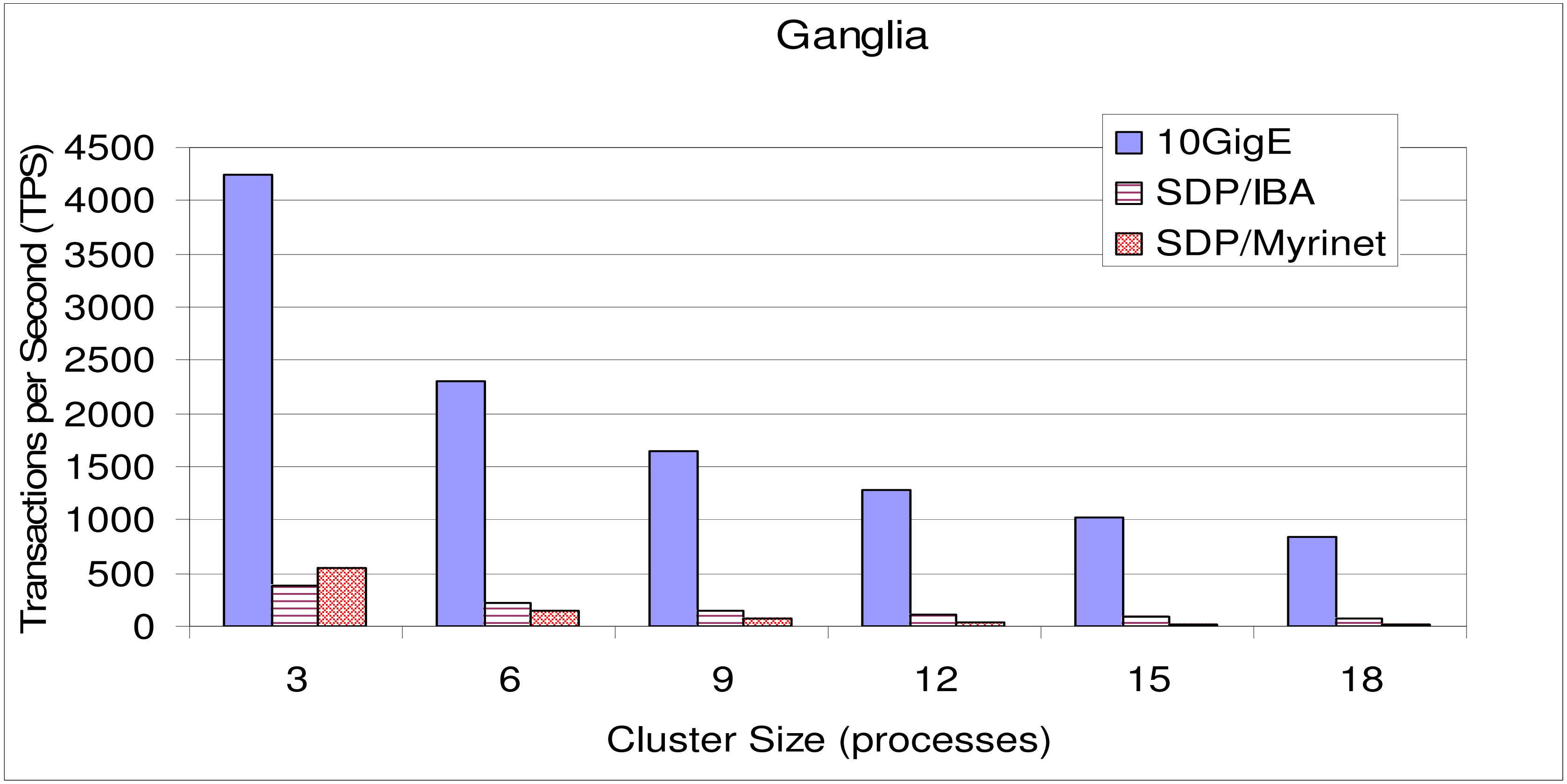
Source: ***Head-to-TOE Evaluation of High-Performance Sockets over Protocol Offload Engines*** **by DK Panda et al**

Test configuration: **4-node cluster connected through 10GbE switch running single connection**

System configuration: **Dual 32-bit Intel Xeon 3.0GHz processors running Red Hat 9.0 Linux kernel 2.4.25smp**

OSU/LANL Benchmarks
10GbE TOE vs IB & Myrinet

- Ganglia Monitoring

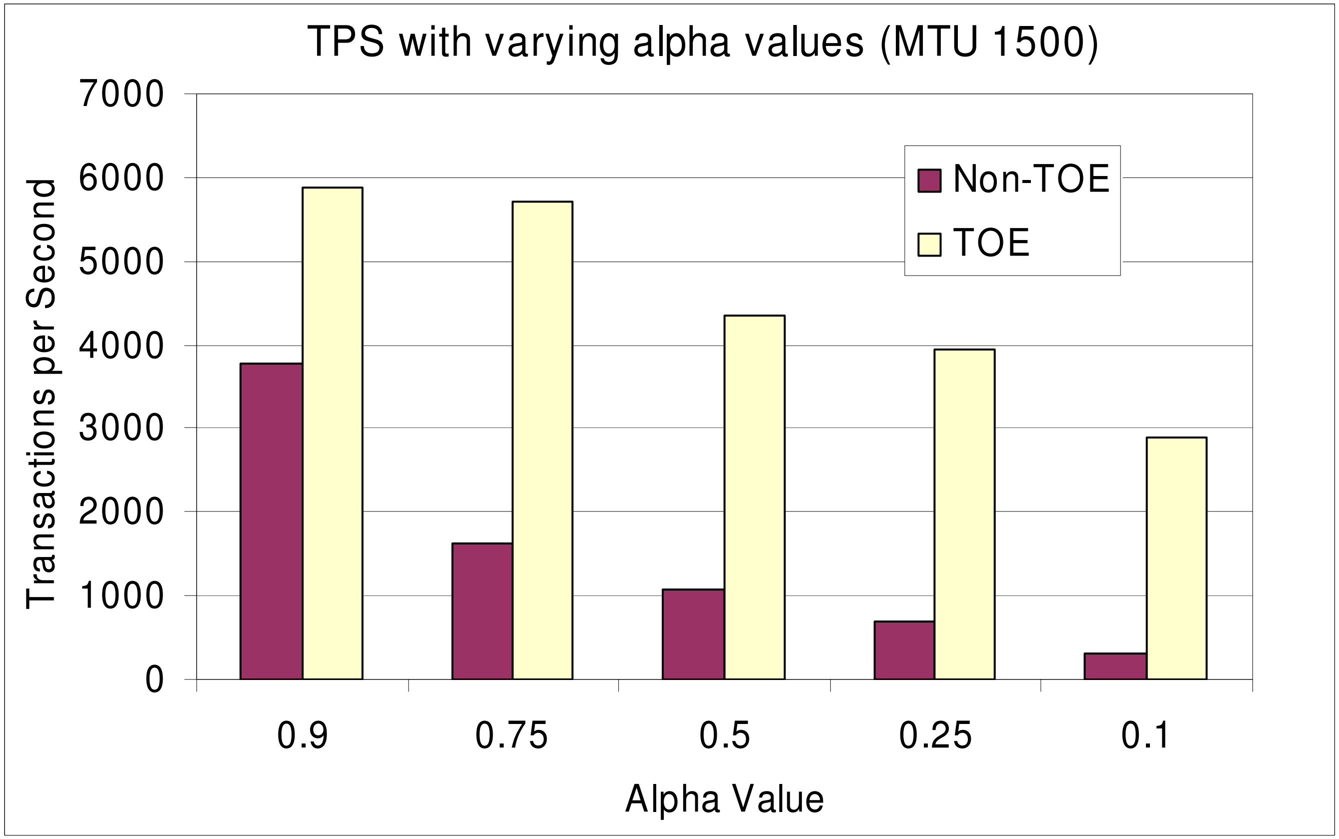


<u>Source</u>: ***Head-to-TOE Evaluation of High-Performance Sockets over Protocol Offload Engines*** **by DK Panda et al**
<u>Test configuration</u>: 4-node cluster connected through 10GbE switch running single connection
<u>System configuration</u>: Dual 32-bit Intel Xeon 3.0GHz processors running Red Hat 9.0 Linux kernel 2.4.25smp

OSU/LANL Benchmarks
10GbE TOE vs 10GbE NIC

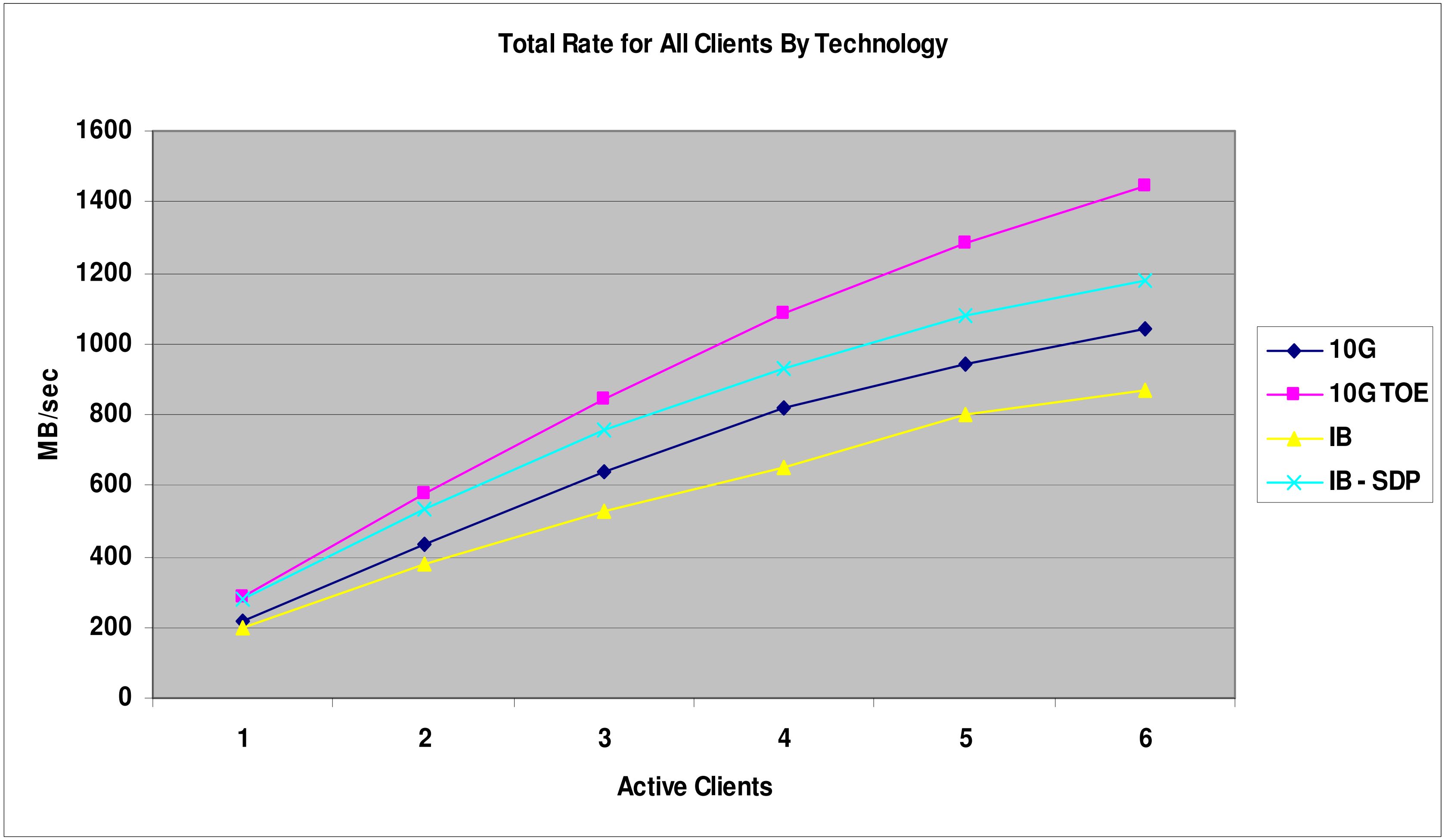
Chelsio Communications Accelerate

- Apache Web Server



Source: ***Performance Characterization of a 10-Gigabit Ethernet TOE*** **by Wu Feng et al**
Test configuration: 4-node cluster connected through 10GbE switch running single connection
System configuration: Quad AMD Opteron 2.0GHz processors running Suse Linux with 2.6.6 stock kernel

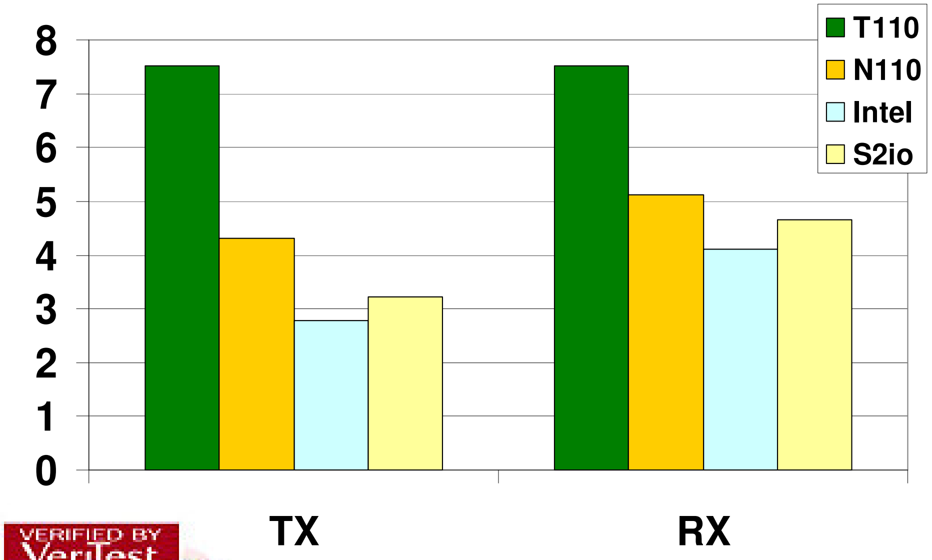
Sandia Benchmarks
10GbE TOE vs IB & 10GbE NIC



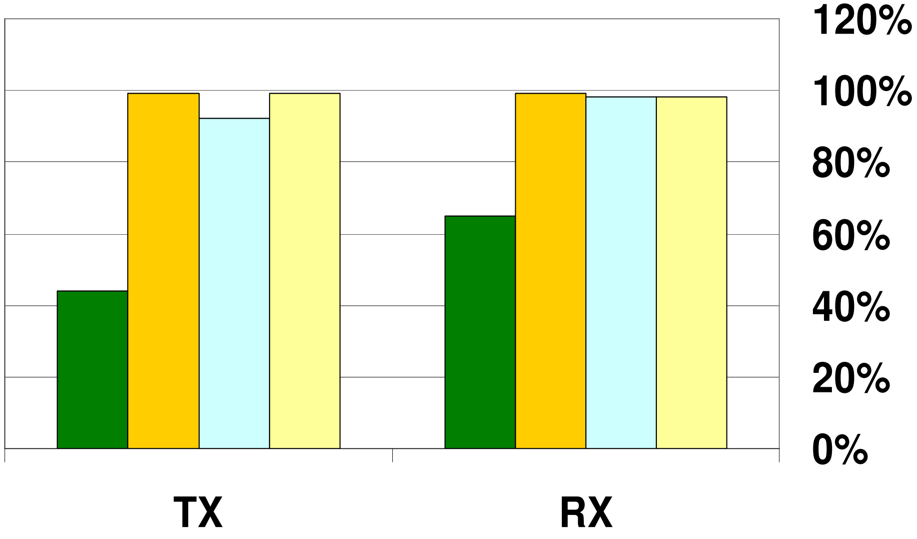
Source: *Infniband and 10-Gigabit Ethernet for I/O in Cluster Computing* by Helen Chen et al
Test configuration: 8-node cluster connected through 10GbE switch running IOzone
System configuration: Dual AMD Opteron 2.2GHz processors running Linux kernel 2.4.25smp

Chelsio Competitive Advantage

Higher Throughput (Gb/s)



Lower CPU Utilization (%)



VERIFIED BY VeriTest.
Testing by the World's Leading Independent Lab

<ins>Source</ins>: Independently verified by VeriTest, Inc.
<ins>Test tool</ins>: netperf
<ins>Test configuration</ins>: 2 systems connected through 10GbE switch running single TCP channel with *1500-byte Ethernet frames*
<ins>System configuration</ins>: AMD Opteron 248 2.2GHz uniprocessor running Linux kernel 2.6.6

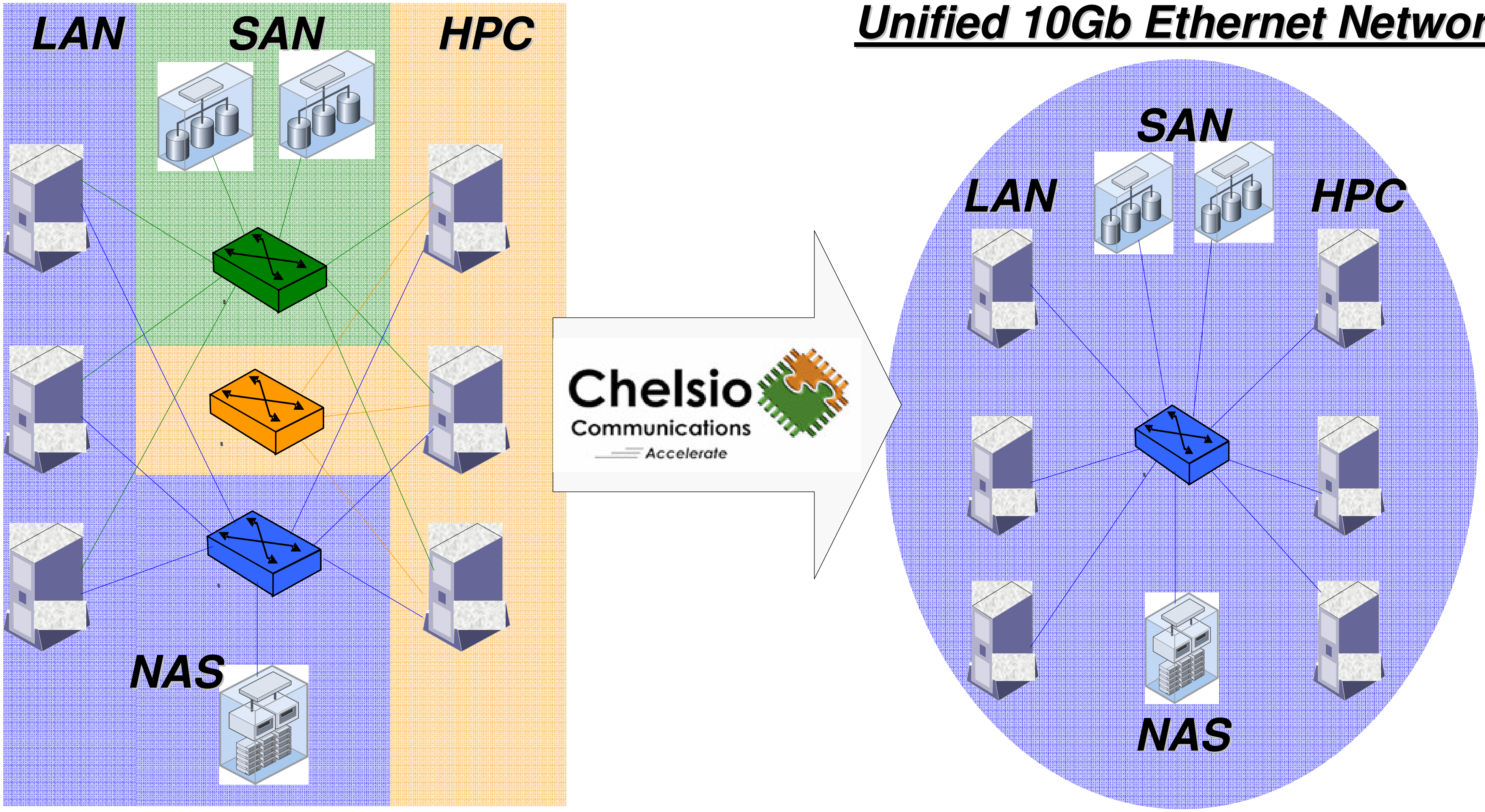
- **T110 achieves 2x network throughput vs basic 10GbE NICs**
- **T110 utilizes only ½ x CPU resources vs basic 10GbE NICs**
- **<ins>RESULT</ins>: T110 delivers <ins>4x performance efficiency</ins> vs NICs**

Chelsio
Communications

Network Convergence & ULP Accelerations

Network Fabric Convergence

Unified 10Gb Ethernet Network

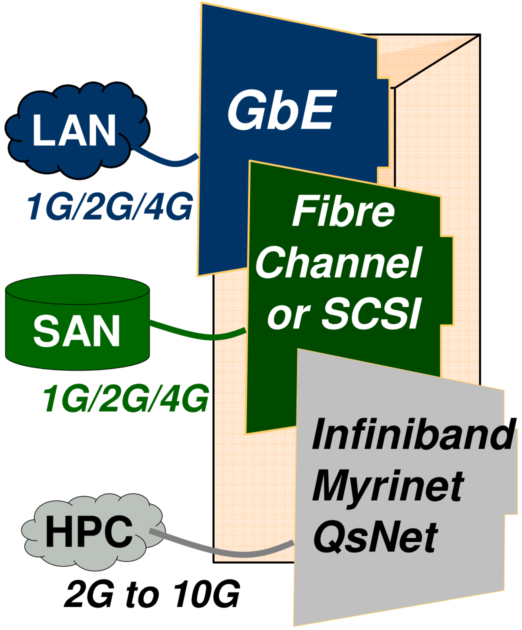


Simplified network architecture – reduced operating costs

Simplified Server Architecture

Chelsio Communications — Accelerate

Current Implementations



Convergence Benefits

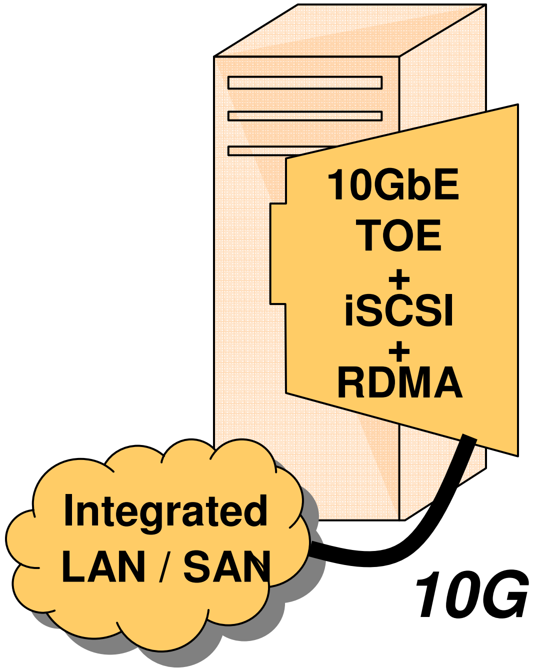
Chelsio's 10GbE Solution

Lower Total Cost of Ownership

- Improves CPU efficiency
- Minimizes software licenses
- Simplifies data center wiring
- Leverages staffing skills & tools

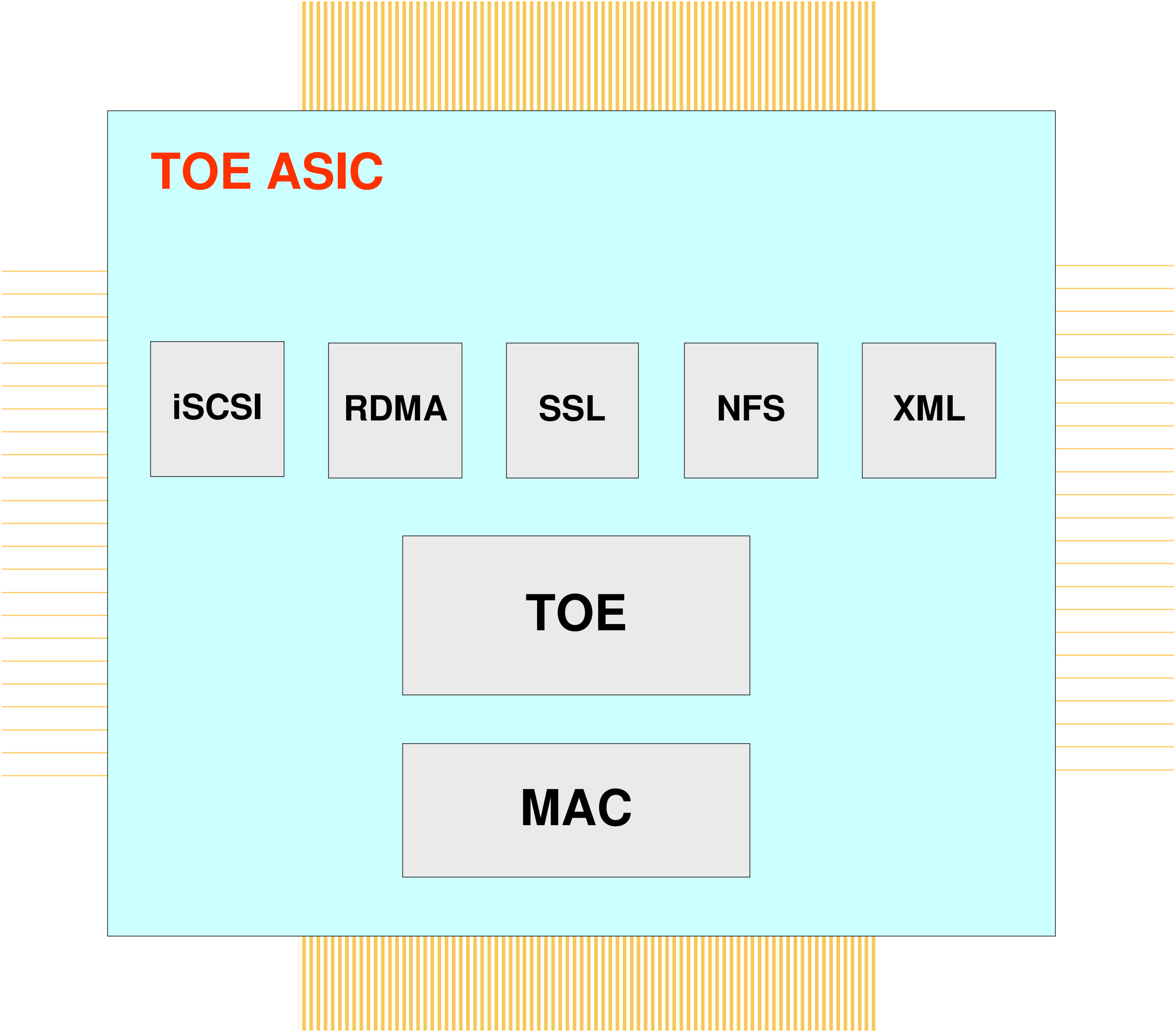
Higher Performance & New Apps

- Improves cluster performance
- Lowers application latency
- Faster backup and recovery
- Enables storage applications

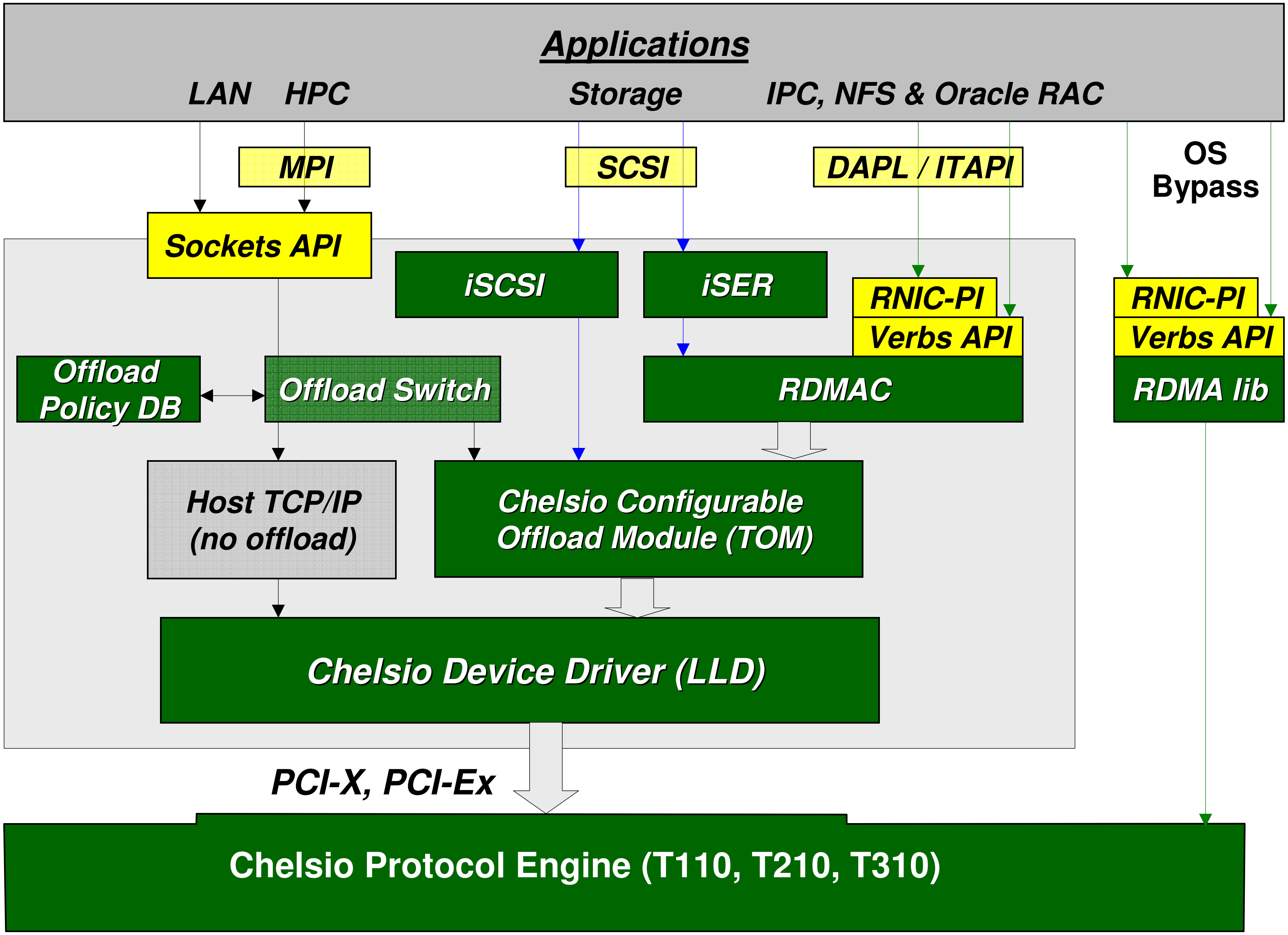


Improved performance – reduced operating costs

TOE Enables ULP Acceleration



Software Architecture

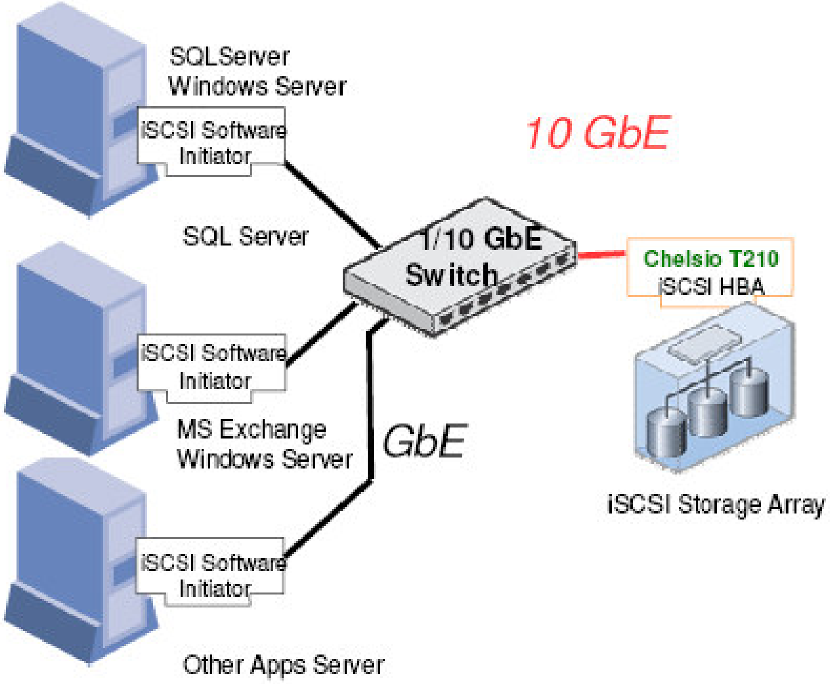


Chelsio Communications

iSCSI over TOE

10GbE Best Suited for Storage

Chelsio Communications
Accelerate

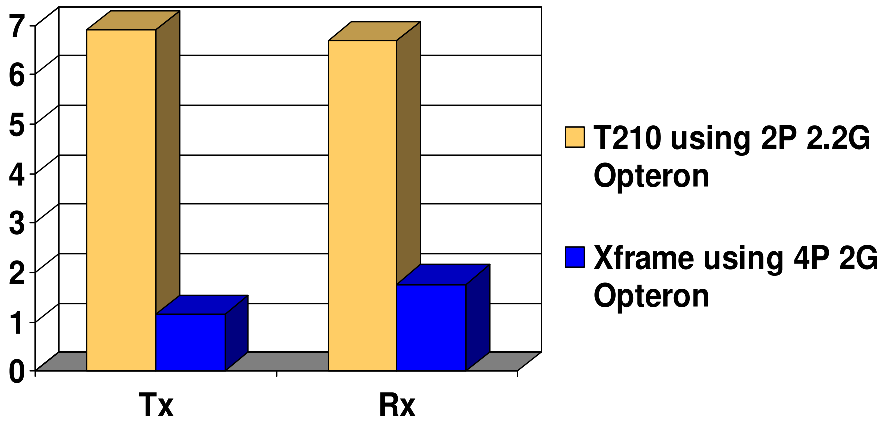


- Free client scalability
  - Free software initiators
  - Free GbE ports with ALL servers
  - HBAs not required for GbE initiators
  - GbE speed adequate for servers
- Similar Target costs to IB, FC
- No change to existing Apps
- Little change to infrastructure

Chelsio iSCSI Performance vs basic 10GbE NICs

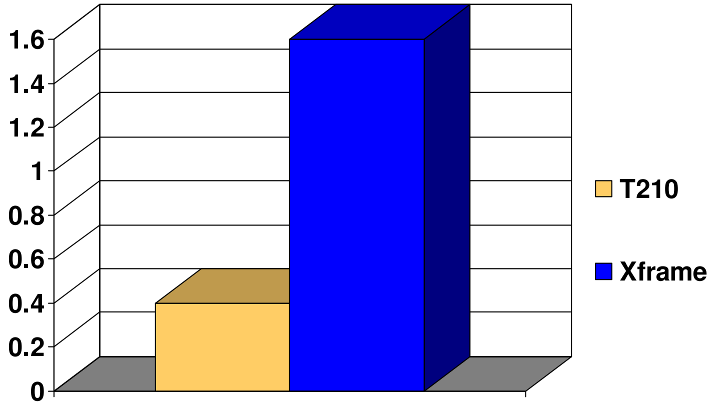
Throughput

Gbps



Average CPU Utilization

x2.2 Ghz Opteron



Sources: T110 iSCSI Performance Analysis by Veritest and Xframe iSCSI Performance Analysis WP Published by Neterion
Note: Charts show performance at 4KB I/O size; iSCSI applications are transactional in nature using 2-4KB I/O sizes

- **T210 achieves 4x iSCSI network throughput vs basic 10GbE NICs**
- **T210 utilizes only ¼ x CPU resources vs basic 10GbE NICs**
- **RESULT: T210 delivers 16x iSCSI performance efficiency vs NICs**

T210 iSCSI Target Performance

	Throughput	Avg. CPU
Read	828MB (TOE)	35%
Write	857MB (TOE)	46%

	IOPS	Avg. CPU
Read	544k (TOE)	88%
Write	539k (TOE)	99%

Target Configuration:

- CPU: 2 x 2.2GHz Opteron
- SW: Linux 2.4.25 and Chelsio Reference iSCSI stack
- IOmeter benchmark
- 28 GbE Microsoft Initiator to one 10GbE Target

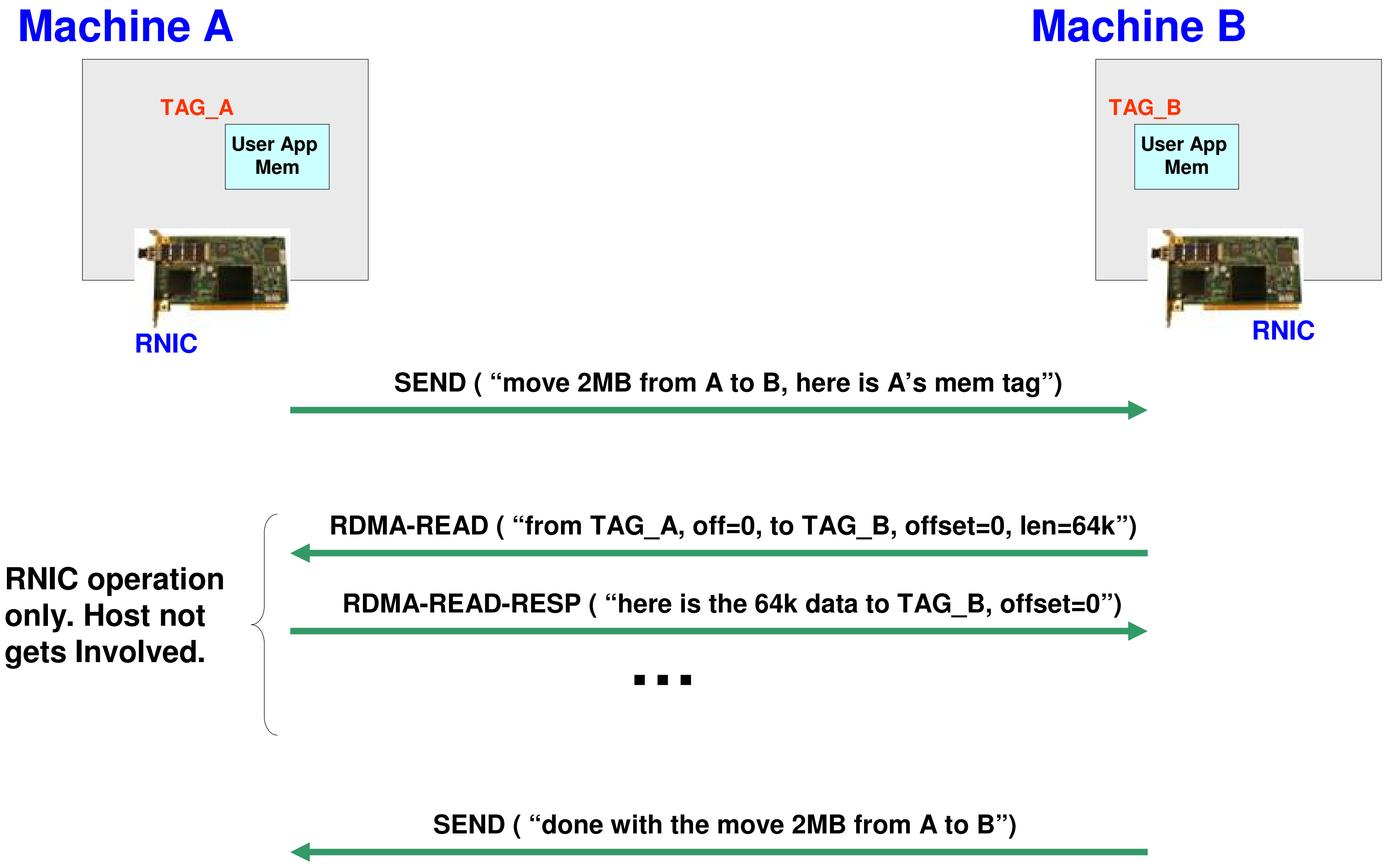
Chelsio
Communications

RDMA over TOE

The Benefit of RDMA

- User space I/O
- OS bypass
- Direct Data Placement (DDP) and zero-copy
- Very low latency
- Very low CPU utilization

RDMA Operations



SEND (“move 2MB from A to B, here is A’s mem tag”)

RNIC operation only. Host not gets Involved.

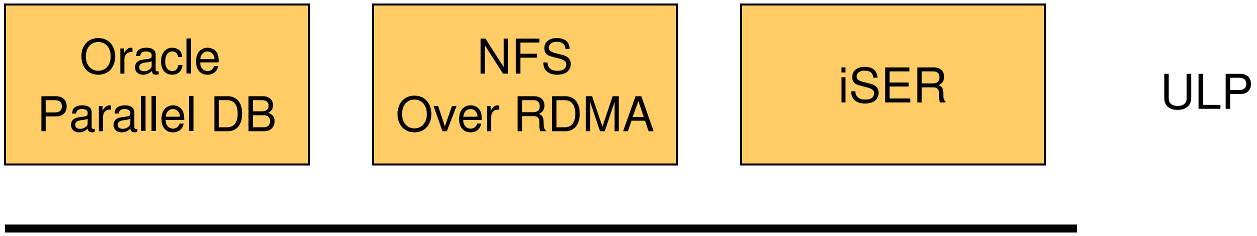
RDMA-READ (“from TAG_A, off=0, to TAG_B, offset=0, len=64k”)

RDMA-READ-RESP (“here is the 64k data to TAG_B, offset=0”)

…

SEND (“done with the move 2MB from A to B”)

RDMA Protocol Stack



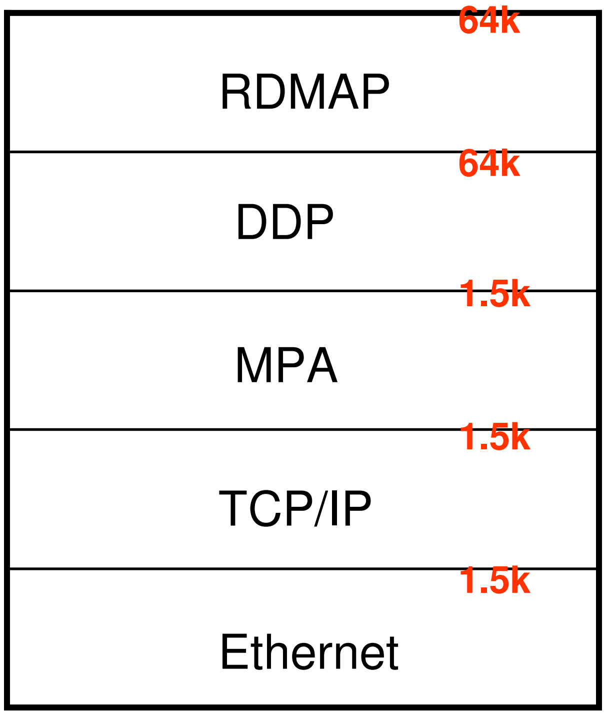
Oracle Parallel DB

NFS Over RDMA

iSER

ULP

RNIC HW



64k

RDMAP

64k

DDP

1.5k

MPA

1.5k

TCP/IP

1.5k

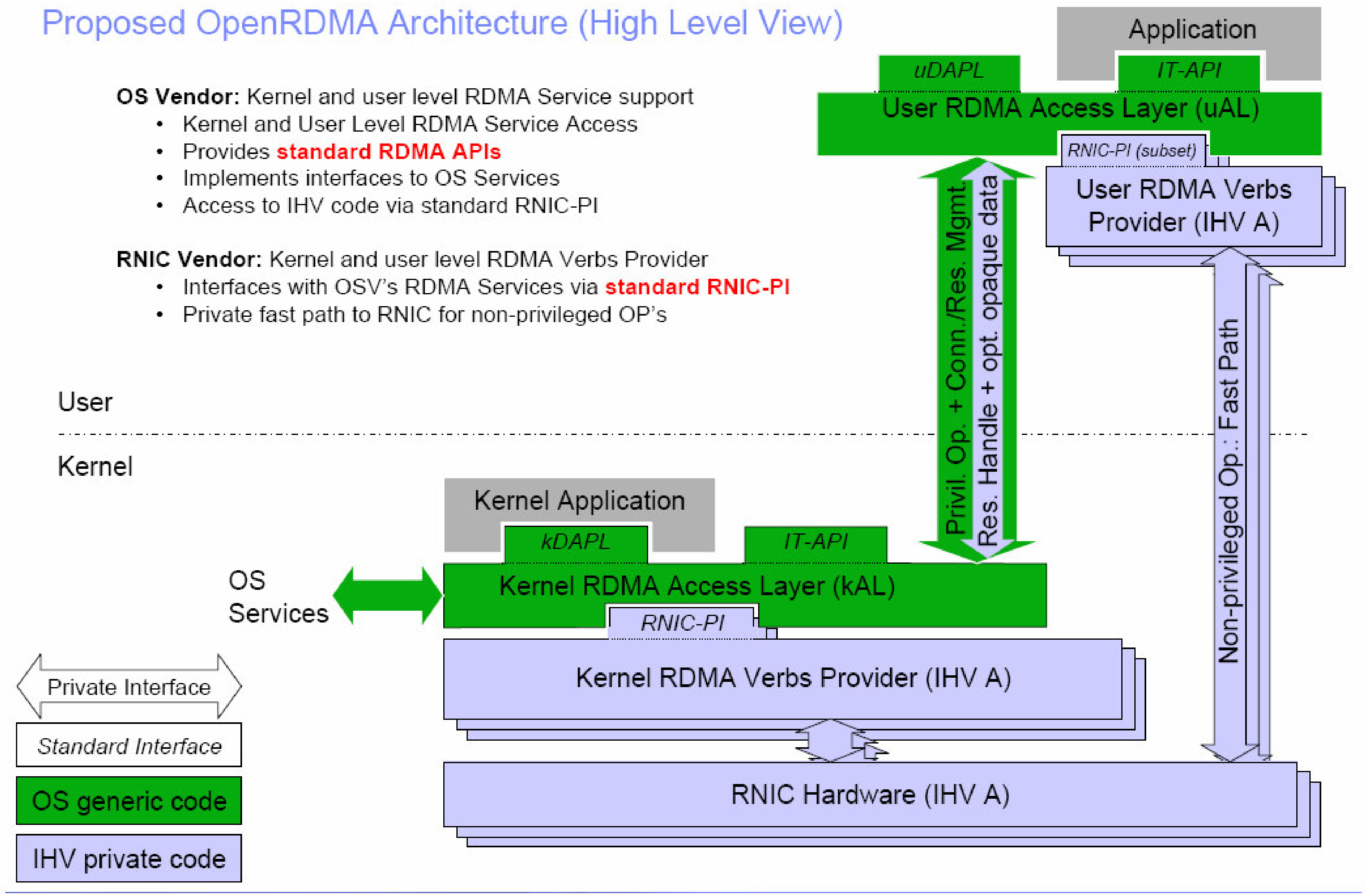
Ethernet

- RDMA Ops:
  RDMA Read, RDMA Read Response
  RDMA Write, Send

- ULP Message segmentation and reassembly
- Out-of-order placement
- In-order delivery

- Framing and CRC
- FPDU aligned with pkt, multiple FPDU in one pkt.
- Marker handling (Start from ISS, every 512B)

RDMA Software Architecture



Proposed OpenRDMA Architecture (High Level View)

OS Vendor: Kernel and user level RDMA Service support

- Kernel and User Level RDMA Service Access
- Provides **standard RDMA APIs**
- Implements interfaces to OS Services
- Access to IHV code via standard RNIC-PI

RNIC Vendor: Kernel and user level RDMA Verbs Provider

- Interfaces with OSV's RDMA Services via **standard RNIC-PI**
- Private fast path to RNIC for non-privileged OP's

Application
uDAPL
IT-API
User RDMA Access Layer (uAL)
RNIC-PI (subset)
User RDMA Verbs Provider (IHV A)
Privil. Op. + Conn./Res. Mgmt.
Res. Handle + opt. opaque data
Non-privileged Op.: Fast Path
User
Kernel
Kernel Application
kDAPL
IT-API
OS Services
Kernel RDMA Access Layer (kAL)
RNIC-PI
Kernel RDMA Verbs Provider (IHV A)
RNIC Hardware (IHV A)
Private Interface
Standard Interface
OS generic code
IHV private code

Thank You!