



STORAGE DEVELOPER CONFERENCE

SNIA ■ SANTA CLARA, 2014

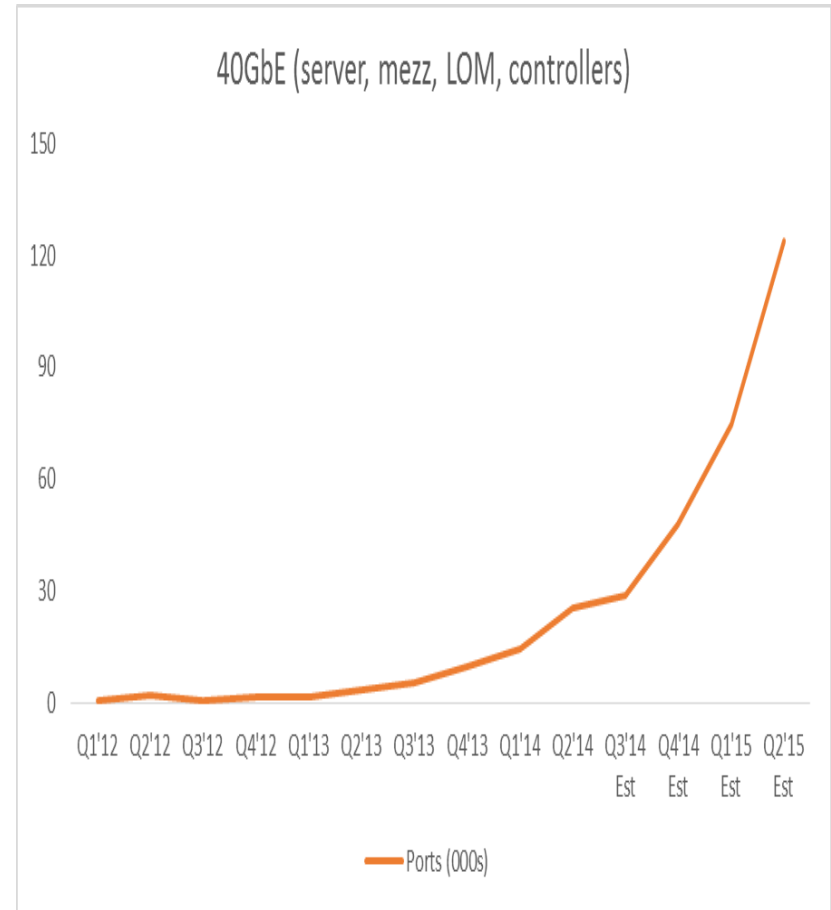
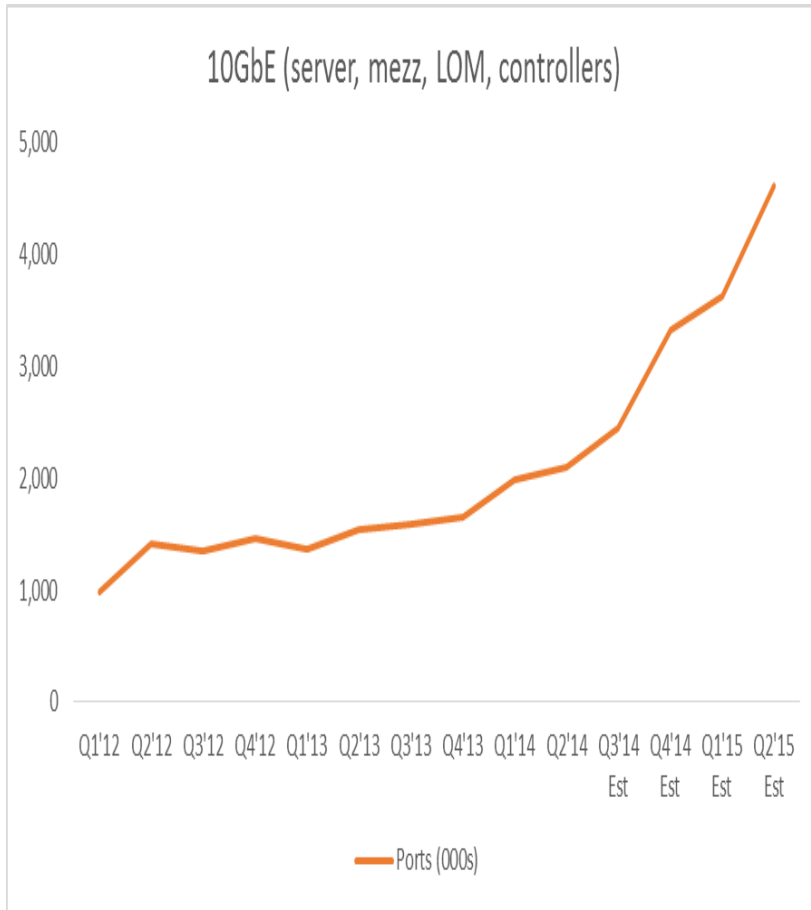
NFS/RDMA over 40Gbps iWARP

Wael Nouredine
Chelsio Communications

Outline

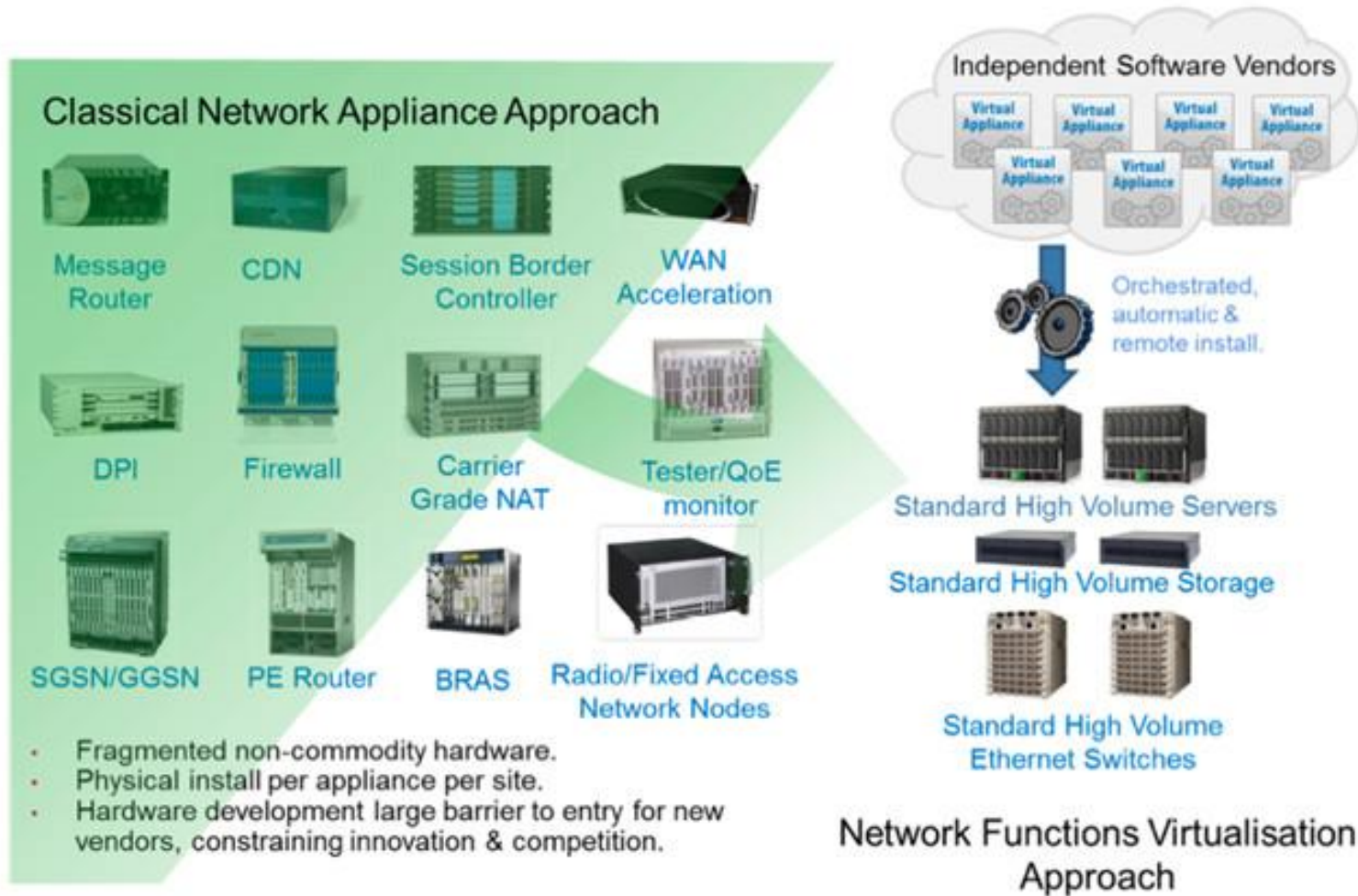
- RDMA
 - Motivating trends
 - iWARP
- NFS over RDMA
 - Overview
 - Chelsio T5 support
 - Performance results

Adoption Rate of 40GbE



Source: Crehan Research - 2Q14 CREHAN Quarterly Market Share Tables

Software Defined Everything



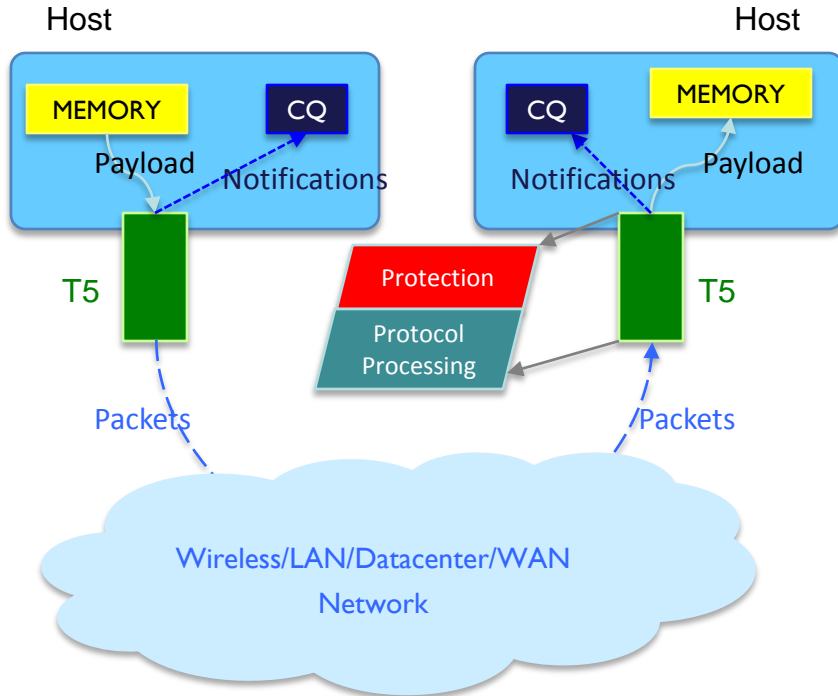
Source: European Telecommunications Standards Institute http://portal.etsi.org/nfv/nfv_white_paper.pdf October, 2012

Motivating Trends

- ❑ Unprecedented curve in 40GbE growth (and pricing)
- ❑ Consolidation and virtualization
 - ❑ Software defined storage (everything) using commodity hardware
 - ❑ Rise of the data center
 - ❑ Power efficiency
- ❑ High speed, ultra low latency SSDs
- ❑ Need for high performance, high efficiency fabric
 - ❑ Ethernet remains the preferred technology
 - ❑ TCP/IP for scalability, reliability, robustness and reach

iWARP RDMA over Ethernet

RDMA Overview



*Performance and efficiency in return
for new communication paradigm*

- ❑ Direct memory-to-memory transfer
- ❑ All protocol processing handled by the NIC
 - ❑ Must be in hardware
- ❑ Protection handled by the NIC
 - ❑ User space access requires both local and remote enforcement
- ❑ Asynchronous communication model
 - ❑ Reduced host involvement
- ❑ Performance
 - ❑ Latency – polling
 - ❑ Throughput
- ❑ Efficiency
 - ❑ Zero copy
 - ❑ Kernel bypass (user space I/O)
 - ❑ CPU bypass

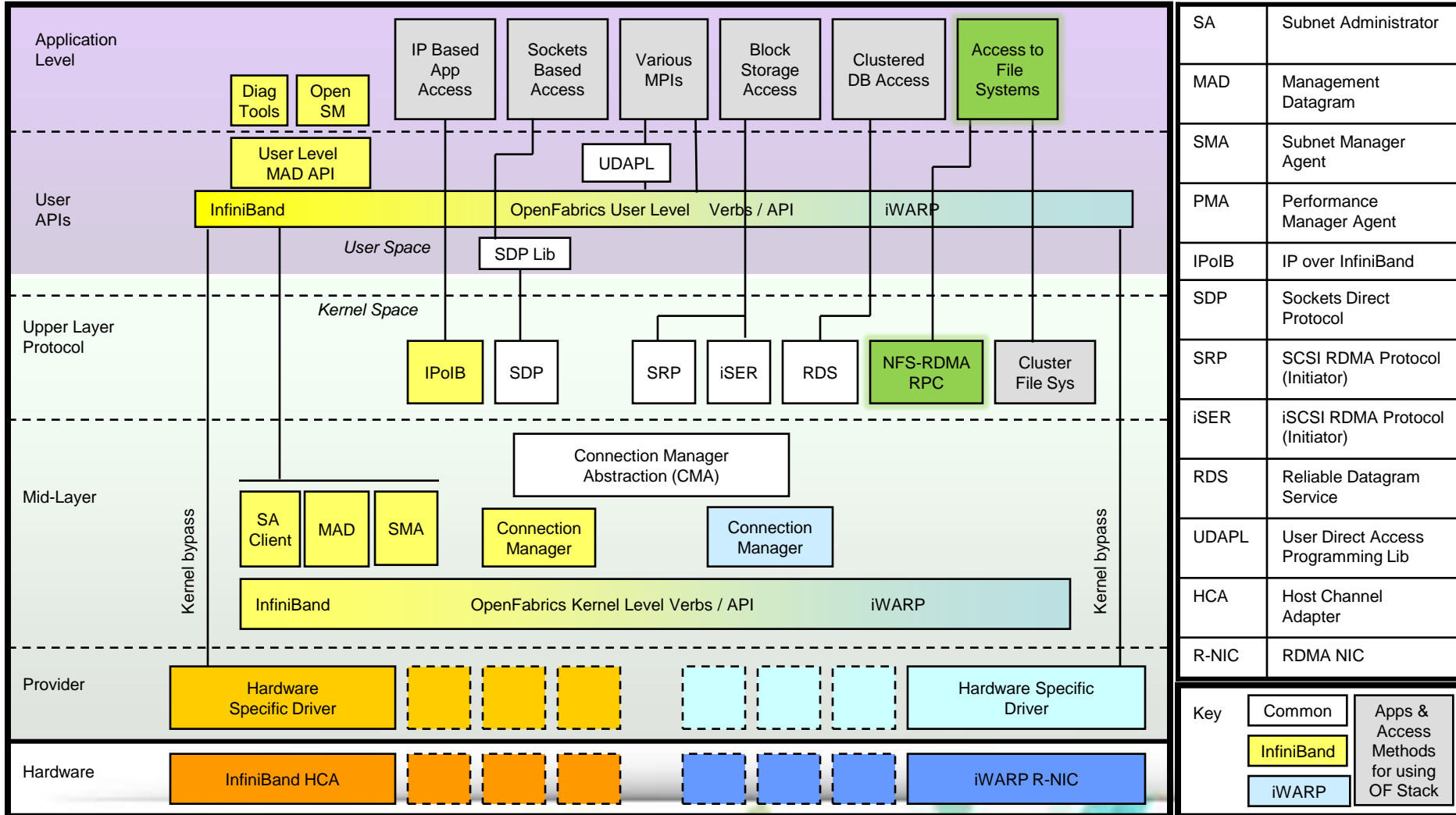
iWARP RDMA over Ethernet

- ❑ IETF RFCs in 2007
 - ❑ Open standard
 - ❑ Multiple vendors
- ❑ Ongoing standardization
 - ❑ Extensions to maintain API uniformity with InfiniBand
 - ❑ Recent RFC 7306 by Broadcom, Chelsio and Intel
- ❑ Mature stack
 - ❑ 3rd generation hardware
- ❑ RDMA over TCP/IP/Ethernet
 - ❑ TCP reliability, scalability, congestion and flow control
 - ❑ IP routability
 - ❑ Ethernet ubiquity
- ❑ Wireless ready
 - ❑ Near 10Gbps, low latency
- ❑ Cloud ready
 - ❑ Standard TCP/IP foundation
 - ❑ No network restrictions
- ❑ Full featured implementation
 - ❑ All RDMA benefits
- ❑ High performance
 - ❑ High packet rate
 - ❑ Low latency (1.5usec user-to-user)
 - ❑ Line rate 40Gb with single connection

iWARP Benefits

- ❑ Convergence
 - ❑ Coexists with all other traffic on same port
 - ❑ No special treatment needed
- ❑ Familiar protocol stack
 - ❑ Standard tools for monitoring/debugging
 - ❑ Standard network function appliances (security, load balancing...)
- ❑ Plug-and-play
 - ❑ No need for lossless network operation
 - ❑ Leverages existing infrastructure
 - ❑ Less expensive network hardware
- ✓ Easy to deploy and manage
- ❑ Leverages decades of TCP/IP experience
 - ❑ Congestion avoidance and control
 - ❑ Critical for network stability
- ❑ Reliability at hardware speeds
 - ❑ Retransmission and re-ordering
- ❑ Routable
 - ❑ Goes wherever IP is spoken
- ❑ Scalable across
 - ❑ Network size
 - ❑ Network architecture
 - ❑ Distance
- ✓ Reliable, robust, scalable

Linux RDMA Architecture



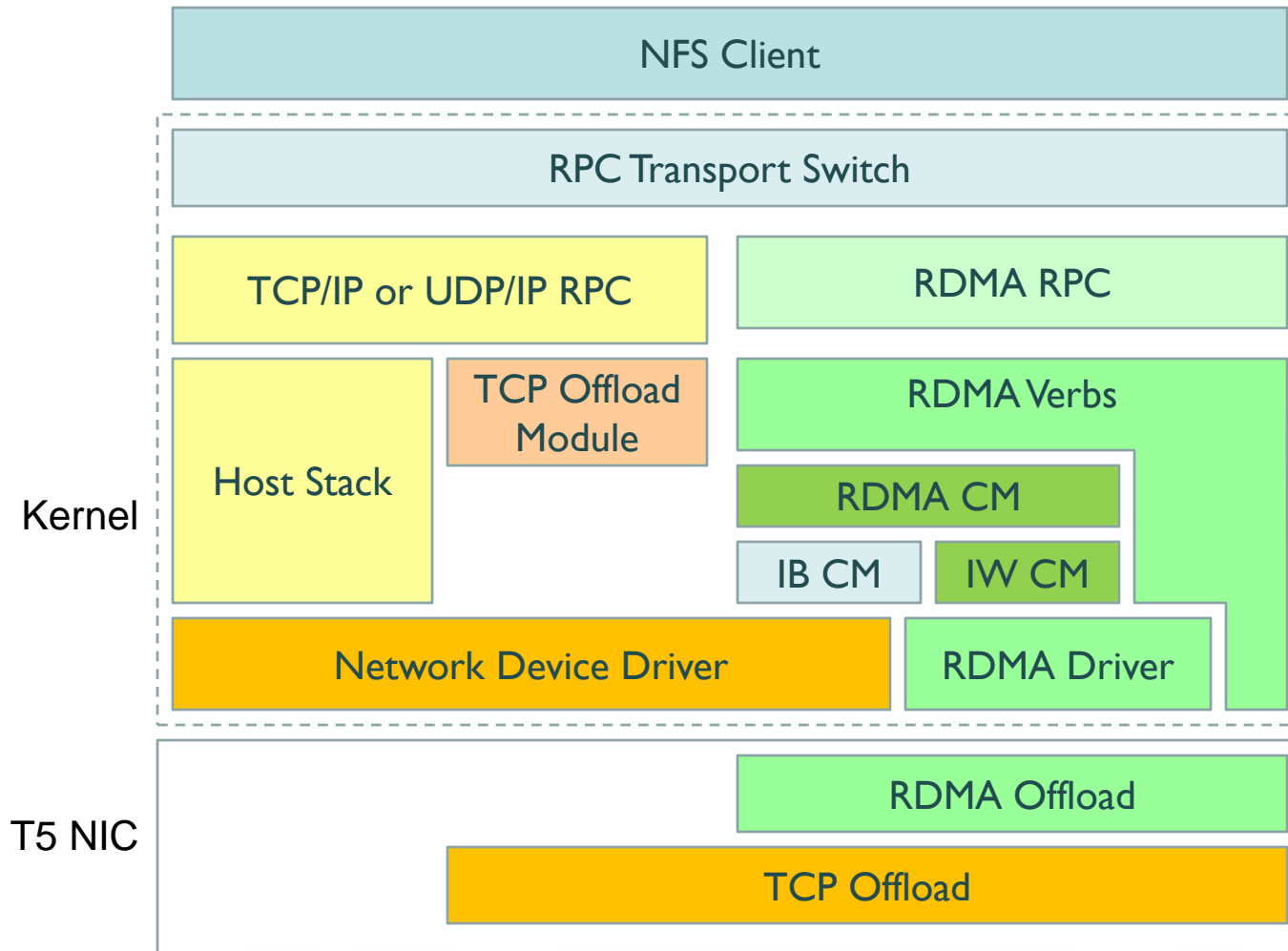
NFS over RDMA Timeline

- ❑ NetApp/Sun 2007
- ❑ IETF RFCs
 - ❑ RFC 5532 problem statement in 2009
 - ❑ RFC 5666 RDMA for RPC in 2010
 - ❑ RFC 5667 NFS DDP in 2010
- ❑ Renewed effort with rise in RDMA interest
 - ❑ Under active development – mostly client side
 - ❑ Chelsio, Emulex, Intel, LANL, Mellanox, NASA, NetApp, Oracle...

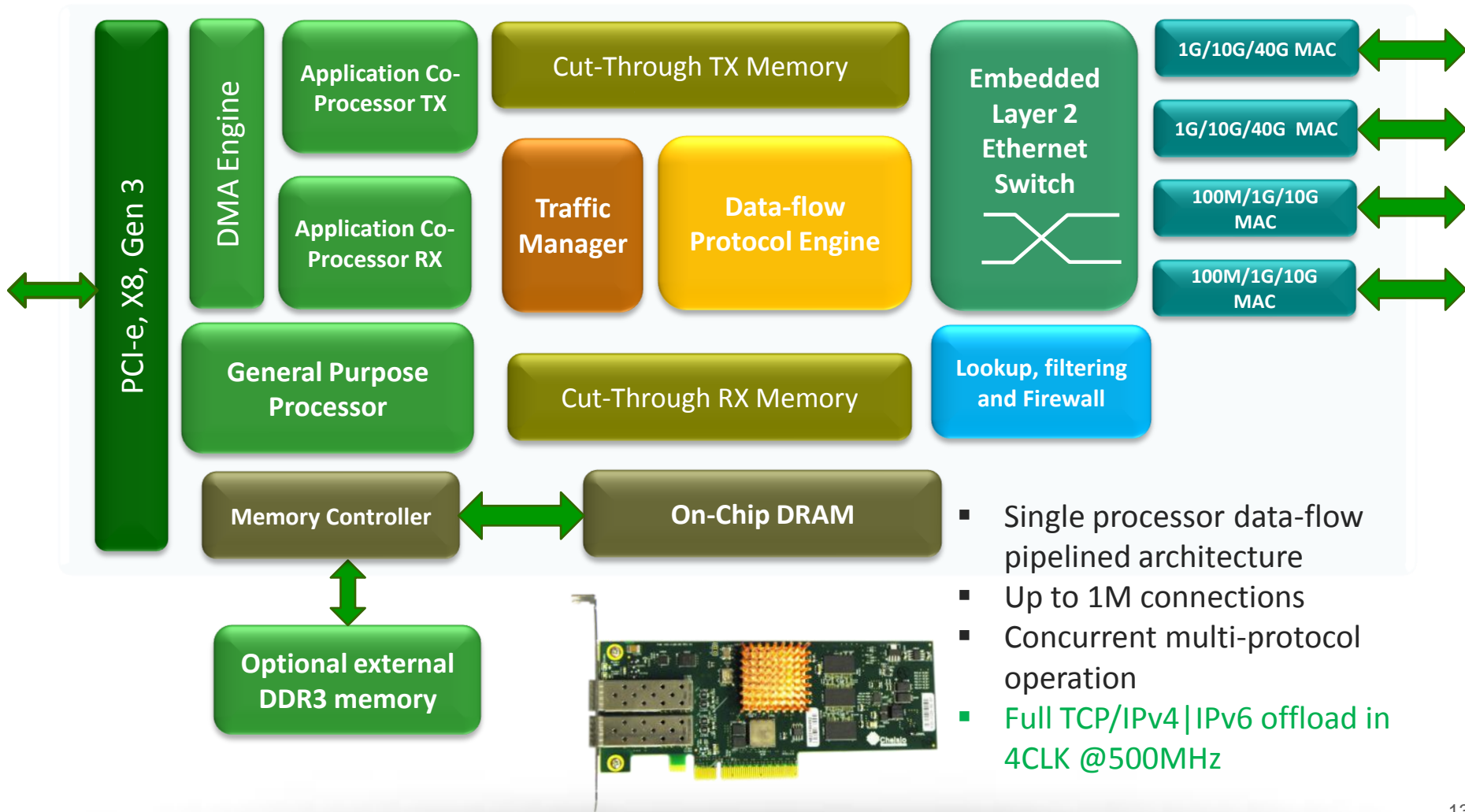
NFS over RDMA Overview

- ❑ NFS extensions to use RDMA fabric (for NFSv2,3,4)
- ❑ Client sends RPC in RDMA messages
- ❑ Server initiates RDMA data transfer transactions
 - ❑ Reduces client side CPU utilization
 - ❑ Eliminates client side data copies
 - ❑ Leverages low latency fabric
 - ❑ Requires NIC with RDMA offload at both server and client ends

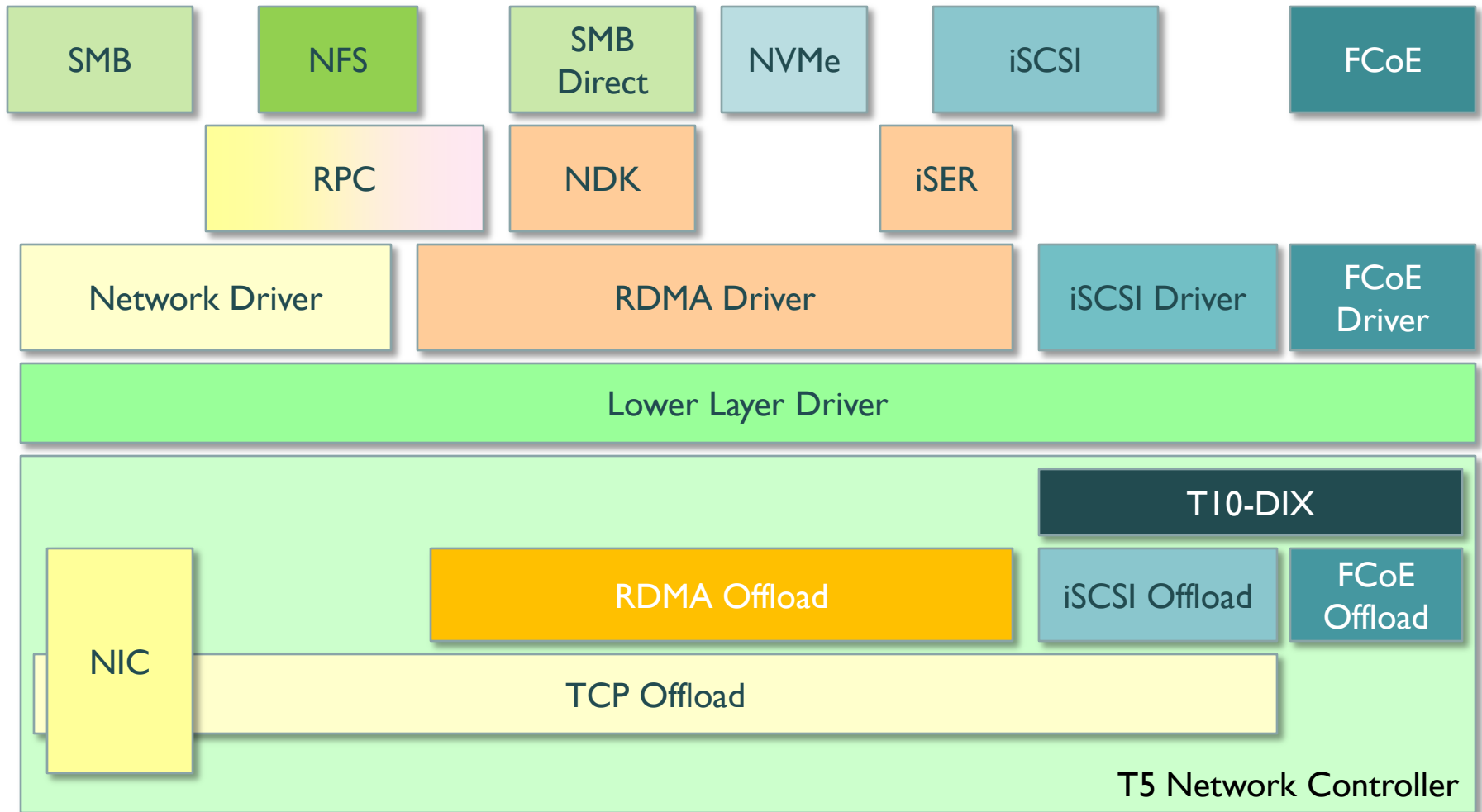
NFS Client Stack



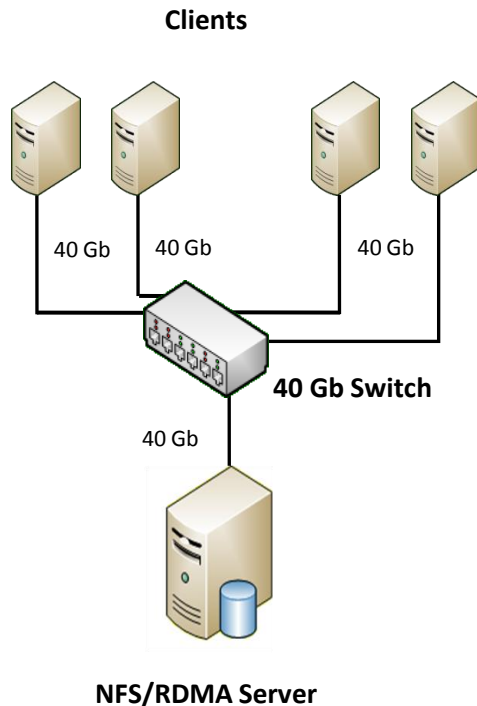
Chelsio T5 Ethernet Controller ASIC



T5 Storage Protocol Support



Test Configuration

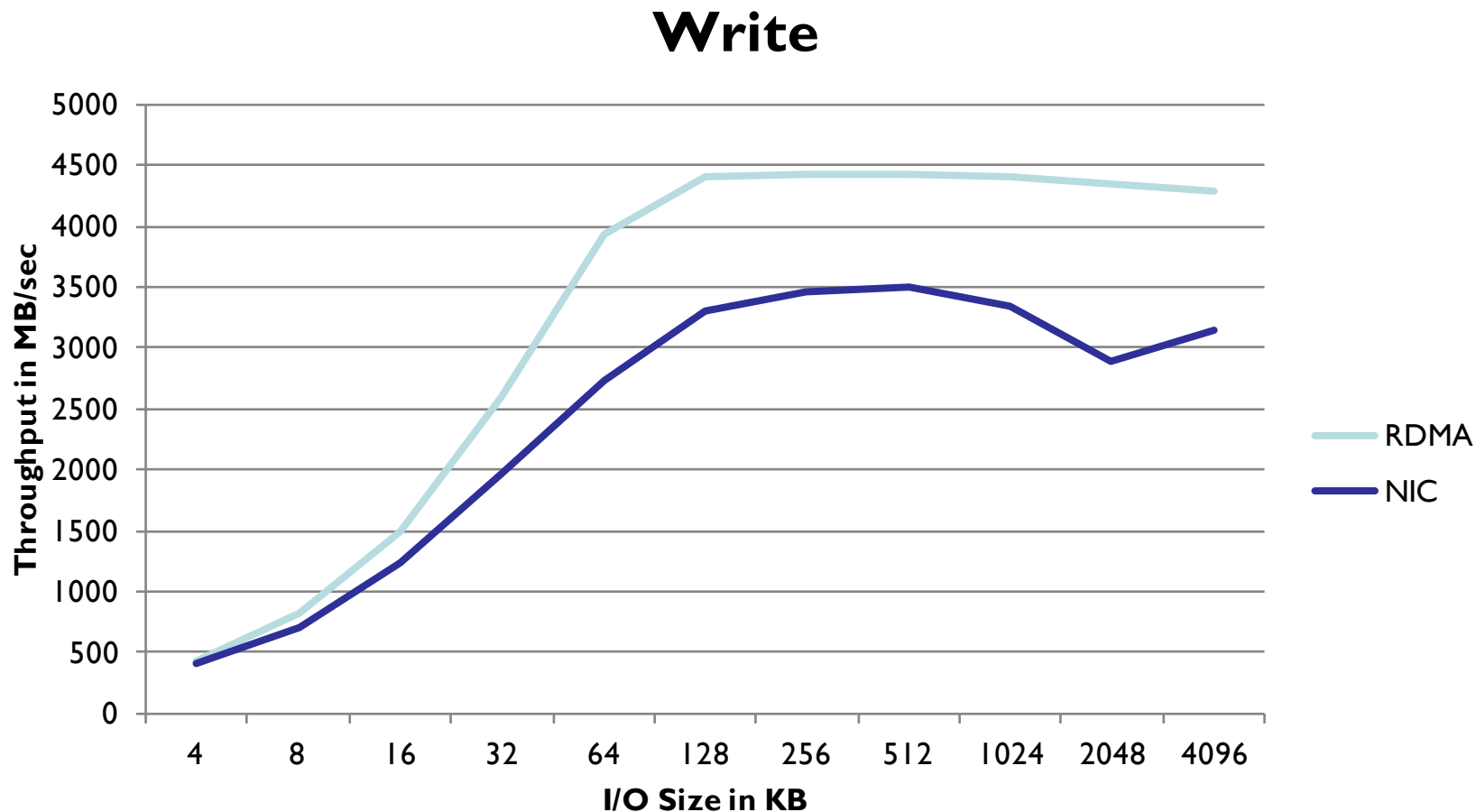


- Clients connected through switch to server with all 40Gbps links
- Sequential I/O direct (no buffer caching)
- Need OFED 3.12+ for 40G iWARP support

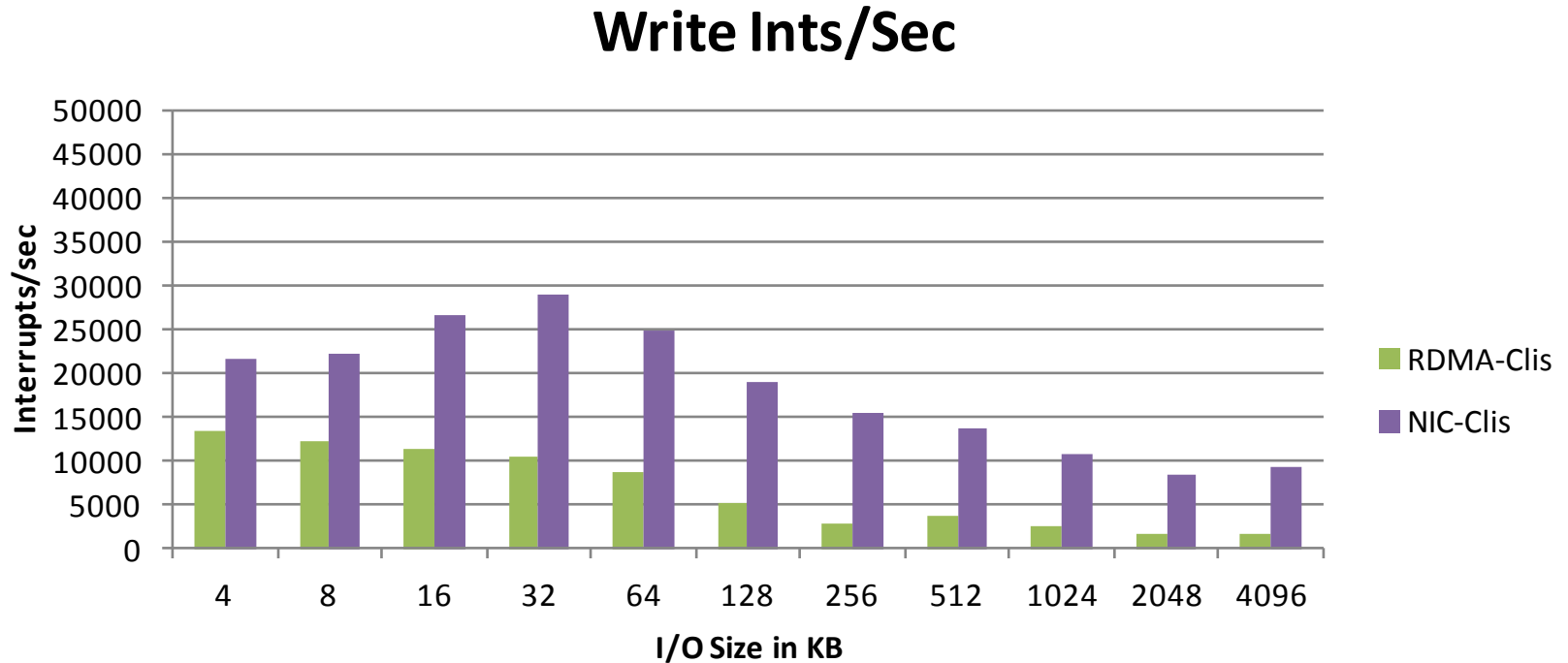
Clients (x4)	
OS	RHEL6.5
Kernel	3.16.0, NFSv4 + latest NFSRDMA fixes
Processor	Intel(R) Xeon(R) CPU E5-2687W v2@3.40GHz
No of Processors	2
No of Cores Total	16 (HT Disabled)
RAM	64 GB
Card Type	T580-CR
Card Core Clock	500MHz

Server	
OS	RHEL6.1
Kernel	3.16.0, NFSv4 + latest NFSRDMA fixes
Processor	Intel(R) Xeon(R) CPU E5-2687W @ 3.10GHz
No of Processors	2
No of Cores Total	16 (HT Disabled)
RAM	64 GB
Card Type	T580-CR
Card Core Clock	500MHz
Share	32GB ramdisk w/ ext2 filesystem.

NFS Write – iWARP vs. L2 NIC

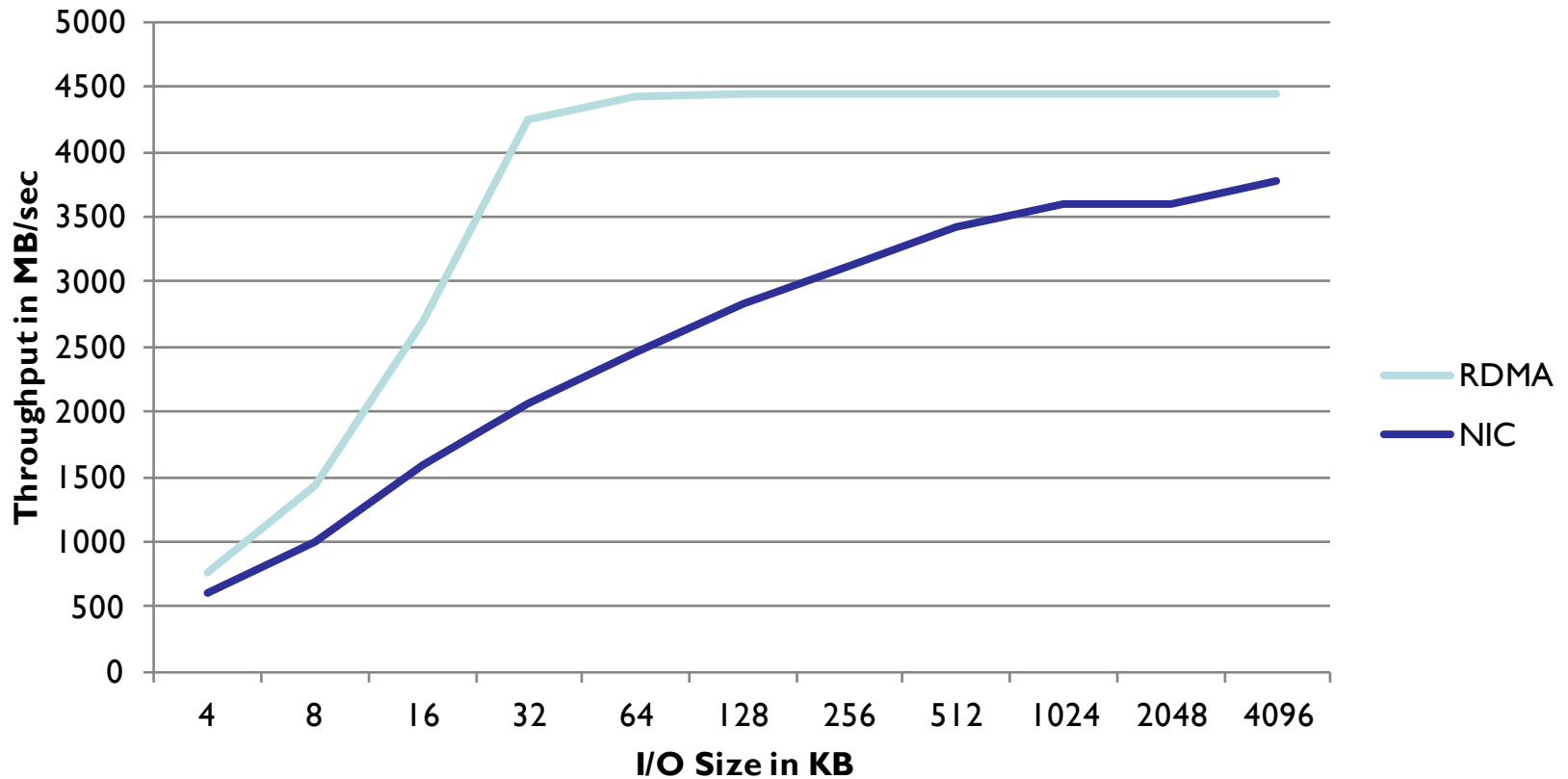


NFS Write Client Ints/sec – iWARP vs. L2 NIC

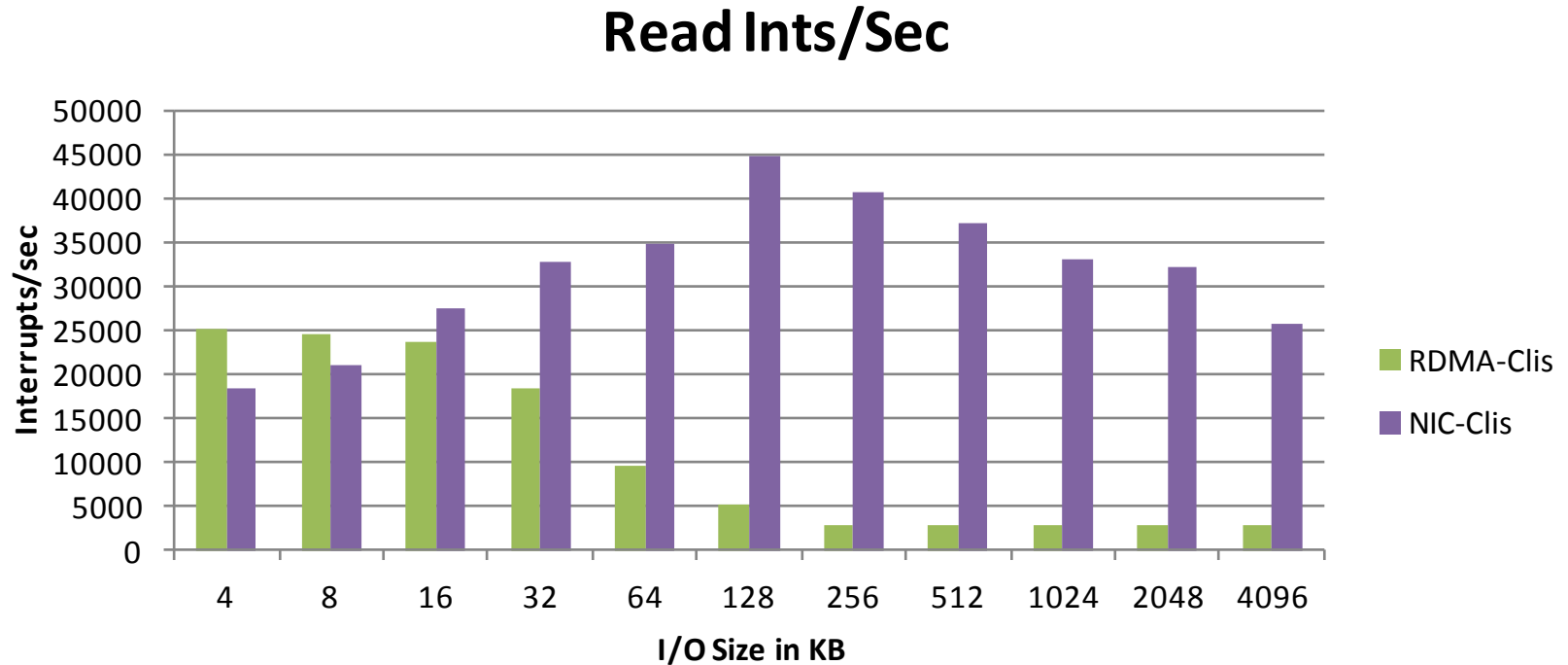


NFS Read – iWARP vs. L2 NIC

Read

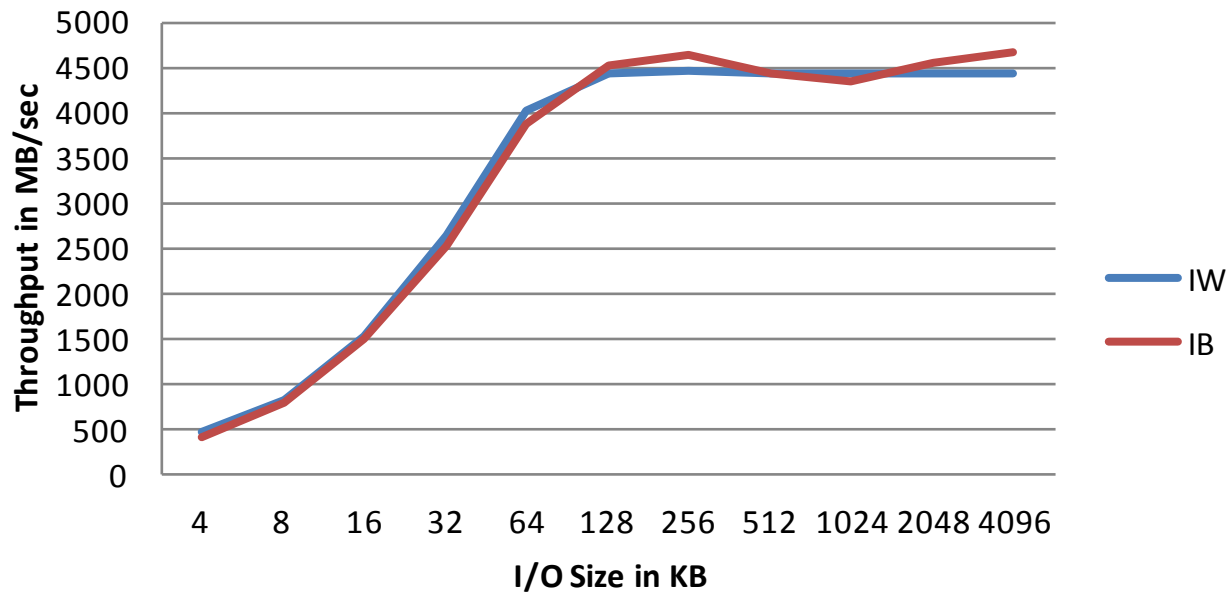


NFS Read Client Ints/sec – iWARP vs. L2 NIC



NFS Write – iWARP vs. InfiniBand

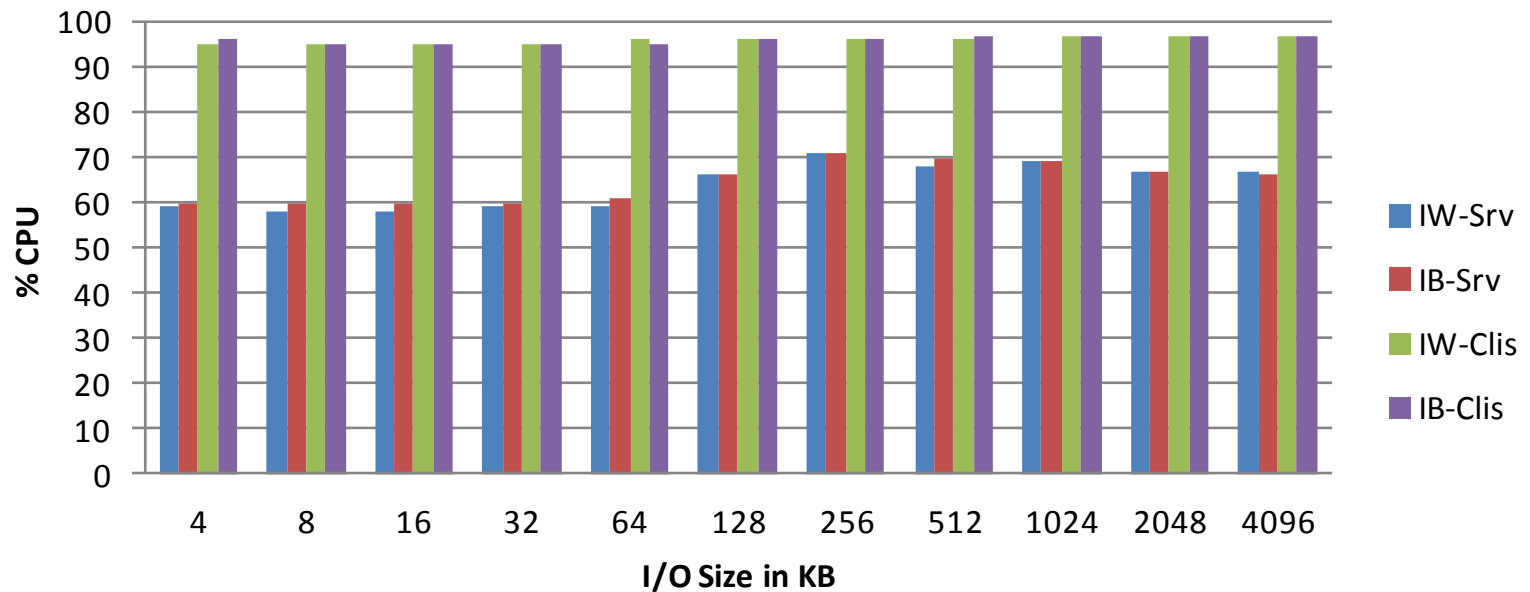
Write Throughput



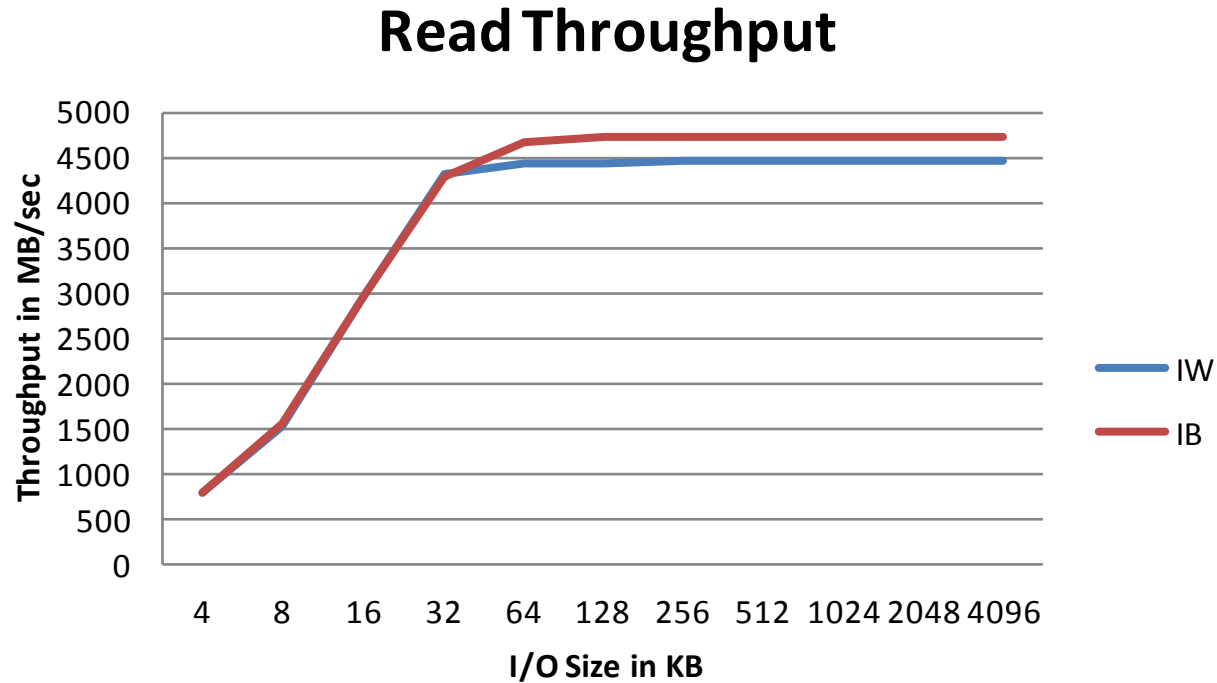
RHEL6.4, NFS Share: 40GB ramdisk, ext2 file system
Kernel: 3.16.0 + NFSv4 + latest NFSRDMA/cxgb4 fixes, default settings
CPU: Intel(R) Xeon(R) CPU E5-2687W 0 @ 3.10GHz 64GB RAM 2 CPUs, 16 cores total, no HT
IW HW: Chelsio Communications Inc T580-LP-CR Unified Wire Ethernet Controller
IB HW: Mellanox Technologies MT27500 Family [ConnectX-3] FDR

NFS Write – iWARP vs. FDR InfiniBand

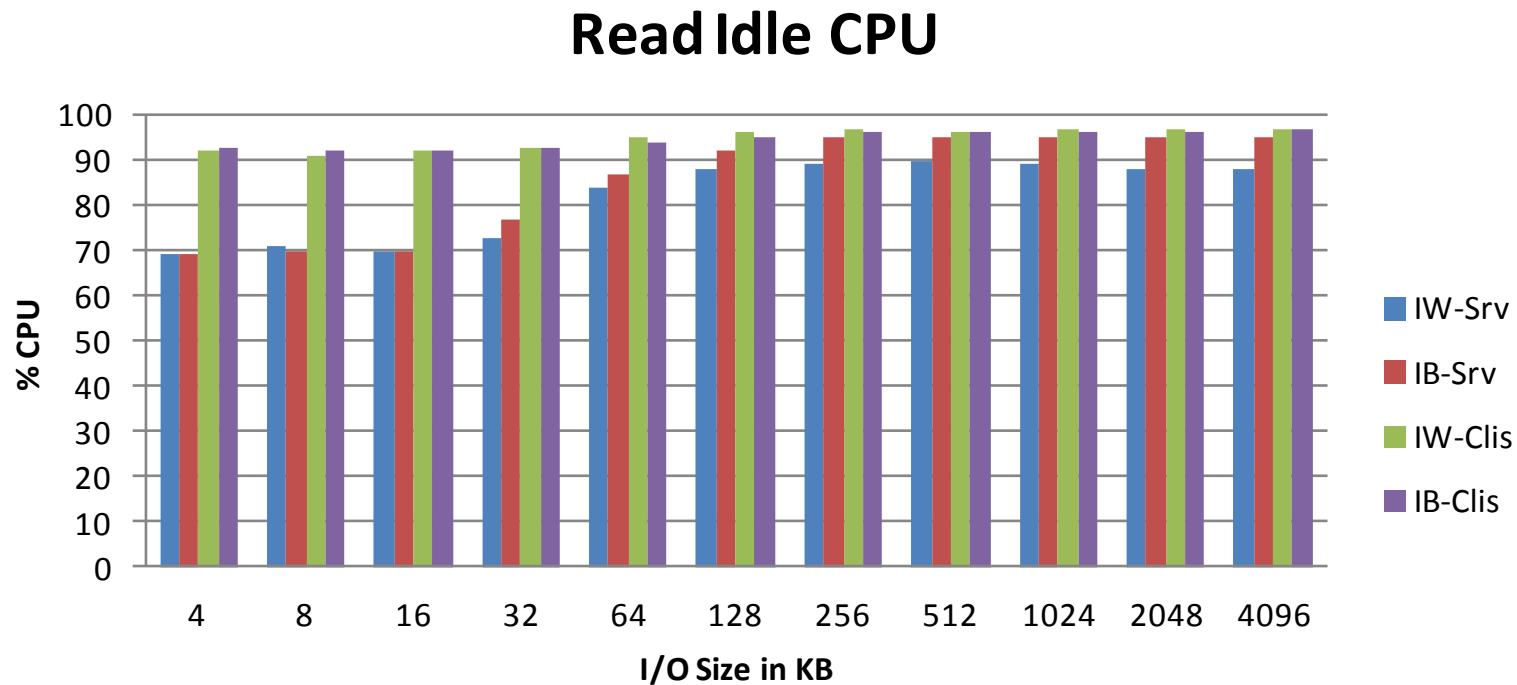
Write Idle CPU



NFS Read – iWARP vs. InfiniBand



NFS Read – iWARP vs. InfiniBand



Conclusions

- ❑ RDMA fabric offers potential for improved efficiency
 - ❑ SMB v3.0 RDMA transport demonstrated significant gains
- ❑ Renewed interest in NFS/RDMA
 - ❑ Work in progress
 - ❑ Performance benefits compared to NIC
- ❑ iWARP RDMA is shipping at 40Gbps
 - ❑ High performance Ethernet alternative to InfiniBand
- ❑ Chelsio adapter enables simultaneous operation of RDMA, NIC, TOE, iSCSI, FCoE...
 - ❑ TCP/IP for Wireless, LAN, Datacenter and Cloud networking
 - ❑ Remains “a great all-in-one adapter”*
- ❑ Call to action
 - ❑ Contribute to RDMA and NFS/RDMA in Linux
 - ❑ Mailing lists linux-rdma and linux-nfs on vger.kernel.org



STORAGE DEVELOPER CONFERENCE

SNIA ■ SANTA CLARA, 2014

Thank You

Please visit www.chelsio.com for more info