iSCSI or iSER?

Asgeir Eiriksson
CTO
Chelsio Communications Inc

Introduction

- iSCSI is compatible with 15 years of deployment on all OSes and preserves software investment
- iSER and iSCSI are layered on top of SCSI
- iSER and iSCSI have built in support for RDMA
  - iSER uses offload NIC on initiator and target sides
  - iSCSI can use software implementations on initiator and target side (soft-iSCSI)

Introduction

- iSER has different reach options
- iSCSI goes where TCP/IP goes
- iSER is on top of verbs RDMA that is used in HPC, HFT, file systems e.g. SMB3 and NVMe over fabrics
- iSCSI offload speed scales the same as iSER

Introduction: SSD and iSCSI and iSER

- Storage API are evolving for optimal use of SSD
  - Will use native API (without SCSI layer)
  - NVMe over fabrics
  - NVM DIMM
- There needs to be a path from iSCSI and iSER to support SSD natively

Introduction: iSCSI vs iSER

- iSER reach options
  - SCSI over iWARP over TCP/IP
  - SCSI over RoCE/IB over UDP/IP over Ethernet
- iSCSI characteristics
  - Over TCP/IP
  - software initiators and/or targets (soft-iSCSI)
  - iSCSI offload devices

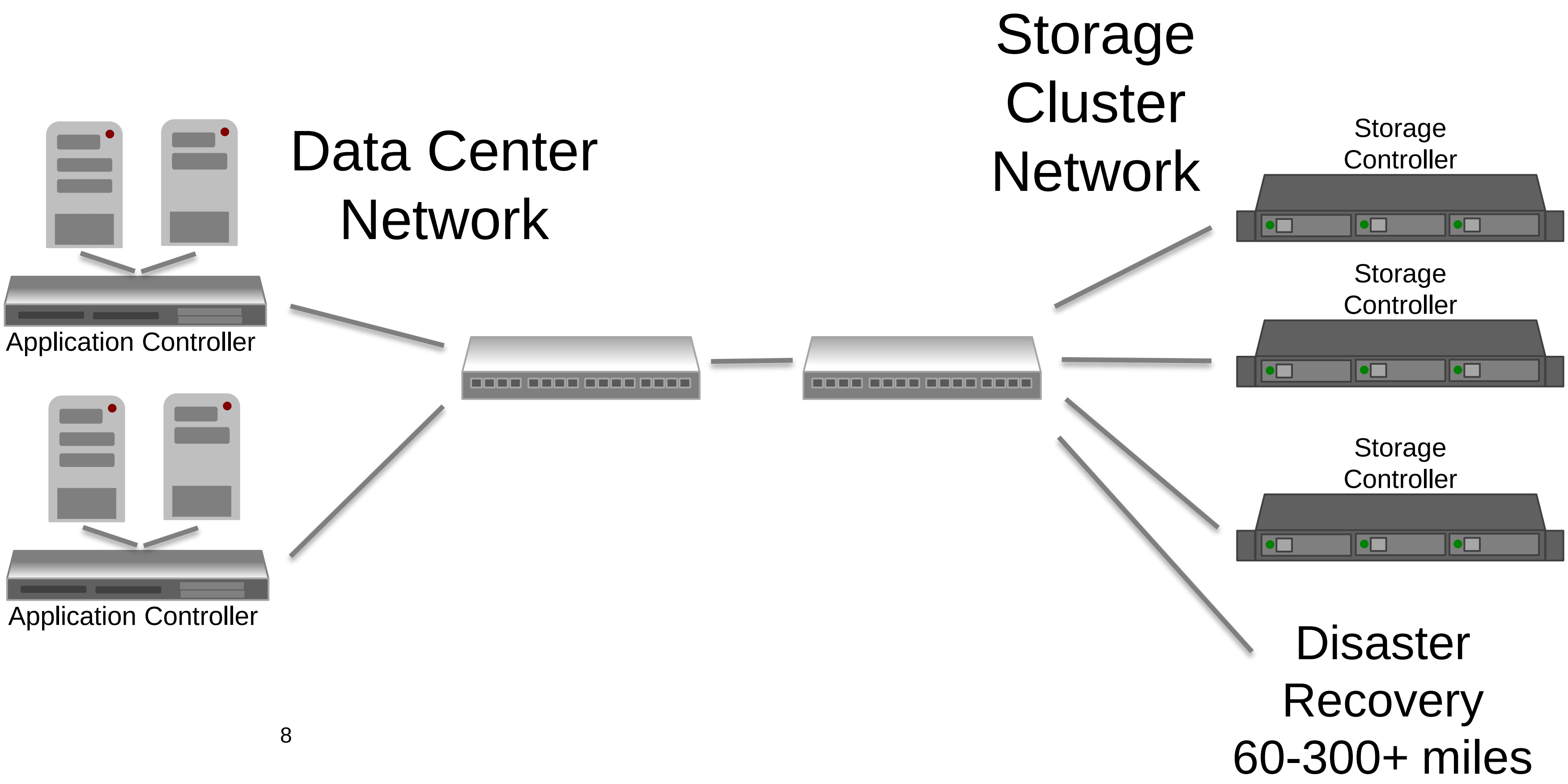
Introduction: iSCSI vs iSER incompatible

- iSER
  - iSER RoCE wire protocol not compatible with RoCEv2 or RoCEv3
  - iSER RoCEvn wire protocol not compatible with iSER iWARP
  - IB wire protocol not compatible with Ethernet
- iSCSI
  - Wire protocol not compatible with iSER

Introduction: speeds and feeds

	Bandwidth (Gbps)	Reach
Ethernet		
iWARP	1, 2.5, 5,10,25,40,50,100	Rack, Data Center, LAN, MAN, WAN
iSCSI		Rack, Data Center, LAN, MAN, WAN
RoCEvn		Rack, Data Center
Infiniband	8, 16, 32, 56, 112	Rack, Data Center

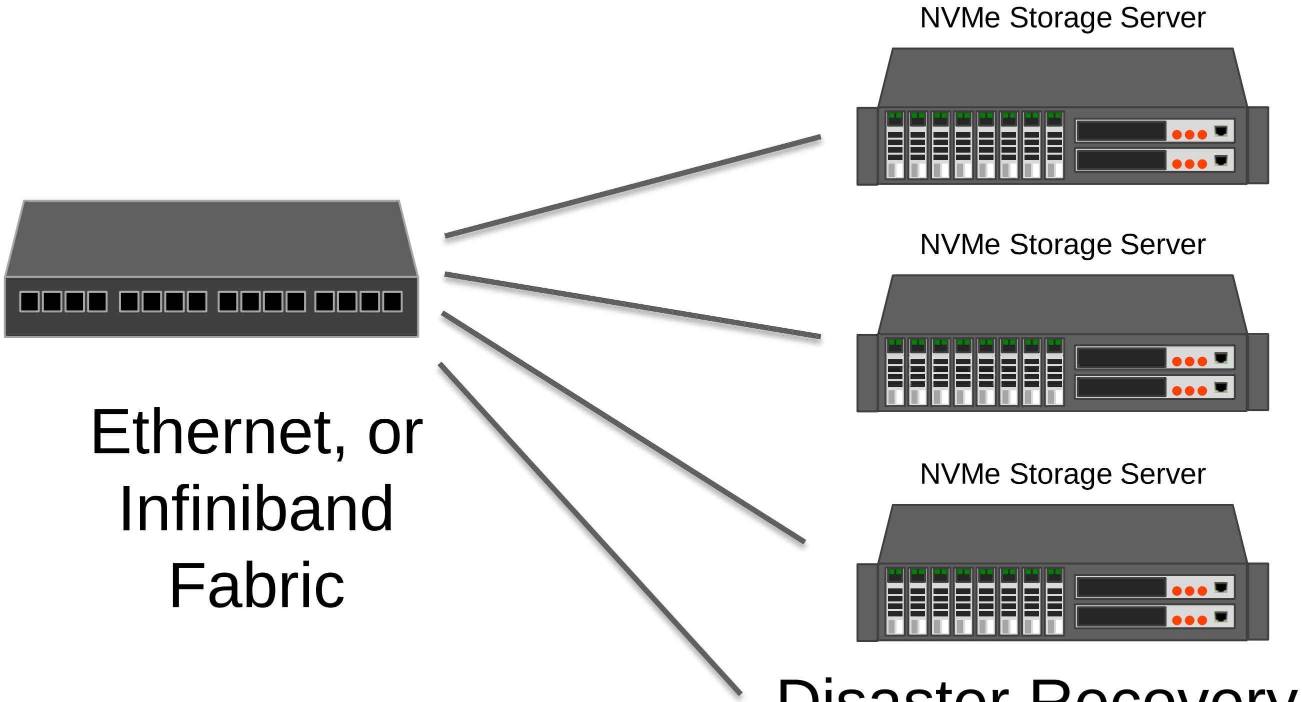
Traditional Scale Out Storage



Traditional Scale Out Storage

- Preserves software investment
- Realizes some of the SSD speedup benefits
  - NVMe over RDMA fabrics over SCSI
- Disaster Recovery (DR) requires MAN or WAN
  - iSCSI
  - iSER iWARP

Shared Server Flash Flash



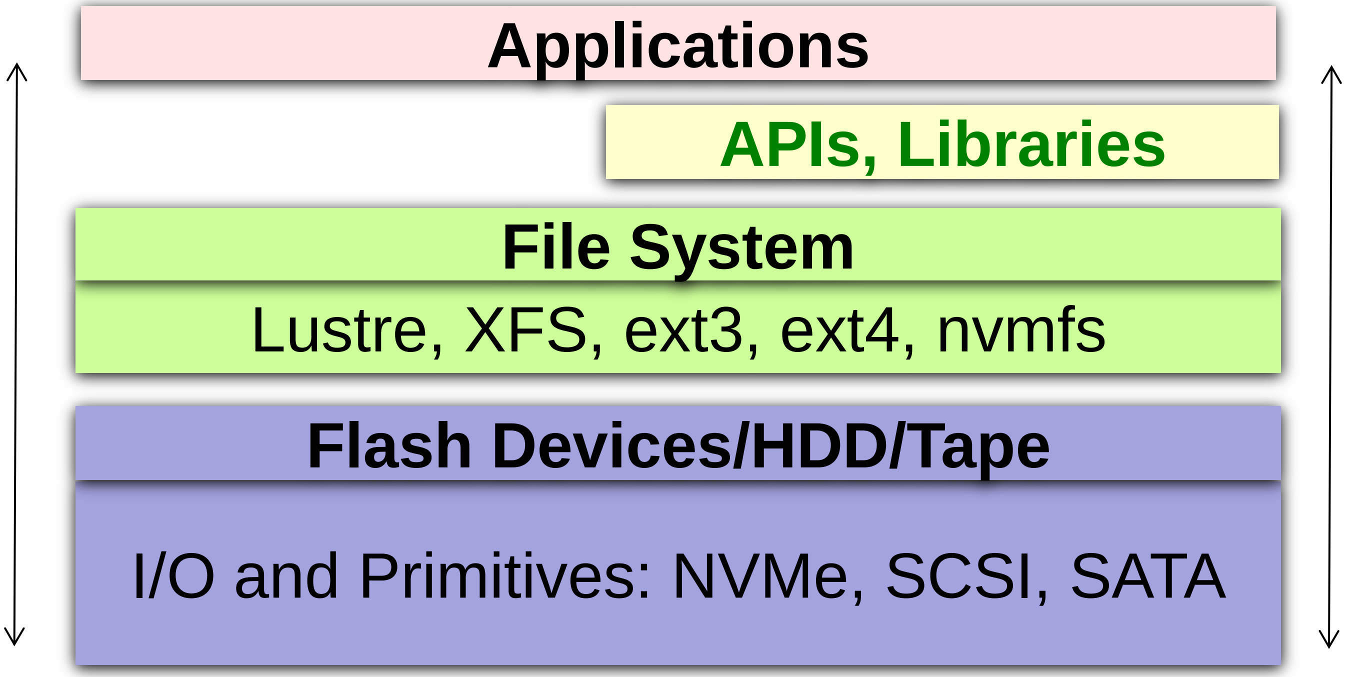
Ethernet, or
Infiniband
Fabric

Disaster Recovery
60-300+ miles

Shared Server Flash

- Ethernet or IB fabric
- RDMA required for sufficient efficiency
  - IB uses RDMA
  - Ethernet has RoCEvn, iWARP and iSCSI with RDMA
- Disaster Recovery (DR) requires MAN or WAN

File and Block Storage API



Applications

APIs, Libraries

File System

Lustre, XFS, ext3, ext4, nvmfs

Flash Devices/HDD/Tape

I/O and Primitives: NVMe, SCSI, SATA

File and Block Storage API

- Preserve software investment
  - Carry forward support for soft-iSCSI
  - Layer SCSI on top of SSD devices e.g. NVMe

- Add support for native NVMe API
- Alternatively jump directly to native SSD API

- Even better: support both at the same time

File and Block Storage API

- Preserving software investment and supporting native SSD storage API is possible with devices that support both iSCSI offload and iSER offload
  - iSER is layered on top of verbs RDMA
  - NVMe over fabrics uses verbs RDMA
- Chelsio T5 supports both iSCSI offload and iSER offload

Ethernet vs Infiniband

- Infiniband
  - Reliable link layer
  - Credit based flow control
- Ethernet is ubiquitous
  - Pause and Prioritized Pause (PPC) for lossless operation that propagates through some switches and fewer routers
  - Flow Control and Reliability at higher layer e.g. TCP, and IB Transport Layer for RoCEvn

Ethernet vs TCP/IP

- iSER over RoCEvn
  - Requires DCB extensions to Ethernet
- iSER over iWARP
  - Goes where TCP/IP goes: wired, wireless, Ethernet, OC-192, rack, cluster, datacenter, LAN, MAN, WAN, space, etc.
- iSCSI goes where TCP/IP goes

Comparing Ethernet Options

	DCB Required	Reach	IP routable	RDMA
FCoE	√	Rack, LAN		√
iSCSI	No	Rack, datacenter, LAN, MAN, WAN Wired, wireless	√	√
iWARP	No	Rack, datacenter, LAN, MAN, WAN Wired, wireless	√	√
RoCEv2	√	Rack, LAN, datacenter	√	√

Comparing Ethernet Options

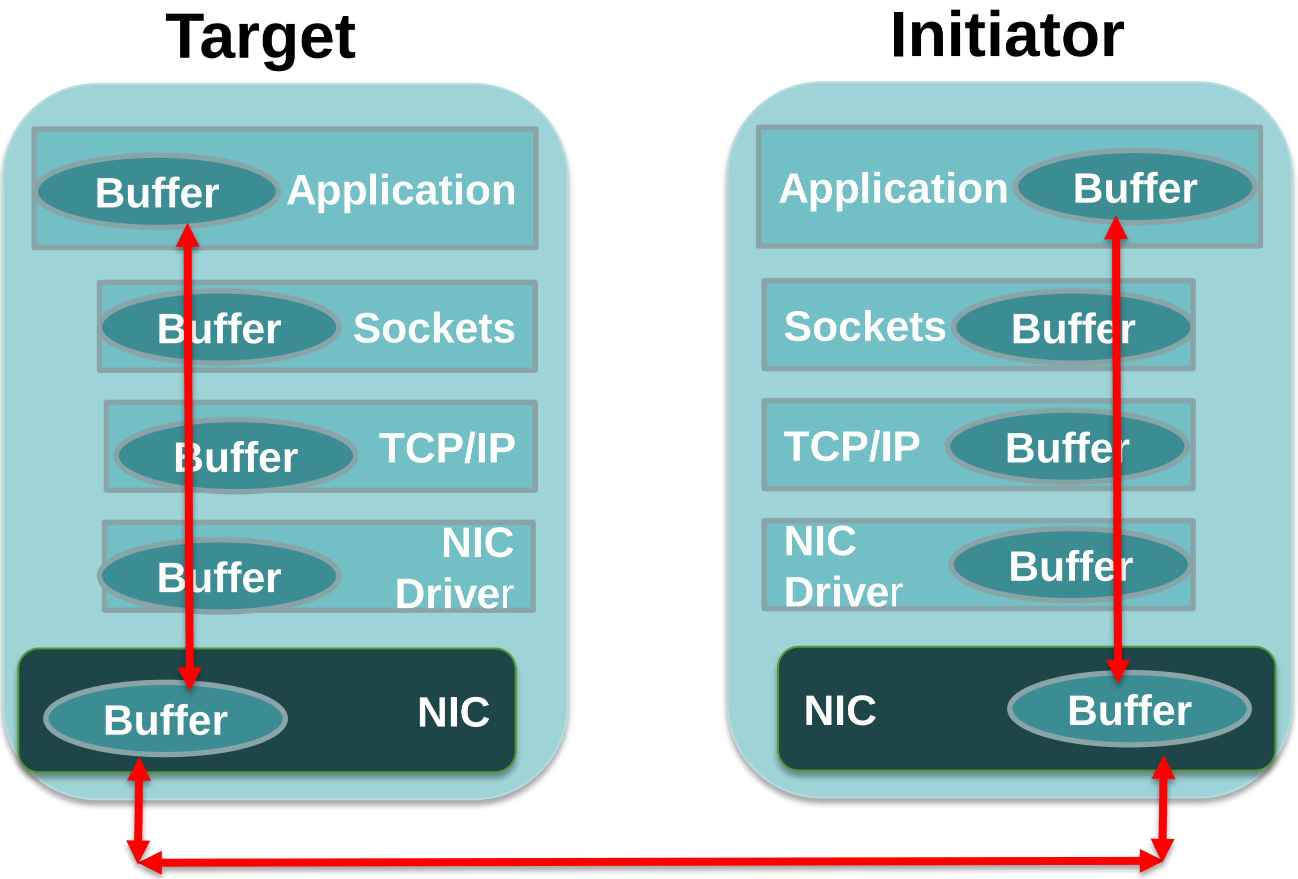
- iSCSI, iWARP
  - Use DCB when it is available but not required for high performance
- iSCSI
  - Has RDMA WRITE and accomplishes RDMA READ by using an RDMA WRITE from other end-point
  - Concurrent support for legacy soft-iSCSI

Comparing Ethernet Options

- RDMA bypasses the host software stack
  - RoCEvn
  - iWARP
  - iSCSI with offload
- soft-iSCSI
  - uses the host TCP/IP stack

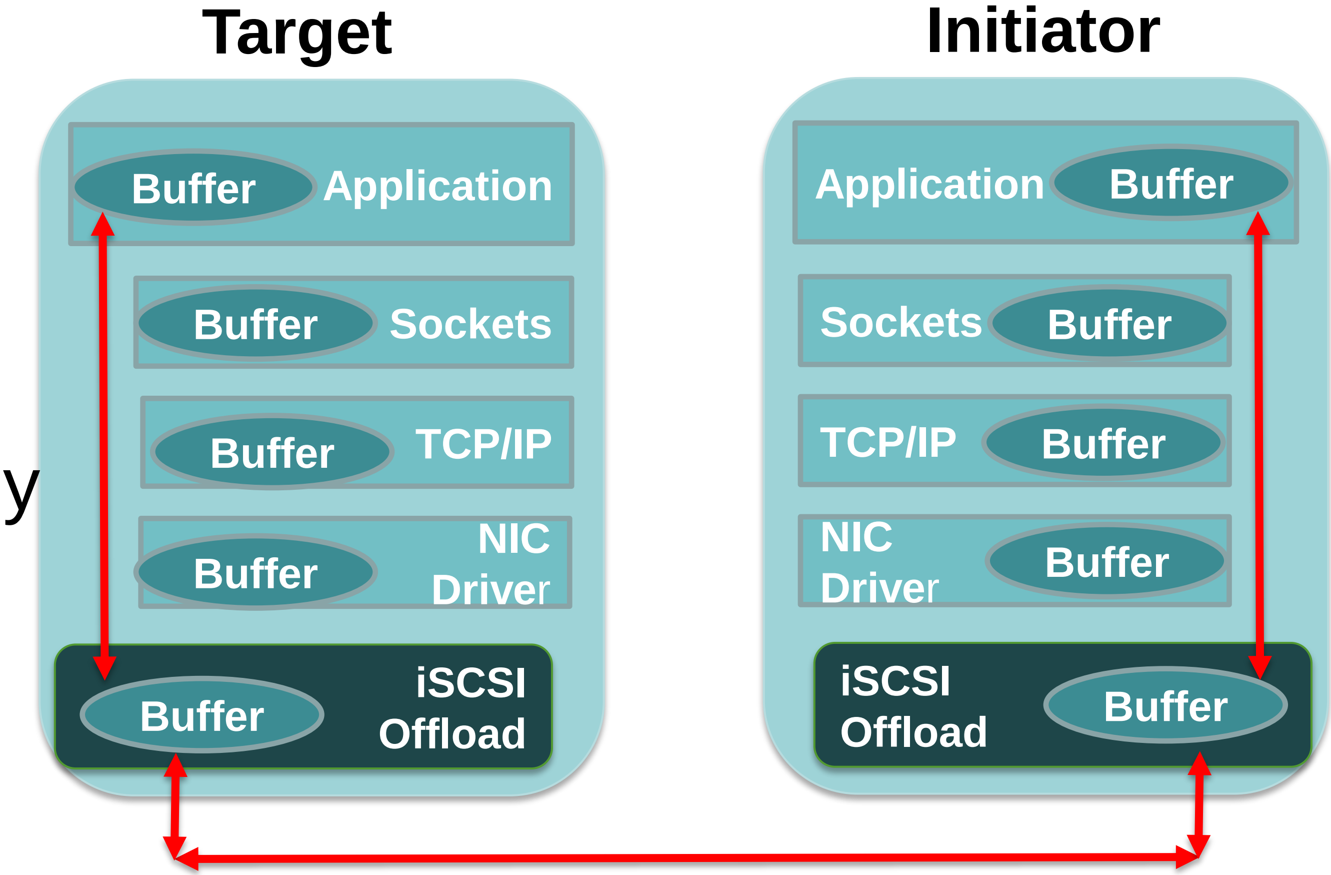
soft-iSCSI

- sw TCP/IP
- Multi-copy send
- Multi-copy receive



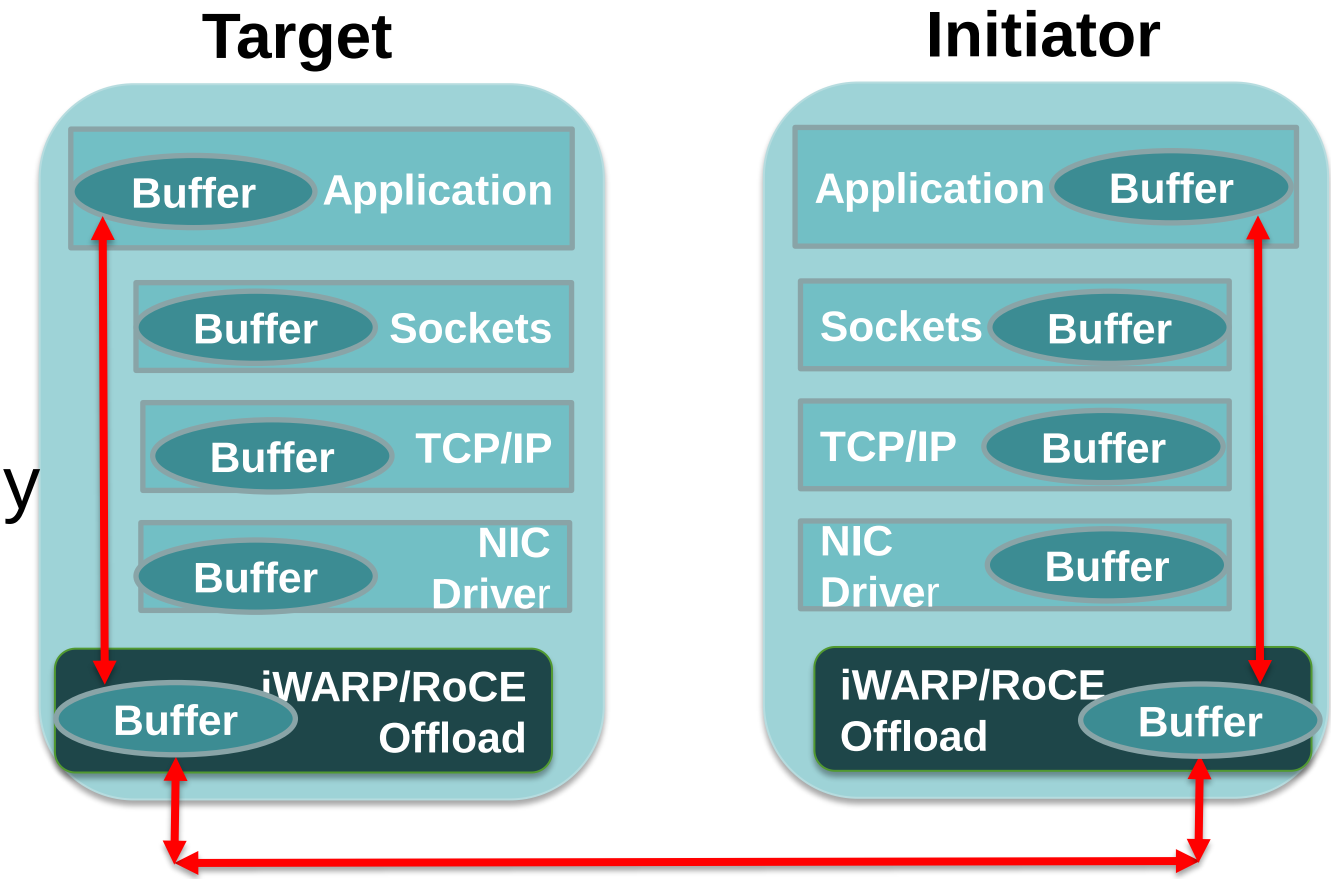
iSCSI offload

- Offload
  - TCP/IP
  - iSCSI
- Bypass
  - zero copy
  - send
  - receive
- RDMA



iSER offload

- Offload
  - TCP/IP
  - UDP/IP
- Bypass
  - zero copy
  - send
  - receive
- RDMA



iSCSI vs iSER scaling

- ❑ Chelsio T5 supports iSCSI and iSER concurrently
  - ❑ 2x40GE/4x10GE support
  - ❑ A storage target using T5 can connect to iSCSI and iSER initiators concurrently
  - ❑ The iSCSI hardware can support hardware initiators and software initiators concurrently
  - ❑ Full TCP/IP offload
  - ❑ Full iSCSI offload or iSCSI PDU offload

iSCSI vs iSER scaling

- Chelsio's iSCSI and iSER implementations scale equally well
  - iSCSI and iSER share the same hardware pipeline
    - Protocols interleave at packet granularity
    - Same hardware is used to implement DDP for iSCSI and iSER
    - Same hardware is used to segment iSCSI and iSER payload
    - Same hardware is used to insert/check CRC for iSCSI and iSER
    - Same hardware TCP/IP implementation
    - Same end-to-end latency for iSCSI and iSER
  - Operation mode is dynamically selected on a per-flow basis

iSCSI vs iSER Performance Comparison

- Use performance numbers for the Chelsio T5 that is a 4x10GE/2x40GE device that supports iSCSI offload, and iSER concurrently
  - 2x40GE performance limited by PCIe 8x Gen3
- In addition supports concurrently FCoE offload, NVMe over iWARP RDMA fabric, and regular NIC operation

Chelsio T5

Storage

- iSCSI, FCoE, T10-DIX
- NAS (TOE) Offload
- Diskless boot
- pNFS (NFS 4.1)
- SMB Direct

Virtualization & Cloud

- SR-IOV
- On-Chip VEB switching
- VEPA, VN-TAG
- Hypervisor offload
- NFV
- SDN

HFT

- WireDirect technology
- Low latency
- High messages/sec
- Wire rate classification

HPC

- iWARP/RDMA
- NFS RDMA
- pNFS (NFS 4.1)

Networking

- 4x10GbE/2x40GbE NIC
- Full Protocol Offload
- Data Center Bridging
- Firewall
- Wire Analytics

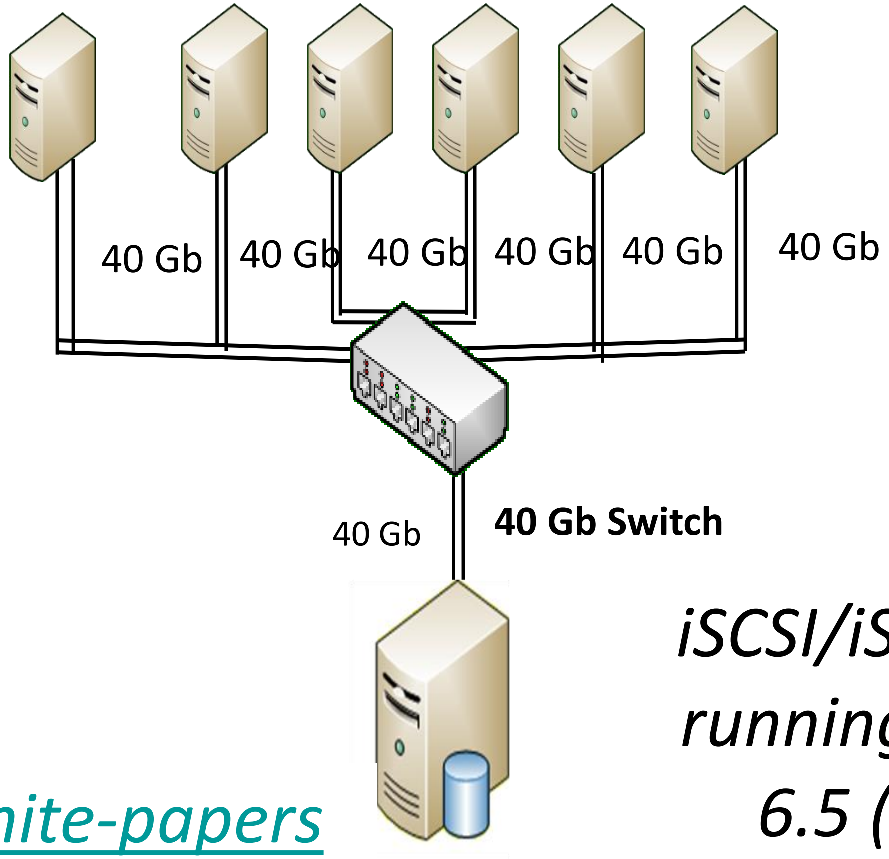
Media Streaming

- Traffic Management
- Video Offload

Chelsio T5-Based Unified Adapter

Performance iSCSI/iSER Offload

iSCSI Initiators with T580-CR adapters

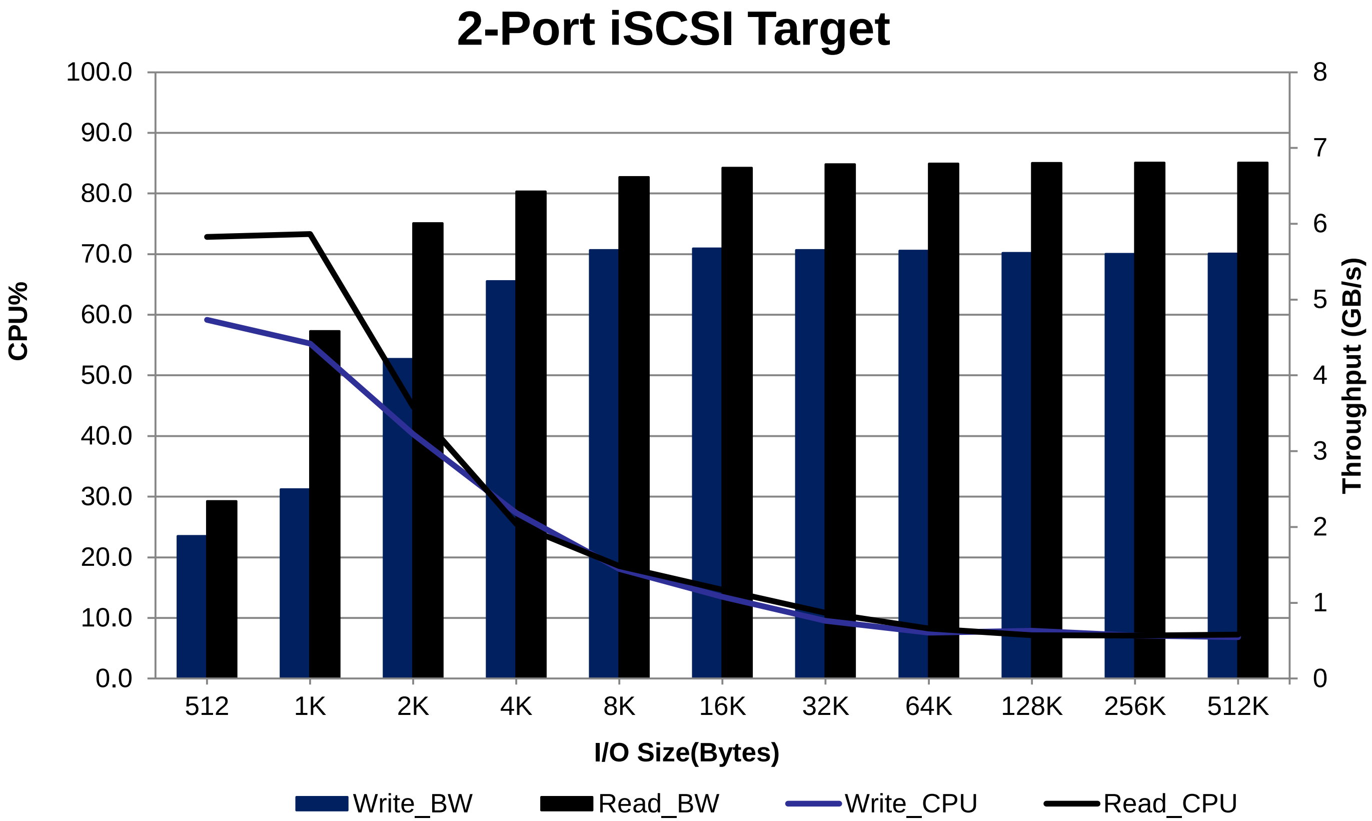


iSCSI/iSER Target running on RHEL 6.5 (3.6.11)

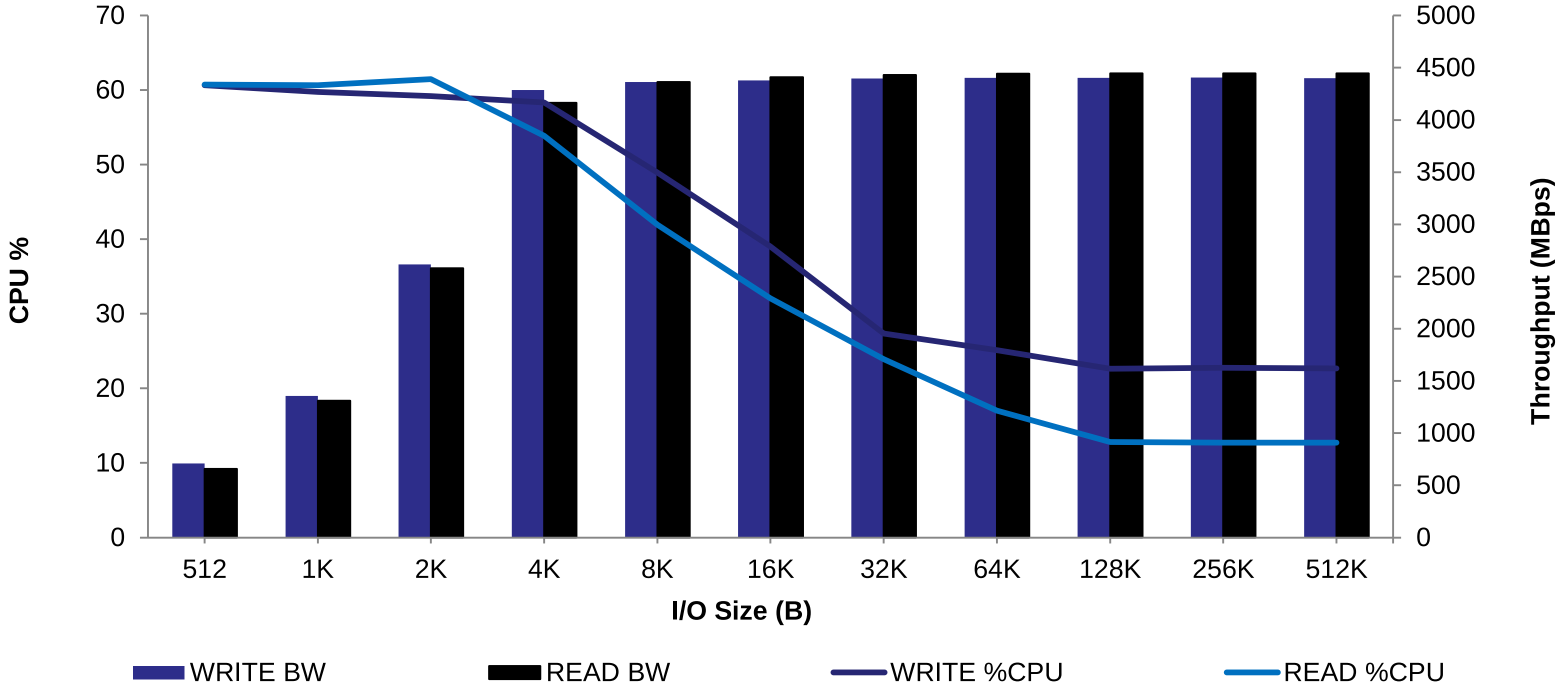
www.chelsio.com/white-papers

Has details

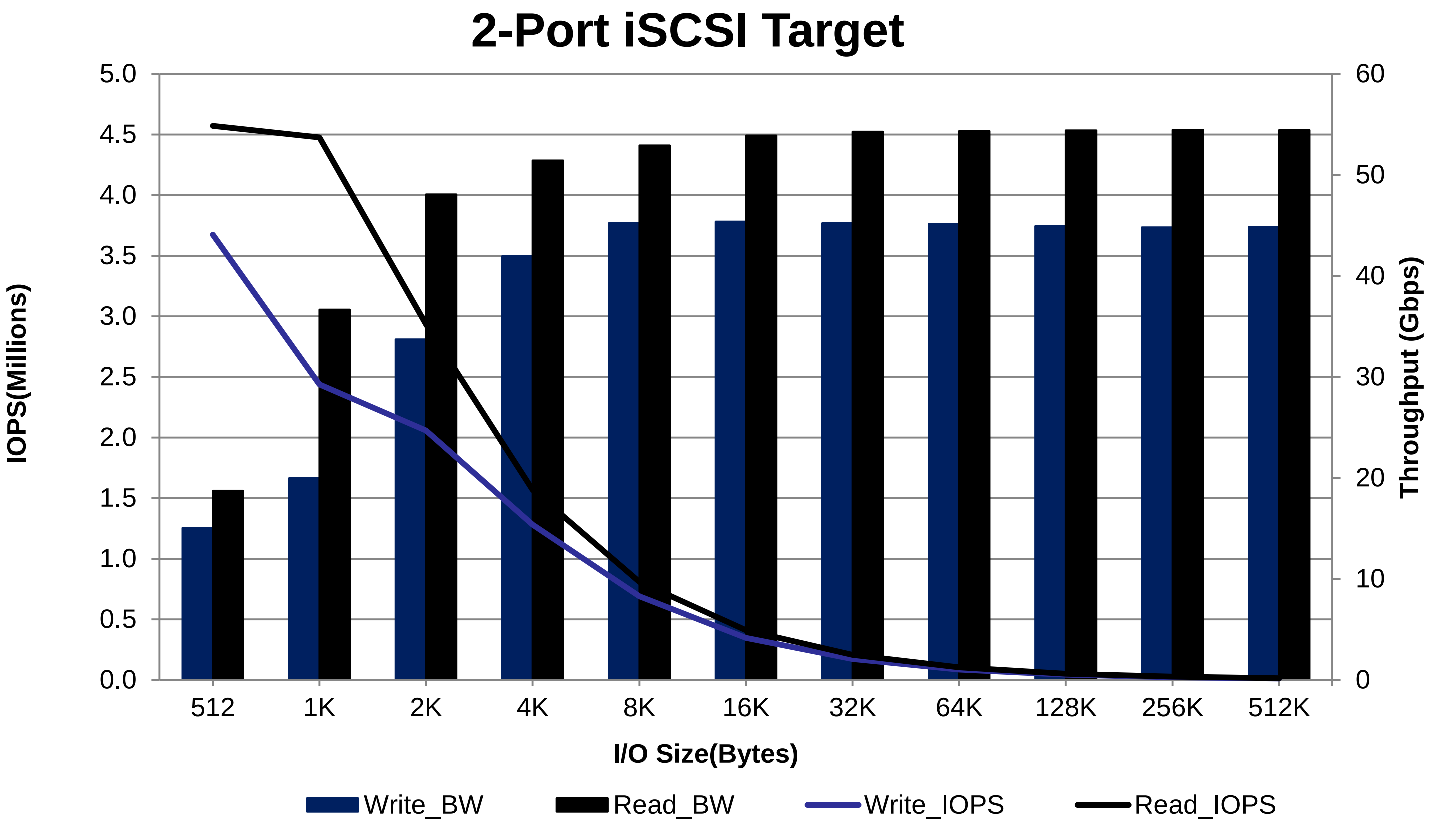
Performance iSCSI 2x40GE offload



Performance 1x40GE iSER



Performance 2x40GE iSCSI IOPS



Conclusions

- iSCSI and iSER layered on top of SCSI protocol which is designed for HDD and tape
  - SSD developing native API, with no SCSI
  - Support for NVMe over RDMA support future proofs investment in devices that support iSCSI and iSER offload

Conclusions

- iSCSI compatible with 15 years of deployment
  - Software initiators on all OSes
- The speed of iSCSI offload scales the same as iSER offload
  - Ethernet speeds have caught up with IB
  - Speed determined by a common SERDES
- iSCSI does not have reach limitations and it goes where TCP/IP goes