Concurrent Support of NVMe over RDMA Fabrics and Established Networked Block and File Storage

Ásgeir Eiriksson

CTO

Chelsio Communications Inc.

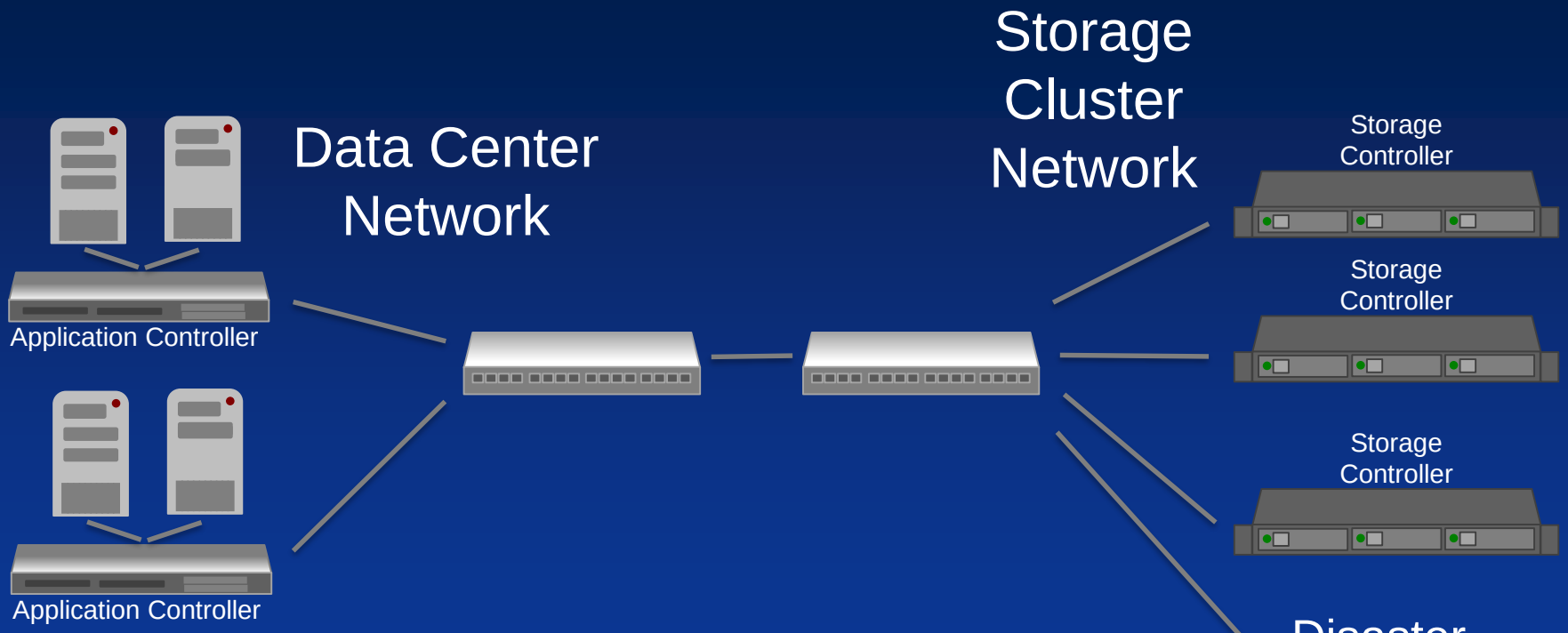
Introduction

- API are evolving for optimal use of SSD
  - NVMe
  - NVMe over RDMA fabrics (NVMf), for networked access
- Huge installed base of SMB, NFS, FC, iSCSI, etc
  - Ideally preserve existing storage product investment
  - Ideally support native NVMf API as ecosystem develops

Introduction

- Chelsio 10/25/40/50/100G Ethernet Adapters
  - Concurrent support for NVMf, SMB 3.X, NFSoRDMA, iSCSI, and FCoE
  - High BW and high IOPS for SMB 3.X, NFSoRDMA, iSCSI and FCoE using NVMe backing store
  - Concurrent High BW and high IOPS and low latency NVMf

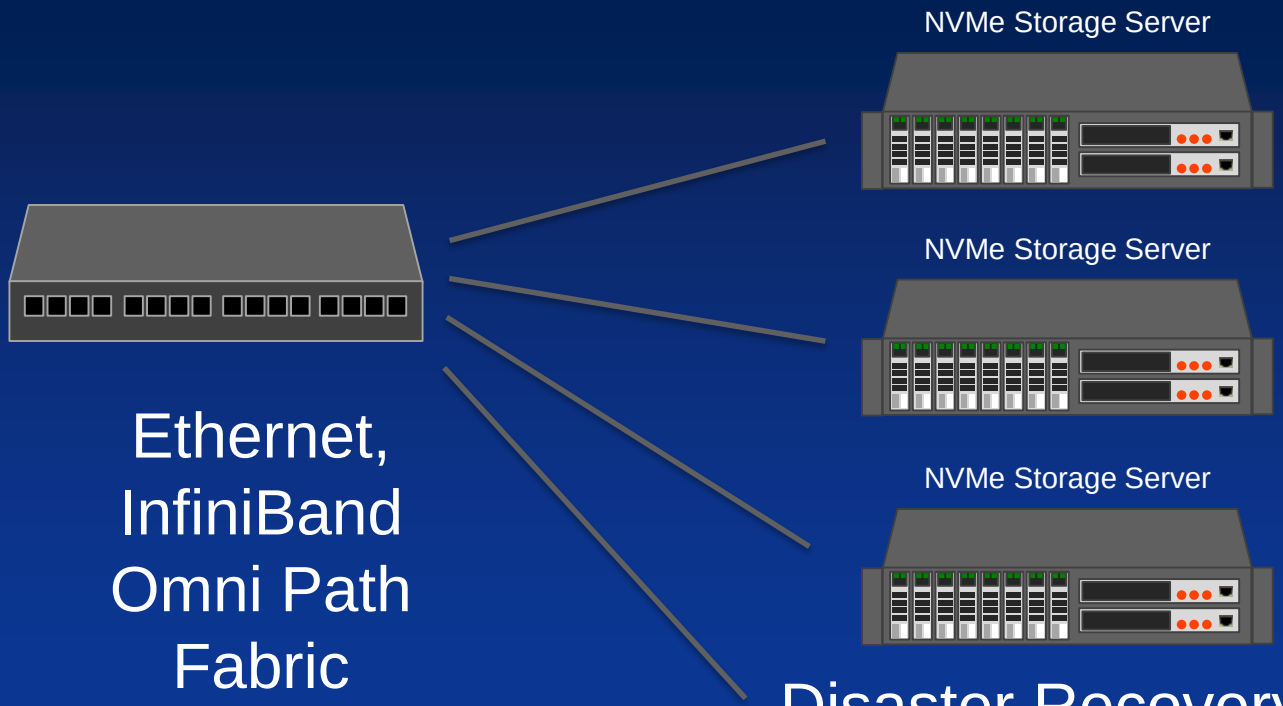
Traditional Scale Out Storage



Traditional Scale Out Storage

- Observations:
  - High BW/IOPS NVMe support preserves software investment, because it keeps existing software price/performance competitive
  - High BW/IOPS NVMe support realizes most of the NVMe speedup benefits
  - Disaster Recovery (DR) requires MAN or WAN

Shared Server Flash



Shared Server Flash

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

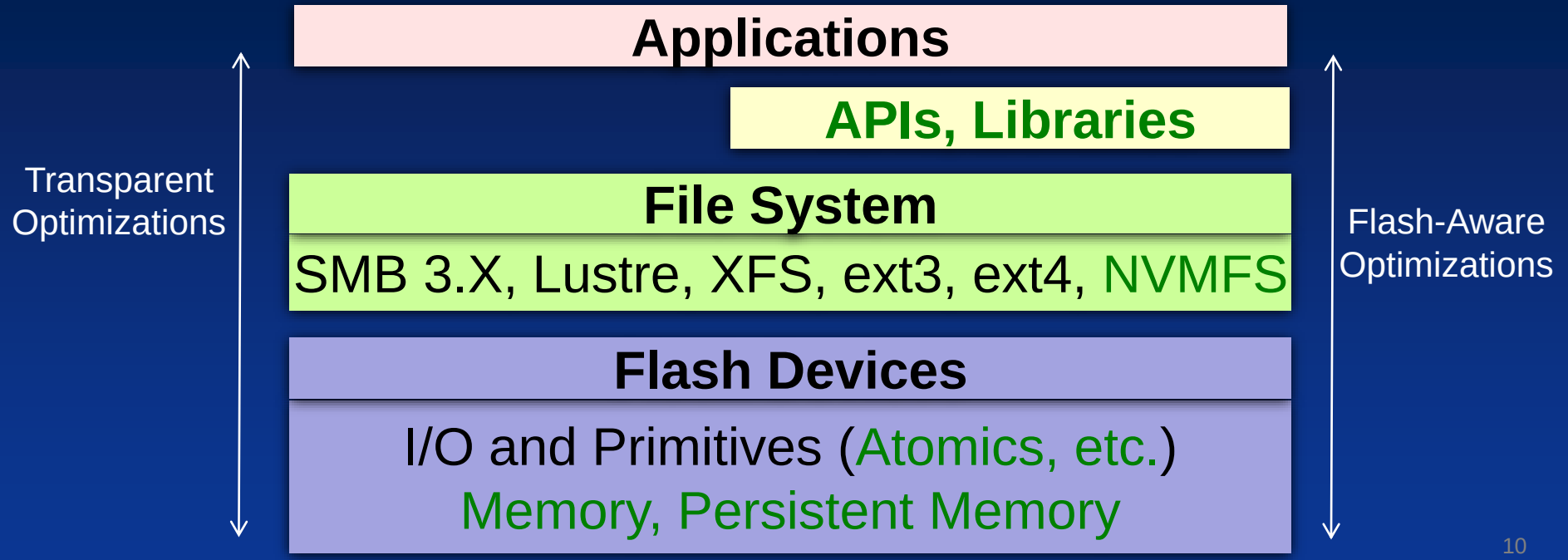
Comparing Ethernet Options

- 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
- RoCEvn
  - Fork uplift of infrastructure required e.g. specialized Ethernet switches, and specialized NIC

Ethernet, Infiniband, OmniPath

- Infiniband, OmniPath
  - Reliable link layer
  - Credit based flow control
- Ethernet
  - 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 RoCE

API


Applications
APIs, Libraries
File System
SMB 3.X, Lustre, XFS, ext3, ext4, NVMFS
Flash Devices
I/O and Primitives (Atomics, etc.)
Memory, Persistent Memory
Transparent Optimizations
Flash-Aware Optimizations

API Decision

- Preserve software investment
- Alternatively jump directly to native NVMe/NVMf API

- Strong preference: preserve investment while at the same time making use of emerging NVMe technology

Comparing Ethernet Options

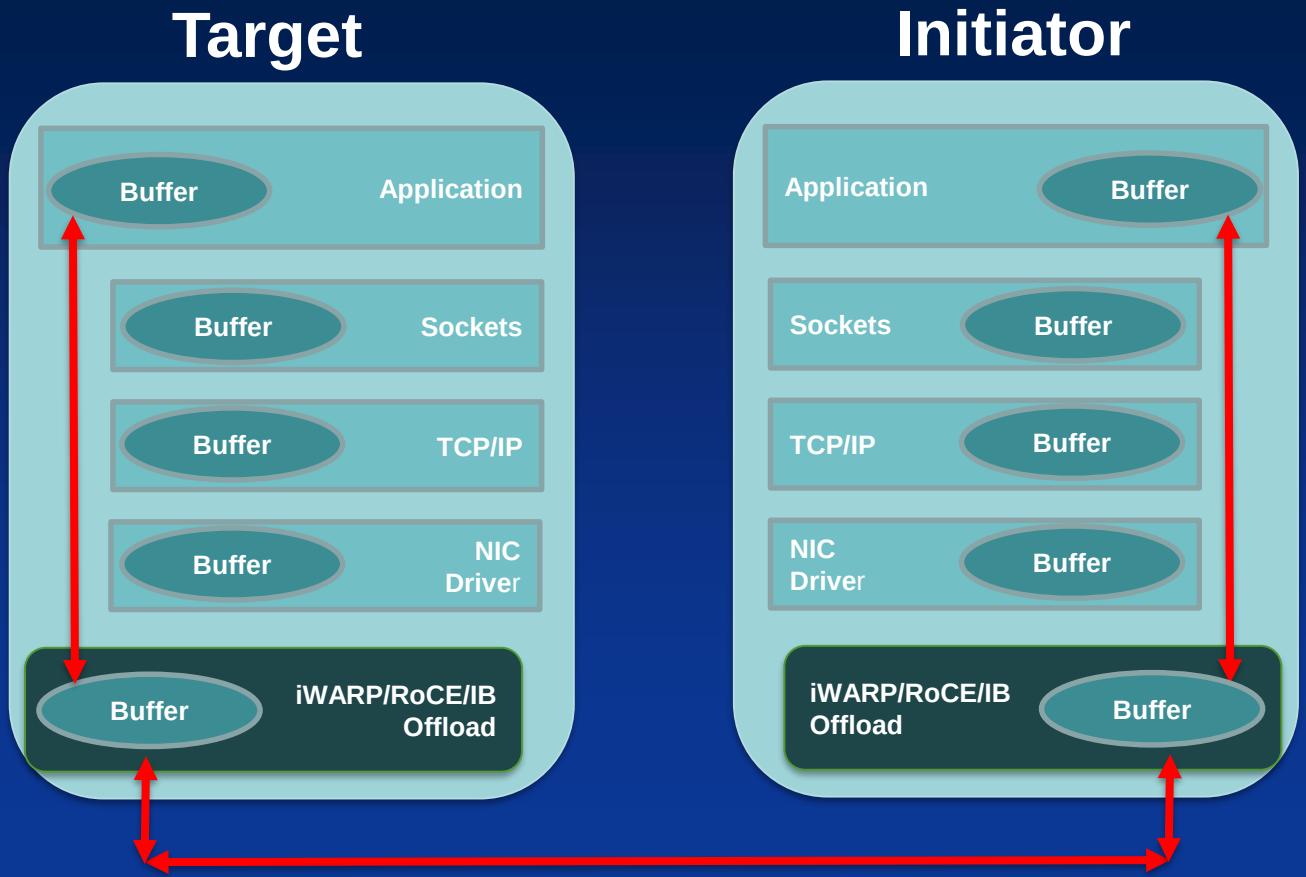
	DCB Required	Reach	IP routable	RDMA
FCoE	√	Rack, LAN		√
iSCSI	No	Rack, datacenter, LAN, MAN, WAN Wired, wireless	√	√
iWARP	No	Rack, datacenter, LAN, MAN, WAN Wired, wireless	√	√
RoCEv2	√	Rack, LAN, datacenter	√	√

Comparing Ethernet Options

- RDMA bypasses the host software stack
  - RoCEvn
  - iWARP
  - iSCSI with offload

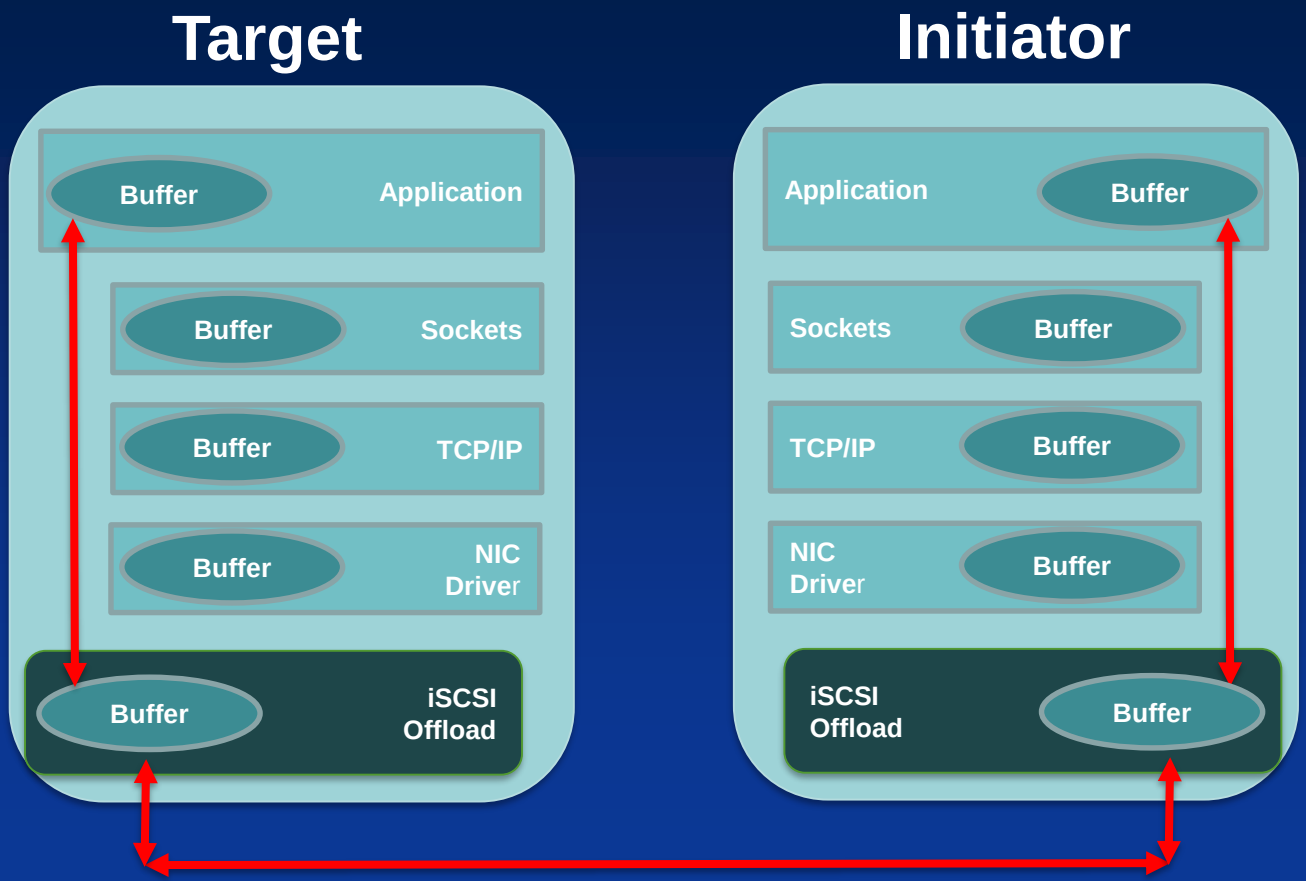
NVMe over RDMA fabrics

- Bypass
- RDMA



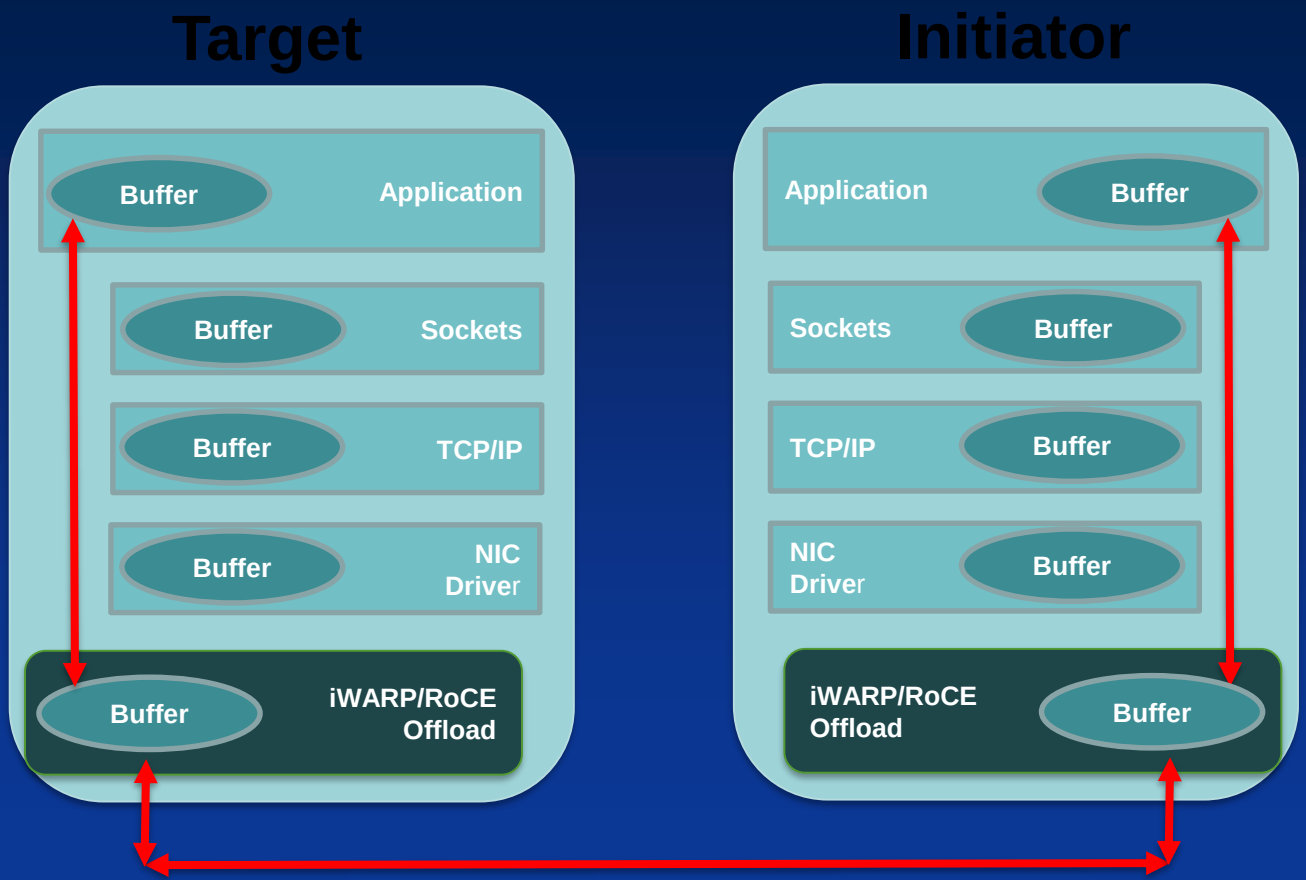
iSCSI with offload

- Bypass
- RDMA



iSER with offload

- Bypass
- RDMA



Chelsio T5 40GE Performance

	BW	IOPS	Latency	Comment
SMBD (SMB 3.X)	40GE			
NFSoRDMA	40GE			
FCoE	40GE			
iSCSI	40GE			
NVMf	40GE		NVMe+8µs	Linux 4.7-rc3

- List of http://www.chelsio.com links to the detailed setup

Summary

- API are evolving for optimal use of networked NVMe devices (NVMf)
  - High BW, High IOPS and low latency
- Chelsio 10/25/40/50/100GE adapters
  - Deliver high BW, High IOPS performance for SMB 3.X, NFSoRDMA, FCoE and iSCSI with NVMe
  - Concurrently: high BW, High IOPS, low latency NVMf

Questions?

Asgeir Eiriksson
asgeir@chelsio.com