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

Ásgeir Eiriksson

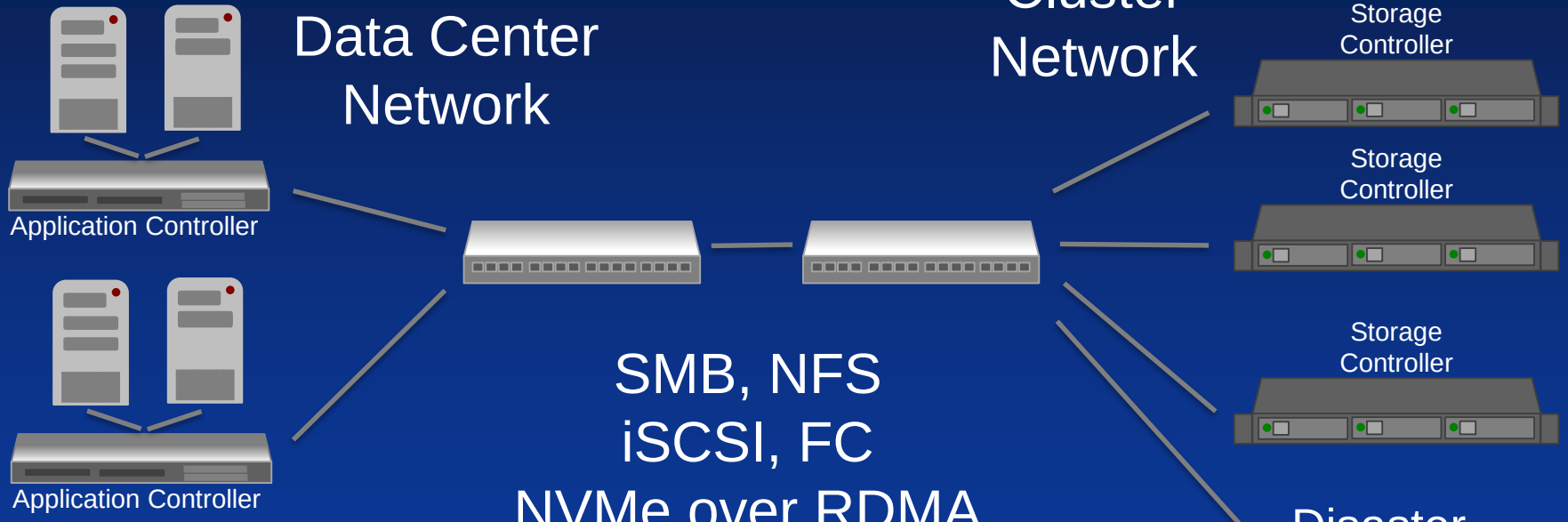
CTO

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.
- Need networking technology
  - that preserves existing storage product investment
  - that supports native NVMf API as ecosystem develops

Traditional Scale Out Storage



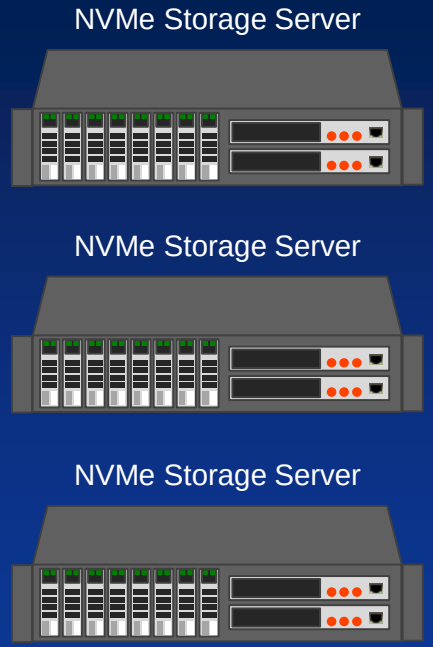
Traditional Scale Out Storage

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

Shared Server Flash

SMB, NFS
iSCSI, FC
NVMe over RDMA

Ethernet,
InfiniBand
Omni Path
Fabric



Disaster Recovery
60-300+ miles

Shared Server Flash

- Observations:
  - RDMA support required for lowest latency
  - Ethernet or IB or OmniPath fabrics possible
    - IB and OmniPath support RDMA
    - Ethernet has RoCEv1-v2, iWARP and iSCSI RDMA options
    - iSCSI offload has built-in RDMA WRITE
  - Disaster Recovery (DR) requires MAN or WAN
    - iWARP, iSCSI only options that support MAN and WAN

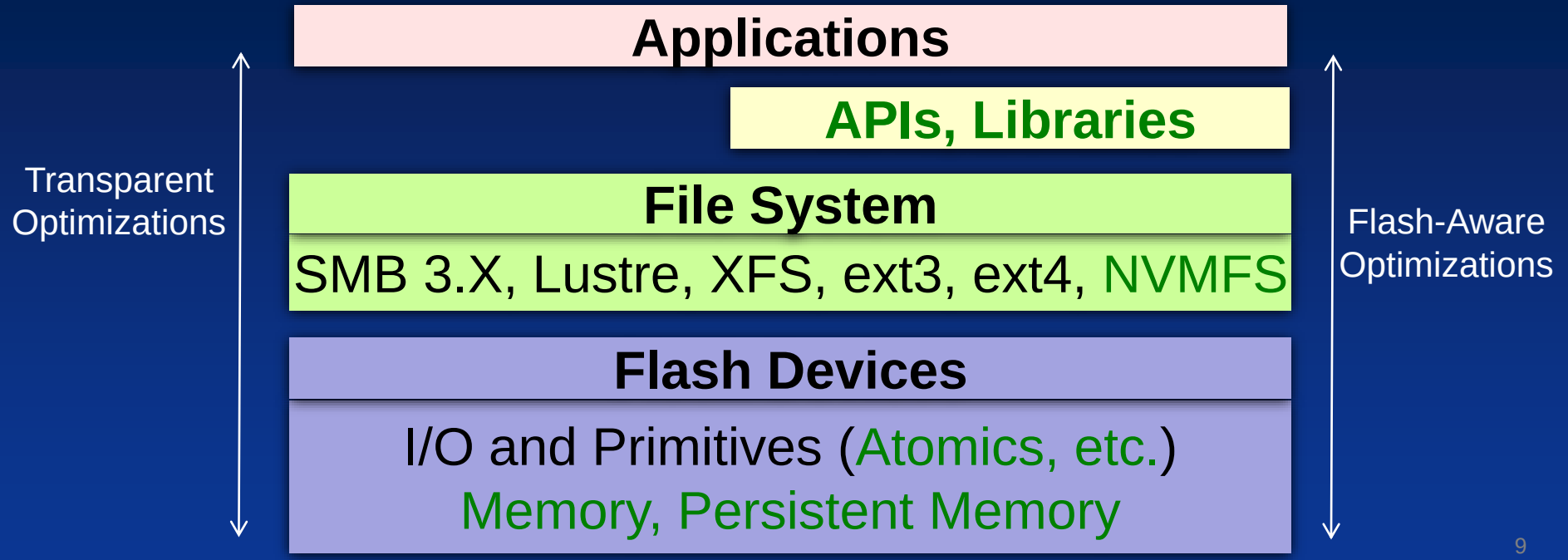
Comparing Ethernet Options

- iSCSI, iWARP
  - Use DCB when it is available but not required for high performance
  - Support for TCP congestion algorithms and IP ECN
- iSCSI
  - Has RDMA WRITE and emulates RDMA READ by using an RDMA WRITE from other end-point
  - Concurrent support for legacy soft-iSCSI
- RoCEv1, RoCEv2
  - 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
  - Rack and LAN technology no MAN or WAN
- Ethernet
  - Ubiquitous
  - Pause and Prioritized Pause (PPC) for lossless operation that is supported by some switches and fewer routers
  - Flow Control and Reliability at higher layer e.g. TCP, and IB Transport Layer for RoCEv1 and RoCEv2

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

- Either: Preserve software investment
- And/or: Alternatively jump directly to native NVMe/NVMf API

- Strong preference: preserve investment while at the same time offer competitive NVMe technology support

Comparing Ethernet Options

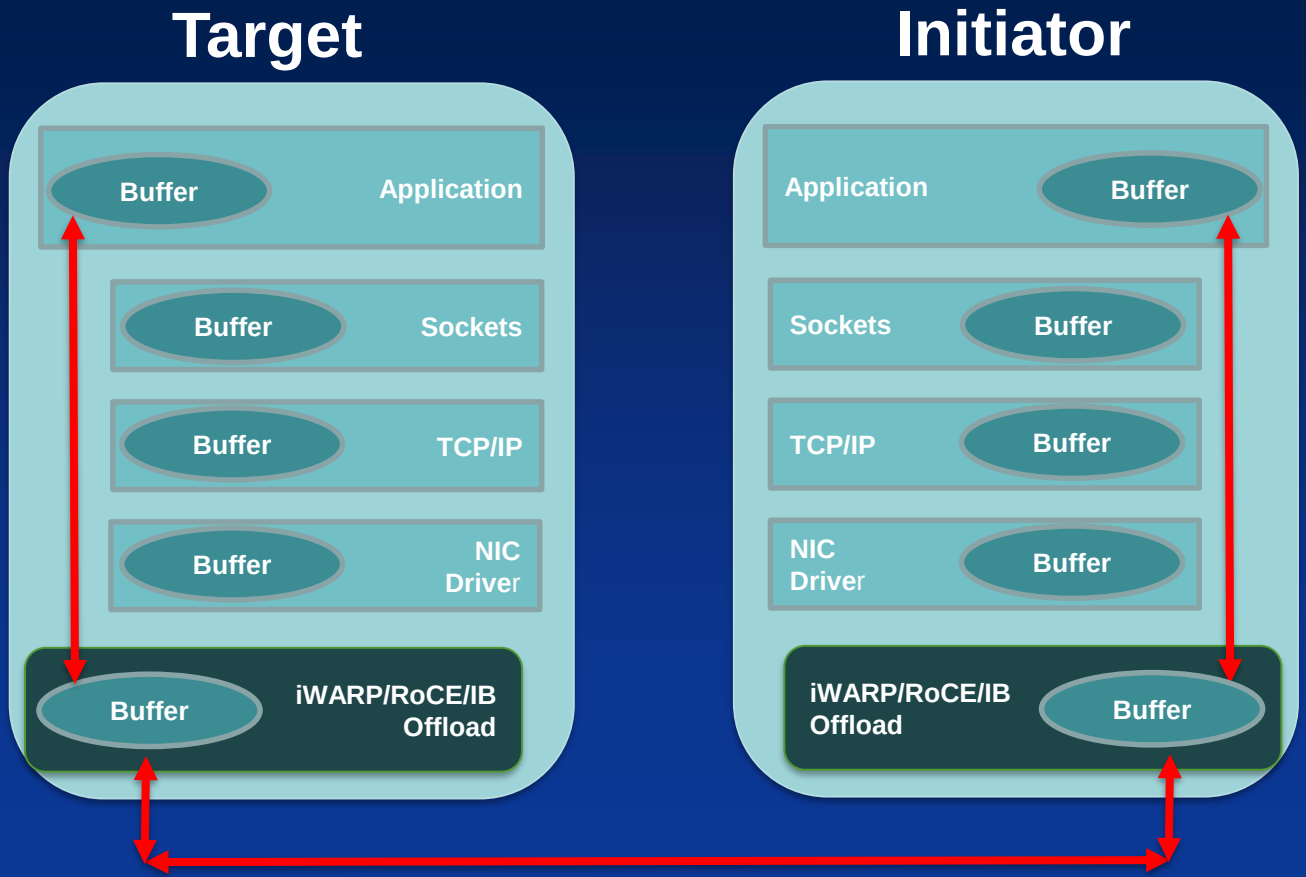
	DCB Required	Reach	IP routable	RDMA
FCoE	√	Rack, LAN		√
iSCSI	No	Rack, datacenter, LAN, MAN, WAN Wired, wireless	√	√
iWARP	No	Rack, datacenter, LAN, MAN, WAN Wired, wireless	√	√
RoCEv2	√	Rack, LAN, datacenter	√	√

Comparing Ethernet Options

- RDMA bypasses the host software network stack
  - RoCEv1, RoCEv2
  - iWARP
  - iSCSI with offload

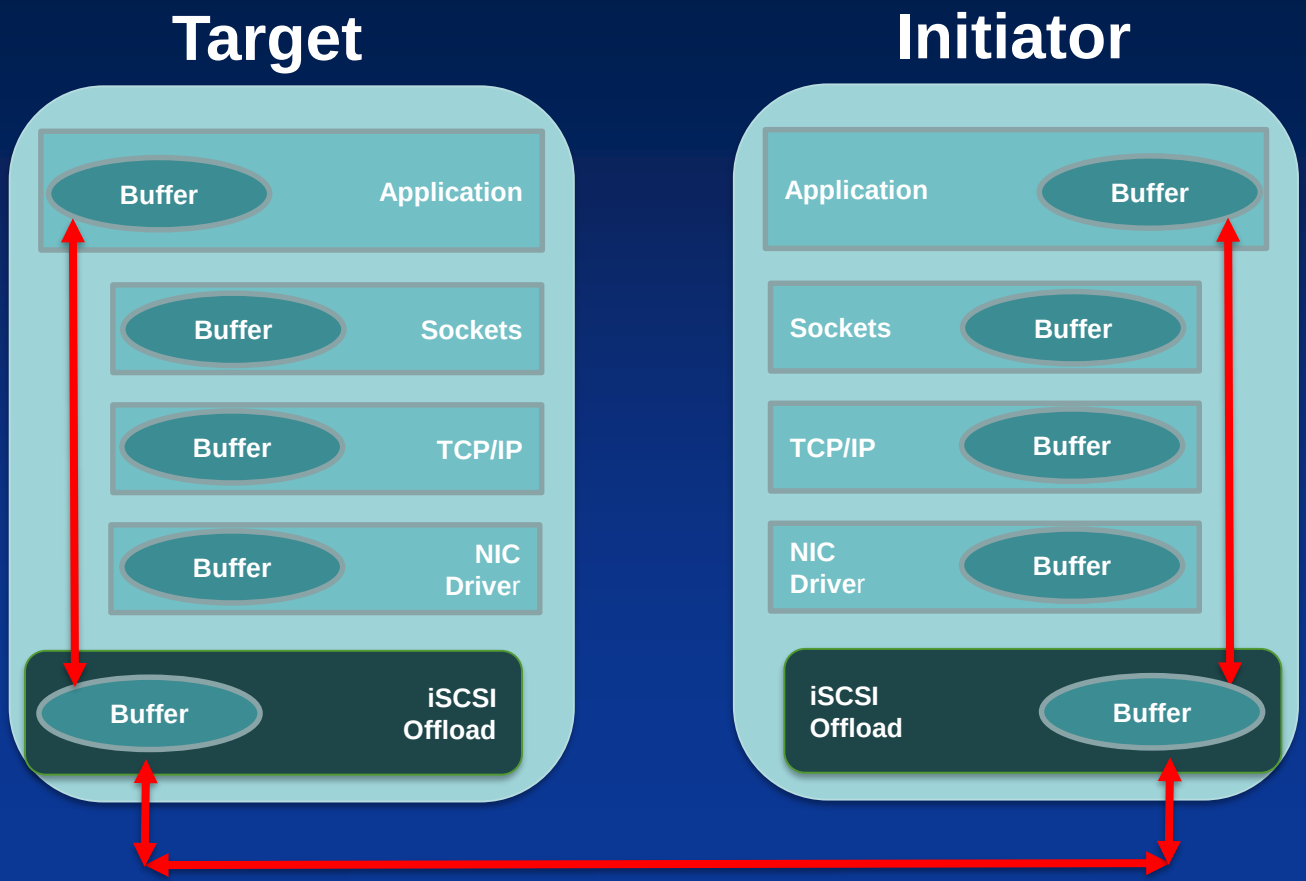
NVMe over RDMA fabrics

- Bypass
- RDMA



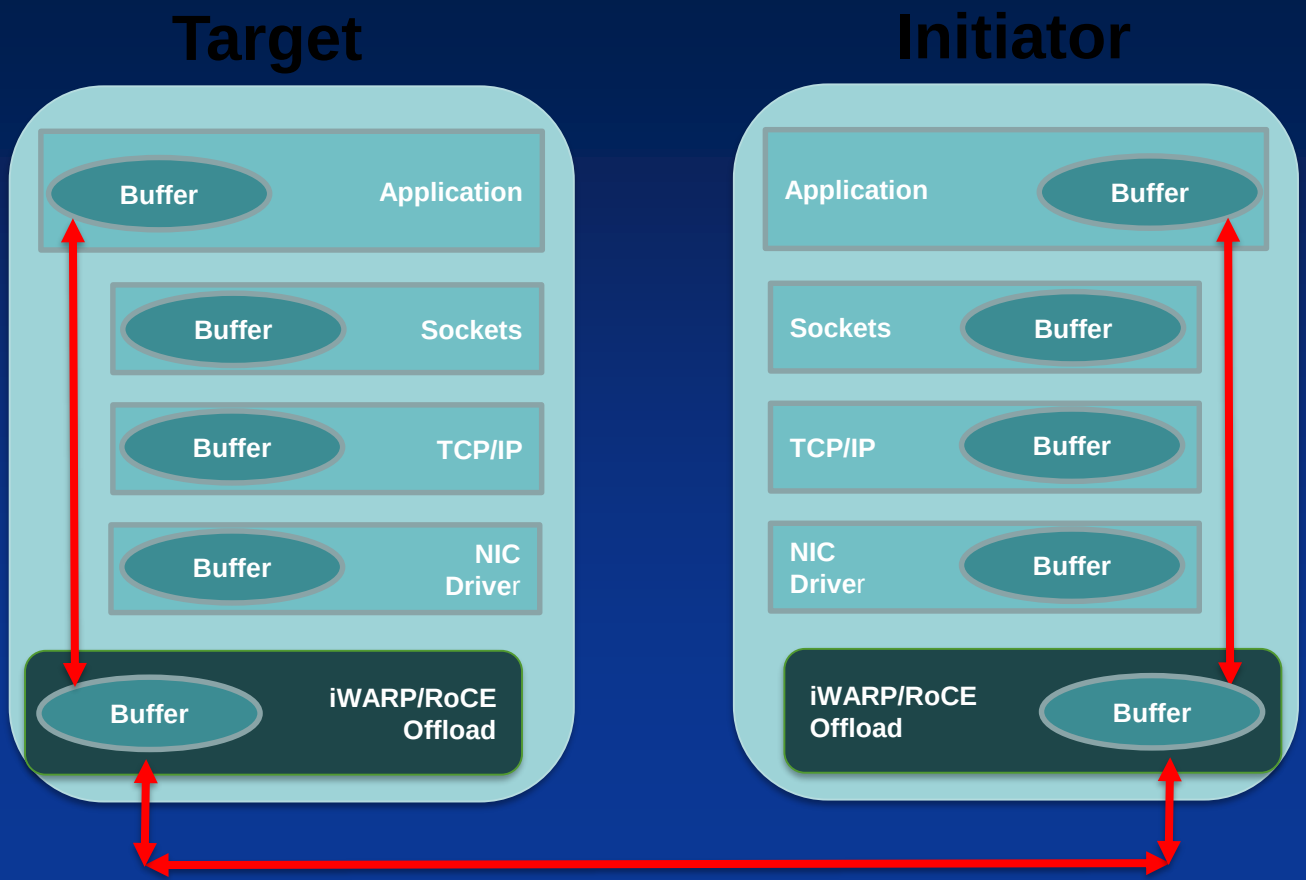
iSCSI with offload

- Bypass
- RDMA



iSER with offload

- Bypass
- RDMA



Chelsio T5 40GE Performance

- Chelsio T5 and T6 Ethernet Adapters
  - 10GE, 25GE, 40GE, 50GE, 100GE support
  - Concurrent support for:
    - NVMf (NVMe over RDMA)
    - SMB 3.X
    - NFSoRDMA
    - iSCSI
    - FCoE

Chelsio T5 40GE Performance

	BW	IOPS	Comment
SMBD (SMB 3.X)	40GE	5M	Microsoft Storage Spaces Direct 16 node aggregate rate with NVMe and SATA SSD. See 1.

1. https://blogs.technet.microsoft.com/filecab/2016/07/26/storage-iops-update-with-storage-spaces-direct/

Chelsio T5 40GE Performance

	BW	IOPS @4KB	Latency	Comment
FCoE	40GE			
iSCSI	40GE	1M		Open-iSCSI
NVMf	40GE	1M	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
- You can preserve investment while adopting NVMf

Questions?

Asgeir Eiriksson
asgeir@chelsio.com