Competing Technologies and Architectures for Networked Flash Storage

Ethernet/InfiniBand/OmniPath/PCI/FibreChannel/SAS

Asgeir Eiriksson

Chelsio Communications

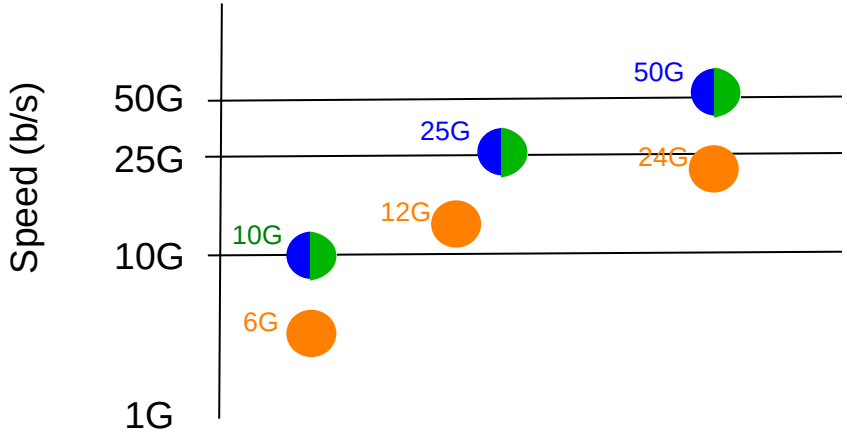
Introduction

- API are evolving for optimal use of SSD
- FC and SAS falling behind the speed curve
- Ethernet, IB and OmniPath on same PHY curve
- PCI on different slightly slower PHY curve
- Ethernet, IB, and OmniPath
  - Have different reach
  - Same protocol stack efficiencies

Introduction: speeds and feeds

	Bandwidth (Gbps)	Reach
SAS	3, 6, 12	Rack
Fibre Channel	4, 8, 16, 32	Rack, Data center
PCI x1/2/4/8/16	8, 16, 32, 64,128	Rack
Ethernet	1, 2.5, 5, 10, 25, 40, 50, 100	Rack, Data Center, LAN, MAN, WAN
Infiniband	8, 16, 32, 56, 112	Rack, Data Center
OmniPath	100, 200	Rack, Data Center

PHY SERDES (single lane) curves

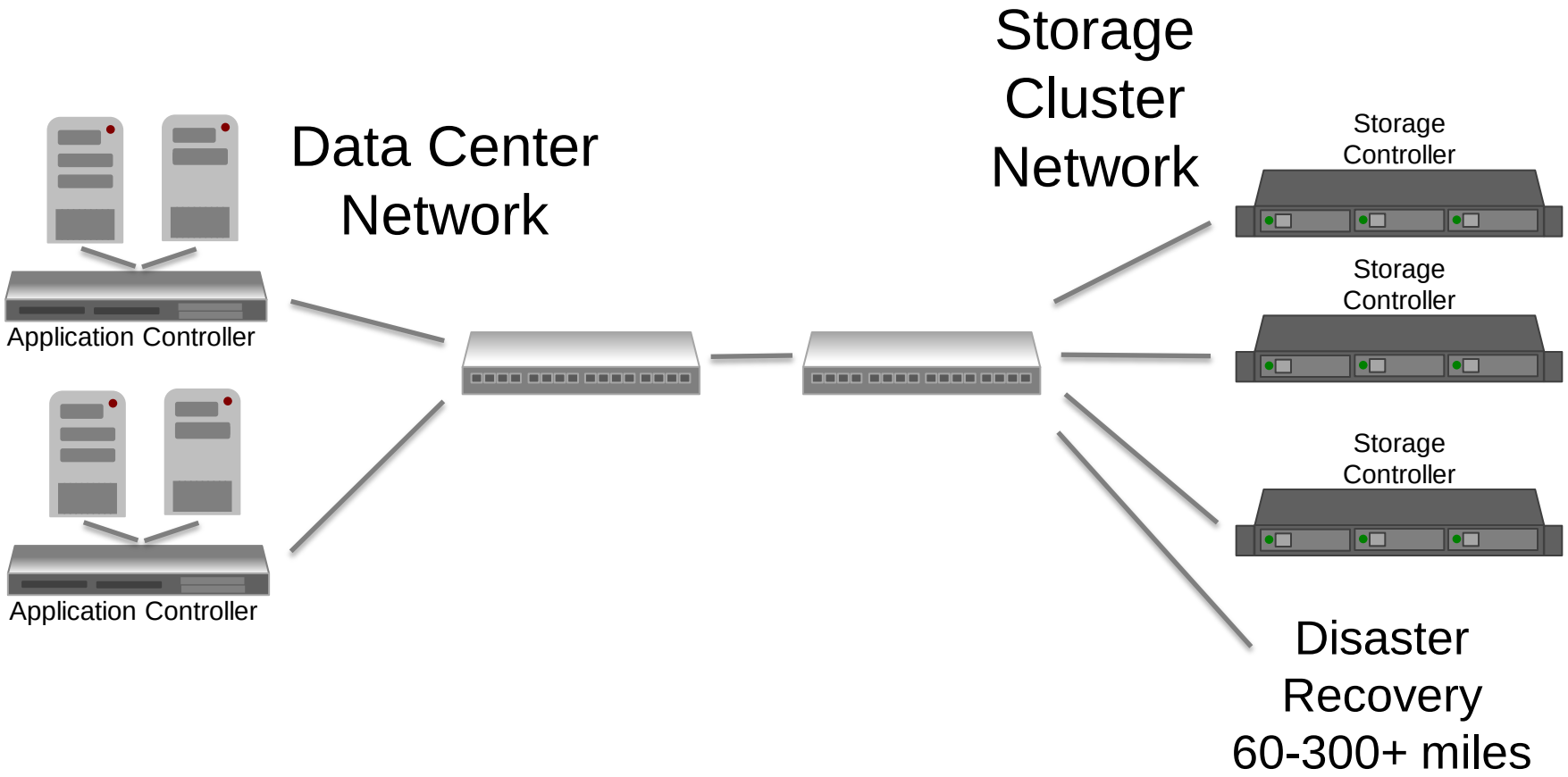


	2010	2015	2018
SAS	6	12	24
Infiniband	10	25	50
Ethernet	10	25	50

Infiniband and Ethernet

- Same PHY curve
- Same speed curve

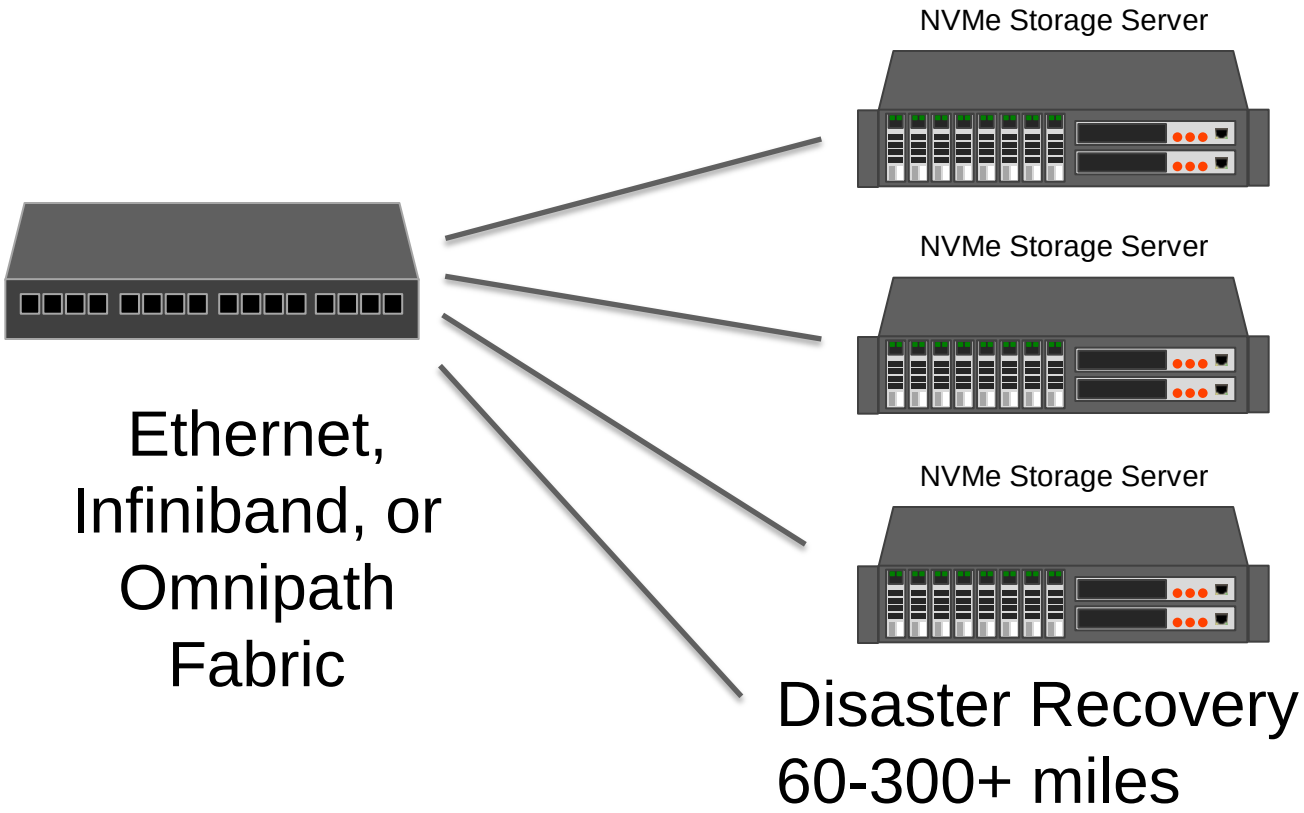
Traditional Scale Out Storage



Traditional Scale Out Storage

- Preserves software investment
- Realizes some of the SSD speedup benefits
- Disaster Recovery (DR) requires MAN or WAN

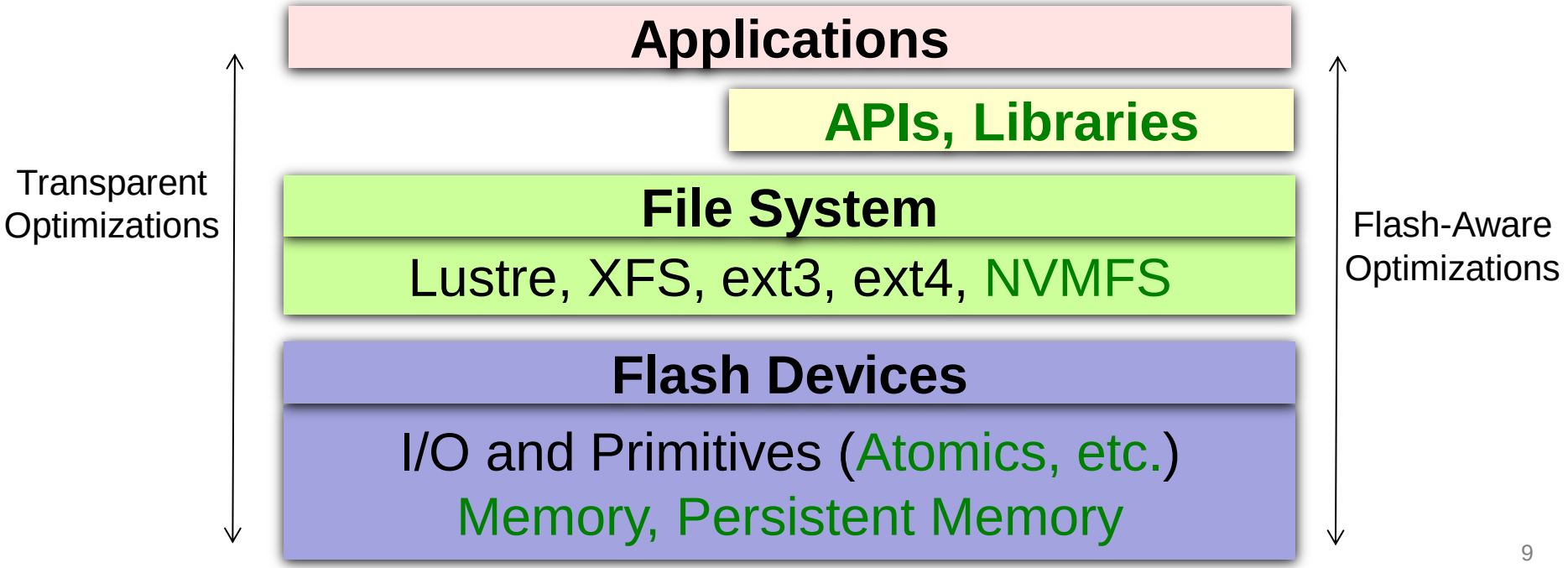
Shared Server Flash



Shared Server Flash

- Ethernet or IB or OmniPath fabric
- PCIe fabric not sufficient reach or scaling
- RDMA required for sufficient efficiency
  - IB and OmniPath use RDMA
  - Ethernet has RoCE, iWARP and iSCSI with RDMA
- Disaster Recovery (DR) requires MAN or WAN

API


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

API

- Preserve software investment
- Alternatively jump directly to native SSD API

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

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

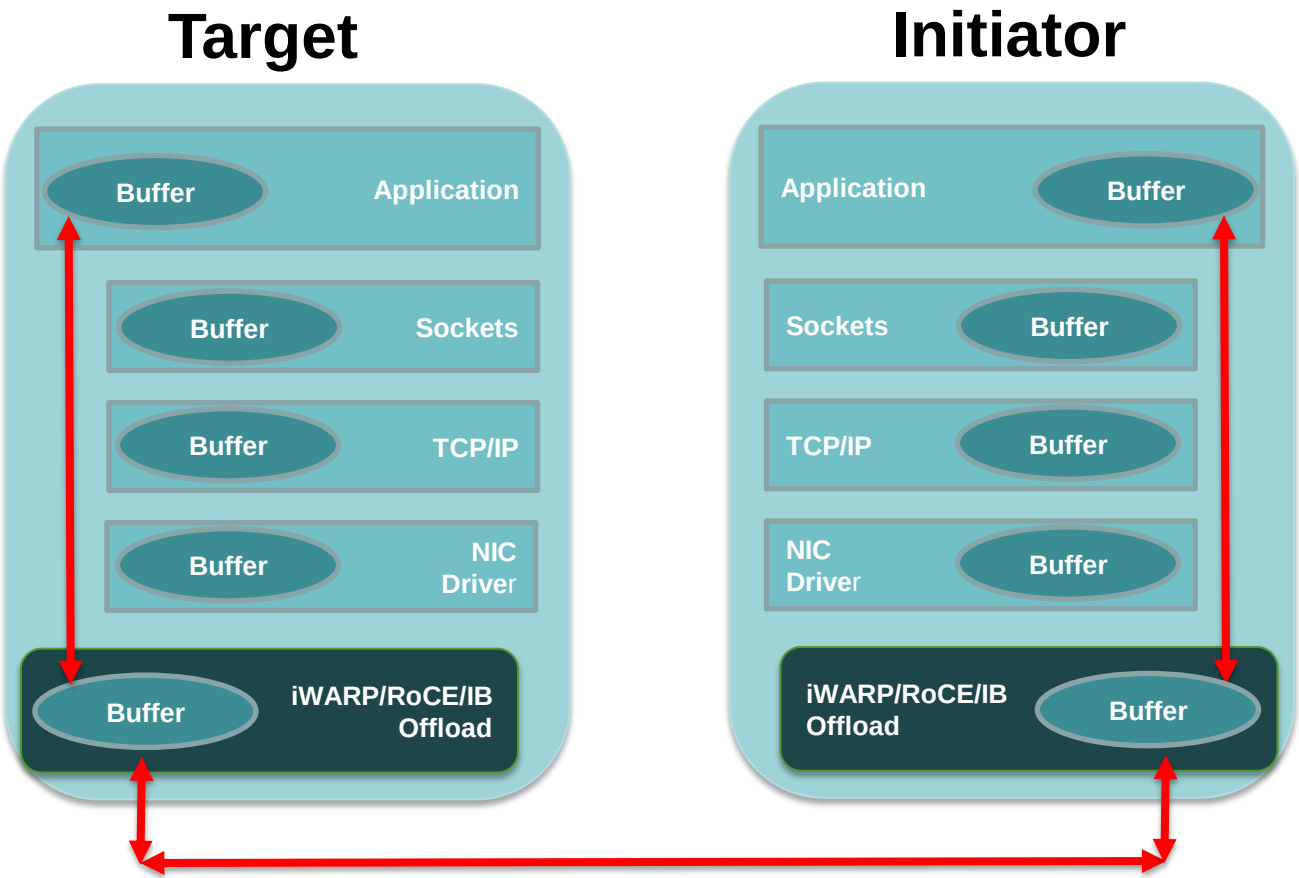
- iSCSI, iWARP
  - Use DCB when it is available but not required for high performance
- iSCSI
  - Has RDMA WRITE and accomplishes RDMA READ by using an RDMA WRITE from other end-point
  - Concurrent support for legacy soft-iSCSI

Comparing Ethernet Options

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

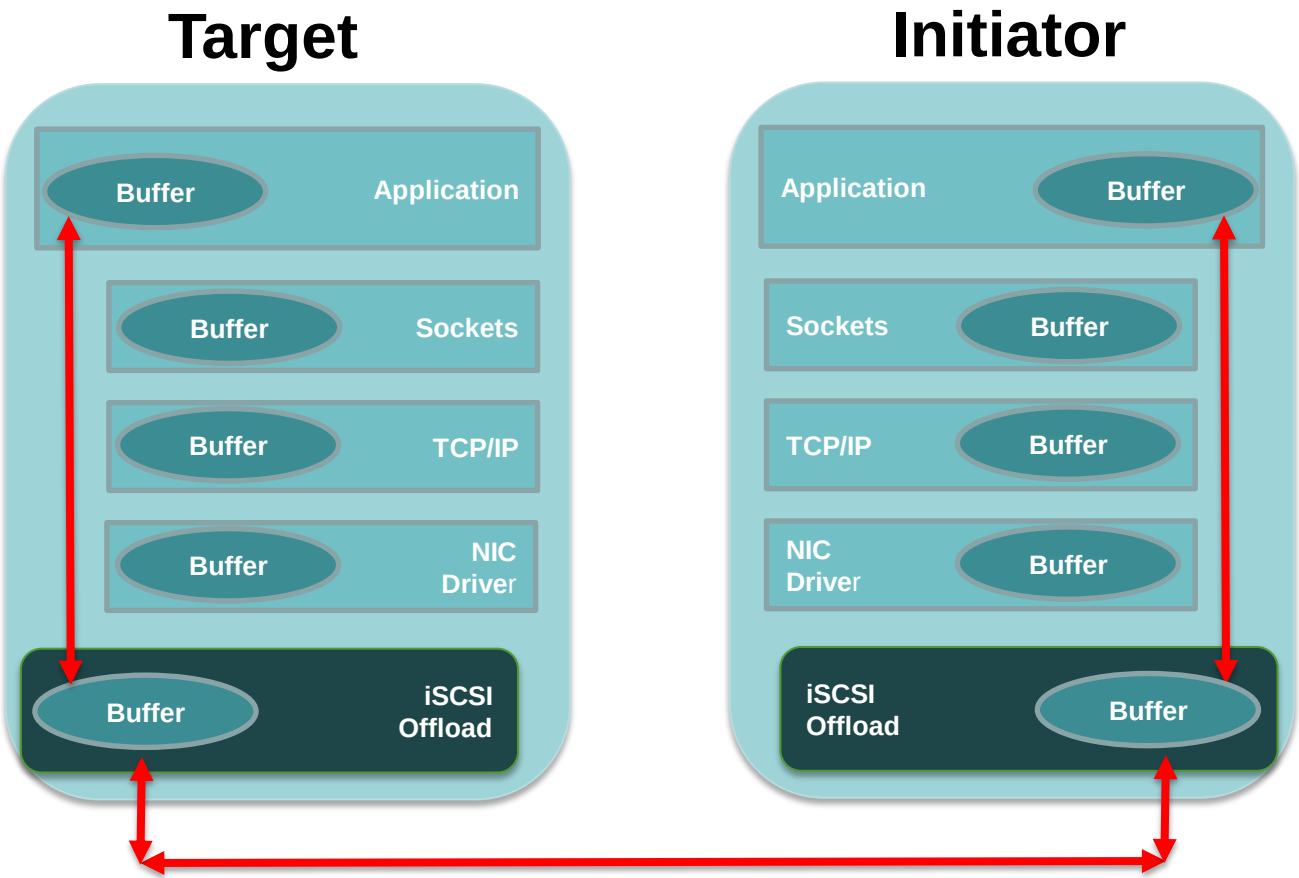
NVMe over RDMA fabrics

- Bypass
- RDMA



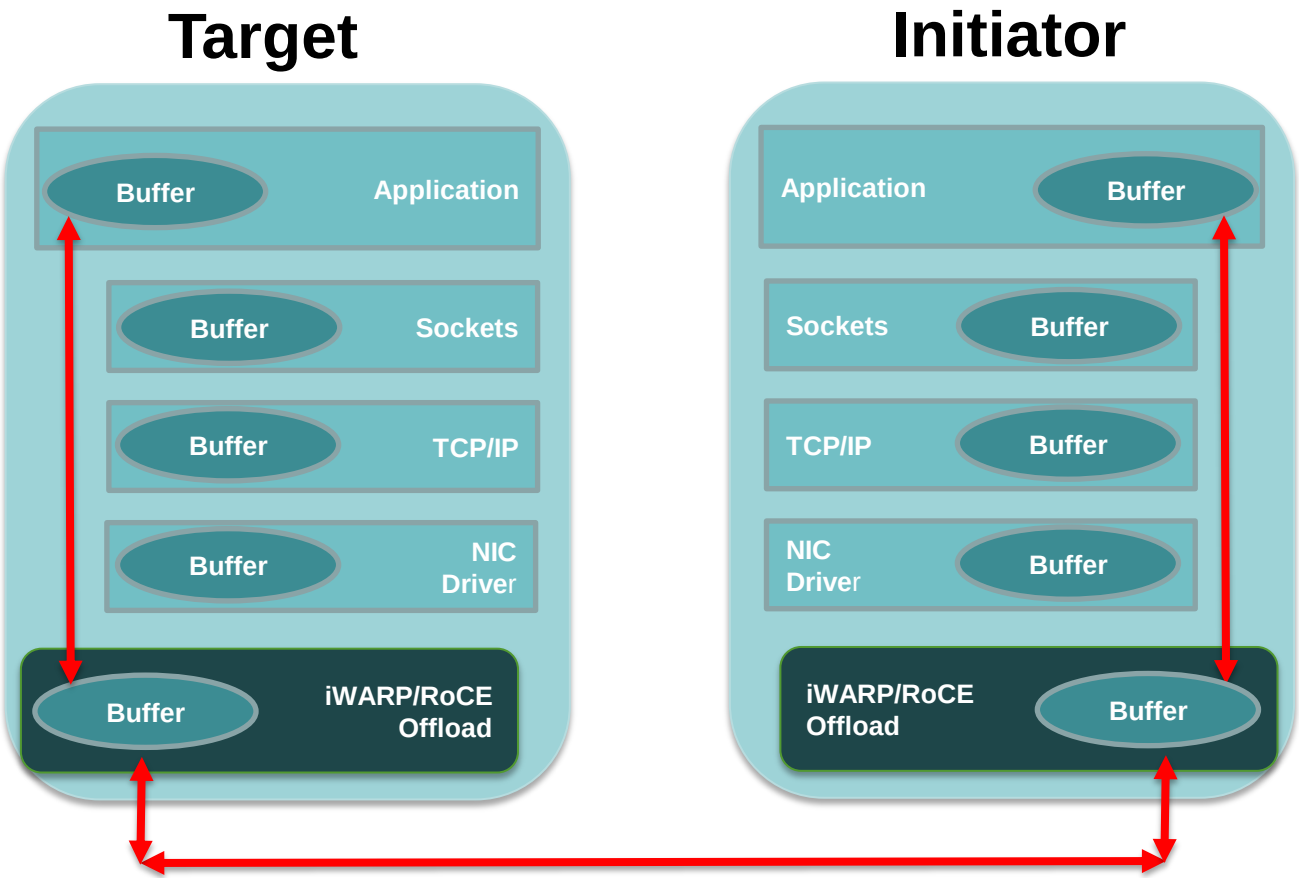
iSCSI with offload

- Bypass
- RDMA

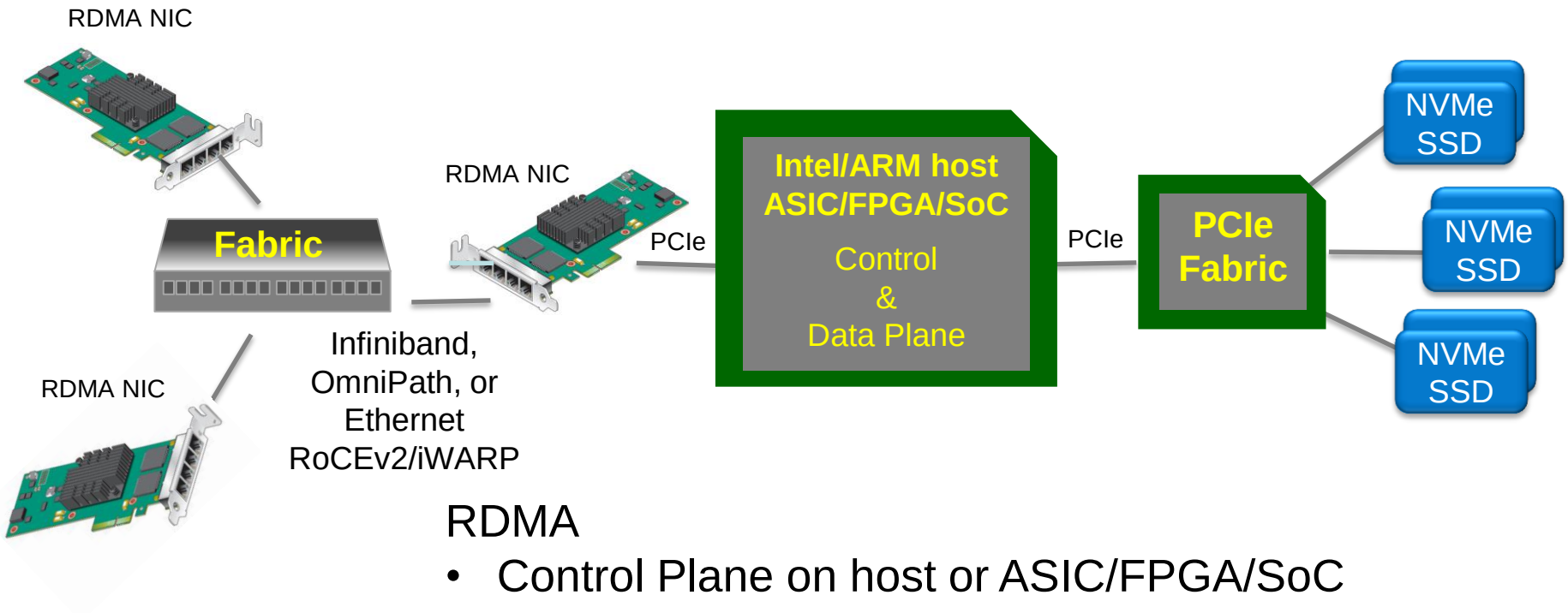


iSER with offload

- Bypass
- RDMA



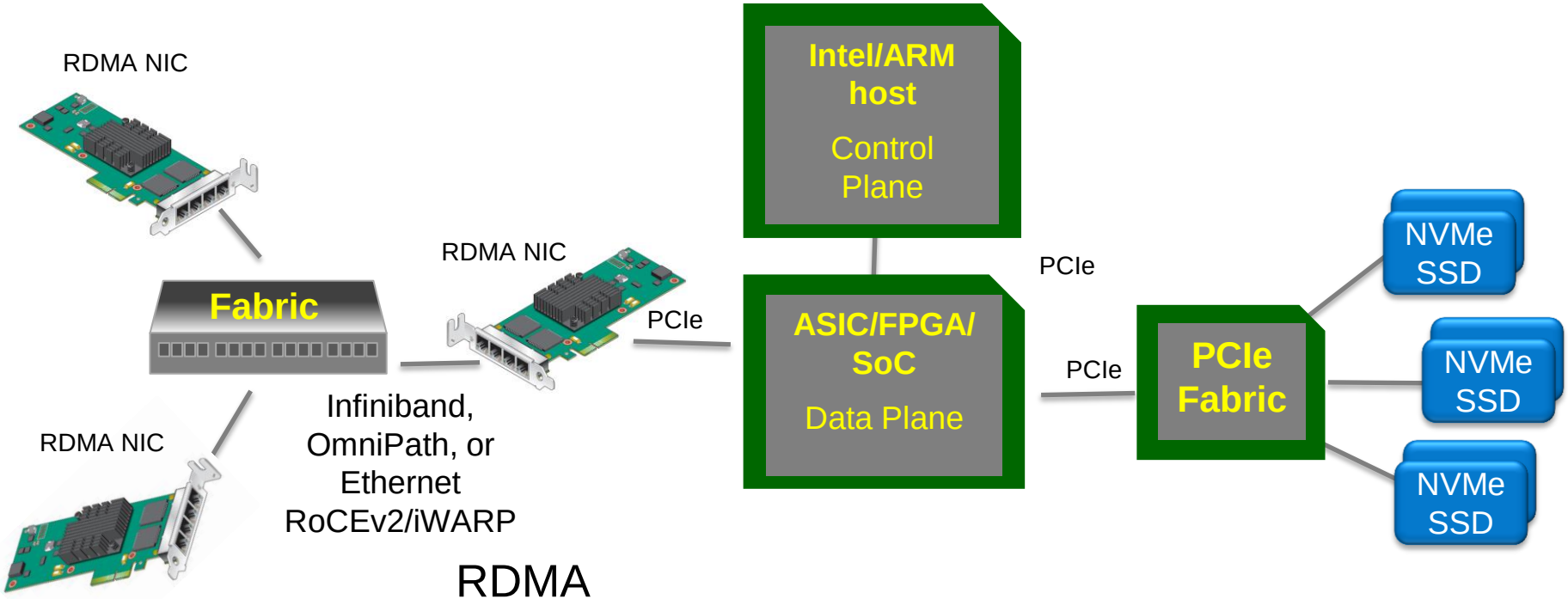
NVMe over fabrics Option 1



RDMA

- Control Plane on host or ASIC/FPGA/SoC
- Data Plane PCIe-host-PCIe or PCIe only

NVMe over fabrics Option 2



RDMA

- Control Plane Intel/ARM host
- Data Plane PCIe-ASIC/FPGA/SoC-PCIe

NVMe over fabrics comparison

- Option 1
  - Flexible
  - Extra latency incurred by copy/copies
- Option 2
  - Minimizes latency by removing host and host memory system from data path

Summary

- API are evolving for optimal use of SSD
- Ethernet, IB, and OmniPath
  - On same SERDES PHY (single lane) curve
  - Have different reach
  - Same protocol stack efficiencies

Questions?

Asgeir Eiriksson

asgeir@chelsio.com