13th ANNUAL WORKSHOP 2017

EXPERIENCES WITH NVME OVER FABRICS

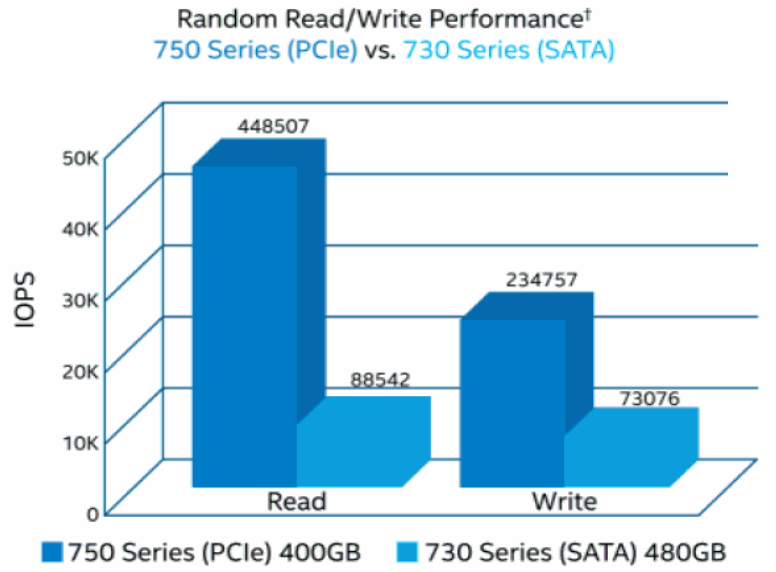
Parav Pandit, Oren Duer, Max Gurtovoy

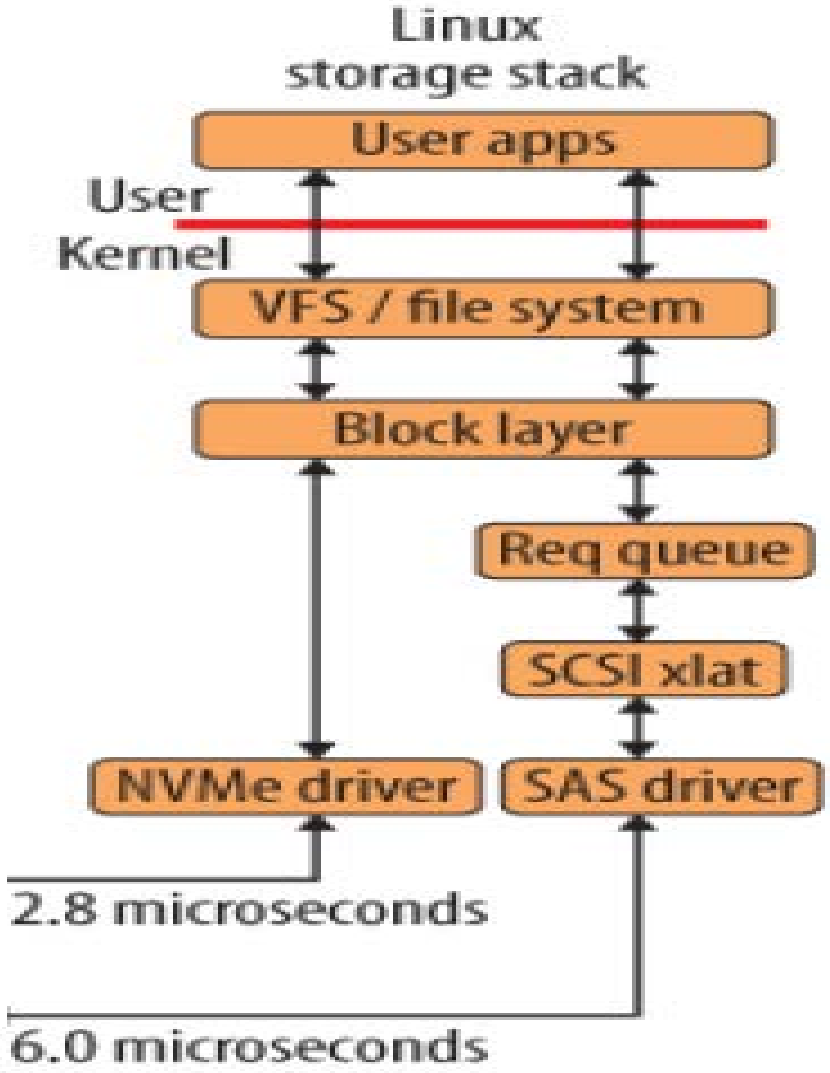
Mellanox Technologies

[31 March, 2017]

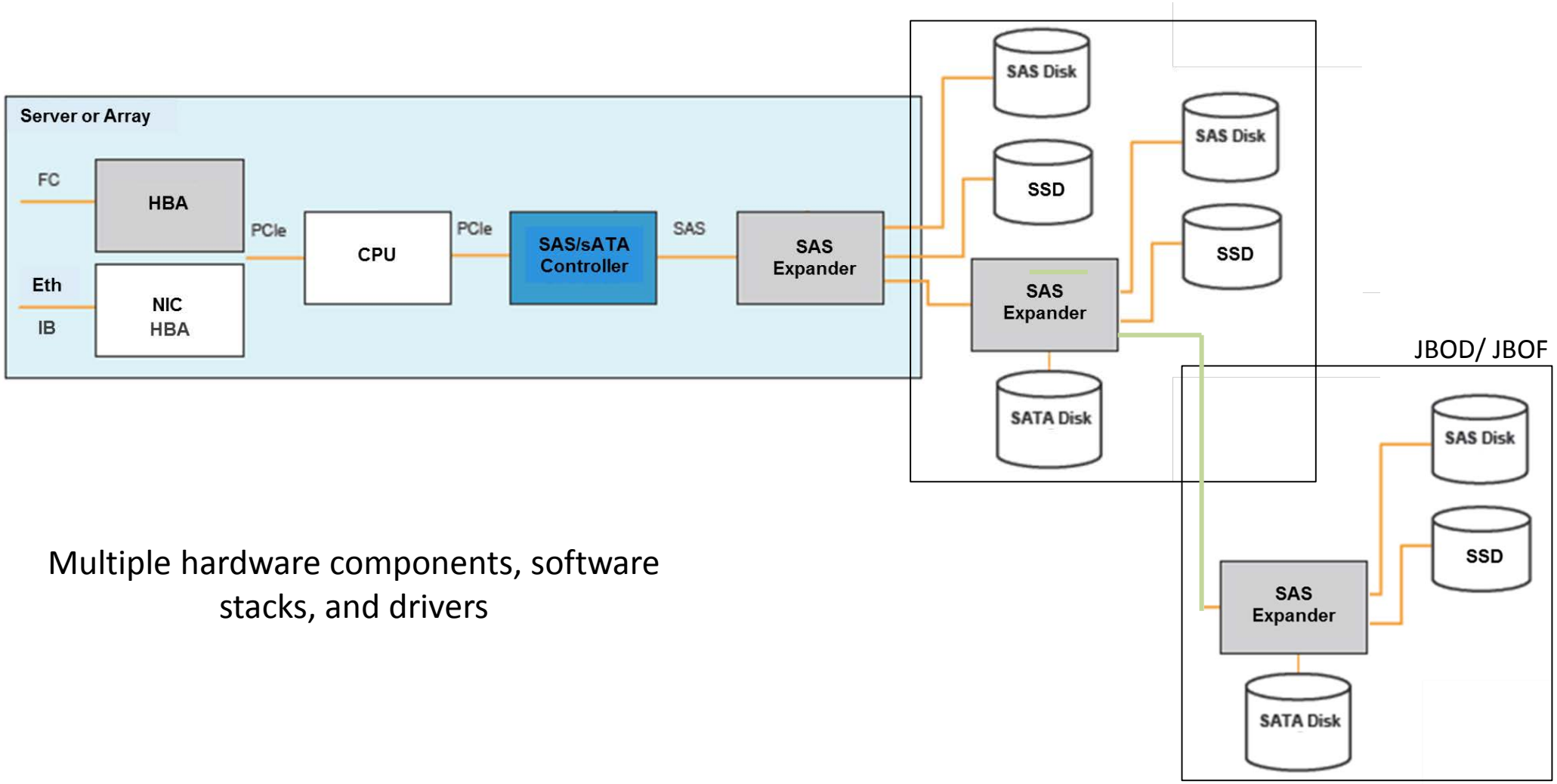
BACKGROUND: NVME TECHNOLOGY

- **Optimized for flash and next-gen NV-memory**
  - Traditional SCSI interfaces designed for spinning disk
  - NVMe bypasses unneeded layers
- **NVMe Flash Outperforms SAS/SATA Flash**
  - 2x-2.5x more bandwidth, 40-50% lower latency
  - Up to 3x more IOPS



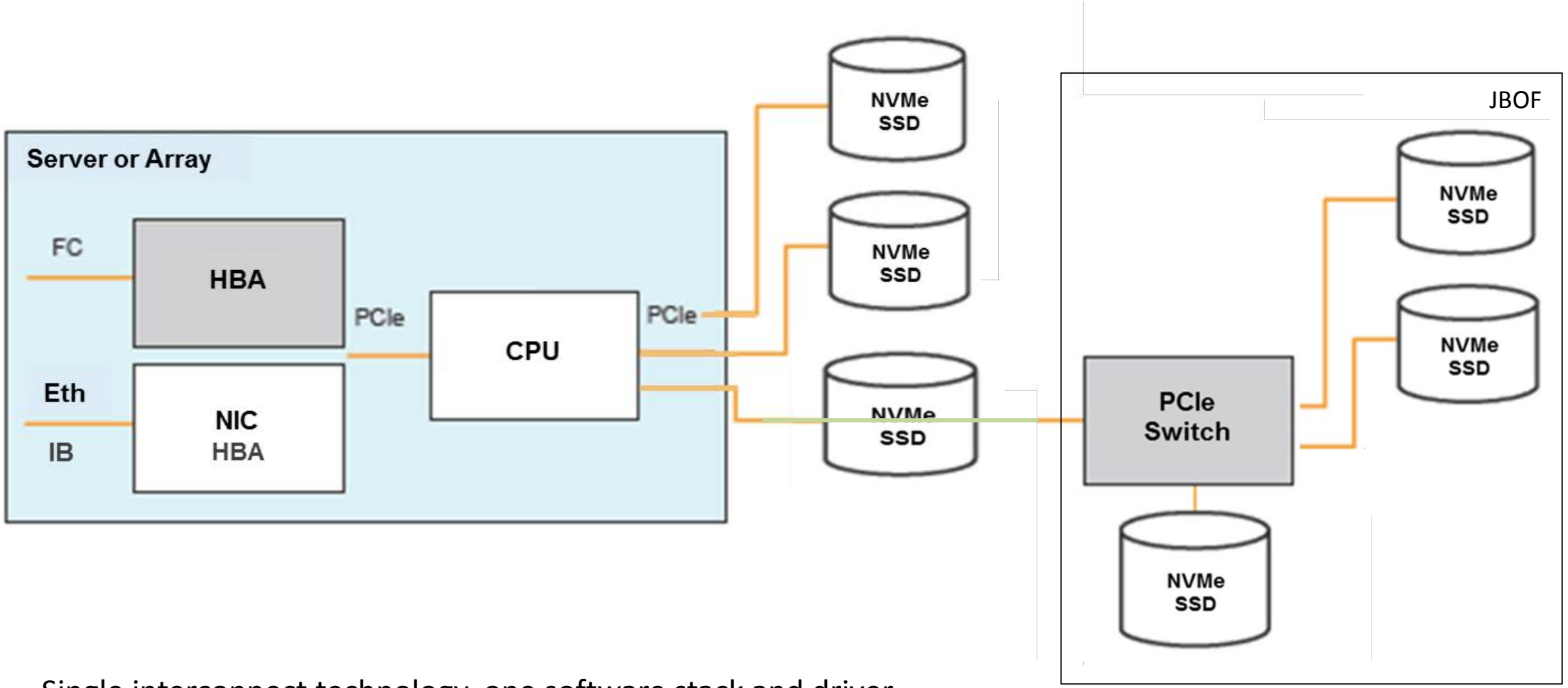


TRADITIONAL SAS/SATA STORAGE ARCHITECTURE



Multiple hardware components, software stacks, and drivers

MORE EFFICIENT AND COST-EFFECTIVE NVME ARCHITECTURE

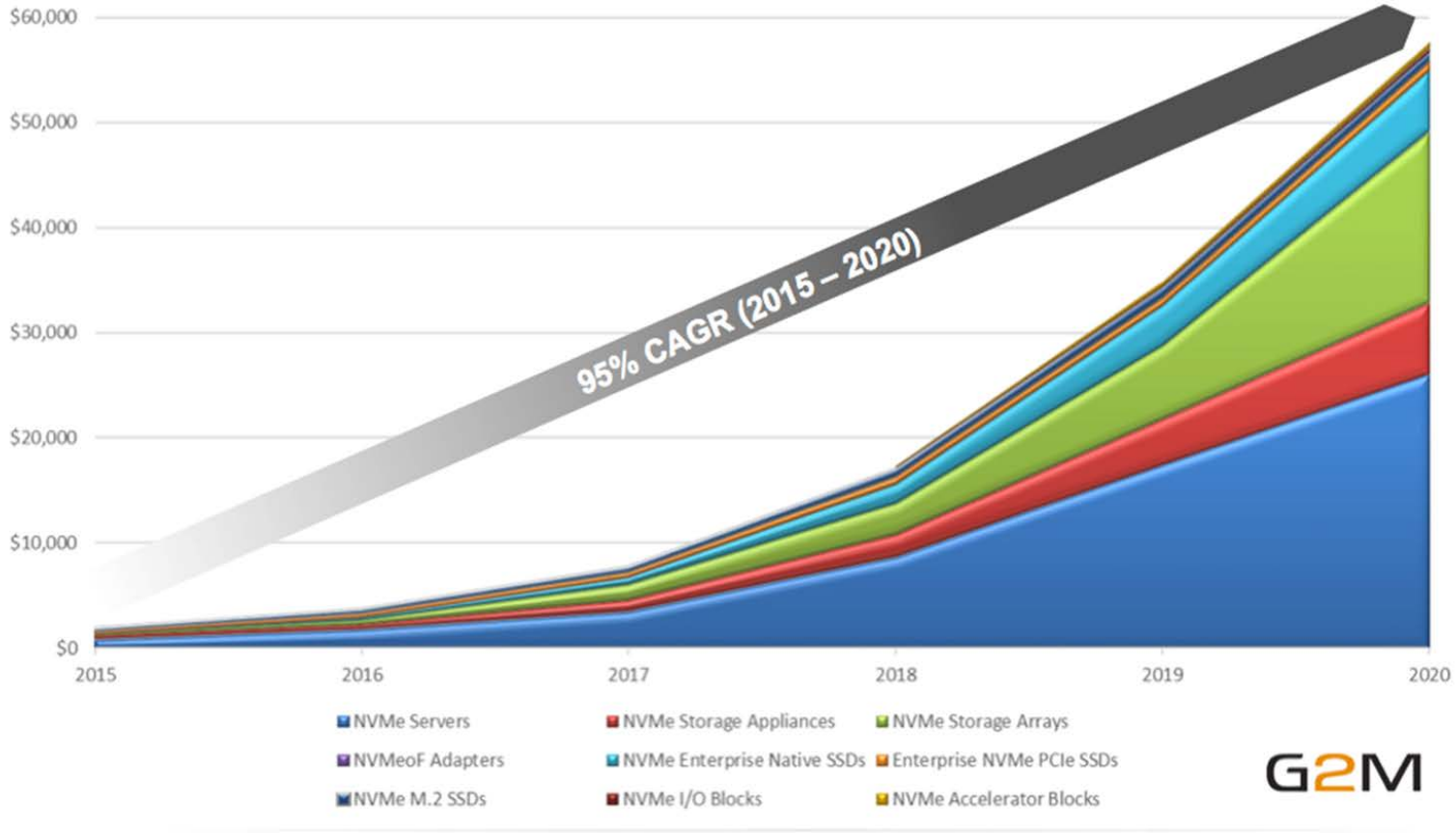


Single interconnect technology, one software stack and driver

NVME MARKET PROJECTED TO GROW TO $57B BY 2020

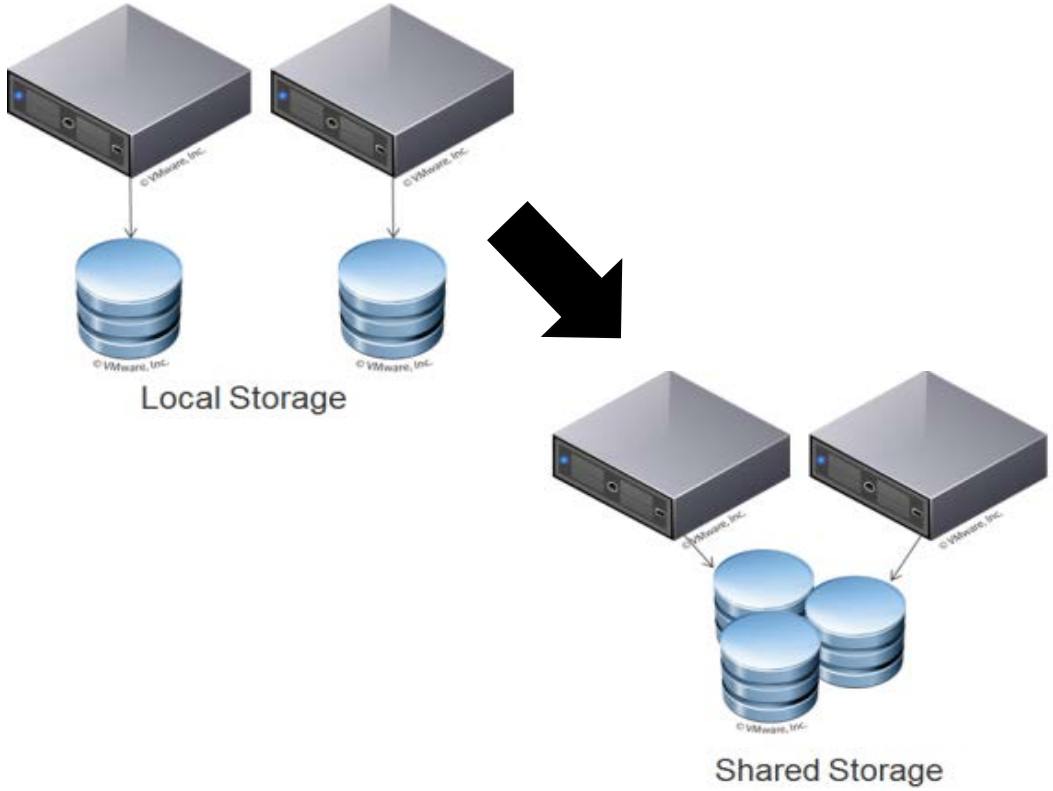
- **>50% of enterprise servers and storage appliances will support NVMe by 2020**
- **~40% of all-flash arrays will be NVMe-based by 2020**
- **Shipments of NVMe SSDs will grow to 25+ million by 2020**

- **740,000 NVMe-oF adapter shipped by 2020**
- **RDMA NICs will claim >75% of the NVMe-oF market**



“NVME OVER FABRICS” ENABLES STORAGE NETWORKING OF NVME DEVICES

- **Sharing NVMe based storage across multiple servers/CPUs**
  - Better utilization: capacity, rack space, power
  - Scalability, management, fault isolation
- **NVMe over Fabrics industry standard developed**
  - Version 1.0 completed in June 2016
- **RDMA protocol is part of the standard**
  - NVMe-oF version 1.0 includes a Transport binding specification for RDMA
  - InfiniBand or Ethernet(RoCE)



SOME NVME-OF DEMOS AT FMS AND IDF 2016

Flash Memory Summit

- **E8 Storage**
- **Mangstor**
  - With initiators from VMs on VMware ESXi
- **Micron**
  - Windows & Linux initiators to Linux target
- **Newisis (Sanmina)**
- **Pavilion Data**
  - in Seagate booth

Intel Developer Forum

- **E8 Storage**
- **HGST (WD)**
  - NVMe-oF on InfiniBand
- **Intel: NVMe over Fabrics with SPDK**
- **Newisis (Sanmina)**
- **Samsung**
- **Seagate**

SOME NVME-OF PRODUCTS IN THE MARKET TODAY

Pavilion Data Systems

Excelero

E8 storage
Rack-Scale Flash. No Compromise.

HUAWEI

Tue, Aug 30, 2016

Mellanox and Huawei Advance RDMA Technology with Leading-Edge NVMe Over Fabrics Solution

SHANGHAI, CHINA – Aug. 31, 2016 – Mellanox® Technologies, Ltd. (NASDAQ: MLNX), a leading supplier of high performance cloud and storage networking solutions, today announced that it will preview a new leading-edge NVMe-oF™ (NVM Express® over Fabrics) solution, jointly developed by Mellanox and Huawei, at this week's HUAWEI CONNECT Conference (HCC 2016) in Shanghai, Aug. 31 – Sept. 2.

Mangstor

E8 STORAGE AND MELLANOX COLLABORATE TO DRIVE HIGH PERFORMANCE CENTRALIZED NVME STORAGE ARRAYS

August 8, 2016

E8 Storage and Mellanox® Technologies Ltd. (NASDAQ: MLNX) today announced a technology partnership designed to deliver an end-to-end shared NVMe solution for high-performance enterprise storage applications. The integration of E8 Storage's rack scale flash architecture with the Mellanox Remote Direct Memory Access (RDMA) network adapters enables converged networking with very low latency and very high throughput and bandwidth. This announcement coincides with today's launch of E8 Storage's D24 flash appliance – the industry's first centralized, highly available NVMe solution.

Mangstor and Mellanox Demonstrate All-Flash SSD Storage Array achieving 10GB/s using NVMe over Fabrics

April 14, 2015 / in Press Releases / by Support

Ethernet Technology Summit, Santa Clara, CA – April 14, 2015 – Mangstor, a leading developer of Intelligent Storage Solutions for Web-Scale and large enterprise data centers, today announced the company is working with Network vendor, Mellanox Technology, to create the next generation of NVM Express™ shared storage solutions. The NMX-Series All-Flash-Array products build on Mangstor's leading performance MX6300 SSDs, combined with Mellanox's industry leading line of ConnectX® VPI adapters with support for InifiniBand and RDMA over Converged Ethernet (RoCE) providing over 10GB/s bandwidth at latencies comparable to local server based NVMe devices.

200Gbps Data Transfer using NVMe over Fabric at SC16: Liqid, Mellanox and EchoStreams Collaboration

LIQID

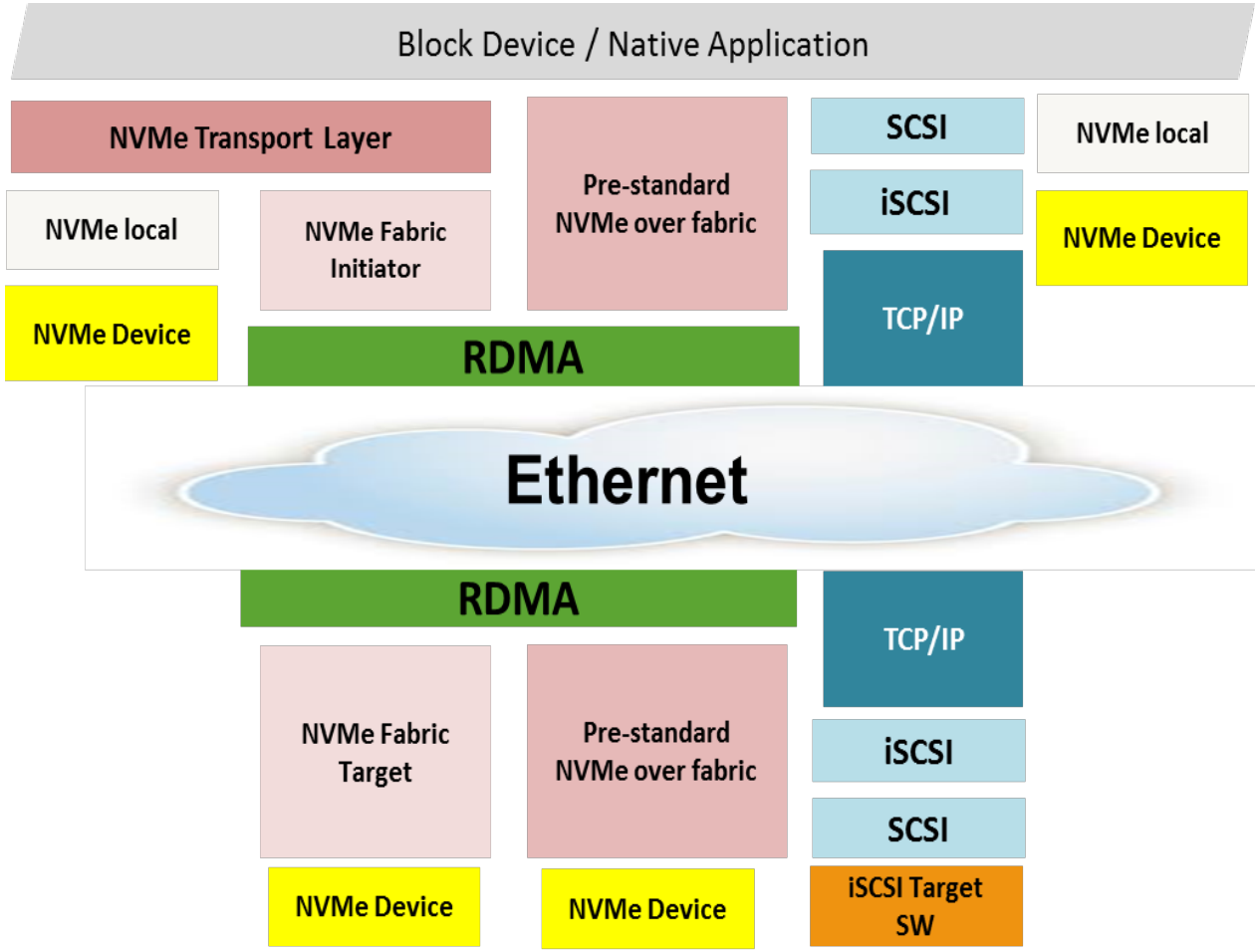
Demonstration at Super Computing 2016 at EchoStreams Booth #2537

November 15, 2016 01:20 PM Eastern Standard Time

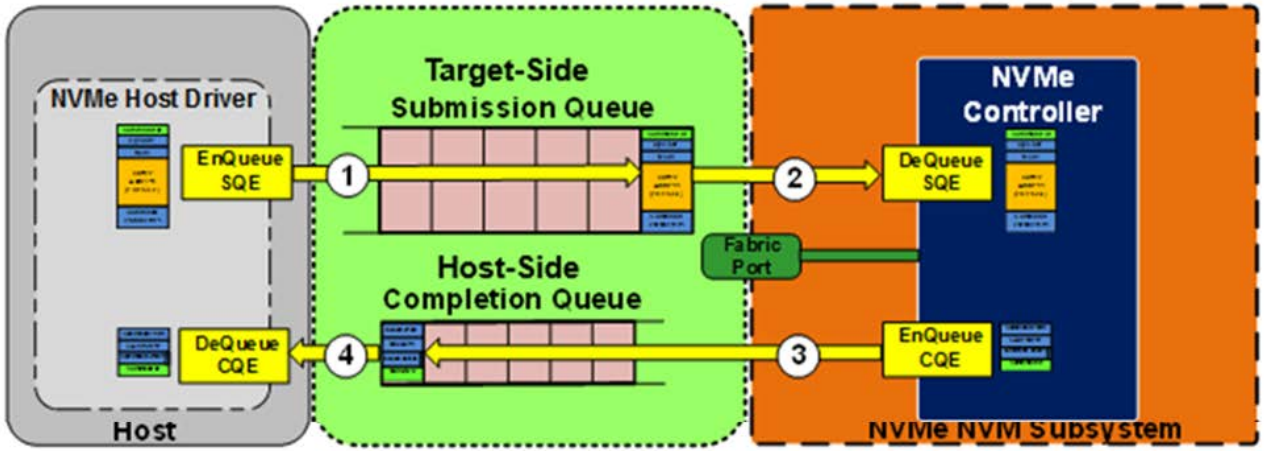
SALT LAKE CITY--(BUSINESS WIRE)--Liqid Inc., the industry leader in NVMe flash performance and PCI Express® (PCIe) based disaggregated infrastructure (DI) solutions, and Echostreams today announce partnership with Mellanox to deliver high performance data transfer nodes of 200Gbps throughput in a compact 1U form factor.

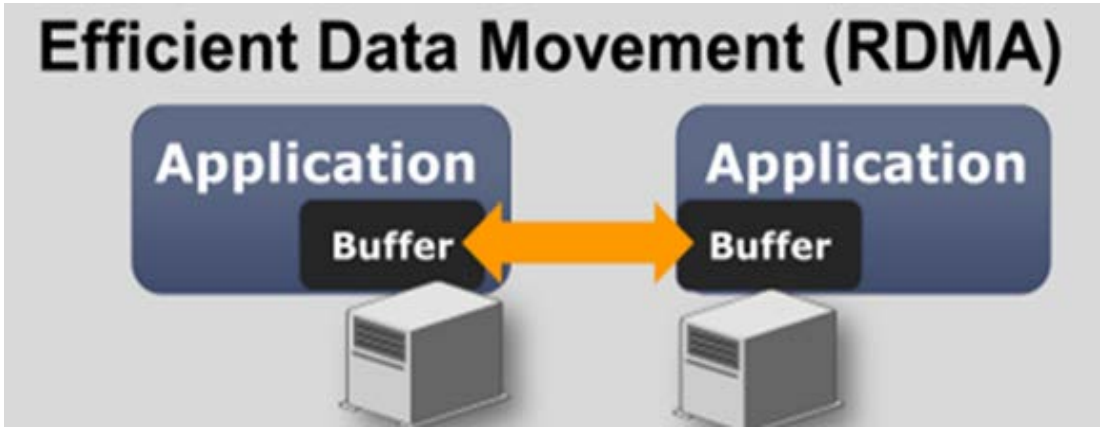
HOW DOES NVME OVER FABRICS MAINTAIN NVME PERFORMANCE?

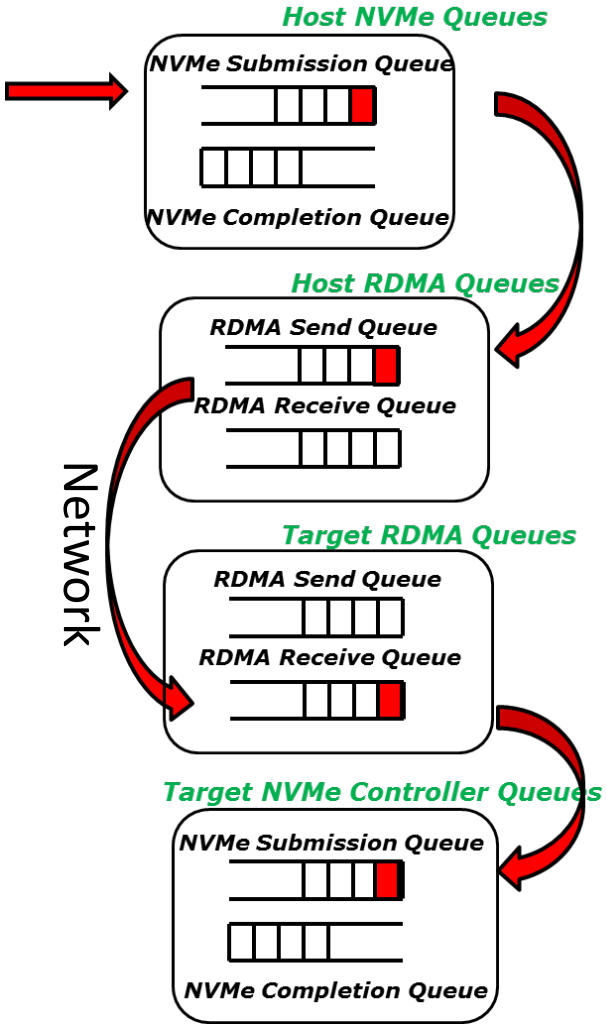
- **Extends NVMe efficiency over a fabric**
  - NVMe commands and data structures are transferred end to end
- **Relies on RDMA for performance**
  - Bypassing TCP/IP
  - Early Pre-Standard version also used RDMA
- **For more Information on NVMe over Fabrics (NVMe-oF)**
  - https://community.mellanox.com/docs/DOC-2186



RDMA & NVME: A PERFECT FIT







NVME OVER FABRICS (NVME-OF) STANDARD 1.0 OPEN SOURCE DRIVER

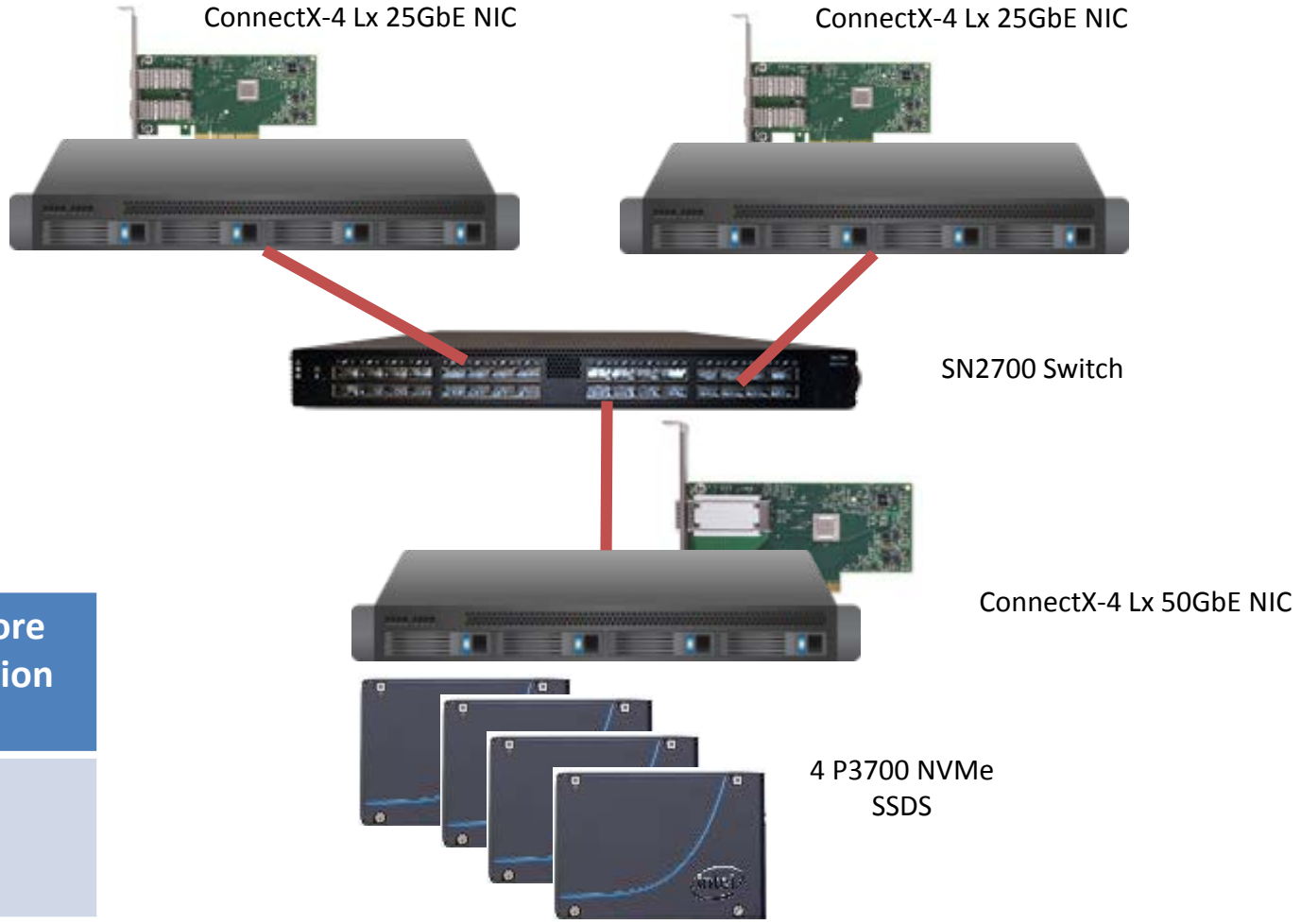
Released with Standard in June 2016

How to use open source driver instructions:
https://community.mellanox.com/docs/DOC-2504

Mellanox
Intel
HGST
EMC
Apeiron Data Systems
Broadcom Corporation
Chelsio Communications, Inc
Excelero
Hewlett Packard Enterprise
Kazan Networks
Kenneth Okin Consulting
Mangstor
NetApp
Oracle America Inc.
PMC
Qlogic Corporation
Samsung
SK hynix Inc.

NVME-OF PERFORMANCE WITH OPEN SOURCE LINUX DRIVERS

Added fabric latency
~12us, BS = 512b



	Bandwidth (Target side)	IOPS (Target side)	Num. Online cores	Each core utilization
BS = 4KB, 16 jobs, IO depth = 64	5.2GB/sec	1.3M	4	50%

INTRODUCTION TO LINUX NVME STACK

- **Originally for PCIe interface, extended for fabrics specification**
- **Host side**
  - Extended for various fabrics – RDMA, FC, Loop
  - Architecture that allows
    - fabric extensions
    - sharing common functionality through common code
    - Extends nvmecli for fabric configuration, discovery
  - Multi queue implementation to benefit from cpu affinity
- **Target side**
  - Kernel implementation that leverages block storage interface
  - Ground up implementation of target side
  - nvmetcli for configuration via configfs
  - sharing common functionality through common code

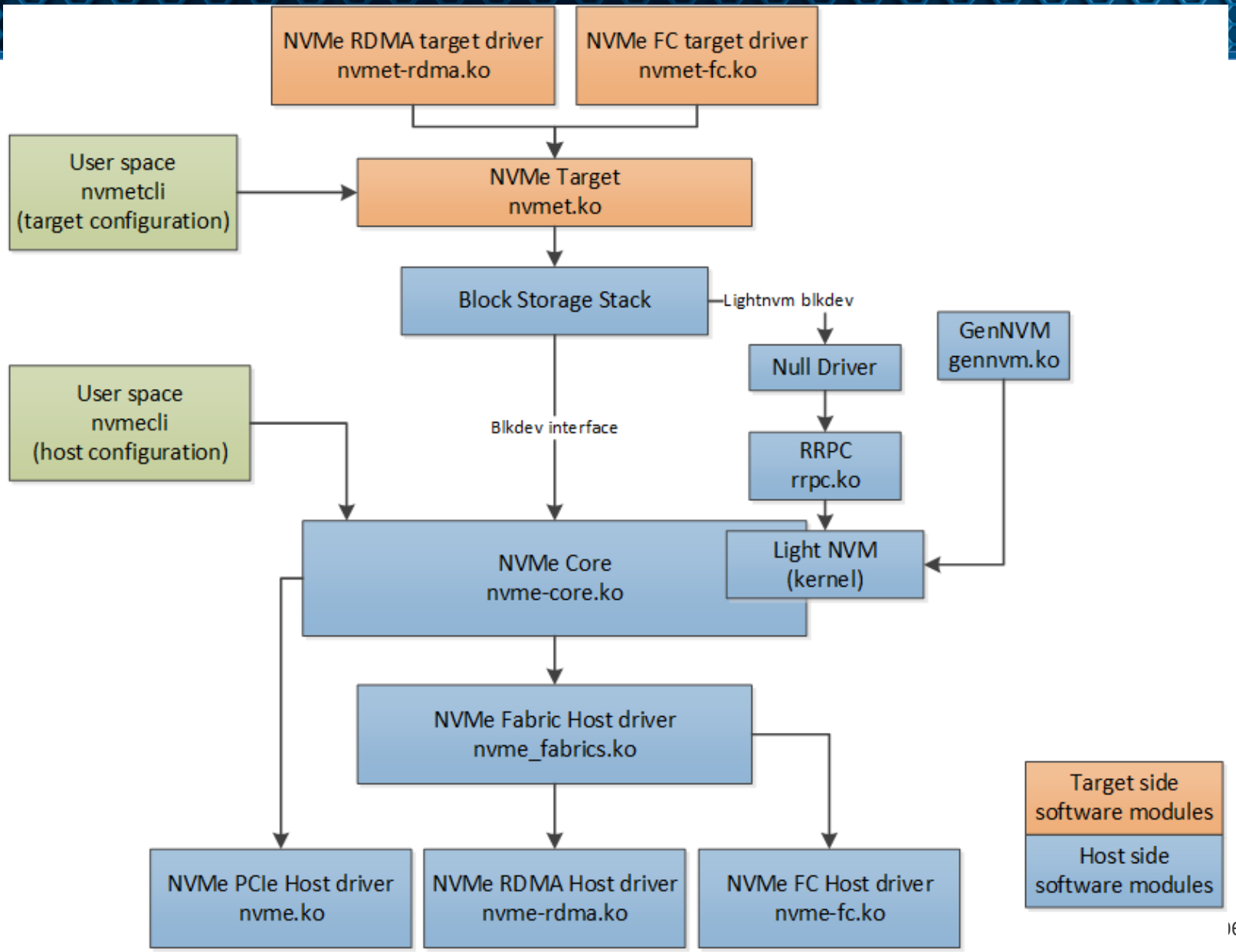
INTRODUCTION TO LINUX NVME RDMA STACK

- **Connection establishment is through standard RDMA CM**
- **Data path and resource setup through kernel IB verbs**
- **Supports target/subsystem defined inline data for write commands**
- **Utilizes common IB core stack for CQ processing, read-write operations**
- **Common code for IB, RoCEv2**

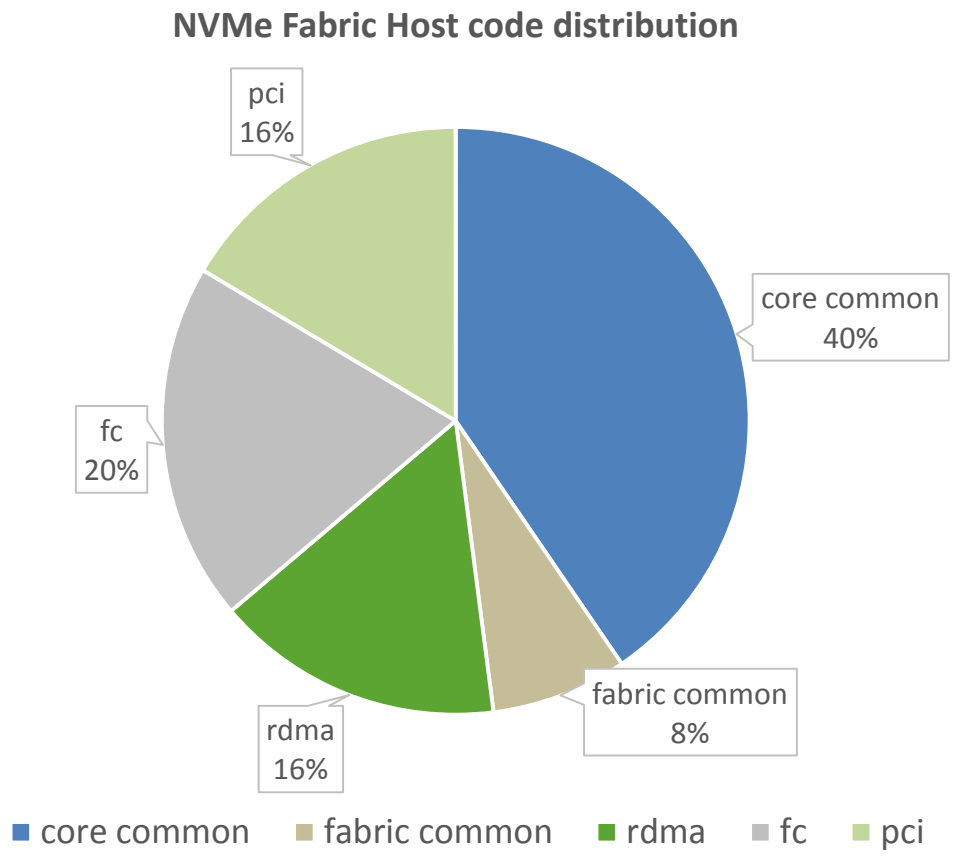
- **NVMe queues map to RDMA queues**
  - NVMe admin queue, IO queue -> RDMA QP, completion queue
  - NVMe completion queue -> RDMA QP, completion queue
- **Block layer SGEs map to RDMA SGE(s), memory region.**

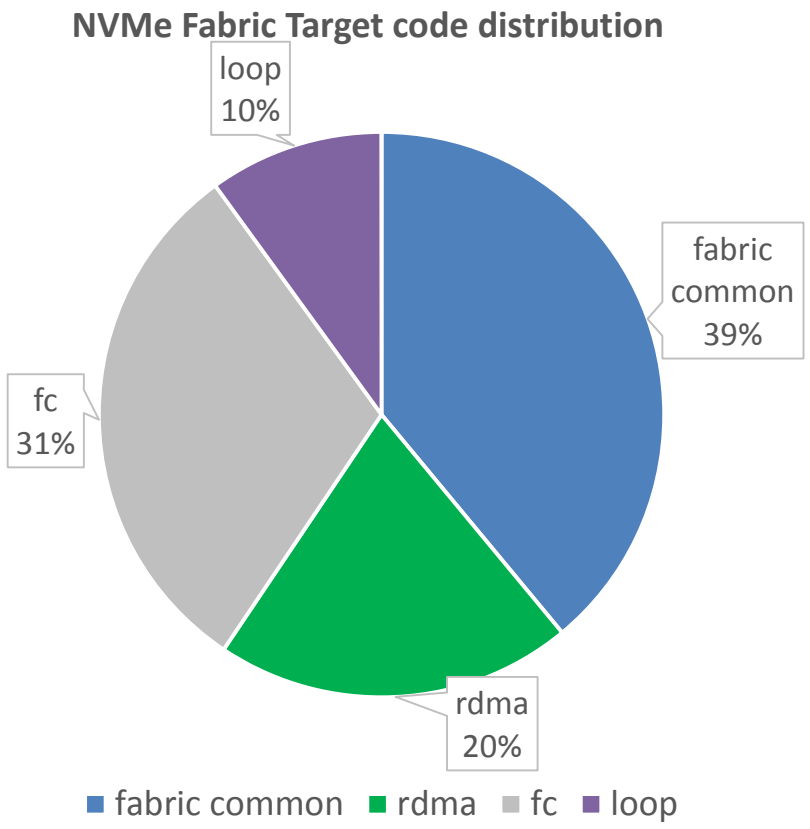
- **NVMe commands and completions are transported through RDMA SQ entries**

LINUX KERNEL NVME SOFTWARE STACK



NVME FABRICS CODE ORGANIZATION



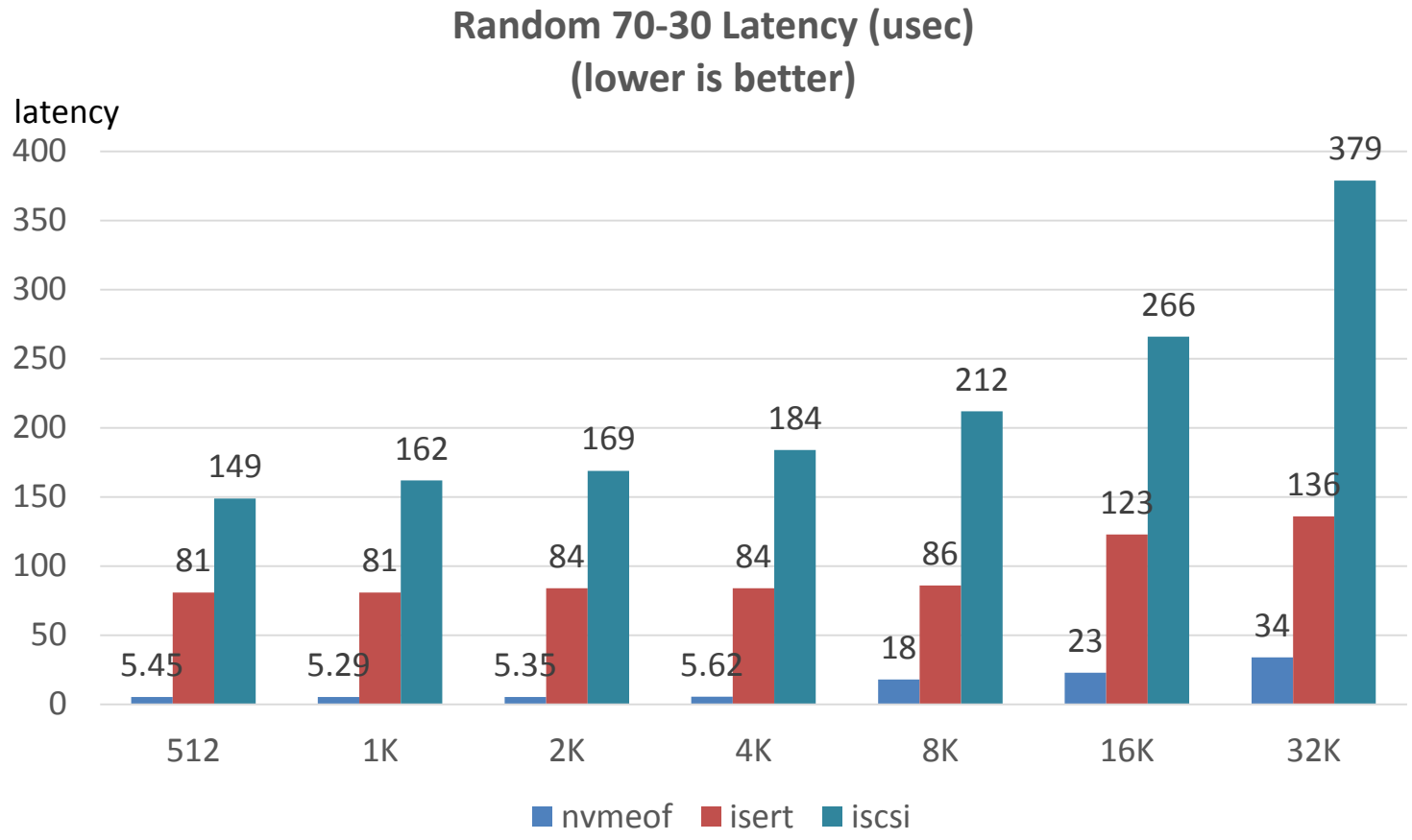


MICRO BENCHMARK SETUP

- **Setup:**
  - Hardware:
    - 64 core x86_64 host and target systems
    - 64Gb RAM
    - 100Gb Ethernet ConnectX-4 NICs
  - Software stack:
    - Linux NVMe host and target software stack with kernel 4.10+.
    - 250GB null target, 4K queue depth, 64 MQs, single LUN or namespace
    - NULL block driver with multiple queues for fabric performance characteristics
  - Tool:
    - fio
    - 16 jobs, 256 queue depth
    - 70% write, 30% read

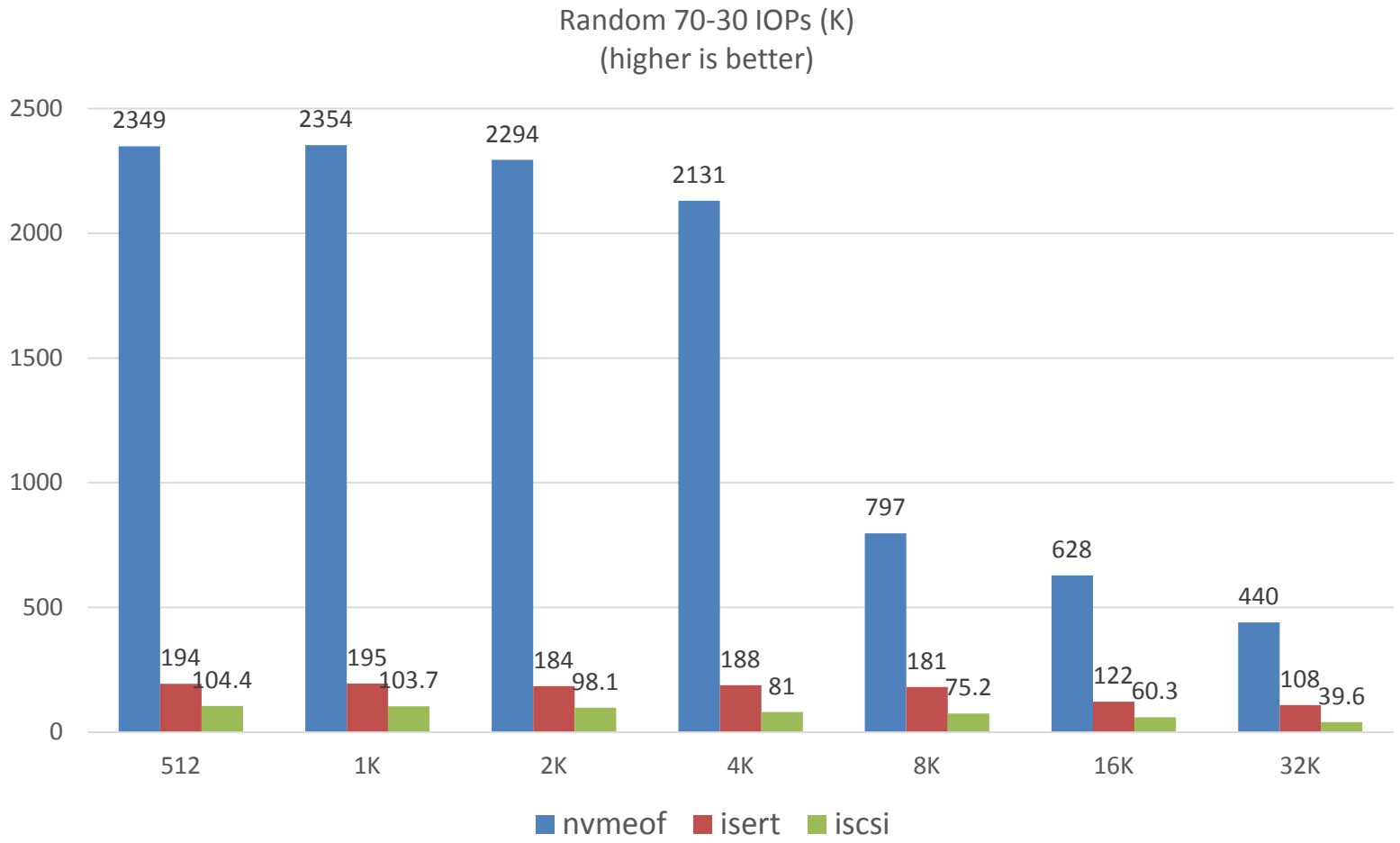
RANDOM READ WRITE 30-70 LATENCY CHART

- **20 times lower latency compare to iSCSI-TCP upto 4K IO Size**
- **10 times lower latency compare to ISER for 8K and higher**
- **2 times lower latency compare to iSER for all IO size**
- **Block layer MQ support comes natively to NVMe**



RANDOM READ WRITE 30-70 IOPS CHART

- **20 times higher IOPs compare to iSCSI-TCP upto 4K size**
- **4 times higher IOPs compare to iSER for size 8K and higher**



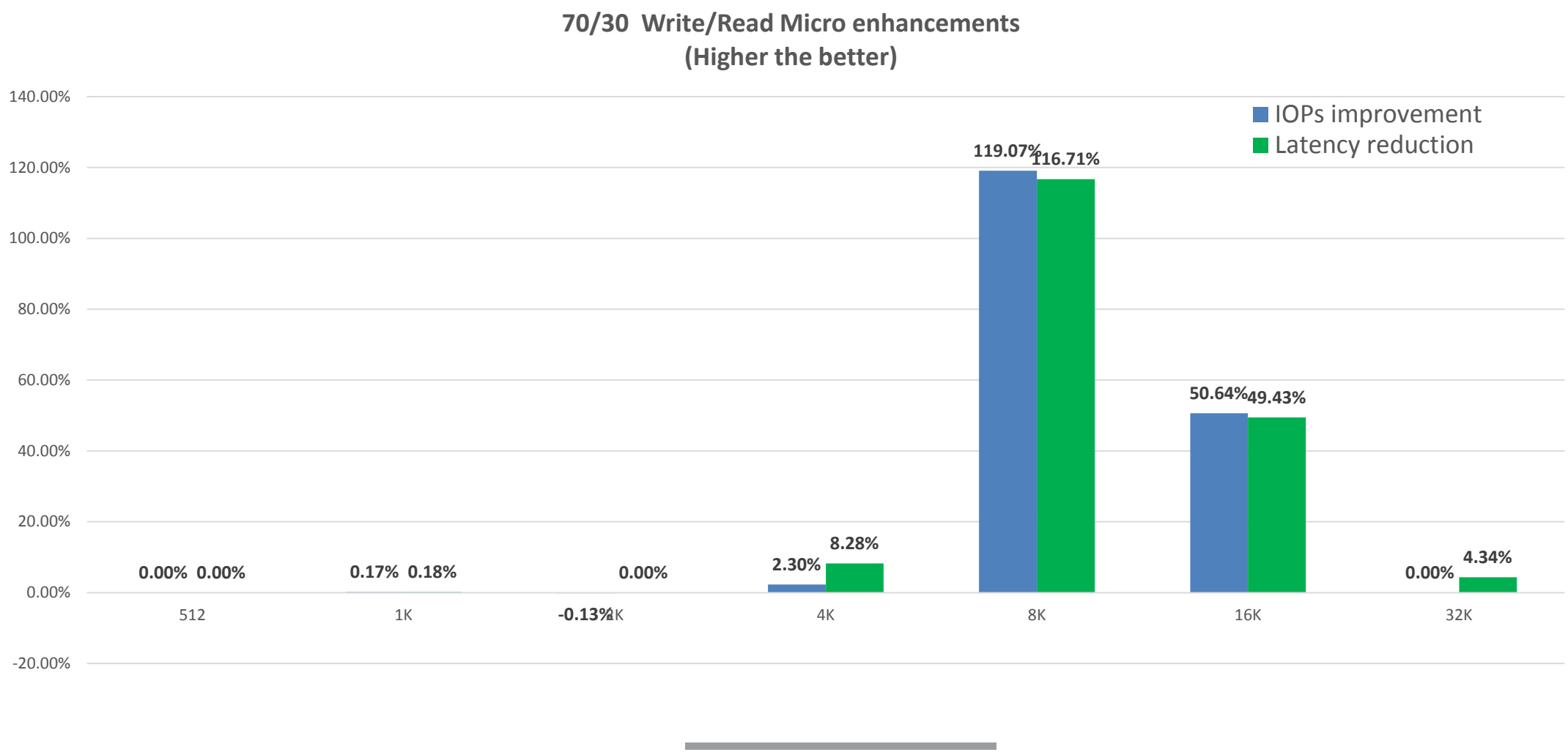
MICRO ENHANCEMENTS FOR RDMA FABRICS

- **Host side**
  - Using more SGEs on host side to handle non contiguous pages
  - Per IOQ setting, might be done based on advertised inline data size

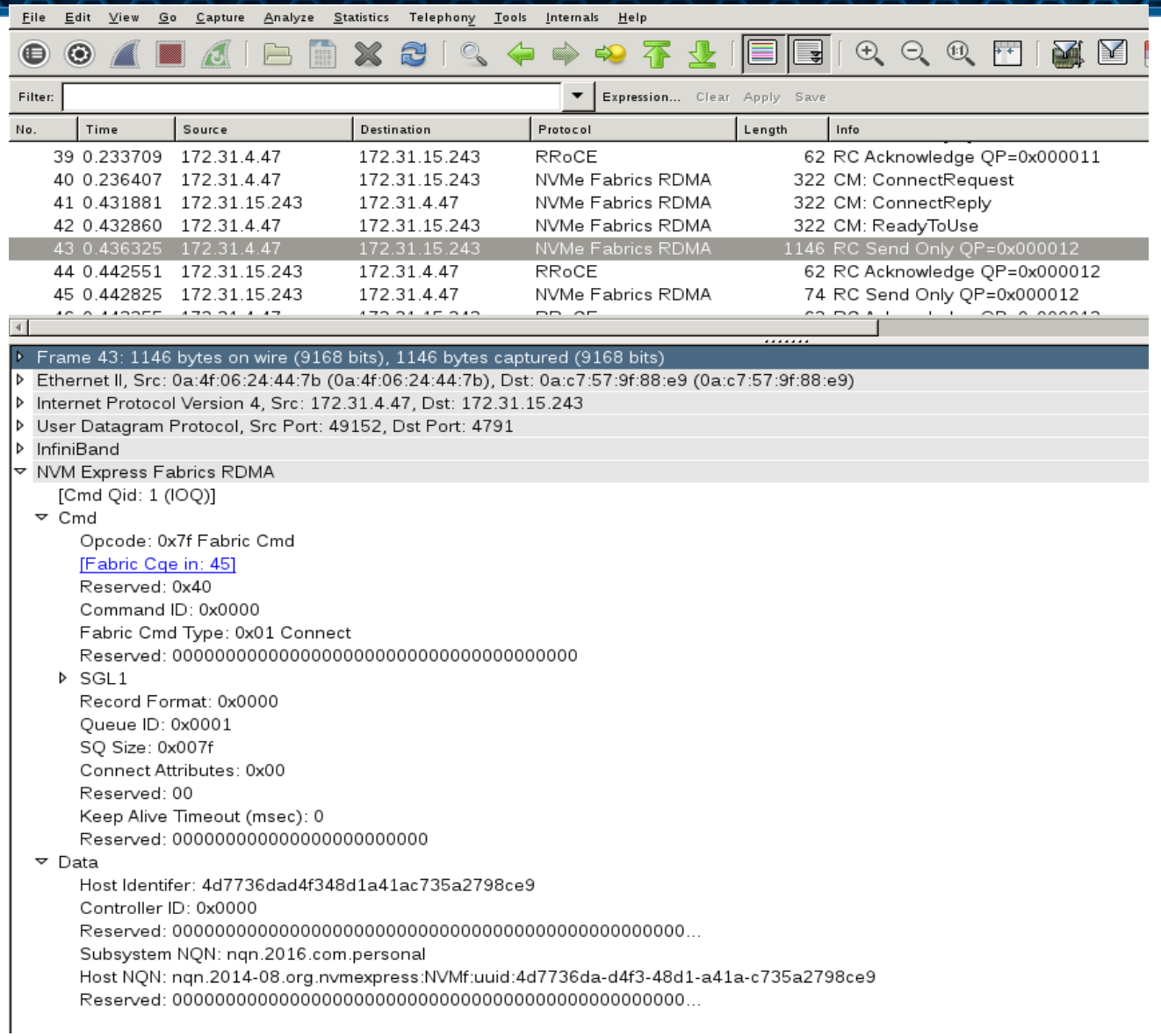
- **Target side**
  - Having higher inline size for incoming write commands
  - Currently static configuration; Will be per host setting in future via configfs

- **CPU utilization is unchanged with these new feature**
- **Patches are under review and update for dynamic configuration**

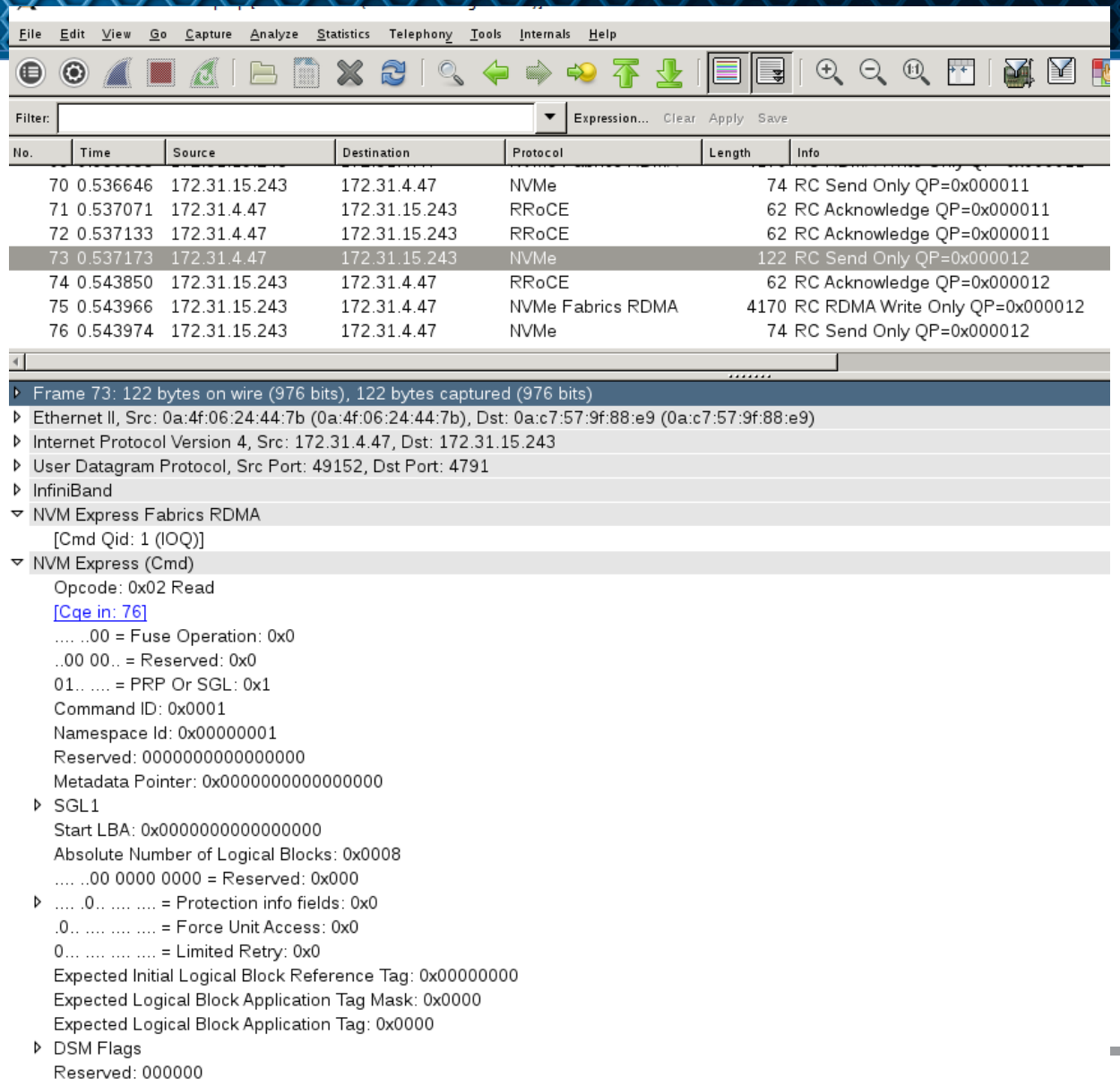
IMPROVEMENTS WITH 16K INLINE SIZE CHANGES

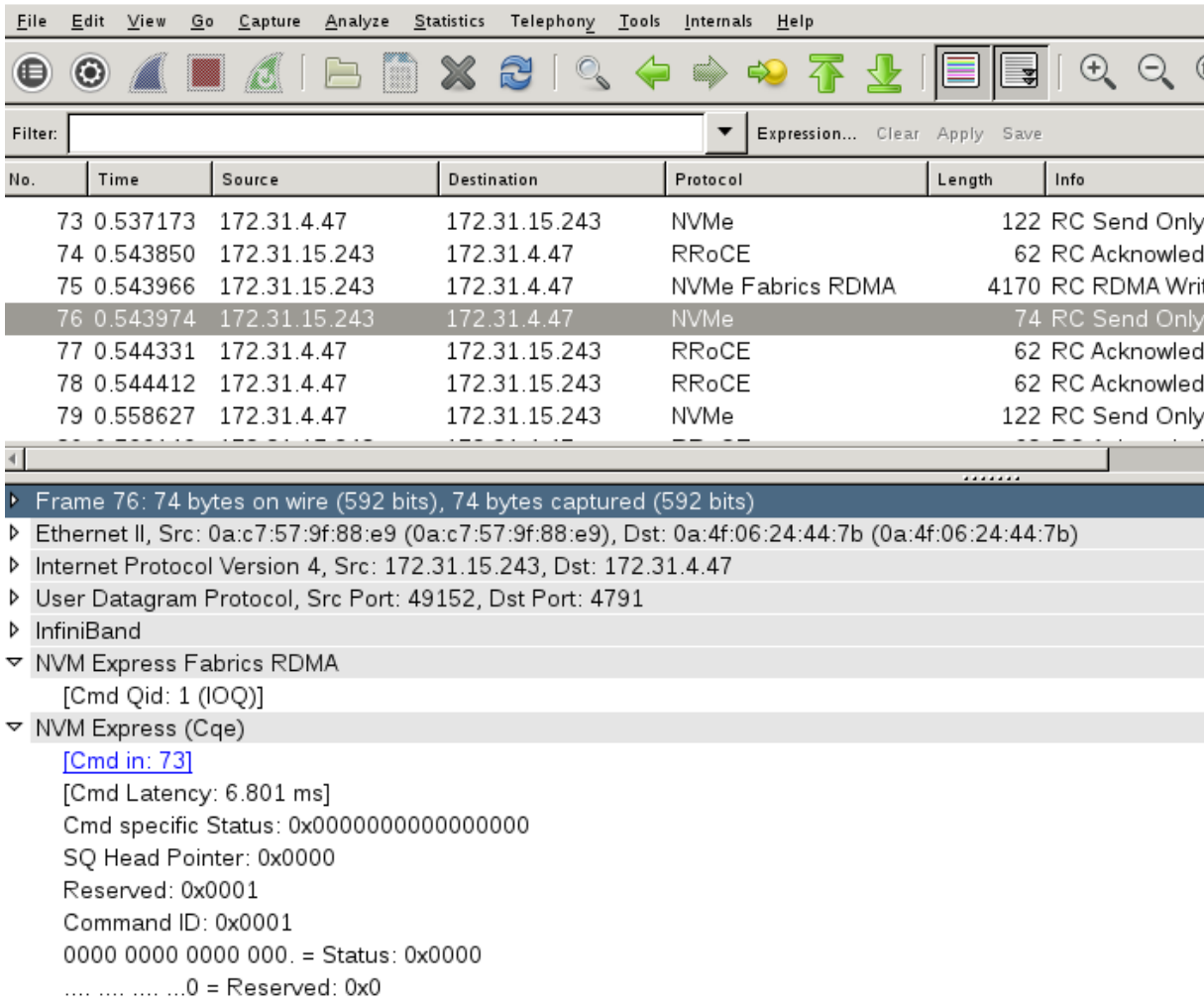


WIRESHARK SUPPORT FOR FABRIC COMMANDS (BASIC)



WIRESHARK TRACES FOR LATENCY ANALYSIS





13th ANNUAL WORKSHOP 2017

THANK YOU

Parav Pandit

Mellanox Technologies