Hardware offloads for SPDK

Sasha Kotchubievsky| Oren Duer | Mellanox Technologies

Apr 2019

© 2019 Mellanox Technologies

Agenda

- What is a SmartNIC
- NVMe emulation using a SmartNIC
- SPDK as NVMe emulation framework
- Offloads
  - NVMf target offload
  - NVMe emulation offload
  - Crypto, compression, RAID, …
- Offloads support in SPDK

© 2019 Mellanox Technologies

What are SmartNIC and Device Emulation

- SmartNIC: NIC and CPU in one SoC
- For networking
  - Software flexibility in the networking path
  - Take software defined decisions per packet, or flow
  - Example: OpenVirtualSwitch (OVS)
- Device Emulation
  - Present to host an arbitrary device
  - Device’s logic implemented in software
  - Accelerations for critical data paths
  - Example: NVMe

© 2019 Mellanox Technologies

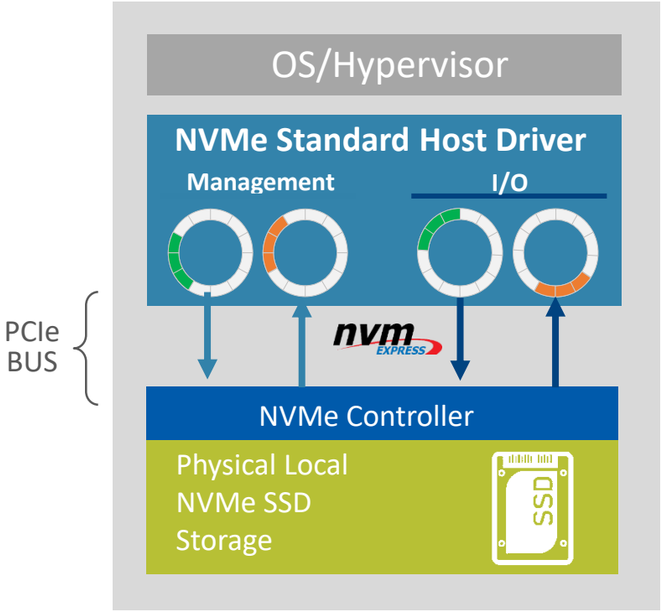
NVMe SNAP:
Local Physical Storage to Hardware Emulated Storage

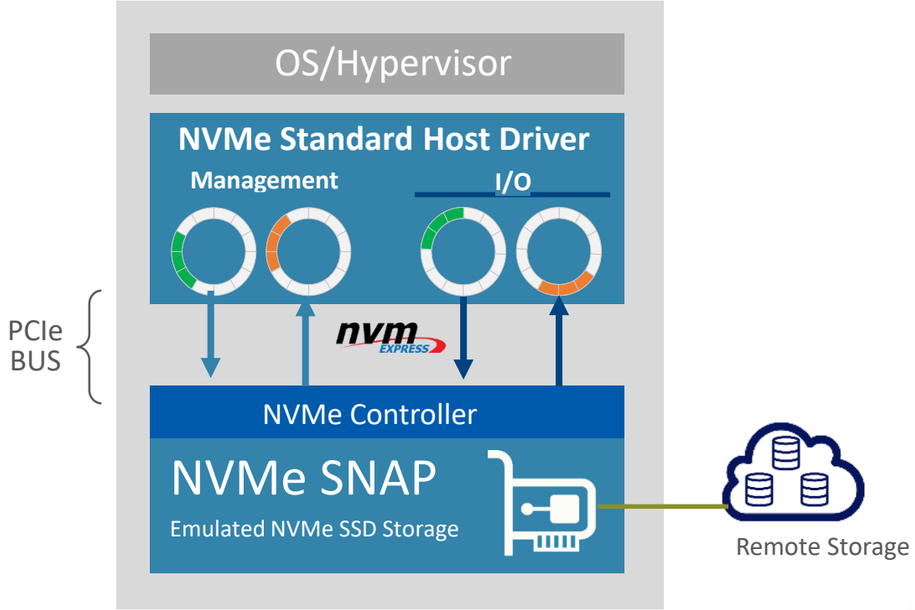
Physical Local NVMe Storage

- ✓ Serving bare-metal and hypervisor/VMs
- ✗ Bound by physical SSDs capacity
- ✗ Under-utilized storage
- ✗ Scalability
- ✗ Over-provisioning bound to compute node

NVMe SNAP Drive Emulation

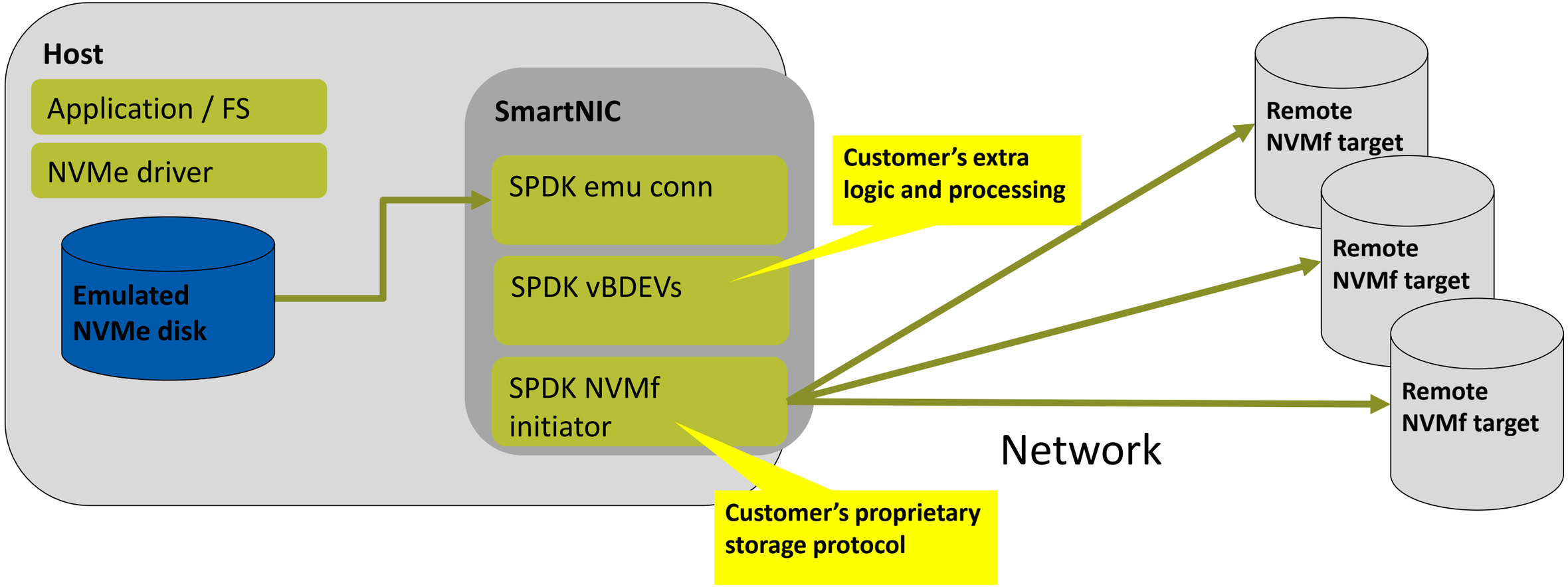
- ✓ Serving bare-metal and hypervisor/VMs
- ✓ Over-provisioning, scaled to rack/cluster
- ✓ Saving OPEX and CAPEX
- ✓ OS-agnostic using inbox standard NVMe driver
- ✓ Supports all network transport types – NVMe-oF, iSCSI, iSER and even proprietary





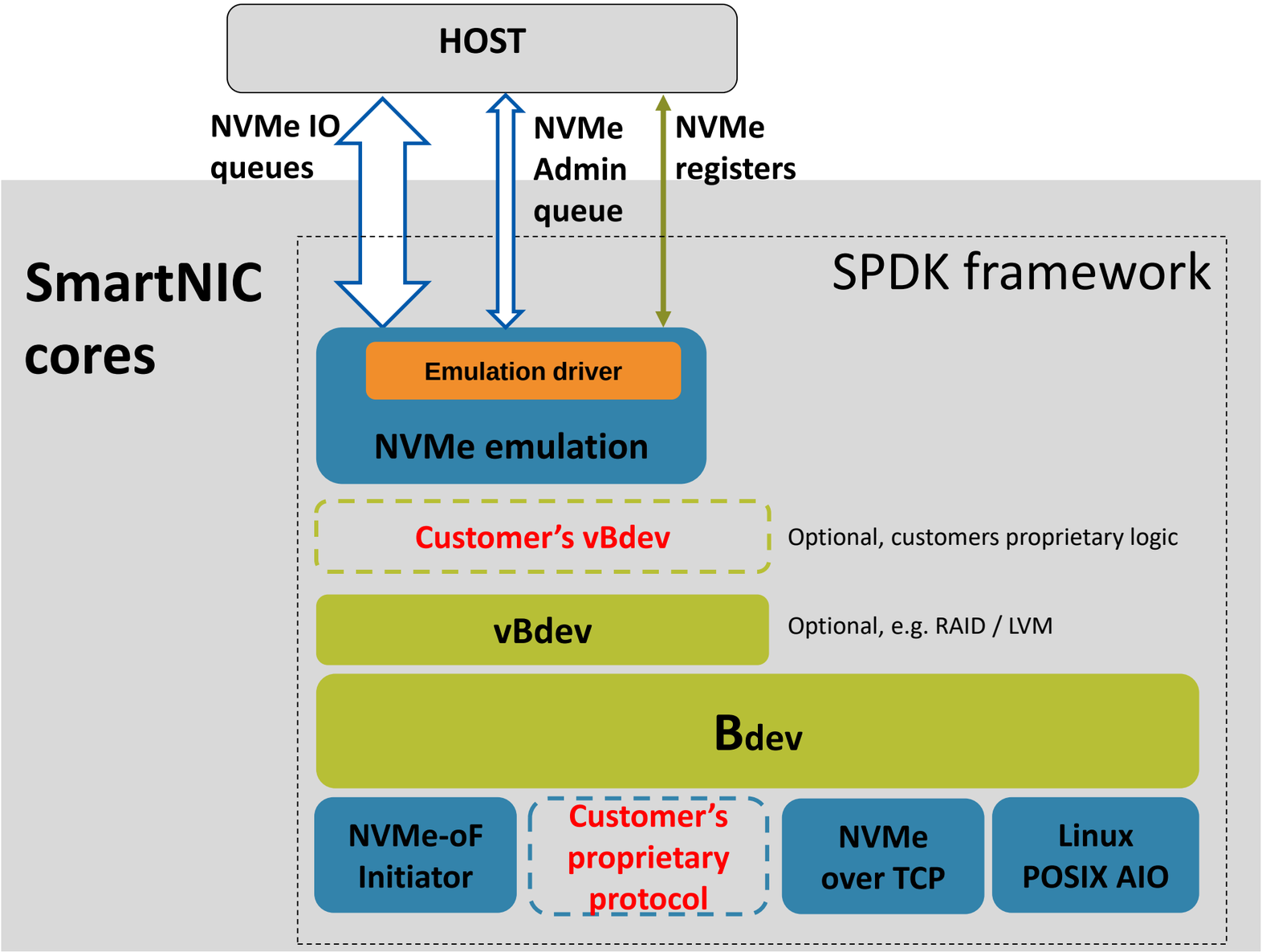
© 2019 Mellanox Technologies

NVMe emulation common use-case



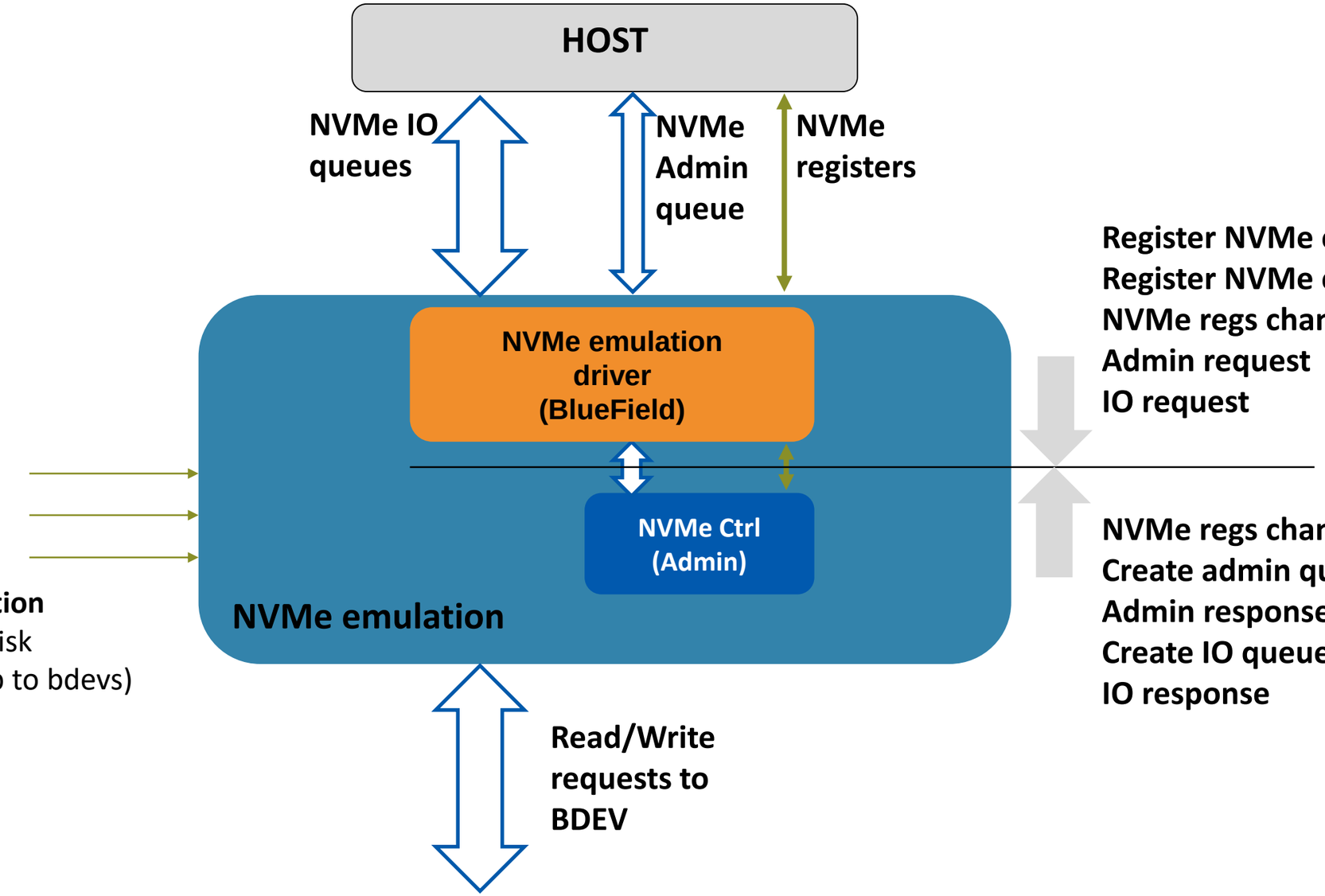
© 2019 Mellanox Technologies

SPDK as NVMe Emulation framework



© 2019 Mellanox Technologies

NVMe emulation over SPDK



Target-like configuration
Subsystem = NVMe disk
Add namespace (map to bdevs)
Remove namespace

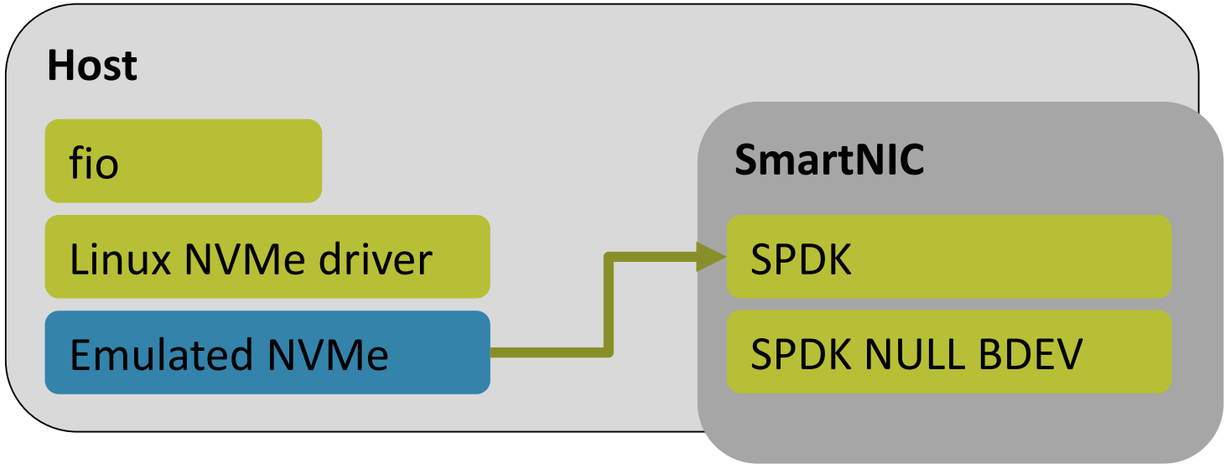
Register NVMe emulation driver
Register NVMe emulated disk
NVMe regs change
Admin request
IO request

NVMe regs change
Create admin queue
Admin response
Create IO queue
IO response

© 2019 Mellanox Technologies

Performance

Initial non-optimized results

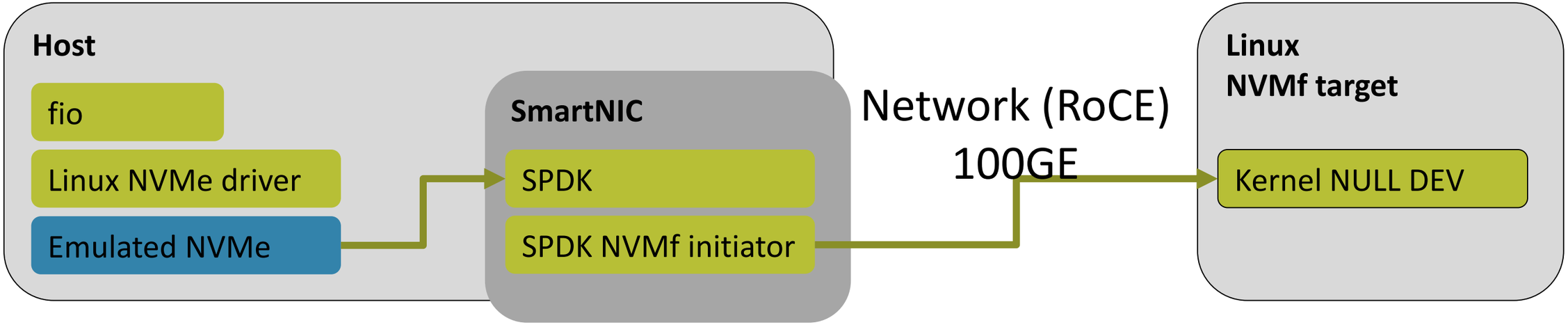


- **2M IOPS @ 4K**
  - 100G SmartNIC
  - Using 15 cores on SmartNIC

© 2019 Mellanox Technologies

Performance

Initial non-optimized results



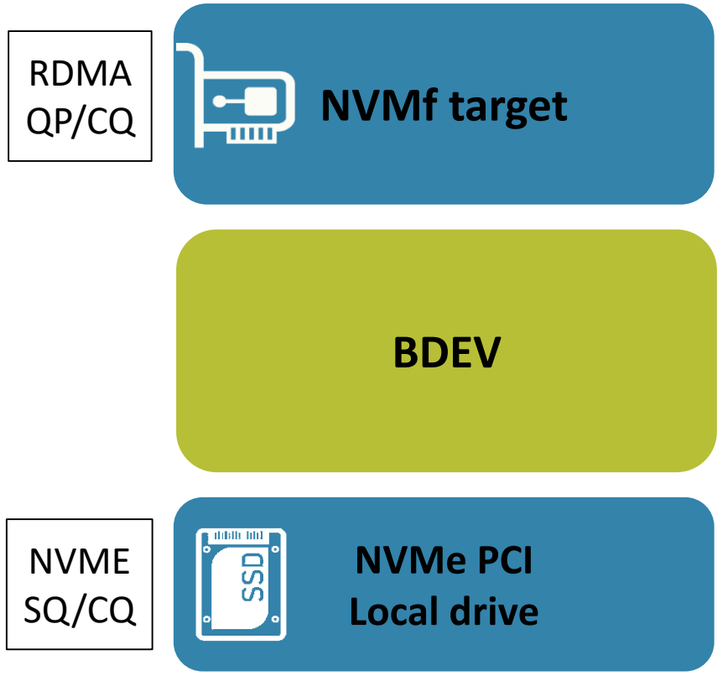
- **1M IOPS @ 4K**
  - 100G wire speed
  - Using 10 cores on SmartNIC

© 2019 Mellanox Technologies

SPDK offloads

Target offload

- RDMA NIC can “talk” directly to an NVME device
- Need to pass NVME SQ/CQ to RDMA NIC
- SPDK tasks (in different layers)
  - Allow configuring an “offloaded NVMf subsystem”
  - Enforce limitations (can only use NVMe as backend)
  - Communicate NVME SQ/CQ to RDMA NIC
  - Handle exceptions

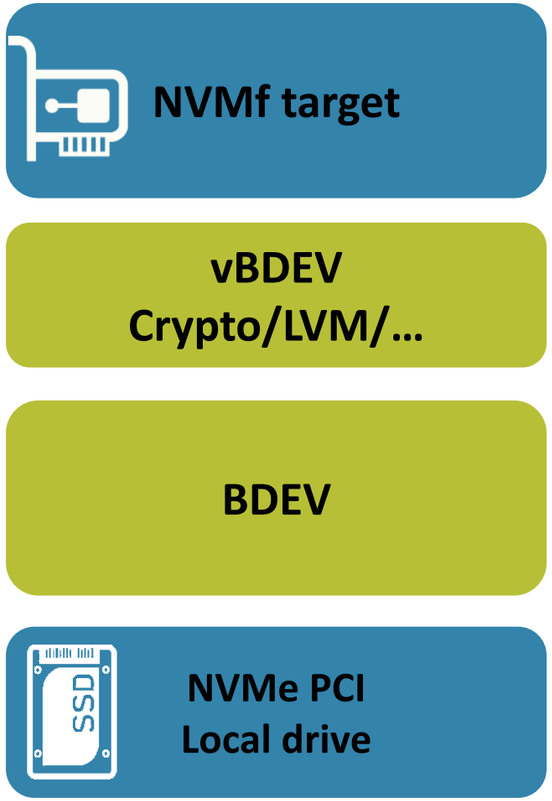


© 2019 Mellanox Technologies

SPDK offloads

vBdev offload

- RDMA NIC may be able to run in-line compute offloads
  - Examples: Crypto, RAIDs, LVM
- Need to pass vBDEV internal configuration info
  - And runtime data…
- SPDK tasks (in different layers)
  - Allow configuration, enforce limitations
  - Allow “interrogating” internal configuration and runtime data
  - Handle exceptions



© 2019 Mellanox Technologies

NVME-OF Target offload

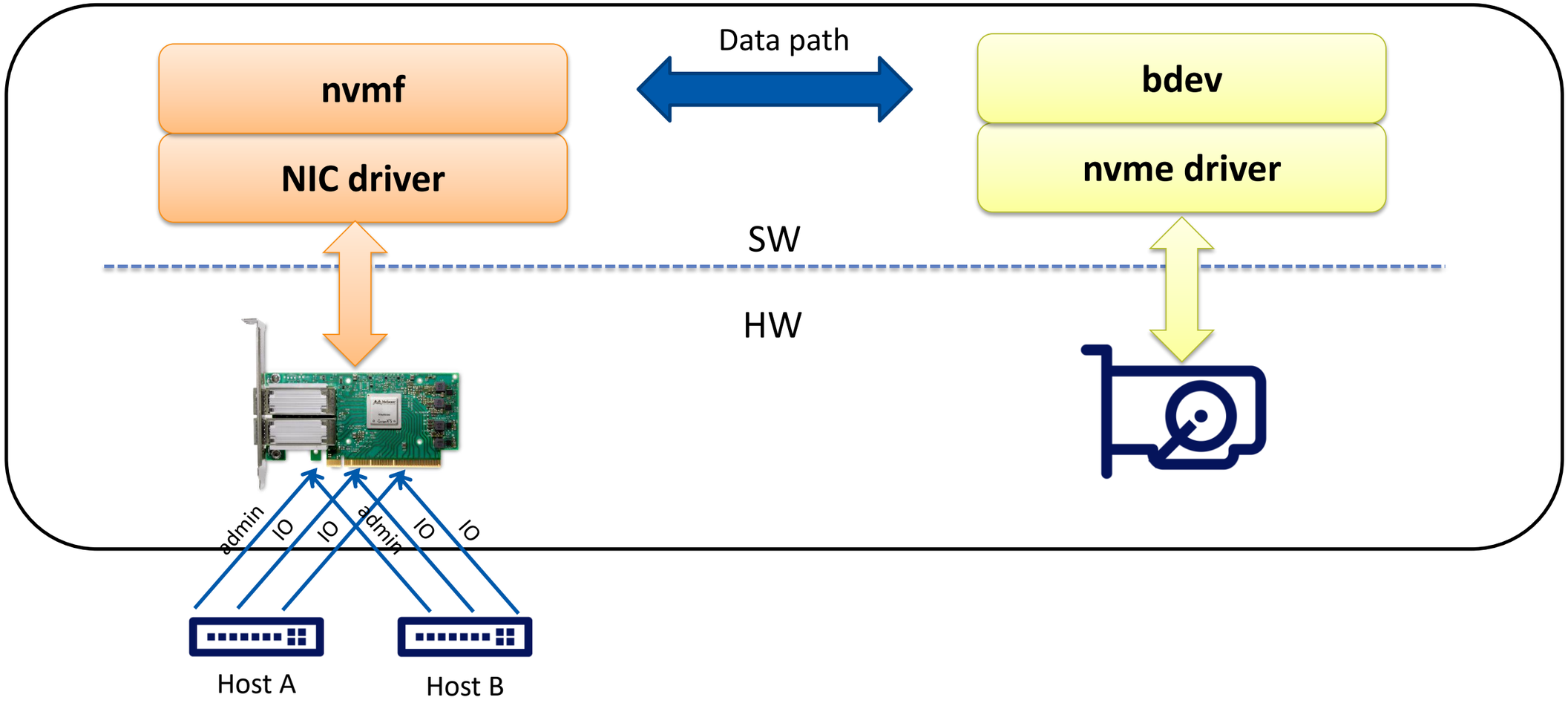
© 2019 Mellanox Technologies

Agenda

- NVME-OF (RDMA) target offload
- Current status
- Performance results

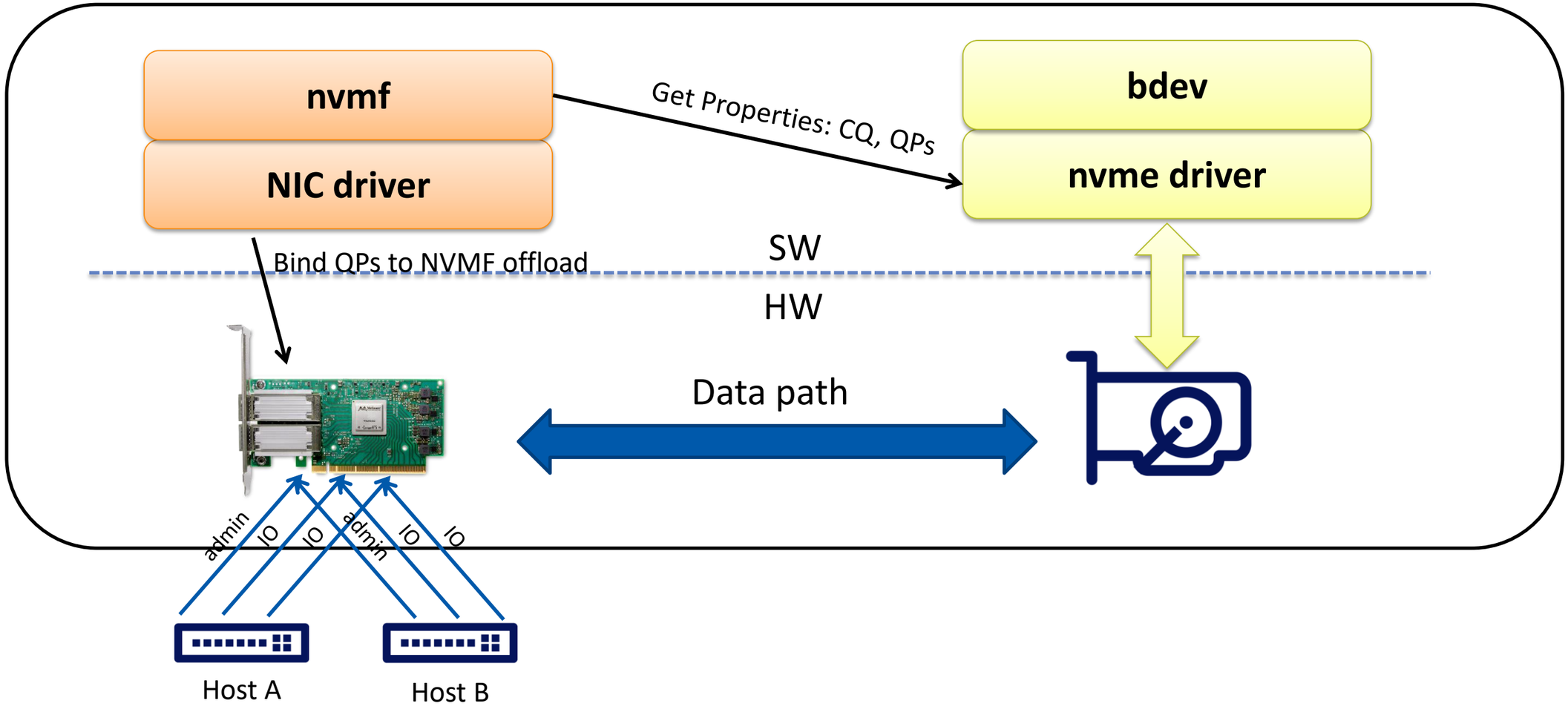
© 2019 Mellanox Technologies

NVMF Target



© 2019 Mellanox Technologies

NVMF Target offload



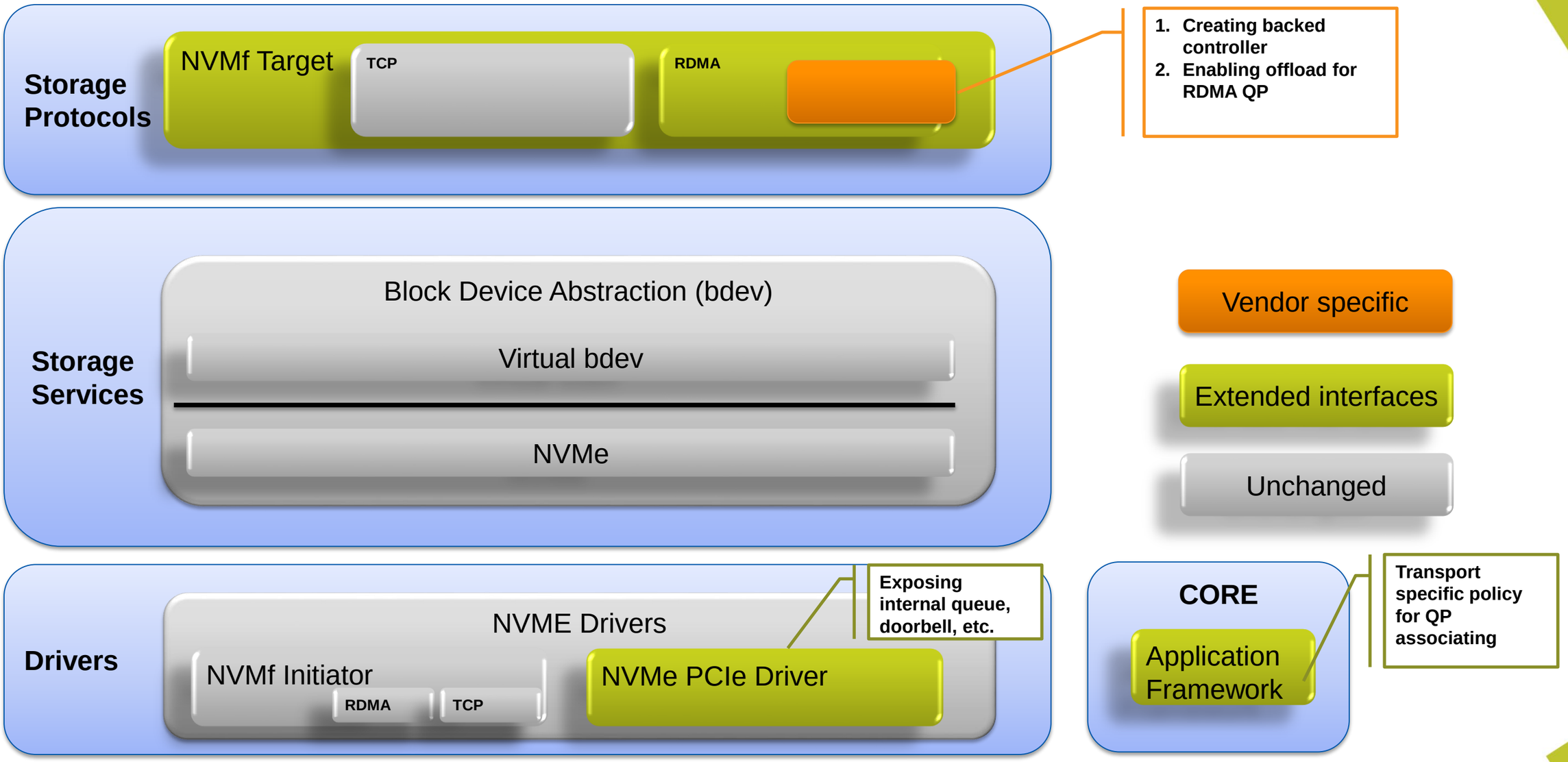
© 2019 Mellanox Technologies

NVMF Target offload

- SPDK is used as “control path”
- Target offload is available in ConnectX-5 and later
- Technical preview is available
- Submit patches to upstream in 2019

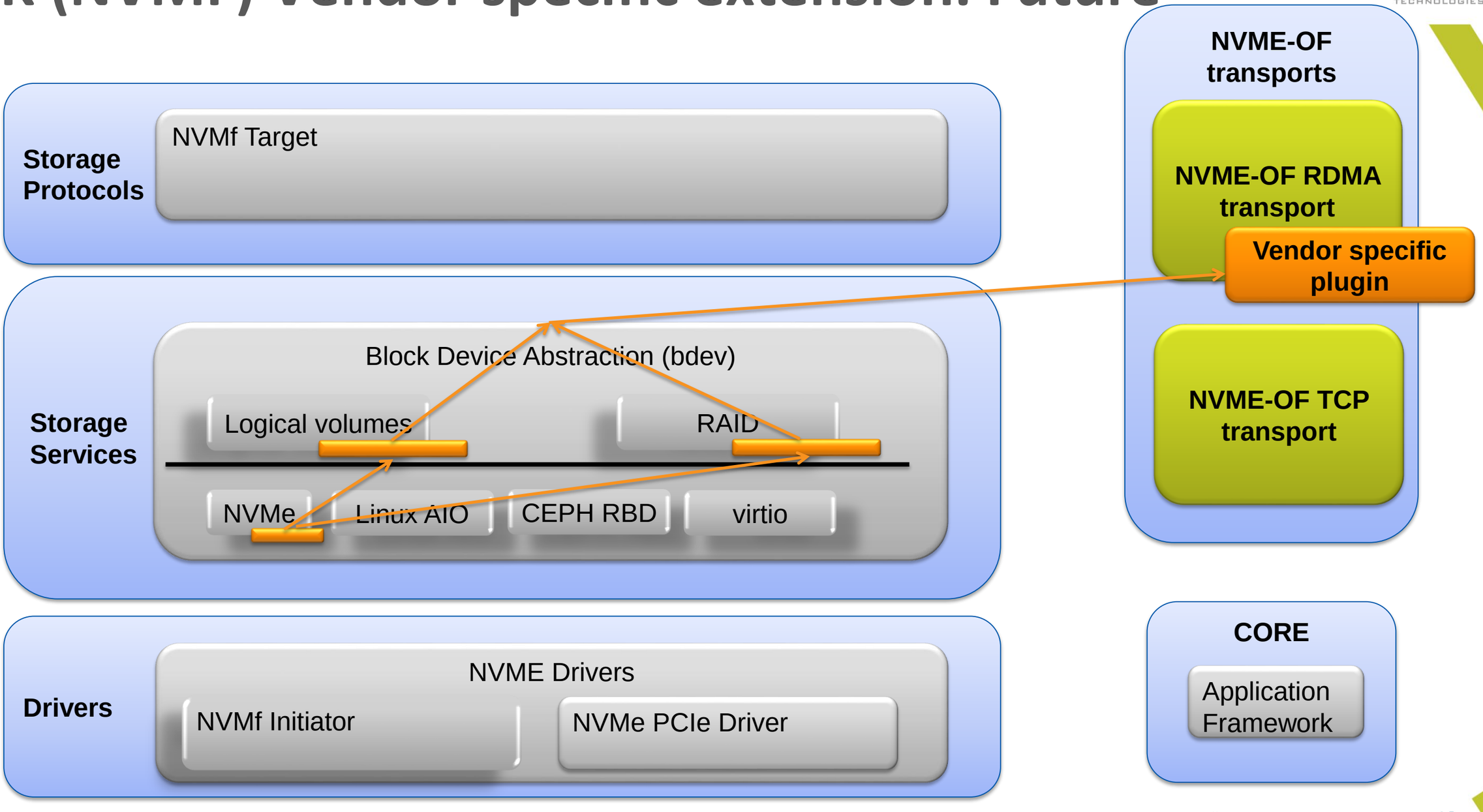
© 2019 Mellanox Technologies

SPDK (NVMF) Vendor specific extension



© 2019 Mellanox Technologies

SPDK (NVMF) Vendor specific extension. Future



© 2019 Mellanox Technologies

Performance results

NVME-OF Target offload

© 2019 Mellanox Technologies

Benchmark Setup

Initiator: x86, 24 cores
SPDK perf

Target: Bluefield , reference platform ,
16 cores, 8xNVME disks (PCIe gen4, x2)
SPDK NVMf target

100Gb, ROCE v2

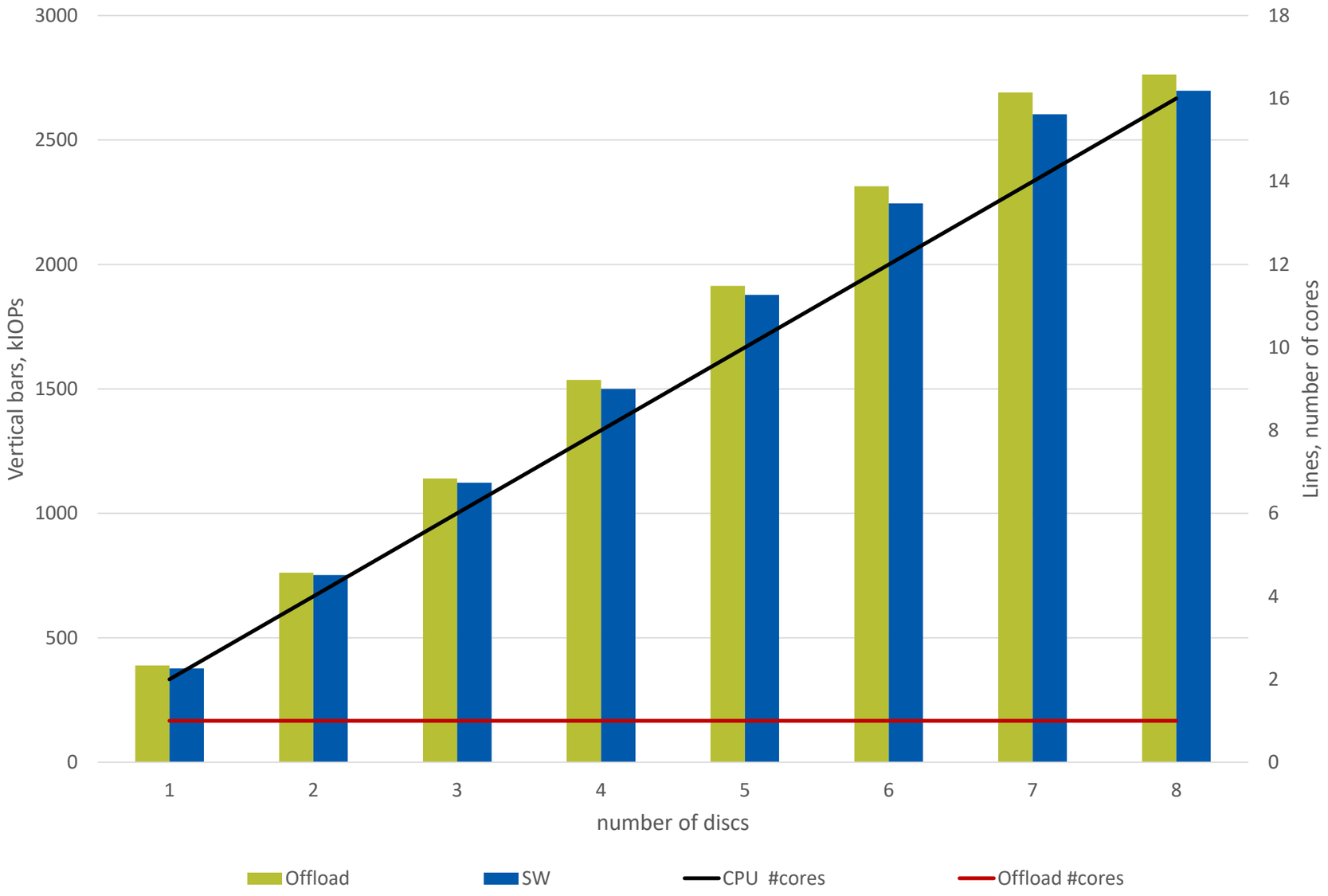
© 2019 Mellanox Technologies

Configuration

- Benchmarking CPU utilization and IOPs multiple disks in NVME-OF offloaded target
- SPDK, NVMF target, Software implementation
  - nvmf spans 16 cores
- SPDK, NVMF target offload
  - nvmf runs on single core
- randread and randwrite, 4K blocks, 16 queue depth

© 2019 Mellanox Technologies

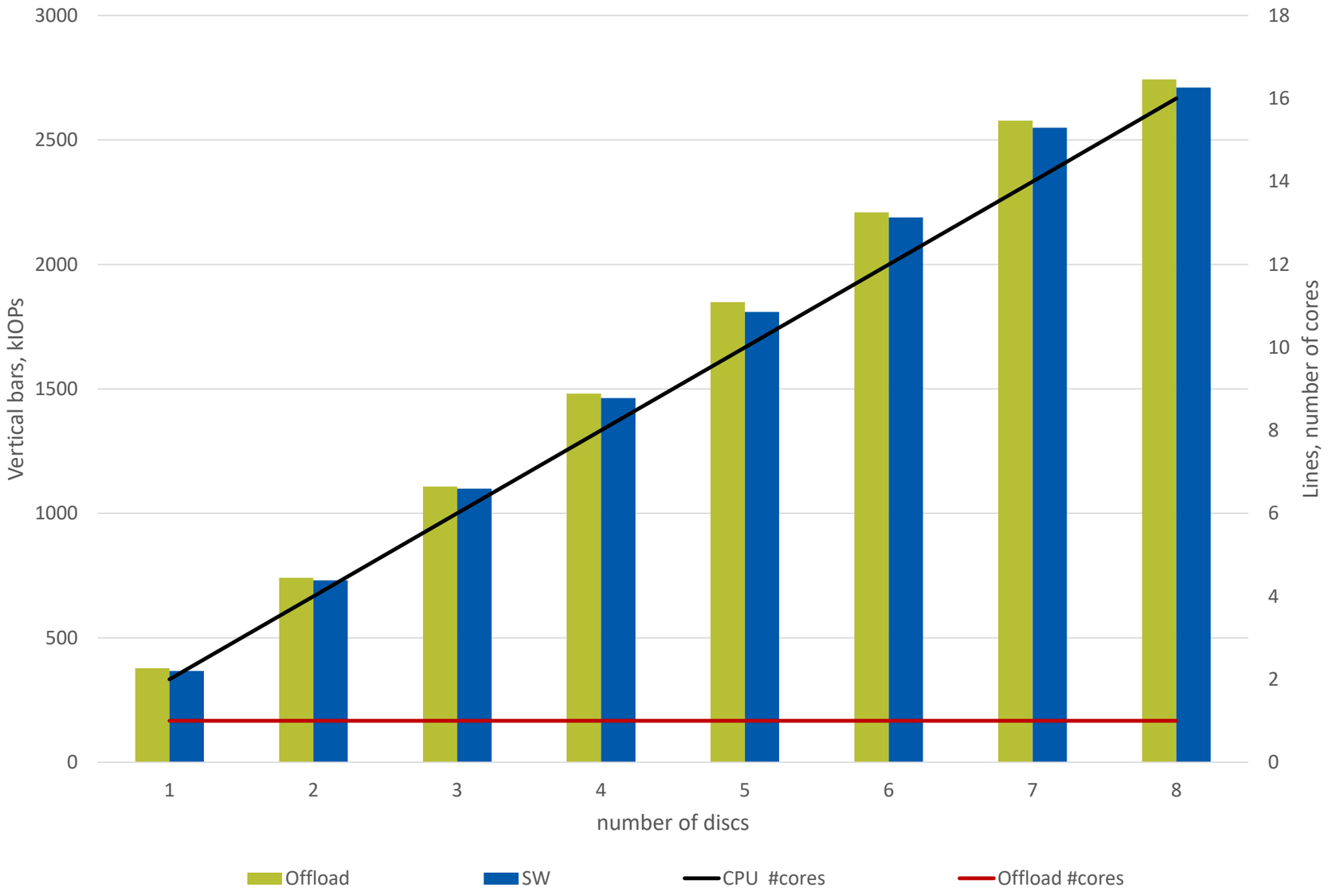
CPU utilization vs number of disks, randread 4K



Vertical bars: IPOs (higher is better)
Lines: number of utilized core (lower is better)

© 2019 Mellanox Technologies

CPU utilization vs number of disks, randwrite 4K



Vertical bars: IPOs (higher is better)
Lines: number of utilized core (lower is better)

© 2019 Mellanox Technologies

Summary

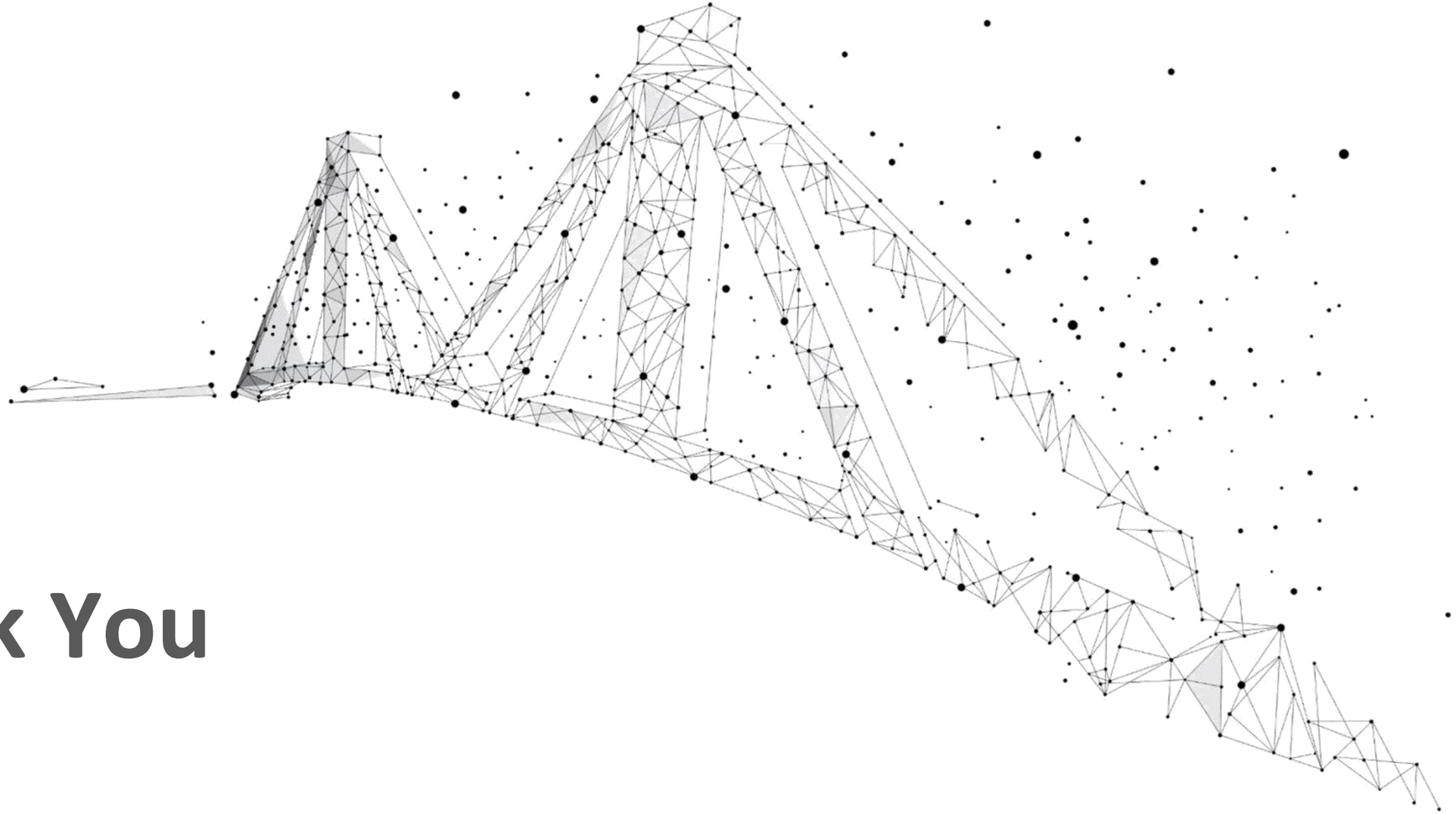
- NMVE-OF target offload demonstrates better performance than software implementation
- HW offload allows full bandwidth utilization without CPU involvement (single core is needed for serving admin commands)

© 2019 Mellanox Technologies

Mellanox®
TECHNOLOGIES

20 YEAR ANNIVERSARY

© 2019 Mellanox Technologies



Thank You

© 2019 Mellanox Technologies

Live Demo

© 2019 Mellanox Technologies

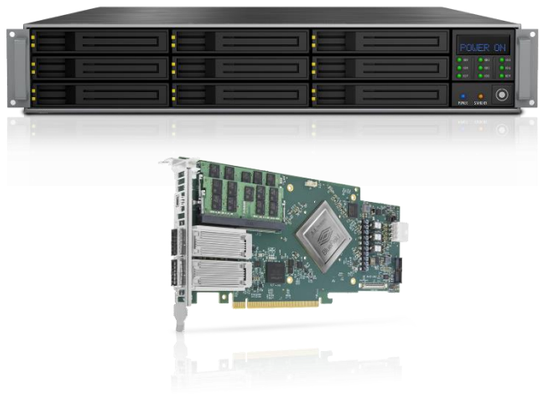
Demo

- Target
  - NVME-OF target offload
- Host
  - Network storage logically presented by SNAP™ as local NVME drives
    - Mellanox NVMe SNAP™ (Software-defined Network Accelerated Processing) technology enables hardware-accelerated virtualization of NVMe storage.
  - SPDK offloads IO processing in SmartNIC
    - Logical Volumes are created using virtual bdev

© 2019 Mellanox Technologies

Demo Setup

Host: x86, MS Windows, BlueField network adapter

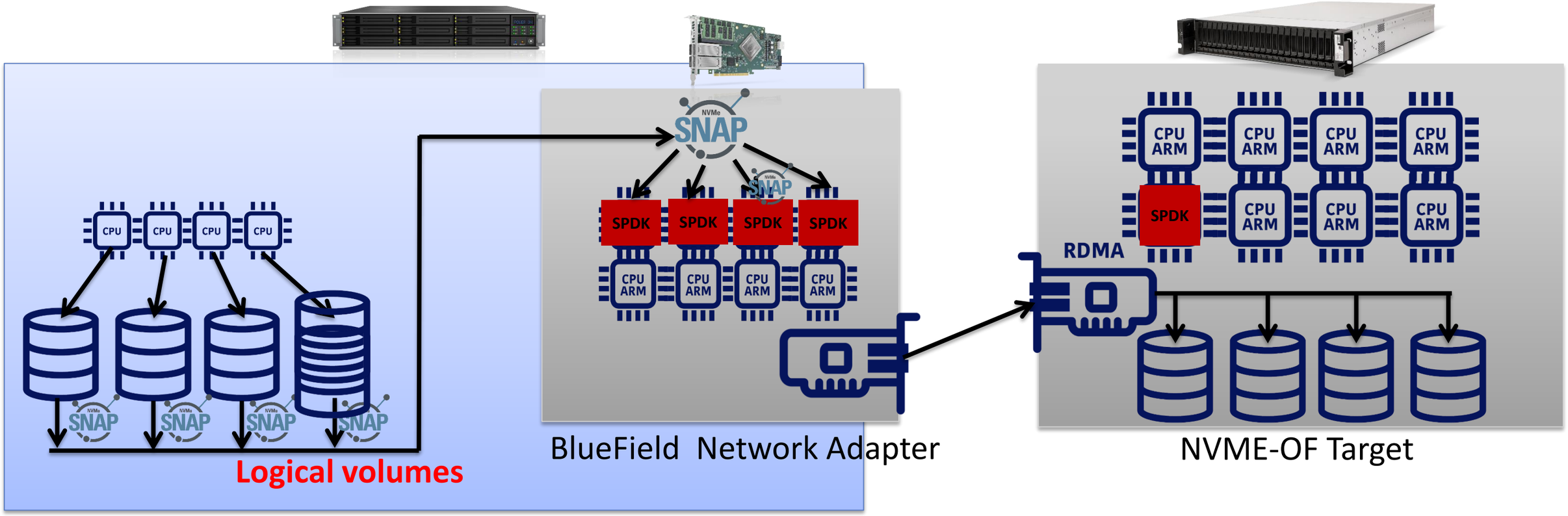


10Gb, ROCE v2

Target: Bluefield , reference platform , 16 cores, 4xNVME disks (PCIe gen4, x2) SPDK NVMf target

© 2019 Mellanox Technologies

Demo setup



© 2019 Mellanox Technologies