SDC

STORAGE DEVELOPER CONFERENCE

SNIA ■ SANTA CLARA, 2017

Low-Overhead Flash Disaggregation via NVMe-over-Fabrics

Vijay Balakrishnan

Memory Solutions Lab.

Samsung Semiconductor, Inc.

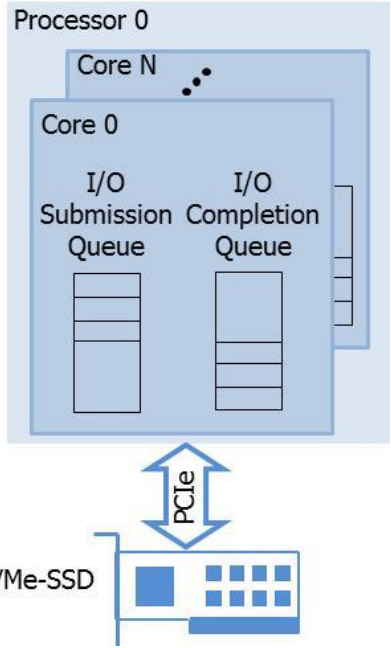
DISCLAIMER

This presentation and/or accompanying oral statements by Samsung representatives collectively, the "Presentation") is intended to provide information concerning the SSD and memory industry and Samsung Electronics Co., Ltd. and certain affiliates (collectively, "Samsung"). While Samsung strives to provide information that is accurate and up-to-date, this Presentation may nonetheless contain inaccuracies or omissions. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of the information provided in this Presentation.

This Presentation may include forward-looking statements, including, but not limited to, statements about any matter that is not a historical fact; statements regarding Samsung's intentions, beliefs or current expectations concerning, among other things, market prospects, technological developments, growth, strategies, and the industry in which Samsung operates; and statements regarding products or features that are still in development. By their nature, forward-looking statements involve risks and uncertainties, because they relate to events and depend on circumstances that may or may not occur in the future. Samsung cautions you that forward looking statements are not guarantees of future performance and that the actual developments of Samsung, the market, or industry in which Samsung operates may differ materially from those made or suggested by the forward-looking statements in this Presentation. In addition, even if such forward-looking statements are shown to be accurate, those developments may not be indicative of developments in future periods

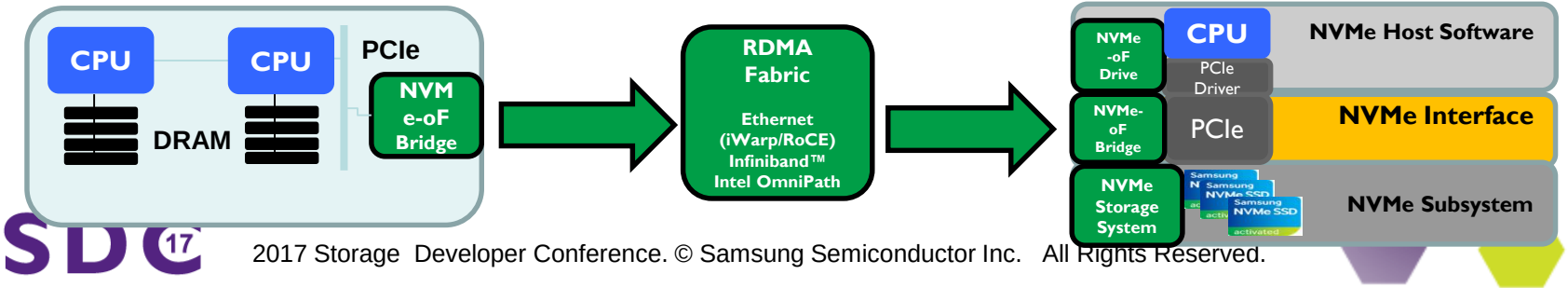
NVMe SSD

- NVMe: High performance, scalable interface for PCIe SSD
- High-performance through parallelization:
  - Large number of deep submission/completion queues
- NVMe-SSDs deliver lots of IOPS/BW
  - 1MIOPS, 6GB/s from a single device
  - 5x more than SAS-SSD, 20x more than SATA-SSD
- Industry standard supported by all major players

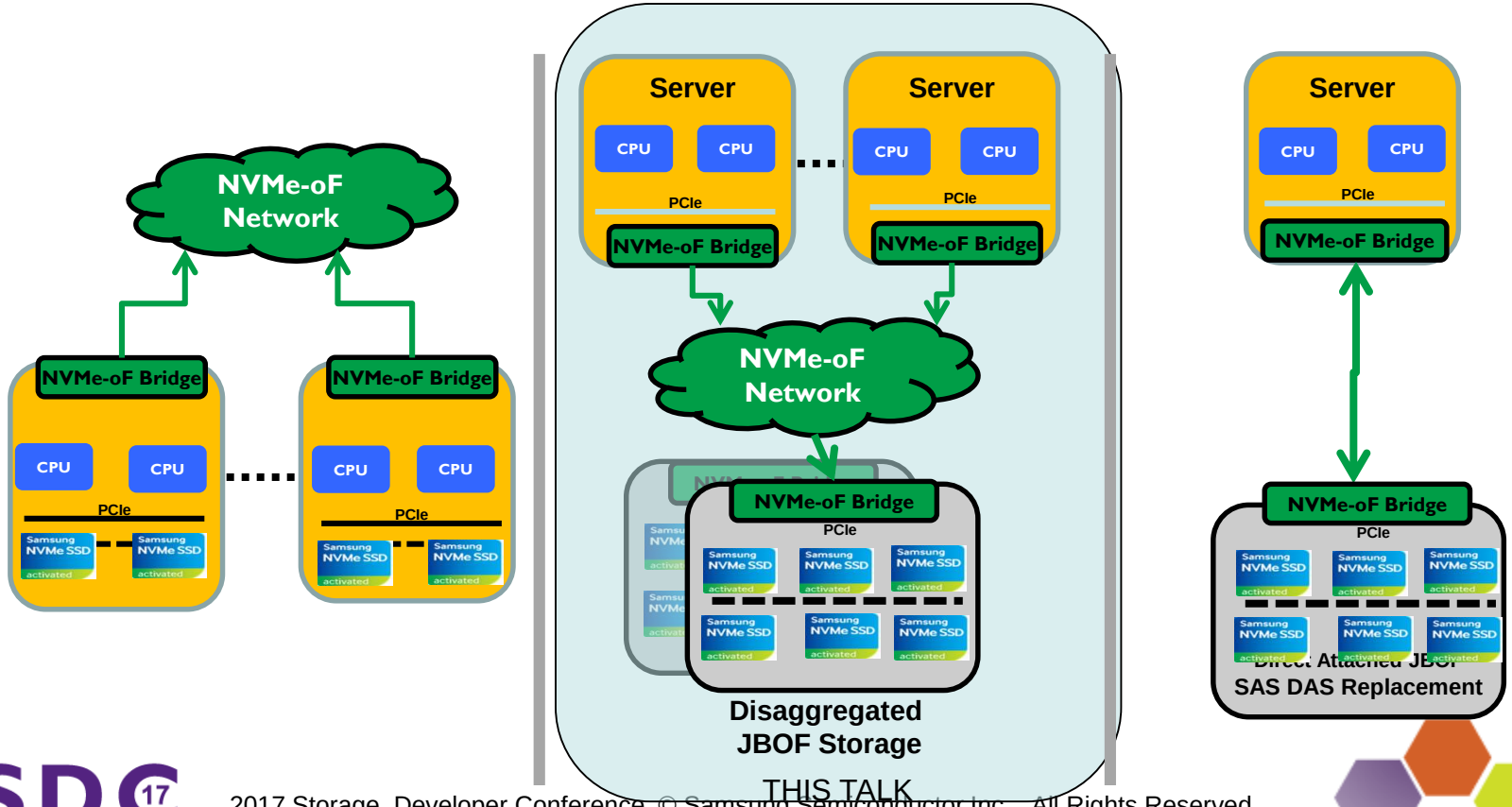


NVMe-over-Fabrics (NVMe-oF)

- A protocol interface to NVMe that enables operation over other interconnects (e.g., Ethernet, InfiniBand™, Fibre Channel, etc.)
- Shares the same base architecture and NVMe Host Software as NVMe
- Parallelism: extends the multiple queue-paired design of NVMe over network
- Enables NVMe scale-out and low latency (<10µs latency target) operations on Data Center Fabrics
- Avoids unnecessary protocol translations

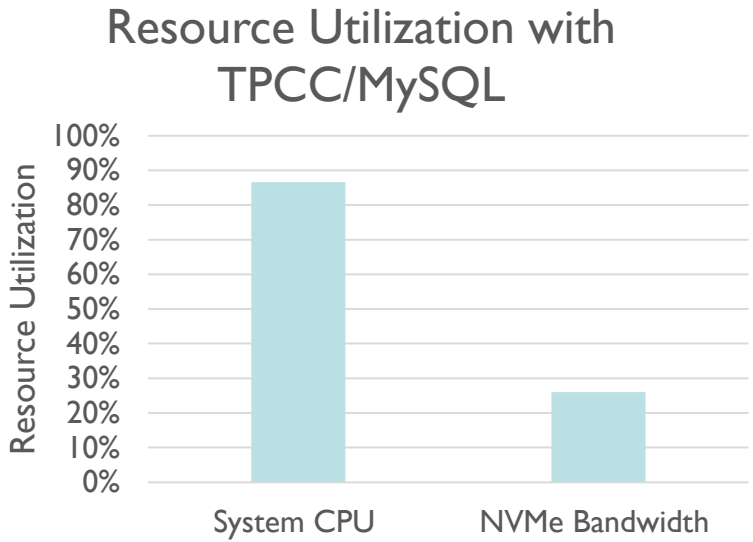


NVMe-oF Use Case Scenarios



NVMe Flash is Underutilized

- Compute saturates before IOPs
  - Need to enable sharing for SSDs
  - Need to scale CPU independently
- Capacity also underutilized
  - Single drive densities are growing

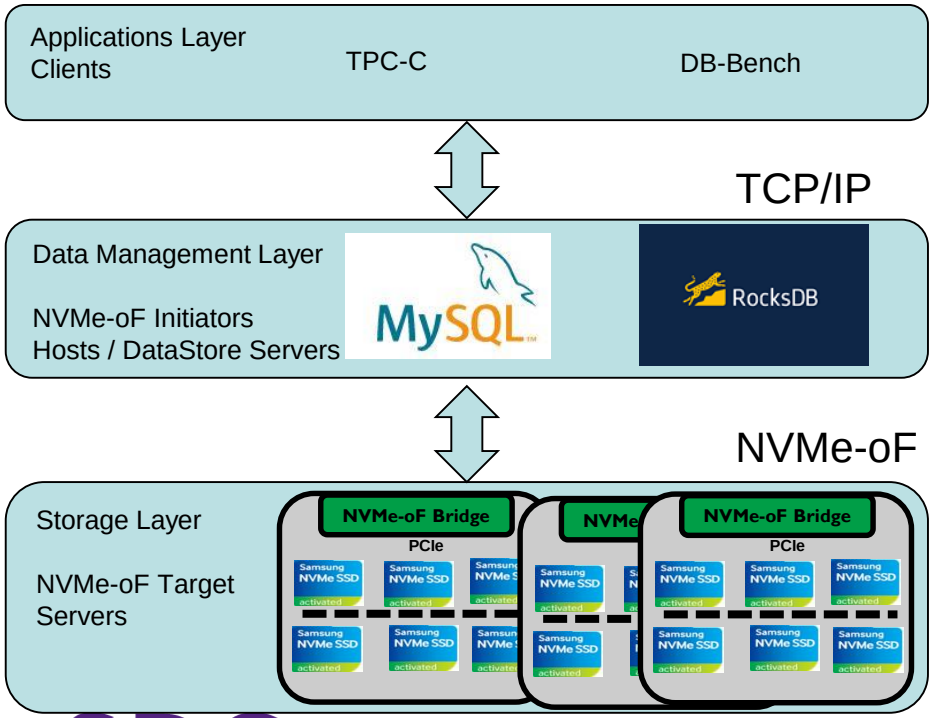


Requirement: Scale & Manage IOPS, Bandwidth and CPU independently

Storage Disaggregation

- Separates compute and storage to different nodes
  - High Speed Networks (25/50/100Gb)
  - H/W accelerated low overhead protocols (RoCE , iWARP)
  - High density flash
- Enables independent resource scaling
  - Allows flexible infrastructure tuning to dynamic loads
  - Reduces resource underutilization
  - Improves cost-efficiency by eliminating waste
- Remote access introduces overhead
  - Additional interconnect latencies
  - Network/protocol processing affect both storage and compute nodes

NVMe SSD Disaggregation



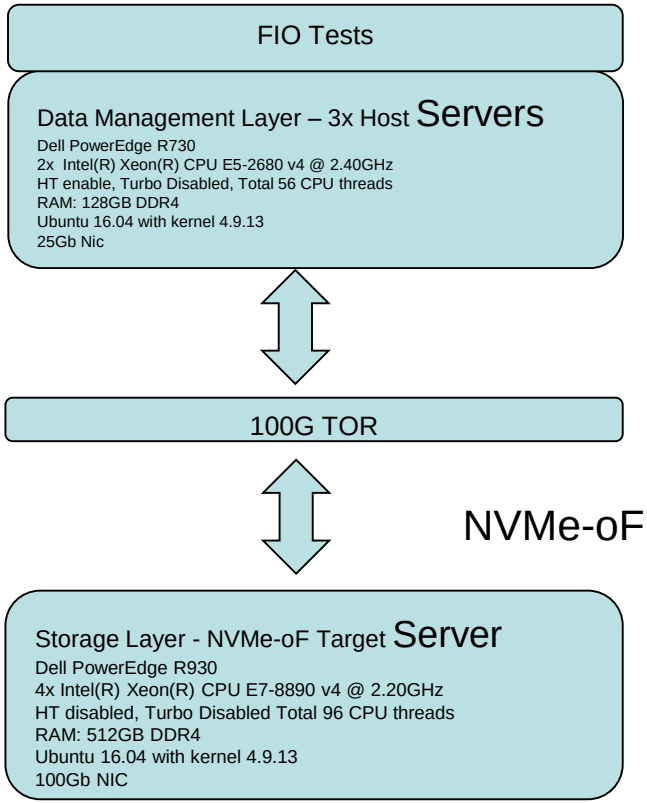
- NVMe disaggregation is more challenging
- ~90µs latency
  - Network/protocol latencies are more pronounced
- ~1MIOPS / Device
  - Protocol overhead tax the CPU and degrade performance

Performance Analysis

- FIO
  - Synthetic test to establish performance limits
- RocksDB
  - Representing KV Stores and NoSQL databases
- MySQL
  - Representing traditional RDBMS

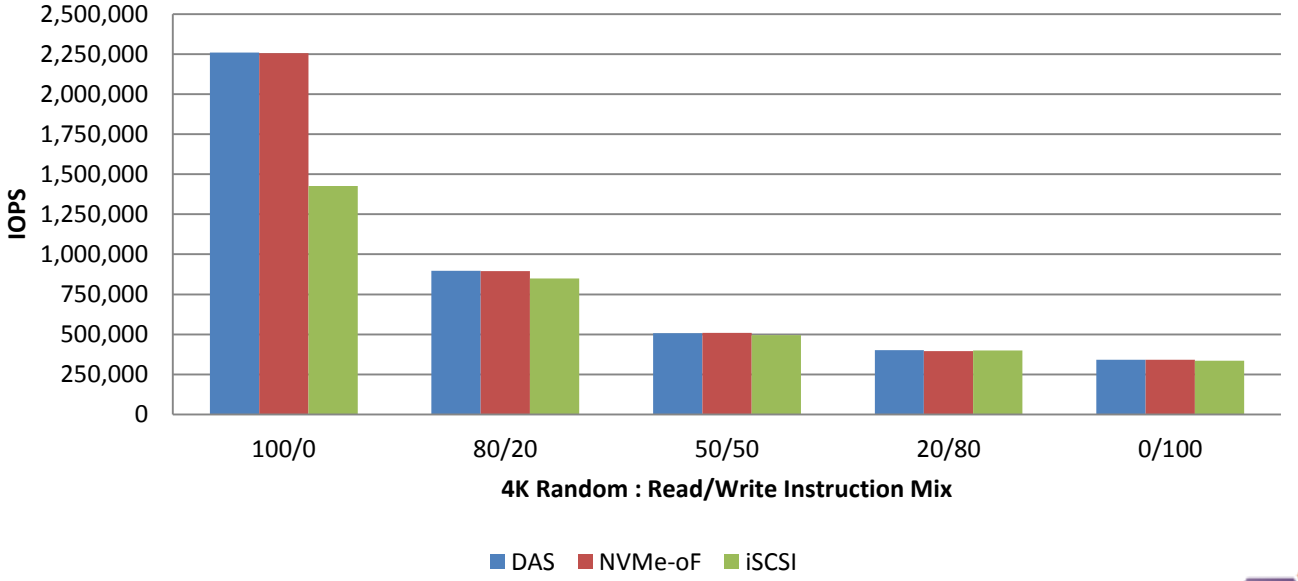
FIO Methodology

- Three configurations:
  1. Baseline: **Local** (direct-attached)
  2. Remote storage with **NVMe-oF** over RoCEv2
  3. Remote storage with **iSCSI**: Tuned by applying best known methods
- Hardware setup:
  - 3 *host* servers (a.k.a. *compute nodes, or datastore* servers)
  - 1 *target* server (a.k.a. *storage* server)
  - Samsung PM1725 NVMe-SSDs (8 drives max)
    - Random: 750/120 KIOPS read/write
    - Sequential: 3000/2000 MB/sec read/write
  - Network:
    - ConnectX-4 100Gb Ethernet NICs with RoCE support
    - 100Gb top-of-rack switch
- FIO
  - Version : 2.1.11
  - Per Host : QD = 32 , Jobs = 16



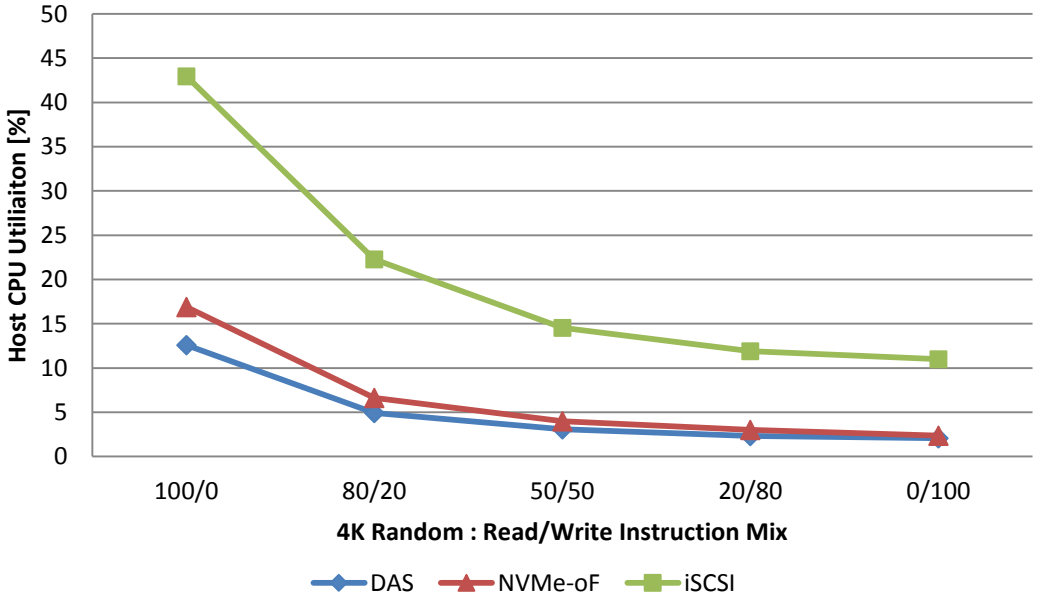
FIO Maximum Throughput

- NVMe-oF throughput is the same as DAS



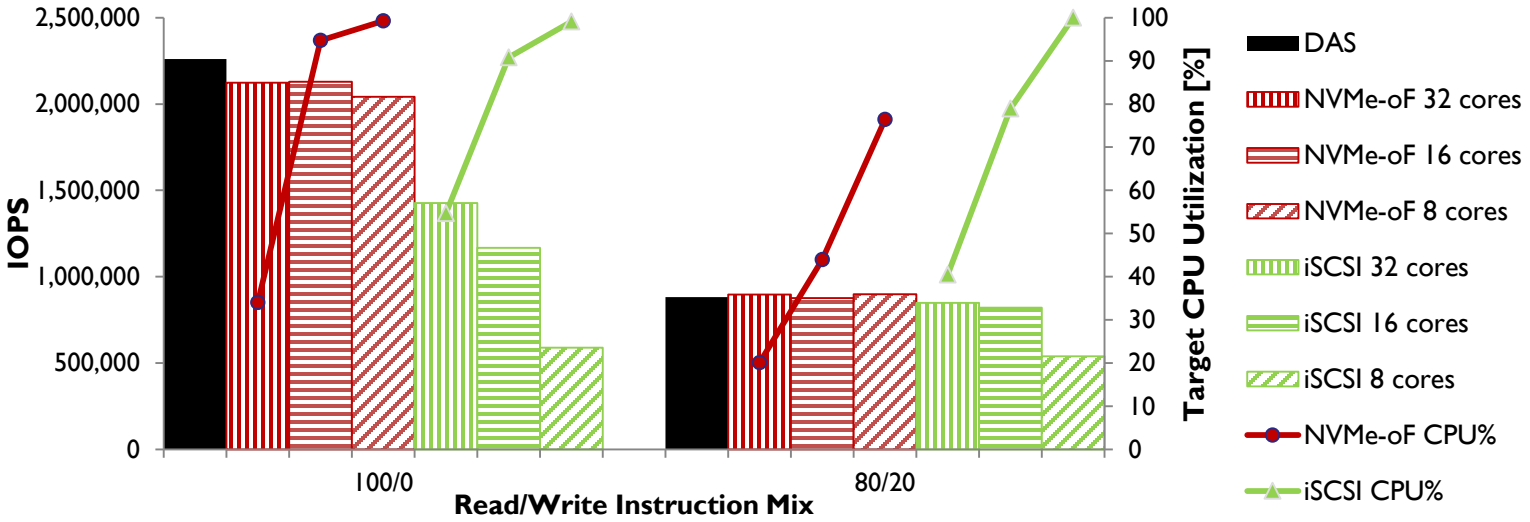
FIO Host CPU Overhead

- CPU processing overhead is minimal



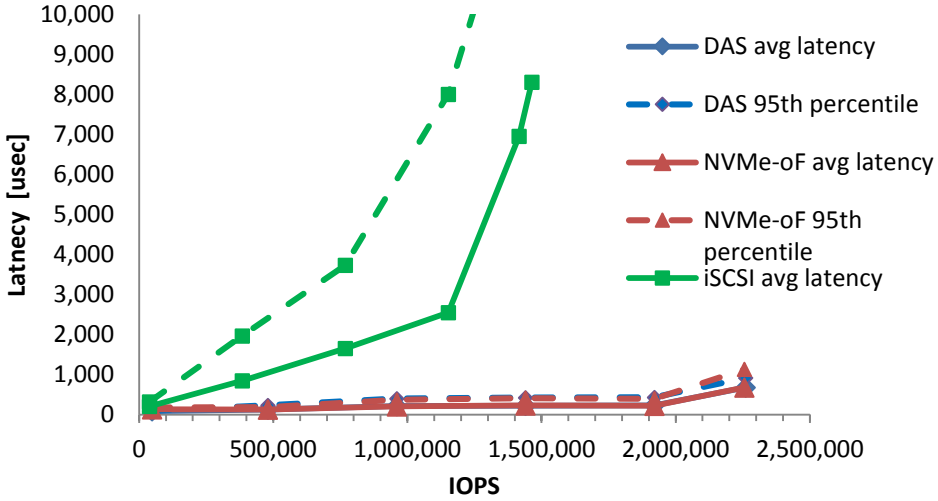
FIO Target Server Overhead

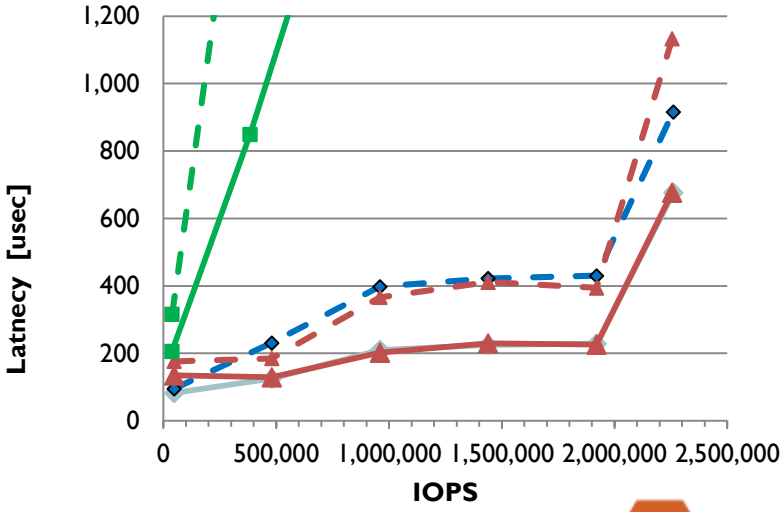
- CPU processing requirements are low
  - 90% of DAS read-only throughput with 1/12th of the target cores
- NVMe-oF requires less CPU to deliver consistent perf



FIO Latency Under Load

- NVMe-oF latencies similar to DAS for all practical loads
  - Both average and tail

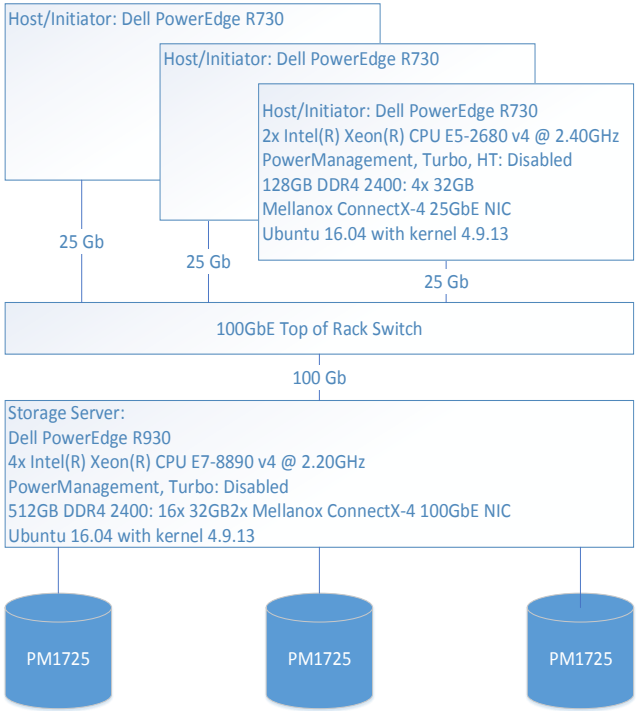




RocksDB

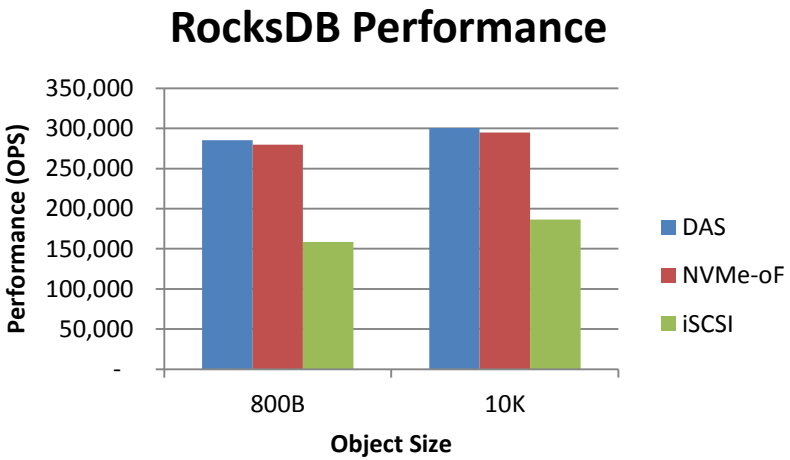
- Persistent Key-Value Store optimized for Flash
- Used as data store in many applications
  - MongoDB, MySQL, Redis-On-Flash, etc.

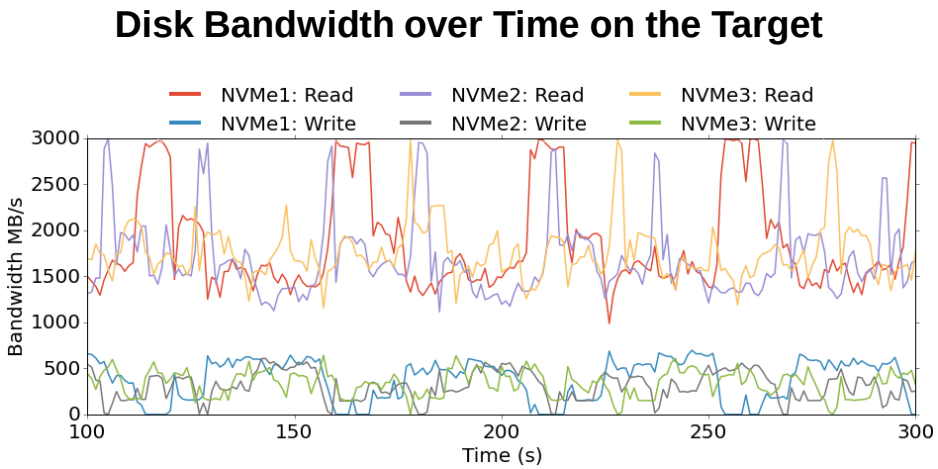
- Benchmark: DB_Bench
  - 800B and 10KB objects
  - 80%-20% read-write mix



RocksDB Throughput

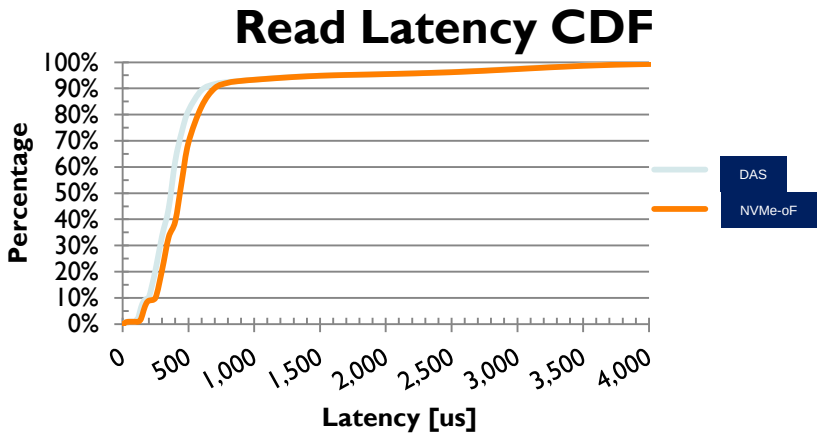
- NVMe-oF performance on-par with DAS
  - 2% throughput difference





RocksDB Latency

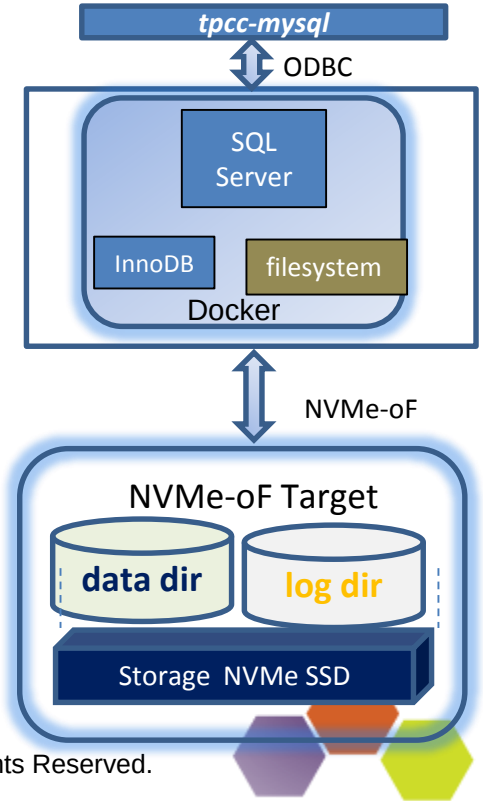
- NVMe-oF performance on-par with DAS
  - 2% throughput difference
  - Average latency increase by 11%, tail latency increase by 2%
    - Average Latency: 507µs ←→ 568µs
    - 99th percentile: 3.6ms ←→ 3.7ms
  - 10% CPU utilization overhead on host



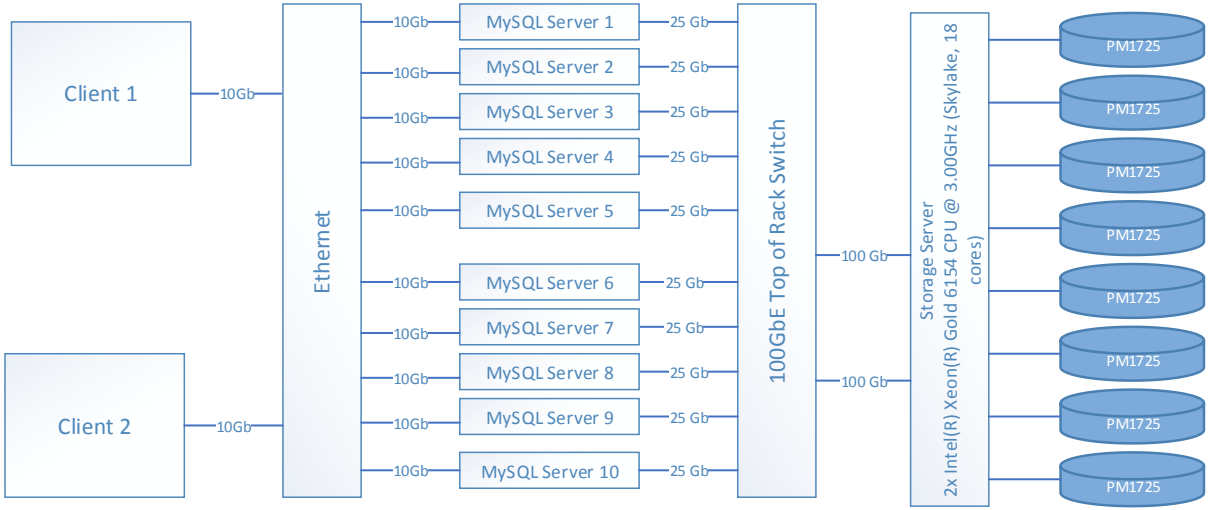
MySQL and TPC-C Setup

MySQL
- Version 5

TPC-C
- 500 Warehouse Setup with 150 Connections



Disaggregated Storage Setup



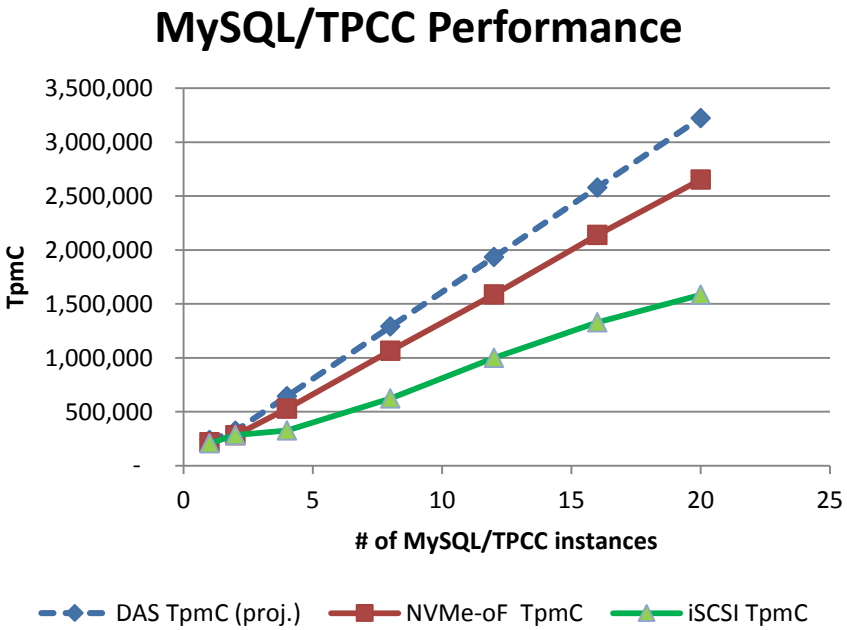
Client & Hosts

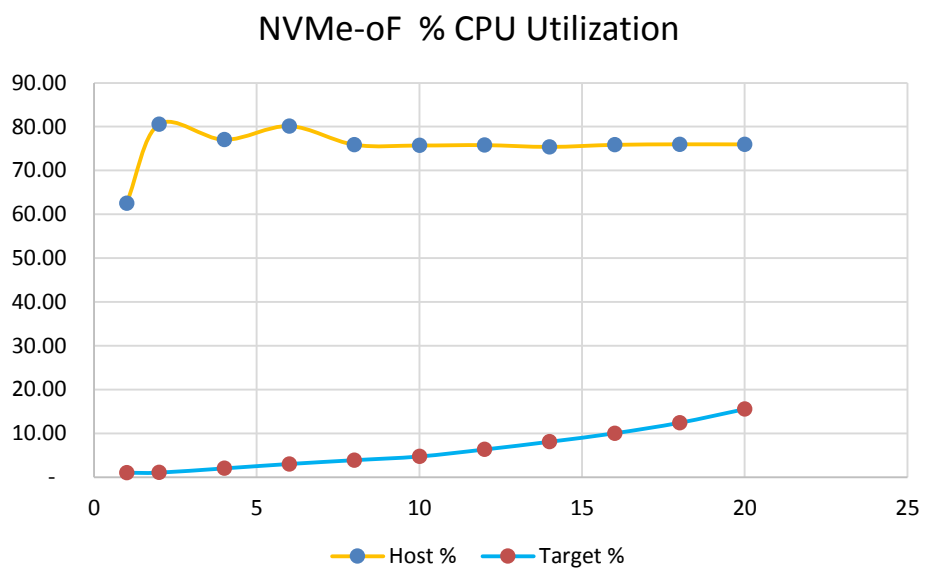
Dell PowerEdge R730
2x Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz
HT enable, turbo enabled, Total 88 CPU threads
RAM: 128GB DDR4.
Ubuntu 16.04 with kernel 4.9.13
Clients 10Gbe NIC
Hosts : 1x Mellanox ConnectX-4 25GbE NIC

Target Server

Supermicro
2x Intel(R) Xeon(R) Gold 6154 CPU @ 3.00GHz
HT enable, turbo enabled, Total 72 CPU threads
RAM: 384GB DDR4.
Ubuntu 16.04 with kernel 4.9.13
2x Mellanox ConnectX-4 100GbE NIC

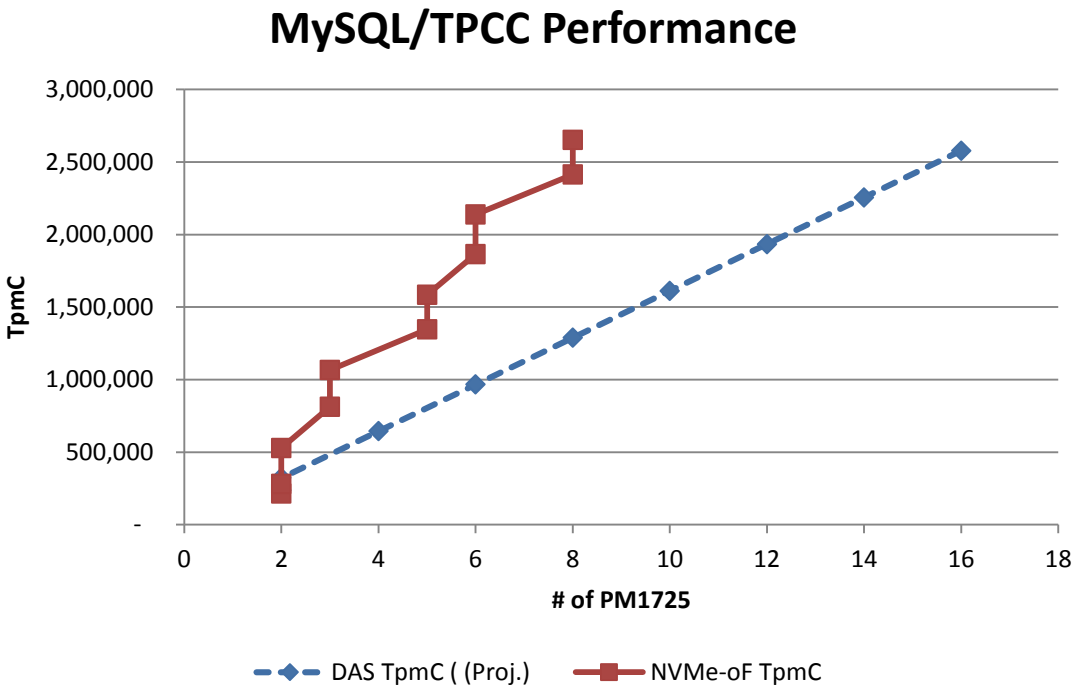
MySQL TPC-C Performance

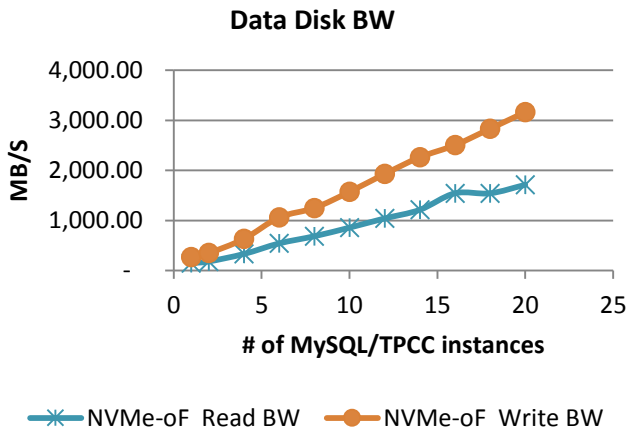


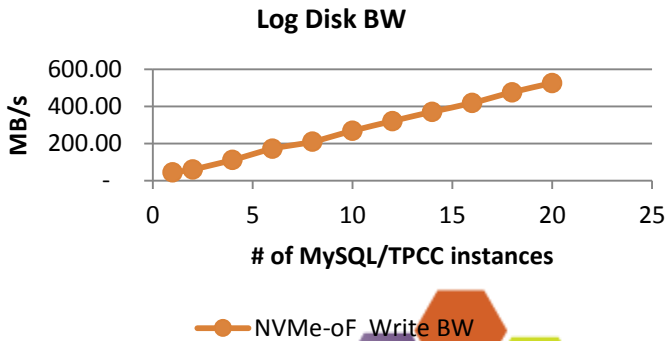


- NVMe-oF delivers scalable performance

MySQL/TPCC: Storage Analysis

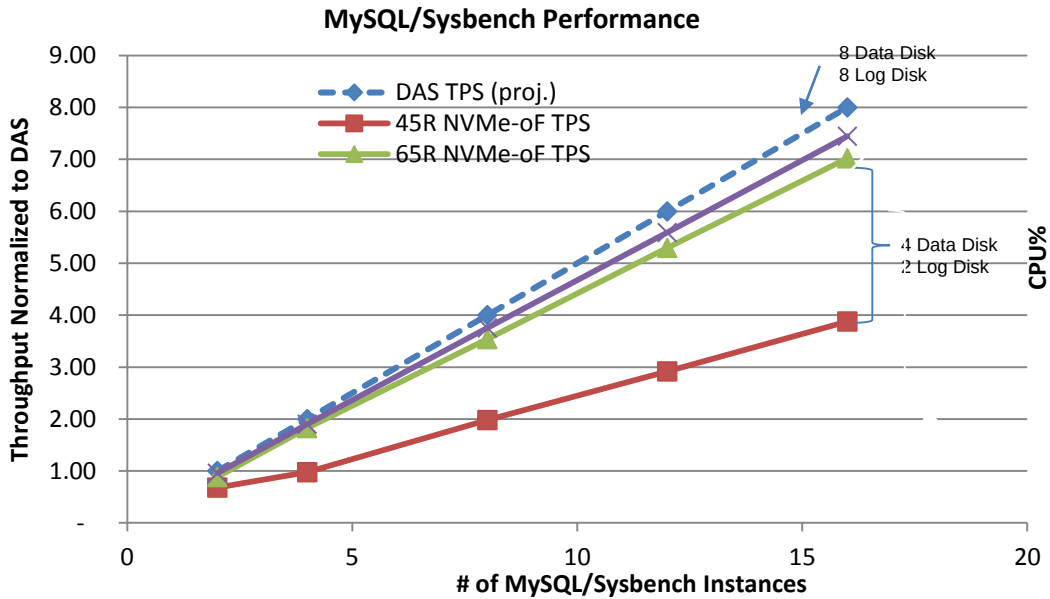


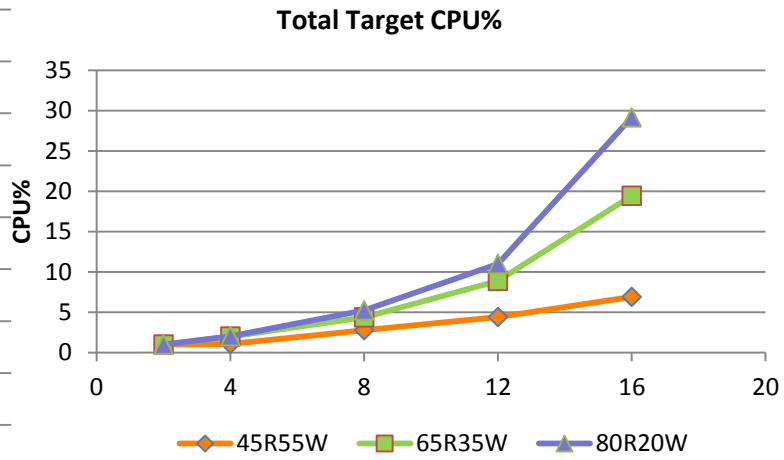




- Fewer Drives : Efficient utilization of NVMe SSDs
- Scalable Performance

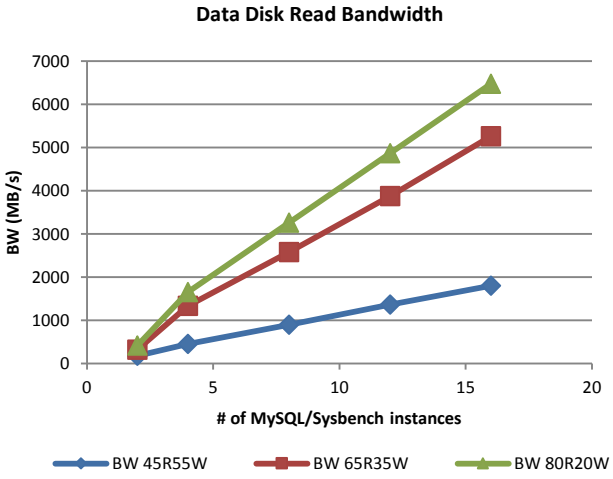
MySQL Sysbench Performance

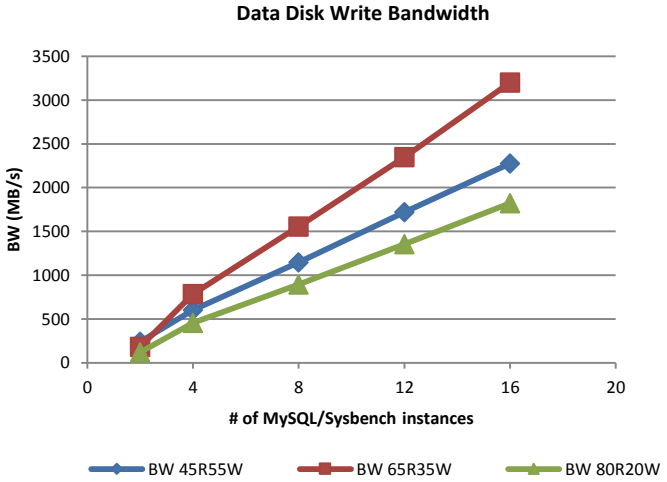




- NVMe-oF delivers scalable performance with fewer drives
- Low target CPU utilization

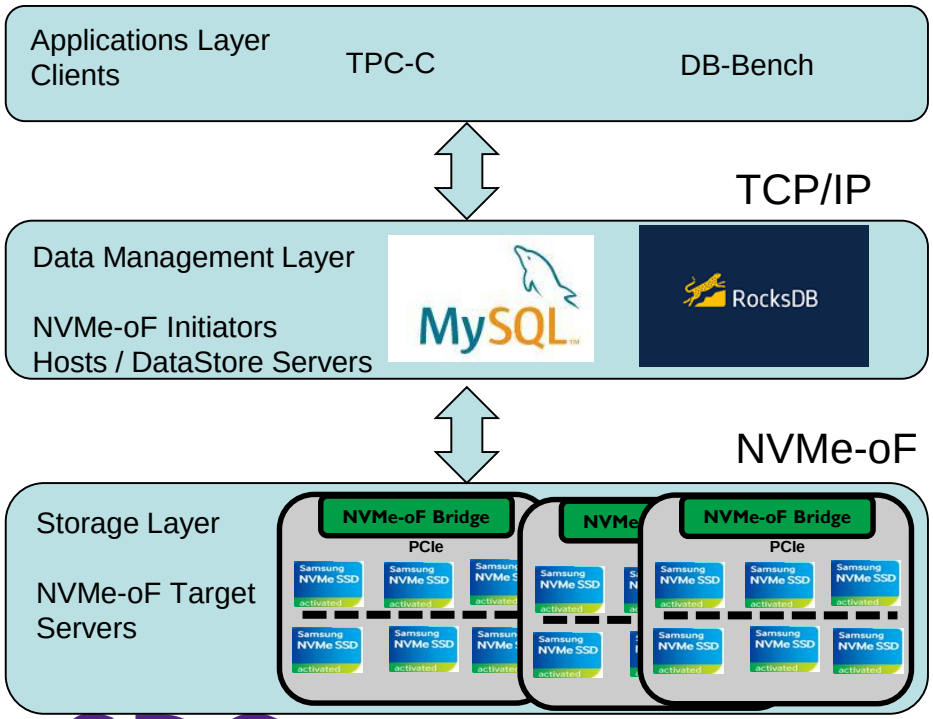
MySQL / Sysbench - Storage Analysis





- Efficient utilization of disk bandwidth
- Scale number of disks as required by application

NVMe-oF Ecosystem Maturing



- Drivers
- Operating Systems

- NVMe SSD
- High Speed Networks
- RDMA Enabled Hardware

Conclusions

- NVMe-oF reduces remote storage overhead to a bare minimum
- Low processing overhead on both host and target
  - Applications (*host*) gets the same performance
  - Storage server (*target*) can support more drives with fewer cores
- NVMe SSD + NVMe-oF enables efficient disaggregation architecture for flash

- Thanks 

- http://www.nvmexpress.org/

vijay.bala@samsung.com