SAMSUNG

NVMe-over-Fabrics Performance Characterization and the Path to Low-Overhead Flash Disaggregation

Zvika Guz, Harry Li, Anahita Shayesteh, and Vijay Balakrishnan

Memory Solution Lab
Samsung Semiconductor Inc.

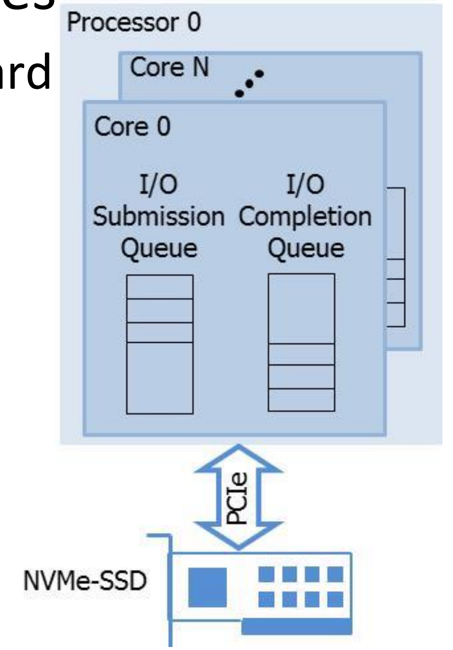
Synopsis

Performance characterization of NVMe-oF in the context of Flash disaggregation

- Overview
  - NVMe and NVMe-over-Fabrics
  - Flash disaggregation
- Performance characterization
  - Stress-testing remote storage
  - Disaggregating RocksDB
- Summary

Non-Volatile Memory Express (NVMe)

- A storage protocol standard on top of PCIe:
  - Standardize access to local non-volatile memory over PCIe
- The predominant protocol for PCIe-based SSD devices
  - NVMe-SSDs connect through PCIe and support the standard
- 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



Storage Disaggregation

- Separates compute and storage to different nodes
  - Storage is accessed over a network rather than locally
- Enables independent resource scaling
  - Allow flexible infrastructure tuning to dynamic loads
  - Reduces resource underutilization
  - Improves cost-efficiency by eliminating waste
- Remote access introduces overheads
  - Additional interconnect latencies
  - Network/protocol processing affect both storage and compute nodes
- HDD disaggregation is common in datacenters
  - HDD are so slow that these overheads are negligible

~~Storage~~ Flash Disaggregation

- NVMe disaggregation is more challenging
  - ~90µs latency → network/protocol latencies are more pronounced
  - ~1MIOPS → protocol overheads tax the CPU and degrade performance
- Flash disaggregation via iSCSI is difficult:
  - iSCSI "introduces 20% throughput drop at the application level"*
  - Even then, it can still be a cost-efficiency win
- We show that these overheads go away with NVMe-oF

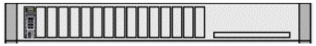
*A. Klimovic, C. Kozyrakis, E. Thereska, B. John, and S. Kumar, "**Flash storage disaggregation**," EuroSys'16

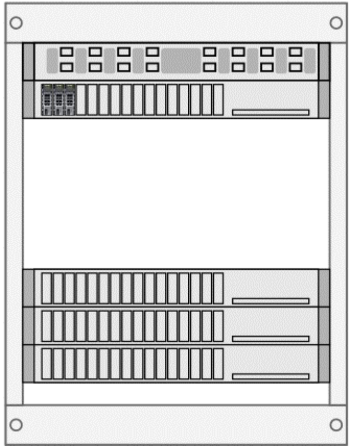
NVMe-oF: NVMe-over-Fabrics

- Recent extension of the NVMe standard
  - Enables access to remote NVMe devices over different network fabrics
- Maintains the current NVMe architecture, and:
  - Adds support for message-based NVMe operations
- Advantages:
  - Parallelism: extends the multiple queue-paired design of NVMe
  - Efficiency: eliminates protocol translations along the I/O path
  - Performance
- Supported fabrics:
  - RDMA – InfiniBand, iWarp, RoCE
  - Fiber Channel, FCoE

Methodology

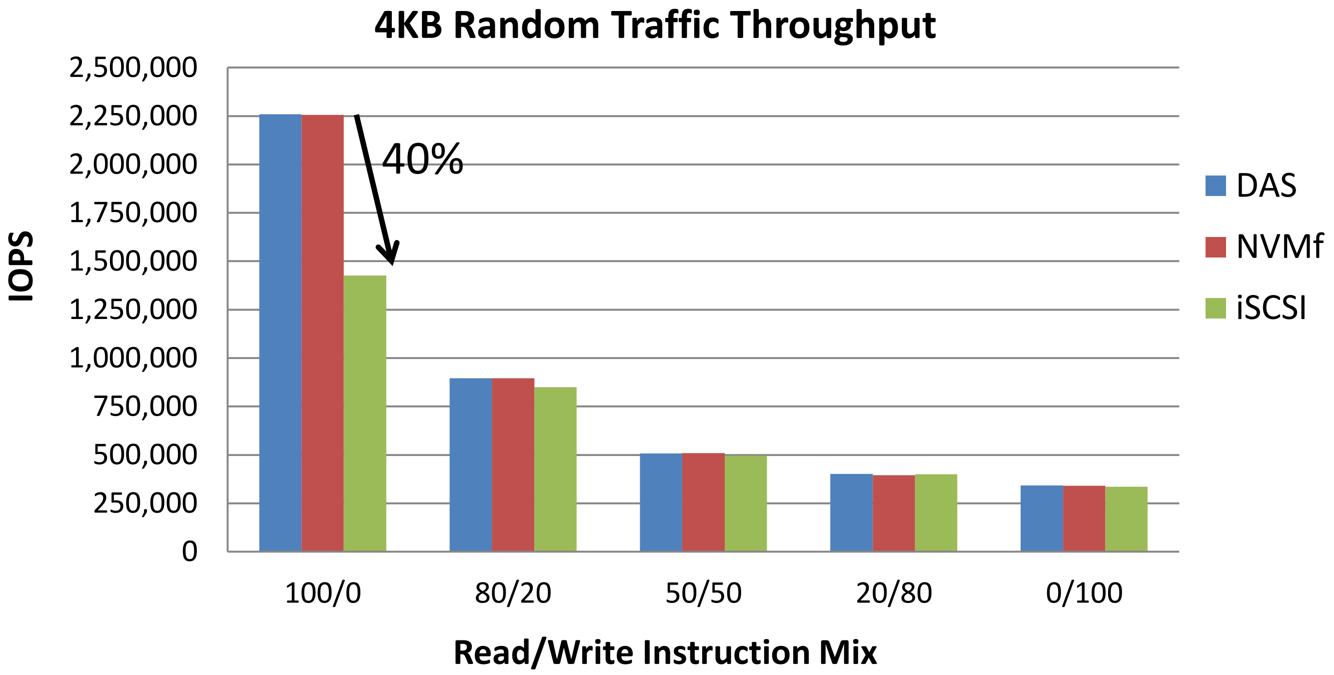
- Three configurations:
  1. Baseline: Local, (direct-attached)
  2. Remote storage with NVMe-oF over RoCEv2
  3. Remote storage with iSCSI
     - Followed best-known-methods for tuning
- Hardware setup:
  - 3 *host* servers (a.k.a. *compute nodes, or datastore* servers)
    - Dual-socket Xeon E5-2699
  - 1 *target* server (a.k.a. *storage* server)
    - Quad-socket Xeon E7-8890
  - 3x Samsung PM1725 NVMe-SSDs
    - 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


Baseline: direct-attached (DAS)


Remote storage setup

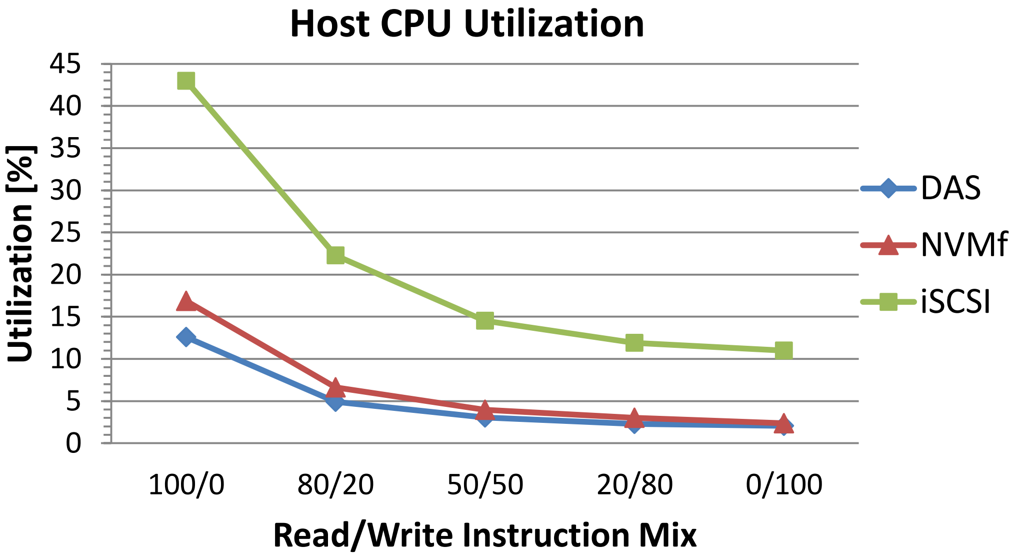
Maximum Throughput

- NVMe-oF throughput is the same as DAS
  - iSCSI cannot keep up for high IOPS rates



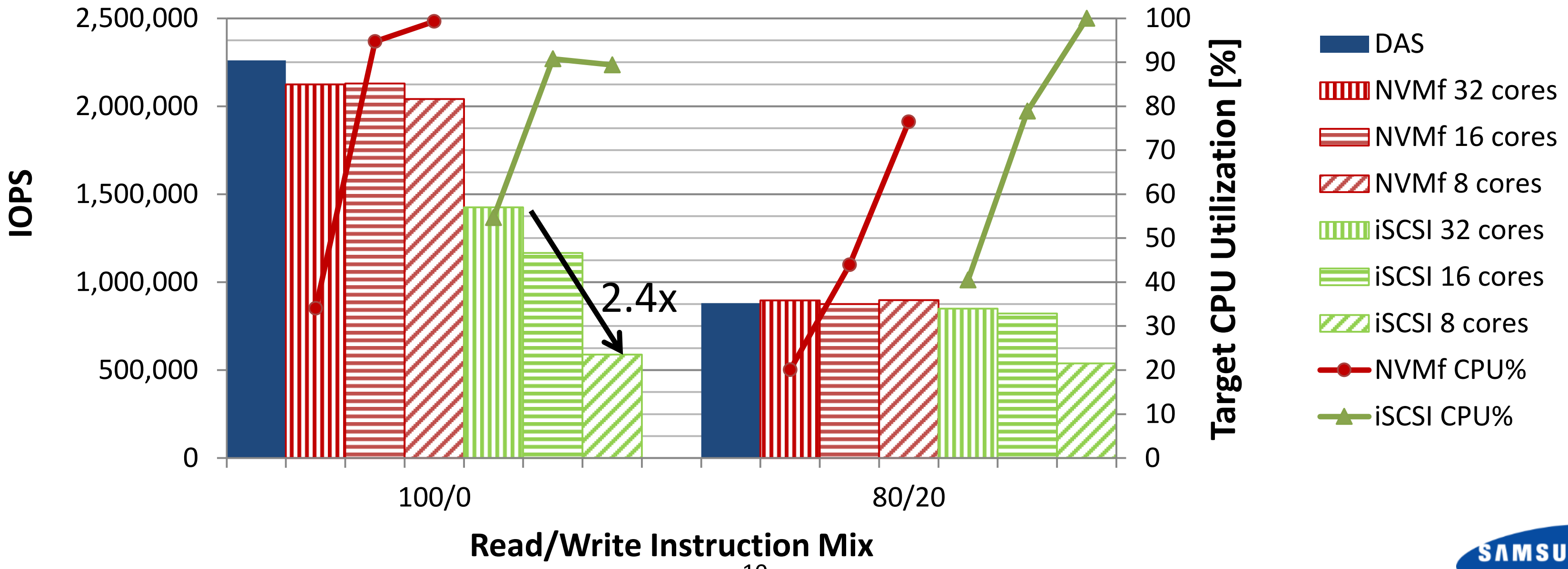
Host CPU Overheads

- NVMe-oF CPU processing overheads are minimal
  - iSCSI adds significant load on the host (30%)
    - Even when performance is on par with DAS



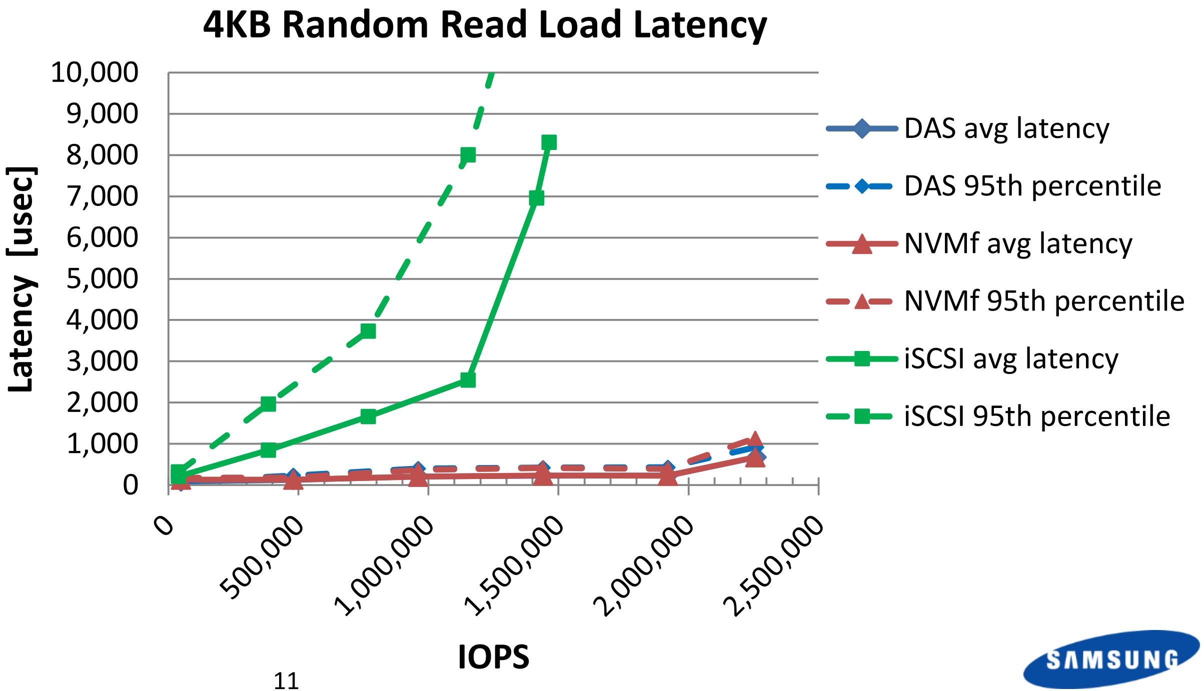
Storage Server CPU Overheads

- CPU processing on target is limited
  - 90% of DAS read-only throughput with 1/12th of the cores
- Cost efficiency win: fewer cores per NVMe-SSD in the server



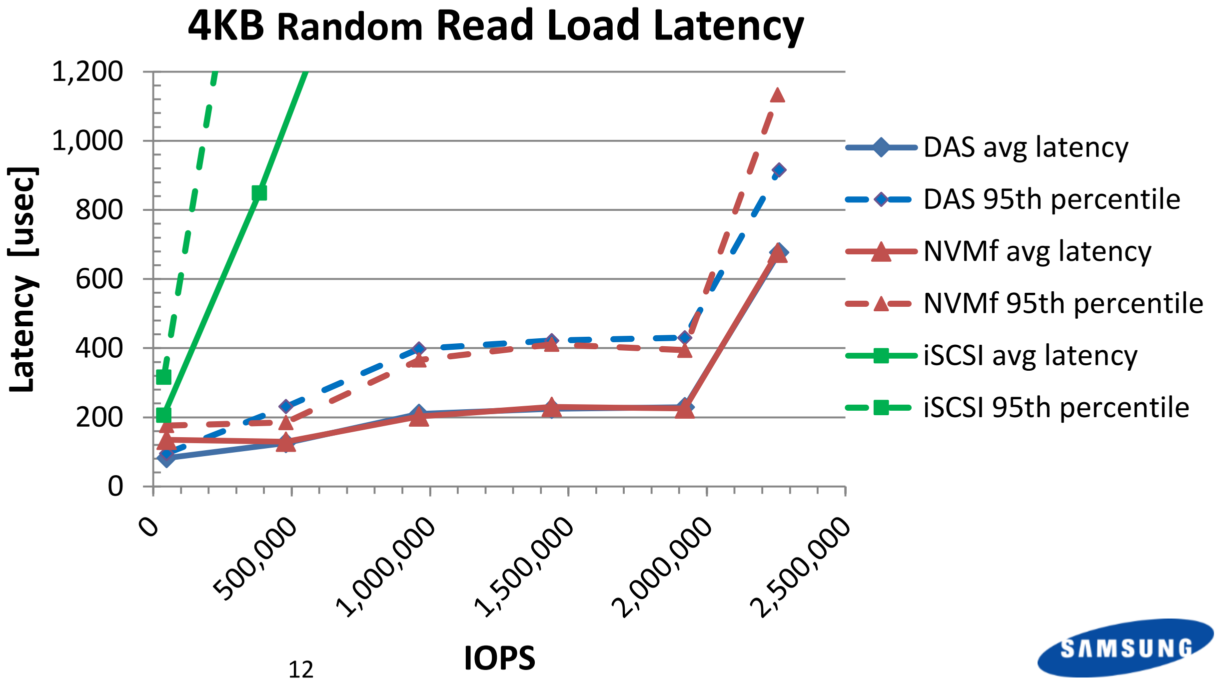
Latency Under Load

- NVMe-oF latencies are the same as DAS for all practical loads
  - Both average and tail
- iSCSI:
  - Saturates sooner
  - 10x slower even under light loads



Latency Under Load

- NVMe-oF latencies are the same as DAS for all practical loads
  - Both average and tail
- iSCSI:
  - Saturates sooner
  - 10x slower even under light loads

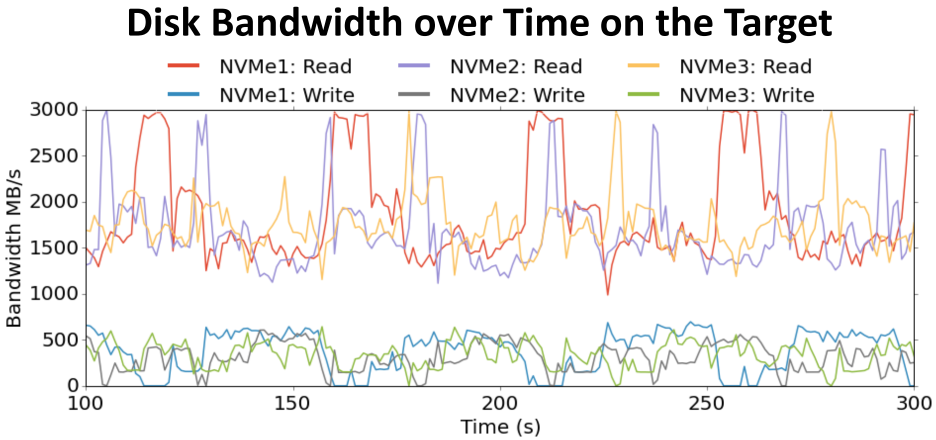


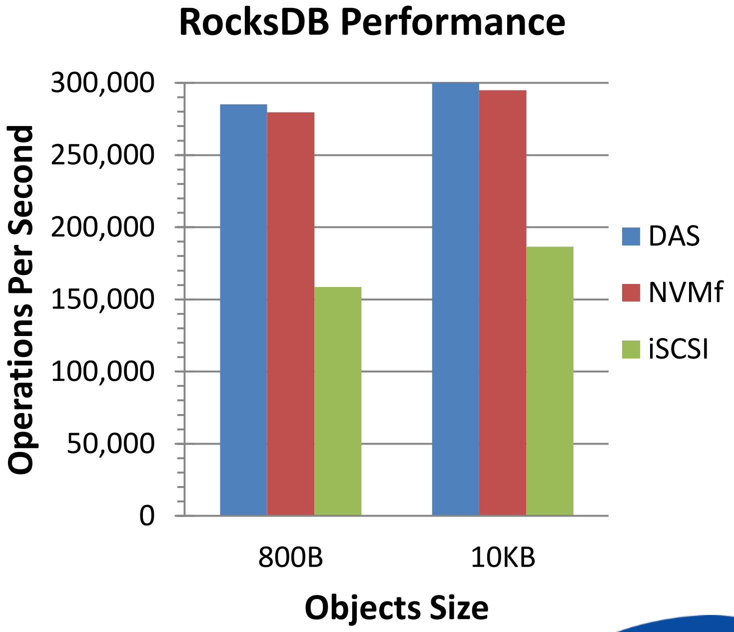
KV-Store Disaggregation (1/3)

- Evaluated using RocksDB, driven with db_bench
  - 3 hosts
  - 3 rocksdb instances per host
  - 800B and 10KB objects
  - 80/20 read-write mix

KV-Store Disaggregation (2/3)

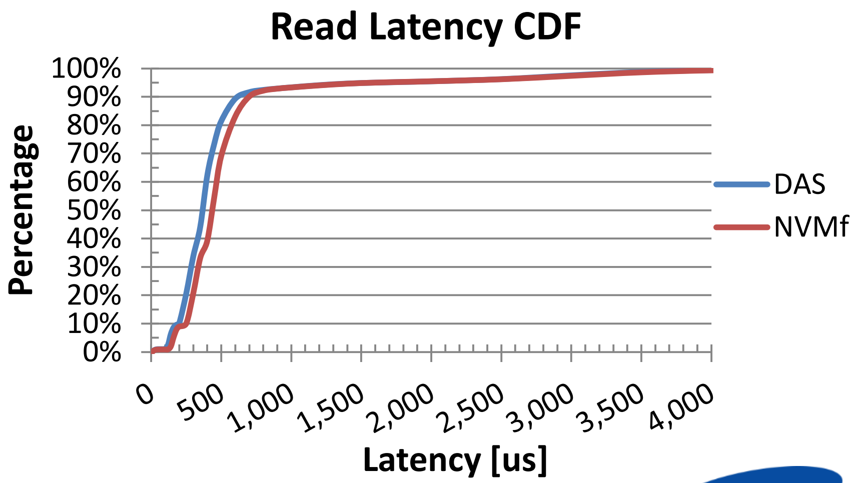
- NVMe-oF performance on-par with DAS
  - 2% throughput difference
    - vs. 40% performance degradation for iSCSI





KV-Store Disaggregation (3/3)

- NVMe-oF performance on-par with DAS
  - 2% throughput difference
    - vs. 40% performance degradation for iSCSI
  - 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



Summary

- NVMe-oF reduces remote storage overheads to a bare minimum
  - Negligible throughput difference, similar latency
  - Low processing overheads on both host and target
    - Applications (*host*) gets the same performance
    - Storage server (*target*) can support more drives with fewer cores
- NVMe-oF makes disaggregation more viable
  - No need to offset iSCSI >>20% performance lose

Thank You!

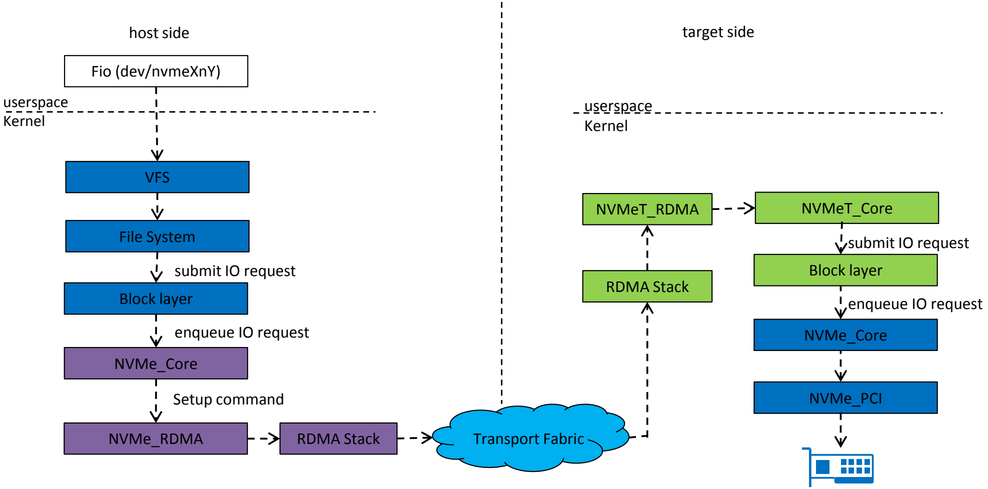
zvika.guz@samsung.com

SAMSUNG

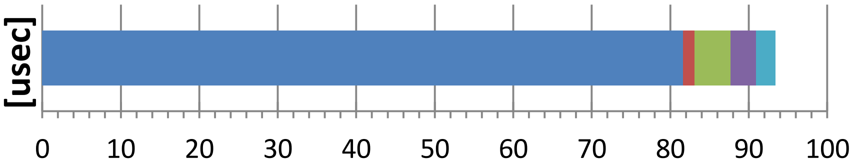
Backup

Unloaded Latency Breakdown

- NVMe-oF adds 11.7µs over DAS access latency
  - Close to the 10µs spec target



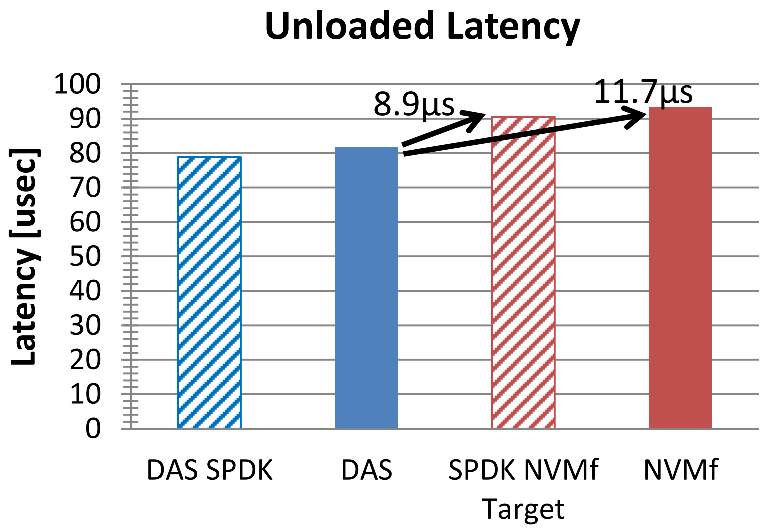
4K Unloaded Read Latency



NVMe DAS Path	81.6
Others	1.52
NVMf Target Modules	4.57
NVMf Host Modules	3.25
Fabric	2.43

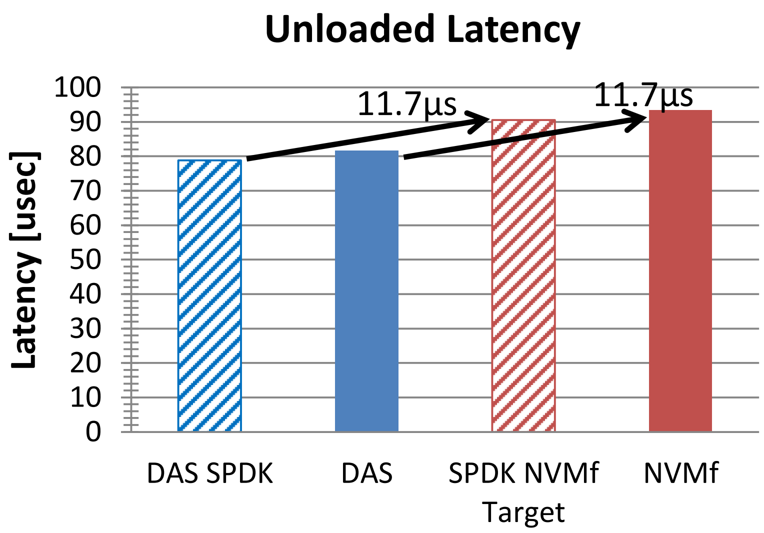
FAQ #1: SPDK

- Storage Performance Development Kit (SPDK)
  - Provides user-mode storage drivers
    - NVMe, NVMe-oF target, and NVMe-oF host
  - Better performance through:
    - Eliminating kernel context switches
    - Polling rather than interrupts
- Will improve NVMe-oF performance
  - **BUT**, was not stable enough for our setup
- For unloaded latency:
  - SPDK target further reduces latency overhead
  - SPDK local ←→ SPDK target similar to local ←→ NVMe-oF



FAQ #1: SPDK

- Storage Performance Development Kit (SPDK)
  - Provides user-mode storage drivers
    - NVMe, NVMe-oF target, and NVMe-oF host
  - Better performance through:
    - Eliminating kernel context switches
    - Polling rather than interrupts
- Will improve NVMe-oF performance
  - **BUT**, was not stable enough for our setup
- For unloaded latency:
  - SPDK target further reduces latency overhead
  - SPDK local ←→ SPDK target similar to local ←→ NVMe-oF



SAMSUNG

FAQ #2: Hyper-convergence vs. Disaggregation

- Hyper-convergence Infrastructure (HCI)
  - Software-defined approach
  - Bundles commodity servers into a clustered pool
  - Abstract underlining hardware into a virtualized computing platform
- We focus on web-scale data centers
  - Disaggregation fits well within their deployment model
    - Several classes of server, some of which are storage-centric
    - Already disaggregate HDD
- NVMe-oF, HCI, and disaggregation are not mutually exclusive
  - HCI on-top of NVMe-oF
  - Hybrid architectures

SAMSUNG