Delivering NoSQL Database Performance with NVMe SSD's

Vijay Balakrishnan
Manu Awasthi
Zvika Guz
Qiumin Xu

Memory Solutions Lab.
Samsung

Disclaimer

This presentation is intended to provide information concerning SSD technology. We do our best to make sure that information presented is accurate and fully up-to-date. However, the presentation may be subject to technical inaccuracies, information that is not up-to-date or typographical errors. As a consequence, Samsung does not in any way guarantee the accuracy or completeness of information provided on this presentation. Samsung reserves the right to make improvements, corrections and/or changes to this presentation at any time.

The information in this presentation or accompanying oral statements may include forward-looking statements. These forward-looking statements include all matters that are not historical facts, statements regarding the Samsung Electronics' intentions, beliefs or current expectations concerning, among other things, market prospects, growth, strategies, and the industry in which Samsung operates. 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 contained in this presentation or in the accompanying oral statements. In addition, even if the information contained herein or the oral statements are shown to be accurate, those developments may not be indicative developments in future periods. The information is provided as a general understanding and not directly representing any product.

Agenda

- NVMe SSD
- Samsung PM1725 NVMe SSD
- Redis-On-Flash with PM1725
  - Deliver >1MOPS @ < 1ms latency consistently
- PM1725 as NVMf target for Cassandra
  - Build efficient remote storage for databases

NVMe Design Advantages

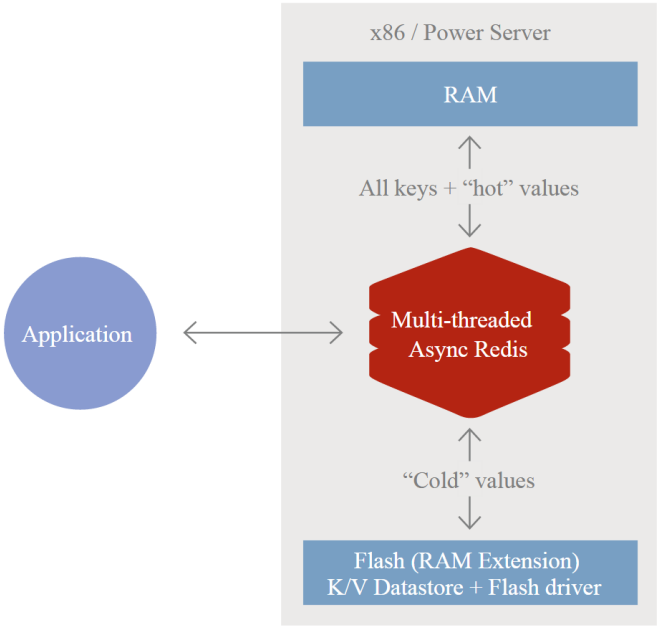
- Lower latency
  - Direct connection to CPU's PCIe lanes
- Higher bandwidth
  - Scales with number of PCIe lanes
- Best in class latency consistency
  - Lower cycles/IO, fewer commands, better queueing
- Lower system power
  - No HBA required

PM1725

Samsung PM1725 Specification	
Form Factor	2.5”
Host Interface	PCIe Gen3 x4
Capacities	800GB, 1.6TB, 3.2TB
Sequential Read	3300 MB/s
Sequential Write	1900 MB/s
Random Read	Upto 840KIOPS
Random Write	Upto 130KIOPS
Read Latency	134 usec
Write Latency	68 usec

- Leverages latest VNAND technology
- Delivers consistent low latency

Redis-on-Flash

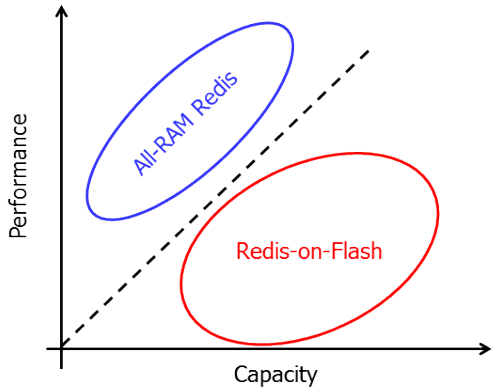


- Closed-source (RLEC Flash)
  - 100% compatible with the open-source Redis
- Uses Flash as RAM extension
  - Increases effective node capacity
- Tiering memory into "*fast*" and "*slow*":
  - RAM saves keys and hot values
  - Flash saves cold values
- Dynamic configuration of RAM/Flash usage
- Uses RocksDB as the storage engine to optimize access to block storage
- Multi-threaded and asynchronous Redis used to access Flash

Get it Here Today: https://redislabs.com/rlec-flash

Why Redis-on-Flash?

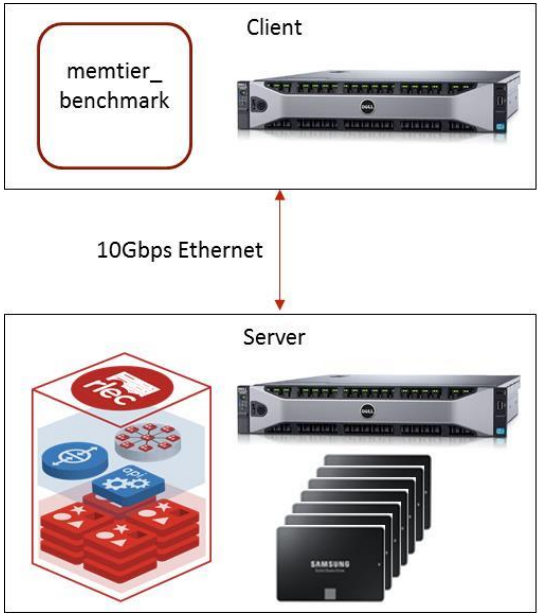
- Optimize price-to-performance for a given workload
  - DRAM is more performant than flash, but $/GB is higher
    - Limited DRAM capacity per server
  - Tiering dramatically reduces $/GB, while preserving good performance ($/ops)
    - Enables orders-of-magnitude more capacity per server
- RoF is suitable for large datasets with skewed access distribution



System Under Test

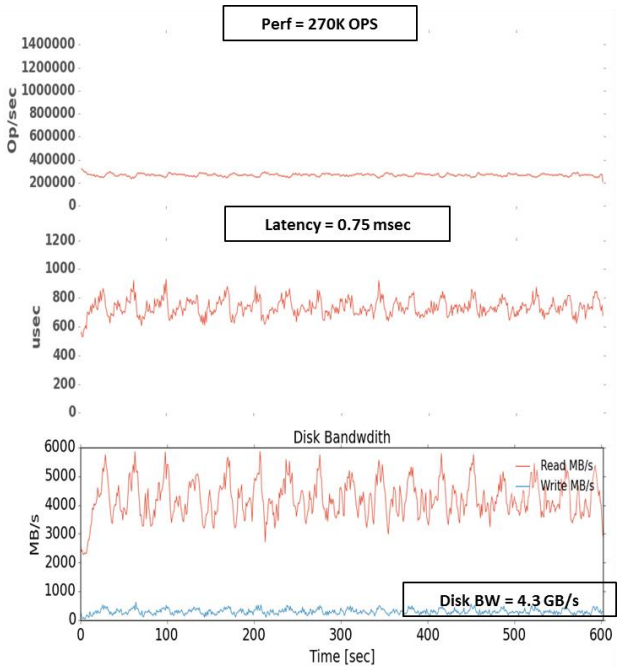
- Single client, single server
  - Industry-standard components, all available today

Server	Dell PowerEdge R730xd, dual-socket	
Processor	2 x Xeon E5-2690 v3 @ 2.6GHz 12 cores, 24 logical processor per CPU 24 cores, 48 logical processor total	
Memory	256GB ECC DDR4	
Network	10GbE	
Storage	4 x Samsung PM1725 NVMe	16 x Samsung 850PRO SATA SSD
Memtier_benchmark	1.2.6	
RLEC version	4.3.0	
Operating System	Ubuntu 14.04	
Linux Kernel	3.19.8	

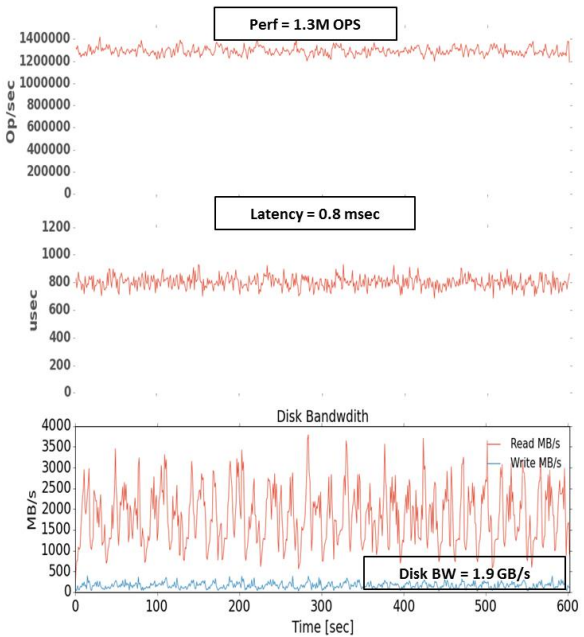


Use case 1: 1KB Objects R/W:80/20

50% RAM-to-Flash Hit ratio



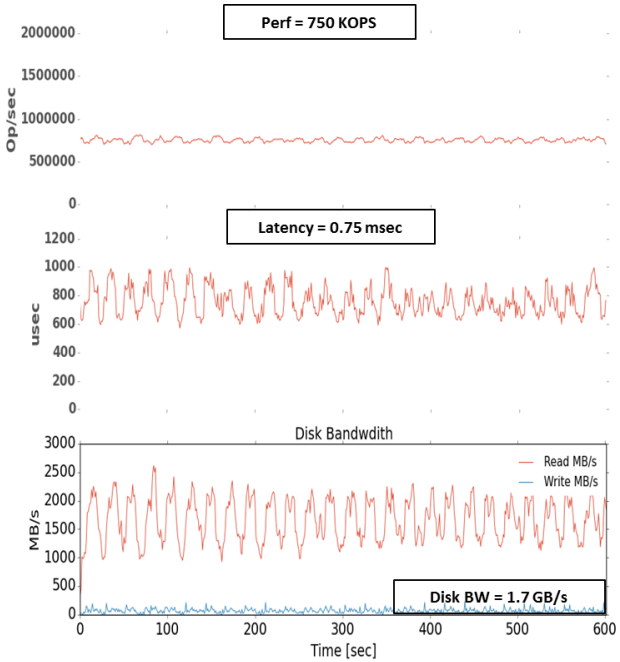
95% RAM-to-Flash Hit ratio



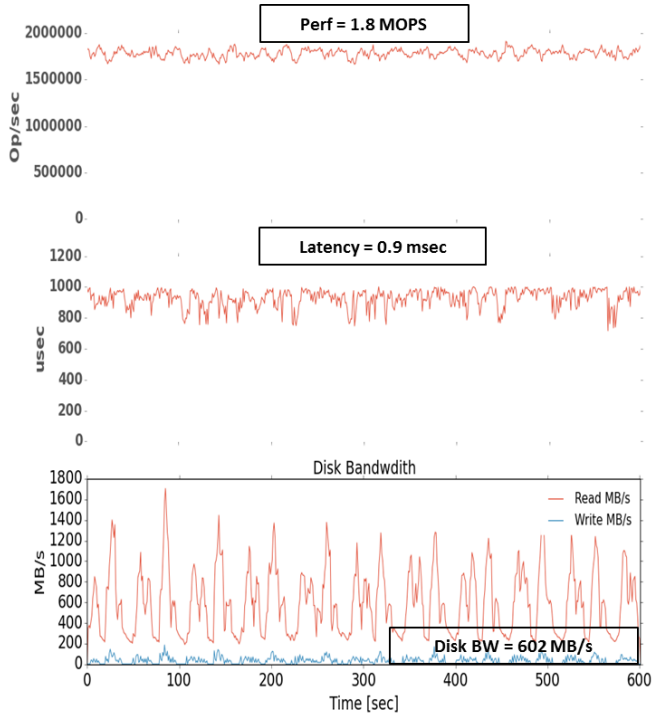
100% of requests served with <1msec latency

Use case 2: 100B Objects R/W : 50/50

50% RAM-to-Flash Hit ratio



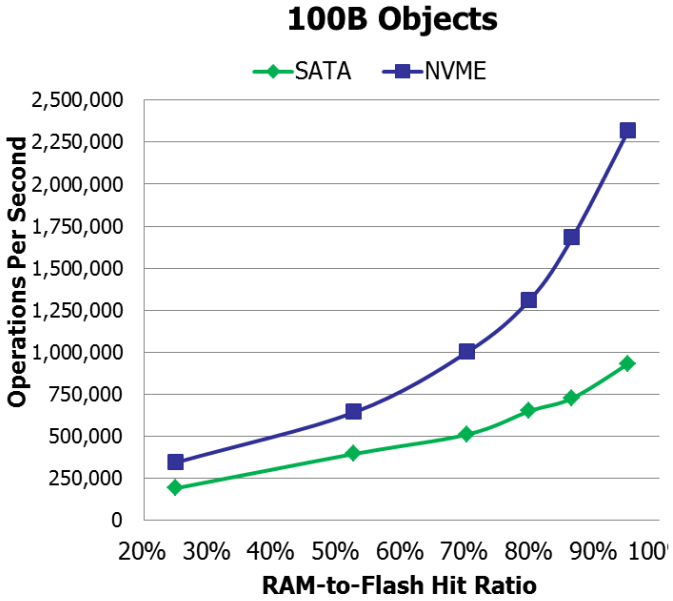
85% RAM-to-Flash Hit ratio

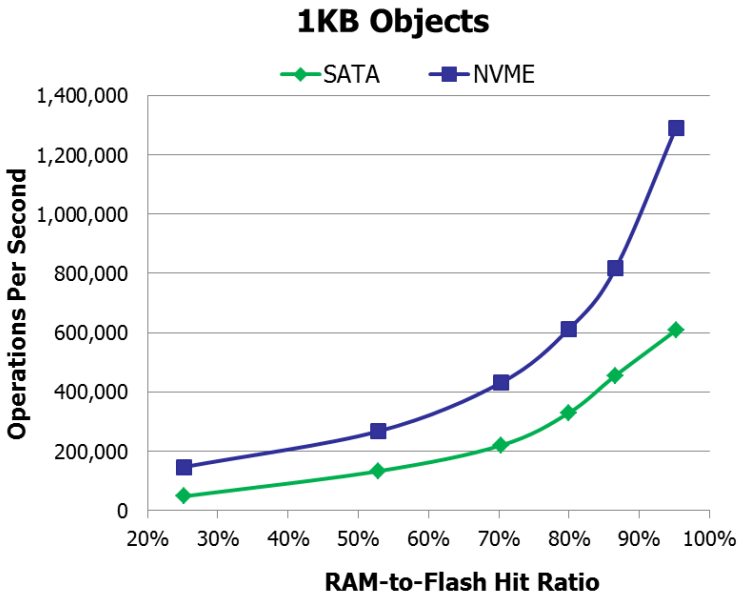


100% of requests served with <1msec latency

Comparison to SATA

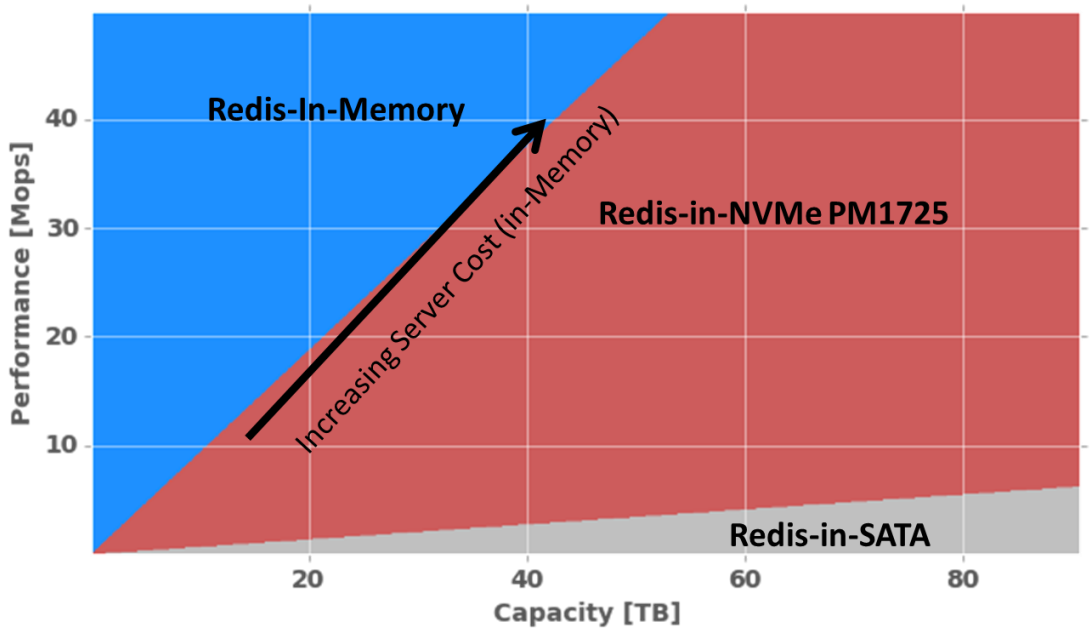
- **80/20 read-write ratio**



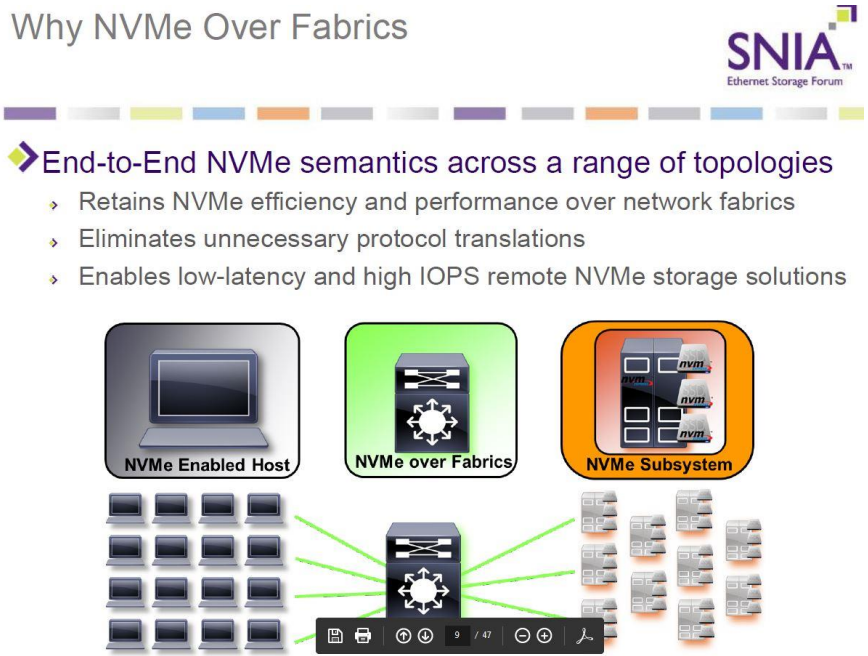


DRAM or Flash?

- Performance and Capacity



NVMe Over Fabrics (NVMf)



- Reference: http://www.snia.org/sites/default/files/ESF/NVMe_Under_Hood_12_15_Final2.pdf

Cassandra on NVMf storage

- Widely used open-source NoSQL
- We know that NVMe drives deliver improved performance & latency
- ➢ However, NVMe drives are underutilized (IOPS and BW)
- ➢ Can we use NVMf to deliver more efficient remote storage?

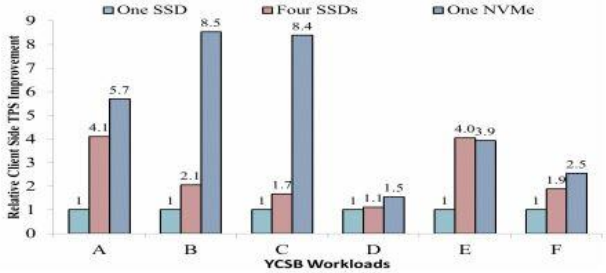
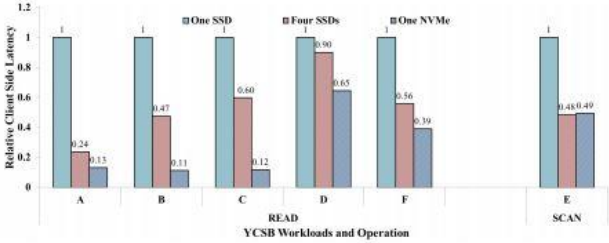


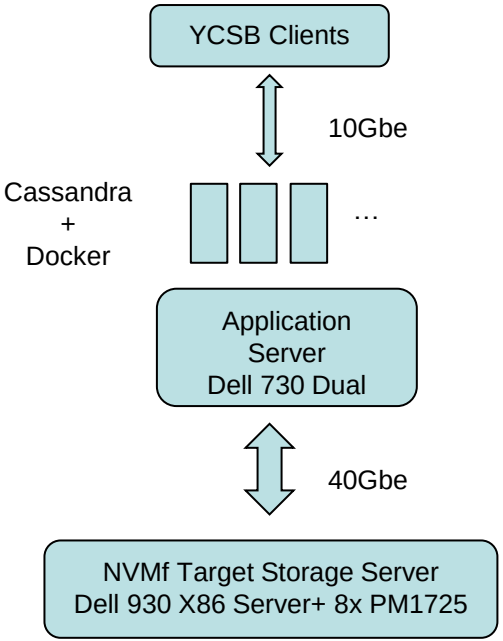
Figure 11: Comparison of Cassandra performance for different storage media.



Performance Analysis of NVMe SSDs and their Implication on Real World Databases

https://www.cs.utah.edu/~manua/pubs/systor15.pdf

System Configuration

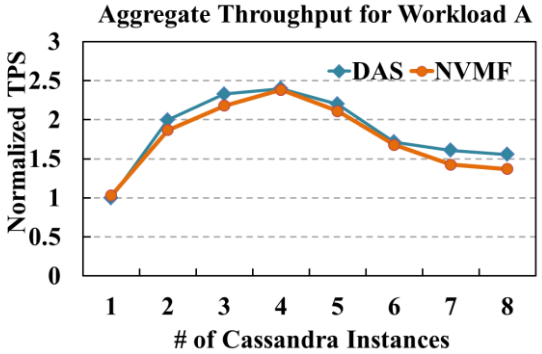


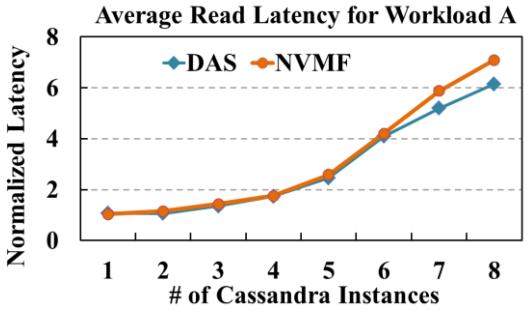
YCSB Workload:

- WorkloadA, 50/50 read/update, zipfian distribution
- WorkloadD, 95/5 read/insert, uniform distribution
- Record count: 100 million records, 100 GB in each database
- Client Thread count: 16

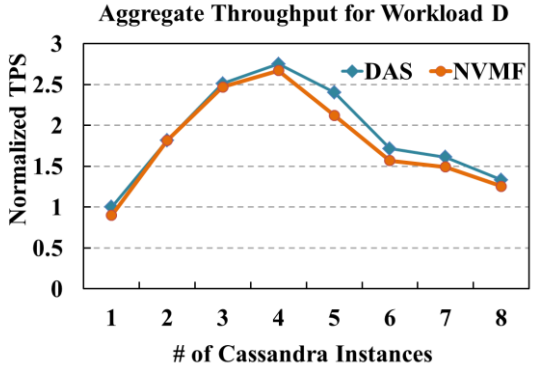
Cassandra Client Performance

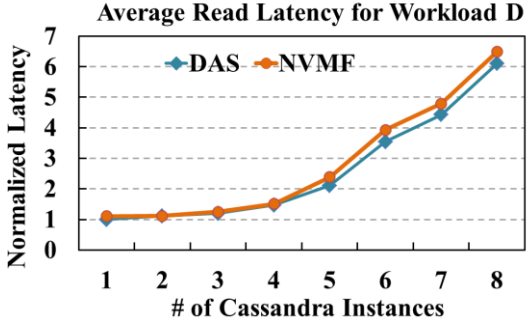
Workload A R50/U50





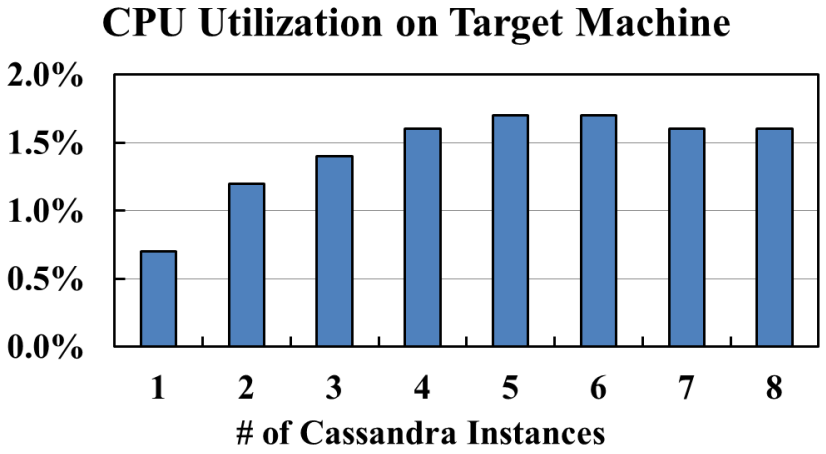
Workload D R95%I5%

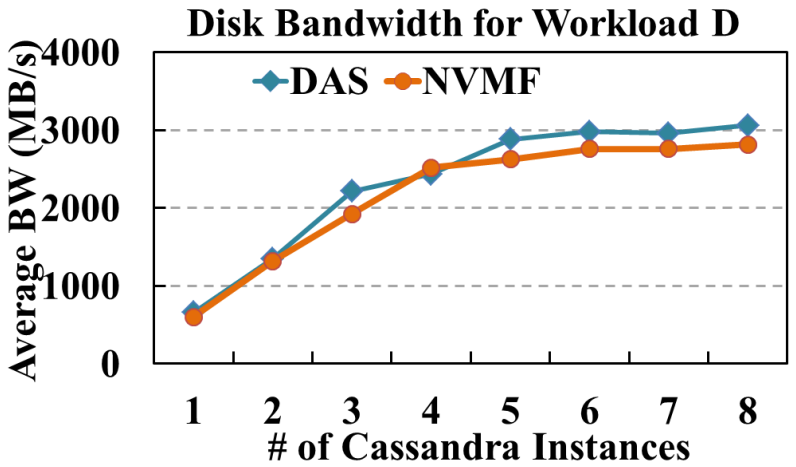




NVMe + NVMf tracks DAS performance with minor differences

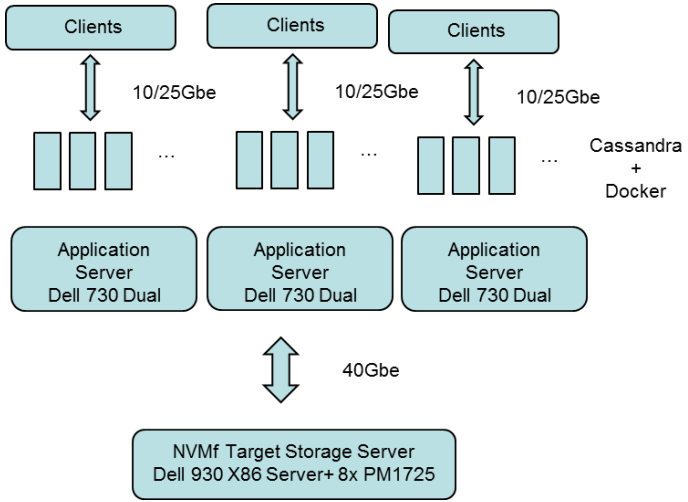
NVMEoF Target Performance





- Low Utilization on Target

Fast and Efficient Storage For Cassandra



- NVMf + PM1725 enables high-performance, efficient disaggregated storage
- Drive higher-utilization of storage systems and NVMe devices
- Call to action:
  - Add reliability features to NVMf
  - More performance improvements

NVMf enables high-performance, low latency remote storage for databases

Conclusions

- RedisOnFlash
  - PM1725 enables larger DBs with fewer servers
  - Maintains consistent < 1ms latency
  - Exceeds 1000K ops/sec for 100B-1000B objects
- Cassandra
  - PM1725 with NVMf target delivers a high performance and scalable NoSQL Solution

Thank You