Challenges of High-IOPS Enterprise-level NVMeoF-based All Flash Array

From the Viewpoint of Software Vendor

Dr. Weafon Tsao
R&D VP, AccelStor, Inc.

Enterprise-level Storage System (ESS)

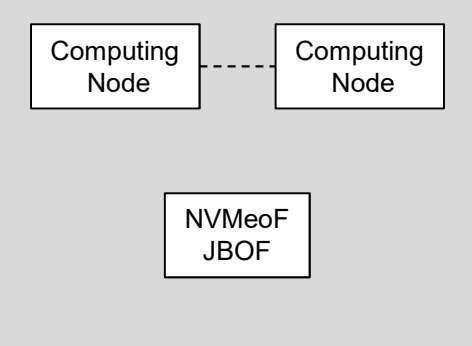
- Data Protection (DP)
- High Availability (HA)
- Snapshot
- Replication (Sync/Async)
- Thin-Provisioning
- Thin-Clone
- Deduplication
- Compression

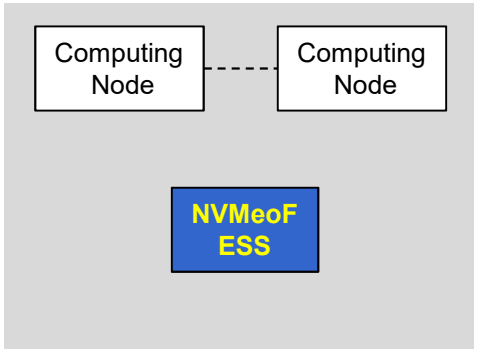
Disaster Recovery (DR)

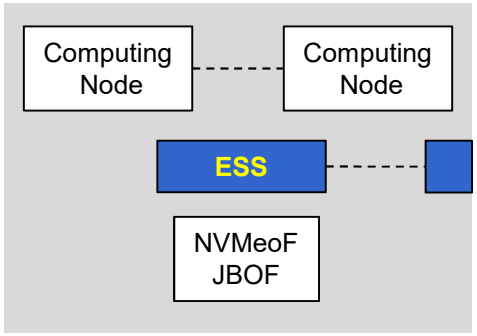
Data Reduction

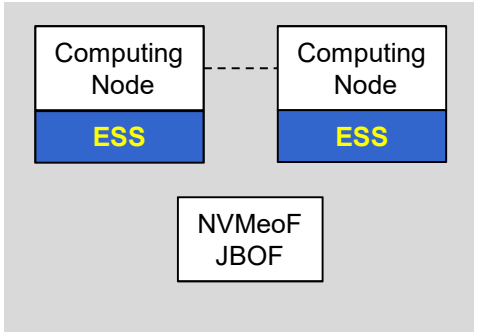
NVMeoF-based ESS

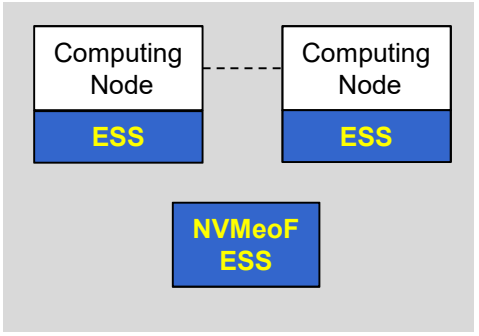
Plain Storage System

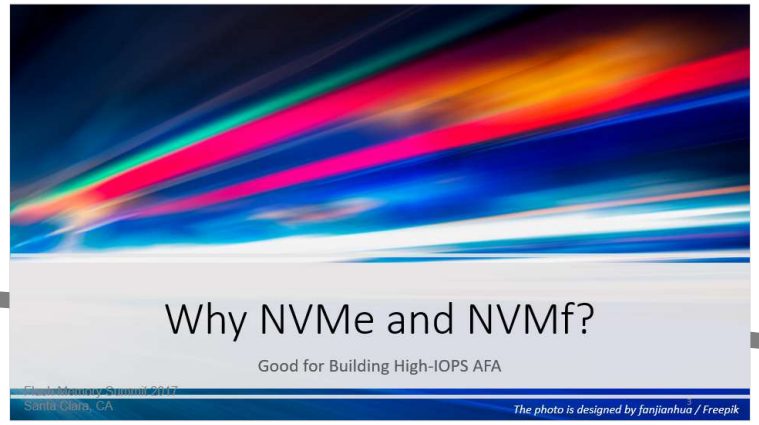


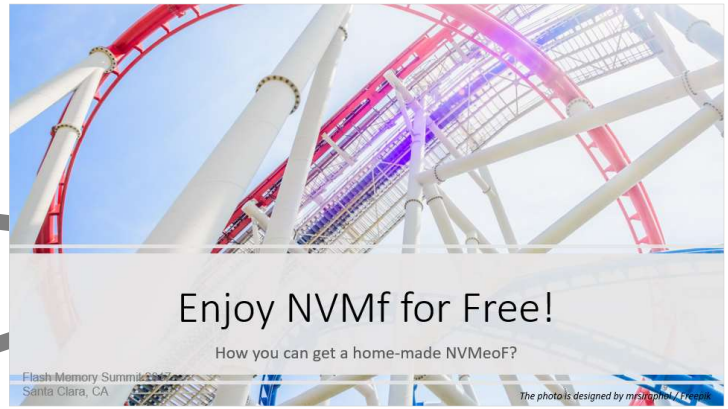


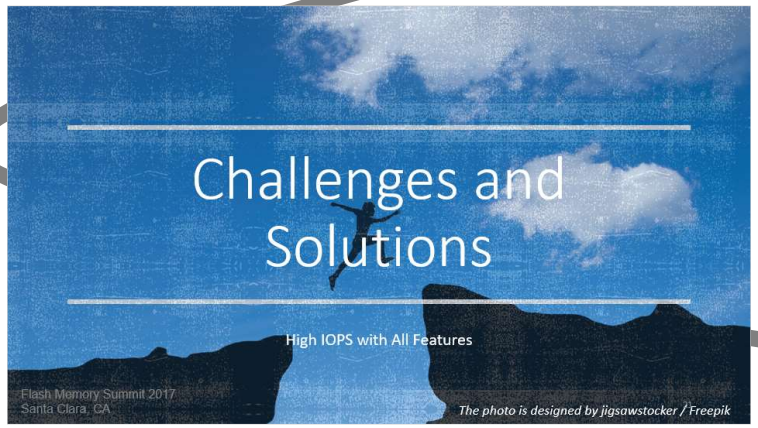







Why NVMe and NVMf?
Good for Building High-IOPS AFA
Flash Memory Summit 2017
Santa Clara, CA
The photo is designed by fanjianhua / Freepik


Enjoy NVMf for Free!
How you can get a home-made NVMeoF?
Flash Memory Summit 2017
Santa Clara, CA


Challenges and Solutions
High IOPS with All Features
Flash Memory Summit 2017
Santa Clara, CA
The photo is designed by jigsawstocker / Freepik

Why NVMe and NVMf?

Good for Building High-IOPS AFA

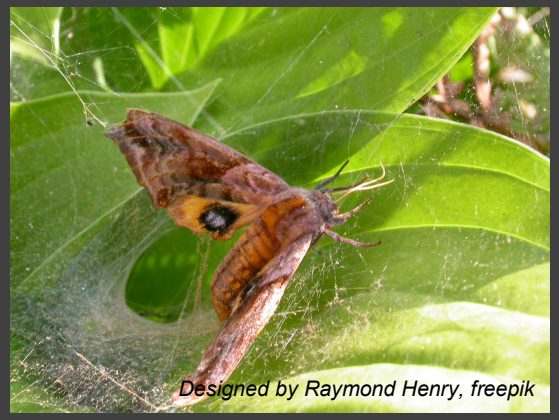
The photo is designed by fanjianhua / Freepik

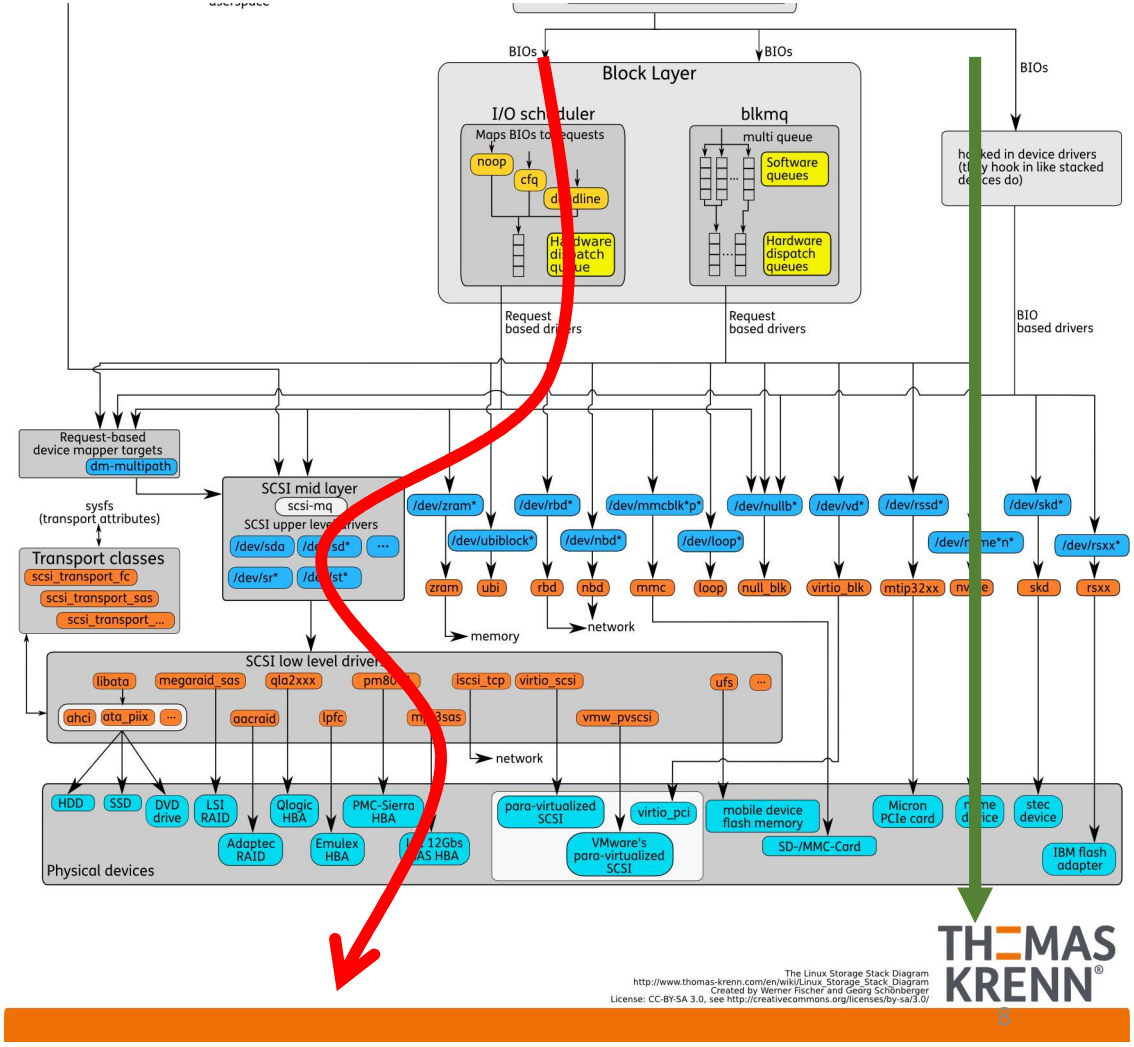
Dad! Dad!
~ an Interrupt

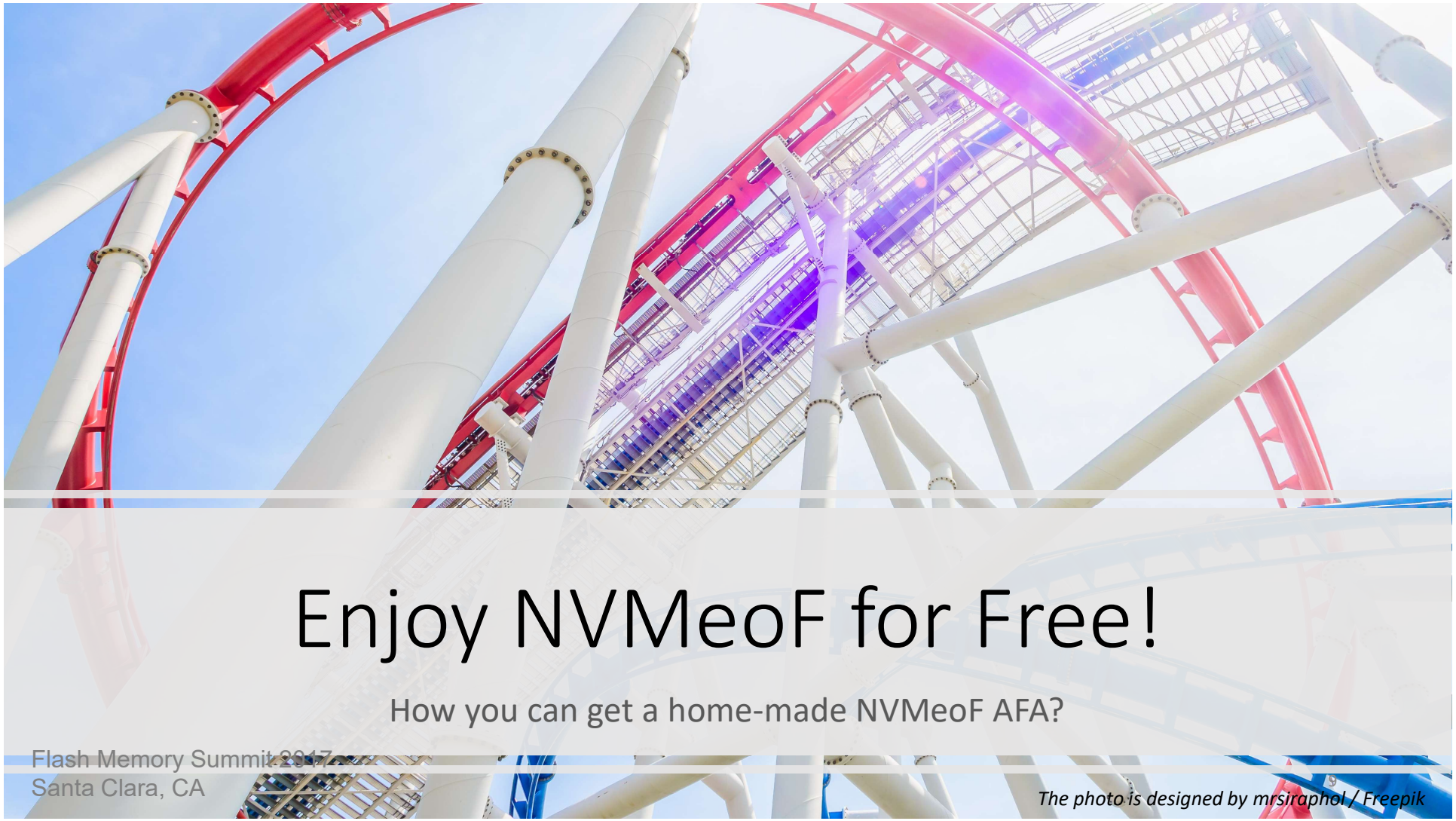
PROFESSOR Robert Kelly's kids invaded a live BBC News interview.

NVMe & NVMf: Active DMA to Move Data

Legacy & Thick Code Stack







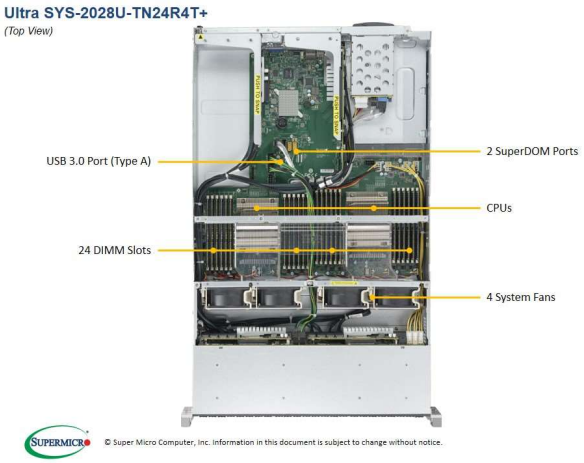
Enjoy NVMeoF for Free!

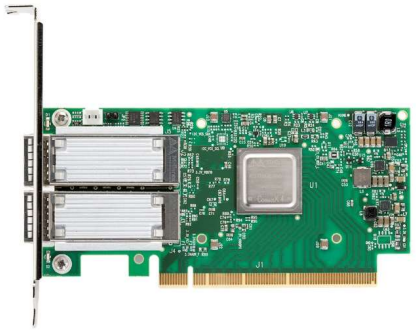
How you can get a home-made NVMeoF AFA?

Flash Memory Summit 2017
Santa Clara, CA

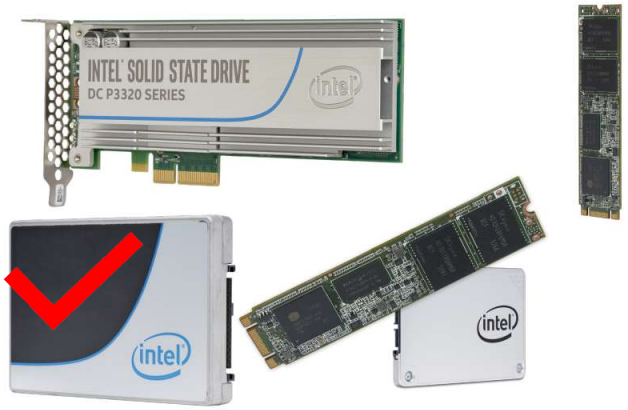
The photo is designed by mrsiraphol / Freepik

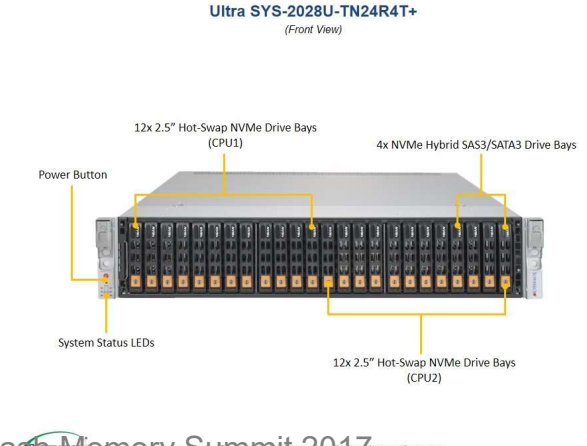

Global News Network 'AVING' by Dooyoung Shin (www.aving.net)
http://us.aving.net/news/view.php?articleId=944262
ASUS
SUPERMICRO
KGUARD
SECURITY
intel
SUPERMICRO
Commodity Hardware





RoCE NIC from Mellanox





3 Keys of HW Components for NVMeoF AFA

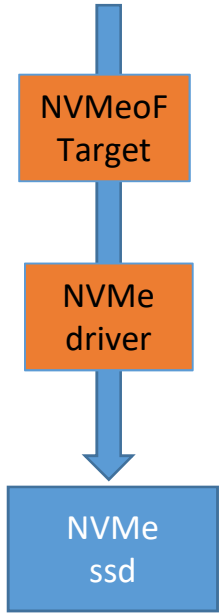
Free Software Resources

- Install Linux Centos 7.2 or above for target-side
- Install Linux Centos 7.2 or above and upgrade kernel to 4.8 for initial-side
- Download SPDK: lock-free software
- Multiple modules in SPDK:
  nvme driver, nvmeof target, iscsi target.
- User-layer application
- Replace the role of the Linux driver for nvme

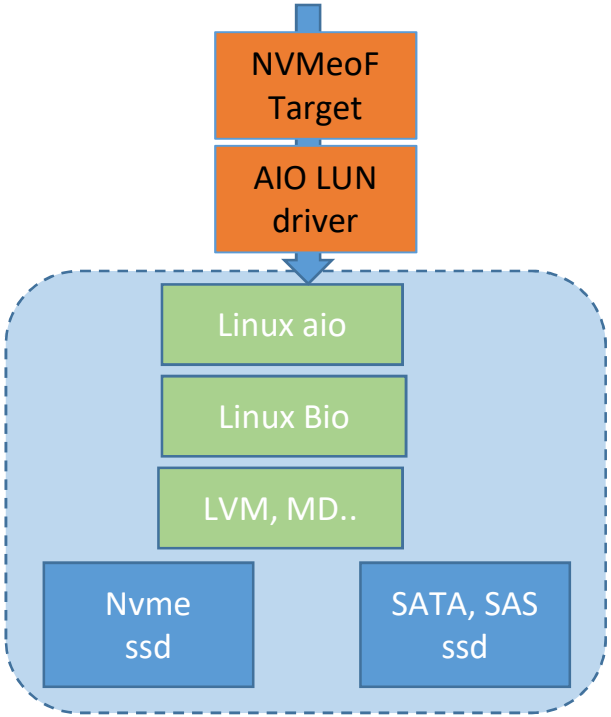
Target Modes: Passthough vs. Linux AIO

- Passthrough
- Memory
- Linux AIO

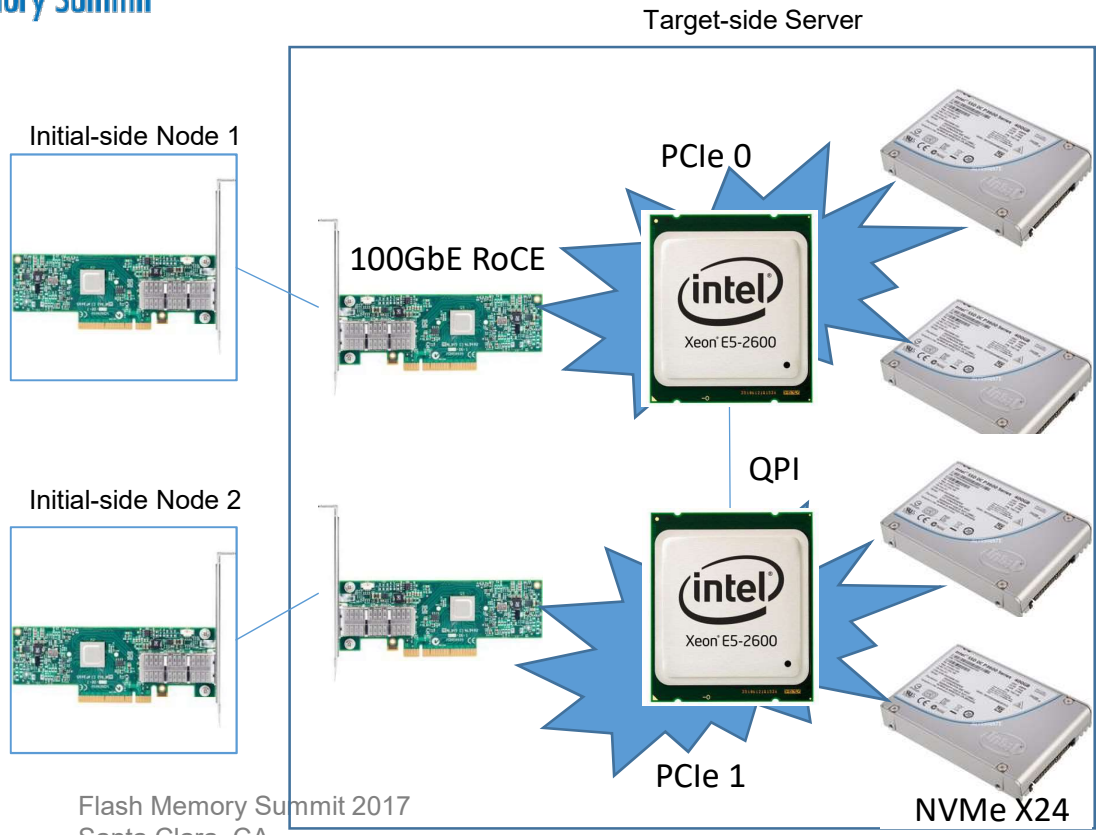
NVMe commands
to a physical nvme device



NVMe commands
to a virtual nvme device



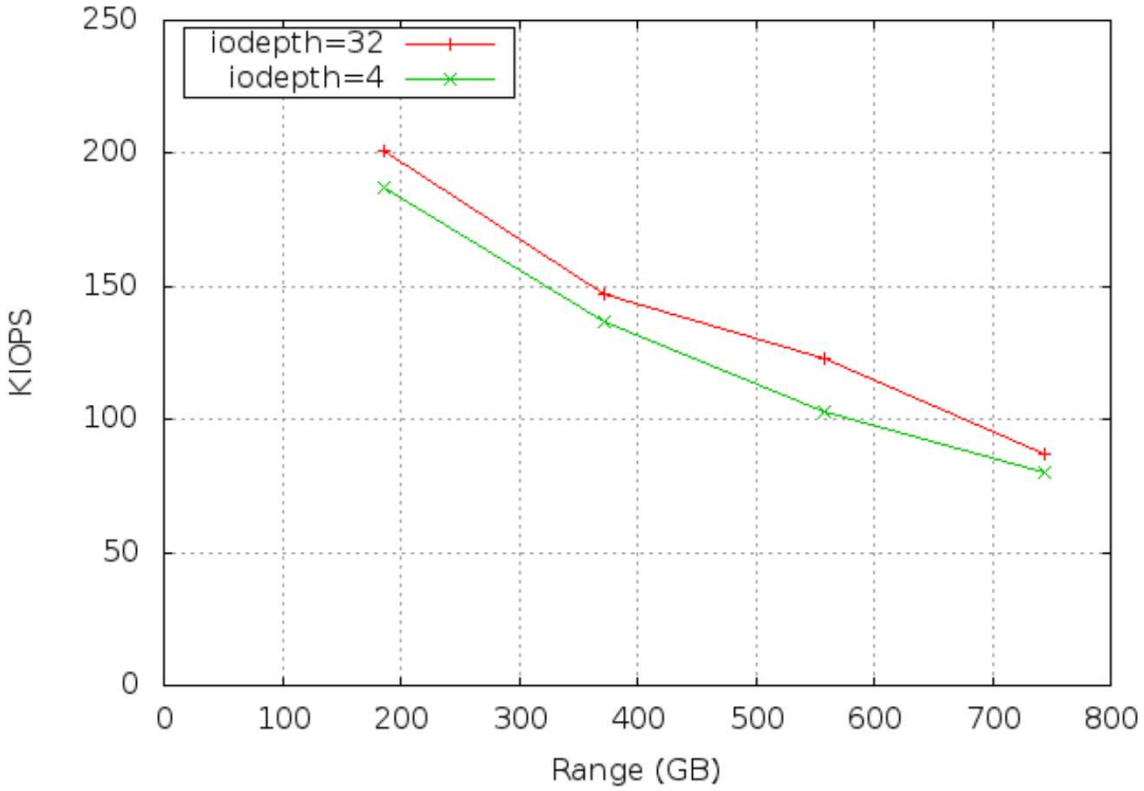
Test Topology for a NVMeoF AFA



target-side server
- *Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz*
- *Dc p3600 x 24*
- *Supermicro*
- *Mellanox connectX-4*

Initial-side node
- *Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz*
- *Supermicro*
- *Mellanox connectX-4*

Performance on Random WRITE over Different Ranges

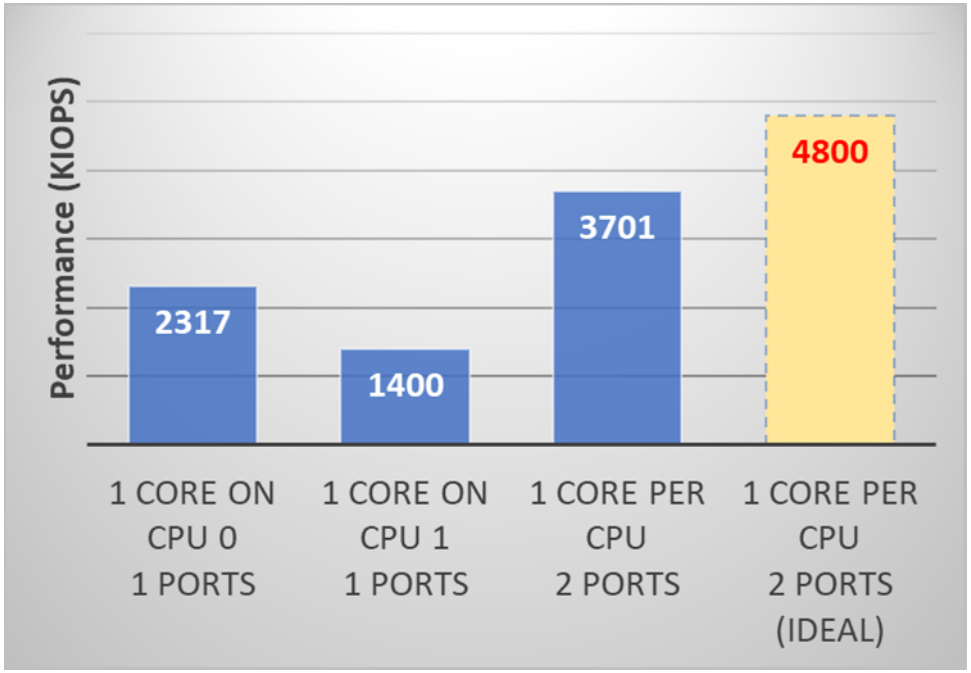


Tested NVMe SSD: Intel DC P3600, 800GB

Measured Maximum Random WRITE IOPS

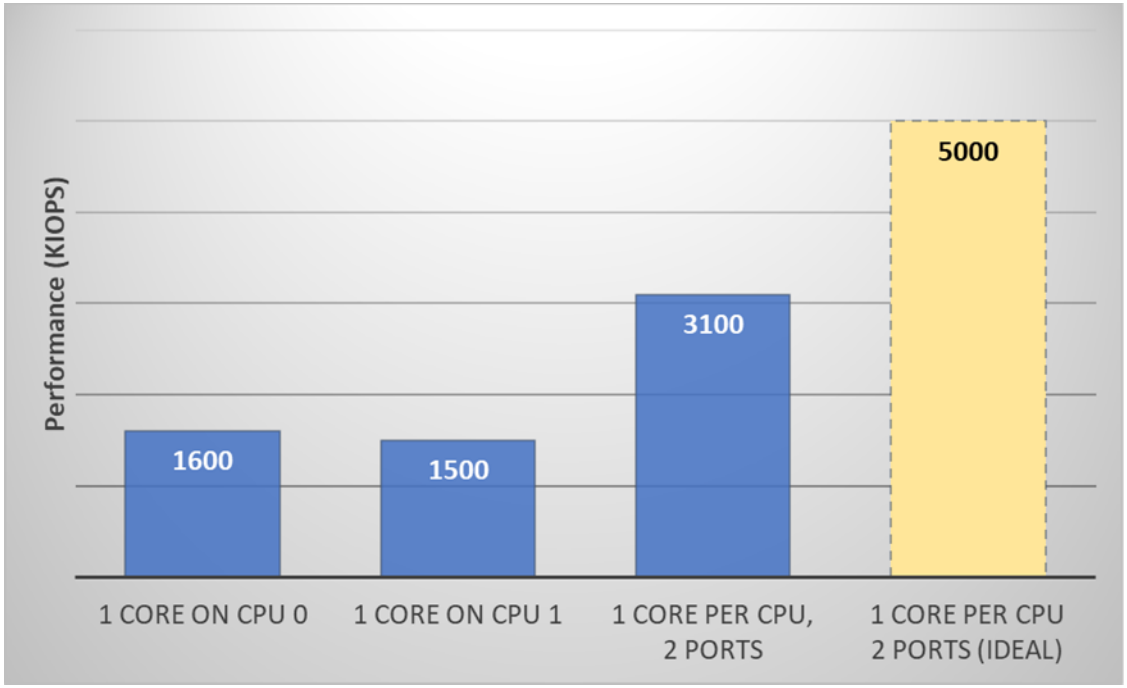
```
ReactorMask 0x0101
AcceptorCore 0
MaxQueuesPerSession 4
MaxQueueDepth 129

[Subsystem1A]
  NQN nqn.2016-06.io.spdk:cnode1
  Core 0
  Mode Direct
  Listen RDMA 50.0.51.1:4420
  NVMe 0000:06:00.0
.....
[Subsystem1B]
  NQN nqn.2016-06.io.spdk:cnode1B
  Core 8
  Mode Direct
  Listen RDMA 50.0.52.1:4420
  NVMe 0000:83:00.0
```

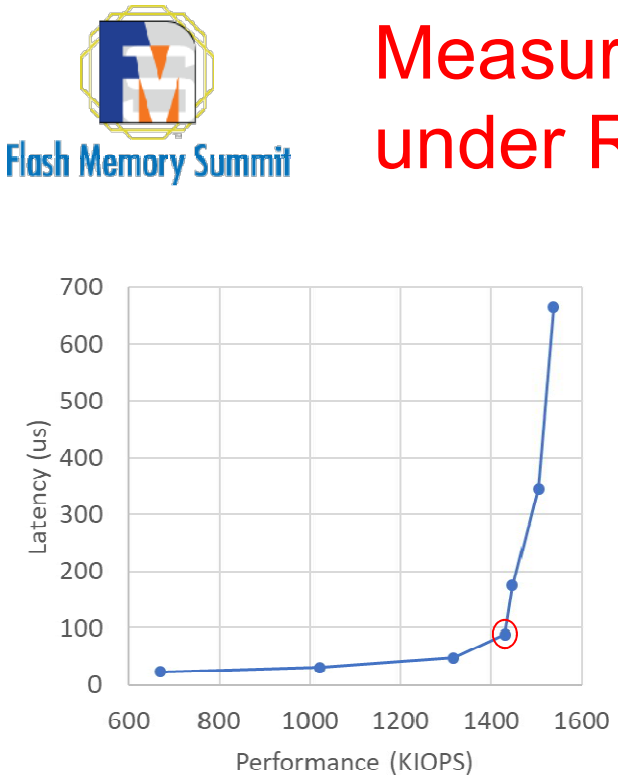


Target-side Node X1: Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz X2
24 nvme drives, 12 for CPU 0 and 12 for CPU 1
Initial-side Node X2: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz X2

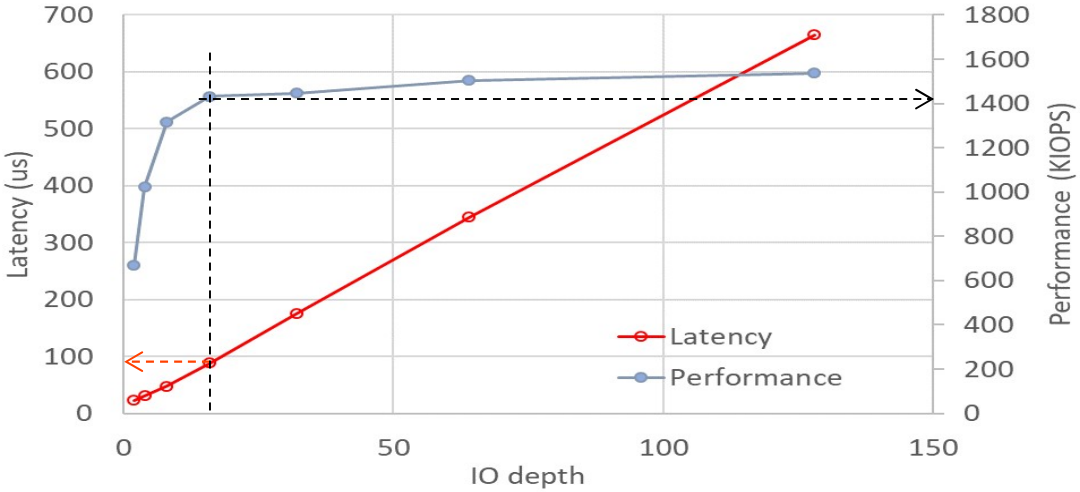
Measured Maximum Random READ IOPS



Target-side Node X1: Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz X2
24 nvme drives, 12 for CPU 0 and 12 for CPU 1
Initial-side Node X2: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz X2



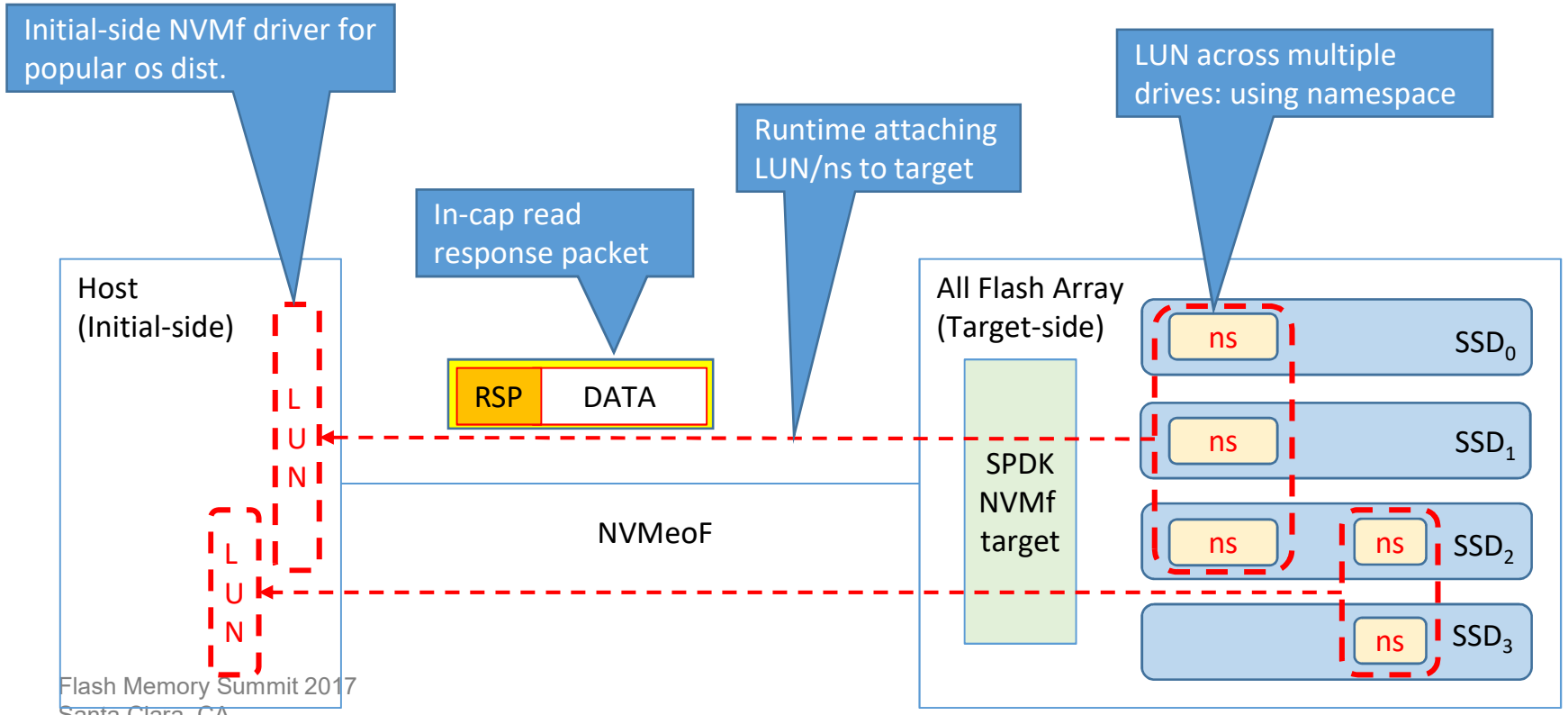
Measured Latency under Random WRITE over NVMeoF



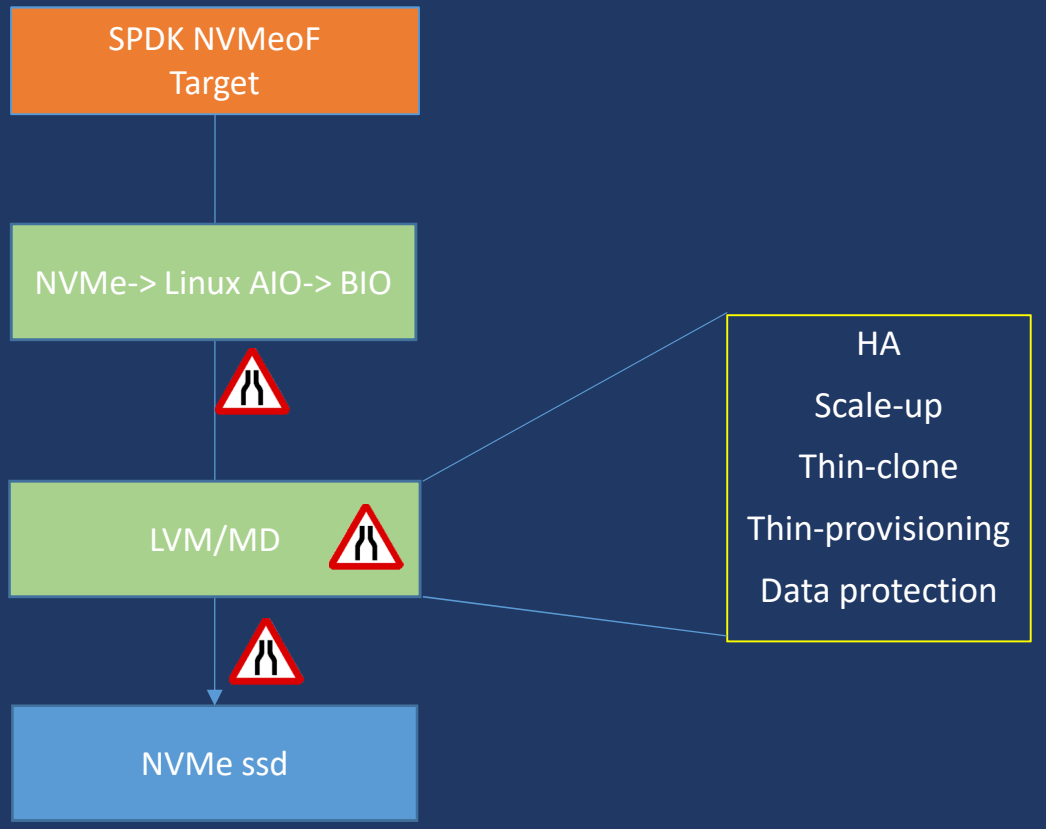
Target-side Node X1: Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz X2
8 nvme drives connecting to CPU 0
Initial-side Node X2: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz X2
Fio run on the initial host with 8 threads through nvmeof interface.
IO depth is configured for each fio thread.


Challenges and Solutions
High IOPS with Enterprise Features
Flash Memory Summit 2017
Santa Clara, CA
19
The photo is designed by jigsawstocker / Freepik

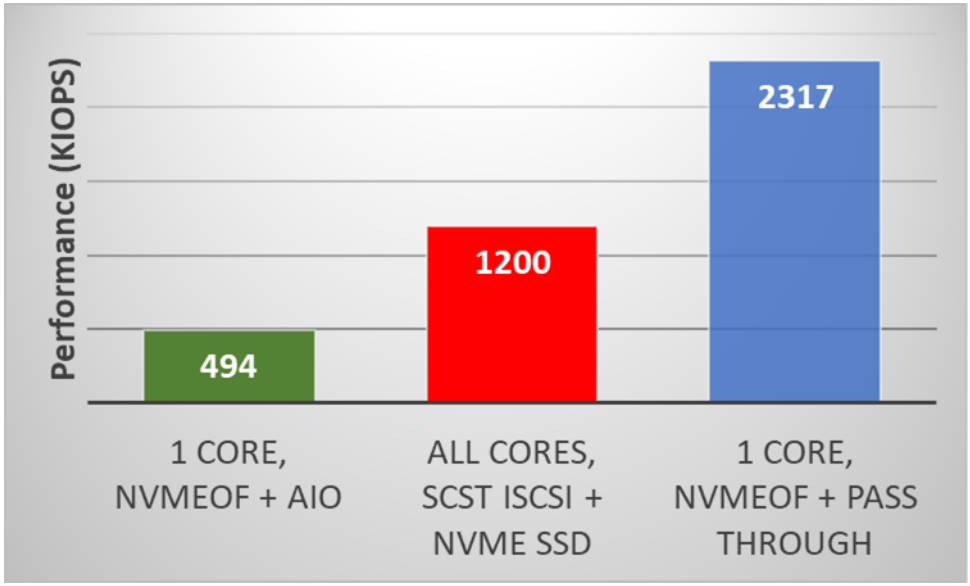
Missed Basic Features



Trade-off between Advanced Features and Performance



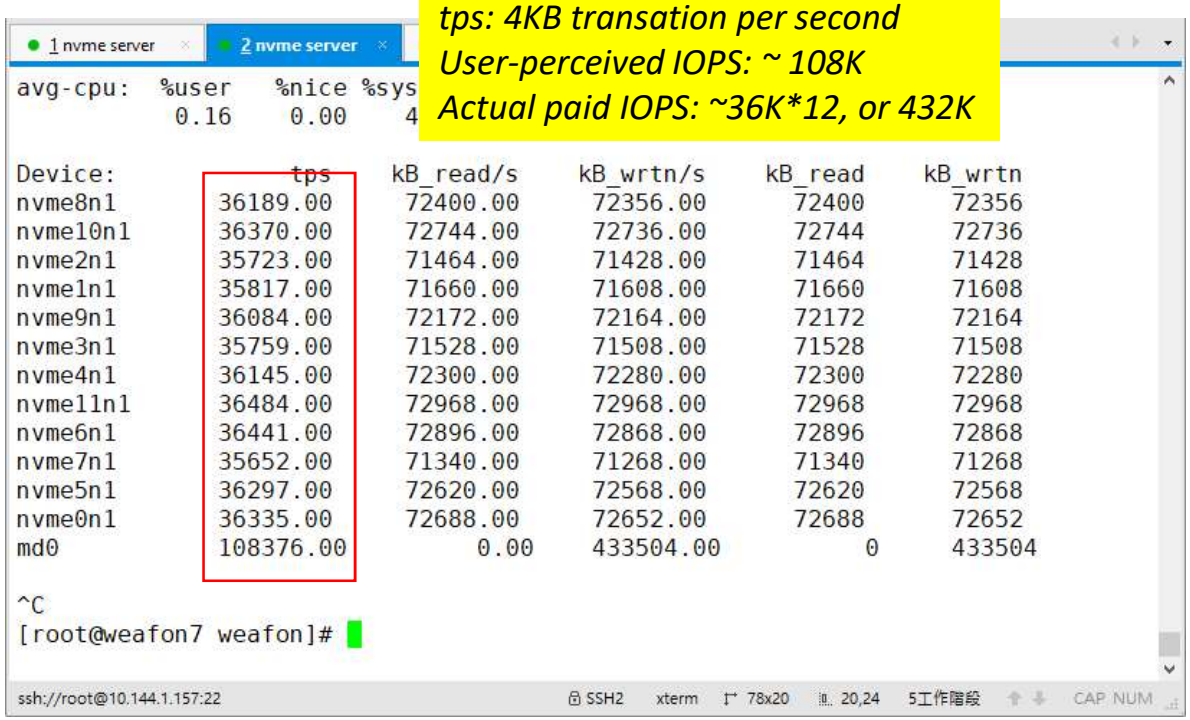
Performance on 4KB Random WRITE IOPS Under Different Target Modes



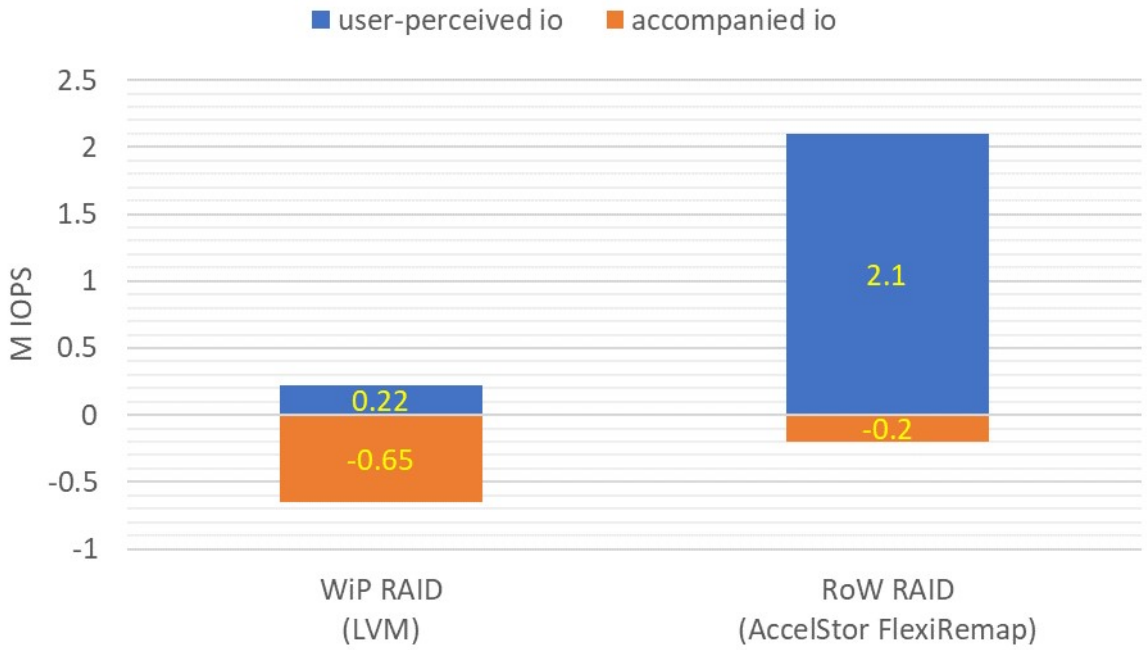
IO Amplification (AMP) – Data Protect Mechanism (DPM) on High IOPS SSD

- Write-in-place RAID -> 2R+2W at least
- Solution:
  - RAID-aware device?
  - multi-copy based?
  - Write Buffer:
    - Additional *dual* log devices
    - Additional effort on hw
    - Faster than nvme? Nvdimm
  - RoW RAID

Screenshot of Linux tool: iostat
4KB random WRITE:
tps: 4KB transation per second
User-perceived IOPS: ~ 108K
*Actual paid IOPS: ~36K*12, or 432K*



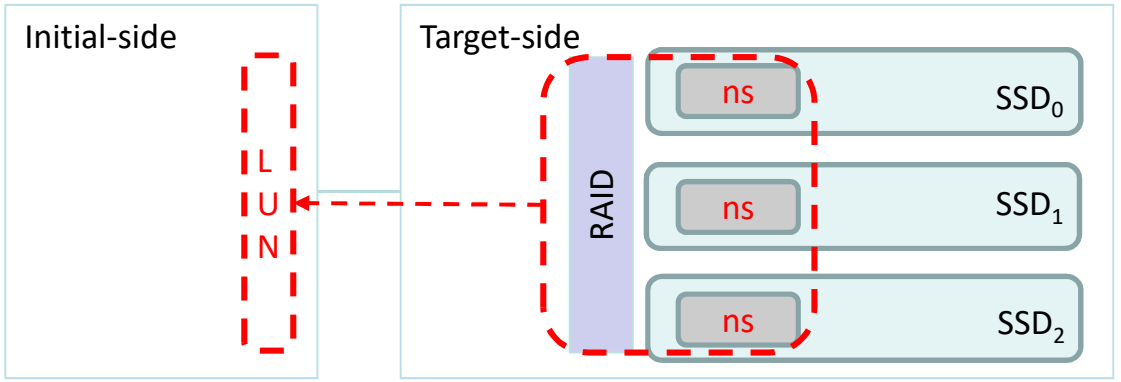
Performance under Different RAID5 over NVMeoF



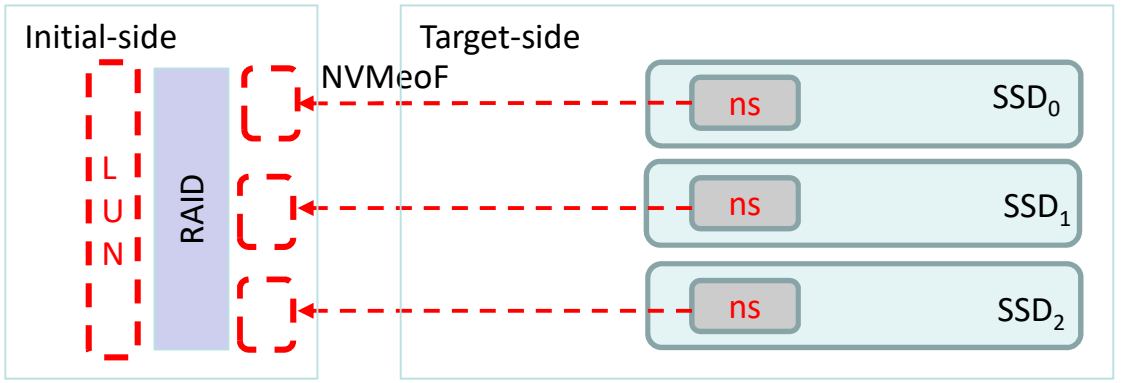
4KB Random WRITE over ¼ range of capacity, No cache before RAID.
consisting of 24 NVMe SSDs where 12 SSDs form a raid group and each SSD provides Read 400K iops, write 200K iops.
Fio run on the initial host through nvmeof interface

Where to Deploy DPM?

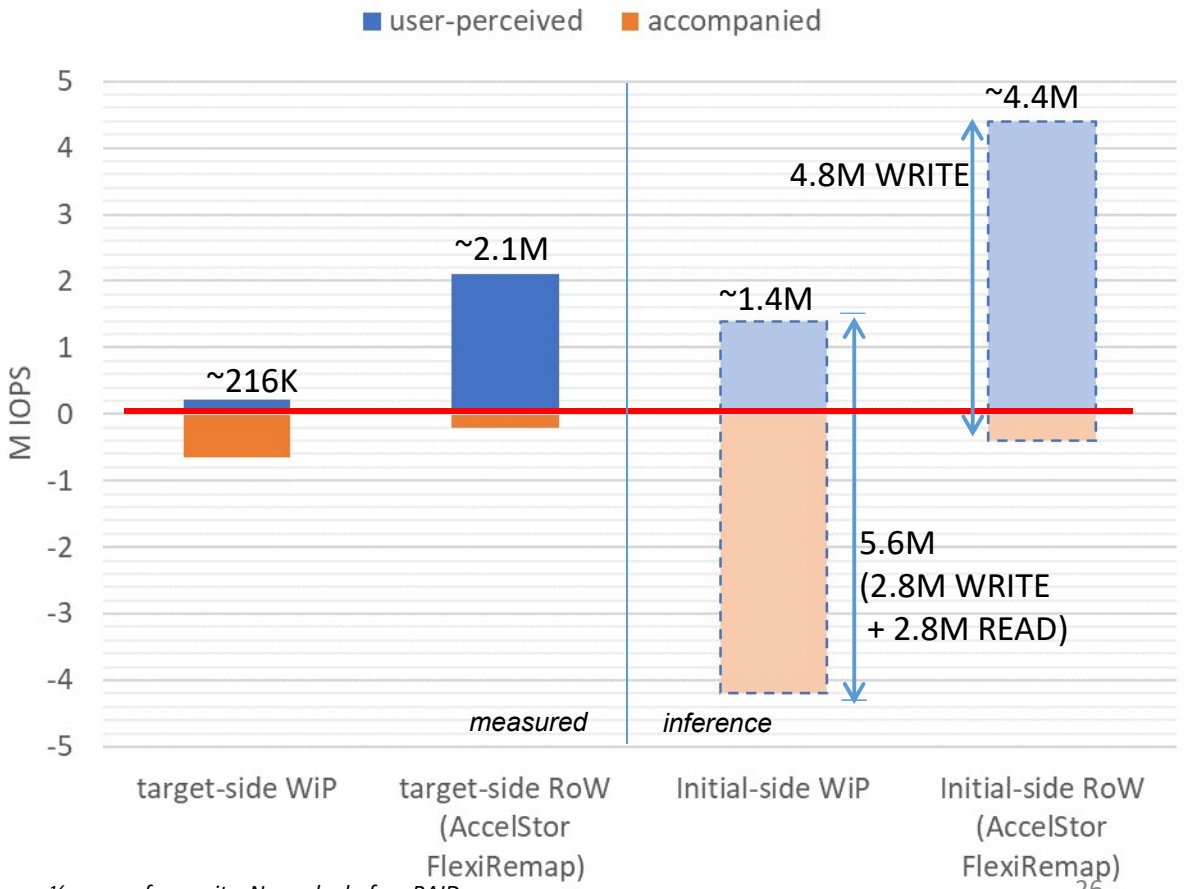
Option 1



Option 2

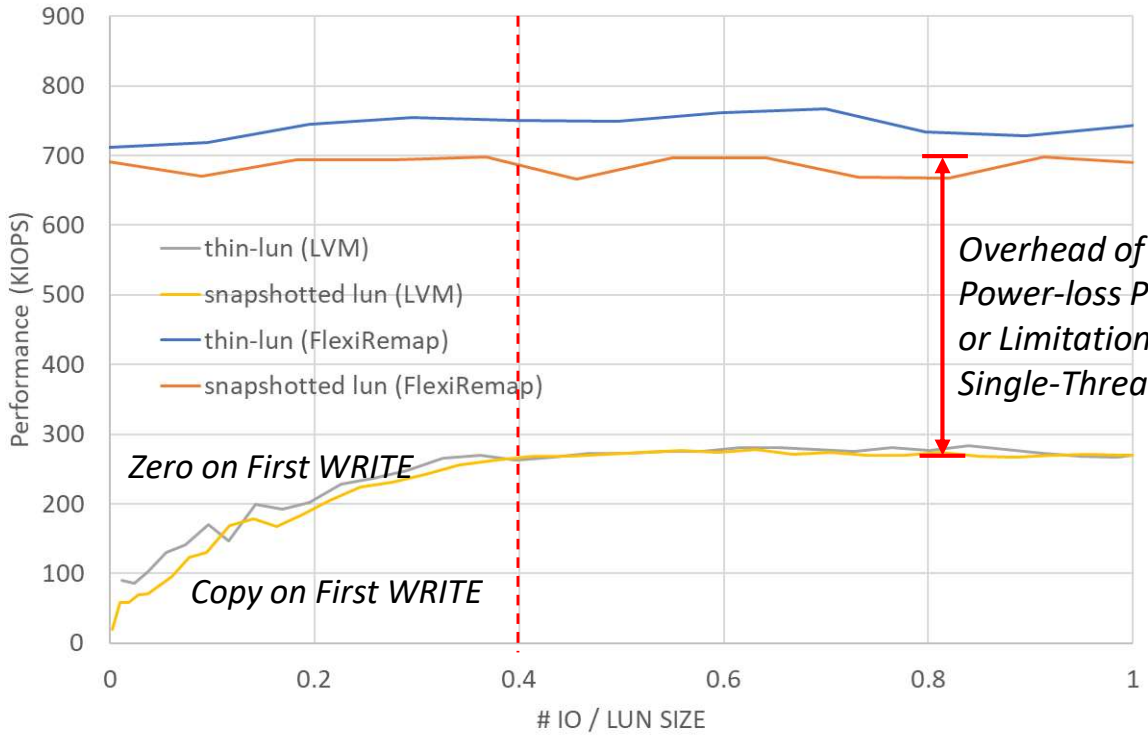


Benefit of Initial-side RAID



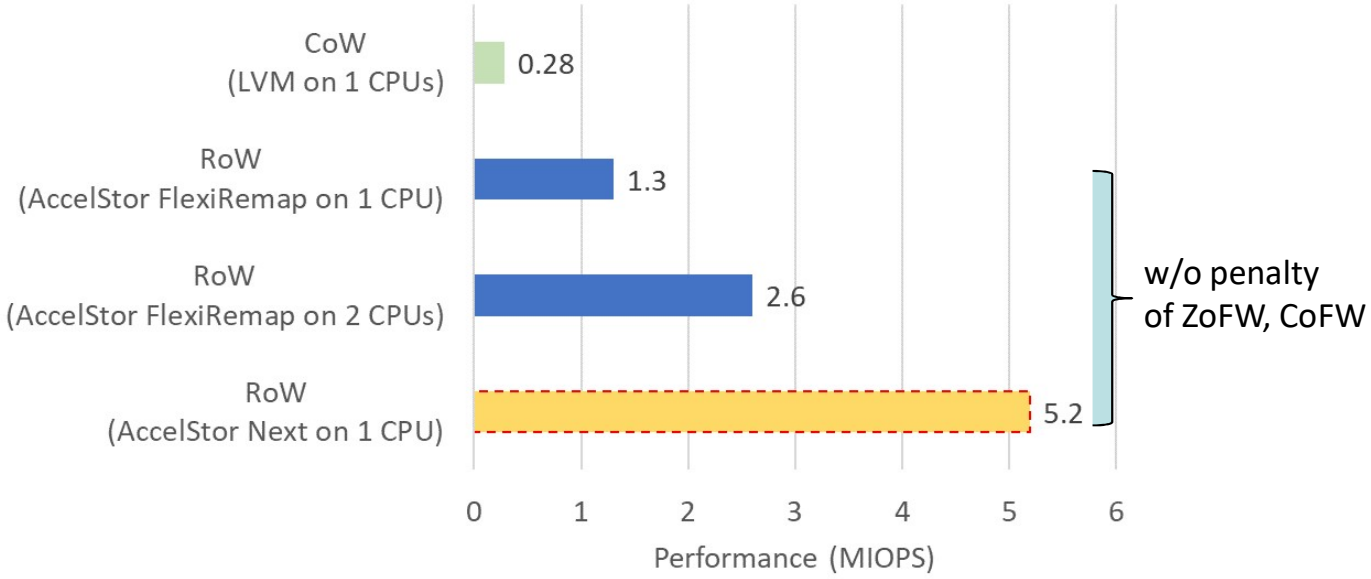
4KB Random WRITE over ¼ range of capacity, No cache before RAID.
consisting of 24 NVMe SSDs where 12 SSDs form a raid group and each SSD provide Read 400K iops, write 200K iops.
Fio run on the initial host through nvmeof interface

Issues of Small WRITE on Thin-Provisioning/Snapshot/ThinClone

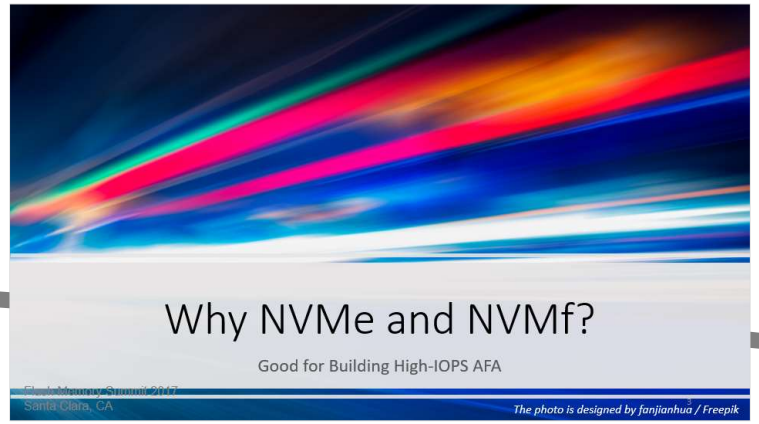


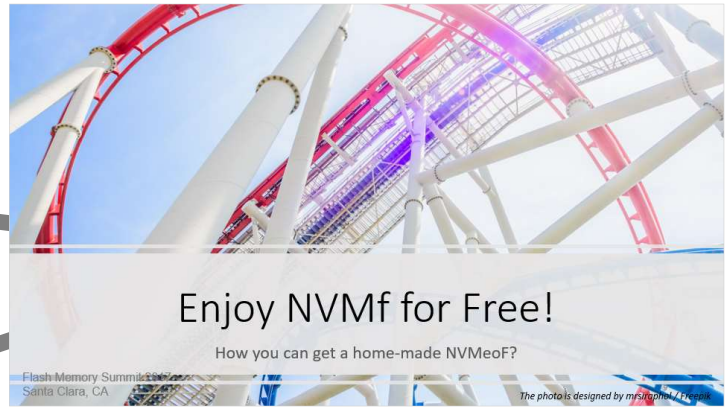
12 intel nvme DC P3600 form a RAID0 group
80GB pool for 300GB thin-lun and snapshot
Fio random 4KB, write range 30GB, 3 threads
LVM: linux lvm, CoW-style, stripe = 12, default
FlexiRemap: AccelStor FlexiRemap, RoW-style
TP: thin-provisioning, TC: thin-clone

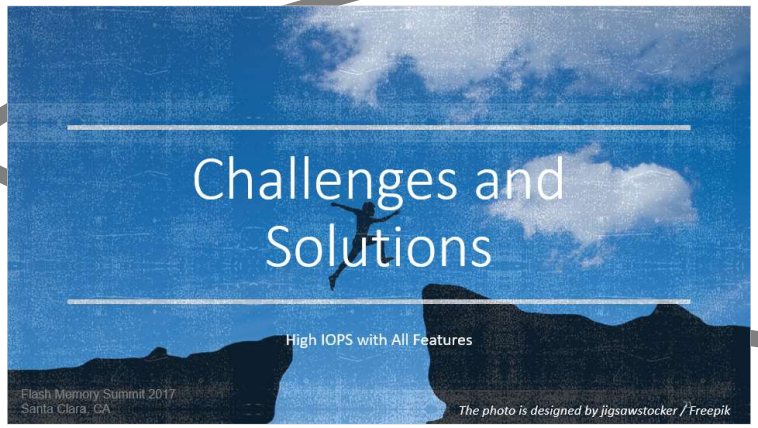
Performance on TP/Snapshot/TC over NVMeoF



AccelStor Next: A patent-pending solution
SSD Vendors are welcome to discuss with us.


Why NVMe and NVMf?
Good for Building High-IOPS AFA
Flash Memory Summit 2017
Santa Clara, CA
The photo is designed by fanjianhua / Freepik


Enjoy NVMf for Free!
How you can get a home-made NVMeoF?
Flash Memory Summit 2017
Santa Clara, CA


Challenges and Solutions
High IOPS with All Features
Flash Memory Summit 2017
Santa Clara, CA
The photo is designed by jigsawstocker / Freepik

Thank You

Weafon.tsao@accelstor.com

AccelStor, Inc.

Booth: 132