

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)

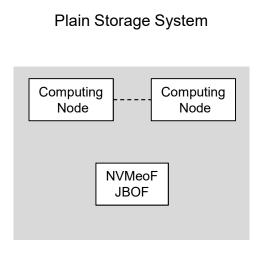
- Data Protection (DP)
- High Availability (HA)
- Snapshot
- Replication (Sync/Async)
- Thin-Provisioning
- Thin-Clone
- Deduplication
- Compression

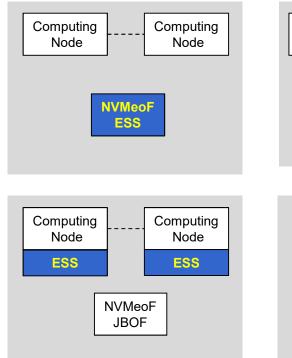
Disaster Recovery (DR)

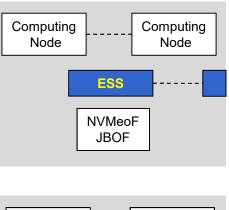
Data Reduction

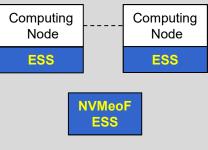


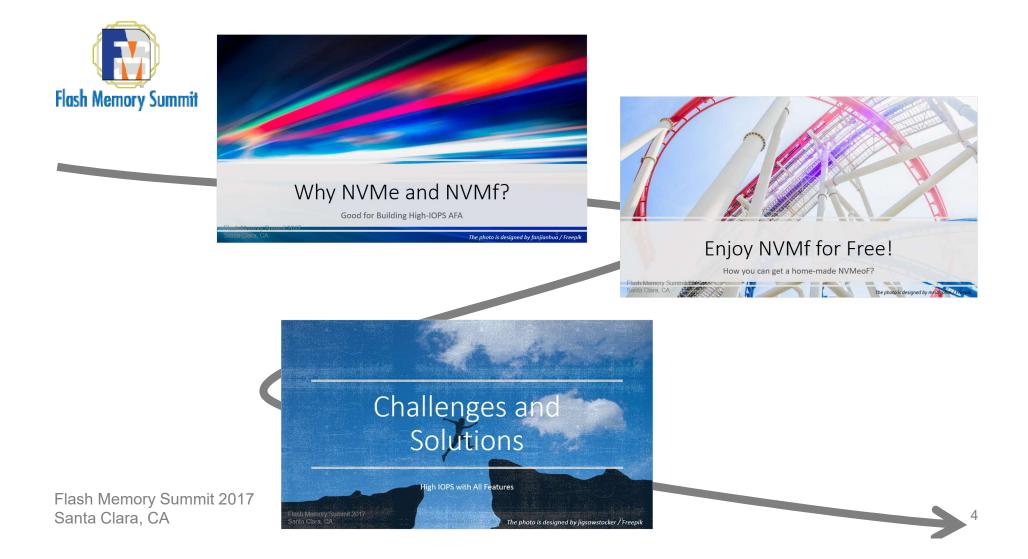
NVMeoF-based ESS













Why NVMe and NVMf?

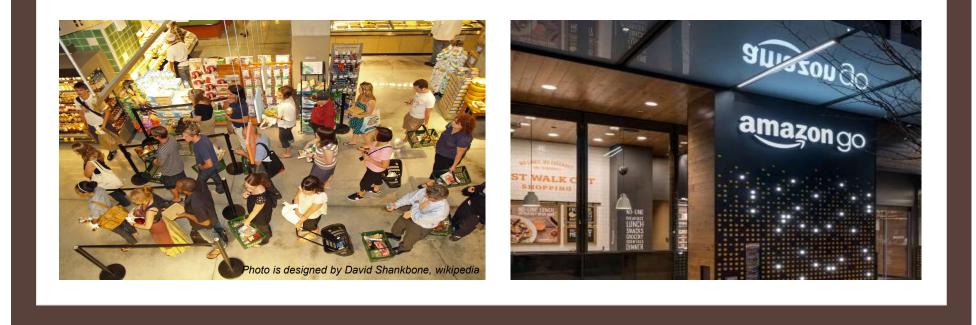
Good for Building High-IOPS AFA

Santa Clara, CA

h Memory-Summit 2017

The photo is designed by fanjianhua / Freepik





NVMe & NVMf: Active DMA to Move Data

Legacy & Thick Code Stack



ascispace BIOs BIOs BIOs Block Layer I/O schuller blkmq multi queue Maps BIOs to equests ij Software queues ked in device drivers y hook in like stacked noop es do) dline Hardware dispatch queues BIO Request based dri Request based drivers based drivers * * * Request-based device mapper targets *** SCSI mid layer sysfs (transport_attributes) /dev/rsso (scsi-mq lev/rbd* dev/mmcblk*p* dev/nullb*) (dev/vd*) /dev/skd SCSI upper leve ivers ne*n* dev/rsxx* dev/sda Transport classes ۷ ¥ ¥ -> network > memory SCSI low level drive ahci ata_piix … vmw_pvscsi > network mobile device flash memory Micron PCIe card stec חחו SSD ne MC-Siel HBA SCSI (virtio_pci) VMware's Adapte RAID mulex HBA SD-/MMC-Card Physical devices adapte **TH_MAS** The Linux Storage Stack Diagram http://www.thomas-krenn.com/en/wiki/Linux_Storage_Stack_Diagram Created by Werner Fischer and Georg Schönberger License: CC-BY-SA 3.0, see http://creativecommons.org/licenses/by-sa/3.0/ **KRENN**

Santa Clara, CA

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





Ultra SYS-2028U-TN24R4T+ (Front View)

12x 2.5" Hot-Swap NVMe Drive Bays (CPU1) 4x NVMe Hybrid SAS3/SATA3 Drive Bays Power Button System Status LEDs (22.5" Hot-Swap NVMe Drive Bays (CPU2)

Flashiemory Summit 20477 revenue voor veel Santa Clara, CA





RoCE NIC from Mellanox

3 Keys of HW Components for NVMeoF AFA



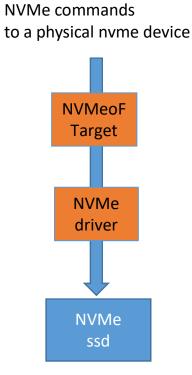
Free Software Resources

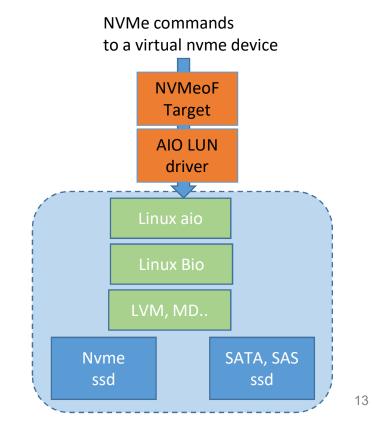
- Install Linux Centos 7.2 or above for target-side
- Install Linux Centos 7.2 or above and upgrade kernel to 4.8 for initial-side
- Download SPDK: lock-free software
- Multiple modules in SPDK: nvme driver, nvmeof target, iscsi target.
- User-layer application
- Replace the role of the Linux driver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



Target Modes: Passthough vs. Linux AIO

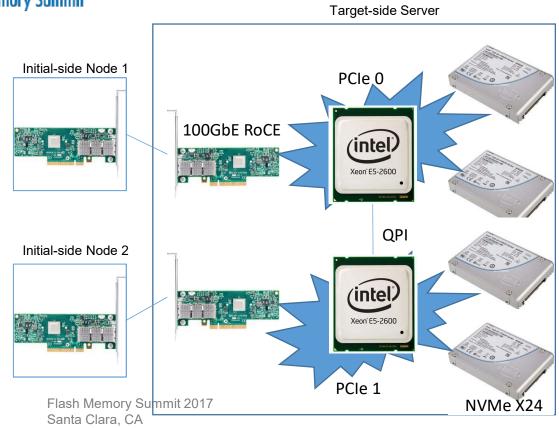
- Passthrough
- Memory
- Linux AlO







Test Topology for a NVMeoF AFA

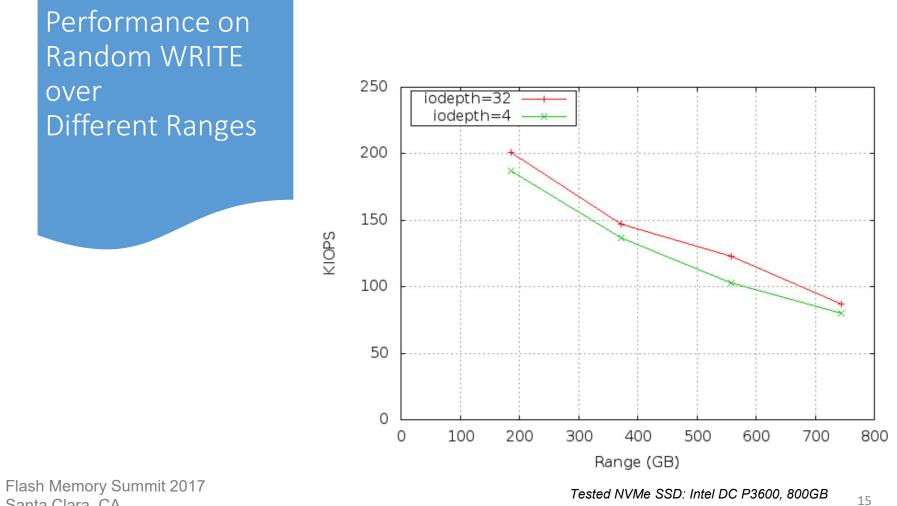


target-side server

- 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



Santa Clara, CA



Measured Maximum Random WRITE IOPS

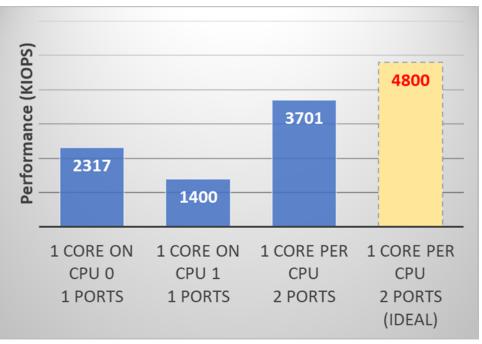
ReactorMask 0x0101 AcceptorCore 0 MaxQueuesPerSession 4 MaxQueueDepth 129

.....

[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

Flash Memory Summit 2017 Santa Clara, CA

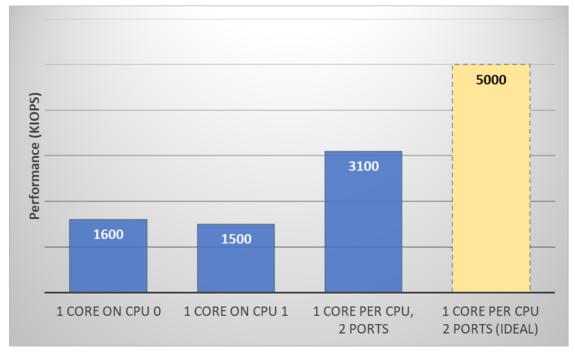


Target-side Node X1: Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz X2 24 nvme drives, 12 for CPU 0 and 12 for CPU 1 Initial-side Node X2: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz X2

16



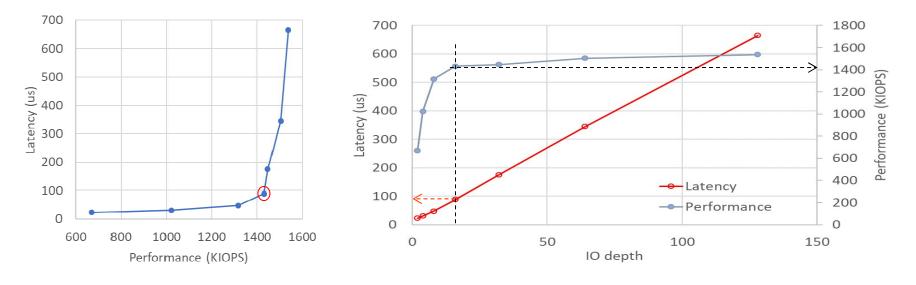
Measured Maximum Random READ IOPS



Flash Memory Summit 2017 Santa Clara, CA 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



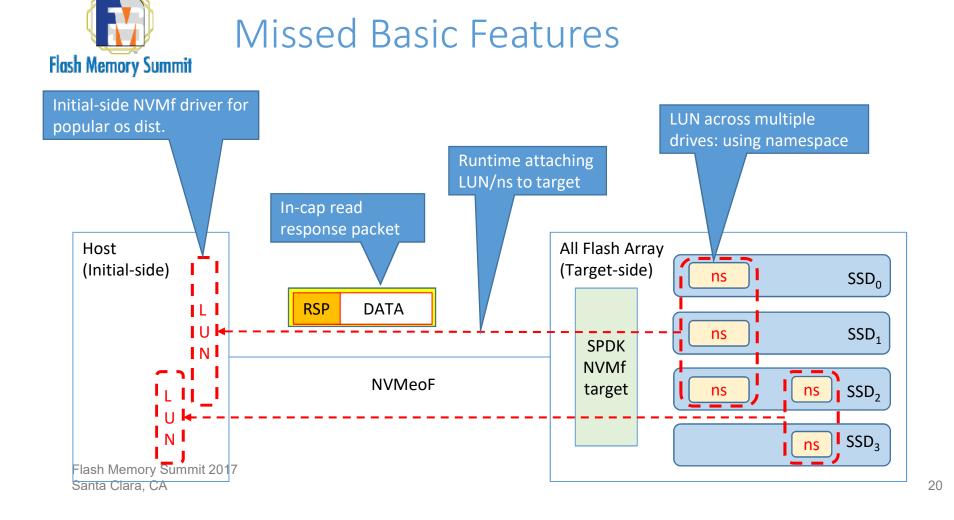
Target-side Node X1: Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20GHz X2 8 nvme drives connecting to CPU 0 Initial-side Node X2: Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz X2 Fio run on the initial host with 8 threads through nvmeof interface. IO depth is configured for each fio thread.

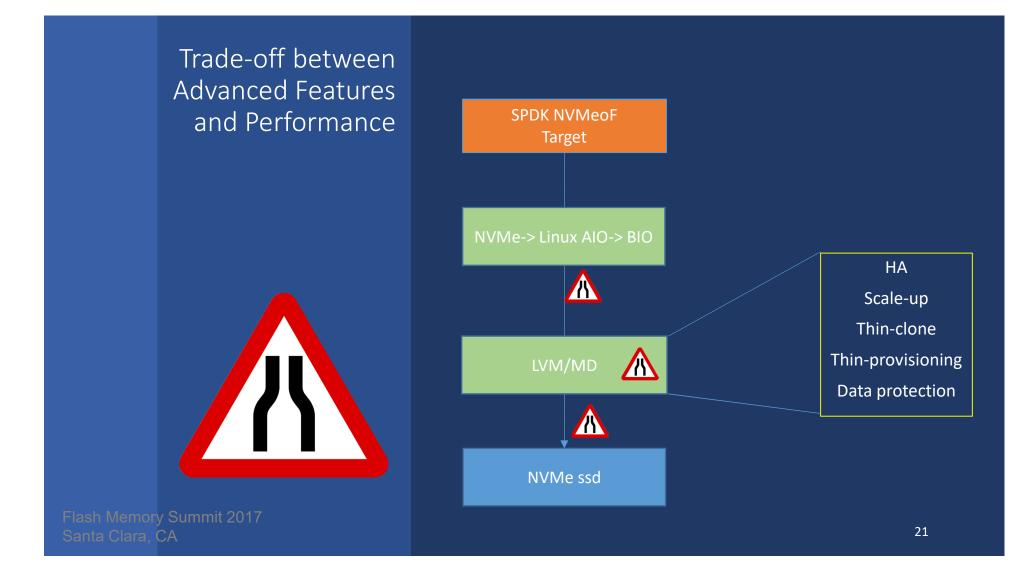
Challenges and Solutions

High IOPS with Enterprise Features

Flash Memory Summit 2017 Santa Clara, CA

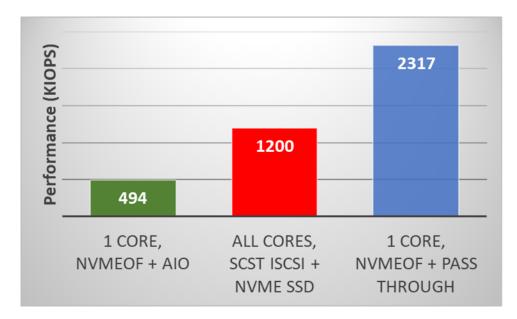
The photo is designed by jigsawstocker / Freepik







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

Flash Memory Summit 2017 Santa Clara, CA

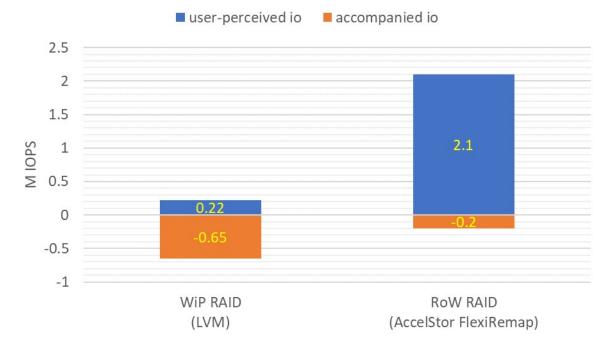
• <u>1</u> nvme server	× <u>2</u> nvme server	100	B transation prceived IOPS			÷Ε
	user %nice ⁹ 0.16 0.00	ksvs	oaid IOPS: ~3		432K	
Device:	tps	kB_read/s	kB_wrtn/s	kB_read	kB_wrtn	
nvme8n1	36189.00	72400.00	72356.00	72400	72356	
nvme10n1	36370.00	72744.00	72736.00	72744	72736	
nvme2n1	35723.00	71464.00	71428.00	71464	71428	
nvmeln1	35817.00	71660.00	71608.00	71660	71608	
nvme9n1	36084.00	72172.00	72164.00	72172	72164	
nvme3n1	35759.00	71528.00	71508.00	71528	71508	
nvme4n1	36145.00	72300.00	72280.00	72300	72280	
nvmelln1	36484.00	72968.00	72968.00	72968	72968	
nvme6n1	36441.00	72896.00	72868.00	72896	72868	
nvme7n1	35652.00	71340.00	71268.00	71340	71268	
nvme5n1	36297.00	72620.00	72568.00	72620	72568	
nvme0n1	36335.00	72688.00	72652.00	72688	72652	
md0	108376.00	0.00	433504.00	0	433504	
^C						
[root@weafo	n7 weafon]#					
ssh://root@10.144.1.1	157:22		ⓓ SSH2 xterm I ⁺	78x20 in 20,24	5工作階段 1 +	CAP NUM

AKB random W/RITE

Screenshot of Linux tool: iostat



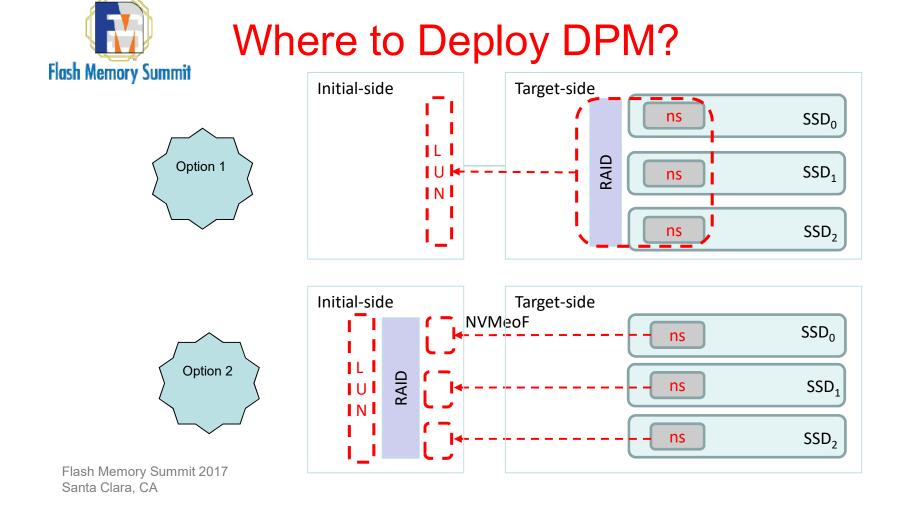
Performance under Different RAID5 over NVMeoF

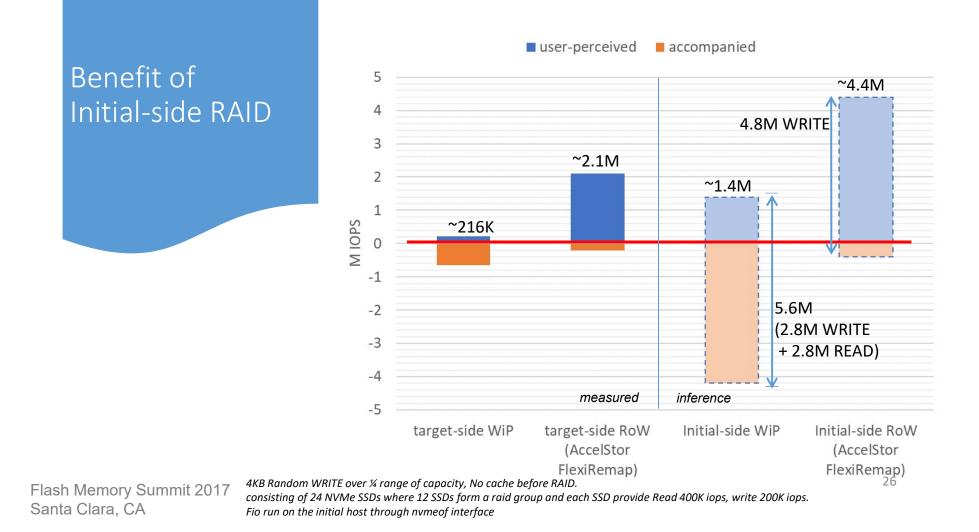


4KB Random WRITE over ¼ range of capacity, No cache before RAID.

Flash Memory Summit 2017 Santa Clara, CA

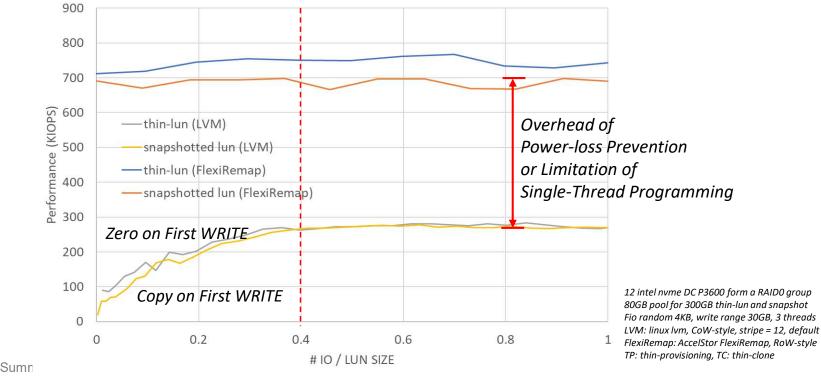
consisting of 24 NVMe SSDs where 12 SSDs form a raid group and each SSD provides Read 400K iops, write 200K iops. Fio run on the initial host through nvmeof interface 24







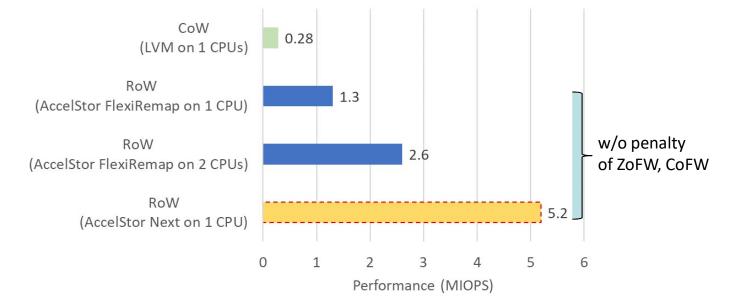
Issues of Small WRITE on Thin-Provisioning/Snapshot/ThinClone



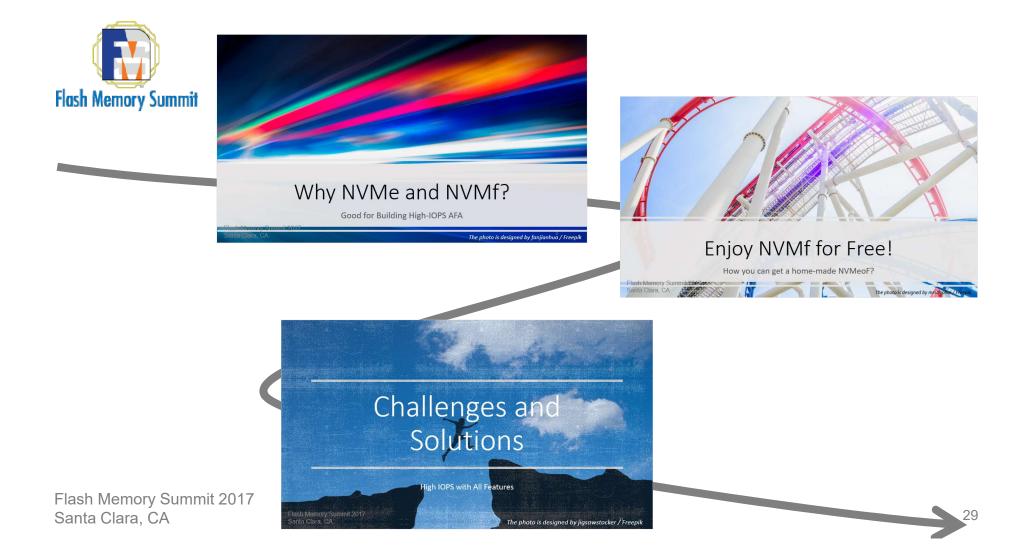
Flash Memory Sumn Santa Clara, CA



Performance on TP/Snapshot/TC over NVMeoF



Flash Memory Summit 2017 Santa Clara, CA AccelStor Next: A patent-pending solution SSD Vendors are welcome to discuss with us.





Thank You

Weafon.tsao@accelstor.com

AccelStor, Inc.

Booth: 132