Analytics @ Rack-Scale with NVMe-oF

Walter Hinton

Pavilion

Flash Memory Summit

Agenda

- About Pavilion
- The Hype Cycle
- Customer A - DAS Aggravation
- Customer B - Workload Unification
- Customer C – Scaling IBM Spectrum Scale™
- Looking Forward

About Pavilion

Flash Memory Summit

The Team:

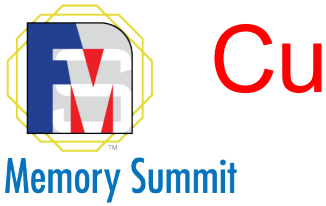
PURE STORAGE, VERITAS, HGST a Western Digital brand, AMD, CISCO, FUSION-IO, Violin MEMORY, VIRIDENT

Our Investors:

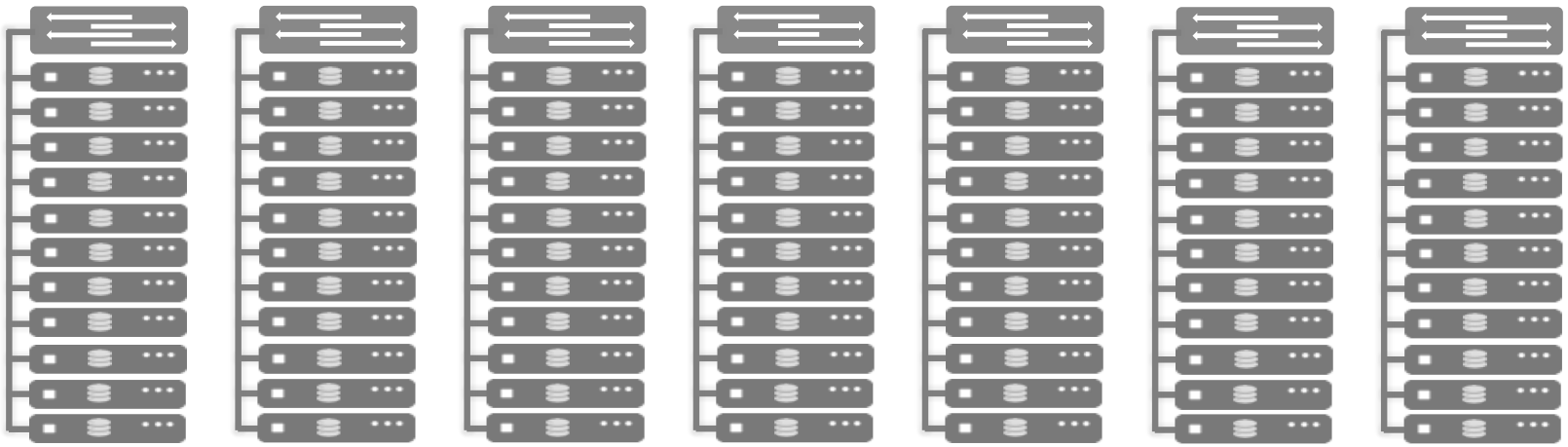
KPCB, ARTIMAN, SK telecom

The World's First Hyperparallel Flash Array

The Gartner Hype Cycle



Customer A – DAS Aggravation

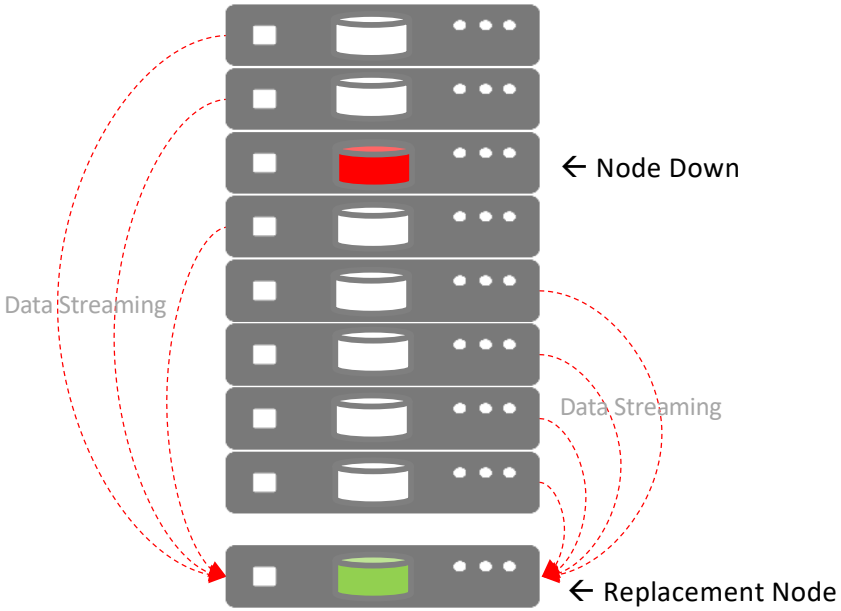


12 hour batch data collection, analytics for ad placement and targeting
72 servers with dual Xeon and large RAM with 1@ 3.2TB NVMe SSDs = 230TB
Server horsepower is adequate, but data collection is doubling every 6 months, requirement to scale to 1PB+
Node recovery for failed drive impacting network traffic for 1 hour or more

Customer A – DAS Aggravation

Unfortunately, SSDs fail!

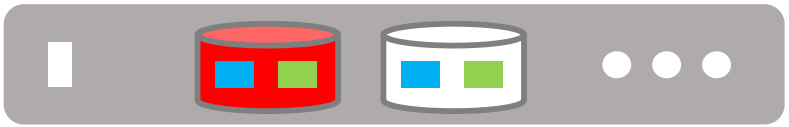
- Node recovery is slow
  - >25 minutes + per TB
  - Worse as SSDs get bigger
  - Worse with more shards
  - Impacts cluster performance



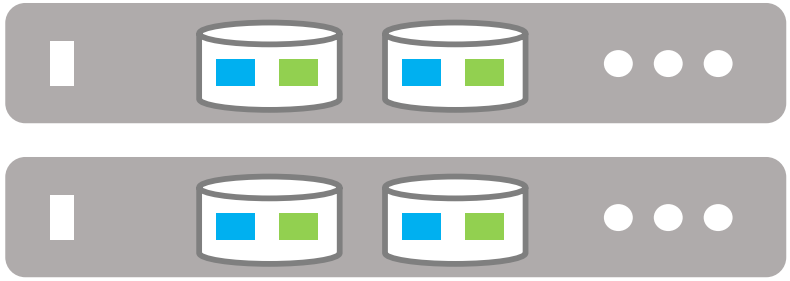
Customer A – DAS Aggravation

So…

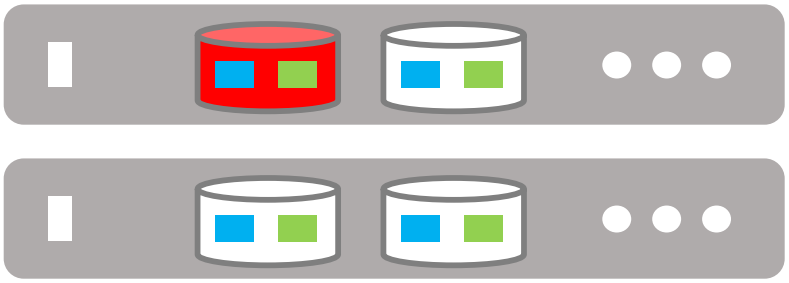
But…



Or…



But…

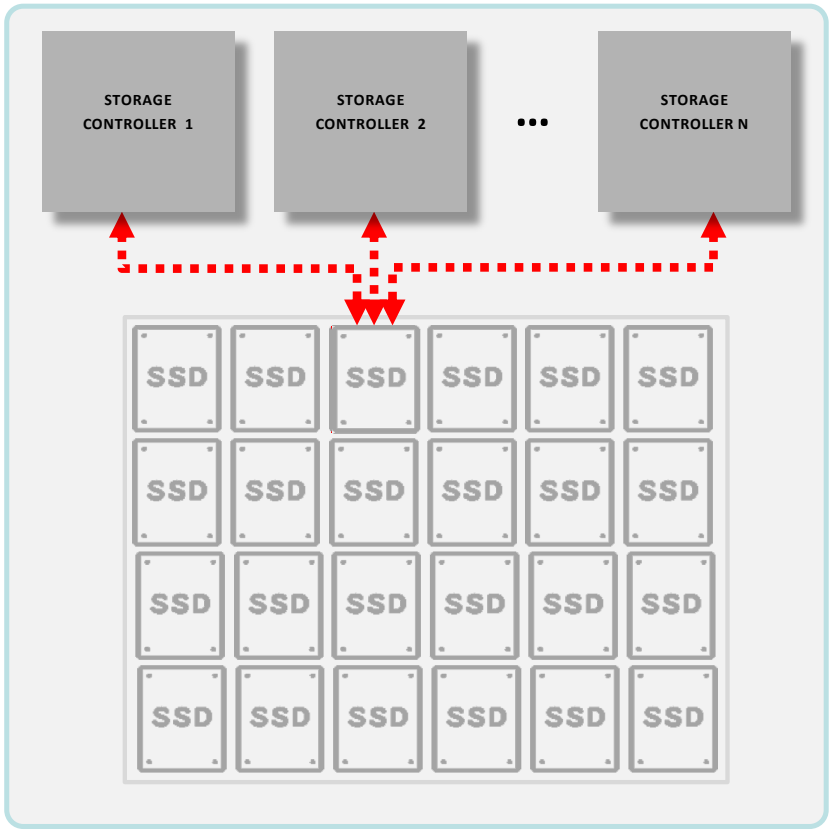


Why double the cost of your most expensive non-volatile storage?

Flash Memory Summit

~~DAS Aggravation~~ Disaggregation

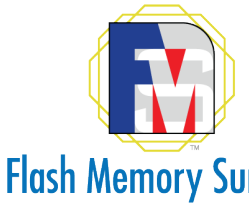
RAID-6 (12% Overhead)



Swarm Recovery (1TB < 5 min)

~~DAS Aggravation~~ Disaggregation

Real-time data collection, analytics for ad placement and targeting
2 NVMe-oF Hyperparallel Flash Arrays
72 @ 15.36TB NVMe SSDs – 1.1PB usable
250GB Read, 250GB Write @ ~ 1ms latency
1+ PB room to expand with options for various DWD choices



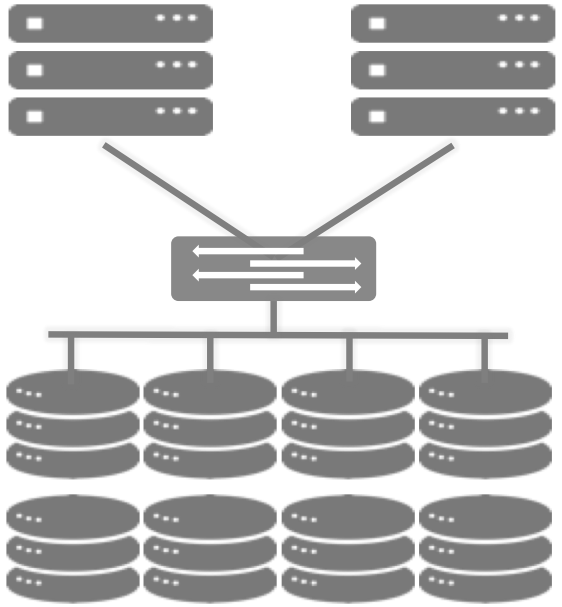
Customer B – Workload Unification

Large & growing population analytics

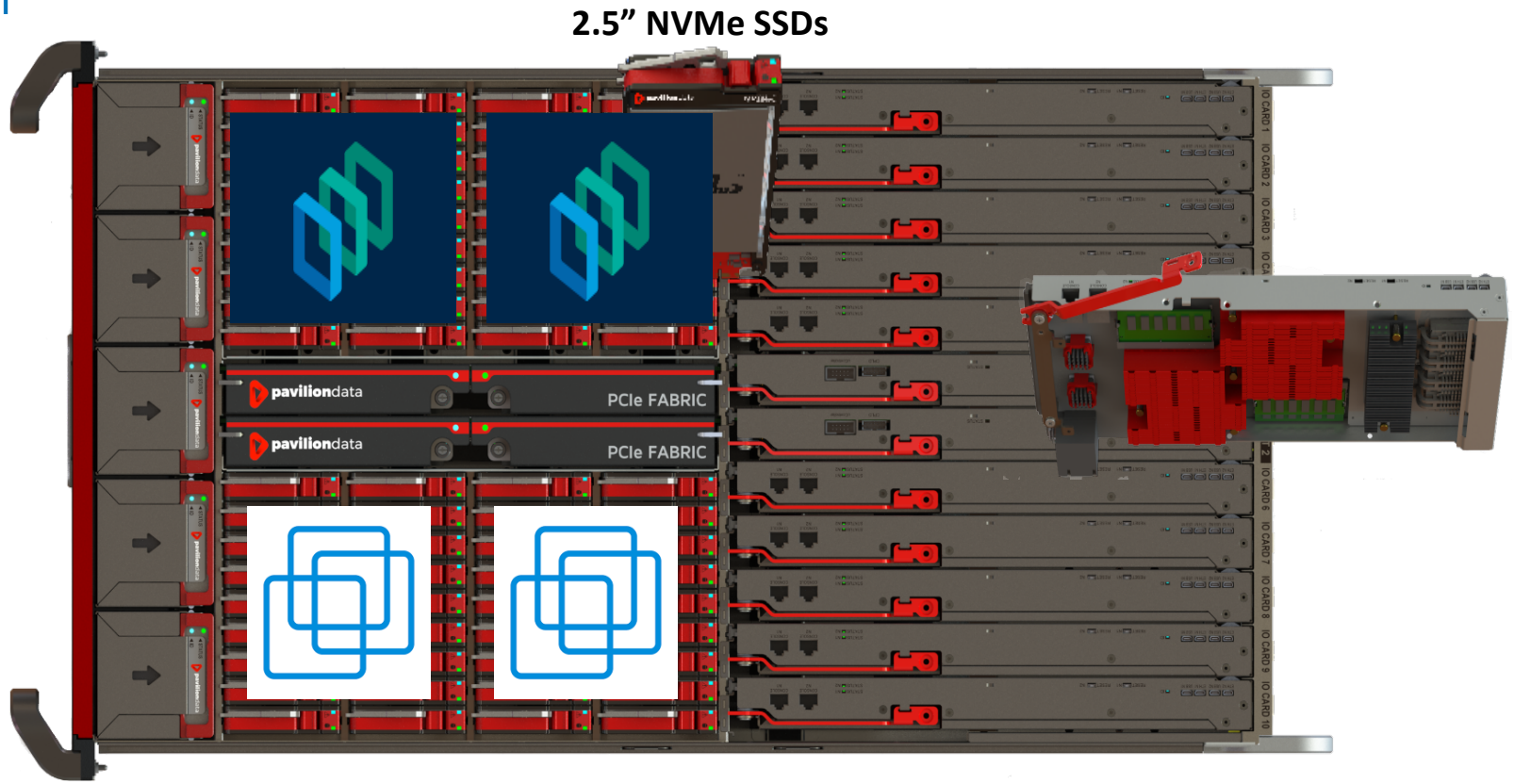
Numerous web portals too slow

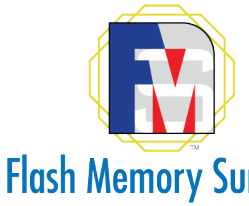
Requirements:

- VMware moving to persistent K8s
- Move from Hybrid SAN to NVMe
- 10M IOPS, 100GB/sec Reads
- Application-based replication
- Limited floorspace



Customer B – Workload Unification

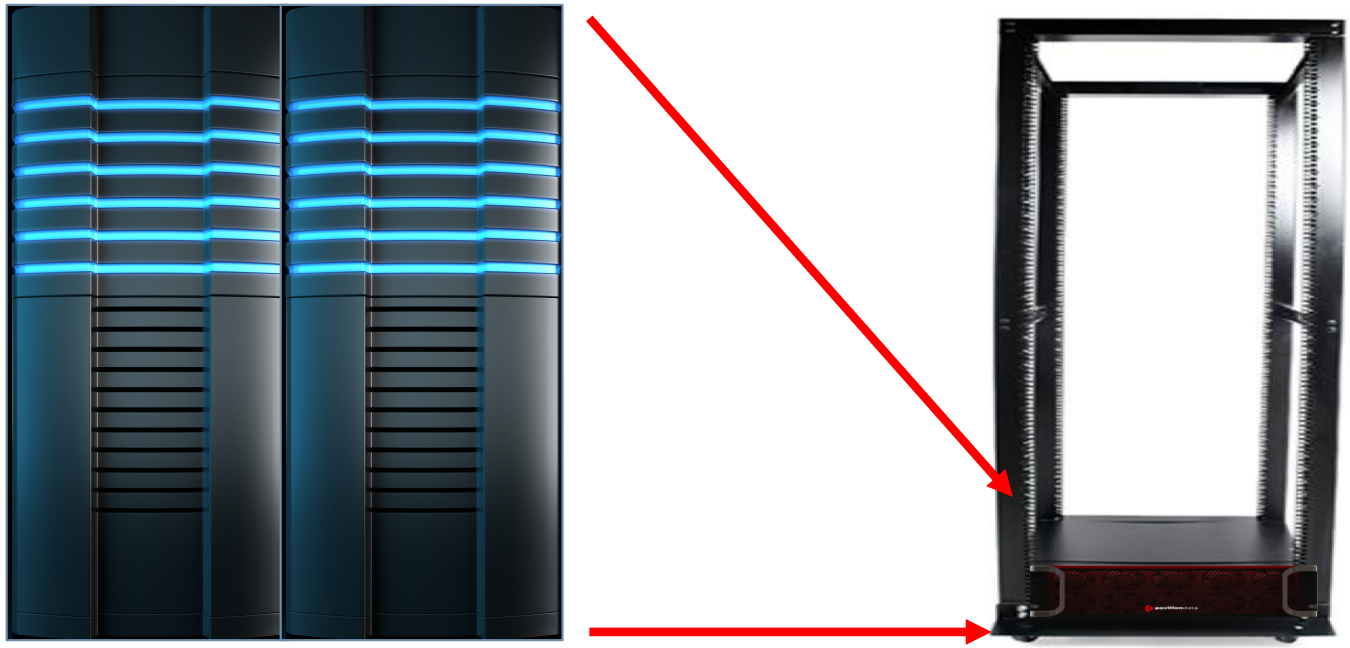




Customer 2 – Workload Unification

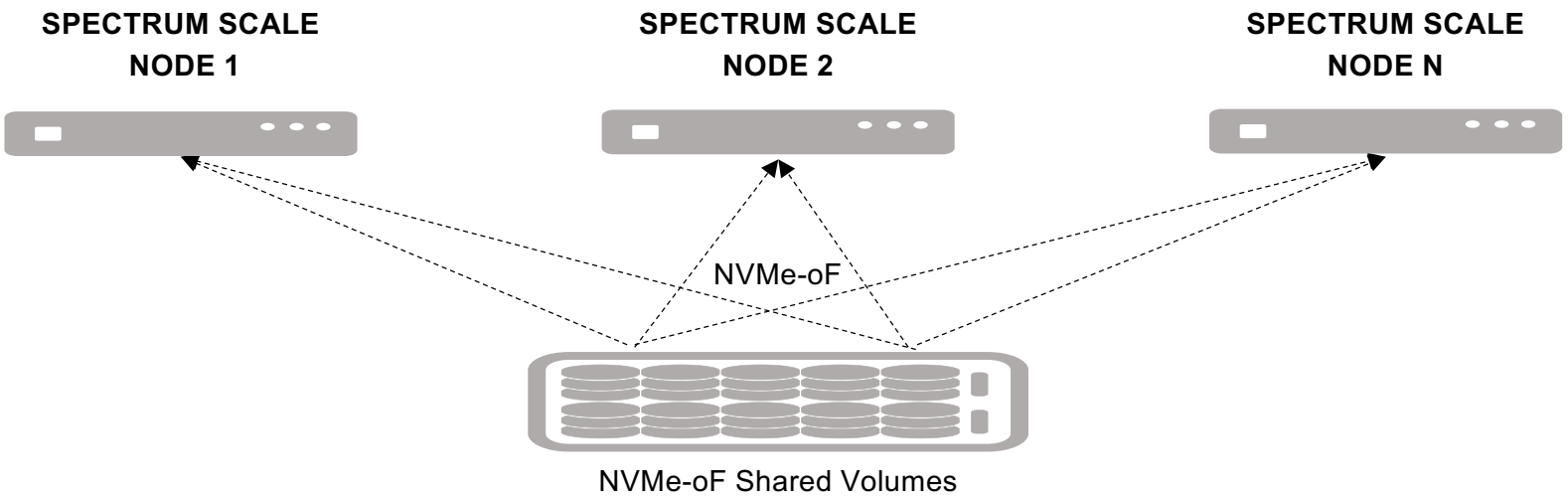
100GB/sec	80 RU	16M IOPs

120GB/sec	4 RU	20M IOPs



Alternatives didn’t meet performance density requirements

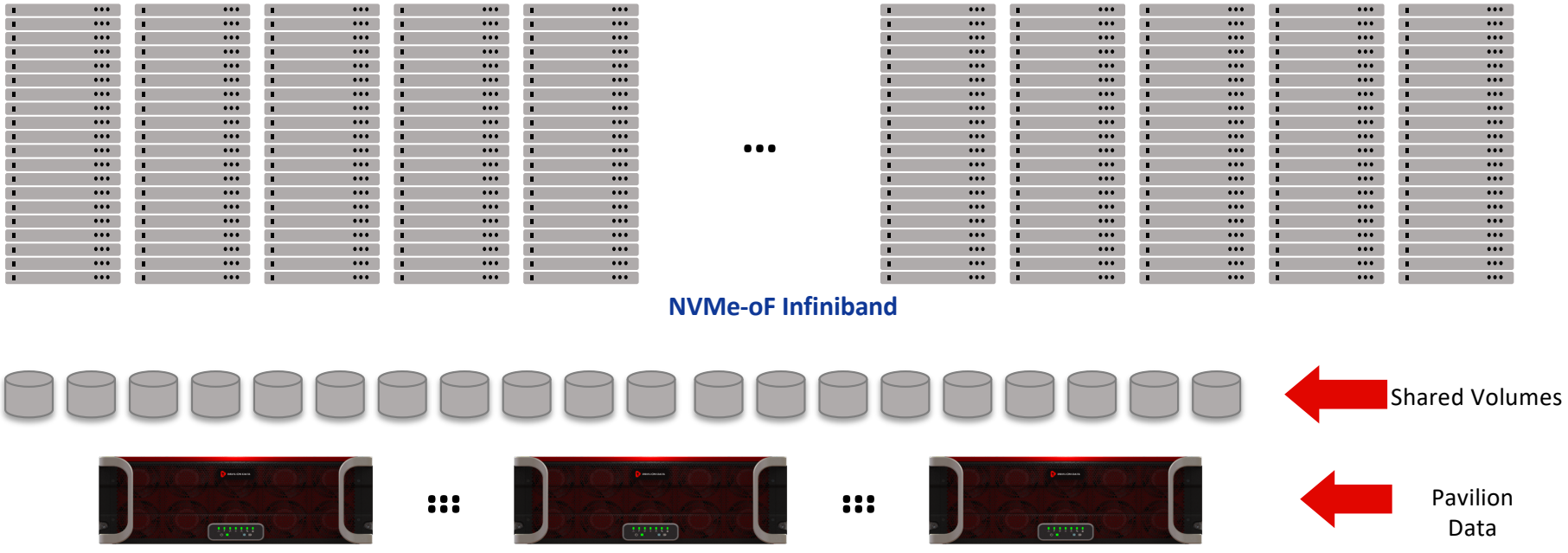
Customer C - Scaling IBM Spectrum Scale™



Massive deployment of facial recognition, correlations to other images and records
Multi-PB, 200+ GB/sec R/W bandwidth
512 Nodes with single namespace minimum
Preferred no host-side software, No SPOF, HA mandate
Wanted to consider elimination of NSD server tier

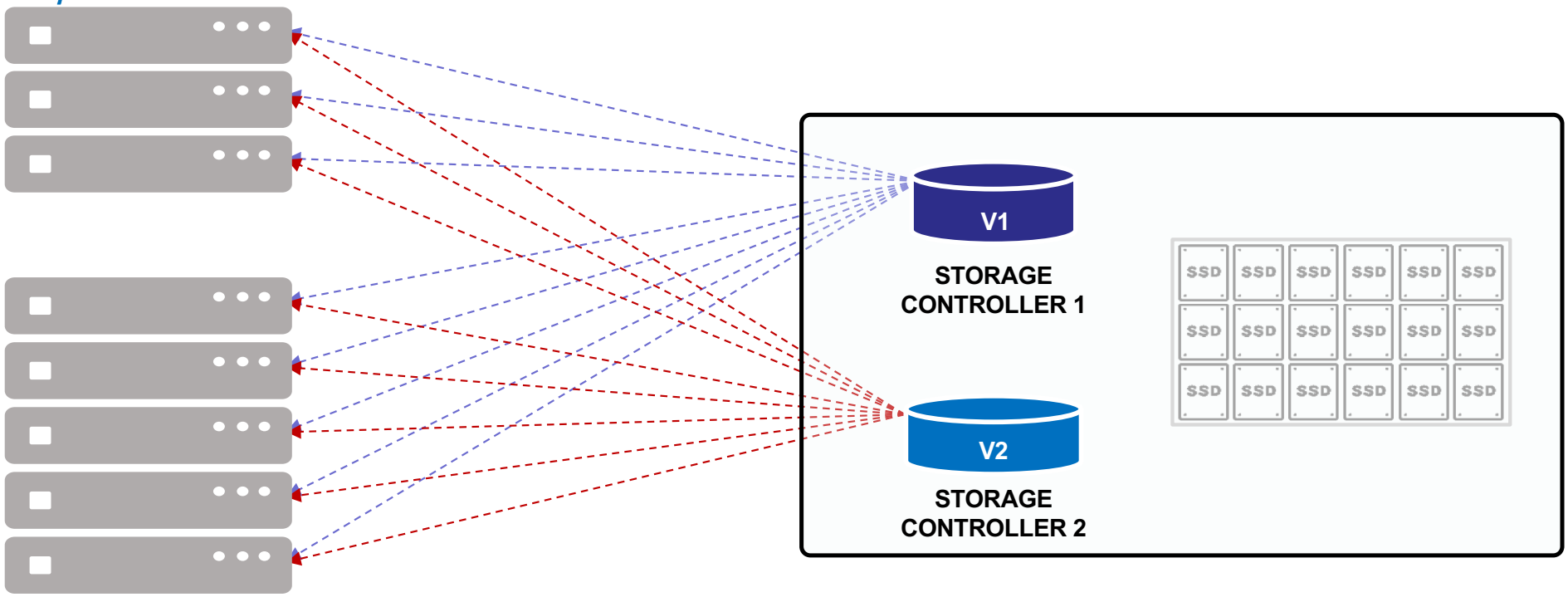
Flash Memory Summit

Customer C - Scaling IBM Spectrum Scale™



Avoided purchasing 20 NSD Servers – Savings paid for ~1 array
Delivering 550GB/sec Read & 400GB/sec Write Bandwidth
Encryption for data at rest
Snapshots for cluster migration and backup

Customer C - Volume Sharing & MPIO



MPIO for NVMe-oF was not yet a standard, it is now shipping in distros
Preferred pathing needs work for NVMe-oF is still not available
No concept of LUN masking in NVMe-oF

Summary

- Hype Cycle is overexaggerated
- But,
  - Challenges with big drives in DAS
  - Limited options for shared NVMe
  - MPIO, Volume Management immaturity
- Yet,
  - Customers are making the impossible possible