

Shared NVMe for High Performance Spectrum Scale Clusters

Zivan Ori

Co-founder and CEO

May 16th 2018

E8 storage

Rack-Scale Flash. No Compromise.



The E8 Storage Difference

A new architecture built specifically for high performance NVMe™

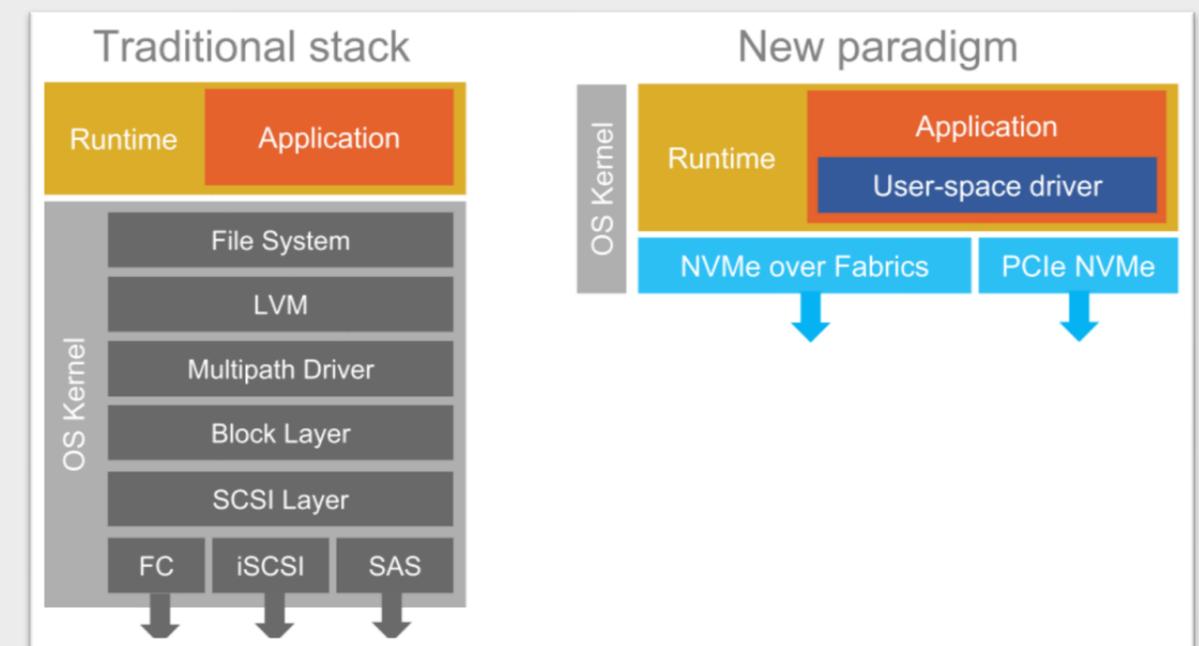
- Extract all performance from NVMe SSDs
- Use off-the-shelf hardware
- Scalable in multiple dimensions
 - Scale hosts for more computing power
 - Scale storage for higher capacity
- Simple, centralized management
- High reliability and availability



| What is NVMe™? (Non-Volatile Memory Express)

Communication protocol designed specifically for flash storage

- High performance, low latency
 - Efficient protocol with lower stack overhead
 - Exponentially more queues / commands than SAS
 - Parallel processing for SSDs vs serial for HDDs
- Support for fabrics (NVMe-oF™)
 - Originally designed for PCIe (internal to servers)
 - Expands support for other transport media
 - RDMA Based: RoCE, iWARP, Infiniband
 - Non-RMDA: FC, TCP
 - Maintains NVMe protocol end to end

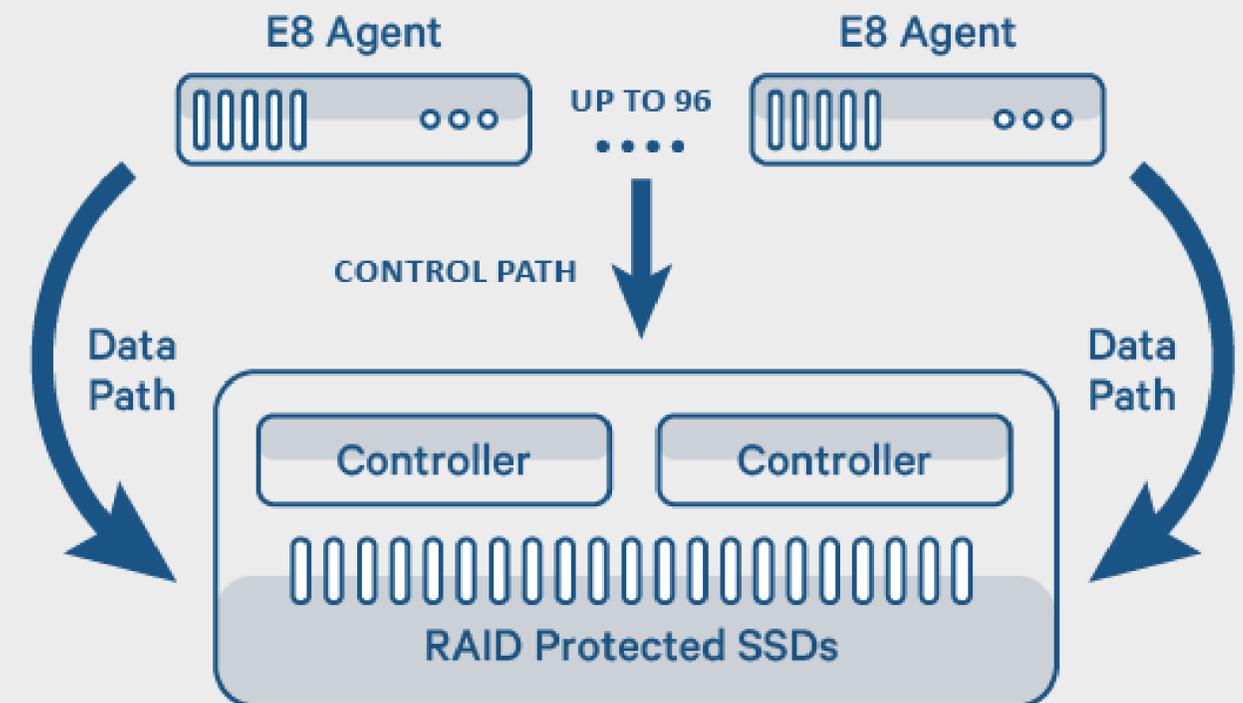


	Connections	I/O Queues	Commands per Queue
SAS	2	1	256
NVMe	128	65,535	64,000

Architected for High Performance NVMe

Separation of data and control; no controller bottleneck

- Centralized control operations
 - E8 Controllers manage all volumes, RAID config
 - Monitoring, management functions
- Distributed data operations
 - Built for IB or RDMA over Converged Ethernet (RoCE)
 - E8 Agents offload 90% of data path operations
 - Auto-discover provisioned volumes
- Leveraging the performance of RDMA
 - Enables direct access to flash as memory via network
 - Bypasses CPU / memory for fast reads



Designed for Availability and Reliability

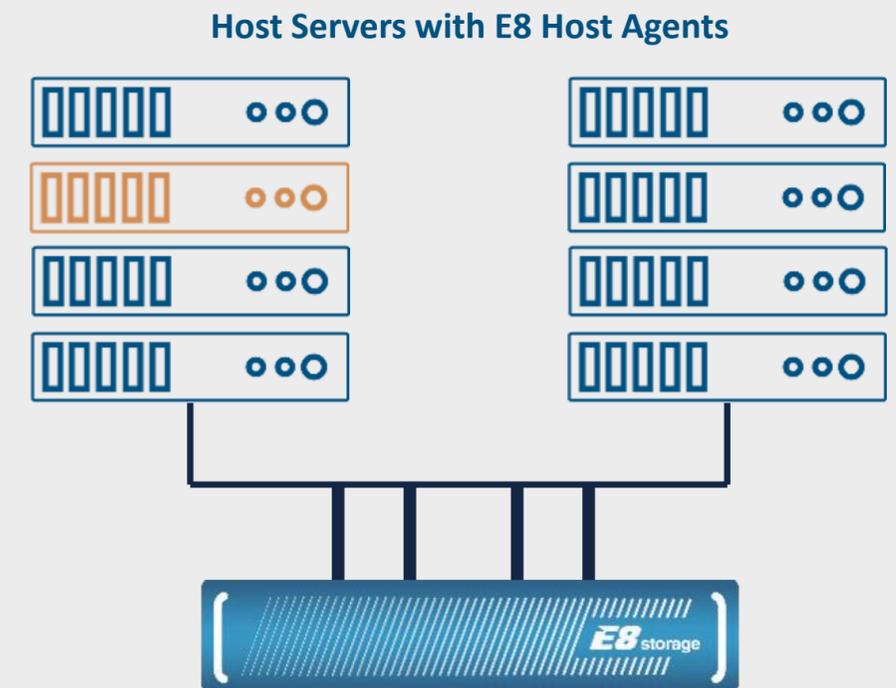
No single point of failure anywhere in the architecture

Hardware

- High-availability off-the-shelf appliances
 - Redundant controllers with auto-failover
 - Redundant power, cooling
 - All parts hot-swappable

Software

- Host agents operate independently
 - Failure of one agent (or more) does not affect other agents
 - Access to shared storage is not impacted
- RAID-6 data protection

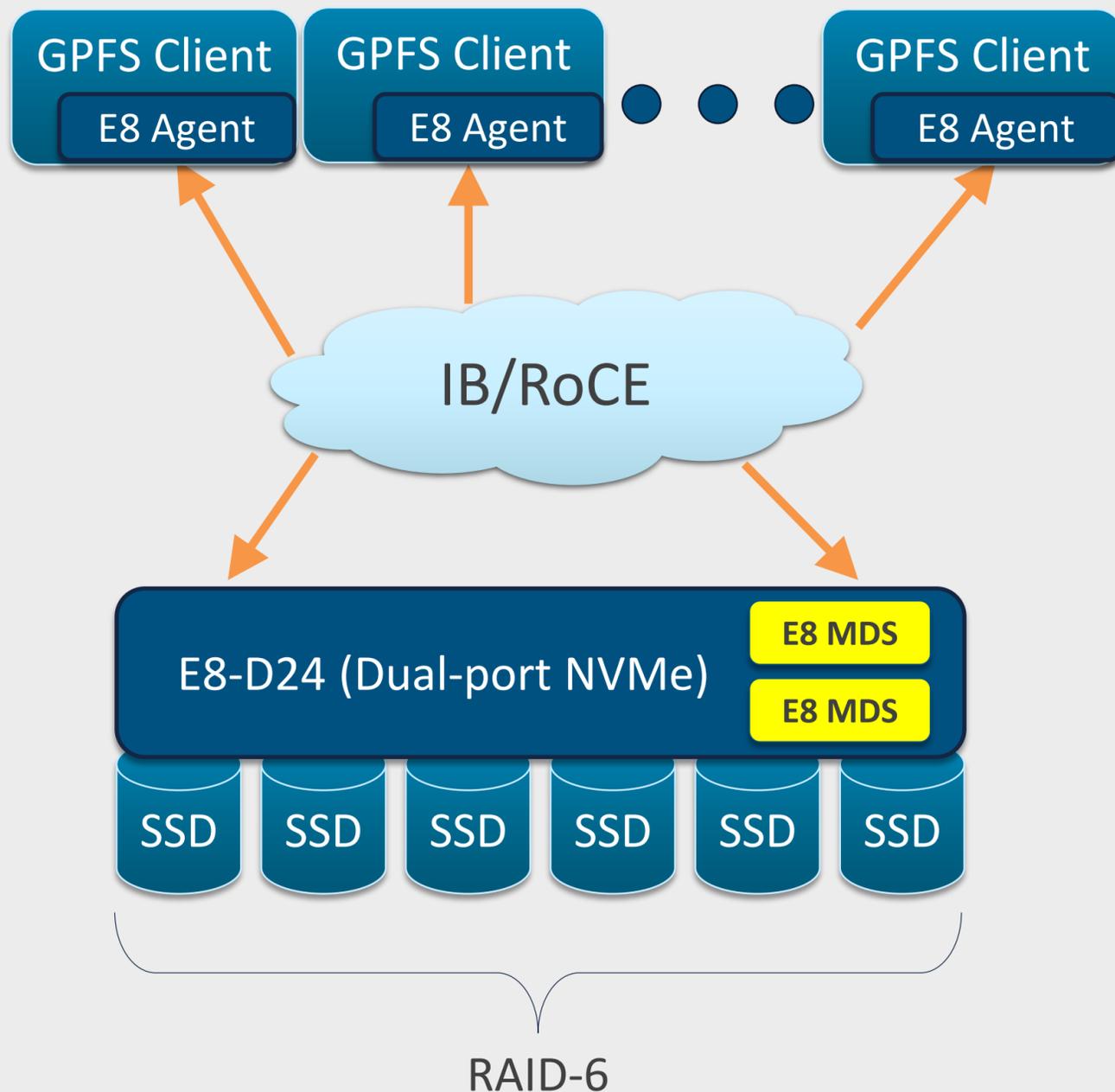


Using E8 with IBM Spectrum Scale

Multiple Deployment Options

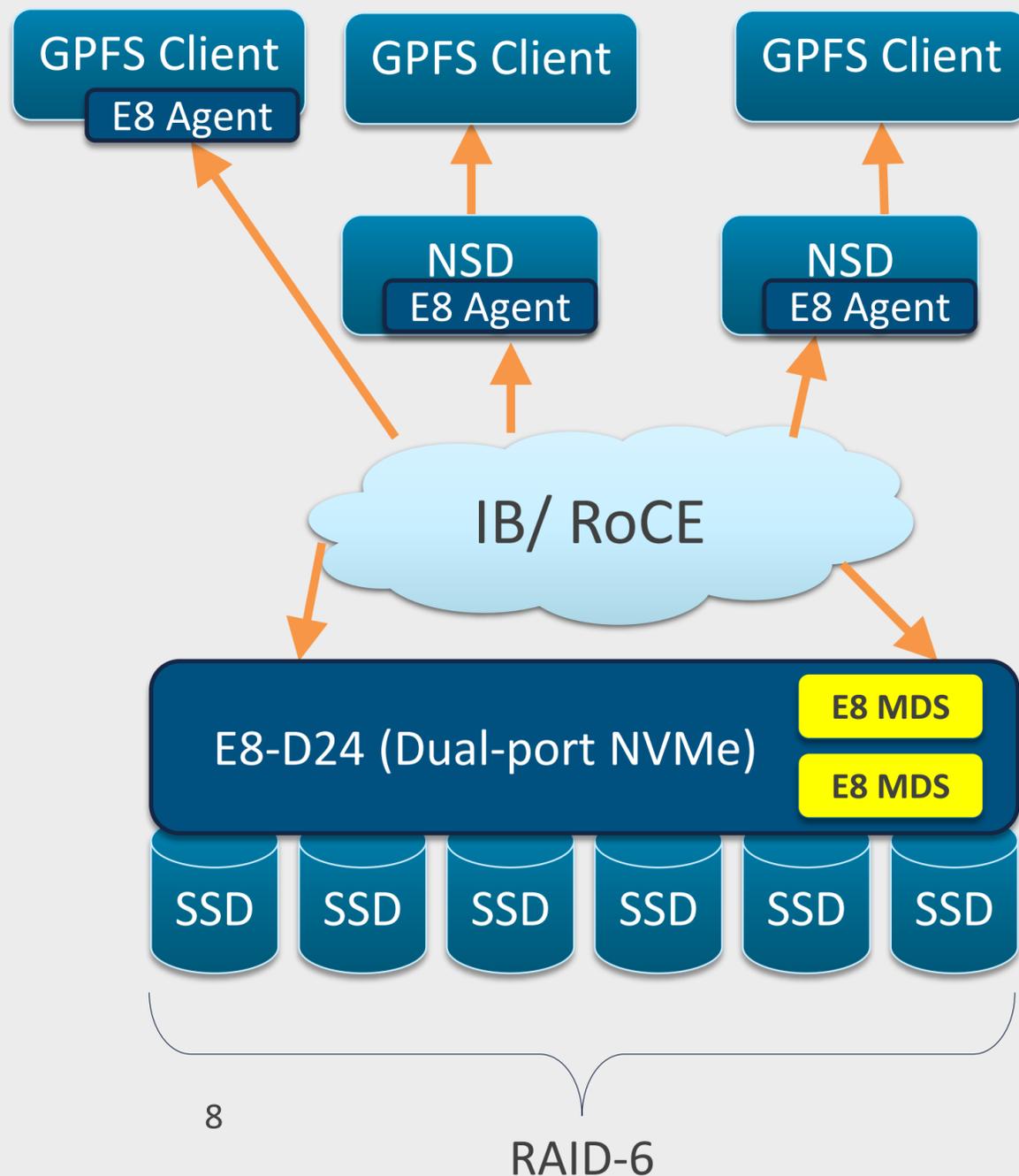
- Standalone pool
- Local Read Only Cache (LROC)
- High Availability Write Cache (HAWC)
- Metadata repository

Deployment – All Clients Connected Directly to Storage



- Scales to over 100 clients
- Direct access to clients, lowest latency
- Standalone pool
 - Shared LUNs
- LROC
 - Non-shared LUNs
- HAWC
 - Non-shared LUNs model

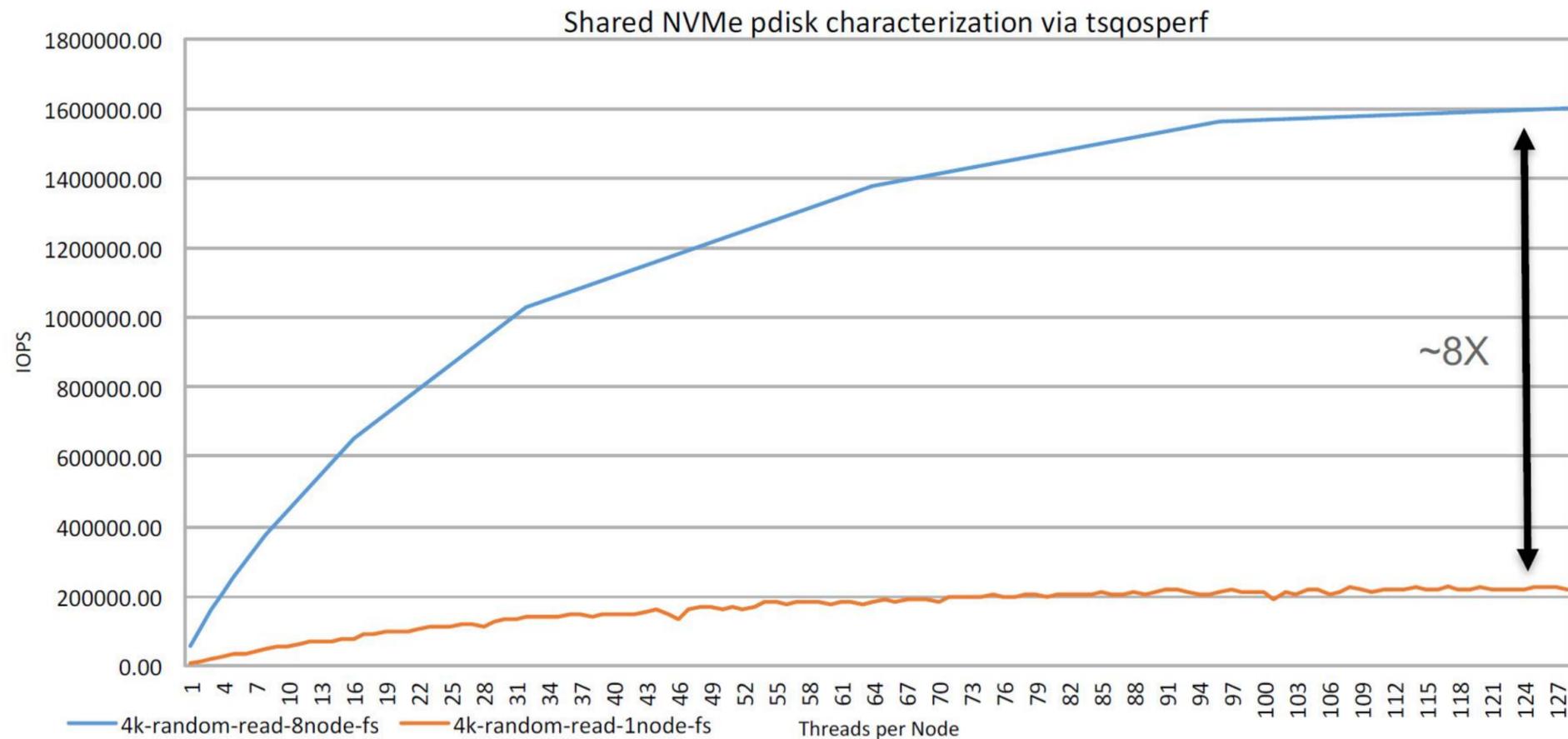
Deployment – Some Nodes Connected via NSD to Storage



- Scalable to larger configurations
 - Can mix connectivity depending on requirements
- Standalone pool
 - Shared LUNs
- LROC
 - Non-shared LUNs (direct connect clients only)
- HAWC
 - Shared LUNs model

Performance - E8 Storage and GPFS

NVMeoF device – 1 vs 8 Nodes – just filesystem access – 4k RR DIO



IBM Systems | 20

From IBM Research Performance Benchmarks of GPFS over E8

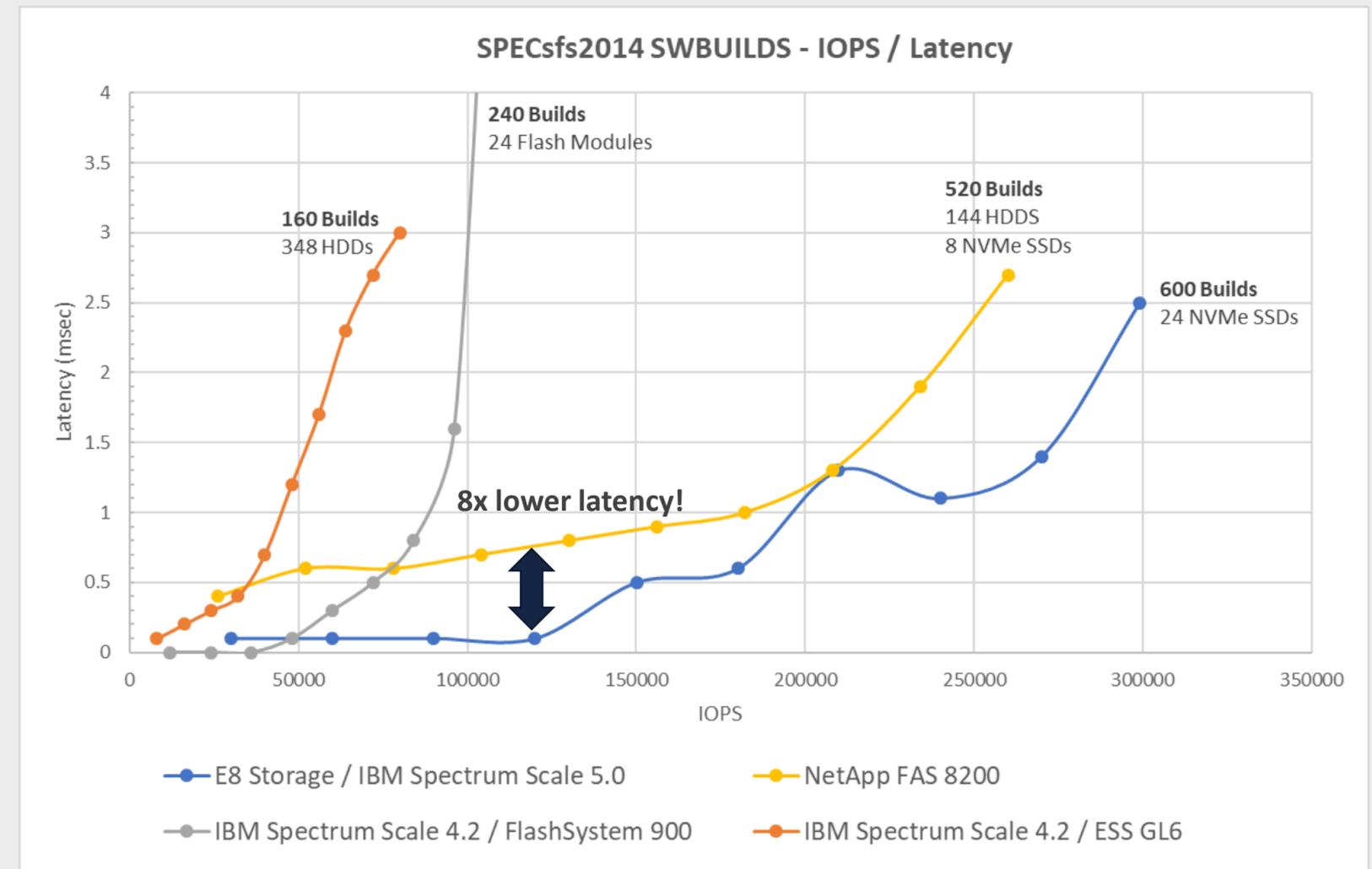
©2018 E8 Storage, Proprietary and Confidential



SPEC SFS® 2014_swbuild Performance*

- #1 performance for GPFS!
 - 2.5x more builds vs IBM all flash array
 - 8x lower latency vs previous record
- The only sub-millisecond ORT!
 - 0.69ms overall response time (ORT)
- More performance, less hardware

E8 Storage	24 NVMe SSDs	2U
NetApp FAS8200	144 HDDs, 8 NVMe SSDs	20U



* As of SPEC SFS® 2014_swbuild results published January 2018. SPEC SFS2014 is the industry standard benchmark for file storage performance. See all published results at <https://www.spec.org/sfs2014/results/>