nvm EXPRESS™

KIOXIA

lightbits

G2M RESEARCH

# Where Has NVMe-oF™ Progressed to in 2021?

Multi-Vendor Webinar

Tuesday June 15, 2021

# Webinar Agenda

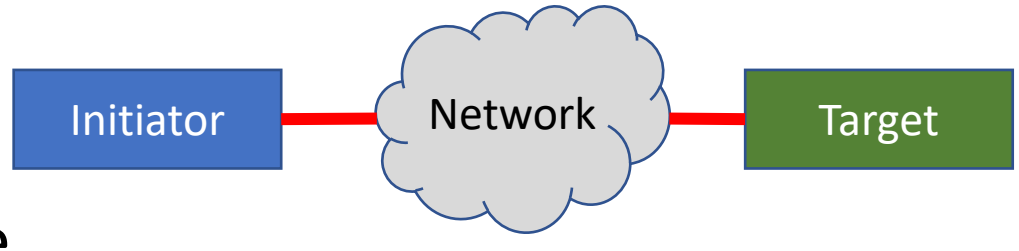
| | |
|---|---|
| **9:00-9:05** | Ground Rules and Webinar Topic Introduction (G2M Research) |
| **9:06-9:30** | Sponsoring Vendor presentations on topic (12 minute each) |
| **9:31-9:36** | Panel Discussion Question #1 |
| **9:37-9:37** | Audience Survey #1 |
| **9:38-9:43** | Panel Discussion Question #2 |
| **9:44-9:44** | Audience Survey #2 |
| **9:45-9:50** | Panel Discussion Question #3 |
| **9:51-9:58** | Audience Q&A (8 minutes) |
| **9:59-10:00** | Wrap-Up |

# G2M Research Introduction and Ground Rules

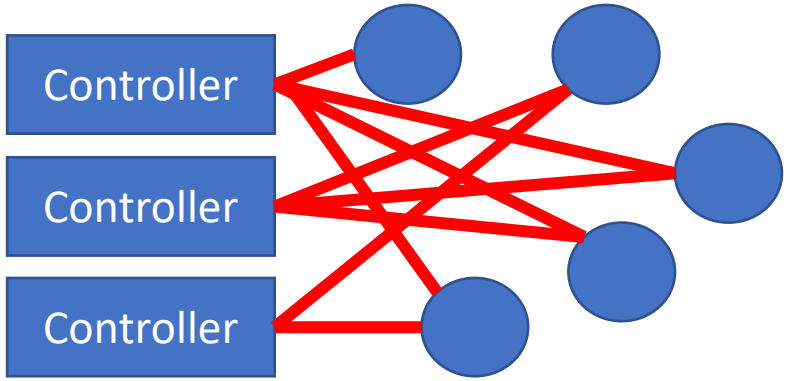
Mike Heumann
(Managing Partner, G2M Research)

# NVMe-oF: The "Conventional" View

- Storage Initiator/Target Use Case
  - Classical connection of storage users (initiators) and storage devices (targets)
  - Provides significantly better performance than SCSI-based protocols



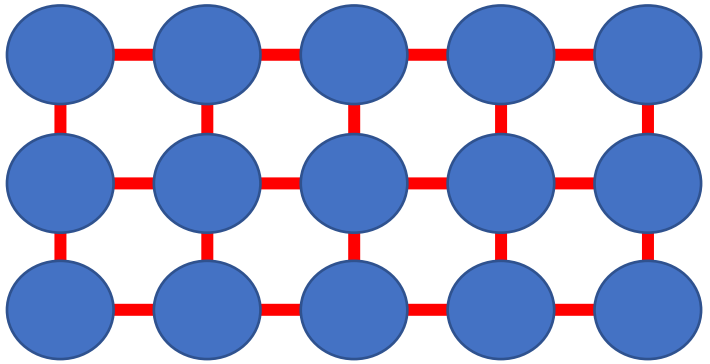

- All-Flash Array (AFA) Back-End Use Case
  - NVMe-oF replaces SAS/SATA "tree" topologies behind network controllers
  - NVMe-oF provides significantly more flexibility



# NVMe-oF: What Has Really Happened

**NVMe-oF has enabled "unconventional" use cases**

- Scale-Out Flash Storage (SOFS)
  - Connects servers and storage appliances with NVMe SSDs into a single namespace
  - Provides DAS-like storage performance, but with the ability to manage storage globally
  - Scale performance and capacity linearly
- Networked Storage Devices
  - NVMe-oF on 100GbE (10Gb/s) is faster than a PCIe 4.0 x4 connection (8Gb/s)
  - NVMe virtual namespaces allows SSD partitioning
  - Potentially eliminates SSD "blast radius" issue



# Panelists

KIOXIA

Matt Hallberg
Sr. Product Marketing Manager
www.kioxia.com

lightbits

Josh Goldenhar
VP, Products
www.lightbitslabs.com

G2M RESEARCH

Mike Heumann
Principal Analyst
www.g2minc.com

# G2M Webinar
# Where has NVMe-oF™ technology progressed to?

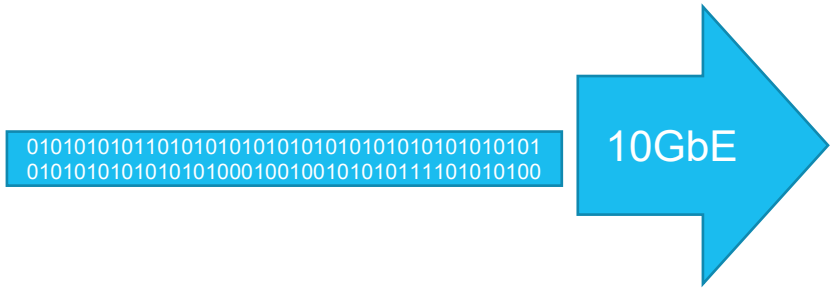
**Matt Hallberg**
**Sr Product Manager**
**Enterprise Storage**



KIOXIA

# NVMe-oF™ and Storage Primer

- Before adoption of NVMe® SSDs into servers, storage arrays and data centers, high bandwidth and low latency networks were not required, as the performance bottleneck was storage (SAS/SATA SSDs and HDDs)




- After adoption of NVMe SSDs, the bottleneck was no longer storage. 10GbE is no longer sufficient when local storage is capable of  millions of IOPs. Network protocols were not designed around the benefits of working with Flash like lower latency


- NVMe over Fabric (NVMe-oF) specification was created to bring high performance and low latency transactions out of the box and into the network

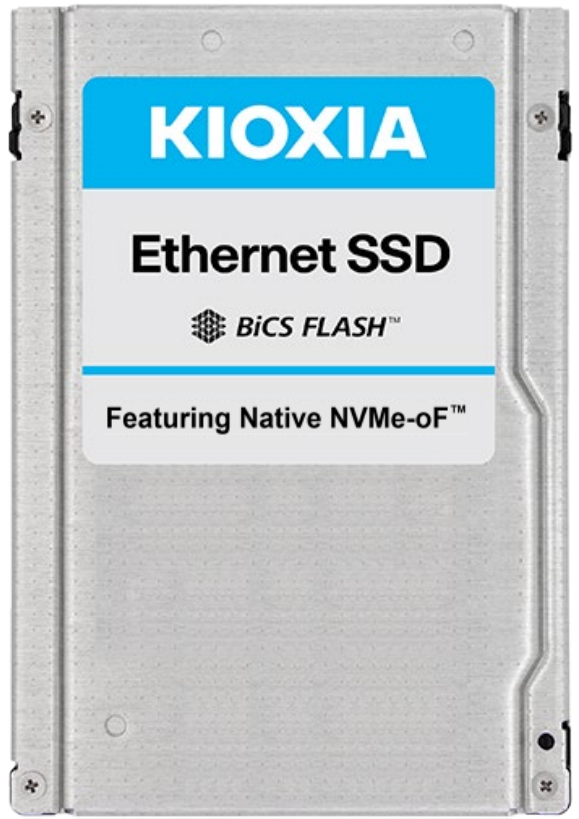
# NVMe-oF™ Adoption is On Track

- Adoption timeline comparable to NVMe®
  - NVMe standard published in 2012
    - Hyperscale adopted quickly, followed by “trickle down” to others
    - Very mainstream 9 years later
  - NVMe-oF specification published 2016
    - Hyperscale adopted quickly
    - On track to become mainstream everywhere
- Ecosystem adoption
  - VMware vSphere® 7.0
  - RHEL® (7.6+) and CentOS™ (7.2+)
  - Azure®, Oracle®, Pure Storage®, etc. all have newer offerings supporting NVMe-oF technology
- Critical improvements and additions to functionality
  - TCP Transport – expands coverage to customers with not using RDMA or Fibre Channel
  - Multipath improvements – (ANA, domains and divisions) ⇨ cross-node scale fault tolerance
  - Discovery and Transport – Persistent controller discovery and I/O queue deletion (gracefully)
  - Support for Controller Memory Buffer (CMB) and Persistent Memory Region (PMR)
  - DMTF Redfish™ / SNIA® Swordfish support (1.2.2) ⇨ Out of band fabric management
- New applications like Ethernet Attached SSDs, Disaggregated Composers (like KumoScale™)

# Ethernet Attached SSD Product Concept

- **World's 1st Native NVMe-oF™ Ethernet SSD**
  - Ethernet derivative of CM6/CD6
  - Best fit for expansion storage of AFA/SDS with EBOF (Ethernet Bunch-Of-Flash)

- **Key Features**
  - Marvell Fabrico-based solution
  - NVMe-oF 1.1 based on NVMe® 1.4
  - Dual 25 GBASE-KR Ethernet, RoCEv2
  - 2.5" SFF 15mmH
  - SFF-9639 Rev 2.1 (Added Native NVMe-oF pinout column Published on December 13, 2019)
  - 1920 GB
  - 1 DWPD



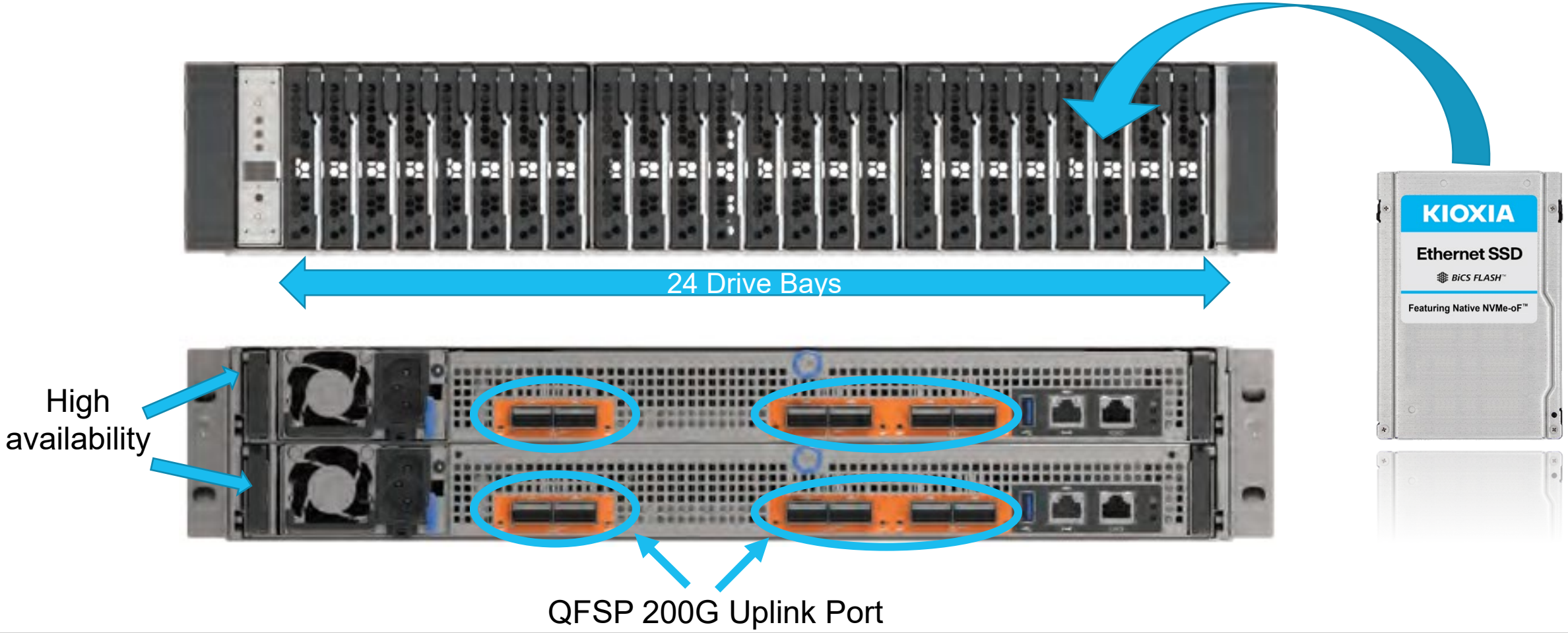

KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1Gb = $2^{30}$ bits = 1,073,741,824 bits, 1GB = $2^{30}$ bytes = 1,073,741,824 bytes and 1TB = $2^{40}$ bytes = 1,099,511,627,776 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

DWPD: Drive Write(s) Per Day. One full drive write per day means the drive can be written and re-written to full capacity once a day every day under the specified workload for the specified lifetime.. Actual results may vary due to system configuration, usage and other factors.

Source: KIOXIA Corporation, as of September 22, 2020 Product image/s may differ from the actual product.

# Ethernet Bunch Of Flash (EBOF)

- 6x 200Gbps high speed network capability
- High performance: 830K IOPS per drive, 20M IOPS per 24 bay EBOF (@4KB random read)

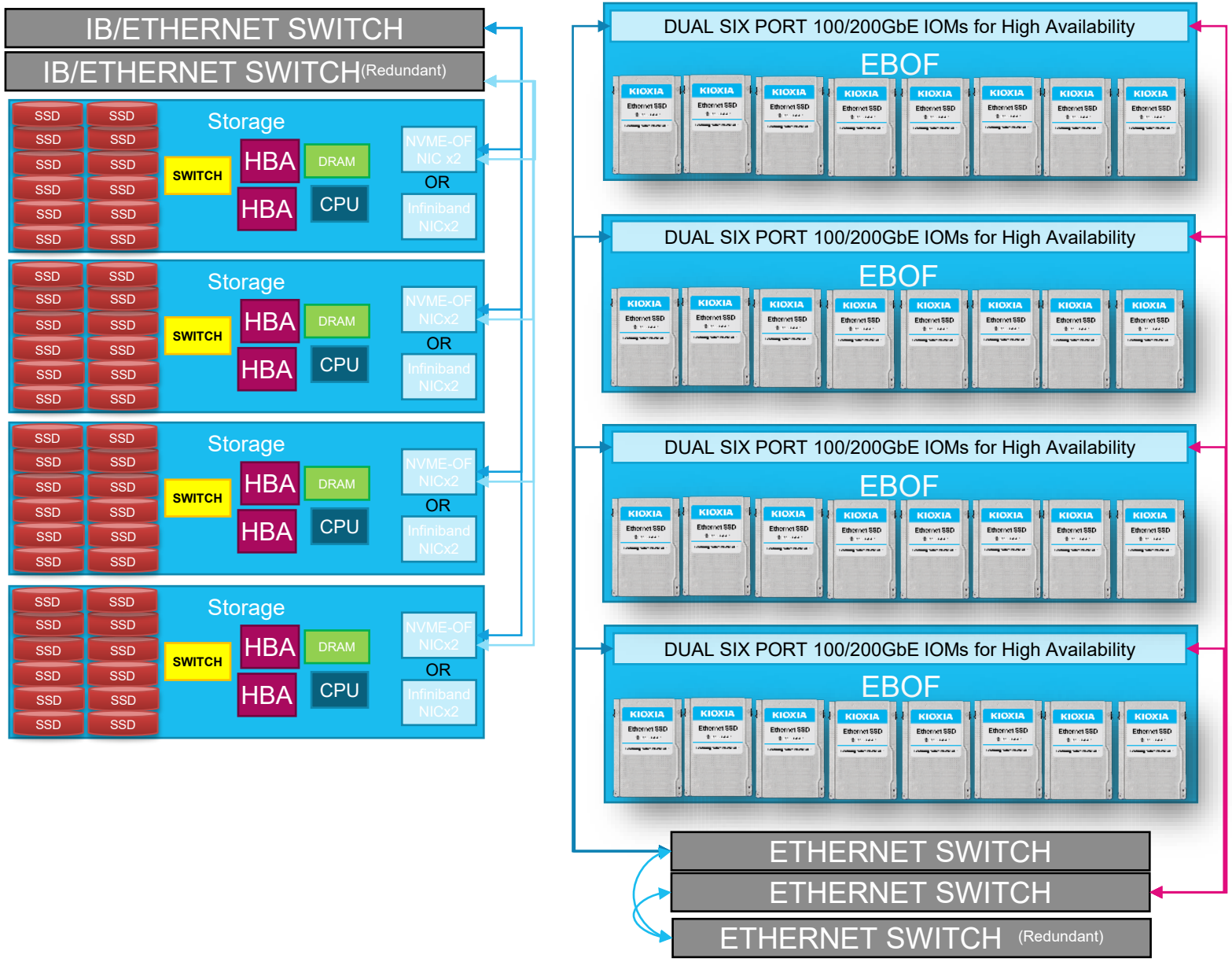





Read and write speed may vary depending on various factors such as host devices, software (drivers, OS etc.), and read/write conditions.

# Use Case: Storage Arrays

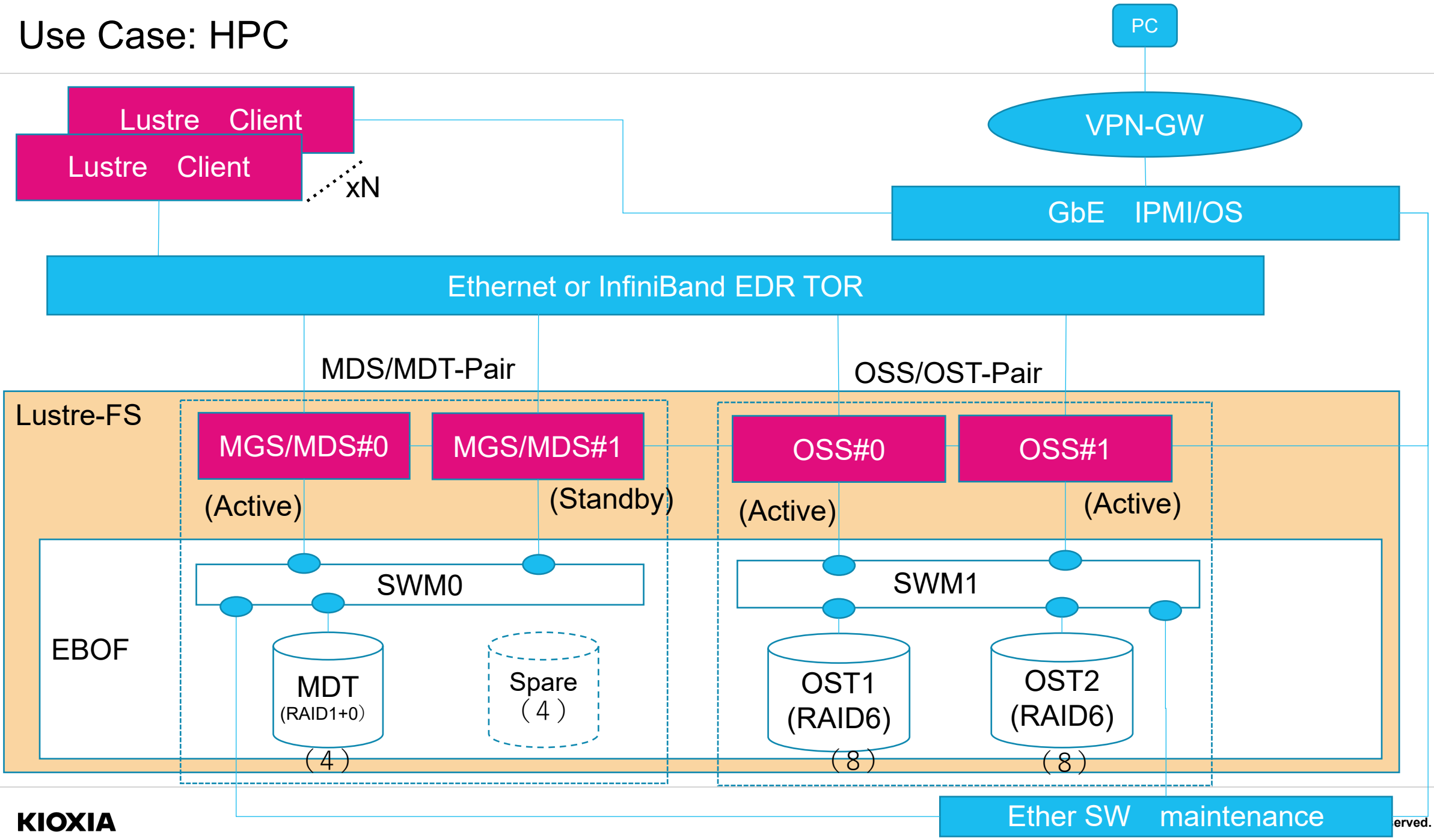
- NVMe-oF™ via RDMA over Converged Ethernet (RoCEv2) is increasing in popularity for Storage arrays as an alternative to InfiniBand with minimal performance tradeoffs
- Currently each storage box requires multiple HBAs and NICs for redundancy
- EBOF architecture “bakes” redundancy / high availability into the box itself
- Cost reduction via
  - System: 2 NICs, 2 HBAs, CPUs, etc.
  - Cooling: Only cooling SSDs!





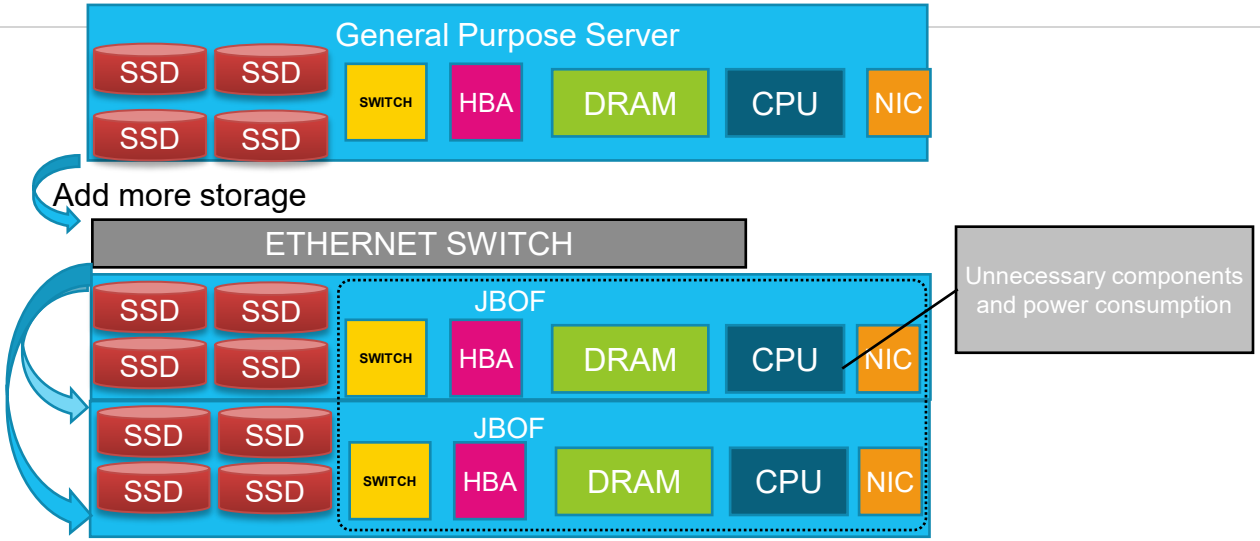

NVMe-oF is a trademark of NVM Express, Inc.

# Use Case: HPC

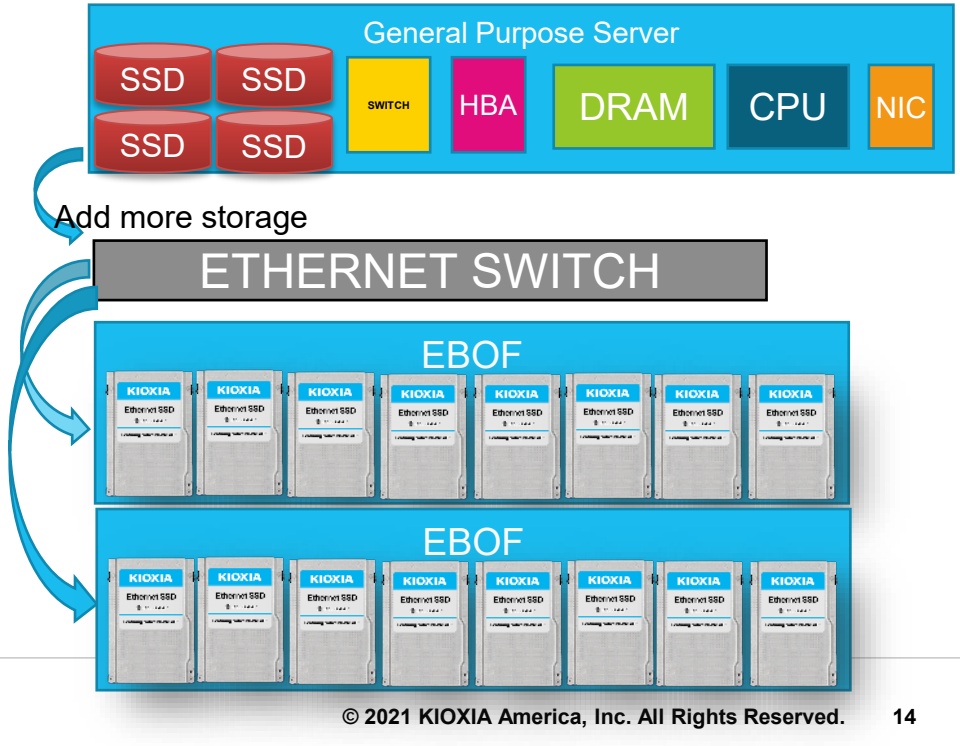


# EBOF Use Case: Data Centers

- Performance and Latency
  - NVMe-oF™ topologies are low latency and high performance
- Efficiency
  - No latency overhead from CPU + components (DRAM, NIC, HBA, etc.)
- Supply
  - Not subject to long lead times on CPUs and other critical components
- Cost
  - System cost is reduced by removing CPU, DRAM, NIC and HBA
  - System cooling cost decreased as only cooling up to 24 drives (@ 2U)
    - Drives are pulling <20W / Ethernet SSD, 24x20W = 480W
    - General purpose servers are rated for a minimum of 750W or more
- Scalable Disaggregated Storage
  - Just add EBOF(s) to RoCEv2 network




**EVOLVE!**

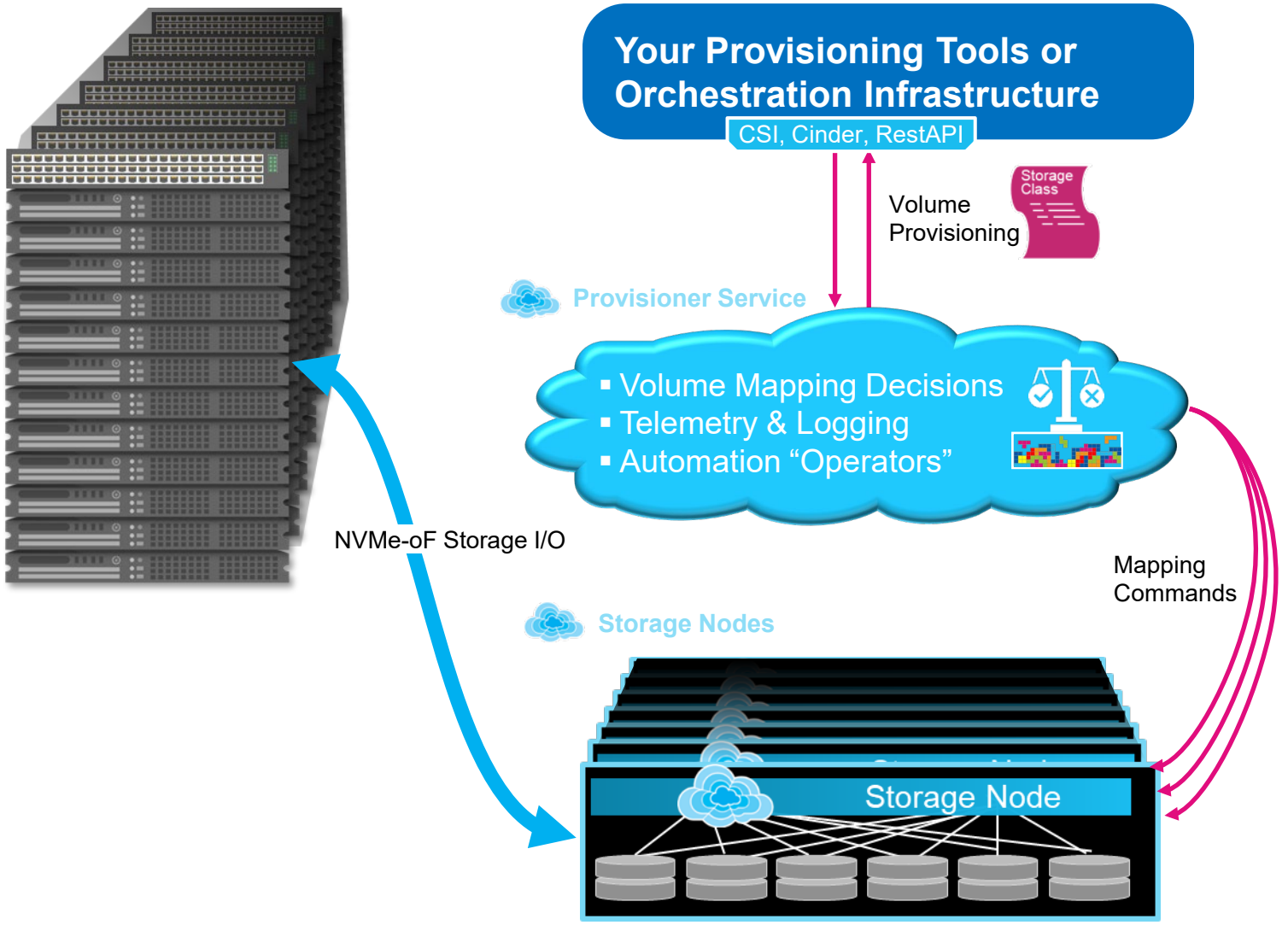





NVMe-oF is a trademark of NVM Express, Inc.

# Requirements to deploy NVMe-oF™ storage at Data Center Scale

KUMOSCALE™

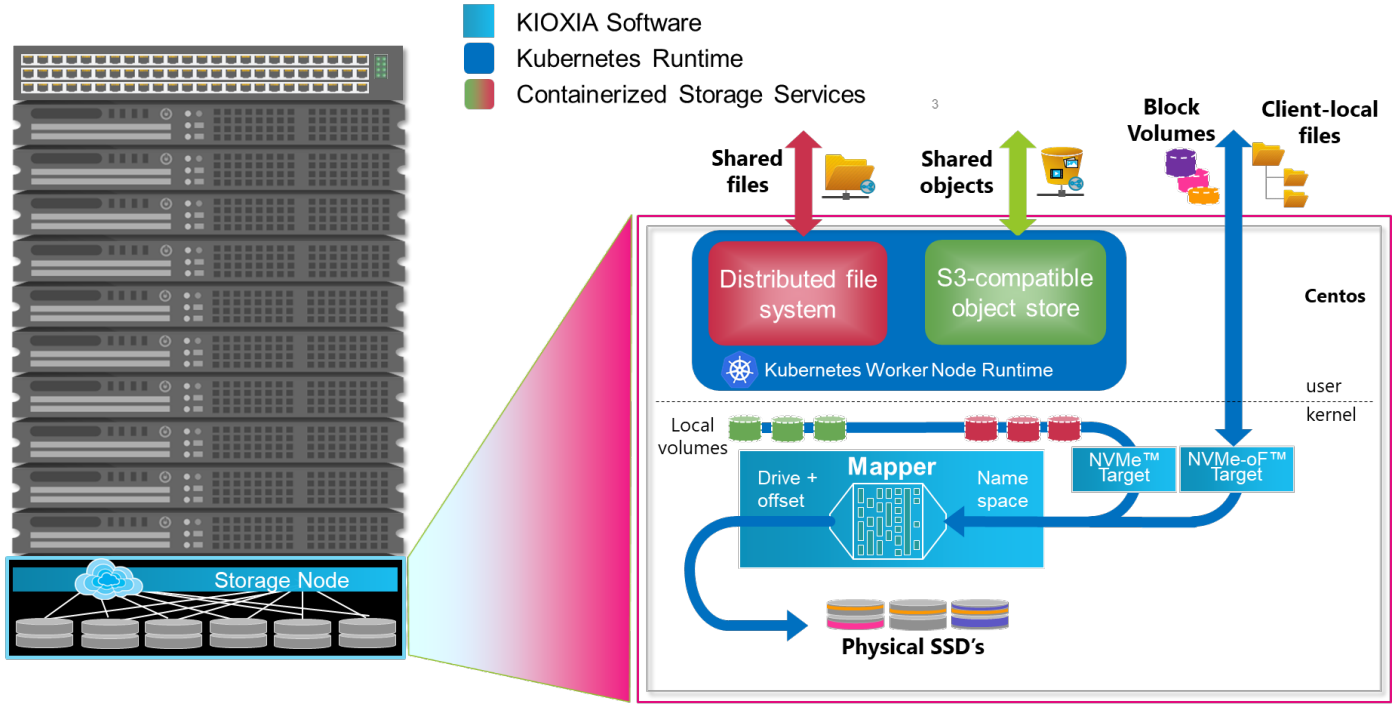



- Federation
  - Unified cluster API
  - Zero-touch deployment at any scale
  - Securely serve multiple tenants/business units from a single cluster.
- State of the Art Networking
  - RDMA or TCP transport
  - eBGP route announcement and failover
  - Zero-trust architecture
- Failure resilience without HA hardware
  - Topology-aware data placement across racks/zones
  - Automated self-healing
- Automation
  - Declarative cluster management via "Operators"
  - Per-application performance SLAs via "Storage Classes"
- Integration
  - Orchestration Frameworks: Kubernetes, OpenStack
  - Automation Frameworks: Ansible, REST API, …
  - Distributed Filesystems: GPFS, Lustre, GlusterFS, …
  - Telemetry & Logging: Prometheus, Graphite, …
- Advanced storage services
  - Thin provisioning
  - Snapshots/clones
  - Online volume resizing & migration

# NVMe-oF™ Architectural Progression (Datacenter use cas

KUMOSCALE™

- First deployments:
  - Drives lack specific NVMe-oF functionality
  - In-band mapping in software: NVMe-oF commands (to virtual volumes) ⇨ NVMe® commands (to physical drives)
- Long-term:
  - Drives “speak” NVMe-oF commands natively
  - Static, out-of-band mapping




Josh Goldenhar
VP Products
www.lightbitslabs.com

# Over a Decade of Innovation

Lightbits Team accomplishments and contributions in the NVMe Space

## NVMe

**Direct-attached**
High-performance PCIe SSDs

**2009 - 2013**

- **First NVMe SSD controller**
- Adopted by top Hyperscalers and all-flash-arrays
- First Linux & VMware drivers

## NVMe-oF

**Rack-scale**
Remote NVMe SSDs on RDMA or FC fabrics.

**2014 - 2016**

- **Defined NVMeoF**
- First NVMe-rack-scale storage solution

## NVMe/TCP

**Cloud & Hyper-scale**
NVMe/TCP across the data-center

**2017 - 2020**

- **Pioneered NVMe/TCP**
- Contributed code to upstream kernel
- First NVMe/TCP SDS
- First Clustered NVMe SDS

## NVMe/TCP

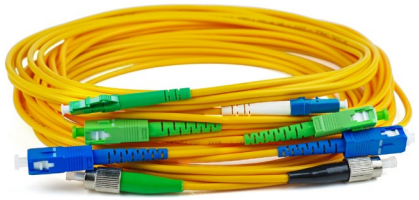
**No Compromise**
Advanced data services, environments and scale

**2021**

- **Snapshots and clones at NVMe speed**
- Support for OpenStack, Kubernetes and More in 2nd ½ 2021
- Continued influence in NVM Express Technical Working Group

# NVMe-oF

- NVMe-oF (RoCE) 2016: RDMA is great – if you need that absolute highest level of performance with the lowest latency,
  - …and are willing to limit your network interface card selections
  - …and are willing to make changes to your switches
  - …and want to re-learn how to do link aggregation and Multi-Chassis Link Aggregation
- NVMe-FC:
  - Well… it requires FC



- NVMe-oF TCP (NVMe/TCP)
  - Standard practices, Ethernet and TCP always win
  - InfiniBand, FDDI, CDDI, FC
  - ATA over Ethernet (AoE) anyone? FCoE?

# 2021: The year of NVMe/TCP?

- 2019
  - Pure Storage announces NVMe-oF (RoCE) and says NVMe/TCP will be coming…
- 2020
  - VMware announces support for NVMe-oF (RoCE)
  - Other large (Dell-EMC) or smaller vendors quietly support NVMe-TCP or annojunce pl ans for NVMe/TCP
  - Lightbits Labs announces full clustered/failover/scalable NVMe/TCP SDS
- 2021?

*Movement to support NVMe/TCP from a major OS or application environment could really move the needle for NVMe/TCP and hence, NVMe-oF in general*

# NVMe/TCP - A Replacement for Fibre Channel and iSCSI SAN

Half the cost, 3 times the performance

## Fiber Channel 32Gbps

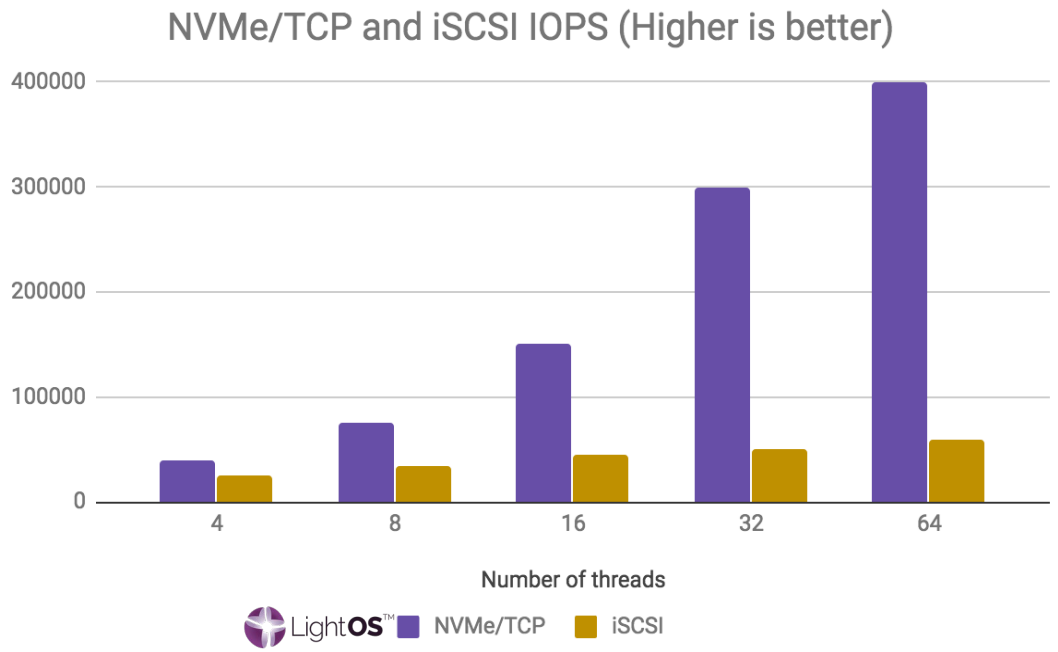
- Expensive
- Specialty Vendors
- Interoperability issues with different speeds
- Requires Fibre infrastructure

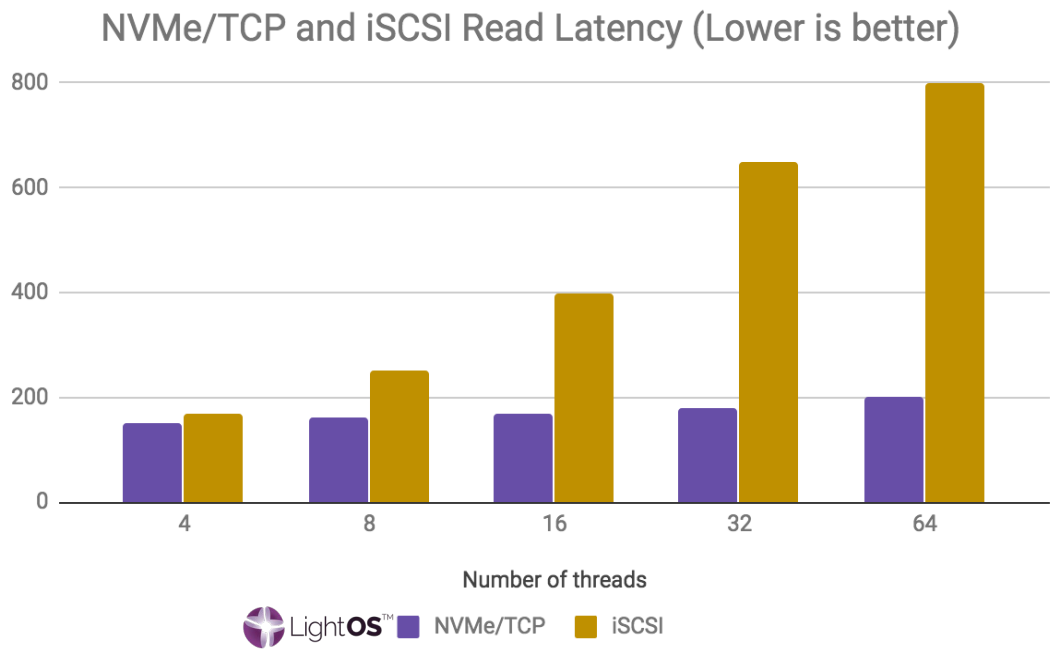
## NVMe/TCP on 100GbE

- **½ the cost of FC32** per-port
- **3X the performance** of FC32
- Commodity - large selection of vendors
- Ethernet and TCP/IP ubiquity

# iSCSI vs. NVMe/TCP

Same hardware, vastly different results

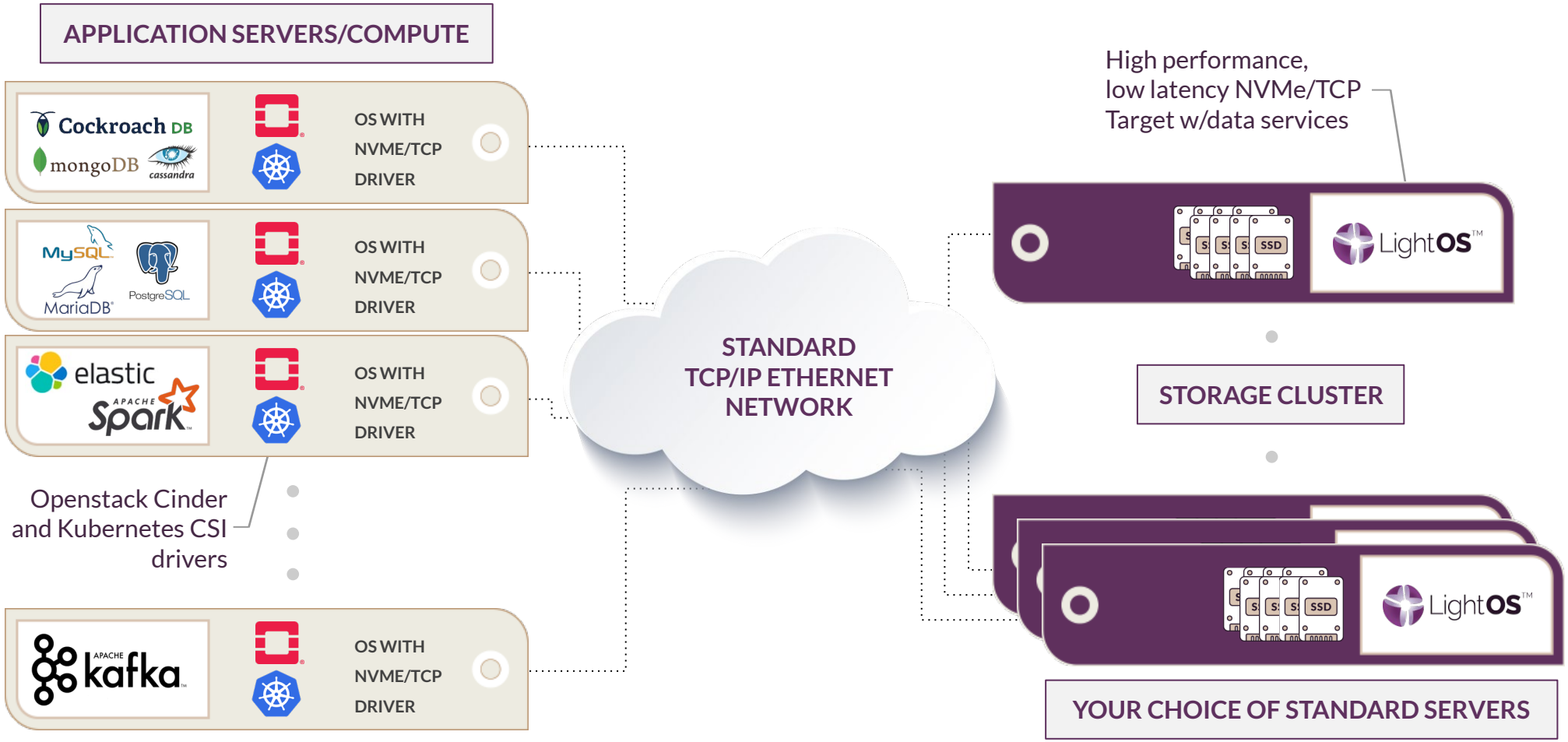



*NVMe/TCP scales linearly with over **6X more IOPs** vs. iSCSI at the same thread counts while attaining as much as **4X lower latency** vs. iSCSI - on the same hardware!*

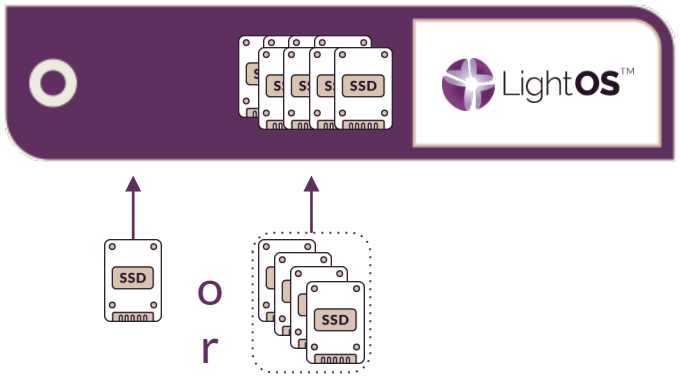
# High Performance Software Defined Storage

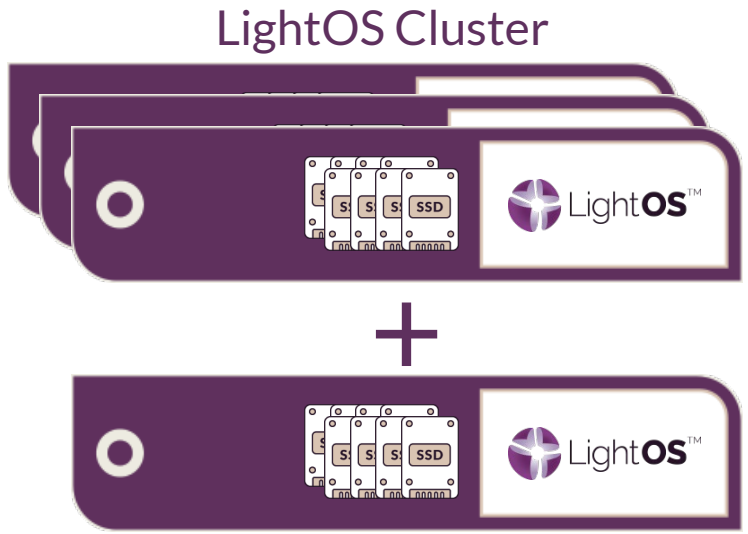
LightOS software + your choice of standard servers, NICs and SSDs



# Flexibility: Independently Scale Storage

Scale up, or out, or both




## Per Target Server

- Start partially populated, add additional NVMe drives at any time
- Add 1 or many at once with no disruption in service

## Per LightOS Cluster

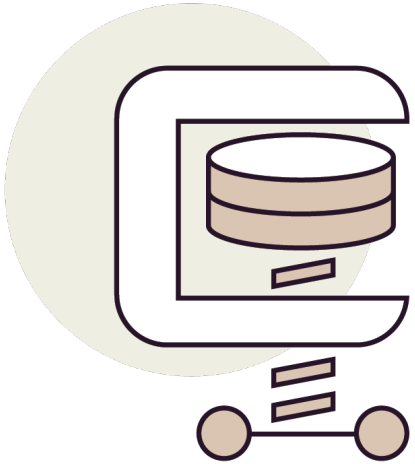
- Start with at least 3 target servers
- Add additional target server to the cluster at any time, online
- Cluster dynamically rebalances

# Rich Data Services

Local flash performance, array-like features

**NVMe virtual block devices (NVMe/TCP):**

- All volumes thin provisioned
- Line rate compression configurable per volume
- Data protection level per volume
- Volume expansion
- Volume Snapshots (read-only) and Clones (read/write) - all thin provisioned
- Low tail latency and consistent response time

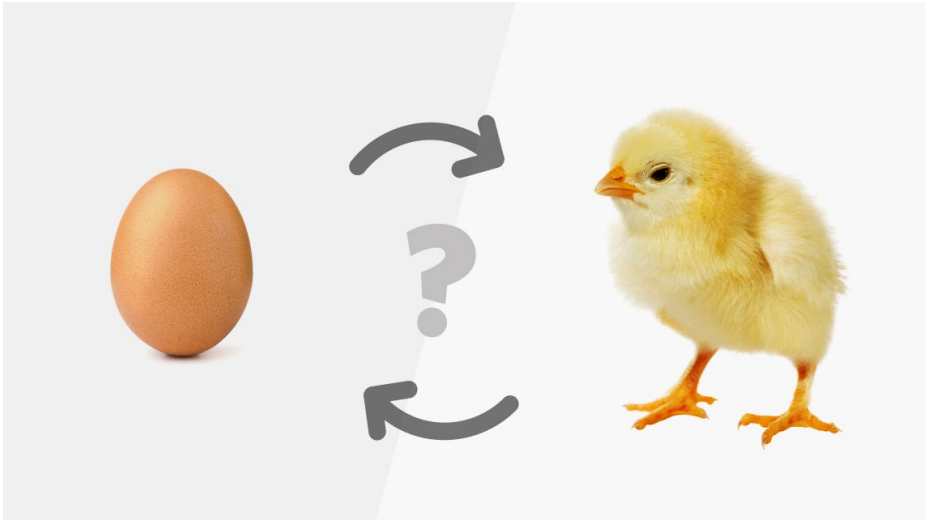


LightOS™

# What's Missing to be "Next Generation SAN"?

- Centralized  and automatic discovery of NVMe-oF Controllers
  - Lightbits' CTO, Sagi Grimberg is still on the NVM Express Technical Working Group
  - Technical Proposals (TP) 8009 and 8010 are underway
- In-Band Authentication
  - TP-8006 is also in the working group, to provide volume (controller) authentication similar to mechanisms for iSCSI

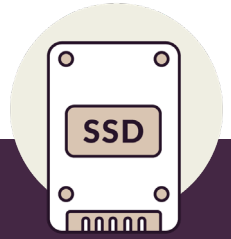
Adoption
Standards
Demand




Demand
Adoption
Standards

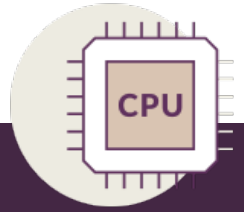
# Summary – A Great Replacement for SAN Today and Tomorrow

Shared, reliable, NVMe block storage that performs like local SSDs




## Bring Your Own Hardware

- Software-defined solution
- Runs on commodity storage server HW
- Standard NVMe SSDs




## Intelligent Flash Management

- Thin provisioning
- Compression
- Snapshots & Thin Clones
- Data protection
- QoS services
- QLC optimized

## No Change in Network (NVMe/TCP)

- TCP/IP network infrastructure - ubiquitous, simple & efficient
- Run on standard Ethernet NICs

## Don't Touch Clients (Target Side Solution)

- No proprietary client software
- High availability targets with ANA
- Standard NVMe/TCP client driver

# Panel Questions and Audience Surveys

# Panel Question #1

- What are the leading storage use cases that really take advantage of NVMe-oF's capabilities?
  - Kioxia
  - Lightbits Labs

# Audience Survey Question #1

- What experience does your organization have with NVMe-oF? (check all that apply):
  - Explored information on its use (conferences, articles, etc.): 58%
  - Talked to NVMe-oF vendors (NW adapters, storage, software, etc.): 27%
  - Defined potential NVMe-oF projects: 21%
  - Started one or more proof-of-concept evaluations: 24%
  - Budgeted for actual production NVMe-oF deployments: 9%
  - Deployed NVMe-oF in production: 27%

# Panel Question #2

- Will NVMe-oF's performance over Ethernet mean the death of "conventional" SAN technologies like Fibre Channel, even with NVMe-oF?
  - Lightbits Labs
  - Kioxia

# Audience Survey Question #2

- Which classes of NVMe-oF use cases have your organization evaluated? (check all that apply):
  - Scale-Out Flash Storage deployments (servers and/or storage appliances with local storage in a single common namespace): 64%
  - Deployment of all-flash arrays with NVMe-oF back-ends: 50%
  - Deploying NVMe-oF into existing or new networked storage configurations: 64%
  - Other use cases (converged infrastructure, etc.) 48%

# Panel Question #3

- Clearly, NVMe-oF enables a number of changes to storage networking over legacy SCSI-based protocols. Where does NVMe-oF go in the future?
  - Kioxia
  - Lightbits Labs


# Audience Q&A


Thank You For Attending!