What’s going on with NVMe?

An examination of new technology adoption

Mike Scriber

Sr. Director, Server Solution Management

9/23/2020

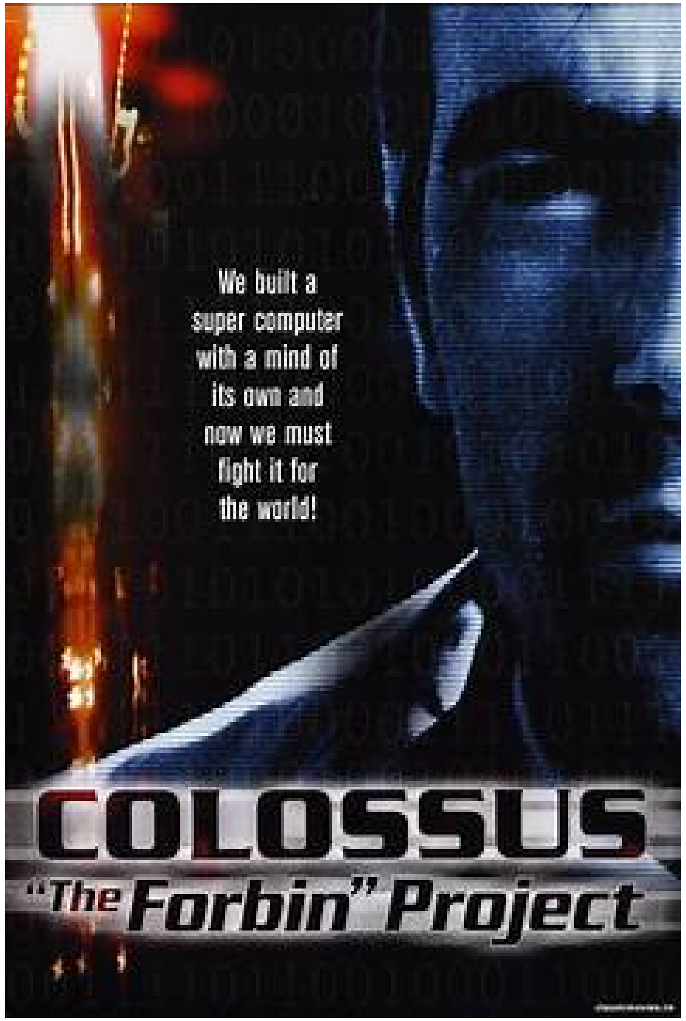
What's going on with NVMe?

- Our Industry Pace
- NVMe Growth
- More is Better
- Where is EDSFF going?
- What is QLC?
- Why NVMeoF?
- GPU Direct

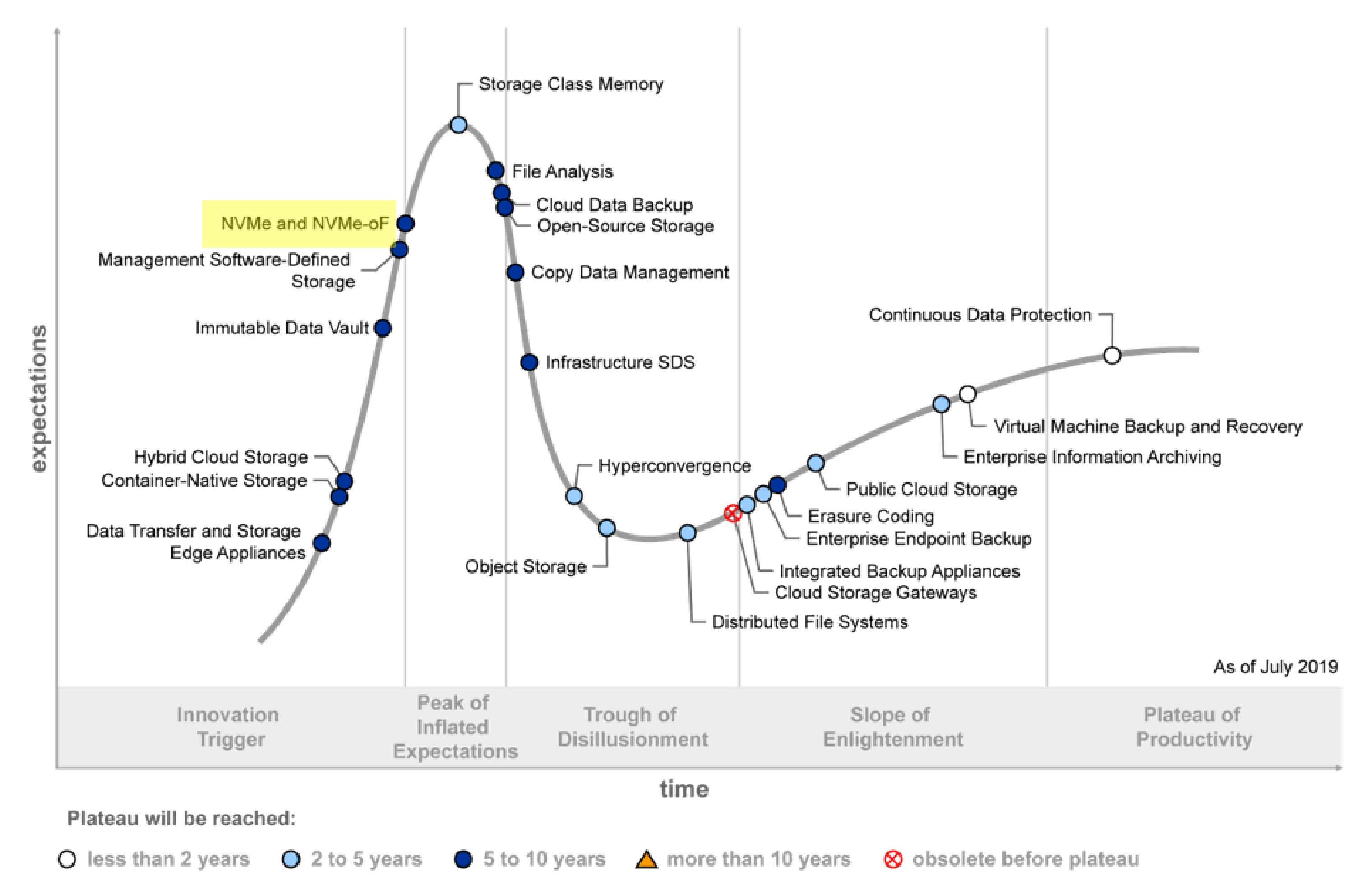
We are driving fast and hard

Our customers are on their own pace

Is technology just science fiction?



Hype Cycle for Storage Technologies 2019



Source : https://www.gartner.com/doc/reprints?id=1-1YH750DY&ct=200225&st=sb

Will Flash Penetrate All Enterprise Storage?

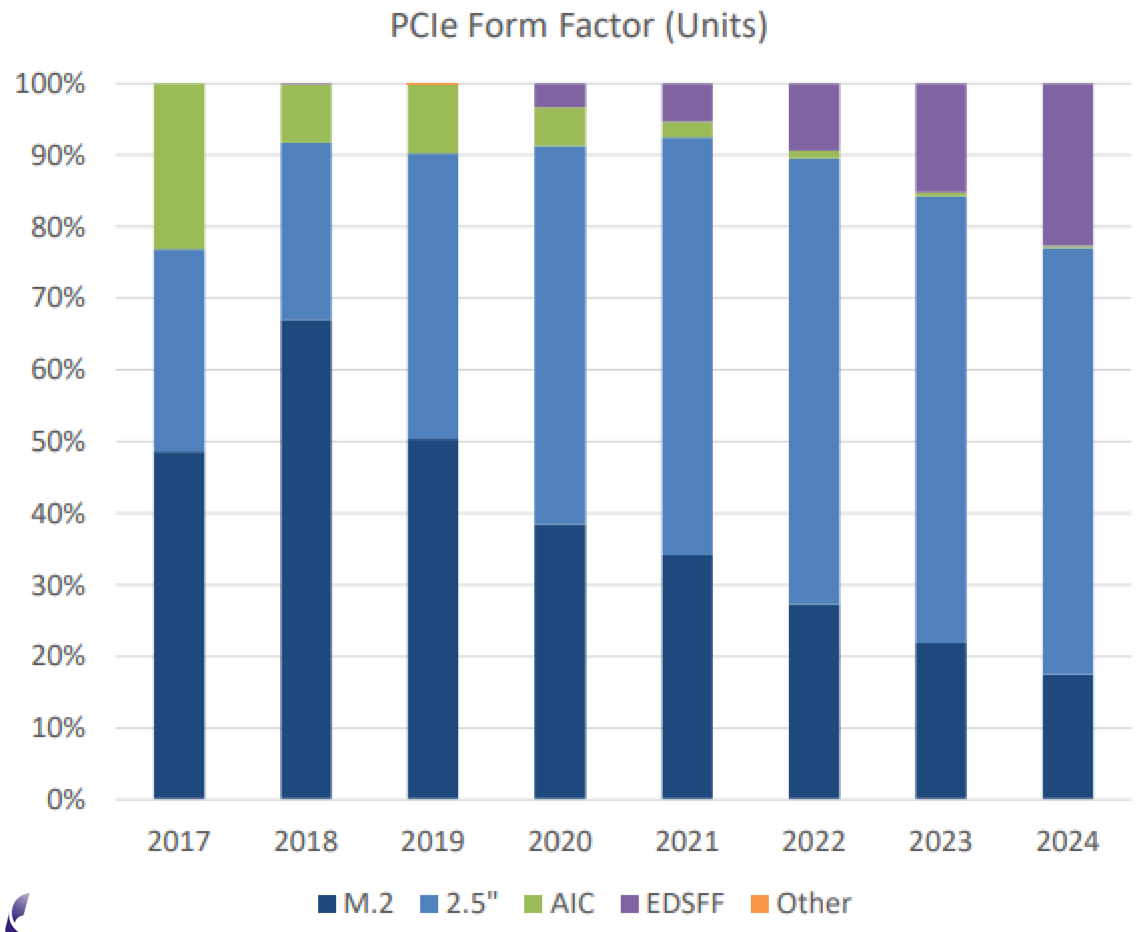
- **IDC** reports by 2019, AFAs were generating almost **80%** of primary external storage revenues.

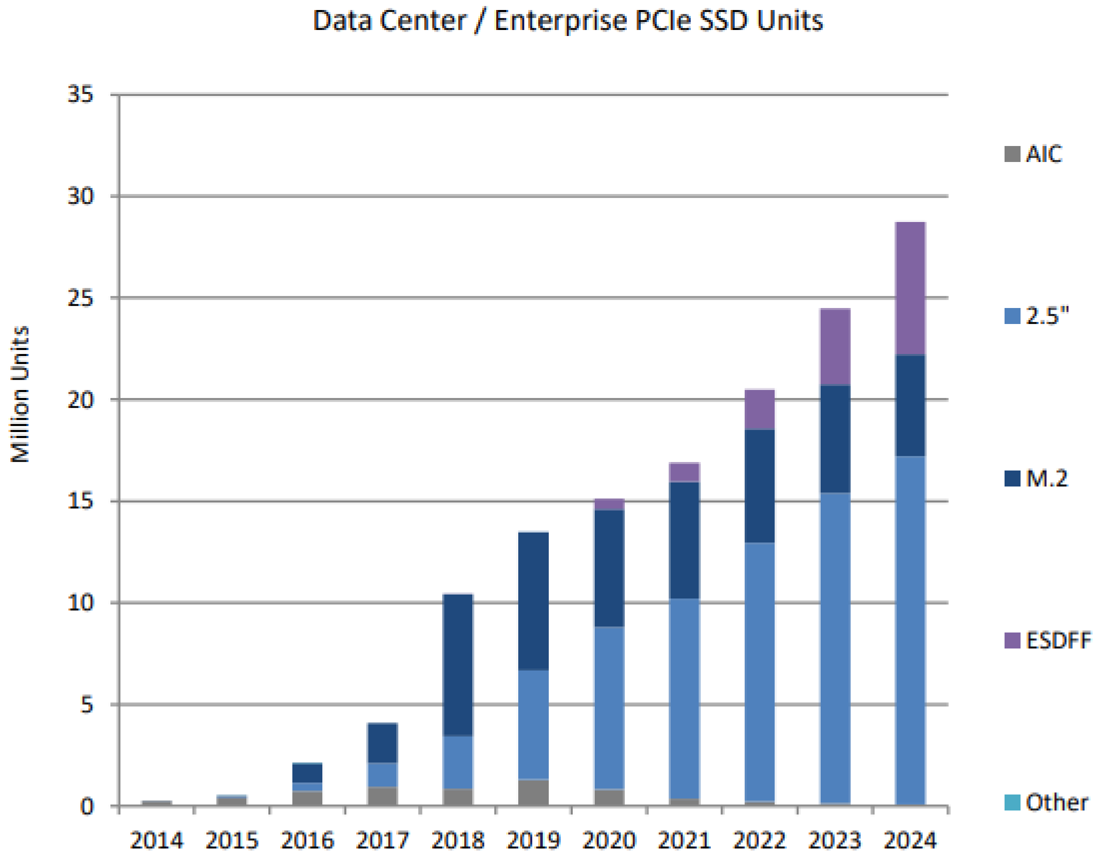
- Flash also brings benefits to the Secondary Storage
  - **Performance**
    - Higher throughput and bandwidth, the ability to move large data sets quickly
  - **Capacity**
    - Increased infrastructure density, reduce the floor space, energy and cooling capacity requirement and improve the overall TCO.
  - **Reliability**
    - No moving parts.

IDC ANALYZE THE FUTURE

https://blogs.idc.com/2019/12/09/will-flash-penetrate-secondary-storage-environments/

Enterprise SSD Form Factor and Unit Trend

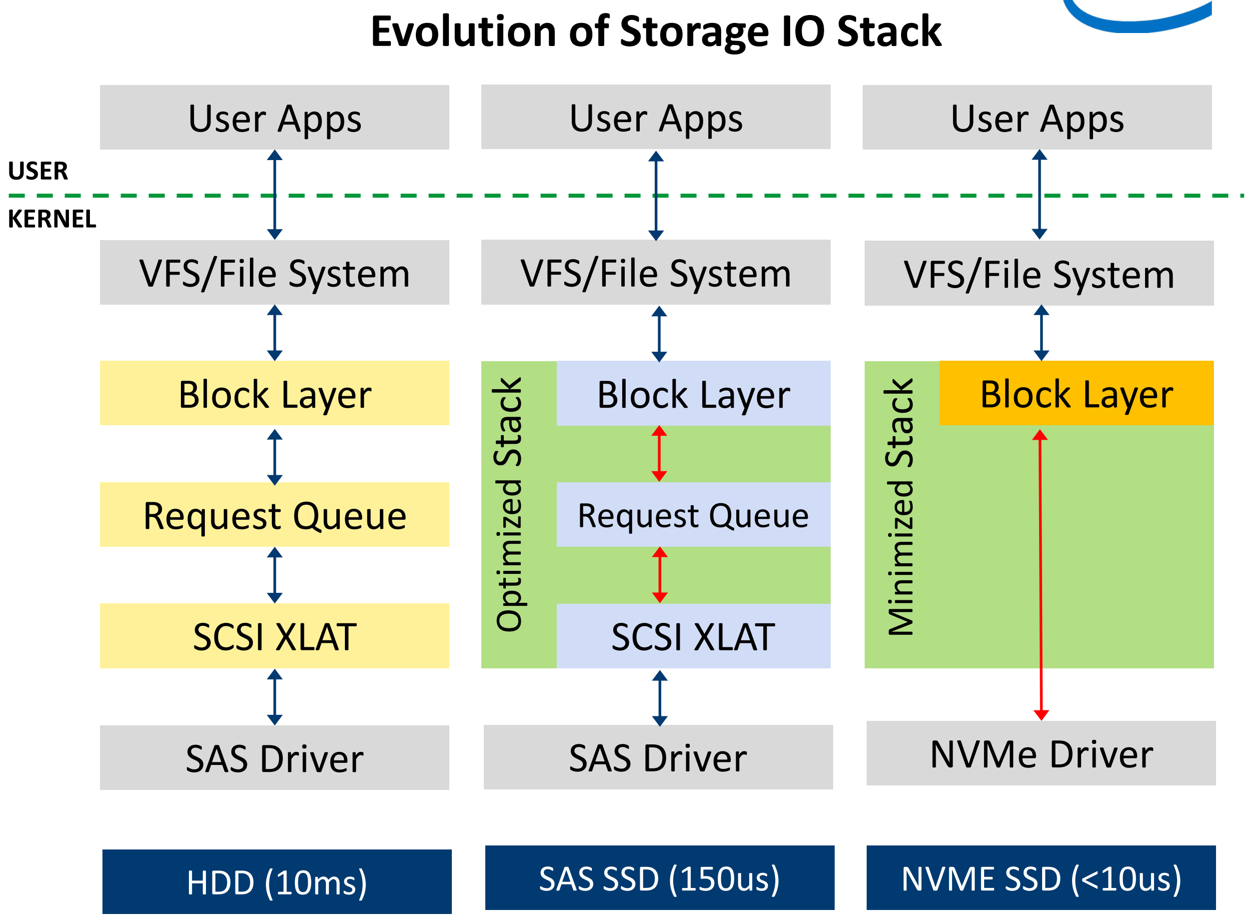




https://pcisig.com/sites/default/files/files/PCI-SIG_Webinar_EDSFF_FINAL.pdf

NVMe Design Principle

- Optimized protocol for NAND flash.
- NVMe bypasses unneeded layers.
- Direct connection to CPU's PCIe lanes.
- Dramatically reducing latency and increasing bandwidth.
- Scales with number of PCIe lanes
- No HBA required.

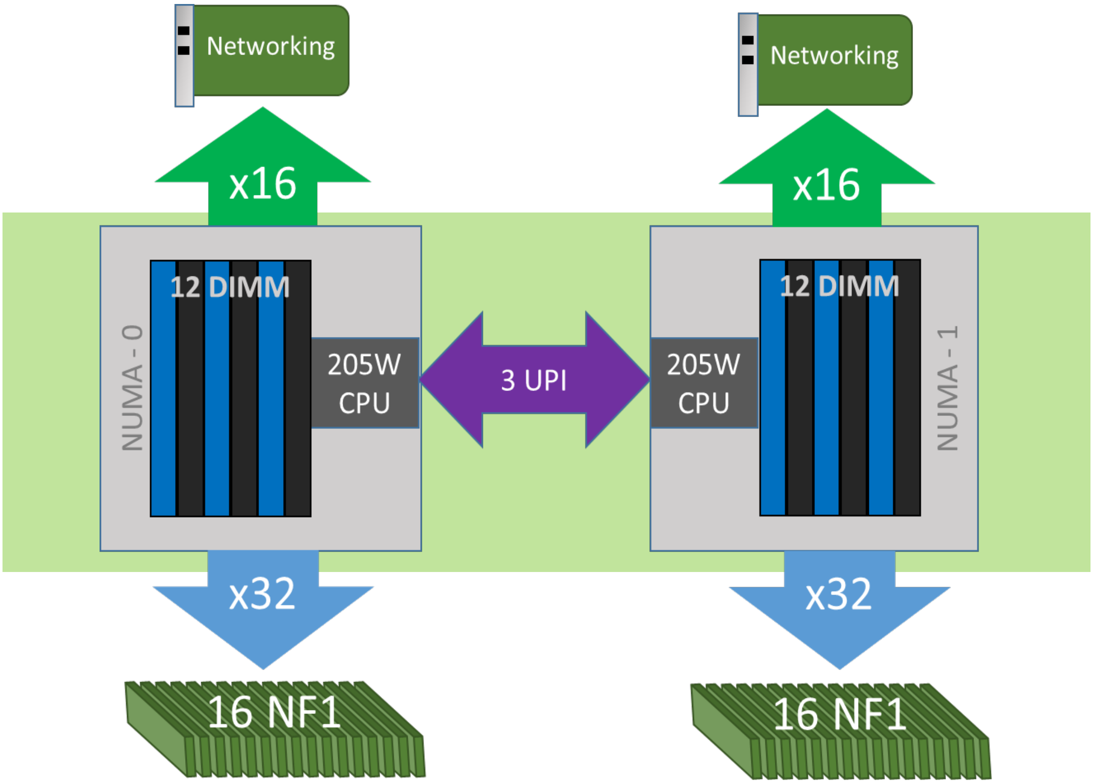


Source : https://www.virtual.com/blog/i-o-i-o-its-nvme-i-go/

More is Better

New CPUs are helping NVME

- More PCIe Lanes
- PCIe Gen 4 and above



X11 1U32 NVMe Optimized Storage Family

Petascale NVMe Solution with Unprecedented Density and Performance

SSG-1029P-NES32R

32 x EDSFF Short (E1.S) NVMe SSD

Super Storage

1U

SSG-1029P-NEL32

32 x EDSFF Long (E1.L) NVMe SSD

SYS-1029P-N32R

32 x U.2 NVMe SSD

SSG-136R-NE32JBF & SSG-136R-N32JBF

32 x E1.L & 32 x U.2 NVMe SSD JBOF

Super JBOF

1U

1U NVMe Petascale Advantages

Economic

- More capacity and faster, less power and space
- Lower TCO with the best operation efficiency (Thermal and Performance per Watt)

Architecture

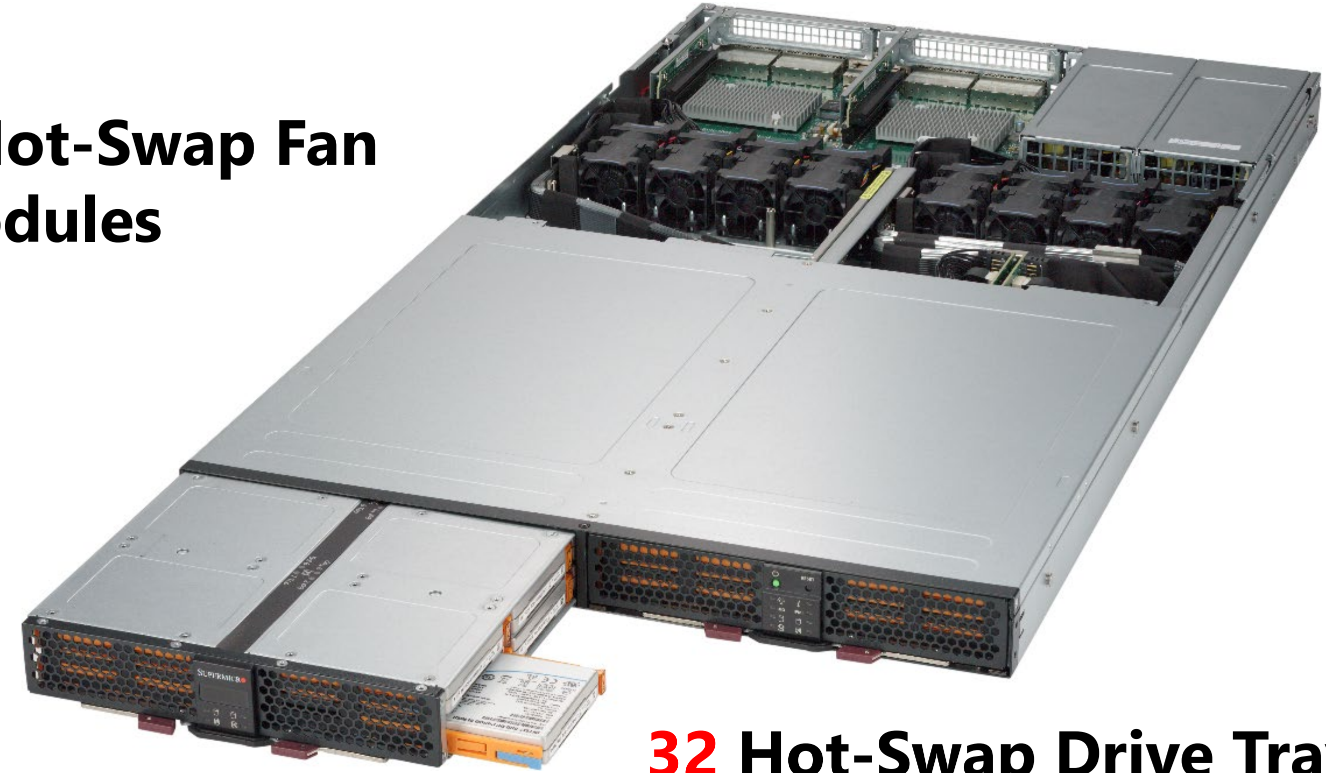
- Highest performance and lowest latency
- NVMe Over Fabric and Disaggregated/Hyperconverged building block

Operation

- Hot-swappable 32 front load NVMe SSD for easy access and service
- Optimized form-factor for heat dissipation and system thermal efficiency

Hot-Swap JBOF Design

8 Hot-Swap Fan Modules

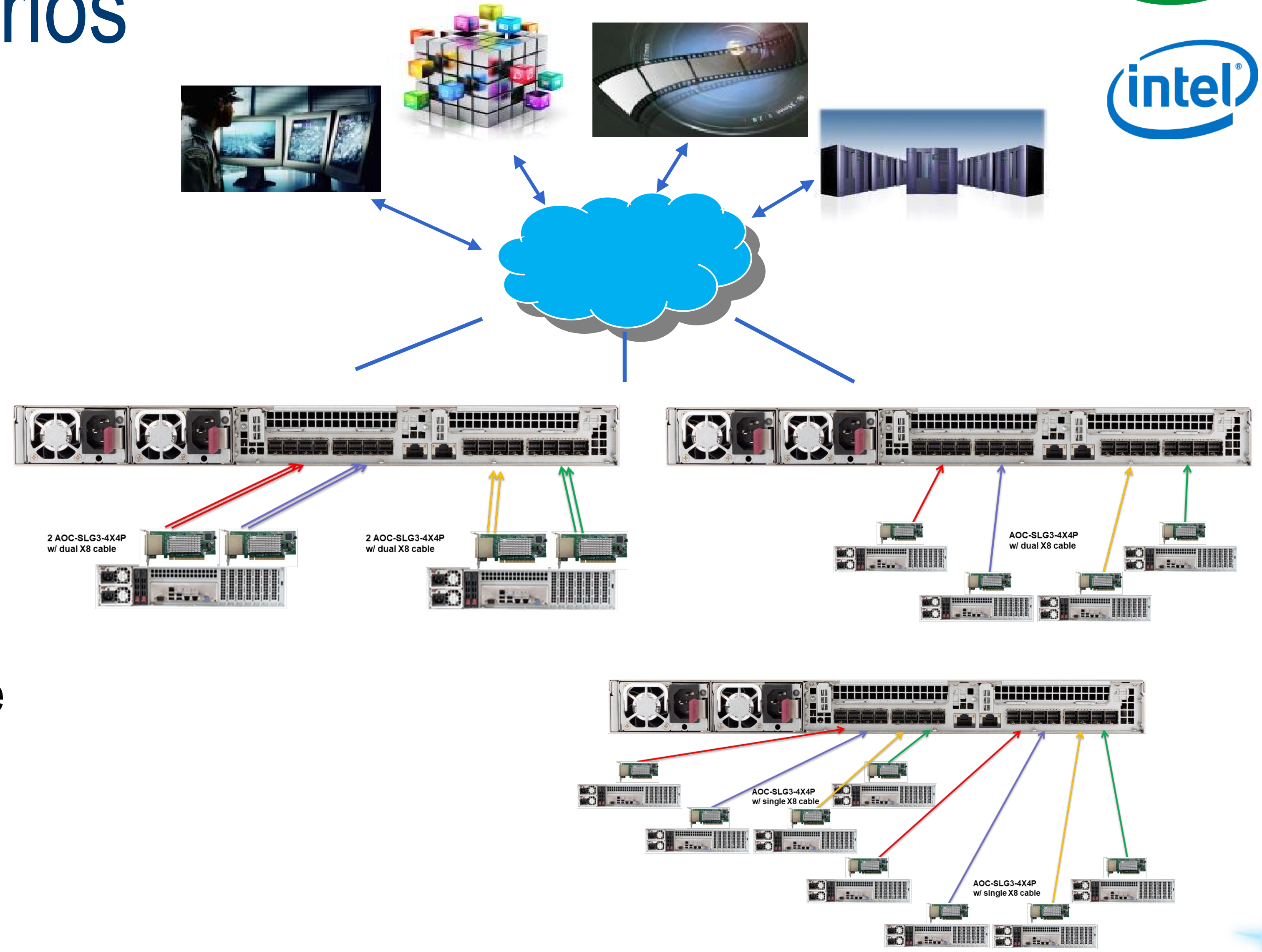


2 Hot-Swap Redundant Power Supplies

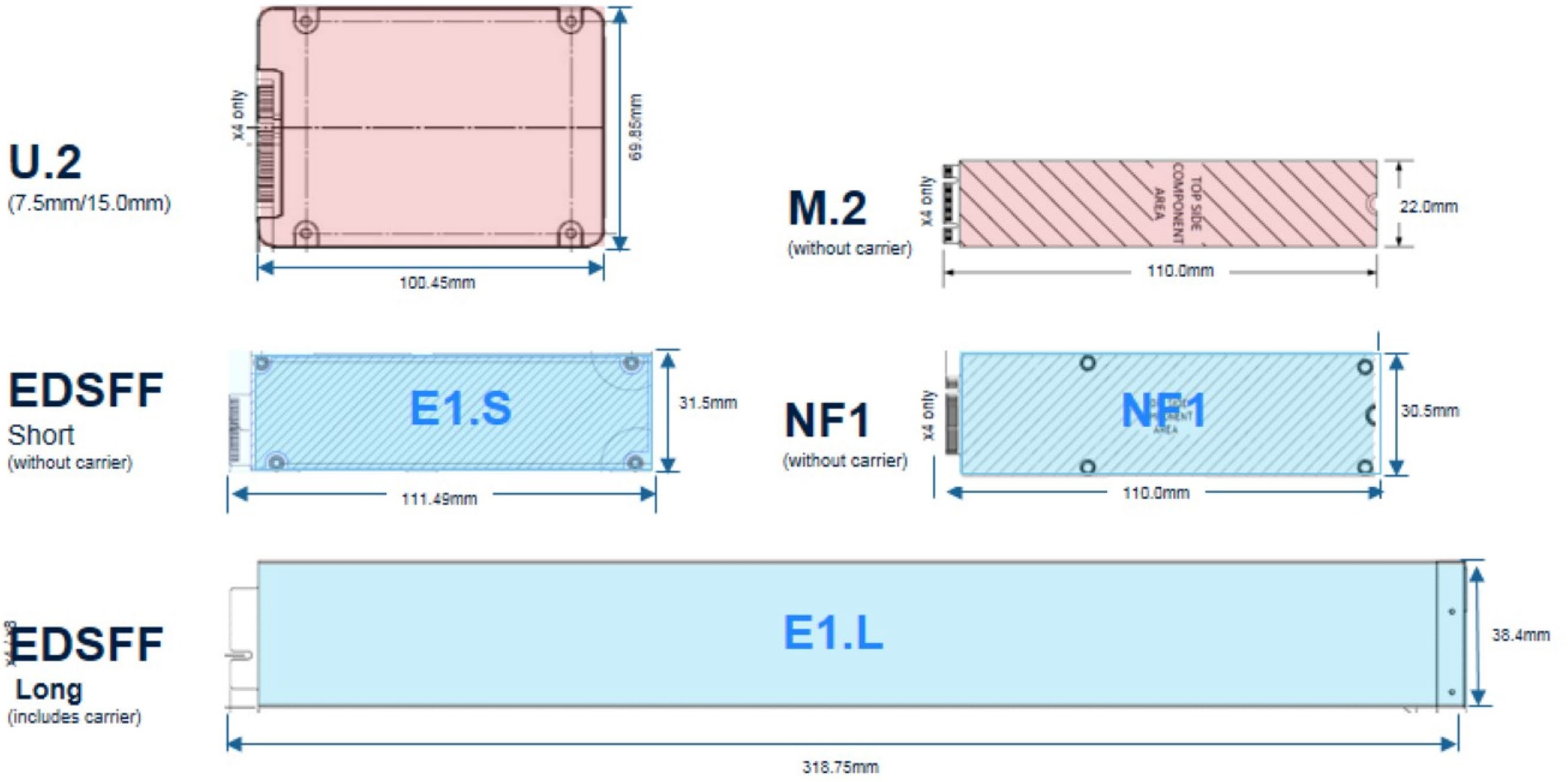
32 Hot-Swap Drive Trays

Application Scenarios

- High capacity storage requirements
  - High Throughput Ingest
  - High Density Hot Storage
  - HPC /Data Analytics
  - Media/Video Streaming
  - Content Delivery Network (CDN)
  - Big Data Top of Rack Storage



NVMe Form Factor Comparison



What is EDSFF*?

1 A group of **15** companies working together[1]

2 Industry standard connector and form factor **optimized** for NVMe*

3 Built for increased operational efficiency and dense storage

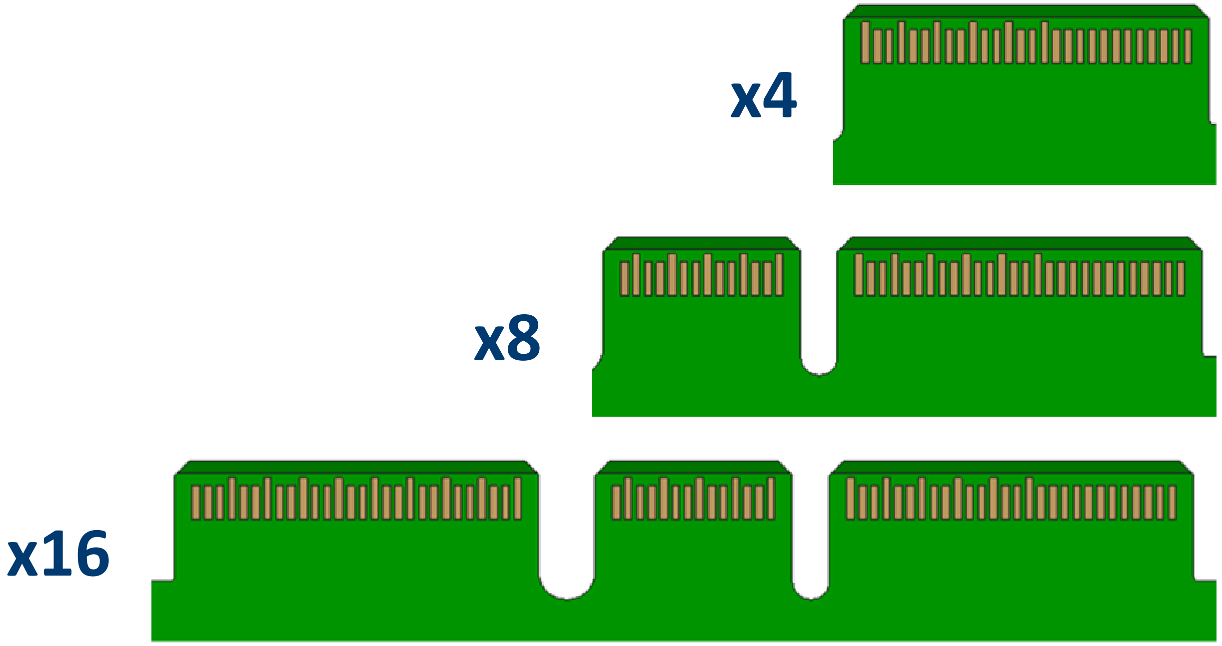
Intel® SSDs with EDSFF* “ruler”

E1.L 18mm

E1.L 9.5mm

E1.S

ALL EDSFF* SSDs support the same:



PCIe* 4.0 and 5.0 ready[7]

1 Connector

Drives high volumes

2 Pinout

Allows interoperability, simplifies backplane design

3 Base Features

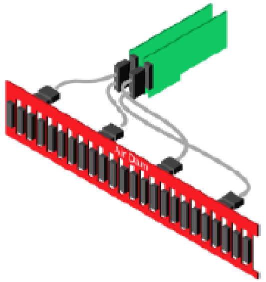
But differentiated by segment and use case

Systems Designed with Flexibility for Storage and Beyond

*Other names and brands may be claimed as the property of others.
Source – Amphenol ICC*. https://www.amphenol-icc.com/connect/cool-edge-high-speed-high-power-card-edge.html. Additional detail: https://EDSFFspec.org/introduction-to-EDSFF/

EDSFF vs. 2.5” Storage Chassis Implementation

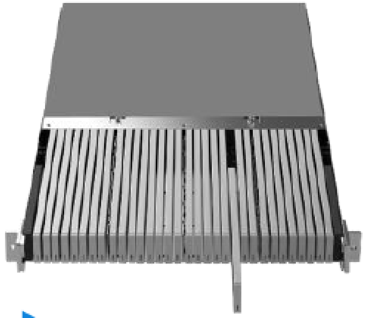
2.5” FORM FACTOR

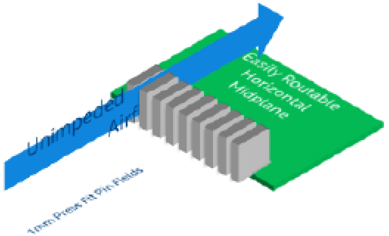


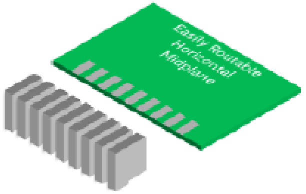
- Backplane requires cut outs to optimize thermals
- Cables add cost and complicate installation, thermals
- LED controller adds failure point

- Drive cages add cost, failure points

RULER FORM FACTOR







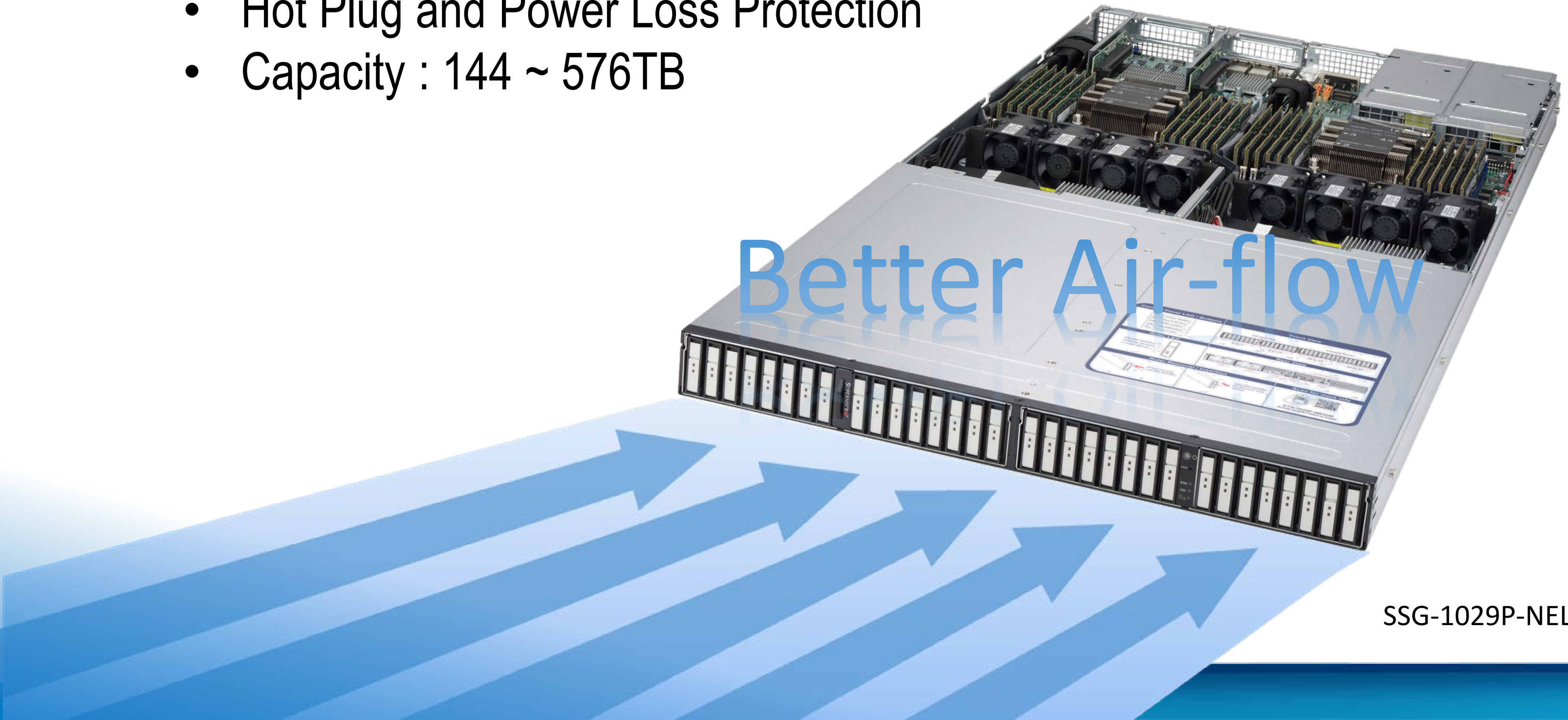
- Eliminate the backplane
- Simplified thermal implementation

- No add in cards required
- No cables to SSDs
- Geographic drive mapping for simplified drive management

Less complicated chassis
Reduced component cost per SSD
Simple hot swap with high density capabilities

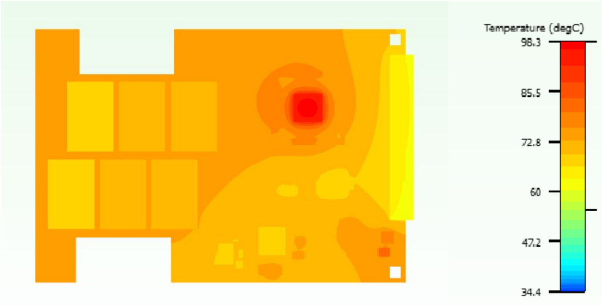
High Efficiency by Design

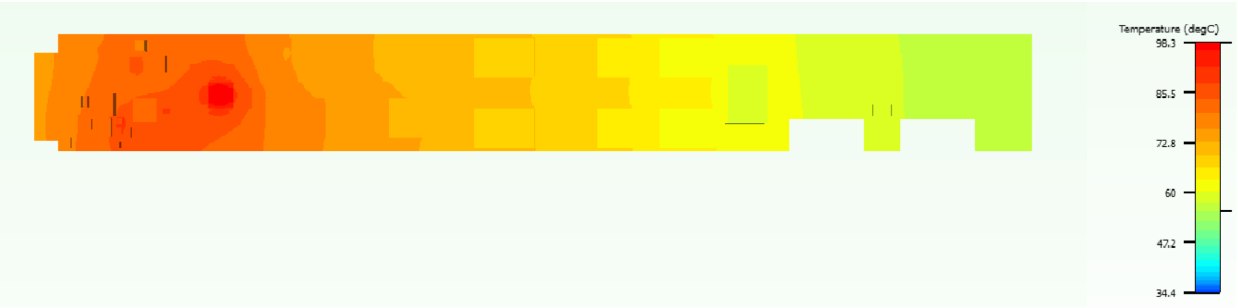
- **Better Air-flow = Better Power Efficiency**
  - Front Loading bays with increased Air-flow
  - High Performance - Up to 10 million IOPS in 1U
  - Hot Plug and Power Loss Protection
  - Capacity : 144 ~ 576TB



SSG-1029P-NEL32R

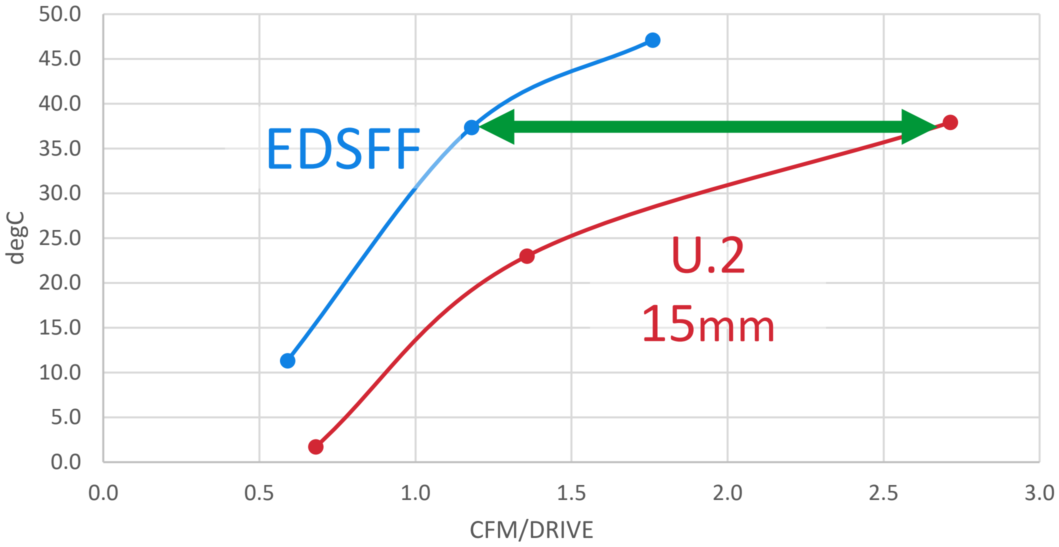
Advantage. Thermal efficiency.



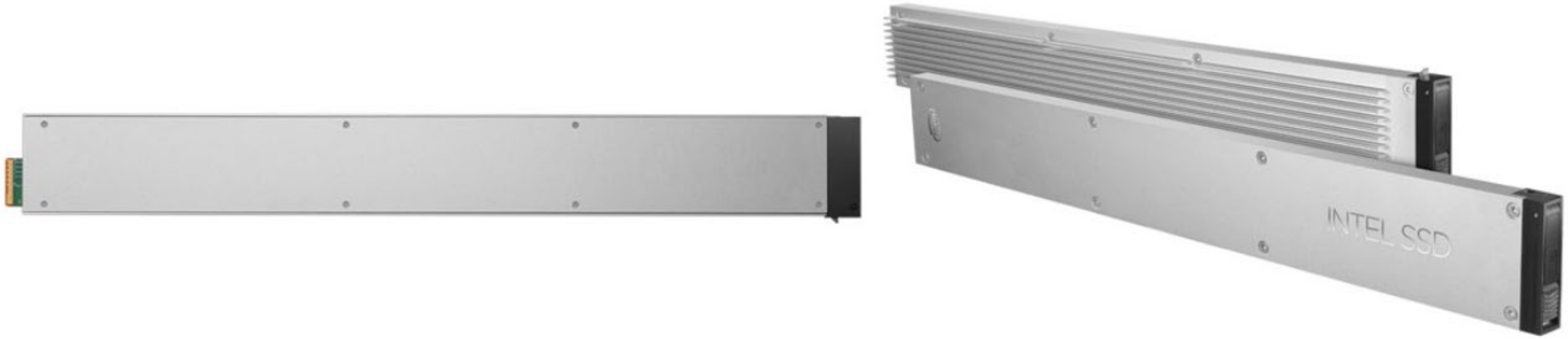


Thermal efficiency

Up to **55%** less airflow[4] vs U.2 15mm



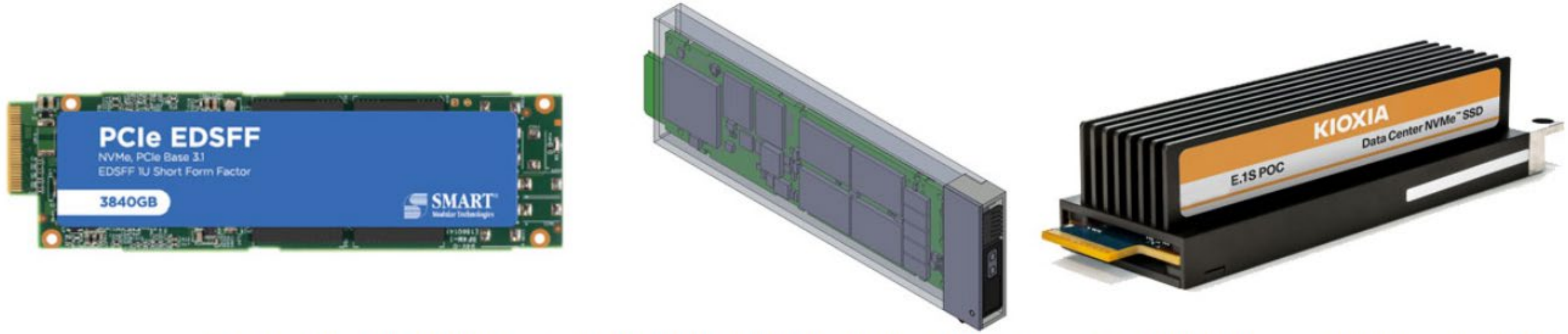
EDSFF Long (E1.L) Form Factors



Illustrations left to right: E1.L 9.5mm (courtesy of Intel); E1.L 18mm (courtesy of Intel)

Type	Width	Length	Thickness
E1.L 9.5mm	up to 25W - 38.4mm	318.75mm	9.5mm
E1.L 18mm	up to 40W - 38.4mm	318.75mm	18mm

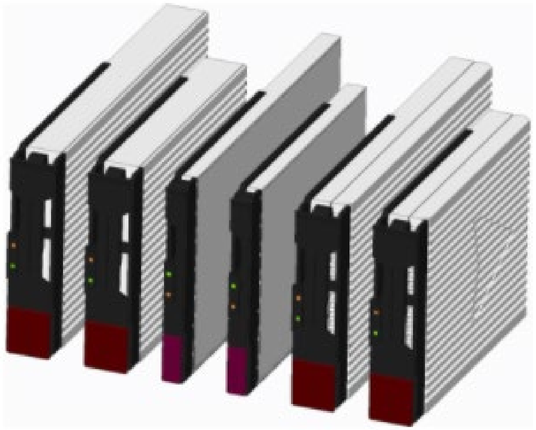
EDSFF Short (E1.S) Form Factors



Illustrations left to right: E1.S 5.9mm (courtesy of SMART Modular Systems); E1.S Symmetric Enclosure (courtesy of Intel); E1.S Asymmetric Enclosure (courtesy of KIOXIA)

Type	Width	Length	Thickness
E1.S 5.9mm	31.5mm	111.49mm	5.9mm
E1.S 8mm heat spreader	31.5mm	111.49mm	8.01mm
E1.S Symmetric Enclosure	33.75mm	118.75mm	9.5mm
E1.S Asymmetric Enclosure	33.75mm	118.75mm	15mm
E1.S Asymmetric Enclosure	33.75mm	118.75mm	25mm

EDSFF 3 (E3) Form Factors



Illustrations left to right: various E.3 configurations (courtesy of Intel)

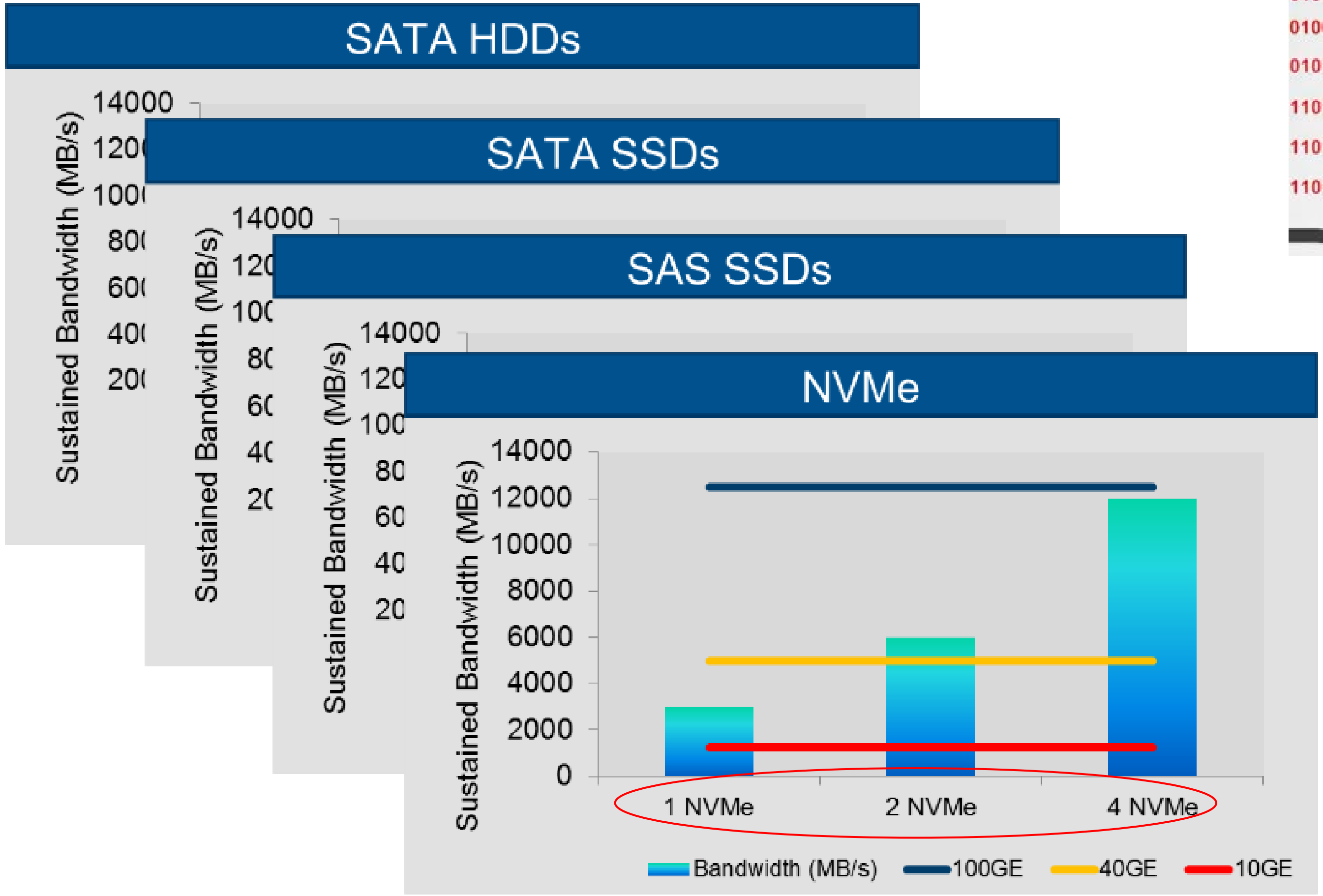
Type	Width	Length	Thickness
E3.S 7.5mm	76mm	104.9mm	7.5mm thickness
E3.S 16.8mm	76mm	104.9mm	16.8mm
E3.L 7.5mm	76mm	142.2mm	7.5mm
E3.L 18mm	76mm	142.2mm	16.8mm

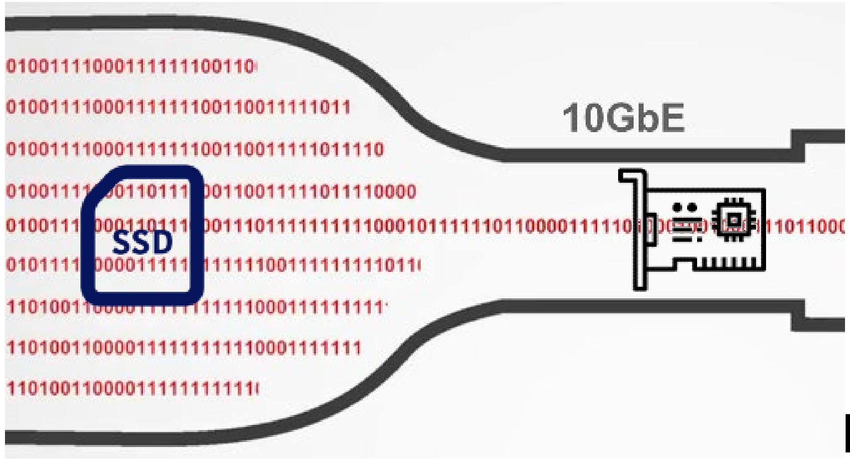
QLC vs TLC

- QLC has 4-bits per cell, while TLC has 3-bits per cell.
  - 33% capacity improvement
- QLC costs less than TLC
  - Closing the price gap between SSDs and HDDs
- QLC EDSFF using 16K block writes
- QLC has slower write performance, but same read performance.
- QLC EDSFF endurance is <0.5 DWPD
  - 8TB drive * 1 DWPD = 8TB per day
  - 16TB drive * .5 DWPD = 8TB per day

QLC is best for read intensive applications

Why NVMe-oF?

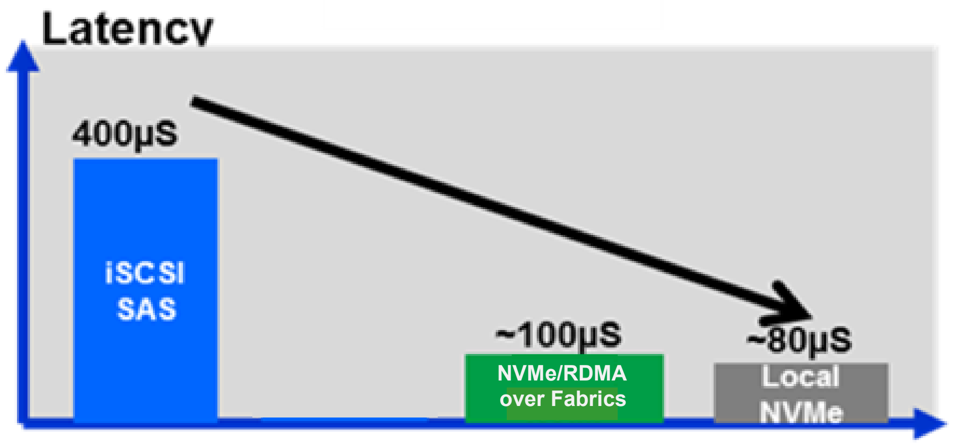


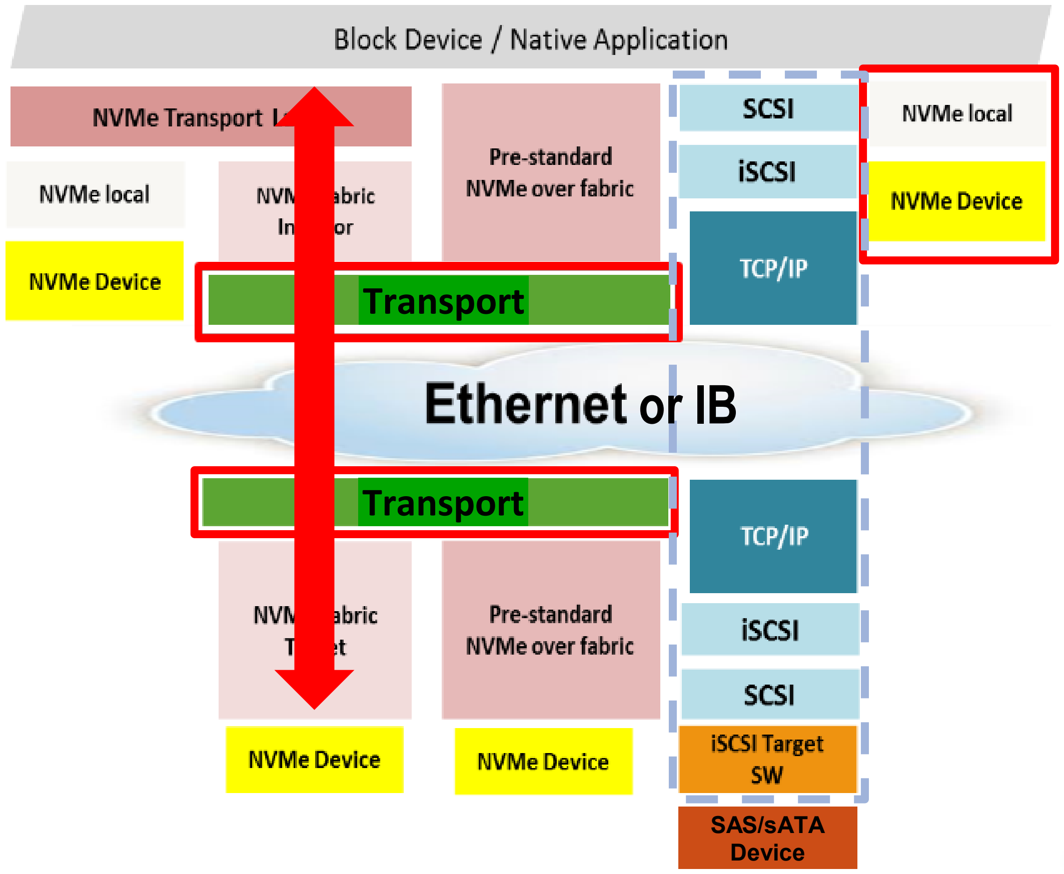


SSDs move the Bottleneck from the Disk to the Network

How Does NVMe-oF Maintain NVMe Like Performance?

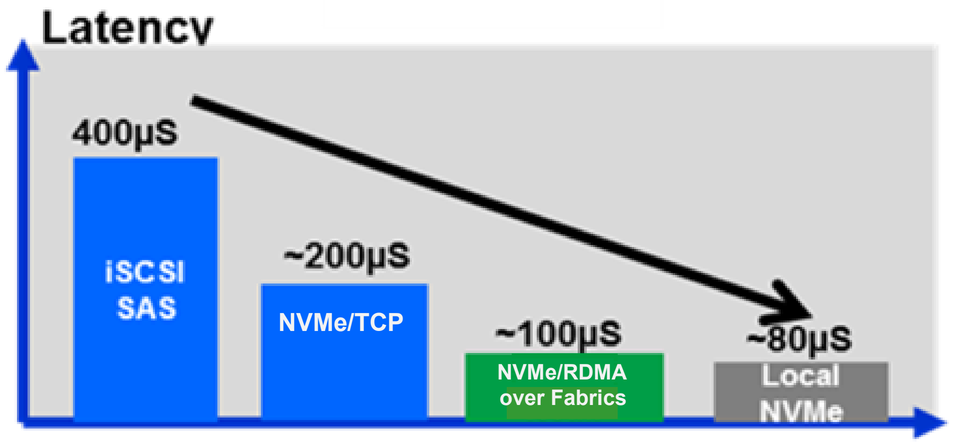
- By extending NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
- Bypassing legacy stacks for performance
- First products all used RDMA
- Performance is impressive

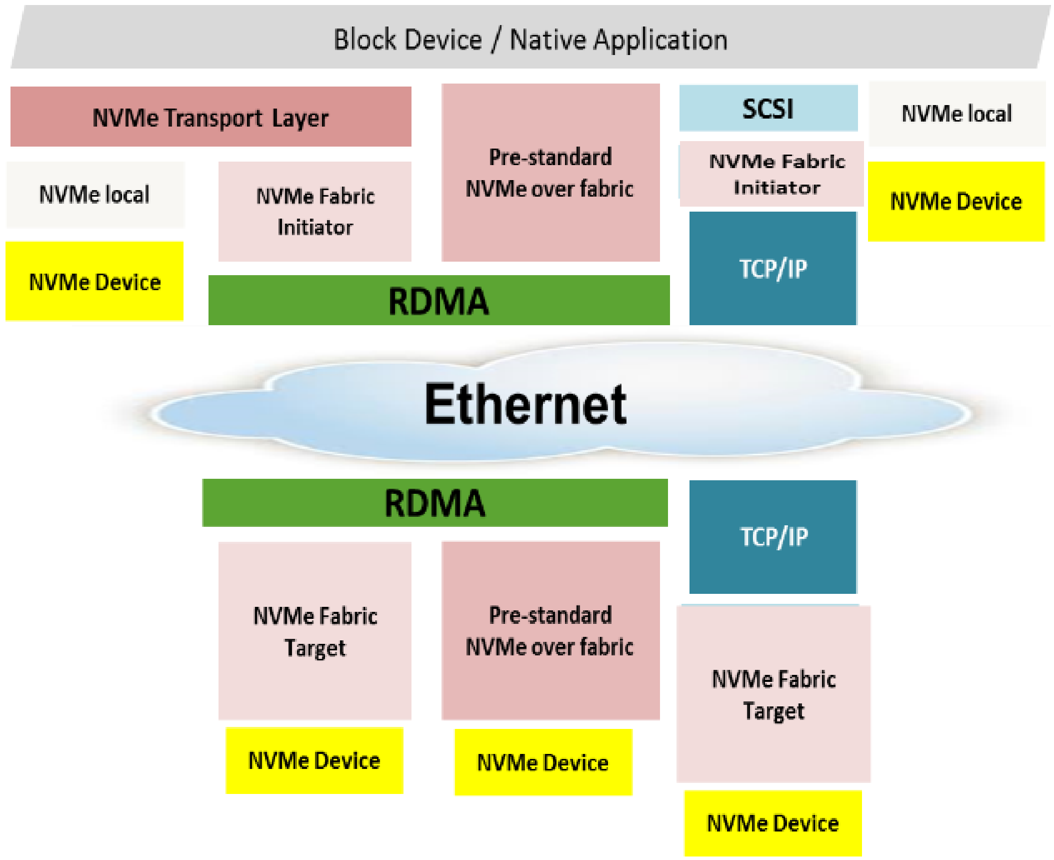




How Does NVMe-oF Maintain NVMe Like Performance?

- By extending NVMe efficiency over a fabric
  - NVMe commands and data structures are transferred end to end
- Bypassing legacy stacks for performance
- First products all used RDMA
- Performance is impressive

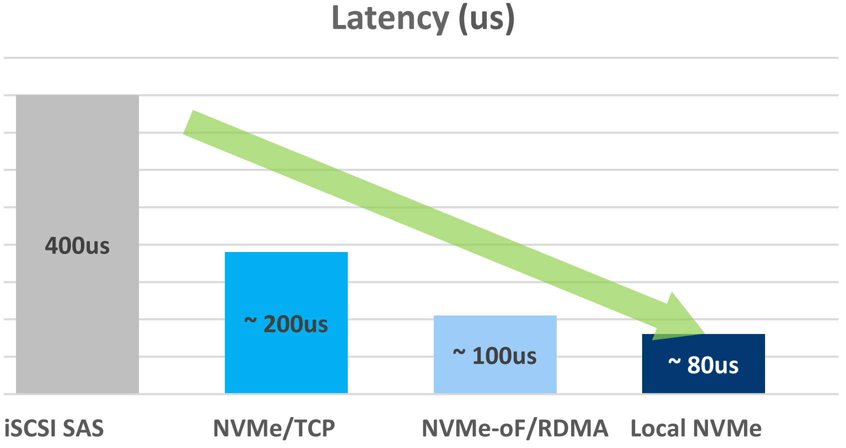


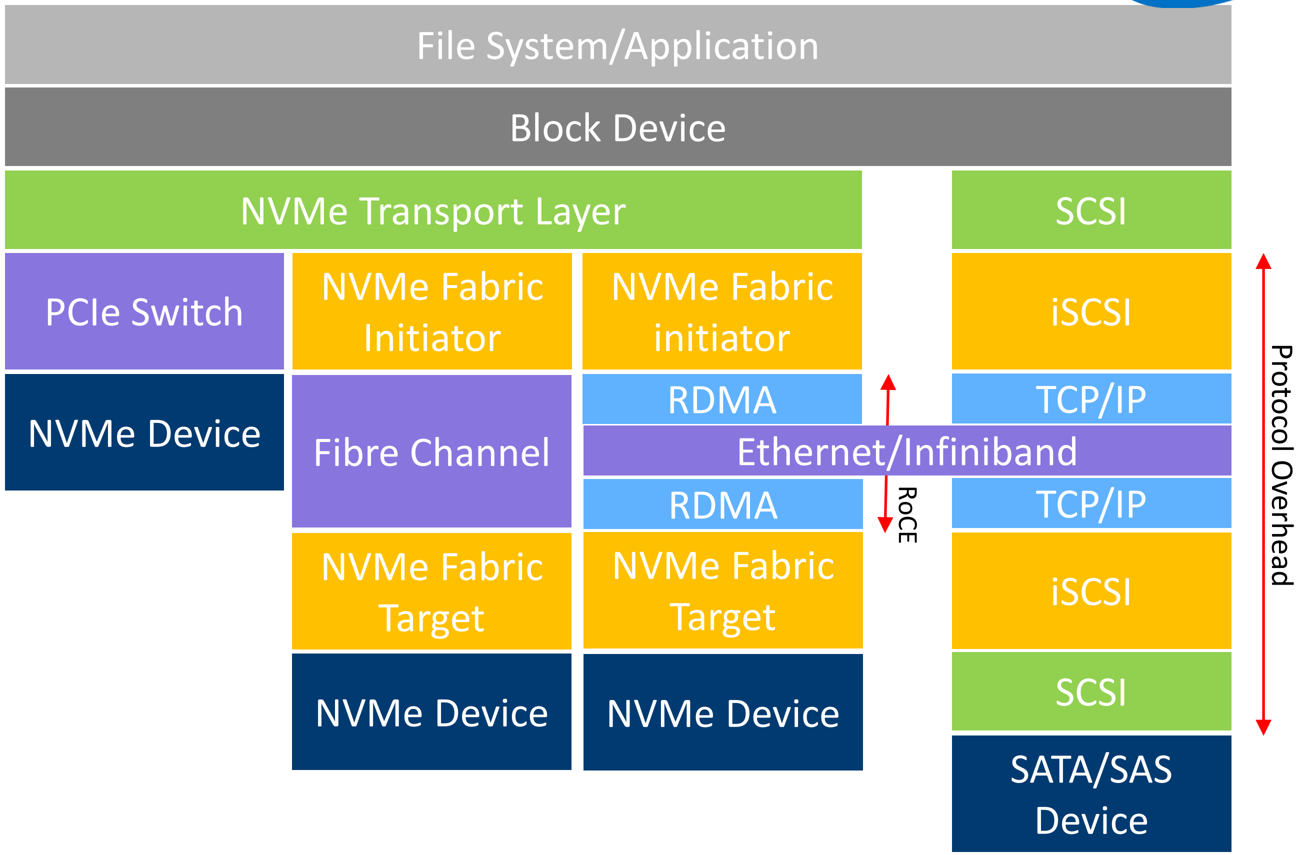


What and Why is NVMe over Fabrics

Number of SSDs to Saturated Network BW

	SATA HDD	SATA SSD	SAS SSD	NVME SSD
10GbE	24	2	1	1
40GbE	100	9	4	2
100GbE	250	24	10	4

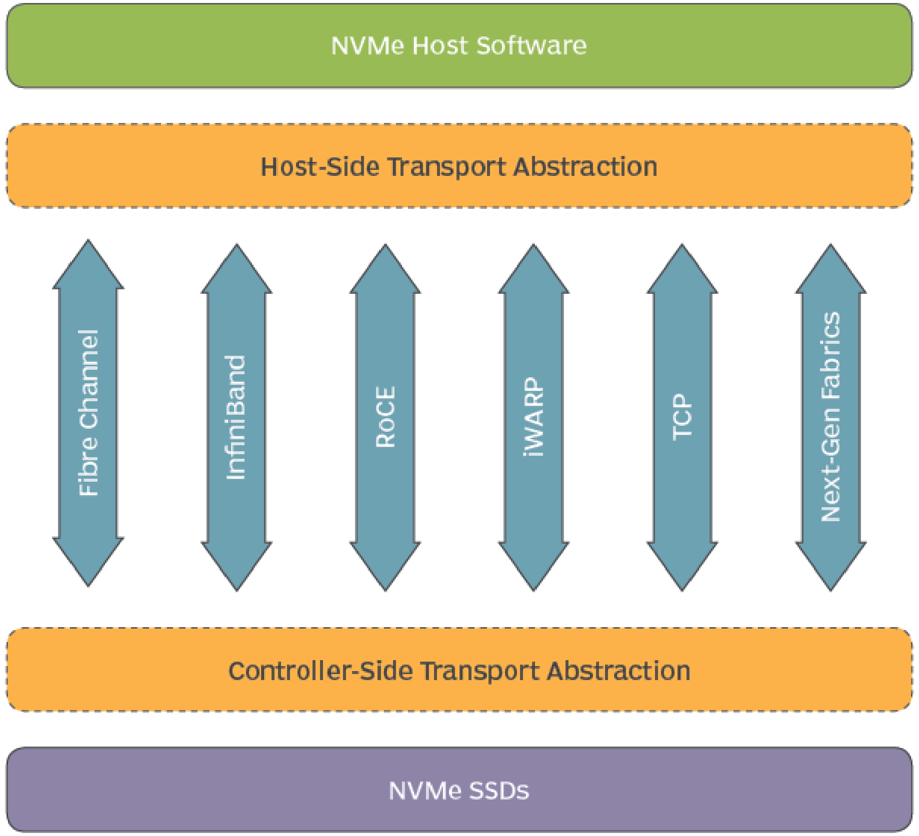




Source : https://www.electronicdesign.com/industrial-automation/article/21805431/nvme-over-fabric-addresses-hyperscale-storage-needs/

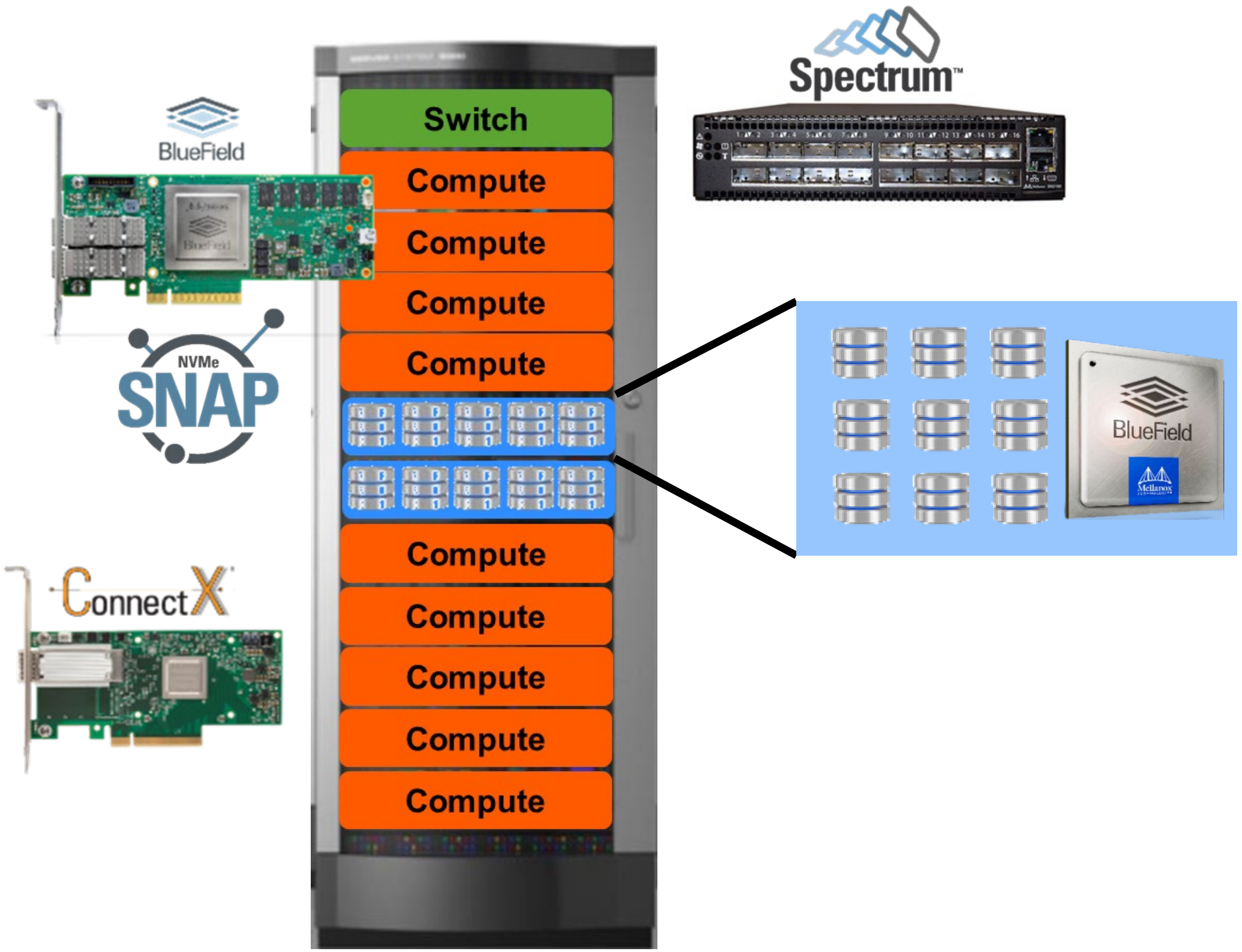
What is NVMe over Fabrics (NVMe-oF)

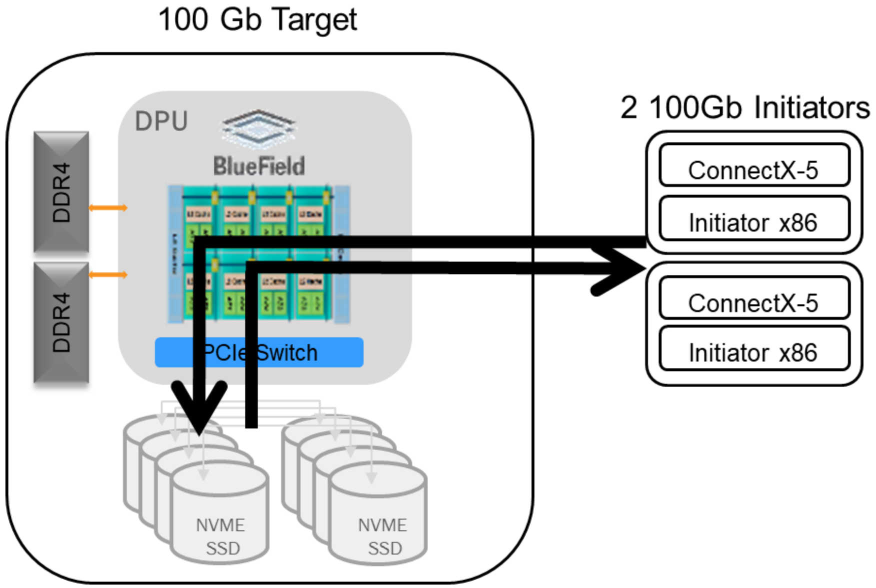
- A protocol interface to NVMe that enable operation over other interconnects (e.g., Ethernet, InfiniBand, Fibre Channel).
- Shares the same base architecture and NVMe Host Software as PCIe.
- Enables NVMe Scale-Out and low latency (<10µS latency) operations on Data Center Fabrics.
- Avoids protocol translation overhead (avoid SCSI)



Source : http://g2minc.com/wp-content/uploads/2018/11/G2M-Research-NVMe-Multi-Vendor-Webinar-on-NVMe-oF-v1-11018.pdf

NVMe-oF Applications - Composable Infrastructure

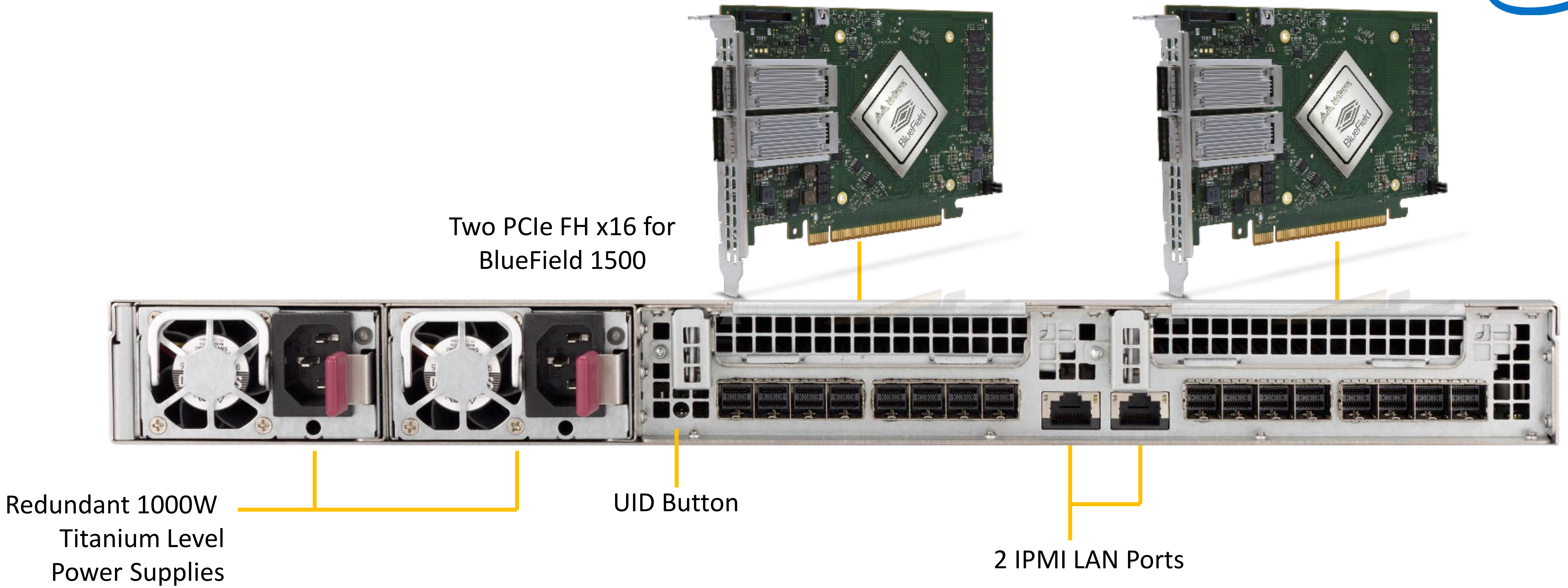




- 8M IOPs, 512B block size
- 5M IOPs, 4K block side
- ~5 usec latency (not including SSD)

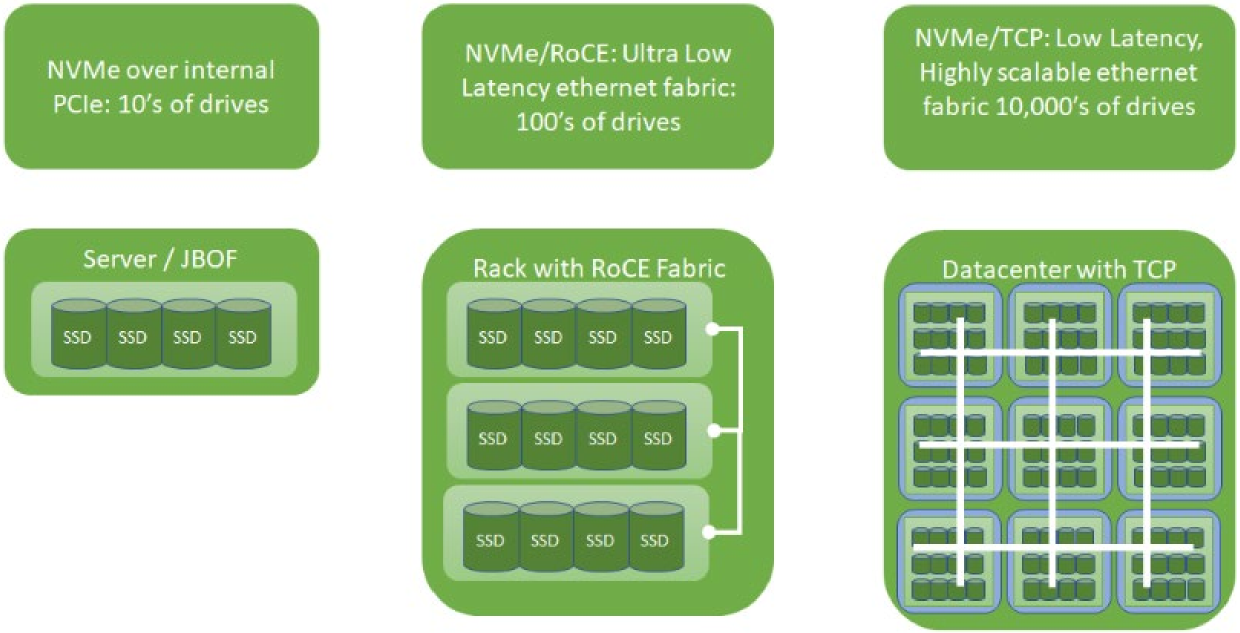
- Nearly local disk performance

NVMeoF JBOF (Rear View)



The Value of Shared Storage and The 'Need for Speed'

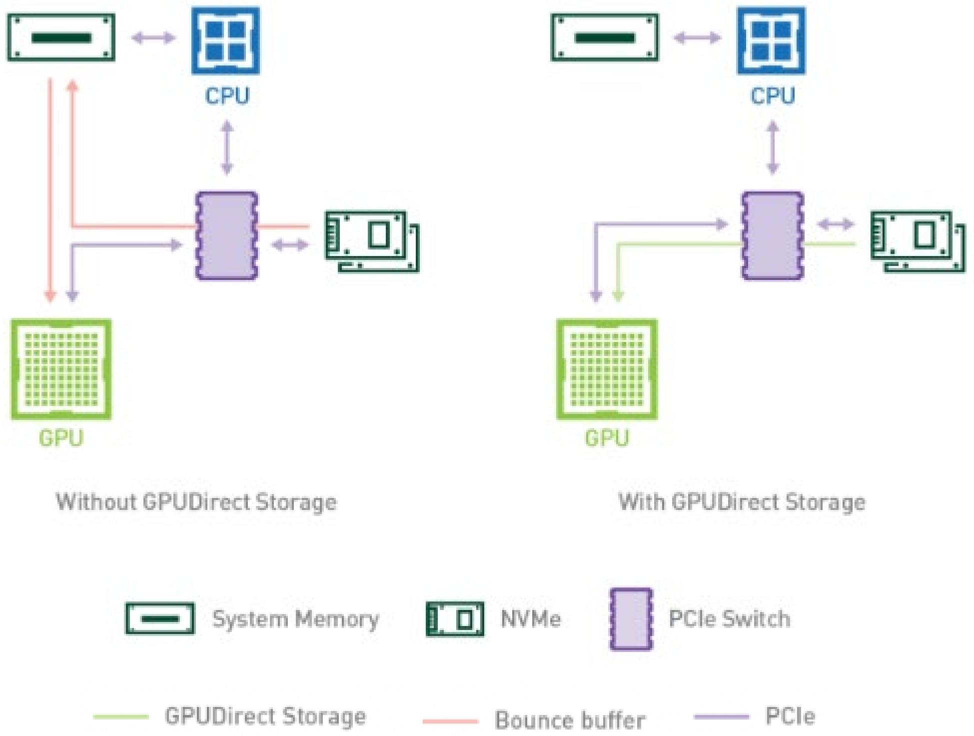
- The cost of data-at-rest is no longer the right metric for storage TCO
  - The value of data is based on how fast it can be accessed and processed
- NVMe over Fabrics increases the velocity of data
  - Faster storage access enables cost reduction through consolidation
  - Faster storage access delivers more value from data
- SSDs are going to become much faster
  - 3D Xpoint Memory, 3D NAND, etc.
  - PMEM, Storage Class Memory, etc



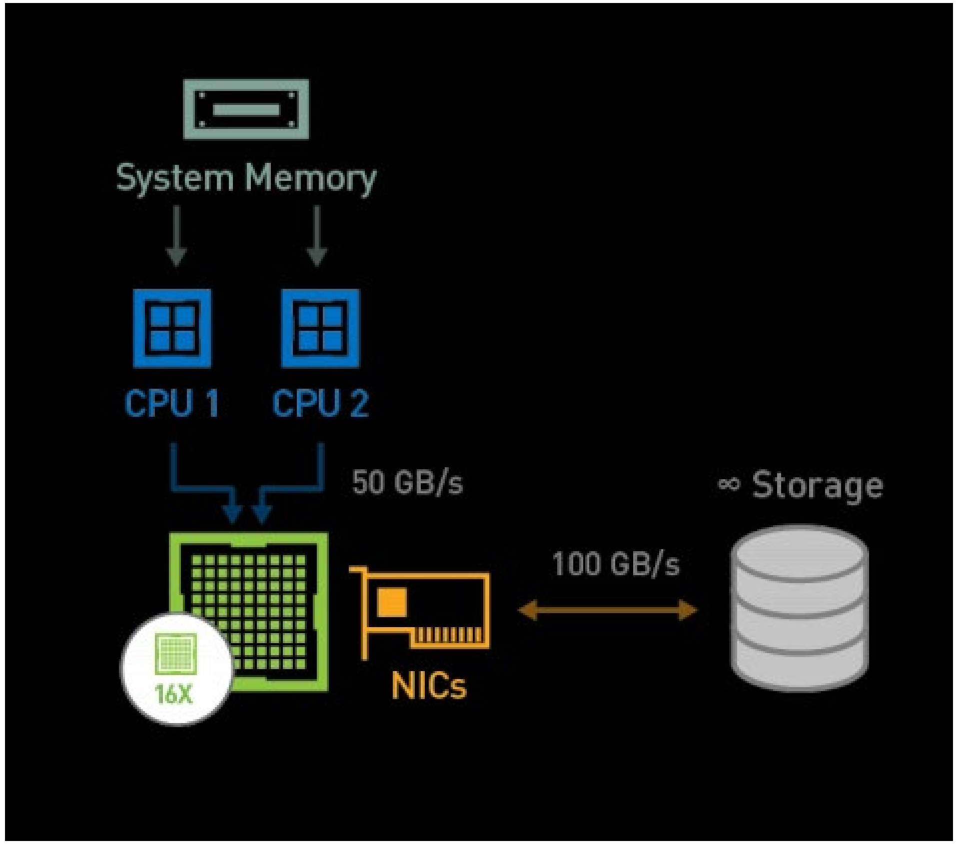
Source : https://www.eetimes.com/nvme-tcp-improves-data-storage/#

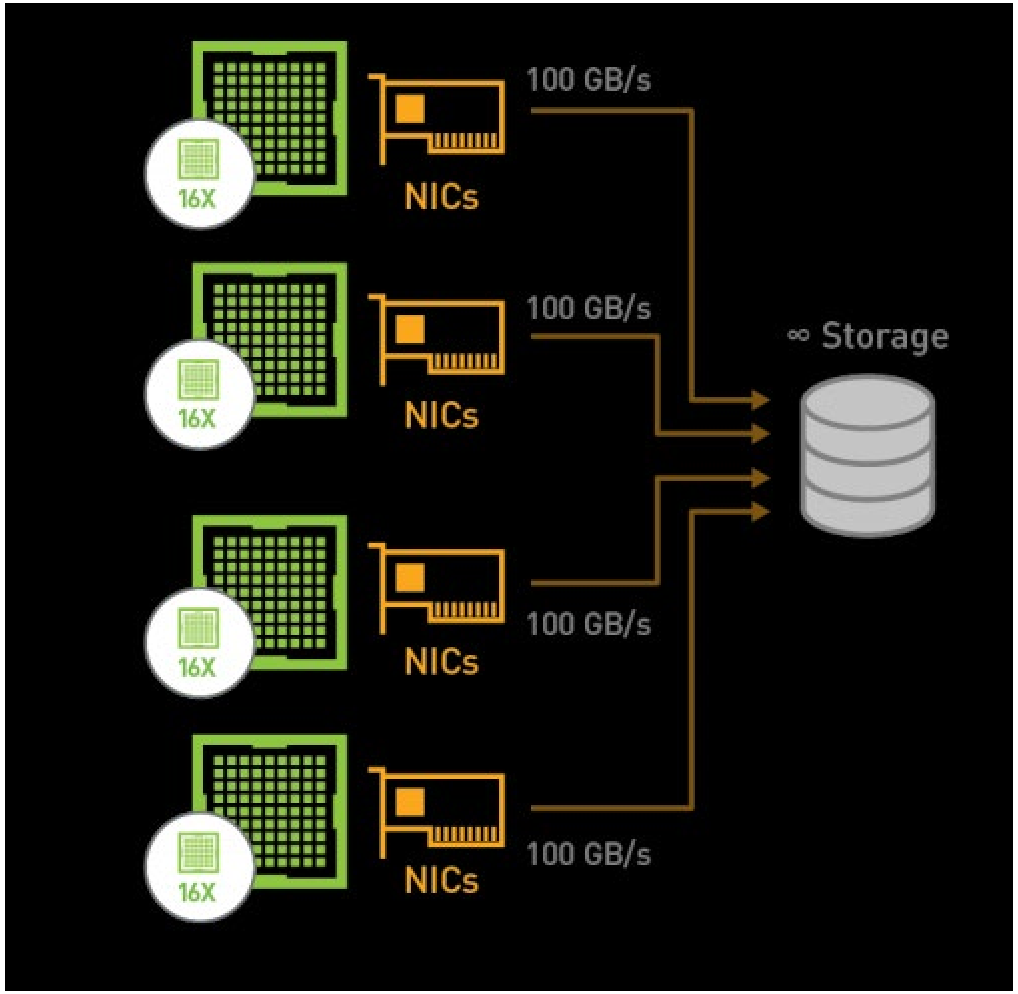
GPUDirect Storage

- Avoid copying through a CPU bounce buffer
- Performance
  - Raw IO bw difference varies by platform , e.g. 2-4X
  - Savings in memory management and utilization can be a force multiplier on top of the
  - Varies by platform
- Broad ecosystem interest, active enabling
- Enabling with broader Linux community
- Coming to a CUDA near you



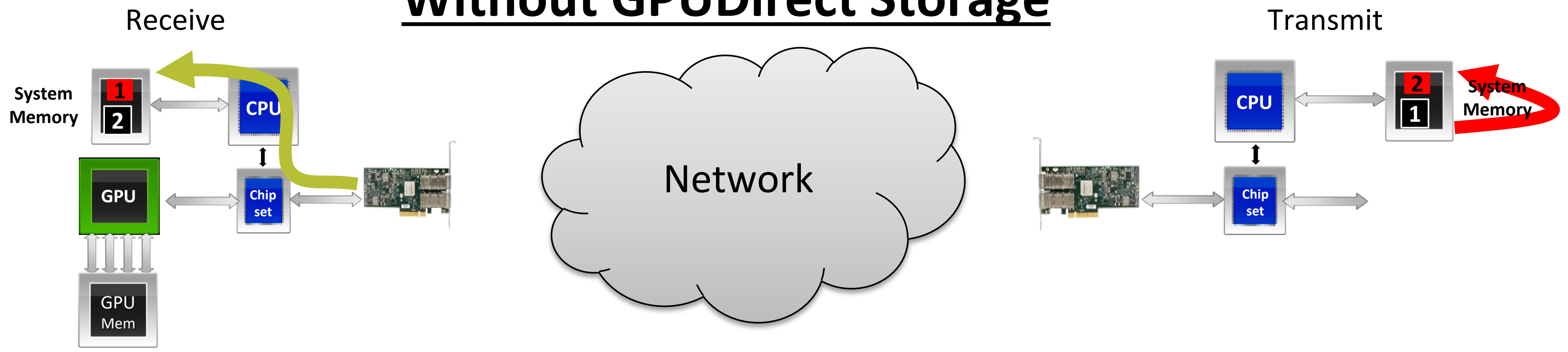
GPUDirect Storage and Cluster



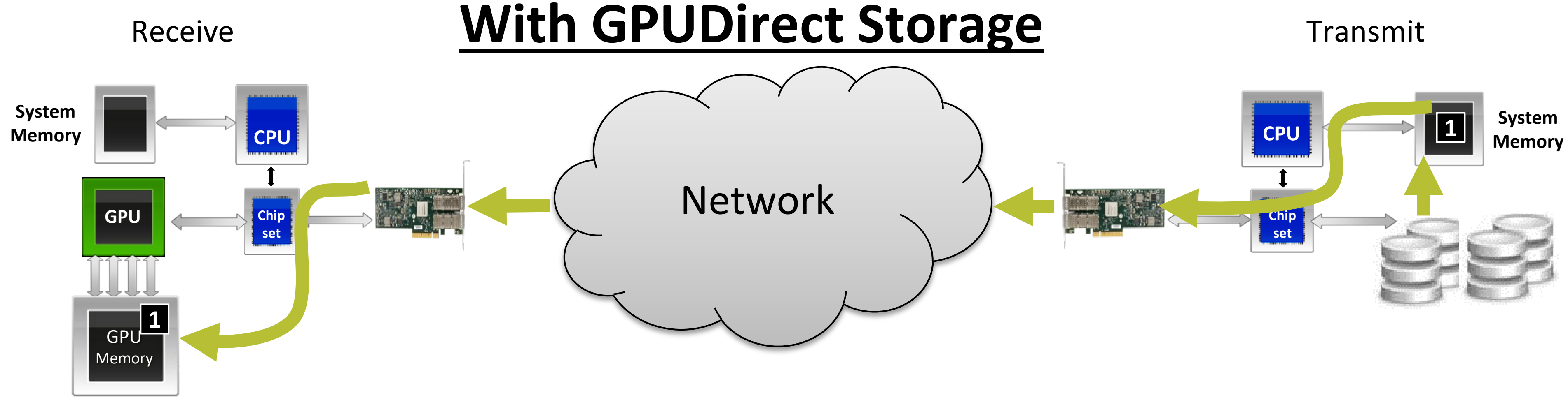


GPUDirect with RDMA

Without GPUDirect Storage



With GPUDirect Storage



NVMe is on the Move with Innovation

- NVMe is growing and changing
- Processors are enabling better NVMe systems
- EDSFF will take over, if we can settle down the spec
- NVMeoF enables low latency transfer of data directly into the drives.
- GPU direct allows access to NVMe drives without the CPU.
- Customers need to know that this is not science fiction.
- Supermicro has products for everything that I have discussed.

DISCLAIMER

Super Micro Computer, Inc. may make changes to specifications and product descriptions at any time, without notice. The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors. Any performance tests and ratings are measured using systems that reflect the approximate performance of Super Micro Computer, Inc. products as measured by those tests. Any differences in software or hardware configuration may affect actual performance, and Super Micro Computer, Inc. does not control the design or implementation of third party benchmarks or websites referenced in this document. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to any changes in product and/or roadmap, component and hardware revision changes, new model and/or product releases, software changes, firmware changes, or the like. Super Micro Computer, Inc. assumes no obligation to update or otherwise correct or revise this information.

SUPER MICRO COMPUTER, INC. MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

SUPER MICRO COMPUTER, INC. SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL SUPER MICRO COMPUTER, INC. BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF SUPER MICRO COMPUTER, Inc. IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

Thank You

www.supermicro.com