

NVMe-oF: Looking Beyond Performance Hero Numbers

Live Webcast

March 25, 2021

10:00 am PT / 1:00 pm ET

Today's Presenters



Alex McDonald
Independent Consultant
Vice Chair SNIA NSF



Erik Smith
Distinguished Member of
Technical Staff
Dell Technologies



Nishant Lodha
Director, Emerging
Technologies
Marvell Semiconductor



Rob Davis
VP of Storage Networking
NVIDIA

SNIA-at-a-Glance



185
industry leading
organizations



2,000
active contributing
members



50,000
IT end users & storage
pros worldwide

Learn more: **snia.org/technical**

 **@SNIA**

Ethernet, Fibre Channel, InfiniBand®

iSCSI, NVMe-oF™, NFS, SMB

Virtualized, HCI, Software-defined Storage

Storage Protocols (block, file, object)

Securing Data

Technologies We Cover

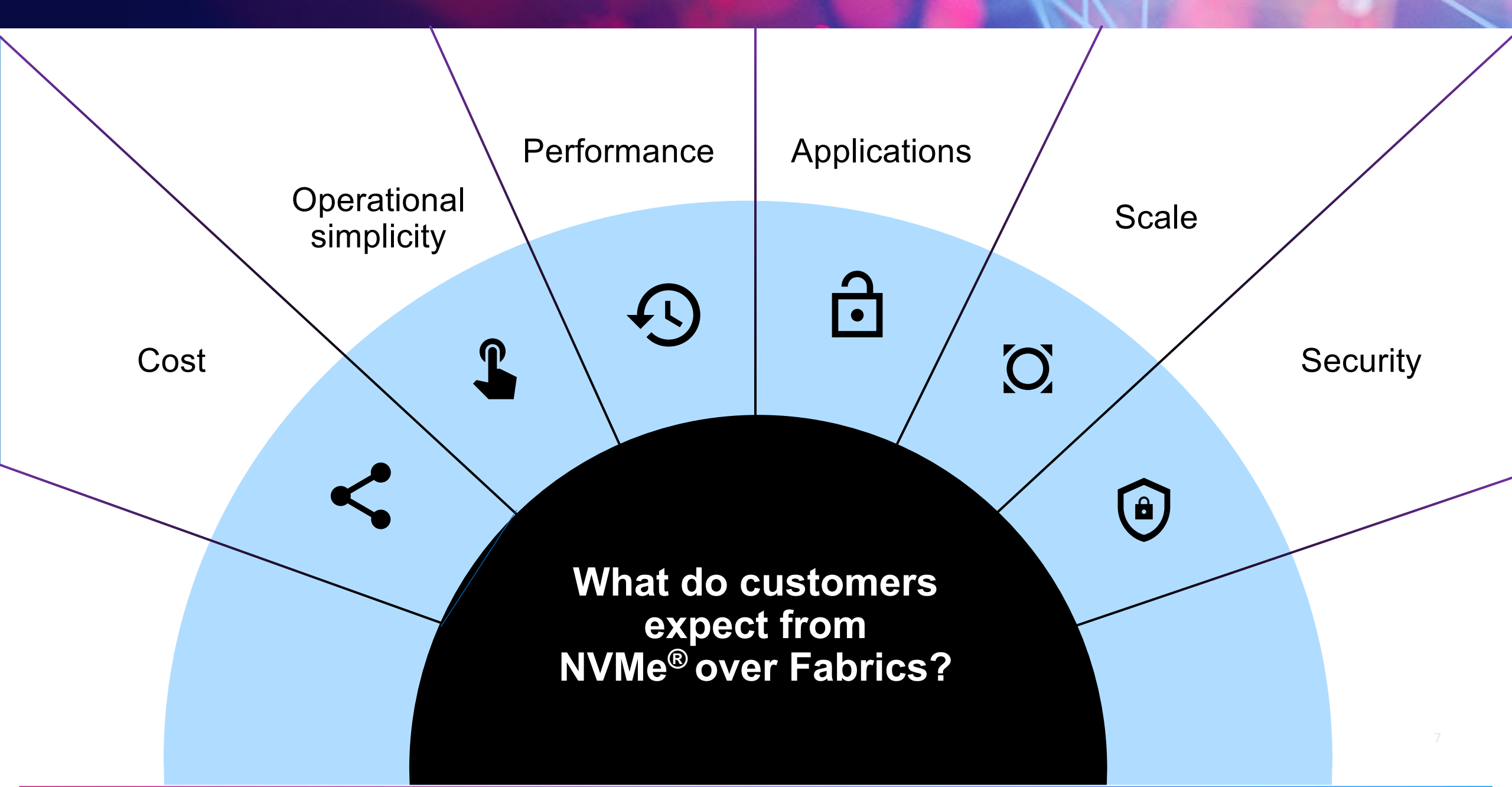
SNIA Legal Notice

- The material contained in this presentation is copyrighted by the SNIA unless otherwise noted.
- Member companies and individual members may use this material in presentations and literature under the following conditions:
 - Any slide or slides used must be reproduced in their entirety without modification
 - The SNIA must be acknowledged as the source of any material used in the body of any document containing material from these presentations.
- This presentation is a project of the SNIA.
- Neither the author nor the presenter is an attorney and nothing in this presentation is intended to be, or should be construed as legal advice or an opinion of counsel. If you need legal advice or a legal opinion please contact your attorney.
- The information presented herein represents the author's personal opinion and current understanding of the relevant issues involved. The author, the presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.

NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.

Defining Hero Numbers

- Hero numbers are:
 - performance metrics that are captured at the extreme ends of the support envelop in order to highlight what is **possible**.
 - usually focus on Latency and IOPS/BW
 - typically captured under ideal conditions (e.g., non-oversubscribed)
- Hero numbers do not:
 - directly correlate to a performance benefit many end-users will experience with their workloads
 - capture the infrastructure ecosystem implications of using one protocol versus another. This is not only about CapEx, OpEx and Security... You need to consider the operating environment as well as the end-devices/appliances and their support for the protocol.

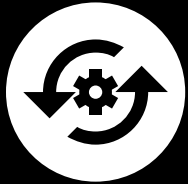


Future of Storage Fabrics

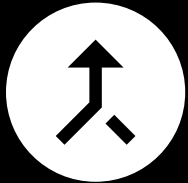
Transition from one **heterogeneous** world to another



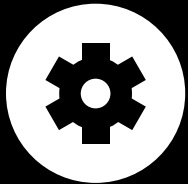
Fibre Channel connectivity
Purpose-build for business-critical apps



iSCSI connectivity
Standard Ethernet-based infrastructure



FCoE connectivity
Blades and converged



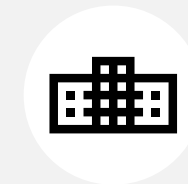
Hyper converged systems
vSAN, AzureStack on standard networking



**Concurrent FC-NVMe
and FCP**

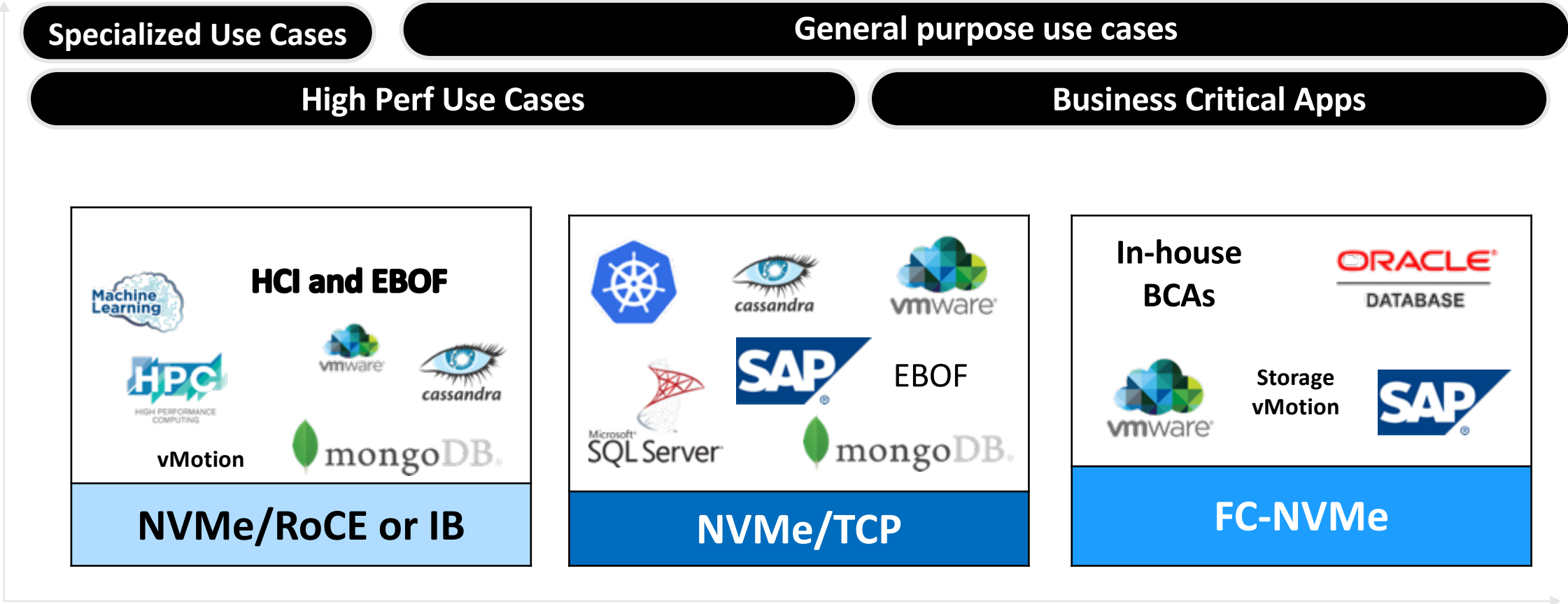


**NVMe/TCP and
NVMe/RoCE or IB**



**RoCE, TCP, iWARP
for HCI**

Workloads and use cases by Fabric



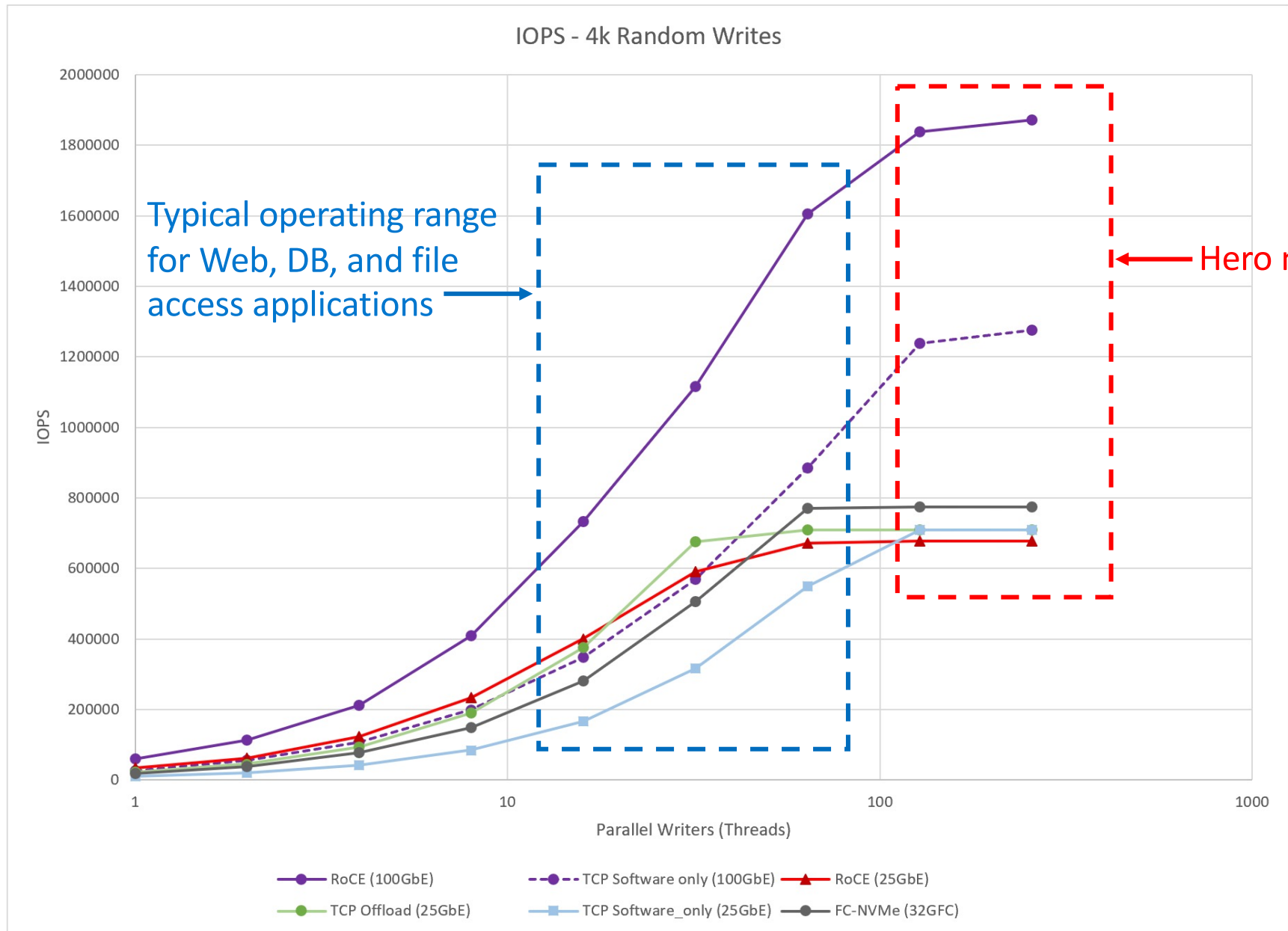
Logos are indicative of workload characteristics only.



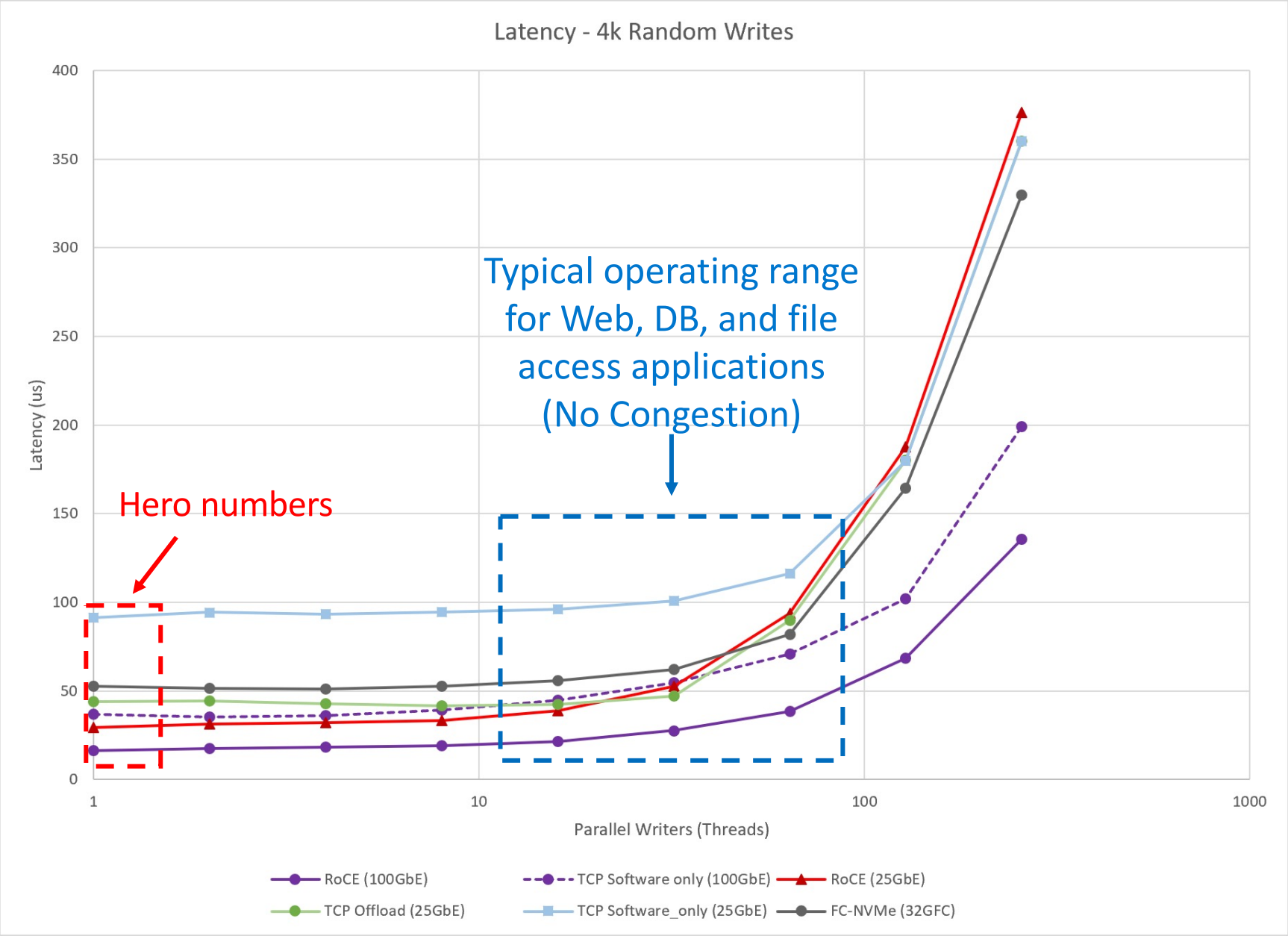
Configuration used for comparison of fabric types

- The test results/data used to generate the following graphs are based on testing performed at Dell Technologies.
- These results are only intended to compare the fabric types that are available for use.
 - In other words, all things being equal on both ends, the results demonstrate the benefit of using one fabric versus another.
- These test results DO NOT represent what end-users will experience when using one fabric versus another to access storage capacity on every storage platform.
 - This will be much more dependent on the storage platform implementation than on the fabric in use.
- The test configuration consisted of:
 1. An EZFIO client running on a physical server running Linux that has an appropriate adapter type installed
 2. All adapters used were either 32GFC, 25GbE or 100GbE based.
 3. A single 32GFC or 100GbE switch was used for “SAN” connectivity
 4. An LIO Target running on a physical server running Linux that is using RAMDISK for storage capacity.
- The Test results shown are all 4k Random Writes performed with varying numbers of threads

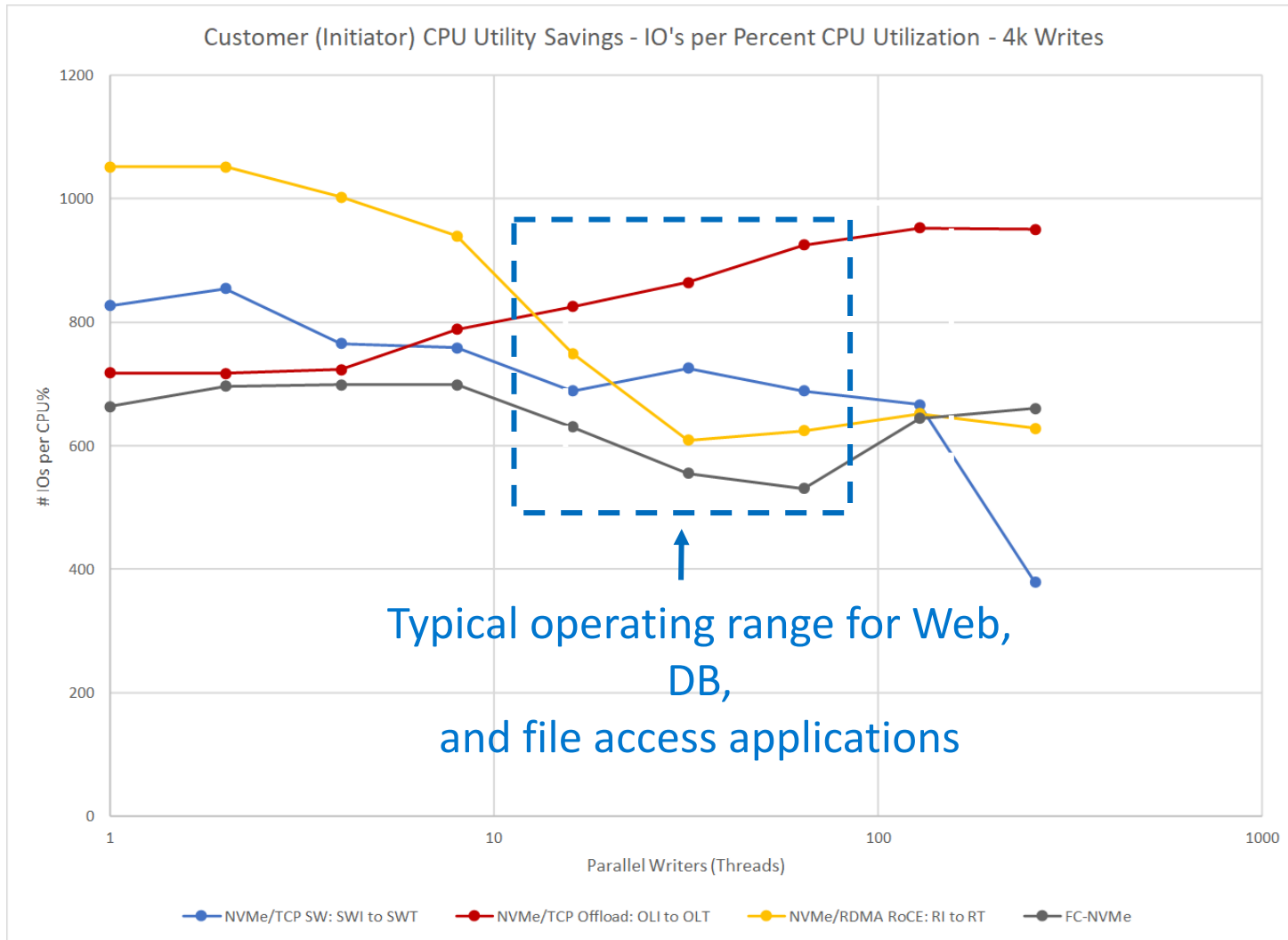
IOPS



Latency

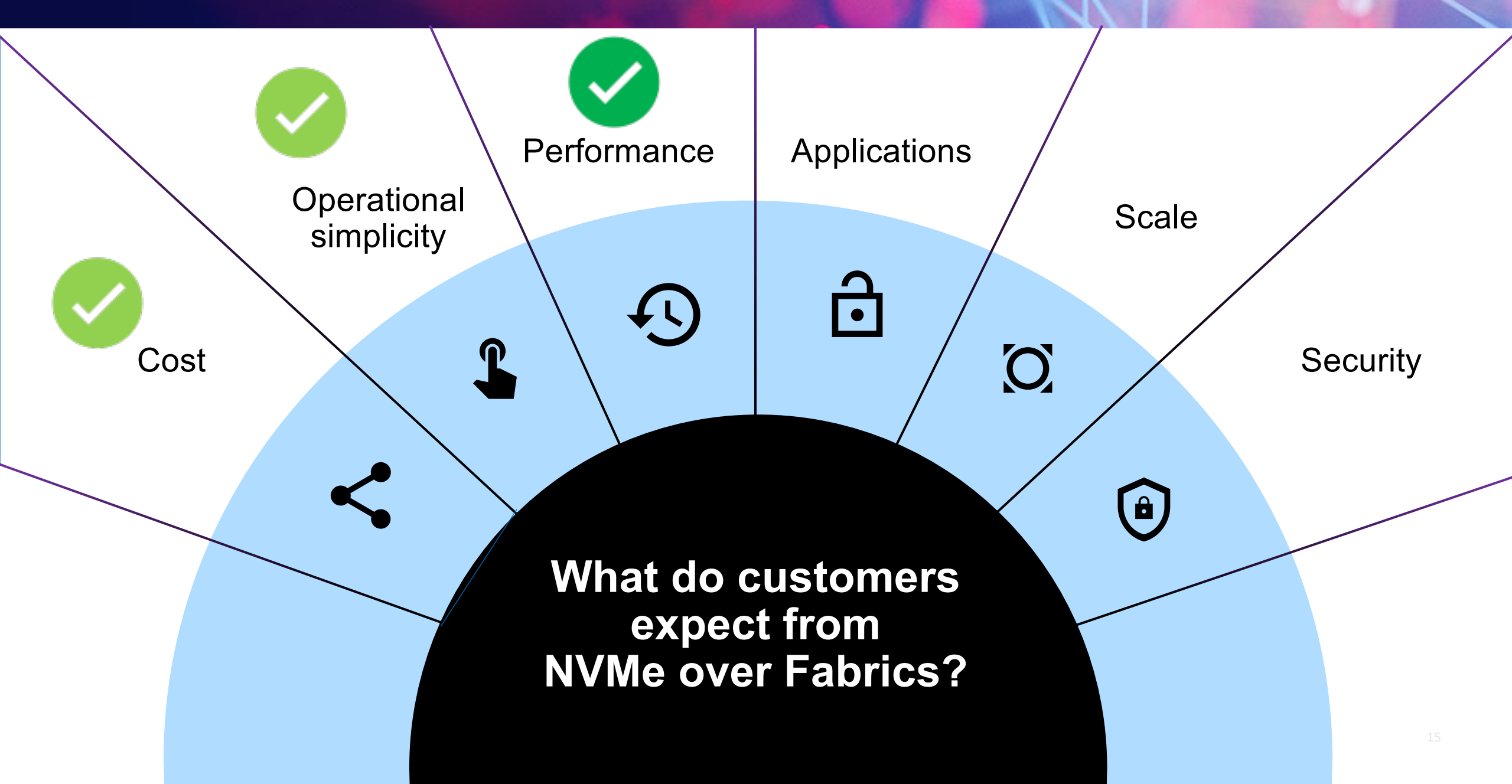


I/Os per Percent CPU Utilization



FC @ 32GFC

TCP & ROCE @ 25GbE













NVMe Fabric type comparison

		FC-NVMe	NVMe/RoCE	NVMe/TCP
OPEX benefits	High Speed Performance	✓	✓	✓
	Software Defined Storage		✓	✓
	Centralized Provisioning ¹	✓	✓	✓
	State Change Notifications	✓	✓	✓
	Edge/Distributed System at Scale			✓
	Cloud Operating Model/Automation			✓
	CapEx Cost Advantage		✓	✓

1. Source: <https://brasstacksblog.typepad.com/brass-tacks/2017/12/nvme-over-fabrics-discovery-problem.html>

Why TCP vs. RoCE?

poor    excellent

Aspect	RoCE	TCP/IP
Performance	Excellent 	Good, getting better 
Interoperability	Fair, getting better 	Excellent 
Interop Testing costs	High (FC-like) 	Moderate 
Network congestion impacts	Visible, unexpected 	Moderate, expected 
Network management impacts	New protocol, missing end-to-end functionality 	Part of normal operations 

- NVMe-oF/Ethernet provides standard, interoperable, high-speed, light-weight, low-latency, cost effective block storage access
- 25GE provides essentially the same throughput as 32G FC, but at a fraction of the cost
- NVMe-oF/TCP delivers better performance and reduced overhead compared to typical iSCSI
- TCP/IP for NVMe-oF transport just works by default
- Specialized configuration of TCP **is not required**
- Realistic performance is similar to RoCE (better as NVMe/TCP offloads emerge)
- TCP/IP is a better fit to Edge, IoT, Client deployments due to price & hardware
- TCP/IP allows a wide variety of network topologies (fully routable and fully flow controlled as needed)

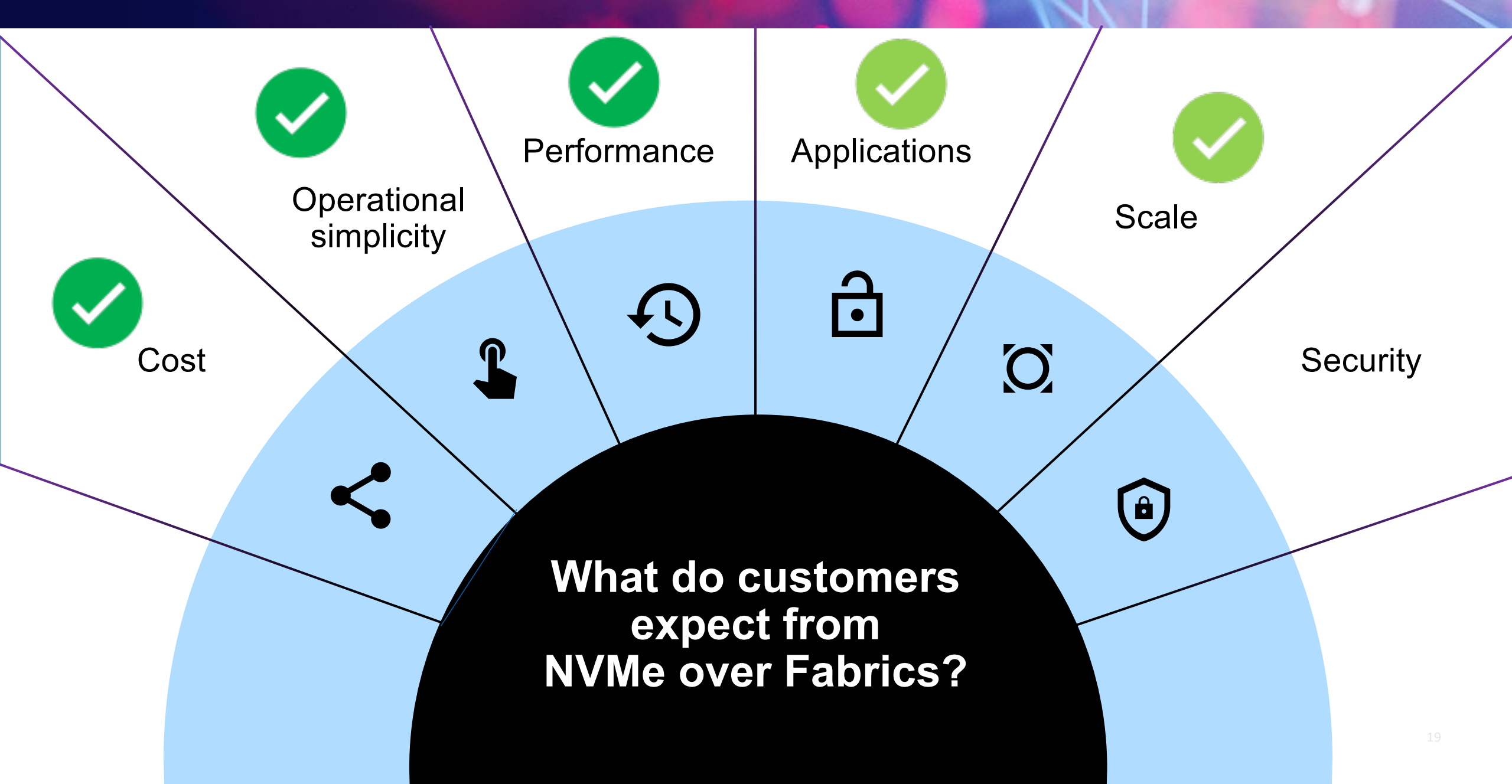
Initial Configuration Steps by Fabric type

Single Host to Single Target

Transport Protocol	Host Steps	Network Steps	Storage Steps	Total
FC	2	5	7	14
FC-NVMe	2	5	7	14
NVMe/TCP	4->1	3	8	15->12
NVMe/RoCE	5->2	7	9	21->18

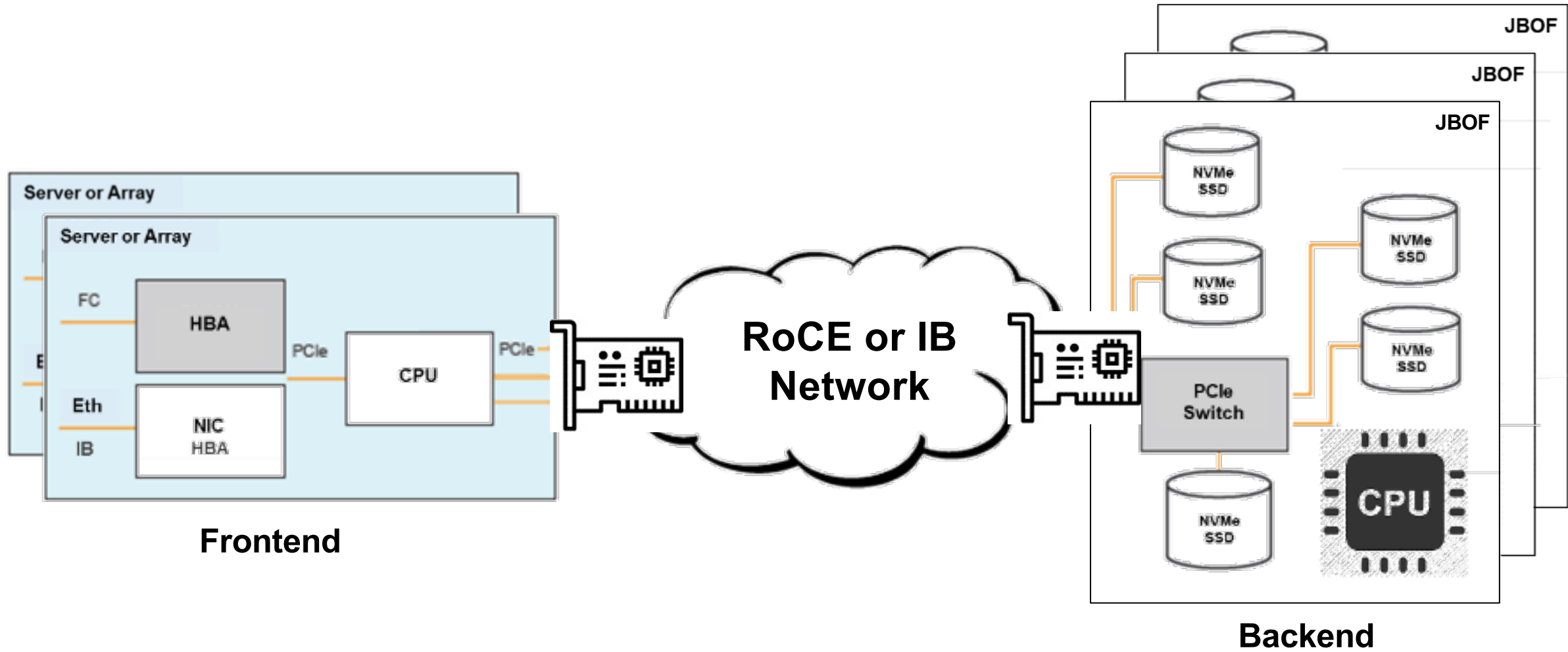
Source: <http://brasstacksblog.typepad.com/brass-tacks/2012/02/fc-and-fcoe-versus-iscsi-network-centric-versus-end-node-centric-provisioning.html>

Source: <http://brasstacksblog.typepad.com/brass-tacks/2015/05/fibre-channel-is-better-than-ethernet.html>

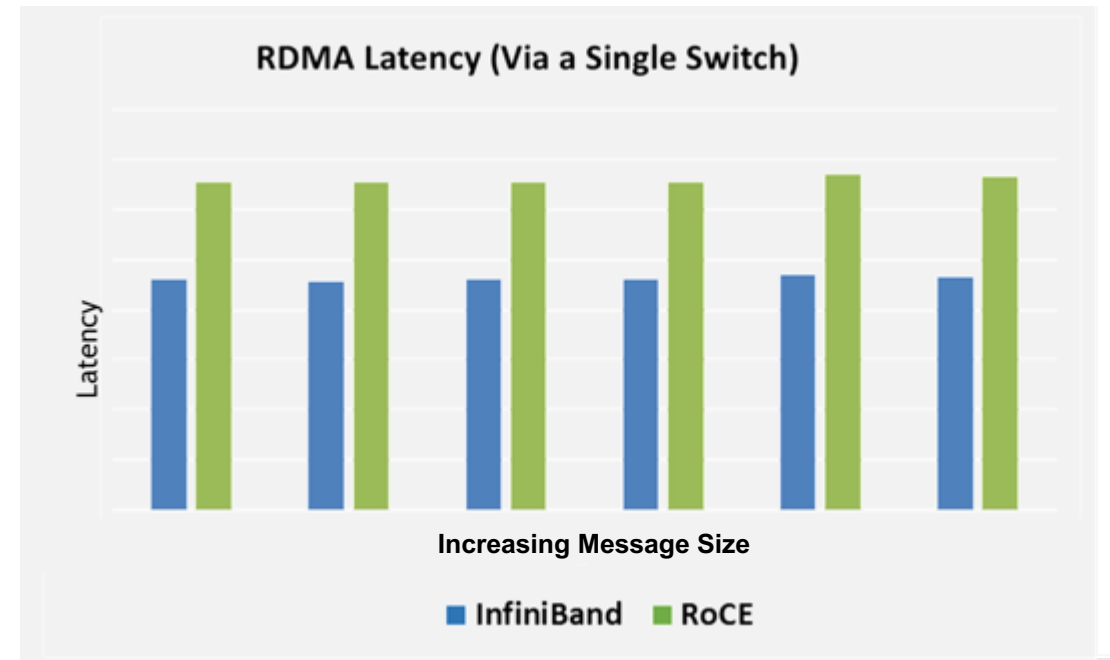
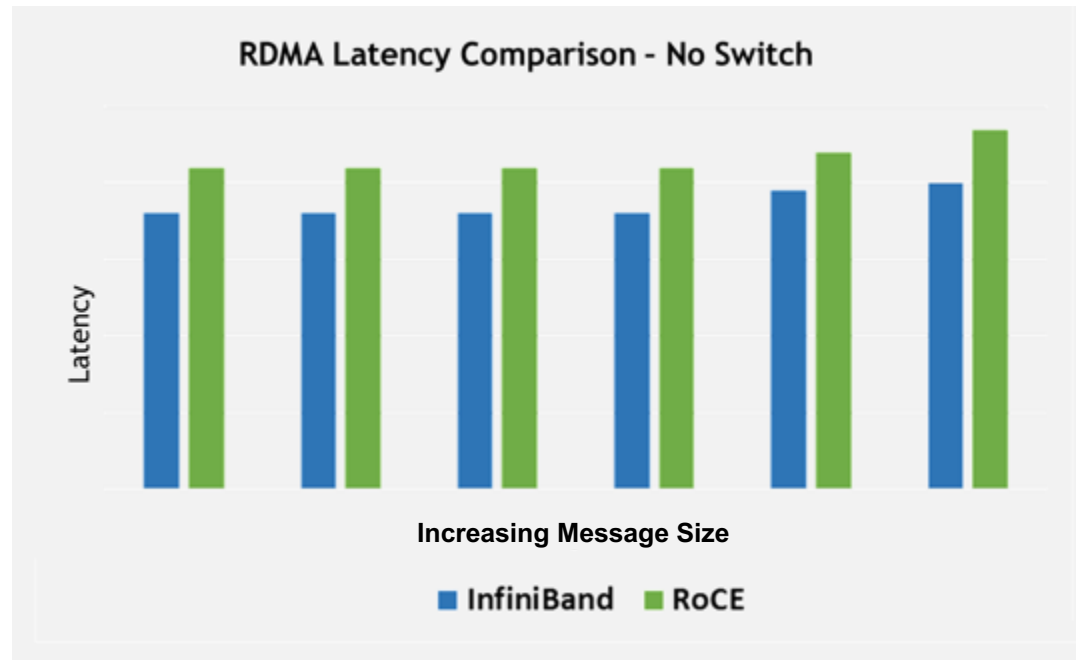


Where is all this NVMe-oF™
performance being used?

Embedded Scale Out Back End

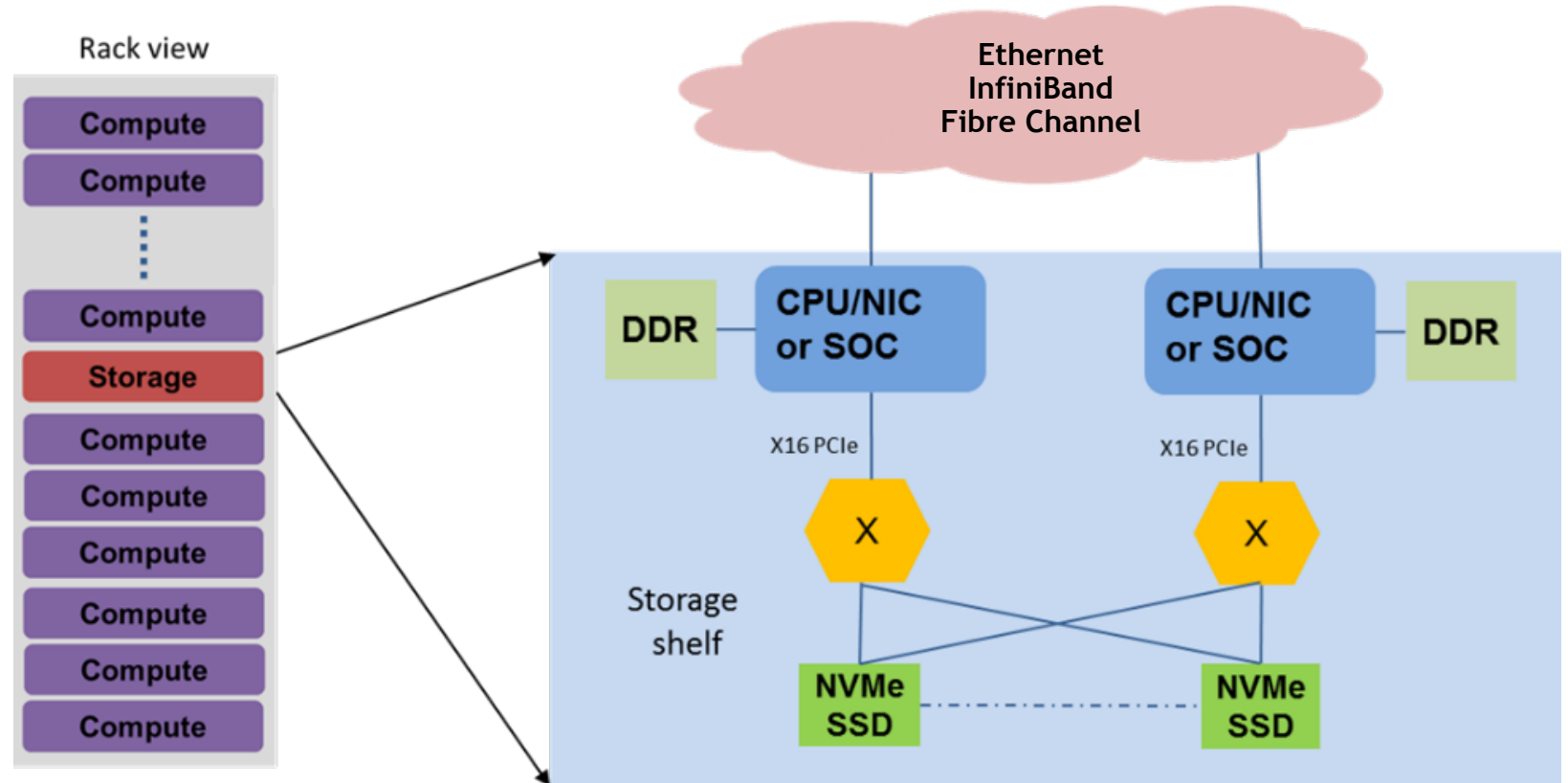


What is the difference between IB and RoCE



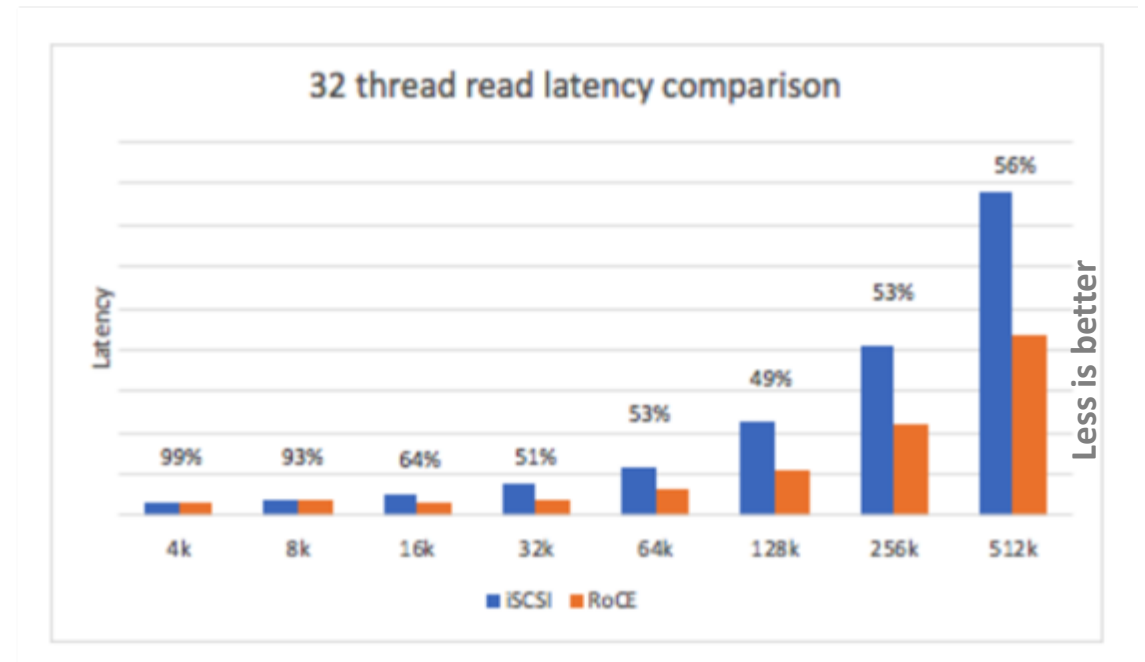
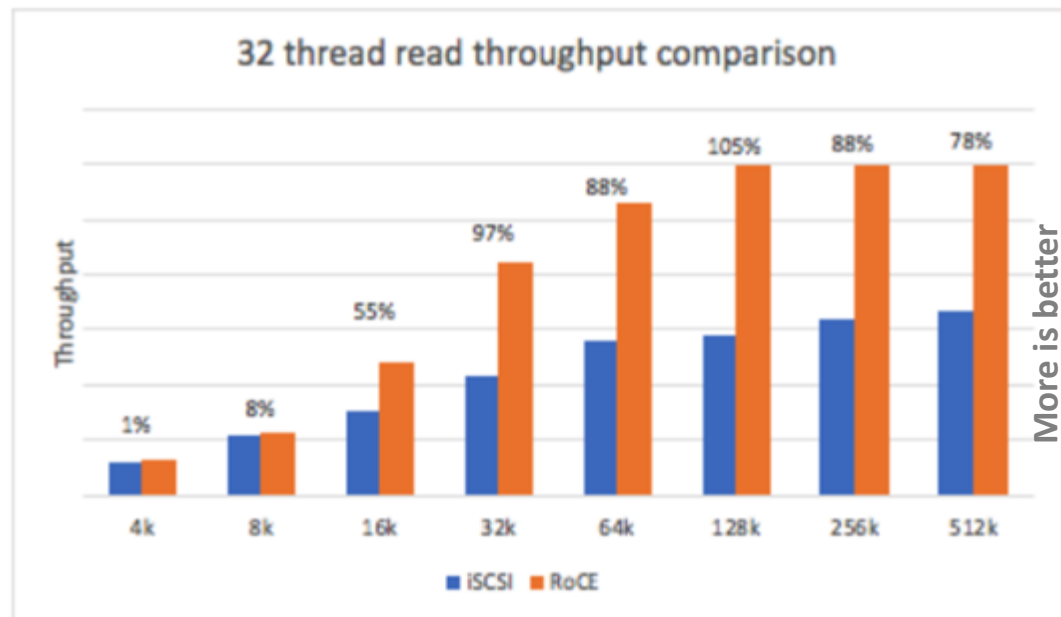
Classic SAN

- SAN features at higher performance
 - Better utilization: capacity, rack space, and power
 - Scalability
 - Management
 - Fault isolation
- Driver support in Linux, VMware and Windows(3rd party)

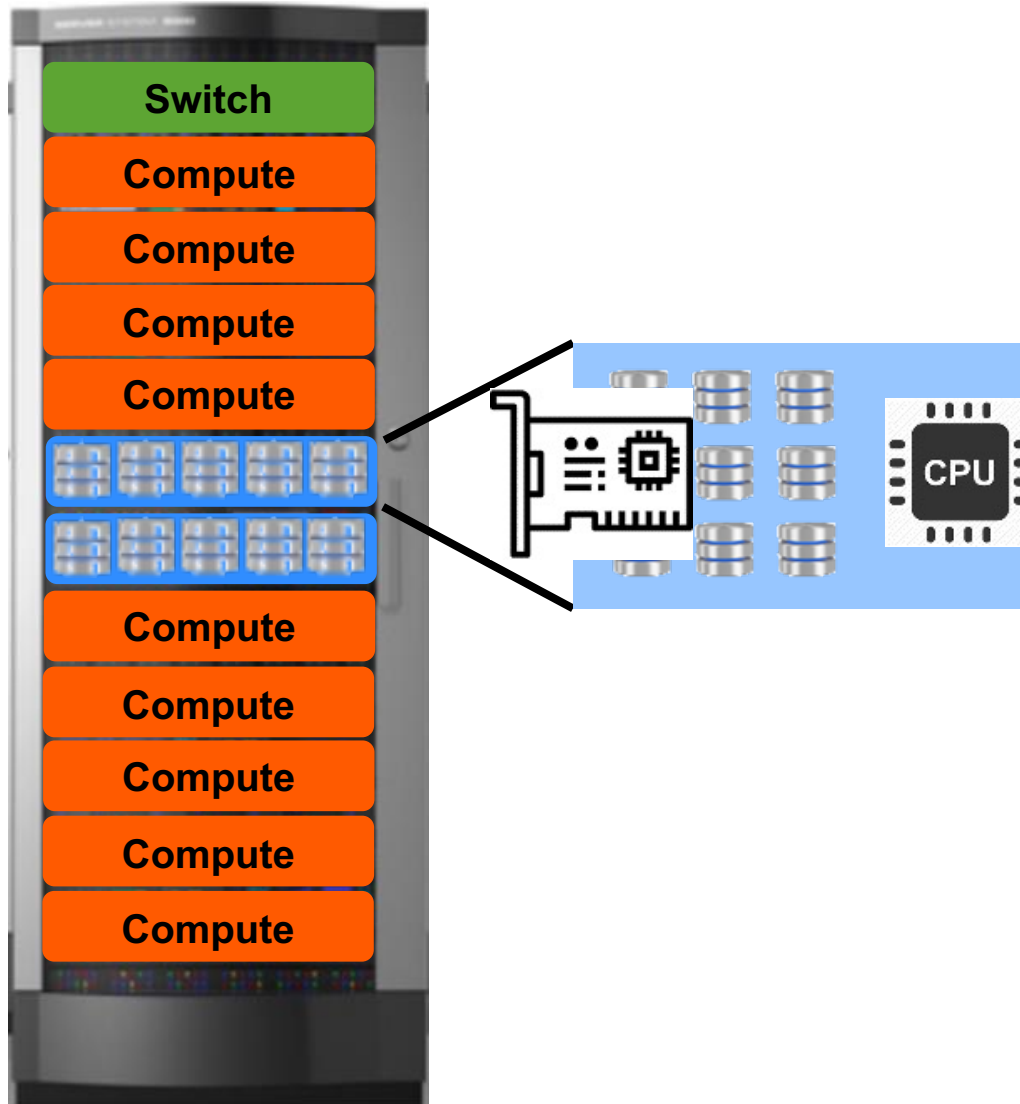


iSCSI and NVMe-oF

Testing with 25GbE on virtual OS



Compute Storage Disaggregation



- Also called Composable Infrastructure and Rack Scale
- NVMe over Fabrics/RoCE enables with nearly local disk performance
- NVMe/TCP may be used depending on performance requirements

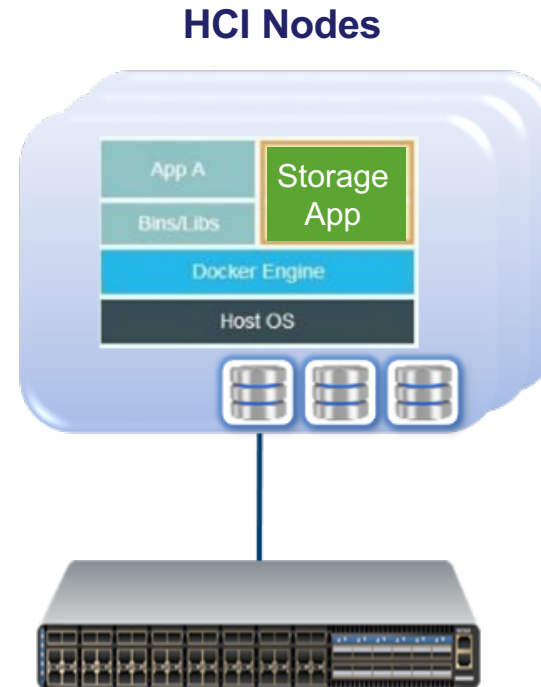
Hyperconverged and Scale-Out Storage

- Scale-out

- Cluster of commodity servers
- Software provides storage functions

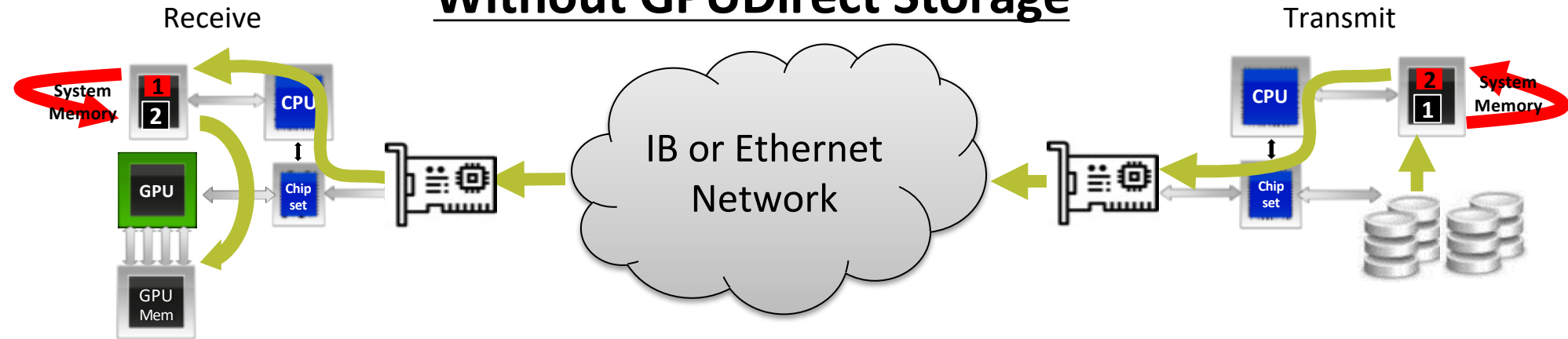
- Hyperconverged collapses compute & storage

- Integrated compute-storage nodes & software
- NVMe/RoCE performs like local/direct-attached SSD
- NVMe/TCP may be used depending on performance requirements

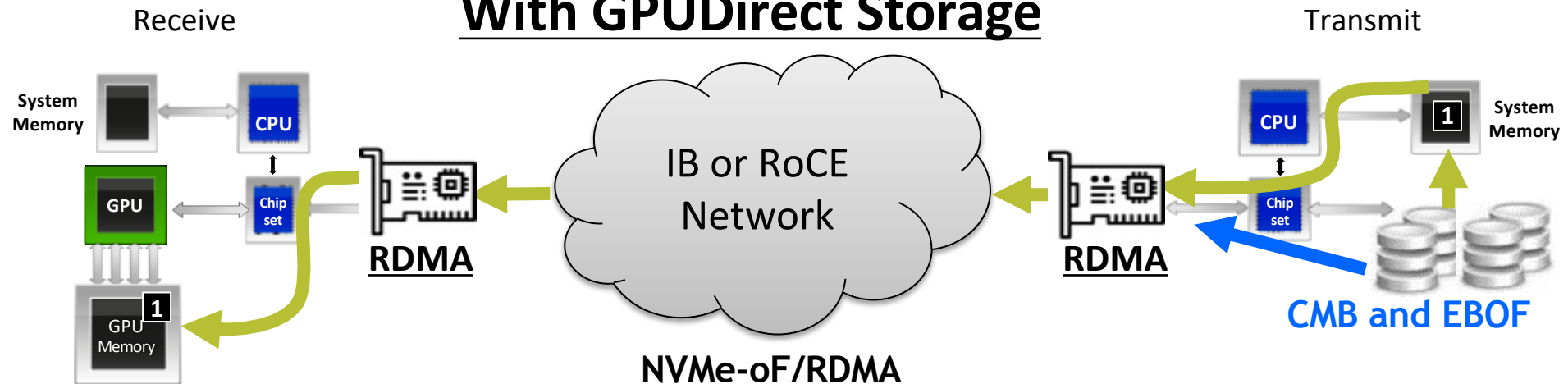


GPUDirect Storage(GDS)

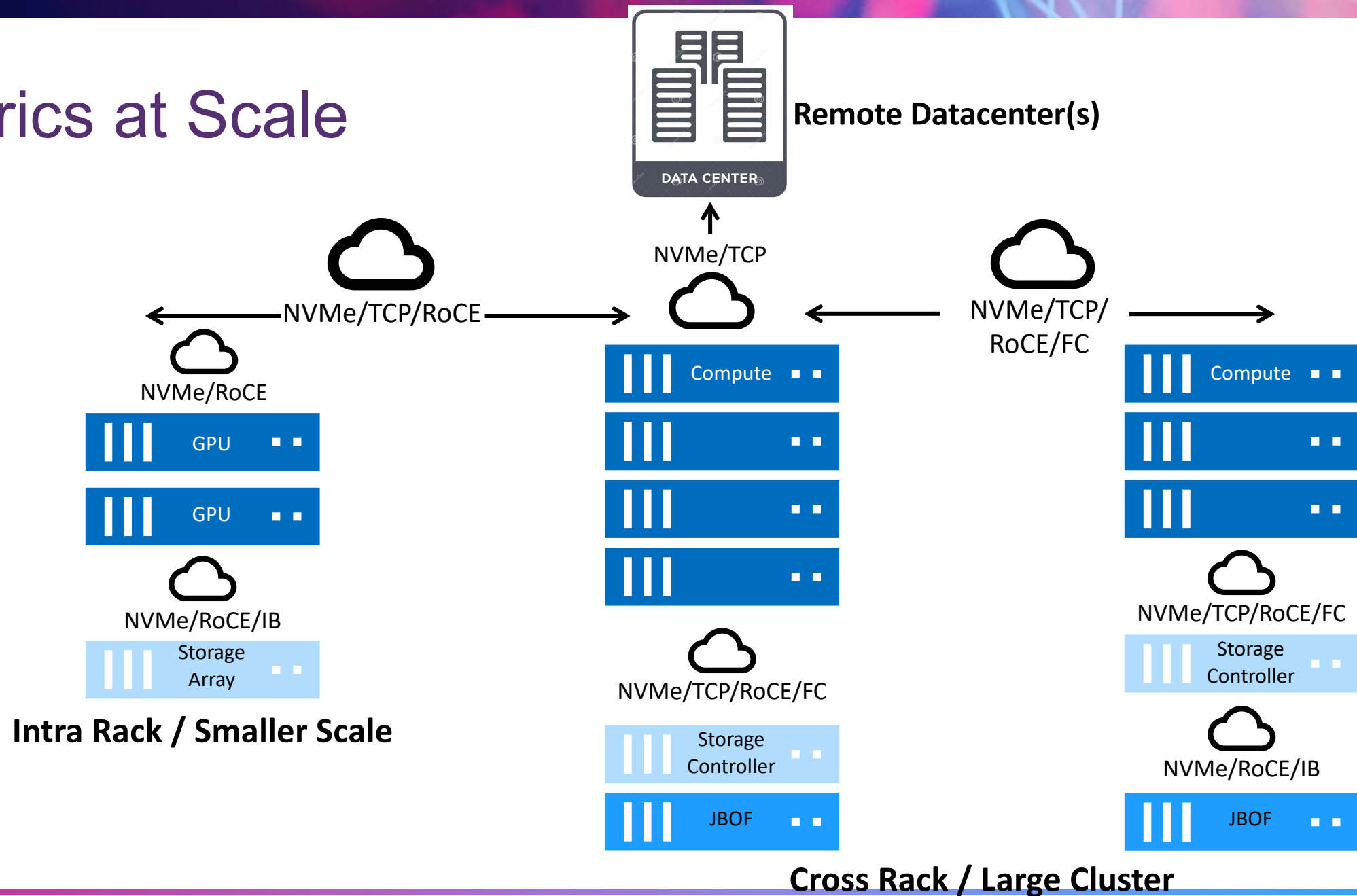
Without GPUDirect Storage



With GPUDirect Storage



Fabrics at Scale



What do customers expect from NVMe over Fabrics?



Cost



Operational simplicity



Performance



Applications



Scale



Security



Potential DC NVMe-oF Security Threats

**Sniffing
Storage Traffic**



**Storage
Masquerading**



**Data
Corruption**



Session Hijacking

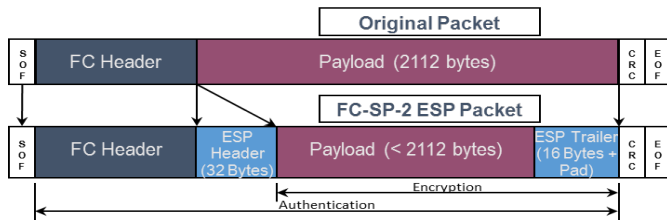


Must secure NVMe Data in flight and rest

Securing NVMe-oF

FC-SP-2 (FC-NVMe)

- FC-SP-2 is an ANSI/INCITS standard (2012) that defines protocols to –
 - Authenticate** Fibre Channel entities
 - Setup** session **encryption keys**
 - Negotiate parameters to ensure per **frame integrity and confidentiality**
 - Define and **distribute security policies** over FCP and FC-NVMe
- Concurrent FCP and FC-NVMe
- HBA provide full offload



TLS for NVMe/TCP

- TLS is a protocol suite defined by the IETF –
 - Privacy and data integrity** between two communicating applications
 - Higher-level protocols like **NVMe** that can layer on top of the TLS protocol transparently.
- NVMe-oF over TCP already supports TLS
- The NVM Express Technical Proposal TP 8011 is adding support for TLS 1.3
- Many NICs can provide full offload

IPsec for NVMe/TCP/RoCE

- IPsec is a secure network protocol suite defined by the IETF –
 - Authenticates and encrypts** the data packets over IP networks
 - Higher-level protocols like **NVMe** that can layer on top of the IPsec protocol transparently.
- NVMe-oF/TCP/RoCE already supports IPSEC
- Many NICs can provide full offload

NVMe over Fabrics



Cost



Operational
simplicity



Performance



Applications



Scale



Security





Questions

Register for Our Next Live NVMe-oF Webcast

Security of Data on NVMe over Fabrics, The Armored Truck Way

May 12, 2021

10:00 am PT / 1:00 pm ET

Register at: <http://bit.ly/SNIANVMeoFSecurity>

After this Webcast

- Please rate this webcast and provide us with your feedback
- This webcast and a copy of the slides will be available at the SNIA Educational Library <https://www.snia.org/educational-library>
- A Q&A from this webcast, including answers to questions we couldn't get to today, will be posted on our blog at <https://sniansfblog.org/>
- Follow us on Twitter [@SNIA NSF](https://twitter.com/SNIA NSF)

Thank You