STORAGE DEVELOPER CONFERENCE
SDC 21
BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

A SNIA Event

Containerized Machine Learning Models using NVMe

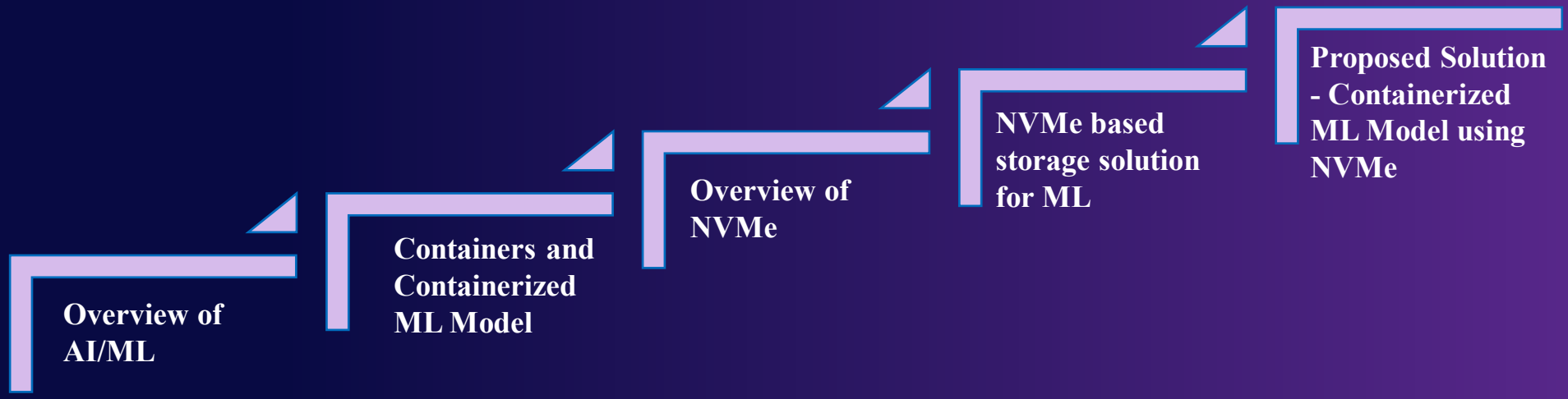
Ramya Krishnamurthy

Ajay Kumar

Ashish Neekhra

HPE

Agenda



 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Overview of AI/ML

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

AI/ML-Overview

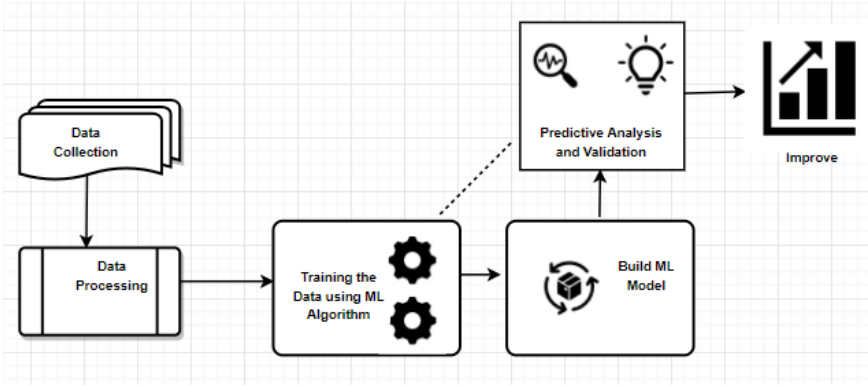
- Machine learning is one of the most strategic technology adoptions in recent years.
- Organizations across nearly every industry are adopting AI
- ML helps companies achieve a competitive edge through improved insights and efficiency
- Also used to automate business processes, prevent and resolve problems through predictive analysis, and improve customer experience

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Fundamentals of ML

Machine Learning Algorithm

- Fundamental logic of a Machine Learning model.
- Set of rules and statistical methodologies for extracting useful data and learning patterns from it.
- Uses the training data to build the ML model
- Machine learning models improve as more training data is collected.
- Identifies patterns that are required to predict the outcome.
- Incorporates corrections and improves its future decision making



 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Criticality of Machine Learning

- Solve complex problems.
- More efficient than manual processing.
- Able to analyze data and find patterns in less than a second.
- Critical for data-driven decisions
- It provides better performance as the scale of data increases.

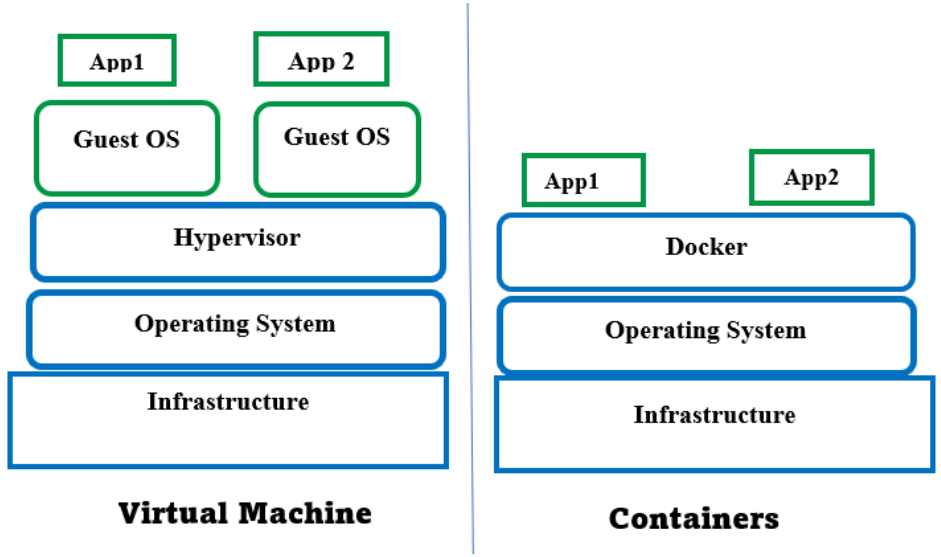
 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Containers and Containerized ML Model

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Containers

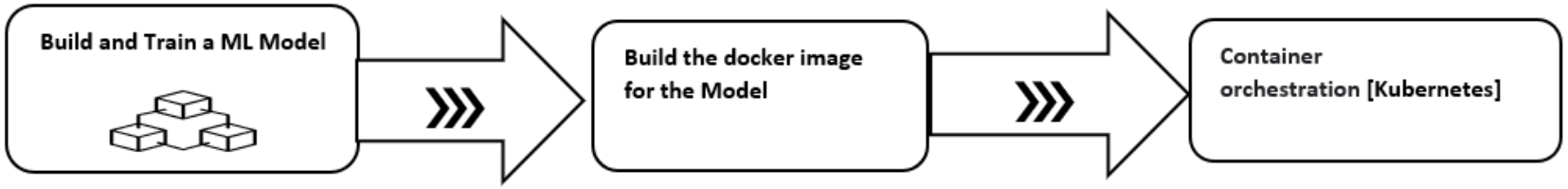
- Lightweight in comparison to virtual machines
- Virtualize the OS and not the hardware
- Efficient, portable, predictable-Applications running in containers can be deployed easily to different operating systems and hardware environments.
- It can be easily and rapidly deployed.
- It is easy to manage and maintain.



 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Containerized ML Model

- Machine learning models can consume a lot of resources and need a lot of computing power to predict, validate, and recalibrate. Standalone machines end up being over utilized, resulting in instability.
- Containerizing ML models is one promising answer to this challenge.
- Containers provide portability and easier dependency management.
- Portability provides the flexibility to spin up multiple containers for simultaneous runs of various training models.
- Containers require fewer system resources than virtual machine environments because they don't include operating system images.
- ML models can be trained faster.
- Containers have a faster start time as they run as a process and do not require an operating system to boot.



 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

NVMe based storage solution for ML

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

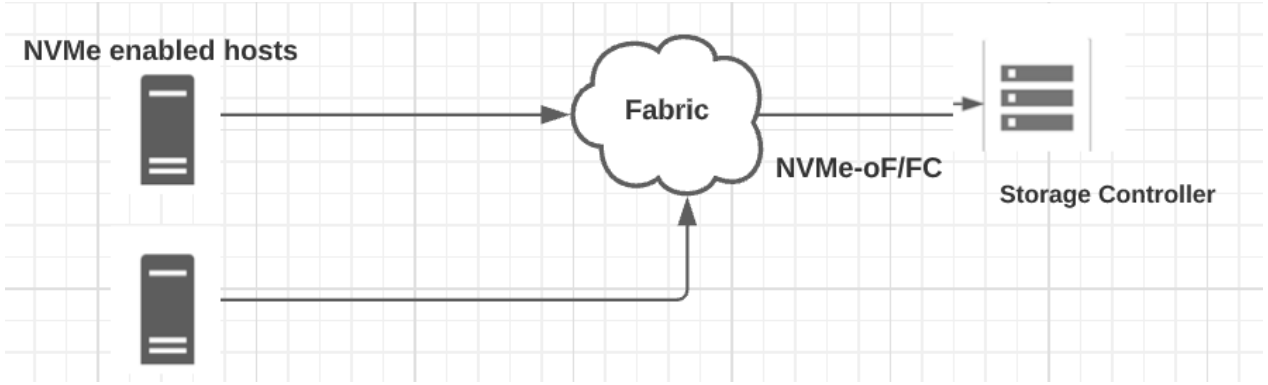
NVMe/Flash

- Delivers data to GPUs at extremely high speeds.
- It eliminates latency bottlenecks and communicates directly with the CPU.
- It outperforms HDDs by up to 1000 times.
- It provides the fastest throughput and response times for flash and next-generation solid-state drives (SSDs).
- Considered to be the best choice for AI/ML training to avoid "starving" GPUs

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

NVMe Over Fabrics

- It is lighter than SCSI and has a leaner driver stack that allows it to run faster while consuming fewer resources.
- Better host performance than SCSI
- No protocol translation is needed in HBA.
- It takes less CPU processing time than FCP.
- Can use the existing Fiber Channel infrastructure for NVMe-oF
- Improved driver performance on the storage controller



 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

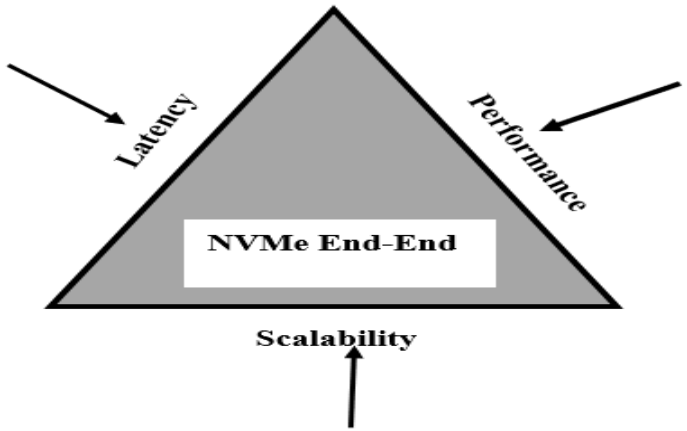
Advantages of using NVMe for AI/ML

It supports NVMe SSDs on the back end, along with NVMe-oF for front-end host connectivity.

- Speeds up data transmission between storage devices and servers.
- It provides the performance, scalability, and low latency needed for higher workloads.
- NVMe-oF reduces the number of single points of failure due to inbuilt multipathing.
- Improved capacity utilization, shared access

Training an AI model is the most resource-intensive stage of machine learning

End to end NVMe is the answer for machine learning workloads!!!!



 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Proposed Solution - Containerized ML Model using NVMe

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

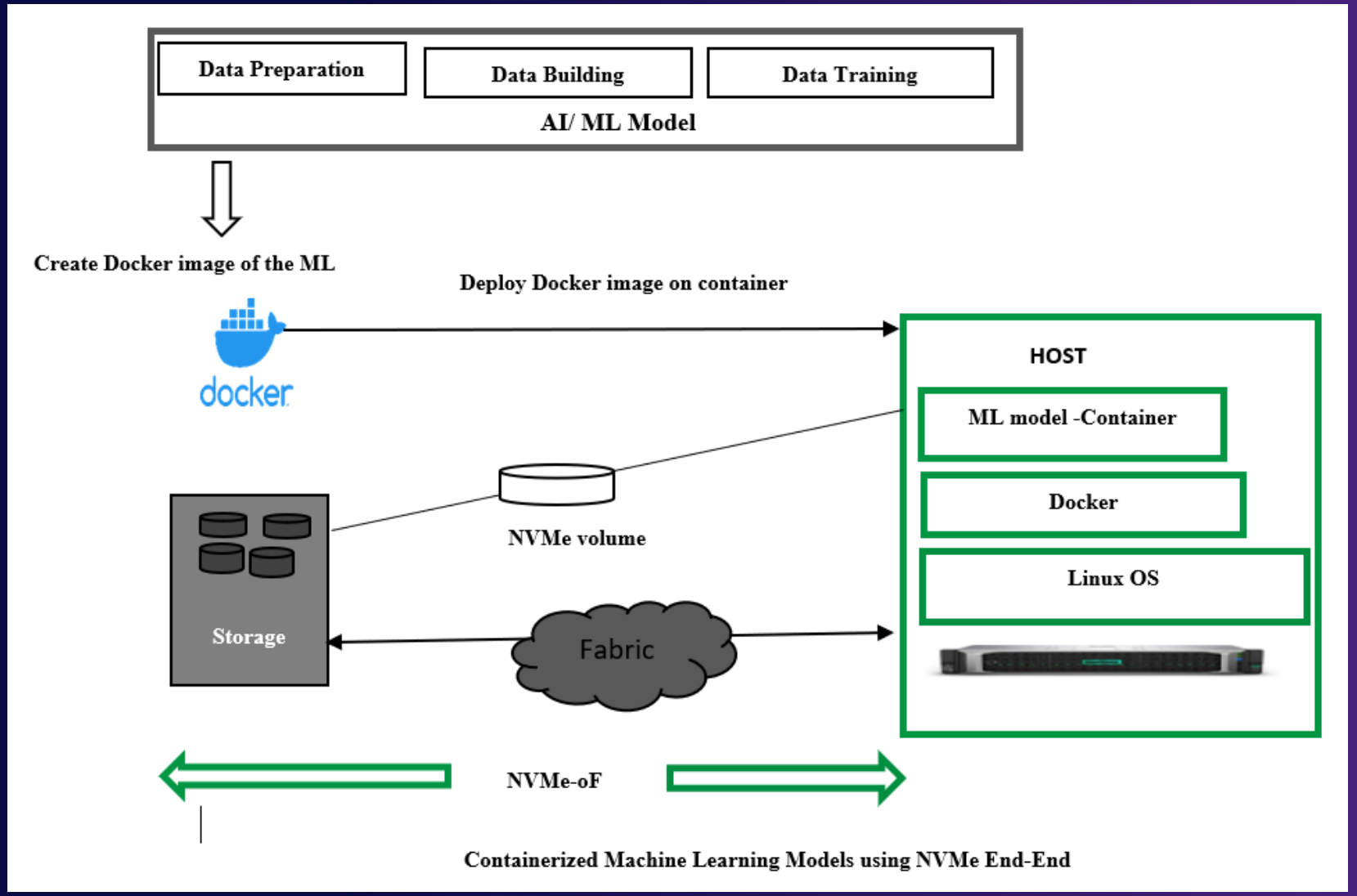
Proposed Solution

Our proposed solution includes

- The server has an x86-based architecture and a dual Intel (R) Xeon (R) processor running at 2.40GHz with 64GB of RAM with SAS and NVMe drives.
- The NVMe-capable storage network infrastructure
- Linux Operating system (O/S) on the server that provides support for NVMe-oF
- The storage controller uses PCIe Gen 3 and supports end-end NVMe [NVMe SSD physical drives and NVMe-oF]
- Both the server and storage system are configured with Gen 6 FC 32Gb HBAs.

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Containerized ML Model using NVMe



- *Build the model.*
- *Create the requirements file containing all the required libraries.*
- *Create the docker file with necessary environment setup*
- *Build the Docker image of the ML model*

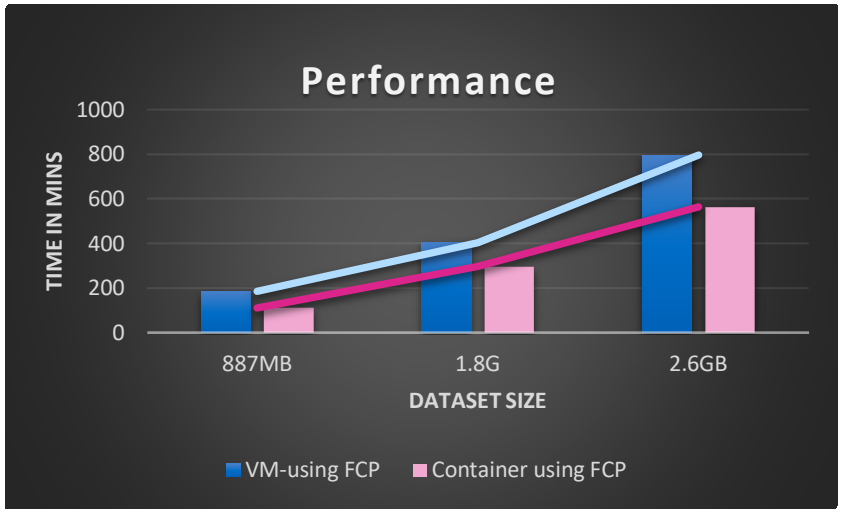
 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Performance Comparison-Phase 1

- As a baseline, we evaluated the performance of an ML model when running on the virtual machine. Which was deployed on an x86 server using NVMe drives from storage over FCP
- The same ML model was subsequently run on a container that was deployed on the x86 server, with storage provided by NVMe drives through FCP

The machine learning model used here is a Deep learning unsupervised model.

The graph here shows a performance comparison for a sample training dataset between virtual machines and containers.
Since containers had lower overhead than virtual machines-we see better performance with containers in comparison to the VM



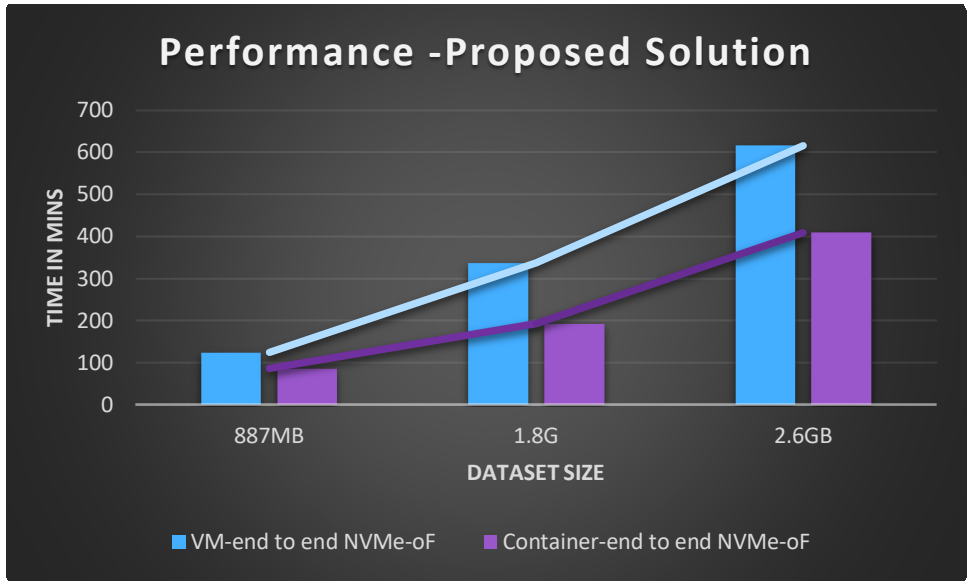
 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Performance Comparison –Proposed solution

Now moving ahead to our proposed solution, we continued the performance evaluation between VM and container with End-End NVMe

- We ran the VM and container on the same x86 server using End-End NVMe [using NVMe drives using NVMe-oF]
- Performance evaluations were conducted for a variety of training dataset sizes

We discovered that the container's performance advantage grew as the dataset size grew and we see more performance gains with the proposed container solution when using end-to-end NVMe.



Time in mins			
Dataset size	887MB	1.8G	2.6GB
VM-NVMe drives over FCP	184	404	795
Container-NVMe drives over FCP	112	296	563
VM-End to End NVMe	124	337	616
Container-End to End NVMe	86	192	409

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

References

HPE Storage
www.hpe.com/storage

HPE Servers
www.hpe.com/servers

AI/ML
https://www.hpe.com/us/en/solutions/artificial-intelligence.html

Dockers
https://www.docker.com/

Containers
https://github.com/containers

NVMe
https://nvmexpress.org/

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

Please take a moment to rate this session.

Your feedback is important to us.

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.

THANK YOU

 ©2021 Storage Networking Industry Association ©. HPE. All Rights Reserved.