



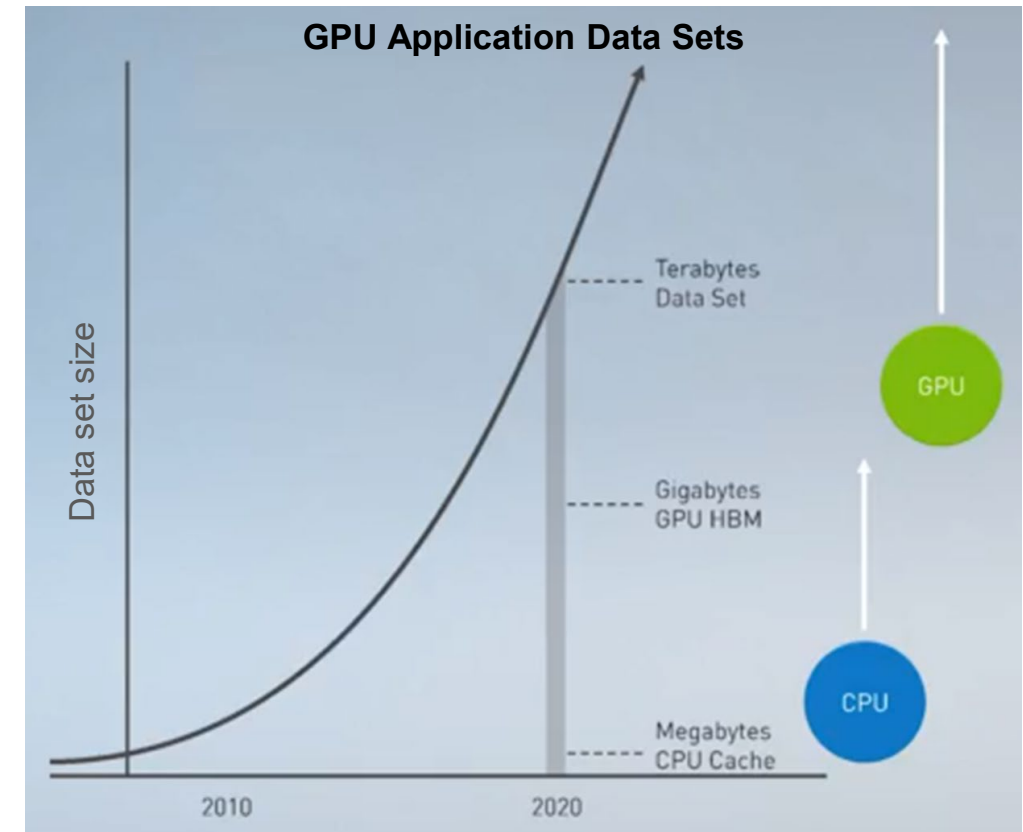
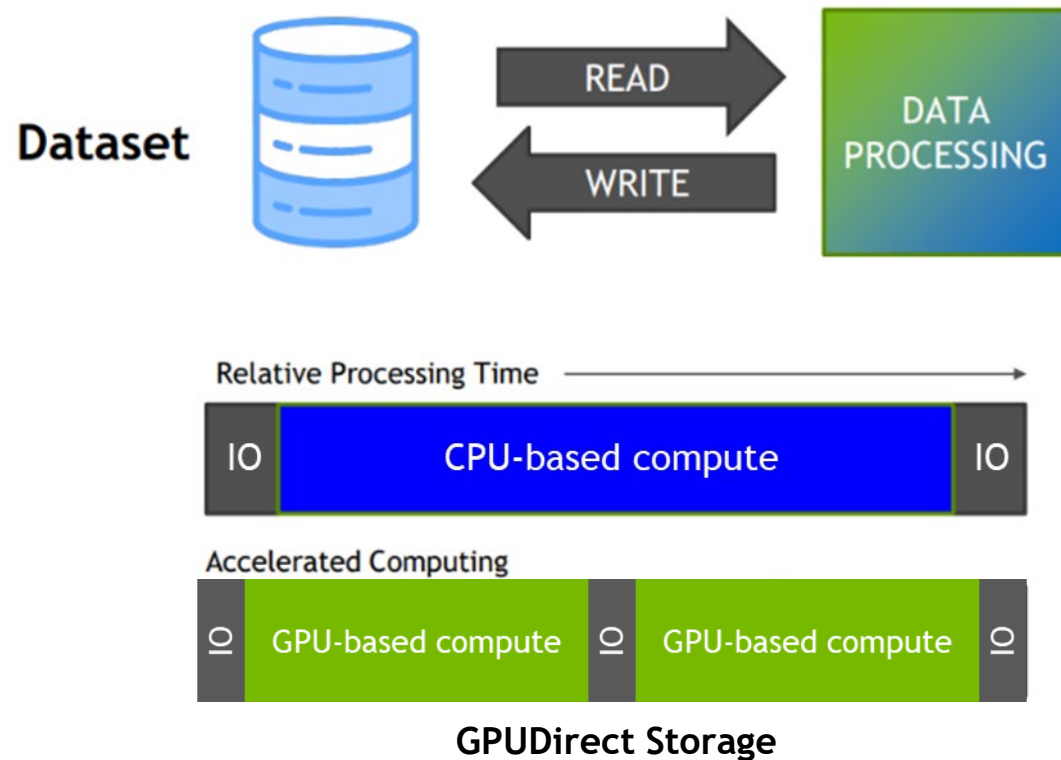
Flash Memory Summit

# The Advantages of Ethernet SSDs in AI

Presented by: Rob Davis and Tyler Nelson

# IO challenges for GPU Storage

## GPU performance & expanding data set sizes





# Examples of Large Data Set AI Applications

- **Mars Lander**

- 150 TB of data
- Visualization of atmospheric effect

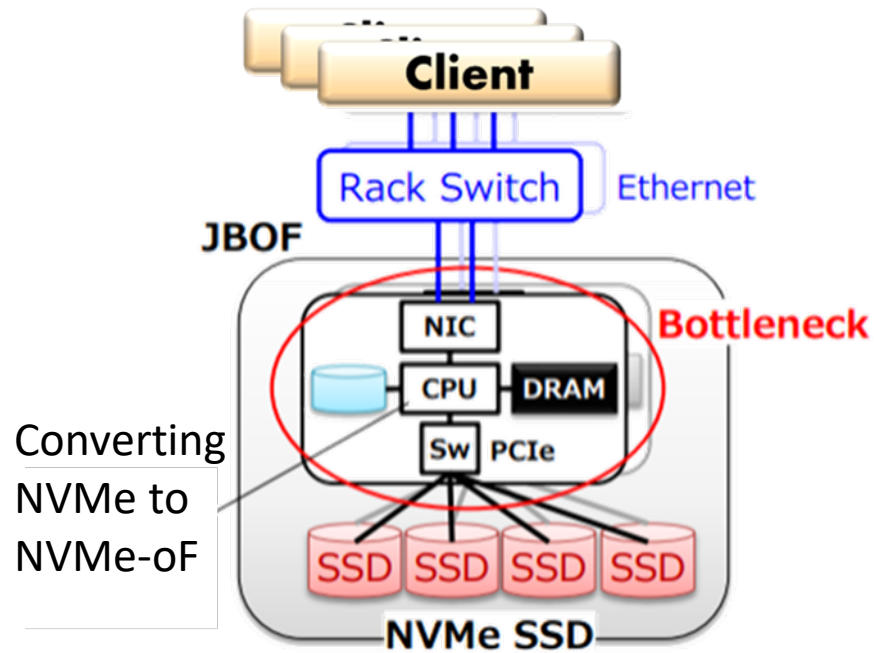
- **Hurricane Simulation**

- AI Algorithm learns weather conditions common in hurricane creations
- It is then fed massive amounts of live weather data it analyses to identify potential for storms

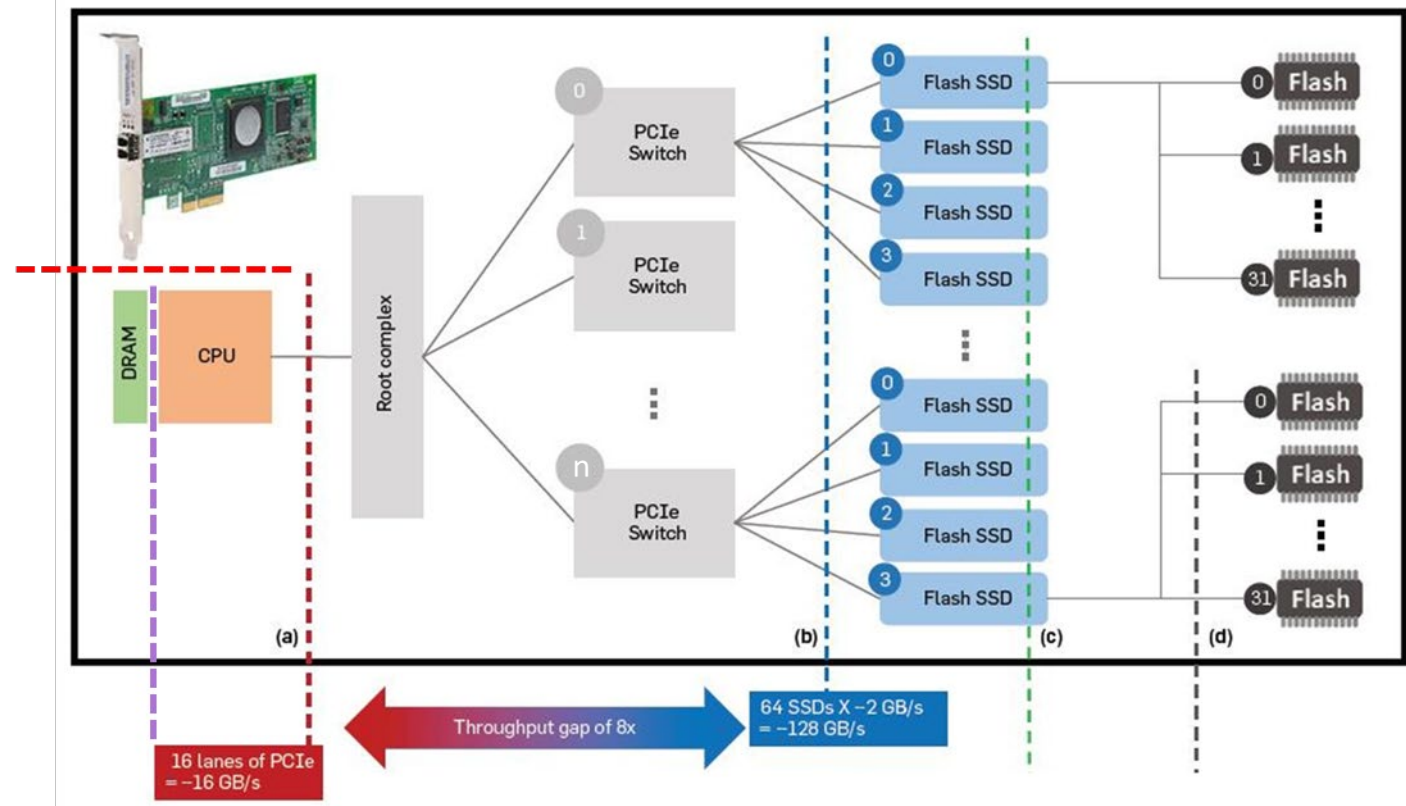


<https://www.youtube.com/watch?v=GAZP1NcdWMO>

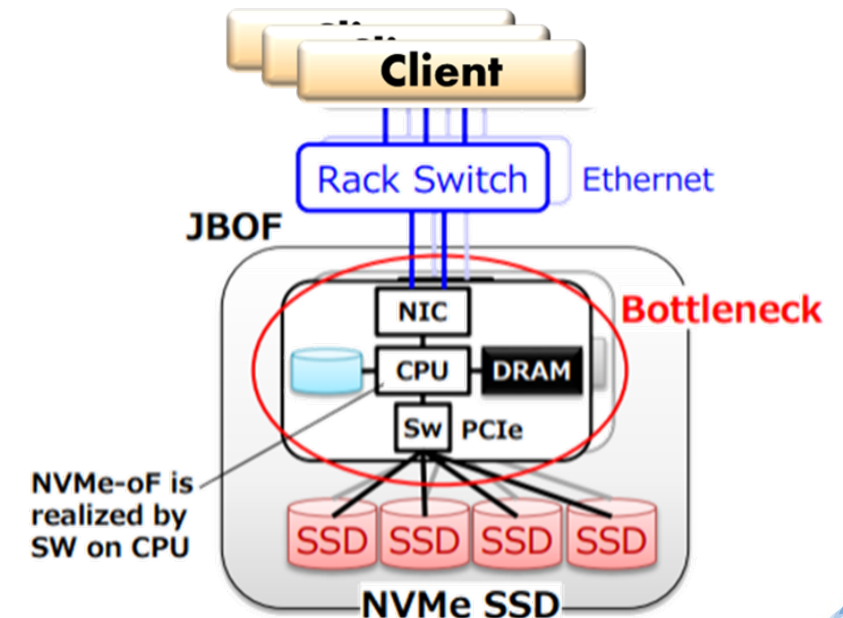
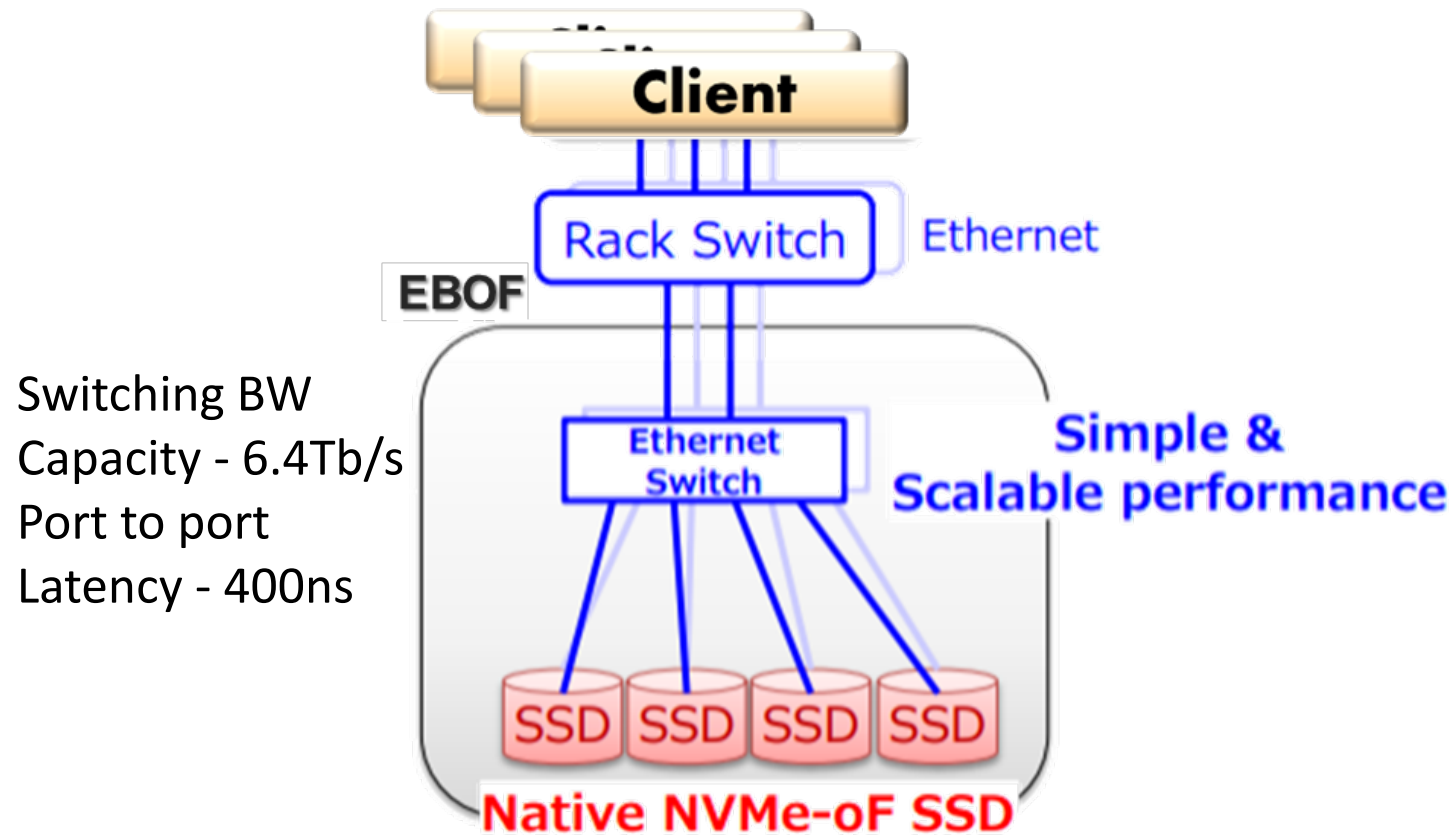
# Most Remote Networked Storage Solutions Limit SSD Performance



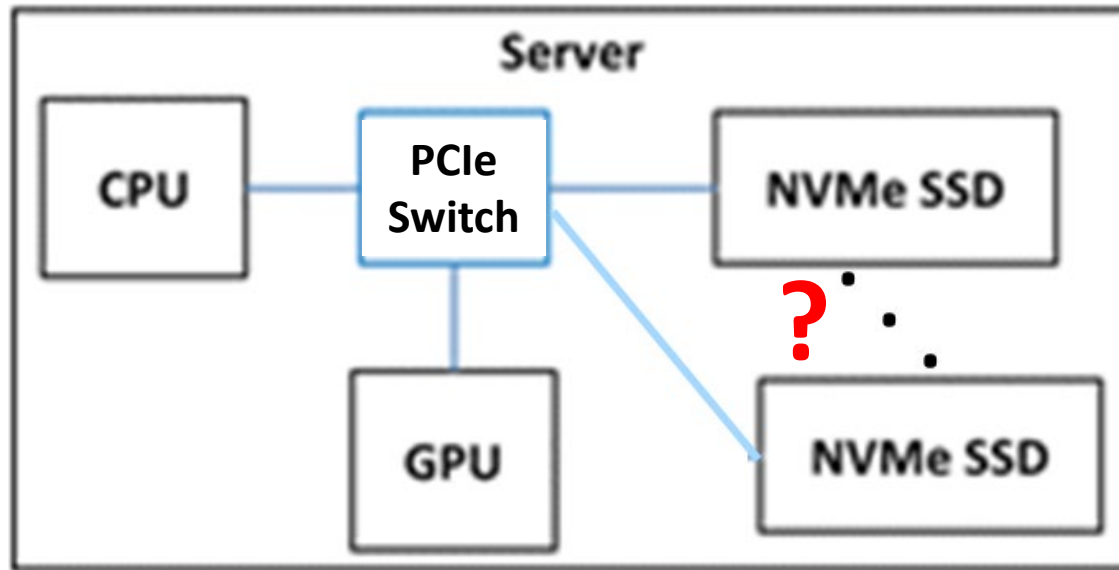
NIC, CPU, DRAM, PCIe Switching all potential bottlenecks



# Ethernet SSD Solutions Remove SSD Performance Bottlenecks



# Local Storage Solutions have Limited Capacity Scaling



A server can only hold so many SSDs

## Local NVMe vs. Remote NVMe-oF SSD

	QD
Random Read	1
	4
	8
	16
	32
	64

IOPS		
Fabric	Native	Fabric vs. Native
13,129	13,924	-6%
51,391	54,411	-6%
99,958	105,602	-5%
188,976	198,764	-5%
337,520	351,569	-4%
536,142	555,102	-3%

	QD
Random Write	1
	4
	8
	16
	32
	64

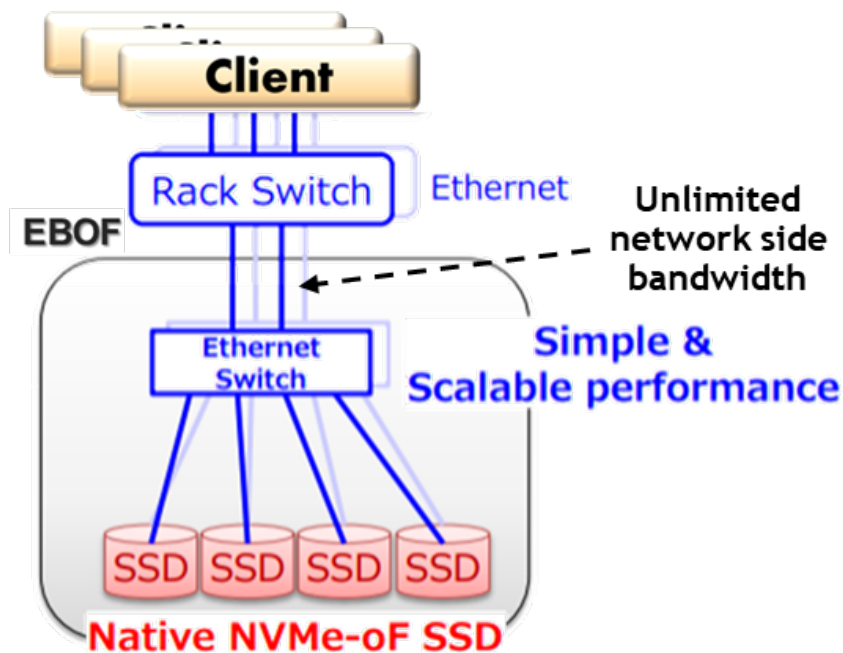
IOPS		
Fabric	Native	Fabric vs. Native
113,270	114,414	1%
122,499	121,531	1%
122,592	121,667	1%
122,617	121,876	1%
122,622	121,889	1%
122,633	122,856	0%

Remote Networked storage has unlimited capacity

# With ESSDs you get Higher Performance and Lower Cost

Two reasons this new architecture is attractive for AI Solutions

- 1) Performance
- 2) Cost



JBOF price comparison  
(Excluding SSD cost)  
NetApp 2019 FMS



JBOF Price per Gbit of performance



\* Supports one 2x200G RNIC connected with x16 PCIe Gen4

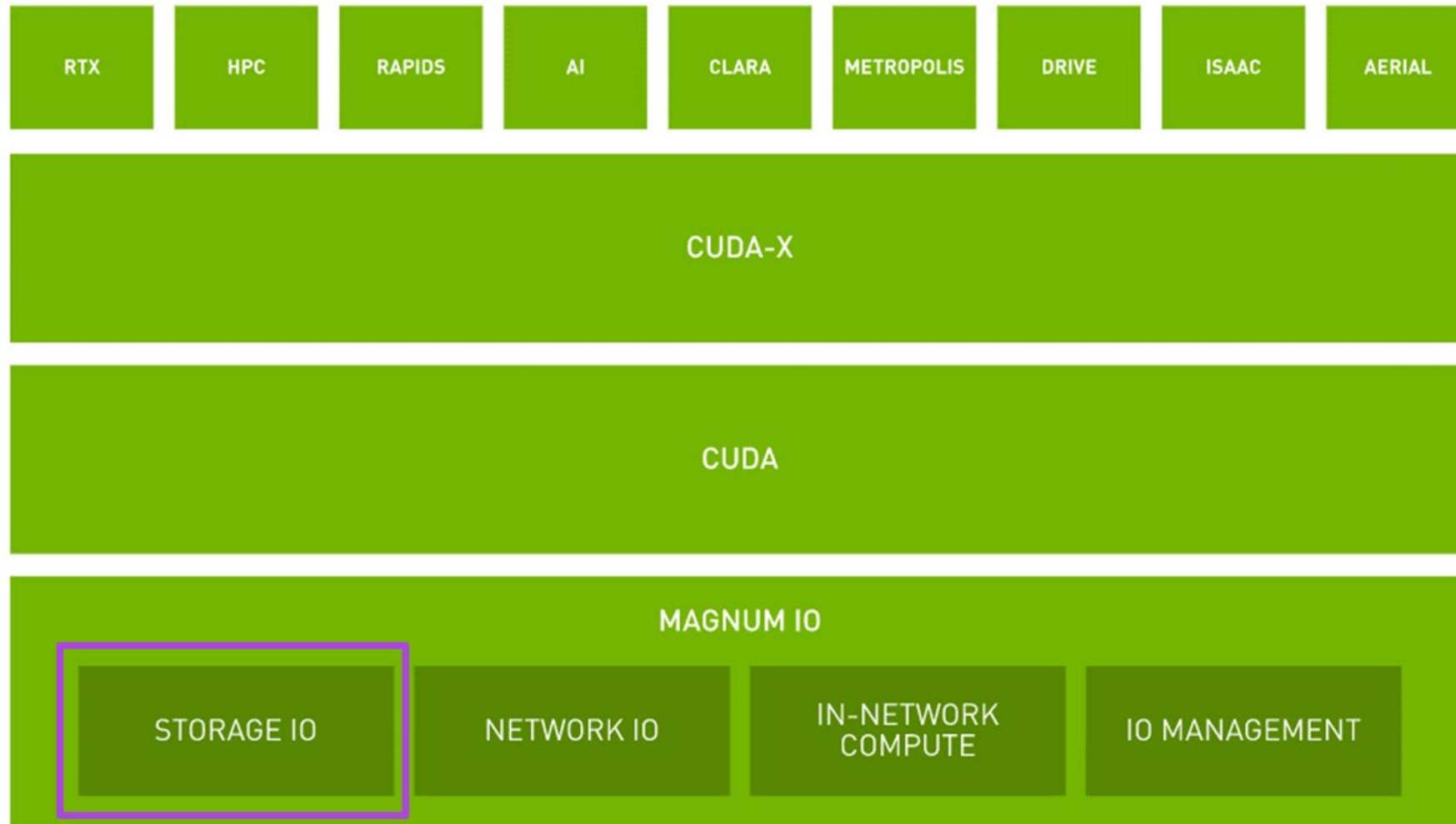
\*\* Supports one 2x200G SOC RNIC connected with x16 PCIe Gen4

\*\*\* Supports three 200G Host connected Ethernet ports





# Magnum IO (MIO)



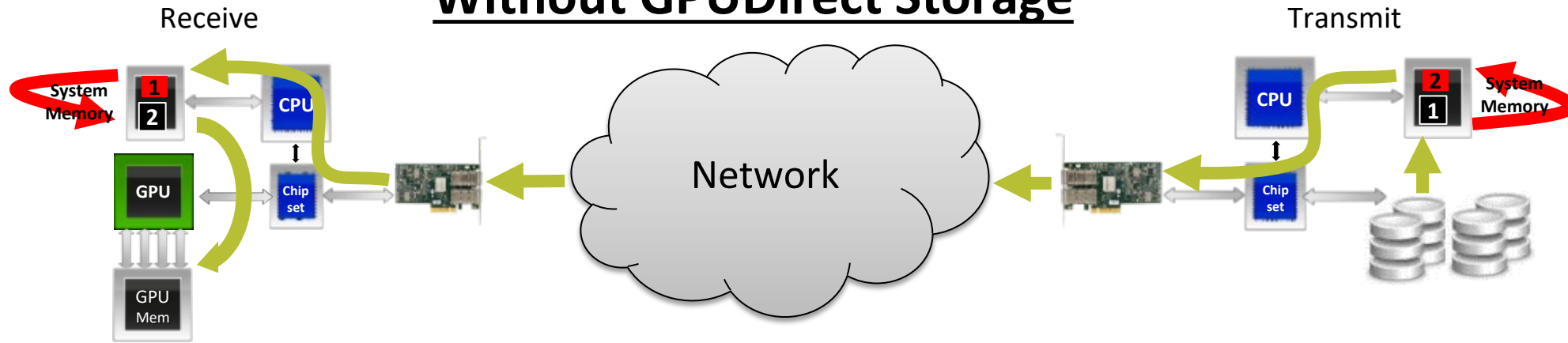
<https://www.nvidia.com/en-us/data-center/magnum-io/>



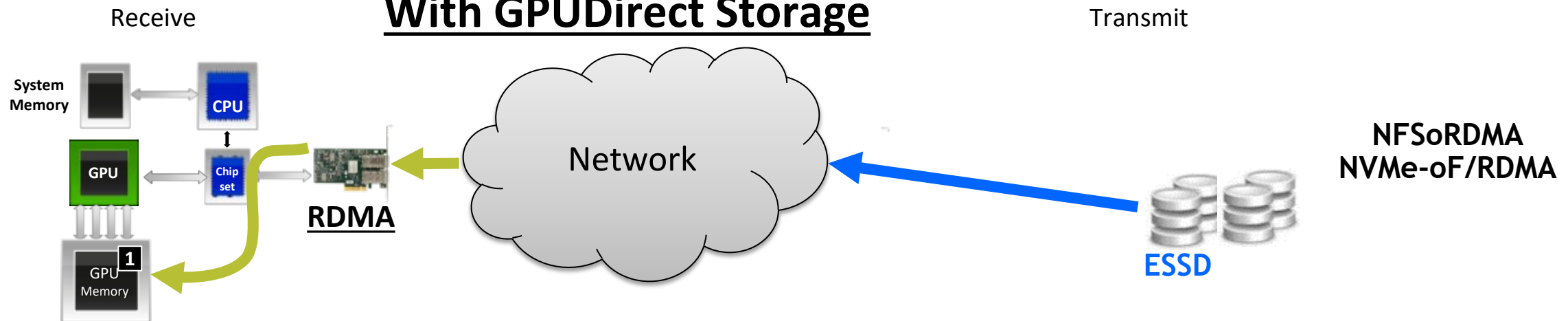


# MIO - GPUDirect storage (GDS)

## Without GPUDirect Storage



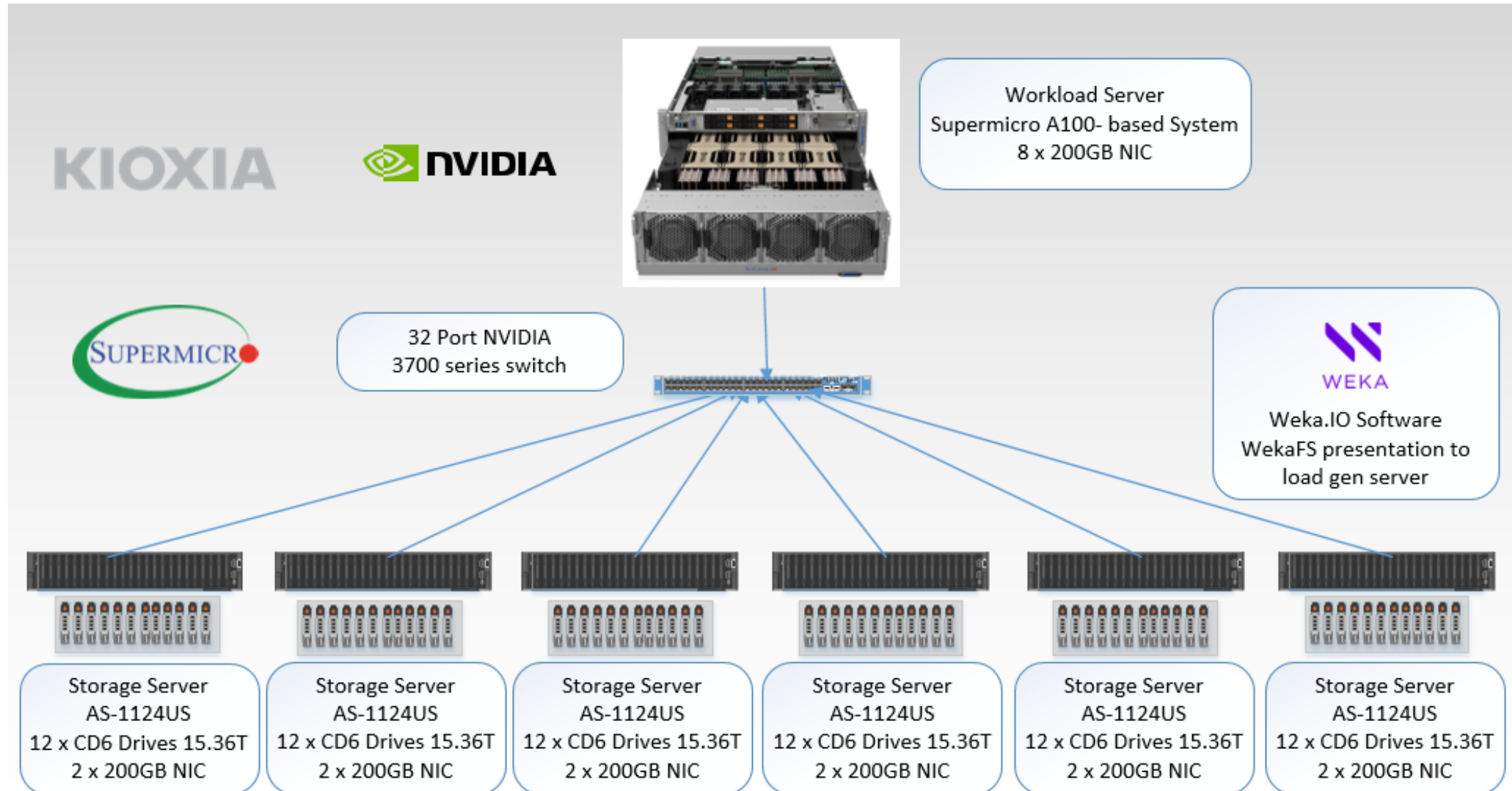
## With GPUDirect Storage



<https://info.nvidia.com/gpudirect-storage-webinar-reg-page.html?ondemandrgt=yes>



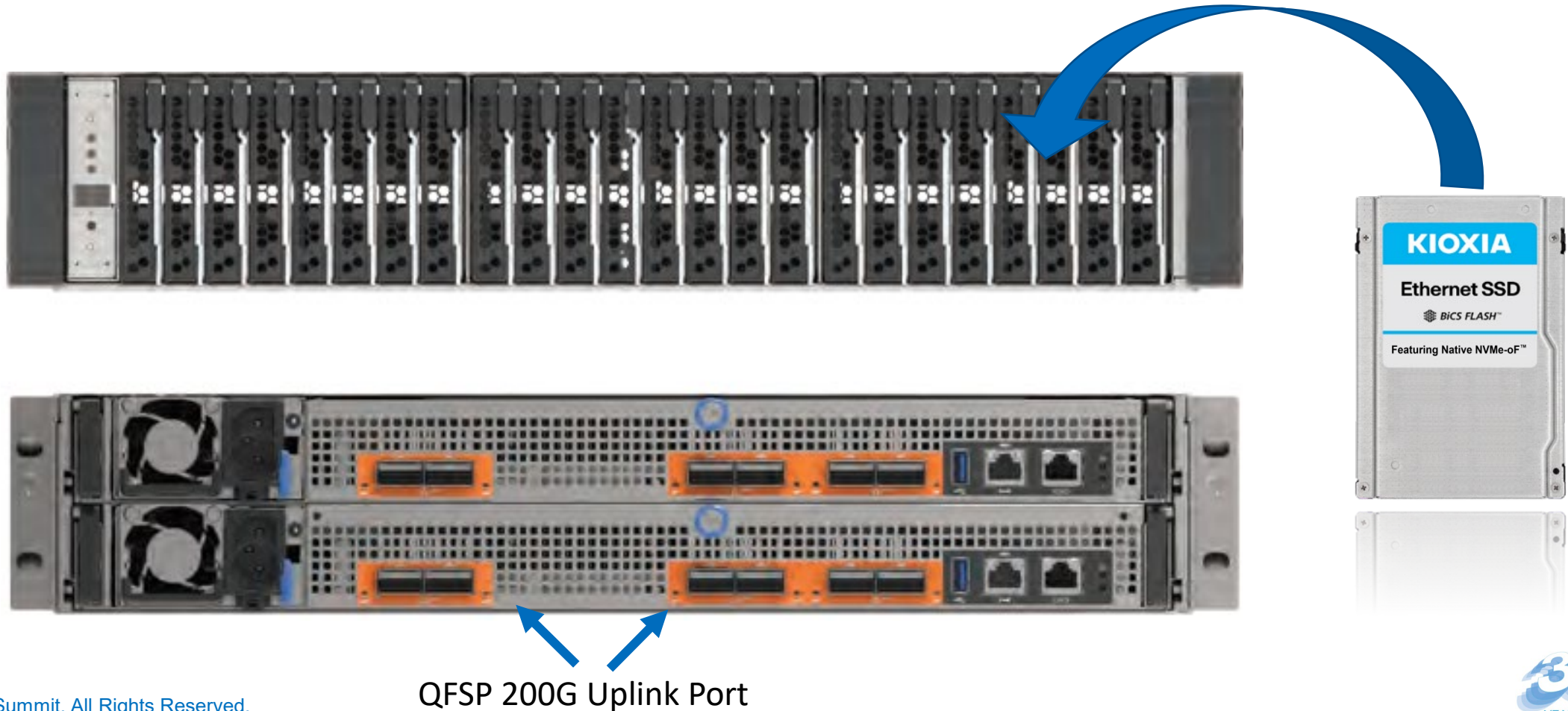
# 1PB of Flash Tech Tips





# EBOF System Specifications

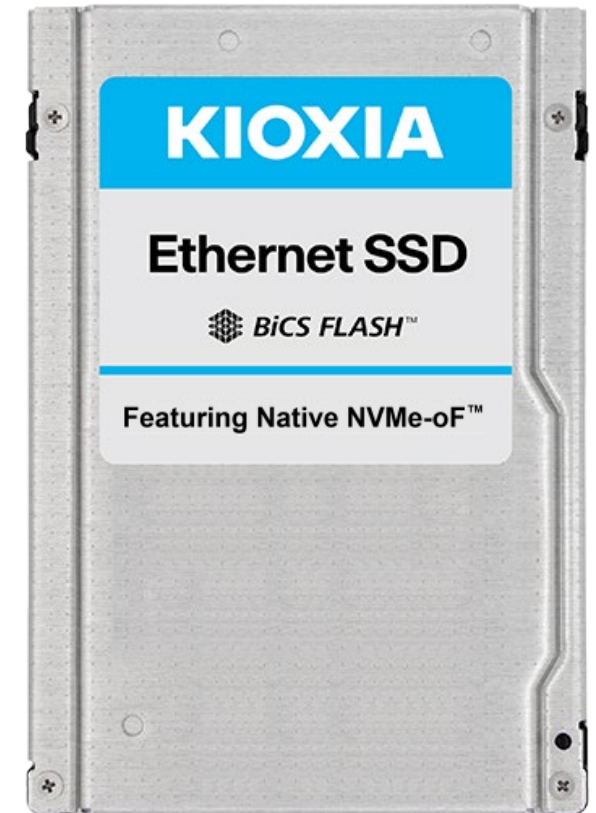
- 6x 200Gbps high speed network capability
- High performance: 830K IOPS per drive, 20M IOPS per 24 bay EBOF (@4KB random read)





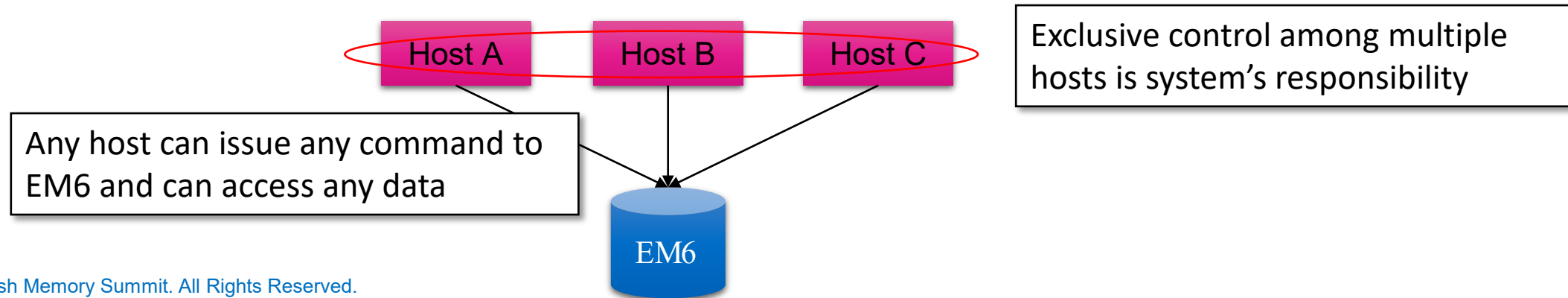
# EM6 SSDs

- World's 1<sup>st</sup> 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 Fabrics-based solution
  - NVMe over Fabrics 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 / 3840 / 7680 GB
  - 1 DWPD



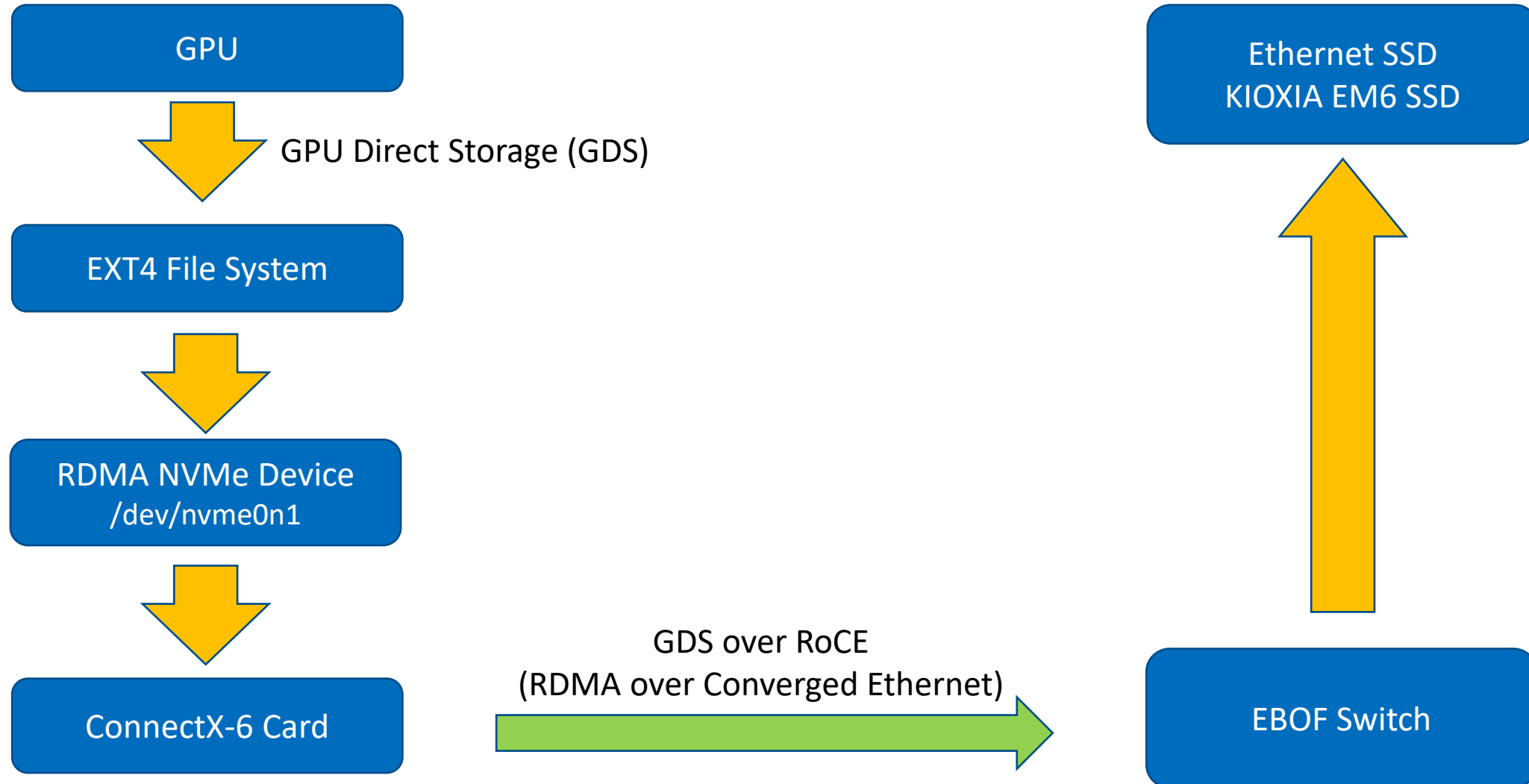
# NVMe-oF Functions

- All NVMe / FTL functionalities are taken over from CM6/CD6
- Fabrics commands are newly added for NVMe-oF operation
  - Connect / Disconnect / Property Set / Property Get
- All NVMe Admin commands and NVM commands are transparent to main SoC via bridge SoC
  - For ES at least, TBD for CS with further security measures leveraging DMTF Redfish
  - ES has no SSD-level exclusive control for multi-host accesses including Admin command - it is supposed to be system's responsibility
    - Any host can issue any Admin command
    - Reservation is not supported
    - Max number reports such as Abort Command Limit must be shared among all hosts





# Data Path for GDS

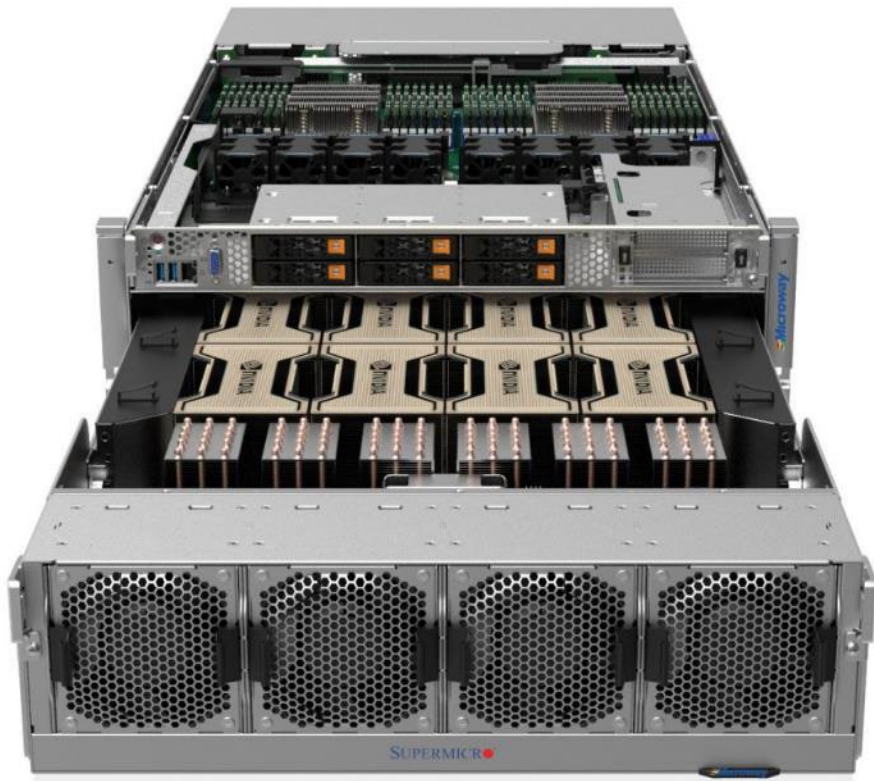






# Layout of Systems

- HGX A100 System as Initiator



- EBOF Chassis for Switching, Carriers and Interposer



- EM6 Ethernet SSD as Target







# Performance Data

- Theoretical maximum of Single 25Gb/s SSD is 3GB/s
- Maximum of a 100Gb/s network connection is 12.5GB/s

# of Drives	XferType	Iosize	IoType	XferMode	TotalThreads	Throughput(GiB/sec)	CPU_USR(%)	CPU_SYS(%)	CPU_IRQ(%)
1	0	64KiB	READ	GPUD	256	2.608061	0.54	0.25	0
1	0	128KiB	READ	GPUD	256	2.631226	0.31	0.16	0
1	0	512KiB	READ	GPUD	192	2.653799	0.17	0.23	0
1	0	1024KiB	READ	GPUD	192	2.650515	0.1	0.26	0
1	0	4096KiB	READ	GPUD	192	2.649202	0.06	0.25	0

# of Drives	XferType	Iosize	IoType	XferMode	TotalThreads	Throughput(GiB/sec)	CPU_USR(%)	CPU_SYS(%)	CPU_IRQ(%)
4	0	64KiB	READ	GPUD	256	10.582392	2.28	1.06	0
4	0	128KiB	READ	GPUD	256	10.587423	1.26	0.7	0
4	0	512KiB	READ	GPUD	192	10.608474	0.48	0.55	0
4	0	1024KiB	READ	GPUD	192	10.615657	0.35	0.52	0
4	0	4096KiB	READ	GPUD	192	10.588682	0.27	0.66	0



# Why Ethernet SSD For AI?

- GPU Direct Storage uses NVMe and RDMA to connect to storage
- Ethernet SSDs connect via the RDMA protocol natively
- Scale out storage has extra cost due to CPU & Server HW
- Ethernet SSD solutions create a native path for the GPUs to access and process data over an extremely fast network link without copying or moving data