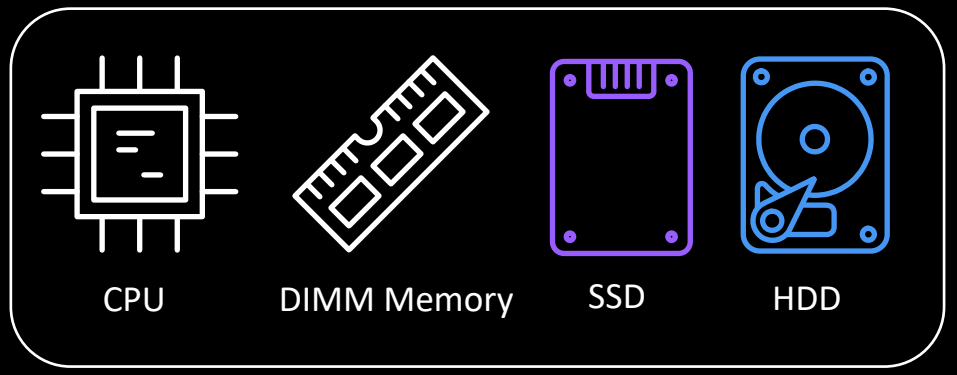
Storage for AI Using GPUs

Leveraging GPUDirect and Western Digital RapidFlex ™ NVMe-oF ™ Controllers to Saturate GPU Bandwidth

The Accelerated Computing Revolution

Traditional Server

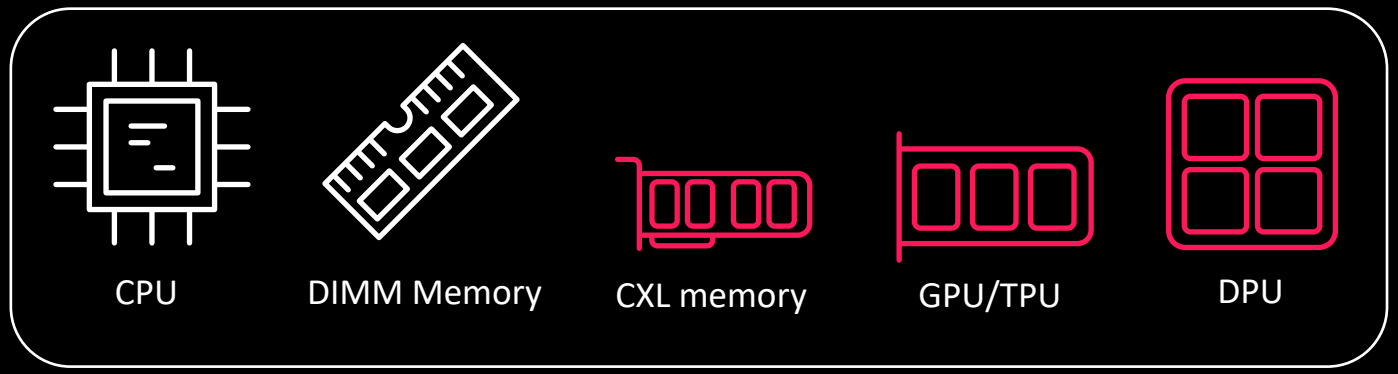


Max 1.5-2kW per 2U server

Server Trends

- CXL enables scalable memory
- GPUs power AI workloads
- DPUs offload CPUs for other specific workloads
- Power density is going up

Accelerated Computer



<10us latency domain

Disaggregated Storage

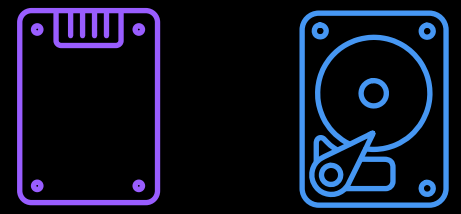
100us latency

10ms latency

NVMe-oF

600W for 24 SSDs

1000W for 100 HDDs



Accelerated Computing Considerations

Cloud-Based ML Challenges

Cost

- **Subscription and Usage Fees**:
  - Pricing models can be complex, involving pay-per-use or tiered services, which can escalate quickly with increased usage
- **Data Transfer Costs**:
  - Moving large datasets into and out of the cloud can incur substantial costs
- **Operational Costs**:
  - Ongoing operational costs for cloud services may include data storage, compute time and additional services like data transformation or transfer

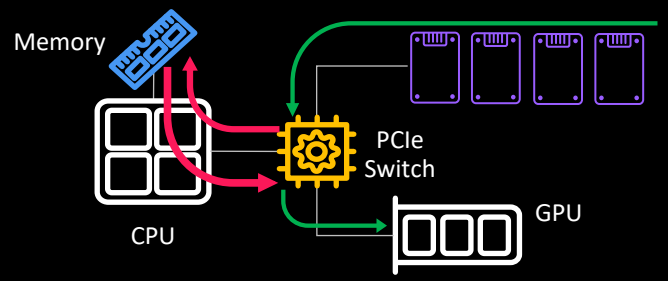
Performance

- **Latency:**
  - Network latency can affect the performance of cloud-based ML models, especially in real-time applications
- **Computational Limits:**
  - More computational power than what is allocated may be required, leading to performance bottlenecks
- **Resource Contention:**
  - Shared resources in the cloud can sometimes lead to contention, impacting performance

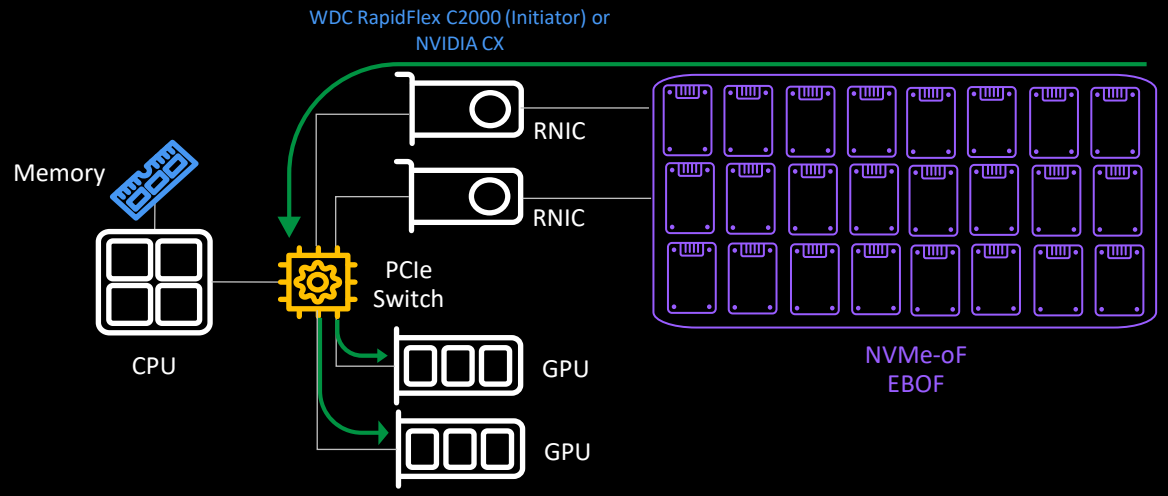
GPUDirect Storage (GDS) with Disaggregation

Using NVIDIA GPUDirect and Western Digital Disaggregated Storage

- Without GDS, GPUs directly read ML data from local SSDs via the CPU complex
- Limits GPU performance and scale



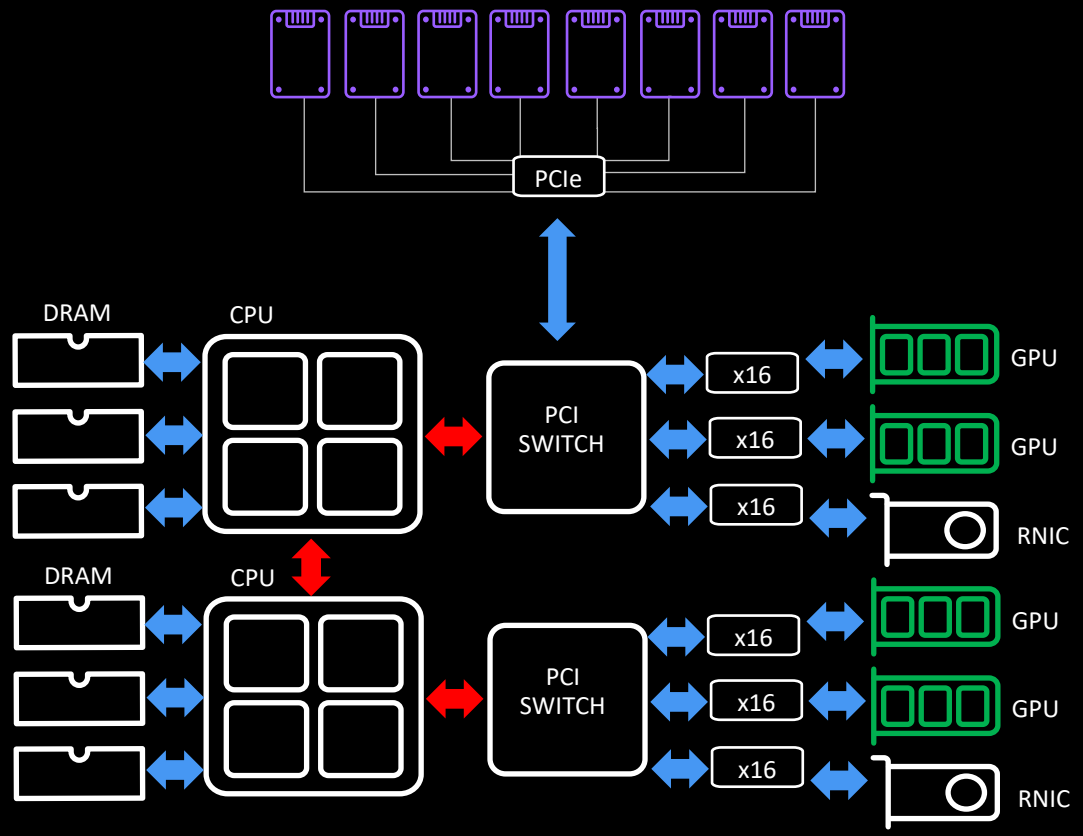
- With GDS: GPUs have a direct path for data exchange instead of going through the CPU
- RapidFlex makes NVMe-oF disaggregated storage look like local NVMe storage
- Allows for linear performance and storage scale



GPUDirect Storage (GDS) Architecture

On-Premise Architectural Considerations

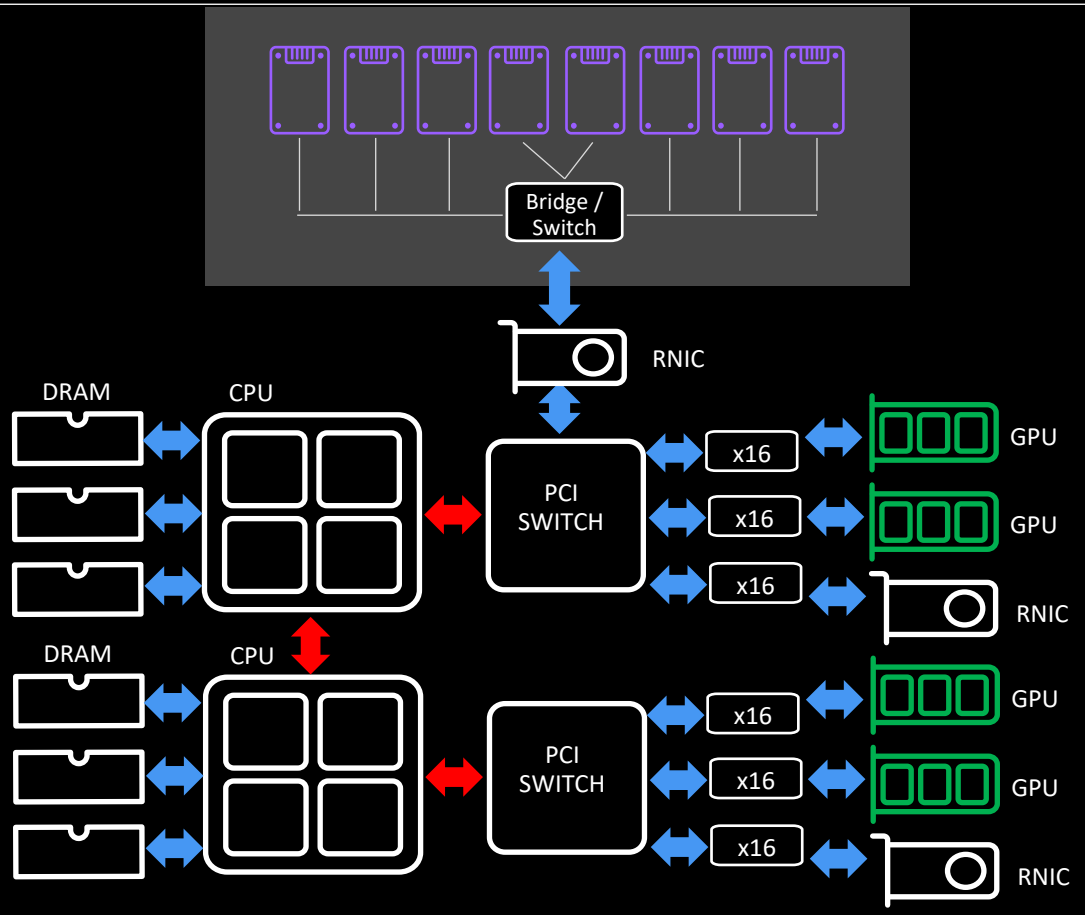
- Many server platforms limit NVMe to GPU access over PCIe with inadequate architecture
  - In-depth platform analysis required for optimal design considerations:
    - i.e. Performance, Storage, Power and cooling
  - It can take 12 to 16x Gen4 NVMe SSDs to saturate a GPU (H100)
  - Local NVMe drive slot availability can limit performance and total ML data set capacity
    - Results in inefficient use of expensive GPU
    - Impacts ML project timeline and cost
    - Impacts infrastructure scalability



Disaggregating Storage

Benefits of Moving Storage Outside the Accelerated Server

- Move SSDs to their own chassis
- Add or use existing RNIC in the server
- Utilize NVMe-oF standard
- Removes 500 – 1000W from the server
- Simplifies sharing SSDs and data
- Enables server upgrade while maintaining existing SSD value
- Independent scaling of compute and storage



AI Demo FMS 2024

OpenFlex™ Data24 - 4200

NVIDIA

H100

H100

x16

x16

PLX Switch Gen5

x16

x16

x16

CPU

CPU

PLX Switch Gen5

x16

x16

x16

GPU Server

ConnectX-7

200 GbE DAC Cables

(12) 100 GbE DAC Cables

NVIDIA SN3700

32-Port 200GbE

Spectrum 2

Western Digital

OpenFlex™ Data24 4200

(24) Ultrastar DC SN655 15.36TB NVMe™ SSD

Ingrasys ES2100

RapidFlex™ C2110 Interposer

NVIDIA

A100

A100

x16 x16

PLX Switch Gen4

x16 x16 x16

CPU

CPU

PLX Switch Gen4

x16 x16 x16

x16 x16

A100

A100

200 GbE DAC Cables

ConnectX-7

Western Digital

RapidFlex Interposer

NVIDIA

SN3700

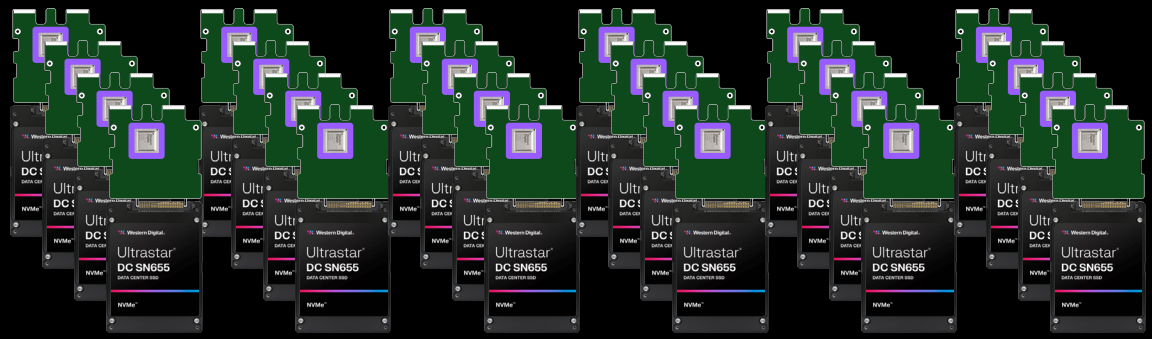
32-Port 200GbE

Spectrum 2

(12) 200 GbE DAC Cables

Ingrasys®

24-Bay ES2100 (Spectrum 2)

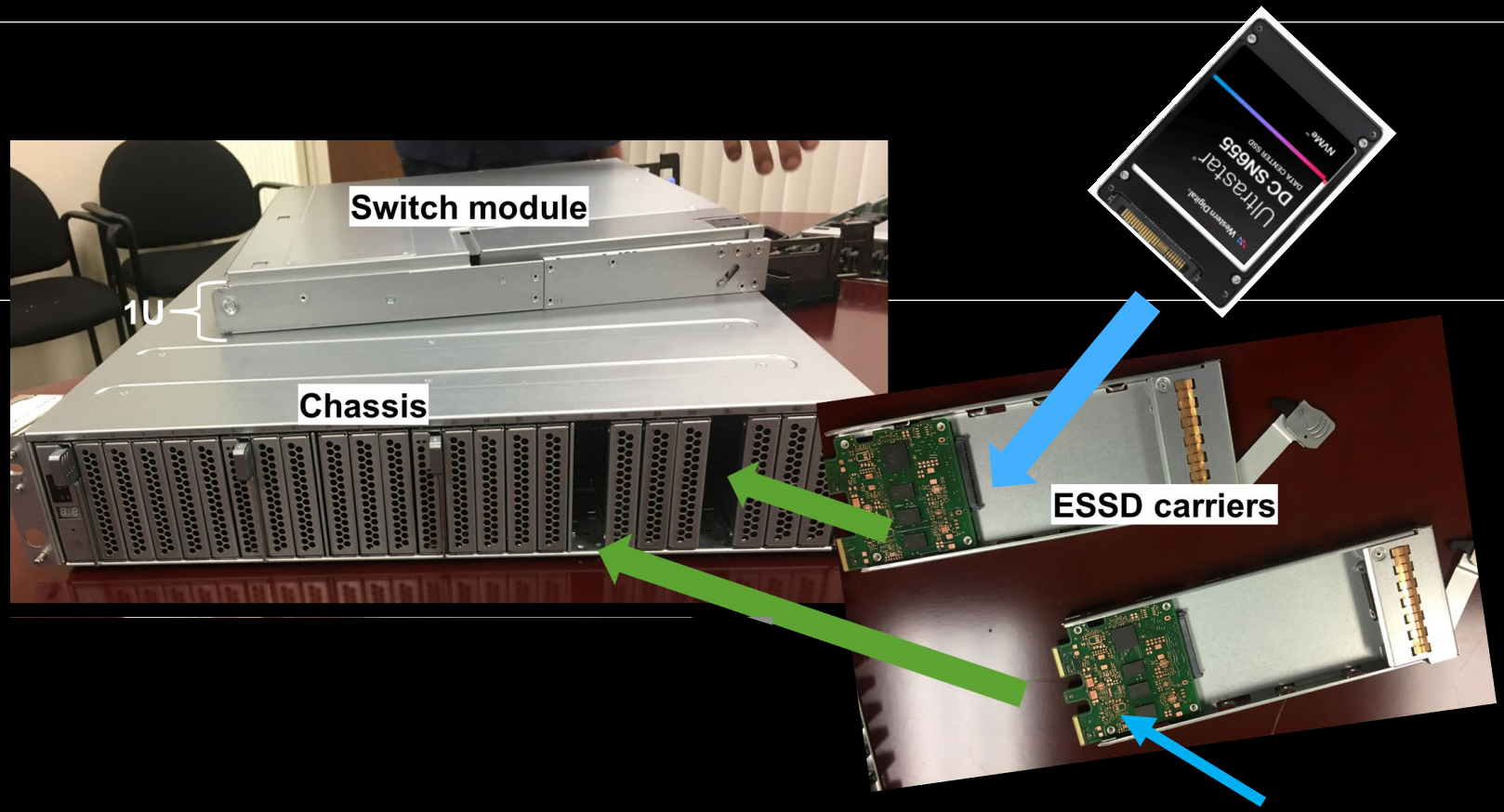


(24) Ultrastar DC SN655 15.36TB NVMe™ SSD

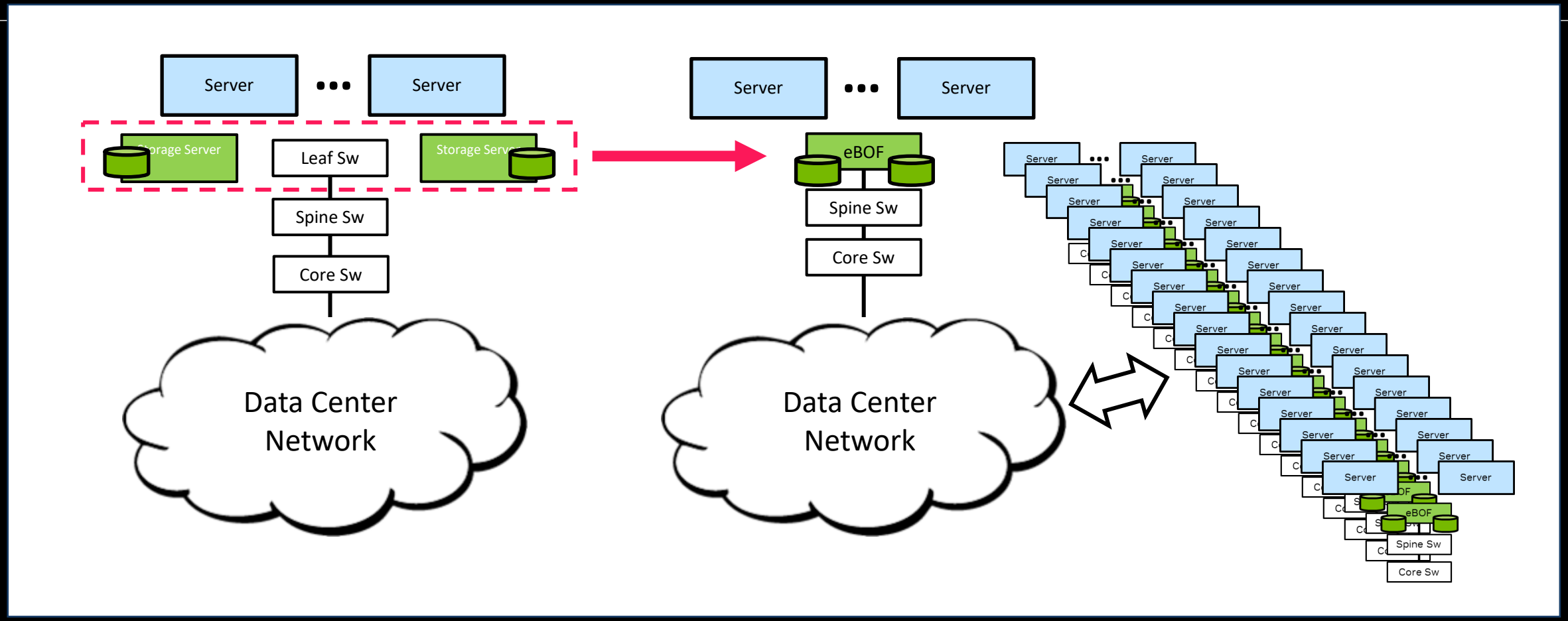
What an EBOF looks like - Ingrasys ES2100

- 12 200GbE external ports
- 24 U.2 SSD 2x50GbE WDC Interposers – 770TB with 32TB SSDs
- NVMe-oF/RDMA or IP Protocol
- Performance = Interposer + SSD + 600ns latency

What's Inside

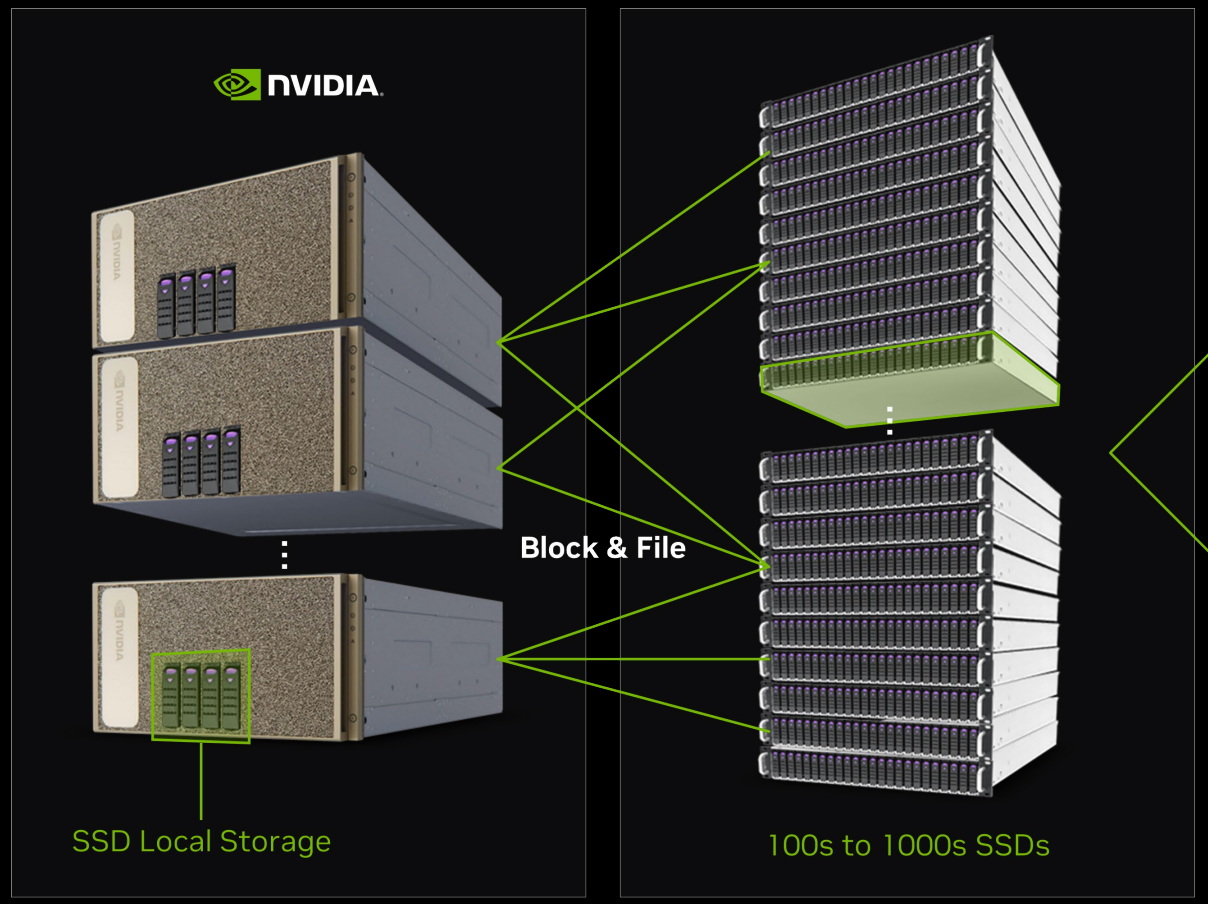


EBOF Can Replace Data Center Leaf Switches



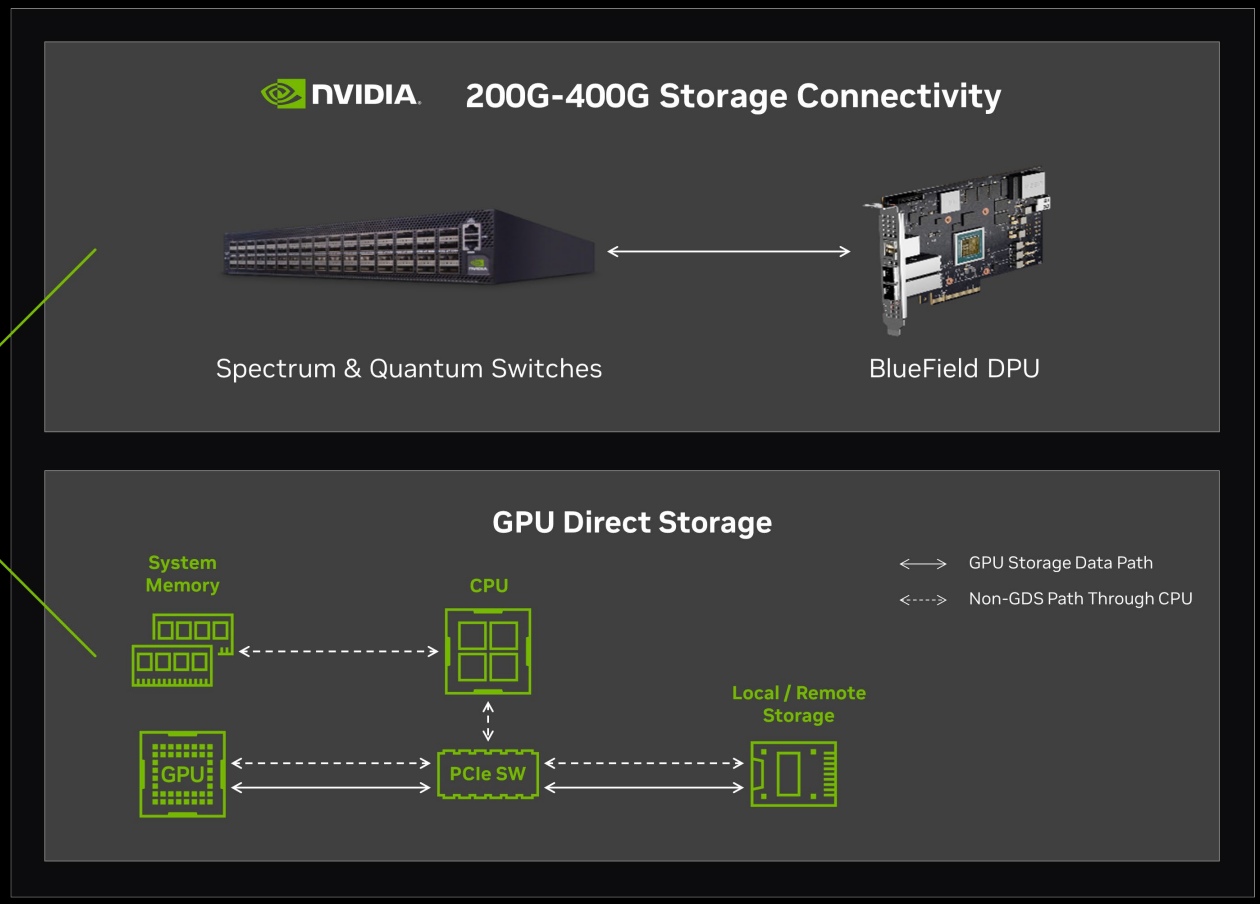
EBOF provides cost, power and space savings at data center scale

GPU Initiated-Storage and EBOF

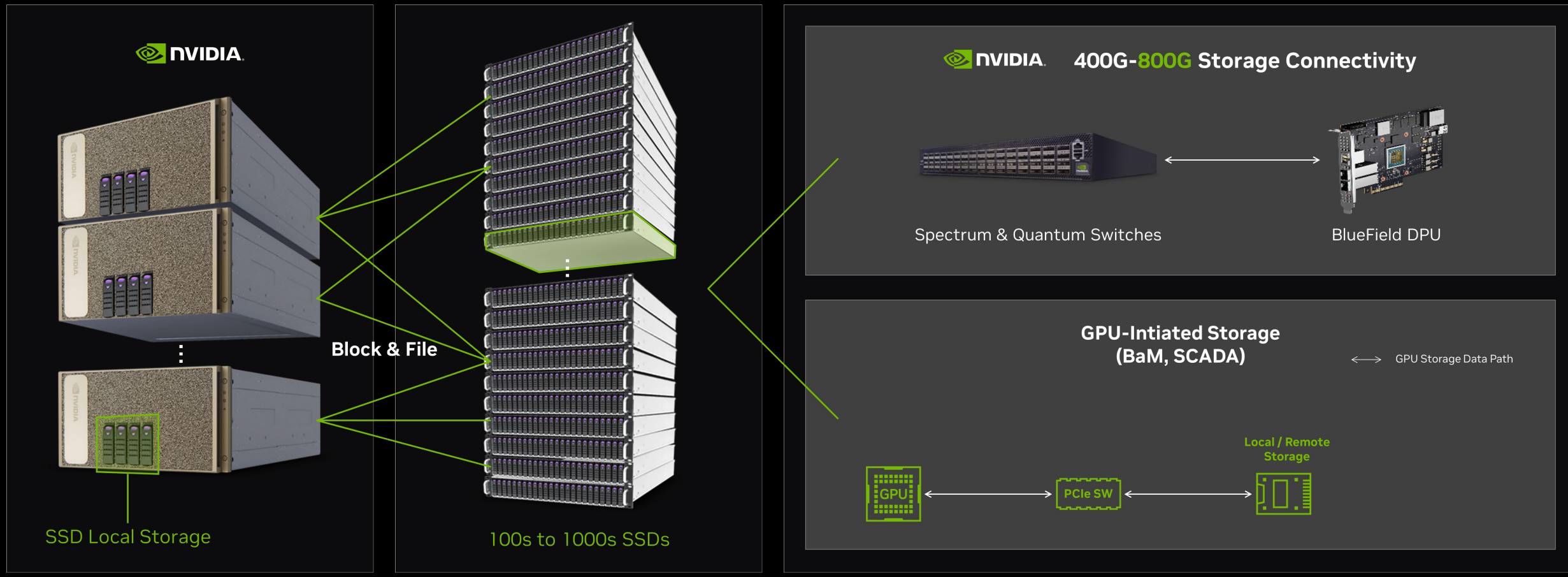


GPU Servers

Networked Scale Out Storage
Network Storage (Expandable)



GPU Initiated-Storage and EBOF

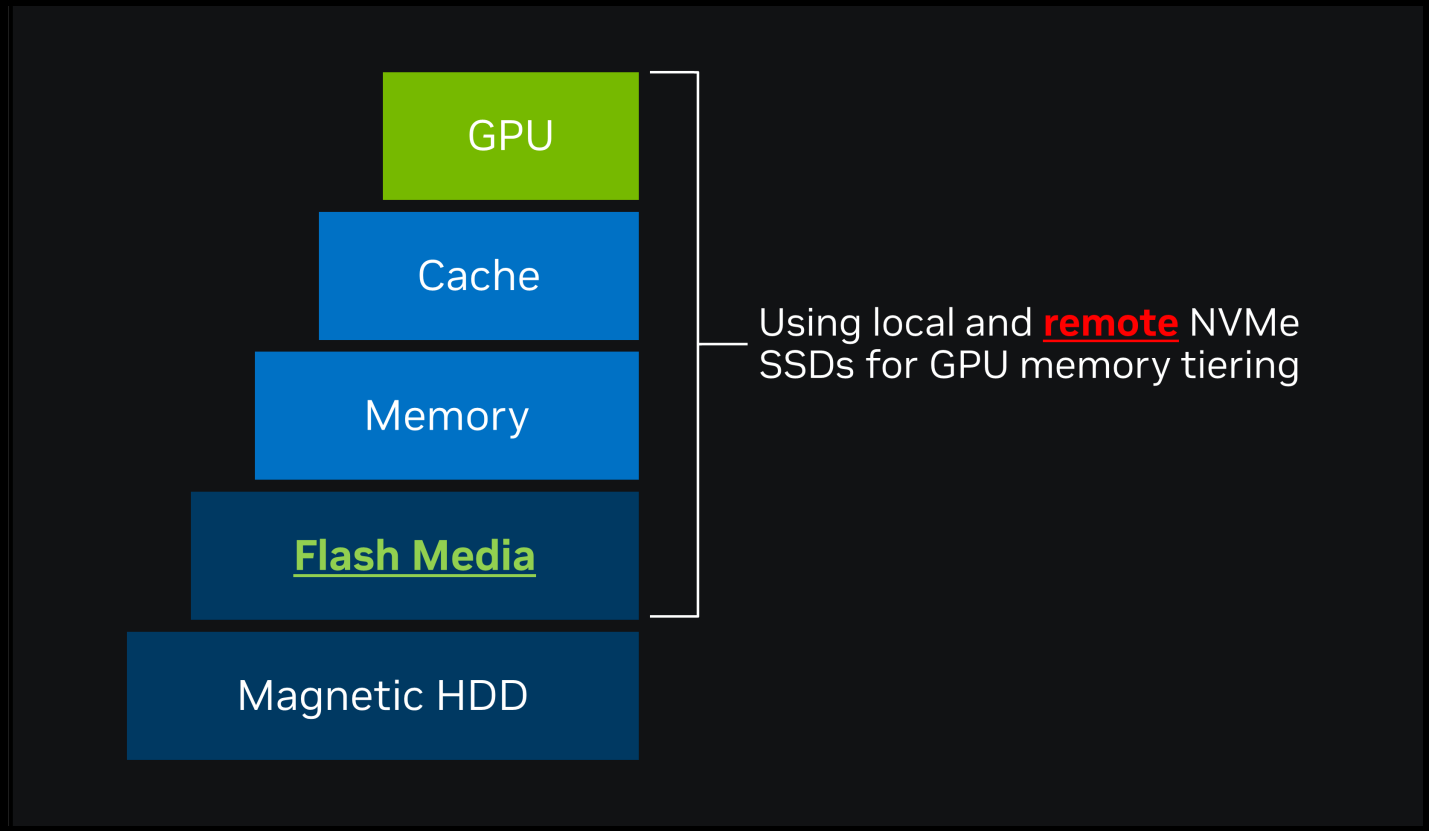


GPU Servers

Networked Scale Out Storage
Network Storage (Expandable)

GPU Initiated-Storage and EBOF

SSDs as a Memory Tier



NVIDIA Index Demo FMS 2024

Description - Visualizing the World's Most Violent Tornadoes

GPUDirect Storage: ON

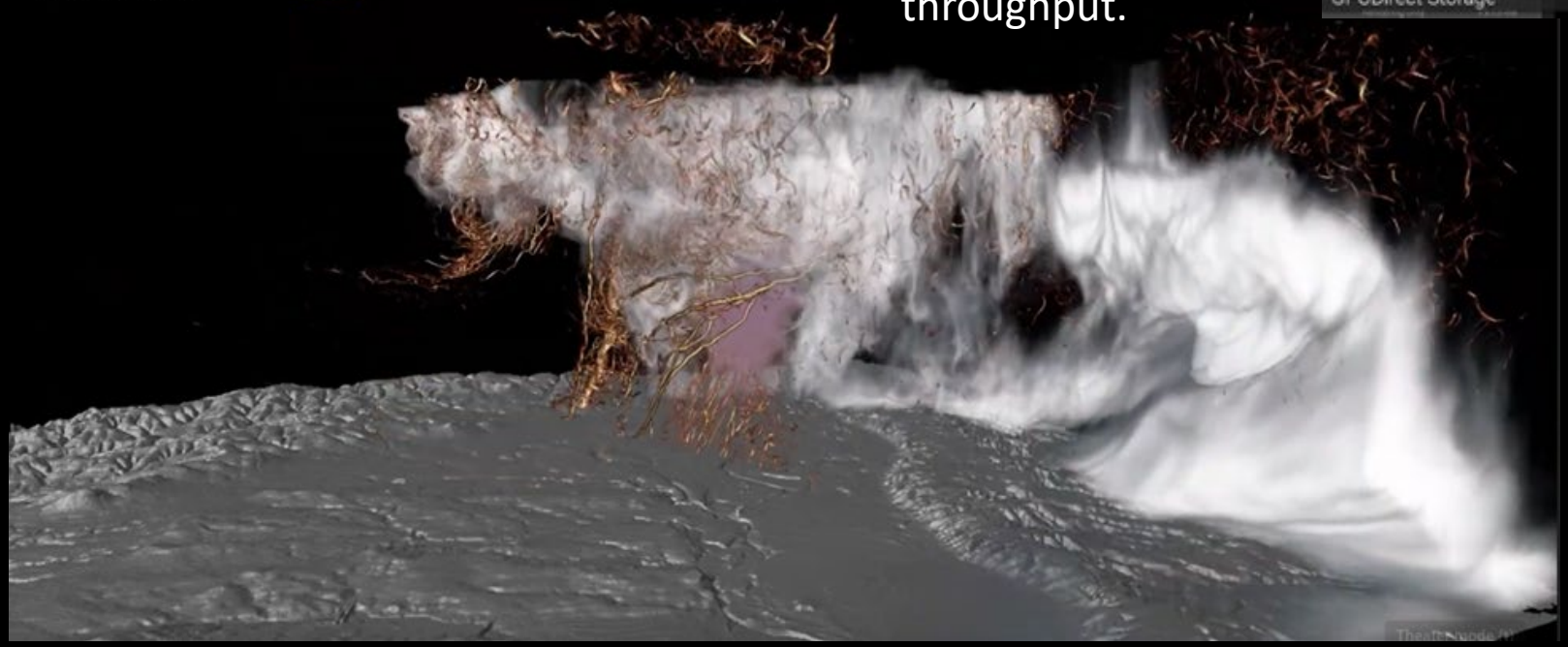
13 FPS with up to 89 GBS of read throughput.

Approximates to 5.9TB of dataset ingest every 65 seconds.

Allows for interactive navigation, on-the-fly parameter adjustments and scrubbing through the simulation with ease.

GPUDirect Storage: OFF

4 FPS with up to 15 GBS of read throughput.



250 billion grid points, each with over a dozen attributes such as rain, hail, pressure and wind speed

AI Demo Objectives FMS24

See This Demonstrated in Western Digital Booth 607

Fast	GPUDirect Storage enables a direct path between NVMe-oF storage and GPU memory.
Scalable	RapidFlex disaggregates NVMe storage and GPU resources to independently and predictably scale ML workloads.
Interoperable	RapidFlex complies with NVMe and NVMe-oF standards to deliver flexibility and choice between compute, GPU and NVMe storage.