

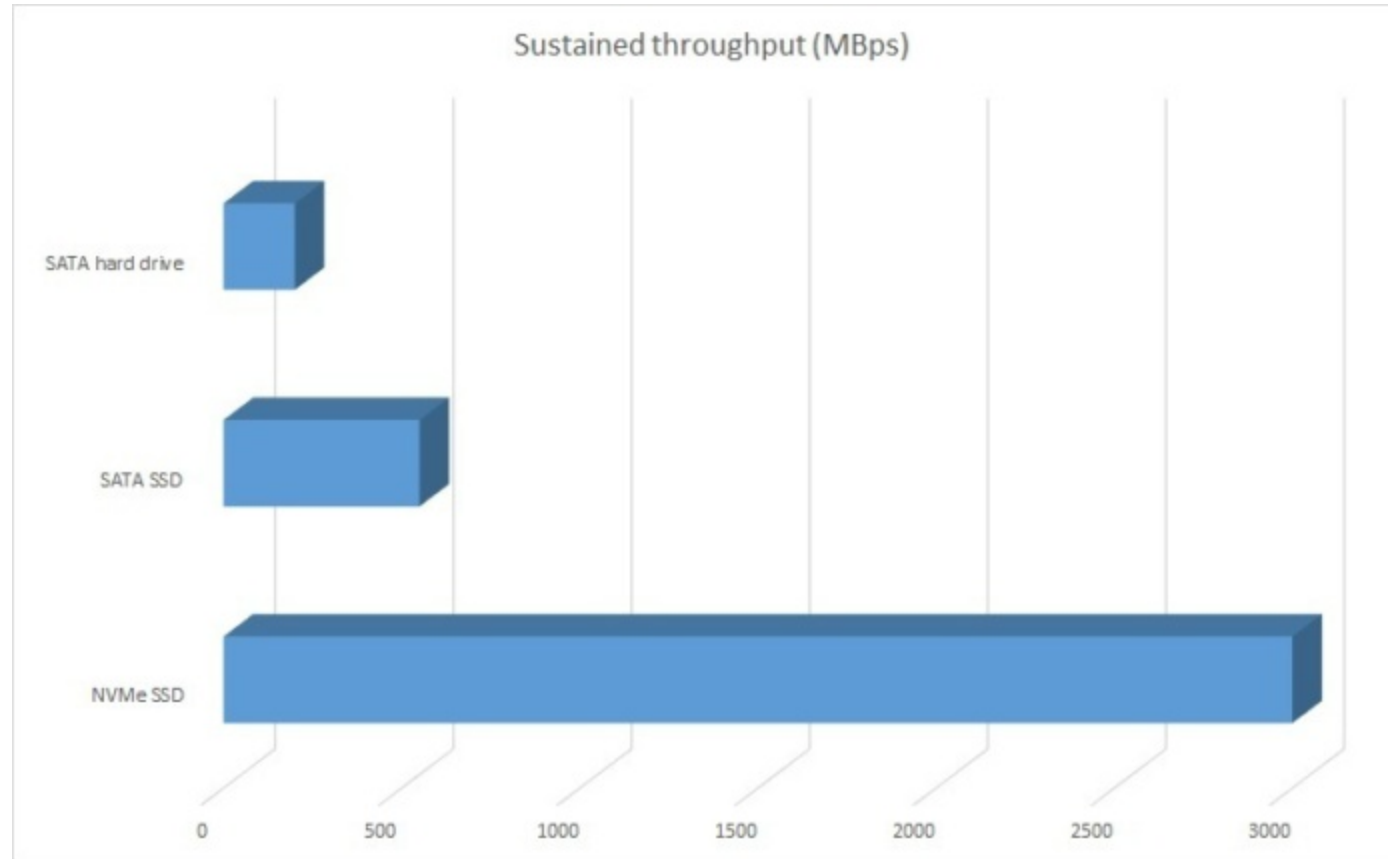
Interoperability as the new frontier



somewhere on the internet, April 24th 2020

Jean-Thomas Acquaviva,
jtacquaviva@ddn.com


Media Performance Evolution



Courtesy Jon L. Jacobi at PCWorld.com

Systems Performance Evolution: IO500

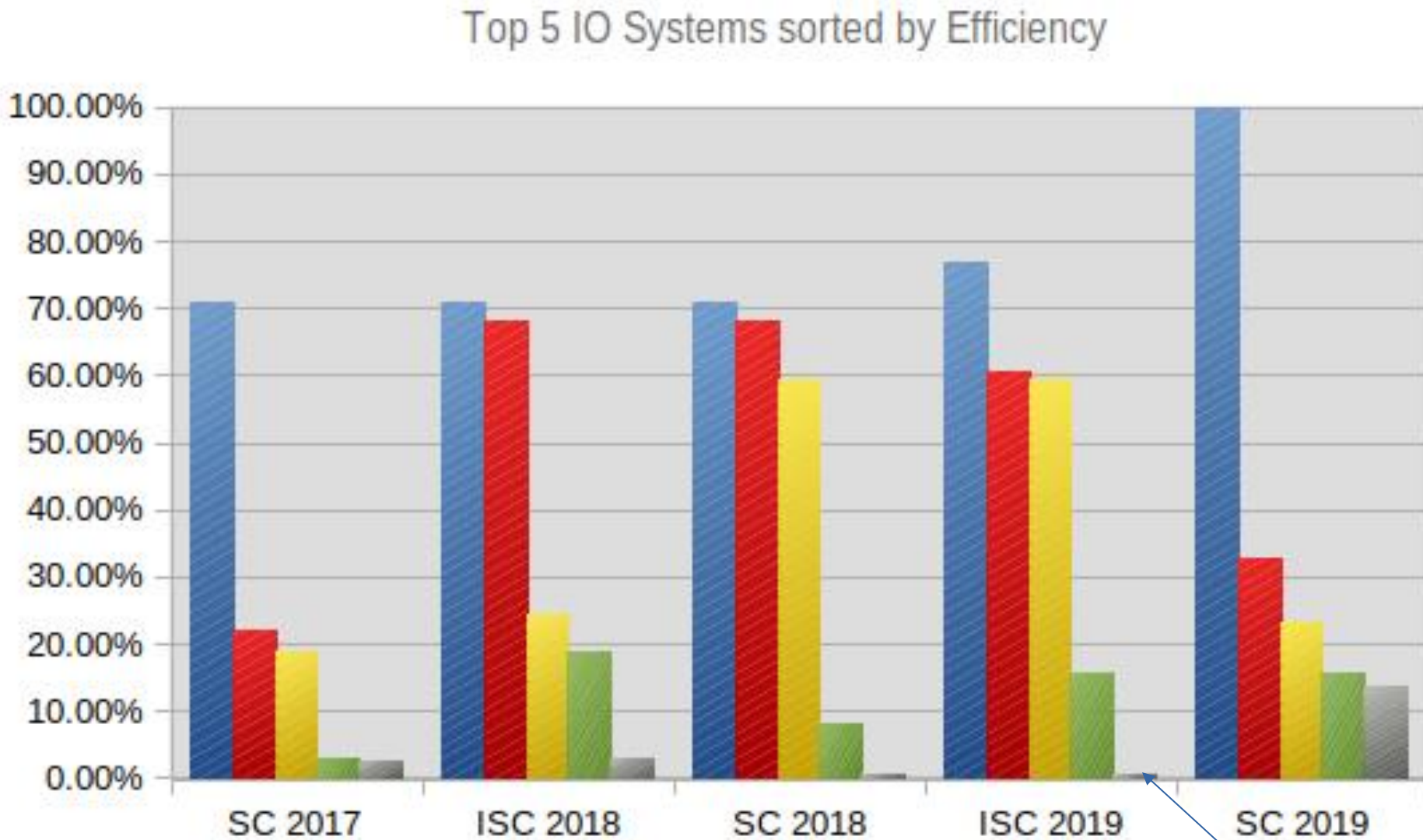
Full List

This is the full list from  [Supercomputing 2019](#). The list shows all submissions.

IO500

#	information			io500				ior			
	list id	system	data	score	bw	md	tot iops	easy write	easy read	hard write	hard read
					GiB/s	kIOP/s	kIOP/s	GiB/s	GiB/s	GiB/s	GiB/s
1	sc19	WekaIO on AWS	zip	938.95	174.74	5045.33	2770.21	298.13	325.62	71.57	134.20
2	sc19	Wolf	zip	933.64	183.36	4753.79	2705.53	108.24	308.91	122.91	275.09
3	sc19	Wolf	zip	516.41	123.89	2152.46	1399.89	108.31	185.12	93.18	126.12
4	sc19	Tianhe-2E	zip	453.68	209.43	982.78	988.79	608.64	613.53	31.87	161.64
5	sc19	DGX-2H SuperPOD	zip	249.50	86.97	715.76	597.16	273.16	286.17	11.16	65.59
6	sc19	Data Accelerator	zip	229.45	131.25	401.13	465.59	341.24	383.01	14.14	160.55

Observation #1 NMVe as a media as delivered

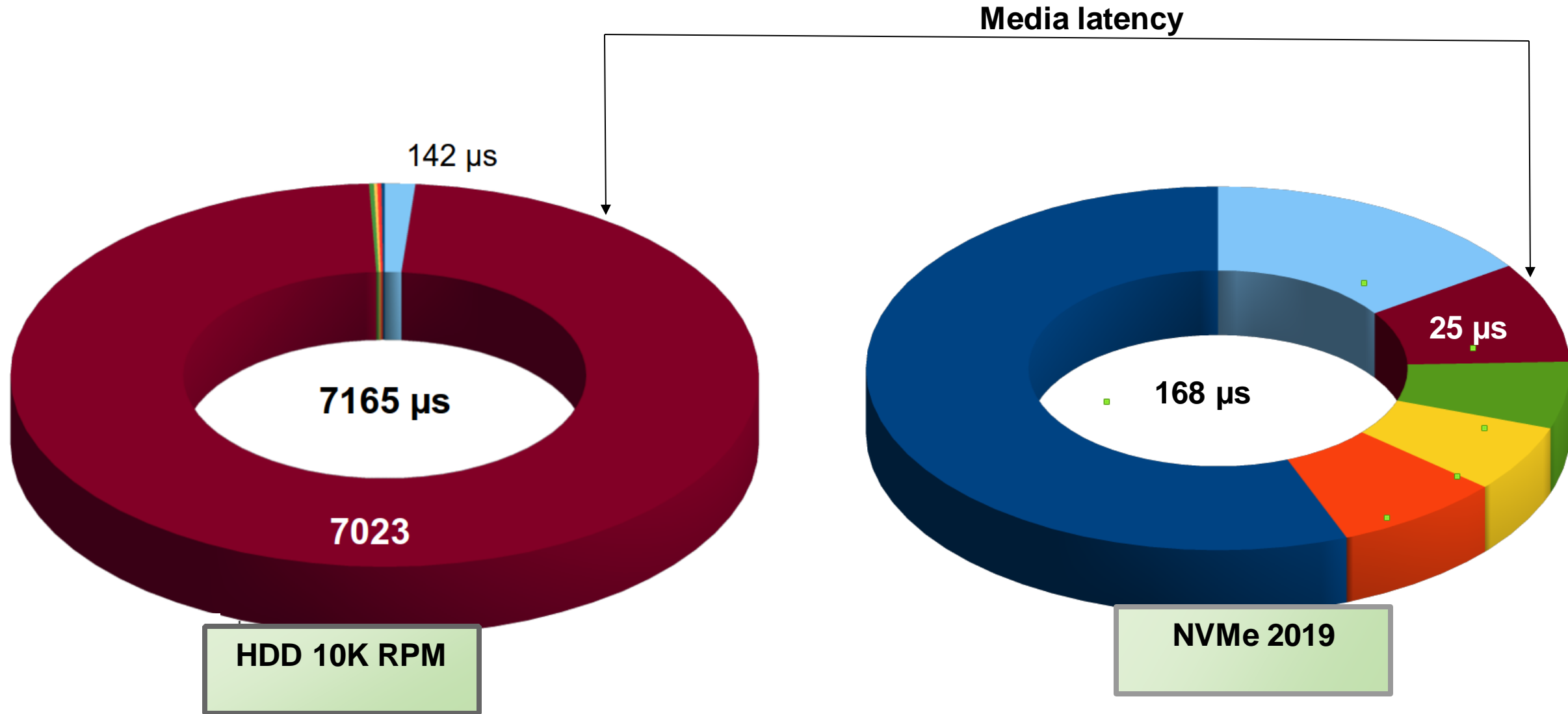


Claim #1 Bandwidth battle is (almost) over

▶ Thanks to new storage media the bandwidth problem is now closed

- If you have pockets deep enough, we know how to build an **efficient** multi TB/sec system
- We deploy TB/sec range flash systems all over the world

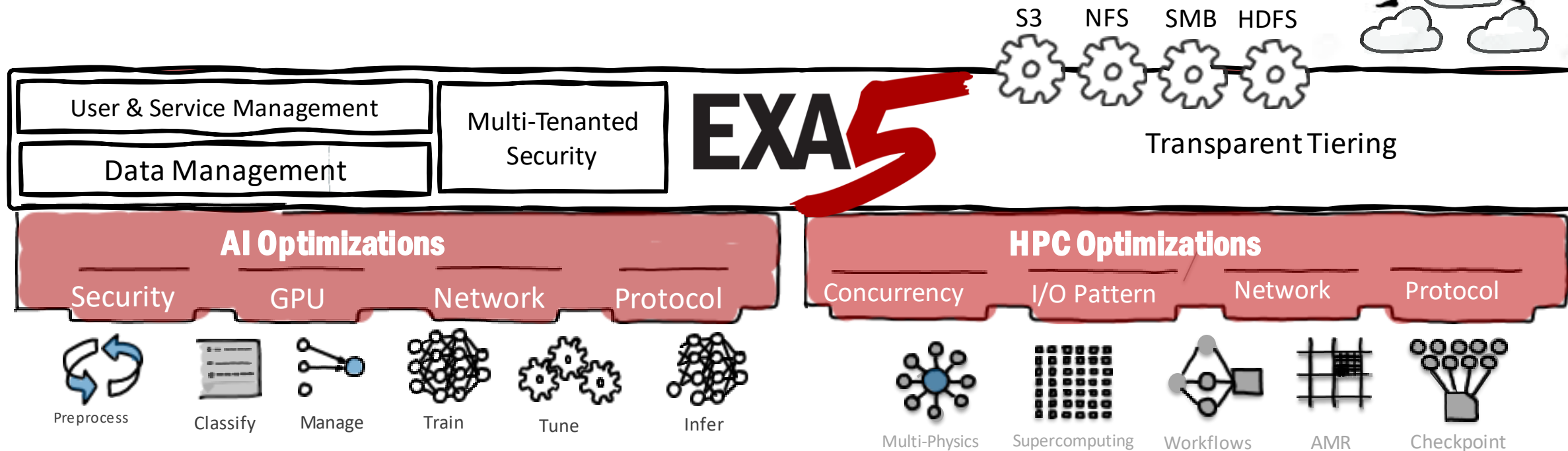
Comment #1 Frontline is now on Latency



Interoperability between Media

- ▶ **Internal data management (tiering) with EXA5**
 - Introduction of hot pools
- ▶ **Decoupling NVMe from Capacity with IME**
 - Use cases for on-prem and in the cloud
 - Highlight of the significant benefits for IME in the cloud
- ▶ **Multi-cloud**
 - Dealing with data gravity in the era of multi-cloud
- ▶ **Data management at scale**

EXA5 - Stretching from Capacity to Fast tier



► Hot pool in Exa5

- File based replication from fast to capacity storage
- Eviction based on watermark policy

Stretching further: Disaggregated NVMe



▶ **Considerations for IME over EXA5 Hot Pools**

- Prefetch on read capabilities enable the use of IME to better behave like a traditional buffer and less like a tier of storage
- Flash native: extend life time of device (i.e. consumer grade)

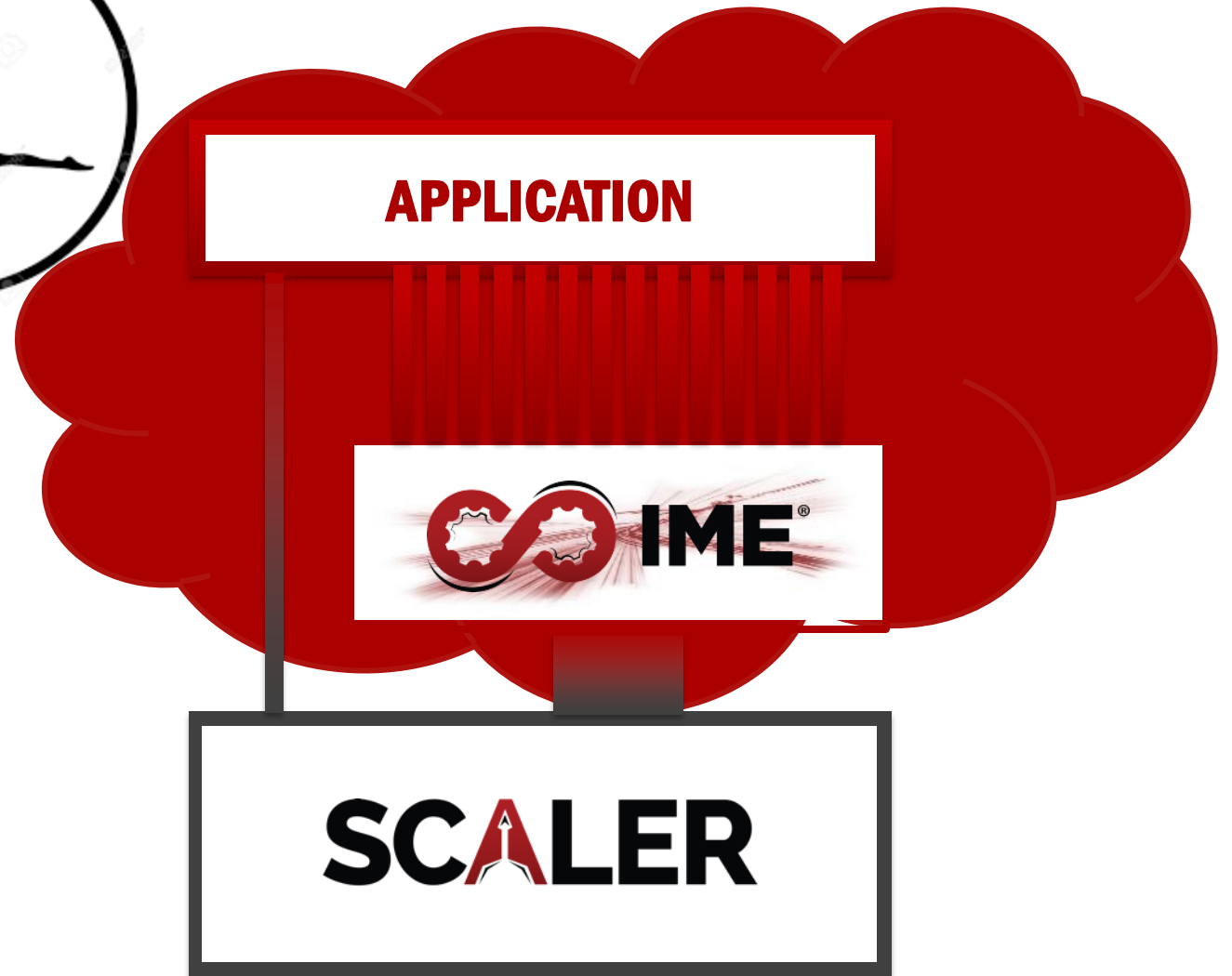
▶ **IME is software defined storage**

- IME Appliances from DDN are superior to other solutions, but IME being software defined provides flexibility
- IME being software defined enables the use of IME in cloud provider environments

▶ **IME simplifies the entry into multicloud**

Jumping to the Cloud

Articulation of on-premise with cloud
or multicloud is the right level of design
For complex workflow.





Data Gravity

Observation #2: Data locality matters

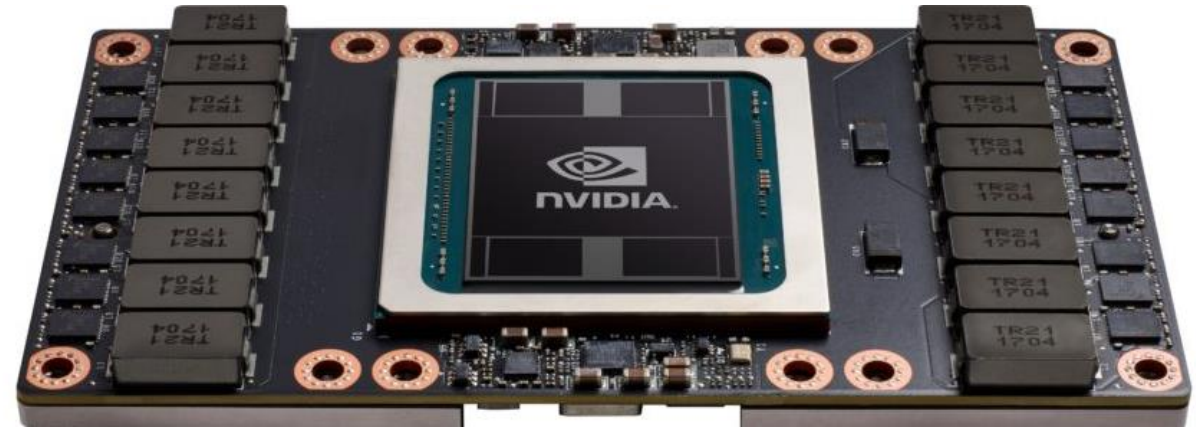
▶ Speed of light: 300.000 Km/sec (300.000.000 m /sec)

▶ 1 m = 3 ns

▶ V100 = 120 TFLOPS (tensor)

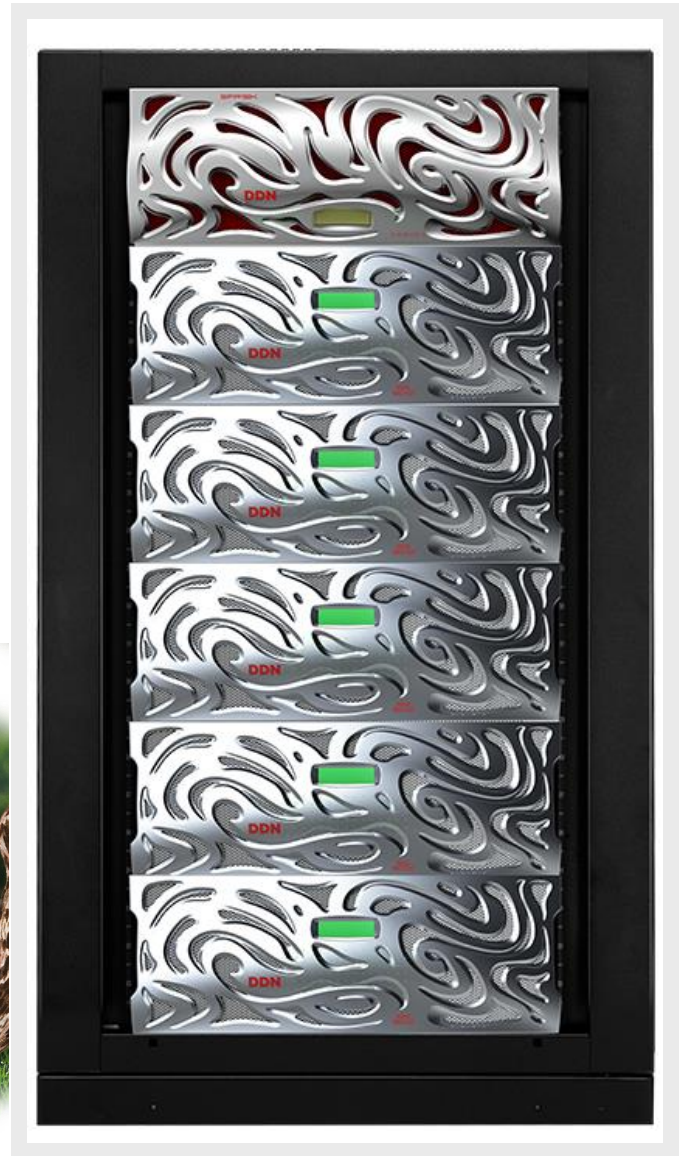
$$120 \cdot 10^{12} / \text{sec} = 120 \cdot 10^3 / \text{ns}$$

Every meter cost 360.000 FLOPS!



Claim #2: Data weight matters

- ▶ 1 PB = 1000 TB = 10^{15} Byte
- ▶ 4 PB = 1 ton
- ▶ Amazon Snowball?!?



Some Cloud Considerations

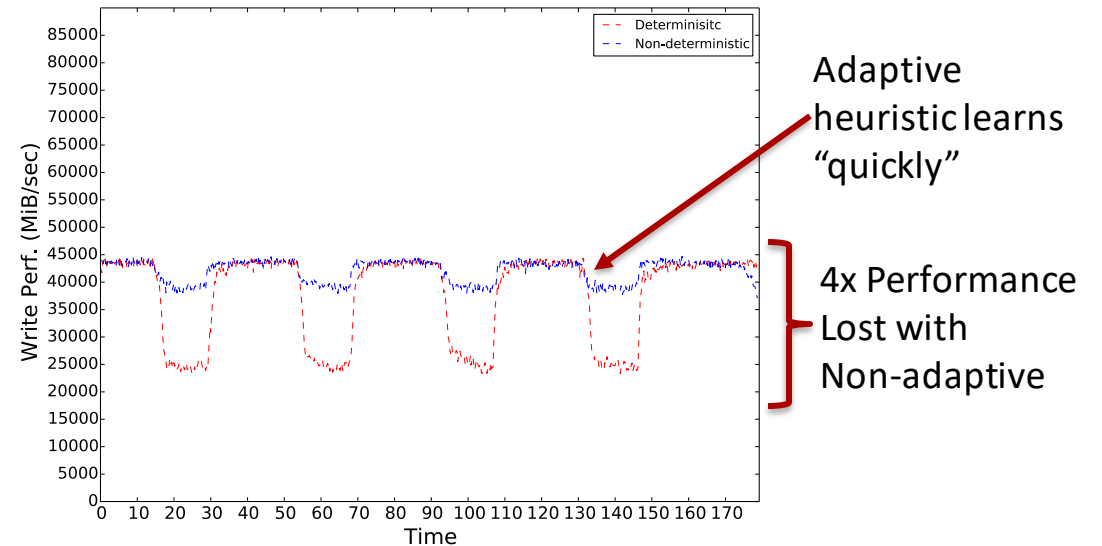
- ▶ Typical customer manually stage data in and out to/from AWS
- ▶ Typical workflow for data transfer is to upload everything, find the files that changed (if any) and then copy the changed files back
 - Has a manual component to it
 - The data transfers are usually single threaded
 - Because of the two-step nature (copy then process) of the data copy, the workflow is prone to risk of serialization
- ▶ Applications that are going to modify data need to have POSIX to avoid serious write amplification
 - Add 4KB to a 1GB file and you have to write out 1GB of data if the data is in object/immutable
- ▶ Most applications don't speak S3 natively
- ▶ AWS Storage Gateway and AWS DataSync attempt to solve this problem, but these solutions for providing on-prem flexibility for data that is stored in AWS (data locking)
- ▶ WAN connection to Cloud needs to be strong with low latency

Smart data movement On-Premise ↔ Cloud

- ▶ Uploading data to the cloud is "free"
- ▶ Egress traffic from the cloud is very expensive and IME will only sync the "dirty" data back to the on-premise system (minimizing cost)
- ▶ Customers with an on-prem data storage solution don't typically want to copy all of their data into the cloud. IME makes the data set portable in a straight forward way on (native to the block)

Software Defined Storage with QoS!

- ▶ Cloud resource is not as stable bare-metal on premise
- ▶ Cloud performance depends on Cloud Provider resources management
- ▶ DDN IME's Non-Deterministic Data Placement further improves performance in the cloud
 - Resilient to network transient degradation
 - Resilient to compute consolidation
 - Avoids write amplification
 - Built-in data protection (no GCP persistent drive)



Multi-cloud

▶ Multi-cloud in HPC today has limited options

- Most multi-cloud solutions are enabled through scheduler integration (where the scheduler copies data).
- A majority of storage providers' approach to enabling multi-cloud causes portions of the data set to be unusable on-prem while it's in the cloud.

▶ DDN's options for enabling multi-cloud

- IME allows a simple extension of a namespace into the cloud, keeping the frequented data in the cloud
- The introduction of filesync and cloudsync in EXA5 provide utilities to copy data (filesync) and then spin up EXA5 in the cloud and import the namespace

DDN IME

FLASH-NATIVE INTELLIGENT DATA CACHE

100% IO500 efficiency

lowest latency

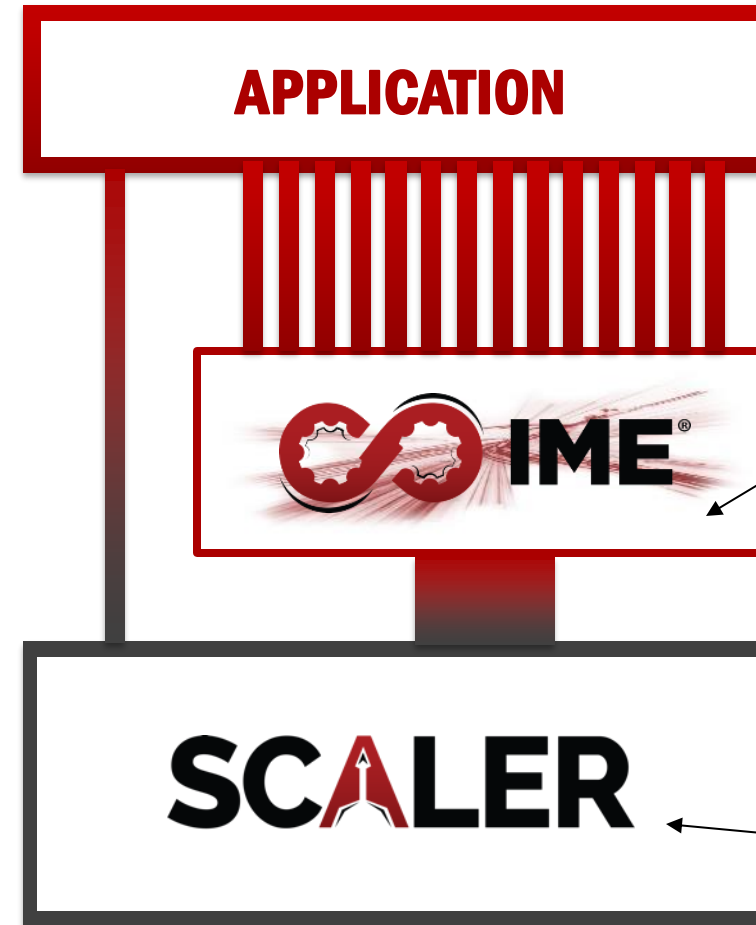
Advanced data protection and resilience

No bottlenecks from POSIX file locking

Coalesce and optimize IO to backend FS

Reduces infrastructure footprint

Easy to deploy



This guy is SDS

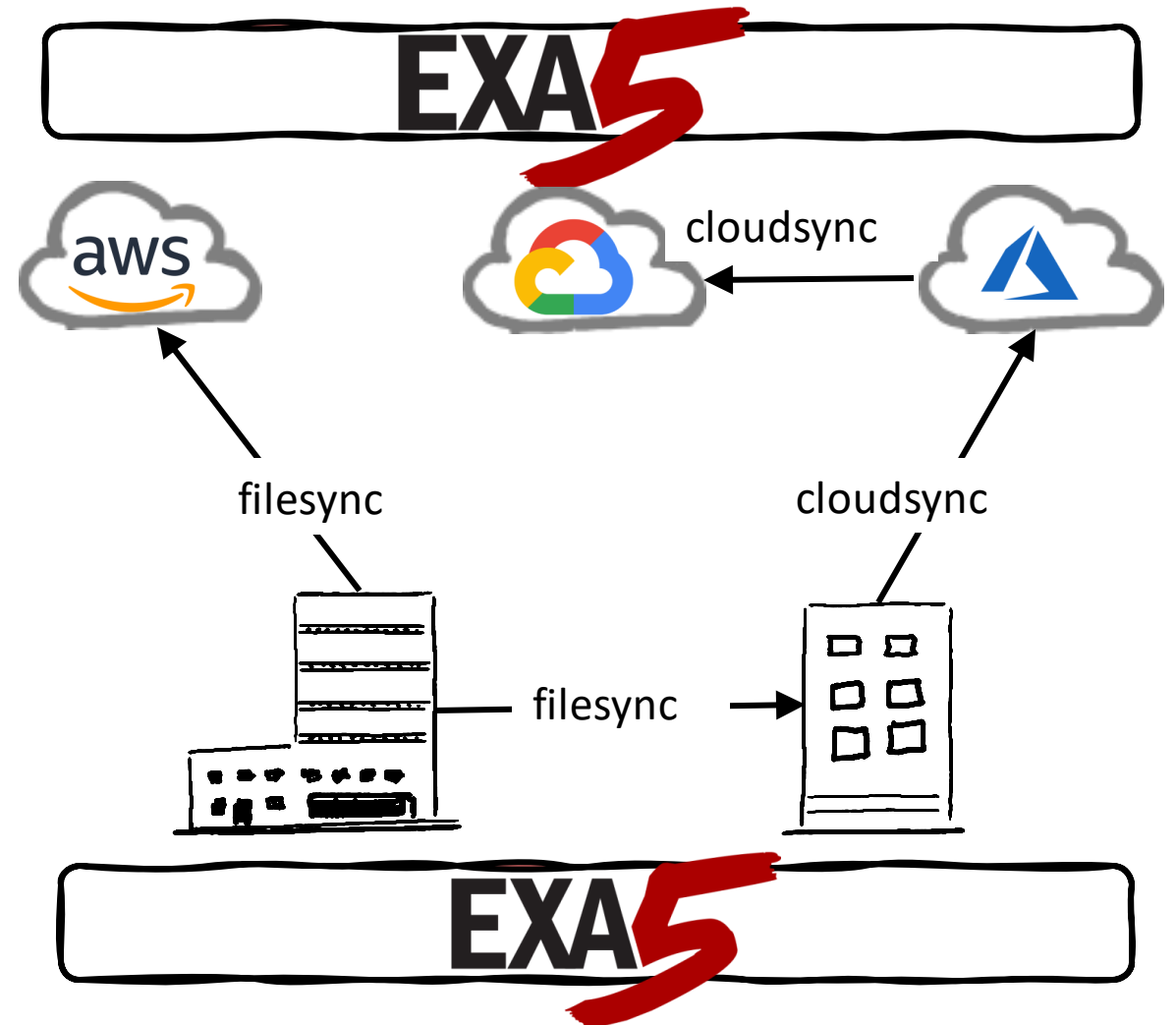
This guy is a
HPC appliance

Autonomous Vehicle Workflow | Ingest, Training & Simulation, Collaboration

Workflow and application	Purpose	Data type	Perf. vs. Capacity driven	Access protocols
Data Ingest, Data Tagging and Cleanup	Store data of car fleet, then prepare data by expanding metadata information including tagging. ETL (extract, transform, load)	Almost all unstructured	Ingest: mostly sequential. Tagging and ETL: random I/O Both require high amounts of capacity available	SMB, NFS, S3 Clients start considering native PFS
Training (DL), Simulation and Resimulation	Train the AI system. Simulate and validate the resulting AI chip and sensors for total car behavior	Mostly unstructured	DL and Resim: extremely high sequential and random IO performance, low latency. Best technology is NVMe/flash	Mostly native PFS, Containers (Docker, Kubernetes, etc.)
Cloud and Collaboration	Distribute data to remote (engineering) sites. Offload data to the cloud (hybrid model)	Mostly unstructured	Capacity driven	S3, Swift Additionally: data movers, as DataFlow

Enabling Hybrid cloud workflow implementations

- ▶ *Stratagem Data Management brings the ability to track and move your data between filesystems and object stores*
- ▶ *Dump your filesystem to S3 for Archive or Data Transfer*
- ▶ *Spin up new Filesystems from S3 at speed reducing your long term cloud costs*



THANK YOU!



Cloud and Collaboration | DDN Solutions and Architectures

▶ IME - Cloud Bursting

- Allows compute bursting into cloud without moving all the data
 - No read / write amplification: Read on demand, Write back deltas
- IME QoS can cope with non-homogeneous Cloud Networks

▶ S3 Data Access

- The quasi-standard, required in many cases

▶ DataFlow

- A powerful data mover tool between different data protocols

▶ IME Remote Caching for two DC Sites

- Similar to Cloud Bursting, but allows to cache data closer to compute across multiple data centers
- This assumes workloads are read-only (e.g. DL, Training)