Is your storage ready for exascale?

Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)

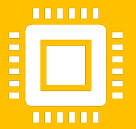
1 zettabyte is equal to
1 billion terabytes.

2010	2015	2018	2020	2025	2030	2035
2	12	33	47	175	612	2,142

Challenges for data storage

Increasing resolution of instruments collecting or producing data

Increasing number of data collectors and producers



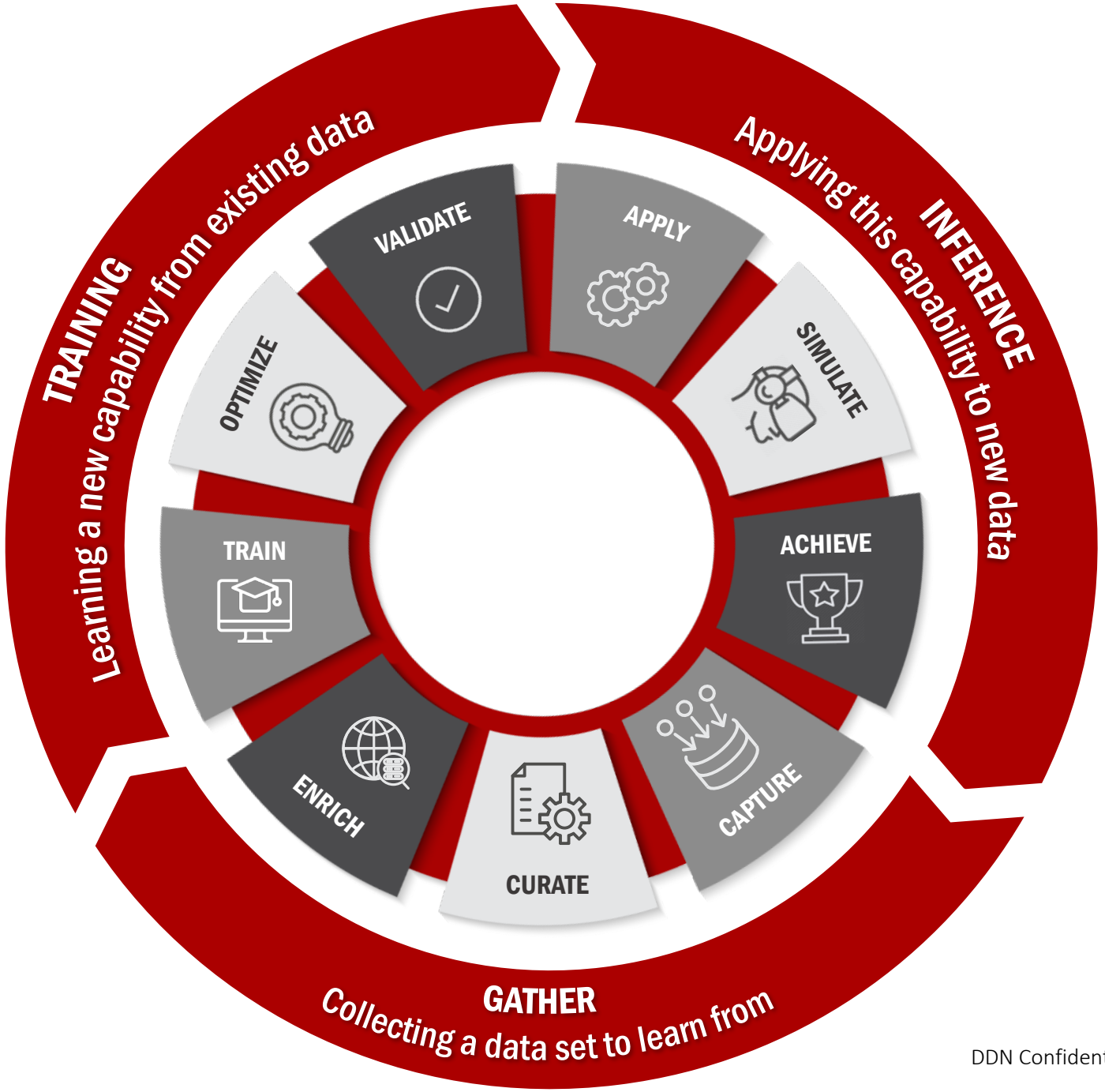
Increasing compute power that can be brought to bear on data to extract information

Increasing value of already-processed data in AI workloads

AI WORKFLOWS ARE CYCLICAL, NOT LINEAR

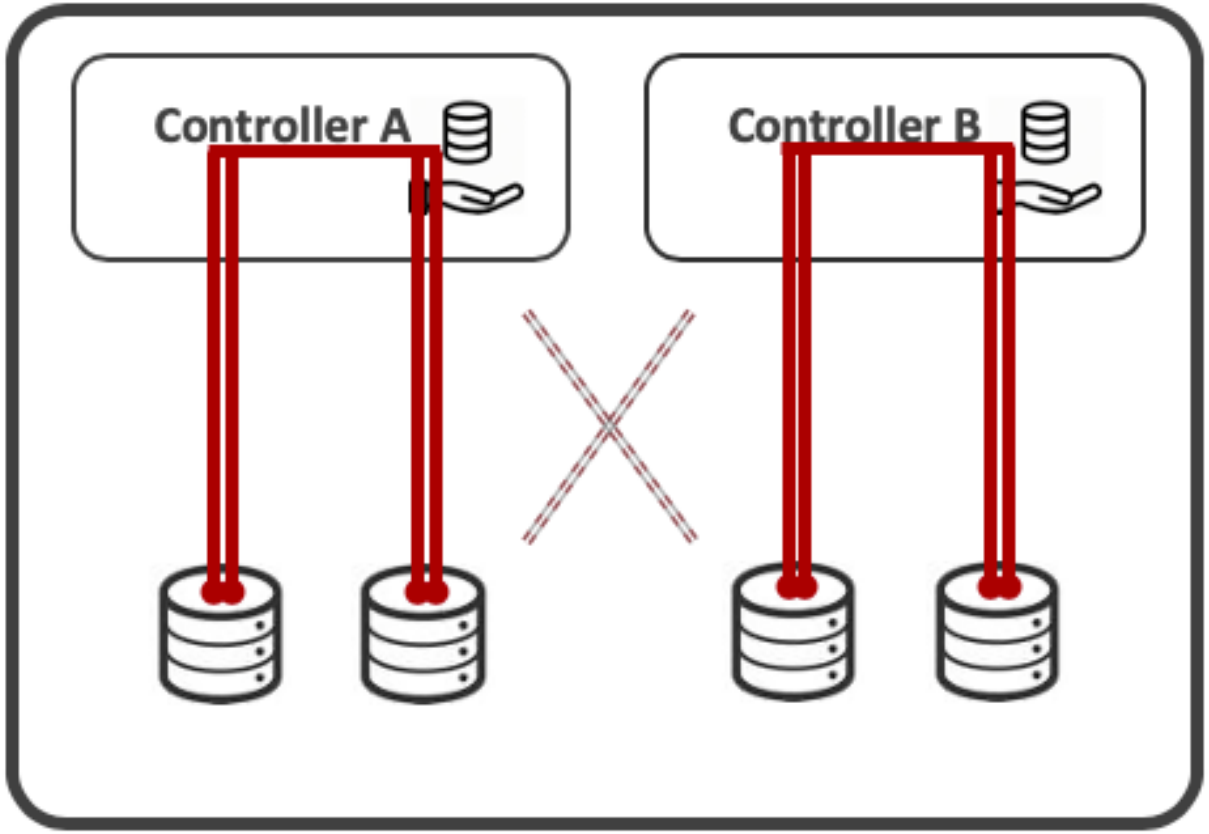
Data is consumed, over and over in random order as models are trained, refined, and tested.

Data may be retained to validate model accuracy, quality control, regulatory compliance, or future refining of the model.



Typical data storage today

- The most common data storage device in HPC today is a 2-controller disk array with nearline SAS drives
- Tape is alive and well, despite being declared dead since 1990



Rotational media

- Rotational media capacity is increasing linearly.
- throughput and IOPS have remained largely unchanged for a decade.

WD RE SAS Specifications

- Model number: WD4001FYYG, WD3001FYYG, WD2001FYYG, WD1001FYYG
- Interface: SAS 6 Gb/s
- Formatted capacity: 4TB, 3TB, 2TB, 1TB
- Host to/from drive transfer rate (sustained): 182 MB/s, 175 MB/s, 170 MB/s, 170 MB/s
- Cache (MB): 32
- Rotational speed (RPM): 7200
- Average drive ready time (sec. from spinup command): 20

Best case rebuild time: 6h for 4 TB drive

Rotational media

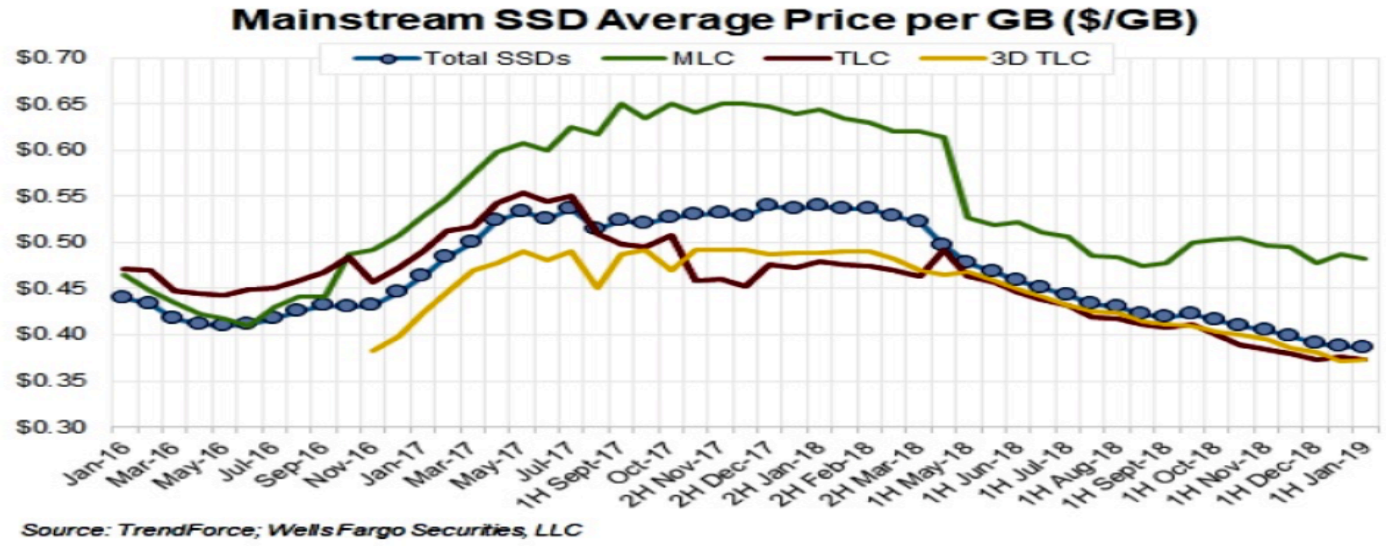
- Rotational media capacity is increasing linearly
- throughput and IOPS have remained largely unchanged for a decade.

Performance		
Data buffer[4] (MB)	512	←
Rotational speed (RPM)	7200	←
Latency average (ms)	4.16	←
Interface transfer rate[5] (MB/s, max)	600	1200
Sustained transfer rate[5]		
(MiB/s, max)	244 / 223	←
(MB/s, max)	255 / 233	←

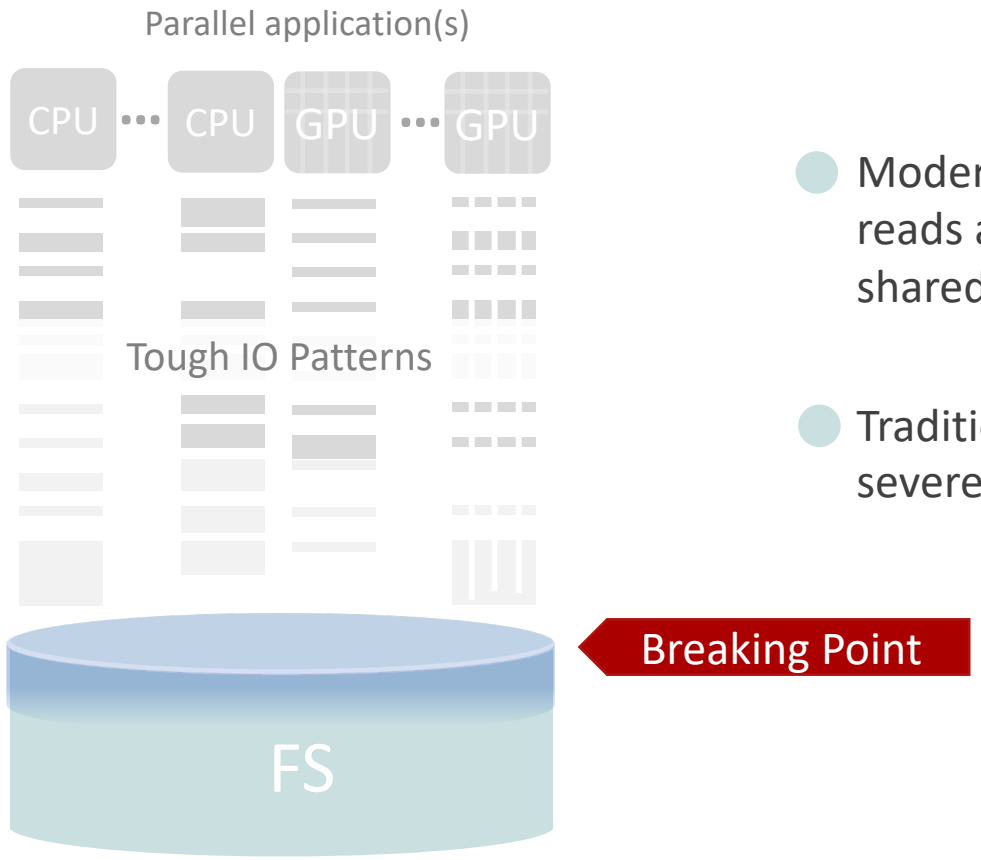
Best case rebuild time: 24 hours for 20 TB drive

Flash media

- Flash capacity growth has been exponential
- Flash media solves the performance problem at the device level, and has good capacity density
- Flash media uses less power than rotational media
- Traditional data protection methods increase rate of burning cells -> reduces hardware lifetime!
- Faster storage devices can hide a lot of antiquated design
- Prices are not dropping nearly as quickly as has been predicted



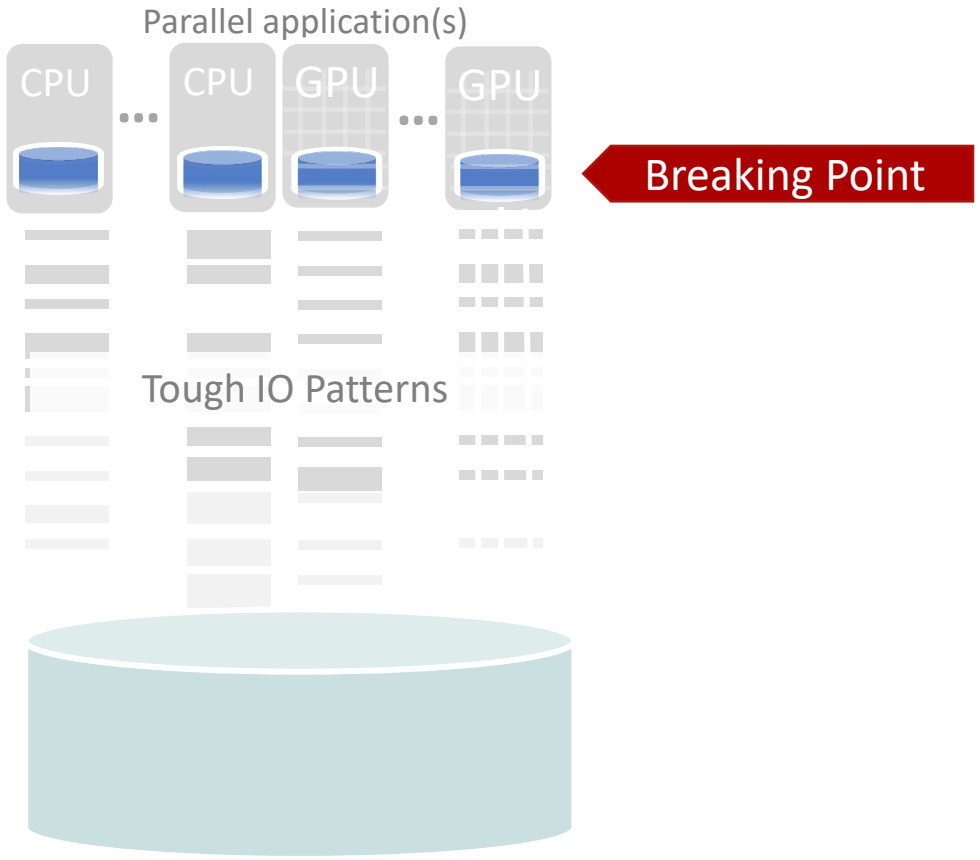
Modern Workloads mean bigger pressures for filesystems



- Modern Workload IO patterns are increasingly mixed and tough: reads and writes, random and sequential, high thread counts, shared file access
- Traditional **Thick File system SW layers** and **fixed data layout** severely restrict performance for tough workloads – even with SSDs

What about NVMe over Fabrics?

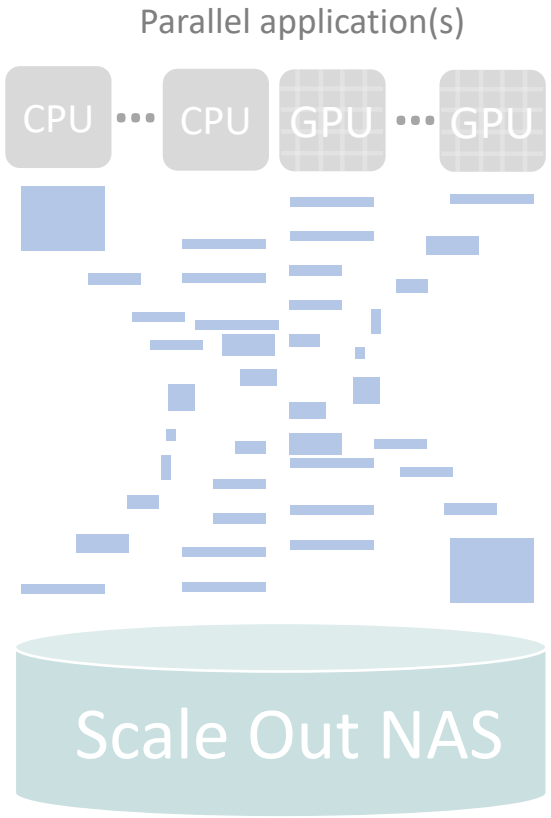
That solves all latency problems, right?



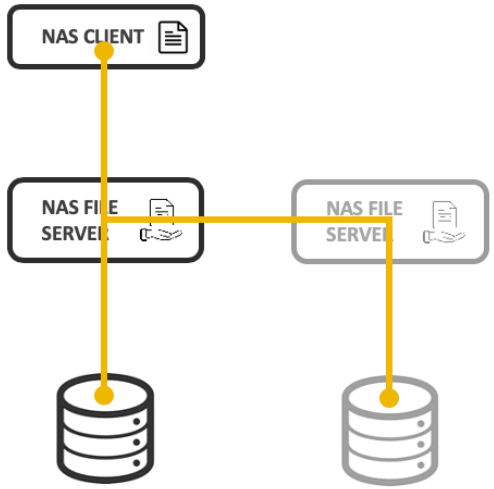
- NVMeoF solves a problem for block, but not for file access. It just moves the bottleneck
- Regardless of the method for providing a block device there is still latency in the filesystem layers
- Your applications only care about filesystem latencies and throughputs

What about Scale out NAS?

Just as Good as a Parallel File System?

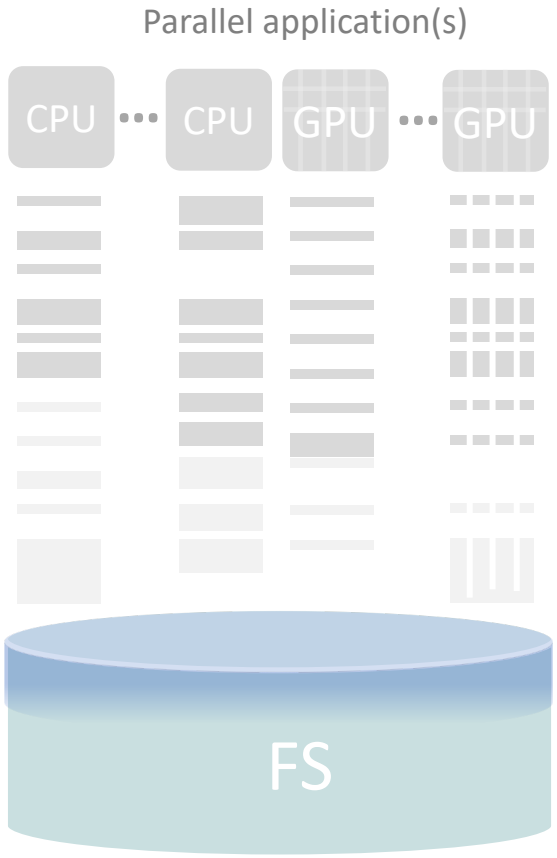


Breaking Point



- Scale out NAS solutions incur a bottleneck from every point outside the storage system to the application.
- The storage system also suffers since there is no scaling assistance from clients

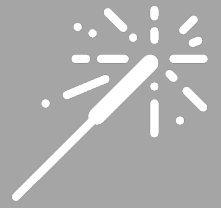
Cloud??



- SLA’s for cloud providers are in the range of 1-2 9’s
- Charged for everything
- Only economical for pilot projects and occasional bursting

So now what?

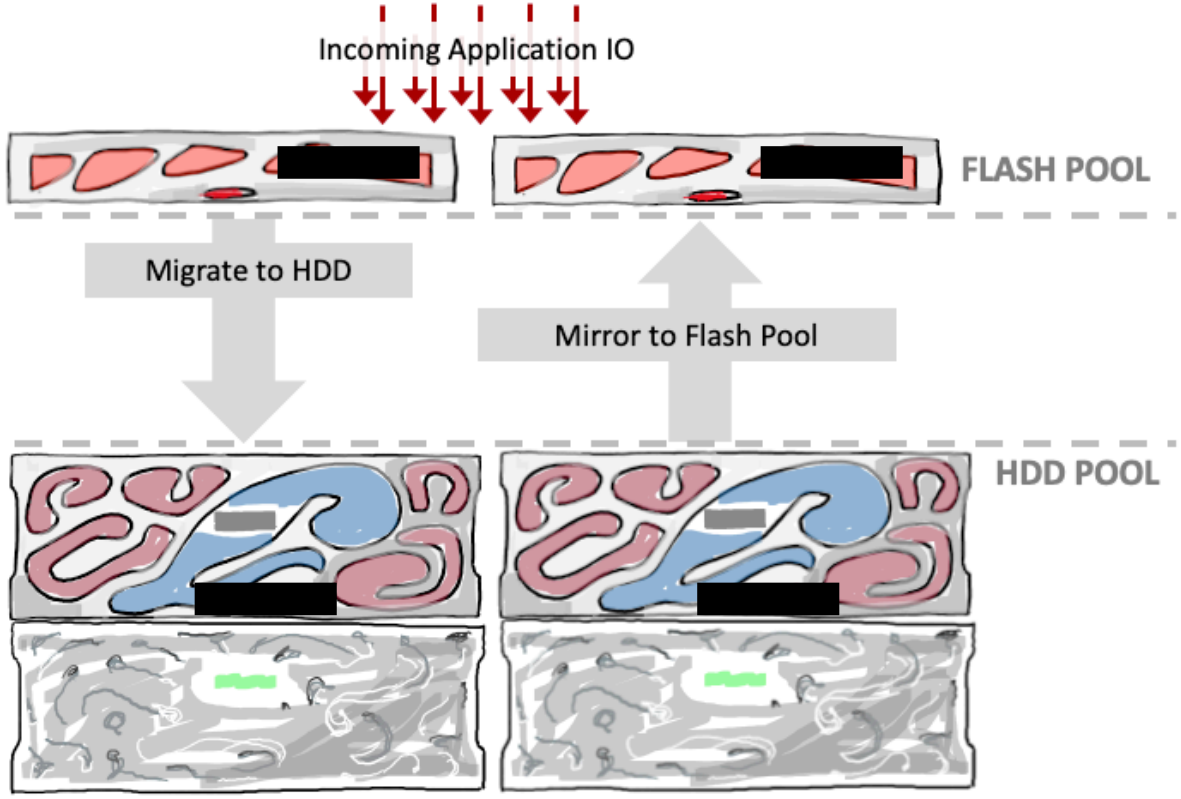
We have too much data and we're too invested in current storage to do a forklift replacement



There is no one magic technology that solves every problem

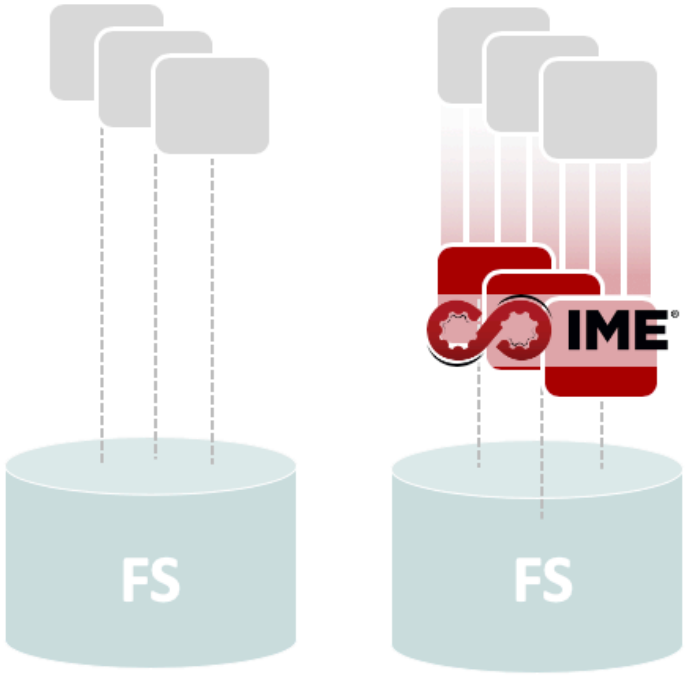
Take data off capacity storage and move to flash for work

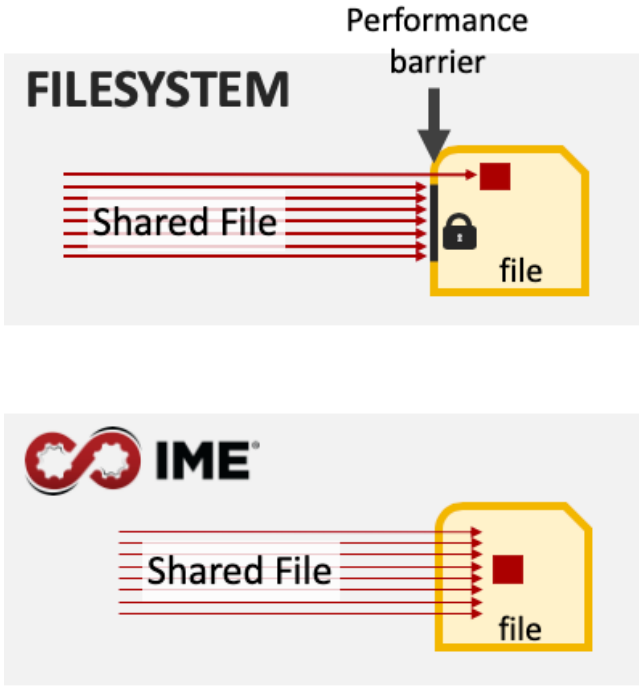
- Use technology in a targeted manner to solve problems, avoid shotgun approach
- Use flash where flash makes sense – to host active datasets



Remove the filesystem entirely when it gets in the way

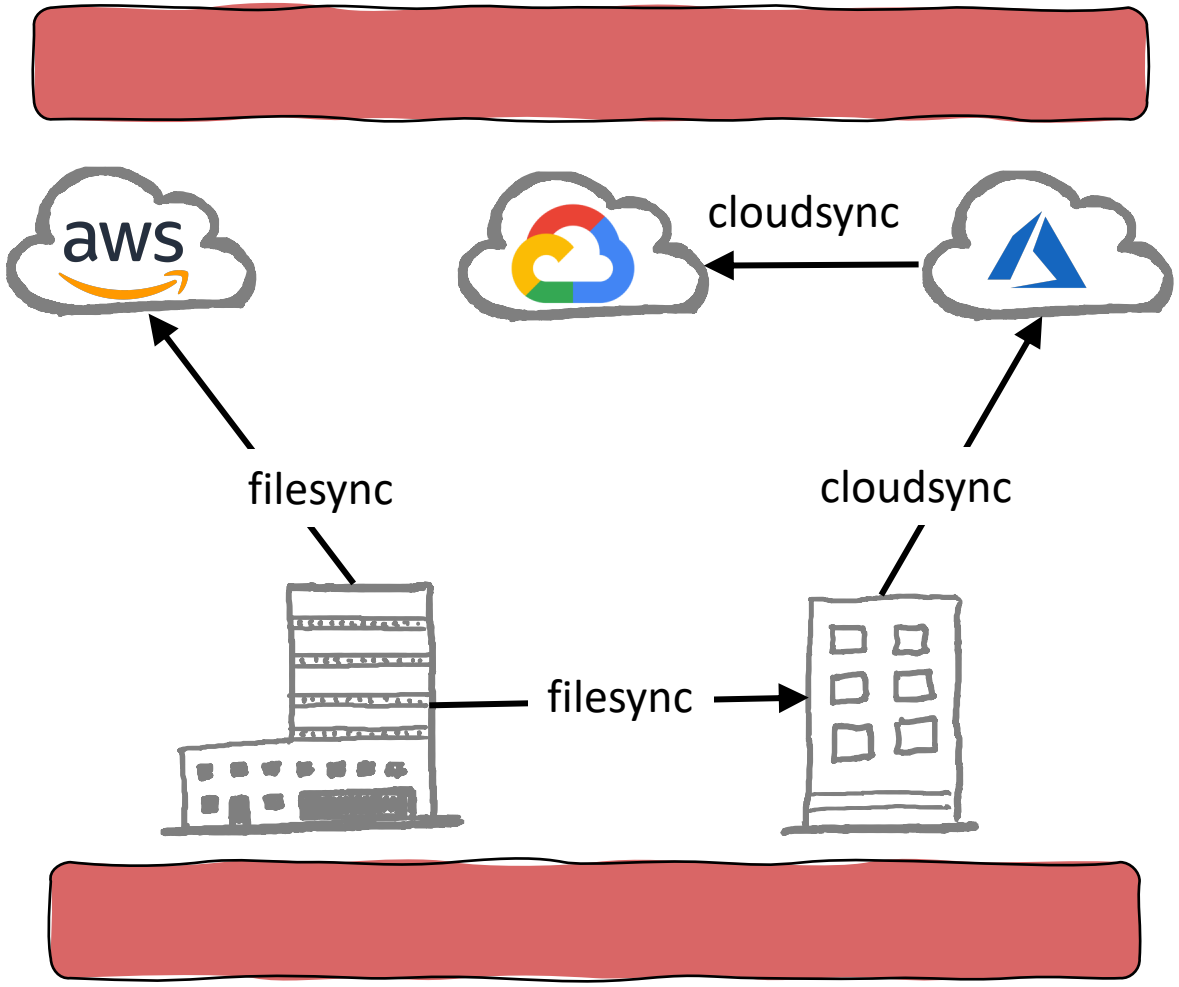
- As great as parallel filesystems are, they still have limitations.
- Use flash as an IO buffer and accelerator
- WEKA, IME, DAOS for accelerating compute jobs.





Enabling Hybrid cloud workflow implementations

- *Use IME to cache your dataset in the cloud. Only dirty cache needs to come back, reducing egress fees*
- *Both EXAScaler and Spectrum Scale bring 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*



This is just a direction – not the final destination

- This is not a roadmap discussion, just some guesses about how data storage may change over the next decade.
- People like me have been predicting the death of tape and rotational media for a long time – and guess what?

Thank you for coming!