Performance requirements for WLCG storage

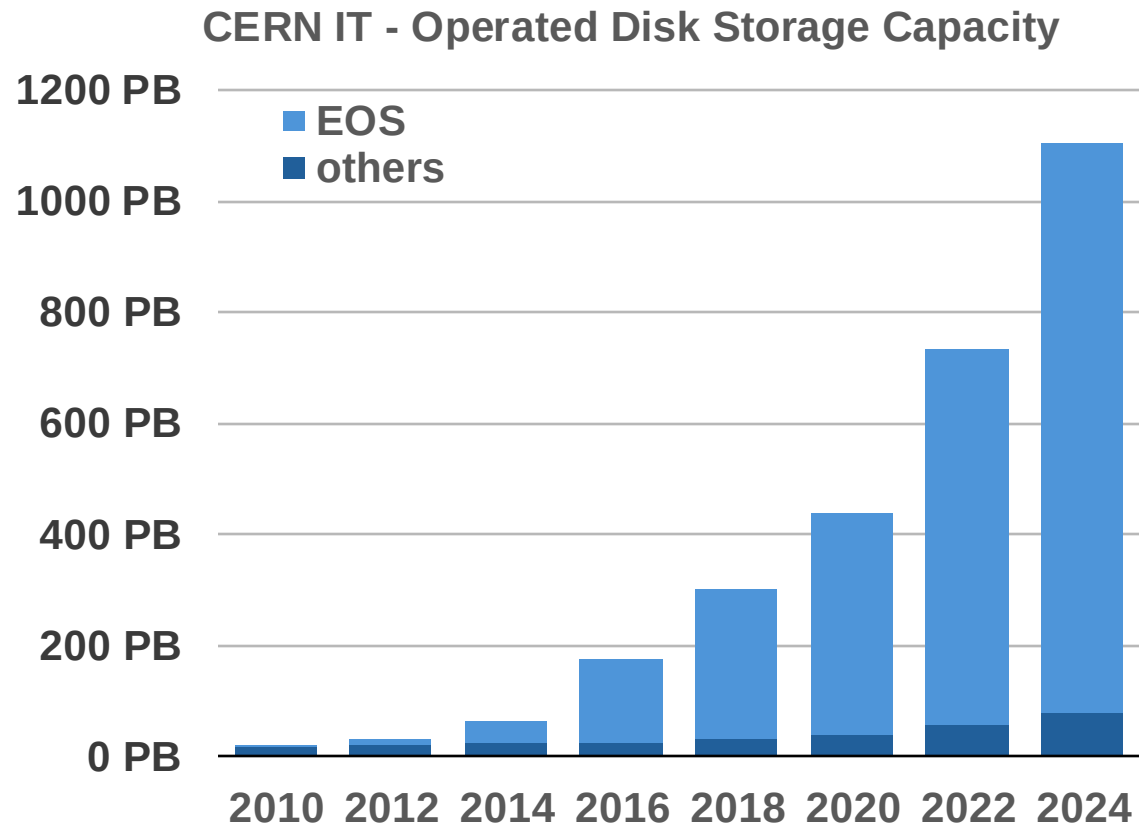
CERN perspective

Luca Mascetti – CERN IT – Physics Data Services

CERN IT Storage Role

Storage Systems are at the core of CERN main businesses:

- Physics Data Recording
  - LHC Data Taking
- Physics Data Processing
- Physics Analysis
- Long-Term Data Archival
- Software Distribution
- General User Storage
- General Infrastructure Storage



Storage Resources

Experiments "*pledge*" storage resources needed in usable space

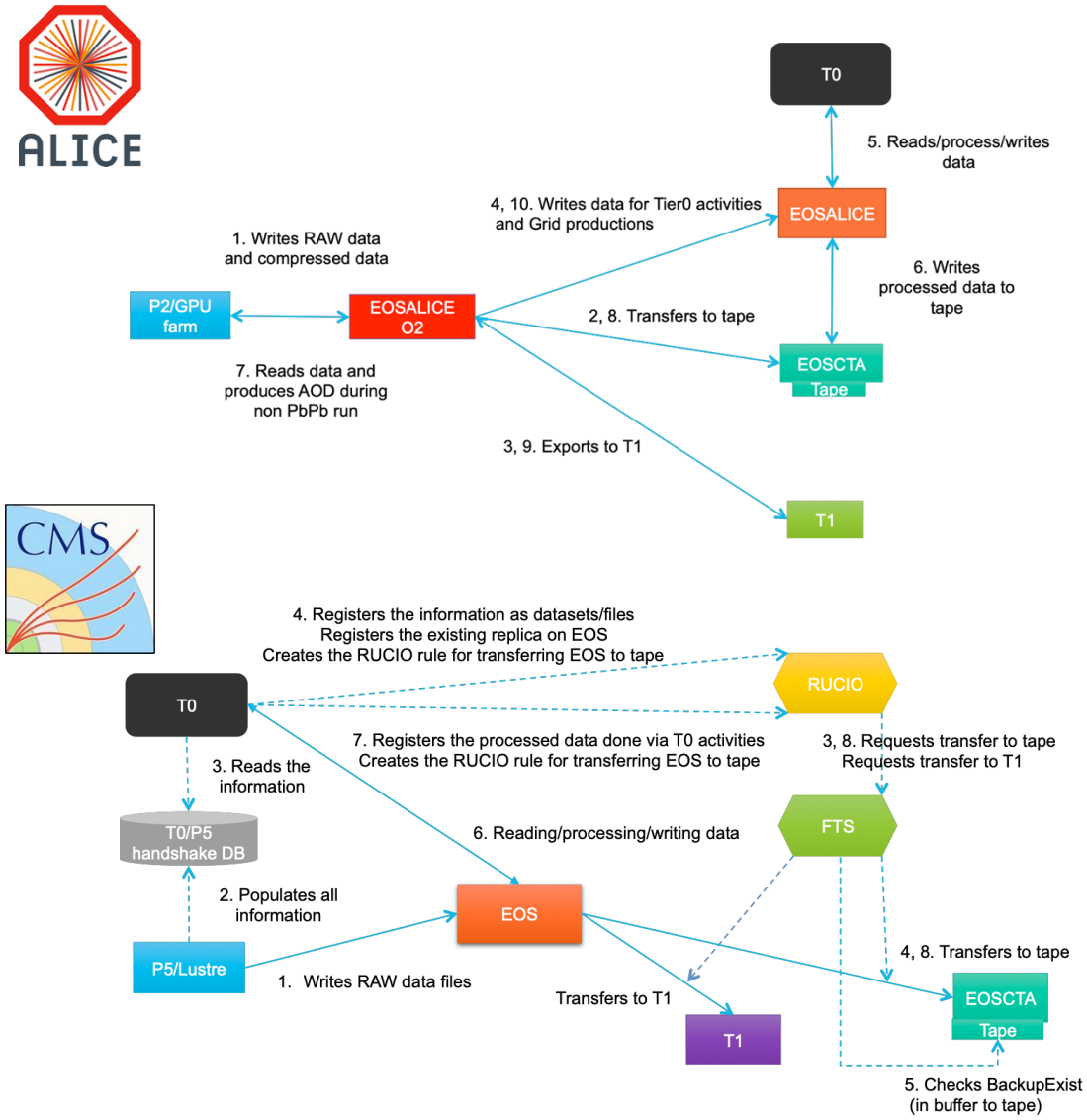
- there is no "*pledge*" for the performance needed by different activities
- only the initial Data-Taking ingestion load is scrutinised by experts
  - This accounts only to a fraction of the ingestion load of our systems
  - This load constantly change every year (even from pp to HI)

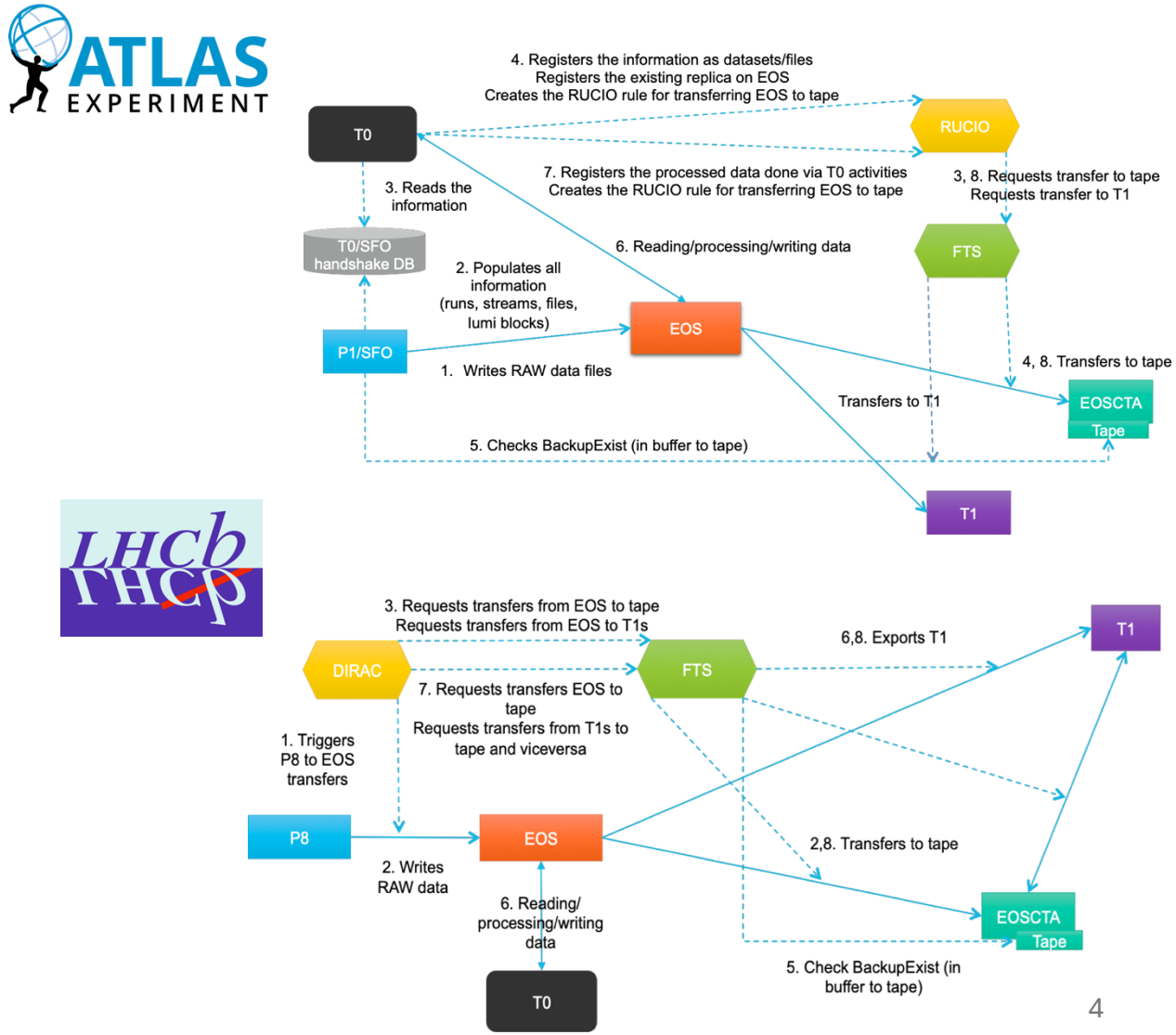
Strategic planning from experts, technology evolution, and investments define the performance that storage system can potentially deliver.

From past experiences we extrapolate and try to predict how the experiment will use our services in the next 5 or more years.

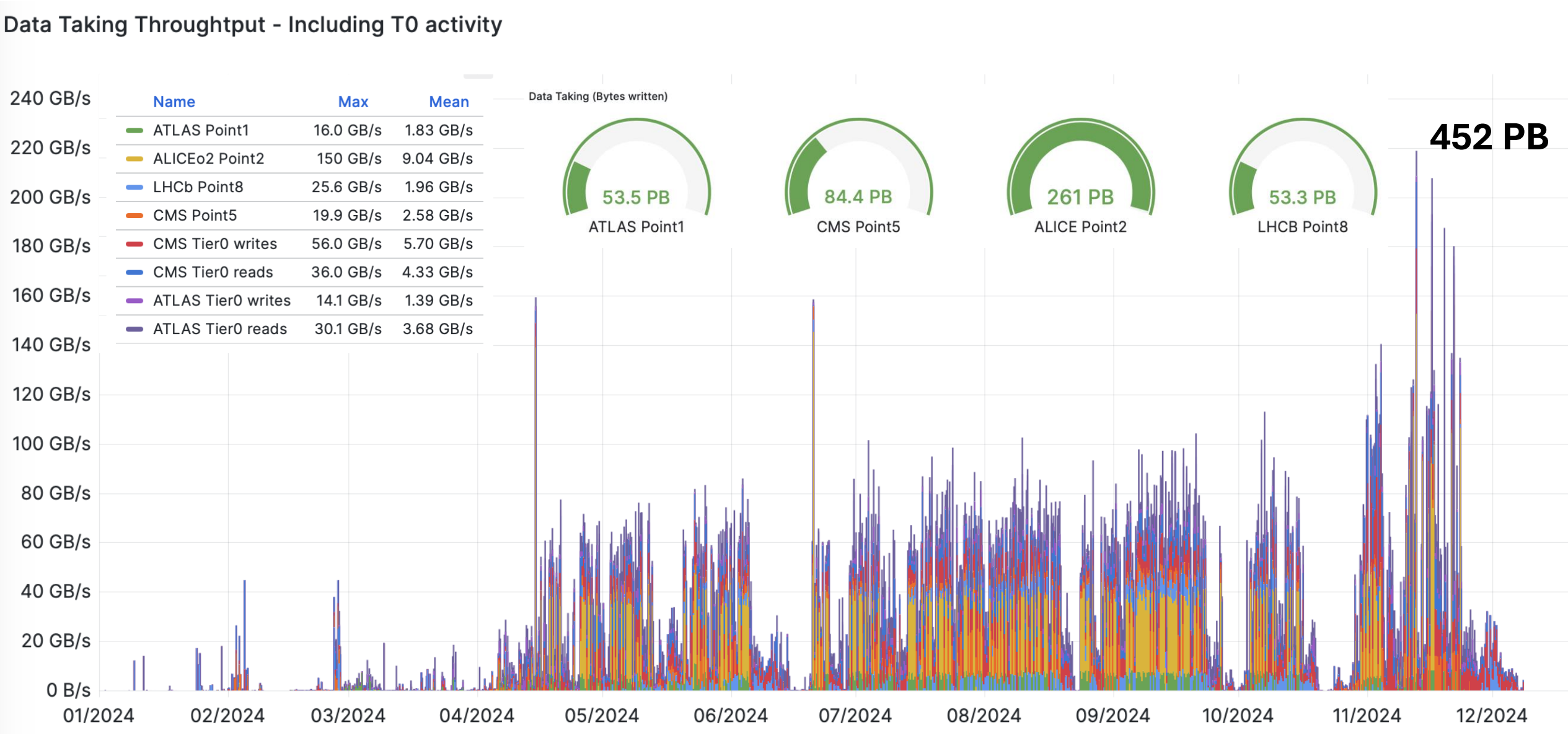
Flexibility (both hardware and software) is extremely important!

LHC Run3 Data Taking Workflows



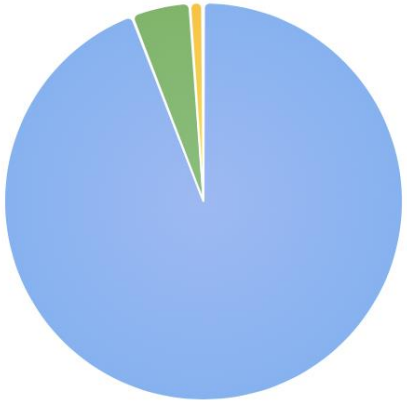


2024 LHC Data Taking



EOS Physics – 2024 Traffic Rates

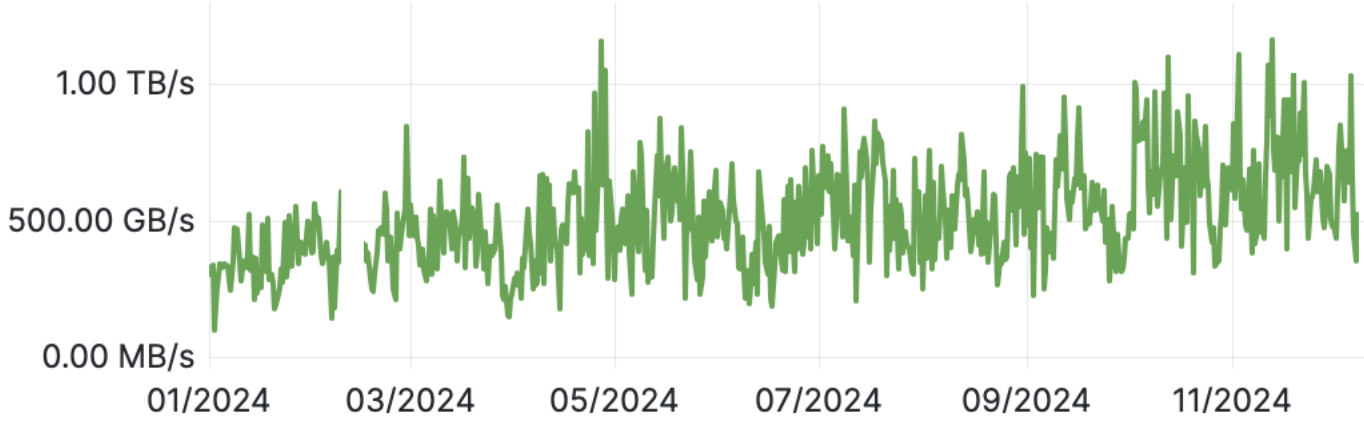
Ingestion per Protocol



	Value	Percent
XROOTD	1.08 EB	94%
HTTP	53.6 PB	5%
FUSE	10.9 PB	1%
GRIDFTP	18.2 TB	0%

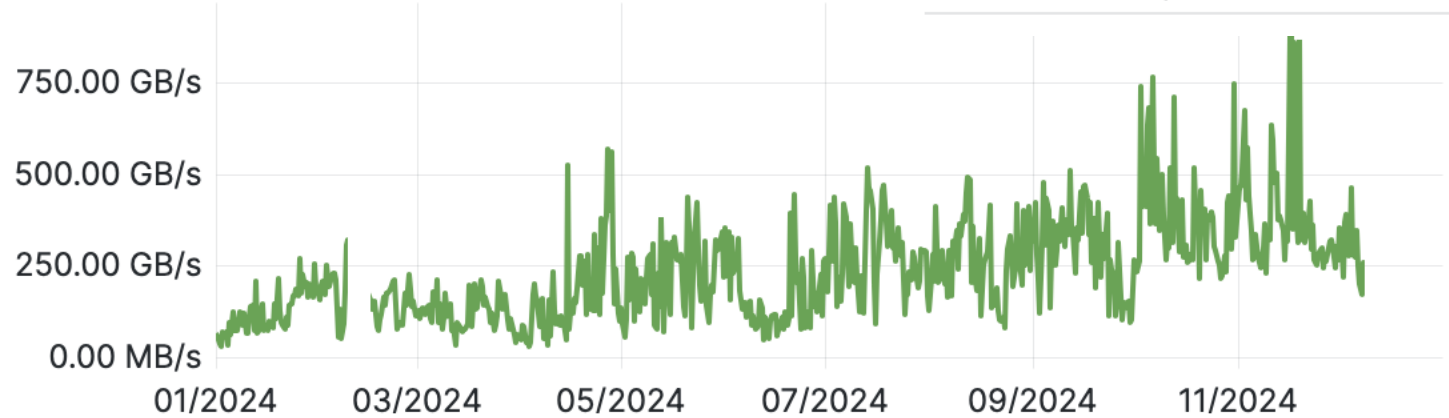
1.15 EB

Maximum: Cluster Network Rates (OUT)



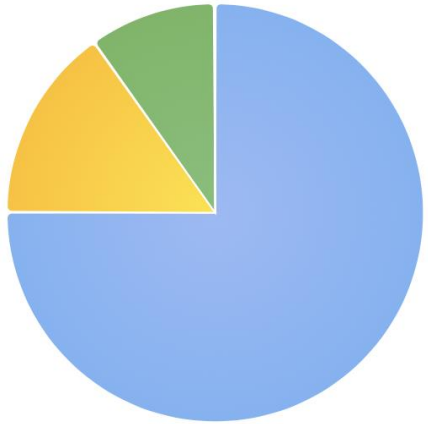
Name	Max	Mean
Maximum peak	1.16 TB/s	512.66 GB/s

Maximum: Cluster Network Rates (IN)



Name	Max	Mean
Maximum peak	881.86 GB/s	237.19 GB/s

Export per Protocol

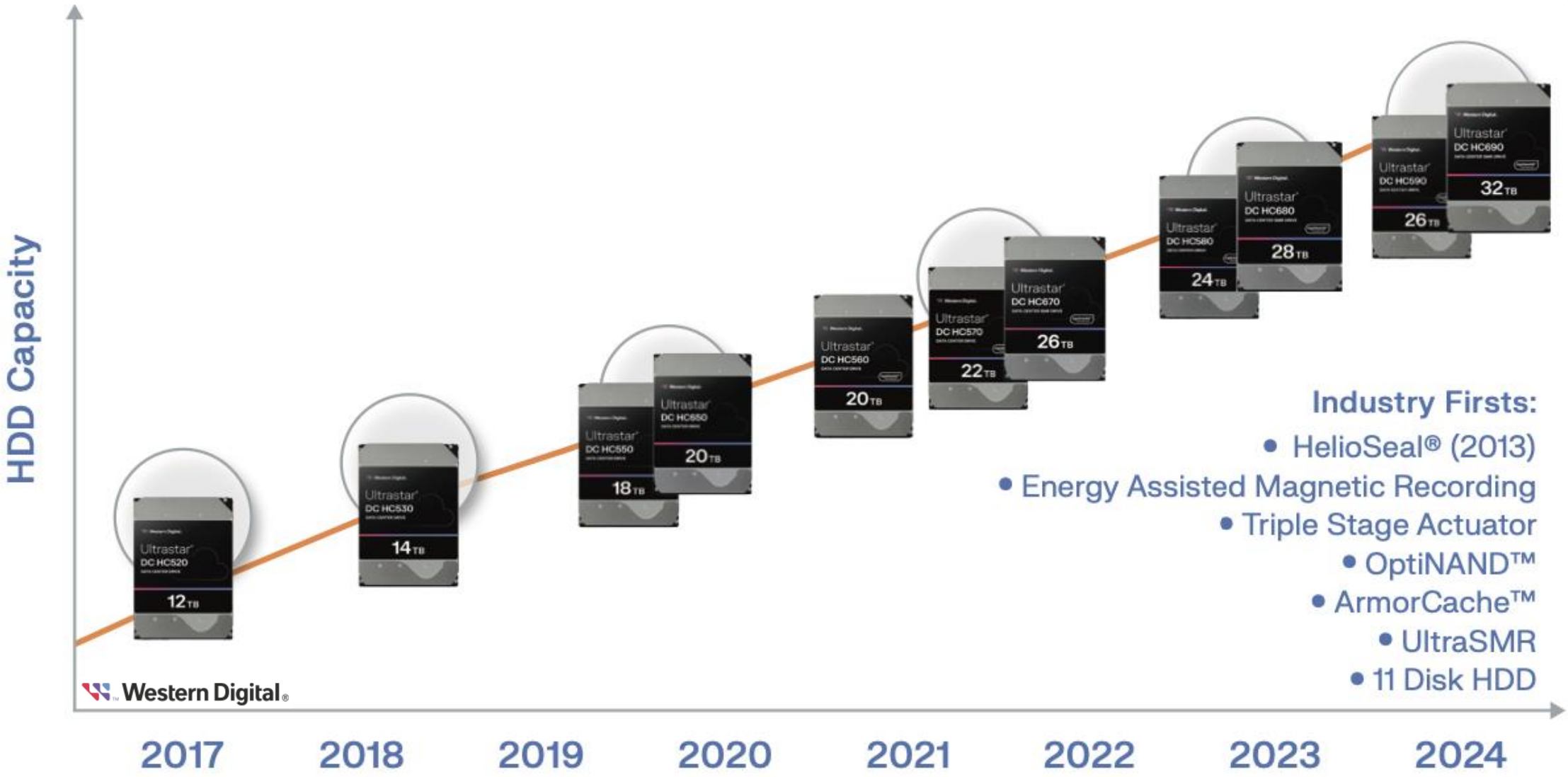


	Value	Percent
XROOTD	5.62 EB	75%
FUSE	1.12 EB	15%
HTTP	728 PB	10%
GRIDFTP	37.3 TB	0%

7.47 EB

Disk Technology Evolution (HDD)

50TB+
2030



Link

Latest WD on Market

All latest technological advancements

- 11-platters
- ePMR
- Triple-stage Actuator
- Helio Sealed
- UltraSMR
- OptiNAND
- ArmorCache

	30TB SATA	30TB SAS	32TB SATA	32TB SAS
Model Number	WSH723200ALxxyz	WSH723200ALxxyz	WSH723220ALxxyz	WSH723220ALxxyz
Formatted capacity[1]	30TB	30TB	32TB	32TB
Recording Technology	SMR	SMR	SMR	SMR
Interface	SATA 6 Gb/s	SAS 12 Gb/s	SATA 6 Gb/s	SAS 12 Gb/s
Format: Sector size (bytes)[2]	512e:512 | 4Kn: 4096	512e:512 | 4Kn: 4096	512e:512 | 4Kn: 4096	512e:512 | 4Kn: 4096
Areal density (Gbits/sq. in.)	1385	1385	1480	1480
Performance				
Data buffer[3] (MB)	512	512	512	512
Rotational speed (RPM)	7200	7200	7200	7200
Latency average (ms)	4.16	4.16	4.16	4.16
Interface transfer rate (MB/s, max)	600	1200	600	1200
Sustained transfer rate[4] (MB/s, max) / (MiB/s, max)	260 / 248	260 / 248	269 / 257	269 / 257

Max ~270 MB/s (1-stream)

Comparison with other products

WD Gold® Enterprise Class SATA HDD

~280-300 MB/s (1-stream)

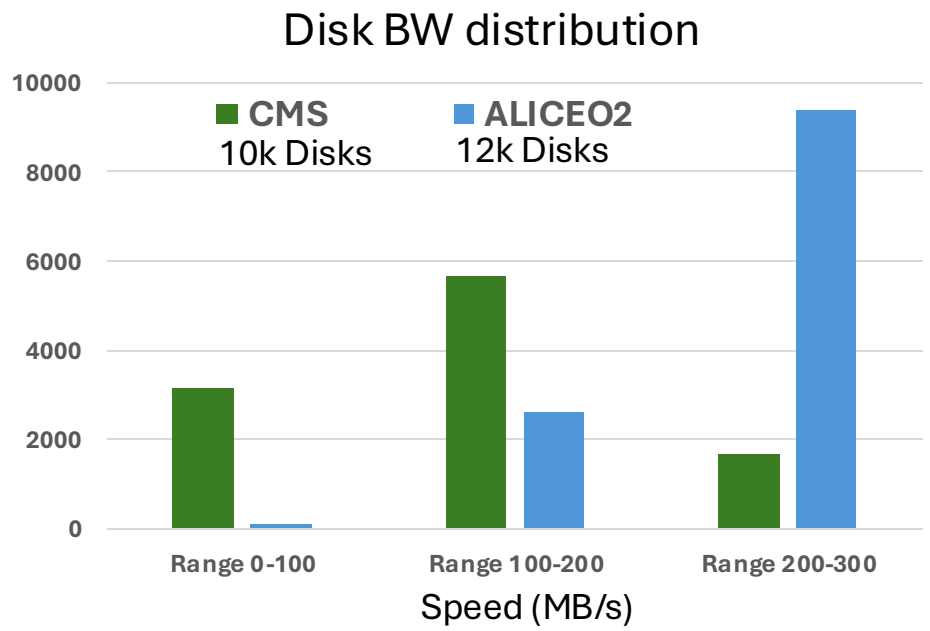
Specifications

Model Number	WD261KRYZ	WD242KRYZ	WD241KRYZ	WD221KRYZ	WD203KRYZ	WD202KRYZ
Formatted capacity[1]	26TB	24TB	24TB	22TB	20TB	20TB
Form factor	3.5-inch	3.5-inch	3.5-inch	3.5-inch	3.5-inch	3.5-inch
Interface	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 Gb/s	SATA 6 GB/s	SATA 6 Gb/s
512n / 512e user sectors per drive[5]	512e	512e	512e	512e	512e	512e
OptiNAND™ technology	Yes	Yes	Yes	Yes	No	Yes
ArmorCache™	Yes	Yes	Yes	Yes	No	No
RoHS compliant[6]	Yes	Yes	Yes	Yes	Yes	Yes
Performance						
Data transfer rate[7] (max sustained)	285MB/s	279MB/s	298MB/s	291MB/s	285MB/s	285MB/s
RPM	7200	7200	7200	7200	7200	7200
Cache (MB)[1,8]	512MB	512MB	512MB	512MB	512MB	512MB

Measuring Real Production

EOS measure* and keep track of several information

- Size and Occupancy
- Current Performance Delivered
  - Bandwidth
  - IOPS
- Short Test of disk performance* at ”booting” time
  - Max Bandwidth
  - Max IOPS



* Measured in a noisy production environment, usually always lower that MAX performance from manufacturer

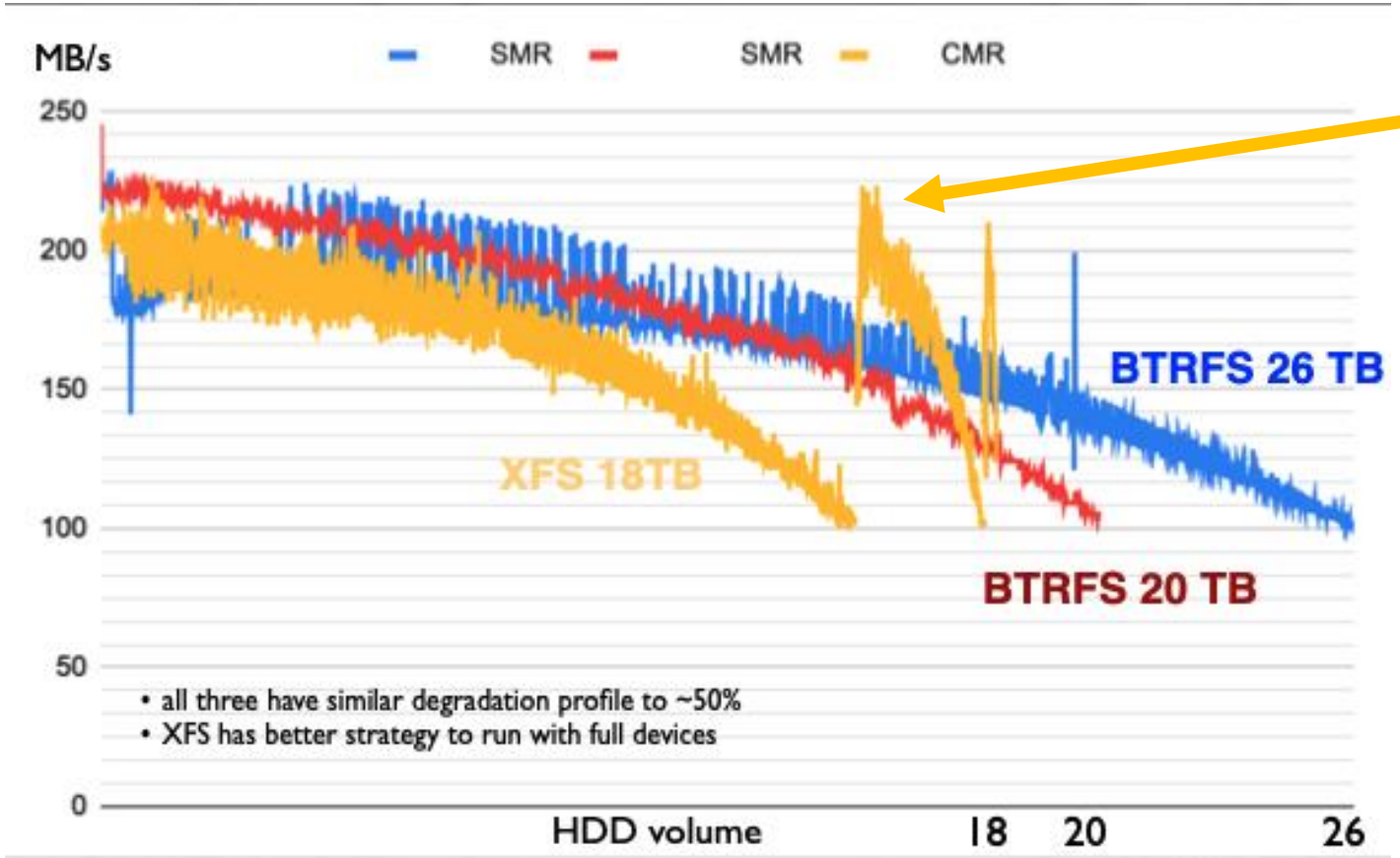
```
[root@eoscms-ns-ip563 (mgm:master mq:master) ~]$ eos fs ls --io | cut -c -60,200-
```

hostport	id	schedgro	max-bytes	used-files	max-files	bal-shd	iops	bw
p06636710d83327.cern.ch:1095	22192	default.1	6.00 TB	412.87 K	293.03 M	0	251	160 MB
p06636710d83327.cern.ch:1095	22193	default.1	6.00 TB	423.62 K	293.03 M	0	245	208 MB
p06636710d83327.cern.ch:1095	22195	default.1	6.00 TB	412.13 K	293.03 M	0	253	207 MB
p06636710w15575.cern.ch:1095	22200	default.1	6.00 TB	402.22 K	293.03 M	0	258	160 MB
p06636710d83327.cern.ch:1095	22201	default.1	6.00 TB	405.65 K	293.03 M	0	192	200 MB
p06636710d83327.cern.ch:1095	22204	default.1	6.00 TB	408.04 K	293.03 M	0	235	210 MB
p06636710w15575.cern.ch:1095	22206	default.	6.00 TB	376.51 K	293.03 M	0	255	173 MB
p06636710d83327.cern.ch:1095	22207	default.1	6.00 TB	418.96 K	293.03 M	0	268	215 MB
p06636710d83327.cern.ch:1095	22209	default.1	6.00 TB	418.06 K	293.03 M	0	271	208 MB

Other HDD performance factors

Single Disk performance depends as well on the following

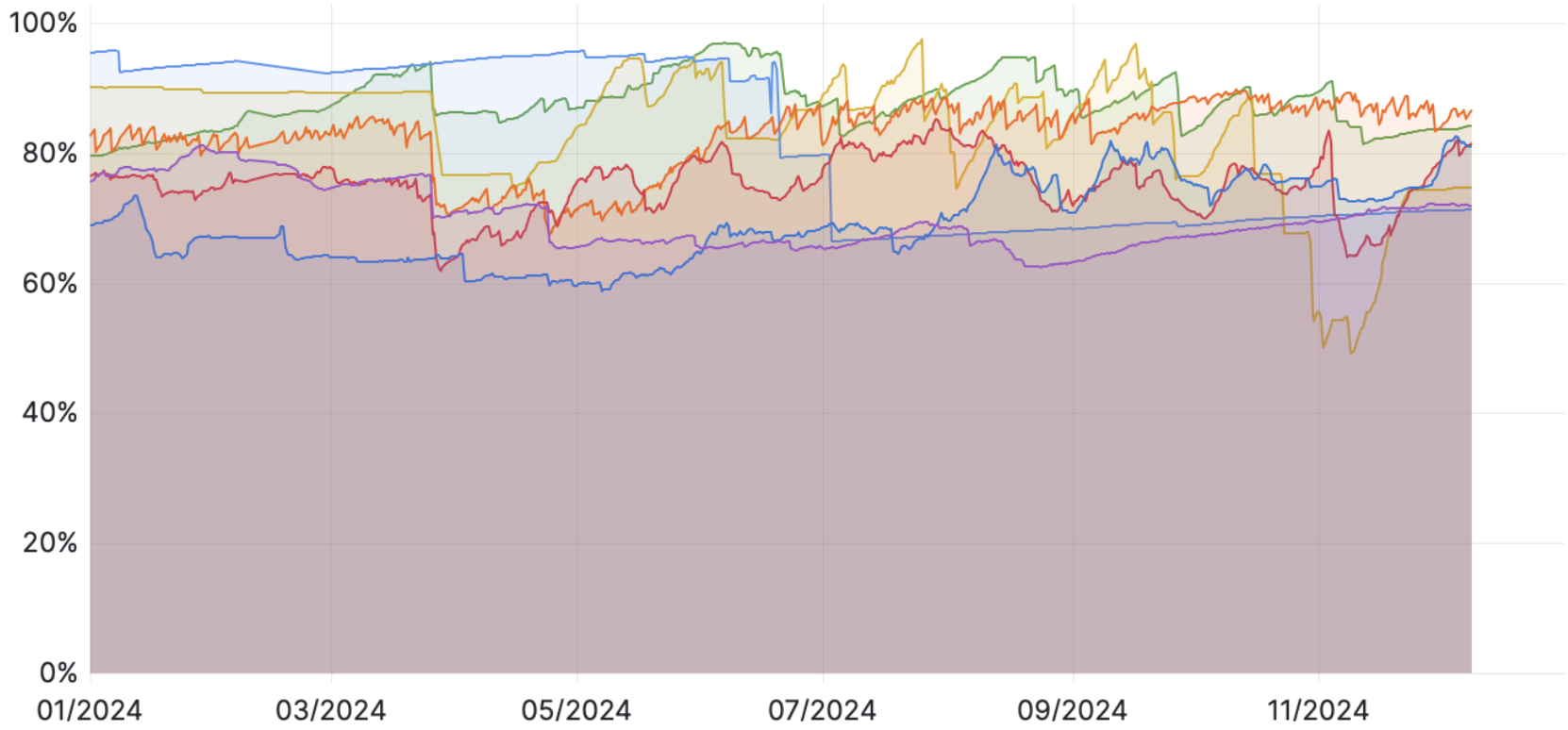
- Disk Fullness
- Holes generated by Write/Delete cycles



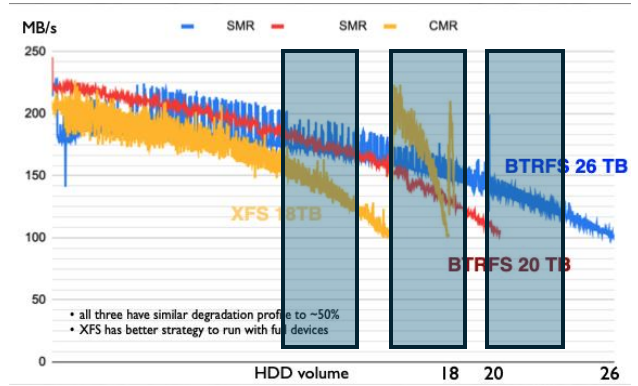
XFS defrag

2024 EOS Physics Storage Usage

Percentage of space utilization (only experiments)



Name	Last *	Max	Mean
eosaliceo2	74.7%	97.5%	83.5%
eosalice	84.2%	97.0%	87.9%
eosams02	71.3%	95.9%	81.0%
eosatlas	86.6%	89.8%	82.8%
eoscms	81.6%	85.2%	75.2%
eoslhcb	81.1%	82.5%	69.2%
eospublic	71.9%	81.2%	69.9%



Actual performance ranges for real production workloads

Other HDD performance factors

Single Disk performance depends as well on the following

- Amount of parallel streams
  - Disk BW 2.8 times slower with 10 sequential streams!
- Stream activity
  - Sequential vs. random read/random write
  - Only 25% of the original performance!

Example:

300 MB/s Disk delivers ~110 MB/s with 10 parallel sequential streams

300 MB/s Disk delivers ~ 75 MB/s with a rnd read.rnd write workload

Then have both workload in parallel, or run with disks filled at 70-85% ;)

Latest CERN storage server architecture (or how to track bottlenecks)

The actual bottleneck changes over time
Due to hardware evolution and server architectures

e.g. In LHC Run2 we were severely bound by NICs (1Gbps)

Router

Switch

Blocking factor
1:3 1:2

Network Interface
25Gb→ 3.1GB/s
100Gb→12.5GB/s

CPU

HBA

SAS Cable

SAS12 cables to multiple JBODs
24GB/s

JBODs **24-30GB/s**

Sum of Disks

Single Disk
250MB/s
(1-stream)

Overall JBODs subsystem performance

# Streams	1	10	50	100
seq.	24 GB/s	8.5 GB/s	5.5 GB/s	1.5 GB/s
rnd rw	6 GB/s	2 GB/s	1.3 GB/s	0.3 GB/s

Disk performance issues will be visible very soon!

- Latest generation Storage server:
  - 120disks x **24TB** drives with 100Gb interfaces
  - Hyper-optimized $/TB

- Currently we provide ~440PB usable to experiments
  - If we would replace the ***whole*** capacity with the latest type of server
    - ~300 servers (instead of ~1k) needed in case of replica x2
    - ~185 servers only in case of Erasure Coding (EC10+2)
      - **~22k disks (instead of ~100k)**

- Disk performance has not changed (and will not) over time… 250MB/s (at best)
  - Additional performance penalty to consider!

- EOS Physics currently deliver around **220k parallel transfers**
  - This will translate NOW to 10 streams in each disk (instead of the current avg. of 2)
    - **For HL-LHC this would be even worse, we expect a factor 10x increase on workloads**

100 PB comparison over HW generations

	HW Generation 2017 10Gbps 48x6TB	HW Generation 2018-2019 25Gbps 96X12TB	HW Generation 2020-2022 100Gbps 96x18TB	HW Generation 2024-2025 100Gbps 120x24TB
Number of servers	347	87	58	**35**
Sum of NICs	433 GB/s	272 GB/s	725 GB/s	425 GB/s
Disk Speed (est.)	150MB/s per disk	200MB/s per disk	250MB/s per disk	250MB/s per disk
Sum of Disk BW (1-stream)	~2.5TB/s	~1.67TB/s	~1.39TB/s	~1TB/s
Sum of Disk BW (10-streams)	~890 GB/s	~600 GB/s	~500 GB/s	~360 GB/s

Summary

Disk Industry is driving the road toward a 50TB+ drives

IO and BW performance remains ~stable over time

Overall Performance/TB is decreasing

On our side we need to review the storage server architecture!

Outlook

In general we should expect higher costs for storage if we want to maintain current levels of performance

Storage flexibility is key!

- **New hardware technologies exploration and assessment**
  - **NAND-based storage will become the backbone of storage solutions**
- **Dedicated software development to help in reducing costs**
  - **E.g. Auto-Caching, Tiering, Conversion Policies…**

Need collaborative efforts across teams!