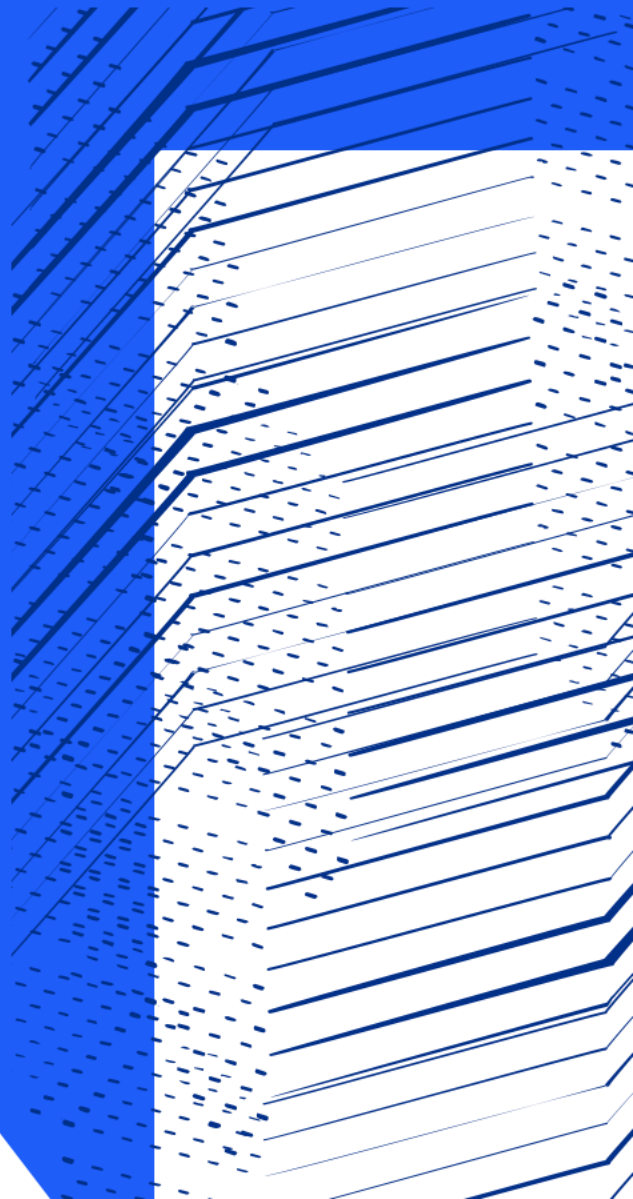
Science and Technology Facilities Council

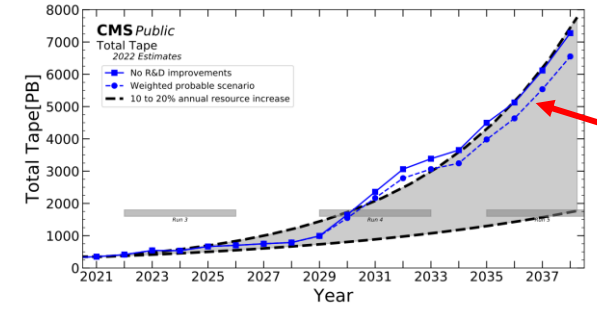
Performance requirements for WLCG storage - RAL perspective

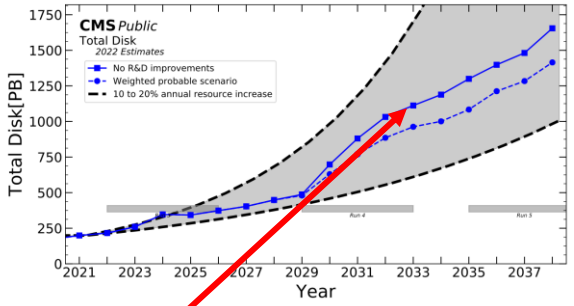
Alastair Dewhurst



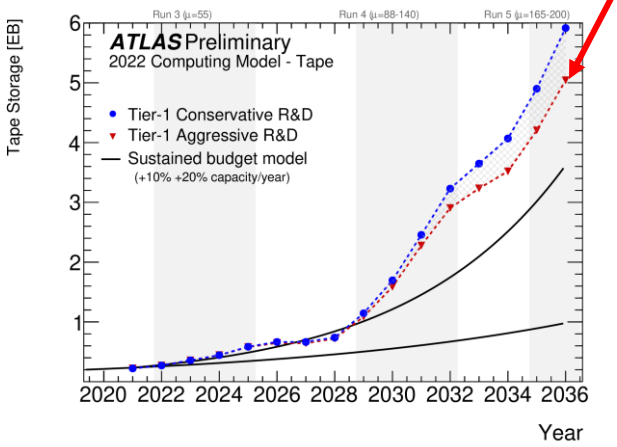
HL-LHC Predictions

- ATLAS and CMS have predictions on their capacity requirements for HL-LHC.
  - The capacity requirements are difficult but there nothing is mentioned on performance.

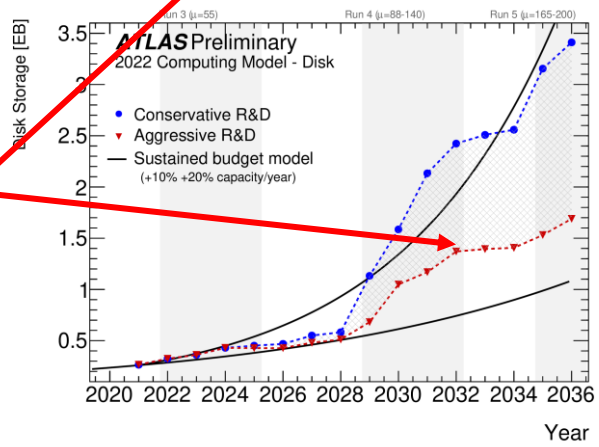




This is not a flat cost tape model!



Disk capacity looks ok, but what about performance?



Science and Technology Facilities Council

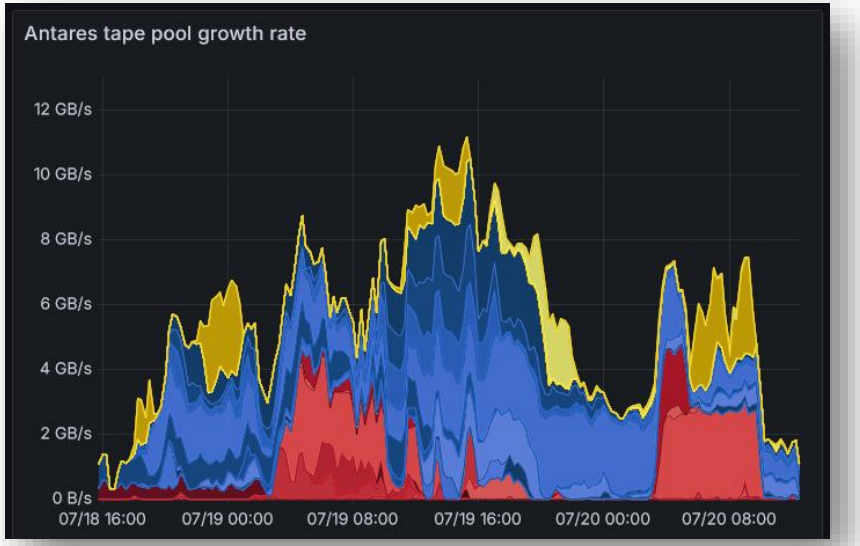
Tape

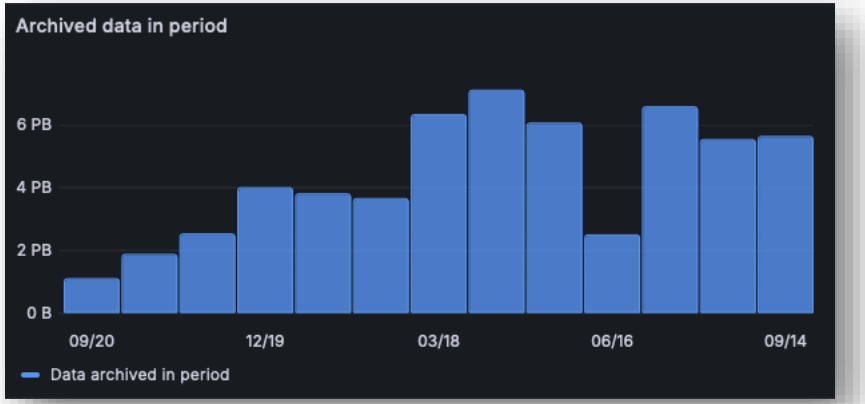
Tape performance vs cost

- For the ~20 years between the start of Run 1 and start of HL-LHC:
  - 100 x capacity improvement per Tape (26% growth per year).
  - 6 x throughput improvement per Tape Drive (9% growth per year).
- Tape drive costs are rising faster than inflation:
  - 2018 = £8k for TS1160 drive
  - 2024 = £20k for TS1170 drive
  - Technology is getting more complex / shrinking market.
- Currently with LTO-9:
  - Increase capacity by 10PB = £65,000
  - Increase throughput by 1GB/s = ~£25,000

Tape Throughput

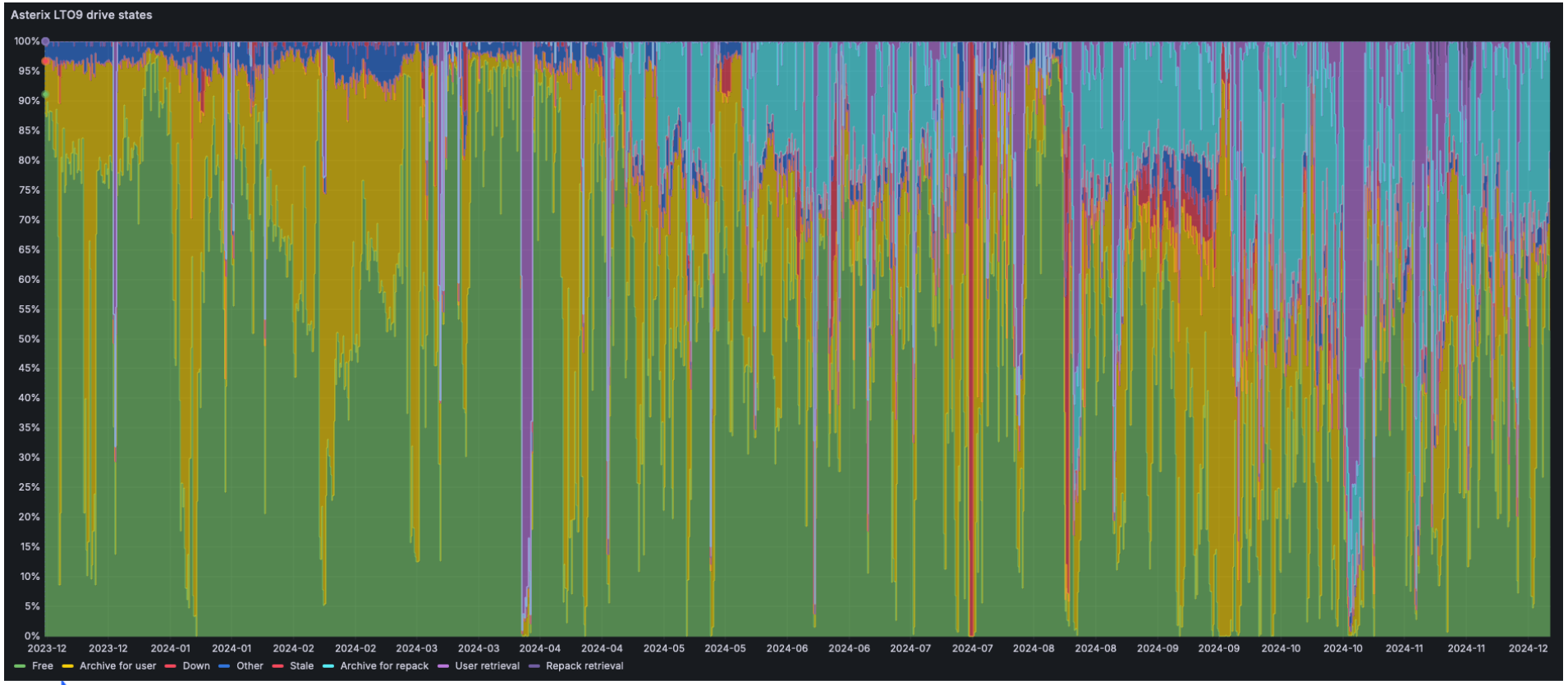
- At RAL we have 36 x Tape Drives with a maximum throughput of 400MB/s each.
- 14GB/s maximum theoretical throughput which during peak load we are getting close to.
- Since data taking started this year, we have averaged over 1PB a week written to tape.





Tape utilization

- Plot shows the percentage utilization of tape drives for the Tier-1 at RAL over the last year.
- We are running a repack campaign which is using ~25% of drives
- Free drives are becoming less common but usage is still spikey.



Green = Free
Yellow = Archival
Purple = Retrieval
Blue = Repack
Red/Orange = Down

Implications

- Due to the nature of Archival Storage there is relatively little R&D could do to improve the situation for HL-LHC.
- If we need more tape capacity than modelled via 10 – 20% annual growth then we need to allocate money from Disk/CPU.
- At RAL, to keep throughput scaling with capacity we would need to increase costs by 10 – 15%.
  - Increase of ~£250k on a total hardware budget for tape of ~£2million over the next 4 years.

Potential mitigations

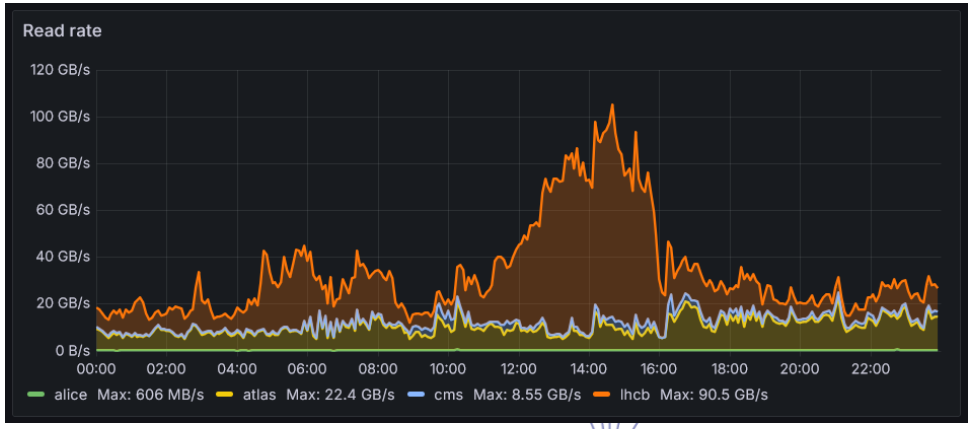
- To mitigate potential performance bottlenecks we can:
  - Delete less (only write what you know you want to keep)
  - Read less
  - Wait longer
- ATLAS currently have ~240PB stored on MCTAPE.
  - Theoretically this could all be regenerated (with some spare GPUs on an HPC).
- Could VOs cope if the average time to recall files in future increased by a factor of 2 – 5?
- Tape remains comfortably the cheapest way to store data (especially long term).
  - Maybe we just have to accept that capacity growth will be slower as more money needs to be spent on throughput.

Science and Technology Facilities Council

Disk

Disk performance

- Since 2016 HDD performance (both IOPs and throughput) has barely changed.
  - Capacity has risen by a factor of 3 (15% increase per year).
- HDD are now considered nearline storage.
  - If you gave a vendor WLCG performance requirements they would offer a flash dominated solution.
- For certain workloads RAL already see the storage (almost) hitting its I/O limits.
  - For optimal workflows we still have factor of 2 – 3 before we reach throughput limits.
- Waiting for disk will become normal.

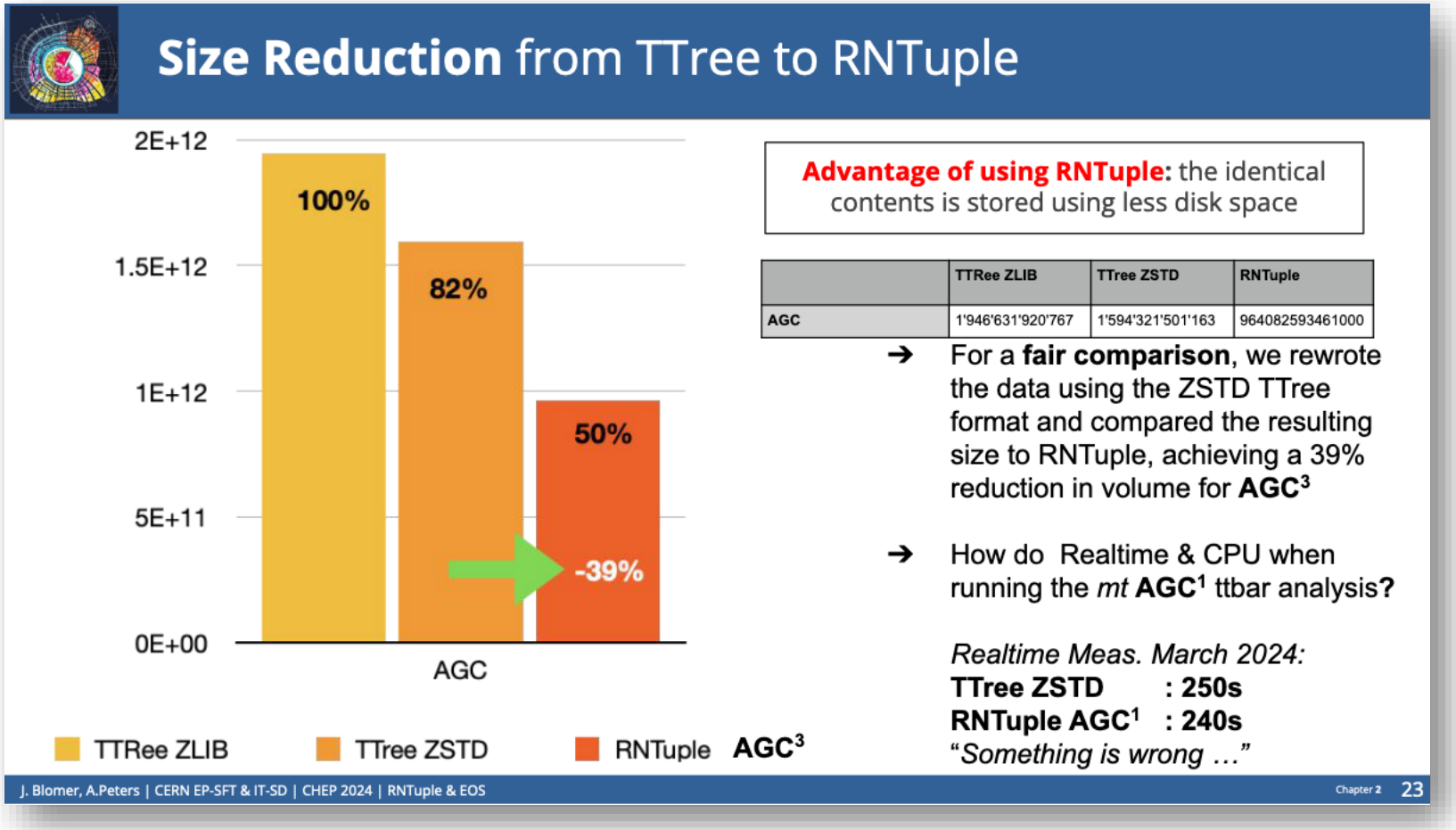


Disk Performance Challenges (1)

[CHEP talk on ROOT RNTuple and EOS: The Next Generation of Event Data I/O]

Amazing improvements have been made in storing events to disk more efficiently.

This does not reduce the usage so we would expect ~60% more usage for the same capacity.



Disk Performance Challenges (2)

- HPC platforms are expected to provide a significant fraction of future compute.
- In general HPC provide a much higher ratio of compute to storage/network compared to a Grid site.
  - Work with less demanding I/O requirements is likely to be sent to HPC.

Site	CPU	GPU	Power (MW)	SSD Storage (PB)	HDD Storage (PB)	External Network (Gb/s)
RAL Tier-1	768	0	0.4	0	73	400 + 200 LHCOPN
Frontier ORNL	9472	37,888	21	11.5	679	400
Leonardo	1536	13,824	6	4	110	100
Isambard-AI	5448		5	25	0	20

Flash to the rescue?

- An Enterprise NVMe drive can do:
  - 1million IOPs (vs 200 for HDD)
  - 3GB/s throughput (vs 250MB for HDD).
- Flash storage is currently more expensive:
  - Quotes RAL receive indicate factor of 3.8 larger upfront cost.
- Flash power usage is significantly lower than HDD
  - Flash drives are much bigger than HDD while energy usage remains roughly constant.
  - TCO difference is reduced to factor 2.5 – 3.0 depending on energy costs + data centre costs.
- On a 5 – 10 year timescale we need to be using Flash.

Science and Technology Facilities Council

Proposal

Pledge Proposal

- For the tape pledge we should add a performance requirement:
  - Minimum 1GB/s nominal throughput per 30PB of Tape media.
  - Recommended 1.5GB/s nominal throughput per 30PB of Tape media.
- For the disk pledge we should split it into two:
  - HDD
  - Flash
- Initially there would be no requirement to pledge flash storage but sites could pledge this if they have it deployed.
  - In future VOs are likely to request some and may have a ratio of Disk to Flash they accept.
- For pledges we should use nominal capacity (i.e. what the vendor tells us) as that is easier to measure by funding bodies.

Science and
Technology
Facilities Council

Questions?

IBM Slide from March 2023

Tape Drive History and Roadmap

LTO Generations	LTO-5	LTO-6	LTO-7	LTO-8	LTO-9	LTO10
Capacity (Native)	1.5 TB	2.5 TB	6 TB	12.0 TB	18 TB	Up to 36 TB
Other Format Capacities	800 GB (400 GB R/O)	1.5 TB (L5) (800 GB R/O)	2.5 TB (L6) (1.5 TB R/O)	9 TB (M8) 6 TB (L7)	12 TB (L8)	Up to 18 TB (L9)
Native Data Rate	140 MB/s	160 MB/s	300 MB/s	360 MB/s	Up to 400 MB/s	Up to 500 MB/s

2008 2010 2011 2013 2014 2015 2017 2017 2021 2018 2023 LTO10+6-12 months

	TS1130	TS1140	TS1150	TS1155	TS1160	TS1170	TS1180
New Format Capacity (Native)	1 TB (JB) 640 GB (JA)	4 TB (JC) 1.6 TB (JB)	10 TB (JD) 7 TB (JC)	15 TB (JD)	20TB (JE) 15 TB (JD) 7 TB (JC)	Up to 50 TB (JF)	Up to 120-80 TB (JG) Up to 50 TB (JF)
Other Format Capacities (Native)	700 GB (JB) 500 GB (JA) 300 GB (JA)	1 TB (JB) 700 GB (JB) (All JA R/O)	4 TB (JC)	7 TB (JC) 4 TB read only (JC)	10 TB (JD) 7 TB (JC) 4 TB (JC)		JE / JF
Native Data Rate	160 MB/s	250 MB/s	360 MB/s	360 MB/s	400 MB/s FC-16	Up to 500 MB/s FC-16	Up to 1000 MB/s FC-32, 25 GibE

TS1170 now released with 50TB capacity and 400MB/s performance

Media Investment Protection
next tape drive generation re-writes JE-media with 50% more capacity **and reduce €/GB by 50%**

Full presentation available: https://www.spectrumscaleug.org/wp-content/uploads/2023/03/SSUG23DE-Sustainability-with-Spectrum-Scale-Spectrum-Archive-and-IBM-Tape.pdf

Appendix – Tape performance

- Why 1GB/s per 30PB?
- 1GB/s for 1 year = 31.5PB of data moved. i.e. we can write a complete copy of the data per year.
- In a typical year:
  - Write 20% new data.
  - Delete 5%, repack and re-write = 15%
  - Recall 10% data
  - Re-pack 15% data = 30% tape usage (read+write)
- This assumes perfect efficiency and would average 75% tape drive usage for the entire year.

Ways to implement Flash?

Cache	Standalone storage
Transparent to VOs	Requires VO management
Difficult to account for	Easy to account for
Some benefit to all workflows	Full benefit to some workflow
No VO Control	Full VO control

We will (obviously) need a combination of both however development effort towards caches is currently more advanced.