

HEPiX Spring 2021 Workshop

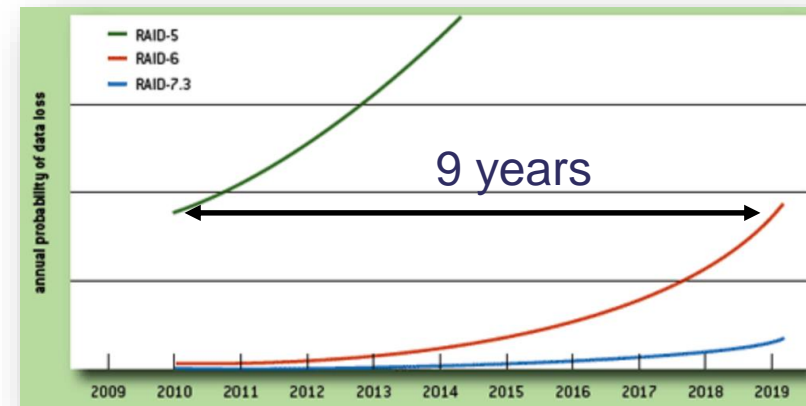
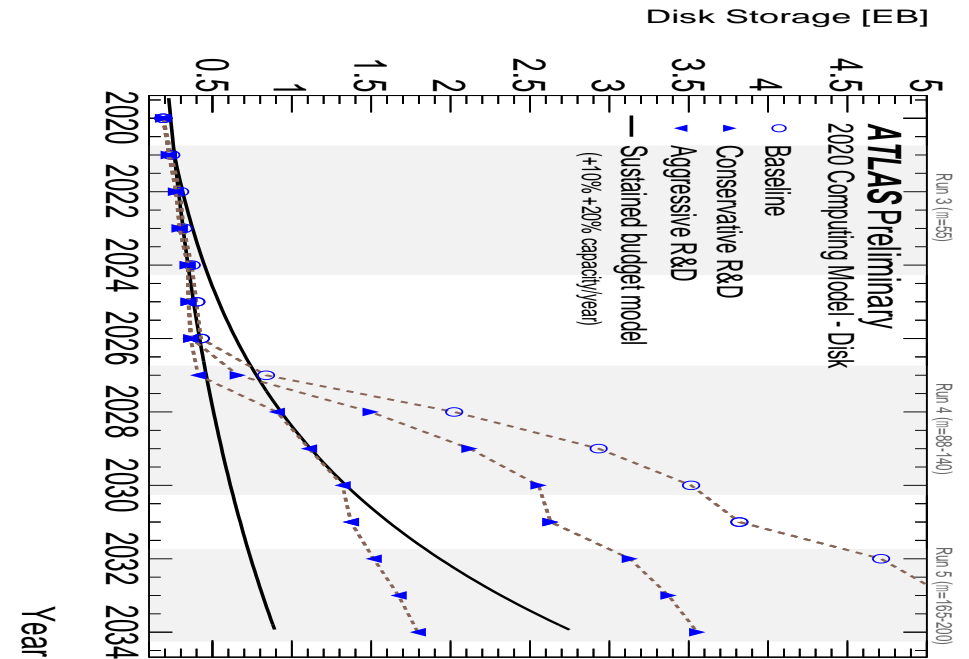
Erasure Coding WG

Alastair Dewhurst, Shikegi Masawa, Andreas-Joachim Peters



Context

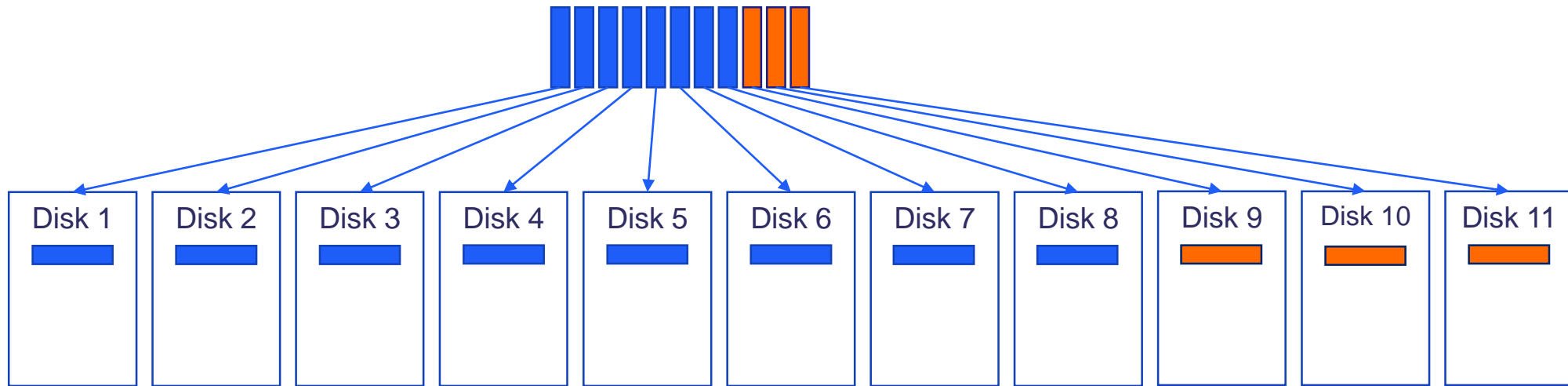
- Predicted HL-LHC data rates mean that even with an aggressive R&D program it could be difficult to store all the data needed.
- A lot of sites use RAID6 for data resilience, although this is now running into scaling issues.
- [I gave a talk at the HSF / WLCG Workshop in November 2020.](#)
 - Erasure Coding can save money over Replication.
 - Erasure Coding is not trivial to implement.



<http://queue.acm.org/detail.cfm?id=1670144>

What do we mean by Erasure Coding?

- Erasure Coding is a very general term.
 - RAID6 is a type of Erasure Code.
- Erasure Coding takes a file and splits into k chunks.
- It generates m additional chunks such that the original file can be reconstructed from any k out of the $k+m$ chunks.
- Each of the chunks are stored on **different** storage nodes.



EC WG mandate

- The HEPiX Erasure Coding Working Group purpose is to help solve some of the data challenges that will be encountered during HL-LHC by enabling sites to store data more efficiently and robustly using Erasure Coding techniques.
 - To provide a forum to allow sites to identify the best underlying storage for their use cases.
 - To provide recommendations on how to configure storage to effectively use Erasure Coding.
 - To work with VOs to ensure that their workflows run efficiently when accessing data stored via Erasure Coding.

Forum

- We have created a mail list:
 - hepex-erasure-coding-wg@hepex.org
- We have created a Indico group:
 - <https://indico.cern.ch/category/13757/>
- We want to get feedback from the community regarding what they need.
- We don't intend to organize regular meetings, small updates could be discussed at other meetings (e.g. DOMA)
- We would like to arrange more detailed discussions (~quarterly).
 - Potentially, focus each meeting on a particular topic.
 - Schedule them with other things (e.g. pre-GDB, EOS Workshop etc).
- First meeting in June this year?

Performance

- Erasure Coding has practical implications:
 - ~Double (internal) network traffic as data is stored and retrieved on different servers.
 - Potentially, high latencies as data may need to be re-assembled.
 - Modifying files can have a significant overhead if the parity chunks need to be all re-calculated.
- Erasure Coding has been traditionally associated with Object Stores but that isn't required.
- It is vital that we stay engaged with the VO to make sure workflows can be run efficiently.

Summary

- Erasure Coding can provide reliable storage which can be used to replace RAID6 at a significant cost reduction compared to replication.
 - It is however not (yet) straight forward for sites to implement Erasure Coding.
- We have setup a new HEPiX working group to focus on Erasure Coding.
 - We welcome feedback from the community as to the best way forward.

Backup

RAL Operational experience

- Most common failure is a bad sector on a drive.
 - Daily occurrence.
- Complete disk failures are the next most common type of failure.
 - 1 - 2 a week. ~1.5% failure rate.
 - We have had a bad batch of drives. ~200 failures in a year out of 2200 total. ~10% failure rate.
- An entire node needing an intervention ~once a month.
 - Being able to easily remove hardware allows regular rolling upgrades, which will provide better reliability.
- As a site you don't want to be operating with no resilience. i.e. you want to be able to survive another failure.
 - $M = 3$ has double the comfort zone compared to $M = 2$ (or RAID6)
- We have comfortably managed failures that would have resulted in significant data loss with hardware RAID.

Cost?

- First order:
 - EC 8 + 3 means 72.7% usable space.
 - 2 x Replication means 50% usable space. } Need to purchase 45% more capacity with Replication
- Erasure Coding has higher CPU and Memory requirements compared to Replication. Assume:
 - 2 x CPU
 - 1.5 x Memory } Adds 3 - 5% to cost of hardware for Erasure Coding
- Assume Erasure Coding requires larger overhead ~5%
- Upfront costs for same amount of usable storage with Replication ~25% more than Erasure Coding.
- Power cost over 5 years ~40% upfront cost.
 - Additional ~20% cost over the lifetime of the hardware for Replication .
- For RAL, Total Cost of ownership: Erasure Coding is about 70% the cost of Replication.