Maintaining a large scale, very active tape archive

MSST 2018

Stephen Richards – Senior Analyst

Stephen.Richards@ecmwf.int

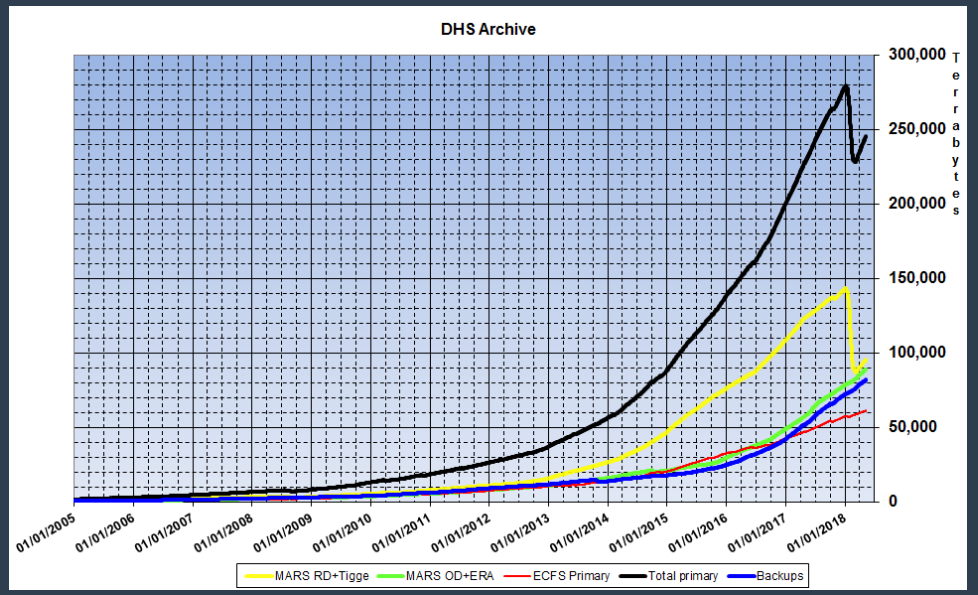
ECMWF

ECMWF

- The two roles of the Centre
  - Run the operation model to produce medium and seasonal forecasting, ie predict the weather
  - Research meteorology in general, mostly with an aim to improve its global IFS model.
- 2 Cray XC40s each with ~130,000 cores locally connected to Sonexion Lustre via Infiniband (FDR)
- HPC2020 remit
  - More novel architectures being considered
  - Reconsidering how the storage layers work between HPC and the archive
- Move of the Data Centre in 2019 to Bologna in Italy

The Current Archive

- Size: 245PB over 320 million files
- Growth today: 275TB/day, next two phases (~605TB & 1.3PB/day)
- MARS (Metrological structured data) and ECFS (Posix) archives
- Disk front-end: 7.5PB
- Tape environments:
  - Primary
    - 4 SL8500 libraries
    - 170 T10kD and 56 T10kC tape drives
    - 34,500 T2 tape media
  - Secondary (DR)
    - 1 TS3500
    - 12 LTO-7, 18 LTO-6



Current Archive continued

- Most drives are active throughout the day, predominately on read requests
- Robotic activity rates (average 840 exchanges per hours, peeking at 1320/hr)
- Deletion of 60PB over last few months

Oracle’s decision to withdraw from Enterprise tape

- Evaluated to two remaining options (TS11xx(Jaguar) or LTO).
- Building costing models for both
- Why disk alone is not an option
  - Cost
  - Level of defence against malicious attack

TS1150/55 and LTO Benchmarks

- T10000D – Used as a baseline
- TS11xx – Similar but mostly a better performance than T10000D
- LTO-7 – Struggled with small file reads

TS1150/55 and LTO testing

- Using 30 drives LTO-7 tested with MARS experiments
- Findings
  - Possible to achieve close to T10000D performance, but at a cost in reading much more data and avoiding many more read/skip operations on the tape
  - Consequential increase in data flows to HPSS client (MARS)
- Decision to go to tender for TS11xx and LTO based libraries
  - Initially for TS11xx drives but in the future having the option to switch to LTO, if necessary

Move of Data Centre from the UK to Italy

- Options for moving the archive
  - Existing archive only moved to Italy once operations are switched to Bologna
  - Running two archives and two underlying HPSS instances one on each site.
  - Either merge the metadata of the two HPSS instances into one large instance
  - Or possibility maintain two instances long-term, an operational and a research instances

- Benefits/drawbacks
  - Expected archive size 400PB – 0.5EB (late 2020)
  - Time & effort to copy data
  - Licensing
  - Development/Testing time
  - Available network bandwidth
  - Risks of moving tape media and equipment

Site (Artists impression)

Site (Asis)

Futures / Concerns

- Making use of RAO on tape
- Easier for ECFS than MARS
- Data Integrity
- Cloud for DR
- Direct to tape connectivity
- Worrying trend in the need for additional drives
- 67% of data on a single tape volume
- Disk/Flash alternatives
- Daily involvement of tape support
- Limited gene pool of tape vendors