Data management and curation at DiRAC

IRIS meeting

Jan 2023

Alastair Basden

Durham University

DiRAC data storage

- Mostly Lustre parallel file systems
  - Typically one per service per generation
    - i.e. each new system procures new storage
  - ~30-40PB total
- Different management approaches at different sites
  - In some cases, storage dies with the service (at EoL)
  - In some cases, storage outlives the service

Storage allocation

- A RAC process allocates storage to projects
- PIs are expected to remove their data within 3 months after the end of the project
  - Often this is not possible
  - Some projects last for 10 years or more
  - Curation of this data is therefore required

Data curation: current approach

- A mature, responsible user base
  - We like to be kind to them which pays dividends when we have requests
- Where possible, data is kept as required
  - users are aided if required to move data to a new service
    - Sometimes copying is done automatically
- Tape archives are available at some sites
  - For archiving upon request (and eventually self-archiving/retrieval)
- Some data still exist from the start of DiRAC (>10 years)
  - e.g. the Eagle simulation outputs on COSMA are still in active use

Making data FAIR

- The Virgo Database
  - A (now ageing) web interface to Virgo datasets
  - Enabling SQL queries to cosmological data
- Sciserver
  - Modern web portal based on Kubernetes
  - Access to data via Jupyter and other tools
    - Bringing the compute to the data

DiRAC Data Curation project

- An ongoing project to provide data curation for >15 years
  - Including tape archives where appropriate
  - Cross-site storage
  - Global distributed storage via StorJ
    - Good bandwidth at any location
  - FAIR principles
  - Automatic metadata tagging and stripping
  - Funded scheduled hardware refreshes
  - Funded staff effort
  - Integration with SKA