iRODS on OSDC

Exploration of iRODS as a data-management and archival store running on OSDC infrastructure.

Many groups now use iRODS to manage, store and preserve research data (www.irods.org).
iRODS can work with many file systems and with S3 to manage data in its storage services.
The work would first develop a mapping of iRODS storage services onto OSDC infrastructure.
It would then look at rules for distributing and replicating files as they arrive on to OSDC servers, so that these data can be used in distributed parallel contributions, such as map-reduce workloads.
An interesting research question is whether throughput can be optimised by distributing files according to some of their properties.

iRODS has a catalogue service that itself can be replicated, for doing the data management, for mapping from metadata to files and for triggering rules.
A next step would be to move this onto OSDC and then to investigate both file accession rates and file lookup rates as a function of file population on a typical OSDC cluster.
This could then be used to populate stores, e.g. using high-performance data transfers, with research data collections.

Rosa Filguera in the Edinburgh DIR group has made extensive use of data compression and data marshalling to accelerate transfers. iRODS rules triggered on file ingest and on file read, could apply similar strategies in the context of iRODS data. Exploring appropriate choices of when to compress/decompress, etc., and how to select compression methods, as part of iRODS rules could discover substantial performance gains.
This should lead to a model, applicable on any OSDC cluster to optimise file ingest and access according to file type or file content. The threshold parameters would vary between clusters and a good research goal would be to be able to learn these for each site automatically.