Standard (Alternative) data management tools

Daniel Traynor, QMUL
GridP 40, 10/04/2018

iRODS

- integrated Rule-Oriented Data System (iRODS). Open source project traces history back to 1995.
- Storage virtualisation of different disk and tape storage systems (sits on top of other storage systems e.g.posix, lustre,dcache). HSM capable. Load balancing. Concurrent access.
- A logical namespace across storage locations. Federation of sites.
- Controlled access through a number of different authentication mechanisms: Kerberos, secure username/password and Grid Security Infrastructure (GSI).
- A rule engine to automate data management and access according to defined policies.
- Client interfaces: web(idrop), cli, GUI (cyberduck!), gridFTP, webdav.

Policy examples

- Automatically replicate a file added to a collection into 3 geographically distributed sites.
- Periodically check integrity of files in a collection and repair/replace if needed/possible.
- Automatically pick a certain storage location based on user or collection or size or type.

iRODS - T2K

- Make data quality files available to world wide collaborators to be checked as soon as produced.
- This requires a data distribution system. iRODS rules used for access, replication and backup.
- Access the files is through a web or cli interface, with a read–only account

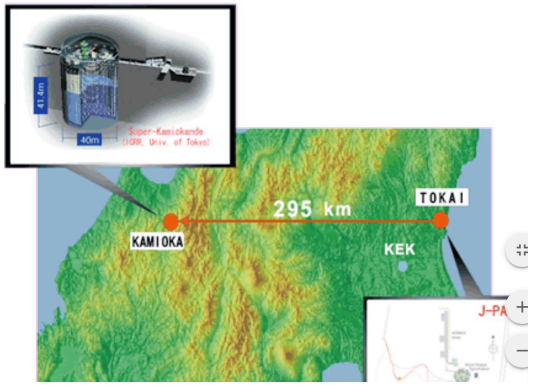
First iRODS Experience in a Neutrino Experiment

F. Di Lodovico[a], A. Hasan[b], Y. Iida[c], T. Sasaki[c]

[a]Queen Mary University of London, London, UK
[b]University of Liverpool, Liverpool, UK
[c]High Energy Accelerator Research Organisation (KEK), Tsukuba, Japan

Abstract

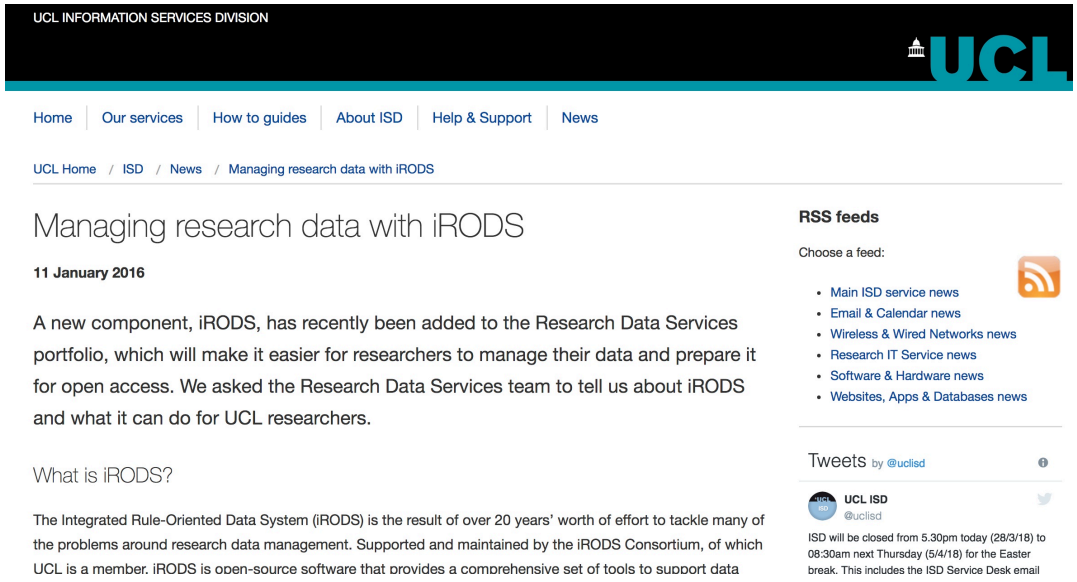
The T2K experiment is a long-baseline neutrino experiment with the primary aim of observing muon into electron neutrino appearance. An important task is the determination of the quality of the collected data. In this paper we describe the first experience gained in managing the T2K data quality data for the near detector such that it is stored and accessible to collaborators world-wide in a timely fashion. We also describe a set of rules developed for the project that have much more general applicability. This is the first use of iRODS in a neutrino experiment.



https://irods.org/uploads/2011/DiLodovico-Neutrino-T2K-paper.pdf

iRODS - UCL

- The push towards making data openly, If data is going to be discoverable and effectively re-used it is essential that adequate metadata is generated that provides a full and accurate description of where the data came from and how it should be understood. As data repositories grow in size and complexity, a solution is needed which will connect everything together and enable researchers to search across multiple types of repository to identify and access relevant data.



UCL member of iRODs consortium

iRODS - others

- eudat (european open science cloud): used in bsafe storage product. Provide federated access to data for collaborations.
- sanger: use iRODS to manage petabytes of genomics data with diverse scientific needs, rules for backup to slough,
- emedlab: posable use for allowing secure data access from VMs to local GPFS storage.
- Panasas(ActivStore array), HGST(Active Archive System),

CDMI

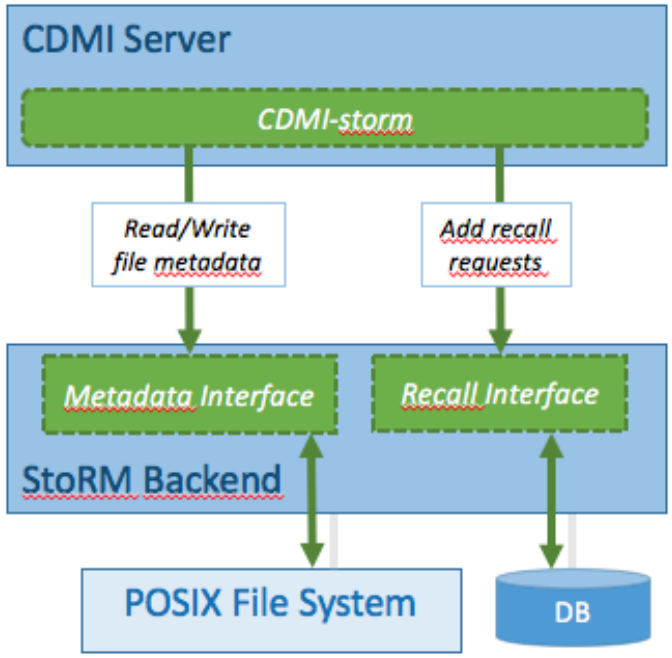
Cloud Data Management Interface

Industry standard implemented in several products.
https://www.snia.org/cdmi

Cloud Data Management Interface

The **Cloud Data Management Interface** (**CDMI**) is a SNIA standard that specifies a protocol for self-provisioning, administering and accessing cloud storage.[1]

CDMI defines RESTful HTTP operations for assessing the capabilities of the cloud storage system, allocating and accessing containers and objects, managing users and groups, implementing access control, attaching metadata, making arbitrary queries, using persistent queues, specifying retention intervals and holds for compliance purposes, using a logging facility, billing, moving data between cloud systems, and exporting data via other protocols such as iSCSI and NFS. Transport security is obtained via TLS.



Functionality similar to Amazon's S3 cloud storage interface.

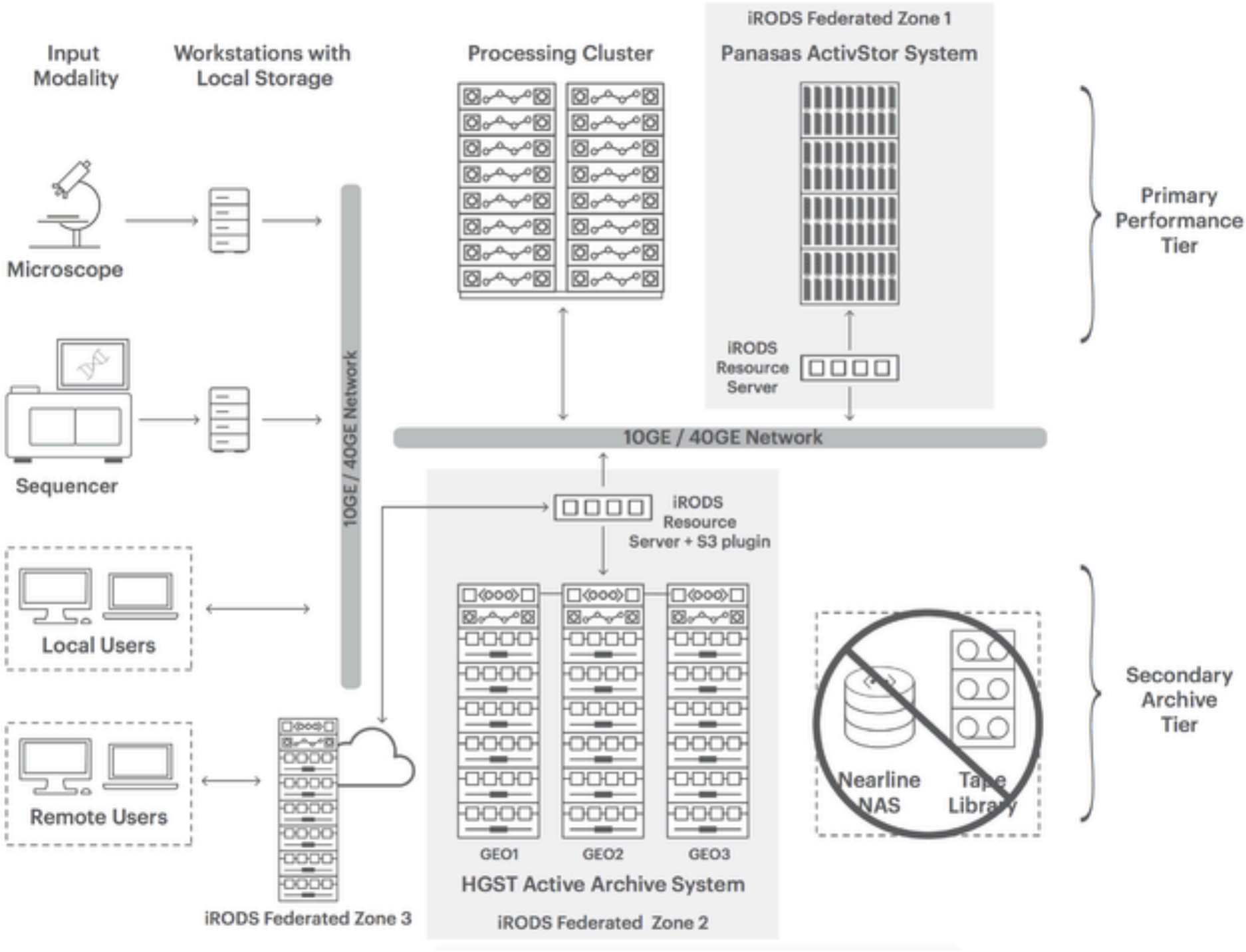
dcache and storm plugins

CDMI StoRM implements the Java Service Provider Interface cdmi-spi to provide a plugin for the INDIGO DataCloud CDMI server in order to support StoRM as storage back-end and allow users to negotiate the Quality of Service of stored data.

Comments

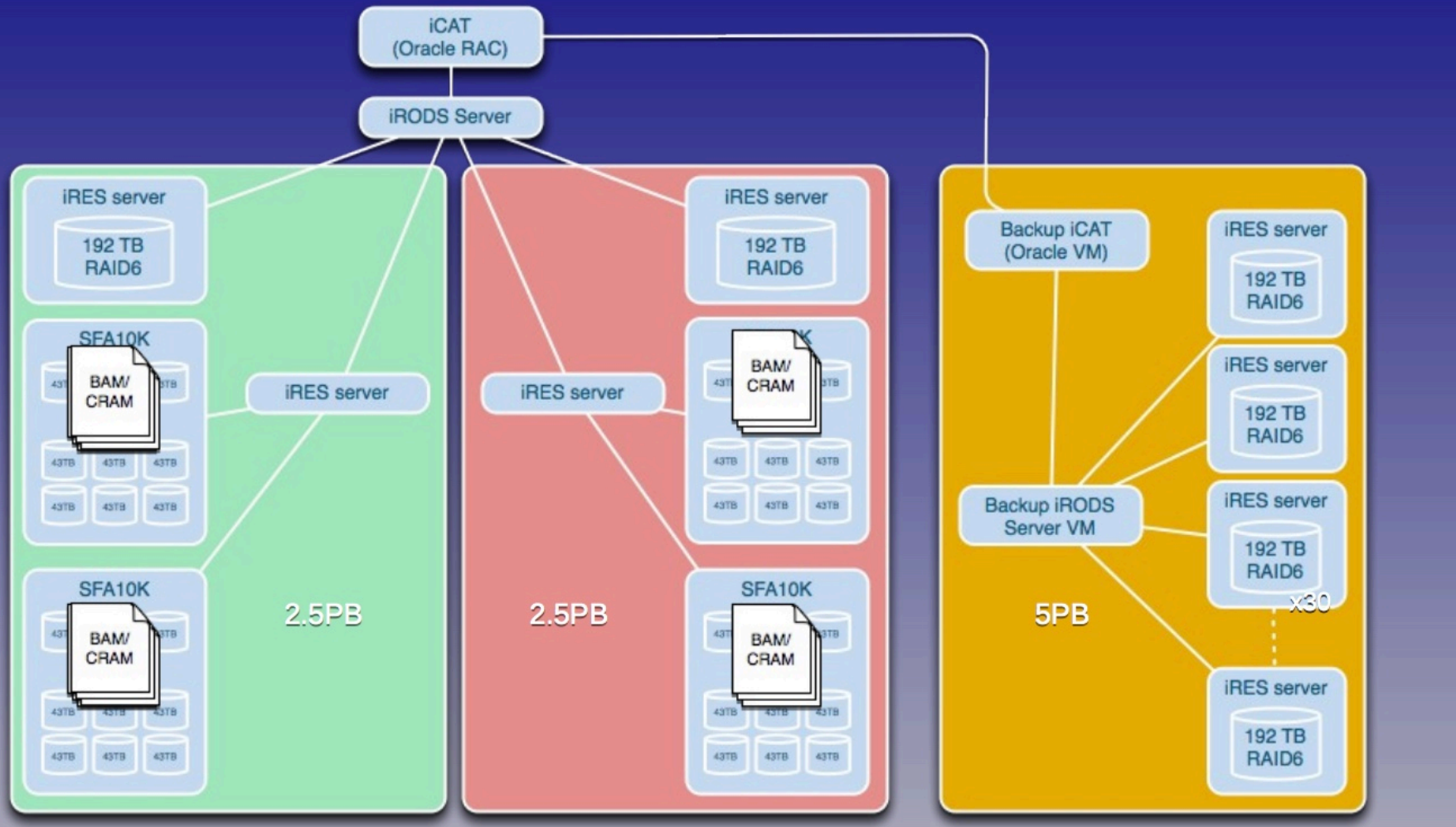
- iRODS mature product under active development and used widely for data management. Worth investigating further. Lustre and dcache plugins.
- iRODs may become a “must” to support a wider science base.
- CDMI pugins for dcache and StoRM make cloud and grid more interoperable. Support and usage not so clear.

backup

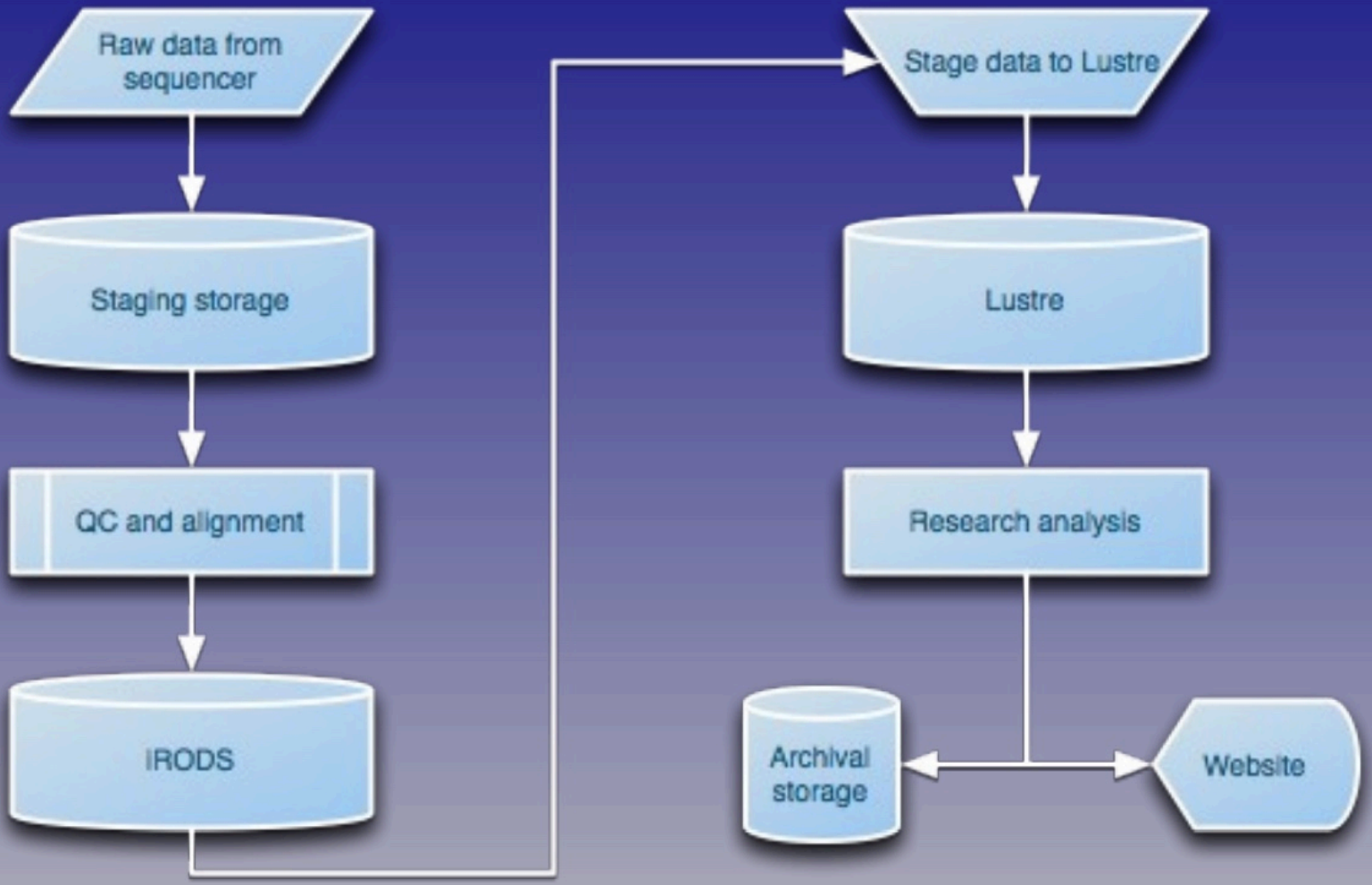


https://www.theregister.co.uk/2016/10/07/building_irods_for_saving_life_sciences_back/

Capacity Expansion and Offsite Migration



Sanger NGS Data Flow



wellcome trust sanger institute

emedlab

iRods – Running inside OpenStack VMs

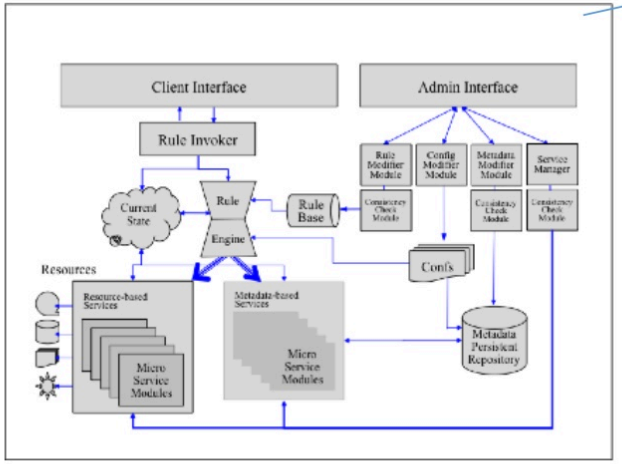
tranSMART v1.2

Arvados

i2b2

elasticluster

Linux



Provides a proper research data management

layer with an API-based user interface

Make data on GPFS available to VMs