Science & Technology Facilities Council
Rutherford Appleton Laboratory

JASMIN: the Joint Analysis System for big data.

JASMIN is designed to deliver a shared data infrastructure for the UK environmental science community. We describe the hybrid batch/cloud environment and some of the compromises we have made to provide a curated archive inside and alongside various levels of managed and unmanaged cloud ... touching on the difference between backup and archive at scale. Some examples of JASMIN usage are provided, and the speed up on workflows we have achieved. JASMIN has just recently been upgraded, having originally been designed for atmospheric and earth observation science, but now being required to support a wider community. We discuss what was upgraded, and why.

Bryan Lawrence

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Science & Technology Facilities Council
Rutherford Appleton Laboratory

Institutional Landscape

RESEARCH COUNCILS UK

Science & Technology Facilities Council

(Biotechnology and Biology) BBSRC 20 Years of Pioneering Great British Bioscience

Arts & Humanities Research Council

EPSRC Engineering and Physical Sciences Research Council

MRC Medical Research Council

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

NERC SCIENCE OF THE ENVIRONMENT

British Geological Survey NATURAL ENVIRONMENT RESEARCH COUNCIL

National Oceanography Centre NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science NATURAL ENVIRONMENT RESEARCH COUNCIL

British Antarctic Survey NATURAL ENVIRONMENT RESEARCH COUNCIL

CEH Centre for Ecology & Hydrology NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation NATURAL ENVIRONMENT RESEARCH COUNCIL

+ Universities, big and small ...

National Centre for Atmospheric Science NATURAL ENVIRONMENT RESEARCH COUNCIL

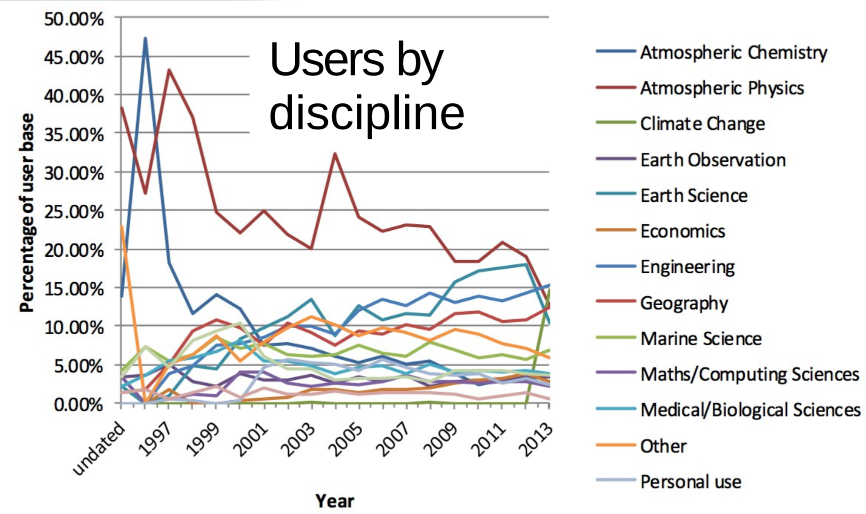
National Centre for Earth Observation NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival SCIENCE AND TECHNOLOGY FACILITIES COUNCIL NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival

Exist: *"to support environmental science, further environmental data archival practices, and develop and deploy new technologies to enhance access to data."*

-> Curation and Facilitation



Curation: Four Data Centres

- British Atmospheric Data Centre
- NERC Earth Observation Data Centre
- IPCC Data Distribution Centre
- UK Solar System Data Centre

(BADC, NEODC, IPCC-DDC, UKSSDC)

Over 23,000 registered users!

+ active research in curation practices!

Facilitation:

- Data Management for scientists (planning, formats, ingestion, vocabularies, MIP support, ground segment advice etc)
- Data Acquisition

  (archiving 3rd party data for community use)
- JASMIN Support

  (Group Workspaces, JASMIN Analysis Platform, Cloud Services, Parallelisation)

is-enes INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

LTDP Long Term Data Preservation

contrail open computing infrastructures for elastic services

Copernicus The European Earth Observation Programme

UK CLIMATE PROJECTIONS USER INTERFACE

Data policy

www.nerc.ac.uk/research/sites/data/policy/data-policy.pdf

The environmental data produced by the activities funded by NERC are considered a public good and they **will** be made openly available for others to use. NERC is **committed to supporting long-term environmental data management** to enable continuing access to these data.

NERC **requires** that all environmental data of **long-term value** generated through NERC-funded activities must be **submitted** to NERC for long-term management and dissemination.

… DMP … All NERC-funded projects must work with the appropriate NERC Data Centre to implement the data management plan, ensuring that data of long-term value are submitted to the data centre in an agreed format and accompanied by all necessary metadata;

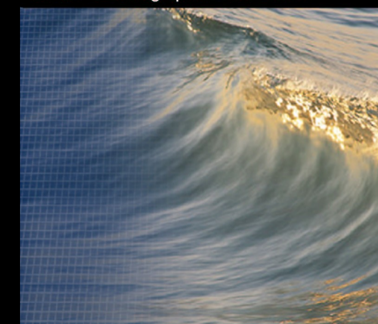
NERC Data Centres

Hydrology:
National Water Archive

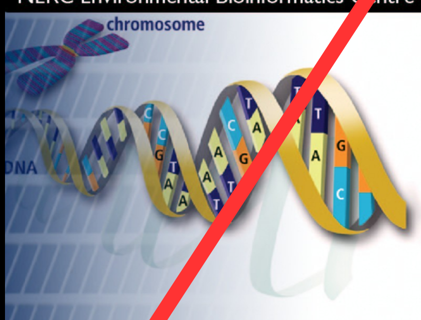
Atmosphere:
British Atmospheric Data Centre

Earth observation:
NERC Earth Observation Data Centre

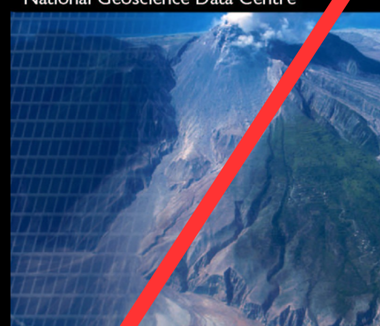
Ocean & marine:
British Oceanographic Data Centre



Bioinformatics:
NERC Environmental Bioinformatics Centre



Earth:
National Geoscience Data Centre



Terrestrial & freshwater:
Environmental Information Centre

Polar:
Antarctic Environmental Data Centre

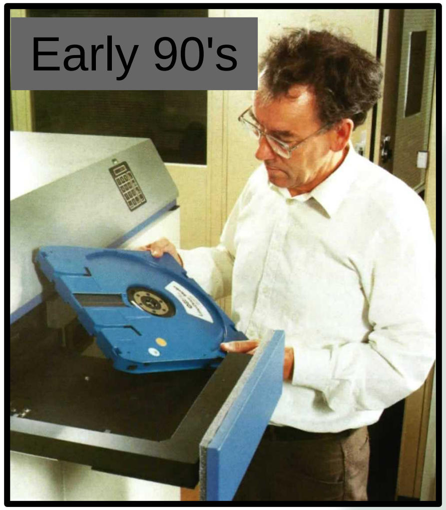
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

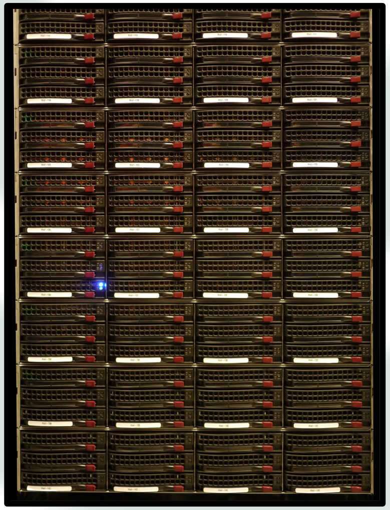
National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

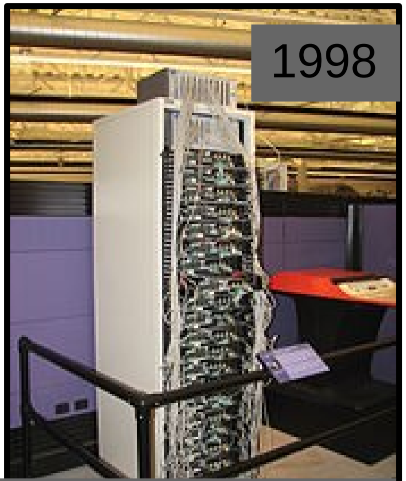
Curation … and … Facilitation

CEDA Evolution

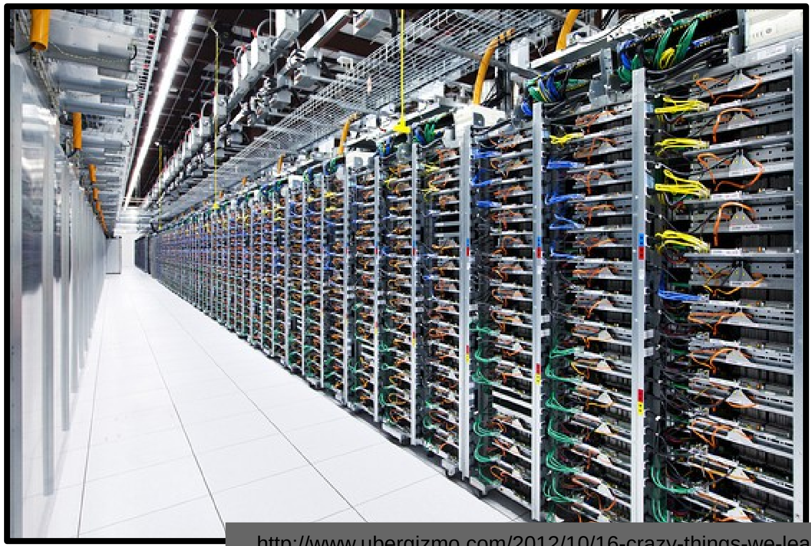


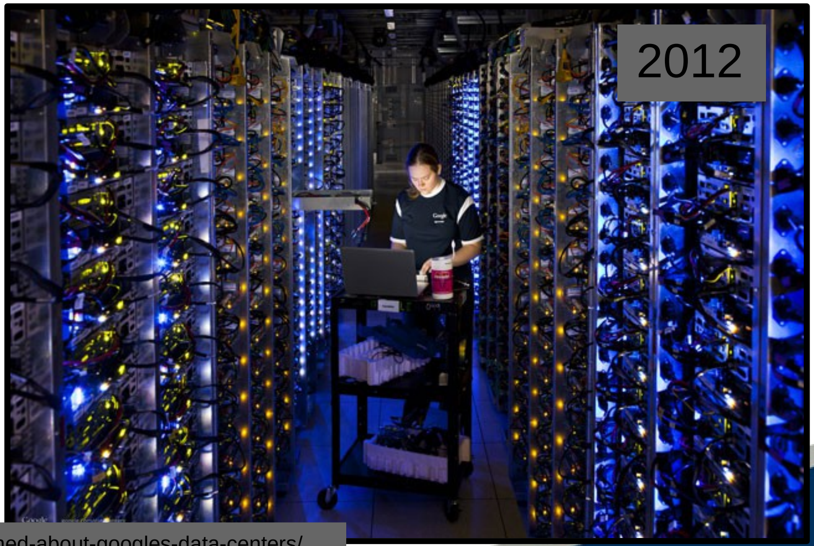


Eerily similar to Google's Evolution :-)



http://infolab.stanford.edu/pub/voy/museum/pictures/display/GoogleBG.jpg Wikipedia





http://www.ubergizmo.com/2012/10/16-crazy-things-we-learned-about-googles-data-centers/,
http://blogs.wsj.com/digits/2012/10/17/google-servers-photos/

Status Quo: UK academic climate computing

Data sources:

- ARCHER (national research computer)
- MONSooN (shared HPC with the Met Office – JWCRP)
- PRACE (European supercomputing)
- Opportunities (e.g. ECMWF, US INCITE programme etc)
- ESGF (Earth System Grid Federation)
- Reanalysis
- Earth Observation
- Aircraft
- Ground Based Observations

=> Big Data Everywhere!

Context: Climate Modelling!

“Without substantial research effort into new methods of storage, data dissemination, data semantics, and visualization, all aimed at bringing analysis and computation to the data, **rather than trying to download the data** and perform analysis locally, it is likely that the data might become frustratingly inaccessible to users”

A National Strategy for Advancing Climate Modeling, US National Acadamy, 2012

Solution 1: Take the (analysis) compute to the (distributed) data

How? All of:

(1) System: Programming libraries which access data repositories more efficiently;

(2) Archive: Flexible range of standard operations at every archive node;

(3) Portal: Well documented workflows supporting specialist user communities implemented on a server with high speed access to core archives;

(4) User: Well packaged systems to increase scientific efficiency.

(5) Pre-computed products.

Solution 1: Take the (analysis) compute to the (distributed) data

How? All of:

(1) System: Programming libraries which access data repositories more efficiently;

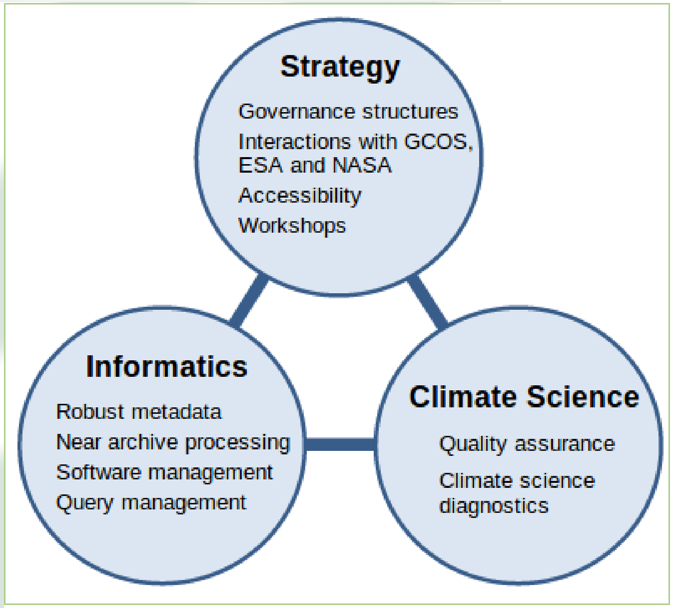
(2) Archive: Flexible range of standard operations at every archive node;

(3) Portal: Well documented workflows supporting specialist user communities implemented on a server with high speed access to core archives;

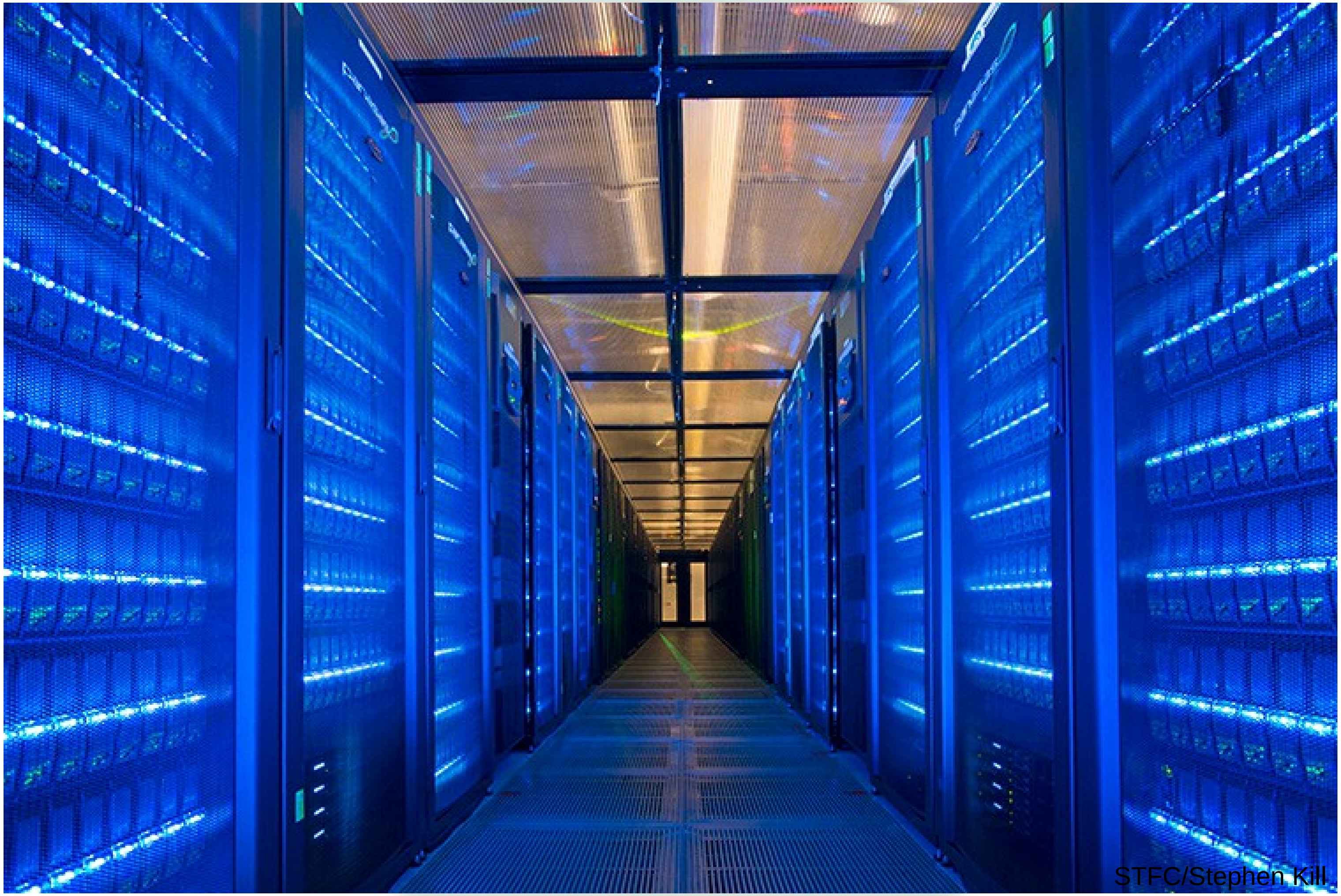
(4) User: Well packaged systems to increase scientific efficiency.

(5) Pre-computed products.

ExArch: Climate analytics on distributed exascale data archives (Juckes PI, G8 funded)



Solution 2: Centralised Systems for Analysis at Scale



Berkeley Dwarfs – Well defined targets for s/w and algorithms Similarity in Computation and Data Movement


Structured Grids


Unstructured Grids


Spectral Methods


Dense Linear Algebra


Sparse Linear Algebra


Particles (N-Body)


MapReduce (inc Monte Carlo)

Originally seven, from Philip Coella, 2004.

Nearly all involved in environmental modelling!

Subsequently another six added:

- Combinational Logic
- Graph Traversal
- Dynamic Programming
- Backtrack and Branch-and-Bound
- Graphical Models
- Finite State Machines

http://view.eecs.berkeley.edu/wiki/Dwarf_Mine

Data Ogres: Commonalities/Patterns/Issues

1) Different Problem Architectures, e.g:
- Pleasingly Parallel (e.g. retrievals over images)
- Filtered pleasingly parallel (e.g. cyclone tracking)
- Fusion (e.g. data assimilation)
- (Space-)Time Series Analysis (FFT/MEM etc)
- Machine Learning (clustering, EOFs etc)

2) Important Data Sources, e.g:
- Table driven (eg. RDBMS+SQL)
- Document driven (e.g XMLDB+XQUERY)
- Image driven (e.g. HDF-EOS + your code)
- (Binary) File driven (e.g. NetCDF + your code)

3) Sub-Ogres: Kernels & Applications, e.g:
- Simple Stencils (Averaging, Finite Differencing etc)
- 4D-Variational Assimilation/ Kalman Filters
- Data Mining Algorithms (classification/clustering) etc
- Neural Networks

(Not intended to be orthogonal or exclusive)

Modified from
Jha et al 2014
arXiv:1403.1528[cs]

Cloud 101: need to understand in order to exploit

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction." – NIST SP800-145

5 essential characteristics

- On-demand self-service
- Broad network access
- Resource pooling
- Rapid elasticity
- Measured service

3 service models

- IaaS (Infrastructure as a Service)
- PaaS (Platform as a Service)
- SaaS (Software as a Service)

4 deployment models

- Private cloud
- Community cloud
- Public cloud
- Hybrid cloud

Cloud: A method of having access to "some" of "a" computer (which you might configure) "all" of the time.

Batch Compute: (aka HPC) A method of having access to "all" of a (preconfigured) computer "some" of the time.

(Subtle differences in performance, not so subtle if you care about data flow, internal or in-bound!)

JASMIN:Joint Analysis System

J is for Joint

Jointly *delivered* by STFC:

CEDA (RALSpace) and SCD.

Joint *users* (initially):

Entire NERC community & Met Office

Joint *users* (target):

Industry (data users & service providers)

Europe (wider environ. academia)

A is for Analysis

Private (Data) Cloud

Compute Service

Web Service Provision

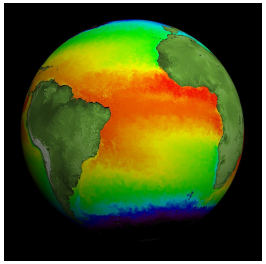
For

Atmospheric Science

Earth Observation

Environmental Genomics

… and more.



S is for System

£10m investment

at RAL

#1 in the world for big data analysis capability?

Opportunities

JASMIN is a collaboration platform!

within NERC (who are the main investor)

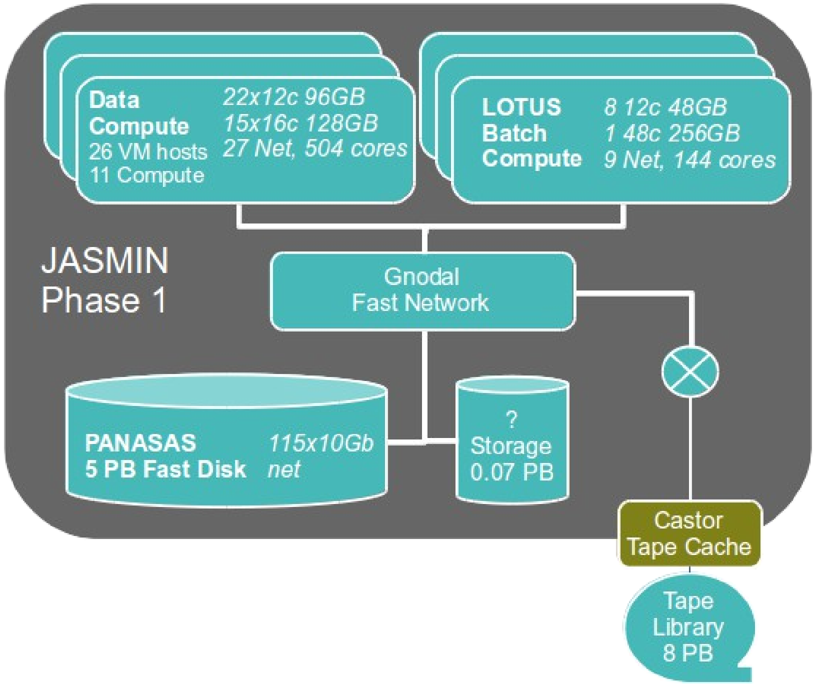
with UKSA (& the S.A. Catapult via CEMS)

with EPSRC (joined up national e-infrastructure)

with industry (as users, cloud providers, SMEs)

(CEMS:the facility for Climate and Environmental Monitoring from Space)

JASMIN as it was

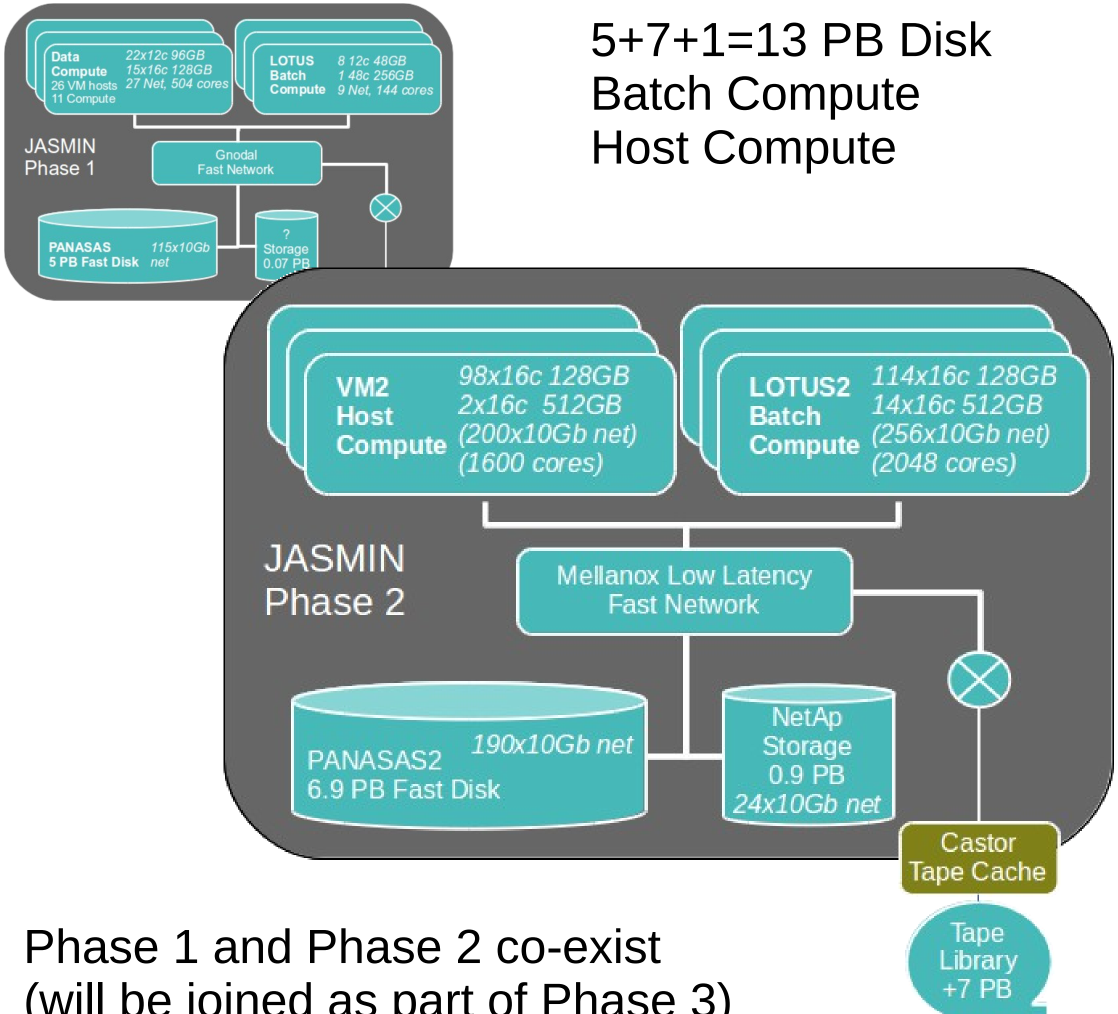


JASMIN is configured as a storage and analysis environment.

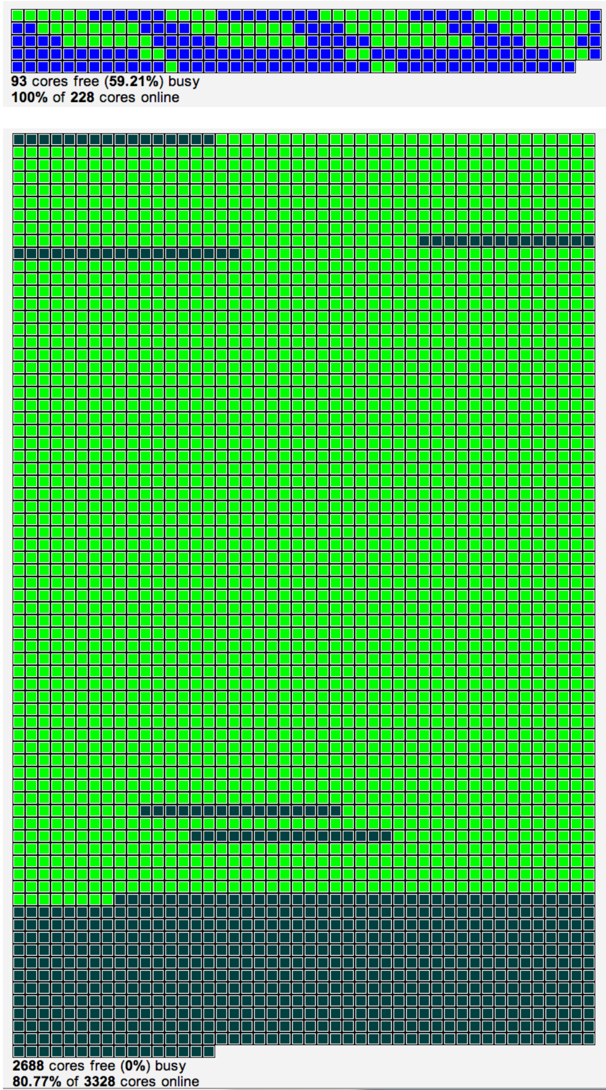
As such, it is configured with two types of compute: a virtual/cloud environment, configured for flexibility, and a batch compute environment, configured for performance.

Both sets of compute connected to 5 PB of parallel fast disk

JASMIN Now



5+7+1=13 PB Disk
Batch Compute
Host Compute



Phase 1 and Phase 2 co-exist
(will be joined as part of Phase 3)

JASMIN I/O Performance

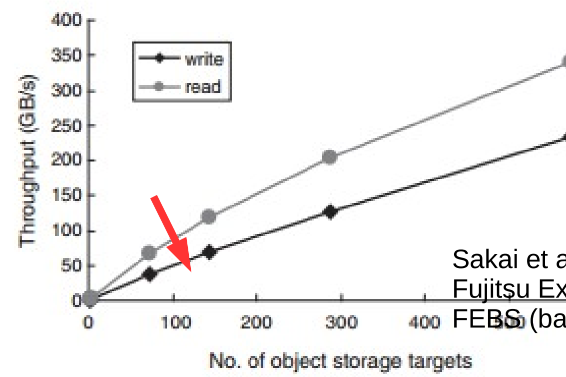
JASMIN Phase 2

- 7 PB Panasas (usable)
- 100 Nodes hypervisors
- 128 Nodes Batch
- Theoretical I/O performance

  Limited by Push: 240 GB/s (190x10 Gbit)
- Actual Max I/O (measured by IOR) using ~ 160 Nodes
  - 133 GB/s Write
  - 140 GB/s Read
  - cf K-Computer 2012, 380 GB/s (then best in world, Sakai, et al, 2012)
  - Performance scales linearly with bladeset size.

(JASMIN phase 1 is in production usage, so we can't do a "whole system" IOR, but if we did, we might expect to double up to ~ 300 GB/s overall – with JASMIN phase 3

to come later this year.)



Sakai et al 2012
Fujitsu Exabyte Filesystem, FEBS (based on Lustre)

Figure 7
Throughput performance (IOR benchmark).

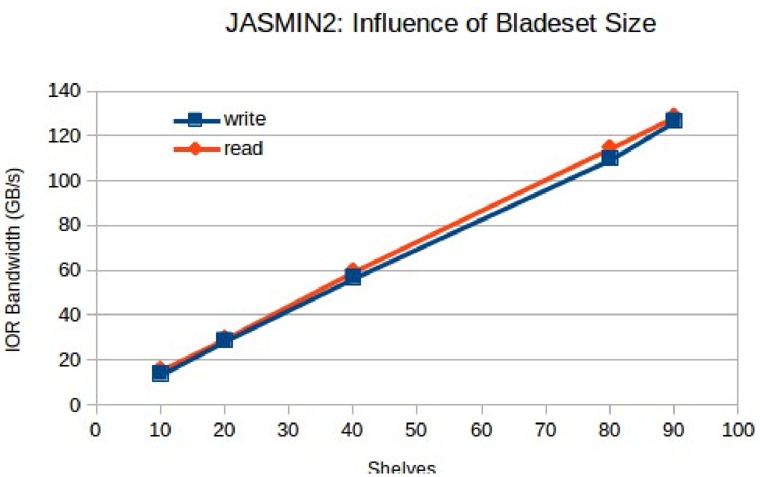
Performance v Reliability

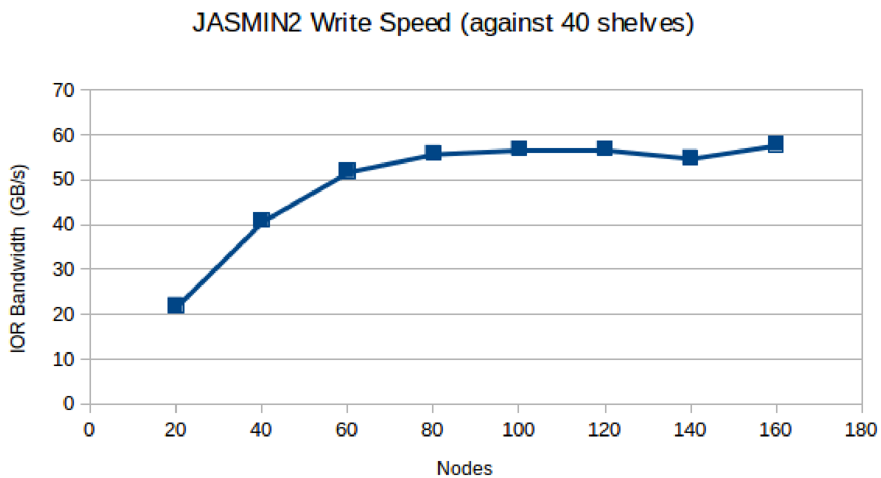
In a Panasas file system we can create "bladesets" (which can be thought of as "RAID domains", but note RAID is file based).

Trade-off (per bladeset) between performance, and reliability:

- Each bladeset can (today) sustain one disk failure (later this year, two with RAID6).
- The bigger the bladeset, the more likely we are to have failures.
- In our environment, we have settled on max o(12) shelves ~ 240 disks per bladeset. In JASMIN2 that's ~ 0.9PB

  (0.7 in JASMIN1, with 3 TB disks cf J2, 4 TB)
- Typically, we imagine a virtual community maxing out on a bladeset, so per community, we're offering o(20 GB/s).



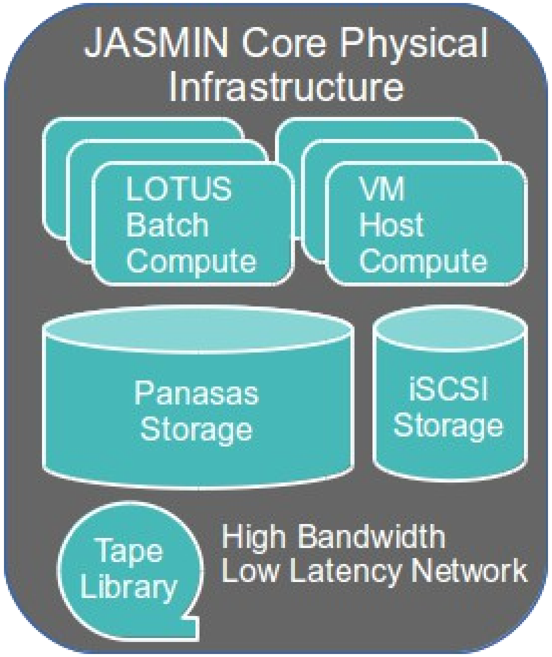


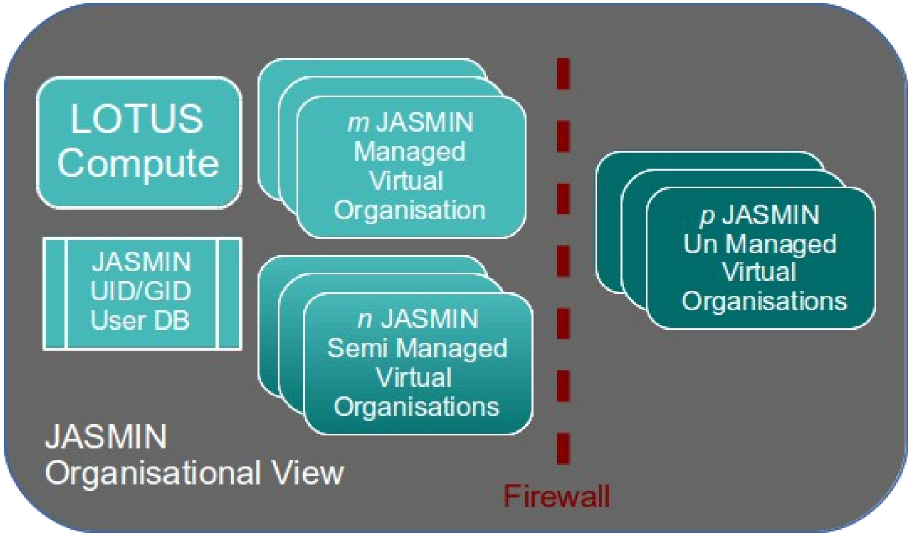
A subliminal message:

Did you notice that we could thrash a state of the art HPC parallel file system to within an inch of it's life with just o(100) nodes?!

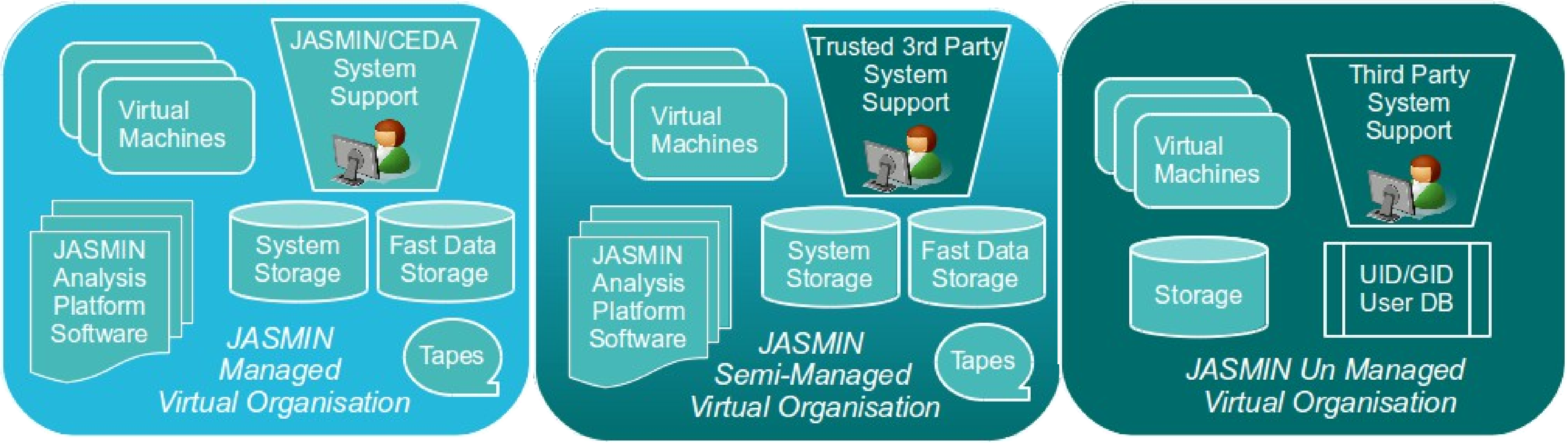
From a simulation point of view: our file systems are nowhere near keeping pace with our compute!

Physical and Organisational Views





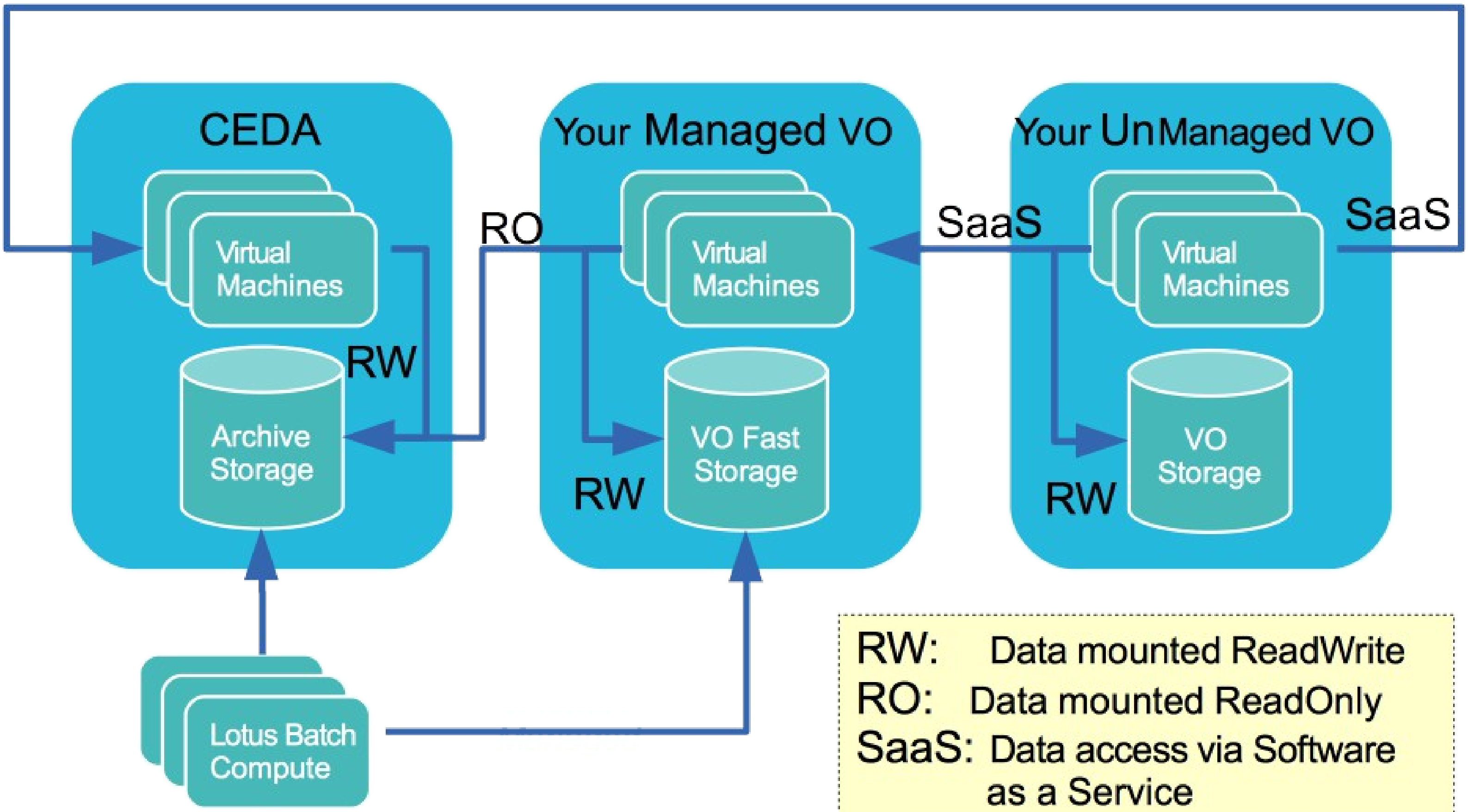
Managed, Semi- and Un-managed Organisations



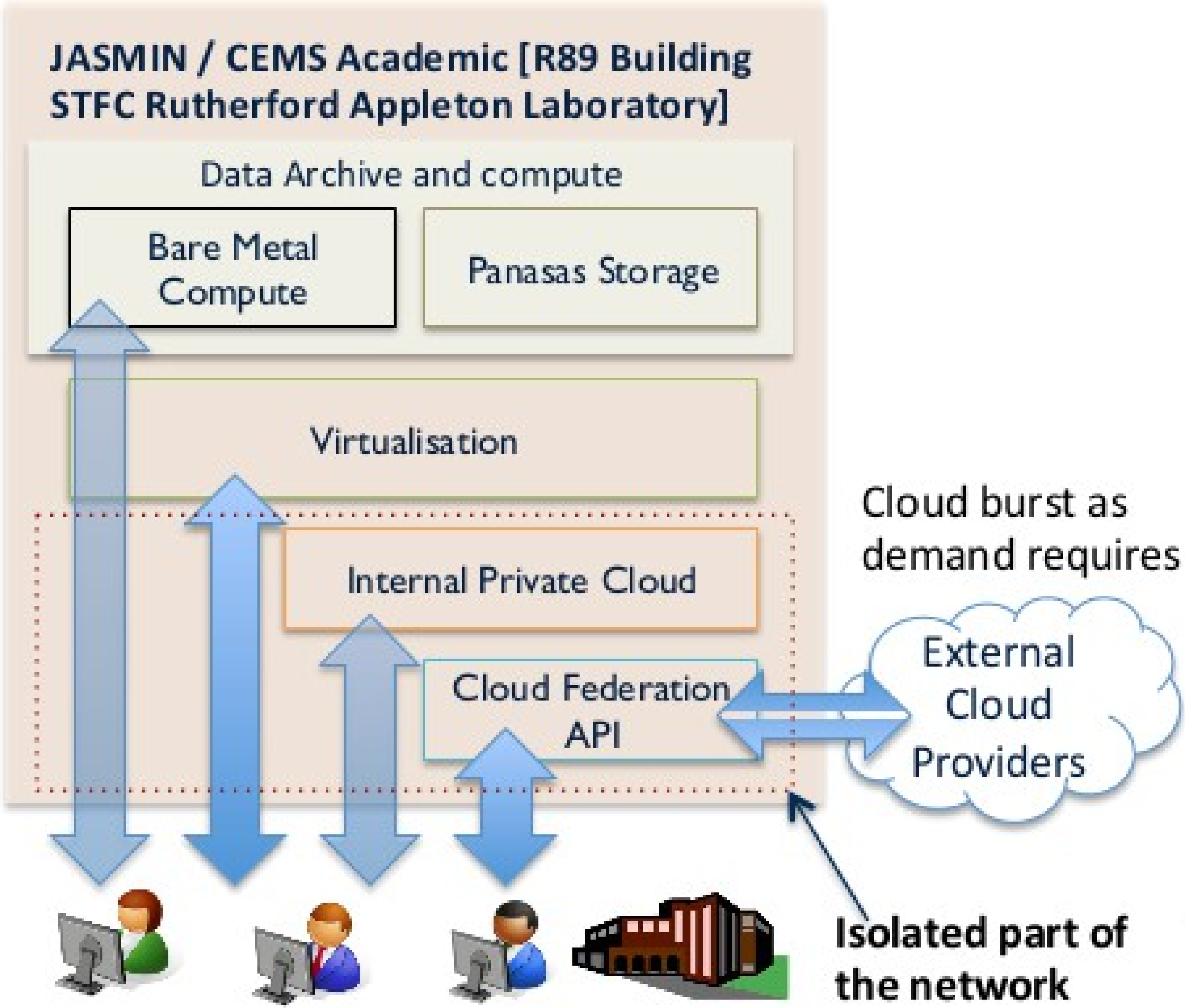
Platform as a Service (Paas) -----------------> Infrastructure as a Service (IaaS)

Science & Technology Facilities Council
Rutherford Appleton Laboratory

Secure and Constrained Access



RW: Data mounted ReadWrite
RO: Data mounted ReadOnly
SaaS: Data access via Software as a Service



Direct access to the data archive - Hosted processing and analysis environments

JASMIN Science

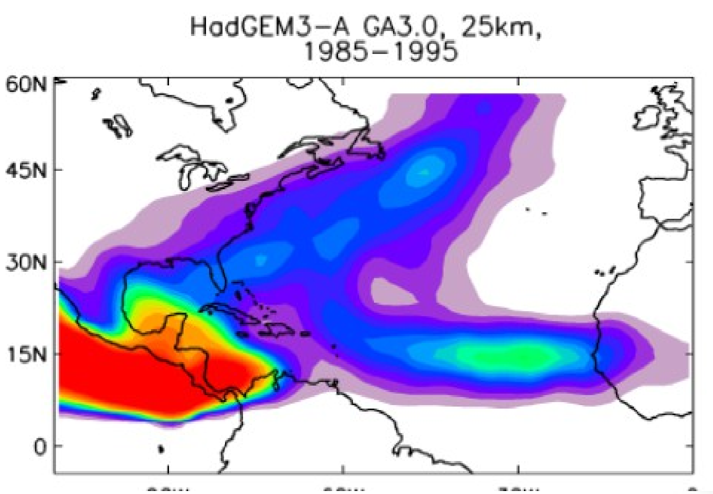
UPSCALE (NCAS + UKMO)

~ 350 TB currently stored

Calculation of "eddy or E-vectors" (Novak, Ambaum,Tailleux) … used at least 3 TB of storage and would have taken an estimated **3 months** on a dedicated, high-performance workstation

Breaking up the analysis task into ~2,500 chunks and submitting them to the LOTUS cluster finished in less than **24 hours** !

Similar speedups on cyclone tracking:



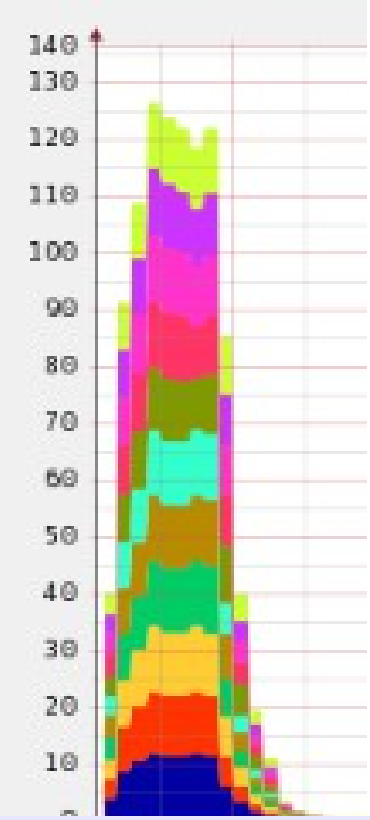
ATSR Reprocessing (RAL)

Last reprocessing took place in 2007-2008.

Used 10 dedicated servers to process data and place product in archive

Previous reprocessing complicated by lack of sufficient contiguous storage for output and on archive.

Reprocessing of 1 month of ATSR2 L1B data using original system took ~**3 days**: using JASMIN-HPC Lotus: **12 minutes**.



132 cores flat out (NOT I/O bound) for 12 minutes)!

Summary

Joint Analysis System providing a platform for:

(1) Curating the data held in the Centre for Environmental Data Archival (and it's constituent data centres).

(2) Facilitating the access to, and exploitation, of large environmental data sets (so far primarily for atmospheric science and earth observation, but from April this year, to the wider environmental science community.

This is done by delivering a very high performance flexible data handling environment with both cloud and a traditional batch computing interfaces.

We have a lot of happy users!

See: B. N. Lawrence, V. L. Bennett, J. Churchill, M. Juckes, P. Kershaw, S. Pascoe, S. Pepler, M. Pritchard, and A. Stephens, “Storing and manipulating environmental big data with JASMIN,” in 2013 IEEE International Conference on Big Data, 2013, pp. 68–75, doi:10.1109/BigData.2013.6691556

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Centre for Environmental Data Archival
SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL