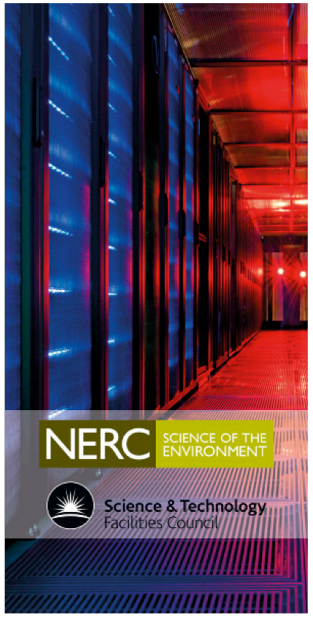
Science & Technology Facilities Council
Rutherford Appleton Laboratory



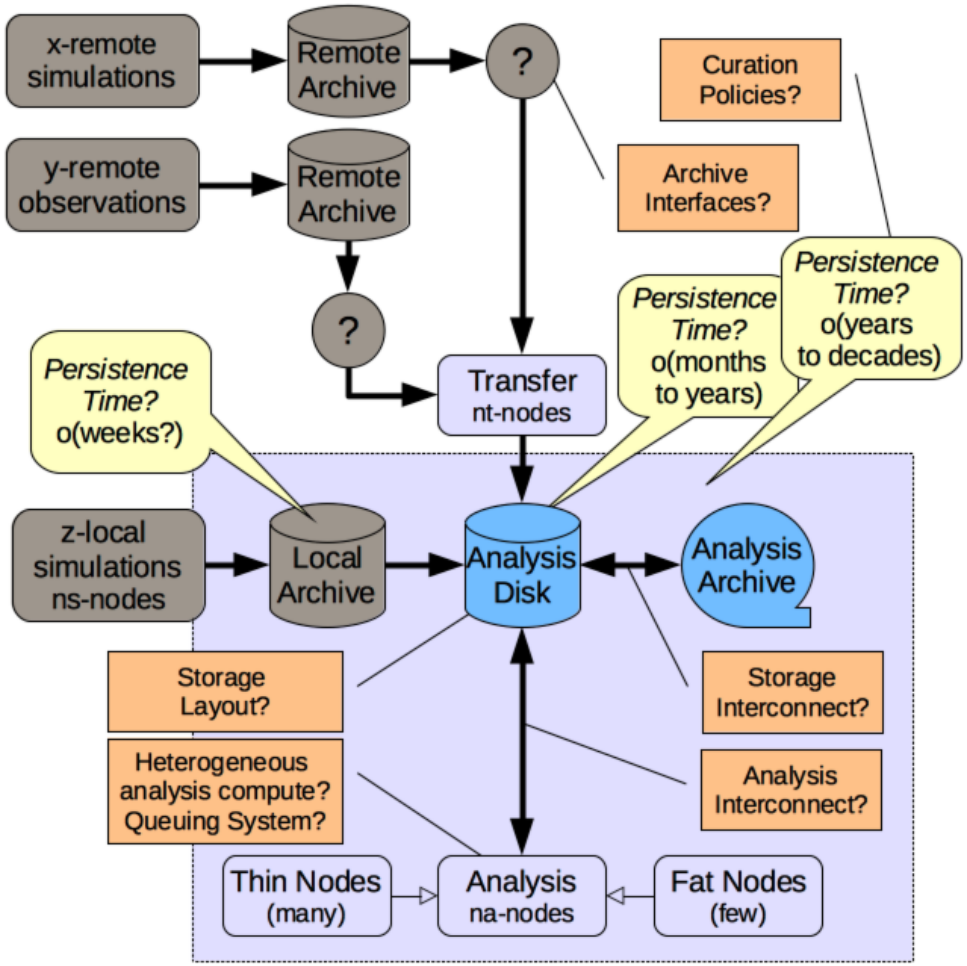
Beating the tyranny of scale with a private cloud configured for Big Data

B.N. Lawrence, V. Bennett, J. Churchill, M. Juckes, P.J. Kershaw, S. Pepler, M. Pritchard, A. Stephens.

STFC. NCAS, NCEO
University of Reading

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

Infrastructure Context



Multiple remote data sources: can't bring the compute to all data, SO: bring all data to one place, and bring the compute to that! (Avoid `n x n` data transfer!)

Need to worry about:

- Storage Layout
- Scheduling
- Curation Policies
- Interfaces
- Storage and analysis interconnect

Organisational Viewpoint

CEDA AS (once BADC)

CEDA EO (once NEODC)

CEDA Solar (once UKSSDC)

IPCC DDC

etc

NERC Managed Analysis Computing

(CEMS + Shared Systems for NCAS, MetO, NOC etc)

NERC Cloud Analysis Computing

(EOS Cloud, Env WB etc)

etc

CEDA Archive Services

Data Centres, Curation, DB systems

User management, External Helpdesk

CEDA Compute Services

Compute Cloud:

PaaS (JAP +Generic Science VMs + User Management), IaaS

External Helpdesk

JASMIN Compute and Storage

Lotus + Private Cloud + Tape Store + DMZ for data transfer

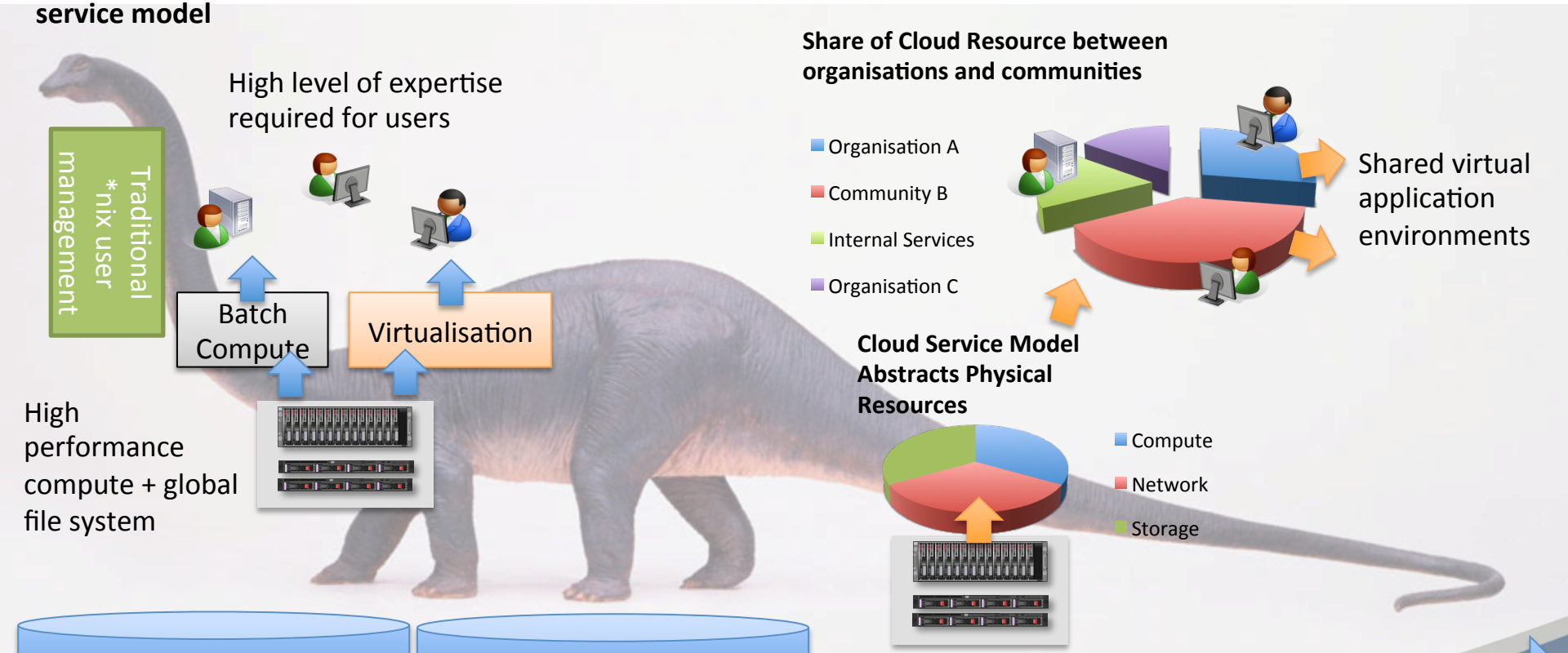
Internal Helpdesk

NERC SCIENCE OF THE ENVIRONMENT

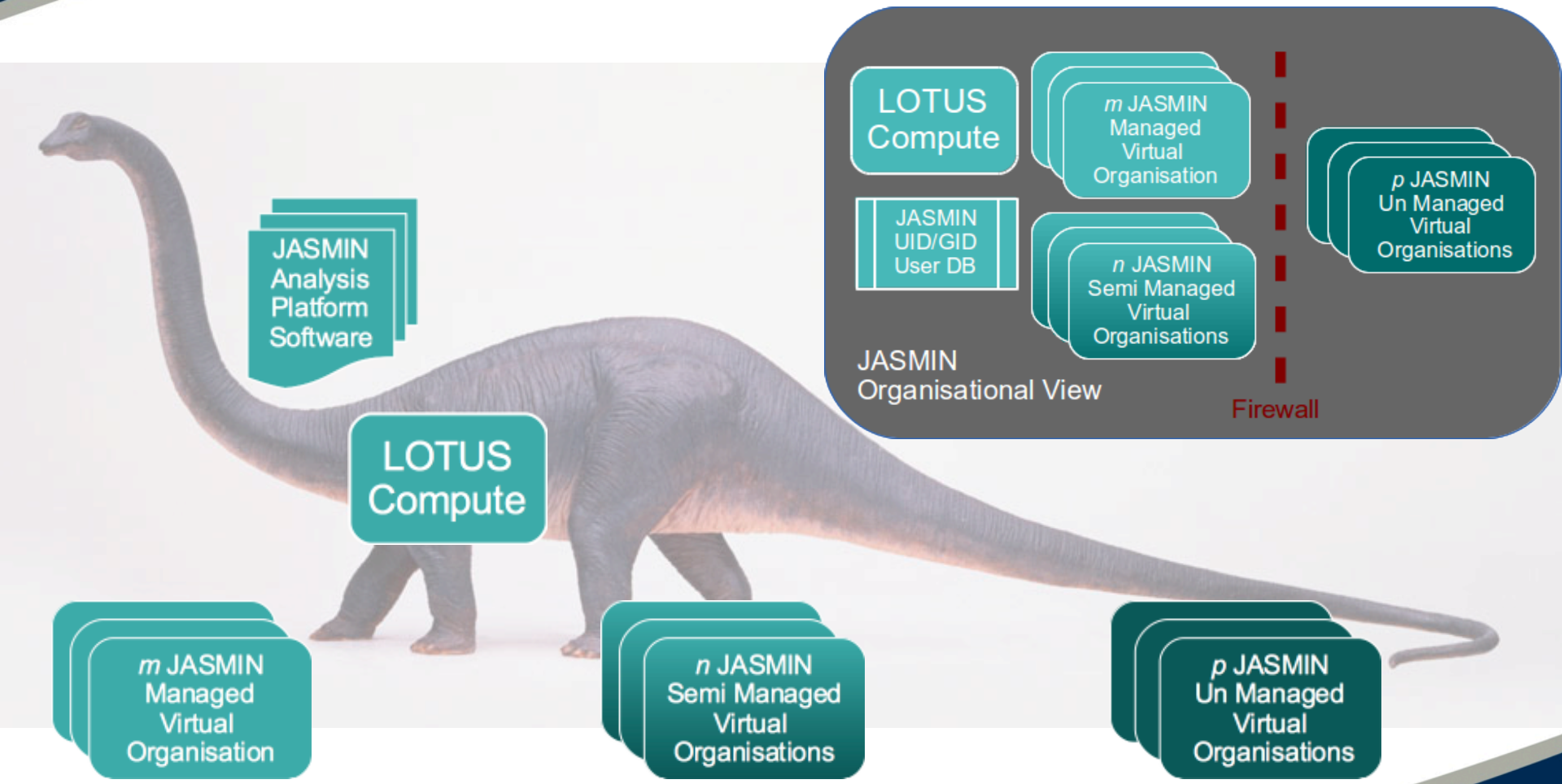
JASMIN and the 'long tail' of Science

Raw infrastructure power (data available all the time, next to the compute) but more constrained service model

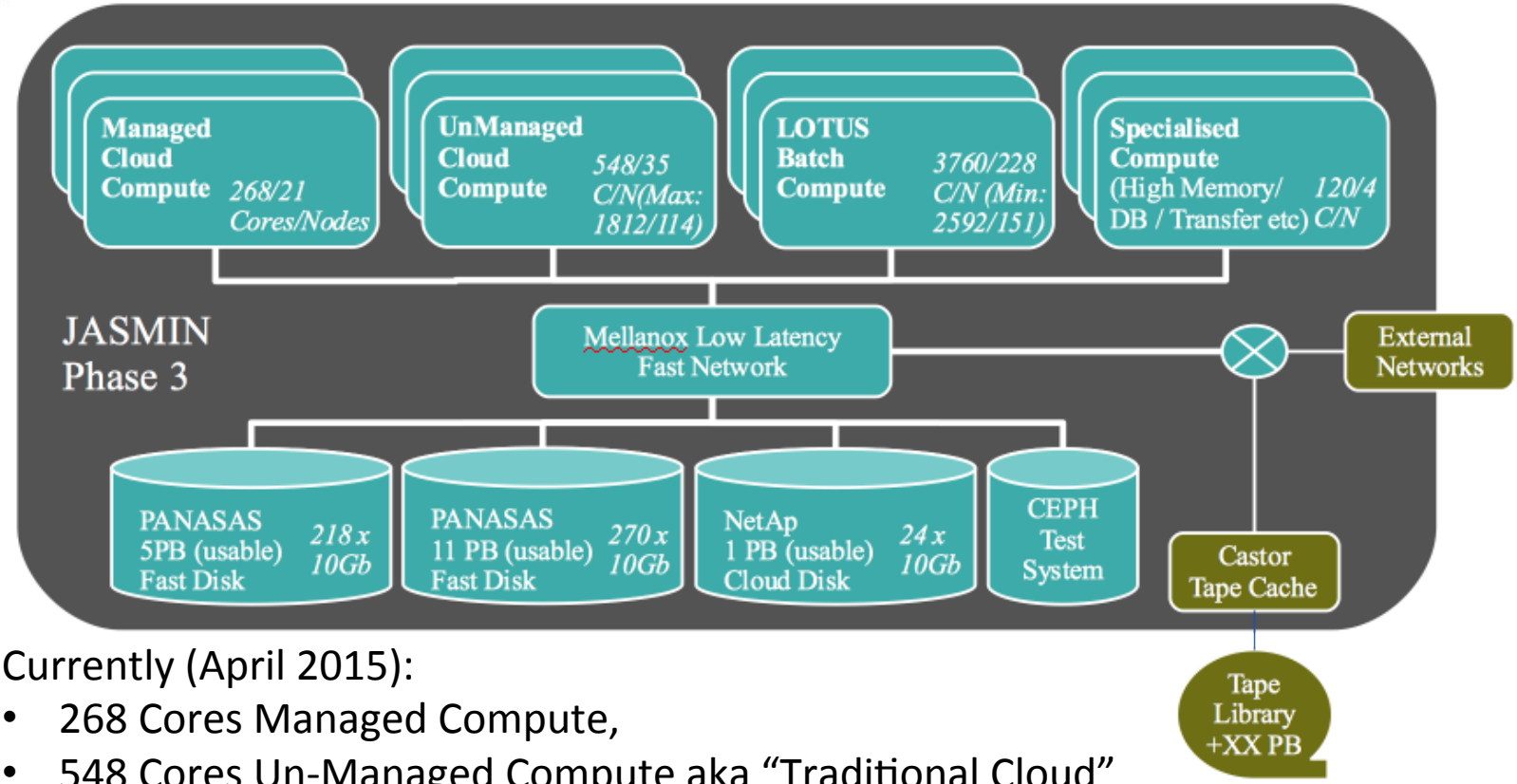
Rich and flexible service model allows establishment of domain specific collaborative environments



Supporting Communities



Engineering Viewpoint



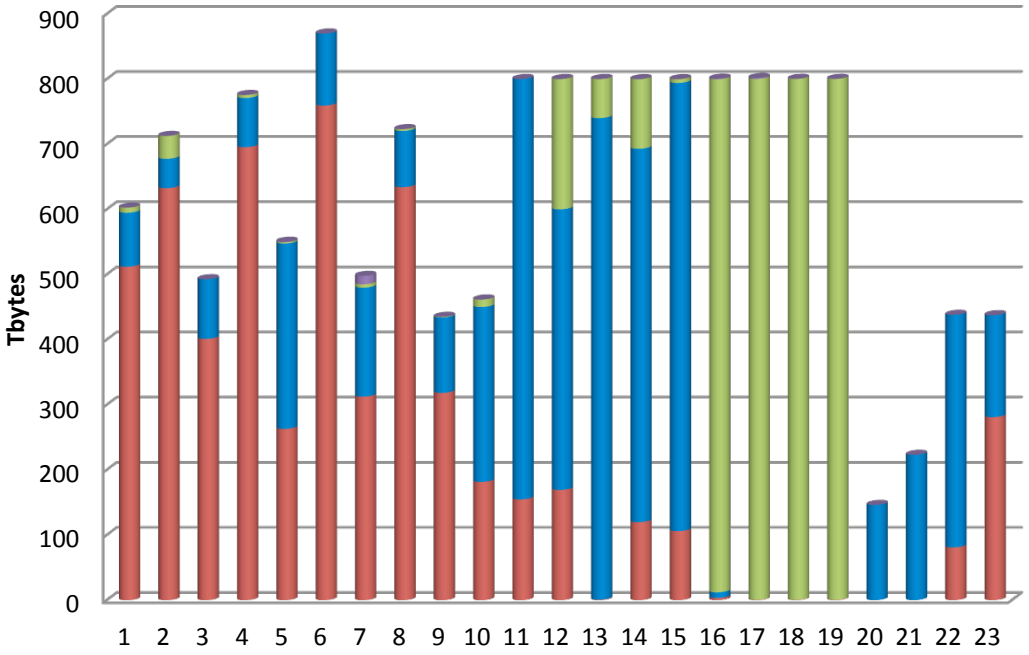
Currently (April 2015):

- 268 Cores Managed Compute,
- 548 Cores Un-Managed Compute aka “Traditional Cloud”
- 3760 Cores Batch Compute
- 120 Cores Specialised Compute
- 17 PB of disk! Note balance of network interfaces in storage and compute!
- Yet to benchmark full I/O, probably in excess of 3 Tb/s?

JASMIN Operations

- 500 JASMIN users
- > 80 projects
- 4.9 PB allocated as Group Workspace; 2.8 PB CEDA archives
- Over 1.5 million processing jobs

Academic CEMS Usage (Nov '14)	
GWS	25 ; 1900 TB
Managed VMs	54
Login users	81



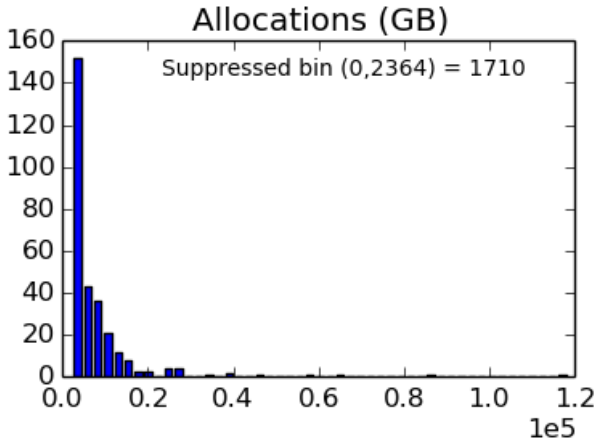
JASMIN usage October 2014.
Blue: allocated but not yet used.
Red: used.
Green: as yet unallocated

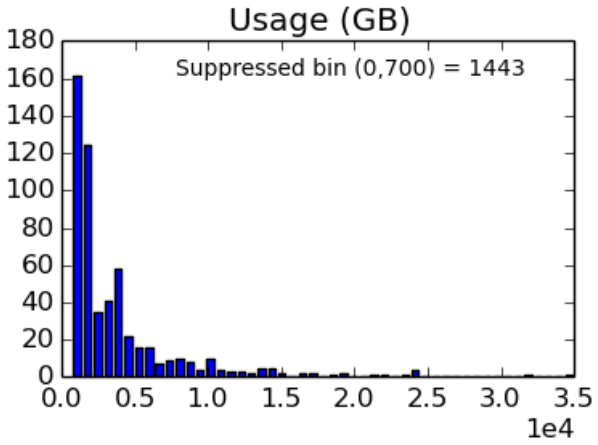
Data Management at Scale (1)

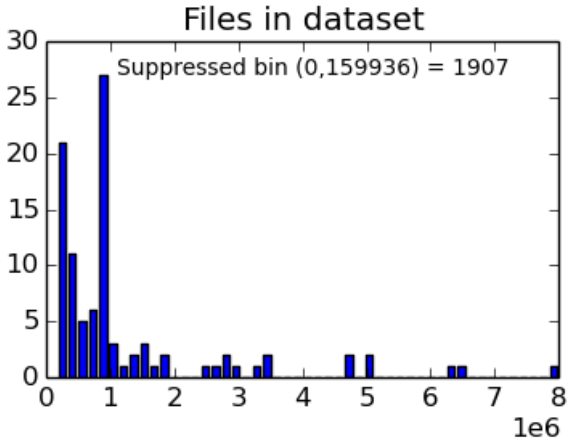
CEDA Archive Snapshot

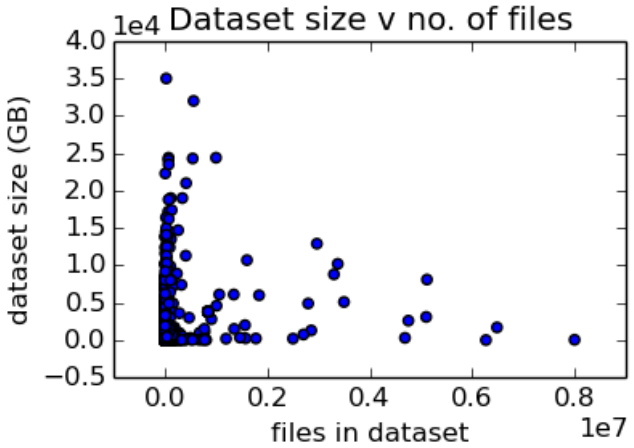
- 3.0 PB of allocated archive, 2.3 PB used in 2,176 **filesets** totalling 152M files.
- 1 copy on disk, at least one on tape near line, and one offsite
- Long tail in both dataset size and number of files.
- Volume and number of files not correlated, although the high volume datasets tend not to have the most files.

How do we test for data integrity?









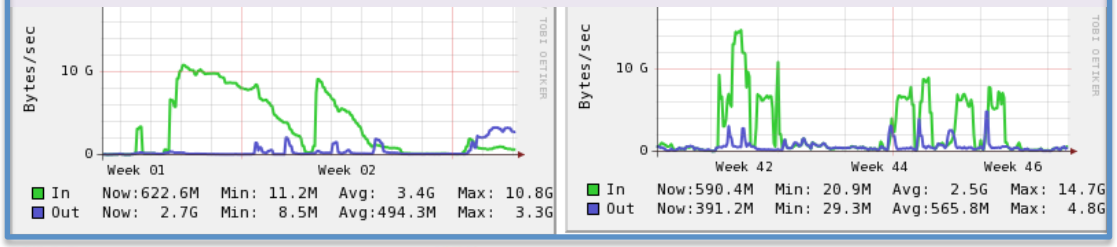
Snapshot data 01/12/2014 via Sam Pepler.

Data Management at Scale (2)

```
for i in range(number_to_do):
  fileset = CEDADB.next_audit()
  # EITHER METHOD A
  # checkm_file = fileset.create_checkm()
  # OR METHOD B
  filelists=fileset.make_jobs()
  for fs in filelists:
    results[fs]=fs.create_checkm()
  checkm_file = combine(results)
  # EITHER WAY:
  CEDADB.store_anal_notify(checkm_file)
  # Yes this is "poor man's Map Reduce"
```

- Doing audits in batches as often useful to only do some at a time.
- CEDADB is a restful service to work out which audit to do next and store result.
- Method A: 1 LOTUS job per fileset. Some filesets are bigger than others so small ones finish fast and larger ones drag on for days.
- Method B makes multiple LOTUS jobs each with no more than a certain volume and no more than a certain number of files.

JASMIN network Ganglia plots: Green net input are proxy for read
Audit jobs done quicker & more efficiently using method 2 (right panel)



It turns out that not only is a lot of data curation embarrassingly parallel (and amenable to map-reduce) but so is a lot of science!

Example uses of CEMS-JASMIN for global land surface products

Jan-Peter Muller, Said Kharbouche (NCEO, UCL)

Objective 1: Re-project BRDF files from SIN-coordinates to lat/lon using an Energy Conservation method

- **Challenge:** the projected SIN-Tiles into lat/lon results for non-rectangular shapes, with different SIN tiles
- **Solution:** SIN and Lat,Lon Cells are represented by geometry polygons rather than simple points and then the process is based on ratios of common area rather than on simple distance
- **Challenge**: huge number of polygons to be spatiality indexed and processed. **This process requires massive RAM and usually takes a very long time**
- **Solution**: Use Cloud-computing system on CEMS-JASMIN (~100 times faster than 224-core in house linux cluster!)

Example uses of CEMS-JASMIN for global land surface products

Jan-Peter Muller, Said Kharbouche (NCEO, UCL)

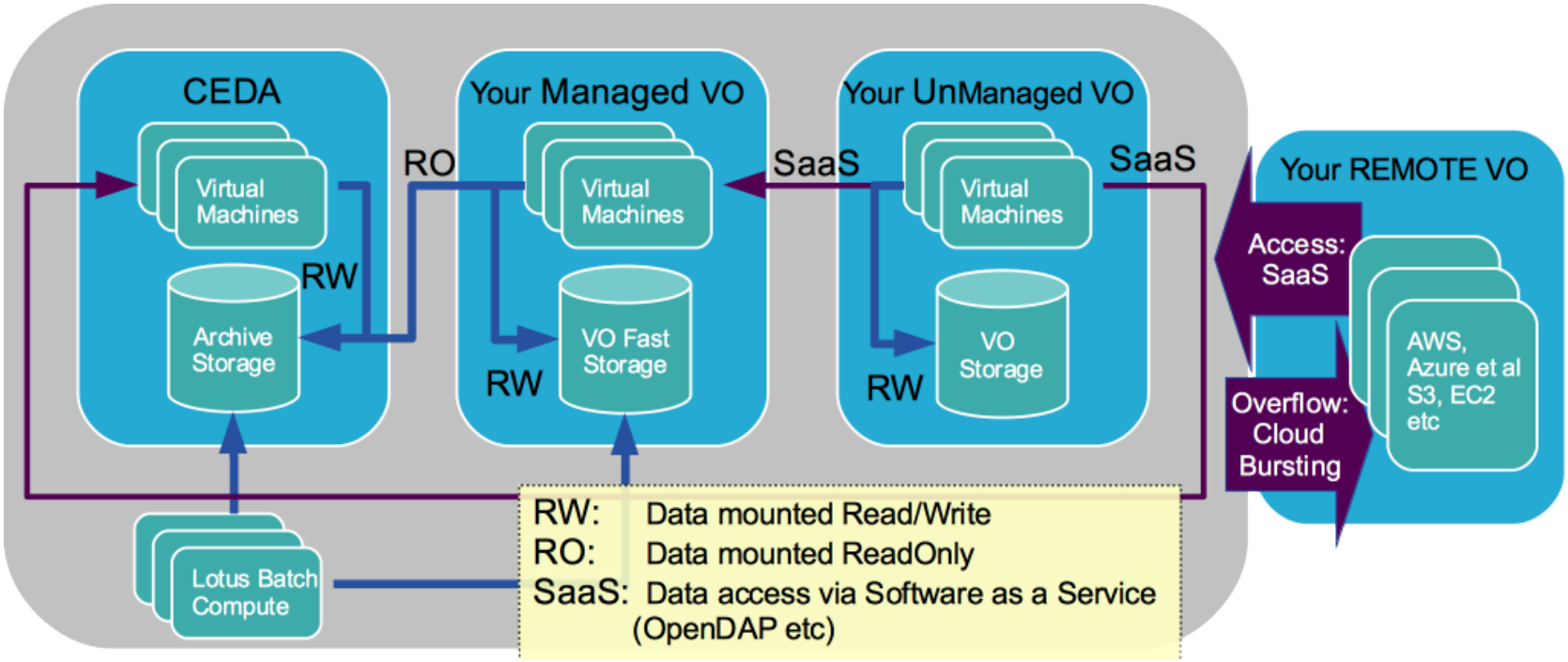
Objective 2: Create specific albedo products for computation of 8-daily LAI/fAPAR between 2002 and 2011 at 3 different resolutions: 1km, 5km and 25km

- **Challenge:** Upscale big data BRDF (50TB) from 1km to 5km and 25km using energy conservation method, and then create separate Albedo-Snow_only and Albedo-Snow_Free products: **This process is extremely time consuming!**
- **Solution**: Cloud-computing system in CEMS-JASMIN (~100 times faster than 224-core in house linux cluster)

Also use Science DMZ for data transfers from NASA

- Achieved rates up to 28 TB/day

Issues



Curated environment is one virtual organisation alongside o(100) other organisations.
Key issues include:

(1) How to provide high performance data access and analytics in the managed and semi-managed environment for multiple users, multiple workflows, all intersecting in some of the data.
(2) How to support high performance data transfer and job migration between the different tiers of infrastructure,
(3) All in a context of extreme data growth.

Workflow and Scheduling Issues (1)

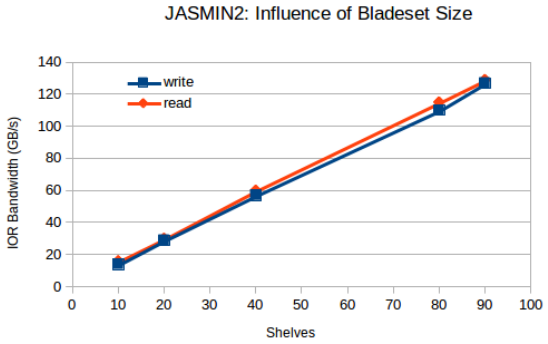
The seven deadly sins of ~~cloud~~ computing ~~research~~ (Schwarzkopf, Murray and Hand, Hotcloud, 2012)

Pick five, all in play:

- **Unnecessary distributed parallelism:** We need to support (nicely) high memory and other nodes inside our environment.
- **Assuming performance homogeneity**. This is a real problem for us in a mixed VM/batch environment ... Help.
- **Forcing the abstraction** (Map-Reduce, HADOOP or bust) We avoid this by having a parallel file system, but how do we know we are getting value?.
- **Unrepresentative workloads.** We really don't know how to optimise our jobs (yes, we can give people exclusive access to nodes, but it's harder to give them exclusive I/O bandwidth).
- **Assuming perfect elasticity.** We haven't worked out how to schedule to use our resources, or how to cloud burst properly.

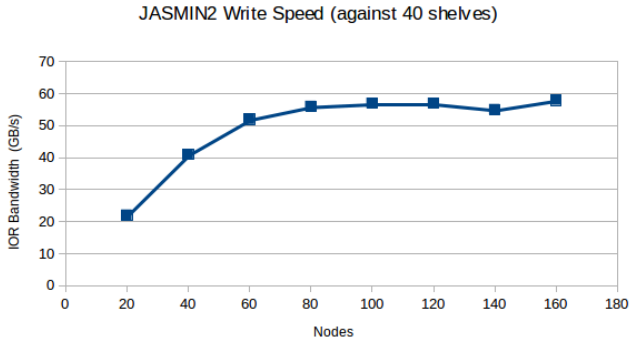
We need work on understanding all these things

Pick one issue: I/O and Storage



(IOR) **Do we understand the performance at the user/app level?**

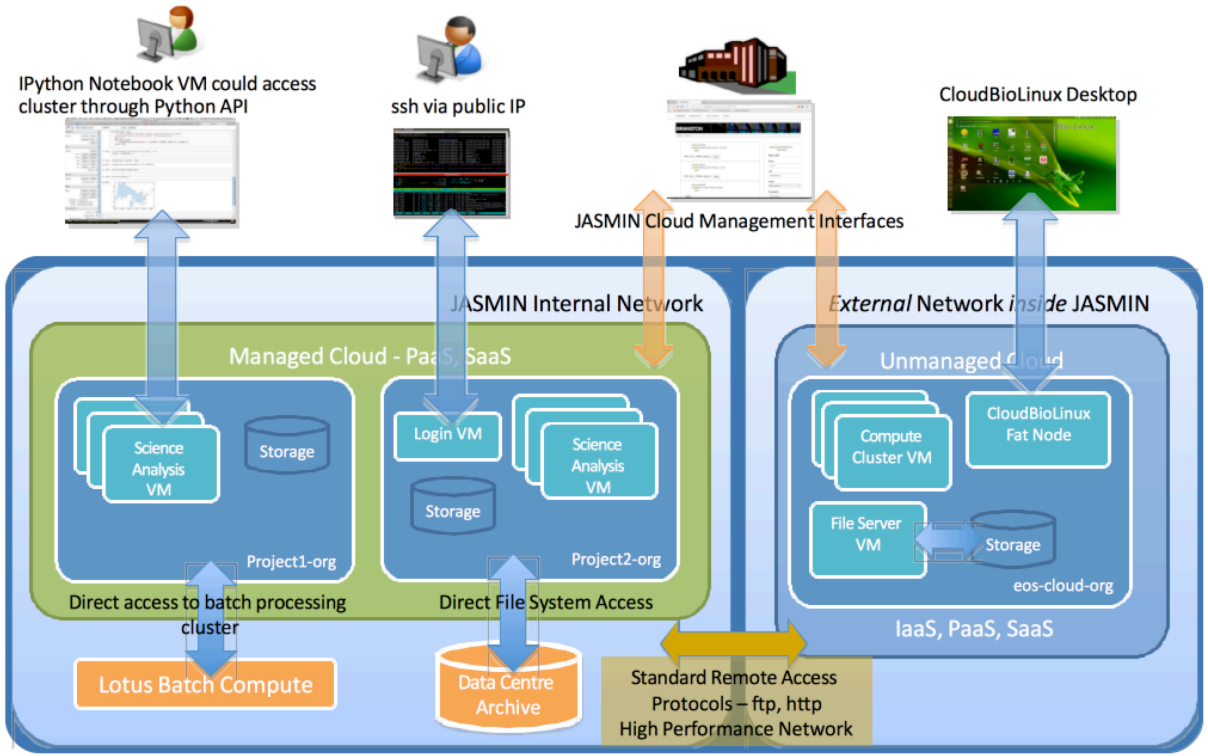
We can break our file system up into pools ("blade sets") in Panasas. Give communities access to resources on one blade set. Now their I/O does not interfere with VOs using other blade sets.



This isn't very flexible! We can still nail a PB bladeset with 80 nodes! **How do we get more and flexible I/O?**

When we run out of physical space for disk, how are we going to efficiently use tape in our workflows?

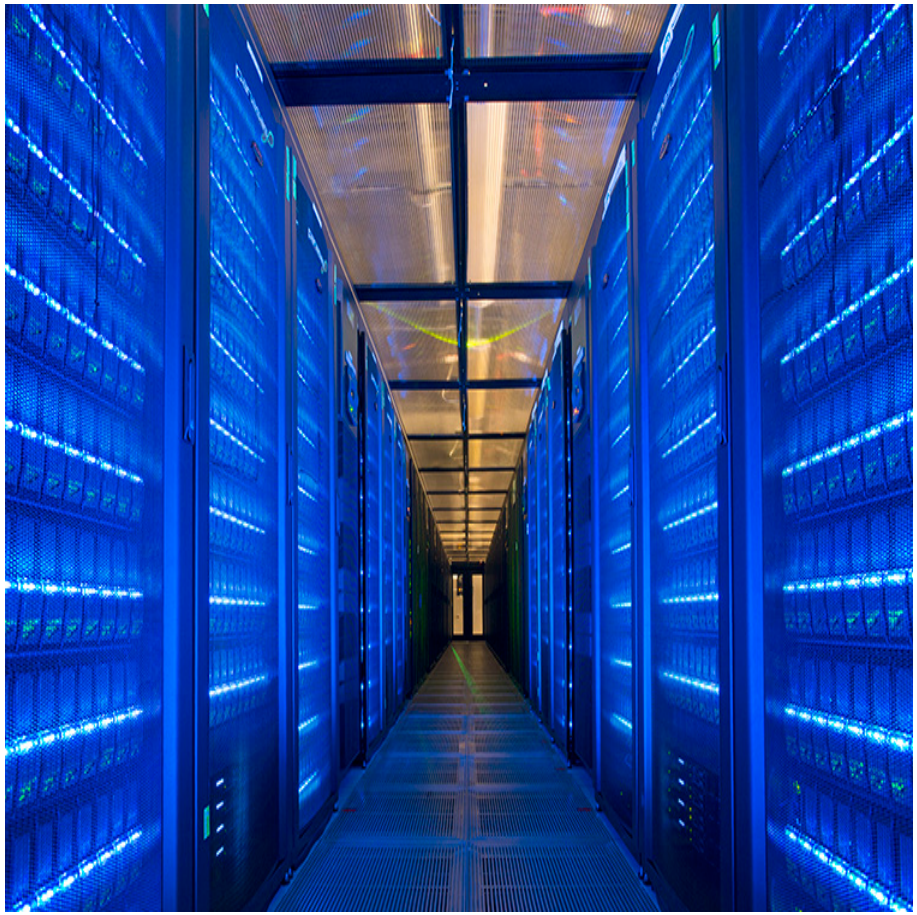
Workflow and Scheduling Issues (2)



This currently works because we have spare capacity, and relatively few users in the un-managed cloud and the ipython notebook environment.

We don't know how to do the scheduling here, the hypervisior/VM paradigm is banging up against the batch system job. Interactive is banging up against resource. The sixth deadly sin: there is not perfect elasticity. We are offering cloud bursting (to Amazon and Azure, we hope), but then there needs to be more work on data pipelines.

Further info



JASMIN

- http://www.jasmin.ac.uk

Centre for Environmental Data Archival

- http://www.ceda.ac.uk

JASMIN Context:

Lawrence, B.N. , 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.** *Proceedings of IEEE Big Data 2013, p68-75,* doi:10.1109/BigData.2013.6691556