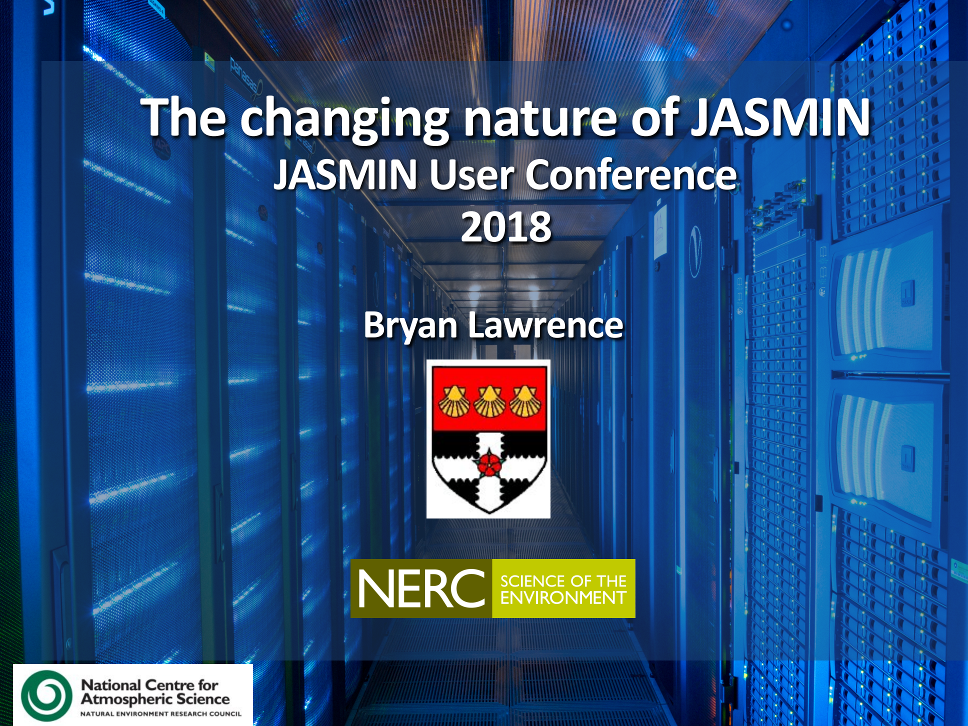


The changing nature of JASMIN

JASMIN User Conference
2018

Bryan Lawrence

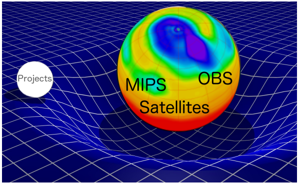
NERC SCIENCE OF THE ENVIRONMENT

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Outline

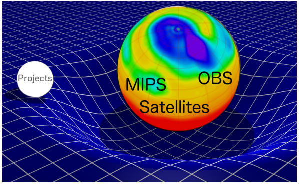
1. Reminder of the functional components of JASMIN
2. Some analysis of usage (context for upgrades):
   - Compute
   - Data Movement
   - Storage
3. Headline description of phase 4 upgrade.
4. Plans for phase 5 and beyond.

JASMIN — The Data Commons

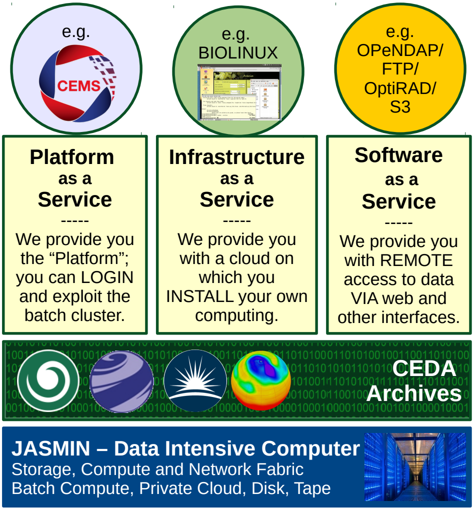


- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the “archive”).
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.

JASMIN — The Data Commons



- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the “archive”).
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.



e.g. CEMS

e.g. BIOLINUX

e.g. OPeNDAP/ FTP/ OptiRAD/ S3

Platform as a Service

We provide you the “Platform”; you can LOGIN and exploit the batch cluster.

Infrastructure as a Service

We provide you with a cloud on which you INSTALL your own computing.

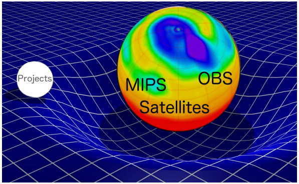
Software as a Service

We provide you with REMOTE access to data VIA web and other interfaces.

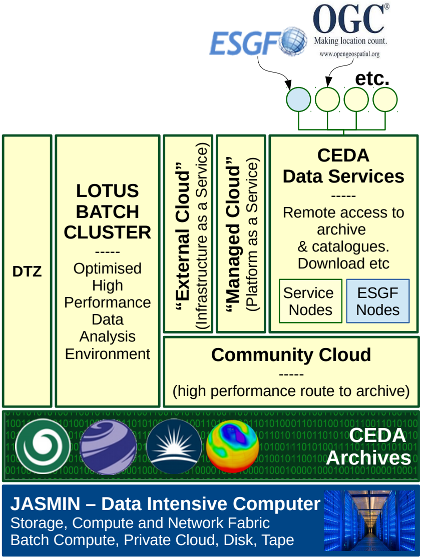
CEDA Archives

JASMIN – Data Intensive Computer
Storage, Compute and Network Fabric
Batch Compute, Private Cloud, Disk, Tape

JASMIN — The Data Commons



- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the “archive”).
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational environment.



ESGF

OGC® Making location count. www.opengeospatial.org

etc.

DTZ

LOTUS BATCH CLUSTER

Optimised High Performance Data Analysis Environment

“External Cloud”
(Infrastructure as a Service)

“Managed Cloud”
(Platform as a Service)

CEDA Data Services

Remote access to archive & catalogues. Download etc

Service Nodes

ESGF Nodes

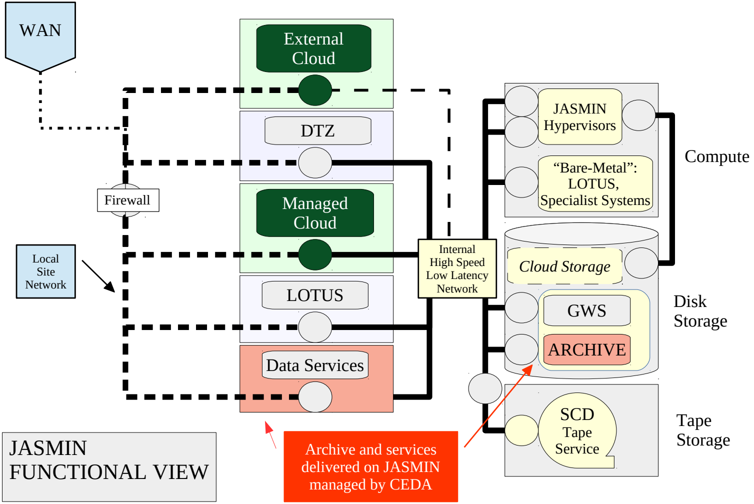
Community Cloud

(high performance route to archive)

CEDA Archives

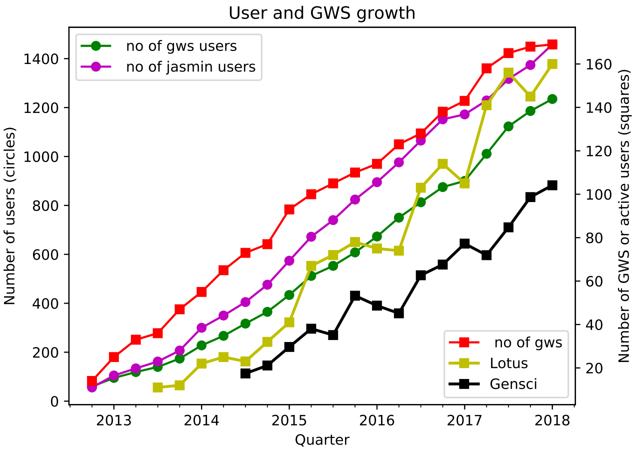
JASMIN – Data Intensive Computer
Storage, Compute and Network Fabric
Batch Compute, Private Cloud, Disk, Tape

A Functional View of JASMIN (pre-phase4)

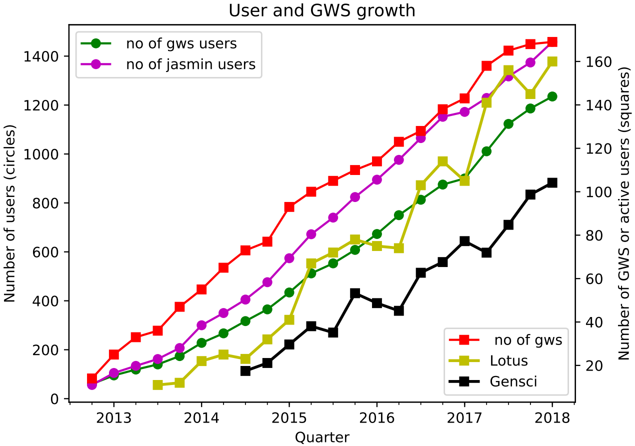


Archive and services delivered on JASMIN managed by CEDA

Users and GWS

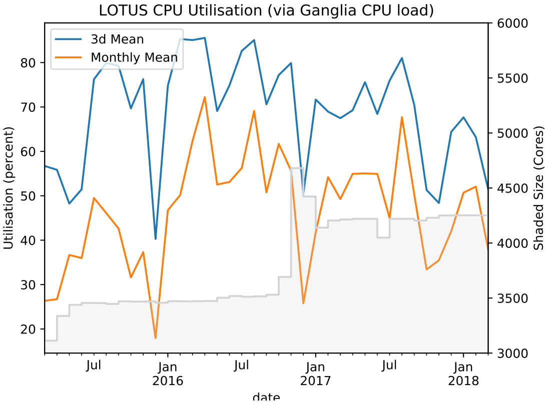


Users and GWS



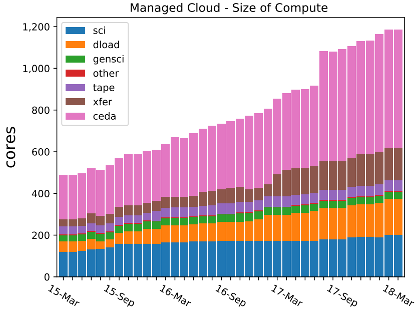
All that growth, while adding lots of kit …

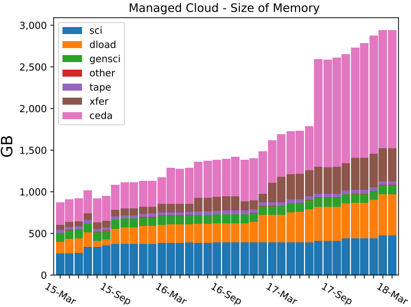
Lotus



Note the changing size of the cluster, and the influence of weekends on utilisation: LOTUS is not a traditional HPC platform.

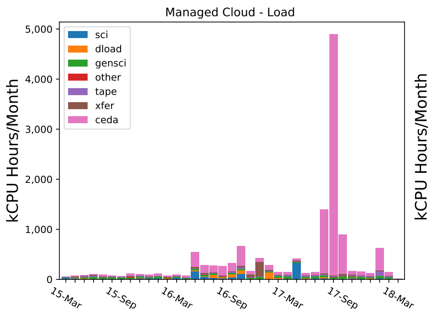
Managed Cloud

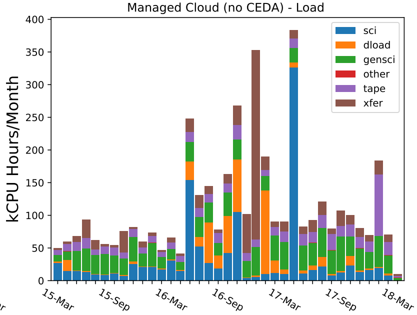

Managed Cloud - Size of Compute


Managed Cloud - Size of Memory

Note that the **gensci** machines are a very small proportion of the managed cloud, even though they are problably the most visible to the most users!

Managed Cloud

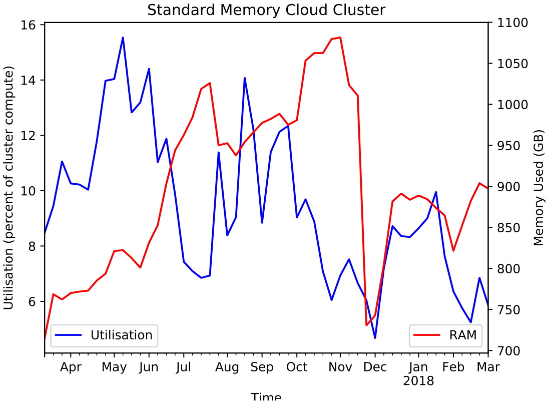




Compute load in the managed cloud:

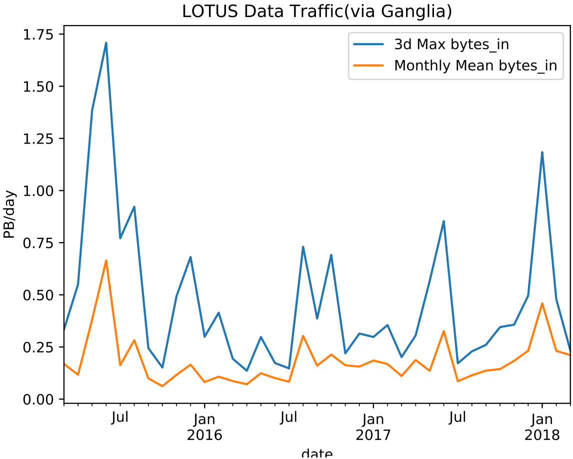
- Data management can be computationally demanding.
- Advantages of virtualisation: persistence versus demand.

External Cloud

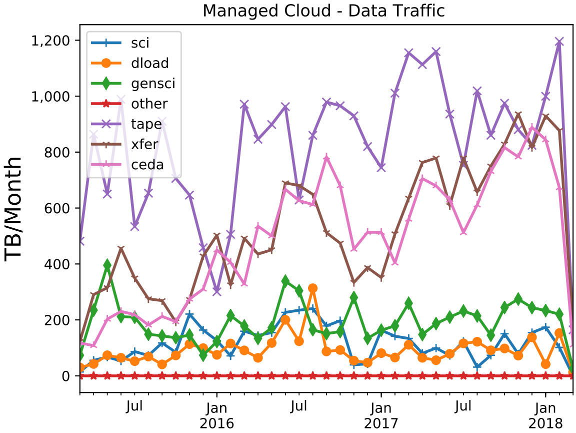


Hard to decide what good utilisation looks like: on-demand, versus persistence? How can we provide the baseload and exploit public cloud for the variability?

Lotus

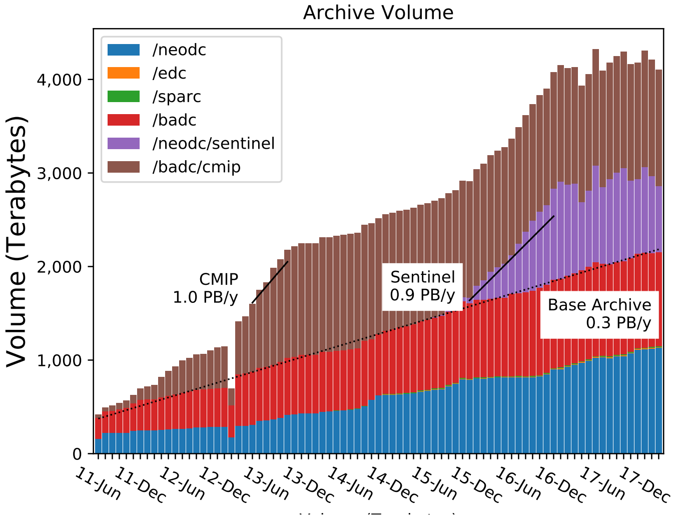

LOTUS Data Traffic(via Ganglia)
3d Max bytes_in
Monthly Mean bytes_in
PB/day
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Jul
Jan 2016
Jul
Jan 2017
Jul
Jan 2018
date

Managed Cloud

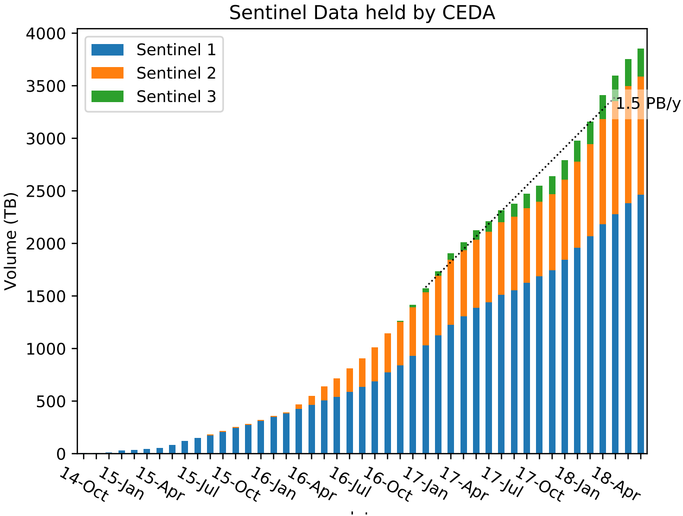


Even the smaller looking lines really add up: in the year to April 2018, 800 TB were downloaded from CEDA archives, by 18,900 users in 13 million files!

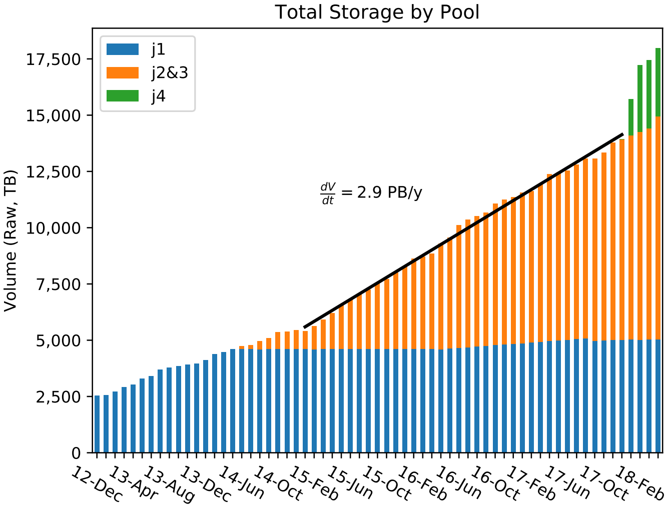
Archive Growth



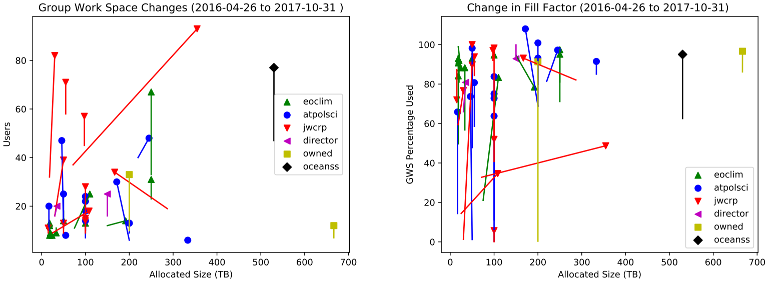
Sentinel Growth



Total Storage Growth

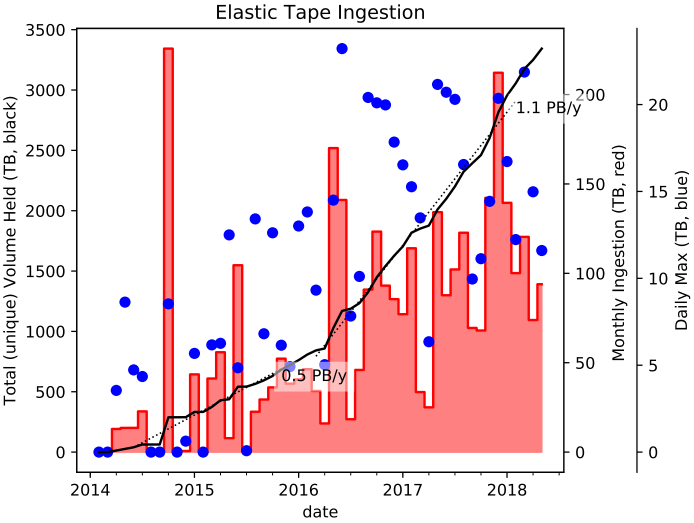


Impact on GWS



The changing nature of selected group work spaces over eighteen months: The left hand panel shows the change in size and number of registered users for all GWS with over 5 users and 10 TB which existed in April 2016. The right hand panel shows the change in fill fraction of these GWS over the period. In both cases, the solid markers denote the end of the period, and the line segments begin at the values at the beginning of the period.

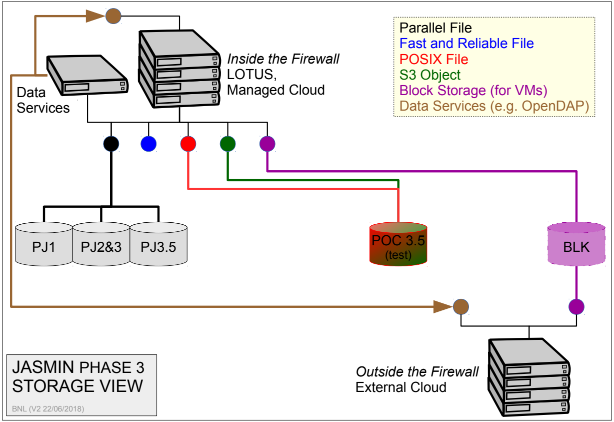
Elastic Tape Growth



Phase 4 Headlines

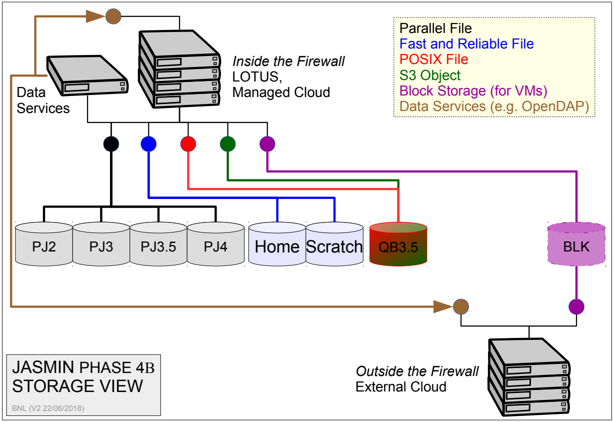
- Aiming to grow disk storage as much as feasible, recognising that the *next* upgrade would be focused on tape.
- Aiming to grow the ability for our cloud to be *suitably* elastic, with future elasticity (and perhaps more) coming from the public cloud.
  - Adding 200 new compute hypervisors
- Introducing new types of storage based on underlying “software defined storage” paradigm:
  - Scale out File system (from Quobyte) - 30 PB
  - Object Store (from Caringo) - 5 PB
  - Retiring, addding, and resizing our parallel file system (from Panasas) - 8.2 PB
  - New home and high performance (for small files) - 0.5 PB.
- A new Proof of Concept Hybrid Test Cluster (10 servers)
- Upgrades to service systems (for the Data Transfer)
- …and lots of underlying network and environment changes. A big deal to deliver, but hopefully invisible to users.

Logical View of DISK Storage at Phase 3



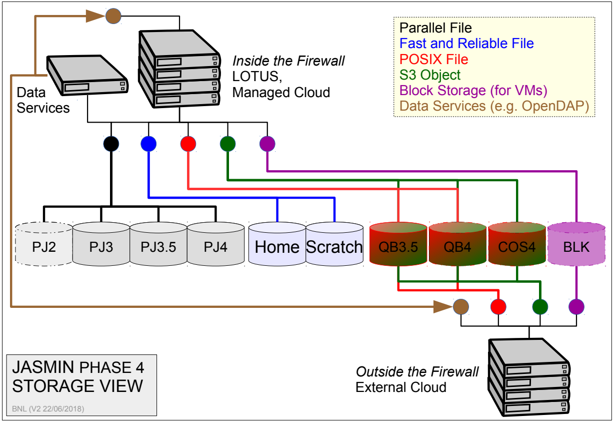
- Panasas Parallel Disk visible inside the firewall, with local block storage in the external cloud.
- Proof of concept new storage being tested.

Logical View of Disk Storage at Phase4 as it was last week



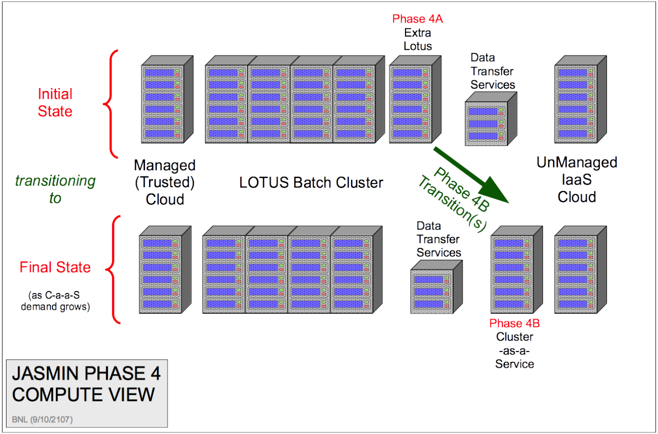
- PJ1 Panasas has gone, addition of PJ4.
- New home and scratch! POC has become Q3.5 and is in use!

Logical View of Disk Storage at Phase4



- Addition of Q4 and OS4. PJ2 to be retired this year.
- Eventual support for access to Q3.5, Q4 and OS4 outside firewall

Logical View of Phase4 Compute Transitions



- Initial deployment of new compute in Lotus (or possibly an additional cluster).
- Eventual migration into the cloud as cloud demand grows.

The next few years

- Historical 2.9 PB/year storage growth despite fluctuations in archive growth. Growth has been somewhat self-limited. Limitation has been achieved by:
  - Considerable use of the RDF in Edinburgh.
  - Considerable use of tape by some groups (Elastic Tape for GWS, Storage-D by the Archive).
- Looking forward we can see three significant perturbations:
  - New updated reanalysis products
  - Current sentinel growth 1.5 PB/y ... plus a bit more for S3B.
  - CMIP6 …2 PB in 18 months from the UK? …10 PB in 24 months?
  - New HPC platform (from beginning of 2020).
- Last year we noted that quite a lot of the disk data was cold (not touched in the last three months).
- Real ongoing question: what data should be where?

Phase 4 Storage Reality

Disk Storage growth!

- 24 PB allocated at beginning of 12018, so realistically we are adding 20 PB of storage.
- If we only grew at 2.9 PB/year, that'd be great …but we expect that growth rate to double …
- Realistically expect this storage to fill up within 4 years.
- What then? Just keep buying disk? (Lots of cold data rememember)?

Phase 4 Storage Reality

Disk Storage growth!

- 24 PB allocated at beginning of 12018, so realistically we are adding 20 PB of storage.
- If we only grew at 2.9 PB/year, that'd be great …but we expect that growth rate to double …
- Realistically expect this storage to fill up within 4 years.
- What then? Just keep buying disk? (Lots of cold data rememember)?

Financial Reality

- Disk price is not falling as fast as our consumption is increasing!
- The UK not as wealthy as it was, and …
- Need to make more use of tape in our workflow! (Tape still cheaper, and most assessments suggest it will remain so!)

Tape and Phase5

Requirements

- Two classes of users: the archive, and end-users.
- Archive needs backup copies.
- Archive needs to grow beyond disk but allow users to effectively cache the archive data on archive disk (not in GWS)!
- Users can overflow GWS: Need sufficient information about what is on tape for users to be able to interrogate their tape holdings with low latency.

Constraints

- Need new hardware. Existing vendor is going out of the tape business.
- JASMIN requirements need to dovetail with STFC requirements.
- We have (primarily) capital funding: not good for software development.
- However, we can build on, and exploit EC Funding.
- This will be a multi-year activity.

…and more

Other phase 5 objectives:

- Finish the Phase 4 storage and compute deployment.
- Continue to improve the cloud environment, including rolling out
  - Jupyter Notebook Service
  - Cluster as a Service (SLURM, DASK, SPARK)
  - Improving access to archive data from within the cloud
  - …and many underlying systems.
- Improve our support for parallel data analysis software