The UK JASMIN Environmental Commons

Bryan Lawrence

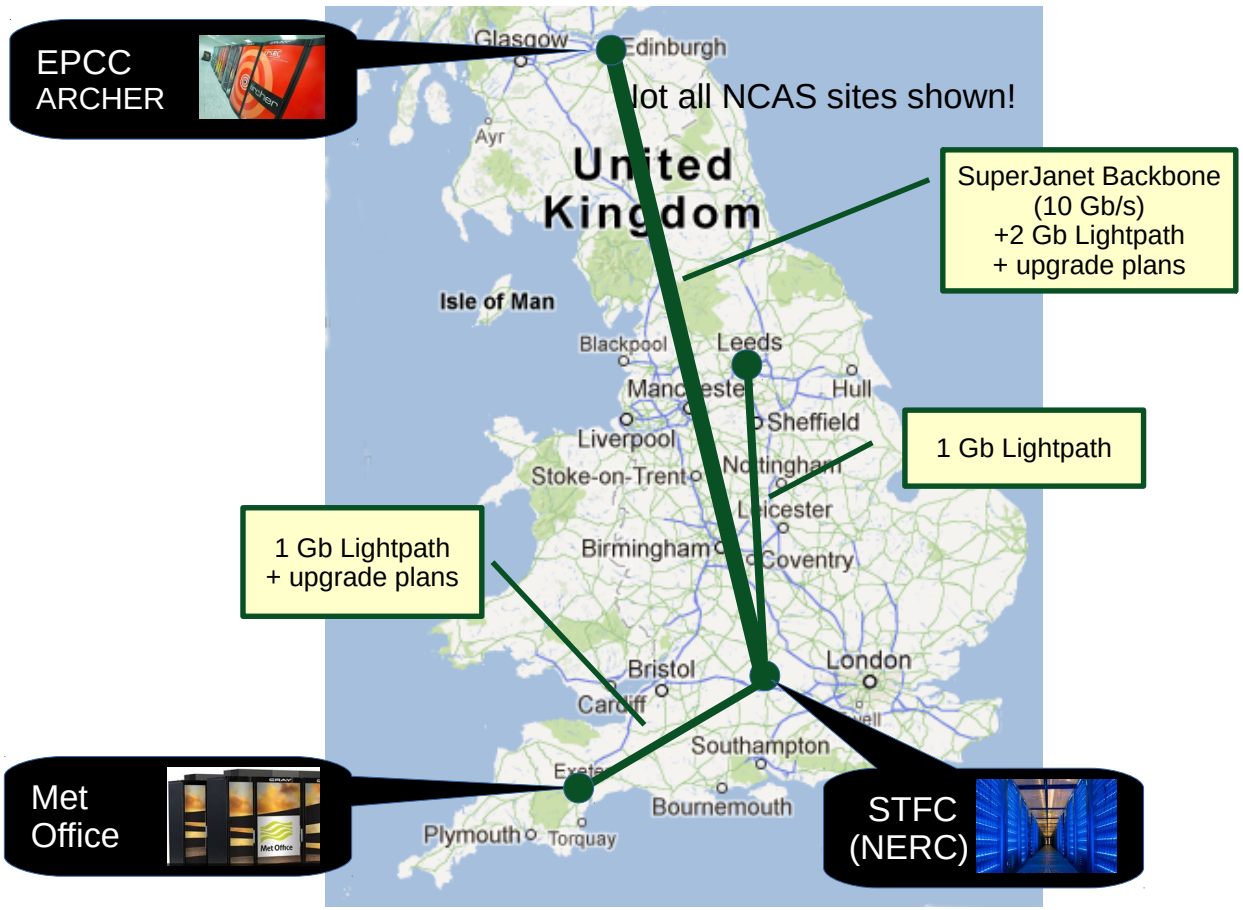
NERC SCIENCE OF THE ENVIRONMENT

Science & Technology Facilities Council

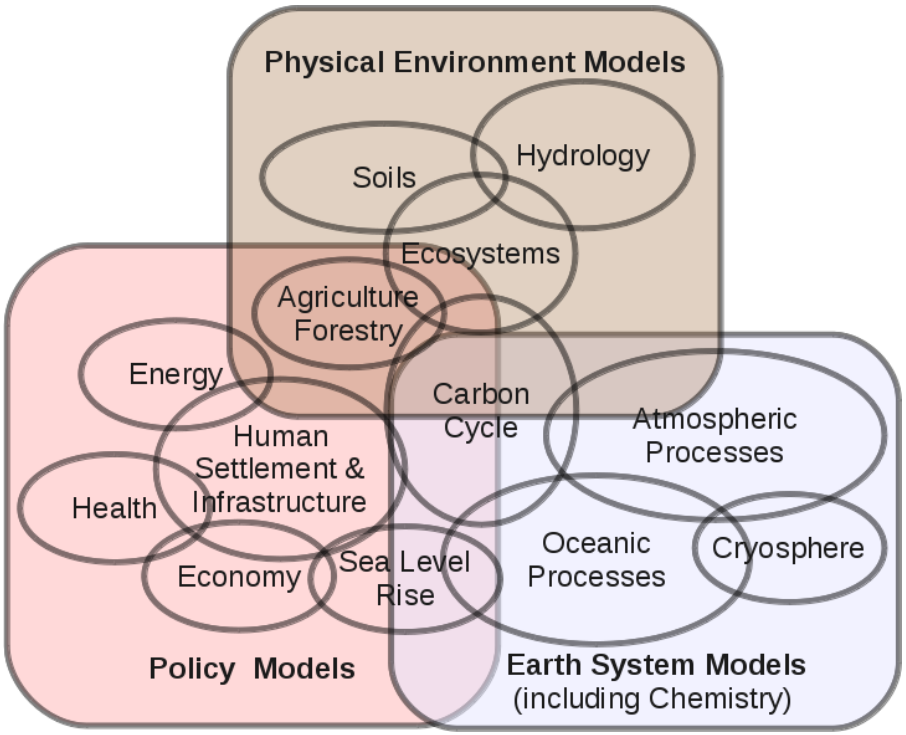
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

Computation and Networks



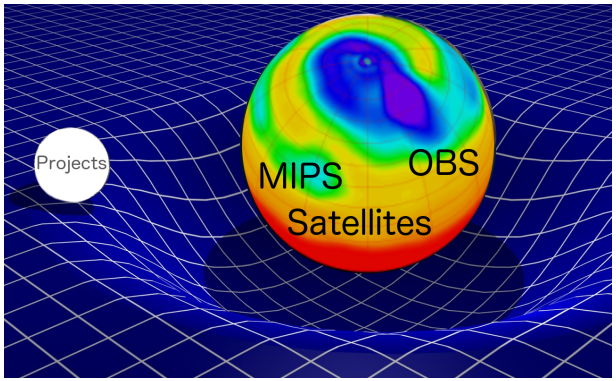
Growing range of interacting communities



Many interacting communities, each with their own software, compute environments, observations etc.

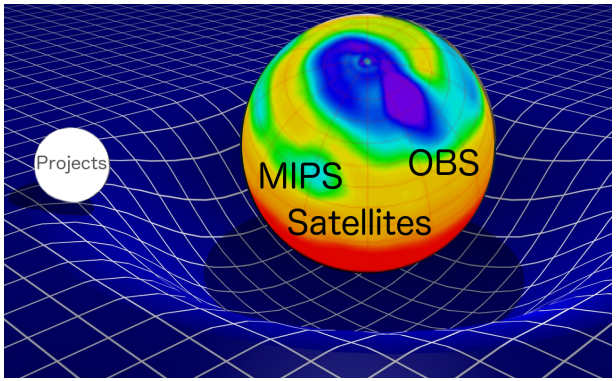
Figure adapted from Moss et al, 2010

JASMIN — The Data Commons

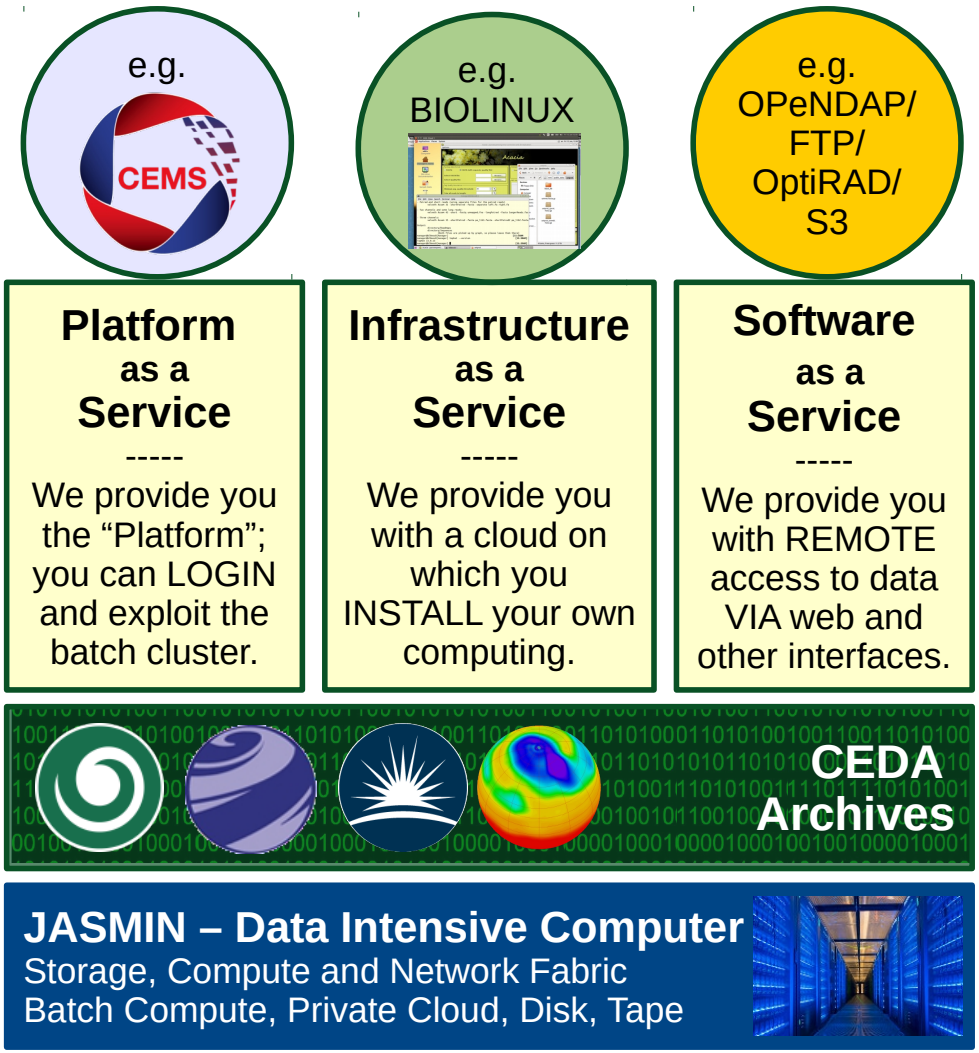


- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational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

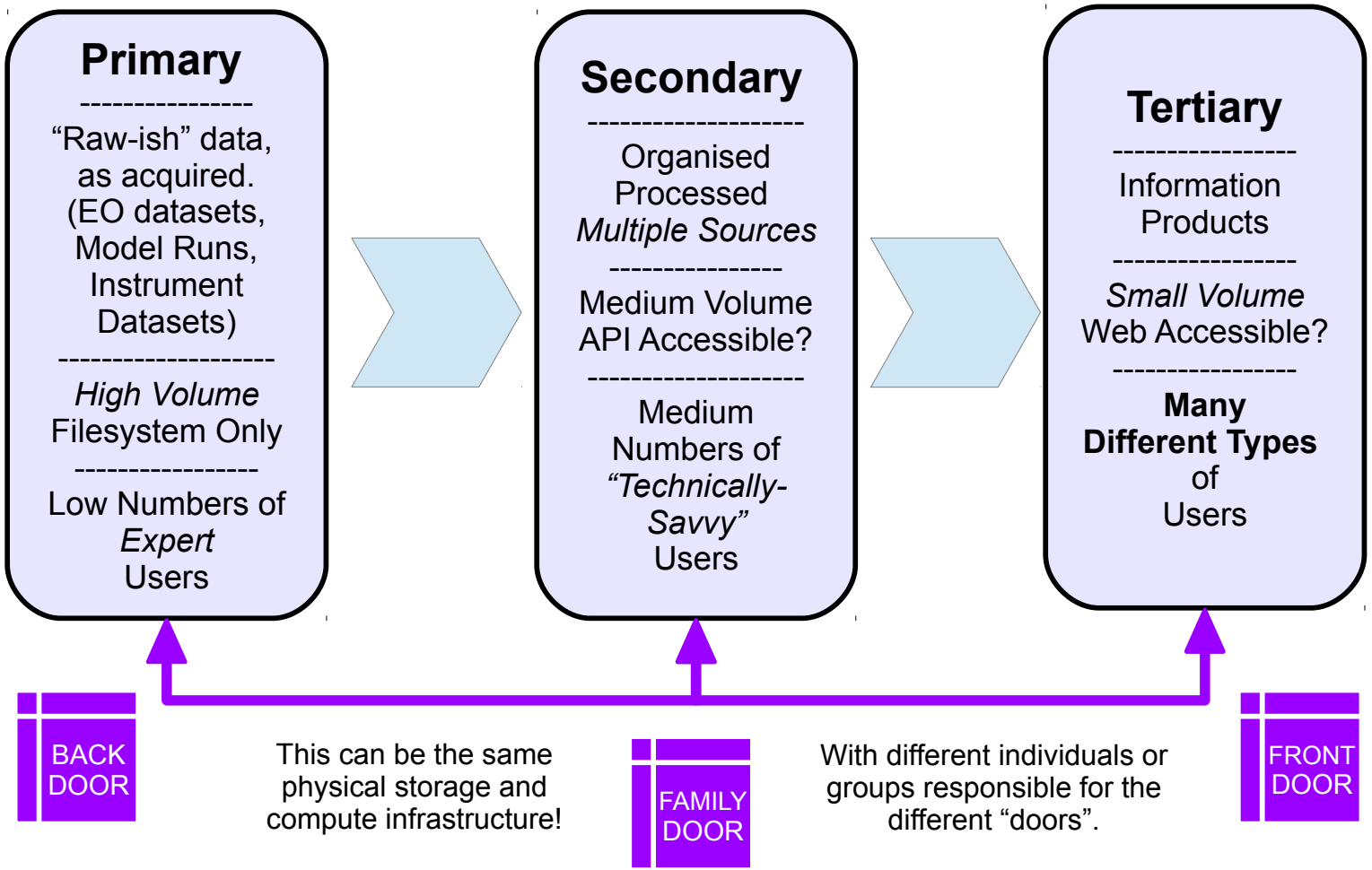
JASMIN — The Data Commons



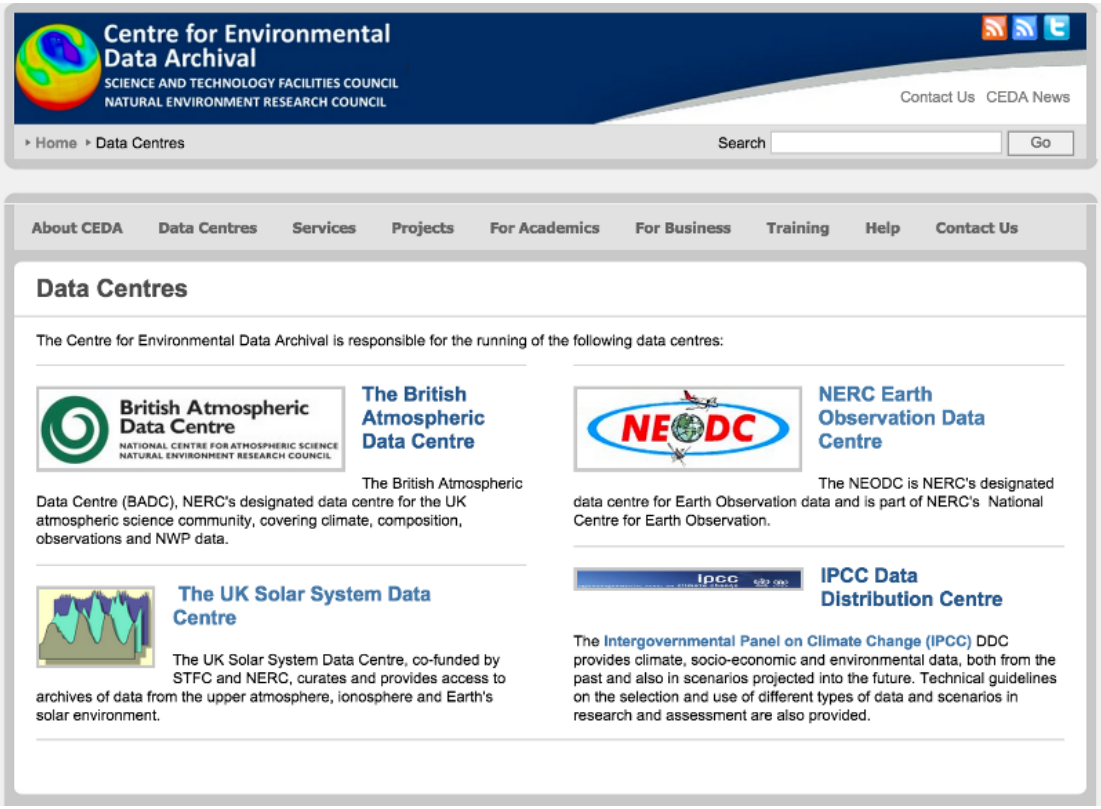
- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational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



Transforming data into information



CEDA

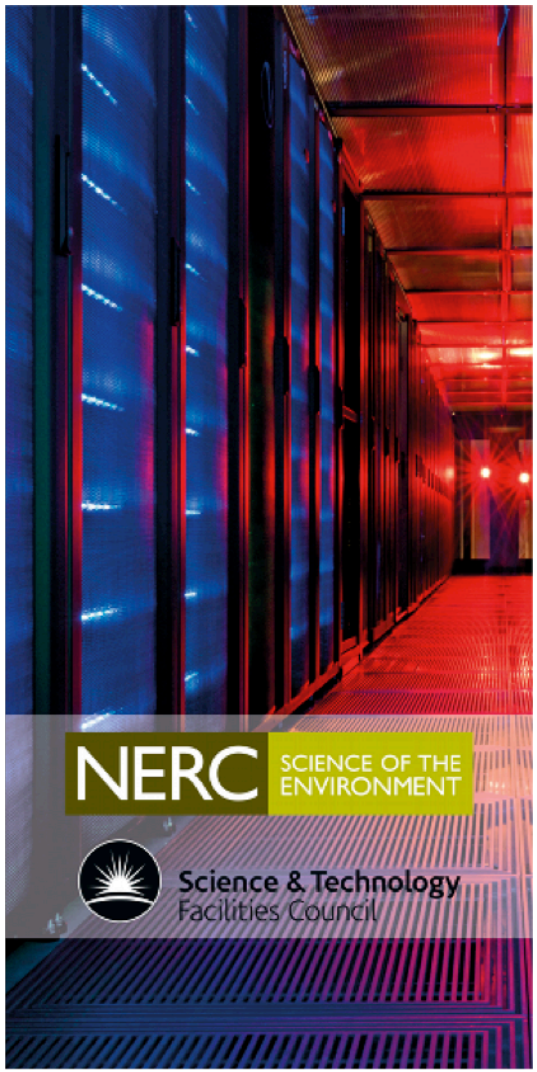


Four internal data centres: `http://ceda.ac.uk`
Acquiring and Curating Data Archives

- Provides the initial mass for the "gravity well", by feeding in both NERC and third party data products, available through the "back door".
- An example of a tenant organisation in its own right, delivering services through the "front door".
- Supports groups delivering customised services through "family doors".

Other data centres could be tenants and contribute to the data commons in the same way.

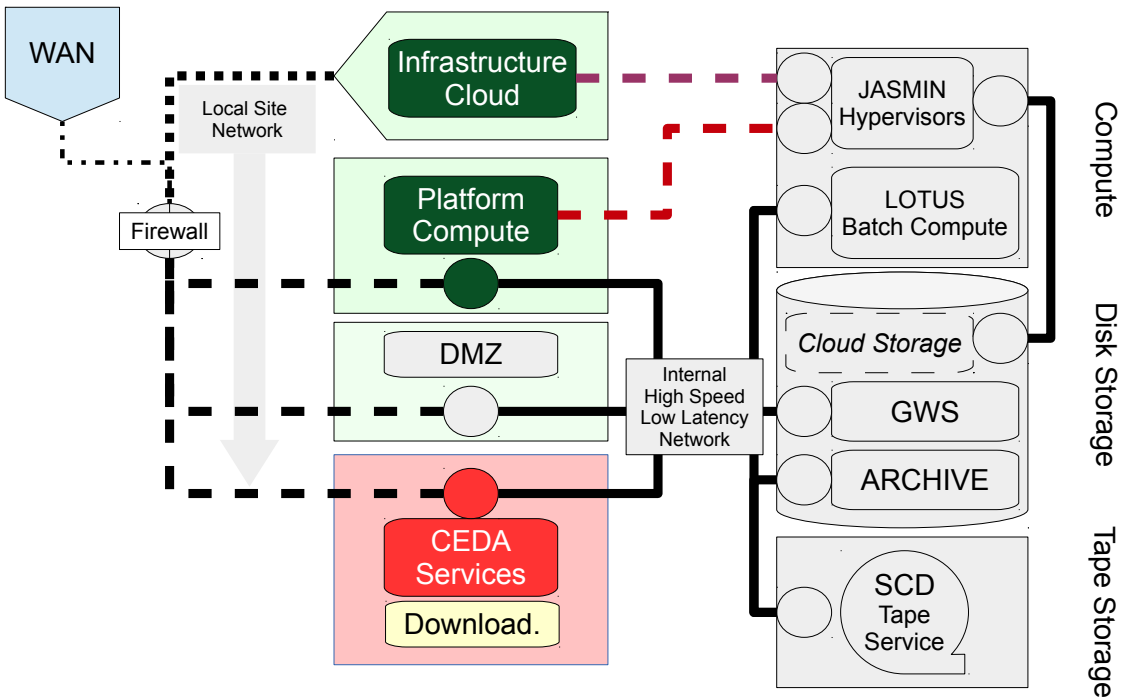
JASMIN



- 16 PB of fast disk; 0.5 PB of bulk disk (for virtual compute); >30 PB of tape.
- 5000 compute cores (cluster and hypervisors); dedicated high memory and transfer machines.

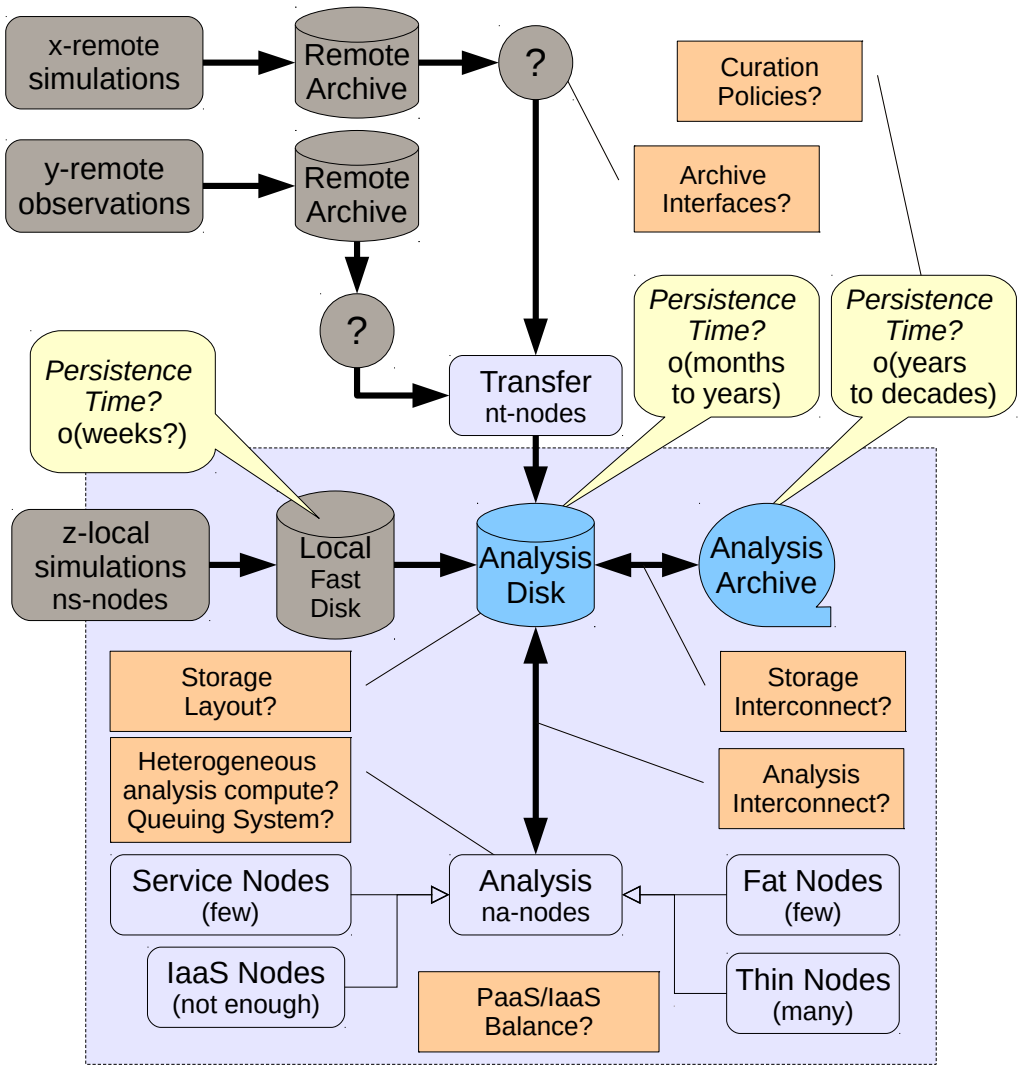
- **The Archive** - curated data directly available to local compute.
- **Group Work Spaces** — fast storage with tape accessible via the "Elastic Tape" service.
- **Generic Platform Compute** — machines configured for generic scientific analysis and data transfer.
- **Hosted Platform Compute** — bespoke machines deployed in the "Managed Cloud".
- **Infrastructure Compute** — private cloud portal and customised compute in the "Un-Managed Cloud".
- **Lotus Batch Cluster** — managed cluster with a range of node configurations (processor and memory).

Architecture



- JASMIN Internal Network: 10 Gbit non-blocking ethernet with low-latency (Mellanox) switches.
- JASMIN Compute: primarily deployed in the LOTUS batch cluster, although two shared private clouds deployed on the hypervisor systems.
- JASMIN Storage: primarily the Panasas fast parallel file system supporting the archive and group work spaces. Fast non-blocking network the heart of JASMIN!
- Tape Support for the archive (Storage-D) and the GWS (Elastic Tape).

Issues in Play

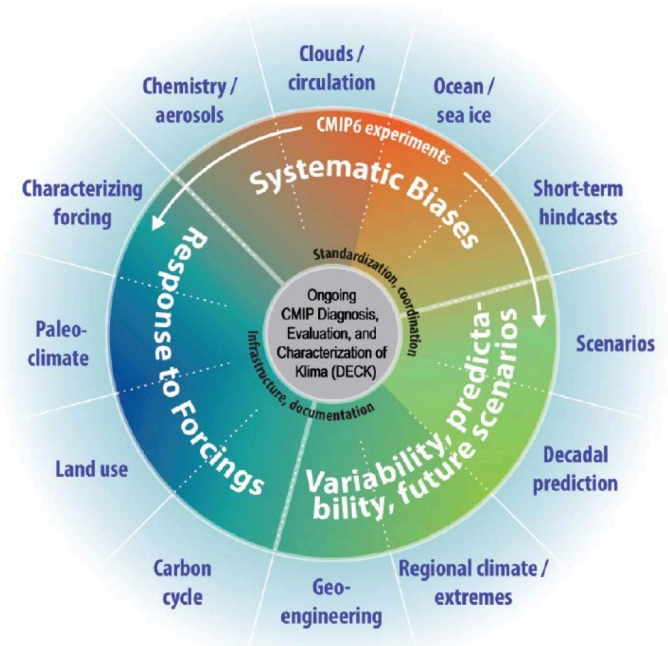


In the petascale era, we're handling petabytes of storage with terabytes in each of hundreds of workflow.

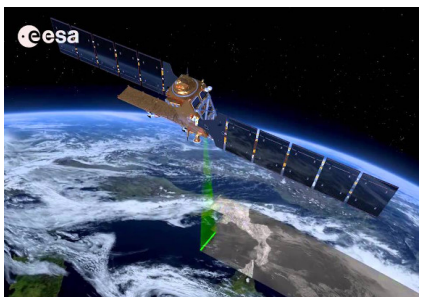
In the exascale era, we'll have exabytes of storage, with petabytes in hundreds of workflows!

But we don't know much about those workflows, now, let alone in the future!

The Organised Data Deluge



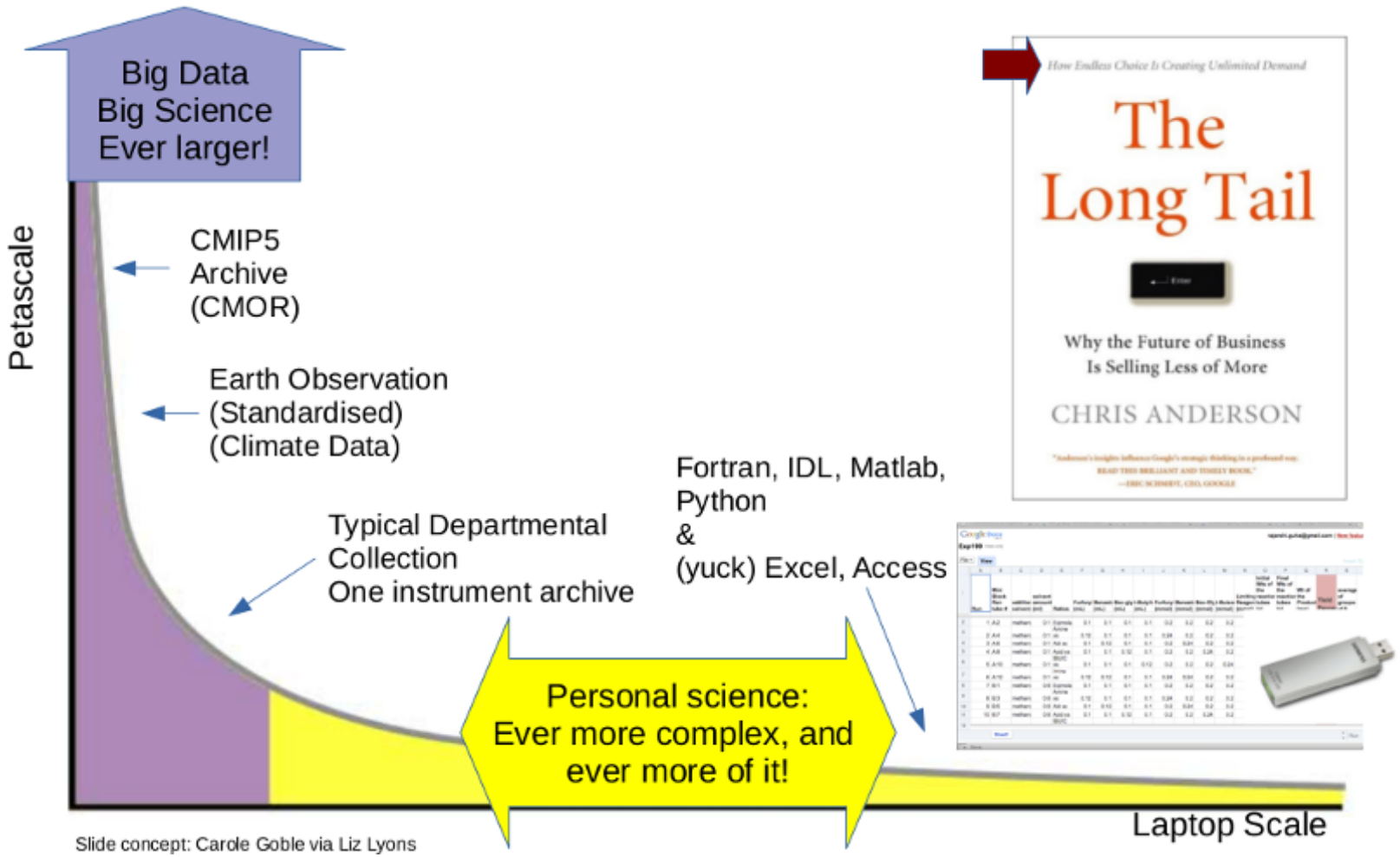
CMIP6 data volumes and data rates not yet known, but the European contribution to HiresMIP alone is expected to exceed 2 PB.



Sentinel 1A (2014), 1B (2016)
Sentinel 2A (2015) 2B (2017?)
Sentinel 3A (2016) 3B (2018?)

Data rate: o(6) PB/year

The unorganised data deluge



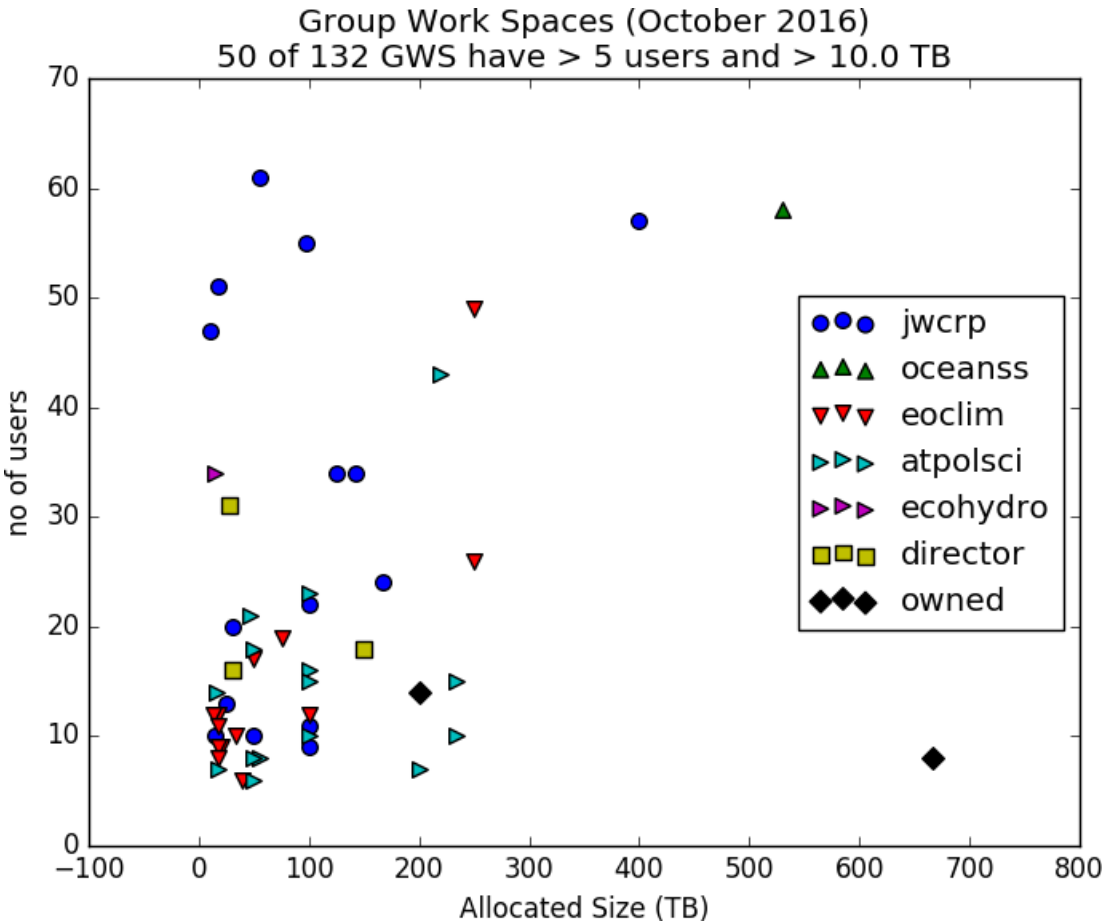
The 50 Largest Group Workspace (GWS) Tenancies on JASMIN

GWS > Consortia > Tenancies.

Tenancies get GWS resources

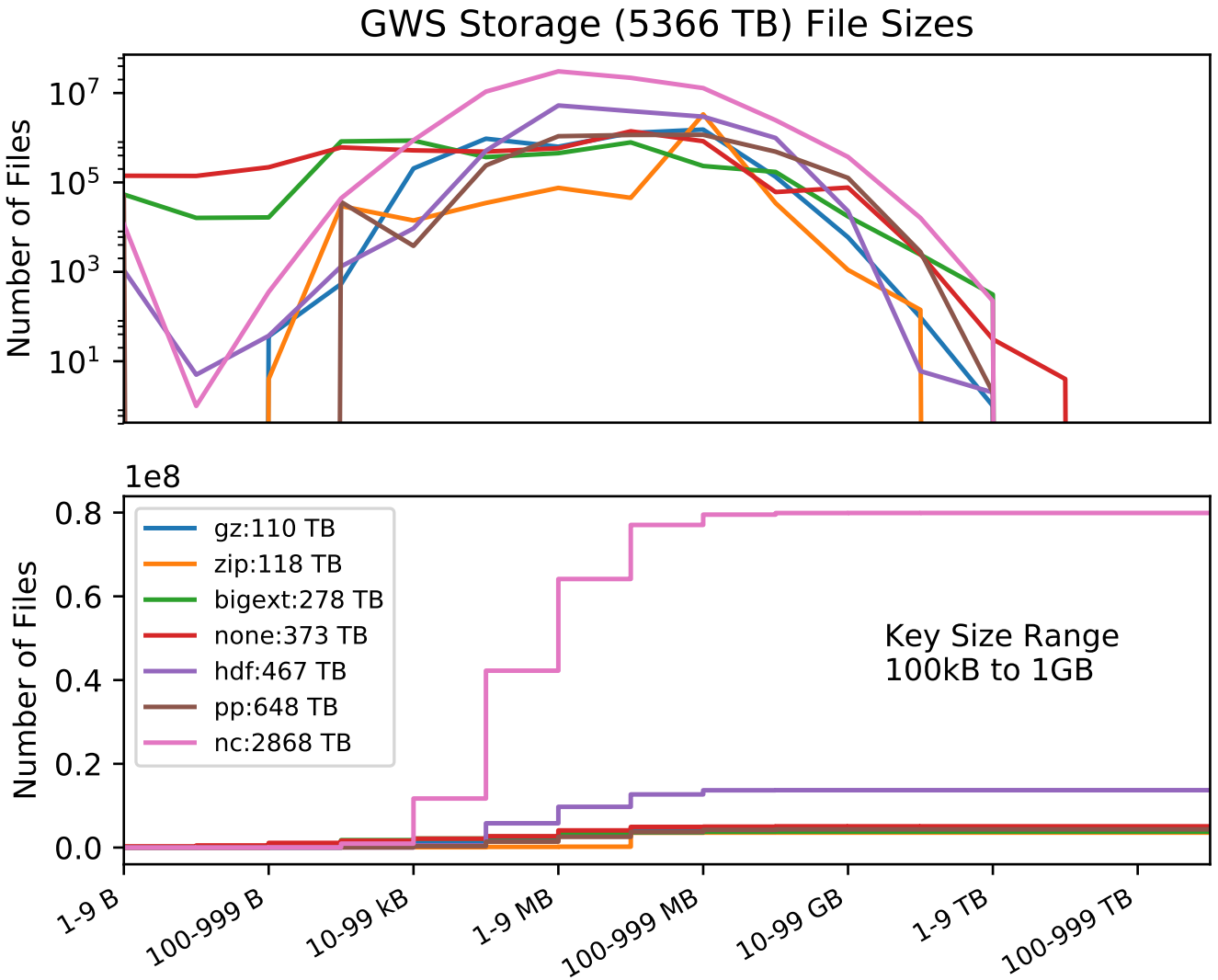
The largest consortia:

atpolsci	Atmospheric and Polar Science
director	Director's allocation (mainly supporting H2020)
ecohydro	Ecology and Hydrology
eoclim	Earth observation and climate services
jwcrp	Joint Weather and Climate Research Programme
oceanss	Oceans and Shelf Seas
owned	Resources owned by third parties within the JASMIN partnership



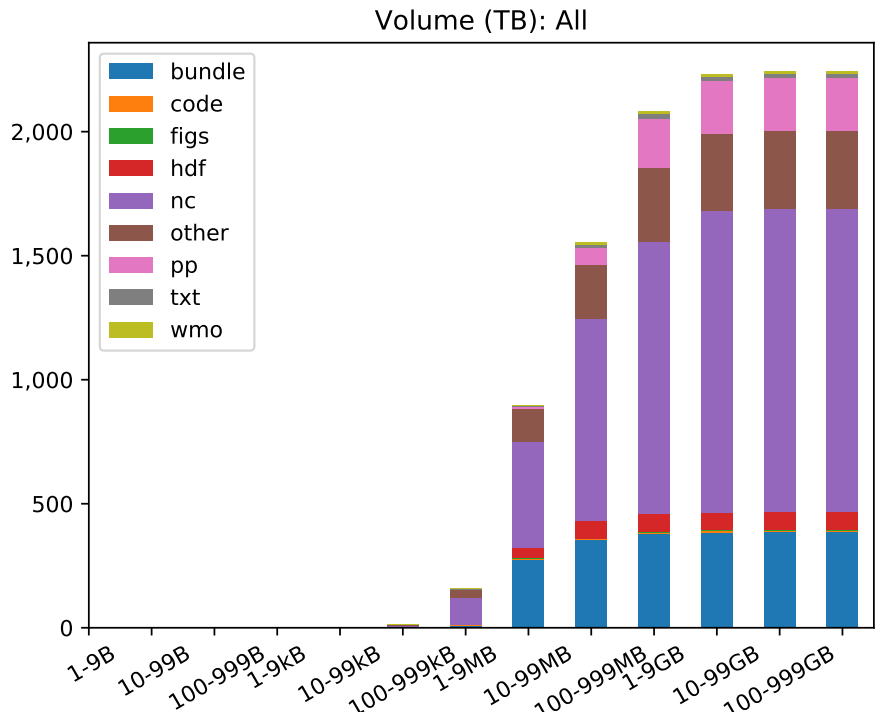
132 group workspace tenancies, supporting 822 users (of 1059 with login access). Most of the rest (< 10 TB) have a handful of users, but there are also 7 with (< 10TB) and (> 10 users); even for relatively small data volumes, sharing and co-location is important.

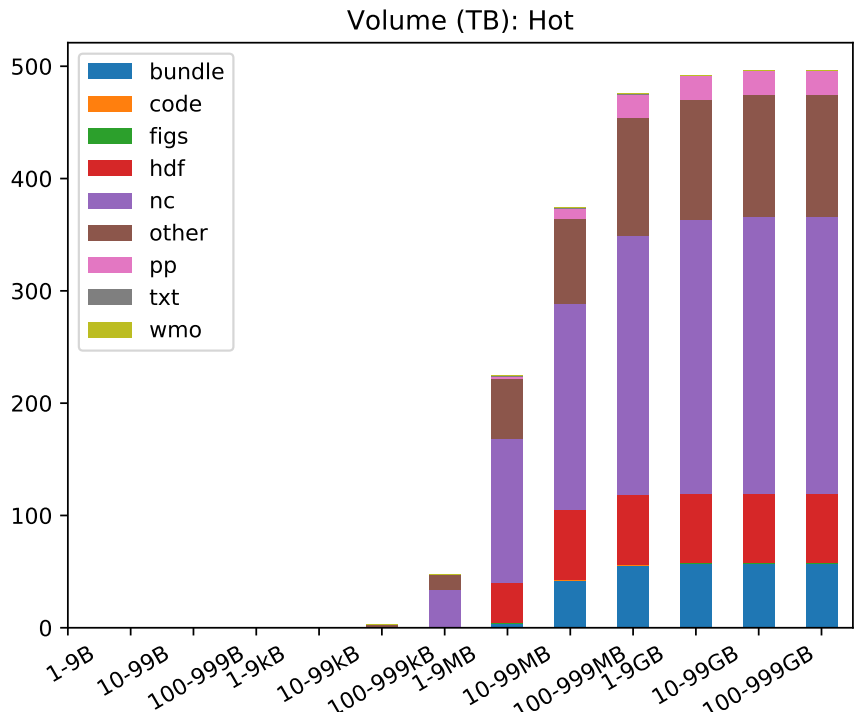
What do the contents of the GWS look like?



Hot or Not? (1)

Data from a *sample* of the GWS (but a **big** sample!).

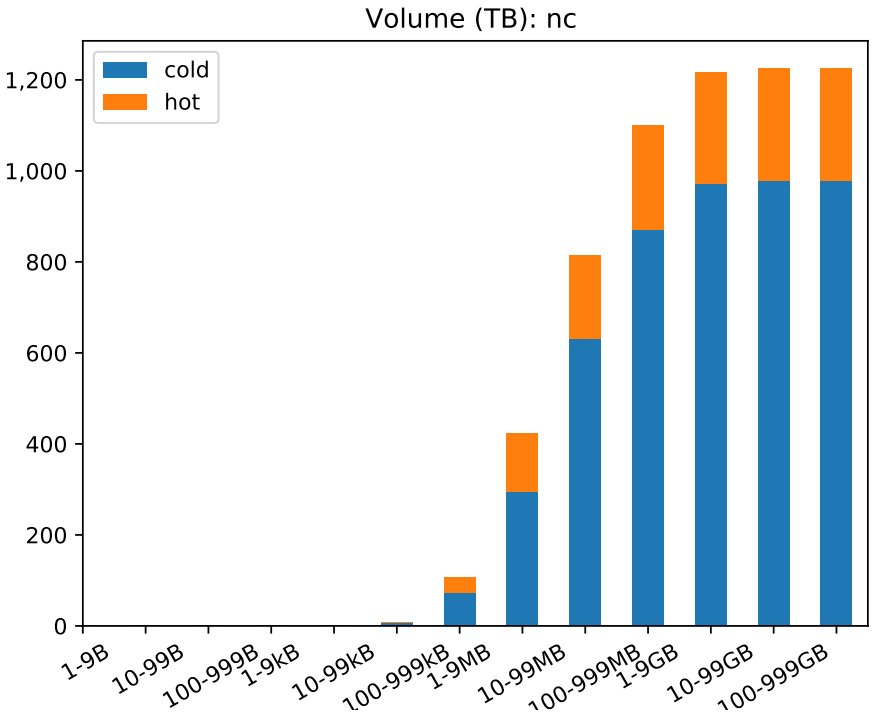


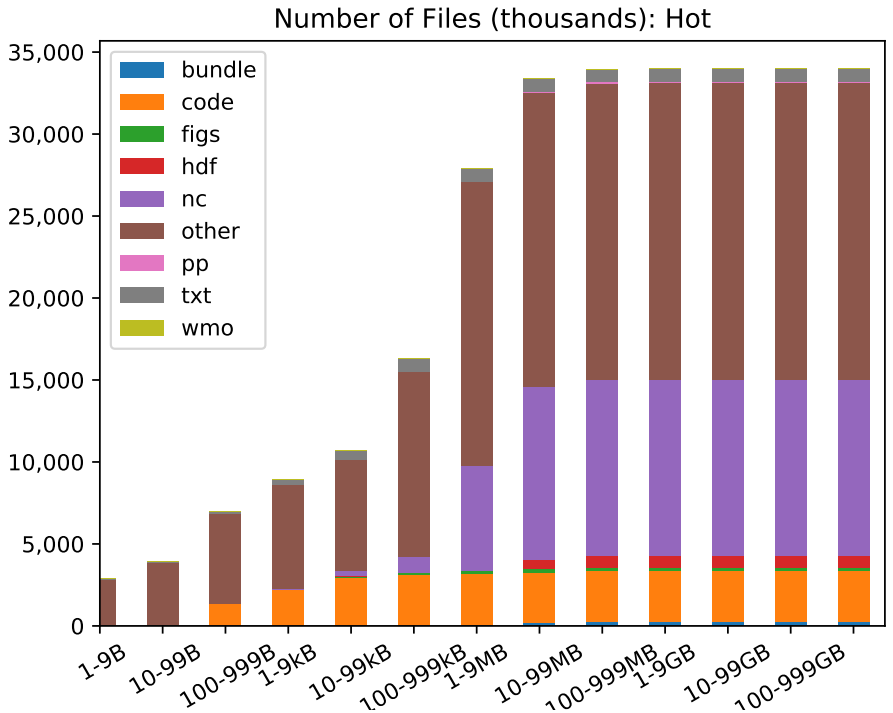


20-25% of the data on our GWS are "hot". Most of that is NetCDF.
Hot = touched in the last 3 months.

(There are no policies in place that mean these figures will have been gamed!)

Hot or Not? (2)

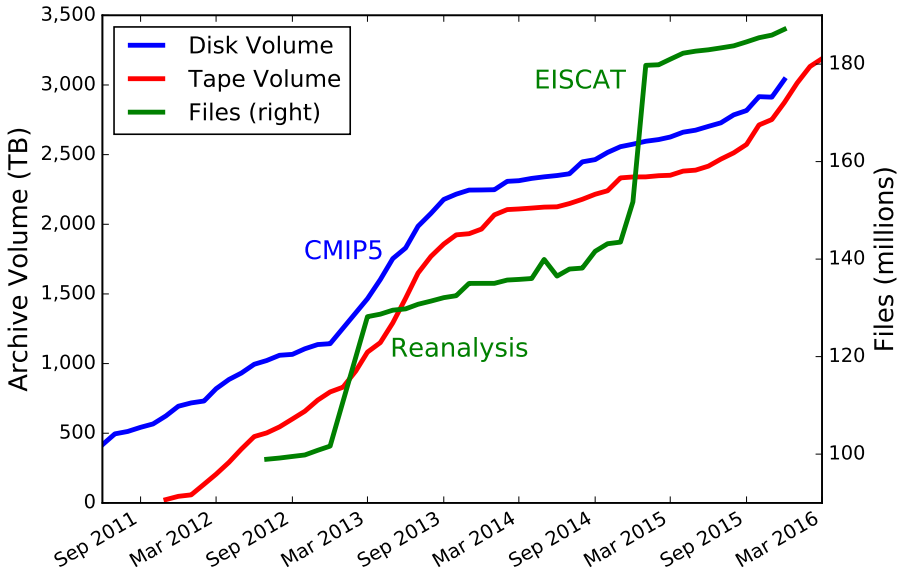




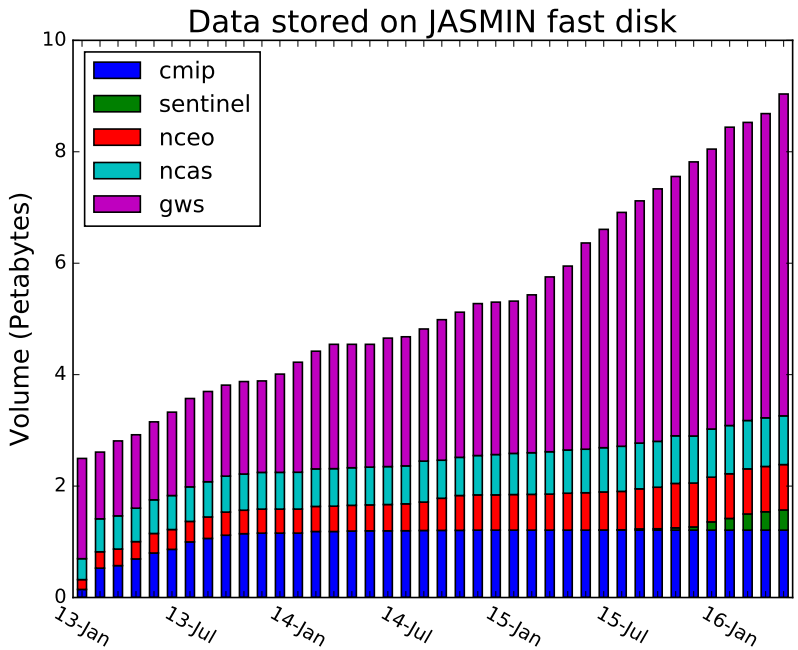
Most NetCDF data is cold though!

Obviously there are a lot of hot files we don't know anything about (tens of millions) but the volumes are modest (a couple of hundred TB at most in this sample).

Storage Volumes

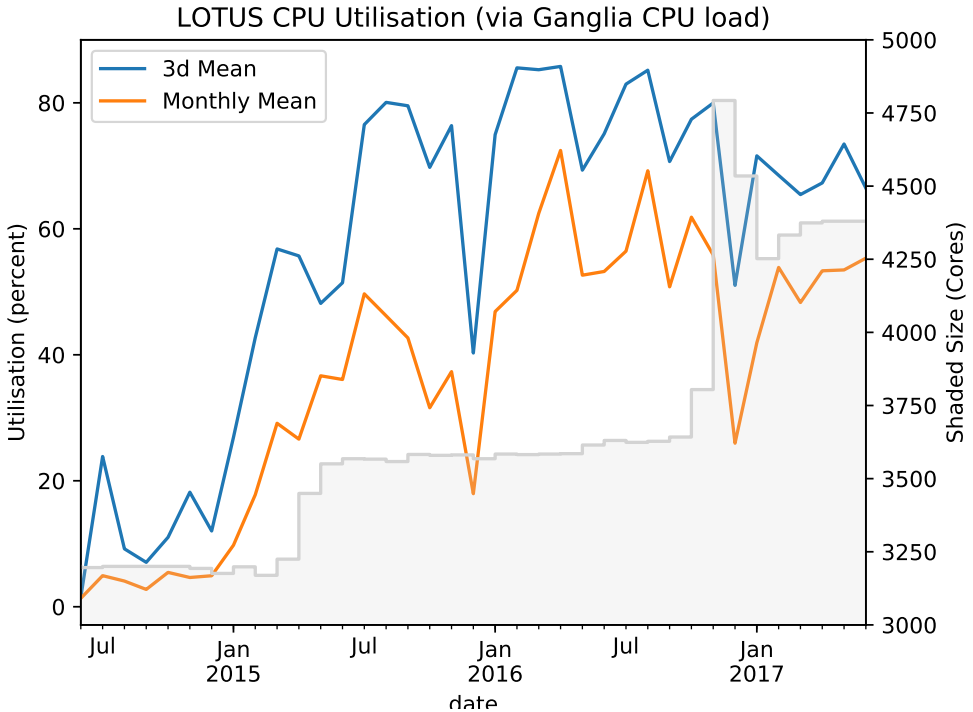


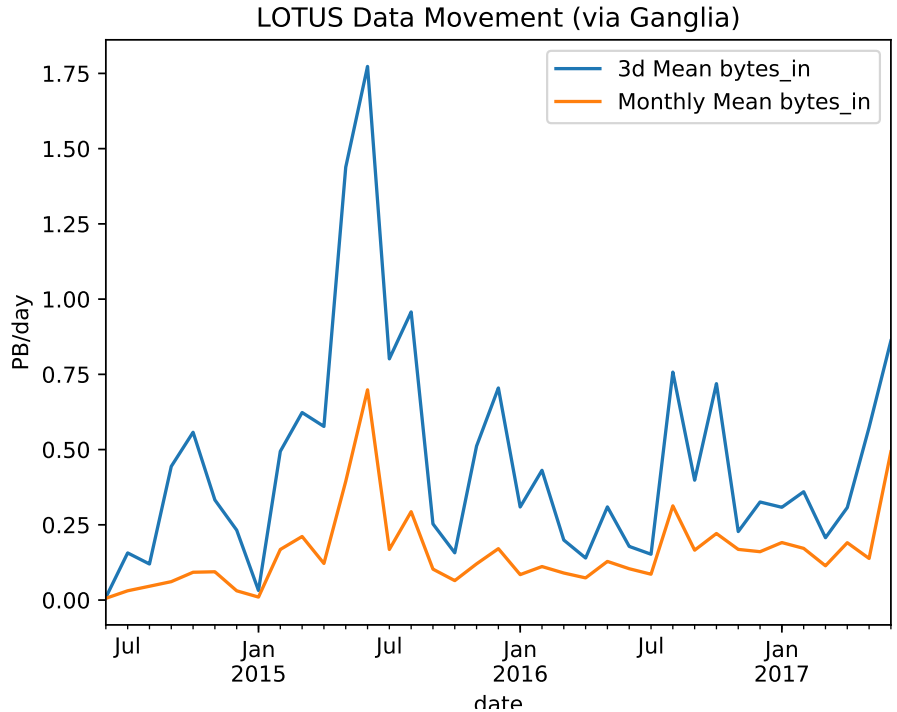
Archive growth is episodic!



GWS growth is continuous!

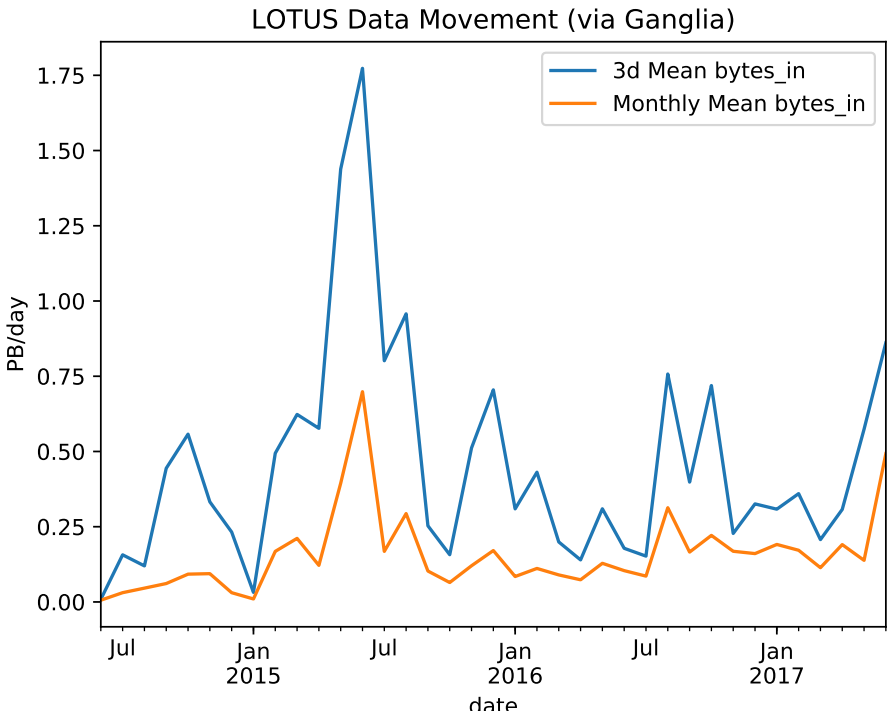
Batch Cluster Usage

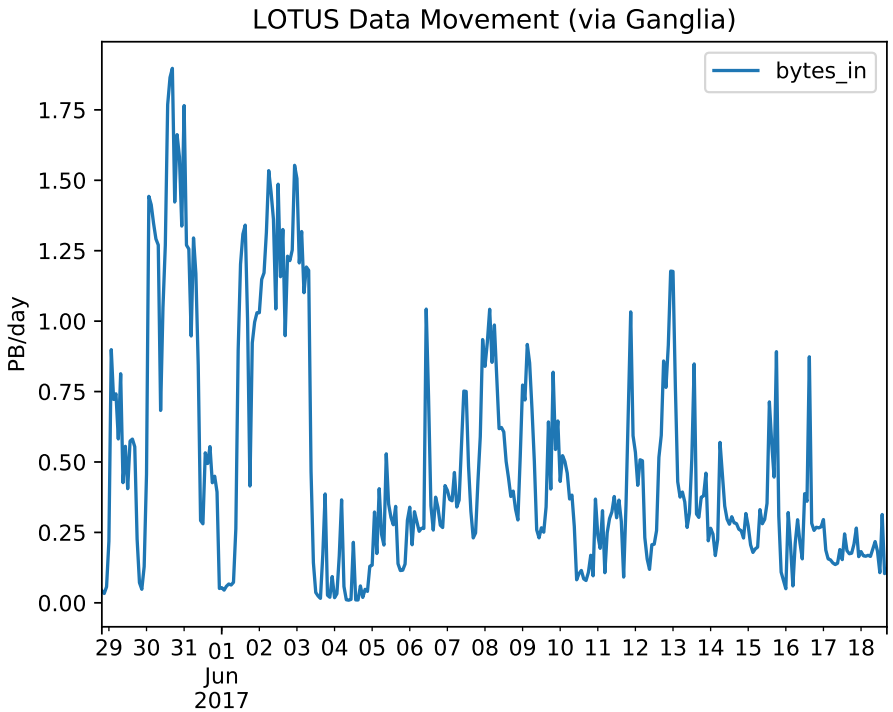




- CPU utilisation is not the priority; CPU availability is key for an analysis environment. Batch is primarily a convenience for parallelisation not a method for filling the machine.
- Data movement requirements are episodic

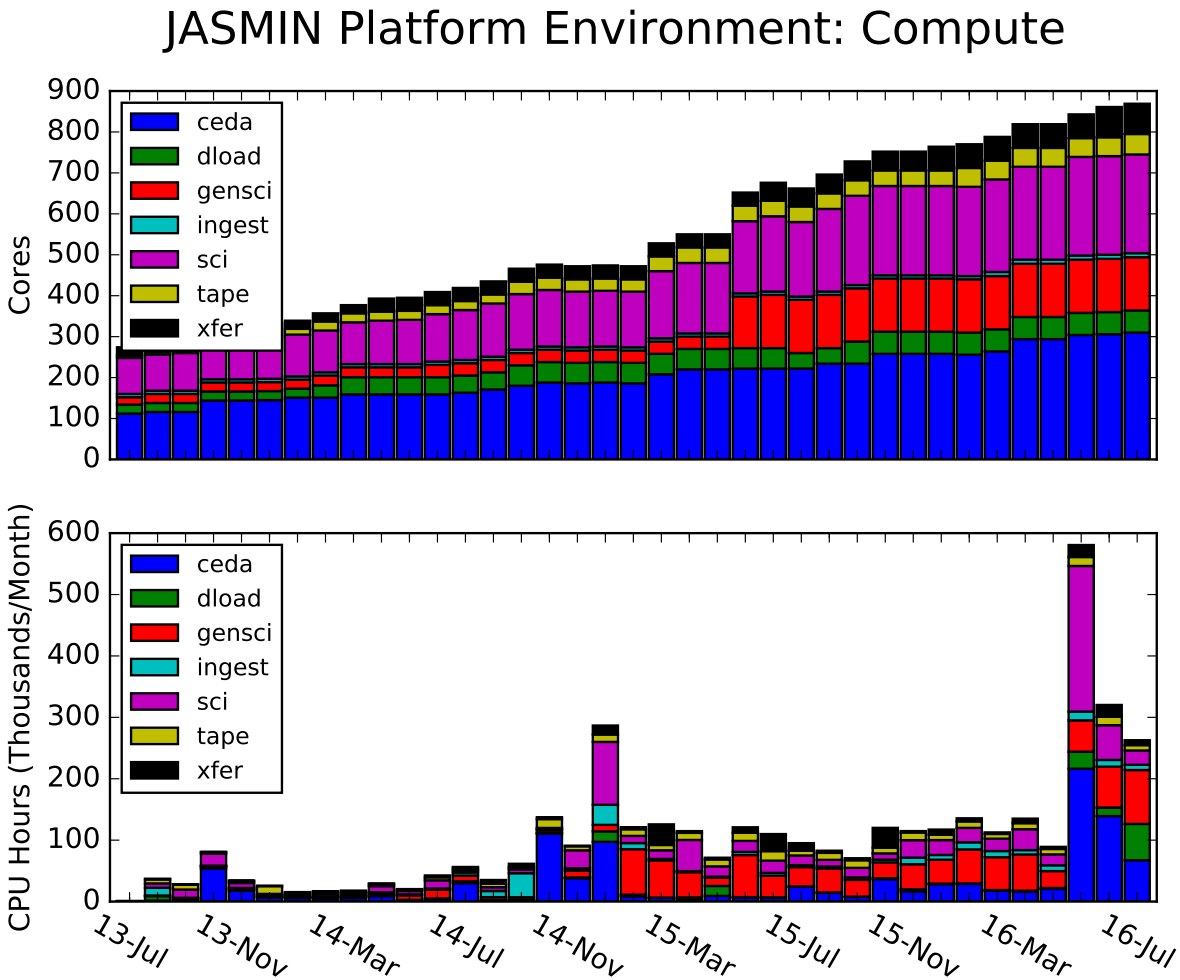
Batch Cluster Usage





- CPU utilisation is not the priority; CPU availability is key for an analysis environment. Batch is primarily a convenience for parallelisation not a method for filling the machine.
- Data movement requirements are episodic, on long and short term timescales.

PaaS Usage



Continual process of adding capacity to support new use cases.

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

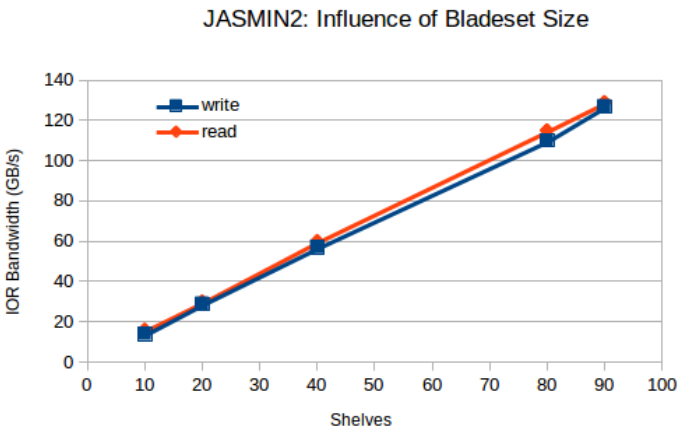
Pick four, all in play:

- *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)* For data, we avoid this by having a parallel file system (we think). What will happen when we don't have a parallel file system?
- *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).

We need work on understanding all these things!

Pick one issue: I/O Workload - System View

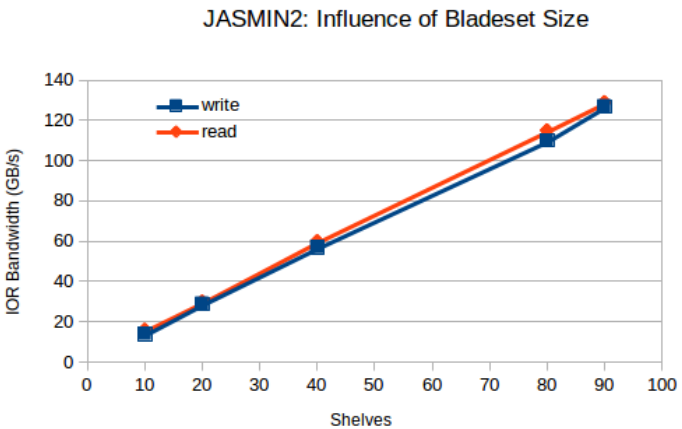
JASMIN2 (6.5 PB, about a third of the system)



- J2 delivers around 1 Tb/s for IOR.
- Whole system IOR now probably 3 Tb/s, but we deploy our Panasas system in bladesets to optimise performance, durability, and rebuild times.

Pick one issue: I/O Workload - System View

JASMIN2 (6.5 PB, about a third of the system)

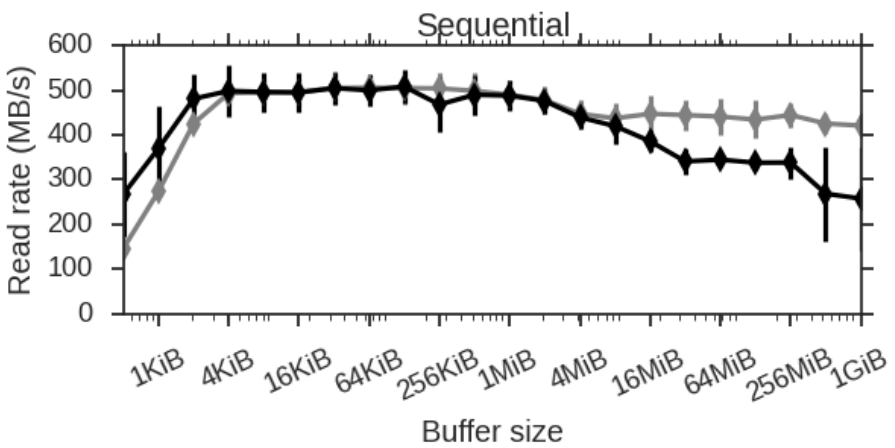


- J2 delivers around 1 Tb/s for IOR.
- Whole system IOR now probably 3 Tb/s, but we deploy our Panasas system in bladesets to optimise performance, durability, and rebuild times.

Delivering (more) I/O performance to communities:

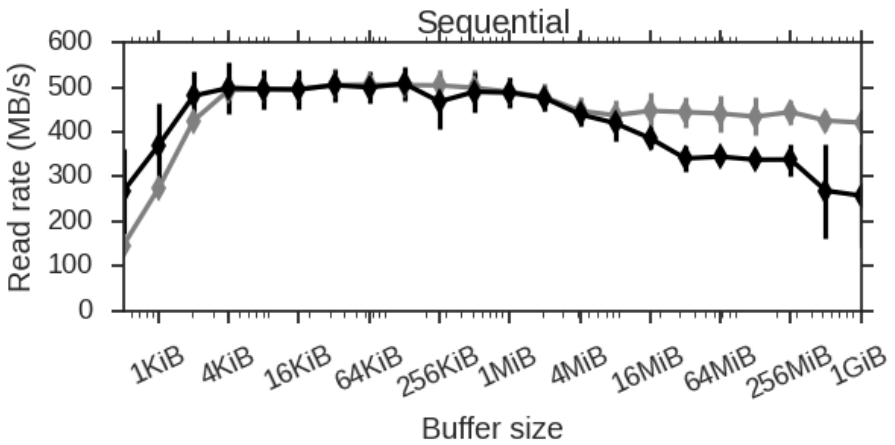
- Average bladeset size is approximately 10-12 shelves.
- IOR suggests we can get 15 GB/s read performance for 10 shelves.
- Typically tenants get GWS which are deployed on one or at most two bladesets.
- Multiple tenants per bladeset!
- Tenants contend with each other on a bladeset, but not with users on other bladesets (non-blocking network!)
- Archive is deployed across multiple bladesets, and archive bladesets are not typically shared with GWS tenants.

Pick one issue: I/O (READ) Workload - User View

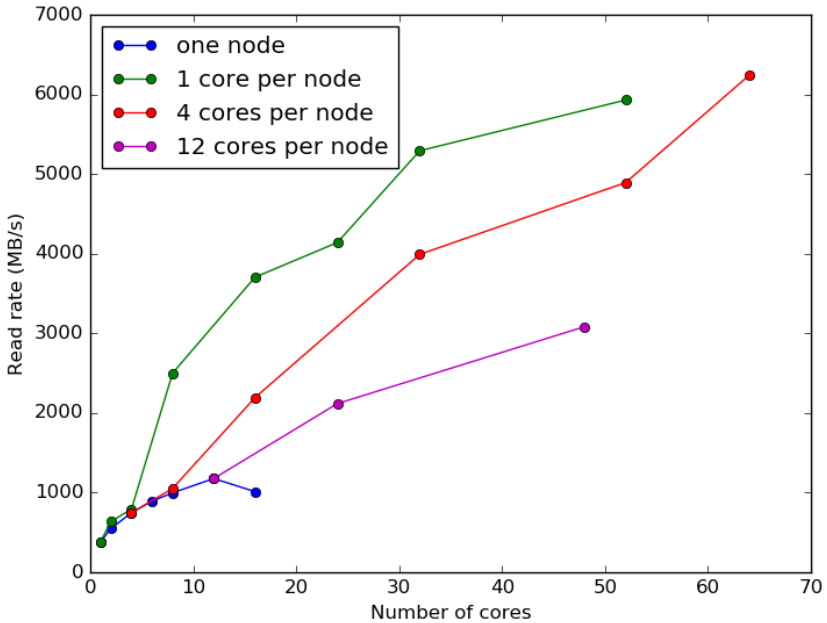


- Testing single threaded read performance on a 256 GB file using default Panasas settings.
- Sequential read through the entire file (black lines are using Python, grey lines using C).
- These results are comparable to those seen on a Lustre file system at Archer.
- Tunable file-system parameters can make significant differences, but such a priori choices may not meet the full range of read-based use cases.

Pick one issue: I/O (READ) Workload - User View



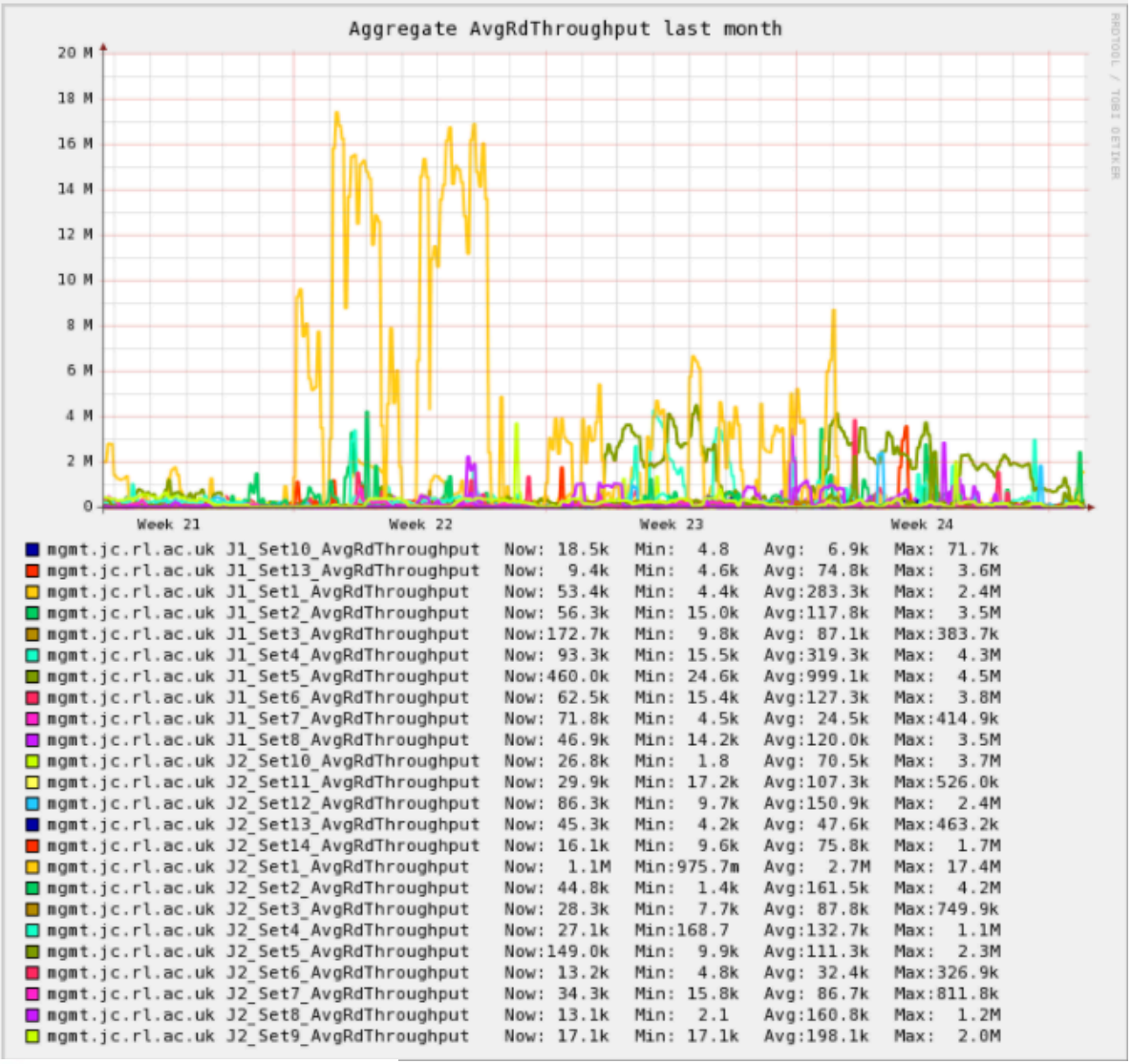
- Testing single threaded read performance on a 256 GB file using default Panasas settings.
- Sequential read through the entire file (black lines are using Python, grey lines using C).
- These results are comparable to those seen on a Lustre file system at Archer.
- Tunable file-system parameters can make significant differences, but such a priori choices may not meet the full range of read-based use cases.



- Can get a significant percentage of the theoretical bandwidth reading from a single 256 GB file using multiple client nodes!
- A significant optimisation problem to work out how many cores per node . . .
- . . . hard to evaluate in the presence of contention with other users of the client nodes.

PhD work of Matt Jones, University of Reading.

Bladeset Usage and Performance



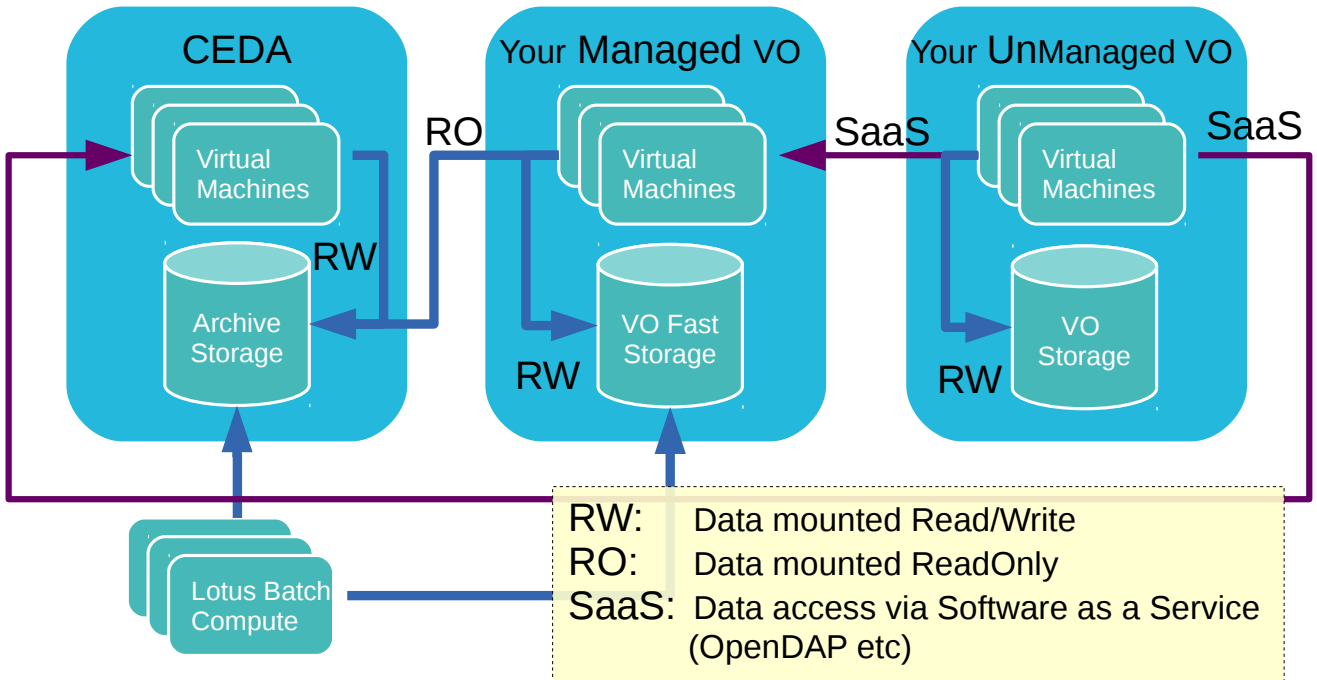
Last month of bladeset READ throughput from PanFS!

- The large yellow curve in the first couple of weeks shows the sustained usage of an "owned" bladeset by an earth observation research group.
- The dark green curve in the last couple of weeks is CEDA operational usage (ingest, data processing etc).

(These results are with a prototype information system, fed through Ganglia. At this point the units are wrong, but the differences are interesting anyway . . .)

What about our cloud?

Objective is to provide an environment with high performance access to curated data archive **and** a high performance data analysis environment - both directly mounted and indirectly accessed!



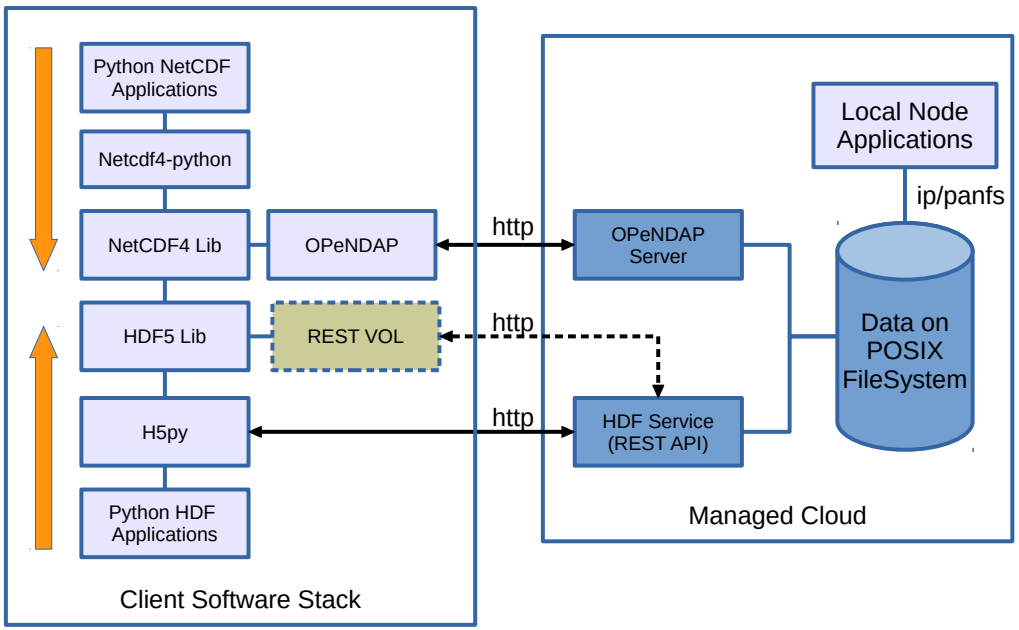
All this in the presence of data growth that exceeds the Kryder rate (that is, data growing faster than storage costs are falling)!

Some guiding principles for our storage environment

- Demand for storage is growing — fast.
  - We can't afford to solve the problem with parallel disk.
  - Object store has some interesting properties beyond price, not least more efficient purchasing/migration/maintenance strategies.
  - We can't afford to solve the problem with any sort of disk alone.
- We need to support a range of deployment environments (with containers, internal and external cloud, batch clusters.
- The future will be tiered.
- Experience with optimising code suggests that domain specific knowledge leads to optimal solutions.
  - We know that HDF (and NetCDF) is a huge part of our workload.
  - We know that we have got (and can have more) use of semantic conventions.
  - We know that not much of the data is really that hot, but file-based tape systems are not that efficient.

Working with Services - Today - OPeNDAP and HDF Server

Working with Services - Today - OPeNDAP and HDF Server



Currently one has two routes to "relatively generic" data access services:

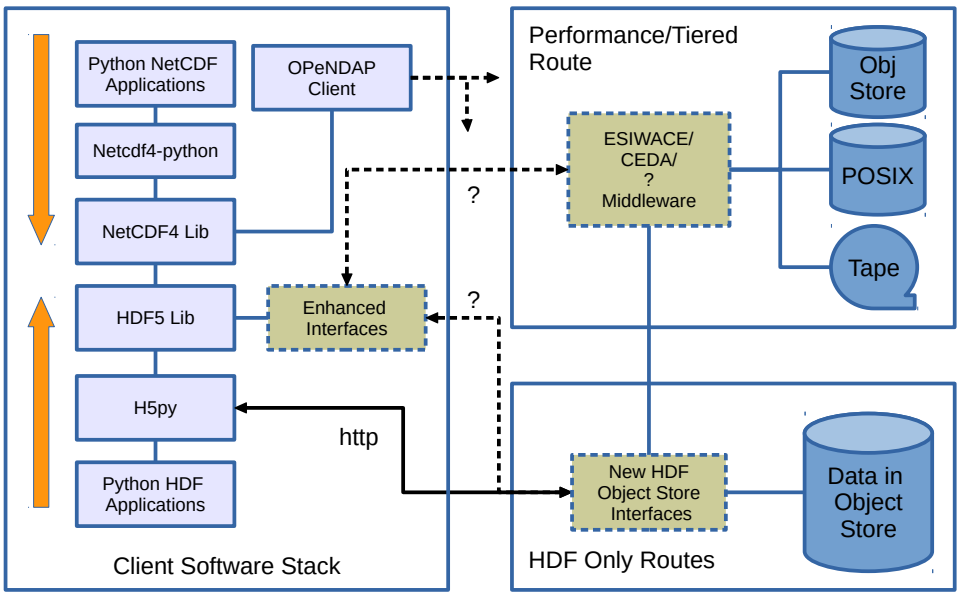
- OPeNDAP, and
- H5Serv access to data on a filesystem which can be accessed via the H5PY client library.

OPeNDAP is relatively well established, and we are rolling out services:

- External facing "archive access", and
- (Primarily) internal facing "high performance" based on 100 Gbit/s physical servers.

https://github.com/HDFGroup/h5serv

Working with Services - It's all about interfaces



Interfaces client-side (almost certainly based on HDF), and Interfaces server-side (can we get performance from http)?

- The HDF Group are working on a range of interesting new interfaces to object stores, some of which are designed for performance, some for fidelity.
- We† are working on range of projects:
  - ESIWACE (with DKRZ, Seagate, THG, CMCC and DDN): new middleware for tiered storage,
  - Our own internal lightweight tiered storage system, and
  - engaging with others: the right way forward isn't really known, but we know it probably needs to take advantage of our domain specific knowledge!

† We means CEDA (= STFC, NCAS & NCEO)!

Summary

1. UK JASMIN system provides an environmental data analysis commons, for observations and simulations from multiple sources.
2. Current hardware environment supports both interactive and batch cluster access.
   - There is a lot of data movement in both.
   - The network is not stressed, and we provide tenancies bandwidth isolation for their own data — but there could be contention for archive access, and it is difficult for users to get the bandwidth that exists anyway!
3. We can't afford to carry on with parallel disk, and we don't think tape alone is a solution, so we are investigating object stores, and object store interfaces.
4. (We have a PB of object store in testing now, with plans to purchase PB's more this year) ... but software and middleware *which does not yet exist*— will be crucial to the success of these plans.