JASMIN - A NERC Data Analysis Environment

Bryan Lawrence
NCAS, STFC & The University of Reading

Outline

Enterprise View

J is for Joint

Jointly *delivered* by STFC:
- CEDA (RALSpace) and SCD.

Joint *users* (initially):
- NERC community & Met Office

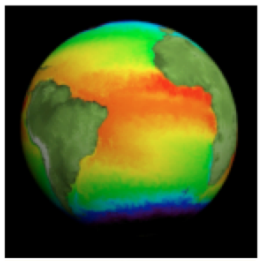
Joint *users* (target):
- Industry (data users & service providers)
- Europe (wider environ. academia)

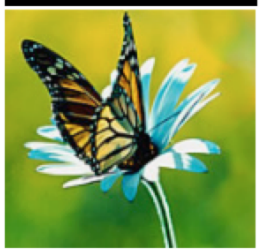
A is for Analysis

- 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)

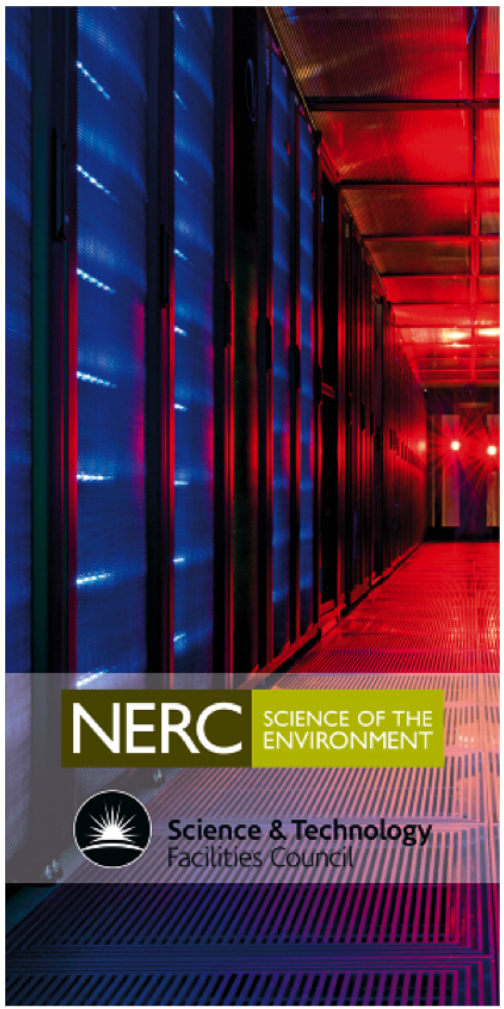
with UKSA (& the Space Catapult via CEMS)

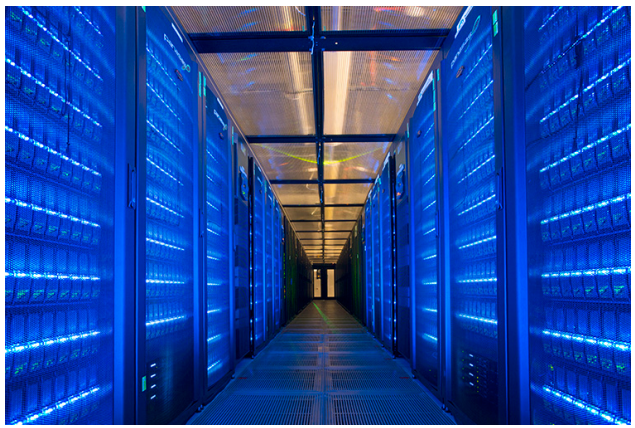
with EPSRC (joined up national e-infrastructure)

with industry (cloud providers, SMEs)

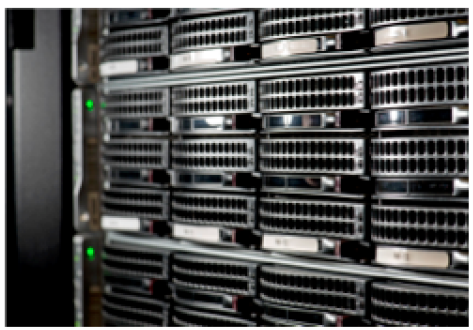
(CEMS:the facility for Climate and Environmental Monitoring from Space)

Gratuitous Photos

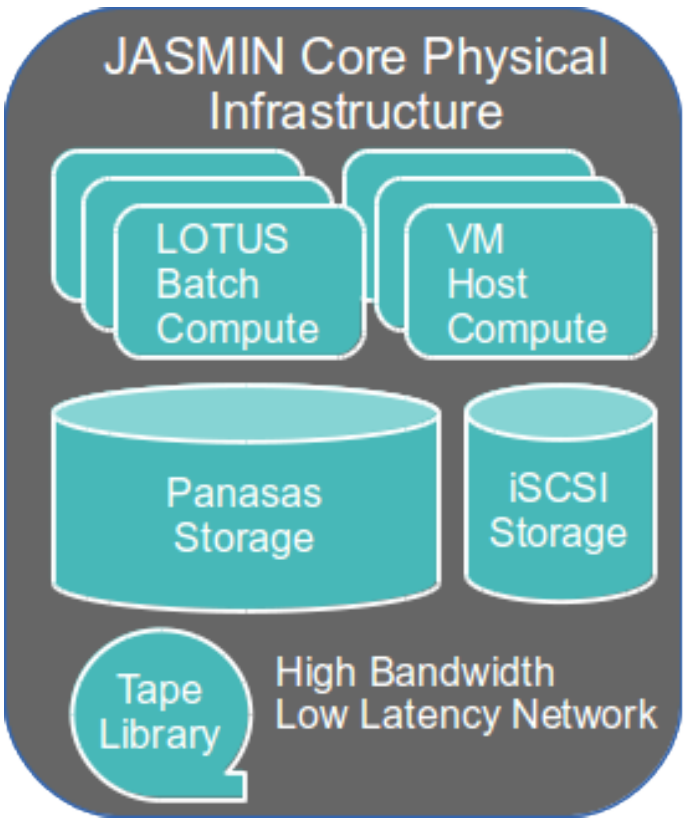


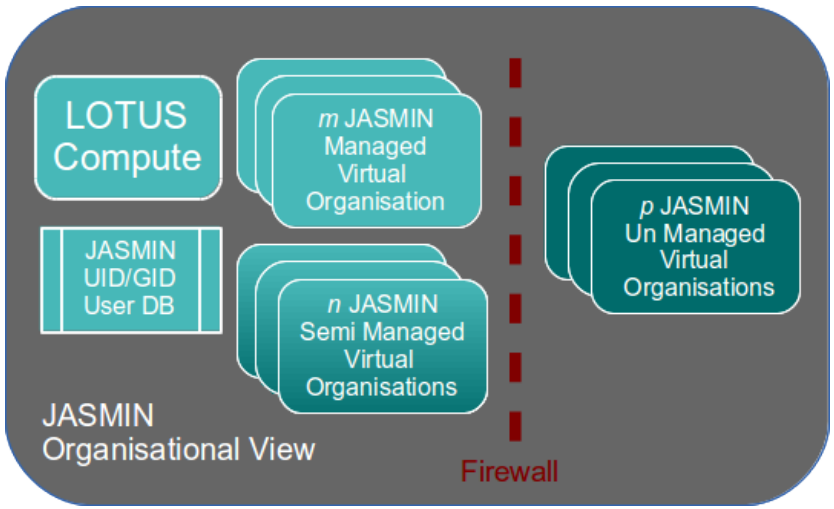


- 12 PB Fast Storage
- 1 PB Bulk Storage
- Elastic Tape
- 4000 compute cores: half deployed as hypervisors, half as the "Lotus" batch cluster.



Physical and Organisational Views

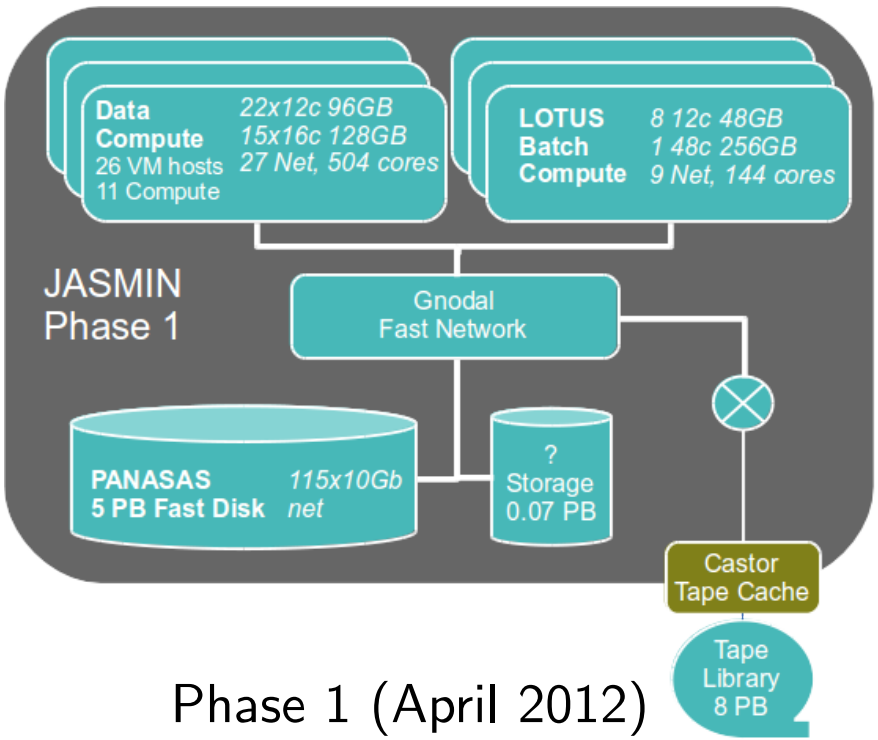




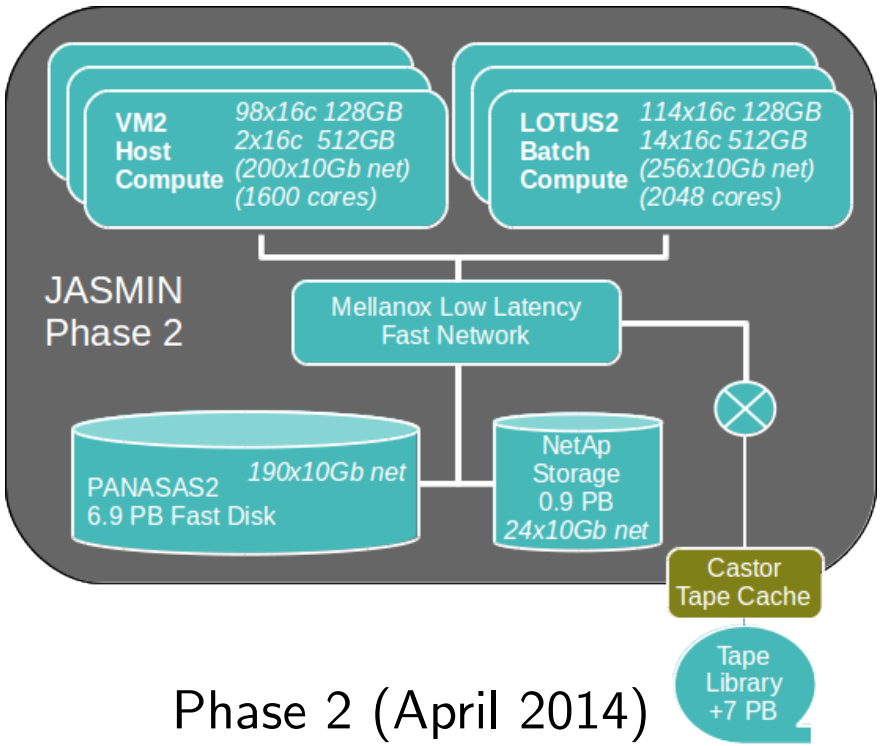
(We'll come back to this view)

JASMIN Phases

Three phases (so far):



Phase 1 (April 2012)



Phase 2 (April 2014)

Phase 1 & 2 to be joined together as part of Phase 3
Phase 3 in procurement now, deployment March/April 2015.

JASMIN LOTUS Compute

Model	Processor	Cores	Memory
194 x Viglen HX525T2i	Intel Xeon E5-2650 v2 "Ivy Bridge"	16	128GB
14 x Viglen HX545T4i	Intel Xeon E5-2650 v2 "Ivy Bridge"	16	512GB
6 x Dell R620	Intel Xeon E5-2660 "Sandy Bridge"	16	128GB
8 x Dell R610	Intel Xeon X5690 "Westmere"	12	48GB
3 x Dell R610	Intel Xeon X5675 "Westmere"	12	96GB
1 x Dell R815	AMD Opteron	48	256GB

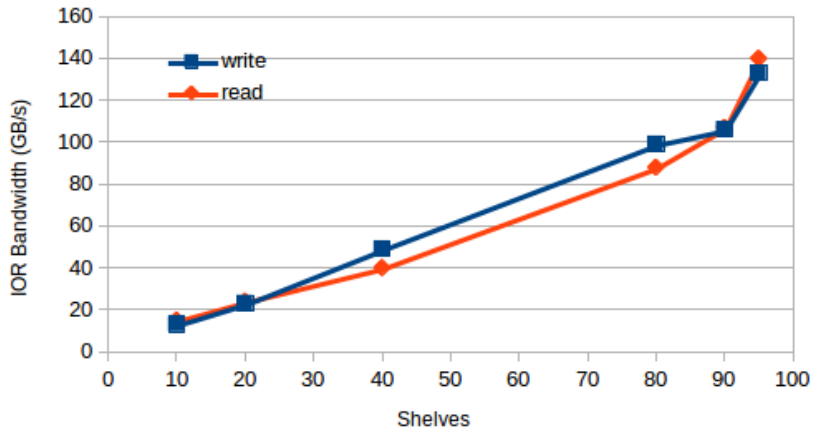
- 226 bare metal hosts, each with 2 NICs; 3556 cores!
- 17 large memory hosts
- Easily reconfigured between hypervisor and lotus roles!

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 $\approx$ 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 add another 1/3 performance to take us up to 200 GB/s overall ? certainly in the top-10, with JASMIN phase 3 to come.)

JASMIN2 Panasas I/O performance



Sakai et al performance (cf storage targets):

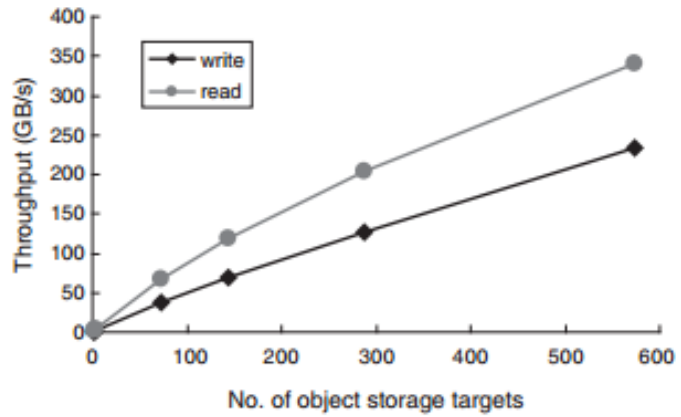
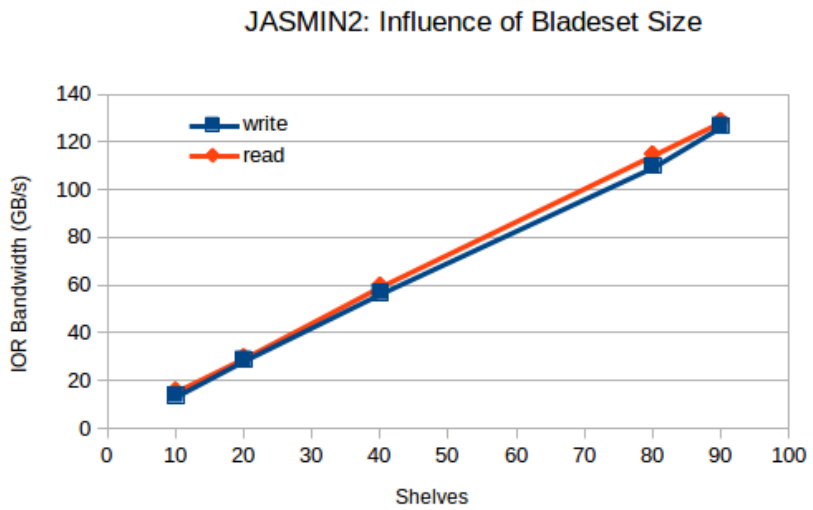


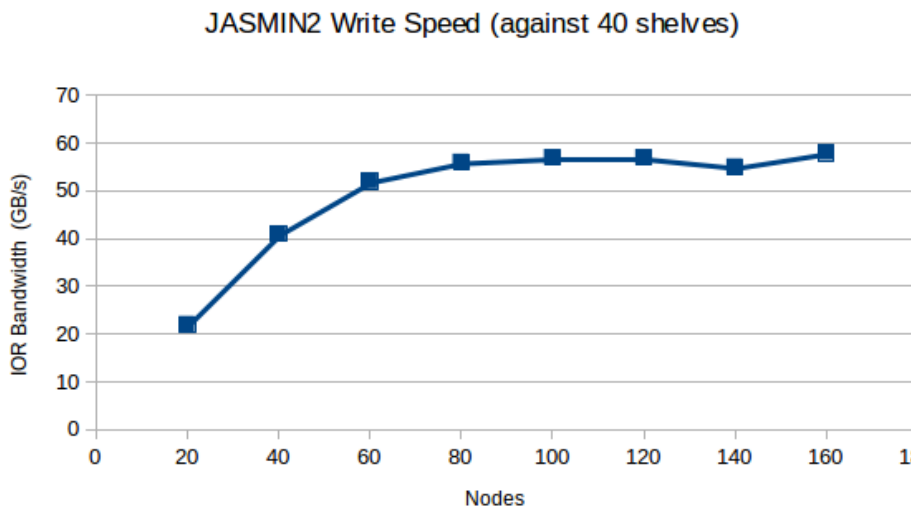
Figure 7
Throughput performance (IOR benchmark).

Performance and 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, contention, 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 $\approx$ 240 disks per bladeset. In JASMIN2 that's $\approx$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.





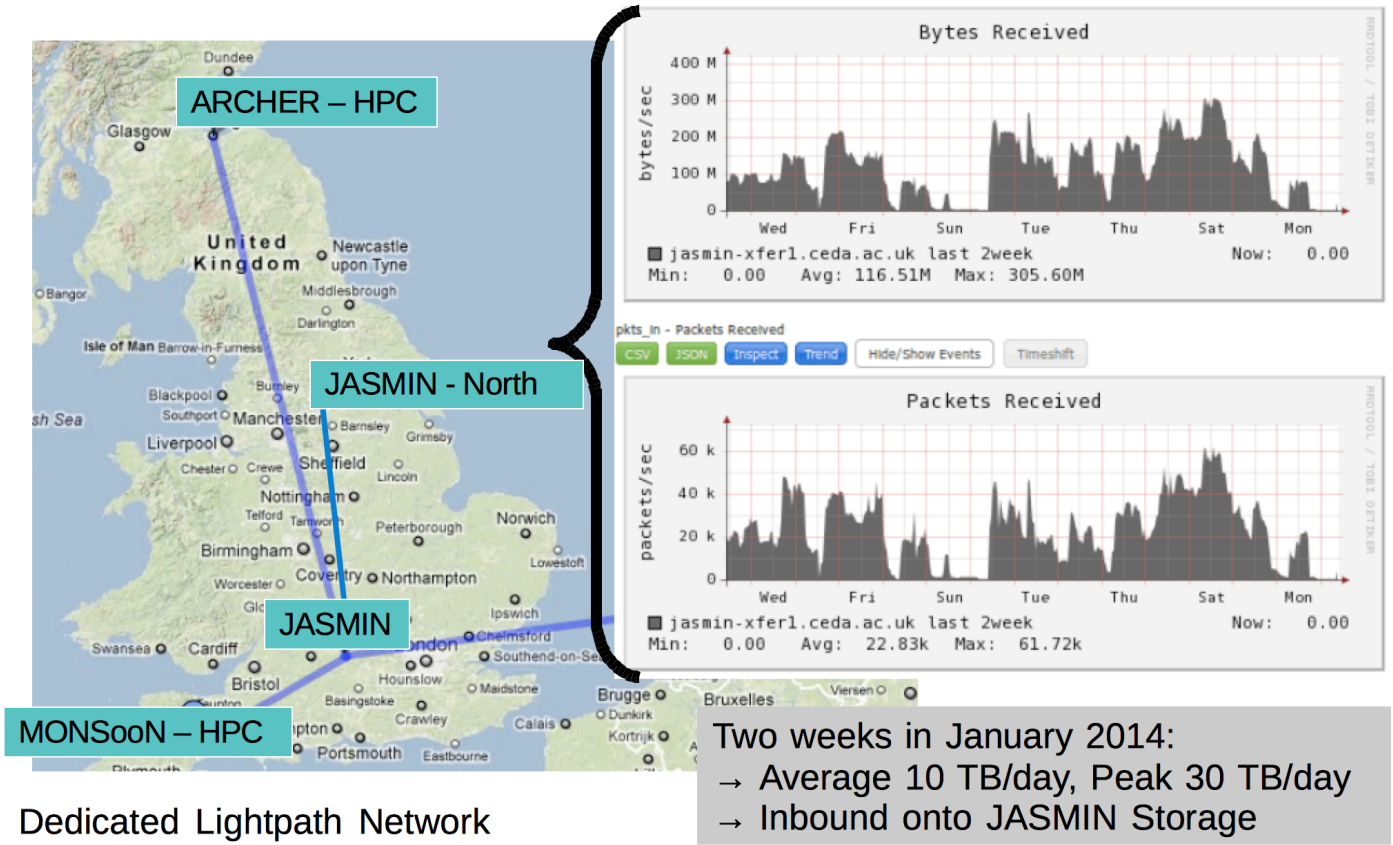
A subliminal message:

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?!

Our file systems are nowhere near keeping pace with our compute!

(Looking to future technologies ...)

Making use of the WAN bandwidth

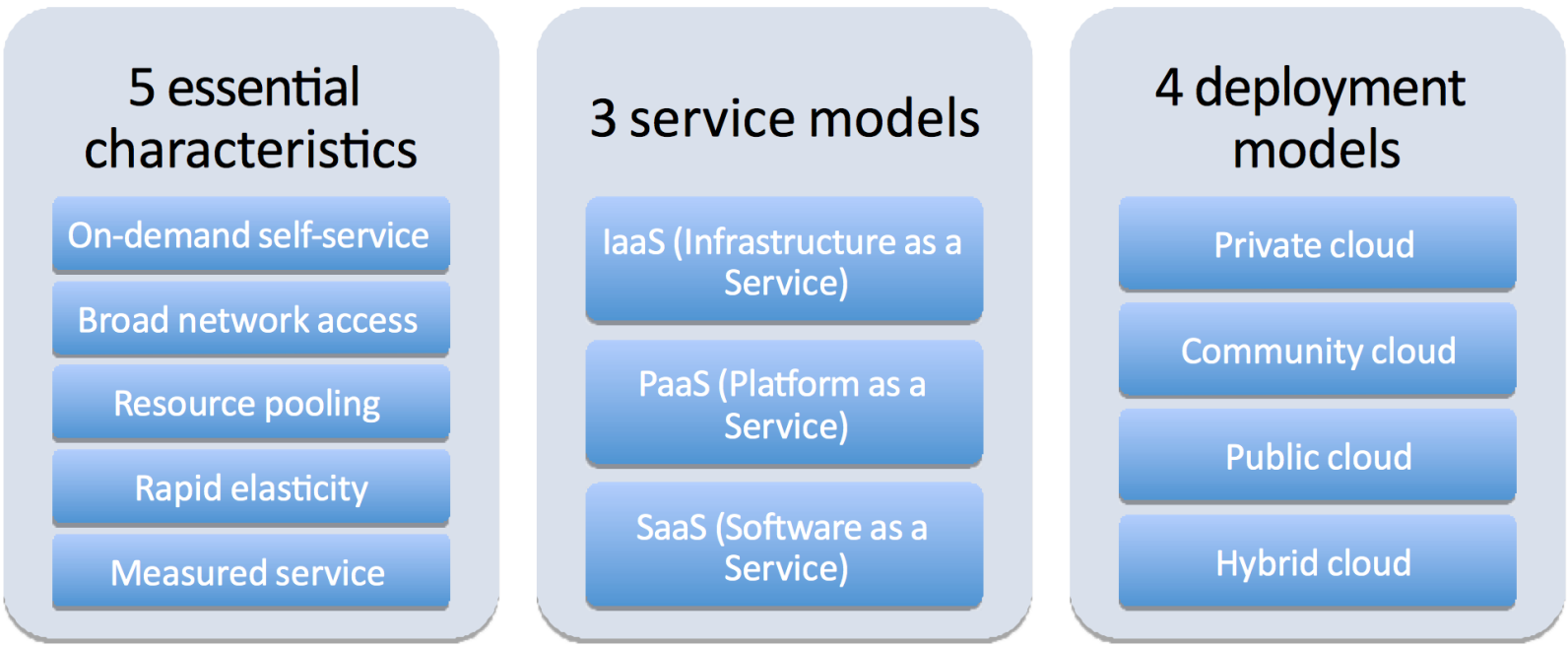


We've had some network upgrades since then. The bottom line is we expect (and see) TBs per day - to JASMIN at least.

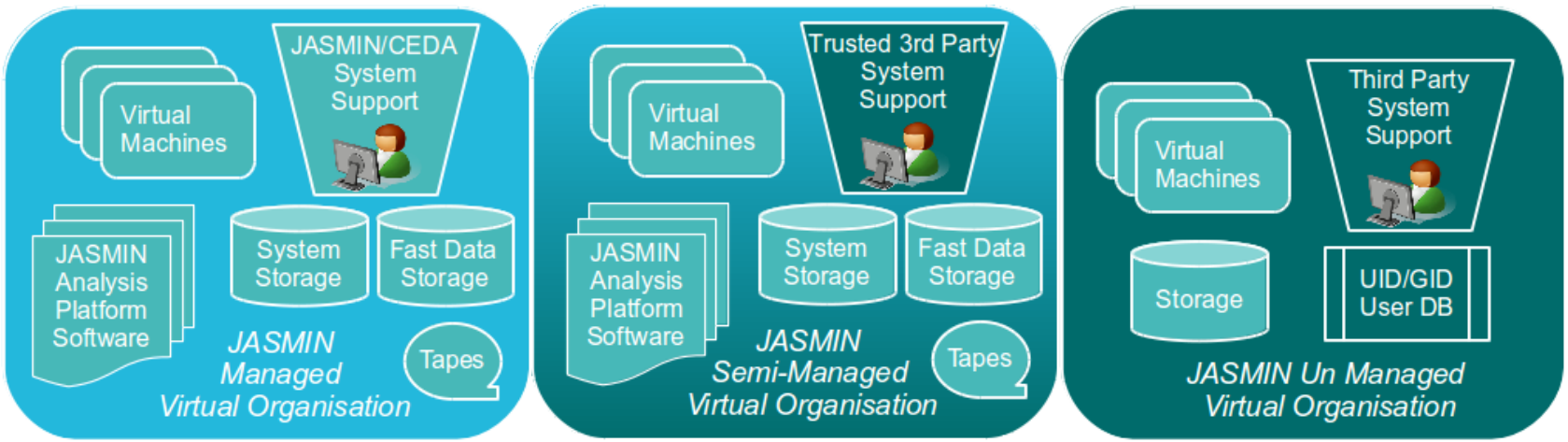
An introduction to the cloud

Why cloud? We're supporting individuals in a range of places who are used to a range of computing environments!

> *"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*



JASMIN Virtual Organisations



Platform as a Service $\longrightarrow$ Infrastructure as a Service

Some Special Virtual Organisations

CEDA: Centre for Environmental Data Archival

- Will provide archival services for the community.
- Data held in the archive will be managed, and made available to all the managed and semi-managed V.O.s directly (and indirectly to the un-managed V.O.s).
- Will provide "generic" access platforms for virtual organisations that do not wish to manage their own platforms and users who do not belong to specific virtual organisations.

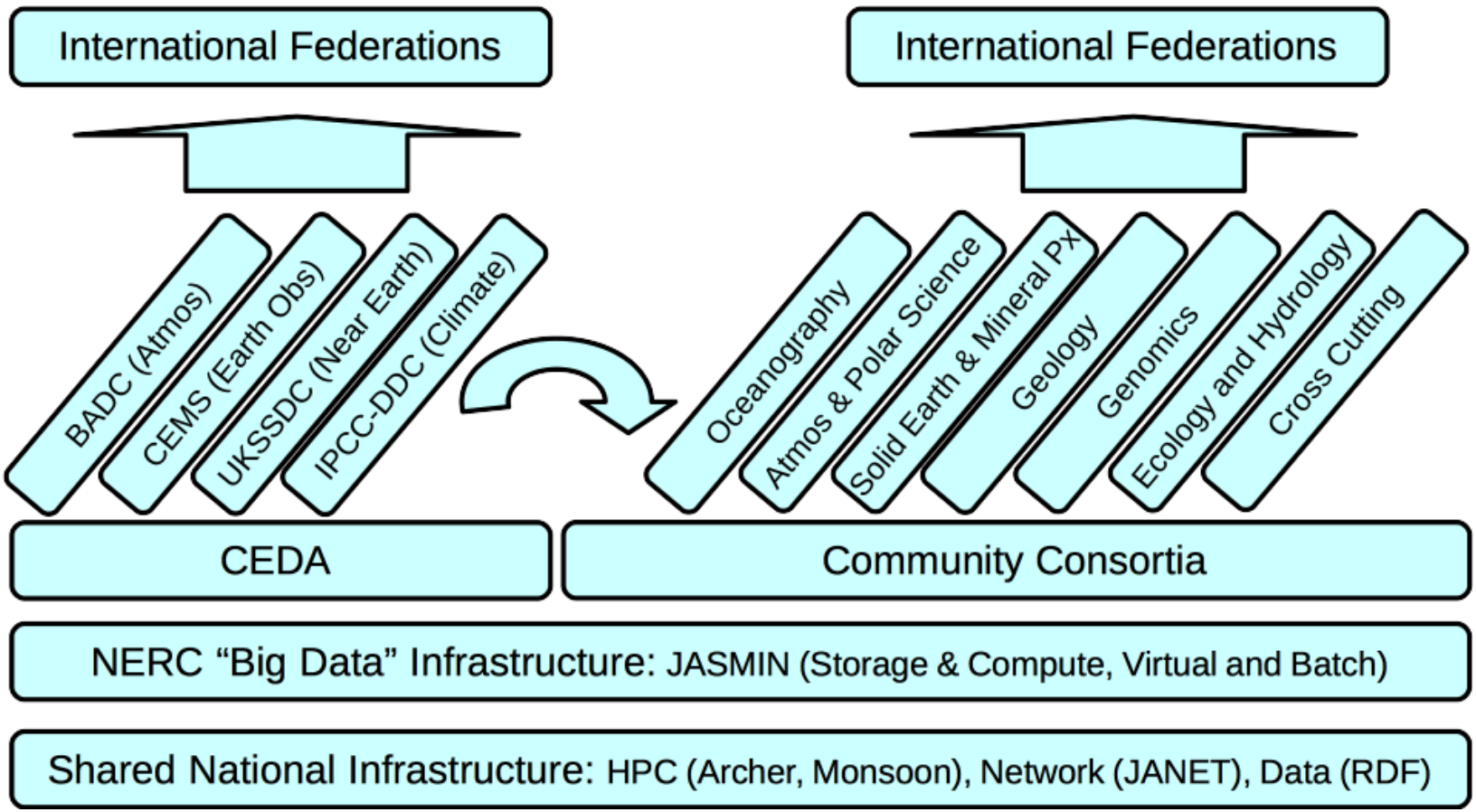
EOS Cloud

- Cloud services for the environmental 'omics community
- Delivered by JASMIN on behalf of the Centre for Ecology and Hydrology

CEMS: The facility for Climate, Environment and Monitoring from Space

- Will acquire and archive (via CEDA) key third party datasets needed by the NERC science community.
- Will provide services for the Earth Observation Community, in particular, in partnership with Satellite Applications catapult (SAC), the UK and European space industry.
- The academic component will run on JASMIN, the bulk of the industrial component, in the SAC, with access to CEDA data.

The "headline" virtual organisations



Environmental Research Workbench

Cloud-enabled User Appliances (applications/services)

A simple hydrological model | A data mash-up | JULES LSM | 3D animation | & lots more!!

Global AOGC model | A model ensemble | Multi-layered statistical integration | A data set

Data manipulation service | TopModel Monte'Carlo simulation for a catchment

Populate the ERW and applied from within it

The "Basic" Workbench

...sits atop the JASMIN Private cloud... a user friendly interface enabling the easy exploitation of the underlying resources "

But can also be applied directly to the private JASMIN-Cloud

Or any other cloud

Public Cloud

JASMIN Private Cloud

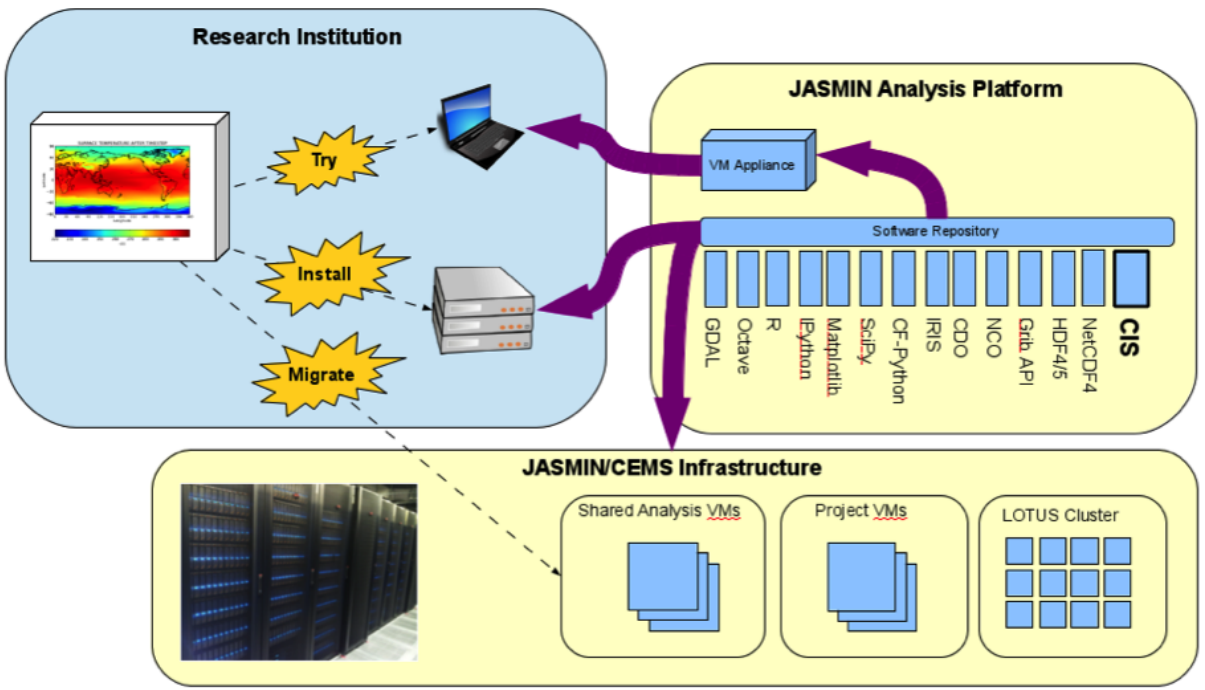
Cloud-bursting when required

Centre for Ecology & Hydrology
NATURAL ENVIRONMENT RESEARCH COUNCIL

NERC SCIENCE OF THE ENVIRONMENT

Platform as a Service: The JASMIN Analysis Platform

- Multi-node infrastructure requires a way to install tools quickly and consistently
- The community needs a consistent platform where ever they need them.
- Users need help migrating analysis to JASMIN.



http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/AnalysisPlatform

What JAP Provides

Standard Analysis Tools

- NetCDF4, HDF5, Grib
- Operators: NCO, CDO
- Python Stack: Numpy, SciPy, Matplotlib, IRIS, cf-python, cdat_lite, IPython
- GDAL, GEOS
- NCAR Graphics, NCL
- R, octave
- IDL (. . . but)
- . . .

Parallelisation and Workflow

- Python MPI bindings
- Jug (simple python task scheduling)
- IPython notebook
- IPython-parallel
- JASMIN Community Intercomparrison Suite

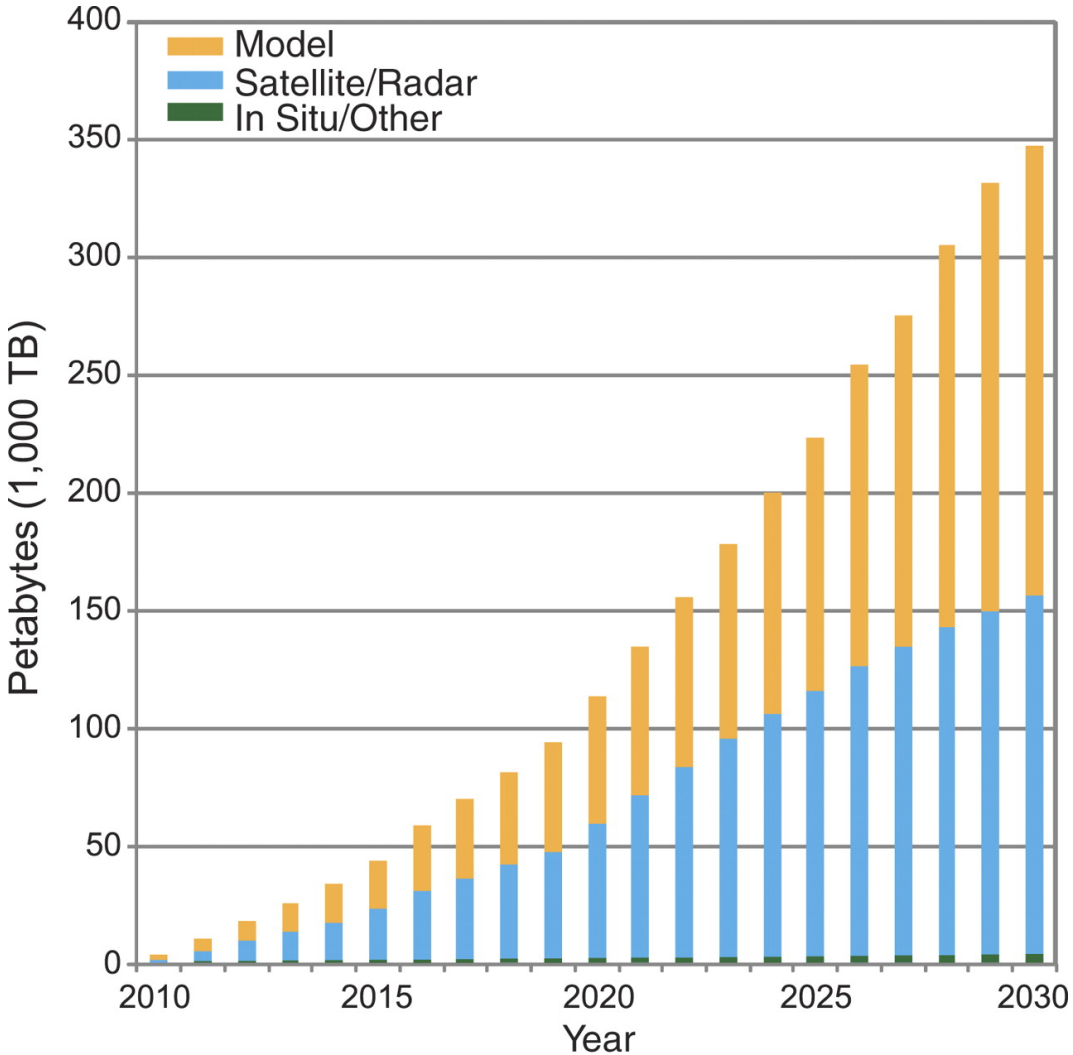
Science Codes

- JASMIN Community Intercomparrison Suite
- . . . soon: validation tooling (e.g ESMVal)

Global Data Archival

Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

(BNL: Even if you are?)



J T Overpeck et al. Science 2011;331:700-702

Causes of Data Growth - Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

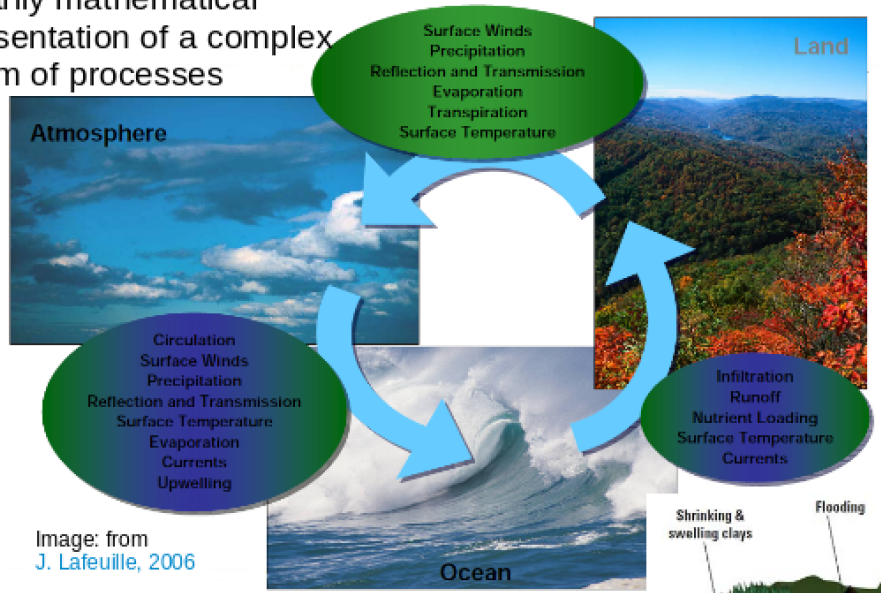
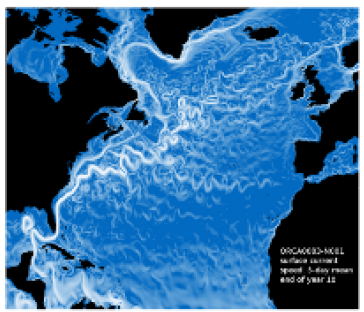
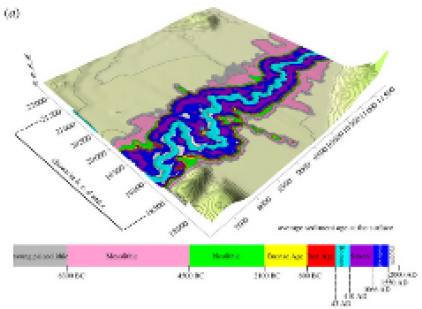
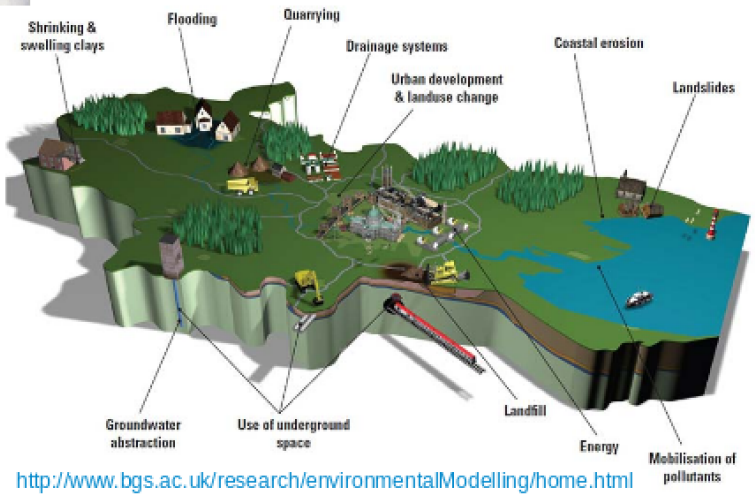


Image: from J. Lafeuille, 2006





Coulthard and Van De Wiel IDoi: 10.1098/rsta.2011.0597



http://www.bgs.ac.uk/research/environmentalModelling/home.html

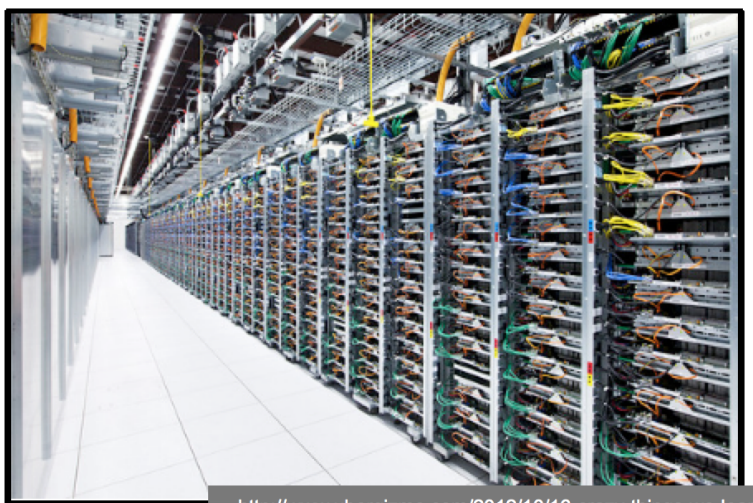
We want to observe and simulate the world at ever higher resolution! More complexity!

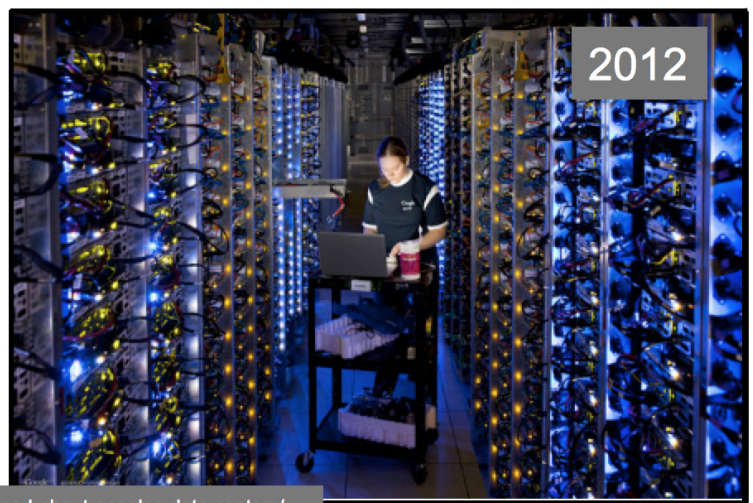
CEDA Evolution

Eerily similar to Google

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/

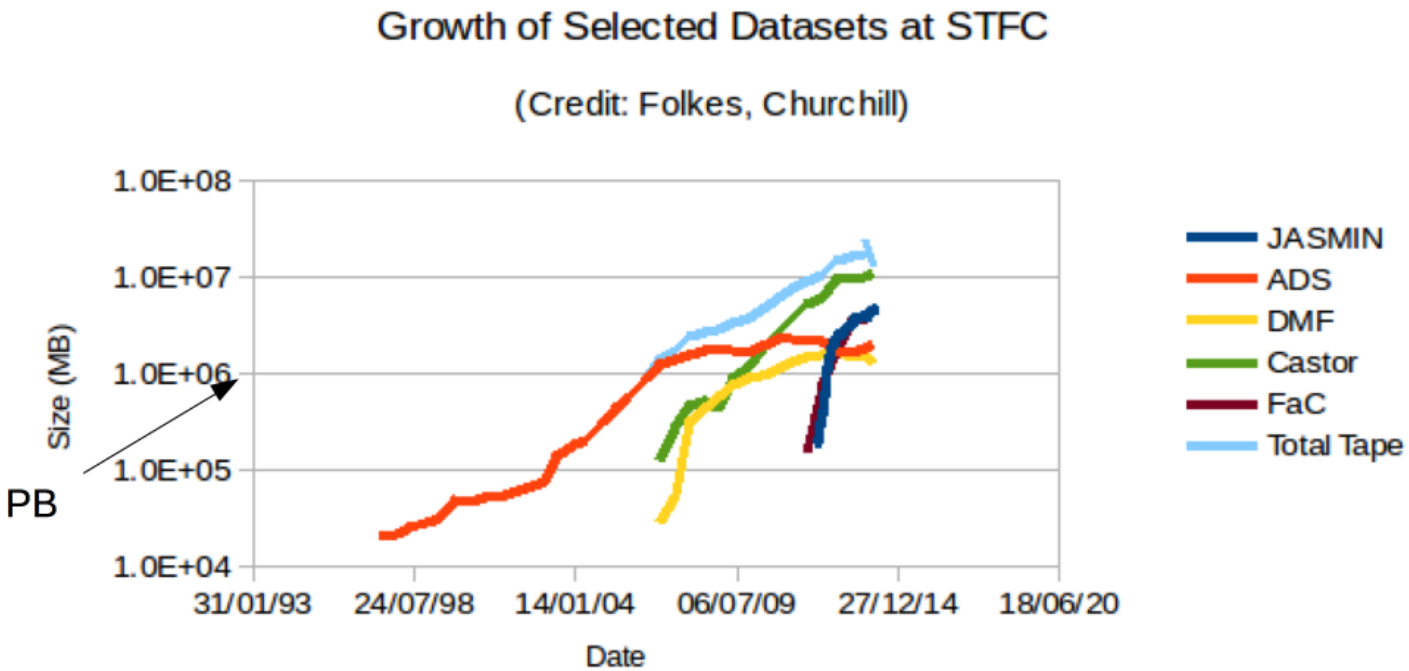
Not so subliminal message:

As we move to exascale storage, not everyone will be able to scale from a few machines to one (or more) massive machine rooms.

Actual subliminal message:

As well as hardware, one needs an awful lot of software to manage and exploit data at scale. Much of it will be bespoke!

STFC and CEDA



Predictions for JASMIN in 2020? 30 — 85 PB of unique data[1]!
But we think we could only fit only 30 PB disk in the physical space available!

([1] Not including CMIP6, which might be anything from 30-300 PB, but we hope at the lower end!)

Tape and Backup

At petascale we can't do automatic backup!

(We have users who can create a 100 TB dataset one day, and trash it the next because it wasn't quite right .... there is no sensible way to manage that automatically!)

Nearly every large site ends up building their own bespoke tape management system (e.g. Met Office/MASS, ECMWF/MARS, CERN/Castor).

We are providing the managed VOs access to an "elastic tape" service; "elastic" in the cloud sense, a VO can keep adding tape beyond what we allocate them if they want to spend their own money!

- Layered on the CASTOR tape service run at STFC.
- VO managers can read and write data without knowing about the tape system, they simply get a job number to go with a list of files, and can retrieve the list of files at a later date.
- There is much to do ... including working out a solution for the un-managed cloud!

Exploiting Parallel Data Analysis

University of Reading

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

STFC Scientific Computing Department

CRAY

PRACE

Met Office

NATURAL ENVIRONMENT RESEARCH COUNCIL

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Joint Weather and Climate Research Programme
A partnership in climate research

UPSCALE and JASMIN

High resolution climate modelling supported by a super-data cluster

M. S. Mizielinski, **P. L. Vidale [PI] , M. J. Roberts, R. Schiemann, M.-E. Demory and J. Strachan**

Supported by T. Edwards, A. Stephens, B. N. Lawrence, M. Pritchard, P. Chiu, A. Iwi, J. Churchill, C. del Cano Novales, J. Kettleborough, W. Roseblade, P. Selwood, M. Foster, M. Glover, and A. Malcolm

UPSCALE

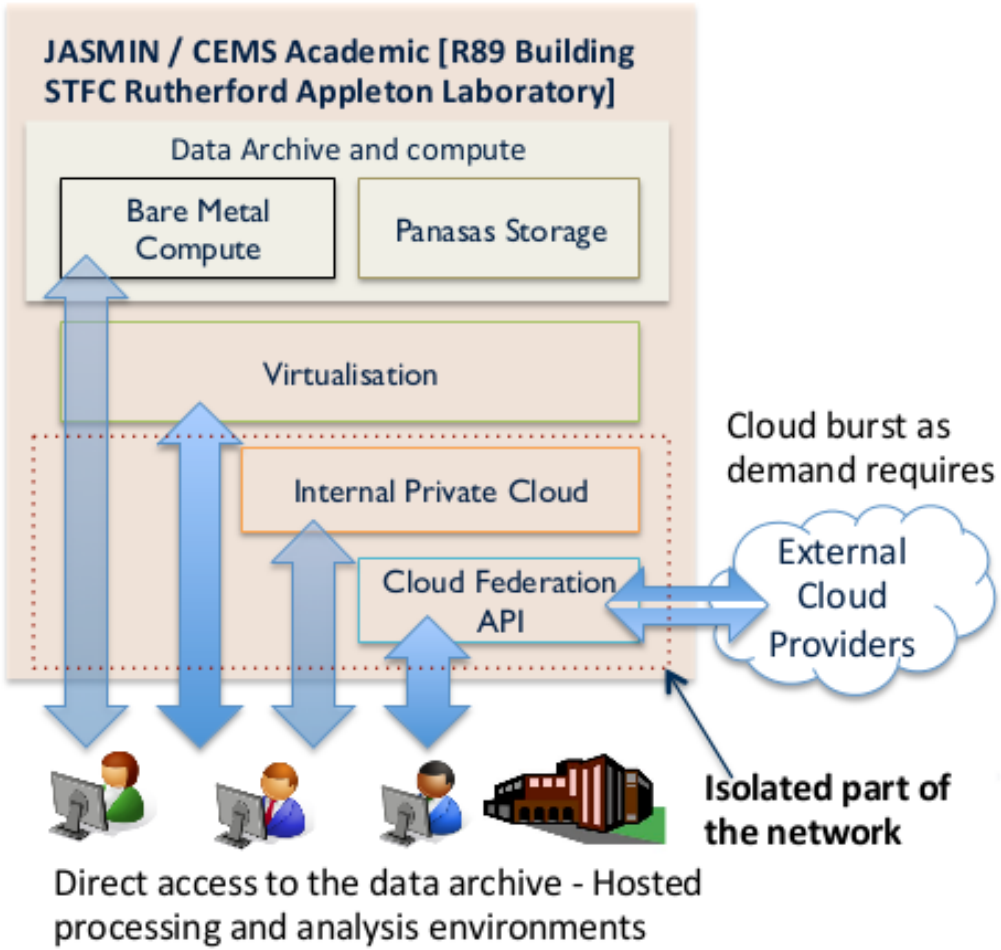
UPSCALE: **U**K on **P**RACE — weather resolving **S**imulations of **C**limate for glob**AL** **E**nvironmental risk

- Ensembles of global atmospheric climate simulations at weather forecasting resolution
- Required more than 30 times the computing time available to our team on UK supercomputer HECToR
- Successfully applied for a 144 million core hour from PRACE lasting for 1 year on HERMIT in Germany
- Produced more than 400 TB of data over 10 months, which was shipped to JASMIN and the Met Office archives

One example of dozens of ongoing projects analysing the data:

- The influence of atmospheric eddies on the north Atlantic storm track, and jet stream, can be investigated by computing "eddy vectors" (L. Novak, M. Ambaum, R. Tailleux, University of Reading) from wind and temperature data
- Analysis across UPSCALE data set uses 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
- JASMIN and LOTUS help the team work around technical challenges, leaving them to focus on the science

The cloud view

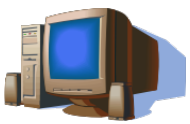


Managed Services

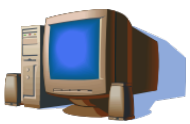
- Group Workspaces (large disk, 2—500 TB)
- Generic Virtual Machines (transfer, login, analysis)
- Dedicated Virtual Machines (for projects/groups)
- Common software environment
- LOTUS compute cluster
- High Performance Data Transfer server
- "Elastic" tape — backups
- Connections to key sites (lightpaths)

Managed and Semi-VMs

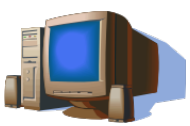
jasmin-login1.ceda.ac.uk; acts as a gateway to other JASMIN nodes; only one; no functionality.



jasmin-xer1.ceda.ac.uk;for copying data in/out; currently SCP & RSYNC; GridFTP; read-write to GWS.



jasmin-sci1,2.ceda.ac.uk; for general scientific analysis; common software build; access to GWSs and archive.



XXX.ceda.ac.uk; requested by specific projects/users; ROOT access for trusted partners; read-write access to GWS.

NOTE: CEMS equivalents also exist of these VM types...but they are fundamentally the same.

CEMS
Climate, Environment & Monitoring from Space

NERC consortia

JASMIN resources will be allocated in two steps:

- The broad distribution of resources will (eventually) be controlled by the NERC HPC committee, who will govern the distribution between the following seven consortia:

  Atmospheric and Polar Sciences, Solid Earth and Mineral Physics, Oceans and Shelf Seas, Geology, Genomics, Ecology and Hydrology, and Earth Observation;

  along with a director's allocation to support strategic and development projects.
- Each consortium will have a manager who can allocate resources within their overall quote (and perhaps negotiate borrowing resources from other consortia).

(It is likely that the bulk of the resource will remain in support of Earth Observation and Climate Science ... writ large, covering much of NERC ... but this still means there will be petascale storage available on JASMIN for other disciplines!)

Resources in the un-managed cloud

Responsibilities of CEDA/STFC:

- Providing and operating the JASMIN Unmanaged Cloud platform.
- Supporting resource allocation by NERC consortia managers - who themselves get their allocations from the NERC HPC committee.
- Providing initial support for setup of a virtual organisation - but thereafter what happens inside the VO is entirely their business.
- We are looking into providing access to tape media from the un-managed VOs, but this is non-trivial ...
- We are working on the ability to cloud-burst into the commercial cloud if compute resources on JASMIN are inadequate.
- *All of this is dependent on core funding from NERC direct to STFC which has yet to be confirmed after this year.*

Responsibilities of tenant organisations:

- To work with their consortia managers to obtain appropriate maximum resource limits (storage, compute etc),
- Providing their own staff to manage and run whatever services it chooses commensurate with the available resource, in particular, to
  1. Provide a named person as the sysadmin point of contact,
  2. Manage their own platform as if it were physically present on their site, including being fully responsibility for the information security of their systems. (These systems will not be inside the RAL firewall.)

It will not have escaped your attention that the effective use of much of the JASMIN resource will depend on the people in this room ...

Summary

- JASMIN is a very large computer system, optimised for data storage and analysis.
- JASMIN is configured for high performance usage by many different segments of the community - utilising their own compute environments, but still getting the benefit of high performance.
- (Not really covered: JASMIN is already changing the nature of some of the science problems we can confront ... a "game-changer" for earth observation!)
- Effective usage of JASMIN will depend on the existing NERC IT community — JASMIN can only enable more science if the IT environments remain familiar and exploitable!

More information

- JASMIN Documentation:
  - http://www.jasmin.ac.uk
- JASMIN Documentation - services:
  - http://www.jasmin.ac.uk/services/
- Scientific Context:
  - Storing and manipulating environmental big data with JASMIN. Lawrence et.al., *2013 IEEE conference on big data.* 10.1109/BigData.2013.6691556.