Infrastructure for Environmental Supercomputing: beyond the HPC!

Bryan N Lawrence

Professor of Weather and Climate Computing, University of Reading
Director of Models and Data, NCAS
Director of the Centre for Environmental Data Archival, STFC

University of Reading
NERC SCIENCE OF THE ENVIRONMENT
Centre for Environmental Data Archival
Science and Technology Facilities Council
Natural Environment Research Council

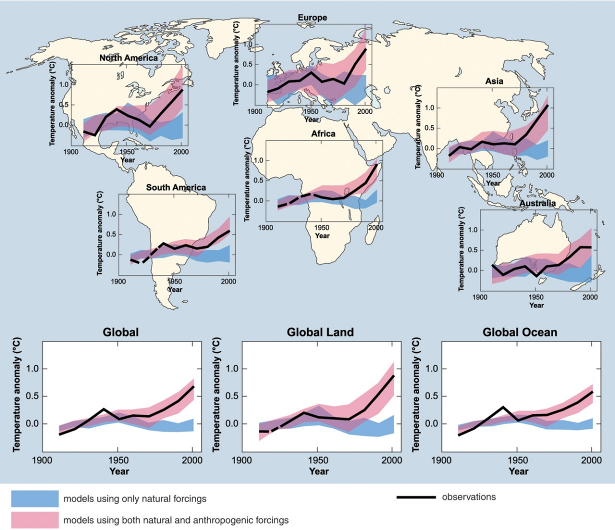
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

Outline

- Motivation
- Drivers
- Background Trends
- Collaboration
- JASMIN
- Summary

From the Large

Fig 2.5
AR4
Synthesis
Report



To the Small

How will climate change affect the global distribution of malaria?

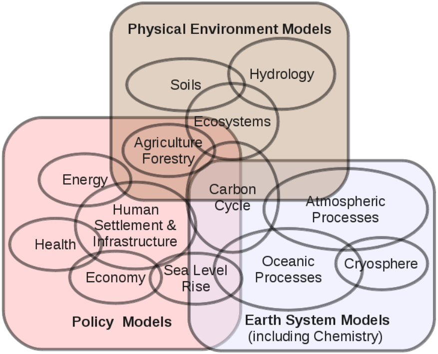
July 2007 Tewkesbury flood: 3B€ loss!

Can we predict risk into the future?

What would be the impact of leakage from an oil and gas well in UK waters on the national economy, coastal and marine biodiversity and the well-being of the population affected?

How will climate change affect the incidence of road and rail closures due to landslides?

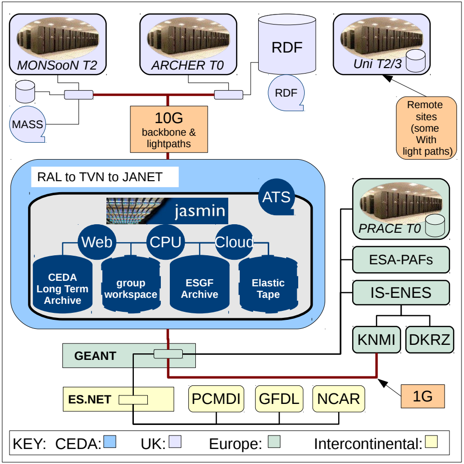
Communities



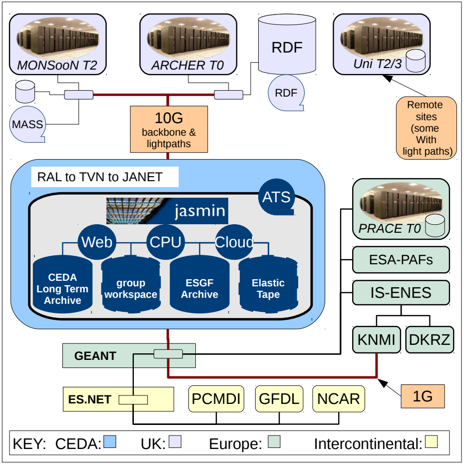
Many interacting communities, each with their own software, compute environments etc.

Figure adapted from Moss et al, 2010

Infrastructure



Infrastructure



- The network view is the easy view!
- What are the data policies? What are the (possible) data residence times?
- What agreements are in place?
- What can we rely on in this picture? For example, who has to agree to upgrade something (a network link for example)?
- How do **community** science drivers/requirements lead to infrastructure provision.

Sharing

Science across scales

Lots of interacting communities

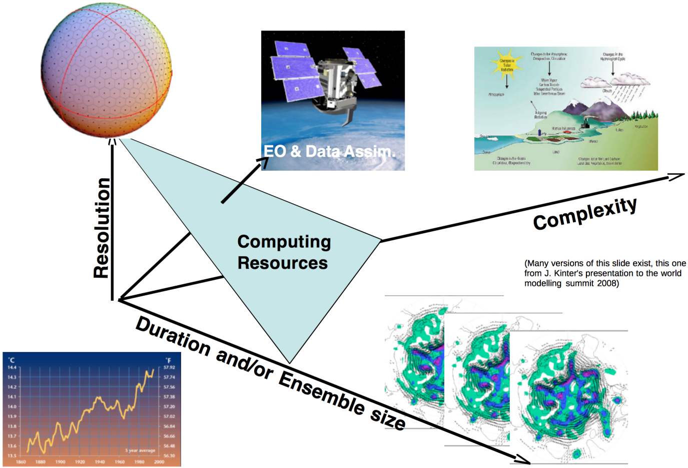
Lots of infrastructure

Can we share infrastructure?

Between communities?

Between nations?

Give me more computing? Global Climate Modelling



Give me more computing - Direct Numerical Simulation

Primarily mathematical representation of a complex system of processes

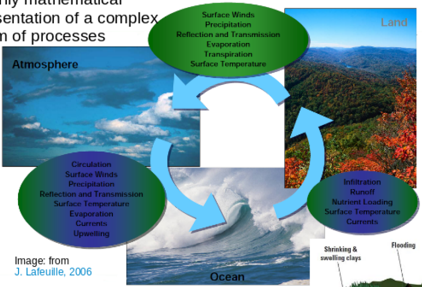
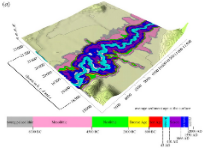
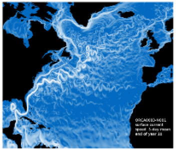
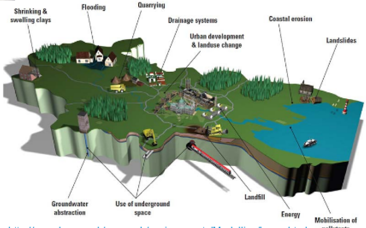


Image: from J. Lafeuille, 2006



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



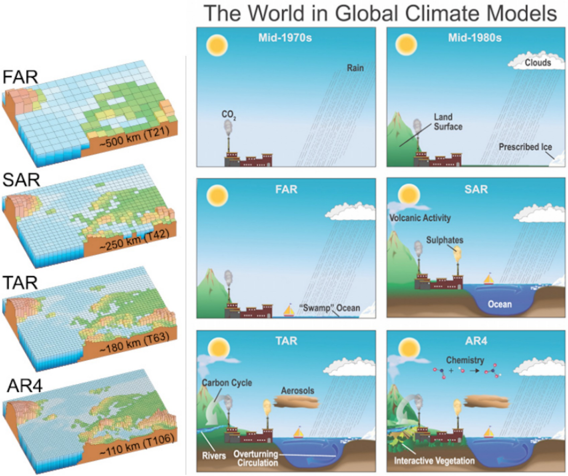


http://www.bgs.ac.uk/research/environmentalModelling/home.html

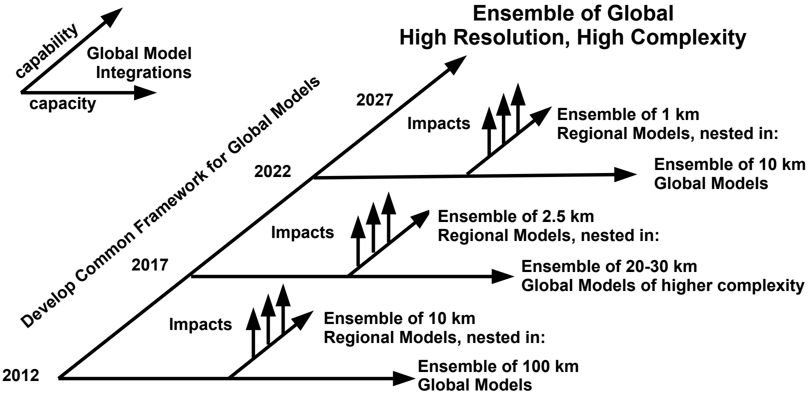
We want to observe and simulate the world at ever higher resolution! More complexity!

Give me more computing? How this has gone

FAR:1990
SAR:1995
TAR:2001
AR4:2007
AR5:2013



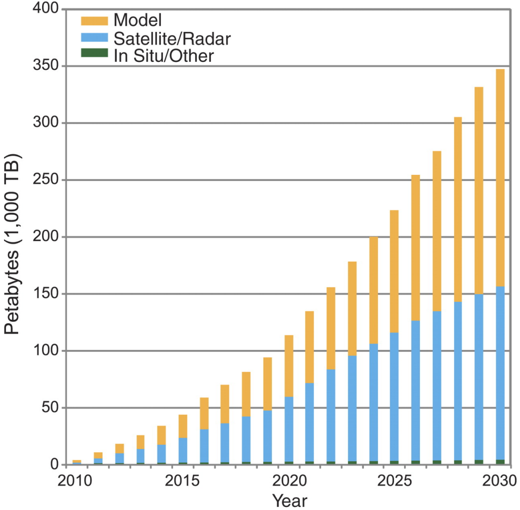
Give me more computing? Where is this going



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.

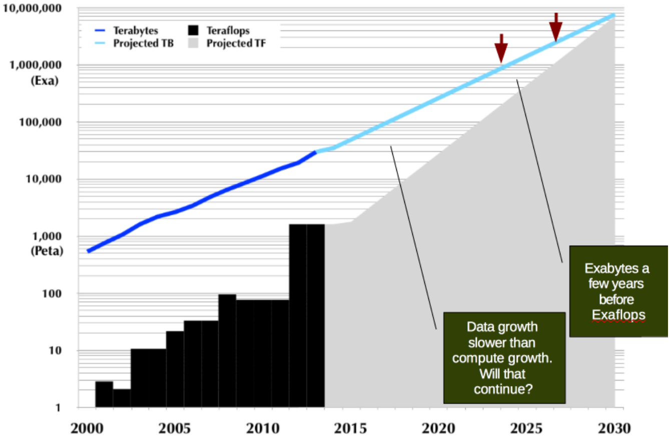
(BNL: Even if you are?)



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

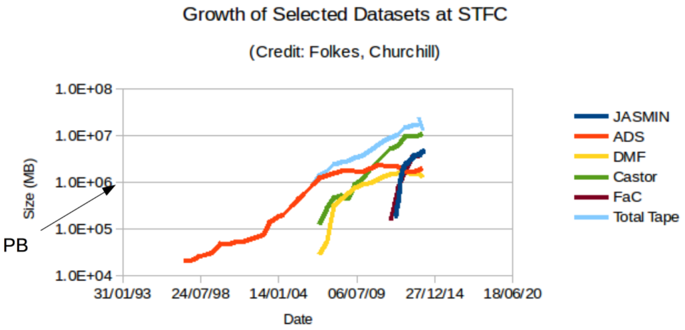
Institutional - NCAR

Storage, and power for storage, will dominate NCAR's compute budget within a few years! (Rich Loft, 2014).



NCAR Compute and Data (courtesy Gary Strand)

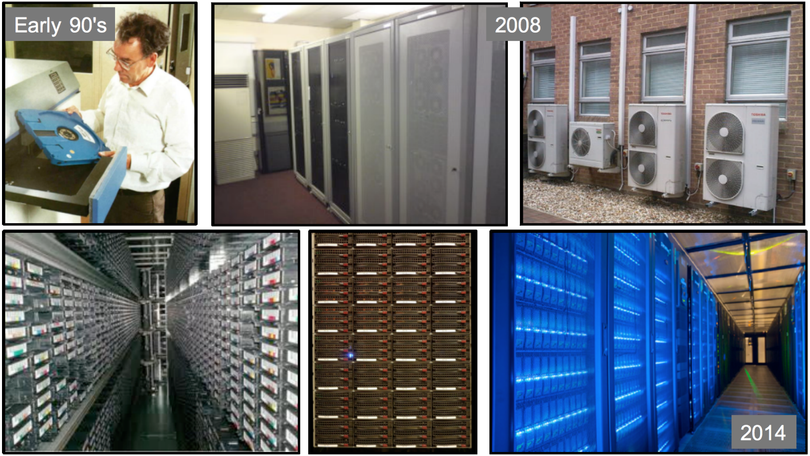
Institutional - STFC and CEDA



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[2]!

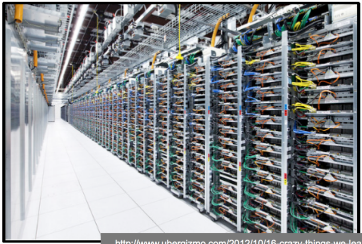
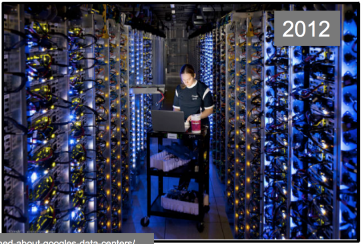
([1] Not including CMIP6, which might be anything from 30 PB up. [2] Unless we can throw out the CERC Tier1 centre with whom we share!)

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/

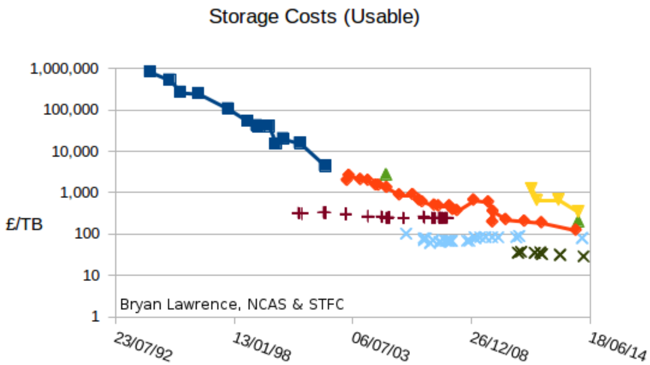
Not so subliminal message:

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!

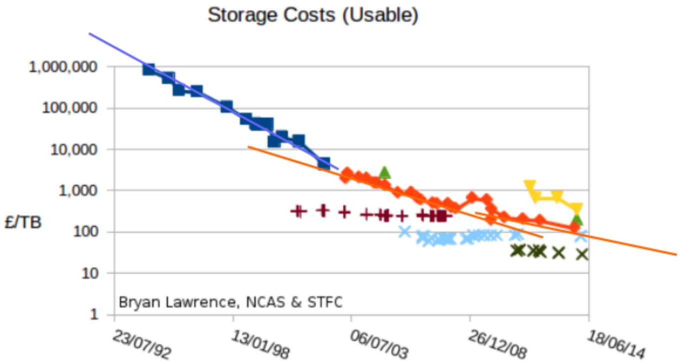
Kryder's Law



Solid objects: colours are different generations of disk. Crosses: different generations of tape.

(Data from Peter Chiu, Jonathan Churchill and Tim Folkes, STFC)

Kryder's Law



Solid objects: colours are different generations of disk. Crosses: different generations of tape.

Kryder's Law definitely slowing down! Plenty of mileage still in tape though!

U.S. National Academy

"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, 2012

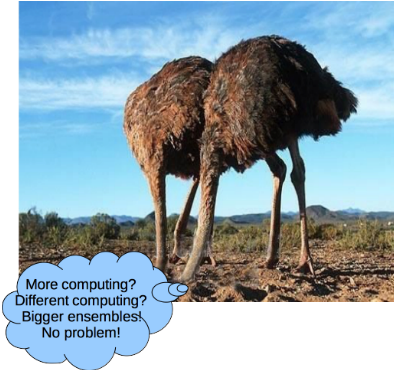
Semantic Analysis: "substantial research effort" "new methods" "computation to data" "rather than trying to download" "frustratingly inaccessible" (to whom?)

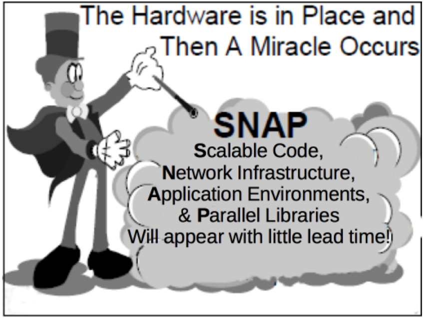
What about software?

According to Ken Batcher, “A supercomputer is a device for turning compute bound problems into I/O bound problems.”

What about software?

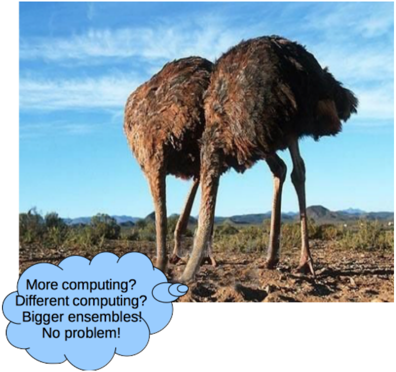
According to Ken Batcher, "A supercomputer is a device for turning compute bound problems into I/O bound problems."

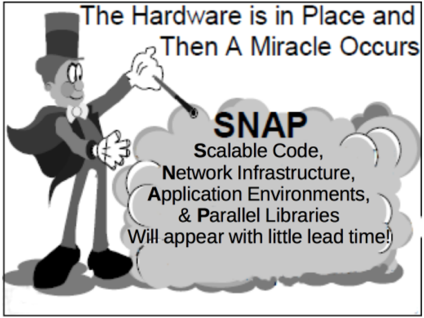




What about software?

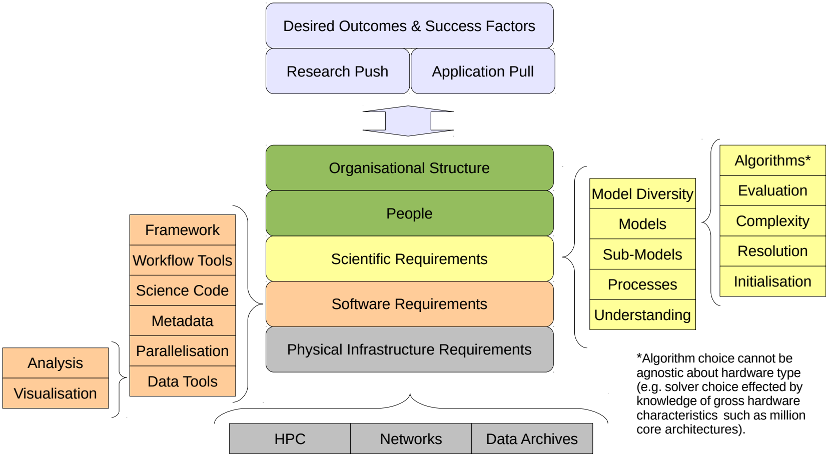
According to Ken Batcher, "A supercomputer is a device for turning compute bound problems into I/O bound problems."





. . . which is a little unfair, but I think it is fair to say that (some of) the community underestimates the effort ahead!

Putting it all together



*Algorithm choice cannot be agnostic about hardware type (e.g. solver choice effected by knowledge of gross hardware characteristics such as million core architectures).

Summary so far

The technology drivers are tending towards infinitely cheap computing and infinitely expensive data systems!

(?tending?: tending, I just said tending, nothing ever asymptotes ok!)

However, while the computing might be (relatively) cheap, exploiting it is likely to become harder and harder

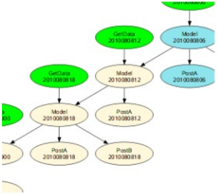
The solution involves collaboration . . .

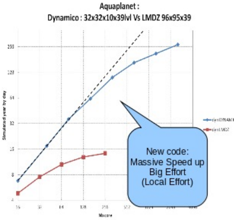
Better Software - 1

Four areas to consider:

- Workflow (e.g. CYLC)
- Simulation (The codes themselves)
- Analysis (CDO, NCO, IRIS, CF-Python etc)
- Data Management (I/O libraries, Tools to document data)

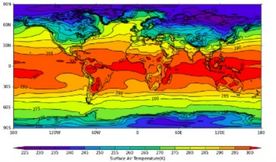
Software Is Infrastructure!





cfplot homepage

cfplot is a set of Python routines for making the common contour and vector plots that climate researchers use. The data to make a contour plot can be passed to cfplot using cf-python as per the following example.




```
import cf, cfplot as cfp
f=cf.read('/opt/graphics/cfplot_data/tas_A1.nc')[0]
cfp.con(f.subspace(time=15))
```

NIWA
Taihoro Nukurangi

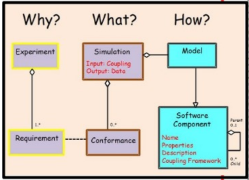
T.Dubos, S.Dubesh, Yann Meurdesoif(LSCE-IPSL)
Results presented at IS-ENES2 workshop, March 2014

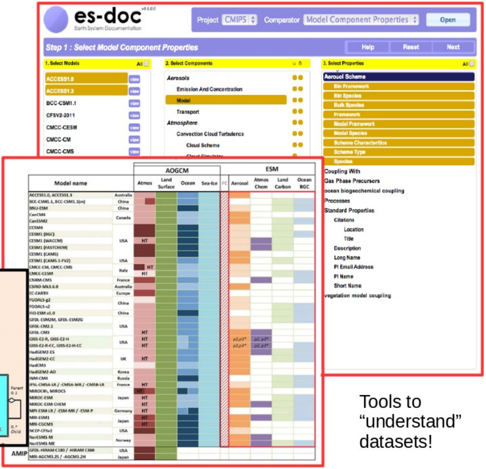
I have deliberately chosen Kiwi, French and British examples: Global activities!

(Major European initiatives - IS-ENES1 and IS-ENES2 ...)

Better Software - 2

CMIP5 (23/05/13):
- 101 experiments
- 61 model variants
- 590,000 datasets!
- 4.5 million files
- 2 PB in global archive
- Unknown PB locally!



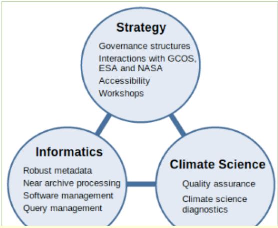


Tools to "understand" datasets!

(Major global initiatives - esdoc!)

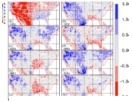
The ExArch Project - Taking compute to the data!

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

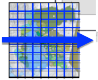


Martin Juckes, V. Balaji, B.N. Lawrence,
M. Lautenschlager, S. Denvil, G. Aloisio, P. Kushner, D. Waliser,
S. Pascoe, A. Stephens, P. Kershaw, F. Laliberte, J. Kim, S. Fiore

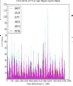
Regional Climate Model Evaluation System (RCMES)



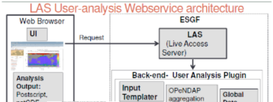
Observation/Model rainfall



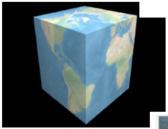
Map over a basin using an area-matching method



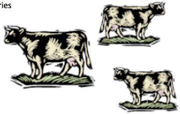
Get basin-mean time series



NOAA – PMEL "Live Access Server"



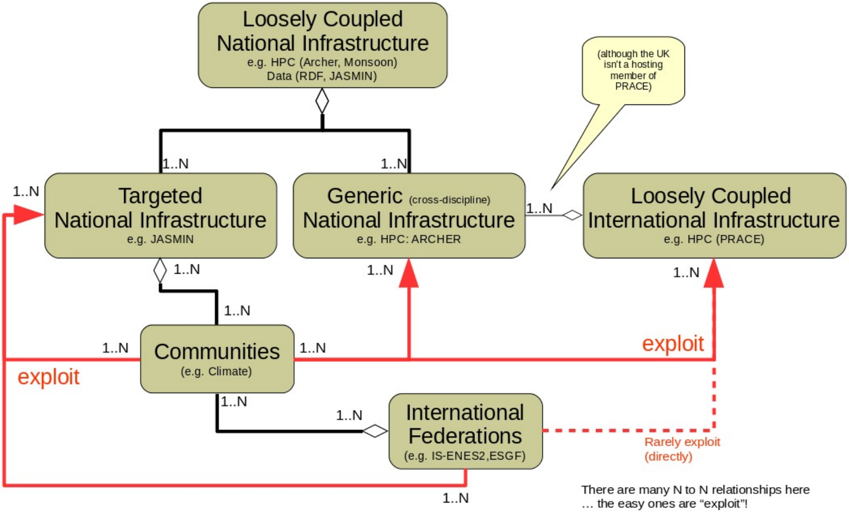
CMCC parallel data analytics framework



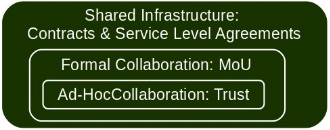
CEDA OGC Web Services

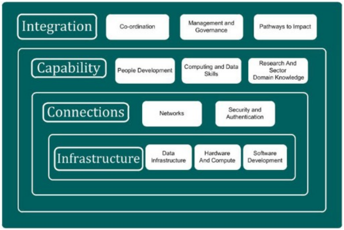
http://climate4impact.eu/

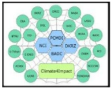
Infrastructure Relationships



Infrastructure and Agreements

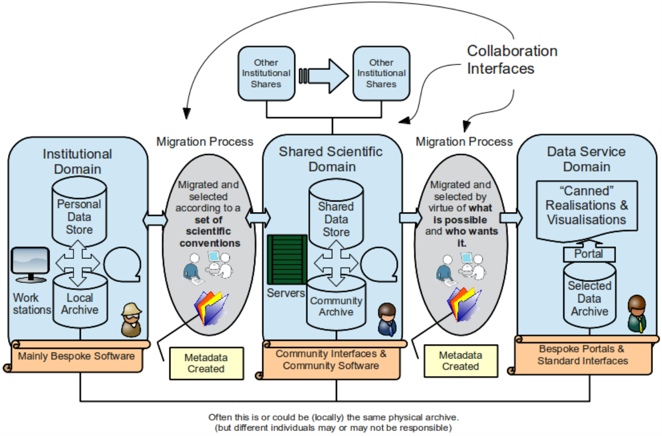






How do we progress from here?

Collaboration Interfaces

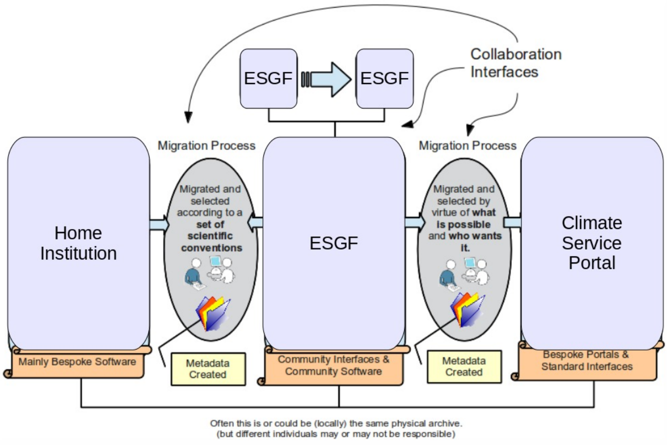


Start to understand the important interfaces!

(Already simplified because we have taken out the generation of the data!)

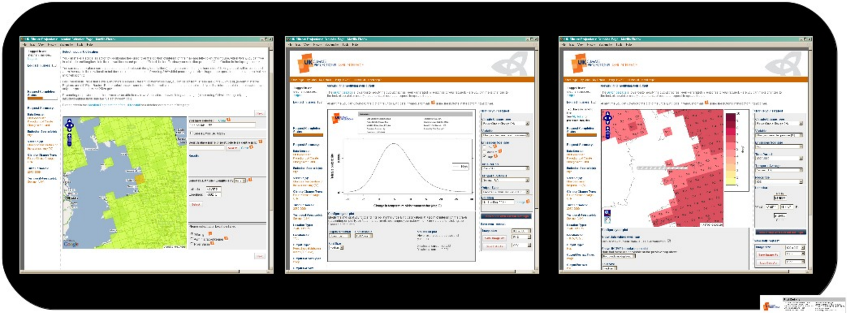
Collaboration Interfaces

Consider three cases: institutional, federated, and served domains!

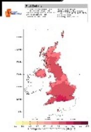


Customised Portals, e.g: UKCIP

... with dedicated hardware:

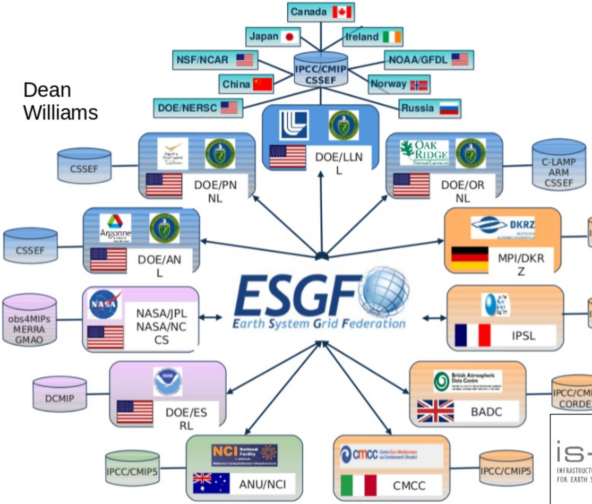


UK Climate Projections: Sophisticated user interface with optimised hardware, to support hundreds of simultaneous users dynamically interacting with data.



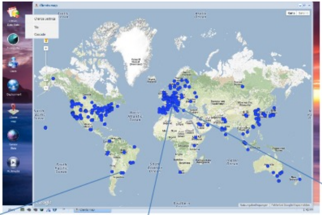
ESGF

Dean Williams



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

The trend



Slide courtesy of Stefan Kindermann, DKRZ and IS-ENES2

Individual End Users

- Limited resources (bandwidth, storage,..)

Organized User Groups

- Organize a local cache of required files
- Most of group don't access ESGF, use cache instead!

Data Centre Service Group

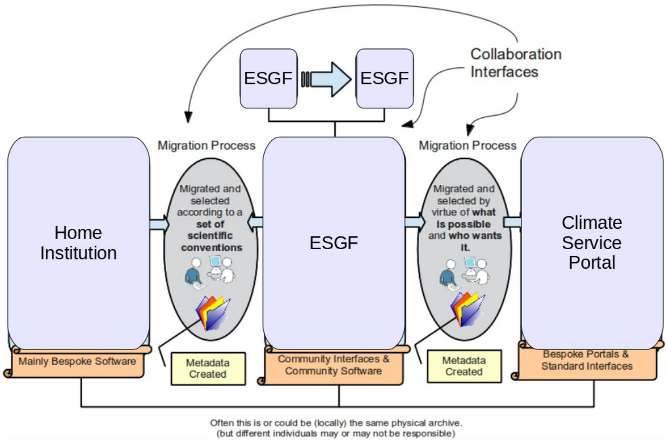
- Provides access to ESGF replica cache
- May also provide access to data near compute resources
- (BADC, DKRZ, IPSL, KNMI, UC)

Trend

Needed: Replacement for „*Download and Process at Home*“ Approach

(Reminder) Collaboration Interfaces

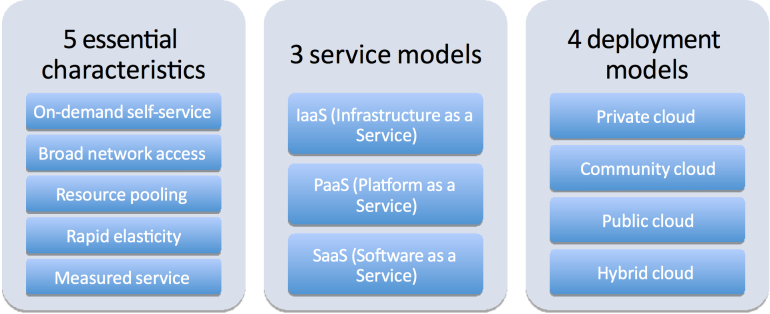
Consider three cases: institutional, federated, and served domains!



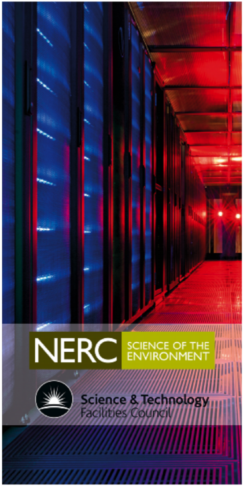
An introduction to the cloud

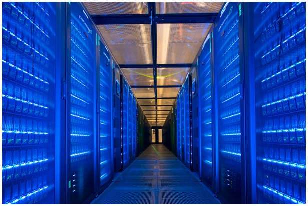
Why cloud? Remember all this communities, with their own software environments?

> *"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 management effort or service provider interaction."* — *NIST SP800-145*



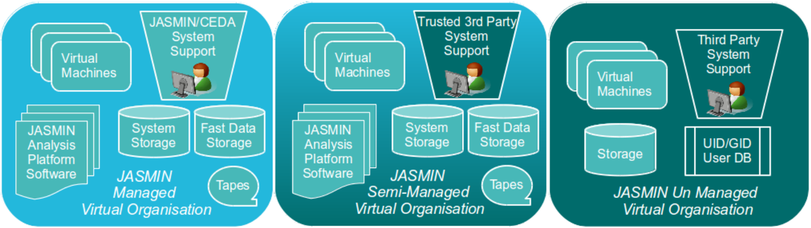
So we have built an "HPC-data" cloud: JASMIN





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

Virtual Organisations



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

NCAS itself will run a semi-managed virtual organisation (with multiple group work spaces), but large groups within NCAS can themselves also run virtual organisations.

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 ...

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

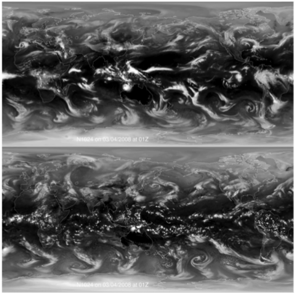
- Will acquire and archive (via CEDA) key third party datasets needed by the NERC science community.
- Will provide services for the Earth Observation Community, in particular, in partnership with Satellite Applications catapult (SAC), the UK and European space industry.
- The academic component will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

UPSCALE

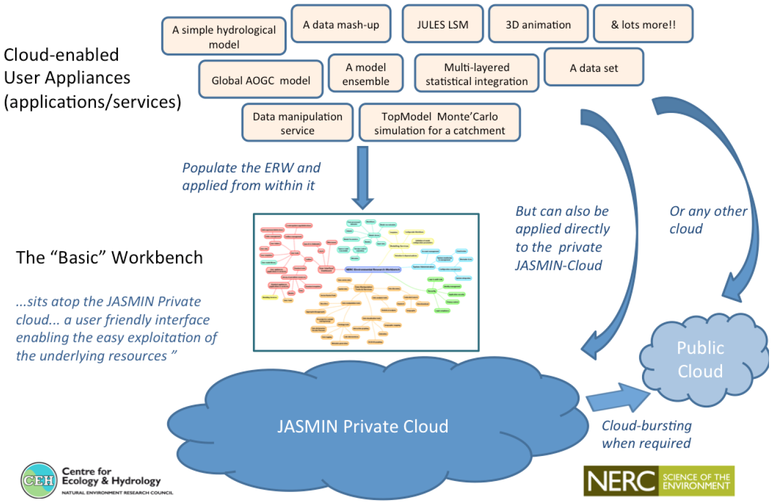
UPSCALE: **U**K on **P**RACE — weather resolving **S**imulations of **C**limate for glob**AL** **E**nvironmental risk.

- Ensembles of global atmospheric climate simulations at weather forecasting resolution.
- Used a one-year 144 million core-hour PRACE allocation on HERMIT (1 PFlop Cray XE6, typically running with up to 50K/115K cores).
- Produced more than 400 TB of data over 10 months, shipped to JASMIN.
- UPSCALE GWS accessed via two VMs: one managed by the met office, one by NERC, with 25 & 33 users respectively — a total of 50 unique GWS users (11/2014).

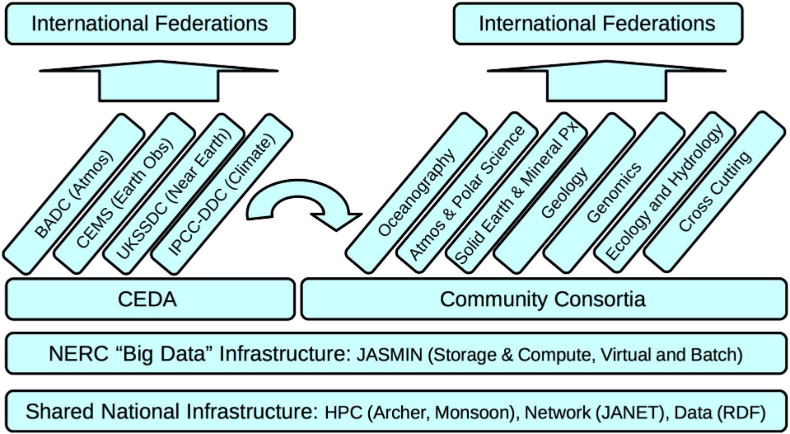
(Vidale/Roberts - NCAS/Met-Office)



Environmental Research Workbench

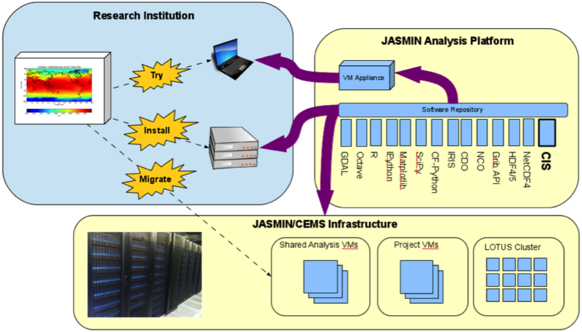


The "headline" virtual organisations



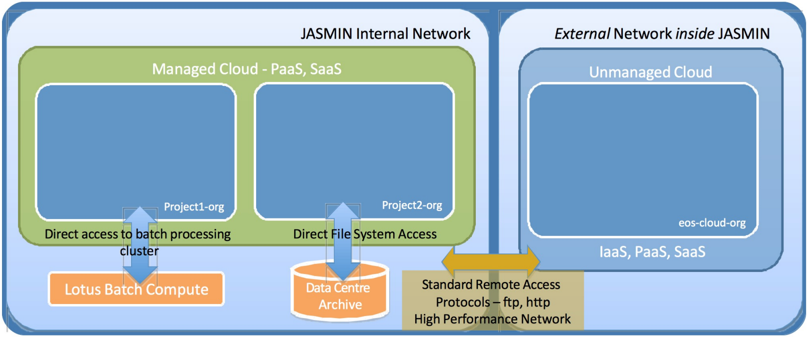
Platform as a Service: The JASMIN Analysis Platform

- 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.

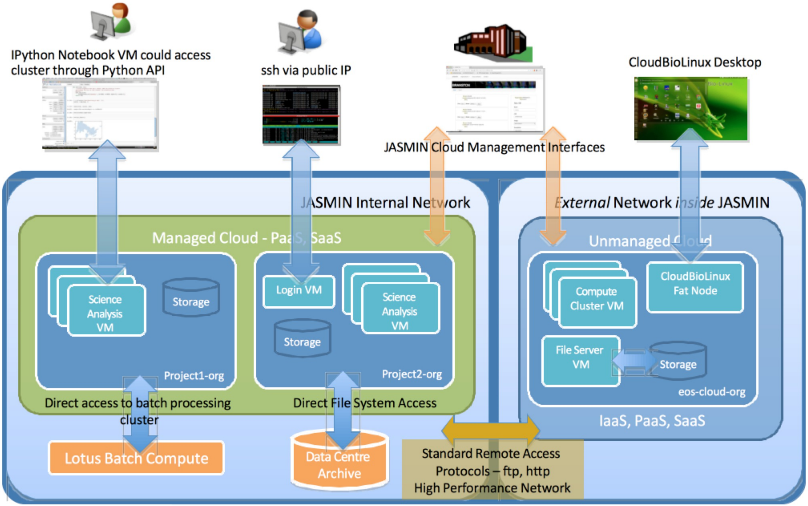


http://proj.badc.rl.ac.uk/cedaservices/wiki/JASMIN/AnalysisPlatform

Integrated Cloud Provisioning



Integrated Cloud Provisioning



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

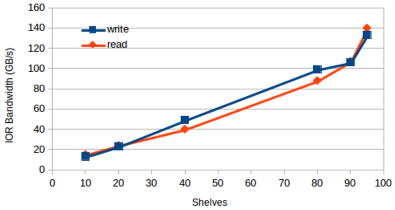
- 226 bare metal hosts, each with 2 NICs; 3556 cores!
- 17 large memory hosts
- Easily reconfigured between hypervisor and lotus roles!

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 ≈ 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):

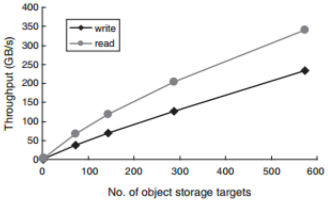
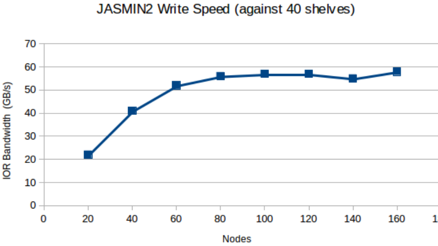


Figure 7
Throughput performance (IOR benchmark).

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 ≈ 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?!

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

(Looking to future technologies ...)

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!

Making use of the bandwidth

Dedicated Lightpath Network

Two weeks in January 2014:
→ Average 10 TB/day, Peak 30 TB/day
→ Inbound onto JASMIN Storage

We've had some network upgrades since then. The bottom line is that you should be able to move TBs per day - to JASMIN at least.

Drivers and Trends

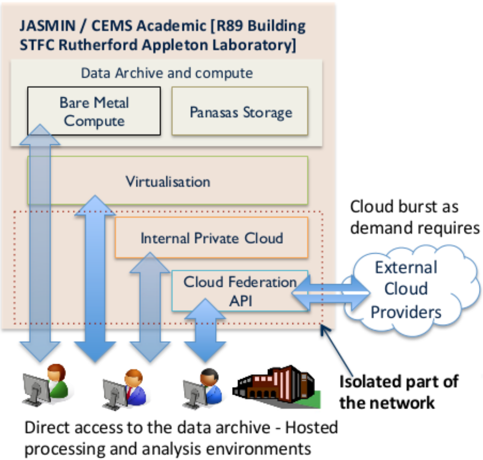
- Scientific (ish)
  - Increasing resolution
  - Increasing interdisciplinarity
  - ... more complexity, more communities involved!
- Technology
  - Cheaper computing ...
  - ... (relatively) more expensive storage.
  - Need to better exploit tape
- Consequence:
  - Frustration
  - More concentration onto community facilities!

Collaboration Infrastructure

- Bigger communities sharing software . . . without necessarily having the necessary understanding of how to "share" software development!
- Drive to bringing compute to the data . . . but where is the data, and is the infrastructure ready for "that particular" compute requirement (software, resource etc)?
- Infrastructures result which are "dedicated", "generic" and trying to cross national boundaries . . . but we haven't really understood all the interfaces and agreements necessary.
- Need to consider institutional, disciplinary requirements in terms of collaboration interfaces as well as software interfaces!

The JASMIN cloud

An attempt to address the "bringing compute to the data" issue:



Final Remarks

- When we consider the entire workflow associated with environmental simulation, we realise that the "time in the supercomputer" **doing** simulation, is only a small part of the entire workflow.
- When we look at the trend in the balance of hardware spending at *weather and climate* supercomputing sites we see a trend towards a greater proportion of the funding on the storage, but

Final Remarks

- When we consider the entire workflow associated with environmental simulation, we realise that the "time in the supercomputer" **doing** simulation, is only a small part of the entire workflow.
- When we look at the trend in the balance of hardware spending at *weather and climate* supercomputing sites we see a trend towards a greater proportion of the funding on the storage, but
- We have yet to see a comensurate trend towards the spend for an appropriate software infrastructure for data, and
- We have yet to see a real understanding of the data handling implications at the generic national and international facilities, although they're all beginning to recognise there might be a problem!

Final Remarks

- When we consider the entire workflow associated with environmental simulation, we realise that the "time in the supercomputer" **doing** simulation, is only a small part of the entire workflow.
- When we look at the trend in the balance of hardware spending at *weather and climate* supercomputing sites we see a trend towards a greater proportion of the funding on the storage, but
- We have yet to see a comensurate trend towards the spend for an appropriate software infrastructure for data, and
- We have yet to see a real understanding of the data handling implications at the generic national and international facilities, although they're all beginning to recognise there might be a problem!

The bottom line: Getting our models to run on (new) supercomputers is hard. Getting them to run performantly is hard. Analysing, exploiting and archiving the data is (probably) **now** even harder!