

Data handling on the path to exascale

Bryan Lawrence



Martin Jukes, Jonathan Churchill
and many others



BADC and CEDA? www.ceda.ac.uk

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

Home Data Centres Data Services **Projects** For Academics For Business About CEDA CEDA News

Data Centres

The Centre for Environmental Data Archival is responsible for the running of the following data centres:

- British Atmospheric Data Centre**
NATIONAL CENTRE FOR ATMOSPHERIC SCIENCE
NATURAL ENVIRONMENT RESEARCH COUNCIL
The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.
- The British Atmospheric Data Centre**
The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.
- NERC Earth Observation Data Centre**
The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.
- IPCC Data Distribution Centre**
The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided. UK Climate Projections.
- The UK Solar System Data Centre**
The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

CEDA: Leading the clip-c Copernicus precursor

Projects

CEDA aims to carry out projects that develop and enhance data and information services. This allows us to actively participate in emerging digital curation and informatics fields, and ultimately provide a better service for environmental science.

CEDA is presently involved in the following projects. For further information please select the title to go to the relevant website or contact CEDA for more information.

- CHARME** - Characterisation of metadata to enable high-quality climate applications and services - CHARME
CHARME is a 2 year FP7 funded project aiming to link commentary metadata (e.g. annotations, supporting information about the data) and datasets. The project will deliver repositories of commentary metadata with interfaces for users to populate and interrogate the information. This will enable users to assess if the of climate data are fit for purpose.
CEDA is working with 8 other UK and European partners, and has key roles on the data model, software development, implementation in archives, and application to climate services.
- LTDP** - ESA Long-Term Data Preservation (LTDP)
CEDA is supporting the European Space Agency (ESA) in its programme for Long Term Data Preservation (LTDP), providing management support and technical expertise to activities associated with European LTDP implementation and framework coordination.
CEDA co-ordinates this project providing expertise on systems technologies, methodologies and approaches which support post mission exploitation of earth observation data.
- Contrail** - Open Computing Infrastructure for Elastic Services
Contrail is a three year FP7 funded project led by INRIA to develop a complete open source cloud computing platform. This will include a solution for federating cloud providers enabling users to seamlessly integrate and scale application across multiple clouds. CEDA is contributing expertise in federated identity management supporting STFC Science who are leading the security work package.
- Climate Information Portal for Copernicus (CLIPC)**
The CLIPC platform will complement existing GMES/Copernicus pre-operational components by providing access on decadal to centennial climate variability data to a wide variety of users. The data will include satellite and in-situ observations, climate models and re-analyses, transformed data products to enable impacts assessments and climate change impact indicators. Supporting data quality and related information will also be made available.
CEDA is leading the project, coordinating a consortium of 22 partners, and leads the access to climate data work package. This work package will provide the software infrastructure to create a single point of access for climate model data from various sources: climate model data, in situ and satellite observations, and re-analyses.
- IS-ENES** - Infrastructure for the European Network for Earth System Modelling - Phase 2 (IS-ENES II)
Peer Review for Publication & Accreditation of Research Data in the Earth sciences (PREPARDE)

Lots more

Key roles in IS-ENES

Mission: Curation AND facilitation!

Outline

Motivation

- Increasing compute means increasing data
- New problems in storing, handling and manipulating “big data”

The Solutions – Such as they are:

- Taking the Compute to the Data (e.g: Exarch)
- **Big Iron (e.g: JASMIN)**
- Getting by with “less” data

The Future

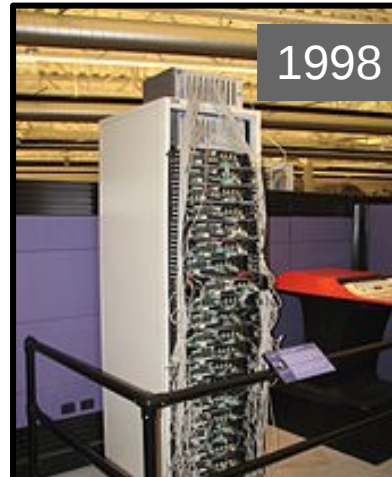
- Fixing the I/O
- Fixing the workflow

Summary

Google's Evolution

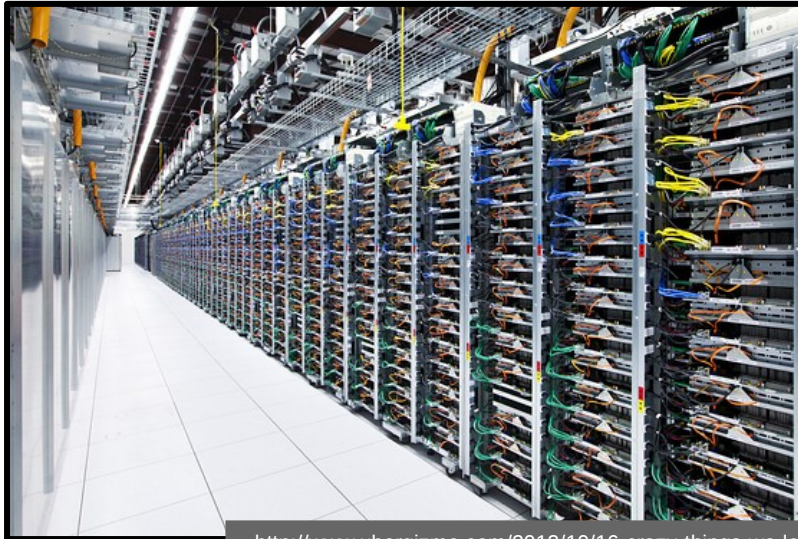


<http://infolab.stanford.edu/pub/voy/museum/pictures/display/GoogleBG.jpg>

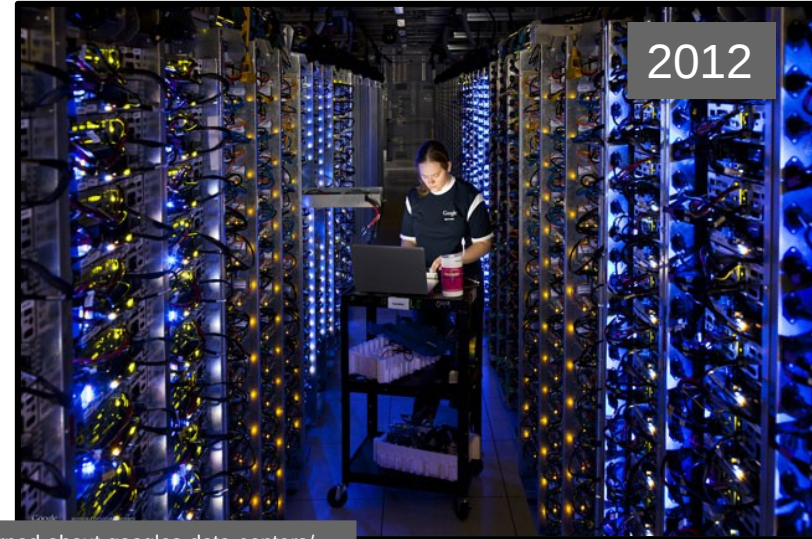


1998

Wikipedia



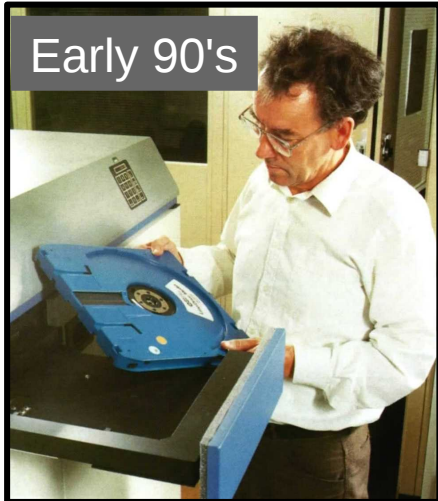
<http://www.ubergizmo.com/2012/10/16-crazy-things-we-learned-about-googles-data-centers/>,
<http://blogs.wsj.com/digits/2012/10/17/google-servers-photos/>



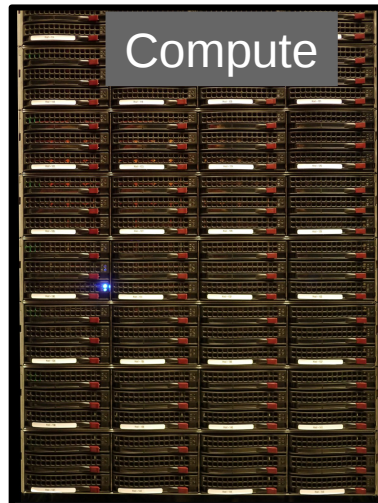
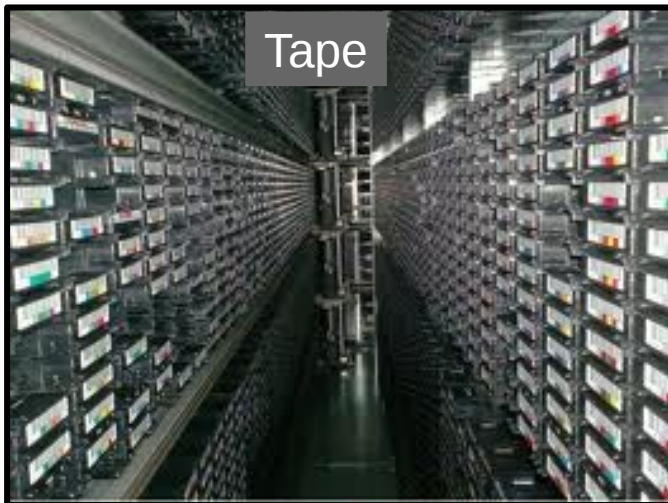
2012

CEDA Evolution

(We're older than Google, but we can do blue lights too!)



2008



Not so subliminal message:

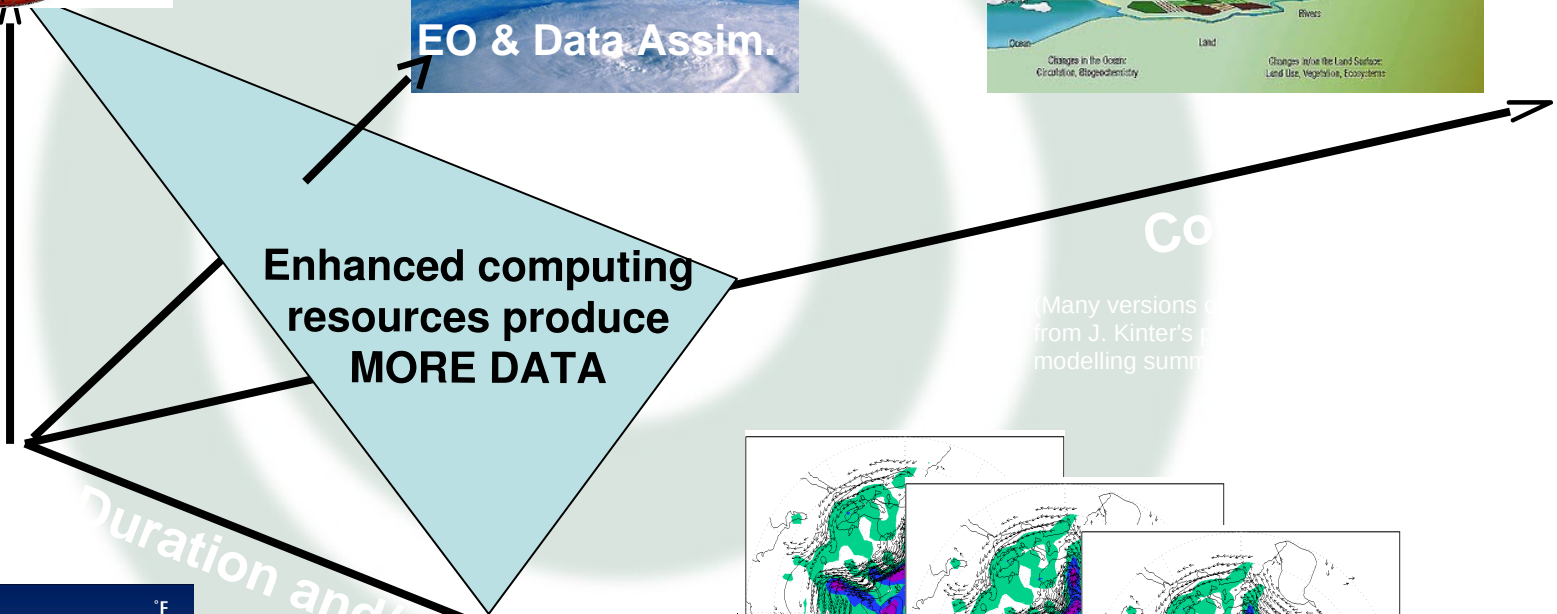
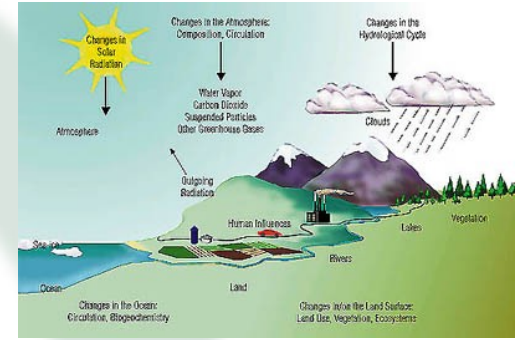
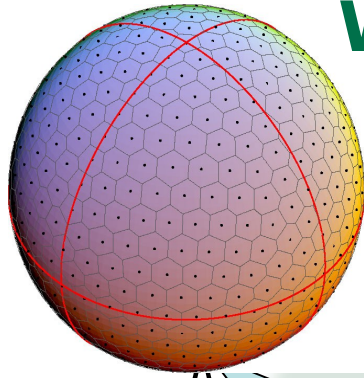
As we move to exascale storage, not everyone will be able to scale from a few machines to one (or more) massive machine rooms.

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!

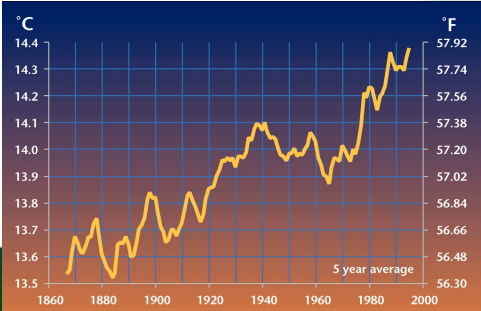
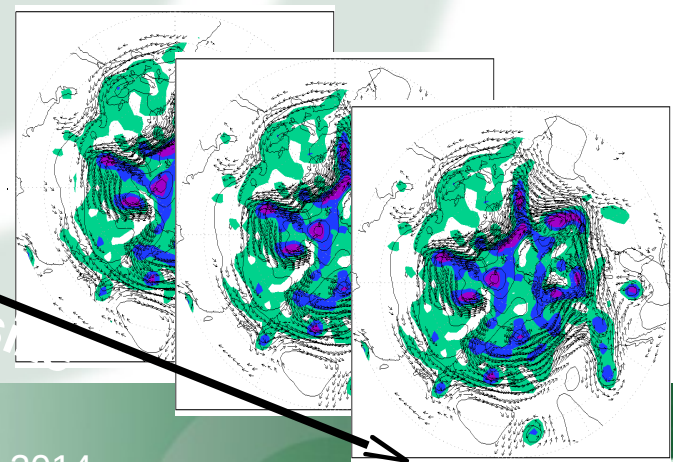


Give me more computing: Whither Numerical Modelling?



Duration and/or Ensemble size

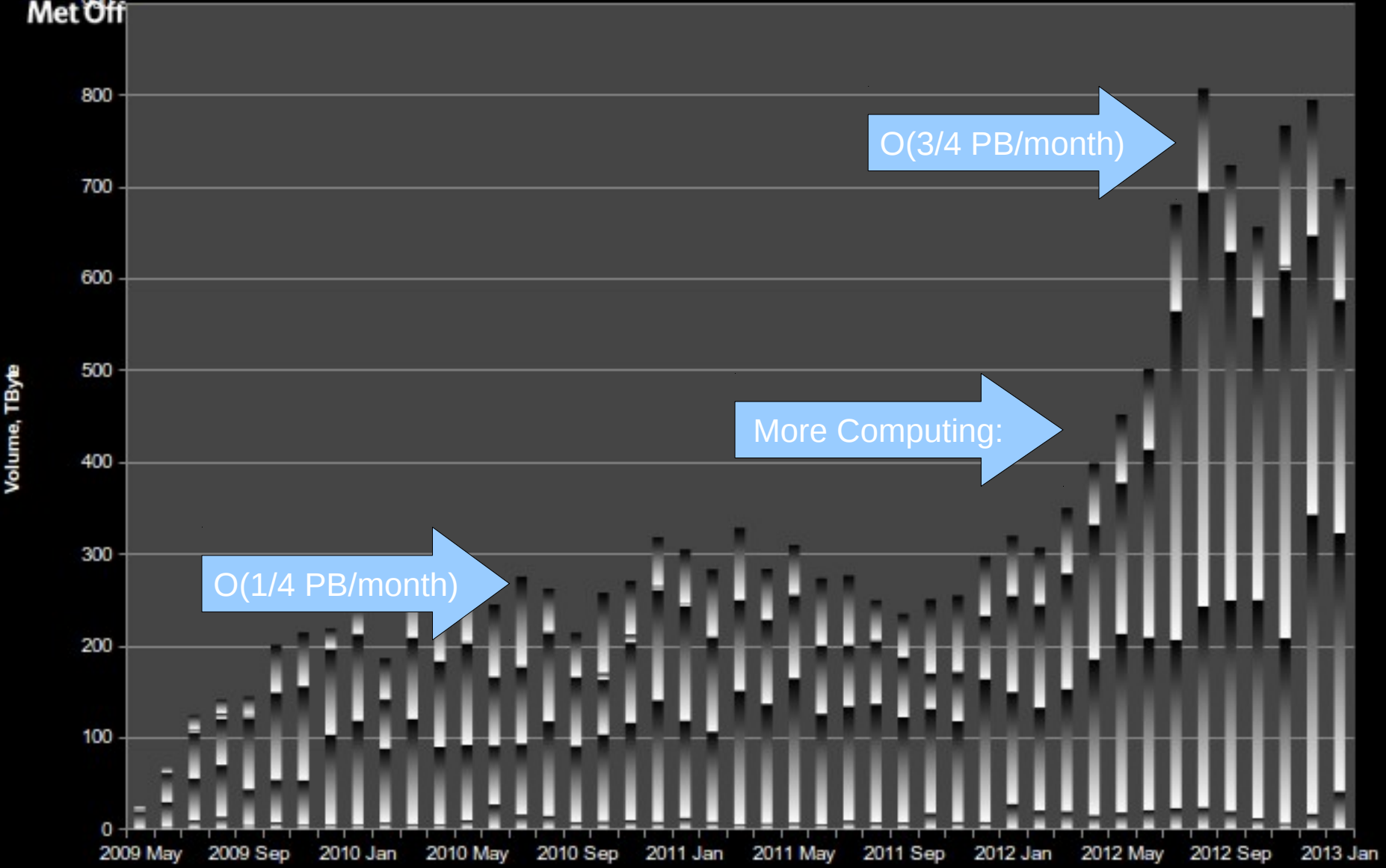
Co
(Many versions of...
from J. Kinter's...
modelling summ...



MASS Monthly Traffic: Archivals

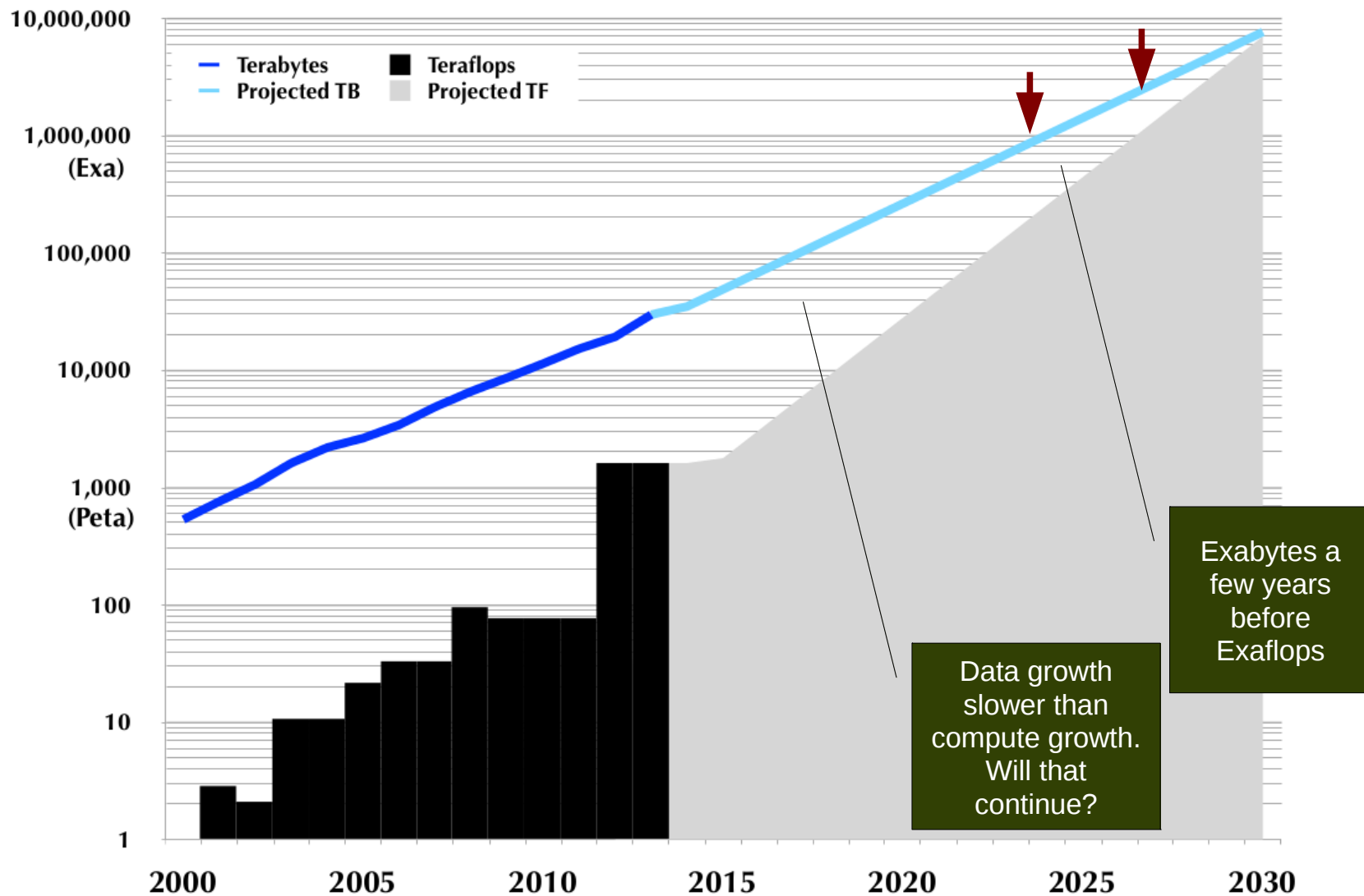


Met Office



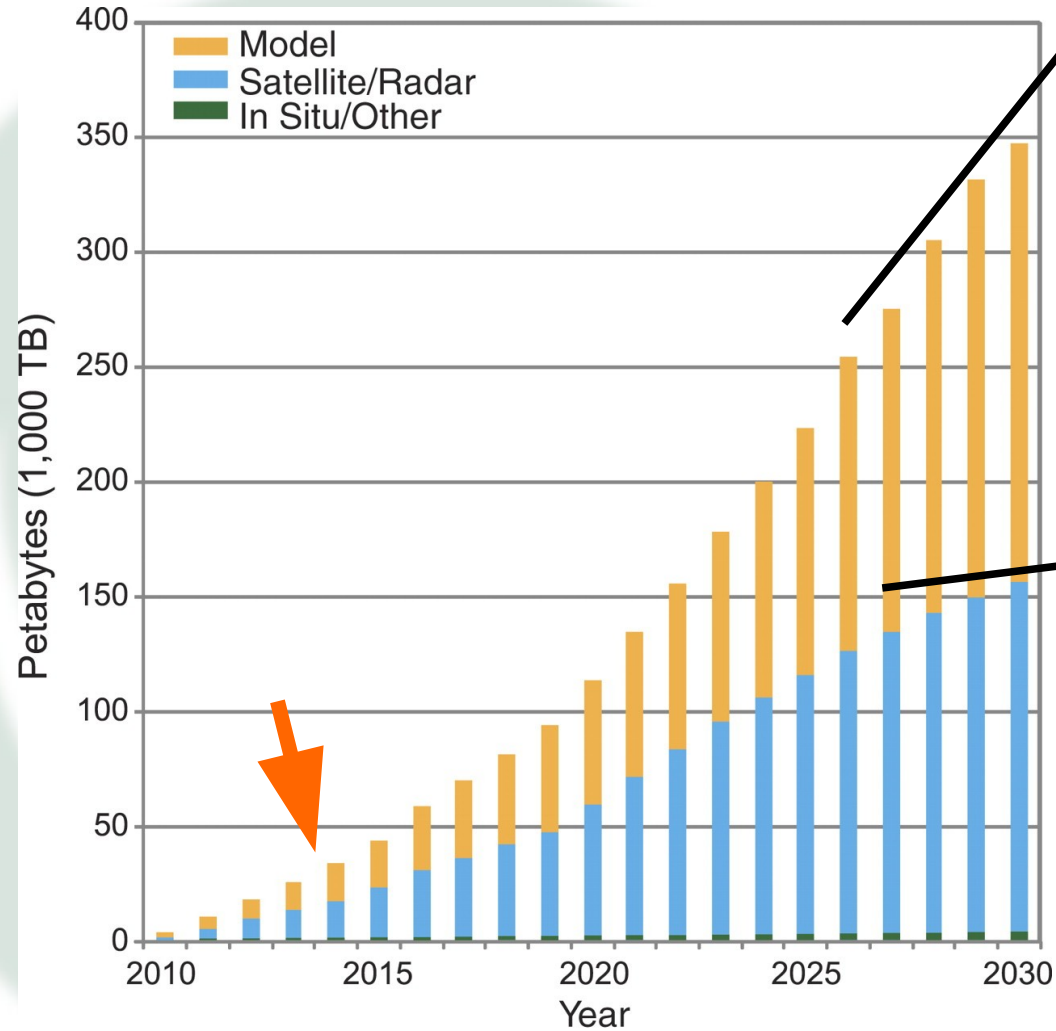
(Mick Carter's Data, my Interpretation)

Data and Compute at NCAR (courtesy Gary Strand)



Gross underestimates ?!

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 "big data" specialist (who is?)



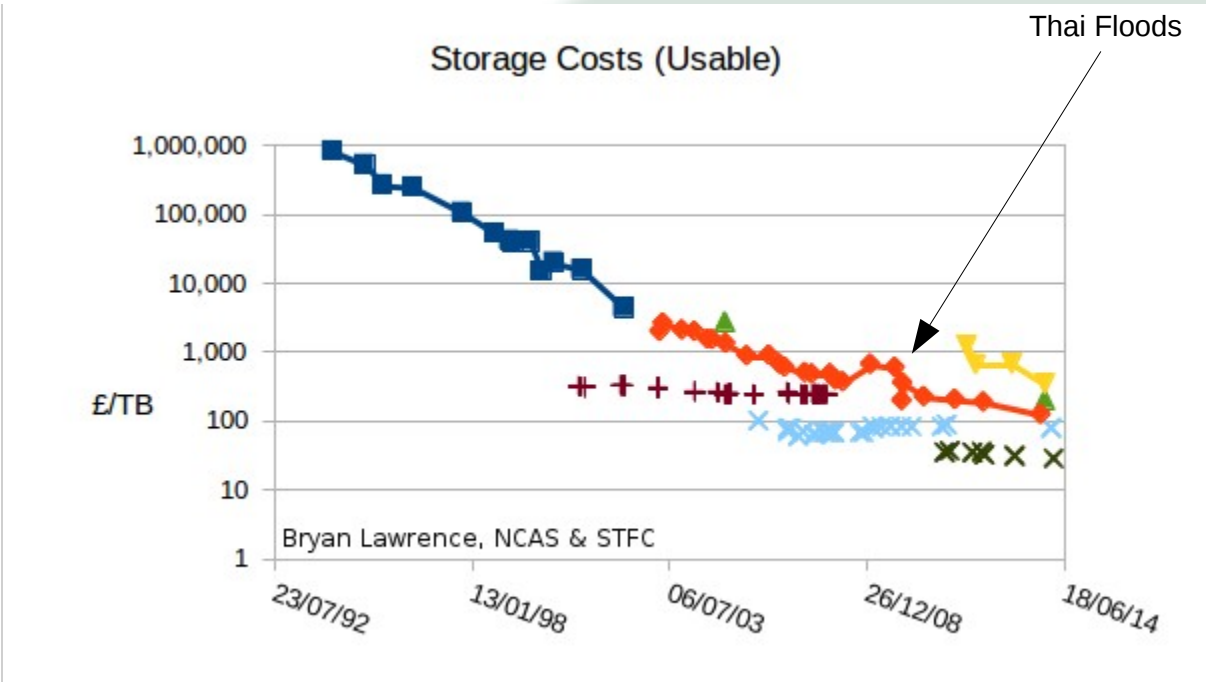
NB
Climate >
EO



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

Cost of storage likely to increase!

Actual costs from STFC:



Filled characters and lines: different generations and disk technologies.

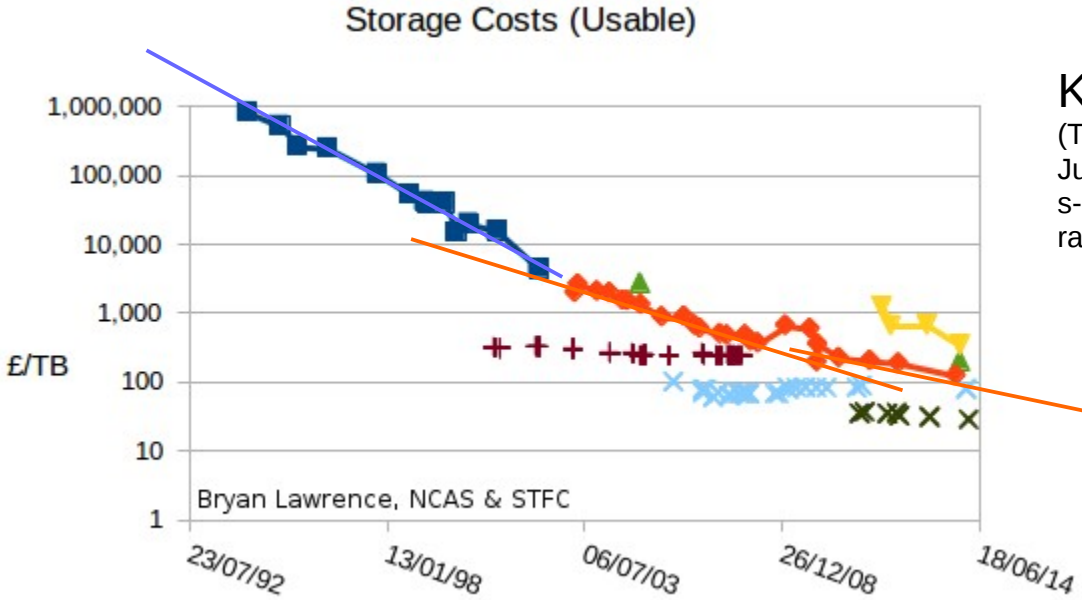
- Yellow is PanFS

Crosses: different tape technologies.

Data courtesy of Peter Chiu, Jonathan Churchill and Tim Folkes

Cost of storage likely to increase!

Actual costs from STFC:



Kryder's Law slowing down!
(There is no such thing as an exponential in real life, Just the growth part of an s-curve, or several s-curves. NB same three "eras" even when we use raw storage costs before RAID and friends.)

Tape technology looks like it has a lot to give us yet, while disk technology is struggling (for the moment a bit like Fusion, the next technology is "just over the horizon").

Whatever, cost of disk is increasing faster than the cost of compute! Especially the cost of "usable" disk.

Data courtesy of Peter Chiu, Jonathan Churchill and Tim Folkes

A quick (broken) calculation

Consider a grand ensemble: Let's say it's something like EC-earth running at 10 years/day, at 25 km grid resolution on 5000 cores.

Not too far from reality?

It would be entirely reasonable (scientifically) to consider a 50 member initial condition ensemble. Let's add in a few variants to try and capture structural (model) uncertainty. Let's say 4.

That's a million core experiment. Feasible with our models today!

But not feasible with our existing HPC(*caveat), but let's pretend we had a million core machine, which EC-earth could use with “similar” core performance.

Now run that “grand ensemble” for 25 years: **2.5 days in the machine**. Only 60 million core hours!

Output?

A 1.25 degree (actually T159L62) model produces roughly 9 GB of data per simulation month in a real application (Colin Jones). Let's say 10 GB to make life easy. (Conservative!)

This simulation could produce (10x5x5: 250 GB model month). The grand ensemble output has: 25x50x4x12=60K months.

So, that's 60K x 250 GB ~15 PB = 6 PB/day (=0.5 Tbit/s!)



2.5 Days in the machine. For climate, years of analysis, during which time, much of the data needs to be online, not nearline!

Data Lifetime is much much longer than the time the model spends in the HPC!

Conclusion: You can't share disk in an archive in the same way as you share an HPC – although obviously you can share support for *your* disk! And you can use tape, but if you expect whole dataset analysis, tape isn't much use!



An Aside: To Recompute or Save?

I'm often asked why we save data, when it might be cheaper to recompute?

There are two issues to consider:

- Can we recompute? (Will it be the “same” simulation? Does it need to be the “same” simulation? When can I get access to the machine again? What is the opportunity cost to the workflow - i.e. I want to do my analysis now, not next month!)
- What are the real costs in play here?

Actually in most cases, e.g. scientific publication workflow of o(some years), model intercomparison, or “frontier experiments, like UPSCALE”, we can't recompute, but let's assume for a moment we can.

- 15 PB in today's money cost (if I put it on tape) ~ 3000 5-TB tapes ~ $3000 \times 50 = \text{£}150,000$
- Today 1000 core-hours cost $>\text{£}10\text{-}20+$, so for a 60 million core hours, ~£1M

That equation will change in the future ... maybe in favour of recompute, but we're not there now, at the moment, even disk storage is cost effective, but the other issue will always be in play.



Consequences?

There's a lot wrong with those calculations,

But the bottom line is that a relatively modest improvement in our software, with a pretty significant improvement in hardware, and hardware availability, along with some realistic efforts to attack uncertainty, will create a data nightmare!

Without improving our dynamic cores (which we will do anyway).

Sure, we can make choices about what to write out - in this future we can start to think about the FLOPS as free, and the BYTES as the significant cost! But we can't recompute unless we can analyse quickly. So analysis is the place to invest!

So, that's the future ... what about now?



Status Quo: UK academic climate computing

Data sources:

- ARCHER (national research computer)
- MONSooN (shared HPC with the Met Office – JWCRP)
- PRACE (European supercomputing)
- Opportunities (e.g. ECMWF, US INCITE programme etc)
- ESGF (Earth System Grid Federation)
- Reanalysis
- Earth Observation
- Ground Based Observations

=> **Big Data
Everywhere!**

Nor any I should clone?

“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

Solution 1: Take the (analysis) compute to the data

How? All of:

- (1) System: Programming libraries which access data repositories more efficiently;
- (2) Archive: Flexible range of standard operations at every archive node;
- (3) Portal: Well documented workflows supporting specialist user communities implemented on a server with high speed access to core archives;
- (4) User: Well packaged systems to increase scientific efficiency.
- (5) Pre-computed products.



Solution 1: Take the (analysis) compute to the data

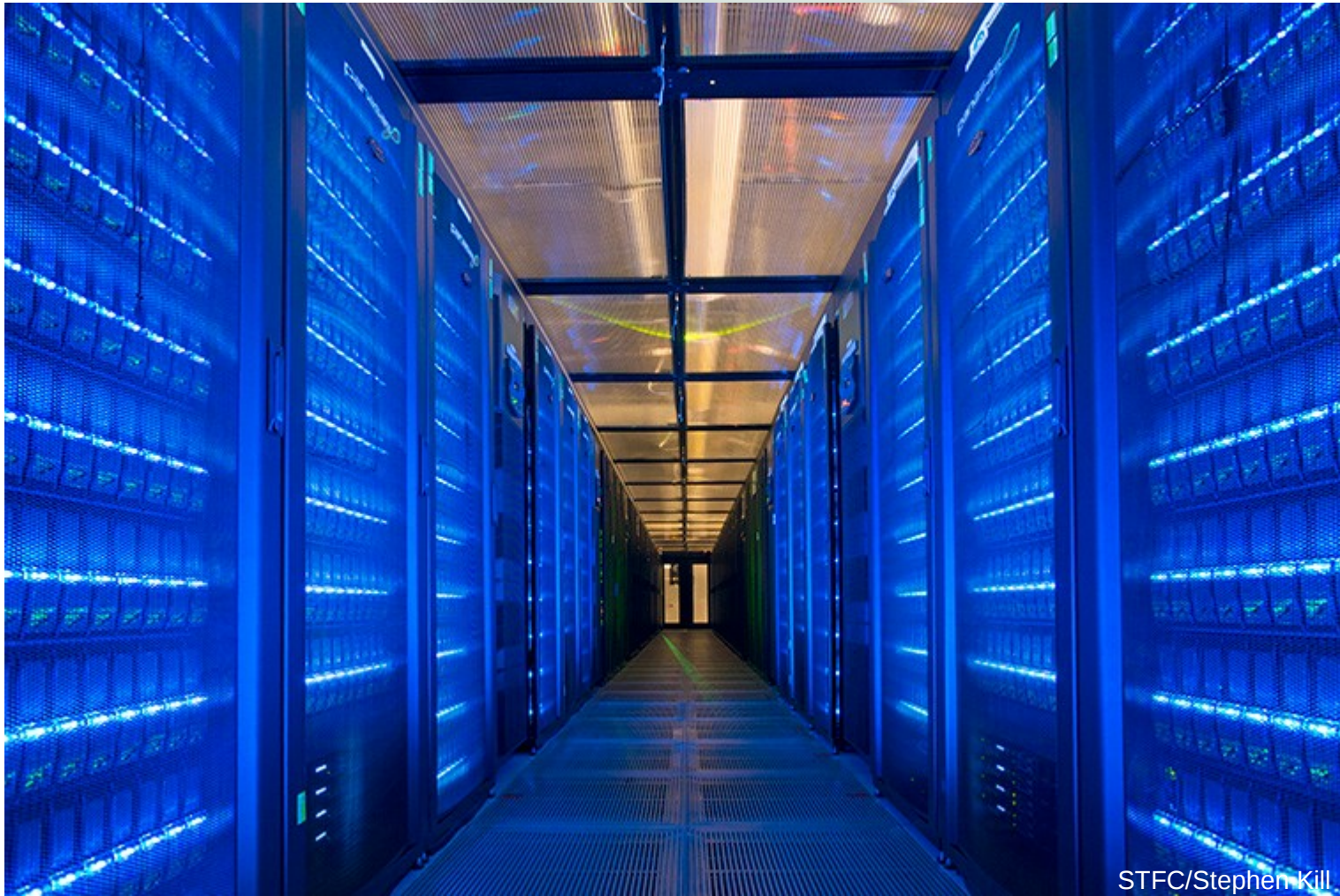
How? All of:

- (1) System: Programming libraries which access data repositories more efficiently;
- (2) Archive: Flexible range of standard operations at every archive node;
- (3) Portal: Well documented workflows supporting specialist user communities implemented on a server with high speed access to core archives;
- (4) User: Well packaged systems to increase scientific efficiency.
- (5) Pre-computed products.

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



Solution 2: Centralised Systems for Analysis at Scale



STFC/Stephen Kill



**National Centre for
Atmospheric Science**
NATURAL ENVIRONMENT RESEARCH COUNCIL

ECMWF, Apr 2014

www.ncas.ac.uk

JASMIN: Joint Analysis System

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

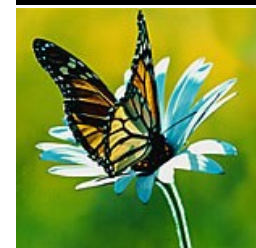
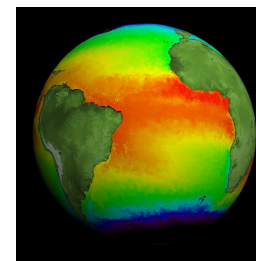
For

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?**

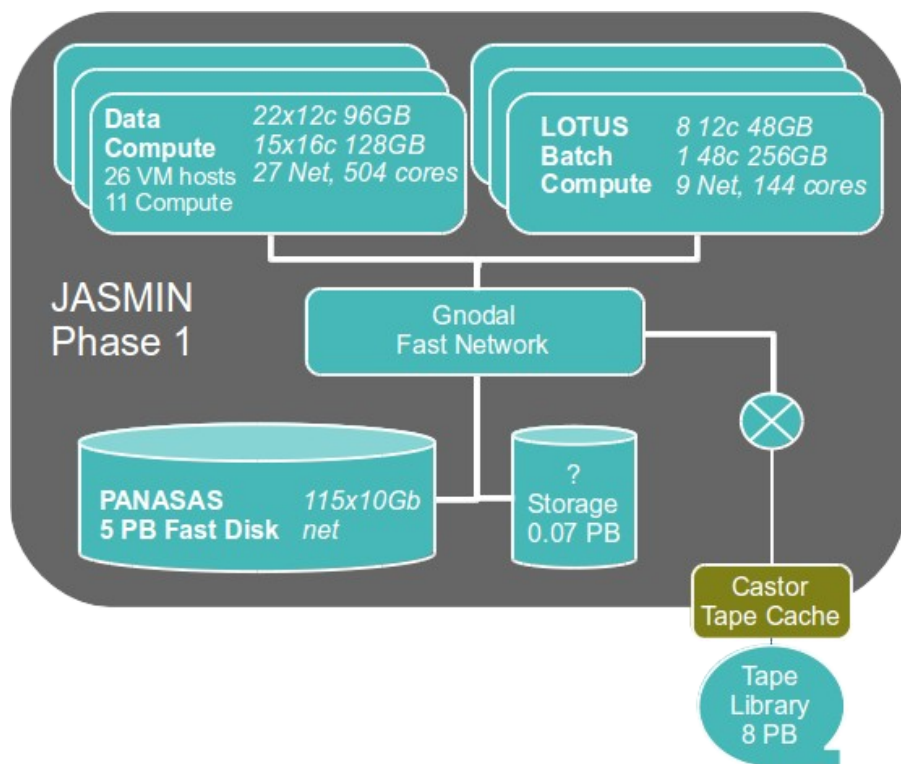


Opportunities

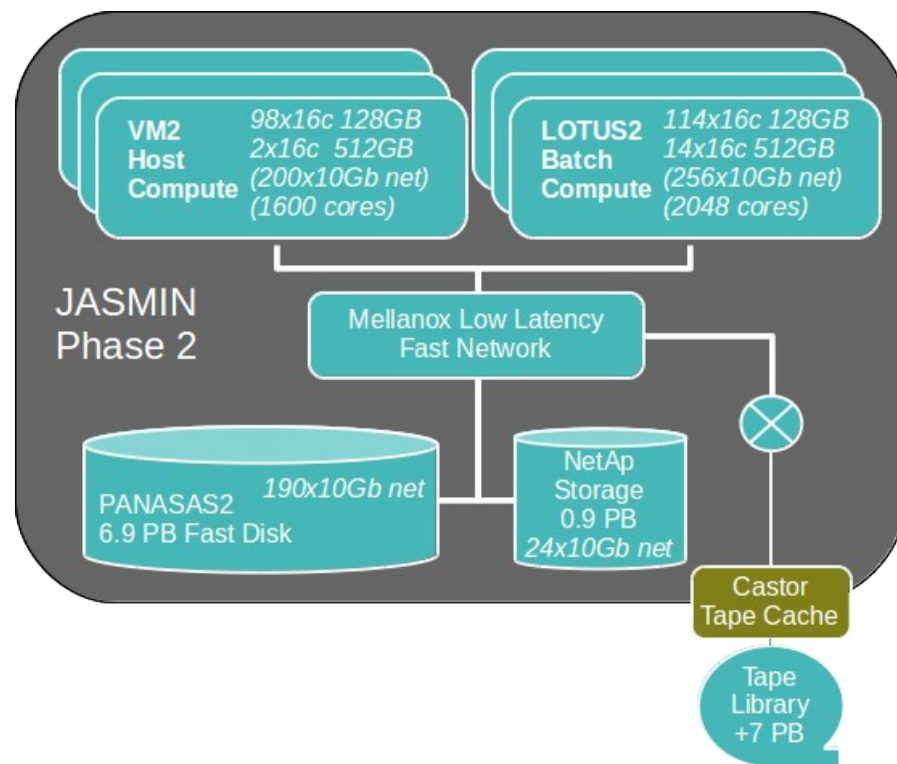
JASMIN is a collaboration platform!
within NERC (who are the main investor)
with UKSA (& the Space Catapult via CEMS)
with EPSRC (joined up national e-infrastructure)
with industry (cloud providers, SMEs)

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

JASMIN Phase 1 and Phase 2 (Co-existing)



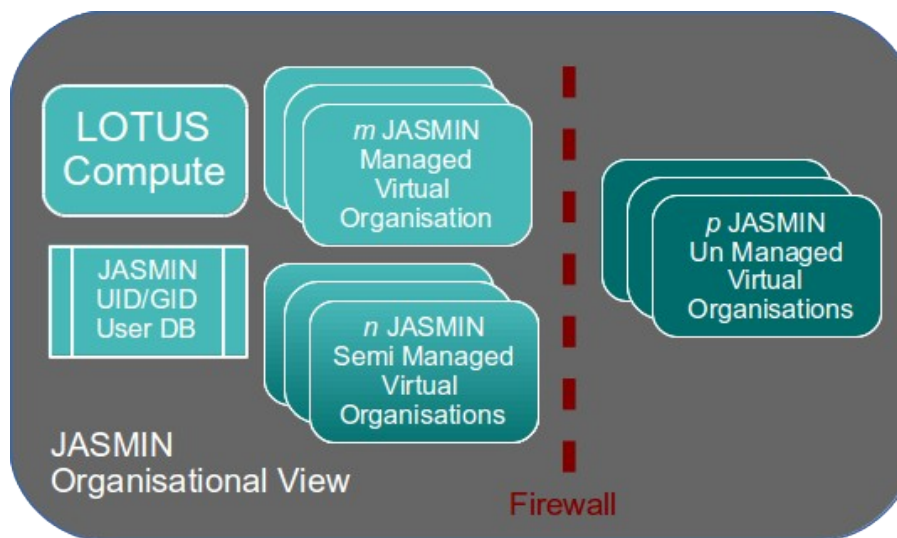
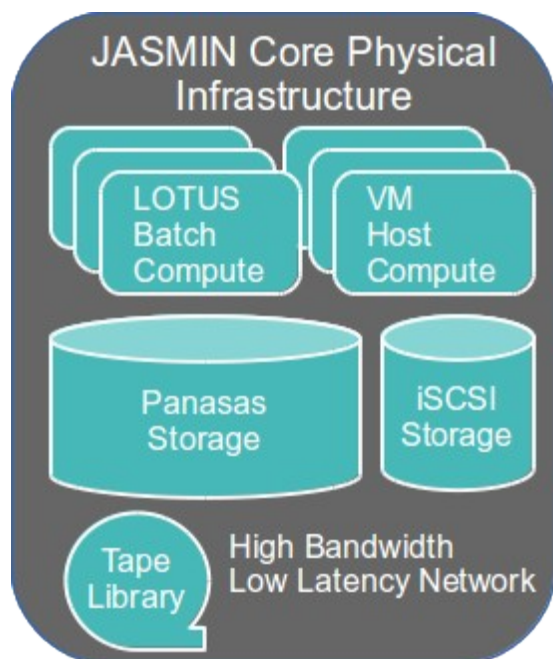
Deployed April 2012



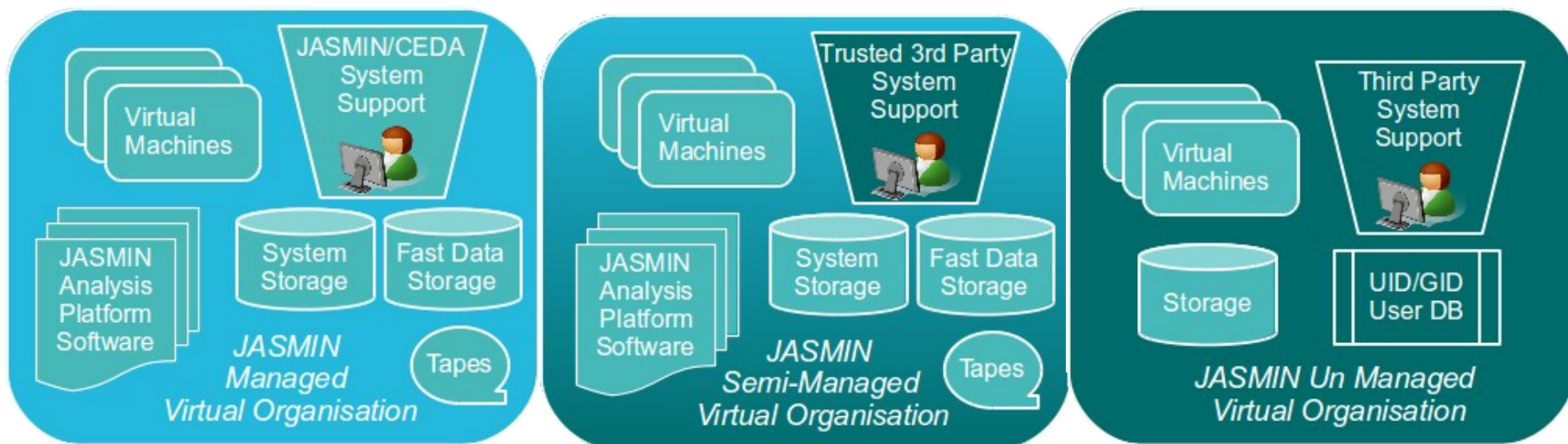
Additional hardware!
- Deployed April 2014

Phase 2 includes h/w for Muller (UCL), Wright (Leeds), Kerridge (STFC), Field (CEH)

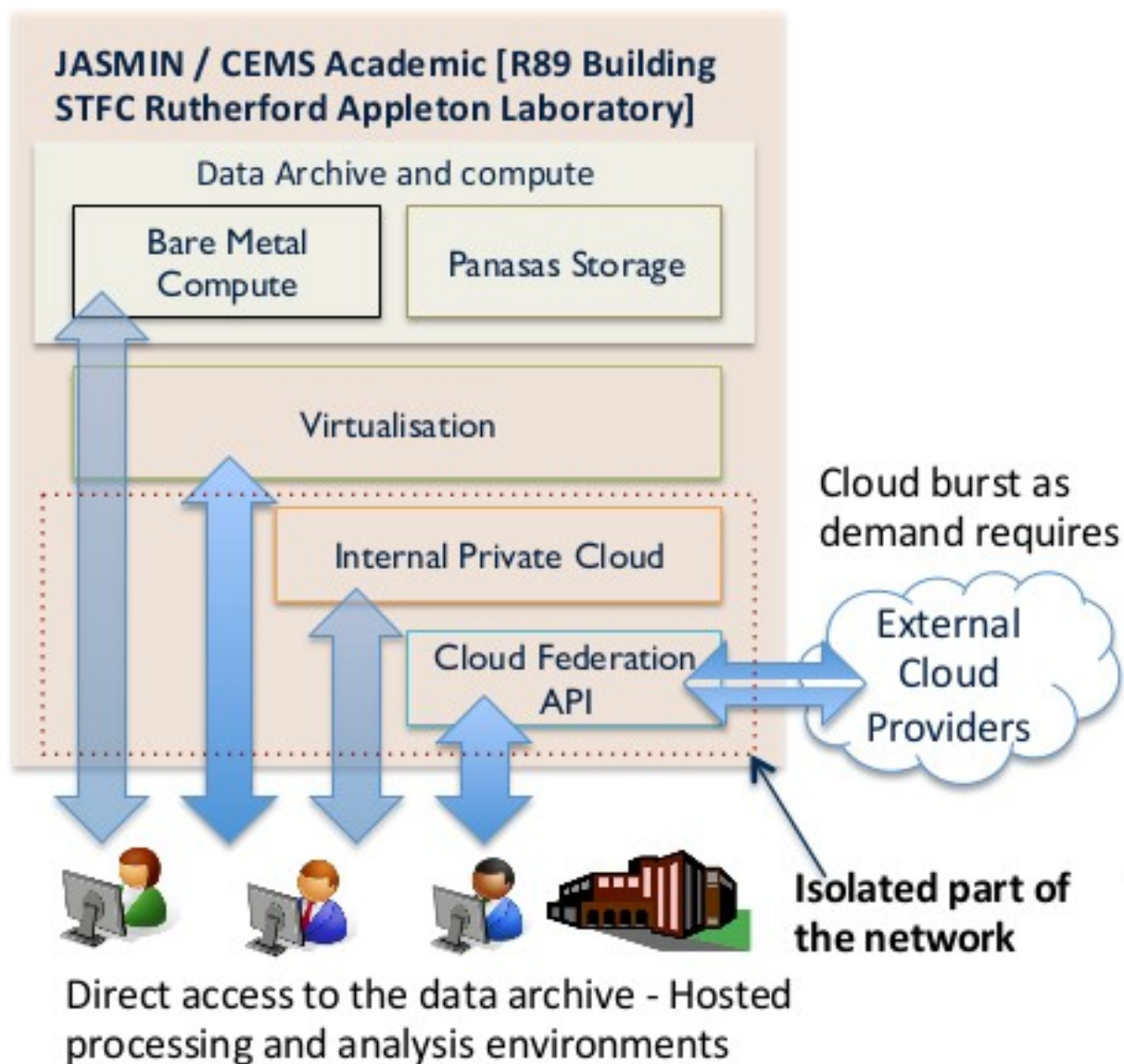
Physical and Organisational Views



Managed, Semi- and Un-managed Organisations



Platform as a Service (Paas) -----> Infrastructure as a Service (IaaS)



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 later this year.)

JASMIN2 Panasas I/O Performance

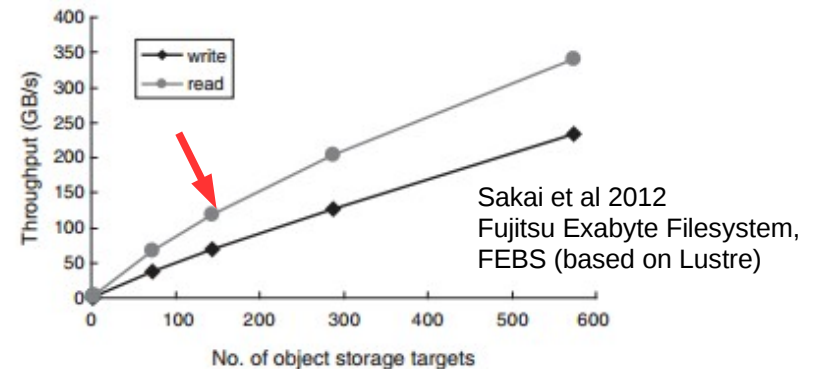
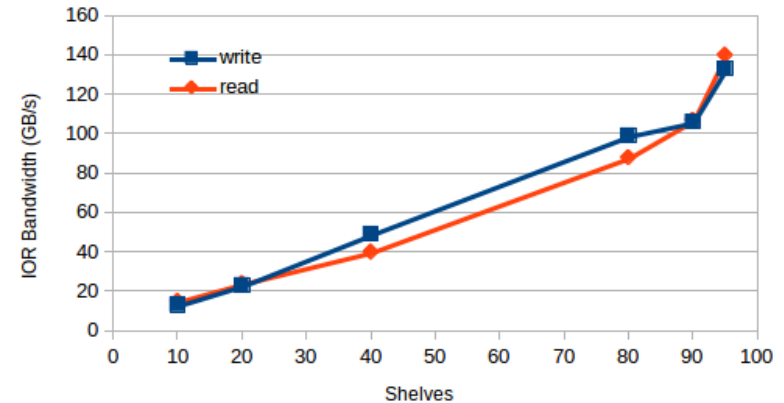


Figure 7
Throughput performance (IOR benchmark).

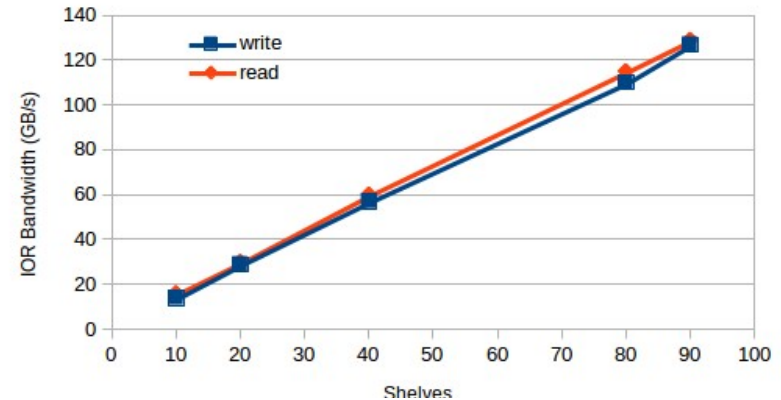
Performance v Reliability

In a Panasas file system we can create “bladesets” (which can be thought of as “RAID domains”, but note RAID is file based).

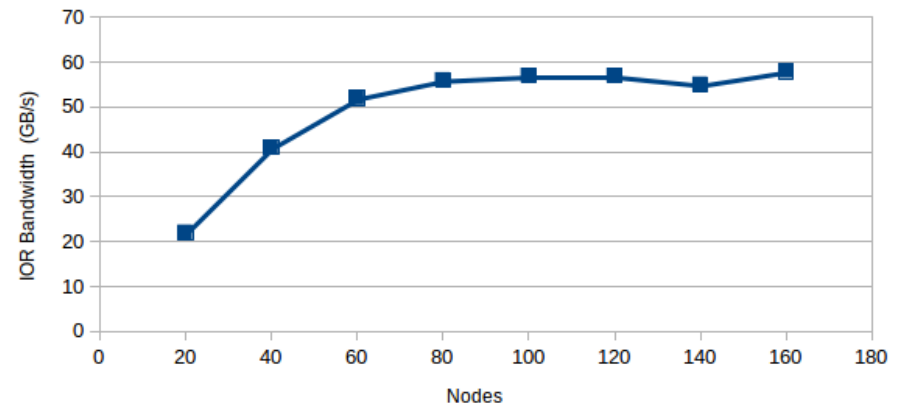
Trade-off (per bladeset) between performance, and reliability:

- 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 \sim 240 disks per bladeset. In JASMIN2 that's \sim 0.9PB (0.7 in JASMIN1, with 3 TB disks of J2, 4 TB)
- Typically, we imagine a virtual community maxing out on a bladeset, so per community, we're offering $o(20)$ GB/s).

JASMIN2: Influence of Bladeset Size



JASMIN2 Write Speed (against 40 shelves)



Another 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?!

From a simulation point of view: our file systems are nowhere near keeping pace with our compute!

Reliability at Exascale?

We're currently (last year) swapping a part (blade, power supply, battery) a fortnight, on our o(5 PB) phase 1 store.

By 2020 we're planning on having roughly 30 PB of disk (funding permitting).

What failure rate might we expect?

- Simple extrapolation, one every couple of days (not all disk failures!).
- Probably safe enough with build time of o(hours) for our bladesets.
- It'll probably be better than that, because we'll have bigger disks (and we think the failure rate is a function of both spindles and bytes).

It would be much worse on commodity hardware (our Panasas kit is at least a factor of two more reliable than the commodity kit used a couple of isles across our machine room in the LHC Tier-1 data centre).

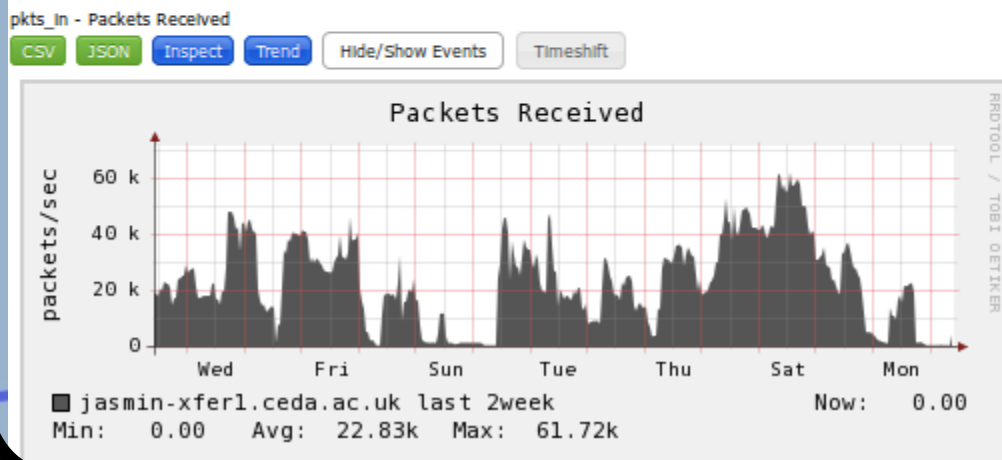
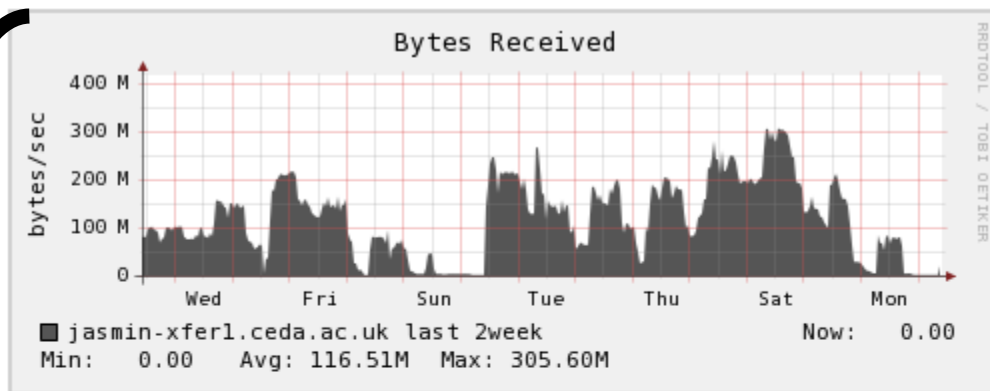
What about exascale?

Simple extrapolation suggests 20 times more failures, or a couple a day.

But we won't have 20 times more communities, so we're likely to increase our "Raid Domain" sizes. Much higher risk?

Tape failures will also start to hit us more?

We are only just starting to think about this ...



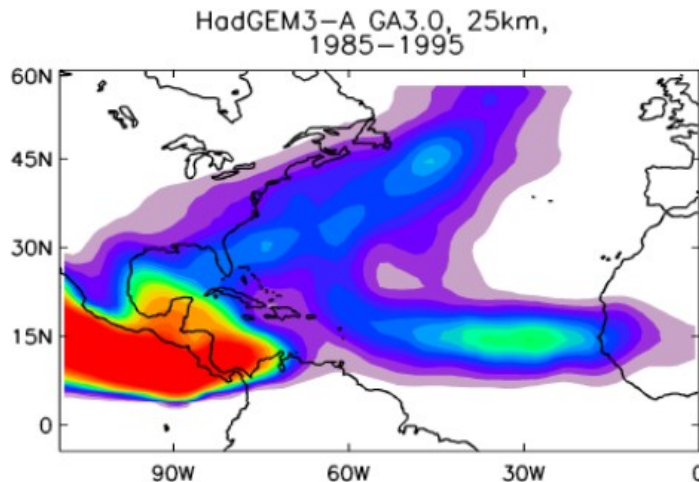
Two weeks in January 2014:
 → Average 10 TB/day, Peak 30 TB/day
 → Inbound onto JASMIN Storage

Dedicated Lightpath Network

JASMIN Science

UPSCALE (NCAS + UKMO)
~ 350 TB stored
(Peaked at over 500TB)

Cyclone tracking algorithm running on MONSooN postprocessor took 55h, on JASMIN, 22h to process one 7 month season of the 25km N512 model
Still not as quick as the original simulation (on PRACE), but we have yet to parallelise this!



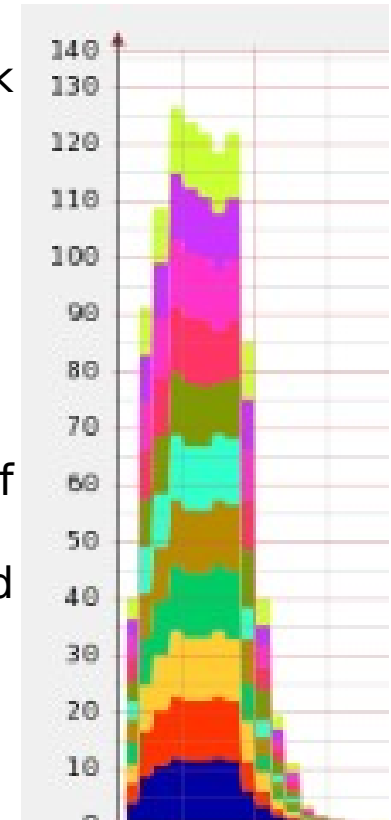
ATSR Reprocessing (RAL)

Last reprocessing took place in 2007-2008.

Used 10 dedicated servers to process data and place product in archive

Previous reprocessing complicated by lack of sufficient contiguous storage for output and on archive.

Reprocessing of 1 month of ATSR2 L1B data using original system took ~**3 days**:
using JASMIN-HPC Lotus: **12 minutes**.



132 cores flat out (NOT I/O bound) for 12 minutes!

Solutions, 3: Less data, better data logistics!

Towards Exascale: It won't be just about the underlying infrastructure, it'll be about the experiment planning ... about data logistics as much as model performance!

Resolution:

Grid resolution and science resolution are different concepts. Why do we write out data at grid resolution?

(Confusion of restart with analysis?)

Ensemble Members:

If we are sampling a PDF, do we need to keep lots of “similar” instances? Can we just keep some?

Which ones? Need “in-flight” ensemble diagnostics.

Model Democracy?

If we knew what a good model was, can we afford to *keep* the output from bad models? (... but ... skill? ...)

(we already throw lots of our own data out ...)

(more use of) Temporal Slicing

“Campaign Periods” of higher resolution data within long duration runs.

Better (and agreed) use of

Compression, e.g. WGDOS for NetCDF!



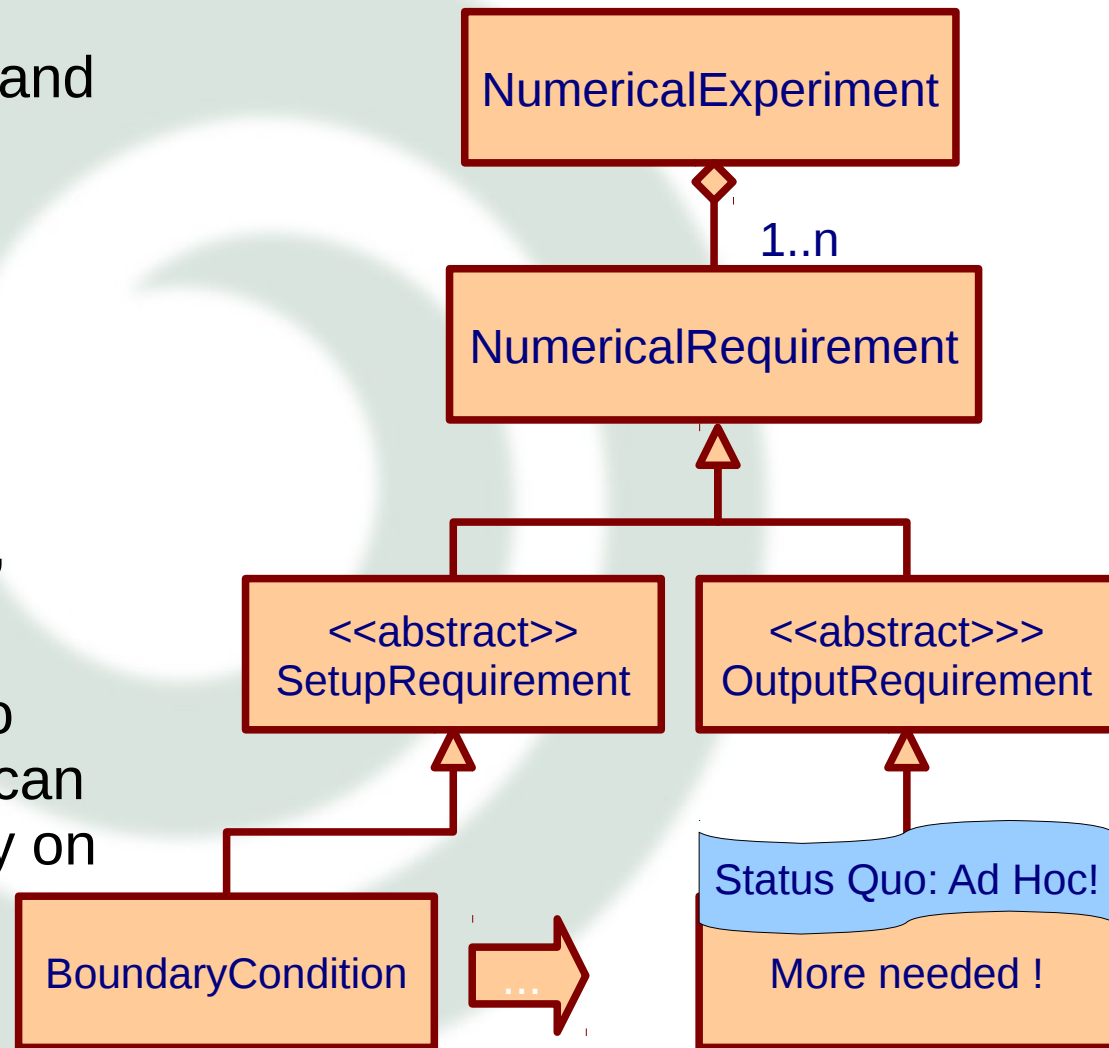
All of which is about “Experimental Design”

Requires much clearer definition of (numerical) experiment requirements and constraints.

Experiment requirements include comprehensive output specifications.

Duration, Resolution (space/time), Parameters, Statistics (?)

All machine readable, so model **workflow tooling** can “just get it right”. Can't rely on humans!



Summary

- We will hit exabytes long before we have exaflops! We will spend more of our “compute budget” on storage.
- We have a scalability problem with storage (and I/O ...)
- We have a scalability problem with our workflow.
- Solution is some combination of new software and dedicated analysis hardware.
 - Both need more investment!
- We can probably expect to see more customised data analysis environments such as JASMIN.
- We need to automate our experiment definitions in order to automate our workflow.

