Two approaches to beating data bottlenecks in weather and climate science

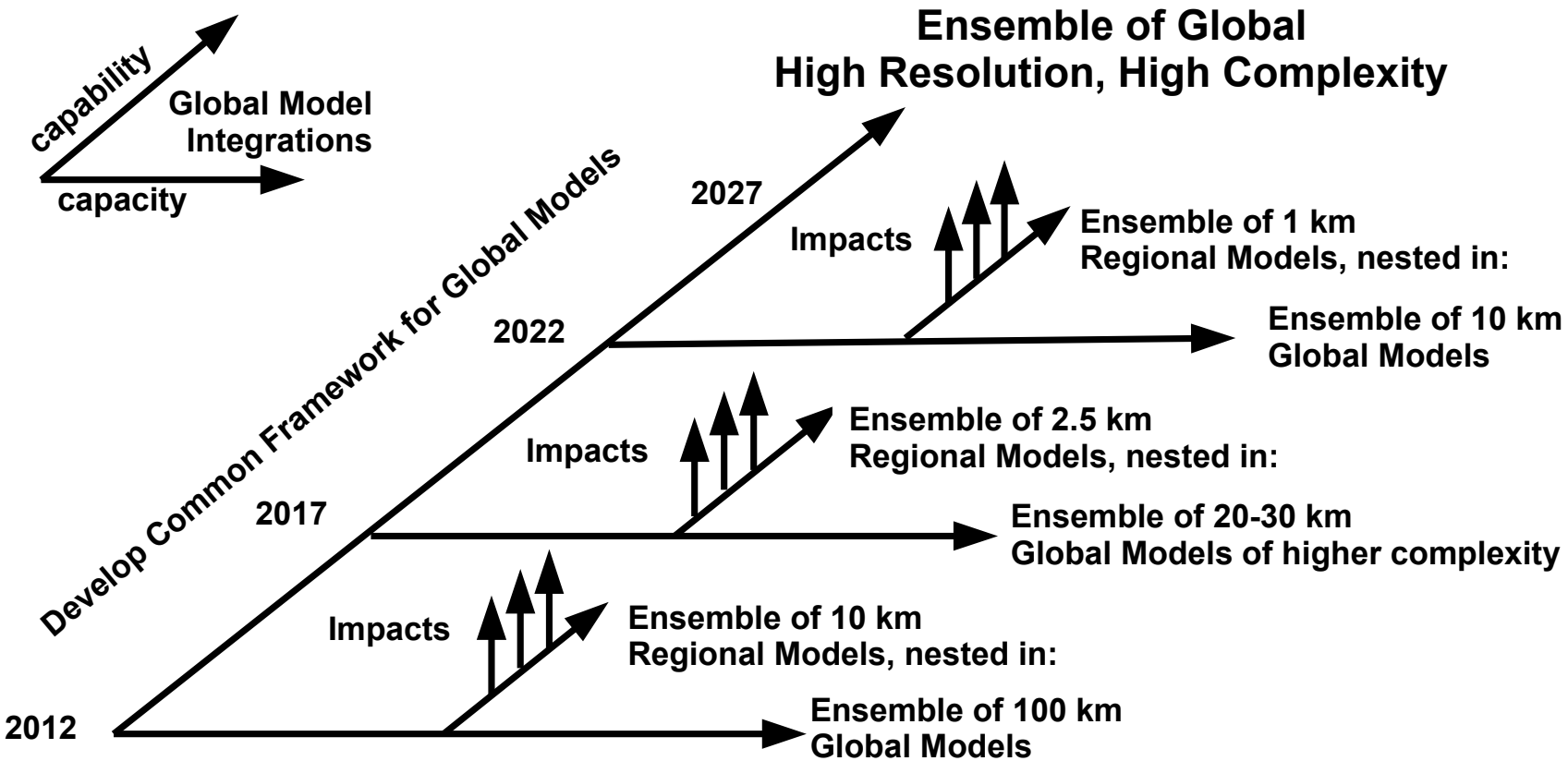
Bryan Lawrence[+†] and Julian Kunkel[†]

[+]NCAS & [†]University of Reading

Jülich, 18/09/18

This presentation includes work from the ESiWACE project funded via the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191.

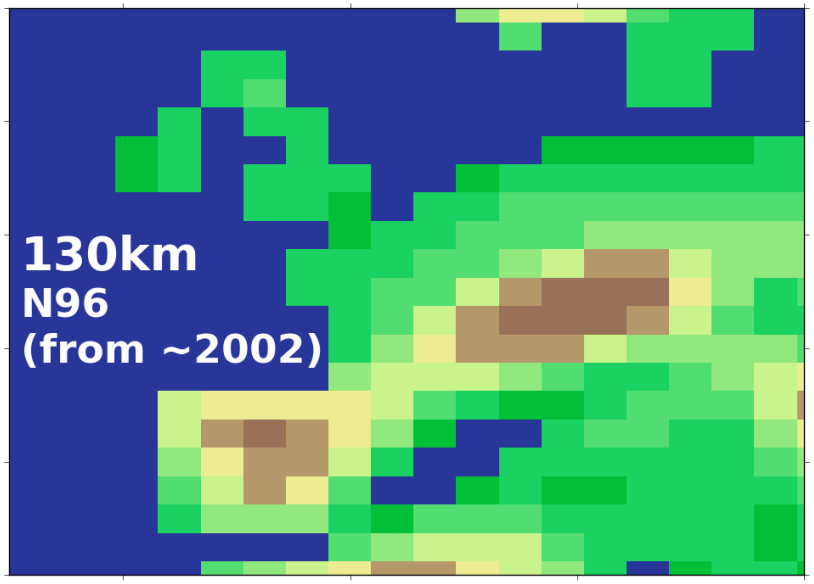
Climate Goals

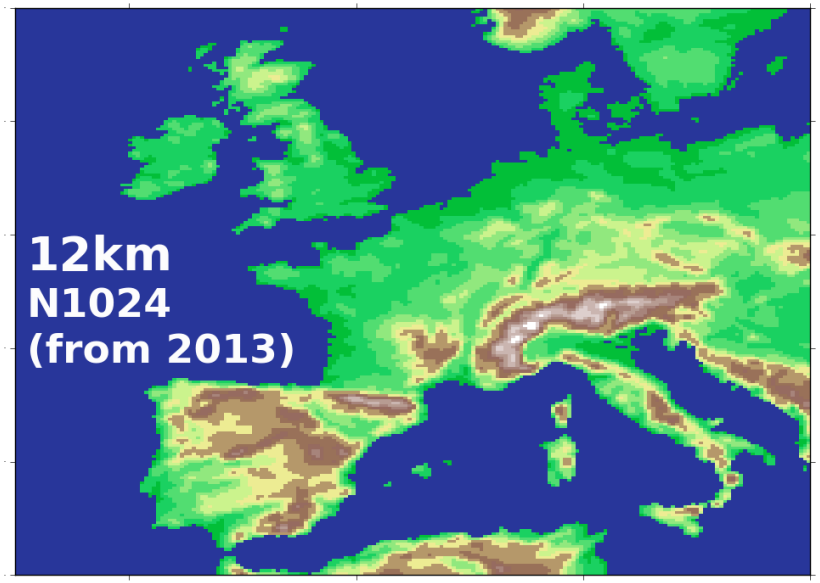


(From "Infrastructure Strategy for the European Earth System Modelling Community" 2012-2022, Mitchell et al, 2012.)

Ever increasing data production

Europe within a global model …





One ”field-year” — 26 GB

1 field, 1 year, 6 hourly, 80 levels
1 x 1440 x 80 x 148 x 192

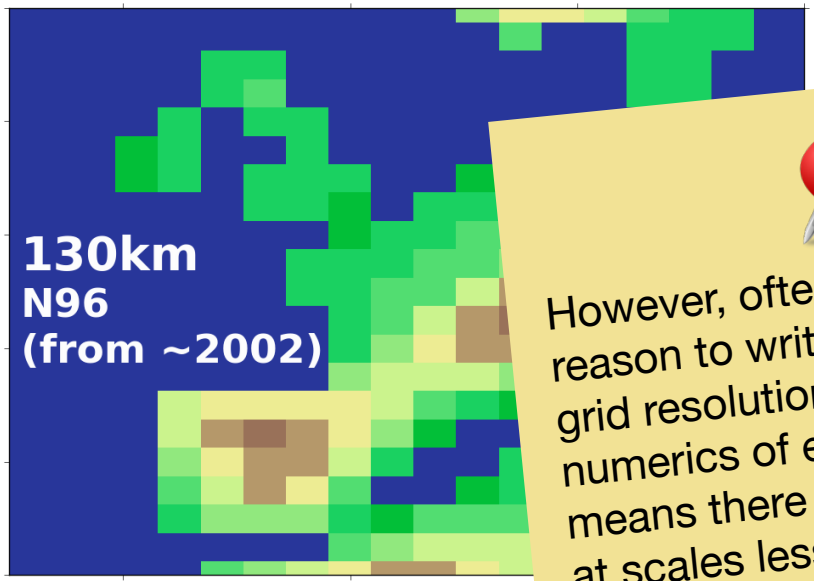
One ”field-year” — 6 TB

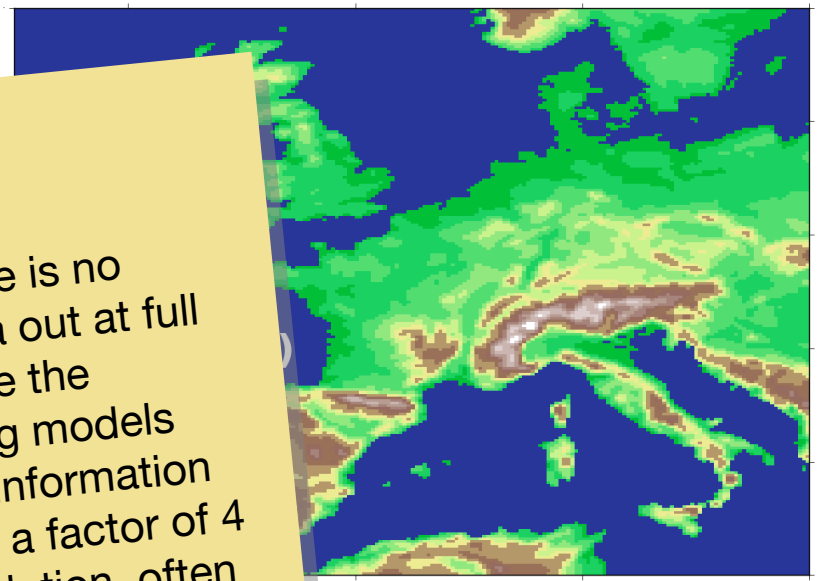
1 field, 1 year, 6 hourly, 180 levels
1 x 1440 x 180 x 1536 x 2048

(Just one axis of data production;
ensembles produce even greater data problems.)

Ever increasing data production

Europe within a global model …





However, often there is no reason to write data out at full grid resolution since the numerics of existing models means there is no information at scales less than a factor of 4 times the grid resolution, often more, so a “one-off” considerable saving is possible for most use cases.

One ”field-year” — 26

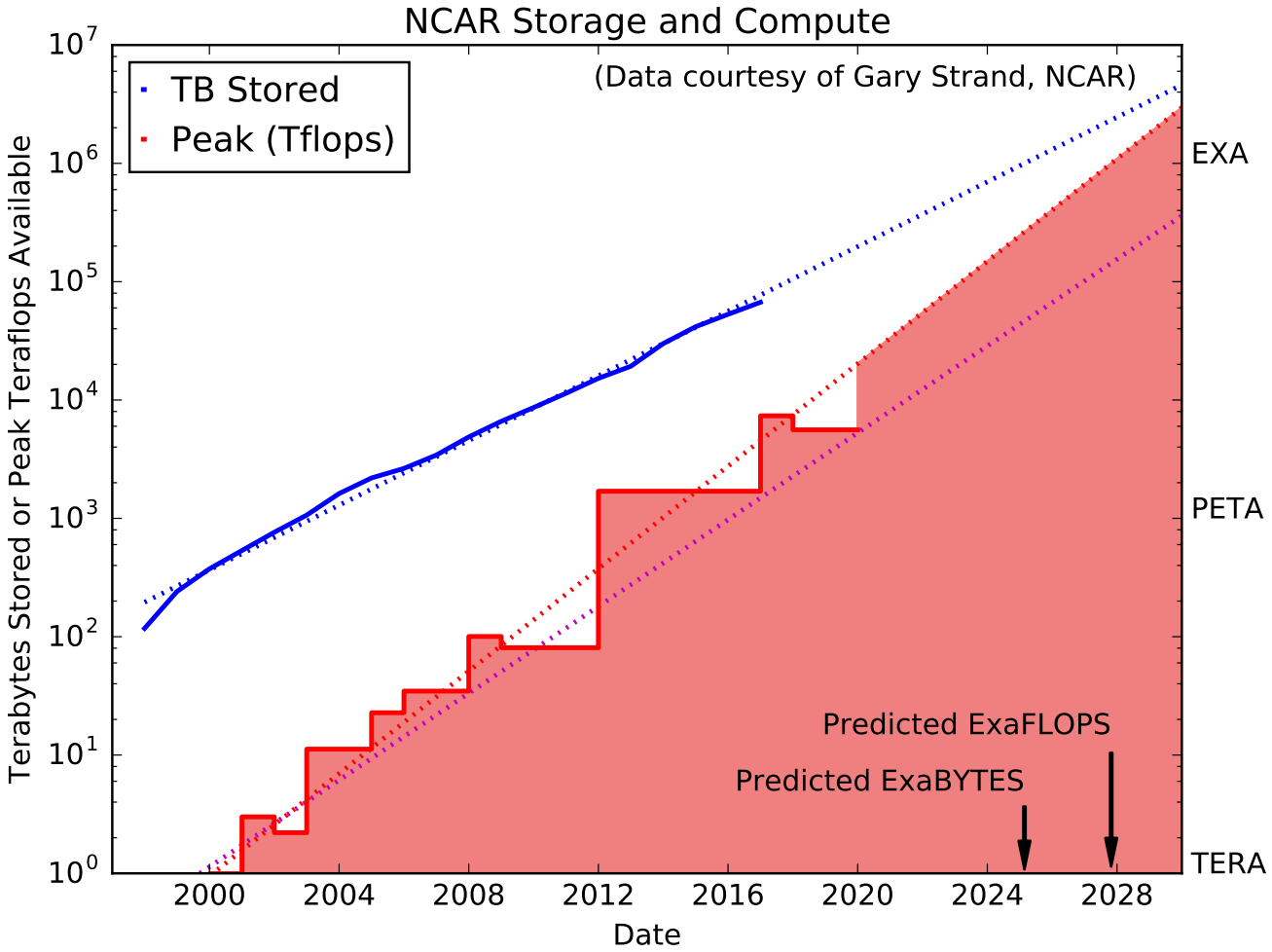
1 field, 1 year, 6 hourly, 80 level
1 x 1440 x 80 x 148 x 192

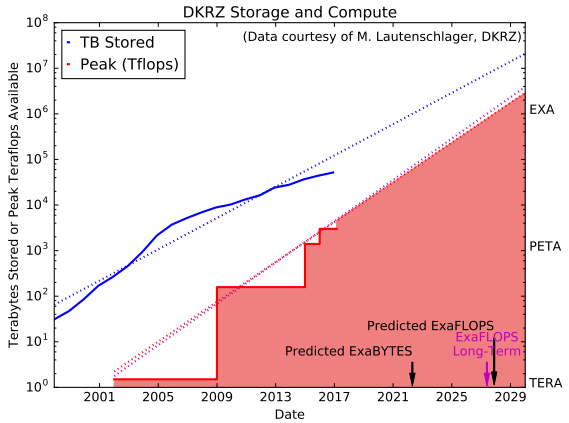
…ear” — 6 TB

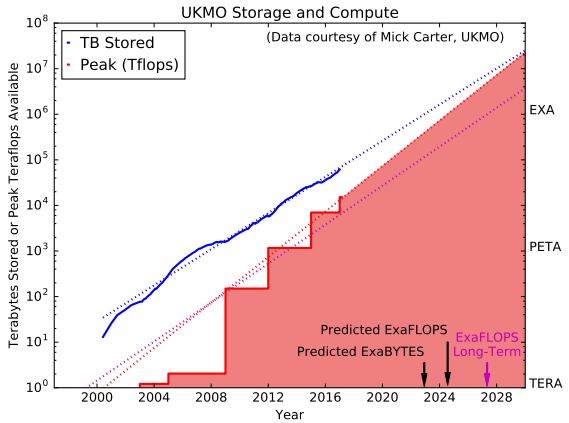
…ld, 1 year, 6 hourly, 180 levels
1 x 1440 x 180 x 1536 x 2048

(Just one axis of data production;
ensembles produce even greater data problems.)

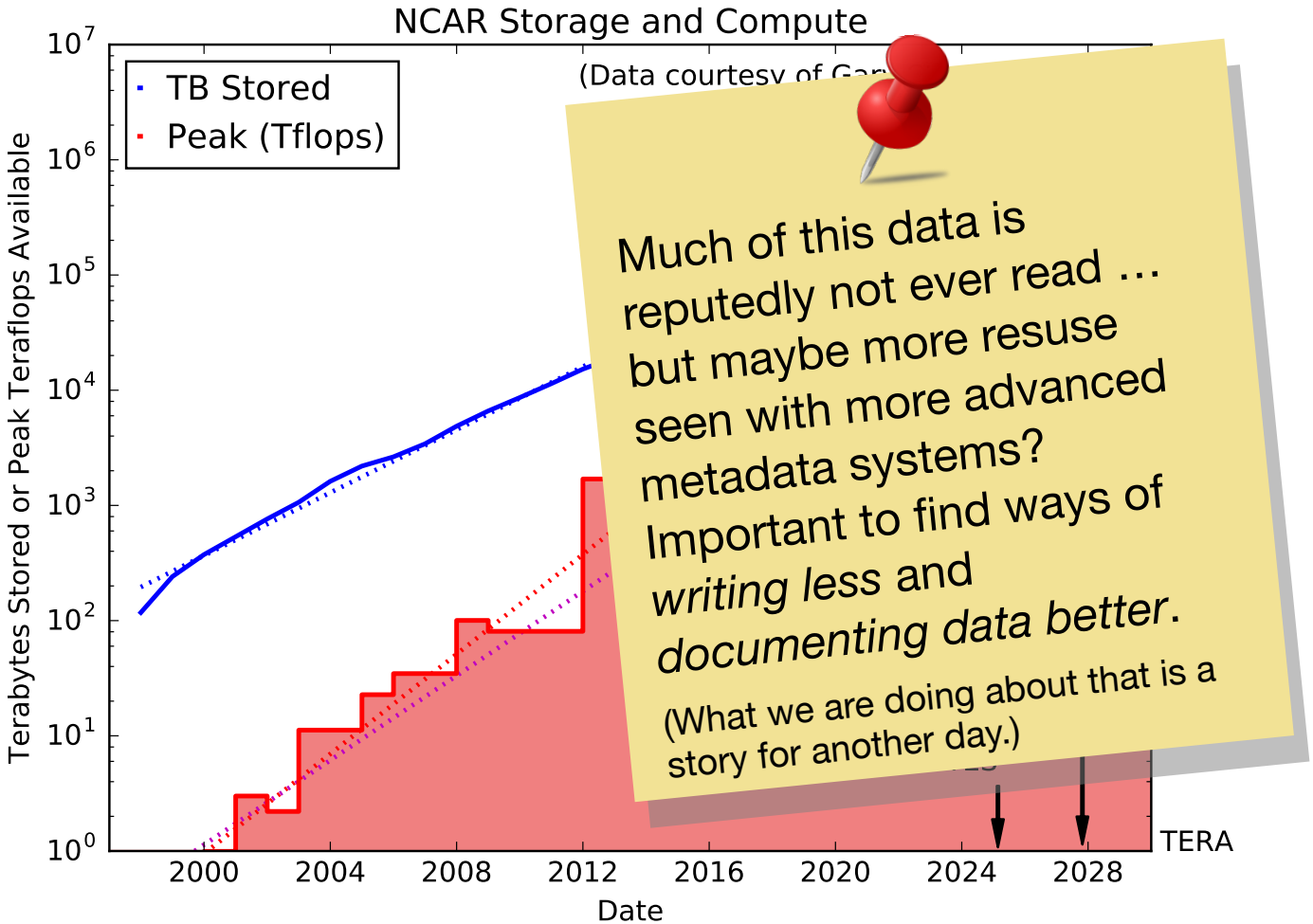
Consequences of increasing data …

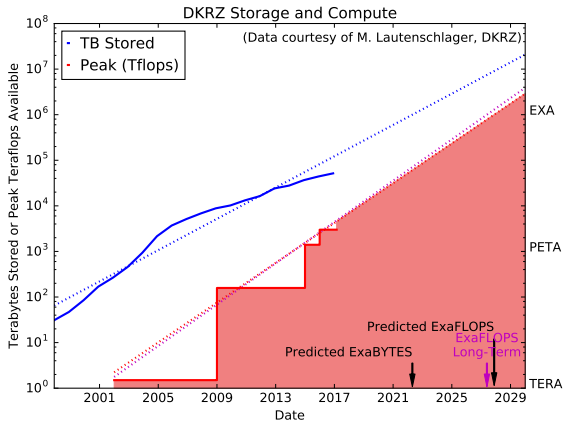


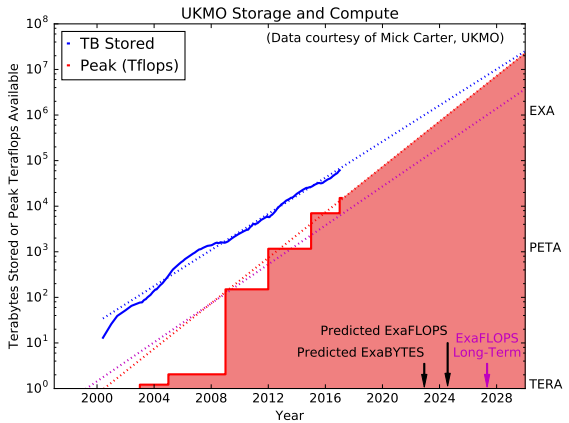




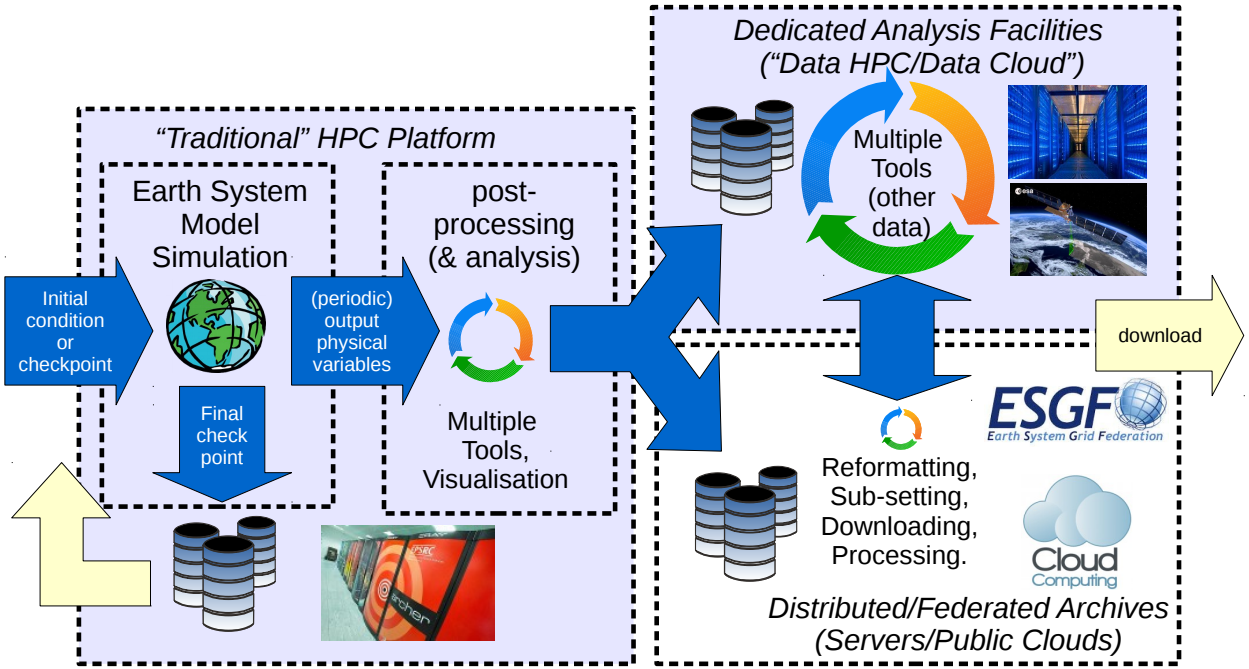
Consequences of increasing data …







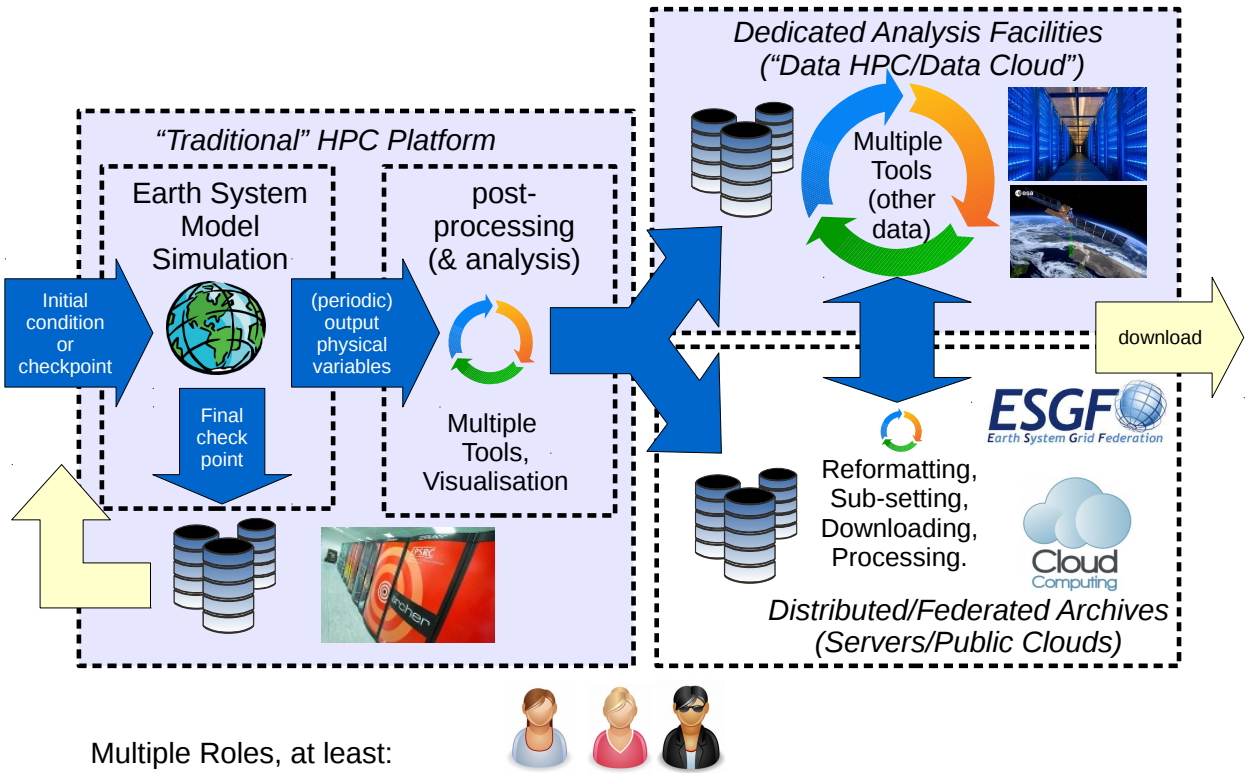
Heterogeneity in the Workflow Environment



Multiple Roles, at least:
Model Developer, Model Tinkerer, Expert Data Analyst, Service Provider, Data User

A range of data handling challenges — there will not be one ring to rule them all!

Heterogeneity in the Workflow Environment



Multiple Roles, at least:
Model Developer, Model Tinkerer, Expert Data Analyst, Service Provider, Data User

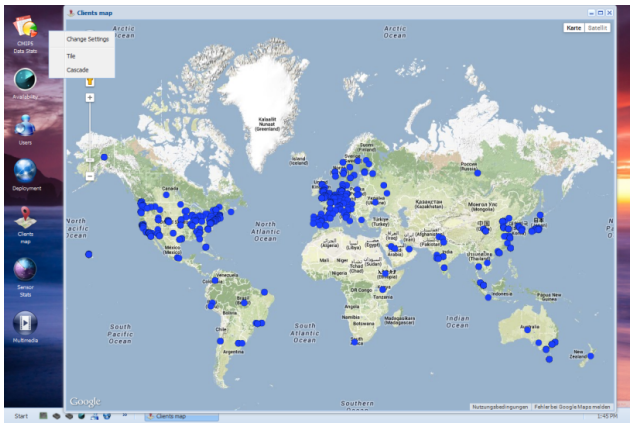
A range of data handling challenges — there will not be one ring to rule them all!

Challenges

1. Not all necessarily on one site!
2. Speed up checkpoint handling (initialisation, intermediate) (burst buffers, I/O servers)
3. Speed up, minimise output, for analysis (in-situ analysis not a *sufficient* condition; burst buffers, I/O servers).
4. Efficient data analysis (optimise, algorithm changes)
5. Disseminate products (not data)

The consequences of data at scale — download doesn't work!

Earth System Grid Experience



Slide content courtesy of Stephan Kindermann, DKRZ and IS-ENES2

Started with **Individual End Users**

- Limited resources (bandwidth, storage)

Moved to **Organised User Groups**

- Organize a local cache of files
- Most of the group don't access ESGF, but access cache.

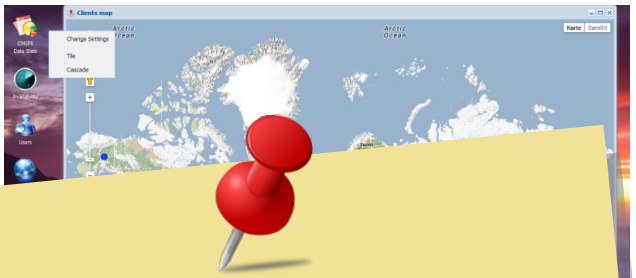
Then **Data Centre Services**

- Provide access to a replica cache
- May also provide compute by data
- CEDA, DKRZ, etc

Trend from download at home, to exploit a cache, to exploit a managed cache with compute!

The consequences of data at scale — download doesn't work!

Earth System Grid Experienc



Much of this was predictable, but CMIP5 was designed without much thought about how the data would be handled (and funded). Some aspects of CMIP6 will be better, but …**Climate Modellers need to take Data Management Plans more seriously!**

Slide content courtesy of Stephan Kindermann, DKRZ and IS-ENES2

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

Started with **Individual End Users**

- Limited resources (bandwidth, storage)

Then **Data Centre Services**

- Provide access to a replica cache
- May also provide compute by data
- CEDA, DKRZ, etc

Trend from download at home, to exploit a cache, to exploit a managed cache with compute!

NERC HPC

NERC Supercomputing

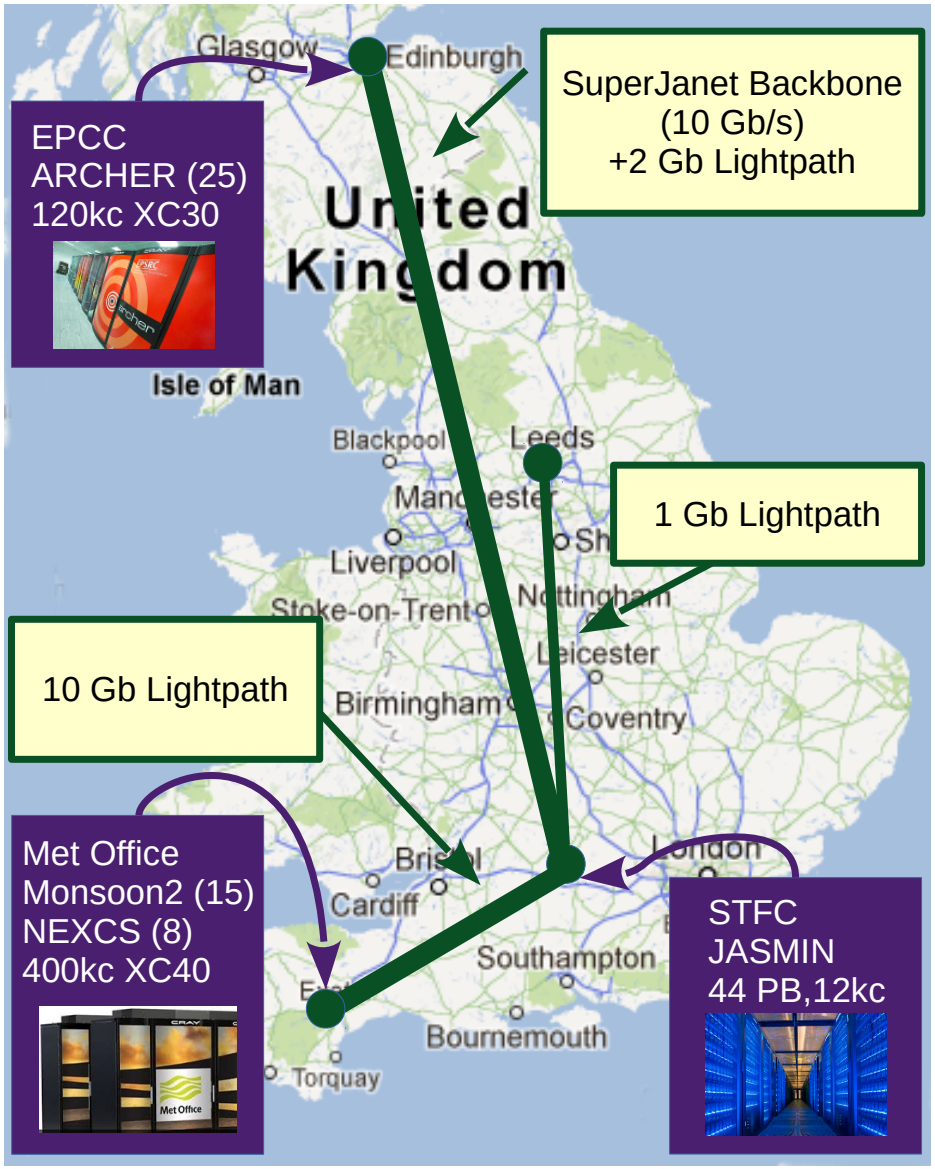
Three Simulation Platforms:

- ARCHER (EPCC in Edinburgh, roughly quarter of the machine)
- Monsoon2 and NEXCS (UKMO in Exeter, similar size resource to ARCHER, much bigger platform).

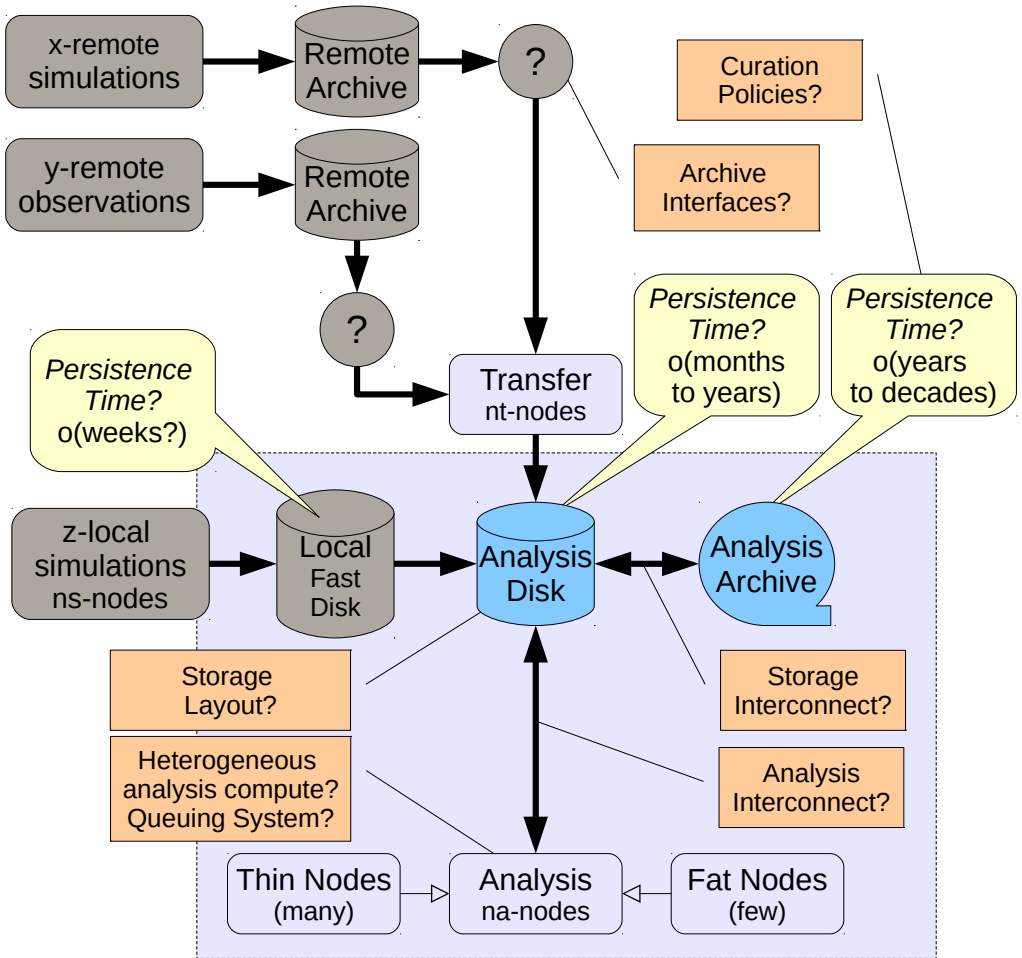
One Analysis and Archive Platform:

- JASMIN (44 PB of spinning disk plus 12K cores plus tape)

Fast *and* reliable *and* fat network links!

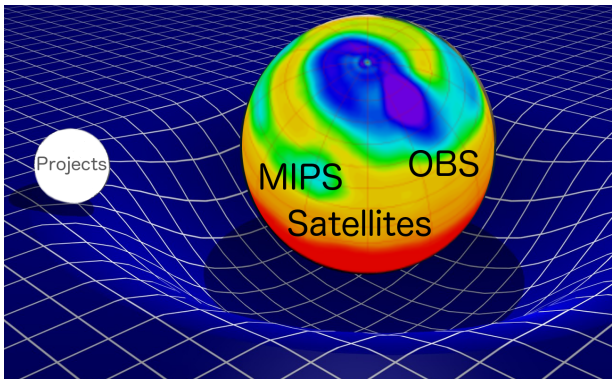


An abstract view



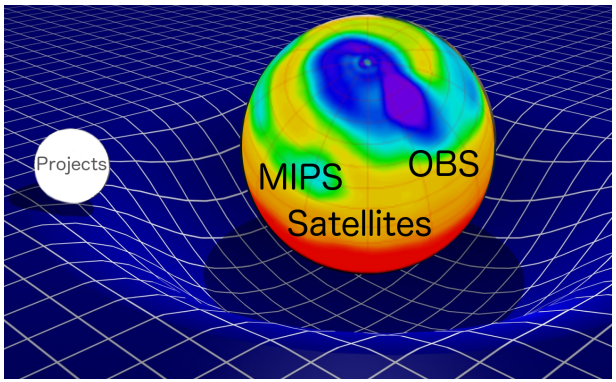
- (Potentially) many different remote simulation sources. How long can the data remain at source?
- Interesting problems moving the data to a common location?
- How long can the data reside on disk at the analysis location? What about in the archive?
- How should we best organise the data?
- What are the best ways to organise analysis compute?
- What are the best ways to address analysis interconnect and I/O bandwidth?

JASMIN — The Data Commons



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

JASMIN — The Data Commons



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

e.g. CEMS

Platform as a Service

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

e.g. BIOLINUX

Infrastructure as a Service

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

e.g. OPeNDAP/ FTP/ OptiRAD/ S3

Software as a Service

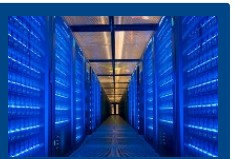
We provide you with REMOTE access to data VIA web and other interfaces.

CEDA Archives

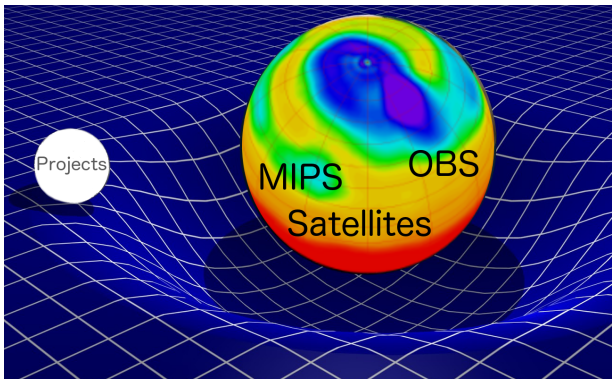
JASMIN – Data Intensive Computer

Storage, Compute and Network Fabric

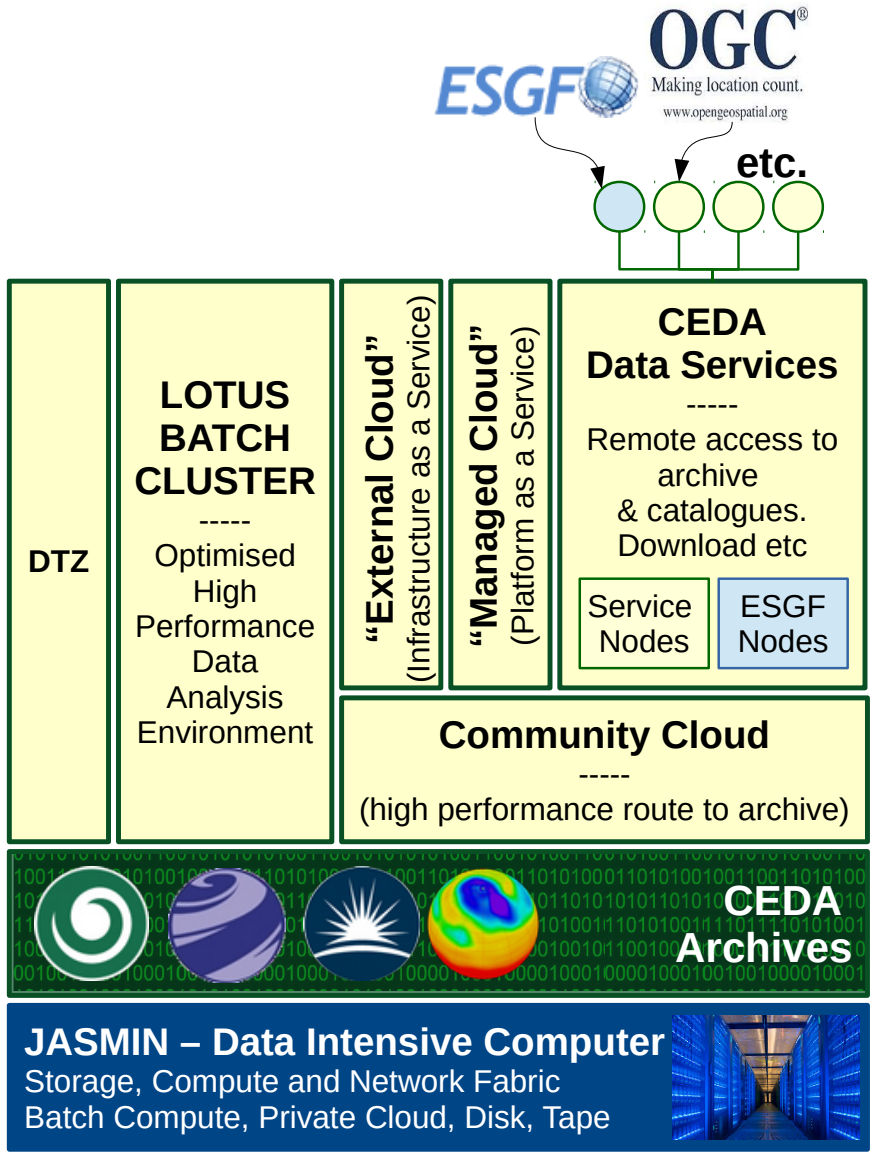
Batch Compute, Private Cloud, Disk, Tape



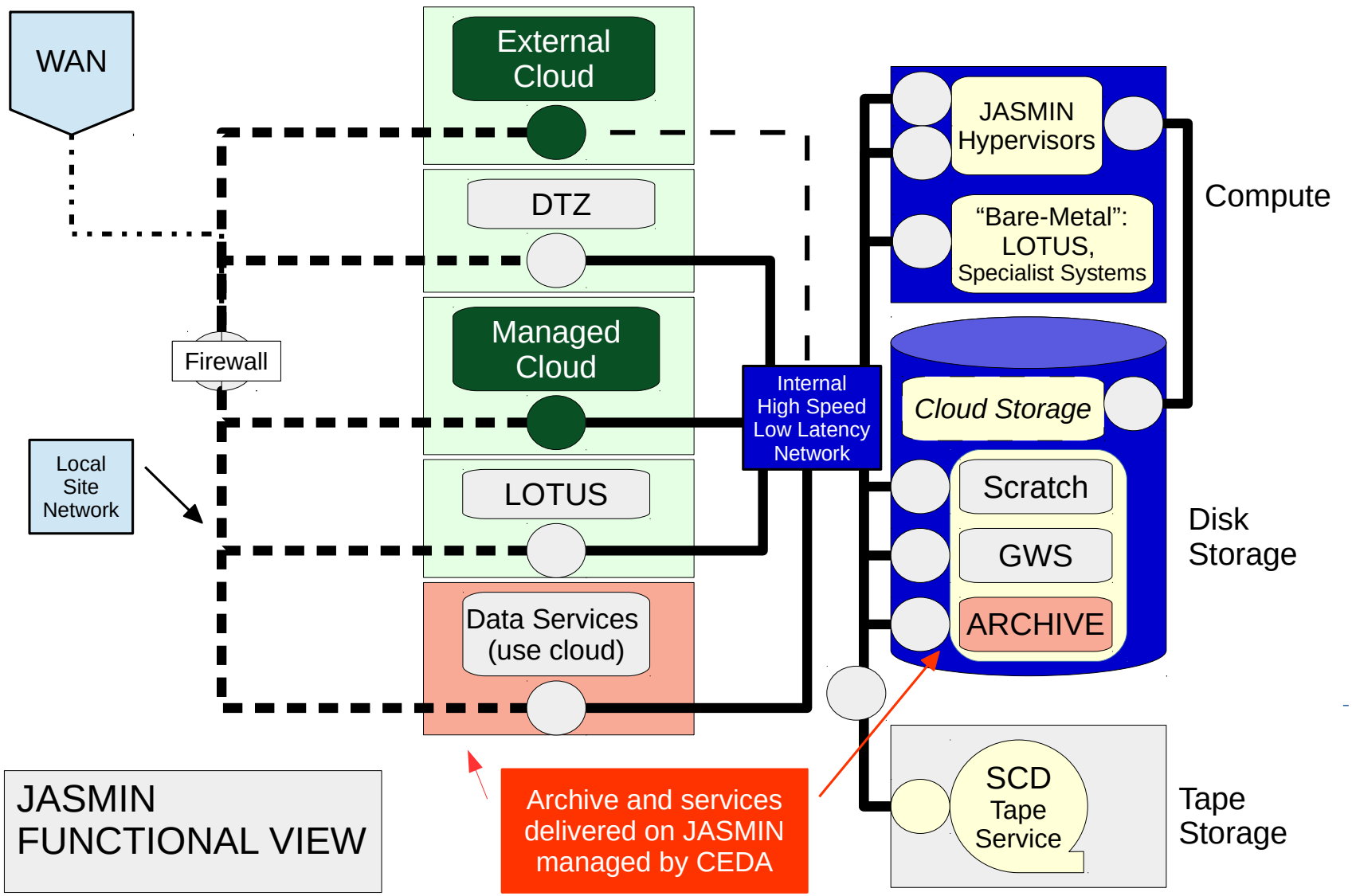
JASMIN — The Data Commons



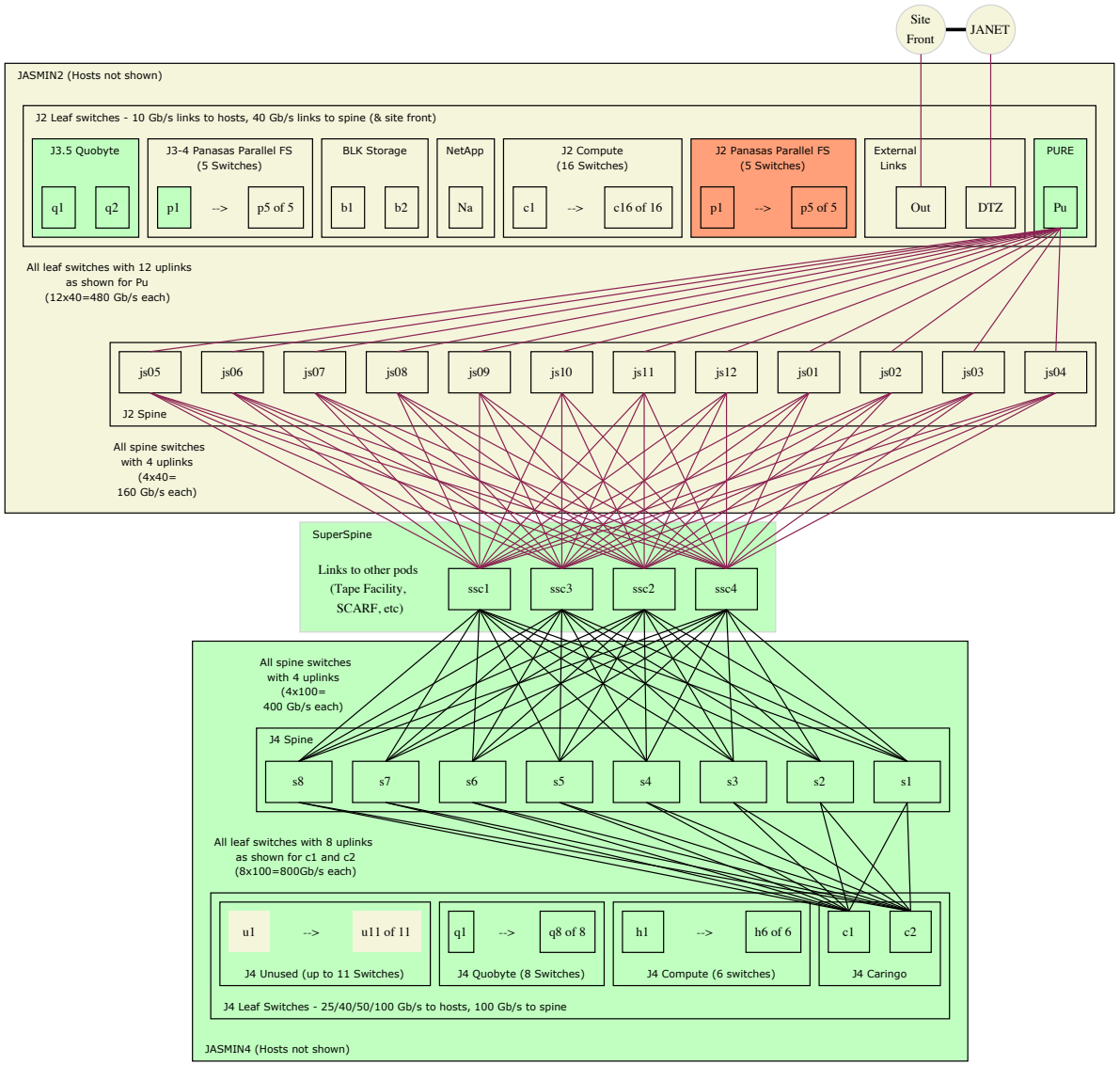
- Provide a state-of-the art storage and computational environment
- Provide and populate a managed data environment with key datasets (the "archive").
- Encourage and facilitate the bringing of data and/or computation alongside/to the archive!
- Provide FLEXIBLE methods of exploiting the computational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



JASMIN Functional View

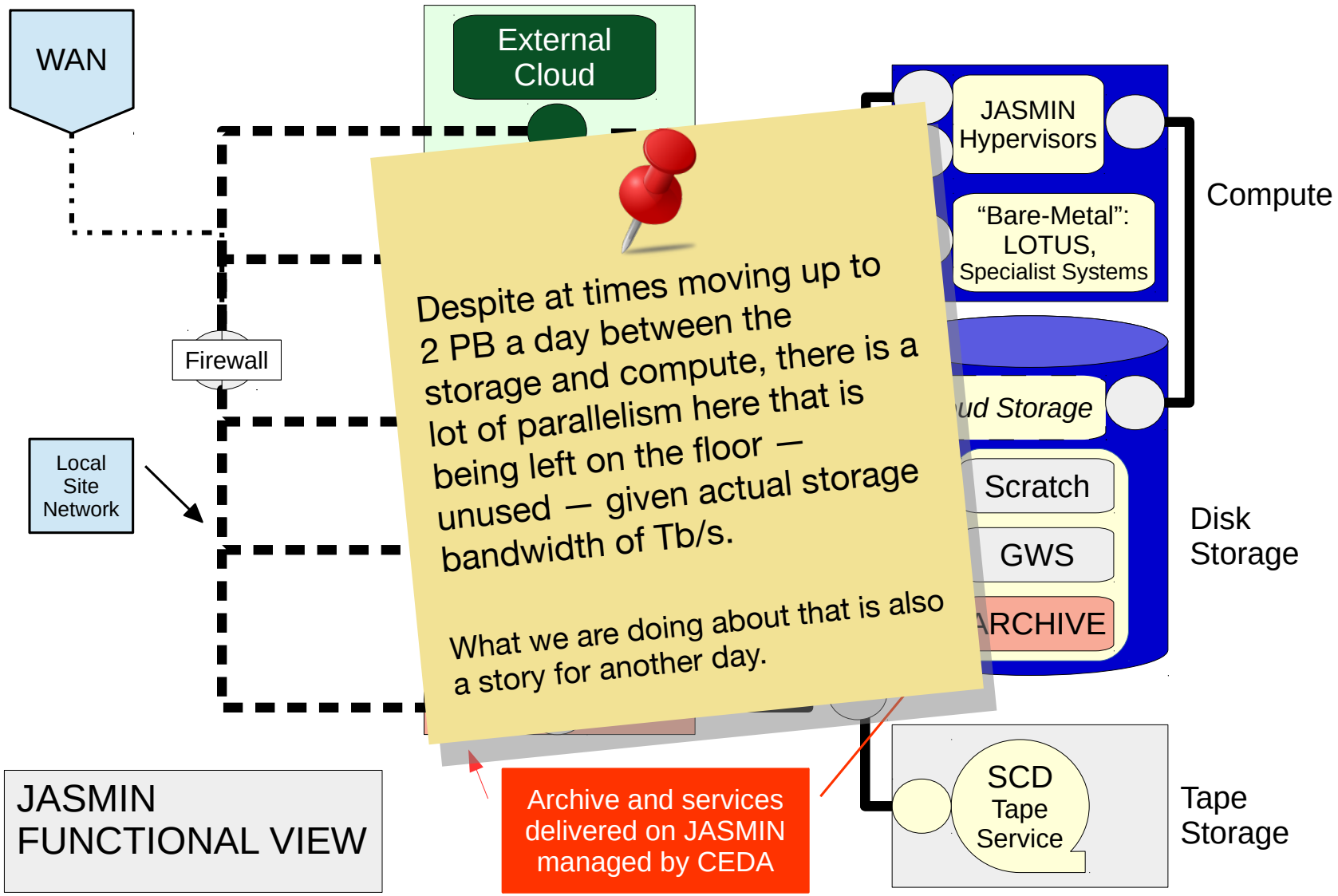


JASMIN Internal Network

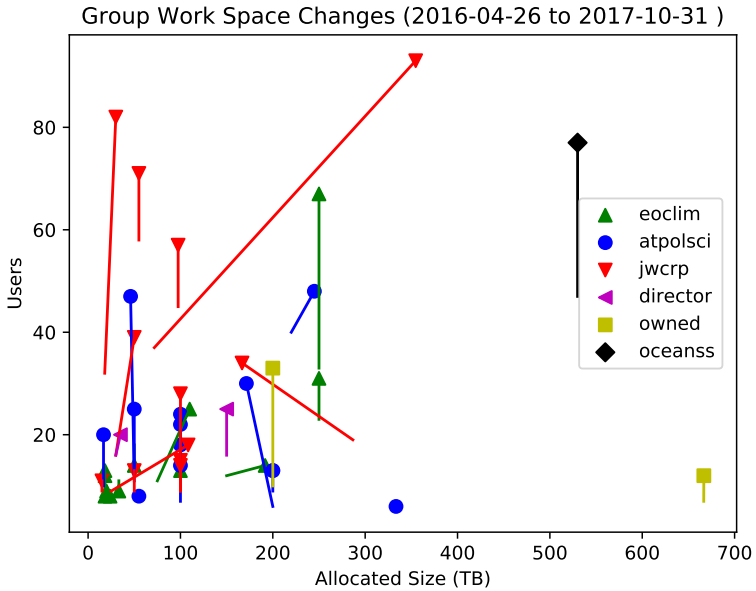


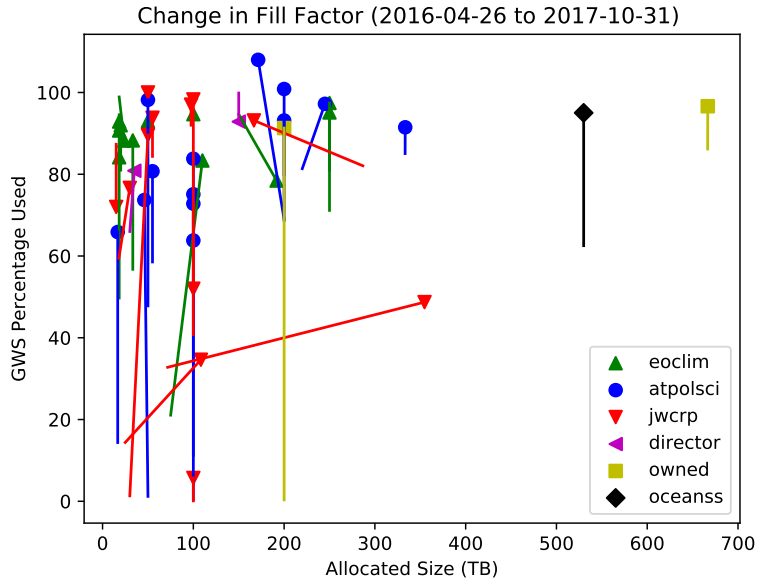
- Pod design with five layer CLOS network connecting pods via a superspine.
- Evolving: JASMIN 2 injection bandwidth into superspine ≈ 2 Tbit/s; JASMIN 4 $>$6 Tbit/s.
- (Inspired by Facebook)

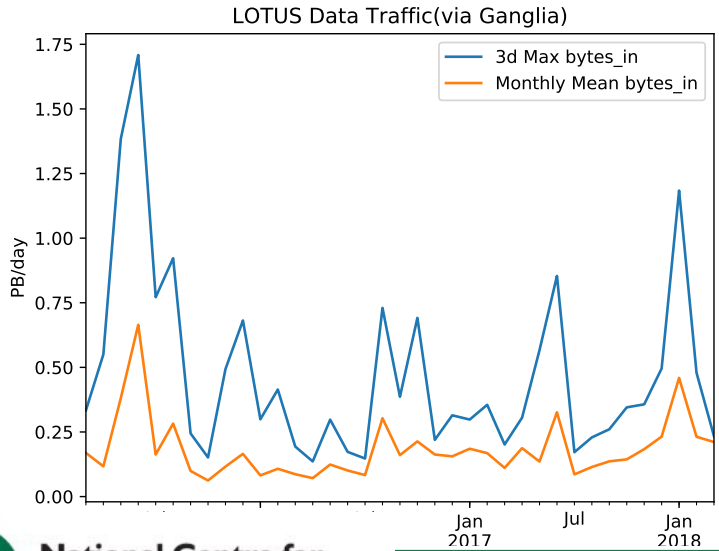
JASMIN Functional View



Users and Usage



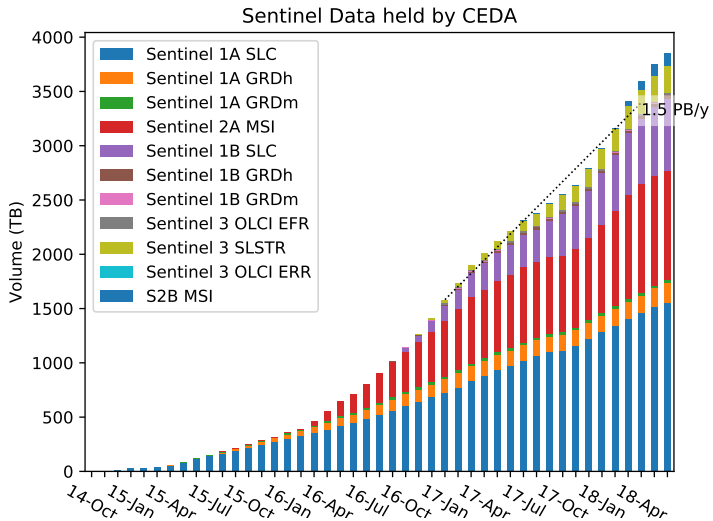


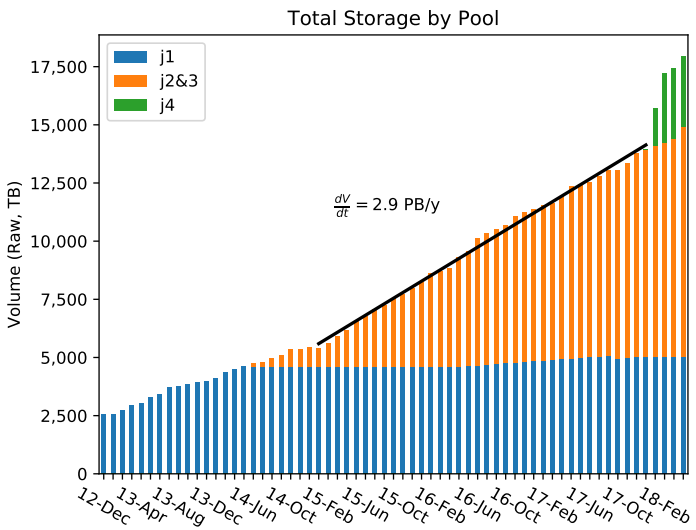


All users and selected GWS

- All commmunities filling available storage.
- Traffic dominated by specific use cases.
- Conclusion: Data storage and handling a pervasive problem!

I: Storage Problems: Infinite Disk?





Disk or Storage Growth?

- We will have exabytes at major centres soon.
- Even in JASMIN we have inexorable growth (looks linear, but isn't).
- Groups each have their own requirements for hot, warm, and cold data.
- How much online high performance disk is right?

JASMIN Phases 4 and 5

- Phase 4: 2017-18+, doubling disk storage, more types, more tiering.
- Phase 5: 2018-19+, new tape systems and new tape software.

II: Handling Problems: Better Software — Problem Space

Challenges in the domain of climate/weather

- Large data volume and high velocity
- Suboptimal performance & performance portability
  - Cannot properly exploit the hardware / storage landscape
  - Tuning for file formats and file systems necessary at the *application* level
- Data conversion is often needed
  - To combine data from multiple experiments, time steps, ...
- Data management practice does not scale & not portable
  - Cannot easily manage file placement and knowledge of what file contains.
  - Hierarchical namespaces does not reflect use cases.
  - Bespoke solutions at every site!

Issues addressed by the ESiWACE Earth System Data Middleware, ESDM

Challenges in the domain of climate/weather

- Large data volume and high velocity
- Suboptimal performance & performance portability
  - Cannot properly exploit the hardware
  - Tuning for file formats and file systems necessary at the *application* level
- Data conversion is often needed
  - To combine data from multiple experiments, time steps, ...
- Data management practice does not scale & not portable
  - Cannot easily manage file placement and knowledge of what file contains.
  - Hierarchical namespaces does not reflect use cases.
  - Bespoke solutions at every site!

(Namecheck: Jakob Luettgau, DKRZ)

Approach

Design Goals of the Earth System Data Middleware

1. Ease of use and deployment.
2. Relaxed access semantics, tailored to scientific data **generation**
   - Understand application data structures and scientific metadata
   - Reduce penalties of **shared** file access (i.e. deliver “lock-free” writes in parallel applications).
3. Site-specific (optimized) data layout schemes providing flexible mapping of data to multiple storage backends
4. Support for multiple data instances to support different **read** patterns.

Approach

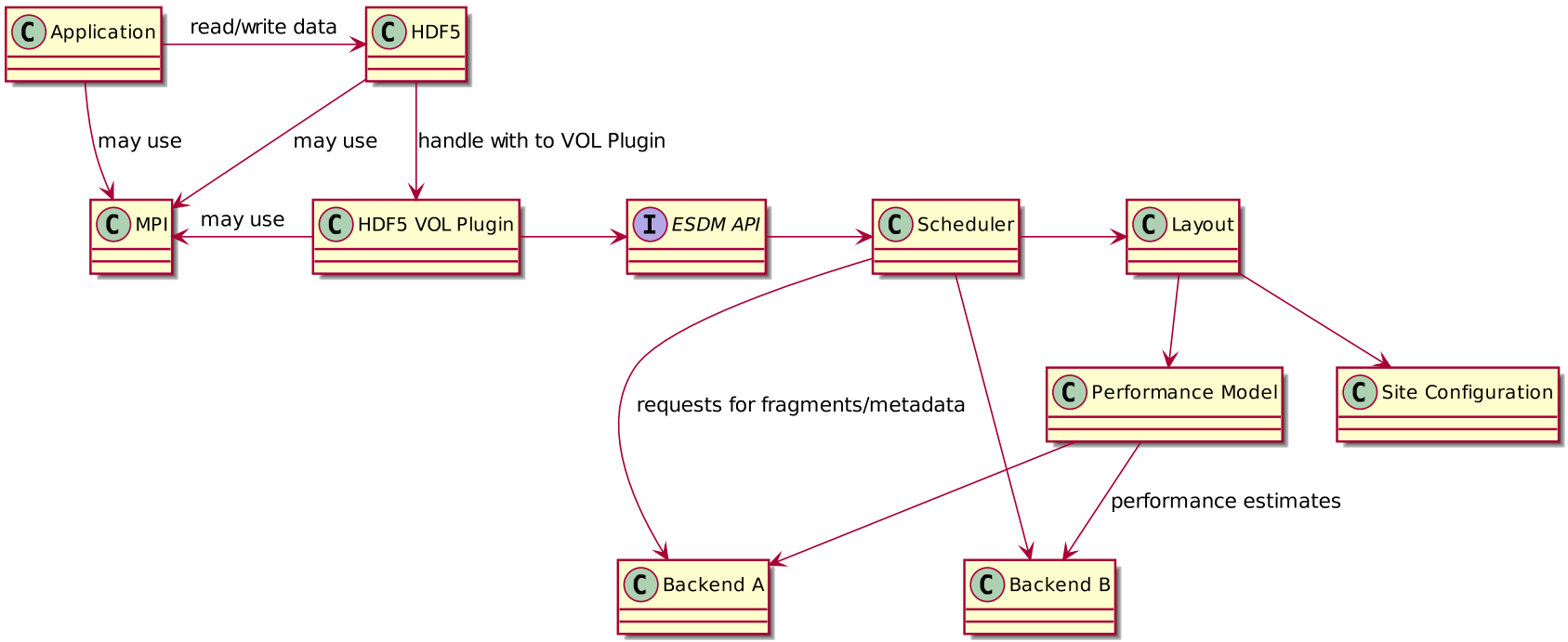
Design Goals of the Earth System Data Middleware

1. Ease of use and deployment.
2. Relaxed access semantics, tailored to scientific data **generation**
   - Understand application data structures and scientific metadata
   - Reduce penalties of **shared** file access (i.e. deliver “lock-free” writes in parallel applications).
3. Site-specific (optimized) data layout schemes providing flexible mapping of data to multiple storage backends
4. Support for multiple data instances to support different **read** patterns.

Architecture

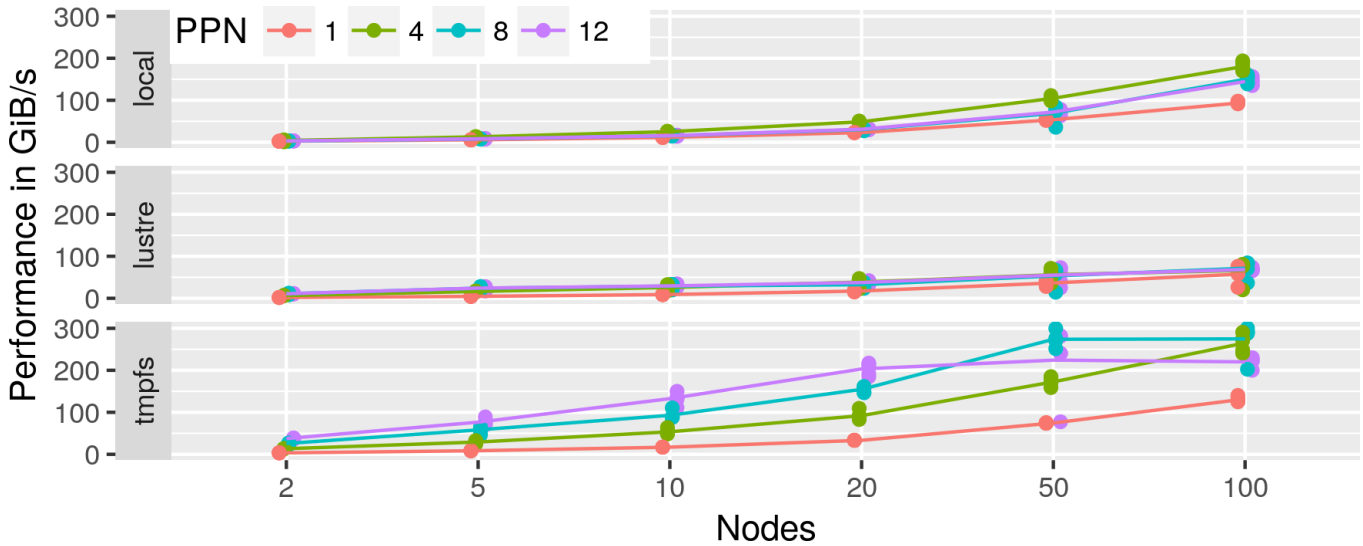
- Middleware: Supports any application that links to a customised version of a normal HDF library (including using NetCDF4 etc)
- A “layout component” lies between the HDF interface and the storage, allowing data to be optimally written using information about local storage components and (limited) information about performance.
- Tools for ingress and to create regular HDF/NetCDF4 on egress.

Architecture supports backend Specific optimization



Interplay of a IO scheduler, a layout component and storage specific performance models.

Status: Prototype exists and performance evaluations underway

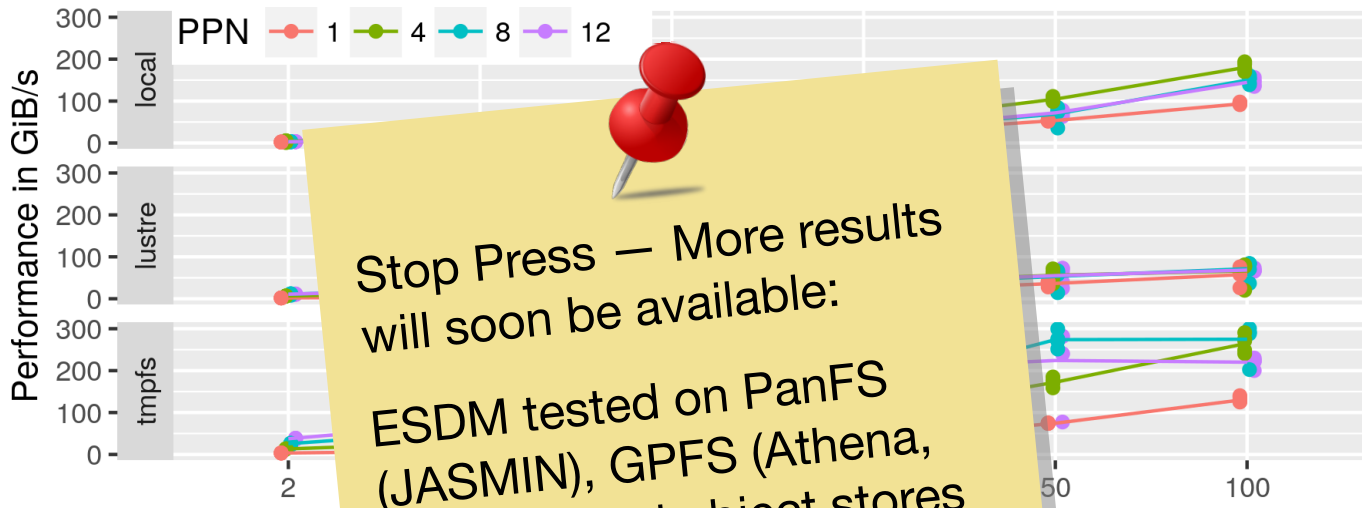


(Fixed Write, 26.8 GiB,coloured lines show mean values for proccesses per node, individual points showing five individual repetitions of each test.)

POSIX - DKRZ's MISTRAL

- Three storage types: *Lustre*, Memory (*tmpfs*), and *local* node SSD.
- Write results as expected: huge advantage *at scale* of local memory writes (up to ≈300Gb/s), signifcant advantage to local disk writes (up to ≈200Gb/s, poorest performance using traditional parallel file system (<100 Gb/s).

Status: Prototype exists and performance evaluations underway



(Fixed Write, 26.8 GiB,coloure … per node, individual points show … t.)

Stop Press — More results will soon be available:

ESDM tested on PanFS (JASMIN), GPFS (Athena, CMCC) , and object stores (DDN WOS at CMMCC, and Seagate Clovis at Seagate).

(ESiWACE collaborators: Sandro Fiore, Alessandro D'Anca and Huang Hua.)

POSIX - DKRZ's MISTR…

- Three storage types: … (*tmpfs*), and *local* node SSD.
- Write results as expected: huge advantage *at scale* of local memory writes (up to ≈300Gb/s), signifcant advantage to local disk writes (up to ≈200Gb/s, poorest performance using traditional parallel file system (<100 Gb/s).

Issues addressed by EWiWACE Semantic Storage ~~Library~~ Tools

Challenges in the domain of climate/weather

- Large data volume and high velocity
- Suboptimal performance & performance portability
  - Cannot properly exploit the hardware
  - Tuning for file formats and file systems necessary at the *application* level
- Data conversion is often needed
  - To combine data from multiple experiments, time steps, ...
- Data management practice does not scale & not portable
  - Cannot easily manage file placement and knowledge of what file contains.
  - Hierarchical namespaces does not reflect use cases.
  - Bespoke solutions at every site!

Namecheck: Neil Massey, STFC

Client Tools - Design Goals and Architecture Principles

Design Goals of the Semantic Storage ~~Library~~ Tools

1. Provide a portable library to address user management of data files on disk and tape which
   - does not *require* significant sysadmin interaction, but
   - can make use of local customisation if available/possible.
2. Exploit current and likely future storage architectures (tape, disk caches, POSIX and object stores).
3. Exploit existing metadata conventions.
4. Can eventually be backported to work with the ESDM.

Client Tools - Design Goals and Architecture Principles

Design Goals of the Semantic Storage ~~Library~~ Tools

1. Provide a portable library to address user management of data files on disk and tape which
   - does not *require* significant sysadmin interaction, but
   - can make use of local customisation if available/possible.
2. Exploit current and likely future storage architectures (tape, disk caches, POSIX and object stores).
3. Exploit existing metadata conventions.
4. Can eventually be backported to work with the ESDM.

Architecture: Exploit CF convention and CFA Framework

1. Fully general and based purely on CF metadata (`https://cfconventions.org`) and
2. CF Aggregation framework (`https://goo.gl/DdxGtw`).
3. Define how multiple CF fields may be combined into one larger field (or how one large field can be divided).

Two client tool components: S3NetCDF and CacheFace (working titles)

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Tiered Storage in User Tools

- Goal: Easy use of non-POSIX (especially Object) storage in existing workflows.
- Solution: Drop in replacement for NetCDF4-python — **S3NetCDF**
- Status: Prototype exists.
- Exploits CF aggregation to store an aggregated view of sub-files in a NetCDF file using **JSON** string content to point at aggregated files (which could be objects in an OS).
- Fragmentation opaque to user if desired!

Portable tool for Data Management

- Goal: Portable tool for users to manage data migration to less accessible storage tiers but maintain semantic information of stored content (beyond filenames).
- Solution: New command line tool: **CacheFace** (?name?) which includes both migration and catalog sub-components.
- Status: The data migration component (JASMIN Data Migration App, JDMA) is going operational at JASMIN shortly — work to be done to get into userspace. Catalog in FY19/20.

Two client tool components: S3NetCDF and CacheFace (working titles)

Tiered Storage in User Tools

- Goal: Easy use of n… R… (especially Object) existing workflows.
- Solution: Drop in re… for NetCDF4-pythor… **S3NetCDF**
- Status: Prototype ex…
- Exploits CF aggregat… store an aggregated … sub-files in a NetCDF… **JSON** string content t… aggregated files (whic… could be objects in an OS).
- Fragmentation opaque to user if desired!

Underlying Philosophy from JASMIN experience: Need to hand control of storage tiering to users and user workflows, cannot rely on system environment to know what should be where, and how to optimise use of storage.

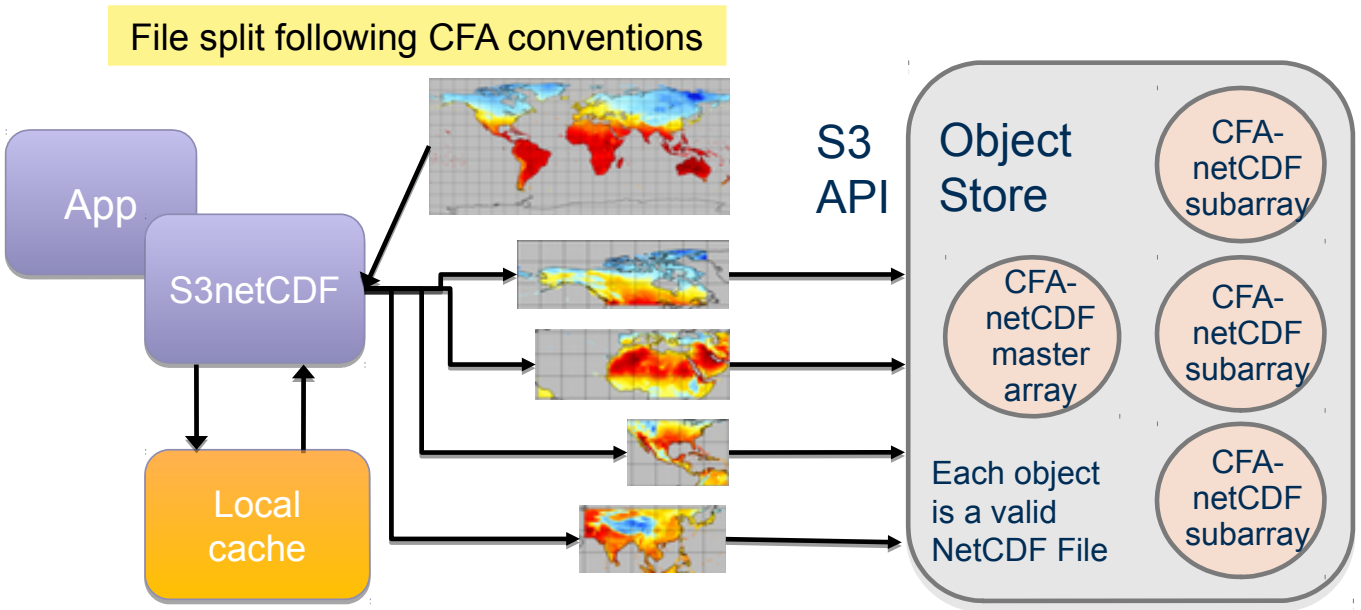
Portable tool for Data Management

…ortable tool for users to … data migration to less …ble storage tiers but … semantic information …d content (beyond …s).

New command line …**heFace** (?name?) …ludes both migration …g sub-components.

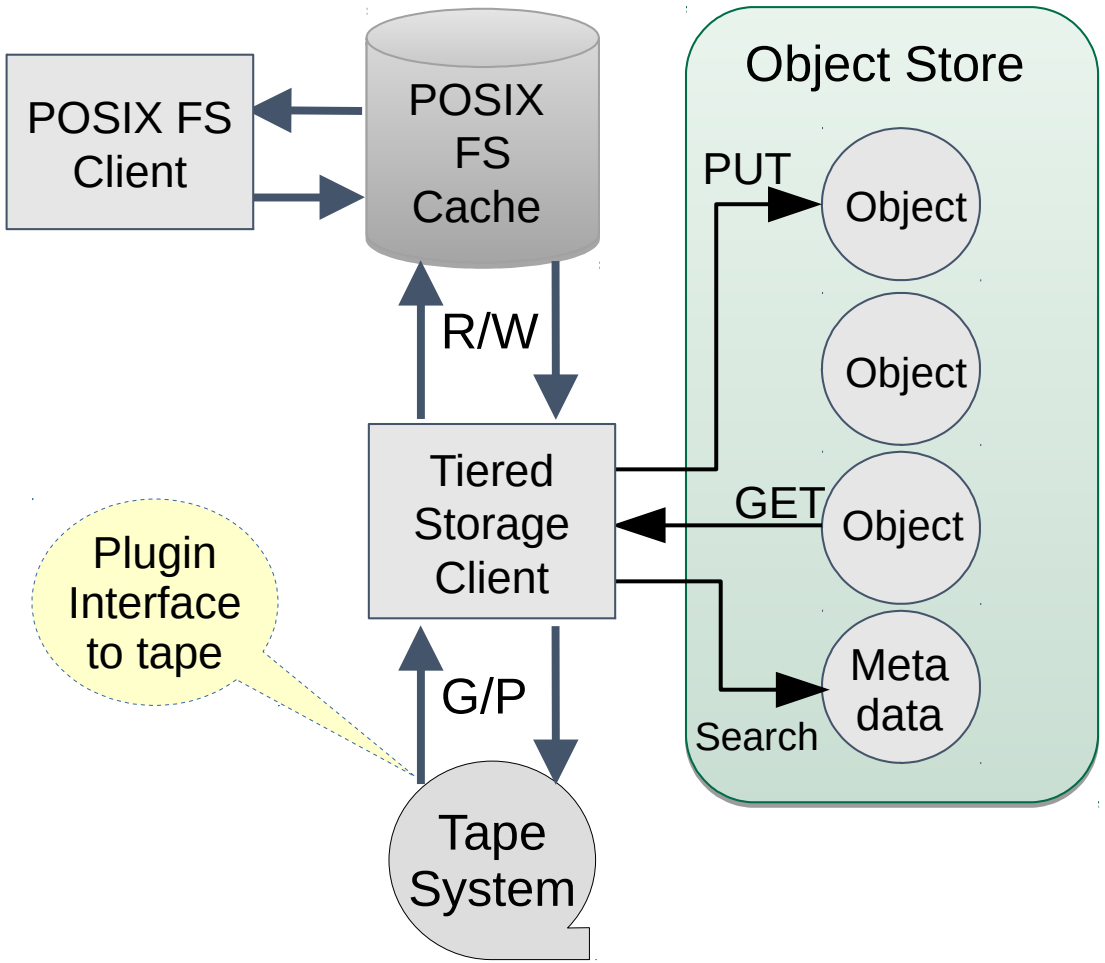
…e data migration component (JASMIN Data Migration App, JDMA) is going operational at JASMIN shortly — work to be done to get into userspace. Catalog in FY19/20.

S3NetCDF (working title)



- Master Array File is a NetCDF file containing dimensions and metadata for the variables (including URLs to fragment file locations).
- Master Array File can be in persistent memory or online, nearline, etc
- NetCDF tools can query file CF metadata content without fetching them.
- *Currently serial, work on parallelisation underway.*

CacheFace (working title)



CacheFace Status

Most pieces exist in at least prototype form.

- Simple metadata system designed.
- Cache system designed and prototype built that can use Minio interface to object store.
- Data migration component developed, and about to go operational (JDMA)
- Another cache system built which depends on our tape environment (ElasticTape).
- Work on integration and developing plugin concept with (portable) replacement for ElasticTape, to begin next year.

Summary: Two approaches to beating bottlenecks

Smarter Hardware

- Workflow demands customised data analyis environments, with
  - specialised hardware, and
  - user configurable software environments (virtualisation, containerisation, cloud).
- JASMIN is the current UK solution to these needs, but
- storage demands cannot be met by buying more disk alone, and
- More sophisticated parallel analysis software is needed for users!

Smarter Software

- Data volume and velocity need addressing throughout workflows.
- ESiWACE solutions are under development for
  - high performance middleware, and
  - user tools to support data migration and cataloging.
- We didn't have time for
  - Ensemble data handling (now in ESiWACE2),
  - ESDOC (in ISENES3),
  - Cluster as a Service and Notebook Services (JASMIN)

…but the science community has to own some of this problem too: everyone needs better a priori understanding of data handling issues!