Supercomputers are no longer all the same and it will get worse; a climate modelling perspective.

Bryan Lawrence & a cast of thousands

NCAS &
University of Reading: Departments of Meteorology and Computer Science

UoR, 11 Feb 19

Outline

- An introduction to climate modelling …
- and the data handling workflow.
- The JASMIN super data computer, and some examples of JASMIN cloud usage.
- The end of Moore's Law
- What next? Maths, computer science, and some of our research directions.

We want to simulate our world

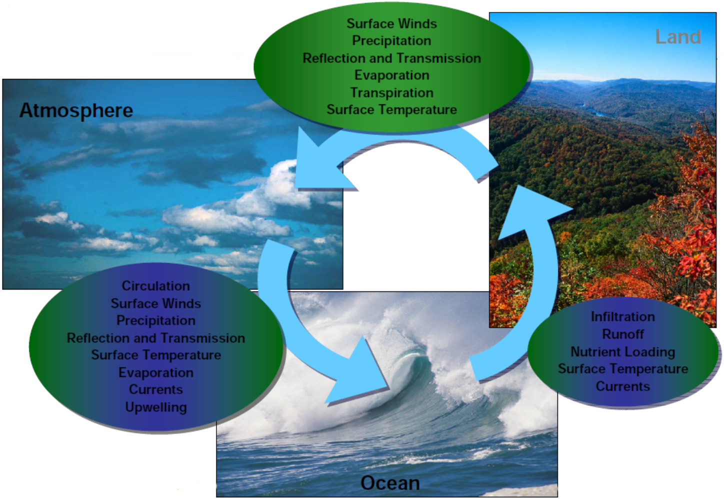


Image: from J. Lafeuille, 2006

Basic Fluid Equations (for the atmosphere)

State Variables:
u, v, w — wind
Π — Exner function (non-dimensional pressure)
Θ — Potential temperature

Coordinates:
r, ϕ, λ — radial position, latitude, longitude

Things that cause change:
$\frac{D}{Dt}$ — time derivative following motion
S — External Forcing (radiative heating etc)

Newton's second law

$$\frac{D_r u}{Dt} - \frac{uv\tan\phi}{r} - 2\Omega\sin\phi v + \frac{c_{pd}\theta}{r\cos\phi}\frac{\partial\Pi}{\partial\lambda} = -\left(\frac{uw}{r} + 2\Omega\cos\phi w\right) + S^u$$

$$\frac{D_r v}{Dt} + \frac{u^2\tan\phi}{r} + 2\Omega\sin\phi u + \frac{c_{pd}\theta}{r}\frac{\partial\Pi}{\partial\phi} = -\left(\frac{vw}{r}\right) + S^v$$

$$\frac{D_r w}{Dt} + c_{pd}\theta\frac{\partial\Pi}{\partial r} + \frac{\partial\Pi}{\partial r} = \left(\frac{u^2 + v^2}{r}\right) + 2\Omega\cos\phi u + S^w$$

mass continuity

$$\frac{D_r}{Dt}\left(\rho_d r^2\cos\phi\right) + \rho_d r^2\cos\phi\left[\frac{\partial}{\partial\lambda}\left(\frac{u}{r\cos\phi}\right) + \frac{\partial}{\partial\phi}\left(\frac{v}{r}\right) + \frac{\partial w}{\partial r}\right] = 0$$

thermodynamics

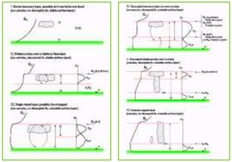
$$\frac{D_r\theta}{Dt} = S^\theta$$

Objective is given knowledge of the external forcing S and the state (u, v, w, Π, Θ) at time t, to advance knowledge of the state variables to time $t + \Delta t$, where Δt is the **timestep**.

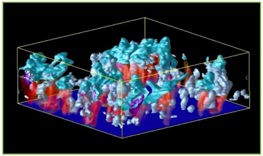
Many of the forcing terms come from parameterisations

Slide Images from Slingo, 2013

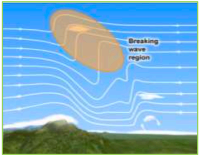
Boundary layer turbulence and mixing



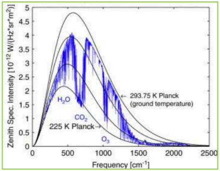
Cumulus convection



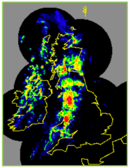
Effects of mountains



Radiation



Precipitation

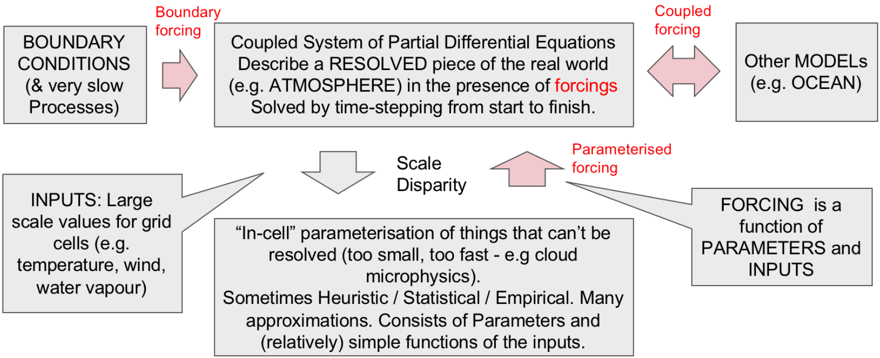


Clouds and microphysics

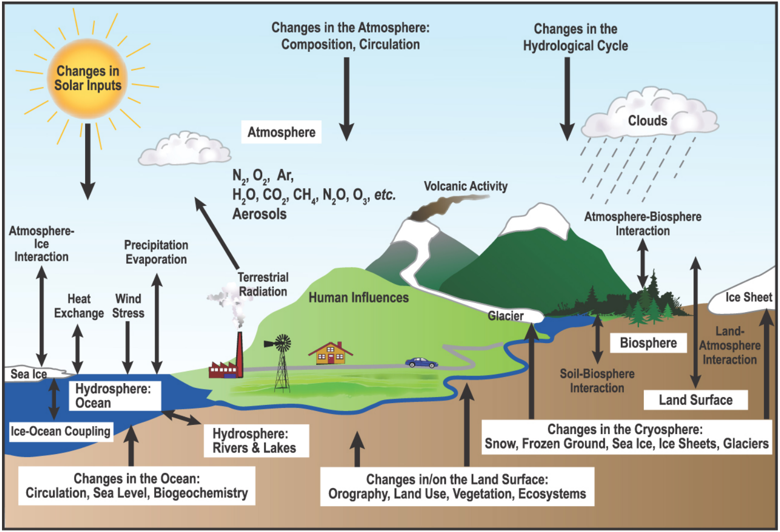
Atmospheric composition

Many sub-grid scale processes which have to be parameterised (that is, approximated, and their “grid-scale” affect is represented by functions of the grid-scale variables and some knowledge of the sub-grid, e.g. orography).

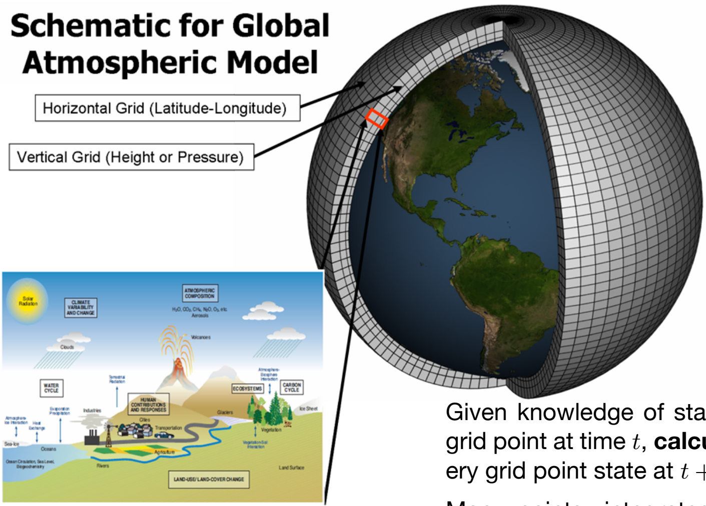
One slide introduction to numerical modelling



beyond the fluid atmosphere - Adding more processes



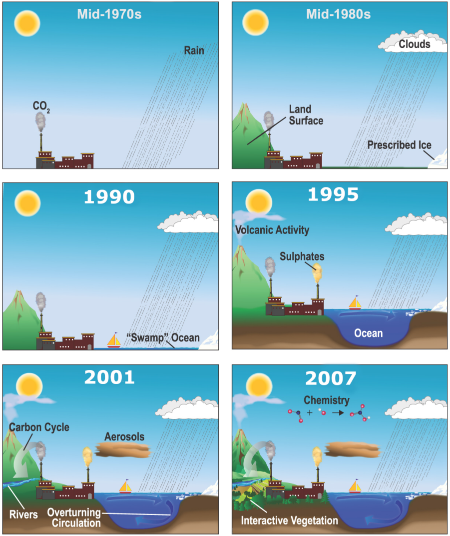
Everything is solved on a grid



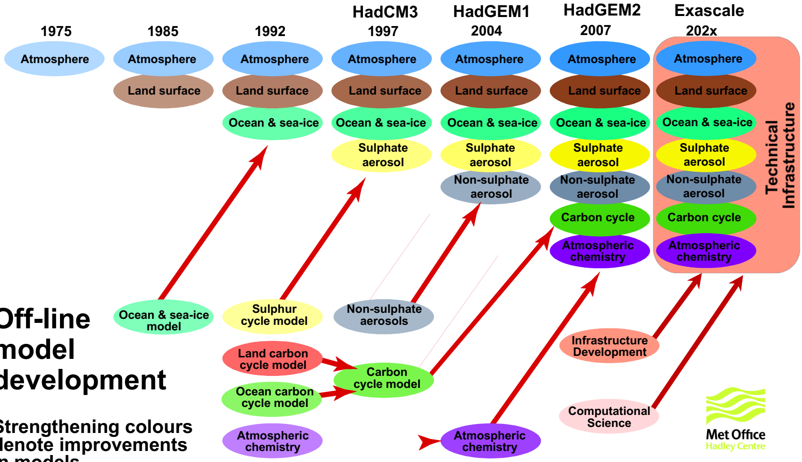
Given knowledge of state at every grid point at time t, **calculate** at every grid point state at $t + \Delta t$.

Many points, integrated for years with timestep of *o(minutes)*!

The Changing World in Climate Models



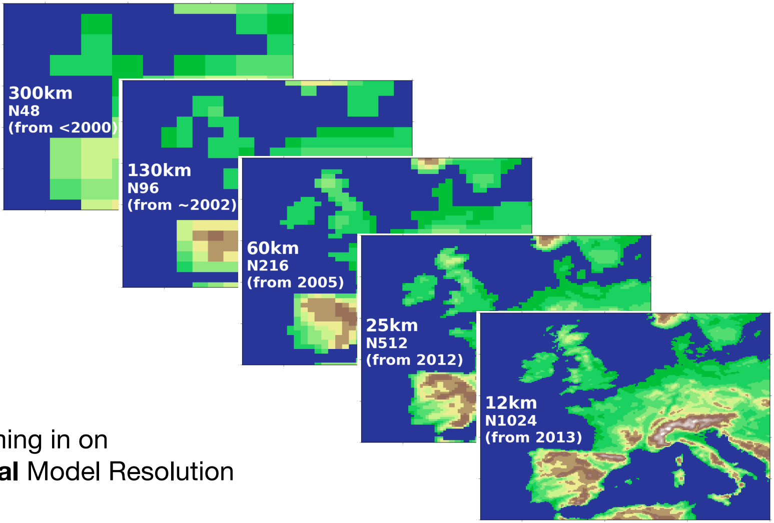
Evolution of Complexity



Off-line model development

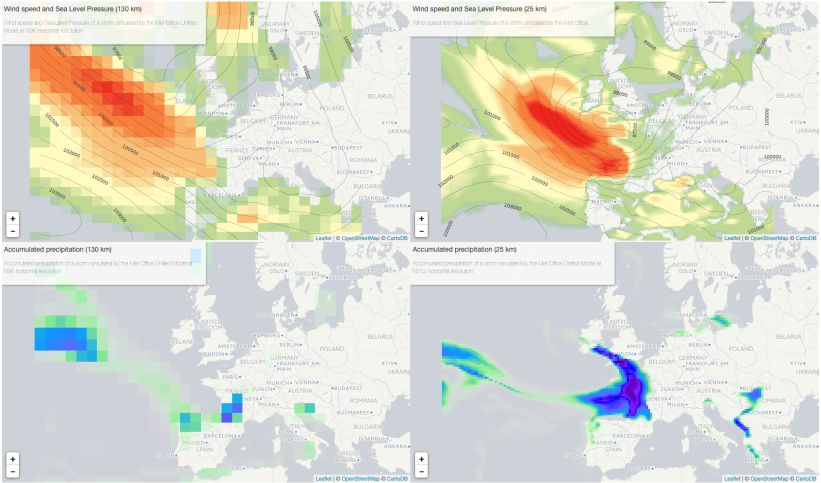
Strengthening colours denote improvements in models

The Evolution of Resolution: A better global microscope!



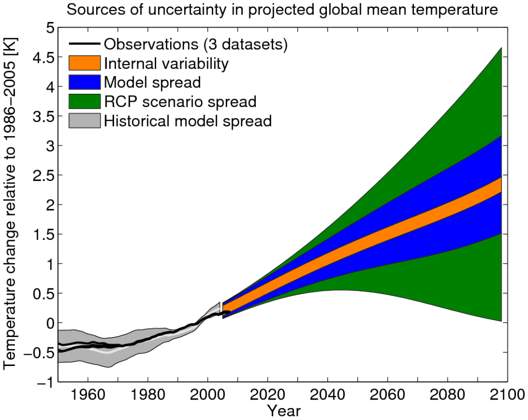
Zooming in on
Global Model Resolution

The influence of resolutioon on simulations of extratropical cyclones



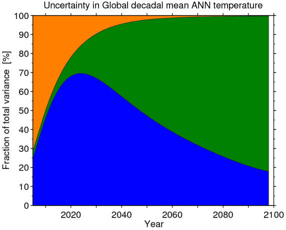
As simulated by the Met Office
https://uip.primavera-h2020.eu/storymaps/extra-tropical-cyclones

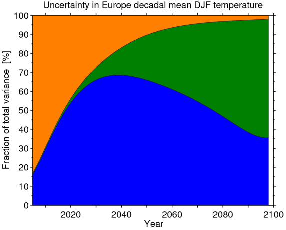
Global Climate Simulation Uncertainty as expressed in AR5



For the global big picture: model uncertainty is not the biggest problem: humanity chooses the pathway!

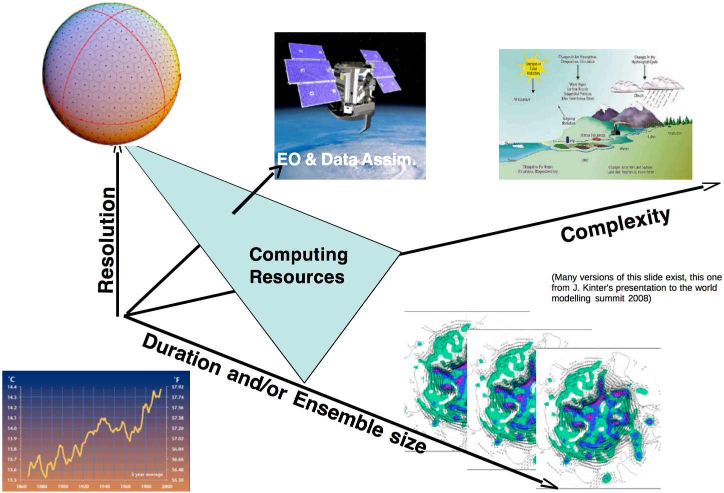
Source: Kirtman et.al., 2013: Near-term Climate Change: Projections and Predictability. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F. et.al. (eds.)]. Cambridge University Press.





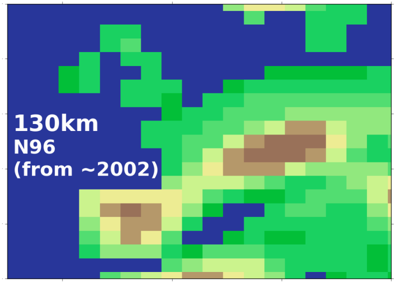
Models are more uncertain at regional scales.

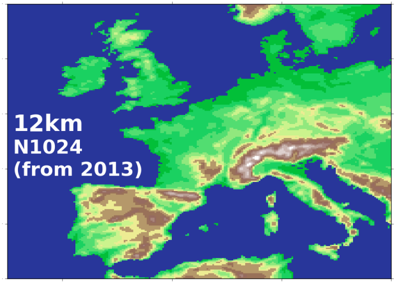
Give me more computing?



(Many versions of this slide exist, this one from J. Kinter's presentation to the world modelling summit 2008)

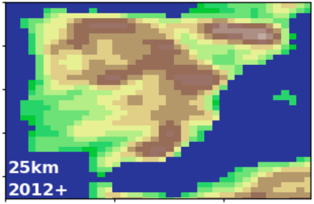
A modest (?) step ...

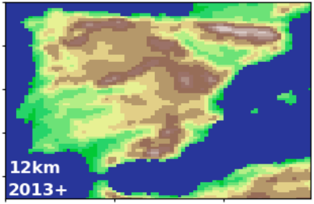


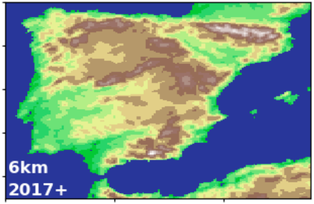


One "field-year" — 26 GB	One "field-year" — >6 TB
1 field, 1 year, 6 hourly, 80 levels 1 x 1440 x 80 x 148 x 192	1 field, 1 year, 6 hourly, 180 levels 1 x 1440 x 180 x 1536 x 2048

Volume — the reality of global 1km grids







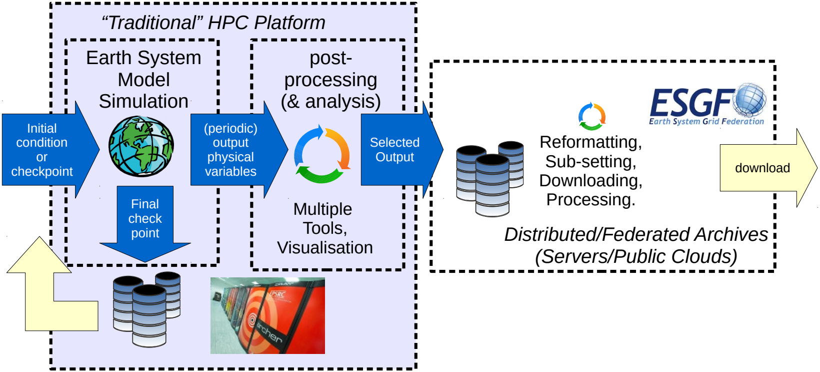
What about 1km? That's the current European Network for Earth System Modelling (ENES) goal!

Consider N13256 (1.01km, 26512x19884)):

- 1 field, 1 year, 6 hourly, 180 levels
- 1 x 1440 x 180 x 26512 x 19884 = 1.09 PB
- 760 seconds to read one 760 GB (xy) grid at 1 GB/s
- but it's worse that that: 10 variables hourly, $>$ 220 TB/day!

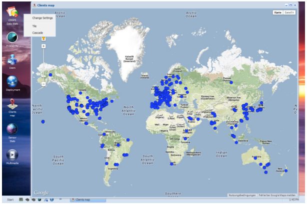
Can no longer consider serial diagnostics, and even parallelised is a challenge for the I/O system!

How we used to do it: from supercomputer to download



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

Earth System Grid Experience



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)

Moved to **Organised User Groups**

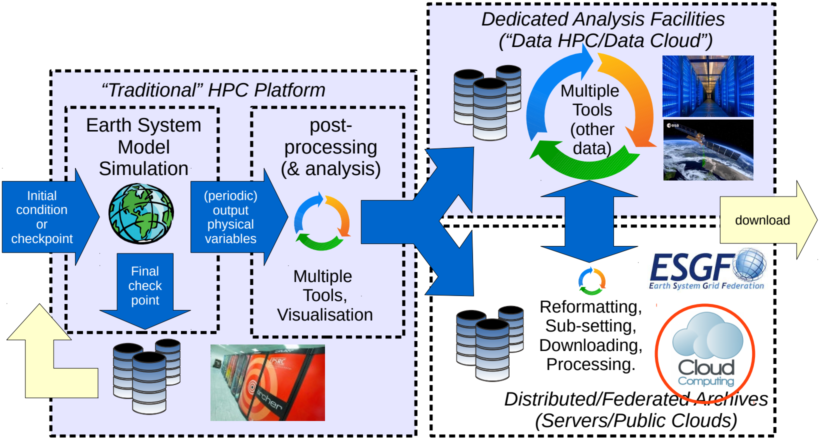
- Organize a local cache of files
- Most of the group don’t access ESGF, but access cache.

Then **Data Centre Services**

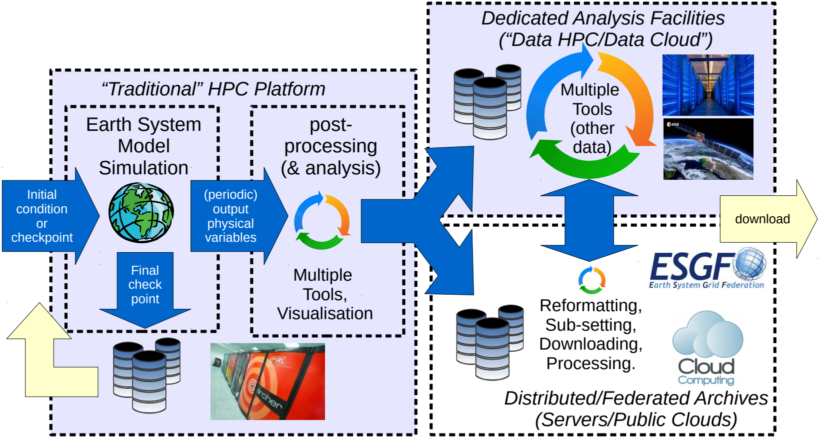
- Provide access to a replica cache
- May also provide compute by data
- CEDA, DKRZ, etc

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

Many different supercomputing environments



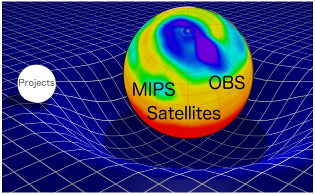
Many different supercomputing environments



Multiple Roles, at least:

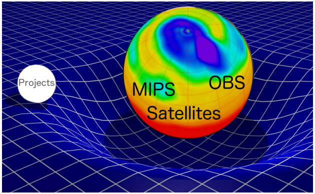
Model Developer, Model Tinkerer, Runner, Expert Data Analyst, Service Provider, Data Manager, Data User

JASMIN — 4 steps in exploiting data gravity to deliver a data commons

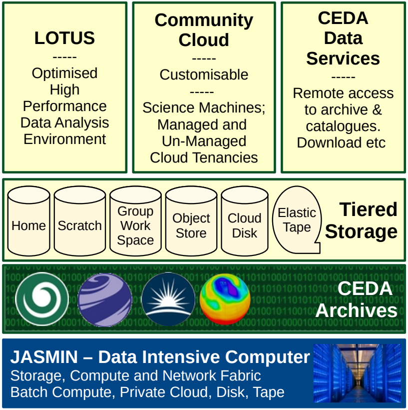


1. Provide and populate a managed data environment with key datasets (the "archive").
2. Encourage and facilitate the bringing of data and/or computation alongside/to the archive!

JASMIN — 4 steps in exploiting data gravity to deliver a data commons

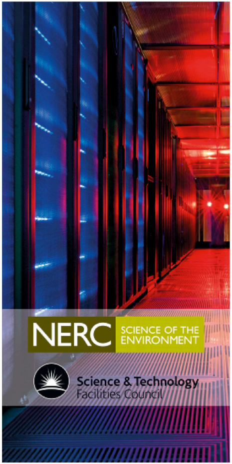


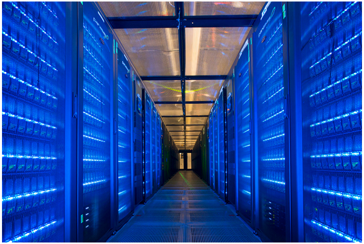
3. Provide a state-of-the art storage and computational environment
4. Provide FLEXIBLE methods of exploiting the computational environment.



1. Provide and populate a managed data environment with key datasets (the “archive”).
2. Encourage and facilitate the bringing of data and/or computation alongside/to the archive!

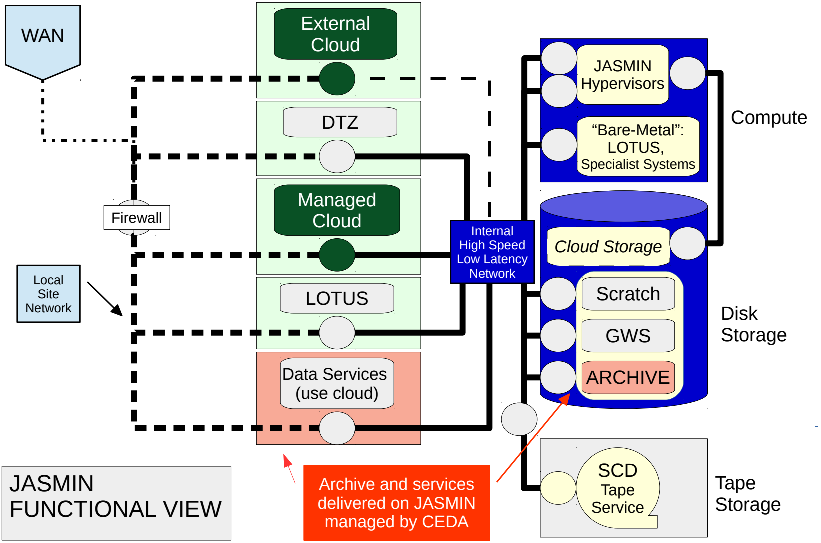
JASMIN: A Data Intensive Computing System





- Customised Fast Network.
- 44 PB Disk Storage.
- Tape Robot and “Elastic Tape Service”.
- 12000 compute cores:
  The “Lotus” batch cluster; hosted compute; cloud.
- Some high memory nodes. Some GPU systems from Q2 2019.

JASMIN: A functional View

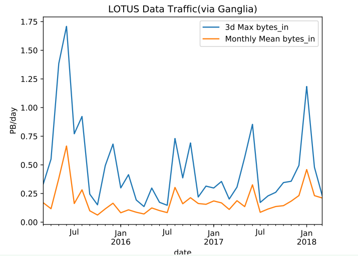


Lotus

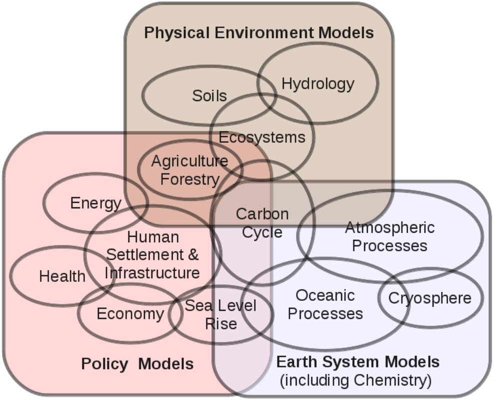
Traditional Batch Cluster

- (Feb'19):8100 cores, 5000 deployed to support single core jobs.
- Very mixed estate, with a range of processors and memory.

Untraditional Usage - Very large dataflows!



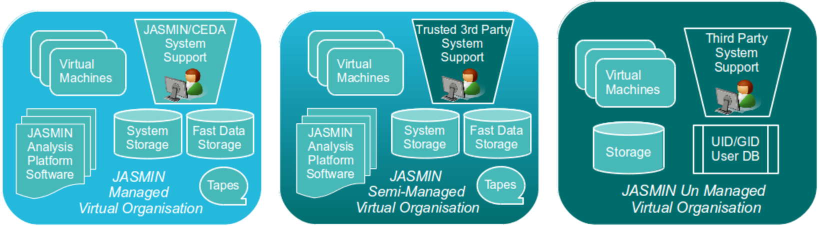
Communities



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

Figure adapted from Moss et al, 2010

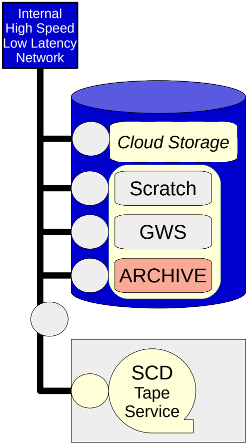
Virtual Compute and Virtual Organisations



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

Example: NCAS as as a big organisation can run a semi-managed virtual organisation (with multiple group work spaces), but large groups within NCAS can themselves setup a virtual organisation to run their own clusters in the un-managed environment.

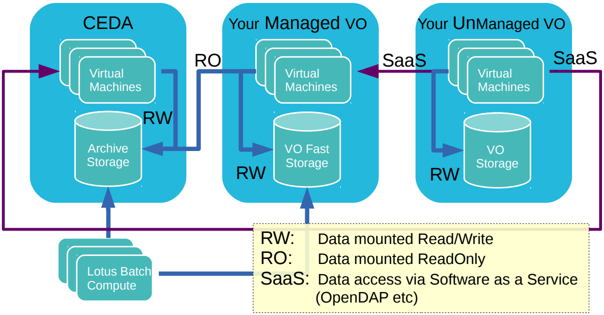
JASMIN Tiered Storage Requirements



There is not one storage system to rule them all

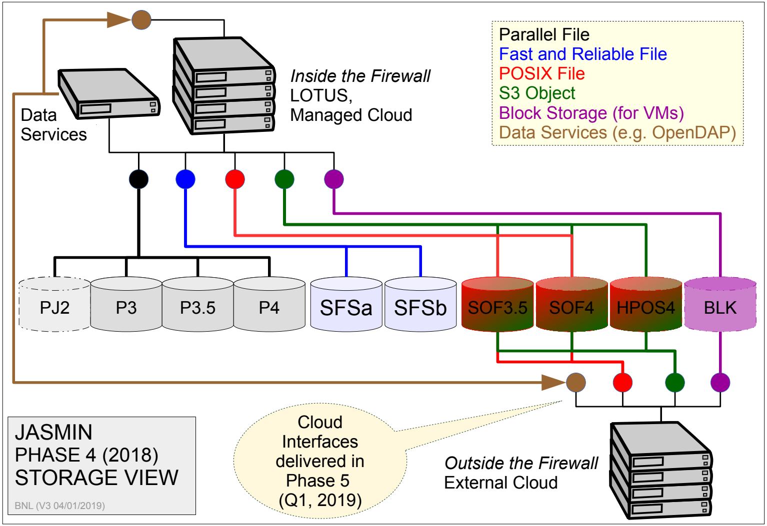
- Tape is (relatively) cheap. Tape is faster than you think. But tape latency is bad.
- Filesystems come with constraints: bandwidth, reliability, scalability, consistency, access control issues. You can't have it all!
- Cloud Storage:
  - Block storage: build their own file systems.
  - Object Storage: Scalable, simple, flexible access control.
- Shared file system requirements:
  - Scratch: fast, but trade-off between fast for large volume, and fast for small files.
  - Group Work Spaces: Community shared storage; not necessarily high performance.
  - Archive: long-term persistent, shared access, reliable.

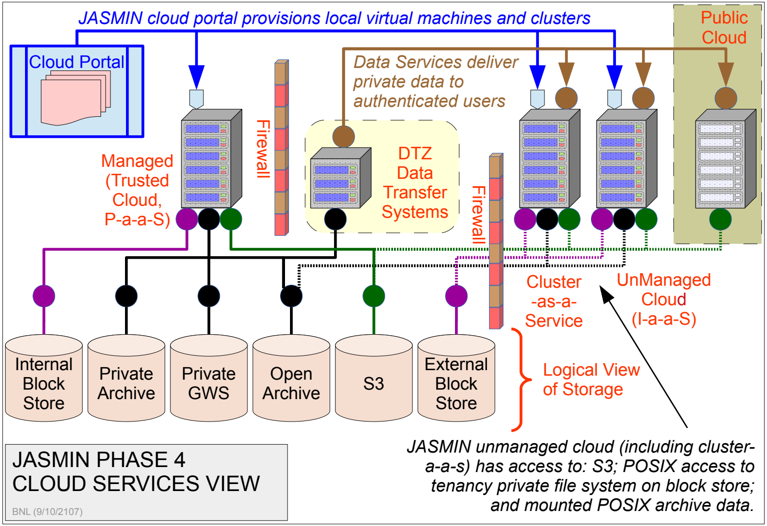
Objective is to provide an environment with high performance access to curated data archive **and** a high performance data analysis environment!



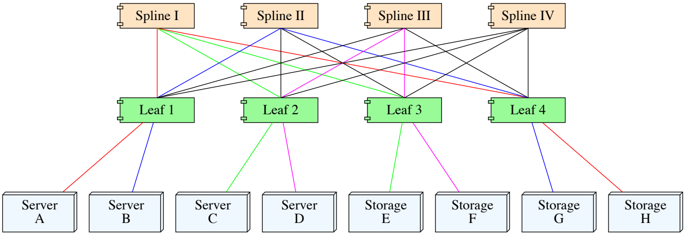
CEDA is one virtual organisation within o(100) such virtual organisations. Key issues include:

- how to provide high performance data access in the managed environment for multiple users, multiple workflows, intersecting in some of the data, and
- between unmanaged (infrastructure as a service) and the data held in the managed environment.


Parallel File
Fast and Reliable File
POSIX File
S3 Object
Block Storage (for VMs)
Data Services (e.g. OpenDAP)
Inside the Firewall
LOTUS,
Managed Cloud
Data
Services
PJ2
P3
P3.5
P4
SFSa
SFSb
SOF3.5
SOF4
HPOS4
BLK
Cloud
Interfaces
delivered in
Phase 5
(Q1, 2019)
Outside the Firewall
External Cloud
JASMIN
PHASE 4 (2018)
STORAGE VIEW
BNL (V3 04/01/2019)


JASMIN cloud portal provisions local virtual machines and clusters
Cloud Portal
Public Cloud
Data Services deliver private data to authenticated users
Firewall
Managed (Trusted Cloud, P-a-a-S)
DTZ Data Transfer Systems
Firewall
Cluster -as-a- Service
UnManaged Cloud (I-a-a-S)
Internal Block Store
Private Archive
Private GWS
Open Archive
S3
External Block Store
Logical View of Storage
JASMIN unmanaged cloud (including cluster-a-a-s) has access to: S3; POSIX access to tenancy private file system on block store; and mounted POSIX archive data.
JASMIN PHASE 4
CLOUD SERVICES VIEW
BNL (9/10/2107)

An introduction to CLOS networks



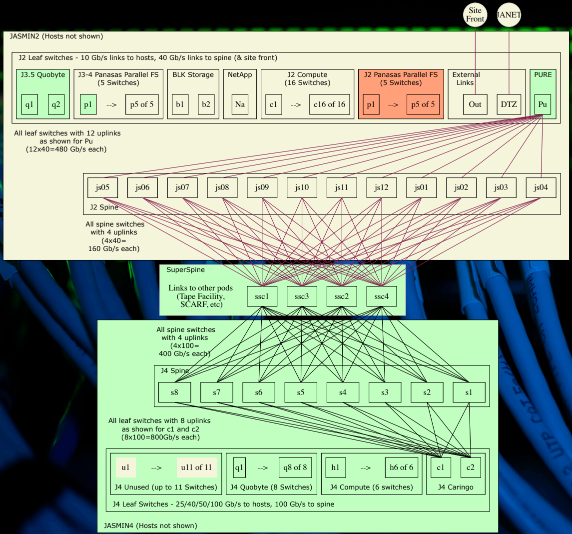
Why? We want:

- Any part of the network to be able to talk to any other part of the network: "East-West" (rather than "North-South" aka server-client).
- Predictable, affordable performance. Scalability.
- Low, but not extremely low latency (allowing more smaller switches, rather than fewer bigger switches).

E.g: Three-Layer CLOS network

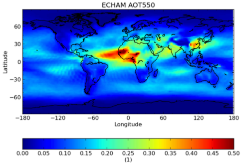
- With r links into each leaf, there needs to be r leaves and r spines for non-blocking links.
- In a blocking network, there are less uplinks into the spine than there are uplinks into the leaves (less spine switches than leaf switches)
- In this case, the leaves are under-populated, We could support two more systems per leaf switch.
- Could scale by adding more leaf and spine switches (and more servers per leaf) up r of each (the maximum r-links supported by each switch) …then more layers.

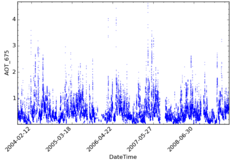
JASMIN Internal Network

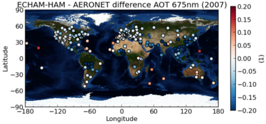


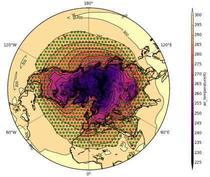
- Pod design with five layer CLOS network connecting pods via a superspine.
- Some blocking into the superspine.
- Evolving:
  - JASMIN 2 injection bandwidth into superspine ≈ 2 Tbit/s;
  - JASMIN 4 >6 Tbit/s.
- More pods possible.
- Designed by Jonathan Churchill, STFC, Inspired by Facebook.

Community Software: JASMIN Analysis Platform *et. al.*

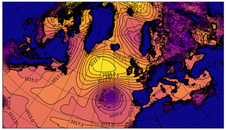


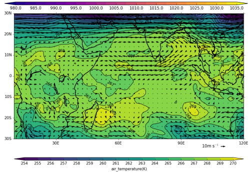







```
import cf, cfplot as cfp
f=cf.read('/opt/graphics/cfplot_data/tas_A1.nc')
g=f.subspace(time=15)
cfp.gopen()
cfp.cscale('magma')
cfp.mapset(proj='npstere')
cfp.con(g)
cfp.stipple(f=g, min=265, max=295, size=100, color='#00ff00')
cfp.gclose()
```





cf-python: https://cfpython.bitbucket.org
cf-plot: https://ajheaps.github.io/cf-plot

...and many more ... all shared and (hopefully) kept up to date on the JAP: http://www.jasmin.ac.uk/services/jasmin-analysis-platform/.

Community Intercomparison Suite: https://www.cistools.net/
Watson-Paris et al, 2016 (doi:10.5194/gmd-2016-27)

Uncommon (and inappropriate?) software solutions

Multiple tools

Contrast between two very types of workflow:

- Build Once: Many analysis tasks are build once, use once, throwaway. No room for optimisation (or MPI). *Need efficient libraries.*
- Repeatable: "build", "run", "move", "reduce/reformat", "analyse". *Much room for automation.*.

What to use? Plethora of architectures and tools out there

Uncommon (and inappropriate?) software solutions

Multiple tools

Contrast between two very types of workflow:

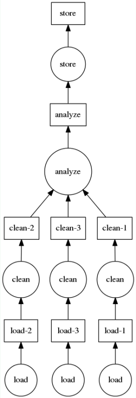
- Build Once: Many analysis tasks are build once, use once, throwaway. No room for optimisation (or MPI). *Need efficient libraries.*
- Repeatable: “build”, “run”, “move”, “reduce/reformat”, “analyse”. *Much room for automation.*.

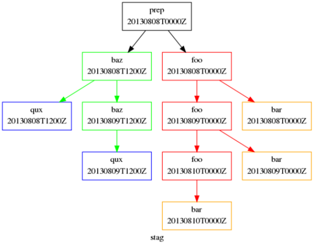
What to use? Plethora of architectures and tools out there

APACHE Spark DASK Pandas

Exploiting Concurrency

Whatever tools, need to get used to generating, understanding, and exploiting concurrency in more complicated ways:

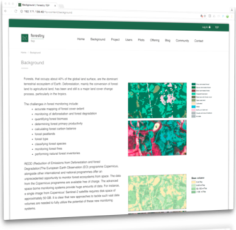




Much to do to harness tools to accelerate workflows!

(These two examples: dask, and cylc, representing bespoke analysis and scheduling, reduction and proliferation.)

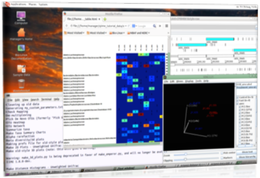
Virtual Research Environments on JASMIN hosted cloud



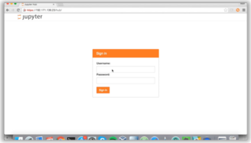
Thematic Exploitation Platforms for ESA

CCI Open Data Portal for ESA

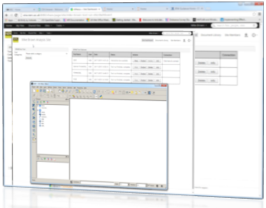
MAJIC interface to JULES model



EOS Cloud — Desktop-as-a-Service for Environmental Genomics

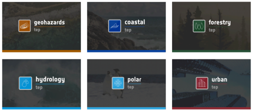


Hosted Ipython Notebooks



NERC Environmental Workbench

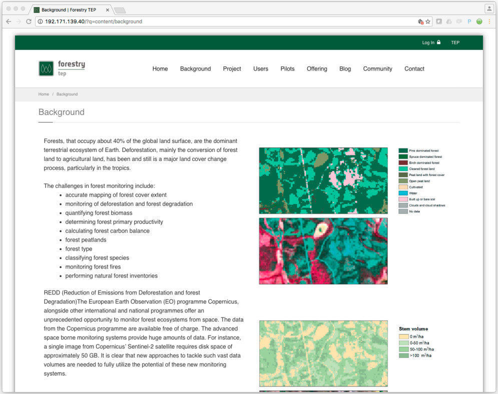
Thematic Exploitation Platforms for ESA



Forestry TEP

- A one-stop shop for forestry remote sensing services for the academic and commercial sectors.
- Offers access to pre-processed satellite and ancillary data, computing power, and software access and hosting.

…built by VTT Technical Research Centre & Arbonaut (FIN), CGI IT & STFC (UK), and Spacebel (BEL).

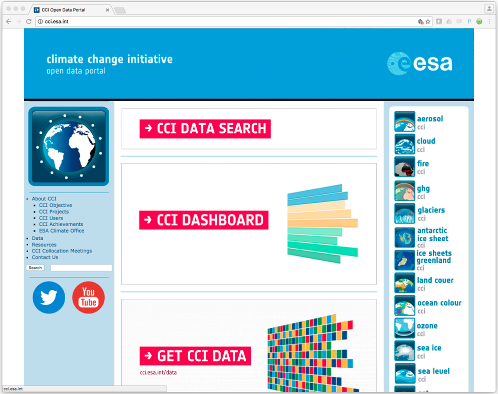


CEDA is supporting the Forestry and Polar TEPS on the JASMIN un-managed cloud.

CCI Open Data Portal for ESA

The Climate Change Initiative

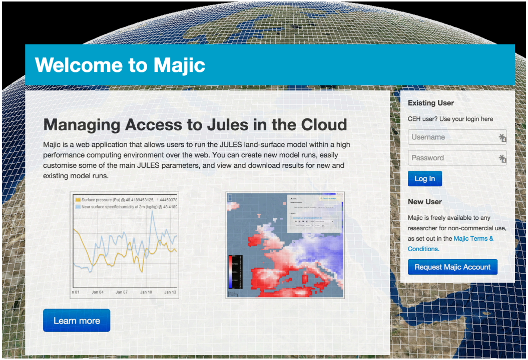
- Exploiting Europe's EO space assets to generate robust long-term global records of essential climate variables such as greenhouse-gas concentrations, sea-ice extent and thickness, and sea-surface temperature and salinity.
- The CCI Open Data Portal is hosted on JASMIN and exploits a near complete copy of the CCI datasets held in the CEDA archive.



MAJIC: Managing Access to JULES in the cloud

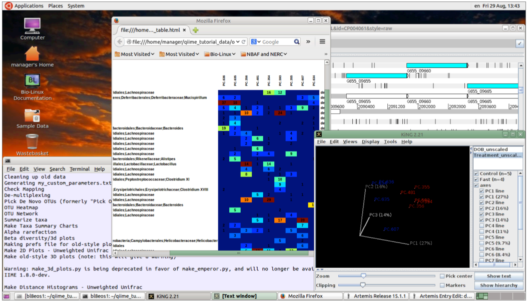
JULES
Joint UK Land Environment Simulator

- JULES is a community land surface model incorporating processes such as surface energy balance, the hydrological cycle, carbon cycle, dynamic vegetation etc.
- MAJIC provides a web portal running in the un-managed cloud which allows users to configure JULES to run on the JASMIN/LOTUS batch cluster and return results.



EOS Cloud: Desktop as a Service for Environmental Genomics

- The EOS cloud is a facility to support NERC omics researchers running on Bio-Linux.
- The DaaS service allows researchers to run Bio-Linux instances in the JASMIN cloud, with the additional function of (nearly) dynamically changing their memory requirements - allowing efficient use of large memory machines by multiple desktop users.

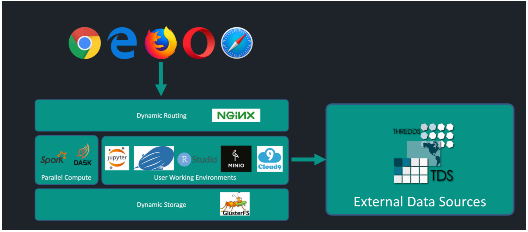


NERC Data Labs

Goals:

- Lower the barrier of entry to collaborative analysis tools
- Faster, repeatable results with higher quality deliverables
- Reduce per-project infrastructure procurement, management and running costs.

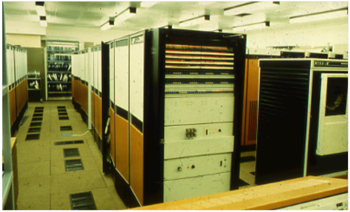
Browser based interface:



Multiple Technologies:

Faster Compute

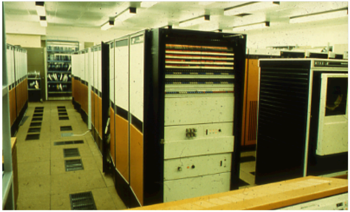
1981: ICL Dist.Array.Proc. (20 MFlops)



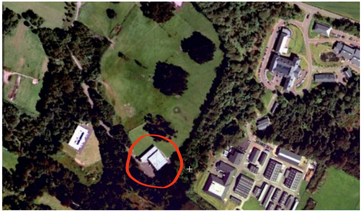
2014: Archer (then 1.4 PFlops)

Faster Compute

1981: ICL Dist.Array.Proc. (20 MFlops)



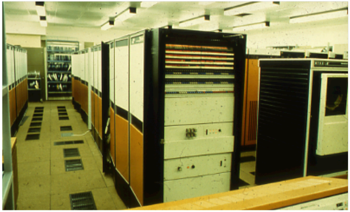
EPCC Advanced Computing Facility, 2014



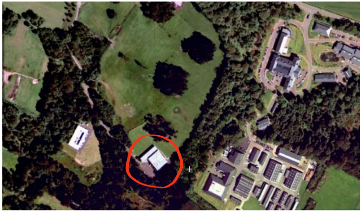
2014: Archer (then 1.4 PFlops)

Faster Compute

1981: ICL Dist.Array.Proc. (20 MFlops)



EPCC Advanced Computing Facility, 2014



2014: Archer (then 1.4 PFlops)

From 1981, without Moore's Law

Slide content courtesy of Arthur Trew:

Moore's Law and Friends

Moore's Law

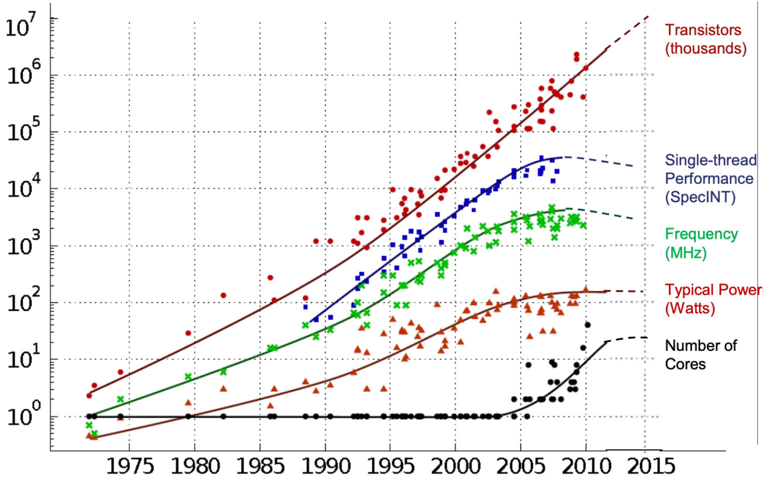
More often misquoted and misunderstood:

- Original, Moore, 1965: The complexity for minimum component costs has increased at a rate of roughly a factor of two per year.
- House (Intel) modified it to note that: The changes would cause computer performance to double every 18 months
- Moore (Modified 1975): The number of transistors in a dense integrated circuit doubles about every two years

Dennard Scaling

- The performance per watt of computing is growing exponentially at roughly the same rate (doubling every two years).
- (Increasing clock frequency as circuits get smaller, but this stopped working around 2006, too much power too small, means meltdown!)

The end of Dennard Scaling



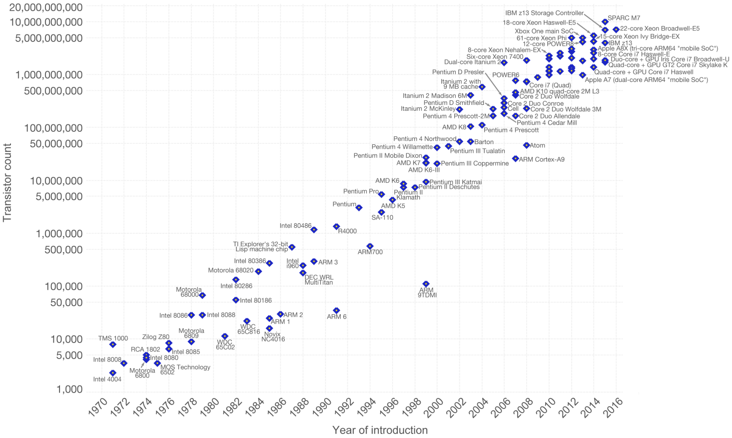
Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten

Moores's Law

Moore's Law – The number of transistors on integrated circuit chips (1971-2016)

Our World in Data

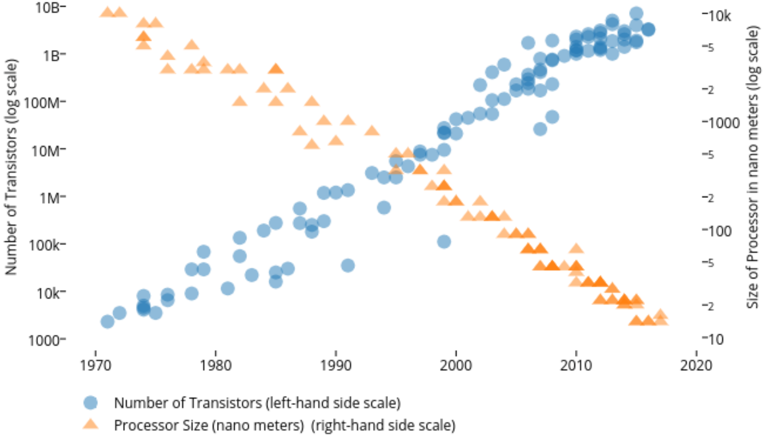
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important as other aspects of technological progress – such as processing speed or the price of electronic products – are strongly linked to Moore's law.



Data source: Wikipedia (https://en.wikipedia.org/wiki/Transistor_count)
The data visualization is available at OurWorldinData.org. There you find more visualizations and research on this topic.

https://en.wikipedia.org/wiki/Transistor_count

Moores's Law



https://www.yaabot.com/31345/quantum-computing-neural-chips-moores-law-future-computing/

Moore's 2nd Law aka Rock's Law

- The cost of a semiconductor chip fabrication plant doubles every four years.
- Noyce, 1977: "…further miniaturization is less likely to be limited by the laws of physics than by the laws of economics."


The Register®
Biting the hand that feeds IT

NTRE SOFTWARE SECURITY DEVOPS BUSINESS PERSONAL TECH SCIENCE

Personal Tech

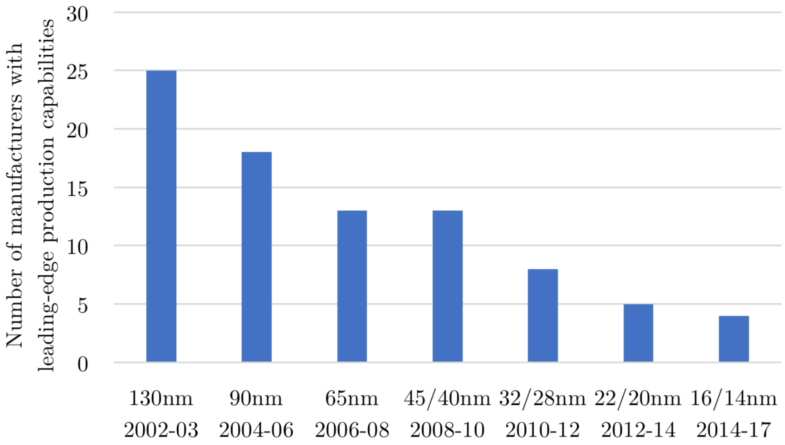
GlobalFoundries scuttles 7nm chip plans claiming no demand

AMD promptly dumps it and hires TSMC for next-gen chips

By Shaun Nichols in San Francisco 27 Aug 2018 at 23:55 18 SHARE ▲

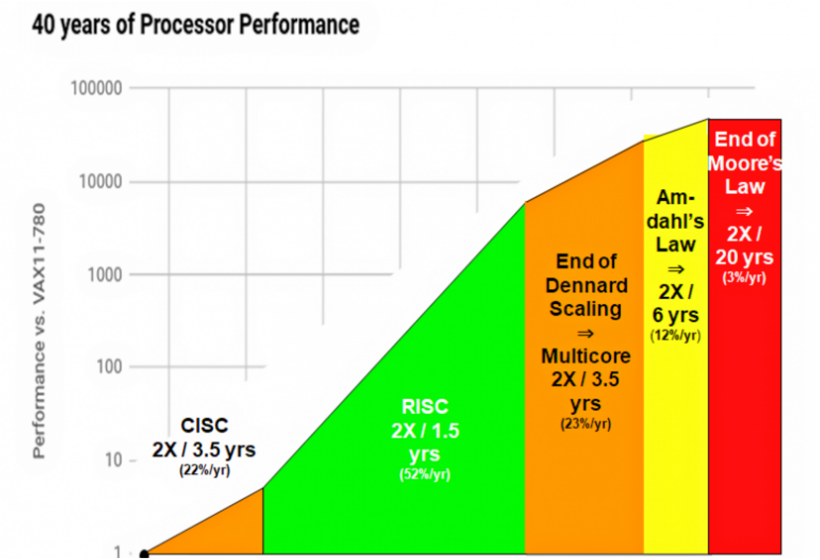
- …to shift resources (including R&D) to the 14 and 12nm efforts where …most of their chip customers …are planning to stay with the current-gen architectures and squeeze performance out by other means.

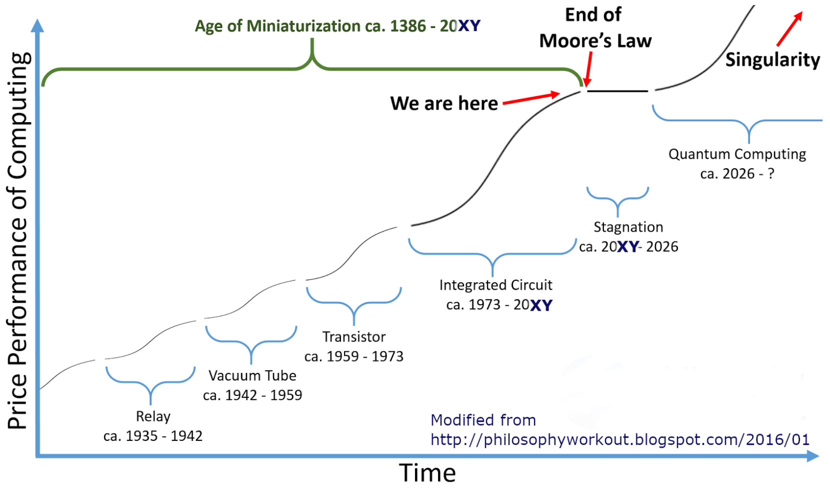
- 7nm is expensive, it's cheaper and easier to improve the performance and density of 12nm, and hardware accelerators and custom chips …



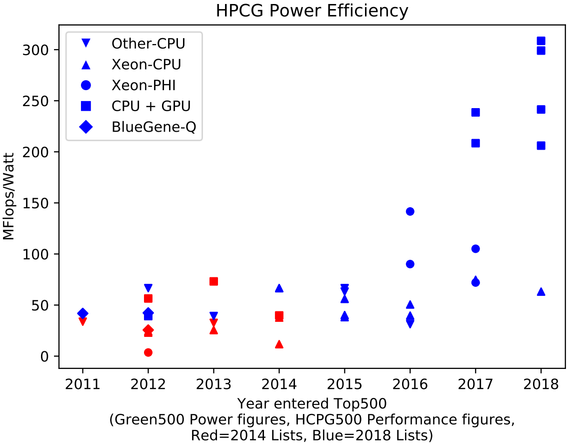
https://www.nextplatform.com/2019/02/05/the-era-of-general-purpose-computers-is-ending/

The Evolving Moore's Law




Age of Miniaturization ca. 1386 - 20XY
End of Moore's Law
Singularity
We are here
Quantum Computing ca. 2026 - ?
Stagnation ca. 20XY- 2026
Integrated Circuit ca. 1973 - 20XY
Transistor ca. 1959 - 1973
Vacuum Tube ca. 1942 - 1959
Relay ca. 1935 - 1942
Price Performance of Computing
Time
Modified from http://philosophyworkout.blogspot.com/2016/01

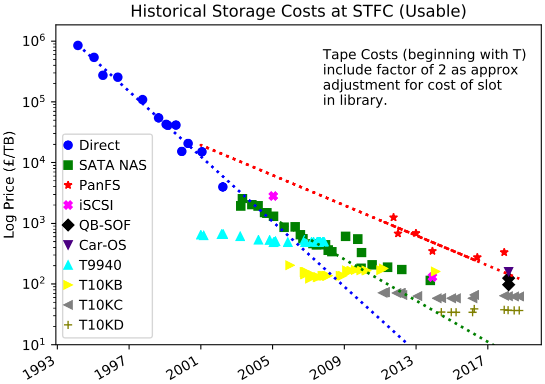
Power Consumption and Performance



Real experience with Kryder's Law!

Kryder's Law

- The assumption that disk drive density, also known as areal density, will double every thirteen months. (Hasn't for some time!)
- The implication of Kryder's Law is that as areal density improves, storage will become cheaper:



- Relative cost of **disk** storage going up: each new generation of disk has a "shallower Kryder rate".
- Each new generation of **tape** is cheaper, and price stable over the lifetime.
- Tape has better technical future prospects than disk!

Smarter Maths? Techniques!

Parallel Time-Stepping

Not radical (in principle):

$$\mathbf{X}_{t+1}(x,y,z,t) = f(\mathbf{X}_{t-1}, \mathbf{X}_t)$$

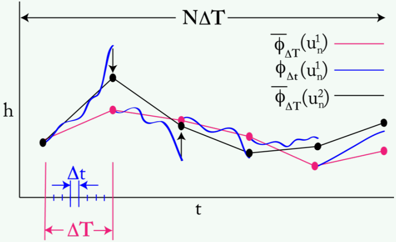
The function f can involve several steps (iterates) or some sort of prediction/correction.

$$\begin{aligned} \text{Predictor:} \quad \mathbf{X}^p_{t+1} &= f_p(\mathbf{X}_{t-1}, \mathbf{X}_t) \\ \text{Corrector:} \quad \mathbf{X}_{t+1} &= f_c(\mathbf{X}^p_{t+1} + \mathbf{X}_t) \end{aligned}$$

There is scope to do some of this in parallel with several methods discussed in the literature.

Parallel in Time

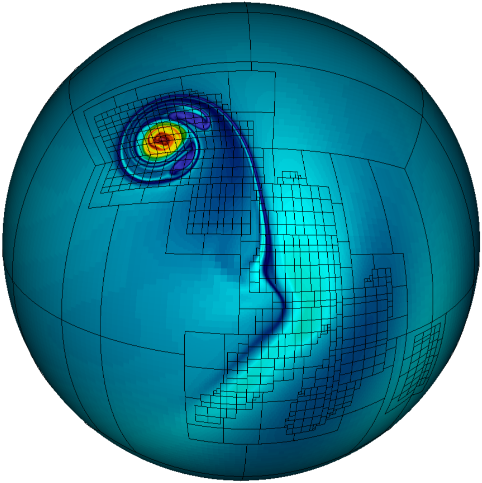
Quite radical:



Predict using a coarse model with long timesteps. Correct in parallel with a finer resolution model.
Some experiments in the literature …

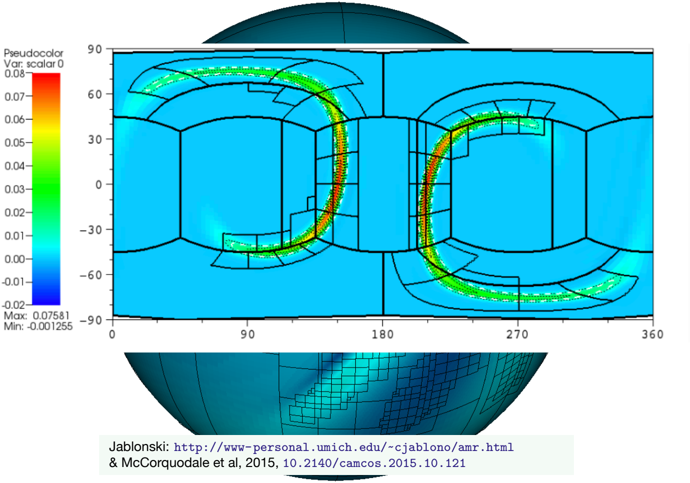
Smarter Maths? - Adaptive Grids

If we can't have ever increasing uniform grids:



Smarter Maths? - Adaptive Grids

If we can't have ever increasing uniform grids:



Jablonski: http://www-personal.umich.edu/~cjablono/amr.html
& McCorquodale et al, 2015, 10.2140/camcos.2015.10.121

Growing impact of Machine Learning and Artificial Intelligence

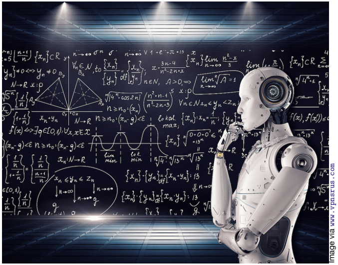


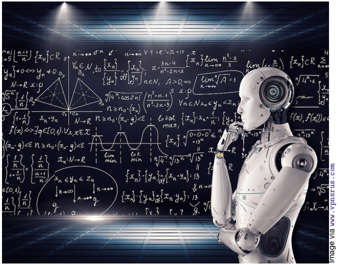
Image via www.vpnsrus.com

Gratuitous “robots are coming” image

Expect ML and AI to have major implications for both

- HPC architectures, and
- Algorithms, in use before, during, and after simulation (analytics)!

Growing impact of Machine Learning and Artificial Intelligence



Gratuitous "robots are coming" image

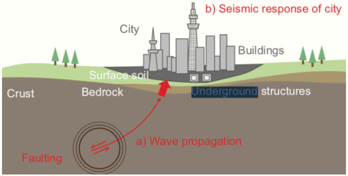
Expect ML and AI to have major implications for both

- HPC architectures, and
- Algorithms, in use before, during, and after simulation (analytics)!

Initial emphasis on climate services, parameter estimation (for parameterisations) and emulation (potentially avoiding avoid long spin-up runs).

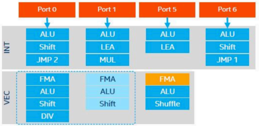
Two interesting examples contributed to the Gordon Bell competition this year:

- Preconditioning implicit solvers using artificial intelligence — ground breaking (!) simulations of earthquakes and building response : Ichimura et al 2018.

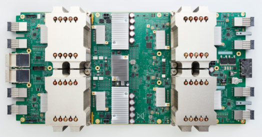


- Exascale Deep Learning for Climate Analytics - Extracting weather patterns from climate simulations: Kurth et al 2018, co-winner of 2018 Gordon Bell prize.

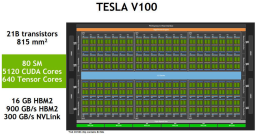
From decades of the same to a Cambrian Explosion



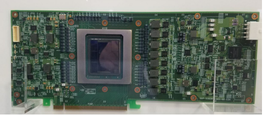
Vector Processors on Intel Zeon



Google's Tensor Programming Unit



GPUs from NVIDIA and AMD



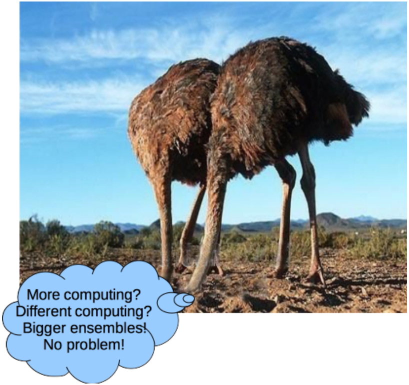
Vector Processing Units from NEC

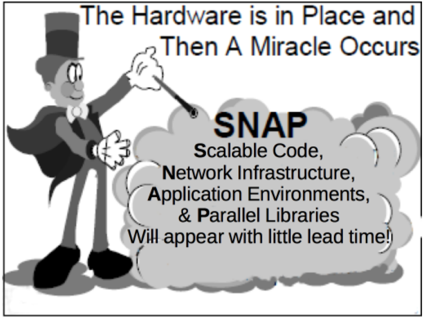
Server chips based on ARM designs

FPGA from many sources

The end of Moore's Law means more specialisation: all with very different programming models!

What about software?





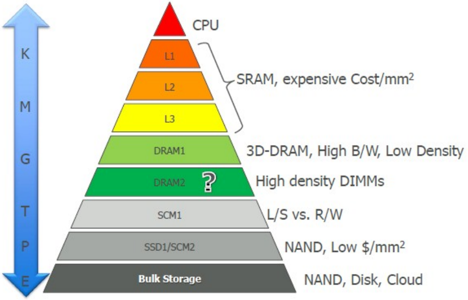
Some people have a very naive idea about the relationship between the hardware and the software!

Too many levels of parallelism

- Vector Units (on chip)
- Parallelism Across Cores
- Shared Memory Concurrency
- Distributed Memory
- Numerical Method Concurrency
- Internal Component Concurrency
- Coupled Component Concurrency
- I/O and Diagnostic Parallelism
- (Storage System Parallelism)

Too many levels of parallelism

- Vector Units (on chip)
- Parallelism Across Cores
- Shared Memory Concurrency
- Distributed Memory
- Numerical Method Concurrency
- Internal Component Concurrency
- Coupled Component Concurrency
- I/O and Diagnostic Parallelism
- (Storage System Parallelism)



Nearly everything is processor/system dependent! (except green layers on left).

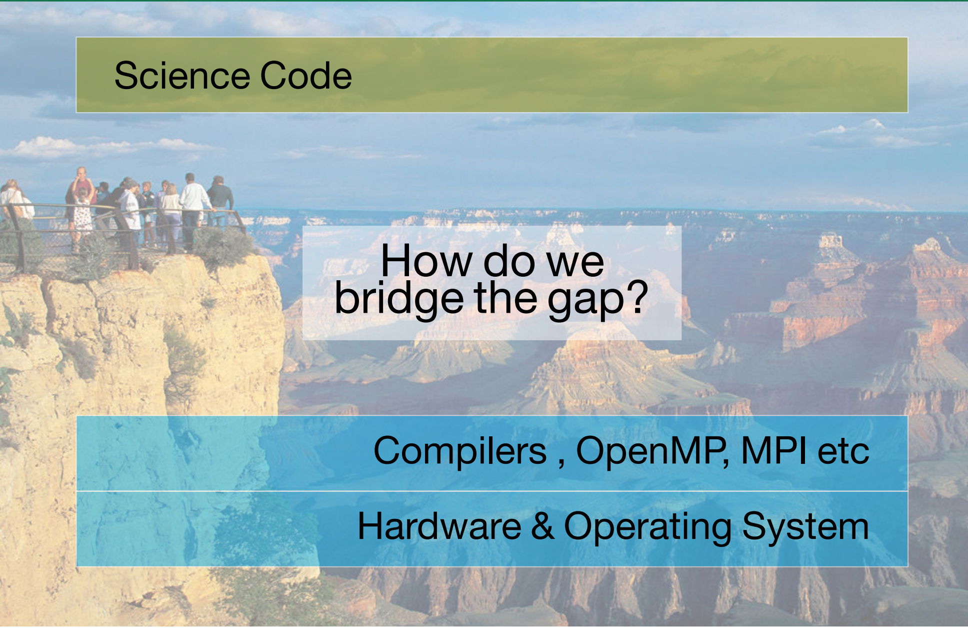
Entirely new programming models are likely to be necessary, with entirely new* constructs such as thread pools and task-based parallelism possible. Memory handling will be crucial!

*New in this context!

Software changing slowly & slowing!

Hardware changing rapidly & accelerating!

How far is it between our scientific aspiration and our ability to develop and/or rapidly adapt our codes to the available hardware?


Science Code
How do we bridge the gap?
Compilers , OpenMP, MPI etc
Hardware & Operating System

Route 1: The Massive Edifice

- No group has enough effort to do all the work needed.
- No group has **all** the relevant expertise.

Route 2: Incremental Advances

- The peril of the local minimum
- Any given span/leap may not be sufficient to cross the next gap!

Route 1: The Massive Edifice

- No group has enough effort to do all the work needed.
- No group has **all** the relevant expertise.

Route 2: Incremental Advances

- The peril of the local minimum
- Any given span/leap may not be sufficient to cross the next gap!

Route 3: Assemble Components

- Share Requirements; Share Development.
- Define Interfaces and Connections.

Science Code

Defined Interfaces and Contracts

High Level Libraries and Tools

Defined Interfaces and Contracts

Libraries and Tools

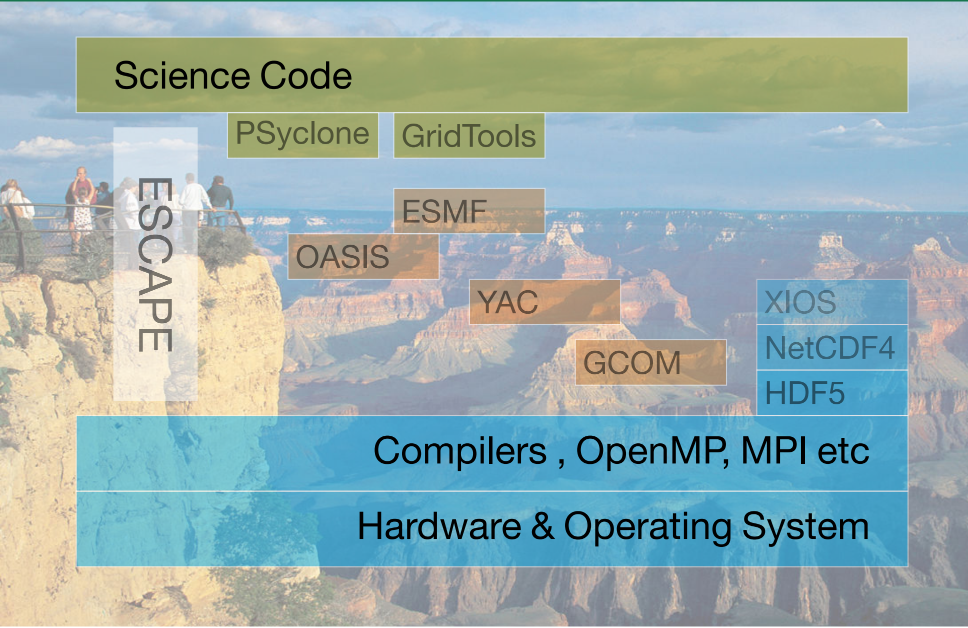
Defined Interfaces and Contracts

Low-Level Libraries and Tools

Defined Interfaces and Contracts

Compilers , OpenMP, MPI etc

Hardware & Operating System


Science Code
PSyclone
GridTools
ESCAPE
ESMF
OASIS
YAC
GCOM
XIOS
NetCDF4
HDF5
Compilers , OpenMP, MPI etc
Hardware & Operating System

Why and What is a Domain Specific Language (DSL)?

Why?

- Humans currently produce the best optimised code!
- Humans can inspect an algorithm, and exploit domain-specific knowledge to reason how to improve performance – but a compiler or generic parallelisation tool doesn't have that knowledge.
- Result: Humans better than generic tools every time, but it's big slow task and mostly not portable!

Why and What is a Domain Specific Language (DSL)?

Why?

- Humans currently produce the best optimised code!
- Humans can inspect an algorithm, and exploit domain-specific knowledge to reason how to improve performance – but a compiler or generic parallelisation tool doesn't have that knowledge.
- Result: Humans better than generic tools every time, but it's big slow task and mostly not portable!

What?

- A domain specific compiler, with a set of rules!
- Exploits a priori knowledge, e.g.
  - Operations are performed over a mesh,
  - The same operations are typically performed independently at each mesh point/volume/element,
  - the meshes themselves typically have consistent properties.
- Leave a much smaller task for the humans!

DSLs in the Wild — two major projects:

- GridTools (formerly Stella)
- PSyclone (from Gung Ho)

Both are DSELs ... domain specific embedded languages.

- Embedded in C++
- Originally targeted finite difference lat-lon Limited Area Model.
- Backends (via human experts) mapped to the science description via C++ templates.

- Embedded in Fortran
- Originally targeted finite element irregular mesh.
- A recipe of optimisations (via human experts) is used by PSyclone to produce targeted code.

In both cases the DSL approach allows mathematical experts to do their thing, while HPC experts do their thing, and the DSL provides a separation of concerns.

Whither the DSL?

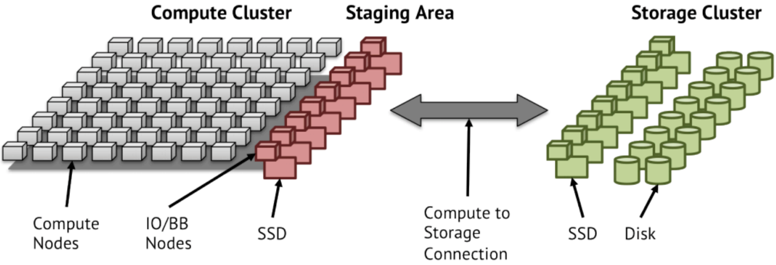
- DSLs are becoming more common across disciplines.
- The Domains are more or less specific …
  - the more specific, the cleaner a domain specific separation of concerns, but the larger the technical debt (maintaining the code and the teams of experts for the backends
  - the more generic, the less the DSL can do for you, and the less the separation of concerns.

Whither the DSL?

- DSLs are becoming more common across disciplines.
- The Domains are more or less specific …
  - the more specific, the cleaner a domain specific separation of concerns, but the larger the technical debt (maintaining the code and the teams of experts for the backends
  - the more generic, the less the DSL can do for you, and the less the separation of concerns.

- The holy grail is to add further separation of concerns inside the DSL …e.g. can we imagine a GridTools *and a* PSyclone front end to a vendor managed intermediate DSL compiler?
  - compare with MPI: successful because vendors manage their own specific backends with a defined API that we all work with to develop our own libraries (e.g. GCOM, YAXT etc)!

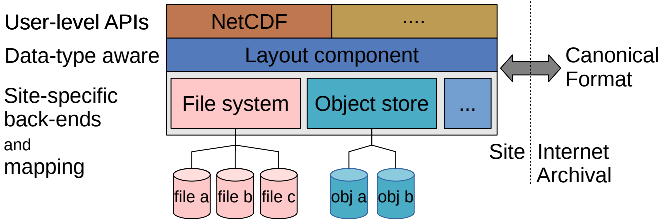
Parallelism in Storage - Getting to and From



Existing filesystems are limiting

- Storage Architecture is complex.
- Difficult to initialise models (takes too long to read and distribute initial data)
- Difficult to get sufficient performance from hundreds of nodes writing to a file system!

Earth System Data Middleware



Key Concepts

- Applications work through existing application interfaces (currently: NetCDF library)
- Middleware utilizes layout component to make placement decisions
- Data is then written/read efficiently avoiding file system limitations (e.g. consistency constraints)
- Potential for deploying with an active storage management system.

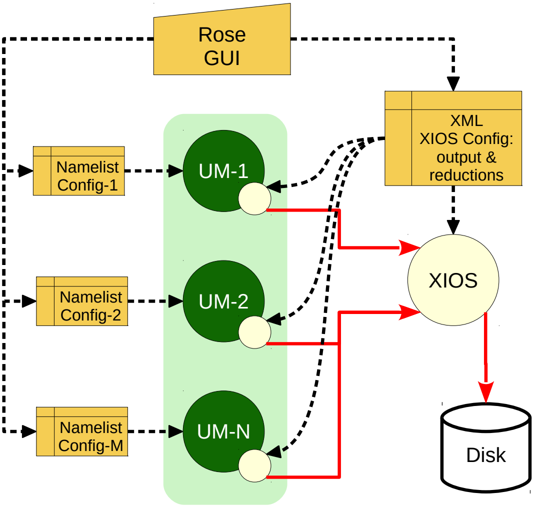
In-Flight Parallel Data Analysis

An ensemble is a set of simulations running different instances of the same numerical experiment. We do this to get information about uncertainty.

Dealing with too much ensemble data

Instead of writing out all ensemble members and doing all the analysis later:

- Calculate ensemble statistics on the fly.
- Only write out some ensemble members.
- (Which ones? A tale for another day, see Daniel Galea's Ph.D work.)



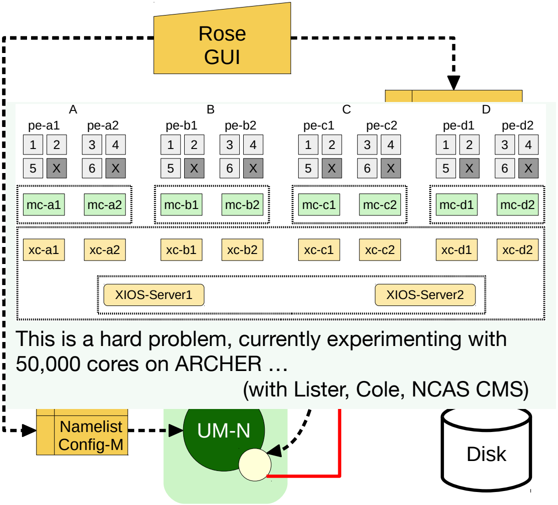
In-Flight Parallel Data Analysis

An ensemble is a set of simulations running different instances of the same numerical experiment. We do this to get information about uncertainty.

Dealing with too much ensemble data

Instead of writing out all ensemble members and doing all the analysis later:

- Calculate ensemble statistics on the fly.
- Only write out some ensemble members.
- (Which ones? A tale for another day, see Daniel Galea's Ph.D work.)



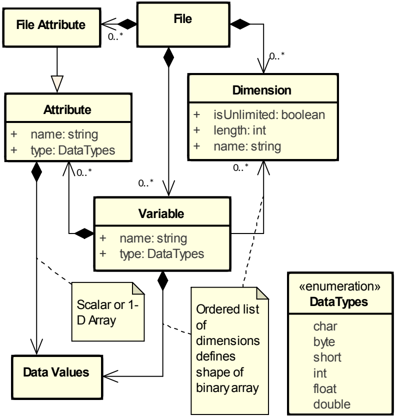
This is a hard problem, currently experimenting with 50,000 cores on ARCHER …

(with Lister, Cole, NCAS CMS)

Climate Forecast Conventions and Data Model

Formats and Semantics

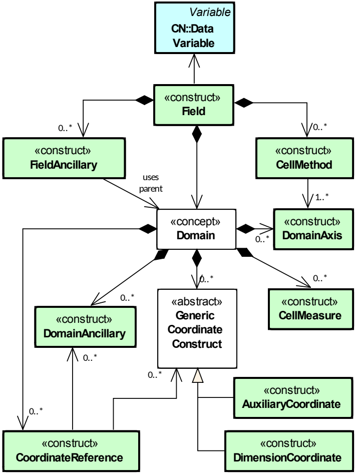
- A file format describes how bits and bytes are organised in some sequence on disk.
- Storage Middleware (e.g. NetCDF) has an implicit or explicit data model for what things are stored in that file.
- The Climate-Forecast conventions describe how coordinates and variable properties are stored in NetCDF.
- We have developed an explicit data model so that these can be used for any storage format.



Climate Forecast Conventions and Data Model

Formats and Semantics

- A file format describes how bits and bytes are organised in some sequence on disk.
- Storage Middleware (e.g. NetCDF) has an implicit or explicit data model for what things are stored in that file.
- The Climate-Forecast conventions describe how coordinates and variable properties are stored in NetCDF.
- We have developed an explicit data model so that these can be used for any storage format.

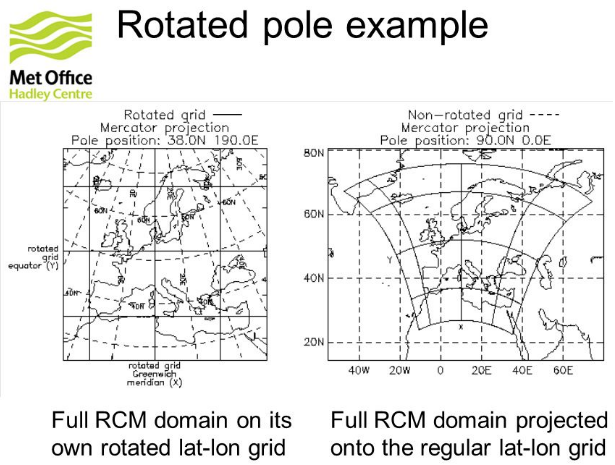


Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., and Taylor, K. E.: A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1), Geosci. Model Dev., 10, 4619-4646, https://doi.org/10.5194/gmd-10-4619-2017, 2017.

CF Conventions in Action

```
print(t)
Field: air_temperature (ncvar\%ta)
---------------------------------
Data               : air_temperature(atmosphere_hybrid_height_coordinate(1),
                     grid_latitude(10), grid_longitude(9)) K
Cell methods       : grid_latitude(10): grid_longitude(9):
                     mean where land (interval: 0.1 degrees) time(1): maximum
Field ancils       : air_temperature standard_error(grid_latitude(10),
                     grid_longitude(9)) = [[0.81, ..., 0.78]] K
Dimension coords: time(1) = [2019-01-01 00:00:00]
                : atmosphere_hybrid_height_coordinate(1) = [1.5]
                : grid_latitude(10) = [2.2, ..., -1.76] degrees
                : grid_longitude(9) = [-4.7, ..., -1.18] degrees
Auxiliary coords: latitude(grid_latitude(10),
                  grid_longitude(9)) = [[53.941, ..., 50.225]] degrees_N
                : longitude(grid_longitude(9),
                  grid_latitude(10)) = [[2.004, ..., 8.156]] degrees_E
                : long_name=
                    Grid latitude name(grid_latitude(10)) = [--, ..., kappa]
Cell measures   : measure:area(grid_longitude(9),
                  grid_latitude(10)) = [[2391.9657, ..., 2392.6009]] km2
...
```

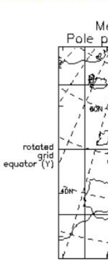
CF Conventions in Action

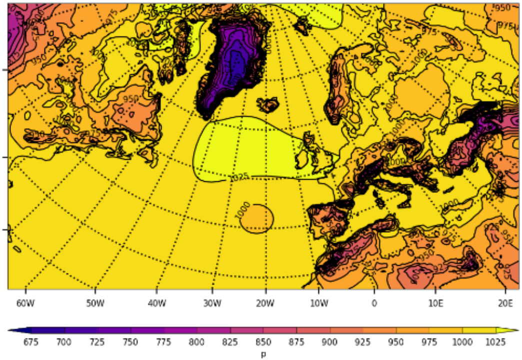



```
oordinate(1),
ne(1): maximum
de(10),
1.5]
ees
rees
]] degrees_N
 degrees_E
[--, ..., kappa]
grid_latitude(10)) = [[2391.9657, ..., 2392.6009]] km2
...
```

CF Conventions in Action

Rotated pole example



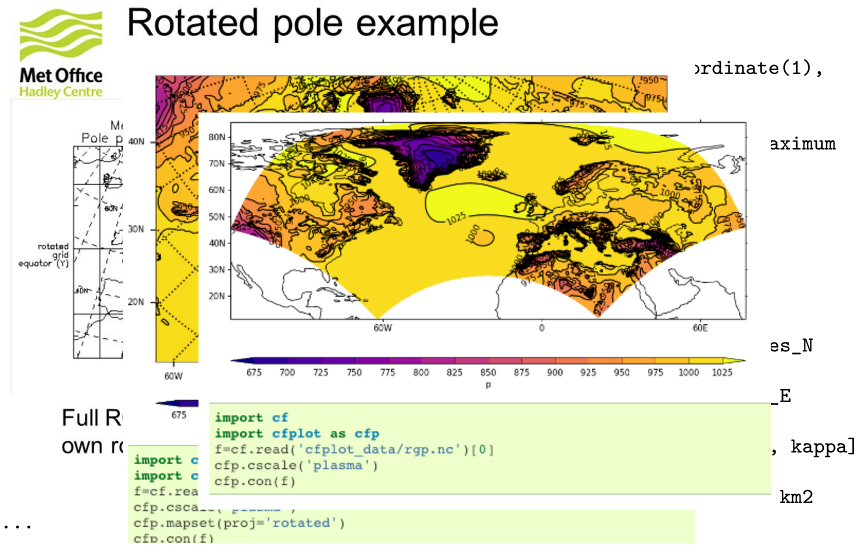


Full R
own r

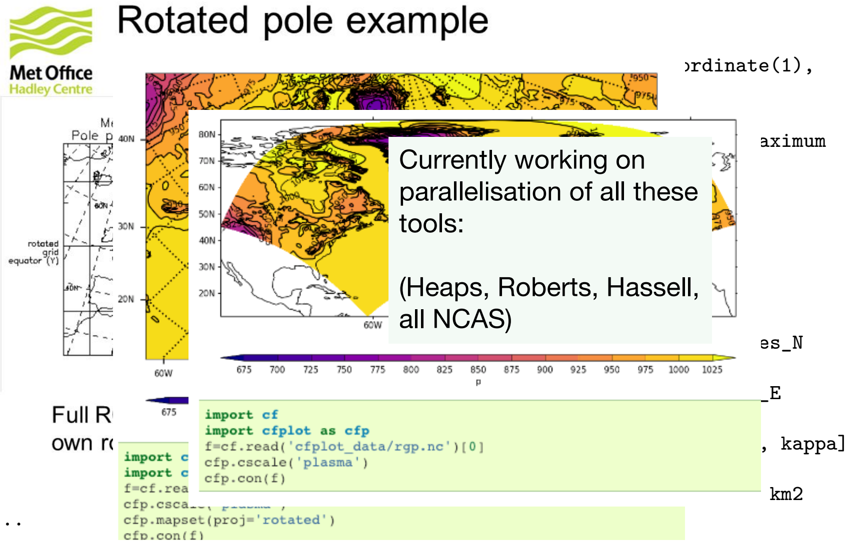
```
ordinate(1),
e(1): maximum
e(10),
.5]
es
ees
| degrees_N
degrees_E
--, ..., kappa]
.6009]] km2
...
```

```
import cf
import cfplot as cfp
f=cf.read('cfplot_data/rgp.nc')[0]
cfp.cscale('plasma')
cfp.mapset(proj='rotated')
cfp.con(f)
```

CF Conventions in Action



CF Conventions in Action



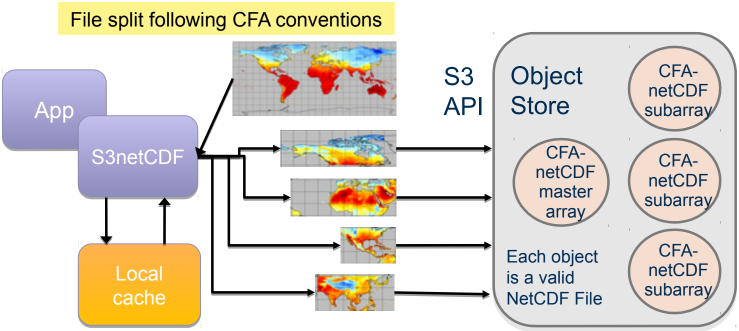
Currently working on parallelisation of all these tools:

(Heaps, Roberts, Hassell, all NCAS)

```
import cf
import cfplot as cfp
f=cf.read('cfplot_data/rgp.nc')[0]
cfp.cscale('plasma')
cfp.con(f)
```

Semantic Storage Layer

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



Architecture (with Massey & Jones, STFC)

- Master Array File is a NetCDF file containing dimensions and metadata for the variables including URLs to fragment file locations
- Master Array file optionally in persistent memory or online, nearline, etc. NetCDF tools can query file CF metadata content without fetching them

Summary

- Climate modelling is one of the grand computational challenges

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.
  - Data handling platforms include many diverse components to serve diverse communities and their workflows.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.
    - Data handling platforms include many diverse components to serve diverse communities and their workflows.
    - Tiered storage is necessary, but complicated in a cloud environment.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.
  - Data handling platforms include many diverse components to serve diverse communities and their workflows.
  - Tiered storage is necessary, but complicated in a cloud environment.
- The traditional route to more computing, via Moore's Law is ending.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.
  - Data handling platforms include many diverse components to serve diverse communities and their workflows.
  - Tiered storage is necessary, but complicated in a cloud environment.
- The traditional route to more computing, via Moore's Law is ending.
- New ways forward need to be found: from new maths, new techniques such as ML and AI, to new ways of programming, and new methods of data handling.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.
  - Data handling platforms include many diverse components to serve diverse communities and their workflows.
  - Tiered storage is necessary, but complicated in a cloud environment.
- The traditional route to more computing, via Moore's Law is ending.
- New ways forward need to be found: from new maths, new techniques such as ML and AI, to new ways of programming, and new methods of data handling.

Summary

- Climate modelling is one of the grand computational challenges
- Data handling is challenging, and getting more so.
- We need customised computing platforms such as JASMIN to handle the workflow.
  - Data handling platforms include many diverse components to serve diverse communities and their workflows.
  - Tiered storage is necessary, but complicated in a cloud environment.
- The traditional route to more computing, via Moore's Law is ending.
- New ways forward need to be found: from new maths, new techniques such as ML and AI, to new ways of programming, and new methods of data handling.

There is a lot for Computer Scientists to do!
aces.cs.reading.ac.uk

ACES