How will scientific, organisational and technical challenges impact on ENES and IS-ENES3?

Bryan Lawrence

NCAS & University of Reading

Paris, 9 Jan 19

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824084.

We live in difficult times

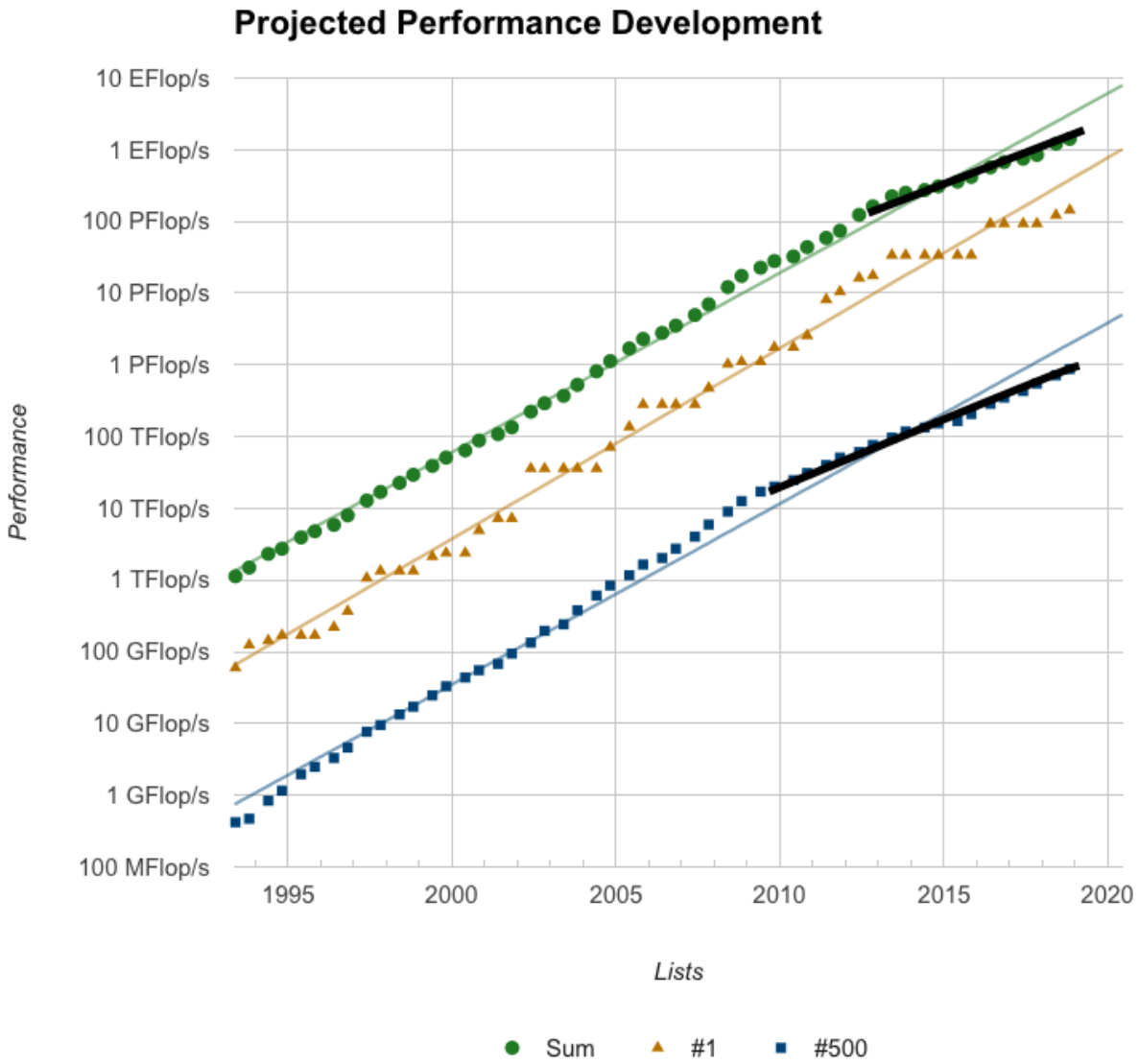
We live in difficult times

We live in difficult times

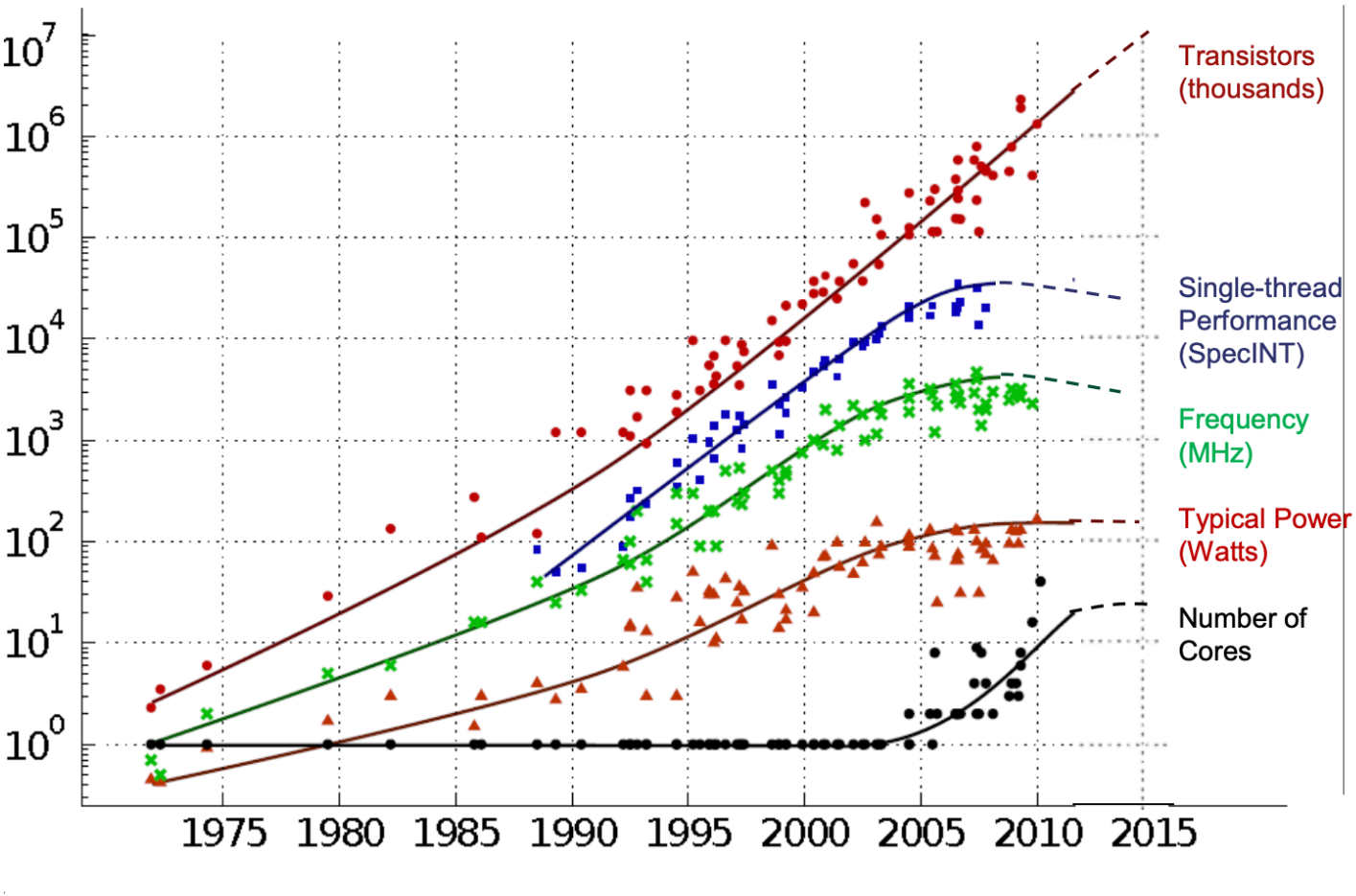
Not that!

This: Moore's Law …slowing

(from top500 November 2018, with my trend-lines)



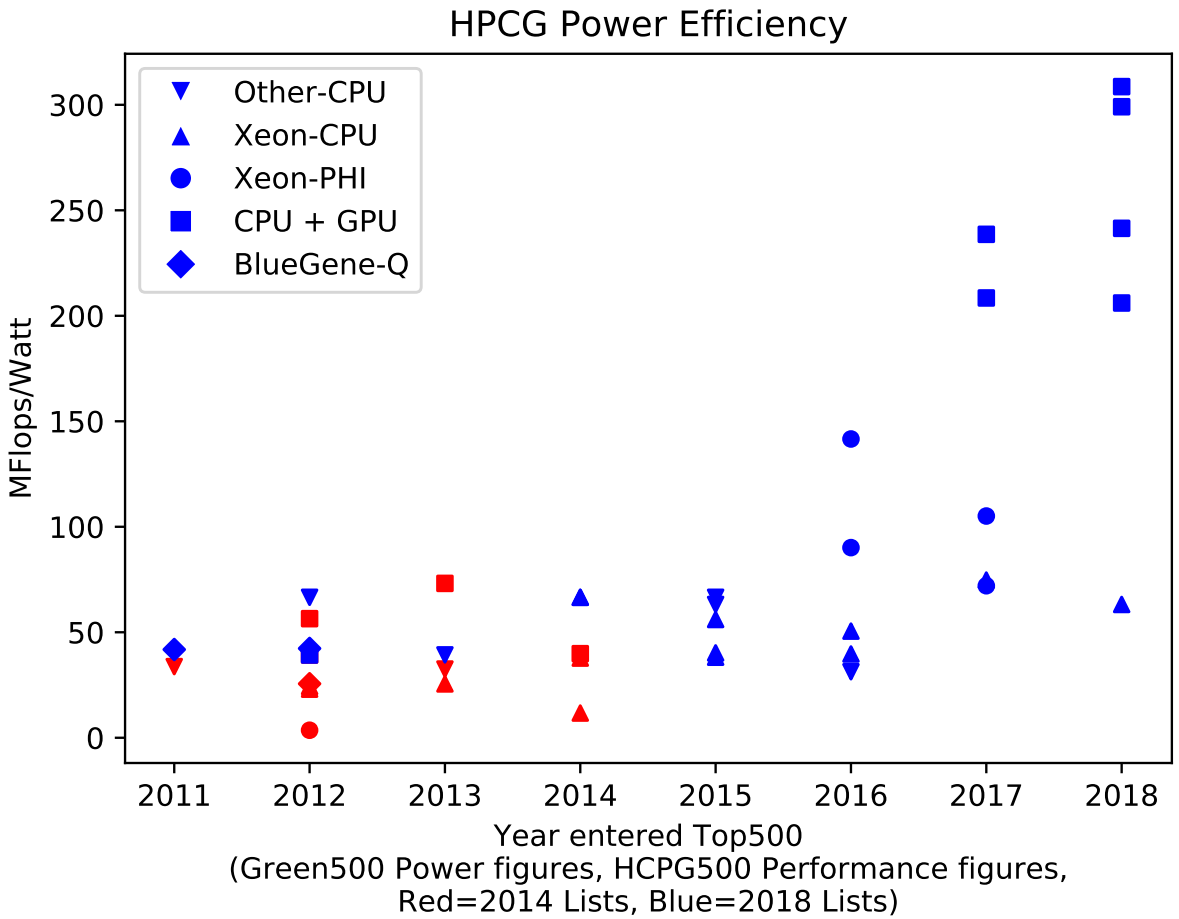
What's going on?



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

End of Dennard Scaling: Frequency can't keep increasing because power can't be removed from the chips, they'd melt!

One of many consequences: Efficiency



Note that the CPU efficiency hasn't change much over the decade!

Not a good story for climate science projects which want more compute!

Caveats and notes

- HPL power consumption and HPCG performance: apples and oranges.
- HPCG strong dependency on memory bandwidth (Marjanovíc et al, 2015).
- HPCG strong dependency on compiler even for same performance (Jäger et al, 2018).

Take home points

Moore's Law Myths and Legends

- Moore's Law is not (yet) dead, but it is slowing, and physics suggests it will die. How long? o(10) years??
- Harder to exploit increasing transistor density for CPU computing, easier for GPU computing (25x performance in last 5 years)!
  - For GPUS, simplicity might lead to more mileage in Moore's Law.
  - Memory bandwidth bigger constraint for HPCG (and us) than raw compute …

Speed and Power

- We spent a decade optimising for vector computers.
- We will have spent two decades optimising for massively parallel MPI.
- We will spend the next decade optimising GPUs/high-bandwidth-memory/Vector Units (maybe FPGA).
- From then on, IF we have strong scaling, we can only get more speed from using more power (electricty!)
- (... and we are running out of (have run out of?) strong scaling!)

Take home points

Moore's Law Myths and Legends

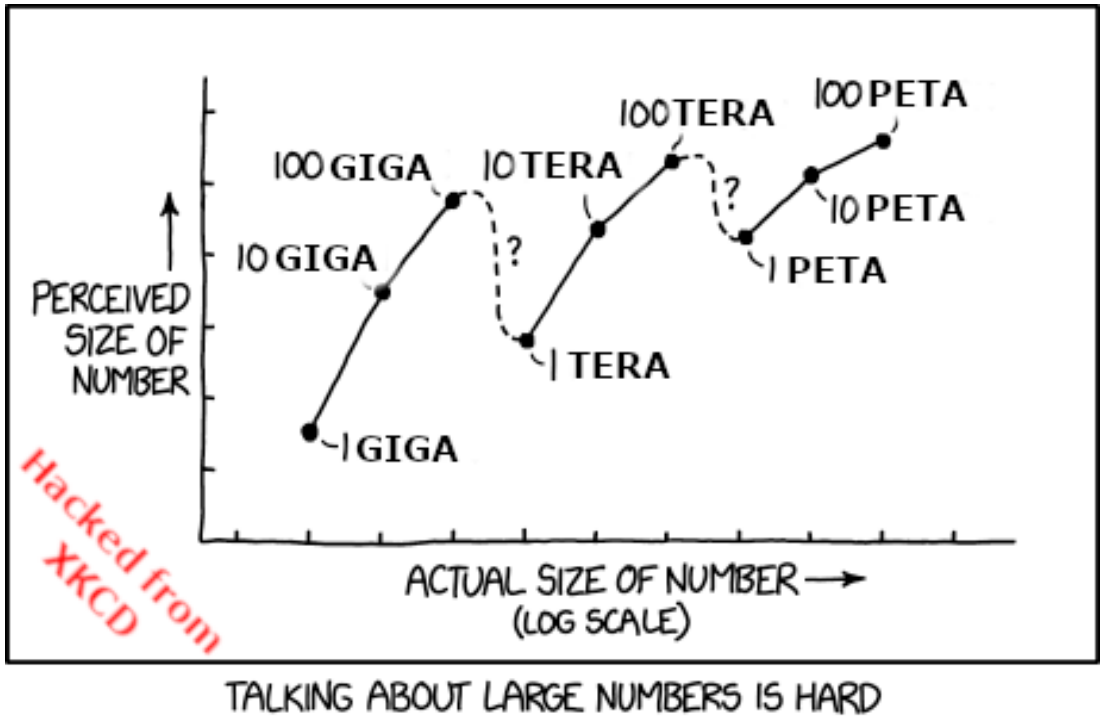
- Moore's Law is not (yet) dead, but it is slowing, and physics suggests it will die. How long? o(10) years??
- Harder to exploit increasing transistor density for C computing, easier for G computing (25x perform last 5 years)!
  - For GPUS, simplici might lead to more mileage in Moore's
  - Memory bandwidth constraint for HPCG (and us) than raw compute …

- Similar (if not worse) problems for storage cost and I/O performance!

Speed and Power

- We spent a decade optimising for vector computers.
- We will have spent two decades optimising for ssively parallel MPI.
- will spend the next decade mising GPUs/high-dwidth-memory/Vector s (maybe FPGA).
- then on, IF we have g scaling, we can only get speed from using more power (electricty!)
- (... and we are running out of (have run out of?) strong scaling!)

Climate Scientists don't *respect* big numbers!

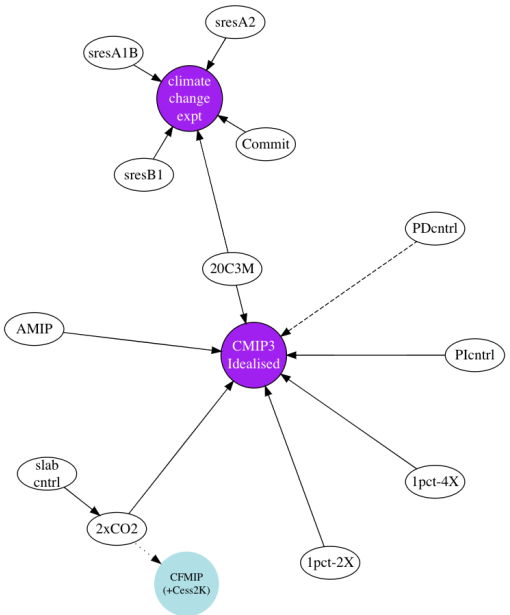


In experimental design, many underestimate:

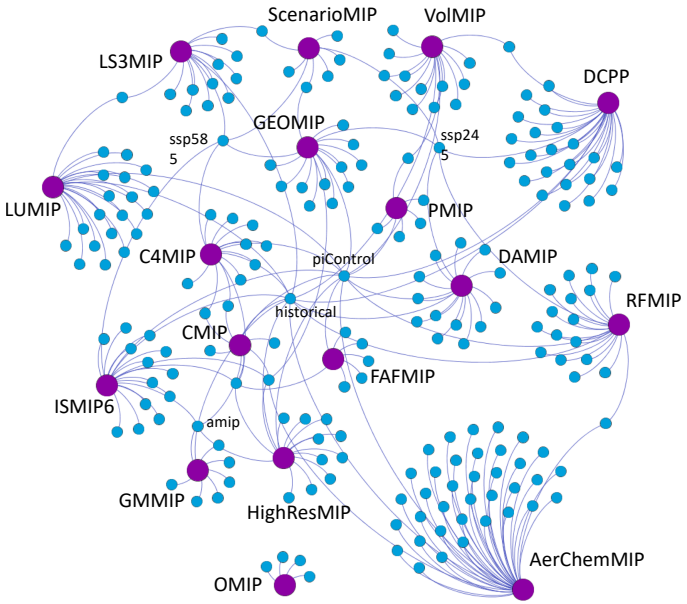
- The energy demands and costs of computing associated with their experiments, and
- The difficulty in managing, disseminating, and utilising large volumes of data!

This is only going to become worse unless we do something about it — but this is not a popular message!

The CMIP Evolution: from CMIP3 to CMIP6



to

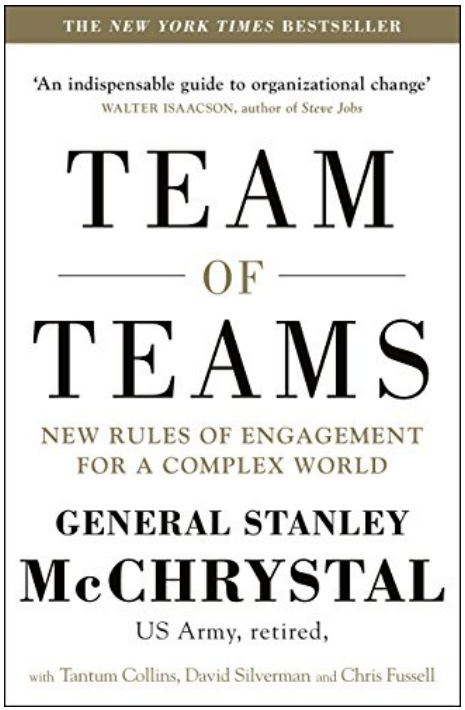


The Logistics of Collaboration

- In HPC we know that the larger the number of cores, the more the communications cost …
- these communications costs need to paid for large scale scientific collaboration too!

From experimental design, to the data request, the (ESGF) dissemination infrastructure, and to the analysis systems; we need to invest more in the supporting infrastructure, and respect the constraints — but this is not a popular message!

But: Tension between efficiency and adaptability



- The pursuit of "efficiency" - getting the most with the least investment of energy, time, or money - was once a laudable goal, but being effective in today's world is less a question of optimising for a known (and relatively stable) set of variables than optimising for a constantly shifting environment.
- Adaptability, not efficiency, must become our central competence.

The CMIP process is neither efficient* nor adaptable!

Has the "usability" requirement gone too far?

*for the modelling centres or the infrastructure.

Climate Scientists don't *respect* big numbers!

CMIP6 : Timescales, Volumes, Costs

- Designed without a budget.
- Designed without respect for the (energy/financial) cost — Does anyone know of any other scientific endeavour of this scale which has no clear cost/benefit analysis/justification?
- Designed without respect for infrastructure requirements.
  - Advent of the WIP reflects acceptance that there *are* infrastructural *issues*.
- As of today: we still don't know the timing and/or volumes of data delivery into the ESGF.
  - This is obviously problematic for data movement, data management, and the overall performance of the system.
  - "Download at home" did not work for CMIP5, yet there appears to be no real understanding by the designer/user community as to the consequences of the factor of ten in volumes expected for CMIP6.

Climate Scientists don't *respect* big numbers!

CMIP6 : Timescales, Volumes, Costs

- Designed without a budget.
- Designed without respect for the (energy/financial) cost — Does anyone know of any other scientific endeavour of this scale which has no clear cost/benefit analysis/justification?

No Hardware Budget

- We are presumed to be funded to deal with this, yet IS-ENES3 funding is relatively small, and does not reflect the underlying hardware costs.

Unrealistic Budget

- We will continue to be pilloried for not delivering an operational service from what is still a research (scale) budget.

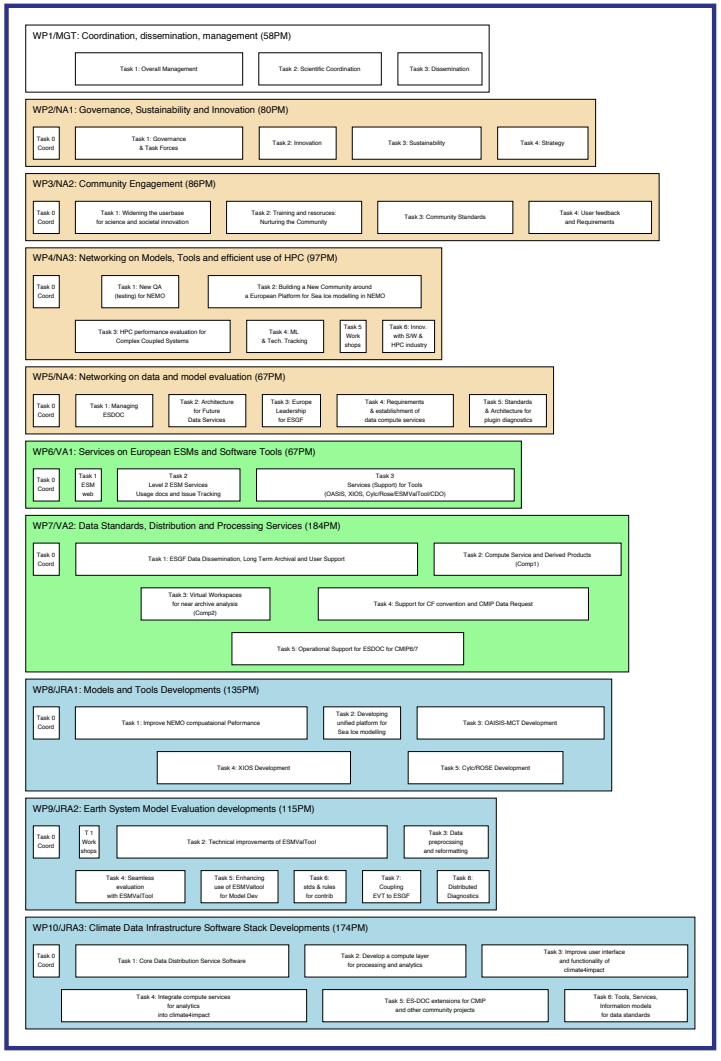
More proactivity needed!

- We will have to be much more proactive in demanding cost/benefit analysis for future activities! This will not be popular.

- "Download at home" did not work for CMIP5, yet there appears to be no real understanding by the designer/user community as to the consequences of the factor of ten in volumes expected for CMIP6.

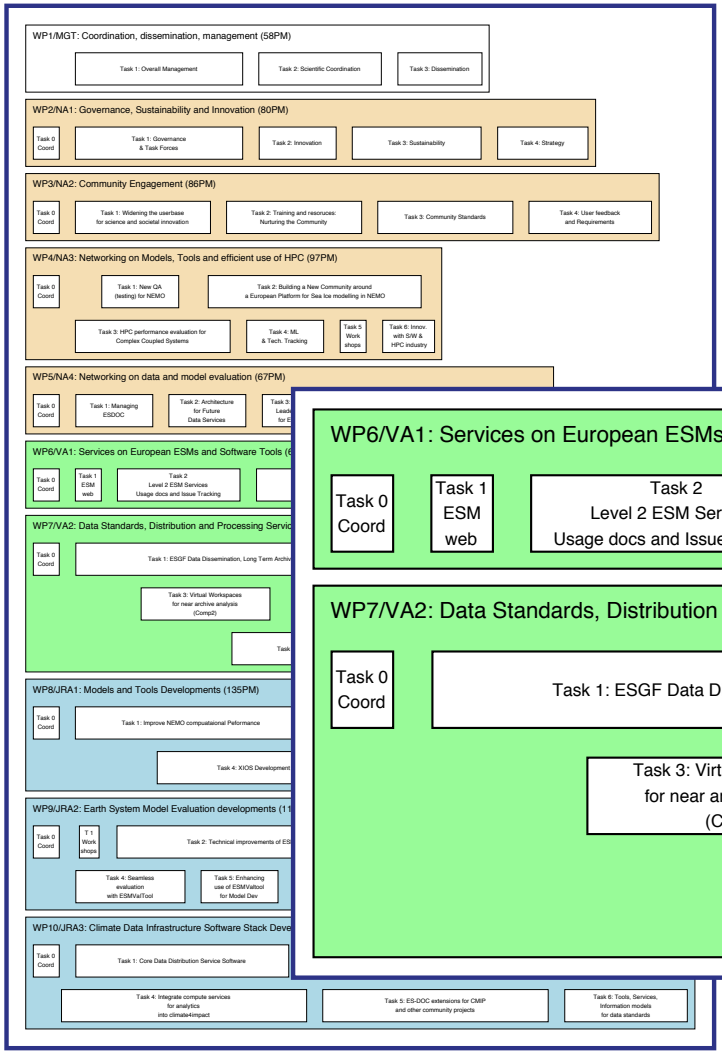
IS-ENES3: Internal Collaboration?

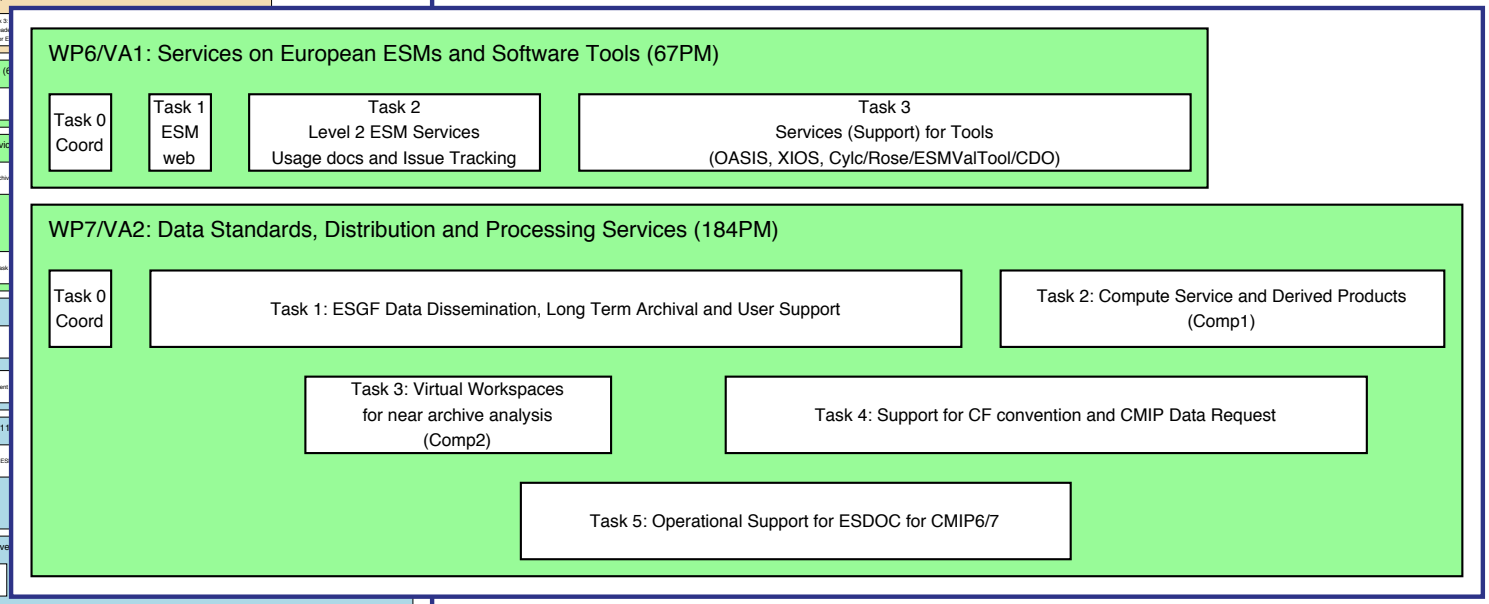
project



IS-ENES3: Internal Collaboration?

project



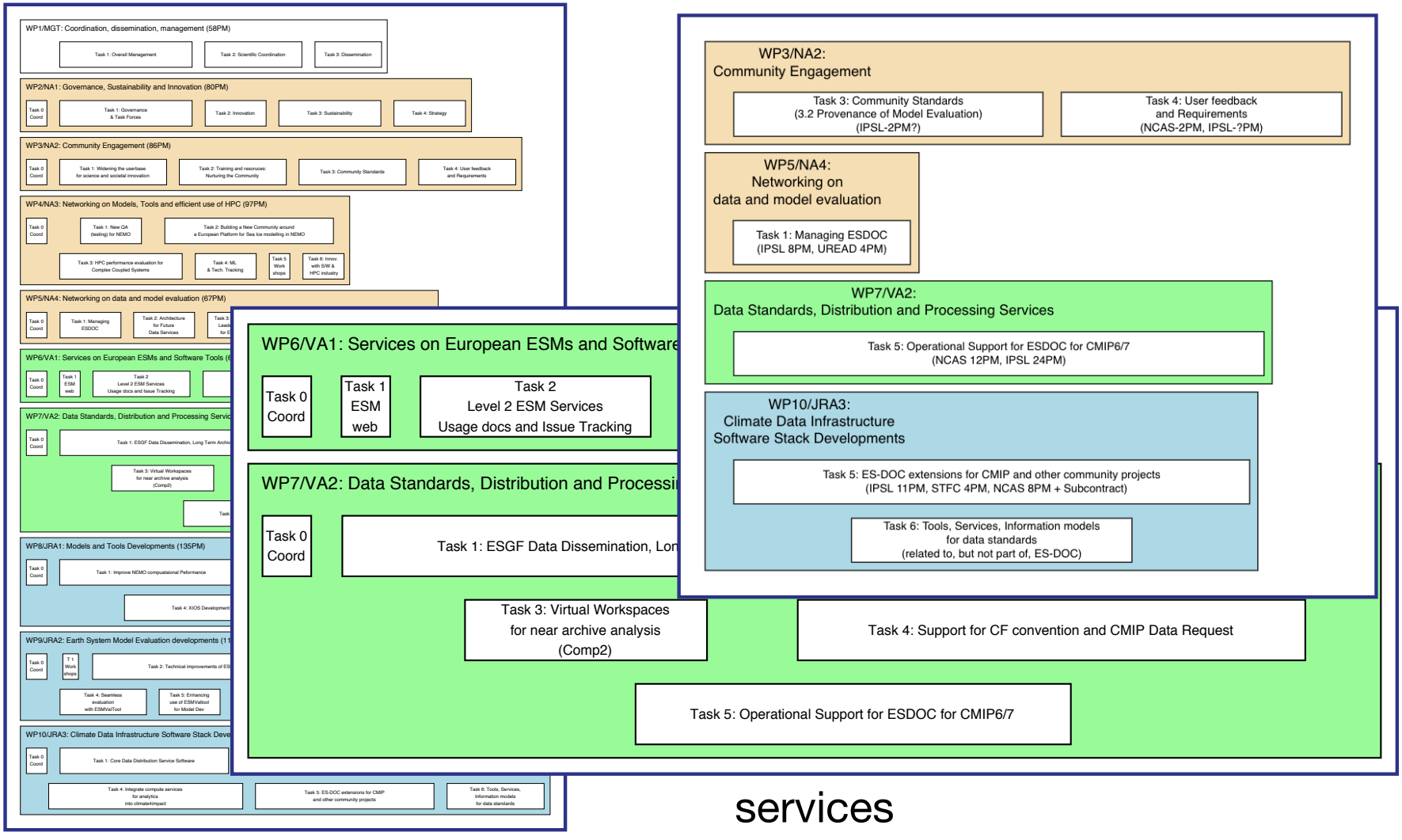


services

IS-ENES3: Internal Collaboration?

project

virtual wp - esdoc

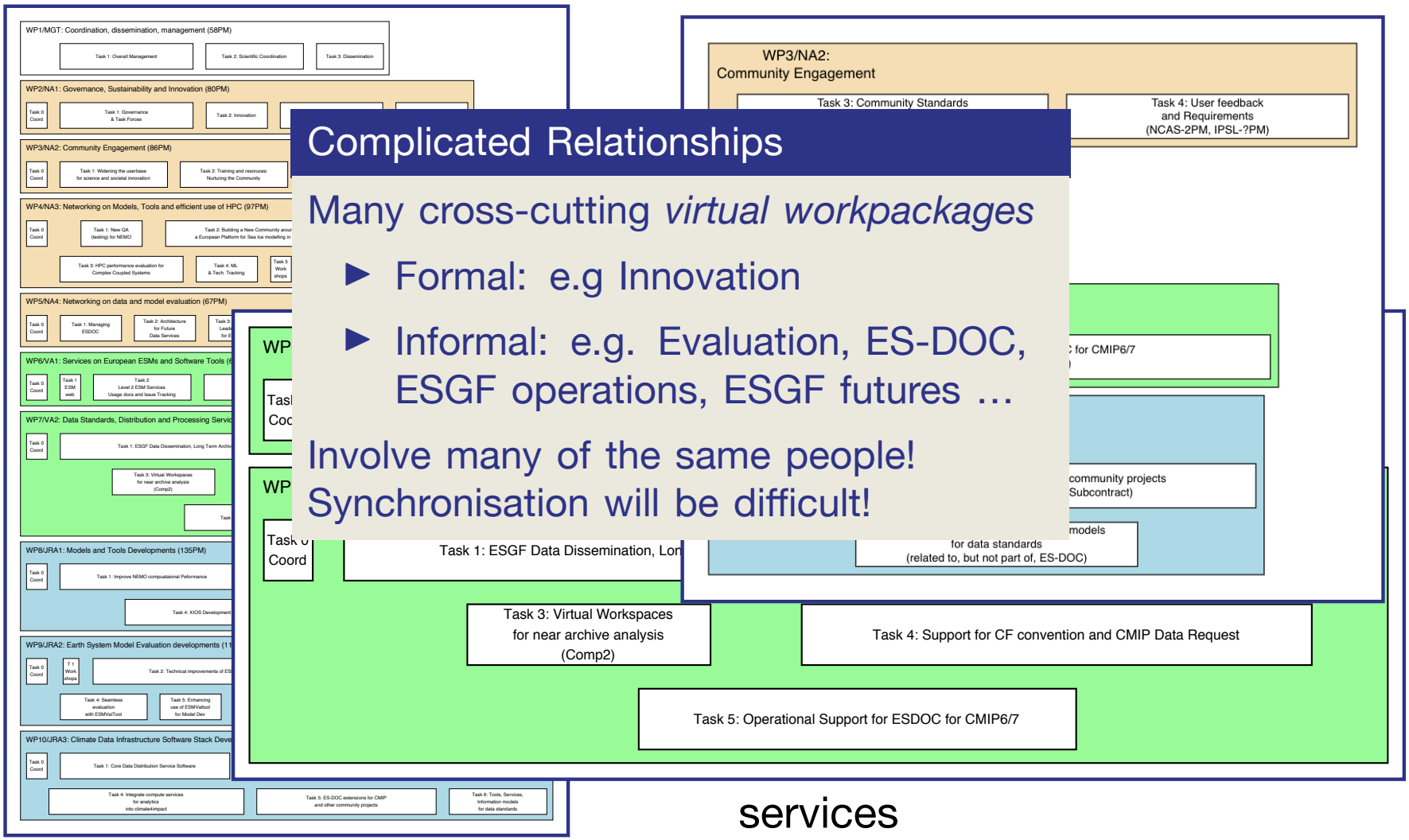


services

IS-ENES3: Internal Collaboration?

project

virtual wp - esdoc



Complicated Relationships

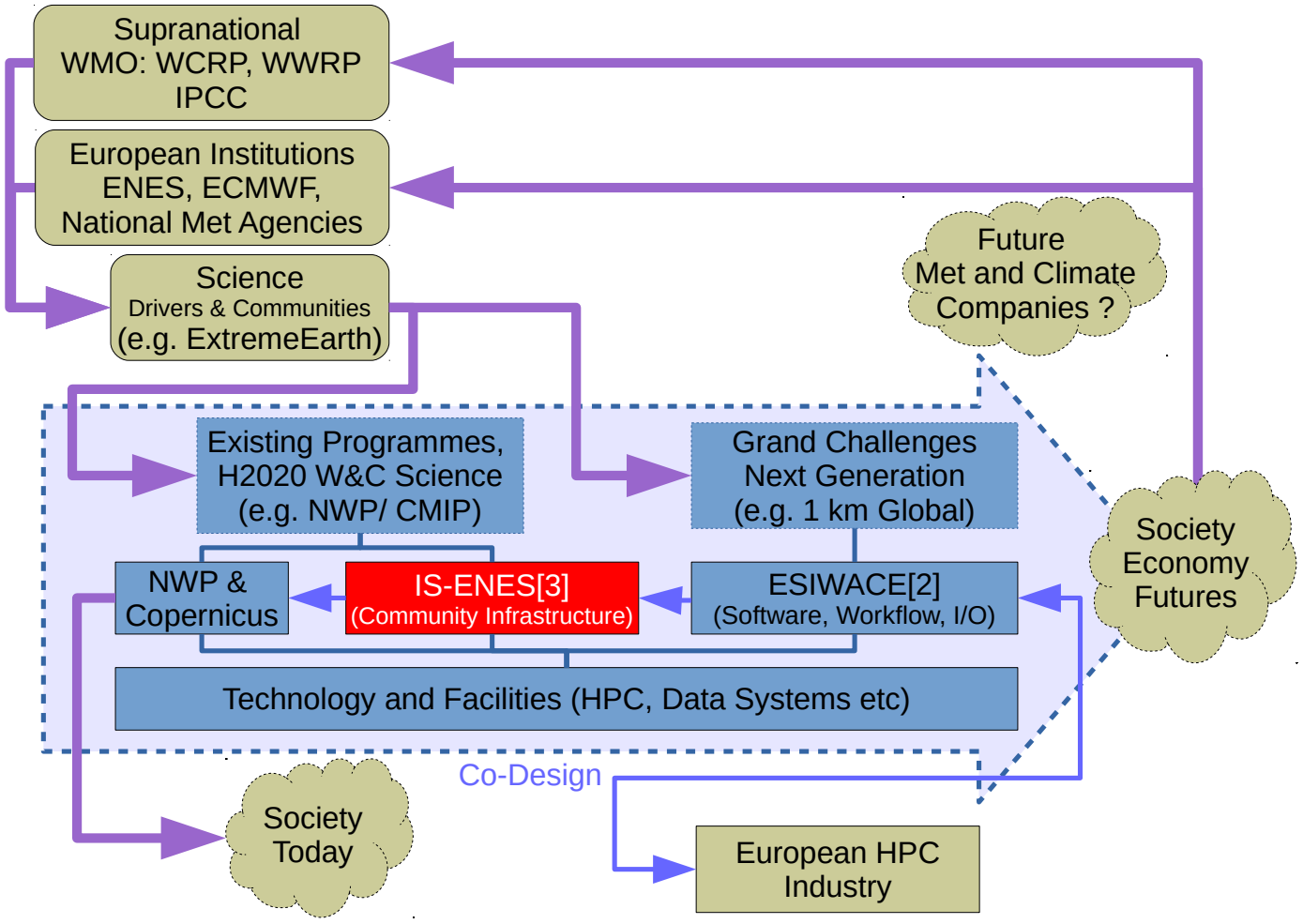
Many cross-cutting *virtual workpackages*

- Formal: e.g Innovation
- Informal: e.g. Evaluation, ES-DOC, ESGF operations, ESGF futures …

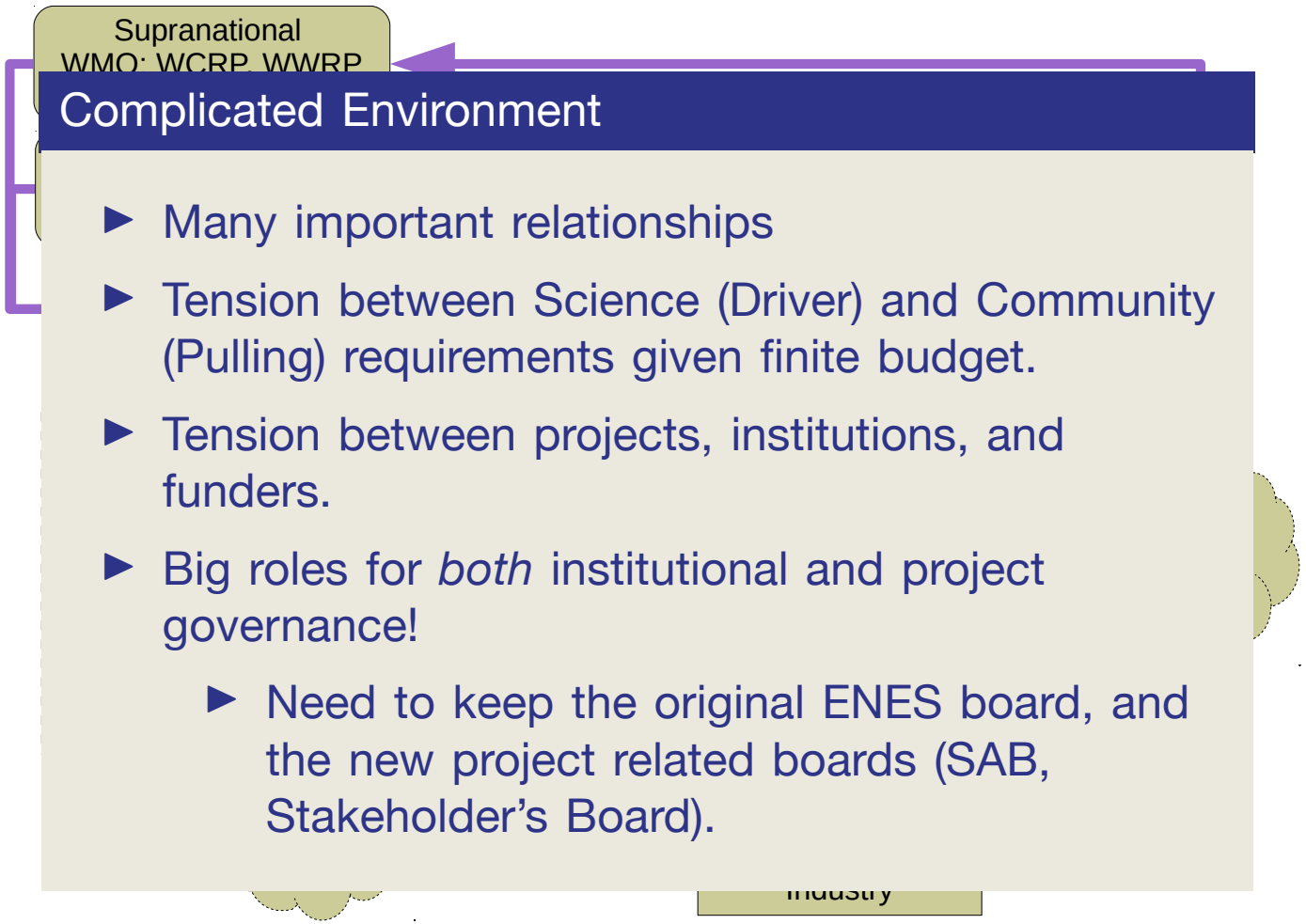
Involve many of the same people!
Synchronisation will be difficult!

services

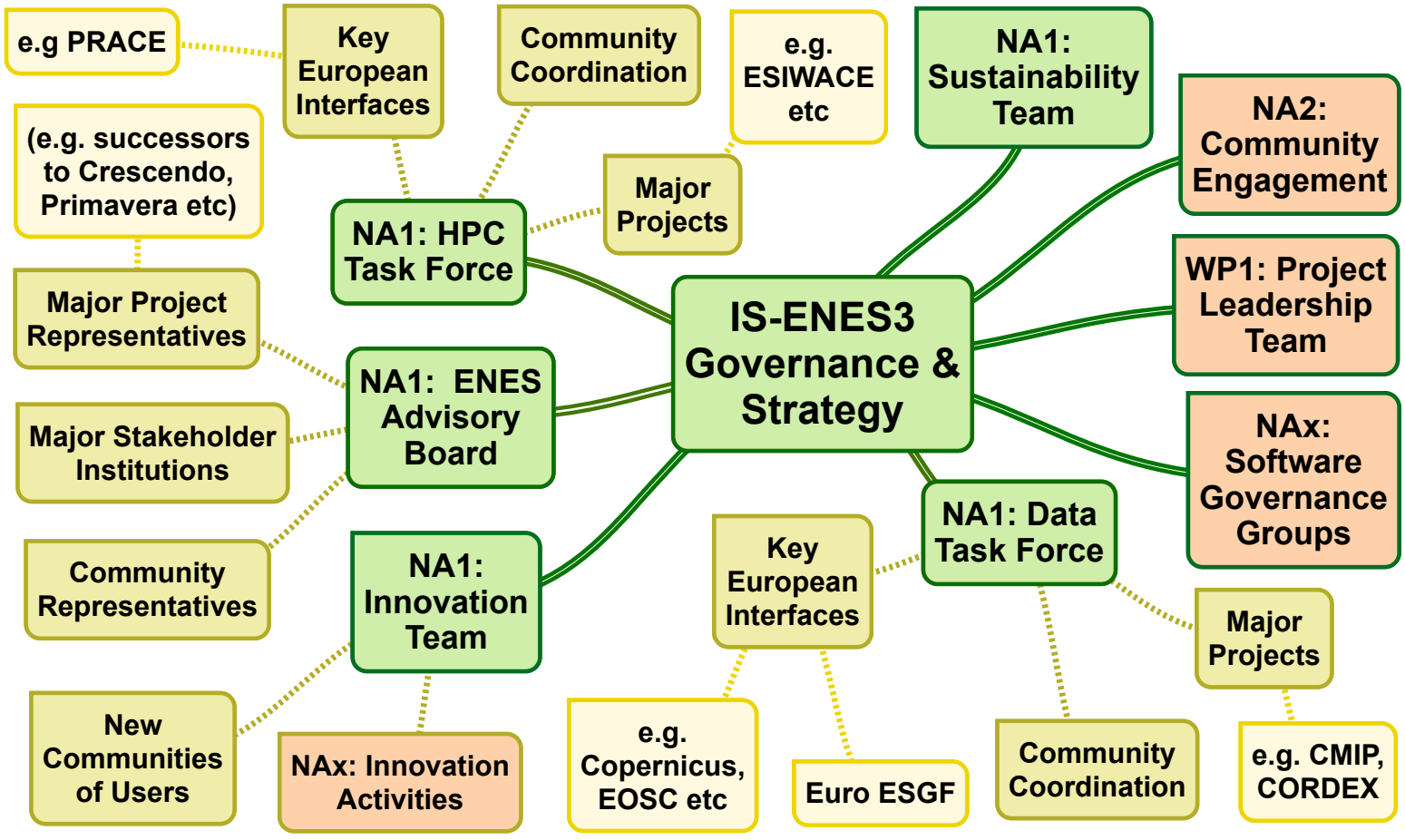
ENES: Projects and Institutions?



ENES: Projects and Institutions?



Everyone else: Stakeholders and Big Players



Everyone else: Stakeholders and Big Players

e.g PRACE

Key European Interfaces

Community Coordination

e.g. ESIWACE etc

NA1: Sustainability Team

NA2: Community Engagement

(e.g. successors to Crescendo, Primavera etc)

WP1: Project Leadership Team

Major Project Representatives

Major Stakeholder Institutions

NAx: Software Governance Groups

Community Representatives

Major Projects

New Communities of Users

NAx: Innovation Activities

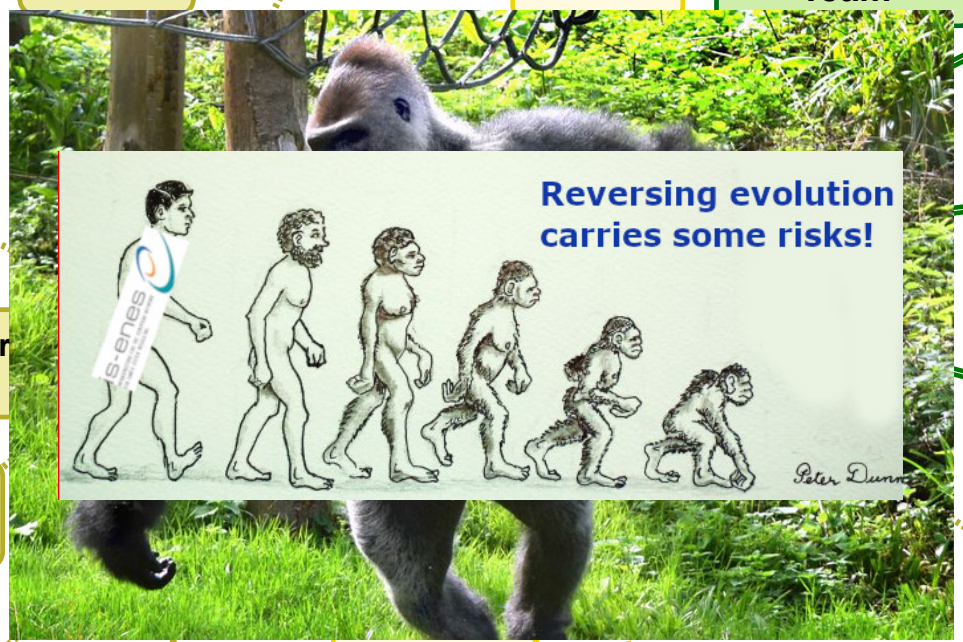
e.g. Copernicus, EOSC etc

Euro ESGF

Community Coordination

e.g. CMIP, CORDEX

Everyone else: Stakeholders and Big Players



e.g PRACE

Key European Interfaces

Community Coordination

e.g. ESIWACE etc

NA1: Sustainability Team

NA2: Community Engagement

(e.g. successors to Crescendo, Primavera etc)

WP1: Project Leadership Team

Major Project Representatives

Major Stakeholder Institutions

NAx: Software Governance Groups

Community Representatives

Major Projects

New Communities of Users

NAx: Innovation Activities

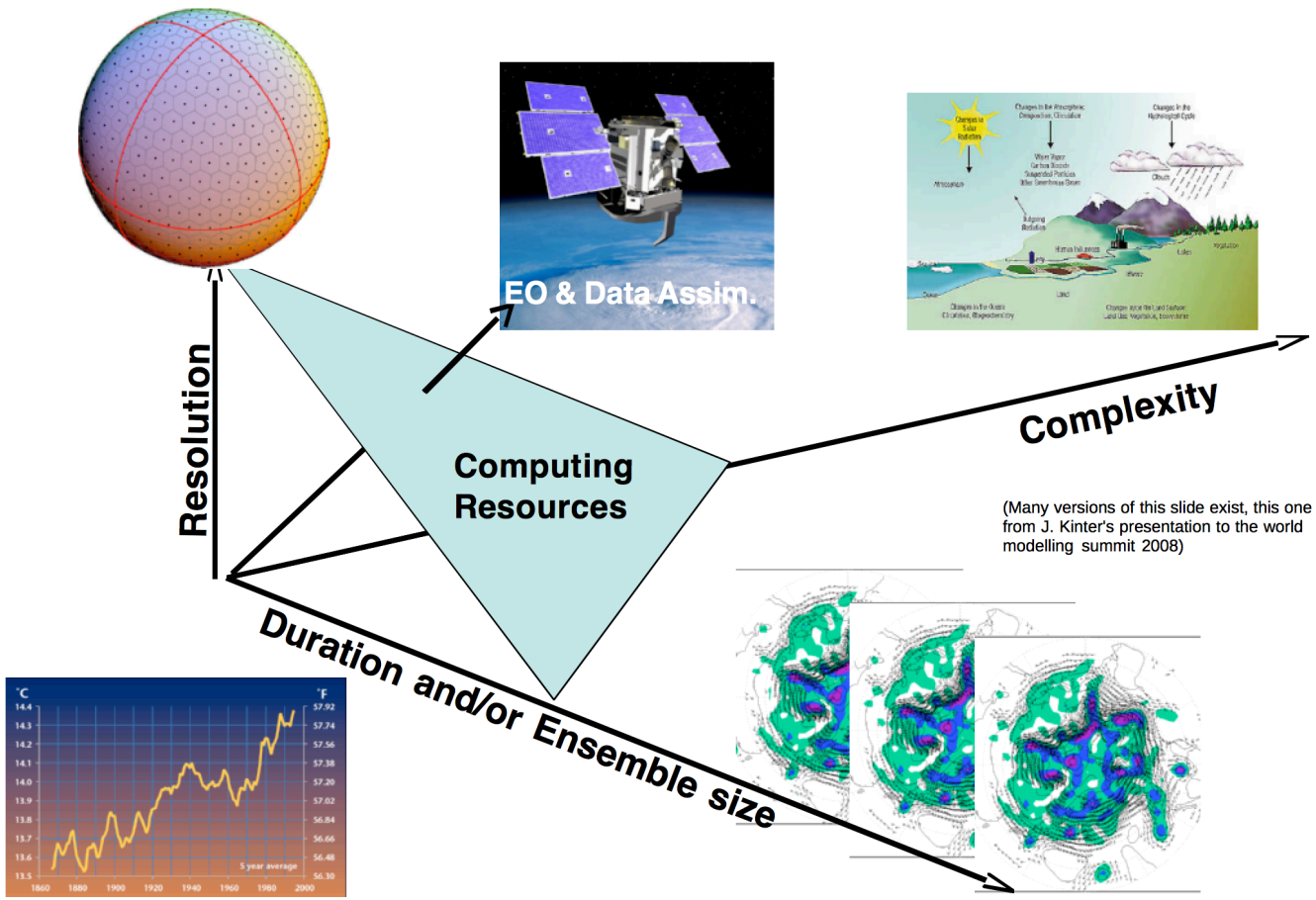
e.g. Copernicus, EOSC etc

Euro ESGF

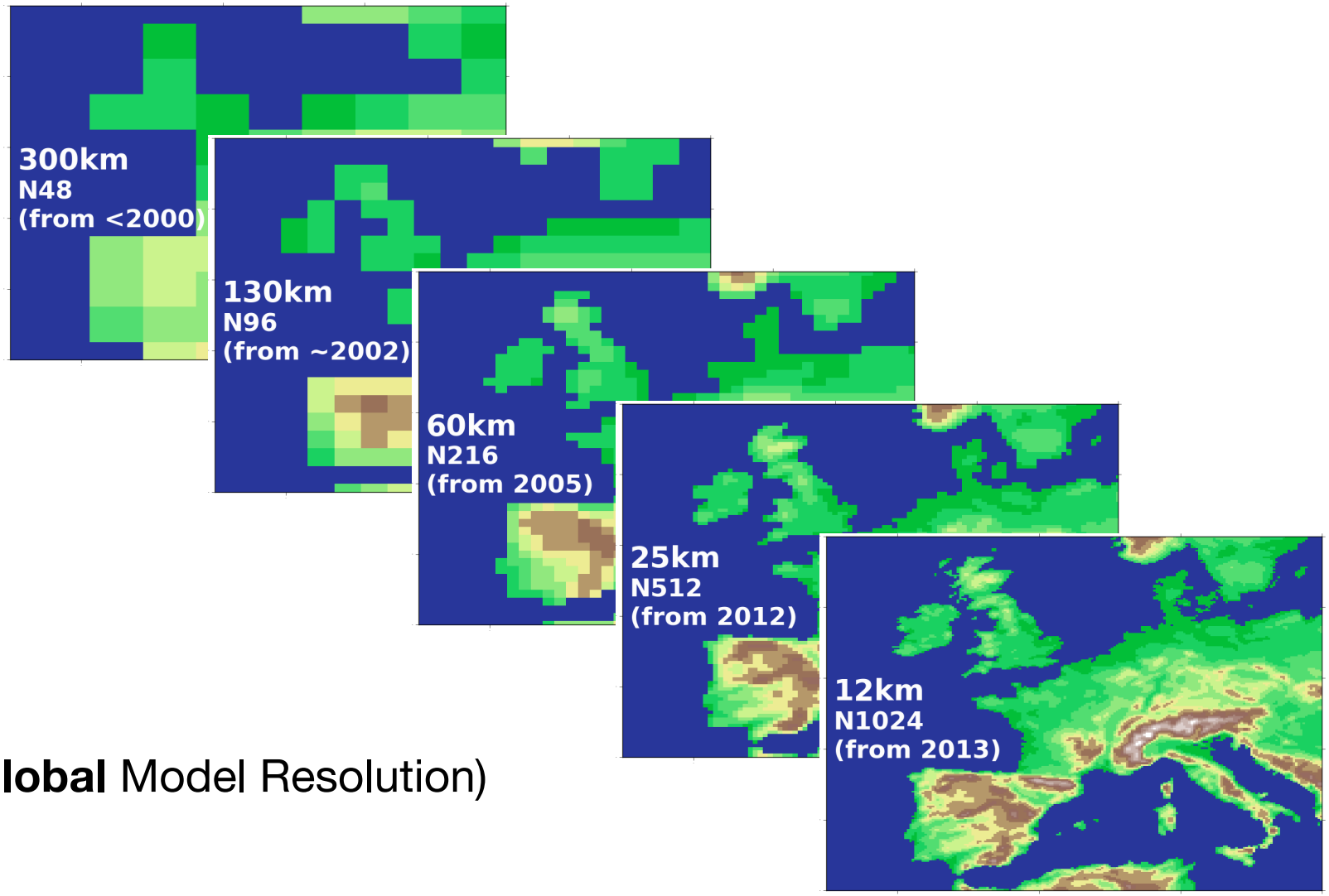
Community Coordination

e.g. CMIP, CORDEX

We are used to thinking like this:

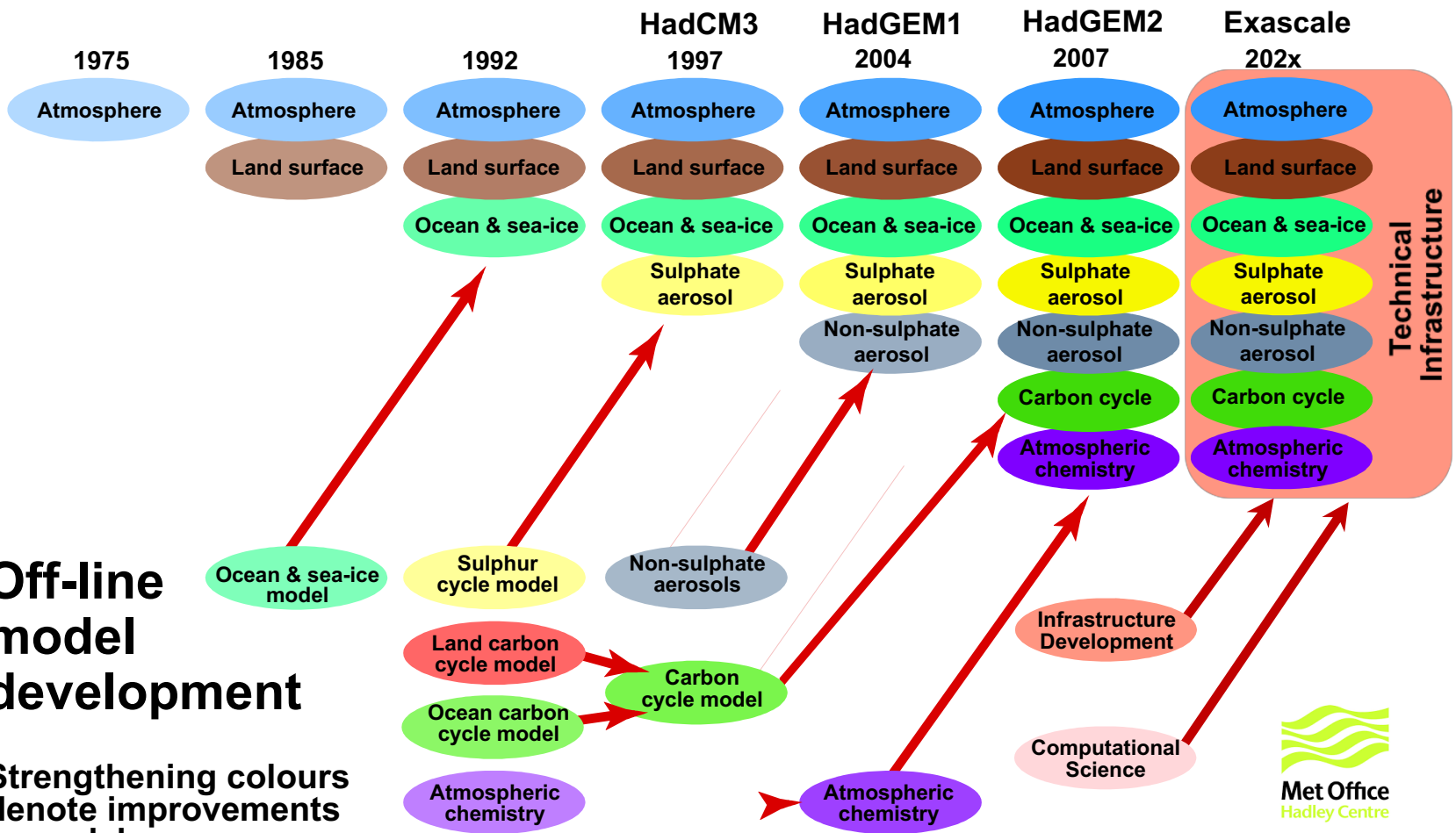


Evolution of Resolution

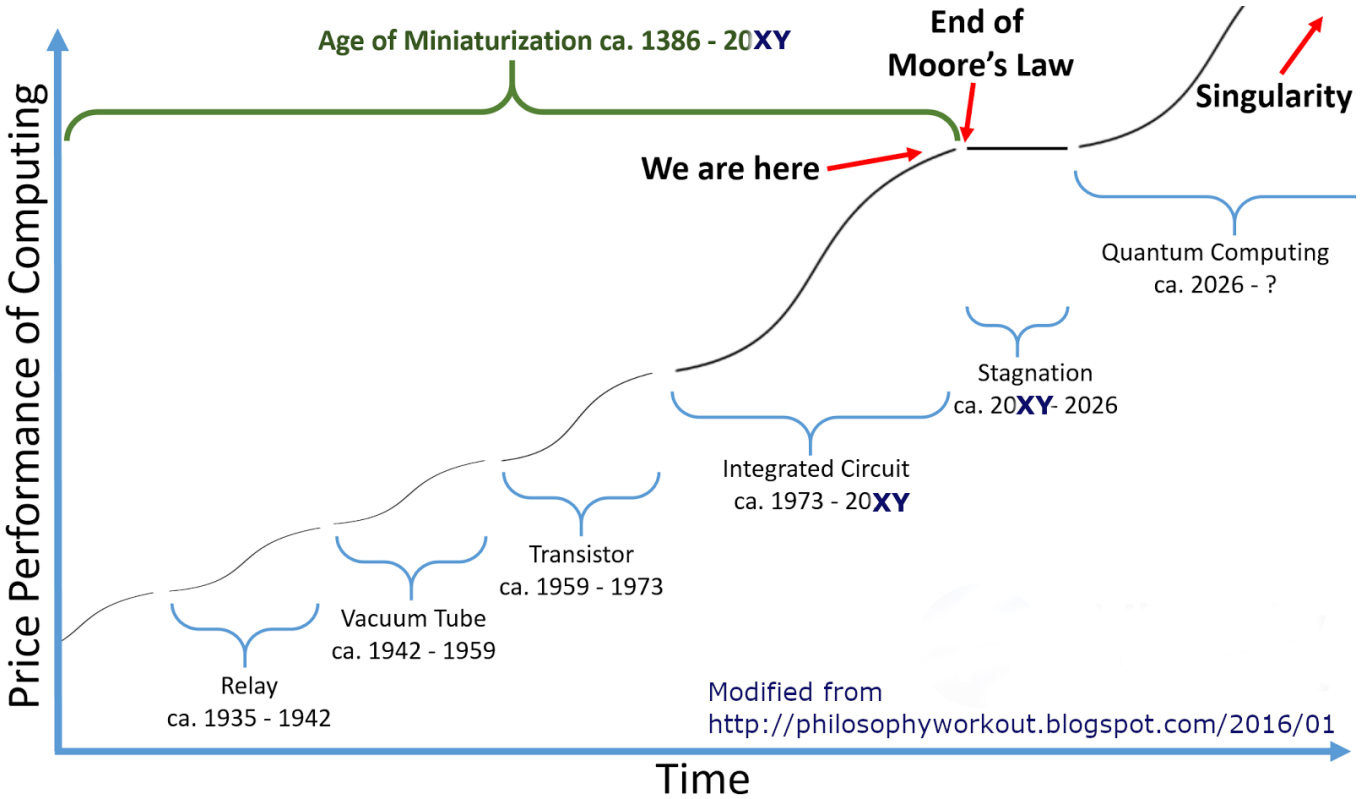


(**Global** Model Resolution)

Evolution of Complexity

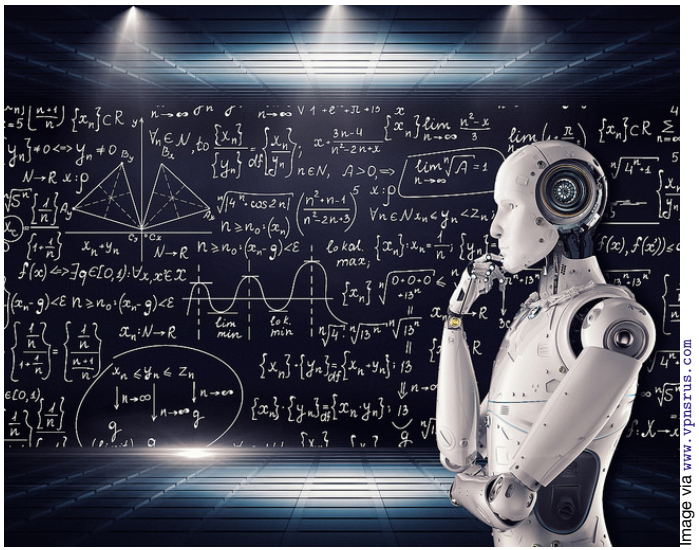


What next then?



The solution involve mathematics, algorithms and software, not hardware!

The rise of AI and ML.

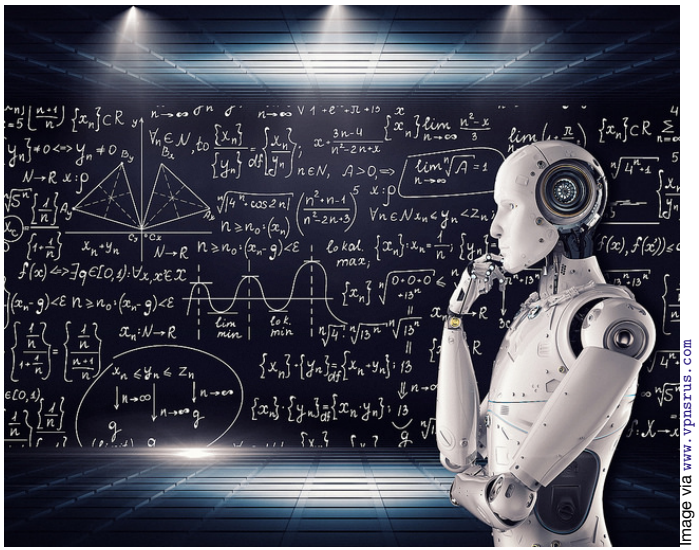


Gratuitous “robots are coming” image

Expect ML and AI to have major implications for both

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

The rise of AI and ML.



Gratuitous “robots are coming” image

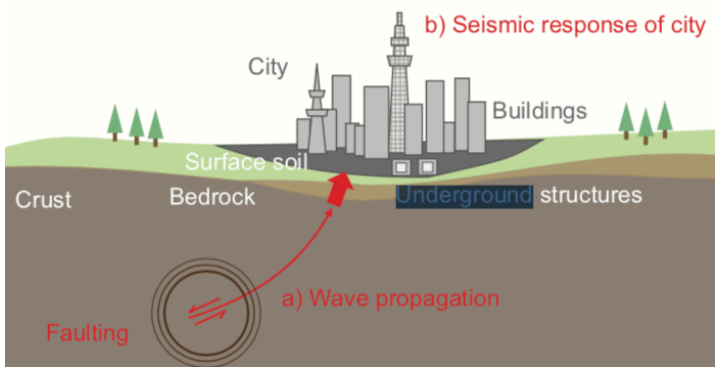
Expect ML and AI to have major implications for both

- 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.

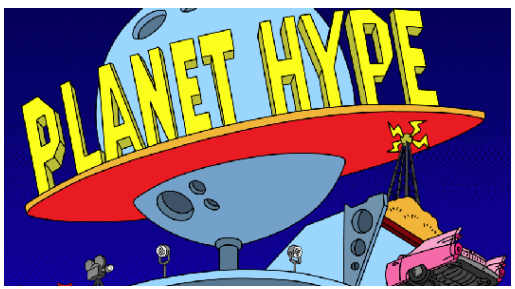
Morphing to a new world based on AI/ML?

A new way of modelling rising from the ashes of the old word?

Morphing to a new world based on AI/ML?

A new way of modelling rising from the ashes of the old word?

OR



The challenge for ENES and IS-ENES3 …

- We cannot ignore the opportunities (and risks)!

For ENES:

- When and how do we (or even, *should we*) move from individual investigations of possibilities to collaborative endeavours?

For IS-ENES3:

- When do these activities become infrastructure? (Not within IS-ENES3 timeline?)
- How do we balance innovation opportunities within the project, with those outside?
- Does this belong to ESiWACE? Is it possible for an EC project like IS-ENES3 to “give up” the exciting opportunities to another project and still be relevant?

Summary

Lots of challenges

- We live in computationally challenging times …
- We need to minimise the challenge of big numbers, and
- …communicate that the current *methodology* is not sustainabile.
- We have a complicated project with lots of moving parts, with
- Many stakeholders and (potentially) existential threats elsewhere.
- Unlike old dogs, we will need new tricks (maybe ML/AI), but which ones are not yet obvious (maybe not).
- It is difficult being *an infrastructure* during periods of great change.
- I said nearly nothing about users; users are challenging, but working on an infrastructure without users is called archaeology.

Summary

Lots of challenges

- We live in computationally challenging times …
- We need to minimise the challenge of big numbers, and
- …communicate that the current *methodology* is not sustainabile.
- We have a complicated project with lots of moving parts, with
- Many stakeholders and (potentially) existential threats elsewhere.
- Unlike old dogs, we will need new tricks (maybe ML/AI), but which ones are not yet obvious (maybe not).
- It is difficult being *an infrastructure* during periods of great change.
- I said nearly nothing about users; users are challenging, but working on an infrastructure without users is called archaeology.
- If it wasn't hard it wouldn't be worth doing it …