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

EuroHPC: Requirements from Weather & Climate

ECMWF

enes
EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

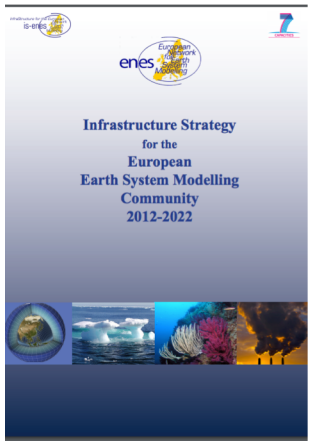
Bryan Lawrence
NCAS, University of Reading

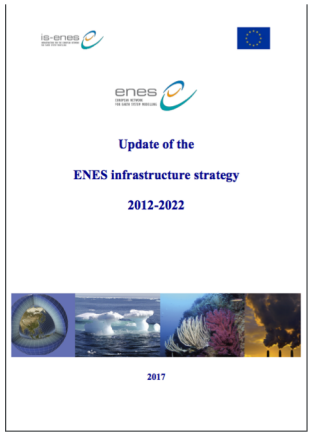
National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

Contributing lessons learned from within our community where we have considerable experience running, using, benchmarking, and procuring leading HPC systems.

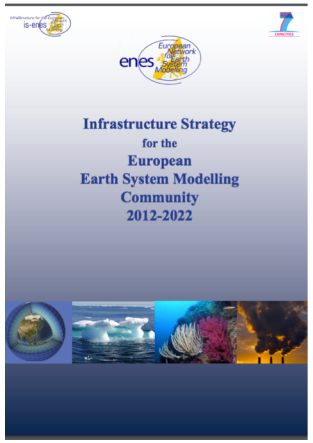
ENES Infrastructure Strategy (2012, updated 2017)

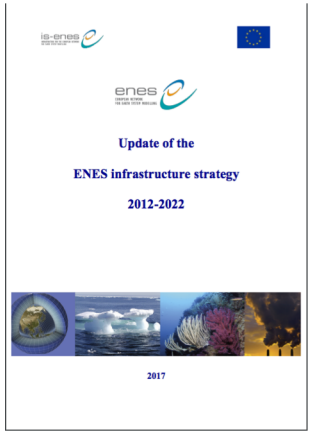




- The **2017 recommendations** are:
- **1. On models**: Support **common development** and sharing of software and **accelerate the preparation for exascale computing** by **exploiting next generation hardware** and developing appropriate algorithms, software infrastructures, and workflows.
- **2. On HPC: Exploit a blend of national and European high-performance facilities** to support current and next generation science and work toward obtaining sustained access to **world-class resources and next generation architectures**.
- **3. On model data:** Evolve towards a sustained data infrastructure providing data that are easily available, well-documented and quality assured, and further invest in research into data standards, workflow, **high performance data management and analytics**.
- **4. On physical network:** Work to **maximize the bandwidth between the major European climate data and compute facilities** and ensure that documentation and guidance on tools and local network setup are available to users.

ENES Infrastructure Strategy (2012, updated 2017)



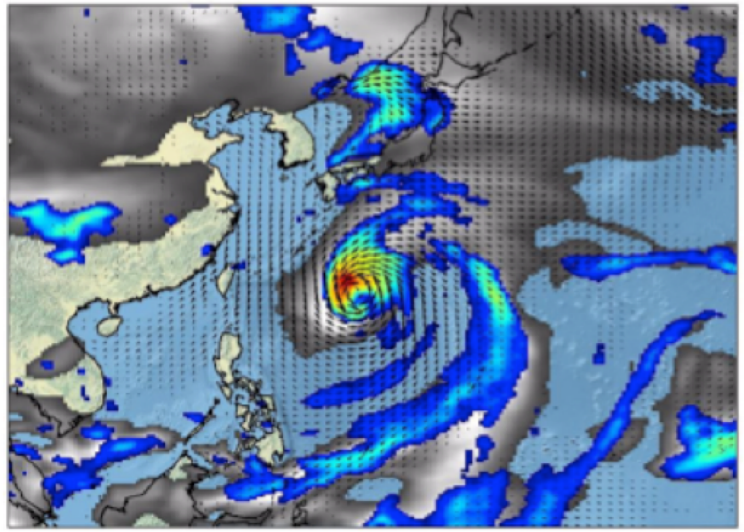


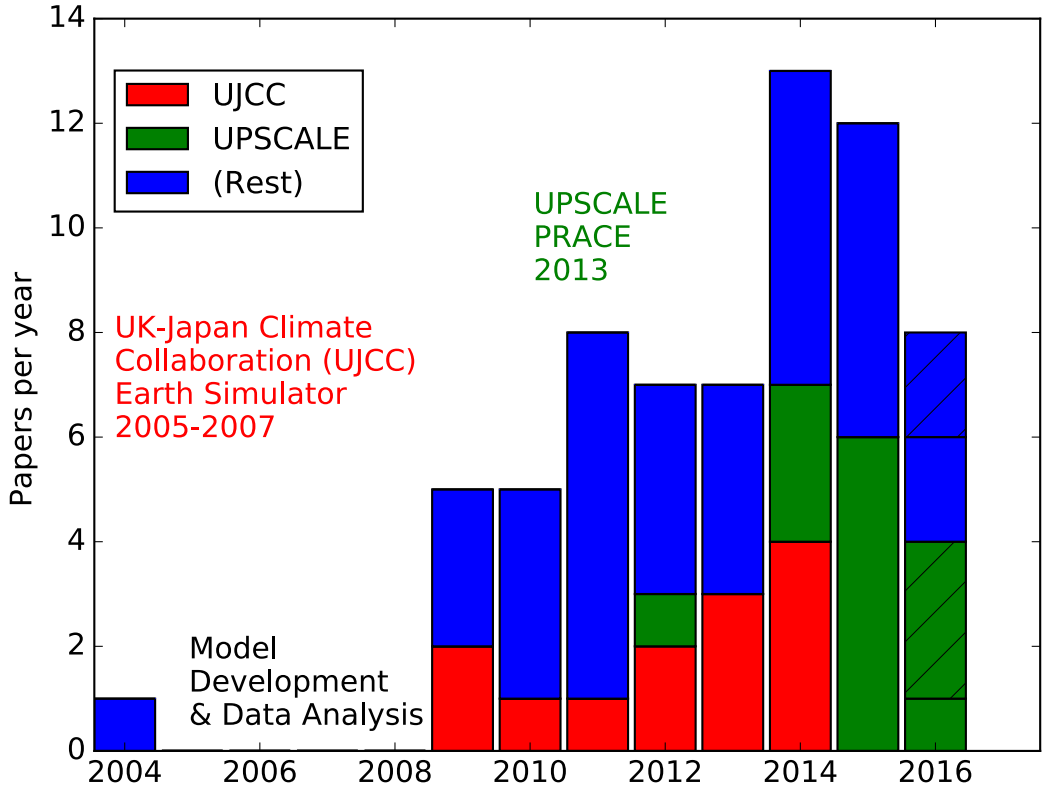
- **5. On people:** Grow the numbers of skilled scientists and software engineers in the ENES community, increase opportunities for training at all levels, and strengthen networking between software engineers**.**
- **6. On model evaluation (new):** Enhance sharing of common open source diagnostics and model evaluation tools, implement governance procedures, and **expand data infrastructure to include computational resources** needed for more systematic evaluation of model output.
- **7. On infrastructure sustainability (new):** Sustain the cooperation necessary to develop future model and data technology **and support international reference experiments** programmes, and **strengthen collaboration with other European actors providing services** to, or using services from, ENES.

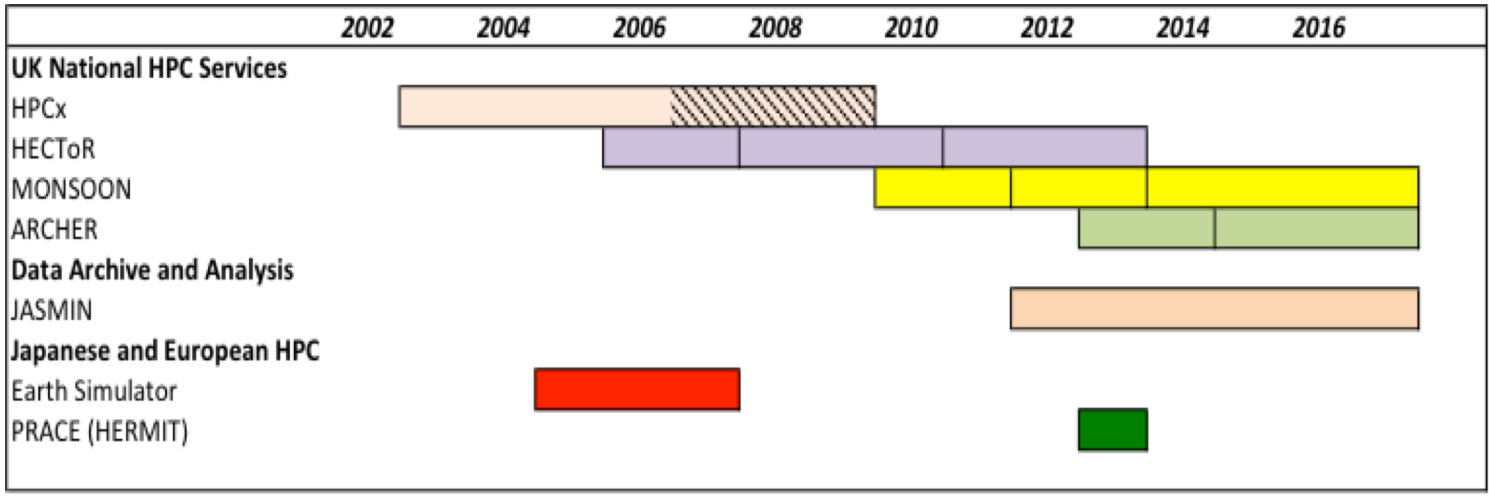
Logistics Example: High Resolution Climate (UPSCALE)

UPSCALE project was a large chunk of a tier-0 machine (HERMIT) for a year (2012)

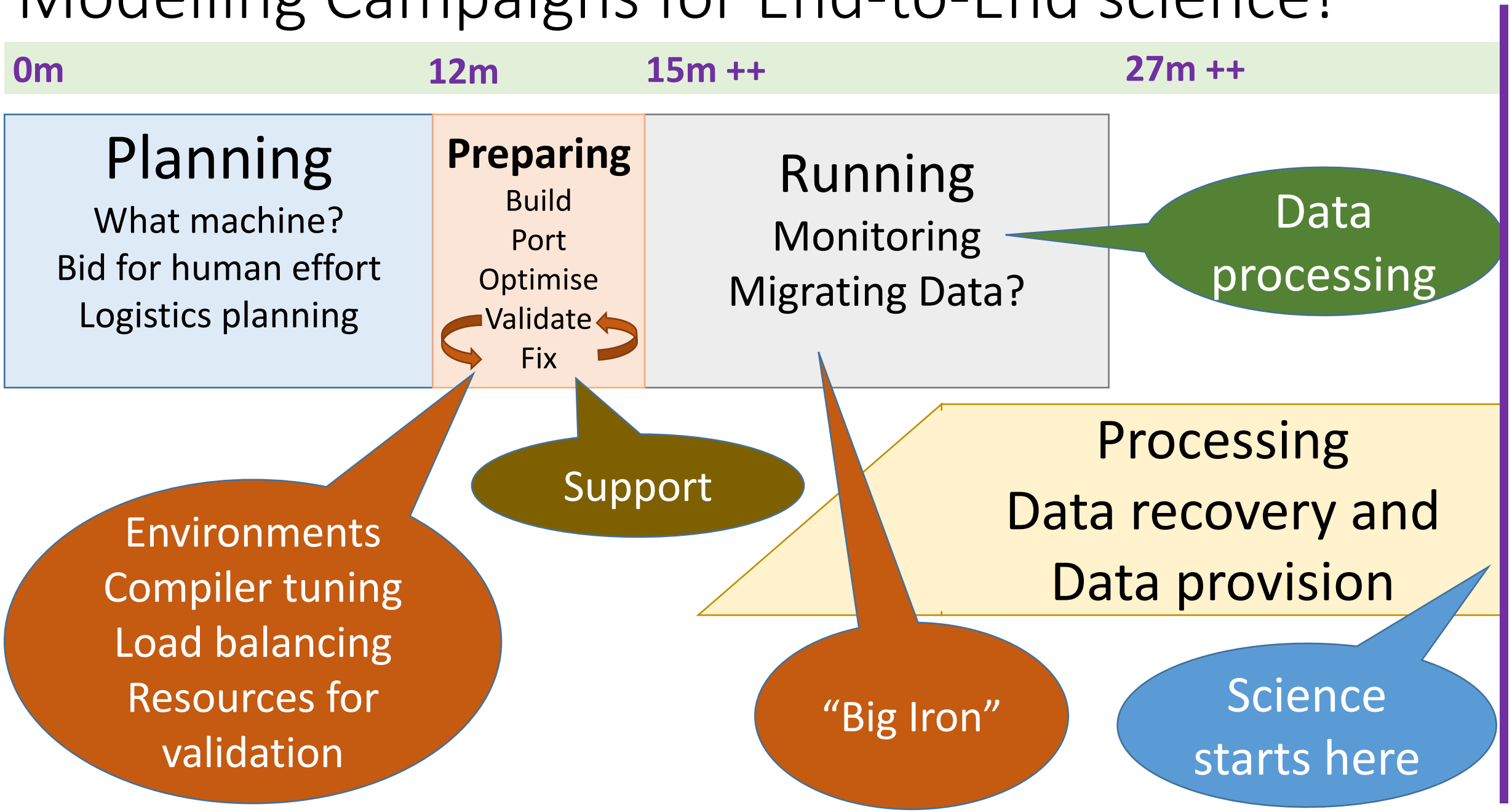
- But it was part of a long term programme,
- Depended on data logistics (and a data platform that was serendipitous, not planned)
- The data ***USER COMMUNITY*** is still exploiting the data and writing publications.







Modelling Campaigns for End-to-End science!



Planning

Context
National and International

Human Resources
Dev, Config
Run, Migrate

Data Logistics
Analysis
Archival

Planning and Logistics

Allocation mechanism must recognize simulations belong in **wider context**. Benefits can come from:

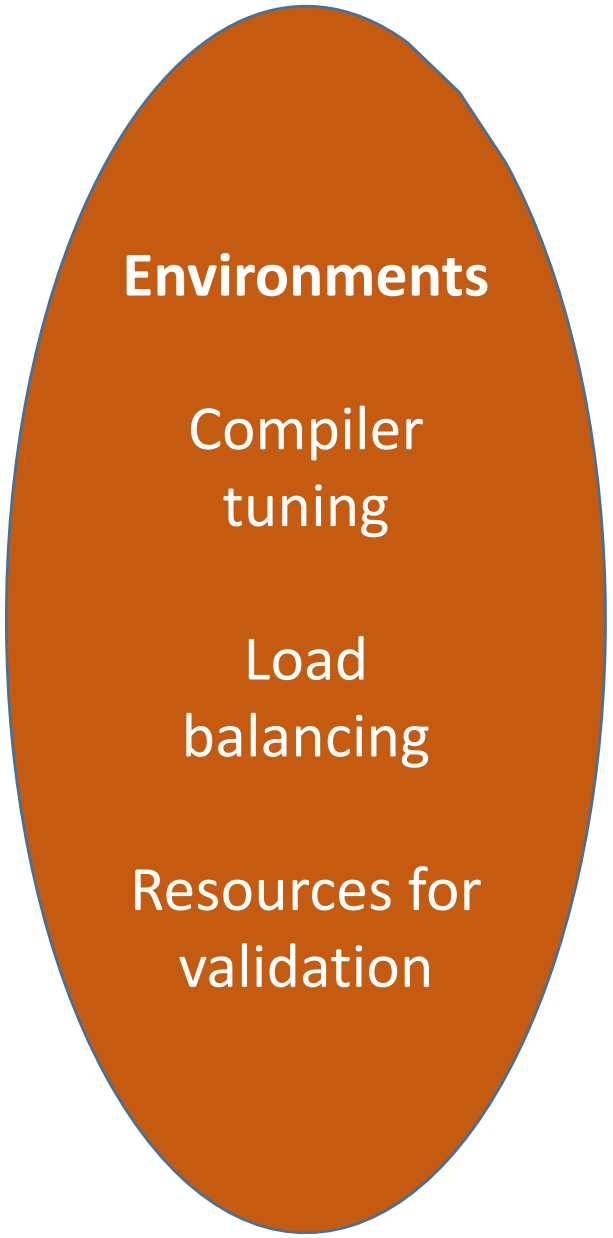
- Bigger picture: WCRP grand challenges, CMIP etc;
- Exploitation of data by much wider audience (c.f. satellites).

Large scale simulation programs are logistic campaigns:

- Need dedicated **human resources** to develop(modify) code for target platform, configure experiments, manage the runs (over months), (potentially) migrate data.
- **Needs HPC allocation mechanism to be synchronized with science grant mechanism** (e.g. H2020 – HPC without funded science projects *and vice versa* does not work).

Data is the output of the first stage of a programme. Science analysis is the objective – need to **plan and support data analysis systems and logistics**.

- Where is the analysis going to be carried out? What other data is necessary to analyse the simulations (e.g. observations, other simulations). Where is the data to be archived, and how documented?



Environment

- Must support *multiple executables* – we couple together *multiple codes* which *exchange data* at run time!
- Require high performance, well maintained Fortran, C, C++, MPI, hybrid MPI/OpenMP.
- Needs to reflect need for *stable compiler environment* and *rapid response to compiler issues* (must support modules or equivalent).
- Batch queue environment needs to be visible to a *persistent workflow scheduler* running onsite or offsite (e.g cylc).
- Need to deal with *large volumes of both input and output data.*

"Big Iron"

Performance Metrics

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

We are primarily interested in **Simulated Years Per (real) Day (SYPD)** – the key measure of **speed**. This is a function of:

- Resolution (number of degrees of freedom in grid)
- Complexity (number of real world parameters simulated)

We are interested in **node-hours per simulated year (NHSY)** and **joules per simulated year (JPSY)** – both measures of **cost**. They are a trade off between

- raw speed and throughput (given imperfect scaling).

Key factors which influence these are:

- *memory bandwidth, raw flops, interconnect (both latency and bandwidth)*.

Our models have significant **Data Intensity** (measured in GB/core-hour),

- *file system performance* matters too!

Balaji et .al.: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geosci. Model Dev., 10, 19-34, https://doi.org/10.5194/gmd-10-19-2017, 2017.

"Big Iron"

Performance Metrics (2)

The influence of file system performance can be measured by the **Data Output Cost.** This is the normalized difference between a standard run and a run without an output, normalized:

$$(NHSY - NHSY_{no\_output})/ NHSY \quad \text{and} \quad SYPD - SYPD_{no\_output})/SYPD$$

Our codes always run faster in SYPD without I/O and ***benchmarking without I/O is very misleading!***

Model simulations run for days to months, and restarting with short queues is expensive.

- The system environment influences the most important measure of speed: **Actual Simulated Years Per Day (ASYPD).**
- This is a function of *system intermittency* (whether planned or involuntary), *queue wait time*, *queue duration*, and *issues with the model workflow* (e.g. can the output data be migrated to archive fast enough, can the initial conditions be kept on storage visible to the compute nodes etc).

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

Community Strategy

Planning

What Machine?

1. For pre-exascale we have no choice: Traditional models have to be adapted to the technology we expect in 2020-21 – this cannot be disruptive – EVOLUTION
   -> Nearly all (all?) production earth system models and global NWP models are CPU based, and there is not enough time to evolve them …

2. For exascale we need a game changer: FET projects will create elements. Much more needed - this can be REVOLUTION.

"Big Iron"

Benchmarking (1 – PreExascale)

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

- Key need to demonstrate ***science throughput*** of new machine, not FLOPS or any arbitrary measure.
- When we talk about "speed" of new machine, it needs to reflect the ratio of
  - "work done on reference machine"/"work possible on new machine"
  - in *science units (e.g. SYPD) at a fixed required speed for a given resolution (with full I/O).*
- *Expect to see benchmarks which are a representative bag of science codes weighted against expected workload. One of which should be from our community – community can select appropriate code!*
- Need to take care to interpret any speed up achieved by vendors in code optimisation.
  - Will this achieve results in production? Are they scientifically valid?
  - Are these speed ups achievable in the normal development production cycle?

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

Community Strategy

Planning

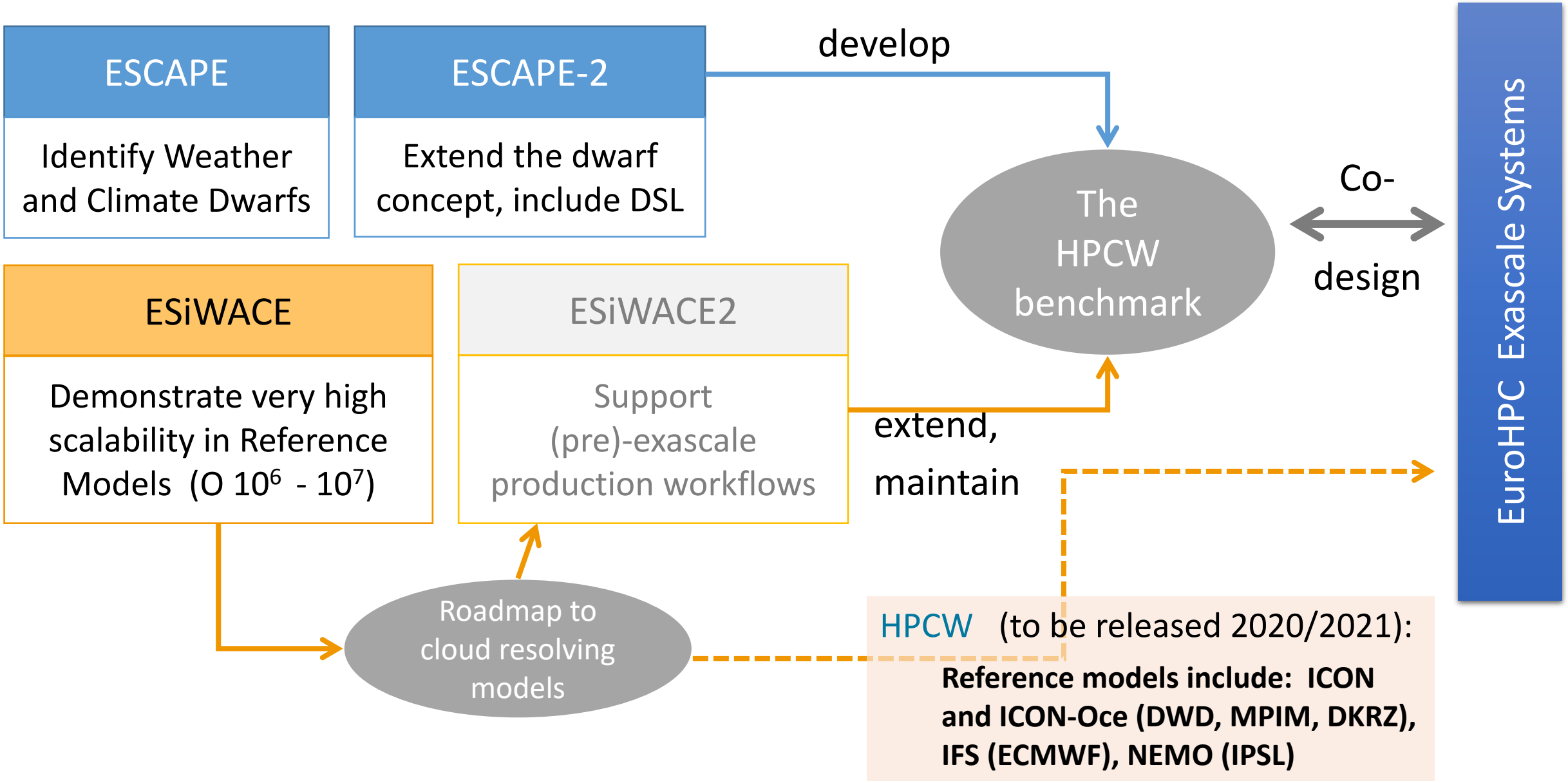
What Machine?

1. For pre-exascale we have no choice: Traditional models have to be adapted to the technology we expect in 2020-21 – this cannot be disruptive – EVOLUTION

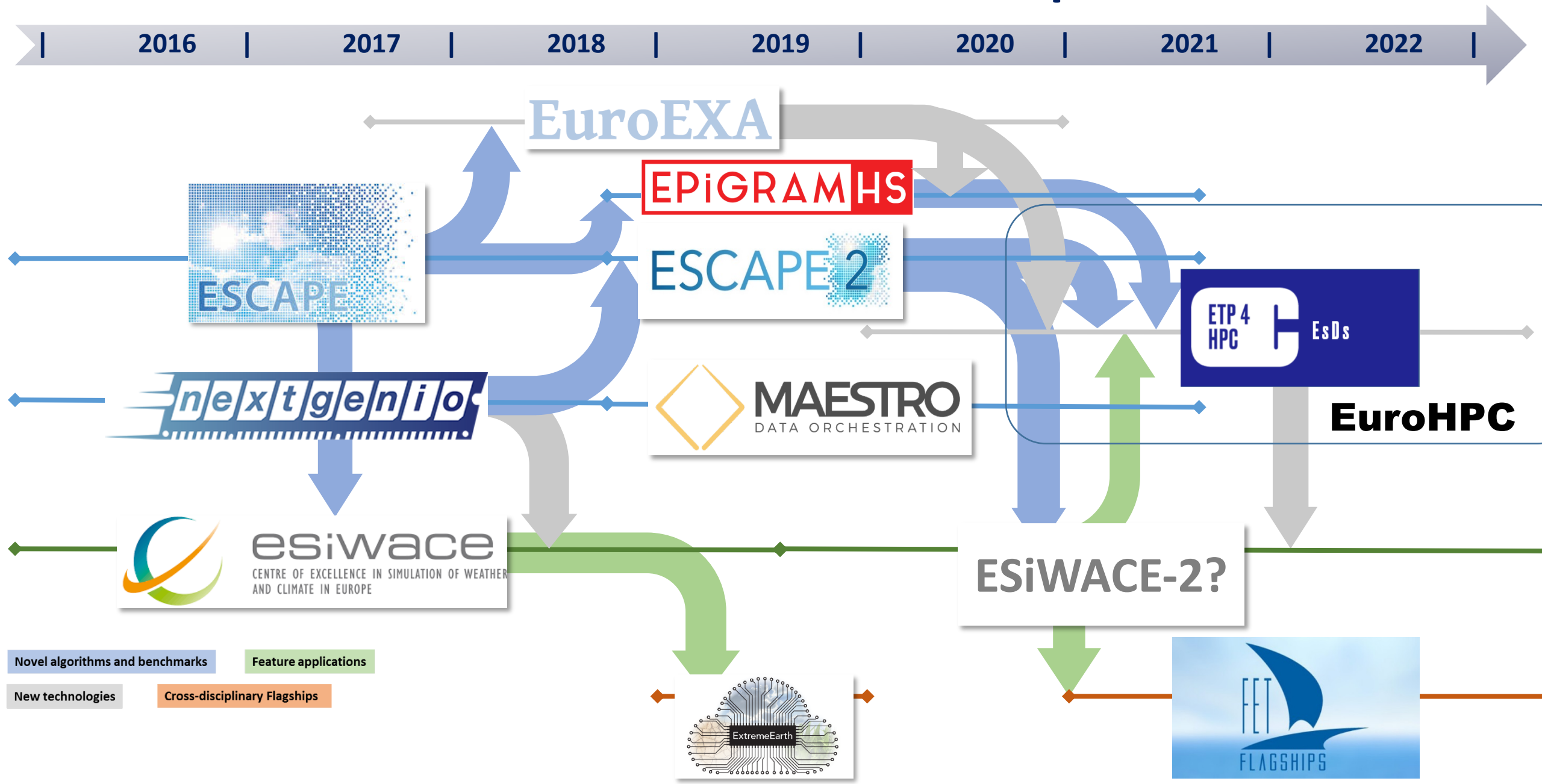
   -> Nearly all (all?) production earth system models and global NWP models are CPU based, and there is not enough time to evolve them …

2. For exascale we need a game changer: FET projects will create elements. Much more needed - this can be REVOLUTION.
   - Revolution needs not only new hardware, but massive investment in people and software.

Benchmarking (2) – Exascale: The HPCW benchmark



Weather and Climate roadmap in H2020



Simulations produce data - not knowledge!

Data processing platform

Science Analysis Platform

Persistent Services on Data

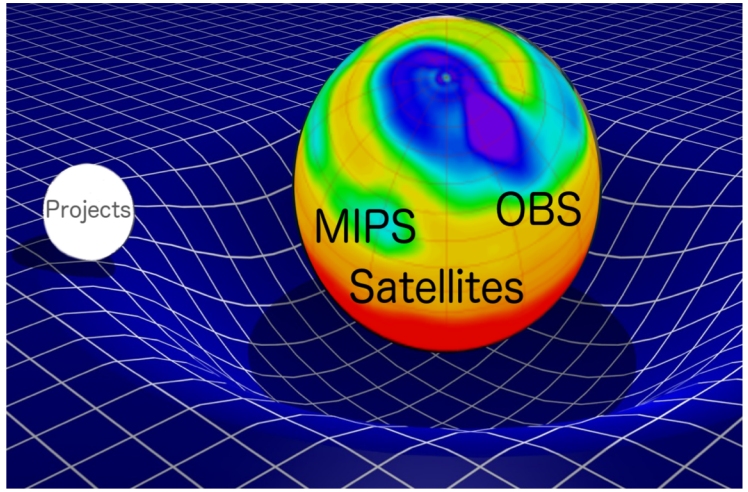
EuroHPC may not need to support the entire data ecosystem (especially persistent services on data), but Europe must – and EuroHPC will need to interface with it.

- ***Open Data Infrastructure at Scale (> 20% of HPC Budget)***!

  Examples: UK Joint Analysis System JASMIN (40 PB disk; tape curated archive; 11K cores; dedicated to analysis, not simulation). French IPSL mesocentre with dedicated network links.

- ***Growing role for AI/ML/Analytics***

Importance of "Data Gravity" to turning data into knowledge!



Platform as a Service	**Infrastructure as a Service**	**Software as a Service**
---	---	---
Traditional "login" and "batch"	Containers and Virtual machines	Persistent services: visualization, download etc

"Big Iron"

The right HPC environment

- High memory bandwidth
- (relatively) high memory per core
- Raw FLOPS
- Interconnect latency and performance
- CPU based in the first phase
  - For the initial EuroHPC timeframe we will not have prevalent European production science codes that can exploit GPU or accelerators

(Lawrence et.al., Geosci. Model Dev., 11, 1799-1821, https://doi.org/10.5194/gmd-11-1799-2018, 2018.)

- Excellent bandwidth and latency to storage, and then offsite.
  - Possible role for innovative memory and storage configurations, e.g. burst buffers etc.
- Petascale persistent disk
  - Persistent means persistent for months to years (more if local)!
  - Petascale means PBs if remote analysis, 10s of PB if local analysis.
  - … and then offsite means: 10s of Gbit/s with no firewall slowdown.
- Rich environment (compilers, long-duration queues, persistent service support for workflow etc).

Exascale Requirements go beyond hardware!

The transition to exascale will require a revolutionary approach to software, especially if it involves a significant element of co-design and/or a move away from CPUs (both of which seem likely)!

This leads to the requirement for

- Massive investment in revolutionary new codes, which will themselves require investment in a bigger workforce with a new set of skills.
- This requirement is reflected
  1. In the **fifth recommendation of the ENES Strategy: On people**: Grow the numbers of skilled scientists and software engineers in the ENES community, increase opportunities for training at all levels, and strengthen networking between software engineers.
  2. **The existence of the FET Flagship ExtremeEarth proposal!**

Acknowledgments

- Significant content from Peter Bauer, Joachim Biercamp, Mick Carter, Marie-Alice Foujois , Sylvie Joussaume, members of the ENES HPC Task force.
- Input from attendees at a relevant panel session at the 5th ENES HPC Workshop (Lecce, Italy, May 2018).

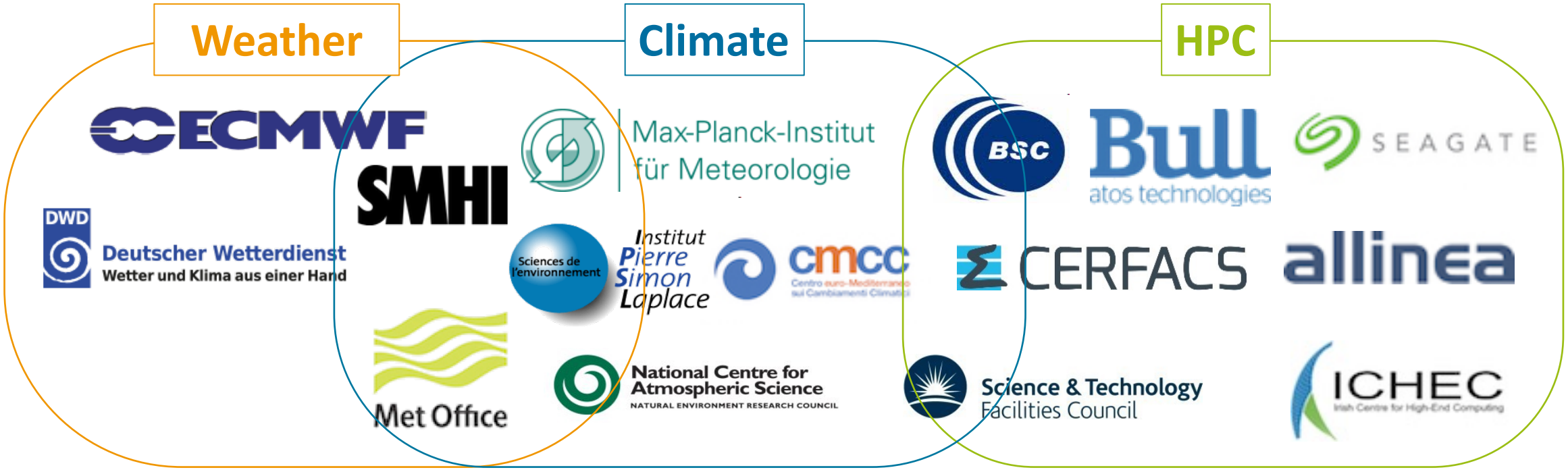
The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

Coordinator: DKRZ



The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programmed under grant agreement No 675191

edci European Extreme Data & Computing Initiative