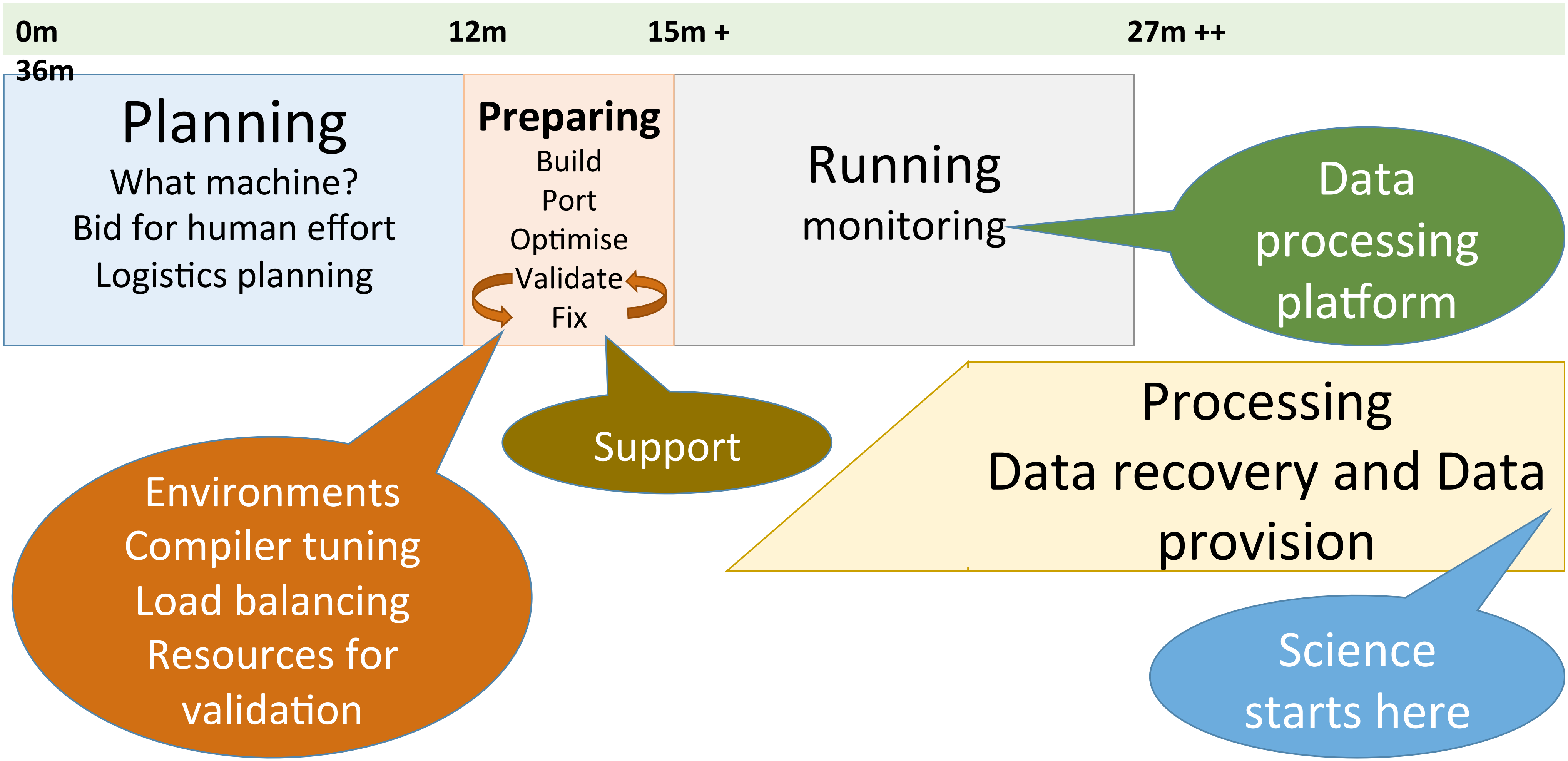
Data Centre Requirements for Weather & Climate Experiments

Mick Carter
Met Office Hadley Centre

Thanks to Professor Pier Luigi Vidale, Dr Grenville Lister, Dr Malcolm Roberts

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

End-to-End science?



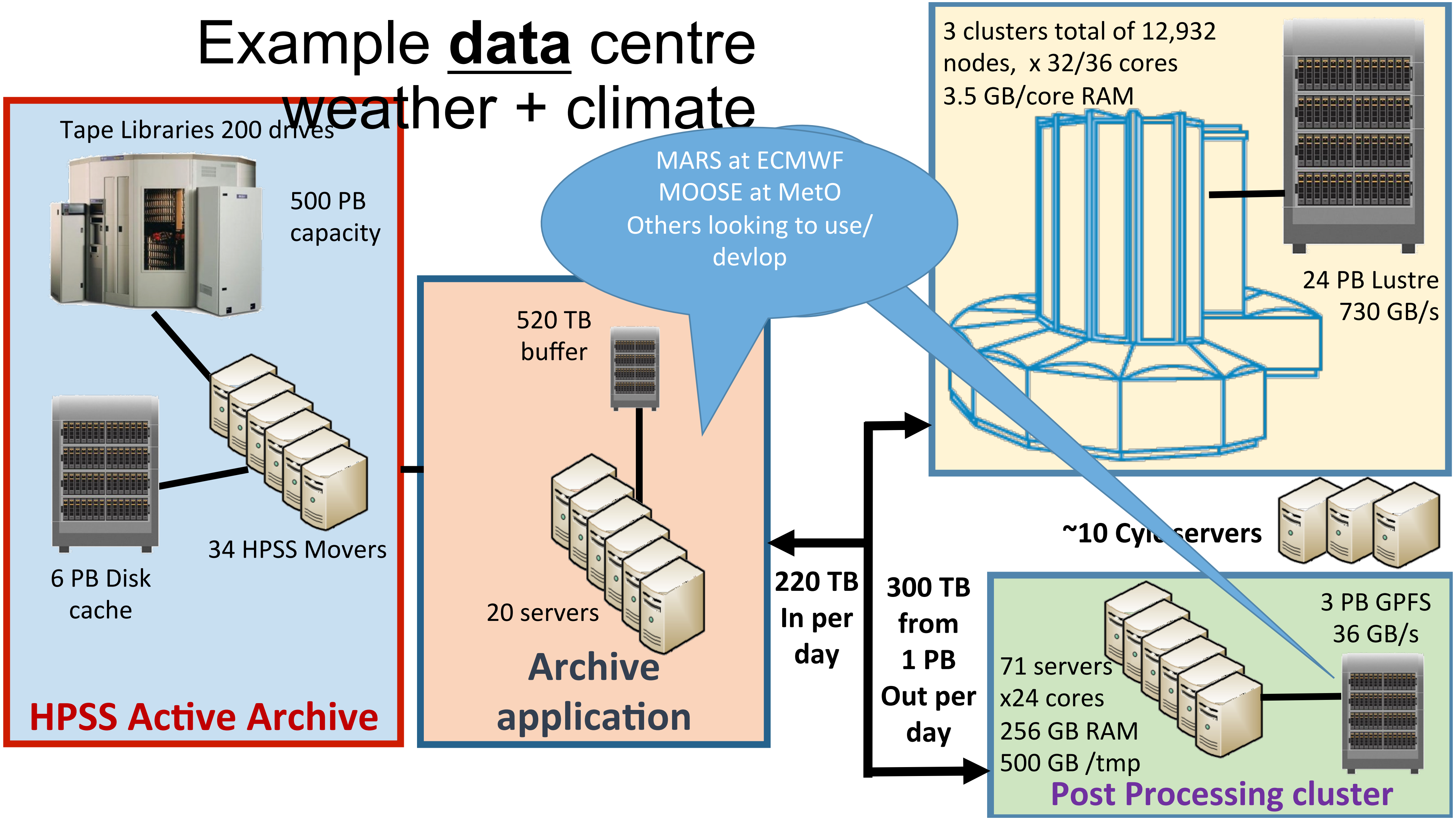
The right HPC Environment

- High memory bandwidth
- IO performance
- Interconnect latency and performance
- High memory per core
- Rich environment
  - High performance, well maintained Fortran, C, C++
  - Ability to run services for suite management (Cylc)
    - Cylc client installed. Somewhere to run persistent user service for month which can submit batch jobs and talk to the cylc server.
- Moderate resources for long periods, not everything for short periods

Other environment issues

- Simple access control & Security
- Support: Optimisation and problem solving
- Stable environments for duration of the project
    - Bit-comparable results or revalidation is required
- Well resources compilation service
- Queues
    - Development: rapid turn around, large resources
    - Production: Long job duration minimise checkpoints. Close to 24x7 access
- Good network connectivity
    - Input data requirements
    - Results measure 100s TBytes

Example **data** centre weather + climate



The right type of call:

- Scientific excellence for something like CMIP6 is hard to argue
- Benefits come from a broader context than the bid
  - Reviews need to take a wider strategy into account – eg CMIP6
  - A bid might be one part of a bigger picture such as IPCC Assessment report
  - Benefits from wide exploitation of the wider dataset by a wider audience
- Have been impossible to coordinate with H2020 projects
  - HPC without funded science projects does not work
  - Funded science projects without the HPC does not work
  - Not possible to match the two up

Summary

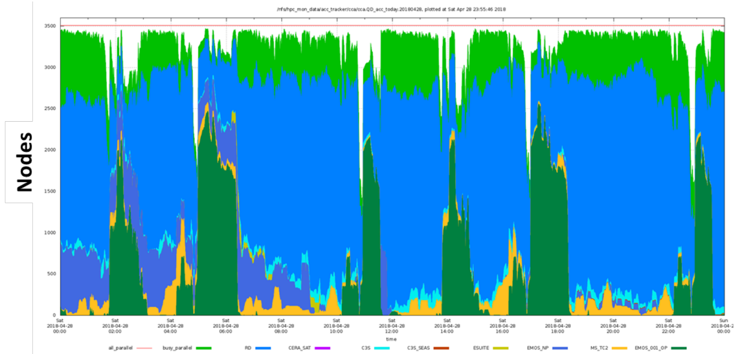
- Our communities can:
  - Provide datasets from models that feed wider communities
  - Organise to deliver results with wide societal benefit
  - Exploit significantly more HPC than we have control over
- PRACE has the capacity to make significant difference
- But we:
  - Need access over longer timeframes with stable access to a platform
    - Prepare, run and recover data
  - Need a different bidding process
    - Suitable for context of wider programmes
    - Recognising benefits of delivery outside the project
  - Need to coordinate with funding for people

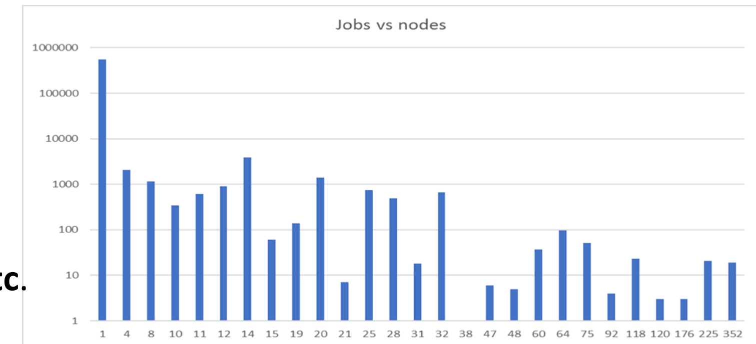
General requirements for home based HPC facility today*

Input from M Hawkins (ECMWF):

- **HPCF needs to cover the entire requirements for HPC and what is needed to make it work**:
  - Operations (24/7) + Research + ECMWF Member States + Copernicus

- ***Facility* means**:
  - Separate self-sufficient & self-contained systems for resilience and maintainability (separate compute, multiple (cross-mounted) file systems)
  - Enough performance to :
    - produce time-critical forecasts
    - trial ambitious research experiments
  - Compute nodes
  - Storage
  - Service (management, network connections, scheduling)
  - Pre/post-processing nodes
  - Login/interactive nodes
  - Power and cooling connections to facility





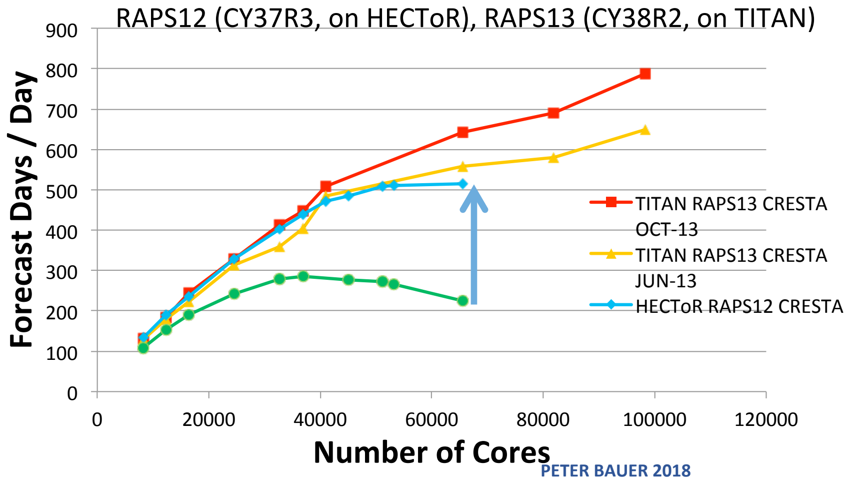
→ **HPCF is <u>not</u> a computing research system ≠ Scalability work like ESCAPE etc.**
→ **Scaling this up to next-generation models is not enough**

*** assuming that ECMWF requirements are representative for <u>present</u> operational weather prediction community**

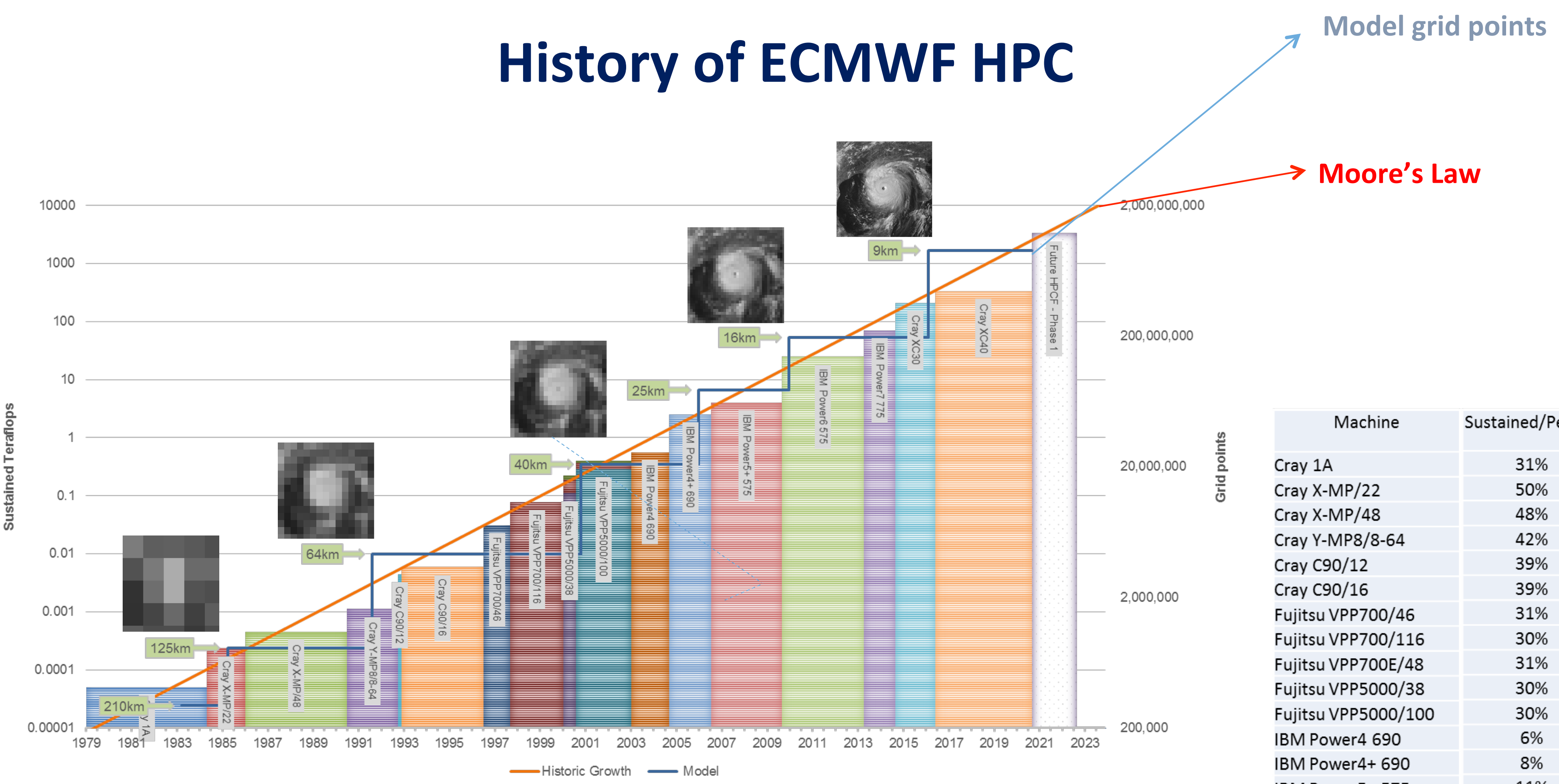
General requirements for external HPC access today*

Input from S Saarinen (ECMWF):

- Connection via ssh, preferably without port numbers; compilation resource
- Download/upload transfer speeds for input/output files of minimum 10 MB/s sustained
- Shared disk (e.g. Lustre, GPFS) space for input files 0.5 TB, with long retention periods
- POSIX compliant shared disk (e.g. Lustre, GPFS), space for runtime files 10TB ballpark
- Batch queuing system PBS or SLURM preferred, fast turn around
- Cray or Intel compilers (with tuned LAPACK & FFTW libraries), and GNU compiler on AMD
- x86_64 compliancy helps (also with AMD but not IBM and ARM)
- Robust & performant interconnect (Mellanox, OPA, Aries)



History of ECMWF HPC



Machine	Sustained/Peak %
Cray 1A	31%
Cray X-MP/22	50%
Cray X-MP/48	48%
Cray Y-MP8/8-64	42%
Cray C90/12	39%
Cray C90/16	39%
Fujitsu VPP700/46	31%
Fujitsu VPP700/116	30%
Fujitsu VPP700E/48	31%
Fujitsu VPP5000/38	30%
Fujitsu VPP5000/100	30%
IBM Power4 690	6%
IBM Power4+ 690	8%
IBM Power5+ 575	11%
IBM Power6 575	8%
IBM Power7 775	5%
Cray XC30	6%
Cray XC40	4%

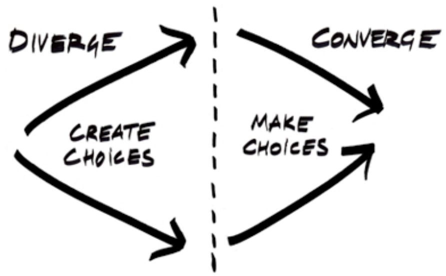
Divergence no.1: Sustained – peak performance
Divergence no.2: Earth-system model degrees of freedom – Moore's law

Dilemmas

1. How do we develop advanced models, prediction systems, workflows with HPC infrastructures lagging at least 5 years behind?
    - We have large-scale infrastructures with 'recent' technology (both software and hardware), but need to develop future systems currently full of gaps in software stack (eg. programming), technology (eg. memory hierarchy)

2. How do we manage the transition of advanced components into operational work streams?
    - We need to <u>incrementally advance</u> operational systems and <u>revolutionize</u> at the same time

3. How do we procure new facilities?
    - Procurements are supporting operations and incremental progress, but not radically new applications



4. How do we manage knowledge access across European community and beyond?
    - One <u>system</u> for all vs a co-developed <u>core set of tools</u> for all

ESM ACTIVITIES IN JÜLICH

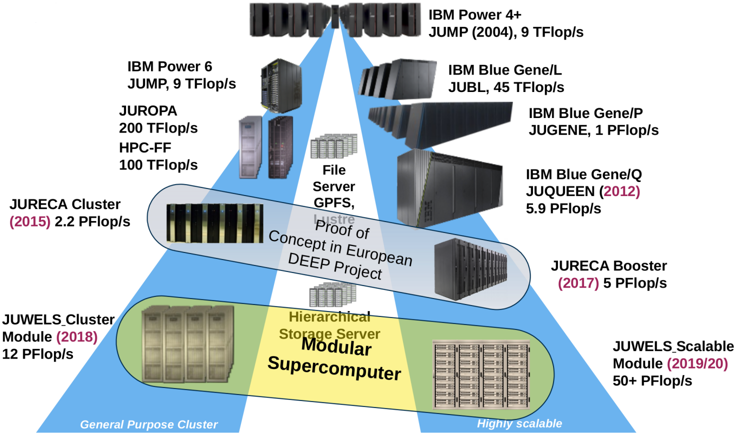
HPC facility and support

May 14, 2018 | Lars Hoffmann | Jülich Supercomputing Centre (JSC)

JÜLICH
Forschungszentrum

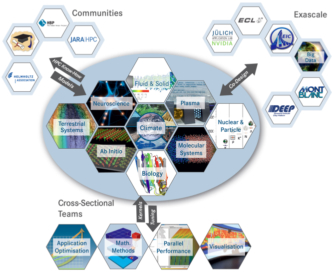
JÜLICH HPC FACILITY

- Evolution towards modular computing...



- JUWELS gets a dedicated ESM partition...

SUPPORT FOR ESM COMMUNITY



- Simulation Labs ‘Climate Science’ and ‘Terrestrial Systems’
  - interface between users of HPC facility and local IT experts
  - porting of community codes (e. g., ICON, WRF) to facility
  - integration of new tools and technologies (GPU, KNL)
  - enable “frontier simulations”
- Research group ‘Earth System Data Exploration’
  - data exploration by machine learning techniques
  - hosting of data services
- Contributions to infrastructures and projects:
  - EU: DEEP, EoCoE, EUDAT, POP, PRACE
  - Germany: HD(CP)^2, Helmholtz ESM project, HDF

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

The Barcelona Supercomputing Center

Kim Serradell
Computational Earth Sciences

19/05/2018

5th ENES HPC Workshop

Mare Nostrum 4

Compute

General Purpose, for current BSC workload

More than 11 Pflops/s

3,456 nodes of Intel Xeon v5 processors

Emerging Technologies, for evaluation of 2020 Exascale systems

3 systems, each of more than 0,5 Pflops/s

with **KNL/KNH, Power9+NVIDIA, ARMv8**

Storage

14 PB of GPFS

Elastics Storage System

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Evaluation of 2020 Exascale systems

- Four different architectures
- Sharing HPC disk
- Easier to deploy the same code
- Versatile workflow manager tool needed

Data Centre Support for Weather and Climate Models and Workflows

S. Requena (GENCI) and X. Delaruelle (CEA/TGCC)

GENCI' FEEDBACK WRT CLIMATE COMMUNITY

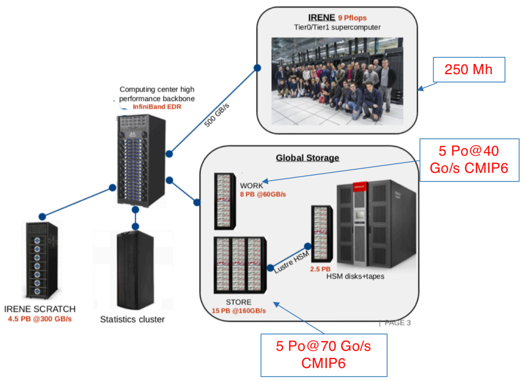
What we learnt

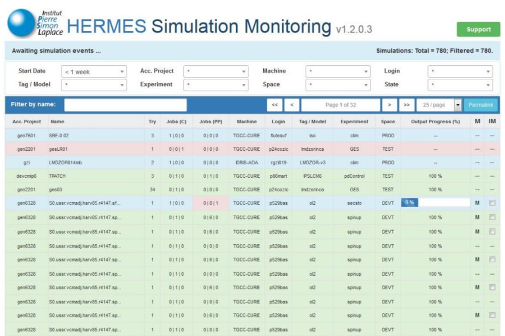
❑ GENCI and Climate / NWP at a glance

- GENCI : 4 supercomputers on 3 national centers (CINES, IDRIS, TGCC) → 14 PF
- MéteoFrance has a dedicated HPC center but others NWP/Climate → Genci
  - NWP/Climate = 13% of projects, 9% hours allocated (250Mh/yr), **#1 storage**

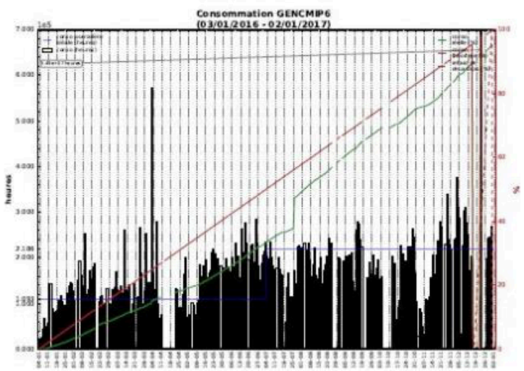
❑ Focus on the CMIP6 **production** exercise (2016-2018)

- Strong collaboration between TGCC, IDRIS, IPSL, GENCI and Renater
  - Dedicated CPU quota (300Mh) and storage (14 PB TGCC, 4 PB IDRIS for ESGF)
- A lot of preliminary and ongoing meetings for knowing/working each other !
  - Requirements in terms of storage capacity, #inodes, type of files, …
  - Data localisation across different filesystems, new Lustre R&D (DNE, OST pools), …
  - Fine monitoring of the simulation, users/job management, priorities, accounting, ….
  - Key of success = to have a dedicated interface centres <-> IPSL
- Balanced & stable HW configuration and 3-year allocation



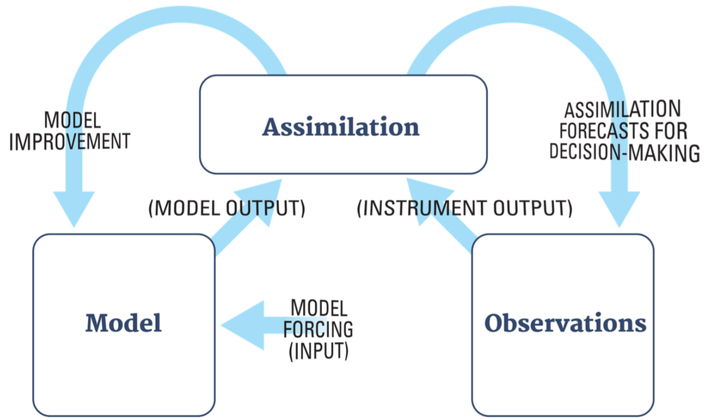


5TH



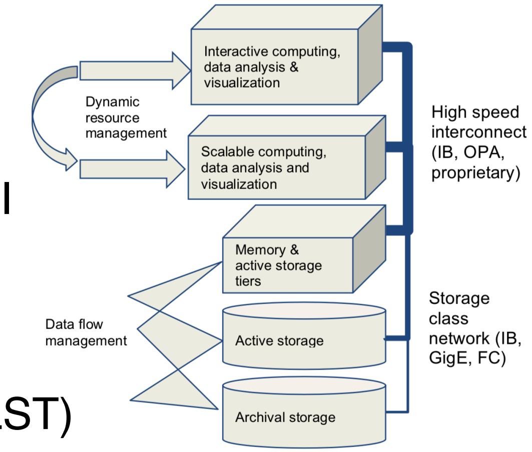
TOWARD A FEDERATED HPC AND DATA RESEARCH INFRASTRUCTURE

Some common challenges



1. Address the HPC/HPDA/AI convergence
   - Support end to end workflows « from the edge to the tape »
   - Deploy/maintain containers and ensure security
   - Go beyond batch : stream/interactive, elastic access modes
   - Urgent computing : decision making during extreme events

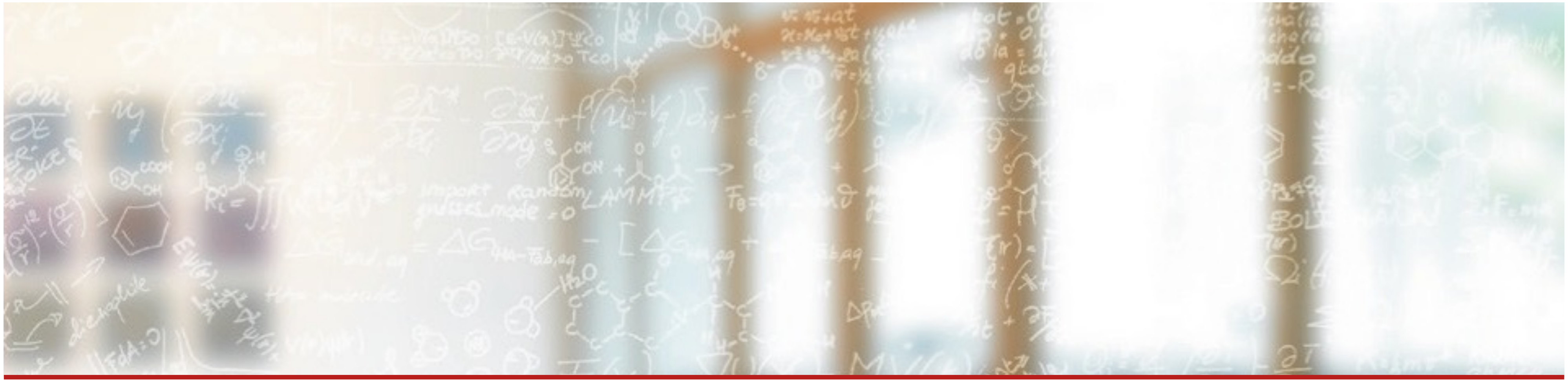
2. Optimise/minimise data movement
   - Power consumption as a major issue for Exascale
   - Cache, prefecth, co locate, compress, … share results
   - Develop in-situ post processing/compression (XIOS) supported by AI



3. Prepare new steps : Exascale Science and CMIP7
   - Foster collaboration between CoE and HPC centers (ex: PRACE HLST)
   - Use Climate/NWP apps, mini apps, kernels for benchmark
   - Allow access to early prototypes for co design
   - Promote EU standard tools : scalable couplers (OASIS-MCT…), pre/post processing (XIOS, …), workflows, UQ frameworks, DSLs, …
   - Training to new prog languages, data analytics, AI, …

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH zürich



CSCS current/future support of weather and climate models and workflow

Will Sawyer, Thomas Schulthess
Swiss National Supercomputing Centre (CSCS)

5th ENES HPC Meeting,
Panel Discussion: “Data centre support for weather/climate models and workflow”

May 18, 2018, Lecce, Italy

CSCS current climate/NWP support

- *"Kesch/Escha" (CS-Storm w/ fat GPU-nodes) for MCH 1km forecast+2.2km ensembles*
  - *12x (2x Haswell / 8x K80), 3x login, 5x post-proc nodes + full backup (Carlos)*
  - *Workflow with extensive shell scripting: delicate and failure prone*
  - *Complex: multiple executables running on single node; co-design with Cray*
- *"Daint" (Cray XC50/40, w/ 5320x P100 GPUs): crCLIM project (Carlos)*
  - *high-res Euro-domain climate runs (on GPU)*
  - *Climate Science: how does increased resolution affect "forecast quality"?*
  - *Comp. Science: reproducible restart capability (= less data storage)*
- *Partnership for Advanced Scientific Computing (PASC) Initiative*
  - *GridTools ecosystem: separation of science and implementation: C++ DSL for atmospheric dynamics (physics) components*
  - *PASCHA: transition of COSMO to GridTools, add Xeon Phi backend*
  - *ENIAC: GPU-port of ICON with CLAW source-to-source translator (Valentin)*

CSCS future support

- *Containerization with Docker/Shifter*
  - *Fundamental limitations in storage scalability; CSCS data storage will grow slowly*
  - *Requirement: reproducibility of model runs without archiving data (2-5 years)*
  - *ESiWACE-2 proposal (Joachim): containerize 4 models*
  - *Compiler versions introduce instability: programming environment must be included*
  - *But: IP issues in putting PEs into containers*
- *Workflow*
  - *CTO office looking into technologies: Common Workflow Language, Eclipse, (others?)*
  - *Python is good bet for scripting workflow (e.g. Thomas)*
  - *PhD: Python atm. model quick prototyping framework for dycore + single column physics*
  - *Extensions to GridTools: GT4Py generate kernels utilizing GridTools 'backend'*
- *Co-design of new platforms for climate/NWP*
  - *Climate/NWP strategic for system design because it leads to systems that are more usable by other domains as well (low arithmetic intensity)*
  - *HPL-optimized platforms non-optimal for almost anything. We don't care about exaflop/s scale and we really mean it while others don't; talk about goals, not exascale.*
  - *partnership with industry and scientific community*

Potential discussion topics: red