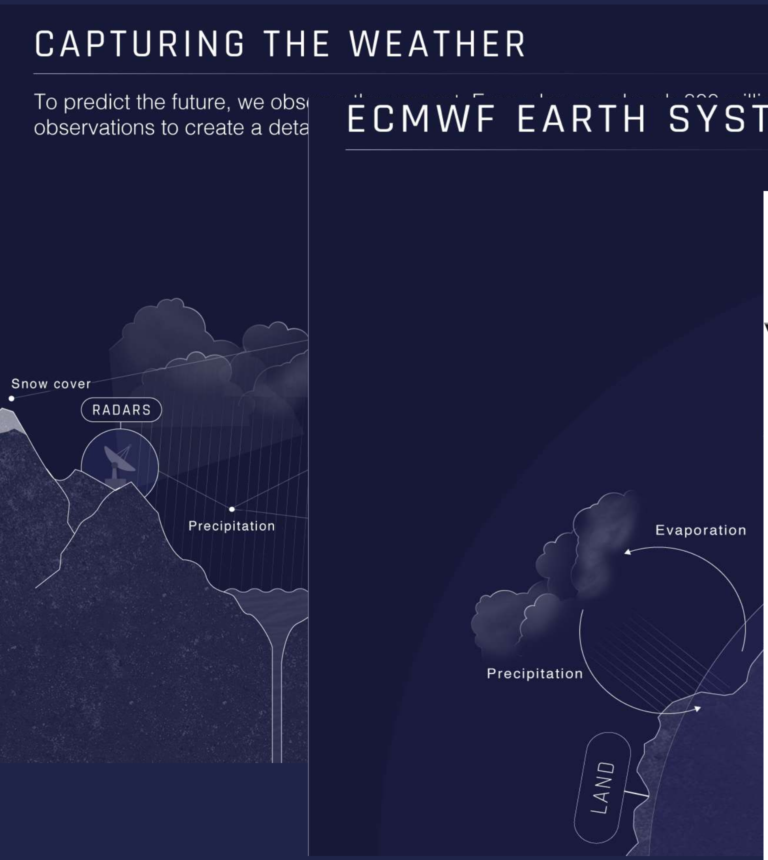
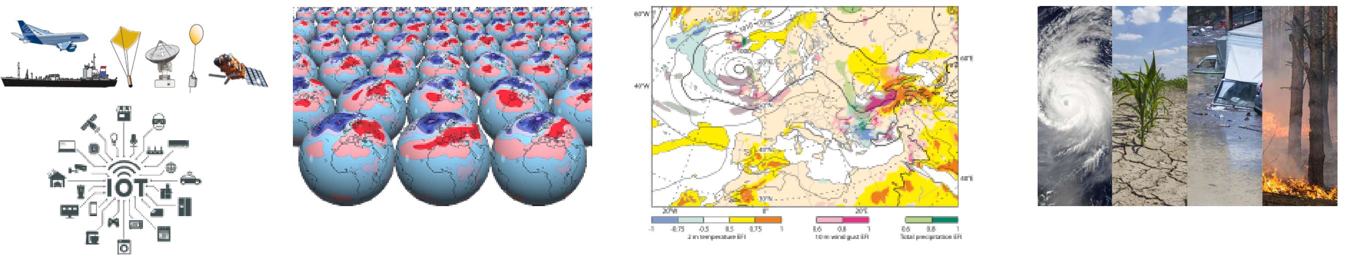
# ECMWF
# Computational challenges in a data-rich environment

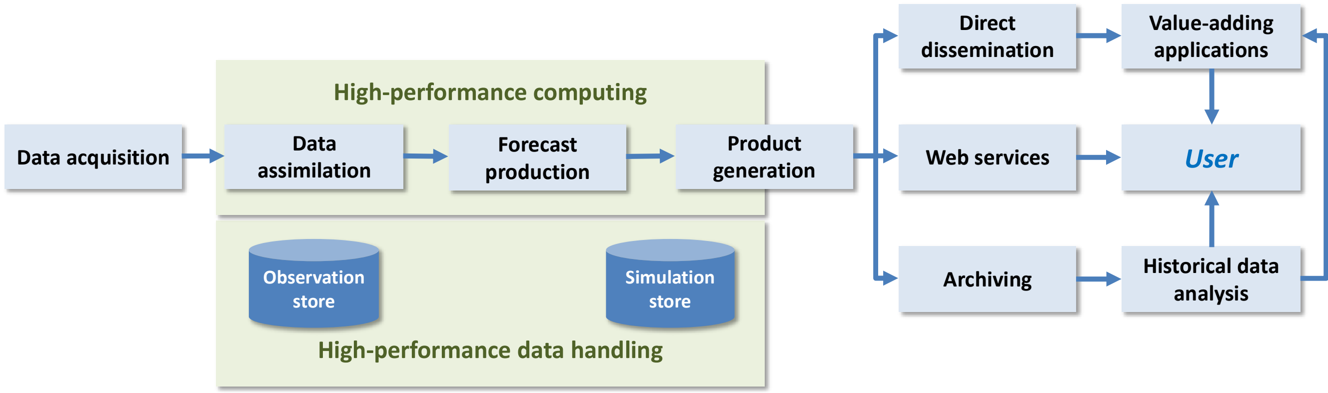
Thomas Geenen
Technology partnership lead for Destination Earth
thomas.geenen@ecmwf.int

# ECMWF operational models: Simulations in a data-rich environment



**Observations** + **Simulation models** = **Information based...** **... Decision making**





Data acquisition

High-performance computing

Data assimilation

Forecast production

Product generation

Observation store

Simulation store

High-performance data handling

Direct dissemination

Value-adding applications

Web services

*User*

Archiving

Historical data analysis

# High resolution simulation is essential but why?

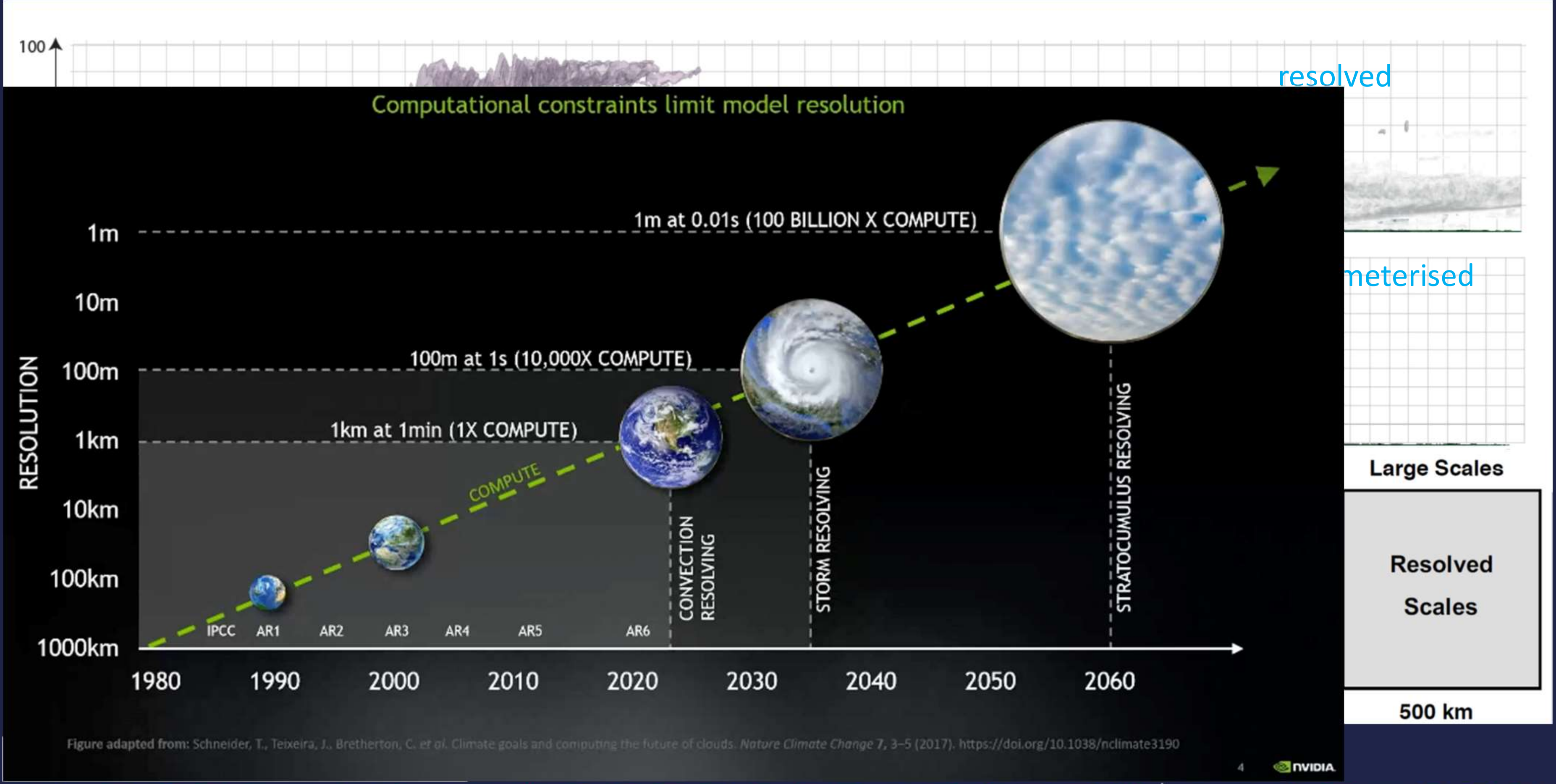
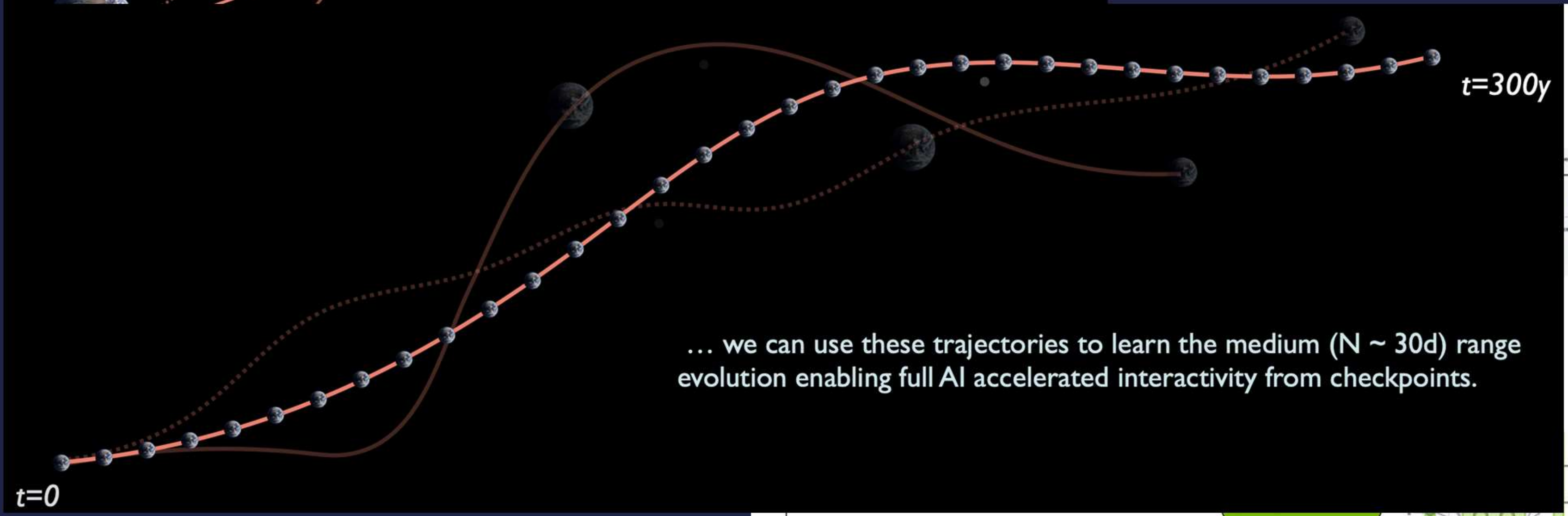
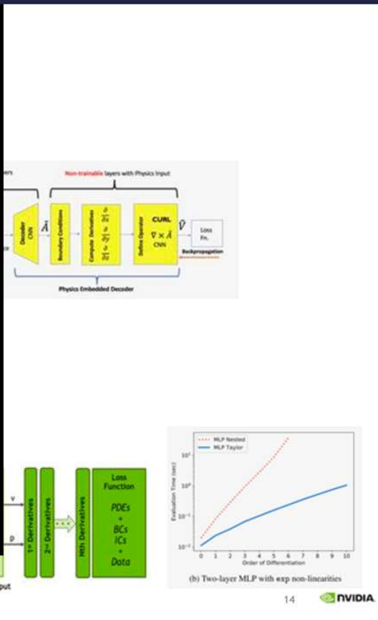


Figure adapted from: Schneider, T., Teixeira, J., Bretherton, C. *et al.* Climate goals and computing the future of clouds. *Nature Climate Change* 7, 3–5 (2017). https://doi.org/10.1038/nclimate3190

# Can we cut any corners?



… we can use these trajectories to learn the medium (N ~ 30d) range evolution enabling full AI accelerated interactivity from checkpoints.

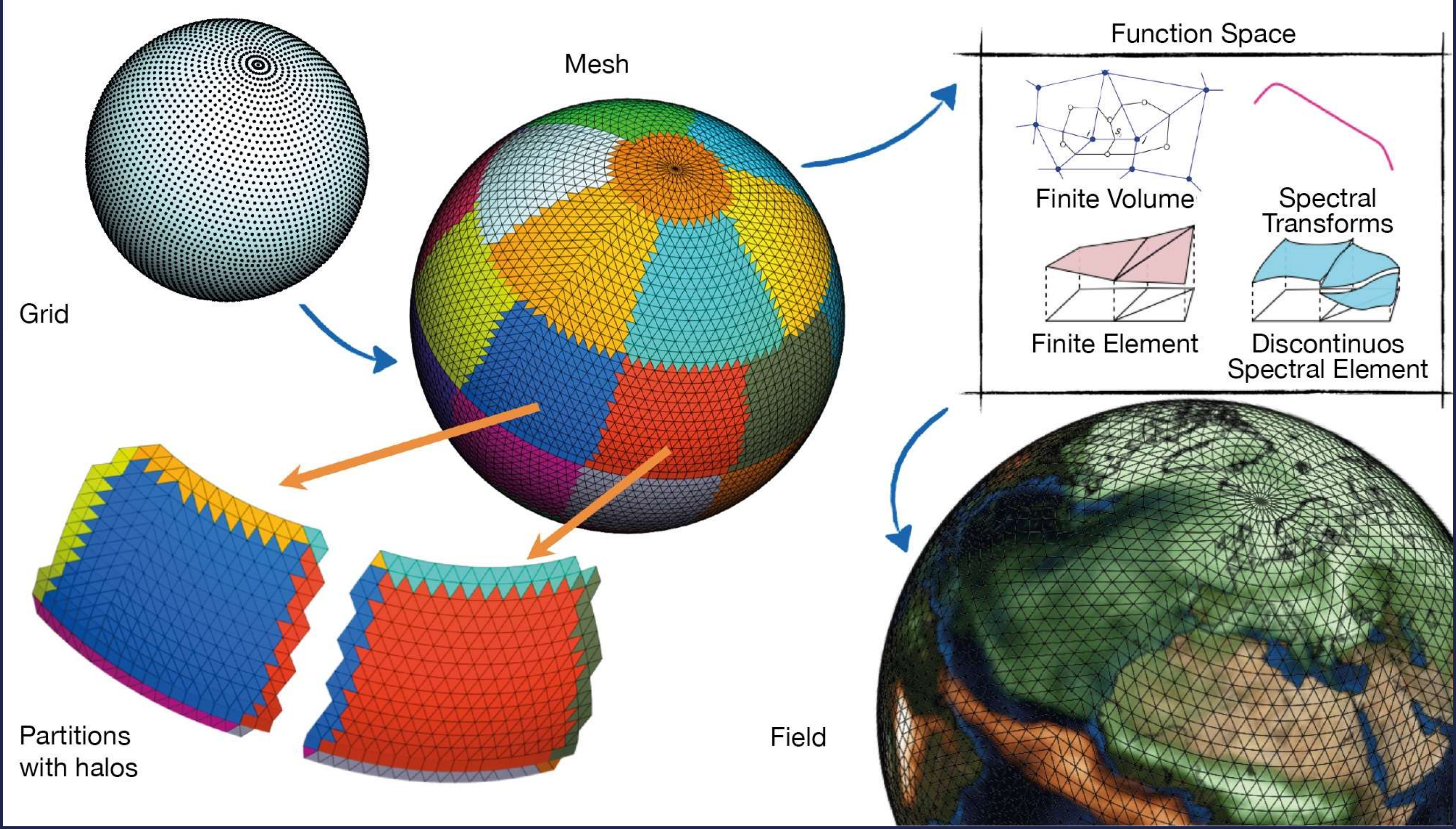


(b) Two-layer MLP with exp non-linearities

Physics

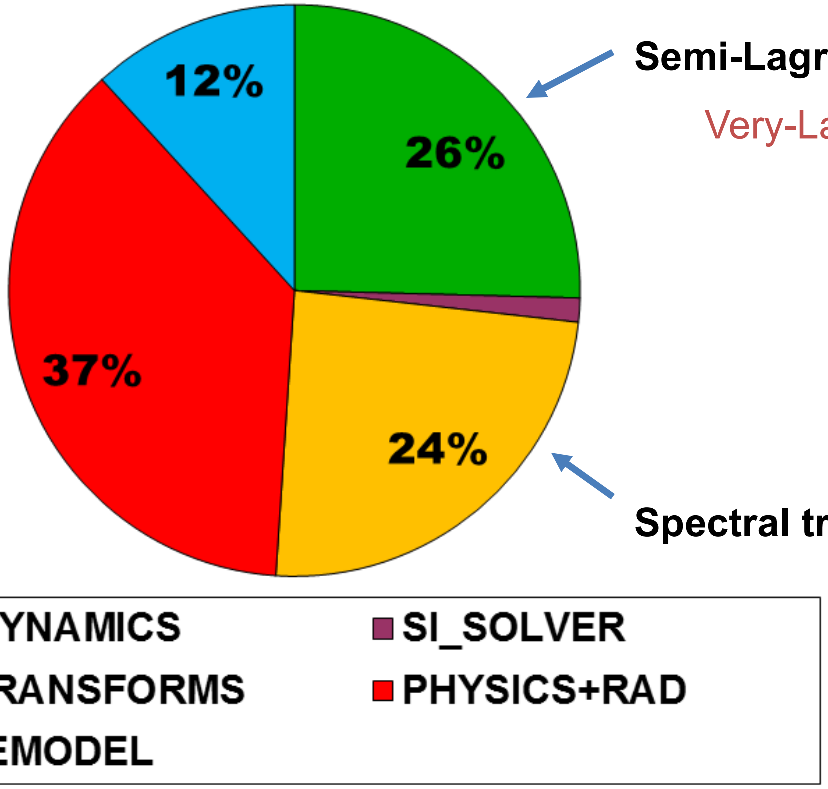
Courtesy Bjorn Stevens

# So What do our methods look like



# Where are the compute cycles burned

## Operational model configuration (9 km)



**Semi-Lagrangian advection**

Very-Large-Halo Exchange (MPI)

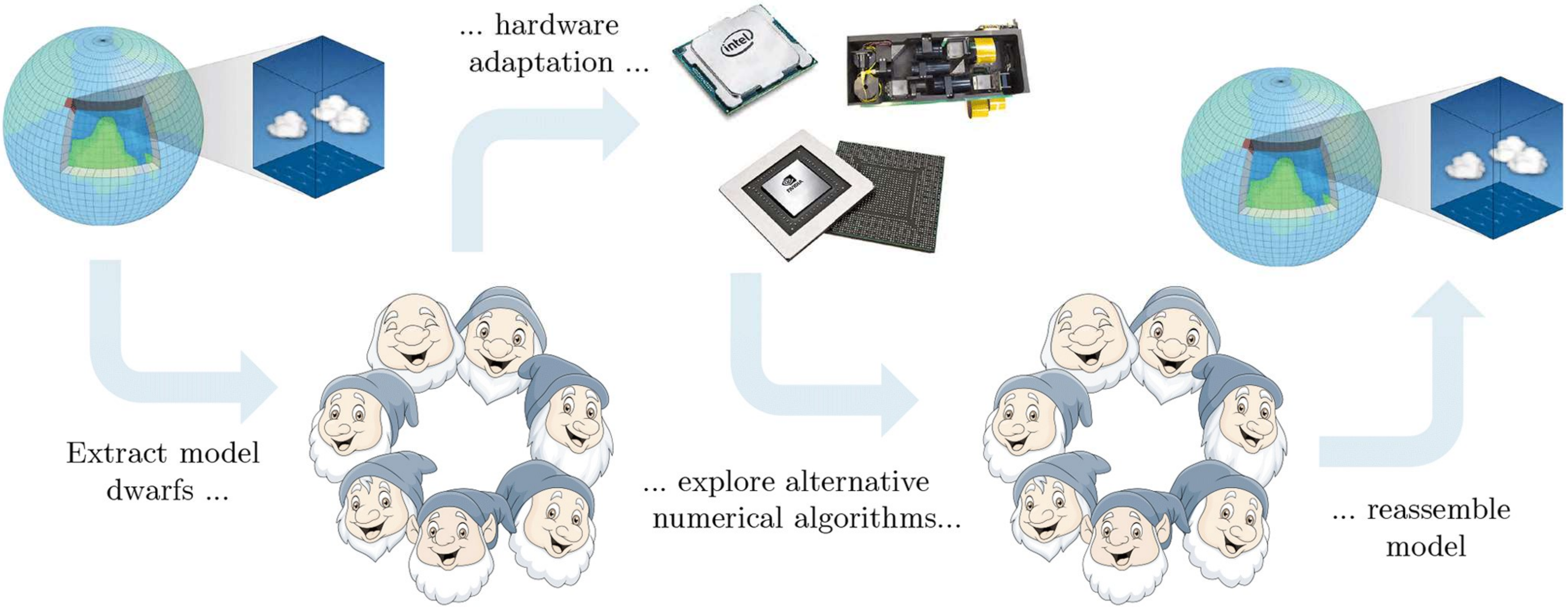
**Cloud physics parameterization**
**Radiation parameterization**

Complex branching

**Spectral transforms (FT/LT and bi-FT)**

Global transpositions
( All-to-All MPI )

# ECMWF approach to optimization



Extract model dwarfs ...

... hardware adaptation ...

... explore alternative numerical algorithms...

... reassemble model

Earth illustration: used under license from GraphicsRF/Shutterstock.com.
Dwarf illustrations: used under license from Teguh Mujiono/Shutterstock.com

# Separation of concerns

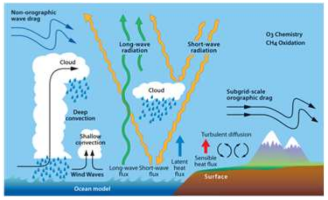
Domain scientist:

- Controls grid, resolution, …
- Maintains single source code!
- No hardware specifics!
- No parallelisation specifics!
- No memory layout concerns

DSL Toolchain

- Provides performance portability across a variety of hardware
- Provides parallelisation
- Memory layout
- **Introspection**

Domain science


Physics

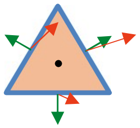
$$\rho\dot{\boldsymbol{u}} = -\nabla p + \rho g - 2\Omega \times (\rho\boldsymbol{v}) + \boldsymbol{f}$$

$$\dot{p} = -\left(\frac{c_{pd}}{c_{vd}}\right) p \nabla \cdot \boldsymbol{u} + \left(\frac{c_{pd}}{c_{vd}} - 1\right) Q_h$$

$$\rho c_{pd}\,\dot{T} = \dot{p} + Q_h$$

Mathematical description

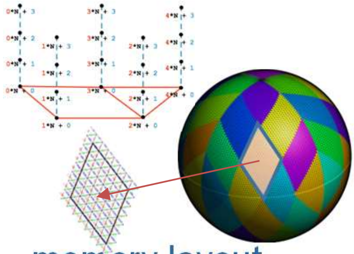
$$\nabla \cdot \mathbf{v} := \frac{1}{A} \sum_{k \in \mathcal{E}} \mathbf{v}_k \cdot \mathbf{l}_k$$


Algorithm development

```
on_edges( sum_reduction, v() * l() ) / A()
```

Domain specific language

Multidisciplinary Abstractions


memory layout, parallelisation, & data structures

OpenACC
Directives for Accelerators

OpenMP

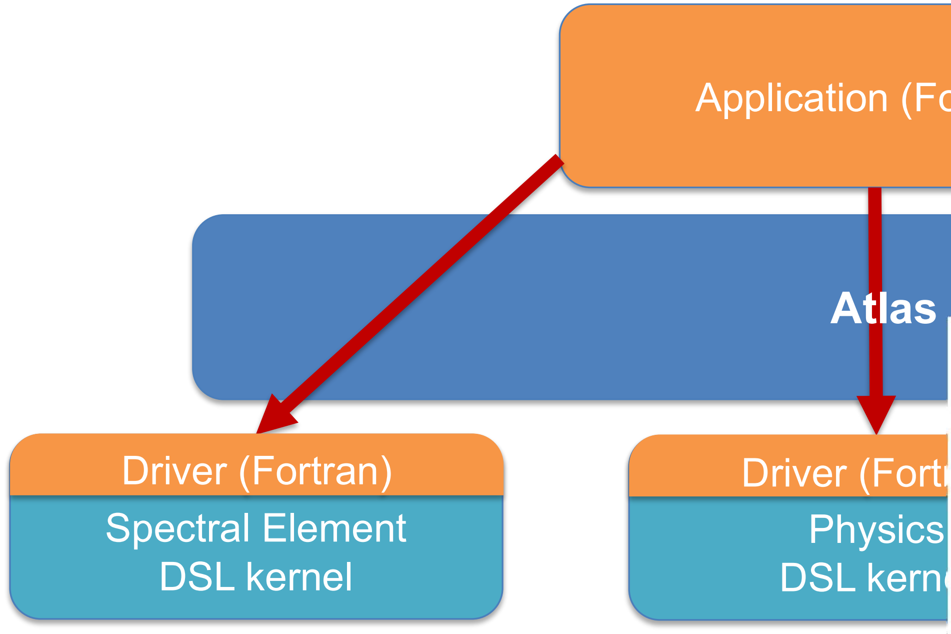
NVIDIA CUDA

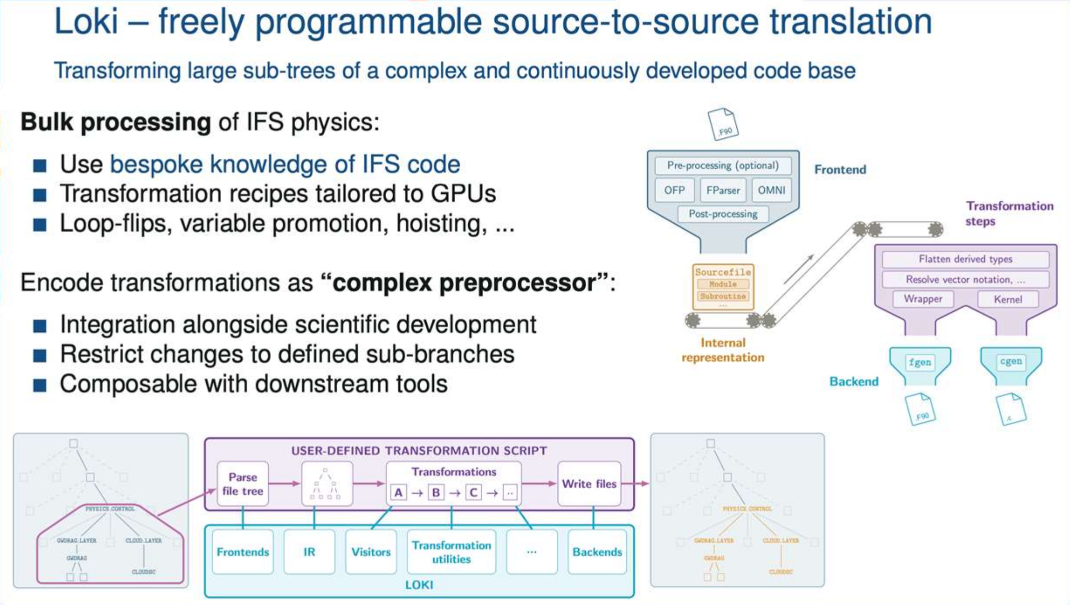
MPI

Programming models & libraries


Hardware specific instructions

# ECMWF approach to optimization





## Loki – freely programmable source-to-source translation

Transforming large sub-trees of a complex and continuously developed code base

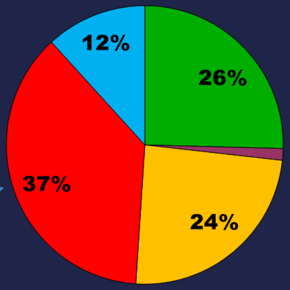
**Bulk processing** of IFS physics:

- Use bespoke knowledge of IFS code
- Transformation recipes tailored to GPUs
- Loop-flips, variable promotion, hoisting, ...

Encode transformations as “**complex preprocessor**”:

- Integration alongside scientific development
- Restrict changes to defined sub-branches
- Composable with downstream tools

# Dealing with the complex physics



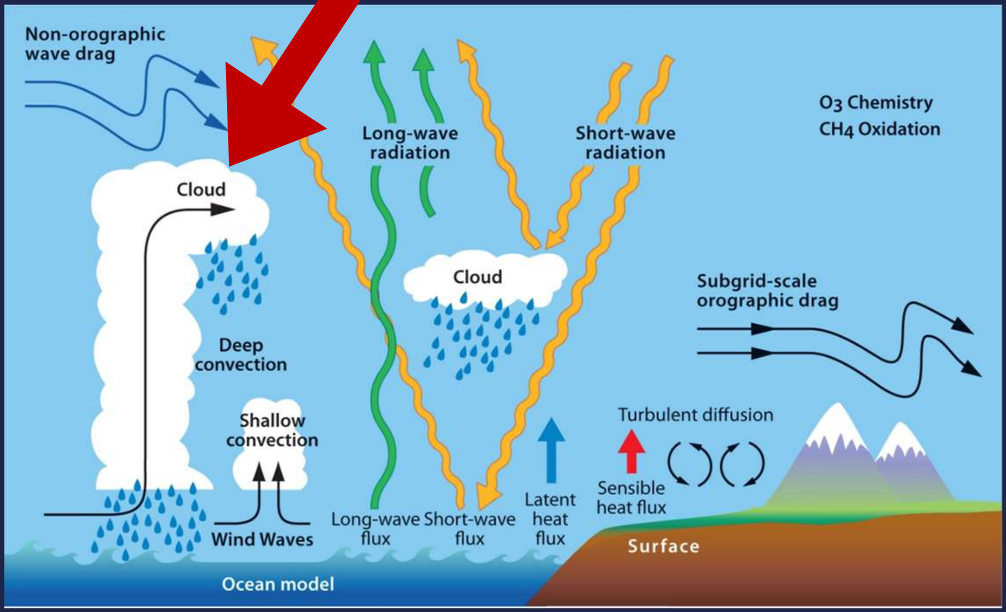
Cloud physics parameterization
Radiation parameterization

Use FPGA to explore arbitrary hardware platforms

**Done for a particularly hard physics routine**

- Cloud physics (representative for other physics)
- High computational intensity per grid point: >3

- Known limitations from CPU and GPU
  - Register pressure
  - Cache misses (complex branching)



Significant speedups on FPGA (2-3x)
Means potential for speedup on new hardware

# In summary

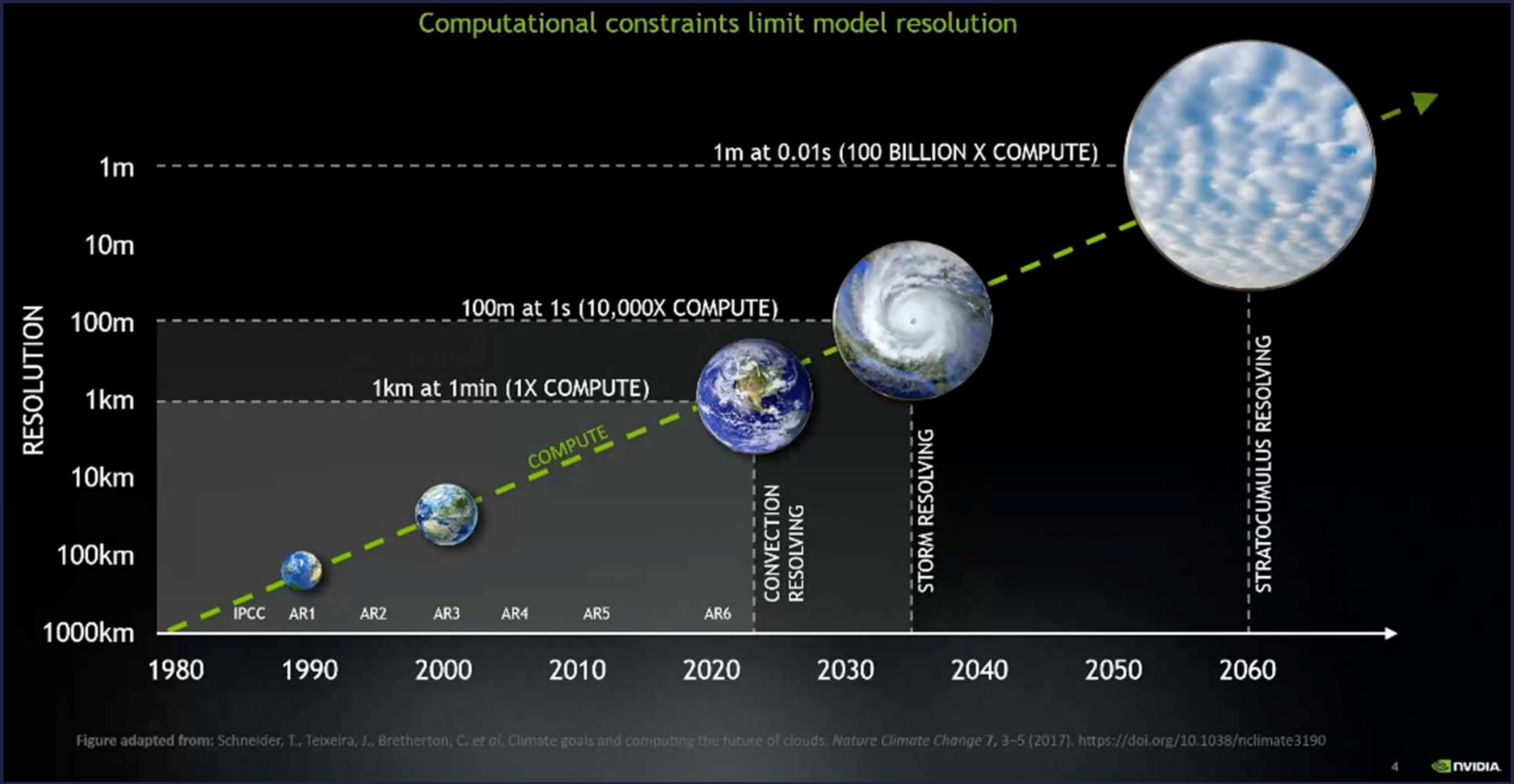
ECMWF has explored many architectures and concepts
We are looking into domain specific languages (DSL) to exploit platform specific optimization

- Our applications are memory bandwidth bound (data movement on all levels)
- We suffer from register pressure (both on GPU and CPU , deep variable stacks)
- We suffer from cache misses (complex branching)

- We are looking at stencil operations for FVM
- We can also use stencil operations for postprocessing and model coupling
- We try to compute as much as possible when we have the data on a grid-point

- We converted from double precision to single and we are exploring lower and mixed precision

# Numerical weather prediction is a zetascale problem



Computational constraints limit model resolution

**Figure adapted from:** Schneider, T., Teixeira, J., Bretherton, C. *et al.* Climate goals and computing the future of clouds. *Nature Climate Change* **7**, 3–5 (2017). https://doi.org/10.1038/nclimate3190

# Thank You


ECMWF | EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
ECMWF

# ECMWF
# Predicting trends to support weather and climate forecast

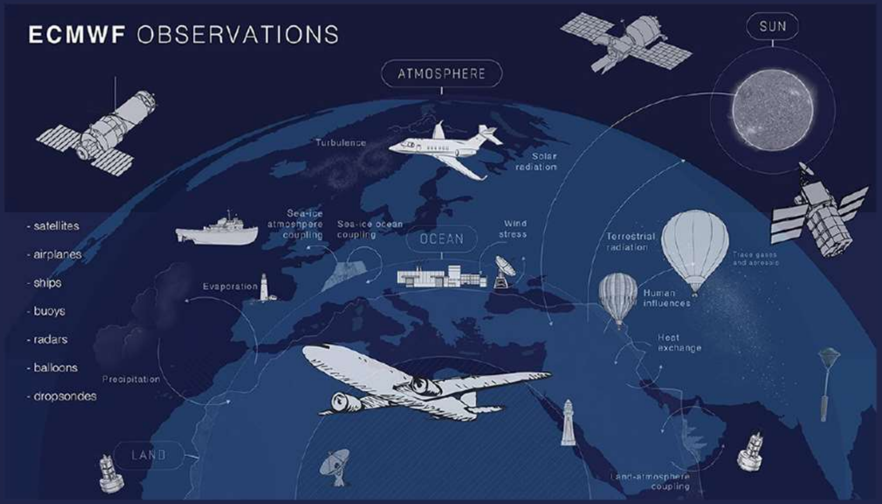
Christine Kitchen
Deputy Director of Computing Department, ECMWF
Christine.Kitchen@ecmwf.int

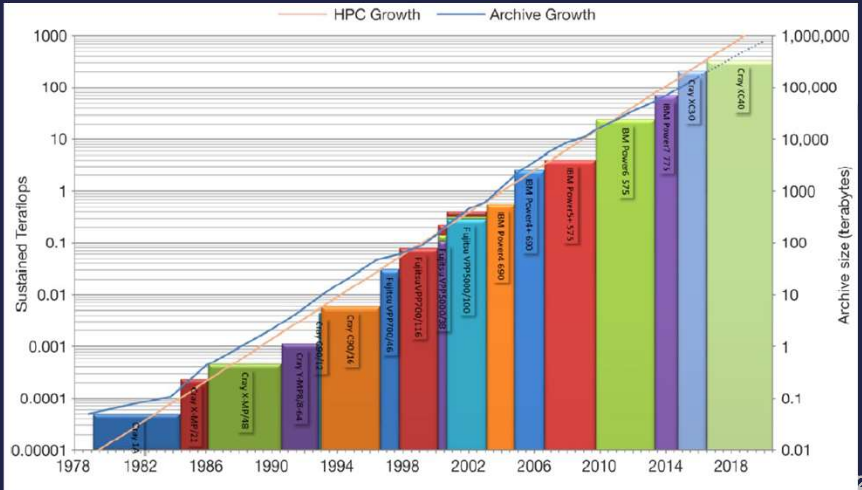
# Current Service – HPCF2020 and DHS

- Our Current high-resolution forecast system has near two billion degrees of freedom
- NWP models are massively parallel – challenge is data heavy and processing hundreds of millions of observations and generating tens of terabytes of model output every day

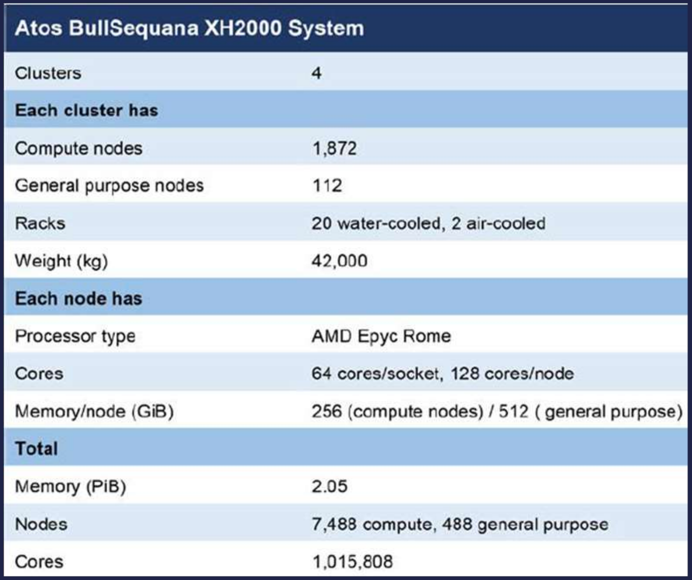
Supercomputing Environment is constantly evolving – driven in part by the 'demise' of Moore's Law

- Increasingly heterogeneous hardware – CPUs → GPUs → FPGAs → ASICs
- Massive Parallelisation
- Power Consumption
- Strong Scaling limit

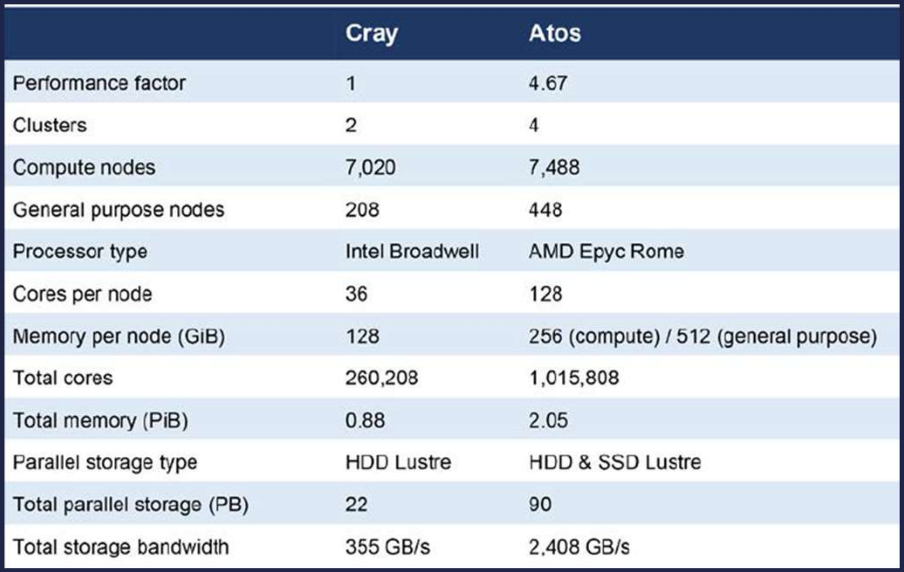




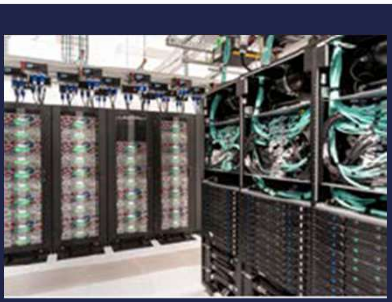
# Current Service – HPCF2020



| Atos BullSequana XH2000 System | |
|---|---|
| Clusters | 4 |
| **Each cluster has** | |
| Compute nodes | 1,872 |
| General purpose nodes | 112 |
| Racks | 20 water-cooled, 2 air-cooled |
| Weight (kg) | 42,000 |
| **Each node has** | |
| Processor type | AMD Epyc Rome |
| Cores | 64 cores/socket, 128 cores/node |
| Memory/node (GiB) | 256 (compute nodes) / 512 ( general purpose) |
| **Total** | |
| Memory (PiB) | 2.05 |
| Nodes | 7,488 compute, 488 general purpose |
| Cores | 1,015,808 |

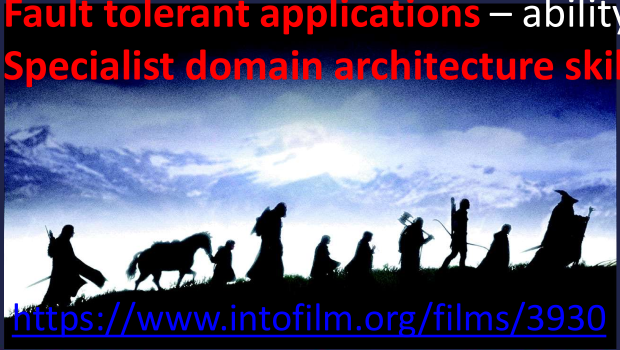


| | Cray | Atos |
|---|---|---|
| Performance factor | 1 | 4.67 |
| Clusters | 2 | 4 |
| Compute nodes | 7,020 | 7,488 |
| General purpose nodes | 208 | 448 |
| Processor type | Intel Broadwell | AMD Epyc Rome |
| Cores per node | 36 | 128 |
| Memory per node (GiB) | 128 | 256 (compute) / 512 (general purpose) |
| Total cores | 260,208 | 1,015,808 |
| Total memory (PiB) | 0.88 | 2.05 |
| Parallel storage type | HDD Lustre | HDD & SSD Lustre |
| Total parallel storage (PB) | 22 | 90 |
| Total storage bandwidth | 355 GB/s | 2,408 GB/s |



# The less “glamorous side” of service provision

- Benefits of production and research environments in close proximity
  - Access to vast corpus of observational data / rapid exchange of code and replicated environments
- Balance of enterprise production quality SLA driven (time critical) vs Reactive & Responsive development environment
  - Containerisation providing greater degrees of flexibility rather than reconfiguring software stacks?
- **Complexity of Environment** due to increasingly more heterogenous solutions
  - System Engineers (administrators/analysts/operators) – **generalist vs specialist skills**
  - Middleware – **rich environment of tools** to support software development and revision control
  - **Maturity of cluster management** tools – ability to identify faults in systems comprising millions of components (filtering noise) / predictive or pre-emptive fault monitoring / diagnosis information – Integration of supplier packages into a coherent interface
  - **Fault tolerant applications** – ability to seamlessly migrate?
  - **Specialist domain architecture skills** – niche sector reliant on a small number of skilled individuals?



https://www.intofilm.org/films/3930

© NEW LINE CINEMA ALL RIGHTS RESERVED

It’s not just the dwarves that need to work together…

# Not just the compute flavour – the entire recipe ingredients are required

- Important to consider in the context of the entire ecosystem – all components need to fully integrate (operating system / cluster management /software development tools & compilers / storage / datacentres)
- Appropriate lead time to establish the skills across the community to support these developments
    - IDENTIFICATION OF NEXT GENERATION OF SKILLS across the landscape
        - Silicon (processor) designers
        - programmers
        - middleware application designers
        - system administrators
    - Global skills shortage - consider in parallel to help address this
    - Ensuring the skills are in place attract & build communities to support?
- Datacentre ecosystem – power / cooling / form factors – cannot be too radical as need to integrate into sites datacentres or associated hosting centres?
- Transitioning between services – is the future a gradual phasing of new technology into core infrastructure rather than large procurement cycles?

*Image: https://www.behance.net/gallery/79780971/Recipe-for-Success*

# The Wish List?

- Diverse range of algorithms used in weather and climate that do not necessarily fit a single platform - some can be readily adapted to fully exploit next generation architectures, but we still have legacy operations to support.
- Weather predictions involve communications and data staging – memory layout and hierarchy provide opportunities to exploit? Cache misses have major performance implications on IFS
- Domain Specific - Cannot afford to be ‘too niche’ – essential **criteria is affordable and sustainable**
  - Identification of other research domain challenges that benefit from architecture performance improvements
  - Ensuring sufficient support network to effectively exploit next generation systems - essential
- Set of agreed **Standards / Frameworks** – varying examples of success to date.
  - Gap analysis to help identify missing components?
- Evaluation
  - Ability to test at scale? Prototypes and early silicon not viable?
  - Potential to **co-develop a meaningful benchmarking** test suite that is representative of various domains?
  - ECMWF interested in building weather algorithms to evaluate new generations of technology to help benchmark early prototypes through a suite of applications representative of the future research requirements – create a performance suite particularly testing memory architecture on processors
- Responsive support network – deferring fixes to future releases impacts on developer and release to market
- **Technology Readiness Levels** – opportunities to engage at design stages to identify core dwarves / algorithm development to exploit and/or influence system design to benefit applications with memory and high I-O demands – understand the impact of workflows and tool chains. In parallel, any features that are not in place provides developers with opportunities to understand out-of-package components to redesign software profiles .

# Summary

- Weather and Climate Community have a complex software stack and environment involving multiple applications
- In supporting Weather and Climate challenges,  considering existing processor architectures – **memory architecture** (high memory bandwidth) is a recongised bottleneck.
- Current generation file system **I/O performance** causes performance degradation and instabilities – must be able to support time-critical activities to meet stringent SLA demands
- Weather and Climate workflows are already **ported** to multiple different system architectures (using distributed networks of supercomputers) – **data formats and fast interconnects** become the challenge!
  - Distributed architectures: future investment strategies must consider nationwide terabit networks?
  - Diverse portfolio of algorithms used in weather and climate prediction – can adapt some of these to exploit new processors (reverse engineer to the platform?)
- Important caveat: cannot afford to create **a niche marketplace** through a bespoke architecture – it **MUST** be **sustainable** – both in the community and longevity to justify investment in developing applications and the lead times to achieving this.
- A holistic environment must be established to ensure adoption and effective exploitation (limited 'specialist' skills in developing and supporting these environments from a service provider perspective).
- New processors architectures: imperative these are compatible and built around **a set of standards and frameworks** to ensure integration / alignment with existing infrastructure.
- ECMWF established track record  as a prime collaborator in a number of EU technology projects -  expertise and awareness of the diverse technology challenges facing the NWP community that have significant social & economic impact.

# Thank You


ECMWF | EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS
ECMWF