ECMWF's IO Challenges and the path to Exascale Numerical Weather Prediction

Tiago Quintino, B. Raoult, S. Smart, A. Bonanni, F. Rathgeber, P. Bauer, N. Wedi

ECMWF

tiago.quintino@ecmwf.int

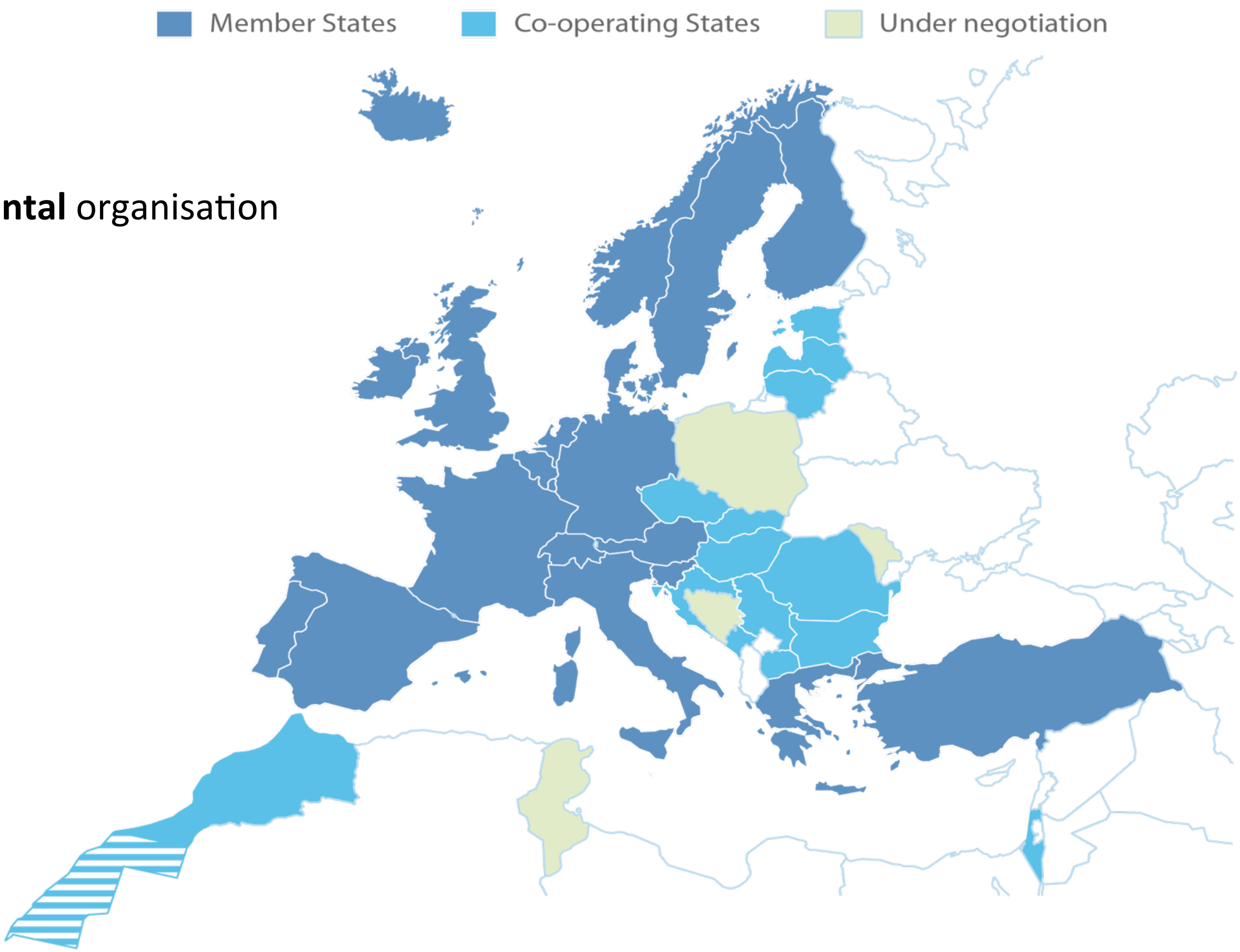
ISC’17, Workshop on Performance and Scalability of Storage Systems

© ECMWF June 22, 2017

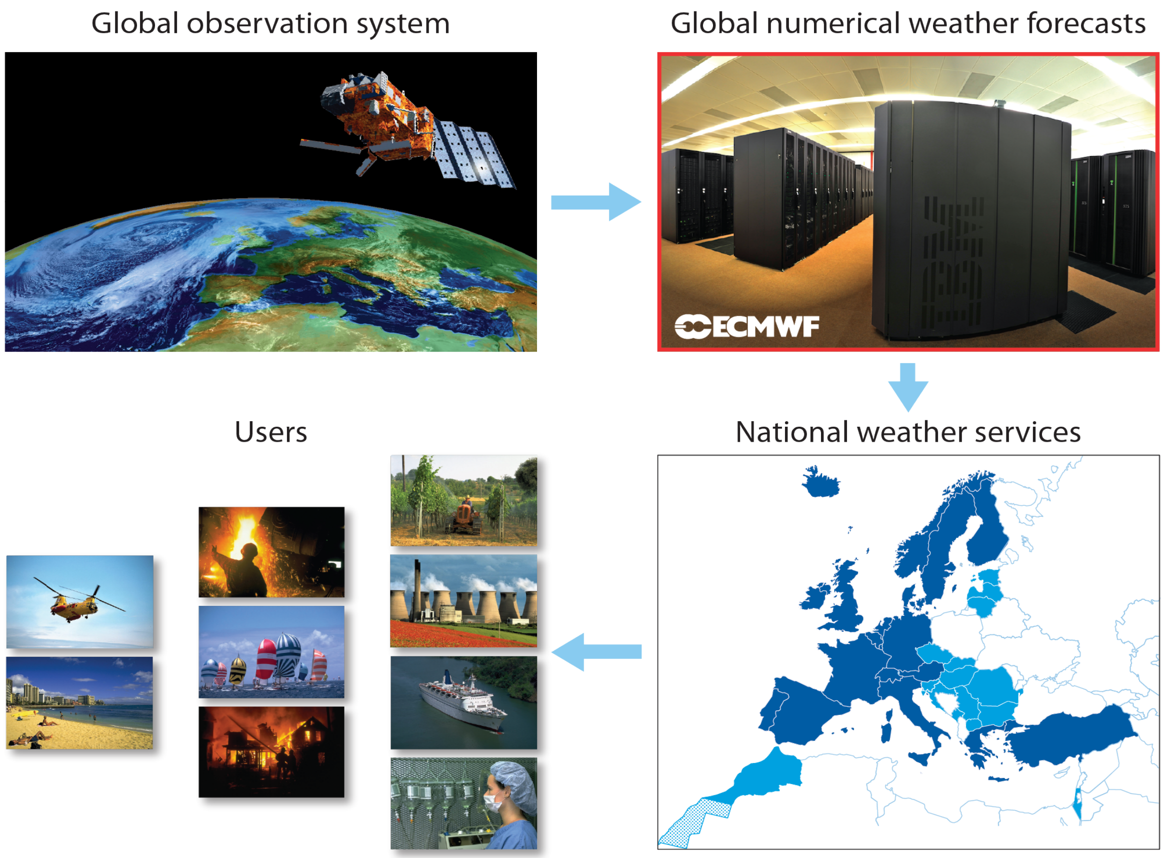
ECMWF

An independent **intergovernmental** organisation

21 Member States
13 Co-operating States



Numerical Weather Prediction @ ECMWF



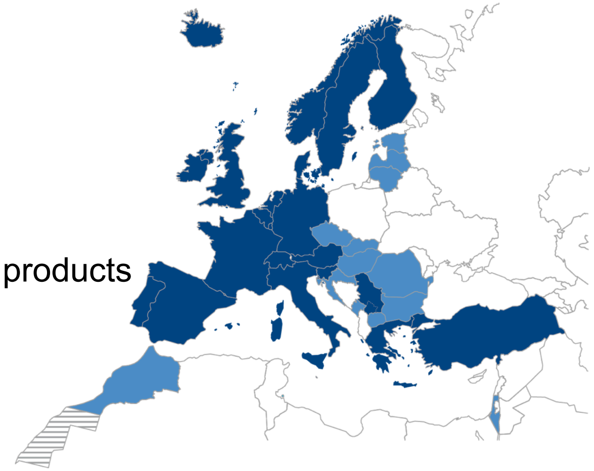
Global observation system

Global numerical weather forecasts

National weather services

Users

ECMWF’s HPC Targets



What do we do?

Operations – Time Critical

- Operational runs – 2 hours from observation cut-off to deliver forecast products
- 10 day forecast twice per day, 00Z and 12Z
- Boundary Conditions 06Z and 18Z, monthly, seasonal, etc.

Research – Non Time Critical

- Improving our models
- Climate reanalysis, etc

HPC Facility Targets

- Capability, minimise the time to solution of Model runs
- Capacity, maximise the throughput of research jobs per day

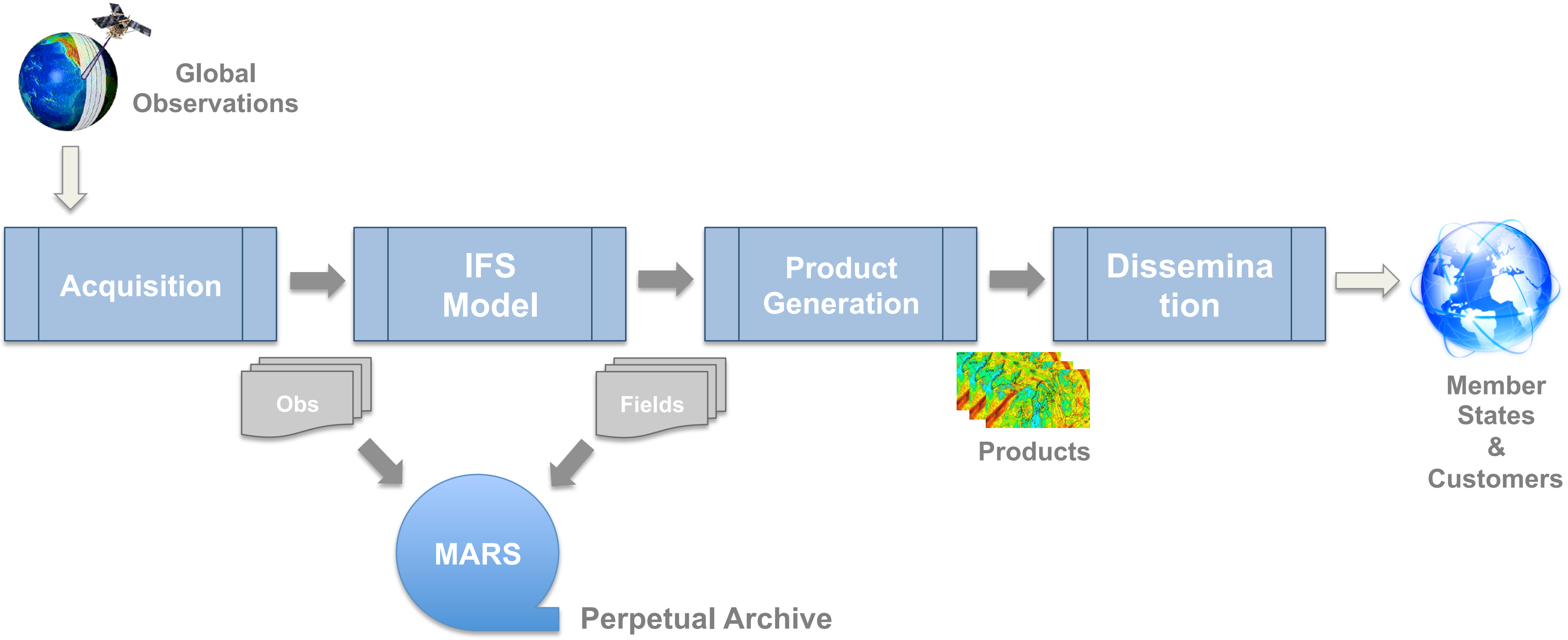
Tension

Time Critical vs. Non Time Critical

Capacity vs. Capability

***Challenge**: design our HPC system to optimise these goals, minimising TCO?*

ECMWF's Production Workflow



Global Observations

Acquisition

IFS Model

Product Generation

Dissemina tion

Member States & Customers

Obs

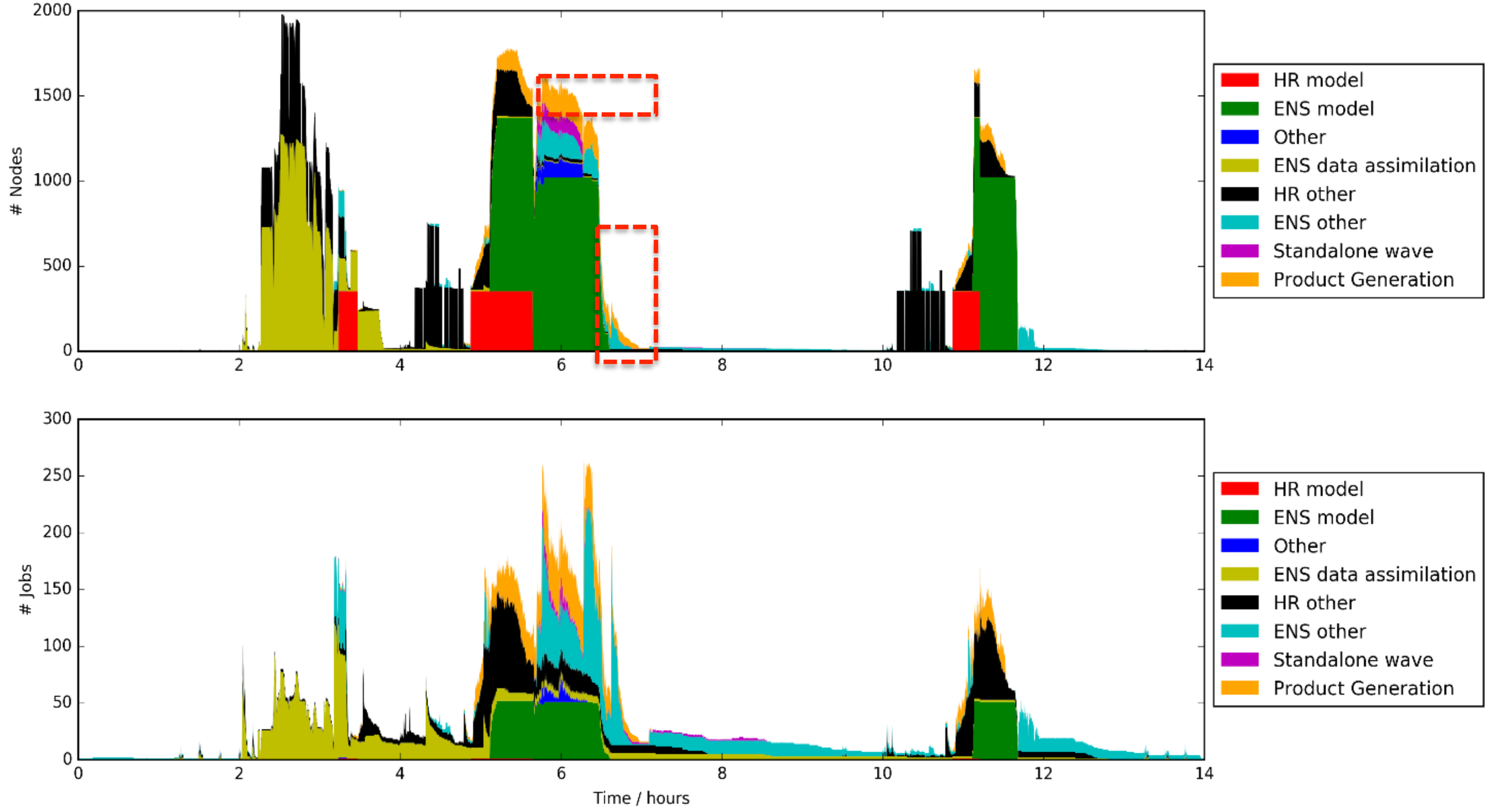
Fields

Products

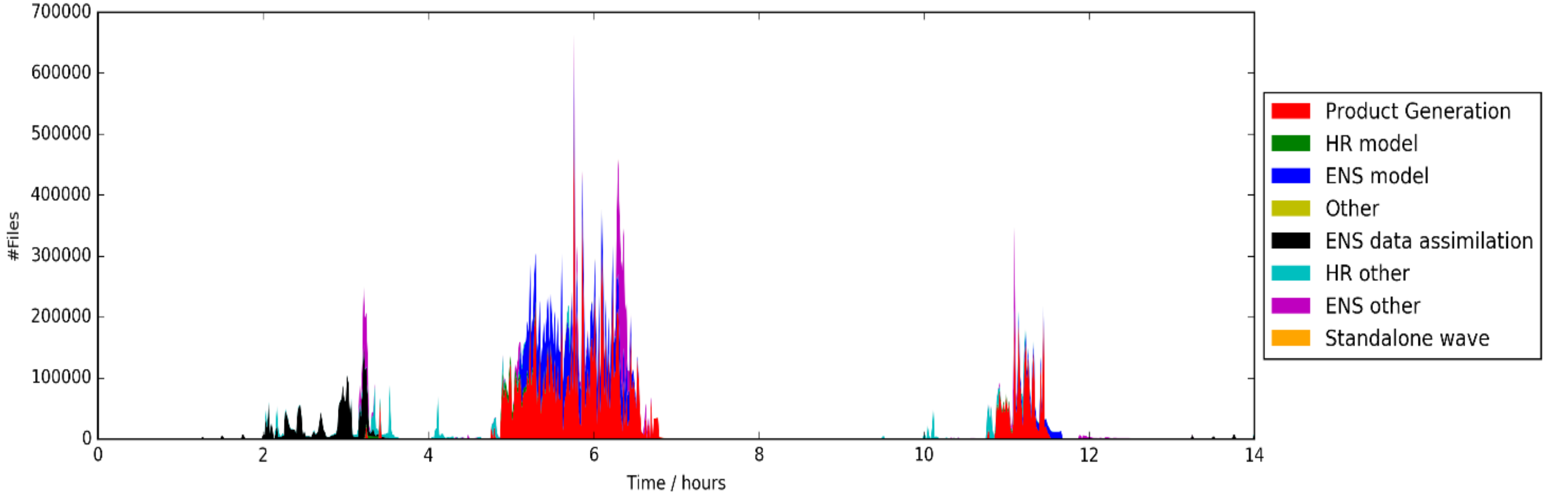
MARS

Perpetual Archive

Operational workload: Job allocation (1 cycle)

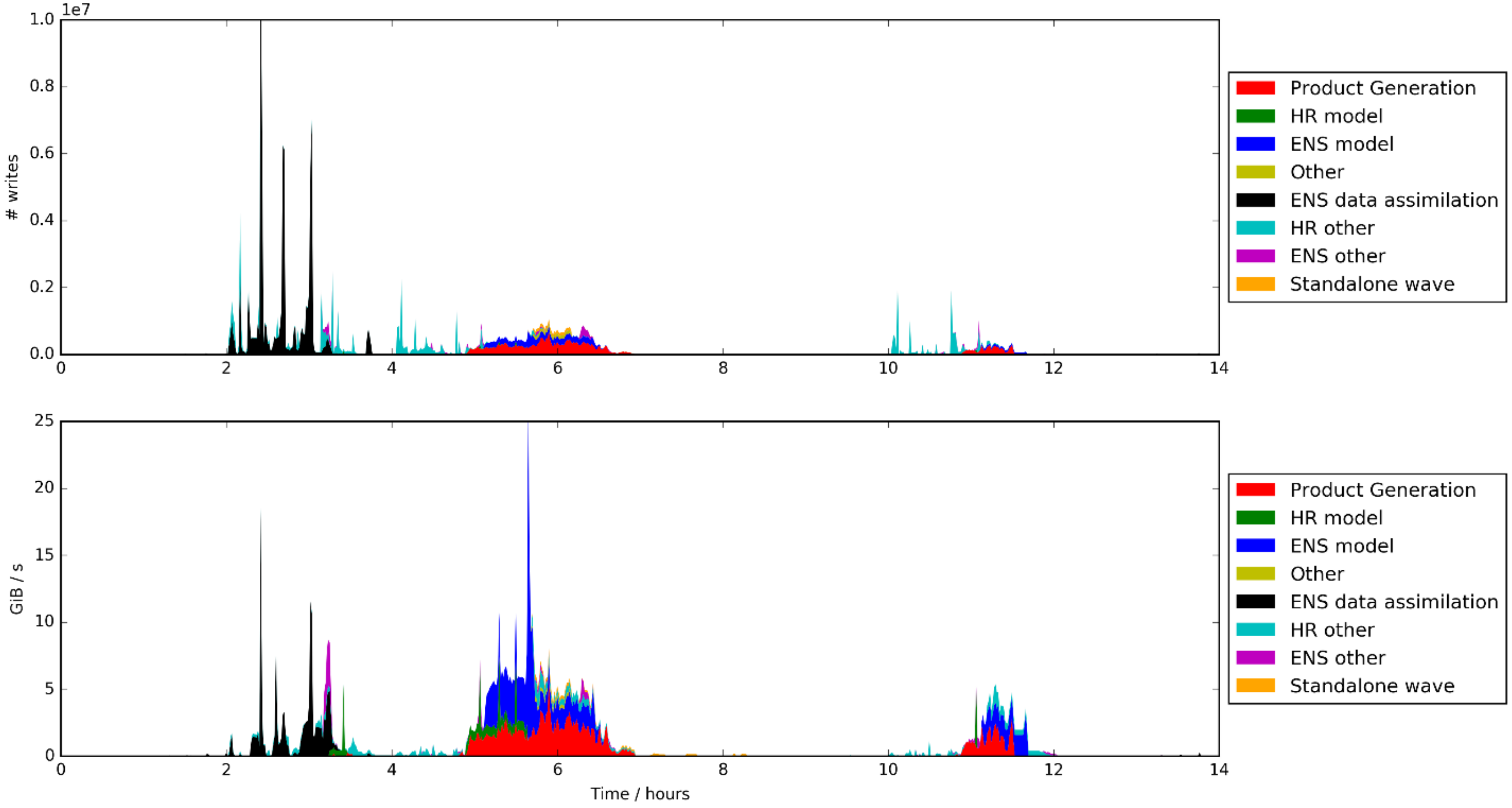


Operational workload: Files opened (1 cycle)

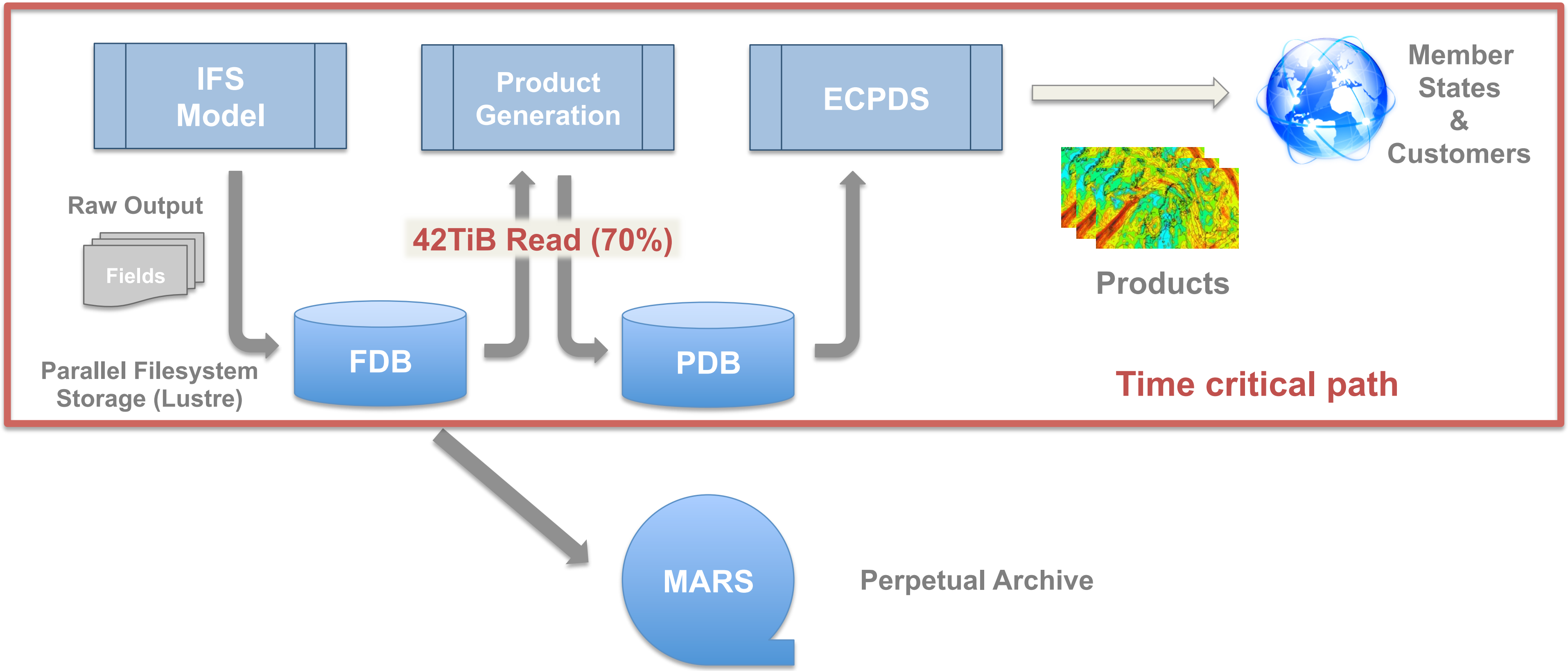


Target Files = # Users x # Steps x # Ranks

Operations workload: Output written (1 cycle)



ECMWF's Production Workflow



Estimated Growth in Model IO

2015

16km, 137 levels

Time critical

- 21 TB/day written
- 22 Million fields
- 85 Million products
- 11 TB/day send to customers

Non-time critical

- 100 TB/day archived
- 400 research experiments
- 400,000 jobs / day

2020

Increase: 2 horizontal, 1 upper air

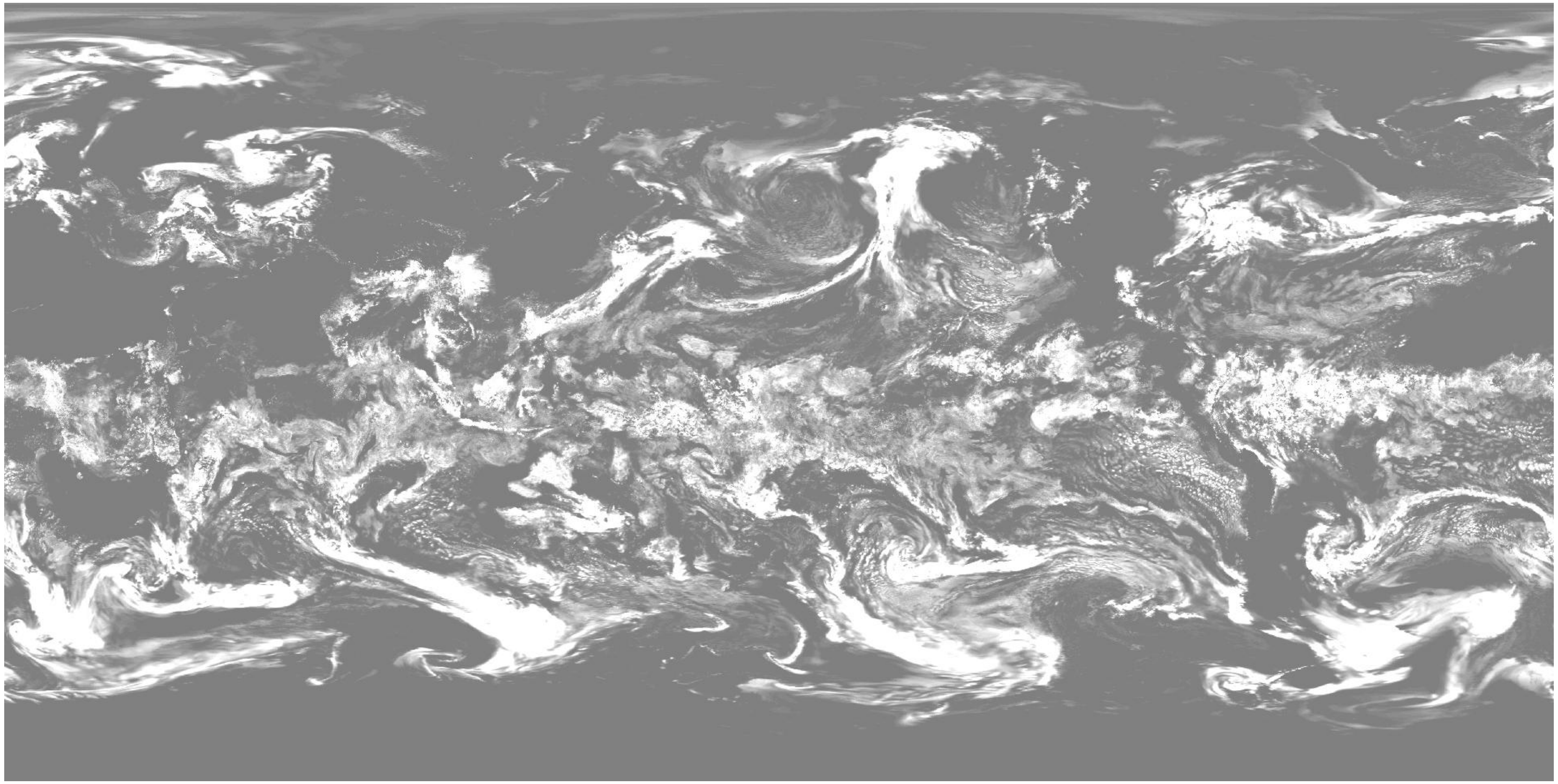
Time critical

- 128 TB/day written
- 90 Million fields
- 450 Million products
- 60 TB/day send to customers

Non-time critical

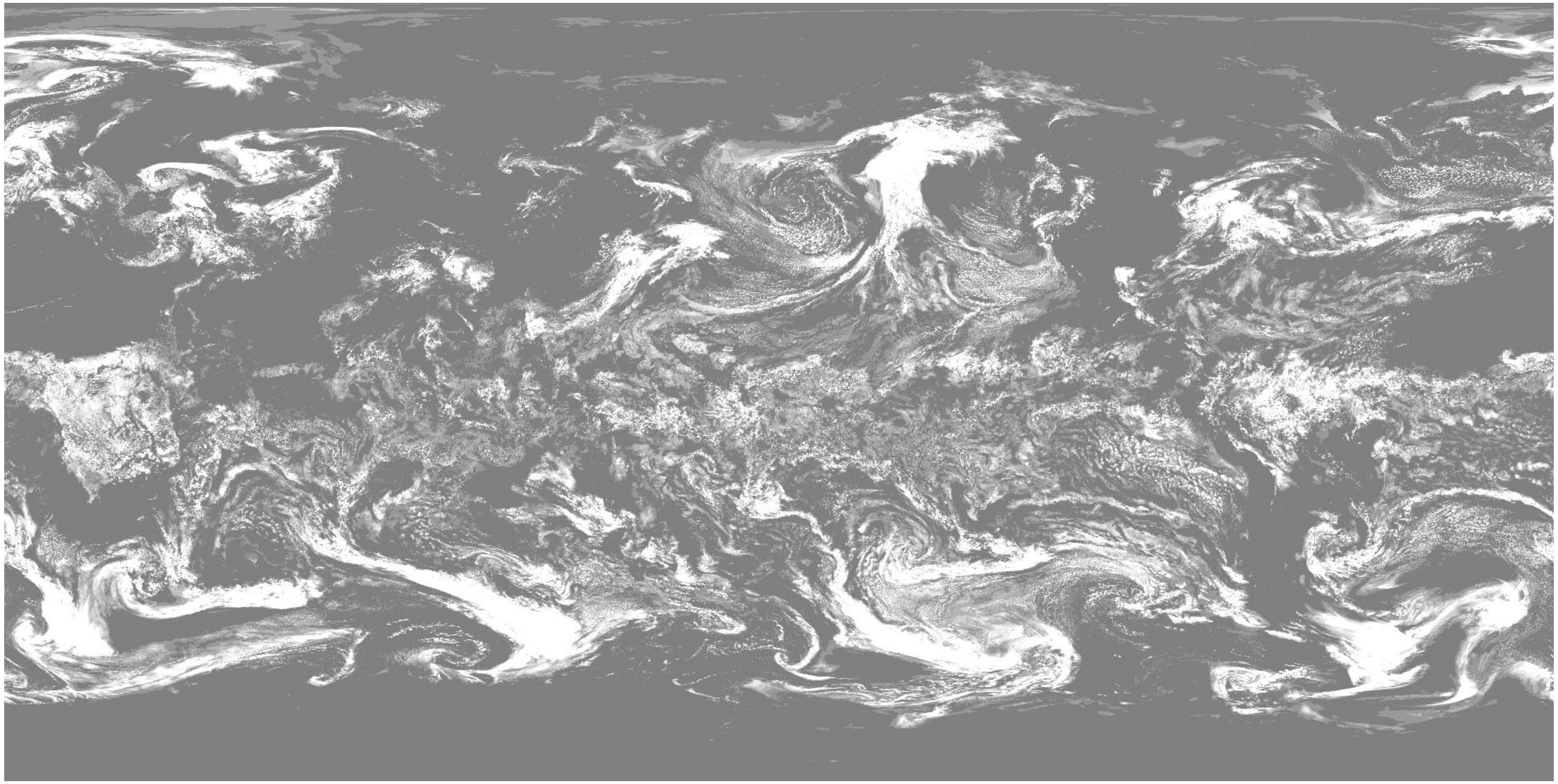
- 1 PB/day archived
- 1000 research experiments

TCo1279 (~9km) a 6.6 Megapixel camera



(12h forecast, *hydrostatic*, with deep convection parametrization, 450s time-step, 240 Broadwell nodes, ~0.75s per timestep)

TCo7999 (~1.25km) 256 Megapixel

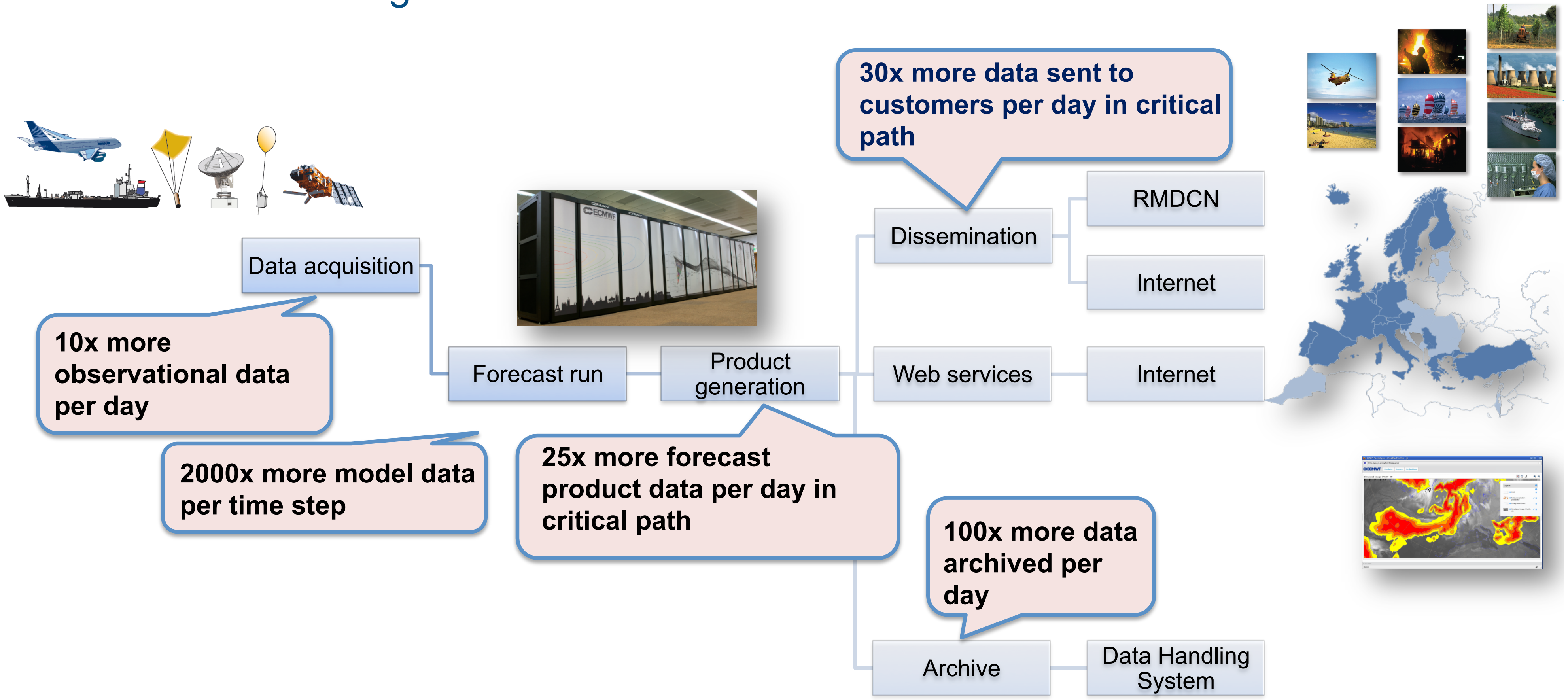


(12 h forecast, *hydrostatic*, no deep convection parametrization, 120s time-step, 960 Broadwell nodes, ~10s per timestep)

History and Future of Resolution Upgrades

Resolution	Grid size	Grid Points	Field Size (in memory)
T319	62.5 km	204 k	1.6 MB
T511	39 km	524 k	4 MB
T799	25 km	1.2 M	9.6 MB
T1279	16 km	2.1 M	16.8 MB
Tco1279	**9 km**	**6.6 M**	**50.4 MB**
Tco1999	5 km	16.1 M	122.6 MB
Tco3999	2.5 km	64 M	490 MB
Tco7999	*1.25 km*	*256 M*	***1909 MB***

10-Year Challenge



Data acquisition

10x more observational data per day

Forecast run

2000x more model data per time step

Product generation

25x more forecast product data per day in critical path

Dissemination

30x more data sent to customers per day in critical path

RMDCN

Internet

Web services

Internet

Archive

100x more data archived per day

Data Handling System

What is NextGenIO?

Integrated into ECMWF's Scalability Programme

Exploring new NVRAM technologies to minimise Exascale I/O bottlenecks

Partners

- EPCC (Proj. Leader)
- Intel
- Fujitsu
- T.U. Dresden
- Barcelona S.C.
- Allinea Software
- ARCTUR
- ECMWF

Project Aims

- Build an HPC prototype system with Intel 3D XPoint technology
- Develop tools and systemware to support application development
- Design scheduler startegies that take NVRAM into account
- Explore how to best use this technology in I/O servers

ECMWF Tasks

- Provide requirements and use cases
- Develop a I/O Workload Simulator
- Explore interation with I/O server layer in IFS
- Test and assess the system scalability

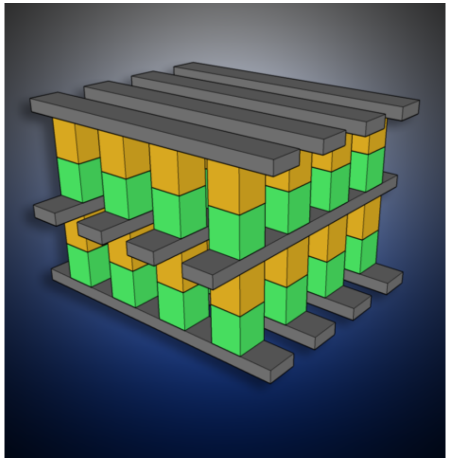
http://www.nextgenio.eu - EU funded H2020 project, runs 2015-2018

NVRAM Intel 3D XPoint

Key characteristics:

- storage **density similar** to NAND flash memory
- **better durability**
- **speed and latency better** than NAND, though slower than DRAM
- priced between NAND and DRAM

Source: https://en.wikipedia.org/wiki/3D_XPoint



"3D XPoint" by Trolomite
Own work. Licensed under CC BY-SA 4.0

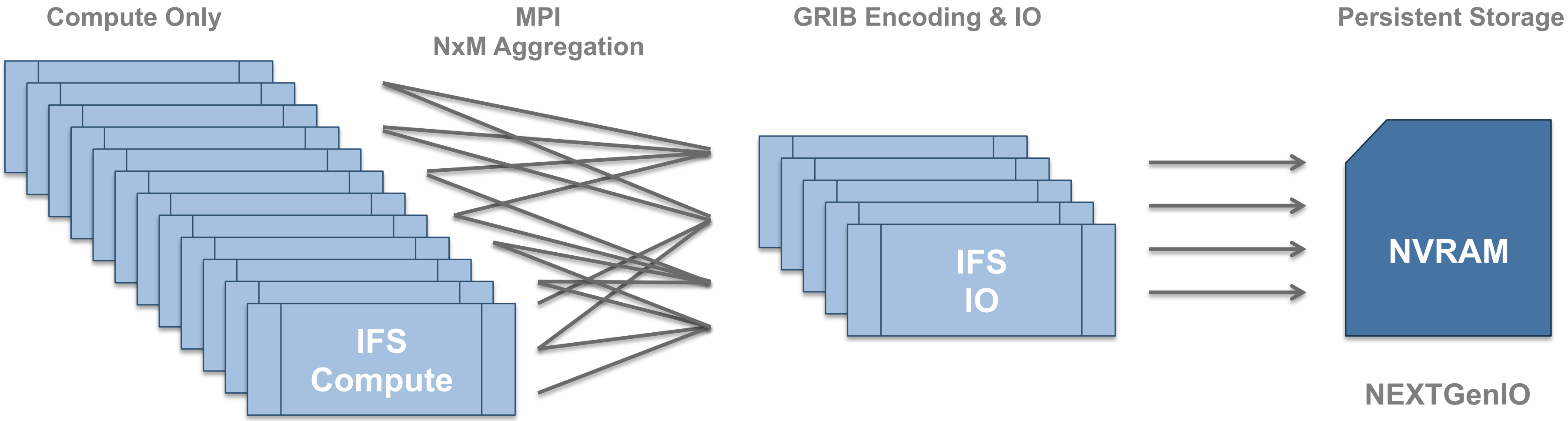
How is ECMWF planning to use this technology?

- **large buffers** for **time critical** applications
  - similar to *burst buffers* but in application space
- **persistence** until archival, for **non time critical**
  - adding a new layer in the hierarchical storage system view

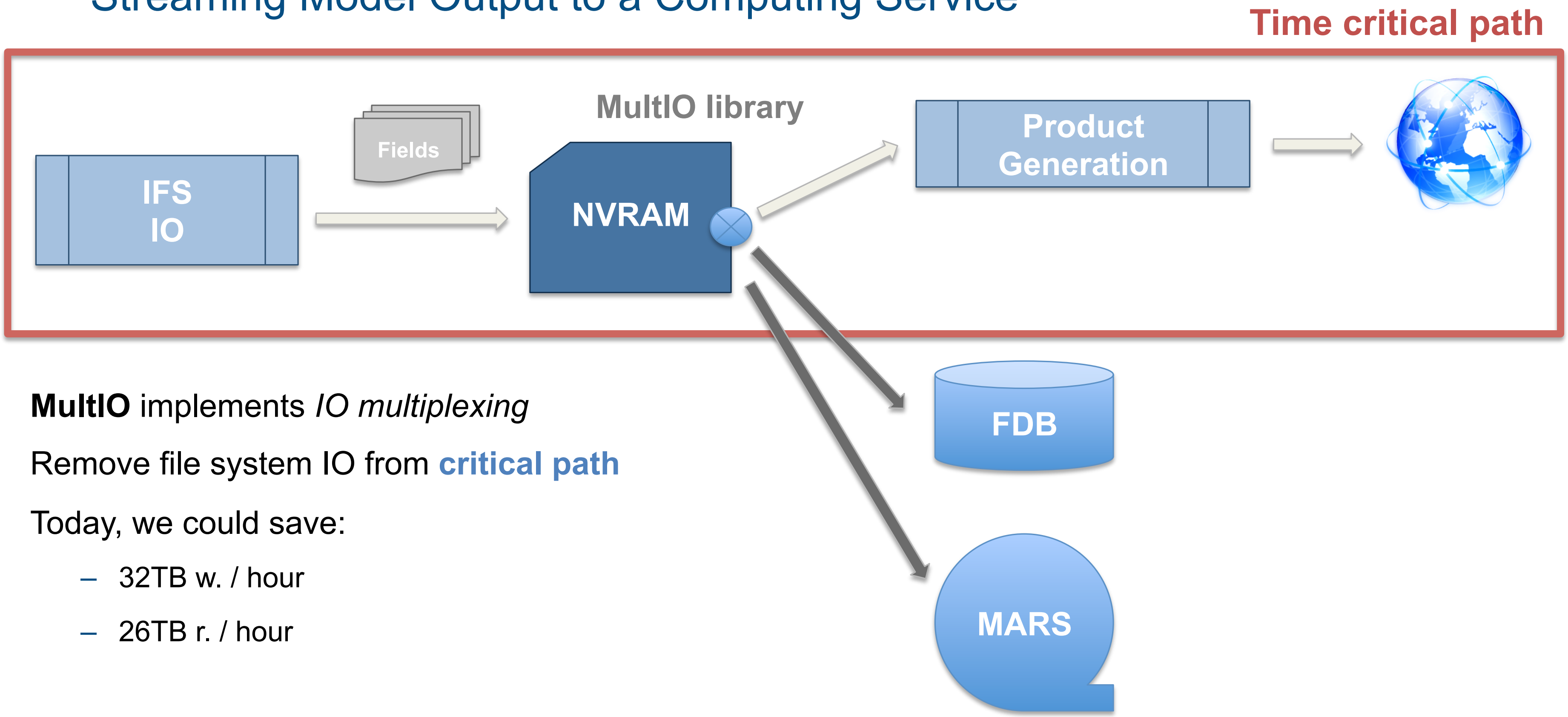
Key Point: High Density at very low latency

IFS IO Server

- Based on MeteoFrance IO server for IFS
- Entered production in March 2016



Streaming Model Output to a Computing Service



MultIO implements *IO multiplexing*

Remove file system IO from **critical path**

Today, we could save:

- 32TB w. / hour
- 26TB r. / hour

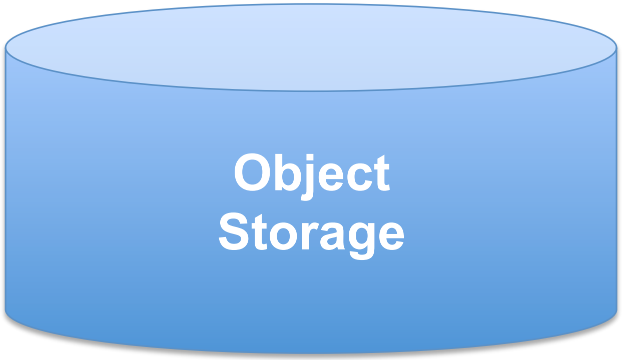
How to store all model output in NVRAM?

Object Store

- Key-Value stores offer **scalability**
  - Just add more instances to increase capacity and throuput
- **Transaction** behavior with minimal synchronization
- Growing popularity, namely due to **Big Data Analytics**

Key: date=12012007, param=temp

Value: 101001...100101010110010



Object Storage

But ECMWF has been using key-value store for 30 years...

MARS

MARS Language

```
RETRIEVE,
   CLASS    = OD,
   TYPE     = FC,
   LEVTYPE  = PL,
   EXPVER   = 0001,
   STREAM   = OPER,
   PARAM    = Z/T,
   TIME     = 1200,
   LEVELIST = 1000/500,
   DATE     = 20160517,
   STEP     = 12/24/36
```

```
RETRIEVE,
   CLASS    = RD,
   TYPE     = FC,
   LEVTYPE  = PL,
   EXPVER   = ABCD,
   STREAM   = OPER,
   PARAM    = Z/T,
   TIME     = 1200,
   LEVELIST = 1000/500,
   DATE     = 20160517,
   STEP     = 12/24/36
```

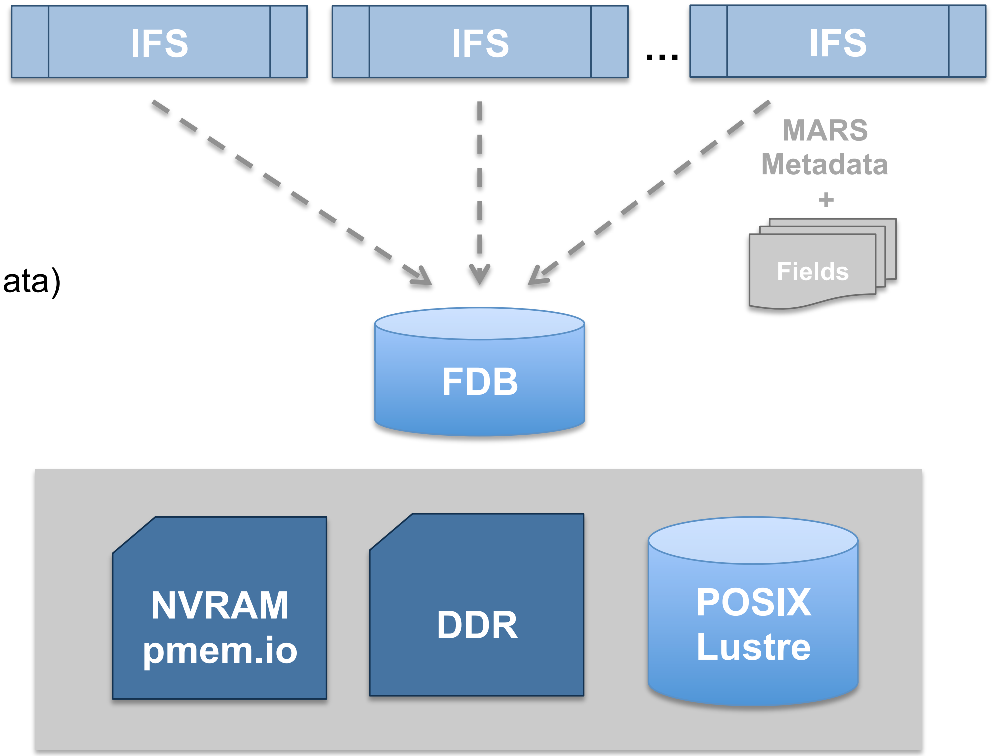
Unique way to describe all ECMWF data both
Operational and **Research**

FDB (version 5)

- Domain specific (NWP) object store
- Transactional, No synchronization
- Key-value store
  - Keys are scientific meta-data (MARS Metadata)
  - Values are byte streams (GRIB)

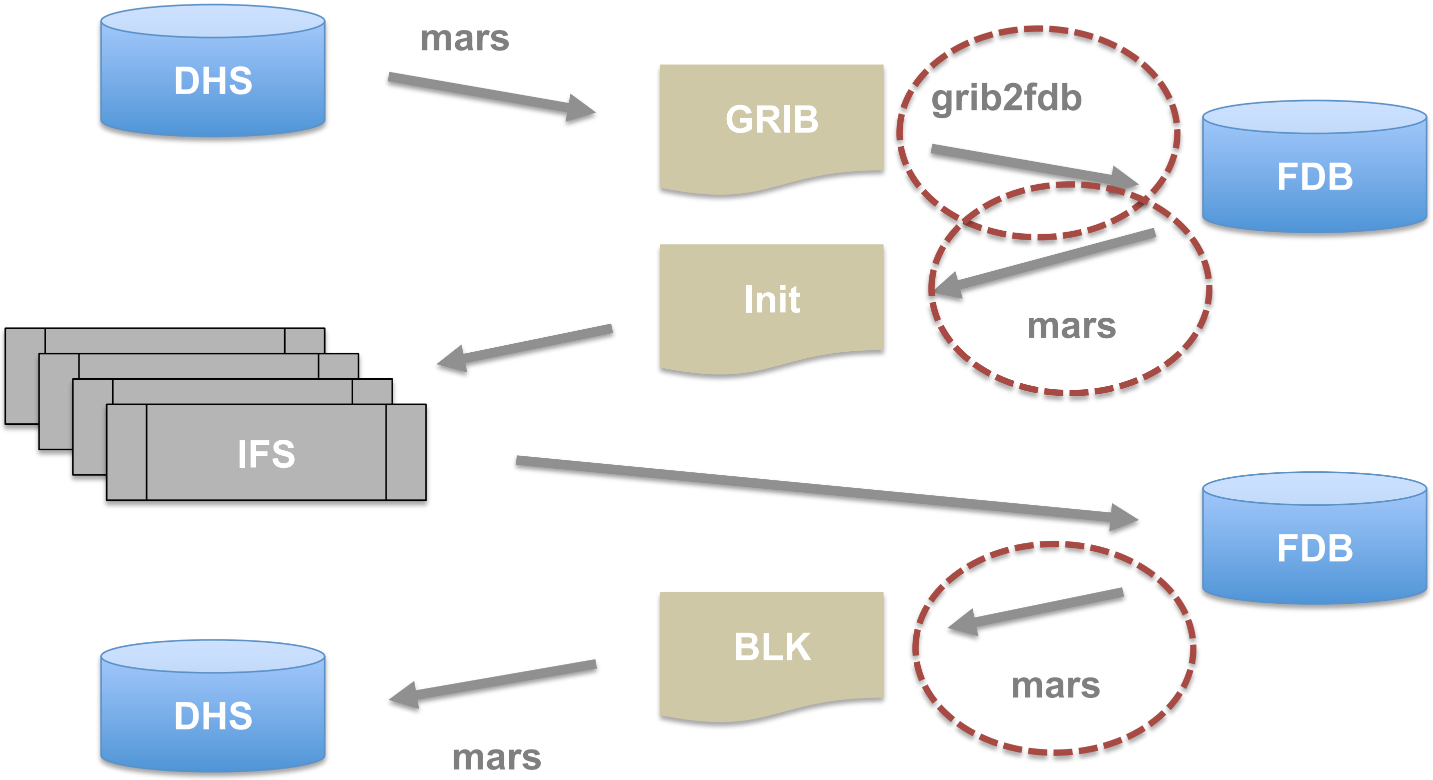
- Support for multiple back-ends:
  - POSIX file-system (currently on Lustre)
  - 3D XPoint using pmem.io library
  - Could explore others:
    - Intel DAOS, Cray DataWarp, etc.

- Supports wild card searches, ranges, data conversion, etc…

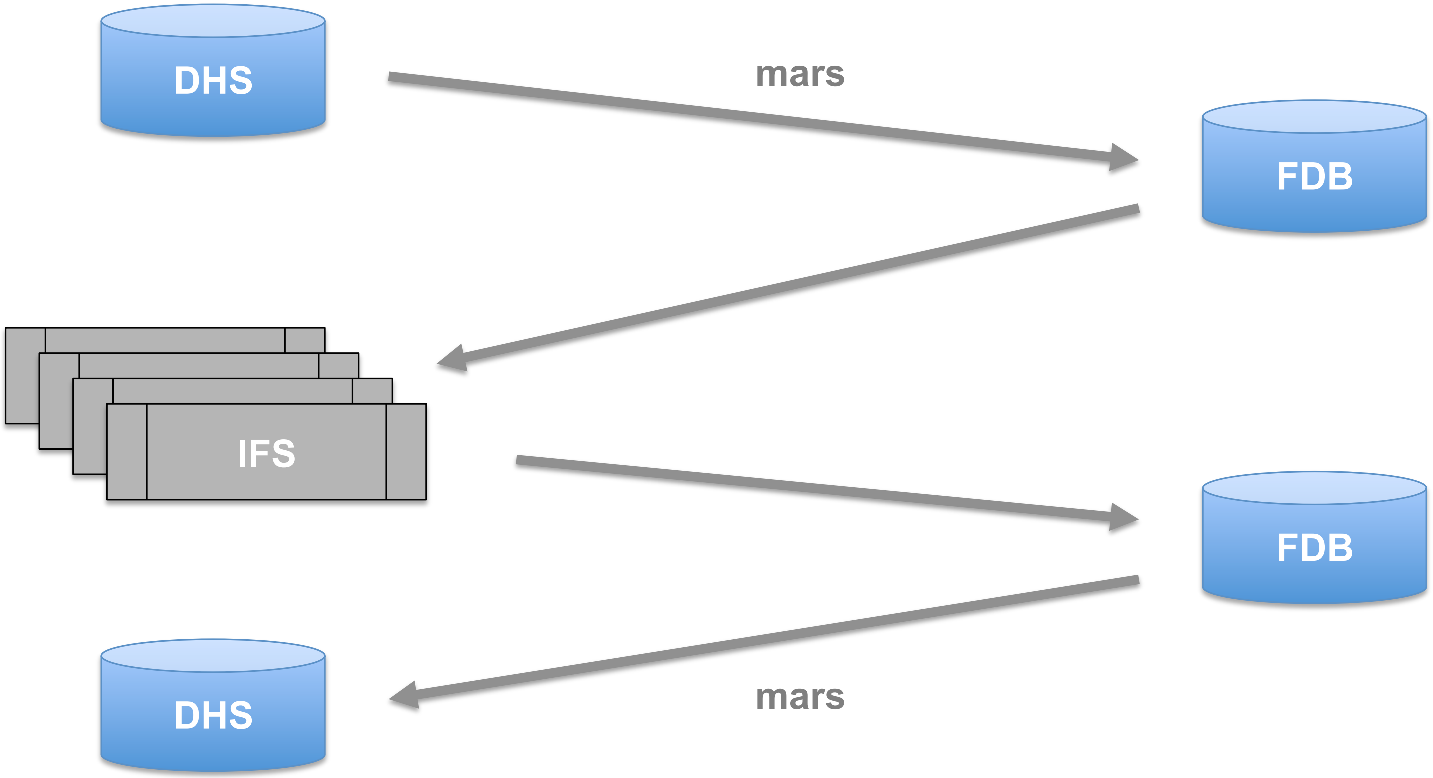


```
param=temperature/humidity,
levels=all,
steps=0/240/by/3
date=01011999/to/31122015,
```

Current Workflow



New Workflow

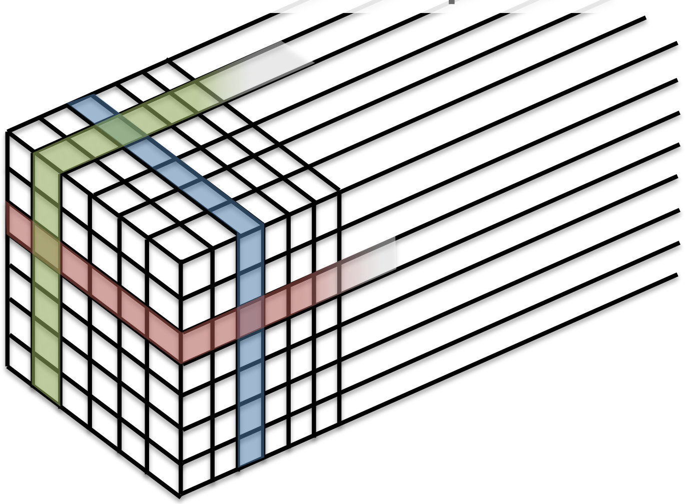


Data Axis

Byte Addressable Hypercubes

- Longitude (3600)
- Latitude (1800)
- Atmospheric levels, Physical parameters (~200)
- Time steps (~100)
- Probabilistic pertubations (50)

Clients want to do **different** analytics across **multiple** axis



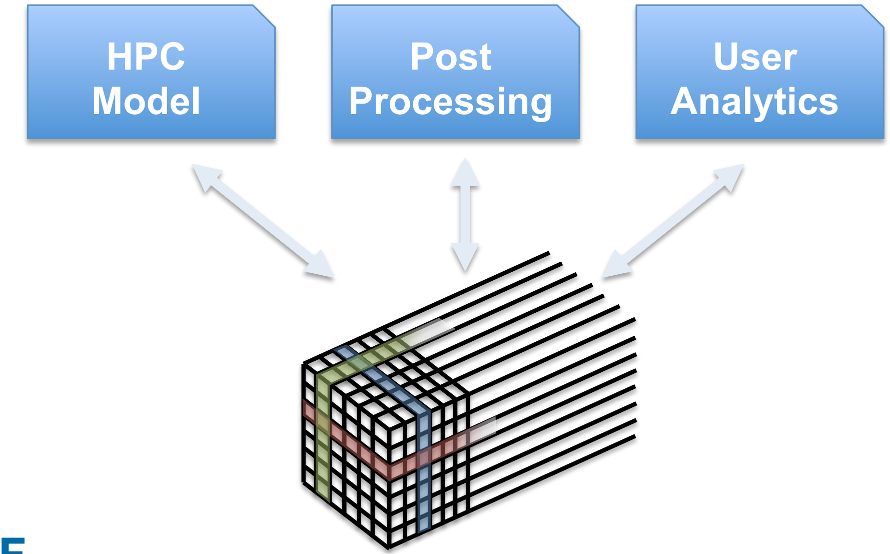
@ double precision

- 9km **48 TiB**
- 5km **192 TiB**
- 1.25km **1.82 PiB**

Not included: *historical observations, multiple models, etc...*

Data Centric Computing

- **Producer-Consumer** model, where *HPC is producer*
- Use data while is **hot**
- Bring **users** to the data, ship ***functions***
- Don't use **files,** use **science to communicate**, use **rich metadata**
- Need to **build shared components** amongst the communities...

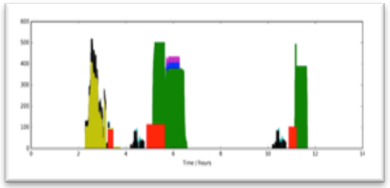


Conclusions & Questions

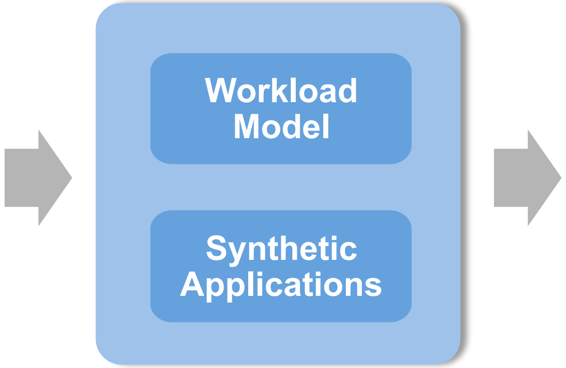
- NWP has had I/O **exponential growth** for many years.
- What is different?
  - Moving from **compute centric to data centric** paradigm
  - Minimise data movement and bring compute to data
- Update our **legacy codes and workflows** to this new paradigm
- How to **adapt upcomming technologies** for complex workflows?
  - Burst Buffers
  - NVRAM
  - Storage-side compute
  - Object stores
- Can we move **beyond the filesystem**? How intrusive should that be?
  - Interpreting scientific data as objects
  - Challenges in data modelling and data curation

Kronos HPC Workload Benchmarking

- **Kronos benchmarks HPC systems by deploying realistic workloads**
  - A **workload model** is generated from **HPC workload profiling data**
  - Generate **a easily-portable workload with synthetic applications**
  - Models and benchmarks **Compute, Interconnect, I/O subsystems**

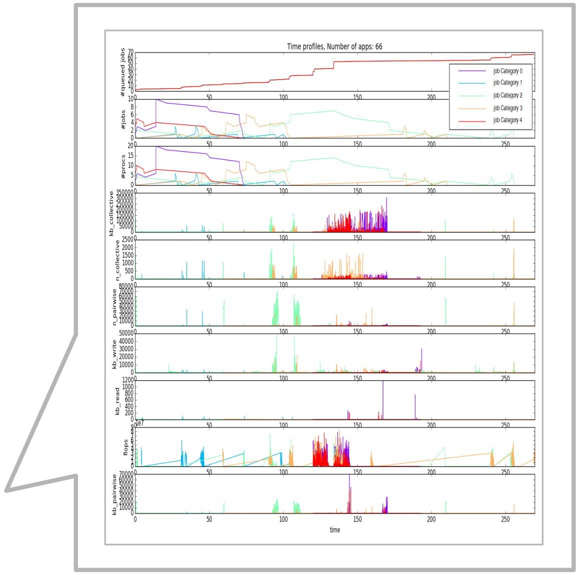


HPC Workload Profile Data



Kronos

HPC Prototype



CPU, Interconnect, I/O metrics retrieved and analysed

Messages To Take Home

Ensemble data sets are growing quadratically to cubically in size,
and this brings an ***I/O crisis*** *for time critical applications*

New technologies (Burst Buffers, SSD's, NVRAM)
are filling in the ***I/O Gap***
but *will change the way we use and store data*

ECMWF is adapting its workflow to take advantage of these upcoming technologies

What would you do differently,
if your persistent storage would be 10,000x faster?

NEXTGenIO has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 671951

Questions?