ECMWF's Extreme Data Challenges on the HPC and Cloud systems

T. Quintino, B. Raoult, S. Smart, J. Hawkes, P. Bauer

ECMWF

tiago.quintino@ecmwf.int

Extreme Data Workshop, Jülich.

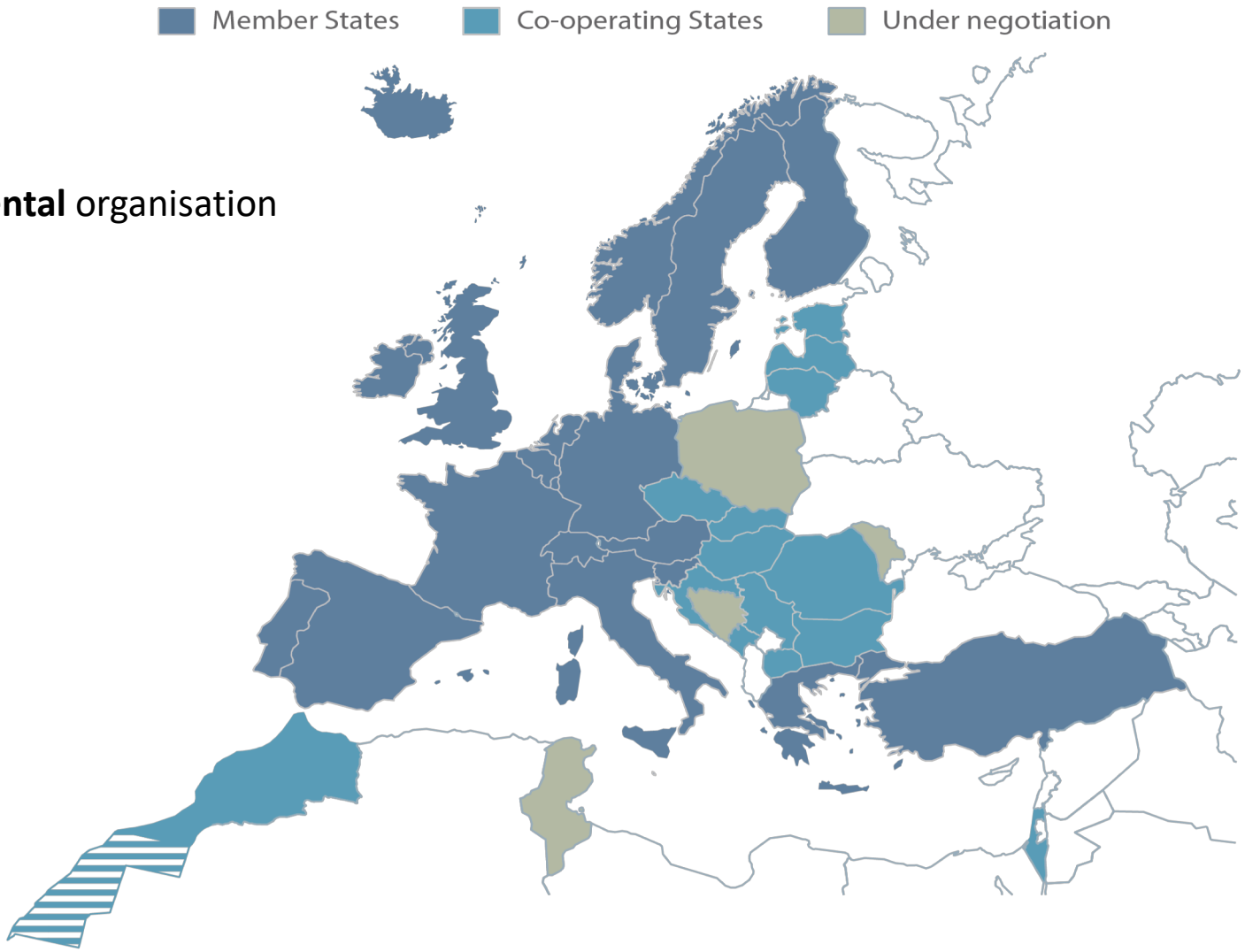
18th September 2018

ECMWF

ECMWF

An independent **intergovernmental** organisation

21 Member States
13 Co-operating States



ECMWF

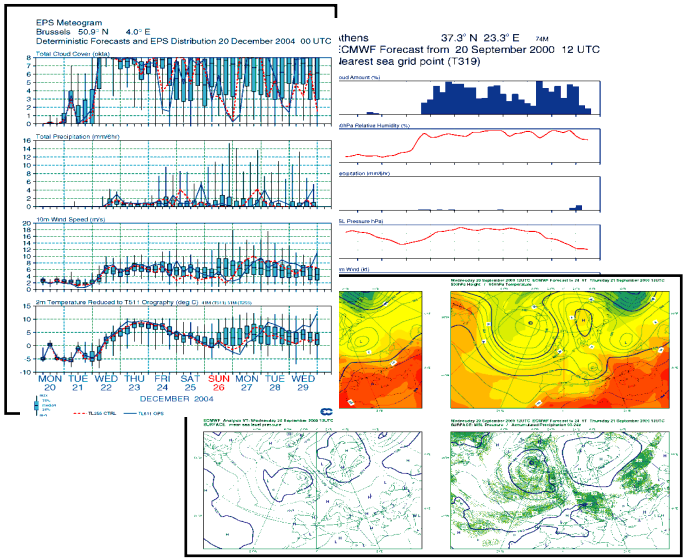
Weather Forecasts — We produce **global weather forecasts**

Medium-Range — 15 days ahead + monthly and seasonal forecasts

Additional Missions — Research and develop weather forecast methods
Maintain a Meteorological Data Archive
of historical ***observations and forecasts***

Additional Services — EU Copernicus Climate Change Service
EU Copernicus Atmospheric Monitoring Service

Reading, United Kingdom



What do we have to achieve this?

People — About 300 staff,
specialists and contractors

Equipment — State-of-the-art supercomputers and data handling systems

Budget — £50 million per year

ECMWF's Forecasting Systems

What do we do?

Operations – **Time Critical**

- Operational runs – 2 hours from satelite cut-off to deliver forecast products
- HRES 0-10 day, 00Z+12Z
  - O1280 (9km) 137 levels
- ENS 0-15 day, 00Z+12Z
  - O640 (18km) 91 levels
- ENS extended 16-46 day, twice weekly
  - O320 (36km) 91 levels
- BC 06Z and 18Z
  - hourly post-processing 0-5 days

Research – **Non Time Critical**

- Experiments to improving our models
- Reforecasts, Climate reanalysis, etc

Global weather forecasting system @ 2018

200 million daily observations

150 TiB daily archived

Data assimilation

First guess

Archive

Initial conditions

Model next step

Intermediate results

Laws of physics

Prediction

Atmosphere is divided in small cells

60 TiB daily products

ECMWF's Forecasting Systems

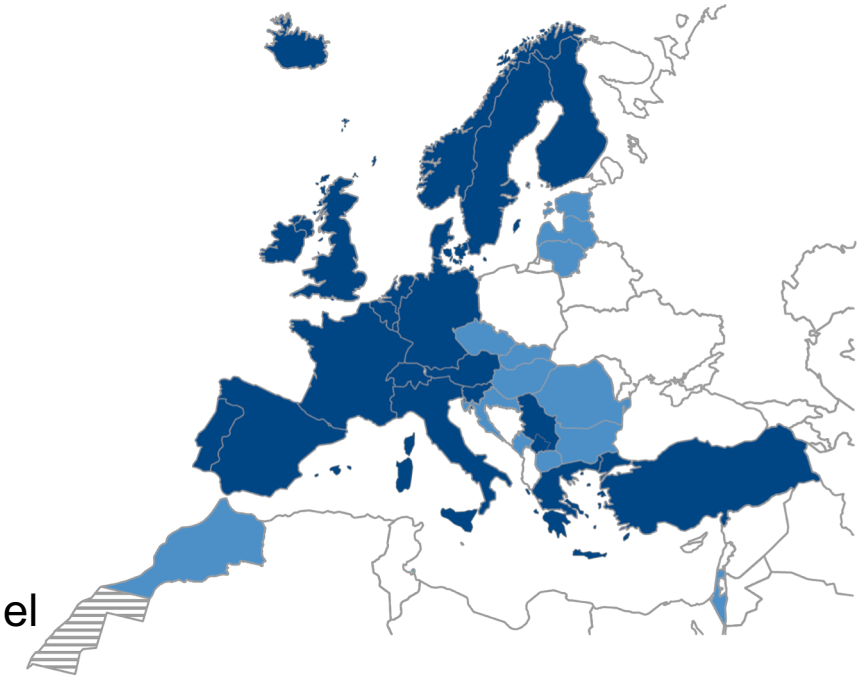
Where do we want to go to?

Stragegy 2025

- **5 km** global ensemble

2020-21

- global x-member ensemble @ 9 km
- 0-15 day in critical path
- coupled to land, ¼ degree ocean and a sea-ice model
- including prognostic atmospheric composition
- initialized with x-member hybrid variational/ensemble analysis



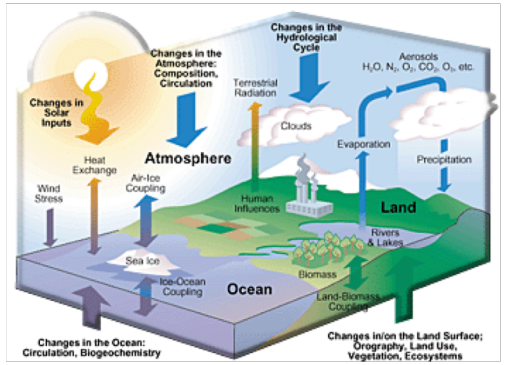
Multiple dimensions

➔Reliability

Ensembles

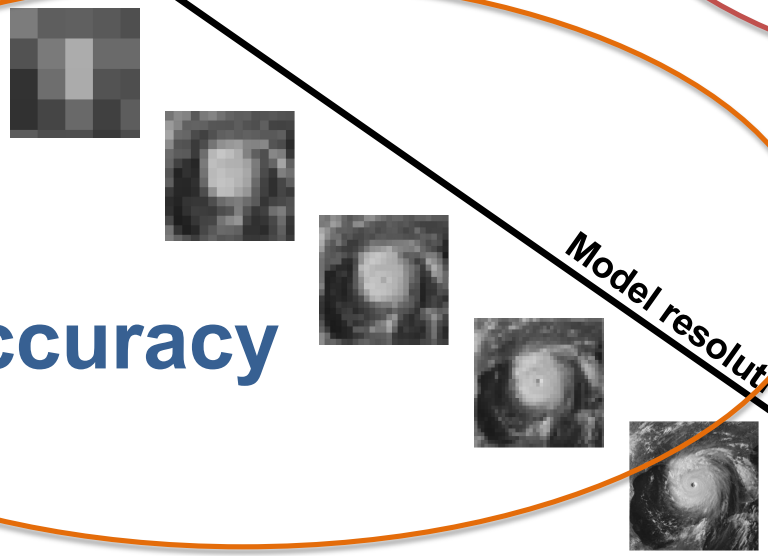
Traditional weather science domain

Model complexity



➔Range

Traditional climate science domain



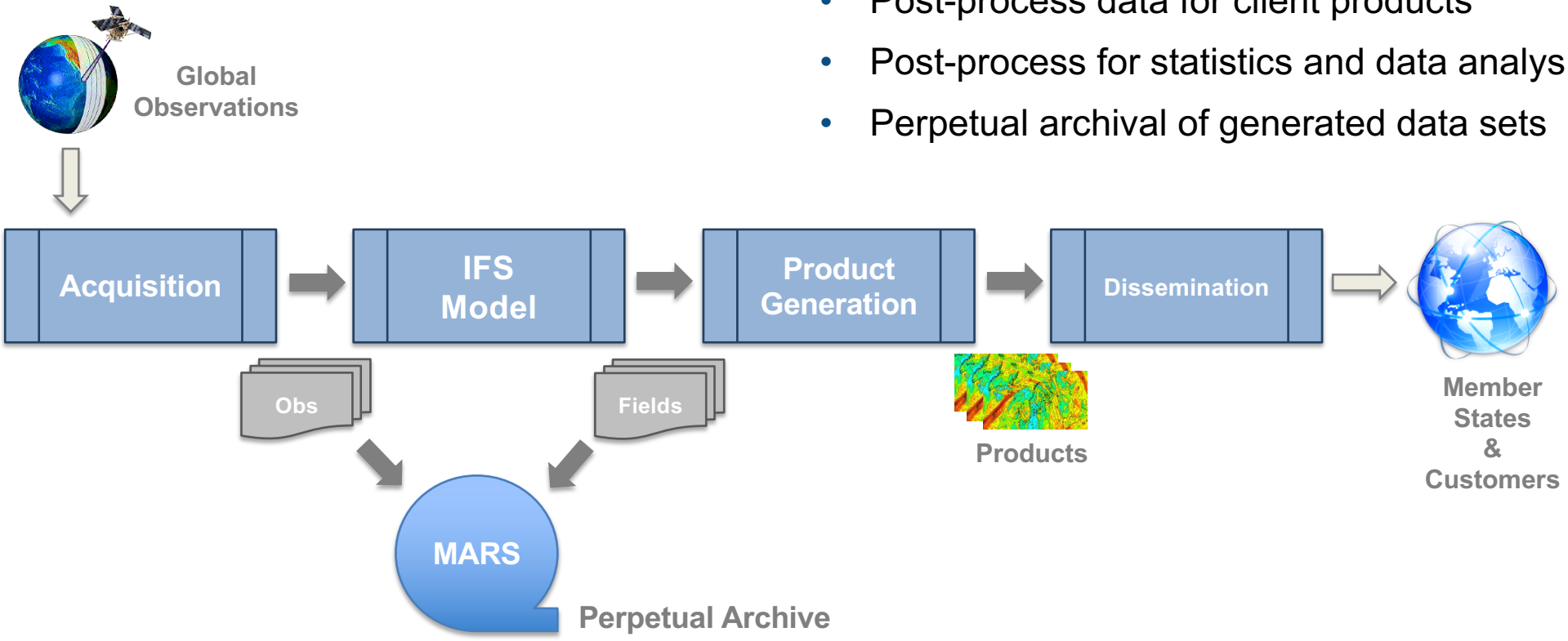
➔Accuracy

Model resolution

Today: it needs high-resolution, 'Earth system' model ensembles to perform at all scales!

Challenges

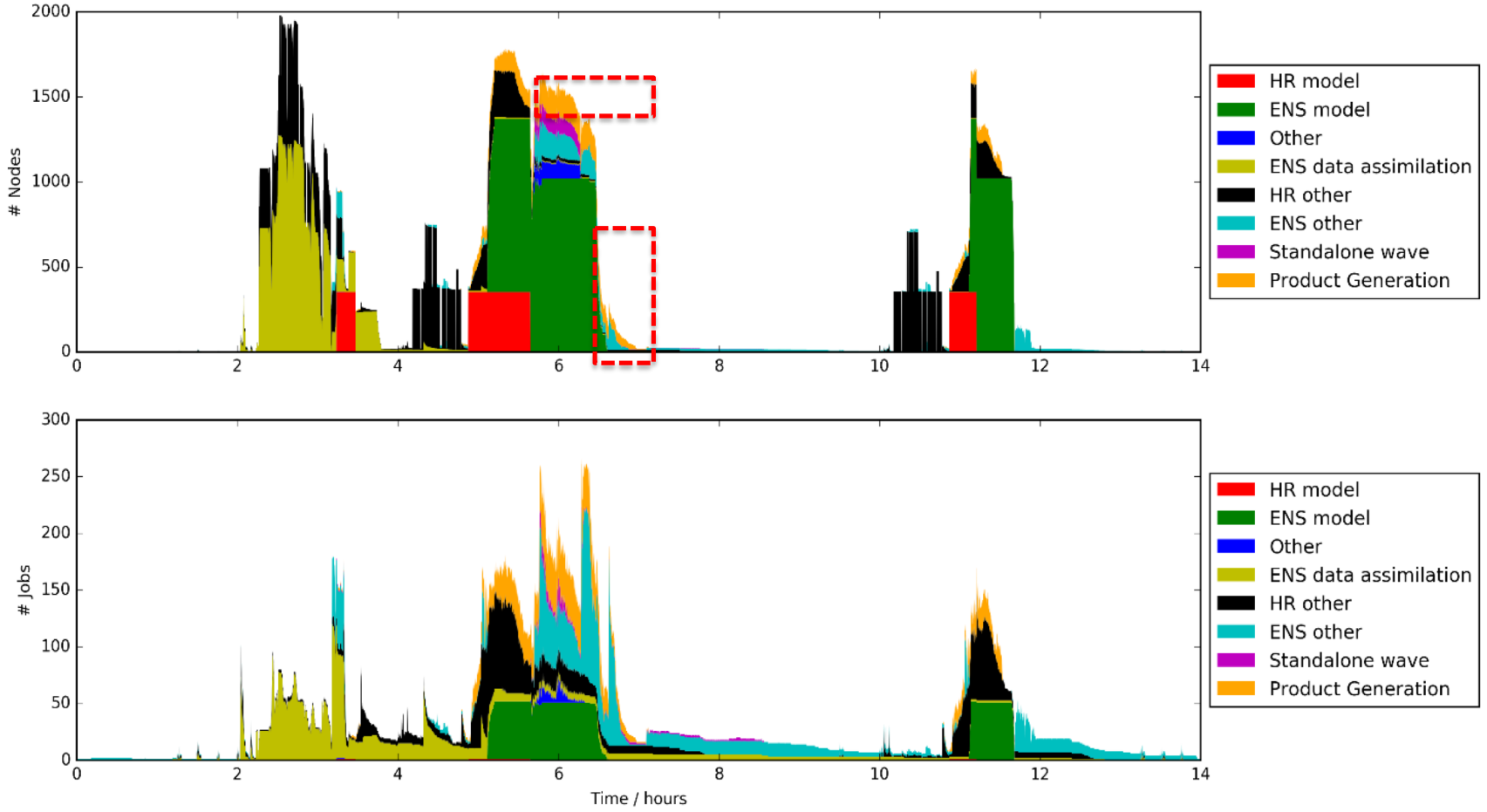
ECMWF's Production Workflow



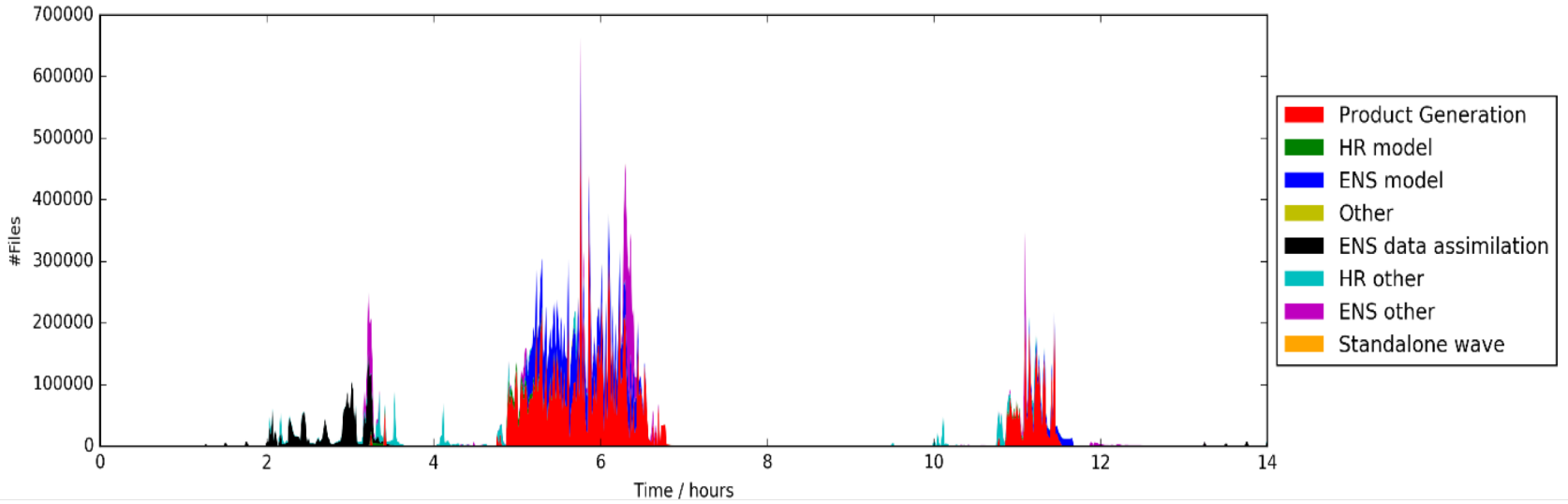
Data Workflow

- Post-process data for client products
- Post-process for statistics and data analysis
- Perpetual archival of generated data sets

Operational workload: Job allocation (1 cycle)

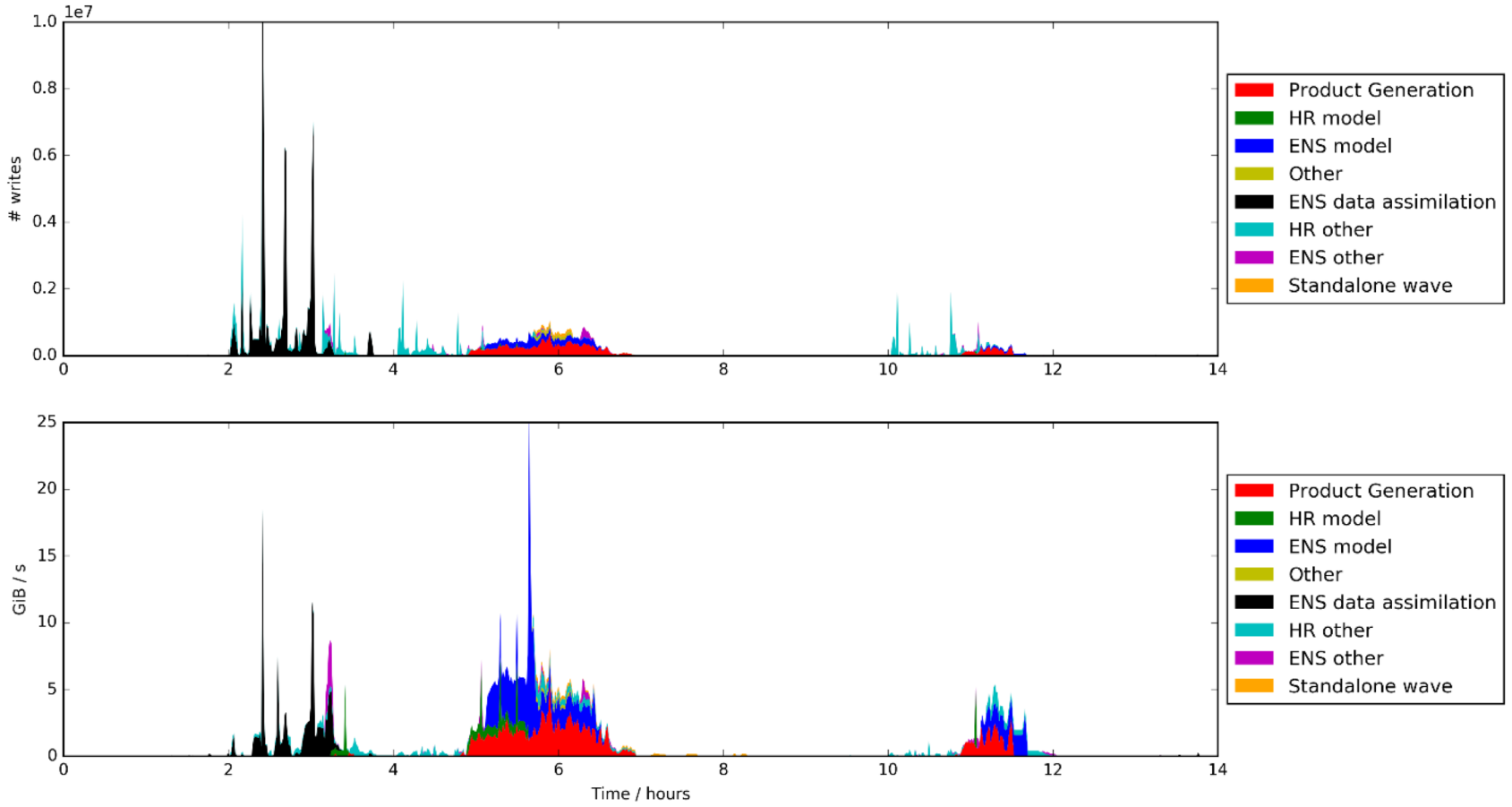


Operational workload: Files opened (1 cycle)

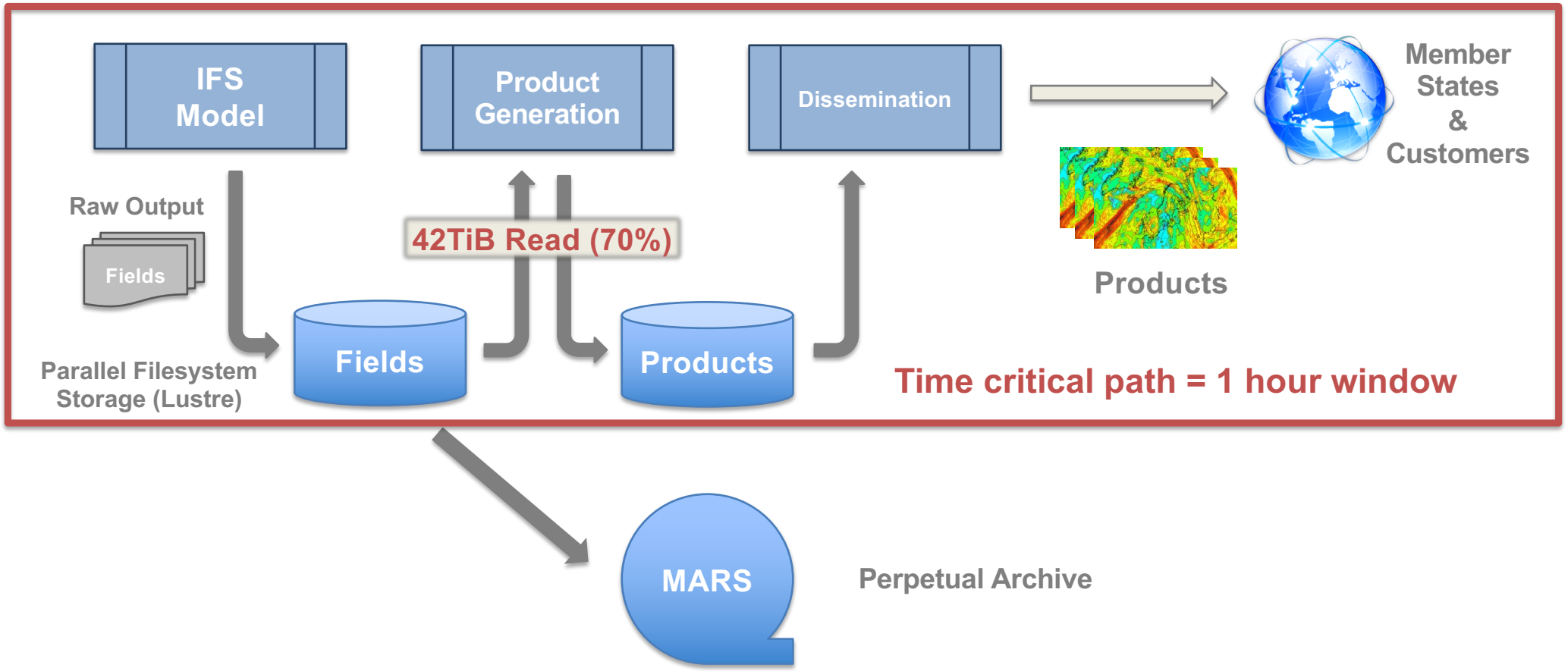


Target Files = # Users x # Steps x # Ranks

Operations workload: Output written (1 cycle)



ECMWF's Production Workflow



Effects of Product Generation

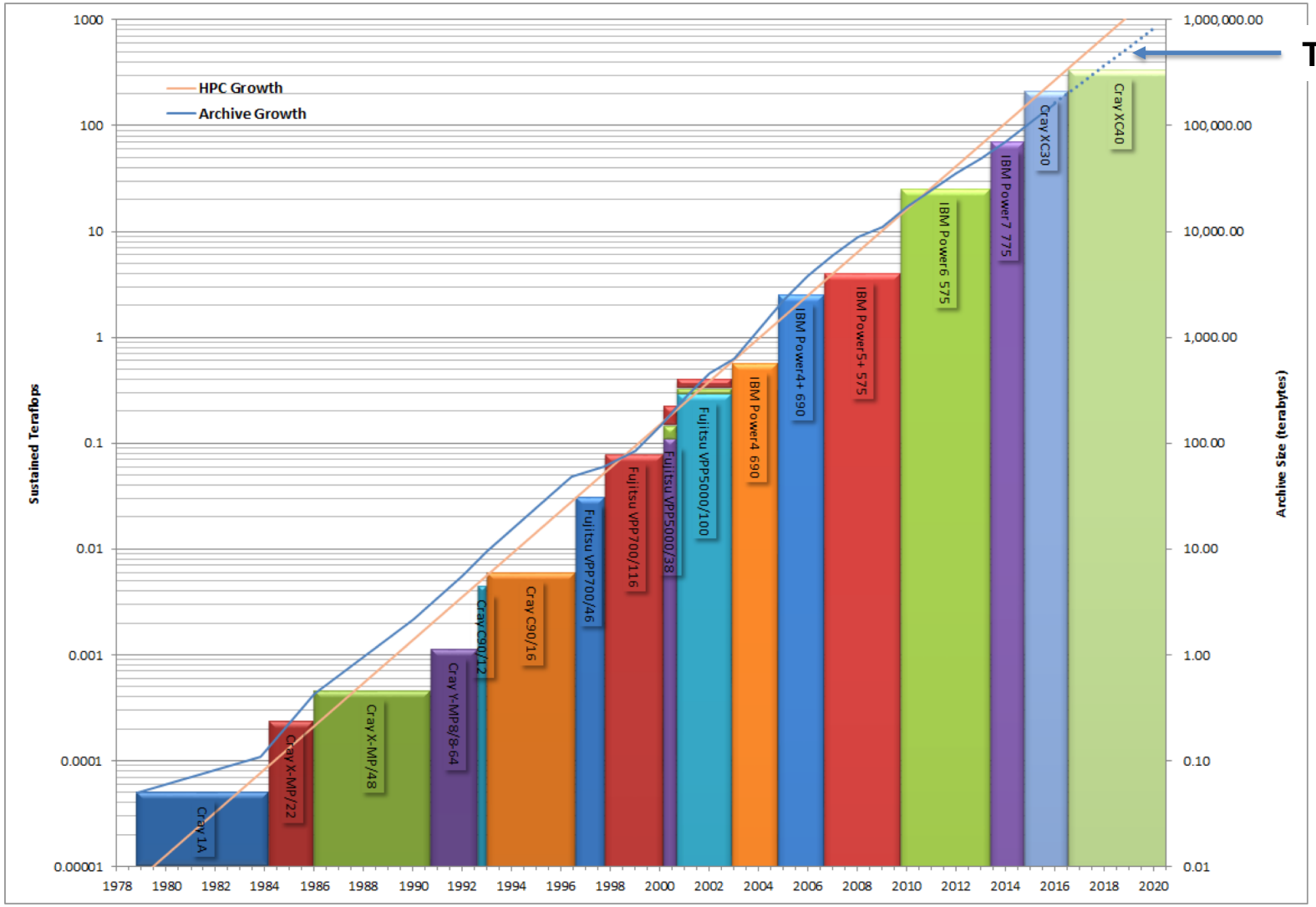
	Model	Model + I/O	Model + I/O + PGen
Nodes	2440	2776	2926
Run time [s]	5765	6749	7260
Relative	-	**+ 17%**	**+ 26%**

Broadwell nodes 2x18 cores
Cray XC40 Aries interconnect
Lustre FS IOR 90GiB/s

Estimated Growth in Model IO

2015	2018	2020
16km, 137 levels	**9km, 137L**	**Increase: 2 horizontal, 1 upper air**
Time critical		**Time critical**
• 21 TiB/day written	**60 TiB/day**	• 128 TiB/day written
• 22 Million fields		• 90 Million fields
• 85 Million products		• 450 Million products
• 11 TiB/day send to customers		• 60 TiB/day send to customers
Non-time critical		**Non-time critical**
• 100 TiB/day archived	**150 TiB/day**	• 1 PiB/day archived
• 400 research experiments		• 1000 research experiments
• 400,000 jobs / day		

Archive growth versus HPC performances (log scale)



Today 200 PiB

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***

The tendency of memory per core diminishing …

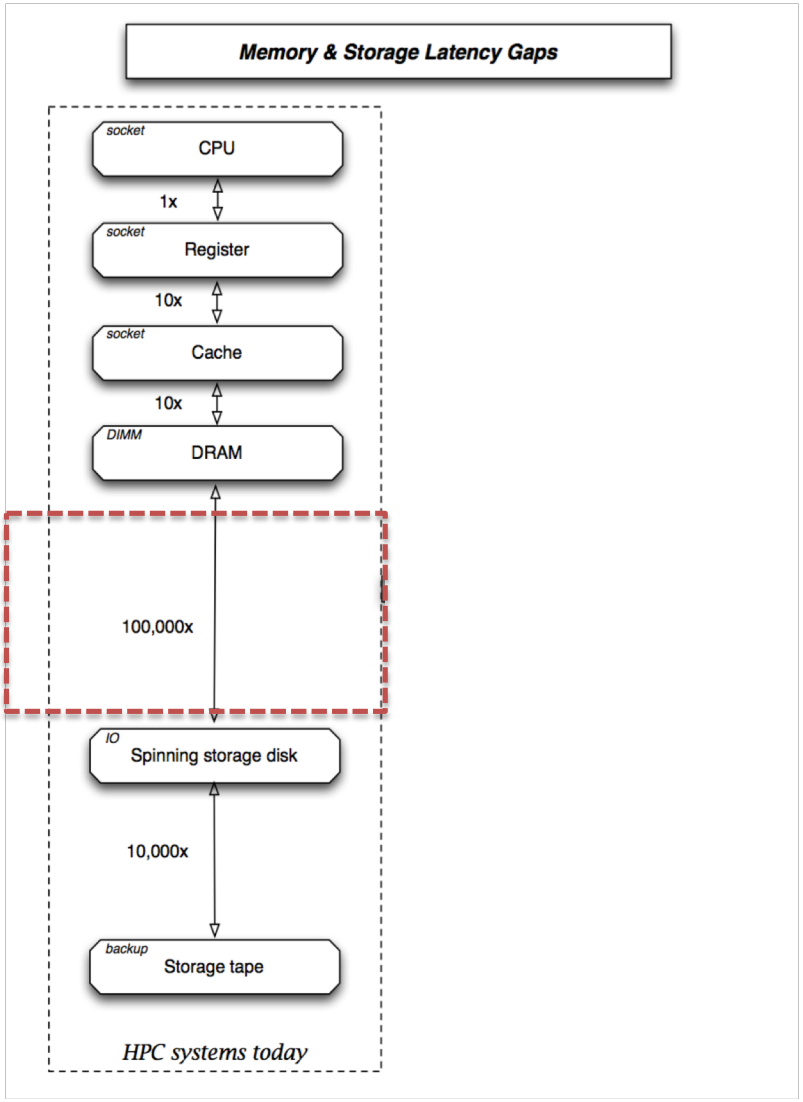
… is likely to have serious implications on the post-processing workflows!

"A supercomputer is a device for turning ***compute-bound*** problems into ***I/O-bound*** problems."

-- Kenneth E. Batcher, Prof. Emeritus, Kent State Univ.

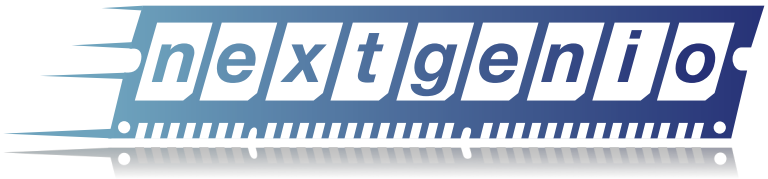
Feeling the Byte?

I/O Gap



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

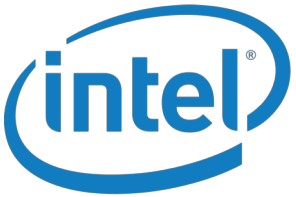
- Build an HPC prototype system with Intel 3D XPoint technology
- Develop tools and systemware to support application development
- Design scheduler startegies that take NVRAM into account
- Explore how to best use this technology in I/O servers

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

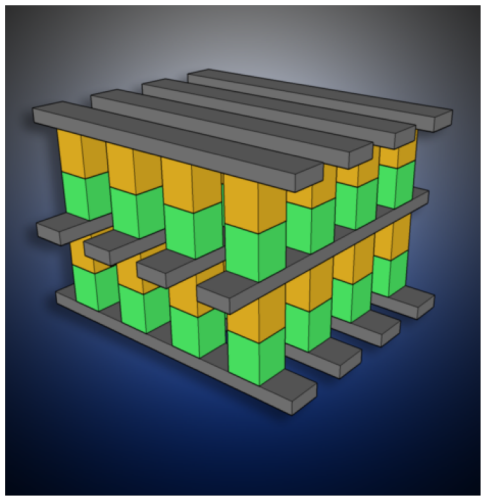
NVRAM Intel 3D XPoint



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

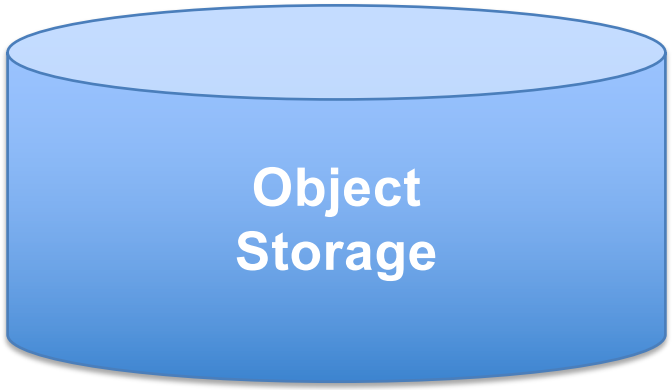
Key Point: High Density at very **low latency**,
Byte Addressable!

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



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

MARS

MARS Language

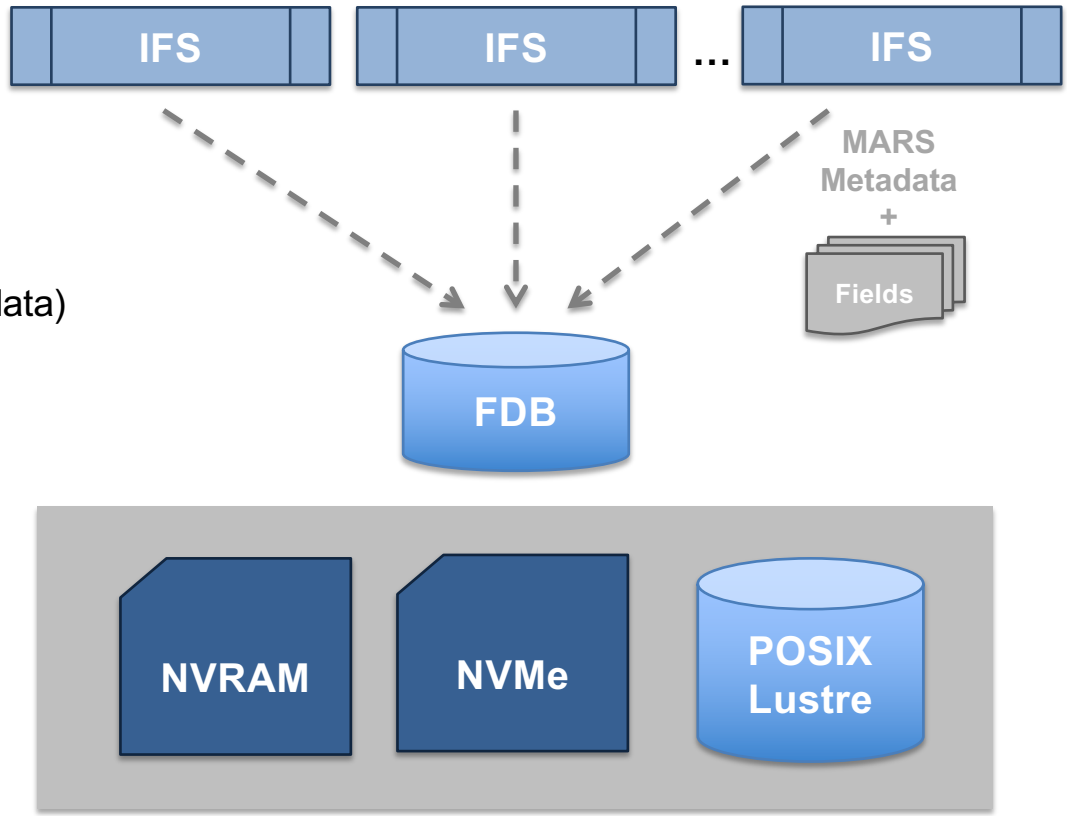
```
RETRIEVE,
    CLASS     = OD,
    TYPE      = FC,
    LEVTYPE   = PL,
    EXPVER    = 0001,
    STREAM    = OPER,
    PARAM     = Z/T,
    TIME      = 1200,
    LEVELIST  = 1000/500,
    DATE      = 20160517,
    STEP      = 12/24/36
```

```
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)

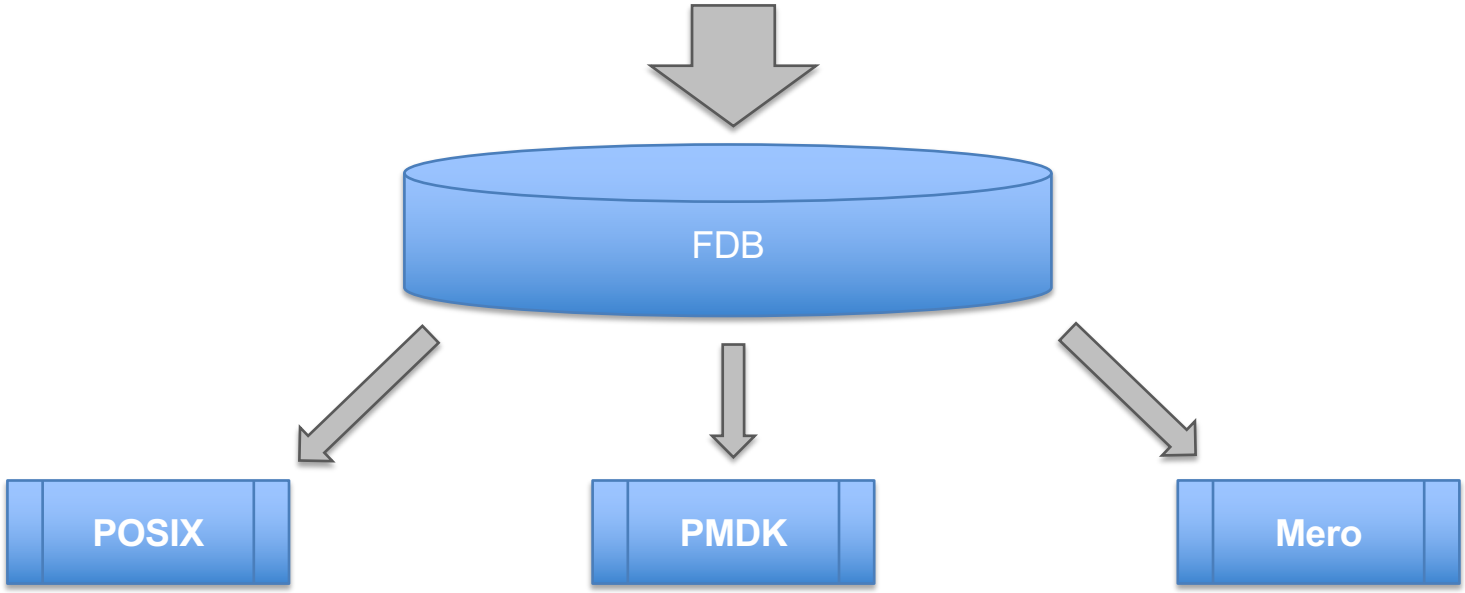
- Domain specific (NWP) object store
- Transactional, No synchronization
- Key-value store
  - Keys are scientific meta-data (MARS Metadata)
  - Values are byte streams (GRIB)



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

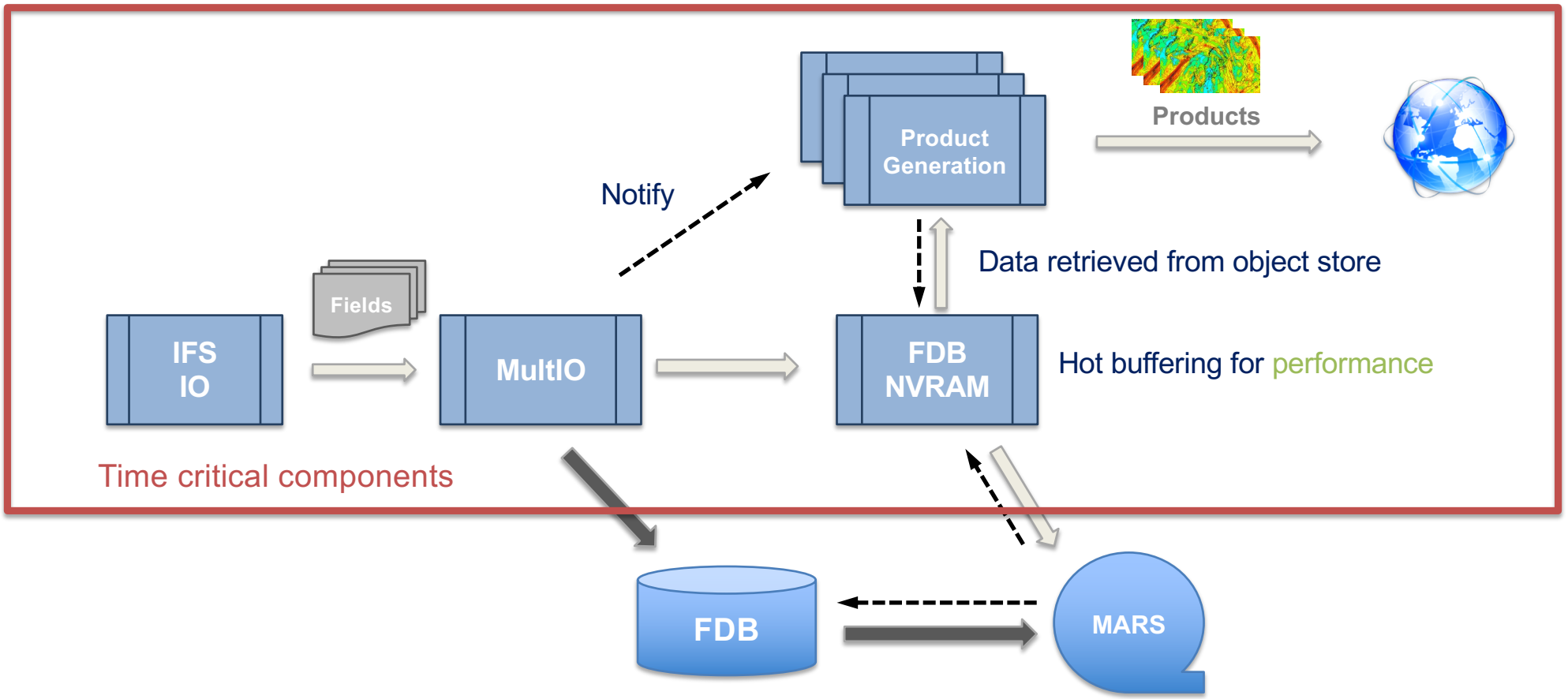
- Supports wild card searches, ranges, data conversion, etc…

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



New **technologies** bring
new **opportunities** to adapt data workflows

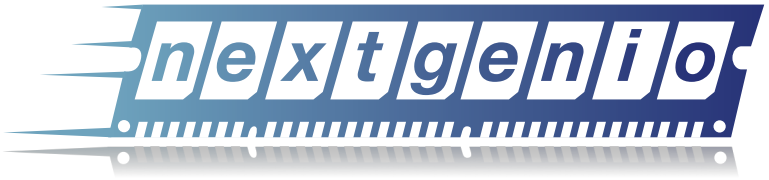
Summary : Overall Infrastructure Plan



Some Results with NVMe

(while we wait for final NVDIMM hardware)

Preliminary numbers – fresh out of the oven …



Two target nodes, with spinning disks (network) attached.

Connected to test cluster via dual 10 Gbps ethernet

Processors	Nodes	Fields	Data [GiB]	Aggregate Fields per second	Aggregate Rate [MiB / s]	Server side Per-process Rate [MiB / s]
1	1	12600	38.51	59.83	187.23	**295.13**
2	2	25200	77.02	126.17	394.84	**392.66**
4	4	50400	154.03	196.80	615.90	**220.70**
8	8	100800	308.06	226.26	708.10	**75.81**
16	16	201600	616.13	345.17	1080.22	**43.73**
32	32	403200	1232.25	331.64	1037.89	**22.46**
64	32	806400	2464.51	316.28	989.80	**9.81**
128	32	806400	2464.51	295.04	923.34	**5.07**
256	32	752640	2300.21	292.33	914.86	**2.66**

Preliminary numbers – fresh out of the oven … (2)

Four target nodes, with NVMe SSDs.

Connected to test cluster via dual 10 Gbps ethernet – but further from the cluster

Processors	Nodes	Fields	Data [GiB]	Aggregate Fields per second	Aggregate Rate [MiB / s]	Server side Per-process Rate [MiB / s]
1	1	12600	38.51	57.14	178.82	**556.48**
2	2	25200	77.02	112.94	353.46	**505.27**
4	4	50400	154.03	212.67	665.55	**493.11**
8	8	100800	308.06	212.29	664.37	**528.82**
16	16	201600	616.13	217.85	681.76	**549.71**
32	32	403200	1232.25	182.42	570.87	**558.43**
64	32	806400	2464.51	204.59	640.27	**561.34**
128	32	806400	2464.51	196.91	616.24	**557.43**
256	32	752640	2300.21	188.48	589.85	**549.81**

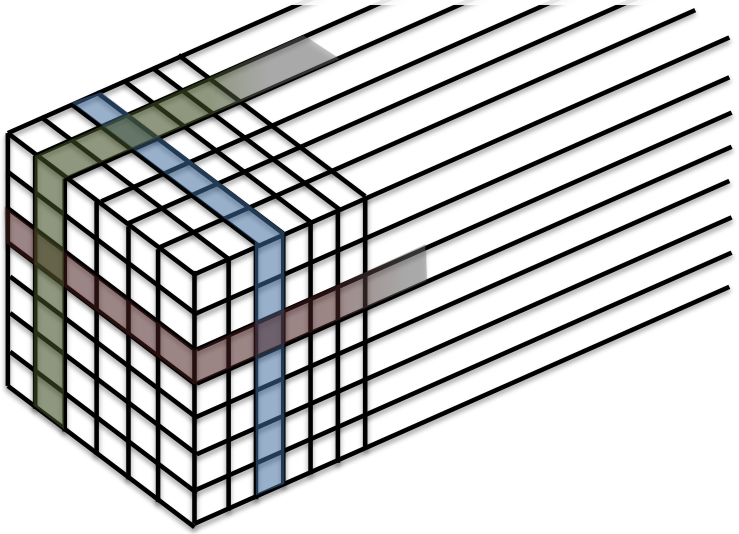
Where are we going?

Impacts of NVRAM on Data Access

Byte Addressable Hypercubes (6D)

- Longitude (3600)
- Latitude (1800)
- Variables (~1000)
  - Atmospheric levels (~ 8 x 100)
  - Physical parameters (~200)
- Time steps (~100)
- Probabilistic perturbations (50)

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



@ double precision

- 16km **80 TiB**
- 9km **235 TiB** **@ 3TiB / node = ~80 nodes**
- 5km **583 TiB**

Novel Data Flows – Data Centric Computing

Producer

Cloud

Consumer

- Bring **users** to the data
- Use data while it is **hot**
- Access using **scientifically meaningful metadata**

NVRAM (?)
Data Hypercubes

Archive

PFS

HDD

Tape

MARS

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

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

Thanks for your attention

Questions?