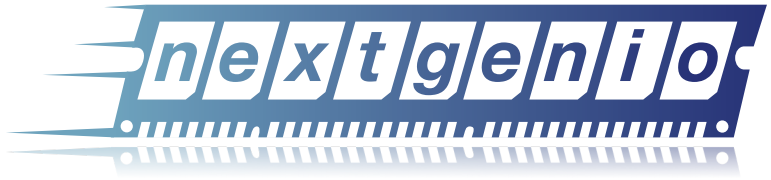


ECMWF's IO and Storage Challenges in the path to Exascale Numerical Weather Prediction

Tiago Quintino, Simon Smart, Baudouin Raoult

ECMWF

tiago.quintino@ecmwf.int

PASC 2019, Zurich

12-14th June 2019

ECMWF

ECMWF's Forecasting Systems

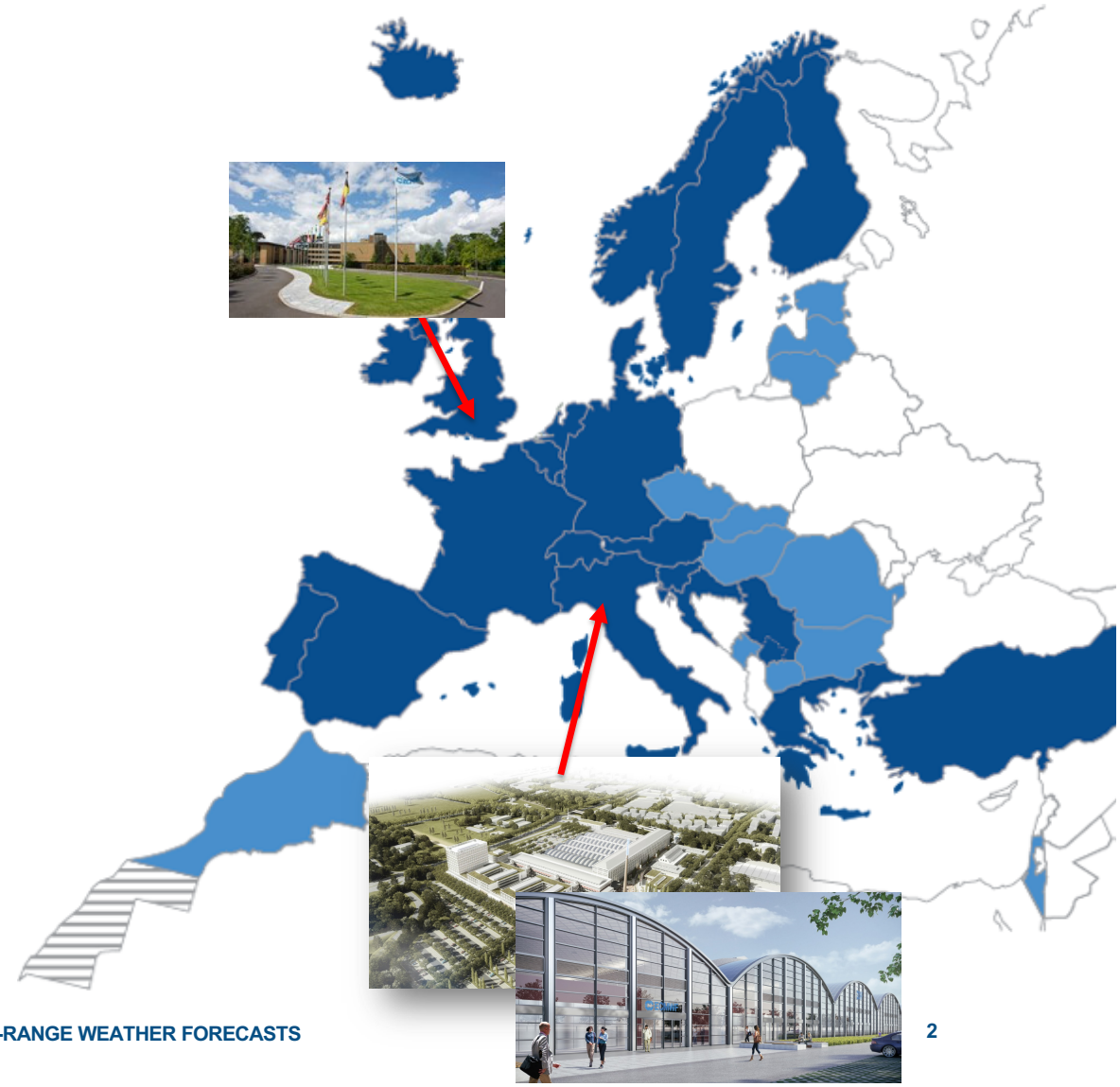
What do we do?

Operations – **Time Critical**

- 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**

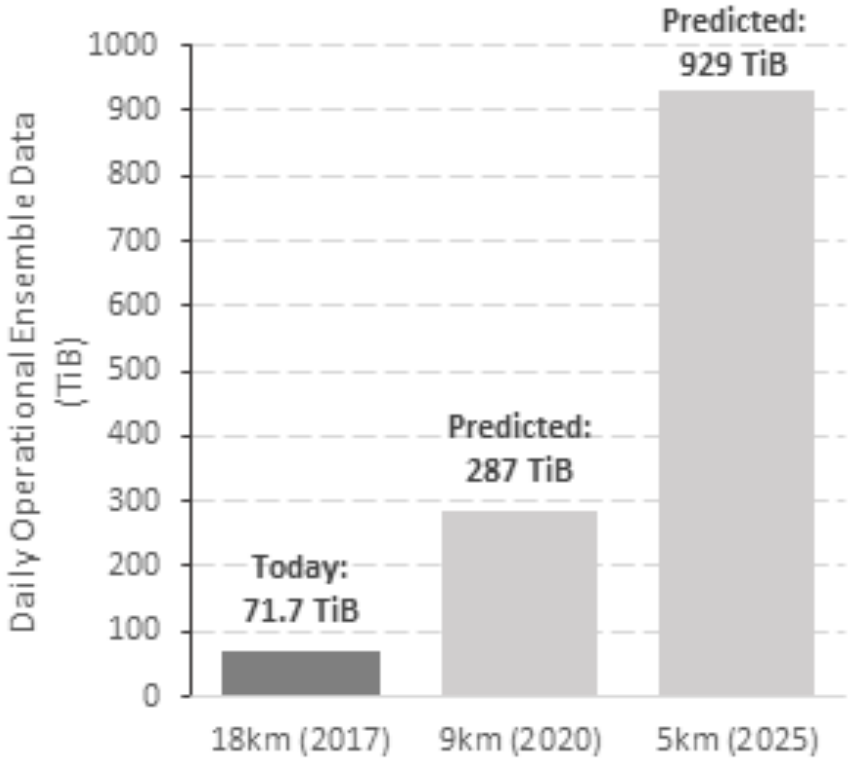
- Experiments to improving our models
- Reforecasts, Climate reanalysis, etc



Challenges

Data Growth – History and Projections

Historical Growth of Generated Products



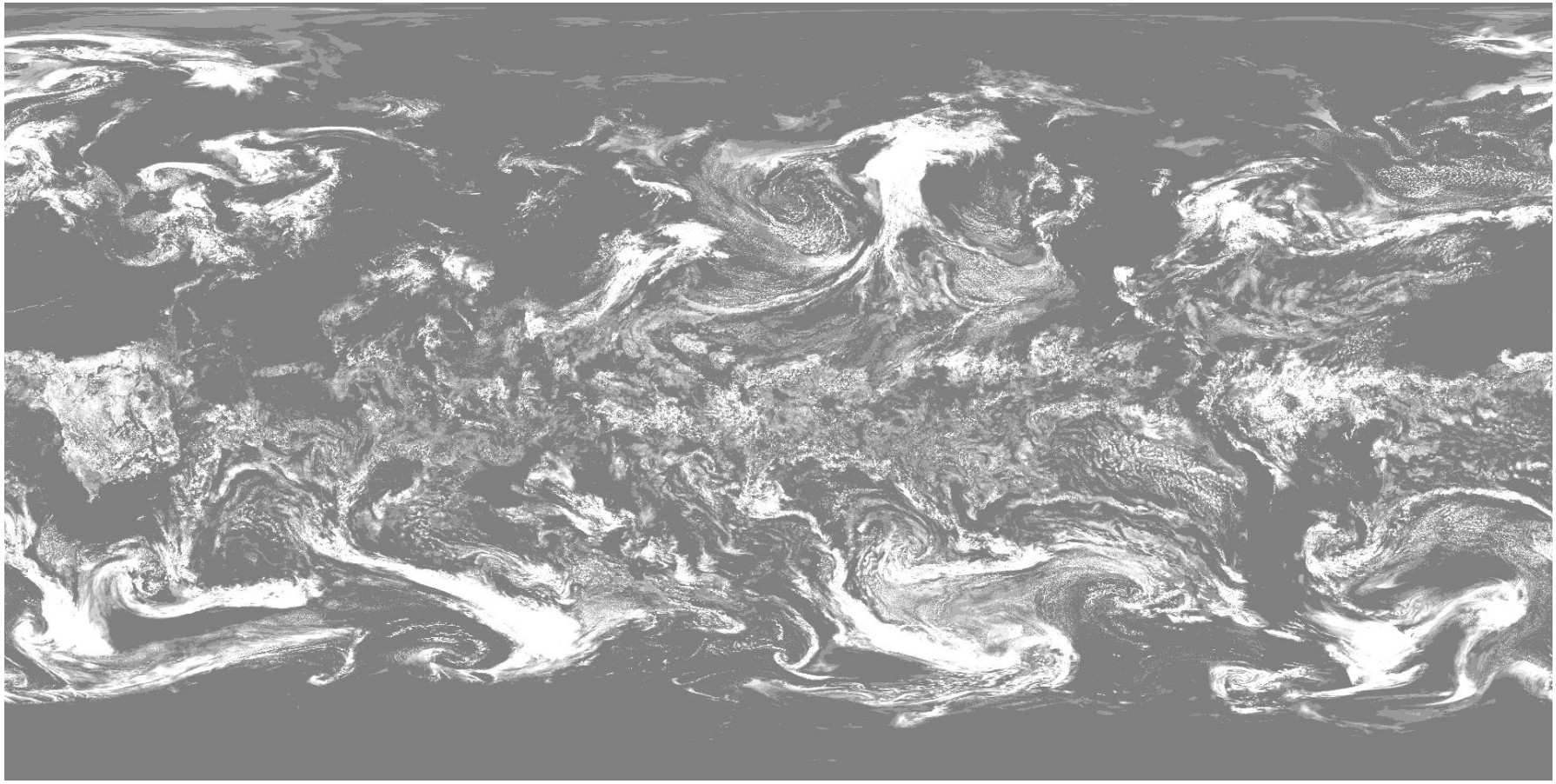
Model Output Projected Growth

History and Future of Resolution Upgrades

Resolution	Grid size	Grid Points	Field Size (in memory)	Vertical Levels	YEAR
T319	62.5 km	204 k	1.6 MB	L31	1998
T511	39 km	524 k	4 MB	L60	2000
T799	25 km	1.2 M	9.6 MB	L91	2006
T1279	16 km	2.1 M	16.8 MB	L91	2010
Tco1279	**9 km**	**6.6 M**	**50.4 MB**	**L137**	**2016**
Tco1999	5 km	16.1 M	122.6 MB	**L160**	***2025***
Tco3999	2.5 km	64 M	490 MB		
Tco7999	*1.25 km*	*256 M*	***1909 MB***	***L180***	***2030***

ECMWF

TCo7999 (~1.25km) 256 Megapixel



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

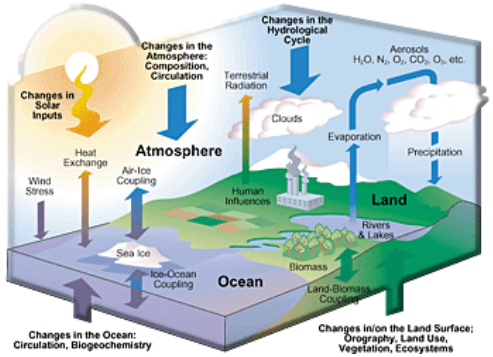
Multiple dimensions

➔Reliability

Ensembles

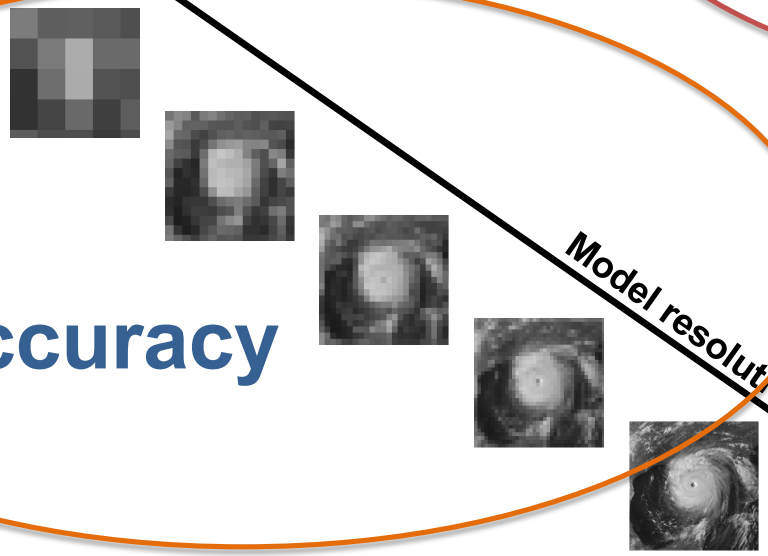
Traditional weather science domain

Model complexity



➔Range

Traditional climate science domain



➔Accuracy

Model resolution

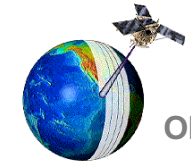
Today: it needs high-resolution, 'Earth system' model ensembles to perform at all scales!

How large is a 1.25 km ensemble forecast?

- 50 member ensemble forecast
- *Compressed* GRIB2 data @ 16bit & 24bit
- @ 9km O1280 — 21 TiB
- Resolution @ 5km O1280 → O1999 — x 3.3
- Upgrade levels 137 → 200 — x 1.46
- Resolution @ 2.5km O1999 → O3999 — x 3.3
- Resolution @ 1.25km O3999 → O7999 — x 3.3

21 TiB x 52.5 = 1102 TiB

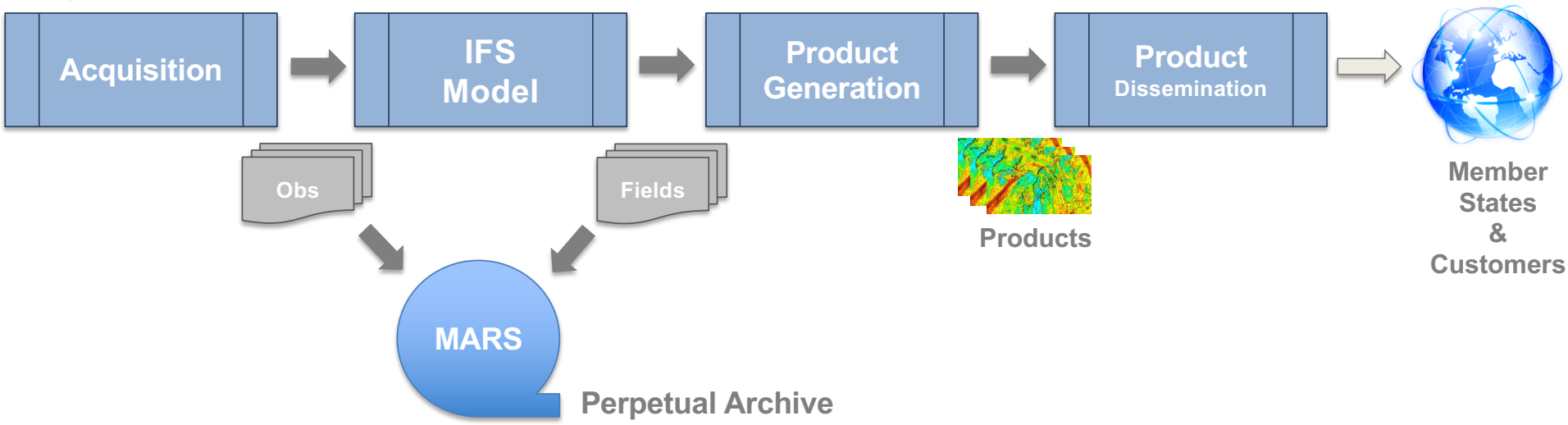
ECMWF's (Simplified) Operational Workflow



Global Observations

Data Workflow

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

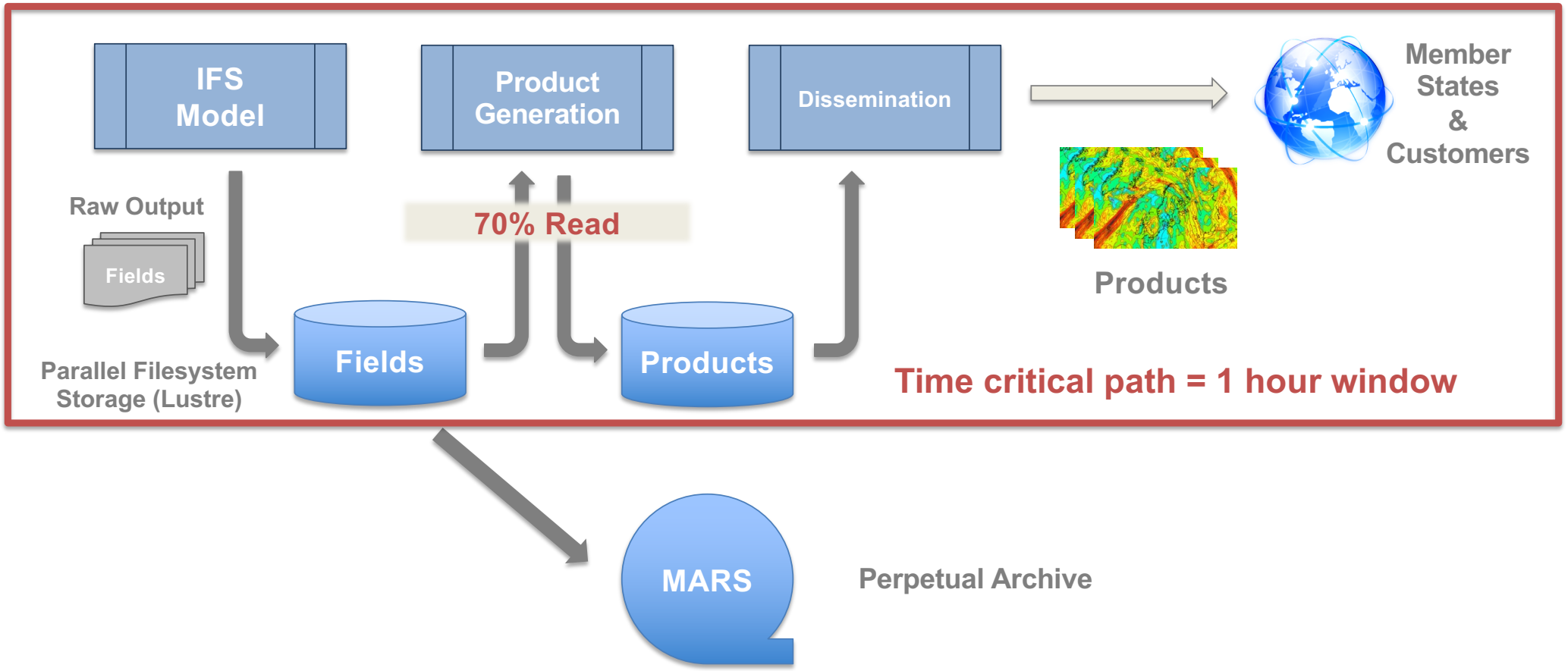


Effects of Product Generation

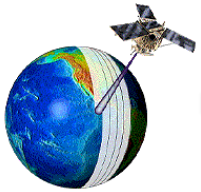
	IFS Model	Model + I/O	Model + I/O + PGen
Nodes	2440	2776	2926
Run time [s]	5765	6749	7260
Relative	-	**+ 17%**	**+ 26%**

9Km 50 member ensemble
Broadwell nodes 2x18 cores
Cray XC40 Aries interconnect
Lustre FS IOR 90GiB/s

ECMWF's Production Workflow



Storage View of Workflow



Observations

Acquisition

Product Dissemination

Member States & Customers

Acquire

Disseminate

Produce

IFS Model

Parallel FS

Archive

Modify

Product Generation

MARS

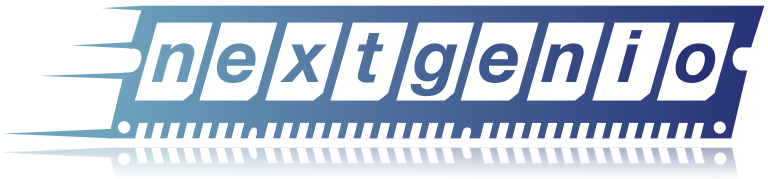
Perpetual Archive

Data is Central!

What have we done so far?

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

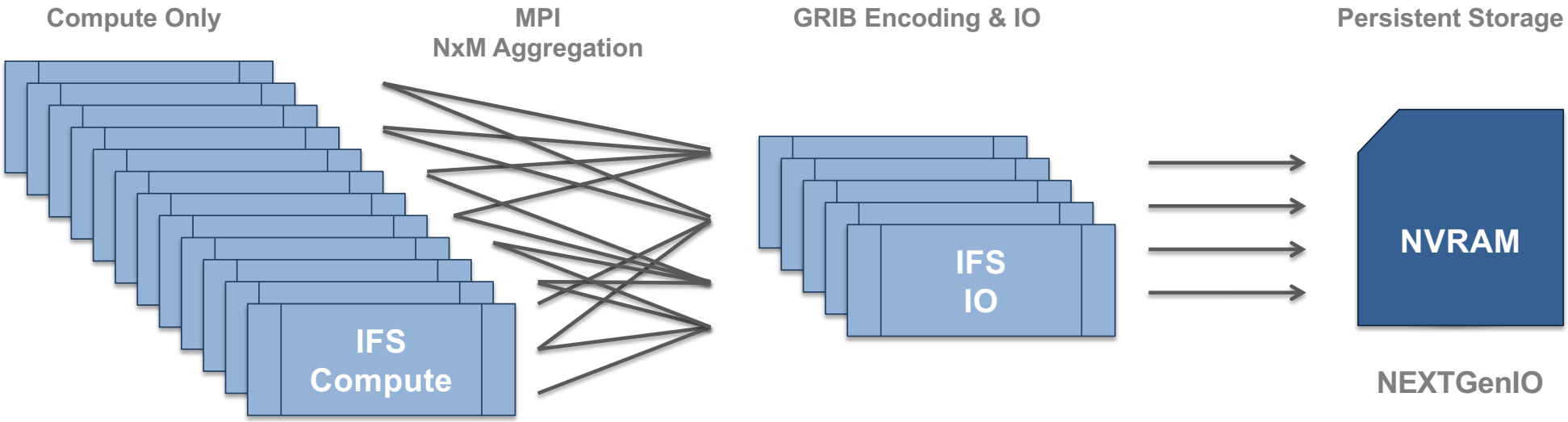
ECMWF Tasks

- 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

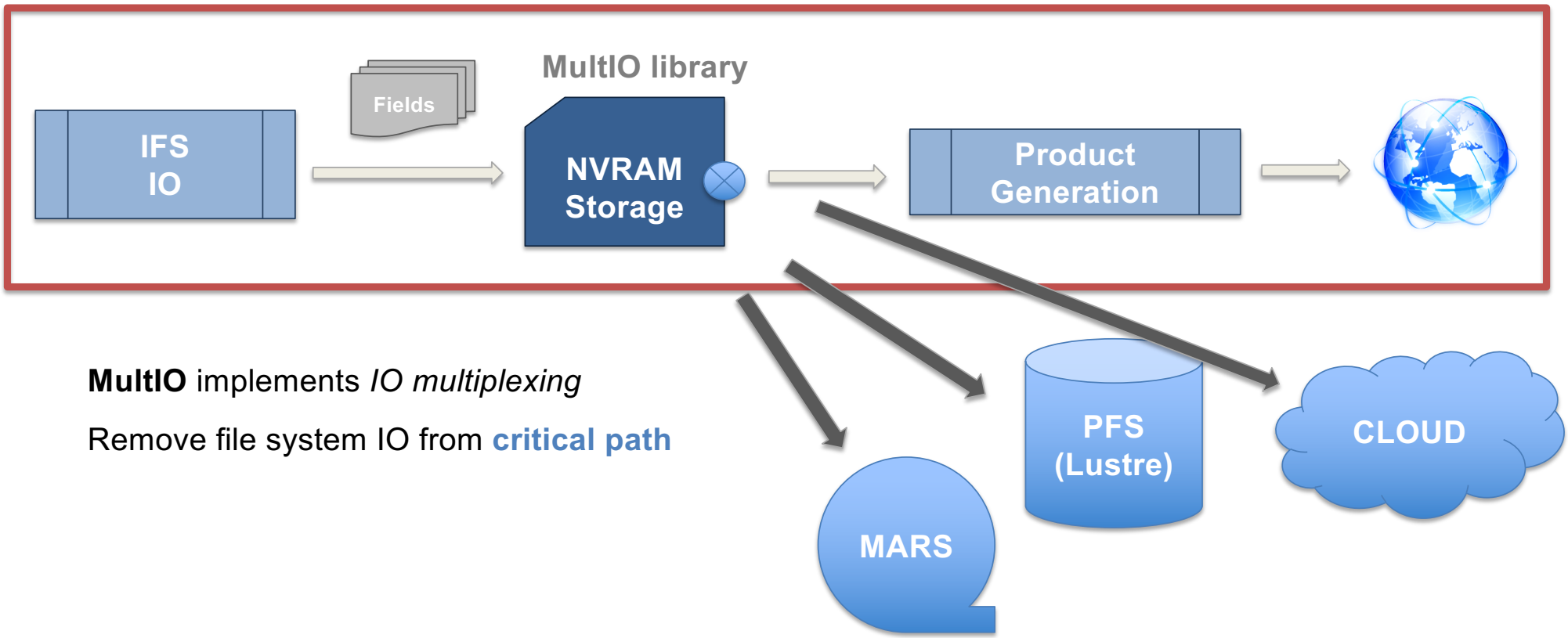
http://www.nextgenio.eu - EU funded H2020 project, runs 2015-2018

IFS IO Server

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



Streaming Model Output to Product Generation



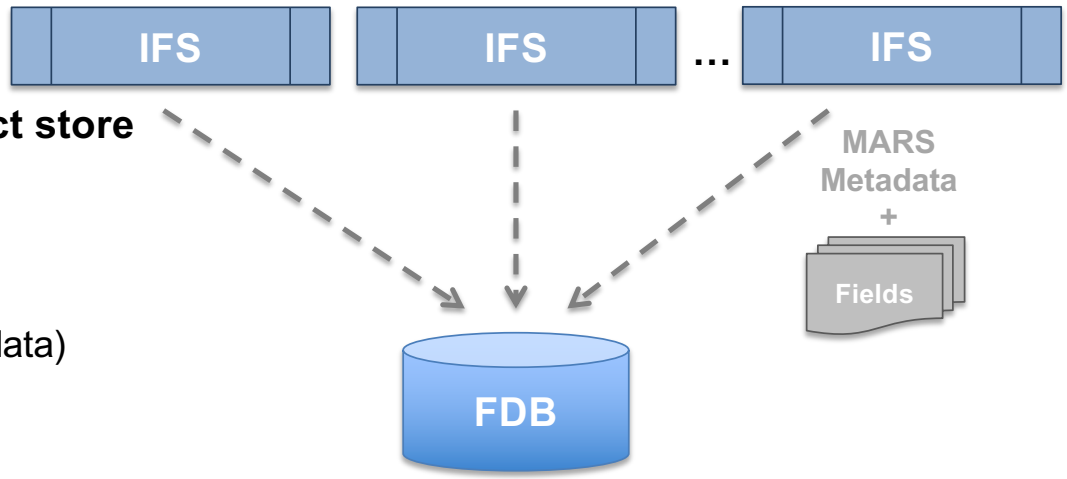
MultIO implements *IO multiplexing*

Remove file system IO from **critical path**

How to store all model output in NVRAM?

FDB (version 5)

- **Domain specific (NWP) Distributed 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

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

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

Paper Presentation:

A High-Performance Distributed Object-Store for Exascale Numerical Weather Prediction and Climate

Simon Smart, T. Quintino, B. Raoult, PASC'2019

FDB5

- ***Full rewrite***
- ***All workflows use this storage software***
- ***3 years preparing for this…***
- ***Roll out to operations 3 days ago (2019.06.11)***
  - ***As we were flying into Zurich*** ✈️

Mini-symposium Presentation (MS54 - The Exabyte Data Challenge)

ECMWF's Extreme Data Challenges Towards a Exascale Weather Forecasting System

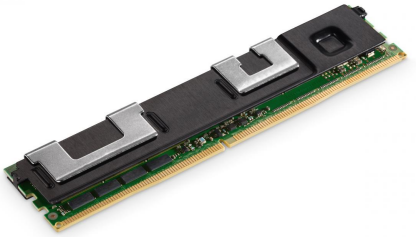
Tiago Quintino, S. Smart, J. hawkes, B. Raoult, PASC'2019

Product Generation (PGen)

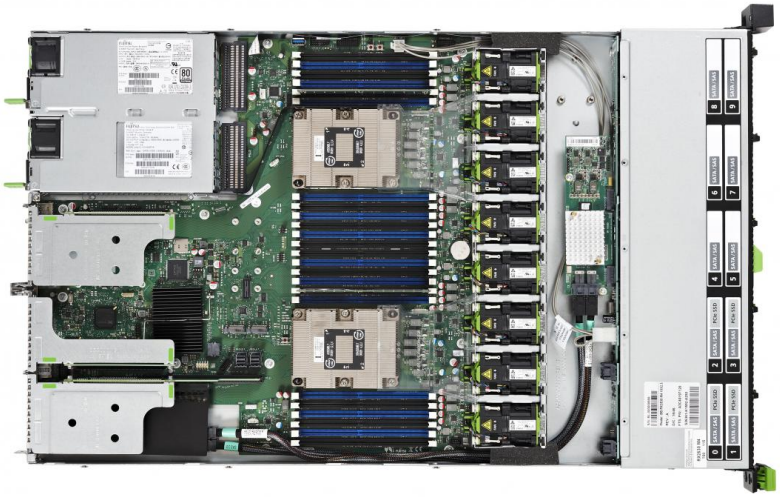
- ***Full rewrite***
- ***All products are computed from this software***
- ***3 years preparing for this…***
- ***Roll out to operations during last 6 months***

NextGenIO Prototype

- Read all @ www.nextgenio.eu
- Development of an HPC node by **with Intel 3D Xpoint**
- Dual-CPU Intel® Xeon® SP nodes
- OmniPath network
- 192GB DRAM
- **3TiB of NVRAM DIMMs**



- **Prototype system**
  - **34 compute nodes**
  - **Hosted @ EPCC, Edinburgh**



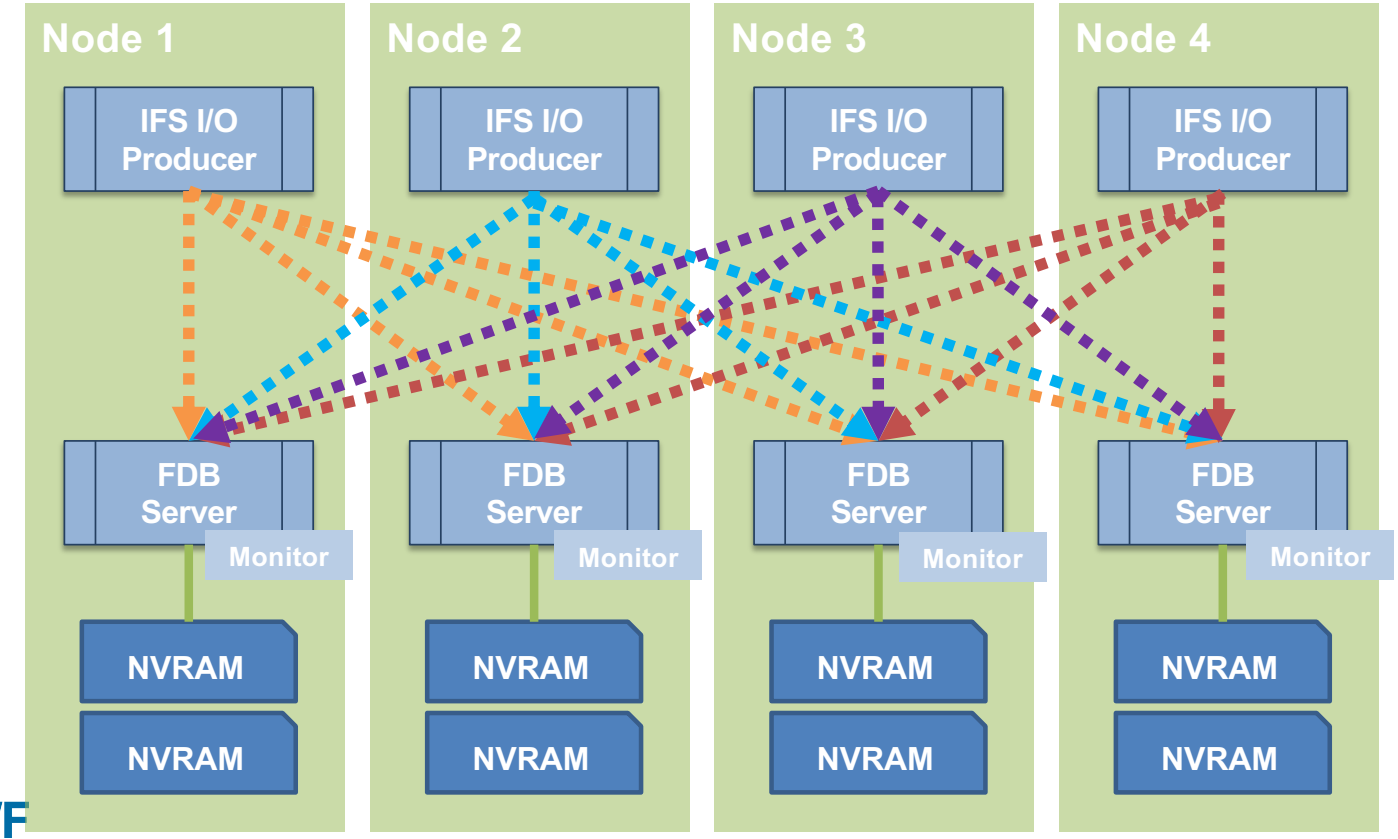
Running the forecast model

	Model + I/O	Model + I/O + PGen
Run time (Lustre) [s]	1793	1928
Run time (Distributed) [s]	1610	1599

NextGenIO prototype. 32 nodes
Intel OmniPath2 interconnect
IFS 6 ensemble members @ O640 (16km)

Data Flow Schematic

- All I/O operations are asynchronous, so computation can continue
- Distributed to all servers using a ***Distributed Hash***, *so no synchronisation needed*



ECMWF

Preliminary Results

ECMWF Operational Filesystem

- Sonexion snx11061
- OST Nodes: **288**
- 20TiB per node (10 disks)
- **4PiB** capacity
- Measured 165GiB/s (IOR)

- Sustained IFS runs: R 22.4 GiB/s + W 22.0 GiB/s = **44.4 GiB/s** ***application data***

NEXTGenIO + Distributed FDB

- Nodes: **34**
- 3TiB per node (12 DIMMs)
- **108 TiB** capacity

- Not yet optimised!
- Measured **sustained** **40 GiB/s R + 40 GiB/s W** ***application data***

ECMWF

Can we handle the 1.25 km ensemble forecast?

- 50 member ensemble forecast
- *Compressed* GRIB2 data @ 16bit & 24bit
- @ 1.25km 7999 1102 TiB
- Required to read 70% x 1.70
- @ 1.25km 7999 1874 TiB
- Time to solution 1 hour 1874 TiB / 3600 = 533 GiB/s
- NextGenIO performance 80 GiB/s
- Required Nb Prototypes 533 / 80 = x 6.7

(by 2030)

Looking ahead

Impacts of NVRAM on Data Access

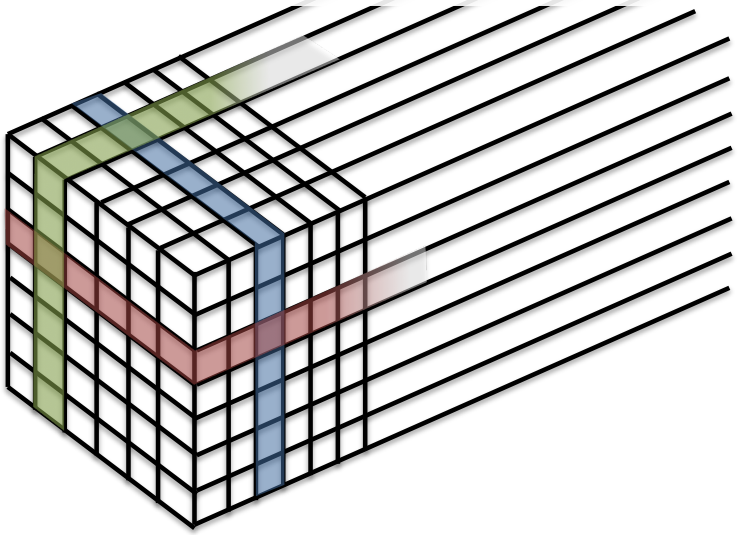
Byte Addressable Hypercubes

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

@ double precision

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

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



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

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

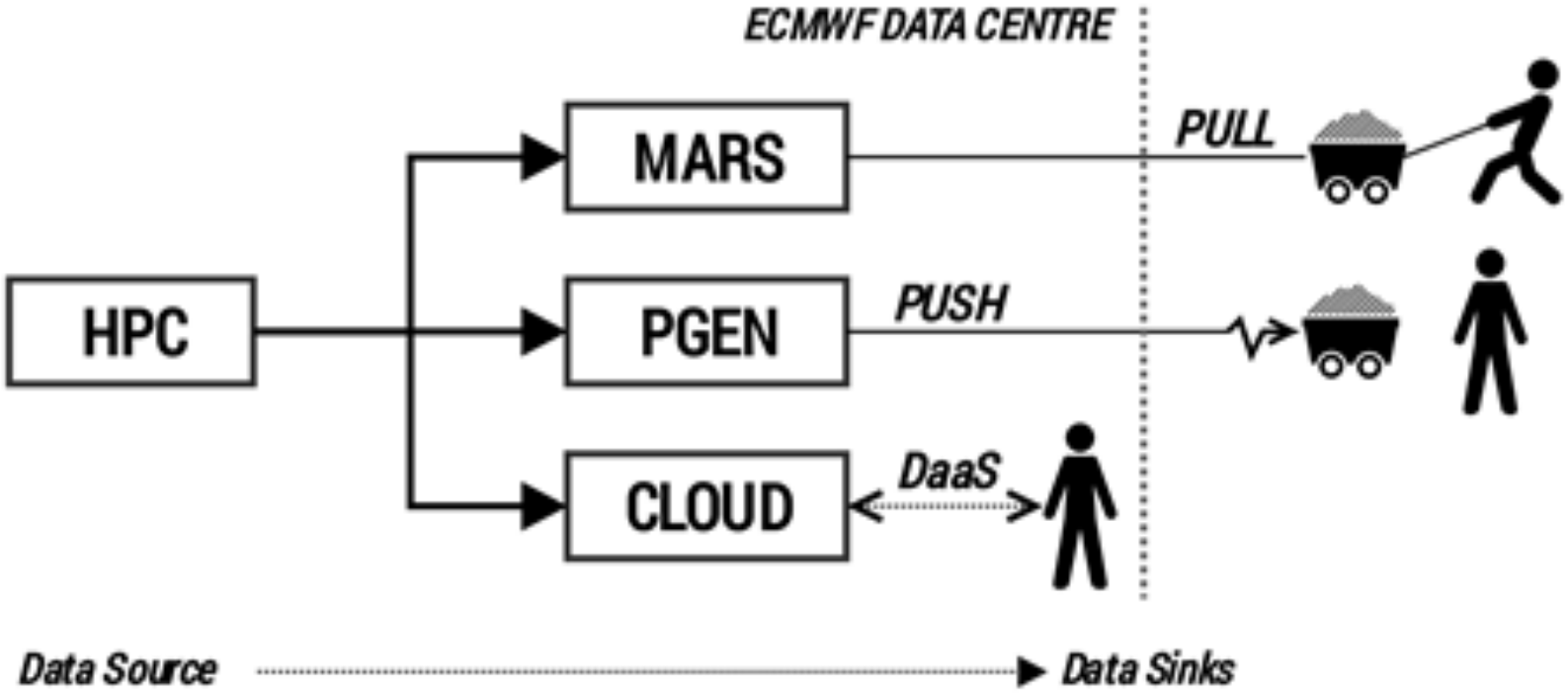
PFS

HDD

Tape

MARS

Novel Data Flows – Multiple Pathways to Serve Data

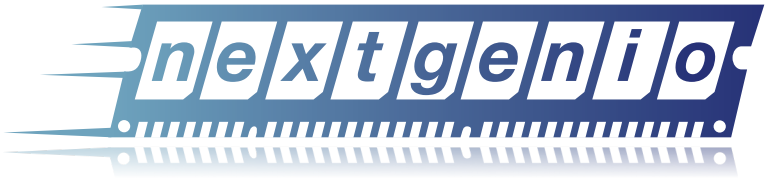


Messages To Take Home

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

*New technologies in the **horizon***
***but** will change the way we use and store data*

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



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?