Running ECMWF’s Workflow on the NextGenIO Prototype

Tiago Quintino, Simon Smart, Antonino Bonanni, Olivier Iffrig, James Hawkes, Domokos Sarmani, Baudouin Raoult

ECMWF

simon.smart@ecmwf.int, tiago.quintino@ecmwf.int

ECMWF’s Forecasting Systems

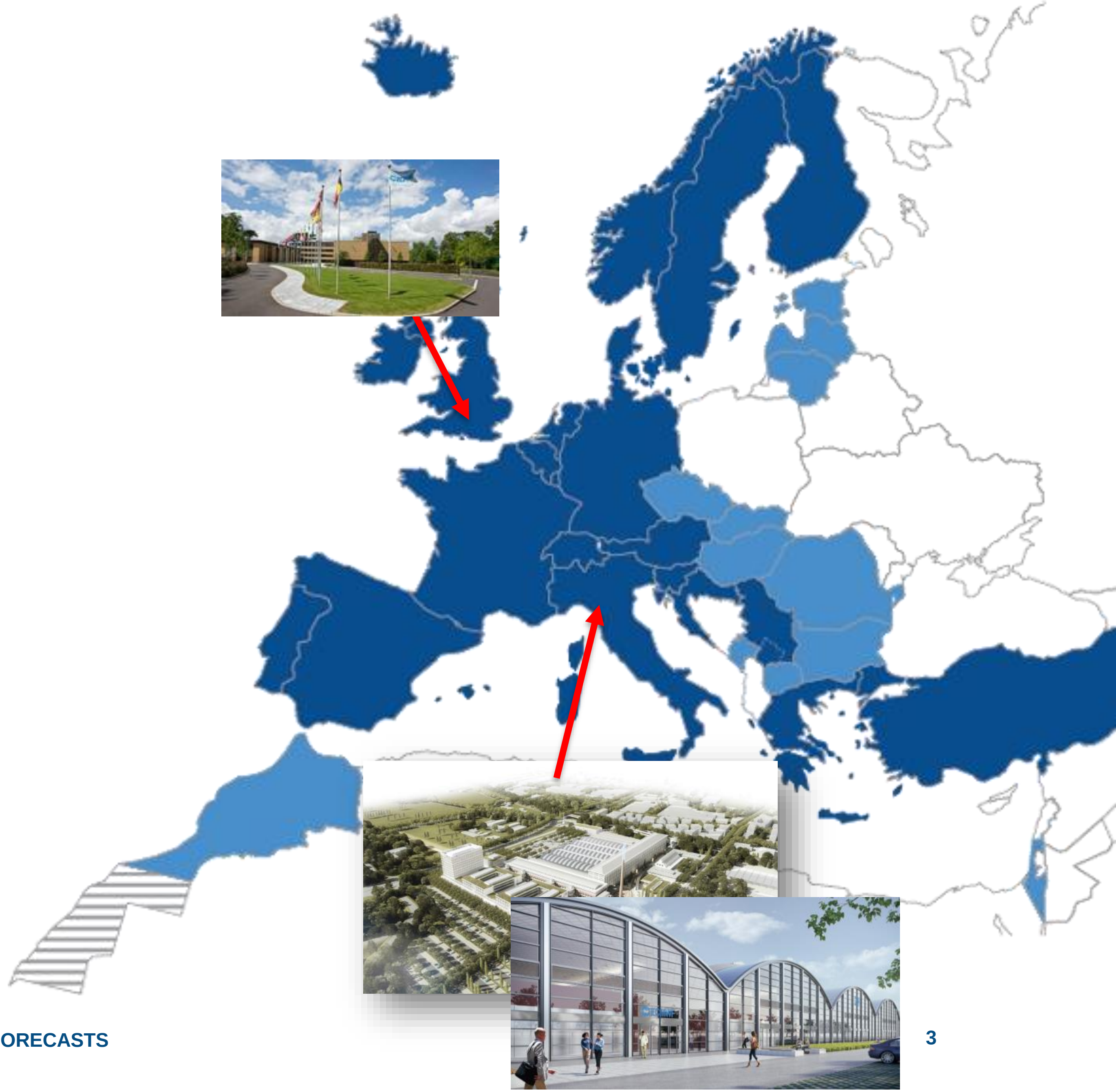
What do we do?

Operations – **Time Critical**

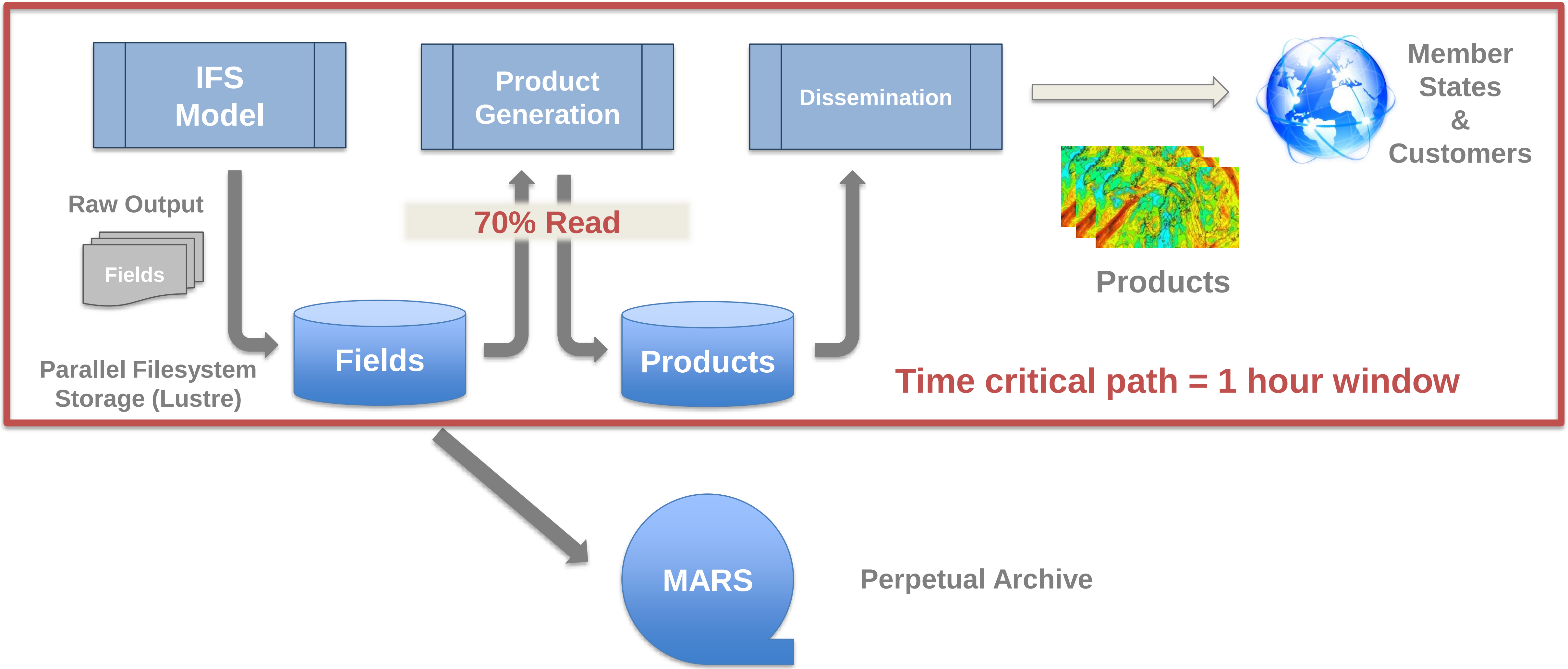
- HRES 0-10 day, 00Z+12Z
  - O1280 (9km) 137 levels
- ENS 0-15 day, 00Z+12Z
  - O640 (18km) 91 levels
- ENS extended 16-46 day, twice weekly
  - O320 (36km) 91 levels
- BC 06Z and 18Z
  - hourly post-processing 0-5 days

Research – **Non Time Critical**

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



ECMWF's Production Workflow

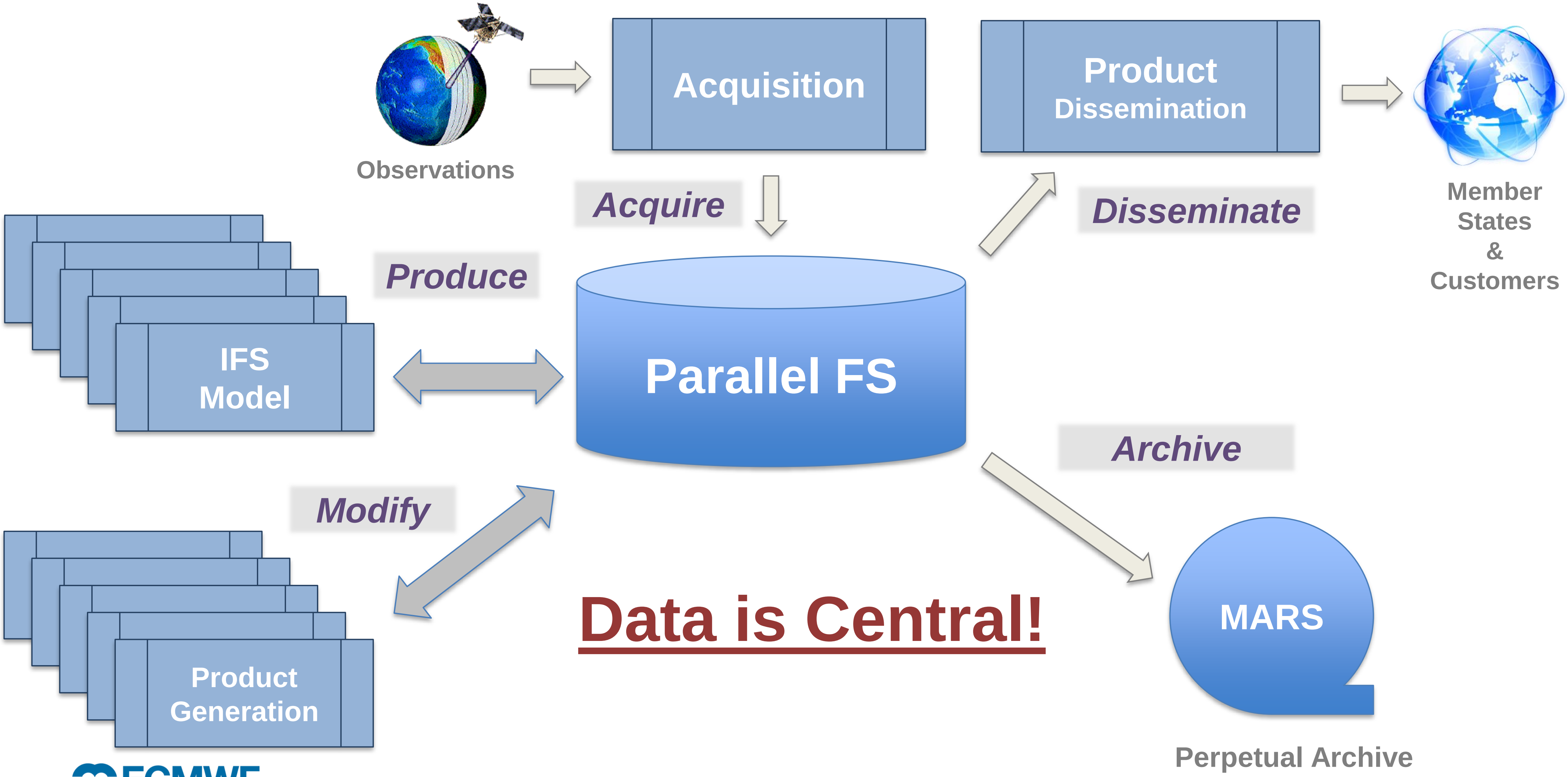


Effects of Product Generation

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

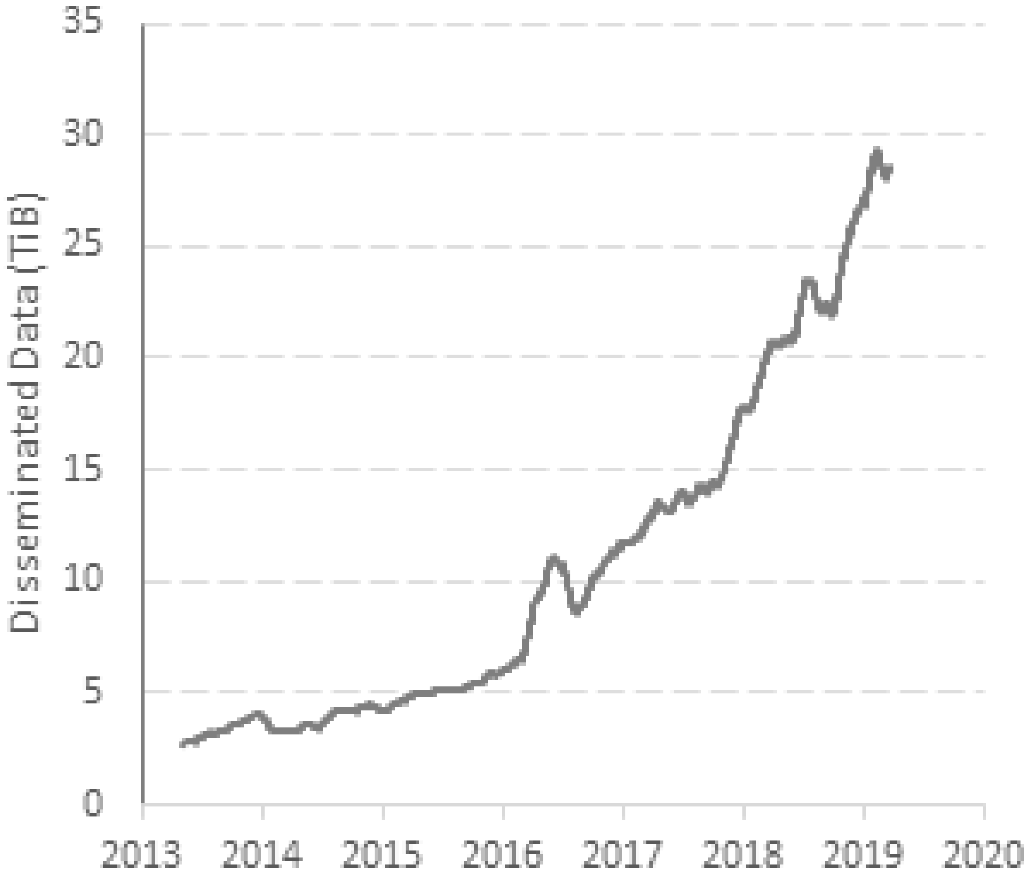
9Km 50 member ensemble
Broadwell nodes 2x18 cores
Cray XC40 Aries interconnect
Lustre FS IOR 90GiB/s

Storage View of Workflow

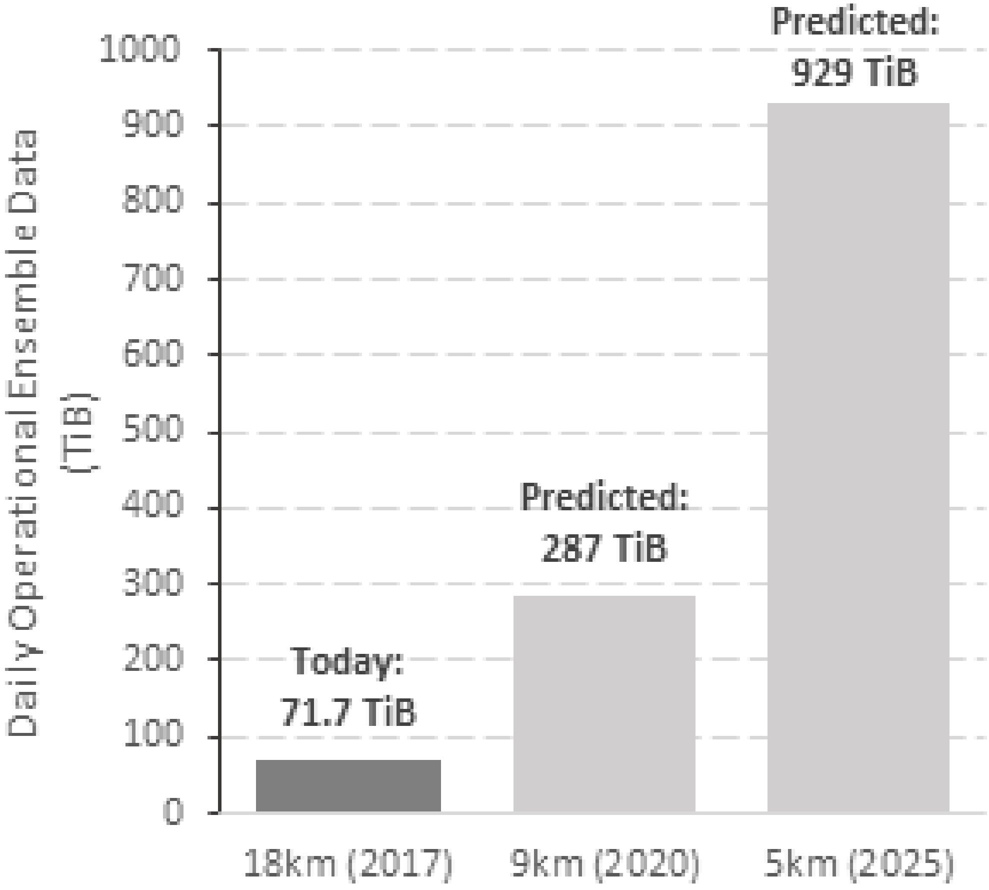


Data is Central!

Data Growth – History and Projections

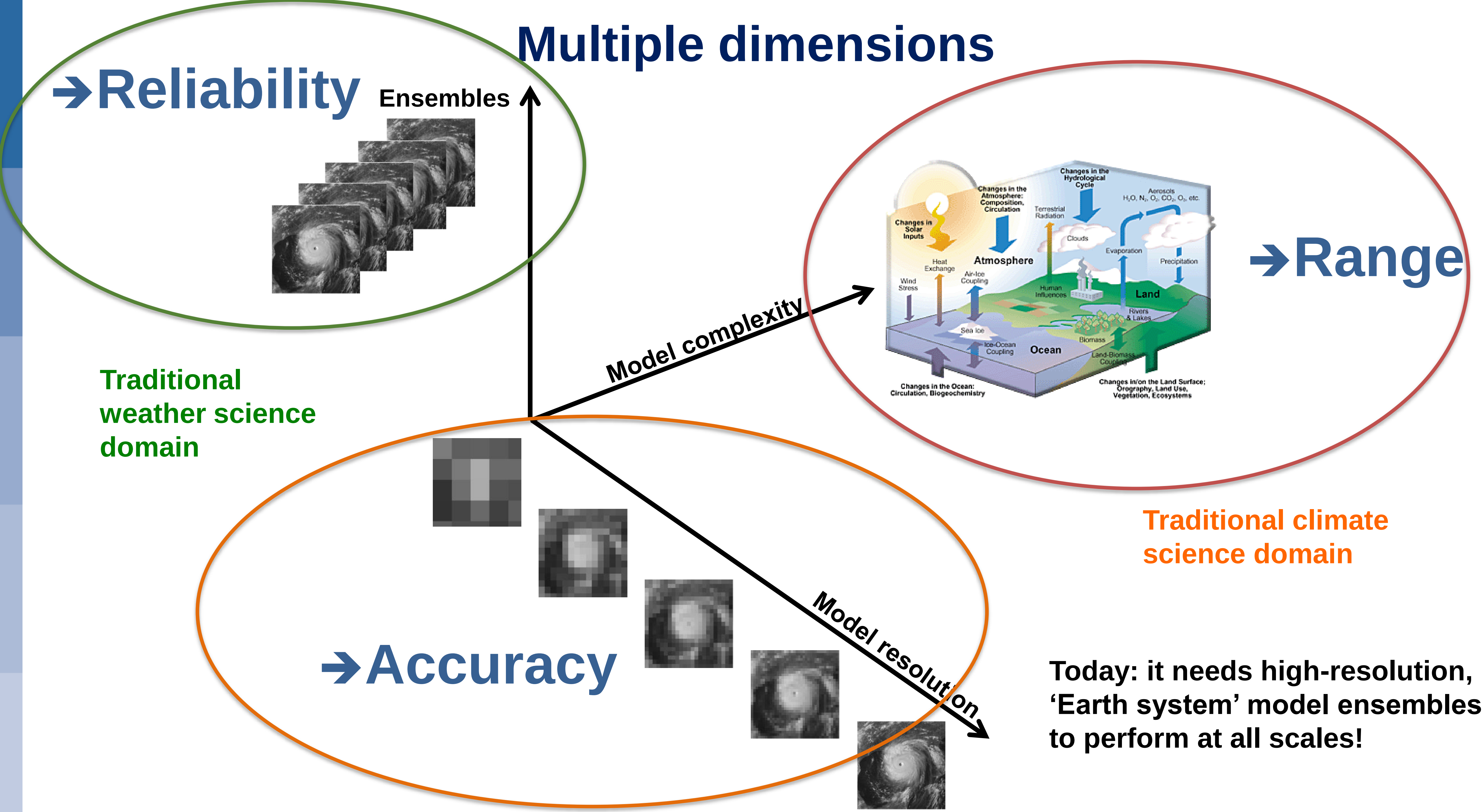


Historical Growth of Generated Products



Model Output Projected Growth

Multiple dimensions


Changes in Solar Inputs
Changes in the Atmosphere: Composition, Circulation
Changes in the Hydrological Cycle
Aerosols H2O, N2, O2, CO2, O3, etc.
Terrestrial Radiation
Clouds
Evaporation
Precipitation
Atmosphere
Heat Exchange
Wind Stress
Air-Ice Coupling
Human Influences
Land
Rivers & Lakes
Sea Ice
Biomass
Ice-Ocean Coupling
Ocean
Land-Biomass Coupling
Changes in the Ocean: Circulation, Biogeochemistry
Changes in/on the Land Surface; Orography, Land Use, Vegetation, Ecosystems

→Reliability

Ensembles

Model complexity

→Range

Model resolution

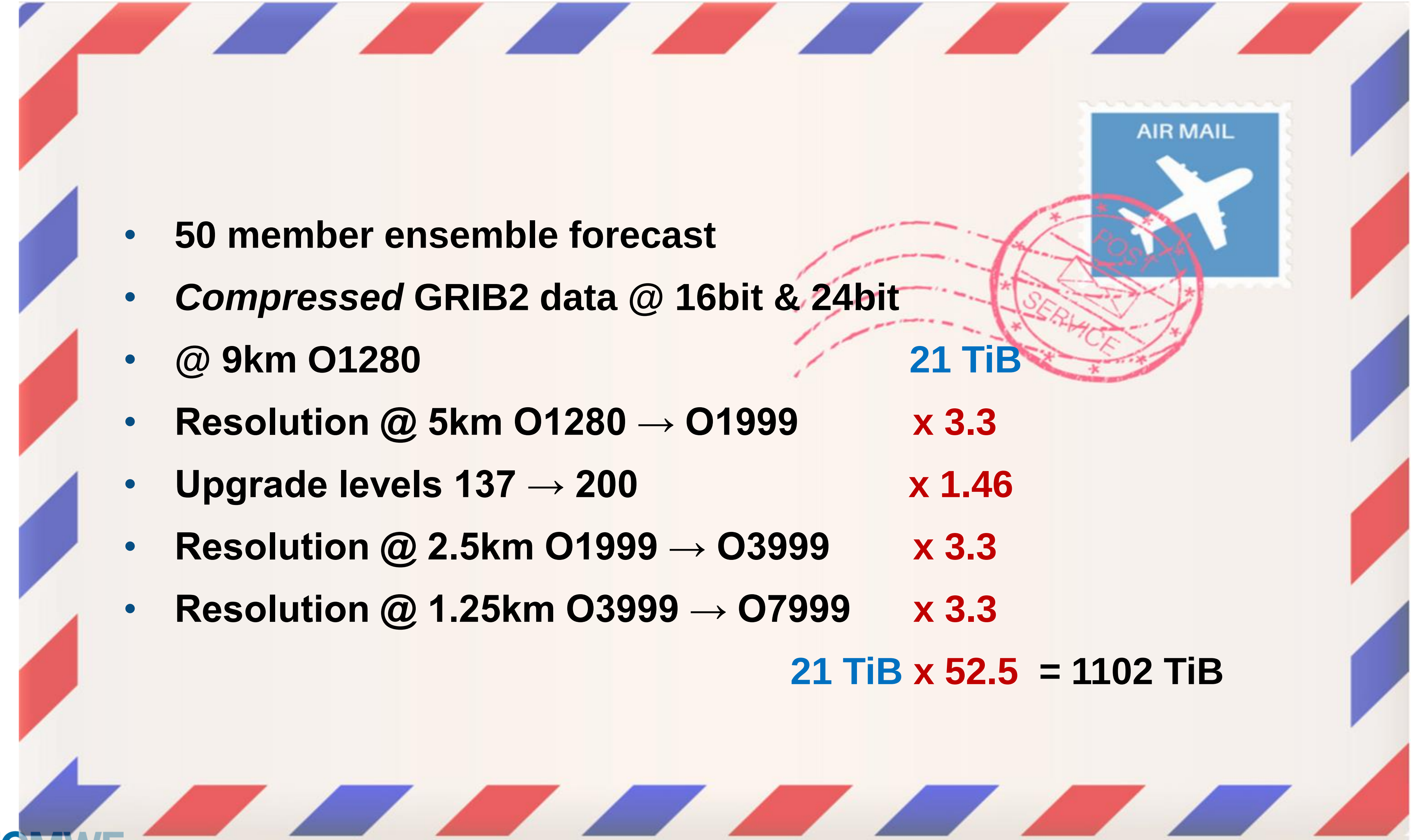
→Accuracy

Traditional weather science domain

Traditional climate science domain

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

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

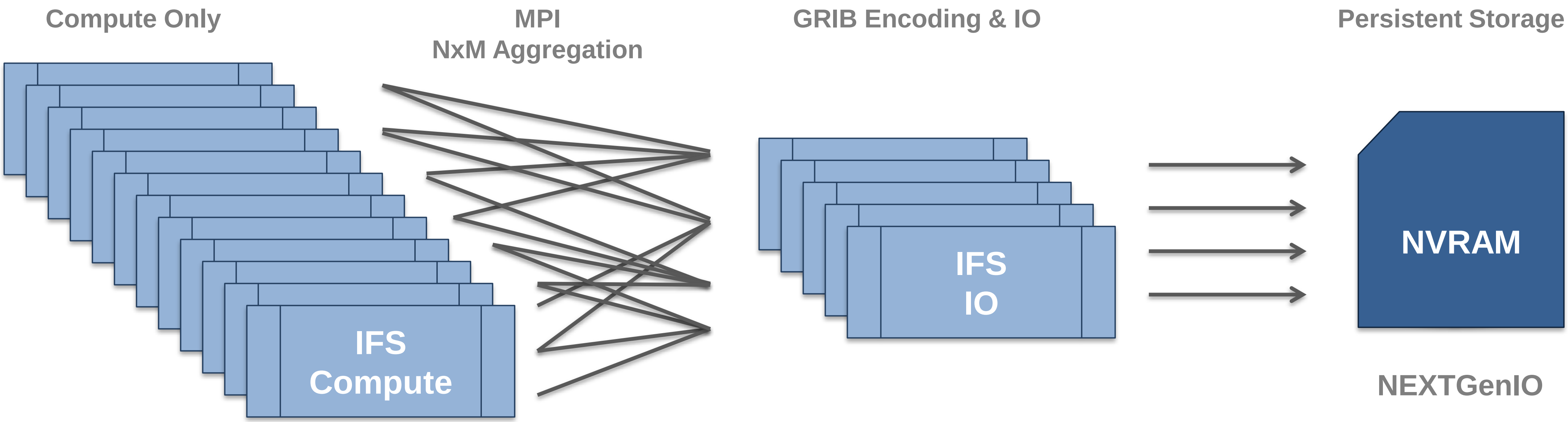
ECMWF Tasks

- 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

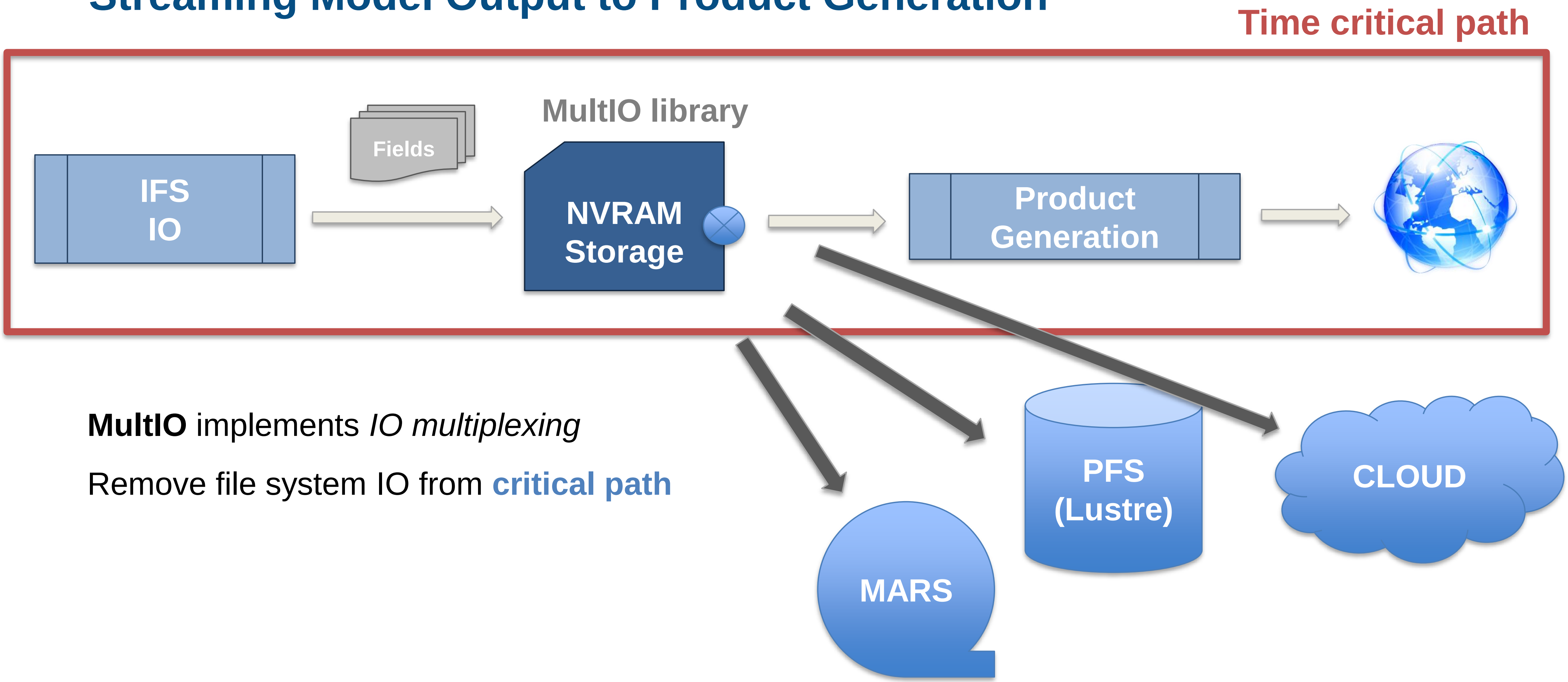
http://www.nextgenio.eu - EU funded H2020 project, runs 2015-2018

IFS IO Server

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



Streaming Model Output to Product Generation

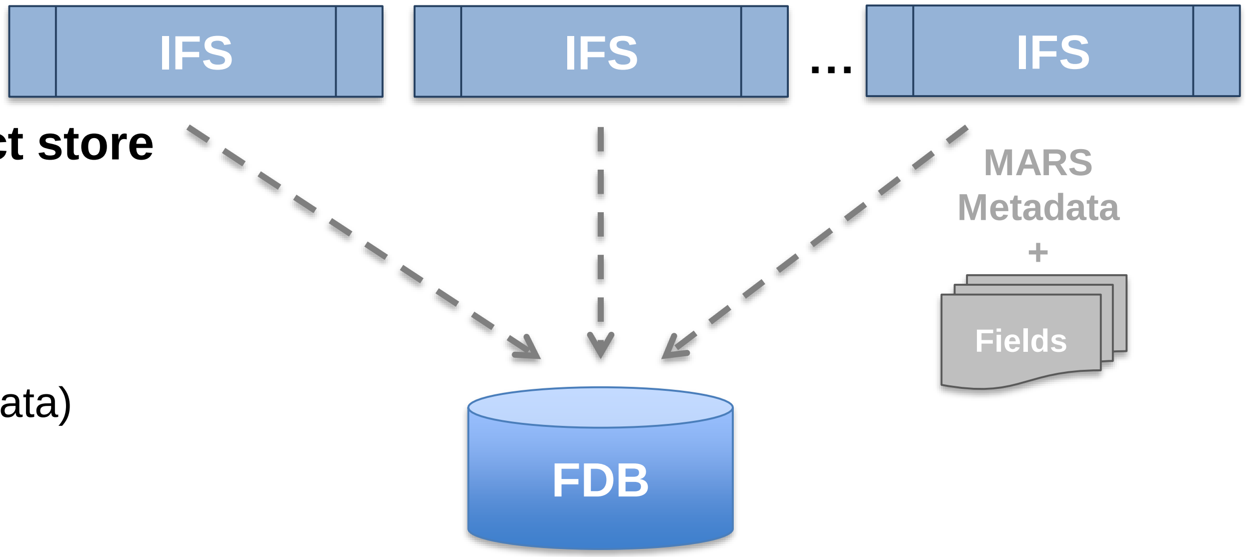


MultIO implements *IO multiplexing*

Remove file system IO from **critical path**

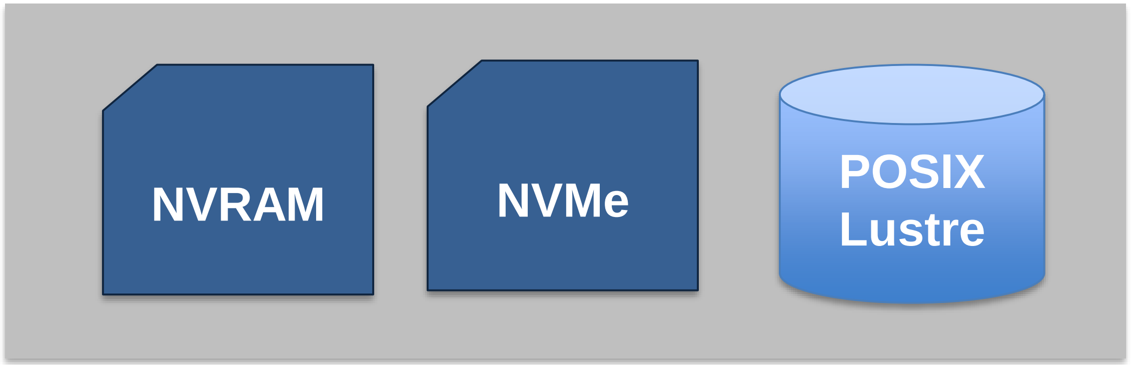
How to store all model output in NVRAM?

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,

Preliminary Results

ECMWF Operational Filesystem

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

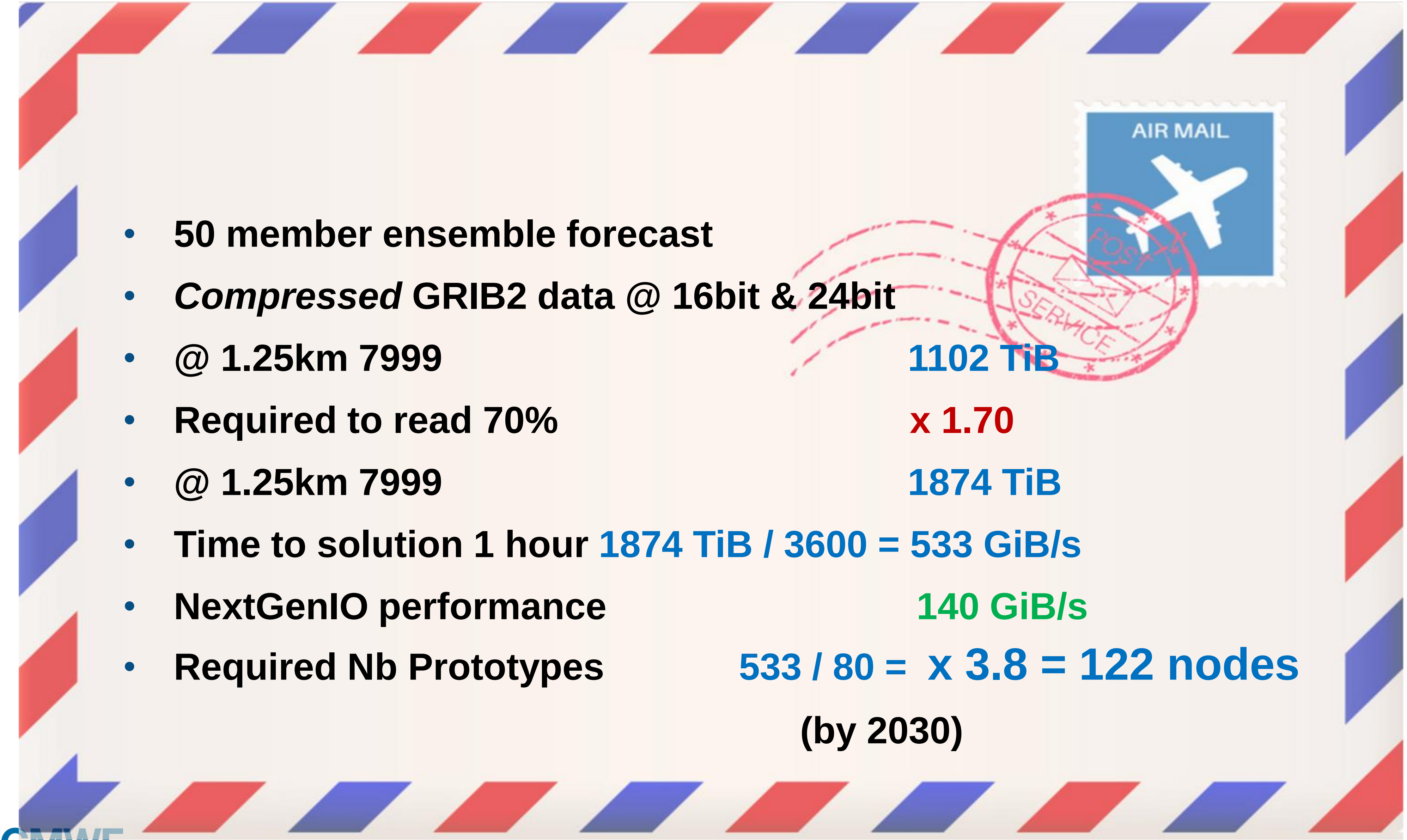
- Sustained IFS runs: R 22.4 GiB/s + W 22.0 GiB/s = **44.4 GiB/s** ***application data***

NEXTGenIO + Distributed FDB

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

- Not yet optimised!
- Measured **sustained 72 GiB/s W** ***application data (16 nodes)***

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 140 GiB/s
- Required Nb Prototypes 533 / 80 = x 3.8 = 122 nodes

(by 2030)

So, Why a ***Domain Specific*** Object Store?

Flexibility

- Many new technologies (H/W and S/W) coming to market
- Existing system is tied to POSIX

Consistency

- Data is presented in the same manner to applications
- Access is through semantically meaningful metadata

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 and **semantic** way to describe all ECMWF data

Semantics

1. ACID – Transactional.
2. Write blocks until data handed over
3. `flush()` blocks until data is visible
4. Visible data is immutable
5. Data can be masked

Into operations…

FDB5 into time-critical operations on Tuesday 11th June!

```
% fdb-stats class=od,date=20190612,expver=0001
Summary:
========

Number of databases                : 58
Fields                             : 83,747,723
Size of fields                     : 104,493,002,498,506 (95.0358 Tbytes)
Duplicated fields                  : 1,316,502
Size of duplicates                 : 2,668,035,857,106 (2.42656 Tbytes)
Reacheable fields                  : 82,431,221
Reachable size                     : 101,824,966,641,400 (92.6093 Tbytes)
Databases                          : 58
TOC records                        : 89,329
Size of TOC files                  : 191,427,584 (182.56 Mbytes)
Size of schemas files              : 949,228 (926.98 Kbytes)
TOC records                        : 89,329
Owned data files                   : 89,271
Size of owned data files           : 104,506,303,059,882 (95.0479 Tbytes)
Index files                        : 89,271
Size of index files                : 13,677,232,128 (12.7379 Gbytes)
Size of TOC files                  : 191,427,584 (182.56 Mbytes)
Total owned size                   : 104,520,172,668,822 (95.0605 Tbytes)
Total size                         : 104,520,172,668,822 (95.0605 Tbytes)
```

Front-ends and API

- Determines where the data is stored …
  - Run-time configurable
  - Implement data collocation policies
  - Manage data pools
  - Implements a simple interface:

Metadata:

```
CLASS    = OD,
TYPE     = FC,
LEVTYPE  = PL,
EXPVER   = 0001,
STREAM   = OPER,
PARAM    = 130,
TIME     = 1200,
LEVELIST = 500,
DATE     = 20190614,
STEP     = 12
```

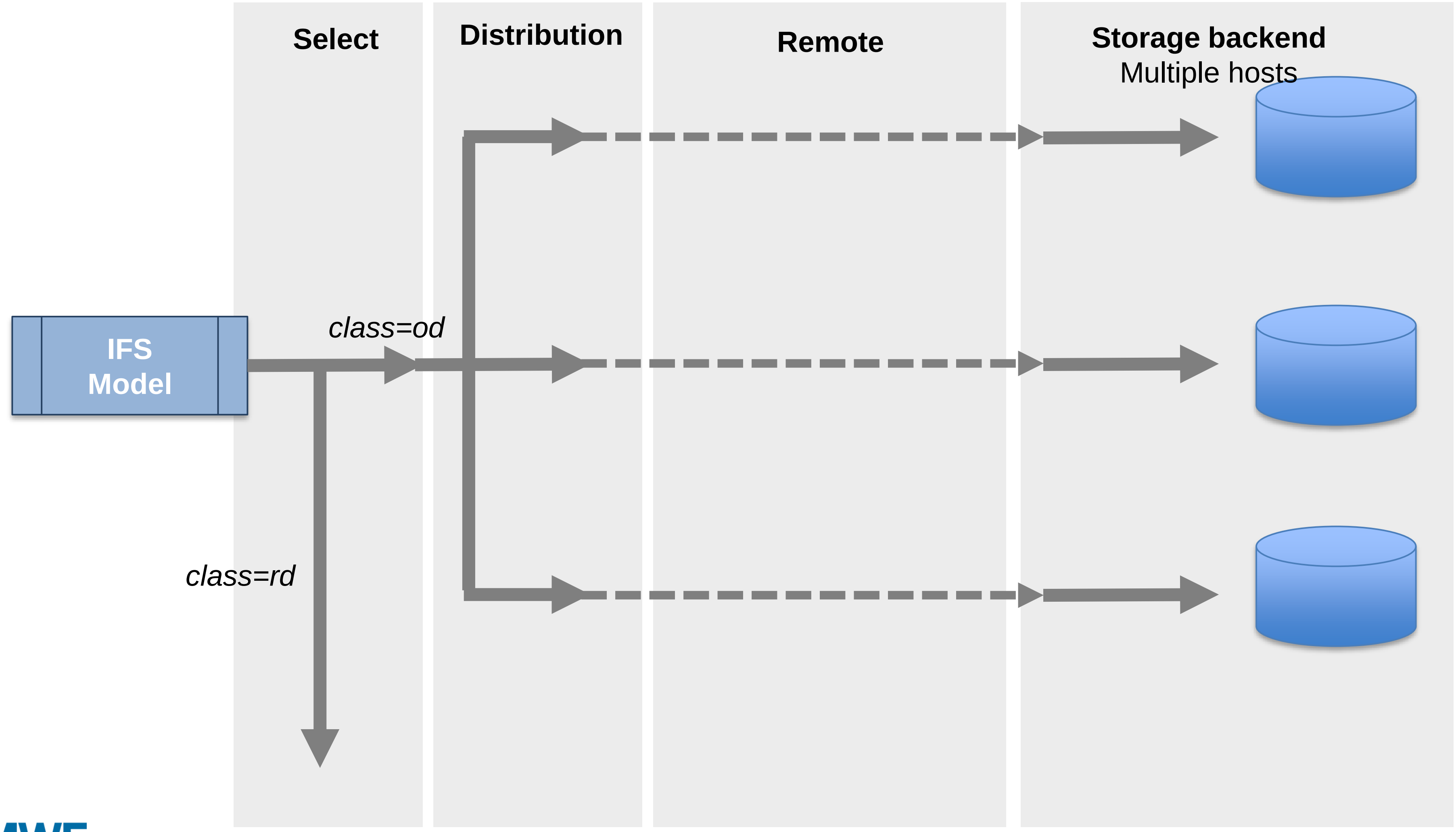
```
archive(Metadata key, void* data, size_t length);

retrieve(Metadata key, void* data, size_t& length);

flush();
```

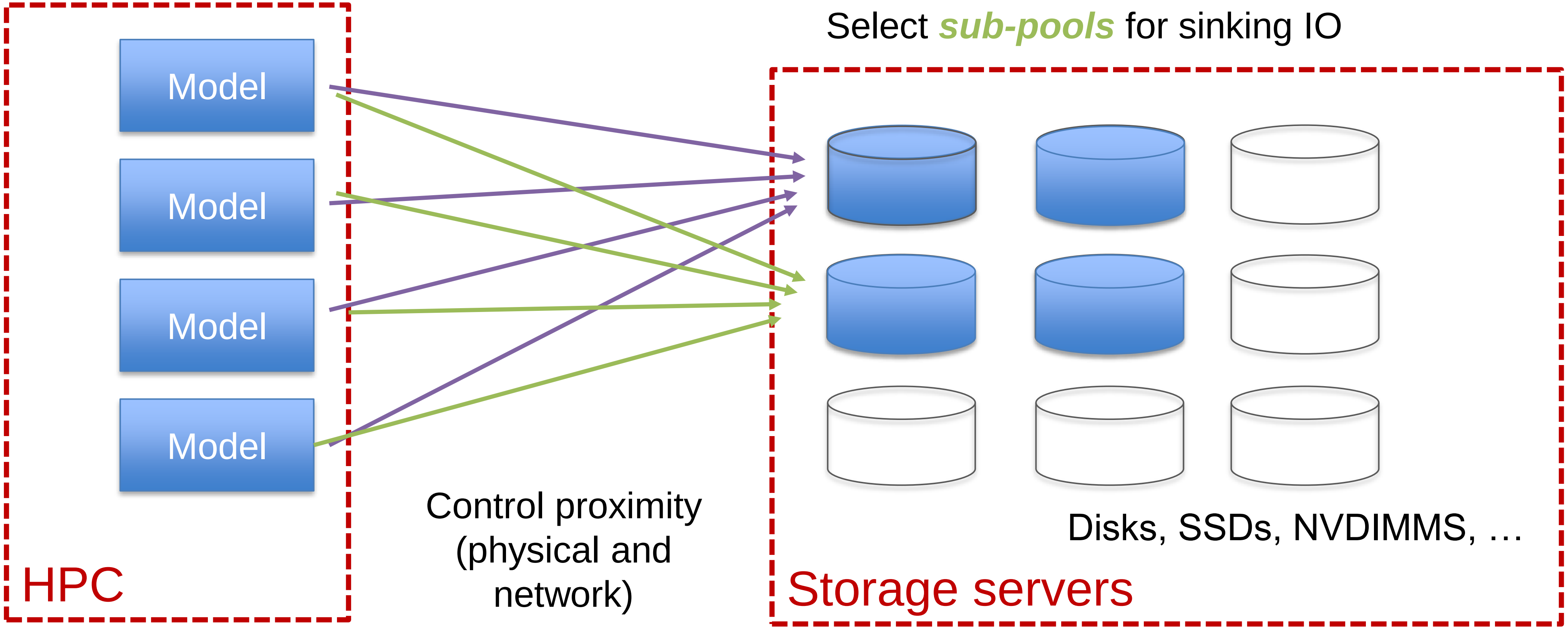
FDB5 Data Routing

- Meta-data controlled routing
- Fully asynchronous I/O
- Remote access TCP/IP



ECMWF

Capability vs Capacity



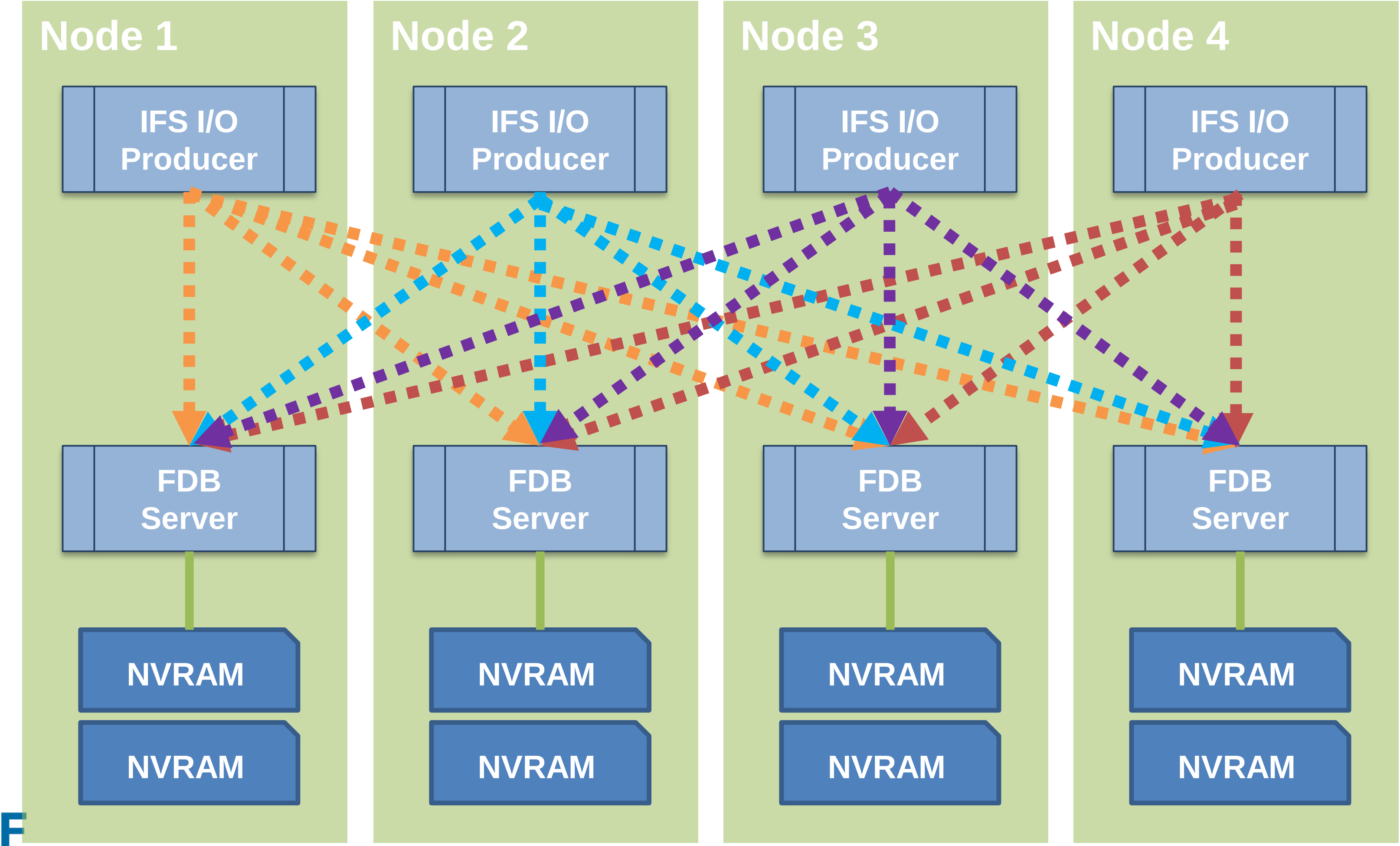
Select ***sub-pools*** for sinking IO

Control proximity (physical and network)

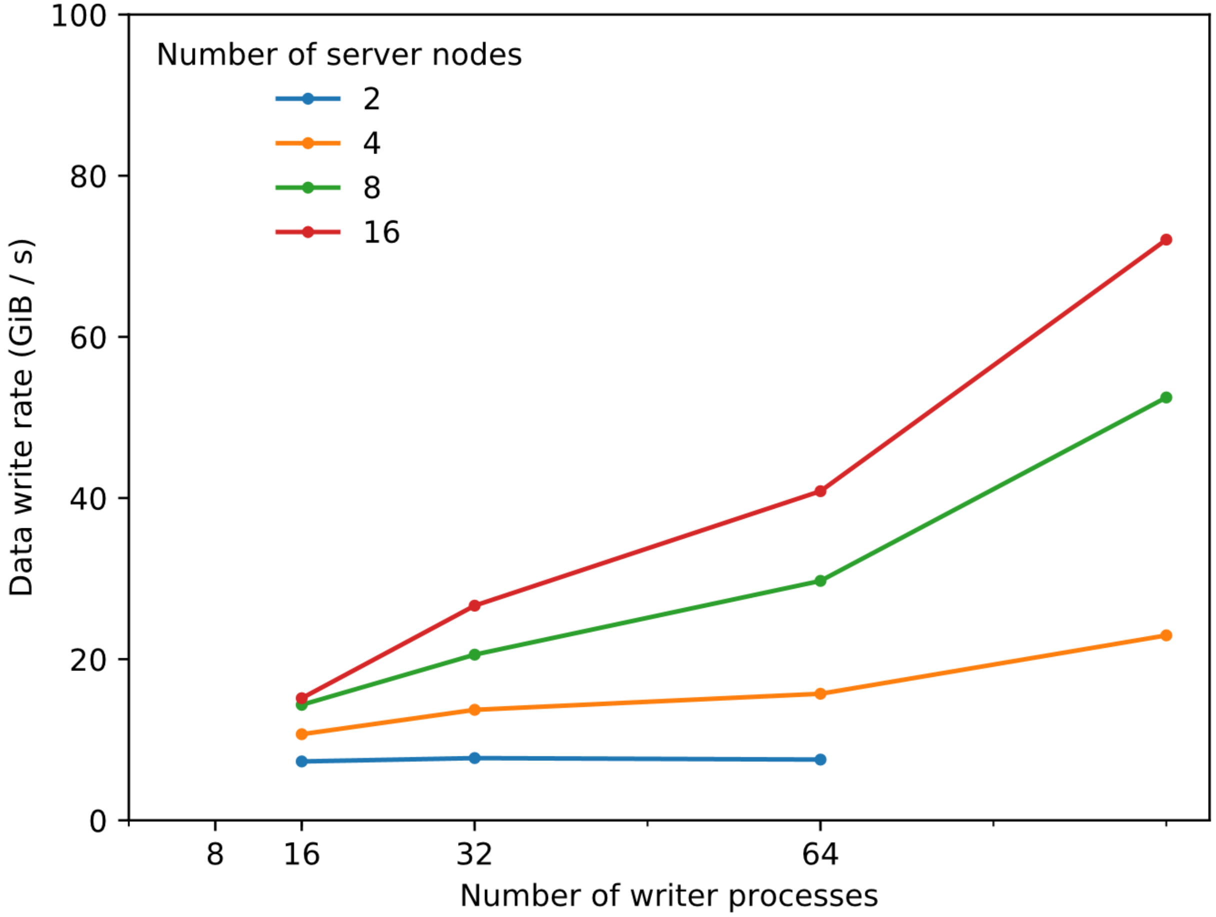
Disks, SSDs, NVDIMMS, …

Data Flow Schematic

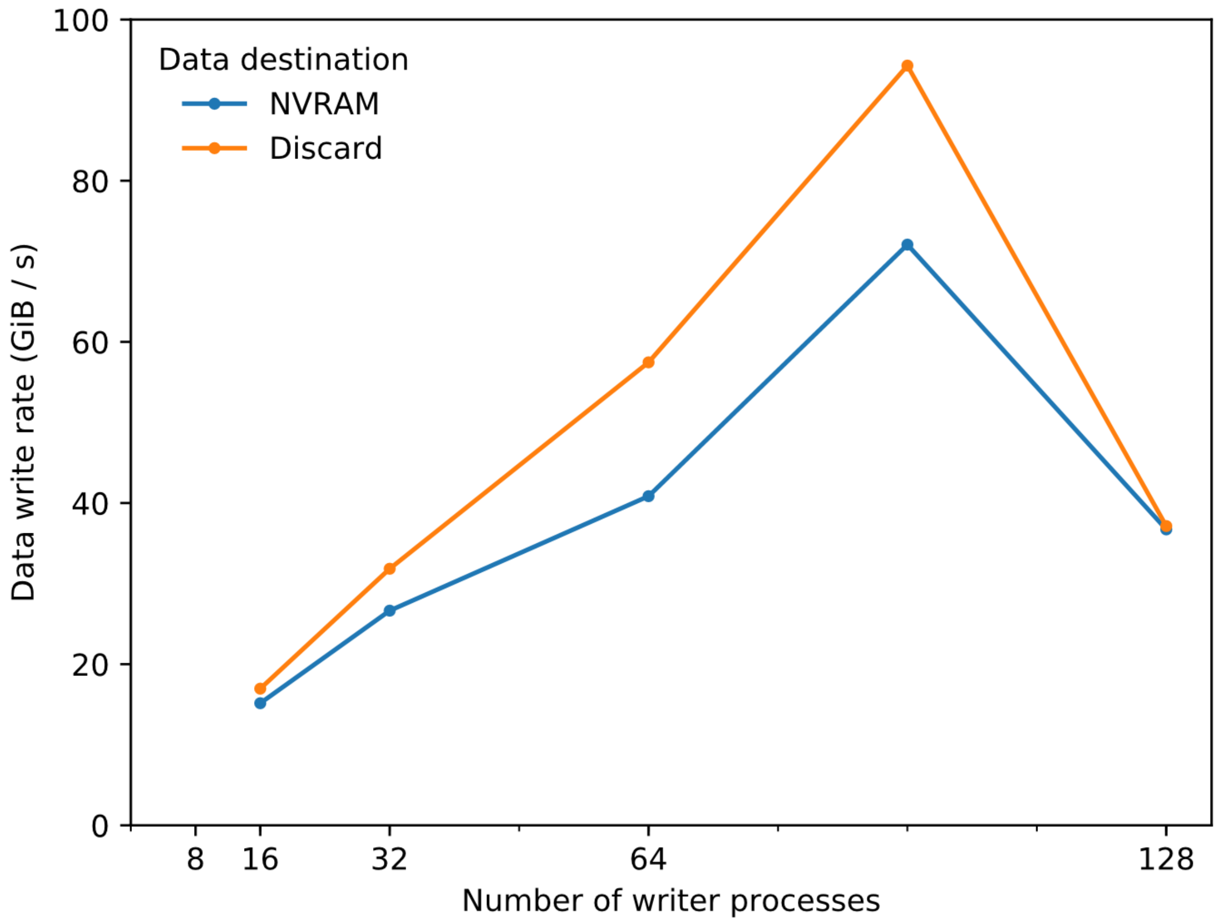
- All I/O operations are asynchronous, so computation can continue
- Distributed to all servers using a ***Distributed Hash****, so no synchronisation needed*



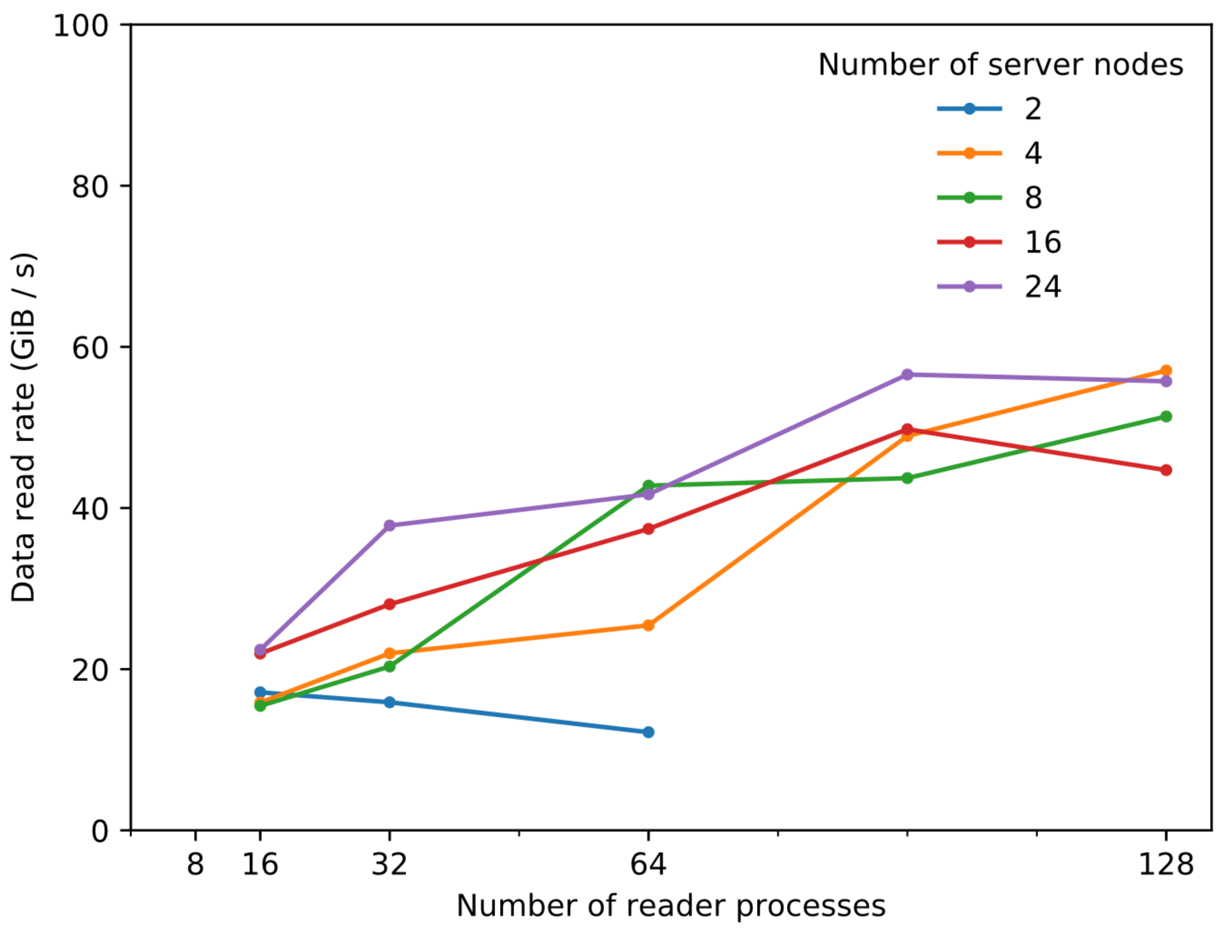
FDB5 Remote Write Performance (DCPMMs)



FDB5 Remote Performance



FDB5 Remote Read Performance (DCPMMs)



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
6 ensemble members

Messages To Take Home

Ensemble data sets are growing quadratically to cubically in size.
A challenge for time critical applications

Storage Class Memories will change the way we use and store data

ECMWF has adapted its workflows to take advantage of these upcoming technologies

*Semantic data access provides an abstraction under which **new technologies** can be introduced, and **performance** can be gained.*

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

nextgenio