ECMWF’s Exascale IO challenges

From inside the HPC to a whole Data archive migration

T. Quintino, S. Smart, O. Iffrig, J. Hawkes, J. Hanley, N. Manubens, E. Danovaro,
A. Bonnani, D. Sarmani, B. Raoult, P. Bauer

ECMWF

tiago.quintino@ecmwf.int

SIGIO/UK

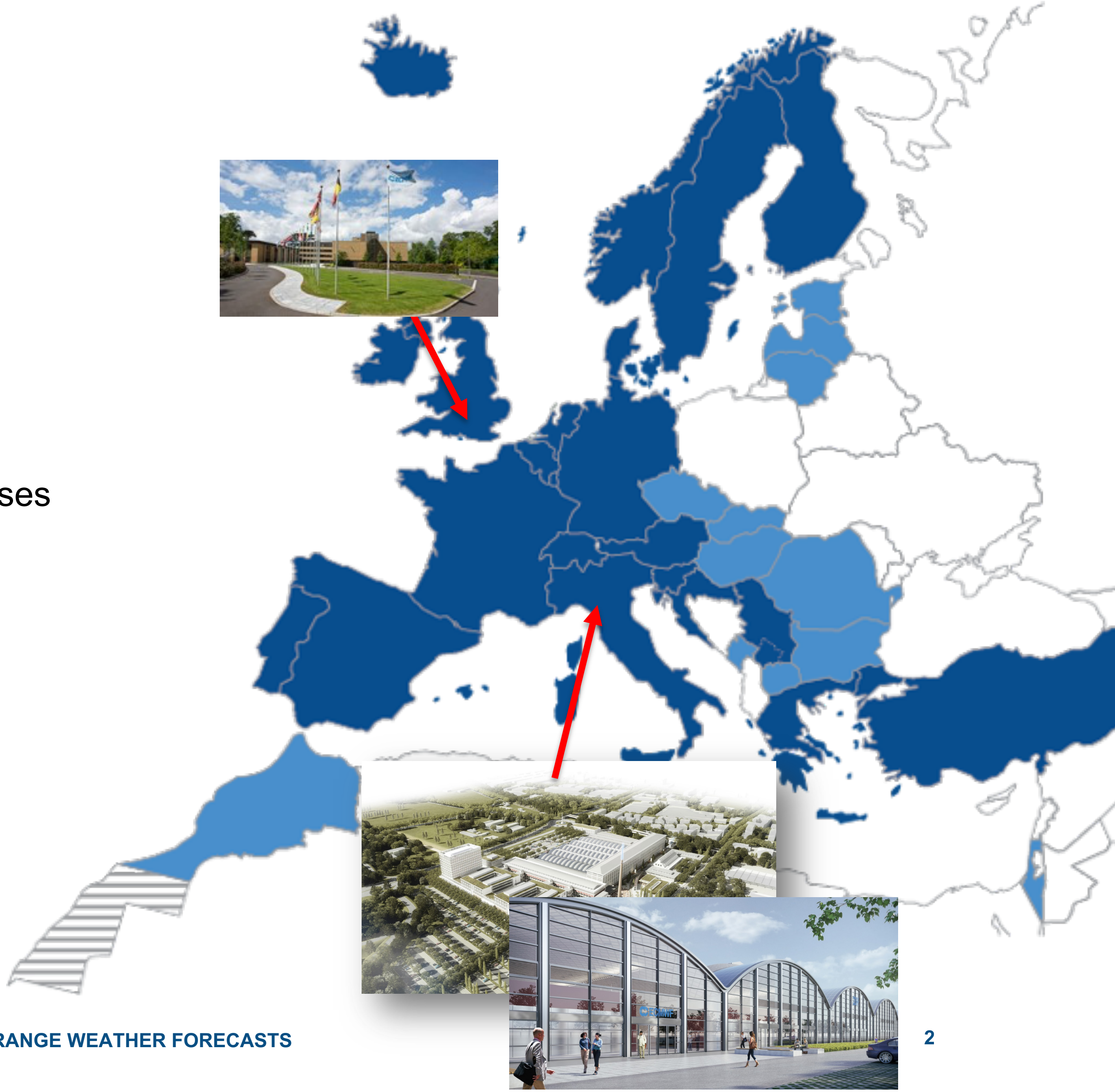
Workshop on Storage Challenges in the UK

23rd April 2020

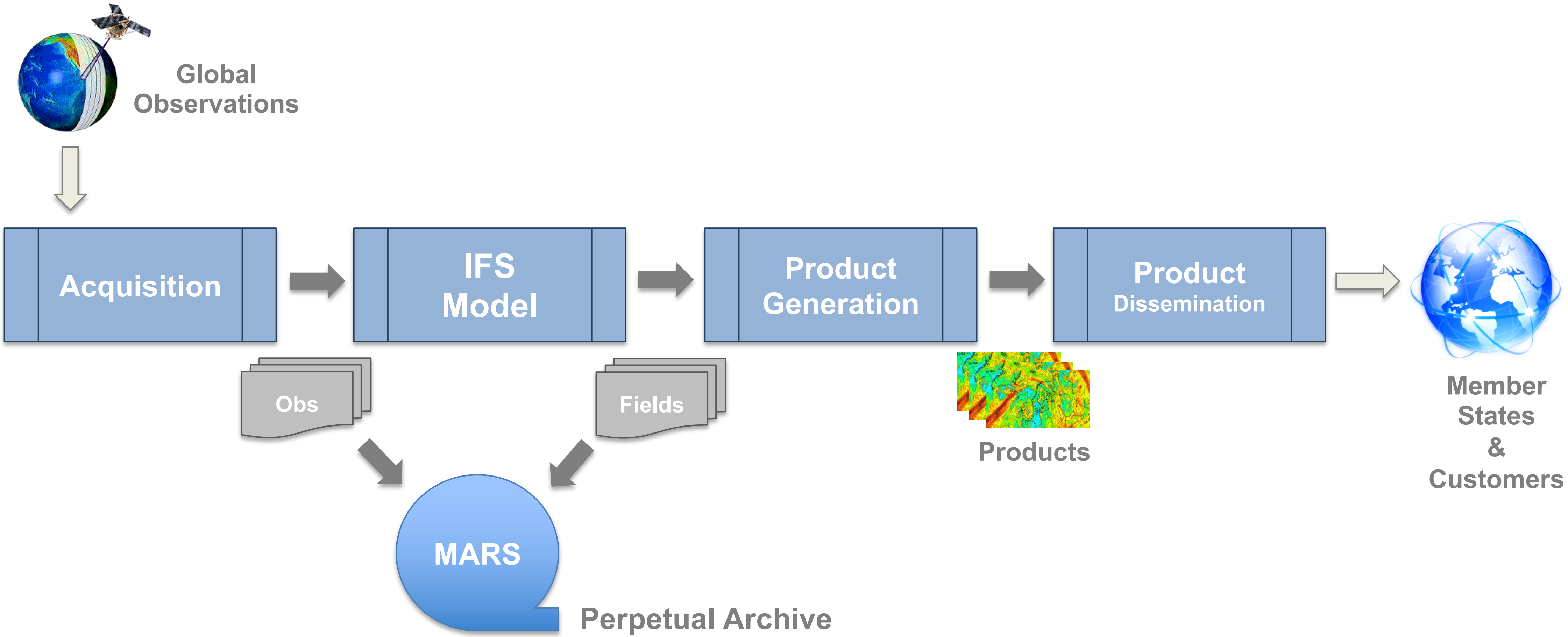
ECMWF

ECMWF’s Forecasting Systems

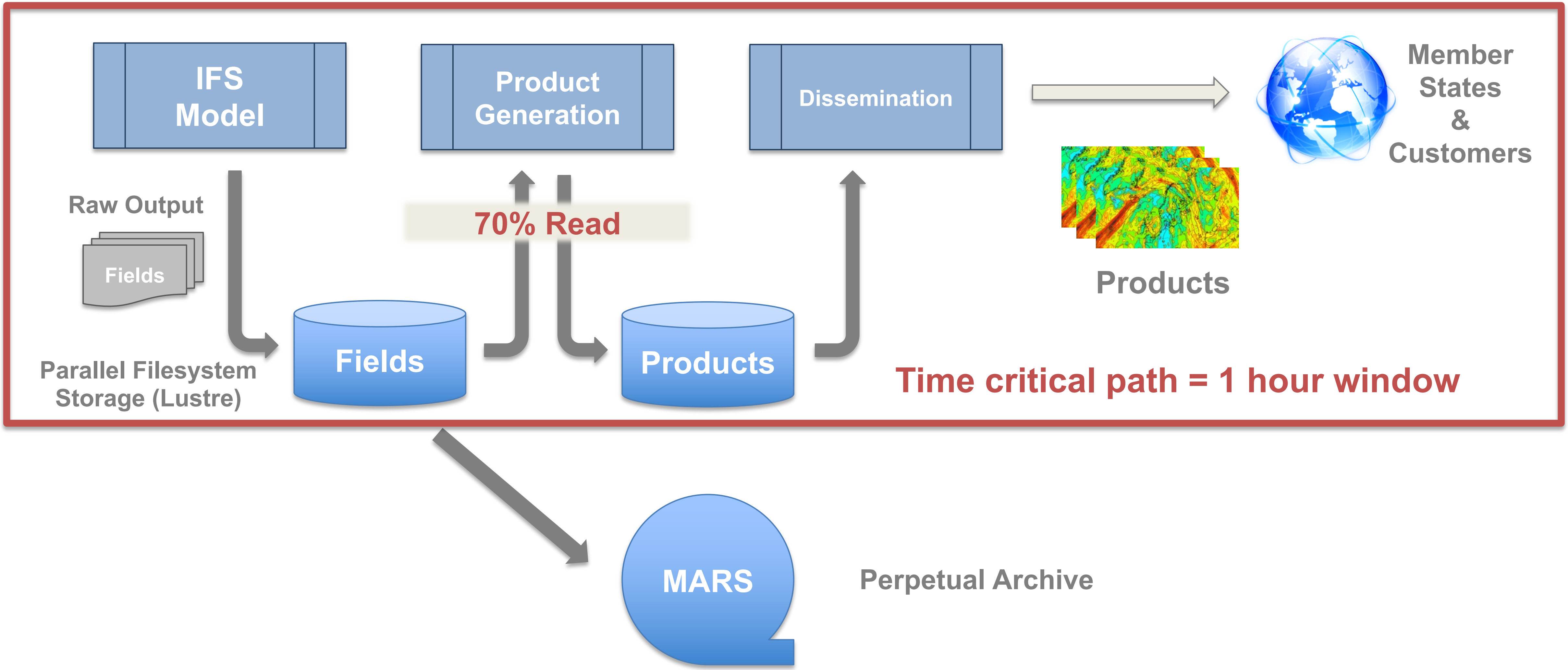
- Established in 1975.
- Intergovernmental Organisation
    - 22 Member States | 12 Cooperation States
    - 350+ staff
- 24/7 operational service
    - Operational NWP centre
    - Supporting NWS (coupled models) and businesses
- Research institution
    - Closely connected with researchers worldwide
- Operates two Copernicus Services
    - Climate Change Service (C3S)
    - Atmosphere Monitoring Service (CAMS)
- Supports Copernicus Emergency Management Service (CEMS)



ECMWF’s Production Workflow



ECMWF's Production Workflow



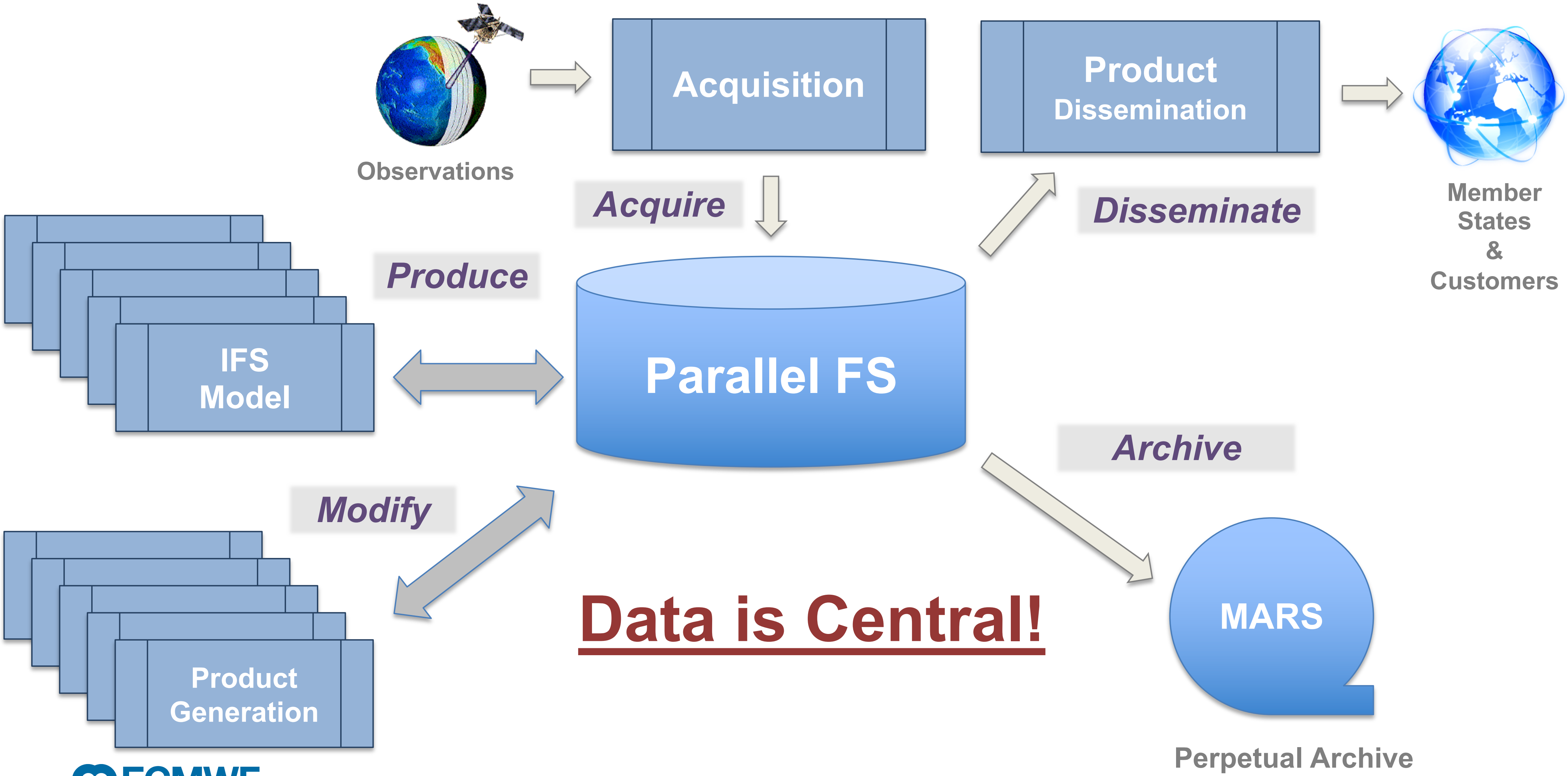
Effects of Product Generation using Parallel Filesystem

	IFS Model (No I/O)	IFS Model + I/O	IFS Model + I/O + PGen
Nodes	2440	2776	2926
Run time [s]	5765	6749	7260
Relative	-	**+ 17%**	**+ 26%**

Runtimes affected by the existence of another parallel job in the system:
Product Generation reading the data the model is writing
"Coupling" via the file system!

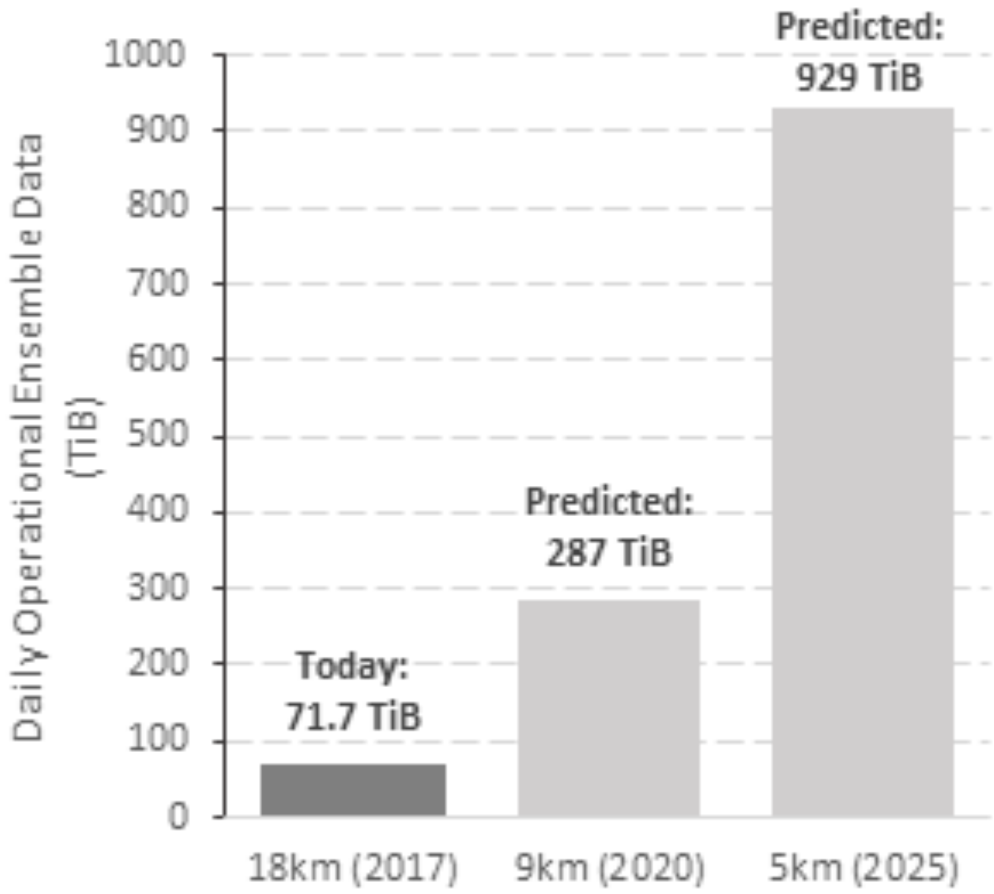
9Km 50 member ensemble
Broadwell nodes 2x18 cores
Cray XC40 Aries interconnect
Lustre FS IOR 90GiB/s

Storage View of Workflow



NEW HPC Facility + New HPC system

Historical Growth of Generated Products

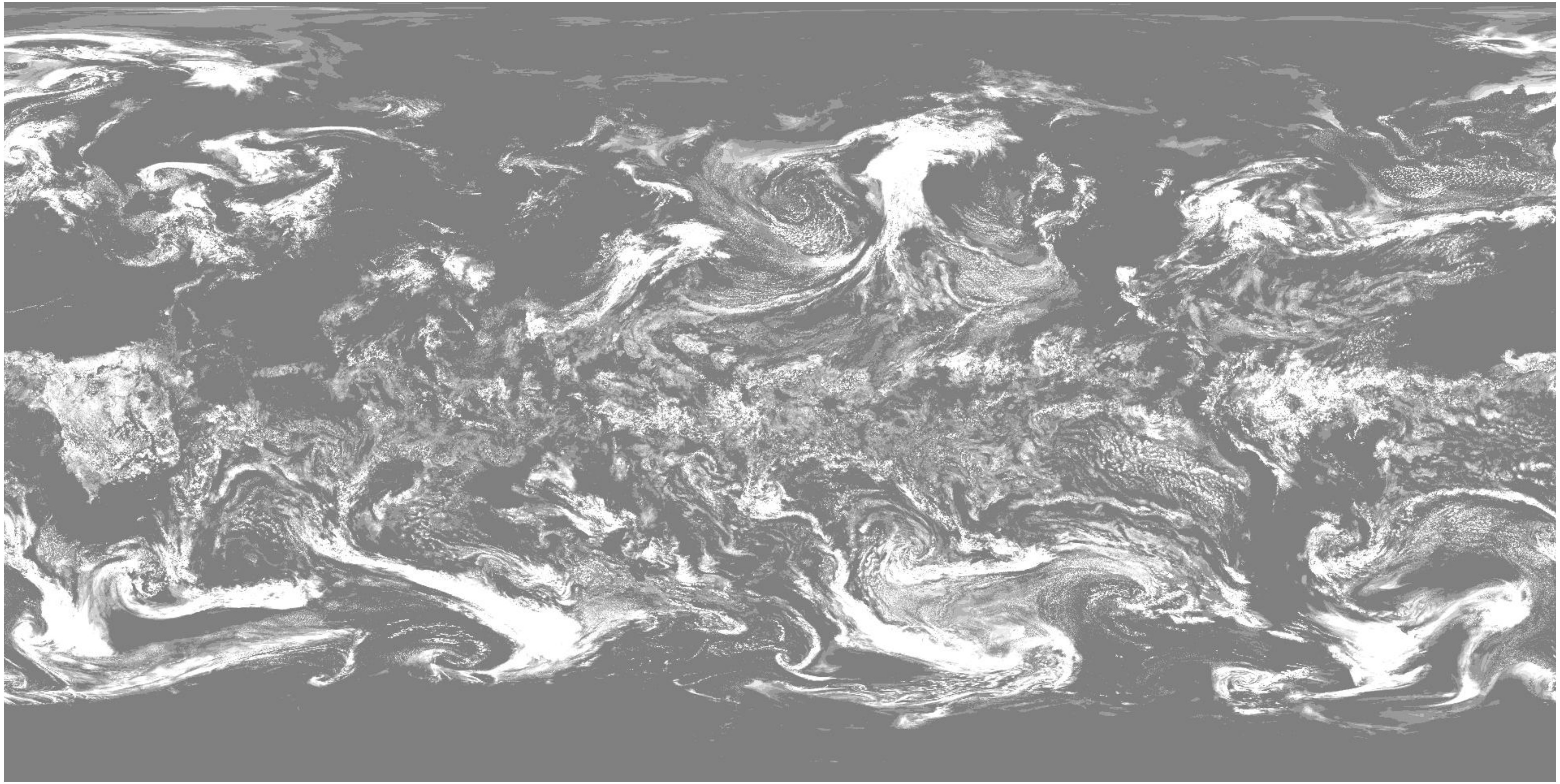


Model Output Projected Growth

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***

TCo7999 (~1.25km) 256 Megapixel



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

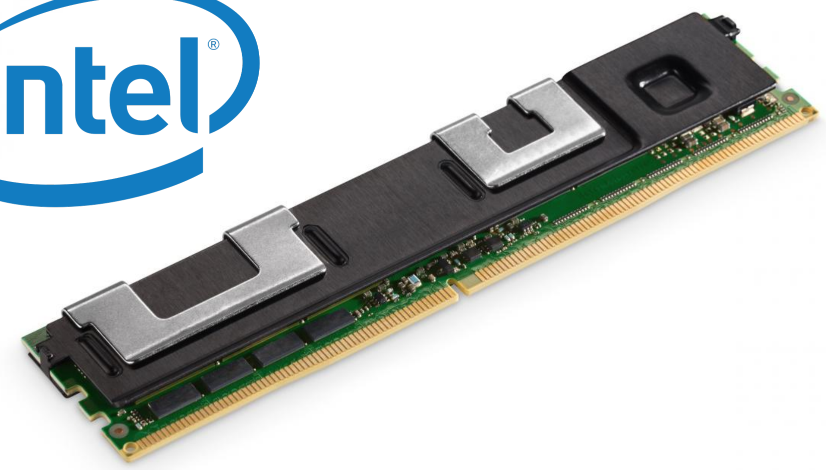
Storage and I/O @ Exascale

How large is a 1.25 km x50 ensemble forecast?

- 50 member ensemble forecast
- *Compressed* GRIB2 data @ 16bit & 24bit
- @ 18km O640 — 21 TiB
- Resolution @ 9km O640 → O1280 — x 3.3
- 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

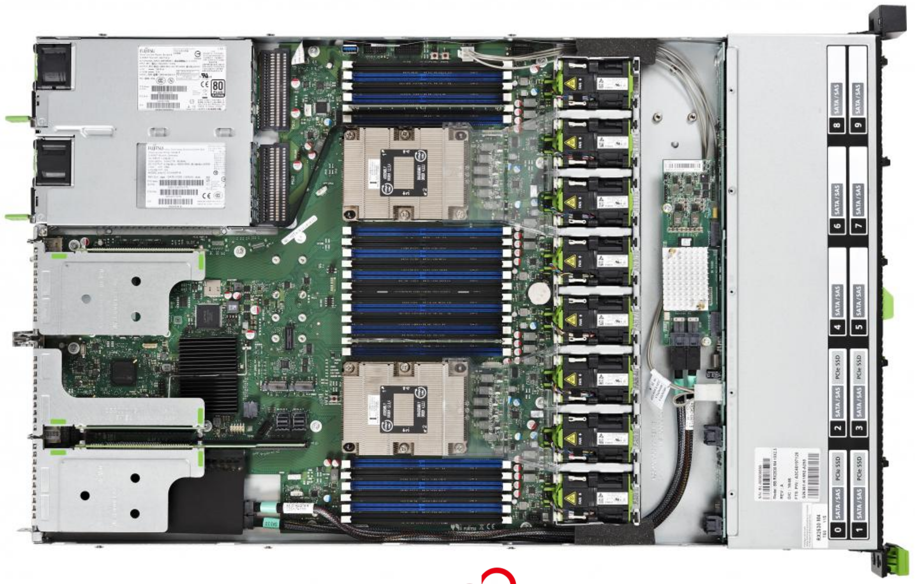
21 TiB x 173.2 = 3638 TiB

NextGenIO Prototype



- Read all @ www.nextgenio.eu
- Development of an HPC node by **with Intel Optane DCPMM**
- Dual-CPU Intel® Xeon® SP nodes (48 cores)
- OmniPath network
- 192GB DRAM
- **3TiB of NVRAM DIMMs (max 6 TiB)**

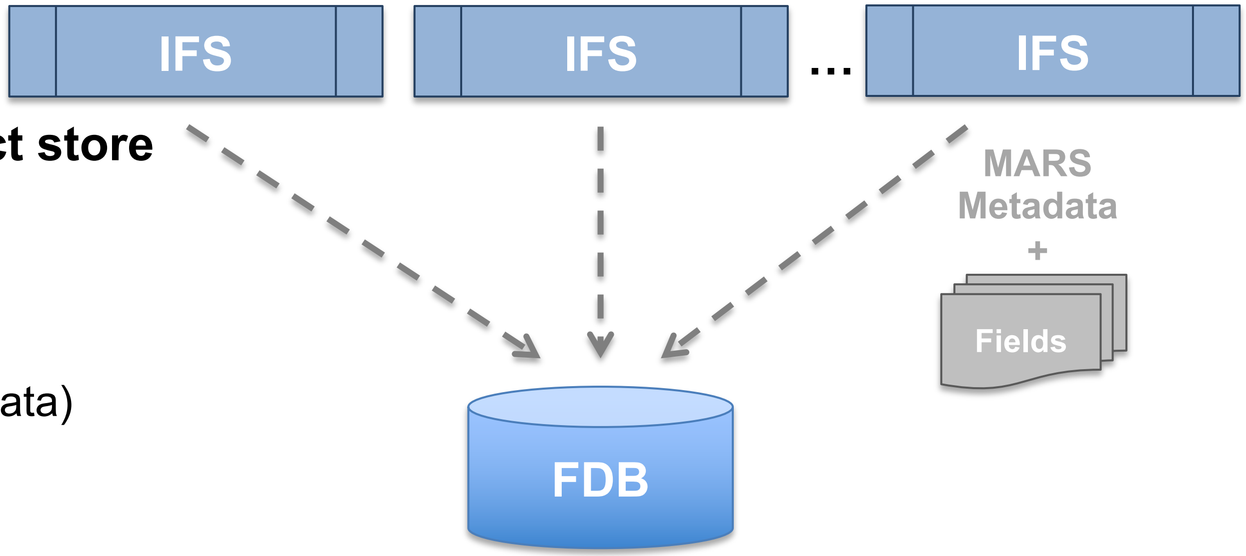
- **Prototype system**
  - **34 compute nodes**
  - **Hosted @ EPCC, Edinburgh**



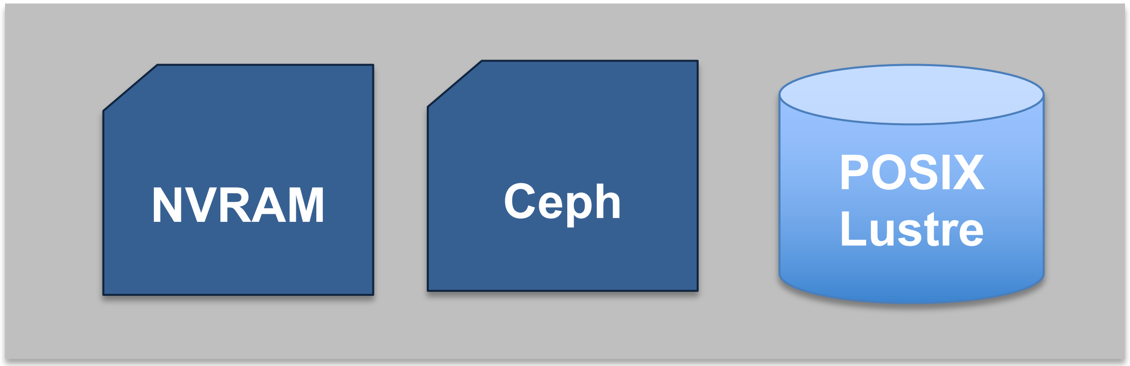
34 x 3 TiB Byte Addressable Storage

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)
  - NVRAM 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,
```

FDB 5 Semantics

1. ACID – ***Transactional***
2. Write blocks until data handed over – ***Asynchronous***
3. `flush()` blocks until data is visible – ***Consistent***
4. Write-once, don't overwrite - ***Immutable***
5. Data can be masked – ***Versioned***

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

Front-ends and API

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

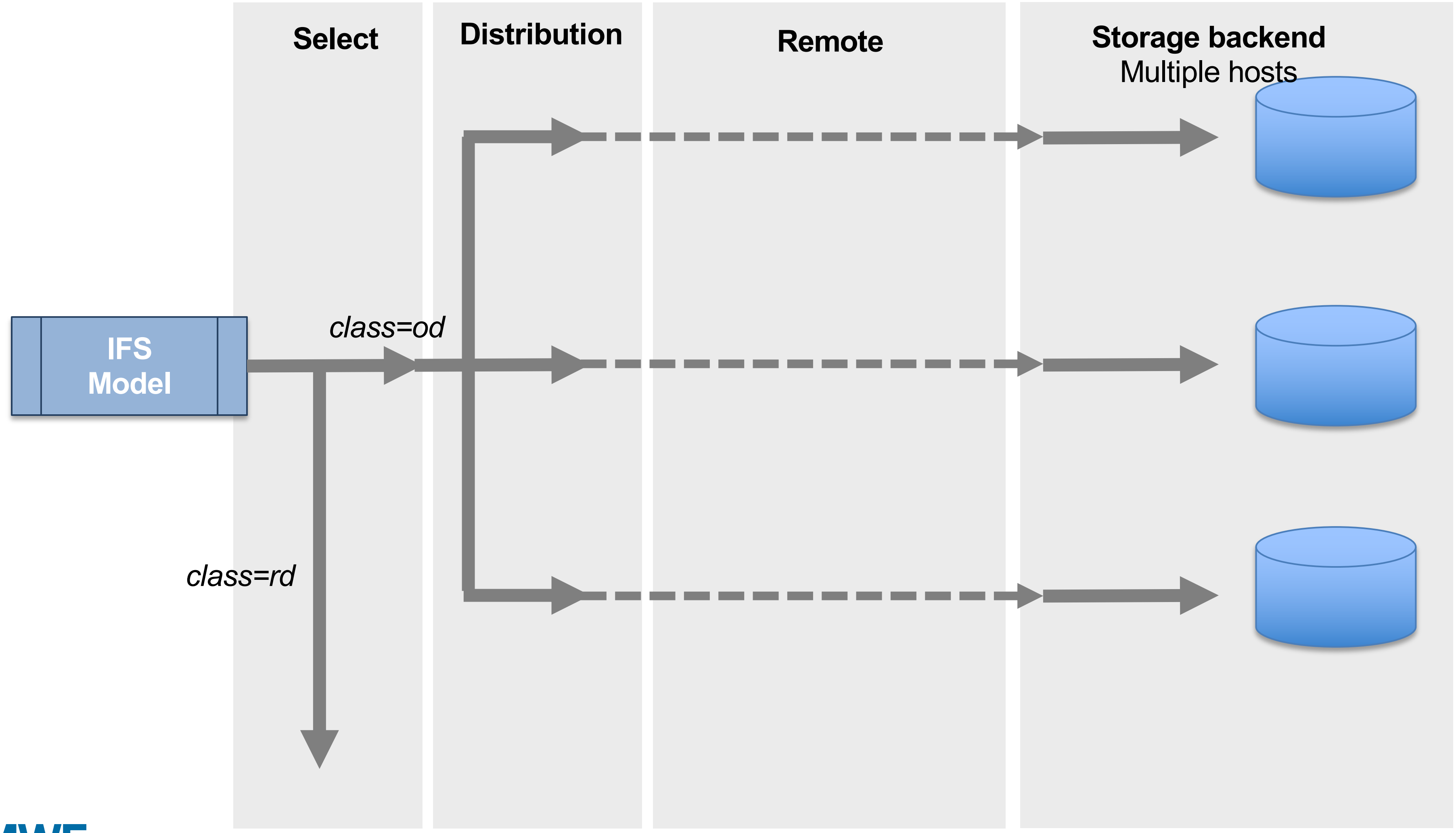
Metadata:

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

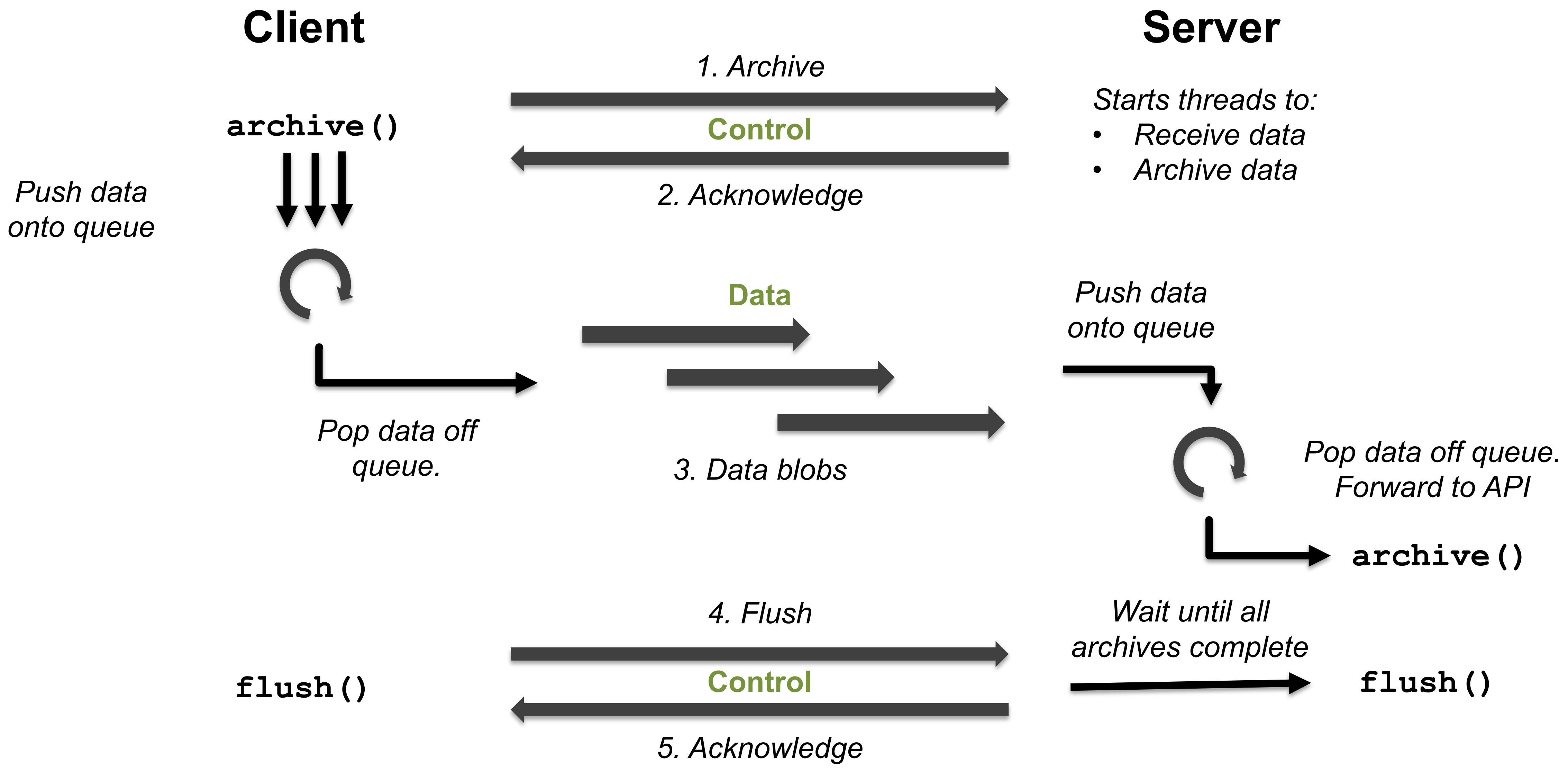
```
archive(Metadata key, void* data, size_t length);

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

flush();
```

FDB5 Data Routing

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



Asynchronous Archiving Data

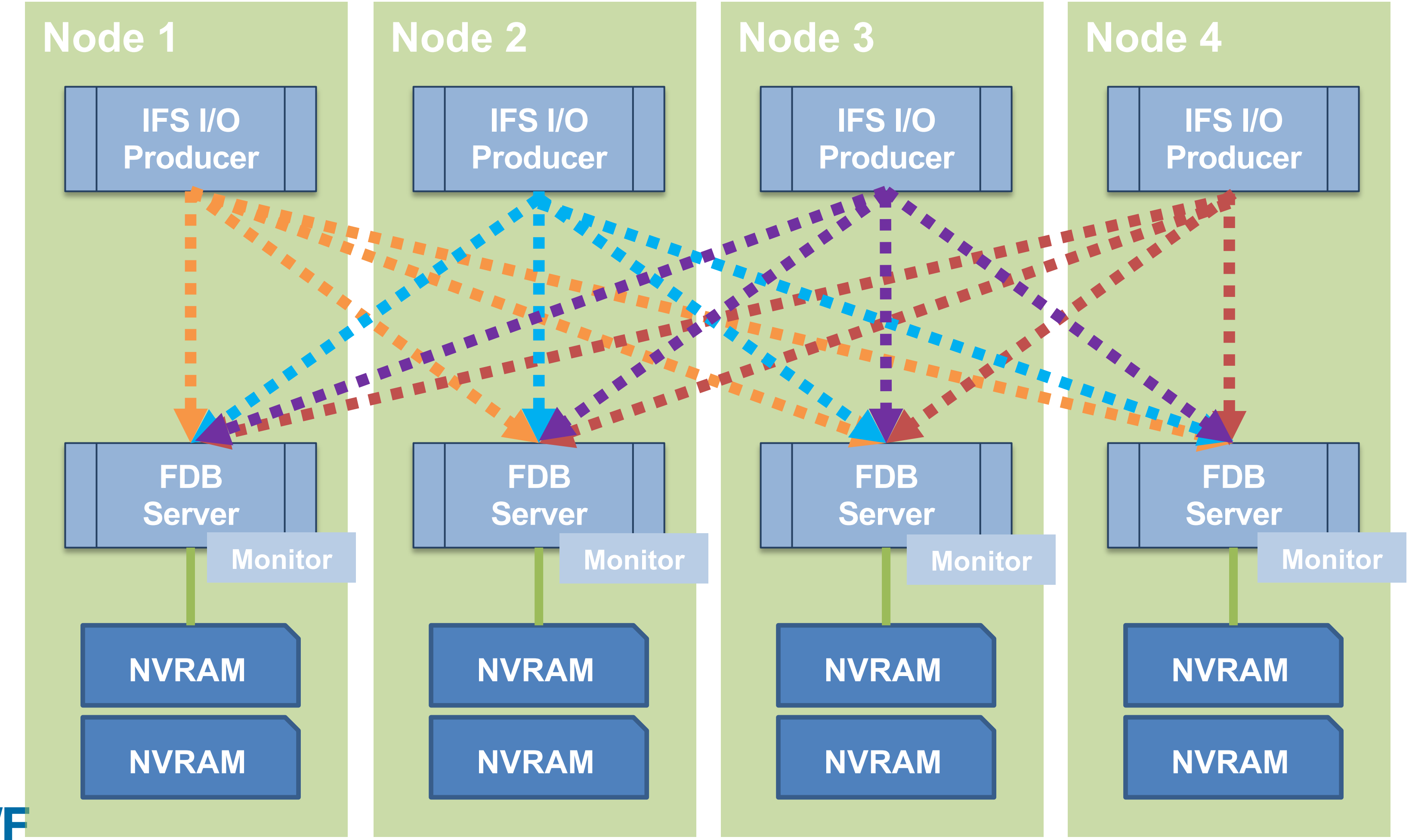


Into operations…

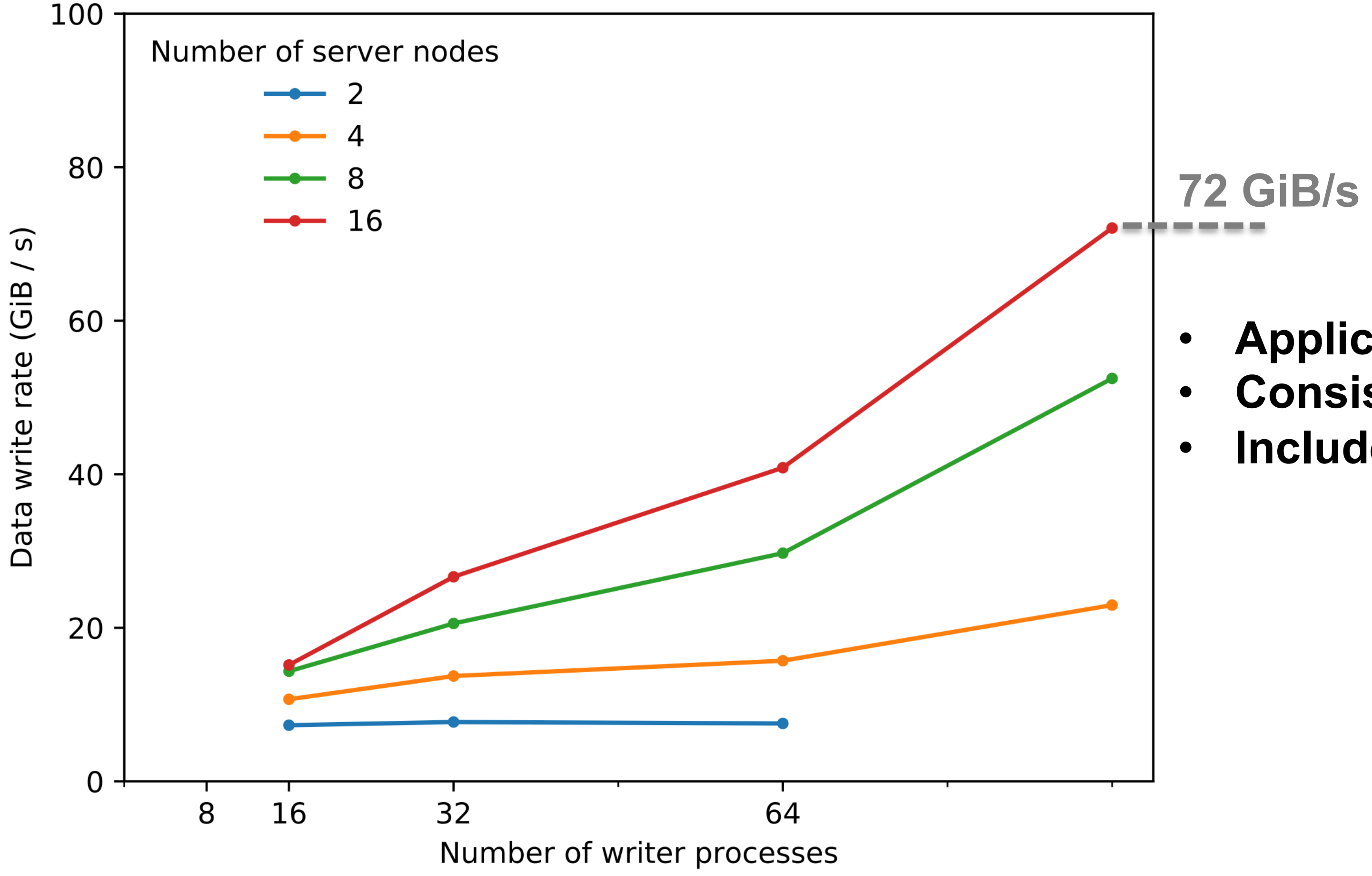
FDB5 into time-critical operations on Tuesday 11th June!

```
% 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)
```

Performance Benchmark Test

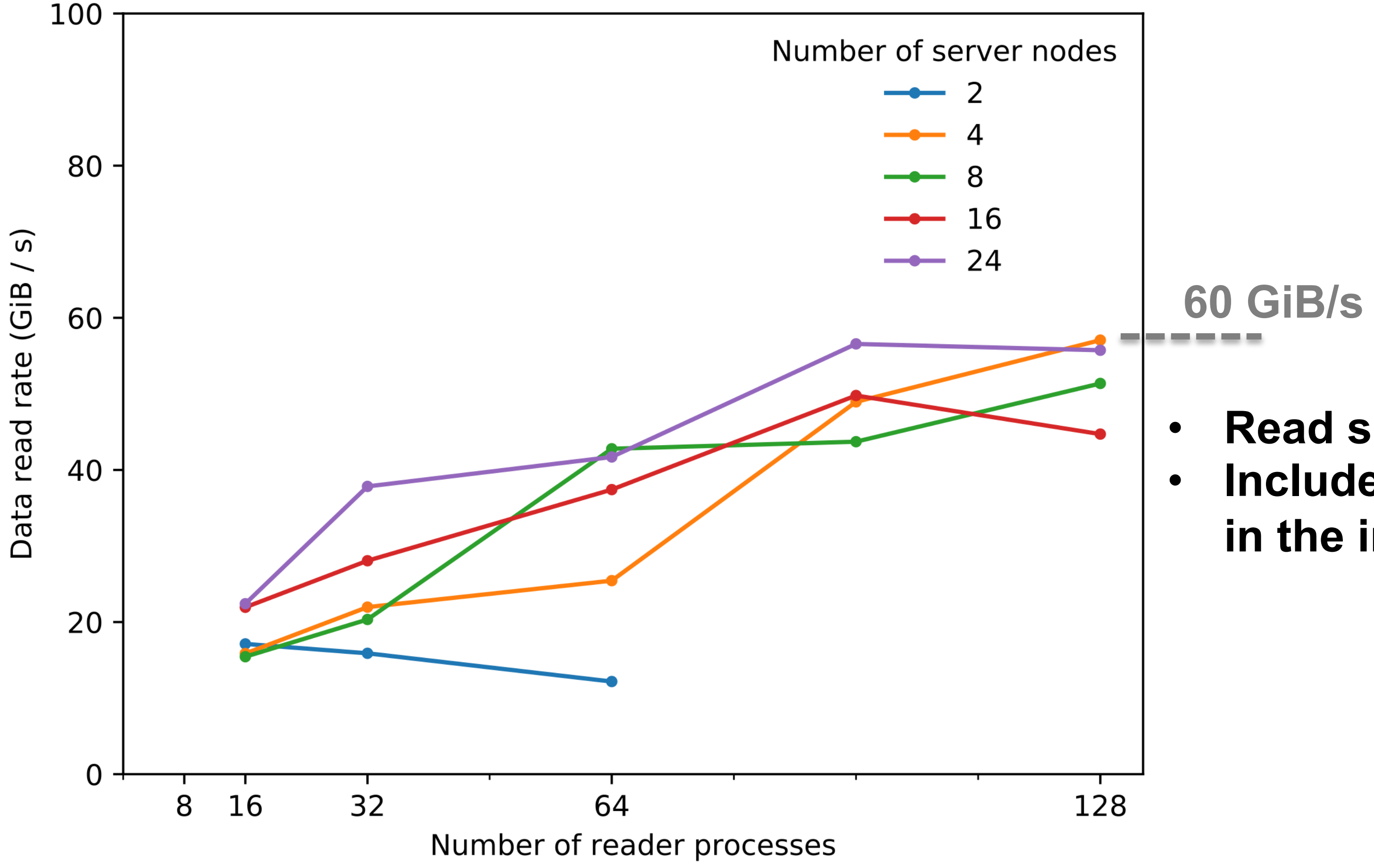


FDB 5 Parallel Write Performance to DCPMMs



- **Application data measured**
- **Consistency semantics**
- **Includes wait for *flush()***

FDB 5 Parallel Read Performance to DCPMMs



- **Read slower than write**
- **Includes the *data lookup* in the indexes**

Running the forecast model

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

Runtimes no longer affected by the Product Generation!!!

NextGenIO prototype. 32 nodes
Intel OmniPath2 interconnect
6 ensemble members

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)***

ECMWF

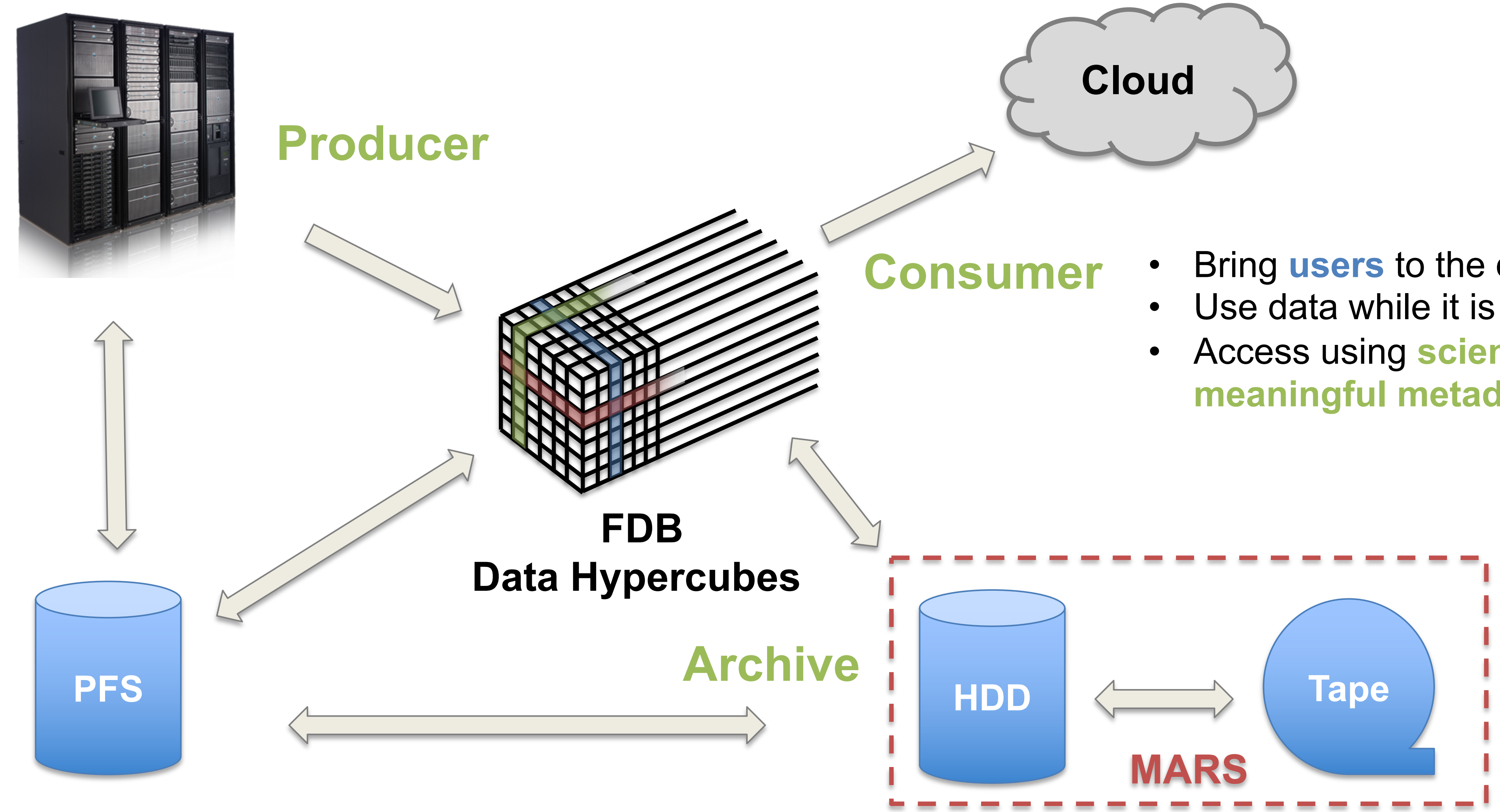
Can we handle the 1.25 km ensemble forecast?

- 50 member ensemble forecast
- *Compressed* GRIB2 data @ 16bit & 24bit
- @ 1.25km 7999 3638 TiB
- Required to read 70% x 1.70
- @ 1.25km 7999 6185 TiB
- Time to solution 1 hour 6185 TiB / 3600 = 1759 GiB/s
- NextGenIO performance (16 nodes) 132 GiB/s
- Required Nb Prototypes 1759 / 132 x 16 = **213 nodes**

NextGenIO x 6.7 (by 2035)

ECMWF Novel Data Flows

Data Analytics / Machine Learning

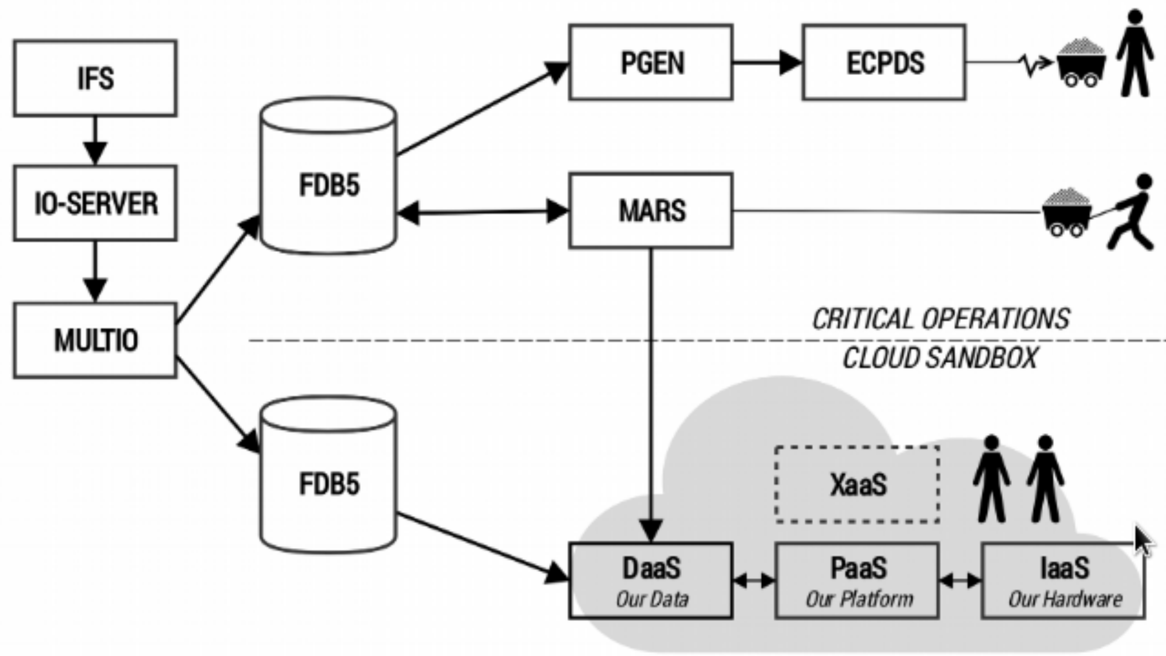


- Bring **users** to the data
- Use data while it is **hot**
- Access using **scientifically meaningful metadata**

Providing ECMWF Data to a Cloud

Requirements:

1. Bring users to the data and avoid moving the data out of the data centre
2. Provide users with computing resources collocated with data
3. Data-centric approach ***"move the compute, not the data"***



How to enable this:

1. Mechanism to pull/push data from ECMWF
2. Mechanism to run custom post-processing
3. Mechanism to explore & discover data

New development: Polytope
Watch this space ;-)

Messages To Take Home

Ensemble data sets are growing quadratically to cubically in size,
How can we best serve this high-resolution data?

New technologies in the ***horizon***
NVRAM and other Storage Class Memories

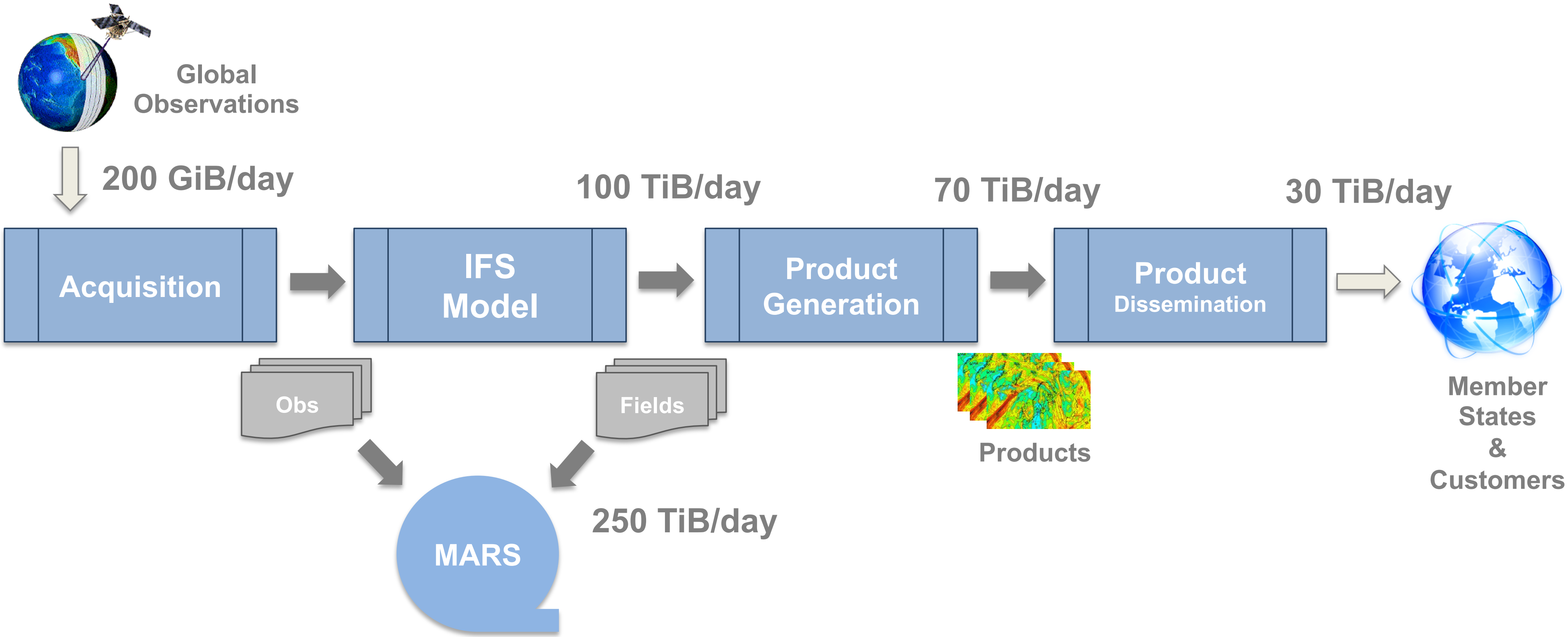
ECMWF is adapting its workflow to take advantage of these upcoming technologies

Developed a distributed object store for Weather and Climate

Working to serve these datasets out of the HPC to Data Analytics Platforms

How about that move of Data Centre?

ECMWF's Production Workflow



Moving a Data Centre

How to move a 24x7 data center?

- *Run **weather forecast 4x per day***
- *Still produce **~ 100TiB/day***
- ***Obtain a new HPC** and install in place*

- *Main issue is **Data Handling System (DHS)***
  - ***350 PiB growing @ 1PiB / 4 days***

Transfer?

350 PiB @ 100Gbips network = 339 days

350 PiB @ 300TiB/day tape access = 1194 days

Code Digression

- *How to do a multi-threaded transactional swap …*

S = 350
B = 0

Lock(S), Lock(B)

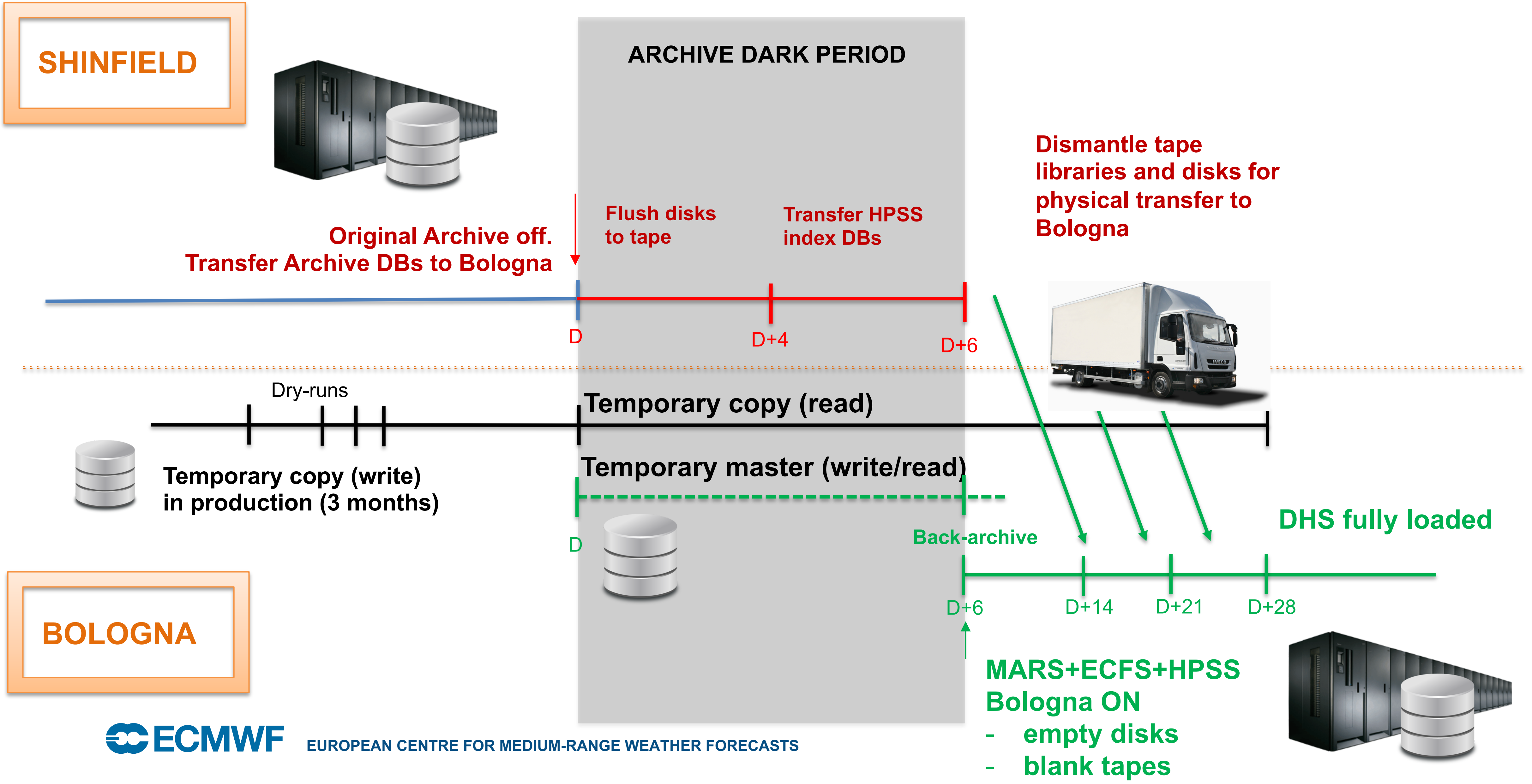
LOCKED PERIOD

TMP = S
S = B
B = TMP

Unlock(B), Unlock(S)

clean(TMP)

DHS Service Transition plan



THANK YOU !

QUESTIONS ?

This work has been supported by NextGenIO project and partly funded by European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement 671951

This work has been supported by LEXIS project and partly funded by European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement 824115