Data-Centric Workflows in Exascale Weather Forecasting

ECMWF's approach in Destination Earth Programme

Tiago Quintino, J. Hawkes, S. Smart, E. Danovaro, N. Manubens, O. Iffrig, D. Sarmani, B. Raoult, P. Bauer

ECMWF

tiago.quintino@ecmwf.int

7th ENES HPC Workshop

ECMWF

ECMWF's Forecasting Systems

Established in 1975, Intergovernmental Organisation

- 22 Member States | 12 Cooperation States
- 350+ staff

24/7 operational service

- Operational NWP – 4x HRES+ENS forecasts / day
- Supporting NWS (coupled models) and businesses

Research institution

- Experiments to continuously improve our models
- Reforecasts and Climate Reanalysis

Operate 2 EU Copernicus Services

- Climate Change Service (C3S)
- Atmosphere Monitoring Service (CAMS)
- Support Copernicus Emergency Management Service (CEMS)

Reading, GB

Bonn, DE

Bologna, IT

EU Destination Earth Programme

Started January 2022

- ECMWF
- European Space Agency (ESA)
- EUMETSAT

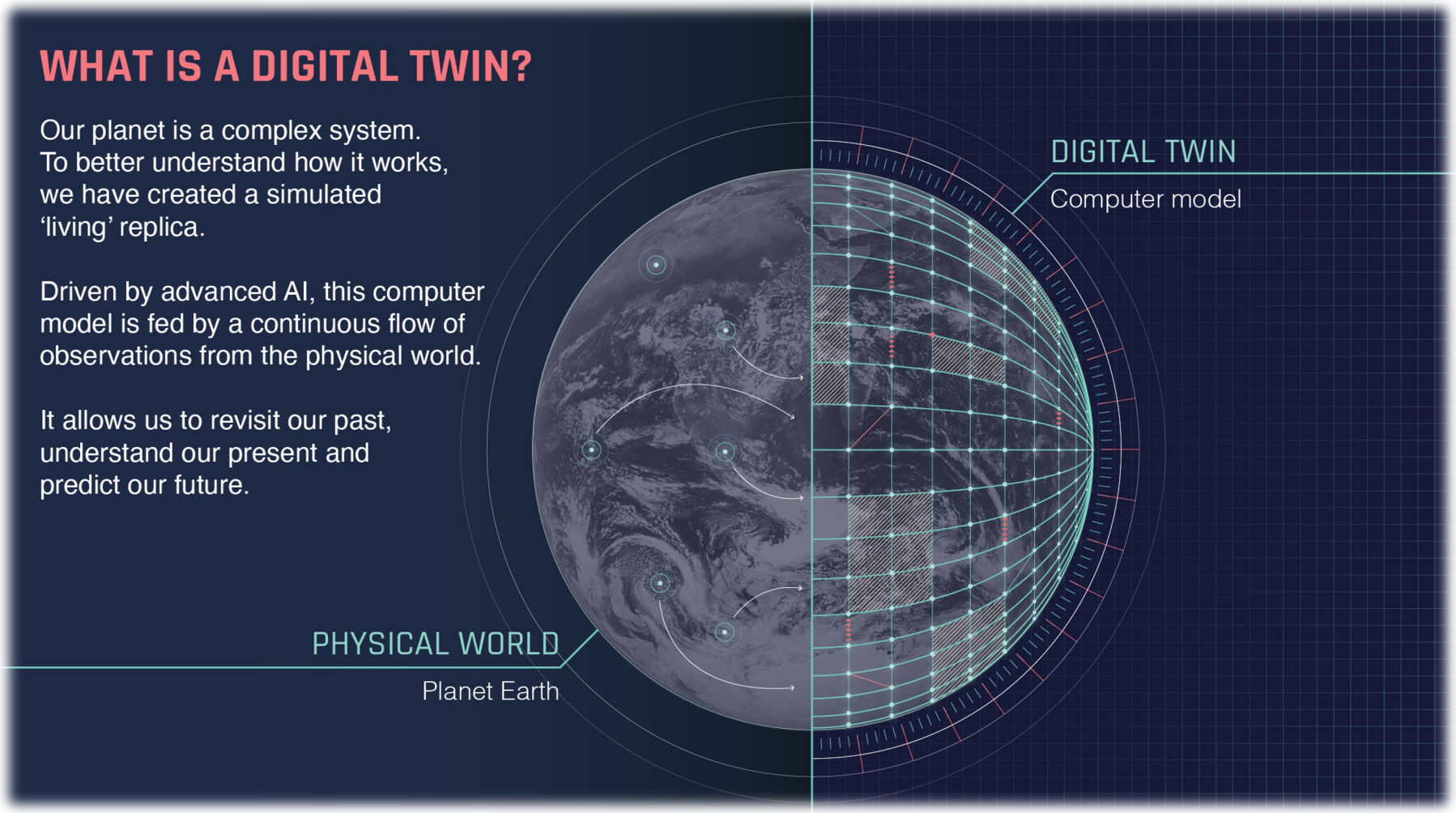
Part of EU's Green Deal and Digital Strategy

Build 2 Digital Twins:

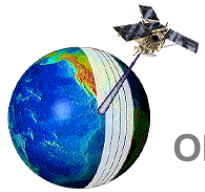
- Weather Extremes
- Climate

Weather Extreme DT

- daily run(s)
- Km-scale resolutions



ECMWF's Production Workflow



Global Observations

200M obs/day

Destination Earth Data Set

1000 TiB/day

45TiB/day

~350 world destinations

Acquisition

Earth System Model

Product Generation

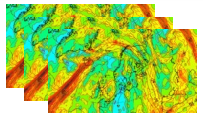
Product Dissemination

Obs

Fields

Archive

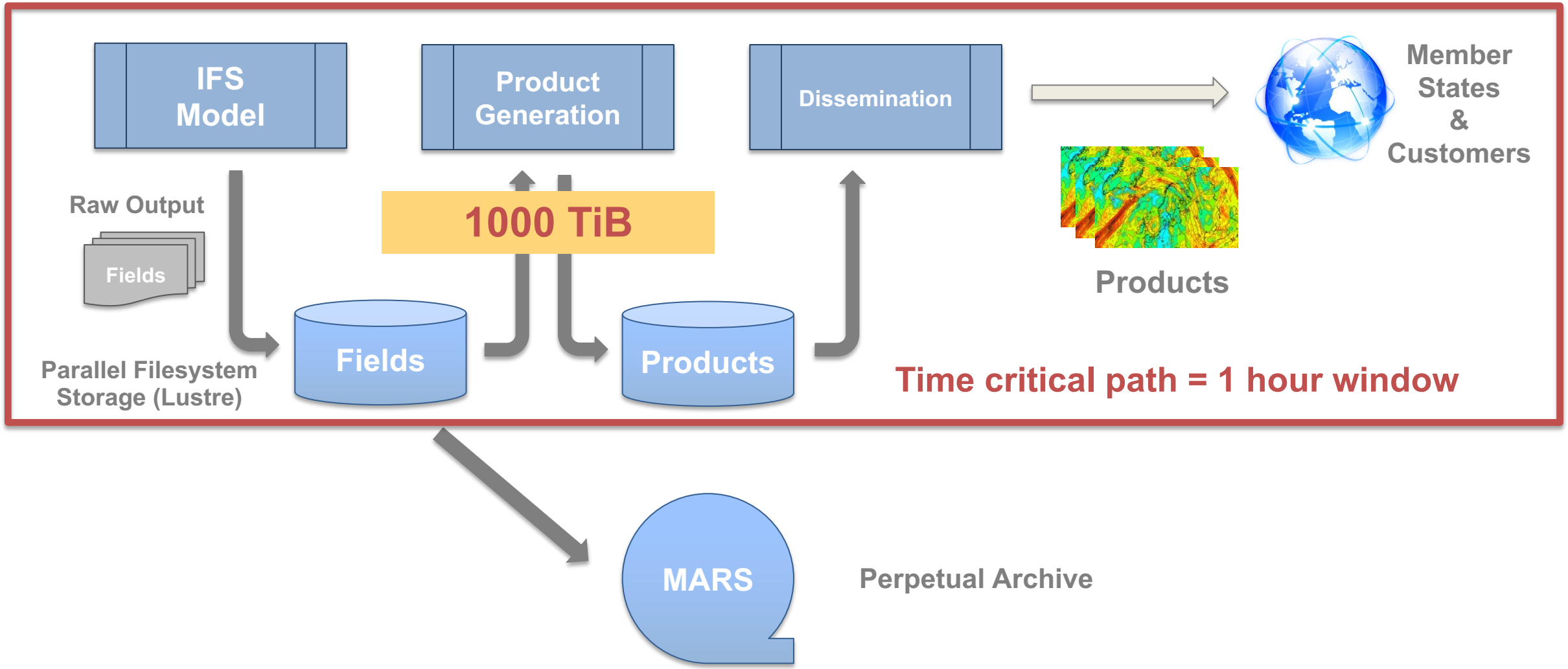
Perpetual Archive



Products

Member States & Customers

ECMWF's Production Workflow - Challenges



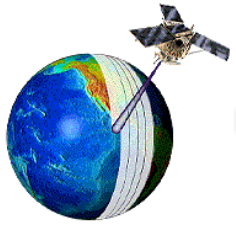
Effects of Product Generation using Parallel Filesystem

	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



Observations

Acquisition

Product Dissemination

Member States & Customers

Acquire

Disseminate

Produce

Earth Sys Model

Storage Parallel FS

Analyse

ML + Analytics

Archive

Modify

Product Generation

Data is Central!

Perpetual Archive

Data is only as useful as what you do with it

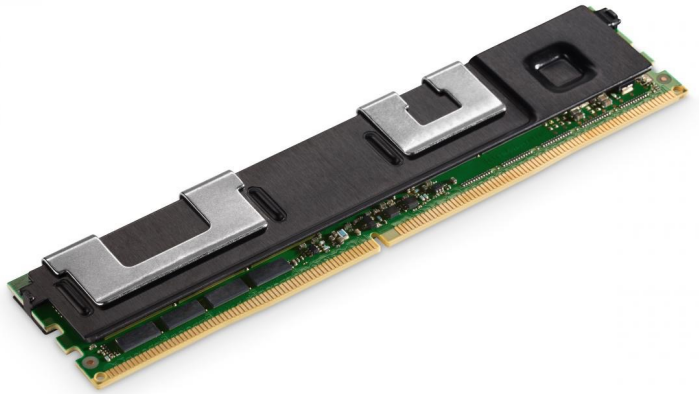
How to make data more useful?

Destination Earth Engine

- Framework for Earth System Model Workflows
- Think of a *Game Engine* but for Earth Systems…
  - It's a Framework – not model specific
  - Series of API's and services
  - Opt-in Components
- **Components:**
  - Workflow manager
  - Data structures and Parallelization library
  - Model Plugin architecture
  - **Key-Value Object Storage with Semantic Data access**
  - **IO-Server**
  - Data Notification system
  - **Data Cube API**
  - Post-Processing API

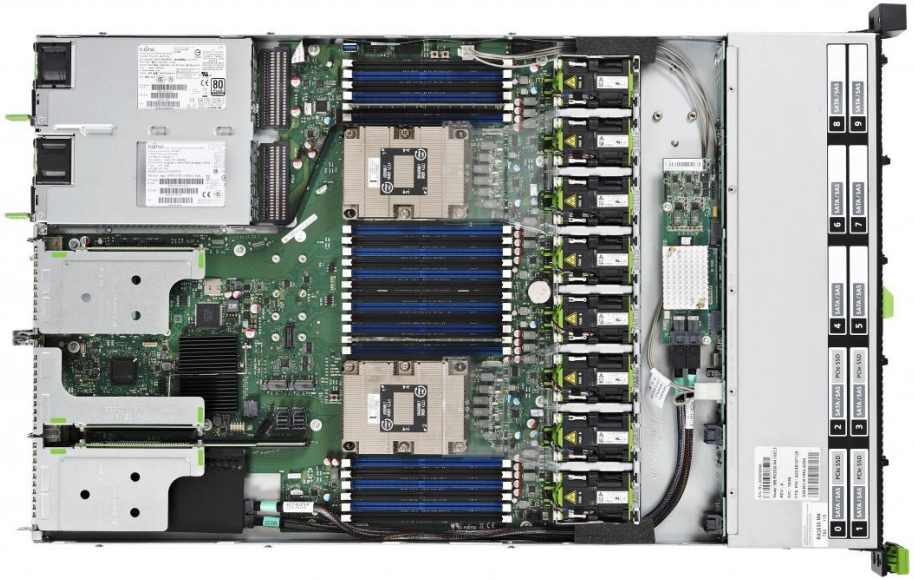
FDB – Domain Specific Object Store

- Read all @ www.nextgenio.eu (finished 2019)
- 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)**



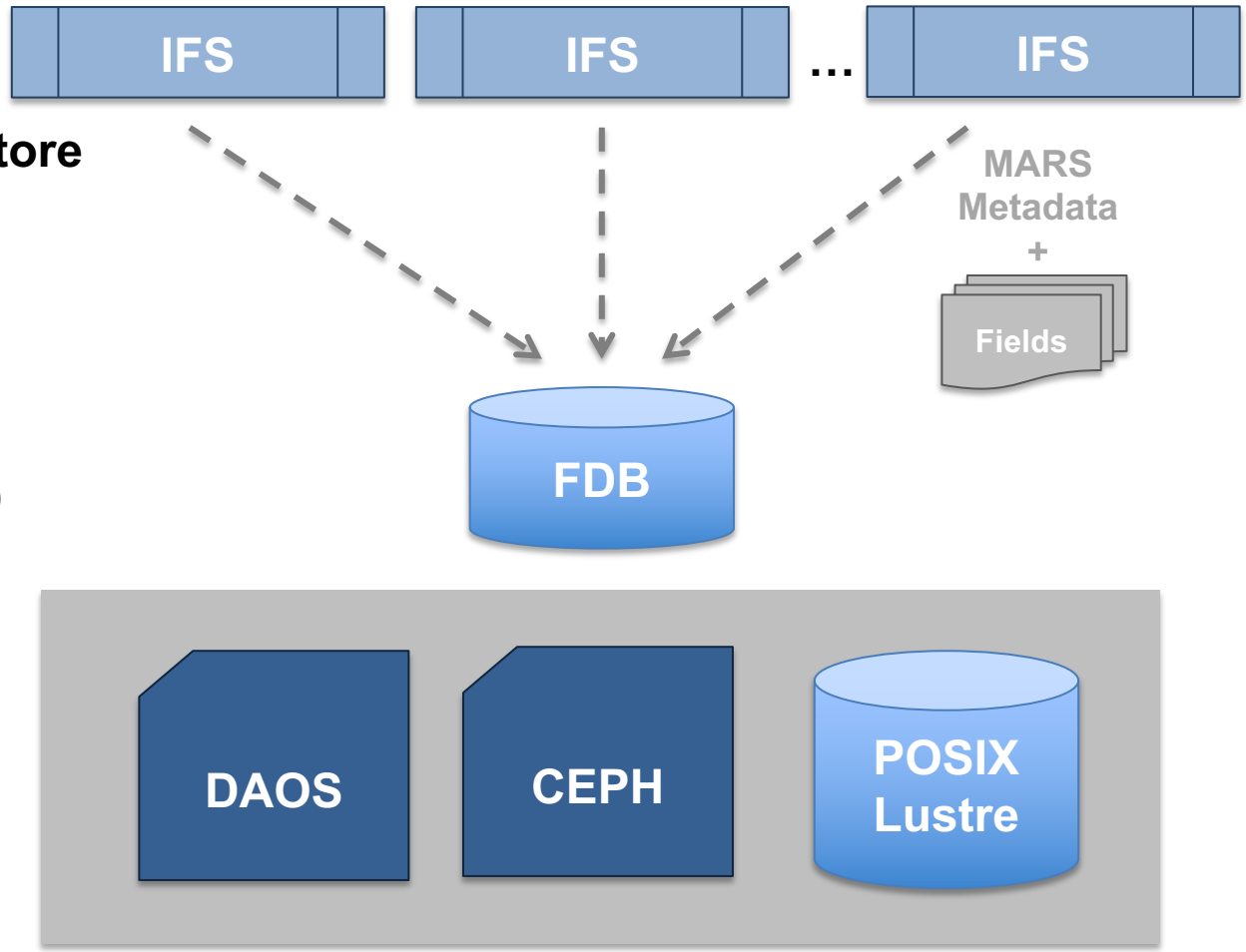
- **Prototype system**
  - **34 compute nodes**
  - **Hosted @ EPCC, Edinburgh**

FUJITSU



34 x 3 TiB Byte Addressable Storage

FDB (version 5)



- **Domain specific (NWP) Distributed object store**
- Transactional, No synchronization
- Semantic access to data
- 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)
  - Intel DAOS (under development)
  - CEPH (Cloud suited object store)
- Supports wild card searches, ranges, data conversion
  - ***In the semantic/scientific language of the user:***

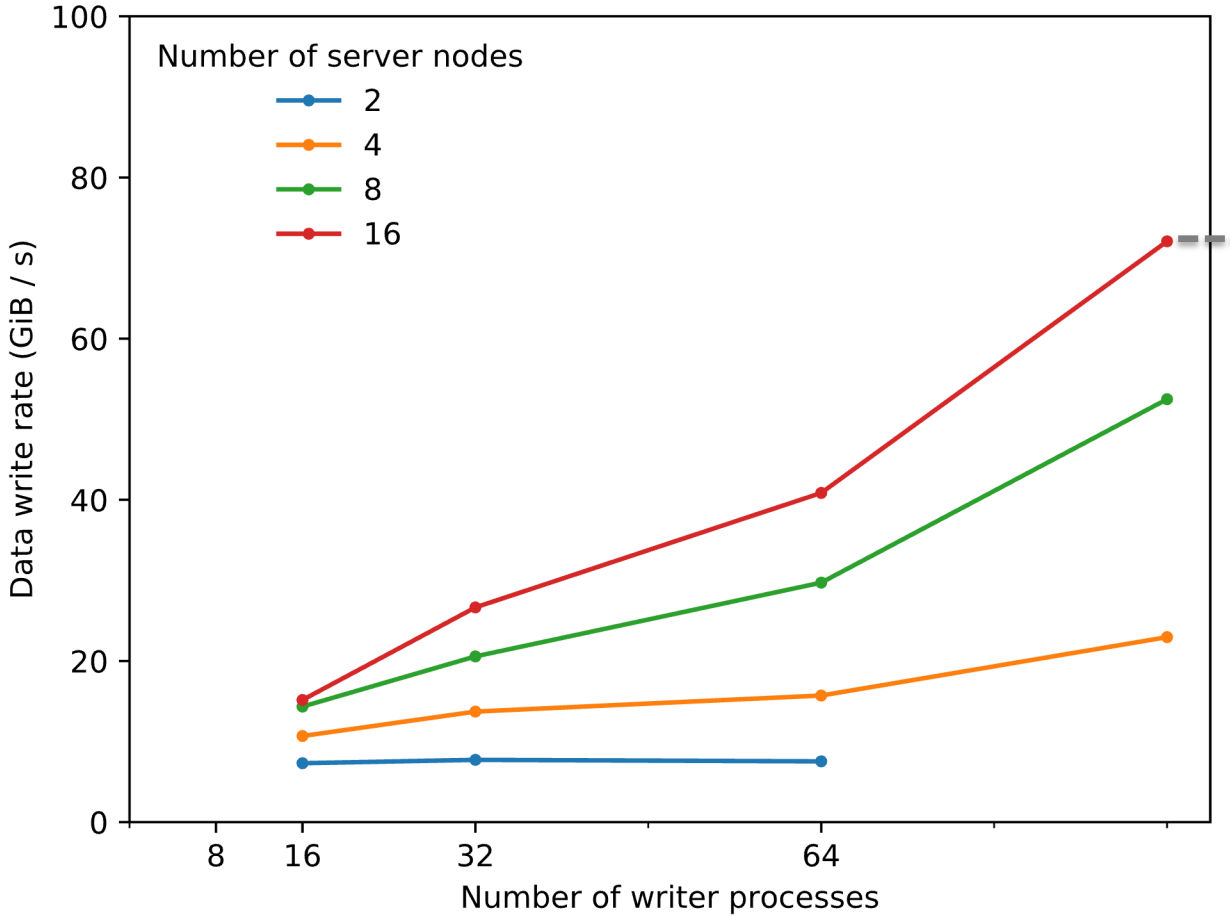
```
param=temperature/humidity,
levels=all,
steps=0/240/by/3
date=01011999/to/31122015,
```

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***

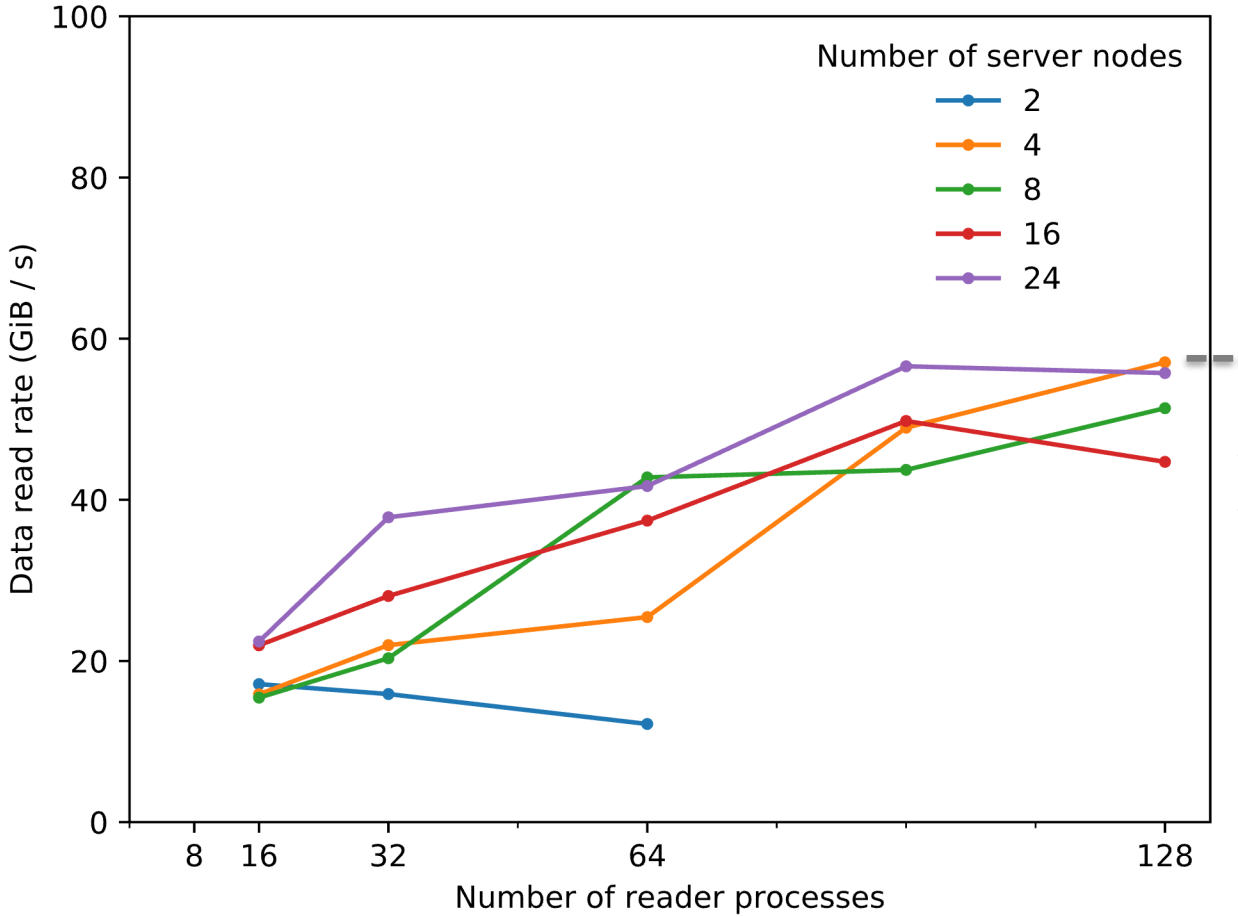
- All I/O operations are asynchronous, so computation can continue
- Distributed to all servers using a ***Rendezvous Hash***, *so no synchronisation needed*

FDB 5 Parallel Write Performance to NVRAM DIMMs



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

FDB 5 Parallel Read Performance to NVRAM DIMMs



60 GiB/s

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

FDB 5 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!!!

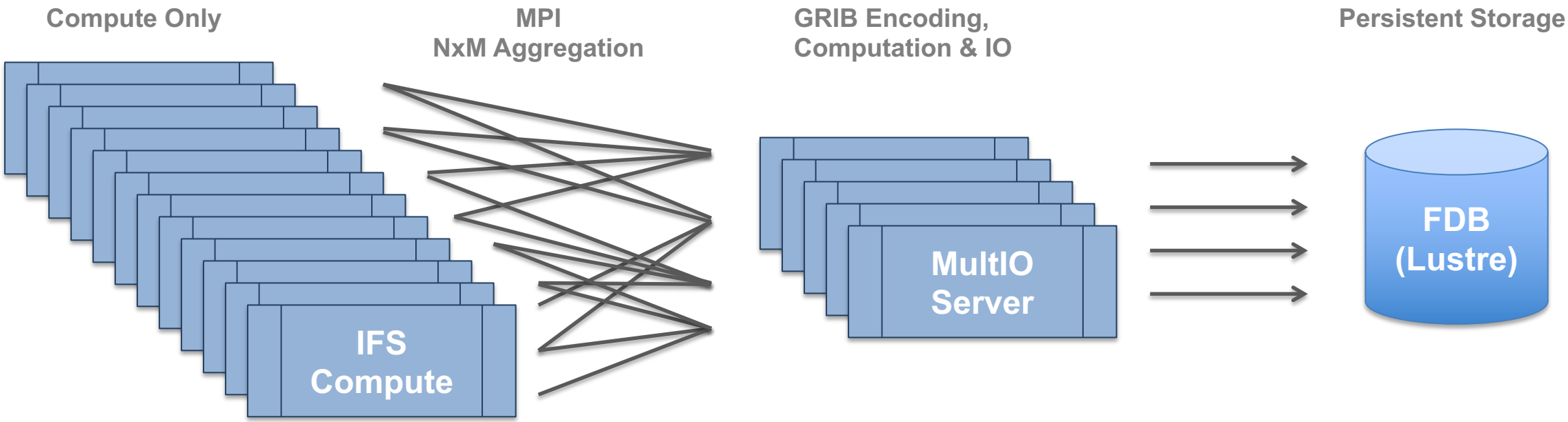
NextGenIO prototype. 32 nodes
Intel OmniPath2 interconnect
6 ensemble members

- Currently operational @ ECMWF
- Open sourced: github.com/ecmwf/fdb
- Read about it:

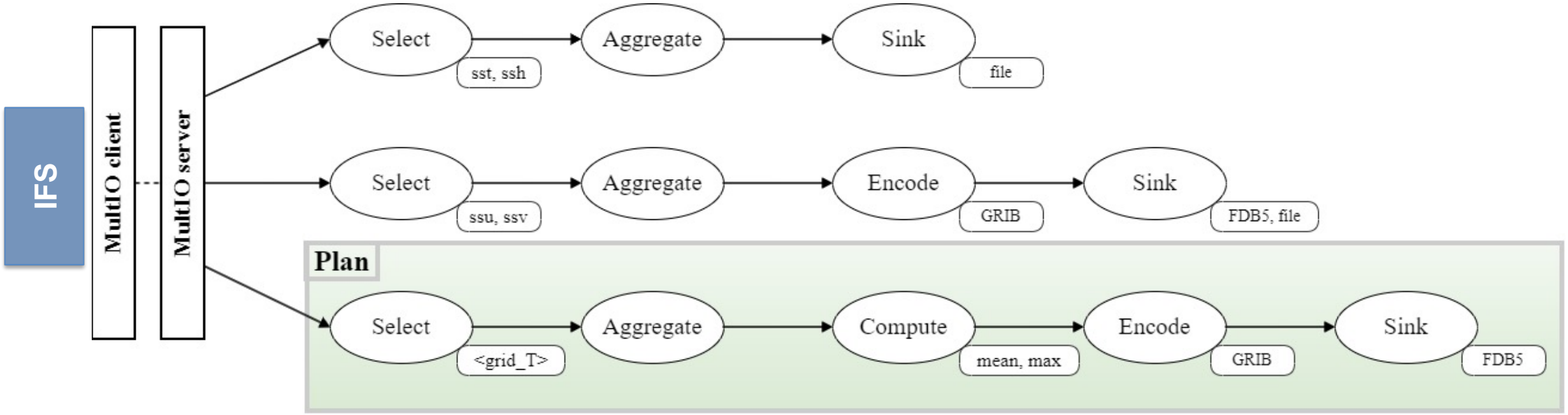
A High-Performance Distributed Object-Store for Exascale Numerical Weather Prediction and Climate
https://doi.org/10.1145/3324989.3325726

MultIO Server

- Currently under development: github.com/ecmwf/multio
- Completed adaptation of NEMO v4 model

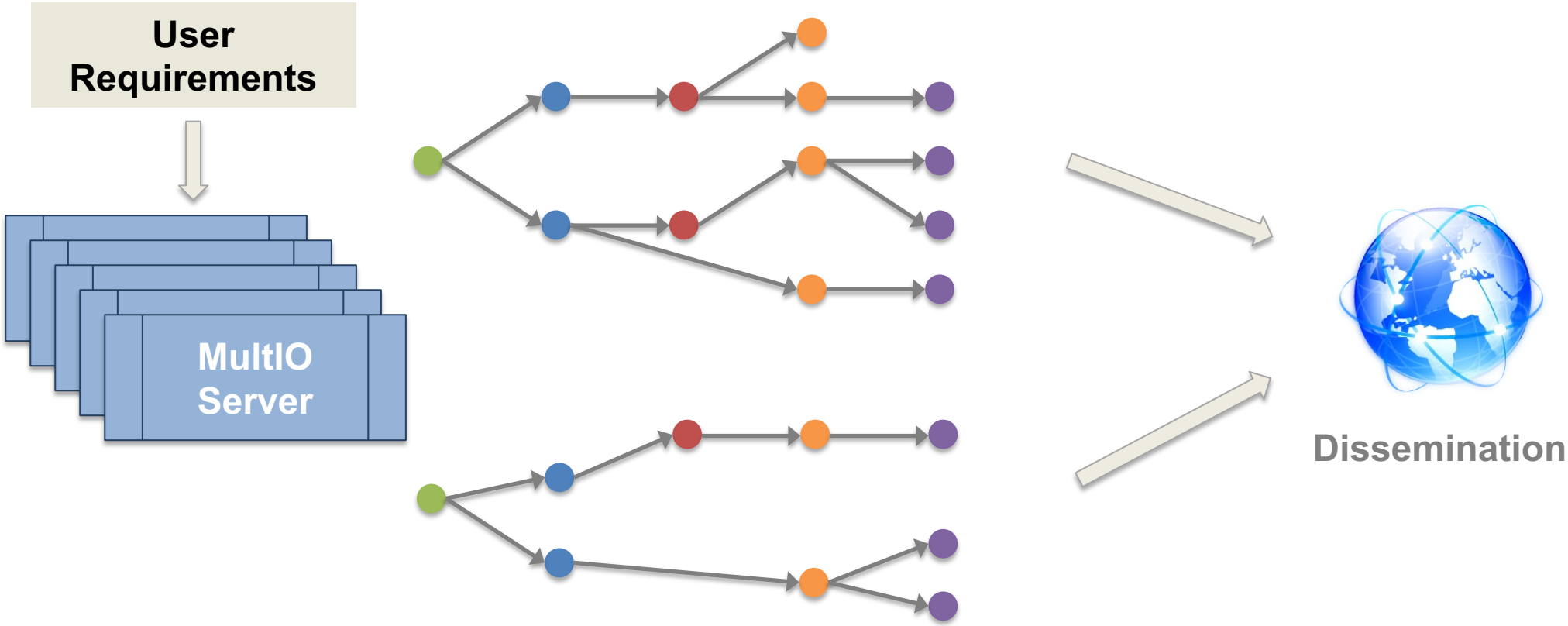


MultIO Server - Programmable Pipeline



- A generic I/O-server, user-programmable pipeline of actions
- Messages that contain **Fields** are passed to the **Plans**
- Messages are ***routed*** along multiple pipelines
- Easily extendable to new domains & grids & models

MultIO Server – On-the-fly Product Generation



All this in-memory, where the data is located

Accessing Data as Hypercubes

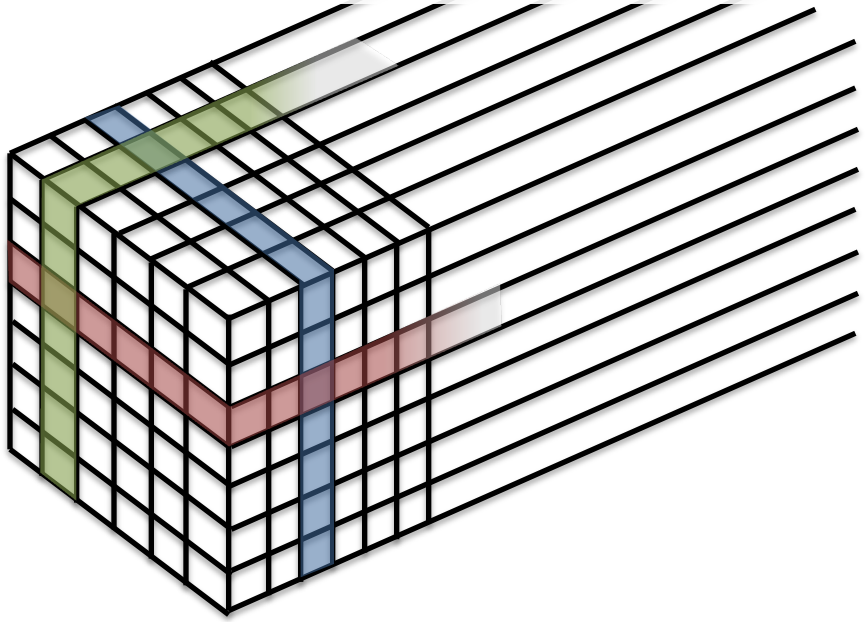
Hypercubes (6D)

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

@ double precision

- 16km **80 TiB**
- 9km **235 TiB**
- 5km **690 TiB**

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



ECMWF Novel Data-Centric Workflows

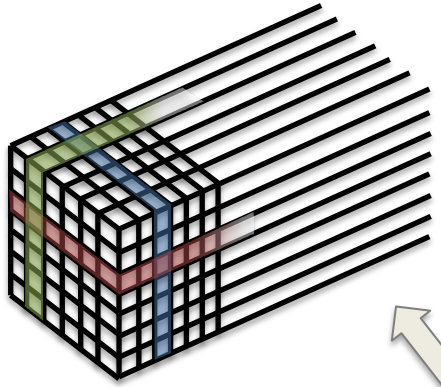
Data Analytics / Machine Learning

HPC Model Producer

Data Consumer

Cloud

- Bring **users** to the data
- Use data while it is **hot**



Object Store
Data Hypercube

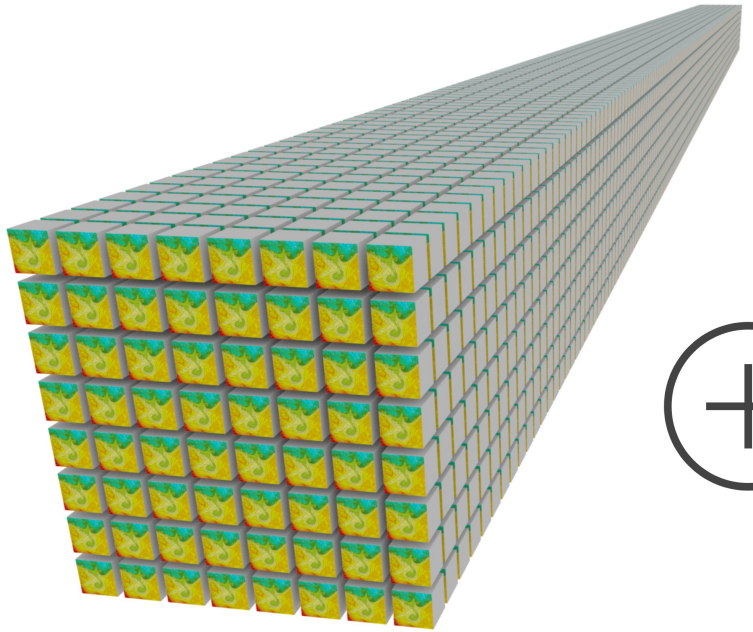
Parallel FS

Data Archive

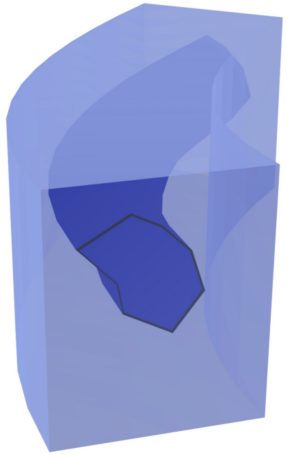
HDD

Tape

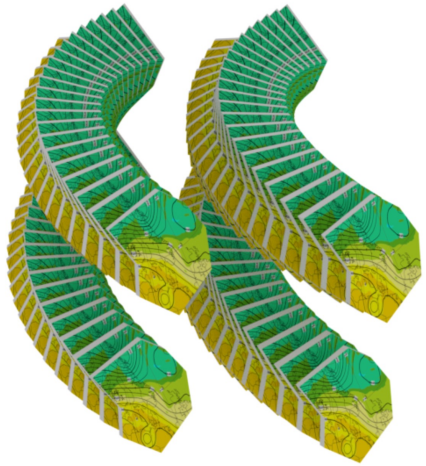
Hypercube Feature Extraction at Scale (~1 PiB)



Real-time Model Output
(Datacube on FDB)



Polytope Query



Resultant Data
Extraction

Polytope (under development)

Messages To Take Home

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

*ECMWF is engaged in building **Km-scale Digital Twin** for Weather Extremes*

*ECMWF is adapting to **data centric workflows** for Exascale weather forecasting, exploring **in-situ data analysis** and **hypercube feature extraction***

ECMWF is refactoring software stack end-to-end to enable Exascale datasets in Weather Forecasting

Work partially funded by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreements 825532 (LEXIS), 801101 (MAESTRO) and 955648 (ACROSS)