On the Convergence of HPC, Cloud and Data Analytics for Exascale Weather Forecasting

ECMWF Present and Future

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

ECMWF

tiago.quintino@ecmwf.int

19th ECMWF Workshop on HPC in Meteorology

ECMWF

ECMWF's Forecasting Systems

Established in 1975, Intergovernmental Organisation

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

24/7 operational service

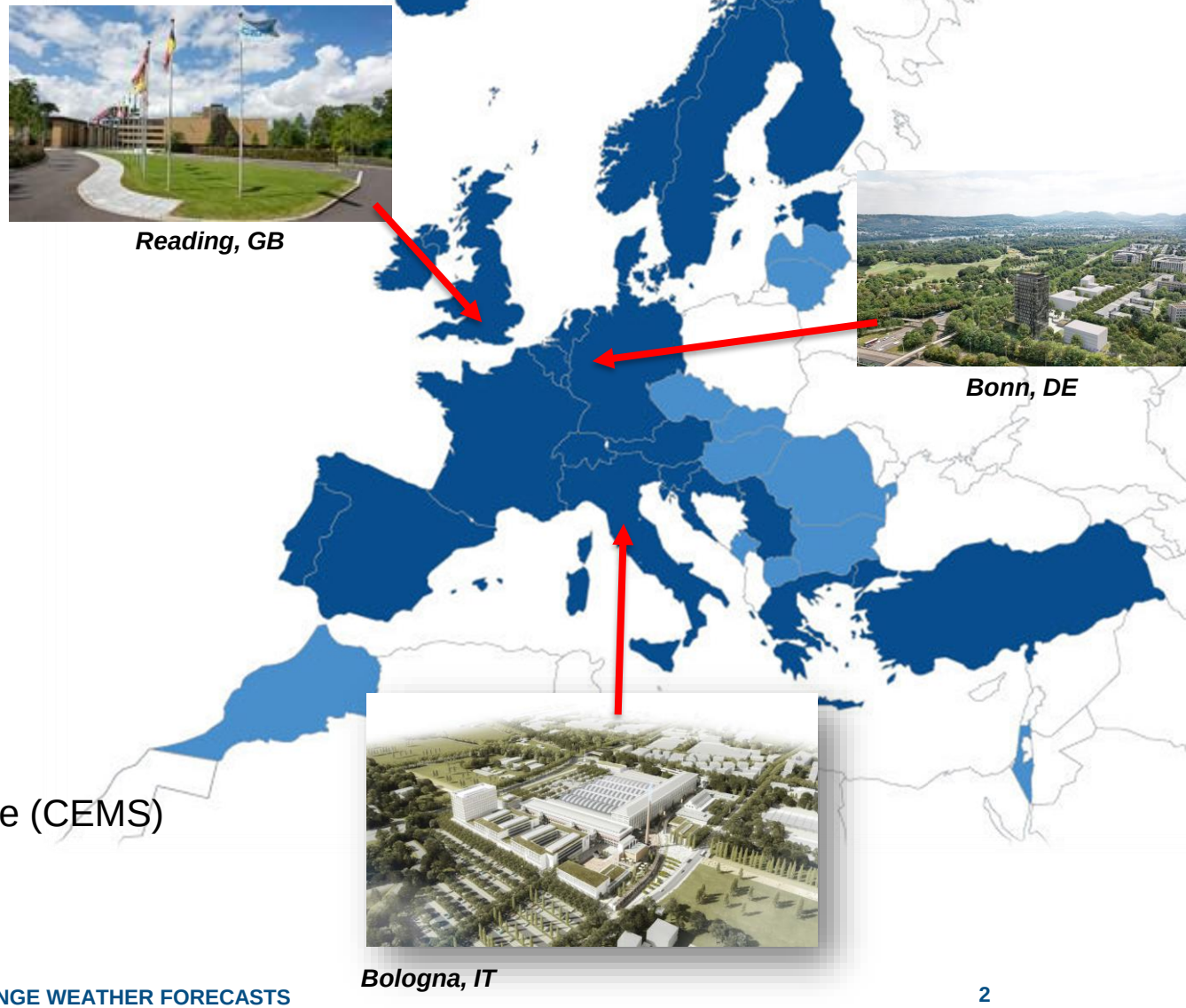
- Operational NWP – 4x HRES+ENS forecasts / day
- Supporting NWS (coupled models) and businesses

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)

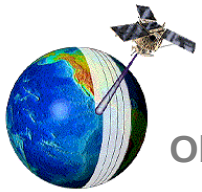


Reading, GB

Bonn, DE

Bologna, IT

ECMWF's Production Workflow



Global Observations

200M obs/day

120TiB/day

45TiB/day

~350 world destinations

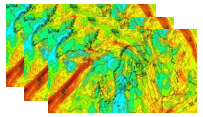
Acquisition → Earth System Model → Product Generation → Product Dissemination → Member States & Customers

Obs

Fields

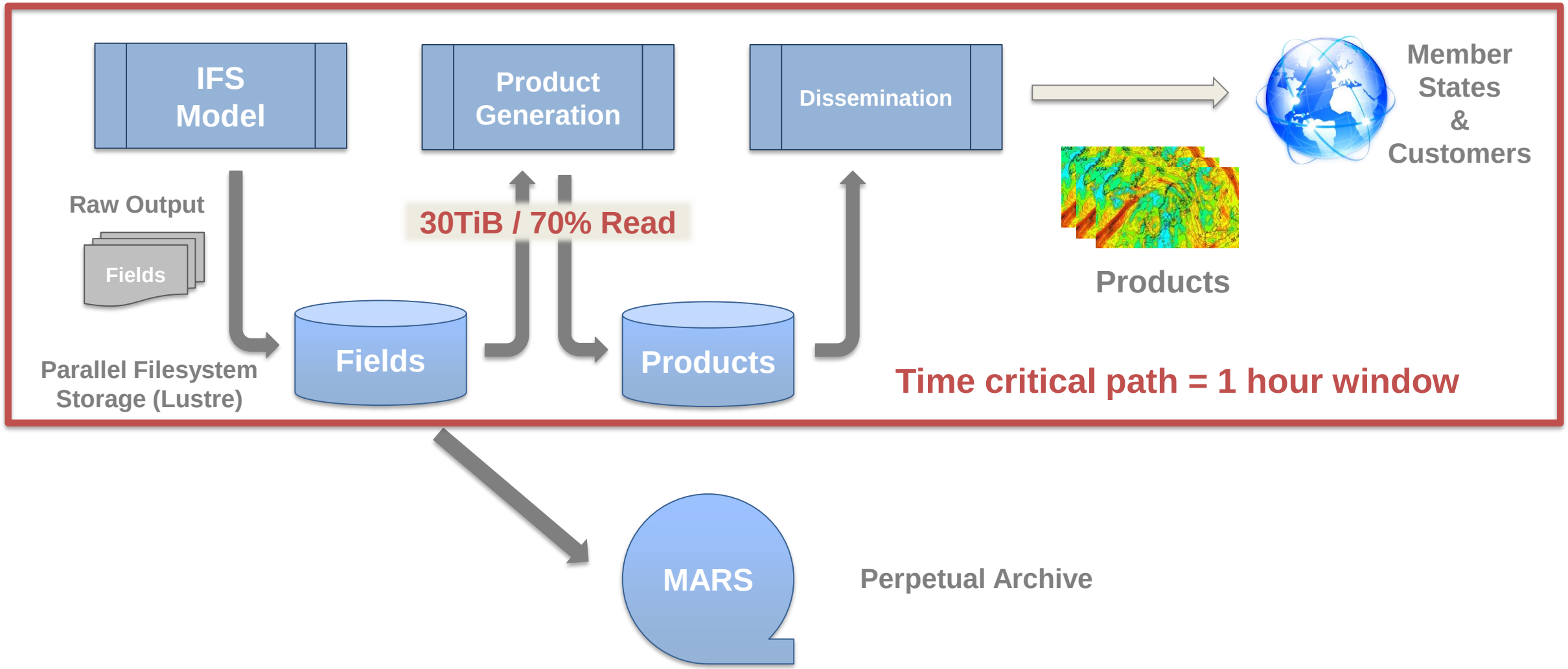
Archive

Perpetual Archive



Products

ECMWF's Production Workflow - Challenges



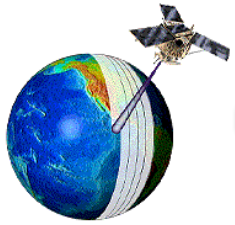
Effects of Product Generation using Parallel Filesystem

	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

Archive

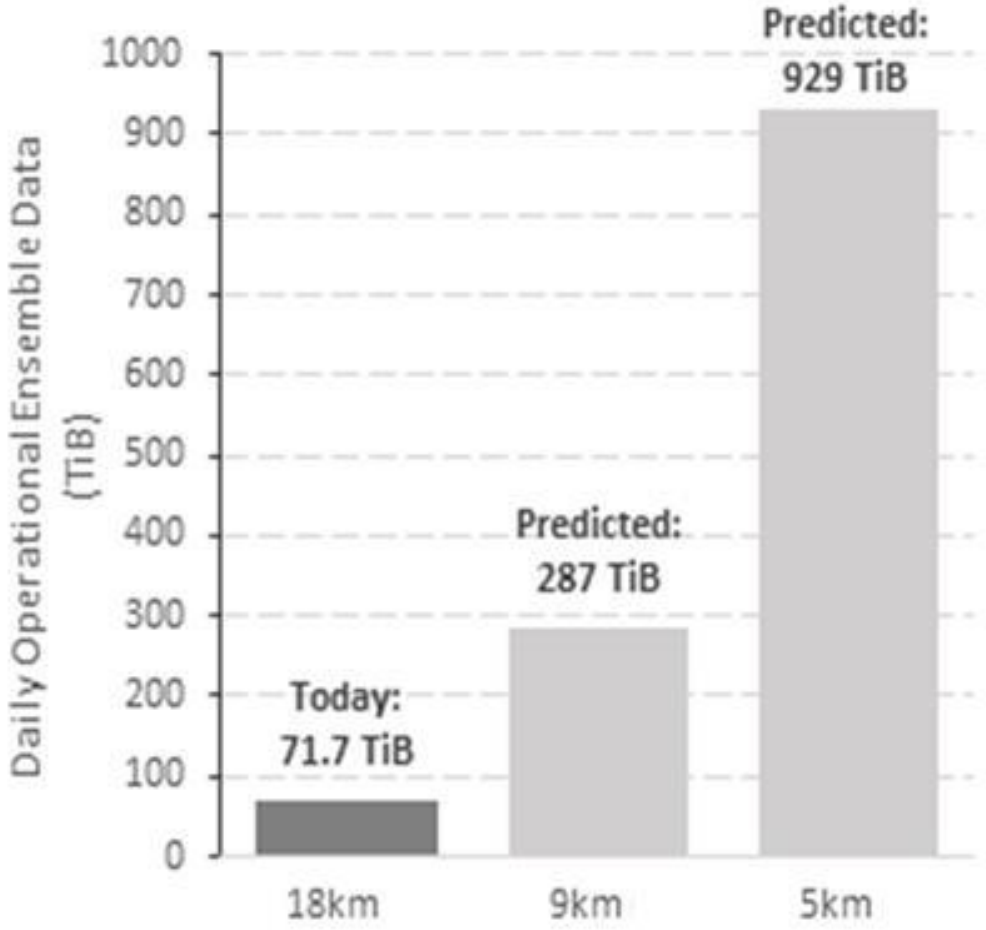
Modify

Product Generation

Perpetual Archive

Data is Central!

Data Growth – History and Projections



Model Output Projected Growth
40% compound yearly

Historical Growth of Disseminated Products

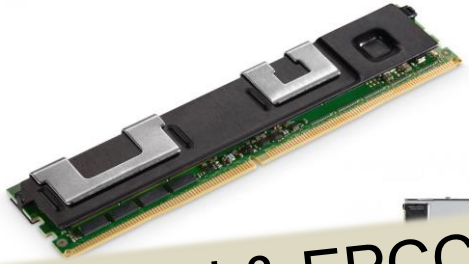
How large is a 1.25 km ensemble forecast?

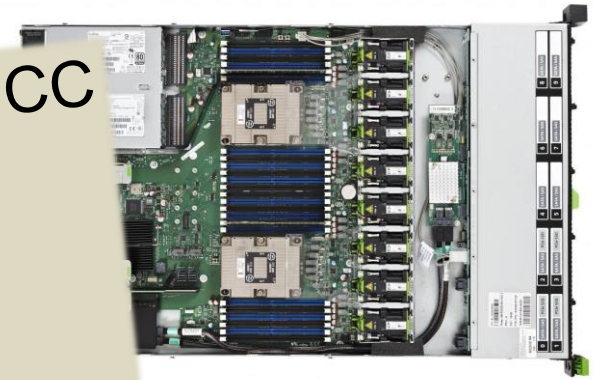
- **50-member ensemble forecast**
- ***Compressed* GRIB2 data @ 16bit & 24bit**
- **@ 9km O1280** **25 TiB**
- **Resolution @ 5km O1280 → O1999** **x 4**
- **Upgrade levels 137 → 200** **x 1.5**
- **Resolution @ 2.5km O1999 → O3999** **x 4**
- **Resolution @ 1.25km O3999 → O7999** **x 4**

25 TiB x 96 = 2400 TiB

FDB – Domain Specific Object Store

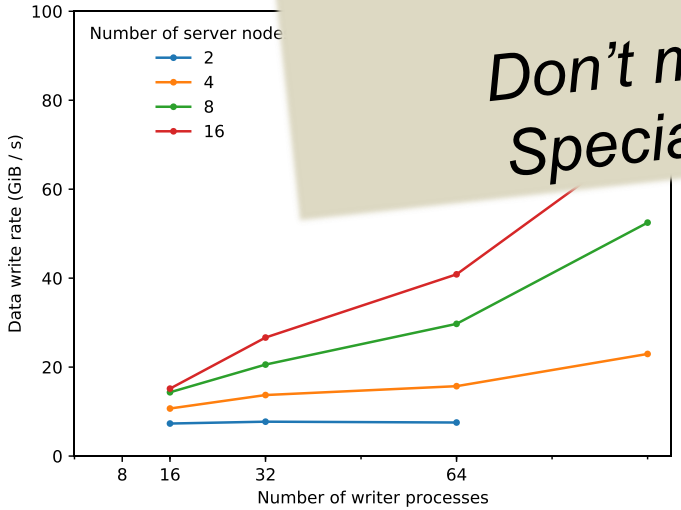
- **Domain specific (NWP) Distributed object store**
- Transactional, No synchronization
- Semantic / Scientific access to data
- Key-value store
- Support f

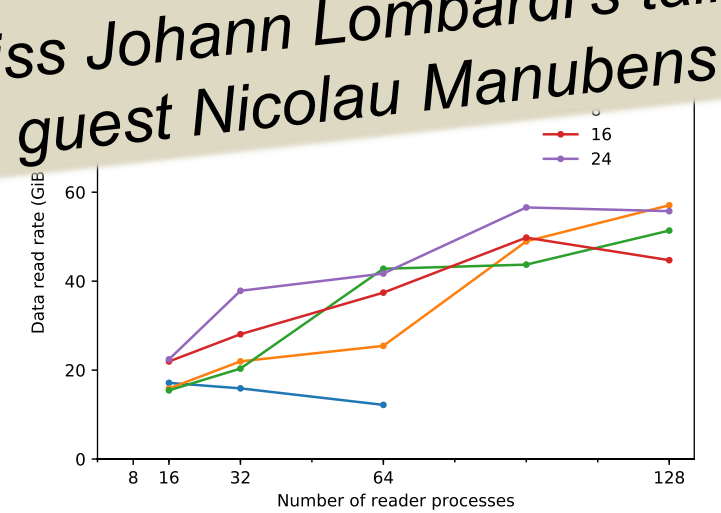




Currently developing an **FDB backend** with Intel & EPCC based on **DAOS & Optane**

Don't miss Johann Lombardi's talk on Friday
Special guest Nicolau Manubens (ECMWF)





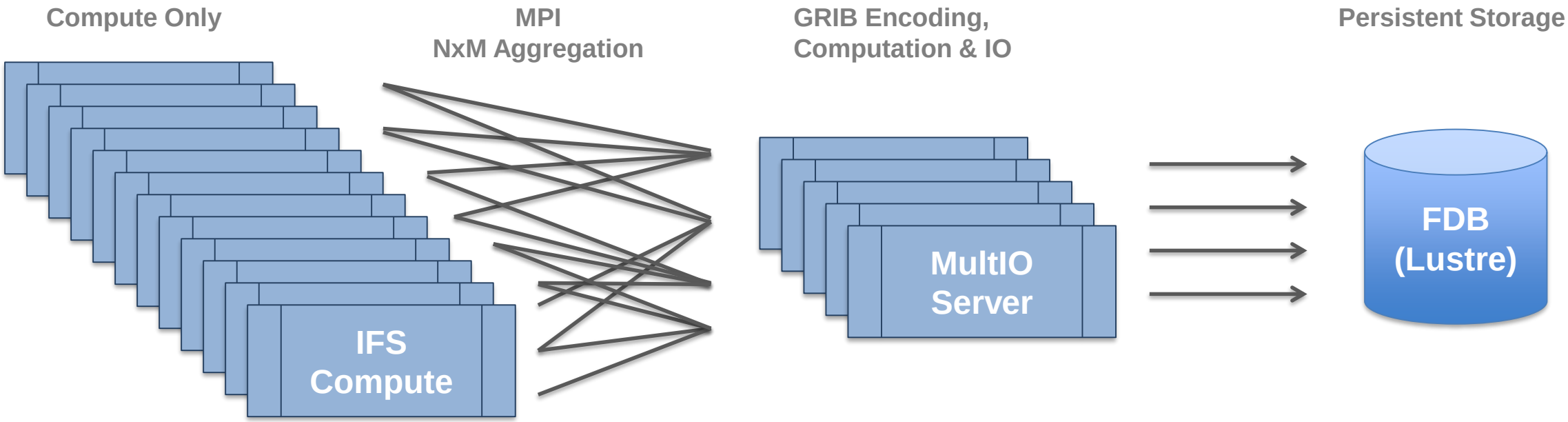
24 Nodes

- **3TiB NVMe DIMMs / Node**
- **Application data measured**
- **Full consistency semantics**

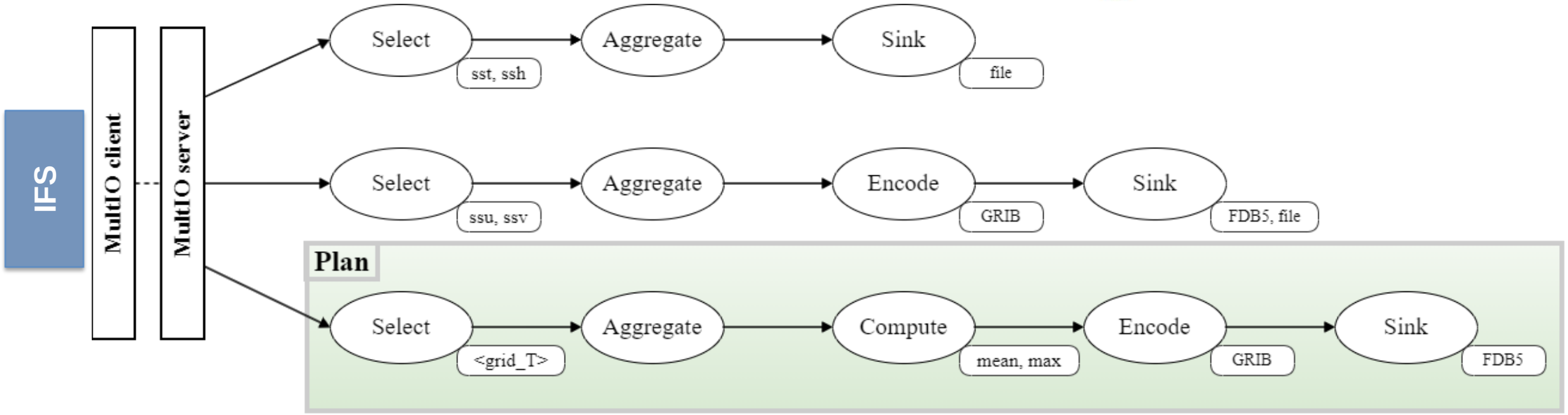
A High-Performance Distributed Object-Store for Exascale Numerical Weather Prediction and Climate
https://doi.org/10.1145/3324989.3325726

MultIO Server

- Currently under development
- Completed adaptation of NEMO v4 model

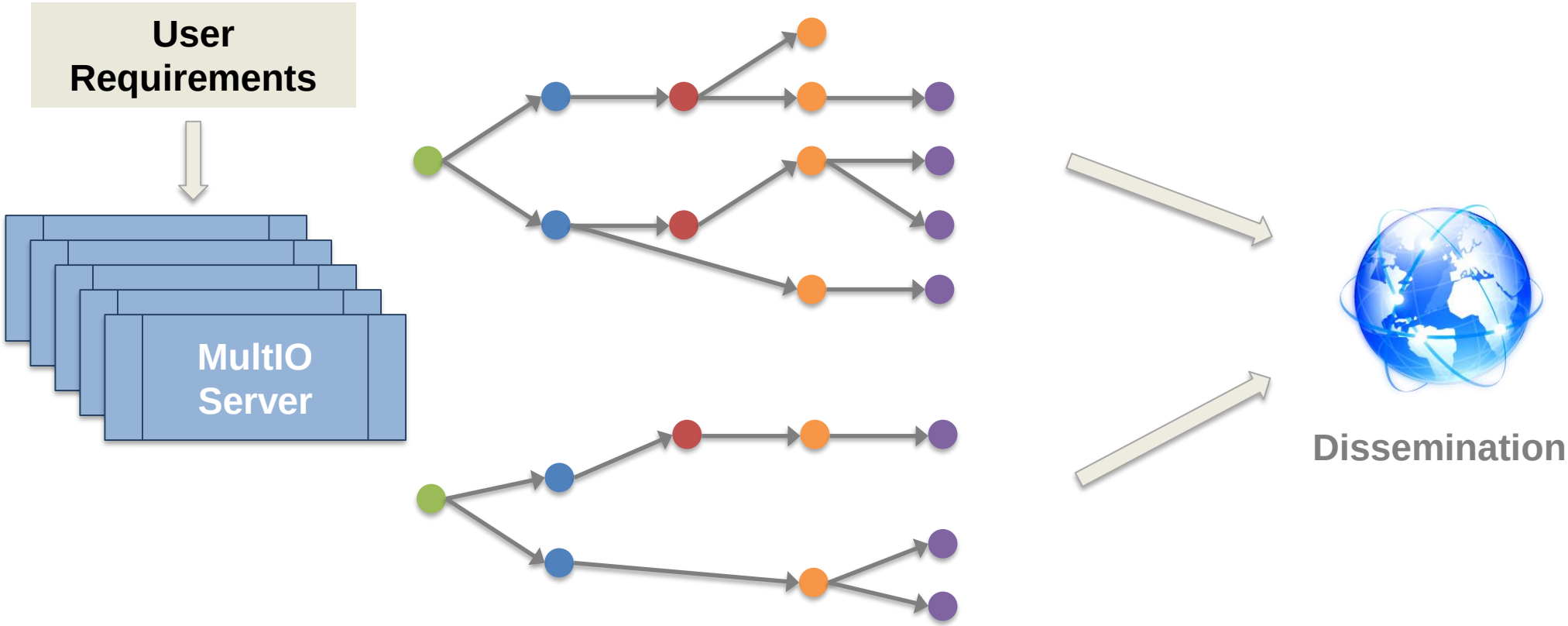


The MultIO Programmable Pipeline



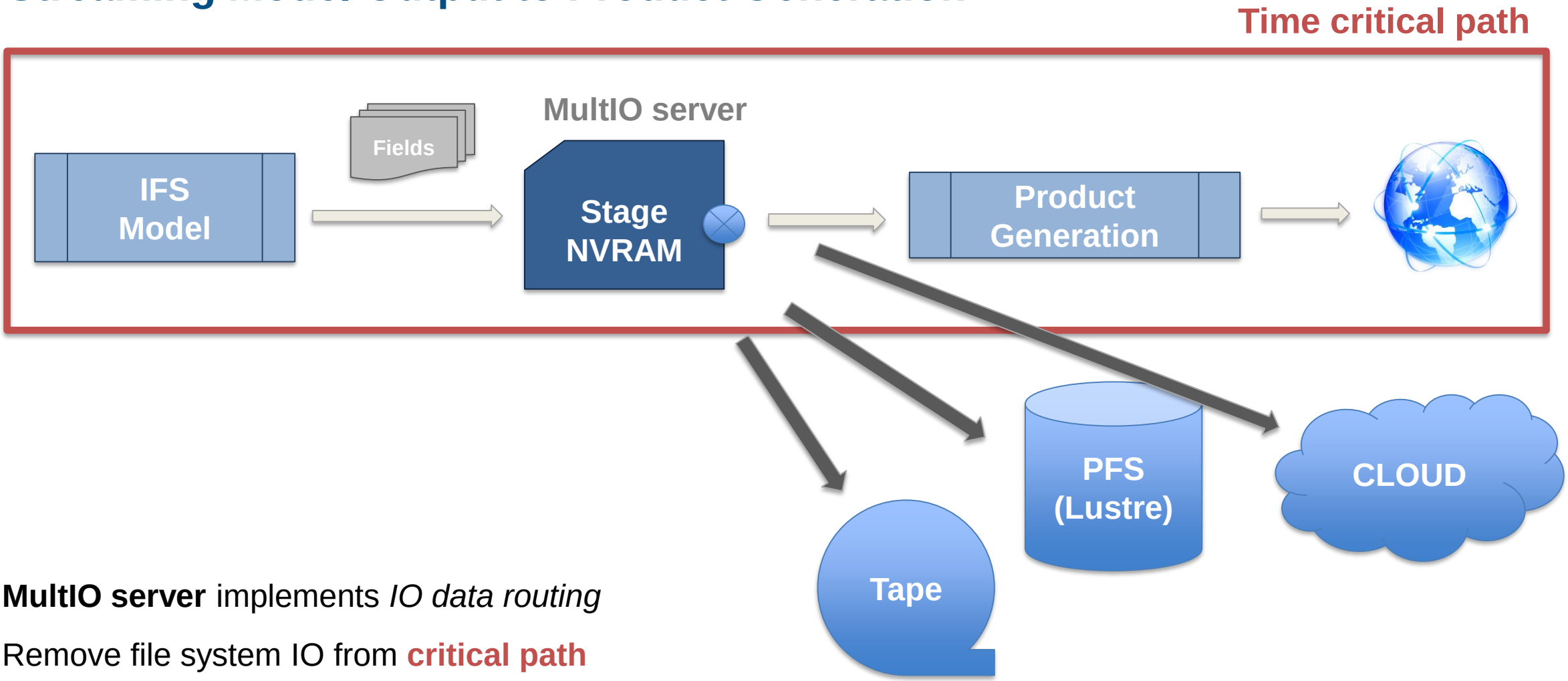
- A generic I/O-server, user-programmable pipeline of actions
- Messages that contain **Fields** are passed to the **Plans**
- Messages are ***routed*** along multiple pipelines
- Easily extendable to new domains & grids & models

On-the-Fly (or In-Place) Product Generation



All this in-memory, where the data is located

Streaming Model Output to Product Generation



MultIO server implements *IO data routing*

Remove file system IO from **critical path**

Product Generation **inside MultIO** with in-situ post-processing

Impacts of NVRAM on Data Access

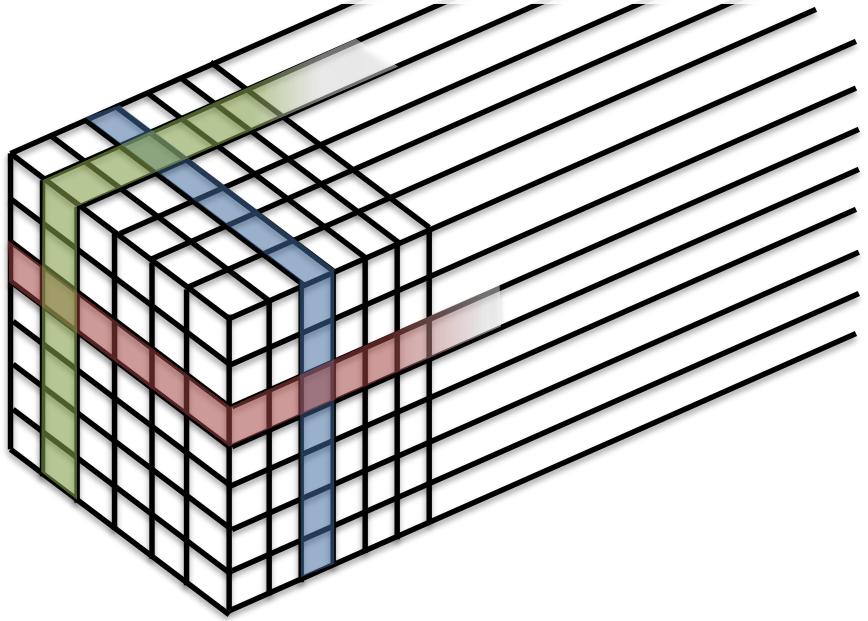
Byte Addressable Hypercubes (6D)

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

@ 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

Producer

Cloud

Consumer

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

FDB
Data Hypercubes

PFS

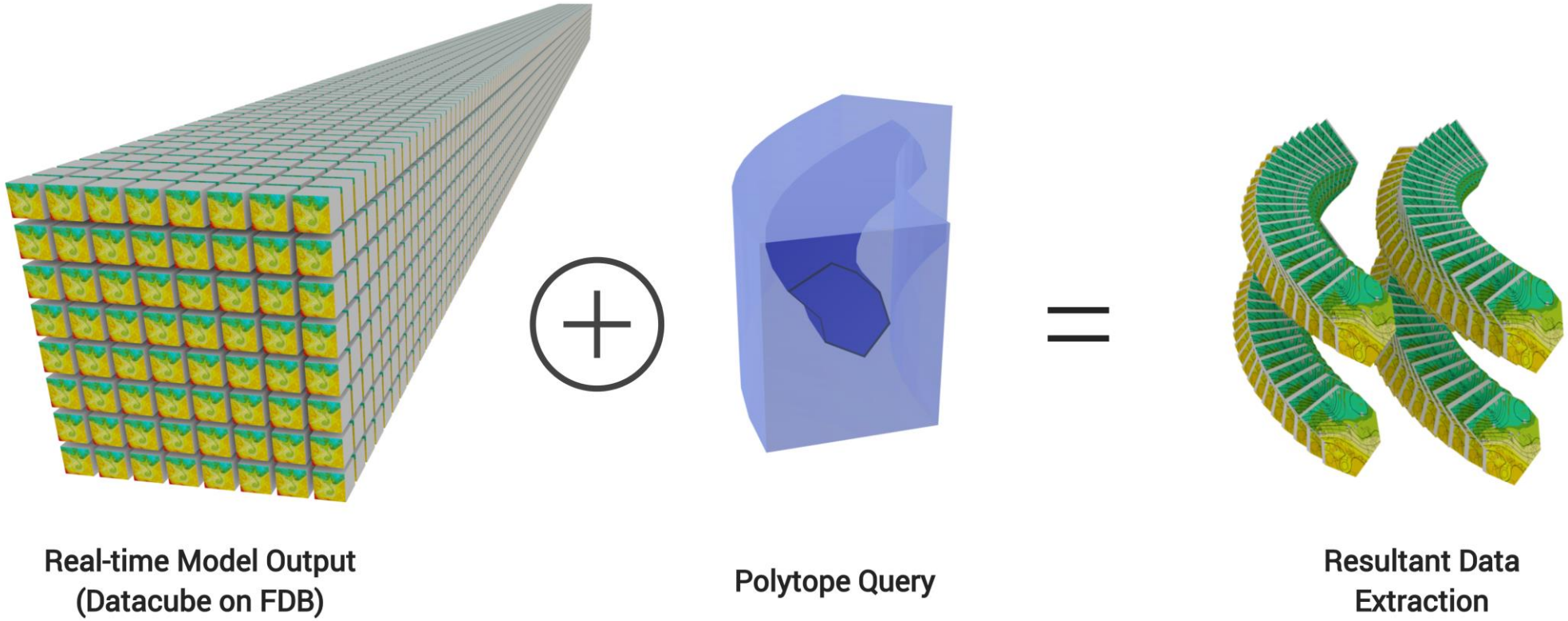
Archive

HDD

Tape

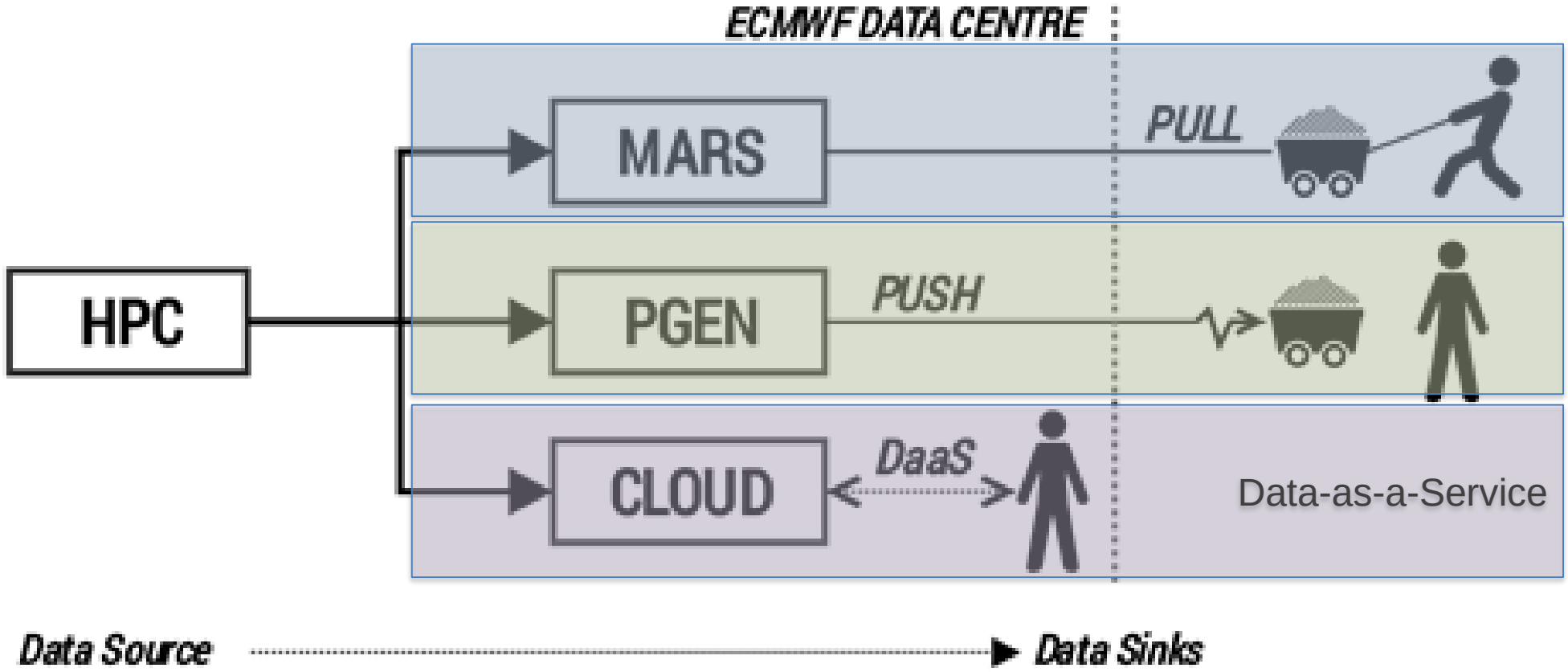
MARS

Datacube Access at Scale (~1 PiB)



Polytope (under development): http://polytope.ecmwf.int/openapi

Novel Data Flows – Multiple Pathways to Serve Data



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 analyze data*

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

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)