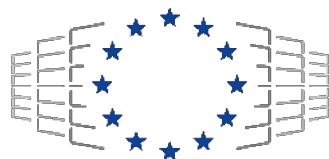


# Addressing data storage constraints in numerical weather prediction

Olivier Iffrig, James Hawkes, Simon Smart, Tiago Quintino



**EuroHPC**  
Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955811. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, the Czech Republic, Germany, Ireland, Sweden, and the United Kingdom.

# ECMWF's Forecasting Systems

## Established in 1975, Intergovernmental Organisation

- 23 Member States | 12 Cooperating 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)

## Destination Earth

- Operates two Digital Twins
- Operates the Digital Twin Engine (DTE)



Reading, GB

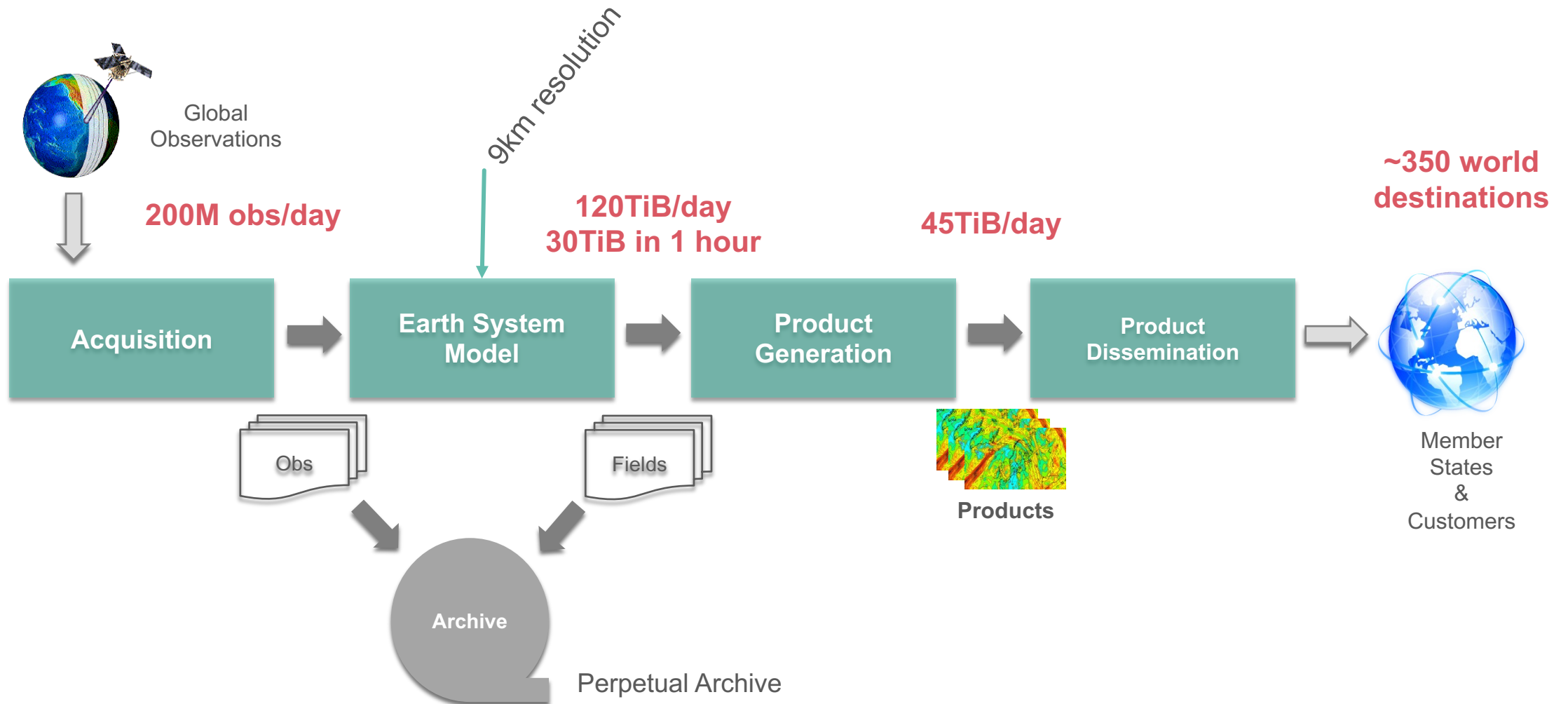


Bonn, DE

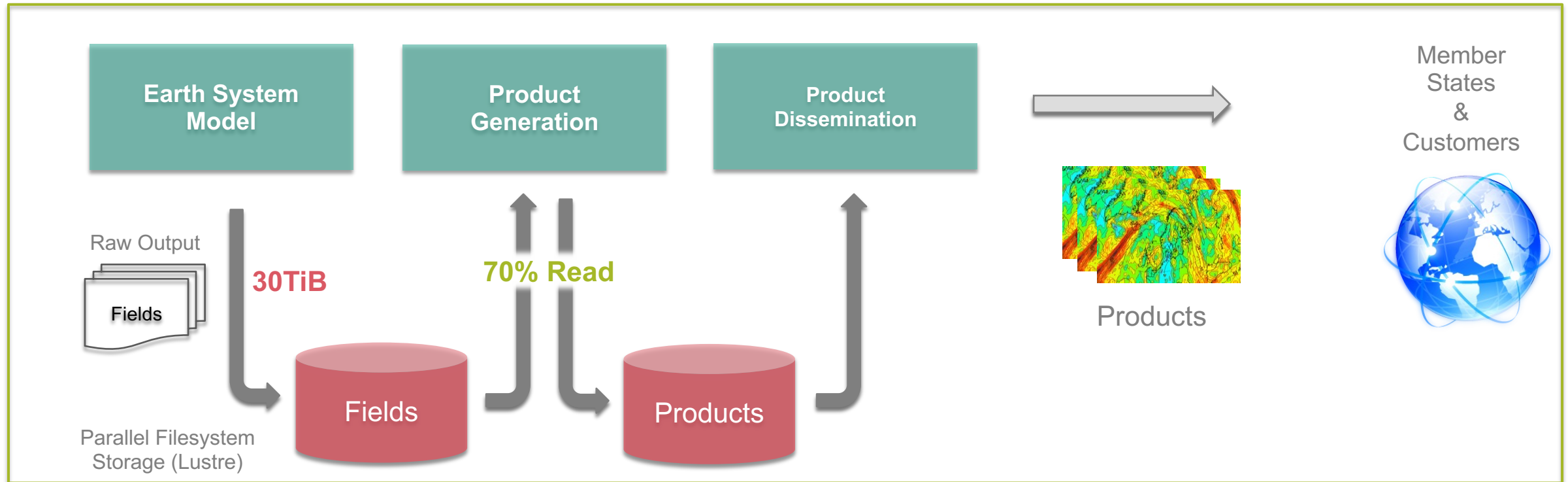


Bologna, IT

# The ECMWF operational workflow



# The ECMWF operational workflow



Time critical path = 1 hour window

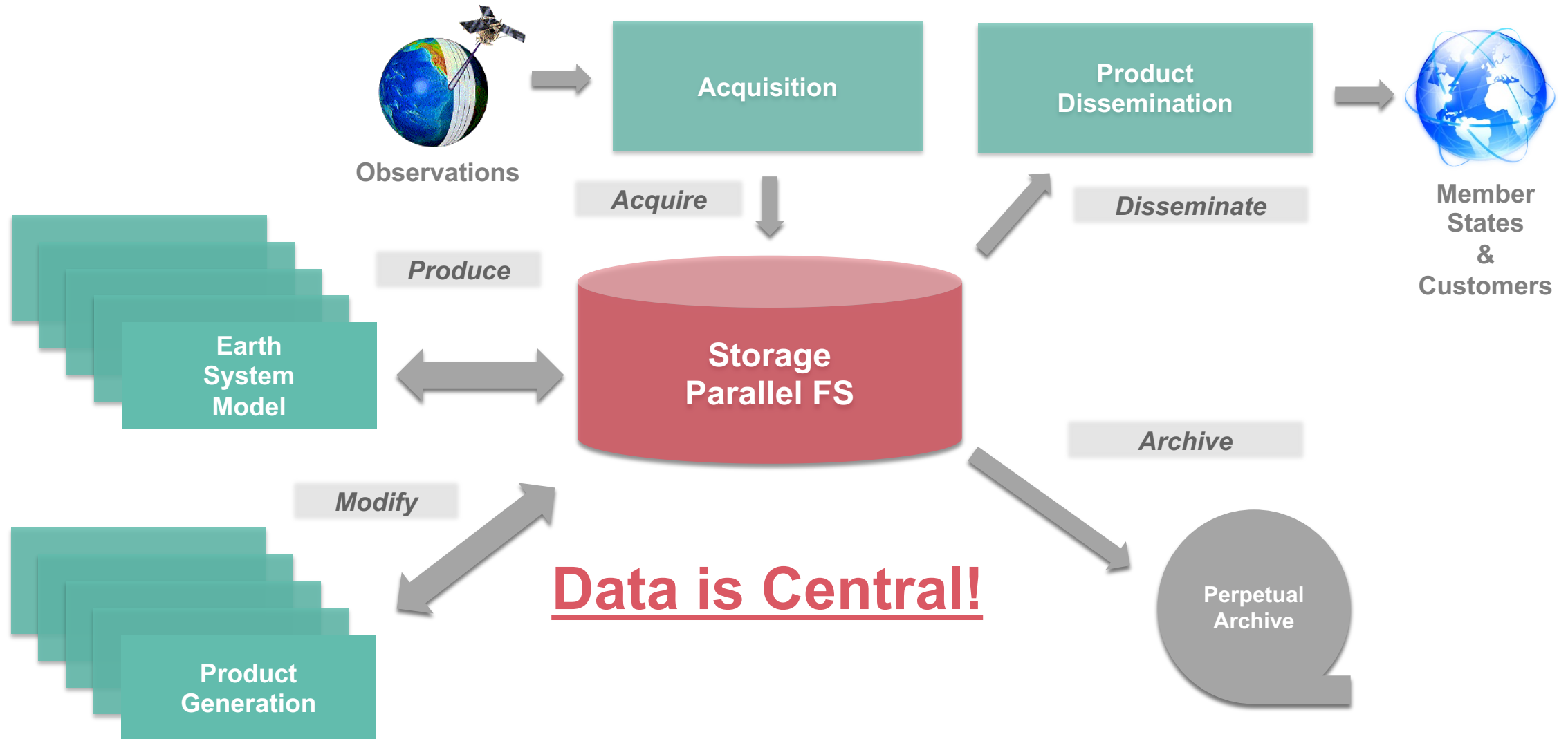
# The ECMWF operational workflow

	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%

**“Coupling” via the file system!**

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

# The ECMWF operational workflow



# Semantic Data Access

- Data is indexed by its scientific metadata, according to a hierarchical schema
- The key used to index data carries scientific meaning
  - Not just a UUID
  - Not just storing metadata with data
  - The metadata is **used to index and uniquely identify the data**
- ECMWF archive from 1975-2022 (>400PiB) is all addressed with the same data language
- Data is implicitly **discoverable and findable (FAIR)** when you know the domain schema

## ~~Non-semantic key~~

~~8s09sno5tdylonj92asy23~~

## Semantic key

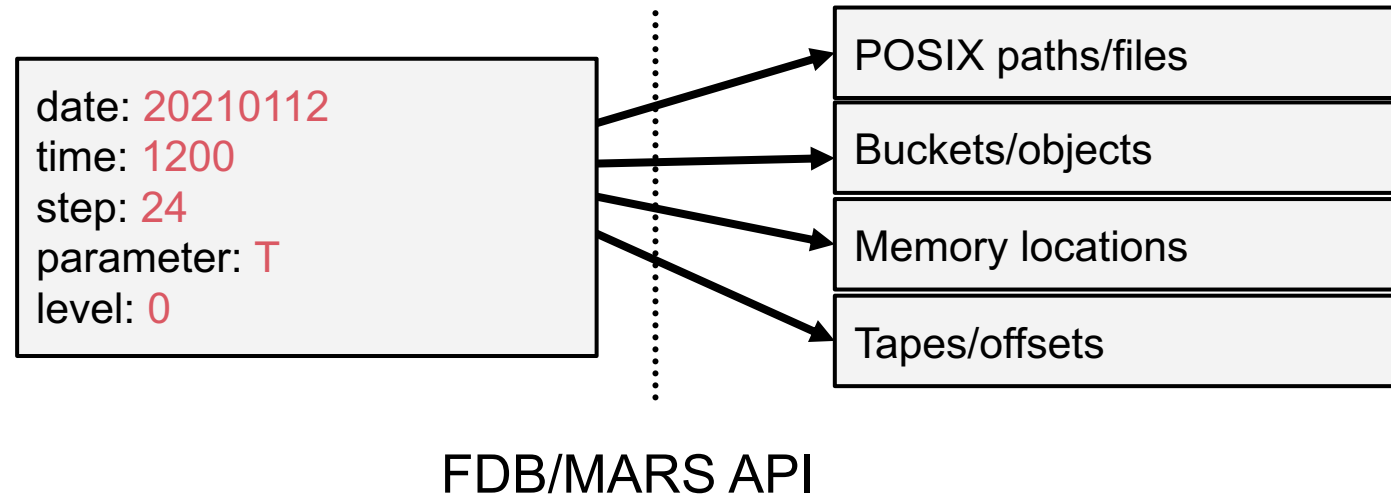
date: 20210112  
time: 1200  
step: 24  
parameter: T  
level: 0

# Semantic Data Access

- The most basic semantic data access can be done with files...

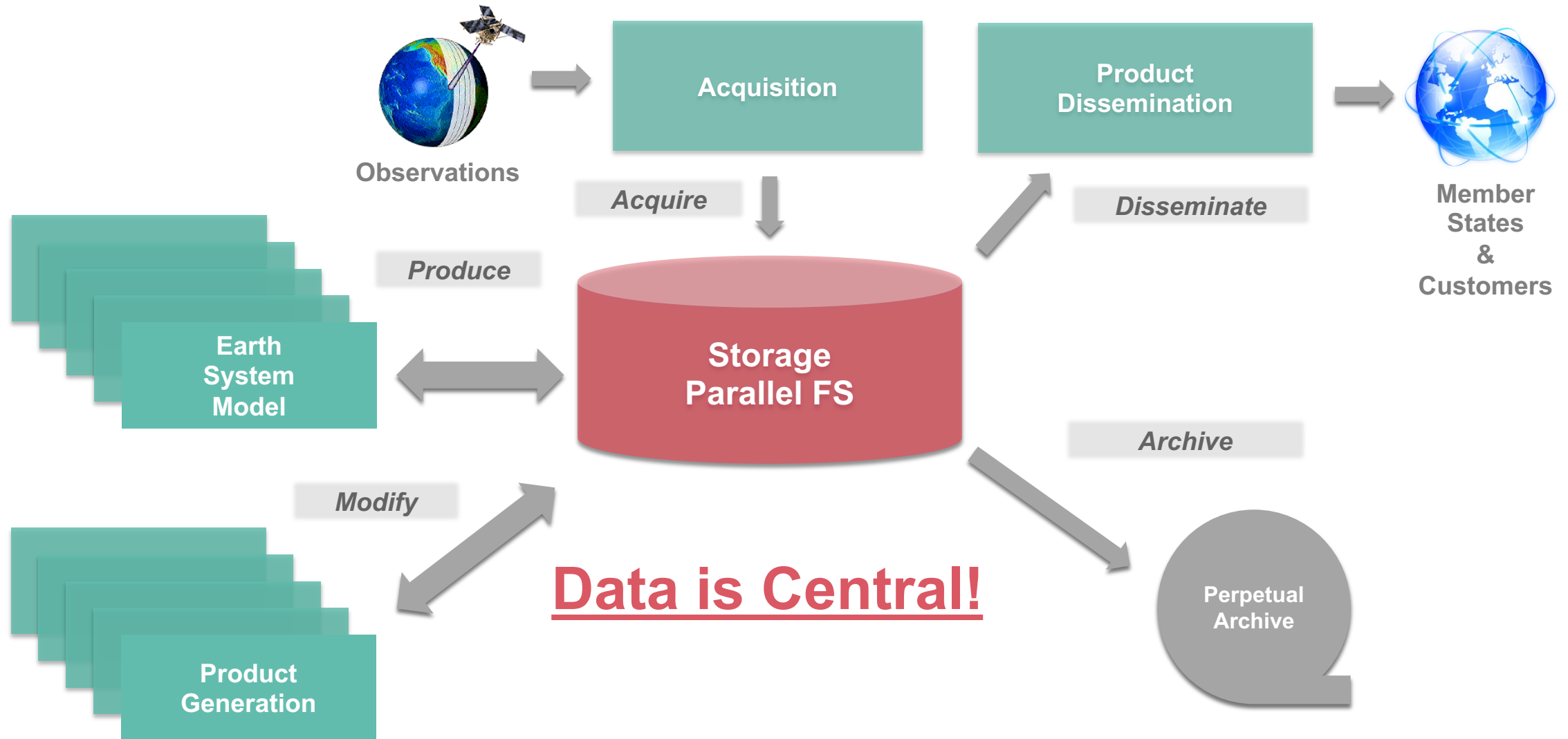
```
../20210112/1200/24/0/T/...
```

- ... but a more scalable implementation decouples the scientific identification from the storage resource

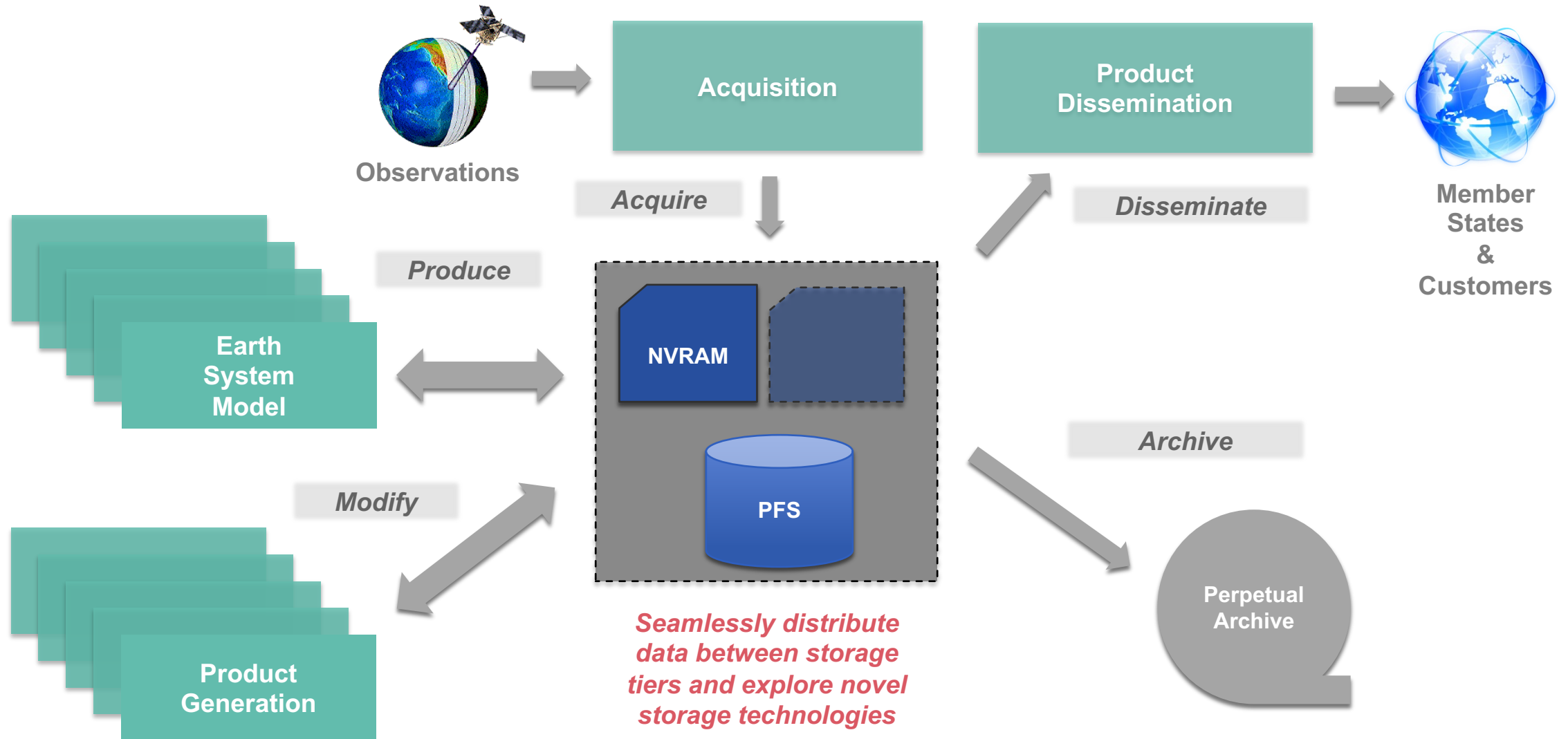


- ... and the applications don't need to care how the objects are stored.

# The ECMWF operational workflow

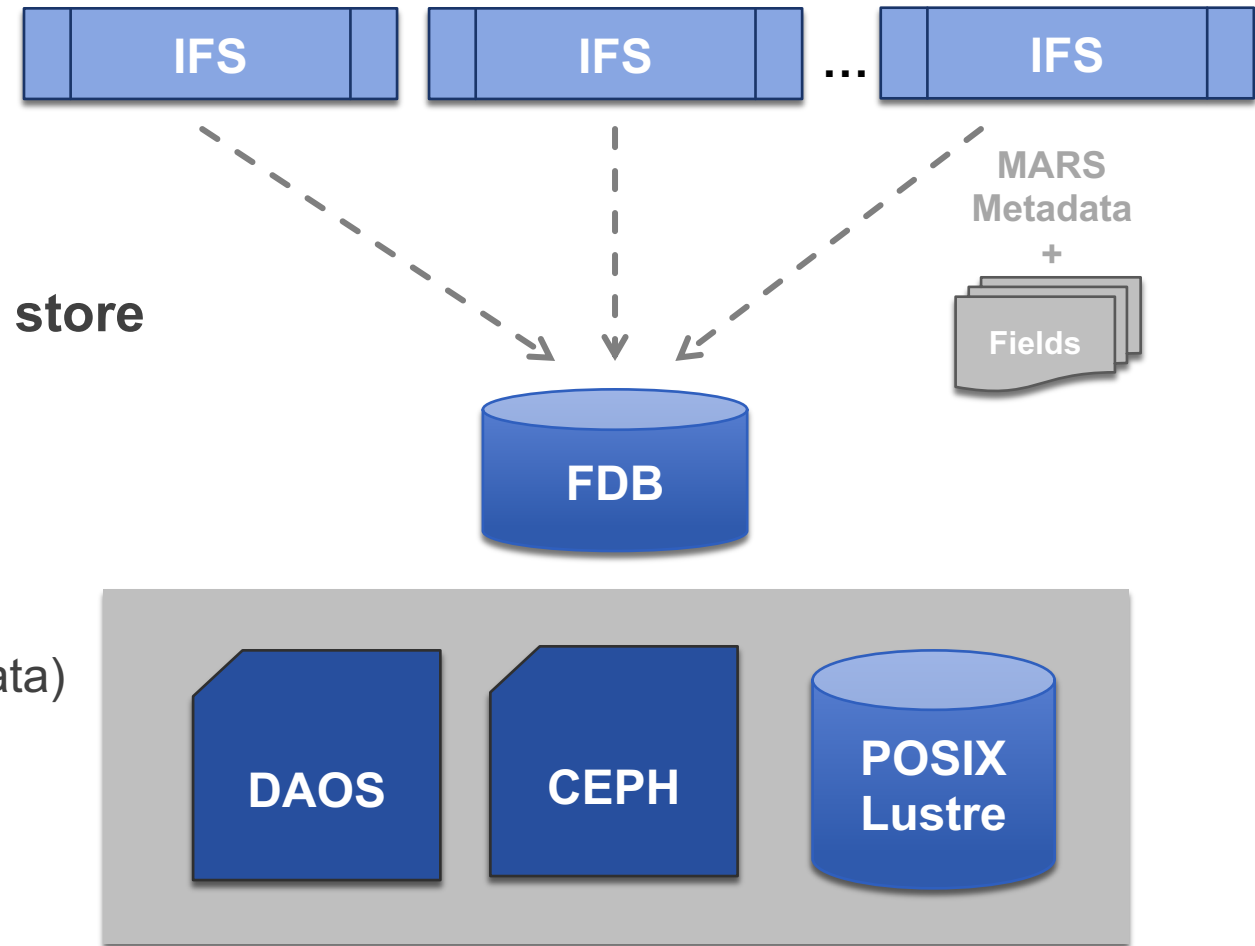


# The ECMWF operational workflow

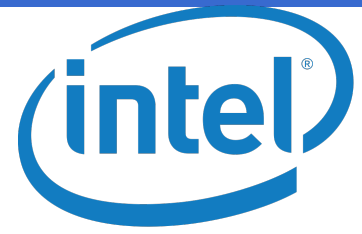


# FDB (version 5)

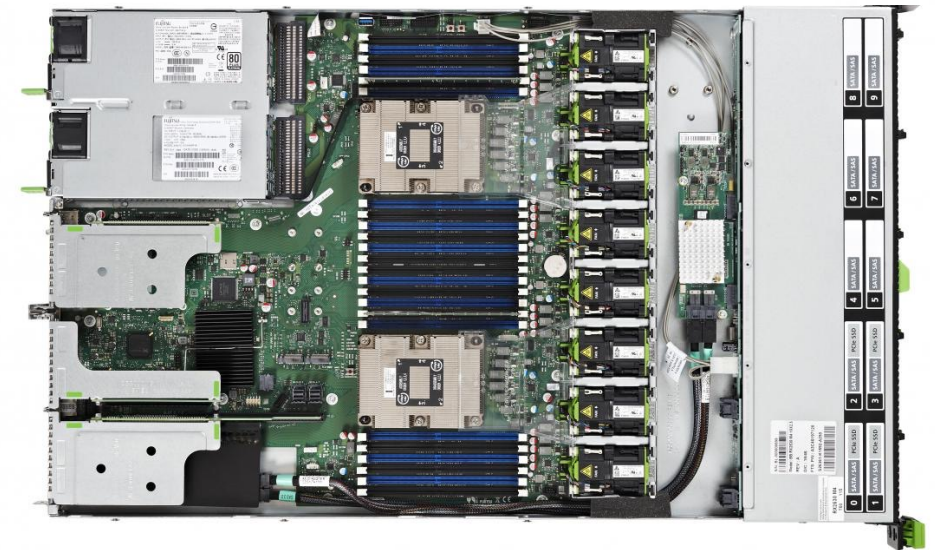
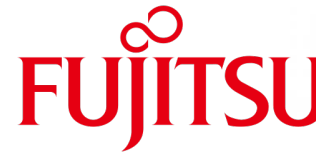
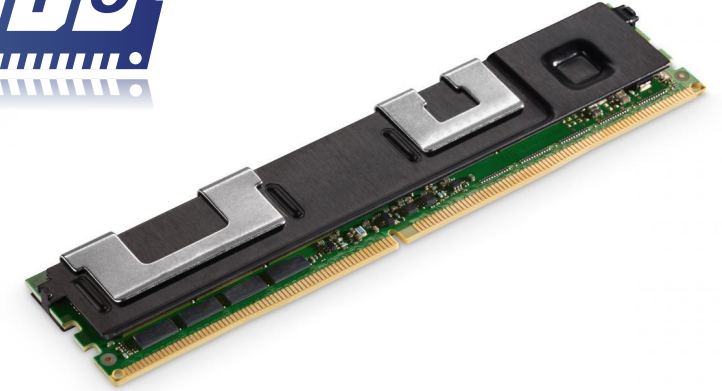
- **Domain specific (NWP) Distributed object store**
- Transactional, no synchronisation
- Semantic access to data
- Key-value store
  - Keys are **scientific meta-data** (MARS Metadata)
  - Values are **byte streams** (GRIB)
- Support for multiple backends:
  - POSIX file-system (currently on Lustre)
  - Intel DAOS (under development)
  - CEPH (Cloud suited object store)



# Exploring improved data processing nodes

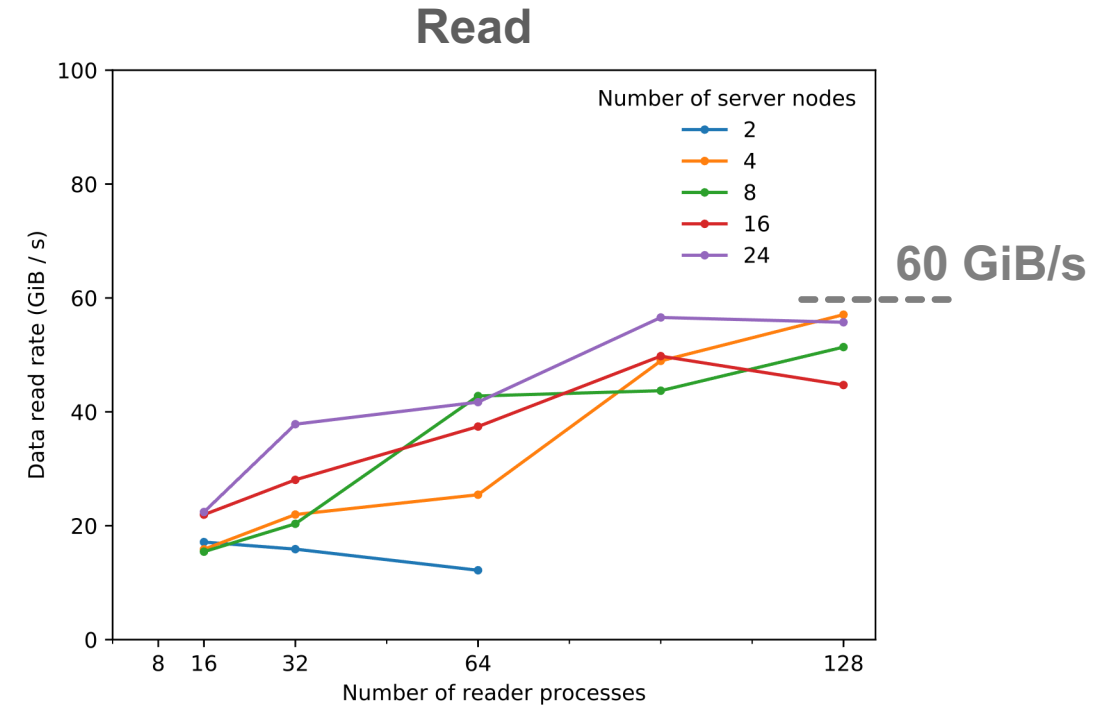
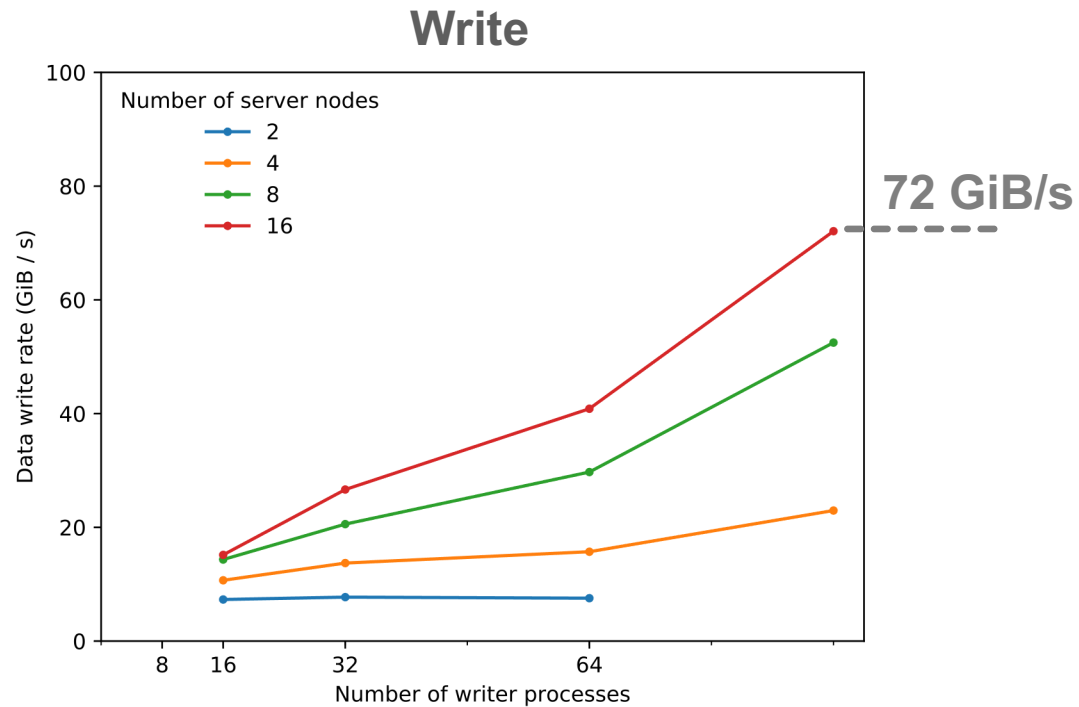


- Read all @ [www.nextgenio.eu](http://www.nextgenio.eu) (finished 2019)
- Development of an HPC node by **with Intel Optane DCPMM**
- 34 x 3 TiB NVRAM DIMMs
- Prototype system
  - 34 compute nodes
  - Hosted @ EPCC, Edinburgh



# Distributed FDB on NVRAM

I/O stack



Full model

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

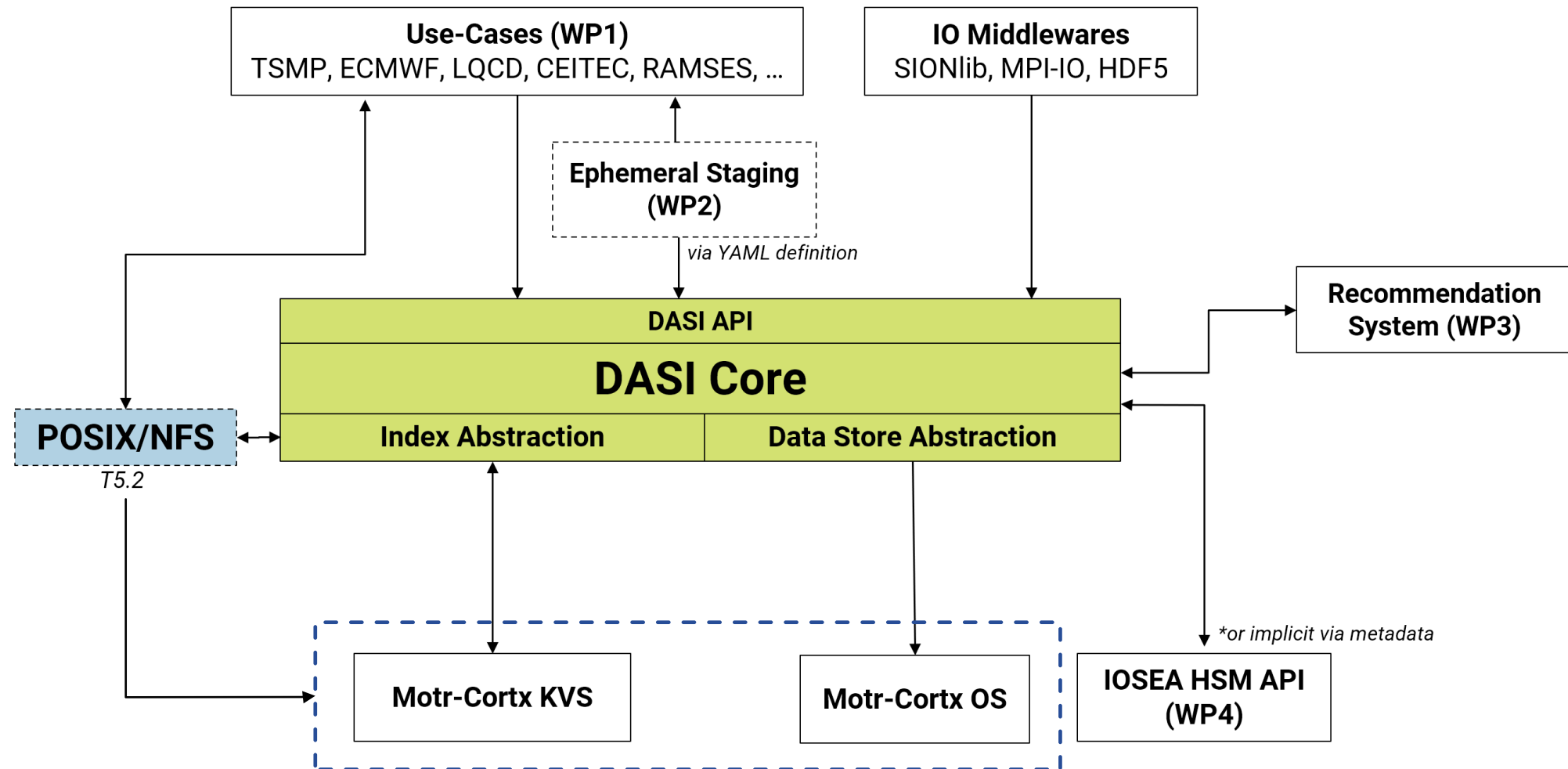
**Runtimes no longer affected by the Product Generation!**

# Enter IO-SEA

- Hierarchical Object Storage
  - From NVMe to tapes
  - A contender for meteorological data storage
- Ephemeral services and data nodes
  - In-situ data processing (PGen)
  - Ability to perform IO read-write using hot storage
- Semantic Data Access and Storage Interface (DASI)
  - A “domain-agnostic FDB”

# DASI

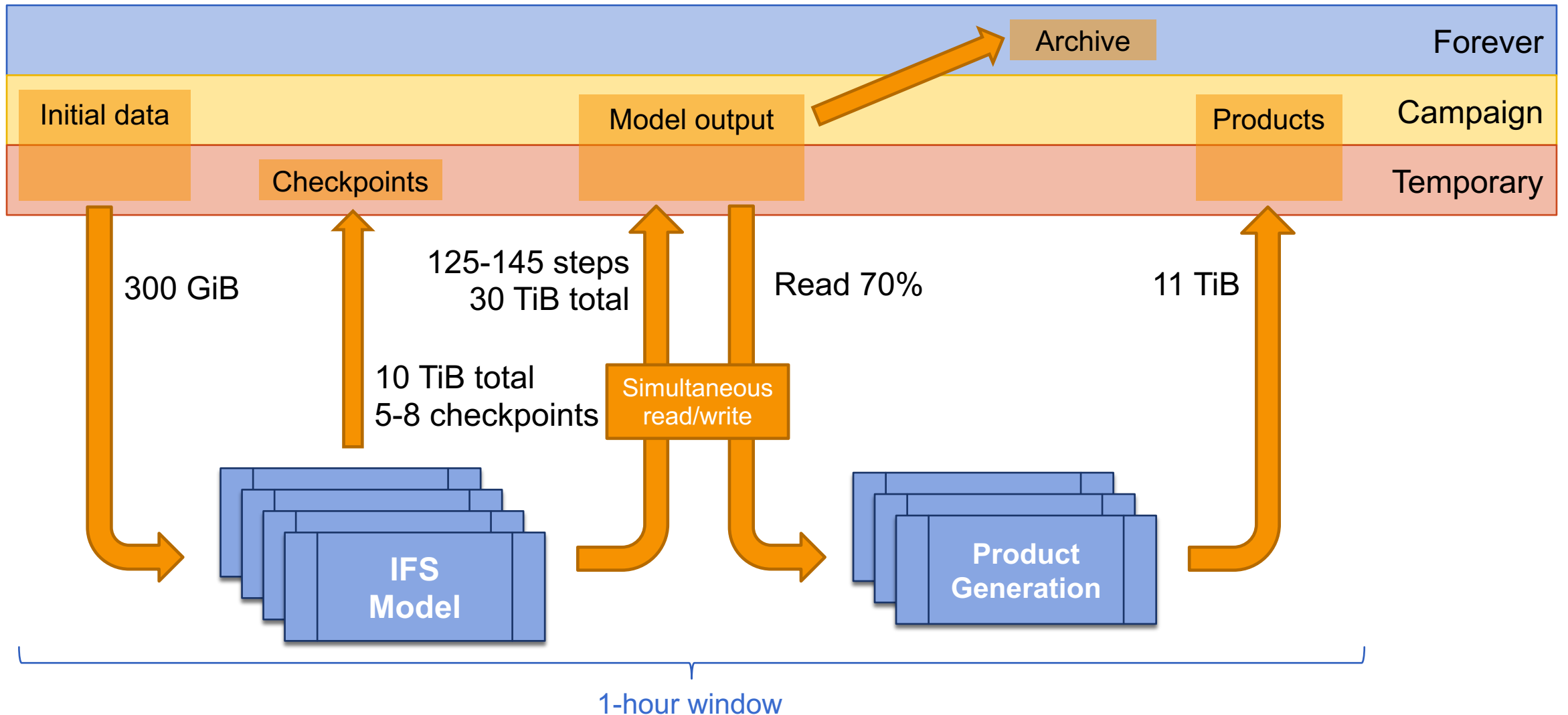
- A scientist-friendly object store, abstracting use-cases with complex IO systems



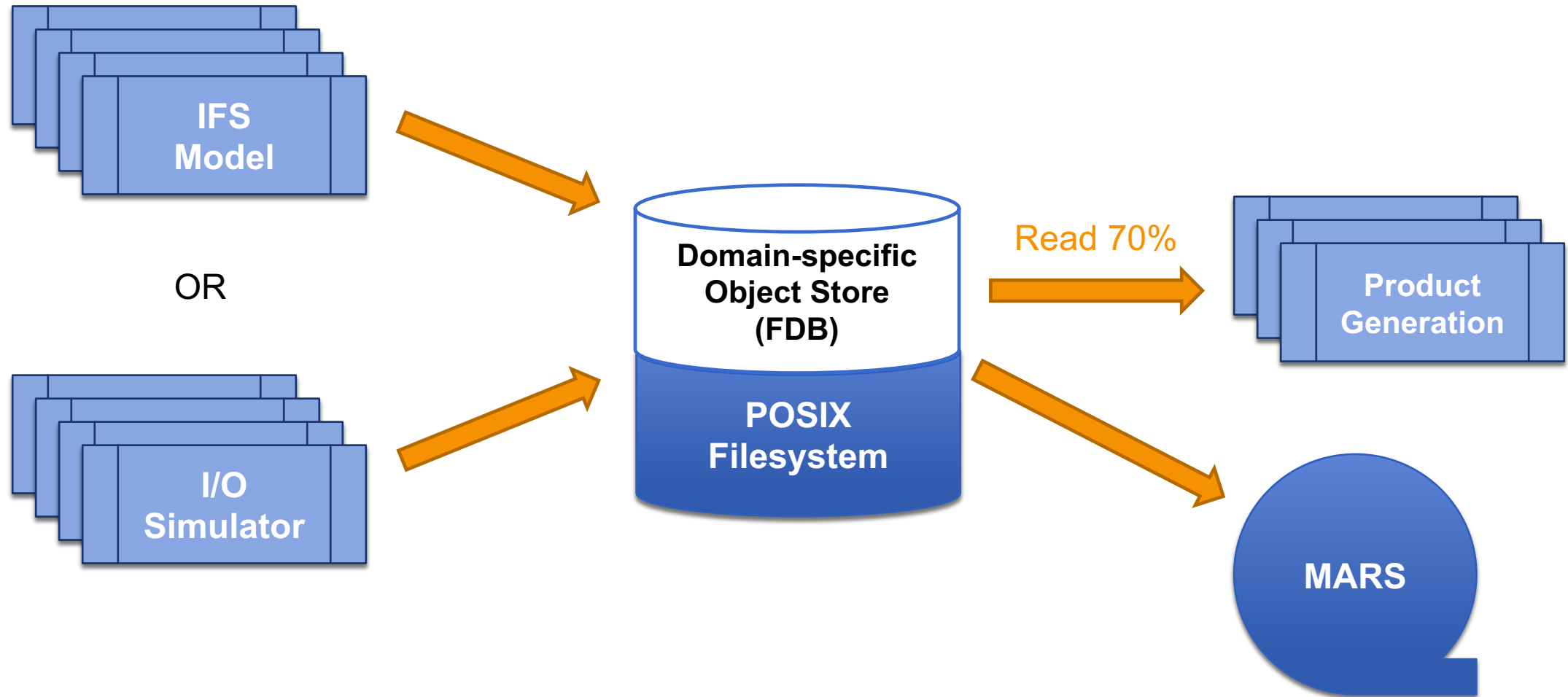
# DASI

- Bringing lessons learned from ECMWF developments
  - 40 years of MARS language
  - Experience in building data storage abstractions
- ... but domain-agnostic, not tied to meteorological data
  - Allows uptake by other scientific domains
  - Allows collaboration on building data storage backends
- Will bring IOSEA HSM storage to use-cases in IOSEA, and beyond...

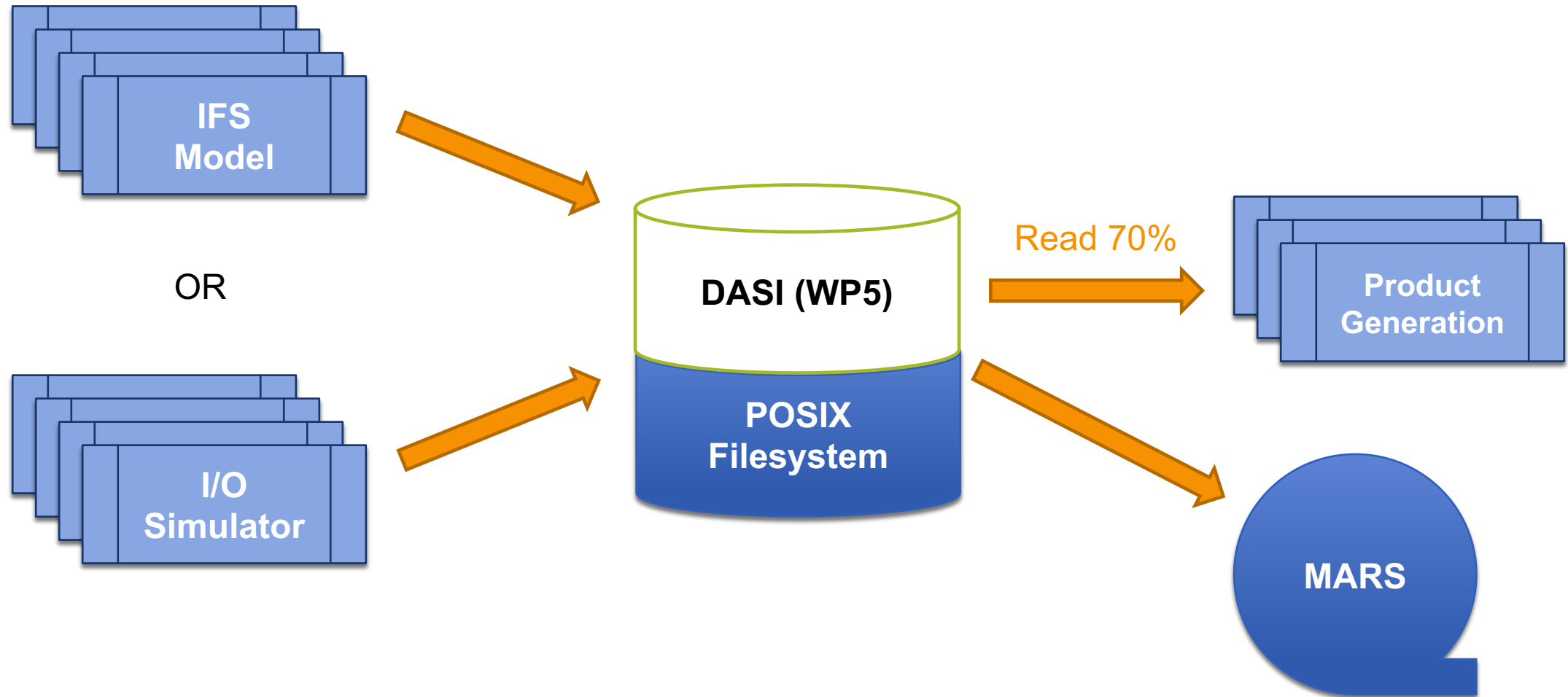
# ECMWF workflow in IOSEA



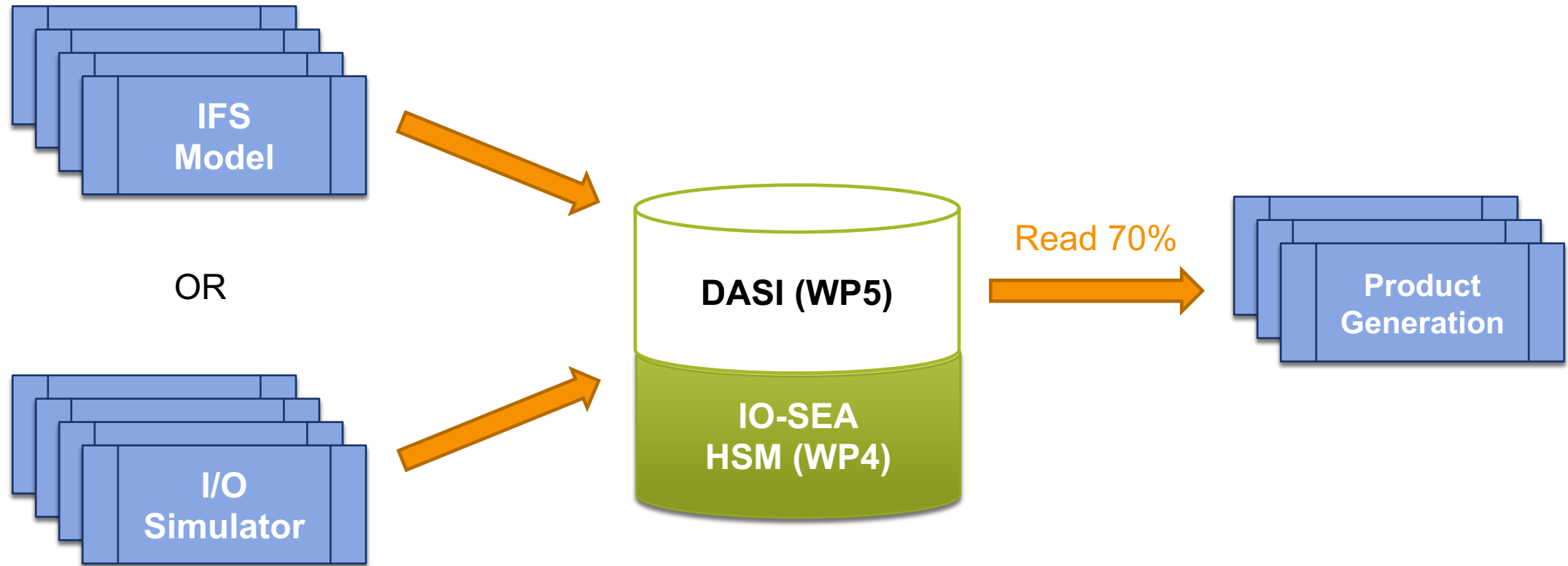
# Current System



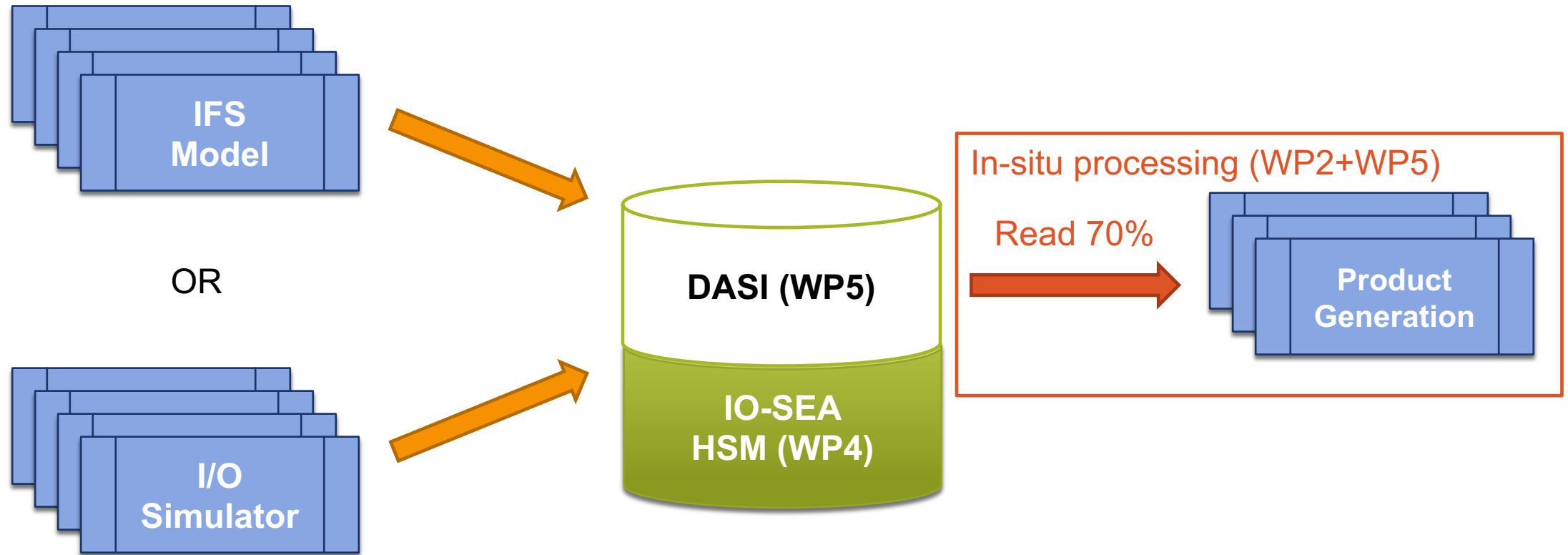
# Future System



# Future System



# Future System



# Highlights

- ECMWF runs time-critical weather forecasts with **IO read-write bottlenecks**
- We need to explore **new hardware/software solutions** for IO as our forecast resolution increases
- 40+ years of experience with **semantic access to data** and abstraction of data storage
  - Allows us to switch IO solution easily
  - We will bring this experience to IOSEA with **DASI**
  - **Collaborate** with scientific domains and IO solution providers