A Distributed Object Store for Meteorological and Climate Data

Simon Smart, Tiago Quintino, Baudouin Raoult

ECMWF

simon.smart@ecmwf.int

ECMWF

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

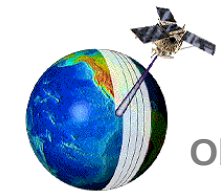
Research – **Non Time Critical**

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

ECMWF's (Simplified) Operational Workflow

Data Workflow

- Post-process data for client pr
- Post-process for statistics and
- Perpetual archival of generate



Global Observations

Acquisition

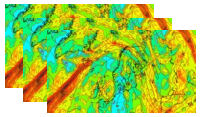
IFS Model

Product Generation

Product Dissemination

Obs

Fields

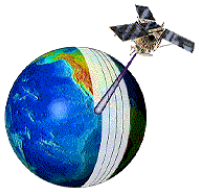


Products

MARS

Perpetual Archive

(Simplified) Storage view of workflow



Observations

Acquisition

Product
Dissemination

IFS
Model

Lustre

Product
Generation

MARS

Perpetual A

History and Future of Resolution Upgrades

Resolution	Grid size	Grid Points	Field Size (in memory)
T319	62.5 km	204 k	1.6 MB
T511	39 km	524 k	4 MB
T799	25 km	1.2 M	9.6 MB
T1279	16 km	2.1 M	16.8 MB
Tco1279	**9 km**	**6.6 M**	**50.4 MB**
Tco1999	5 km	16.1 M	122.6 MB
Tco3999	2.5 km	64 M	490 MB
Tco7999	*1.25 km*	*256 M*	***1909 MB***

ECMWF

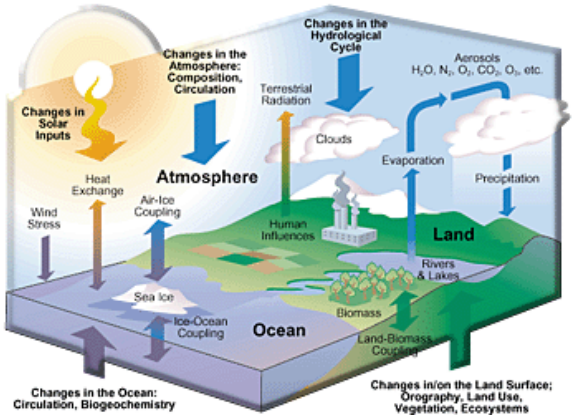
Multiple dimensions

Reliability

Ensembles

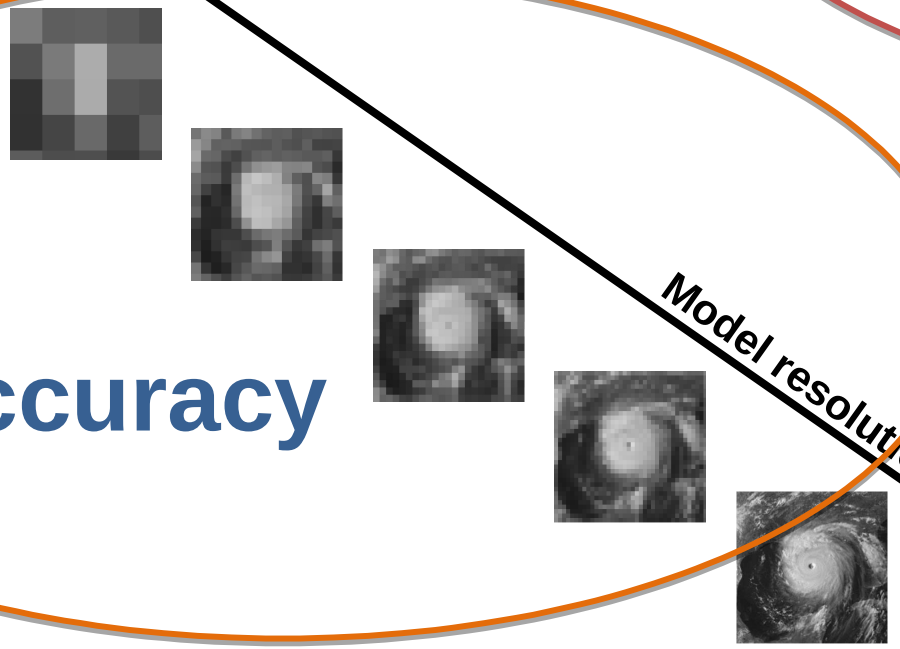
Traditional weather science domain

Model complexity



Traditio
science

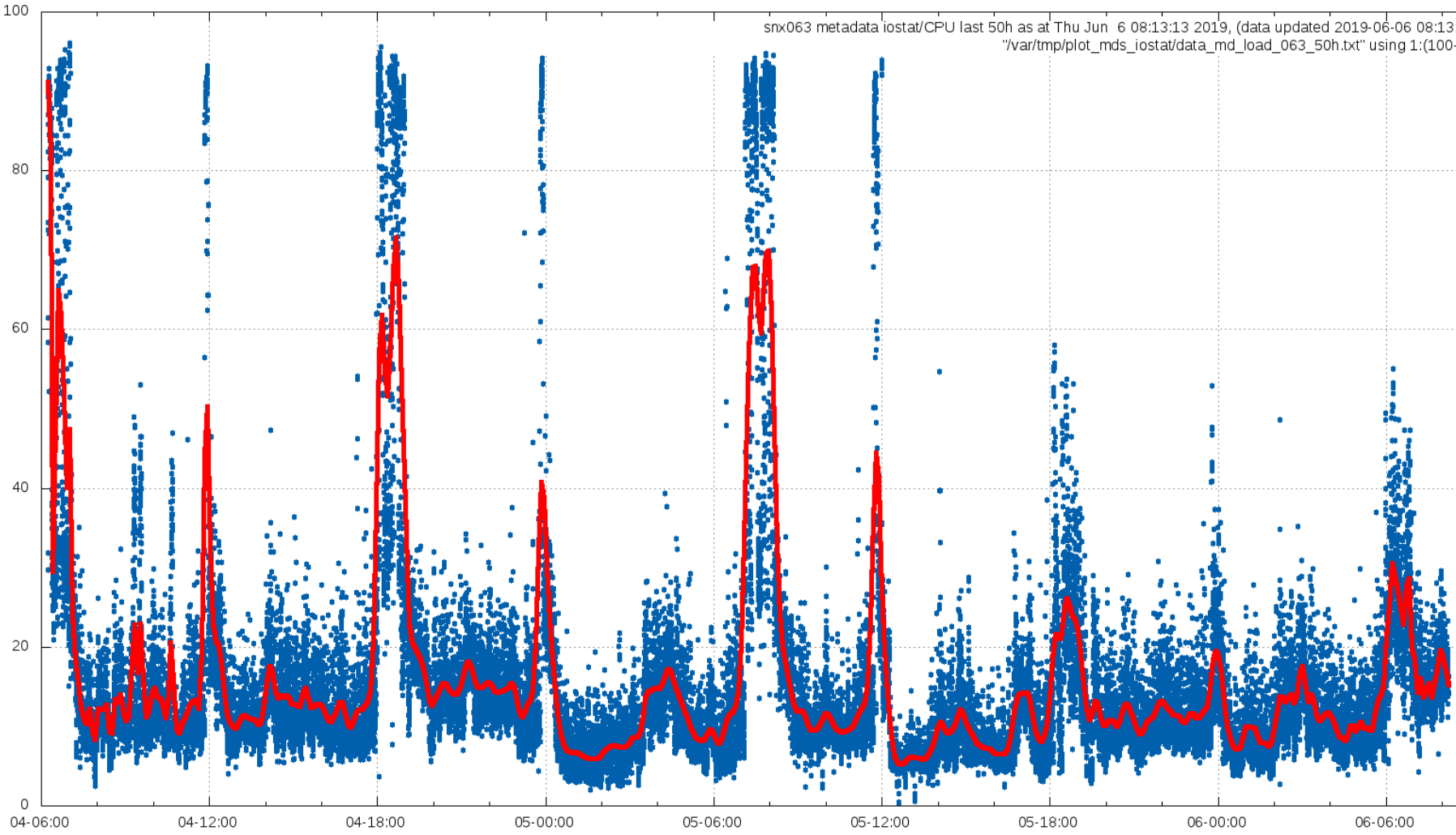
Accuracy



Model resolution

Today: it need
'Earth system'
to perform at a

Impact of Metadata Optimisation



Effects of Product Generation

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

Broadwell nodes 2x18 cores
Cray XC40 Aries interconnect
Lustre FS IOR 90GiB/s

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  = 1000/50
    DATE      = 2016051
    STEP      = 12/24/3
```

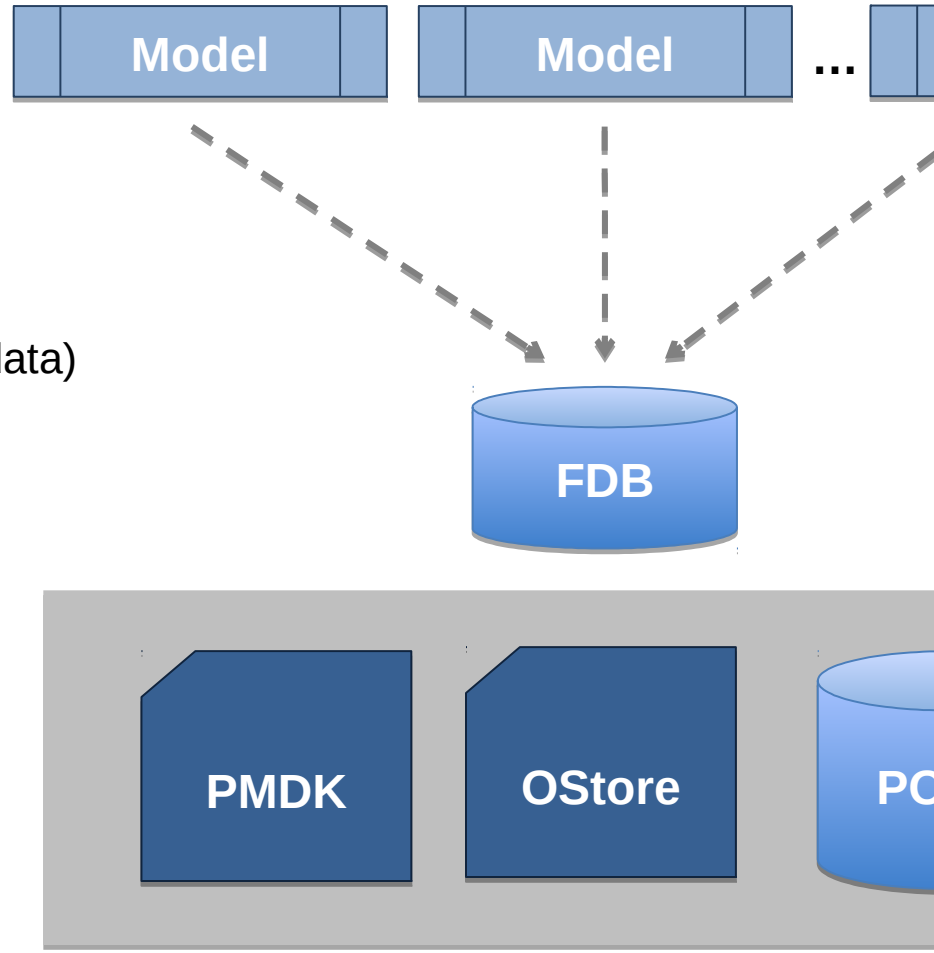
Unique and **semantic** way to describe all ECMWF data

Semantics

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

FDB (version 5)

- Domain specific (NWP) object store
- Transactional, No synchronization, No MPI
- 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)
  - DCPMMs using pmdk library
  - Could explore others:
    - Intel DAOS, Cray DataWarp, MERO, etc.

- Supports wild card searches, ranges, data conversion, etc…

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

Into operations…

```
% 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      = 2019061
STEP      = 12
```

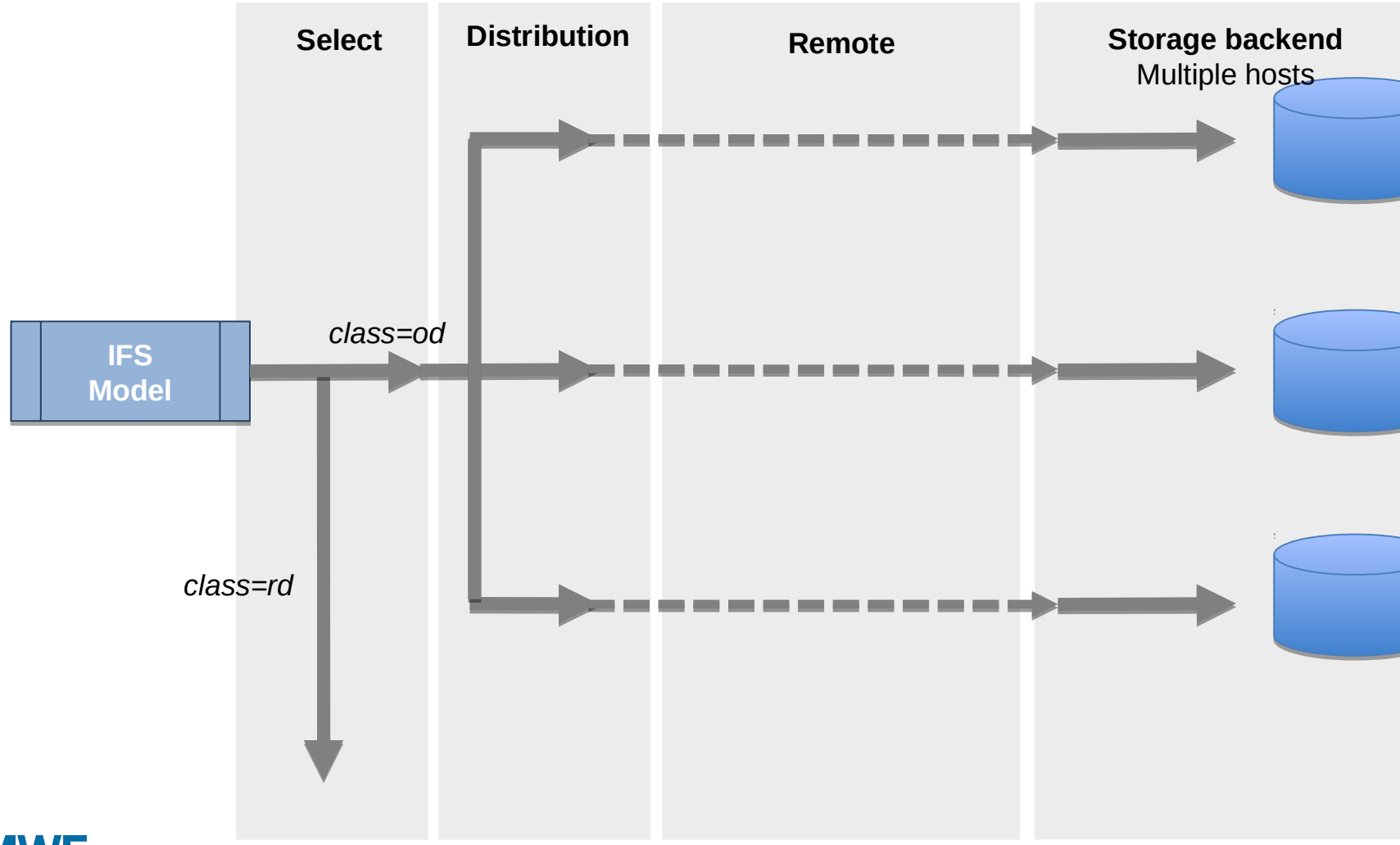
```
archive(Metadata key, void* data, size_t length);

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

flush();
```

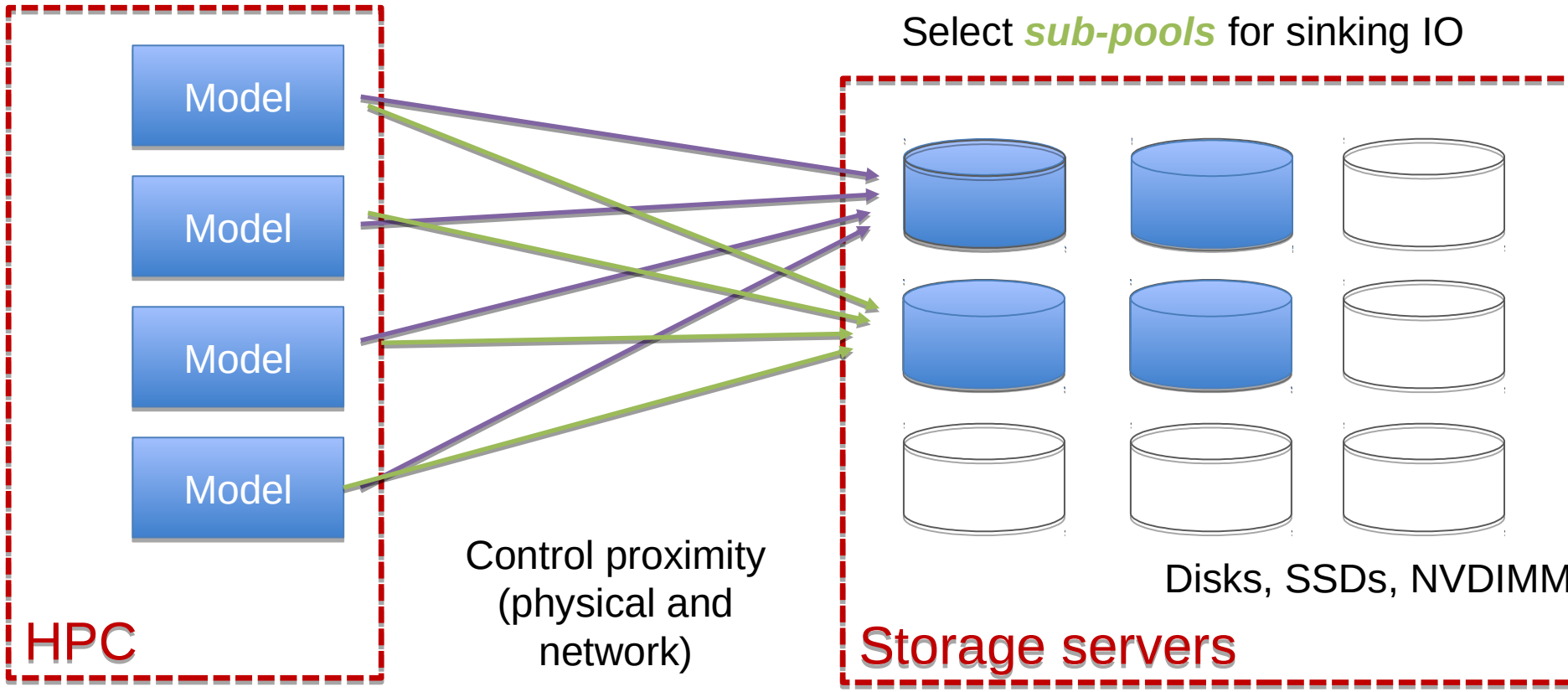
FDB5 Data Routing

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

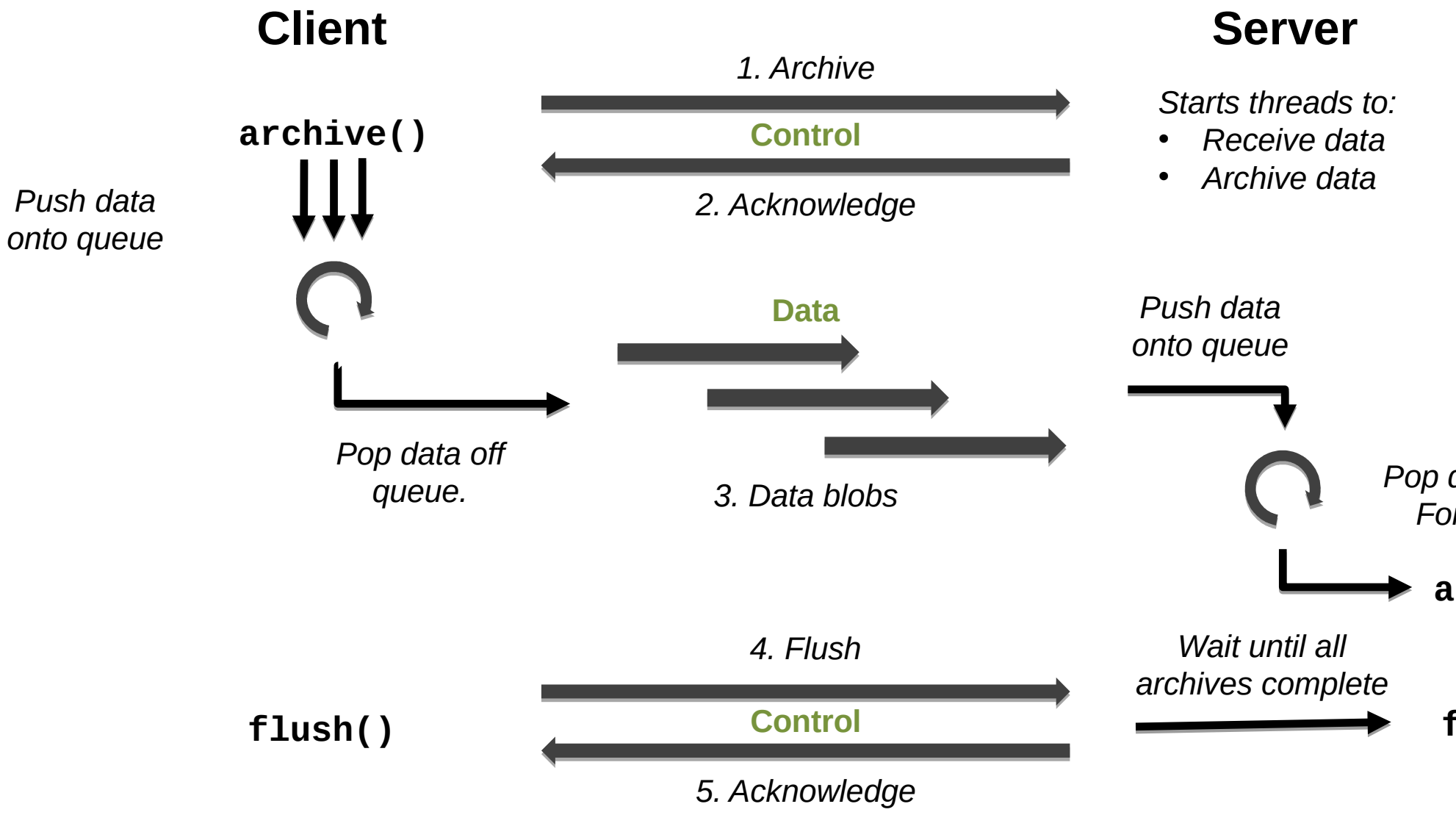


ECMWF

Capability vs Capacity

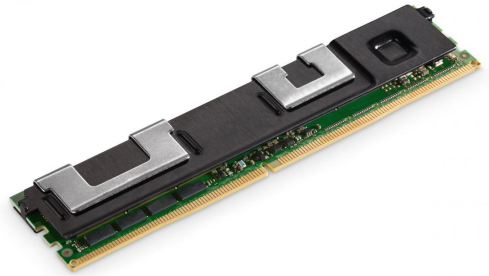


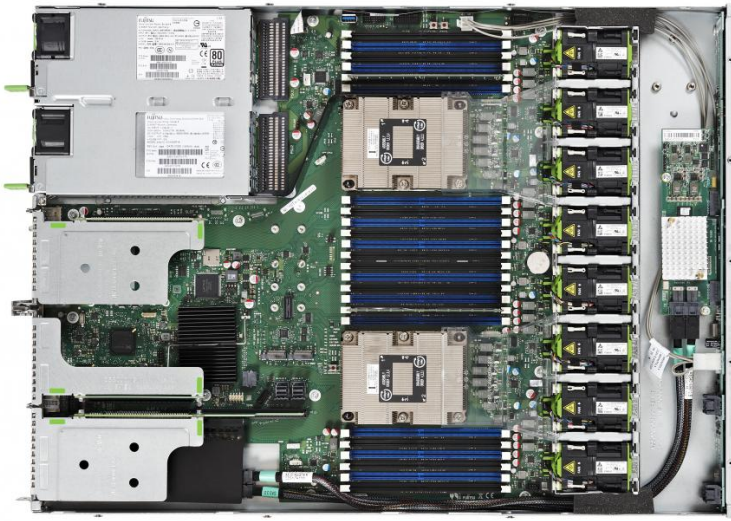
Archiving Data



NextGenIO Prototype

- Read all @ www.nextgenio.eu
- Development of an HPC node by **with Intel 3D Xpoint**
- Dual-CPU Intel® Xeon® SP nodes
- OmniPath network
- 192GB DRAM
- **3TiB of NVRAM DIMMs**

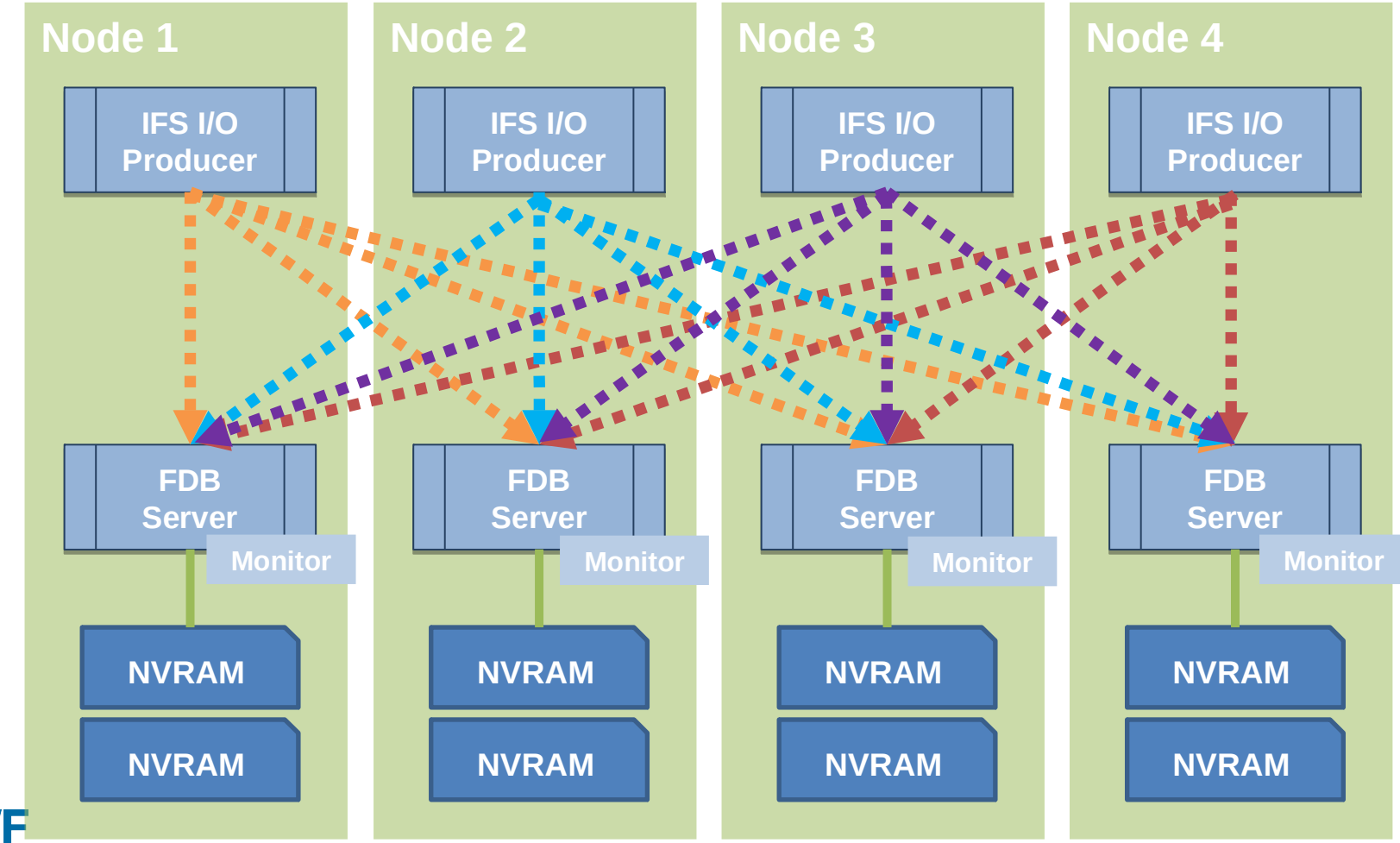




- **Prototype system**
  - **34 compute nodes**
  - **Hosted @ EPCC, Edinburgh**

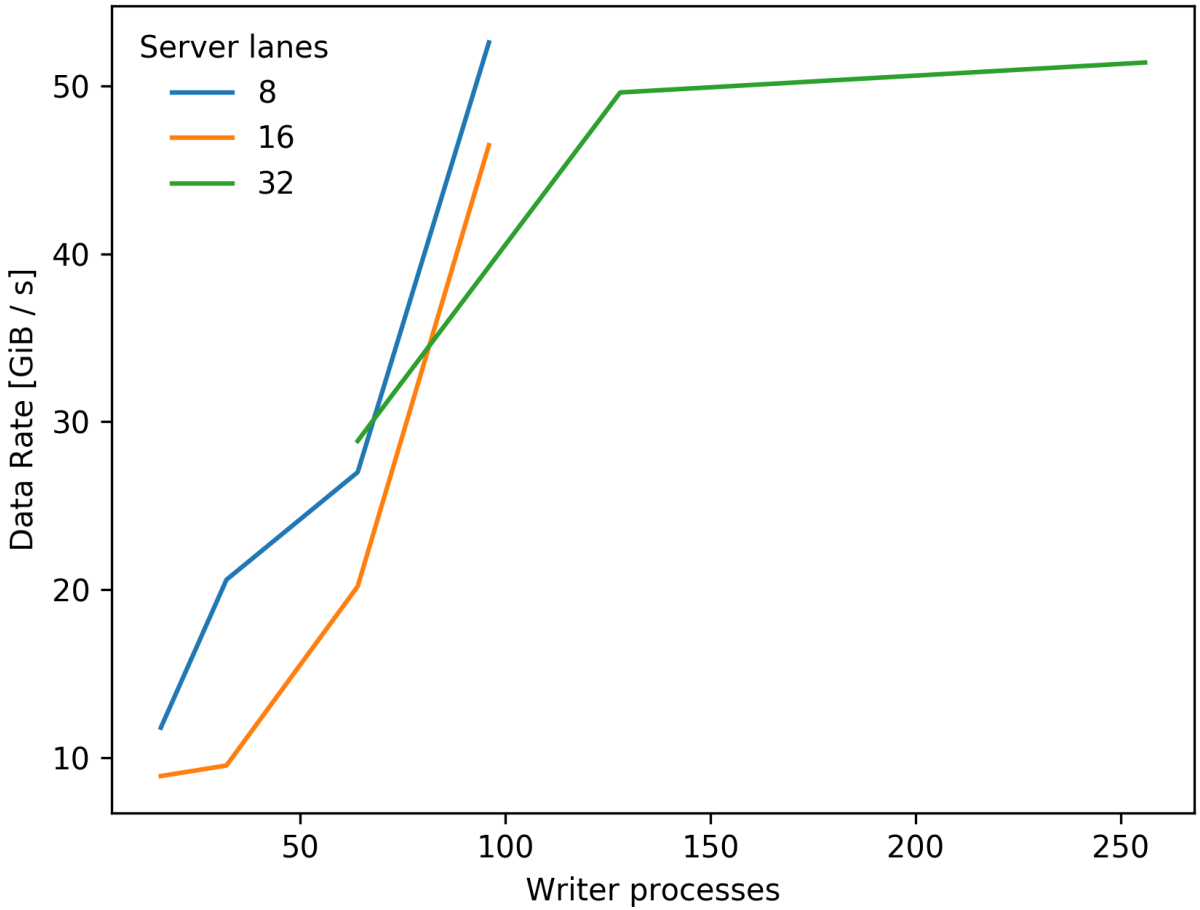
Data Flow Schematic

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

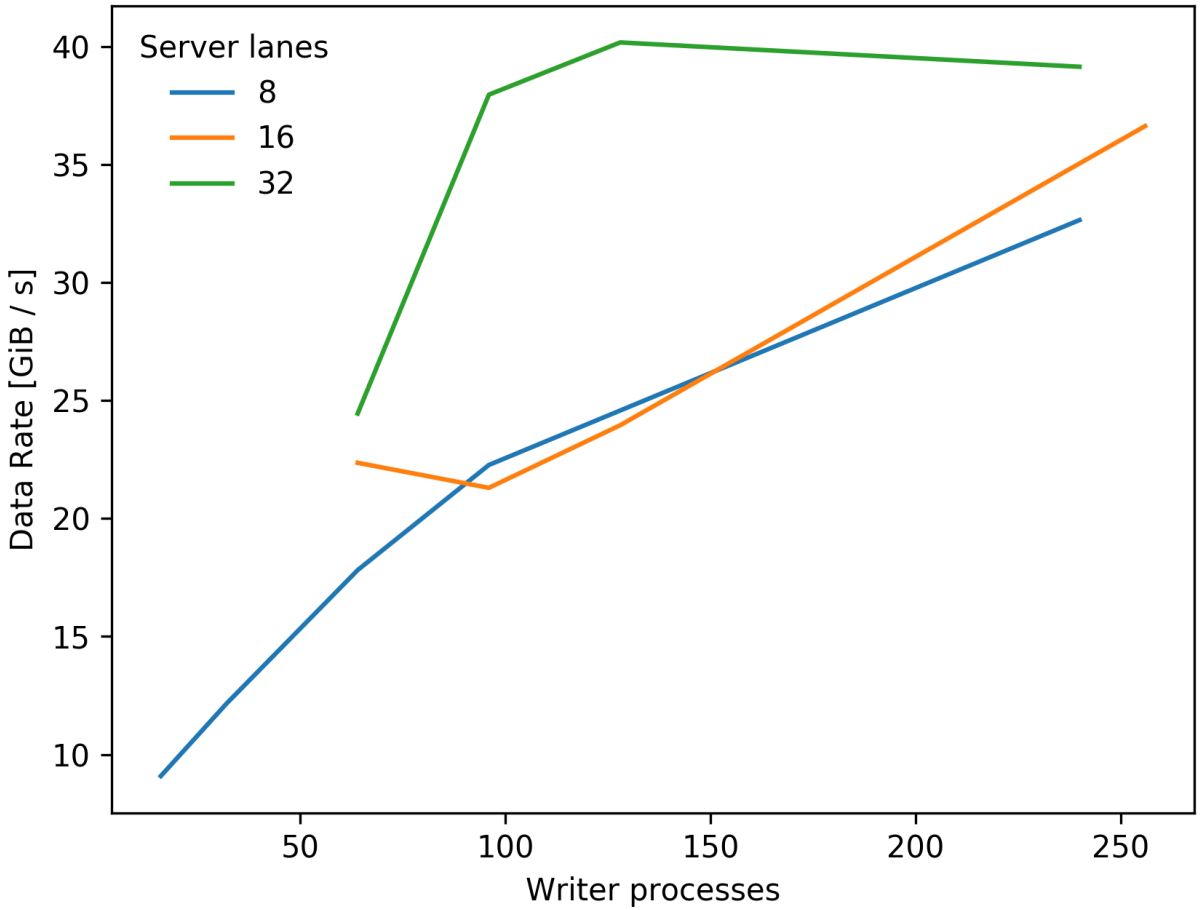


ECMWF

FDB5 Remote Performance



FDB5 Remote Performance (DCPMMs)

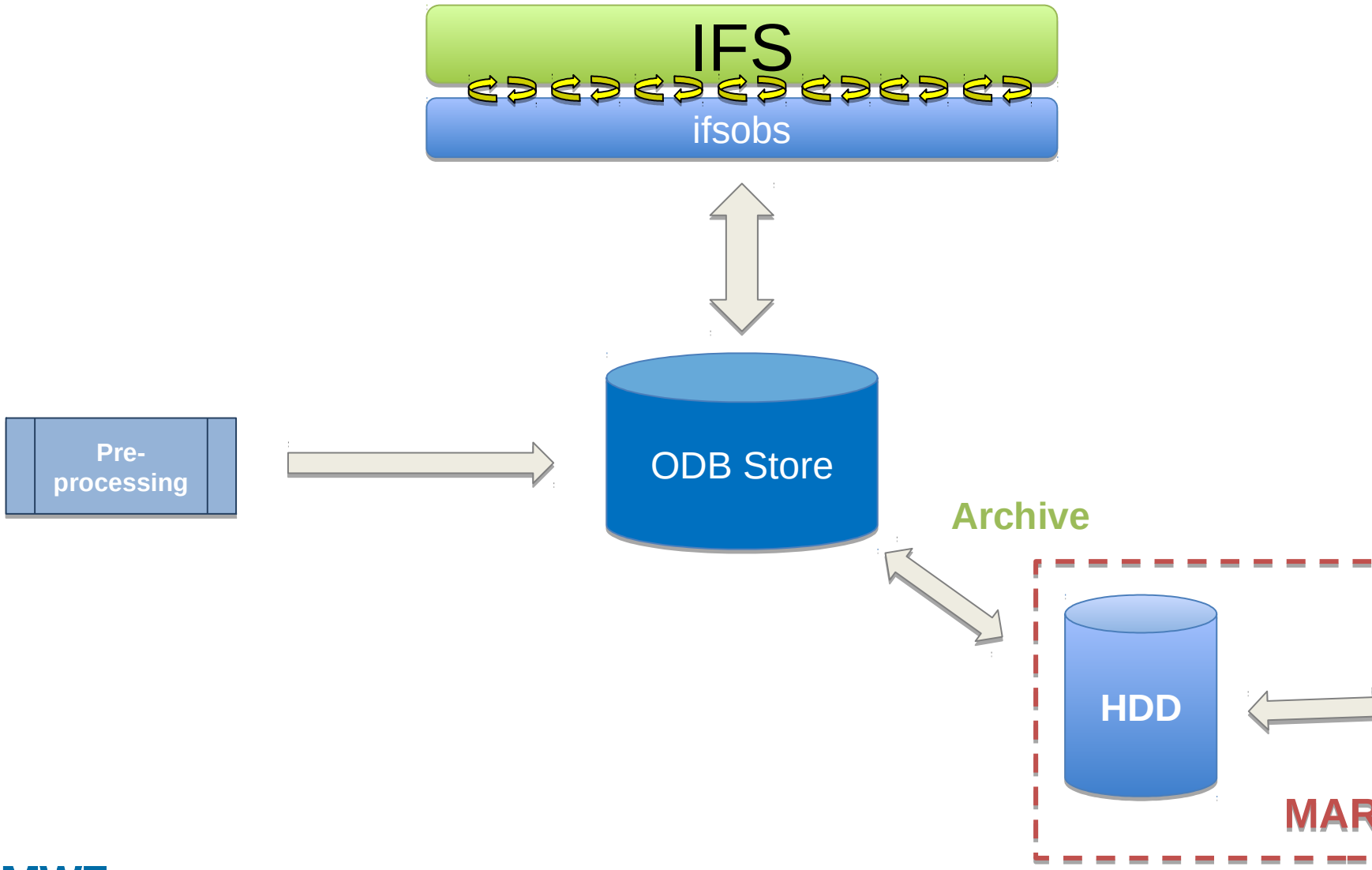


Running the forecast model

	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

ODB Store



ECMWF

Novel Data Flows – Data Centric Computing

Producer

Cloud

Consumer

- Bring use
- Use data
- Access us meaningf

NVRAM (?)
Data Hypercubes

Archive

PFS

HDD

MARS

Messages To Take Home

Domain-specific APIs *give powerful tools:*

Domain specific semantics

Consistent, science-oriented access

Separation of concerns

ECMWF has retooled its operation I/O with FDB5

Non-Volatile memory and other novel storage devices and hierarchies are coming

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

ECMWF EUROPEAN CENTRE FOR MEDIUM-RANGE WEATHER FORECASTS