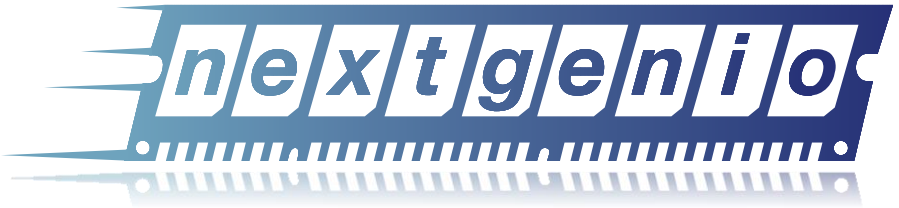
Development of a Domain Specific Distributed Object-Store

For Numerical Weather Prediction and Climate Data

S. Smart, T. Quintino, B. Raoult, P. Bauer

ECMWF

simon.smart@ecmwf.int

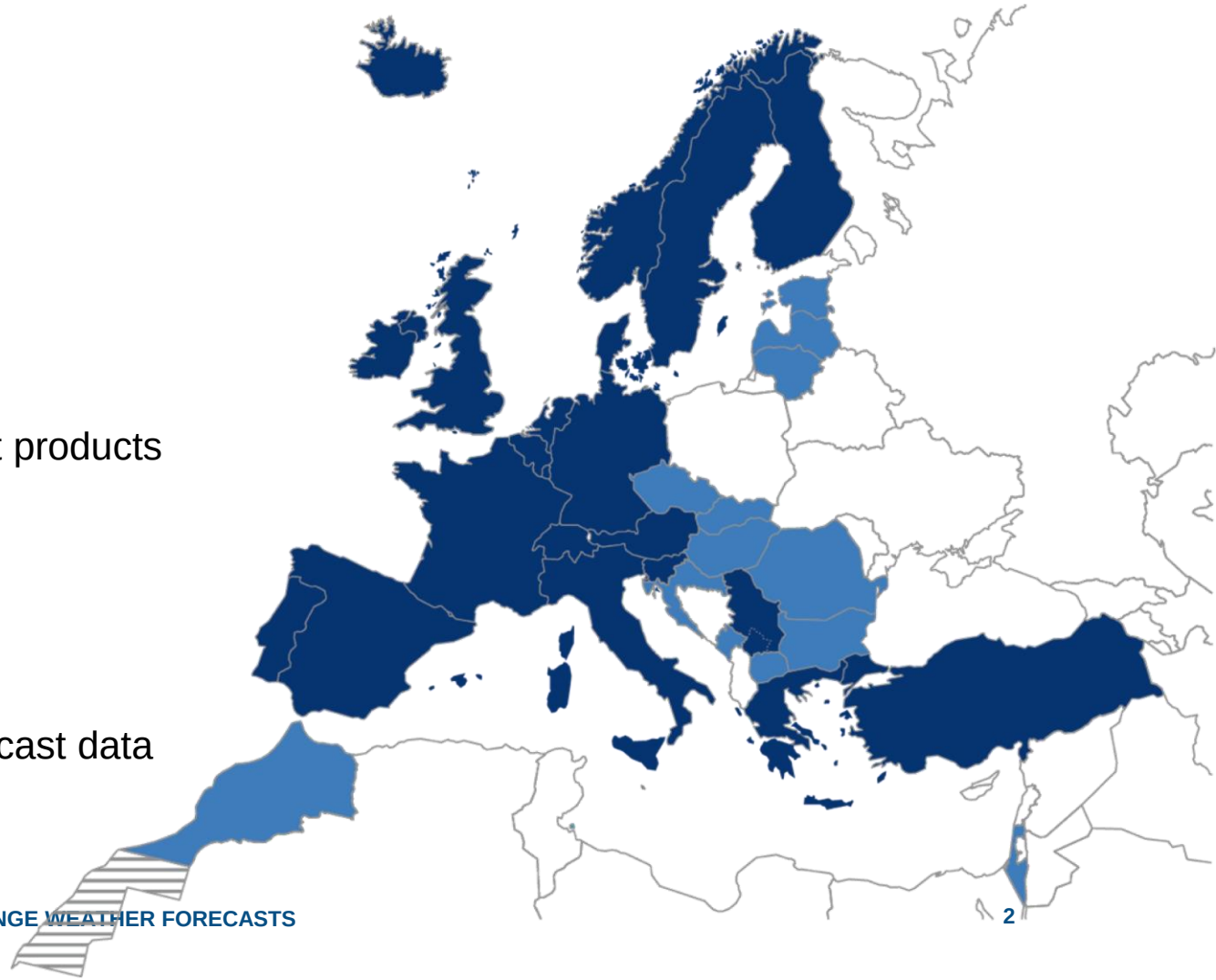


ECMWF

European Centre for Medium Range Weather Forecasts (ECMWF)

An independent **intergovernmental** organisation

21 Member States
13 Co-operating States



What do we do?

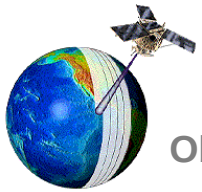
Operational forecasts – **Time Critical**

- 2 hours from satellite cut-off to deliver forecast products
- Twice per day, 00Z and 12Z

Research – **Non Time Critical**

- Vast majority of the workload
- Re-uses current and historic analysis and forecast data

ECMWF's (Simplified) Operational Workflow



Global Observations

Data Workflow

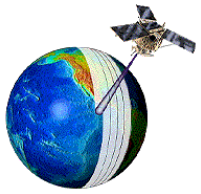
- Post-process data for client products
- Post-process for statistics and data analysis
- Perpetual archival of generated data sets

Acquisition → IFS Model → Product Generation → Product Dissemination → Member States & Customers

Obs, Fields → MARS (Perpetual Archive)

Products

(Simplified) Storage view of workflow



Observations

Acquisition

Product
Dissemination

Member
States
&
Customers

IFS
Model

Lustre

Product
Generation

MARS

Perpetual Archive

MARS Language

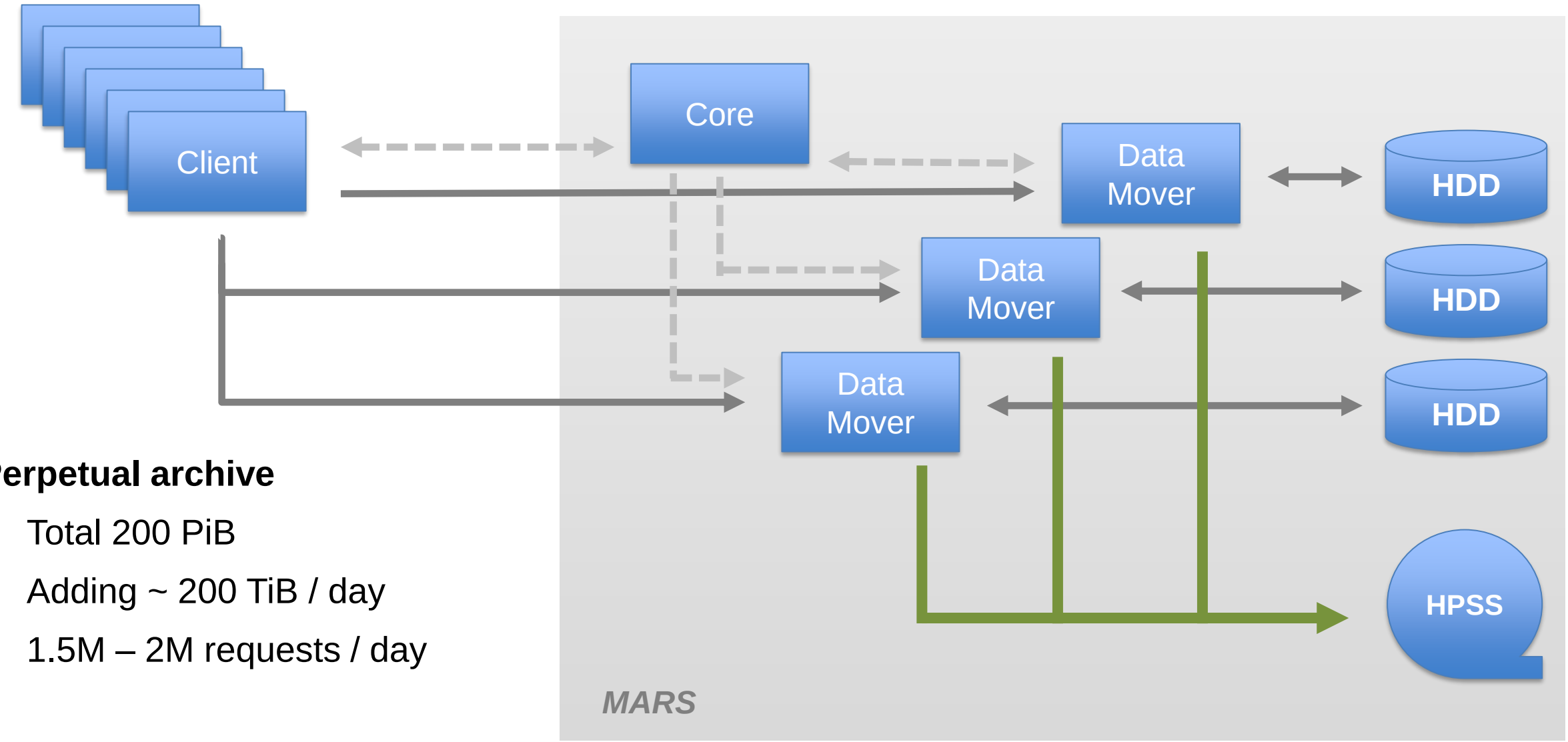
```
RETRIEVE,
    CLASS     = OD,
    TYPE      = FC,
    LEVTYPE   = PL,
    EXPVER    = 0001,
    STREAM    = OPER,
    PARAM     = Z/T,
    TIME      = 1200,
    LEVELIST  = 1000/500,
    DATE      = 20160517,
    STEP      = 12/24/36
```

```
RETRIEVE,
    CLASS     = RD,
    TYPE      = FC,
    LEVTYPE   = PL,
    EXPVER    = ABCD,
    STREAM    = OPER,
    PARAM     = Z/T,
    TIME      = 1200,
    LEVELIST  = 1000/500,
    DATE      = 20160517,
    STEP      = 12/24/36
```

Unique way to describe all ECMWF data both
Operational and **Research**

Separate ***where*** data is stored
from ***how*** data is stored.

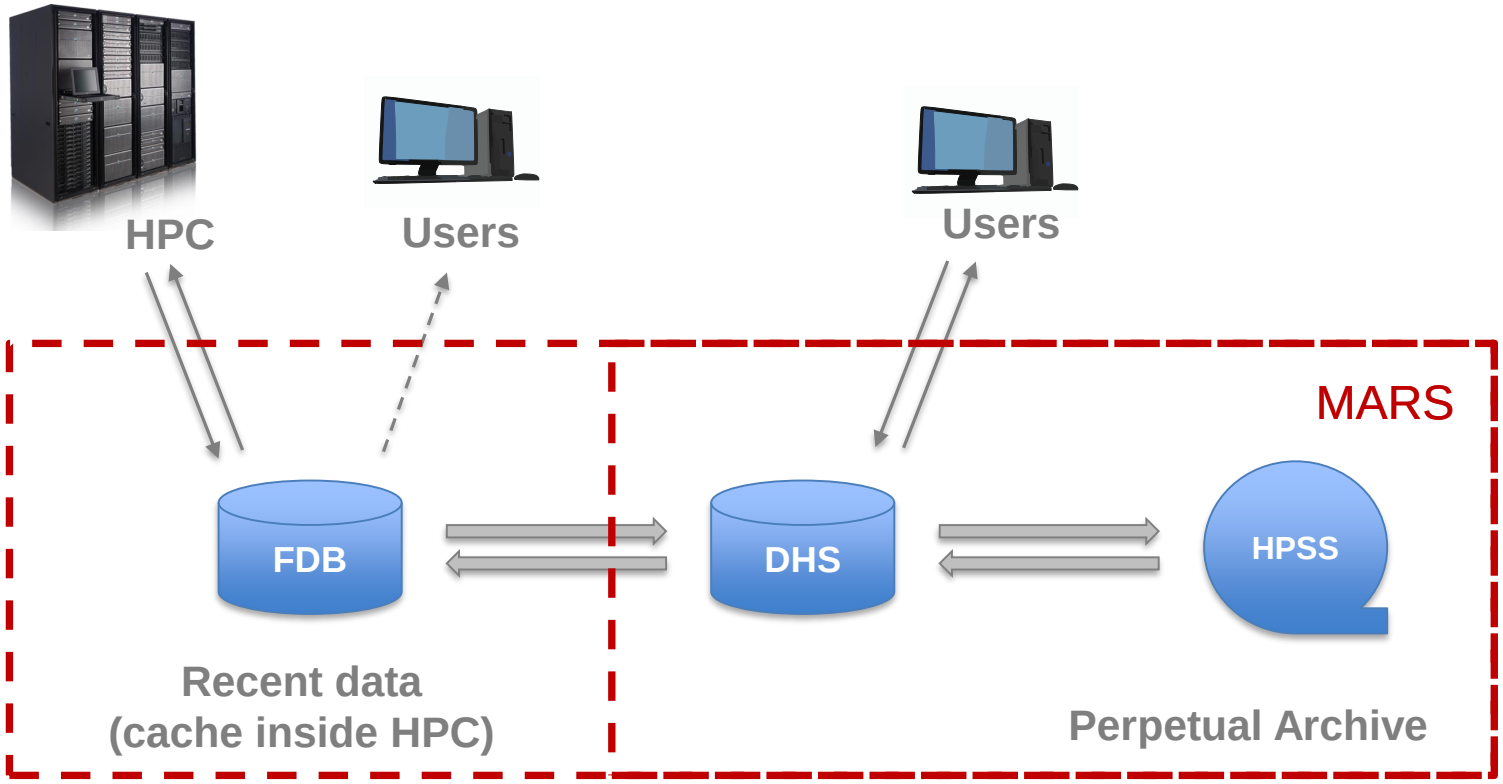
Meteorological Archival and Retrieval System (MARS)



Perpetual archive

- Total 200 PiB
- Adding ~ 200 TiB / day
- 1.5M – 2M requests / day

What is the Fields Database (FDB)?



Hit-Ratio:	80%	16%	4%

Challenges

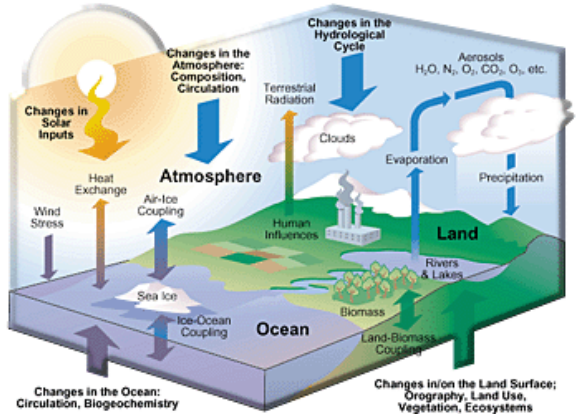
Multiple dimensions

➔Reliability

Ensembles

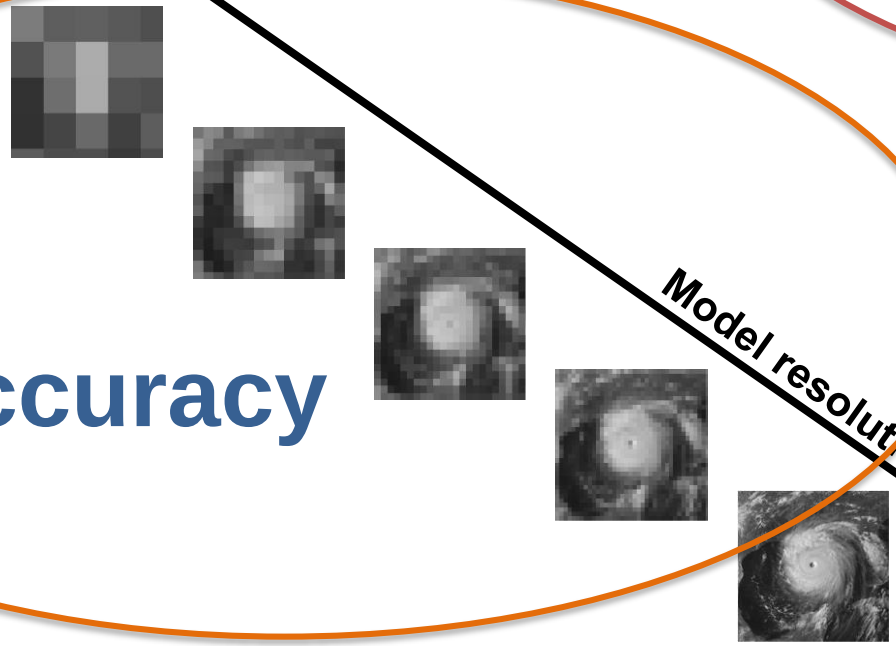
Traditional weather science domain

Model complexity



➔Range

Traditional climate science domain



➔Accuracy

Model resolution

Today: it needs high-resolution, 'Earth system' model ensembles to perform at all scales!

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***

So, Why a ***Domain Specific*** Object Store?

Flexibility

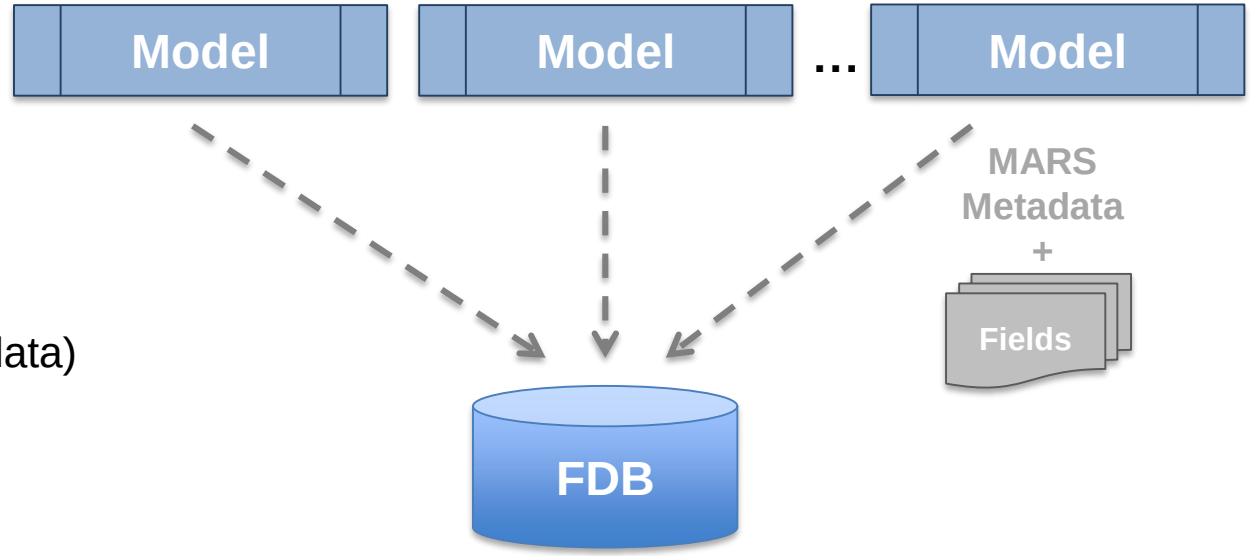
- Many new technologies (H/W and S/W) coming to market
- Existing system is tied to POSIX

Consistency

- Data is presented in the same manner to applications
- Access is through semantically meaningful metadata

FDB (version 5)

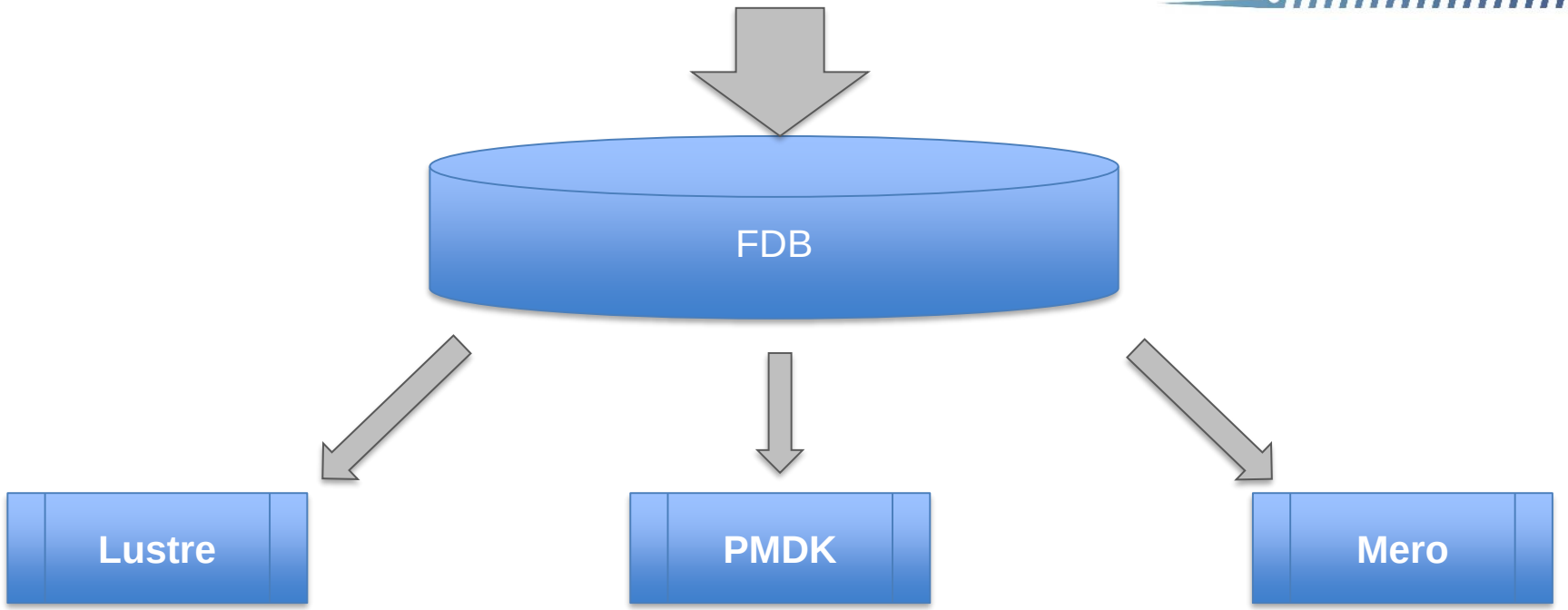
- Domain specific (NWP) object store
- Transactional, No synchronization, No MPI
- Key-value store
  - Keys are scientific meta-data (MARS Metadata)
  - Values are byte streams (GRIB)



- Support for multiple back-ends:
  - POSIX file-system (currently on Lustre)
  - 3D XPoint using pmdk library
  - Could explore others:
    - Intel DAOS, Cray DataWarp, MERO, etc.

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

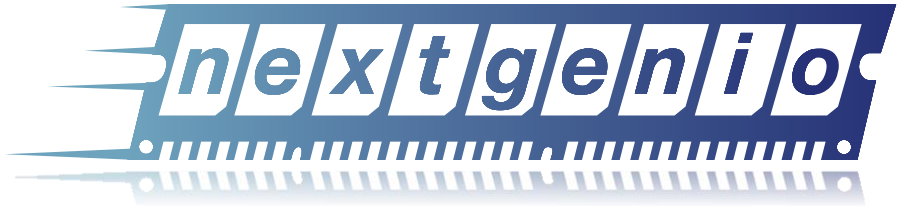
param=temperature/humidity,
levels=all,
steps=0/240/by/3
date=01011999/to/31122015,



New **opportunities** to adapt data workflows

What is NextGenIO?

Integrated into ECMWF's Scalability Programme



Exploring new NVRAM technologies to minimise Exascale I/O bottlenecks

Partners

- EPCC (Proj. Leader)
- Intel
- Fujitsu
- T.U. Dresden
- Barcelona S.C.
- Allinea Software
- ARCTUR
- ECMWF

Project Aims

- Build an HPC prototype system with Intel 3D XPoint technology
- Develop tools and systemware to support application development
- Design scheduler startegies that take NVRAM into account
- Explore how to best use this technology in I/O servers

ECMWF Tasks

- Provide requirements and use cases
- Develop a I/O Workload Simulator
- Explore interation with I/O server layer in IFS
- Test and assess the system scalability

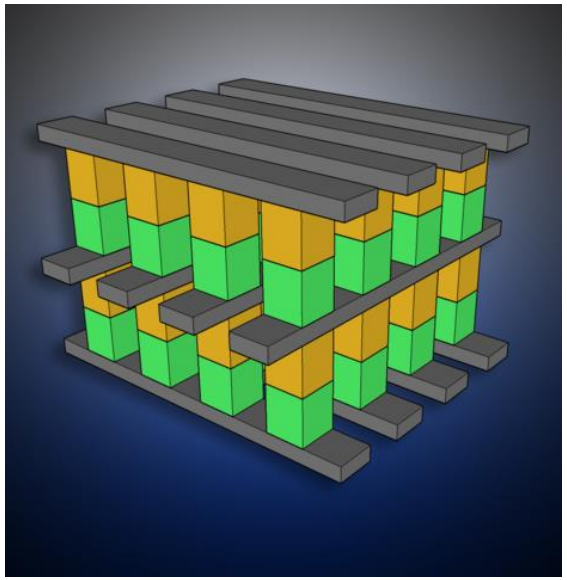
http://www.nextgenio.eu - EU funded H2020 project, runs 2015-2018

NVRAM Intel 3D XPoint

Key characteristics:

- storage **density similar** to NAND flash memory
- **better durability**
- **speed and latency better** than NAND, though slower than DRAM
- priced between NAND and DRAM

Source: https://en.wikipedia.org/wiki/3D_XPoint



"3D XPoint" by Trolomite
Own work. Licensed under CC BY-SA 4.0

How is ECMWF planning to use this technology?

- **large buffers** for **time critical** applications
  - similar to *burst buffers* but in application space
- **persistence** until archival, for **non time critical**
  - adding a new layer in the hierarchical storage system view

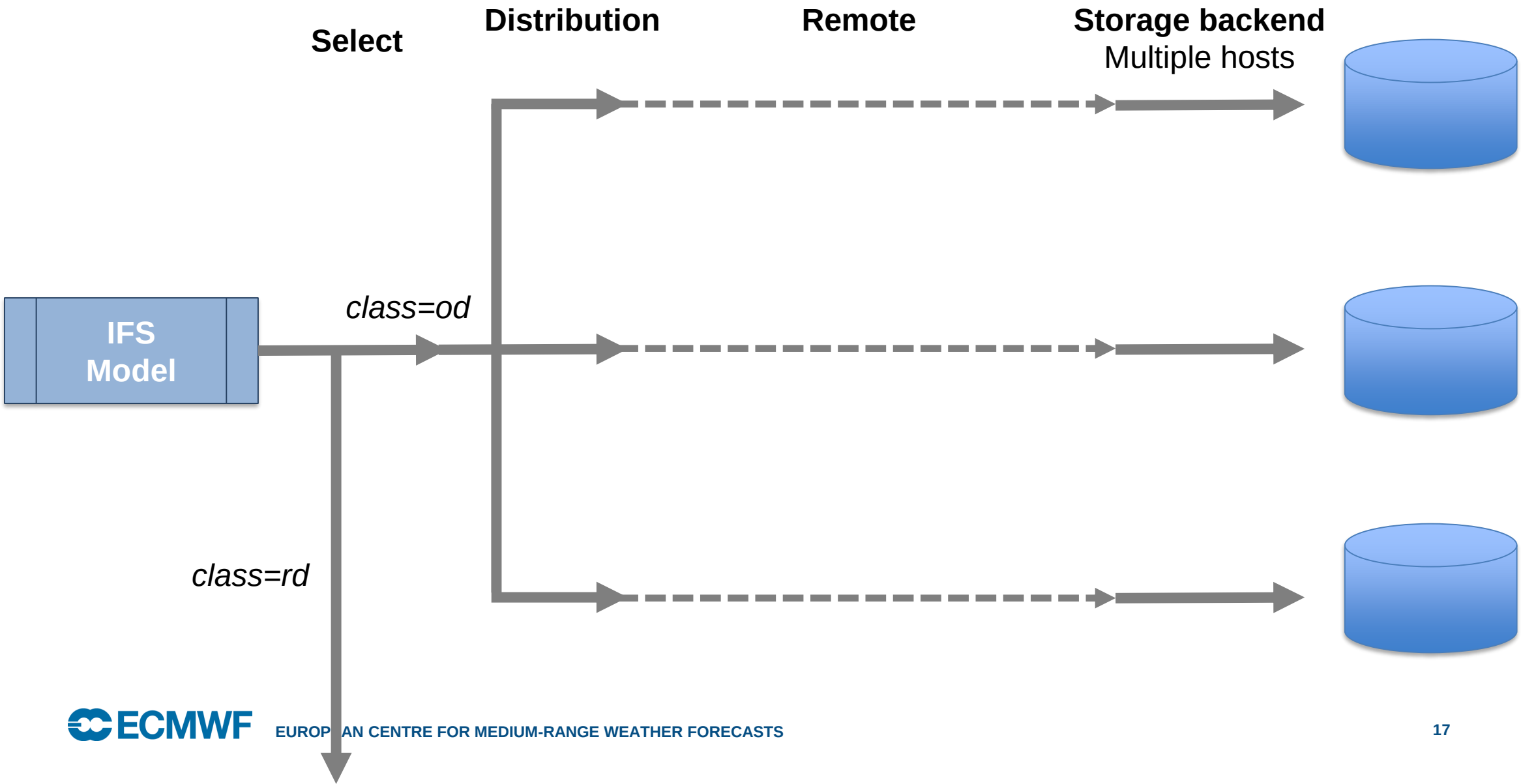
Key Point: High Density at very low latency

Front-ends and API

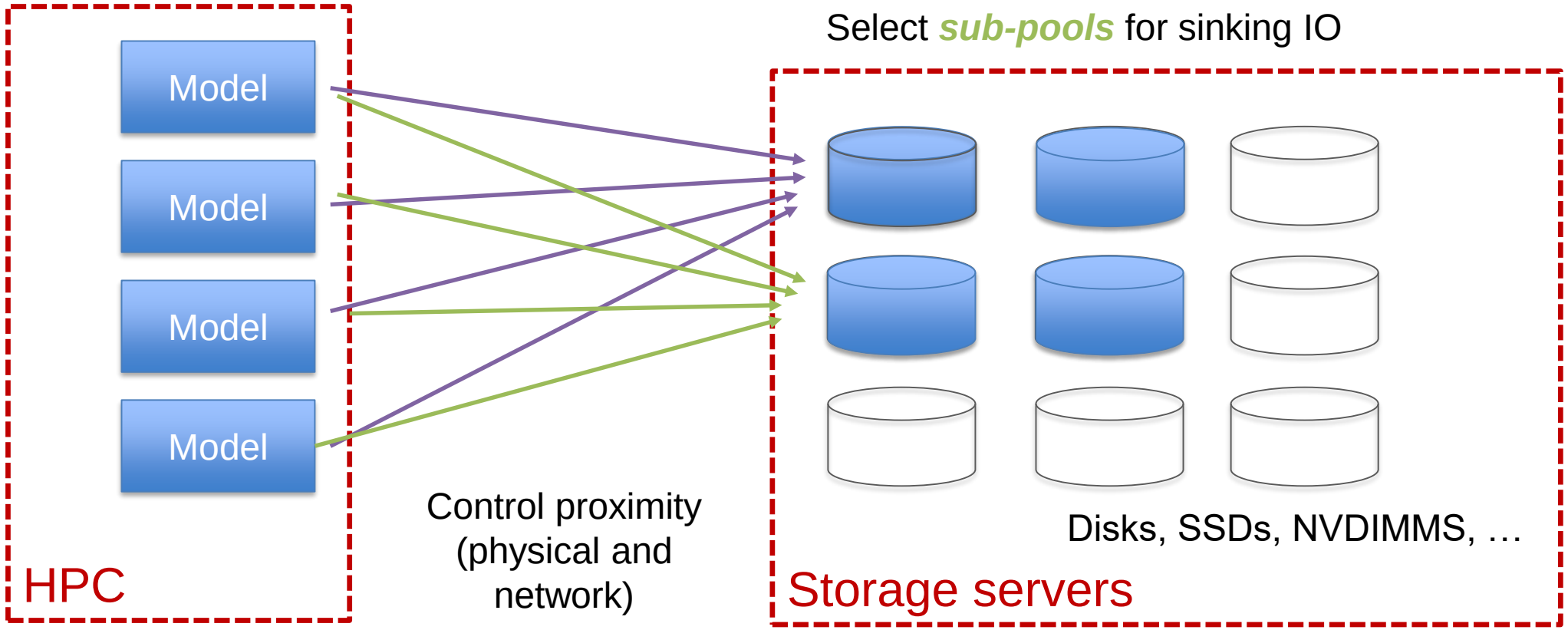
- Determines where the data is stored …
  - Run-time configurable
  - Implement data collocation policies
  - Manage data pools
  - Implements a simple interface:

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

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

flush();
```

Flexible data storage



Capability vs Capacity



Some Results

(while we wait for NVDIMM hardware)

Preliminary numbers – fresh out of the oven …

Two target nodes, with spinning disks (network) attached.

Connected to test cluster via dual 10 Gbps ethernet

Processors	Nodes	Fields	Data [GiB]	Aggregate Fields per second	Aggregate Rate [MiB / s]	Server side Per-process Rate [MiB / s]
1	1	12600	38.51	59.83	187.23	295.13
2	2	25200	77.02	126.17	394.84	392.66
4	4	50400	154.03	196.80	615.90	220.70
8	8	100800	308.06	226.26	708.10	75.81
16	16	201600	616.13	345.17	1080.22	43.73
32	32	403200	1232.25	331.64	1037.89	22.46
64	32	806400	2464.51	316.28	989.80	9.81
128	32	806400	2464.51	295.04	923.34	5.07
256	32	752640	2300.21	292.33	914.86	2.66

Preliminary numbers – fresh out of the oven … (2)

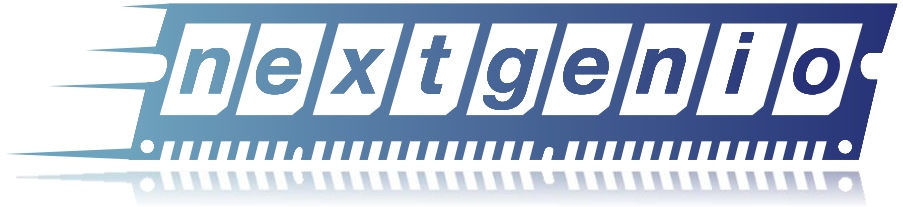
Four target nodes, with NVMe SSDs.

Connected to test cluster via dual 10 Gbps ethernet – but further from the cluster

Processors	Nodes	Fields	Data [GiB]	Aggregate Fields per second	Aggregate Rate [MiB / s]	Server side Per-process Rate [MiB / s]
1	1	12600	38.51	57.14	178.82	556.48
2	2	25200	77.02	112.94	353.46	505.27
4	4	50400	154.03	212.67	665.55	493.11
8	8	100800	308.06	212.29	664.37	528.82
16	16	201600	616.13	217.85	681.76	549.71
32	32	403200	1232.25	182.42	570.87	558.43
64	32	806400	2464.51	204.59	640.27	561.34
128	32	806400	2464.51	196.91	616.24	557.43
256	32	752640	2300.21	188.48	589.85	549.81

Where are we going?

Impacts of NVRAM on Data Access



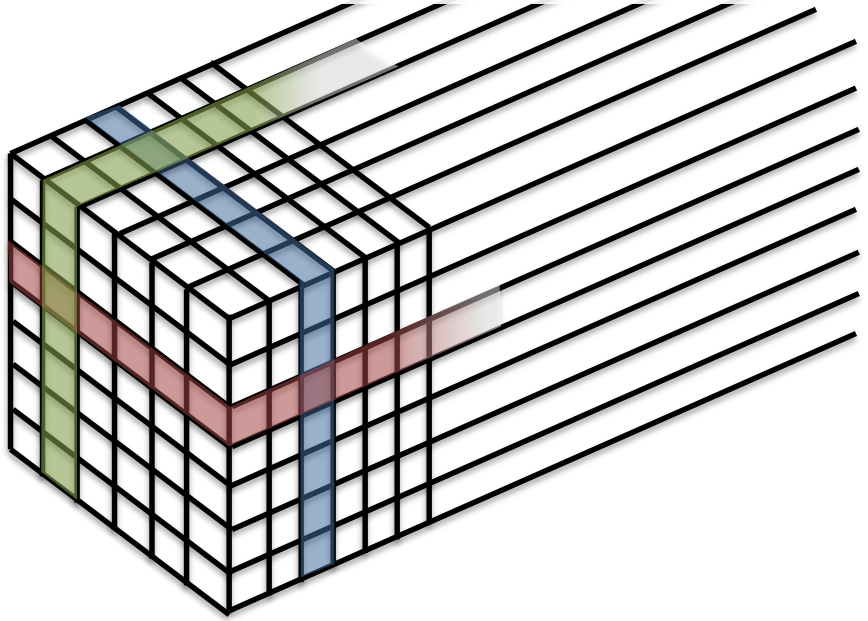
Byte Addressable Hypercubes

- 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 **583 TiB**

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



Excluding: historical observations, multiple models, etc...

Novel Data Flows

Producer

Cloud

Consumer

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

NVRAM
(FDB)

Archive

PFS

HDD

Tape

MARS

Conclusions

1. Diverse hardware and diverse workflows are coming
   - ***Require flexible storage architecture for future-proofing***

2. We are not limited to *coping* with increases in data volumes
   - ***Enable new workflows and data usage patterns***

3. ECMWF is bringing a new Meteorological Object Store into Production
   - ***Adding an abstraction between application and IO stack***
   - ***Hoping to make use of upcoming technologies: Mero, DAOS, pmdk, etc…***

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