ECMWF's Next Generation IO for the IFS Model and Product Generation

Future workflow adaptations

Tiago Quintino, B. Raoult, S. Smart, A. Bonanni, F. Rathgeber, P. Bauer

ECMWF

tiago.quintino@ecmwf.int

ECMWF 17th Workshop on High Performance Computing in Meteorology
Reading, UK

ECMWF

ECMWF's HPC Targets

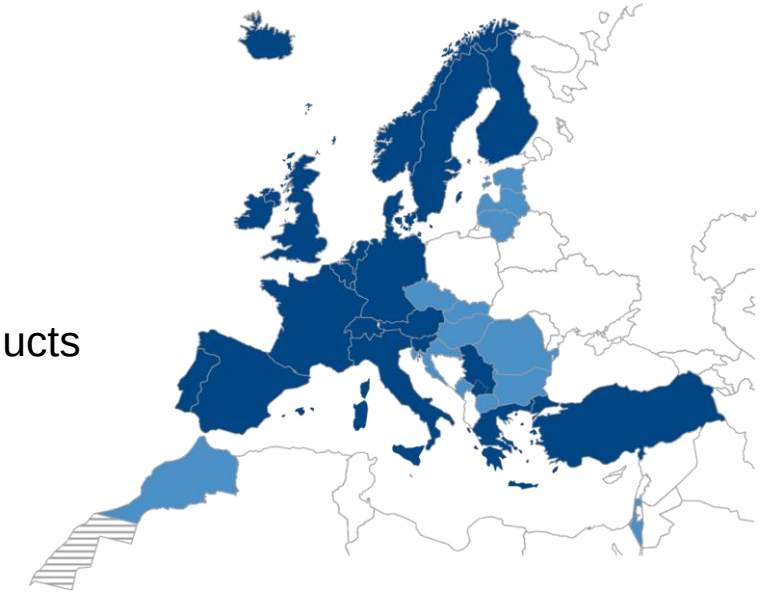
What do we do?

Operations – **Time Critical**

- Operational runs – 2 hours from satelite cut-off to deliver forecast products
- 10 day forecast twice per day, 00Z and 12Z
- Boundary Conditions 06Z and 18Z, monthly, seasonal, etc.

Research – **Non Time Critical**

- Improving our models
- Climate reanalysis, etc



HPC Facility Targets

- **Capability**, minimise the time to solution of Model runs
- **Capacity**, maximise the throughput of research jobs per day

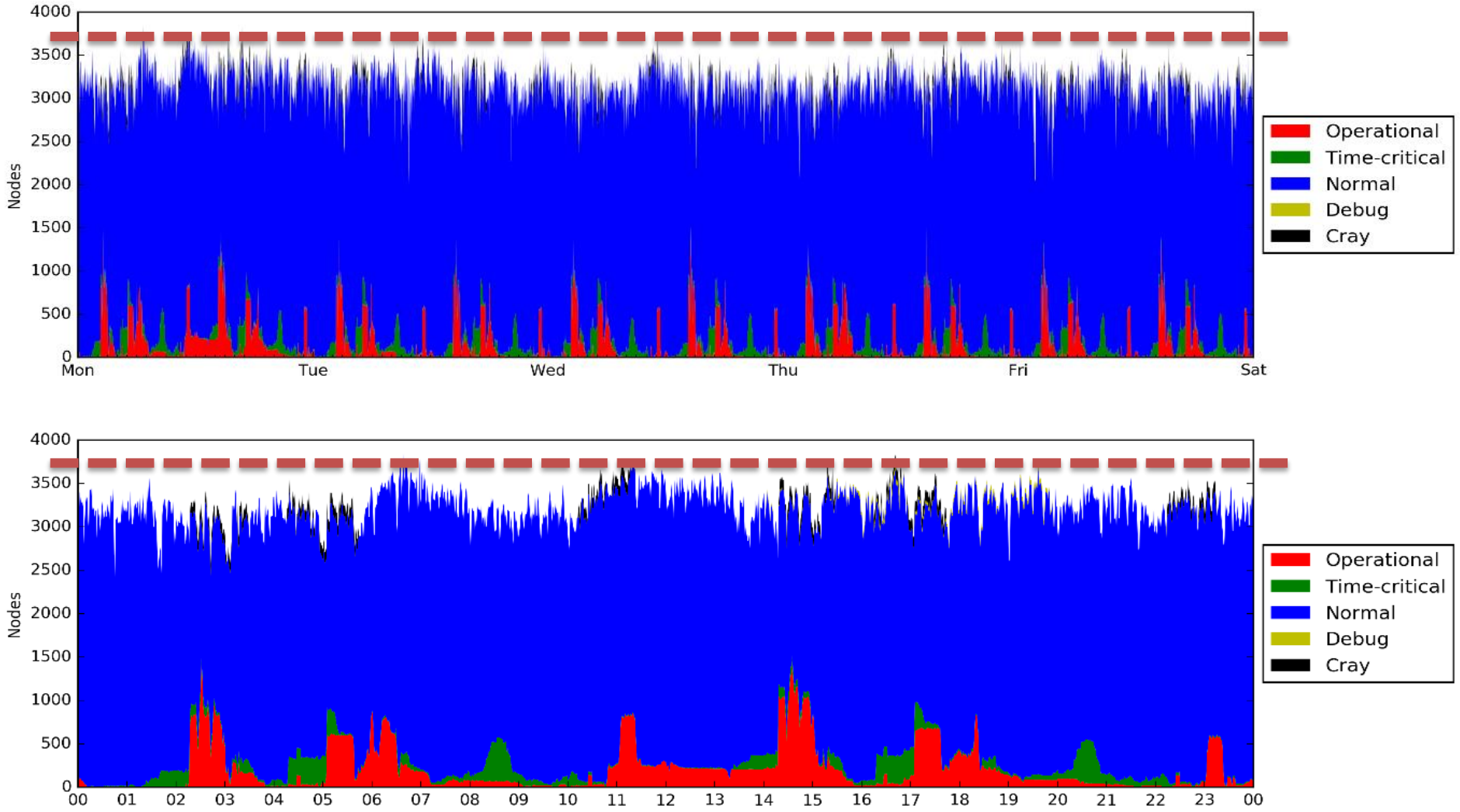
Tension

Time Critical vs. **Non Time Critical**

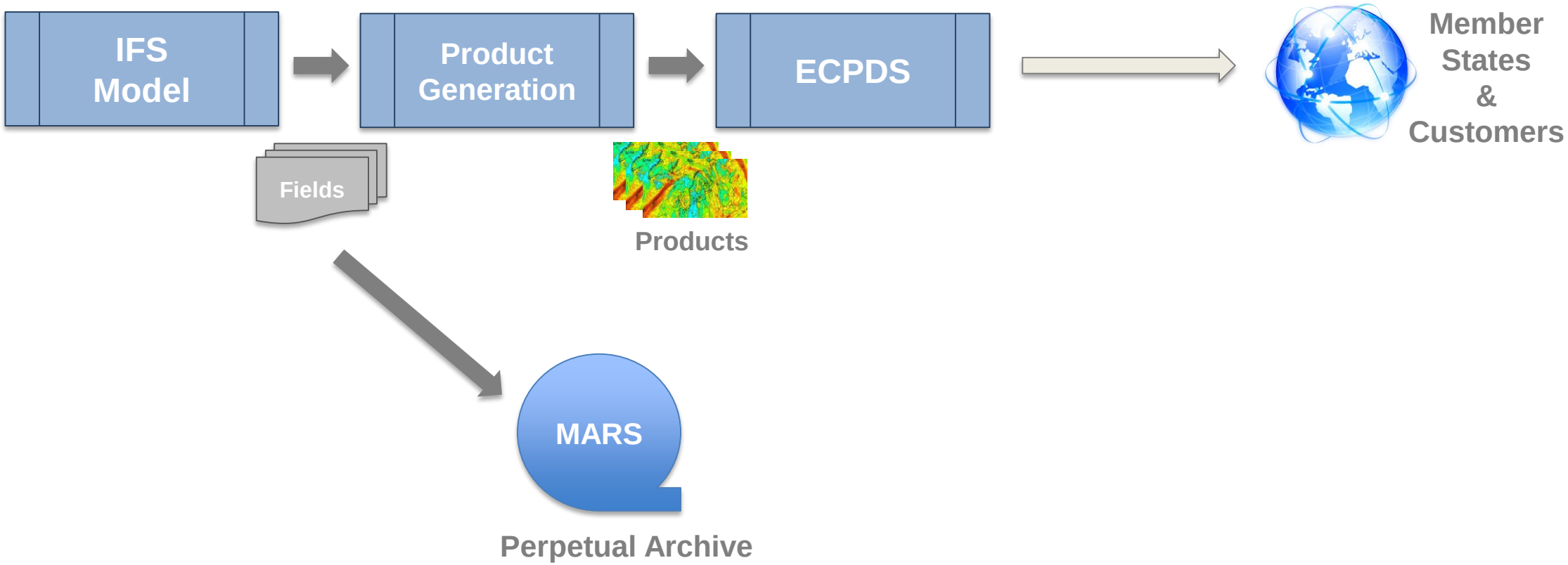
Capacity vs. **Capability**

***Challenge**: design our HPC system to optimise these goals, minimising TCO?*

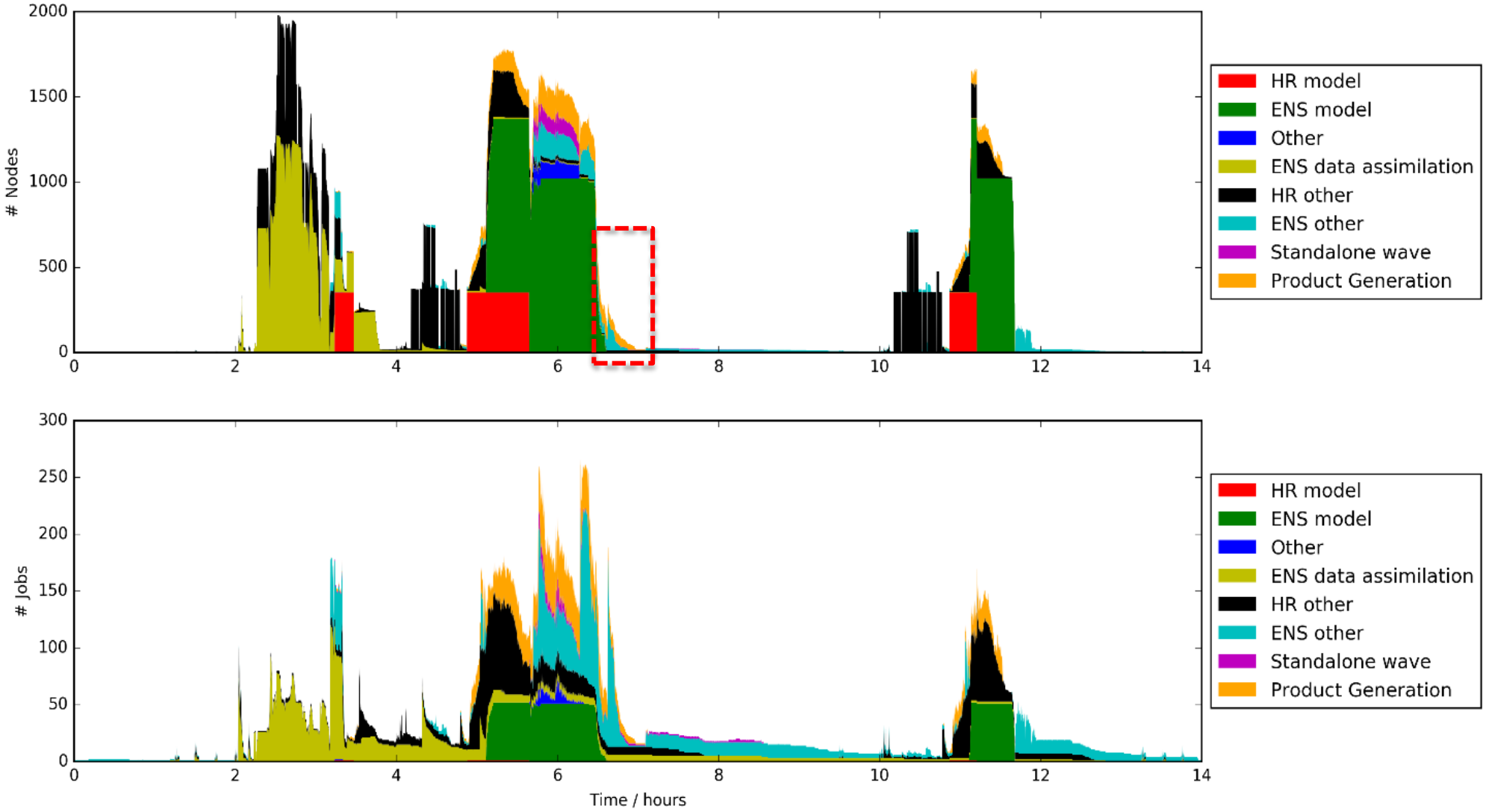
ECMWF HPC Job profile

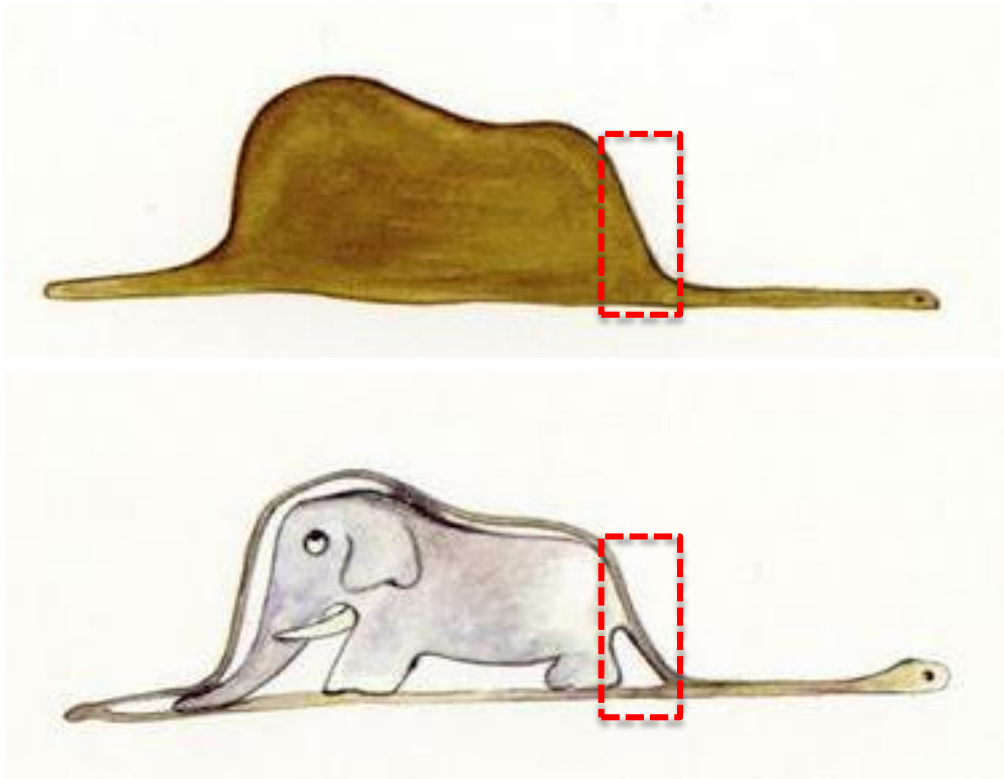


ECMWF's Production Workflow



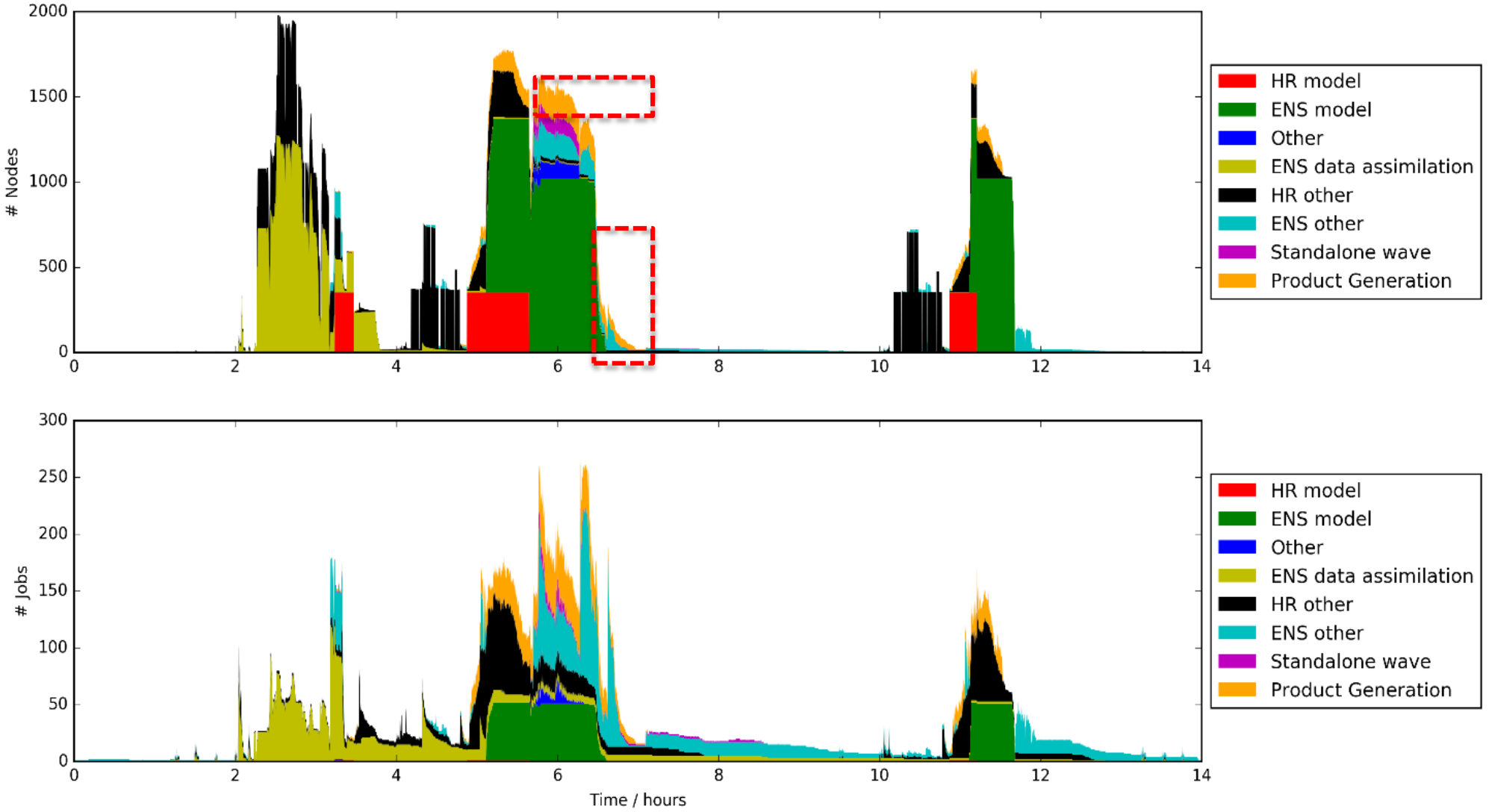
Operational workload: Job allocation (1 cycle)



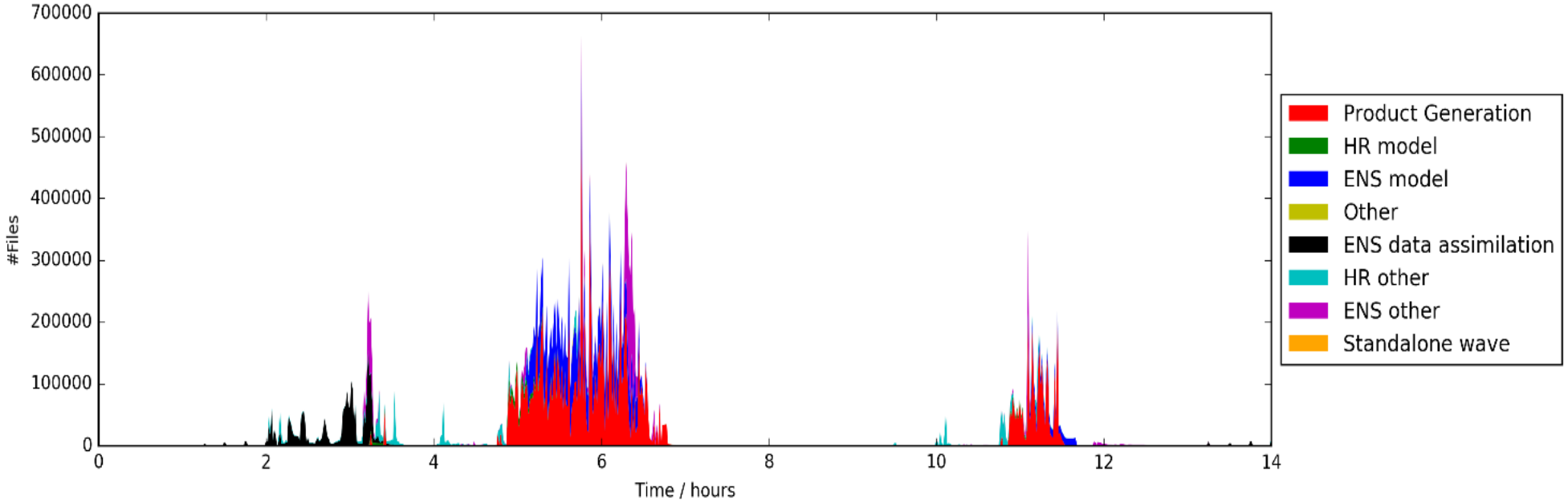


Le Petit Prince, Antoine de Saint-Exupéry

Operational workload: Job allocation (1 cycle)

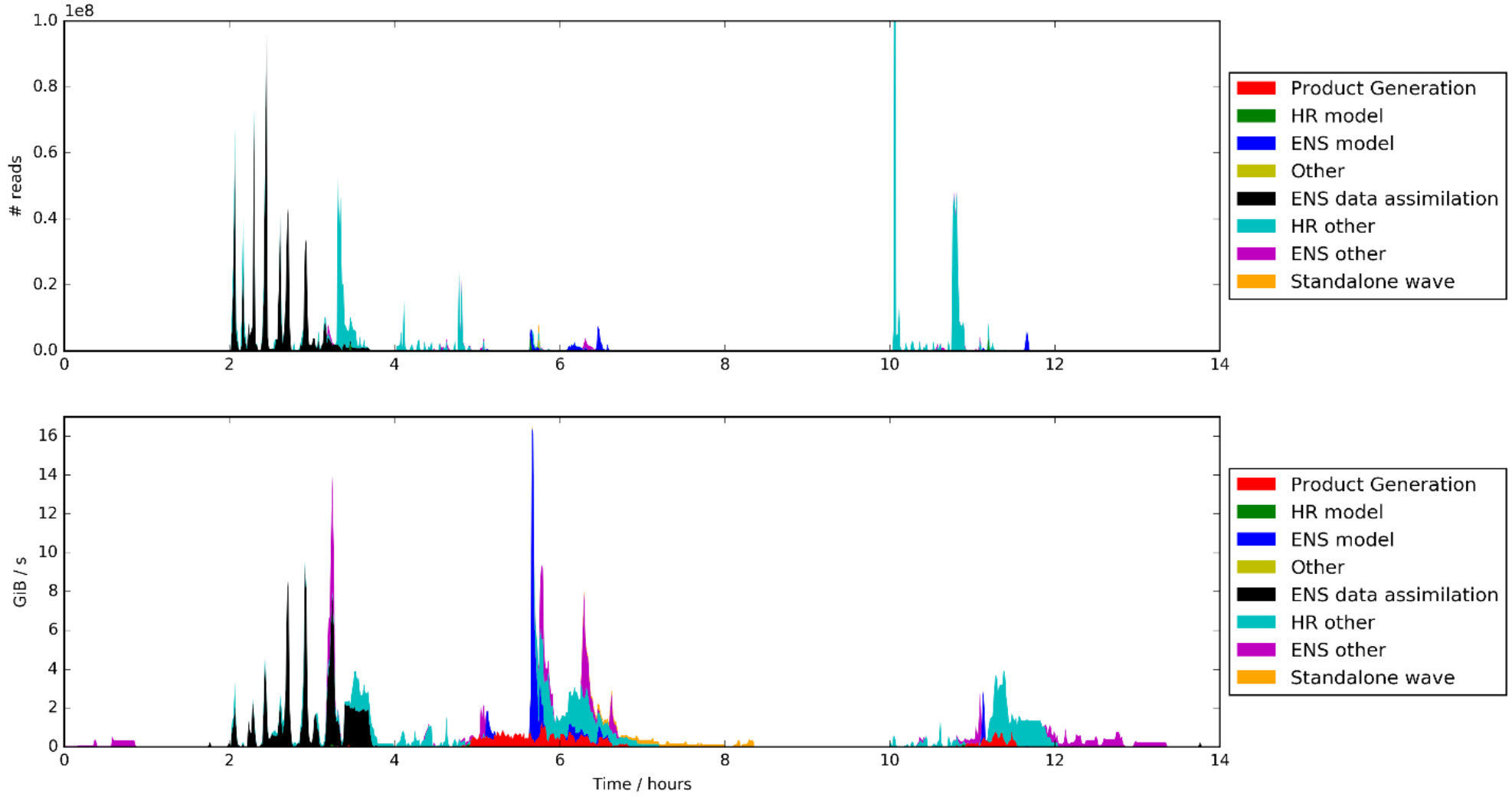


Operational workload: Files opened (1 cycle)

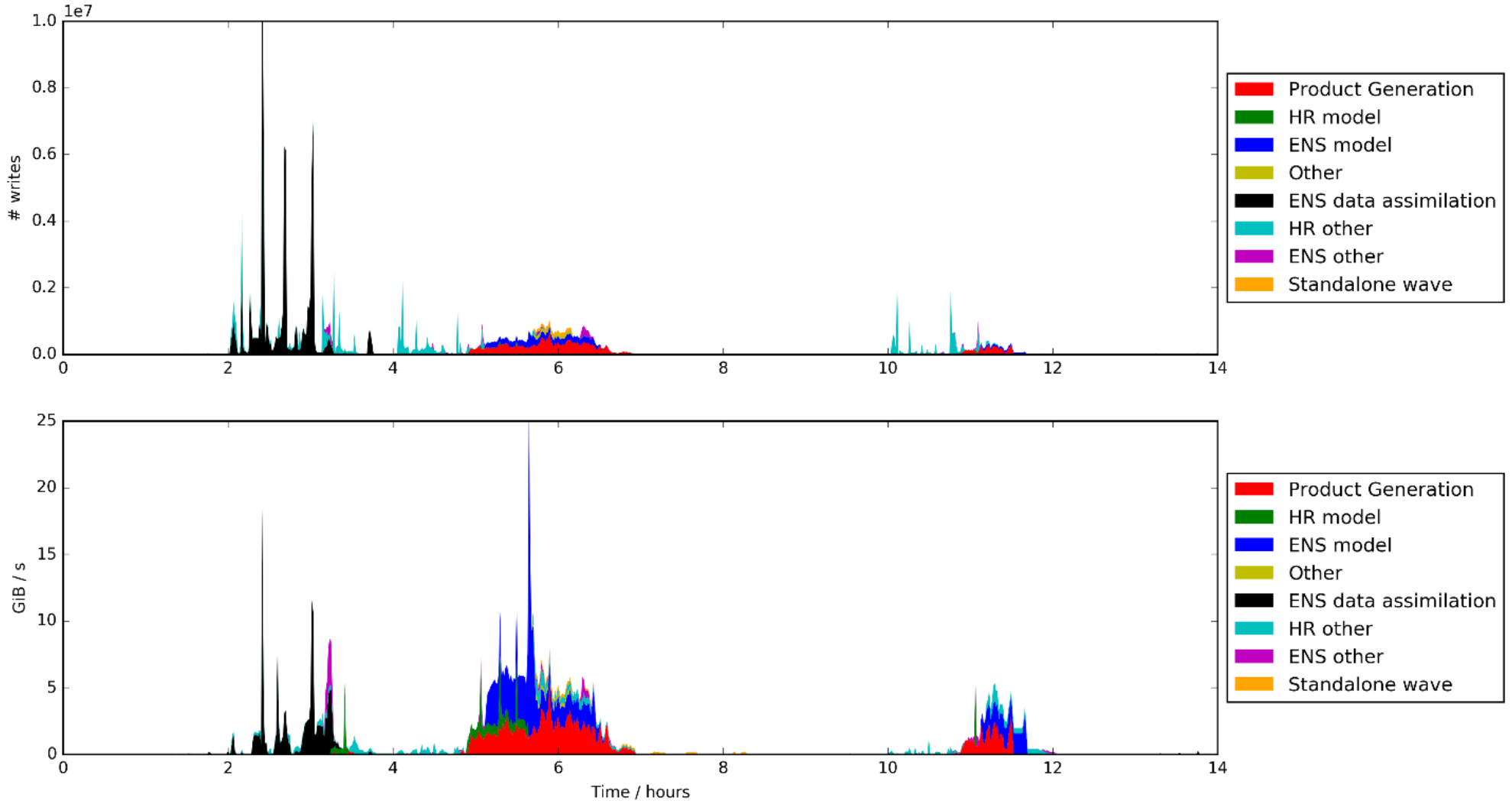


Target Files = # Users x # Steps x # Ranks

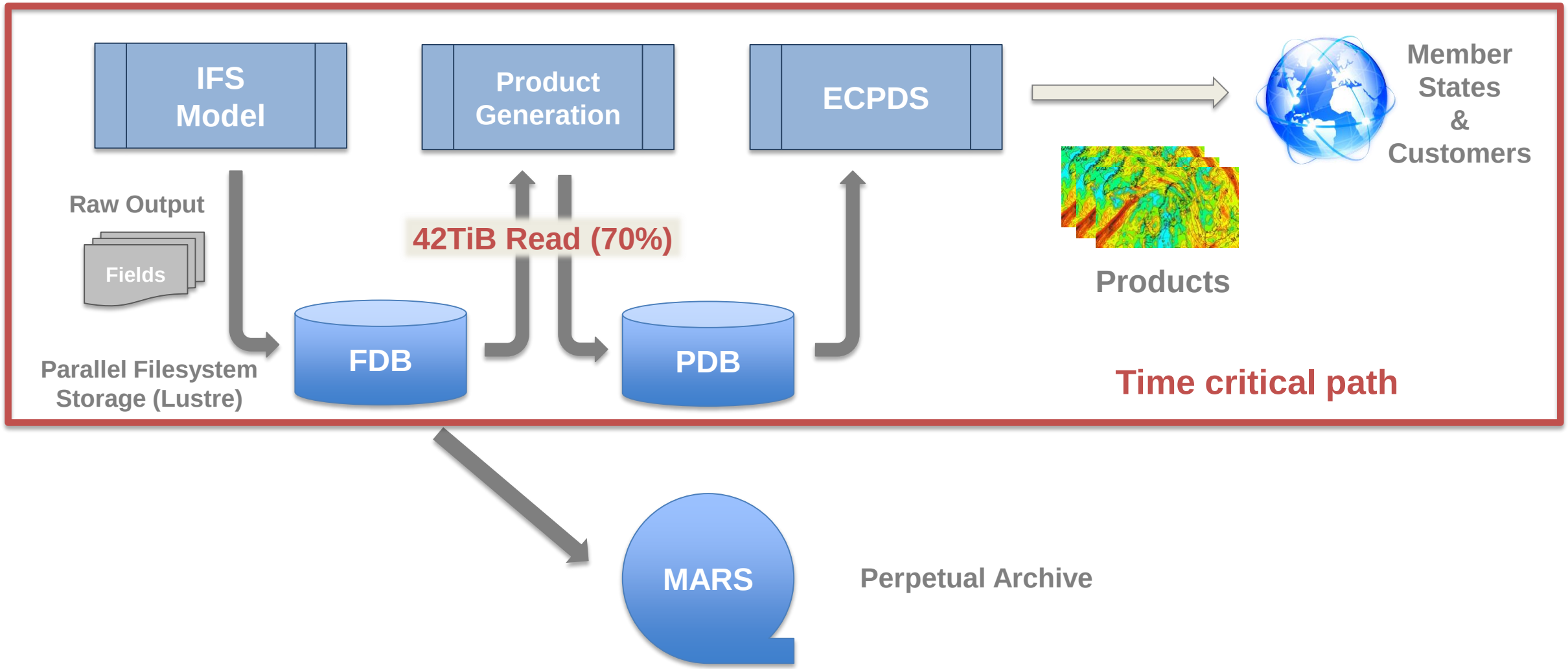
Operational workload: Input Read (1 cycle)



Operations workload: Output written (1 cycle)



ECMWF's Production Workflow



Estimated Growth in Model IO

2015

16km, 137 levels

Time critical

- 21 TB/day written
- 22 Million fields
- 85 Million products
- 11 TB/day send to customers

Non-time critical

- 100 TB/day archived
- 400 research experiments
- 400,000 jobs / day

2020

Increase: 2 horizontal, 1 upper air

Time critical

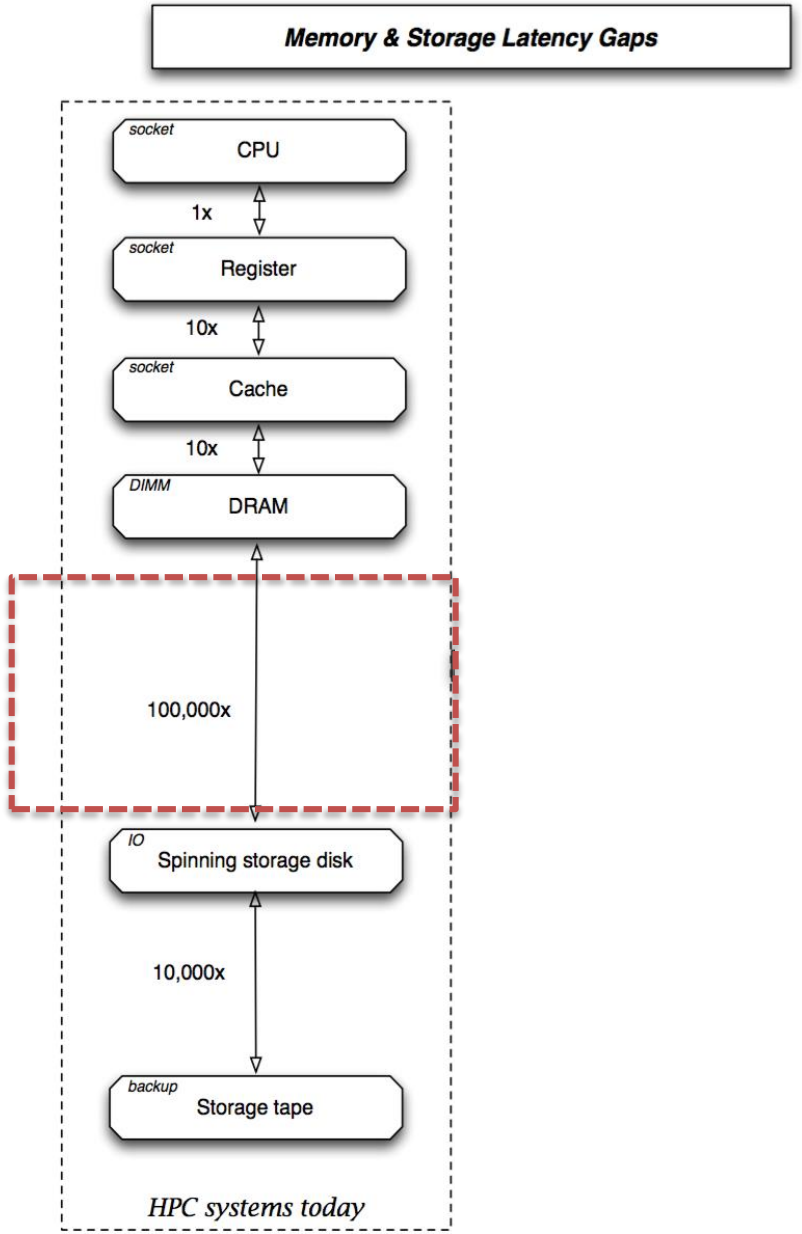
- 128 TB/day written
- 90 Million fields
- 450 Million products
- 60 TB/day send to customers

Non-time critical

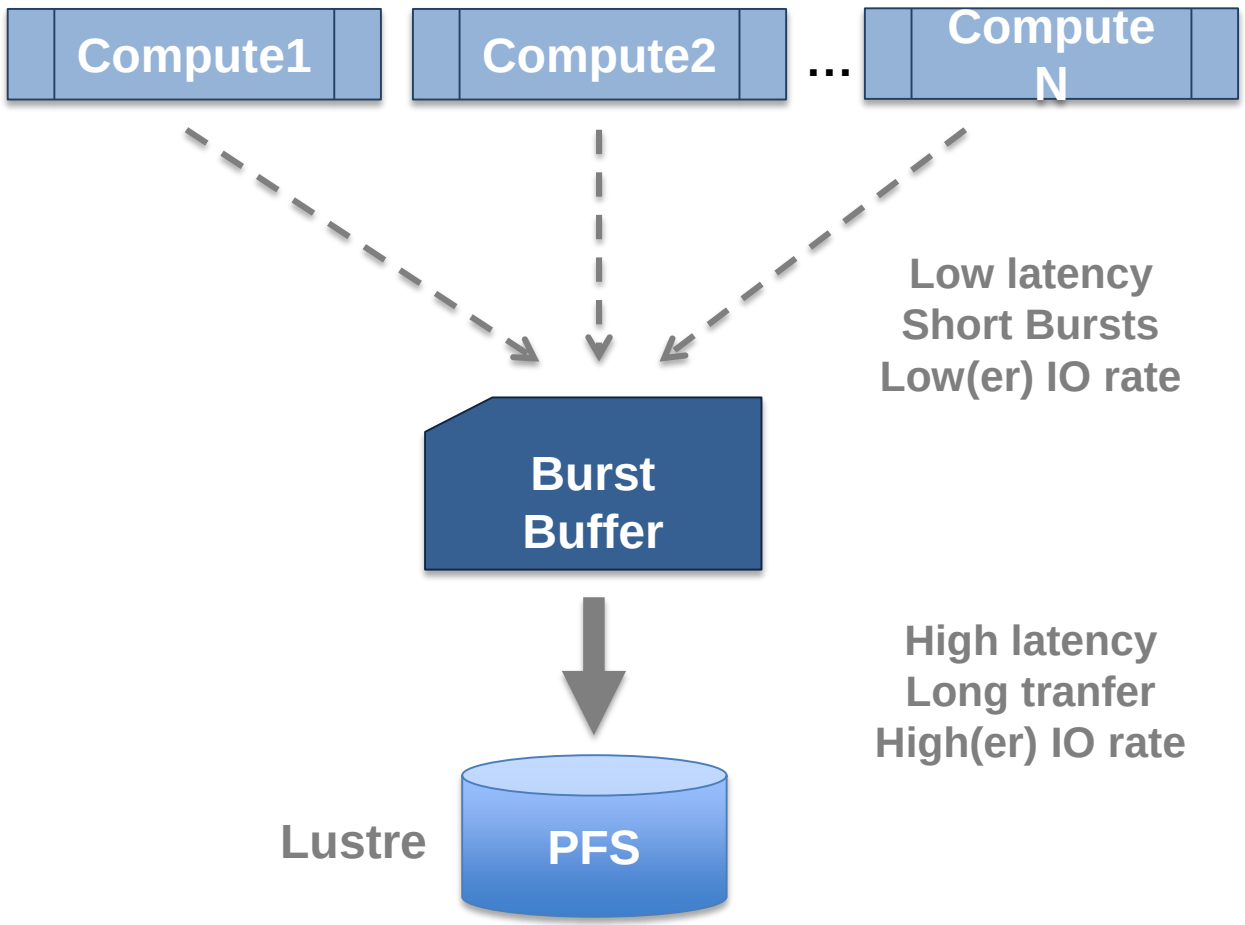
- 1 PB/day archived
- 1000 research experiments
- 1,000,000 jobs / day

Feeling the Byte?

I/O Gap



Burst Buffers



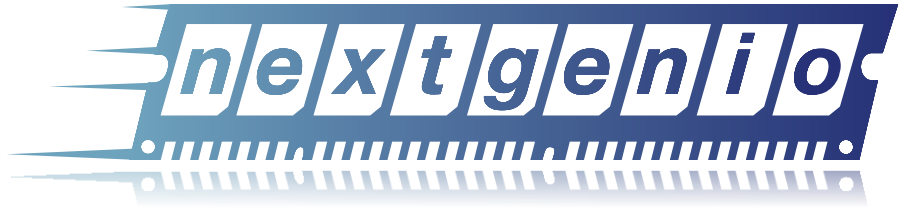
- Initially designed for check pointing
- Used to absorb IO peaks
- Layered on top of PFS
- Application sees **persistence** with low latency

- Concerns …
  - **Sharding vs Consistency vs Shared**
  - **Data replication (resilience)**
  - **POSIX file system interface (still)**

What if we could **change** the application?

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

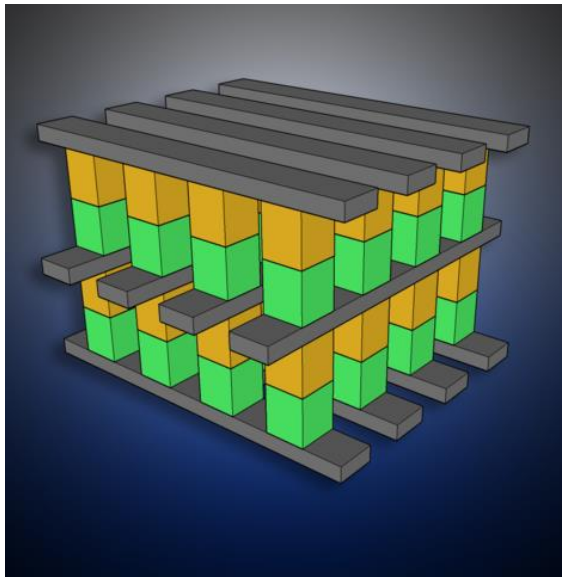
http://www.nextgenio.eu - EU funded H2020 project, runs 2015-2018

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

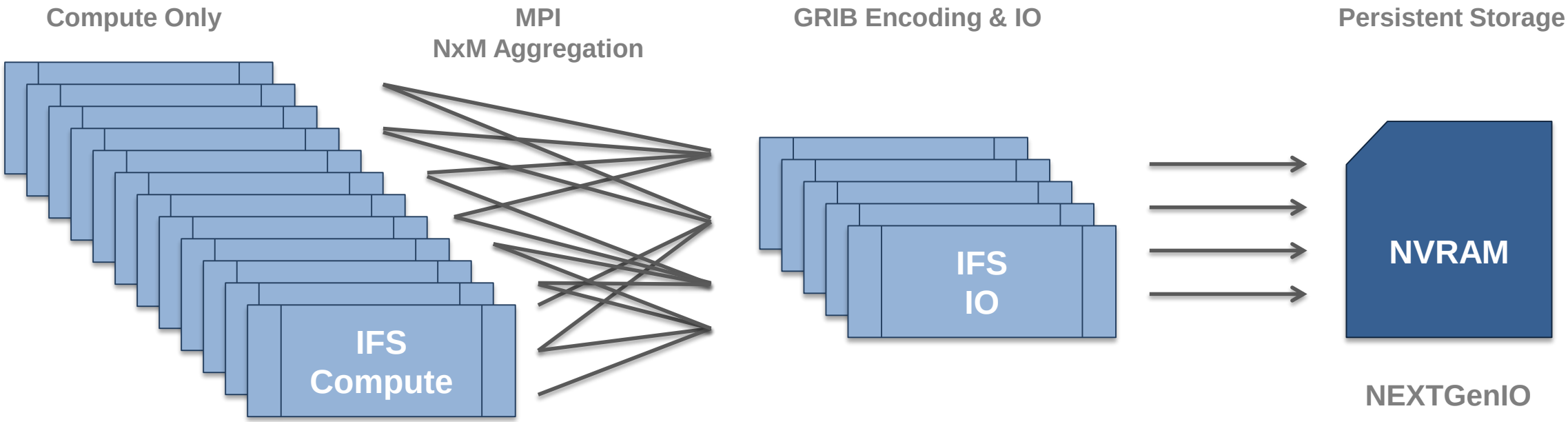
How is ECMWF planning to use this technology?

- **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

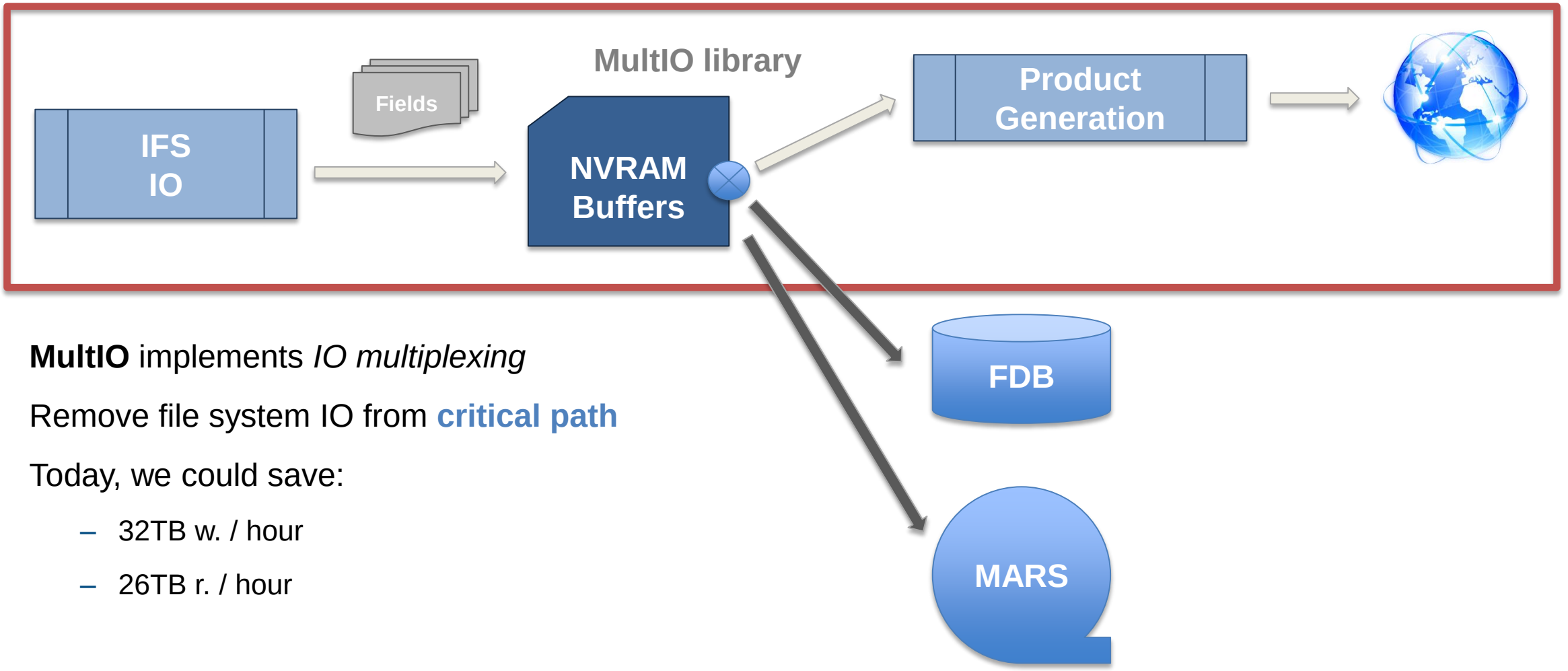
Key Point: High Density at very low latency

IFS IO Server

- Based on MeteoFrance IO server for IFS
- Entered production in March 2016



Streaming Model Output to a Computing Service



MultIO implements *IO multiplexing*

Remove file system IO from **critical path**

Today, we could save:

- 32TB w. / hour
- 26TB r. / hour

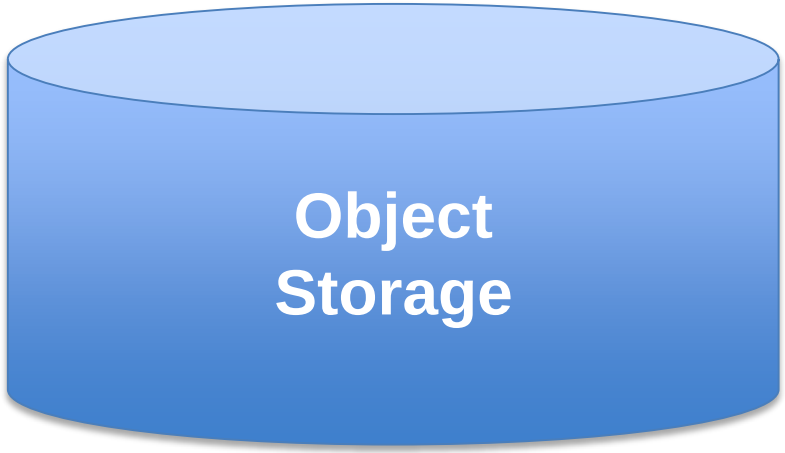
How to store all model output in NVRAM buffers?

Object Store

- Key-Value stores offer **scalability**
  - Just add more instances to increase capacity and throuput
- **Transaction** behavior with minimal synchronization
- Growing popularity, namely due to **Big Data Analytics**

Key: date=12012007, param=temp

Value: 101001...100101010110010



But ECMWF has been using key-value store for 30 years...

MARS

MARS Language

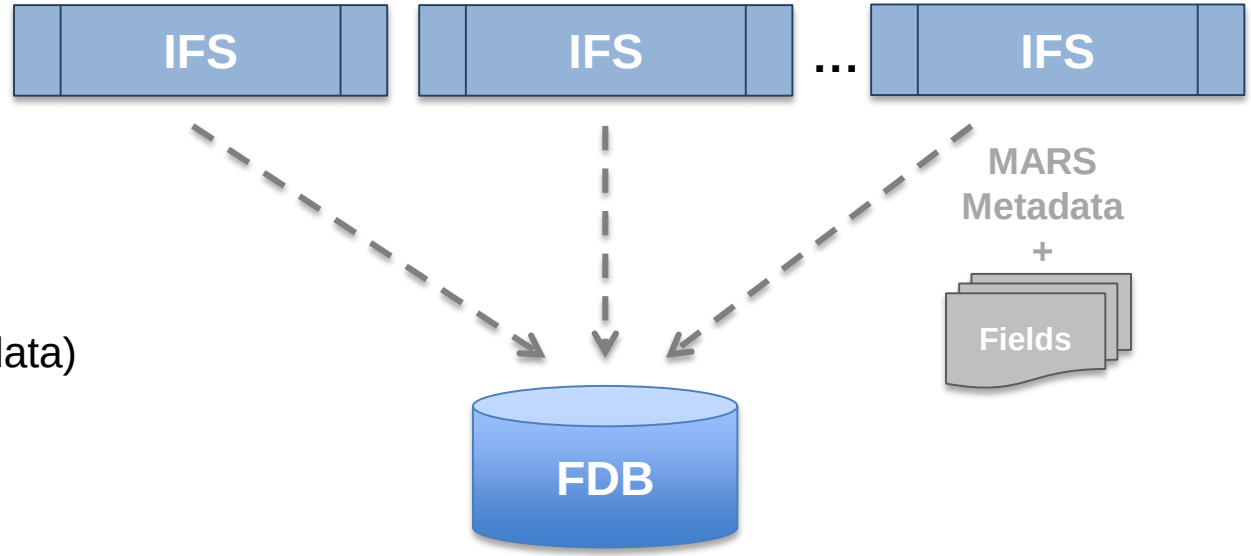
```
RETRIEVE,
    CLASS     = OD,
    TYPE      = FC,
    LEVTYPE   = PL,
    EXPVER    = 0001,
    STREAM    = OPER,
    PARAM     = Z/T,
    TIME      = 1200,
    LEVELIST  = 1000/500,
    DATE      = 20160517,
    STEP      = 12/24/36
```

```
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**

FDB (version 5)

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



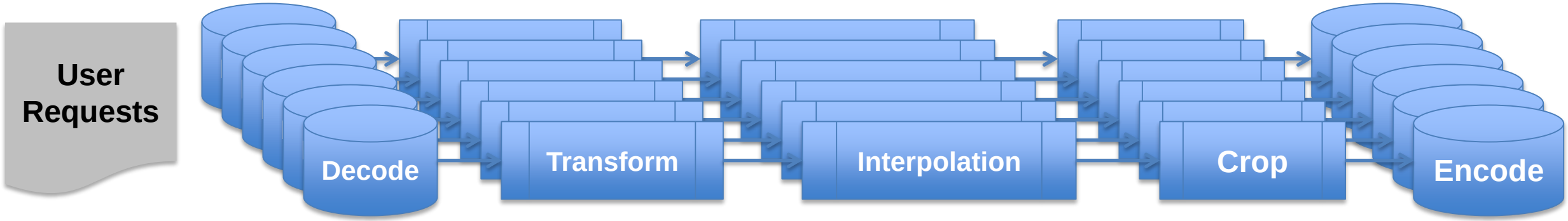
- Support for multiple back-ends:
  - POSIX file-system (currently on Lustre)
  - 3D XPoint using pmem.io library
  - Could explore others:
    - Intel DAOS, Cray DataWarp, etc.

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

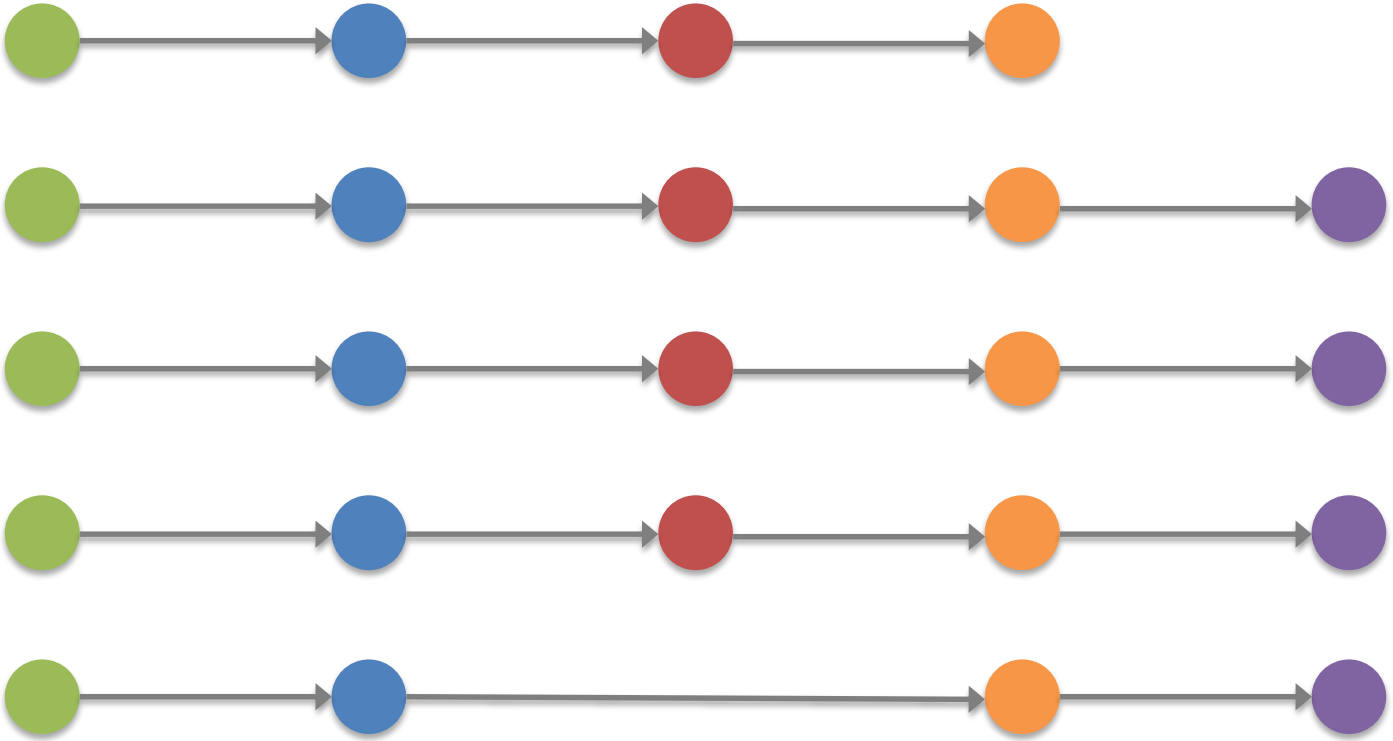
param=temperature/humidity,
levels=all,
steps=0/240/by/3
date=01011999/to/31122015,

Product Generation - PGen

- Rewrite in C++
- Based on ...
  - New interpolation software (**MIR**)
  - **Caching** algorithms for operators
- (Explicit) **Task Graph** analysis
  - ***Remember: users can update requests daily***
  - Factorise common tasks
  - Batch and Reorder execution
  - Compute time-series on-the-fly

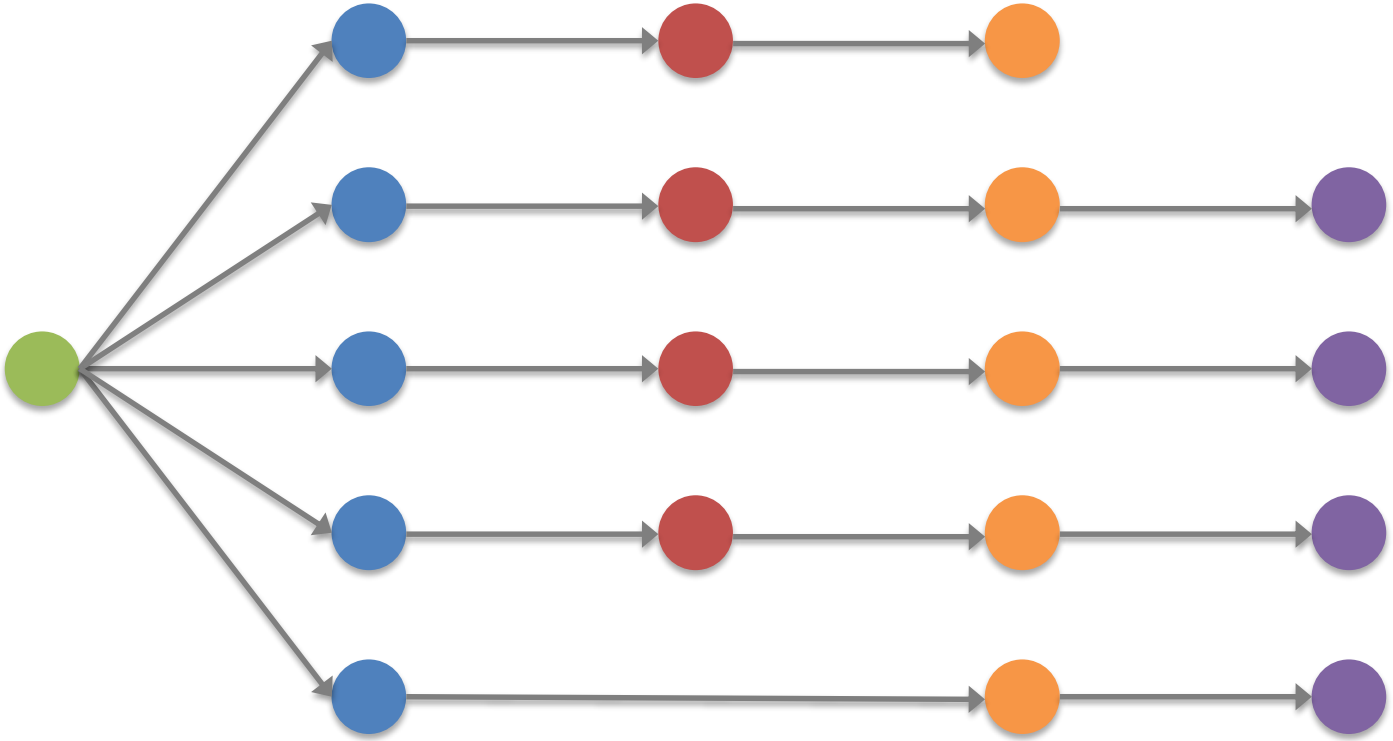


PGen – Task Graph Analysis



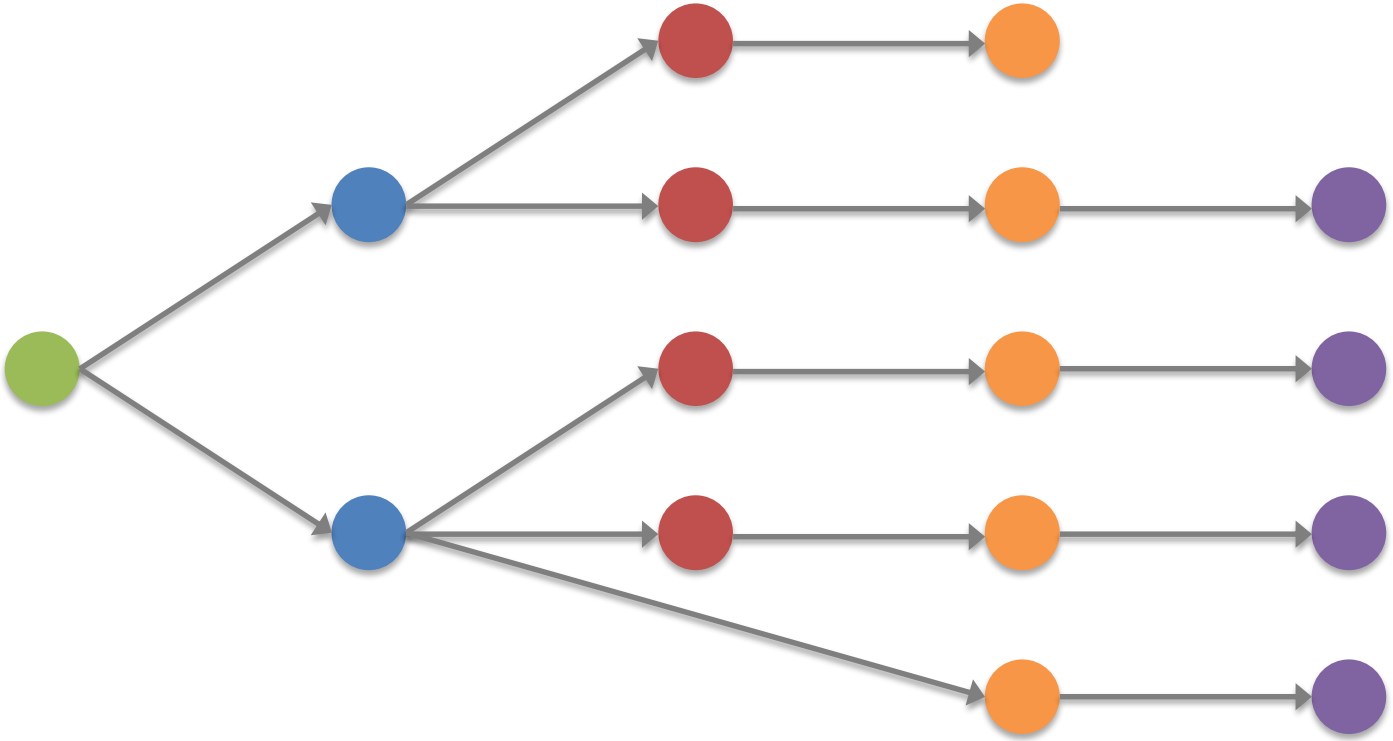
PGen – Task Graph Analysis

Merge same input



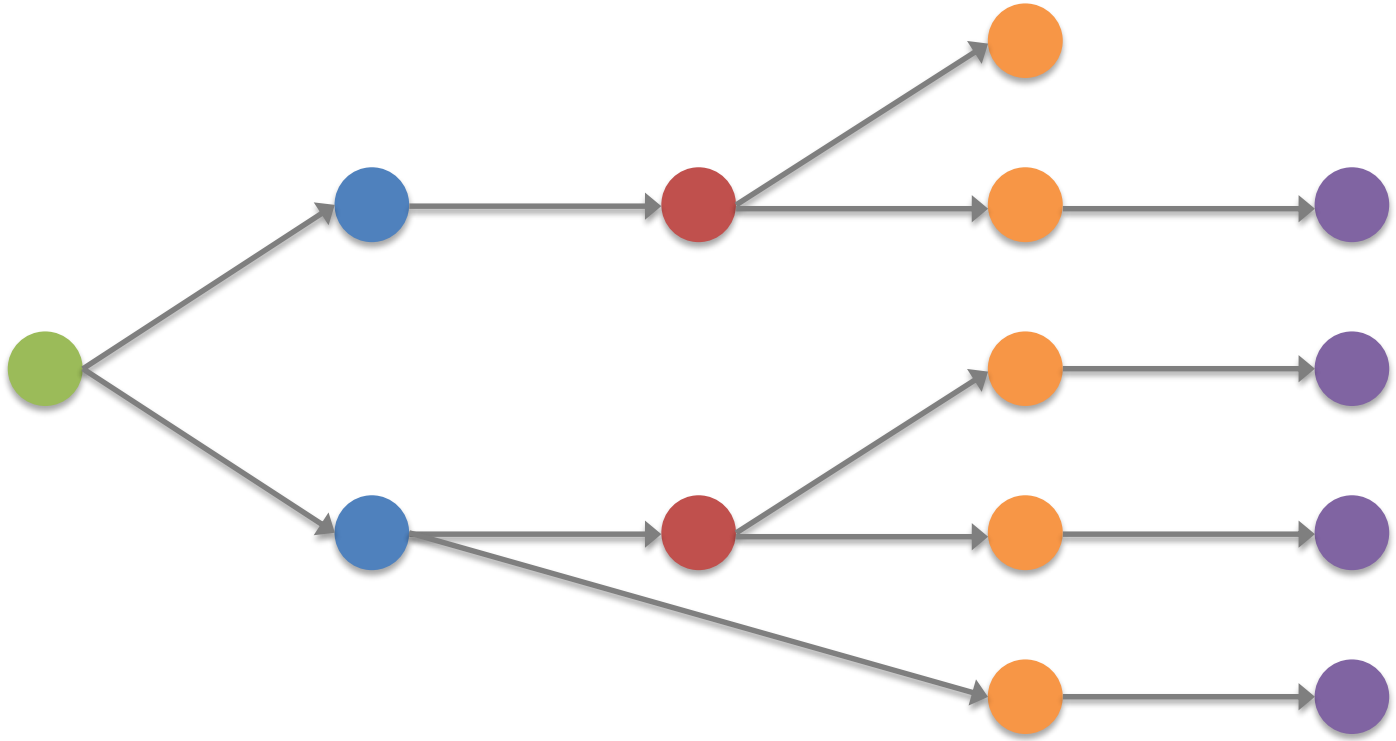
PGen – Task Graph Analysis

Merge same SH transforms



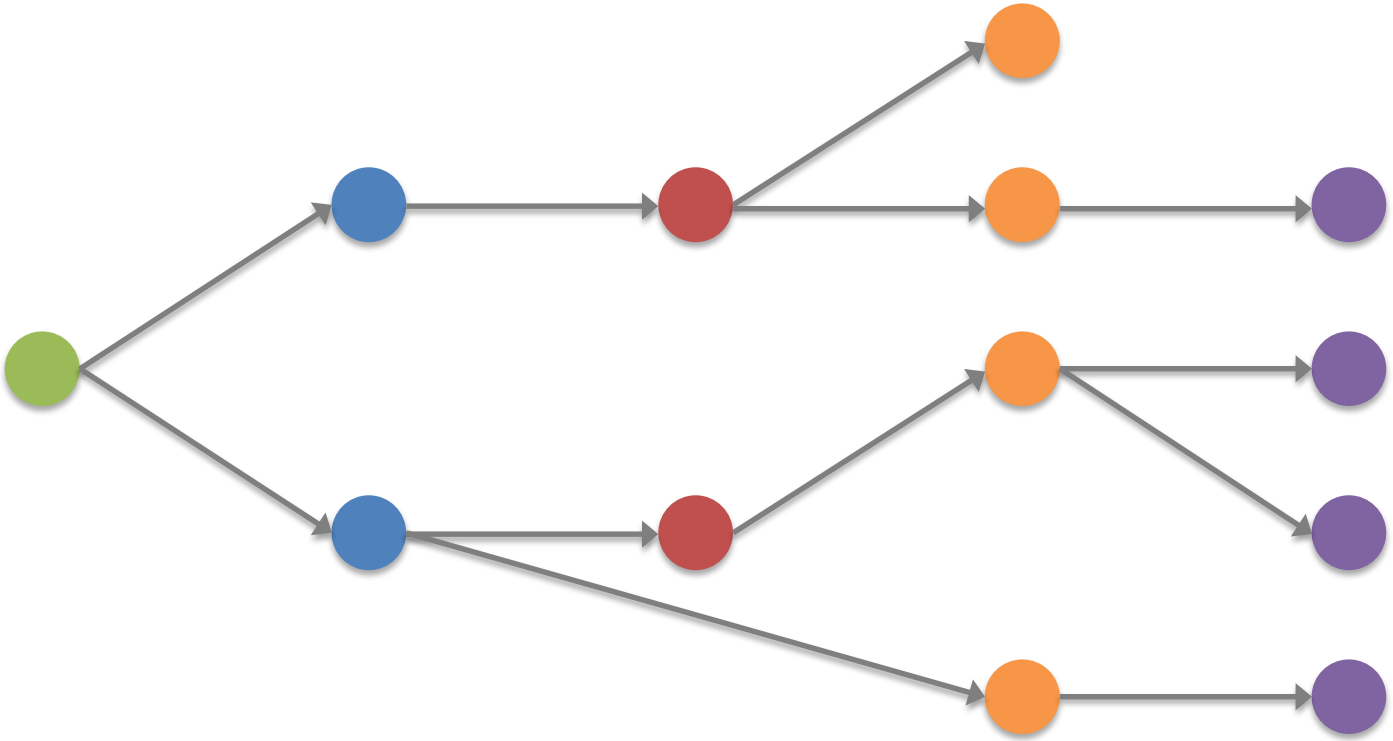
PGen – Task Graph Analysis

Merge same interpolation target grids



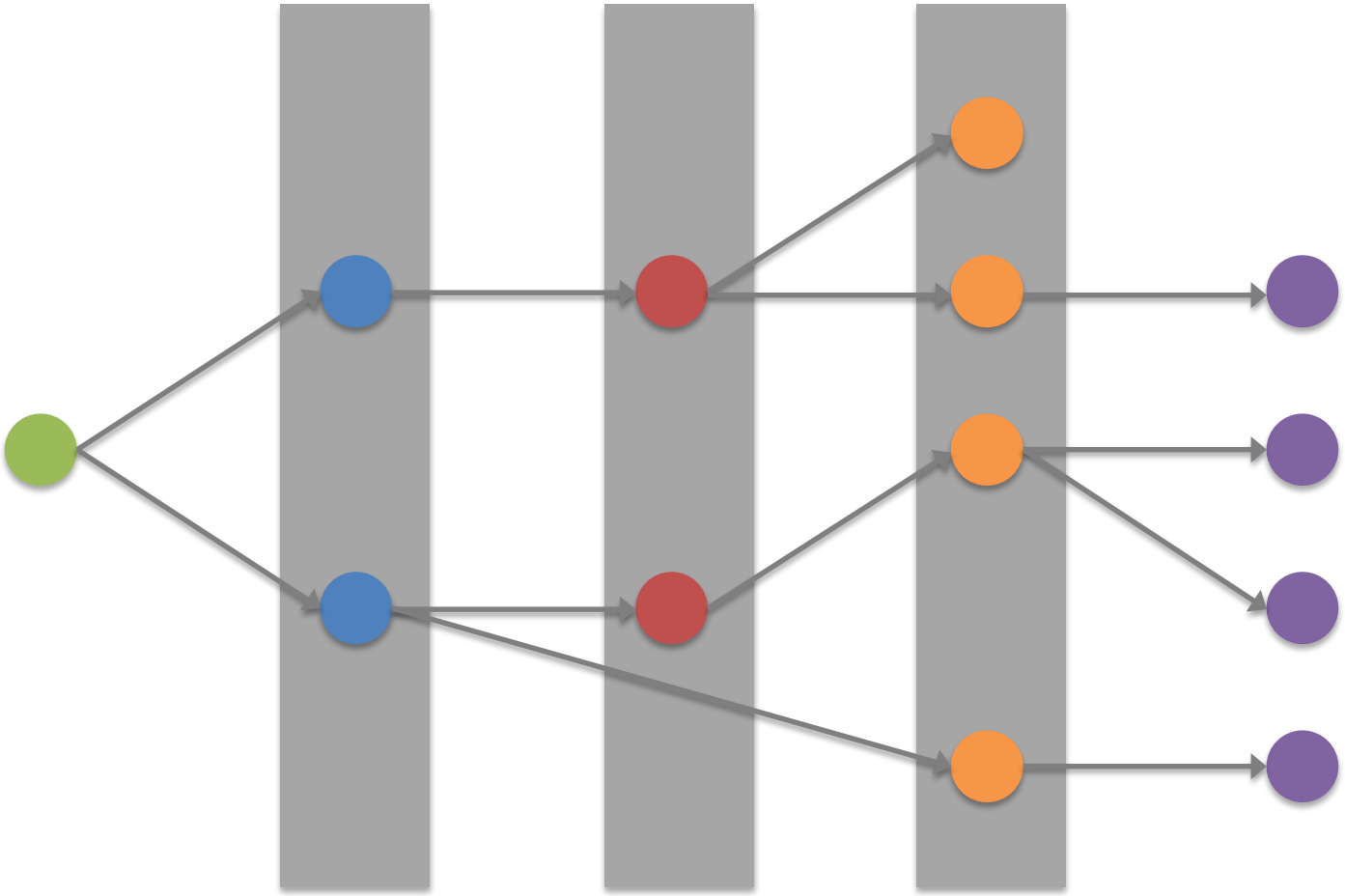
PGen – Task Graph Analysis

Merge same rotation and cropping

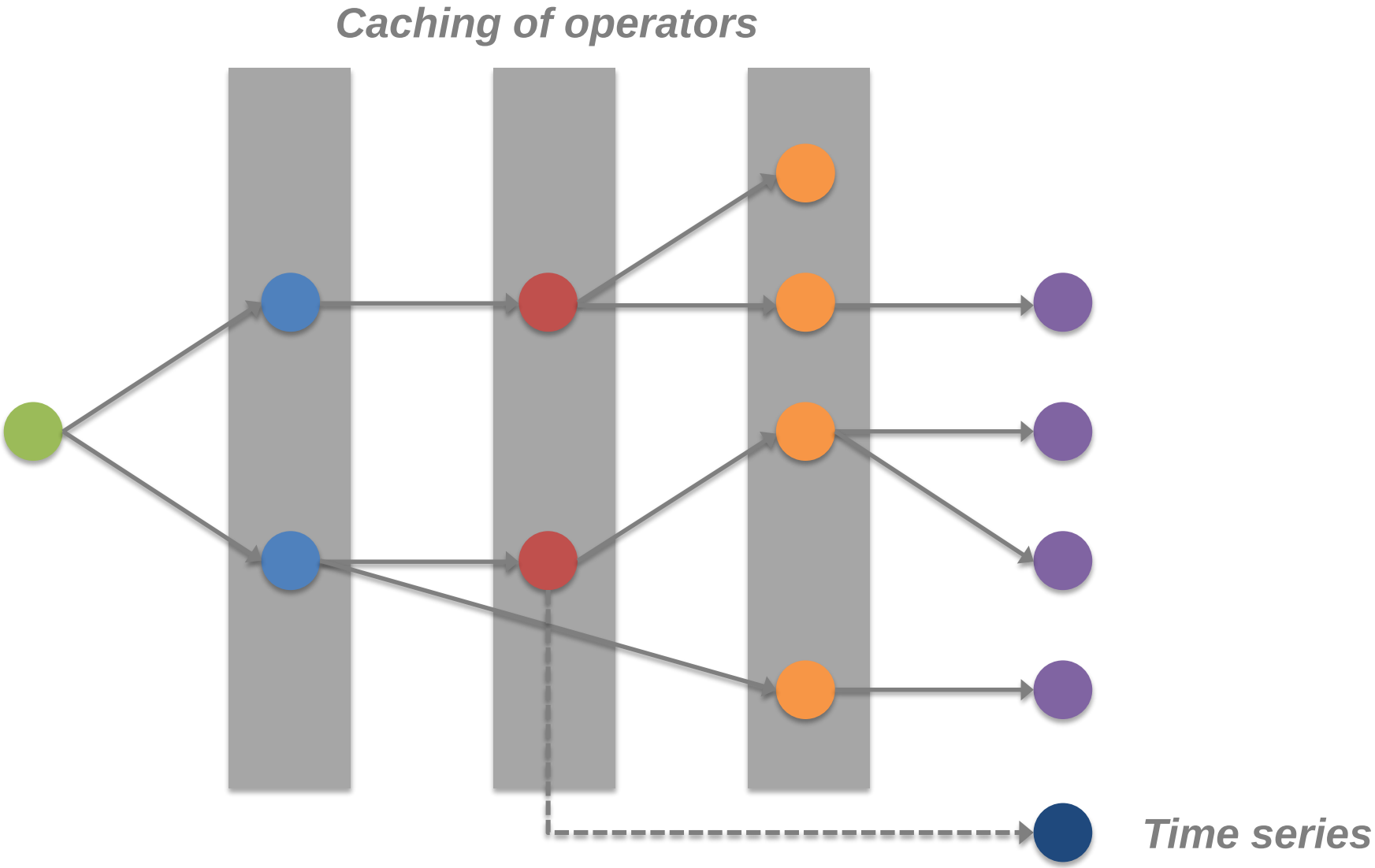


PGen – Task Graph Analysis

Caching of operators

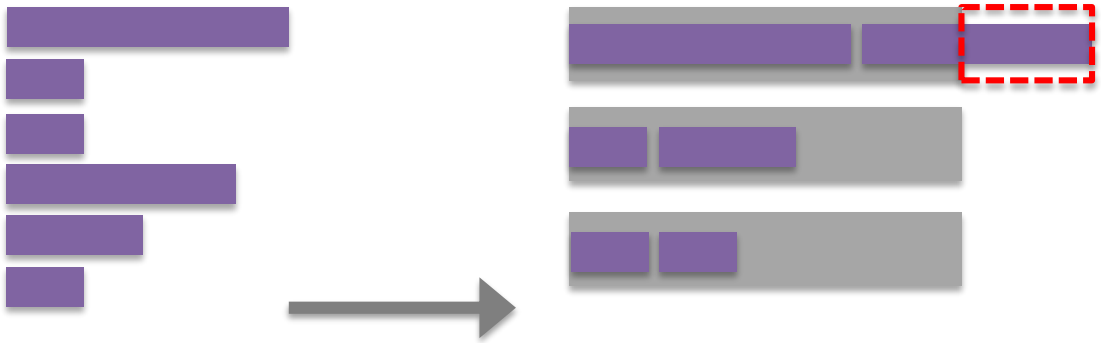


PGen – Task Graph Analysis

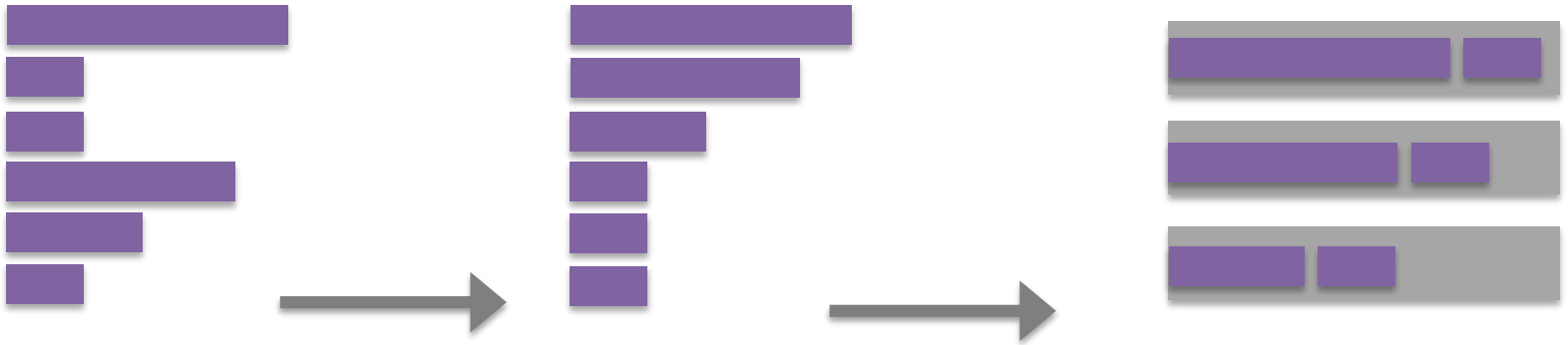


PGen – Task Reordering for Dynamic Load Balancing

Example: 6 tasks, 3 workers



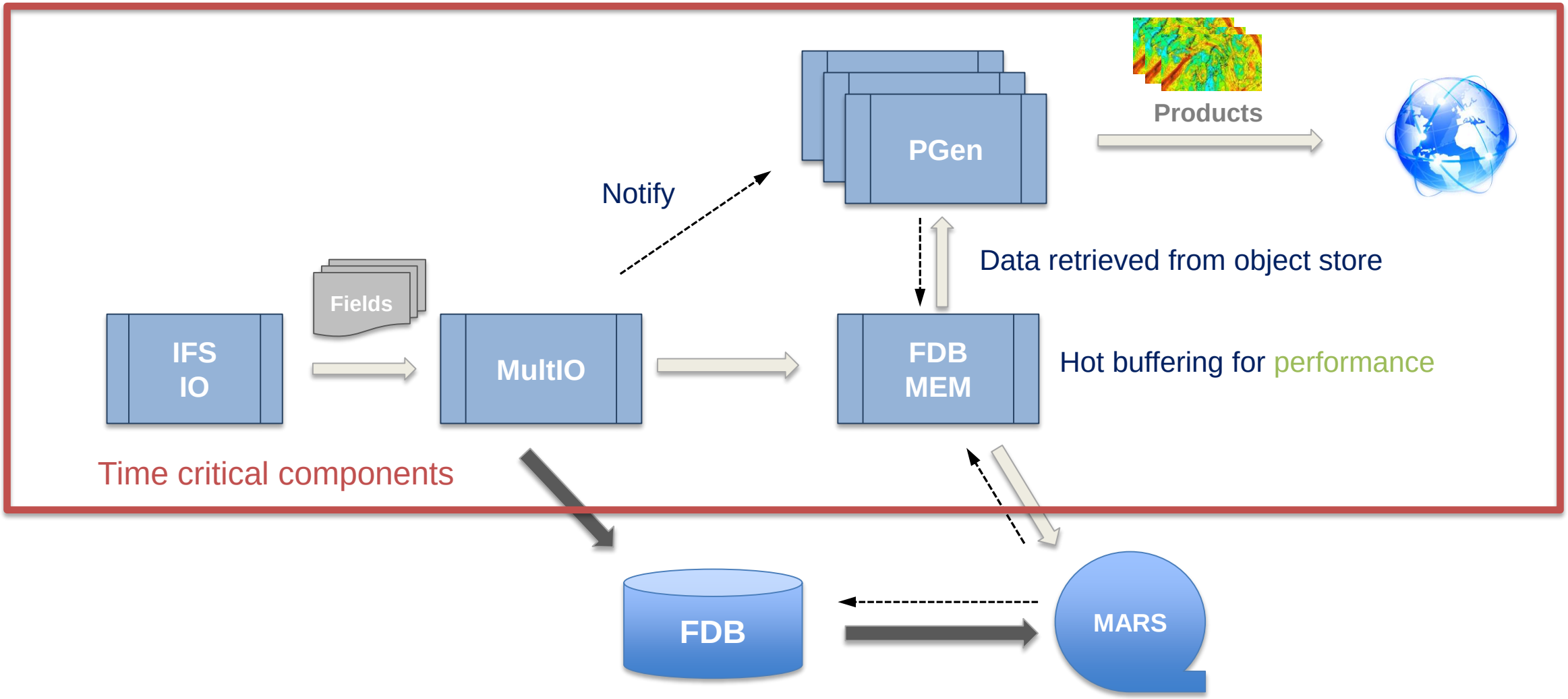
FIFO dispatch



Heuristic Reorder

FIFO dispatch

Summary : Overall Infrastructure Plan



Messages To Take Home

*Burst Buffers, SSD's, NVRAM, are filling in the **I/O Gap***
and will change the way we use and store data

ECMWF is adapting its workflow to take advantage of these upcoming technologies (MultIO, FDB5, MIR, PGen)

What would you do differently,
if your persistent storage would be 10,000x faster?

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