

ECMWF's Next Generation IO for the IFS Model

Part of ECMWF's Scalability Programme

Tiago Quintino, B. Raoult, P. Bauer

ECMWF

tiago.quintino@ecmwf.int

ECMWF's HPC Targets

What do we do?

Operations – **Time Critical**

- operational runs – 2 hours from satellite 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**

IO Profile

- Daily IO profile includes peaks due to **time critical** runs
- Peaks require an otherwise *oversized parallel storage* system

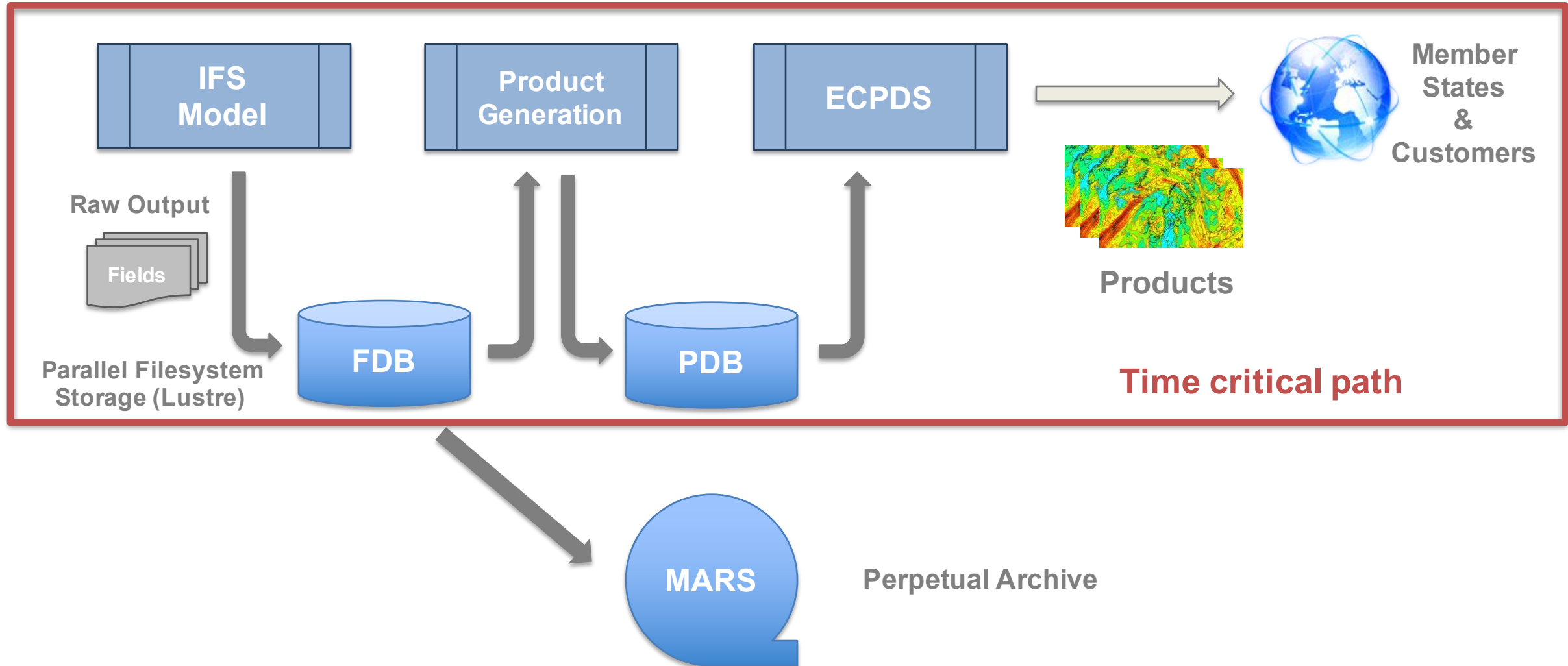
- **Objective:**

- soften these peaks by optimizing application IO
- co-design HPC system with vendors
- maintain **capability** (reduce TCO)
- increase **capacity** (more research runs)

Can we **reduce** the need for *oversized parallel storage* system?



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

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

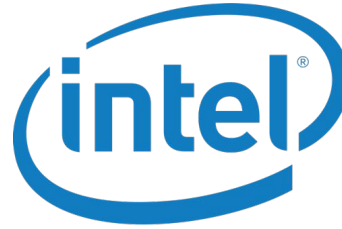
- Build an HPC prototype system with Intel 3D XPoint technology
- Develop tools and systemware to support application development
- Design scheduler strategies that take NVRAM into account
- Explore how to best use this technology in I/O servers

ECMWF Tasks

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

<http://www.nextgenio.eu> - EU funded H2020 project

NVRAM Intel 3D XPoint



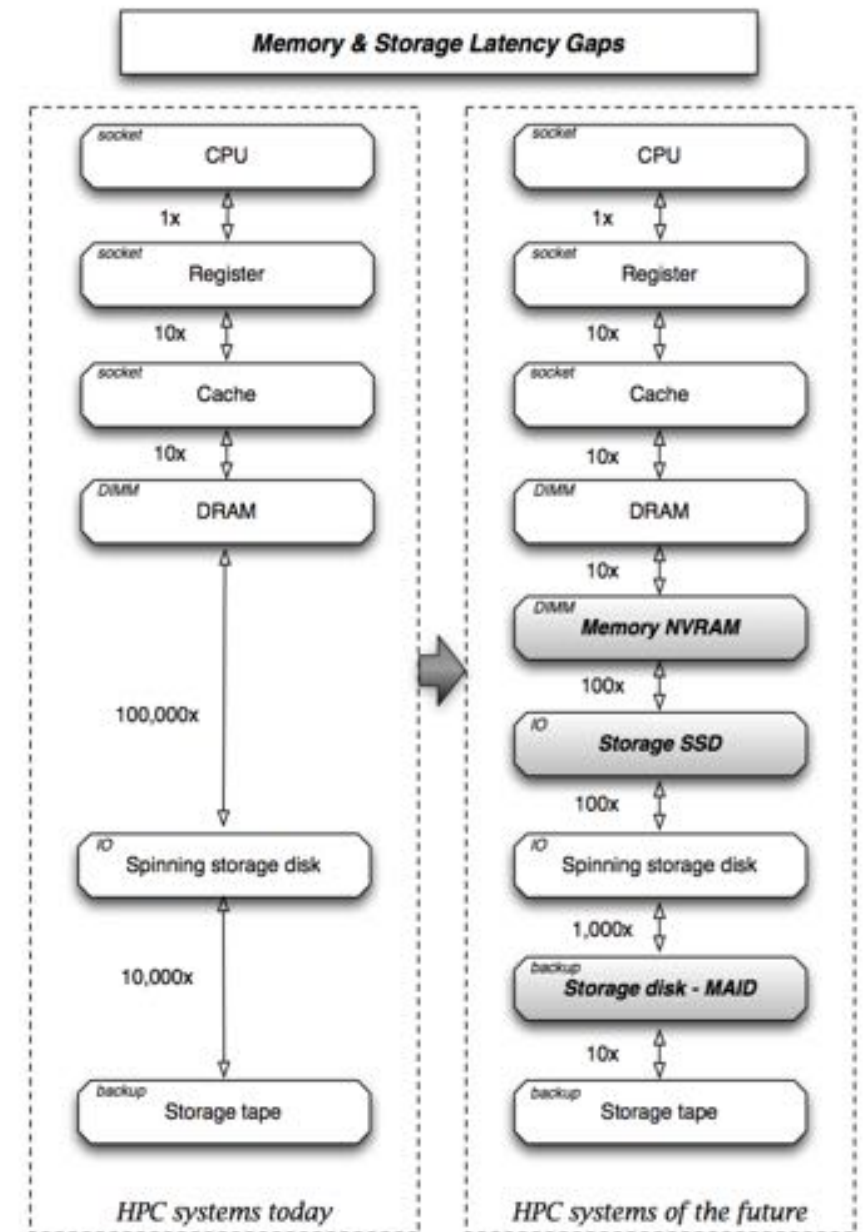
Key characteristics:

- 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

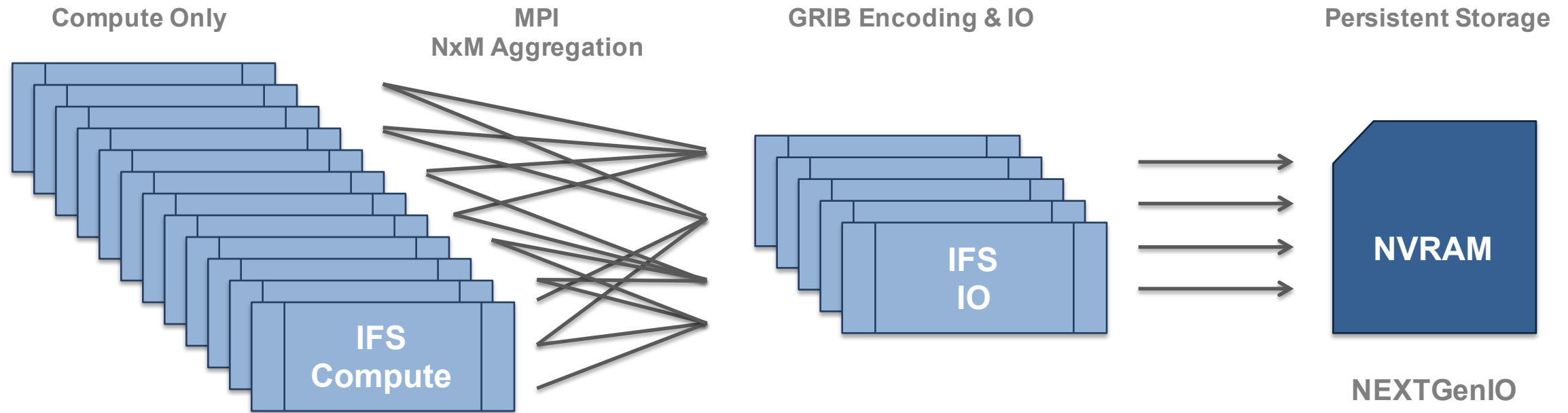
How is ECMWF planning to use this technology?

- **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



IFS IO Server

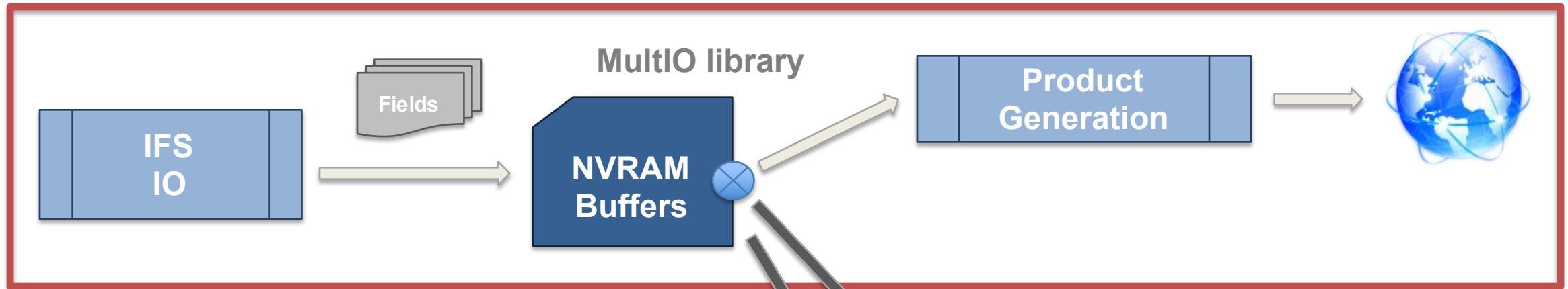
- Based on MeteoFrance IO server for IFS
- Entering production in March 2016



What if some node fails?
Issues with **resilience** in **time critical path**

Streaming Model Output to a Computing Service

Time critical path



MultIO implements *IO multiplexing*

Remove filesystem IO from **critical path**

Today, we could save:

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

Still needs **checkpoints** to filesystem for **resilience**

Summary

- Filesystem IO is a bottleneck to **time critical capability** computing
- Need to increase **capacity** with same **TCO** (run more experiments)
- **NextGenIO** is a co-design project with Intel & Fujitsu
 - Adapting 3D Xpoint technology to HPC, in particular to NWP (IFS)
- Stream Model Output to Product Generation
 - **Minimize** filesystem IO in the **critical path**
 - **NVRAM** can be used as application controlled *burst buffers*
 - Develop a **multiplexing IO library** to dispatch output fields to multiple destinations

Messages To Take Home

NVRAM will change the way we use and store data

***What would you do differently,
if your disk access 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