Approaches to I/O Scalability Challenges in the ECMWF Forecasting System

PASC'16, June 9 2016

Florian Rathgeber, Simon Smart,
Tiago Quintino, Baudouin Raoult, Stephan Siemen, Peter Bauer

Development Section, Forecast Department,
European Centre for Medium-Range Weather Forecasts (ECMWF)

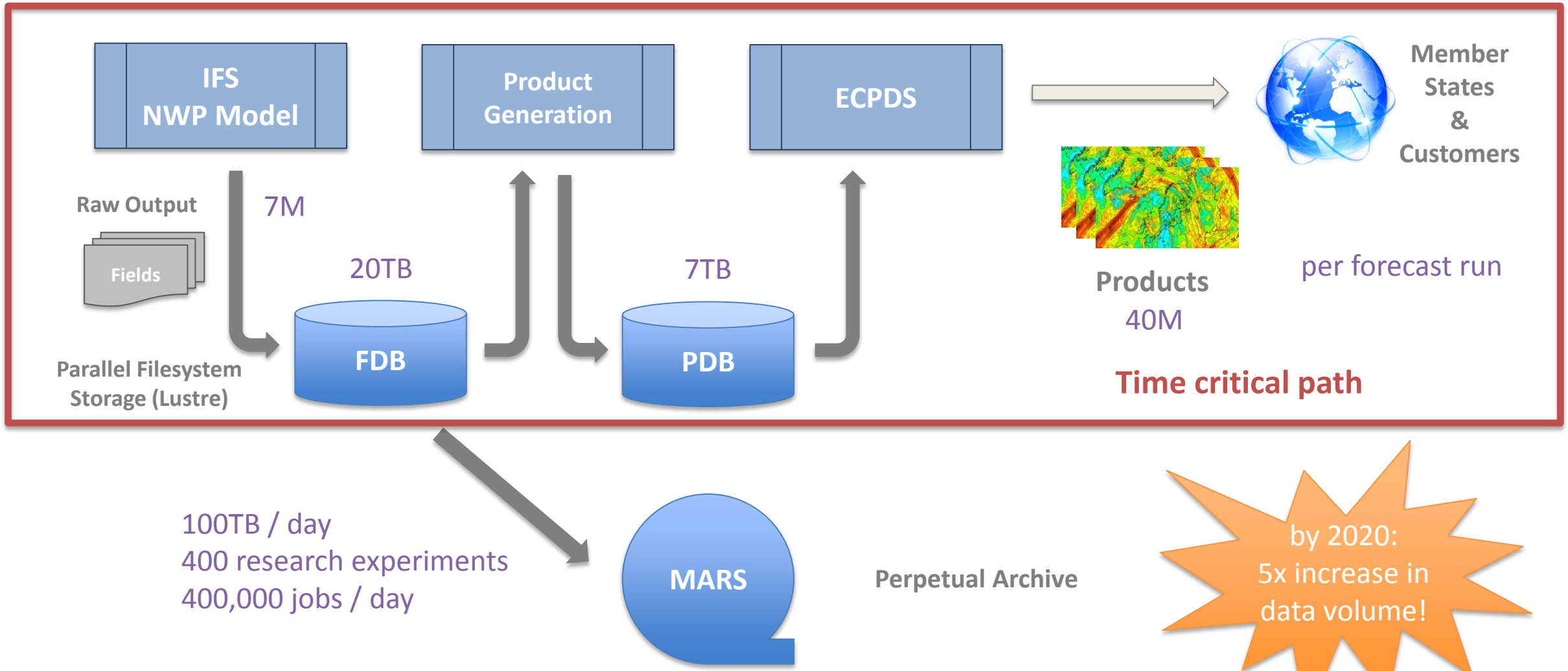
http://ecmwf.int | florian.rathgeber@ecmwf.int

ECMWF

nextgenio

ECMWF's Production Workflow

2016: 9km resolution, 137 levels



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

MARS

Perpetual Archive

by 2020:
5x increase in
data volume!

Hardware constraints

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

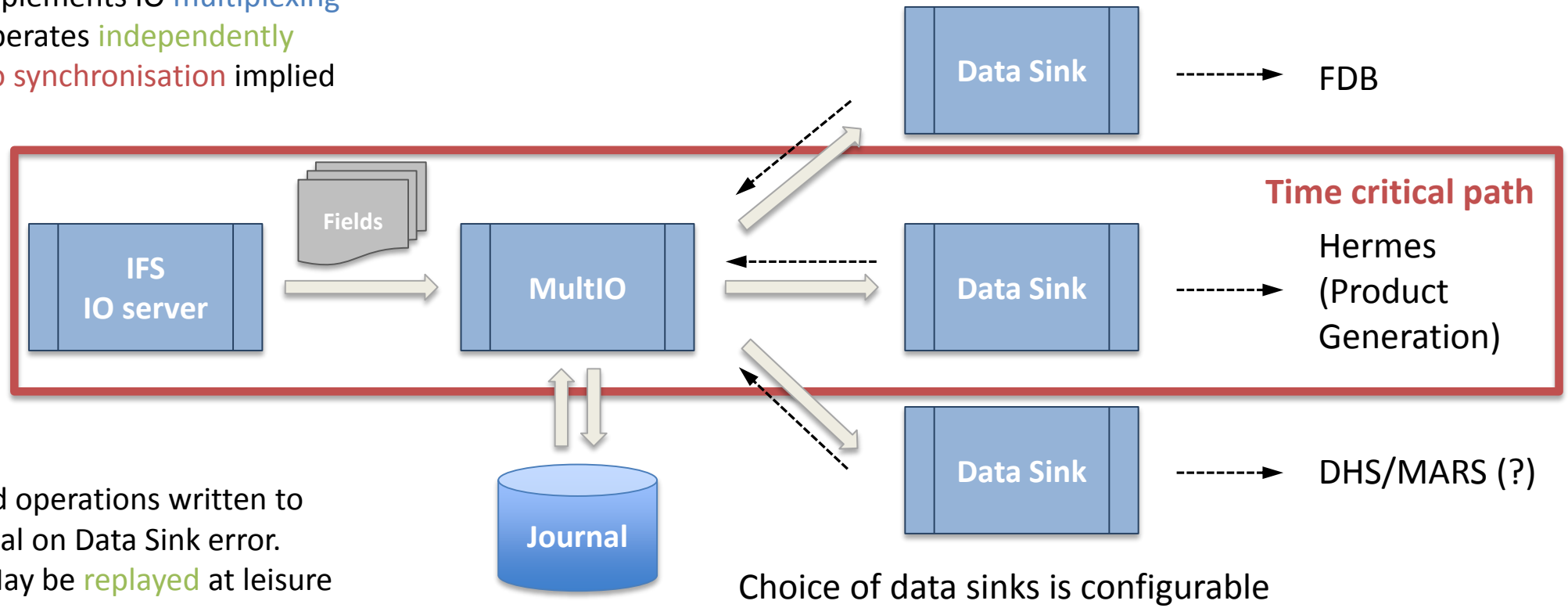
Can we **avoid** hitting disk so often?

Can we **stream** data directly from the Model to Dissemination?

Redirecting the Data Flows: MultIO Library Structure

Each IO server node has a MultIO system

- implements IO multiplexing
- Operates independently
- No synchronisation implied



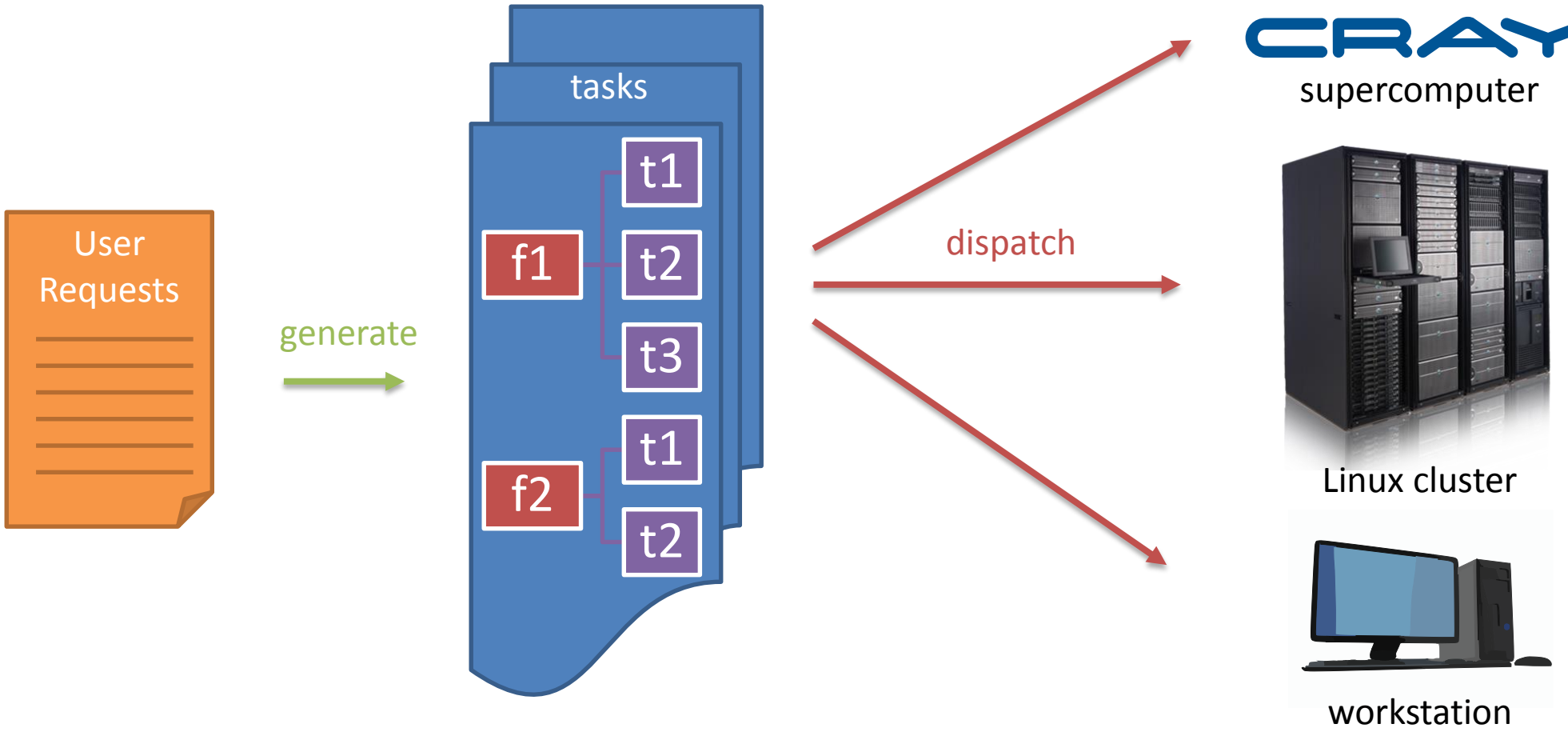
Failed operations written to journal on Data Sink error.

- May be replayed at leisure

Choice of data sinks is configurable

- May be asynchronous
- Errors caught and reported to MultIO

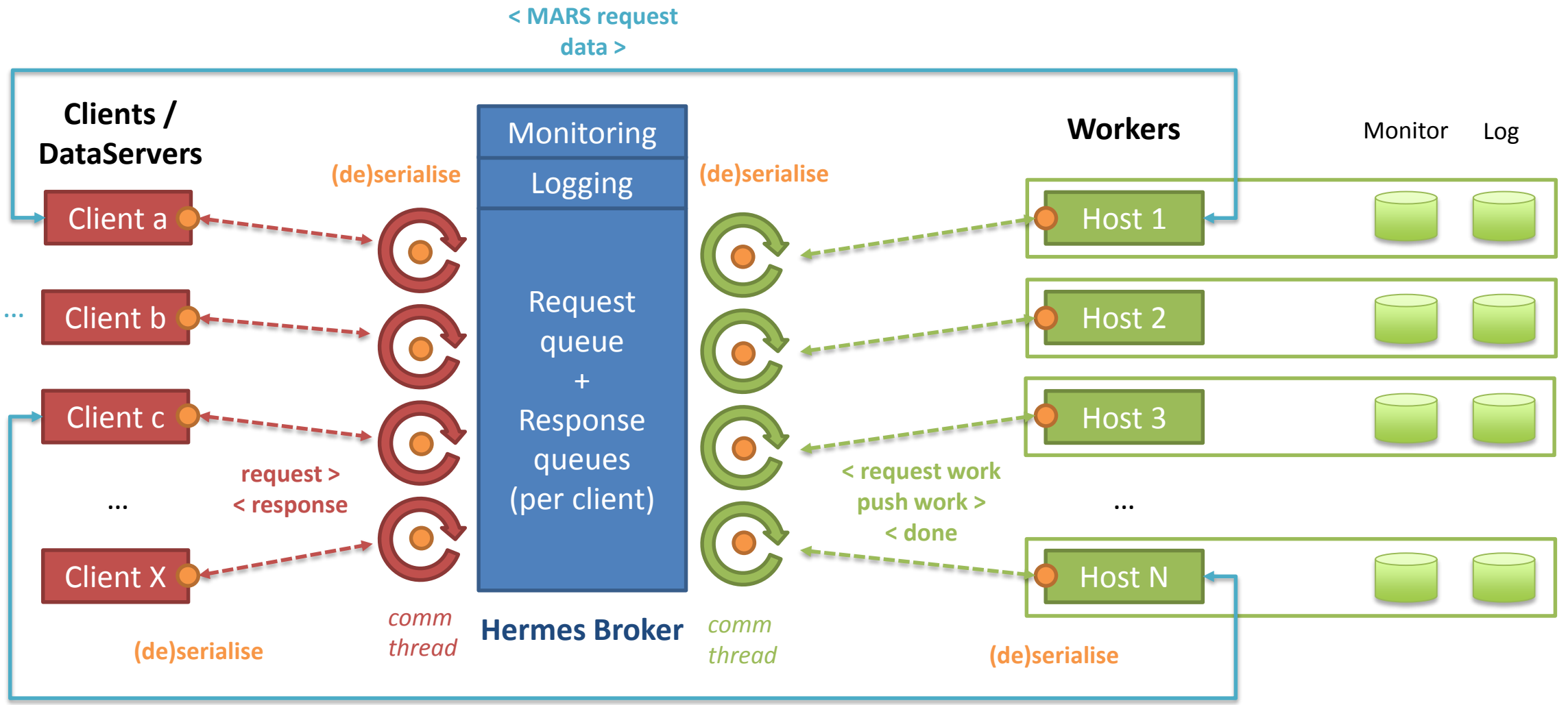
Hermes Computation Service for ECWMF Product Generation



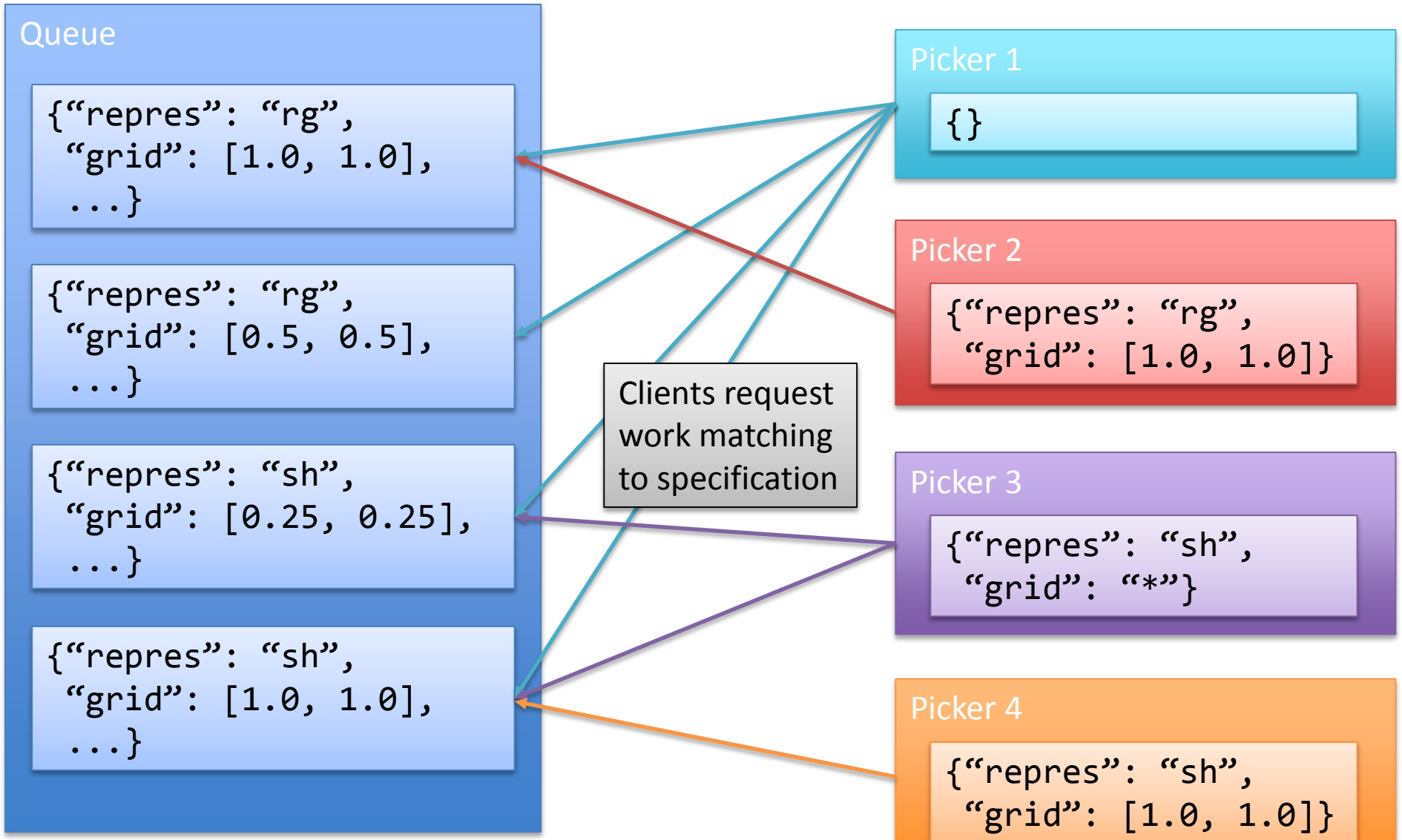
Frontend (definition)

Backend (execution)

Hermes Computation Service: Broker Architecture



Hermes Computation Service: Work Selection



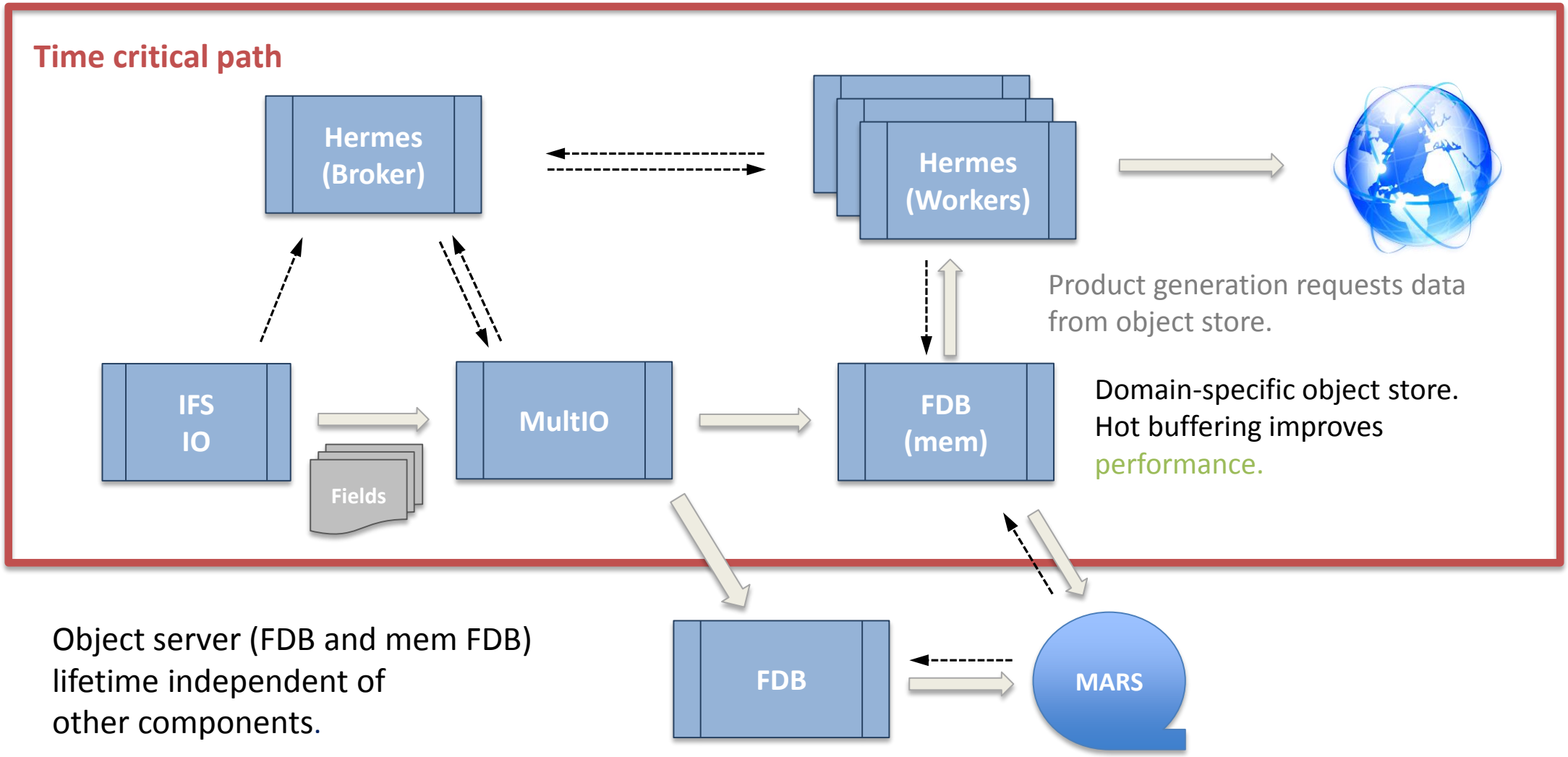
Generic worker

Optimise caching of interpolation matrix

Optimise for any SH transforms

Optimise loading of SH coefficients

Possible Infrastructure for ECMWF Production Workflow



Object server (FDB and mem FDB) lifetime independent of other components.

Summary

- Filesystem IO is a bottleneck to **time critical** computing
- Stream Model output to Product Generation
  - **Minimize** filesystem IO in the **critical path**
  - Upcoming **NVRAM** can be used to temporarily store model output
- Flexible hardware and architecture requirements for product generation

Thank you! Questions?

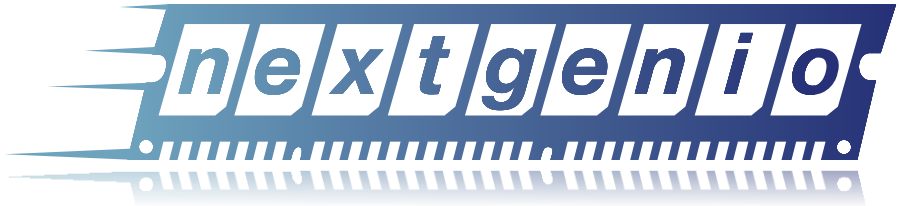
ECMWF: http://ecmwf.int

NextGenIO: http://nextgenio.eu

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

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