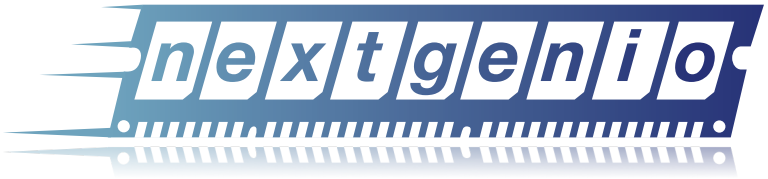


ECMWF's Extreme Data Challenges towards a Exascale Weather Forecasting System

Tiago Quintino, Simon Smart, James Hawkes, Baudouin Raoult

ECMWF

tiago.quintino@ecmwf.int

PASC 2019, Zurich

12-14th June 2019

ECMWF

ECMWF's Forecasting Systems

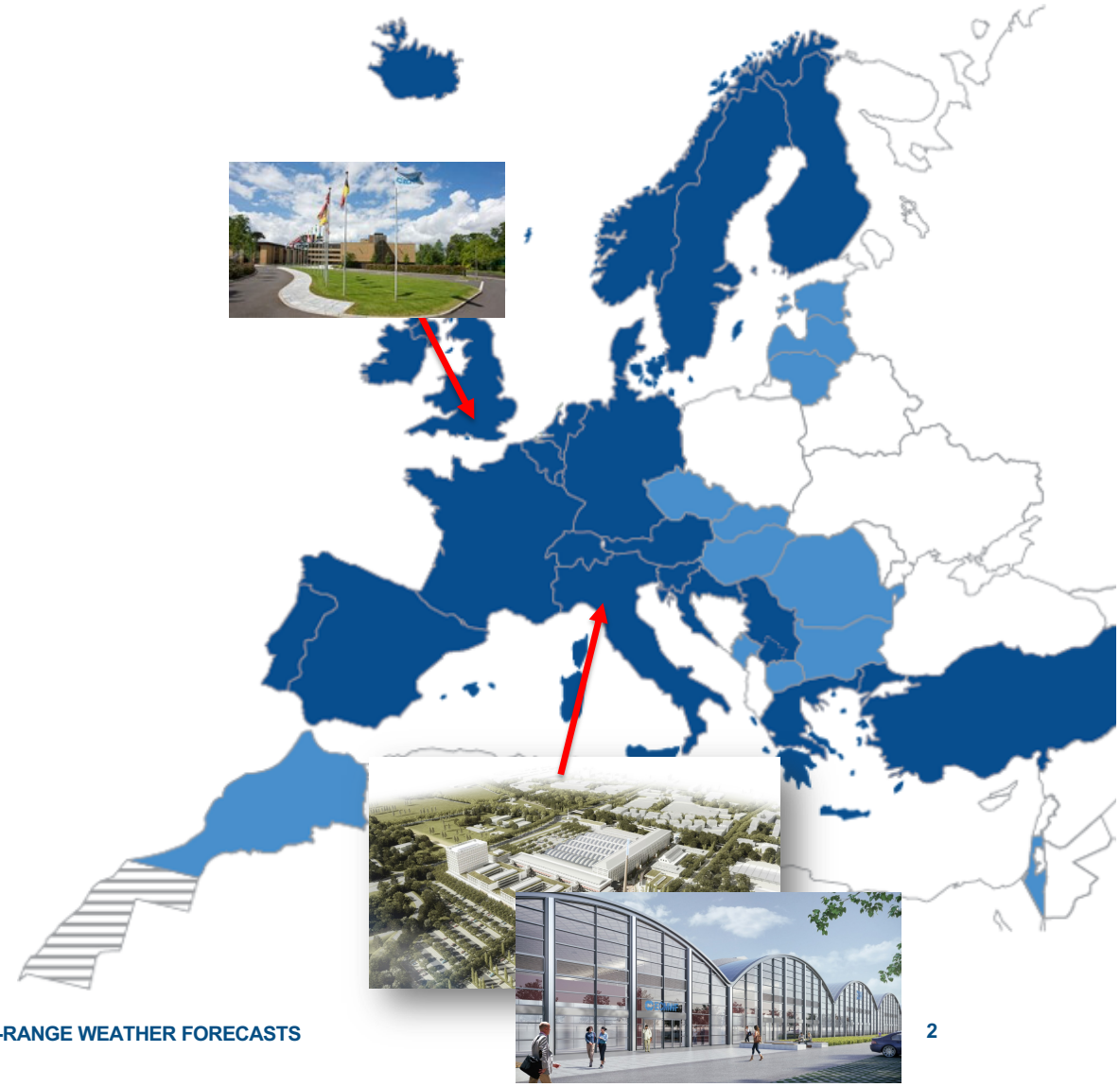
What do we do?

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**

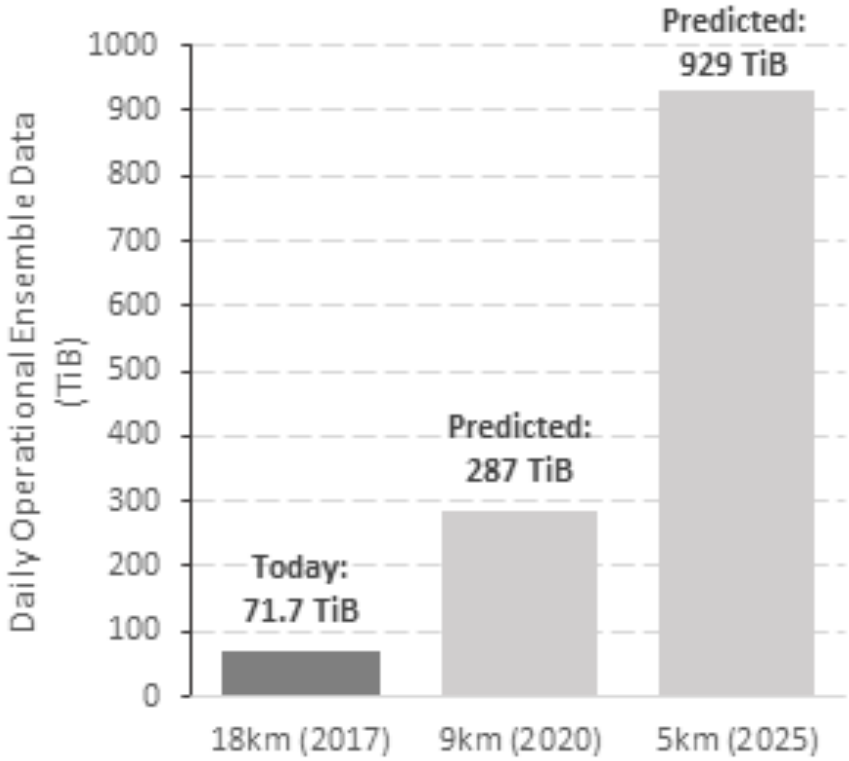
- Experiments to improving our models
- Reforecasts, Climate reanalysis, etc



Challenges

Data Growth – History and Projections

Historical Growth of Generated Products



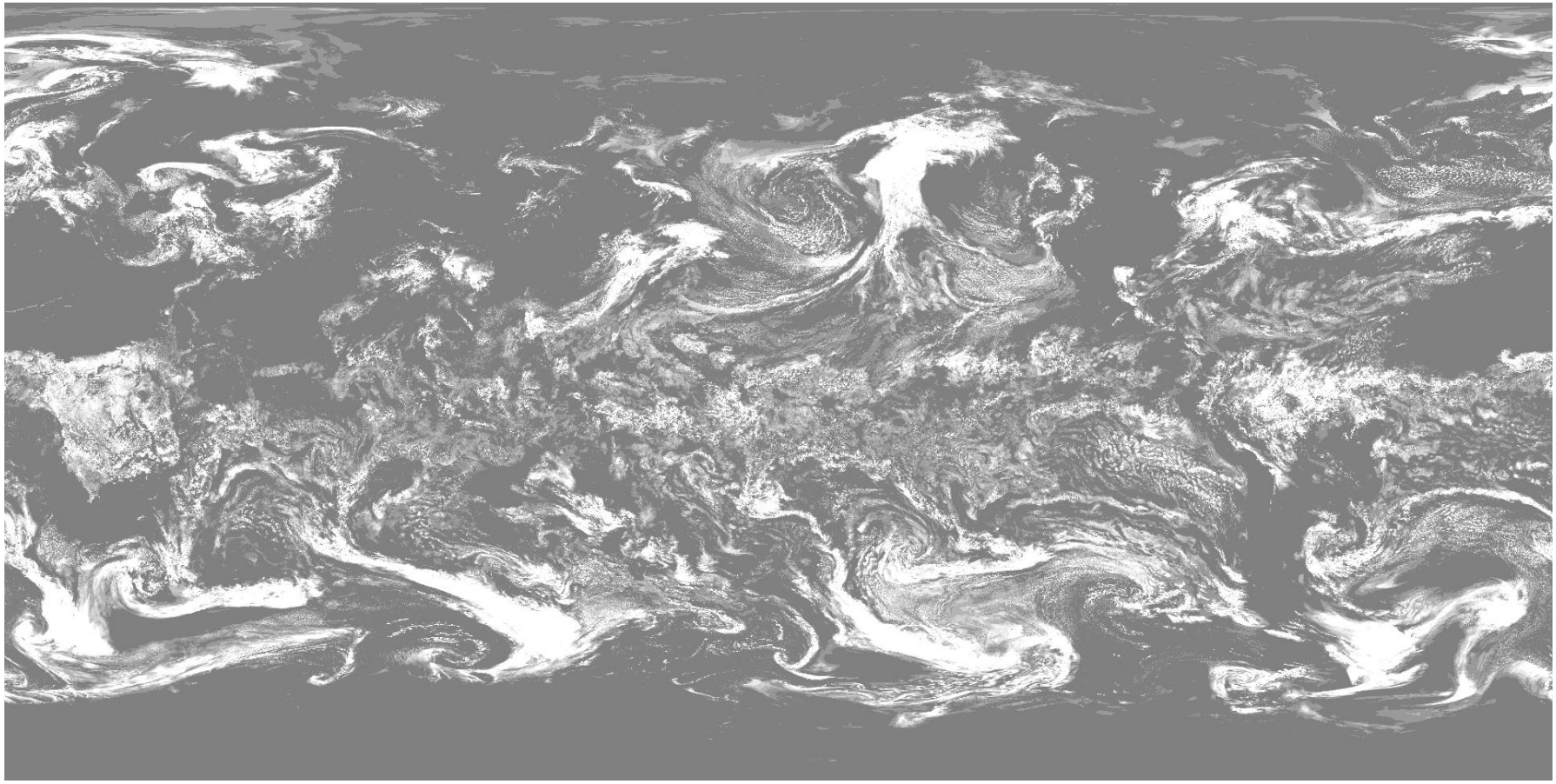
Model Output Projected Growth

History and Future of Resolution Upgrades

Resolution	Grid size	Grid Points	Field Size (in memory)	Vertical Levels	YEAR
T319	62.5 km	204 k	1.6 MB	L31	1998
T511	39 km	524 k	4 MB	L60	2000
T799	25 km	1.2 M	9.6 MB	L91	2006
T1279	16 km	2.1 M	16.8 MB	L91	2010
Tco1279	**9 km**	**6.6 M**	**50.4 MB**	**L137**	**2016**
Tco1999	5 km	16.1 M	122.6 MB	**L160**	***2025***
Tco3999	2.5 km	64 M	490 MB		
Tco7999	*1.25 km*	*256 M*	***1909 MB***	***L180***	***2030***

ECMWF

TCo7999 (~1.25km) 256 Megapixel



(12 h forecast, *hydrostatic*, no deep convection parametrization, 120s time-step, 960 Broadwell nodes, ~10s per timestep)

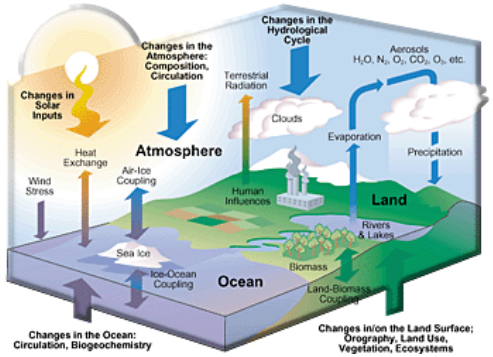
Multiple dimensions

➔Reliability

Ensembles

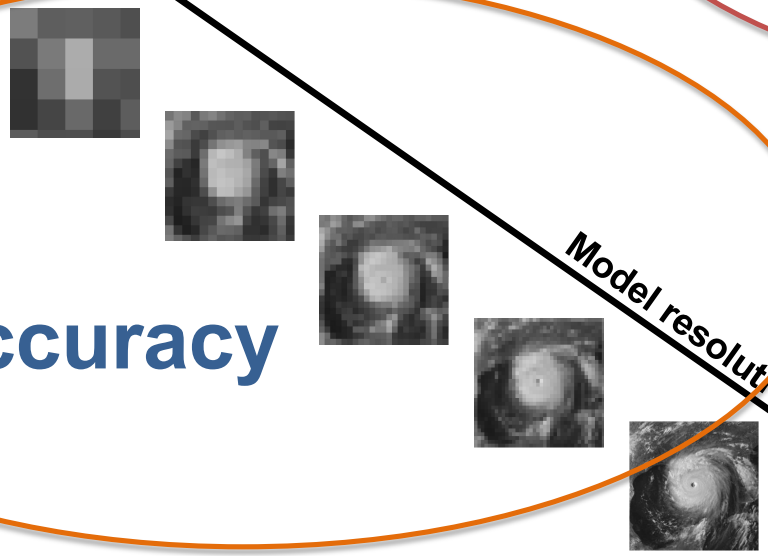
Traditional weather science domain

Model complexity



➔Range

Traditional climate science domain

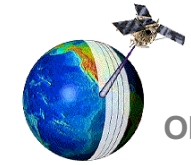


➔Accuracy

Model resolution

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

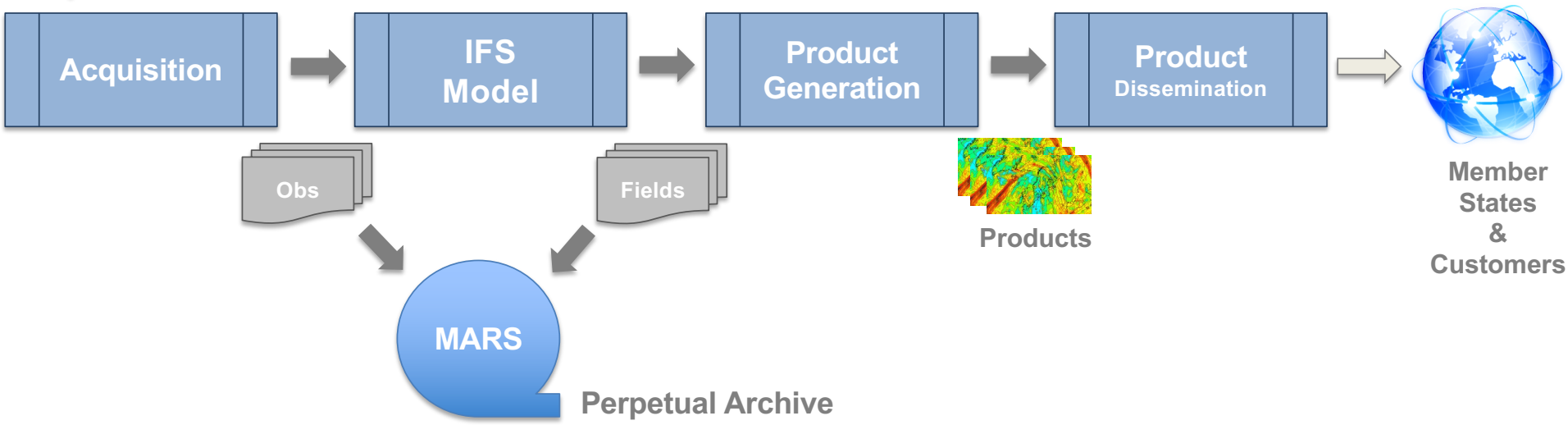
ECMWF's (Simplified) Operational Workflow



Global Observations

Data Workflow

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

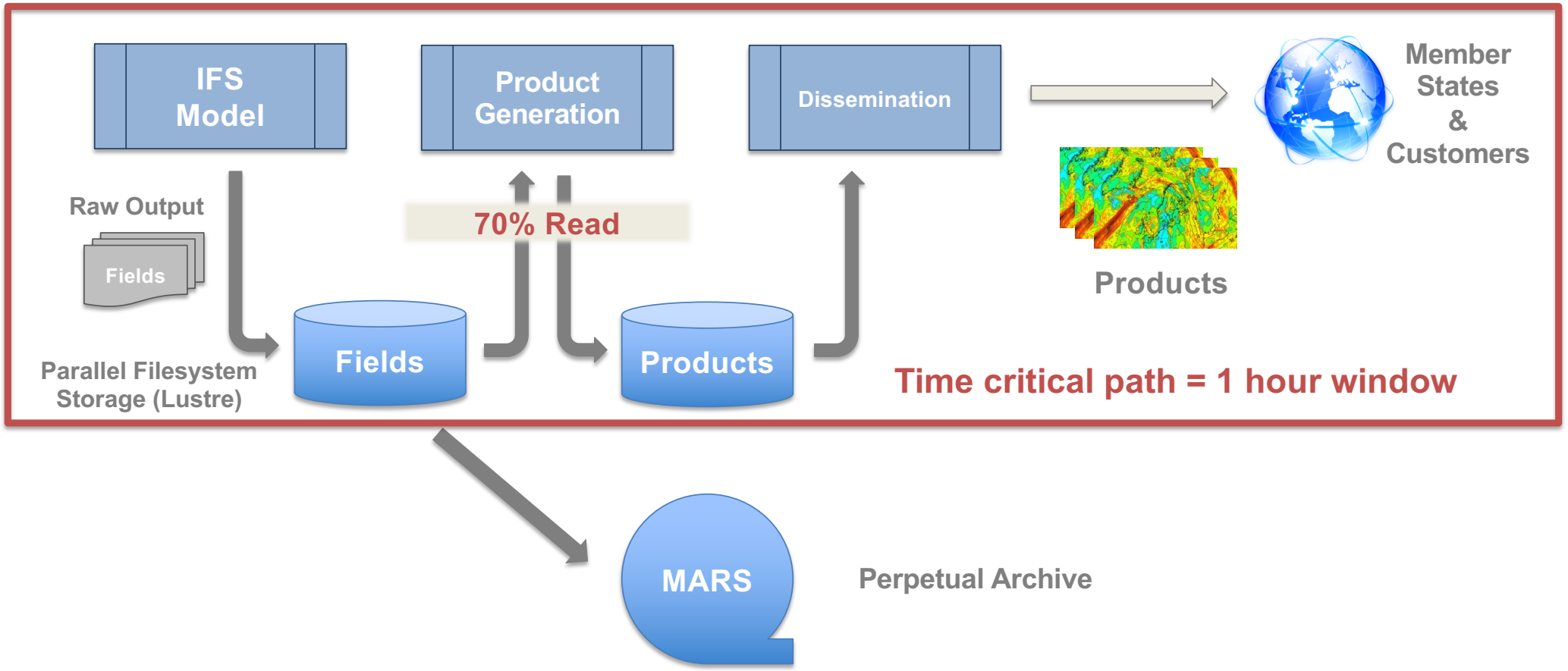


Effects of Product Generation

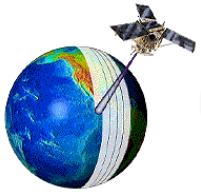
	IFS Model	Model + I/O	Model + I/O + PGen
Nodes	2440	2776	2926
Run time [s]	5765	6749	7260
Relative	-	**+ 17%**	**+ 26%**

9Km 50 member ensemble
Broadwell nodes 2x18 cores
Cray XC40 Aries interconnect
Lustre FS IOR 90GiB/s

ECMWF's Production Workflow



Storage View of Workflow



Observations

Acquisition

Product Dissemination

Member States & Customers

Acquire

Disseminate

Produce

IFS Model

Parallel FS

Archive

Modify

Product Generation

MARS

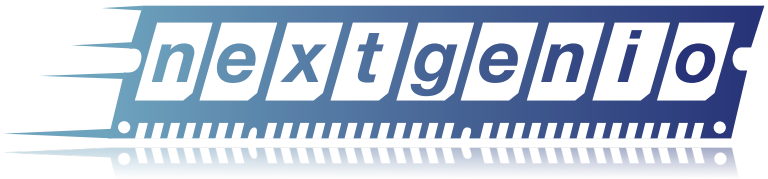
Perpetual Archive

Data is Central!

What have we done so far?

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

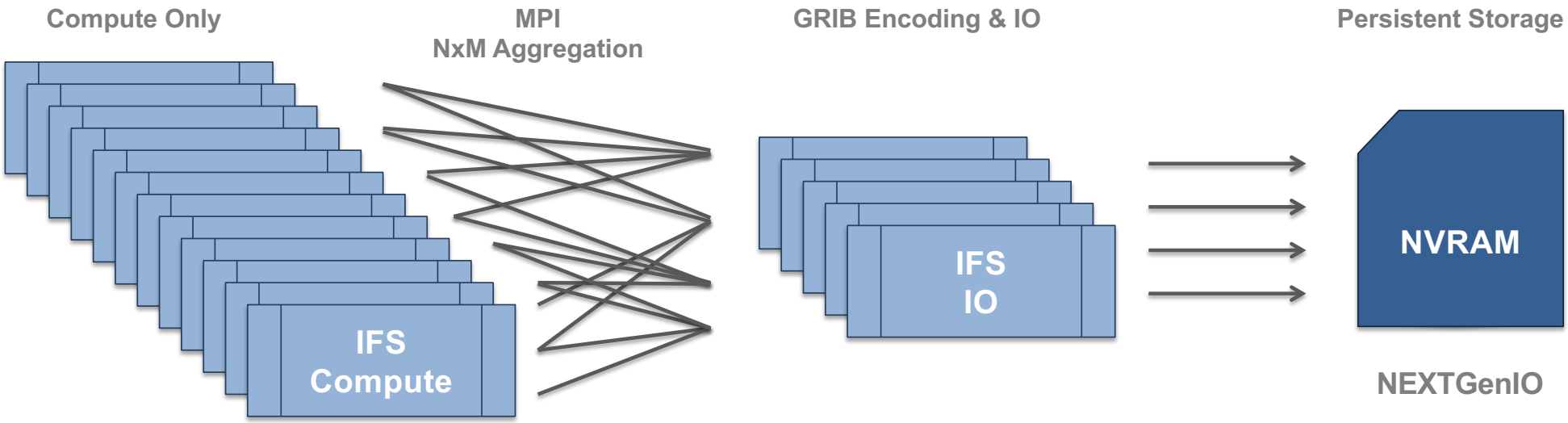
ECMWF Tasks

- 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

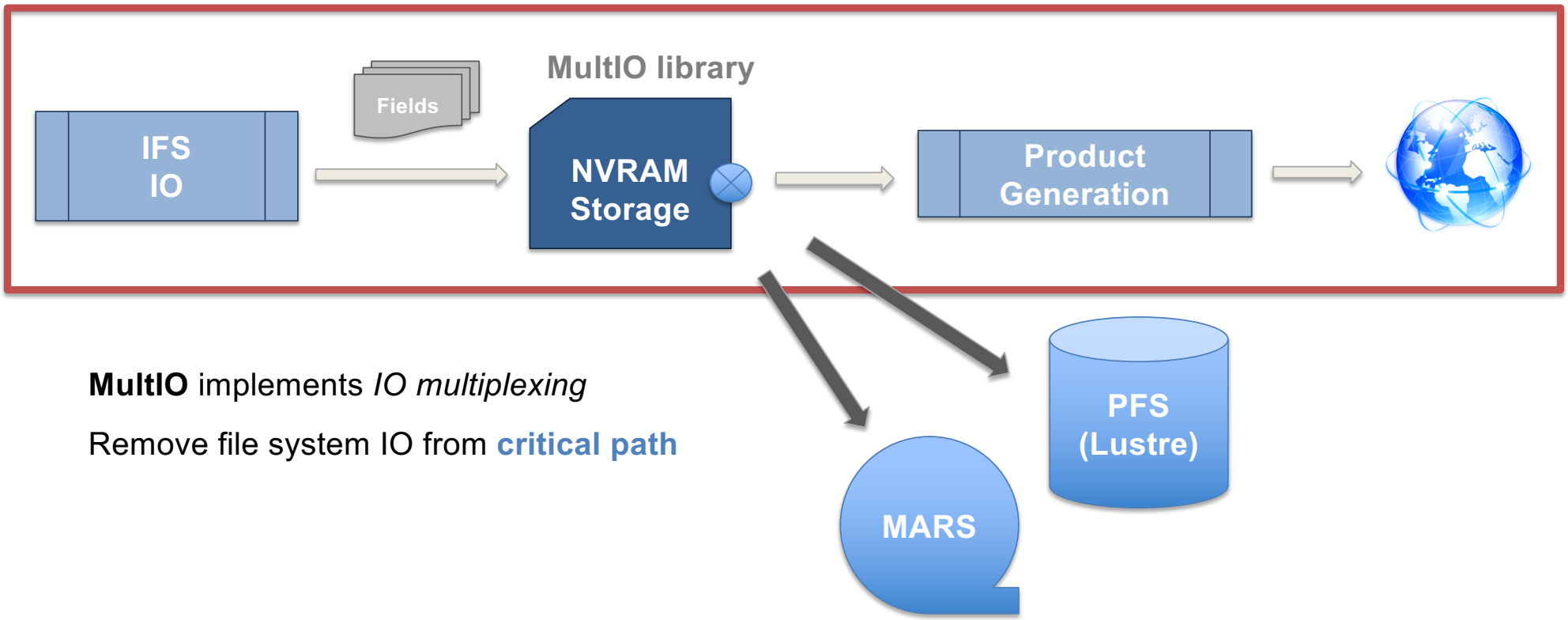
http://www.nextgenio.eu - EU funded H2020 project, runs 2015-2018

IFS IO Server

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



Streaming Model Output to Product Generation



MultIO implements *IO multiplexing*

Remove file system IO from **critical path**

How to store all model output in NVRAM?

Paper Presentation:

A High-Performance Distributed Object-Store for Exascale Numerical Weather Prediction and Climate

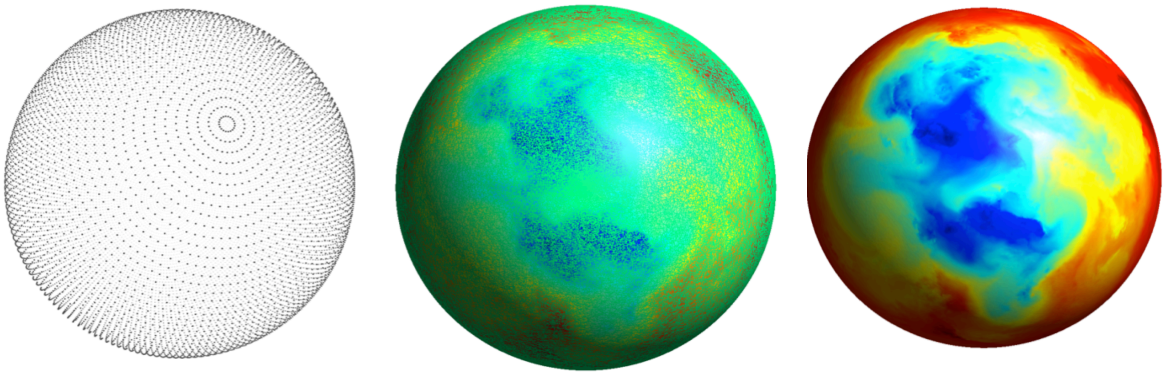
Simon Smart, T. Quintino, B. Raoult, PASC'2019

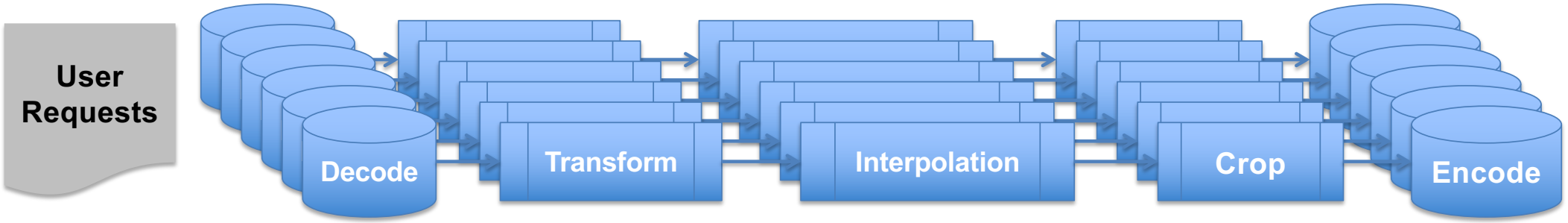
FDB5

- ***Full rewrite***
- ***All workflows use this storage software***
- ***3 years preparing for this…***
- ***Roll out to operations 3 days ago (2019.06.11)***
  - ***As we were flying into Zurich*** ✈️

Product Generation – PGen / MIR

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





Upgrading the Interpolation Package

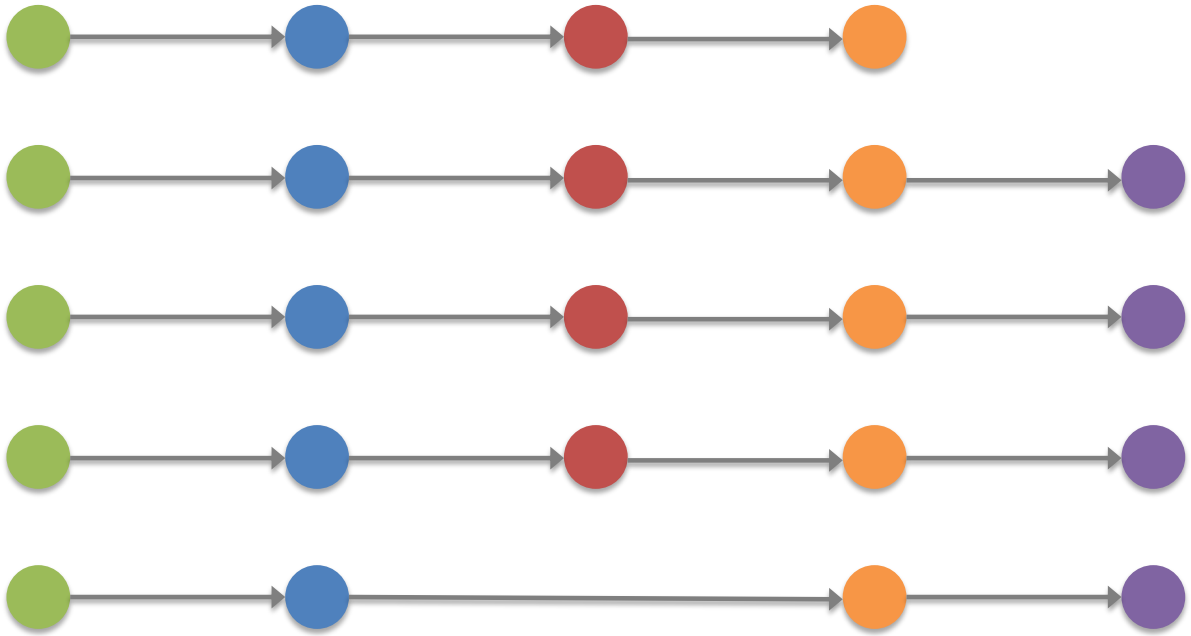
Interpolation is **pervasive**:

- Product genetation
- Access to data archive (MARS)
- Visualisation of products
- Web services

Used by many operational systems at ECMWF

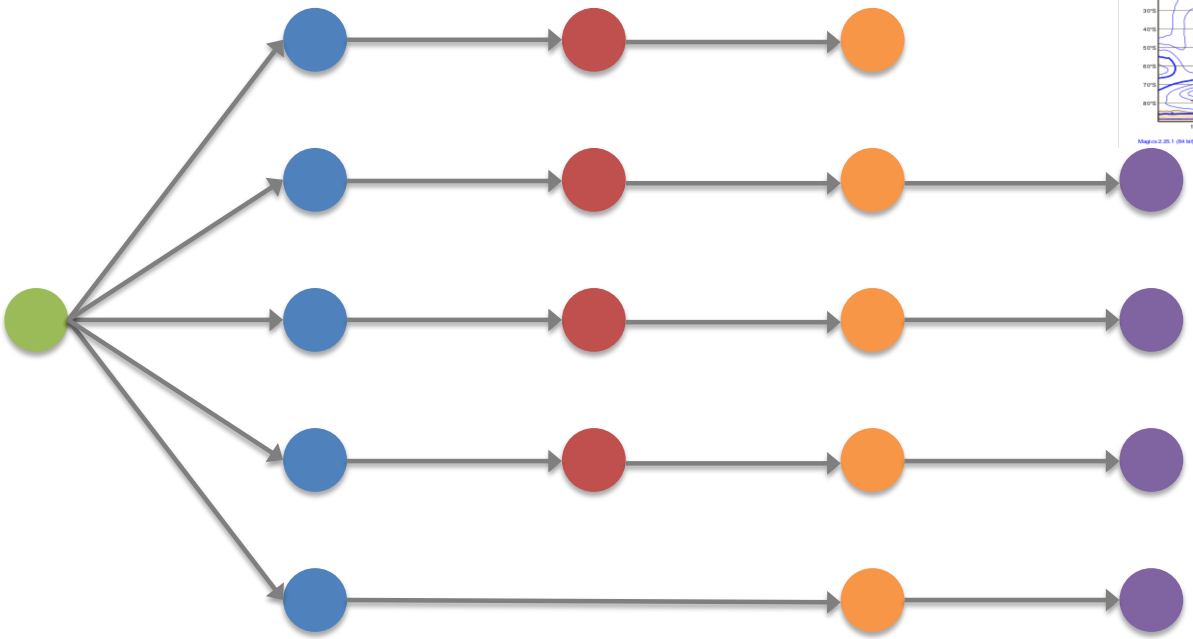
ECMWF

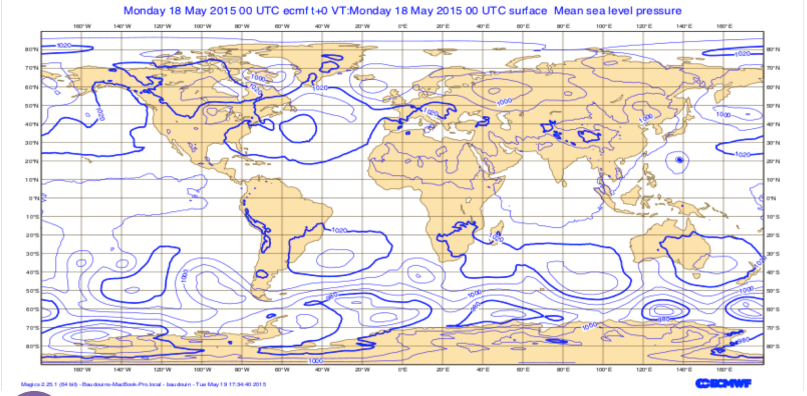
PGen – Task Graph Analysis



PGen – Task Graph Analysis

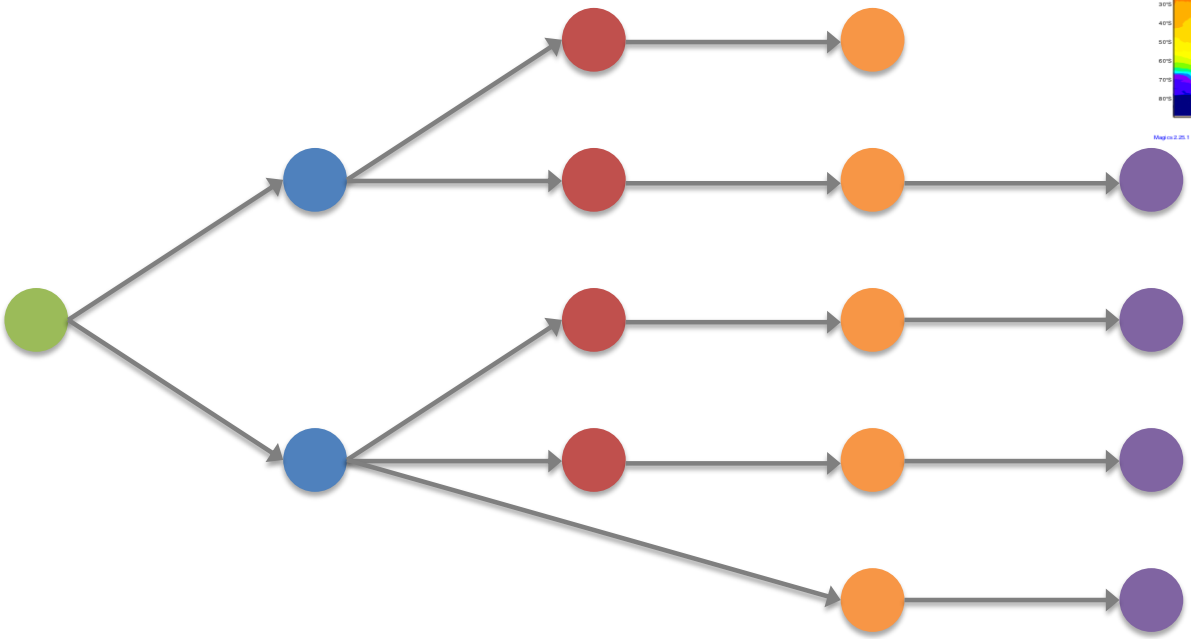
Merge same input

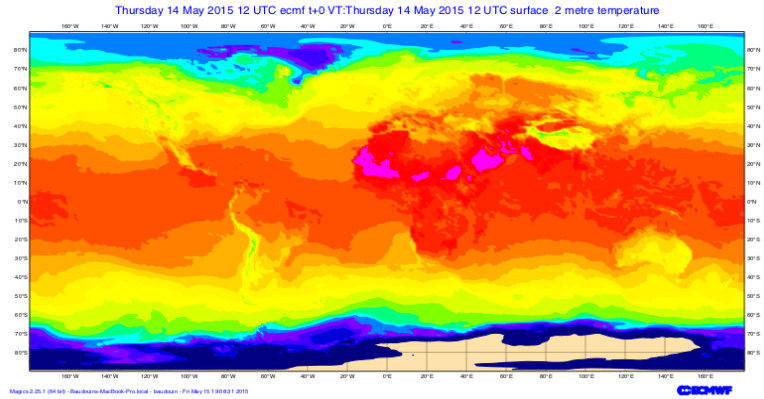




PGen – Task Graph Analysis

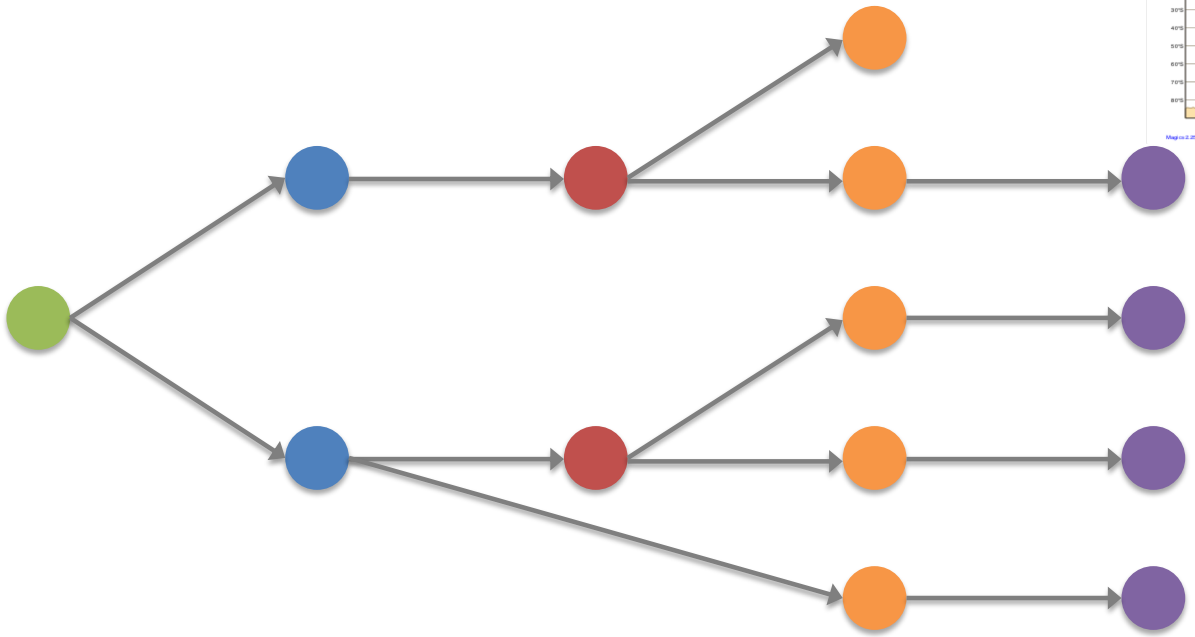
Merge same interpolation target grids

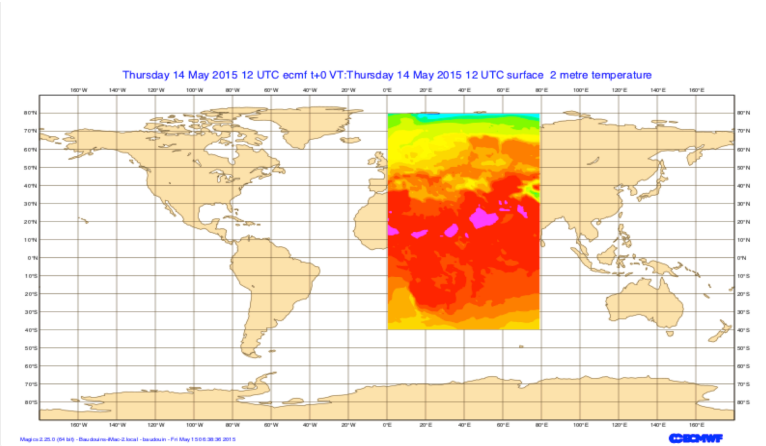




PGen – Task Graph Analysis

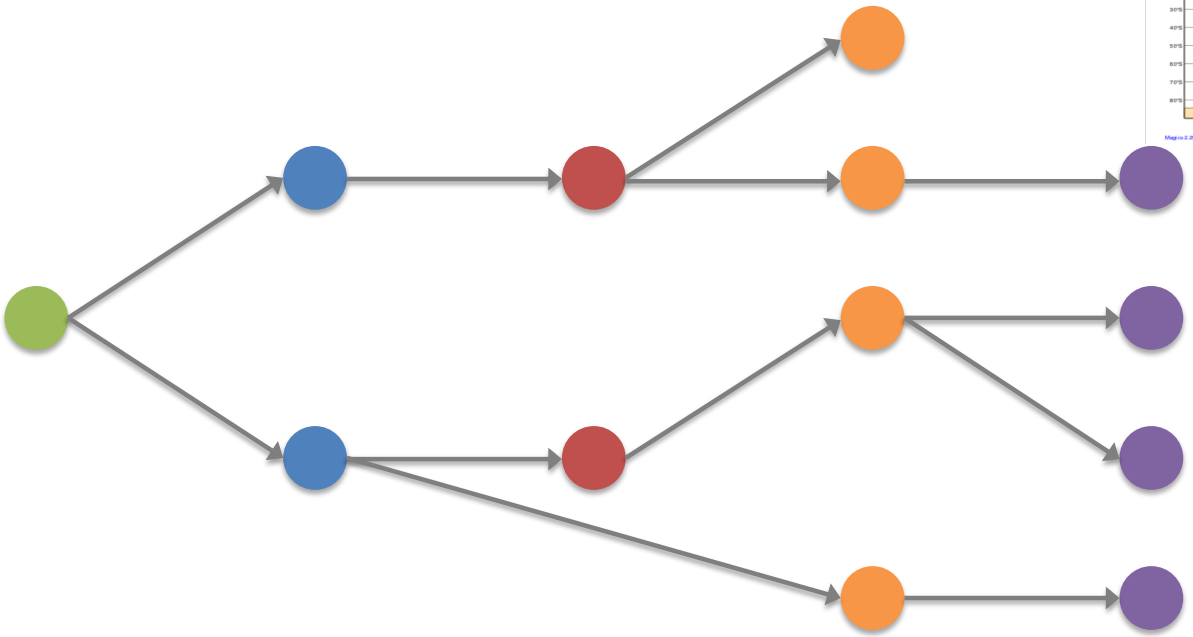
Merge same local area cropping

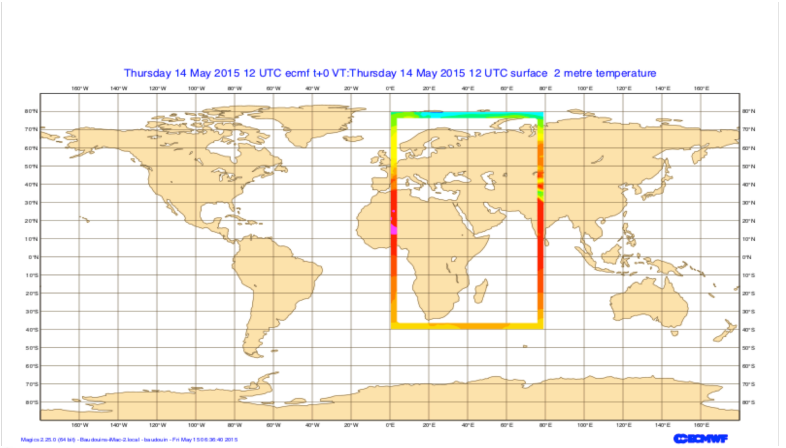




PGen – Task Graph Analysis

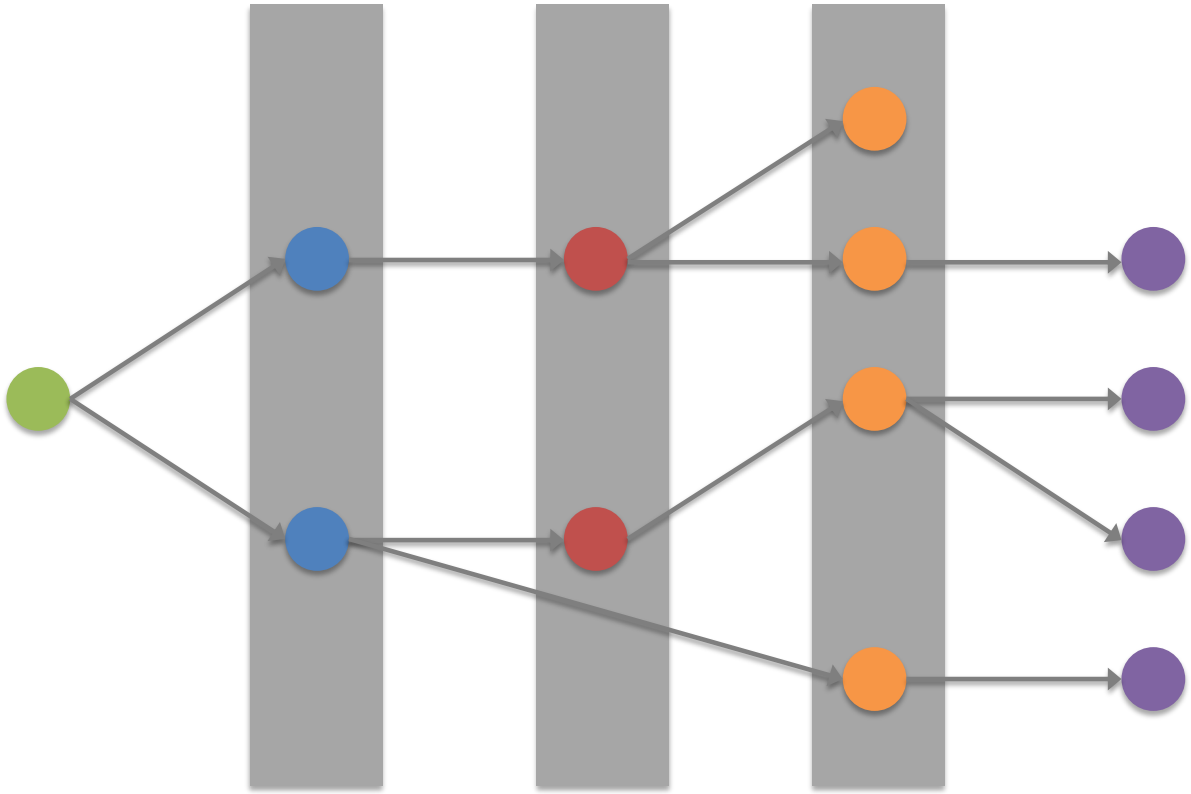
Merge same local area frames



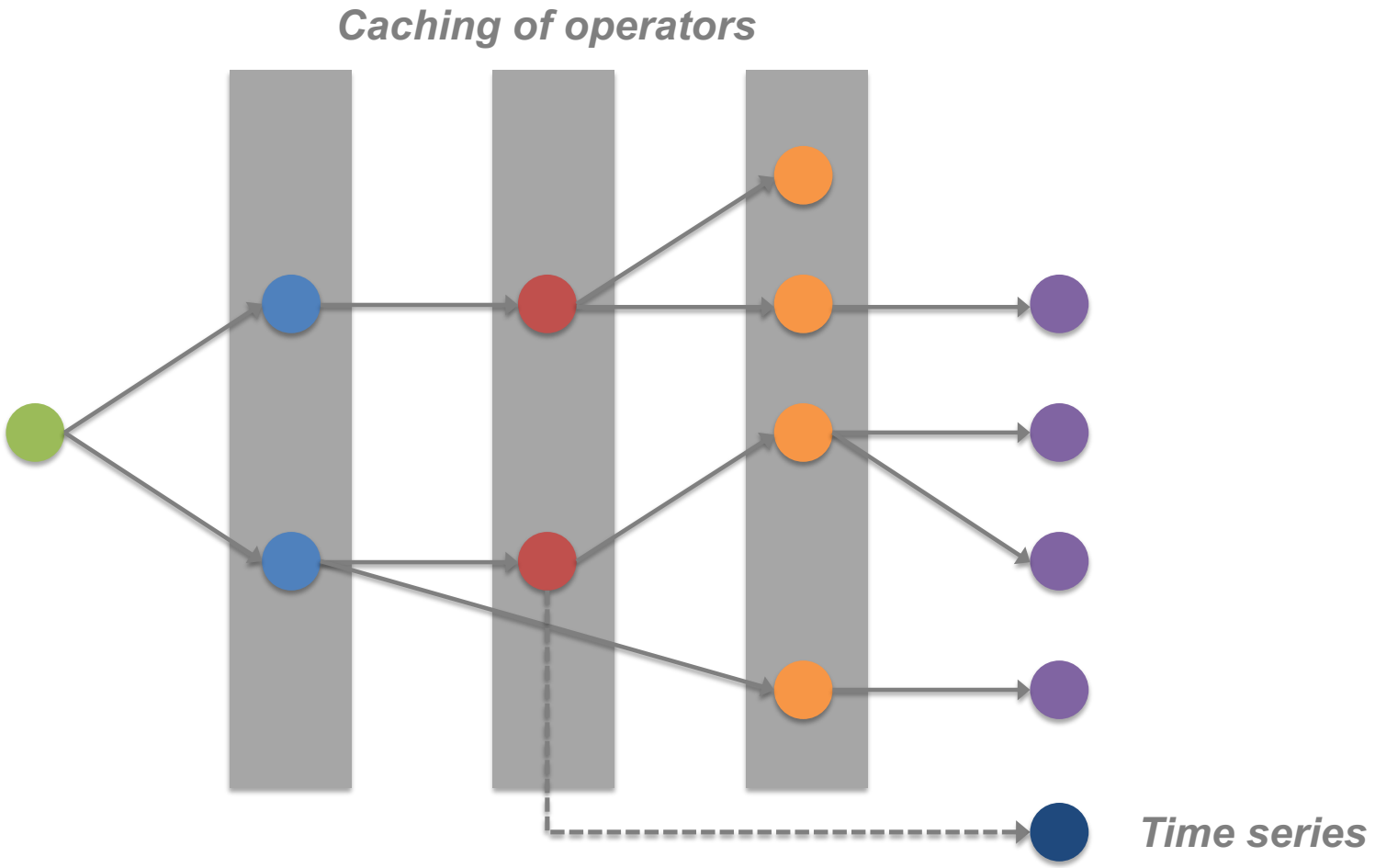


PGen – Task Graph Analysis

Caching of operators

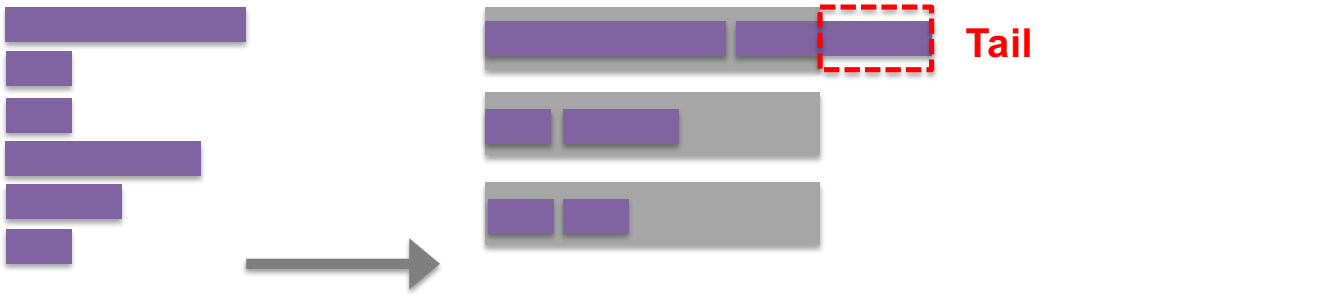


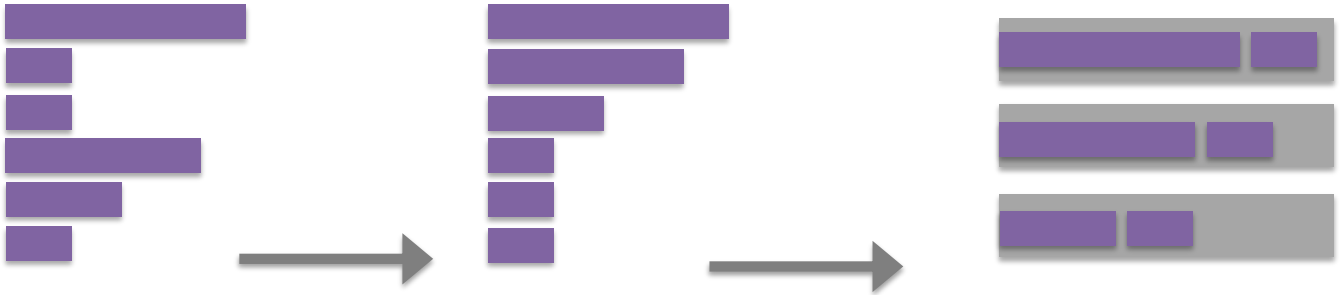
PGen – Task Graph Analysis



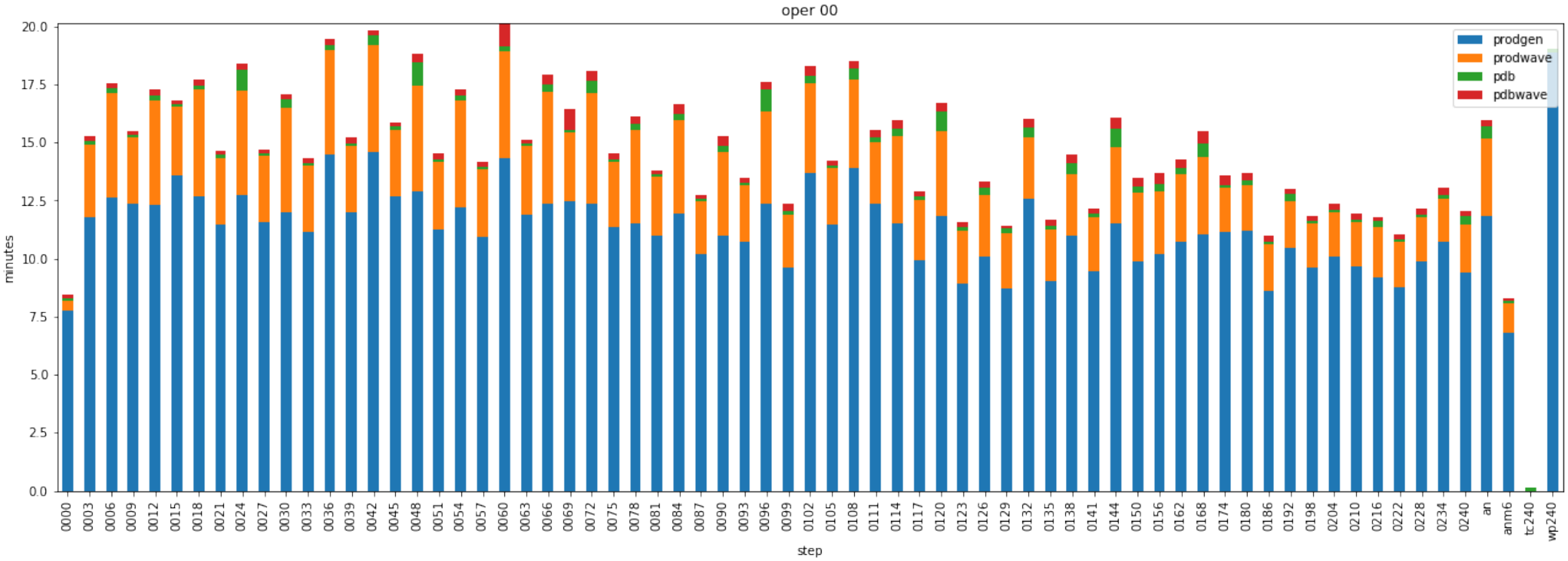
PGen – Task Reordering for Dynamic Load Balancing

Example: 6 tasks, 3 workers



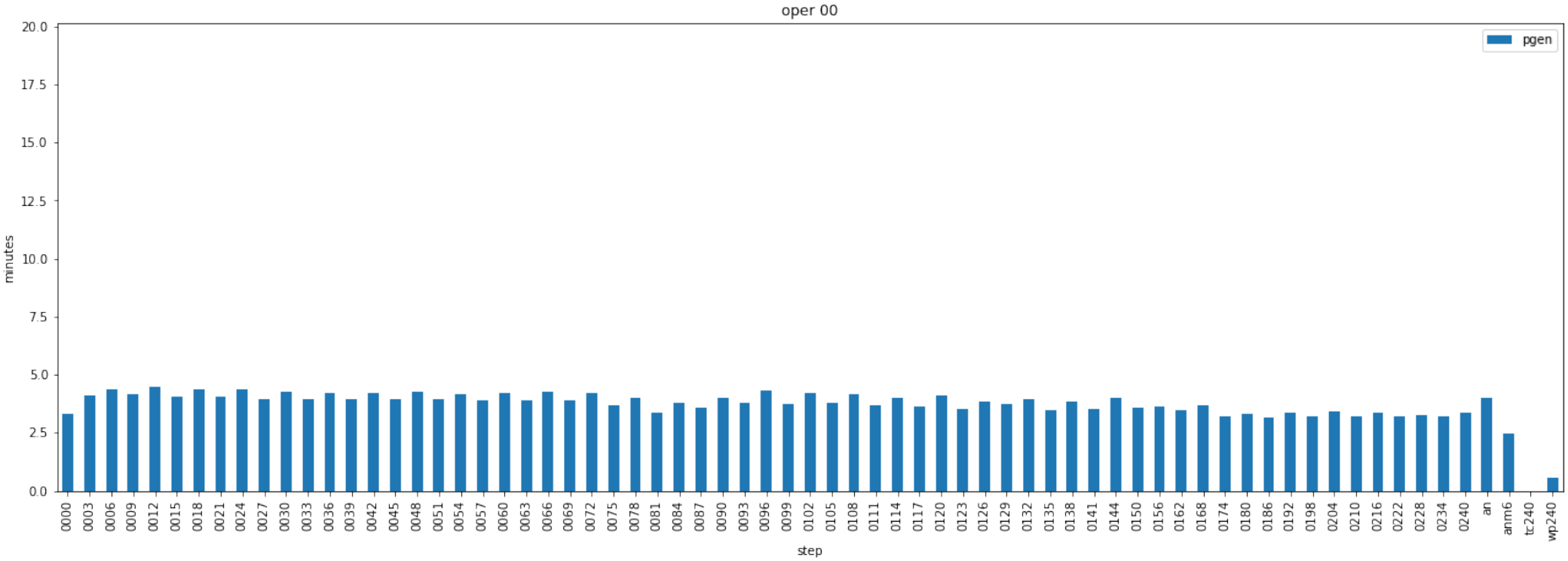


Performance Analysis – Oper Stream 00Z run



e.g. step 24 ~ 18 min

Performance Analysis – Oper Stream 00Z run



e.g. step 24 ~ 4.3 min = 412% faster

New software in workflow

- MIR - New interpolation software. Operational Mar 2018
- PGEN – New product generation software. Operational Mar-June 2018
- FDB5 – New object storage software. Operational 11 June 2018

- Previous Operational Product Generation
  - 20-25min per forecast step
- PGEN + MIR into
  - 6 min per step
  - 4 min interpolation + 2 min reading data
- FDB5 in Ops (3 days ago)
  - 6 min per step
  - 4 min interpolation + 1 min reading data (x2 improve in performace)

All SAME Hardware!

Looking ahead

Impacts of NVRAM on Data Access

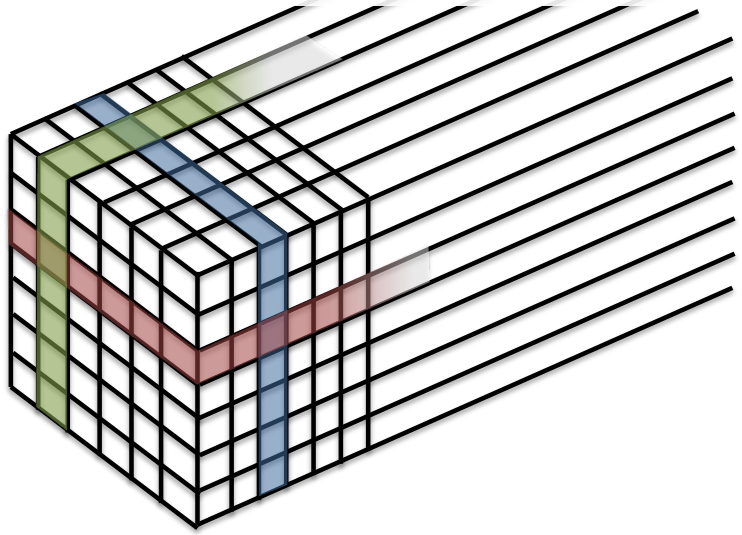
Byte Addressable Hypercubes

- Longitude (3600)
- Latitude (1800)
- Atmospheric levels, Physical parameters (~200)
- Time steps (~100)
- Probabilistic pertubations (50)

@ double precision

- 9km **48 TiB**
- 5km **192 TiB**
- 1.25km **1.82 PiB**

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



Not included: *historical observations, multiple models, etc...*

Novel Data Flows – Data Centric Computing

Data-as-a-Service

Cloud

Producer

Consumer

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

NVRAM (?)
Data Hypercubes

Archive

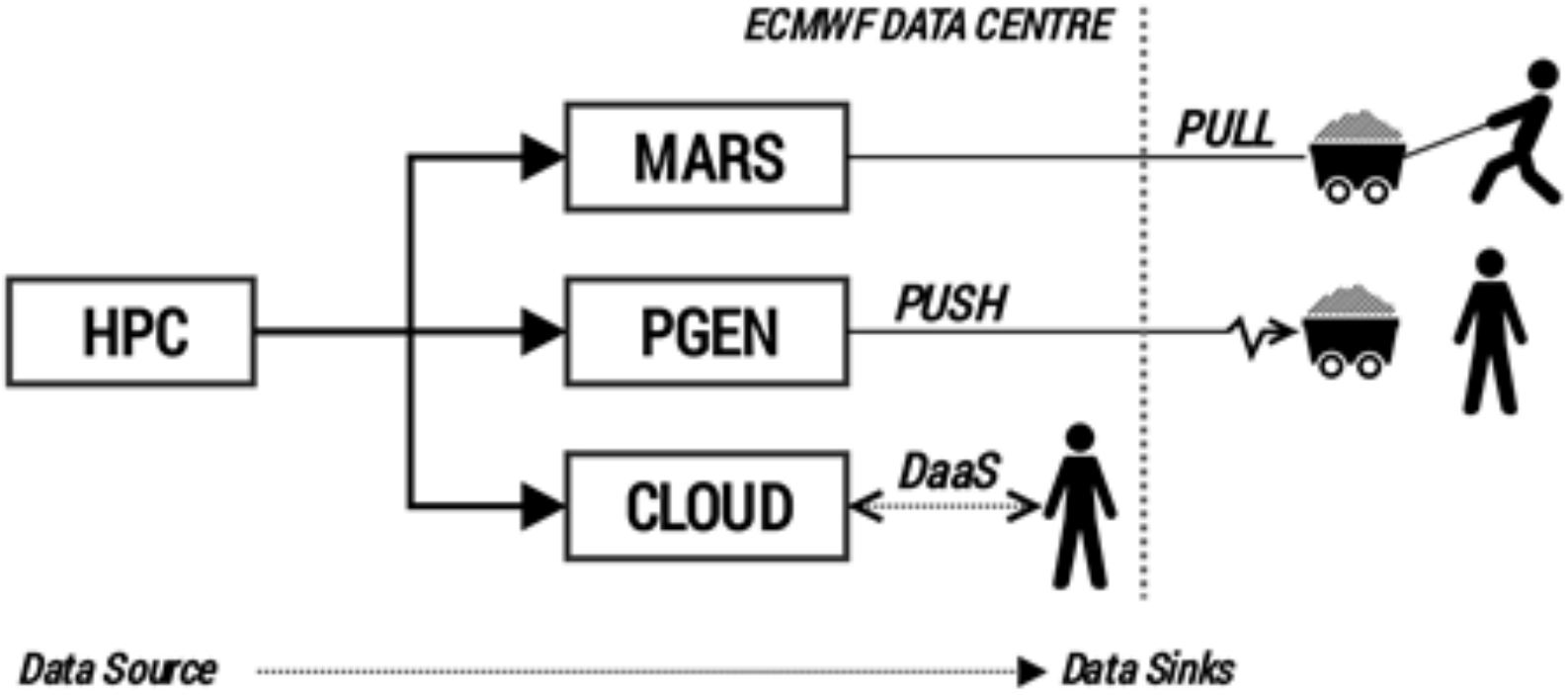
PFS

HDD

Tape

MARS

Novel Data Flows – Multiple Pathways to Serve Data

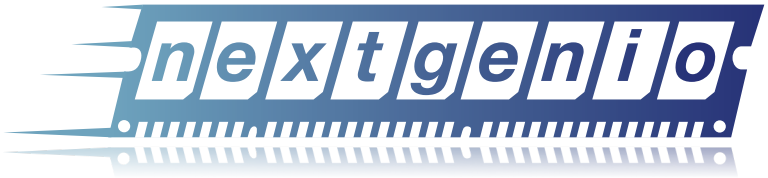


Messages To Take Home

Ensemble data sets are growing quadratically to cubically in size,
Brings an I/O crisis for time critical applications

*New technologies in the **horizon***
***but** will change the way we use and store data*

ECMWF is adapting its workflows to take advantage of these upcoming technologies



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

Thanks for your attention

Questions?