Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Improving the I/O scalability for the next generation of Earth system models: OpenIFS-Cassandra integration as a case study

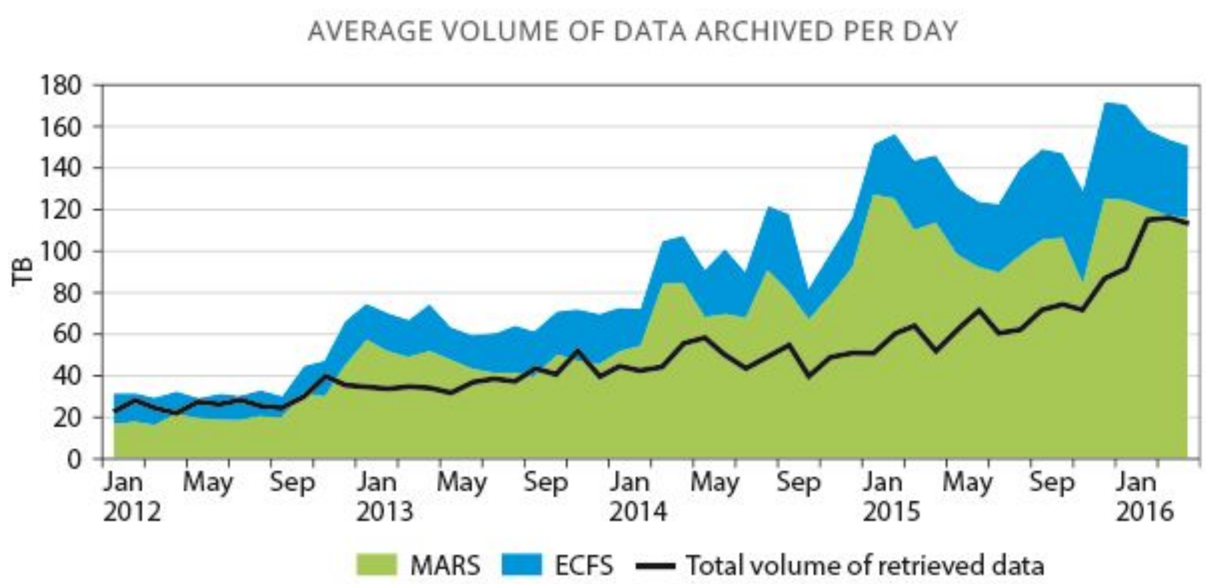
Authors: Xavier Yepes-Arbós, Pol Santamaria
Co-authors: Kim Serradell, Yolanda Becerra
Contributors: Jesus Labarta, Harald Servat

26/09/2019

NEXTGenIO Workshop on applications of NVRAM storage to exascale I/O, ECMWF, UK

Introduction

- Earth system models have benefited of the exponential growth of supercomputing power
- This allows to use more complex computational models to find more accurate solutions
- As a consequence, the generated amount of data has grown considerably



Source: Annual Report 2015, ECMWF

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Introduction

- However, since the I/O was not significant enough in the past, not much attention was paid to improve it

- Due to this reason, some Earth system models output data using inefficient sequential I/O schemes

- This type of scheme requires a serial process:
  - Gather all data in the master process of the model
  - Then, the master process sequentially writes all data

- This is not scalable for higher grid resolutions, and even less, for future exascale machines

- This is the case of the I/O scheme of the OpenIFS model

OpenIFS model

- OpenIFS is a free and simplified version of the Integrated Forecasting System (IFS), available under a license from the European Centre for Medium-Range Weather Forecasts (ECMWF)
- IFS is a global data assimilation and forecasting system which includes the modelling of the atmospheric composition
- Originally developed for weather forecasting, IFS writes using the GRIB format

GRIB format

- General Regularly-distributed Information in Binary form (GRIB) is a concise data format commonly used for Numerical Weather Prediction (NWP) output

- It is designed to be:
  - Self-describing
  - Compact
  - Portable across computer architectures

- In addition, it offers high performance for I/O operations to meet critical time-to-solution requirements

Post-processing in OpenIFS

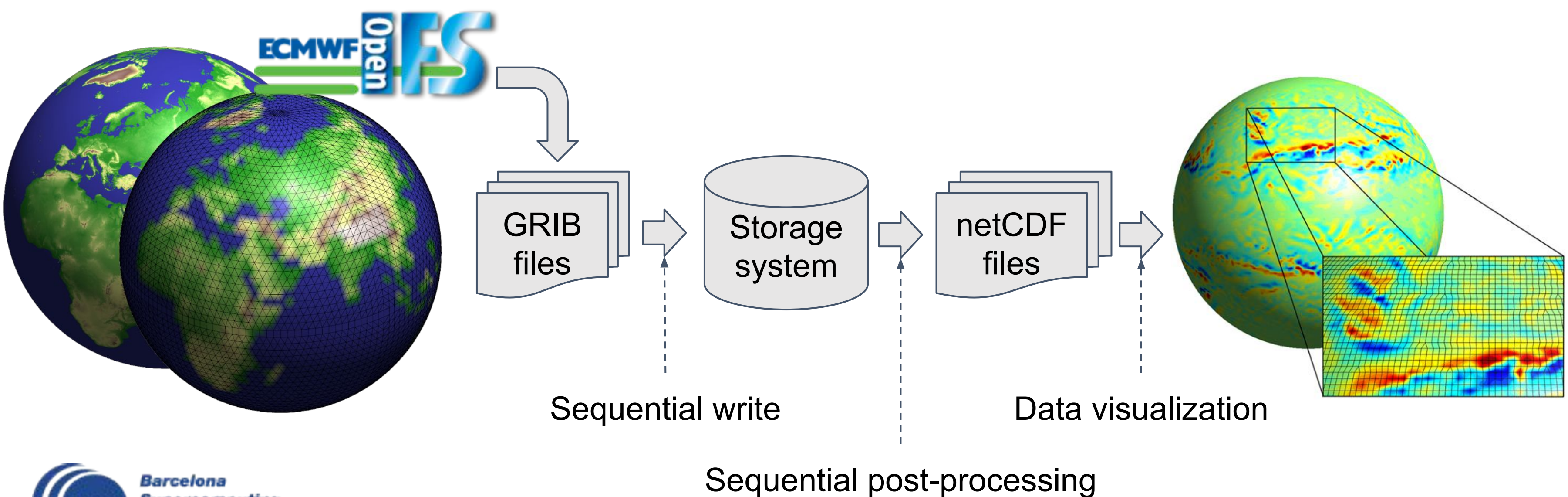
- OpenIFS can be also used for climate modelling if coupled into an Earth system model, such as EC-Earth

- However, it is necessary to perform some post-processing, including, but not limited to:
  - Convert GRIB files to netCDF files
  - Regridding
  - Transform spectral fields to grid-point fields
  - Compute diagnostics

- Post-processing turns into an expensive process, which will increase in cost, complexity and size as a result of increasing the resolution

Queries in OpenIFS

- Global-scale simulations over a large period of time produce a huge amount of data
- OpenIFS data is written “as it is” into the storage system
- However, most of the time users are not interested in all data
- When users want to query a subset of the data, they have to post-process (or filter) entire files
- This is a slow and inefficient process difficult to parallelize



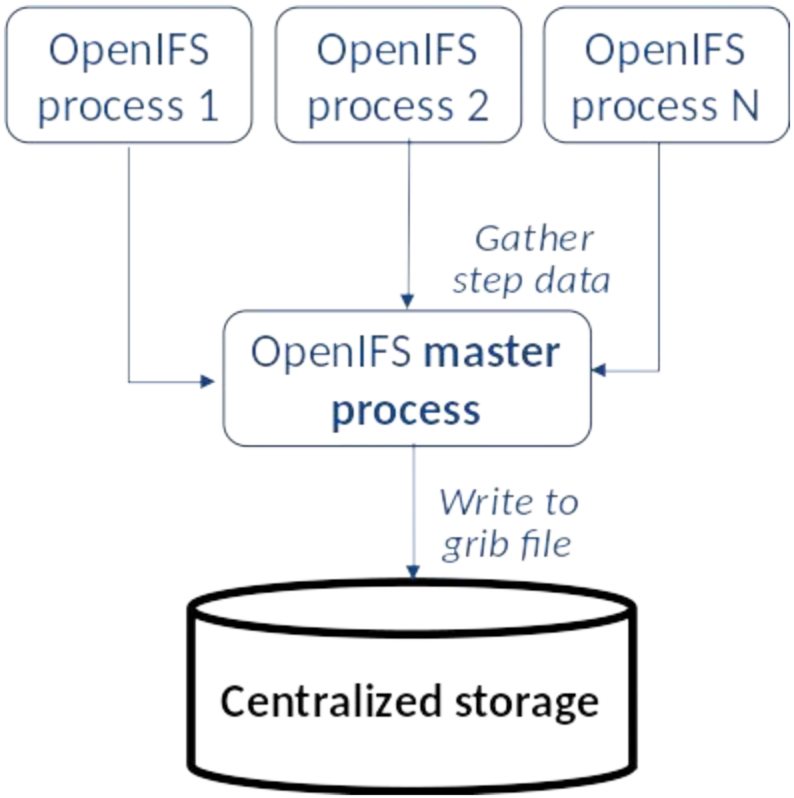
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Proposed I/O scheme

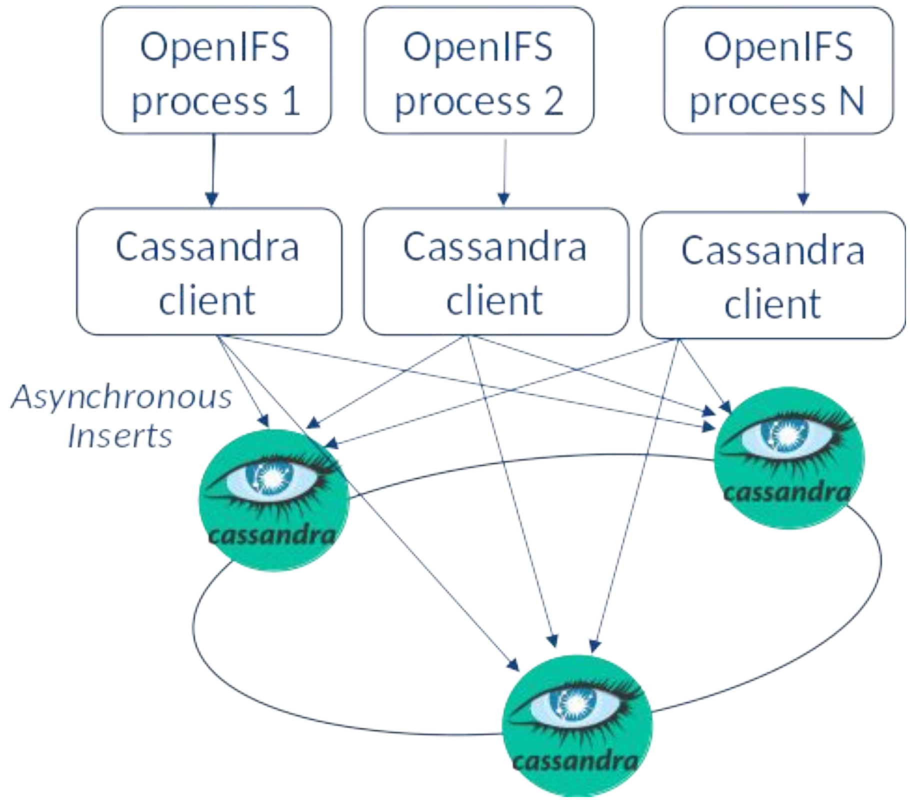
- As said, OpenIFS has a bottleneck in the sequential I/O scheme which uses a master-workers architecture
- This known issue is being addressed in the ESiWACE2 project
- Our solution is to adapt OpenIFS to perform distributed I/O
- To avoid moving the bottleneck from the model to the storage system, we opted for a distributed masterless database to store and query data: Apache Cassandra
- The main reasons are:
  - Resilience and scalability: to petabytes of data and thousands of nodes
  - Indexation: efficient distributed queries
  - Open-source

Distributed architecture

Traditional architecture



Proposed architecture



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Why Cassandra?

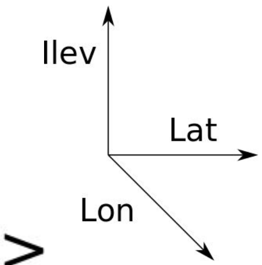
Apache Cassandra is a distributed, highly-scalable, and fault-tolerant Key-Value database.

- Pros:
  - Distributed storage, highly scalable
  - Manages coherency, corruption and concurrent access
- Cons: Bad HPC integration
  - IPoverIB
  - Scales up to few cores
  - Client based on thread pool that interferes with simulation
  - Not integrated with HPC queue systems

cassandra

We tested several data models

1) Represent each grid point as a row to distributed arrays.
   - Pros: Easy to implement
   - Cons: Millions of small insertions → Overhead



$$<lat_i, lon_j, ilev_k, ts_l>, < field0_{ijkl}, field1_{ijkl},...>$$

2) Arrays are partitioned and distributed.

$$<ts_i>, <array[field0_{0..nlat, 0..nlon, 0..nilev, i},...]>$$

- Our tool Hecuba partitions the arrays
  - Transparent to the client → simplifies its implementation
- We chose to partition the array into 4K chunks to avoid congesting Cassandra.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

cassandra

OpenIFS configurations evaluated

- OpenIFS 40r1
- Initial conditions: storm Xaver (1 December 2013) → maximum forecast length of 8 days

	AMIP	HighResMIP	Theoretical
Resolution	T255L91	T639L137	T1279L137
# MPI processes	47	141	752
# OpenMP threads	2	2	2
Time step	2700 seconds	900 seconds	600 seconds

OpenIFS configurations evaluated (2)

	AMIP	HighResMIP	Theoretical
Resolution	T255L91	T639L137	T1279L137
Forecast length	8 days	8 days	4 days*
Output frequency	3 hours (4 time steps)	3 hours (12 time steps)	1 hour (6 time steps)
Total size GRIB files	9 GB	80 GB	795 GB
Total size Cassandra (replication factor 1)	37 GB	355 GB	4224 GB

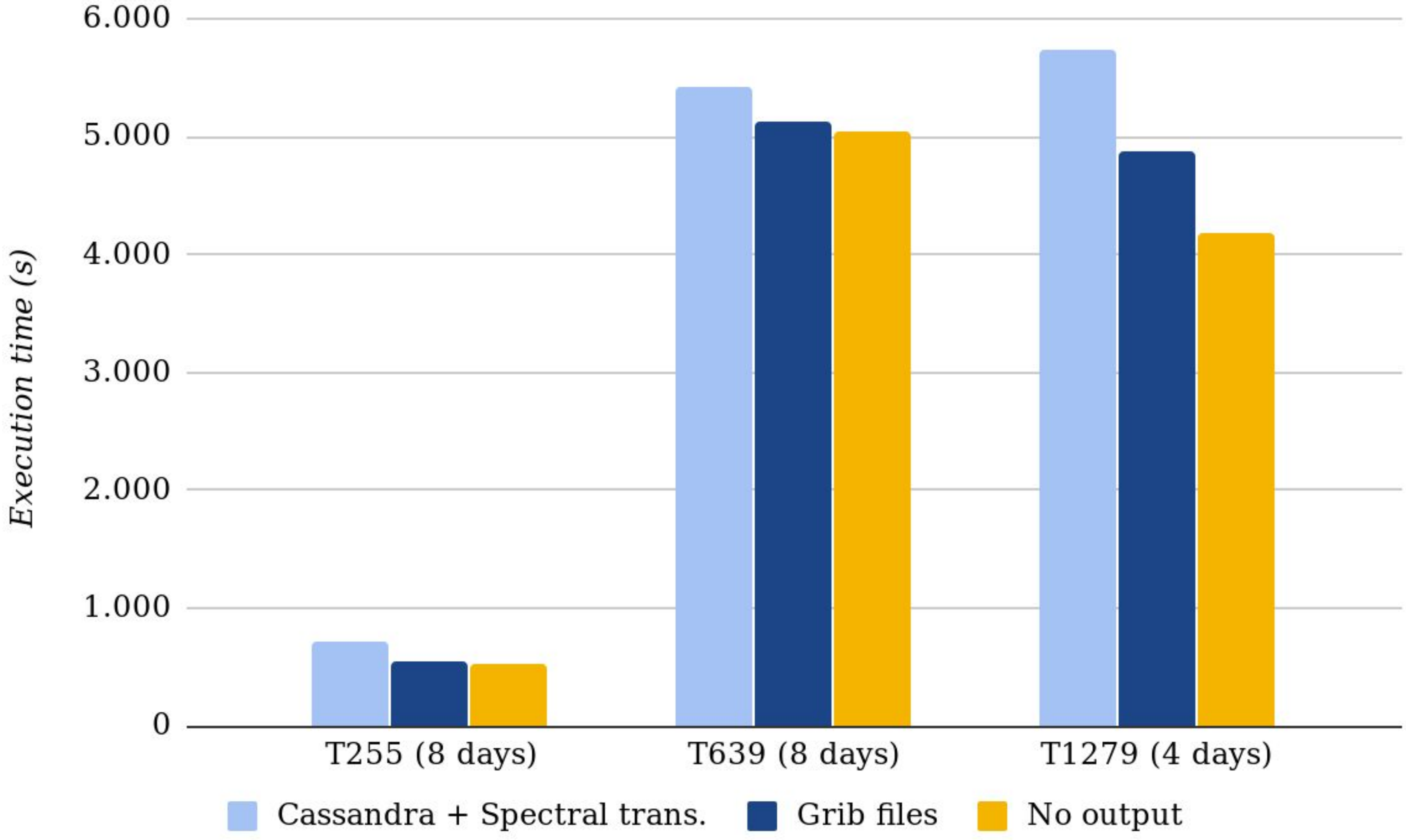
*Quota exceeded issue

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HPC platform used

- MareNostrum 4
- Lenovo system composed of:
  - SD530 Compute Racks
  - Intel Omni-Path high performance network interconnect
  - SuSE Linux Enterprise Server
- Compute nodes are equipped with:
  - 2 sockets Intel Xeon Platinum (Skylake) with 24 cores each (48 cores per node)
  - 96 GB of main memory
  - 100 Gbit/s Intel Omni-Path
  - 10 Gbit/s Ethernet
- IBM GPFS file system

OpenIFS performance comparison



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

cassandra

Performance Analysis

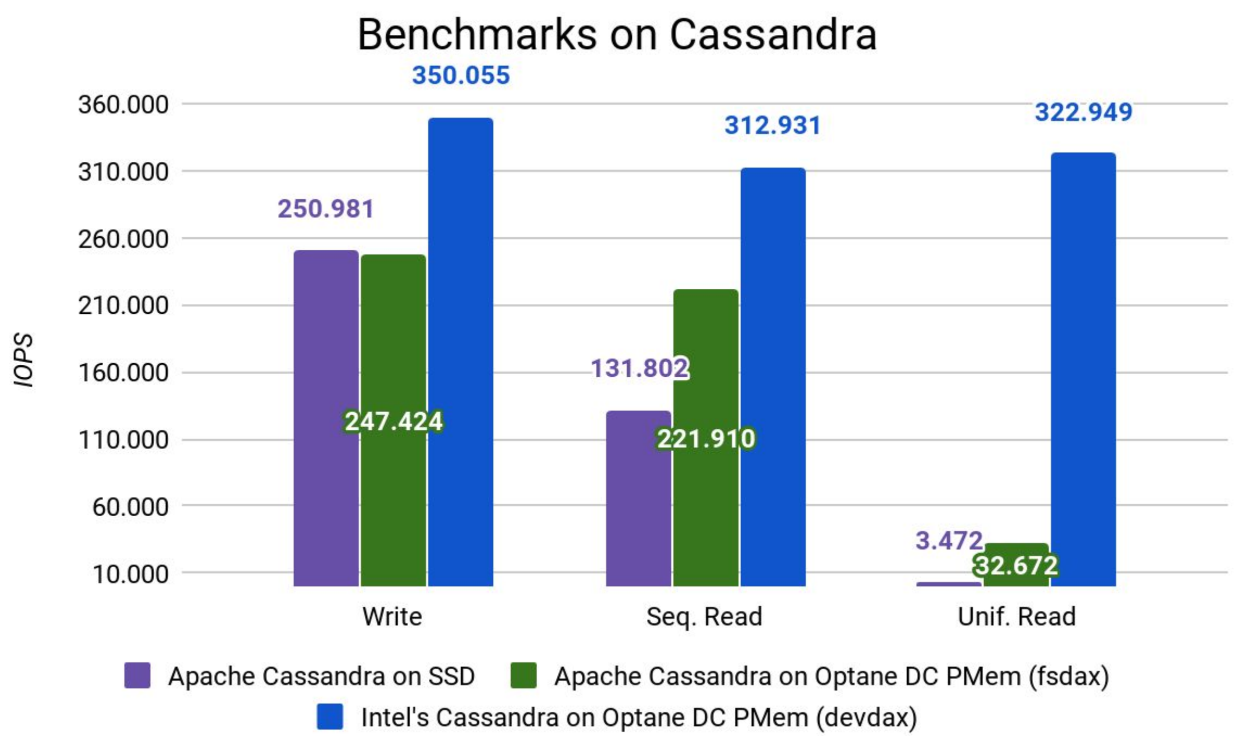
- Different output size (1 to 4.5 ratio GRIB files vs. Cassandra)
  - Data is stored in Cassandra using double precision but GRIB stores using a scale factor + offset to use variable precision
- Spectral fields transformed to grid-point fields prior to output only with Cassandra
- Cassandra integration with HPC hardware could be improved
- Slightly slower than files but the benefits are on queries:
  - Efficient filtering (post-processing operations) and sampling
  - Distributed computing
  - Concurrent access while data is generated

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

cassandra

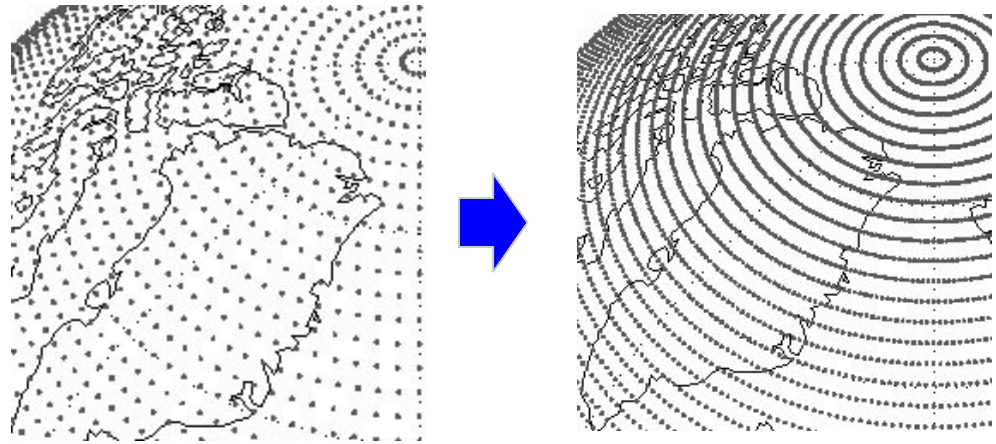
BSC-Intel collaboration

- Apache Cassandra adapted by Intel to run on the Intel(R) Optane(TM) DC Persistent Memory
  - Will boost the performance of both OpenIFS and post-processing applications



Common post-process operations

- Regridding
  Reduced Gaussian to rectangular



- Temporal queries over a period of time: average, maximum, minimum, cumulate, etc

- Spatial queries: regional domain extraction, collapsing in the vertical direction or in the horizontal, etc

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Post-process workflow

- Typical approach for implementing queries in the Earth Science domain:
  - Bash or Python script that manipulates the OpenIFS output files. Filters entire files, and each post-process step generates new files.

- Our proposal: parallel post-processing using PyCOMPSs and Hecuba
  - PyCOMPSs offers an automatic parallelization of sequential code
  - Hecuba offers a transparent access to persistent objects in Cassandra

Conclusions

- Room for improvement: Cassandra's C++ Client makes use of threads, which conflict with MPI

- Potential benefits for the post-processing operations
  - Parallel execution
  - Easy to implement
  - Possibility of using standard libraries to manipulate the data (NumPy)

- Intel's Cassandra is promising: we expect a performance boost thanks to the Intel(R) Optane(TM) DC Persistent Memory

Future work

- Understand the impact on OpenIFS of the Cassandra client threads
- Complete and evaluate the distributed queries
- Write data in double, single or smaller precision depending on the user needs
- Evaluate OpenIFS taking advantage of Intel's Cassandra in a distributed environment

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

xavier.yepes@bsc.es
pol.santamaria@bsc.es