Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

ESIWACE2: OIFS-XIOS benchmarking

Xavier Yepes-Arbós, Mario C. Acosta, Miguel Castrillo, Kim Serradell

Computational Earth Sciences Performance Team

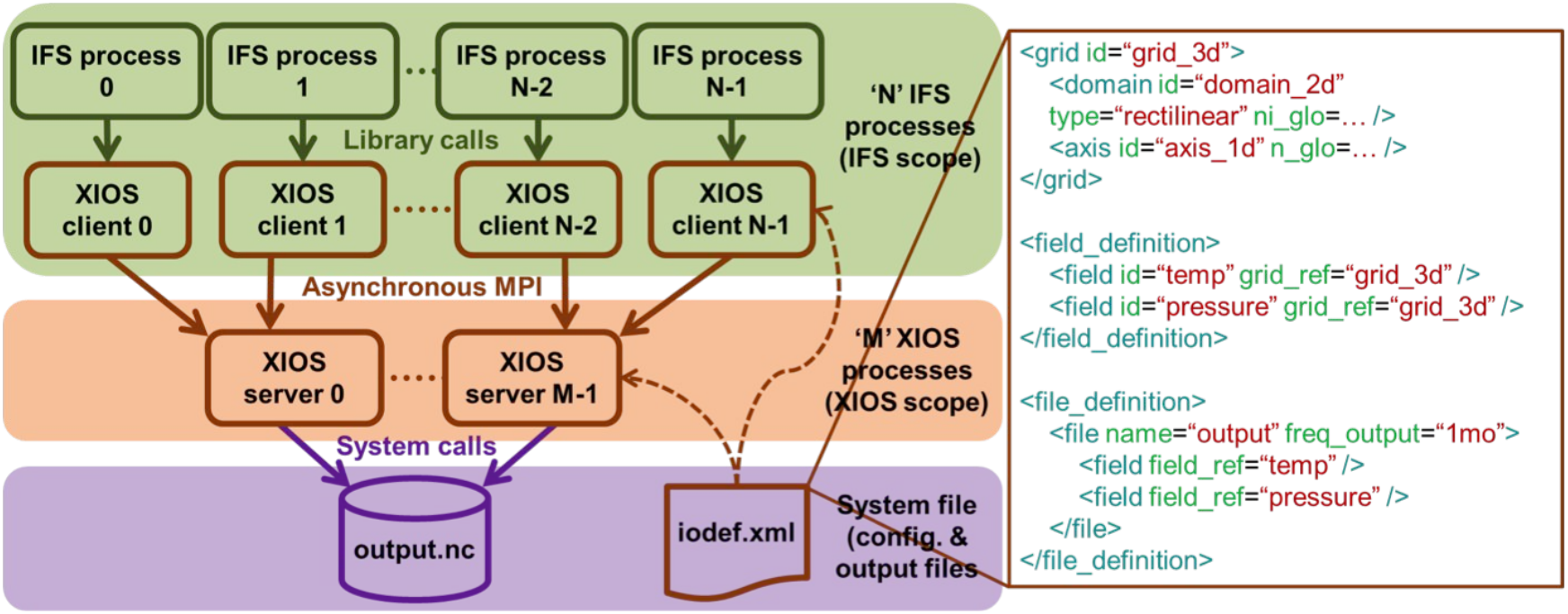
28/07/2020

esiwace2
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

(Open)IFS-XIOS

- Different XIOS features available (listed only some of them):
  - Horizontal interpolations (from reduced Gaussian to rectangular Gaussian)
  - Arithmetic operations
  - Time operations: average, maximum, minimum, etc
- The XIOS integration porting from IFS to OpenIFS 43r3 is done



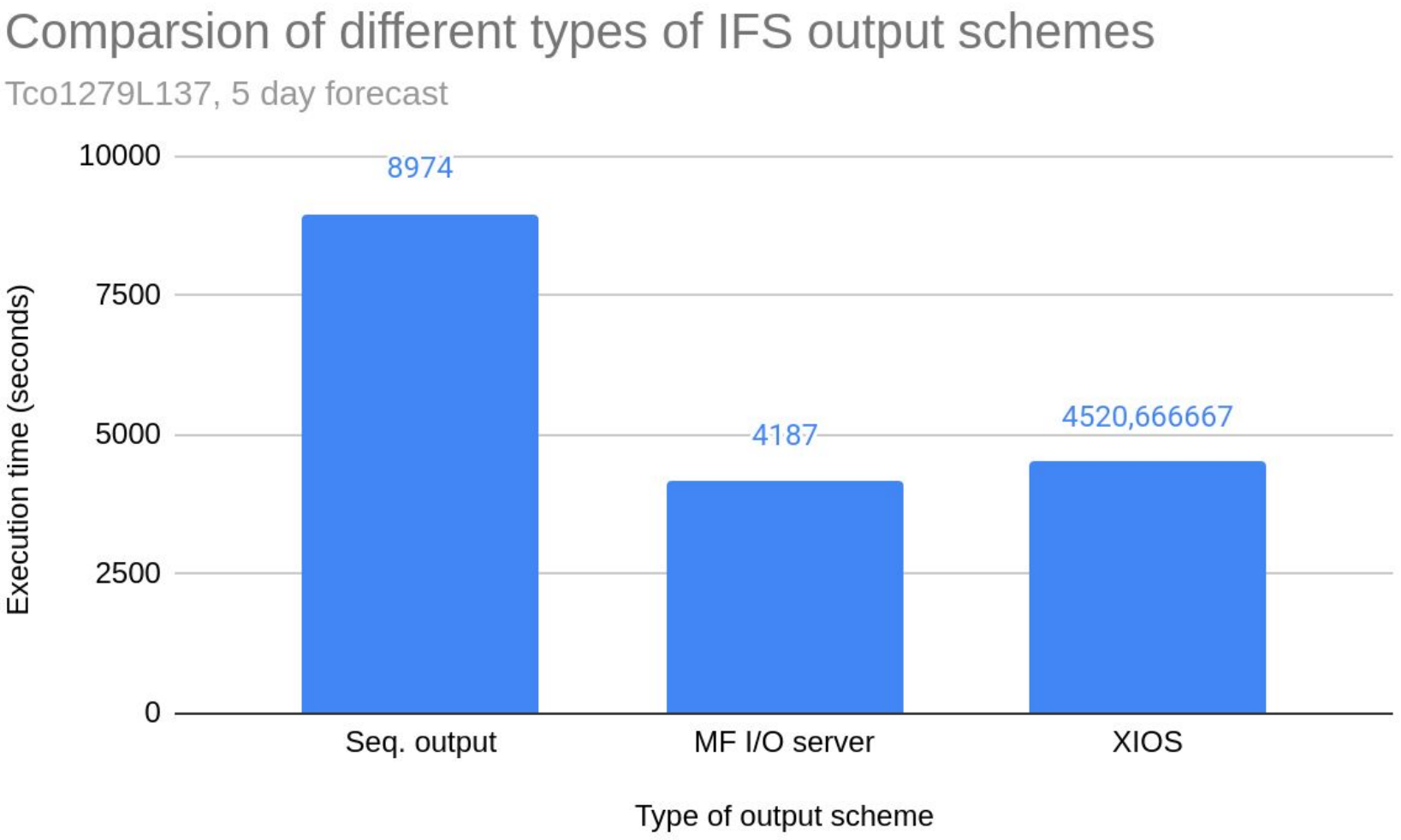
(Open)IFS-XIOS

- IFS-XIOS integration:
  - Both grid-point and spectral fields are supported
  - All 3D and surface fields
  - Different vertical levels are available: model, pressure, theta and PV levels
  - No longer needed to set up the FullPos namelist (NAMFPC)
  - FullPos spectral fitting is available
  - Physical tendencies and fluxes output (PEXTRA fields) are also supported
- Both XIOS 2.0 and 2.5 versions have been tested
- Highlights from the computational performance point of view:
  - In-depth benchmarking: the overhead of outputting data through XIOS is really small
  - A profiling and performance analysis was done to detect potential bottlenecks
  - Two different optimizations are available (switchable in the XIOS XML namelist):
    - Computation and communication overlap
    - Sends from (Open)IFS to XIOS either in double or single precision

(Open)IFS-XIOS

- Tco1279L137 grid (9 km)
- Time step of 600 seconds
- 702 MPI processes and 6 OpenMP threads and hyperthreading enabled
- A 5 day forecast and a huge hourly output configuration to stress the I/O (GRIB output size: 2.4 TB. NetCDF output size: 9.9 TB).
- The MF I/O server uses 30 servers with 3 dedicated nodes and XIOS uses 40 servers with 20 dedicated nodes.

(Open)IFS-XIOS



XIOS Benchmarking

- A final benchmarking will be done through ESiWACE2 to analyze the efficiency of XIOS for expensive configurations as the EC-Earth demonstrator
- Interact with XIOS team while doing the tests to adjust some parameters for performance if needed.
- Explore the problems with a real model (OpenIFS here) and then develop a toy model reproducing the problem to analyse it and work on it. Yann has this toy model environment and it could be run on different machines (e.g. CMCC by Italo).
- **Computing environment that will be used**:
  - MareNostrum 4
    - Computing nodes: 48 cores, 96 GB of main memory and 100 Gbit/s Intel Omni-Path
    - GPFS filesystem

Output scalability

Test how XIOS servers and memory consumption scales by only outputting data and according to these conditions:

- Use two different I/O loads:
  - Real output configuration: Similar to a CMIP6 experiment with both 2D and 3D variables.
  - Very large output configuration: Theoretical experiment with many 3D fields and high frequency.
- Fix a large amount of MPI processes for the model
  - Scale the number of nodes exclusively dedicated to XIOS servers.
  - Scale the number of XIOS servers within the “XIOS node” (nodes exclusively allocated for XIOS servers).
- Fix an amount of XIOS servers and exclusive nodes to be used for them.
  - Scale the number of MPI processes of the model to analyze how XIOS behaves when the number of clients increases.

Output scalability

Test how XIOS servers and memory consumption scales by only outputting data and according to these conditions:

- Use both one_file and multiple_file modes
- Use both one or two level servers (ratio, pools, timeseries)
- Test the affinity of XIOS servers to reduce data movement
- Local storage strategy: Configure XIOS to write chunked netCDF files directly to the local scratch of the XIOS servers nodes.

Post-processing cost

Determine the cost associated to perform common filters:

- Spatial filters (e.g. horizontal interpolation from unstructured to regular).
- Temporal filters (e.g. daily average).
- Arithmetic filters (e.g. compute wind speed and direction from u and v components).
- Compression filter.
- CMCC could contribute by performing the benchmarking in the CMCC Zeus machine of the critical parts by using the toy model framework available in XIOS.

Profiling

Make use of advanced performance analysis tools such as Extrae and Paraver.

- Basic performance analysis (including XIOS part for the MPI scalability curve of the client model).
- Analyze both the computational performance and memory consumption of netCDF parallel writing (one_file mode) and netCDF sequential writing (multiple_file mode).
- Analyze common filters.
- If it is possible, compare the XIOS performance in a GPFS or a Lustre filesystem.

Grand Challenge

Different tests using 3072, 6144, 12K, 24K, 36K, 49K, 79K, 98K, 122K and 148K cores. Without output

- 8 MPI/6 OpenMP per node
- 48 MPI per node

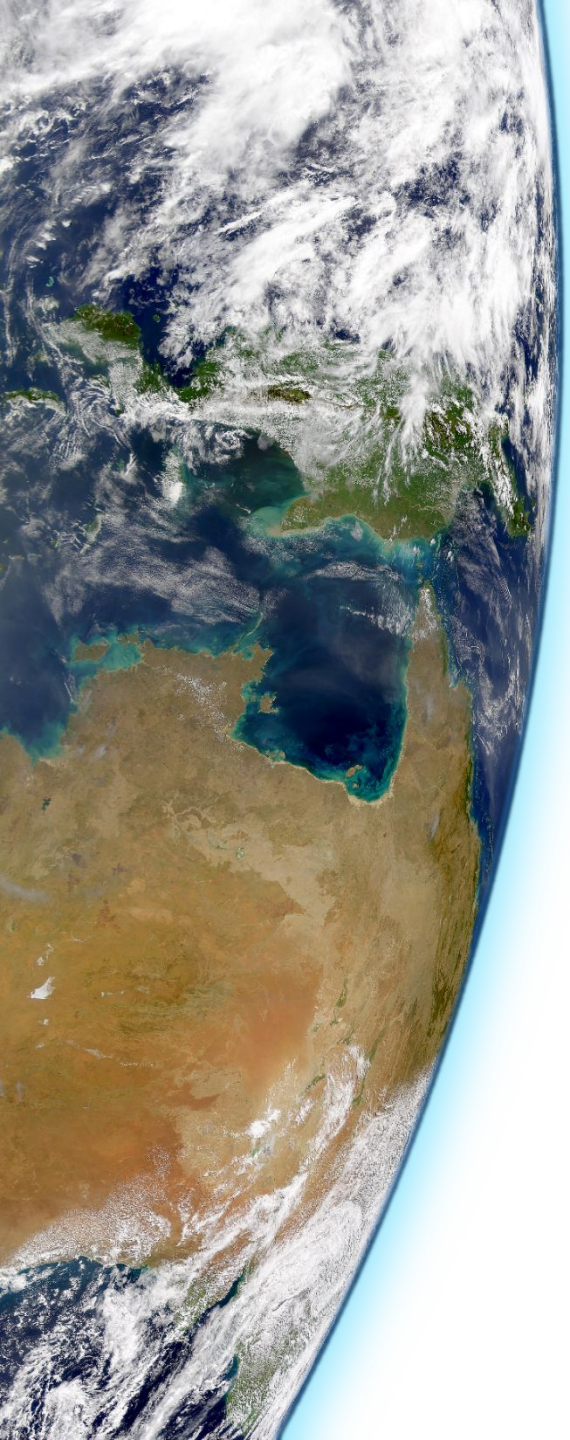
Tests with 3072, 6144, 12K, 24K, 36K, 79K and 122K. With output (HighResMIP)

- Setup 1: 8 MPI-6 OpenMP per node. 1 level of servers, multiple file, 10 nodes for XIOS with 2 servers per node
- Setup 1bis: Same Setup 1 without output (writing process).
- Setup 2: 48 MPI por nodo (without OpenMP). 1 level of servers, multiple file, 10 nodes for XIOS with 2 servers.
- Setup 2bis: Same Setup 2 without output (writing process).
- Setup 3: 47 MPI OpenIFS and 1 XIOS server. 2 levels of servers, multiple file, 10 nodes for the second level of XIOS servers. 1 XIOS server of level 1 per OpenIFS node.

Grand Challenge

Tests with 3072, 6144, 12K, 24K, 36K, 79K and 122K. With output (theoretical)

- Setup 1: 8 MPI-6 OpenMP per node. 1 level of servers, multiple file, 25 nodes for XIOS with 2 servers per node
- Setup 1bis: Same Setup 1 without output (writing process).



Thank you!

mario.acosta@bsc.es