



**Barcelona  
Supercomputing  
Center**

*Centro Nacional de Supercomputación*



# Optimizing Climate Models with Accelerators and emerging Technologies

Mario Acosta, Christian Guzman, Alexey Medvedev and Xavier Yepes

22-10-24

Barcelona, October 2024



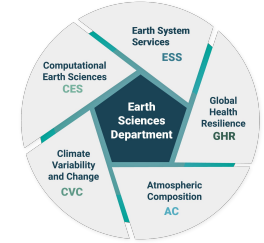
## Earth Sciences Department (BSC)

Develop and implement **environmental modelling** and **forecasting** using process-based and AI models, focusing on **weather, climate, and air quality**



Transferring solutions to support the main **societal & environmental challenges** through service development.

- 190 people
- Funding from Horizon programmes, Copernicus and DestinE, private sector, ESA, Spanish, and regional governments
- Close link to local universities

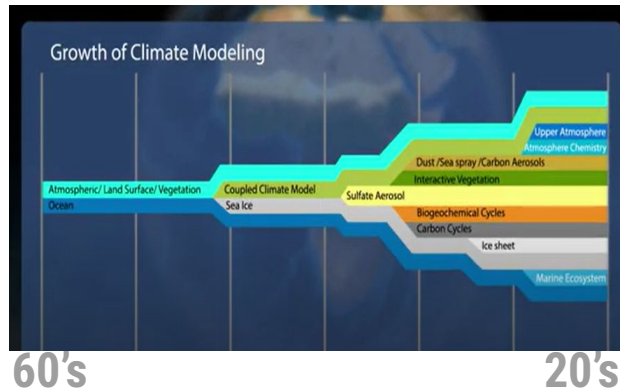


**Barcelona  
Supercomputing  
Center**  
Centro Nacional de Supercomputación

Why HPC and accelerators in  
weather, ocean  
and climate modelling?

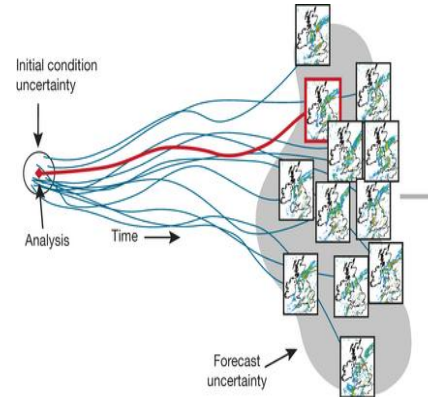
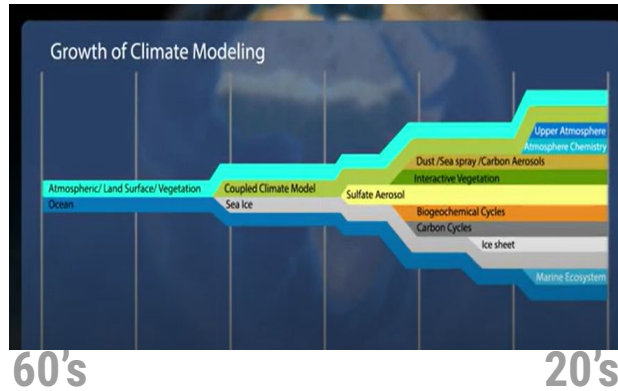
# Why HPC and accelerators

## Increasing complexity of ocean and climate models



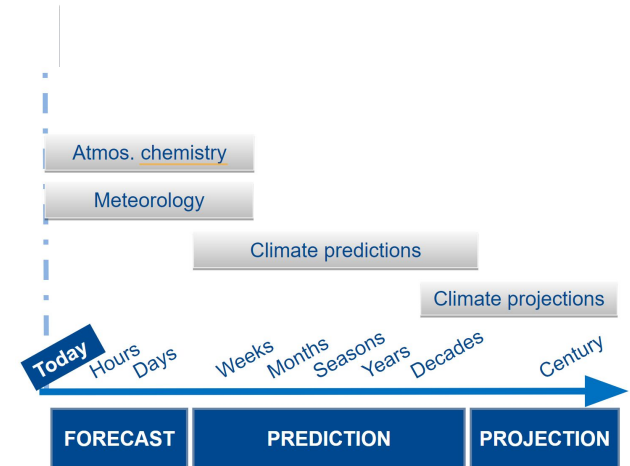
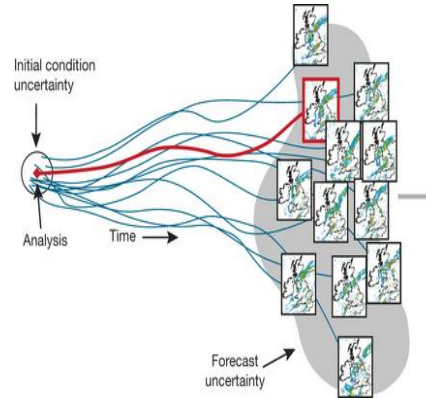
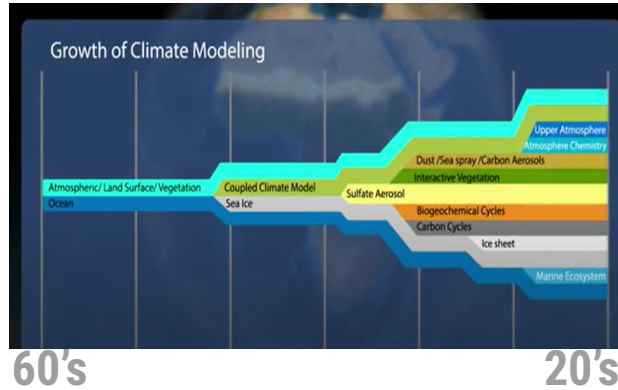
# Why HPC and accelerators

## Increasing complexity of ocean and climate models



# Why HPC and accelerators

## Increasing complexity of ocean and climate models

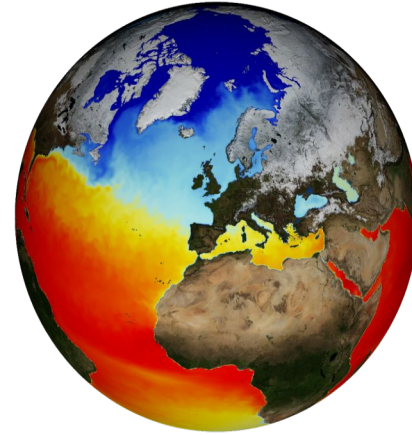


Destination Earth Climate-DT model (IFS-NEMO).

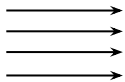


# Why HPC and accelerators


- Computational power is critical in Earth science models.
- Simulations use a **huge amount of computational resources**.
- Complex simulations will need much more resources.
- Mostly of Earth science models use parallelization based on spatial domain decomposition
- Mostly calculations are independent but some interaction is needed among subdomains
- Take advantage of the specific hardware is mandatory
- Optimizations techniques are needed to increase the performance of these models, this is know as High Performance Computing

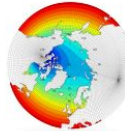


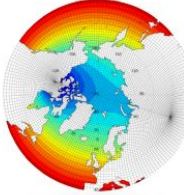
Grid resolution  
Mem needed  
Simulation time  
Outputs size



  
ORCA 2  
550 MB of memory  
8 CPU hours  
10 Gigabytes of output (daily)

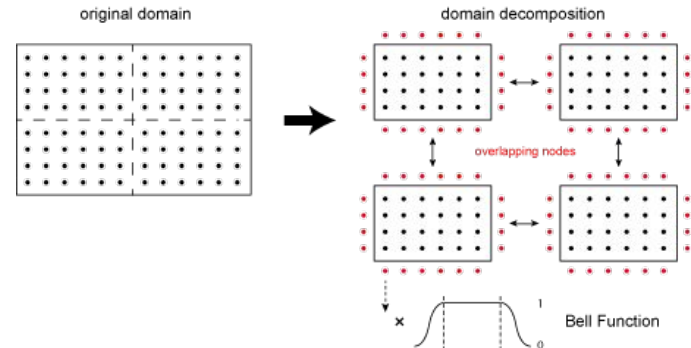
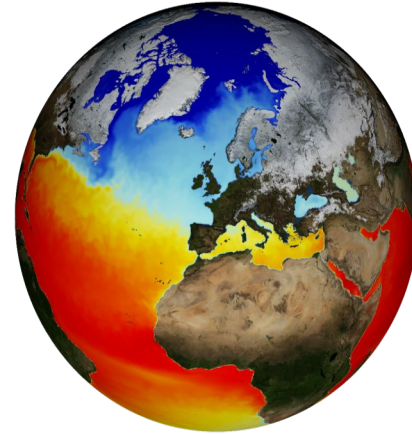
  
ORCA 1/4  
47 Gigabytes of memory  
3500 CPU hours  
120 Gigabytes of output (daily)

  
ORCA 1/12  
414 Gigabytes of memory  
90 000 CPU hours  
1 Terabyte of output (daily )

  
ORCA 1/36  
> 1 Terabytes of memory  
~4 000 000 CPU hours  
> 5 Terabytes of output (daily)

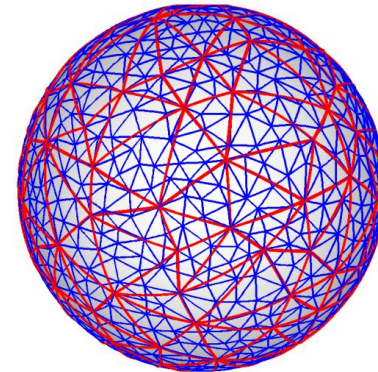
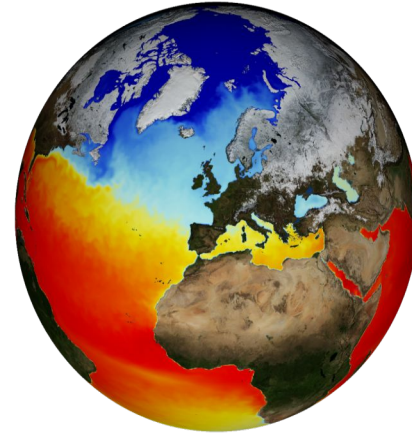
# Why HPC and accelerators

- Especially critical in Earth science models.
- Simulations use a **huge amount of computational resources**.
- Future simulations will need much more resources.
- Mostly of Earth science models use parallelization based on spatial domain decomposition
- Mostly calculations are independent but some interaction is needed among subdomains
- Earth Science Models are very complex
- Take advantage of the specific hardware is mandatory
- Optimizations techniques are needed to increase the performance of these models, this is know as High Performance Computing

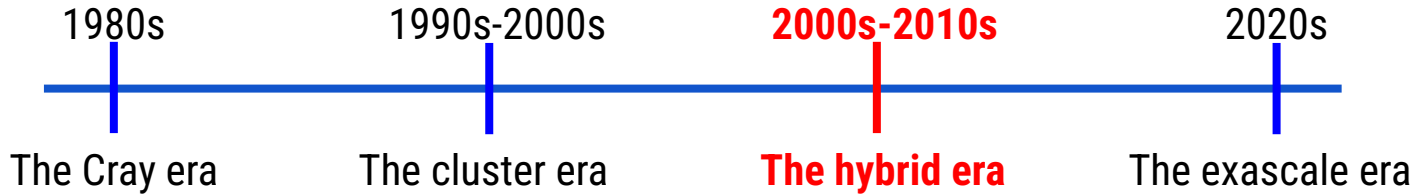


## Why HPC and accelerators

- Especially critical in Earth science models.
- Simulations use a **huge amount of computational resources**.
- Future simulations will need much more resources.
- Mostly of Earth science models use parallelization based on spatial domain decomposition
- Mostly calculations are independent but some interaction is needed among subdomains
- Earth Science Models are very complex
- Take advantage of the specific hardware is mandatory
- Optimizations techniques are needed to increase the performance of these models, this is known as High Performance Computing



# HPC history and challenges



- These systems contain large numbers (hundreds to thousands) of small efficient cores (many-cores)
- Graphics Processing Units (GPUs) were integrated into HPC clusters to accelerate the performance.
  - This approach essentially “offloaded” certain types of operations from the CPU to the GPU.

# HPC history and challenges



- Exascale machines will be more purpose-built
- Extending multi-core/many-core clusters to the Exascale range is hampered by the disconnection between hardware and software.
- Heterogeneous computing and co-design as solution
  - Fixed Accelerators provide order(s) of magnitude more specialized performance
  - The main problem could be complexity and programmability

## The pre- and exascale path

- What we can gain
  - New computing elements (GPUs, FPGAs, AI, Quantum, RISC-V)
  - More parallelism (Million threads)
  - Much more computing (Exaflops)
  - Data streaming
  - More data (Exabytes)
  - More complexity in our models (increased resolution, more parameters or components, more ensembles)
  - Larger Datasets

## The pre- and exascale path

- What we can gain
  - New computing elements (GPUs, FPGAs, AI, Quantum, RISC-V)
  - More parallelism (Million threads)
  - Much more computing (Exaflops)
  - Data streaming
  - More data (Exabytes)
  - More complexity in our models (increased resolution, more parameters or components, more ensembles)
  - Larger Datasets

**Modeling and  
simulation**



**Machine  
Learning**



**Big Data  
Analytics**

- In short time
  - New pre-exascale machines (LUMI, LEONARDO, MareNostrum5).
  - High-resolution “Digital Twins” using EuroHPC hardware
  - European Projects pursuing HPC improvements (ESiWACE3)
  - Hardware and software acceleration (IA, GPUs)
- In medium/long time
  - RISC-V, Quantum, FPGAs
  - Clouding
  - Hardware and software acceleration (IA, GPUs)
  - ...

**Modeling and  
simulation**

# ESiWACE3



# Destination Earth



# EU-PILOT



# ESiWACE3 - Centre of Excellence in Simulation of Weather and Climate in Europe

**ESiWACE3 focuses to support the weather and climate modelling community to reach the excellence regarding exascale supercomputing**



Ocean at different resolutions using the EC-Earth model performed by Oriol Tintó (BSC)

**Coordinated**



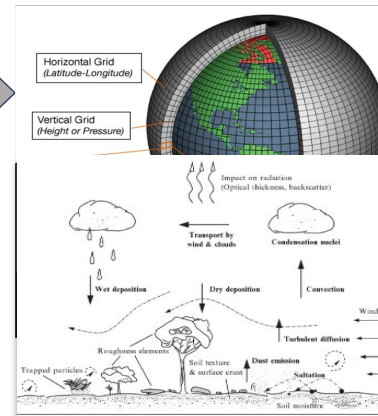
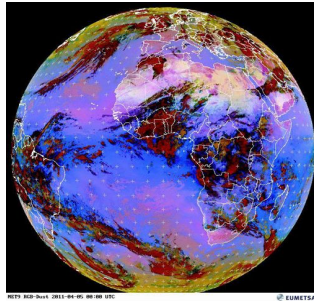
**Consortium of 12  
partners from 8  
different countries**



**Start: 1 January 2023  
End: 31 December 2026**

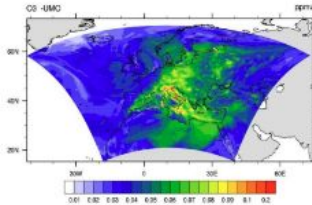
# MONARCH (Multiscale Online Nonhydrostatic Atmosphere Chemistry)

- A chemical weather prediction system.
- In-house developed.
- Provides operational regional mineral dust forecasts for the World Meteorological Organization and aerosol forecasts to the International Cooperative for Aerosol Prediction (ICAP) initiative.
- Accurate simulations depends on the availability of substantial computing power and data storage capacity.



## Chemistry solvers

- Resolution of chemical processes can take up to 90% of the time execution of atmospheric models.
- MONARCH recently integrated CAMP (Chemistry Across Multiple Phases) to enhance accuracy (Dawson et al. GMD 2022).
- CAMP is adapted to integrate latest developments using accelerators (GPUs).
- Chemistry simulations are composed of millions of Ordinary Differential Equations (ODEs), requiring high computational power to ensure high accuracy.



Millions of equations  
( $y' = f(t,y)$ ), one for each  
point of the map

CAMP  
(Solve  
Chemistry)

## Implementation

- The CPU version distributes the workload between MPI cores.
- Each MPI solve a region of the atmosphere.
  - A region is compressed by multiple grid-cells (a map point in the atmosphere)
    - A cell contains multiple chemical concentrations to solve (System of Ordinary Differential Equations)
- **The new GPU version distributes the regions workload between CPU and GPU.**
  - The CPU part is solved as the CPU version.
  - **The GPU component further distributes the workload of solving a chemical concentration between each GPU thread.**

## Results

---

- Up to 8x times faster than the CPU in a node-against-node comparison.
- High performance and memory efficiency, close to the ideal bound.
- Simultaneous usage of GPU and CPU resources near optimal value.

## EUPILOT project

- **Motivation:** The aim of the European PILOT project is to design, build, and validate the first EU-based accelerator platform for HPC
- **Goal:** demonstrate an all European technology by achieving different goals, such as:
  - Open source hardware for HPC
  - Software and hardware co-design
  - System software
  - HPC and HPDA applications
  - ...
- **Weather and climate prediction** models are key HPC applications to be adapted to this new European accelerator
  - Extremely high computational cost to solve very high resolutions

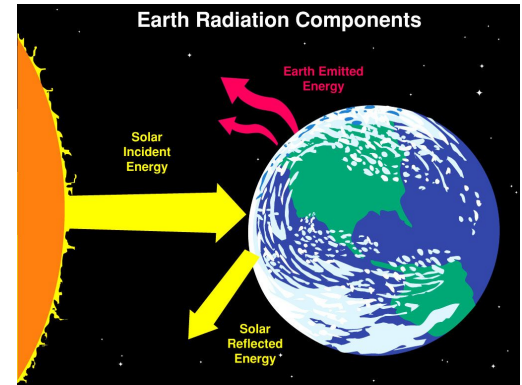
## Target applications

- **Earth system models (ESMs)** are very complex applications, such as EC-Earth
- Instead of the whole ESMs, we focus the analysis on the so-called **dwarfs**, which encapsulate basic algorithmic functionalities
- Some of them are particularly suitable to be run in vector units as they are very **computationally intensive** and **embarrassingly parallel**.

CLOUDSC: cloud  
microphysics  
scheme



ACRANEB2:  
radiation transfer  
scheme



## Auto-vectorization analysis

- Identify the most **time-consuming loops** with Extrae
- Dwarf **compilation** with LLVM Flang for RISC-V architecture + vector unit
- **Emulate** the generated code with Vehave
- Analyse the vector instructions generated (compiler **auto-vectorization** capability) with Paraver
- Code **refactorization** to enable or improve the vectorization
- Provide feedback to help in the **co-design** of the LLVM Flang compiler

# Results

- LLVM, GCC and Fujitsu auto-vectorization comparison of the ten most time-consuming loops of CLOUDSC (pre and post refactorized code)

| Loop | LLVM Flang (RISC-V)                                   | GCC (A64FX) before refactorization   | Fujitsu (A64FX)                                   |
|------|---|--------------------------------------|---|
| 100  | Cannot prove it is safe to reorder memory operations. | Vectorised                           | Vectorised  |
| 116  | The bounds of the array cannot be identified.         | Not vectorised: control flow in loop | Vectorised  |
| 20   | Vectorised  | Vectorised                           | Loop unswitching<br>Vectorised                    |
| 13   | Vectorised  | Vectorised                           | Vectorised  |
| 74   | Control flow cannot be substituted for a select       | Vectorised                           | Vectorised Loop unrolled 2 times                  |
| 10   | The compiler could not apply the transformation.      | Not vectorised: control flow in loop | This loop is not parallelized<br>Loop unswitching |
| 68   | The compiler could not apply the transformation       | Not consecutive access               | This loop is not parallelized                     |
| 38   | Vectorised  | Vectorised                           | Vectorised  |
| 72   | Vectorised  | Vectorised                           | Vectorised  |
| 73   | Vectorised  | Vectorised                           | Vectorised  |



| Loop | LLVM Flang (RISC-V) Refactored                        | GCC (A64FX) before refactorization   | Fujitsu (A64FX)                                   |
|------|---|--------------------------------------|---|
| 100  | Cannot prove it is safe to reorder memory operations. | Vectorised                           | Vectorised  |
| 116  | The bounds of the array cannot be identified.         | Not vectorised: control flow in loop | Vectorised  |
| 20   | Vectorised  | Vectorised                           | Loop unswitching<br>Vectorised                    |
| 13   | Vectorised  | Vectorised                           | Vectorised  |
| 74   | Control flow cannot be substituted for a select       | Vectorised                           | Vectorised Loop unrolled 2 times                  |
| 10   | The compiler could not apply the transformation.      | Not vectorised: control flow in loop | This loop is not parallelized<br>Loop unswitching |
| 68   | The compiler could not apply the transformation       | Not consecutive access               | This loop is not parallelized                     |
| 38   | Vectorised  | Vectorised                           | Vectorised  |
| 72   | Vectorised  | Vectorised                           | Vectorised  |
| 73   | Vectorised  | Vectorised                           | Vectorised  |

## Conclusions

---

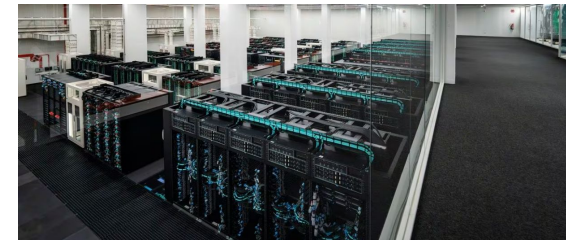
- Dwarfs evaluated and refactorized can exploit the **maximum vector length** of the RISC-V vector unit
  - Computationally intensive functionalities (dwarfs) of the ESMs **ready to run** in the RISC-V vector unit
- The auto-vectorisation capabilities of LLVM Flang are **behind** other state-of-the-art compilers
  - Helped in the compiler **co-design** to improve it
- Pending to run and test auto-vectorisation on the single **FPGA SDV** (not available yet)

# DESTINATION EARTH INITIATIVE

- **Context:** European Commission's programme, part of **Green Deal & Digital Strategy**.
- **Objective:** To develop **digital twins** (DTs) of the **Earth** to support decision-making.
- **Users:** Policy makers and environmentally sensitive sectors.
- **First DTs:** Climate Change Adaptation Digital Twin and Weather Extremes Digital Twin.
- **Computing resources:** Two EuroHPC pre-exascale supercomputers: LUMI (currently in operation in Finland) and MareNostrum 5 (Spain) provided by EuroHPC Joint Undertaking.



**MareNostrum 5**



# NEMO

---



- **NEMO global ocean model**

- Nucleus for **E**uropean **M**odelling of the **O**cean
- State-of-the-art modelling framework for research activities and forecasting services in ocean and climate sciences

- **Role in Destination Earth:**

- Works as an ocean modeling part in the IFS-NEMO coupled atmosphere-ocean model
- Takes 40-50% of compute time of a typical modeling workflow for Digital Twin

# NEMO on GPU

---

- **Objectives:** Create a well-performing and practical NEMO port on GPUs:
  - Not necessarily the entire application
  - Reasonable GPU utilization
  - Focused on obtaining rapid results for European initiatives:
    - a) Destination Earth – Climate Digital Twin
    - b) European Digital Twin of the Ocean (EDITO)
- **Constraints:**
  - **Portability** between different **architectures** and **NEMO** versions:
    - I) AMD and NVIDIA hardware (**LUMI** and **MareNostrum 5**)
    - II) NEMO **4.0** and NEMO **5.0** codebases
  - **Minimal code changes:** Preserve the code structure and keep existing hybrid MPI/OpenMP CPU code performing well

# NEMO on GPU

---

- **Scope of work:** Port the whole SEA-ICE (SI3) module and analyze the speedup
- **What we achieved so far?**
  - The “pilot” region is ported: Prather scheme for Sea-Ice advection
  - **Based on OpenACC** for portability, performance, and code clarity/cleanness
  - We achieved **portability** between NVIDIA and AMD using a subset of OpenACC
  - Can inform **PSyclone developments** and be combined with it in the future (we are already in touch with STFC)
  - We developed a good porting methodology and a good infrastructure supporting it
  - Speedup for adv\_x/adv\_y subroutines (pure calculations in **icedyn\_adv\_pra**):
    - up to **20x** on MareNostrum 5 GPU node (2xIntel 8460Y+ (80 cores) vs. 4xH100)
    - up to **9x** on LUMI-G node (1xAMD EPYC 7A53 (64 cores) vs. 4xM250X)

# NEMO on GPU

---

- **Problems to solve:** Not perfect general speedup and scaling, for the full Sea-Ice advection module (ORCA12, 32 nodes):
  - **2.9x** on LUMI-G nodes; **5.9x** on MareNostrum-5 ACC nodes
  - Counting device-host transfers: **1.8x** on LUMI-G nodes, **3.1x** on MareNostrum-5 ACC nodes
  - The main limiter are the LBCLNK subroutines (they contain MPI exchanges)
- **Conclusions:**
  - The NEMO/GPU porting process is going on and successfully covers 2 platforms: AMD and NVIDIA
  - We hope to port full Sea-Ice module, improve LBCLNK module performance and integrate with another effort of porting NEMO code (PSyclone-based)



**Barcelona  
Supercomputing  
Center**

*Centro Nacional de Supercomputación*

# Optimizing Climate Models with Accelerators and emerging Technologies

[mario.acosta@bsc.es](mailto:mario.acosta@bsc.es)

