www.bsc.es

Reading 26/10/2018

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Performance Study of Climate and Weather Models: Towards a More Efficient Operational IFS

18th Workshop on high performance computing in meteorology

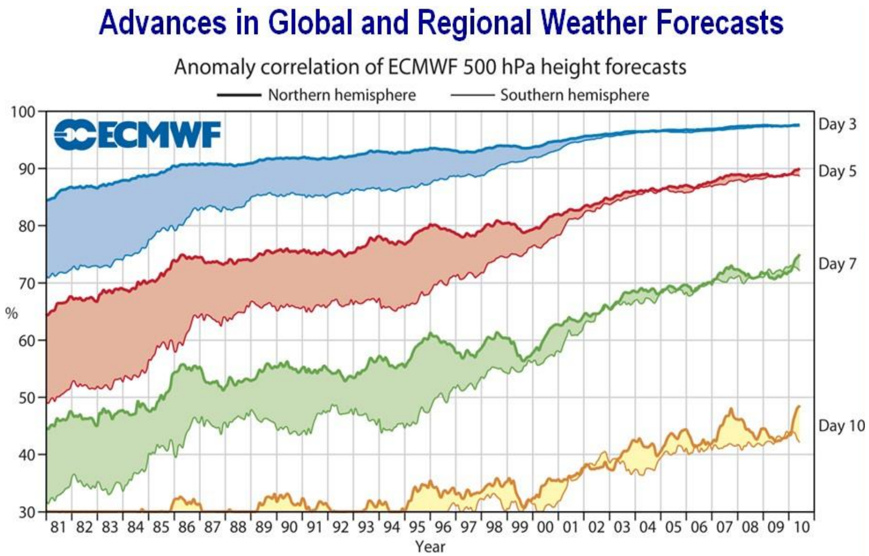
Earth Science Department
Mario Acosta and Performance Team
Computer Science Department
Jesus Labarta

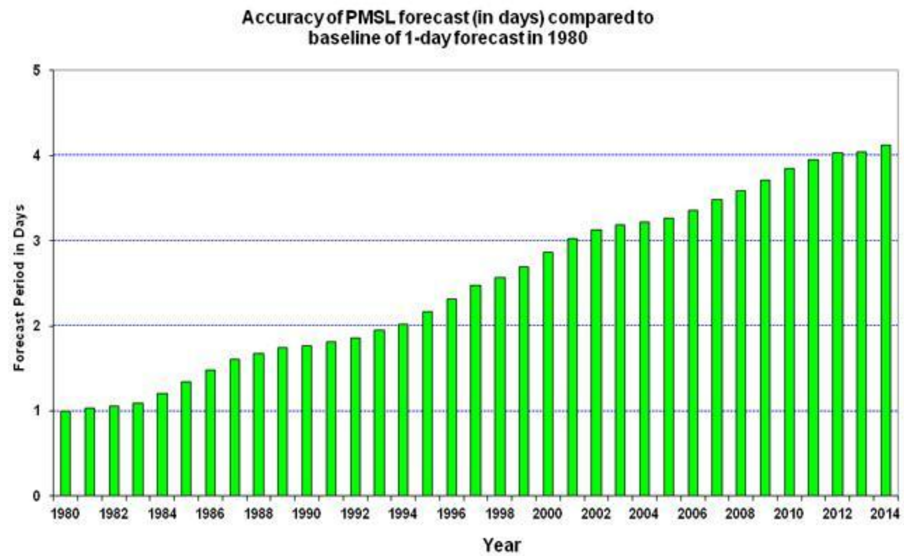
Introduction

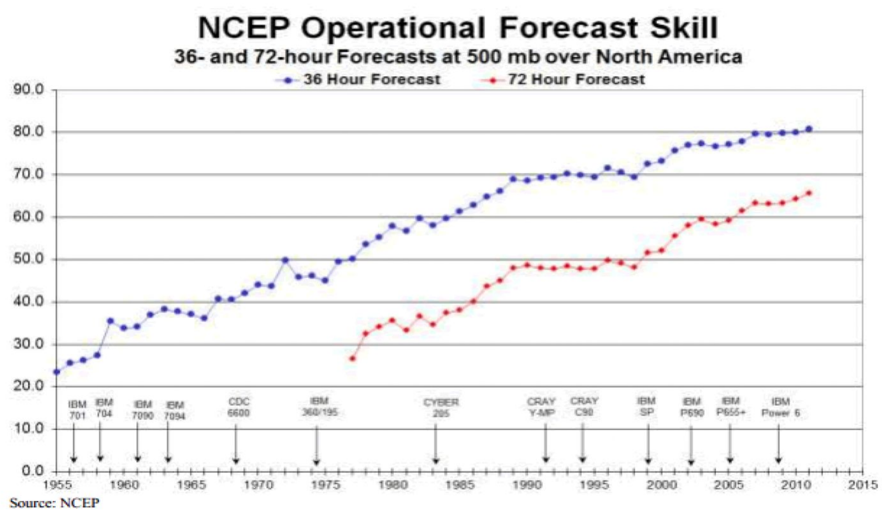
- Projections
- Impact analysis
- Adaptation to climate change.

Currently, **only computational models have the potential** to provide geographically and physically consistent estimates.

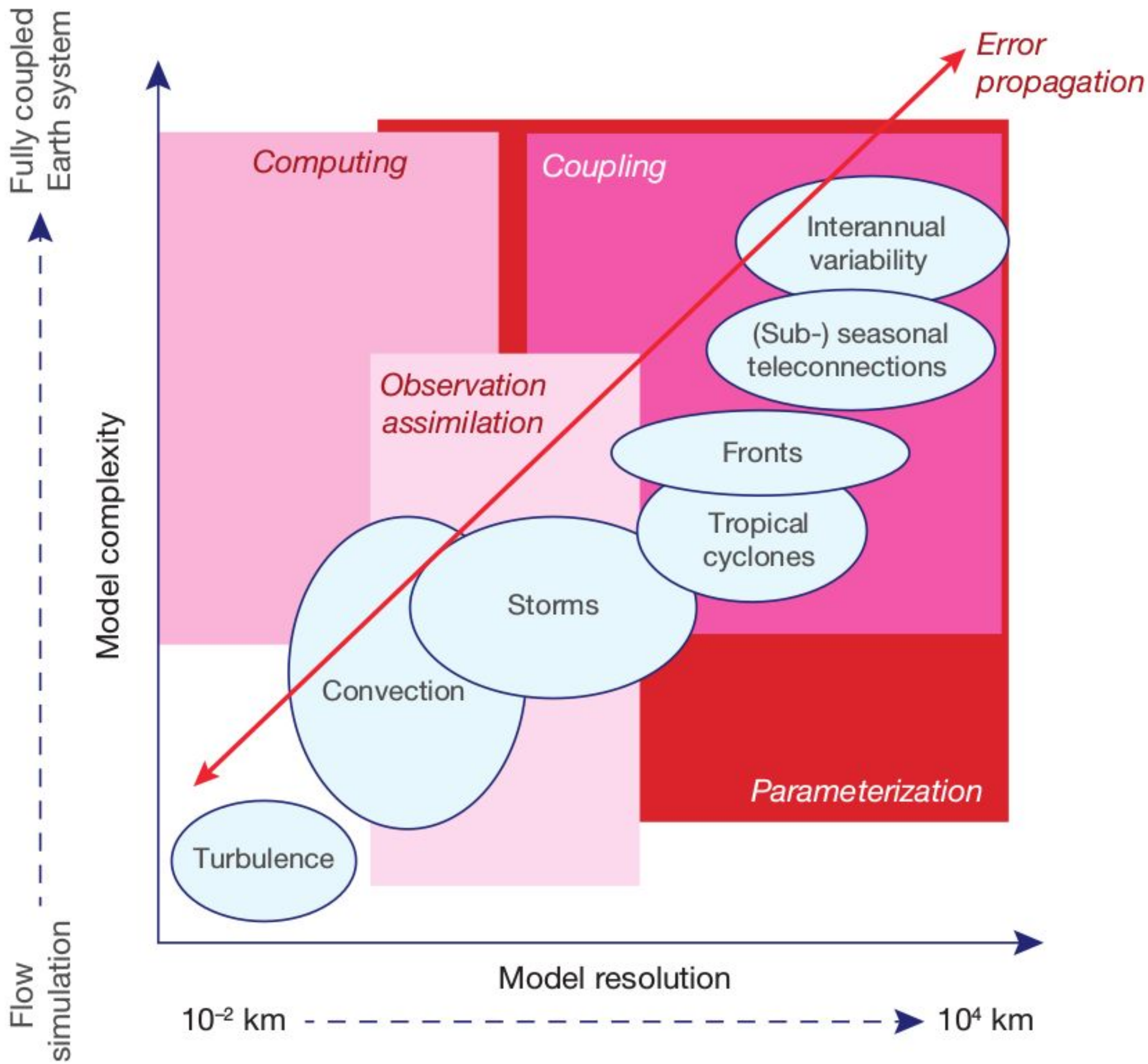
Introduction







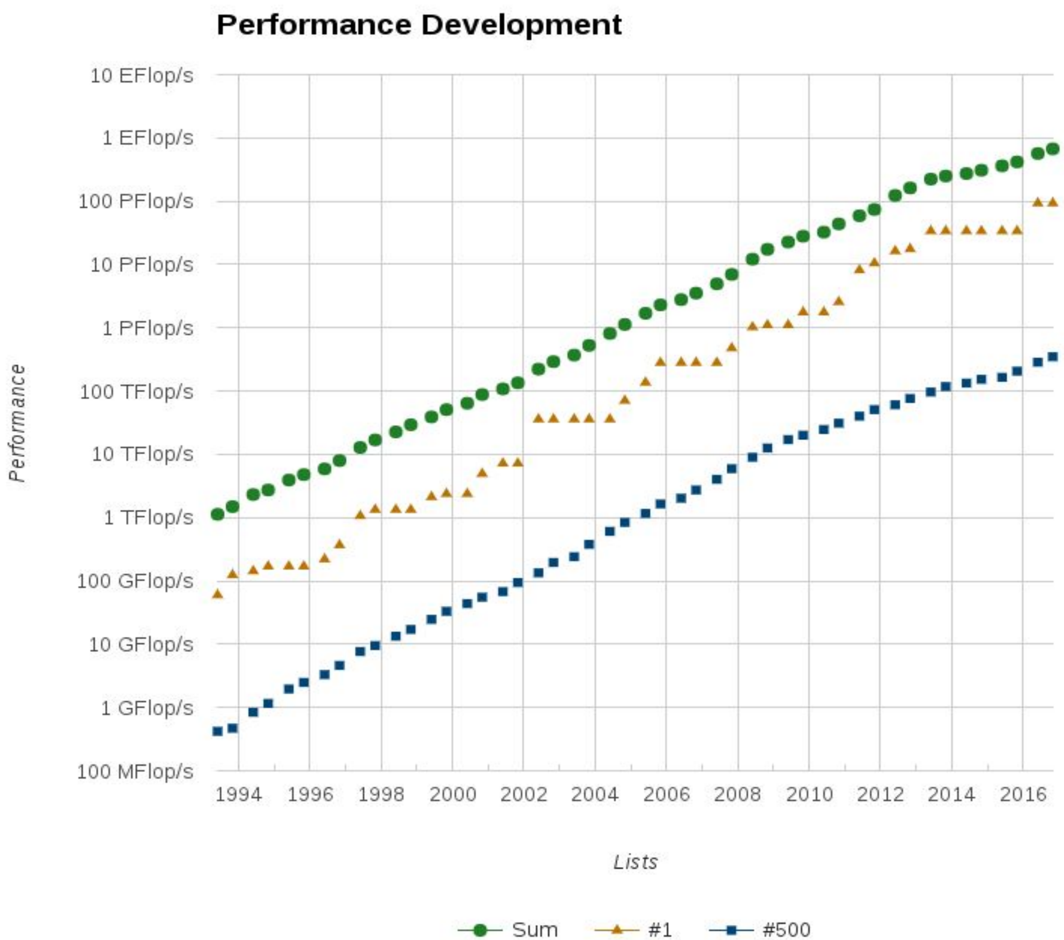
Introduction



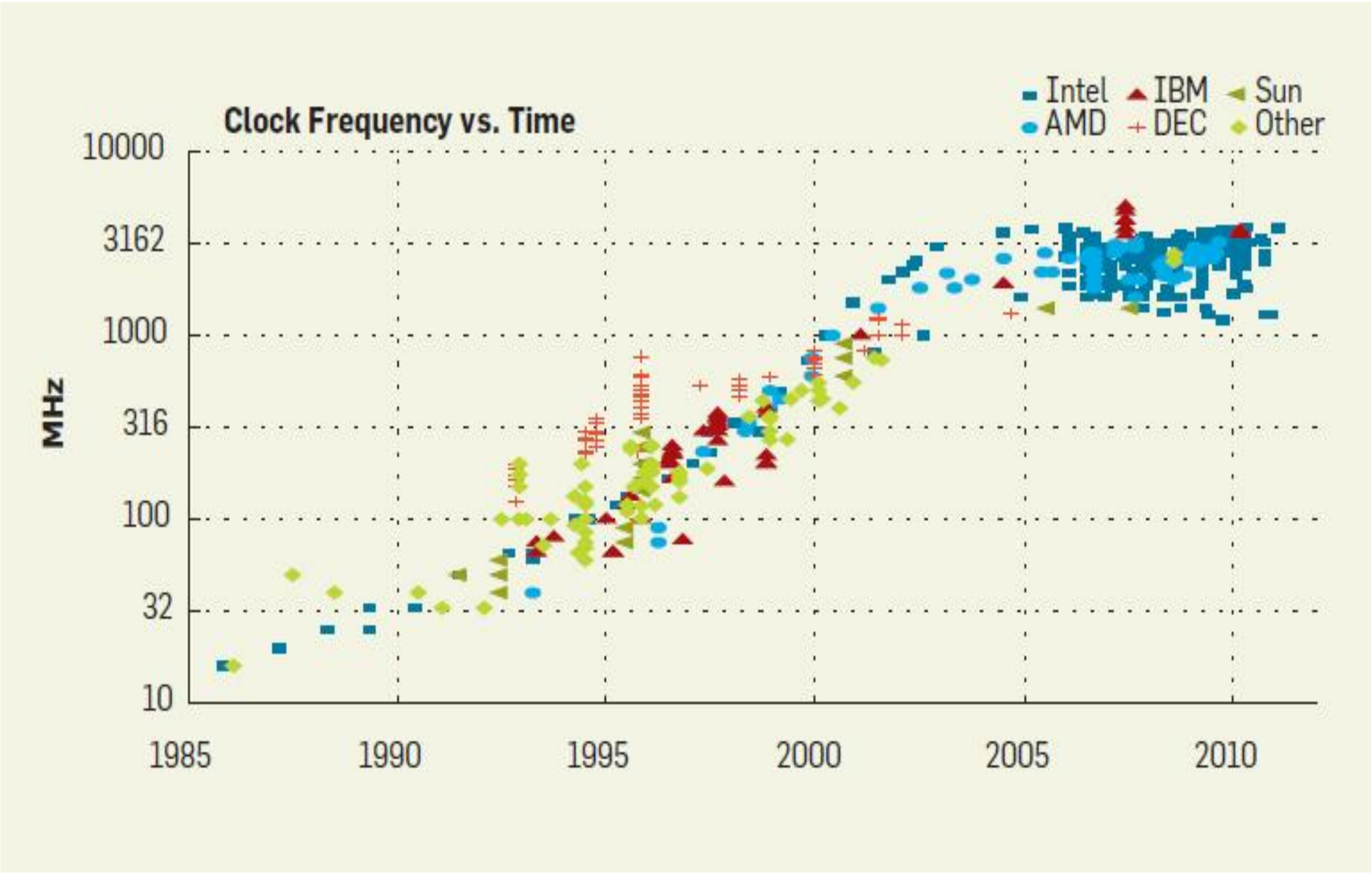
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

TOP500 Ranking of most powerful supercomputers



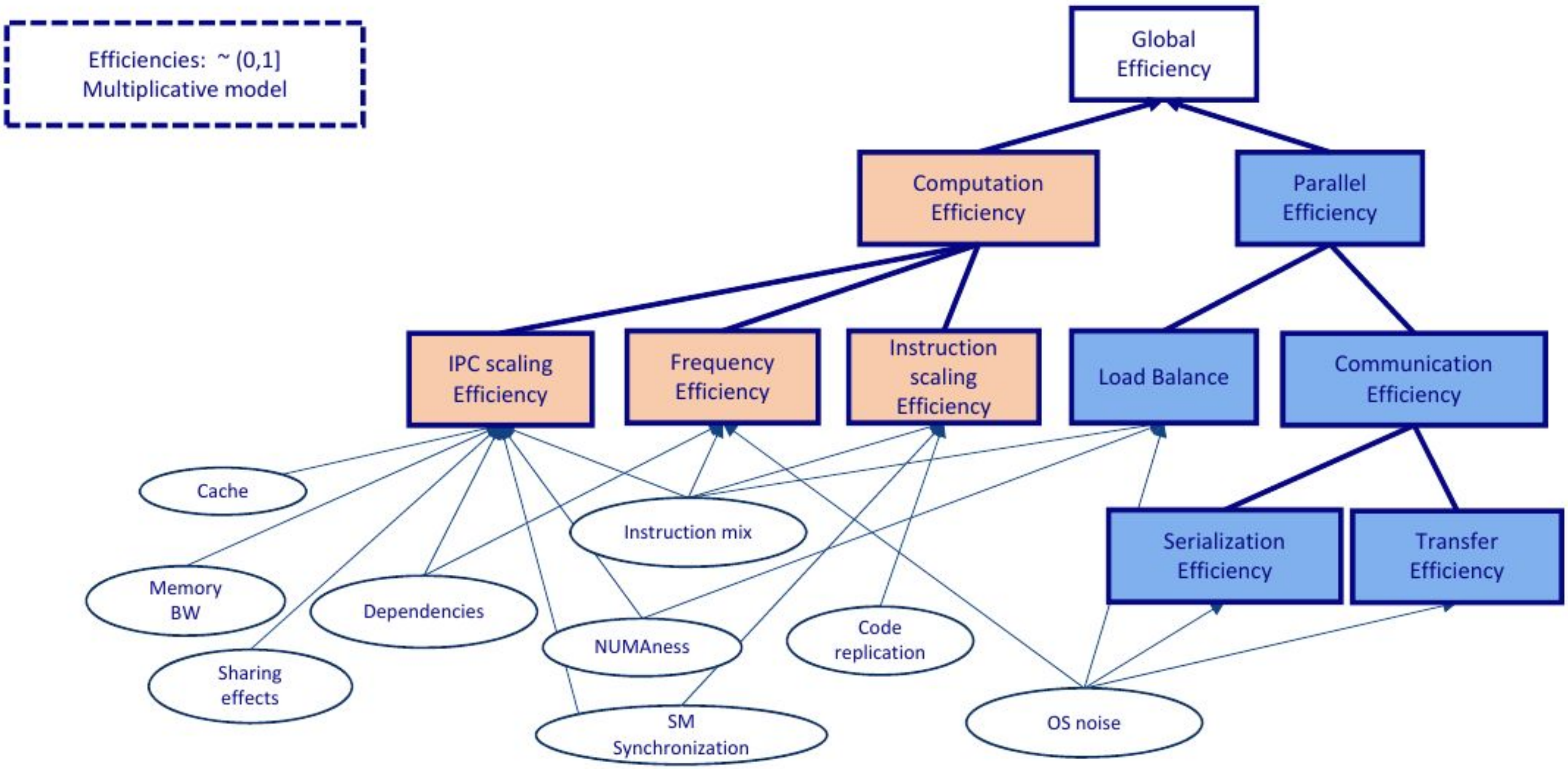
Introduction



HPC Challenges

- To be able to use the computing power of modern supercomputers, applications must exploit parallelism.
- Parallelism produce overhead (extra computation and communications)
  - "*Overhead does not look a problem in my model*" → But if the needs increase (i.e. higher resolutions), a bad implementation will be a problem in some point.
  - We need a method to evaluate the parallelism efficiency of our computational models.
    - When the hardware change
    - When the number of resources change
    - When the model complexity change
    - When the resolution change
    - …

HPC Challenges

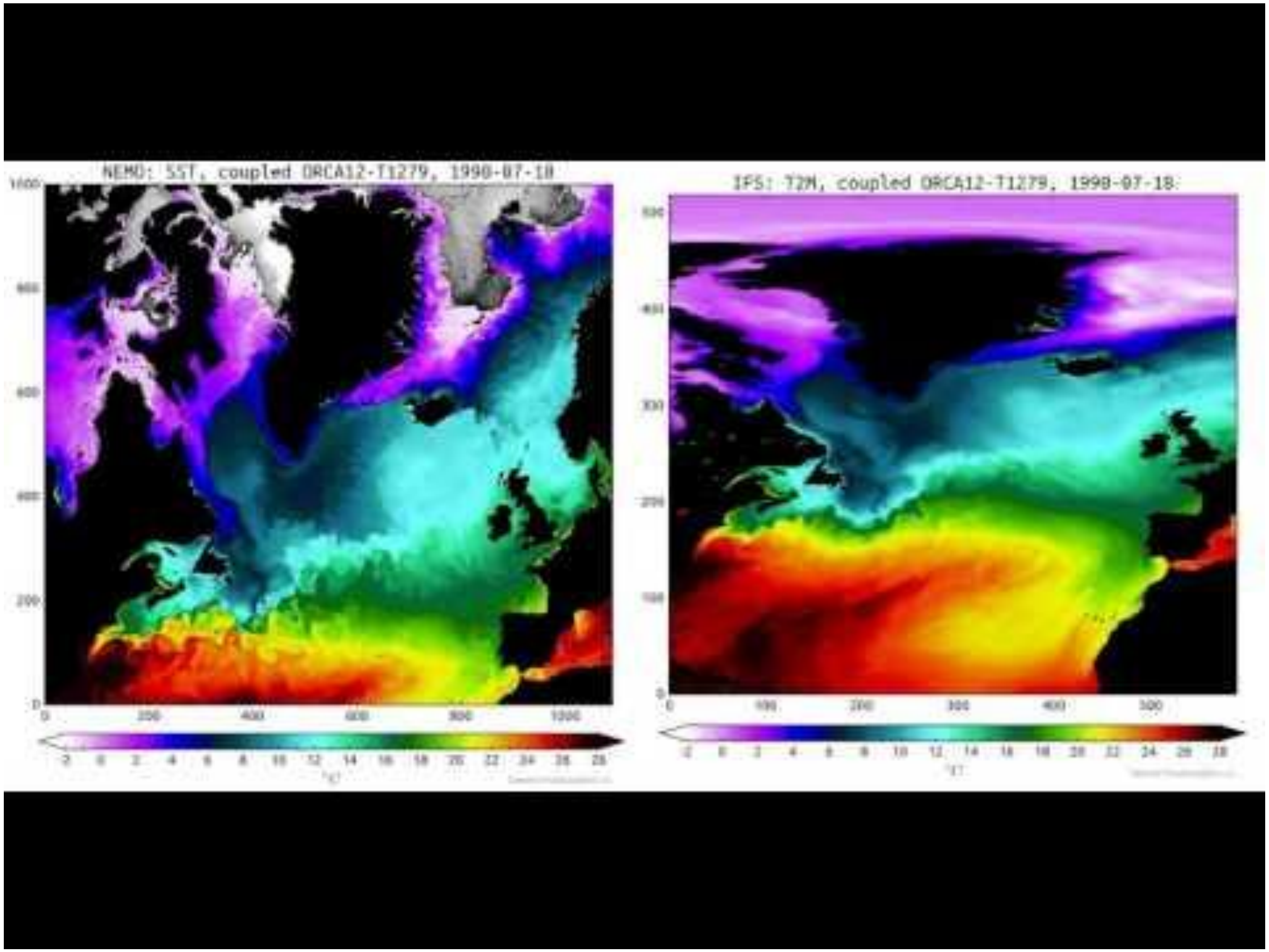


HPC Challenges

- The necessary refactoring of numerical codes is given a lot of attention and is stirring a number of discussions.
  - Computational performance analysis and new optimizations are needed for actual numerical models.
  - Study new algorithms for the new generation of high performance platforms (path to exascale).
- Several European institutions and projects working together in the same direction (ESCAPE → Dwarfs, ESiWACE → EsD’s, ETP4HPC…)
- Future H2020 where we will work → ESCAPE2, ESiWACE2, IS-ENES3

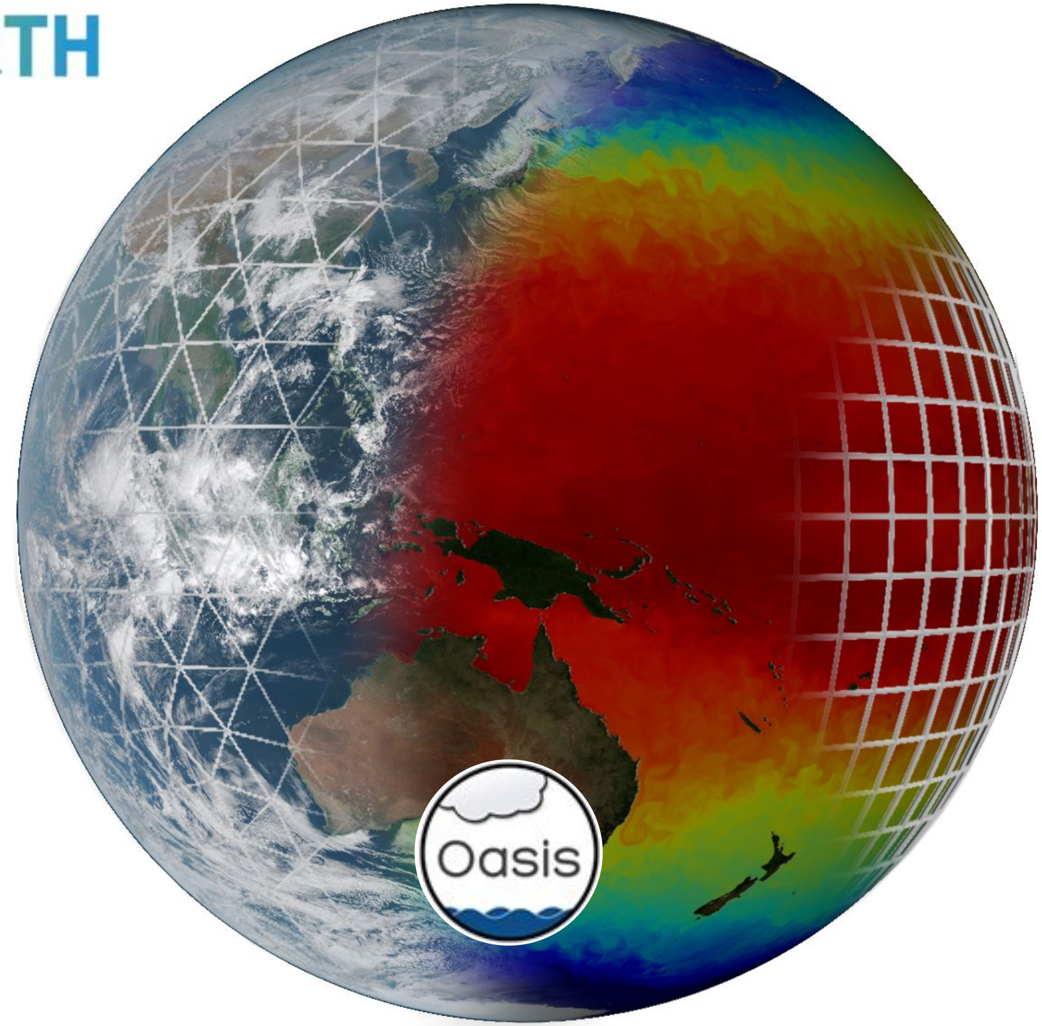
HPC Challenges

The global ultra-high resolution EC-Earth climate model ORCA12-T1279 in action over the North Atlantic.



Sea-surface temperature and sea-ice concentration from the ocean component (NEMO, left panel)

HPC Challenges



Atmosphere:
IFS
ECMWF

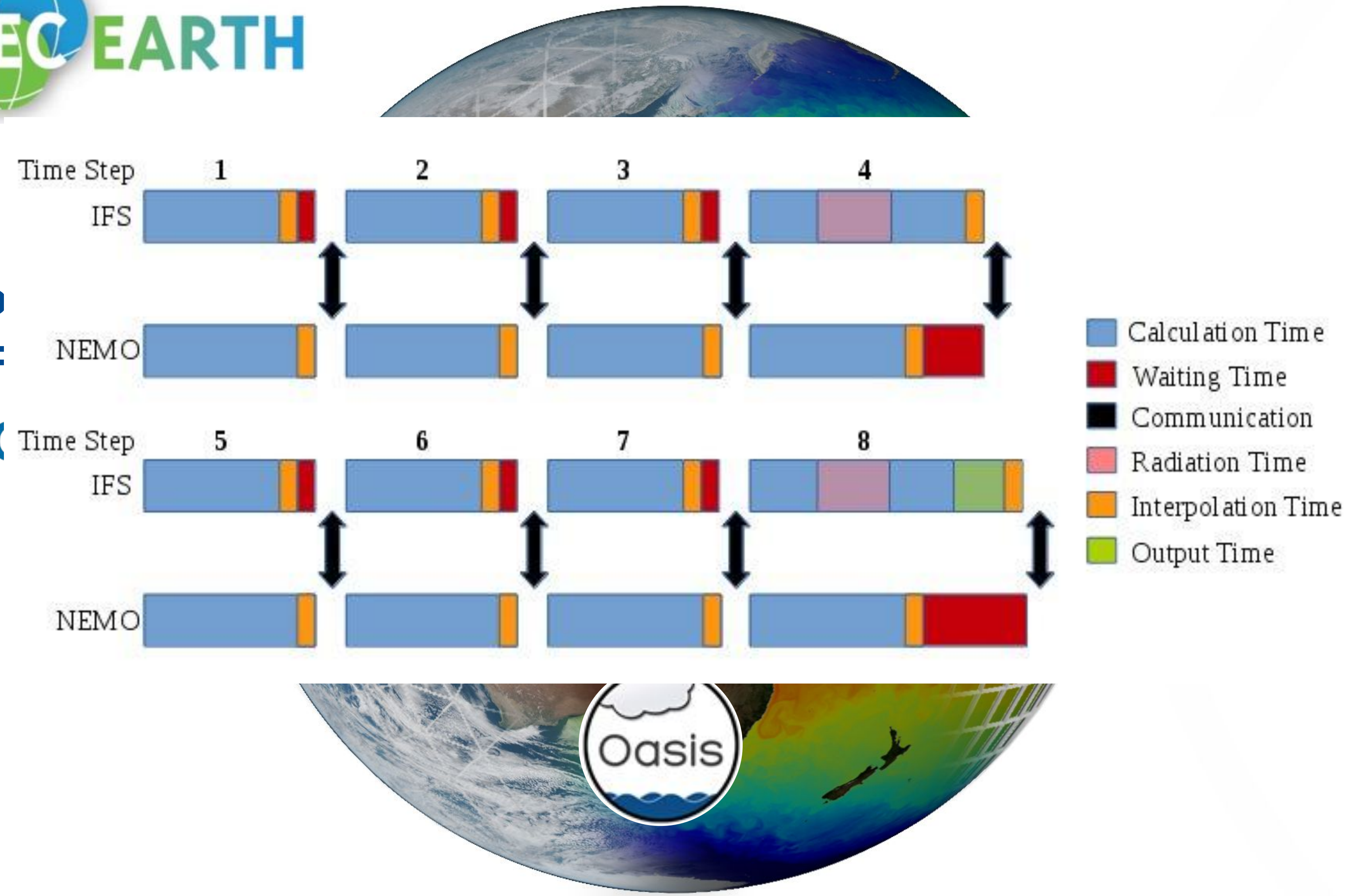
Ocean+ICE:
NEMO

Coupler:

HPC Challenges

Atmo

IF



n:

Coupler:

Optimizing the Computational Model

Weather and Climate Science

High Performance Computing (Services and Research) applied to Earth System Modelling

- Knowledge about the mathematical and computational side of Earth System Applications
- Knowledge about the specific needs in HPC of the Earth System Applications
- Researching about HPC methods specifically used for Earth System Applications

Computer Science

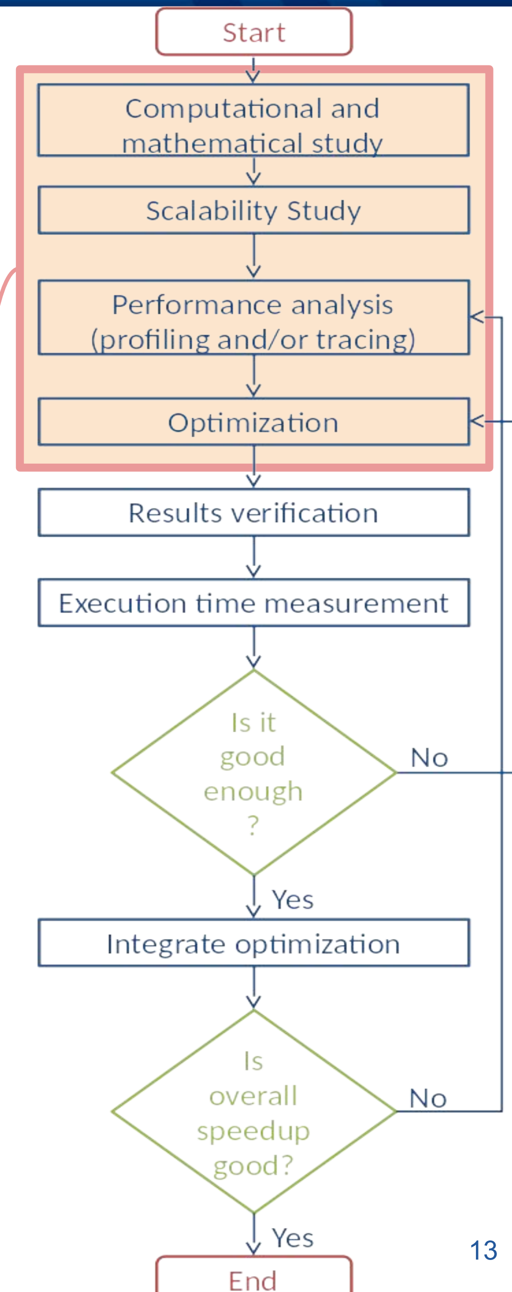
Optimizing the Computational Model

Weather and Climate Science

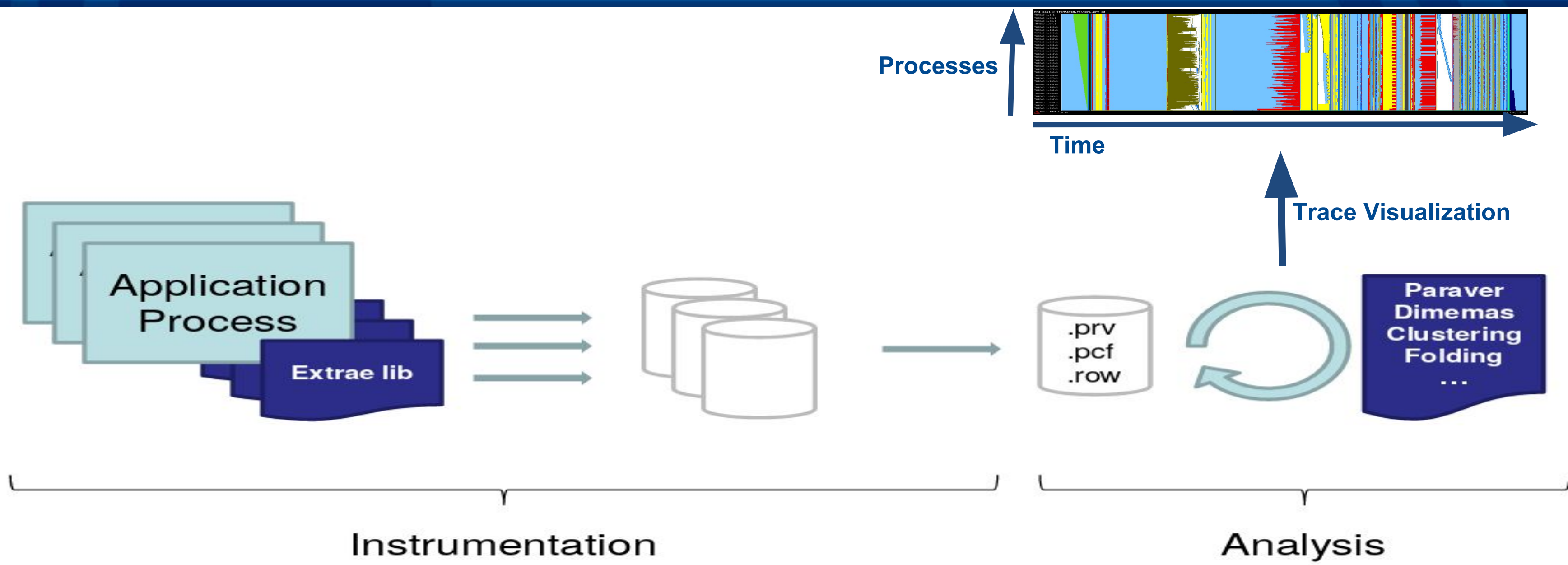
High Performance Computing (Services and Research) applied to Earth System Modelling

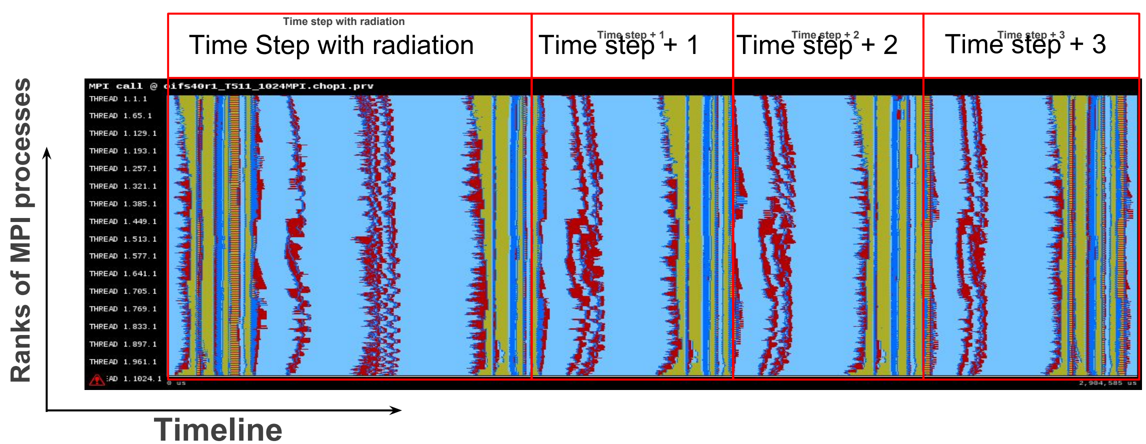
- Knowledge about the mathematical and computational side of Earth System Applications
- Knowledge about the specific needs in HPC of the Earth System Applications
- Researching about HPC methods specifically used for Earth System Applications

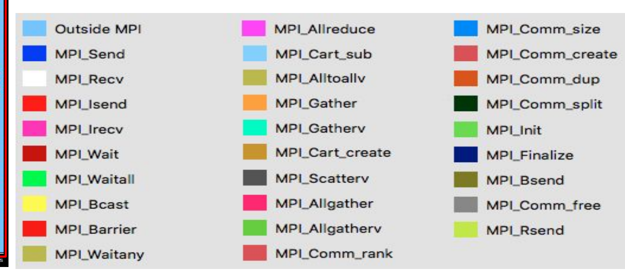
Computer Science



Profiling Analysis







Legend with the colors representation of the MPI functions

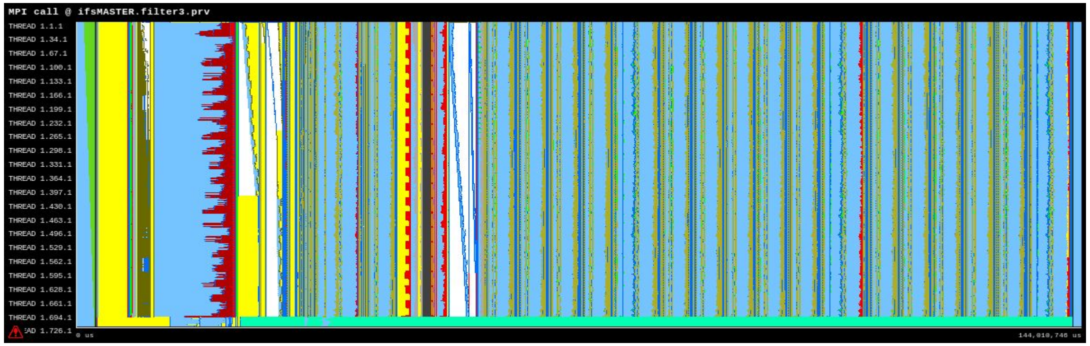
Base trace with MPI events for four time steps of IFS using 512 processes.

IFS Benchmarking

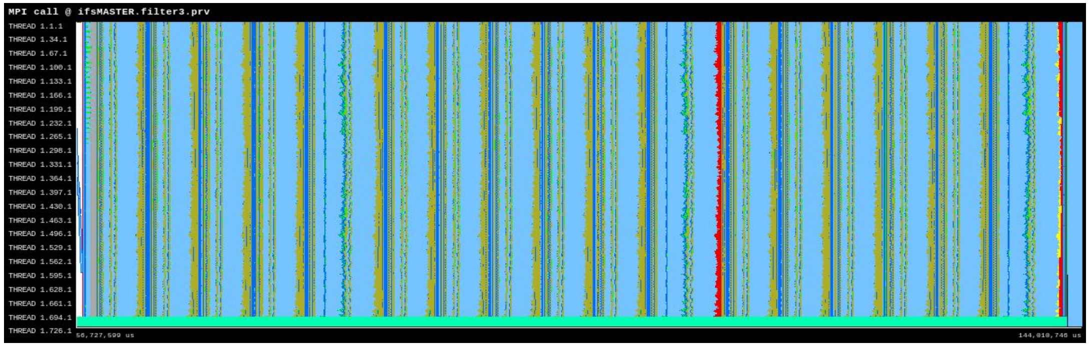
- IFSCY45R1, T1259
- Base Case: 702x6 (702 MPI processes, each one with 6 OpenMP threads)
- 24 hours of simulation (validated to one 10 days simulation)
- Transfer sensitivity, Dimemas Ideal network, Bandwidth $\longrightarrow \infty$
- Strong scaling tests
  - (48x1, 48x2, 48x4, 48x8, 48x9, 48x18)
  - (234x1, 468x1, 702x1, 1170x1)
- Hybrid implementation tests
  - (1403x3, 702x6, 468x9, 234x18, 117x36)

Localize the Study Area

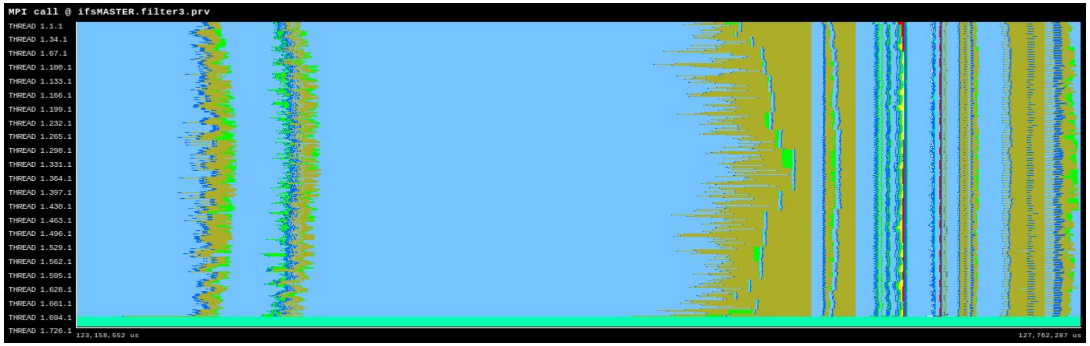
- General MPI profile+Histogram → localize your study area



Complete execution

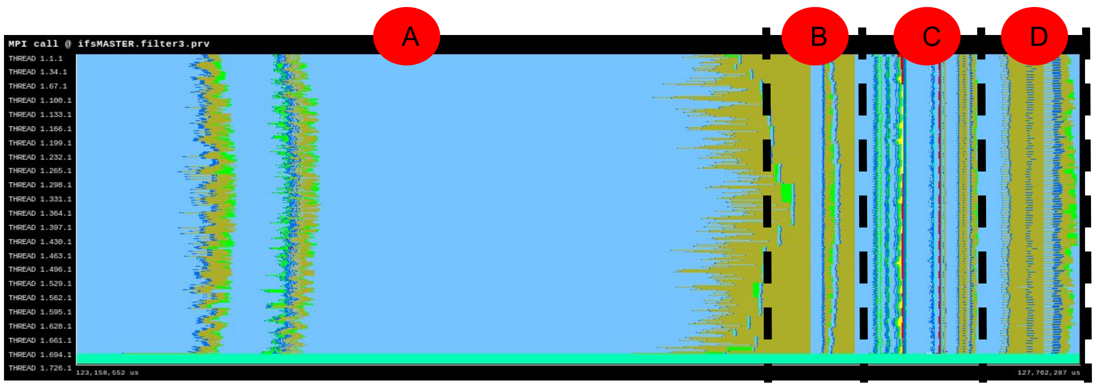


Some time steps



One time step

Localize the Study Area

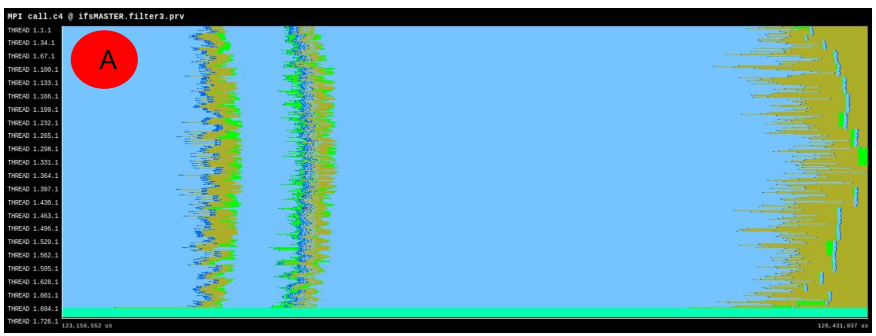


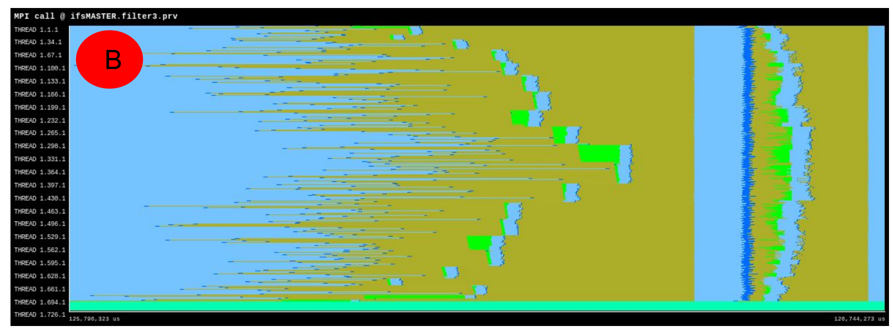
A Grid Point Calculations

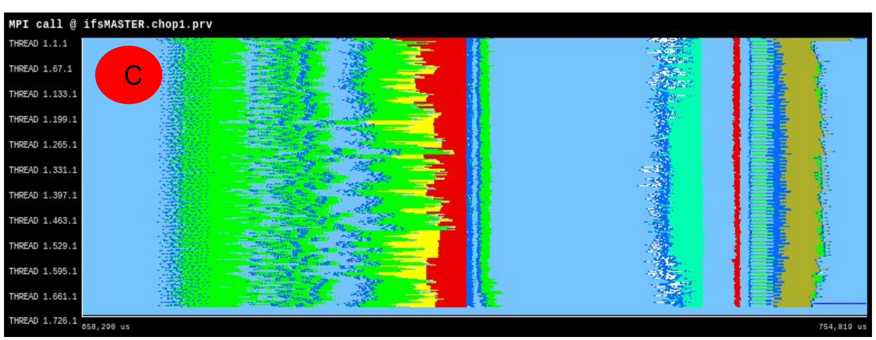
B Transformations and Transpositions (Fourier + Legendre)

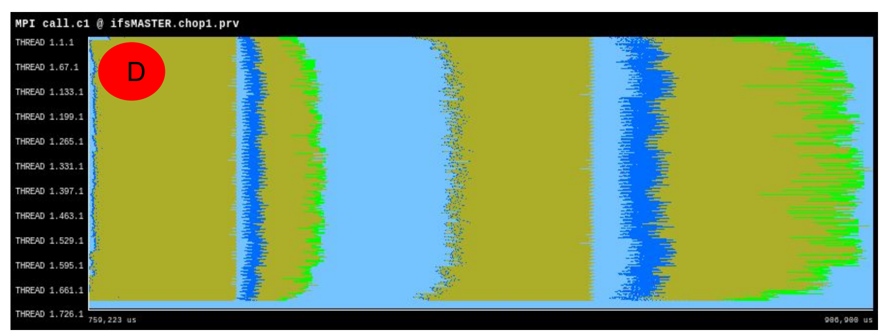
C Spectral Calculations

D Fourier + Legendre Inverse

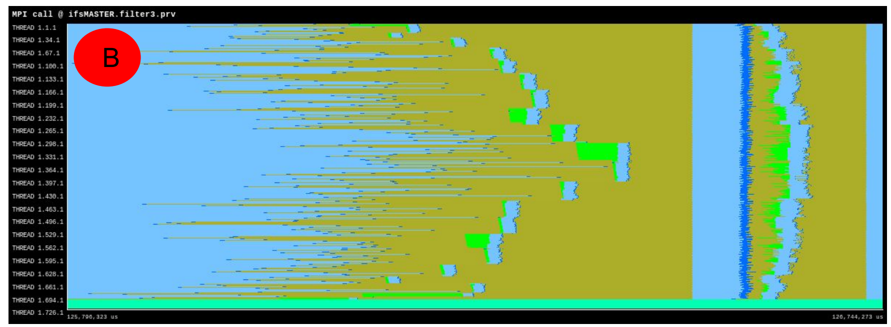








I can see your real form!



IFS Profiling Analysis

- Parallel and Communication efficiency, Global load balance → less than 85%?

Parallel Efficiency

MPI profiling

	Outside MPI	MPI_Send	MPI_Recv	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Barrier	MPI_Alltoallv	MPI_Gatherv	MPI_Comm_rank	MPI_Comm_size	MPI_Bsend	MPI_Waitany
Total	66,578.44 %	1.71 %	773.76 %	646.21 %	239.35 %	12,362.37 %	806.93 %	10,757.31 %	35.56 %	2.49 %	448.23 %	0.81 %	7,746.82 %
Average	66.31 %	0.00 %	0.77 %	0.64 %	0.24 %	12.31 %	0.80 %	10.71 %	0.04 %	0.00 %	0.45 %	0.81 %	7.72 %
Maximum	72.93 %	0.01 %	2.98 %	1.60 %	0.80 %	18.56 %	1.84 %	25.06 %	1.12 %	0.01 %	1.88 %	0.81 %	19.25 %
Minimum	57.05 %	0.00 %	0.01 %	0.08 %	0.07 %	3.11 %	0.00 %	5.25 %	0.00 %	0.00 %	0.16 %	0.81 %	0.31 %
StDev	2.03 %	0.00 %	0.57 %	0.36 %	0.06 %	2.52 %	0.41 %	3.57 %	0.12 %	0.00 %	0.10 %	0 %	3.18 %
Avg/Max	0.91	0.31	0.26	0.40	0.30	0.66	0.44	0.43	0.03	0.34	0.24	1	0.40

MPI_Send

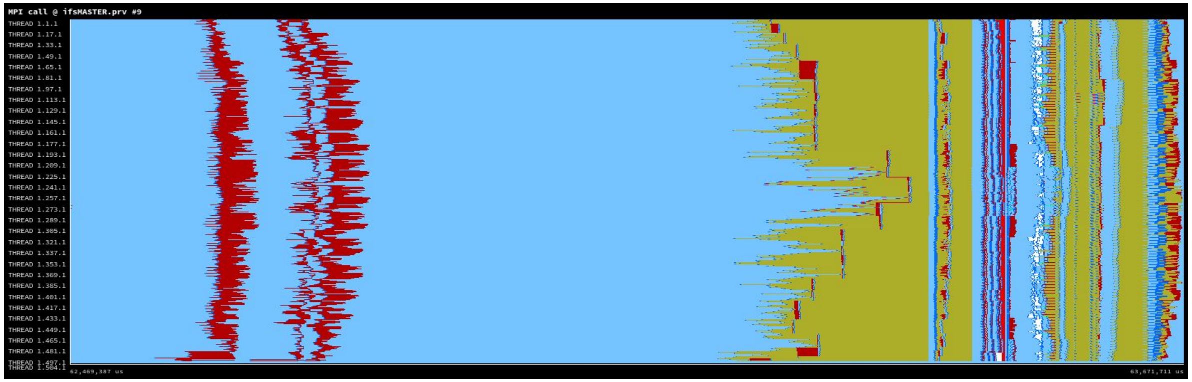
Communication Efficiency

Global Load Balance

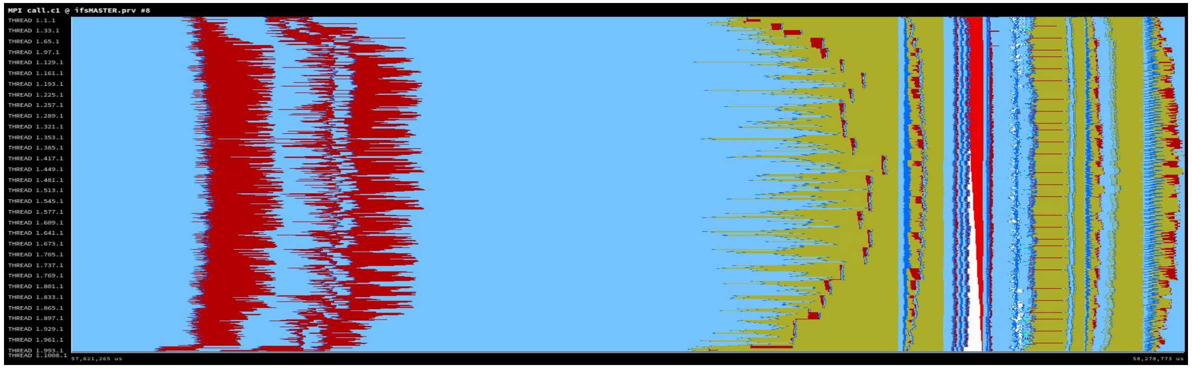
IFS profiling analysis

468 MPI processes

702 MPI processes



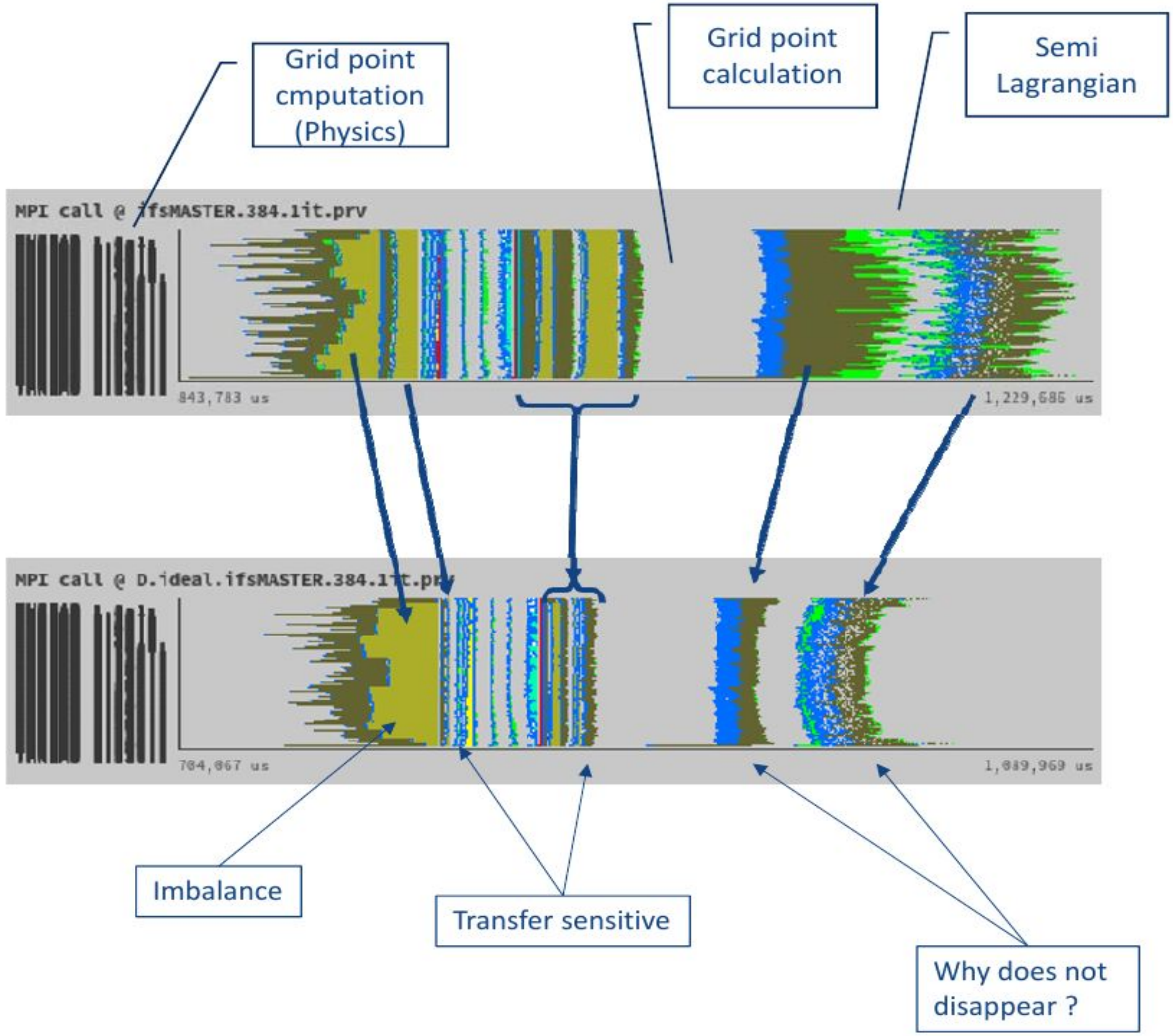
1070 MPI processes



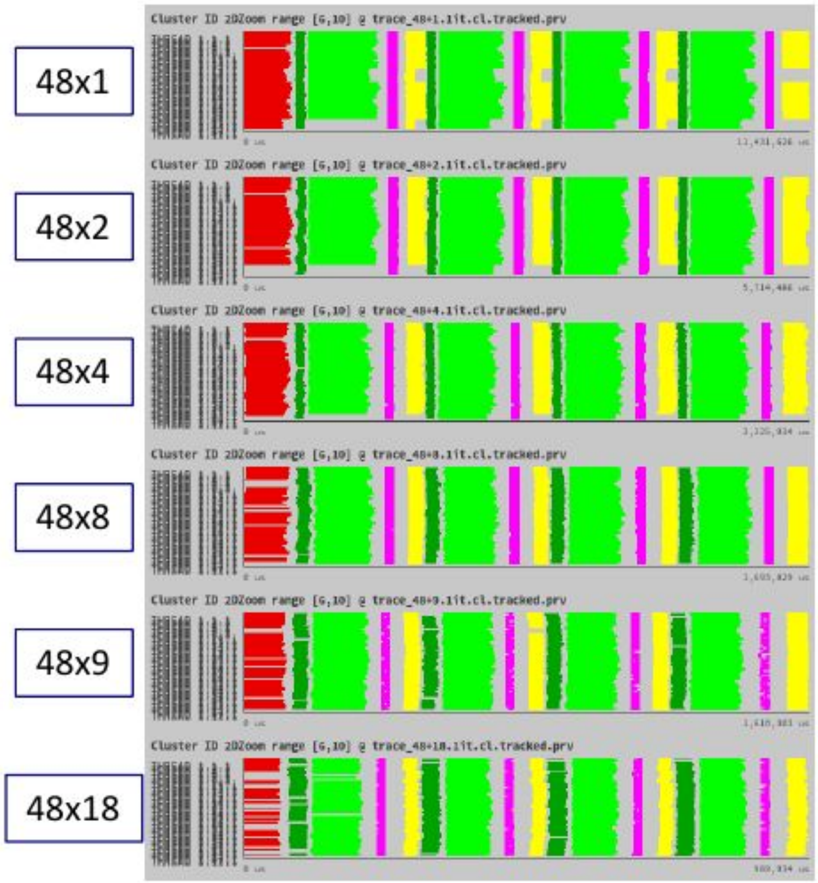
IFS Profiling Analysis

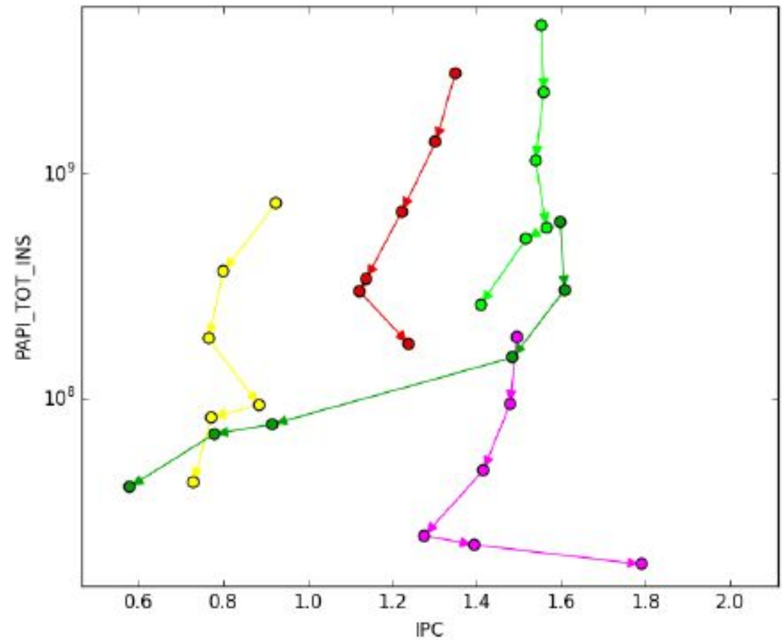
What if

- Actual run
- Ideal network

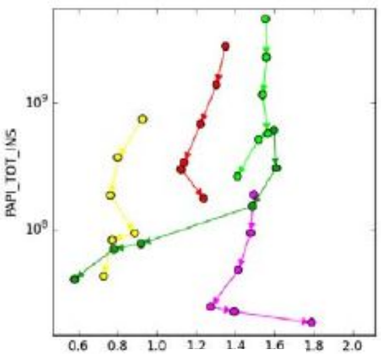


Tracking MPI+OMP Strong Scaling

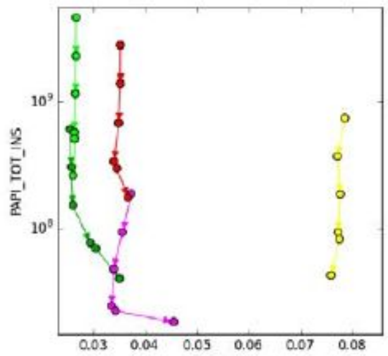




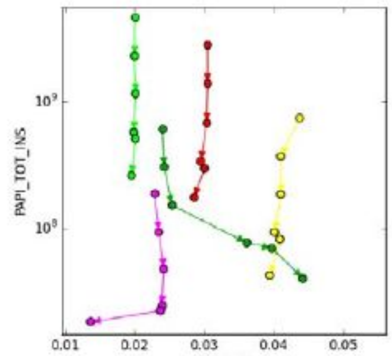
Tracking MPI+OMP Strong Scaling



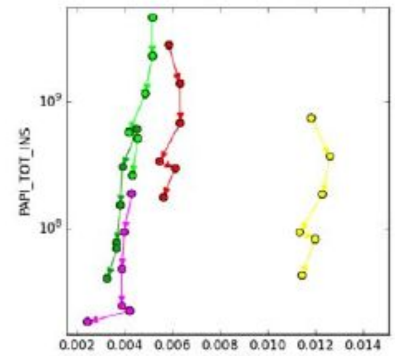
IPC



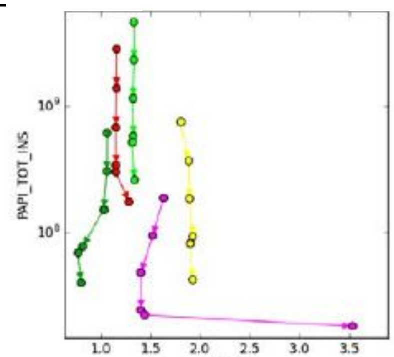
L1_MPKI



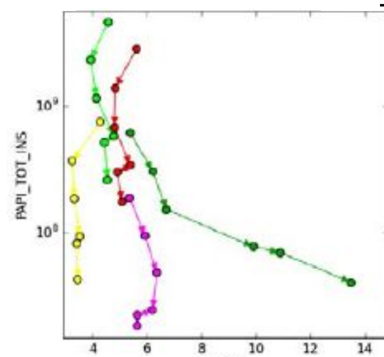
L2_MPKI



L3_MPKI



L1L2



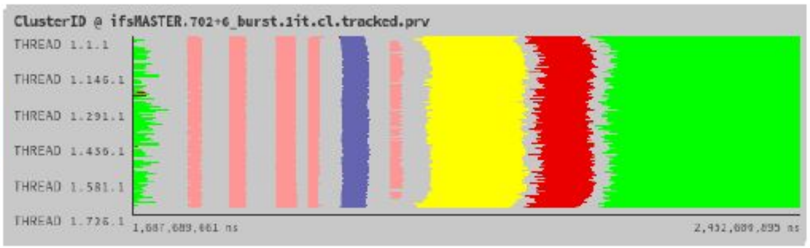
L2L3

IPC: Instructions per Cycle
MPKI: Misses per 1000 Instructions

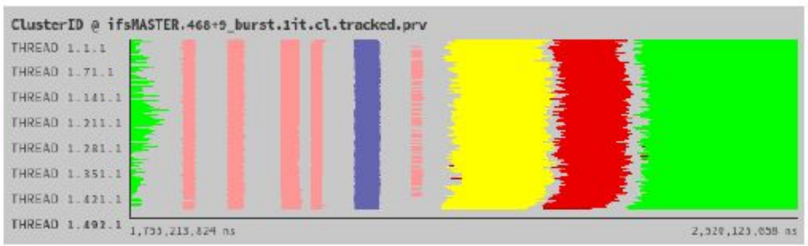
How much Hybrid?

- Fix cores = 4212 → 1402x3, 702x6, 468x9, 234x18 ?

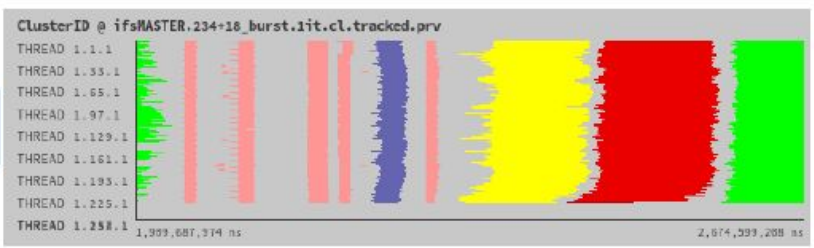
702 x 6

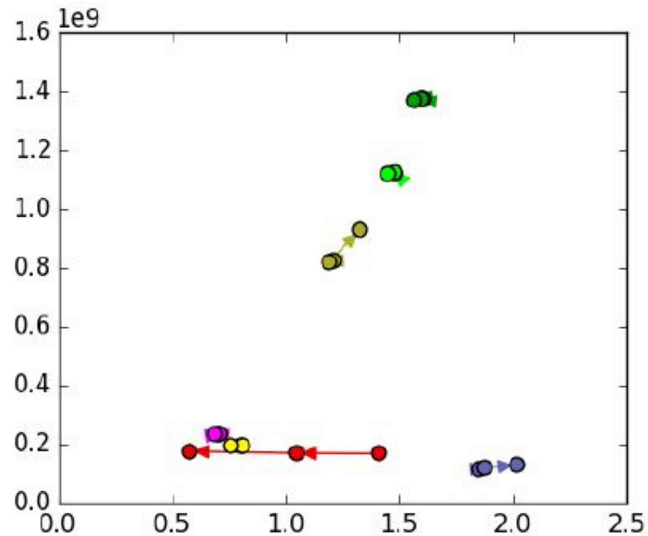


468 x 9

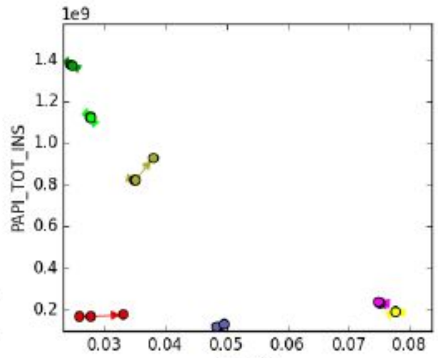


234 x 18

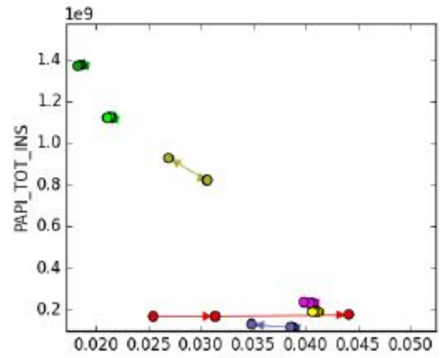




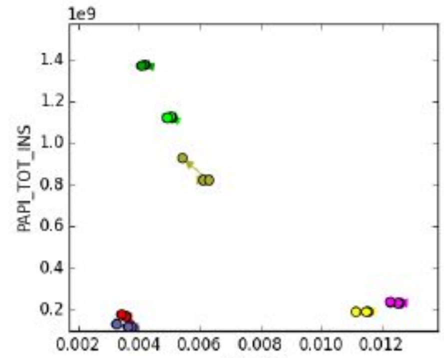
How much Hybrid?



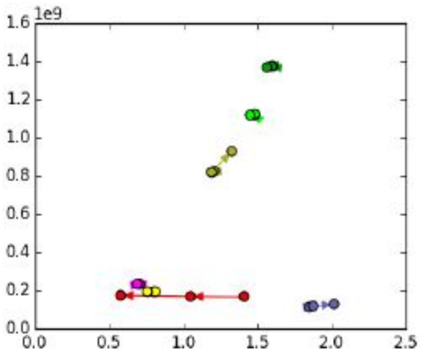
L1_MPKI



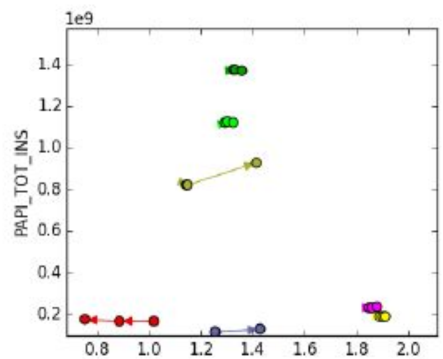
L2_MPKI



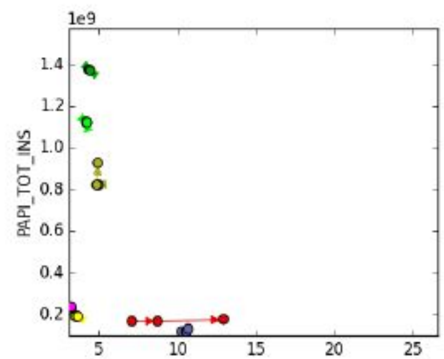
L3_MPKI



IPC



L1L2



L2L3

IPC: Instructions per Cycle
MPKI: Misses per 1000 Instructions

Computational Efficiency

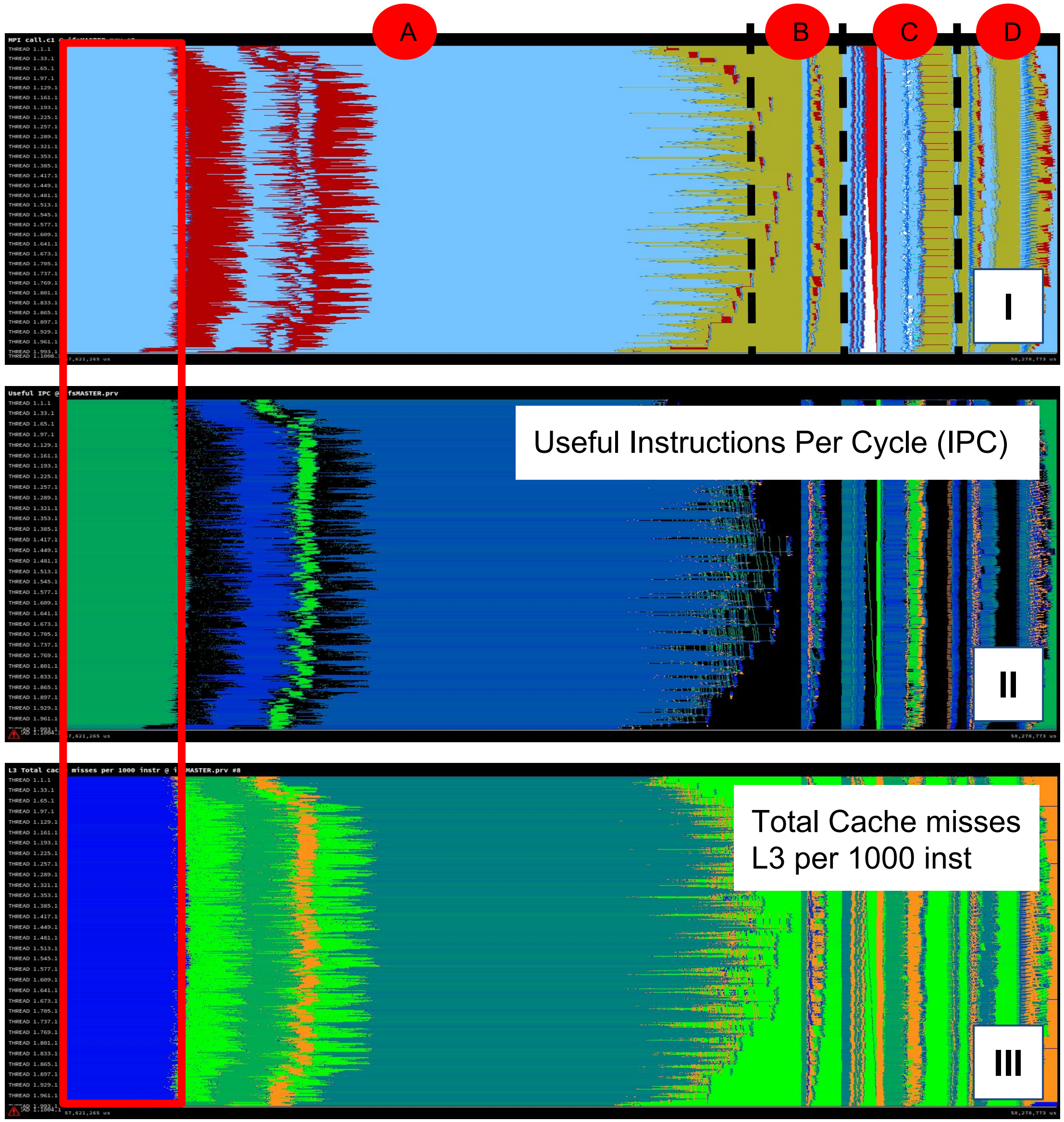
IFS: MPI Events

IPC< 1

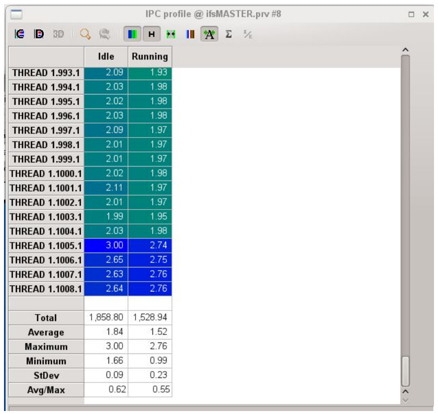
IPC> 1

Total Cache misses < 100

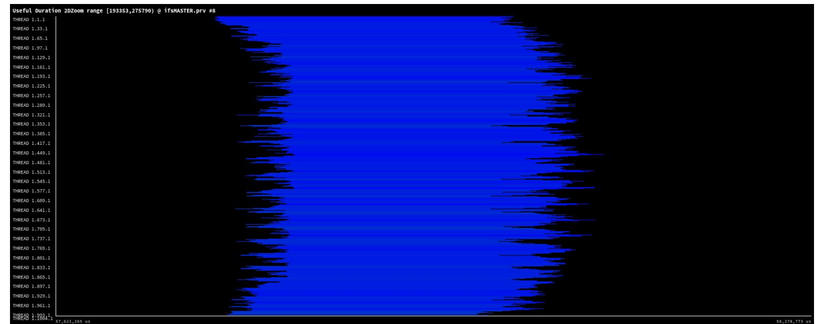
Total Cache misses > 100

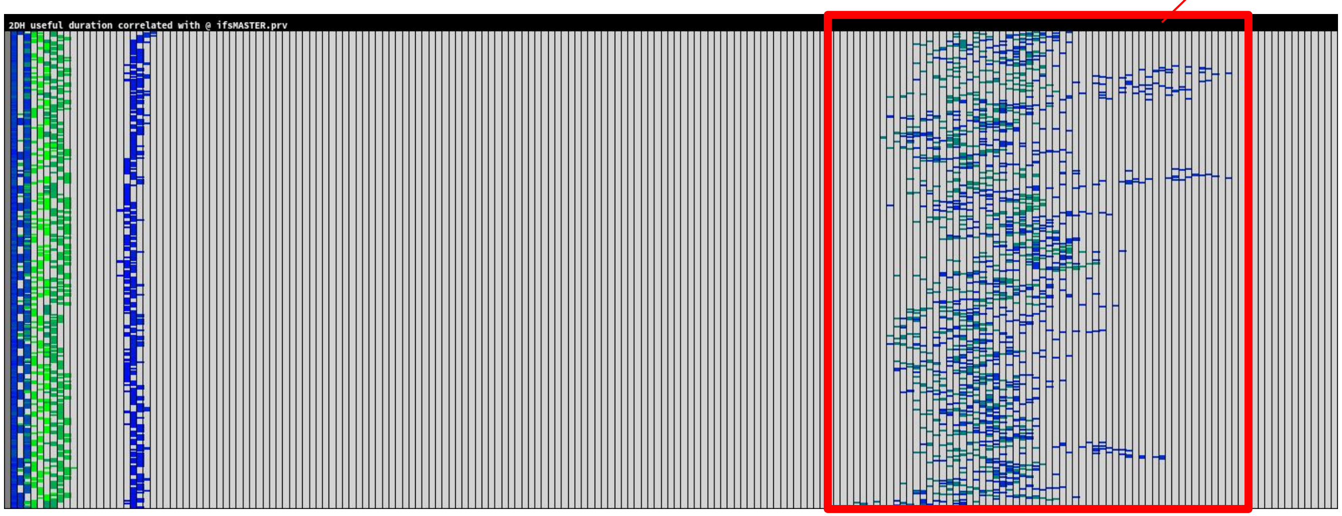


Load imbalance physical calculations


IPC profile @ ifsMASTER.prv #8

	Idle	Running
THREAD 1.993.1	2.09	1.93
THREAD 1.994.1	2.03	1.98
THREAD 1.995.1	2.02	1.98
THREAD 1.996.1	2.03	1.98
THREAD 1.997.1	2.09	1.97
THREAD 1.998.1	2.01	1.97
THREAD 1.999.1	2.01	1.97
THREAD 1.1000.1	2.02	1.98
THREAD 1.1001.1	2.11	1.97
THREAD 1.1002.1	2.01	1.97
THREAD 1.1003.1	1.99	1.95
THREAD 1.1004.1	2.03	1.98
THREAD 1.1005.1	3.00	2.74
THREAD 1.1006.1	2.65	2.75
THREAD 1.1007.1	2.63	2.76
THREAD 1.1008.1	2.64	2.76
Total	1,858.80	1,528.94
Average	1.84	1.52
Maximum	3.00	2.76
Minimum	1.66	0.99
StDev	0.09	0.23
Avg/Max	0.62	0.55



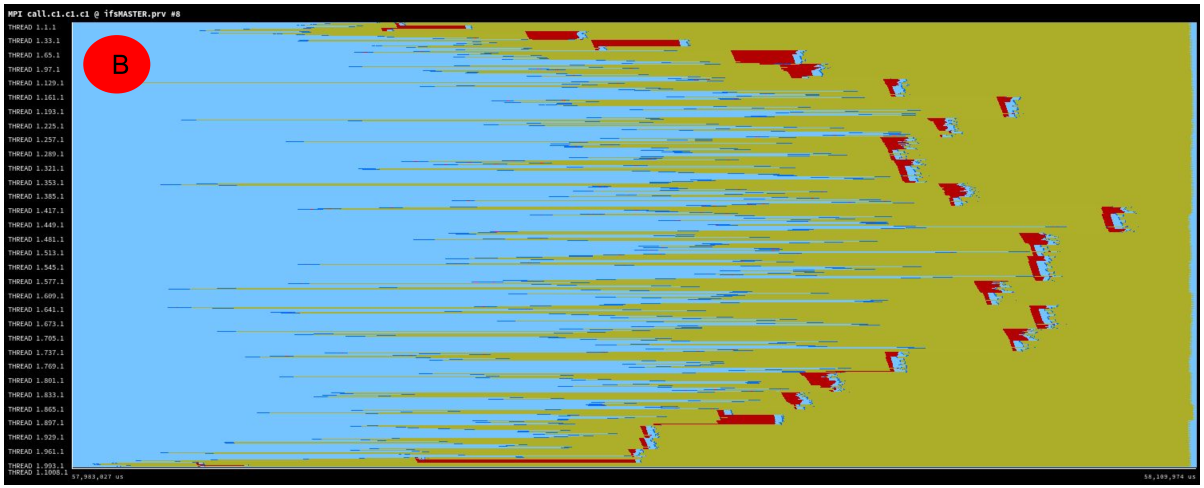


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Load imbalance physical calculations

B Transformations and Transpositions (Fourier + Legendre)



MPI call profile @ ifsMASTER.prv #8

	Outside MPI	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Alltoallv	MPI_Comm_size	MPI_Waitany
Total	51,509.08 %	426.77 %	182.48 %	1,521.25 %	26,180.80 %	338.91 %	20,640.69 %
Average	51.10 %	0.43 %	0.18 %	1.52 %	26.08 %	0.34 %	20.56 %
Maximum	100 %	1.09 %	0.72 %	19.46 %	87.77 %	0.84 %	74.09 %
Minimum	3.40 %	0.02 %	0.03 %	0.05 %	4.62 %	0.07 %	0.13 %
StDev	16.31 %	0.25 %	0.06 %	2.23 %	14.98 %	0.09 %	14.03 %
Avg/Max	0.51	0.39	0.25	0.08	0.30	0.40	0.28

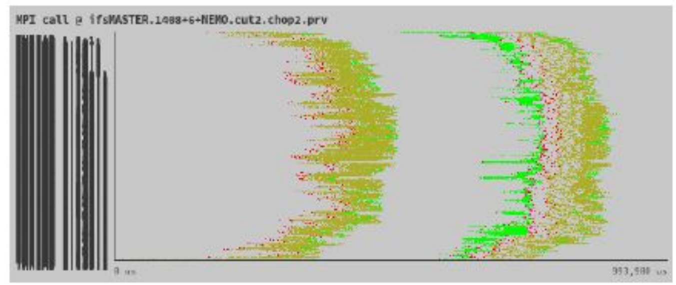
MPI_Send

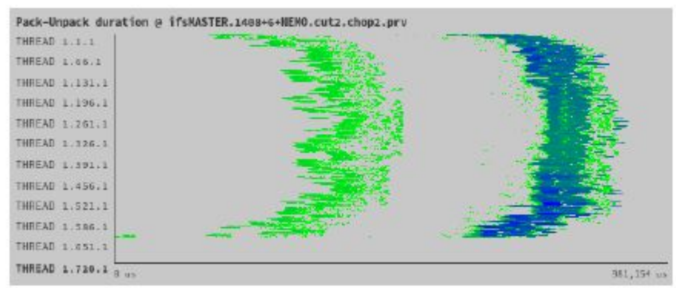
IFS Profiling Analysis

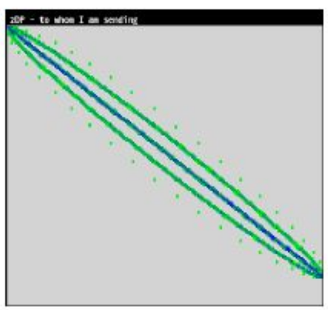
- MPI call sequence?
- Connectivity pattern?

```
for (){ irecv(); }
for (){  pack(); isend(); }
for (){  Waitany(); unpack(); }
Waitall();
```

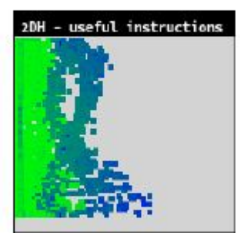
- Questions/ potential considerations:
  - Is packing/unpacking parallelized?
    - fork-join? Taskified?
  - Would it be possible? How far could these ops be overlapped?
  - Would it be worthwhile? How much concurrency to use?



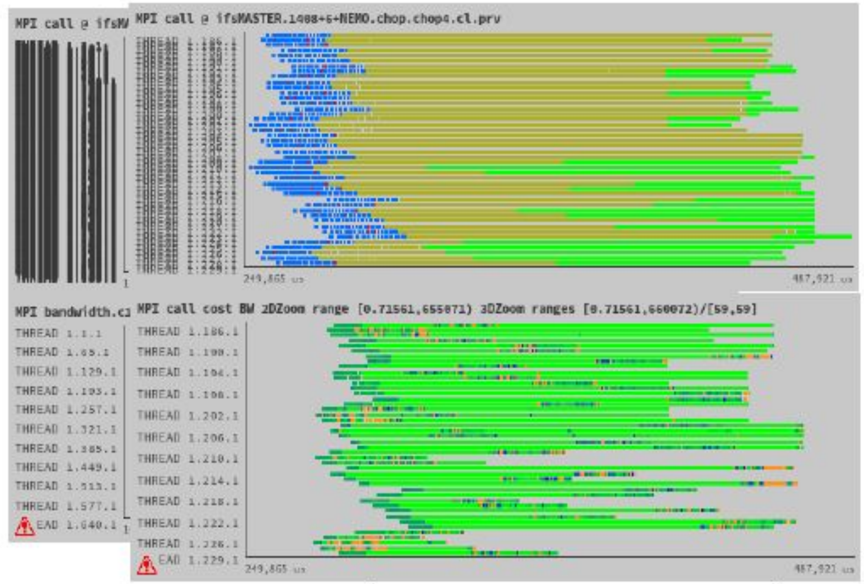




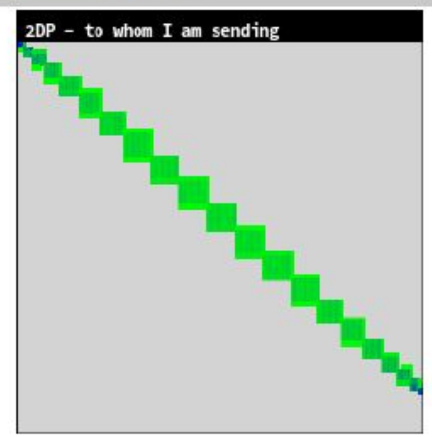
Pack-unpack time
up to 25 ms burst
up to 50 ms aggregated
IPC < 0.3



IFS Profiling Analysis



- MPI call bandwidth
  - bytes/time of the call
- Pattern
  - Some long calls
  - Synchronized ?
- Connectivity pattern?
- Questions/ potential considerations:
  - Is packing/unpacking parallelized?
  - Can reuse non productive MPI time?
    - long waitanys?
    - Waitall for sends?
  - Communication pattern issues ?
    - Endpoint contention?



Transfer to neighbours in Rank order

IFS Profiling Analysis

Big number of comm_size calls and big number of isend/recv,irecv+wait_any calls

MPI call profile @ ifsMASTER.prv #8

	Outside MPI	MPI_Send	MPI_Recv	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Barrier	MPI_Alltoallv	MPI_Gatherv	MPI_Comm_rank	MPI_Comm_size	MPI_Bsend	MPI_Waitany
Total	1,720,717	1,003	19,733	285,726	267,999	329,555	1,004	4,016	2,008	2,008	581,488	1,003	224,170
Average	1,713.86	1	19.65	284.59	266.93	328.24	1	4	2	2	579.17	1,003	223.28
Maximum	6,823	1	1,379	413	566	654	1	4	2	2	3,069	1,003	379
Minimum	467	1	3	76	67	116	1	4	2	2	160	1,003	27
StDev	406.04	0	43.46	61.10	64.21	60.95	0	0	0	0	147.07	0	64.66
Avg/Max	0.25	1	0.01	0.69	0.47	0.50	1	1	1	1	0.19	1	0.59

MPI_Send

Conclusions and possible Optimizations

- Physical computation load imbalance
  - Produced by the different quantity of work depending on the latitude assigned in the domain decomposition
  - Increase the overhead needed for MPI communications during the direct transformation/transpositions
  - Dynamical Load Balance (DLB) using OpenMP tasks could solve the problem
  - Overlapping during the computation creating blocks of latitudes inside one subdomain, these blocks could overlap physical calculations and transformation/transpositions.

Conclusions and possible Optimizations

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

- Semi-Lagrangian stage computation
  - Low IPC (less than one)
  - Unefficient methodology for the MPI communication among neighbors
  - Dynamical Load Balance (DLB) using OpenMP tasks could solve the problem
  - Improving the possible pack/unpack operations before/after sending

Conclusions and possible Optimizations

- Grid-Point Computation
  - Low IPC (less than one) in an area of only computation
  - High rate of memory access fails
  - Evaluate how the locality of the variables in this computation stage and improve it

- Other minor issues
  - Big number of comm_size calls
  - Redundant operations for asynchronal communications
  - Variables containing some information per MPI rank should solve the problem

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you!

mario.acosta@bsc.es

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you!

mario.acosta@bsc.es