
Barcelona Supercomputing Center
Centro Nacional de Supercomputación


# BSC tools to study the computational efficiency of EC-Earth components

Miguel Castrillo, Oriol Tintó, Kim Serradell

# Objectives

# EC-Earth



- Earth System Model
- Reliable in-house predictions of global climate change
- Part of a Europe-wide consortium
- Being used in large European projects
  - EMBRACE
  - EUPORIAS
  - IS-ENES
  - SPECS
- 3.1 version → IFS + NEMO-LIM + OASIS

# Energy Efficiency

- Energy efficiency

$$EE = \frac{Performance}{Power\ consumed}$$

Increase performance

Reduce power

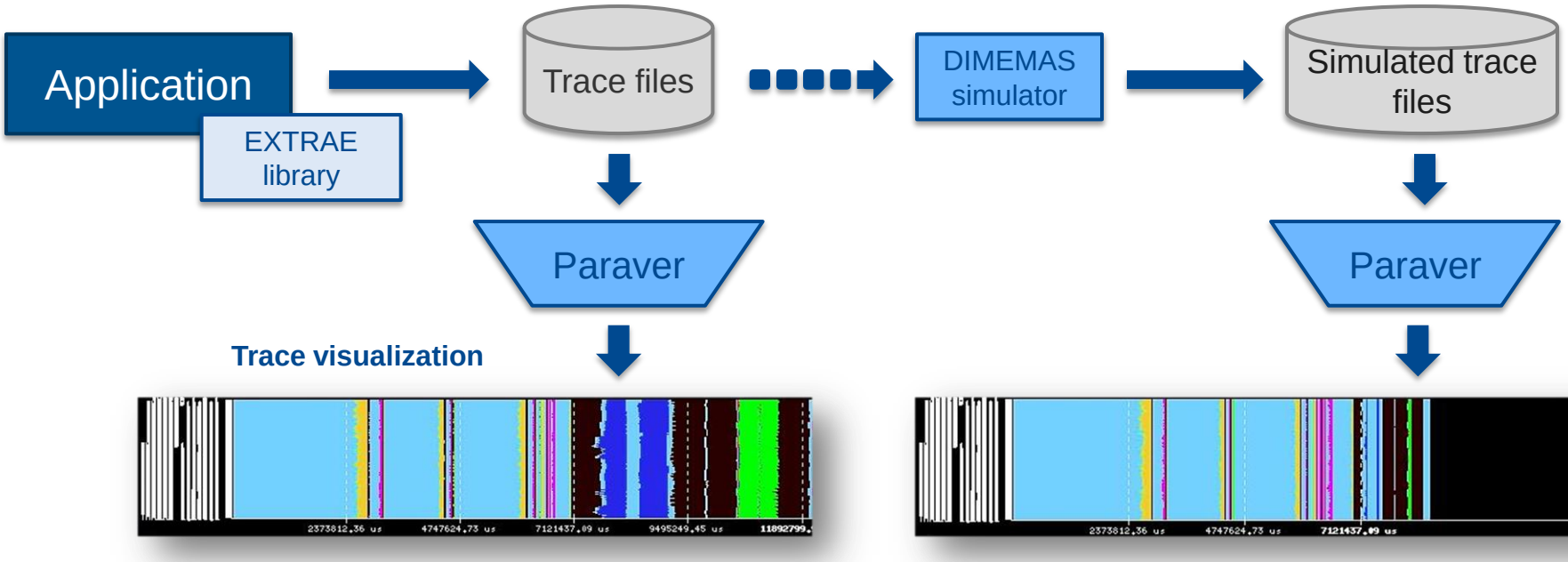
- Performance loss caused by:
  - Bad programming
  - Load imbalance
  - Synchronization
  - Resource contention
  - …


Barcelona Supercomputing Center
Centro Nacional de Supercomputación


# Methodology

# BSC Computer Sciences Performance Tools

- Since 1991
- Based on traces
- Open Source: http://www.bsc.es/paraver
- **Extrae**: Package that generates Paraver trace-files for a post-mortem analysis
- **Paraver**: Trace visualization and analysis browser
  - Includes trace manipulation: Filter, cut traces
- **Dimemas**: Message passing simulator




DIMEMAS generated trace. Target = ideal machine

# Extrae



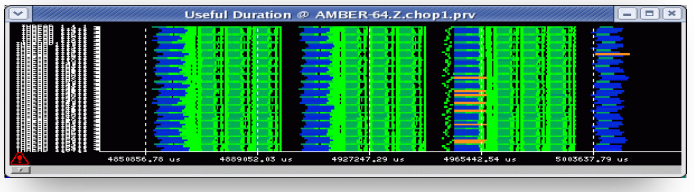
- BSC instrumentation package

- When/Where
  - Parallel programming model runtime
  - Selected user functions
  - Periodic samples
  - User events

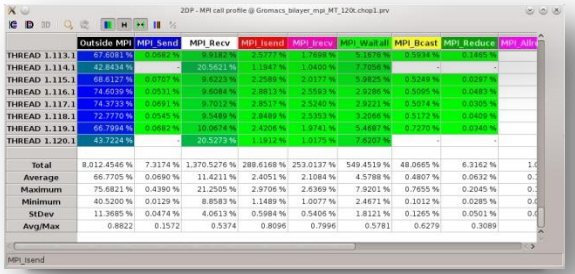
- Additional information
  - Counters

# Paraver

- Every behavioral aspect/metric of a thread can be described as a function of time
- Those functions of time can be rendered into a 2D image




- Statistics can be computed for each possible value or range of values of that function of time
  - Tables: Profiles and histograms



- Types of functions
  - Categorical: State, user function…
  - Logical: In specific user function, in MPI call…
  - Numerical: IPC, L2 miss ratio, duration of computation burst…

# PARAVER trace analysis

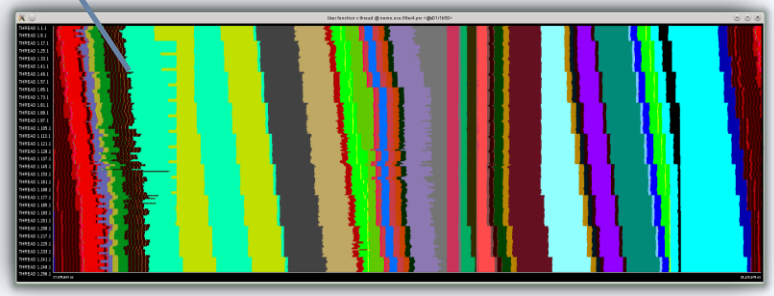
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

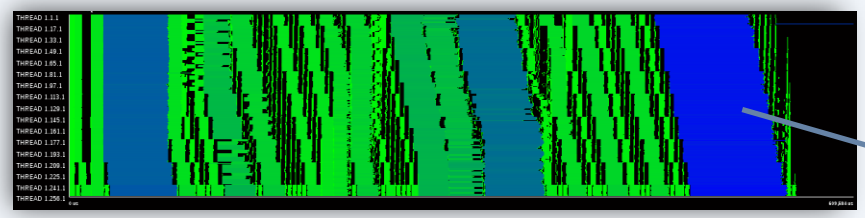
## Serial efficiency

Each color represents a function

### Functions instrumentation

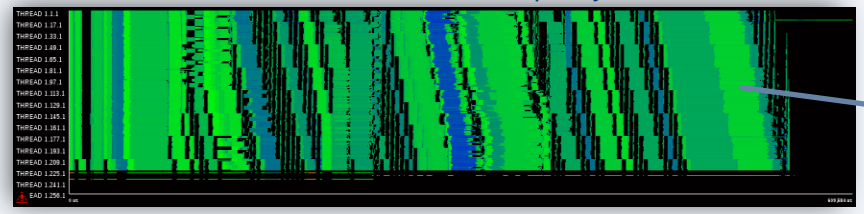


Useful Duration



Darker color = Longer duration

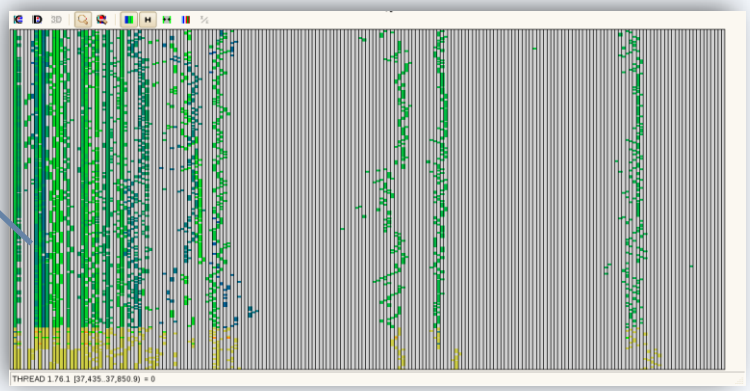
Useful IPC - Instructions per cycle



Lighter color = Less IPC

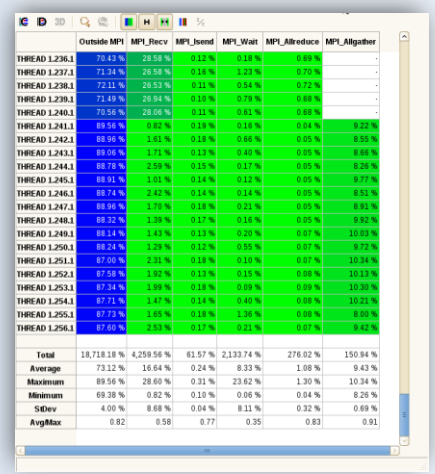
Horizontal axis -> Time component

### Correlation between two functions



Darker color = Higher values

## Parallel efficiency



| | Outside MPI | MPI_Recv | MPI_Isend | MPI_Wait | MPI_Allreduce | MPI_Allgather |
|---|---|---|---|---|---|---|
| THREAD 1.236.1 | 70.43 % | 28.58 % | 0.12 % | 0.18 % | 0.69 % | - |
| THREAD 1.237.1 | 71.34 % | 26.58 % | 0.16 % | 1.23 % | 0.70 % | - |
| THREAD 1.238.1 | 72.11 % | 26.53 % | 0.11 % | 0.54 % | 0.72 % | - |
| THREAD 1.239.1 | 71.49 % | 26.94 % | 0.18 % | 0.79 % | 0.68 % | - |
| THREAD 1.240.1 | 70.56 % | 28.06 % | 0.11 % | 0.61 % | 0.68 % | - |
| THREAD 1.241.1 | 89.56 % | 0.82 % | 0.19 % | 0.16 % | 0.04 % | 9.22 % |
| THREAD 1.242.1 | 88.96 % | 1.61 % | 0.18 % | 0.66 % | 0.05 % | 8.55 % |
| THREAD 1.243.1 | 89.06 % | 1.71 % | 0.13 % | 0.40 % | 0.05 % | 8.66 % |
| THREAD 1.244.1 | 88.78 % | 2.59 % | 0.15 % | 0.17 % | 0.05 % | 8.26 % |
| THREAD 1.245.1 | 88.91 % | 1.01 % | 0.14 % | 0.12 % | 0.05 % | 9.77 % |
| THREAD 1.246.1 | 88.74 % | 2.42 % | 0.14 % | 0.14 % | 0.05 % | 8.51 % |
| THREAD 1.247.1 | 88.96 % | 1.70 % | 0.18 % | 0.21 % | 0.05 % | 8.91 % |
| THREAD 1.248.1 | 88.32 % | 1.39 % | 0.17 % | 0.16 % | 0.05 % | 9.92 % |
| THREAD 1.249.1 | 88.14 % | 1.43 % | 0.13 % | 0.20 % | 0.07 % | 10.03 % |
| THREAD 1.250.1 | 88.24 % | 1.29 % | 0.12 % | 0.55 % | 0.07 % | 9.72 % |
| THREAD 1.251.1 | 87.00 % | 2.31 % | 0.18 % | 0.10 % | 0.07 % | 10.34 % |
| THREAD 1.252.1 | 87.58 % | 1.92 % | 0.13 % | 0.15 % | 0.08 % | 10.13 % |
| THREAD 1.253.1 | 87.34 % | 1.99 % | 0.18 % | 0.09 % | 0.09 % | 10.30 % |
| THREAD 1.254.1 | 87.71 % | 1.47 % | 0.14 % | 0.40 % | 0.08 % | 10.21 % |
| THREAD 1.255.1 | 87.73 % | 1.65 % | 0.18 % | 1.36 % | 0.08 % | 8.00 % |
| THREAD 1.256.1 | 87.60 % | 2.53 % | 0.17 % | 0.21 % | 0.07 % | 9.42 % |
| Total | 18,718.18 % | 4,259.56 % | 61.57 % | 2,133.74 % | 276.02 % | 150.94 % |
| Average | 73.12 % | 16.64 % | 0.24 % | 8.33 % | 1.08 % | 9.43 % |
| Maximum | 89.56 % | 28.60 % | 0.31 % | 23.62 % | 1.30 % | 10.34 % |
| Minimum | 69.38 % | 0.82 % | 0.10 % | 0.06 % | 0.04 % | 8.26 % |
| StDev | 4.00 % | 8.68 % | 0.04 % | 8.11 % | 0.32 % | 0.69 % |
| Avg/Max | 0.82 | 0.58 | 0.77 | 0.35 | 0.83 | 0.91 |

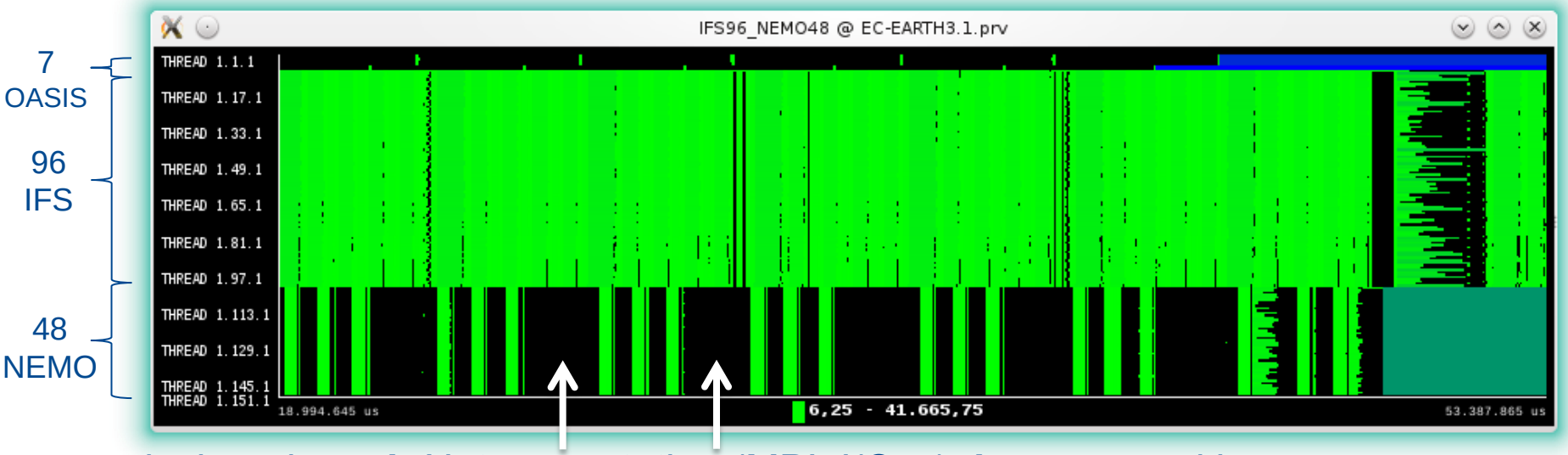
MPI Stats reflect the percentage of time invested in computation for each thread.

Total stats give the communication efficiency and the load balance

* Examples from NEMO 3.4–LIM2 / ORCA025

# An EC-Earth Paraver trace

- Motivation: Finding a good configuration to optimize the resources usage.
- IFS T255L91-ORCA1L46
- Configuration used in production
  - Using 7 cores for OASIS, 96 for IFS and 48 for NEMO
- 1 day simulation traces
- Traces generated in burst mode (only computational regions > 100us)
- Paraver view → Useful duration (displays duration of computational bursts)



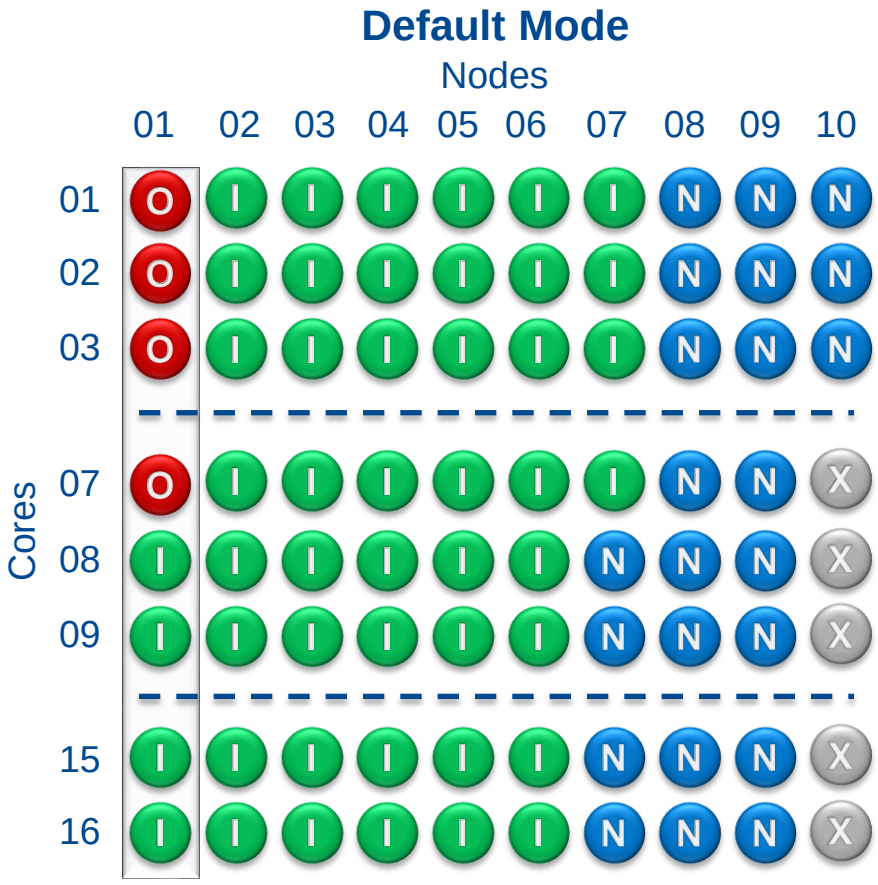

Black regions → Not computation (MPI, I/O…) → NEMO waiting

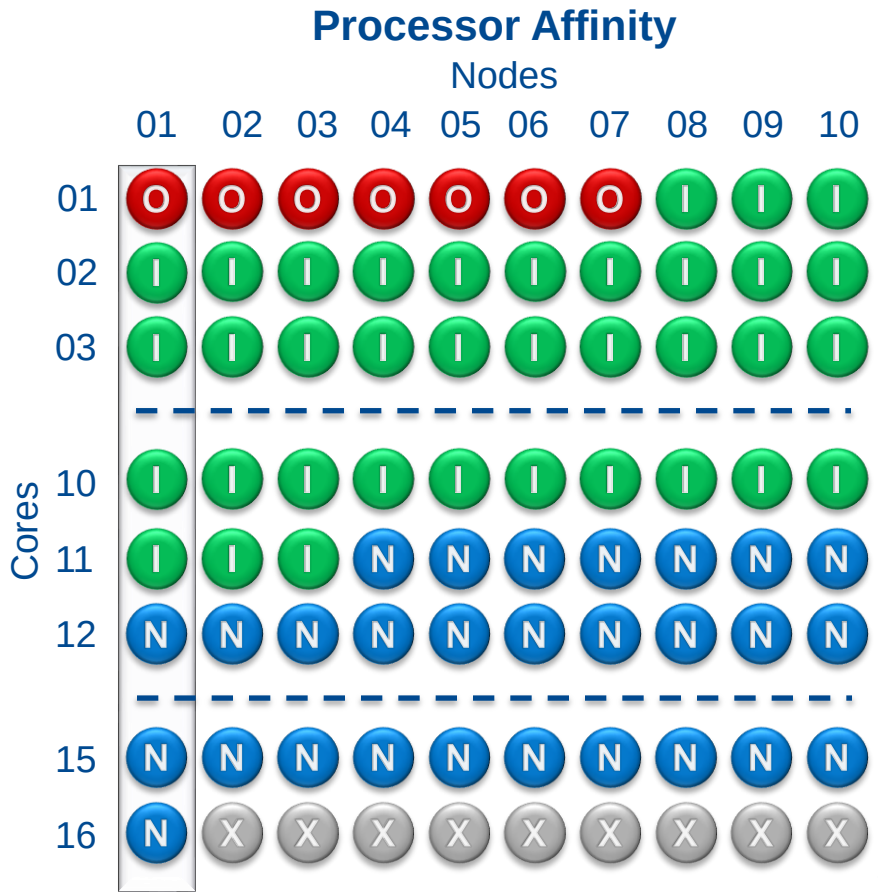
*Time axis*

# Dealing with resource contention

- Resource contention: Conflict over access to shared resources, i.e. memory, storage, buses…
- Distributing the processes among different nodes to mitigate the problem.

O 7 Oasis + I 96 IFS + N 48 NEMO in 10 nodes (16 cores per node)



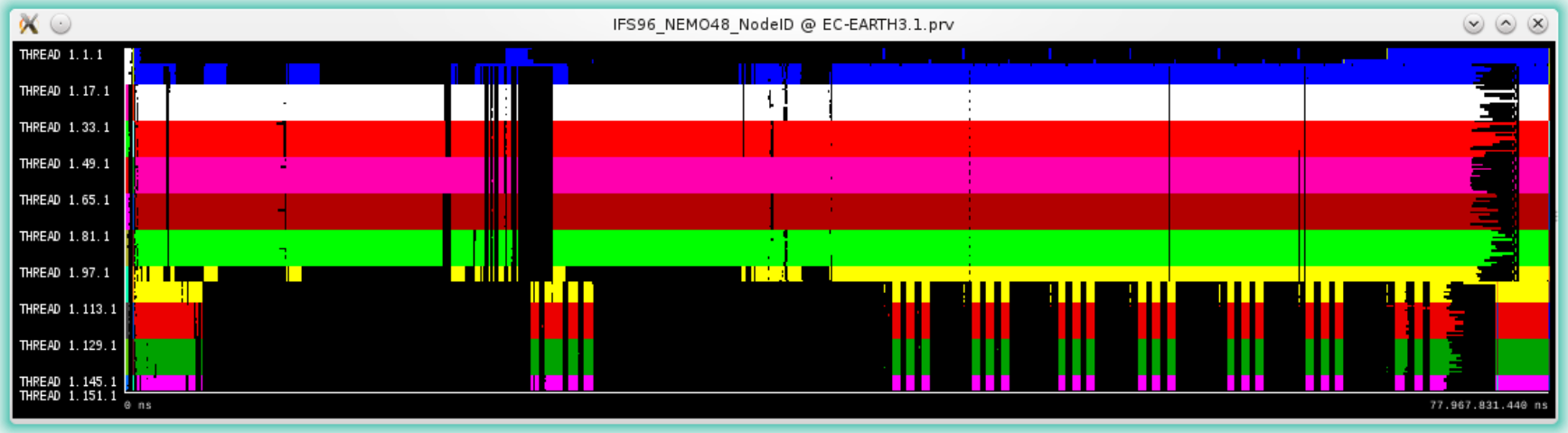



# Dealing with resource contention

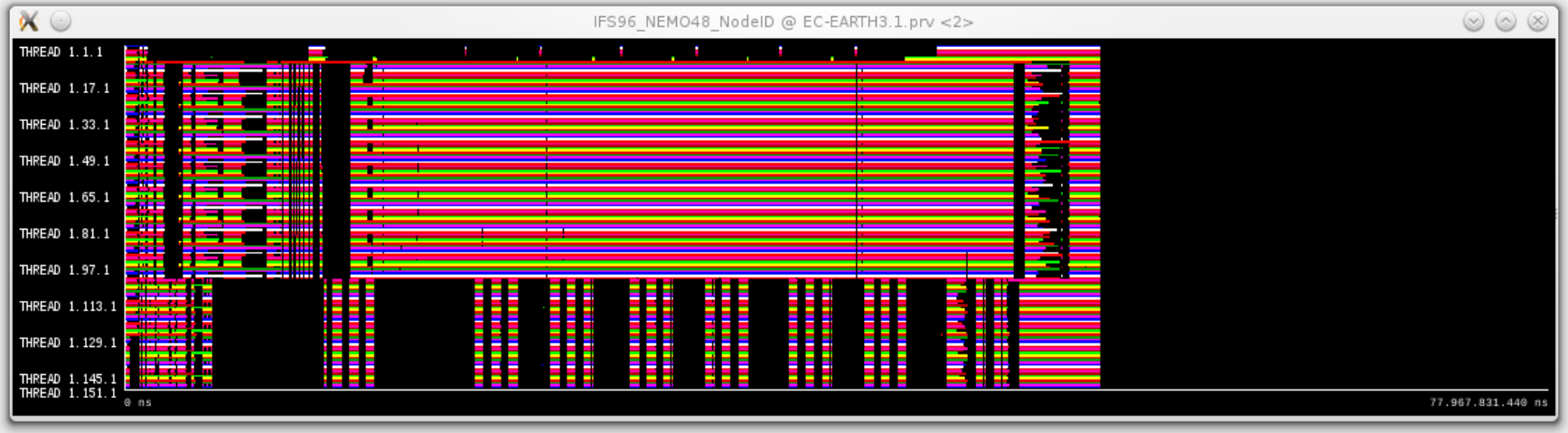


- Color identifies the different nodes

**Default Mode**

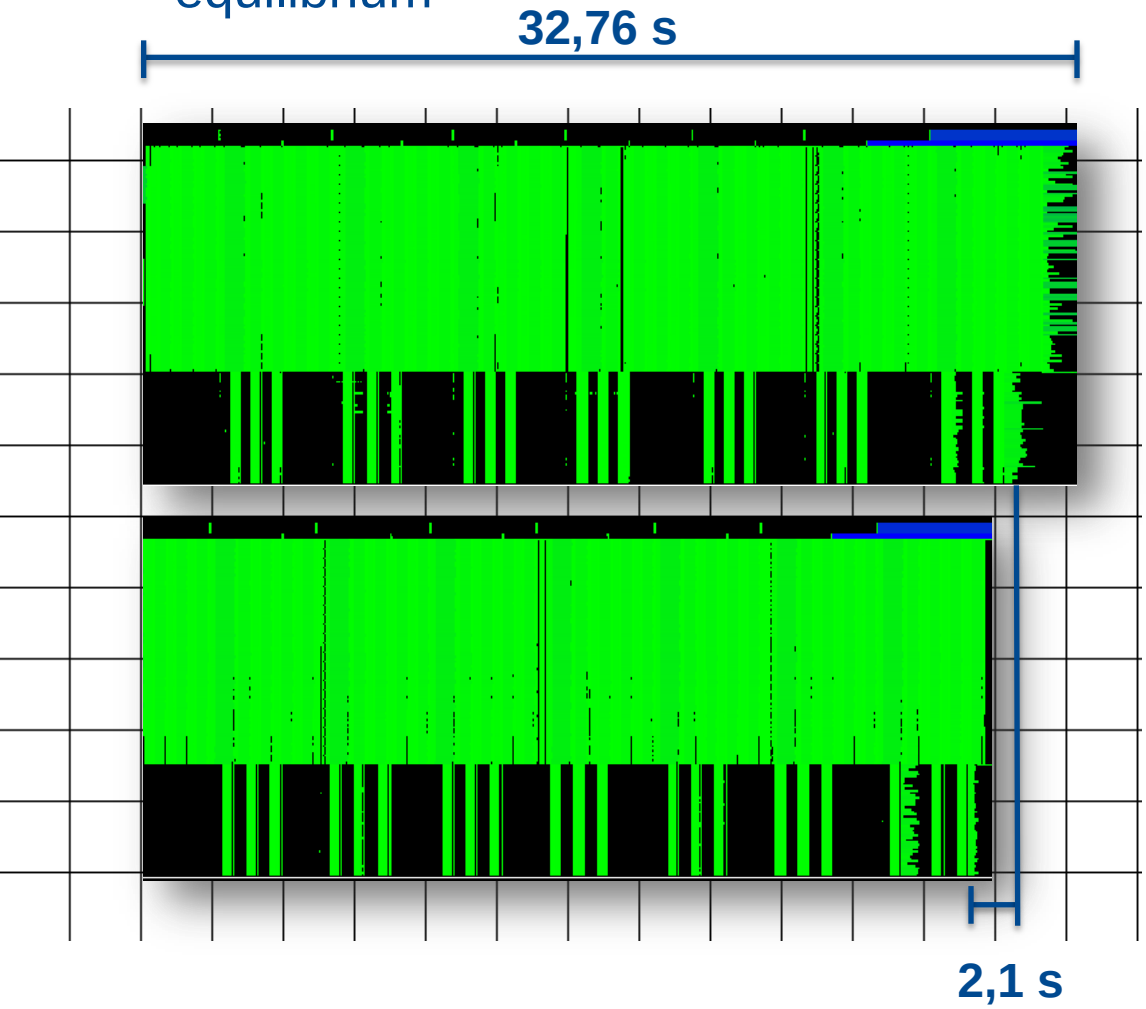



**Processor Affinity**



# Processor affinity effect

- There is an improvement, but NEMO is still waiting a lot for IFS
- Next step: Increase IFS and reduce NEMO resources to find an equilibrium

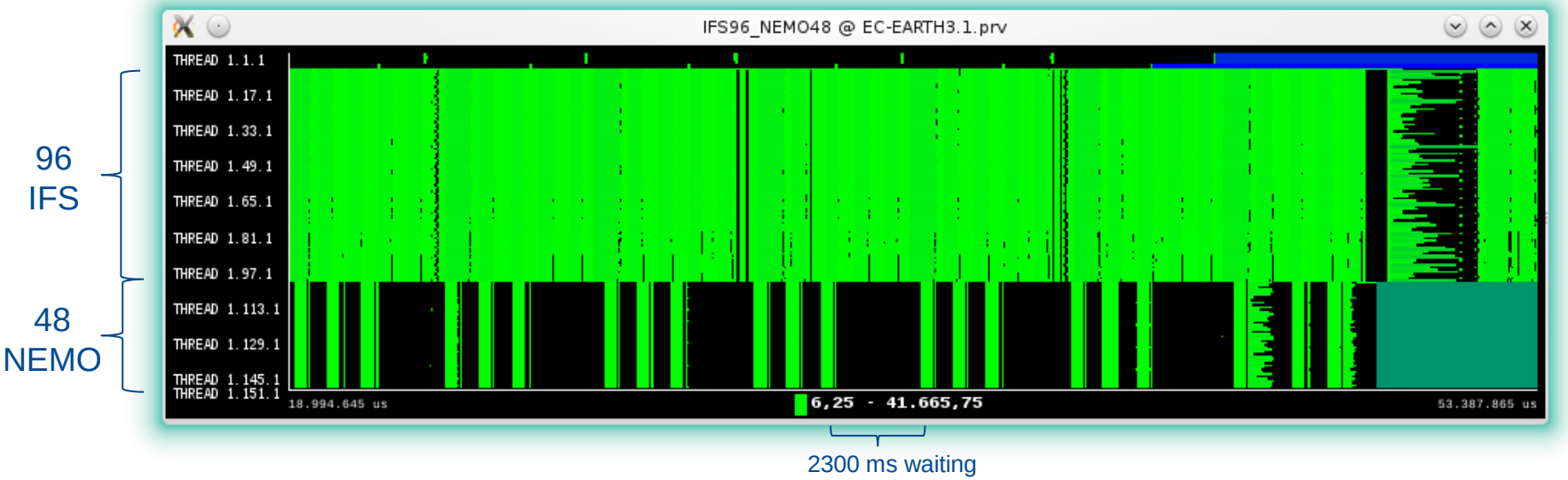



1 day simulation → 2,1s save
10 years simulation → 2,1h save

Only with a change in the order of processes !!

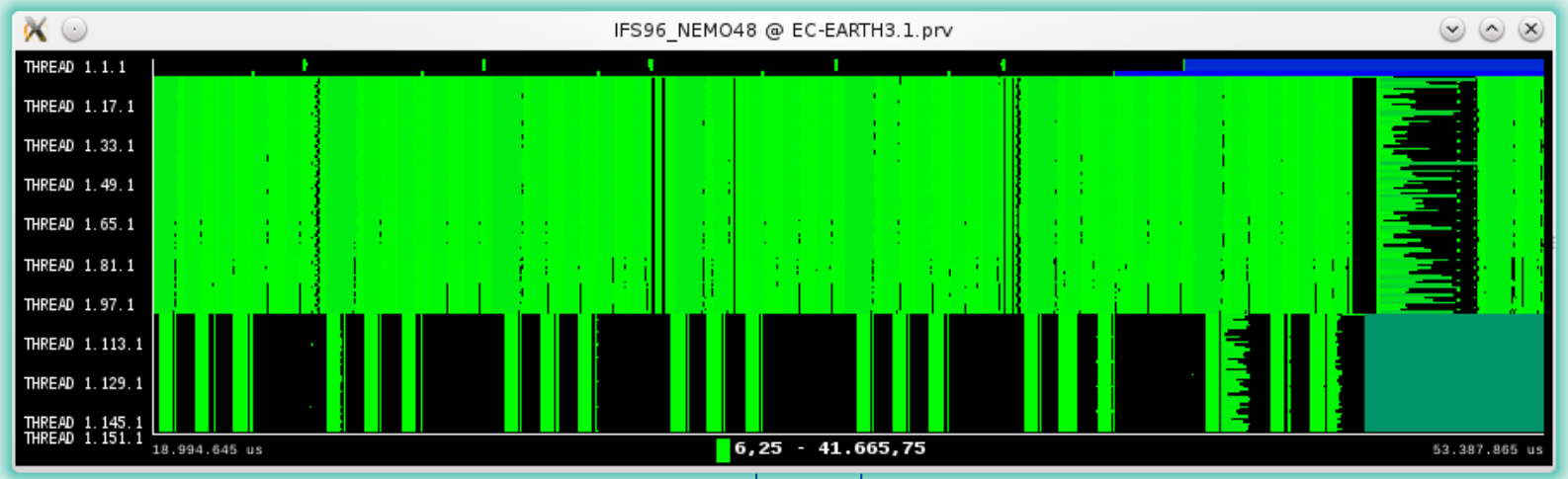
# Finding an optimum configuration

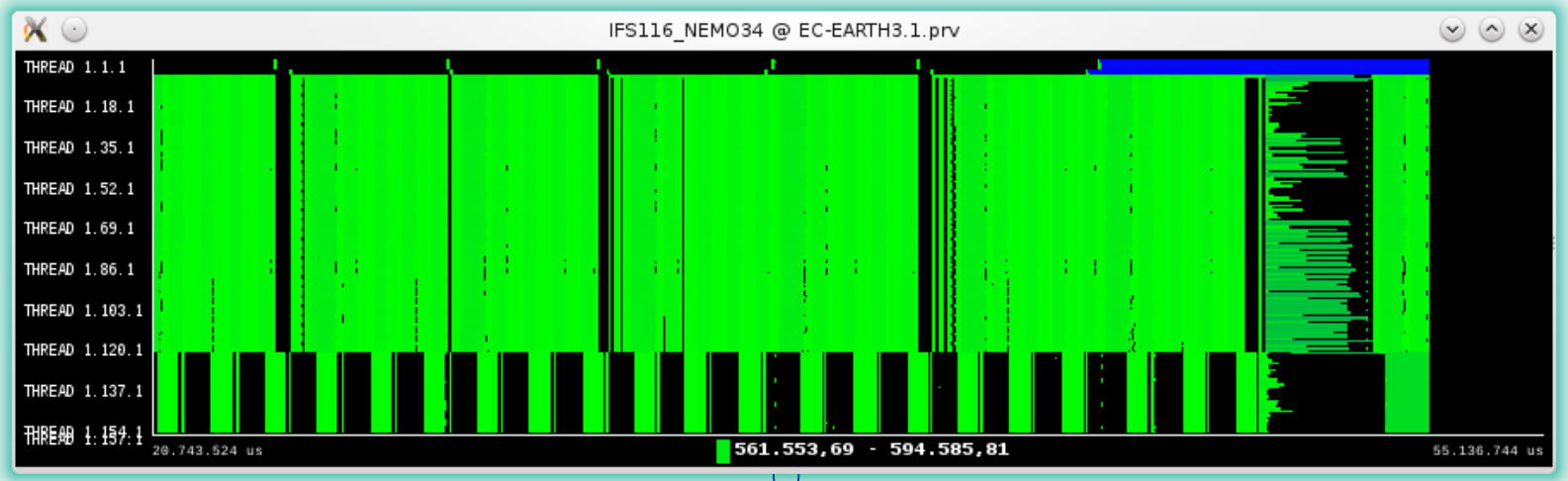
- Finding a better configuration to increase the performance.



# Finding an optimum configuration

- Finding a better configuration to increase the performance.
- Optimum configuration for this resolution: 116 IFS – 34 NEMO

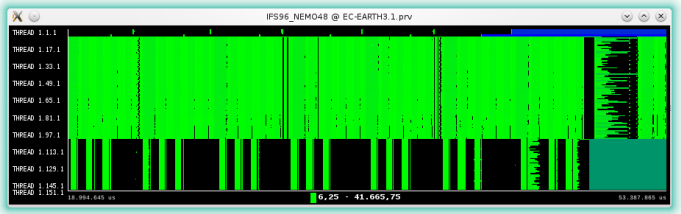


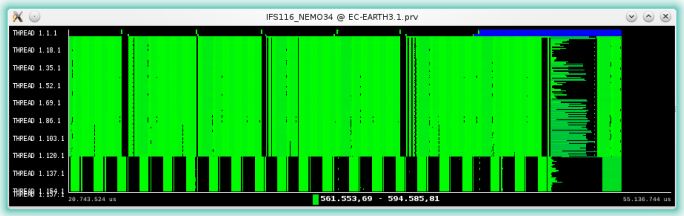
# Finding an optimum configuration

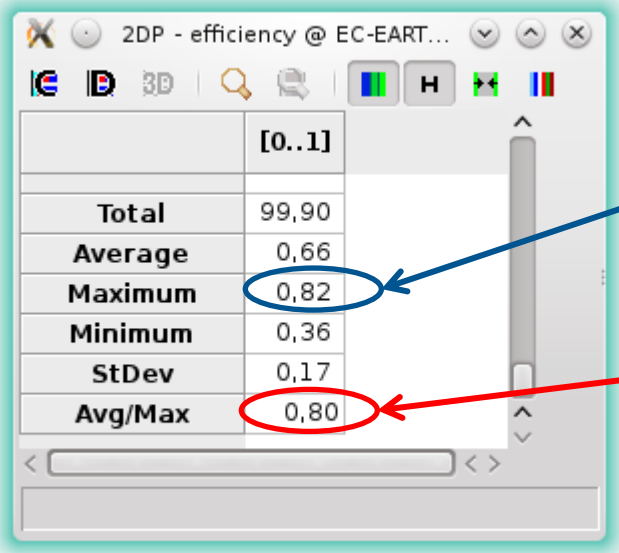
- The time gain comes from a better resource usage
- The workload is better balanced → Can be confirmed in Paraver

7 OASIS - 96 IFS – 48 NEMO



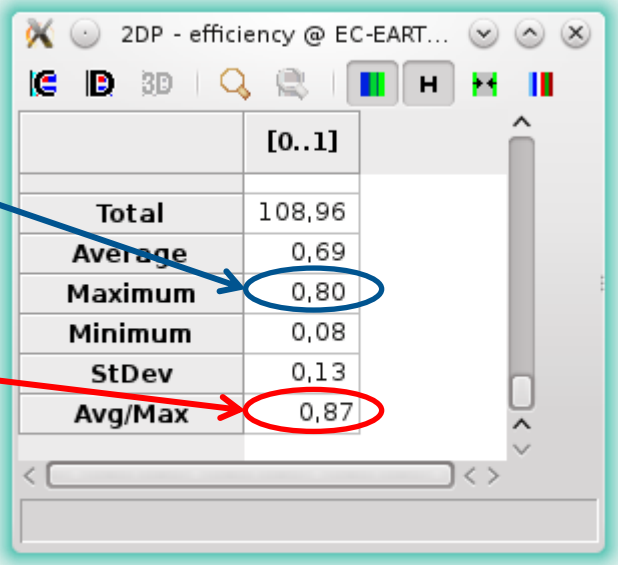
7 OASIS – 116 IFS – 34 NEMO






**Comms. Efficiency**

**Load Balance**



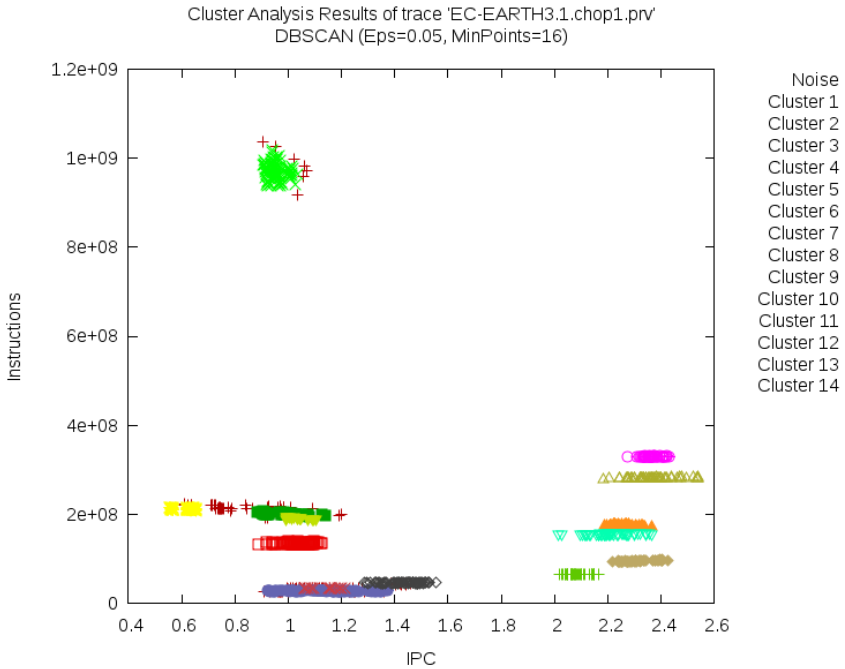

**+0.07 LB efficiency**


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

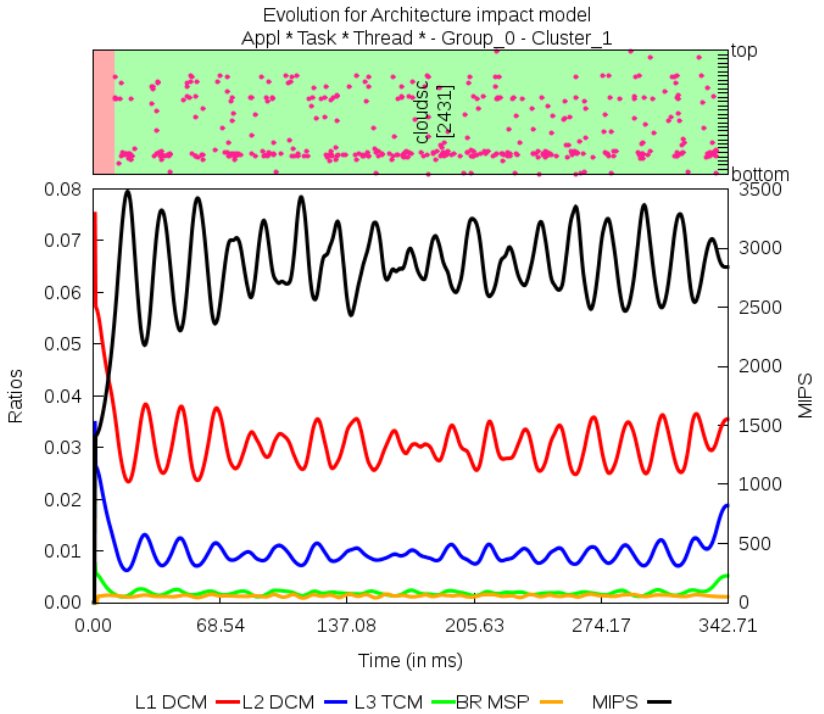

# Ongoing work

# Instantaneous Performance (Clustering + Folding)

- **Clustering** → Data mining technique for data classification
  - Detects different trends in the computation regions (CPU bursts) of an application
  - Similarity intended in terms of duration or hardware counter reduced metrics
- **Folding** → Instantaneous metrics with minimum overhead
  - Instrumentation (delimits regions) + sampling (progression within the regions)
  - Captures performance counters and call-stack references



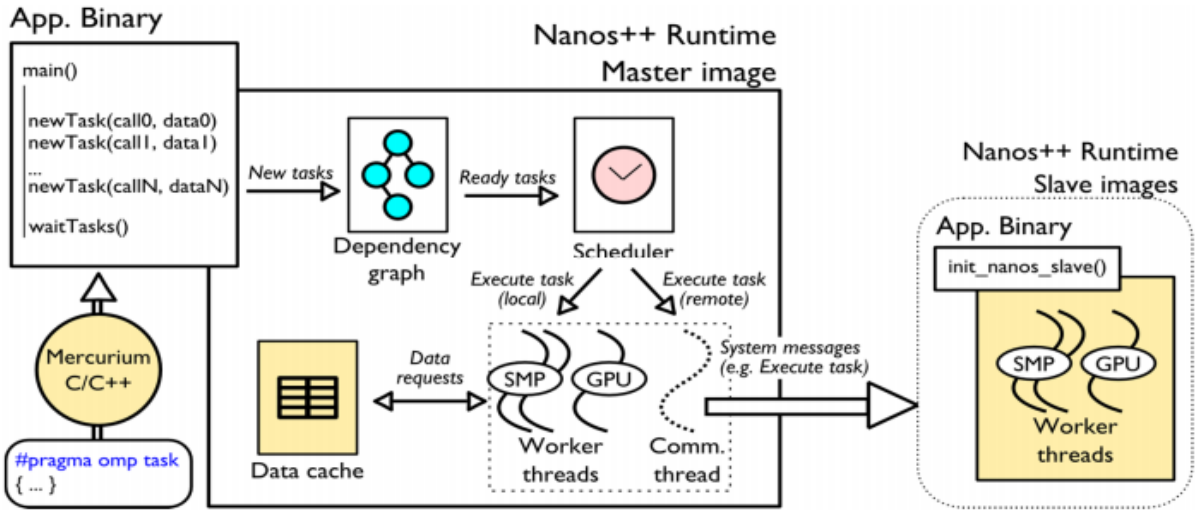

EC-Earth clustering scatter plot




Performance view of a folded cluster

# OmpSs

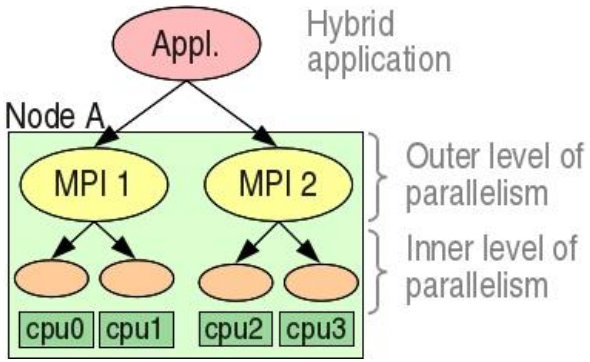
- Parallel Programming Model
- Build on existing standard: OpenMP
- Directive based to keep a serial version
- Targeting: SMP, clusters, and accelerator devices
- Developed at the Barcelona Supercomputing Center (BSC)
  - Mercurium source-to-source compiler
  - Nanos++ runtime system
- https://pm.bsc.es/ompss

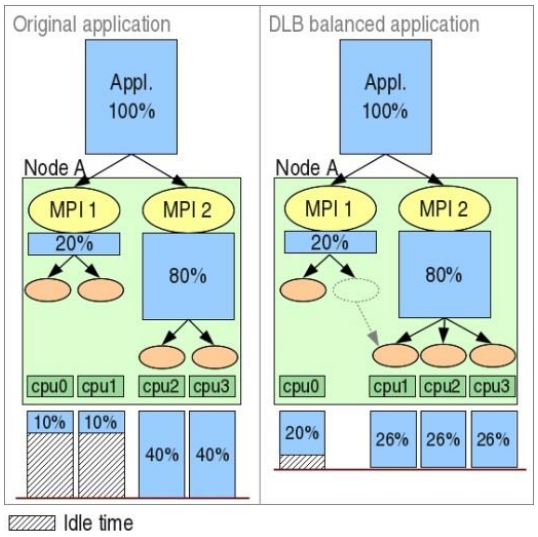


# Dynamic Load Balancing

- Dynamic library
- Speeds up hybrid applications with nested parallelism
  - Improving the load imbalance of the outer level parallelism
  - Different load balancing algorithms
- Load balance within node
- Automatically achieved by the runtime
- LeWI: Lend core When Idle

dynamic load balancing



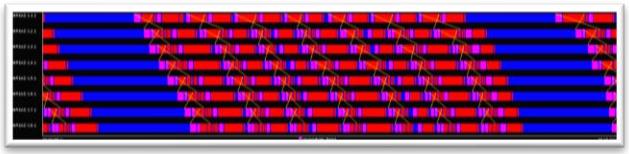



# NEMO 3.6 optimization

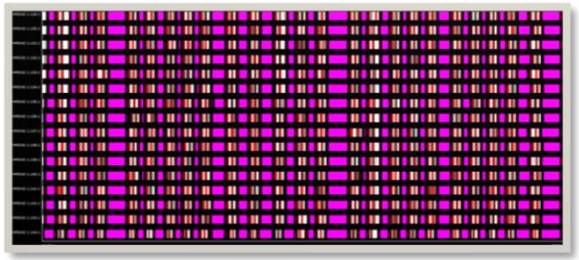


## Communication Bottlenecks identified using ORCA1 domain

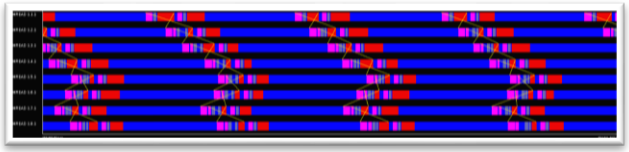
Computation / Communication ratio



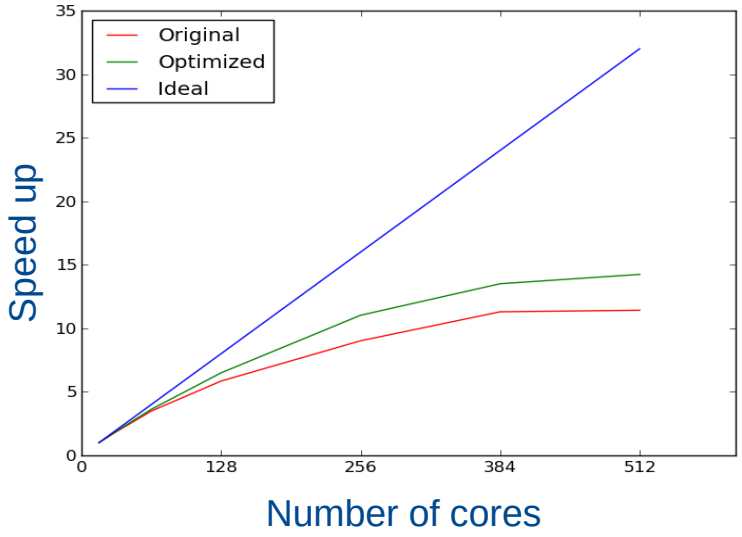
Abuse of Global Communications

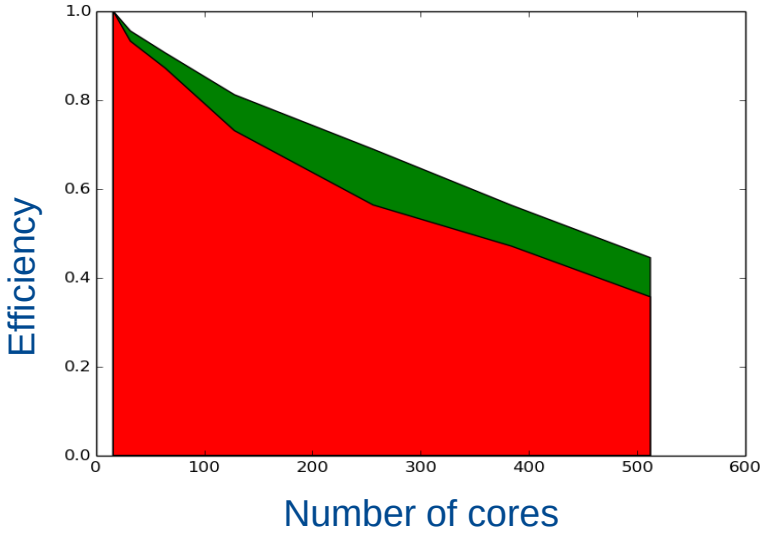


**Message Packing**



**Reduction of Global Communications**




Barcelona Supercomputing Center
Centro Nacional de Supercomputación


# Conclusions

# Conclusions



- Little changes in the configuration can significantly improve the performance.
- Trace analysis can guide the users in performing this task.
- A precise analysis and prediction can generate ideas that direct the restructuring of the application in the most productive way.

www.bsc.es


# Thank you!

For further information please contact
miguel.castrillo@bsc.es