Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Understanding applications with Paraver

Judit Gimenez
judit@bsc.es

Jan 31st 2019

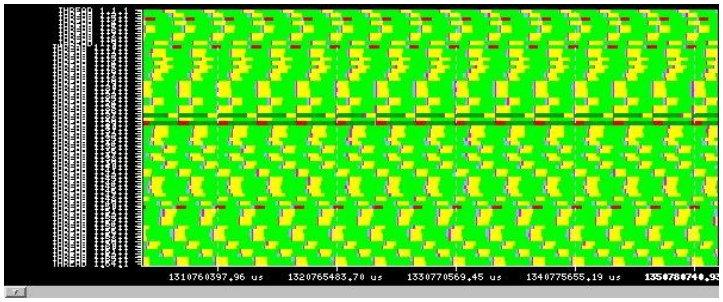
CRHPCS19, San José

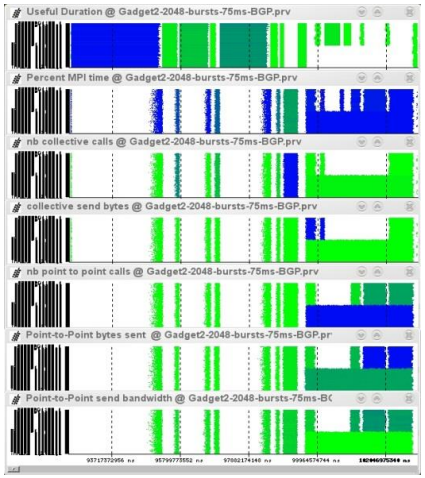
Humans are visual creatures

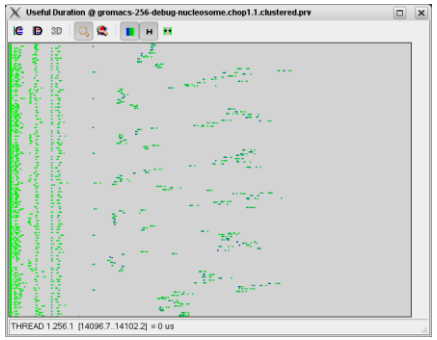
- Films or books? PROCESS
  - Two hours vs. days (months)
- Memorizing a deck of playing cards STORE
  - Each card translated to an image (person, action, location)
- Our brain loves pattern recognition IDENTIFY
  - What do you see on the pictures?

Our Tools

- Since 1991
- Based on traces
- Open Source
- http://tools.bsc.es
- Core tools:
  - Paraver (paramedir) – offline trace analysis
  - Dimemas – message passing simulator
  - Extrae – instrumentation
- Focus
  - Detail, variability, flexibility
  - Behavioral structure vs. syntactic structure
  - Intelligence: Performance Analytics



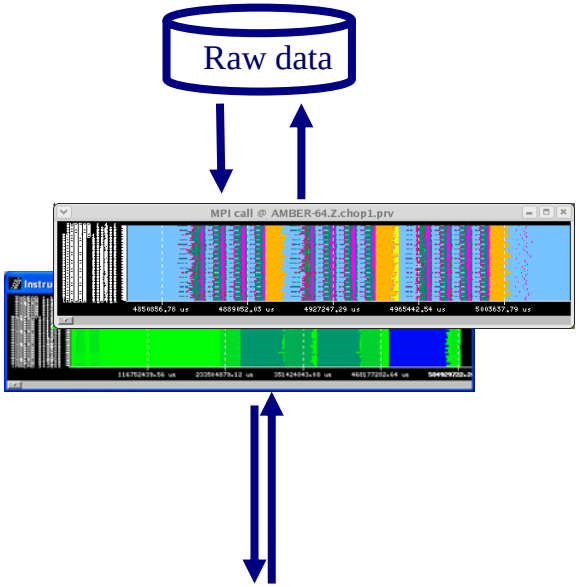




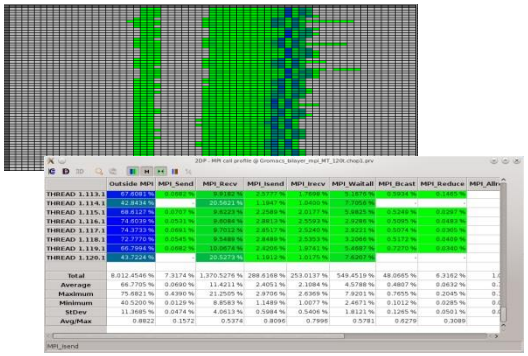
Paraver

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Paraver – Performance data browser



Timelines



2/3D tables (Statistics)

Trace visualization/analysis

+ trace manipulation

Goal = Flexibility

No semantics

Programmable

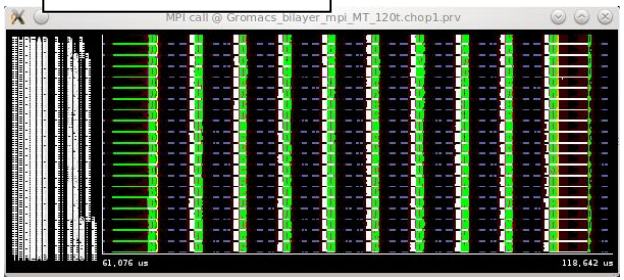
Comparative analyses

Multiple traces

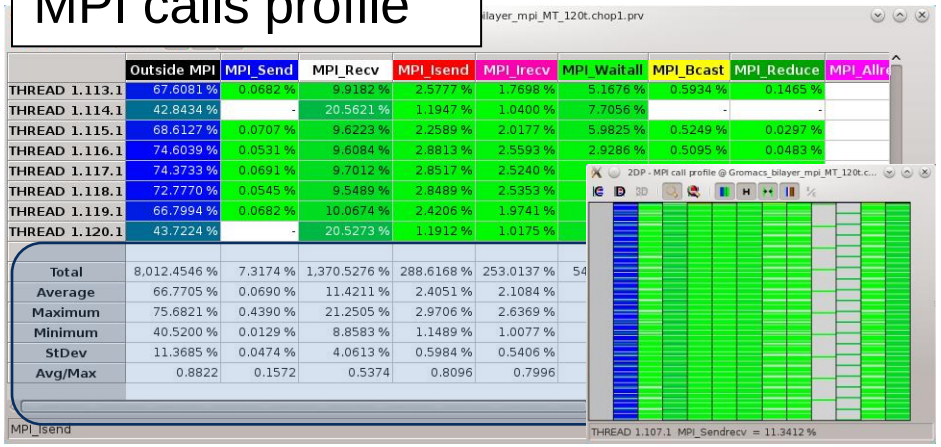
Synchronize scales

From timelines to tables

MPI calls

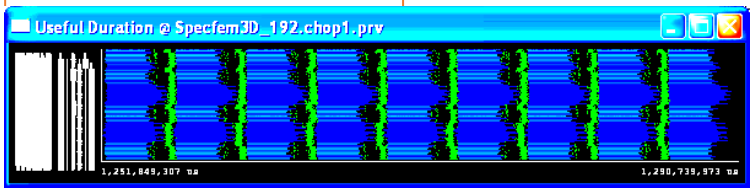


MPI calls profile

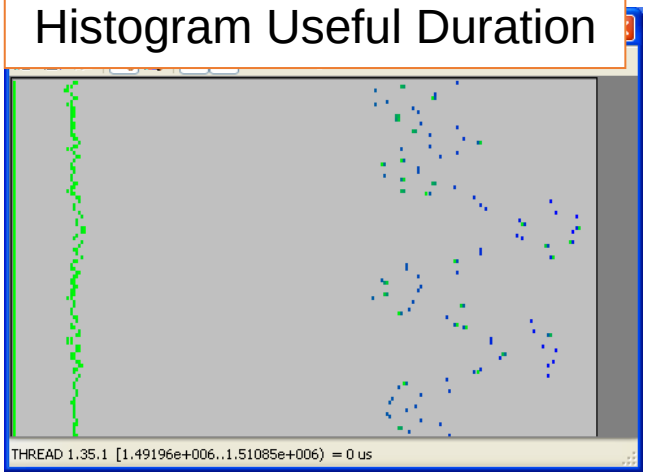


	Outside MPI	MPI_Send	MPI_Recv	MPI_Isend	MPI_Irecv	MPI_Waitall	MPI_Bcast	MPI_Reduce
THREAD 1.113.1	67.6081 %	0.0682 %	9.9182 %	2.5777 %	1.7698 %	5.1676 %	0.5934 %	0.1465 %
THREAD 1.114.1	42.8434 %	-	20.5621 %	1.1947 %	1.0400 %	7.7056 %	-	-
THREAD 1.115.1	68.6127 %	0.0707 %	9.6223 %	2.2589 %	2.0177 %	5.9825 %	0.5249 %	0.0297 %
THREAD 1.116.1	74.6039 %	0.0531 %	9.6084 %	2.8813 %	2.5593 %	2.9286 %	0.5095 %	0.0483 %
THREAD 1.117.1	74.3733 %	0.0691 %	9.7012 %	2.8517 %	2.5240 %			
THREAD 1.118.1	72.7770 %	0.0545 %	9.5489 %	2.8489 %	2.5353 %			
THREAD 1.119.1	66.7994 %	0.0682 %	10.0674 %	2.4206 %	1.9741 %			
THREAD 1.120.1	43.7224 %	-	20.5273 %	1.1912 %	1.0175 %			
Total	8,012.4546 %	7.3174 %	1,370.5276 %	288.6168 %	253.0137 %	54		
Average	66.7705 %	0.0690 %	11.4211 %	2.4051 %	2.1084 %			
Maximum	75.6821 %	0.4390 %	21.2505 %	2.9706 %	2.6369 %			
Minimum	40.5200 %	0.0129 %	8.8583 %	1.1489 %	1.0077 %			
StDev	11.3685 %	0.0474 %	4.0613 %	0.5984 %	0.5406 %			
Avg/Max	0.8822	0.1572	0.5374	0.8096	0.7996			

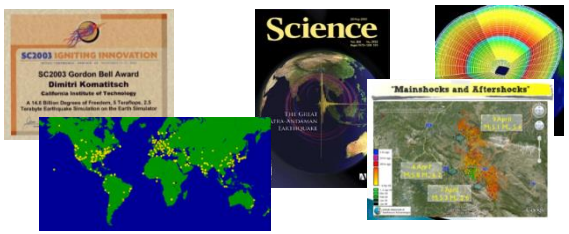
Useful Duration

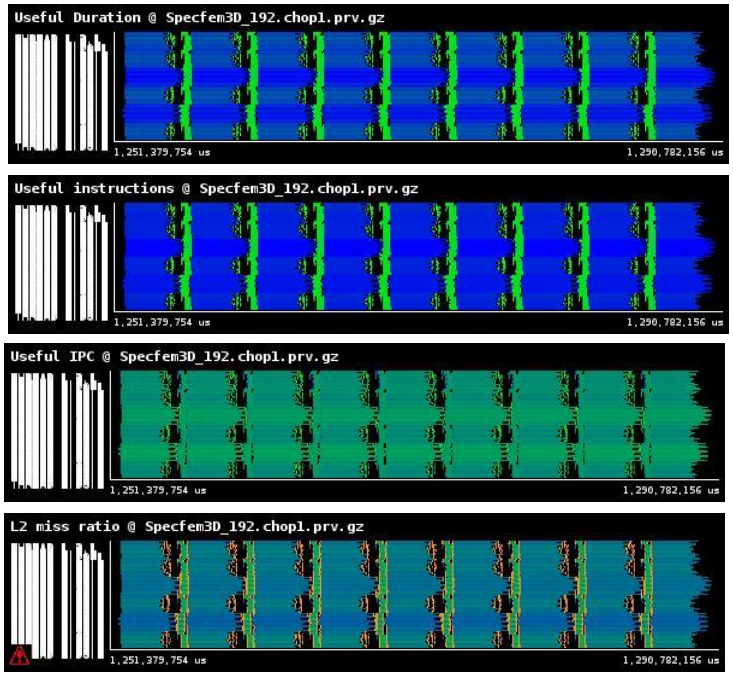


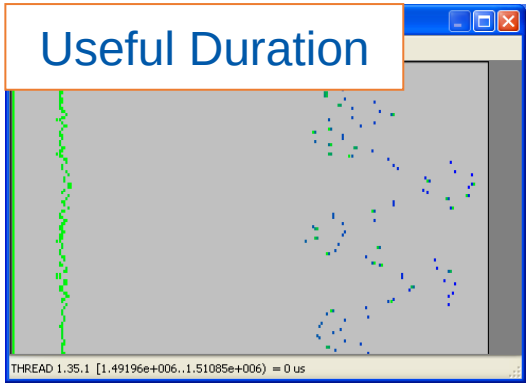
Histogram Useful Duration

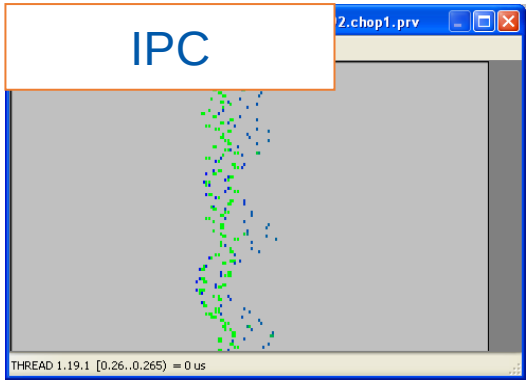


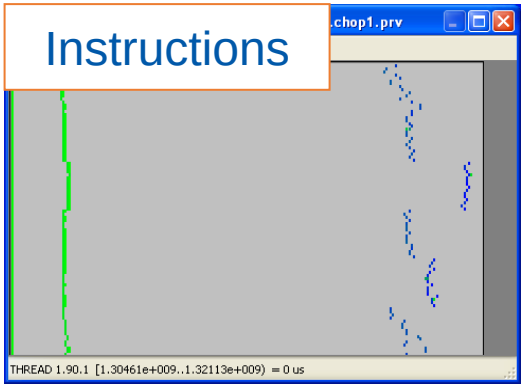
Analyzing variability

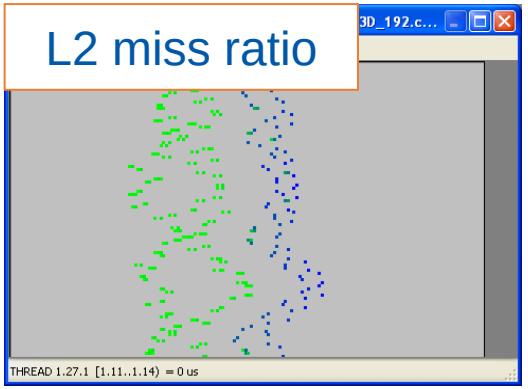






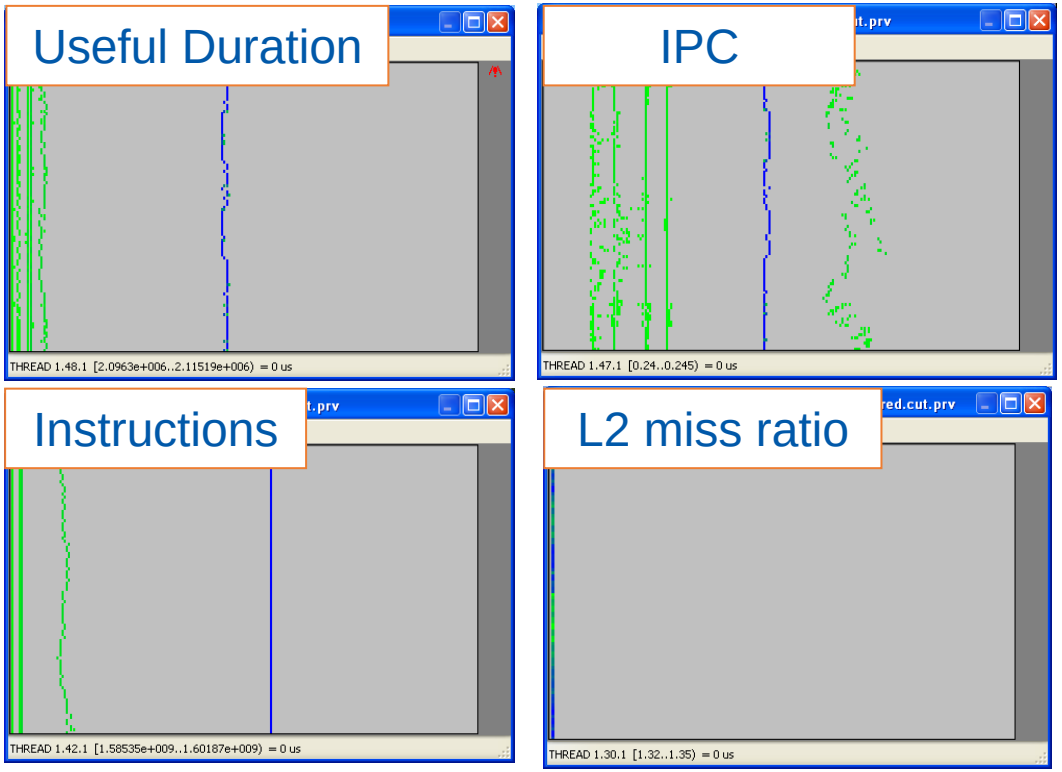






Analyzing variability

- By the way: six months later ….

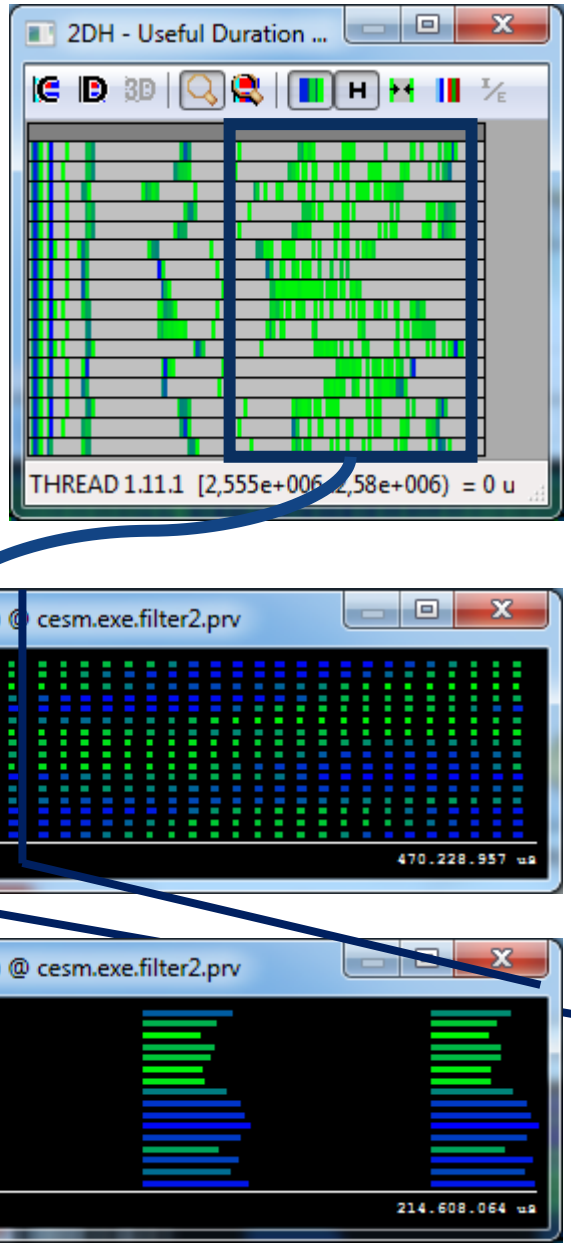


From tables to timelines

CESM: 16 processes, 2 simulated days

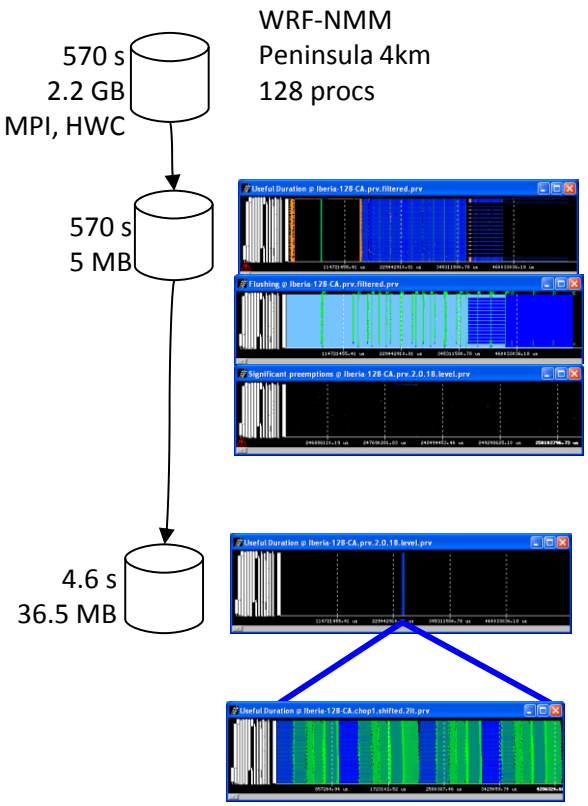
- Histogram useful computation duration shows high variability
- How is it distributed?

- Dynamic imbalance
  - In space and time
  - Day and night.
  - Season ? ☺



Trace manipulation

- Data handling/summarization capability
  - Filtering
    - Subset of records in original trace
    - By duration, type, value,…
    - Filtered trace IS a paraver trace and can be analysed with the same cfgs (as long as needed data kept)
  - Cutting
    - All records in a given time interval
    - Only some processes
  - Software counters
    - Summarized values computed from those in the original trace emitted as new even types
    - #MPI calls, total hardware count,…

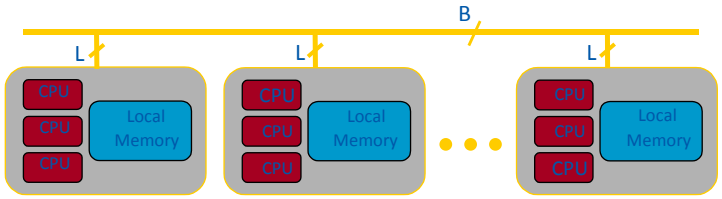


Dimemas

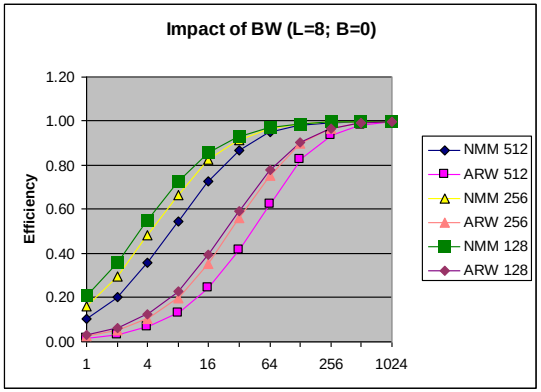

BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Dimemas – Coarse grain, Trace driven simulation

- Simulation: Highly non linear model
  - MPI protocols, resource contention…



- Parametric sweeps
  - On abstract architectures
  - On application computational regions
- What if analysis
  - Ideal machine (instantaneous network)
  - Estimating impact of ports to MPI+OpenMP/CUDA/…
  - Should I use asynchronous communications?
  - Are all parts equally sensitive to network?
- MPI sanity check
  - Modeling nominal

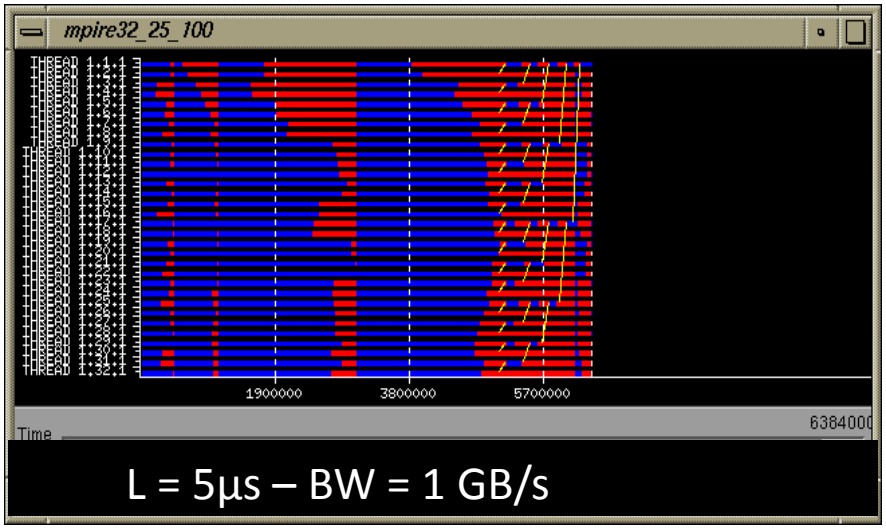


- Paraver – Dimemas tandem
  - Analysis and prediction
  - What-if from selected time window

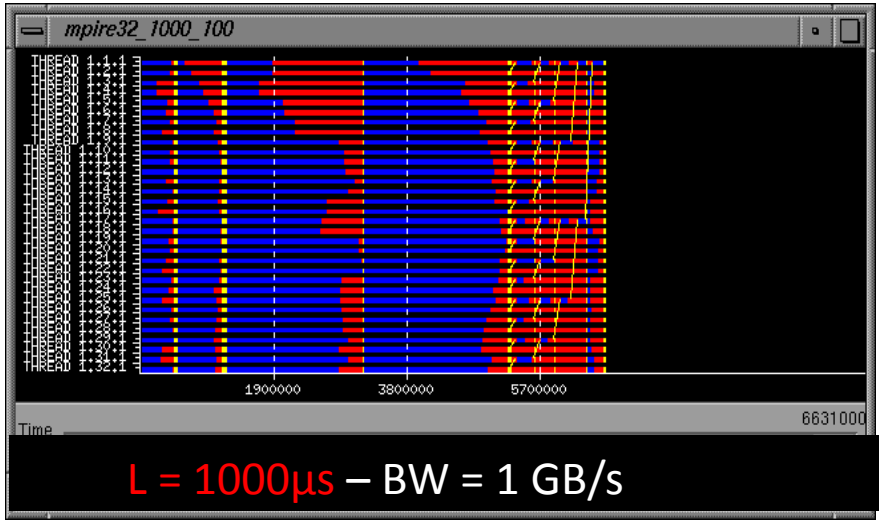
Detailed feedback on simulation (trace)

Network sensitivity

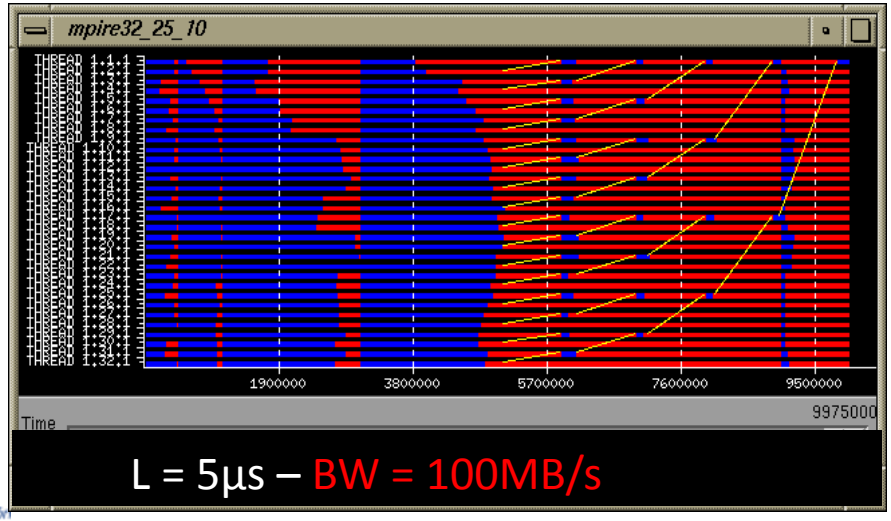
- MPIRE 32 tasks, no network contention



L = 5µs – BW = 1 GB/s



L = 1000µs – BW = 1 GB/s

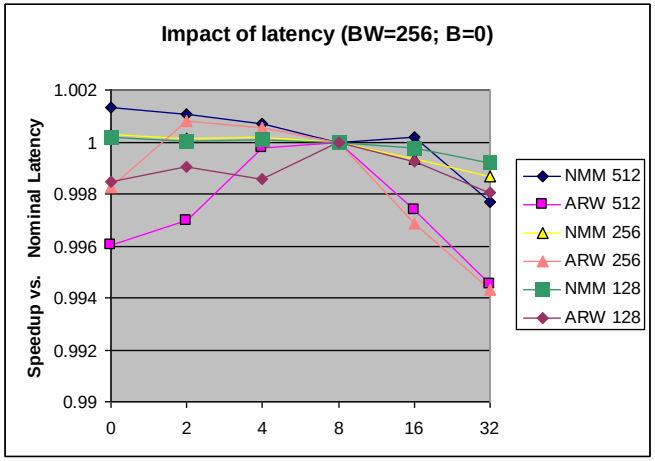


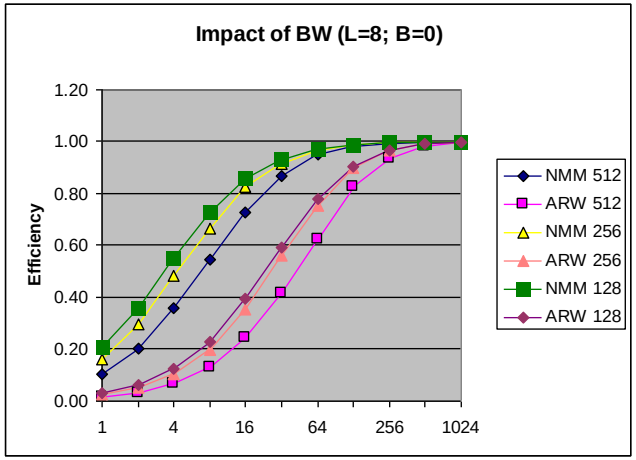
L = 5µs – BW = 100MB/s

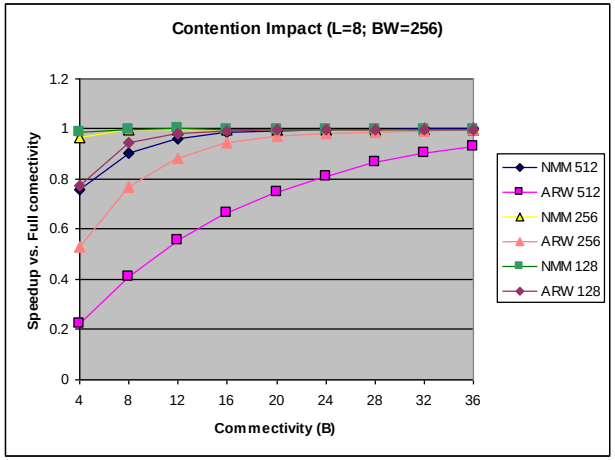
All windows same scale

Network sensitivity

- WRF, Iberia 4Km, 4 procs/node
  - Not sensitive to latency
  - NMM
    - BW – 256MB/s
    - 512 – sensitive to contention
  - ARW
    - BW - 1GB/s
    - Sensitive to contention

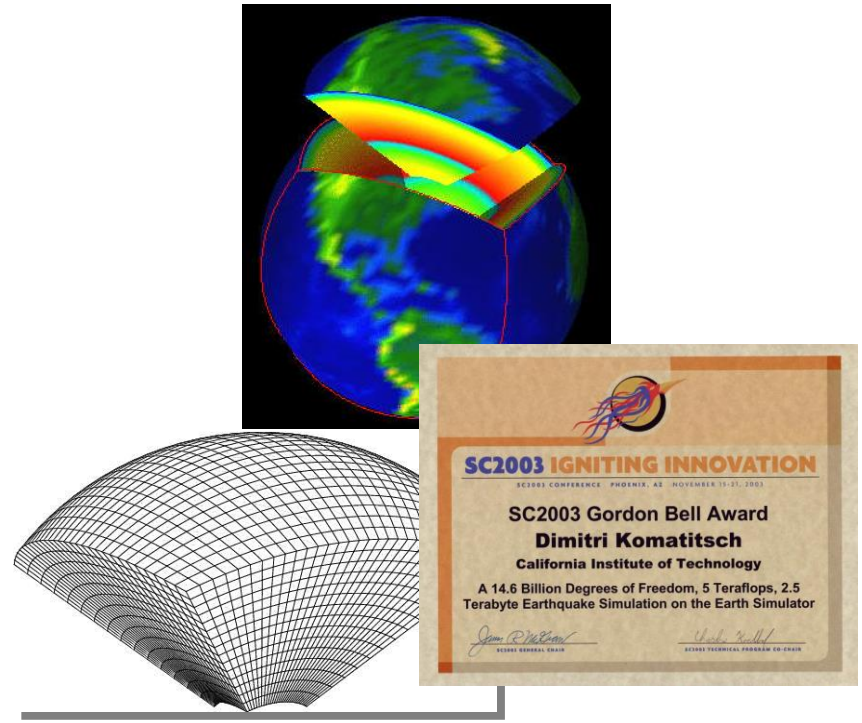




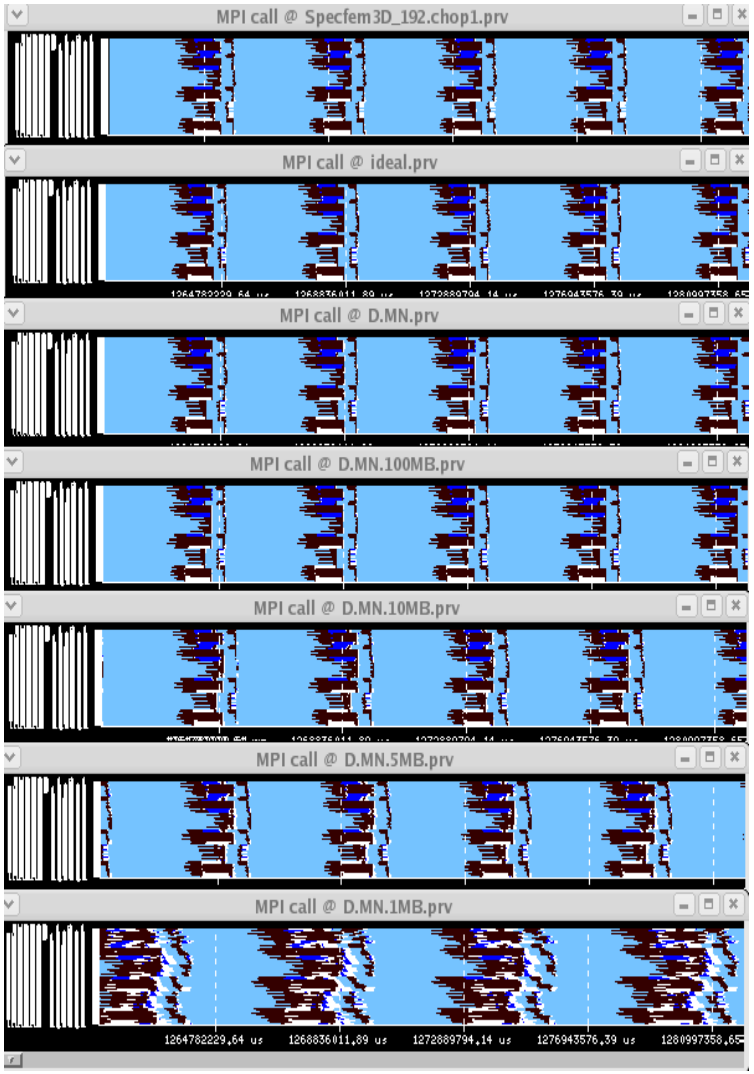


Would I will benefit from asynchronous communications?

SPECFEM3D



Courtesy Dimitri Komatitsch

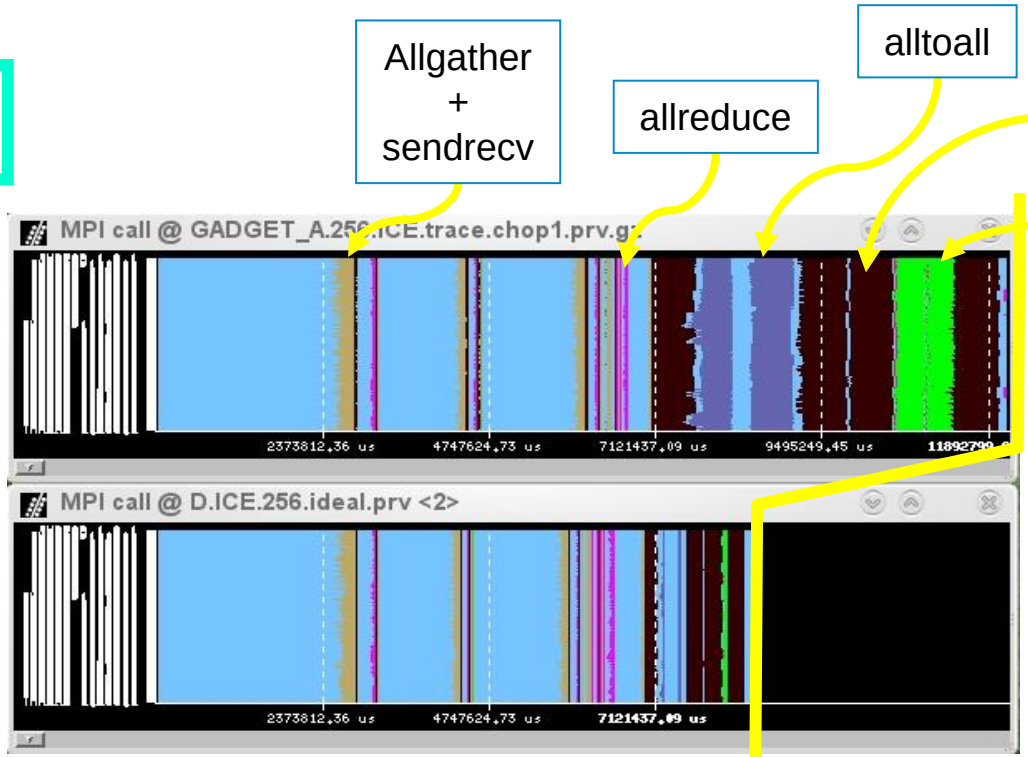


Ideal machine

The impossible machine: BW = ∞, L = 0

- Actually describes/characterizes Intrinsic application behavior
  - Load balance problems?
  - Dependence problems?

GADGET @ Nehalem cluster
256 processes

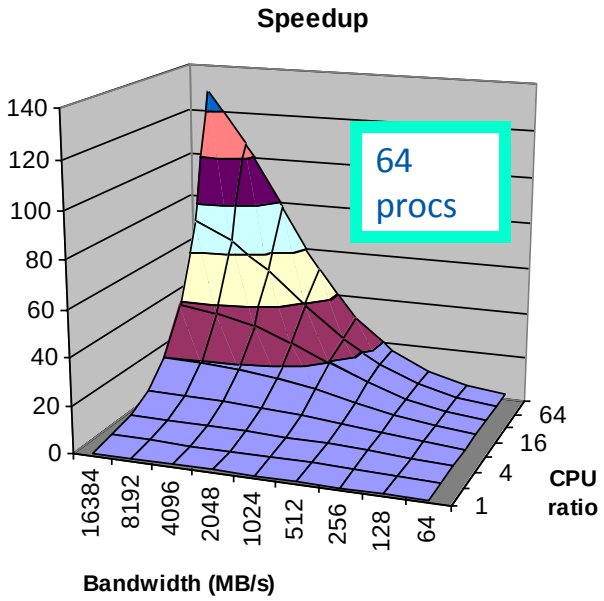


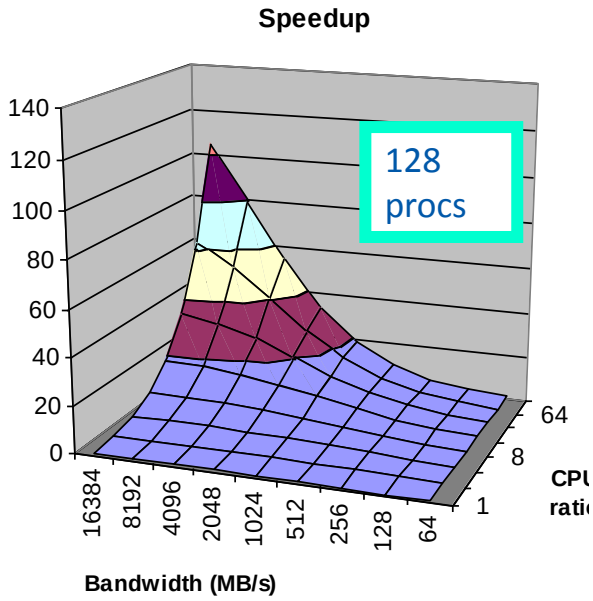
Impact on practical machines?

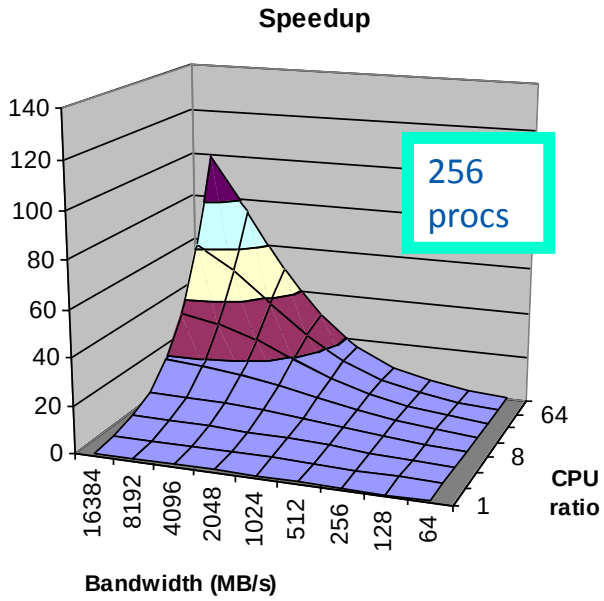
Impact of architectural parameters

- **Ideal speeding up ALL** the computation bursts by the CPUratio factor
  - The more processes the less speedup (higher impact of bandwidth limitations) !!

GADGET

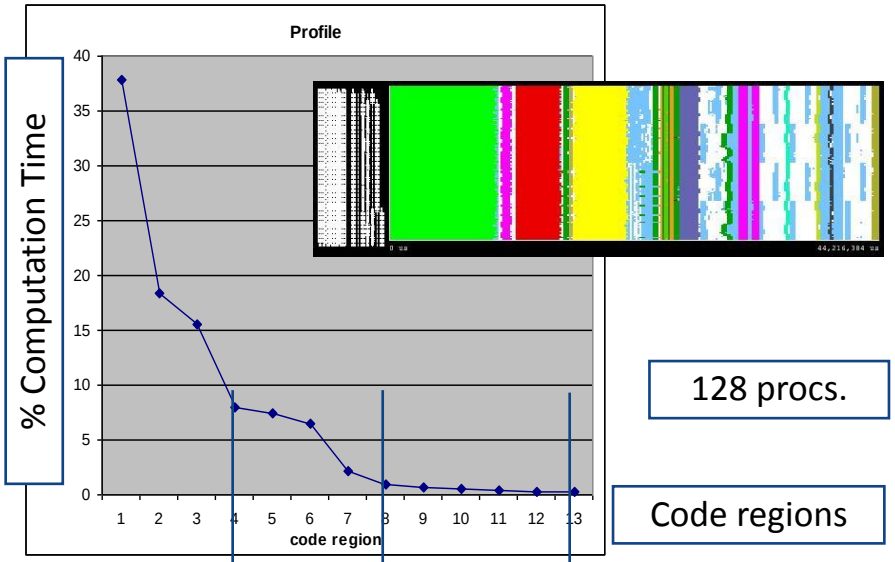


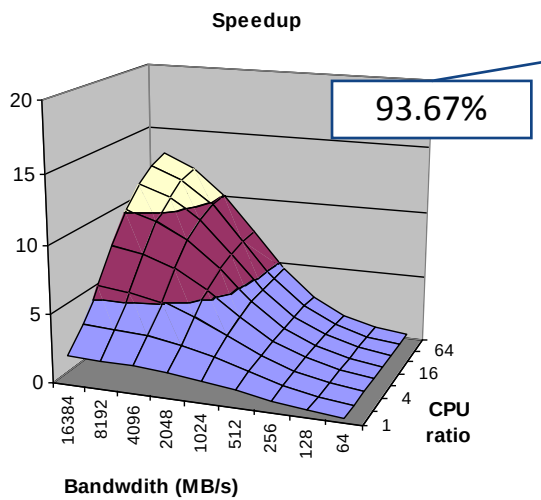


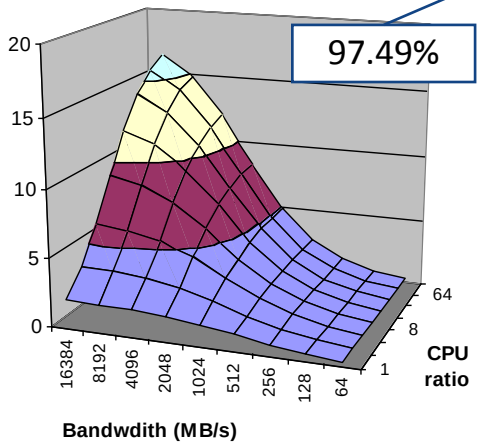


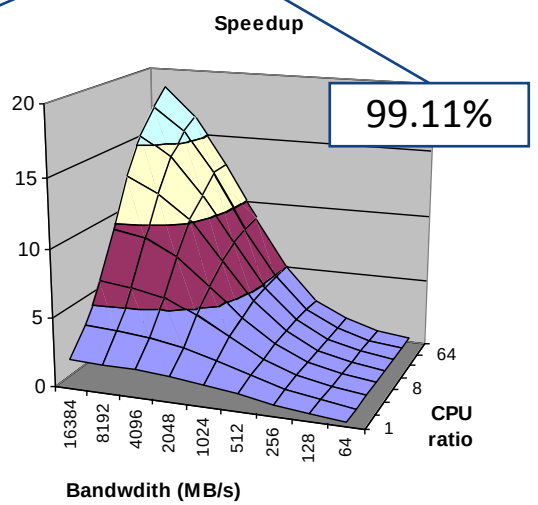
Hybrid parallelization

- Hybrid/accelerator parallelization
  - Speed-up SELECTED regions by the CPUratio factor







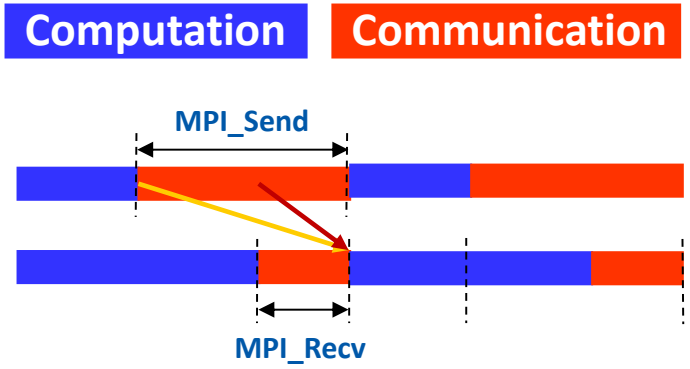


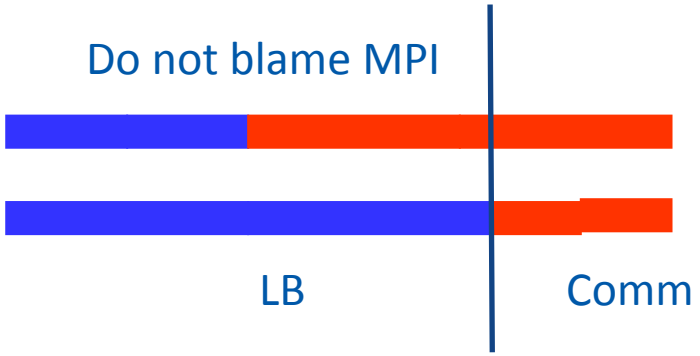
(Previous slide: speedups **up to 100x**)

Efficiency Models

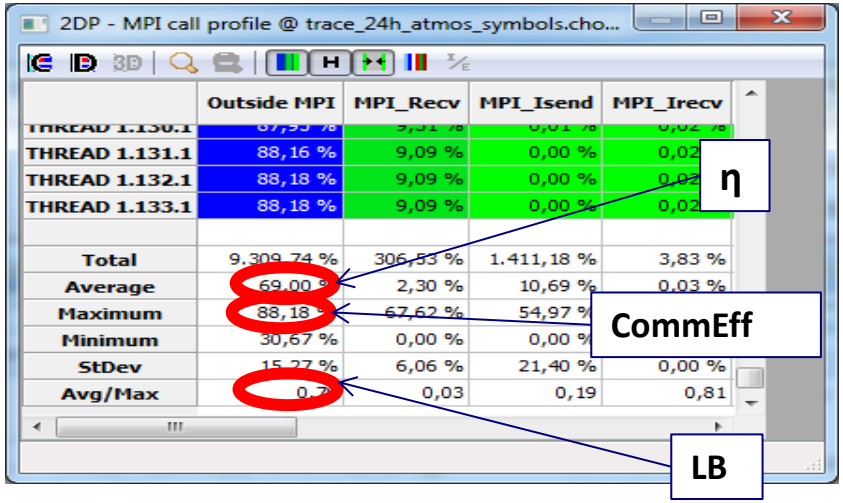
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Parallel efficiency model

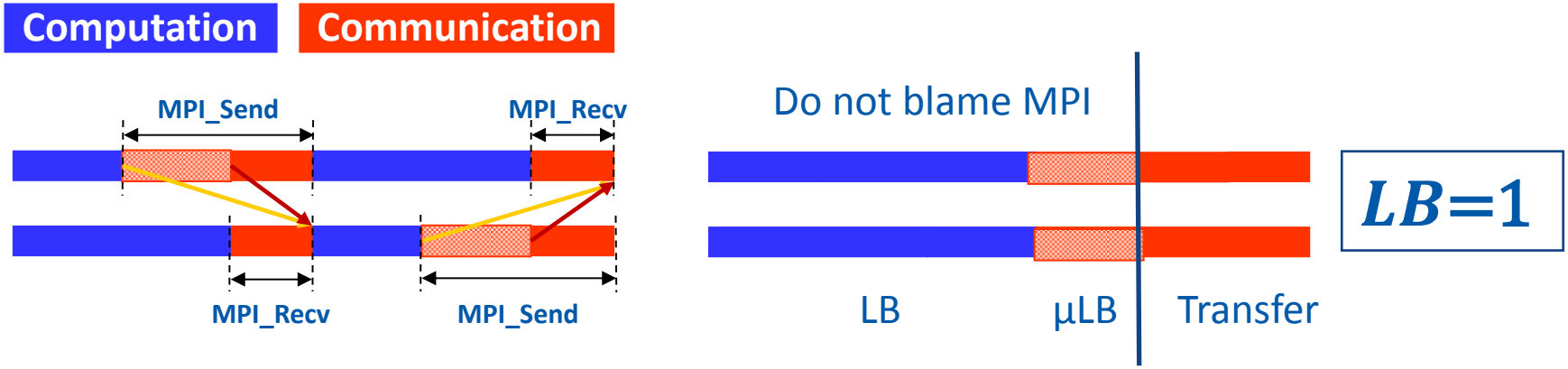




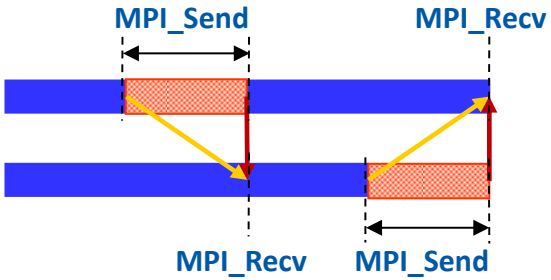
- Parallel efficiency = LB eff * Comm eff



Parallel efficiency refinement: LB * µLB * Tr



- Serializations / dependences (µLB)
- Dimemas ideal network → Transfer (efficiency) = 1

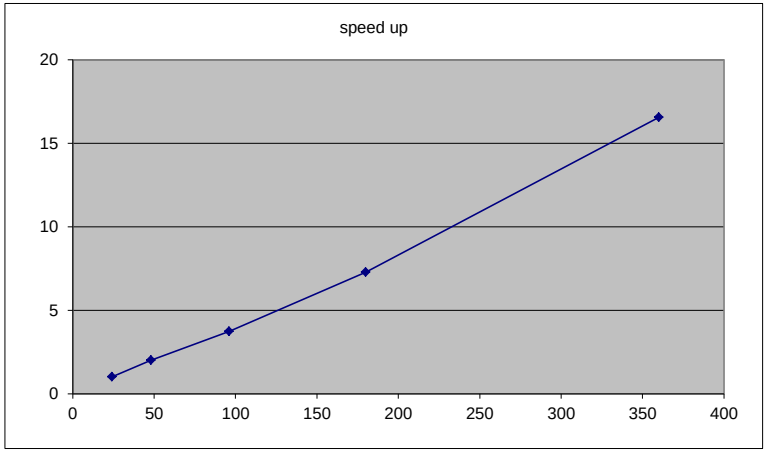


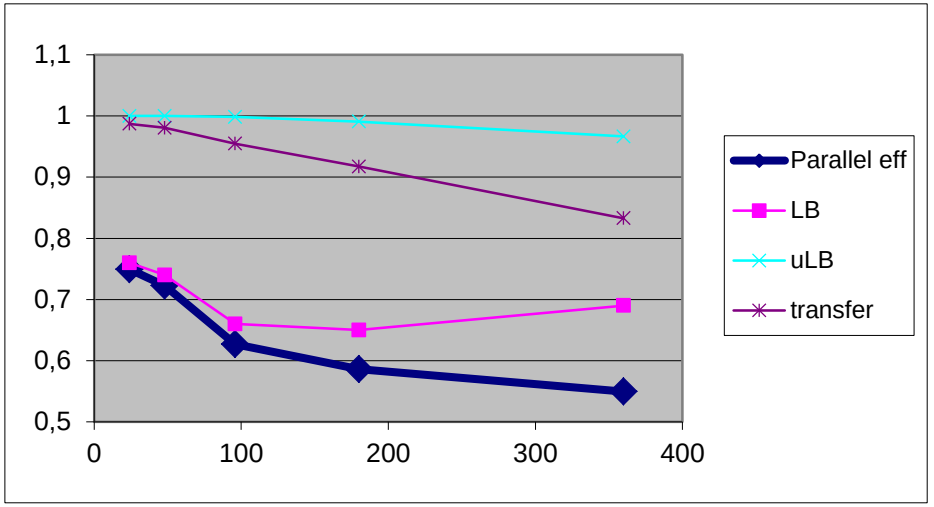
Why scaling?

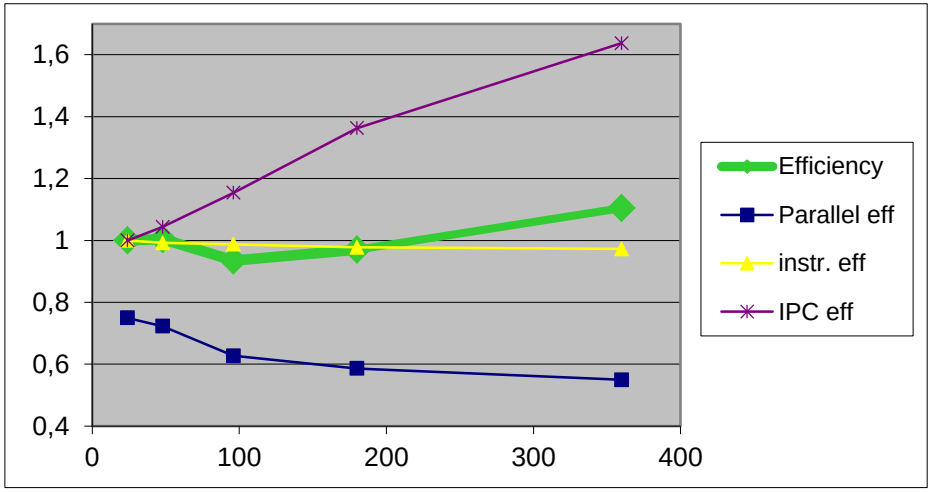
$$\eta_{\parallel} = LB * Ser * Trf$$

CG-POP mpi2s1D - 180x120

Good scalability !!
Should we be happy?





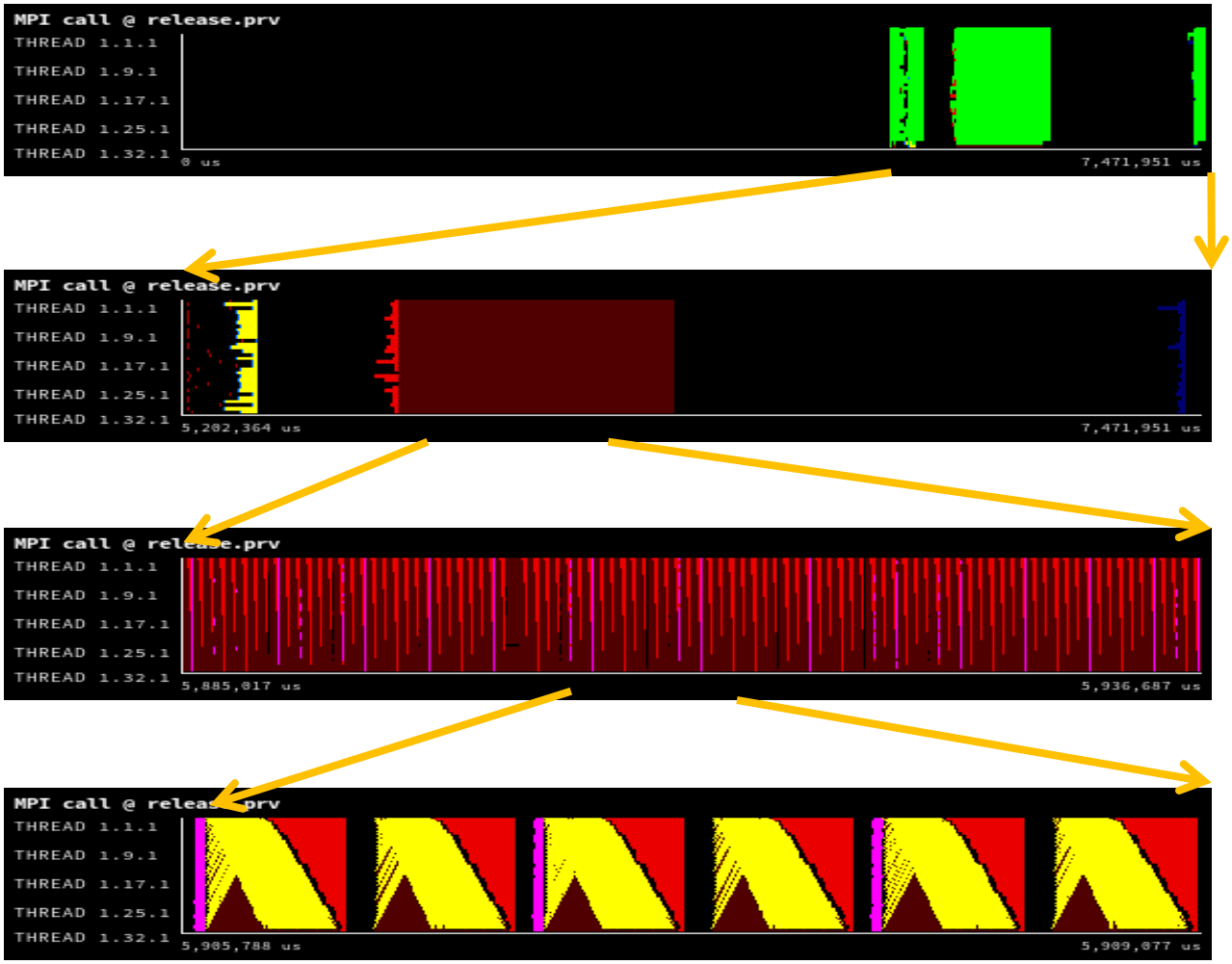


Why efficient?

Parallel efficiency =93.28
Communication = 93.84

Parallel efficiency = 77.93
Communication = 79.79

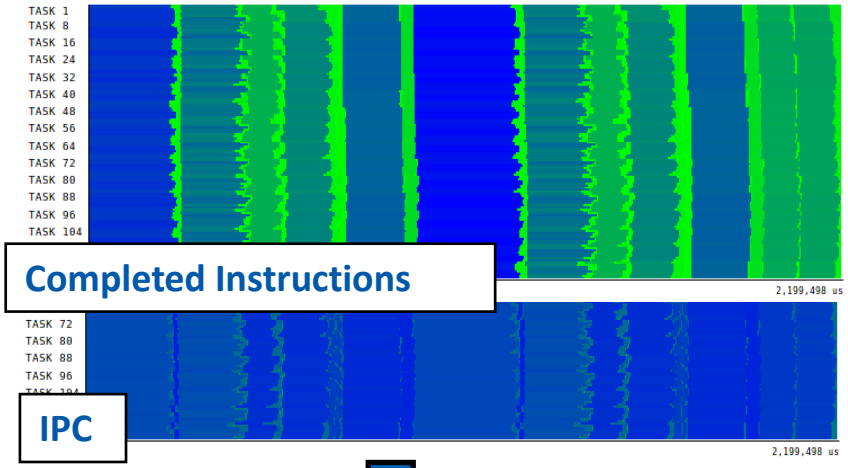
Parallel efficiency = 28.84
Communication eff = 30.42

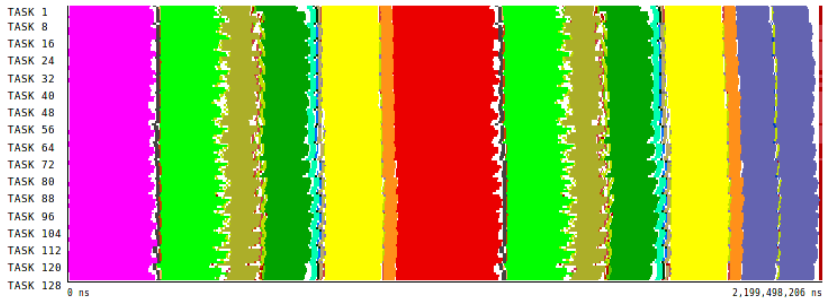


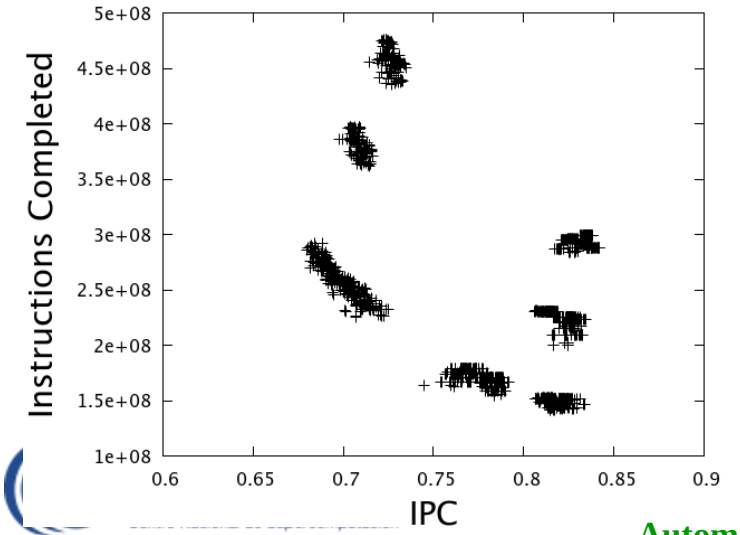
Analytics

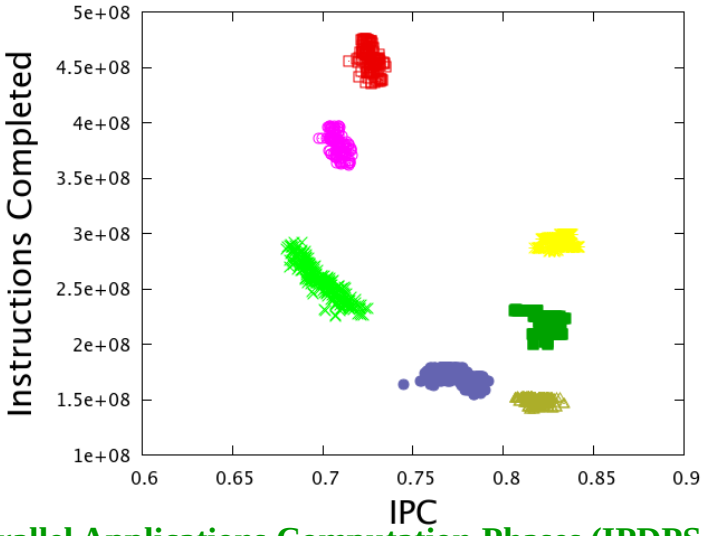
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Using Clustering to identify structure









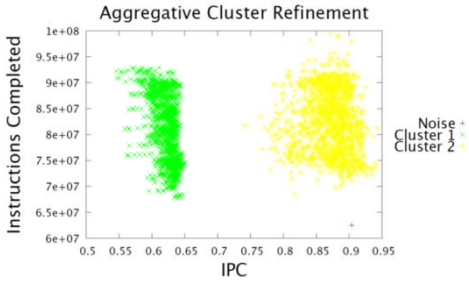
Automatic Detection of Parallel Applications Computation Phases (IPDPS 2009)

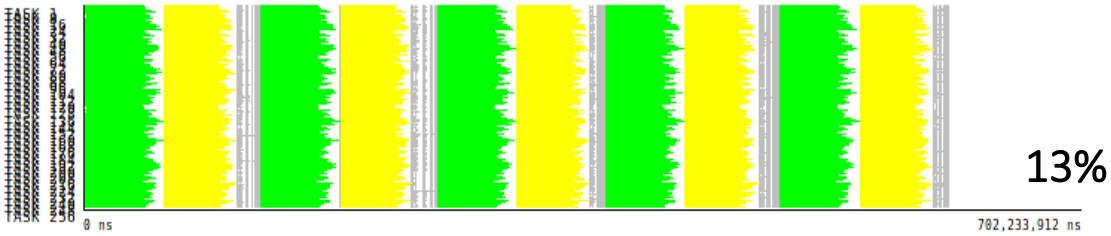
What should I improve?

What if ….

PEPC

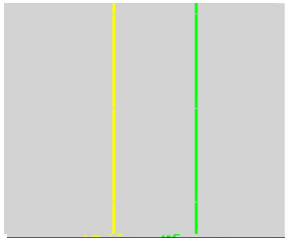
… we increase the IPC of Cluster1?

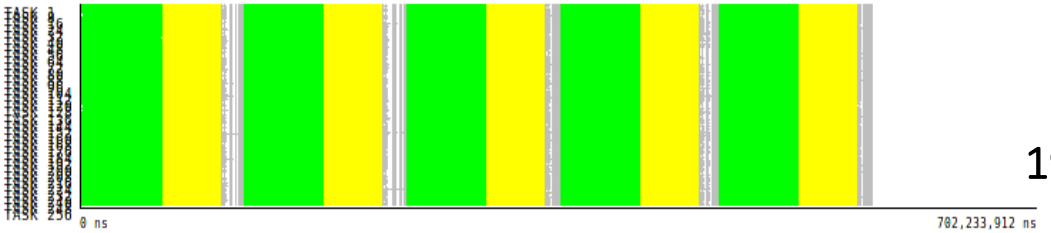




13% gain

… we balance Clusters 1 & 2?

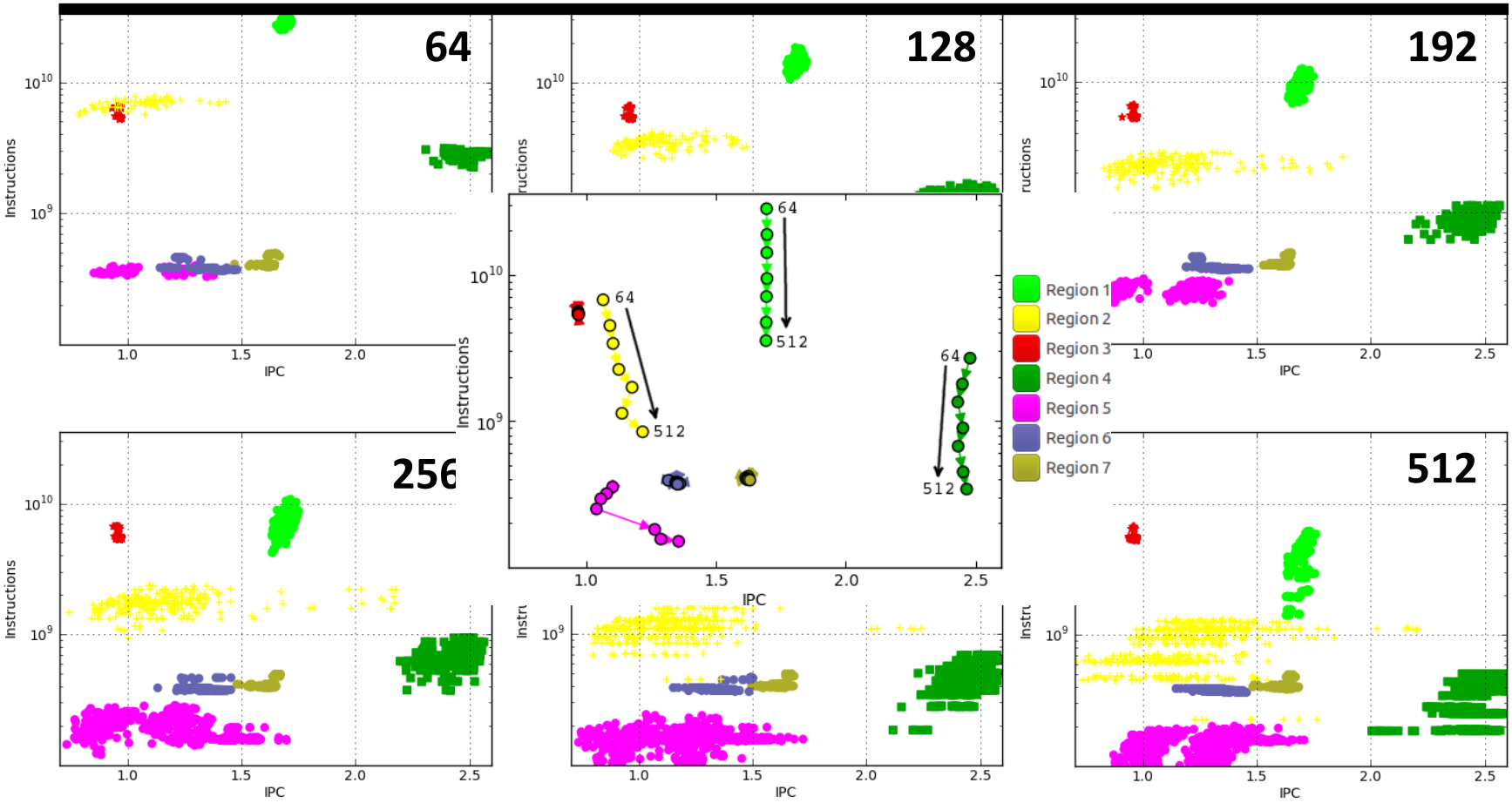




19% gain

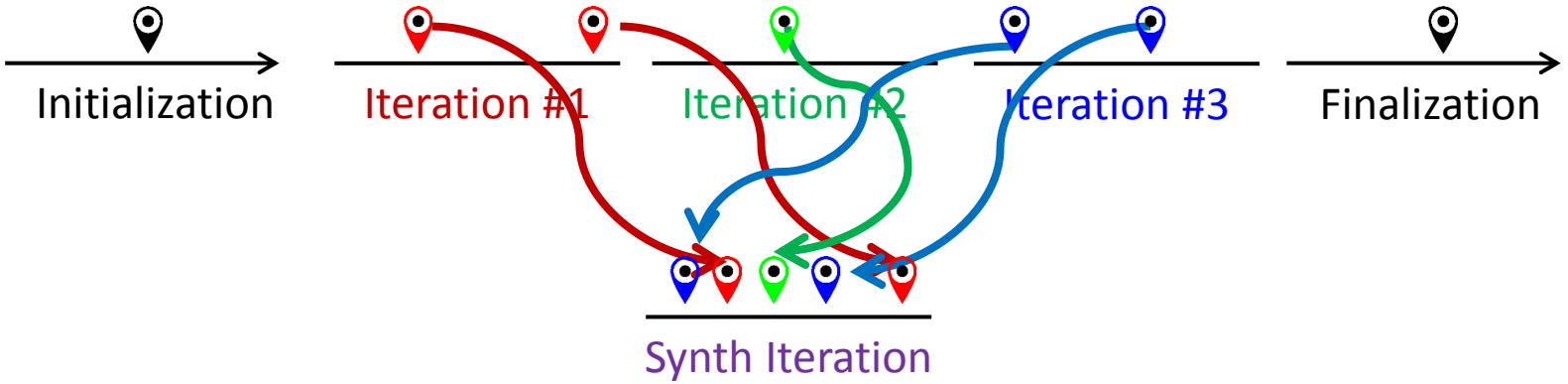
Tracking scability through clustering

- OpenMX (strong scale from 64 to 512 tasks)



Folding

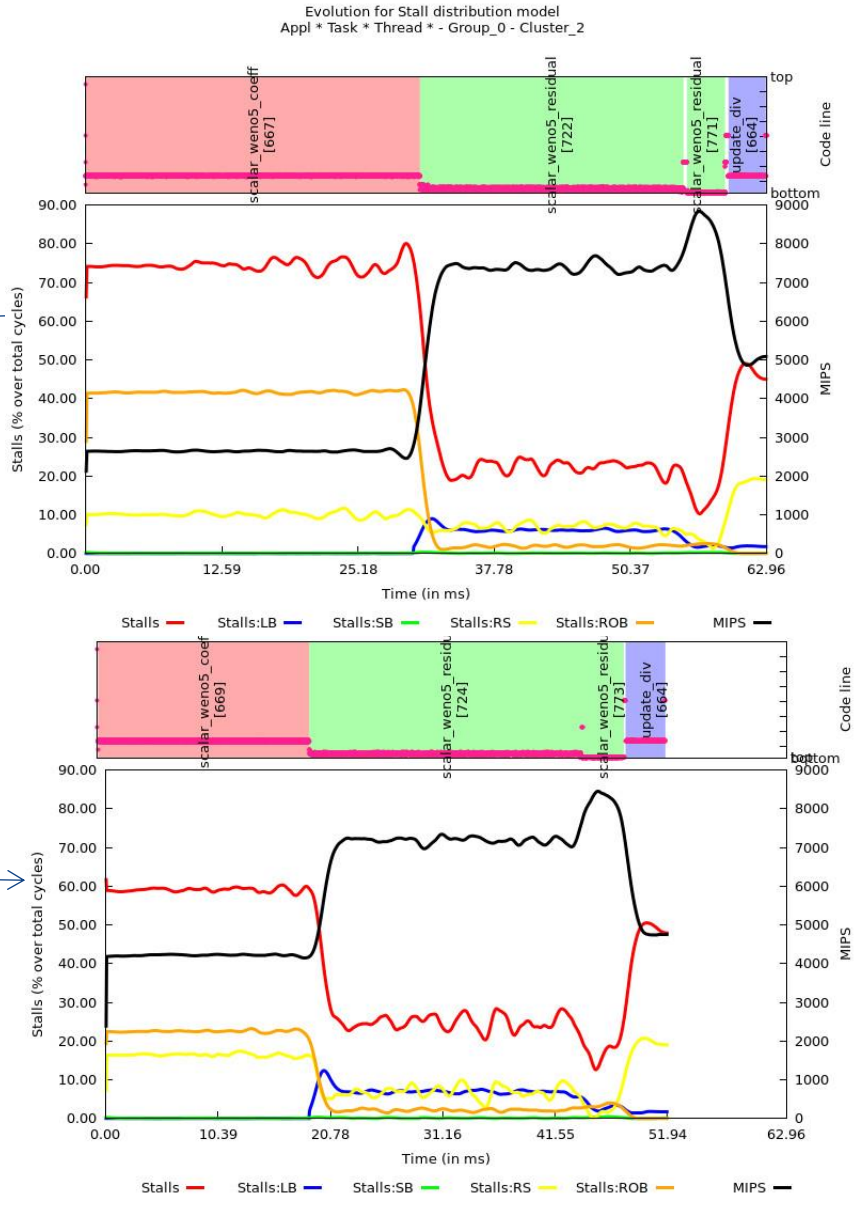
- Instantaneous metrics with minimum overhead
  - Combine instrumentation and sampling
    - Instrumentation delimits regions (routines, loops, …)
    - Sampling exposes progression within a region
  - Captures performance counters and call-stack references



“Blind” optimization

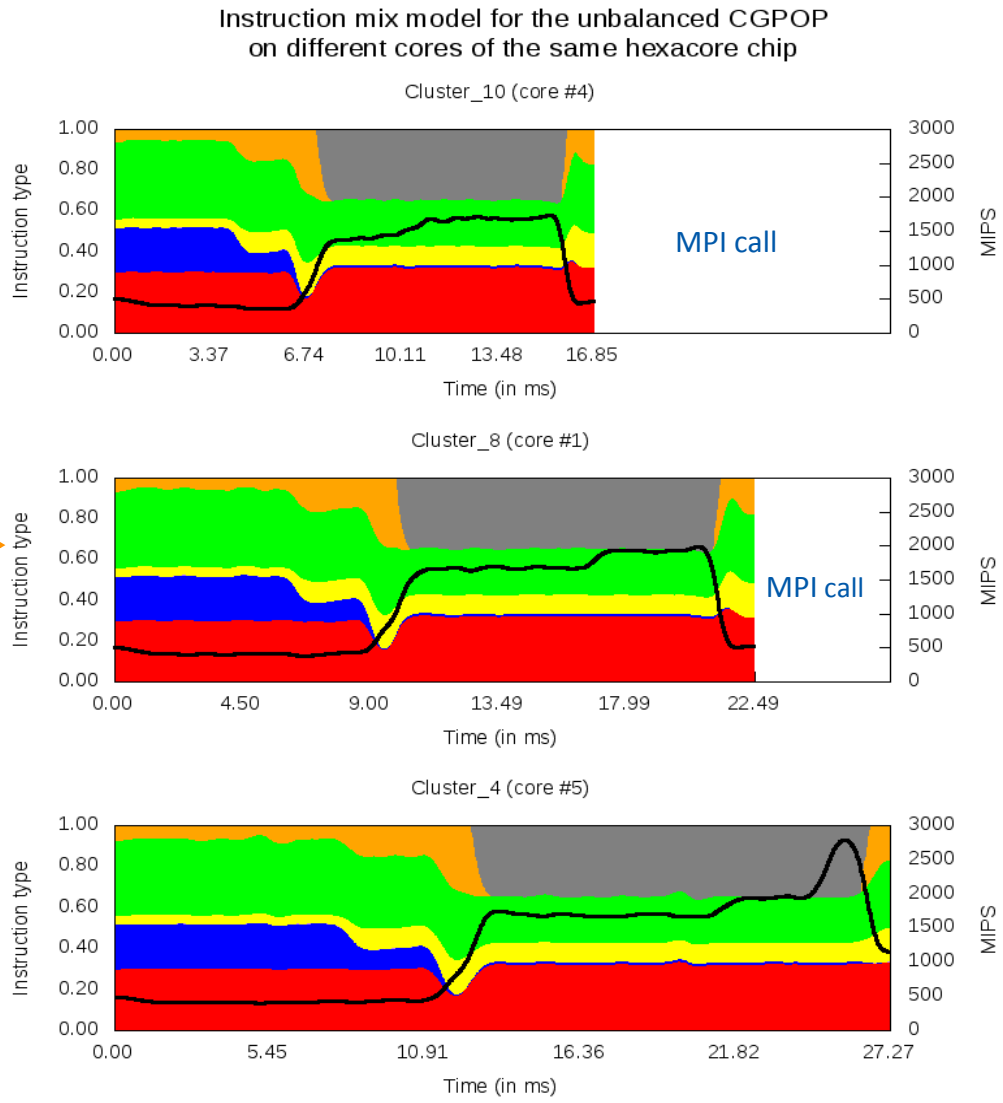
- From folded samples of a few levels to timeline structure of “relevant” routines

Recommendation without access to source code



CG-POP multicore MN3 study

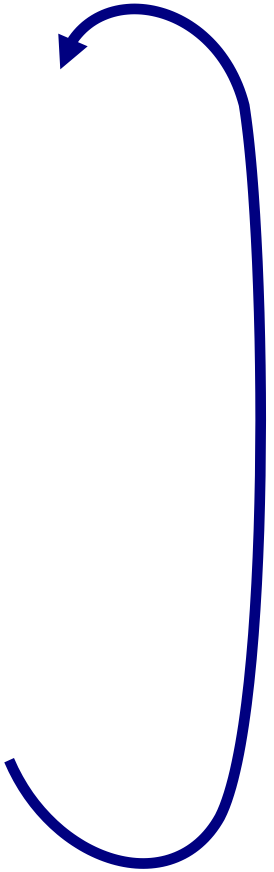
- Unbalanced MPI application
- Same code
- Different duration
- Different performance



Methodology

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Performance analysis tools objective



Help generate hypotheses

Help validate hypotheses

Qualitatively

Quantitatively

First steps

- Parallel efficiency – percentage of time invested on computation
  - Identify sources for “inefficiency”:
    - load balance
    - Communication /synchronization
- Serial efficiency – how far from peak performance?
  - IPC, correlate with other counters
- Scalability – code replication?
  - Total #instructions
- Behavioral structure? Variability?

BSC Tools web site

- tools.bsc.es
  - downloads
    - Sources / Binaries
    - Linux / windows / MAC
  - documentation
    - Training guides
    - Tutorial slides

- Getting started
  - Start wxparaver
  - Help → tutorials and follow instructions
  - Follow training guides
    - Paraver introduction (MPI): Navigation and basic understanding of Paraver operation