BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Understanding applications with Paraver

Judit Gimenez
judit@bsc.es

Sept 16-20, 2019

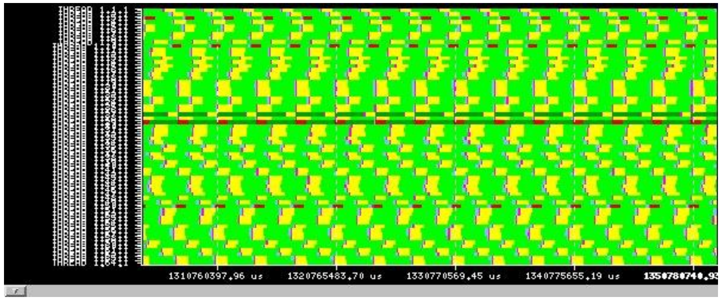
ANF Evaluation workshop

Humans are visual creatures

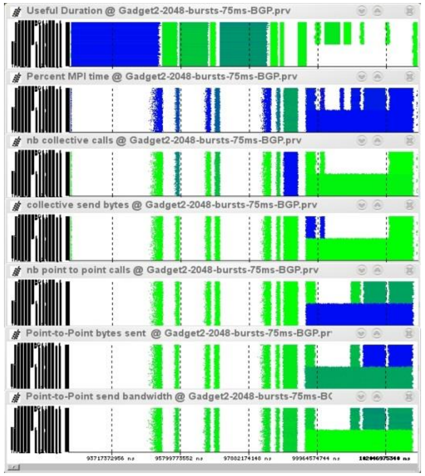
- Films or books? PROCESS
  - Two hours vs. days (months)
- Memorizing a deck of playing cards STORE
  - Each card translated to an image (person, action, location)
- Our brain loves pattern recognition IDENTIFY
  - What do you see on the pictures?

Our Tools

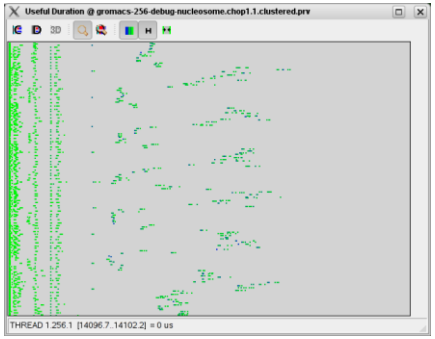
- Since 1991
- Based on traces
- Open Source
- http://tools.bsc.es



- Core tools:
  - Paraver (paramedir) – offline trace analysis
  - Dimemas – message passing simulator
  - Extrae – instrumentation


Useful Duration @ Gadget2-2048-bursts-75ms-BGP.prv
Percent MPI time @ Gadget2-2048-bursts-75ms-BGP.prv
nb collective calls @ Gadget2-2048-bursts-75ms-BGP.prv
collective send bytes @ Gadget2-2048-bursts-75ms-BGP.prv
nb point to point calls @ Gadget2-2048-bursts-75ms-BGP.prv
Point-to-Point bytes sent @ Gadget2-2048-bursts-75ms-BGP.pr
Point-to-Point send bandwidth @ Gadget2-2048-bursts-75ms-BC

- Focus
  - Detail, variability, flexibility
  - Behavioral structure vs. syntactic structure
  - Intelligence: Performance Analytics

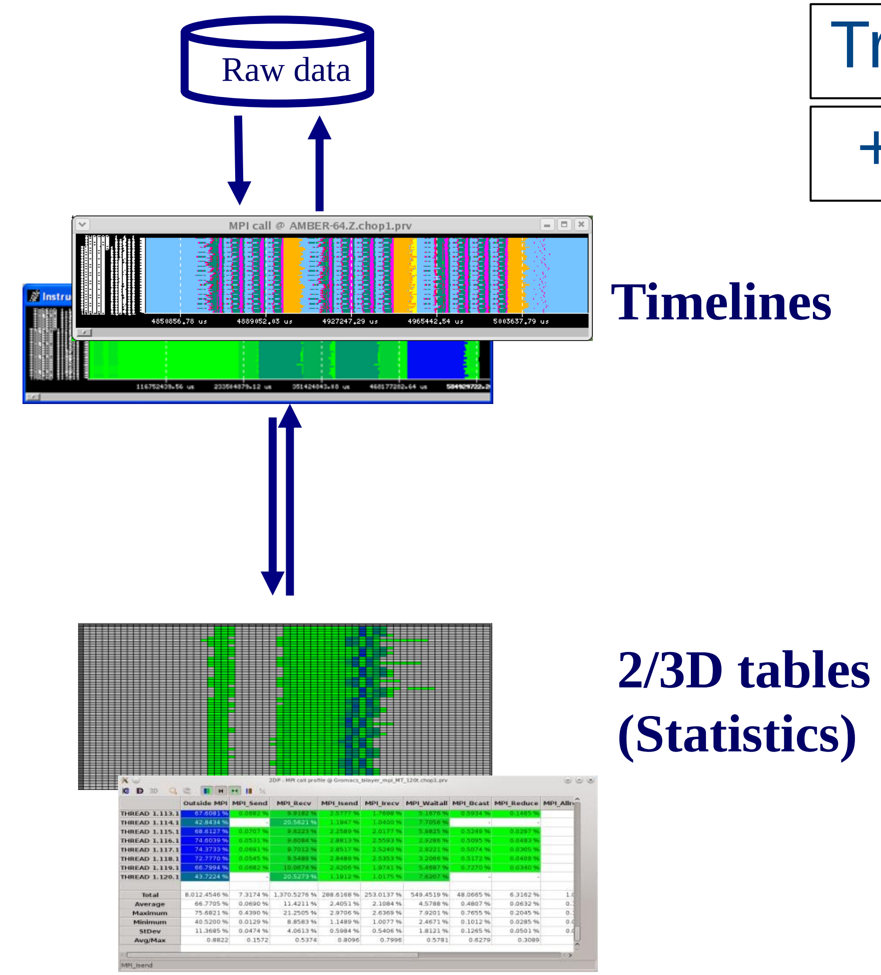

Useful Duration @ gromacs-256-debug-nucleosome.chop1.1.clustered.prv
3D
THREAD 1.256.1 [14096.7..14102.2] = 0 us

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Paraver

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Paraver – Performance data browser



Raw data

Timelines

2/3D tables (Statistics)

2DP - MPI call profile @ Gromacs_bilayer_mpi_MT_120t.chop1.prv

	Outside MPI	MPI_Send	MPI_Recv	MPI_Isend	MPI_Irecv	MPI_Waitall	MPI_Bcast	MPI_Reduce	MPI_Allr
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 %	9.8223 %	2.2589 %	2.0177 %	5.9825 %	0.5249 %	0.0297 %	
THREAD 1.116.1	74.6039 %	0.0531 %	9.8084 %	2.8813 %	2.5593 %	2.9286 %	0.5095 %	0.0483 %	
THREAD 1.117.1	74.3733 %	0.0691 %	9.7012 %	2.8517 %	2.5240 %	2.9221 %	0.5074 %	0.0305 %	
THREAD 1.118.1	72.7770 %	0.0545 %	9.5489 %	2.8489 %	2.5353 %	3.2066 %	0.5172 %	0.0409 %	
THREAD 1.119.1	66.7994 %	0.0682 %	10.0674 %	2.4206 %	1.9741 %	5.4687 %	0.7270 %	0.0340 %	
THREAD 1.120.1	43.7224 %	-	20.5273 %	1.1912 %	1.0175 %	7.6207 %	-	-	
Total	8,012.4546 %	7.3174 %	1,370.5276 %	288.6168 %	253.0137 %	549.4519 %	48.0665 %	6.3162 %	1.0
Average	66.7705 %	0.0690 %	11.4211 %	2.4051 %	2.1084 %	4.5788 %	0.4807 %	0.0632 %	0.
Maximum	75.6821 %	0.4390 %	21.2505 %	2.9706 %	2.6369 %	7.9201 %	0.7655 %	0.2045 %	0.
Minimum	40.5200 %	0.0129 %	8.8583 %	1.1489 %	1.0077 %	2.4671 %	0.1012 %	0.0285 %	0.0
StDev	11.3685 %	0.0474 %	4.0613 %	0.5984 %	0.5406 %	1.8121 %	0.1265 %	0.0501 %	0.0
Avg/Max	0.8822	0.1572	0.5374	0.8096	0.7996	0.5781	0.6279	0.3089	

MPI_Isend

Trace visualization/analysis

+ trace manipulation

Goal = Flexibility

No semantics

Programmable

Comparative analyses

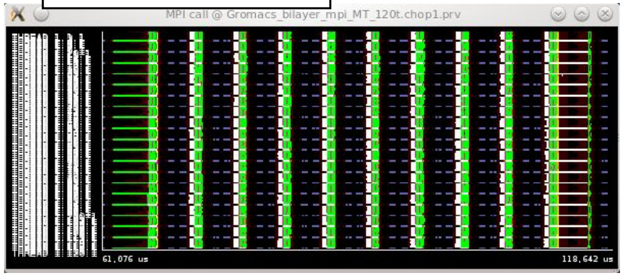
Multiple traces

Synchronize scales

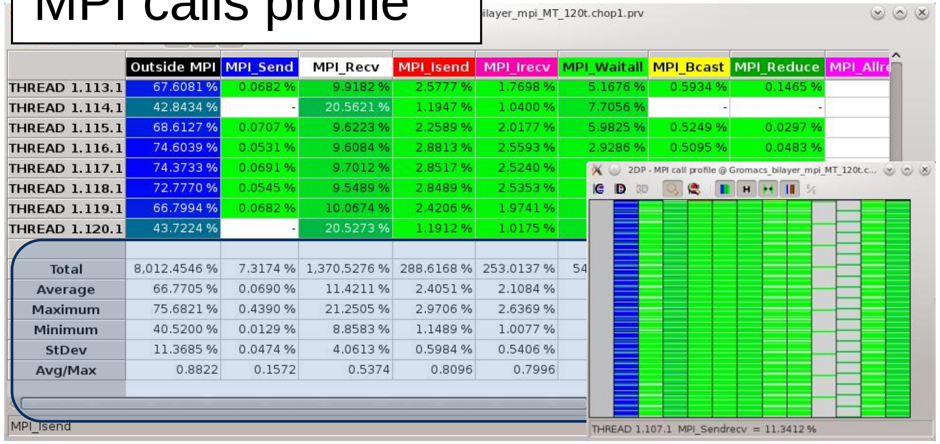
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

From timelines to tables

MPI calls



MPI calls profile

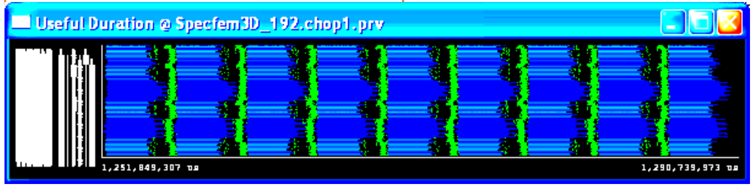

ilayer_mpi_MT_120t.chop1.prv
2DP - MPI call profile @ Gromacs_bilayer_mpi_MT_120t.c...
3D
H

	Outside MPI	MPI_Send	MPI_Recv	MPI_Isend	MPI_Irecv	MPI_Waitall	MPI_Bcast	MPI_Reduce	MPI_Allre
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				

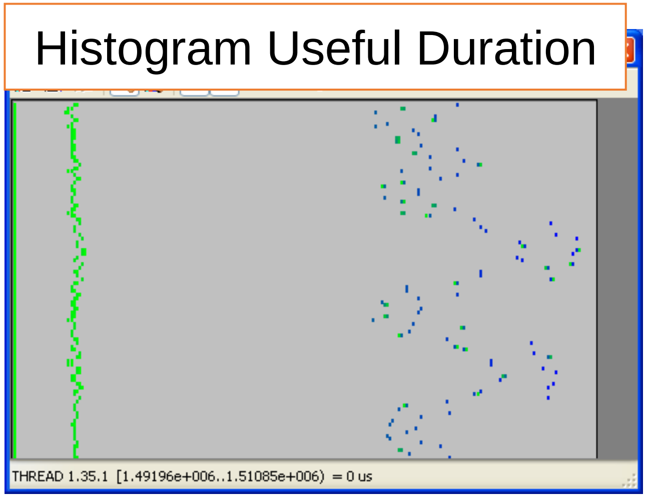
MPI_Isend

THREAD 1.107.1 MPI_Sendrecv = 11.3412 %

Useful Duration

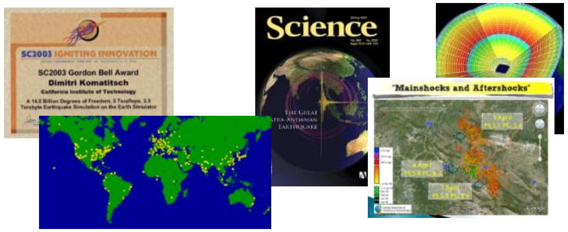


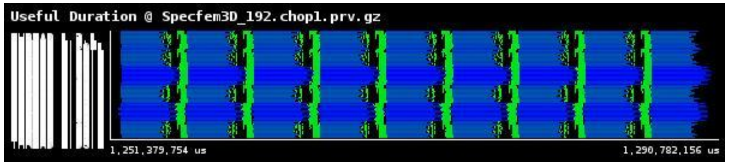
Histogram Useful Duration

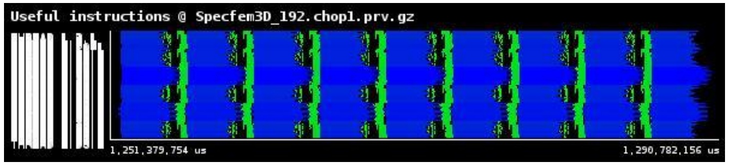


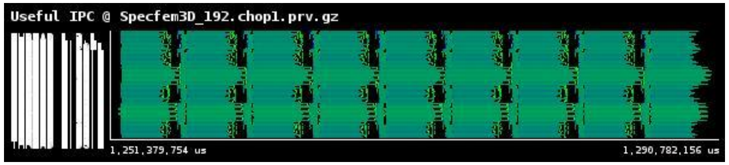
THREAD 1.35.1 [1.49196e+006..1.51085e+006) = 0 us

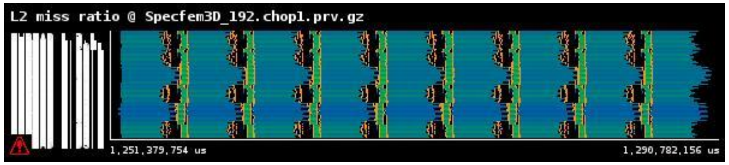
Analyzing variability

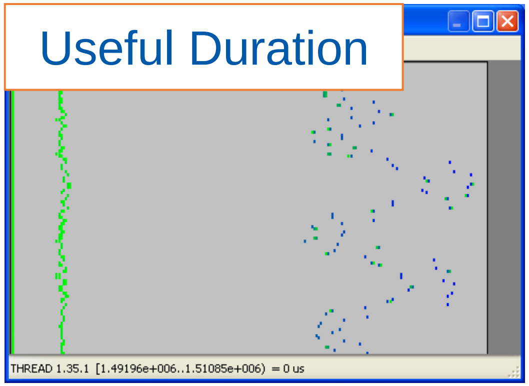

SC2003 IGNITING INNOVATION
SC2003 Gordon Bell Award
Dimitri Komatitsch
Science
THE GREAT SUMATRA-ANDAMAN EARTHQUAKE
"Mainshocks and Aftershocks"


Useful Duration @ Specfem3D_192.chop1.prv.gz

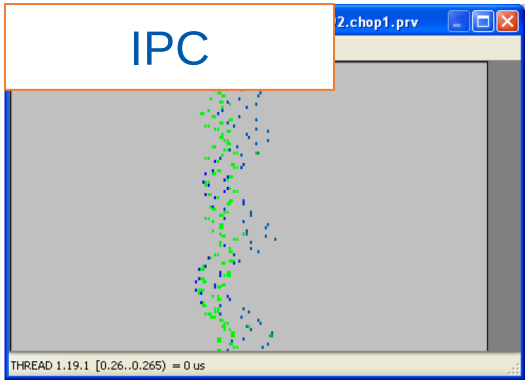

Useful instructions @ Specfem3D_192.chop1.prv.gz


Useful IPC @ Specfem3D_192.chop1.prv.gz

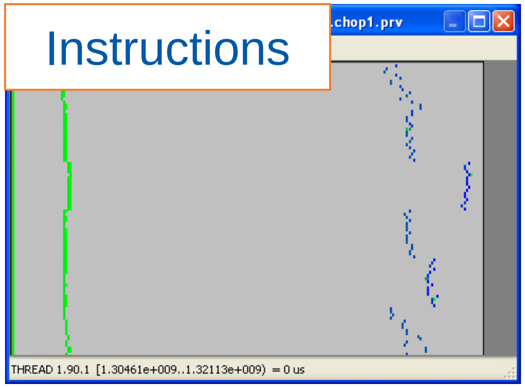

L2 miss ratio @ Specfem3D_192.chop1.prv.gz


Useful Duration

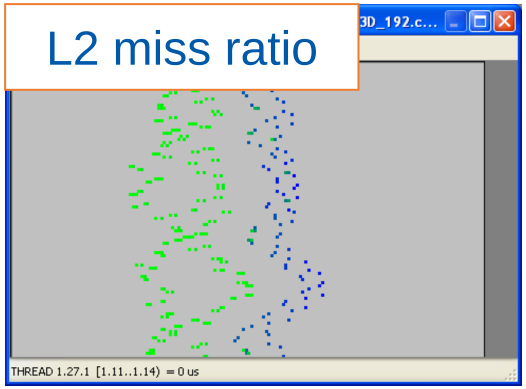
THREAD 1.35.1 [1.49196e+006..1.51085e+006) = 0 us


IPC

THREAD 1.19.1 [0.26..0.265) = 0 us


Instructions

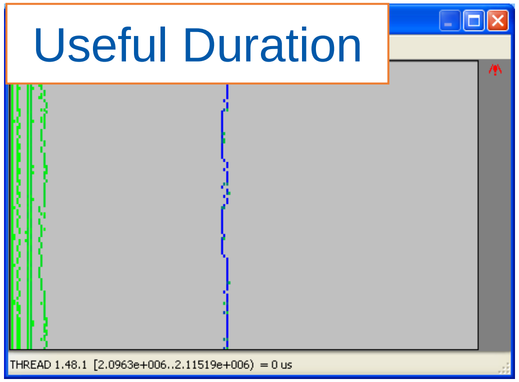
THREAD 1.90.1 [1.30461e+009..1.32113e+009) = 0 us


L2 miss ratio

THREAD 1.27.1 [1.11..1.14) = 0 us

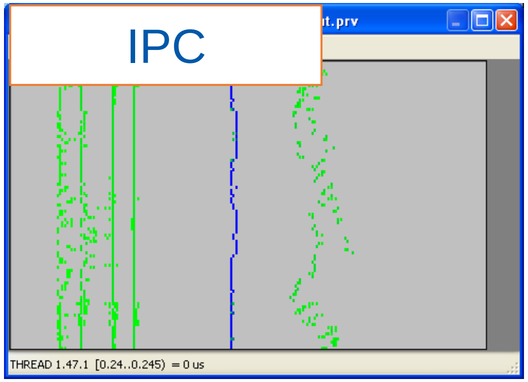
Analyzing variability

- By the way: six months later ….



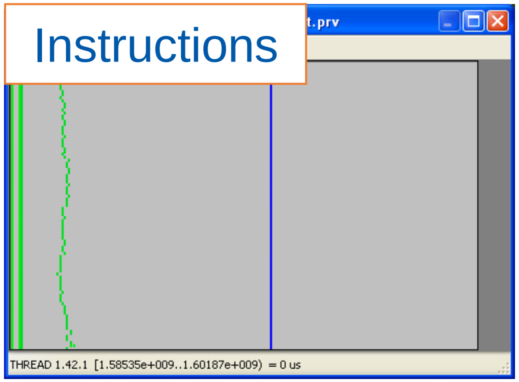
Useful Duration

```
THREAD 1.48.1 [2.0963e+006..2.11519e+006) = 0 us
```



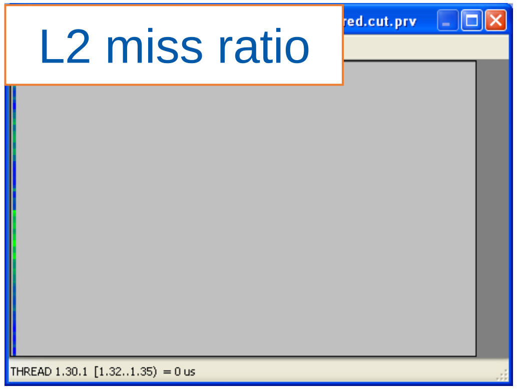
IPC

```
THREAD 1.47.1 [0.24..0.245) = 0 us
```



Instructions

```
THREAD 1.42.1 [1.58535e+009..1.60187e+009) = 0 us
```



L2 miss ratio

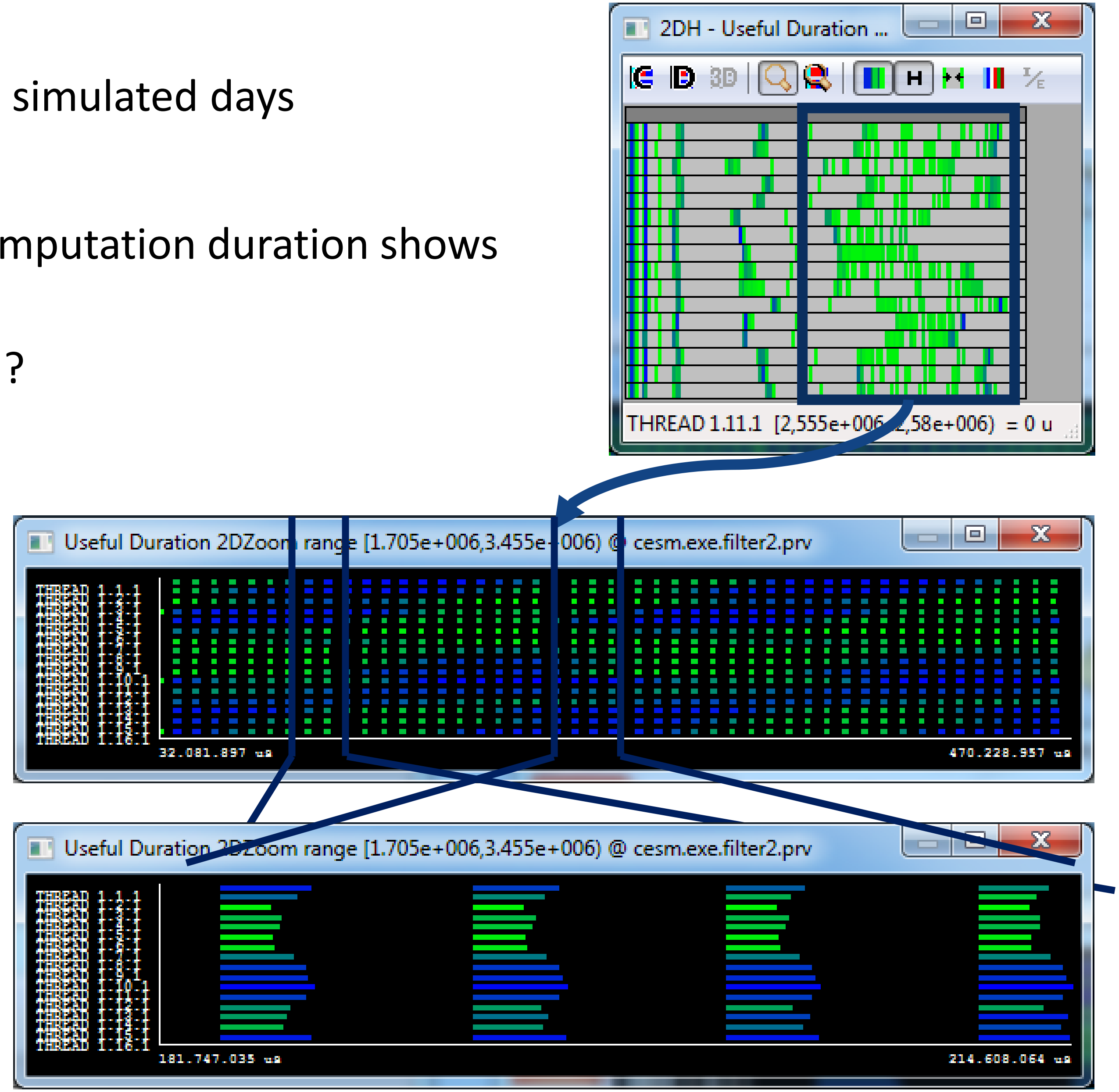
```
THREAD 1.30.1 [1.32..1.35) = 0 us
```

From tables to timelines

CESM: 16 processes, 2 simulated days

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

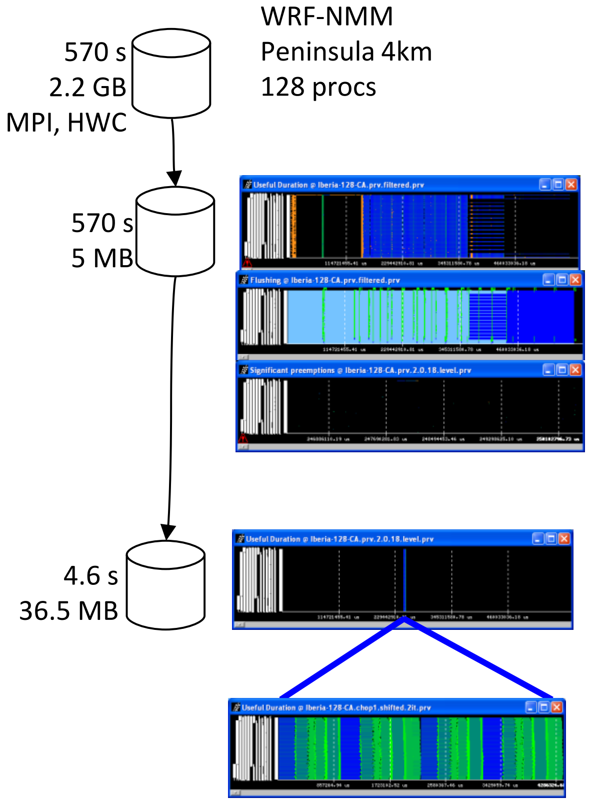
- Dynamic imbalance
  - In space and time
  - Day and night.
  - Season ? ☺


2DH - Useful Duration ...
3D
H
THREAD 1.11.1 [2,555e+006,2,58e+006) = 0 u
Useful Duration 2DZoom range [1.705e+006,3.455e+006) @ cesm.exe.filter2.prv
THREAD 1.1.1
THREAD 1.16.1
32.081.897 us
470.228.957 us
Useful Duration 2DZoom range [1.705e+006,3.455e+006) @ cesm.exe.filter2.prv
THREAD 1.1.1
THREAD 1.16.1
181.747.035 us
214.608.064 us

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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,…

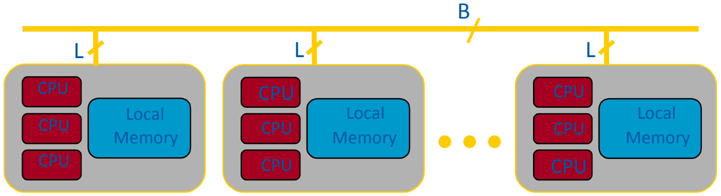

570 s
2.2 GB
MPI, HWC
WRF-NMM
Peninsula 4km
128 procs
570 s
5 MB
Useful Duration @ Iberia-128-CA.prv.filtered.prv
Flushing @ Iberia-128-CA.prv.filtered.prv
Significant preemptions @ Iberia-128-CA.prv.2.0.18.level.prv
4.6 s
36.5 MB
Useful Duration @ Iberia-128-CA.prv.2.0.18.level.prv
Useful Duration @ Iberia-128-CA.chop1.shifted.2it.prv

Dimemas

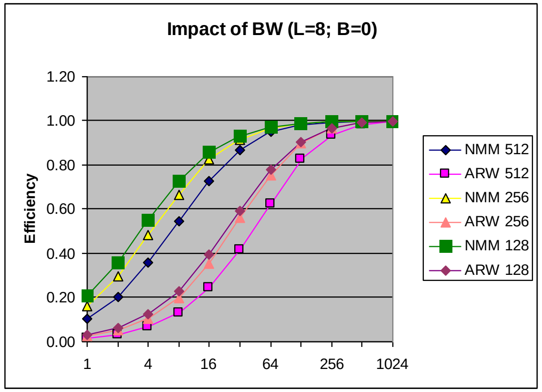
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Dimemas – Coarse grain, Trace driven simulation

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



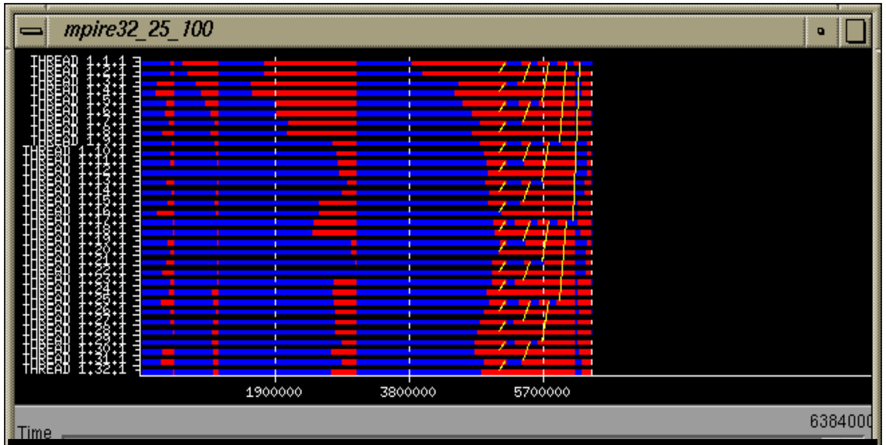
- Parametric sweeps
  - On abstract architectures
  - On application computational regions
- What if analysis
  - Ideal machine (instantaneous network)
  - Estimating impact of ports to MPI+OpenMP/CUDA/…
  - Should I use asynchronous communications?
  - Are all parts equally sensitive to network?
- MPI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



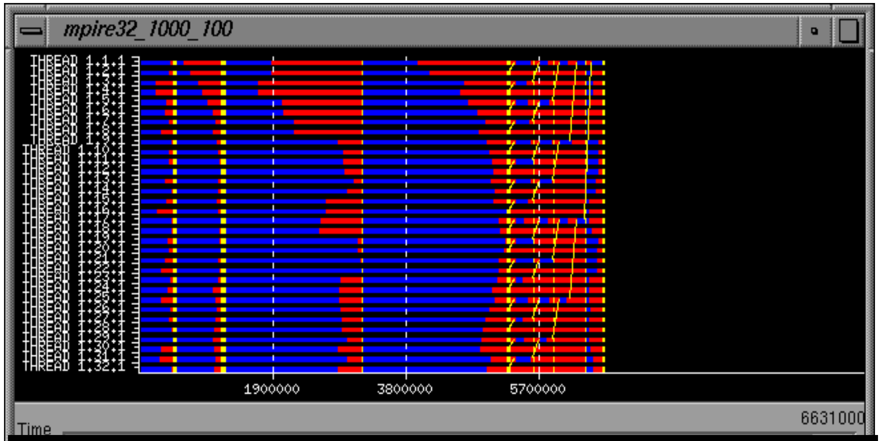
Detailed feedback on simulation (trace)

Network sensitivity

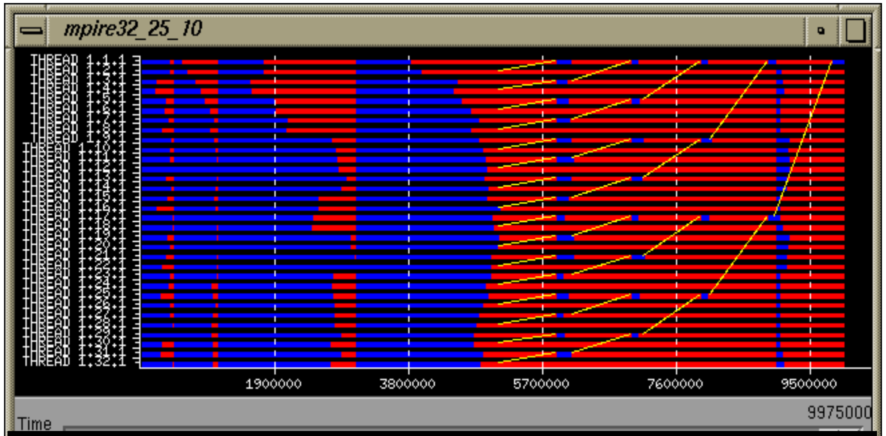
- MPIRE 32 tasks, no network contention



L = 5µs – BW = 1 GB/s



L = 1000µs – BW = 1 GB/s



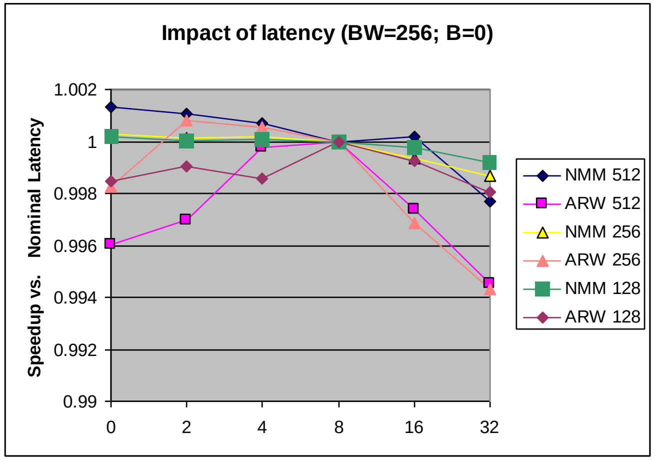
L = 5µs – BW = 100MB/s

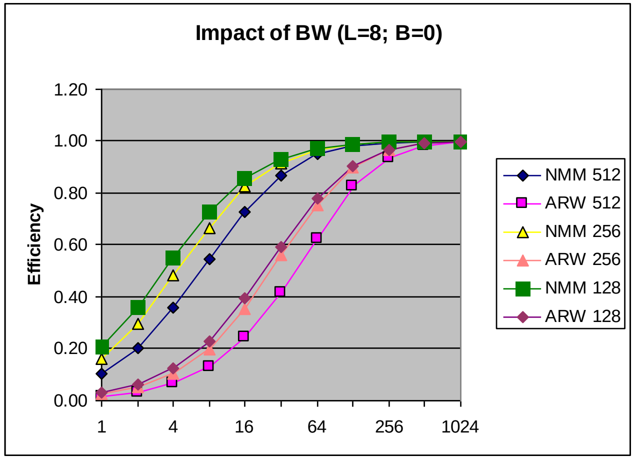
All windows same scale

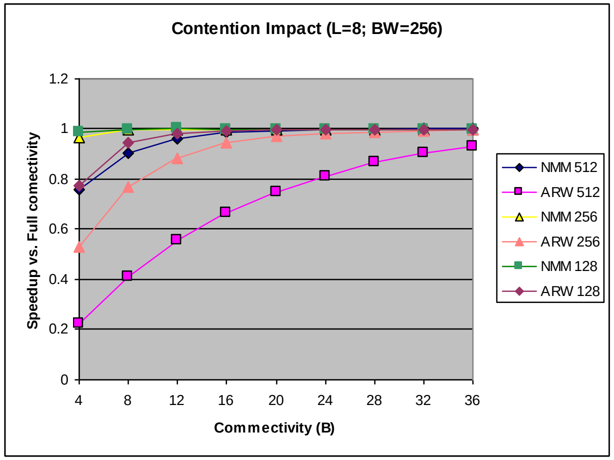
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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



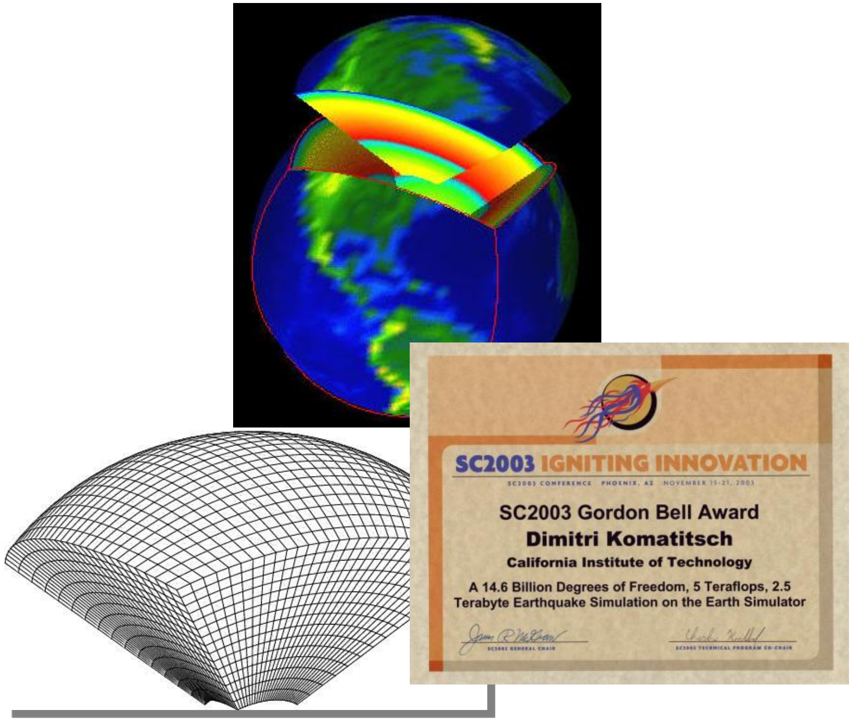




Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Would I will benefit from asynchronous communications?

SPECFEM3D



SC2003 IGNITING INNOVATION

SC2003 CONFERENCE PHOENIX, AZ NOVEMBER 15-21, 2003

SC2003 Gordon Bell Award

Dimitri Komatitsch

California Institute of Technology

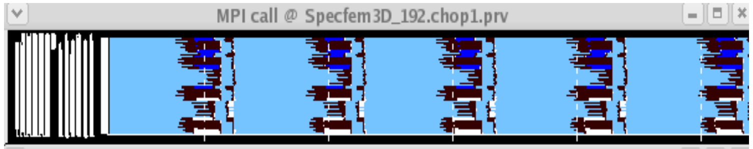
A 14.6 Billion Degrees of Freedom, 5 Teraflops, 2.5 Terabyte Earthquake Simulation on the Earth Simulator

SC2003 GENERAL CHAIR

SC2003 TECHNICAL PROGRAM CO-CHAIR

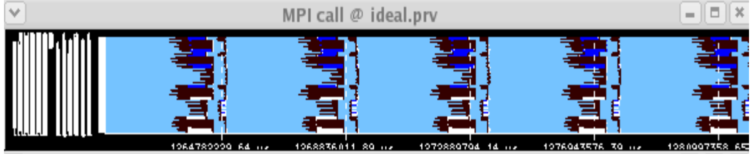
Courtesy Dimitri Komatitsch

Real



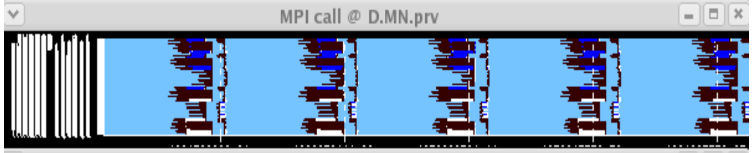
MPI call @ Specfem3D_192.chop1.prv

Ideal



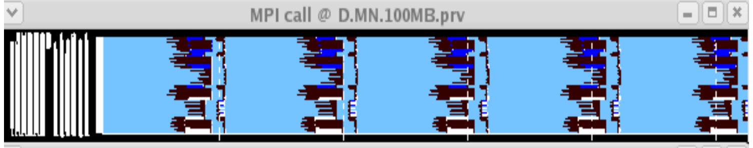
MPI call @ ideal.prv

Prediction MN



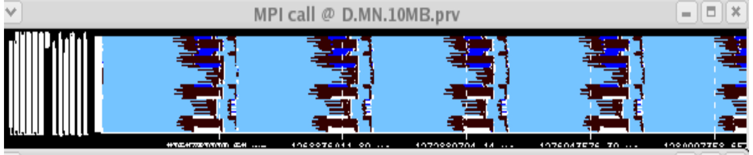
MPI call @ D.MN.prv

Prediction 100MB/s



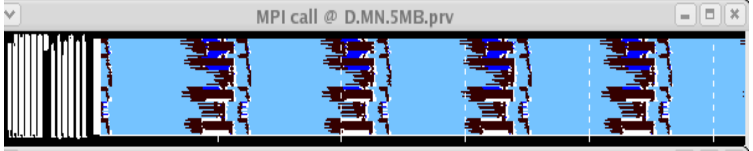
MPI call @ D.MN.100MB.prv

Prediction 10MB/s



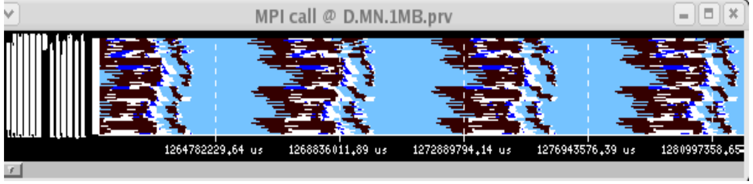
MPI call @ D.MN.10MB.prv

Prediction 5MB/s



MPI call @ D.MN.5MB.prv

Prediction 1MB/s



MPI call @ D.MN.1MB.prv

Ideal machine

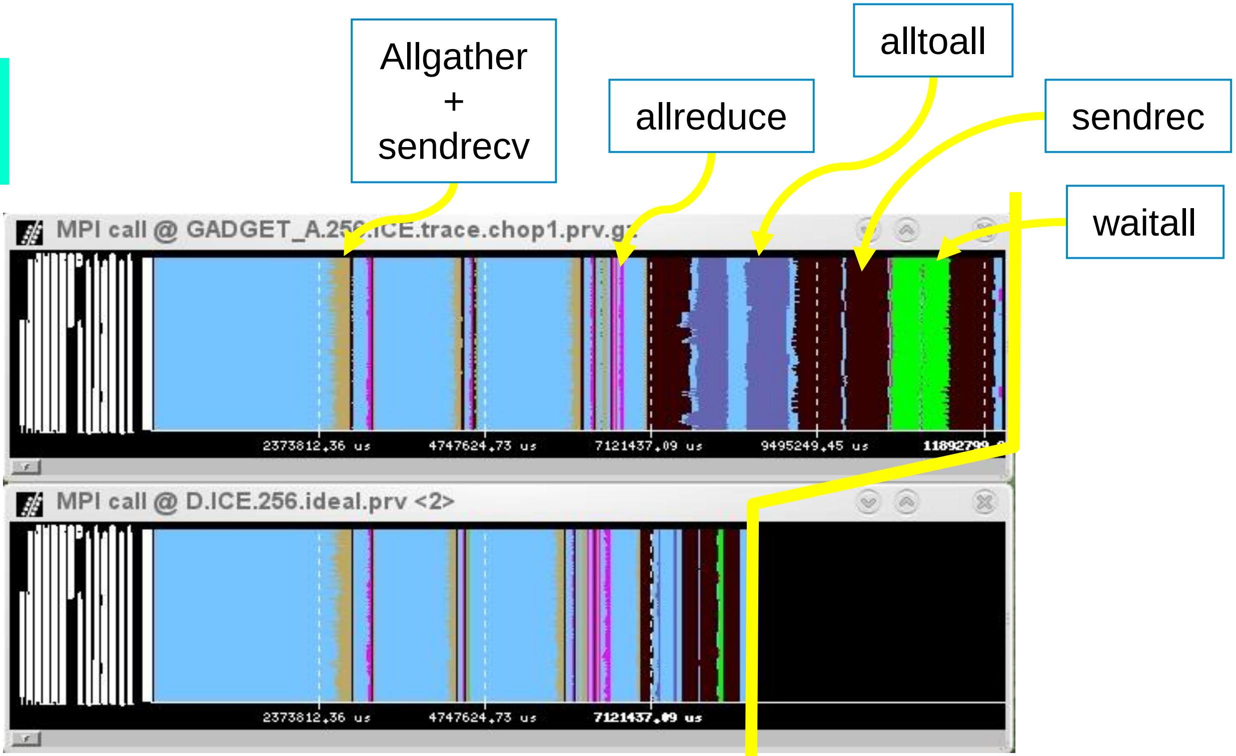
The impossible machine: BW = ∞, L = 0

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

GADGET @ Nehalem cluster
256 processes

Real run

Ideal network



Allgather + sendrecv

allreduce

alltoall

sendrec

waitall

MPI call @ GADGET_A.256.ICE.trace.chop1.prv.gz

2373812,36 us 4747624,73 us 7121437,09 us 9495249,45 us 11892799

MPI call @ D.ICE.256.ideal.prv <2>

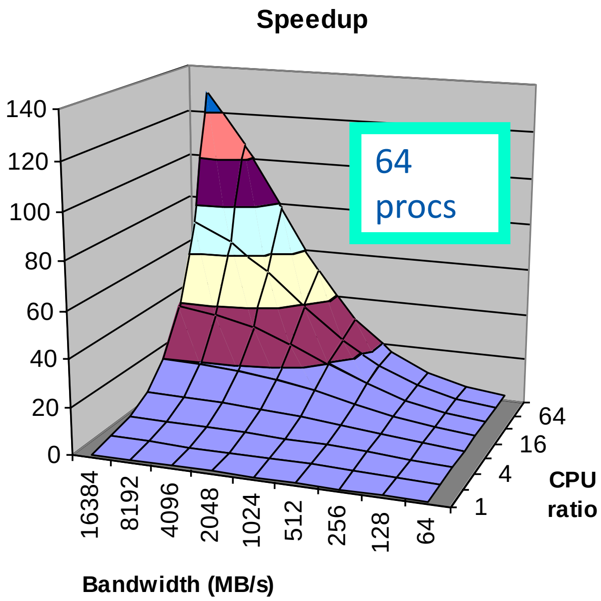
2373812,36 us 4747624,73 us 7121437,09 us

Impact on practical machines?

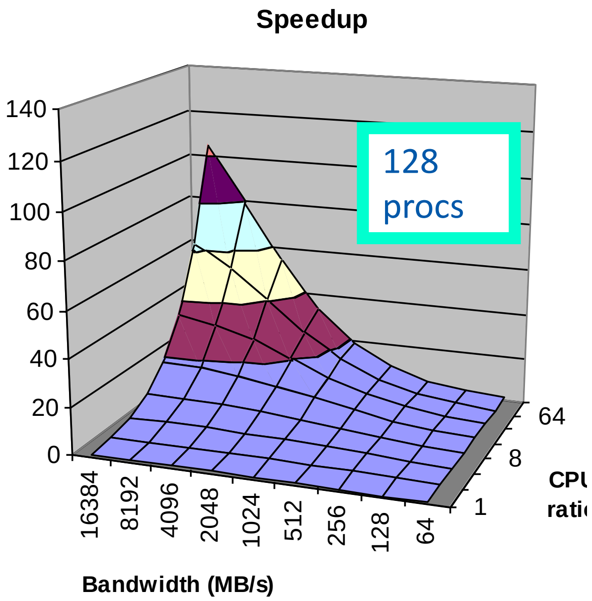
Impact of architectural parameters

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

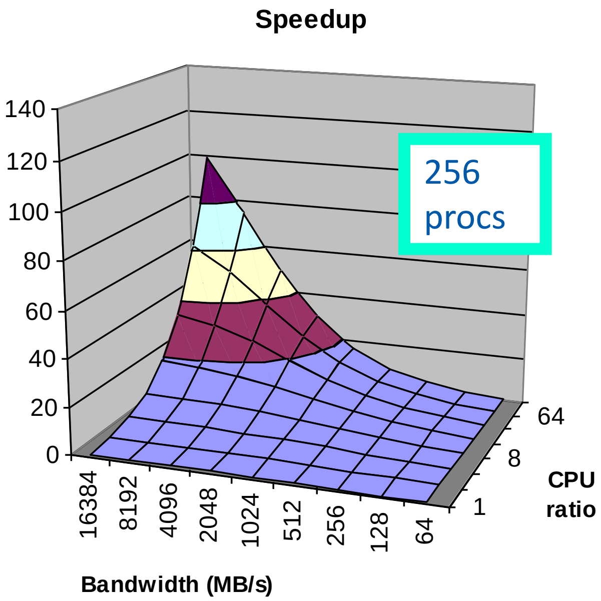
GADGET


Speedup

64 procs


Speedup

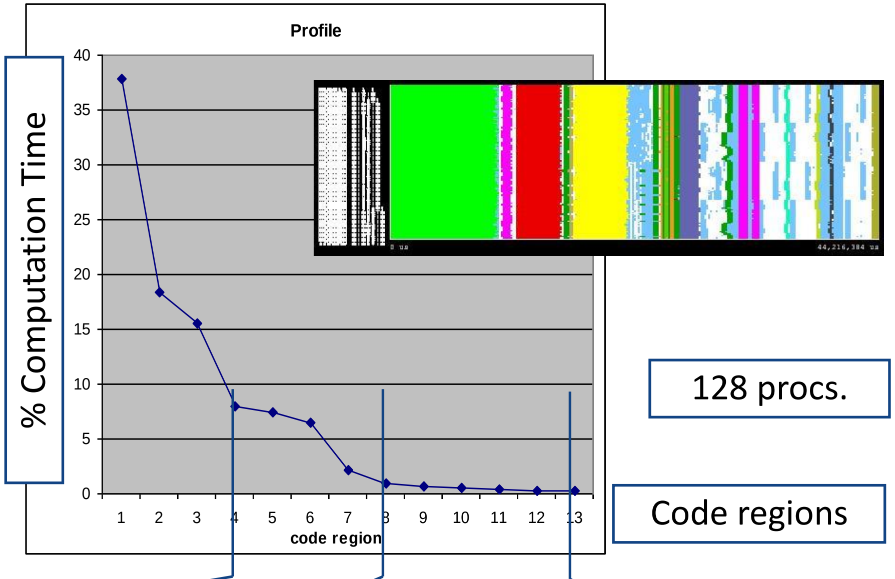
128 procs

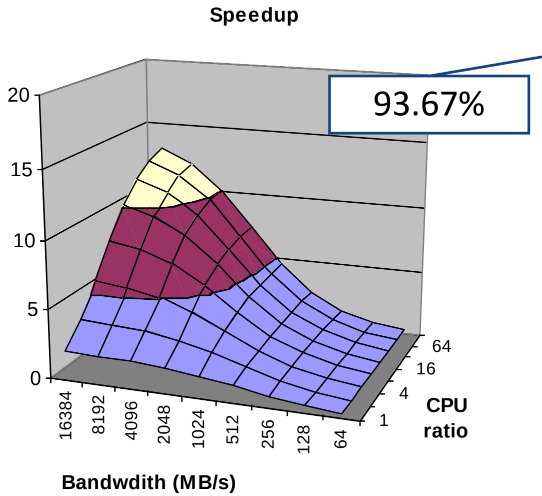

Speedup

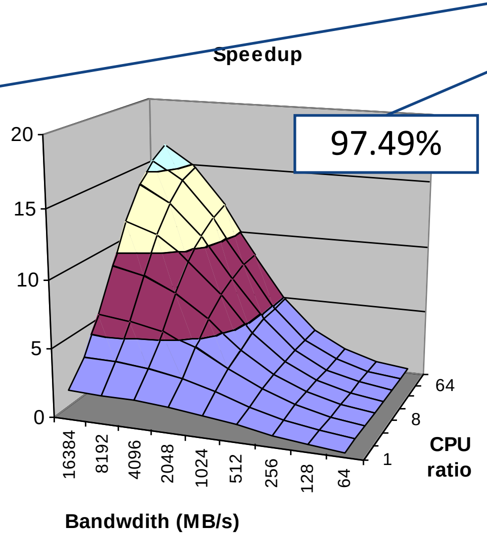
256 procs

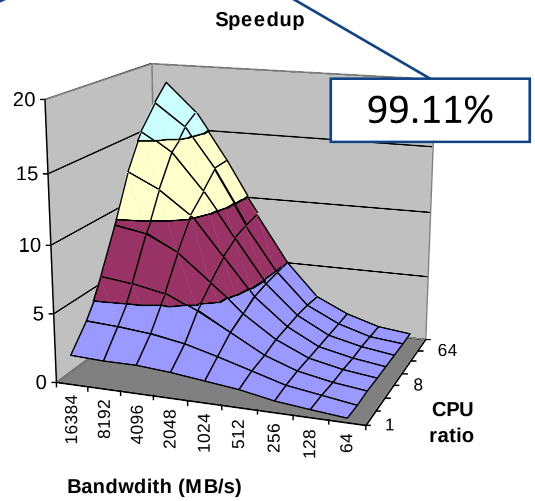
Hybrid parallelization

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







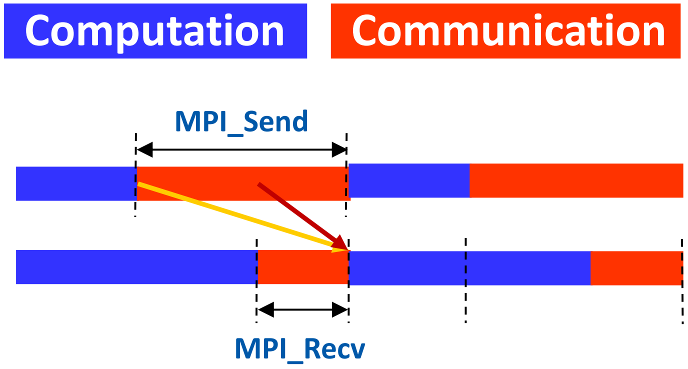


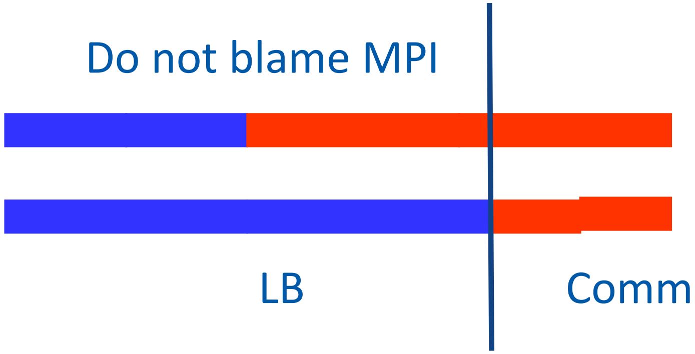
(Previous slide: speedups **up to 100x**)

Efficiency Model

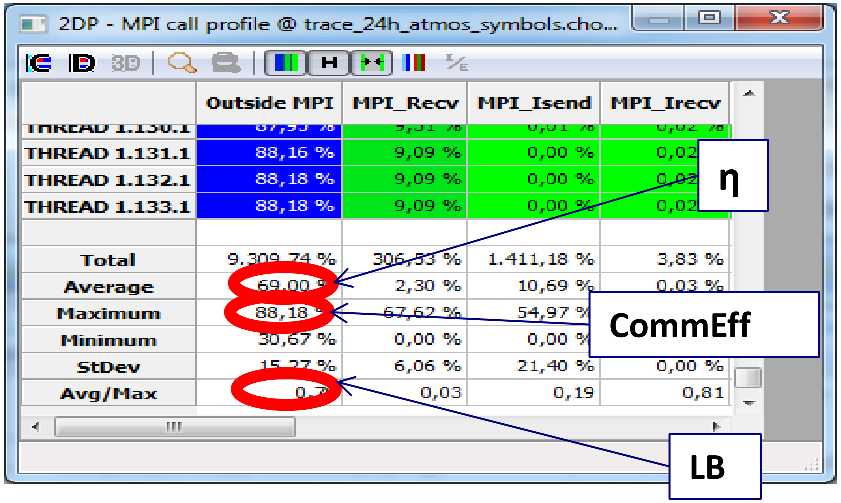
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Parallel efficiency model





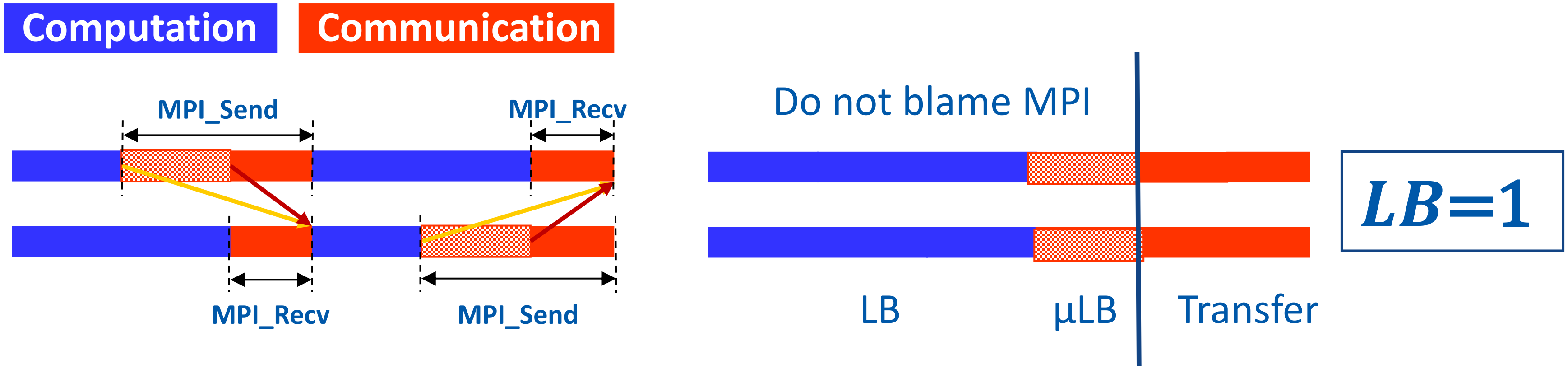
- Parallel efficiency = LB eff * Comm eff



2DP - MPI call profile @ trace_24h_atmos_symbols.cho...

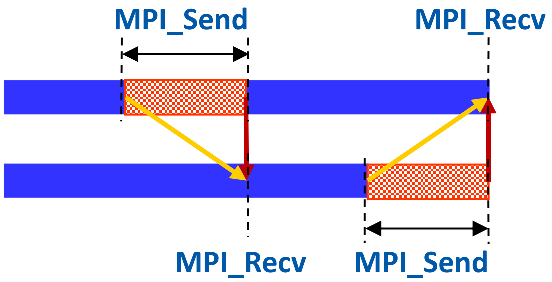
	Outside MPI	MPI_Recv	MPI_Isend	MPI_Irecv
THREAD 1.130.1	87,95 %	9,31 %	0,01 %	0,02 %
THREAD 1.131.1	88,16 %	9,09 %	0,00 %	0,02
THREAD 1.132.1	88,18 %	9,09 %	0,00 %	0,02
THREAD 1.133.1	88,18 %	9,09 %	0,00 %	0,02
Total	9.309,74 %	306,53 %	1.411,18 %	3,83 %
Average	69,00 %	2,30 %	10,69 %	0,03 %
Maximum	88,18	67,62 %	54,97 %	
Minimum	30,67 %	0,00 %	0,00 %	
StDev	15,27 %	6,06 %	21,40 %	0,00 %
Avg/Max	0,7	0,03	0,19	0,81

Parallel efficiency refinement: LB * µLB * Tr



$$\boxed{LB=1}$$

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

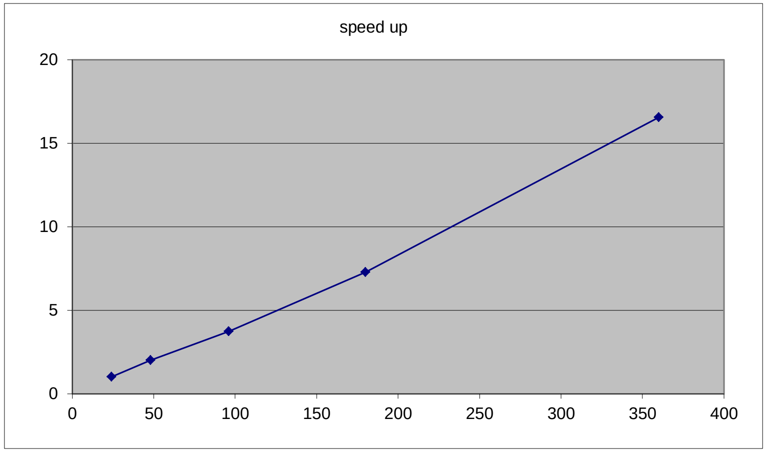


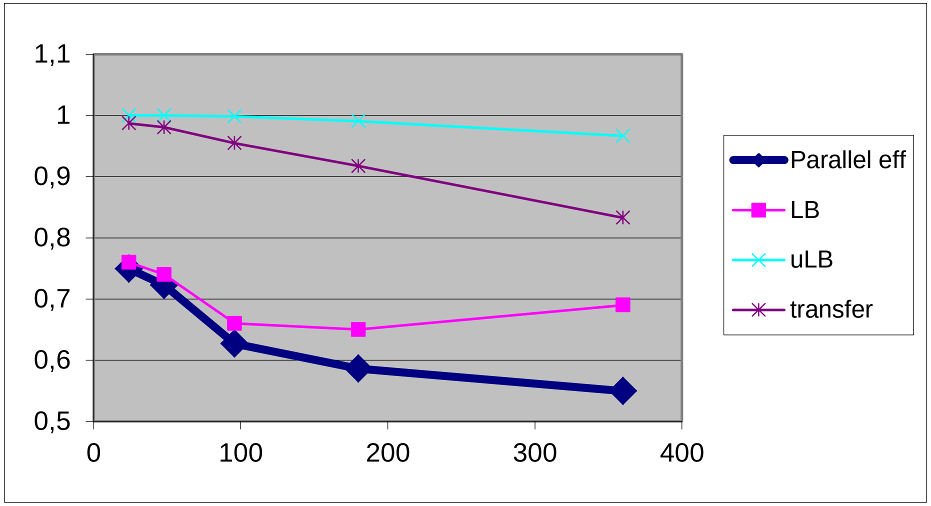
Why scaling?

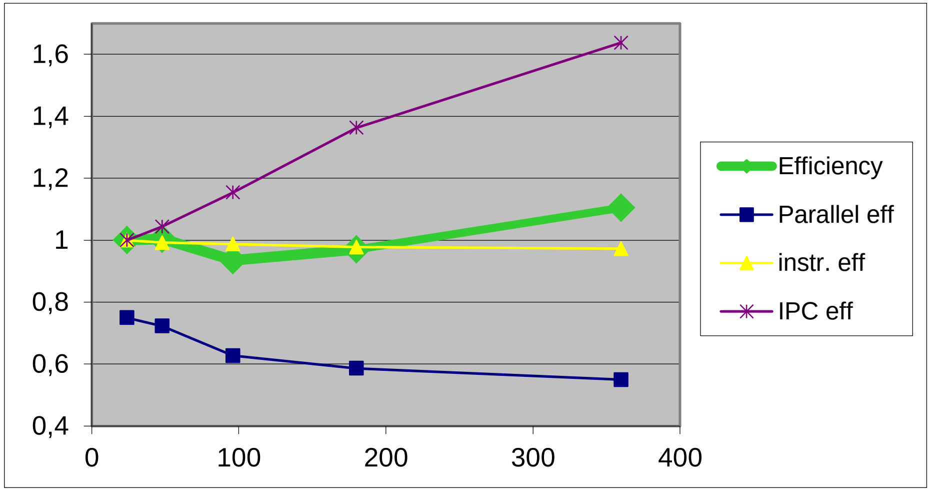
$$\eta_{\parallel} = LB * Ser * Trf$$

CG-POP mpi2s1D - 180x120

Good scalability !!
Should we be happy?







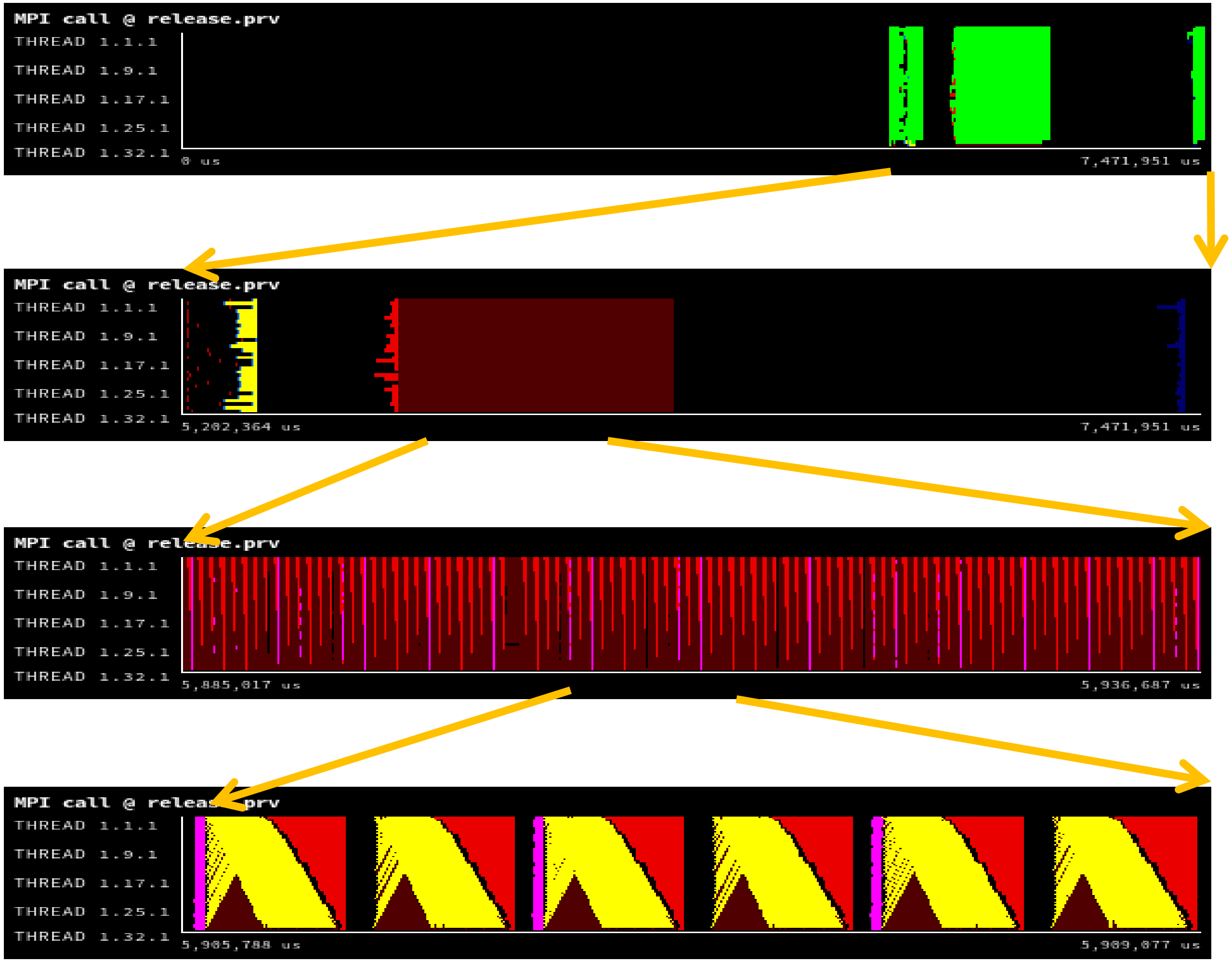
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Why efficient?

Parallel efficiency =93.28
Communication = 93.84

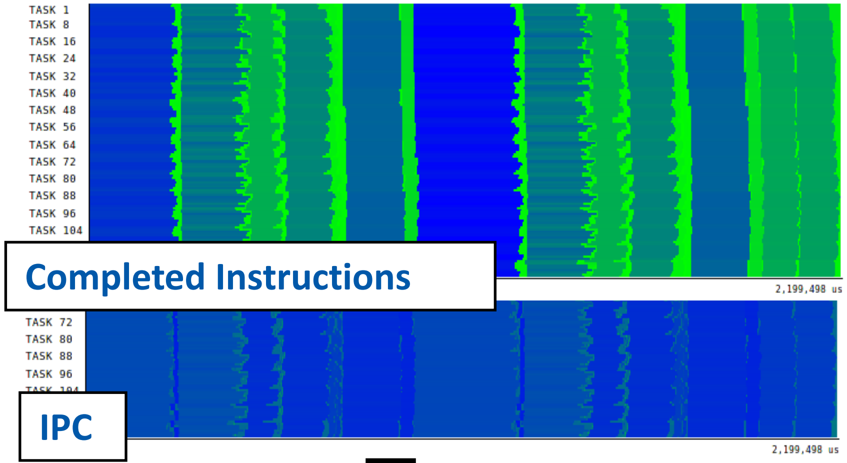
Parallel efficiency = 77.93
Communication = 79.79

Parallel efficiency = 28.84
Communication eff = 30.42



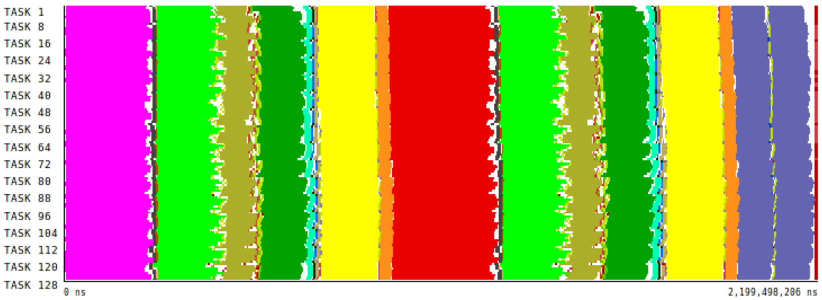
Analytics

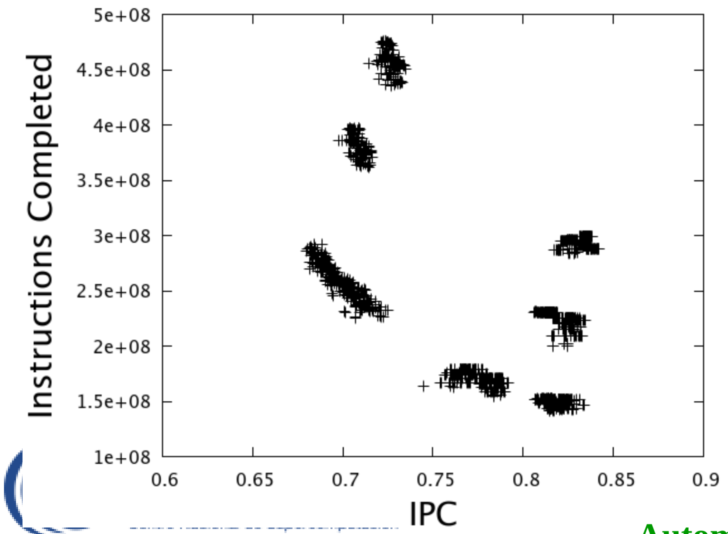
Using Clustering to identify structure

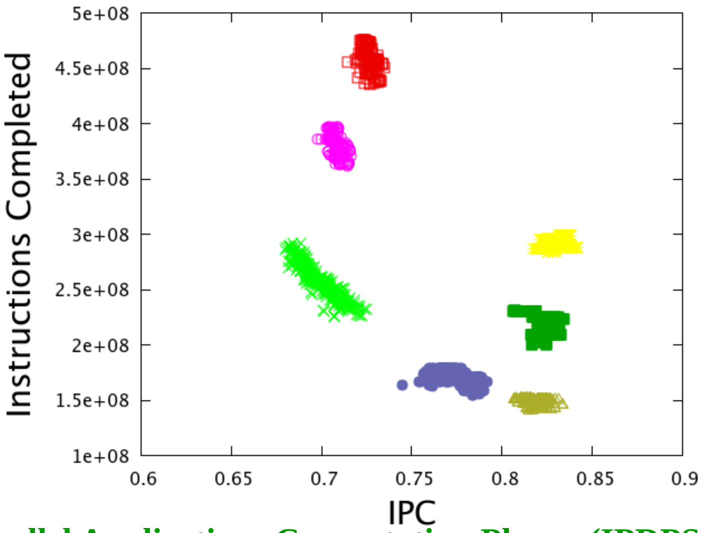


Completed Instructions

IPC





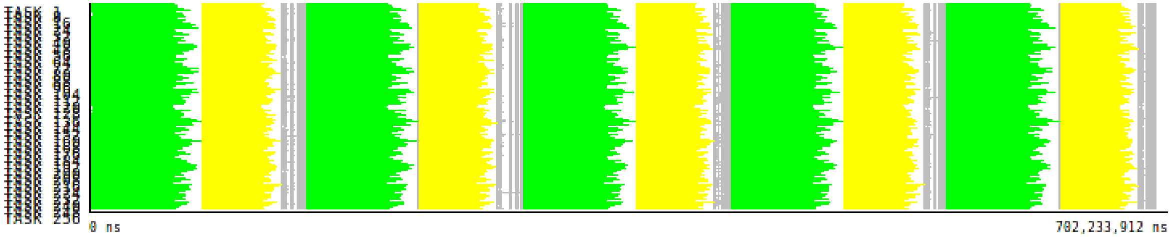


Automatic Detection of Parallel Applications Computation Phases (IPDPS 2009)

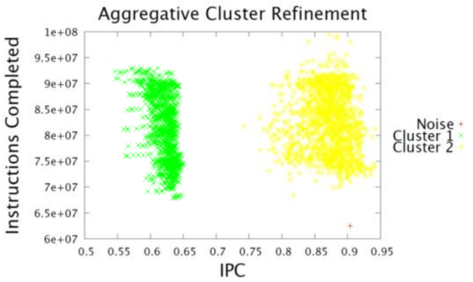
What should I improve?

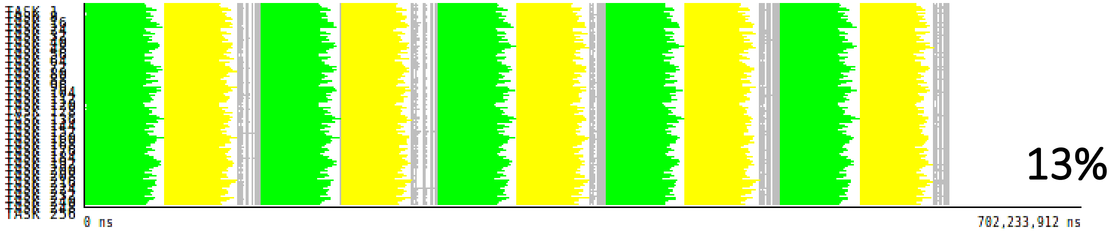
What if ….

PEPC



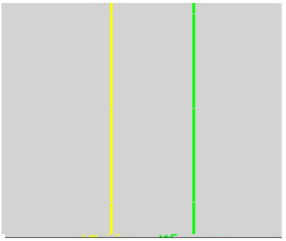
… we increase the IPC of Cluster1?

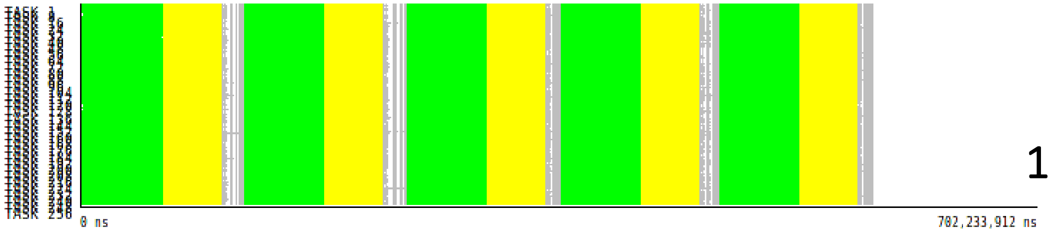




13% gain

… we balance Clusters 1 & 2?

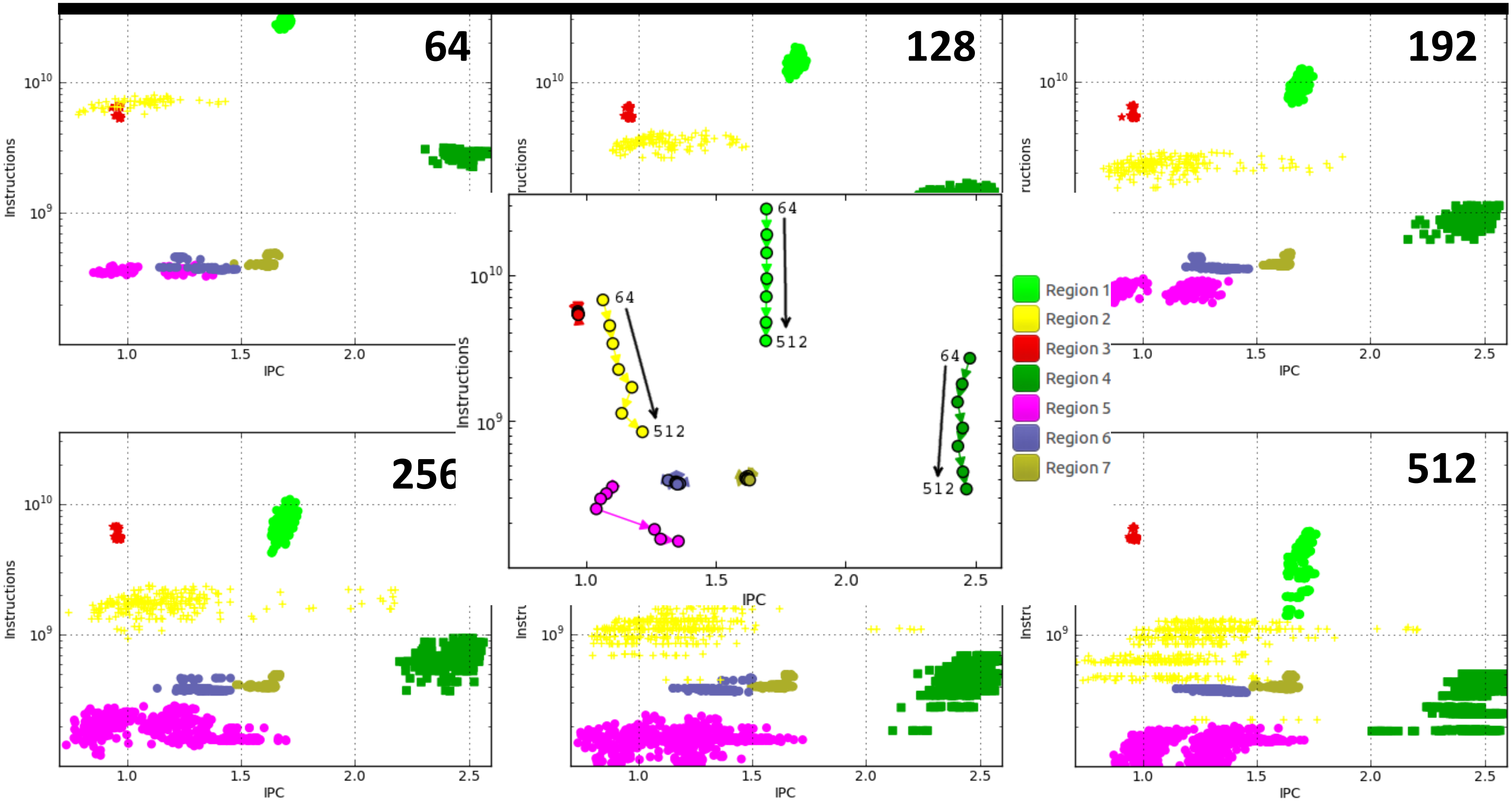




19% gain

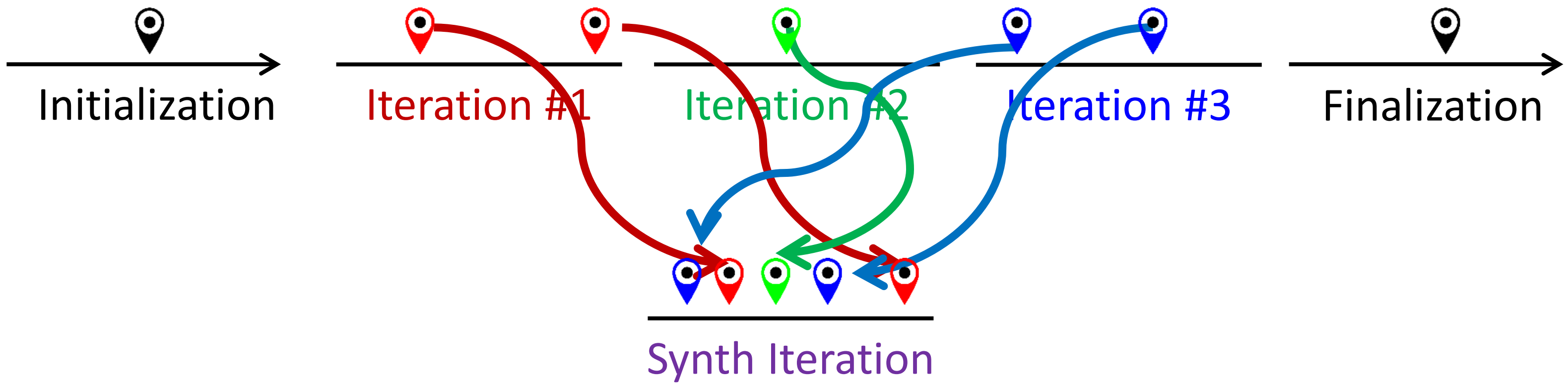
Tracking scability through clustering

- OpenMX (strong scale from 64 to 512 tasks)



Folding

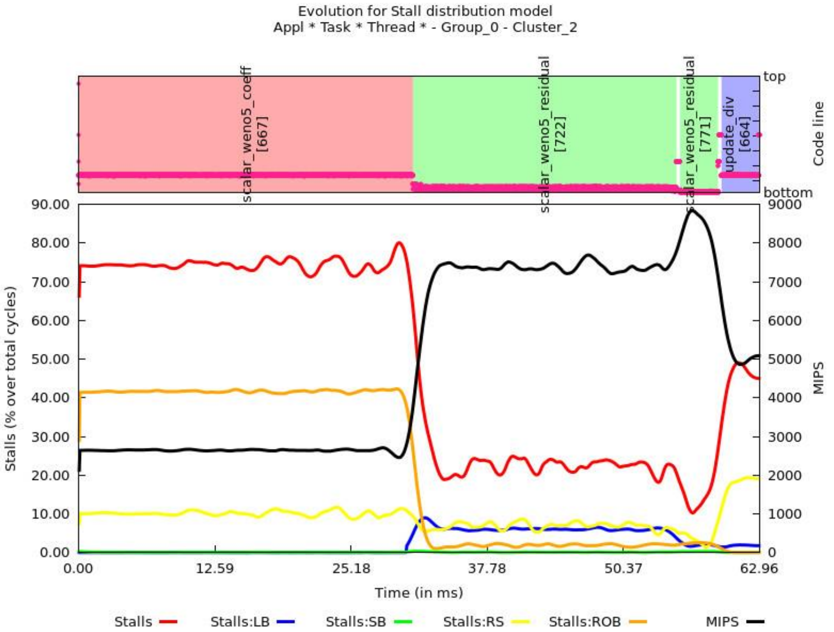
- Instantaneous metrics with minimum overhead
  - Combine instrumentation and sampling
    - Instrumentation delimits regions (routines, loops, …)
    - Sampling exposes progression within a region
  - Captures performance counters and call-stack references

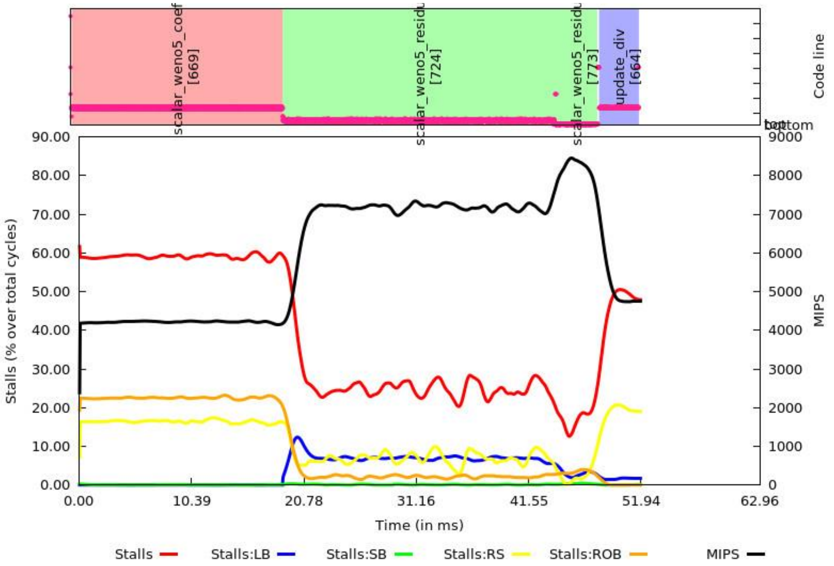


“Blind” optimization

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

Recommendation without access to source code

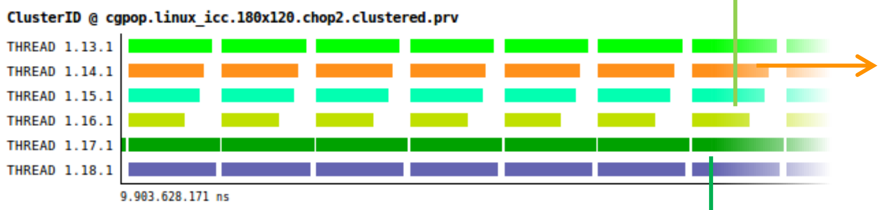




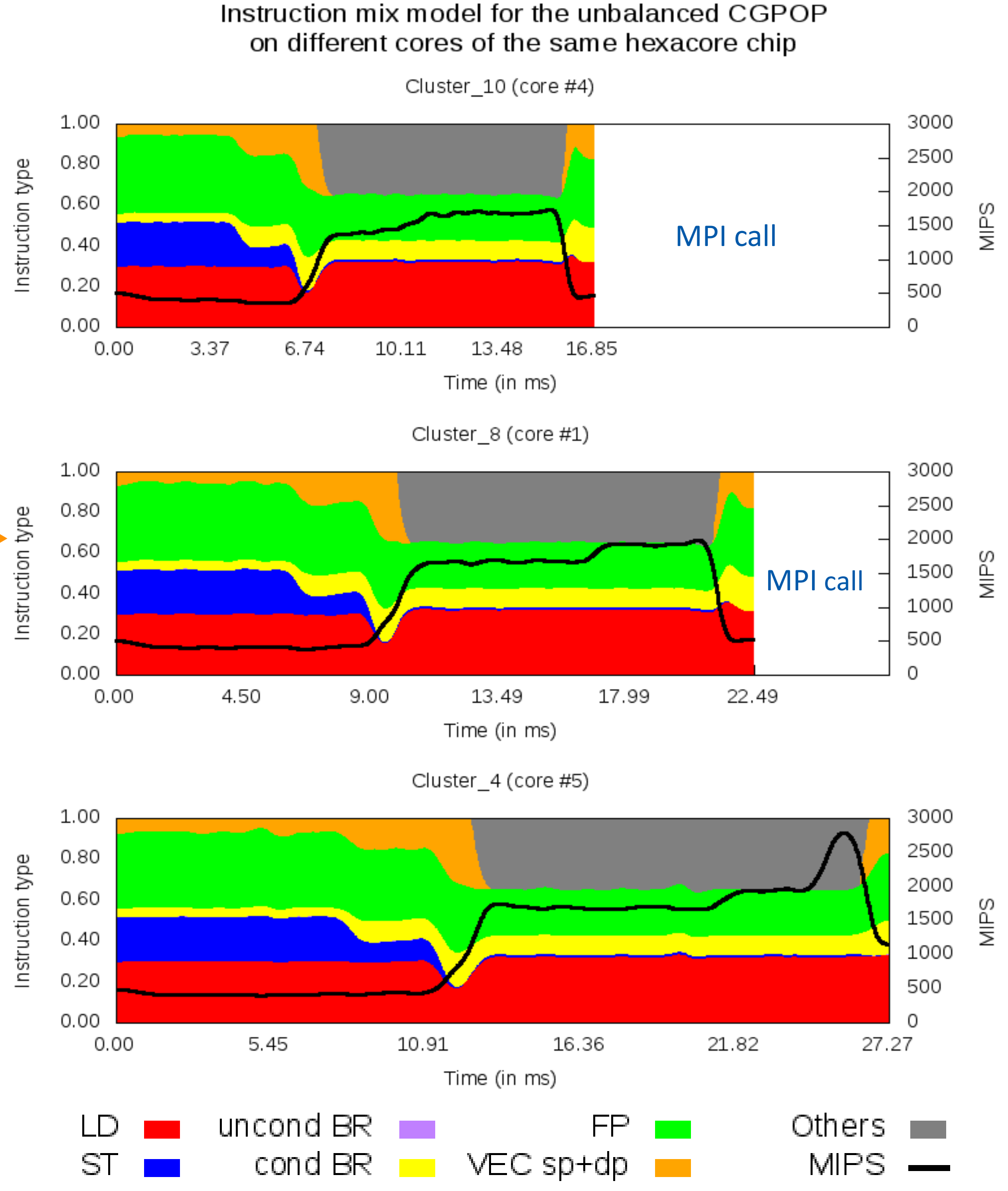
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CG-POP multicore MN3 study

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



```
ClusterID @ cgpop.linux_icc.180x120.chop2.clustered.prv
THREAD 1.13.1
THREAD 1.14.1
THREAD 1.15.1
THREAD 1.16.1
THREAD 1.17.1
THREAD 1.18.1
9.903.628.171 ns
```



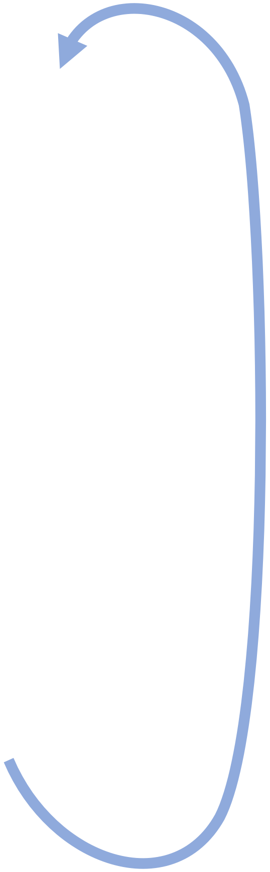
Instruction mix model for the unbalanced CGPOP on different cores of the same hexacore chip

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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?

Paraver Tutorial:

Introduction to Paraver and Dimemas methodology

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

Demo

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Same code, different behaviour

Code	Parallel efficiency	Communication eff.	Load Balance eff.
lulesh@mn3	90.55	**99.22**	91.26
lulesh@leftraru	**69.15**	99.12	**69.76**
lulesh@uv2 (mpt)	70.55	96.56	73.06
lulesh@uv2 (impi)	85.65	95.09	90.07
lulesh@mt	83.68	95.48	87.64
lulesh@cori	90.92	98.59	92.20
lulesh@thunderX	73.96	97.56	75.81
lulesh@jetson	75.48	**88.84**	84.06
lulesh@claix	77.28	92.33	83.70
lulesh@jureca	88.20	98.45	89.57
lulesh@mn4	86.59	98.77	87.67
lulesh@inti	88.16	98.65	89.36
lulesh@archer	88.01	97.95	89.86
lulesh@romeo	89.56	99.01	90.45
lulesh@mn4	**91.02**	98.38	**92.52**
lulesh@ stampede2 (skl)	85.76	97.63	87.84
lulesh@ stampede2 (knl)	89.21	98.42	90.64
lulesh@isambard	90.32	97.16	92.96

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Warning::: Higher parallel efficiency does not mean faster!