



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

Understanding applications with Paraver

tools@bsc.es

24th VI-HPS Tuning Workshop
University of Southampton, Feb 8th 2017

Humans are visual creatures

« Films or books?

- Two hours vs. days (months)

PROCESS

« Memorizing a deck of playing cards

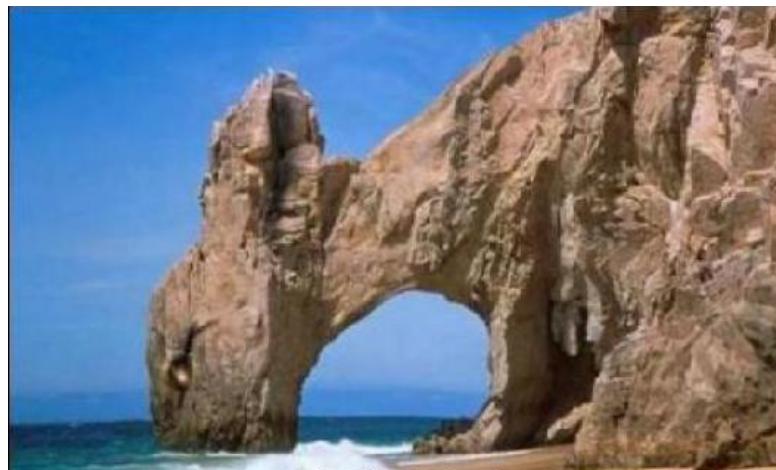
- Each card translated to an image (person, action, location)

STORE

« Our brain loves pattern recognition

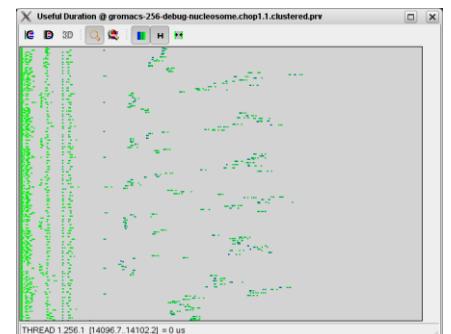
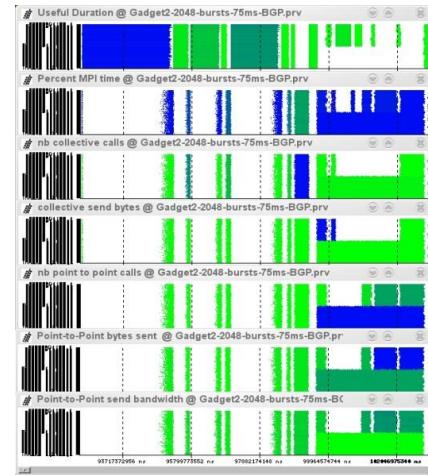
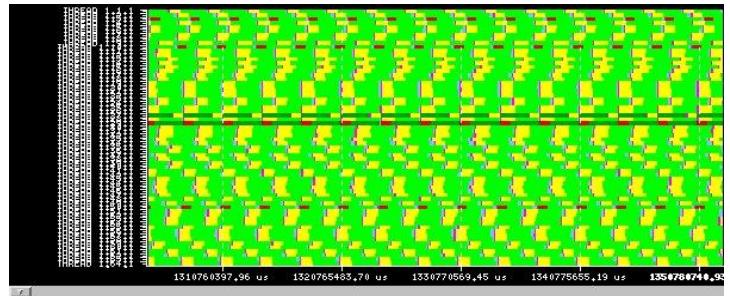
- What do you see on the pictures?

IDENTIFY



Our Tools

- « Since 1991
- « Based on traces
- « Open Source
 - <http://www.bsc.es/paraver>
- « 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



www.bsc.es

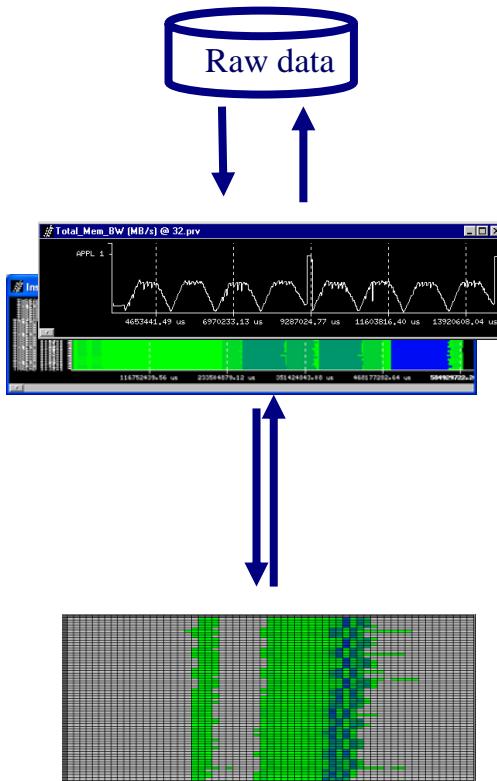


**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

Paraver

Paraver – Performance data browser



Trace visualization/analysis
+ trace manipulation

Timelines

Goal = Flexibility
No semantics
Programmable

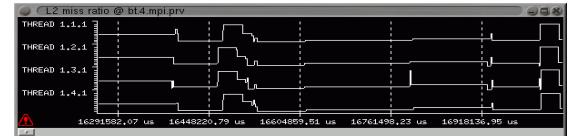
2/3D tables (Statistics)

Comparative analyses
Multiple traces
Synchronize scales

Timelines

« Each window displays one view

- Piecewise constant function of time



$$s(t) = S_i, i \in [t_i, t_{i+1})$$

« Types of functions

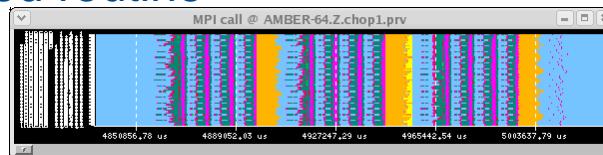
- Categorical

- State, user function, outlined routine

$$S_i \in [0, n] \subset N, \quad n <$$

- Logical

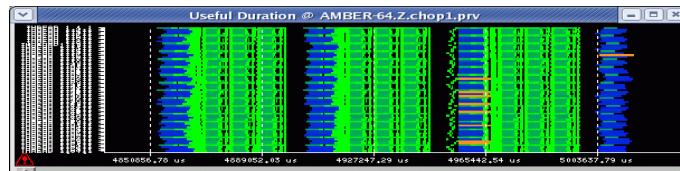
- In specific user function, In MPI call, In long MPI call



$$S_i \in \{0, 1\}$$

- Numerical

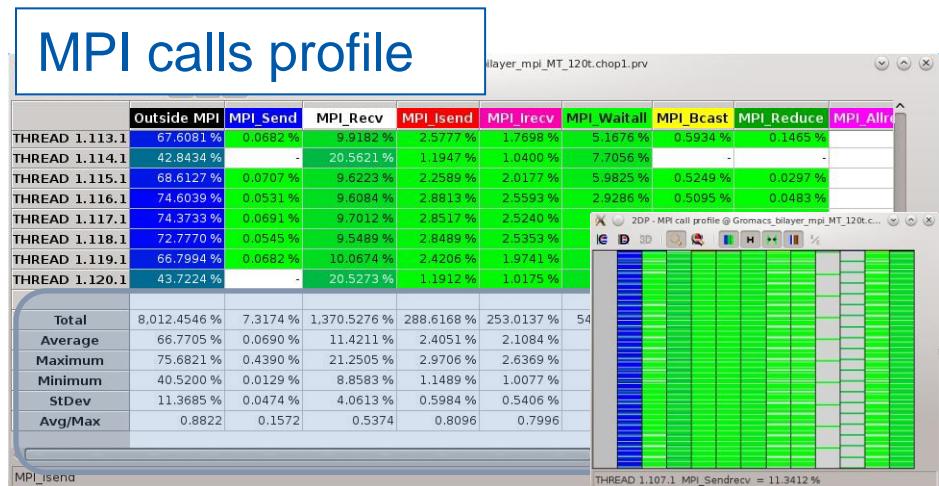
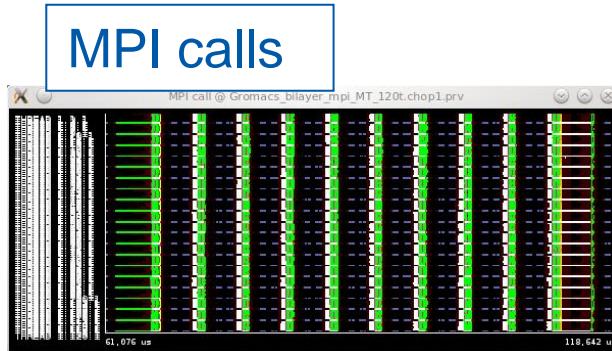
- IPC, L2 miss ratio, Duration of MPI call, duration of computation burst



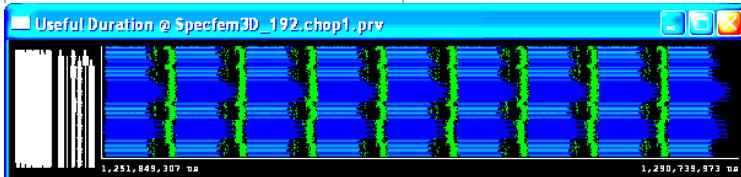
$$S_i \in R$$

Tables: Profiles, histograms, correlations

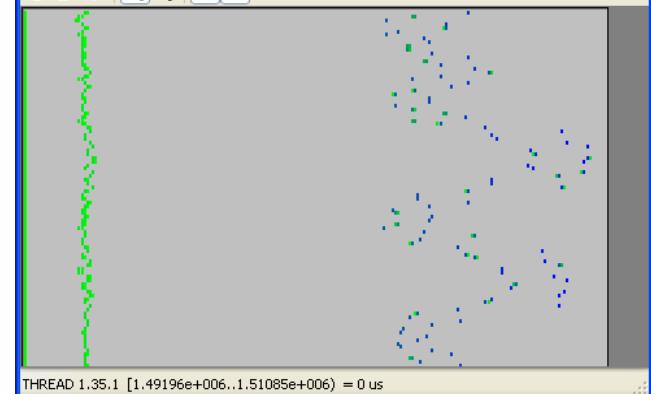
From timelines to tables



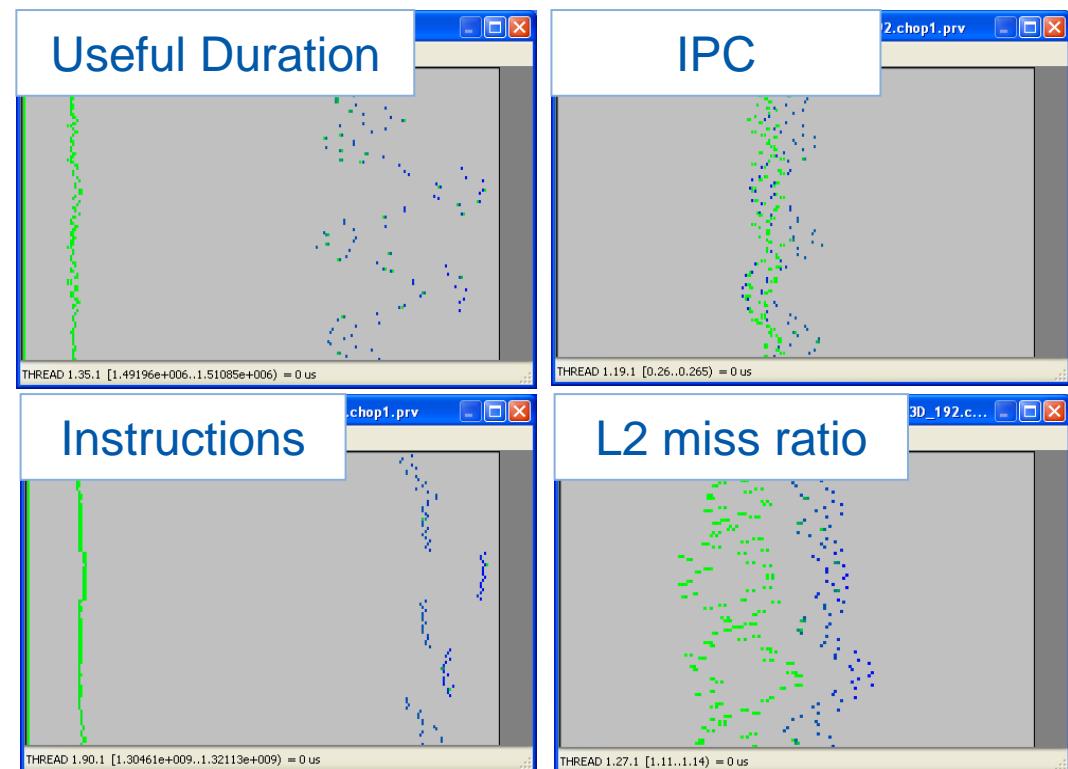
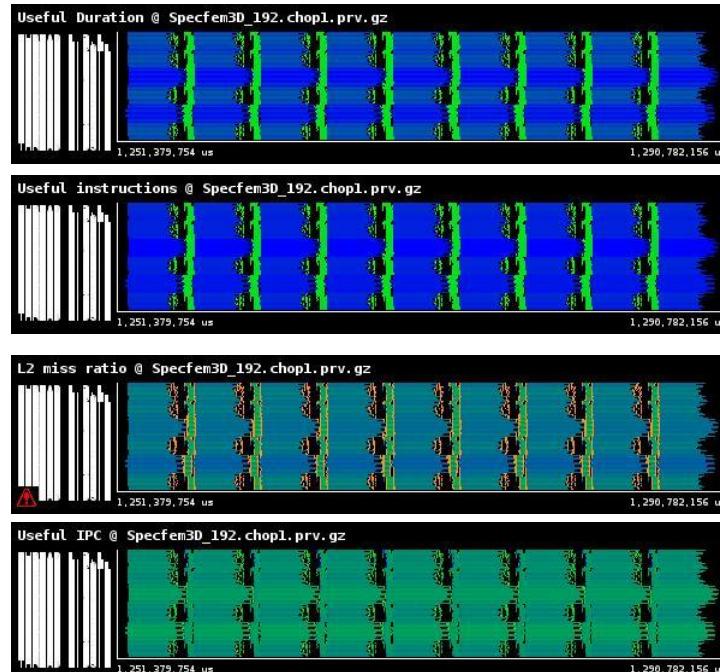
Useful Duration



Histogram Useful Duration

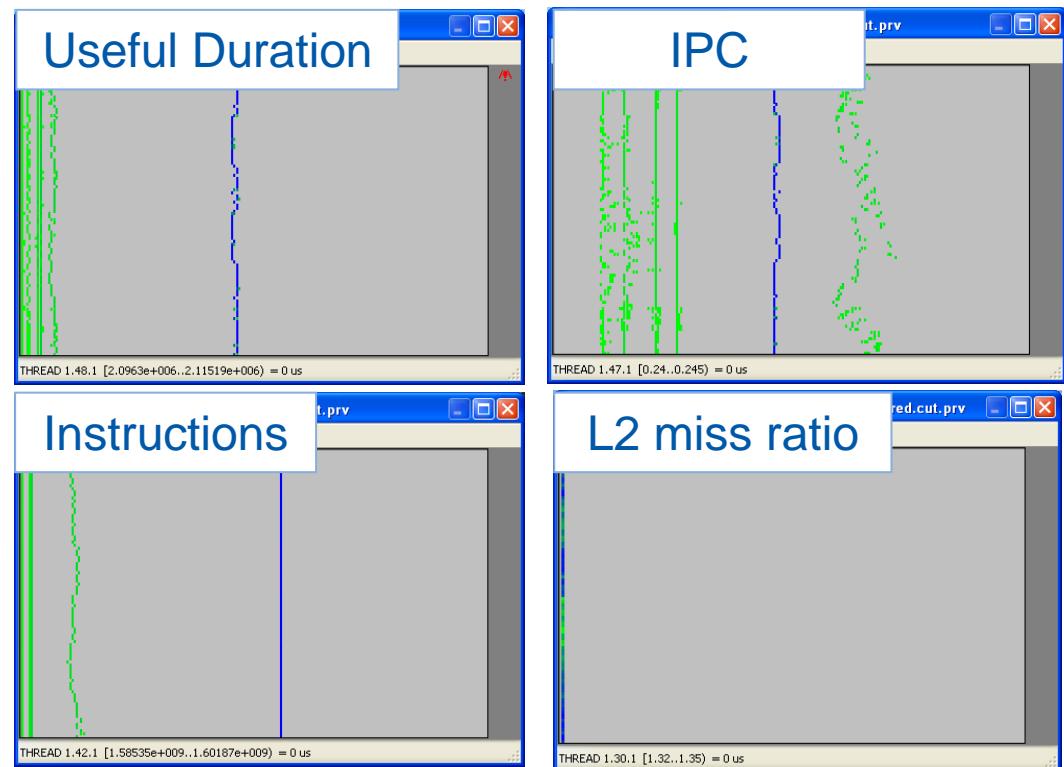


Analyzing variability through histograms and timelines



Analyzing variability through histograms and timelines

« By the way: six months later



From tables to timelines

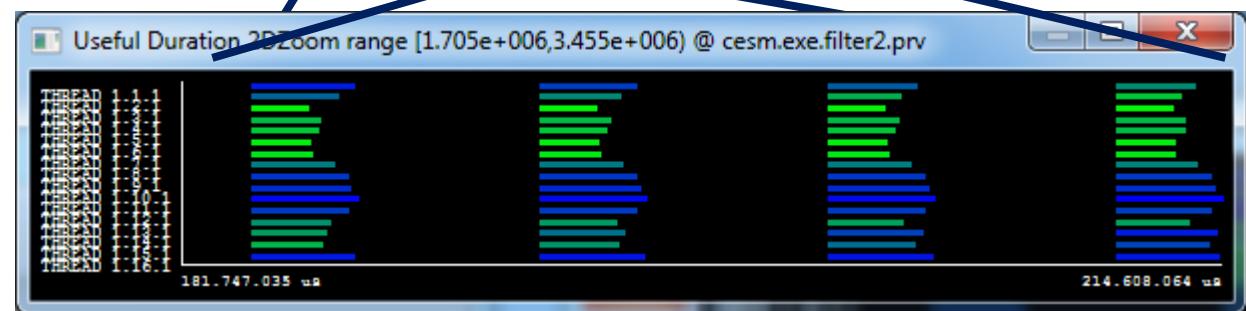
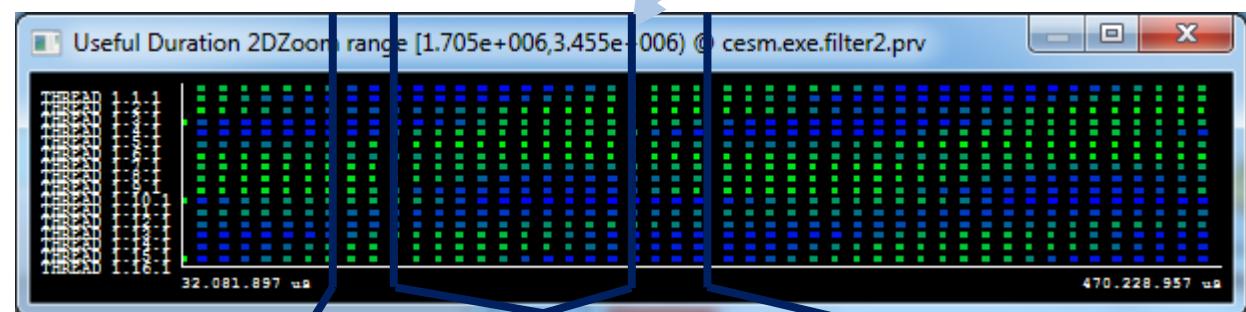
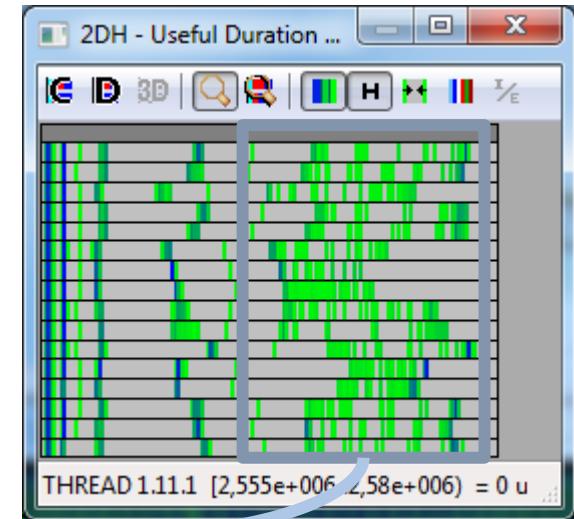
« CESM: 16 processes, 2 simulated days

« 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

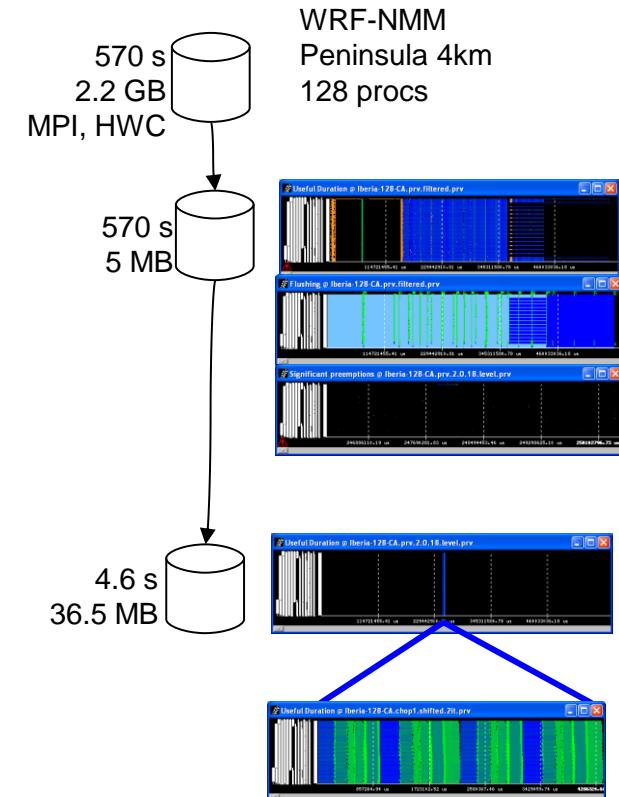
- Subset of records in original trace
- By duration, type, value,...
- Filtered trace IS a paraver trace and can be analysed with the same cfgs (as long as needed data kept)

– 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,...





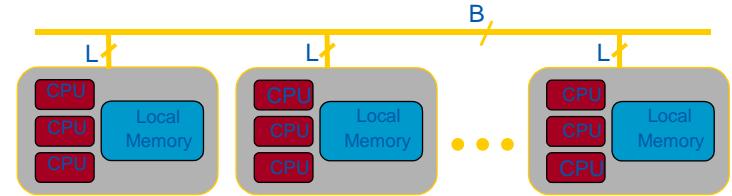
**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

Dimemas

Dimemas: Coarse grain, Trace driven simulation

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

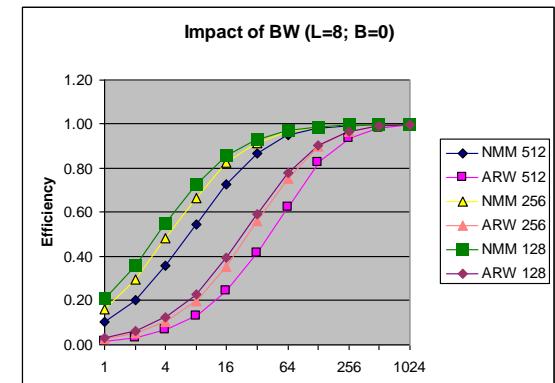


- « Parametric sweeps
 - On abstract architectures
 - On application computational regions

- « 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?

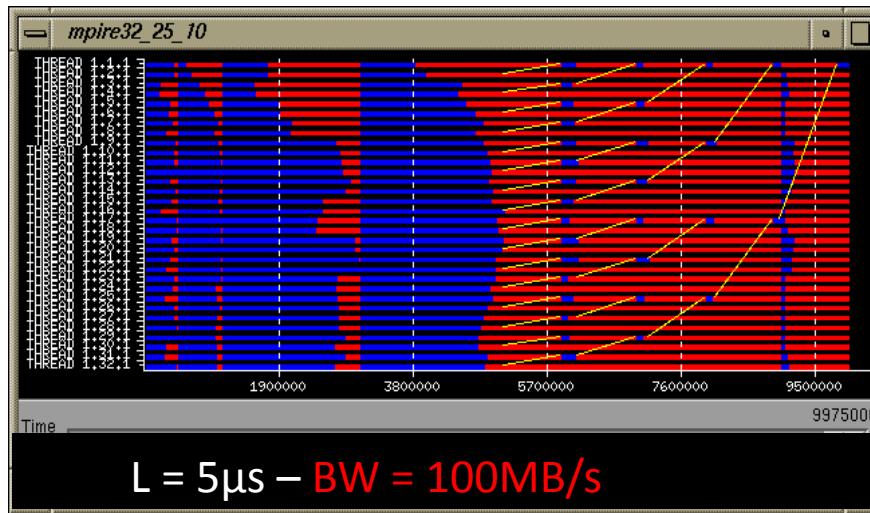
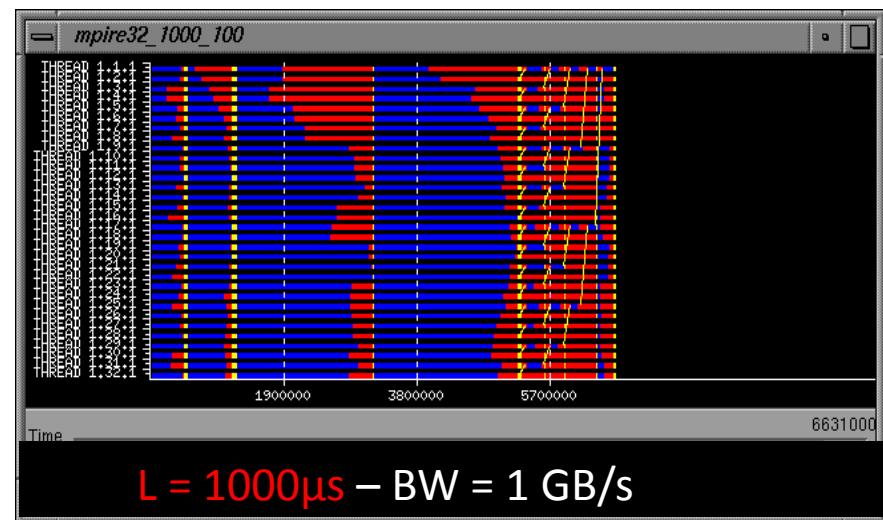
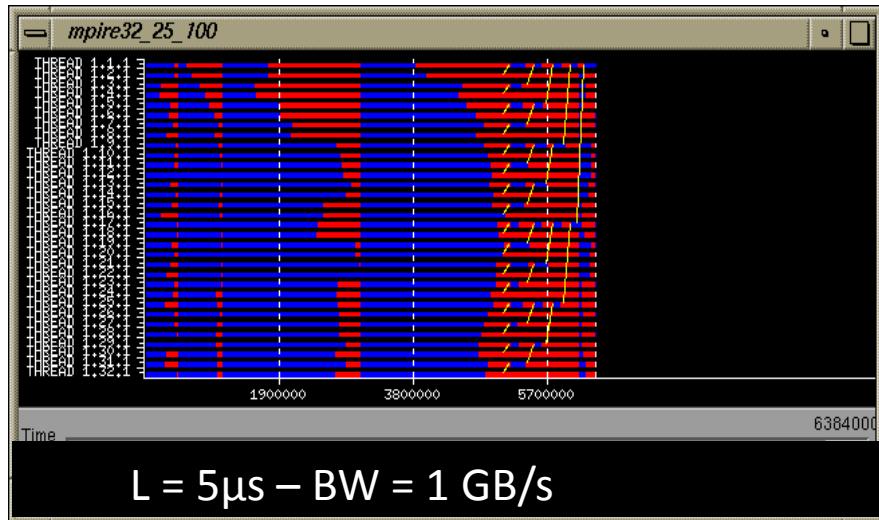
- « MPI sanity check
 - Modeling nominal

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



Network sensitivity

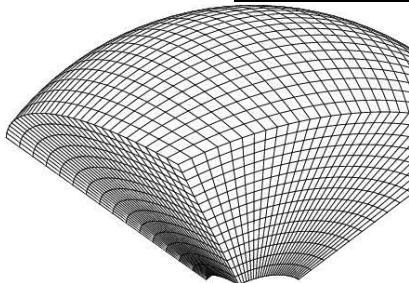
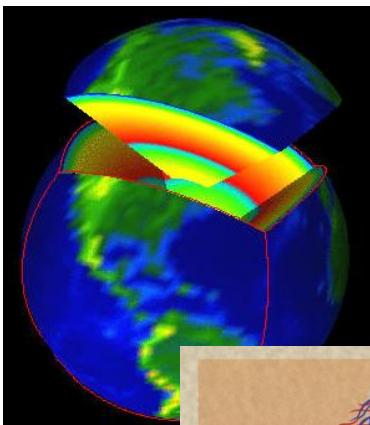
« MPIRE 32 tasks, no network contention



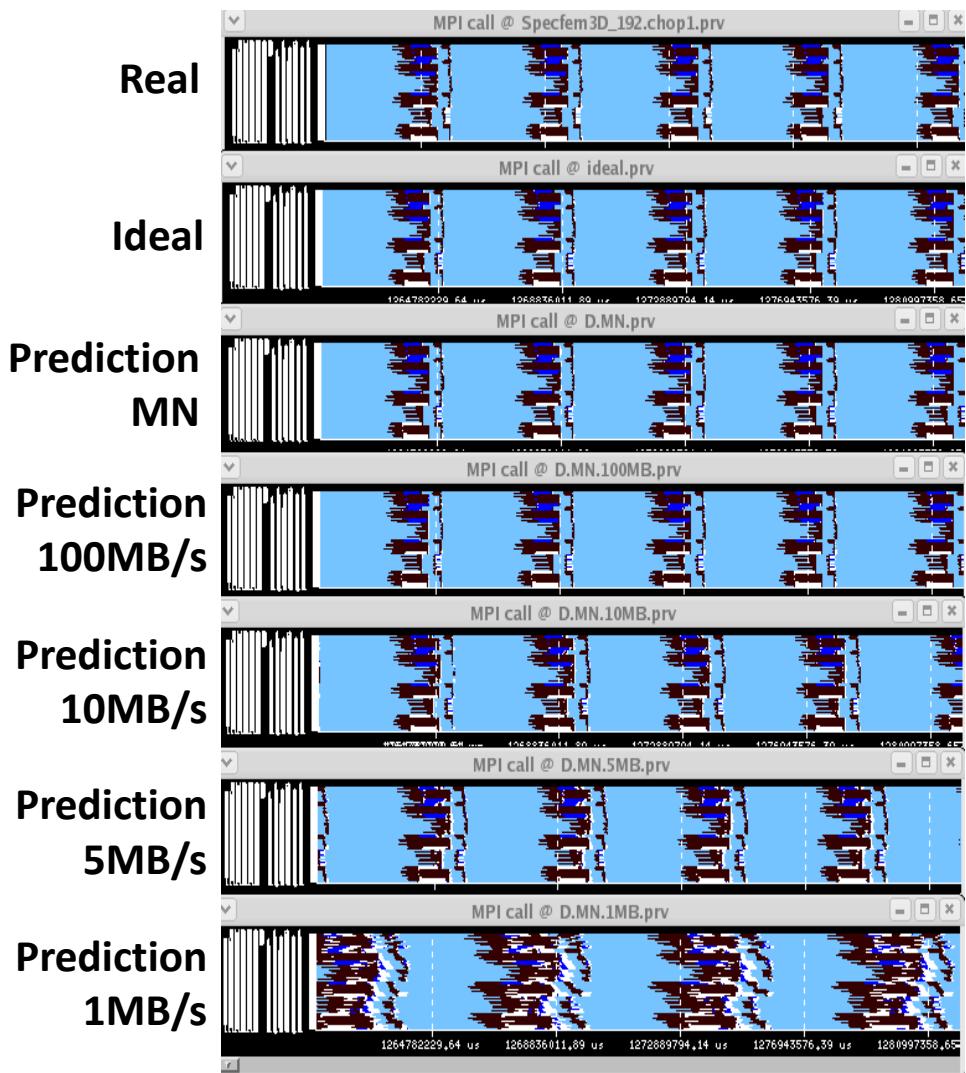
All windows same scale

Would I will benefit from asynchronous communications?

 SPECFEM3D

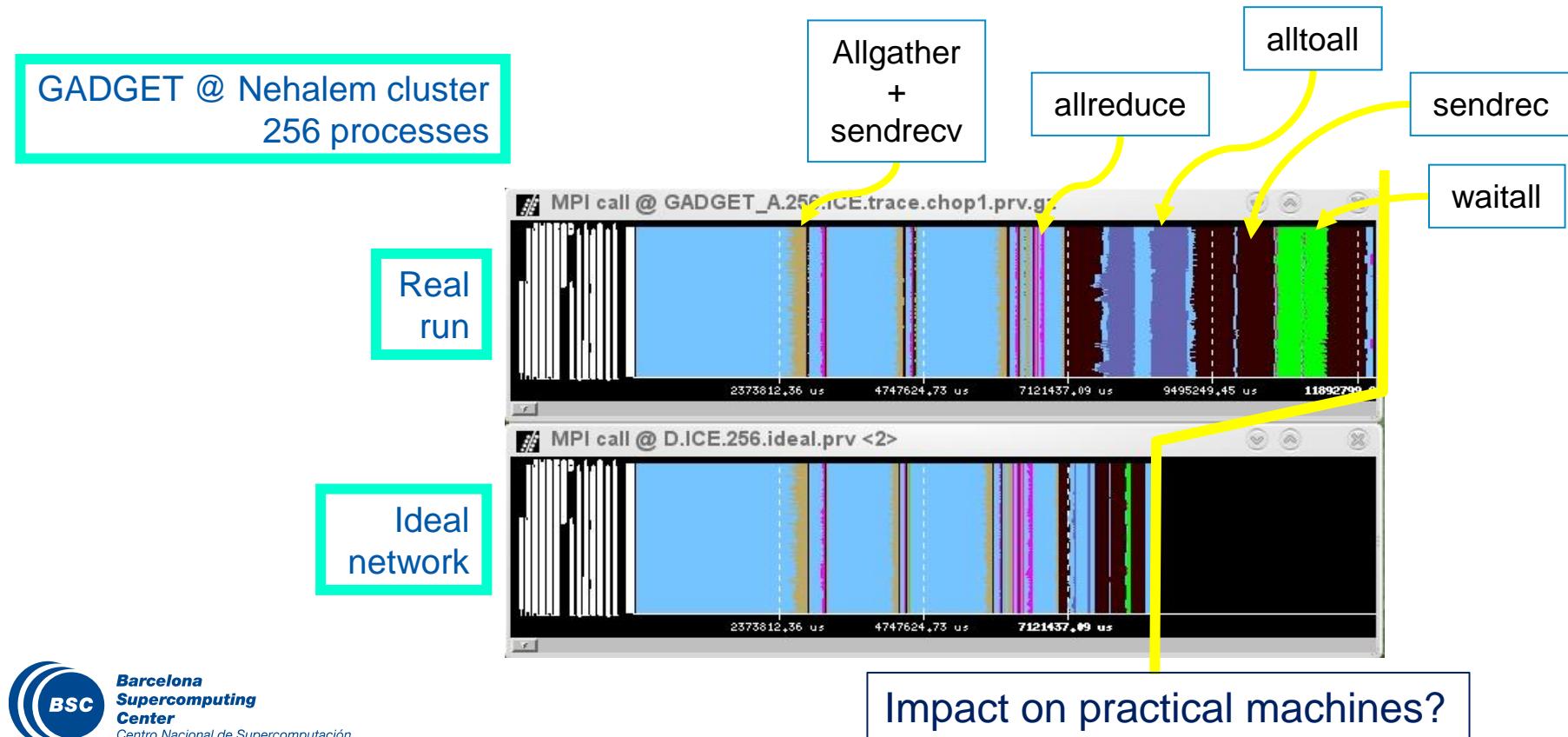


Courtesy Dimitri Komatitsch



Ideal machine

- « The impossible machine: $BW = \infty$, $L = 0$
- « Actually describes/characterizes Intrinsic application behavior
 - Load balance problems?
 - Dependence problems?

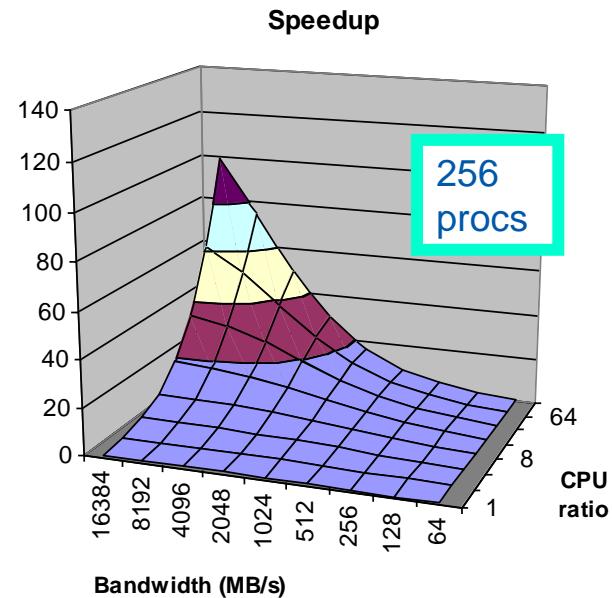
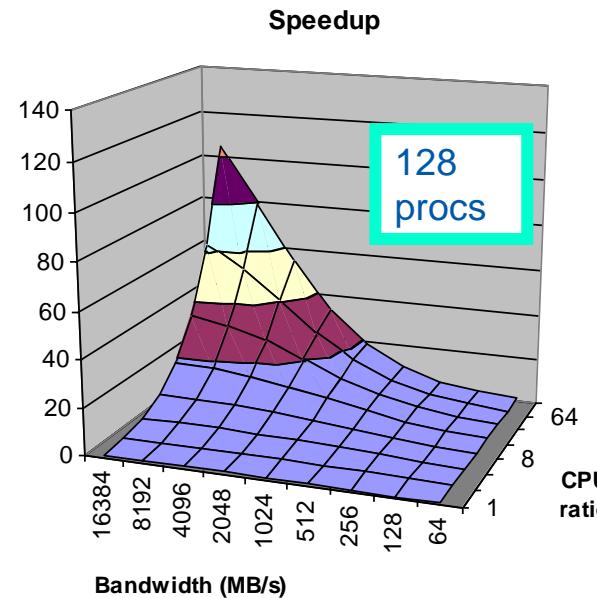
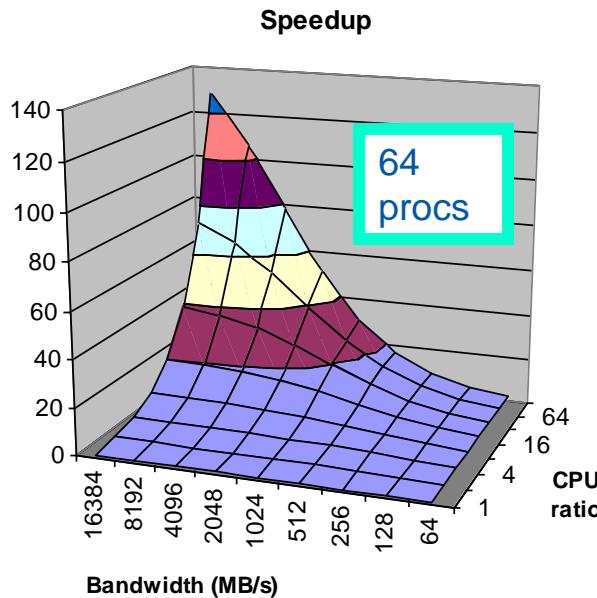


Impact of architectural parameters

« Ideal speeding up ALL the computation bursts by the CPUratio factor

- The more processes the less speedup (higher impact of bandwidth limitations) !!

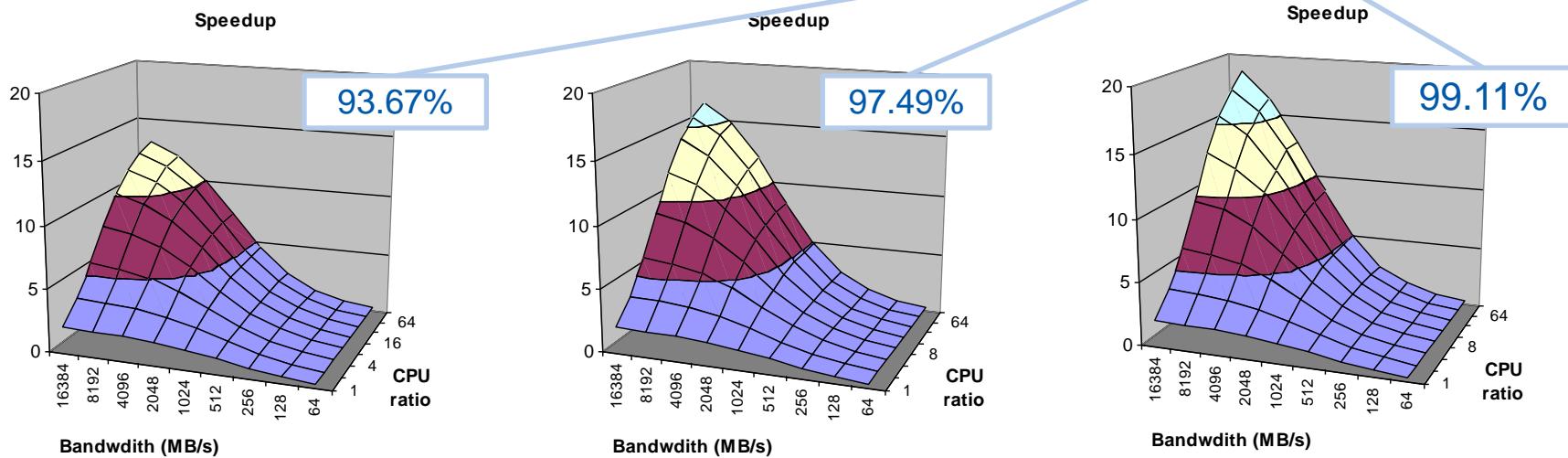
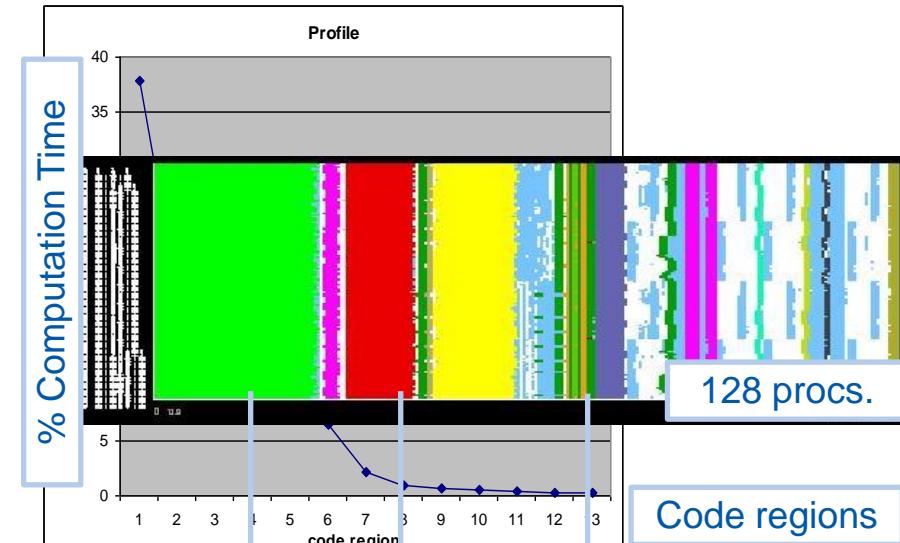
GADGET



Hybrid parallelization

Hybrid/accelerator parallelization

- Speed-up SELECTED regions by the CPURatio factor



(Previous slide: speedups up to 100x)



**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

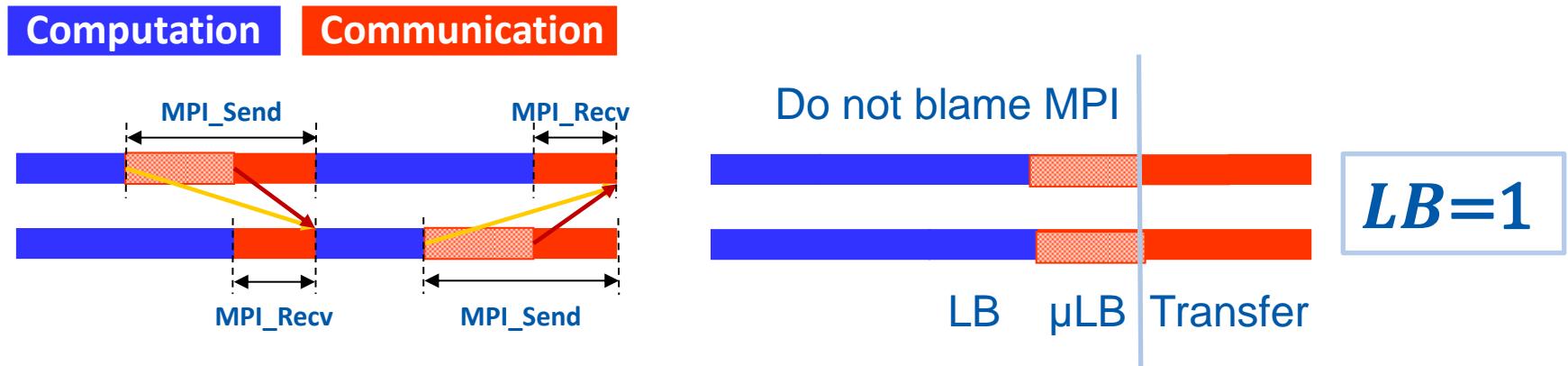
Models and Extrapolation

Parallel efficiency model

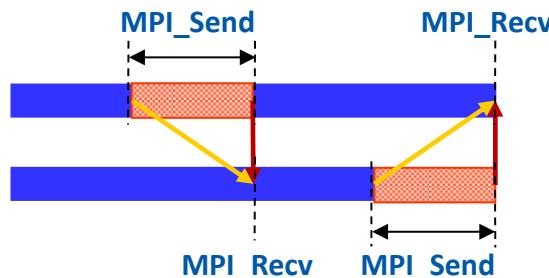


« Parallel efficiency = LB eff * Comm eff

Parallel efficiency refinement: LB * μ LB * Transfer



- « Serializations / dependences (μ LB)
- « Dimemas ideal network → Transfer (efficiency) = 1

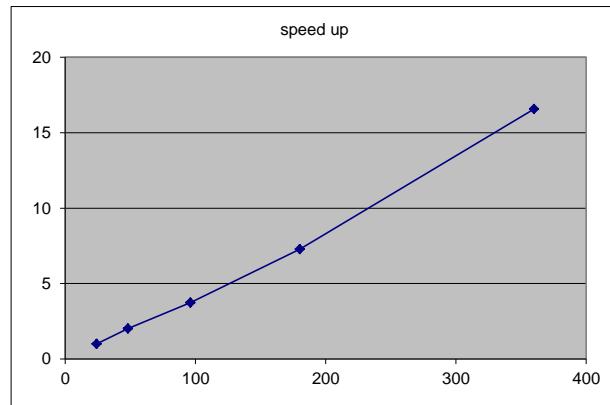


Why scaling?

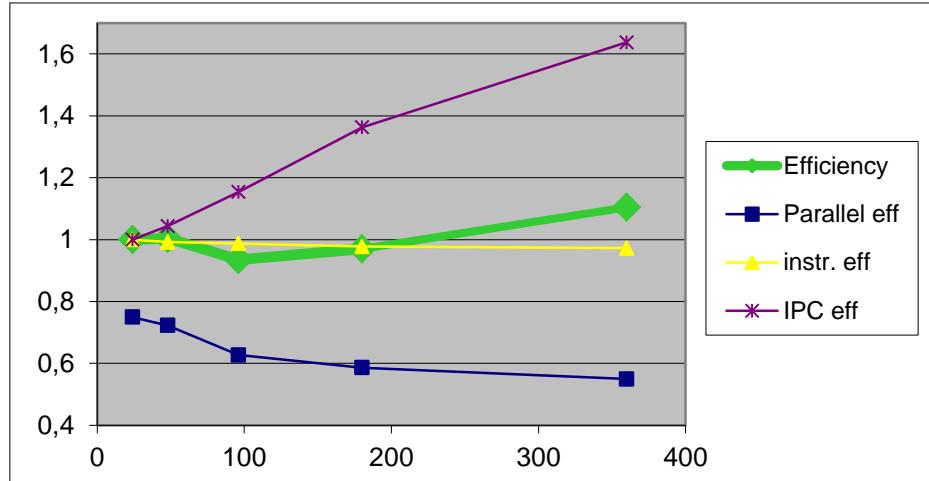
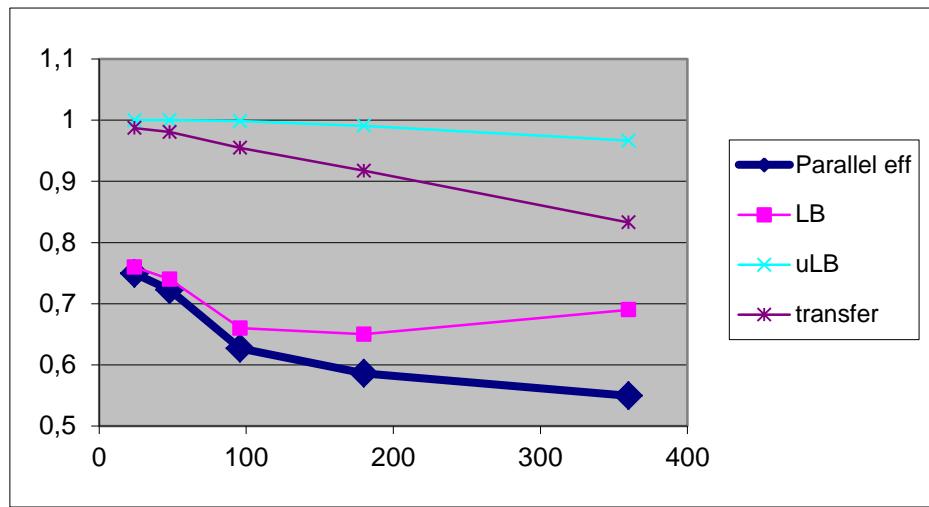
$$\eta_{\parallel} = LB * Ser * Trf$$

CG-POP mpi2s1D - 180x120

Good scalability !!
Should we be happy?



$$\eta = \eta_{\parallel} * \eta_{instr} * \eta_{IPC}$$



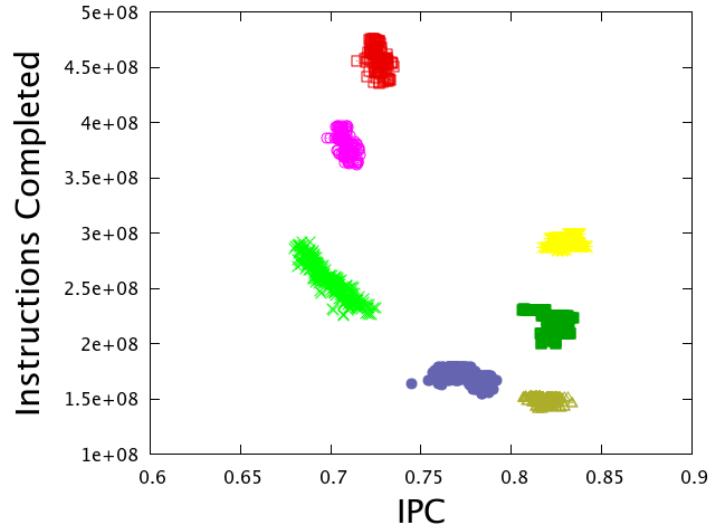
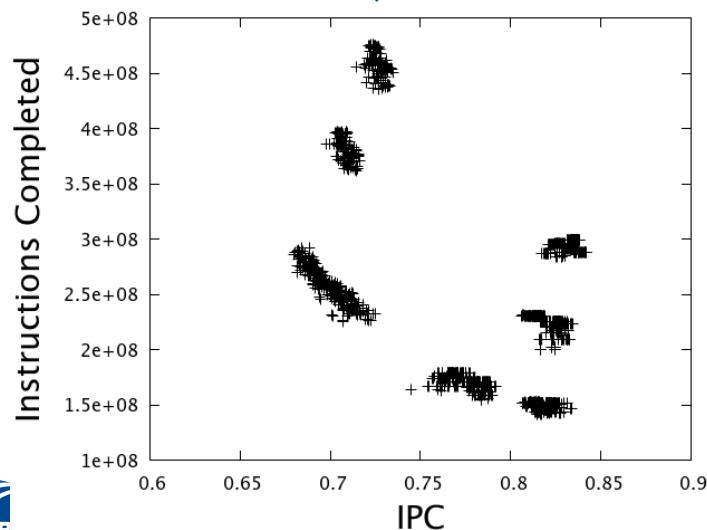
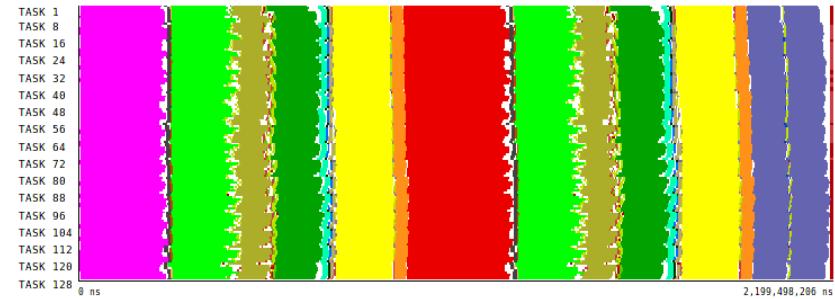
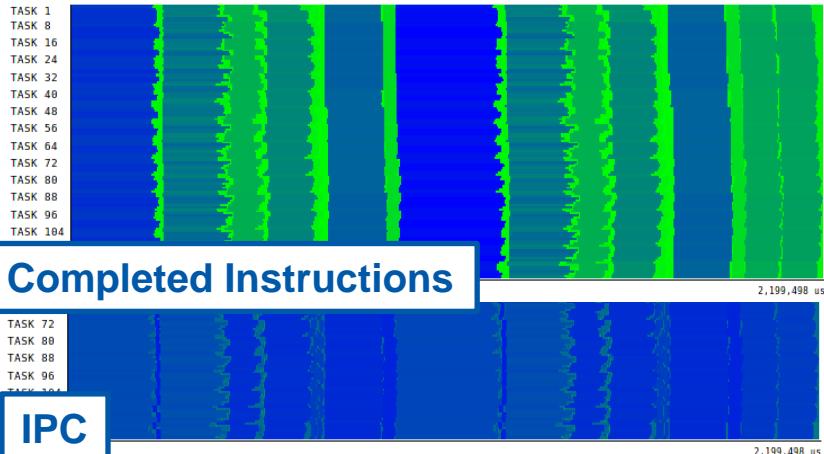


**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

Clustering

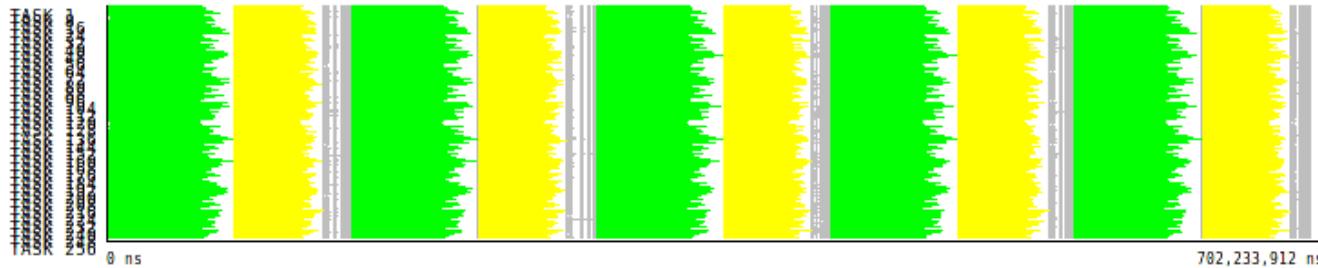
Using Clustering to identify structure



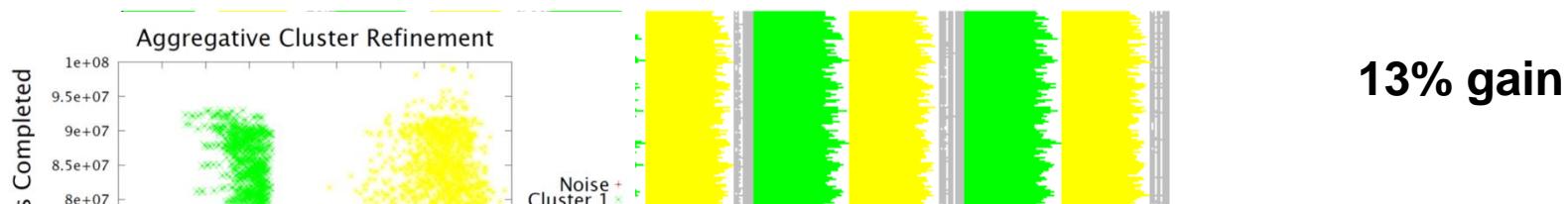
What should I improve?

What if

PEPC

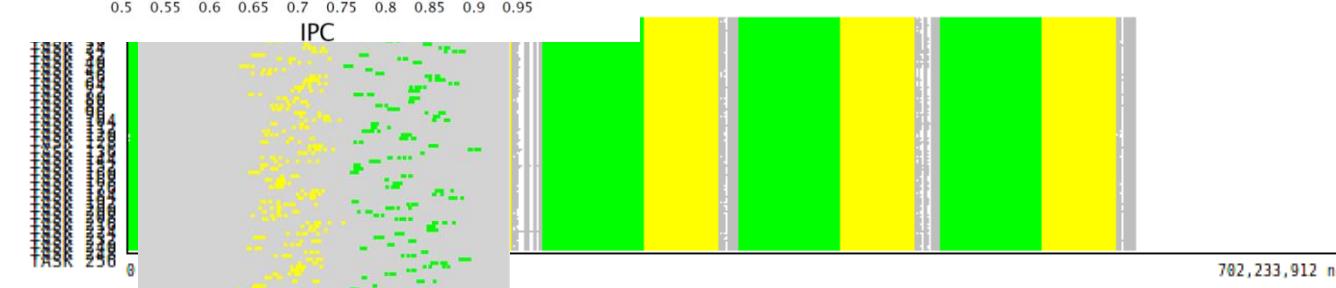


... we increase the IPC of Cluster1?



13% gain

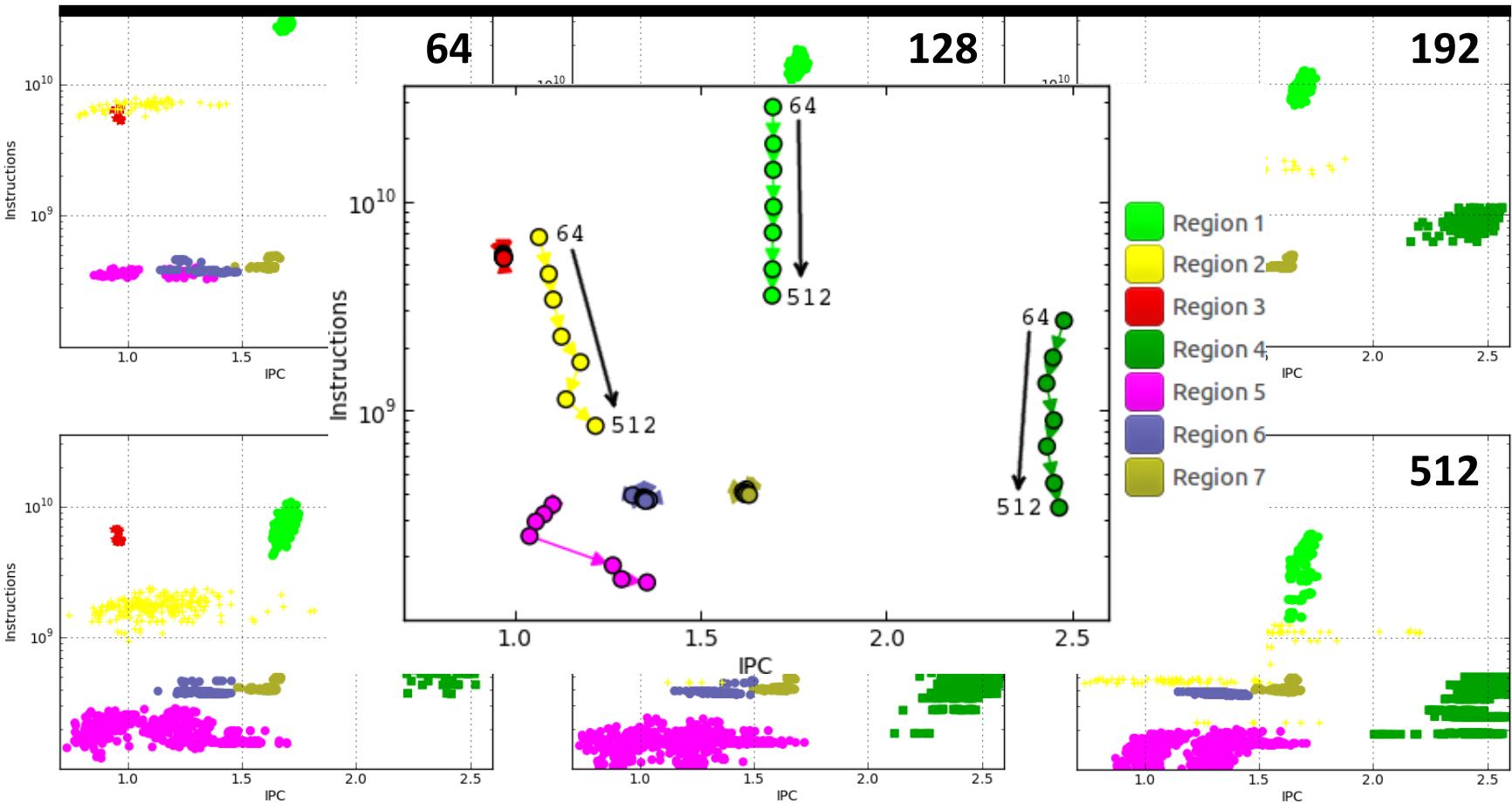
we balance Clusters 1 & 2?



19% gain

Tracking scalability through clustering

OpenMX (strong scale from 64 to 512 tasks)





**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

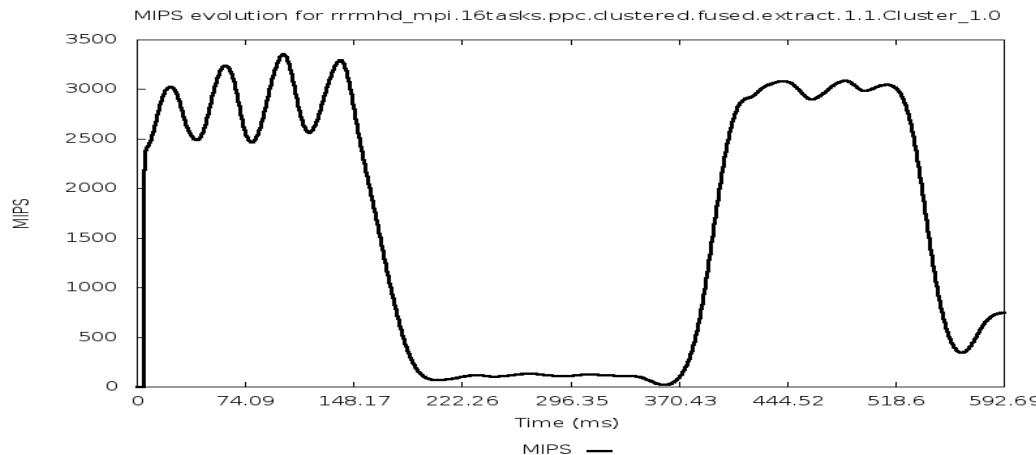
Folding

Folding: Detailed metrics evolution

- Performance of a sequential region = 2000 MIPS

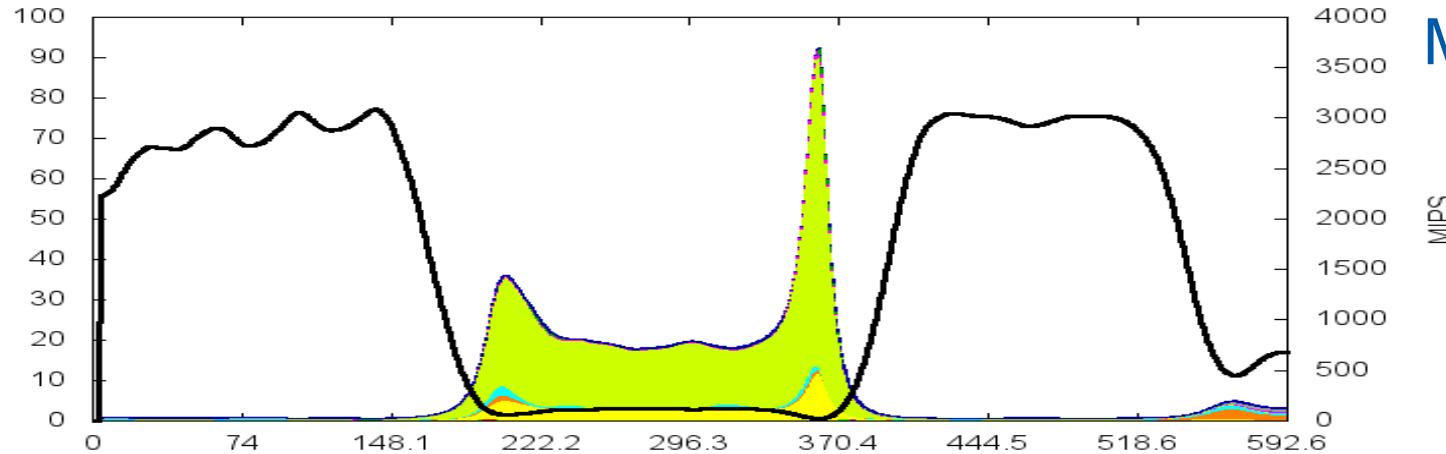
Is it good enough?

Is it easy to improve?



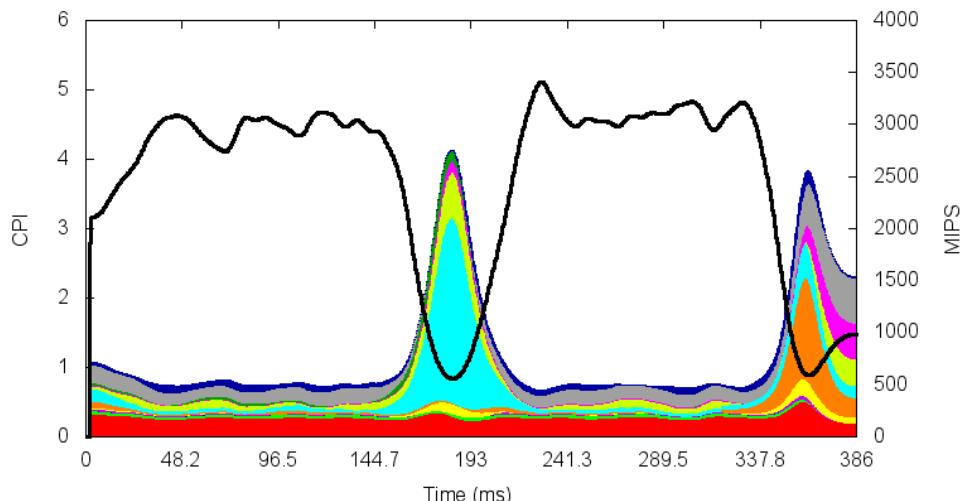
Folding: Instantaneous CPI stack

PowerPC CPI break-down evolution for Cluster 1 of Mr.Genesis



MRGENESIS

PowerPC CPI break-down evolution for Cluster 1 of Mr.Genesis



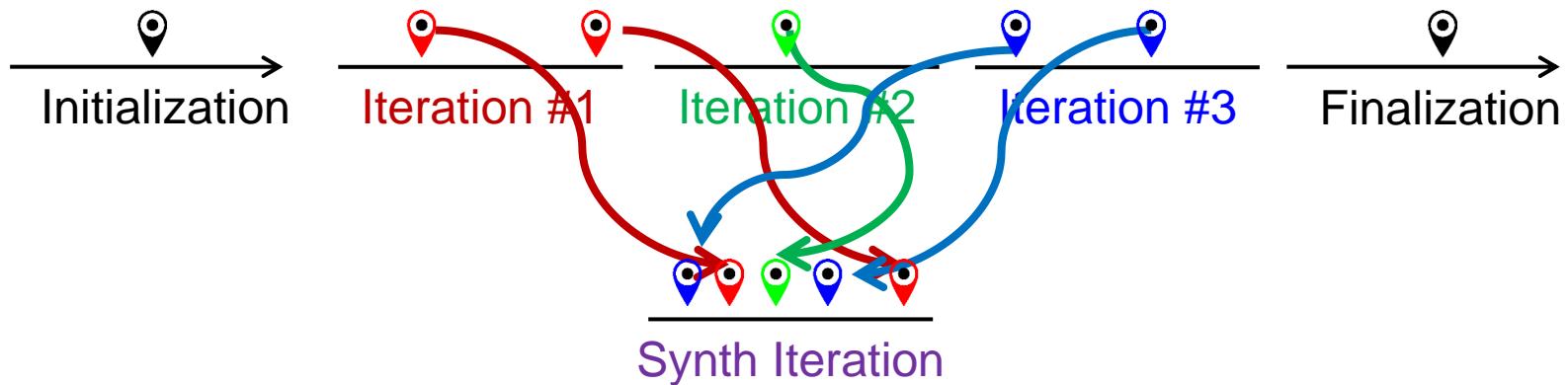
Useful cycles	Red	LSU: Basic latency	Yellow
I-cache miss	Green	FXU: Div/MSTPR/MSFPR	Grey
Branch mispredict	Blue	FXU: Basic latency	Magenta
Flush penalties, etc	Purple	FPU: FDIV/FSQRT	Green
LSU: Translation lookup	Yellow	FPU: Basic latency	Grey
LSU: Other reject	Orange	Other stall cycles	Dark Blue
LSU: D-cache miss	Cyan	MIPS	Black

- Trivial fix.(loop interchange)
- Easy to locate?
- Next step?
- Availability of CPI stack models for production processors?
 - Provided by manufacturers?

Folding

« Instantaneous metrics with minimum overhead

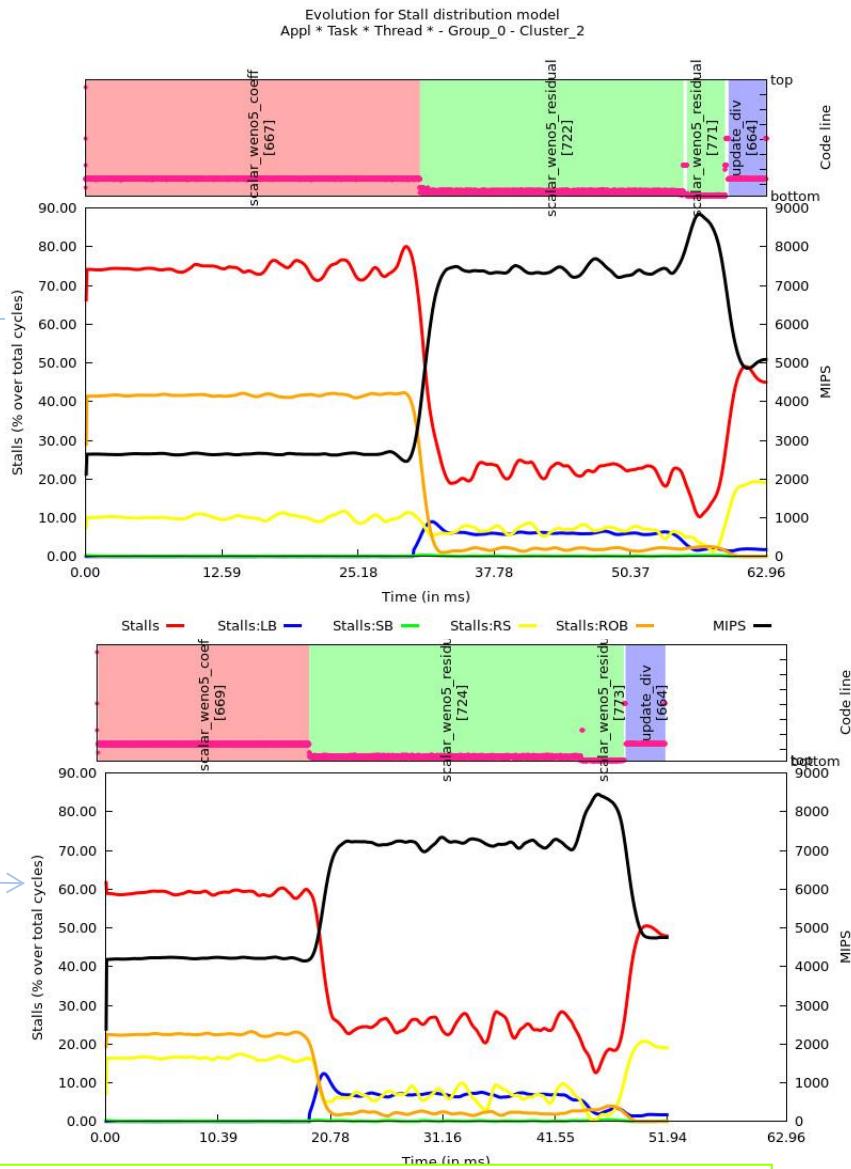
- Combine instrumentation and sampling
 - Instrumentation delimits regions (routines, loops, ...)
 - Sampling exposes progression within a region
- Captures performance counters and call-stack references



“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

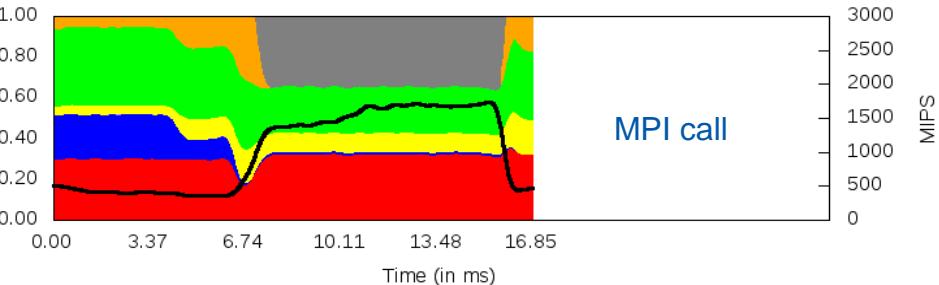
Unbalanced MPI application

- Same code
- Different duration
- Different performance

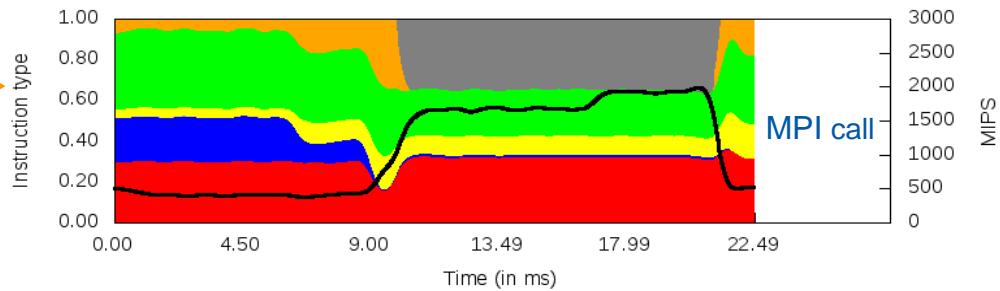


Instruction mix model for the unbalanced CGPOP on different cores of the same hexacore chip

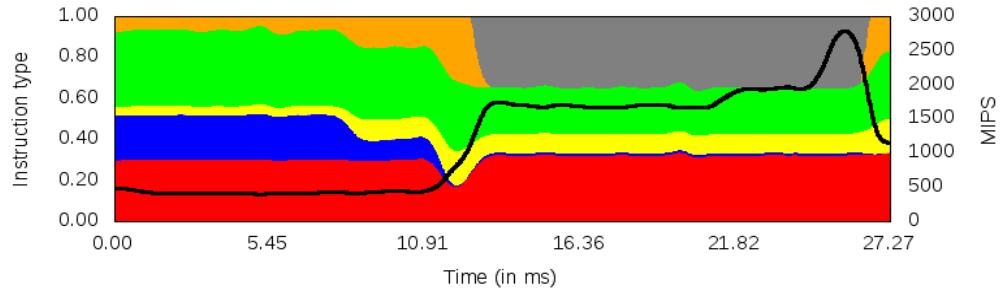
Cluster_10 (core #4)



Cluster_8 (core #1)



Cluster_4 (core #5)



LD ■ uncond BR ■ FP ■ Others
ST ■ cond BR ■ VEC sp+dp ■ MIPS

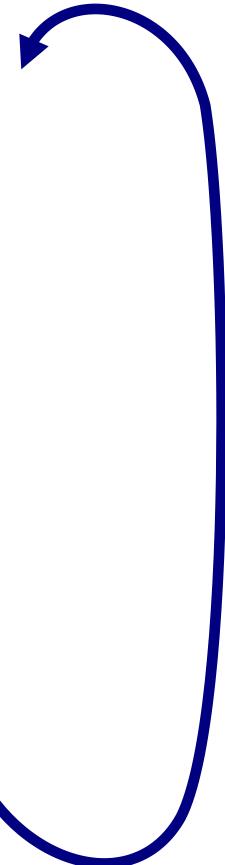


**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

Methodology

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

« www.bsc.es/paraver

- 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

Thanks!

- « Use your brain, use visual tools :)
- « Look at your codes!

www.bsc.es/paraver





**Barcelona
Supercomputing
Center**

Centro Nacional de Supercomputación

Demo

Some examples of efficiencies

Code	Parallel efficiency	Communication efficiency	Load Balance efficiency
Gromacs@mt	66.77	75.68	88.22
BigDFT@altamira	59.64	78.97	75.52
CG-POP@mt	80.98	98.92	81.86
ntchem_mini@pi	92.56	94.94	97.49
nicam@pi	87.10	75.97	89.22
cp2k@jureca	75.34	81.07	92.93
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
icon@mistral	79.86	84.02	95.05
k-Wave@salomon	89.08	92.84	95.96