INTELLECTUAL PROPERTY RIGHTS NOTICE:

- The User may only download, make and retain a **copy of the materials for his/her use for non-commercial** and research purposes.
- The User **may not commercially use the material**, unless has been granted prior written consent by the Licensor to do so; and **cannot remove, obscure or modify copyright notices**, text acknowledging or other means of identification or disclaimers as they appear.
- For further details, please contact **BSC-CNS** patc@bsc.es

PATC training, Barcelona, May 2012 ‹#›

PRACE TRAINING COURSE under PRACE Advance Training Centre at BSC

BSC-CNS http://www.bsc.es/

PRACE project http://www.prace-ri.eu/

PRACE Training Portal http://www.training.prace-ri.eu/

PATC @ BSC Training Program

http://www.bsc.es/marenostrum-support-services/hpc-trainings/prace-trainings

PATC training, Barcelona, May 2012 ‹#›

Performance analysis with

BSC-Tools

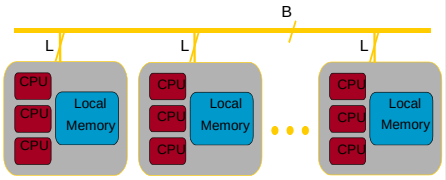
Judit Gimenez (judit@bsc.es)
BSC

Dimemas

Dimemas: Coarse grain trace driven simulator

Key factors influencing performance

- Abstract architecture
- Basic MPI protocols
- No atempt to model details



Objectives

- Simple / general
- Fast simulations

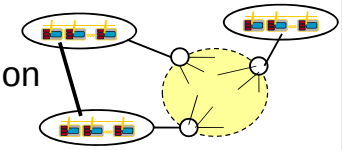
$$T = \frac{MessageSize}{BW} + L$$

Linear components

- Point2point, CPU/block speed

Non-linear components

- Synchronization, resources contention



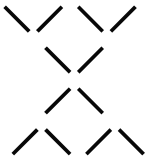
Network of SMPs / GRID

BSC tools @ PATC, Barcelona May 21-22 2012 5

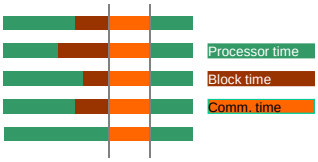
Collective communication model

Generic model

- Barrier / Fan-in / Fan-out
- Cost of communication phase
  - Generic
  - Per call
    - Model factor
      - Lin / log / const
    - Size of message
      - Min over all processes
      - Avg over all processes
      - Max over all processes



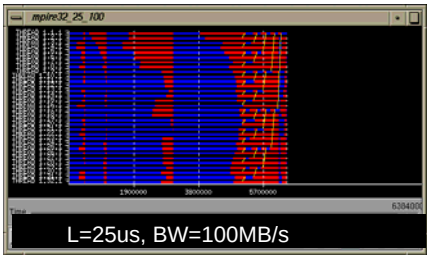
Collective

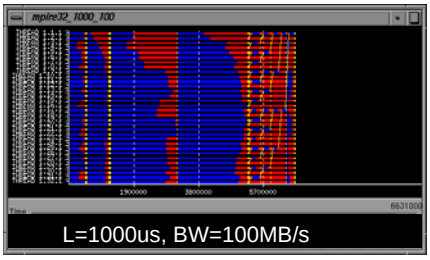


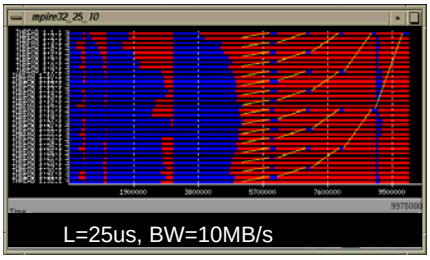
BSC tools @ PATC, Barcelona May 21-22 2012 6

Understanding applications

MPIRE 32 tasks, no network contention



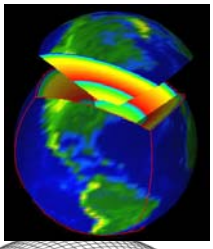




All windows same scale

Predicting performance

SPECFEM3D – asynchronous communication?



Courtesy Dimitri Komatitsch

Real

ideal

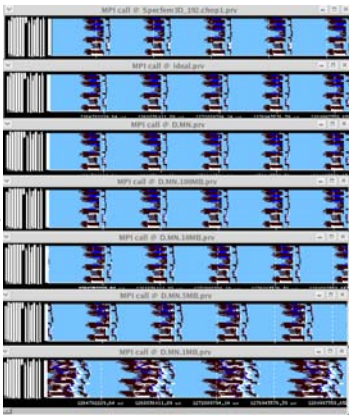
MN prediction

Prediction 100MB/s

Prediction 10MB/s

Prediction 5MB/s

Prediction 1MB/s



Parametric studies – network sensitivity

WRF, Iberia 4Km, 4 procs/node

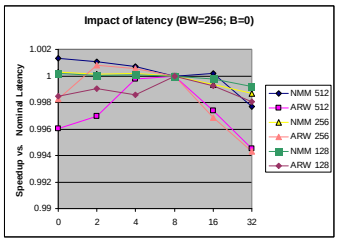
None sensitive to latency

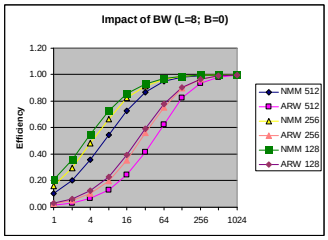
NMM

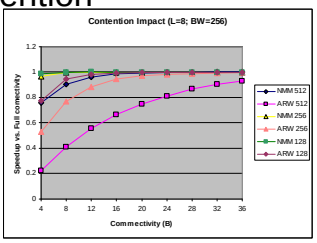
- Bw – 256MB/s
- 512 – sensitive to contention

ARW

- Bw - 1GB/s
- Sensitive to contention







Multiscale simulation: L2 size vs network bw

Left: clusters IPC with different cache sizes

- 64KB - 512MB

Right: execution time with different network bw and cache size

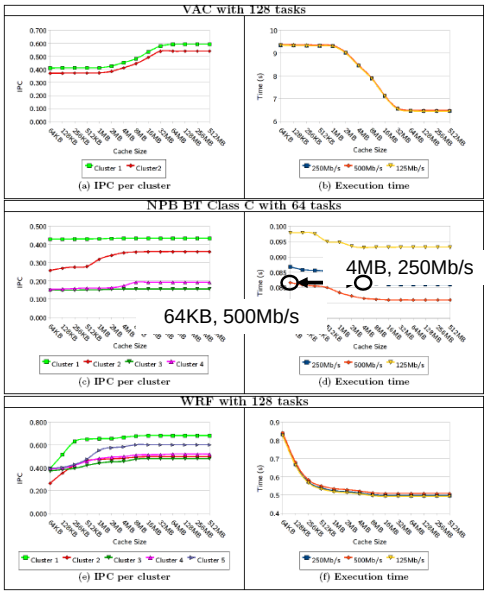
- 125Mb/s - 500Mb/s

NAS BT

- Can compensate cache reduction with more network bw

VAC, WRF

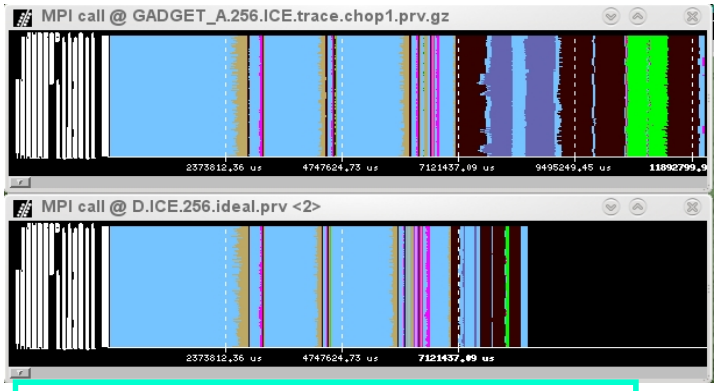
- Dominated by comp.



Application limits

GADGET behavior: load balance / dependences?

Real run



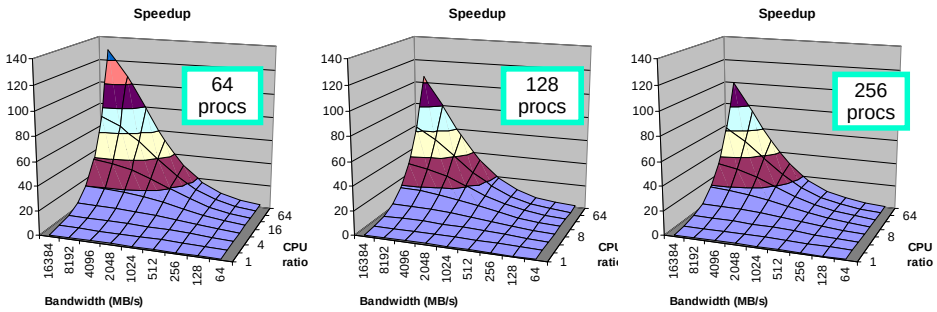
Ideal network: infinite bandwidth, no latency

BSC tools @ PATC, Barcelona May 21-22 2012 11

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

BSC tools @ PATC, Barcelona May 21-22 2012 12

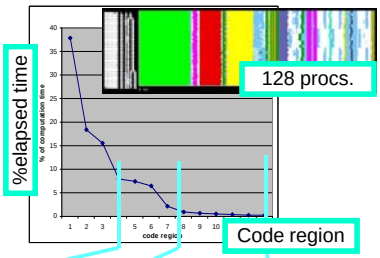
What if: Potential of hybrid parallelization

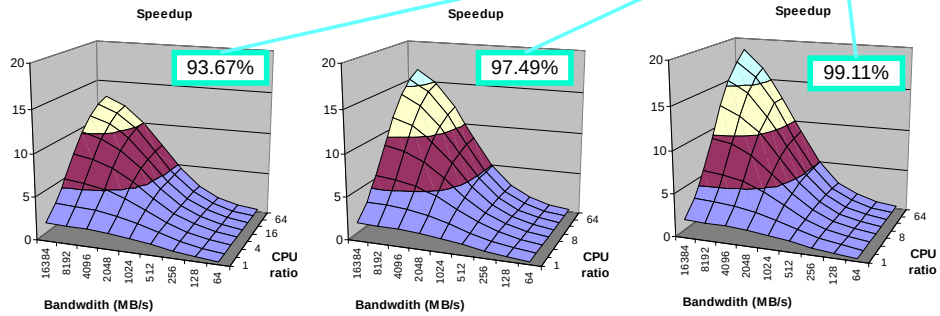
Hybrid/accelerator parallelization

Speedup SELECTED regions by the CPUratio factor

We do need to overcome the **hybrid Amdahl's law**

→ **asynchrony + Load balancing mechanisms !!!**





CG-POP mpi2s1D - 180x120

Comparing 10 iterations – MPI calls view, same time scale

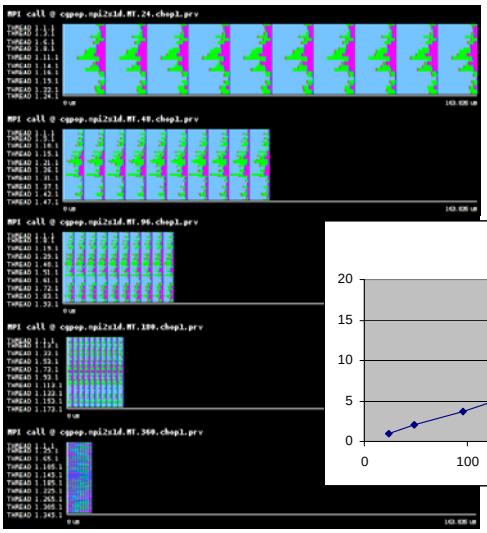
24 tasks

48 tasks

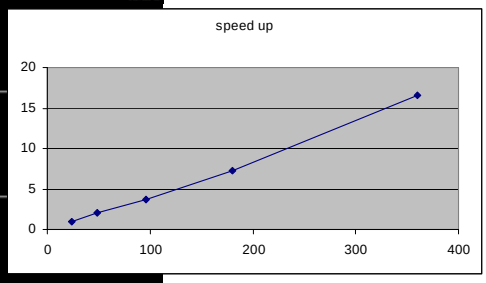
96 tasks

180 tasks

360 tasks



Good scalability



CG-POP mpi2s1D - 180x120

Comparing 1 iterations – **different** time scale

MPI call view | Duration of computing phase

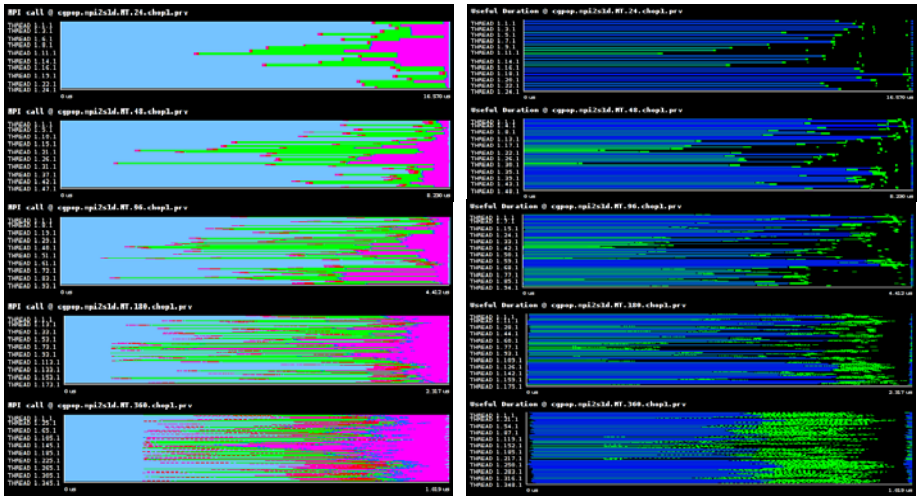
24

48

96

180

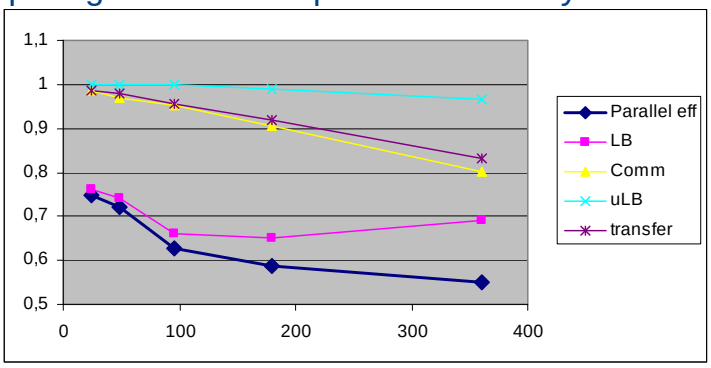
360



BSC tools @ PATC, Barcelona May 21-22 2012 15

CG-POP mpi2s1D - 180x120 efficiency

Comparing 10 iterations parallel efficiency

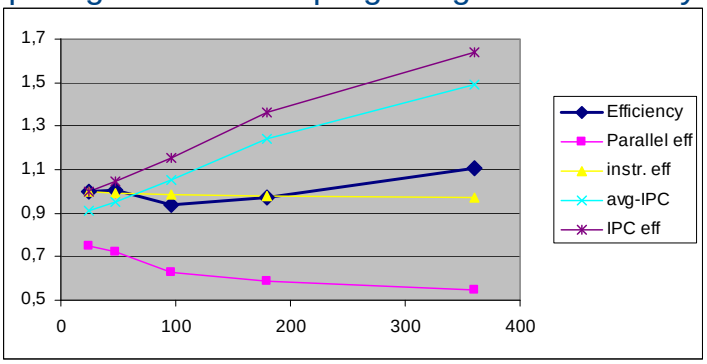


Cores	Parallel eff	LB	Comm	uLB	transfer
24	0,75	0,76	0,99	1,00	0,99
48	0,72	0,74	0,97	1,00	0,98
96	0,63	0,66	0,95	1,00	0,95
180	0,59	0,65	0,90	0,99	0,92
360	0,55	0,69	0,80	0,97	0,83

BSC tools @ PATC, Barcelona May 21-22 2012 16

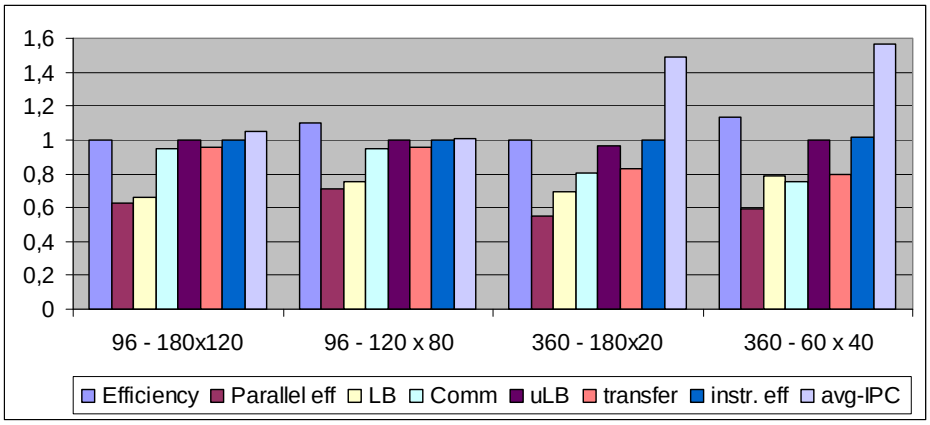
CG-POP mpi2s1D - 180x120 efficiency

Comparing 10 iterations program global efficiency



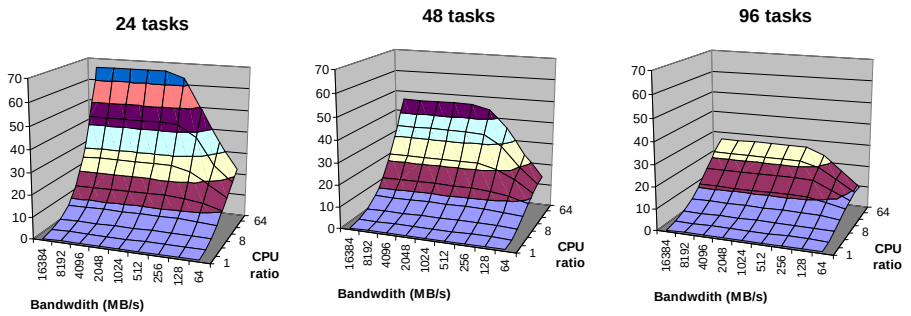
Cores	Time	Parallel eff.	tot-Ins	avg-IPC
24	163804479	0,75	8633761558	0,91
48	81589358	0,72	8701167887	0,95
96	43793988	0,63	8745555664	1,05
180	22508326	0,59	8834298647	1,24
360	9890676	0,55	8877820754	1,49

CG-POP mpi2s1D - Block size impact



CG-POP mpi2s1D - 180x120 sensitivity

Parametric studies – sensitivity to network bandwith and CPU speed. Comparing 24, 48 and 96 (24 tasks scale)

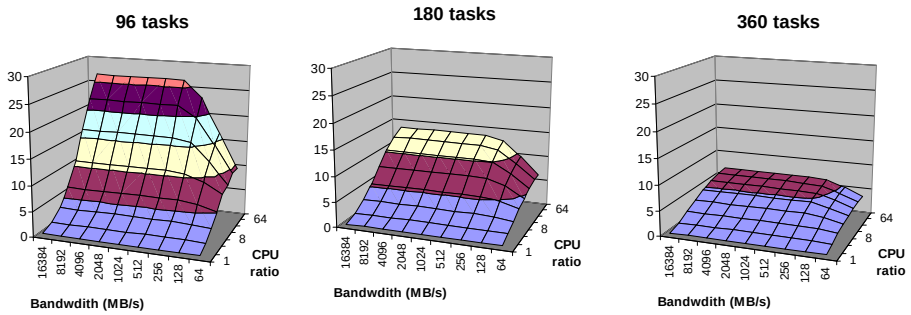


Boundary at 512-1024 MB/s network bandwidth

Speedup potential reduced with the scale

CG-POP mpi2s1D - 180x120 sensitivity

Parametric studies – sensitivity to network bandwith and CPU speed. Comparing 96, 180 and 360 (96 task scale, previous max 70)

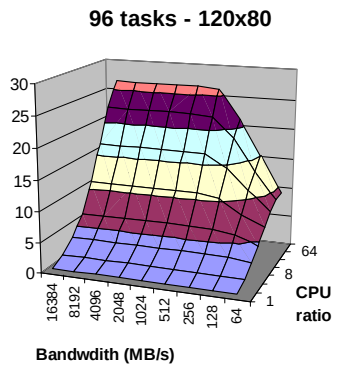


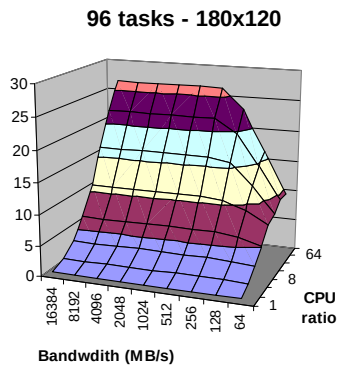
Boundary at 512-1024 MB/s network bandwidth

Speedup potential reduced with the scale

Block size impact

96 tasks – Very similar

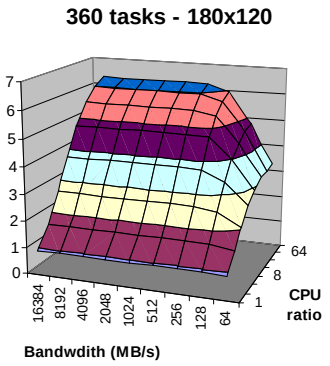


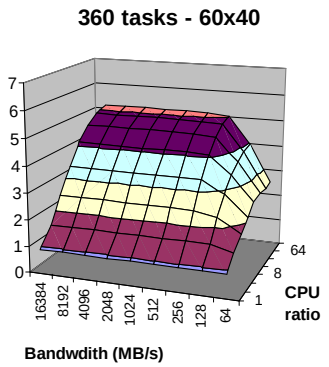


Still place for improvement?

Block size impact

360 tasks – Small impact, worst with small block size

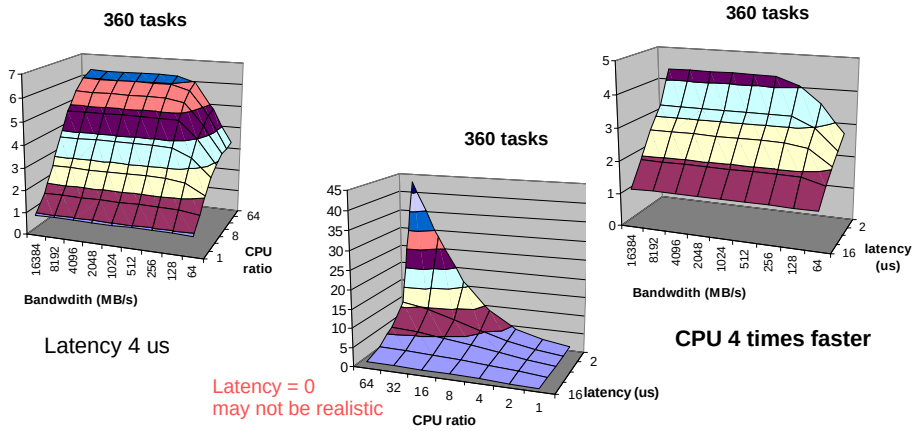




Reaching the plateau?

CG-POP mpi2s1D - 180x120 sensitivity

Sensitivity to CPU, Bandwidth and latency for 360 tasks case



Latency 4 us

Latency = 0
may not be realistic

CPU 4 times faster

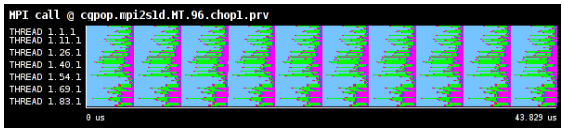
Bw 1024 MB/s

CG-POP mpi2s1D - 180x120 unbalance

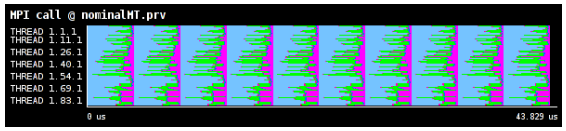
Study impact of balancing main computation

96 tasks – eliminate main unbalanced computation MPI calls view, same time scale

Instrumented run



Nominal simulation

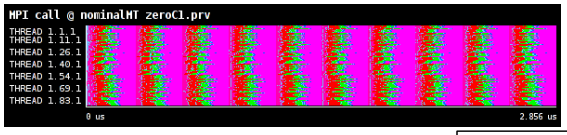


Nominal sim.
Zeroing main computation

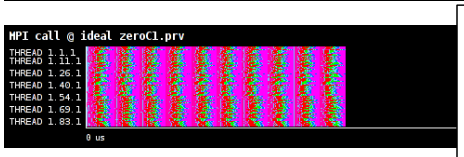
CG-POP mpi2s1D - 180x120 unbalance

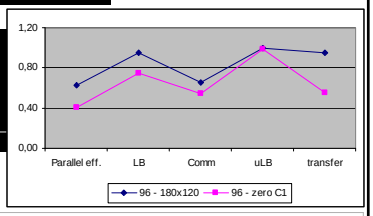
96 tasks – communication phase only: small computations between point 2 point + collective calls, MPI calls view, same time scale

Nominal sim. Zeroing main computation



Ideal network



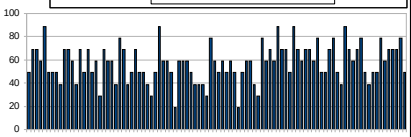


Unbalance in halo exchange

computation (75%), nb messages (65%)

Small amount of serialization

Large transfer time



Conclusions

- Sometimes things are not as they look like
  - Even the scalability is good, the parallel efficiency does not scale but IPC improvements compensate
- Small scale analysis can give hints to identify problems at large scale
  - IPC variability
  - Code replication
  - What-if analysis

INTELLECTUAL PROPERTY RIGHTS NOTICE:

- The User may only download, make and retain a **copy of the materials for his/her use for non-commercial** and research purposes.
- The User **may not commercially use the material**, unless has been granted prior written consent by the Licensor to do so; and **cannot remove, obscure or modify copyright notices**, text acknowledging or other means of identification or disclaimers as they appear.
- For further details, please contact **BSC-CNS** patc@bsc.es

PATC training, Barcelona, May 2012 ‹#›