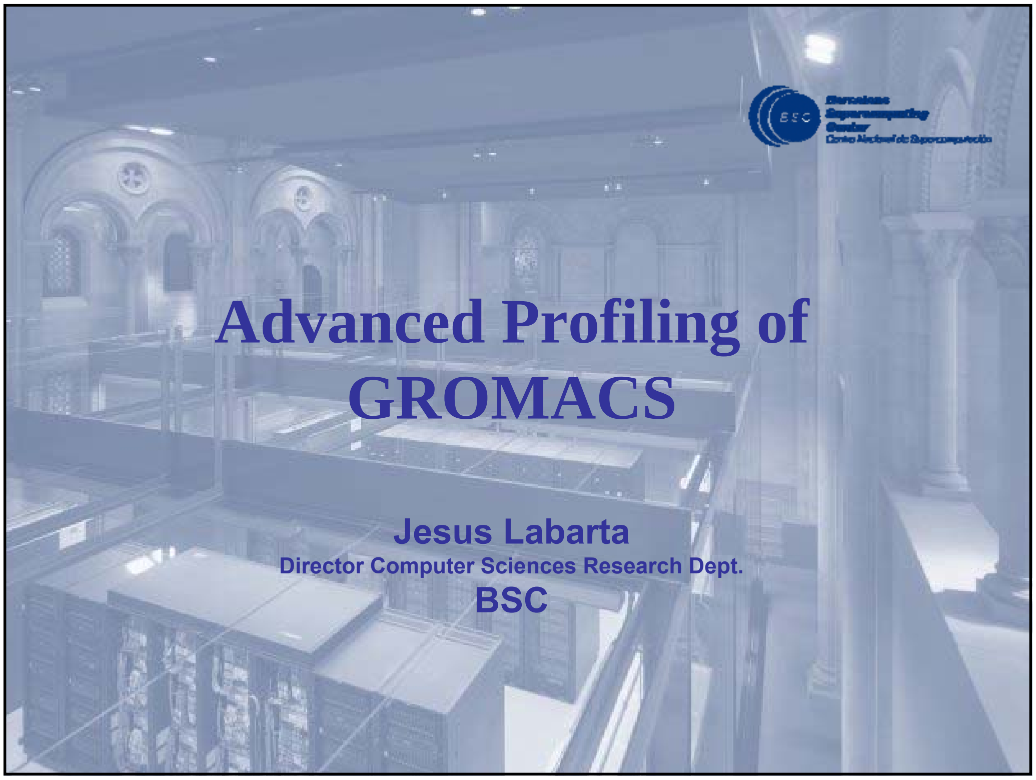


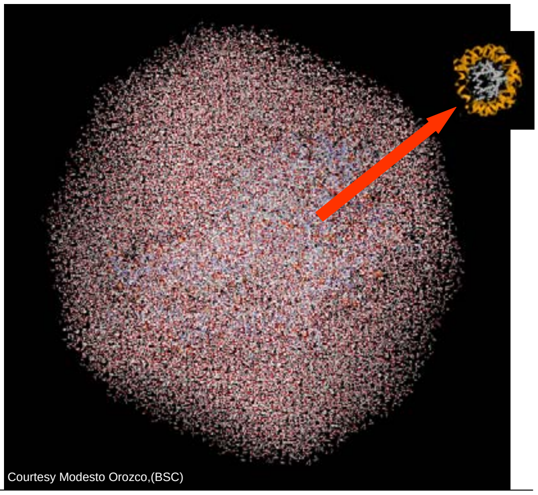
Advanced Profiling of GROMACS

Jesus Labarta
Director Computer Sciences Research Dept.
BSC

All I know about GROMACS

- A Molecular Dynamics application
- Heavily used @ BSC

- Not much ☹

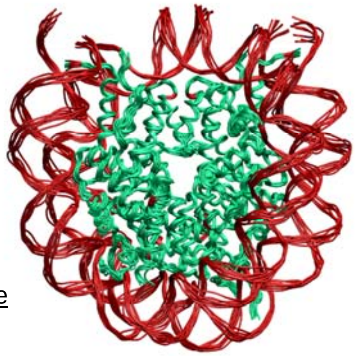


Courtesy Modesto Orozco,(BSC)

My interest

- Efficiency? Scalability?
- Can it be improved?

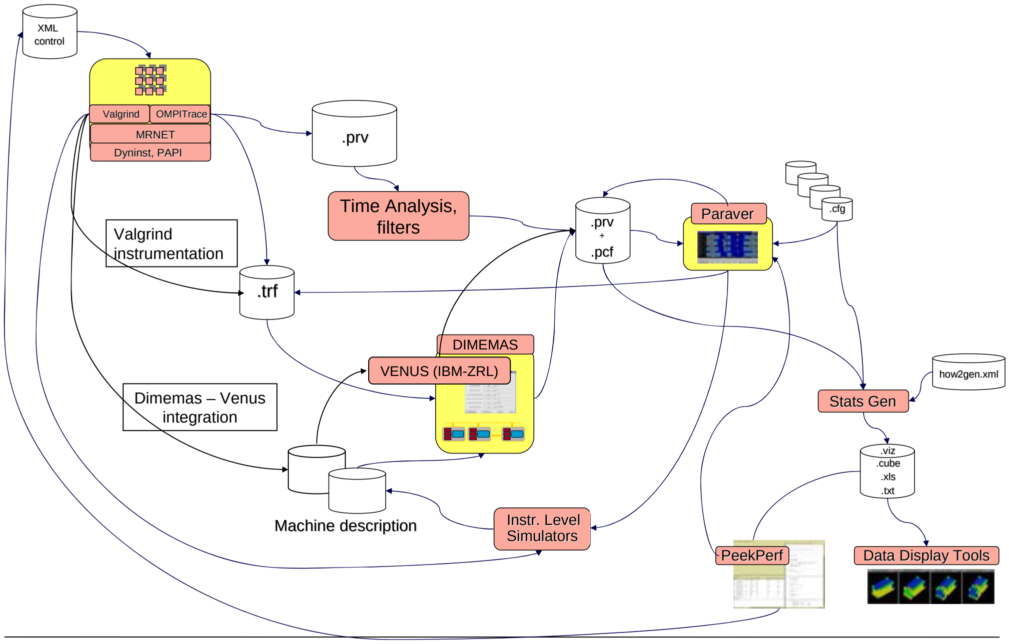
- Test case for our CEPBA-Tools environment
- Feedback/drive or research
  - Performance analysis tools
  - Programming models
  - Automatic Load balancing
  - Interconnects



Courtesy Agnes Noy (UB)

<u>Nucleosome test case</u>
145732 atoms
TIP3P water model
parmBSC0 force-field
Berendsen thermostat and barostat,
2 fs time step
constraints on Hydrogens

Evaluation infrastructure



XML control

Valgrind | OMPITrace

MRNET

Dyninst, PAPI

.prv

Time Analysis, filters

.prv + .pcf

Paraver

.cfg

Valgrind instrumentation

.trf

DIMEMAS

VENUS (IBM-ZRL)

Dimemas – Venus integration

Machine description

Instr. Level Simulators

Stats Gen

how2gen.xml

.viz .cube .xls .txt

PeekPerf

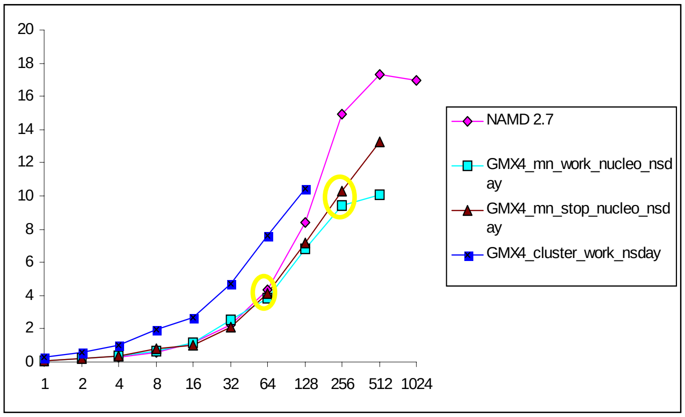
Data Display Tools

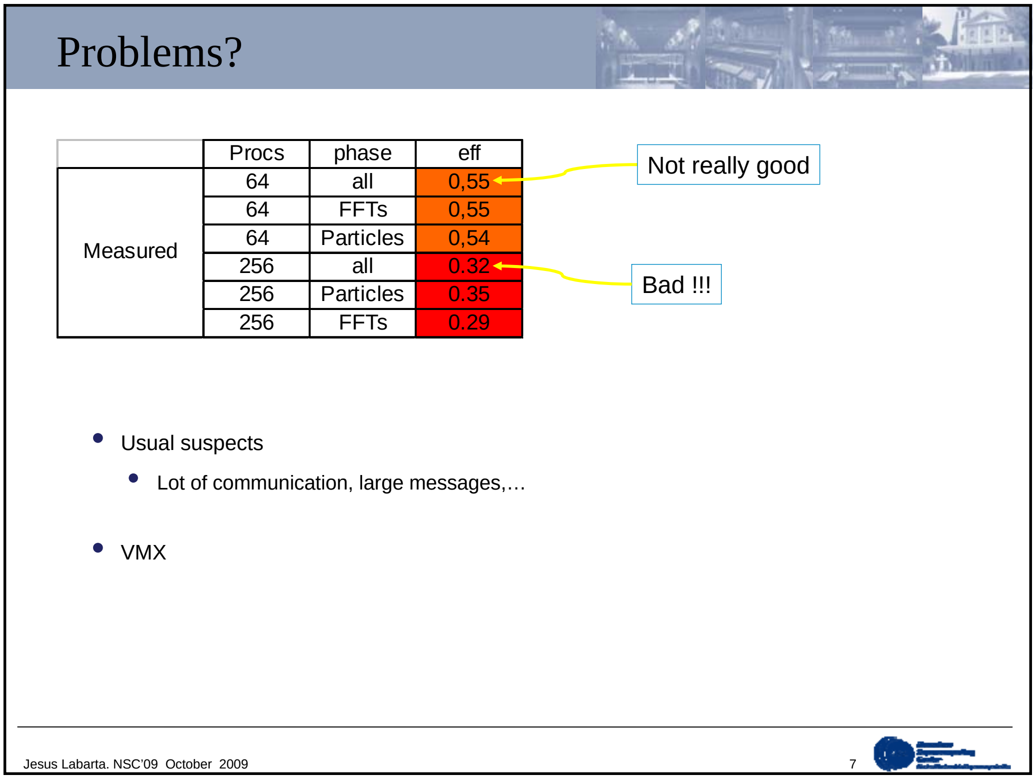
Index

- A story on
  - Performance
  - for (;;)
    - Parallel efficiency
    - Sequential performance (VMX,…)
  - Hybrid programming

Performance

- Raw user metric (ns/day)





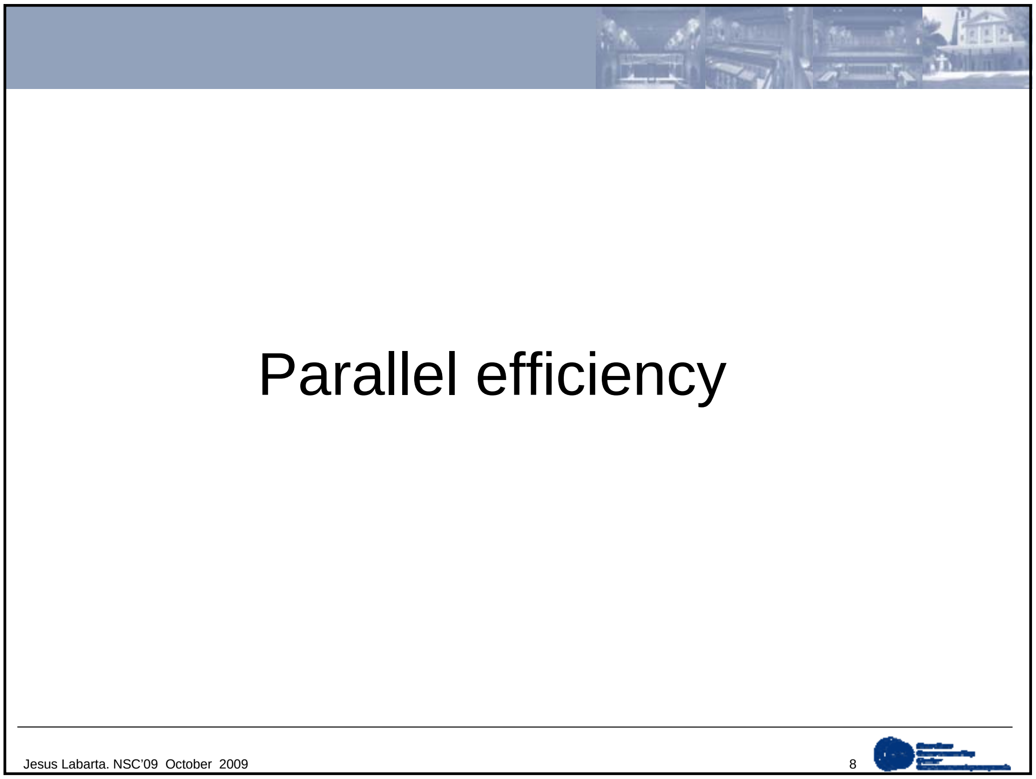
Problems?

	Procs	phase	eff
Measured	64	all	0,55
	64	FFTs	0,55
	64	Particles	0,54
	256	all	0.32
	256	Particles	0.35
	256	FFTs	0.29

Not really good

Bad !!!

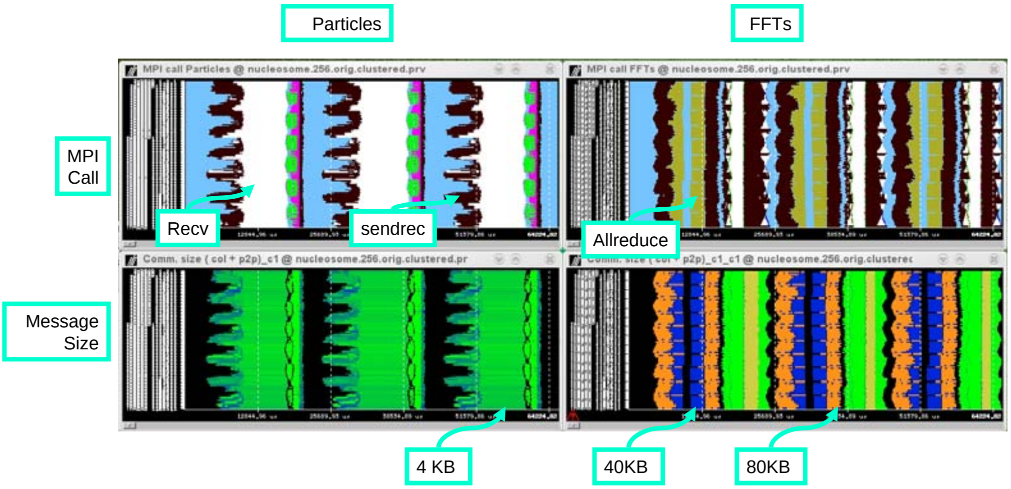
- Usual suspects
  - Lot of communication, large messages,…
- VMX



Parallel efficiency

Lots of communications, message sizes, …

- Suspect → convicted?


MPI call Particles @ nucleosome.256.orig.clustered.prv
MPI call FFTs @ nucleosome.256.orig.clustered.prv
Comm. size (col + p2p)_c1 @ nucleosome.256.orig.clustered.pr
Comm. size (col + p2p)_c1_c1 @ nucleosome.256.orig.clusterec

Particles

FFTs

MPI Call

Message Size

Recv

sendrec

Allreduce

4 KB

40KB

80KB

Convicted

- Who?
  - Machine? Program?
- Average Bandwidth per core ?
  - 12 -15 MB/s
- Ideal interconnect

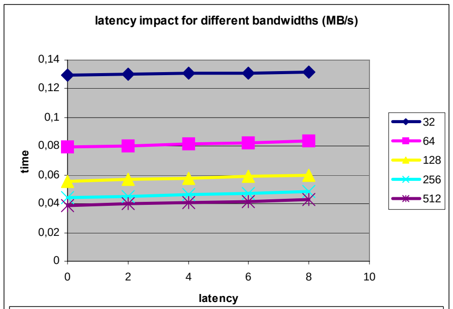
	Procs	phase	eff
Measured	64	all	0,55
	64	FFTs	0,55
	64	Particles	0,54
	256	all	0.32
	256	Particles	0.35
	256	FFTs	0.29
Prediction ideal	256	all	0.59
	256	Particles	0.63
	256	FFTs	0.52

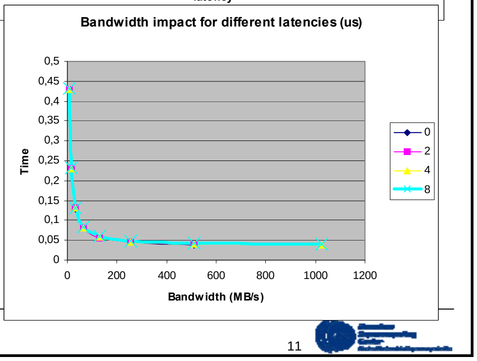
Better !!!!!!

Not that good !!!!!!

Sensitivity to interconnect parameters

	Procs	phase	eff
Measured	64	all	0,55
	64	FFTs	0,55
	64	Particles	0,54
	256	all	0.32
	256	Particles	0.35
	256	FFTs	0.29
Prediction ideal	256	all	0.59
	256	Particles	0.63
	256	FFTs	0.52





- Latency. Need to group messages?
- Sensitivity to network bandwidth?

VMX

VMX

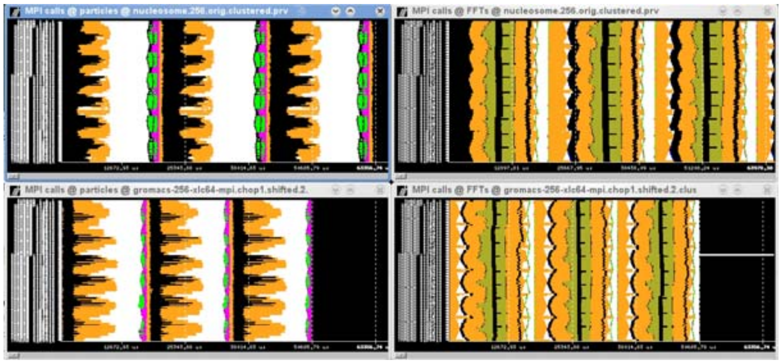
- Not as much gain as expected ???

Particles

FFTs

No VMX

VMX


MPI calls @ particles @ nucleosome.256.orig.clustered.prv
MPI calls @ FFTs @ nucleosome.256.orig.clustered.prv
MPI calls @ particles @ gromacs-256-xlc64-mpi.chop1.shifted.2.
MPI calls @ FFTs @ gromacs-256-xlc64-mpi.chop1.shifted.2.clus

VMX

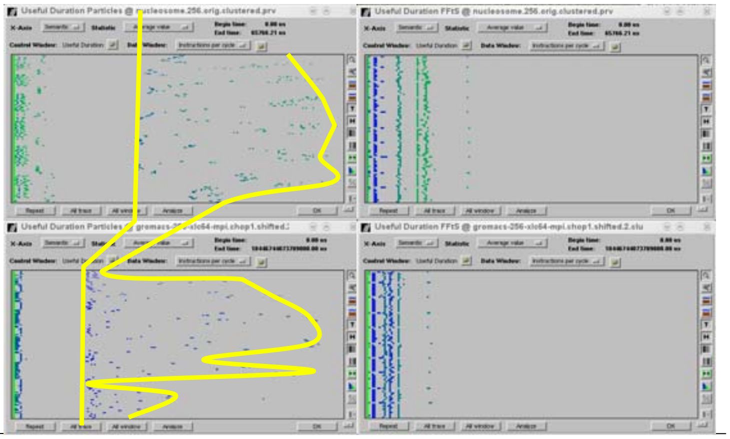
- Histogran of duration: Not uniform impact !!!!
  - Across Particle processes
  - Between FFTs and Particles FFTs

Particles

FFTs

No VMX

VMX

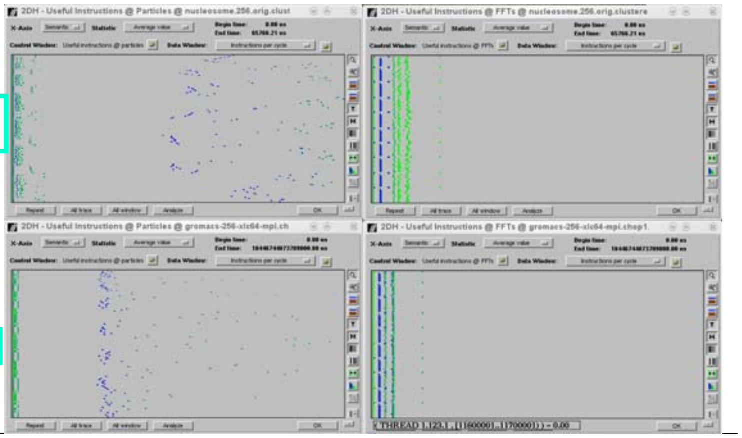

Useful Duration Particles @ nucleosome.256.orig.clustered.prv
Useful Duration FFtS @ nucleosome.256.orig.clustered.prv
Useful Duration Particles @ gromacs-256-xlc64-mpi.chop1.shifted.2
Useful Duration FFtS @ gromacs-256-xlc64-mpi.chop1.shifted.2.clu
X-Axis
Statistic
Control Window:
Data Window:
Begin time:
End time:
Repeat
All trace
All window
Analyze
OK

VMX

- Histograms of Instructions

Particles

FFTs


2DH - Useful Instructions @ Particles @ nucleosome.256.orig.clust
2DH - Useful Instructions @ FFTs @ nucleosome.256.orig.clustere
2DH - Useful Instructions @ Particles @ gromacs-256-xlc64-mpi.ch
2DH - Useful Instructions @ FFTs @ gromacs-256-xlc64-mpi.chop1.
X-Axis
Statistic
Control Window:
Data Window:
Begin time:
End time:
Repeat
All trace
All window
Analyze
OK

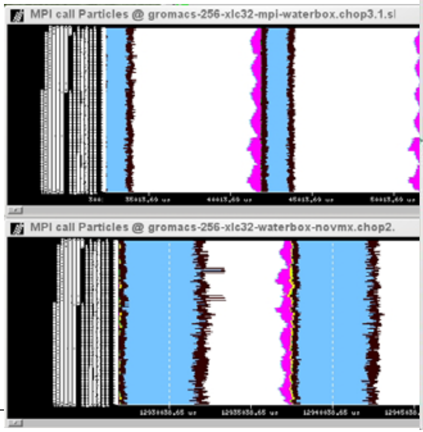
THREAD 1.123.1 ,[11600001..11700001) = 0.00

No VMX

VMX

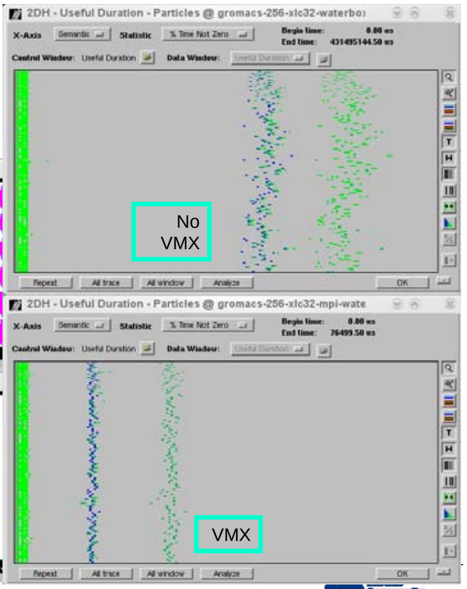
VMX @ waterbox

- Impact depends on actual data !!!


MPI call Particles @ gromacs-256-xlc32-mpi-waterbox.chop3.1.sl
MPI call Particles @ gromacs-256-xlc32-waterbox-novmx.chop2.

VMX

No VMX



No VMX

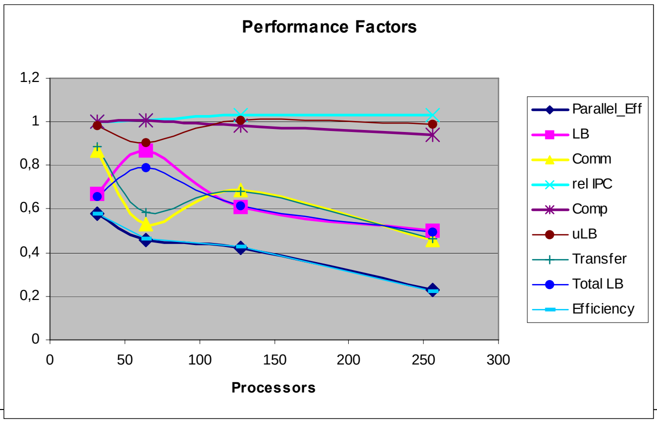
VMX

Back to parallel performance

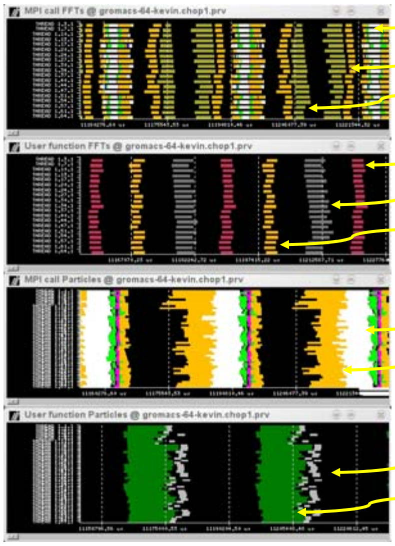
Performance factors

$$Parallel\_\eta = \overbrace{LB * \rlap{\underbrace{}_{Comm}}microLB}^{TotalLB} * Transfer$$

$$\eta = Parallel\_\eta * relIPC * Comp$$



Structure: A few colors


MPI call FFTs @ gromacs-64-kevin.chop1.prv
User function FFTs @ gromacs-64-kevin.chop1.prv
MPI call Particles @ gromacs-64-kevin.chop1.prv
User function Particles @ gromacs-64-kevin.chop1.prv

MPI call FFTs

recv

sendrec

allreduce

User functions @ FFTSs

make_bsplines

gather_f_bsplines

solve_pme

MPI call Particles

recv

sendrec

User functions @ Particles

ewald_LRcorrection

do_nonbonded

Scaling structure

User function @ Particles

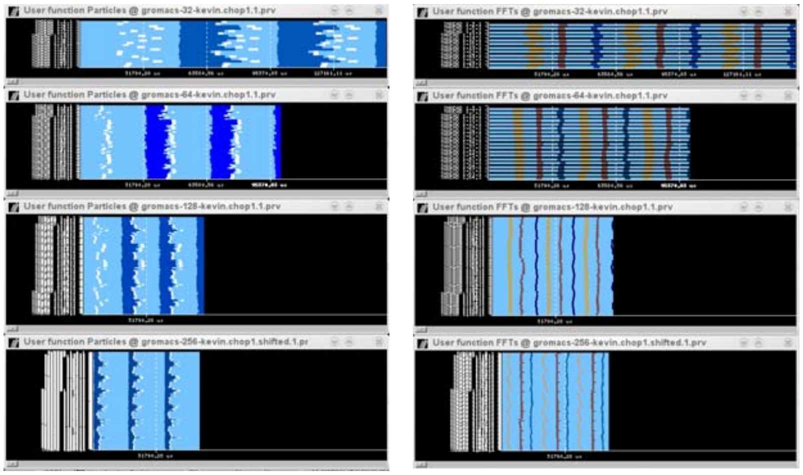
User function @ FFTSs

32

64

128

256


User function Particles @ gromacs-32-kevin.chop1.1.prv
User function FFTs @ gromacs-32-kevin.chop1.1.prv
User function Particles @ gromacs-64-kevin.chop1.1.prv
User function FFTs @ gromacs-64-kevin.chop1.1.prv
User function Particles @ gromacs-128-kevin.chop1.1.prv
User function FFTs @ gromacs-128-kevin.chop1.1.prv
User function Particles @ gromacs-256-kevin.chop1.shifted.1.pr
User function FFTs @ gromacs-256-kevin.chop1.shifted.1.prv

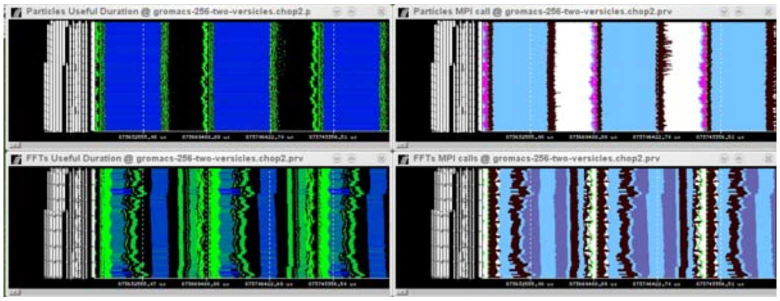
PRACE Test case

Useful Duration

MPI calls

Particles

FFTs


Particles Useful Duration @ gromacs-256-two-versicles.chop2.p
Particles MPI call @ gromacs-256-two-versicles.chop2.prv
FFTs Useful Duration @ gromacs-256-two-versicles.chop2.prv
FFTs MPI calls @ gromacs-256-two-versicles.chop2.prv

- Average point to point bandwidth: 13 MB/s

PRACE Test case

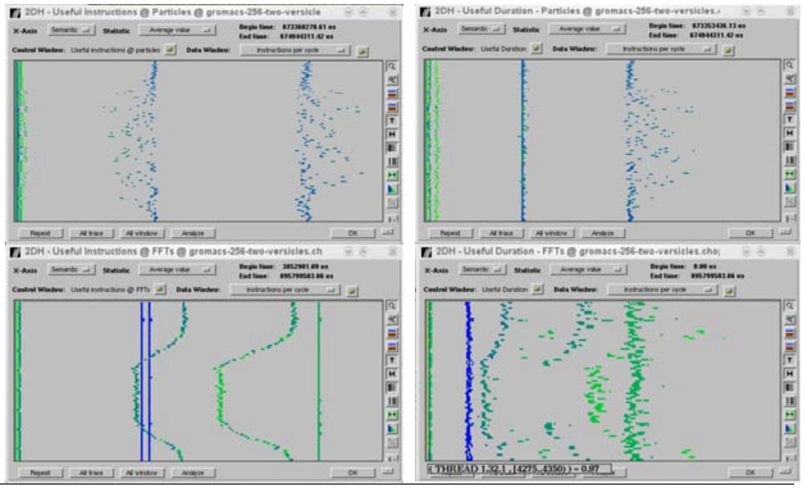
- Effects → opportunities

Instructions

Duration

Particles

FFTs


2DH - Useful Instructions @ Particles @ gromacs-256-two-versicle
2DH - Useful Duration - Particles @ gromacs-256-two-versicles.
2DH - Useful Instructions @ FFTs @ gromacs-256-two-versicles.ch
2DH - Useful Duration - FFTs @ gromacs-256-two-versicles.cho
Repeat
All trace
All window
Analyze
OK

X-Axis: Semantic Statistic: Average value Begin time: 873368278.61 us End time: 874944311.42 us
Control Window: Useful instructions @ particles Data Window: Instructions per cycle

X-Axis: Semantic Statistic: Average value Begin time: 873353436.13 us End time: 874944311.42 us
Control Window: Useful Duration Data Window: Instructions per cycle

X-Axis: Semantic Statistic: Average value Begin time: 3052981.09 us End time: 895799583.86 us
Control Window: Useful instructions @ FFTs Data Window: Instructions per cycle

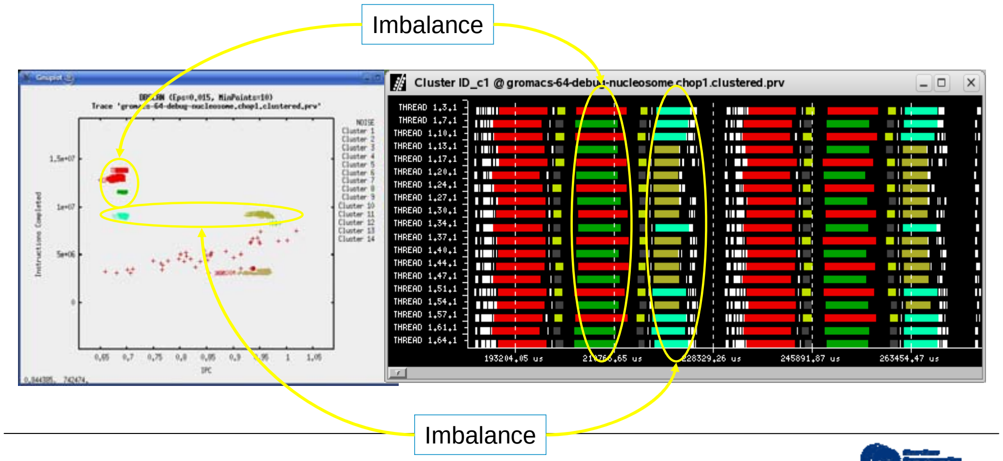
X-Axis: Semantic Statistic: Average value Begin time: 0.00 us End time: 895799583.86 us
Control Window: Useful Duration Data Window: Instructions per cycle

(THREAD 1.32.1 , [4275..4350)) = 0.87

Back to sequential performance

64 processes: FFTs

- Globally balanced but some local imbalance

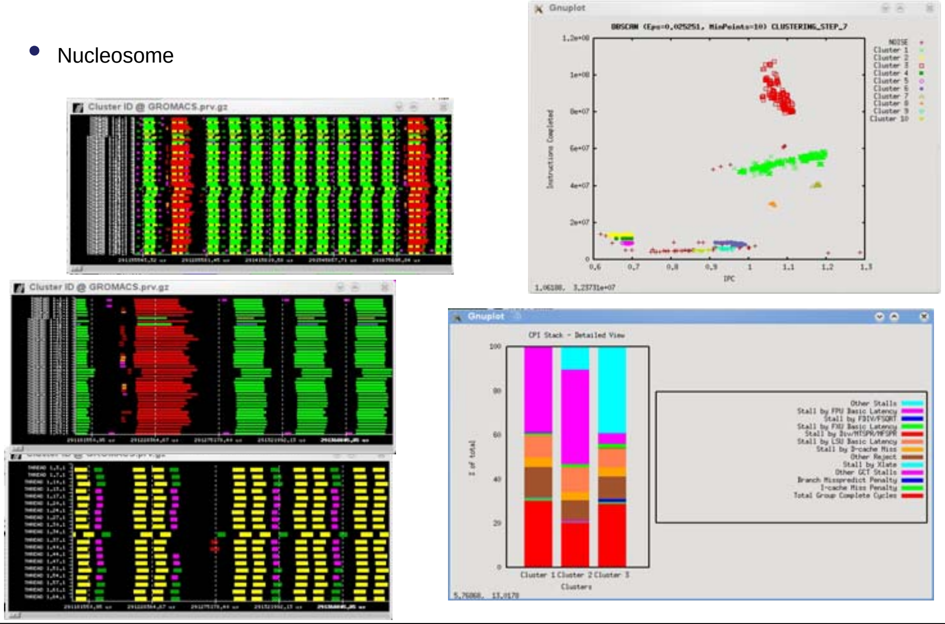


Imbalance

Imbalance

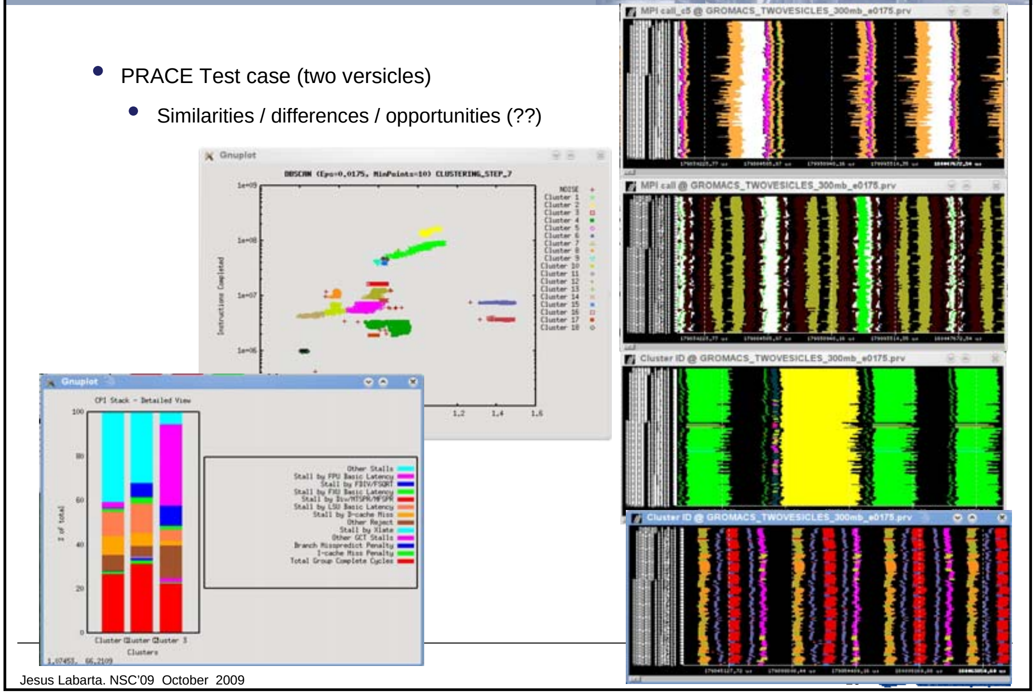
Online analysis

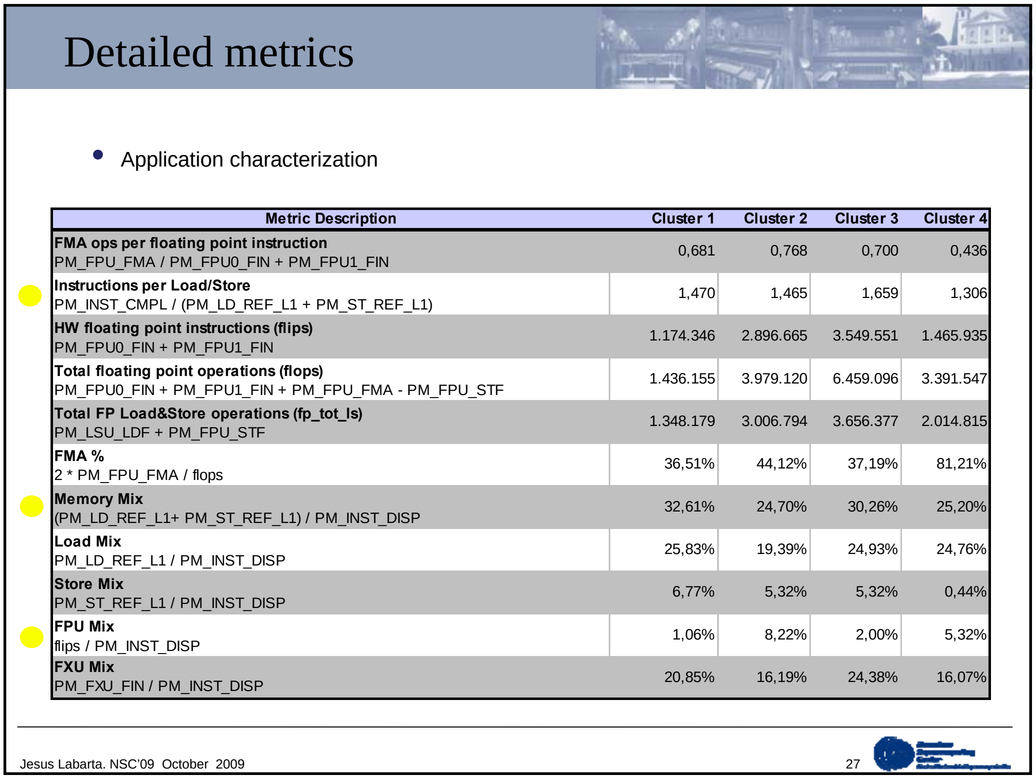
- Nucleosome



Online analysis

- PRACE Test case (two versicles)
  - Similarities / differences / opportunities (??)

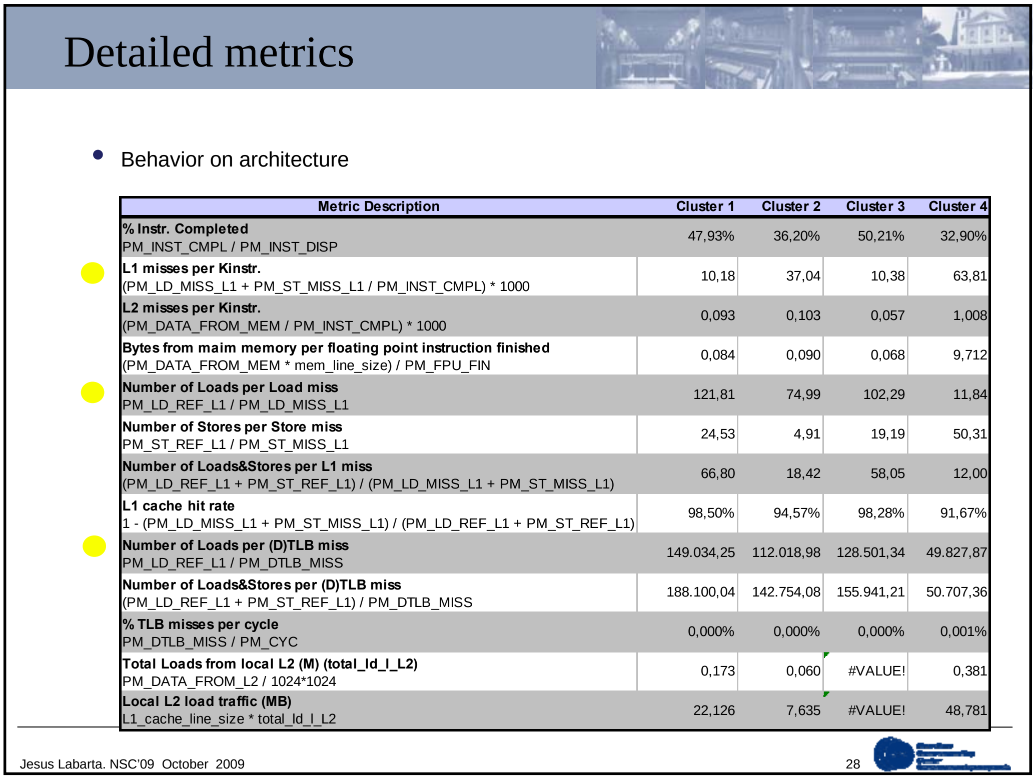

MPI call_c5 @ GROMACS_TWOVESICLES_300mb_e0175.prv
MPI call @ GROMACS_TWOVESICLES_300mb_e0175.prv
Cluster ID @ GROMACS_TWOVESICLES_300mb_e0175.prv
Cluster ID @ GROMACS_TWOVESICLES_300mb_e0175.prv
Gnuplot
DBSCAN (Eps=0,0175, MinPoints=10) CLUSTERING_STEP_7
Instructions Completed
CPI Stack - Detailed View
% of total
Cluster 1
Cluster 2
Cluster 3
Clusters
Other Stalls
Stall by FPU Basic Latency
Stall by FDIV/FSQRT
Stall by FXU Basic Latency
Stall by Div/MTSPR/MFSPR
Stall by LSU Basic Latency
Stall by D-cache Miss
Other Reject
Stall by Xlate
Other GCT Stalls
Branch Misspredict Penalty
I-cache Miss Penalty
Total Group Complete Cycles



Detailed metrics

- Application characterization

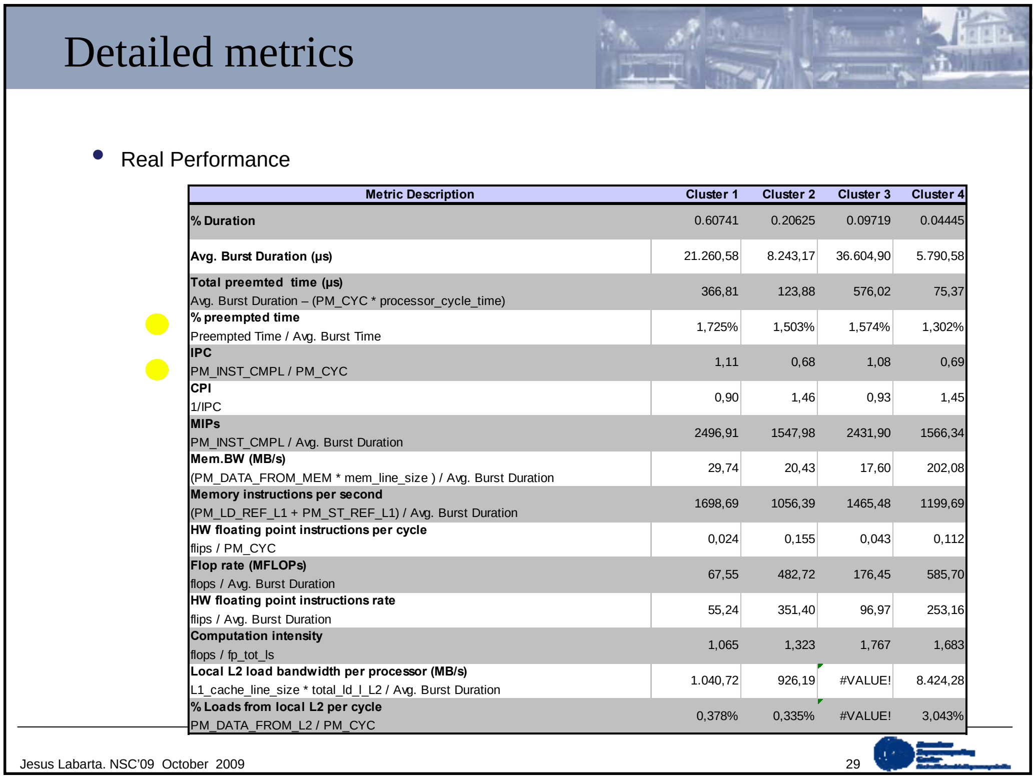
Metric Description	Cluster 1	Cluster 2	Cluster 3	Cluster 4
FMA ops per floating point instruction PM_FPU_FMA / PM_FPU0_FIN + PM_FPU1_FIN	0,681	0,768	0,700	0,436
● **Instructions per Load/Store** PM_INST_CMPL / (PM_LD_REF_L1 + PM_ST_REF_L1)	1,470	1,465	1,659	1,306
HW floating point instructions (flips) PM_FPU0_FIN + PM_FPU1_FIN	1.174.346	2.896.665	3.549.551	1.465.935
Total floating point operations (flops) PM_FPU0_FIN + PM_FPU1_FIN + PM_FPU_FMA - PM_FPU_STF	1.436.155	3.979.120	6.459.096	3.391.547
Total FP Load&Store operations (fp_tot_ls) PM_LSU_LDF + PM_FPU_STF	1.348.179	3.006.794	3.656.377	2.014.815
FMA % 2 * PM_FPU_FMA / flops	36,51%	44,12%	37,19%	81,21%
● **Memory Mix** (PM_LD_REF_L1+ PM_ST_REF_L1) / PM_INST_DISP	32,61%	24,70%	30,26%	25,20%
Load Mix PM_LD_REF_L1 / PM_INST_DISP	25,83%	19,39%	24,93%	24,76%
Store Mix PM_ST_REF_L1 / PM_INST_DISP	6,77%	5,32%	5,32%	0,44%
● **FPU Mix** flips / PM_INST_DISP	1,06%	8,22%	2,00%	5,32%
FXU Mix PM_FXU_FIN / PM_INST_DISP	20,85%	16,19%	24,38%	16,07%



Detailed metrics

- Behavior on architecture

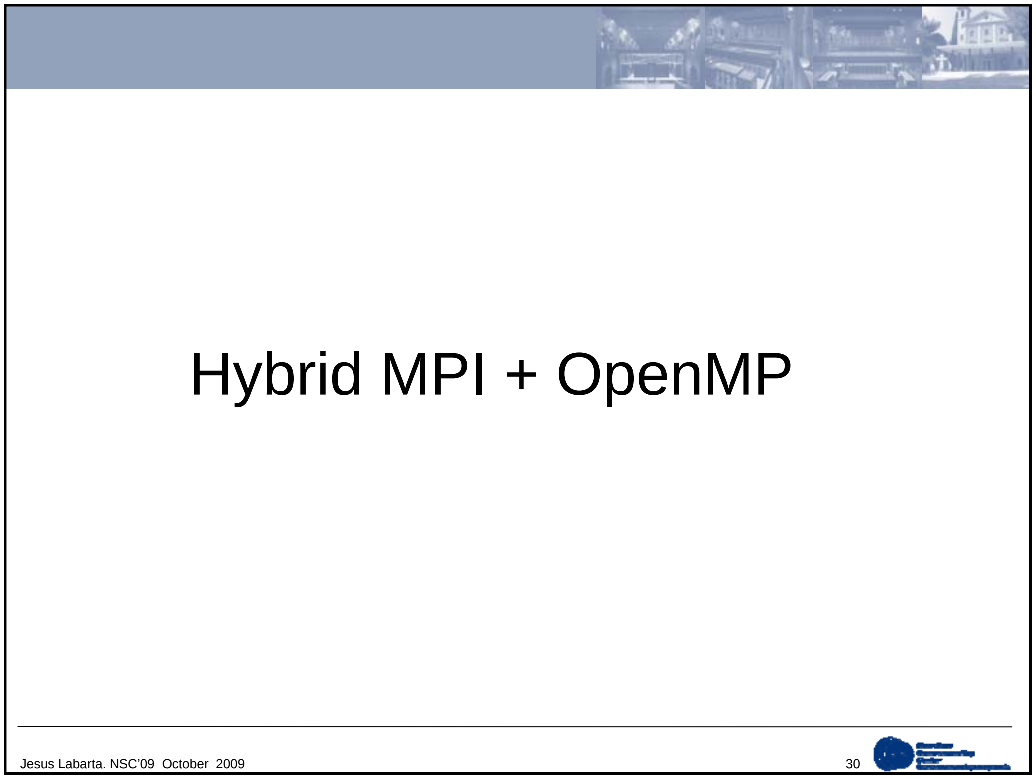
Metric Description	Cluster 1	Cluster 2	Cluster 3	Cluster 4
% Instr. Completed PM_INST_CMPL / PM_INST_DISP	47,93%	36,20%	50,21%	32,90%
● **L1 misses per Kinstr.** (PM_LD_MISS_L1 + PM_ST_MISS_L1 / PM_INST_CMPL) * 1000	10,18	37,04	10,38	63,81
L2 misses per Kinstr. (PM_DATA_FROM_MEM / PM_INST_CMPL) * 1000	0,093	0,103	0,057	1,008
Bytes from maim memory per floating point instruction finished (PM_DATA_FROM_MEM * mem_line_size) / PM_FPU_FIN	0,084	0,090	0,068	9,712
● **Number of Loads per Load miss** PM_LD_REF_L1 / PM_LD_MISS_L1	121,81	74,99	102,29	11,84
Number of Stores per Store miss PM_ST_REF_L1 / PM_ST_MISS_L1	24,53	4,91	19,19	50,31
Number of Loads&Stores per L1 miss (PM_LD_REF_L1 + PM_ST_REF_L1) / (PM_LD_MISS_L1 + PM_ST_MISS_L1)	66,80	18,42	58,05	12,00
L1 cache hit rate 1 - (PM_LD_MISS_L1 + PM_ST_MISS_L1) / (PM_LD_REF_L1 + PM_ST_REF_L1)	98,50%	94,57%	98,28%	91,67%
● **Number of Loads per (D)TLB miss** PM_LD_REF_L1 / PM_DTLB_MISS	149.034,25	112.018,98	128.501,34	49.827,87
Number of Loads&Stores per (D)TLB miss (PM_LD_REF_L1 + PM_ST_REF_L1) / PM_DTLB_MISS	188.100,04	142.754,08	155.941,21	50.707,36
% TLB misses per cycle PM_DTLB_MISS / PM_CYC	0,000%	0,000%	0,000%	0,001%
Total Loads from local L2 (M) (total_ld_l_L2) PM_DATA_FROM_L2 / 1024*1024	0,173	0,060	#VALUE!	0,381
Local L2 load traffic (MB) L1_cache_line_size * total_ld_l_L2	22,126	7,635	#VALUE!	48,781



Detailed metrics

- Real Performance

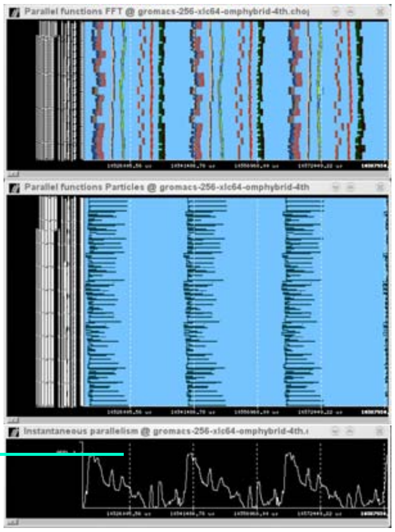
Metric Description	Cluster 1	Cluster 2	Cluster 3	Cluster 4
% Duration	0.60741	0.20625	0.09719	0.04445
Avg. Burst Duration (µs)	21.260,58	8.243,17	36.604,90	5.790,58
Total preemted time (µs) Avg. Burst Duration – (PM_CYC * processor_cycle_time)	366,81	123,88	576,02	75,37
% preempted time Preempted Time / Avg. Burst Time	1,725%	1,503%	1,574%	1,302%
IPC PM_INST_CMPL / PM_CYC	1,11	0,68	1,08	0,69
CPI 1/IPC	0,90	1,46	0,93	1,45
MIPs PM_INST_CMPL / Avg. Burst Duration	2496,91	1547,98	2431,90	1566,34
Mem.BW (MB/s) (PM_DATA_FROM_MEM * mem_line_size) / Avg. Burst Duration	29,74	20,43	17,60	202,08
Memory instructions per second (PM_LD_REF_L1 + PM_ST_REF_L1) / Avg. Burst Duration	1698,69	1056,39	1465,48	1199,69
HW floating point instructions per cycle flips / PM_CYC	0,024	0,155	0,043	0,112
Flop rate (MFLOPs) flops / Avg. Burst Duration	67,55	482,72	176,45	585,70
HW floating point instructions rate flips / Avg. Burst Duration	55,24	351,40	96,97	253,16
Computation intensity flops / fp_tot_ls	1,065	1,323	1,767	1,683
Local L2 load bandwidth per processor (MB/s) L1_cache_line_size * total_ld_l_L2 / Avg. Burst Duration	1.040,72	926,19	#VALUE!	8.424,28
% Loads from local L2 per cycle PM_DATA_FROM_L2 / PM_CYC	0,378%	0,335%	#VALUE!	3,043%



Hybrid MPI + OpenMP

Hybrid parallelization

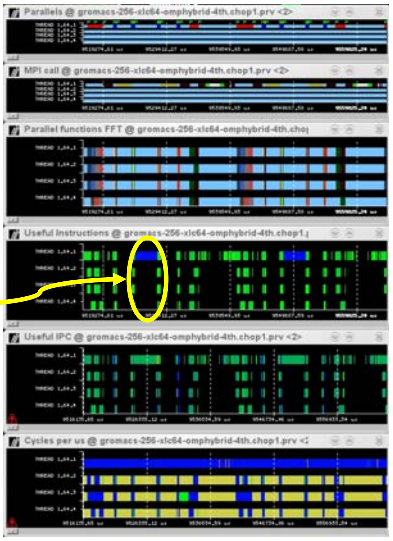
- OpenMP parallelization
  - Particles by Christian Simarro (BSC)
  - FFTs by Sebastian von Alfthan (CSC)
- 256 MPI x 4 openMP


Parallel functions FFT @ gromacs-256-xlc64-omphybrid-4th.cho
Parallel functions Particles @ gromacs-256-xlc64-omphybrid-4th
Instantaneous parallelism @ gromacs-256-xlc64-omphybrid-4th.

130 procs!!!!

Hybrid parallelization

- FFTs
  - Partial parallelization
  - Too synchronous
    - MPI – OpenMP –MPI – OpenMP ….
  - OS scheduling issues


Parallels @ gromacs-256-xlc64-omphybrid-4th.chop1.prv <2>
MPI call @ gromacs-256-xlc64-omphybrid-4th.chop1.prv <2>
Parallel functions FFT @ gromacs-256-xlc64-omphybrid-4th.cho
Useful Instructions @ gromacs-256-xlc64-omphybrid-4th.chop1.
Useful IPC @ gromacs-256-xlc64-omphybrid-4th.chop1.prv <2>
Cycles per us @ gromacs-256-xlc64-omphybrid-4th.chop1.prv <

Not Parallelized

Hybrid parallelization

- Particles

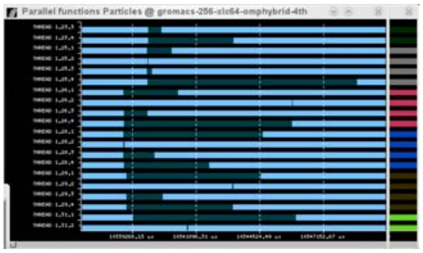
```
for(n=n0; (n<n1);
#pragma omp par              or private(tabledata, nrnb_ind, nlist, outeriter,
                        inneriter, tabletype, kernelptr) shared(fr,nblists)
   for(i=i0       1); i++) {
      …
      if (something)
         compute();
   }
```

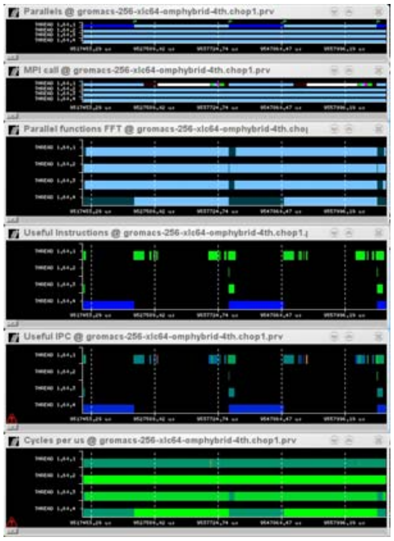
Right place?

Very imbalanced

Few iterations

Big difference across processes

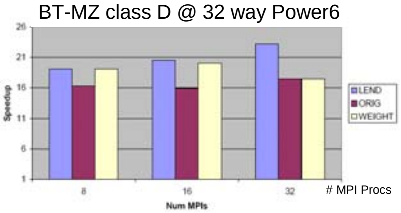

Parallel functions Particles @ gromacs-256-xlc64-omphybrid-4th

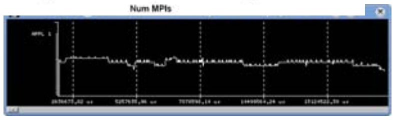

Parallels @ gromacs-256-xlc64-omphybrid-4th.chop1.prv
MPI call @ gromacs-256-xlc64-omphybrid-4th.chop1.prv
Parallel functions FFT @ gromacs-256-xlc64-omphybrid-4th.cho
Useful instructions @ gromacs-256-xlc64-omphybrid-4th.chop1.
Useful IPC @ gromacs-256-xlc64-omphybrid-4th.chop1.prv
Cycles per us @ gromacs-256-xlc64-omphybrid-4th.chop1.prv

Dynamic Load Balancing run time

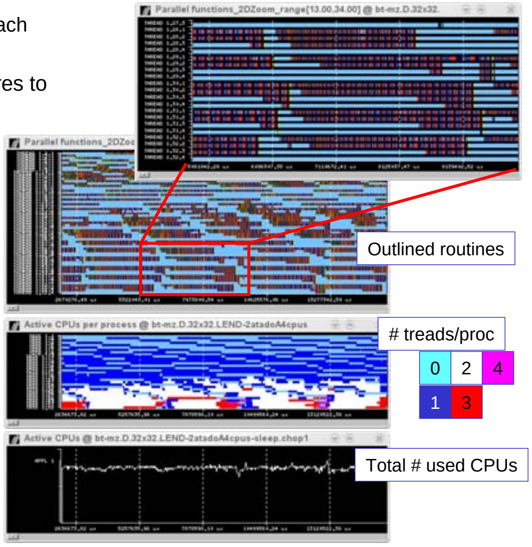
- MPI + OpenMP (StarSs)
  - Job is started with P MPI processes each with 1 OpenMP thread
- Processes blocking at MPI calls lend cores to other processes
- GADGET @ 800 cores: Sup=2.5

BT-MZ class D @ 32 way Power6





BT-MZ class D @ 32 way Power6

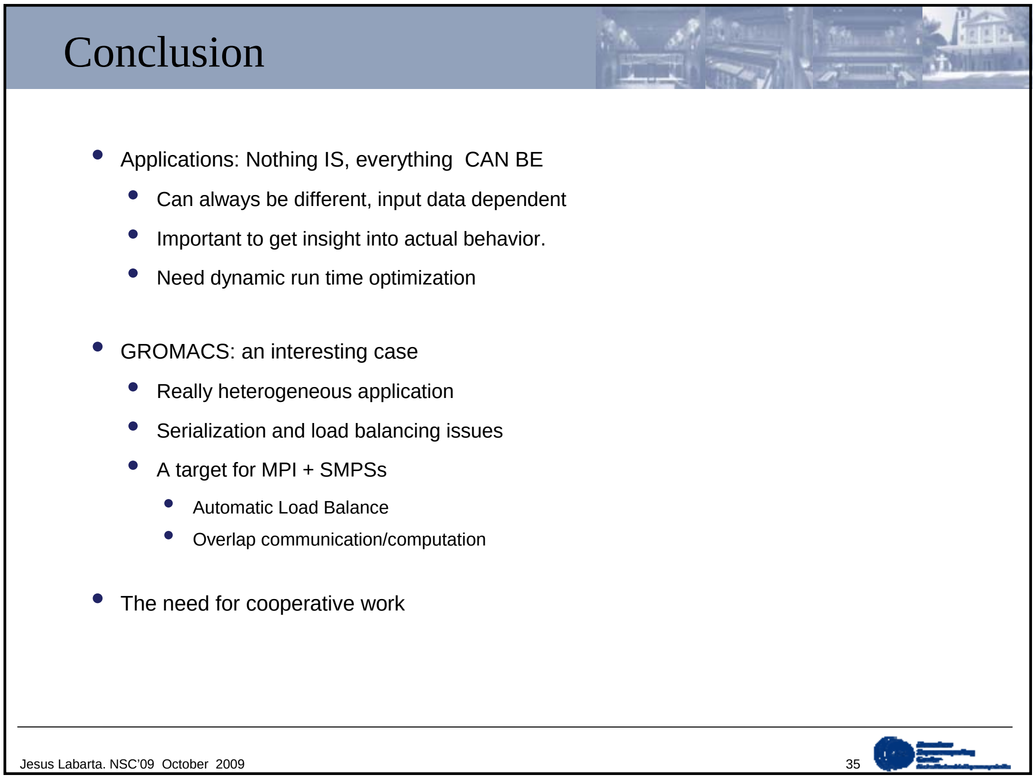

Parallel functions_2DZoom_range[13.00,34.00] @ bt-mz.D.32x32.
Parallel functions_2DZoo
Active CPUs per process @ bt-mz.D.32x32.LEND-2atadoA4cpus
Active CPUs @ bt-mz.D.32x32.LEND-2atadoA4cpus-sleep.chop1

Outlined routines

# treads/proc

0	2	4
1	3	

Total # used CPUs



Conclusion

- Applications: Nothing IS, everything CAN BE
  - Can always be different, input data dependent
  - Important to get insight into actual behavior.
  - Need dynamic run time optimization

- GROMACS: an interesting case
  - Really heterogeneous application
  - Serialization and load balancing issues
  - A target for MPI + SMPSs
    - Automatic Load Balance
    - Overlap communication/computation

- The need for cooperative work