Introduction to BSC-Tools

Germán Llort

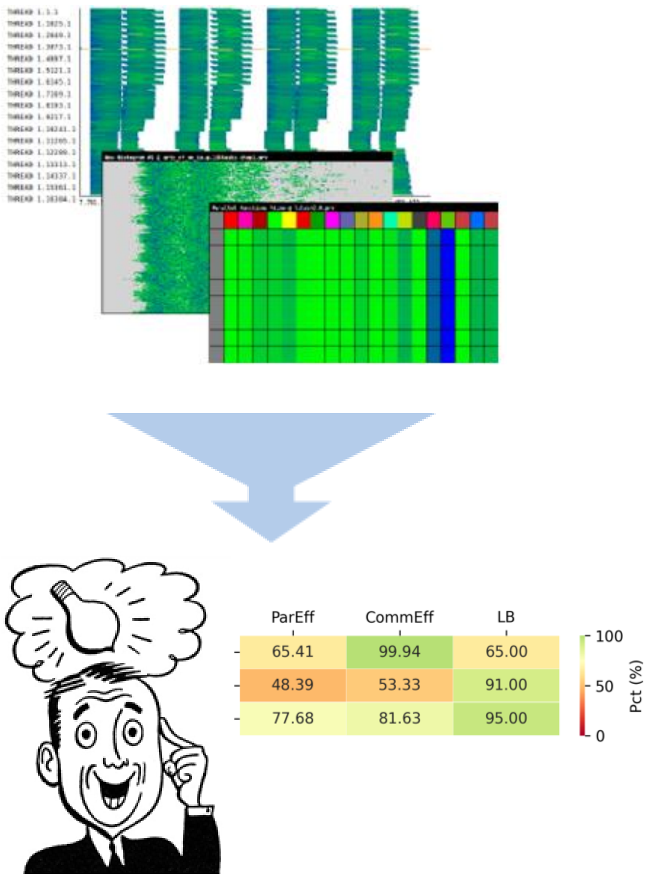
gllort@bsc.es

DEEP-SEA Seminar

05/11/2021

BSC Tools

- Since 1991
- Based on traces
- Open source – https://tools.bsc.es
- Focus: Detail, variability, flexibility
- Core tools
  - Extrae – Instrumentation
  - Paraver – Offline trace analysis
  - Dimemas – Message-passing simulator
- Performance Analytics
  - Leveraging techniques from data analytics
  - Towards insight and models



Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Extrae: Flexible instrumentation…

- Platforms
  - Intel, Cray, BlueGene, MIC, ARM, Android, Fujitsu Sparc …
- Parallel programming models
  - MPI, OpenMP, pthreads, OmpSs, CUDA, OpenCL, Java, Python …
- Performance Counters
  - Using PAPI interface
- Link to source code
  - Callstack at MPI routines
  - OpenMP outlined routines
  - Selected user functions (Dyninst)
- Periodic sampling
- User events anywhere in your program (Extrae API)

No need to recompile nor relink!

… of unmodified binaries…

- Symbol substitution through LD_PRELOAD

```
export LD_PRELOAD=$EXTRAE_HOME/lib/libmpitrace.so
```

  - Specific libraries for each runtime and combinations
    - MPI
    - OpenMP
    - OpenMP+MPI
    - …

Recommended

- Dynamic instrumentation
  - Based on Dyninst (developed by U.Wisconsin / U.Maryland)
    - Instrumentation in memory
    - Binary rewriting

- Static link (i.e., PMPI, Extrae API)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

… with low overhead

	MN4
Punctual event	166 ns
Event + PAPI counters	751 ns
Event + 1-level callstack	2.875 us
Event + 6-levels callstack	6.109 us

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Extrae: Quick start guide

- Sample SLURM jobscript

```
#!/bin/bash
#SBATCH --job-name=lulesh
#SBATCH --output=lulesh_27.out
#SBATCH --error=lulesh_27.err
#SBATCH --ntasks=27
#SBATCH --cpus-per-task=1
#SBATCH --time=00:10:00
#SBATCH --exclusive

srun --cpu_bind=cores
  ../apps/lulesh2.0_mpi -i 10 -p -s 65
```

Request resources

Run the program

Extrae: Quick start guide (II)

```
#!/bin/bash
#SBATCH --job-name=lulesh
#SBATCH --output=lulesh_27.out
#SBATCH --error=lulesh_27.err
#SBATCH --ntasks=27
#SBATCH --cpus-per-task=1
#SBATCH --time=00:10:00
#SBATCH --exclusive


srun --cpu_bind=cores ./trace.sh
  ../apps/lulesh2.0_mpi -i 10 -p -s 65
```

```
#!/bin/bash

# Configure Extrae
export EXTRAE_CONFIG_FILE=./extrae.xml
export EXTRAE_HOME=<path-to-installation>

# Load the tracing library (choose RT/C/Fortran)
export LD_PRELOAD=$EXTRAE_HOME/lib/libmpitrace.so

# Run the program
$*
```

What to trace?

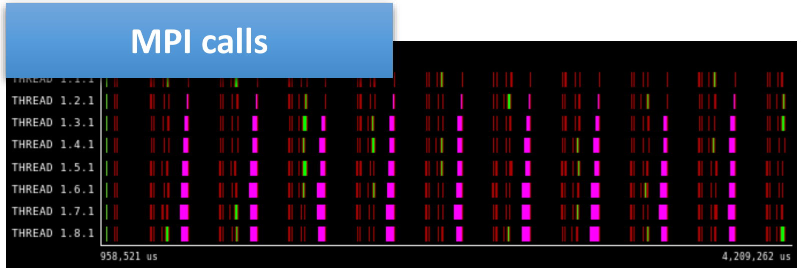
Choose tracing library

- libseqtrace
- libmpitrace[f]
- libomptrace
- libpttrace
- libcudatrace
- libompitrace[f]
- libptmpitrace[f]
- libcudampitrace[f]
- …

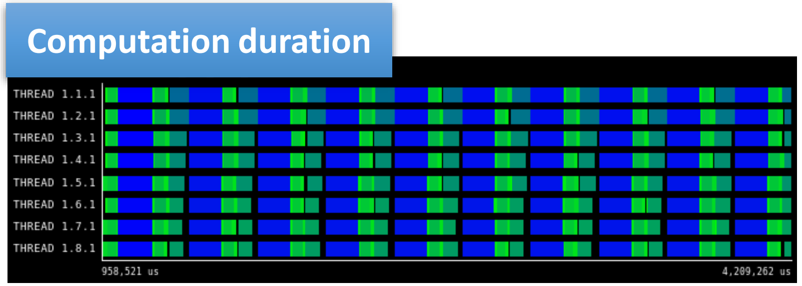
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Understanding applications with Paraver

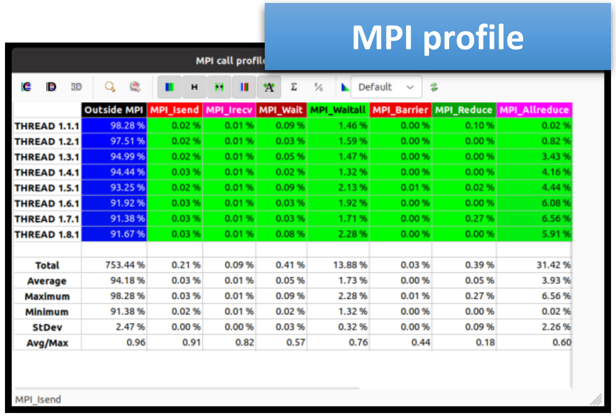
- Timeline views
  - Categorical data – color encoding



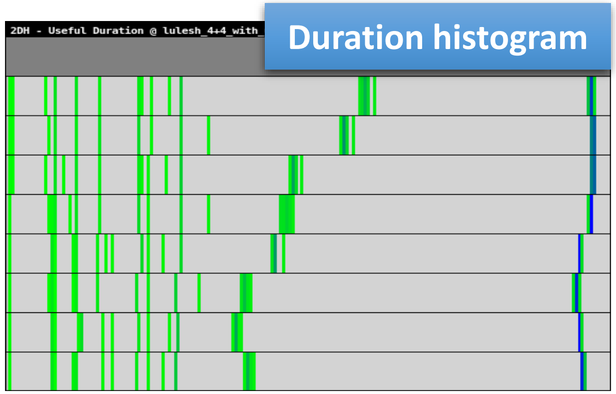
  - Continuous data – gradient (green to blue)



- From timelines to tables

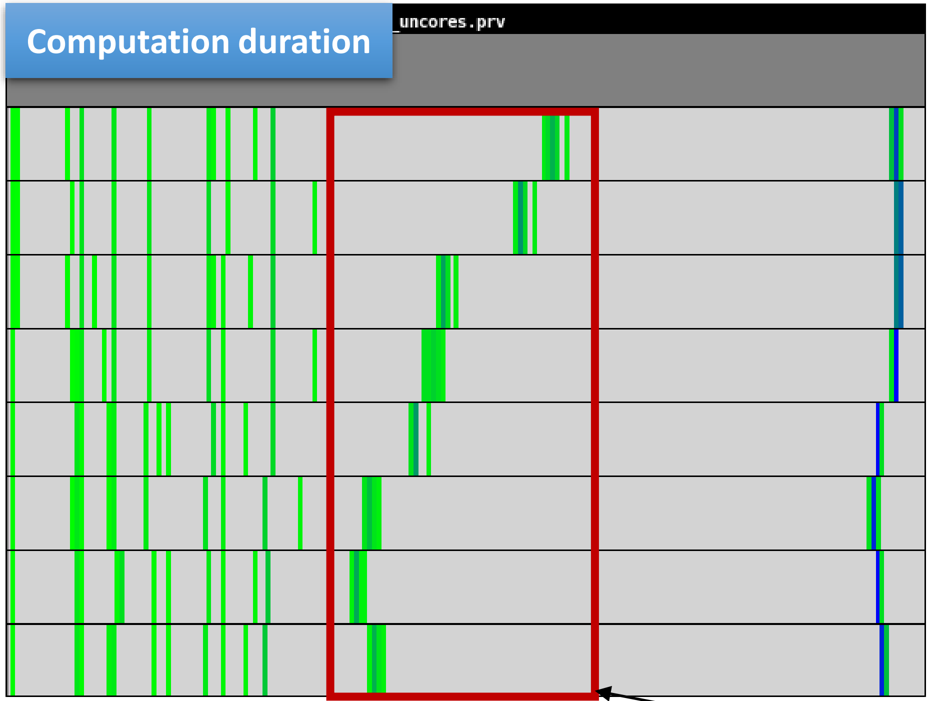


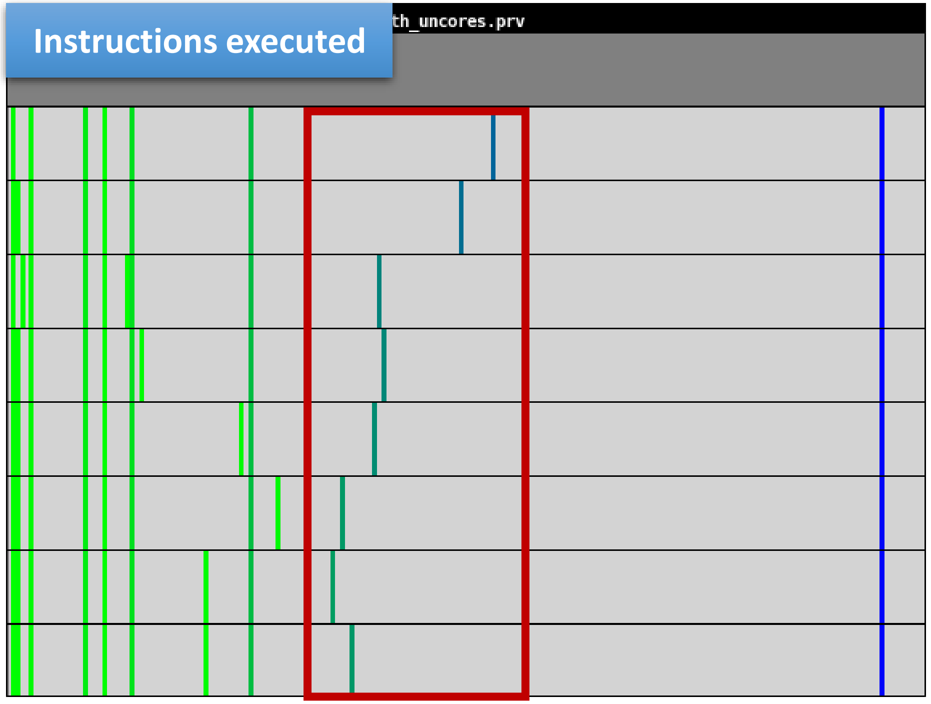
	Outside MPI	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Waitall	MPI_Barrier	MPI_Reduce	MPI_Allreduce
THREAD 1.1.1	98.28 %	0.02 %	0.01 %	0.09 %	1.46 %	0.00 %	0.10 %	0.02 %
THREAD 1.2.1	97.51 %	0.02 %	0.01 %	0.03 %	1.59 %	0.00 %	0.00 %	0.82 %
THREAD 1.3.1	94.99 %	0.02 %	0.01 %	0.05 %	1.47 %	0.00 %	0.00 %	3.43 %
THREAD 1.4.1	94.44 %	0.03 %	0.01 %	0.02 %	1.32 %	0.00 %	0.00 %	4.16 %
THREAD 1.5.1	93.25 %	0.02 %	0.01 %	0.09 %	2.13 %	0.01 %	0.02 %	4.44 %
THREAD 1.6.1	91.92 %	0.03 %	0.01 %	0.03 %	1.92 %	0.00 %	0.00 %	6.08 %
THREAD 1.7.1	91.38 %	0.03 %	0.01 %	0.03 %	1.71 %	0.00 %	0.27 %	6.56 %
THREAD 1.8.1	91.67 %	0.03 %	0.01 %	0.08 %	2.28 %	0.00 %	0.00 %	5.91 %
Total	753.44 %	0.21 %	0.09 %	0.41 %	13.88 %	0.03 %	0.39 %	31.42 %
Average	94.18 %	0.03 %	0.01 %	0.05 %	1.73 %	0.00 %	0.05 %	3.93 %
Maximum	98.28 %	0.03 %	0.01 %	0.09 %	2.28 %	0.01 %	0.27 %	6.56 %
Minimum	91.38 %	0.02 %	0.01 %	0.02 %	1.32 %	0.00 %	0.00 %	0.02 %
StDev	2.47 %	0.00 %	0.00 %	0.03 %	0.32 %	0.00 %	0.09 %	2.26 %
Avg/Max	0.96	0.91	0.82	0.57	0.76	0.44	0.18	0.60



Understanding applications with Paraver (II)

- Correlating multiple views





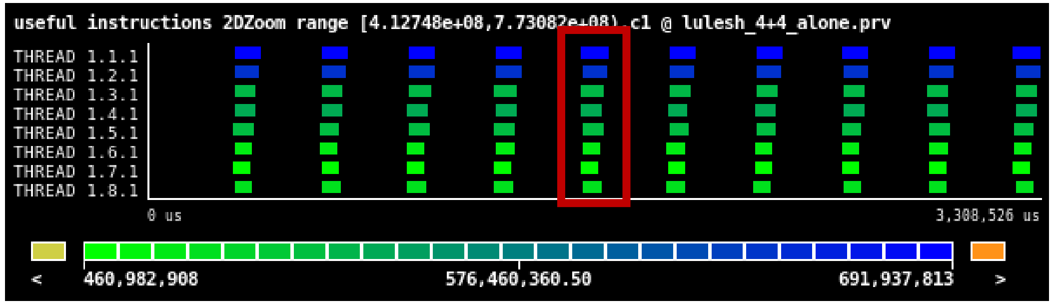
Duration imbalance

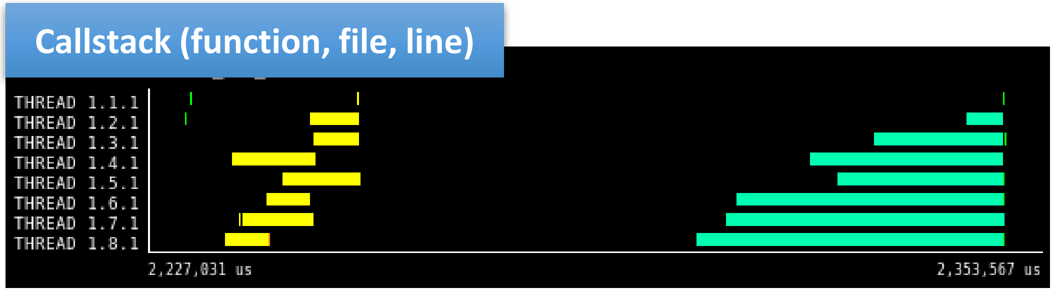
So…

Work imbalance

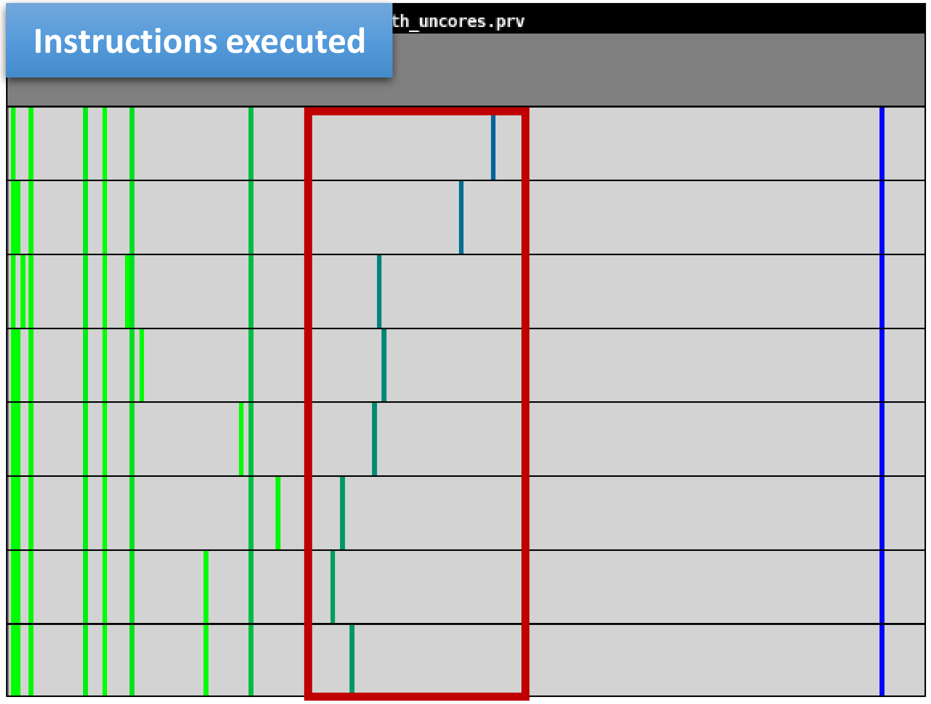
Understanding applications with Paraver (III)

- Going back to the source code



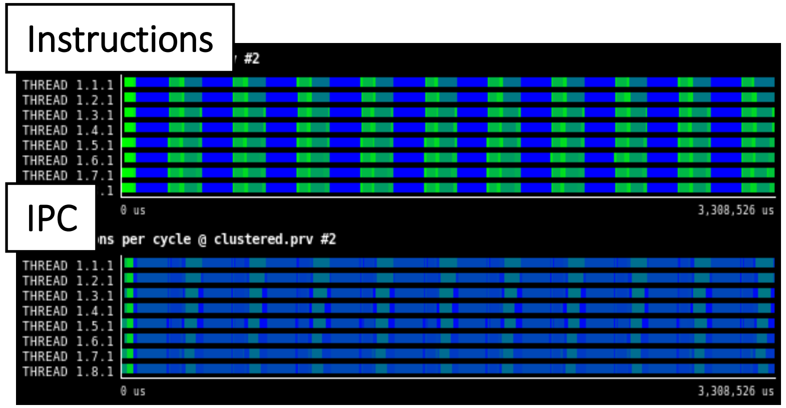


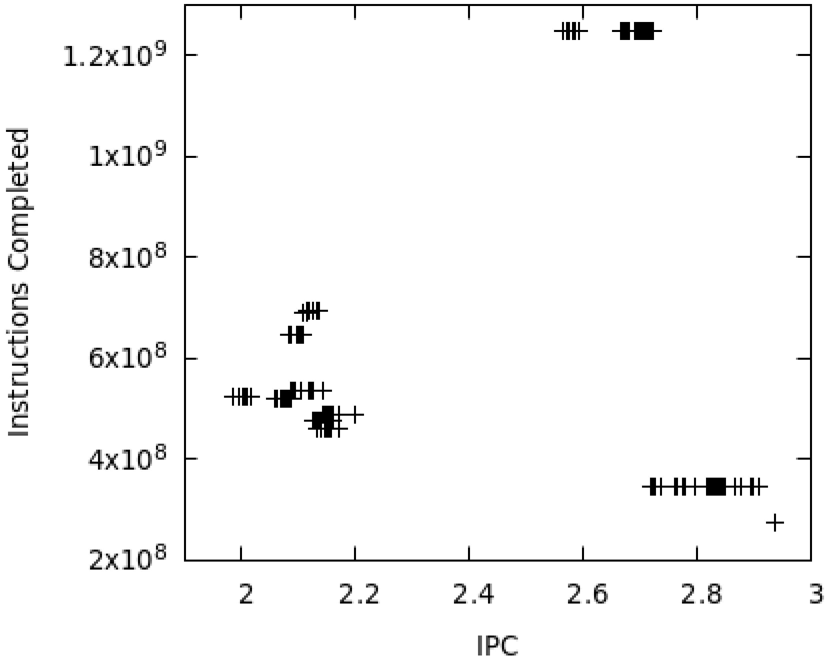
CommSend TimeIncrement



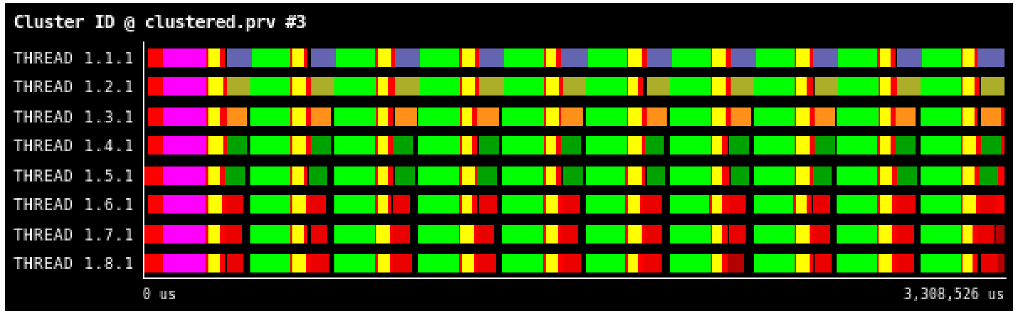
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

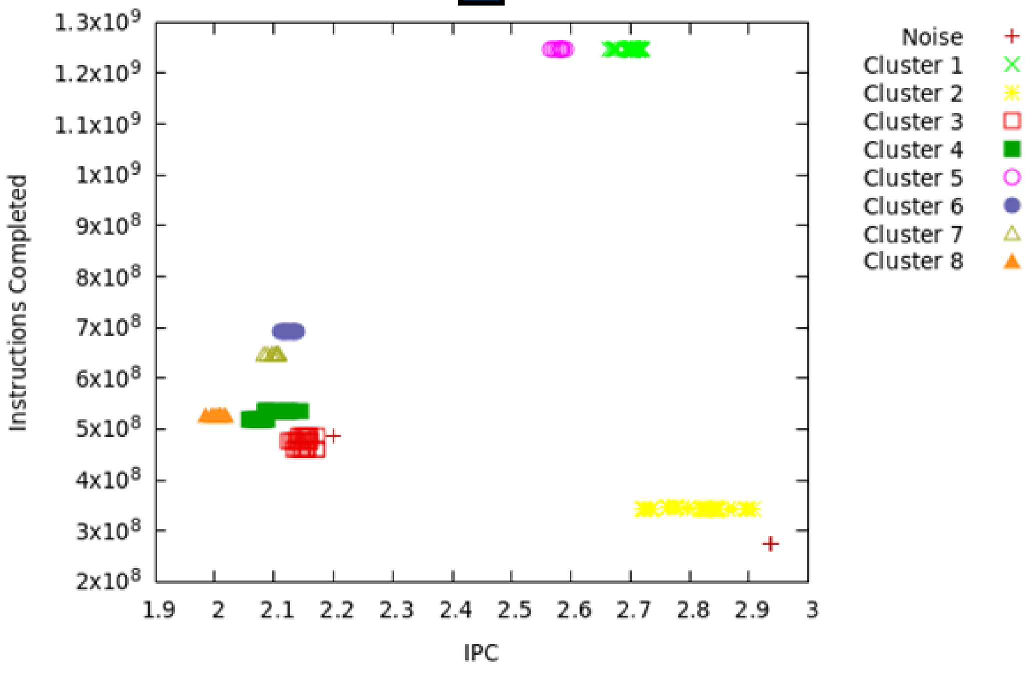
Clustering to identify structure





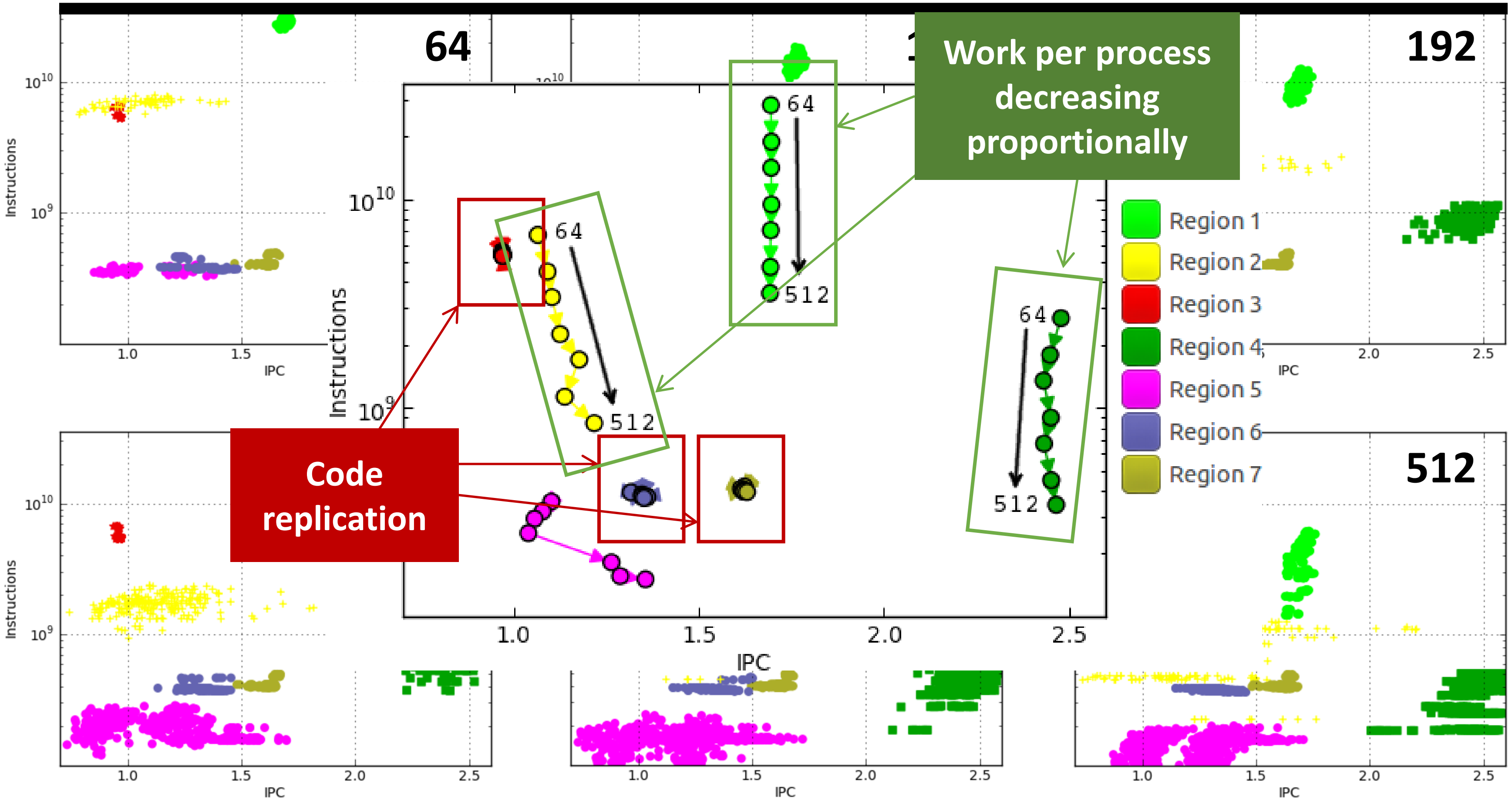
DBSCAN





Tracking to compare experiments

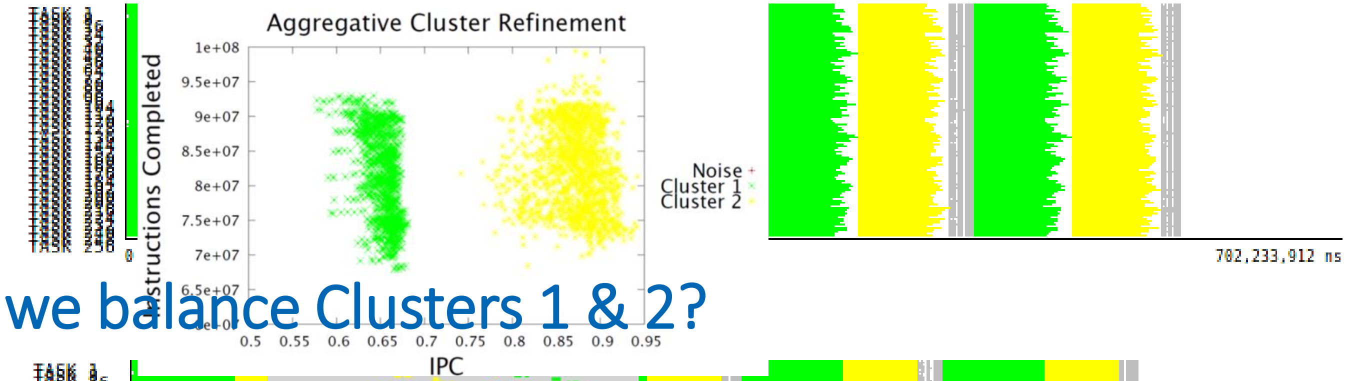
- Use case: Study the scalability of the computing regions → From 64 to 512 cores



Dimemas to predict scenarios

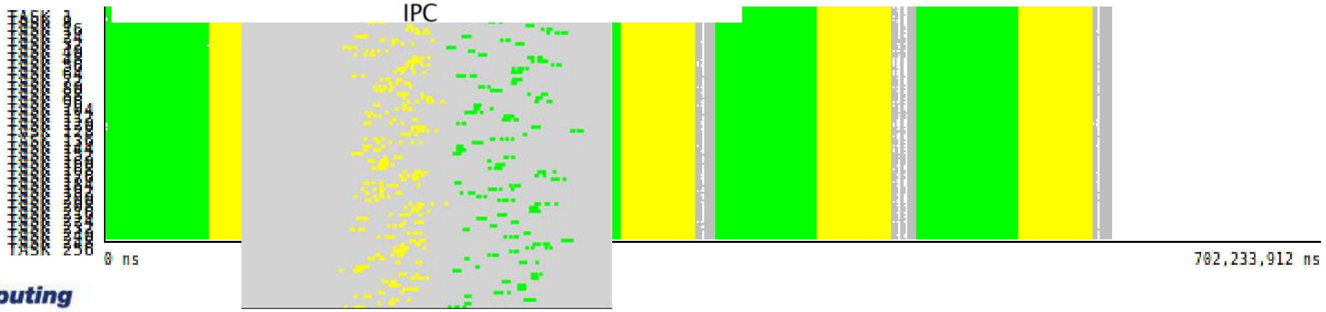
What if ….

… we increase the IPC of Cluster1?



13% gain

… we balance Clusters 1 & 2?

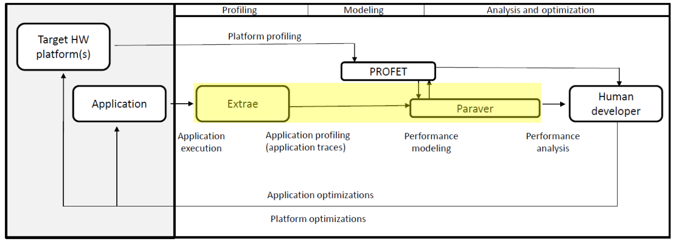


19% gain

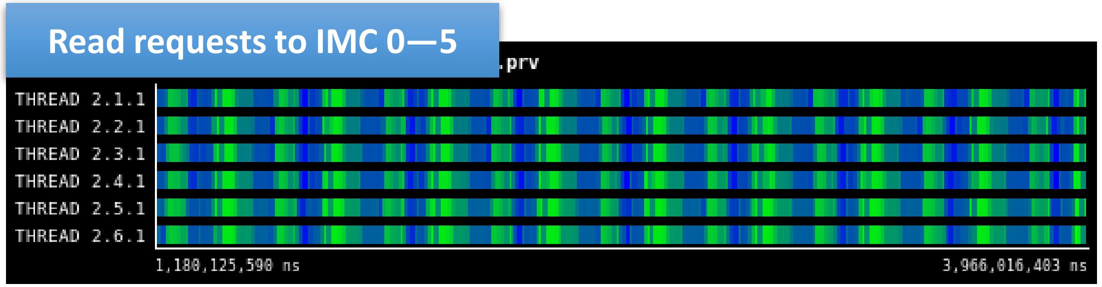
Quick estimates on when is an optimization worth the effort

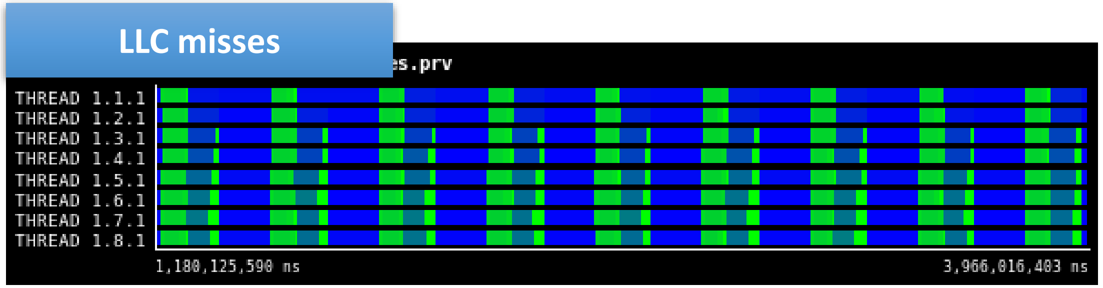
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

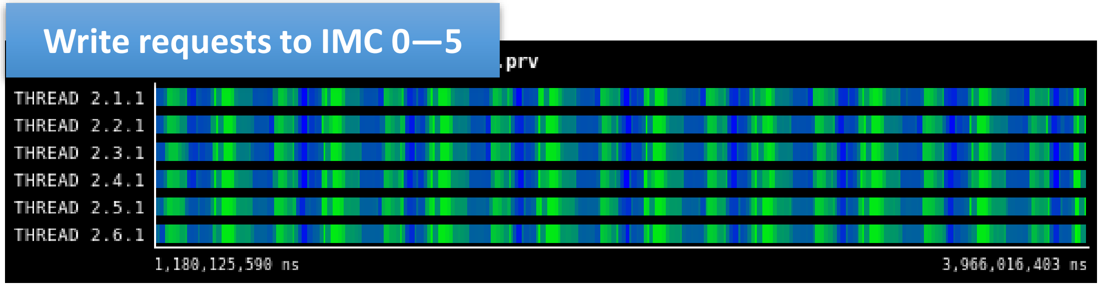
Tools + PROFET: Current progress



- Extrae writes Paraver traces with required counters for PROFET





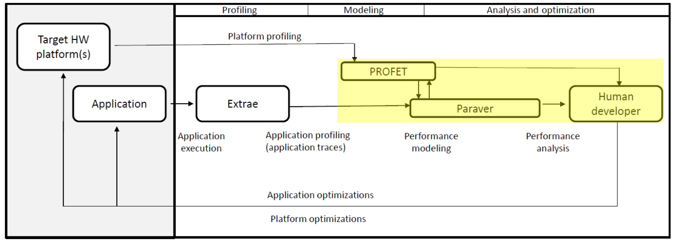


Native counters measured by application threads at instrumentation points

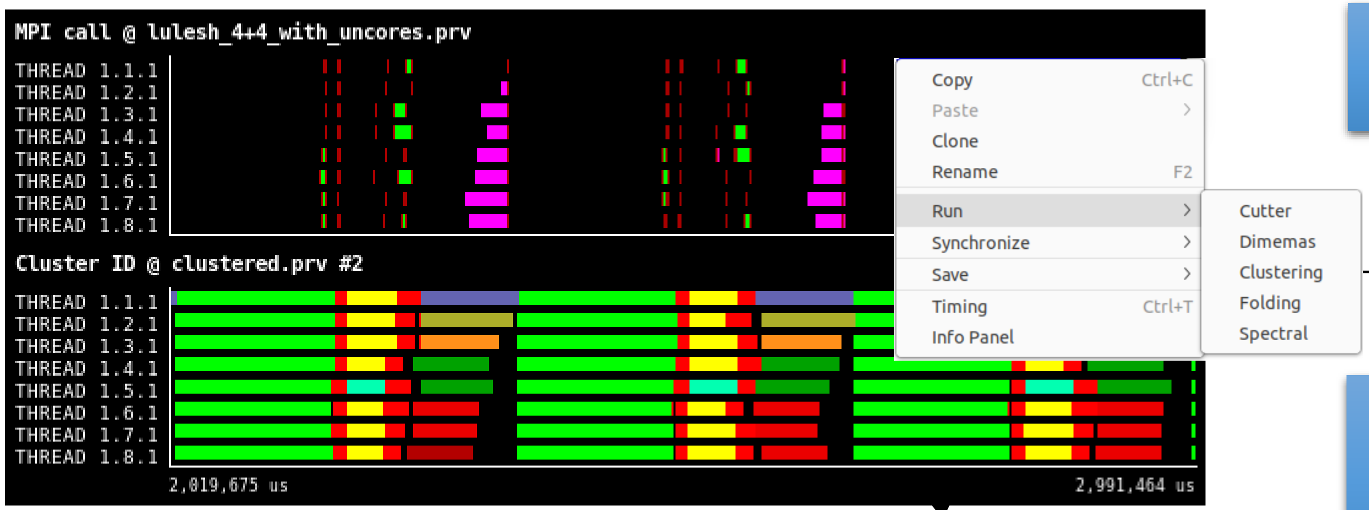
Uncore counters sampled periodically by dedicated threads per IMC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

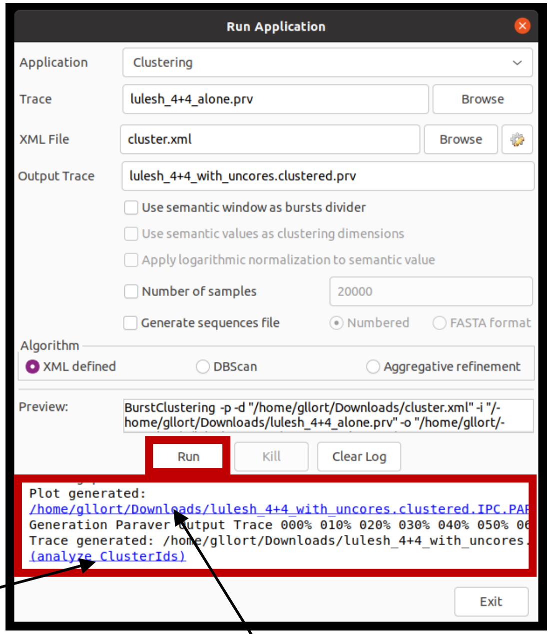
Integrated environment



- Characterise application's memory usage on selected regions directly from Paraver context menu



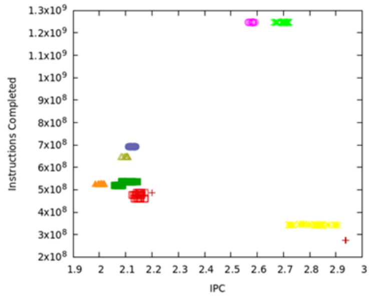
1. Invoke tool with presets



2. Hit 'Run' → Tool output is reported with links

- Same approach for integrating PROFET
    - In-trace feedback
        - Memory usage (latency, bandwidth)
    - Additional model outputs

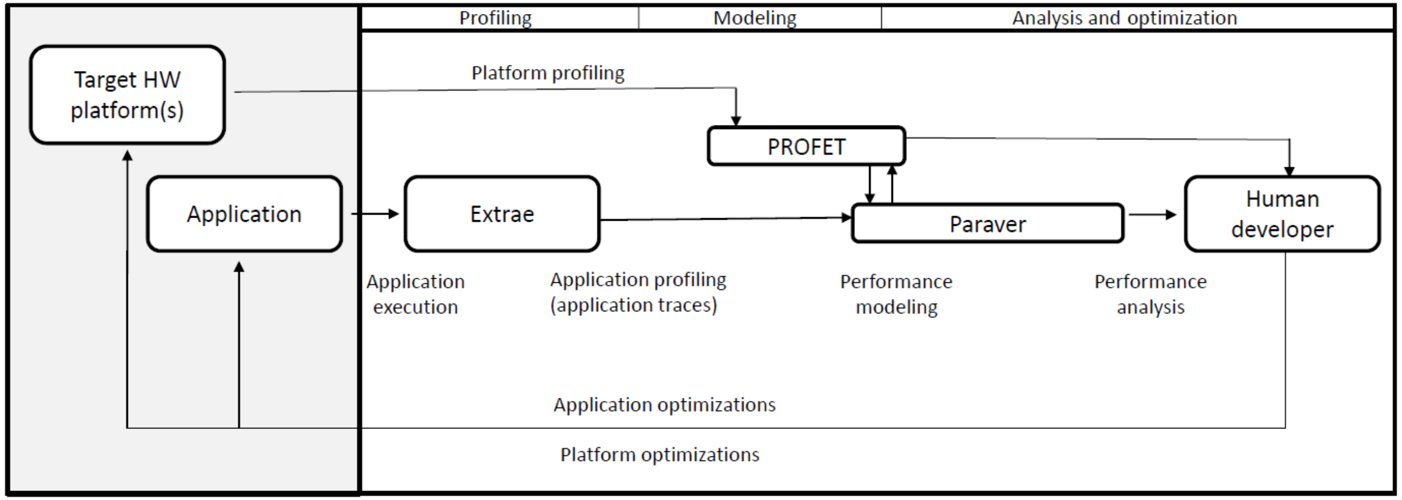
3. Clicking on a link automatically brings up new Paraver windows



4. Or additional reports (e.g. GNUplot)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

App developers: Try our cycle!



- It's easy to use!
- Just add LD_PRELOAD to your jobscript to get a trace
- Directly invoke PROFET from Paraver context menu → Run
- PROFET outputs will jump back to Paraver
- We are happy to assist you!
  - Just need the jobscript and the binary to give you useful insight!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

https://tools.bsc.es