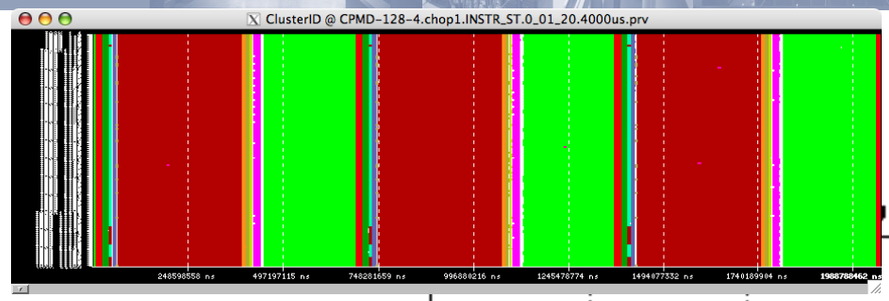
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Clustering

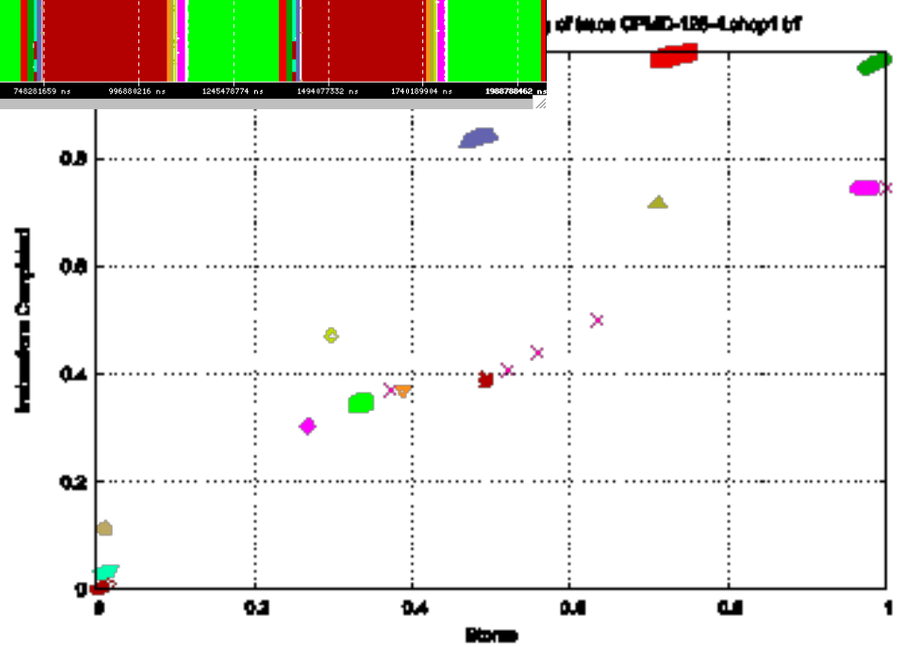
BSC Performance Tools

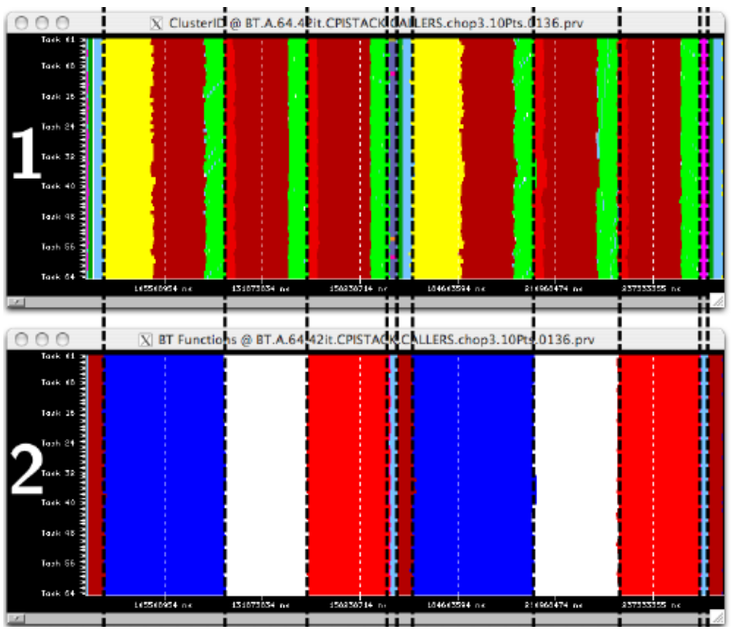
Clustering

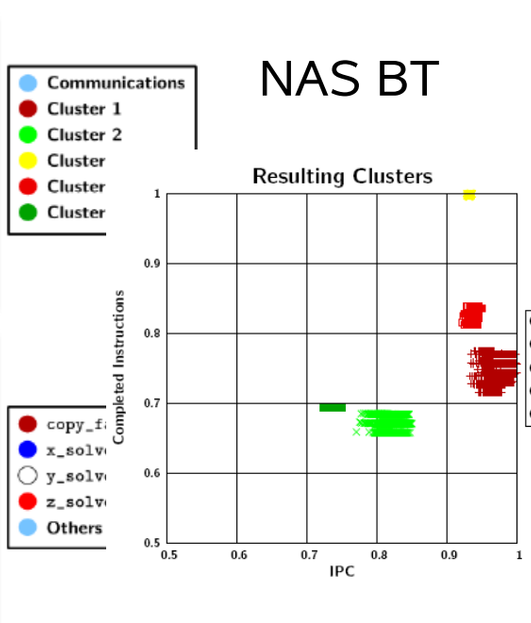
- Identify computation regions of similar behavior
  - Data structure not Gaussian → DBSCAN
  - Similar in terms of duration or hardware counter rediced metrics
    - Different routines may have similar behavior
    - One routine may have different behaviors



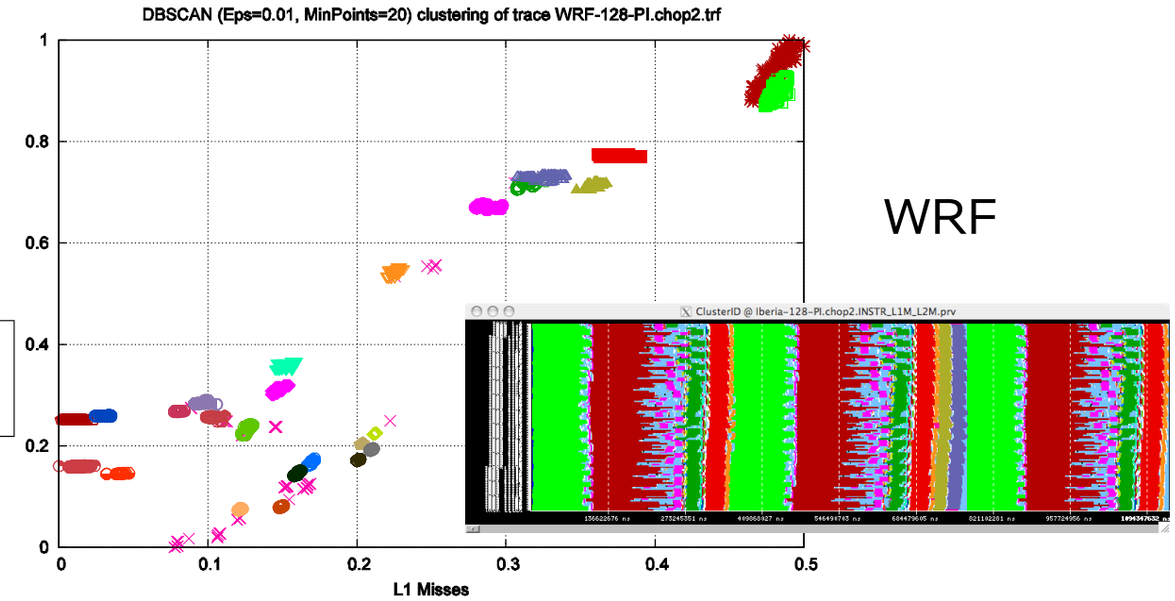
CPMD







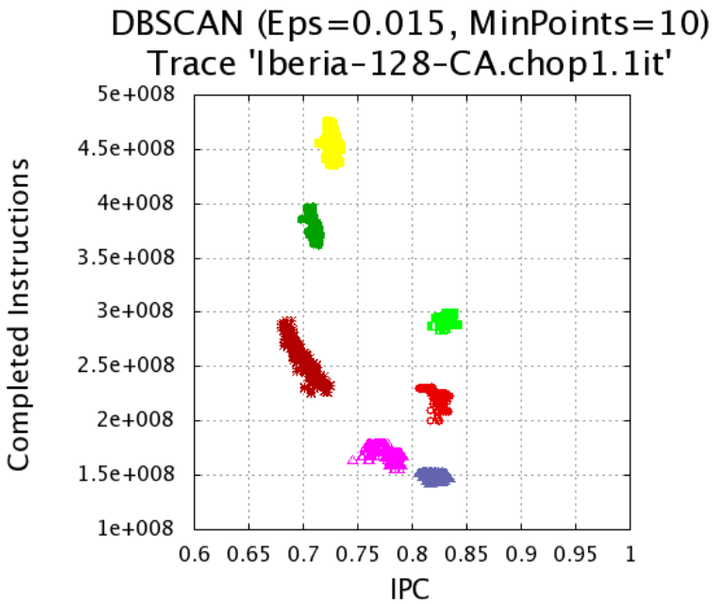
NAS BT



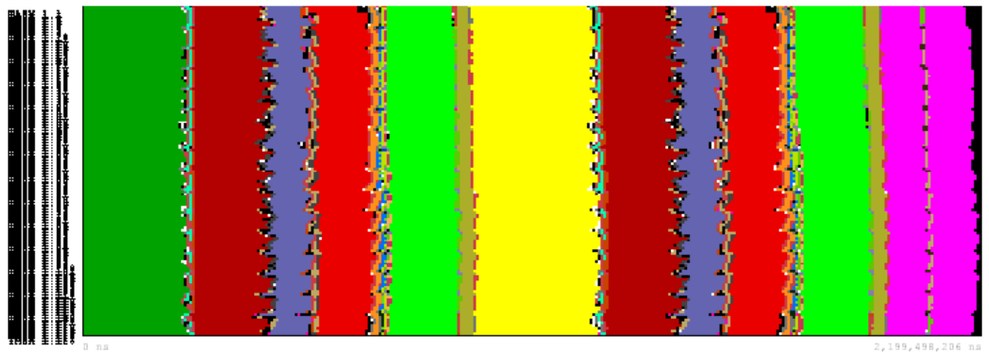
WRF

Clustering results

Scatter Plot of Clustering Metrics

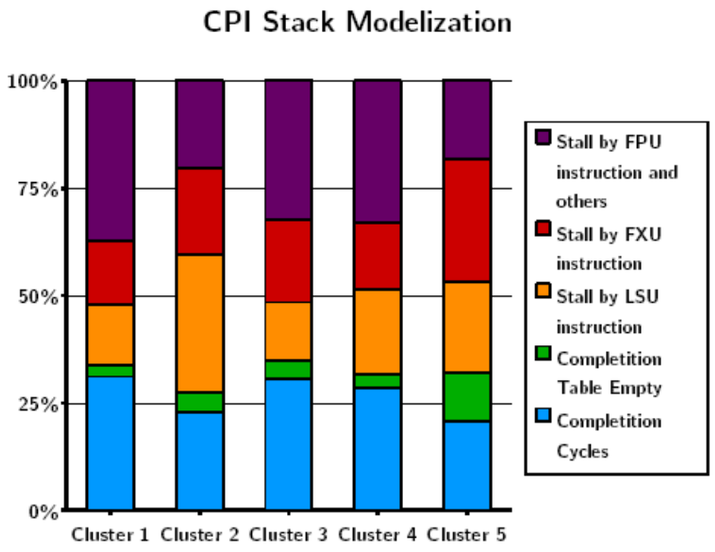


Clusters Distribution Over Time



Clusters Performance

CLUSTER	1	2
% TIME	36.29	29.52
AVG. BURST DUR. (MS)	220.46	177.70
IPC	0.53	0.50
MIPS	1210.07	1164.36
L1M/KINSTR	22.72	32.63
L2M/KINSTR	0.59	1.23
MEM.BW (MB/S)	90.77	182.65



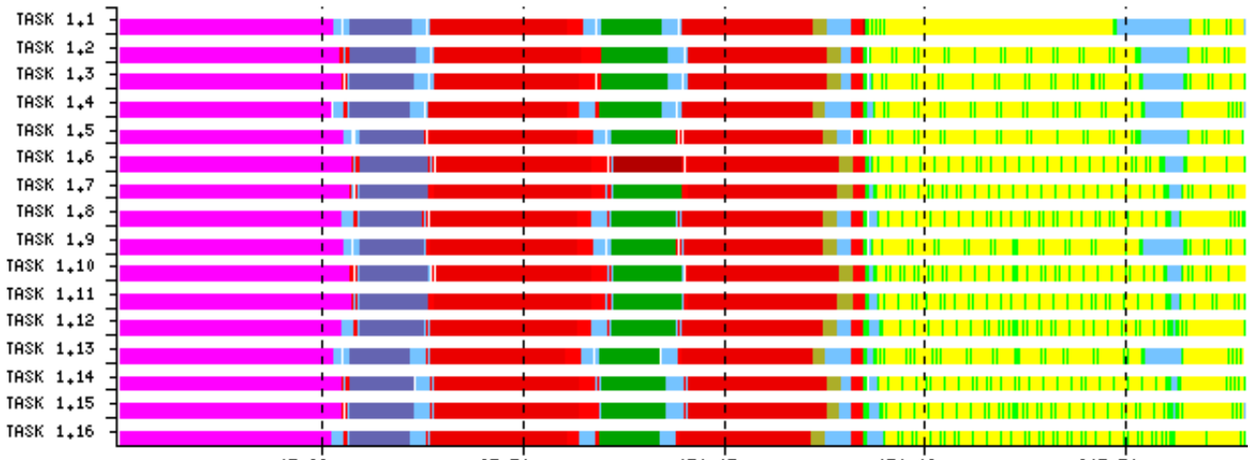
Code Linking

CLUSTER	CODE SECTION
1	solve_nmm.f:[2037 - 2310]
2	solve_nmm.f:[1478 - 1782] solve_nmm.f:[2030 - 1782]
3	solve_nmm.f:[1241 - 1345]
4	solve_nmm.f:[2771 - 2865] solve_nmm.f:[2388 - 2489]
5	solve_nmm.f:[1478 - 1569]
6	solve_nmm.f:[1607 - 1633]

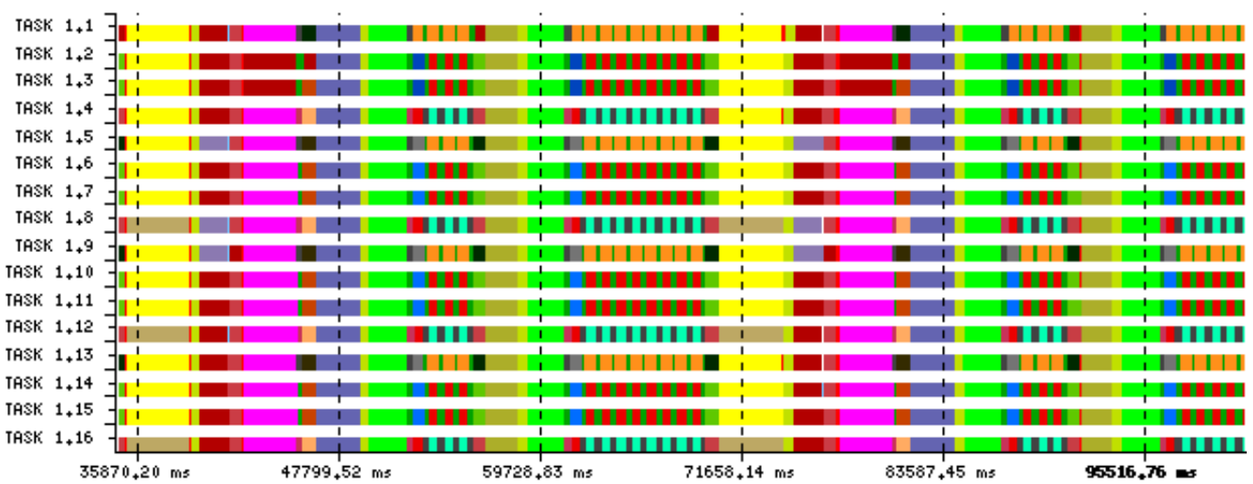
Quality of clustering result

- Expected SPMD behavior
- Load balance/ heterogeneous algorithm issues

LU.A

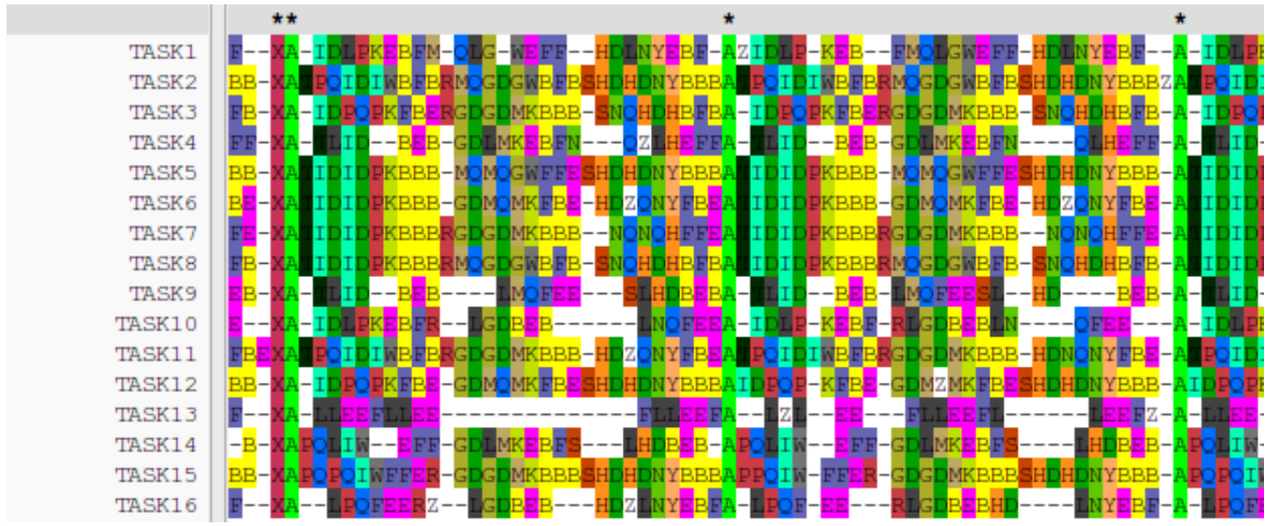


WRF

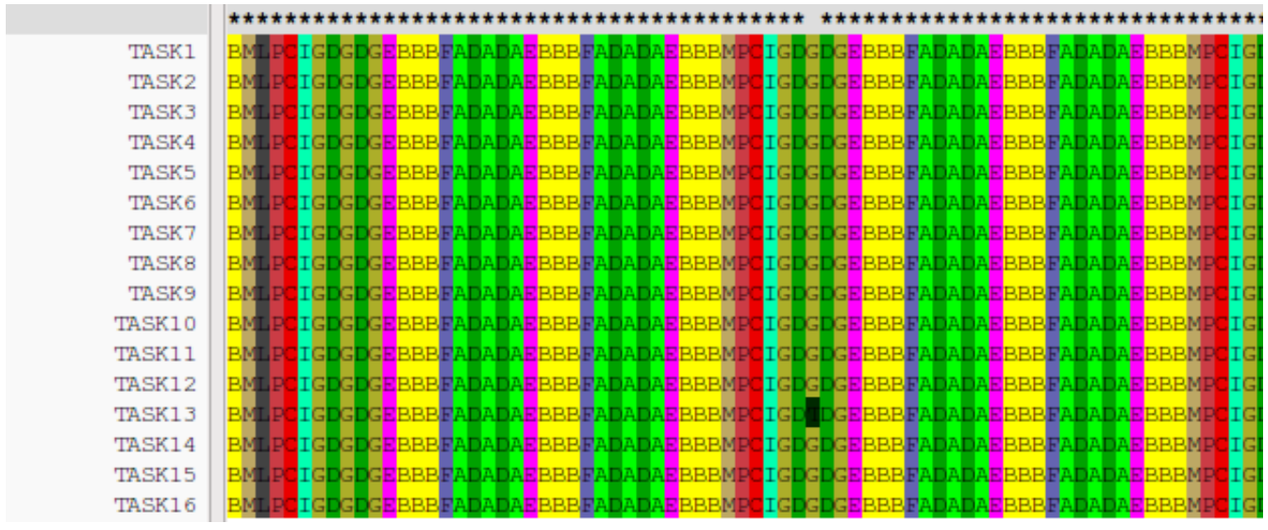


Automatic clustering quality assessment

- Leverage Multiple Sequence alignment tools from Life Sciences
- Process == Sequence of clusters ↔ sequence of amino acids == DMA
- CLUSTAL W, T-Coffee, Kalign2
- Cluster Sequence Score (0..1)
- Per cluster / Global
  - Weighted average



BT.A 0022



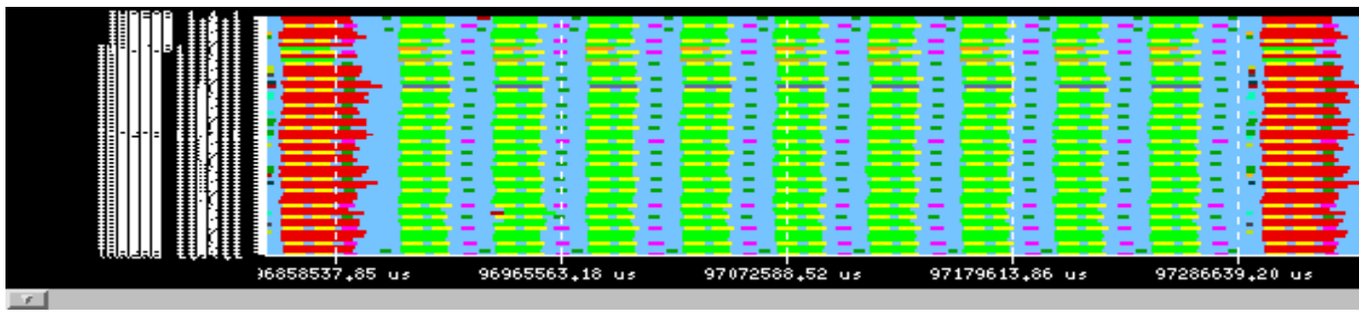
BT.A 0043

Clustering vs. classification

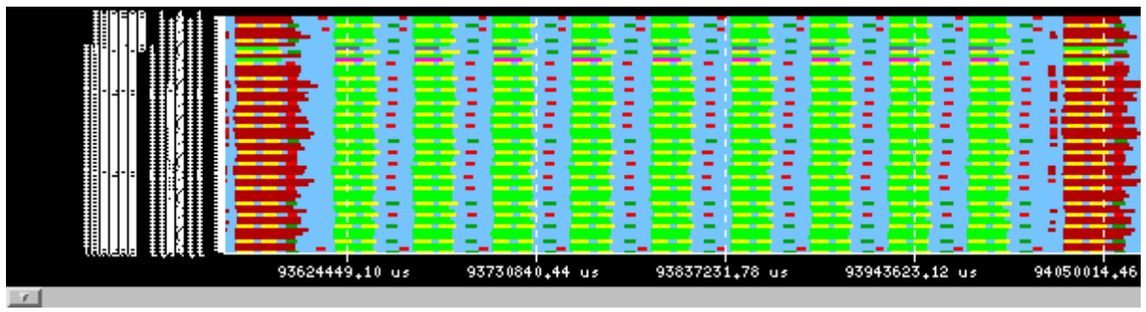
- Clustering time drastically grows with number of points
- Selection of a subset of data to clusterize
  - Space: Select a few processes. Full time sequence
  - Random sampling: wide covering
- Remaining data: “nearest” neighbor classification

GROMACS 64 processes

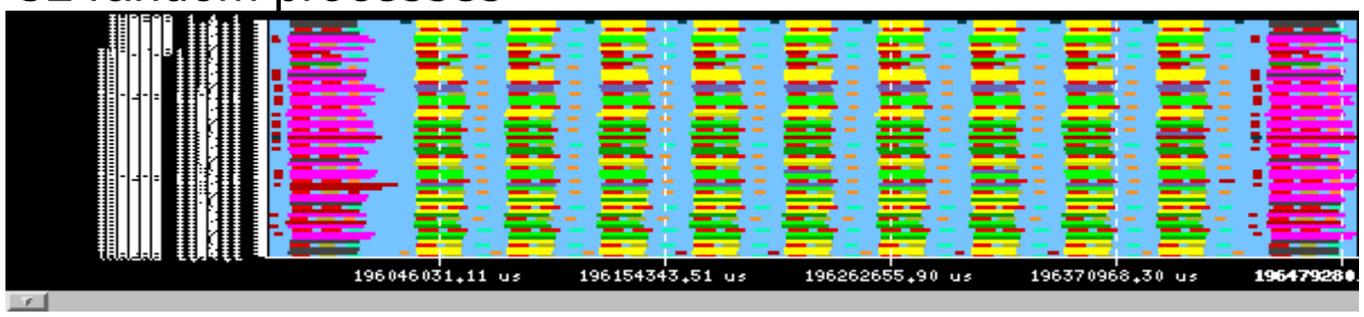
64 processes



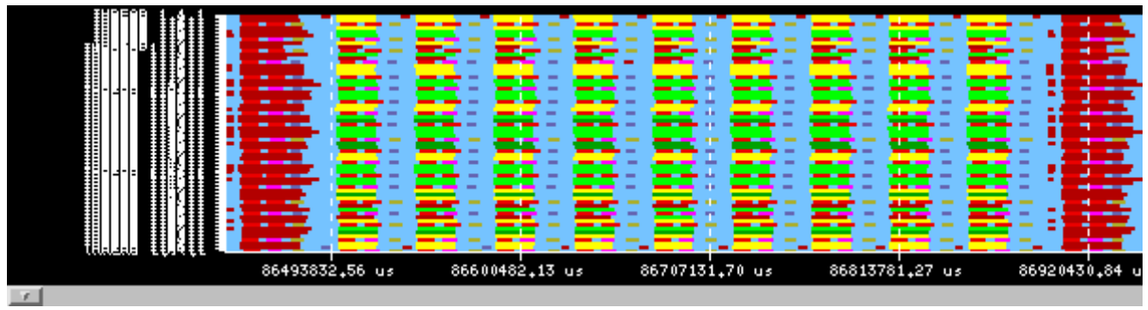
25% random records



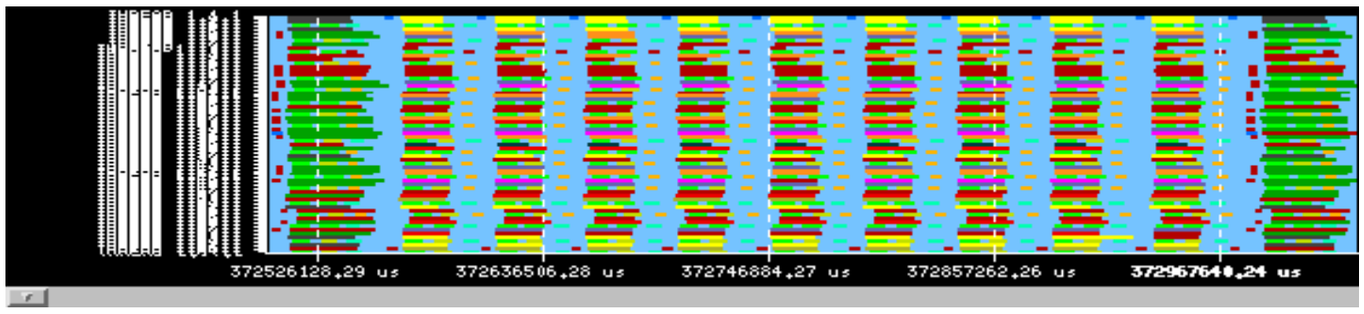
32 random processes



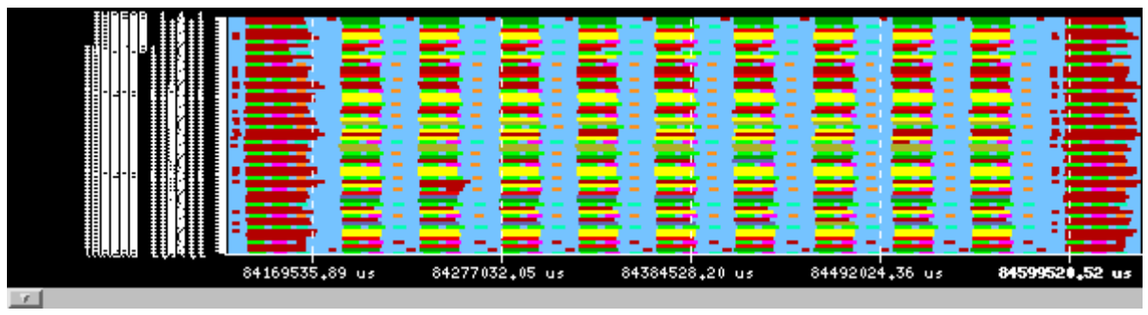
15% random records



16 random processes

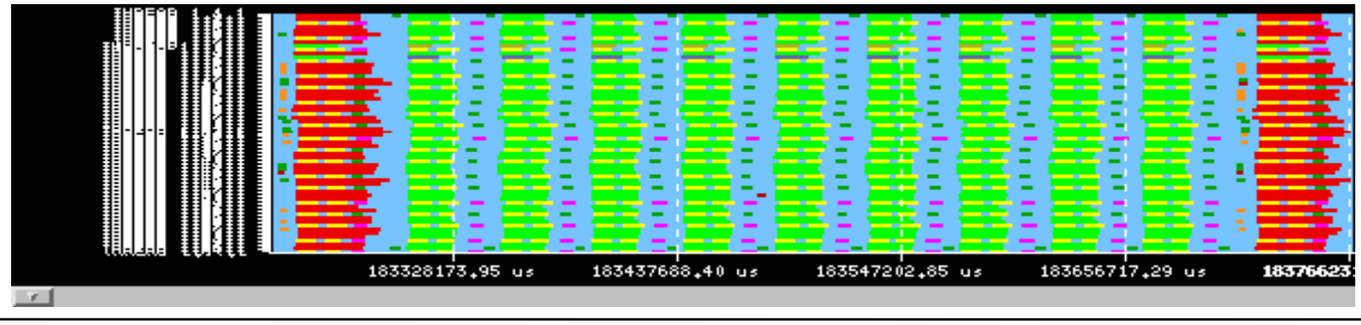


10% random records

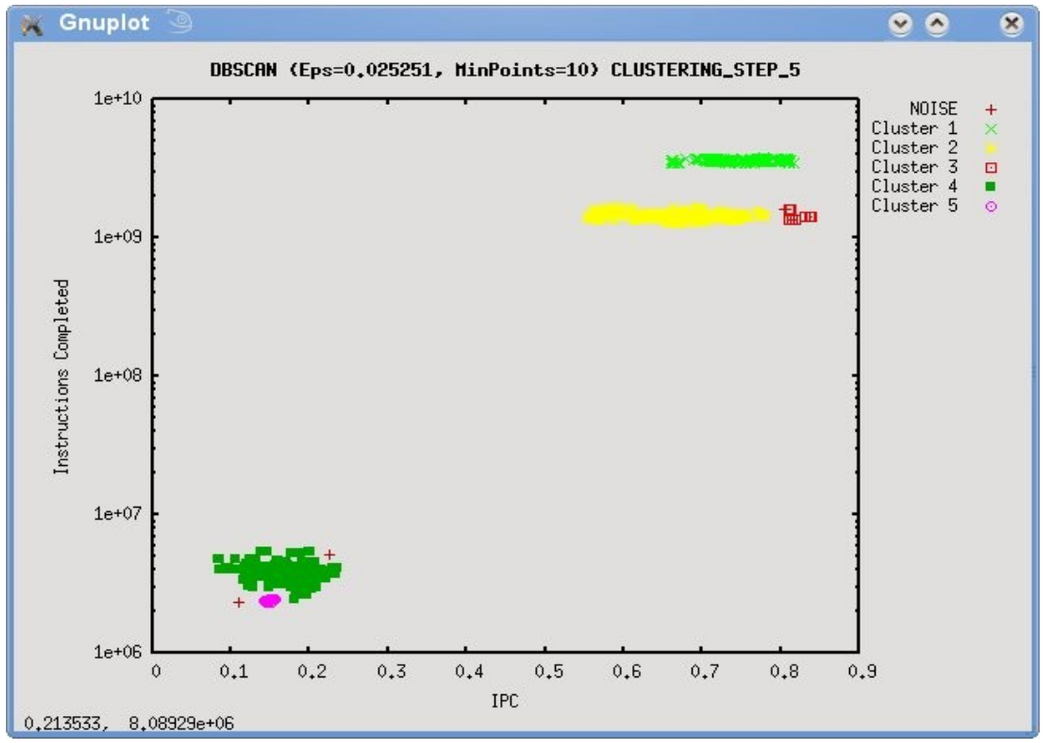


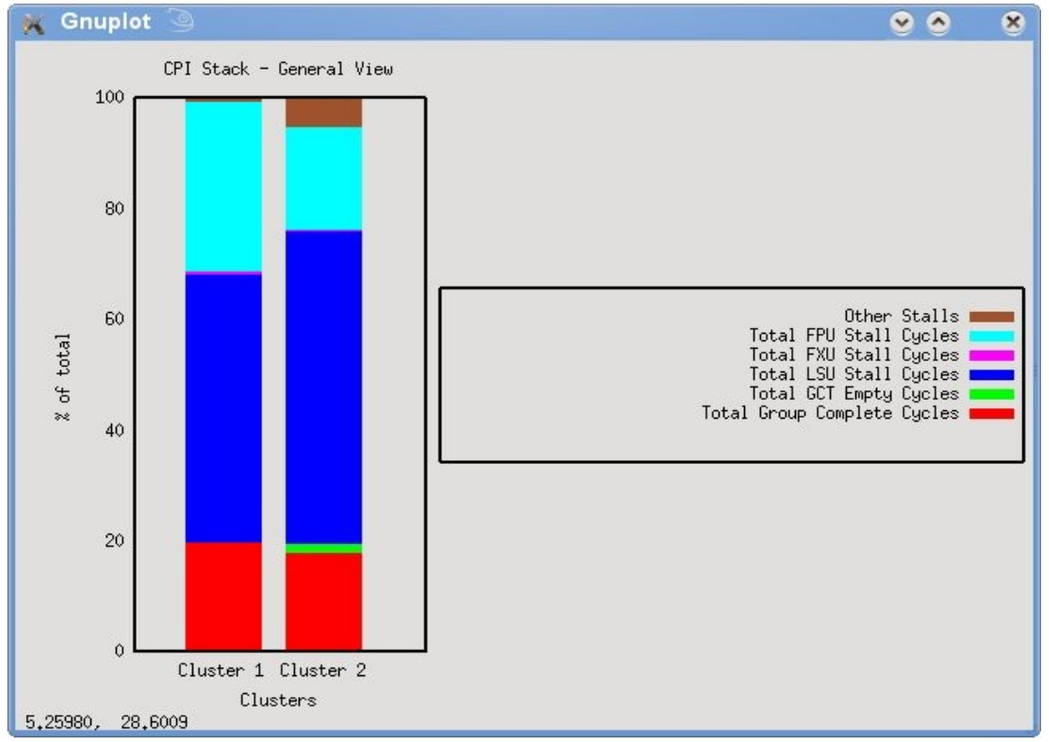
Good quality
Fast analysis

8 random processes + 15% random records

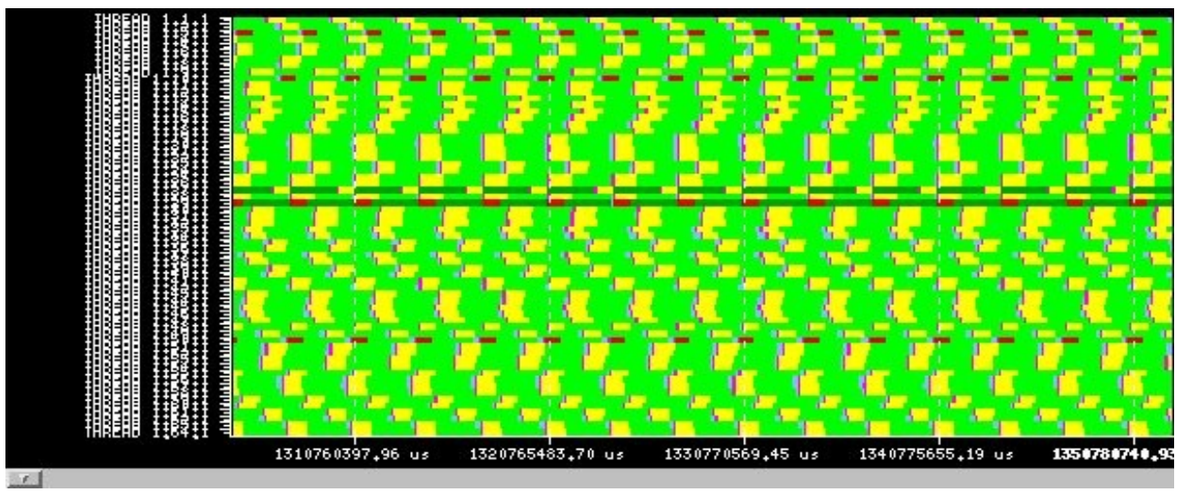


SPECFEM3D (64 tasks)

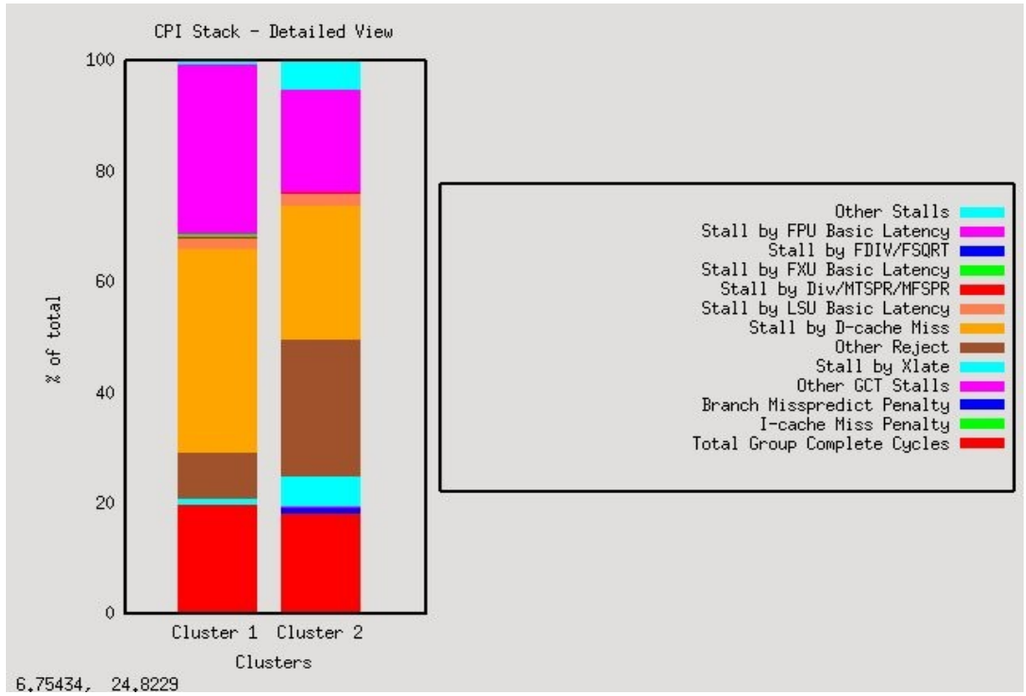




Clusters distribution



CPI STACK model



SPECFEM3D (64 tasks)

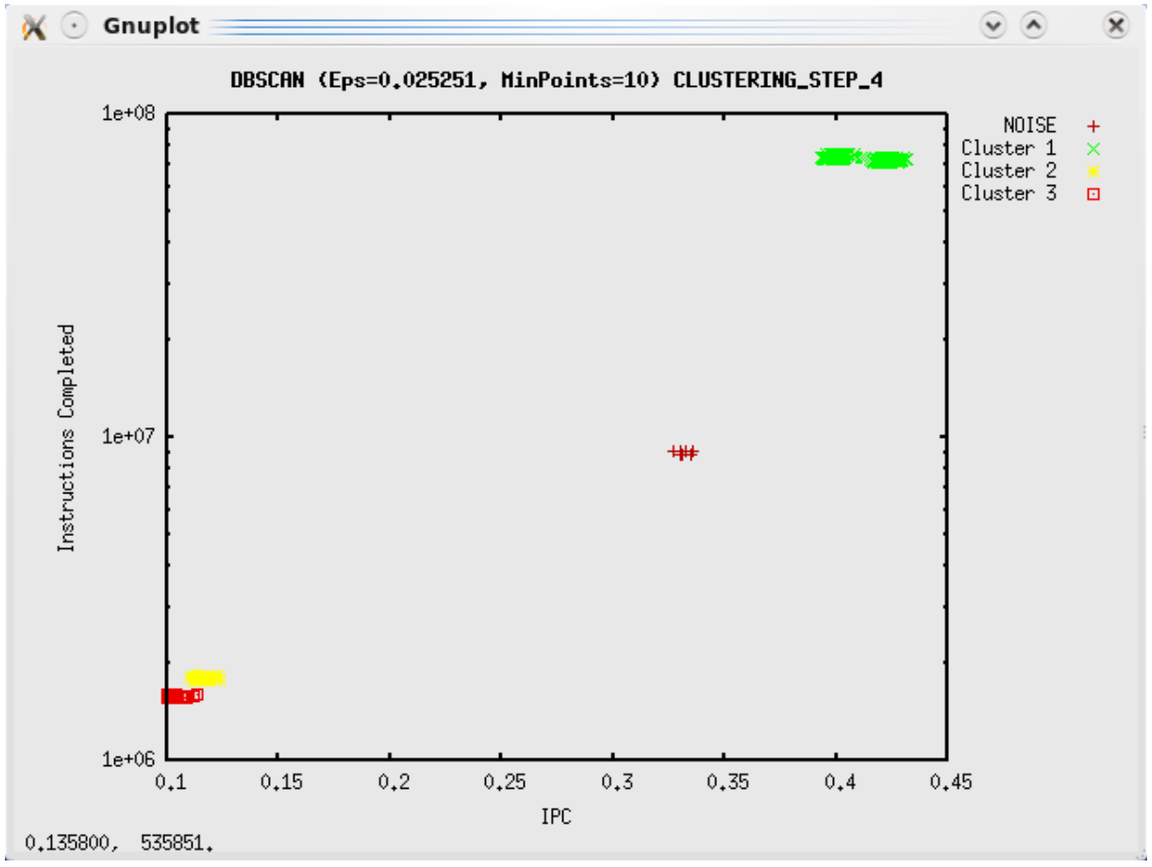
- Other statistics

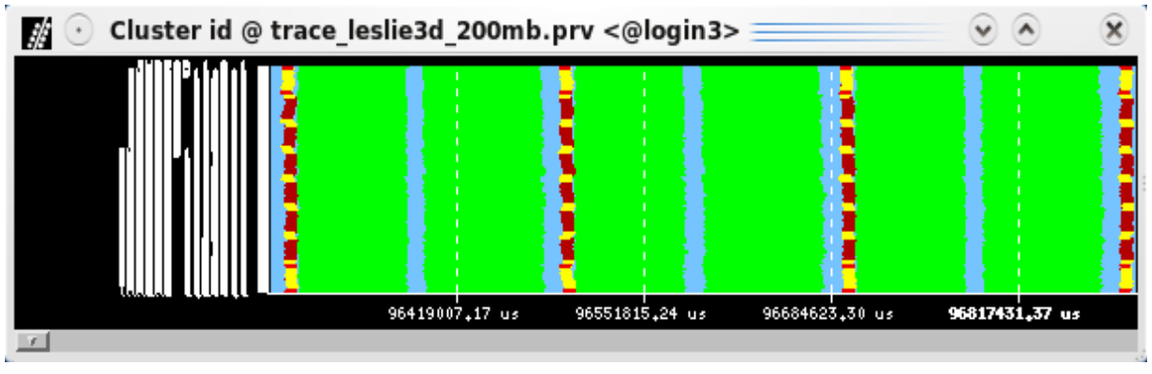
Category	Metric Description	Cluster 1	Cluster 2
Performance	**% Duration**	0.70236	0.26024
Performance	**Avg. Burst Duration (µs)**	2.142.664,27	947.044,53
Performance	**Total preemted time (µs)**	31394,19	13759,21
Performance	**% preempted time**	1,465%	1,453%
Performance	**IPC**	0,75	0,66
Performance	**CPI**	1,33	1,51
Performance	**MIPs**	1702,70	1501,75
Performance	**Mem.BW (MB/s)**	260,22	203,84
Performance	**Memory instructions per second**	1324,80	1740,67
Performance	**HW floating point instructions per cycle**	0,290	0,168
Performance	**Flop rate (MFLOPs)**	1.421,36	555,95
Performance	**HW floating point instructions rate**	656,95	381,15
Performance	**Computation intensity**	2,004	1,128
Performance	**Local L2 load bandwidth per processor (MB/s)**	5.634,58	2.409,52
Performance	**% Loads from local L2 per cycle**	2,039%	0,872%

Category	Metric Description	Cluster 1	Cluster 2
Architecture	**% Instr. Completed**	32,25%	32,22%
Architecture	**L1 misses per Kinstr.**	41,37	34,82
Architecture	**L2 misses per Kinstr.**	1,194	1,060
Architecture	**Bytes from maim memory per floating point instruction finished**	1,649	1,361
Architecture	**Number of Loads per Load miss**	24,81	66,68
Architecture	**Number of Stores per Store miss**	6,74	5,27
Architecture	**Number of Loads&Stores per L1 miss**	18,81	33,28
Architecture	**L1 cache hit rate**	94,68%	97,00%
Architecture	**Number of Loads per (D)TLB miss**	11.403,59	5.873,04
Architecture	**Number of Loads&Stores per (D)TLB miss**	12.945,98	6.425,90
Architecture	**% TLB misses per cycle**	0,005%	0,012%
Architecture	**Total Loads from local L2 (M) (total_ld_l_L2)**	94,320	17,828
Architecture	**Local L2 load traffic (MB)**	12.073,007	2.281,926

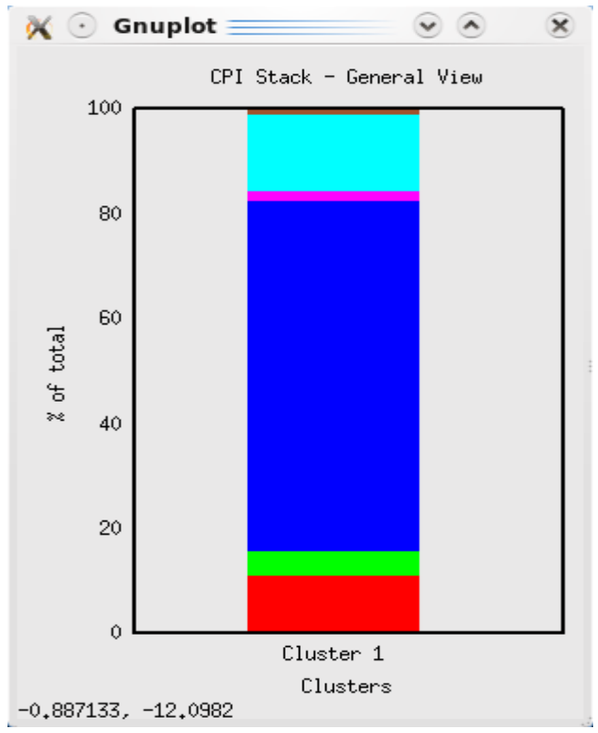
Category	Metric Description	Cluster 1	Cluster 2
Instruction Mix	**FMA ops per floating point instruction**	0,763	0,633
Instruction Mix	**Instructions per Load/Store**	1,285	0,863
Instruction Mix	**HW floating point instructions (flips)**	1.407.628.943	360.968.332
Instruction Mix	**Total floating point operations (flops)**	3.045.492.593	526.507.674
Instruction Mix	**Total FP Load&Store operations (fp_tot_ls)**	1.519.651.784	466.816.330
Instruction Mix	**FMA %**	81,04%	82,15%
Instruction Mix	**Memory Mix**	25,10%	37,35%
Instruction Mix	**Load Mix**	22,11%	34,14%
Instruction Mix	**Store Mix**	2,99%	3,21%
Instruction Mix	**FPU Mix**	12,44%	8,18%
Instruction Mix	**FXU Mix**	3,93%	12,64%

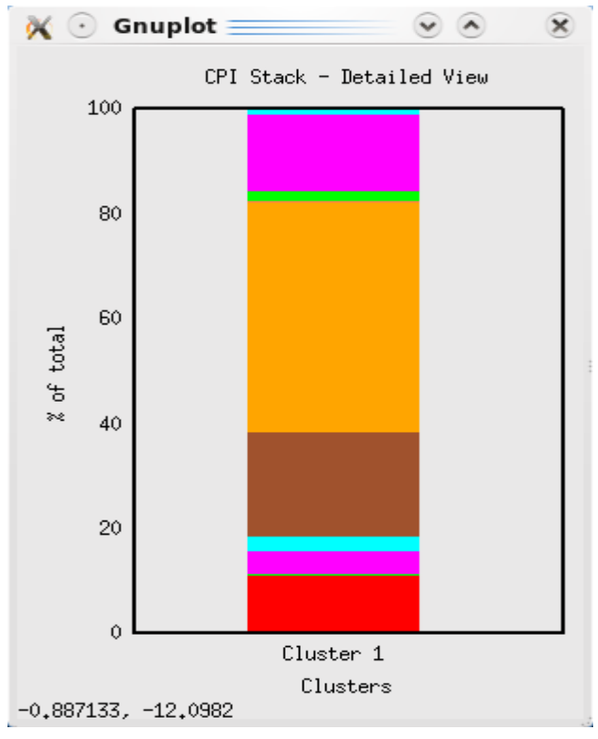
SPECMPI'07 – LESLIE3D (256 tasks)



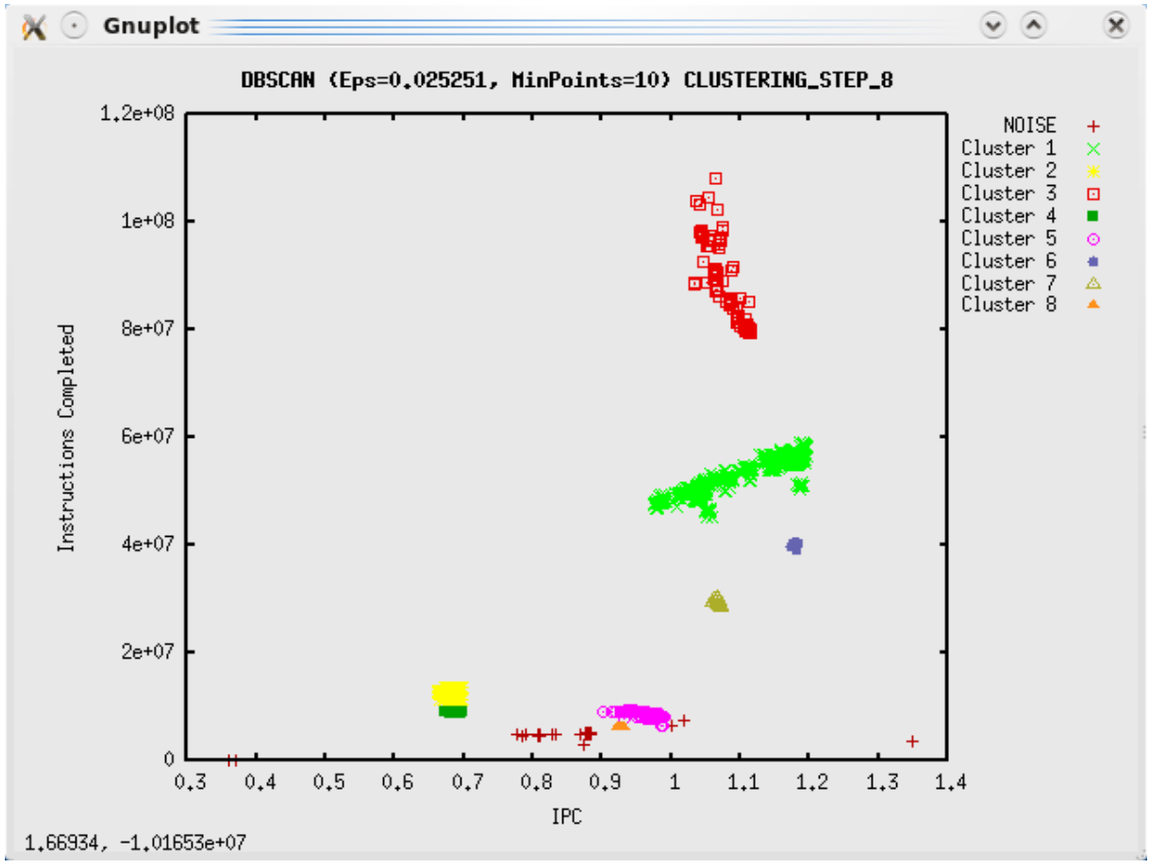


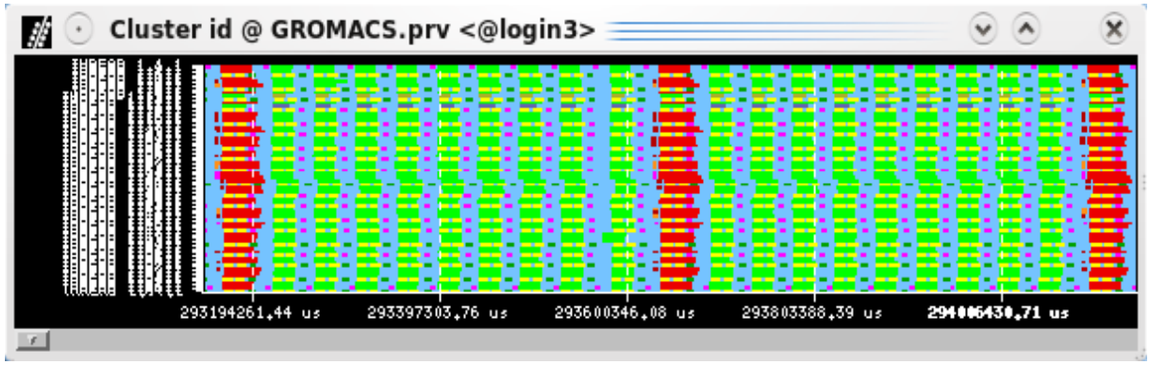
256 tasks, 300 Mb



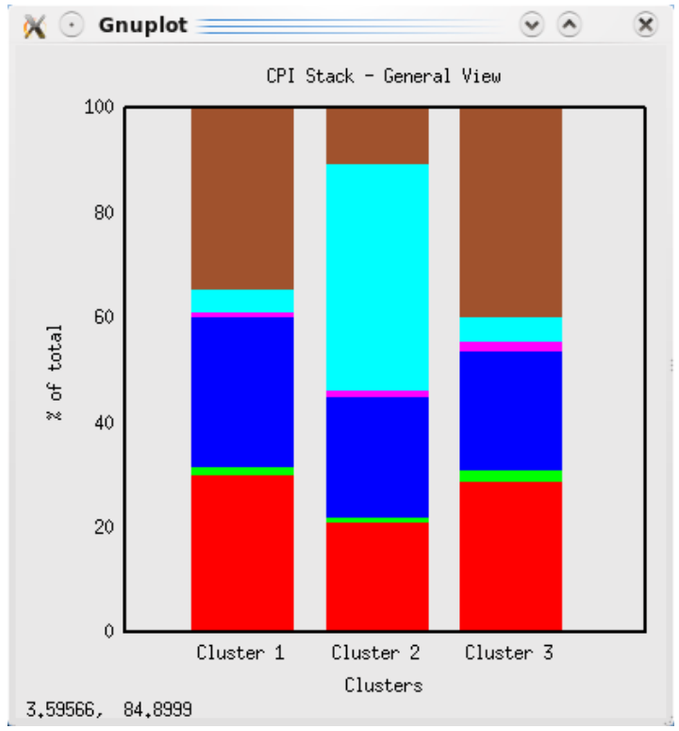


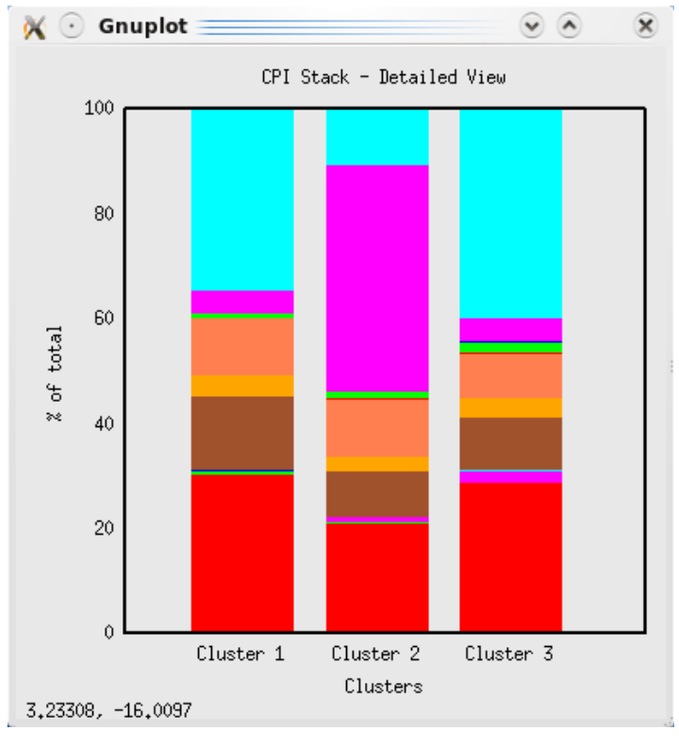
GROMACS (64 tasks)





64 tasks, 150 Mb





Clustering steps

1. Obtain a tracefile
   - If using multiple counter sets, guarantee some common counters
   - If targeting CPIStack – use predefined sets (as example)
2. Define the clustering XML – clustering axis, scatter plot axis, metrics to extrapolate
3. Run the bust-clusterizer tool
4. Analyze the results – scatter plot, clustered paraver tracefile (, cpistack – table of performance metrics)

- Files available @ MN in /gpfs/apps/CEPBATOOLS/PRACE_tutorial/examples/clustering

Clustering steps – Obtaining a tracefile

- extrae_cpistack.xml – Select all the required counter sets to compute CPIStack model

Addapt to your application

```
<counters enabled="yes">
    <cpu enabled="yes" starting-set-distribution="1">
     <set enabled="yes" domain="all" changeat-time="1s">
PM_GCT_EMPTY_CYC,PM_LSU_LMQ_SRQ_EMPTY_CYC,PM_HV_CYC,PM_CYC,PM_1PLUS_PPC_CMP
L,PM_INST_CMPL,PM_GRP_CMPL,PM_TB_BIT_TRANS
      </set>
     <set enabled="yes" domain="all" changeat-time="1s">
PM_GCT_EMPTY_CYC,PM_INST_CMPL,PM_FLUSH_BR_MPRED,PM_BR_MPRED_TA,PM_GCT_EMPTY
_IC_MISS,PM_CYC,PM_GCT_EMPTY_BR_MPRED,PM_L1_WRITE_CYC
      </set>
     <set enabled="yes" domain="all" changeat-time="1s">
PM_LSU0_BUSY,PM_LSU1_BUSY,PM_LSU_FLUSH,PM_FLUSH_LSU_BR_MPRED,PM_CMPLU_STALL
_LSU,PM_INST_CMPL,PM_CMPLU_STALL_ERAT_MISS,PM_CYC
      </set>
     (....)
```

Clustering steps – Clustering specification

- cluster_I_IPC.xml

Addapt starting from a "big" number (0.1)

May need to addapt to reject non relevant computations

```
<?xml version="1.0"?>
<clustering_definition duration_filter="10000" use_duration="no"
apply_log="yes" normalize_data="yes" threshold_filter="0">
  <DBSCAN epsilon="0.020" min_points="10"/>
  <clustering_parameters>
  <single_event apply_log="yes" name="Native_Instr">
    <event_type>42001090</event_type>
    <factor>1.0</factor>
  </single_event>
  <mixed_events apply_log="yes" name="cl_IPC" operation="/">
    <event_type_a>42001090</event_type_a>
    <event_type_b>42001008</event_type_b>
    <factor>1.0</factor>
  </mixed_events>
  </clustering_parameters>
```

Paraver event – instructions counter

Builds IPC from Paraver events

Clustering steps – Clustering specification

- cluster_I_IPC.xml (and 2)

Metrics to extrapolate

Scatter plot axis:
IPC, Instructions

```
<projection_parameters>
    <single_event apply_log="yes" name="PM_INST_DISP">
      <event_type>42001091</event_type>
      <factor>1.0</factor>
    </single_event>
    <single_event apply_log="yes" name="PM_DATA_FROM_MEM">
      <event_type>42001012</event_type>
      <factor>1.0</factor>
    </single_event>
(...)
  <plot_options raw_metrics="yes">
    <x_metric title="IPC">cl_IPC</x_metric>
    <y_metric title="Instructions Completed">Native_Instr</y_metric>
  </plot_options>
</clustering_definition>
```

Clustering steps – Running burst-clusterizer

- clusterize.sh <tracefile>

```
#!/bin/bash
xml=cluster.I.IPC.xml
clusterize=/gpfs/apps/CEPBATOOLS/burst-clusterizer-devel/32/bin/burst-
clusterizer-stable

$clusterize -d $xml -c -i $1.prv $1_clustered.prv
```

Clustering steps – Running burst-clusterizer

```
Usage:
  ./burst-clusterizer-stable -d <clustering_def.xml> [OPTIONS] -i <input_trace>
[<output_trace>]

  -h                          This help
  -s                          Silent mode
  -d <clustering_def_xml>     XML containing the clustering process definition
  -b                          Print "block begin" and "block end" records
                              for each burst on output trace
  -c                          Generate the CPIStack model report for each cluster
                               found.
  -p <k>[,k_end]              Computes the k-neighbour (or range) distance in
                              terms of clustering parameter defined with '-d'.
                              Generates an GNUPlot to easily select the DBScan
                              parameters
  -m <eigen_matrix_file>      CSV file containing an eigenvectors matrix to
                              transform the original space
```

Clustering steps – Running burst-clusterizer

```
-x                              Extract *normalized* data from input file
  -r                              Extract *raw* data from input file to disk
  -a[n]                           Create cluster sequence to compute the alignment
                                  Using '-an' noise points are NOT FLUSHED in the
                                  resulting sequence
  -t                              Generate the file used to create a tree trough
                                  successive clusterings
  -i <input_trace>                Input Dimemas/Paraver trace
  -n                              Do NOT generate output trace (but, the output
                                  trace name is needed)
  <output_trace|output_data>      Output Dimemas/Paraver traces resulting from the
                                  clustering process or output data file if
                                  parameters '-x' or '-r' are used
```

Clustering steps – burst-clusterizer outputs

- Paraver tracefile with events identifying clusterid of the computing regions
  - `output_name.`**`prv`**`, ourput_name.`**`pcf`**

Next running state belongs to cluster 7

```
2:1:1:1:1:214115714:90000001:7
1:1:1:1:1:214115714:392776571:1
2:1:1:1:1:214115714:60000019:3:60000119:3:42001073:9455169:42001145:1907024
0:42001086:4369022:42001008:127463060:42001001:32383493:42001090:103633639:
42001079:32433965:42001211:393318
2:1:1:1:1:392776571:90000001:0
2:1:1:1:1:392776571:90000001:5
```

- Use configuration files to load predefined views as
  - cluster_id.cfg – timeline of the clusters
  - 2dp_cid.cfg – profile analysis of the different clusters
  - 3dh_ipc_cid.cfg – histogram of the IPC distribution for the different clusters

Clustering steps – burst-clusterizer outputs

- Scatter plot (gnuplot and table of points labeled with clustter_id)
  - `output_name.`**`cl_IPC.Native_Instr.gnuplot`**
  - `output_name.`**`clustered.csv`**

```
Native_Instr, cl_IPC, cluster_id
103633639.000000,0.813048,4
208944291.000000,0.515611,6
101473632.000000,0.825092,4
207310140.000000,0.516196,6
101805859.000000,0.826846,4
207311962.000000,0.515731,6
(...)
```

Clustering steps – burst-clusterizer outputs

- Table of projected metrics
  - `output_name.`**`clusters_info.csv`**

WARNING::: Performance metrics and CPI-Stack based on PPC 970 hardware counters

```
Cluster Name,Cluster 1,Cluster 2,Cluster 3,Cluster 4
Density,88,97,95,89
Total duration,19431319793,16573362782,9958380002,6484333782
Avg. duration,220810452,170859410,104825052,72857682
% Total duration,0.37049,0.31600,0.18987,0.12363
PM_CYC,188228124,386472209,236642031,164835232
PM_GRP_CMPL,46409199,55155653,25732210,31108464
PM_GCT_EMPTY_CYC,11076062,1331066,5634077,1828854
PM_GCT_EMPTY_IC_MISS,3935162,654553,747382,380387
(...)
```

- Inport this file to excel template `CPIStackDetailedDefinitiveNEW.{xls | ods}`