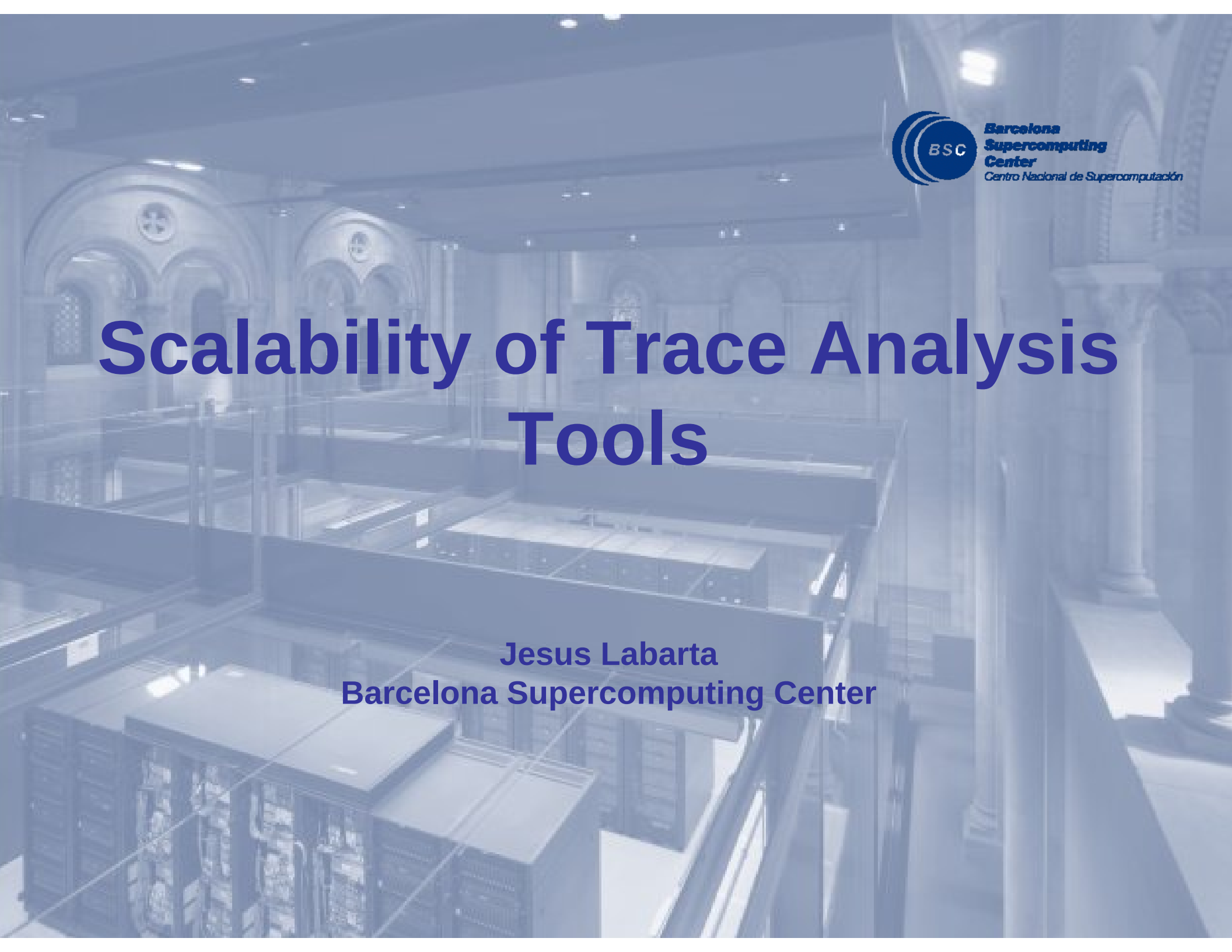


Scalability of Trace Analysis Tools

Jesus Labarta
Barcelona Supercomputing Center

What is Scalability?

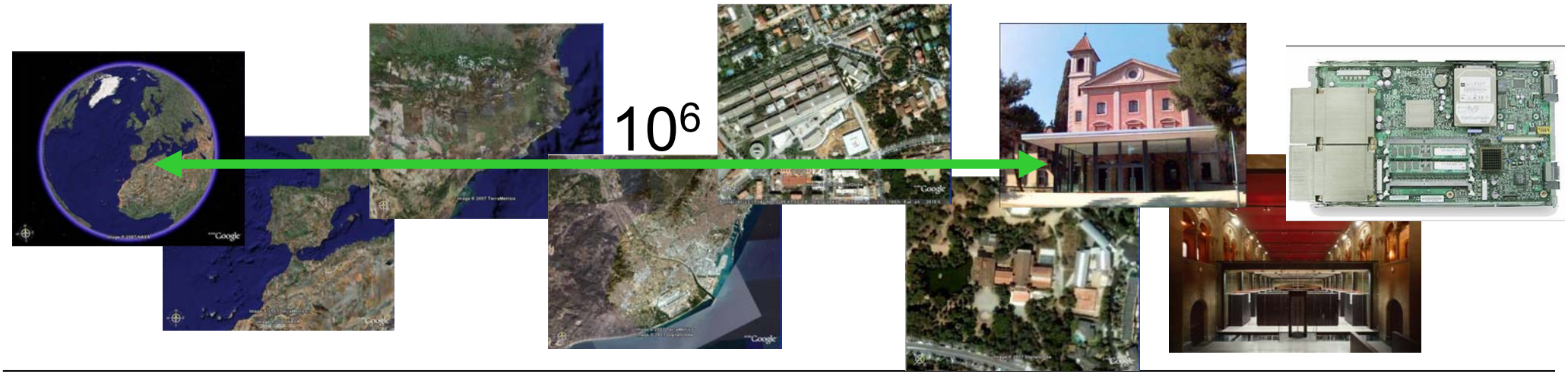
Index

- General view
- Scalability of instrumentation and preprocessing
- Scalability of display
- Dynamic range
- Analysis methodology
- Interoperability

Scalability

- Amount of data vs information

- Dynamic range



Performance analysis universe

Acquisition

Presentation

Profile land

Sampling land

$\int_a^b f(t)dt$

Processors/processes

Time

Event types

Instrumentation land

$f(t, p)$

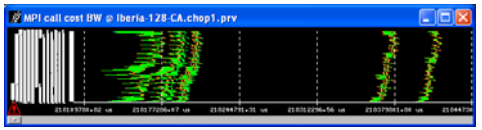
scalasca

Timeline land

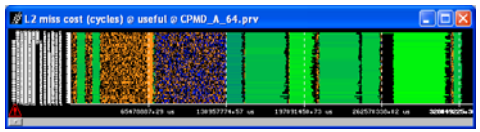
Scalability

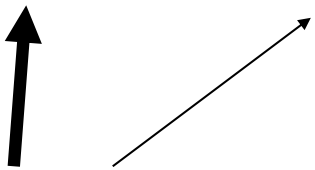
- Mechanisms
  - Separation of engine and display
  - Distributed implementation
  - Data encoding
  - Subset selection
  - Non linear rendering
  - Software counters
- Algorithms
  - Techniques to process the raw data
    - Signal processing, clustering,…
  - Metrics
    - Computation vs. MPI
  - Reported information
    - Counts vs models

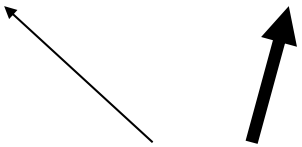
$$MPI\_call\_Cost = \frac{MPI\_call\_duration}{\#bytes}$$

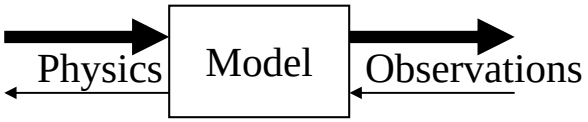


$$L2\_miss\_latency = \frac{\#cycles - \#instr / idealIPC}{\#L2misses}$$









CEPBA-tools towards scalability

Selection
- On-Off (time, processors, space)
- external control file
- events/information emitted (ie. MPI, HWC)
- Limit buffer sizes / duration
- Structure detection (i.e. periodicity)
- Circular buffer (issues: matching, density)
- min. duration states
- software counters (MPI_Probe,#MPIs, size)

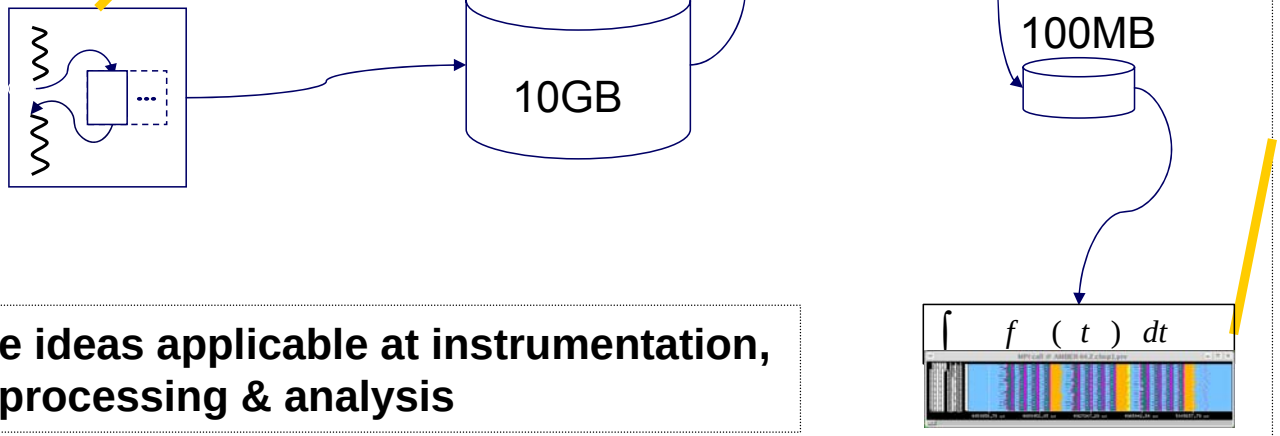
Parallel merge

Software counters
- count original events
- accumulate values (hwc)
- when: pediodic, condition

Subset selection
- time, processors
- trace size limit
- states/comms/events

Manual
- filters/GUI

Automatic



Functionality
- Non linear
- Composition
- Aggregation

Display
- Non linear render
  - what & color
- Generic subset of objects

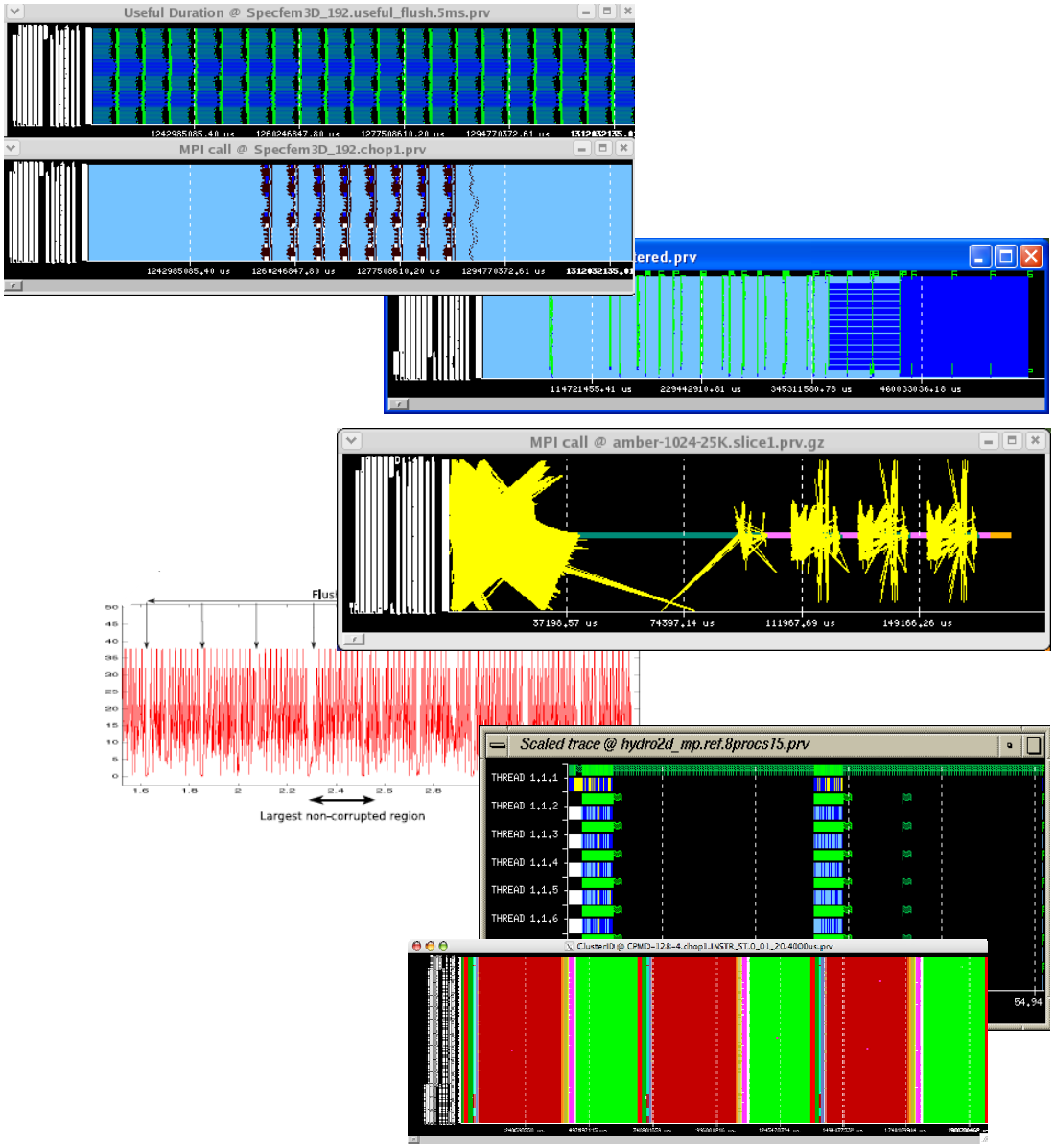
Performance
- Trace loading
- Metric comp. (intervals)
- OpenMP, Distributed

Same ideas applicable at instrumentation, postprocessing & analysis

Scalability of instrumentation and postprocessing

Scalability of instrumentation / preprocessing

- What/where/how
  - Selection
    - States, events
    - Time/space
  - Structure detections
    - Signal processing
    - Clustering
  - Summarization
    - Software counters
- Distributed implementation



Scalability of instrumentation / preprocessing

- XML control specification

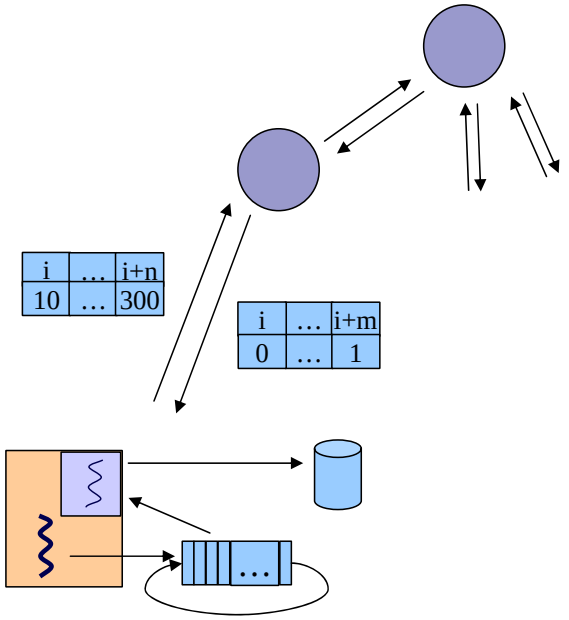
```
<trace enabled="yes" home="/gpfs/apps/CEPBATOOLS/64.hwc">
<mpi enabled="yes">
  <callers enabled="yes">1-3</callers>
  <counters enabled="yes" />
</mpi>

<openmp enabled="
  <locks enabled="n
  <counters enabled
 </openmp>

<user-functions ena
  <max-depth enable
  <counters enabled
 </user-functions>
```

```
<counters enabled="yes">
  <cpu enabled="yes" starting-set-distribution="1">
      <set enabled="yes" domain="all" changeatglobalops="5">
        PM_CYC,PM_DATA_FROM_MEM,PM_GCT_FULL_CYC,PM_INST_CMPL,PM_INST_DISP,PM_LD_MISS_L1,PM_LD_REF_L1,PM_ST_REF_L1
      </set>
      <set enabled="yes" domain="user" changeatglobalops="5">
        PM_BRQ_FULL_CYC,PM_BR_MPRED_CR,PM_BR_MPRED_TA,PM_CYC,PM_GCT_FULL_CYC,PM_INST_CMPL,PM_INST_DISP,PM_LD_MISS_L1
      </set>
  </cpu>
  <network enabled="yes" />
  <resource-usage enabled="yes" />
</counters>
```

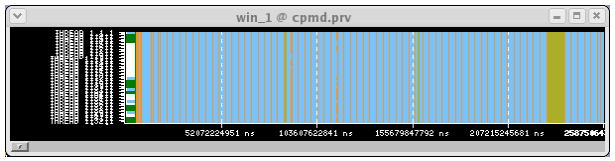
```
<bursts enabled="no">
  <threshold enabled="yes">500u</threshold>
  <counters enabled="yes" />
  <mpi-statistics enabled="yes" />
</bursts>
```

Distributed trace control

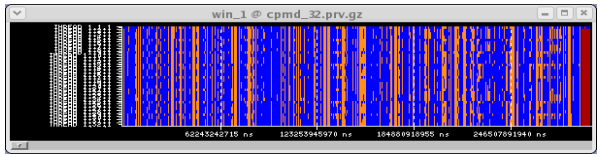
- MRNET based mechanism
  - Local instrumentation on a circular buffer
  - Periodic MRNet front-end initiation of collection process
  - Local algorithm
  - Reduction on tree
  - Selection at root propagated
  - Locally emit trace events



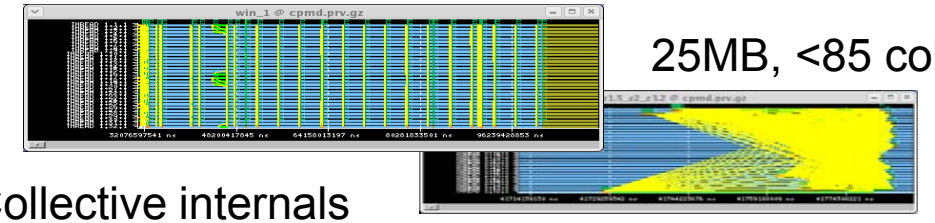
- Algorithm
  - Collective duration threshold



<1MB, <85 col



245MB, >15500 col



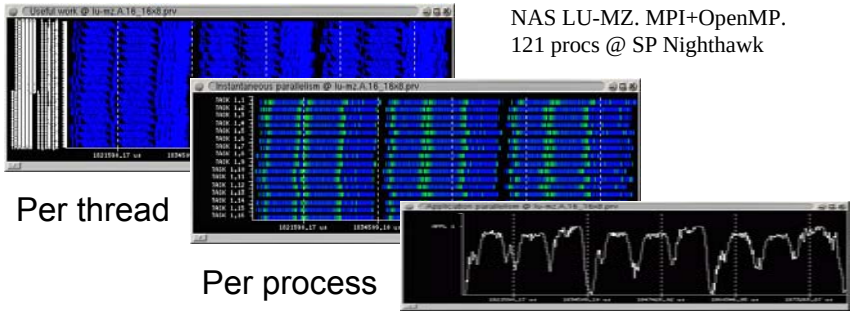
25MB, <85 col

Collective internals

Scalability of display

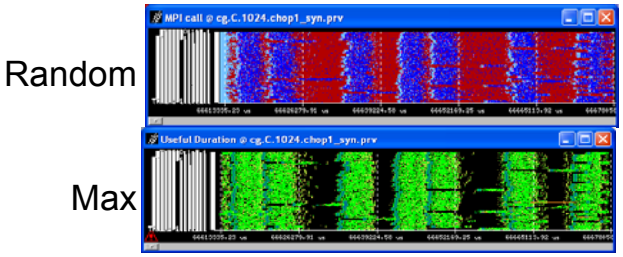
Scalability of Presentation

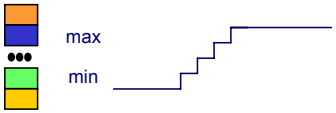
- Aggregation
  - Functional rather than scalability motivation



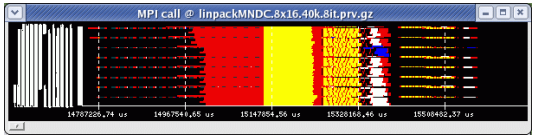
NAS LU-MZ. MPI+OpenMP.
121 procs @ SP Nighthawk

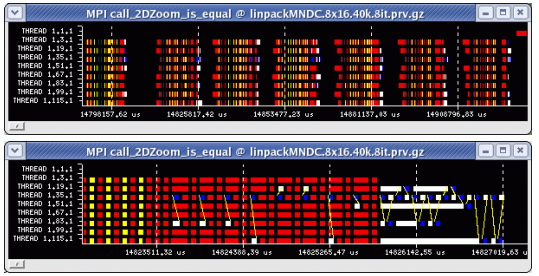
- Display
  - Non linear render
    - Value for pixel
    - Colors





- Objects
  - Any subset





Scalability of Presentation

- Linpack @ MareNostrum

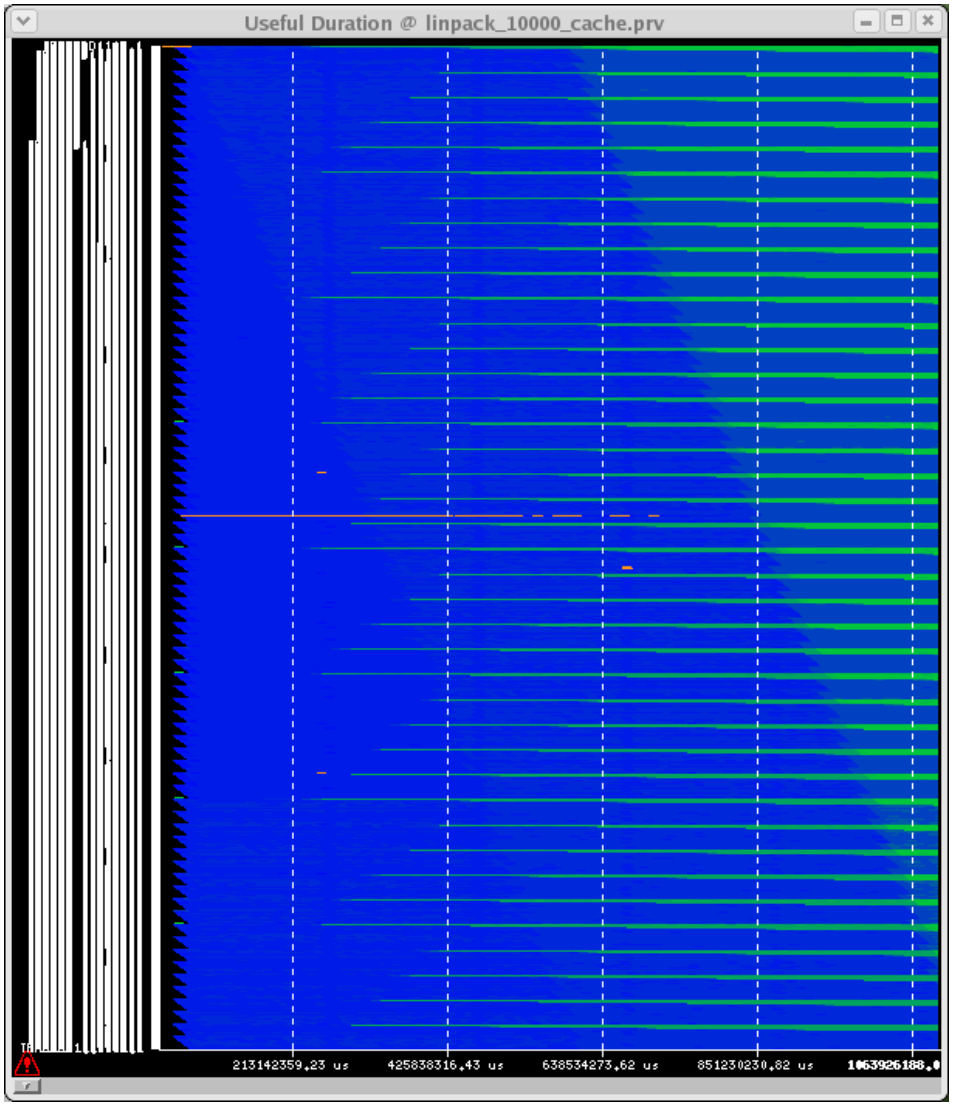
1700 seconds

Dgemm duration

11.8 s

10 s

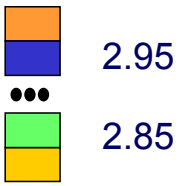
10000 processors



Scalability of Presentation

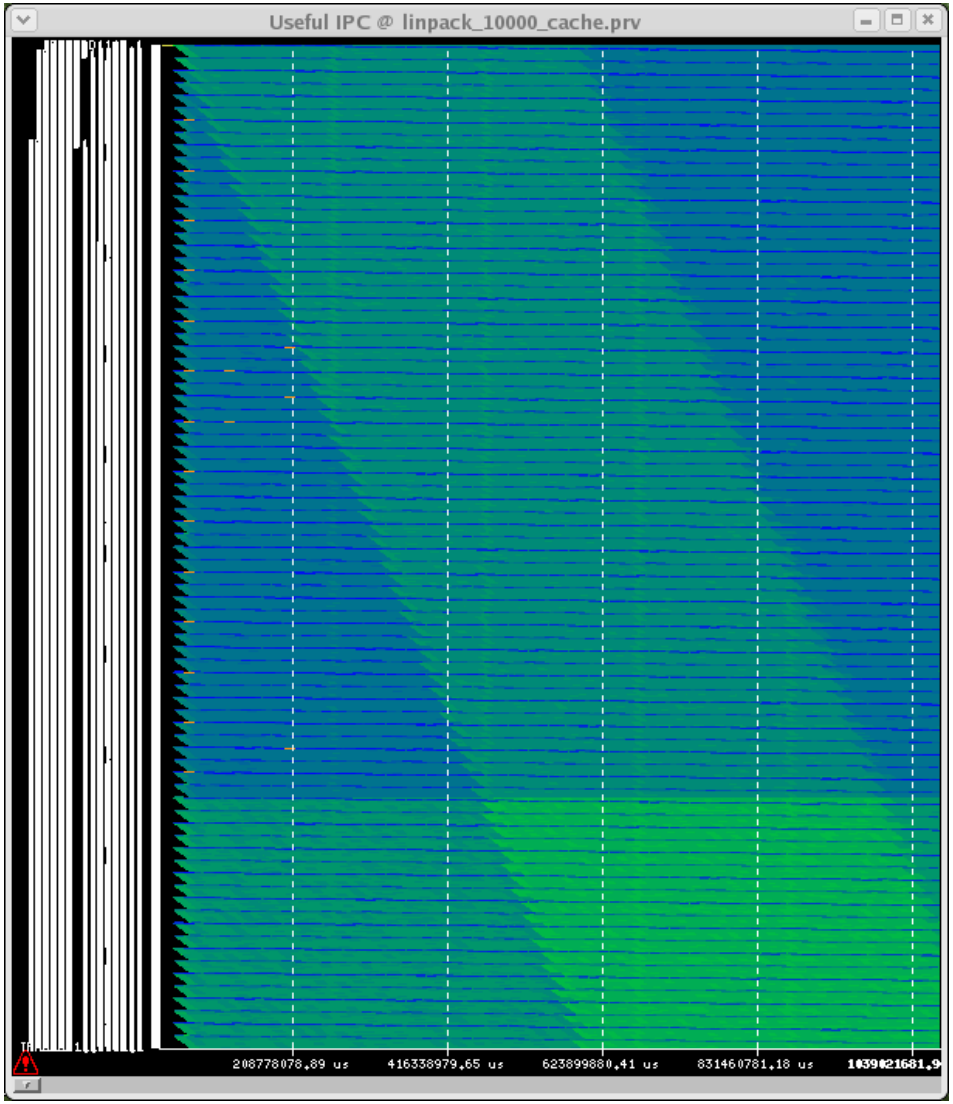
- Linpack @ MareNostrum

Dgemm IPC



1700 seconds

10000 processors



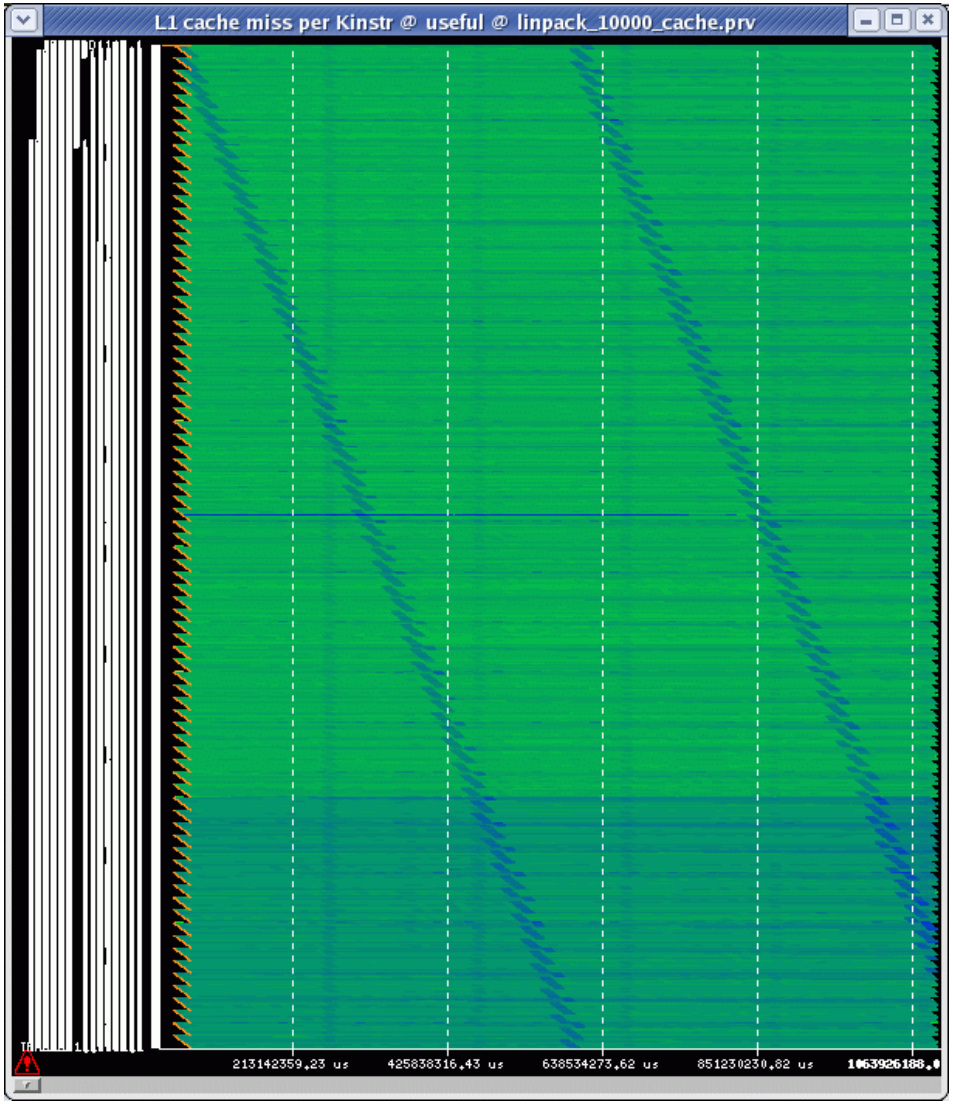
Scalability of Presentation

- Linpack @ MareNostrum

1700 seconds

Dgemm L1 miss ratio

10000 processors

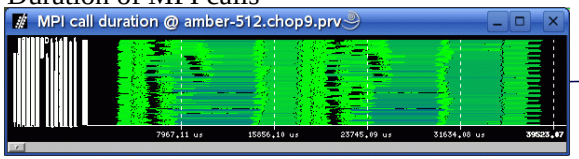


Dynamic range

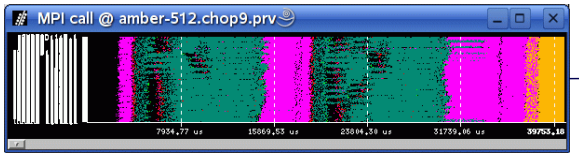
Interoperation between analysis and display

- AMBER @ 512 procs.
  - Which is the longest MPI call? Why?

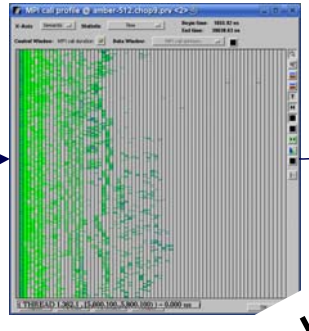
Duration of MPI calls



MPI calls

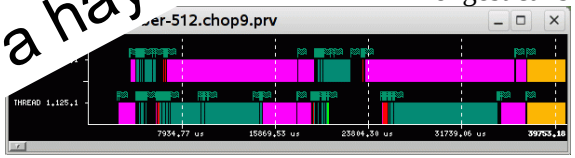


Histogran of duration of MPI calls

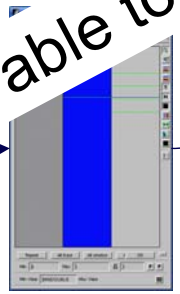


Duration of longest calls

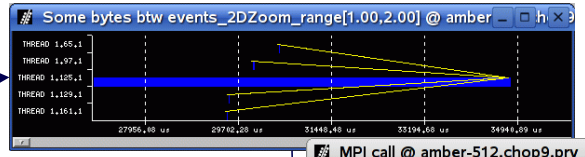
Longest calls



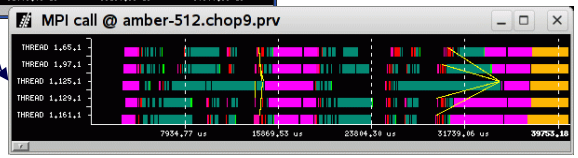
Who sends to 1



Who sends to 125 in selected area

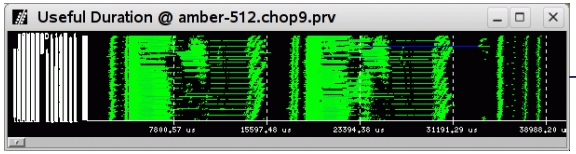


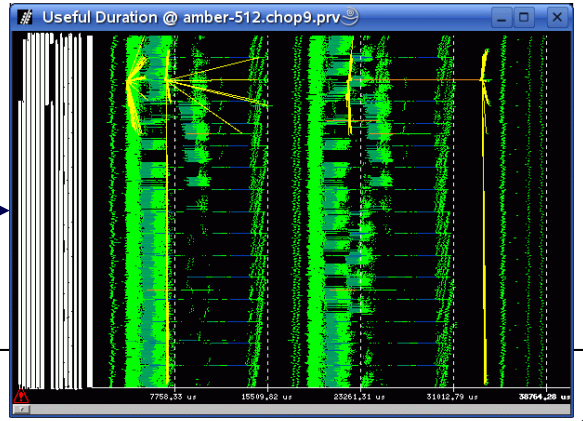
MPI calls form senders to 125



Scalability ≈ being able to find needles in a haystack

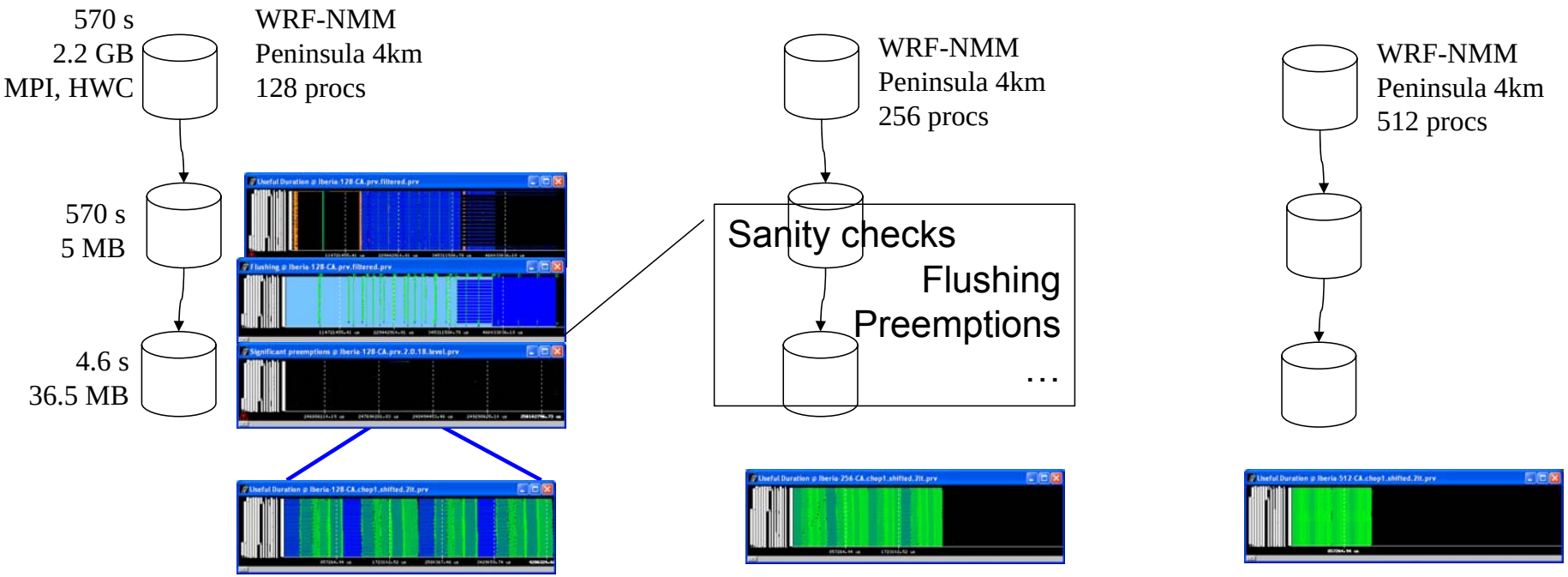
Useful duration



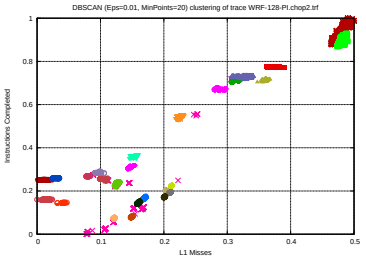


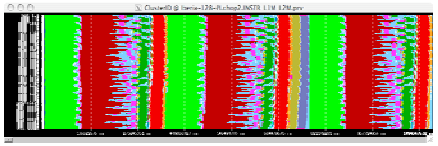
Analysis methodology

Methodology: Automatic analysis

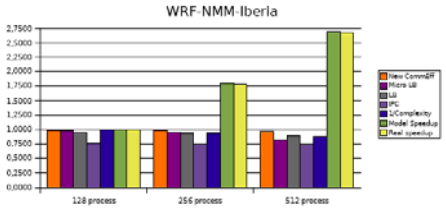


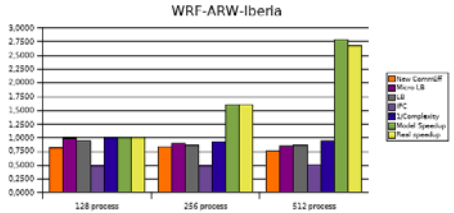
$$Sup = \frac{P}{P_0} * \frac{LB}{LB_0} * \frac{CommEff}{CommEff_0} * \frac{IPC}{IPC_0} * \frac{\#instr_0}{\#instr}$$





Region	IPC	L3D misses per 1000 instr	D TLB misses per 1000 instr	L1D $ misses per 1000 instr	Bytes / Instr
1	0,57	2,34	0,01	75,55	0,30
2	0,54	0,48	0,05	52,6	0,06
3	0,53	1,18	0,14	47,64	0,15
4	0,62	0,38	0,04	43,27	0,05
5	0,42	1,56	0,18	43,84	0,20





Clustering report

ClusterID	001	002	003	004	005	006
* Basic metrics						
|--> IPC	0.93	0.92	0.89	0.97	1.0	0.66
|--> MIPS	2050	2030	1950	2130	2190	1460
|--> MFLOPS	247.5	220.6	264.7	211.3	536.9	47.2
|--> L1 Data Misses/KInstr	0.66	0.97	0.98	0.40	1.14	1.31
|--> L2 Data Misses/KInstr	0.28	0.11	0.46	0.15	0.21	1.32
|--> Memory BW	70.5	26.7	110.3	39.9	54.8	235.0
* CPIStack Model						
|--> Completion Cycles	29.1	31.0	27.7	30.3	29.2	22.9
|--> Completion Table Empty	2.1	3.3	2.4	1.5	3.4	3.7
| |--> I-Cache Miss Penalty	1.3	2.8	1.9	0.5	3.0	3.2
| |--> Branch redirection	0.1	0.2	0.2	0.0	0.2	0.5
| |--> Others 0.7	0.3	0.3	1.0	0.2	-0.0	
|--> Completion Stall Cycles	68.7	65.7	70.9	68.2	67.4	73.4
| |--> Stall by LSU instruction	20.8	14.9	24.0	19.2	16.8	35.2
| | |--> Stall by reject	0.9	0.8	1.0	0.6	1.0	1.7
| | |--> D-Cache miss	9.4	4.2	13.2	6.7	7.8	24.7
| | |--> LSU Basic Latency	10.5	10.0	9.8	11.9	8.0	8.8
| |--> Stall by FXU instruction	16.0	18.7	14.2	18.0	8.0	17.5
| | |--> DIV/MTSPR/ecc	0.0	0.0	0.0	0.0	0.0	0.0
| | |--> FXU basic latency	16.0	18.7	14.2	18.0	7.9	17.5
| |--> Stall by others	32.0	32.1	31.8	31.0	42.6	20.6

Interoperability

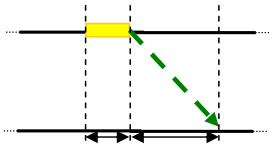
Interoperability

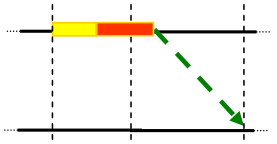
- Between Paraver Modules
  - Search: 2D → timeline
- To standard data analysis tools
  - Excel
  - OpenDX
- Between Performance analysis tools
  - Paraver - Dimemas
  - Fine grain network simulators
    - FSIM
    - IBM -Zurich
  - Compatibility with other trace formats and analyzers
  - Higher level application models:
    - Performance assertions (ORNL)
  - Fine grain processor performance:
    - Instruction level simulator
    - Processor performance models

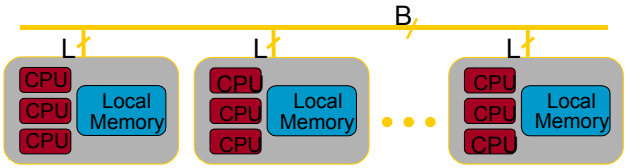
Dimemas: Coarse grain, Trace driven simulation

- Simulation: Highly non linear model
  - Linear components
    - Point to point communication
    - Sequential processor performance
      - Global CPU speed
      - Per block/subroutine
  - Non linear components
    - Synchronization semantics
      - Blocking receives
      - Rendezvous
    - Resource contention
      - CPU
      - Communication subsystem
        - links (half/full duplex), busses

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

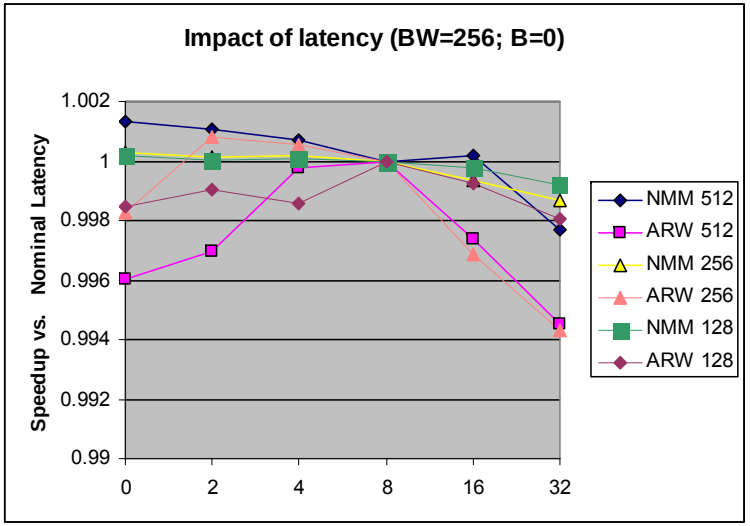


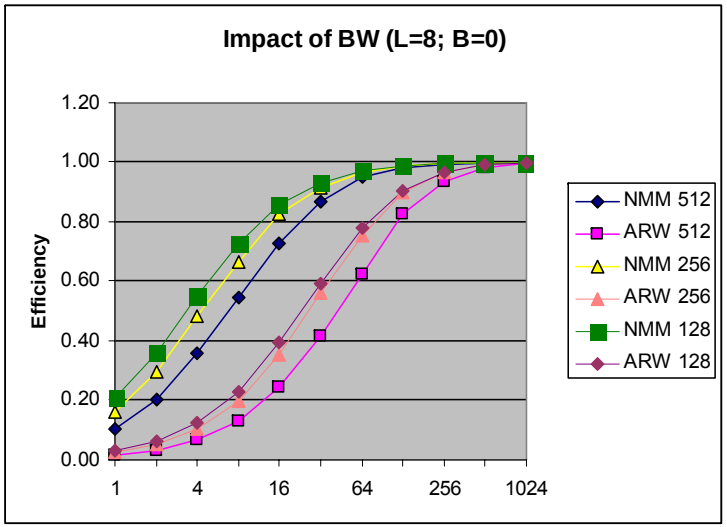


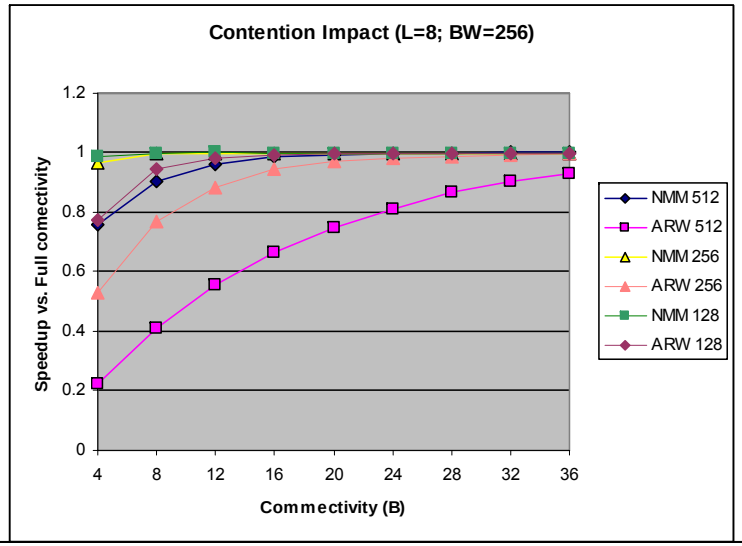


Network sensitivity

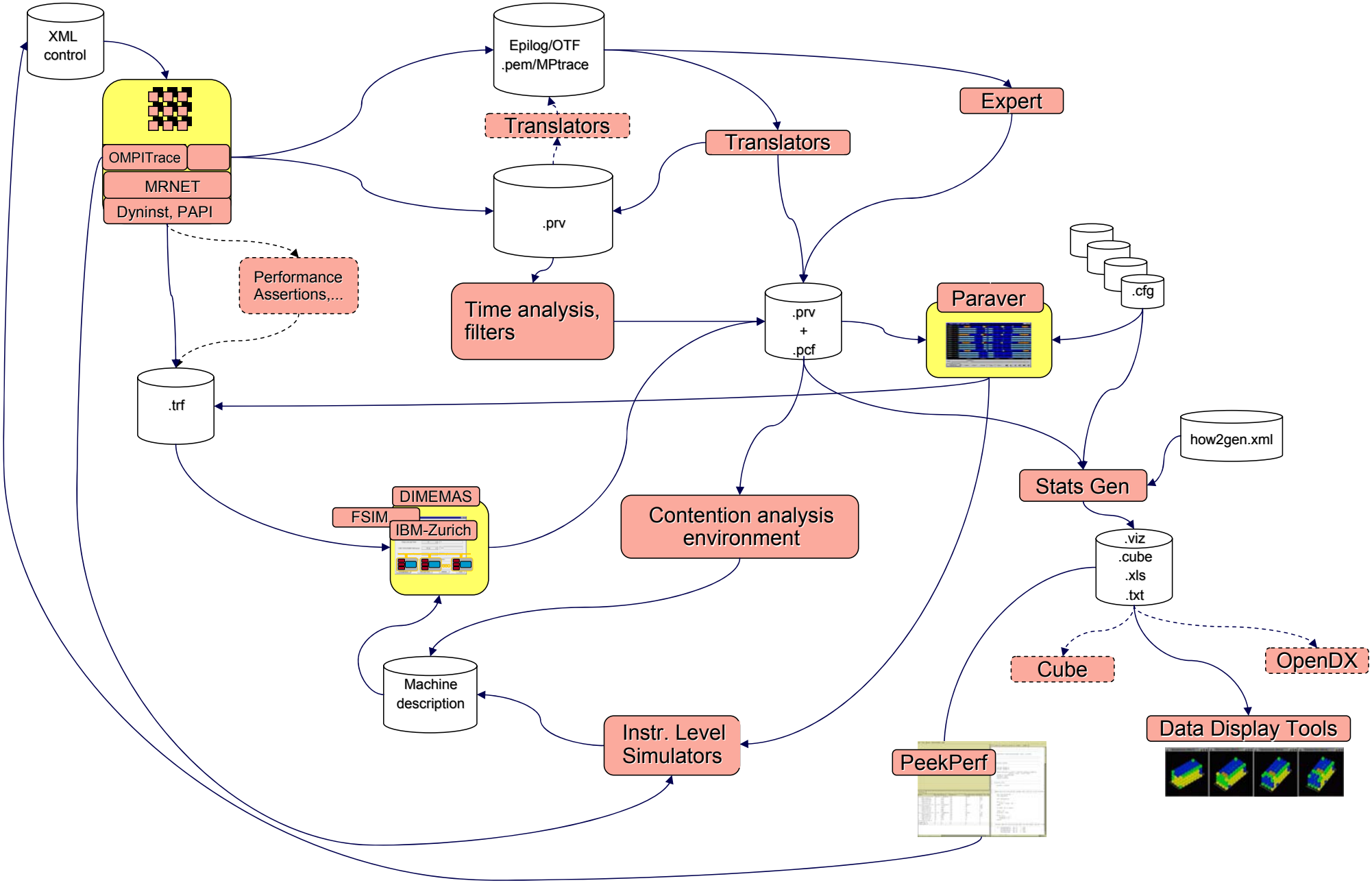
- Simulations with 4 processes per node
- NMM Iberia 4Km
  - Not sensitive to Latency
  - 512 sensitive to contention?
  - 256 MB/s OK
- ARW Iberia 4 Km
  - Not sensitive to Latency
  - sensitive to contention
  - Need 1GB/s







Interoperability



Conclusion

- Traces useful to understand and develop still at large process counts.
- Importance of variance in time and space.
- Trace visualization should support for high dynamic range.
- Importance of algorithms vs mechanisms
- Interest in integration/interoperation:
  - Probe injection mechanisms
  - Hwc
  - Trace control mechanism and specification
  - Profiler output to drive tracing run
  - Scalable acquisition infrastructure
  - Trace formats
  - Control of fine grain instrumentation
  - Signal and data interfaces
  - Profile files format