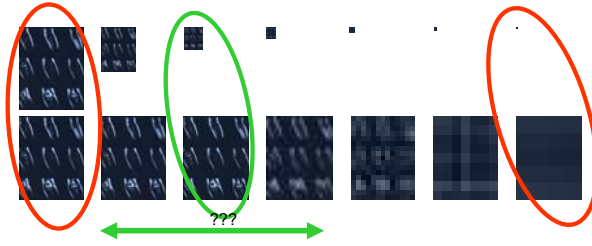


Index

- General view
- Scalability of instrumentation and preprocessing
- Scalability of display
- Scalability, models and automatic analysis
- Summary

Scalability

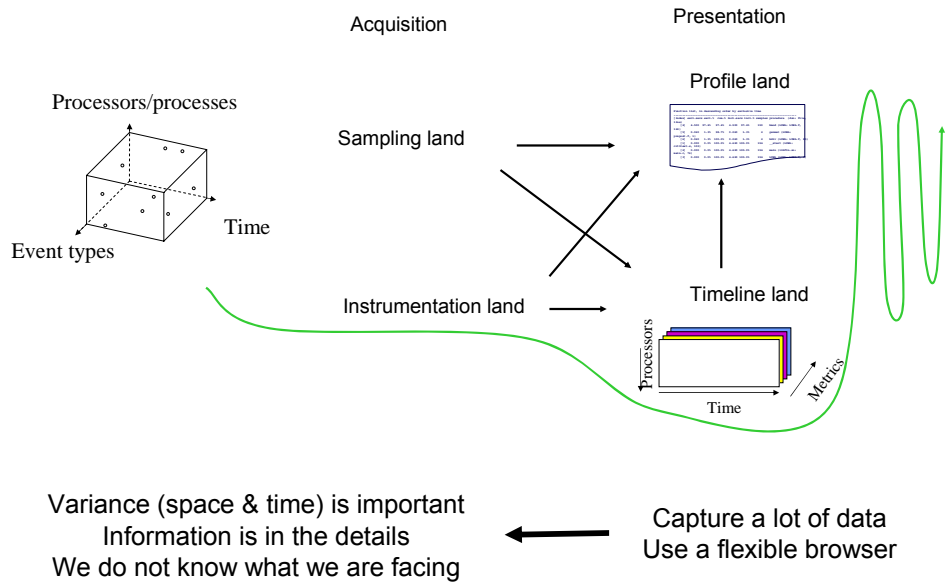
- Amount of data vs information



- Dynamic range



Performance analysis universe



Jesus Labarta, SC-APART, Reno, Nov. 2007

5

CEPBA-tools towards scalability

Selection

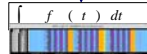
On-Off (time, processors, space)
external control file
events/information emitted (ie. MPI, HWC)
Limit buffer sizes / duration
Structure detection (i.e. periodicity)
Circular buffer (issues: matching, density)
min. duration states
software counters (MPI_Probe, #MPIs, size)

Parallel merge



10GB

100MB



Same ideas applicable at instrumentation, postprocessing & analysis

Software counters

count original events
accumulate values (hwc)
when: periodic, 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

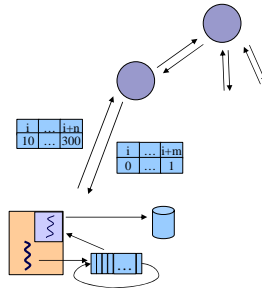
Trace loading
Metric comp. (intervals)
OpenMP, Distributed

Jesus Labarta, SC-APART, Reno, Nov. 2007

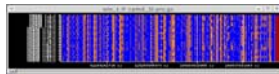
6

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



245MB, >15500 col



<1MB, <85 col



25MB, <85 col

Collective internals

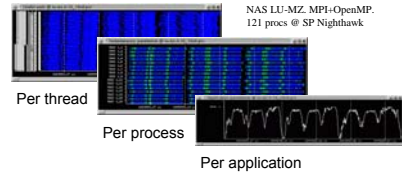


Scalability of display

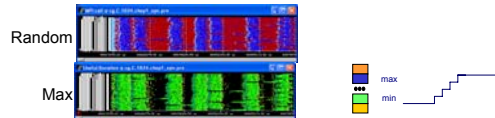


Scalability of Presentation

- Aggregation
 - Functional rather than scalability motivation



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



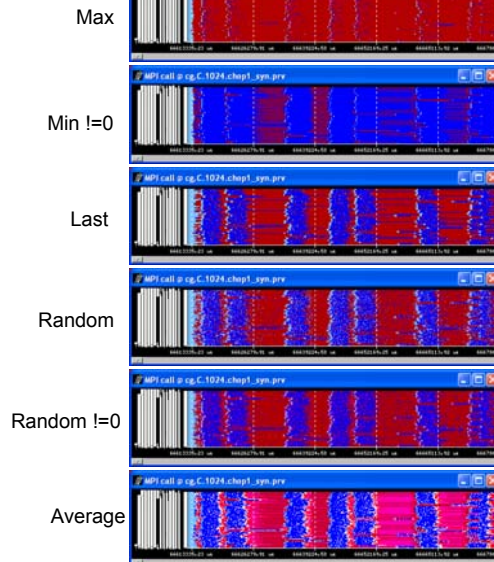
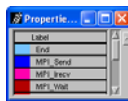
- Objects
 - Any subset



Scalability of display

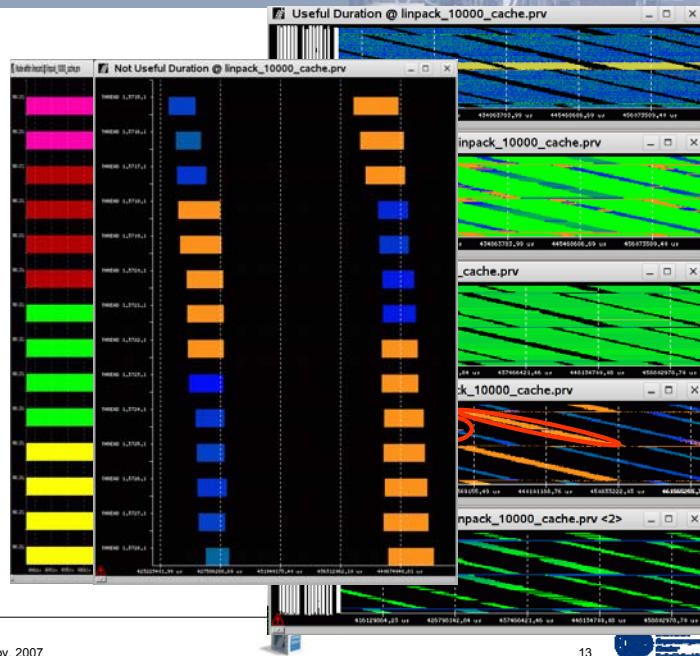
- Non Linear Render

CG.C
1024 CPUs
MPI call

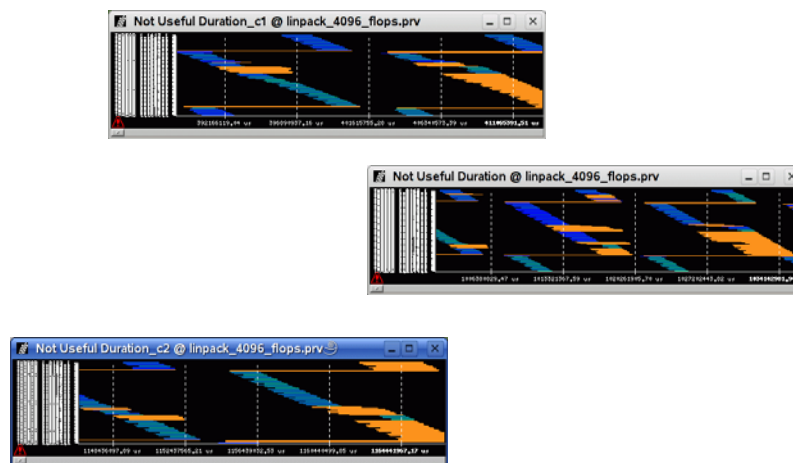


Display and navigation

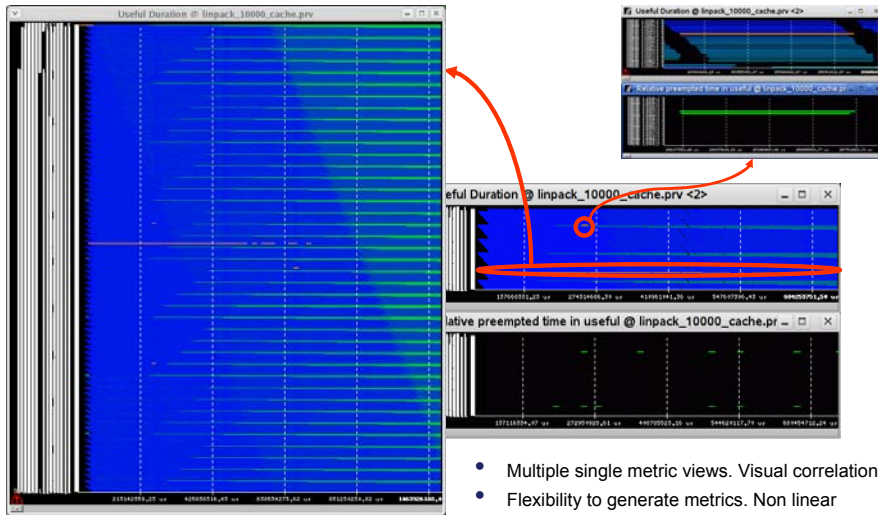
Linpack
@10000 CPUs
3000 seconds
500MB trace



Impact of contention?



Display and navigation

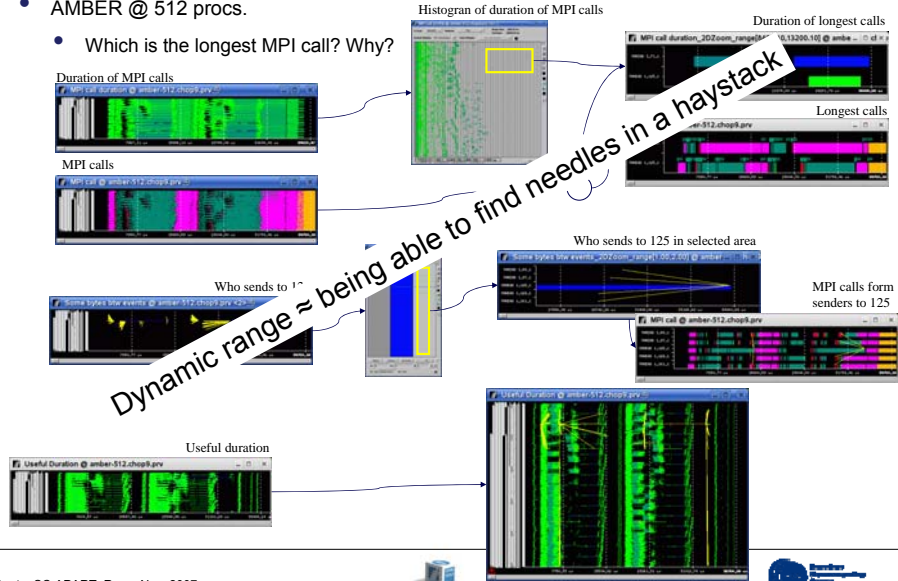


- Multiple single metric views. Visual correlation
- Flexibility to generate metrics. Non linear
- Zooming and synchronization capabilities
- Non linear 2D rendering
- Non linear coloring. "few" levels. Scale tuning

Interoperation between analysis and display

- AMBER @ 512 procs.

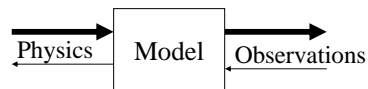
- Which is the longest MPI call? Why?



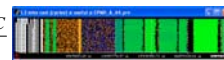
Scalability, models and automatic analysis

Scalability and models

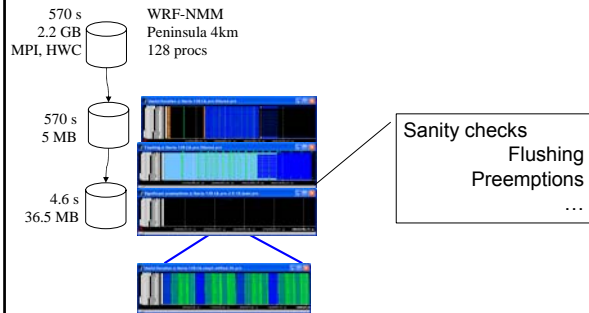
- Models are key for performance analysis
 - **Reference for observed metrics**
 - **Identify key factors that explain behavior**
- Common bad practices
 - We very seldom use them
 - We use then one way:
 - estimate/prediction
 - Seldom for parameter fitting
 - Obsessed by detail instead of modeling just key factors.
 - Obsessed by accuracy of prediction instead of properly capturing trends
- Models:
 - Bounds: nominal/microbenchmarks (MFLOPS, MIPS, IPC, MB/s,...)
 - Analytic: parameters obtained from microbenchmarks and (manual) complexity analysis
 - Simulators of different levels of detail to estimate reasonably achievable performance
 - **Hierarchical convolution**



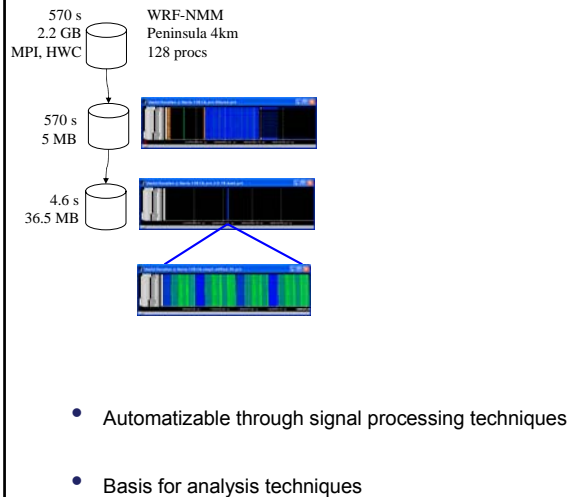
$$L2_miss_latency = \frac{\#cycles - \#instr / idealIPC}{\#L2misses}$$



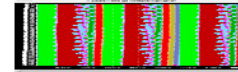
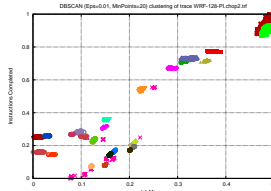
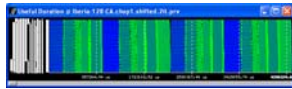
Automatic structure detection



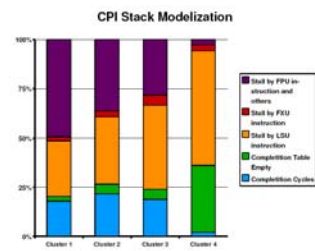
Automatic structure detection



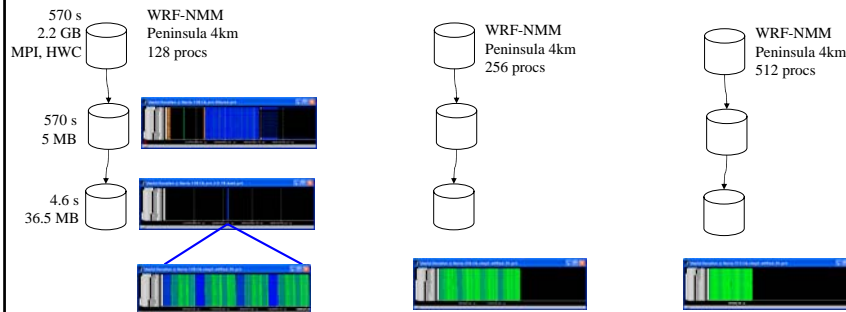
Clustering techniques



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	1.55	0.16	43.84	0.20



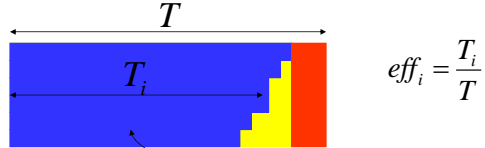
Methodology: Automatic analysis



- Automatizable through signal processing techniques
- Basis for analysis techniques

Scaling model: based on measurements

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



$$CommEff = \max(eff_i)$$

$$LB = \frac{\sum_{i=1}^P eff_i}{P * \max(eff_i)}$$

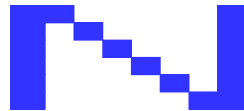
IPC
 $\#instr$

Scaling model: adding simulation capabilities

$$Sup = \frac{P}{P_0} * \frac{macroLB}{macroLB_0} * \frac{microLB}{microLB_0} * \frac{CommEff}{CommEff_0} * \frac{IPC}{IPC_0} * \frac{\#instr_0}{\#instr}$$



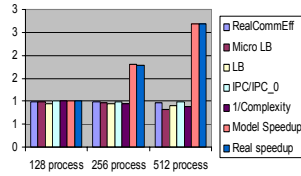
Migrating/local load imbalance



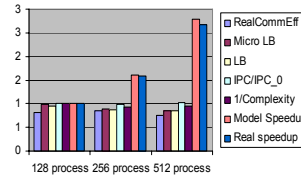
serialization

Scaling model

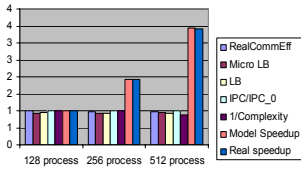
WRF-NMM-Iberia



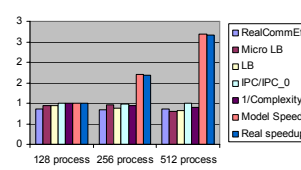
WRF-ARW-Iberia



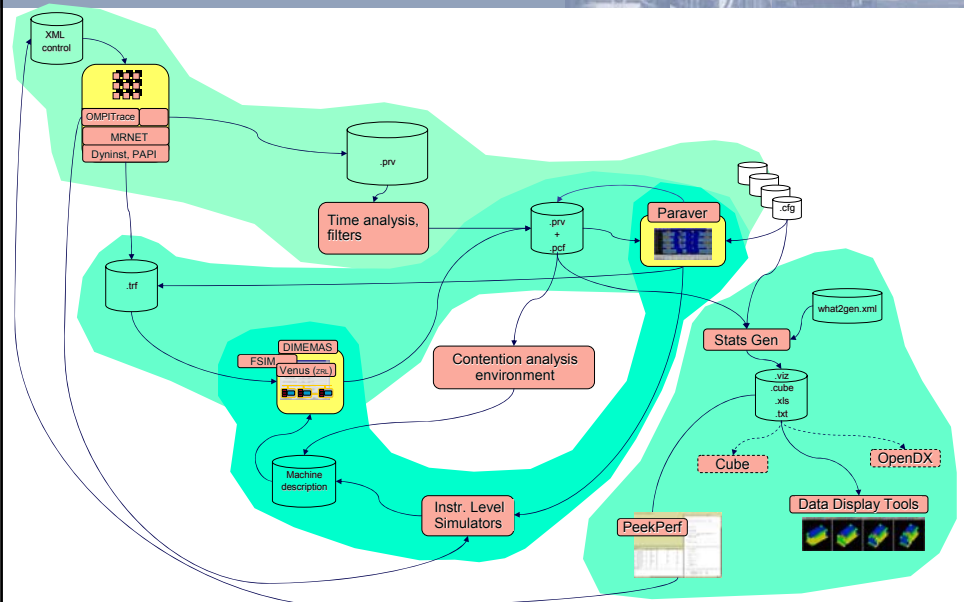
WRF-NMM-Europe



WRF-ARW-Europe



Modular, interoperable infrastructure





Summary



Scalability



- Mechanisms
 - Separation of engine and display
 - Distributed implementation
 - Data encoding
 - Subset selection
 - Non linear rendering
 - Logical pixels (2x2, 3x3, M x N?)
 - Software counters
- Algorithms
 - Techniques to process the raw data
 - Signal processing, clustering,...
 - Metrics
 - Should be useful
 - Understandable by "mortals"
 - lead to right decision making (ie. Computation vs MPI)
 - **Based on models**

↑ ↗
Emphasis
Necessary,
not sufficient

↖ ↑
Importance
Intelligence
Automatic



Thesis

“A **single instrumented run**
captures a **lot of information**
that is essentially **thrown away**
in current parallel programming
practice”

“It is possible to
squeeze the information
in the trace”

An analogy



Use of traces

Huge probe effect

Team work

Multidisciplinary

Correlate different sources

Speculate till arriving to consistent theory

