New techniques and integration efforts in the CEPBA-Tools environment

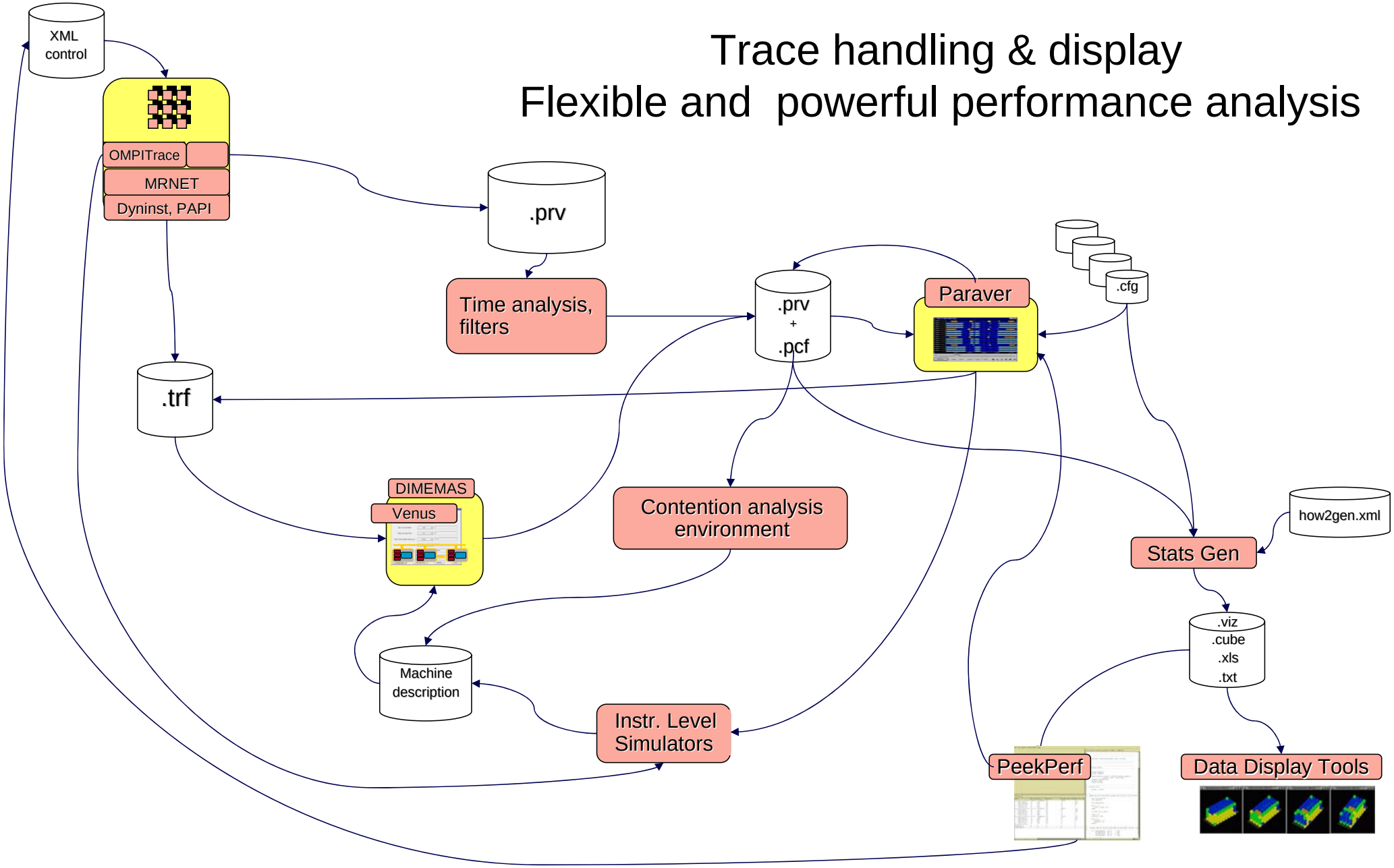
Jesus Labarta
Barcelona Supercomputing Center

Index

- General overview
- Recent developments
  - Time analysis
  - Modelling
  - Clustering
  - Sampling
- Tools integration
  - Peekperf
- Interconnection evaluation
- Hierarchical modelling / prediction

Paraver

CEPBA-Tools Environment



Trace handling & display
Flexible and powerful performance analysis

Offer

- Paraver — OPEN SOURCE by end of the year
- Dimemas. — OPEN SOURCE by end of the year
- Instrumentation — OPEN SOURCE by end of the year

- Structure analysis tools — in development
  - Signal analysis
  - Clustering
  - Sampling + tracing

- OpenMP incl 3.0 tasks — OPEN SOURCE
- StarSs: CellSs / SMPSs /GPUSs — OPEN SOURCE

CEPBA-Tools Environment

Analysis of a run

XML control

OMPITrace

MRNET

Dyninst, PAPI

.prv

Time analysis, filters

.prv + .pcf

Paraver

.cfg

.trf

DIMEMAS

Venus

Contention analysis environment

how2gen.xml

Stats Gen

.viz

.cube

.xls

.txt

Machine description

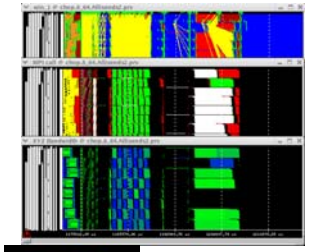
Instr. Level Simulators

PeekPerf

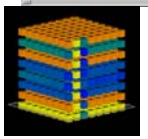
Data Display Tools

Parallel Program Instrumentation: Platforms

BGL



PAPI



Network counters

PERUSE

By Rainer Keller (HLRS) & (UTK)

SX8

By Rainer Keller (HLRS)

MPI Calls

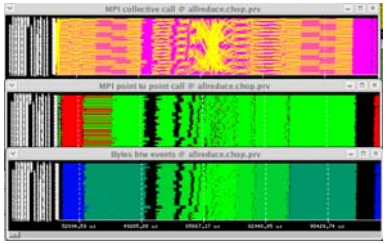
MFLOPs

% Vector instr

Avg vector length

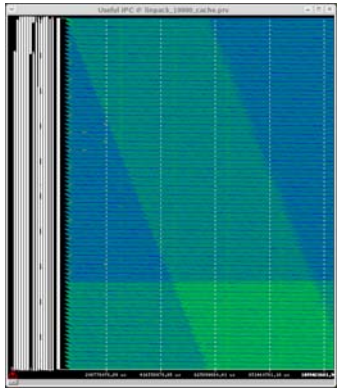
Bank conflicts/us

MPICH collective internals



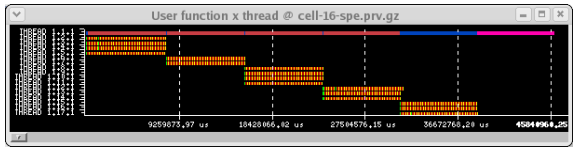
Marenostrum

IPC



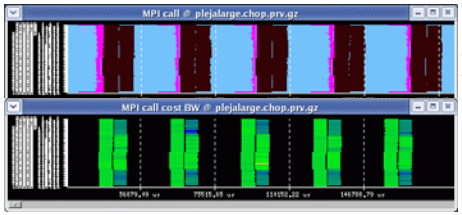
MPI + OpenMP
Lib. Preload
PAPI

Cell BE



Power 5 AIX

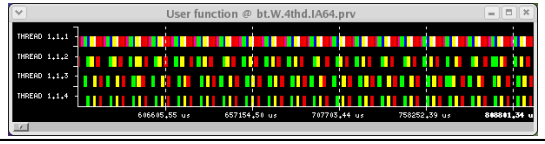
MPI + OpenMP
DPCL
PMAPI



MPI calls

Bandwidth

Altix



MPI + OpenMP
Dyninst
PAPI
(Supported by NASA AMES)

(O)MPI-Trace

- 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>
```

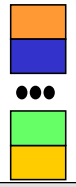
```
<counters enabled="yes">
  <cpu enabled="yes" starting-set-distribution="1">
       <set enabled="yes" domain="all“ changeatglobalops="5">
         PM_CYC,PM_DATA_FROM_MEM,PM_GCT_FULL_CYC,PM_INST_CMPL,PM_INST_DISP,PM_LD_MISS_L1,PM_LD_REF_L1,PM_ST_REF_L1
       </set>
       <set enabled="yes" domain="user"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

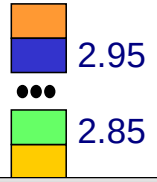
Scalability of Presentation: timelines

- Linpack @ Marenostrum: 10k cores x 1700 s

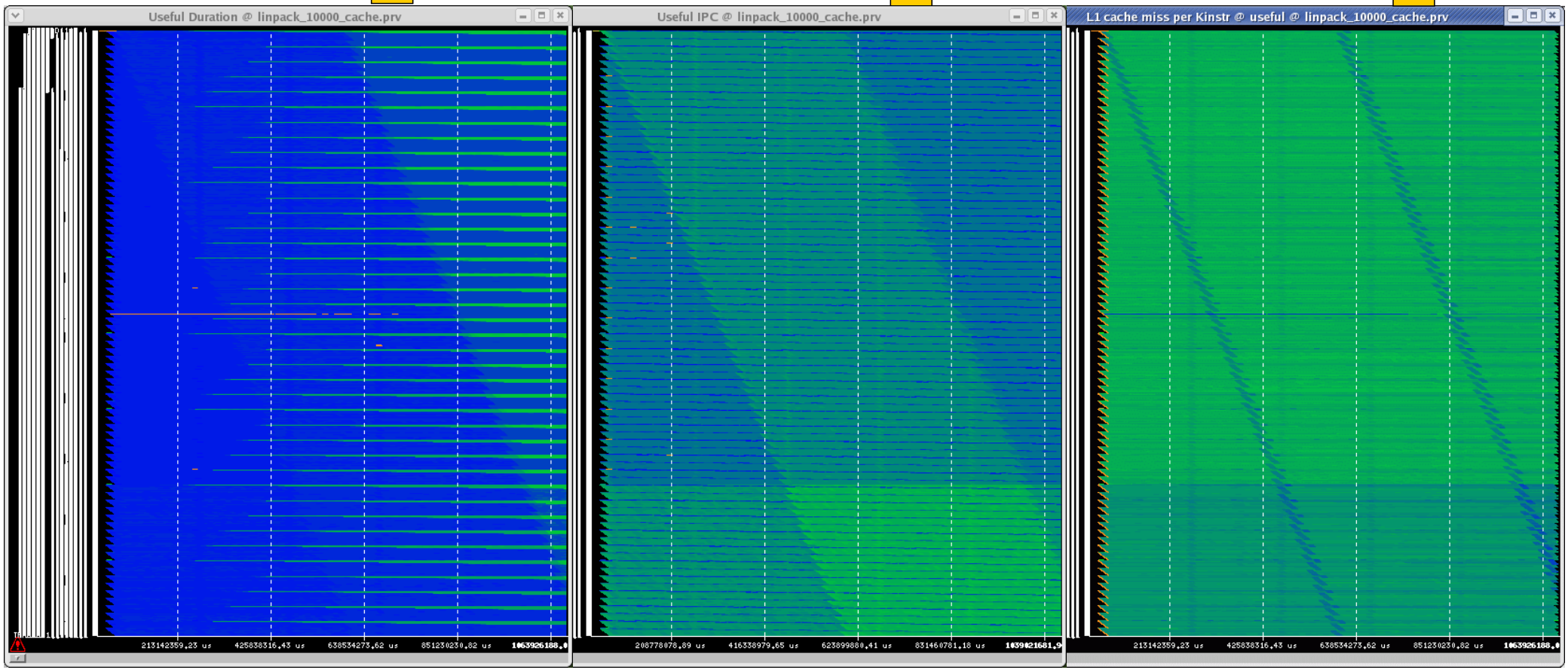
Dgemm duration



Dgemm IPC



Dgemm L1 miss ratio



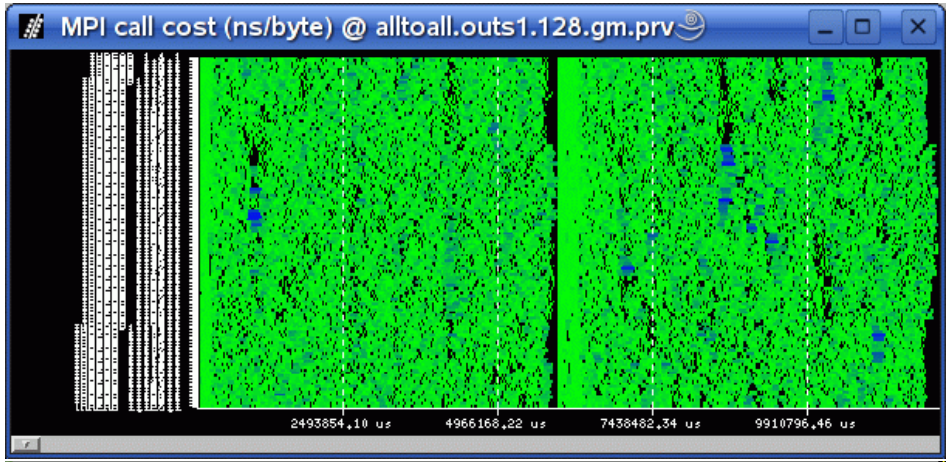
Scalability

- A dynamic range issue

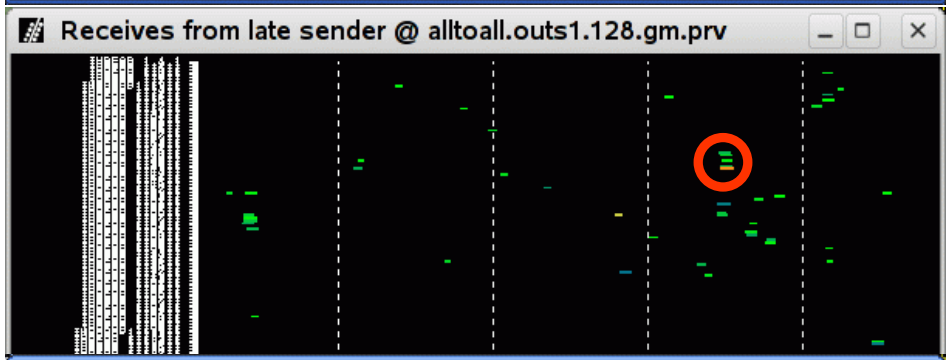
MPI calls

MPI comms.

ns/byte



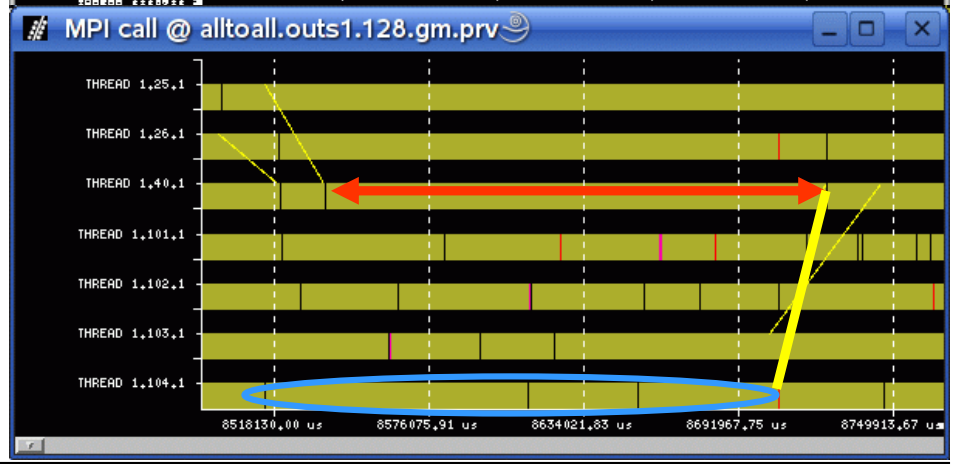
Early receivers



Severe early receiver

- Real causes may be far away from observed effects
- Need to integrate measurement and modeling

Comms severe early receiver

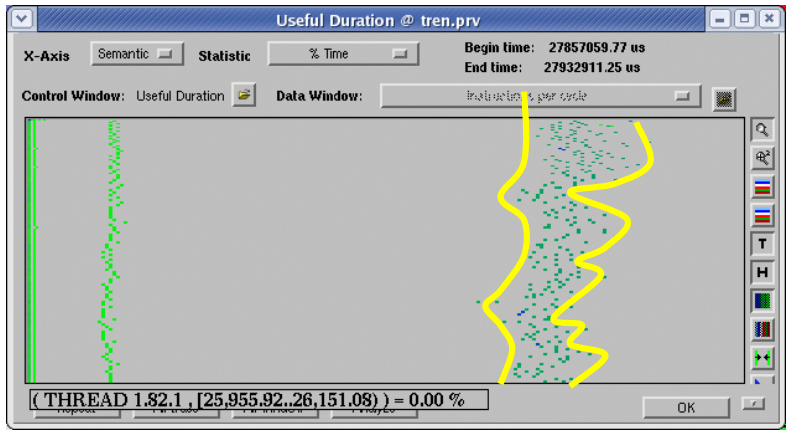


Detail severe early receiver

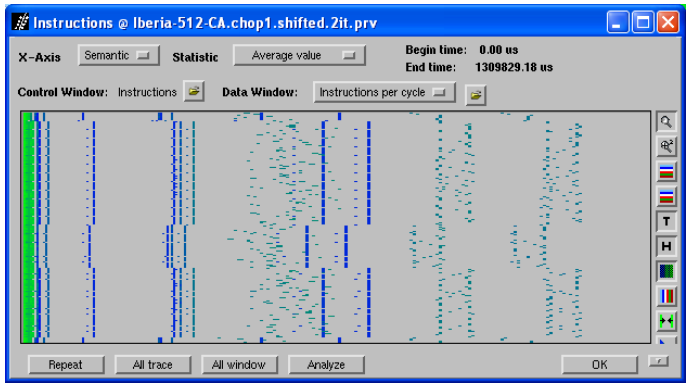
Scalability of Presentation: histograms

- There is more load imbalance than typically aware/accepted

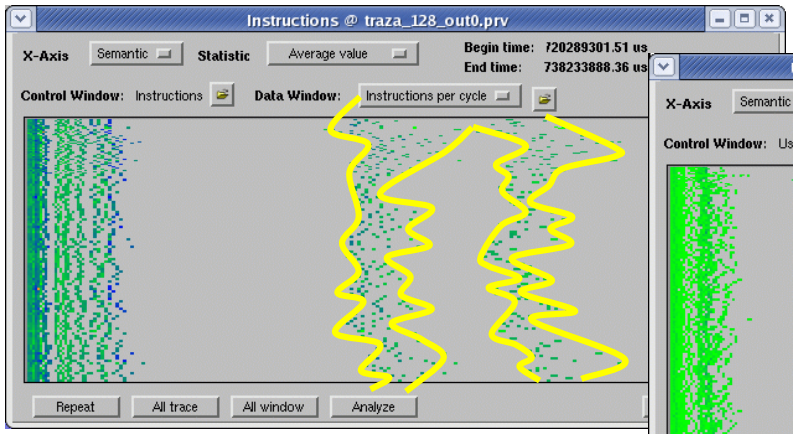
Duration Alya @ 128



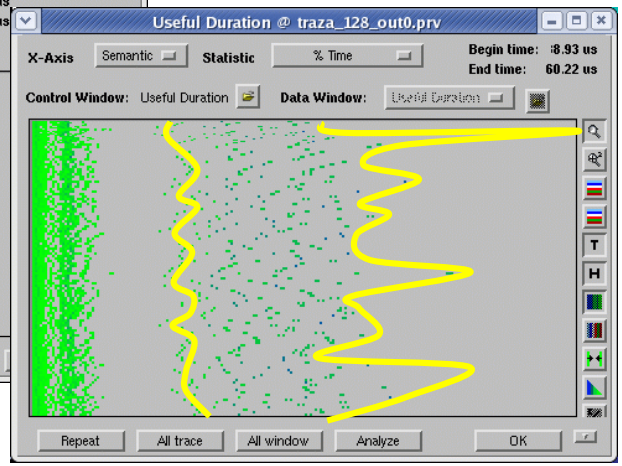
Instructions WRF @ 512 (Iberia12K)



Instructions Airbus @ 128

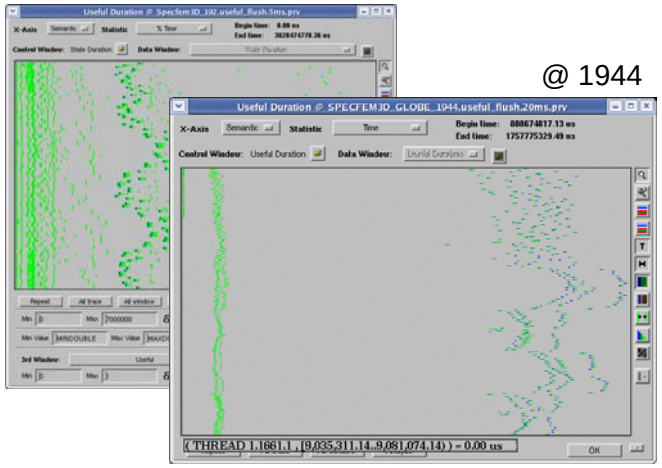


Duration Airbus @ 128

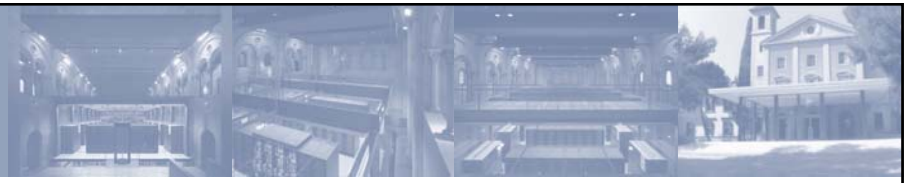


Duration SPECFEM3D @ 192

@ 1944



Trace handling utilities



- Filter
  - Duration, Type of event, Size, Space.
- Cut
- Merge
- Shift
- Software counters
- Paramedir
- Trace format converters: OTF → Paraver, AIXtrace → Paraver, …

- Combinations
  - Minimize backwards comms: optimization loop
    - Paramedir
    - Shift
  - Extract a communication phase sub-trace
    - Shift, cut, shift, cut

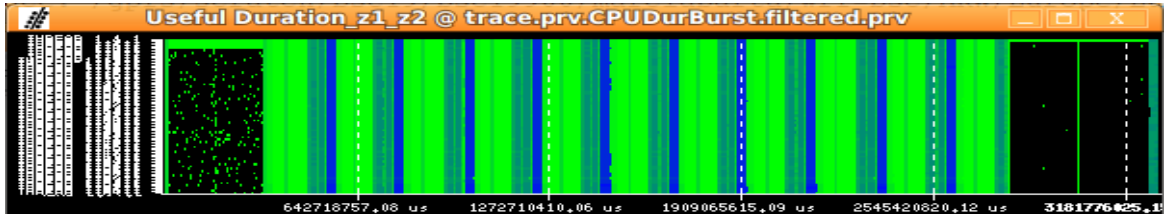
Signal processing

Signal processing and performance analysis

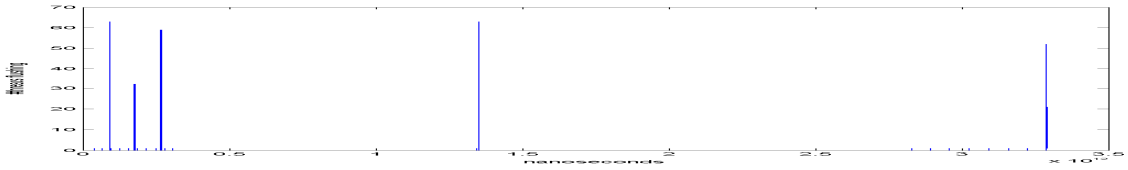
- Signals
  - Flushing processes, %preemted time, #msgs/BW (clogged system)
  - Sum burst duration, #processes in MPI, average IPC,…
- Mathematical morphology
- Spectral analysis
  - Autocorrelation
  - Wavelet transform
- Useful
  - Periodic structure: Reference focus for detailed analysis

An example

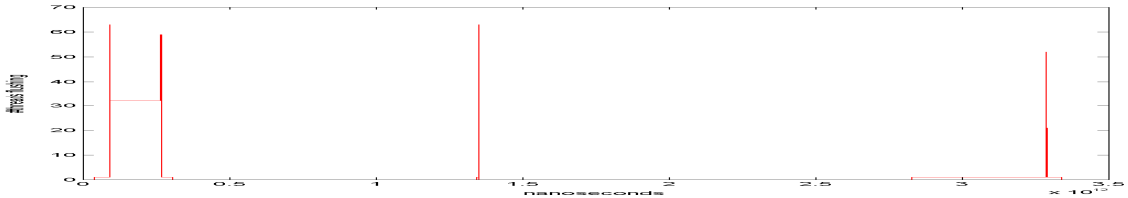
- CPMD



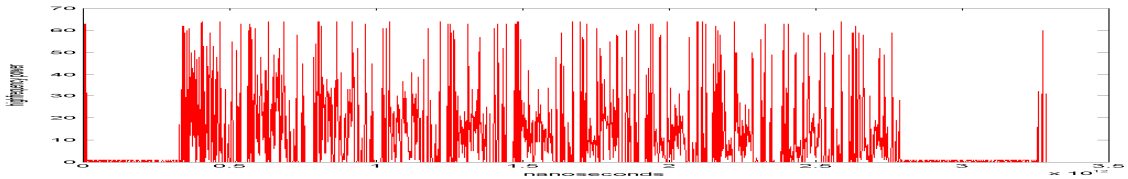
Flushing



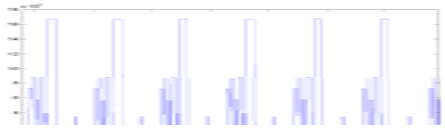
Flushing filtered



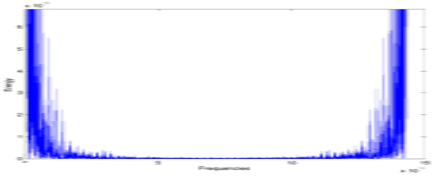
Wavelet
high frequency



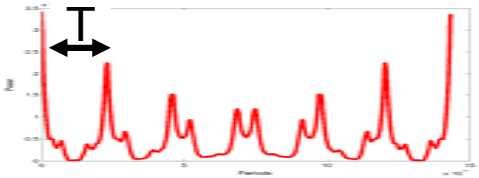
Σ Useful Duration



Spectral density

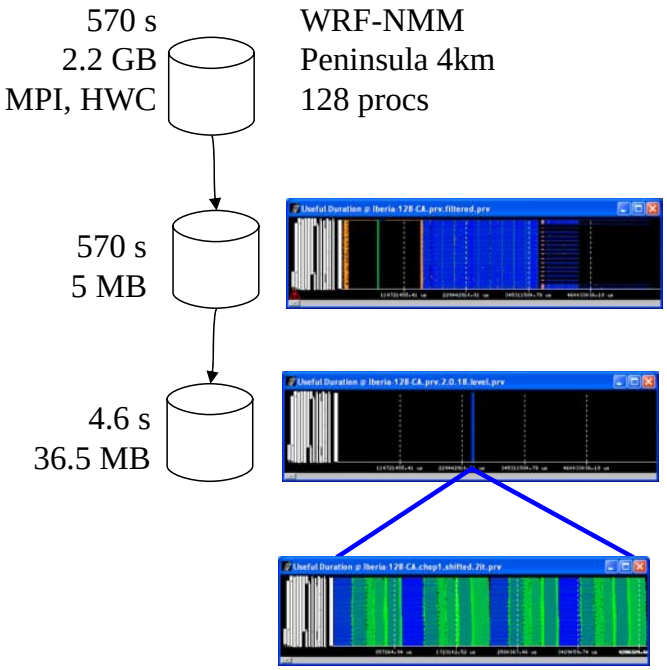


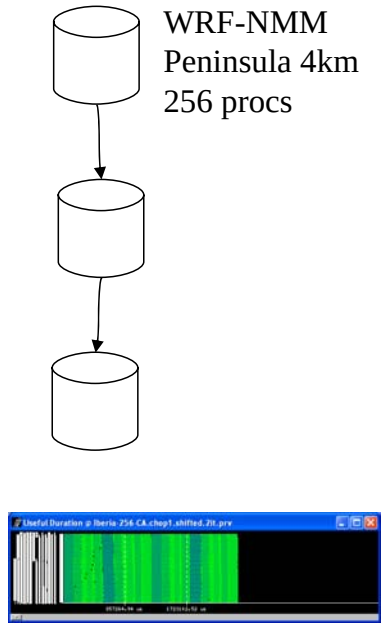
Autocorrelation

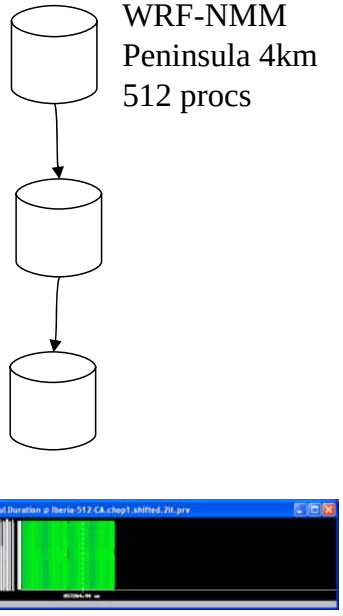


Models

Methodology: automatic speedup analysis

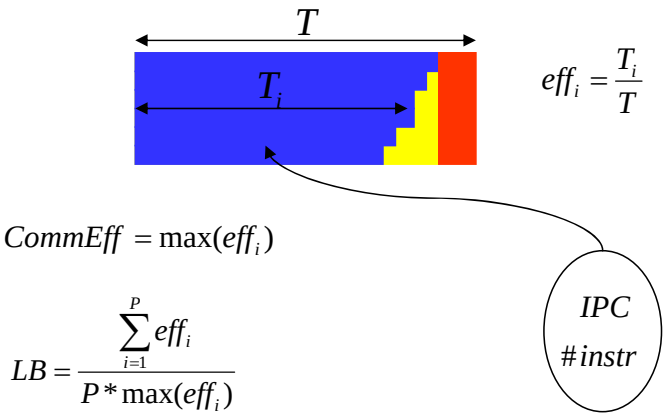






Speedup model

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

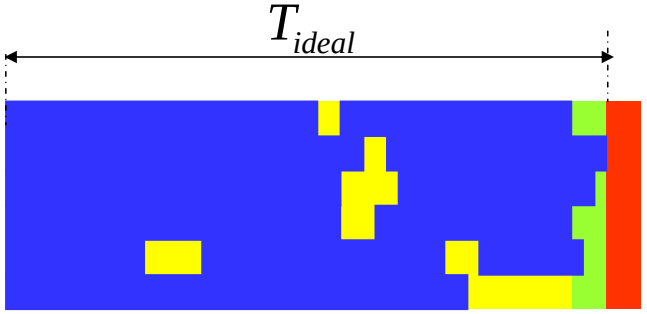


Directly from real execution metrics

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

Requires Dimemas simulation

$$microLB = \frac{\max(T_i)}{T_{ideal}} \qquad CommEff = \frac{T_{ideal}}{T}$$

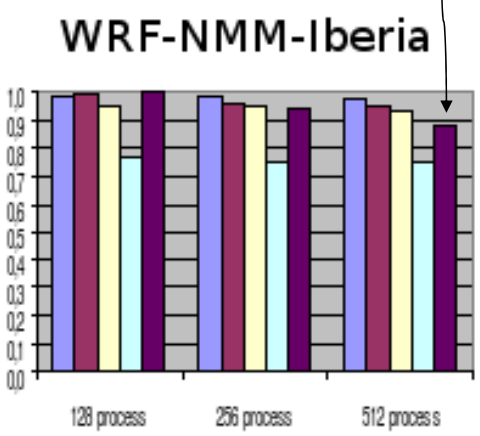


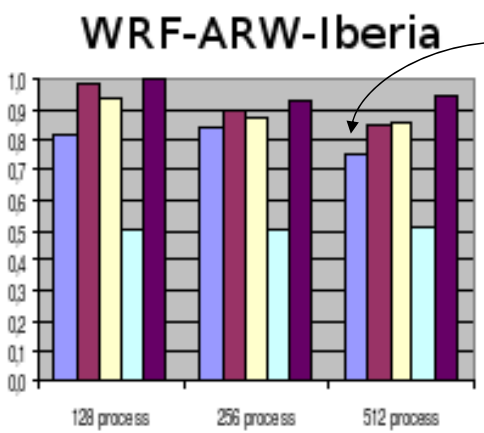
Migrating/local load imbalance
Serialization

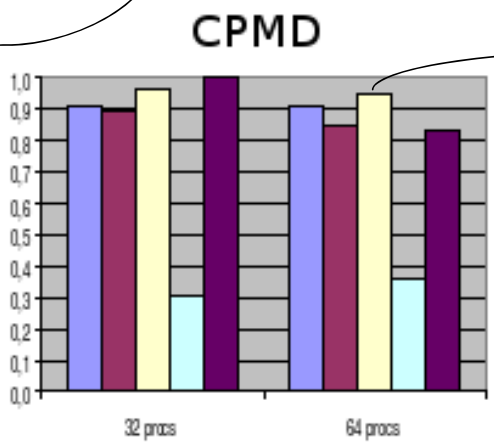
Speedup model

Replication of computation

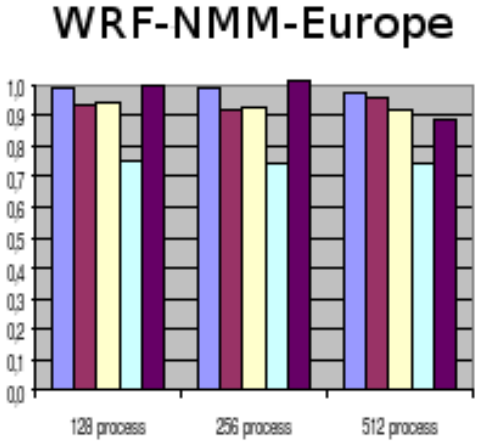
Poor communication efficiency

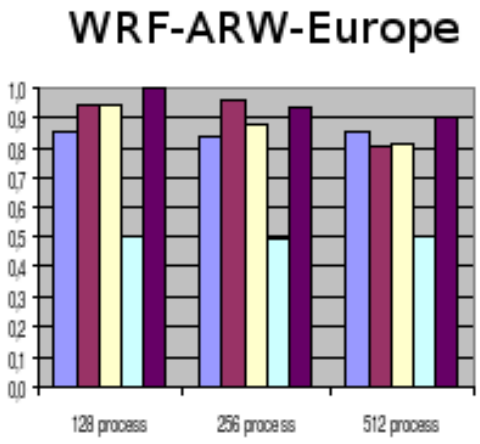


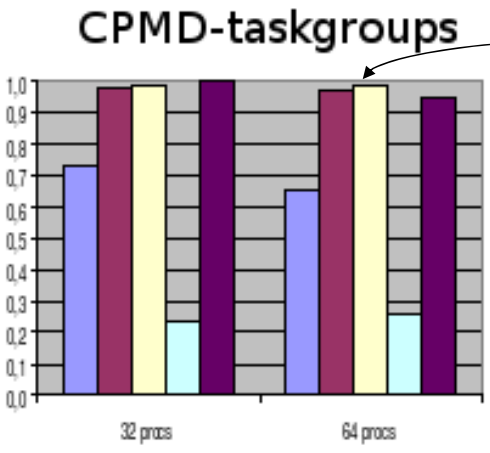




Improved macro and micro load balance







Communcication

Micro Load Bal.

Macro Load Bal

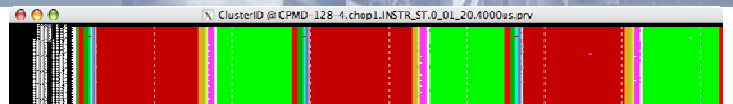
IPC

Computation

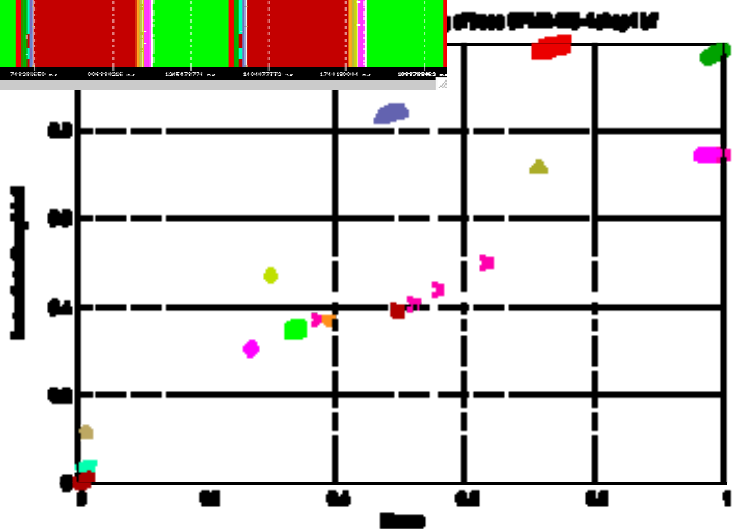
Clustering

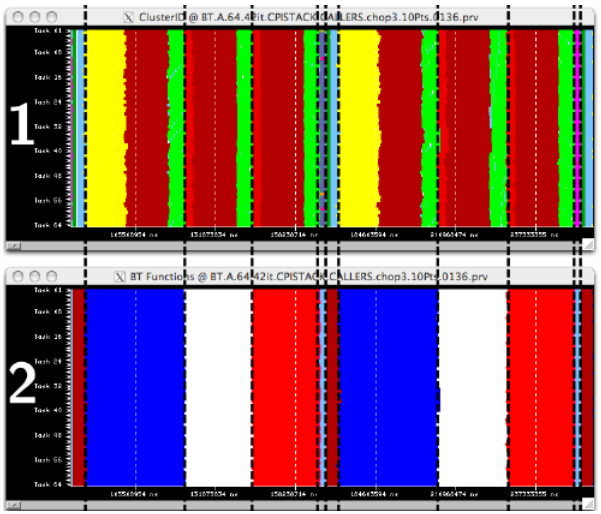
Clustering

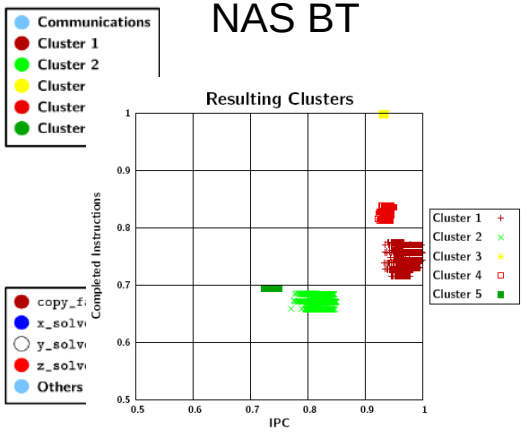
- Useful for
  - Identifying and highlighting structure
    - Cluster information injected in trace
    - Phases within routines
    - Different routines may have similar behavior
  - Compact trace encoding
  - Input to time analysis



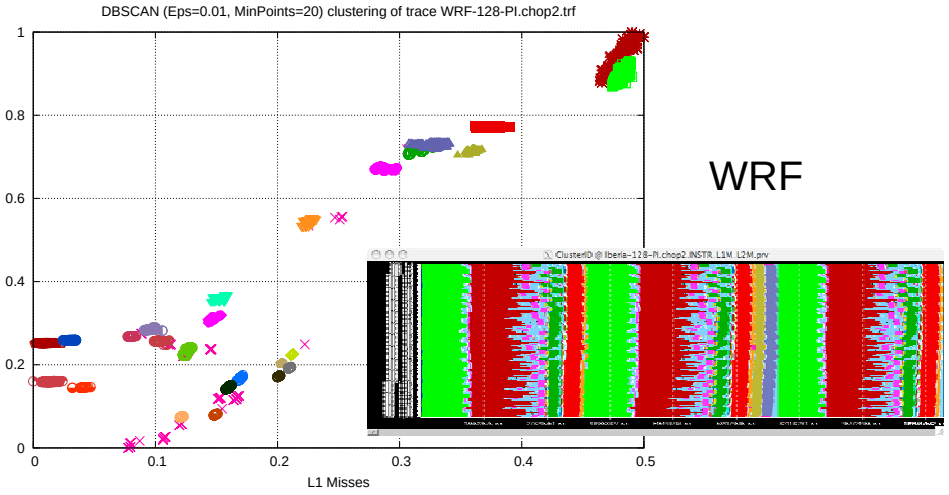
CPMD







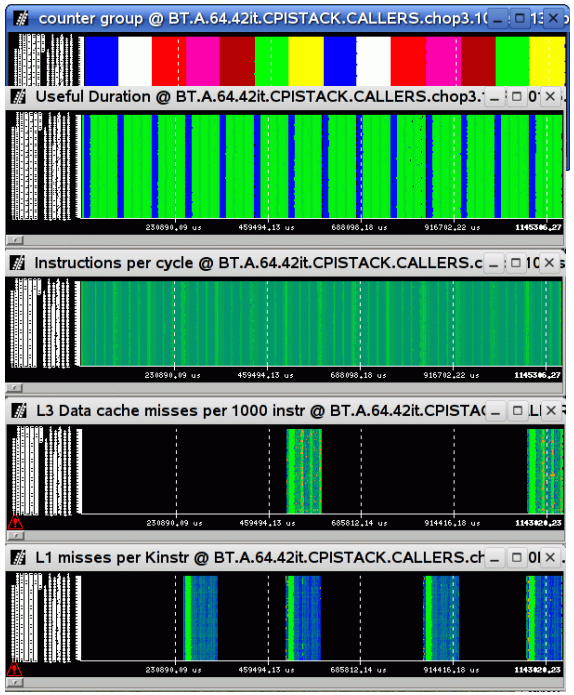
NAS BT



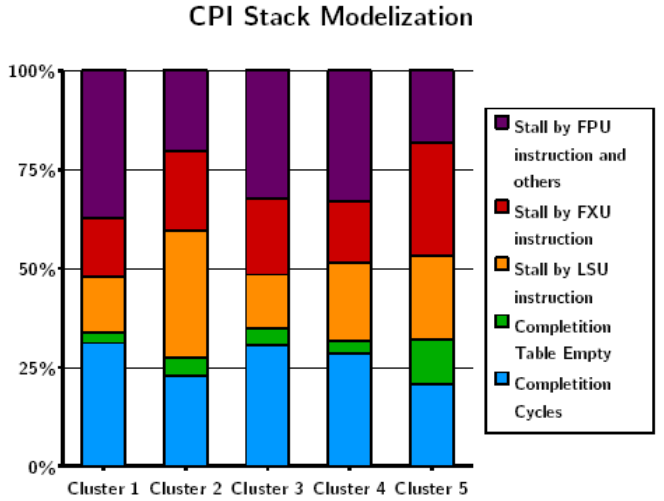
WRF

Clustering

- Useful for
  - Precise projection of hardware counters
    - Statistics
    - CPI stack model



Cluster	1	2	3	4	5
%Time	54.88	17.96	16.90	6.44	1.42
Avg. Burst Dur. (ms)	1.02	0.78	13.14	2.50	1.11
IPC	1.02	0.65	0.89	0.91	0.53
MIPS	2231.8	1423.3	1966.5	2001.8	1163.0
MFLOPS	339.2	46.3	191.6	269.2	23.6
L1M/KInstr	0.92	1.53	1.19	1.17	2.88
L2M/KInstr	0.06	1.26	0.06	0.35	0.21
Mem.BW (MB/s)	16.79	218.47	13.87	85.77	29.76



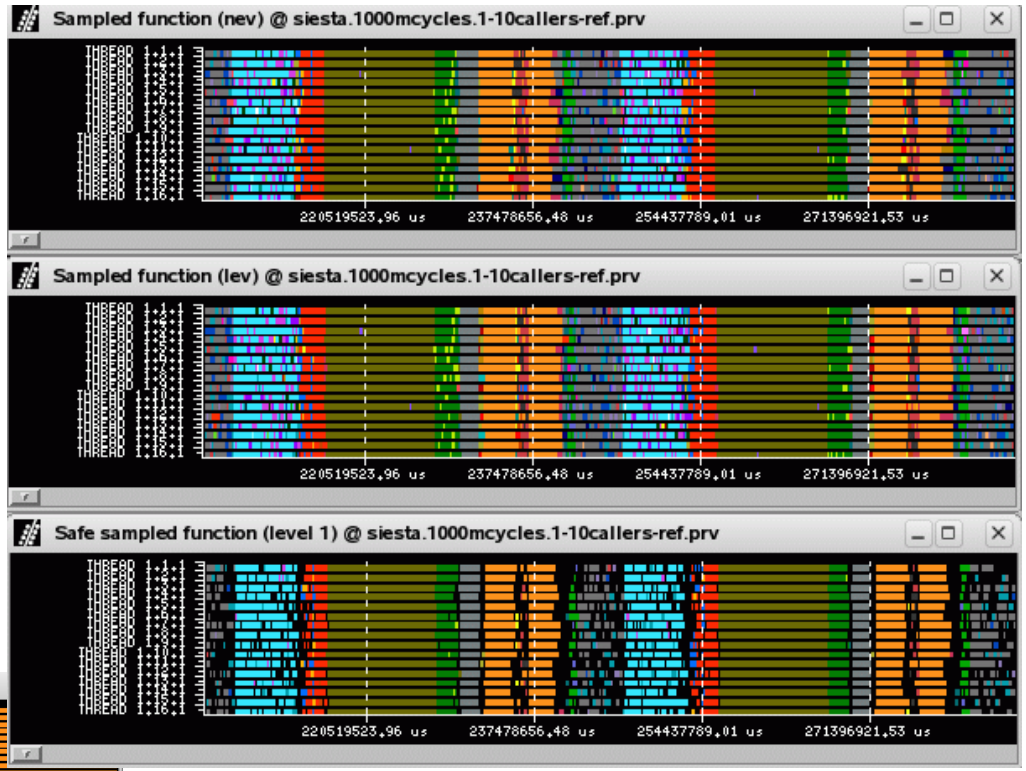
Samping + ….

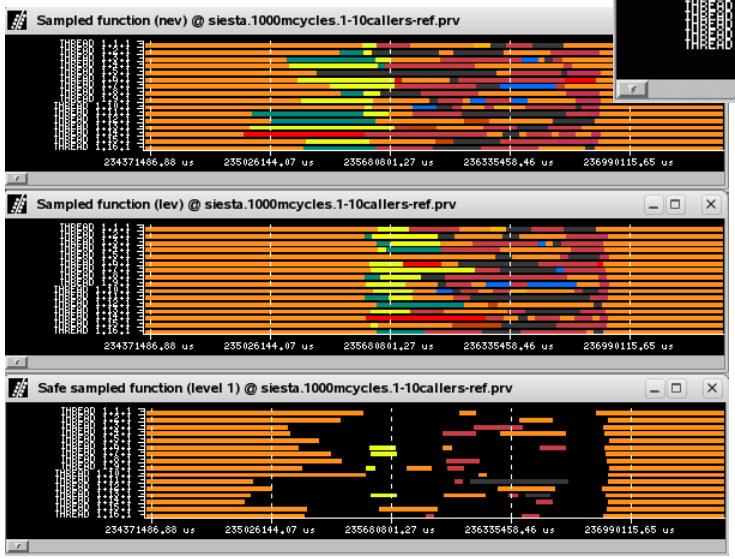
Sampling

- Tesis:
  - Has much more potential than currently used
  - Mixed with: instrumentation, frequency analysis, clustering
- Example questions
  - Can we get Information of time distribution (i.e. same length of all instances)?
  - Can we get very fine grain information with little overhead?
- The issue: How to project information captured by sampling
- Periodic Sampling (cycles counter)
  - High frequency (> Nyquist)
  - Low frequency (< Nyquist)
- “non periodic” sampling
  - Correlated to specific architectural/application hardware counters

High frequency periodic sampling

- Sampling frequency > Nyquist
- Function within interval?
  - Event at beginning
  - Event at end
  - “safe” when both are the same

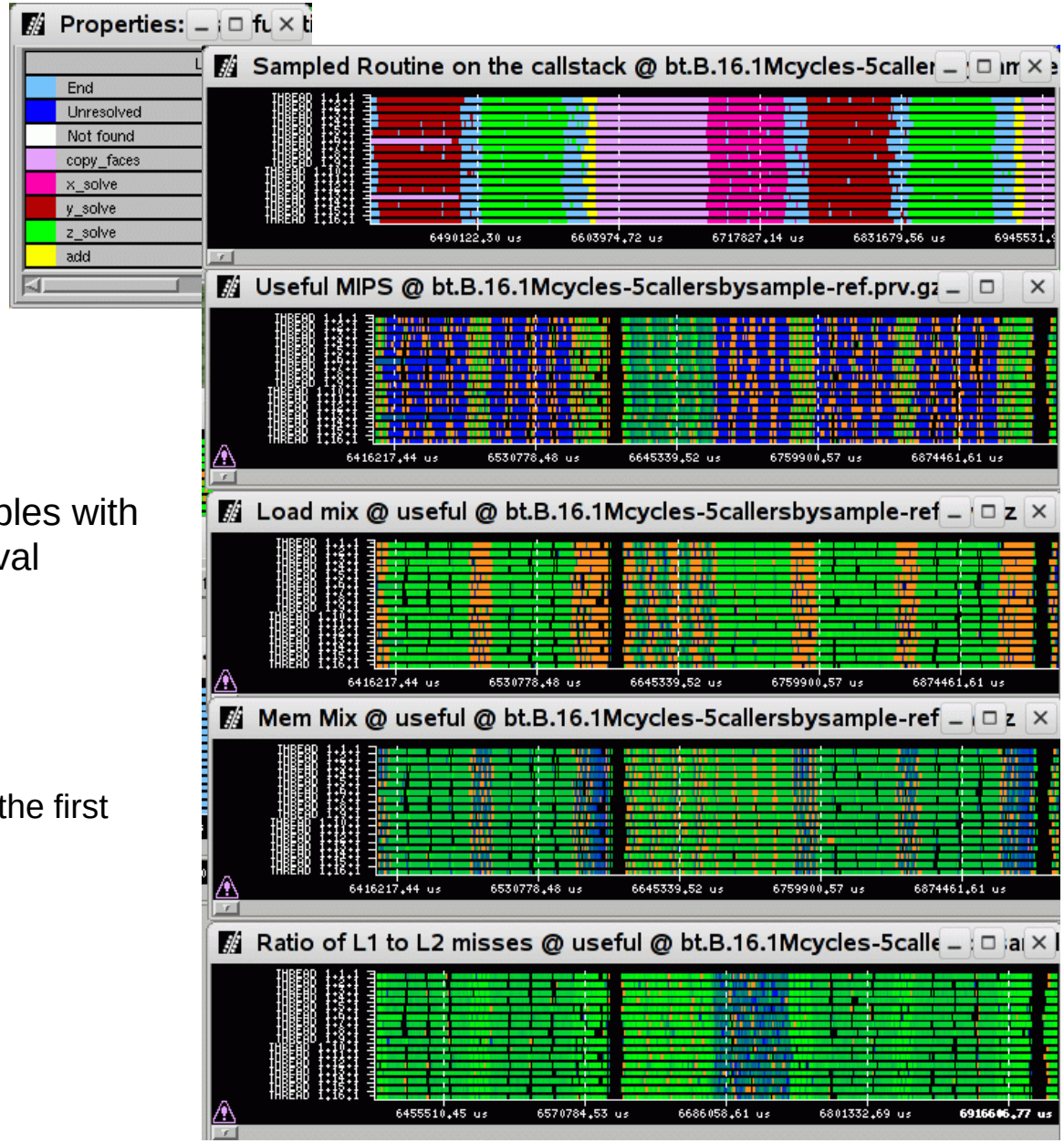




High frequency periodic sampling

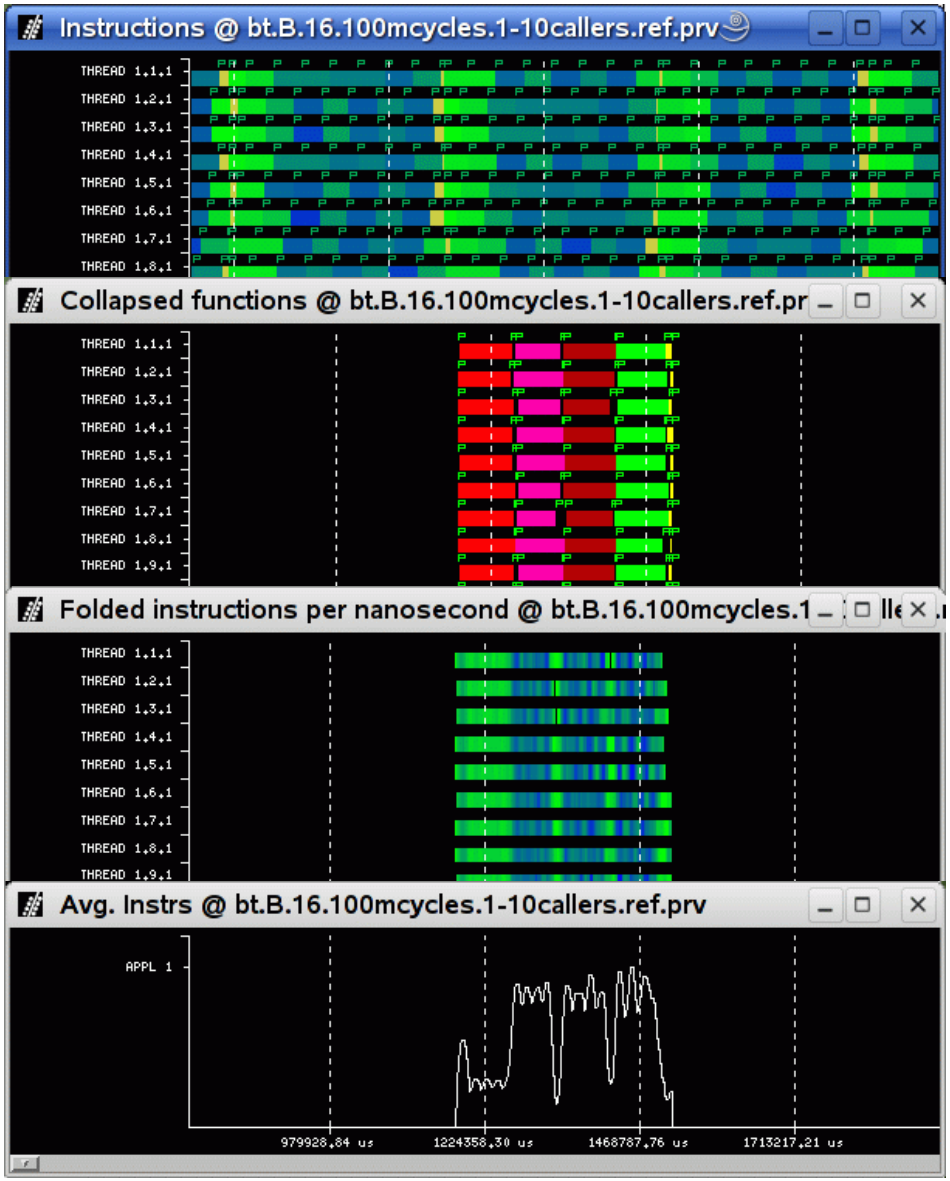
- Sampling frequency > Nyquist
- Shows fine structure

- Identification of function span
  - Assumption: consecutive samples with same call stack → whole interval assigned to function
  - All functions / specific subsets
    - Top of stack
    - Walk the stack searching for the first routine in the target set.



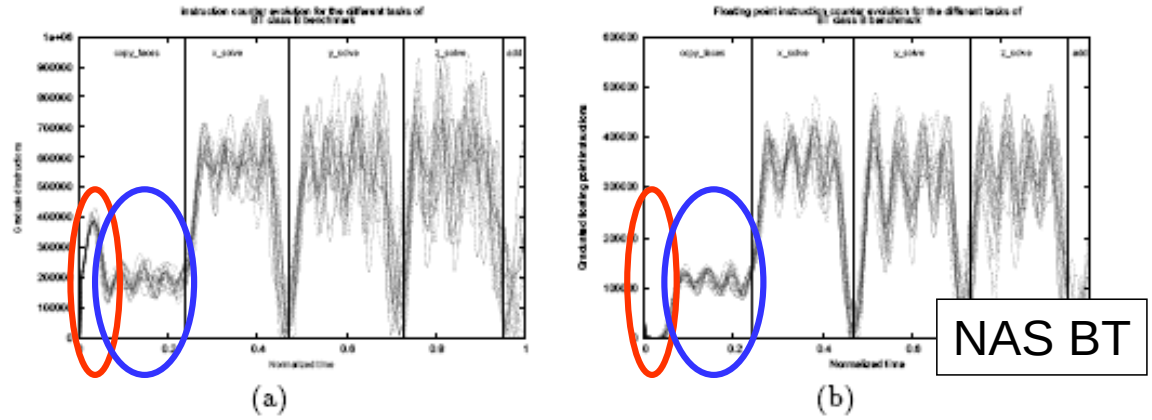
Low frequency periodic sampling

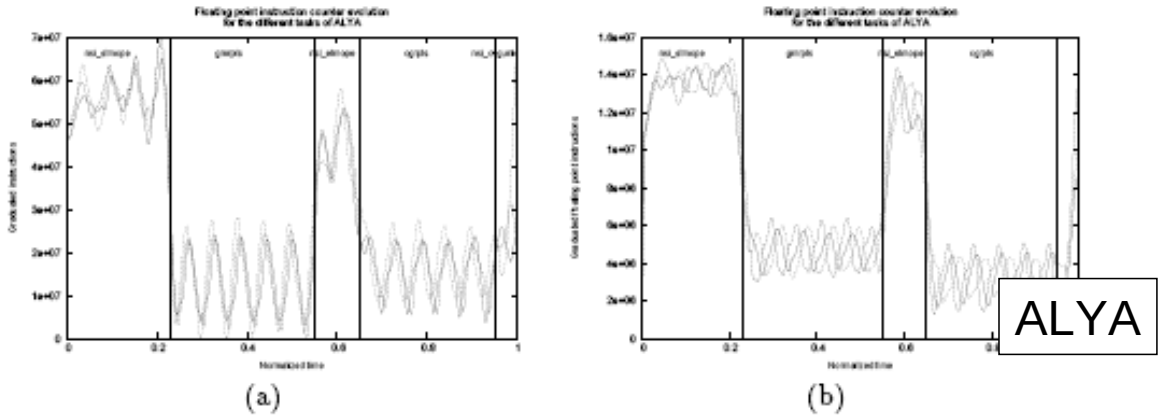
- How to increase precision?
- Folding based on known periodic structure of application (tagged iterations)
- Applies to stationary applications
- Result: trace for one iteration with synthetic paraver events
- Refer counts/timestamps to start of iteration
  - Call stack
    - Search for consecutive sequences of folded samples within same function and generate synthetic events
  - Hardware counters
    - Noise reduction
    - Fit folded samples
    - Sample fitting curve to generate synthetic events.

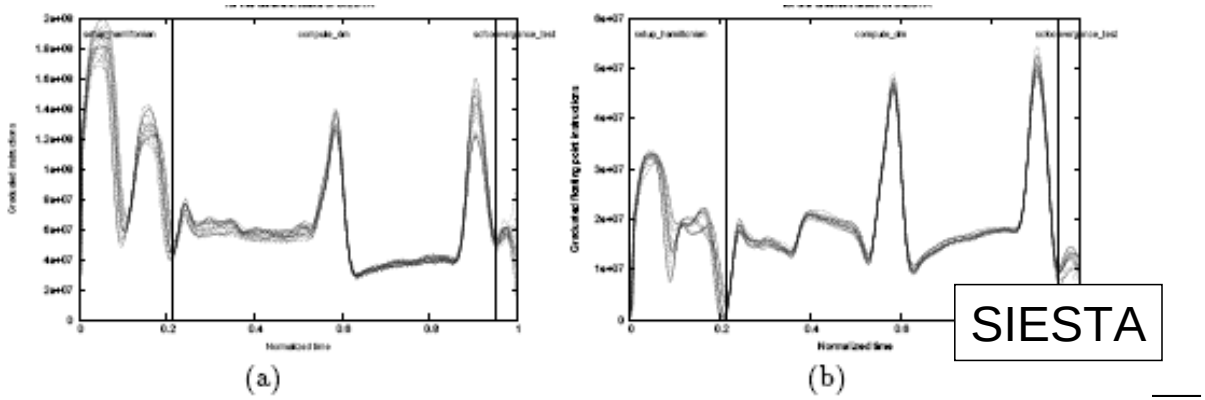


- Low pass filter …
- but still possible to identify loops

- Scalable summarization ...
- but still high level of detail

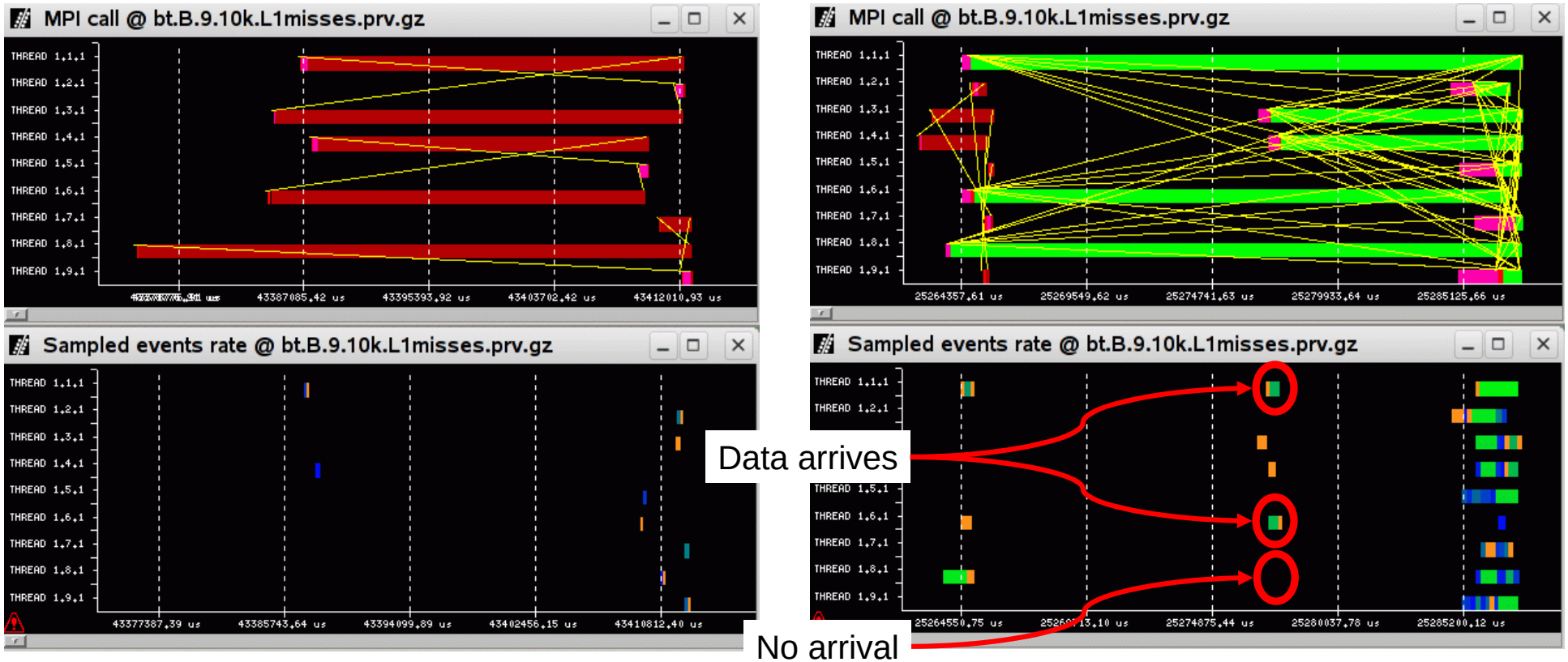






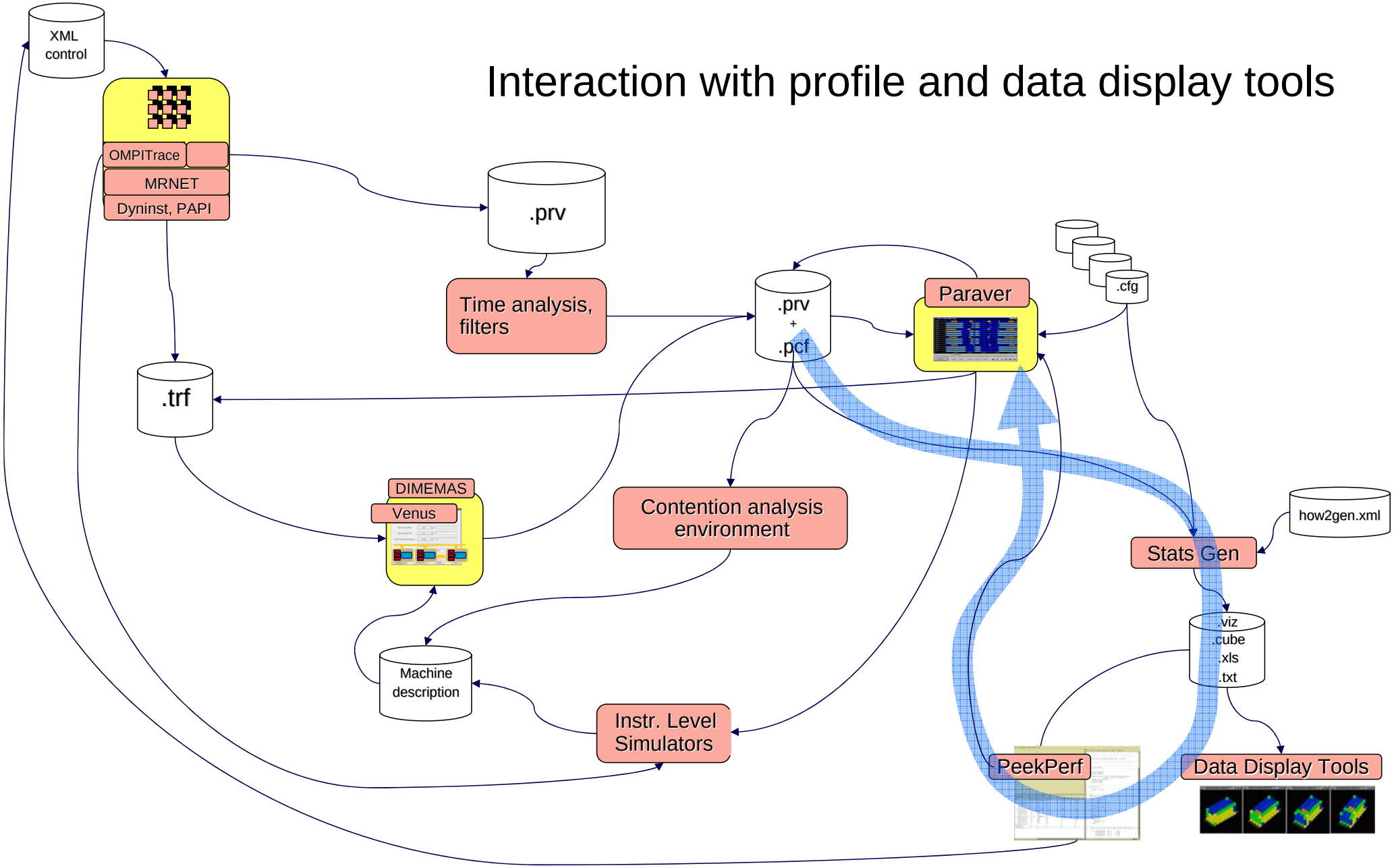
Non periodic sampling

- Correlated with specific architectural/program feature
- Useful to identify
  - Internal behavior
    - Density of L1 misses within MPI in an SMP: when data actually arrives.



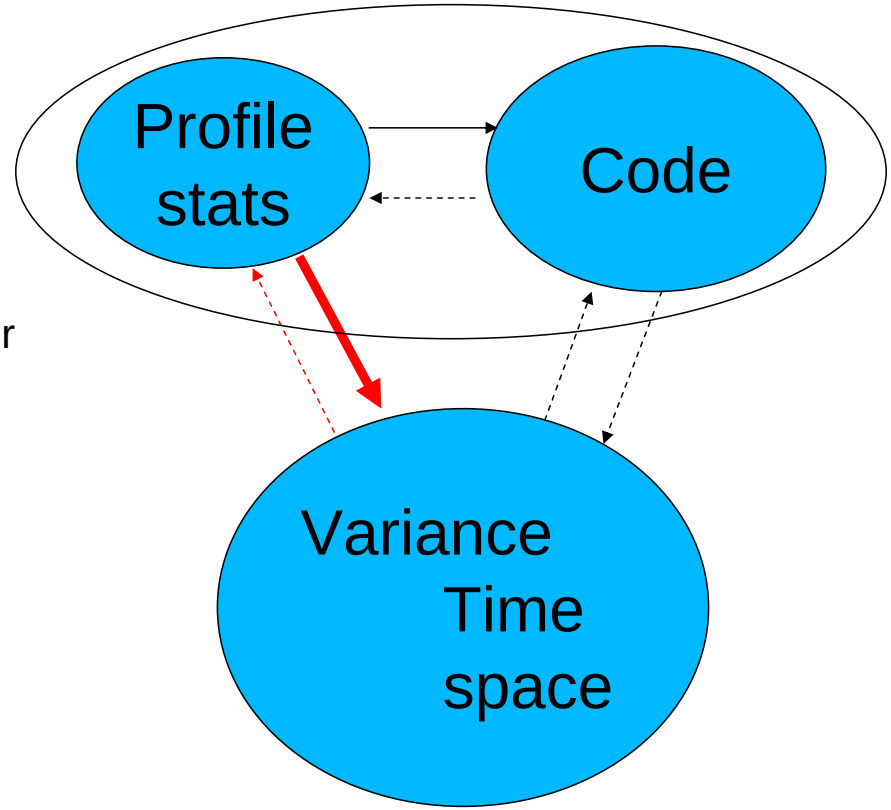
Profiles -- Timelines

CEPBA-Tools Environment



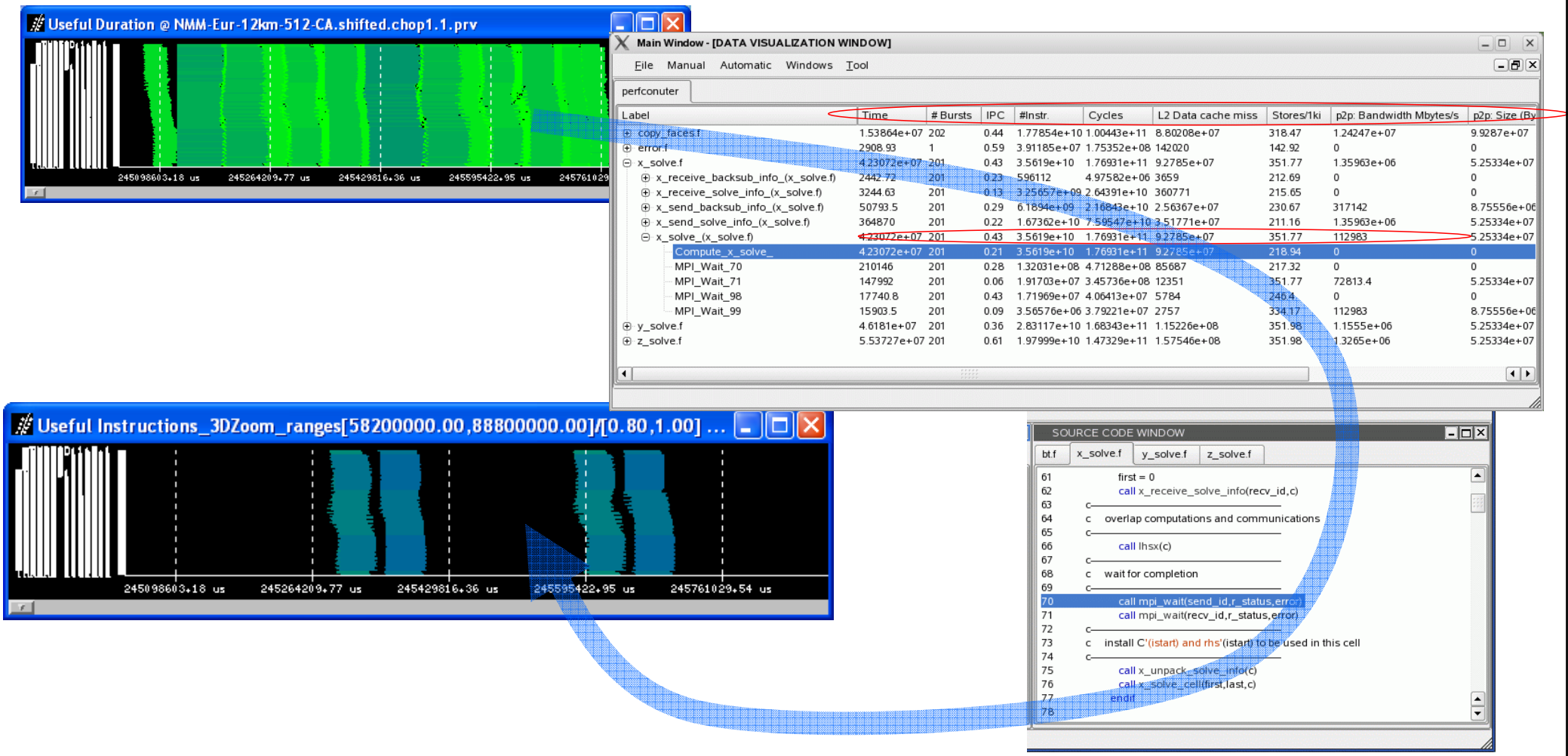
Integration with profile presentation tools

- Profile presentation tools
  - Reduction/aggregation of the performance dimensions
    - Time dimension disappeared / Space dimension sometimes
- Traces
  - “All” data is there
- Library
  - API to access statistics computed with Paramedir
- Translators
  - Configuration: Tool building environment
  - Prototypes
    - Peekperf prototype, Gprof like, …
  - Starting
    - Paraprof, CUBE



Integration Paraver- Peekperf

- Statistics generator
  - Built a set of C++ libraries to access **paramedir** stats.



Tracing???

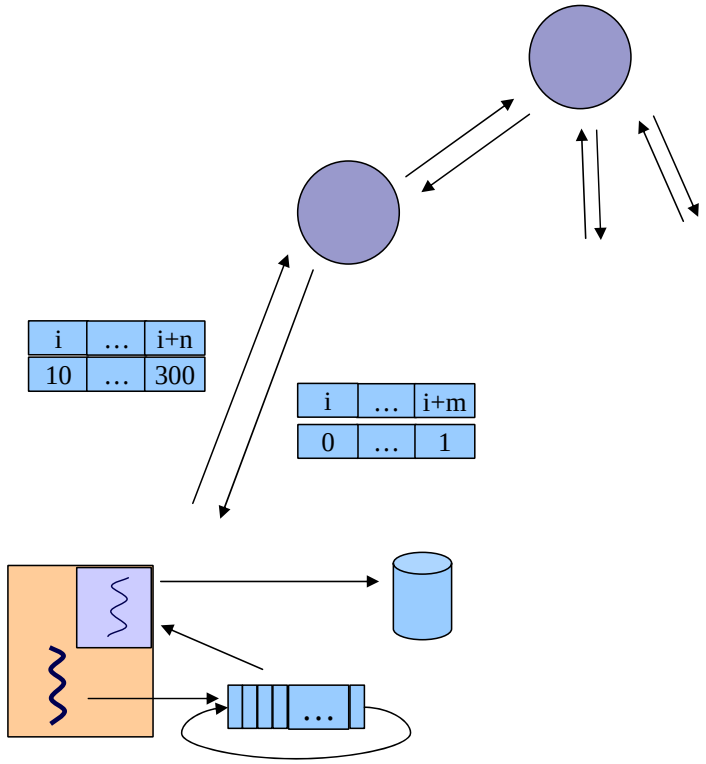
Is it all about traces?

- Traces
  - See - colors
  - Scalable view of metrics and variance
  - Detailed analysis
  - Experience - Understanding - Insight
  - Ideas

- Then
  - Put the **intelligence** in online environments
  - Use scalable infrastructures

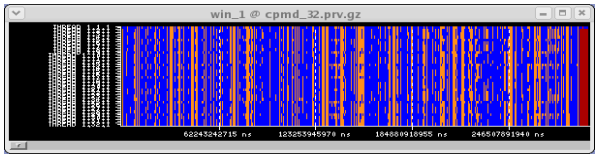
Distributed trace control

- 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

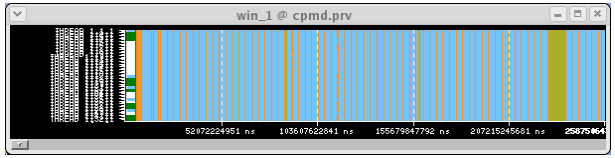


Online analysis

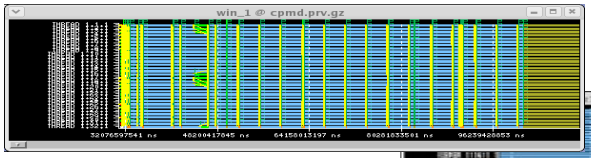
- Collective duration threshold



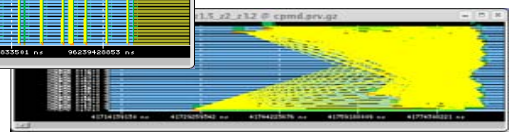
245MB, >15500 col



<1MB, <85 col

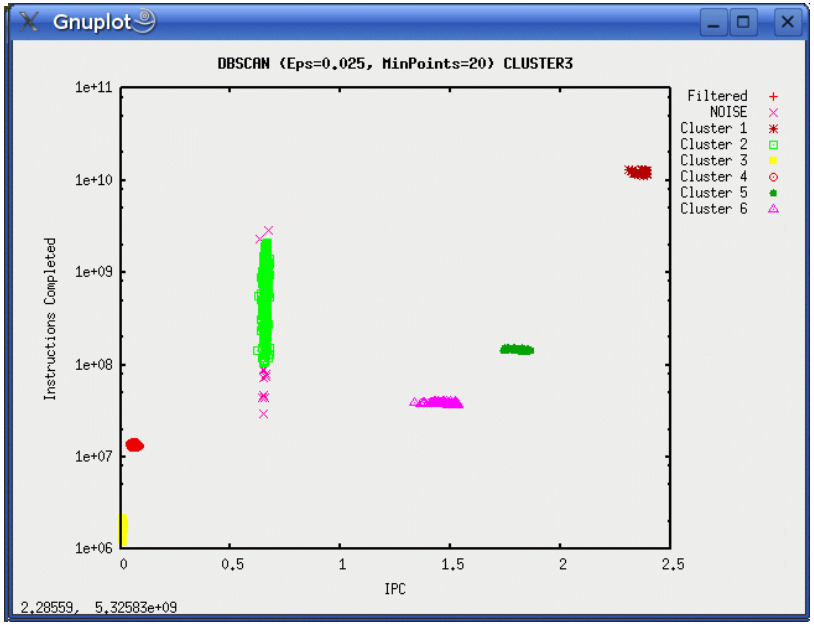


25MB, <85 col



Collective internals

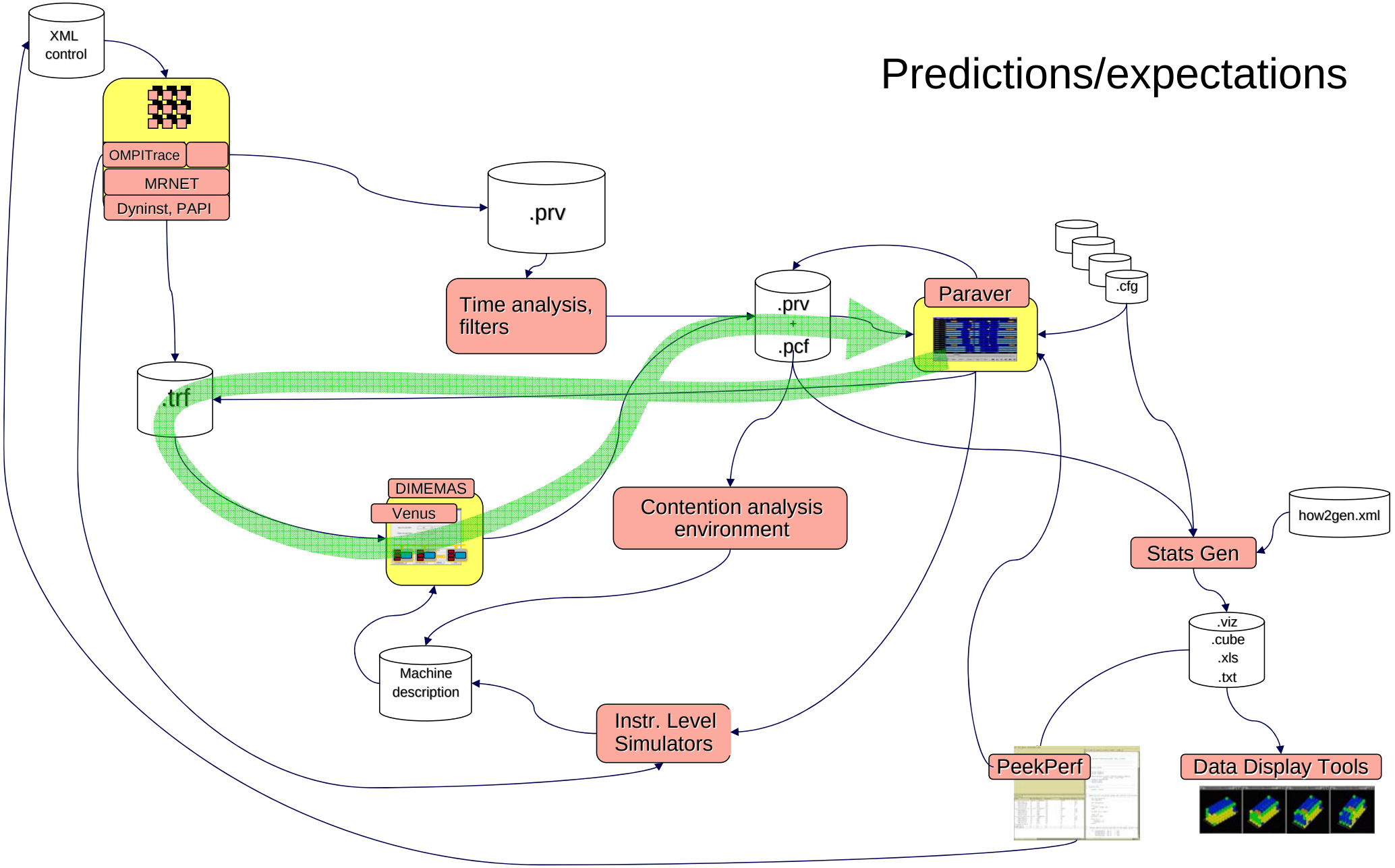
- Periodic clustering snapshots
- Periodic frequency analysis



- Maximize information/data ratio
  - Direct report metrics, ….
  - Activate tracing (user/automatic)
  - Focused traces

Dimemas

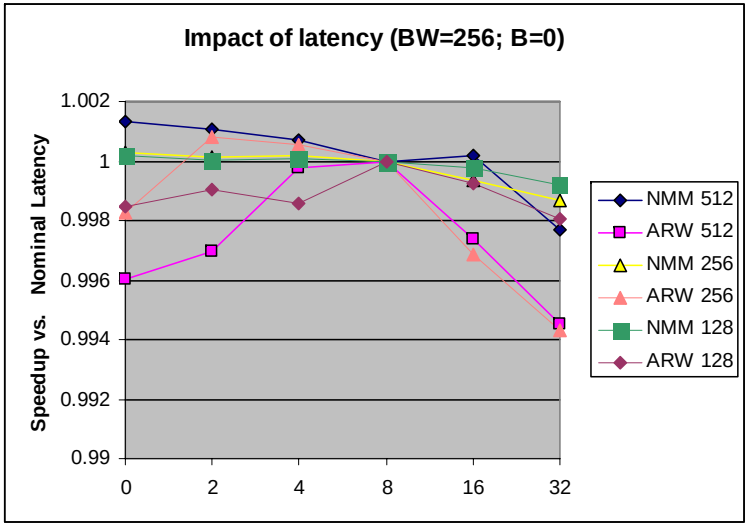
CEPBA-Tools Environment

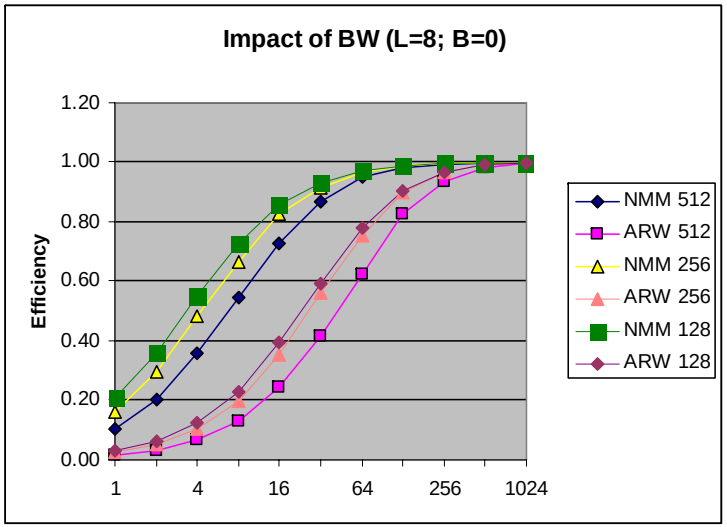


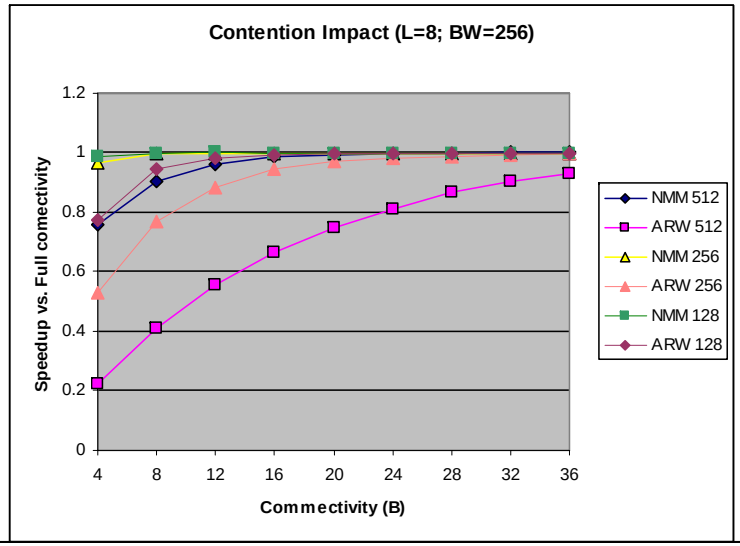
Predictions/expectations

Network sensitivity

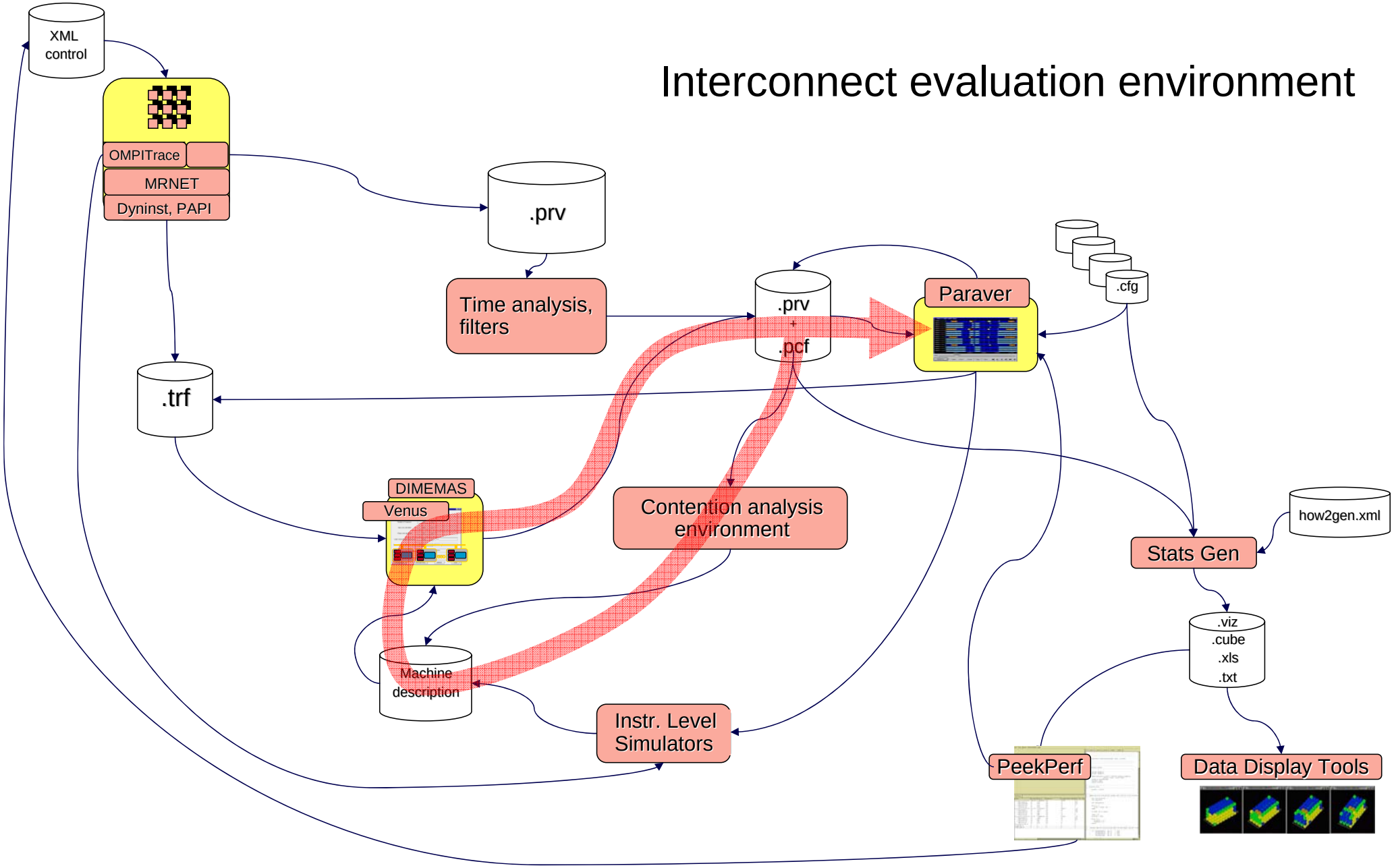
- Simulations with 4 processes per node
- NMM Iberia 4Km
  - Not sensitive to Latency
  - 512 sensitive to contention?
  - 256 MB/s OK
- ARW Iberia 4 Km
  - Not sensitive to Latency
  - sensitive to contention
  - Need 1GB/s







CEPBA-Tools Environment

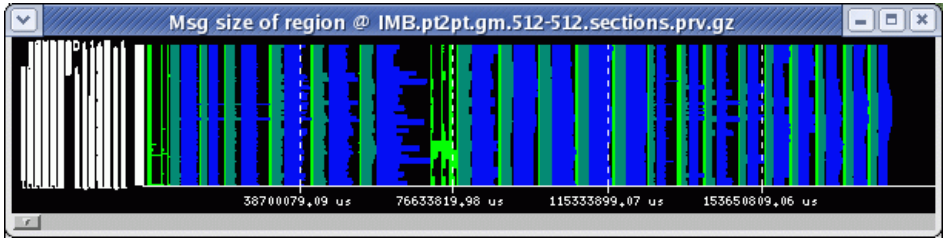


Contention impact

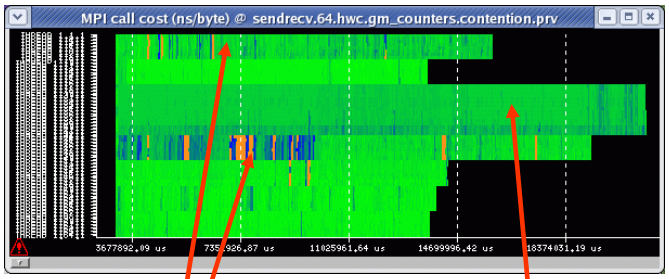
- on large systems?
- In multiuser environments

512 nodes, 4MB

Dependence on appl. phase (comm. Pattern)

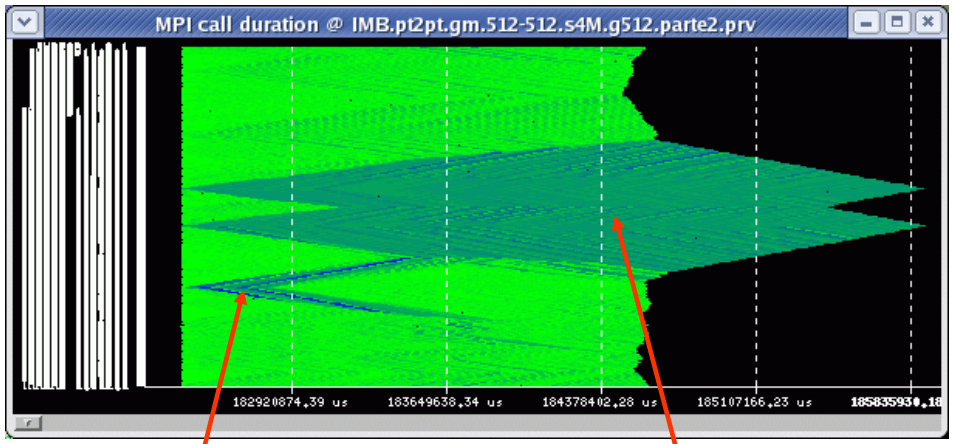


64 nodes, G=8, 4MB



External contention

Internal contention



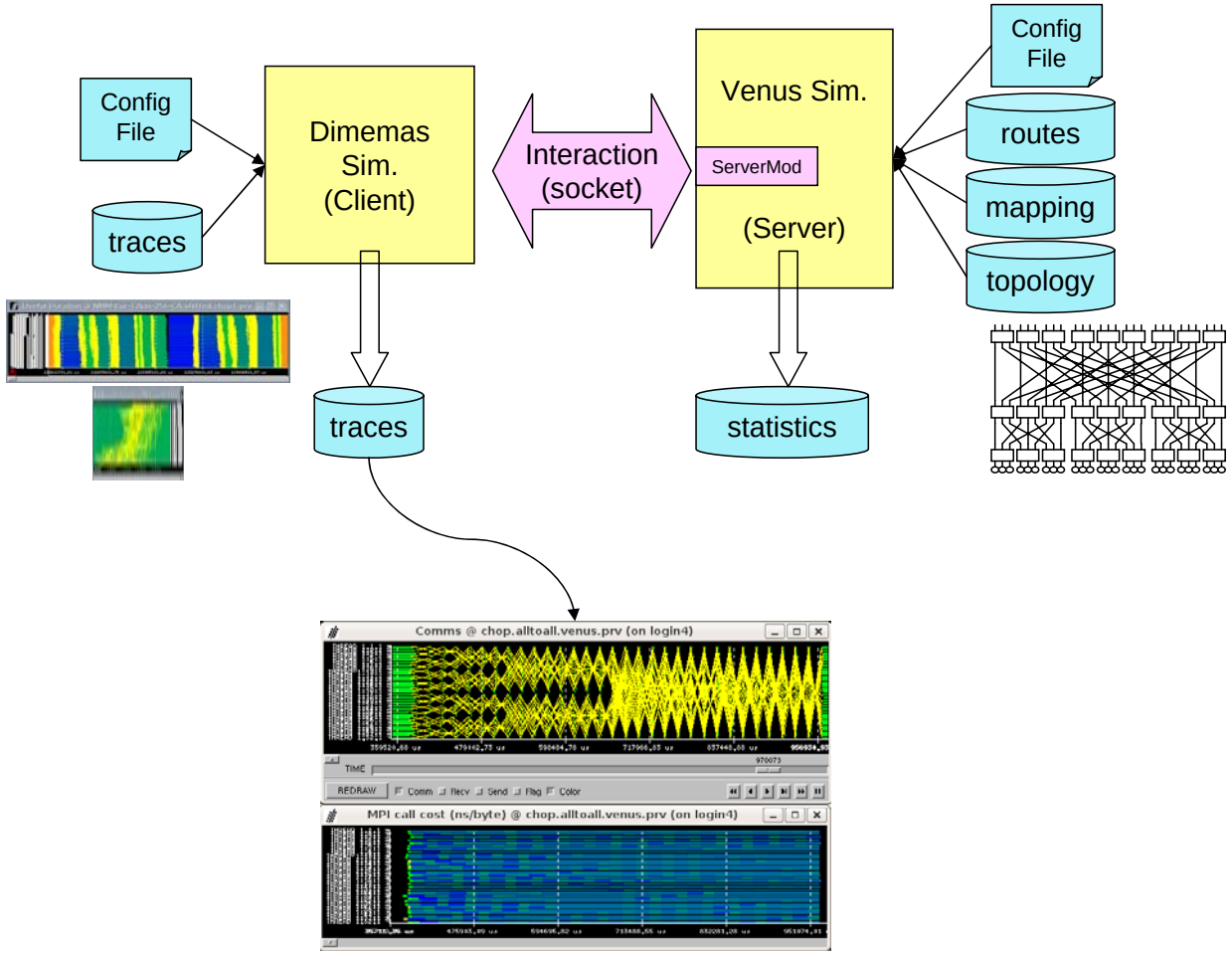
Bubble propagation

Propagation of internal contention

What is the benchmark measuring?
Appropriate number of iterations?

Interconnect simulation environment

- Dimemas
  - Very fast
  - Sensitivity: identify coarse grain factors
  - identify relevant communication phases
- Venus
  - Very detailed network model
- Venus-Dimemas integration
  - Understand application usage of physical comm. resources
  - Communication subsystem design



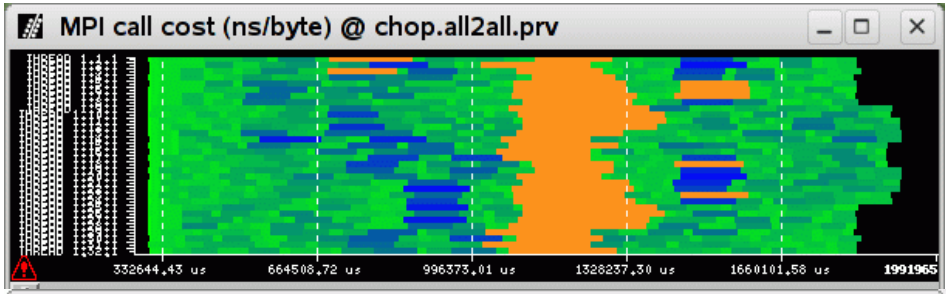
Contention

- P2p application trace chop
- @ collectives

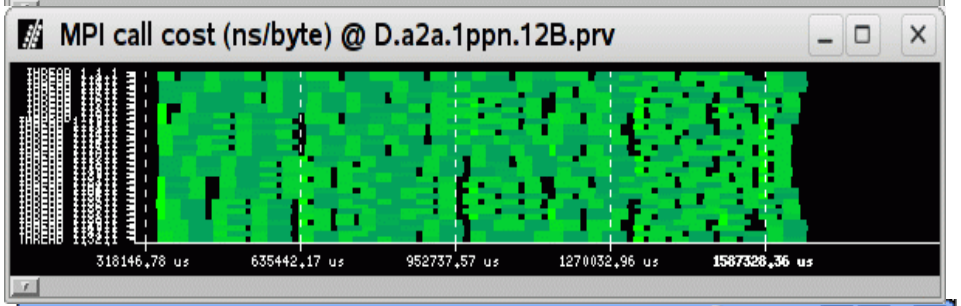
- Evaluating
  - Routing
    - Oblivious
    - Pattern aware
    - Dynamic
  - Topology
    - Slimmed trees
    - Direct networks
    - Process mapping
  - Protocol
    - Eager limit
    - Packet segments

MPI call cost in ns/byte

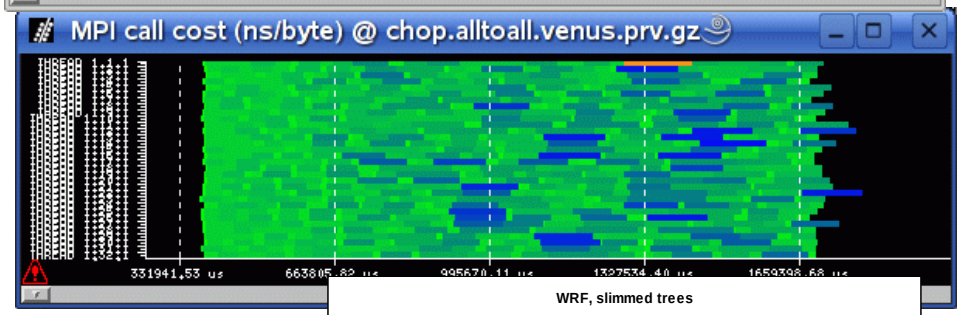
Actual run

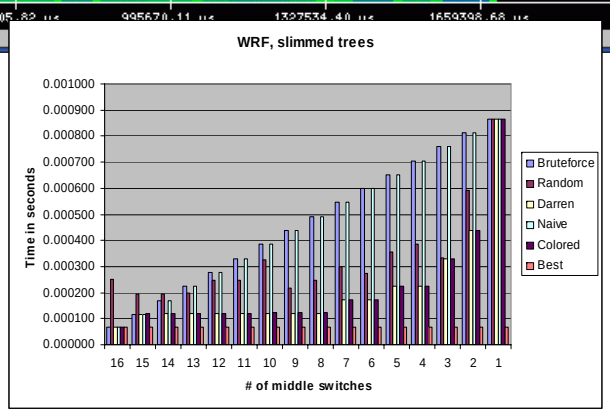


Dimemas

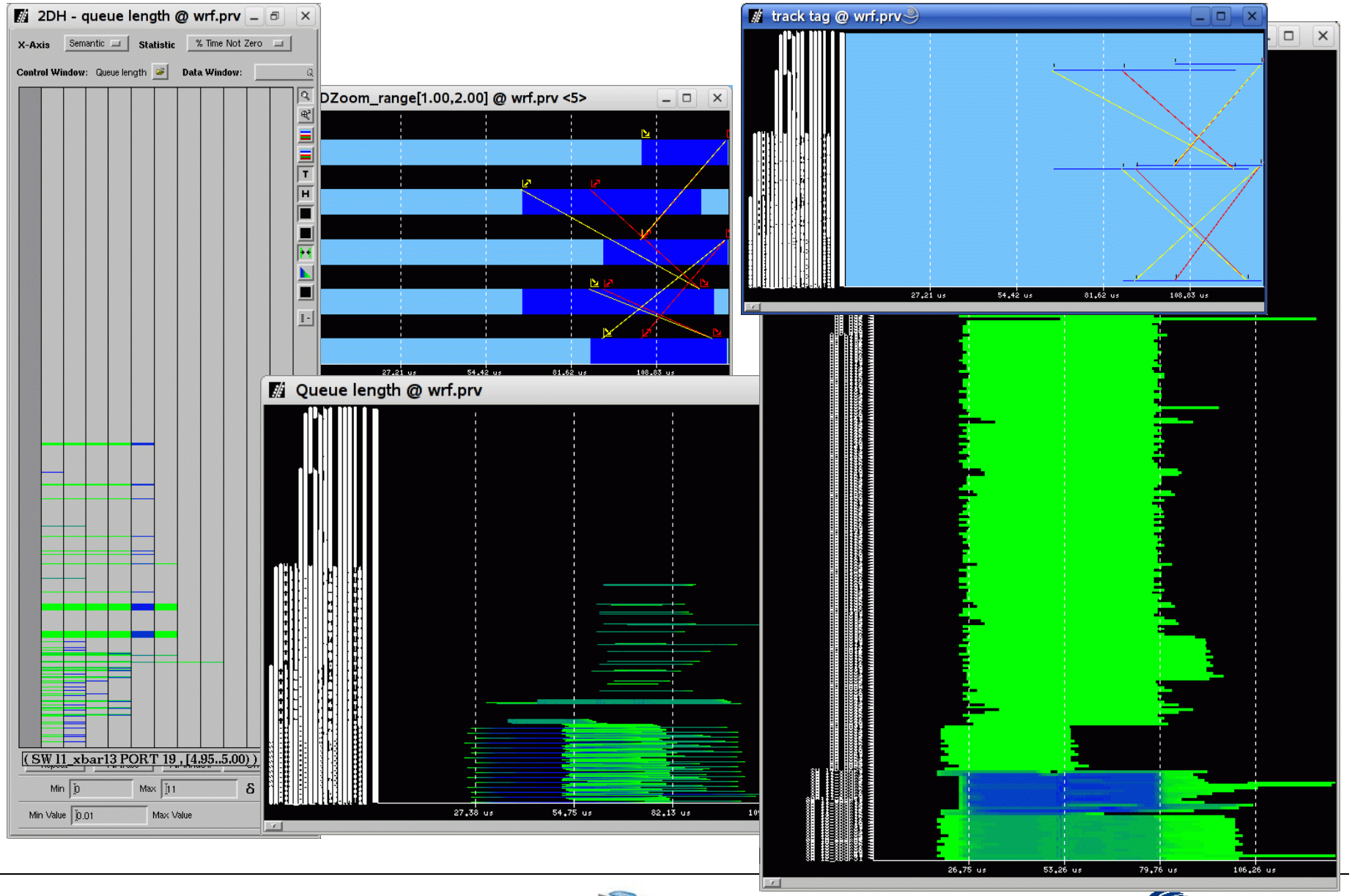


D&V



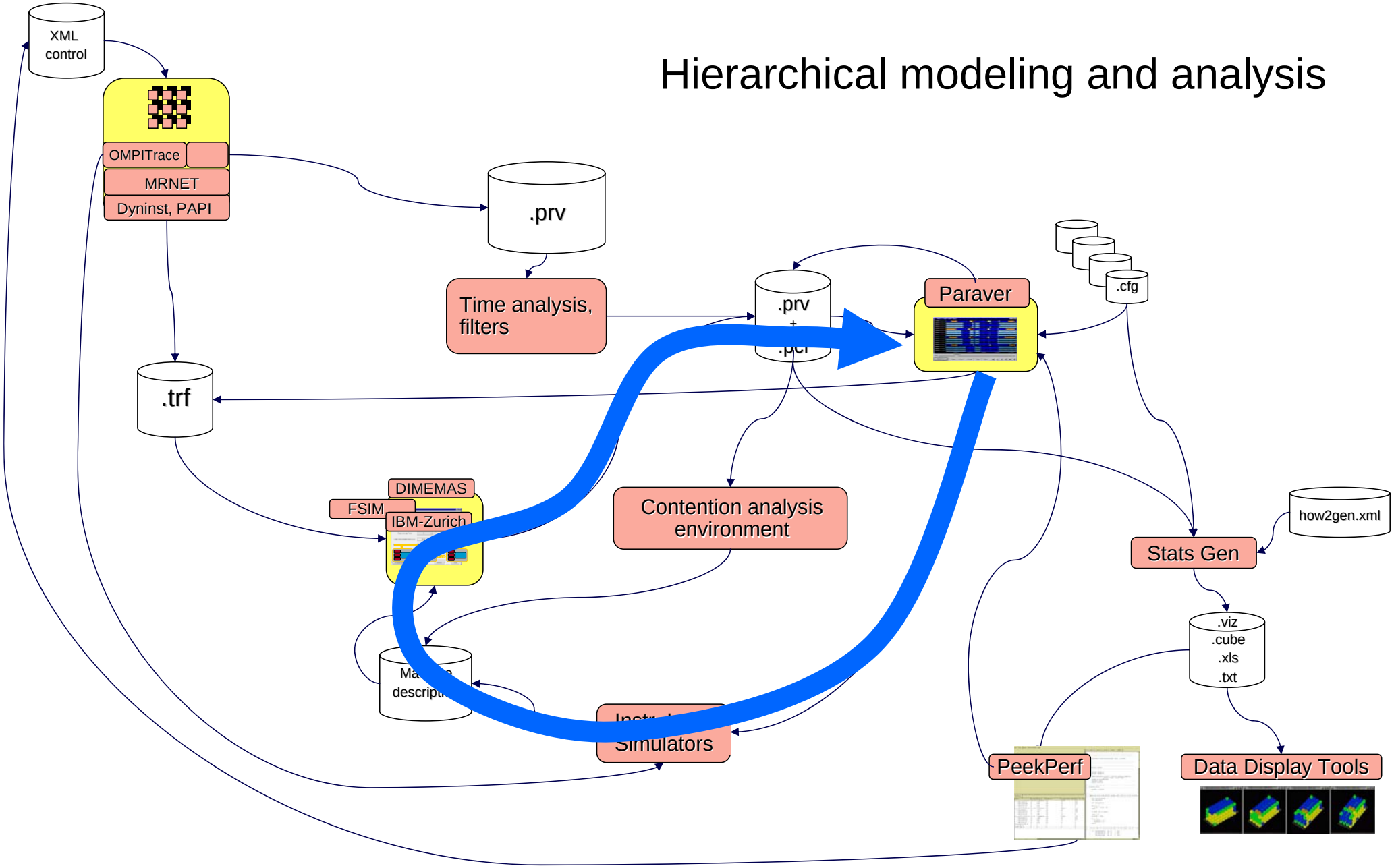


Internals network

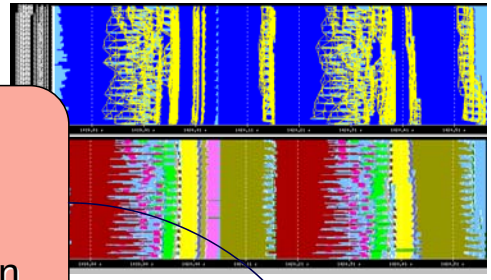


Hierarchical modeling

CEPBA-Tools Environment



Hierarchical modeling and analysis



MPITrace

.prv
+
.pcf

Clustering
Identification
Trf generation

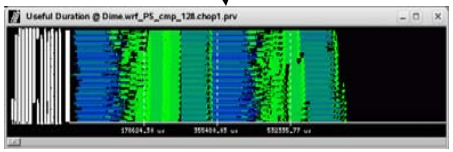
MPI Task	ClusterID	MPI Call	count	Instructions
T7	2	MPI_Allreduce	4	115.000.000
T7	1	MPI_Wait	9	190.000.000
T7	5	MPI_Wait	3	21.200.000
T7	4	MPI_Wait	6	29.500.000
T7	3	MPI_Wait	15	41.900.000

.trf

DIMEMAS
Venus

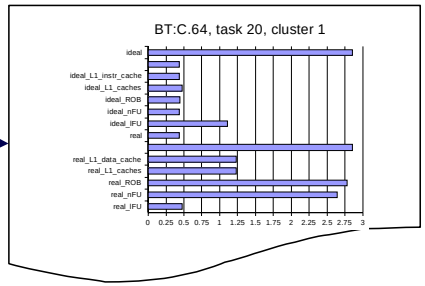
Valgrind

Machine description



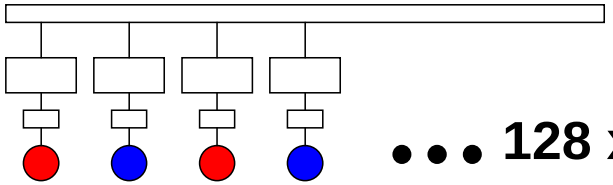
MPSim

.fF



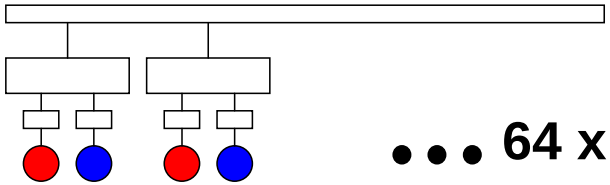
Turandot simulation results for WRF (table)

Single Core / Single Thread



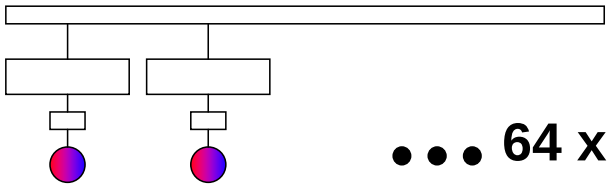
• • • 128 x 1 CPU

Dual Core / Single Thread



• • • 64 x 2 CPU

Single Core / Dual Thread



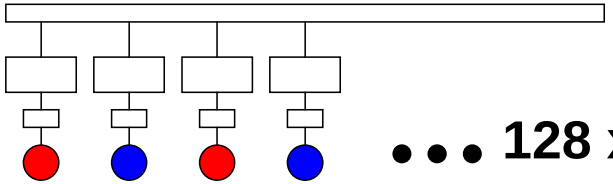
• • • 64 x 1 CPU x 2 th

	IPC			
	Measured	Simulated P5 (TURANDOT)		
Cluster	MN	ST	CMP	SMT
1	0,54	0,75	0,75	0,56
2	0,49	0,72	0,72	0,52
3	0,62	0,83	0,72	0,43
4	0,72	0,94	0,83	0,50
5	0,79	1,07	0,84	0,52

	CPU ratio			
	Measured	Simulated P5 (TURANDOT)		
Cluster	MN	ST	CMP	SMT
1	1	1,38	1,38	1,04
2	1	1,47	1,47	1,05
3	1	1,34	1,16	0,70
4	1	1,30	1,15	0,69
5	1	1,36	1,06	0,66

Dimemas simulation results for WRF

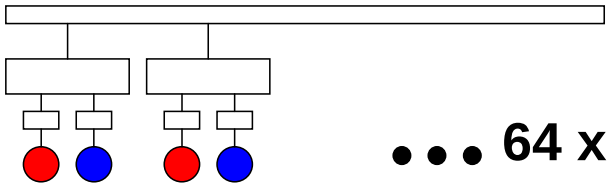
Baseline: ppc970 FX



• • • 128 x 1 CPU

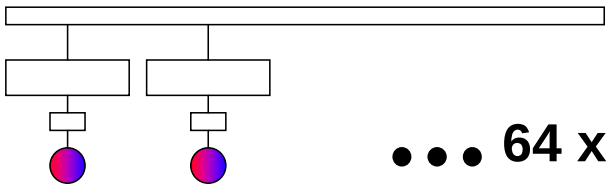
Replace 970 by Power5

Replace 970 by Power5 CMP

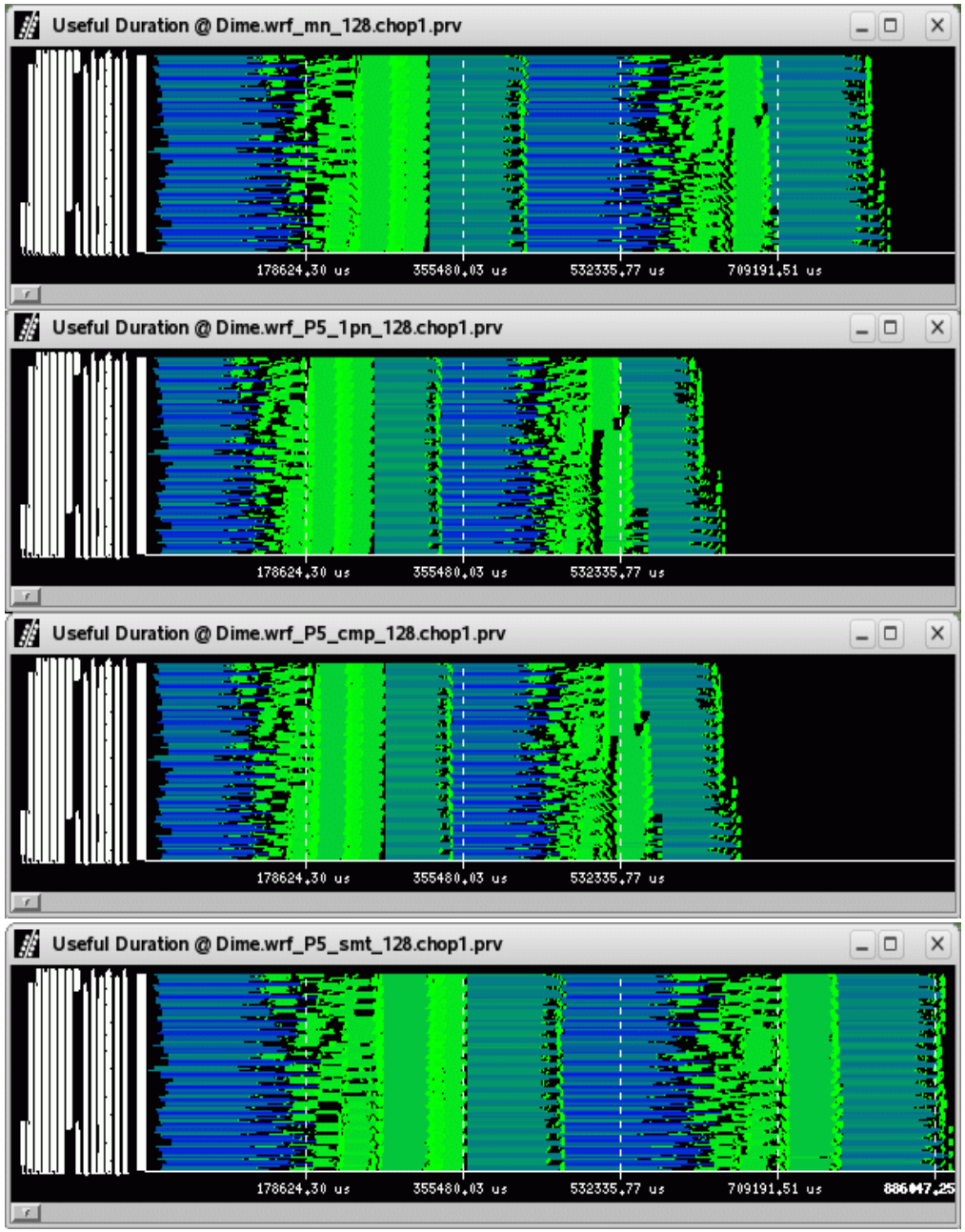


• • • 64 x 2 CPU

Replace 970 by Power5 SMT



• • • 64 x 1 CPU x 2 th



Offers and interests

Offer

• Paraver	OPEN SOURCE by end of the year
• Dimemas.	OPEN SOURCE by end of the year
• Instrumentation.	OPEN SOURCE by end of the year
• Structure analysis tools	in development
◦ Signal analysis	
◦ Clustering	
◦ Sampling + tracing	
• OpenMP incl 3.0 tasks	OPEN SOURCE
• CellSs /SMPSs	OPEN SOURCE

Interest

- HWC
  - PAPI
  - CPIstack models from hardware counters → manufacturers
- Control infrastructure
  - MRNet, LaunchMON
- Instrumentation infrastructure
  - Dyninst, P^{N}MPI
- Profile display tools
  - File formats
  - Link to timeline mechanism
- Instrumentation control. Profiler based control
- Probe injection mechanisms
- Call stack API.
- Compiler information / model on application
- Simple code restructuring