



Analysis of AIX traces with Paraver

Judit Gimenez, Jesus Labarta (CEPBA-UPC)

Terry Jones (LLNL)

Technology Transfer

User Support

Research

Education

Training

HPC Facilities

Mobility of Researchers

Parallel Expertise

Index

- Motivation
- AIXtrace2paraver
- Some Examples
- Conclusions



Motivation

■ AIX Trace @ LLNL

- Very detailed information - good!
- Generate tons of ASCII reports - not so good!
 - ✓ Scripts to extract some info
 - ✓ Lot of details “lost”

■ Paraver

- High potential of analysis
 - ✓ qualitative and quantitative
 - ✓ detailed analysis
- no semantics neither on the tool, nor on the trace format

■ Objective

- Analyze with Paraver the information captured with AIX Trace



Index

- **Motivation**
- **AIXtrace2paraver**
 - Approach
 - Information emitted
 - Other features
- **Some Examples**
- **Conclusions**



Approach: step 1 - AIXtracelauncher

■ This step is **OPTIONAL**

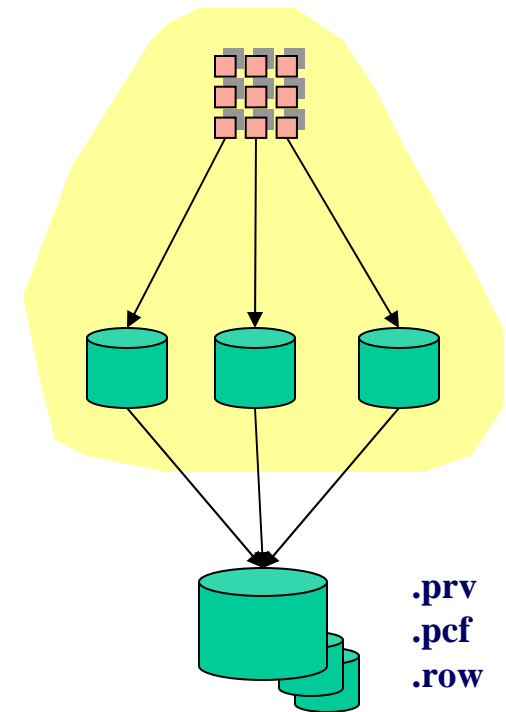
- Not required by the translator

■ Binary starting the **AIX Trace Facility**

- To simplify the launch of the tool
- To read the AIX events that we translate

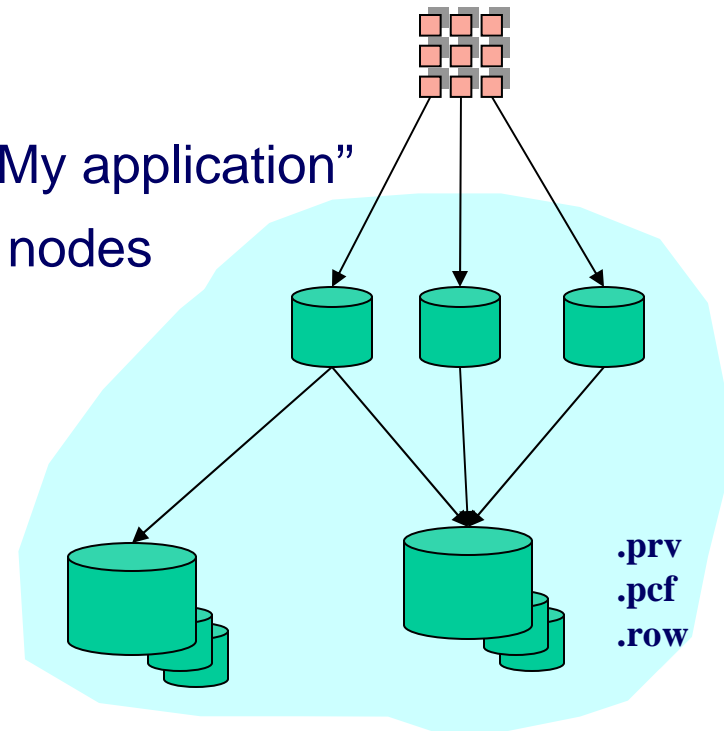
■ **Three modes:**

- Trace node during n seconds
- Trace node during the execution of an application
- Sample mode: trace intervals



Approach: step 2 - AIXtrace2prv

- **Translator from AIXtrace binary format to Paraver format.**
- **Emit to the .prv trace:**
 - All processes in node
 - Only selected processes from node
 - All processes, mark selected ones as “My application”
 - Only selected processes from different nodes
- **Other options**
 - User events mapping
 - Software counters
 - Print list of processes



Information emitted to the Paraver trace

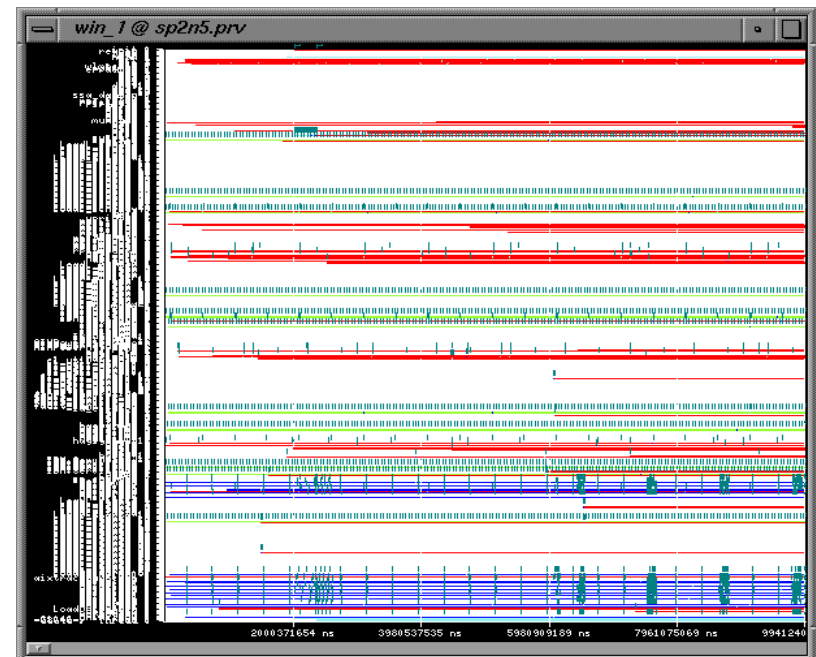
■ Per thread information:

● States and context switches:

- ✓ Not created, no info, running, blocked, stopped, ready, yield
- ✓ On which processor

● Events:

- ✓ System calls
- ✓ Arguments to system call: fd, size
- ✓ Return values of system calls
- ✓ Sockets
- ✓ SCSI driver calls
 - strategy, bstart, iodone
- ✓ **User events**



Other features

■ Traces from multiple nodes

- Synchronized reading the switch clock
- Same configuration files

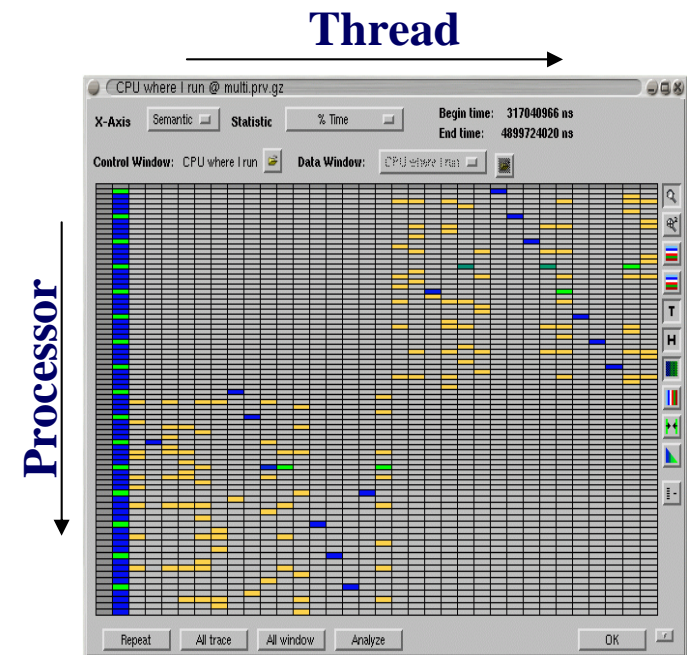
■ Software counters

- When high frequency of system calls
 - ✓ large traces
 - ✓ no need for the details of each call
- Summarization:
 - ✓ at periodic intervals
 - ✓ how many calls of each type

■ Process classification

- My application, Other appl, System procs
- Text file to define system procs names

■ Remove threads with no info



Only for 32-bit kernels!



Changes in Paraver

...This page has been intentionally left blank



Configuration files

- **Some interesting views captured**
- **Provided in two major directories**
 - Node: analyses applicable to all the processes of the node
 - ✓ resources allocation, process mapping, system calls, disk activity, sockets primitives....
 - Application: applicable to the user application only
 - ✓ few generic views
 - ✓ most specific for each application analyzed:
 - aggregate, barrier, NAS-BT



Index

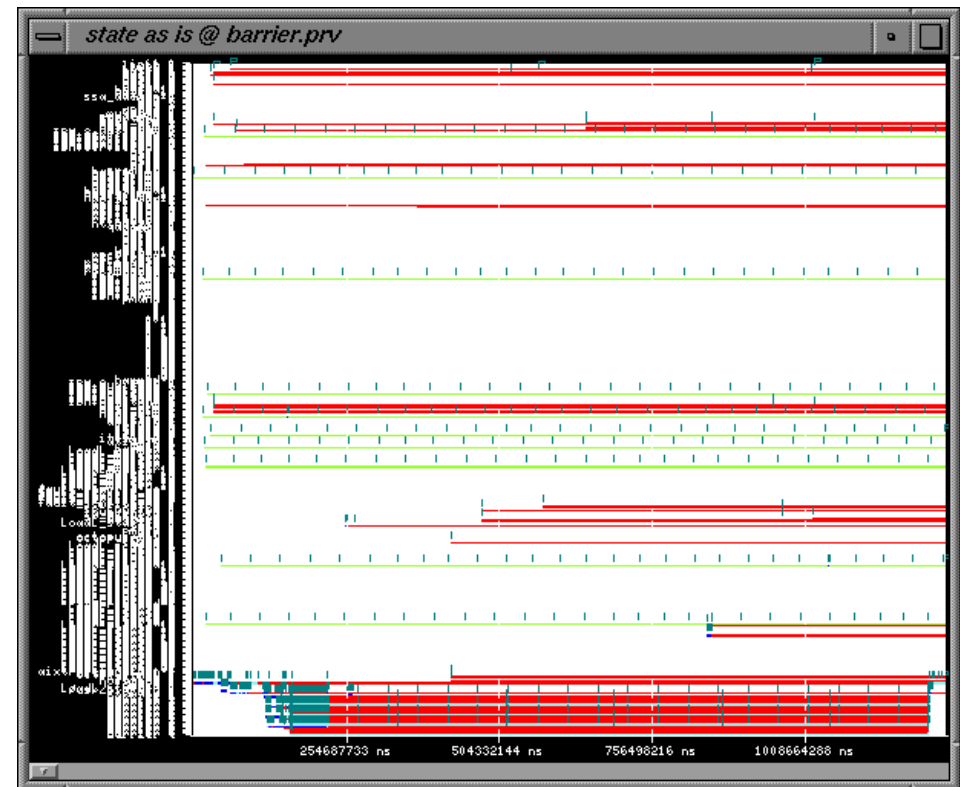
- **Motivation**
- **AIXtrace2paraver**
- **Some Examples**
 - System interferences
 - Analyzing MPI behavior
 - IRS run @ LLNL
- **Conclusions**



System interferences

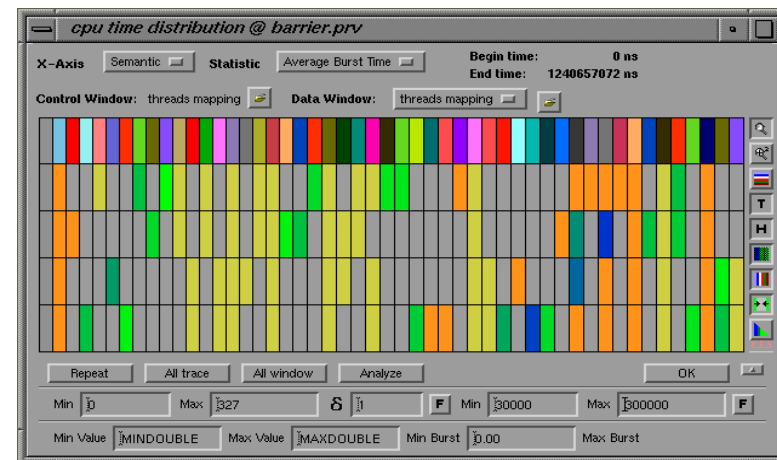
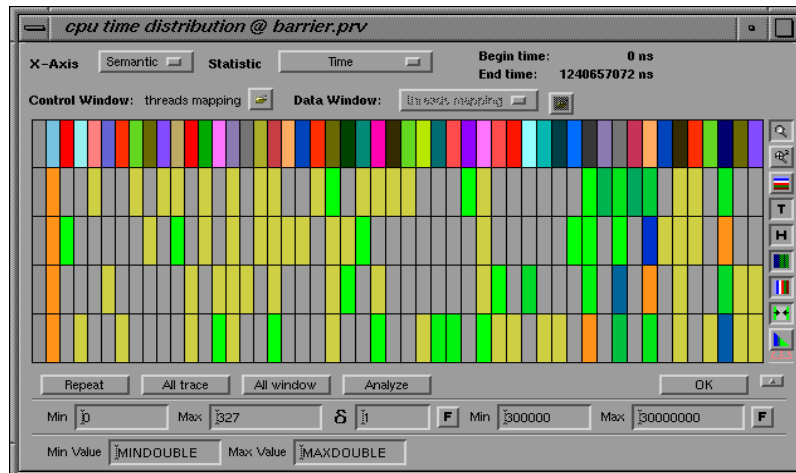
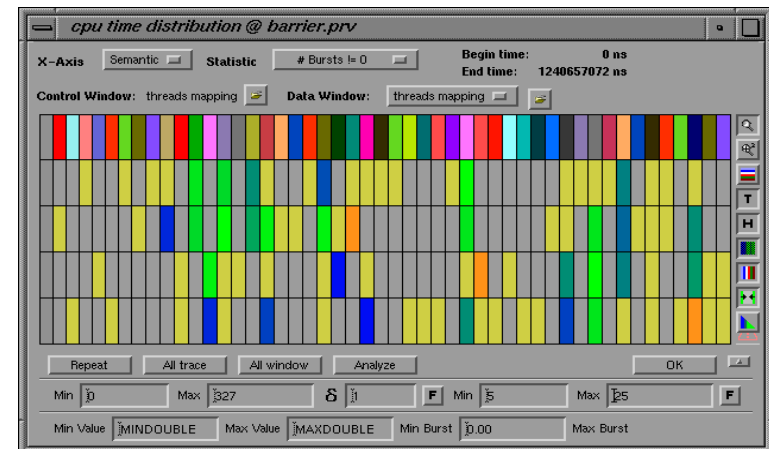
■ Environment

- Very fine grain application
 - ✓ Loop barrier - computation
- 4 tasks in a 4-way node
- No other users



System interferences – CPU time distribution

- Mapping – many processes run on most of the processors
- System processes -Typical runs of few tens of us

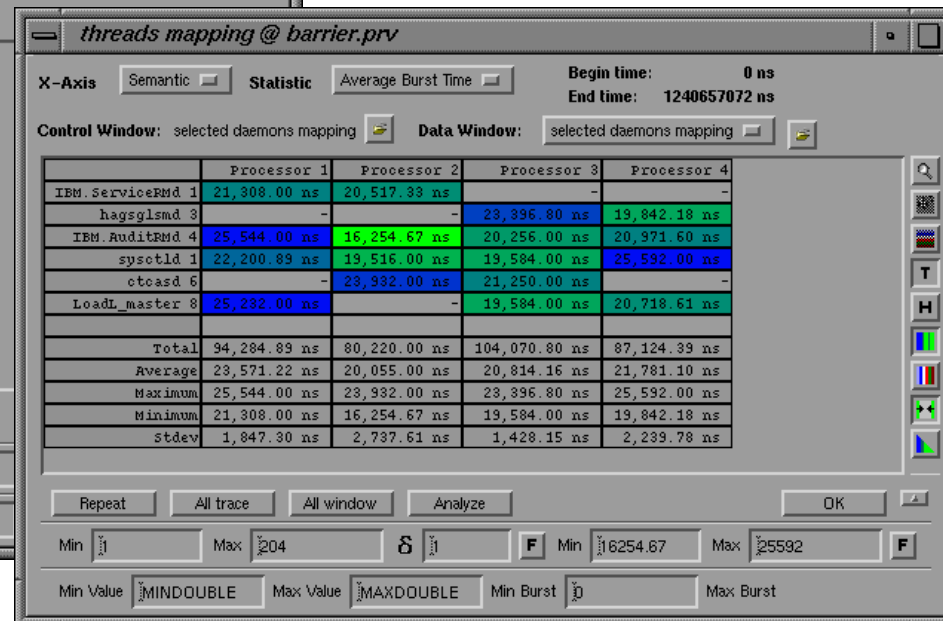
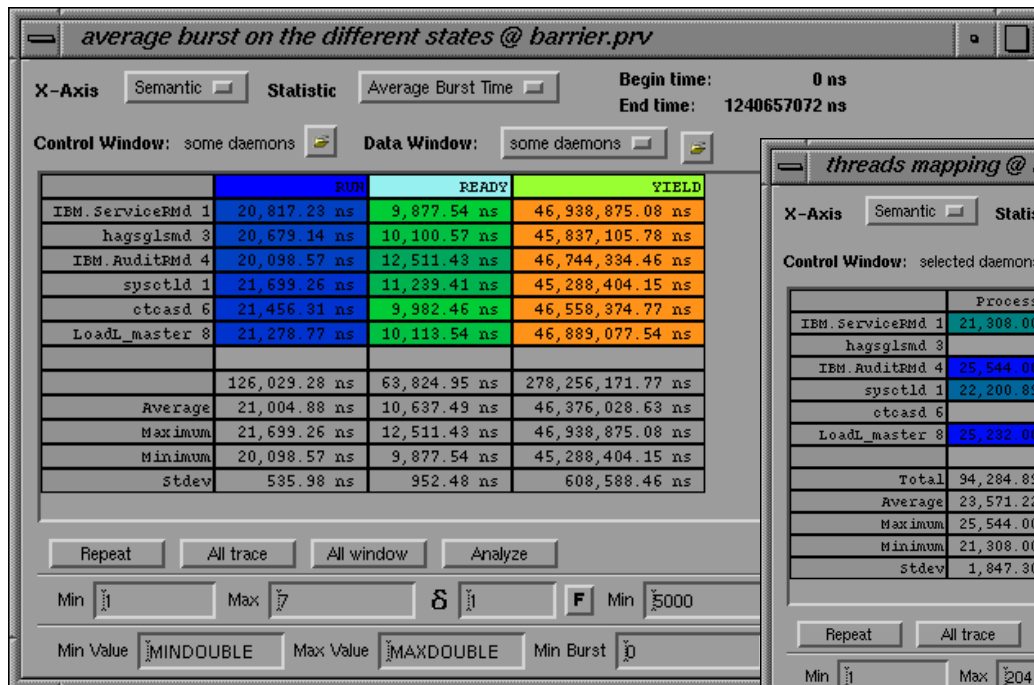
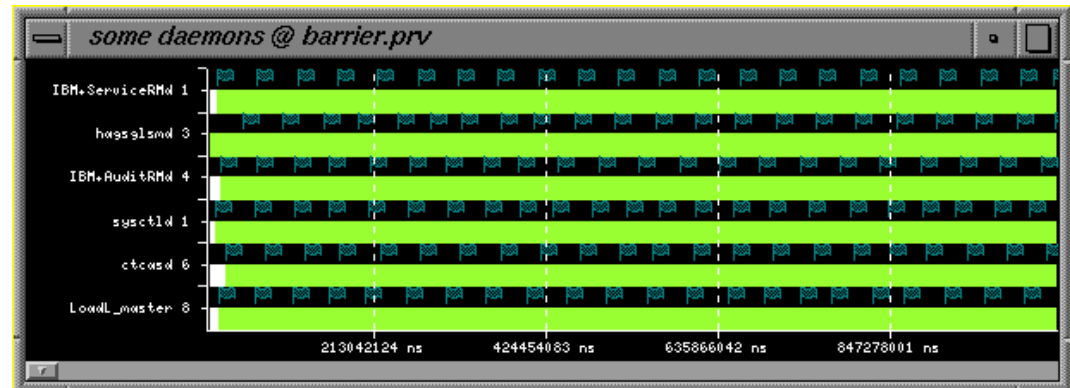


System interferences – system daemons

■ Similar behavior

- Yields for $\cong 46.4$ ms
- Run $\cong 21\mu$ s

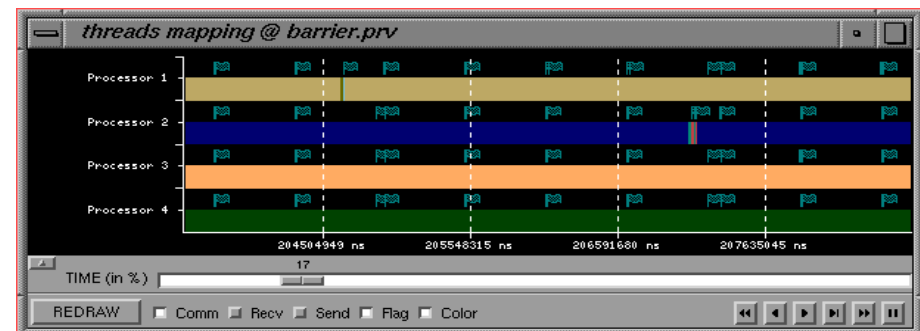
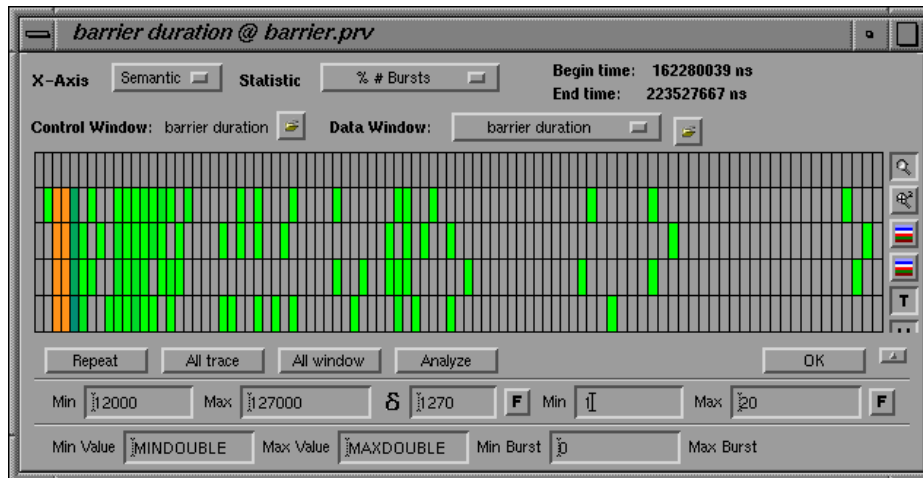
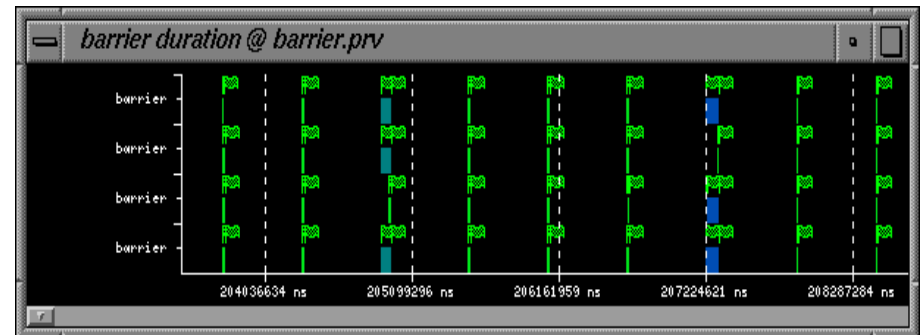
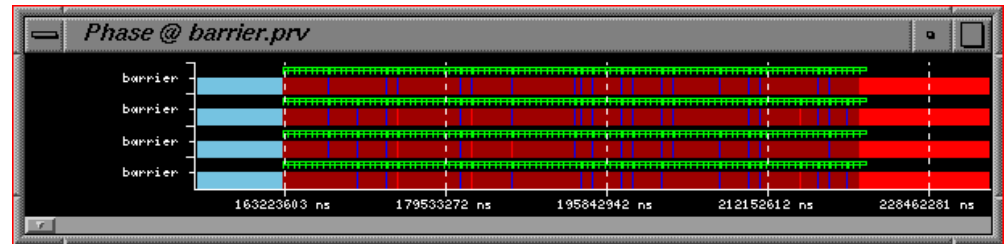
■ Most run on many CPUs



System interferences – impact on the appl.

■ User events

- Some “Very large” barriers
 - ✓ Typical – 14us
 - ✓ Large – range 66-93us
- The cost is paid by all the tasks
 - ✓ 1 task delayed by the system
 - ✓ 3 tasks wait in the barrier



Index

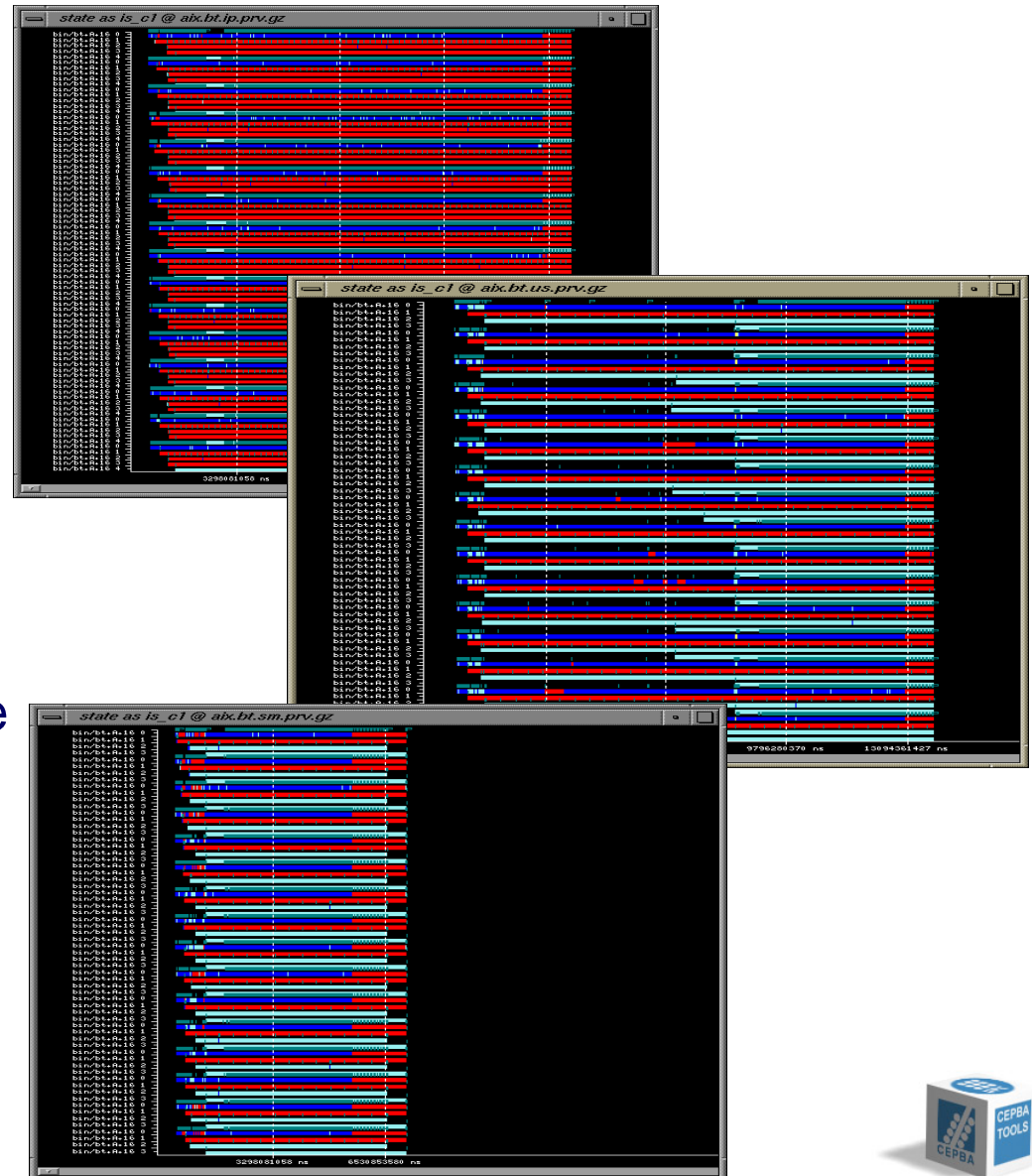
- **Motivation**
- **AIXtrace2paraver**
- **Some Examples**
 - System interferences
 - Analyzing MPI behavior
 - IRS run @ LLNL
- **Conclusions**



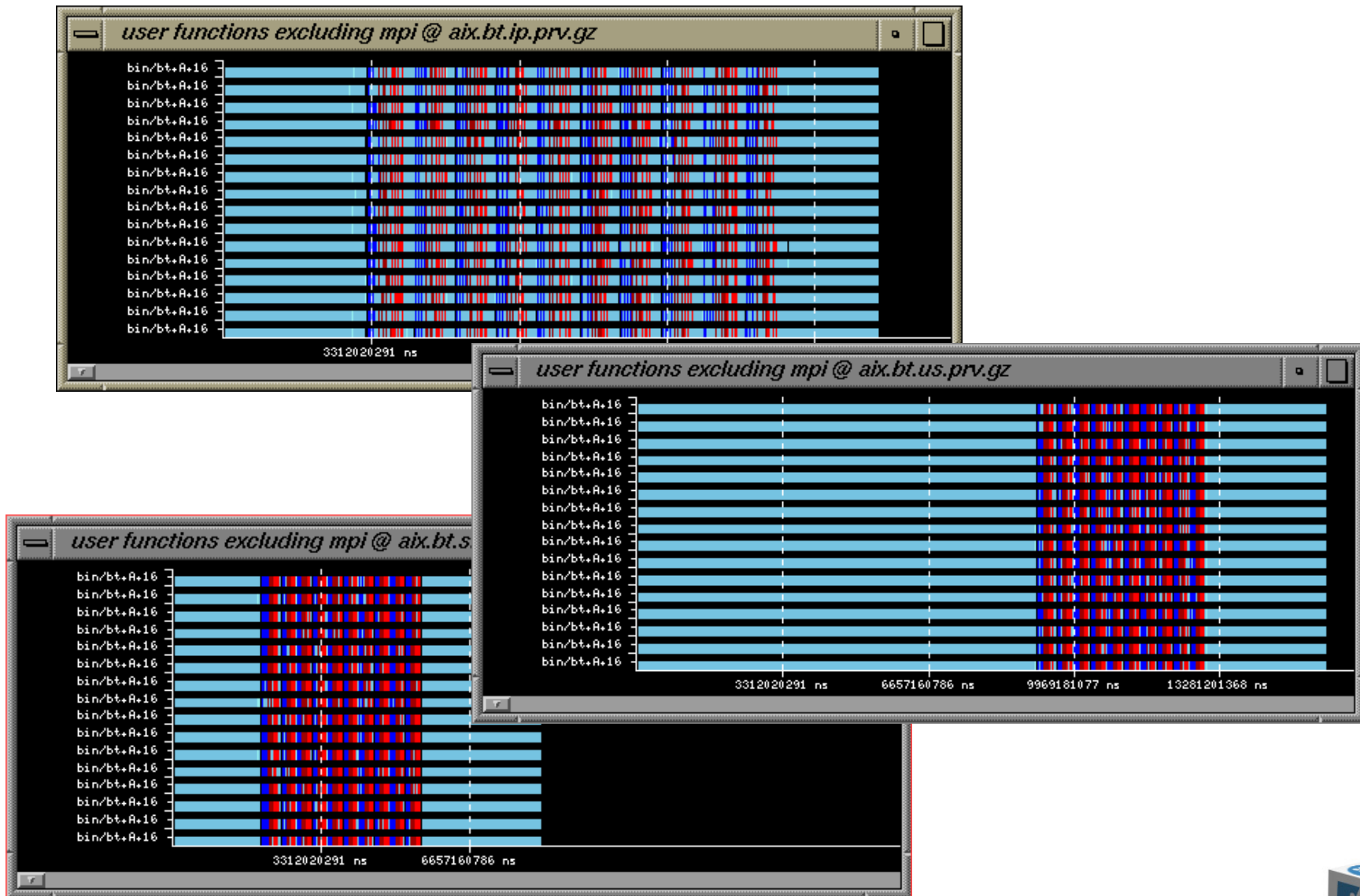
Analyzing MPI behavior

■ Environment

- NAS–BT, class A
- Modified source code to instrument
 - ✓ Some user functions
 - ✓ All mpi calls
- 16 tasks in a 16-way node
- 3 runs: SM, US, IP



Analyzing MPI behavior – user functions



Analyzing MPI behavior – time distribution

user functions excluding mpi @ aix.bt.ip.prv.gz

X-Axis: Semantic **Statistic** Time Begin time: 3135.57 ms End time: 12446.89 ms

Control Window: user functions excluding mpi Data Window: user functions excluding mpi

	exit	x_solve	y_solve	z_solve	copy_faces
bin/bt.a.16	5,653.87 ms	1,136.56 ms	1,062.93 ms	1,153.76 ms	240.21 ms
Total	90,727.79 ms	18,309.84 ms	17,312.34 ms	17,913.31 ms	3,615.34 ms
Maximum	5,670.49 ms	1,144.36 ms	1,082.02 ms	1,119.58 ms	225.96 ms
Minimum	5,769.03 ms	1,355.16 ms	1,261.59 ms	1,312.01 ms	249.19 ms
stdev	5,086.53 ms	1,098.79 ms	1,031.87 ms	1,057.42 ms	208.68 ms
c.v.	155.58 ms	56.92 ms	49.44 ms	59.17 ms	11.86 ms
	0.03 ms	0.05 ms	0.05 ms	0.05 ms	6.33 ms

user functions excluding mpi @ aix.bt.sm.prv.gz

X-Axis: Semantic **Statistic** Time Begin time: 1953.17 ms End time: 5556.73 ms

Control Window: user functions excluding mpi Data Window: user functions excluding mpi

	exit	x_solve	y_solve	z_solve	copy_faces
bin/bt.a.16	359.05 ms	1,037.56 ms	966.11 ms	987.53 ms	203.92 ms
Total	6,224.89 ms	16,546.05 ms	15,293.38 ms	15,558.28 ms	3,214.68 ms
Maximum	389.06 ms	1,034.13 ms	955.84 ms	972.39 ms	200.92 ms
Minimum	452.01 ms	1,074.95 ms	980.86 ms	991.55 ms	206.61 ms
stdev	937.07 ms	943.87 ms	174.60 ms	48.61 ms	1.99 ms
c.v.	12.58 ms	13.98 ms	7.29 ms	1.99 ms	0.04 ms
	0.01 ms	0.01 ms	0.04 ms	0.04 ms	0.04 ms

Min: 48.61 Max: 1074.95

user functions excluding mpi @ aix.bt.us.prv.gz

X-Axis: Semantic **Statistic** Time Begin time: 9073.93 ms End time: 12944.46 ms

Control Window: user functions excluding mpi Data Window: user functions excluding mpi

	exit	x_solve	y_solve	z_solve	copy_faces
bin/bt.a.16	567.17 ms	1,067.84 ms	976.43 ms	995.69 ms	211.36 ms
Total	3,448.64 ms	16,937.09 ms	15,537.67 ms	15,804.49 ms	3,393.31 ms
Maximum	590.54 ms	1,058.57 ms	971.10 ms	987.78 ms	212.08 ms
Minimum	646.35 ms	1,077.97 ms	985.52 ms	1,038.70 ms	215.83 ms
stdev	504.54 ms	1,037.01 ms	953.00 ms	966.59 ms	209.58 ms
c.v.	30.67 ms	10.46 ms	11.30 ms	16.88 ms	1.79 ms
	0.05 ms	0.01 ms	0.01 ms	0.02 ms	0.01 ms

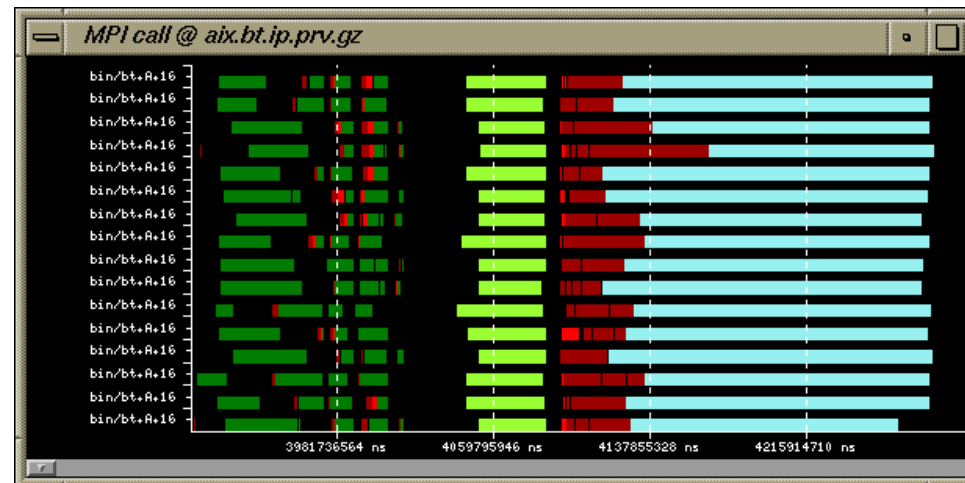
Min: 48.20 Max: 1077.97



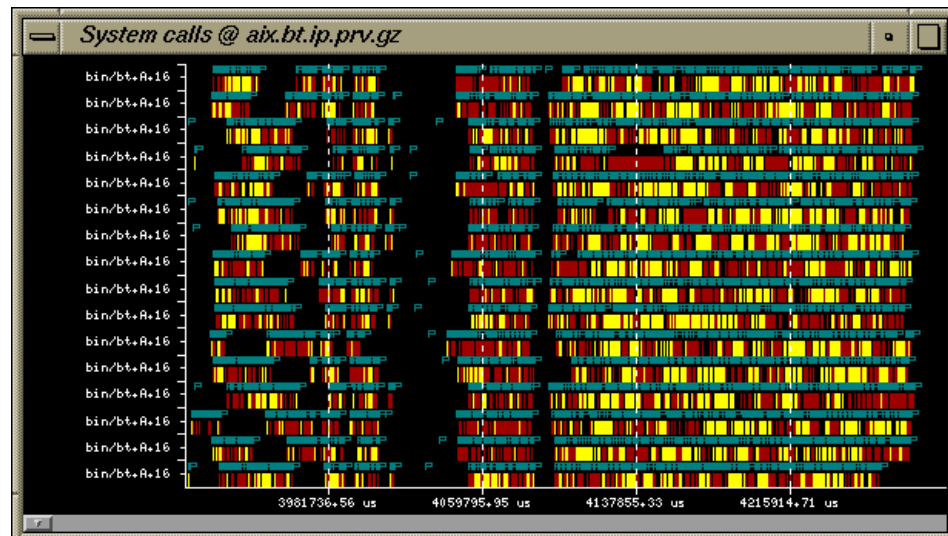
Analyzing MPI behavior – internals of MPI

■ IP implementation

- MPI calls

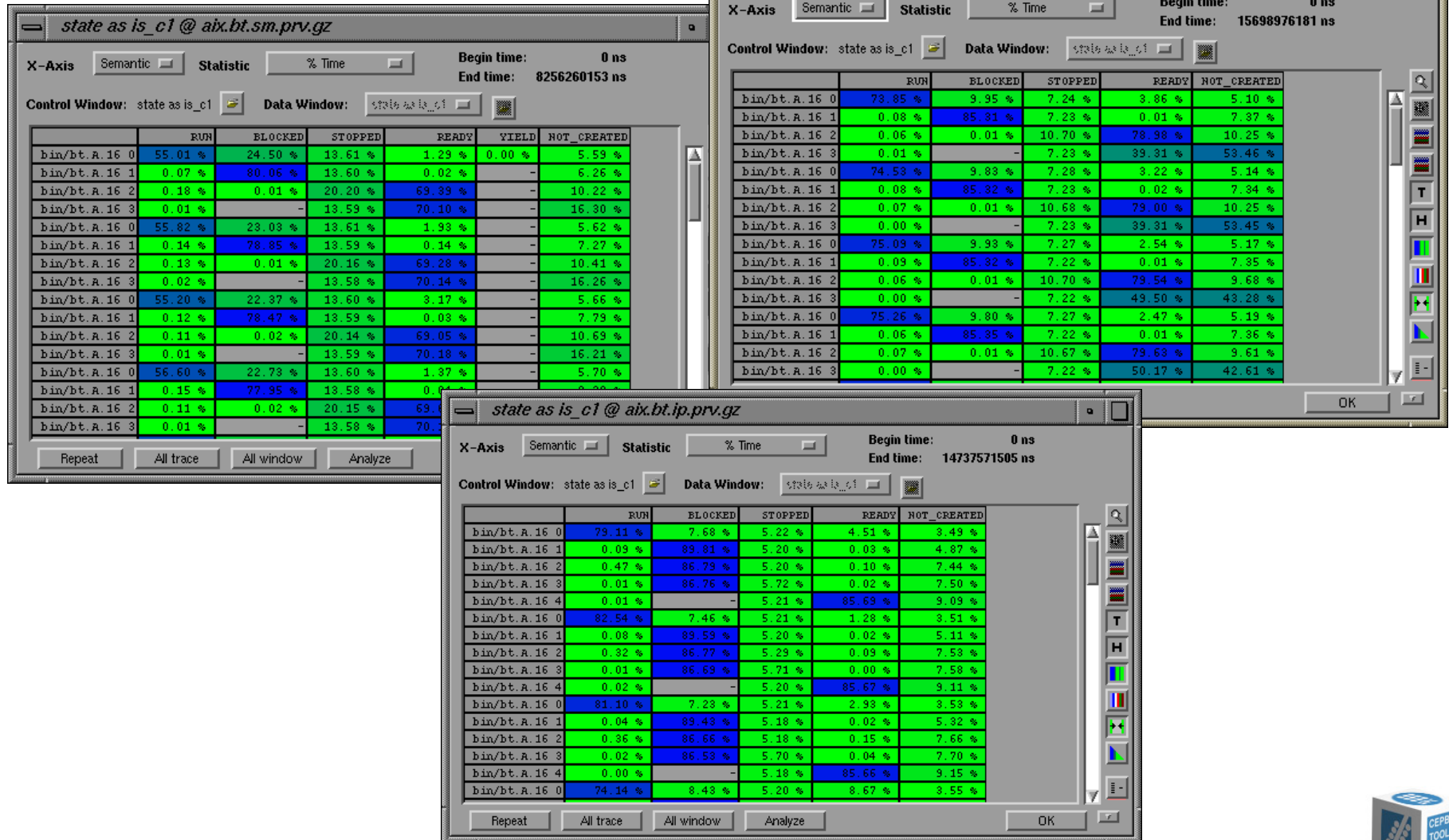


- System calls



Analyzing MPI behavior – internals of MPI

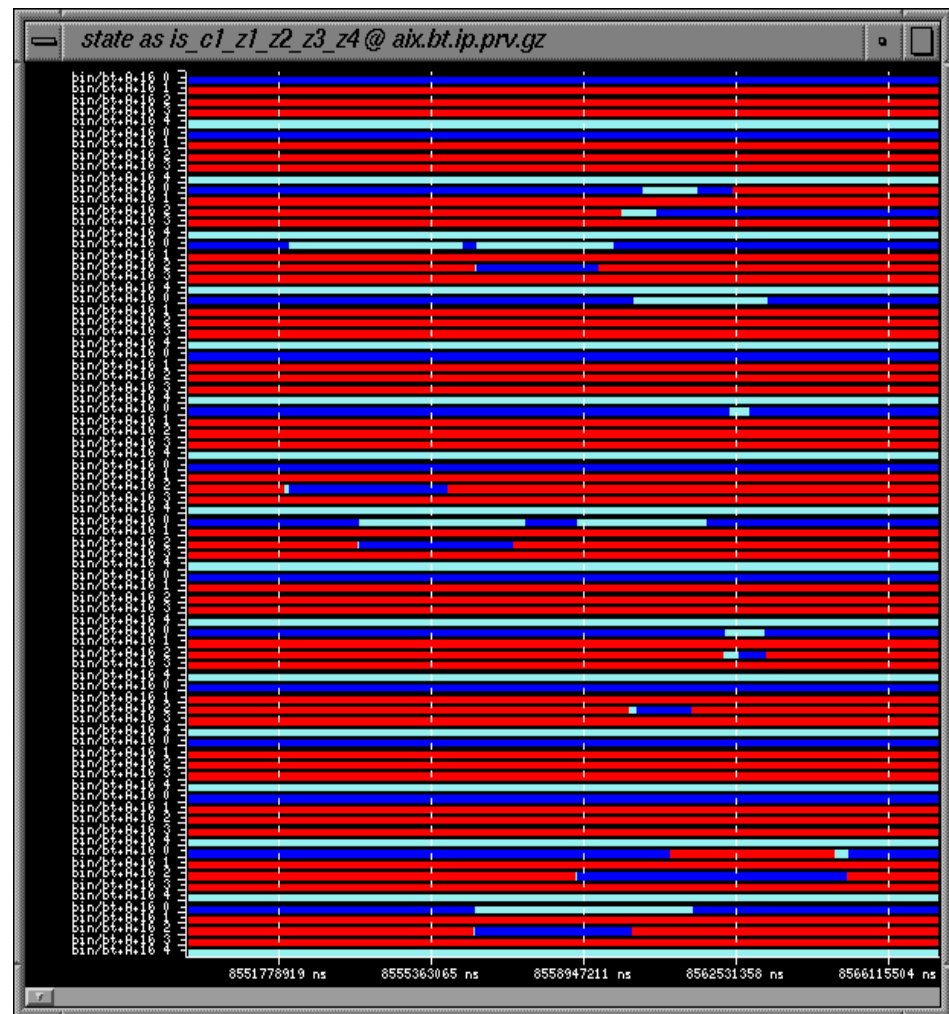
■ MPI internal daemons



Analyzing MPI behavior – internals of MPI

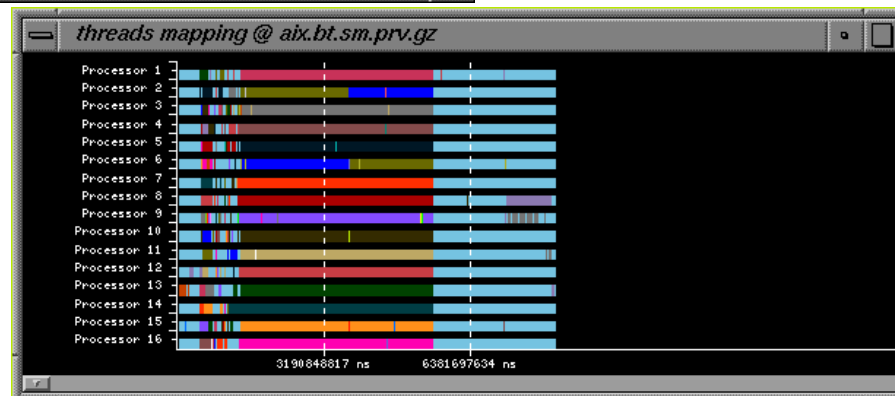
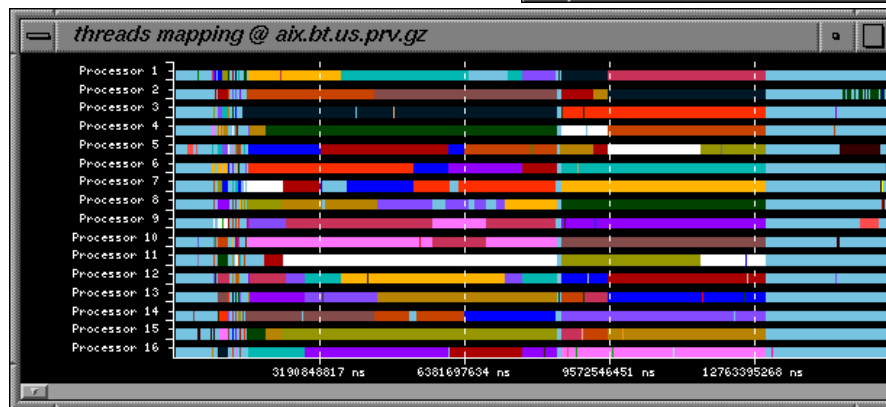
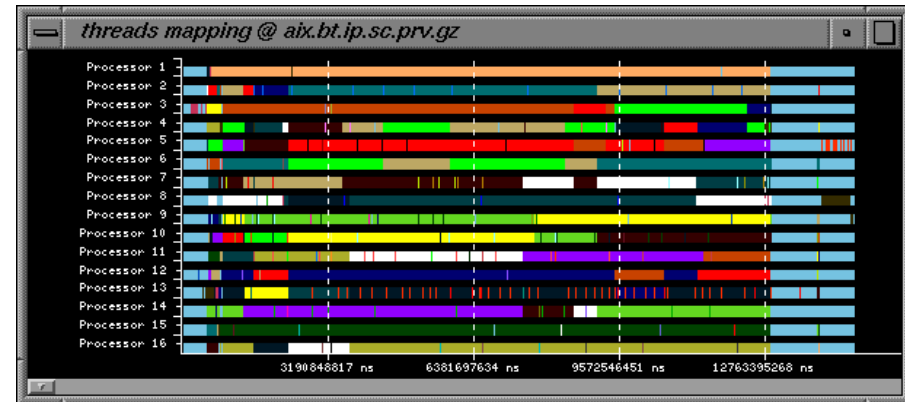
■ MPI internal daemons

- Sometimes interfere their own MPI task
- Sometimes interfere other MPI task



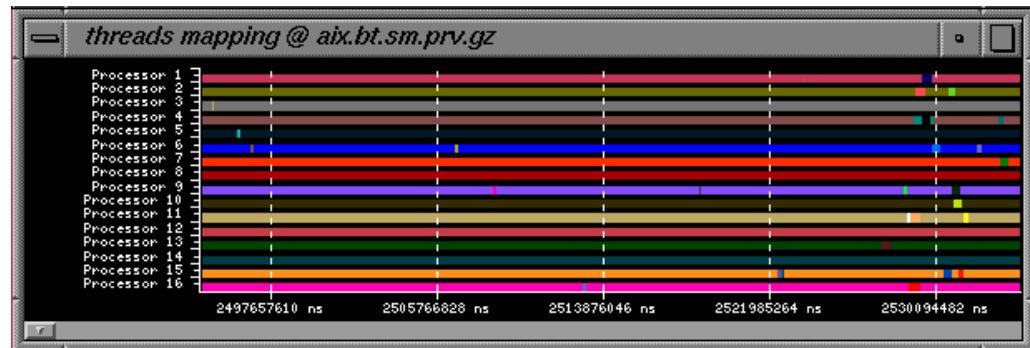
Analyzing MPI behavior – thread mapping

- Process migration
 - Initially very high
 - Not many in stable region

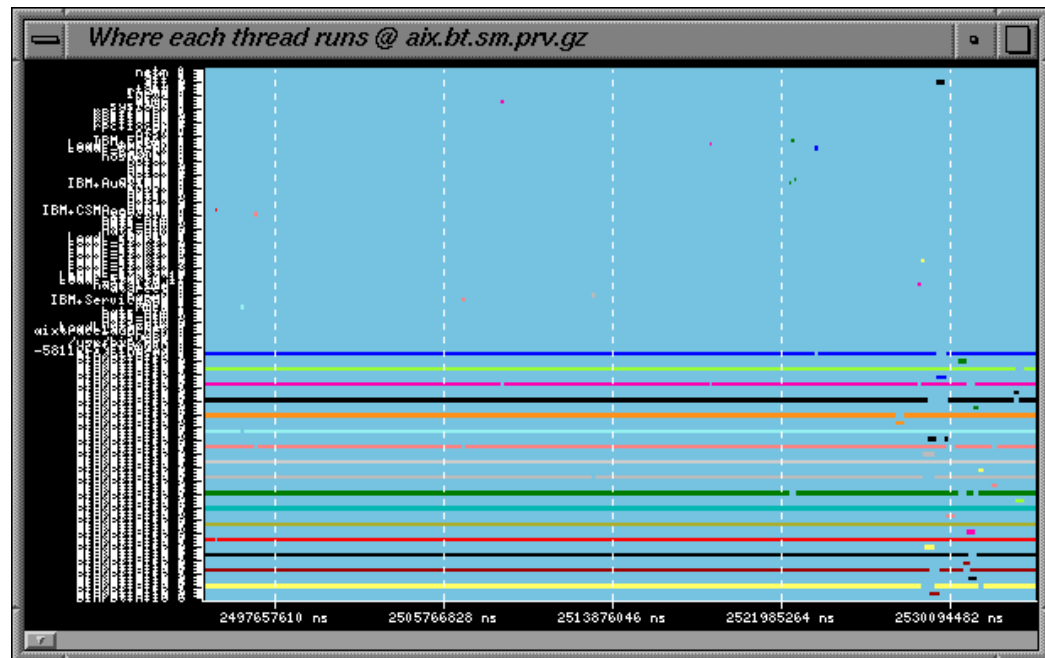


Analyzing MPI behavior – preemptions

- Zooming into stable zone of SM run



- Who ?



Index

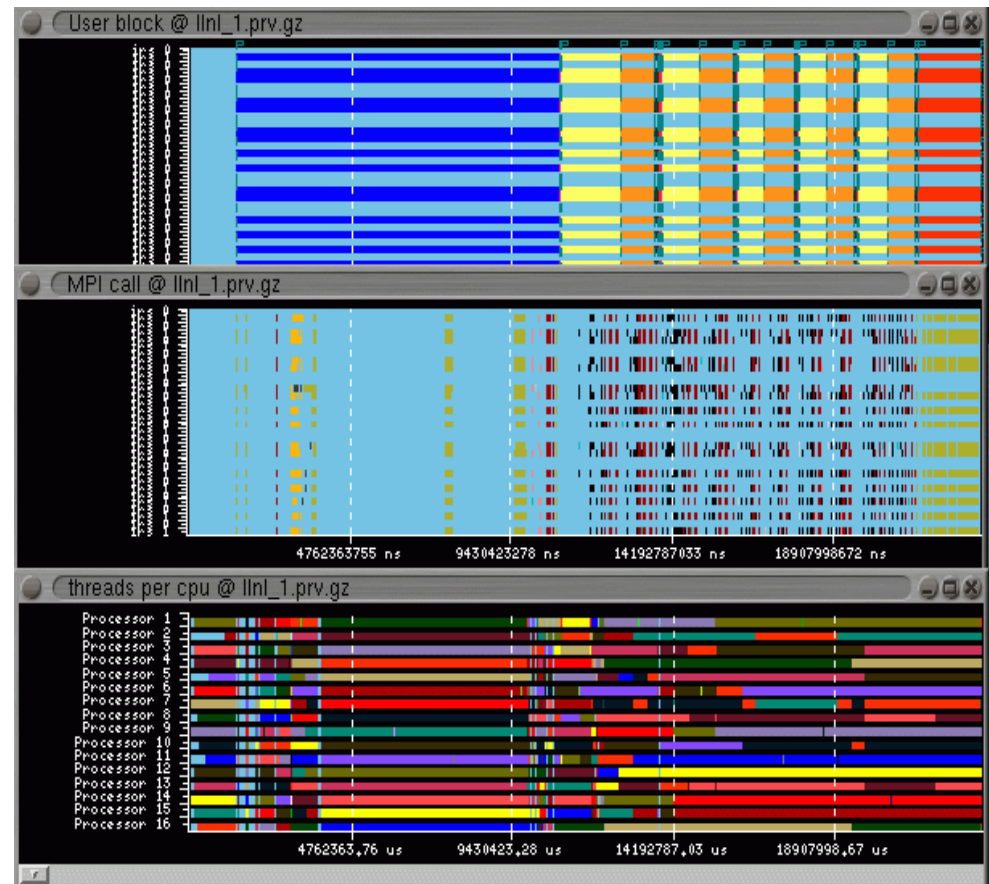
- **Motivation**
- **AIXtrace2paraver**
- **Some Examples**
 - System interferences
 - Analyzing MPI behavior
 - IRS run @ LLNL
- **Conclusions**



IRS run @ LLNL

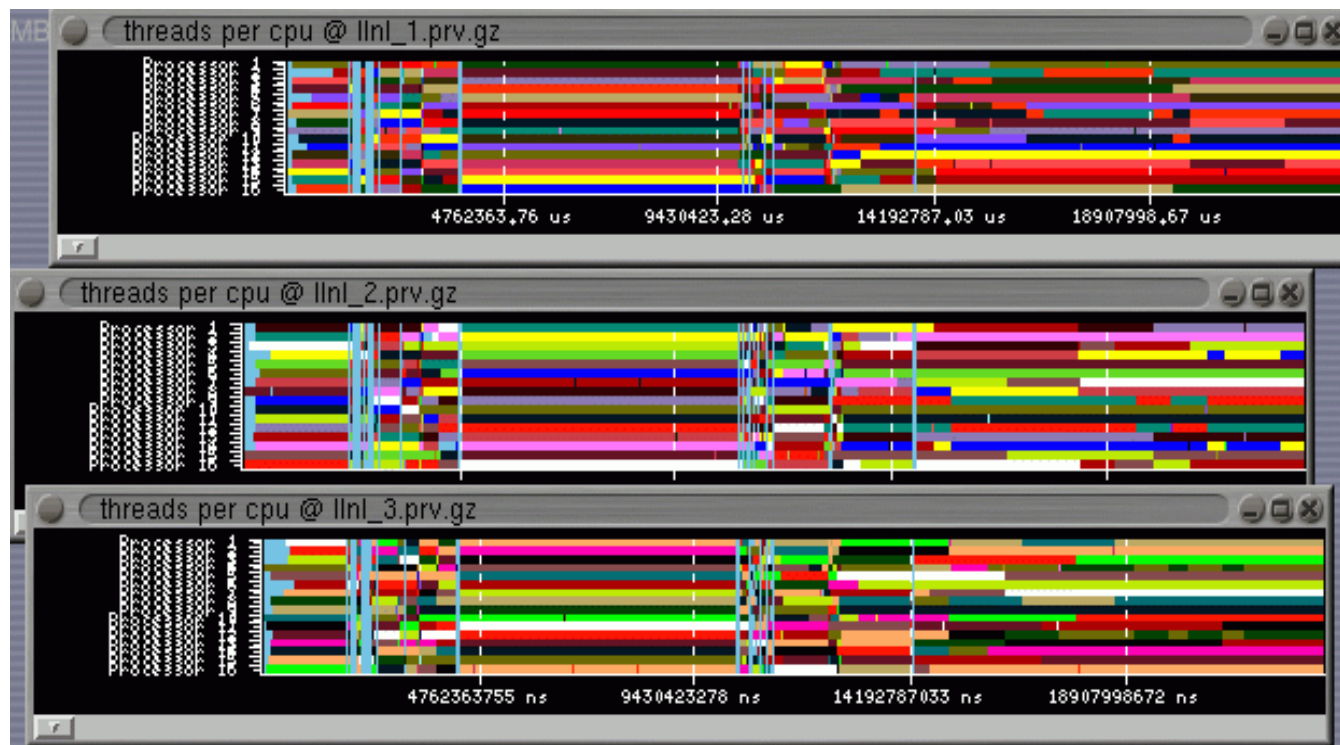
■ Environment

- IRS run on 22 nodes @ LLNL
- Trace obtained
 - ✓ without aixtracelauncher
 - ✓ without dumping switch clock
- Different mapping of the user events



IRS run @ LLNL

- Multiple nodes view
 - with manual alignment
 - Synchronized scheduling effects ?



Index

- **Motivation**
- **AIXtrace2paraver**
- **Some Examples**
- **Conclusions**



Conclusions

- **Description of the translator AIXtrace2prv developed under support from LLNL (Contact: Terry Jones)**
- **Shown the huge potential of combining**
 - The extraordinary amount of data captured by AIX trace
 - The extraordinary flexibility and processing power of Paraver to extract information from raw performance data
- **Porting to 64-bit kernels...?**
- **Mechanism to automatically insert user events...**
- **Available to Paraver users or through an evaluation license (www.cepba.upc.es/paraver)**

