

VI-HPS

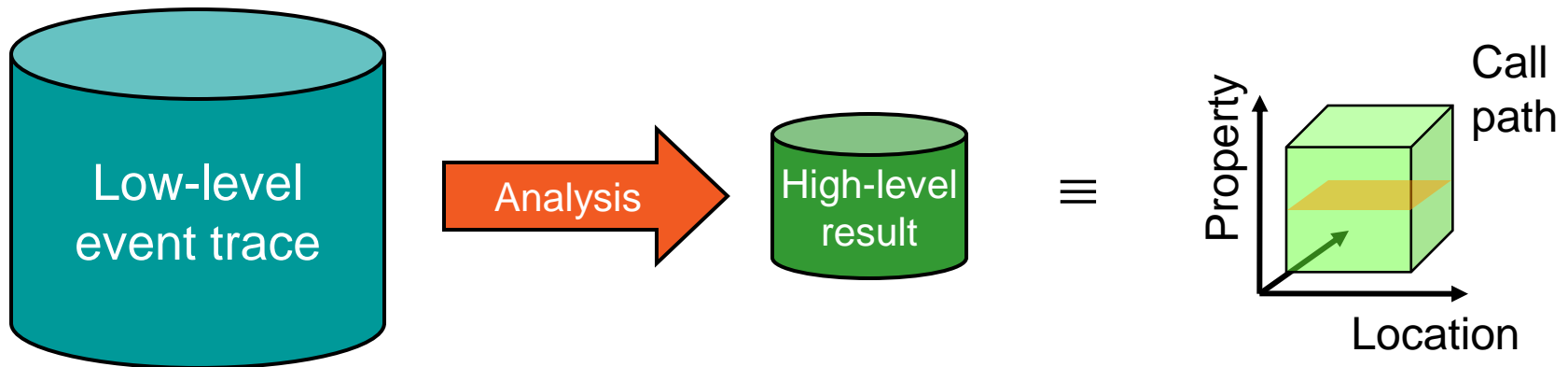


Automatic trace analysis with Scalasca

Markus Geimer
Jülich Supercomputing Centre

scalasca 

- Idea
 - Automatic search for patterns of inefficient behaviour
 - Classification of behaviour & quantification of significance



- Guaranteed to cover the entire event trace
- Quicker than manual/visual trace analysis
- Parallel replay analysis exploits memory & processors to deliver scalability

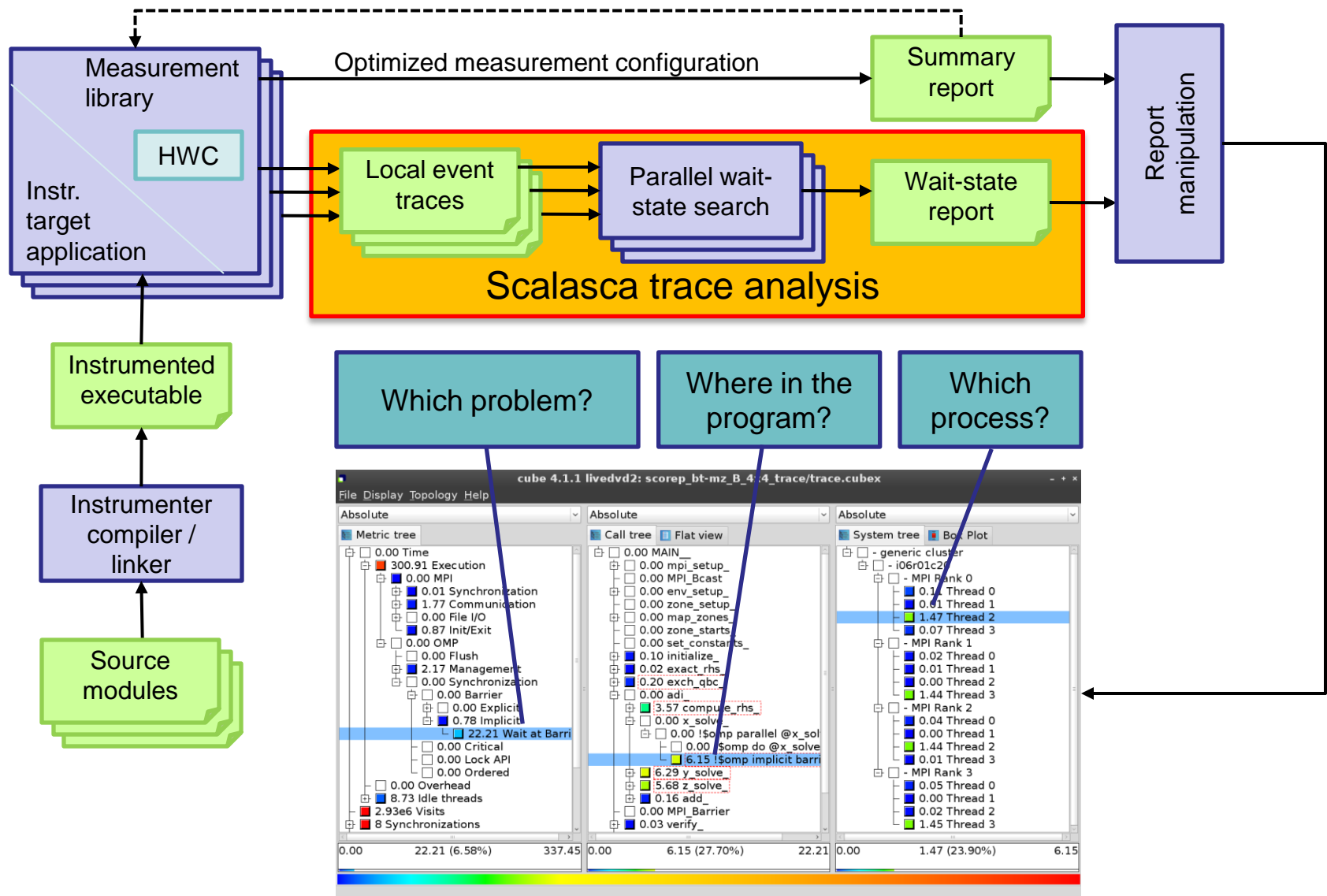
- Project started in 2006
 - Initial funding by Helmholtz Initiative & Networking Fund
 - Many follow-up projects
- Follow-up to pioneering KOJAK project (started 1998)
 - Automatic pattern-based trace analysis
- Now joint development of
 - Jülich Supercomputing Centre
 - German Research School for Simulation Sciences

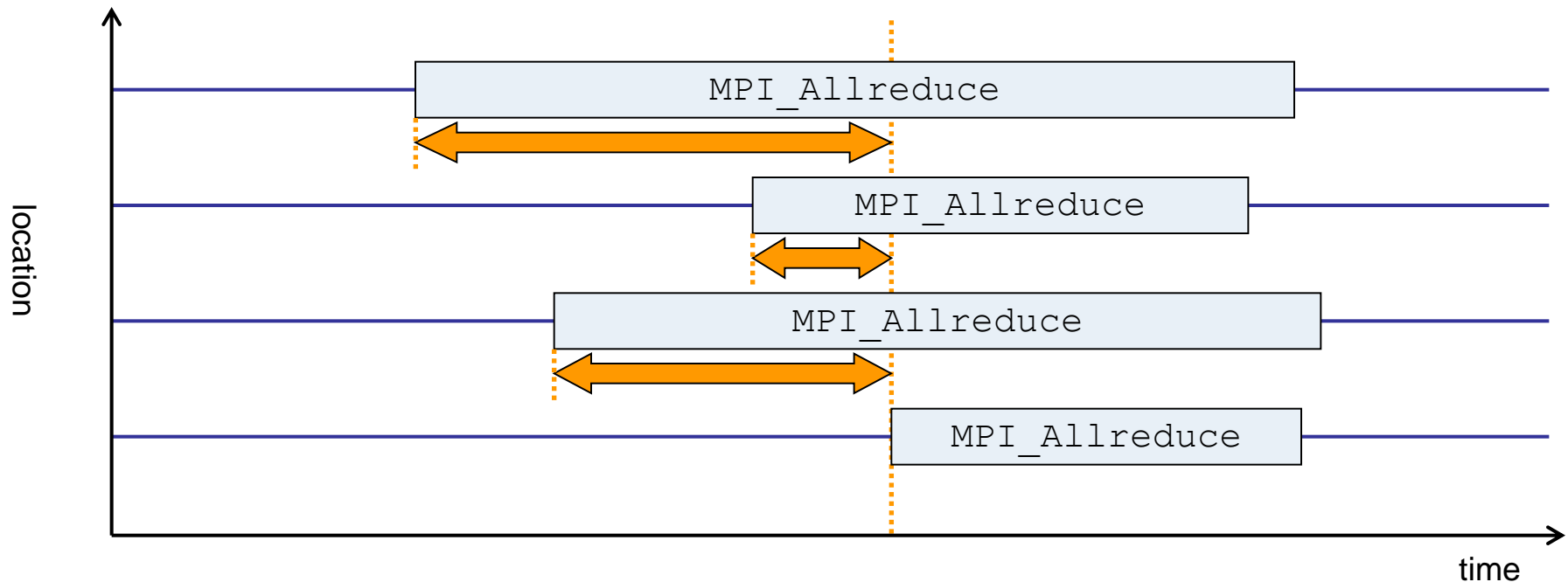


- Development of a **scalable** performance analysis toolset
- Specifically targeting **large-scale** parallel applications
 - such as those running on IBM BlueGene or Cray XT with 10,000s to 100,000s of processes
- Latest release in July 2012: Scalasca v1.4.2
- Here: Scalasca v2.0α with Score-P support
(no release date yet, available on request)

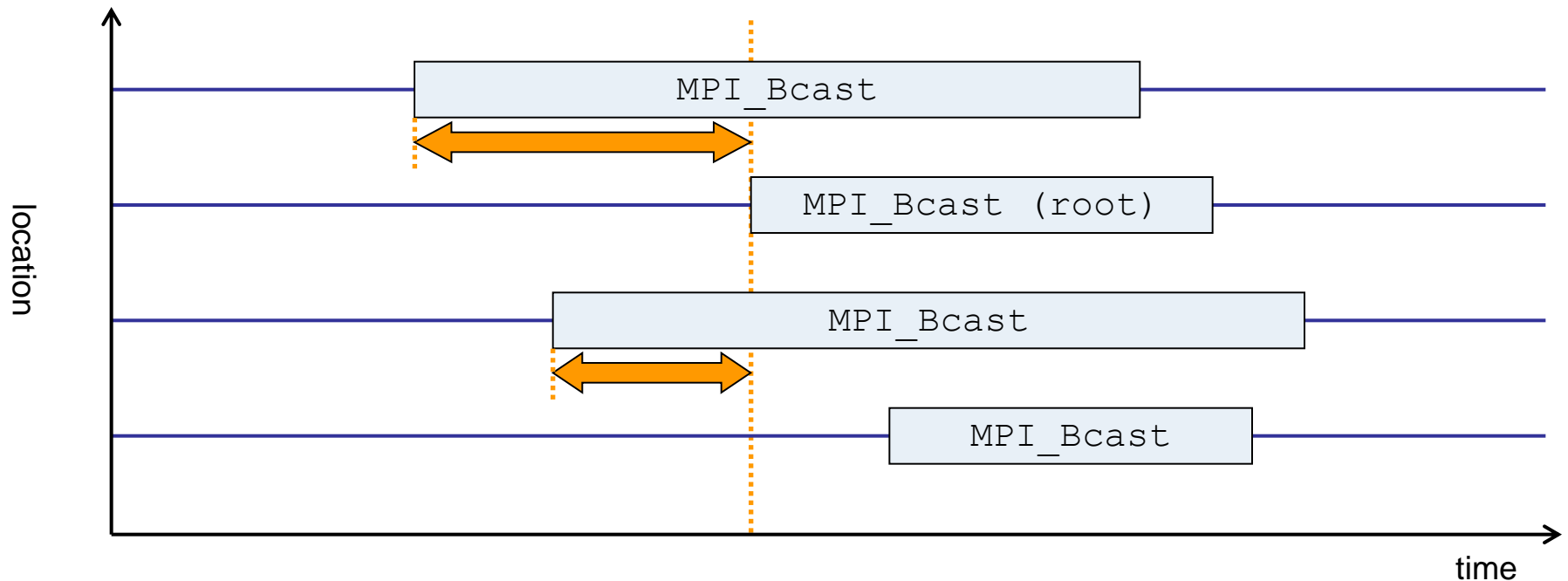
- Open source, New BSD license
- Portable
 - BG/P, BG/P, BG/L, IBM SP & blade clusters, Cray XT, SGI Altix, NEC SX, SiCortex, Solaris & Linux clusters, ...
- Supports parallel programming paradigms & languages
 - MPI, OpenMP & hybrid OpenMP/MPI
 - Fortran, C, C++
- Integrated measurement & analysis toolset
 - Runtime summarization (aka profiling)
 - Automatic event trace analysis

- Open source, New BSD license
- Still aims to be portable
 - But not widely tested yet
- Scalasca 1.4 measurement system superseded by Score-P
 - Scalasca 2.0 focuses on trace-based analyses only
- Supports common data formats
 - Reads event traces in OTF2 format
 - Writes analysis results in CUBE4 format



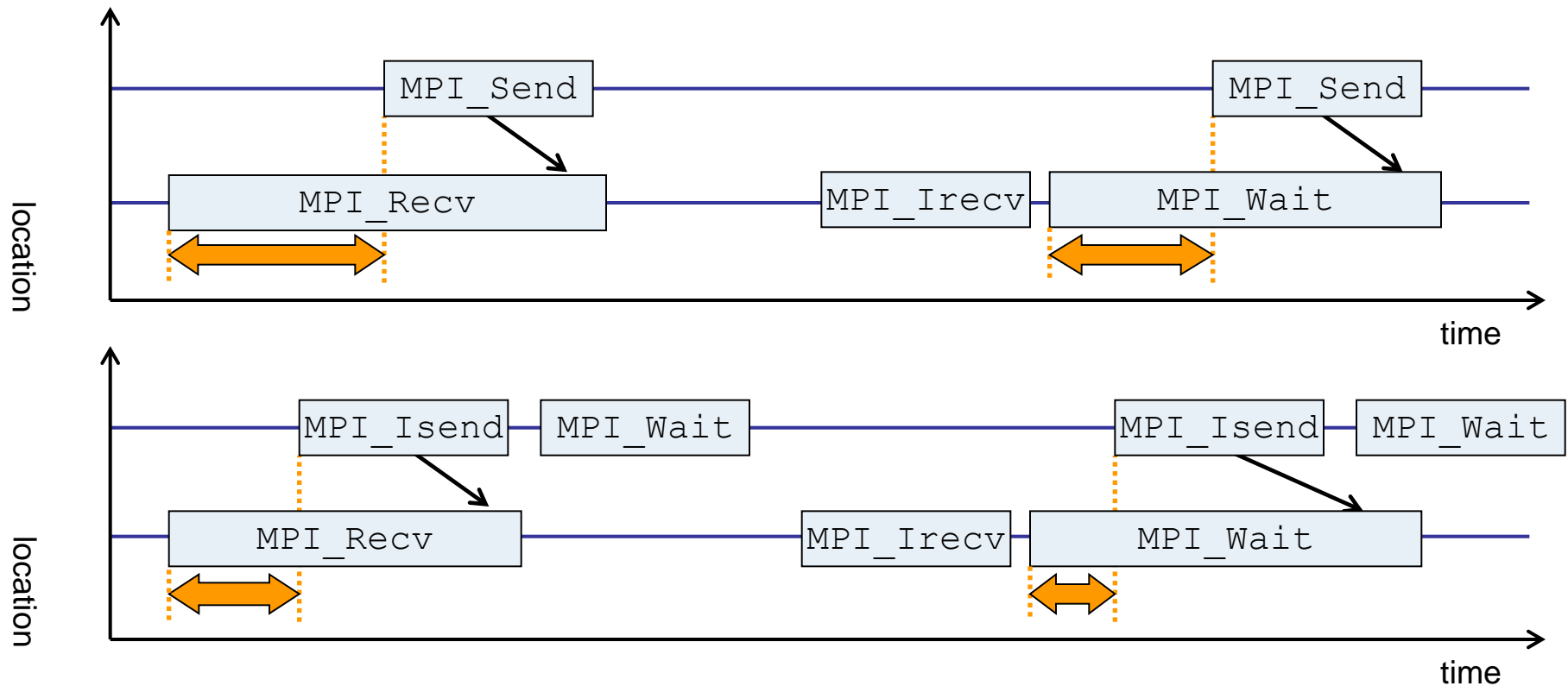


- Time spent waiting in front of synchronizing collective operation until the last process reaches the operation
- Applies to: MPI_Allgather, MPI_Allgatherv, MPI_Alltoall, MPI_Reduce_scatter, MPI_Reduce_scatter_block, MPI_Allreduce



- Waiting times if the destination processes of a collective 1-to-N operation enter the operation earlier than the source process (root)
- Applies to: MPI_Bcast, MPI_Scatter, MPI_Scatterv

Example: Late Sender



- Waiting time caused by a blocking receive operation posted earlier than the corresponding send
- Applies to blocking as well as non-blocking communication

VI-HPS



Hands-on: NPB-MZ-MPI / BT

scalasca

- Scalasca measurement collection & analysis nexus

```
% scan
Scalasca 2.0: measurement collection & analysis nexus
usage: scan {options} [launchcmd [launchargs]] target [targetargs]
      where {options} may include:
  -h      Help: show this brief usage message and exit.
  -v      Verbose: increase verbosity.
  -n      Preview: show command(s) to be launched but don't execute.
  -q      Quiescent: execution with neither summarization nor tracing.
  -s      Summary: enable runtime summarization. [Default]
  -t      Tracing: enable trace collection and analysis.
  -a      Analyze: skip measurement to (re-)analyze an existing trace.
  -e exptdir    : Experiment archive to generate and/or analyze.
                  (overrides default experiment archive title)
  -f filtdir    : File specifying measurement filter.
  -l lockfile   : File that blocks start of measurement.
  -m metrics    : Metric specification for measurement.
```

- Scalasca analysis report explorer

```
% square  
Scalasca 2.0: analysis report explorer  
usage: square [-v] [-s] [-f filtfile] [-F] <experiment archive  
          | cube file>  
-F          : Force remapping of already existing reports  
-f filtfile : Use specified filter file when doing scoring  
-s          : Skip display and output textual score report  
-v          : Enable verbose mode
```

- **scan** configures Score-P by setting some environment variables automatically
 - Precedence order:
 - Command-line arguments
 - Environment variables already set
 - Automatically determines values
- To see the effect, either open a new terminal window or unset all Score-P environment variables from previous runs

```
% unset SCOREP_EXPERIMENT_DIRECTORY  
% unset SCOREP_FILTERING_FILE  
% unset SCOREP_ENABLE_TRACING  
% unset SCOREP_ENABLE_PROFILING  
% env | grep SCOREP
```

- Also, **scan** prevents overwriting experiment directories

- Run the application using the Scalasca measurement collection & analysis nexus prefixed to launch command

```
% cd bin.scorep
% OMP_NUM_THREADS=4 scan -f scorep.filt mpiexec -np 4 ./bt-mz_W.4
S=C=A=N: Scalasca 2.0 runtime summarization
S=C=A=N: ./scorep_bt-mz_W_4x4_sum experiment archive
S=C=A=N: Thu Sep 13 18:05:17 2012: Collect start
mpiexec -np 4 ./bt-mz_W.4

NAS Parallel Benchmarks (NPB3.3-MZ-MPI) - BT-MZ MPI+OpenMP Benchmark

Number of zones:      8 x      8
Iterations: 200      dt:    0.000300
Number of active processes:      4

[... More application output ...]

S=C=A=N: Thu Sep 13 18:05:39 2012: Collect done (status=0) 22s
S=C=A=N: ./scorep_bt-mz_W_4x4_sum complete.
```

- Creates experiment directory `./scorep_bt-mz_W_4x4_sum`

- Score summary analysis report

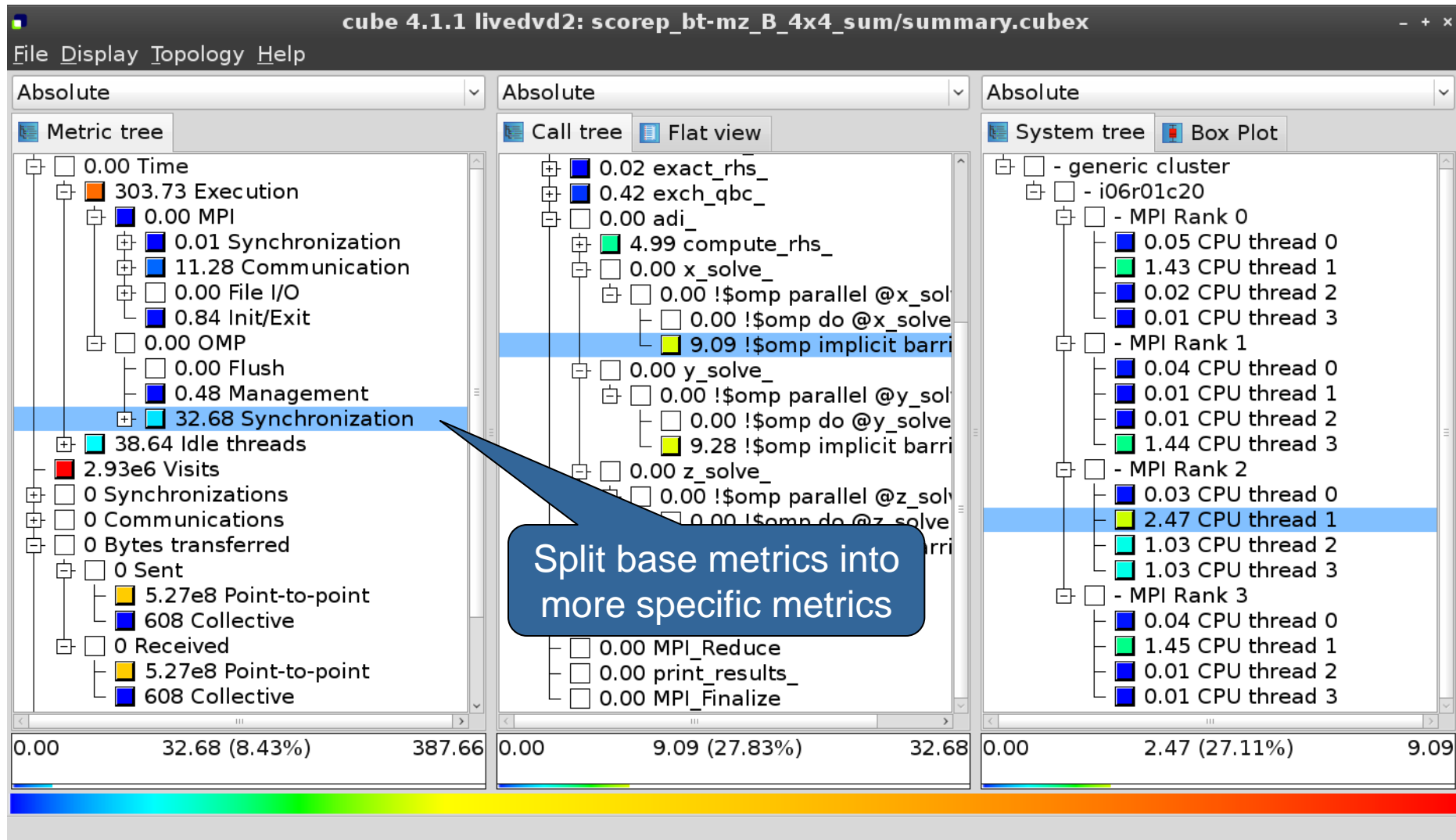
```
% square -s scorep_bt-mz_W_4x4_sum  
INFO: Post-processing runtime summarization result...  
INFO: Score report written to ./scorep_bt-mz_W_4x4_sum/scorep.score
```

- Post-processing and interactive exploration with CUBE

```
% square scorep_bt-mz_W_4x4_sum  
INFO: Displaying ./scorep_bt-mz_W_4x4_sum/summary.cubex...  
  
[GUI showing summary analysis report]
```

- The post-processing generates a metric hierarchy, splitting some base metrics into more specific metrics

Post-processed summary analysis report



- To enable additional statistics and pattern instance tracking, set `SCAN_ANALYZE_OPTS="-i"`

```
% export SCAN_ANALYZE_OPTS="-i"
```

- Re-run the application using Scalasca nexus with `"-t"` flag

```
% OMP_NUM_THREADS=4 scan -f scorep.filt -t mpiexec -np 4 ./bt-mz_W.4
S=C=A=N: Scalasca 2.0 trace collection and analysis
S=C=A=N: ./scorep_bt-mz_W_4x4_trace experiment archive
S=C=A=N: Thu Sep 13 18:05:39 2012: Collect start
mpiexec -np 4 ./bt-mz_W.4
  NAS Parallel Benchmarks (NPB3.3-MZ-MPI) - BT-MZ MPI+OpenMP Benchmark

Number of zones:    8 x    8
Iterations: 200      dt:    0.000300
Number of active processes:    4

[... More application output ...]

S=C=A=N: Thu Sep 13 18:05:58 2012: Collect done (status=0) 19s
[... continued ...]
```

- Continues with automatic (parallel) analysis of trace files

```
S=C=A=N: Thu Sep 13 18:05:58 2012: Analyze start
mpiexec -np 4 scout.hyb -i ./scorep_bt-mz_W_4x4_trace/traces.otf2
SCOUT   Copyright (c) 1998-2012 Forschungszentrum Juelich GmbH
        Copyright (c) 2009-2012 German Research School for Simulation
        Sciences GmbH

Analyzing experiment archive ./scorep_bt-mz_W_4x4_trace/traces.otf2

Opening experiment archive ... done (0.002s).
Reading definition data    ... done (0.004s).
Reading event trace data  ... done (0.669s).
Preprocessing              ... done (0.975s).
Analyzing trace data       ... done (0.675s).
Writing analysis report    ... done (0.112s).

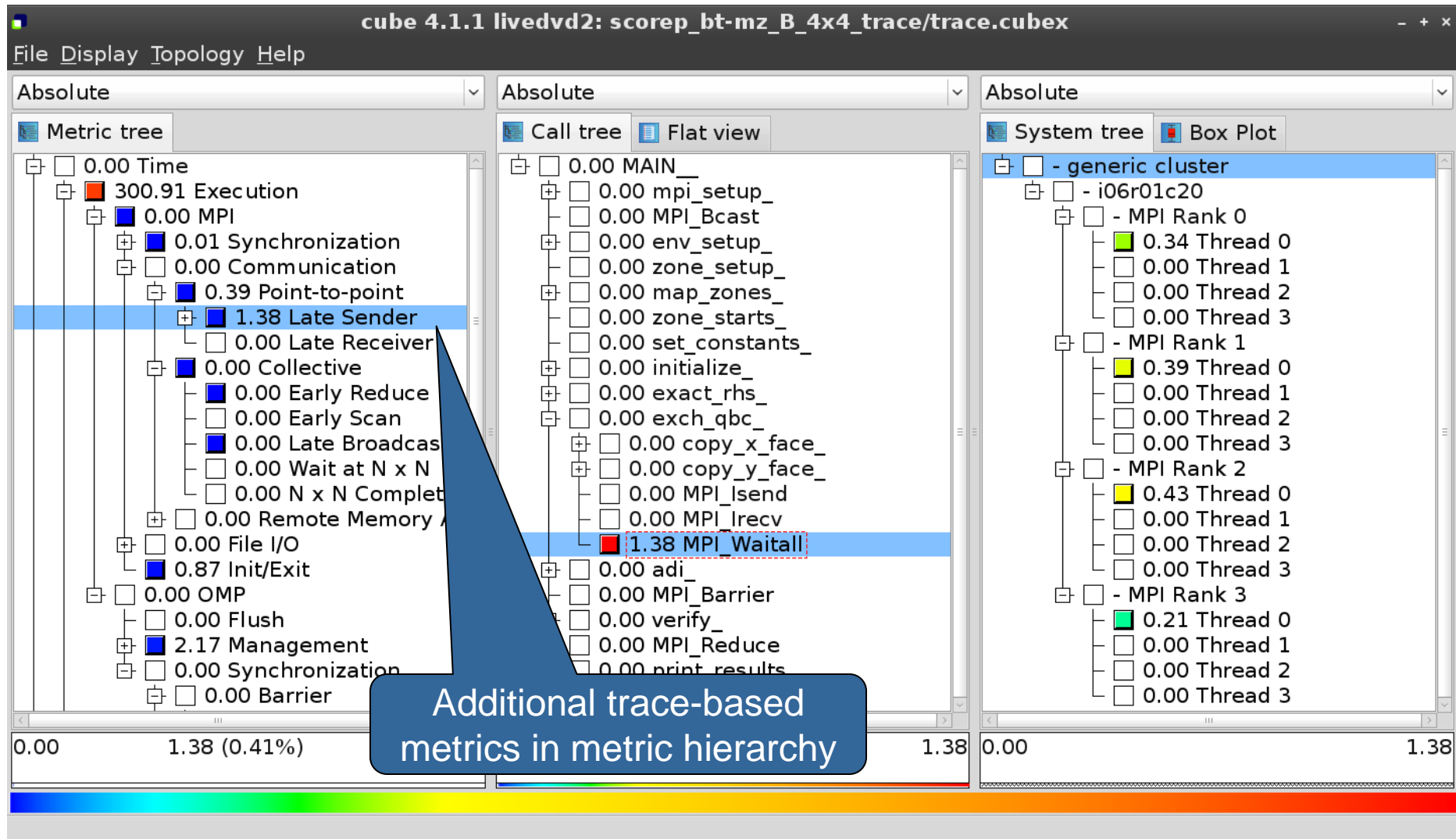
Max. memory usage         : 145.078MB

Total processing time      : 2.785s
S=C=A=N: Thu Sep 13 18:06:02 2012: Analyze done (status=0) 4s
```

- Produces trace analysis report in experiment directory containing trace-based wait-state metrics

```
% square scorep_bt-mz_W_4x4_trace  
INFO: Post-processing runtime summarization result...  
INFO: Post-processing trace analysis report...  
INFO: Displaying ./scorep_bt-mz_W_4x4_sum/trace.cubex...  
  
[GUI showing trace analysis report]
```

Post-processed trace analysis report



The screenshot displays the cube 4.1.1 interface for the file `livedvd2: scorep_bt-mz_B_4x4_trace/trace.cubex`. The interface is divided into three main panels, each with a dropdown menu set to 'Absolute'.

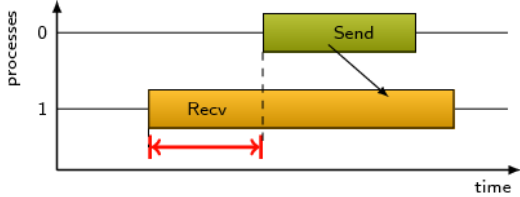
- Metric tree:** Shows a hierarchical view of metrics. The 'Late Sender' item is selected, and a context menu is open over it. The menu options include: Info, Full info, **Online description** (highlighted), Expand/collapse, Find items, Find Next, Clear found items, Copy to clipboard, Create derived metric..., Remove metric..., Statistics, and Max severity in trace browser.
- Call tree:** Shows a hierarchical view of call events. The 'MAIN' item is selected.
- System tree:** Shows a hierarchical view of system components. The 'generic cluster' item is selected.

A blue callout box with the text "Access online metric description via context menu" points to the 'Online description' option in the context menu.

Performance properties

Late Sender Time

Description:
Refers to the time lost waiting caused by a blocking receive operation (e.g., `MPI_Recv` or `MPI_Wait`) that is posted earlier than the corresponding send operation.



If the receiving process is waiting for multiple messages to arrive (e.g., in an call to `MPI_Waitall`), the maximum waiting time is accounted, i.e., the waiting time due to the latest sender.

Unit:
Seconds

Diagnosis:
Try to replace `MPI_Recv` with a non-blocking receive `MPI_Irecv` that can be posted earlier, proceed concurrently with computation, and complete with a wait operation after the message is expected to have been sent. Try to post sends earlier, such that they are available when receivers need them. Note that outstanding messages (i.e., sent before the receiver is ready) will occupy internal message buffers, and that large numbers of posted receive buffers will also introduce message management overhead, therefore moderation is advisable.

Parent:
[MPI Point-to-point Communication Time](#)

Children:

Close

The screenshot displays the VI-HPS interface with the title bar 'cube 4.1.1 livedvd2: scorep_bt-mz_B_4x4_trace/trace.cubex'. The main window shows a 'Metric tree' on the left and a 'Call tree' on the right. The 'Metric tree' is expanded to show the '1.38 Late Sender' pattern. A context menu is open over this pattern, listing various actions. The 'Statistics' option is highlighted. A callout box points to the 'Statistics' option with the text 'Access pattern instance statistics via context menu'. Another callout box points to the 'Statistics info' dialog with the text 'Click to get statistics details'. The 'Statistics info' dialog shows the following data:

Statistics info		
Pattern:	mpi_latesender	
Sum:	1.38	
Count:	832	
Mean:	0.00	5%
Standard deviation:	0.00	13%
Maximum:	0.03	100%
Upper quartile (Q3):	0.00	3%
Median:	0.00	3%
Lower quartile (Q1):	0.00	2%
Minimum:	0.00	0%

The 'Statistics info' dialog also includes a 'To Clipboard' button and a 'Close' button. The 'Metric tree' at the bottom shows a color-coded bar for the '1.38 Late Sender' pattern, with a value of 1.38 (0.41%) and a color bar ranging from blue to red. The status bar at the bottom indicates 'Shows metric statistics'.



The screenshot shows the 'cube 4.1.1' application window. The title bar reads 'cube 4.1.1 livedvd2: scorep_bt-mz_B_4x4_trace/trace.cubex'. The 'File' menu is open, showing options like 'Open...', 'Save as...', 'Close', 'Open external...', 'Close external', 'Connect to trace browser', 'Settings', 'Screenshot...', 'Quit', 'trace.cubex', and 'summary.cubex'. The 'Connect to trace browser' option is highlighted, and a submenu is visible with 'Connect to vampir...' and 'Connect to paraver...'. The 'Connect to vampir' dialog box is open, showing 'Open local file' checked, 'Host: localhost', 'Port: 30000', and 'File: c:/supermuc_expts/scorep_bt-mz_B_4x4_trace/traces.otf2'. The 'Browse' button is highlighted. The background shows a call tree and system tree view. A color bar at the bottom indicates performance metrics.

To investigate most severe pattern instances, connect to a trace browser...

...and select trace file from the experiment directory

Connect to vampir and display a trace file

Show most severe pattern instances

cube 4.1.1 livedvd2: scorep_bt-mz_B_4x4_trace/trace.cubex

File Display Topology Help

Absolute Absolute Absolute

Metric tree Call tree Flat view System tree Box Plot

0.00 Time

- 300.91 Execution
 - 0.00 MPI
 - 0.01 Synchronization
 - 0.00 Communication
 - 0.39 Point-to-point
 - 1.38 Late Sender**
 - 0.00 Late Receiver
 - 0.00 Collective
 - 0.00 Early Reduce
 - 0.00 Early Scan
 - 0.00 Late Broadcast
 - 0.00 Wait at N x N
 - 0.00 N x N Completion
 - 0.00 Remote Memory Access
 - 0.00 File I/O
 - 0.87 Init/Exit
 - 0.00 OMP
 - 0.00 Flush
 - 2.17 Management
 - 0.00 Synchronization
 - 22.99 Barrier

0.00 1.38 (0.41%) 337.45

0.00 MAIN

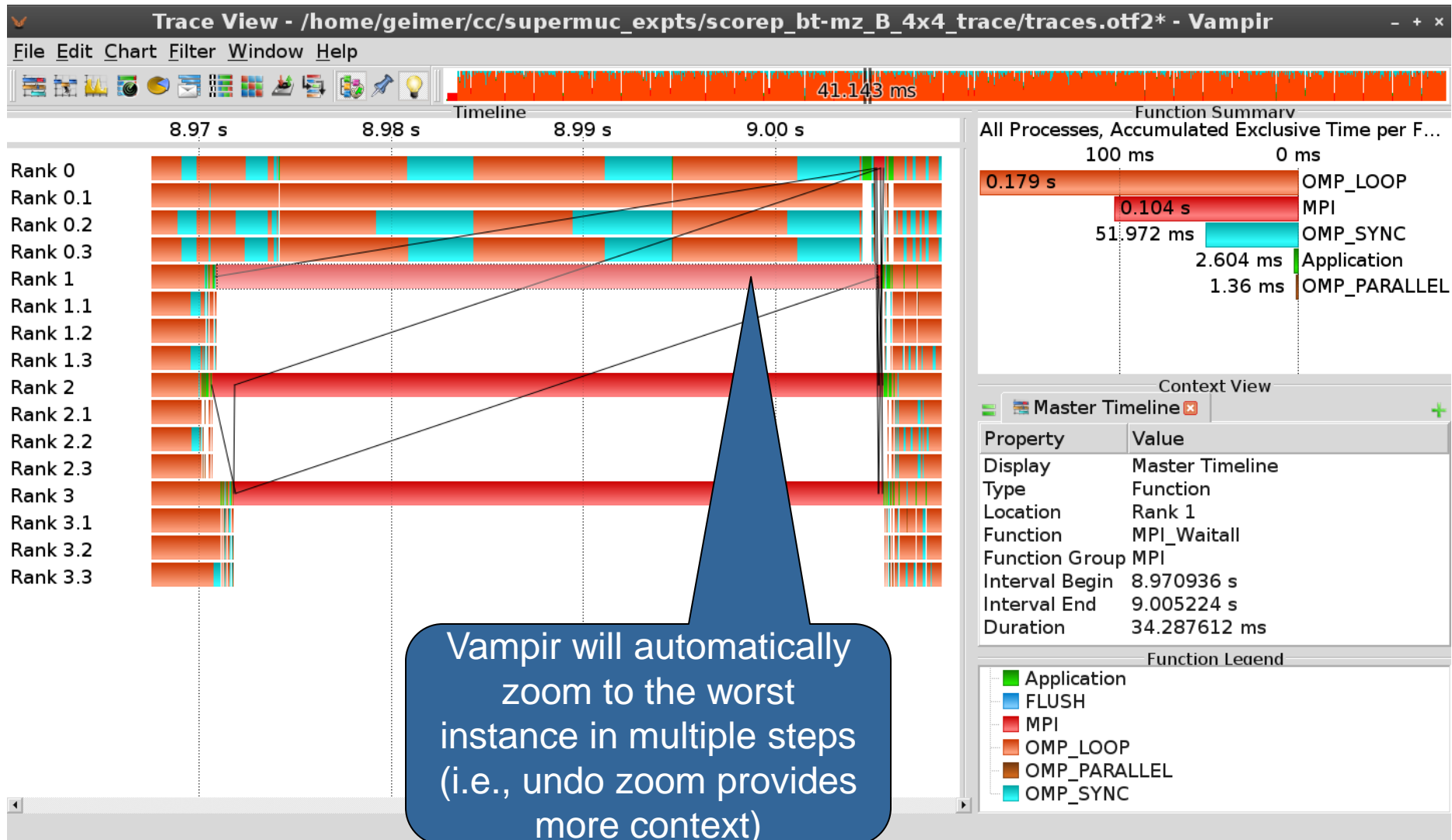
- 0.00 mpi_setenv
- 0.00 MPI_Barrier
- 0.00 env_set
- 0.00 zone_set
- 0.00 map_zone
- 0.00 zone_start
- 0.00 set_constants
- 0.00 initialize
- 0.00 exact_rank
- 0.00 exchange
- 0.00 copy
- 0.00 copy
- 0.00 MPI_Init
- 1.38 MPI_Wait**
- 0.00 adi
- 0.00 MPI_Barrier
- 0.00 verify
- 0.00 MPI_Reduce
- 0.00 print_results
- 0.00

0.00 1.38

Select "Max severity in trace browser" from context menu of call paths marked with a red frame



Investigate most severe instance in Vampir



Scalable performance analysis of large-scale parallel applications

- toolset for scalable performance measurement & analysis of MPI, OpenMP & hybrid parallel applications
- supporting most popular HPC computer systems
- available under New BSD open-source license
- sources, documentation & publications:
 - <http://www.scalasca.org>
 - mailto: scalasca@fz-juelich.de

