International HPC Summer School 2023: Performance analysis and optimization

Automatic trace analysis with Scalasca

Ilya Zhukov
Jülich Supercomputing Centre

Performance analysis steps

- 0.0 Reference preparation for validation

- 1.0 Program instrumentation
- 1.1 Summary measurement collection
- 1.2 Summary analysis report examination

- 2.0 Summary experiment scoring
- 2.1 Summary measurement collection with filtering

- 3.0 Trace measurement with filtering
- 3.1 Event trace examination & analysis

BT-MZ trace measurement collection

```
% cd bin.scorep
% cp ../jobscript/bridges2/scalasca.sbatch.C.8 .
% less scalasca.sbatch.C.8
…

# Score-P measurement configuration
export SCOREP_FILTERING_FILE=../config/scorep.filt
export SCOREP_TOTAL_MEMORY=95M

scalasca -analyze -t mpirun \
"--map-by ppr:$TASKS_PER_SOCKET:socket:pe=$SLURM_CPUS_PER_TASK \
 -n $SLURM_NTASKS" $EXE




% sbatch scalasca.sbatch.C.8
```

- Re-run measurement with new filter configuration and memory requirements

- Submit new job

BT-MZ trace measurement analysis

```
SCOUT   (Scalasca 2.6)
Copyright (c) 1998-2021 Forschungszentrum Juelich GmbH
Copyright (c) 2014-2021 RWTH Aachen University
Copyright (c) 2009-2014 German Research School for Simulation Sciences GmbH

Analyzing experiment archive ./scorep_bt-mz_C_8x6_trace/traces.otf2

Opening experiment archive ... done (0.002s).
Reading definition data    ... done (0.003s).
Reading event trace data   ... done (0.463s).
Preprocessing              ... done (0.339s).
Analyzing trace data       ... done (8.563s).
Writing analysis report    ... done (0.199s).

Max. memory usage         : 1071.035MB

Total processing time     : 9.666s
```

- Continues with automatic (parallel) analysis of trace files

Memory required for trace analysis typically several times size of trace files

BT-MZ trace analysis report examination

```
% ls
…
scalasca.sbatch.C.8                trace-C.8-<jobid>.err
trace-C.8-<jobid>.out              scorep_bt-mz_C_8x6_trace/


% ls scorep_bt-mz_C_8x6_trace/
scorep.filt       scorep.cfg       scorep.log       profile.cubex
traces/           traces.def       trace.stat       scout.cubex
scout.err         scout.log


% square -F -s scalasca_bt-mz_C_8x6_trace/
INFO: Post-processing runtime summarization report...
INFO: Post-processing trace analysis report...
…
```

- Creates experiment directory
  - The analysis report that was collated after measurement (**profile.cubex**)
  - A trace analysis was performed after the measurement (**scout.cubex**)

- The post-processing derives additional metrics and generates a structured metric hierarchy with **square -F -s** (**scalasca -examine -F -s**)

BT-MZ trace analysis report examination (cont.)

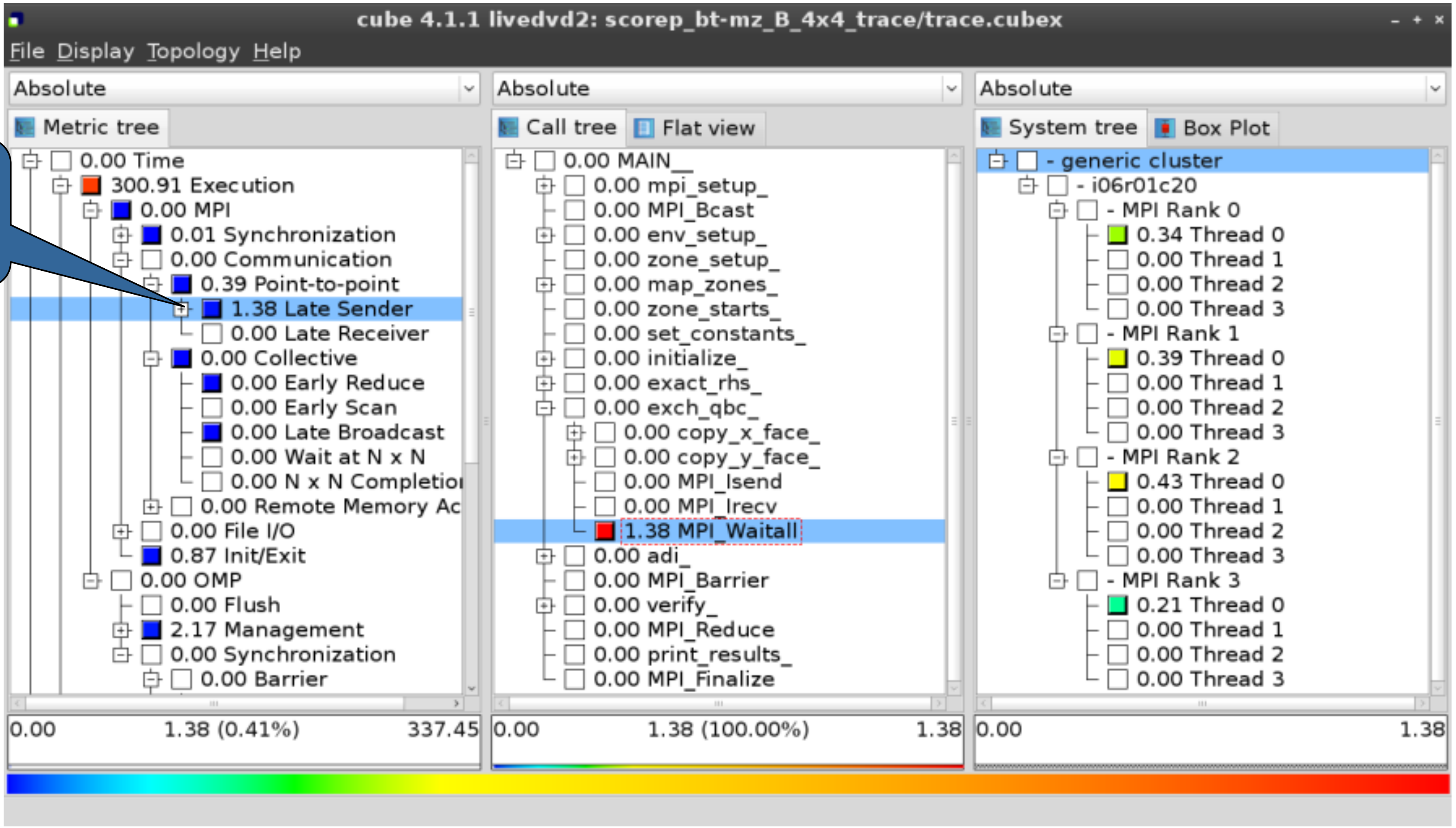
```
% ls scorep_bt-mz_C_8x6_trace/
scorep.filt     scorep.cfg
traces/         traces.def
profile.cubex   traces.otf2
scorep.log      trace.stat
scout.cubex     scout.err
scout.log       summary.cubex
trace.cubex     scorep.score




% cube scorep_bt-mz_C_8x6_trace/trace.cubex


       [cube GUI showing trace analysis report]
```

- Postprocessing
  - Provides scoring report (**scorep.score**)
  - Creates additional metrics and generates a structured metric hierarchy for with profile (**summary.cubex**) and trace analysis metrics (**trace.cubex**)
- Interactive exploration with CUBE

Hint:
Copy '*.cubex' to your laptop using 'scp' to improve responsiveness of GUI

Post-processed trace analysis report

Additional trace-based metrics in metric hierarchy



cube 4.1.1 livedvd2: scorep_bt-mz_B_4x4_trace/trace.cubex

File Display Topology Help

Absolute | Absolute | Absolute

Metric tree:
```
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 Ac
      0.00 File I/O
      0.87 Init/Exit
    0.00 OMP
      0.00 Flush
      2.17 Management
      0.00 Synchronization
        0.00 Barrier
```

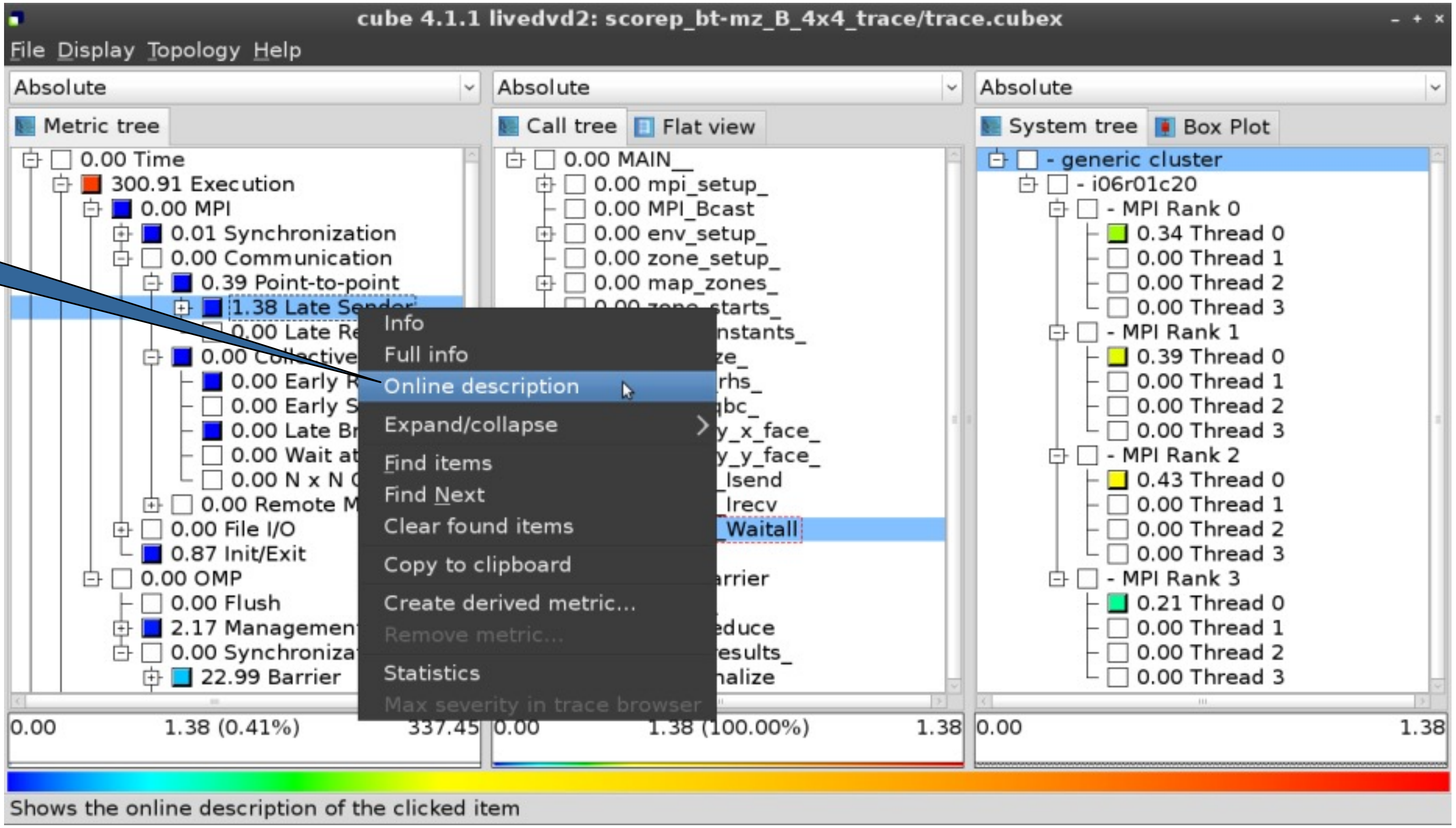
Call tree / Flat view:
```
0.00 MAIN__
  0.00 mpi_setup_
  0.00 MPI_Bcast
  0.00 env_setup_
  0.00 zone_setup_
  0.00 map_zones_
  0.00 zone_starts_
  0.00 set_constants_
  0.00 initialize_
  0.00 exact_rhs_
  0.00 exch_qbc_
    0.00 copy_x_face_
    0.00 copy_y_face_
    0.00 MPI_Isend
    0.00 MPI_Irecv
    1.38 MPI_Waitall
  0.00 adi_
  0.00 MPI_Barrier
  0.00 verify_
  0.00 MPI_Reduce
  0.00 print_results_
  0.00 MPI_Finalize
```

System tree / Box Plot:
```
- generic cluster
  - i06r01c20
    - MPI Rank 0
      0.34 Thread 0
      0.00 Thread 1
      0.00 Thread 2
      0.00 Thread 3
    - MPI Rank 1
      0.39 Thread 0
      0.00 Thread 1
      0.00 Thread 2
      0.00 Thread 3
    - MPI Rank 2
      0.43 Thread 0
      0.00 Thread 1
      0.00 Thread 2
      0.00 Thread 3
    - MPI Rank 3
      0.21 Thread 0
      0.00 Thread 1
      0.00 Thread 2
      0.00 Thread 3
```

```
0.00   1.38 (0.41%)   337.45 | 0.00   1.38 (100.00%)   1.38 | 0.00   1.38
```

Online metric description

Access online metric description via context menu



```
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 Re
        0.00 Collective
          0.00 Early R
          0.00 Early S
          0.00 Late Br
          0.00 Wait at
          0.00 N x N C
        0.00 Remote M
      0.00 File I/O
      0.87 Init/Exit
    0.00 OMP
      0.00 Flush
      2.17 Managemen
      0.00 Synchroniza
        22.99 Barrier
```

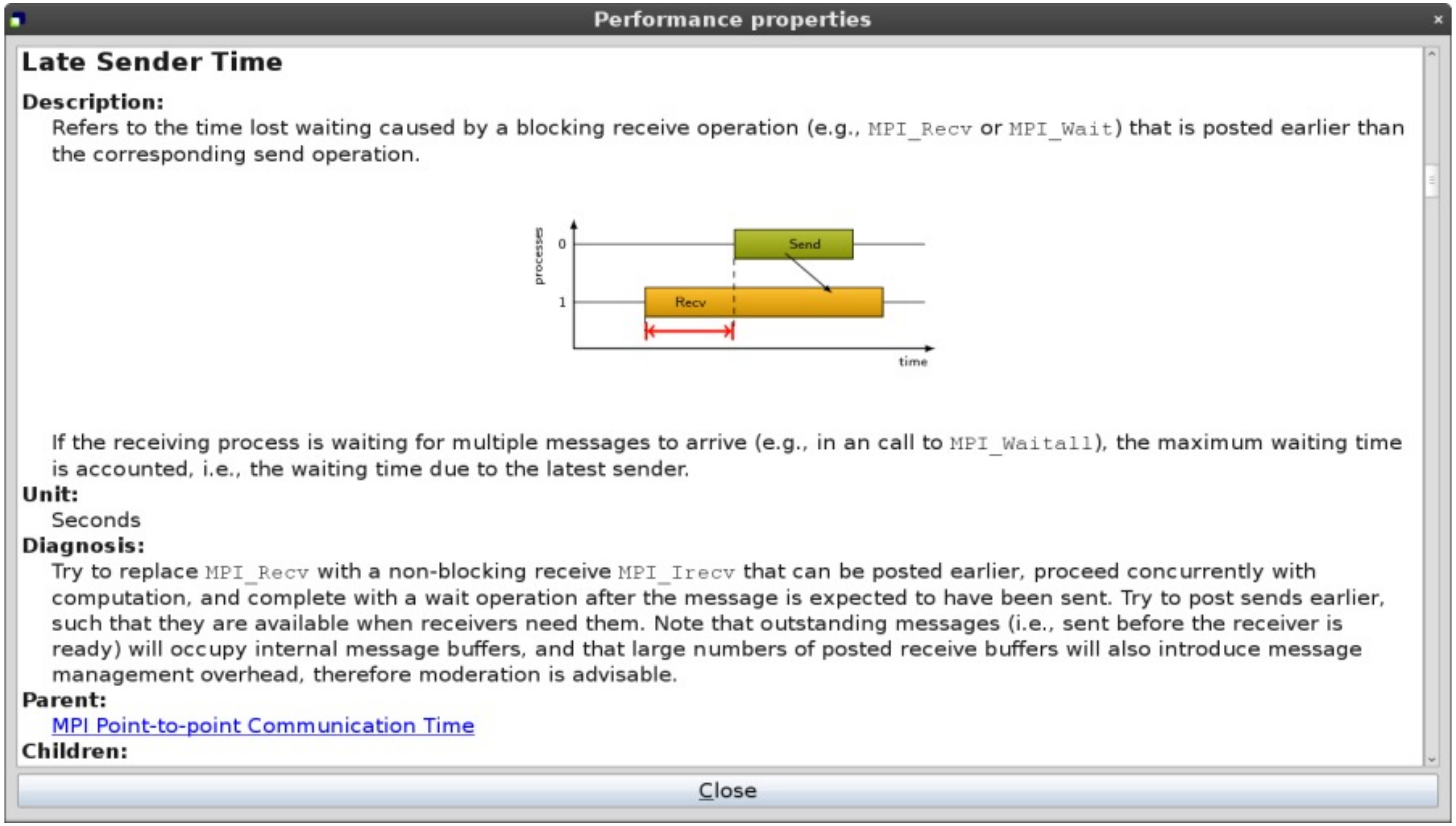
```
0.00 MAIN__
  0.00 mpi_setup_
  0.00 MPI_Bcast
  0.00 env_setup_
  0.00 zone_setup_
  0.00 map_zones_
```

```
- generic cluster
  - i06r01c20
    - MPI Rank 0
        0.34 Thread 0
        0.00 Thread 1
        0.00 Thread 2
        0.00 Thread 3
    - MPI Rank 1
        0.39 Thread 0
        0.00 Thread 1
        0.00 Thread 2
        0.00 Thread 3
    - MPI Rank 2
        0.43 Thread 0
        0.00 Thread 1
        0.00 Thread 2
        0.00 Thread 3
    - MPI Rank 3
        0.21 Thread 0
        0.00 Thread 1
        0.00 Thread 2
        0.00 Thread 3
```

```
0.00    1.38 (0.41%)    337.45 | 0.00    1.38 (100.00%)    1.38 | 0.00    1.38
```

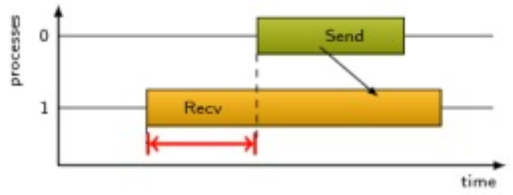
Shows the online description of the clicked item

Online metric description



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

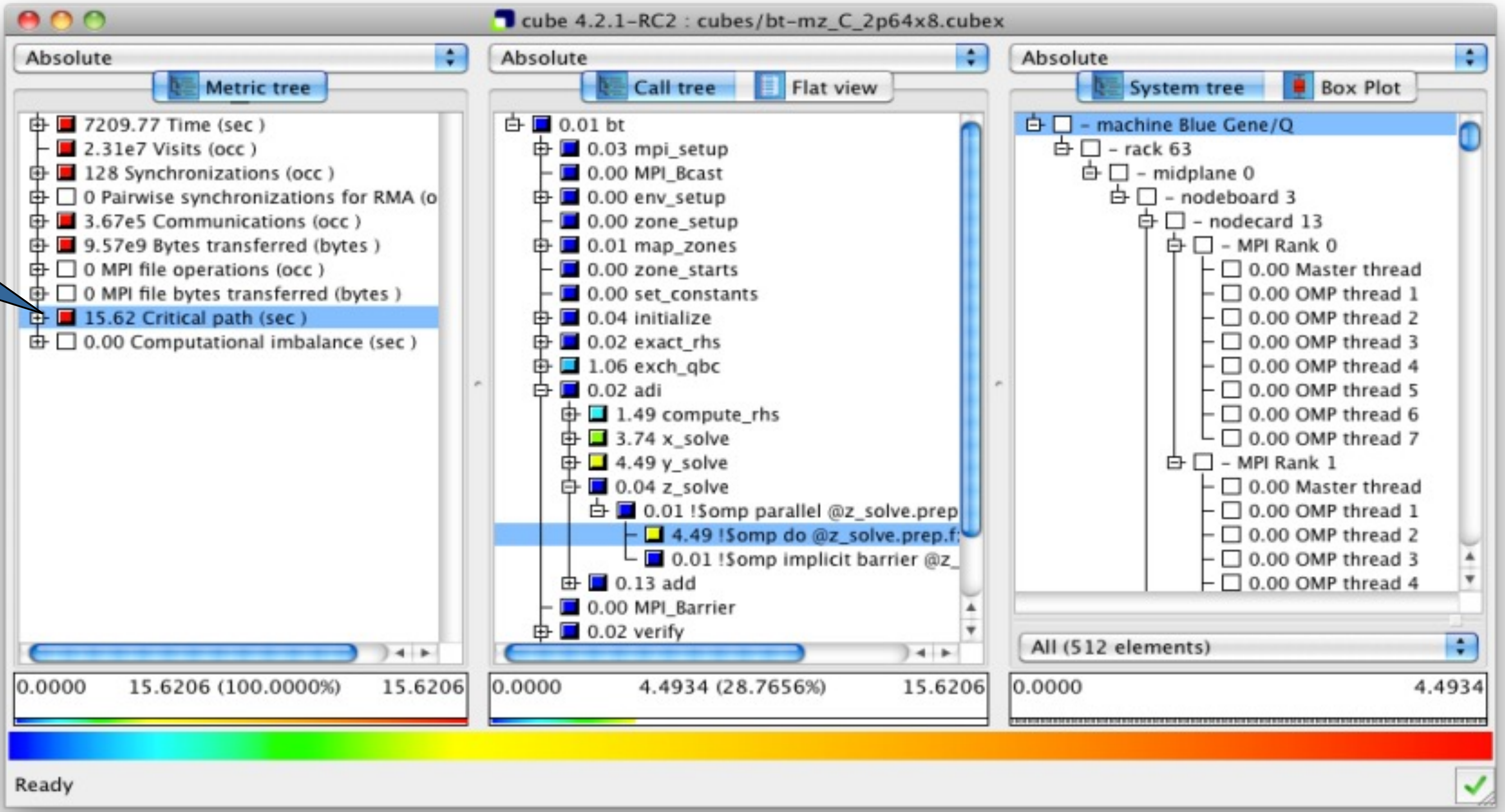
Diagnosis:
Try to replace `MPI_Recv` with a non-blocking receive `MPI_Irecv` that can be posted earlier, proceed concurrently with computation, and complete with a wait operation after the message is expected to have been sent. Try to post sends earlier, such that they are available when receivers need 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:

Critical-path analysis

Critical-path profile shows wall-clock time impact


cube 4.2.1-RC2 : cubes/bt-mz_C_2p64x8.cubex

Absolute — Metric tree

- 7209.77 Time (sec)
- 2.31e7 Visits (occ)
- 128 Synchronizations (occ)
- 0 Pairwise synchronizations for RMA (o
- 3.67e5 Communications (occ)
- 9.57e9 Bytes transferred (bytes)
- 0 MPI file operations (occ)
- 0 MPI file bytes transferred (bytes)
- 15.62 Critical path (sec)
- 0.00 Computational imbalance (sec)

0.0000 15.6206 (100.0000%) 15.6206

Absolute — Call tree | Flat view

```
0.01 bt
  0.03 mpi_setup
  0.00 MPI_Bcast
  0.00 env_setup
  0.00 zone_setup
  0.01 map_zones
  0.00 zone_starts
  0.00 set_constants
  0.04 initialize
  0.02 exact_rhs
  1.06 exch_qbc
  0.02 adi
    1.49 compute_rhs
    3.74 x_solve
    4.49 y_solve
    0.04 z_solve
      0.01 !$omp parallel @z_solve.prep
        4.49 !$omp do @z_solve.prep.f
        0.01 !$omp implicit barrier @z_
    0.13 add
  0.00 MPI_Barrier
  0.02 verify
```

0.0000 4.4934 (28.7656%) 15.6206

Absolute — System tree | Box Plot

```
- machine Blue Gene/Q
  - rack 63
    - midplane 0
      - nodeboard 3
        - nodecard 13
          - MPI Rank 0
            0.00 Master thread
            0.00 OMP thread 1
            0.00 OMP thread 2
            0.00 OMP thread 3
            0.00 OMP thread 4
            0.00 OMP thread 5
            0.00 OMP thread 6
            0.00 OMP thread 7
          - MPI Rank 1
            0.00 Master thread
            0.00 OMP thread 1
            0.00 OMP thread 2
            0.00 OMP thread 3
            0.00 OMP thread 4
```

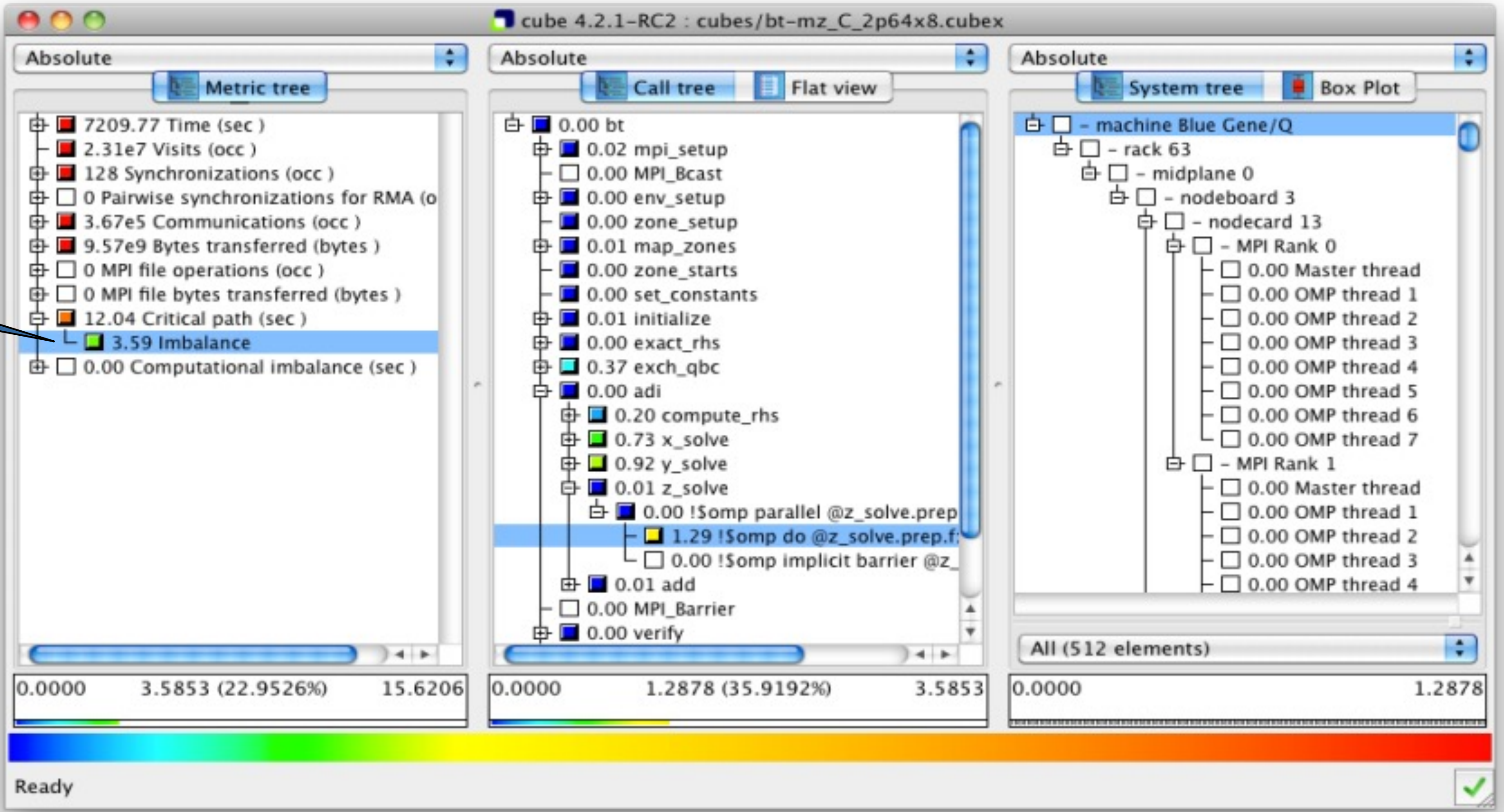
All (512 elements)

0.0000 4.4934

Ready

Critical-path analysis

Critical-path imbalance highlights inefficient parallelism


cube 4.2.1-RC2 : cubes/bt-mz_C_2p64x8.cubex

Absolute | Metric tree
```
7209.77 Time (sec )
2.31e7 Visits (occ )
128 Synchronizations (occ )
0 Pairwise synchronizations for RMA (o
3.67e5 Communications (occ )
9.57e9 Bytes transferred (bytes )
0 MPI file operations (occ )
0 MPI file bytes transferred (bytes )
12.04 Critical path (sec )
  3.59 Imbalance
0.00 Computational imbalance (sec )
```
0.0000   3.5853 (22.9526%)   15.6206

Absolute | Call tree | Flat view
```
0.00 bt
  0.02 mpi_setup
  0.00 MPI_Bcast
  0.00 env_setup
  0.00 zone_setup
  0.01 map_zones
  0.00 zone_starts
  0.00 set_constants
  0.01 initialize
  0.00 exact_rhs
  0.37 exch_qbc
  0.00 adi
    0.20 compute_rhs
    0.73 x_solve
    0.92 y_solve
    0.01 z_solve
      0.00 !$omp parallel @z_solve.prep
        1.29 !$omp do @z_solve.prep.f
        0.00 !$omp implicit barrier @z_
    0.01 add
  0.00 MPI_Barrier
  0.00 verify
```
0.0000   1.2878 (35.9192%)   3.5853

Absolute | System tree | Box Plot
```
- machine Blue Gene/Q
  - rack 63
    - midplane 0
      - nodeboard 3
        - nodecard 13
          - MPI Rank 0
            0.00 Master thread
            0.00 OMP thread 1
            0.00 OMP thread 2
            0.00 OMP thread 3
            0.00 OMP thread 4
            0.00 OMP thread 5
            0.00 OMP thread 6
            0.00 OMP thread 7
          - MPI Rank 1
            0.00 Master thread
            0.00 OMP thread 1
            0.00 OMP thread 2
            0.00 OMP thread 3
            0.00 OMP thread 4
```
All (512 elements)

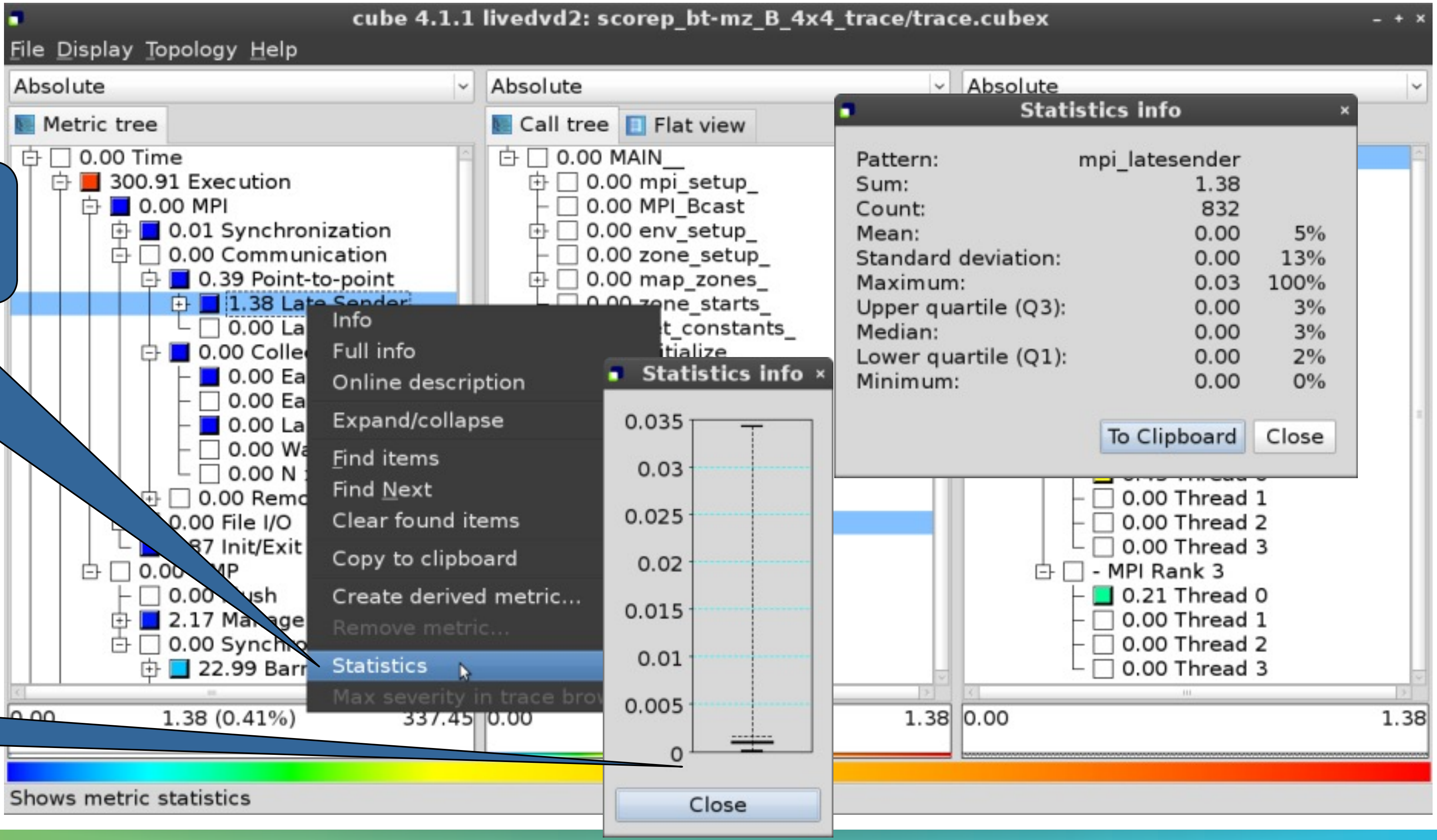
0.0000   1.2878

Ready

Pattern instance statistics

Access pattern instance statistics via context menu

Click to get statistics details



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%

To Clipboard Close

Scalasca: Reference material

Scalasca command – One command for (almost) everything

```
% scalasca
Scalasca 2.3.1
Toolset for scalable performance analysis of large-scale parallel applications
usage: scalasca [OPTION]... ACTION <argument>...
    1. prepare application objects and executable for measurement:
       scalasca -instrument <compile-or-link-command> # skin (using scorep)
    2. run application under control of measurement system:
       scalasca -analyze <application-launch-command> # scan
    3. interactively explore measurement analysis report:
       scalasca -examine <experiment-archive|report>  # square

Options:
   -c, --show-config     show configuration summary and exit
   -h, --help            show this help and exit
   -n, --dry-run         show actions without taking them
       --quickref        show quick reference guide and exit
       --remap-specfile  show path to remapper specification file and exit
   -v, --verbose         enable verbose commentary
   -V, --version         show version information and exit
```

- The '`scalasca -instrument`' command is deprecated and only provided for backwards compatibility with Scalasca 1.x., recommended: use Score-P instrumenter directly

Scalasca compatibility command: skin / scalasca -instrument

```
% skin
Scalasca 2.3.1: application instrumenter (using Score-P instrumenter)
usage: skin [-v] [-comp] [-pdt] [-pomp] [-user] [--*] <compile-or-link-command>
  -comp={all|none|...}: routines to be instrumented by compiler [default: all]
                 (... custom instrumentation specification depends on compiler)
  -pdt:  process source files with PDT/TAU instrumenter
  -pomp: process source files for POMP directives
  -user: enable EPIK user instrumentation API macros in source code
  -v:    enable verbose commentary when instrumenting

  --*:   options to pass to Score-P instrumenter
```

- Scalasca application instrumenter
  - Provides compatibility with Scalasca 1.x
  - **Deprecated! Use Score-P instrumenter directly.**

Scalasca convenience command: scan / scalasca -analyze

```
% scan
Scalasca 2.3.1: 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 filtfile  : File specifying measurement filter.
  -l lockfile  : File that blocks start of measurement.
  -m metrics   : Metric specification for measurement.
```

- Scalasca measurement collection & analysis nexus

Scalasca advanced command:
scout - Scalasca automatic trace analyzer

```
% scout.hyb --help
SCOUT   Copyright (c) 1998-2016 Forschungszentrum Juelich GmbH
        Copyright (c) 2009-2014 German Research School for Simulation
                                Sciences GmbH

Usage: <launchcmd> scout.hyb [OPTION]... <ANCHORFILE | EPIK DIRECTORY>
Options:
  --statistics       Enables instance tracking and statistics [default]
  --no-statistics    Disables instance tracking and statistics
  --critical-path    Enables critical-path analysis [default]
  --no-critical-path Disables critical-path analysis
  --rootcause        Enables root-cause analysis [default]
  --no-rootcause     Disables root-cause analysis
  --single-pass      Single-pass forward analysis only
  --time-correct     Enables enhanced timestamp correction
  --no-time-correct  Disables enhanced timestamp correction [default]
  --verbose, -v      Increase verbosity
  --help             Display this information and exit
```

- Provided in serial (.ser), OpenMP (.omp), MPI (.mpi) and MPI+OpenMP (.hyb) variants

Scalasca advanced command: clc_synchronize

- Scalasca trace event timestamp consistency correction

```
Usage: <launchcmd> clc_synchronize.hyb <ANCHORFILE | EPIK_DIRECTORY>
```

  - Provided in MPI (.mpi) and MPI+OpenMP (.hyb) variants
  - Takes as input a trace experiment archive where the events may have timestamp inconsistencies
    - E.g., multi-node measurements on systems without adequately synchronized clocks on each compute node
  - Generates a new experiment archive (always called ./clc_sync) containing a trace with event timestamp inconsistencies resolved
    - E.g., suitable for detailed examination with a time-line visualizer

Scalasca convenience command: square / scalasca -examine

```
% square
Scalasca 2.3.1: analysis report explorer
usage: square [-v] [-s] [-f filtfile] [-F] <experiment archive | cube file>
   -c <none | quick | full> : Level of sanity checks for newly created reports
   -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
   -n                       : Do not include idle thread metric
```

- Scalasca analysis report explorer (Cube)

Automatic measurement configuration

- scan configures Score-P measurement by automatically setting some environment variables and exporting them
  - E.g., experiment title, profiling/tracing mode, filter file, …
  - Precedence order:
    - Command-line arguments
    - Environment variables already set
    - Automatically determined values
- Also, scan includes consistency checks and prevents corrupting existing experiment directories
- For tracing experiments, after trace collection completes then automatic parallel trace analysis is initiated
  - Uses identical launch configuration to that used for measurement (i.e., the same allocated compute resources)

Further information

Scalable performance **a**nalysis of **la**rge-**sc**ale parallel **a**pplications

- Toolset for scalable performance measurement & analysis of MPI, OpenMP & hybrid parallel applications
- Supporting most popular HPC computer systems
- Available under 3-clause BSD open-source license
- Sources, documentation & publications:
  - http://www.scalasca.org
  - mailto: scalasca@fz-juelich.de