Automatic trace analysis with the Scalasca Trace Tools

Markus Geimer & Brian Wylie
Jülich Supercomputing Centre

trace tools
scalasca

Scalasca Trace Tools

DOI 10.5281/zenodo.4103922

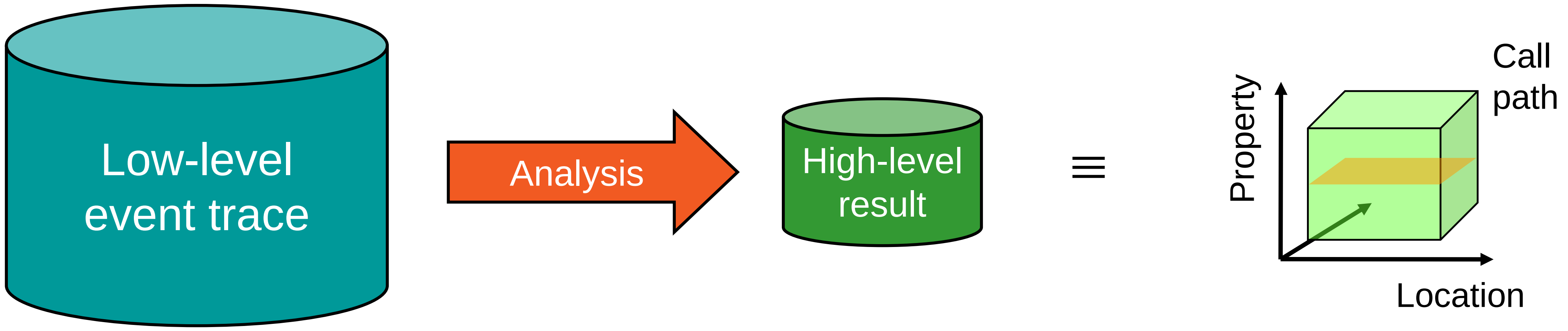
- **Scalable trace-based** performance analysis toolset for the most popular parallel programming paradigms
  - Current focus: MPI, OpenMP, and (to a limited extend) POSIX threads
  - Analysis of traces including *only host-side events* from applications using CUDA, OpenCL, or OpenACC (also in combination with MPI and/or OpenMP) is possible, but results need to be interpreted with some care

- Specifically targeting large-scale parallel applications
  - Demonstrated scalability up to 1.8 million parallel threads
  - Of course also works at small/medium scale

- Latest release:
  - Scalasca Trace Tools v2.6.1 (Dec 2022)

Automatic trace analysis

- Idea
  - Automatic search for patterns of inefficient behaviour
  - Classification of behaviour & quantification of significance
  - Identification of delays as root causes of inefficiencies



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

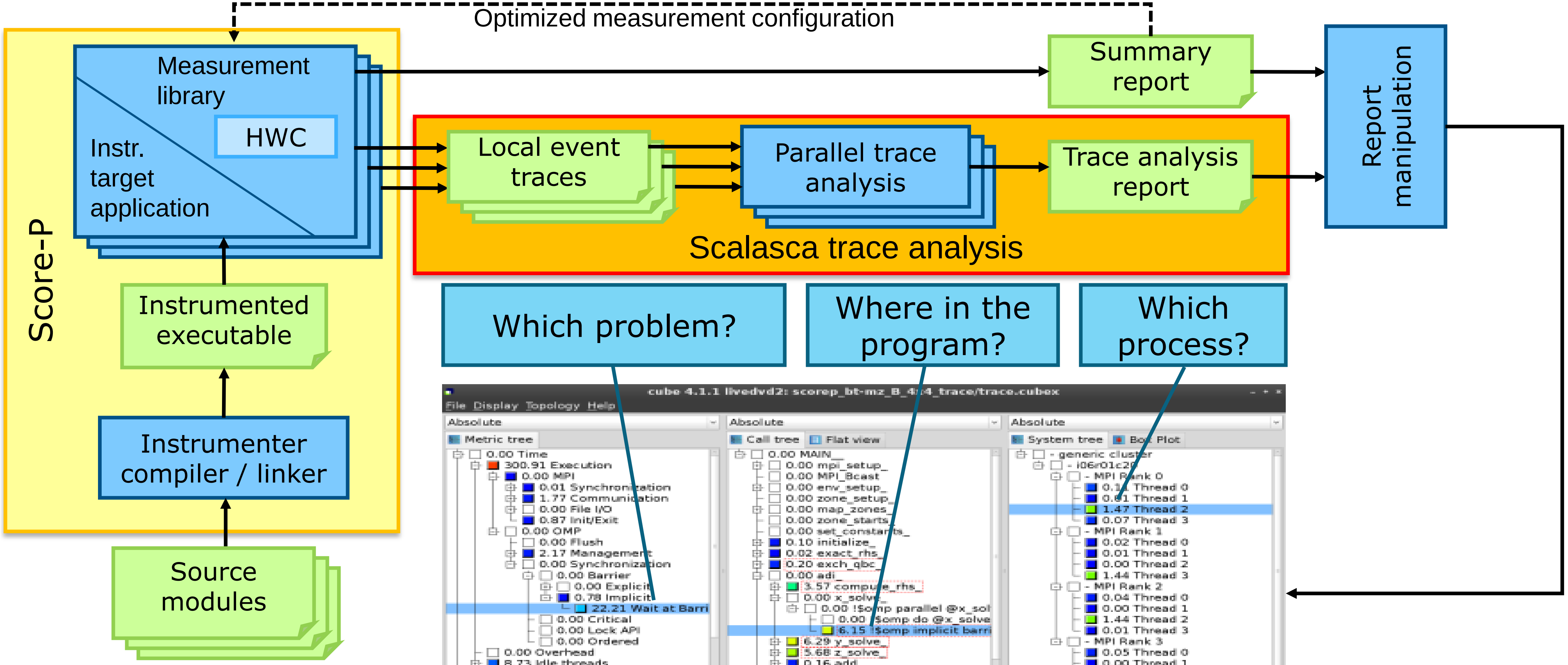
Scalasca Trace Tools: Features

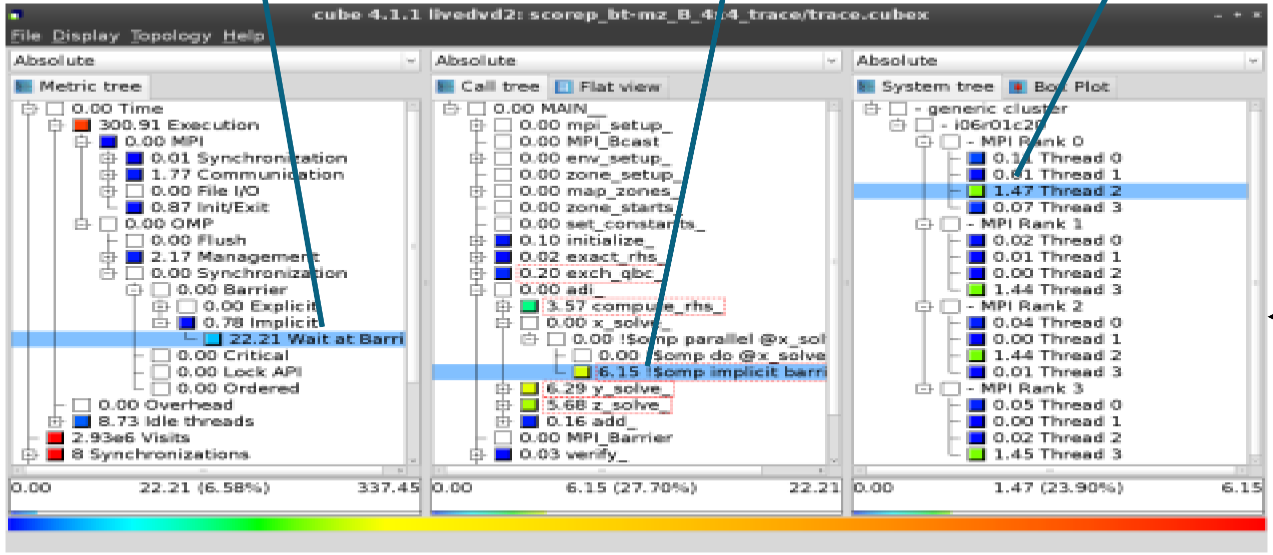
- Open source, 3-clause BSD license
- Supports all major HPC platforms

- Uses Score-P instrumenter & measurement libraries
  - Scalasca v2 core package focuses on trace-based analyses
  - Provides convenience commands for measurement, analysis, and post-processing
  - Supports common data formats
    - Reads event traces in OTF2 format
    - Writes analysis reports in CUBE4 format

- Current limitations:
  - Unable to handle traces …
    - with MPI thread level exceeding MPI_THREAD_FUNNELED
    - containing memory events, CUDA/OpenCL device events (kernel, memcpy), SHMEM, or OpenMP nested parallelism
  - PAPI/rusage metrics for trace events are ignored

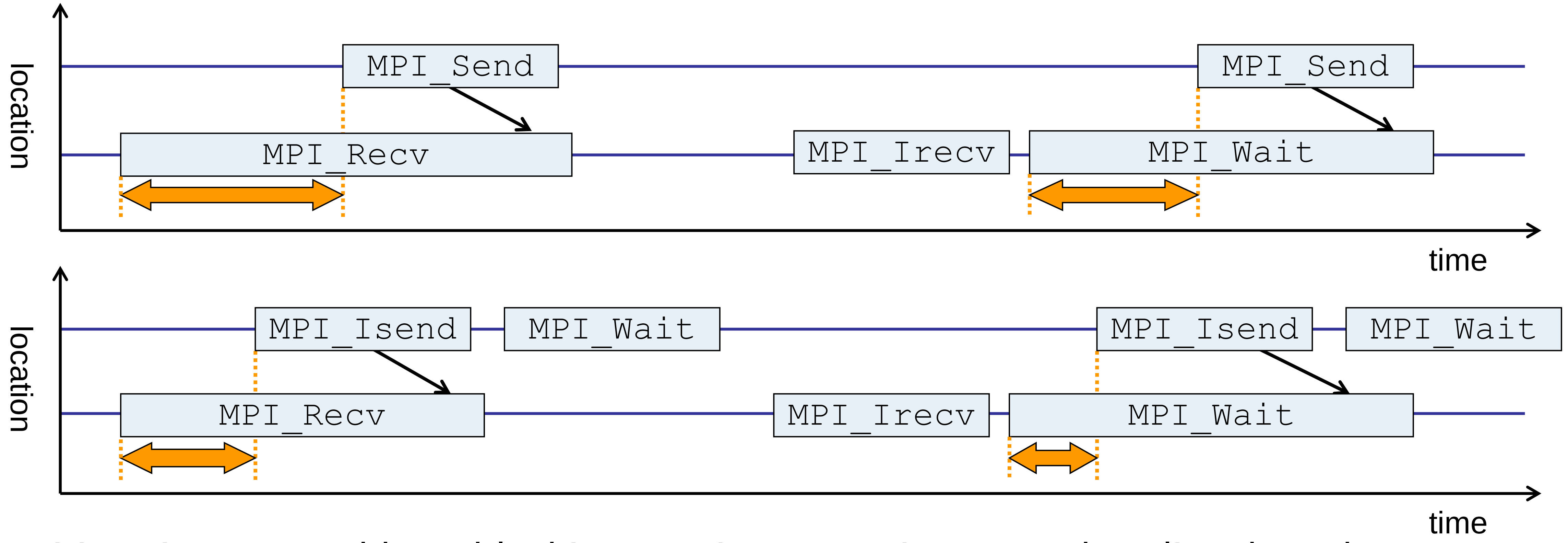
Scalasca workflow





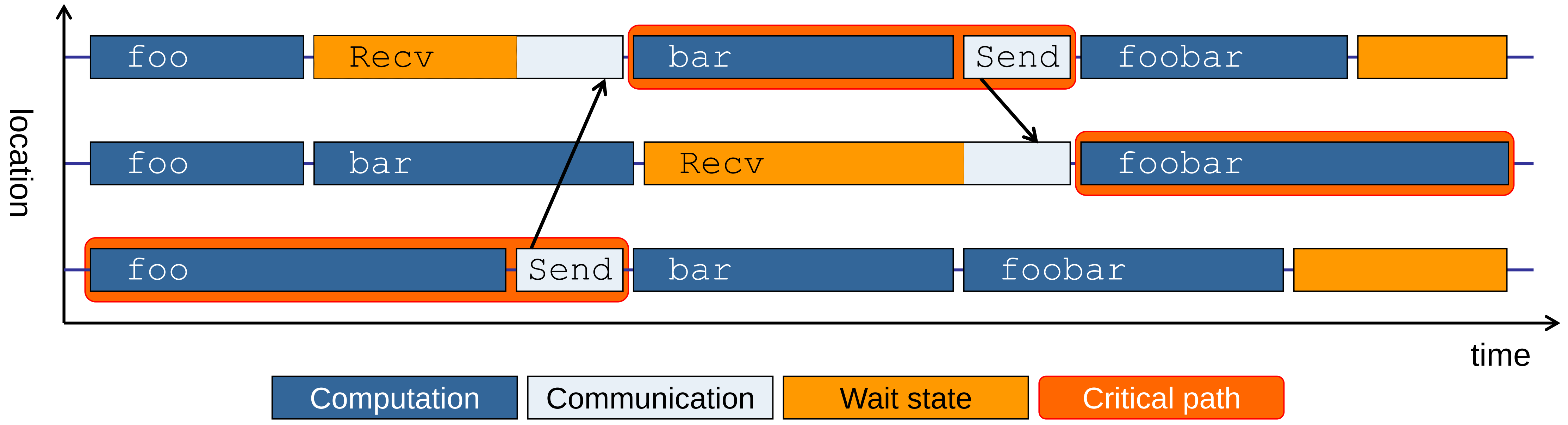
```
cube 4.1.1 livedvd2: scorep_bt-mz_B_4x4_trace/trace.cubex
File  Display  Topology  Help

Absolute
Metric tree
0.00 Time
  300.91 Execution
    0.00 MPI
      0.01 Synchronization
      1.77 Communication
      0.00 File I/O
      0.87 Init/Exit
    0.00 OMP
      0.00 Flush
      2.17 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          0.78 Implicit
            22.21 Wait at Barri
        0.00 Critical
        0.00 Lock API
        0.00 Ordered
    0.00 Overhead
    8.73 Idle threads
  2.93e6 Visits
  8 Synchronizations
0.00          22.21 (6.58%)          337.45

Absolute
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_
  0.10 initialize_
  0.02 exact_rhs_
  0.20 exch_qbc_
  0.00 adi_
    3.57 compute_rhs_
    0.00 x_solve_
      0.00 !$omp parallel @x_sol
        0.00 !$omp do @x_solve
        6.15 !$omp implicit barri
    6.29 y_solve_
    5.68 z_solve_
    0.16 add_
  0.00 MPI_Barrier
  0.03 verify_
0.00          6.15 (27.70%)          22.21

Absolute
System tree   Box Plot
- generic cluster
  - i06r01c20
    - MPI Rank 0
      0.11 Thread 0
      0.01 Thread 1
      1.47 Thread 2
      0.07 Thread 3
    - MPI Rank 1
      0.02 Thread 0
      0.01 Thread 1
      0.00 Thread 2
      1.44 Thread 3
    - MPI Rank 2
      0.04 Thread 0
      0.00 Thread 1
      1.44 Thread 2
      0.01 Thread 3
    - MPI Rank 3
      0.05 Thread 0
      0.00 Thread 1
      0.02 Thread 2
      1.45 Thread 3
0.00          1.47 (23.90%)          6.15
```

Example: "*Late Sender*" wait state



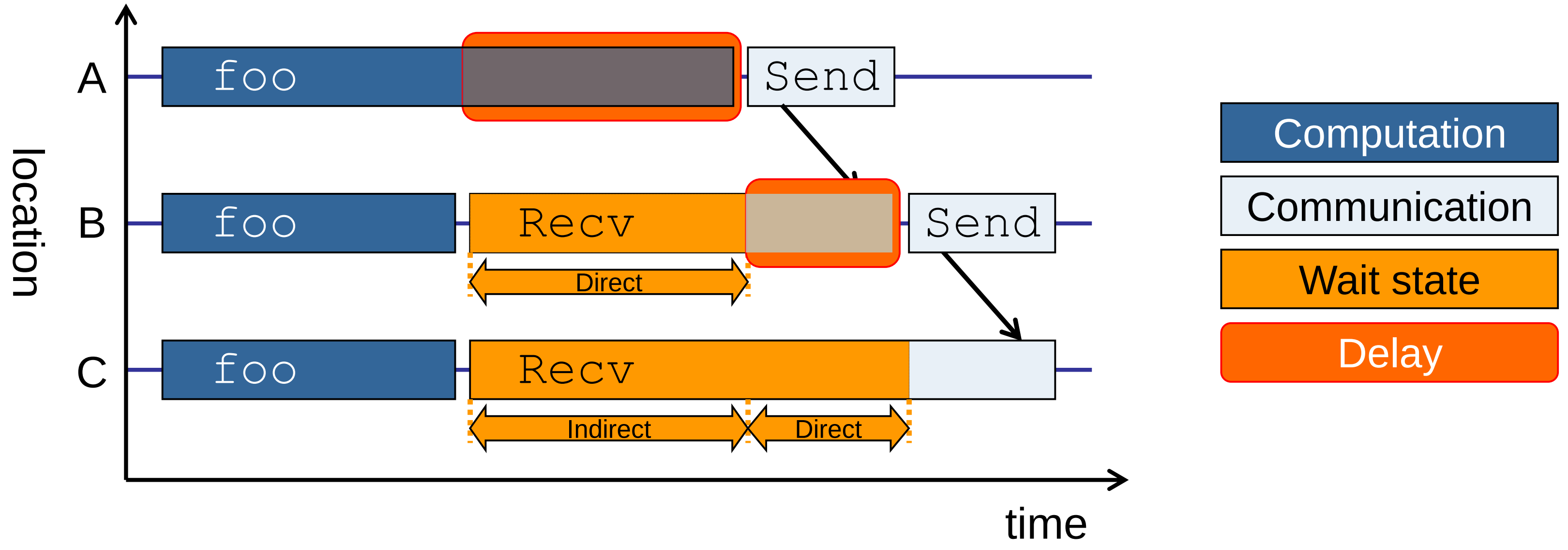
- Waiting time caused by a blocking receive operation posted earlier than the corresponding send
- Applies to blocking as well as non-blocking communication

Example: Critical path



- Shows call paths and processes/threads that are responsible for the program’s wall-clock runtime
- Identifies good optimization candidates and parallelization bottlenecks

Example: Root-cause analysis



- Classifies wait states into direct and indirect (i.e., caused by other wait states)
- Identifies *delays* (excess computation/communication) as root causes of wait states
- Attributes wait states as *delay costs*

Hands-on: CloverLeaf MPI+OpenACC

trace tools
scalasca

Recap: Setup for exercises

- Connect to your account on Turpan (with X11-forwarding)

```
% ssh -X <yourid>@turpan
```

- Set account and default environment (NVHPC + OpenMPI) via helper script

```
% source /tmpdir/vi-hps/opt/setup.sh
```

- Change to directory containing CloverLeaf_OpenACC sources
  - Existing instrumented executable can be reused

```
% cd $WORK
% cd CloverLeaf_OpenACC
```

- Load Scalasca module
  - Depends on (i.e., implicitly loads) Score-P & CubeGUI

```
% module load scalasca
```

CloverLeaf_OpenACC summary measurement collection...

```
% cd bin.scorep
% cat scan.sbatch
...
module load scalasca

# Score-P measurement configuration
export SCOREP_OPENACC_ENABLE=regions,wait,enqueue
export SCOREP_CUDA_ENABLE=default
#export SCOREP_CUDA_ENABLE=none
#export SCOREP_CUDA_ENABLE=kernel,kernel_callsite,idle
#export SCOREP_ENABLE_TRACING=true

# Scalasca analysis configuration
export SCAN_ANALYZE_OPTS="--time-correct"

# Run the application
scan -s  mpiexec ./clover_leaf
```

- Change to directory with the Score-P instrumented executable and edit the job script

Hint:
scan = scalasca –analyze
-s = profile/summary (default)

```
% sbatch scan.sbatch
```

- Submit the job

CloverLeaf_OpenACC summary measurement

```
S=C=A=N: Scalasca 2.6.1 runtime summarization
S=C=A=N: ./scorep_clover_leaf_6_sum experiment archive
S=C=A=N: Wed Oct 25 14:51:20 2023: Collect start
mpiexec ./clover_leaf

   Clover version    1.400

 [... More application output ...]

S=C=A=N: Wed Oct 25 14:52:09 2023: Collect done (status=0) 49s
S=C=A=N: ./scorep_clover_leaf_6_sum complete.
```

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

- Creates experiment directory:
  `scorep_clover_leaf_6_sum`

CloverLeaf_OpenACC summary analysis report examination

- Score summary analysis report

```
% square -s  scorep_clover_leaf_6_sum
INFO: Post-processing runtime summarization result (profile.cubex)...
INFO: Score report written to ./scorep_clover_leaf_6_sum/scorep.score
```

- Post-processing and interactive exploration with Cube

```
% square  scorep_clover_leaf_6_sum
INFO: Displaying ./scorep_clover_leaf_6_sum/summary.cubex...

                         [GUI showing summary analysis report]
```

Hint:
Copy ‘summary.cubex’ to local system (laptop) using ‘scp’ to improve responsiveness of GUI

- The post-processing derives additional metrics and generates a structured metric hierarchy

CloverLeaf_OpenACC trace measurement collection...

```
% cd bin.scorep
% edit scan.sbatch
...
module load scalasca

# Score-P measurement configuration
export SCOREP_OPENACC_ENABLE=regions,wait,enqueue
export SCOREP_CUDA_ENABLE=runtime
#export SCOREP_CUDA_ENABLE=none
#export SCOREP_CUDA_ENABLE=kernel,kernel_callsite,idle
#export SCOREP_ENABLE_TRACING=true

# Scalasca analysis configuration
export SCAN_ANALYZE_OPTS="--time-correct"

# Run the application
scan -t  mpiexec ./clover_leaf
```

- Change to directory with the Score-P instrumented executable and edit the job script

Hint:
`scan` = `scalasca –analyze`
`-t` = trace collection & analysis

```
% sbatch scan.sbatch
```

- Submit the job

CloverLeaf_OpenACC trace measurement ... collection

```
S=C=A=N: Scalasca 2.6.1 trace collection and analysis
S=C=A=N: Wed Oct 25 14:58:52 2023: Collect start
mpiexec ./clover_leaf


   Clover version    1.400

 [... More application output ...]

S=C=A=N: Wed Oct 25 14:59:33 2023: Collect done (status=0) 41s
```

- Starts measurement with collection of trace files …

CloverLeaf_OpenACC trace measurement ... analysis

```
…
S=C=A=N: Wed Oct 25 14:59:33 2023: Analyze start
mpiexec scout.mpi --time-correct ./scorep_clover_leaf_6_trace/traces.otf2

SCOUT   (Scalasca 2.6.1)

Analyzing experiment archive ./scorep_clover_leaf_6_trace/traces.otf2

Opening experiment archive ... done (0.009s).
Reading definition data    ... done (0.008s).
Reading event trace data   ... done (0.695s).
Preprocessing              ... done (0.721s).
Timestamp correction       ... done (0.711s).
Analyzing trace data       ... done (10.108s).
Writing analysis report    ... done (0.137s).

Max. memory usage          : 426.422MB

        # passes         : 1
        # violated       : 0

Total processing time      : 12.489s
S=C=A=N: Wed Oct 25 14:59:49 2023: Analyze done (status=0) 16s
```

- Continues with automatic (parallel) analysis of trace files

CloverLeaf_OpenACC trace analysis report exploration

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

```
% square  scorep_clover_leaf_6_trace
INFO: Post-processing runtime summarization report (profile.cubex)...
INFO: Post-processing trace analysis report (scout.cubex)...
INFO: Displaying ./scorep_clover_leaf_6_trace/trace.cubex...

                [GUI showing trace analysis report]
```

Hint:
Run ‘square -s’ first and then copy ‘trace.cubex’ to local system (laptop) using ‘scp’ to improve responsiveness of GUI

Demo:
TeaLeaf MPI+OpenMP case study

Case study: TeaLeaf MPI+OpenMP

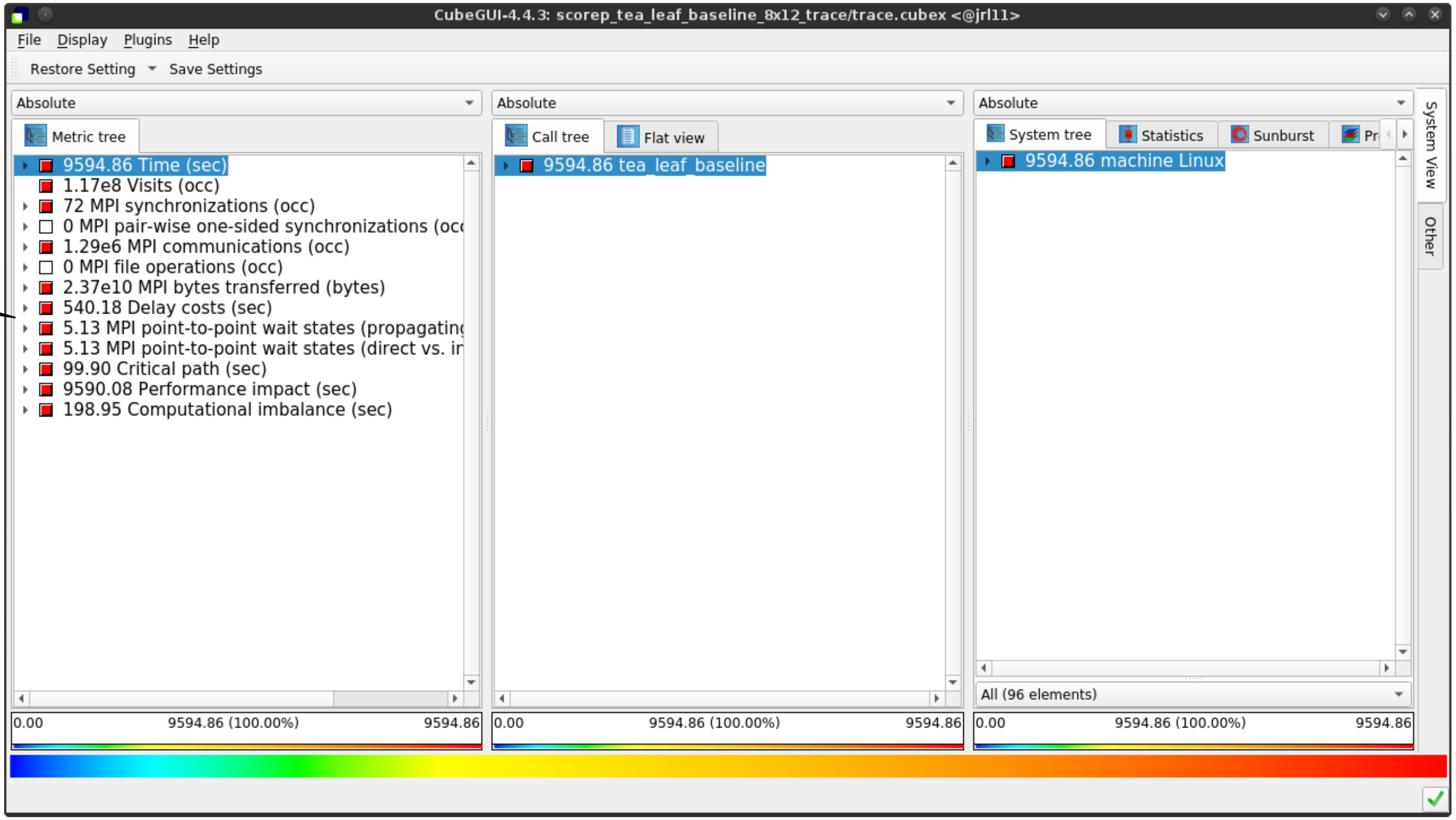
- HPC mini-app developed by the UK Mini-App Consortium
  - Solves the linear 2D heat conduction equation on a spatially decomposed regular grid using a 5 point stencil with implicit solvers
  - Part of the Mantevo 3.0 suite
  - Available on GitHub: https://uk-mac.github.io/TeaLeaf/

- Measurements of TeaLeaf reference v1.0 taken on Jureca cluster @ JSC
  - Using Intel 19.0.3 compilers, Intel MPI 2019.3, Score-P 5.0, and Scalasca 2.5
  - Run configuration
    - 8 MPI ranks with 12 OpenMP threads each
    - Distributed across 4 compute nodes (2 ranks per node)
    - Test problem “5”: 4000 × 4000 cells, CG solver

```
% cube scorep_tea_leaf_baseline_8x12_trace/trace.cubex
                                  [GUI showing post-processed trace analysis report]
```

Scalasca analysis report exploration (opening view)

Additional top-level metrics produced by the trace analysis…


CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>
File Display Plugins Help
Restore Setting Save Settings
Absolute
Metric tree
Call tree
Flat view
System tree
Statistics
Sunburst
System View
Other
All (96 elements)

```
9594.86 Time (sec)
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizations (occ
1.29e6 MPI communications (occ)
0 MPI file operations (occ)
2.37e10 MPI bytes transferred (bytes)
540.18 Delay costs (sec)
5.13 MPI point-to-point wait states (propagatinç
5.13 MPI point-to-point wait states (direct vs. ir
99.90 Critical path (sec)
9590.08 Performance impact (sec)
198.95 Computational imbalance (sec)
```

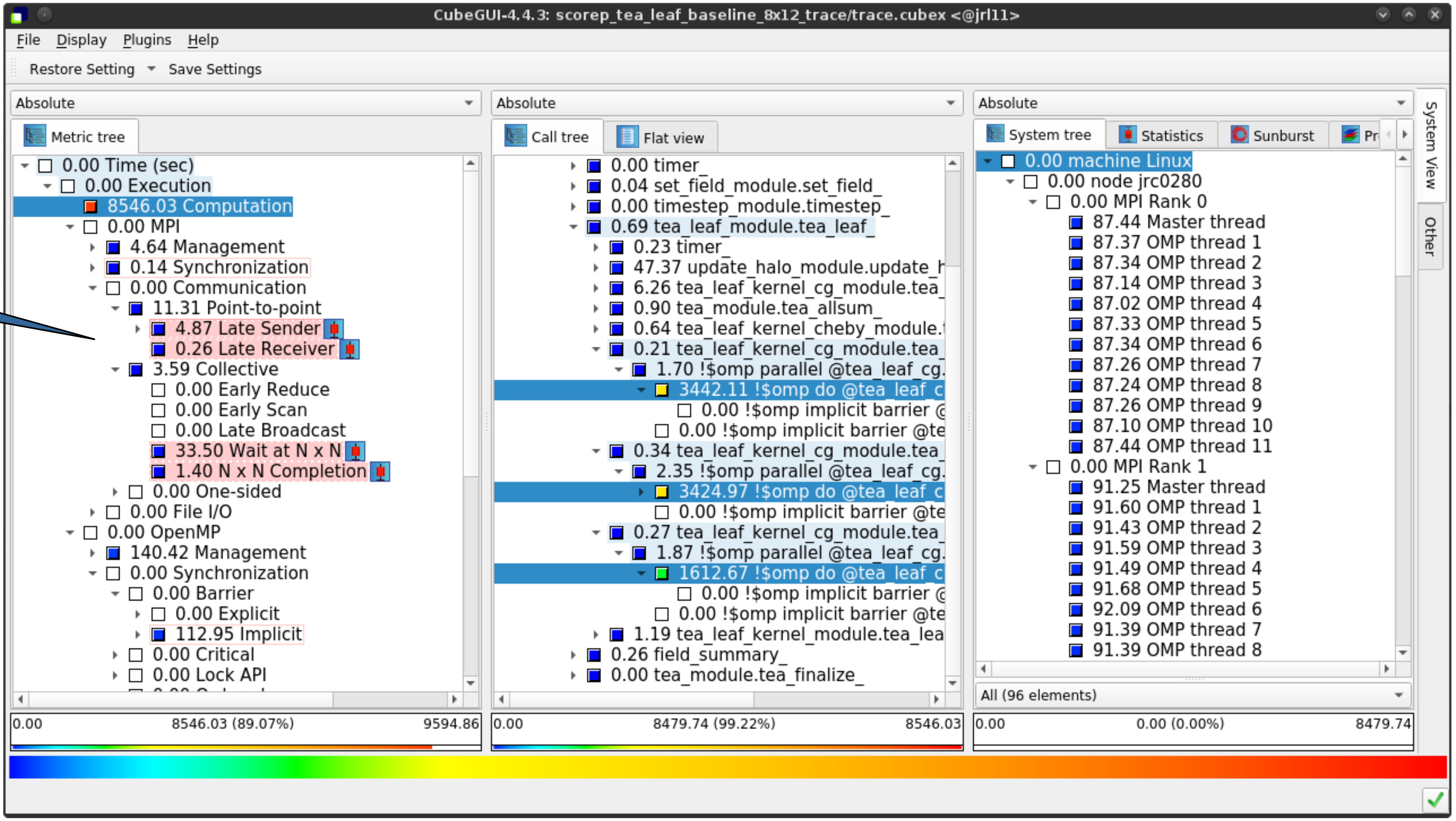
```
9594.86 tea_leaf_baseline
```

```
9594.86 machine Linux
```

```
0.00    9594.86 (100.00%)    9594.86
```

Scalasca wait-state metrics

…plus additional wait-state metrics as part of the “Time” hierarchy

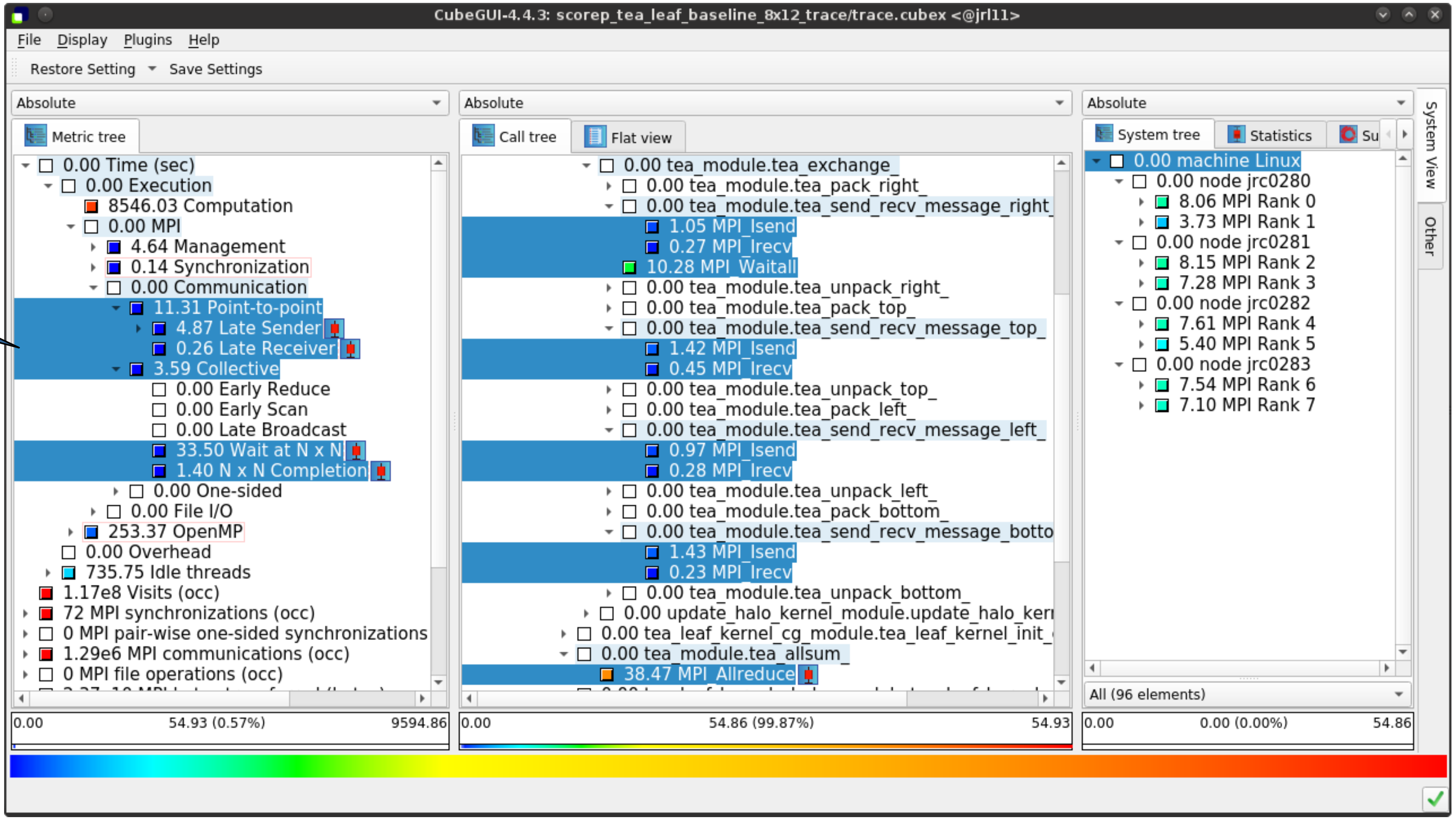


CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

```
Metric tree                          Call tree                                      System tree
0.00 Time (sec)                      0.00 timer_                                    0.00 machine Linux
  0.00 Execution                     0.04 set_field_module.set_field_                 0.00 node jrc0280
    8546.03 Computation              0.00 timestep_module.timestep_                     0.00 MPI Rank 0
    0.00 MPI                         0.69 tea_leaf_module.tea_leaf_                       87.44 Master thread
      4.64 Management                  0.23 timer_                                        87.37 OMP thread 1
      0.14 Synchronization             47.37 update_halo_module.update_h                  87.34 OMP thread 2
      0.00 Communication               6.26 tea_leaf_kernel_cg_module.tea_                87.14 OMP thread 3
        11.31 Point-to-point           0.90 tea_module.tea_allsum_                        87.02 OMP thread 4
          4.87 Late Sender             0.64 tea_leaf_kernel_cheby_module.                 87.33 OMP thread 5
          0.26 Late Receiver           0.21 tea_leaf_kernel_cg_module.tea_                87.34 OMP thread 6
        3.59 Collective                  1.70 !$omp parallel @tea_leaf_cg.                87.26 OMP thread 7
          0.00 Early Reduce                3442.11 !$omp do @tea_leaf_c                   87.24 OMP thread 8
          0.00 Early Scan                    0.00 !$omp implicit barrier @                87.26 OMP thread 9
          0.00 Late Broadcast              0.00 !$omp implicit barrier @te                87.10 OMP thread 10
          33.50 Wait at N x N          0.34 tea_leaf_kernel_cg_module.tea_                87.44 OMP thread 11
          1.40 N x N Completion          2.35 !$omp parallel @tea_leaf_cg.              0.00 MPI Rank 1
        0.00 One-sided                     3424.97 !$omp do @tea_leaf_c                   91.25 Master thread
      0.00 File I/O                        0.00 !$omp implicit barrier @te                91.60 OMP thread 1
    0.00 OpenMP                        0.27 tea_leaf_kernel_cg_module.tea_                91.43 OMP thread 2
      140.42 Management                  1.87 !$omp parallel @tea_leaf_cg.                91.59 OMP thread 3
      0.00 Synchronization                 1612.67 !$omp do @tea_leaf_c                   91.49 OMP thread 4
        0.00 Barrier                         0.00 !$omp implicit barrier @                91.68 OMP thread 5
          0.00 Explicit                    0.00 !$omp implicit barrier @te                92.09 OMP thread 6
          112.95 Implicit              1.19 tea_leaf_kernel_module.tea_lea                91.39 OMP thread 7
        0.00 Critical                0.26 field_summary_                                  91.39 OMP thread 8
        0.00 Lock API                0.00 tea_module.tea_finalize_

0.00   8546.03 (89.07%)   9594.86   0.00   8479.74 (99.22%)   8546.03   0.00   0.00 (0.00%)   8479.74
```

TeaLeaf Scalasca report analysis (I)

While MPI communication time and wait states are small (~0.6% of the total execution time)…



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

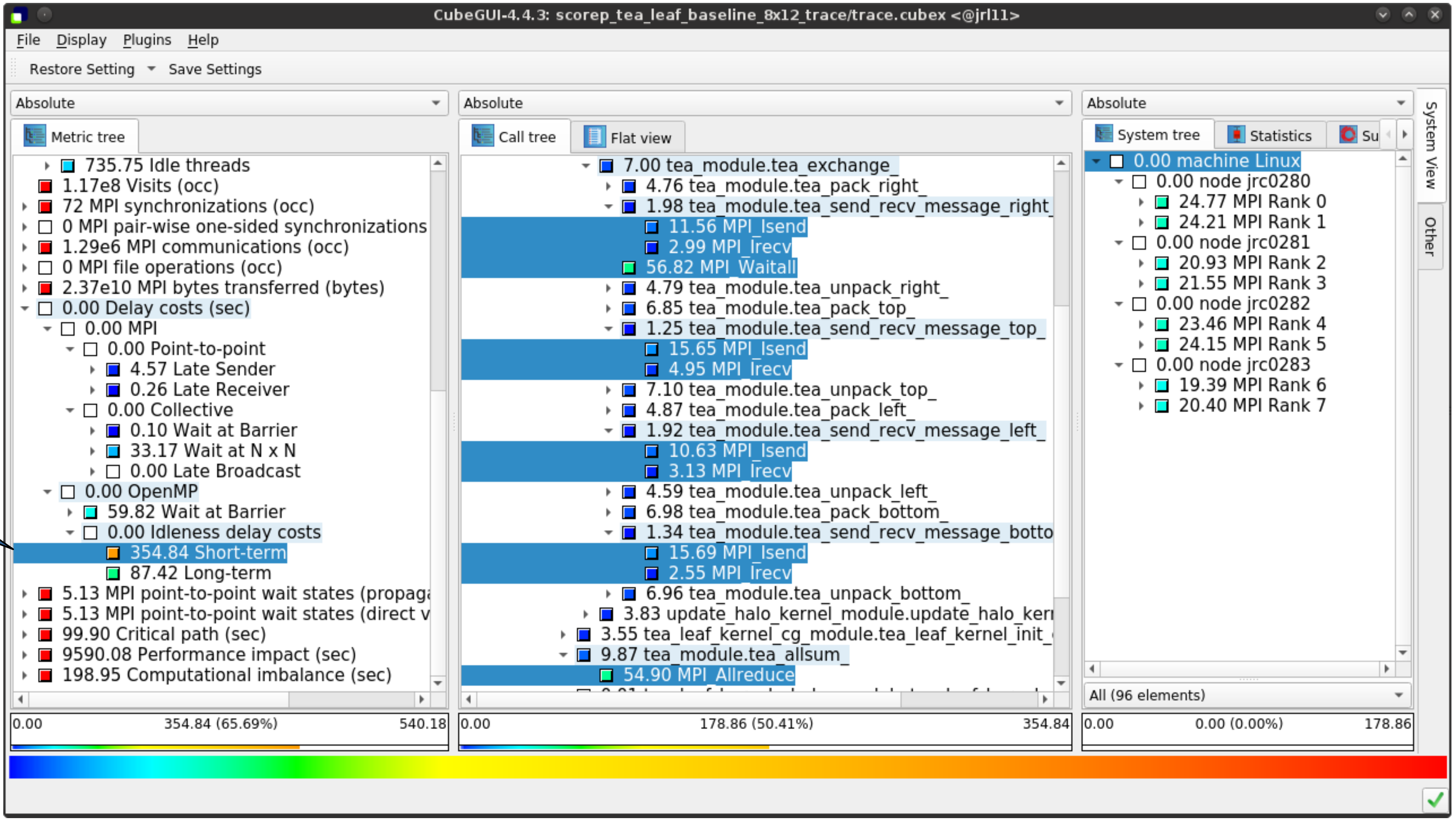
```
Absolute
Metric tree
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Early Reduce
          0.00 Early Scan
          0.00 Late Broadcast
          33.50 Wait at N x N
          1.40 N x N Completion
        0.00 One-sided
      0.00 File I/O
    253.37 OpenMP
  0.00 Overhead
  735.75 Idle threads
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizations
1.29e6 MPI communications (occ)
0 MPI file operations (occ)
0.00        54.93 (0.57%)        9594.86
```

```
Absolute
Call tree | Flat view
0.00 tea_module.tea_exchange_
  0.00 tea_module.tea_pack_right_
  0.00 tea_module.tea_send_recv_message_right_
    1.05 MPI_Isend
    0.27 MPI_Irecv
    10.28 MPI_Waitall
  0.00 tea_module.tea_unpack_right_
  0.00 tea_module.tea_pack_top_
  0.00 tea_module.tea_send_recv_message_top_
    1.42 MPI_Isend
    0.45 MPI_Irecv
  0.00 tea_module.tea_unpack_top_
  0.00 tea_module.tea_pack_left_
  0.00 tea_module.tea_send_recv_message_left_
    0.97 MPI_Isend
    0.28 MPI_Irecv
  0.00 tea_module.tea_unpack_left_
  0.00 tea_module.tea_pack_bottom_
  0.00 tea_module.tea_send_recv_message_botto
    1.43 MPI_Isend
    0.23 MPI_Irecv
  0.00 tea_module.tea_unpack_bottom_
0.00 update_halo_kernel_module.update_halo_kern
0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
0.00 tea_module.tea_allsum_
  38.47 MPI_Allreduce
0.00        54.86 (99.87%)        54.93
```

```
Absolute
System tree | Statistics | Su
0.00 machine Linux
  0.00 node jrc0280
    8.06 MPI Rank 0
    3.73 MPI Rank 1
  0.00 node jrc0281
    8.15 MPI Rank 2
    7.28 MPI Rank 3
  0.00 node jrc0282
    7.61 MPI Rank 4
    5.40 MPI Rank 5
  0.00 node jrc0283
    7.54 MPI Rank 6
    7.10 MPI Rank 7
All (96 elements)
0.00        0.00 (0.00%)        54.86
```

TeaLeaf Scalasca report analysis (II)

…they directly cause a significant amount of the OpenMP thread idleness


CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>
File Display Plugins Help
Restore Setting Save Settings
Absolute
Metric tree
Call tree
Flat view
System tree
Statistics
System View
Other

```
735.75 Idle threads
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizations
1.29e6 MPI communications (occ)
0 MPI file operations (occ)
2.37e10 MPI bytes transferred (bytes)
0.00 Delay costs (sec)
  0.00 MPI
    0.00 Point-to-point
      4.57 Late Sender
      0.26 Late Receiver
    0.00 Collective
      0.10 Wait at Barrier
      33.17 Wait at N x N
      0.00 Late Broadcast
  0.00 OpenMP
    59.82 Wait at Barrier
    0.00 Idleness delay costs
      354.84 Short-term
      87.42 Long-term
5.13 MPI point-to-point wait states (propag
5.13 MPI point-to-point wait states (direct v
99.90 Critical path (sec)
9590.08 Performance impact (sec)
198.95 Computational imbalance (sec)
```

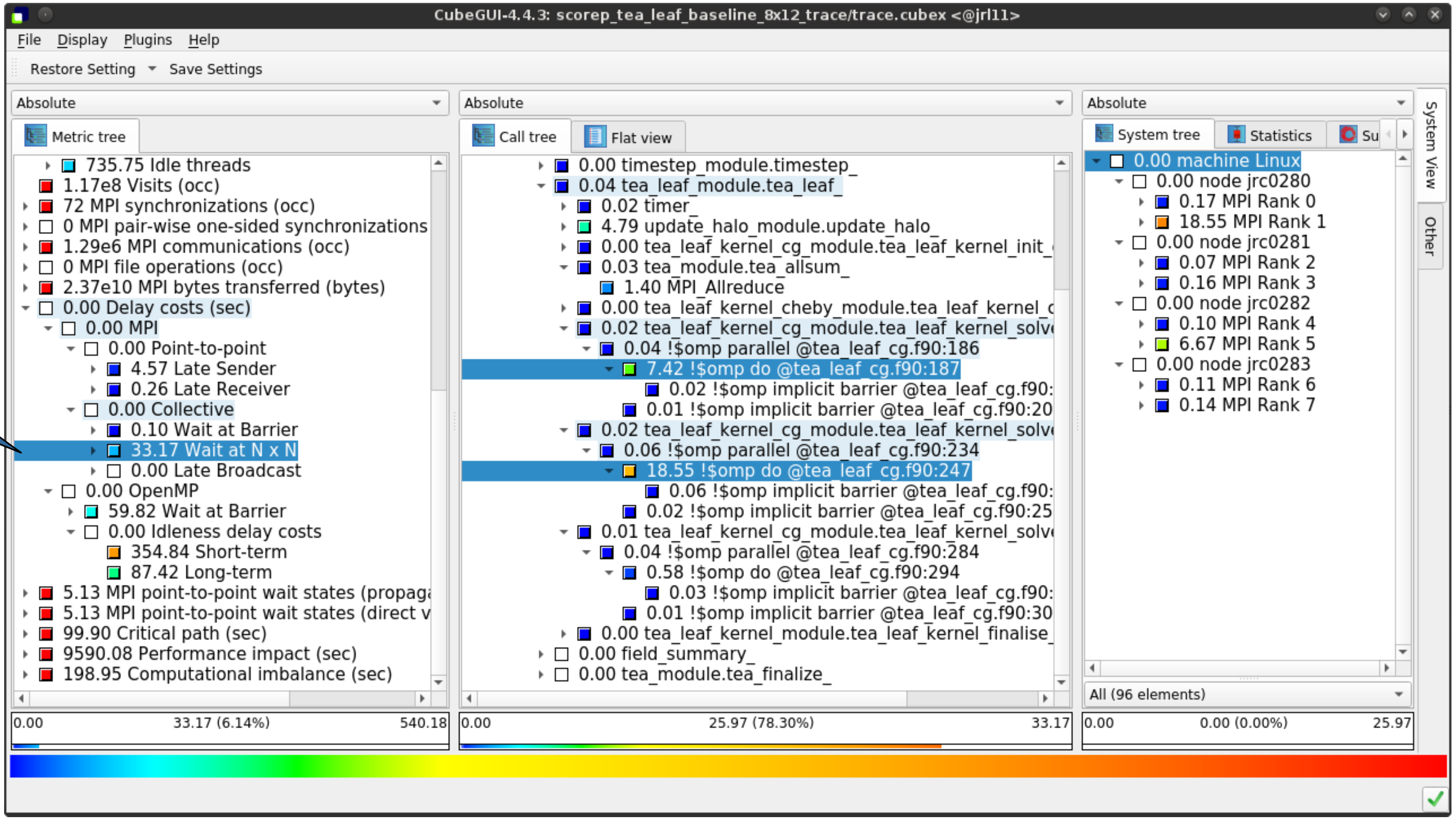
```
7.00 tea_module.tea_exchange_
  4.76 tea_module.tea_pack_right_
  1.98 tea_module.tea_send_recv_message_right_
    11.56 MPI_Isend
    2.99 MPI_Irecv
  56.82 MPI_Waitall
  4.79 tea_module.tea_unpack_right_
  6.85 tea_module.tea_pack_top_
  1.25 tea_module.tea_send_recv_message_top_
    15.65 MPI_Isend
    4.95 MPI_Irecv
  7.10 tea_module.tea_unpack_top_
  4.87 tea_module.tea_pack_left_
  1.92 tea_module.tea_send_recv_message_left_
    10.63 MPI_Isend
    3.13 MPI_Irecv
  4.59 tea_module.tea_unpack_left_
  6.98 tea_module.tea_pack_bottom_
  1.34 tea_module.tea_send_recv_message_botto
    15.69 MPI_Isend
    2.55 MPI_Irecv
  6.96 tea_module.tea_unpack_bottom_
3.83 update_halo_kernel_module.update_halo_kerı
3.55 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
9.87 tea_module.tea_allsum_
  54.90 MPI_Allreduce
```

```
0.00 machine Linux
  0.00 node jrc0280
    24.77 MPI Rank 0
    24.21 MPI Rank 1
  0.00 node jrc0281
    20.93 MPI Rank 2
    21.55 MPI Rank 3
  0.00 node jrc0282
    23.46 MPI Rank 4
    24.15 MPI Rank 5
  0.00 node jrc0283
    19.39 MPI Rank 6
    20.40 MPI Rank 7
All (96 elements)
```

```
0.00   354.84 (65.69%)   540.18
0.00   178.86 (50.41%)   354.84
0.00   0.00 (0.00%)   178.86
```

TeaLeaf Scalasca report analysis (III)

The "Wait at NxN" collective wait states are mostly caused by the first 2 OpenMP `do` loops of the solver (on ranks 5 & 1, resp.)…



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings

Absolute | Absolute | Absolute

Metric tree

```
  735.75 Idle threads
  1.17e8 Visits (occ)
  72 MPI synchronizations (occ)
  0 MPI pair-wise one-sided synchronizations
  1.29e6 MPI communications (occ)
  0 MPI file operations (occ)
  2.37e10 MPI bytes transferred (bytes)
  0.00 Delay costs (sec)
    0.00 MPI
      0.00 Point-to-point
        4.57 Late Sender
        0.26 Late Receiver
      0.00 Collective
        0.10 Wait at Barrier
        33.17 Wait at N x N
        0.00 Late Broadcast
    0.00 OpenMP
      59.82 Wait at Barrier
      0.00 Idleness delay costs
        354.84 Short-term
        87.42 Long-term
  5.13 MPI point-to-point wait states (propaga
  5.13 MPI point-to-point wait states (direct v
  99.90 Critical path (sec)
  9590.08 Performance impact (sec)
  198.95 Computational imbalance (sec)
```

Call tree | Flat view

```
  0.00 timestep_module.timestep_
  0.04 tea_leaf_module.tea_leaf_
    0.02 timer_
    4.79 update_halo_module.update_halo_
    0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
    0.03 tea_module.tea_allsum_
      1.40 MPI_Allreduce
    0.00 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
    0.02 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
      0.04 !$omp parallel @tea_leaf_cg.f90:186
        7.42 !$omp do @tea_leaf_cg.f90:187
          0.02 !$omp implicit barrier @tea_leaf_cg.f90:
        0.01 !$omp implicit barrier @tea_leaf_cg.f90:20
    0.02 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
      0.06 !$omp parallel @tea_leaf_cg.f90:234
        18.55 !$omp do @tea_leaf_cg.f90:247
          0.06 !$omp implicit barrier @tea_leaf_cg.f90:
        0.02 !$omp implicit barrier @tea_leaf_cg.f90:25
    0.01 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
      0.04 !$omp parallel @tea_leaf_cg.f90:284
        0.58 !$omp do @tea_leaf_cg.f90:294
          0.03 !$omp implicit barrier @tea_leaf_cg.f90:
        0.01 !$omp implicit barrier @tea_leaf_cg.f90:30
    0.00 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
  0.00 field_summary_
  0.00 tea_module.tea_finalize_
```

System tree | Statistics | Su

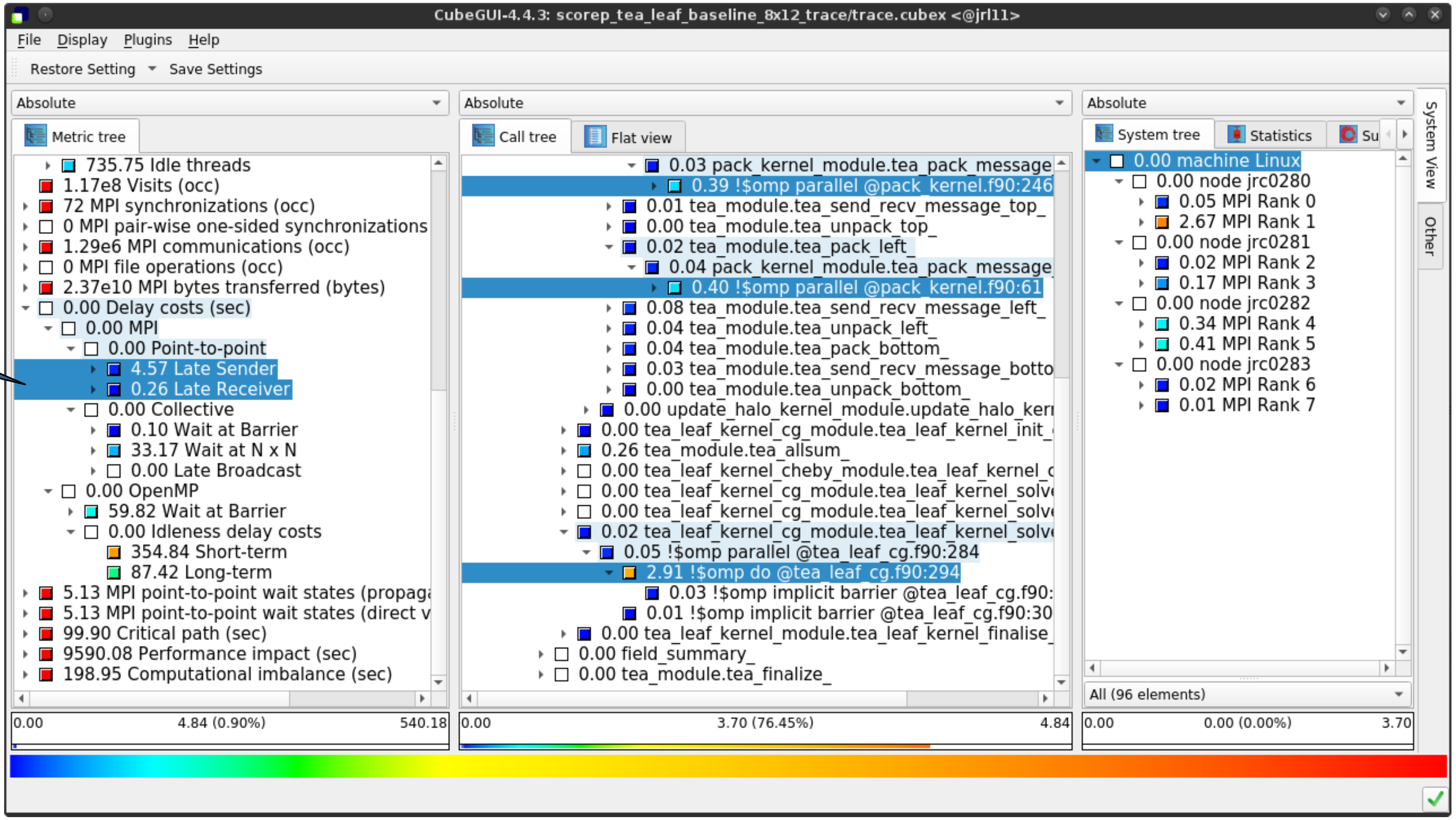
```
0.00 machine Linux
  0.00 node jrc0280
    0.17 MPI Rank 0
    18.55 MPI Rank 1
  0.00 node jrc0281
    0.07 MPI Rank 2
    0.16 MPI Rank 3
  0.00 node jrc0282
    0.10 MPI Rank 4
    6.67 MPI Rank 5
  0.00 node jrc0283
    0.11 MPI Rank 6
    0.14 MPI Rank 7
```

All (96 elements)

```
0.00   33.17 (6.14%)   540.18
0.00   25.97 (78.30%)   33.17
0.00   0.00 (0.00%)   25.97
```

TeaLeaf Scalasca report analysis (IV)

…while the MPI point-to-point wait states are caused by the 3rd solver `do` loop (on rank 1) and two loops in the halo exchange



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

```
735.75 Idle threads
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizations
1.29e6 MPI communications (occ)
0 MPI file operations (occ)
2.37e10 MPI bytes transferred (bytes)
0.00 Delay costs (sec)
  0.00 MPI
    0.00 Point-to-point
      4.57 Late Sender
      0.26 Late Receiver
    0.00 Collective
      0.10 Wait at Barrier
      33.17 Wait at N x N
      0.00 Late Broadcast
  0.00 OpenMP
    59.82 Wait at Barrier
    0.00 Idleness delay costs
      354.84 Short-term
      87.42 Long-term
5.13 MPI point-to-point wait states (propaga
5.13 MPI point-to-point wait states (direct v
99.90 Critical path (sec)
9590.08 Performance impact (sec)
198.95 Computational imbalance (sec)
```

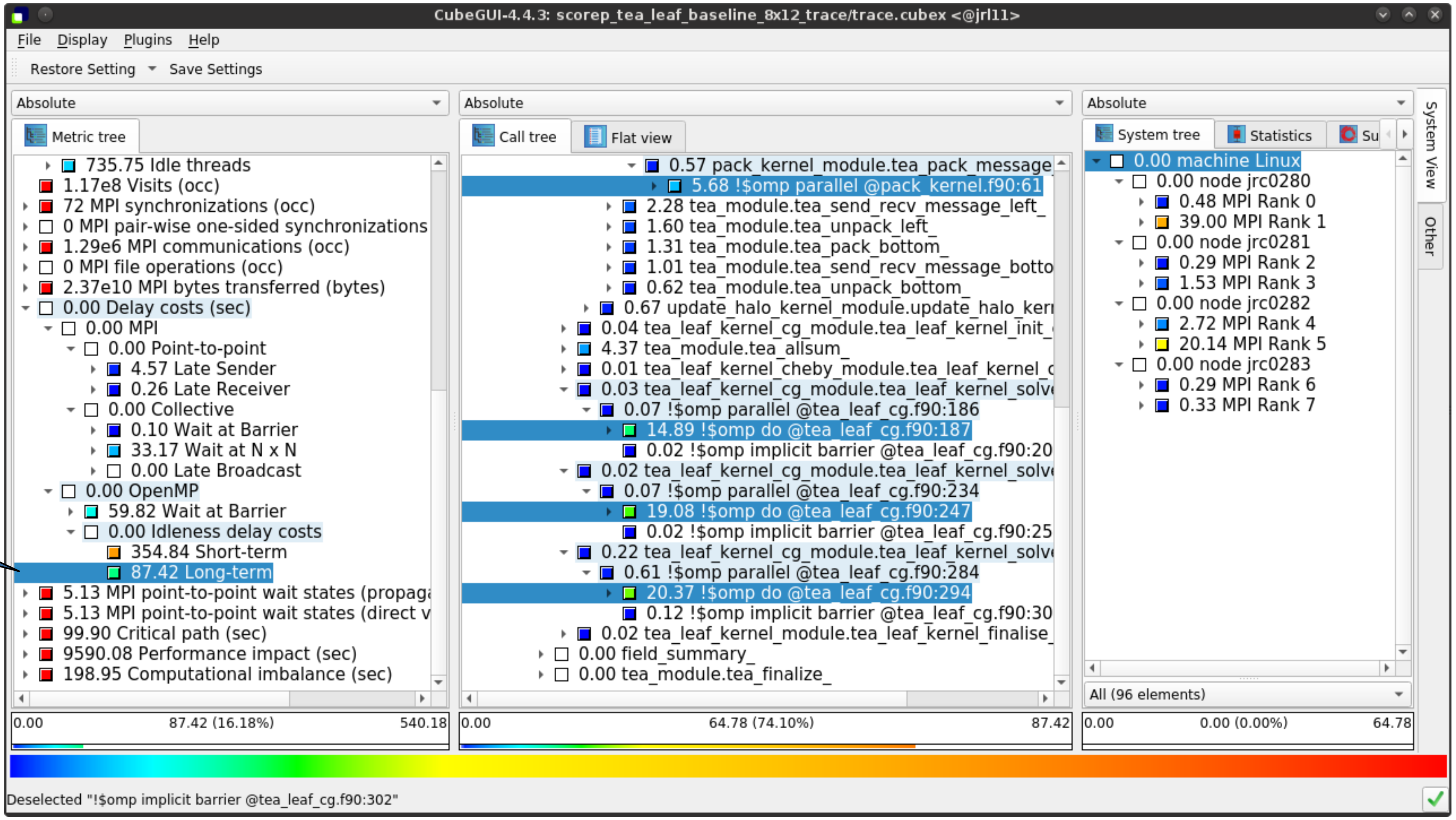
```
        0.03 pack_kernel_module.tea_pack_message
          0.39 !$omp parallel @pack_kernel.f90:246
      0.01 tea_module.tea_send_recv_message_top_
      0.00 tea_module.tea_unpack_top_
      0.02 tea_module.tea_pack_left_
        0.04 pack_kernel_module.tea_pack_message
          0.40 !$omp parallel @pack_kernel.f90:61
      0.08 tea_module.tea_send_recv_message_left_
      0.04 tea_module.tea_unpack_left_
      0.04 tea_module.tea_pack_bottom_
      0.03 tea_module.tea_send_recv_message_botto
      0.00 tea_module.tea_unpack_bottom_
    0.00 update_halo_kernel_module.update_halo_kerr
  0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
  0.26 tea_module.tea_allsum_
  0.00 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
  0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
  0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
  0.02 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.05 !$omp parallel @tea_leaf_cg.f90:284
      2.91 !$omp do @tea_leaf_cg.f90:294
        0.03 !$omp implicit barrier @tea_leaf_cg.f90:
      0.01 !$omp implicit barrier @tea_leaf_cg.f90:30
  0.00 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
0.00 field_summary_
0.00 tea_module.tea_finalize_
```

```
0.00 machine Linux
  0.00 node jrc0280
    0.05 MPI Rank 0
    2.67 MPI Rank 1
  0.00 node jrc0281
    0.02 MPI Rank 2
    0.17 MPI Rank 3
  0.00 node jrc0282
    0.34 MPI Rank 4
    0.41 MPI Rank 5
  0.00 node jrc0283
    0.02 MPI Rank 6
    0.01 MPI Rank 7
```

```
0.00    4.84 (0.90%)    540.18
0.00    3.70 (76.45%)   4.84
0.00    0.00 (0.00%)    3.70
```

TeaLeaf Scalasca report analysis (V)

Various OpenMP do loops (incl. the solver loops) also cause OpenMP thread idleness on other ranks via propagation



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

Metric tree:

```
  735.75 Idle threads
  1.17e8 Visits (occ)
  72 MPI synchronizations (occ)
  0 MPI pair-wise one-sided synchronizations
  1.29e6 MPI communications (occ)
  0 MPI file operations (occ)
  2.37e10 MPI bytes transferred (bytes)
  0.00 Delay costs (sec)
    0.00 MPI
      0.00 Point-to-point
        4.57 Late Sender
        0.26 Late Receiver
      0.00 Collective
        0.10 Wait at Barrier
        33.17 Wait at N x N
        0.00 Late Broadcast
    0.00 OpenMP
      59.82 Wait at Barrier
      0.00 Idleness delay costs
        354.84 Short-term
        87.42 Long-term
  5.13 MPI point-to-point wait states (propag
  5.13 MPI point-to-point wait states (direct v
  99.90 Critical path (sec)
  9590.08 Performance impact (sec)
  198.95 Computational imbalance (sec)
0.00    87.42 (16.18%)    540.18
```

Call tree:

```
        0.57 pack_kernel_module.tea_pack_message
          5.68 !$omp parallel @pack_kernel.f90:61
      2.28 tea_module.tea_send_recv_message_left_
      1.60 tea_module.tea_unpack_left_
      1.31 tea_module.tea_pack_bottom_
      1.01 tea_module.tea_send_recv_message_botto
      0.62 tea_module.tea_unpack_bottom_
    0.67 update_halo_kernel_module.update_halo_kerı
  0.04 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
  4.37 tea_module.tea_allsum_
  0.01 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
  0.03 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.07 !$omp parallel @tea_leaf_cg.f90:186
      14.89 !$omp do @tea_leaf_cg.f90:187
      0.02 !$omp implicit barrier @tea_leaf_cg.f90:20
  0.02 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.07 !$omp parallel @tea_leaf_cg.f90:234
      19.08 !$omp do @tea_leaf_cg.f90:247
      0.02 !$omp implicit barrier @tea_leaf_cg.f90:25
  0.22 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.61 !$omp parallel @tea_leaf_cg.f90:284
      20.37 !$omp do @tea_leaf_cg.f90:294
      0.12 !$omp implicit barrier @tea_leaf_cg.f90:30
  0.02 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
0.00 field_summary_
0.00 tea_module.tea_finalize_
0.00    64.78 (74.10%)    87.42
```

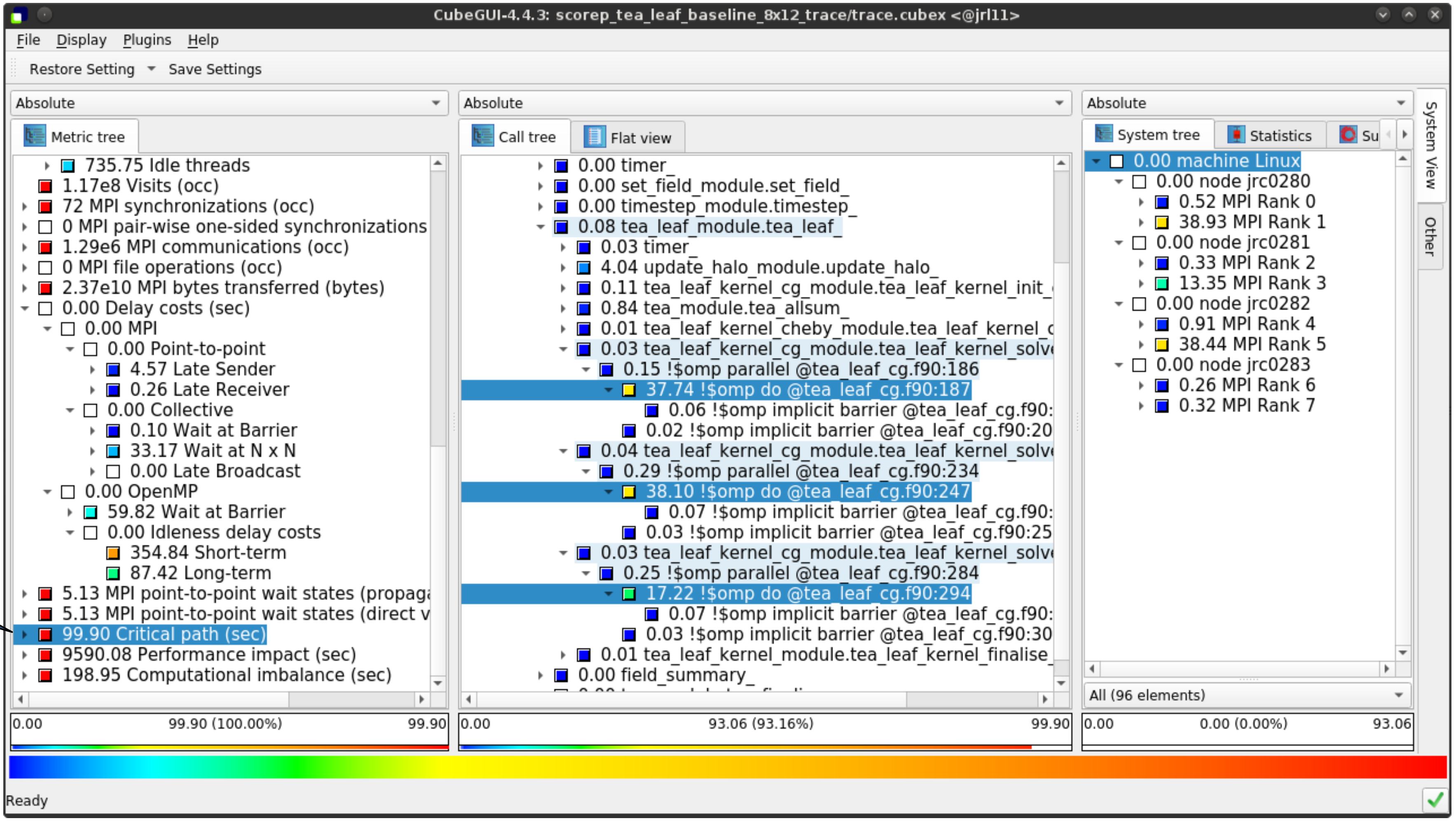
System tree:

```
0.00 machine Linux
  0.00 node jrc0280
    0.48 MPI Rank 0
    39.00 MPI Rank 1
  0.00 node jrc0281
    0.29 MPI Rank 2
    1.53 MPI Rank 3
  0.00 node jrc0282
    2.72 MPI Rank 4
    20.14 MPI Rank 5
  0.00 node jrc0283
    0.29 MPI Rank 6
    0.33 MPI Rank 7
All (96 elements)
0.00    0.00 (0.00%)    64.78
```

Deselected "!$omp implicit barrier @tea_leaf_cg.f90:302"

TeaLeaf Scalasca report analysis (VI)

The Critical Path also highlights the three solver loops…



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings

Absolute | Absolute | Absolute

Metric tree:

```
735.75 Idle threads
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizations
1.29e6 MPI communications (occ)
0 MPI file operations (occ)
2.37e10 MPI bytes transferred (bytes)
0.00 Delay costs (sec)
  0.00 MPI
    0.00 Point-to-point
      4.57 Late Sender
      0.26 Late Receiver
    0.00 Collective
      0.10 Wait at Barrier
      33.17 Wait at N x N
      0.00 Late Broadcast
  0.00 OpenMP
    59.82 Wait at Barrier
    0.00 Idleness delay costs
      354.84 Short-term
      87.42 Long-term
5.13 MPI point-to-point wait states (propaga
5.13 MPI point-to-point wait states (direct v
99.90 Critical path (sec)
9590.08 Performance impact (sec)
198.95 Computational imbalance (sec)
```

0.00 — 99.90 (100.00%) — 99.90

Call tree / Flat view:

```
0.00 timer_
0.00 set_field_module.set_field_
0.00 timestep_module.timestep_
0.08 tea_leaf_module.tea_leaf_
  0.03 timer_
  4.04 update_halo_module.update_halo_
  0.11 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
  0.84 tea_module.tea_allsum_
  0.01 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
  0.03 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.15 !$omp parallel @tea_leaf_cg.f90:186
      37.74 !$omp do @tea_leaf_cg.f90:187
        0.06 !$omp implicit barrier @tea_leaf_cg.f90:
      0.02 !$omp implicit barrier @tea_leaf_cg.f90:20
  0.04 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.29 !$omp parallel @tea_leaf_cg.f90:234
      38.10 !$omp do @tea_leaf_cg.f90:247
        0.07 !$omp implicit barrier @tea_leaf_cg.f90:
      0.03 !$omp implicit barrier @tea_leaf_cg.f90:25
  0.03 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    0.25 !$omp parallel @tea_leaf_cg.f90:284
      17.22 !$omp do @tea_leaf_cg.f90:294
        0.07 !$omp implicit barrier @tea_leaf_cg.f90:
      0.03 !$omp implicit barrier @tea_leaf_cg.f90:30
  0.01 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
0.00 field_summary_
```

0.00 — 93.06 (93.16%) — 99.90

System tree / Statistics / Su:

```
0.00 machine Linux
  0.00 node jrc0280
    0.52 MPI Rank 0
    38.93 MPI Rank 1
  0.00 node jrc0281
    0.33 MPI Rank 2
    13.35 MPI Rank 3
  0.00 node jrc0282
    0.91 MPI Rank 4
    38.44 MPI Rank 5
  0.00 node jrc0283
    0.26 MPI Rank 6
    0.32 MPI Rank 7
```

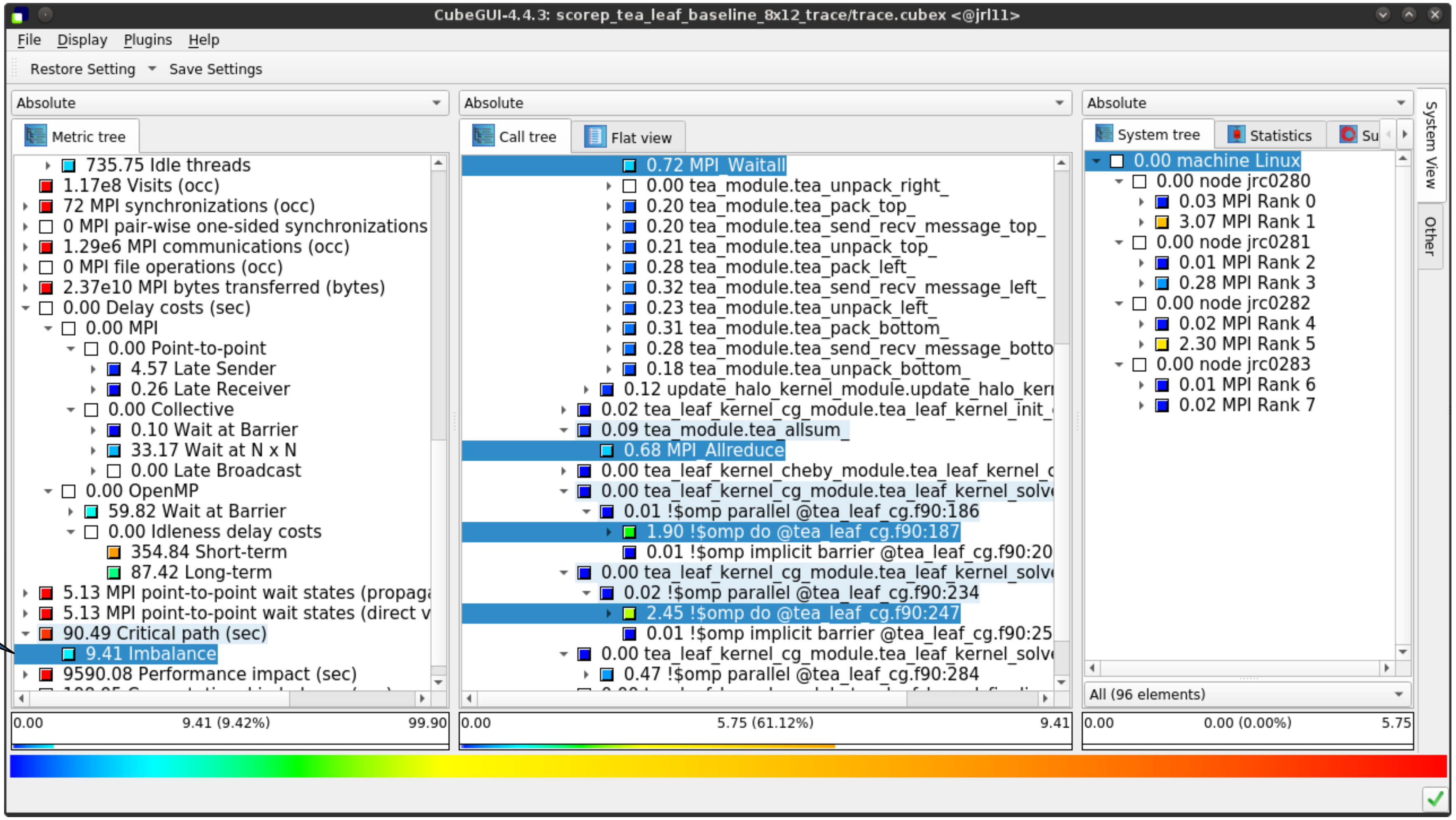
All (96 elements)

0.00 — 0.00 (0.00%) — 93.06

Ready

TeaLeaf Scalasca report analysis (VII)

…with imbalance (time on critical path above average) mostly in the first two loops and MPI communication

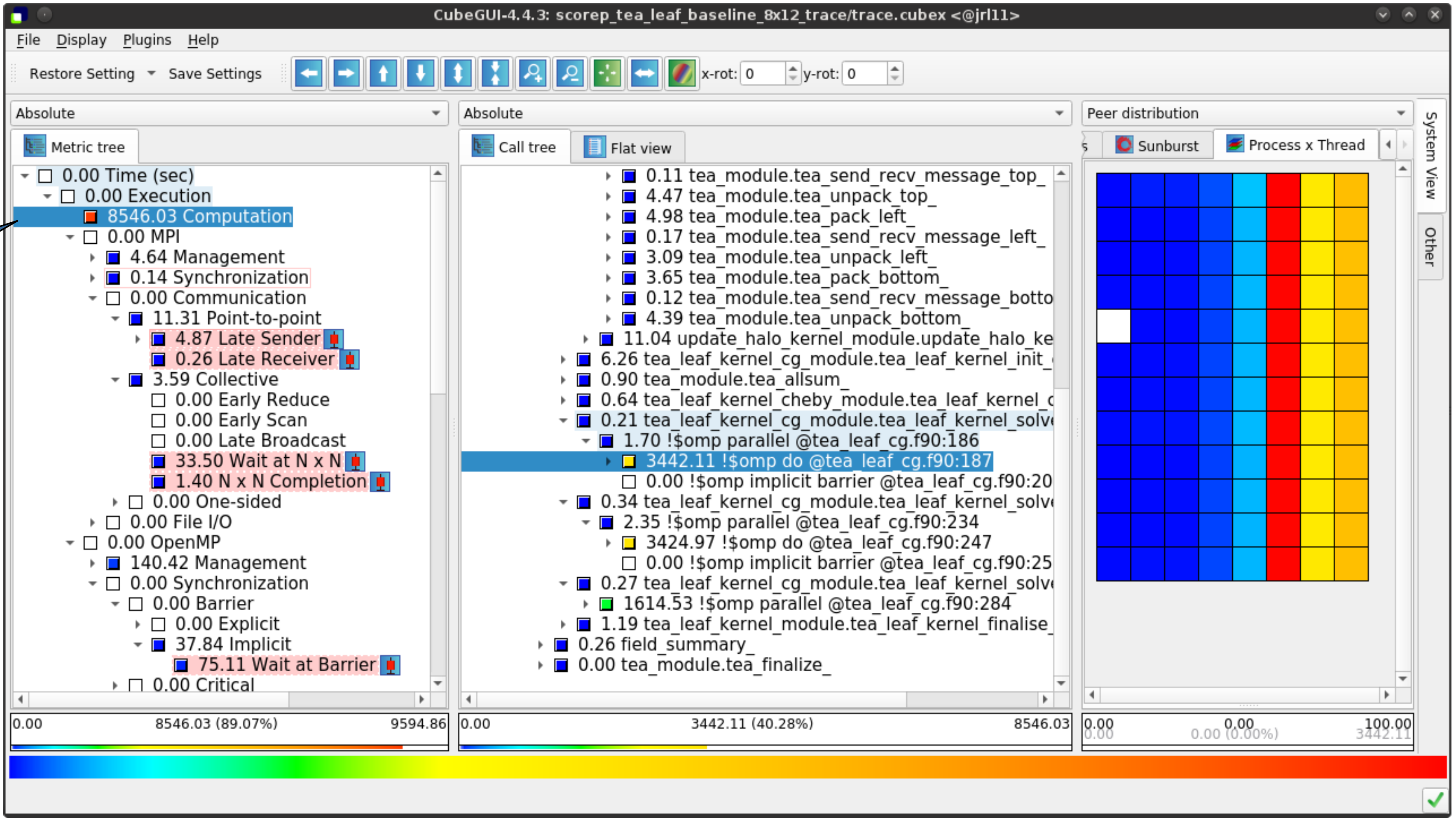


CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

```
Metric tree:
735.75 Idle threads
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizations
1.29e6 MPI communications (occ)
0 MPI file operations (occ)
2.37e10 MPI bytes transferred (bytes)
0.00 Delay costs (sec)
  0.00 MPI
    0.00 Point-to-point
      4.57 Late Sender
      0.26 Late Receiver
    0.00 Collective
      0.10 Wait at Barrier
      33.17 Wait at N x N
      0.00 Late Broadcast
  0.00 OpenMP
    59.82 Wait at Barrier
    0.00 Idleness delay costs
      354.84 Short-term
      87.42 Long-term
5.13 MPI point-to-point wait states (propag
5.13 MPI point-to-point wait states (direct v
90.49 Critical path (sec)
  9.41 Imbalance
9590.08 Performance impact (sec)
0.00          9.41 (9.42%)          99.90

Call tree:
      0.72 MPI_Waitall
      0.00 tea_module.tea_unpack_right_
      0.20 tea_module.tea_pack_top_
      0.20 tea_module.tea_send_recv_message_top_
      0.21 tea_module.tea_unpack_top_
      0.28 tea_module.tea_pack_left_
      0.32 tea_module.tea_send_recv_message_left_
      0.23 tea_module.tea_unpack_left_
      0.31 tea_module.tea_pack_bottom_
      0.28 tea_module.tea_send_recv_message_botto
      0.18 tea_module.tea_unpack_bottom_
    0.12 update_halo_kernel_module.update_halo_ker
  0.02 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
  0.09 tea_module.tea_allsum_
    0.68 MPI_Allreduce
  0.00 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
  0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_solv
    0.01 !$omp parallel @tea_leaf_cg.f90:186
      1.90 !$omp do @tea_leaf_cg.f90:187
      0.01 !$omp implicit barrier @tea_leaf_cg.f90:20
  0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_solv
    0.02 !$omp parallel @tea_leaf_cg.f90:234
      2.45 !$omp do @tea_leaf_cg.f90:247
      0.01 !$omp implicit barrier @tea_leaf_cg.f90:25
  0.00 tea_leaf_kernel_cg_module.tea_leaf_kernel_solv
    0.47 !$omp parallel @tea_leaf_cg.f90:284
0.00          5.75 (61.12%)          9.41

System tree:
0.00 machine Linux
  0.00 node jrc0280
    0.03 MPI Rank 0
    3.07 MPI Rank 1
  0.00 node jrc0281
    0.01 MPI Rank 2
    0.28 MPI Rank 3
  0.00 node jrc0282
    0.02 MPI Rank 4
    2.30 MPI Rank 5
  0.00 node jrc0283
    0.01 MPI Rank 6
    0.02 MPI Rank 7
0.00          0.00 (0.00%)          5.75
```

TeaLeaf Scalasca report analysis (VIII)

Computation time of 1st…


CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

Metric tree (Absolute):

```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Early Reduce
          0.00 Early Scan
          0.00 Late Broadcast
          33.50 Wait at N x N
          1.40 N x N Completion
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at Barrier
        0.00 Critical
```

0.00 — 8546.03 (89.07%) — 9594.86

Call tree (Absolute):

```
0.11 tea_module.tea_send_recv_message_top_
4.47 tea_module.tea_unpack_top_
4.98 tea_module.tea_pack_left_
0.17 tea_module.tea_send_recv_message_left_
3.09 tea_module.tea_unpack_left_
3.65 tea_module.tea_pack_bottom_
0.12 tea_module.tea_send_recv_message_botto
4.39 tea_module.tea_unpack_bottom_
11.04 update_halo_kernel_module.update_halo_ke
6.26 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
0.90 tea_module.tea_allsum_
0.64 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
0.21 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
  1.70 !$omp parallel @tea_leaf_cg.f90:186
    3442.11 !$omp do @tea_leaf_cg.f90:187
    0.00 !$omp implicit barrier @tea_leaf_cg.f90:20
0.34 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
  2.35 !$omp parallel @tea_leaf_cg.f90:234
    3424.97 !$omp do @tea_leaf_cg.f90:247
    0.00 !$omp implicit barrier @tea_leaf_cg.f90:25
0.27 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
  1614.53 !$omp parallel @tea_leaf_cg.f90:284
1.19 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
0.26 field_summary_
0.00 tea_module.tea_finalize_
```

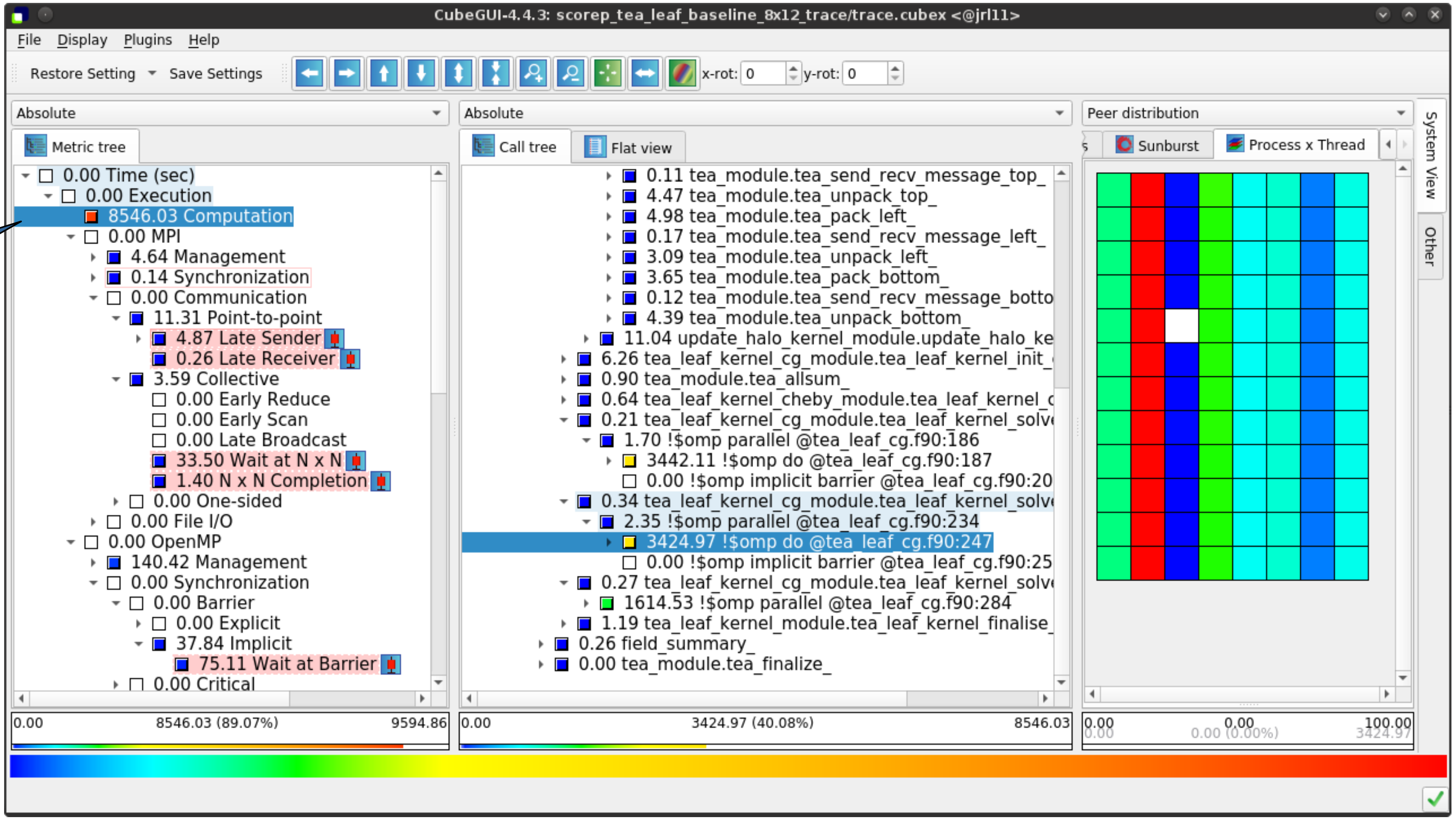
0.00 — 3442.11 (40.28%) — 8546.03

Peer distribution — Sunburst — Process x Thread

0.00 — 0.00 (0.00%) — 100.00 / 3442.11

TeaLeaf Scalasca report analysis (IX)

…and 2nd `do` loop mostly balanced within each rank, but vary considerably across ranks…



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

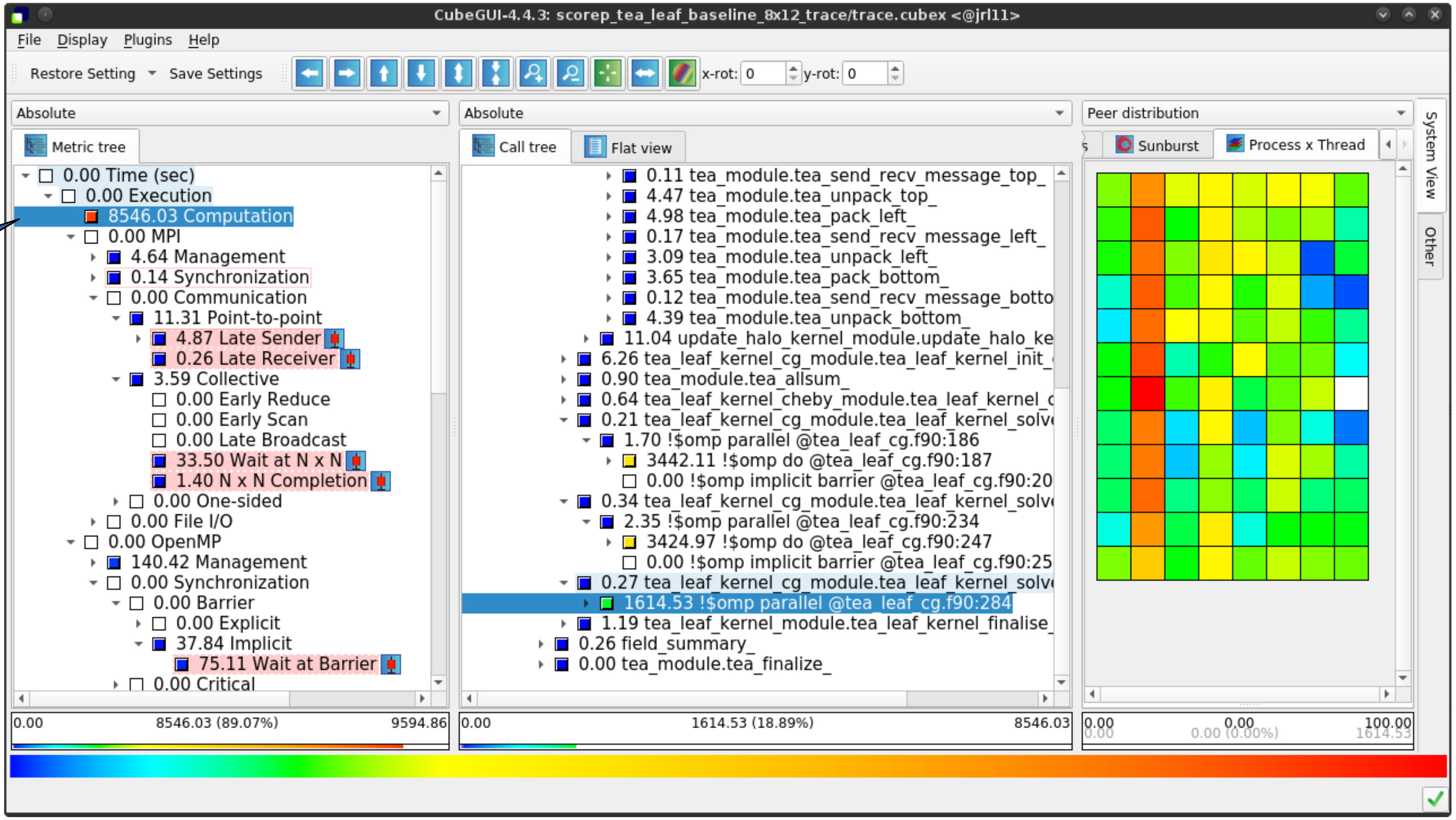
```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Early Reduce
          0.00 Early Scan
          0.00 Late Broadcast
          33.50 Wait at N x N
          1.40 N x N Completion
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at Barrier
        0.00 Critical
```

```
      0.11 tea_module.tea_send_recv_message_top_
      4.47 tea_module.tea_unpack_top_
      4.98 tea_module.tea_pack_left_
      0.17 tea_module.tea_send_recv_message_left_
      3.09 tea_module.tea_unpack_left_
      3.65 tea_module.tea_pack_bottom_
      0.12 tea_module.tea_send_recv_message_botto
      4.39 tea_module.tea_unpack_bottom_
    11.04 update_halo_kernel_module.update_halo_ke
  6.26 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
  0.90 tea_module.tea_allsum_
  0.64 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
  0.21 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    1.70 !$omp parallel @tea_leaf_cg.f90:186
      3442.11 !$omp do @tea_leaf_cg.f90:187
      0.00 !$omp implicit barrier @tea_leaf_cg.f90:20
  0.34 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    2.35 !$omp parallel @tea_leaf_cg.f90:234
      3424.97 !$omp do @tea_leaf_cg.f90:247
      0.00 !$omp implicit barrier @tea_leaf_cg.f90:25
  0.27 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    1614.53 !$omp parallel @tea_leaf_cg.f90:284
  1.19 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
0.26 field_summary_
0.00 tea_module.tea_finalize_
```

```
0.00    8546.03 (89.07%)    9594.86
0.00    3424.97 (40.08%)    8546.03
0.00    0.00 (0.00%)        100.00
0.00                        3424.97
```

TeaLeaf Scalasca report analysis (X)

…while the 3rd do loop also shows imbalance within each rank



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings x-rot: 0 y-rot: 0

Absolute | Absolute | Peer distribution

Metric tree | Call tree | Flat view | Sunburst | Process x Thread | System View | Other

```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Early Reduce
          0.00 Early Scan
          0.00 Late Broadcast
          33.50 Wait at N x N
          1.40 N x N Completion
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at Barrier
        0.00 Critical
```

```
      0.11 tea_module.tea_send_recv_message_top_
      4.47 tea_module.tea_unpack_top_
      4.98 tea_module.tea_pack_left_
      0.17 tea_module.tea_send_recv_message_left_
      3.09 tea_module.tea_unpack_left_
      3.65 tea_module.tea_pack_bottom_
      0.12 tea_module.tea_send_recv_message_botto
      4.39 tea_module.tea_unpack_bottom_
    11.04 update_halo_kernel_module.update_halo_ke
  6.26 tea_leaf_kernel_cg_module.tea_leaf_kernel_init_
  0.90 tea_module.tea_allsum_
  0.64 tea_leaf_kernel_cheby_module.tea_leaf_kernel_c
  0.21 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    1.70 !$omp parallel @tea_leaf_cg.f90:186
      3442.11 !$omp do @tea_leaf_cg.f90:187
      0.00 !$omp implicit barrier @tea_leaf_cg.f90:20
  0.34 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    2.35 !$omp parallel @tea_leaf_cg.f90:234
      3424.97 !$omp do @tea_leaf_cg.f90:247
      0.00 !$omp implicit barrier @tea_leaf_cg.f90:25
  0.27 tea_leaf_kernel_cg_module.tea_leaf_kernel_solve
    1614.53 !$omp parallel @tea_leaf_cg.f90:284
  1.19 tea_leaf_kernel_module.tea_leaf_kernel_finalise_
0.26 field_summary_
0.00 tea_module.tea_finalize_
```

```
0.00   8546.03 (89.07%)   9594.86
0.00   1614.53 (18.89%)   8546.03
0.00   0.00   0.00 (0.00%)   100.00   1614.53
```

TeaLeaf analysis summary

- The first two OpenMP do loops of the solver are well balanced within a rank, but are imbalanced across ranks
  - ➔ Requires a global load balancing strategy
- The third OpenMP do loop, however, is imbalanced within ranks,
  - causing direct “Wait at OpenMP Barrier” wait states,
  - which cause indirect MPI point-to-point wait states,
  - which in turn cause OpenMP thread idleness
  - ➔ Low-hanging fruit

- Adding a `SCHEDULE(guided)` clause reduced
  - the MPI point-to-point wait states by ~66%
  - the MPI collective wait states by ~50%
  - the OpenMP “Wait at Barrier” wait states by ~55%
  - the OpenMP thread idleness by ~11%
  - ➔ **Overall runtime (wall-clock) reduction by ~5%**

Scalasca Trace Tools: Further information

- Collection of trace-based performance tools
  - Specifically designed for large-scale systems
  - Features an automatic trace analyzer providing wait-state, critical-path, and delay analysis
  - Supports MPI, OpenMP, OpenACC, POSIX threads, and hybrid MPI+OpenMP/OpenACC/Pthreads
    - host-side events only (no device/offloaded events) from OpenMP, OpenACC, OpenCL, HIP, CUDA
- Available under 3-clause BSD open-source license

- Documentation & sources:
  - https://www.scalasca.org
- Contact:
  - mailto: scalasca@fz-juelich.de

trace tools
scalasca

Reference material

trace tools

scalasca

Scalasca command – One command for (almost) everything

```
% scalasca
Scalasca 2.6.1
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 will be remove in the next major release
  ⇨ use Score-P instrumenter directly

Scalasca convenience command: scan / scalasca -analyze

```
% scan
Scalasca 2.6.1: 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.
  -R #runs        : Specify the number of measurement runs per config.
  -M cfgfile      : Specify a config file for a multi-run measurement.
  -P preset       : Specify a preset for a multi-run measurement, e.g., 'pop'.
  -L              : List available multi-run presets.
  -D cfgfile      : Check a multi-run config file for validity and dump
                  : the processed configuration for comparison.
```

- Scalasca measurement collection & analysis nexus

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)

Scalasca convenience command: square / scalasca -examine

```
% square
Scalasca 2.6.1: analysis report explorer
usage: square [OPTIONS] <experiment archive | cube file>
  -C <none | quick | full> : Level of sanity checks for newly created reports
  -c <number>              : Consider number of counters when doing scoring (-s)
  -F                       : Force remapping of already existing reports
  -f filtfile              : Use specified filter file when doing scoring (-s)
  -s                       : Skip display and output textual score report
  -v                       : Enable verbose mode
  -n                       : Do not include idle thread metric
  -S <mean | merge>        : Aggregation method for summarization results of
                             each configuration (default: merge)
  -T <mean | merge>        : Aggregation method for trace analysis results of
                             each configuration (default: merge)
  -A                       : Post-process every step of a multi-run experiment
  -I                       : Ignore structural sanity checks and force aggregation
                             of measurements in a multi-run experiment
  -x <scorep-score opt>    : Pass option(s) to scorep-score
```

- Scalasca analysis report explorer (Cube)

Scalasca advanced command:
scout - Scalasca automatic trace analyzer

```
% scout.hyb --help
SCOUT   (Scalasca 2.6.1)
Copyright (c) 1998-2022 Forschungszentrum Juelich GmbH
Copyright (c) 2014-2021 RWTH Aachen University
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

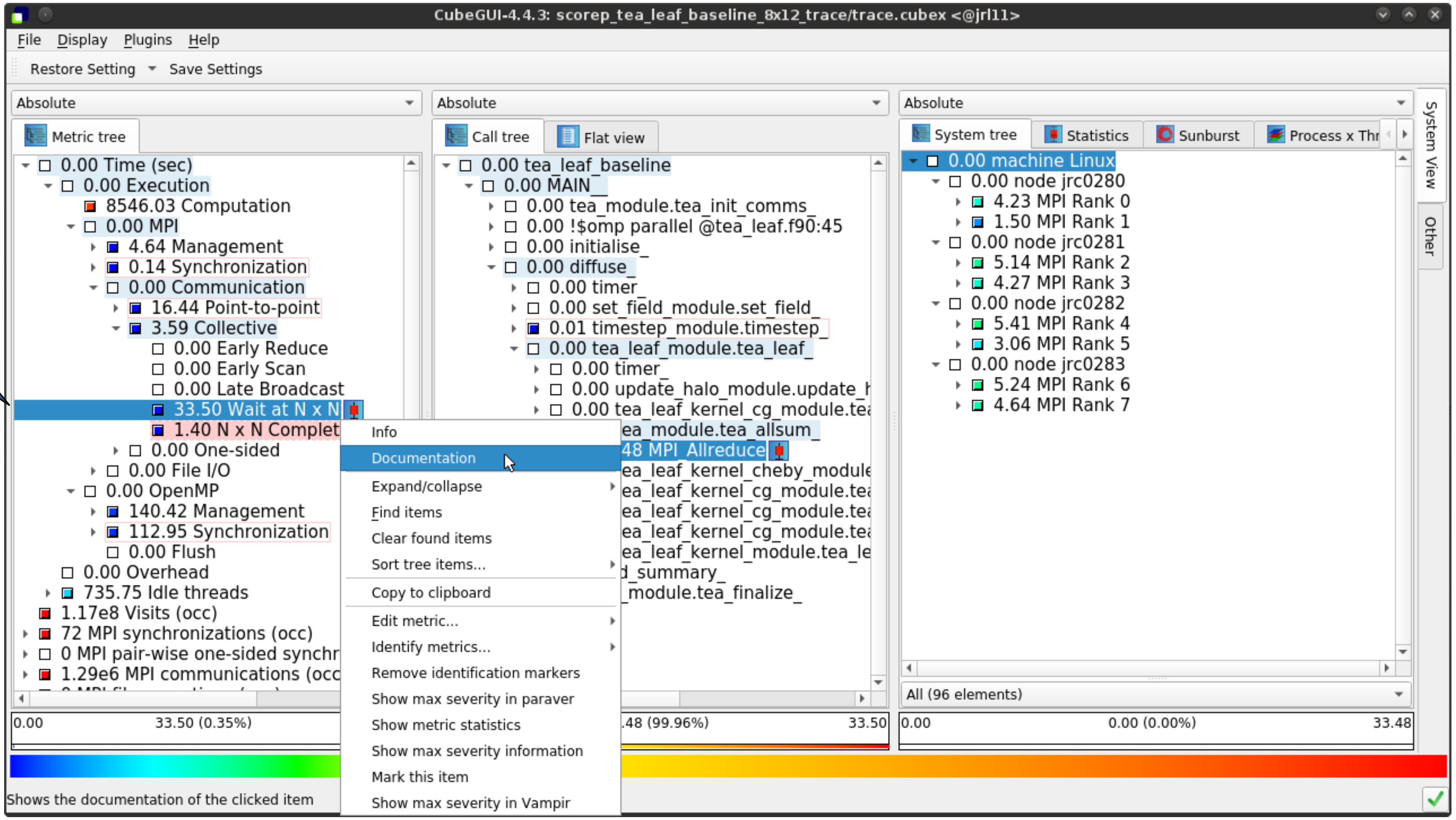
- Scalasca trace event timestamp consistency correction

```
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

Online metric description

Access online metric description via context menu (right-click)



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

0.00 Time (sec)
0.00 Execution
8546.03 Computation
0.00 MPI
4.64 Management
0.14 Synchronization
0.00 Communication
16.44 Point-to-point
3.59 Collective
0.00 Early Reduce
0.00 Early Scan
0.00 Late Broadcast
33.50 Wait at N x N
1.40 N x N Complet
0.00 One-sided
0.00 File I/O
0.00 OpenMP
140.42 Management
112.95 Synchronization
0.00 Flush
0.00 Overhead
735.75 Idle threads
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchr
1.29e6 MPI communications (occ

0.00 tea_leaf_baseline
0.00 MAIN_
0.00 tea_module.tea_init_comms_
0.00 !$omp parallel @tea_leaf.f90:45
0.00 initialise_
0.00 diffuse_
0.00 timer_
0.00 set_field_module.set_field_
0.01 timestep_module.timestep_
0.00 tea_leaf_module.tea_leaf_
0.00 timer_
0.00 update_halo_module.update_h
0.00 tea_leaf_kernel_cg_module.tea
ea_module.tea_allsum_
48 MPI_Allreduce
ea_leaf_kernel_cheby_module
ea_leaf_kernel_cg_module.tea
ea_leaf_kernel_cg_module.tea
ea_leaf_kernel_cg_module.tea
ea_leaf_kernel_module.tea_le
d_summary_
_module.tea_finalize_

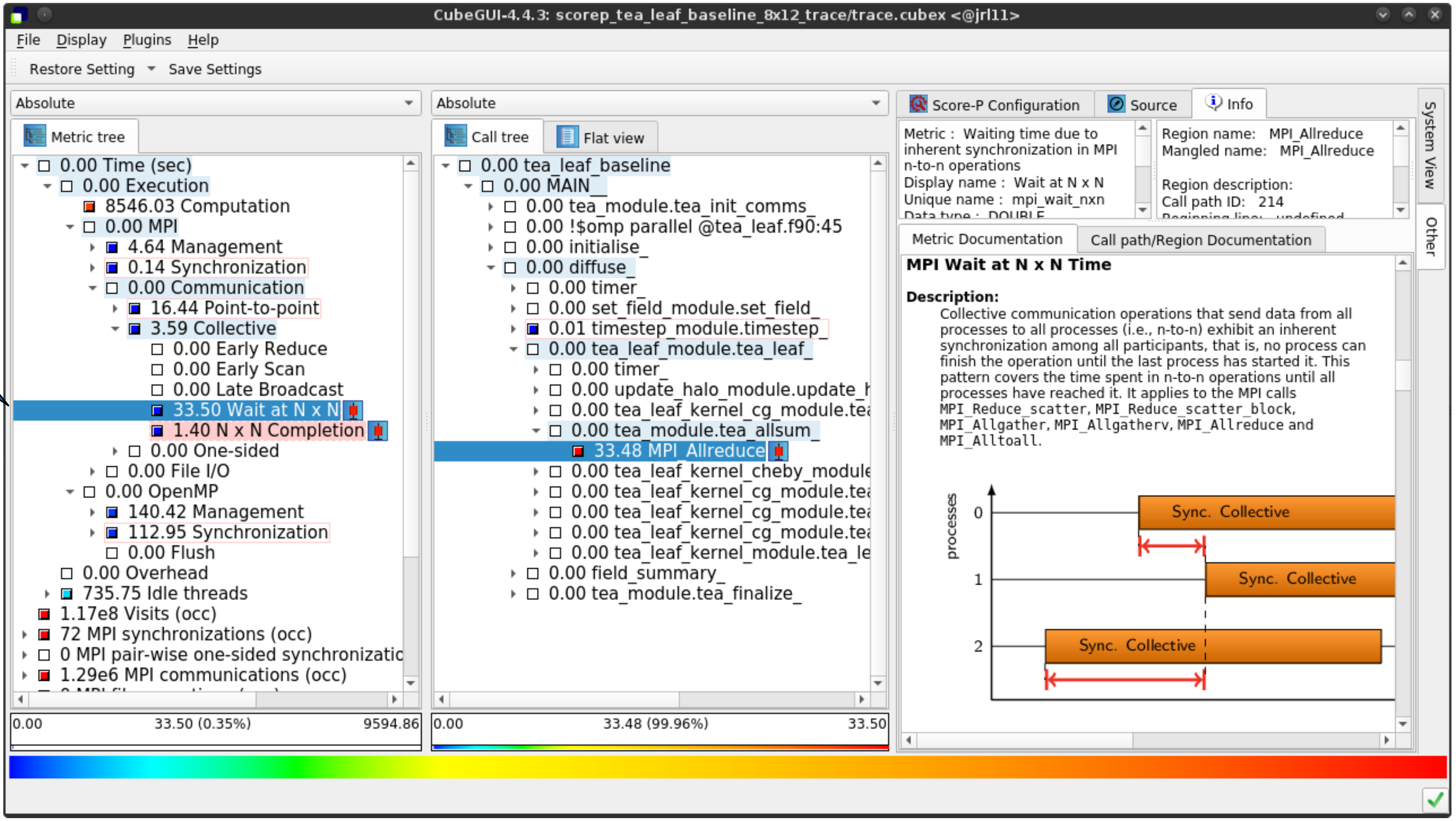
0.00 machine Linux
0.00 node jrc0280
4.23 MPI Rank 0
1.50 MPI Rank 1
0.00 node jrc0281
5.14 MPI Rank 2
4.27 MPI Rank 3
0.00 node jrc0282
5.41 MPI Rank 4
3.06 MPI Rank 5
0.00 node jrc0283
5.24 MPI Rank 6
4.64 MPI Rank 7

0.00 33.50 (0.35%)
.48 (99.96%) 33.50
0.00 0.00 (0.00%) 33.48

Shows the documentation of the clicked item

Online metric description (cont.)

Selection of different metric automatically updates description



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings

```
Absolute                                   Absolute
Metric tree                                Call tree   Flat view

0.00 Time (sec)                            0.00 tea_leaf_baseline
  0.00 Execution                             0.00 MAIN__
    8546.03 Computation                        0.00 tea_module.tea_init_comms_
    0.00 MPI                                   0.00 !$omp parallel @tea_leaf.f90:45
      4.64 Management                          0.00 initialise_
      0.14 Synchronization                     0.00 diffuse_
      0.00 Communication                         0.00 timer_
        16.44 Point-to-point                     0.00 set_field_module.set_field_
        3.59 Collective                          0.01 timestep_module.timestep_
          0.00 Early Reduce                      0.00 tea_leaf_module.tea_leaf_
          0.00 Early Scan                          0.00 timer_
          0.00 Late Broadcast                      0.00 update_halo_module.update_h
          33.50 Wait at N x N                      0.00 tea_leaf_kernel_cg_module.tea
          1.40 N x N Completion                    0.00 tea_module.tea_allsum_
        0.00 One-sided                               33.48 MPI_Allreduce
      0.00 File I/O                                0.00 tea_leaf_kernel_cheby_module
    0.00 OpenMP                                    0.00 tea_leaf_kernel_cg_module.tea
      140.42 Management                            0.00 tea_leaf_kernel_cg_module.tea
      112.95 Synchronization                       0.00 tea_leaf_kernel_cg_module.tea
      0.00 Flush                                   0.00 tea_leaf_kernel_module.tea_le
  0.00 Overhead                                0.00 field_summary_
  735.75 Idle threads                          0.00 tea_module.tea_finalize_
1.17e8 Visits (occ)
72 MPI synchronizations (occ)
0 MPI pair-wise one-sided synchronizatic
1.29e6 MPI communications (occ)

0.00     33.50 (0.35%)     9594.86        0.00     33.48 (99.96%)     33.50
```

Score-P Configuration | Source | Info

Metric : Waiting time due to inherent synchronization in MPI n-to-n operations
Display name : Wait at N x N
Unique name : mpi_wait_nxn
Data type : DOUBLE

Region name: MPI_Allreduce
Mangled name: MPI_Allreduce

Region description:
Call path ID: 214
Beginning line: undefined

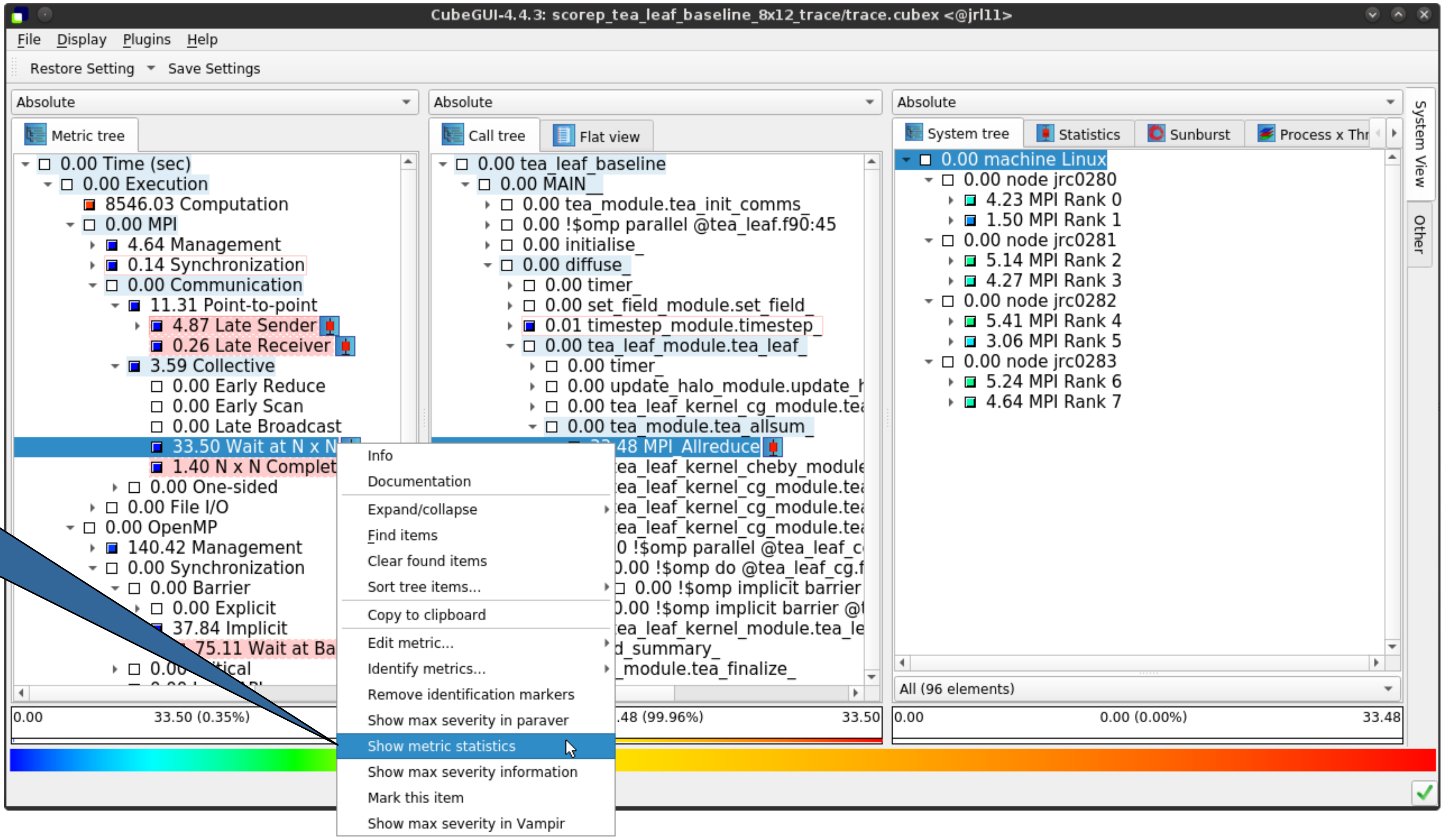
Metric Documentation | Call path/Region Documentation

MPI Wait at N x N Time

Description:
Collective communication operations that send data from all processes to all processes (i.e., n-to-n) exhibit an inherent synchronization among all participants, that is, no process can finish the operation until the last process has started it. This pattern covers the time spent in n-to-n operations until all processes have reached it. It applies to the MPI calls `MPI_Reduce_scatter`, `MPI_Reduce_scatter_block`, `MPI_Allgather`, `MPI_Allgatherv`, `MPI_Allreduce` and `MPI_Alltoall`.

Metric statistics

Access metric statistics for metrics marked with box plot icon from context menu



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings

Absolute | Absolute | Absolute

Metric tree | Call tree | Flat view | System tree | Statistics | Sunburst | Process x Thr

```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Early Reduce
          0.00 Early Scan
          0.00 Late Broadcast
          33.50 Wait at N x N
          1.40 N x N Complet
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at Ba
        0.00 Critical
```

```
0.00 tea_leaf_baseline
  0.00 MAIN__
    0.00 tea_module.tea_init_comms_
    0.00 !$omp parallel @tea_leaf.f90:45
    0.00 initialise_
    0.00 diffuse_
      0.00 timer_
      0.00 set_field_module.set_field_
      0.01 timestep_module.timestep_
      0.00 tea_leaf_module.tea_leaf_
        0.00 timer_
        0.00 update_halo_module.update_h
        0.00 tea_leaf_kernel_cg_module.tea
        0.00 tea_module.tea_allsum_
          33.48 MPI_Allreduce
        ea_leaf_kernel_cheby_module
        ea_leaf_kernel_cg_module.tea
        ea_leaf_kernel_cg_module.tea
        ea_leaf_kernel_cg_module.tea
        0 !$omp parallel @tea_leaf_c
        0.00 !$omp do @tea_leaf_cg.f
        0.00 !$omp implicit barrier
        0.00 !$omp implicit barrier @t
        ea_leaf_kernel_module.tea_le
      d_summary_
    _module.tea_finalize_
```

```
0.00 machine Linux
  0.00 node jrc0280
    4.23 MPI Rank 0
    1.50 MPI Rank 1
  0.00 node jrc0281
    5.14 MPI Rank 2
    4.27 MPI Rank 3
  0.00 node jrc0282
    5.41 MPI Rank 4
    3.06 MPI Rank 5
  0.00 node jrc0283
    5.24 MPI Rank 6
    4.64 MPI Rank 7
```

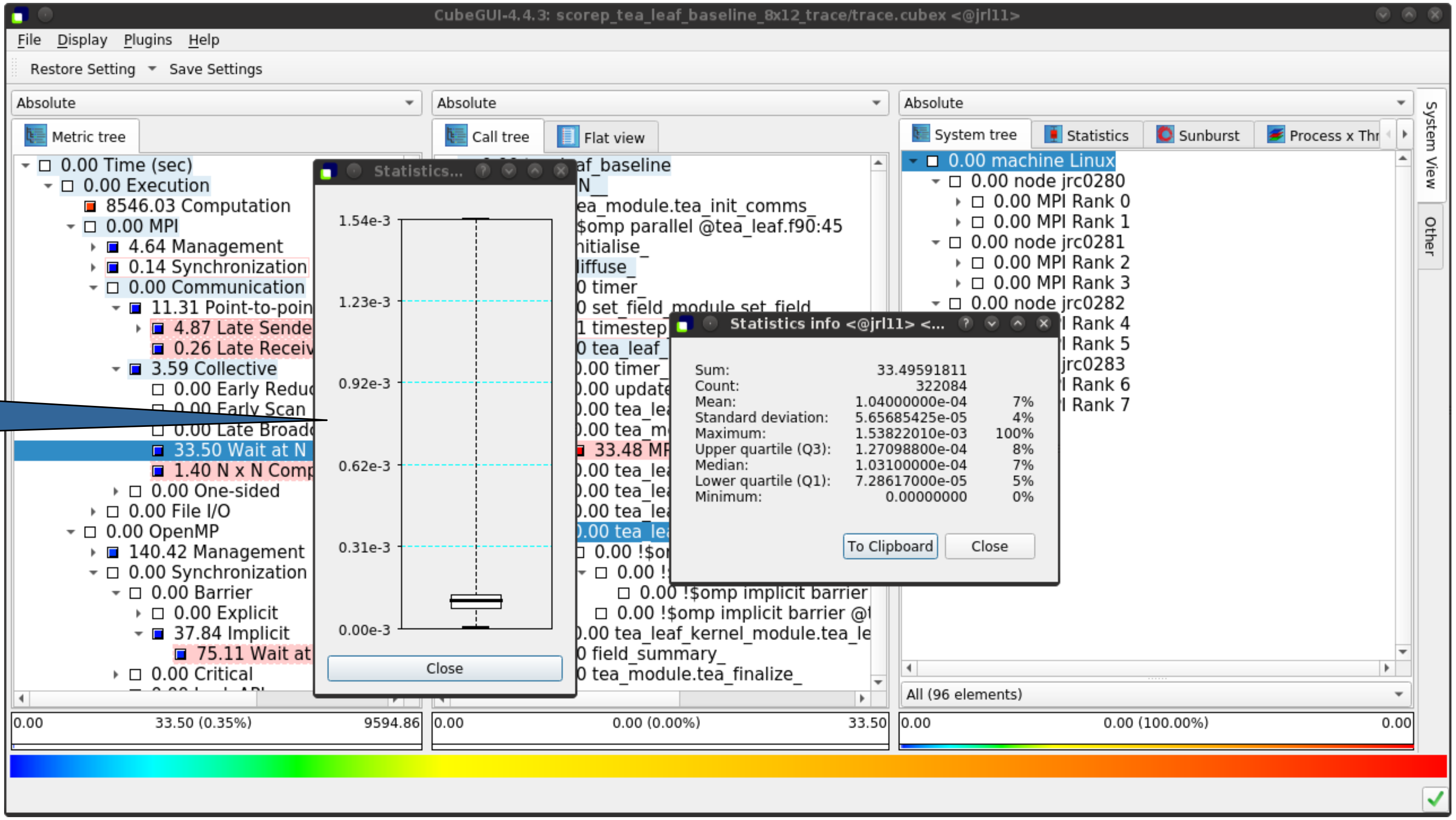
All (96 elements)

0.00 | 33.50 (0.35%) | .48 (99.96%) | 33.50 | 0.00 | 0.00 (0.00%) | 33.48

System View | Other

Metric statistics (cont.)

Shows instance statistics box plot, click to get details



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings

Absolute | Absolute | Absolute

Metric tree | Call tree | Flat view | System tree | Statistics | Sunburst | Process x Thr

```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-poin
          4.87 Late Sende
          0.26 Late Receiv
        3.59 Collective
          0.00 Early Reduc
          0.00 Early Scan
          0.00 Late Broadc
          33.50 Wait at N
          1.40 N x N Comp
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at
        0.00 Critical
```

```
af_baseline
N_
ea_module.tea_init_comms_
$omp parallel @tea_leaf.f90:45
nitialise_
liffuse_
0 timer_
0 set_field_module_set_field_
1 timestep_
0 tea_leaf_
0.00 timer_
0.00 update_
0.00 tea_lea
0.00 tea_mo
33.48 MP
0.00 tea_lea
0.00 tea_lea
0.00 tea_lea
0.00 tea_lea
0.00 !$om
0.00 !
  0.00 !$omp implicit barrier
0.00 !$omp implicit barrier @t
0.00 tea_leaf_kernel_module.tea_le
0 field_summary_
0 tea_module.tea_finalize_
```

```
0.00 machine Linux
  0.00 node jrc0280
    0.00 MPI Rank 0
    0.00 MPI Rank 1
  0.00 node jrc0281
    0.00 MPI Rank 2
    0.00 MPI Rank 3
  0.00 node jrc0282
    I Rank 4
    I Rank 5
  jrc0283
    I Rank 6
    I Rank 7
```

All (96 elements)

0.00 | 33.50 (0.35%) | 9594.86 — 0.00 | 0.00 (0.00%) | 33.50 — 0.00 | 0.00 (100.00%) | 0.00

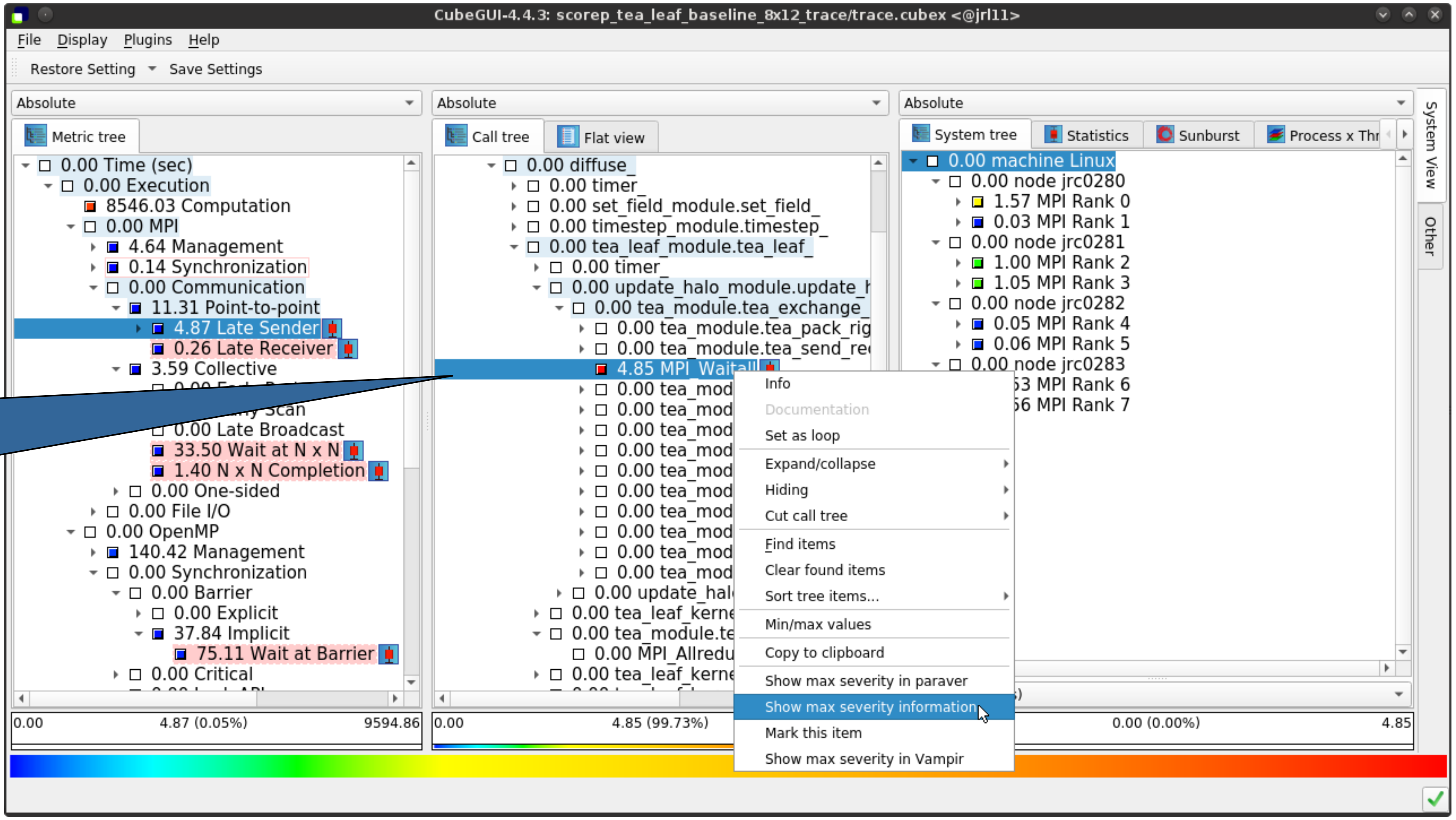
Statistics...

Statistics info <@jrl11> <...

```
Sum:                      33.49591811
Count:                         322084
Mean:                  1.04000000e-04     7%
Standard deviation:    5.65685425e-05     4%
Maximum:               1.53822010e-03   100%
Upper quartile (Q3):   1.27098800e-04     8%
Median:                1.03100000e-04     7%
Lower quartile (Q1):   7.28617000e-05     5%
Minimum:                   0.00000000     0%
```

To Clipboard | Close

Metric instance statistics



CubeGUI-4.4.3: scorep_tea_leaf_baseline_8x12_trace/trace.cubex <@jrl11>

File Display Plugins Help

Restore Setting Save Settings

Absolute | Absolute | Absolute

Metric tree | Call tree | Flat view | System tree | Statistics | Sunburst | Process x Thr

```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Late Broadcast
          33.50 Wait at N x N
          1.40 N x N Completion
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at Barrier
        0.00 Critical
```

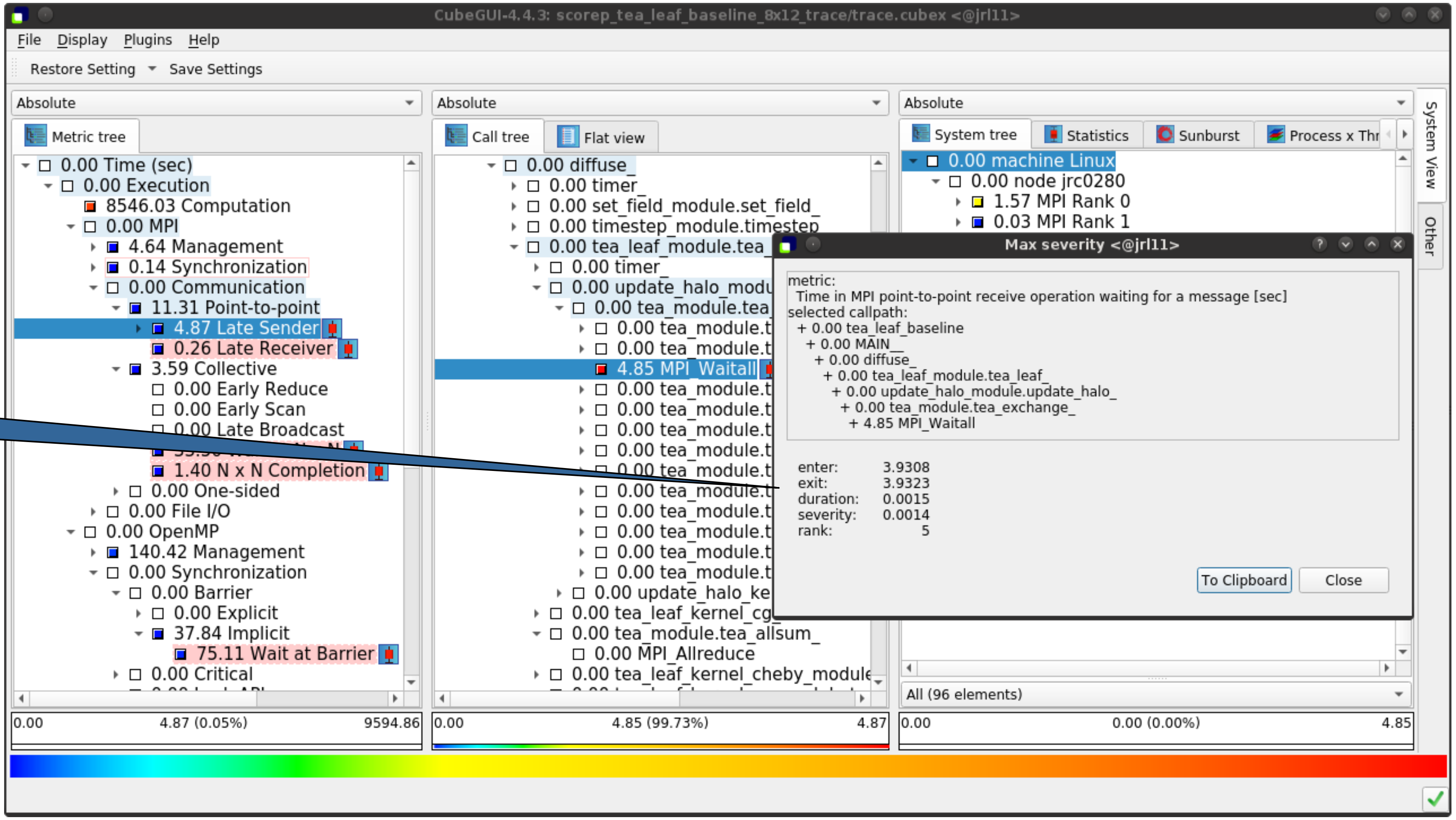
```
0.00 diffuse
  0.00 timer
  0.00 set_field_module.set_field
  0.00 timestep_module.timestep
  0.00 tea_leaf_module.tea_leaf
    0.00 timer
    0.00 update_halo_module.update_h
      0.00 tea_module.tea_exchange
        0.00 tea_module.tea_pack_rig
        0.00 tea_module.tea_send_re
          4.85 MPI_Waitall
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
        0.00 tea_mod
      0.00 update_hal
    0.00 tea_leaf_kerne
    0.00 tea_module.te
      0.00 MPI_Allredu
    0.00 tea_leaf_kerne
```

```
0.00 machine Linux
  0.00 node jrc0280
    1.57 MPI Rank 0
    0.03 MPI Rank 1
  0.00 node jrc0281
    1.00 MPI Rank 2
    1.05 MPI Rank 3
  0.00 node jrc0282
    0.05 MPI Rank 4
    0.06 MPI Rank 5
  0.00 node jrc0283
    53 MPI Rank 6
    56 MPI Rank 7
```

0.00 | 4.87 (0.05%) | 9594.86
0.00 | 4.85 (99.73%)
0.00 (0.00%) | 4.85

Access most-severe instance information for call paths marked with box plot icon via context menu

Metric instance statistics (cont.)



```
0.00 Time (sec)
  0.00 Execution
    8546.03 Computation
    0.00 MPI
      4.64 Management
      0.14 Synchronization
      0.00 Communication
        11.31 Point-to-point
          4.87 Late Sender
          0.26 Late Receiver
        3.59 Collective
          0.00 Early Reduce
          0.00 Early Scan
          0.00 Late Broadcast
          1.40 N x N Completion
        0.00 One-sided
      0.00 File I/O
    0.00 OpenMP
      140.42 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          37.84 Implicit
            75.11 Wait at Barrier
        0.00 Critical
```

```
0.00 diffuse_
  0.00 timer_
  0.00 set_field_module.set_field_
  0.00 timestep_module.timestep_
  0.00 tea_leaf_module.tea
    0.00 timer_
    0.00 update_halo_modu
      0.00 tea_module.tea
        0.00 tea_module.t
        0.00 tea_module.t
        4.85 MPI_Waitall
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
        0.00 tea_module.t
      0.00 update_halo_ke
    0.00 tea_leaf_kernel_cg
    0.00 tea_module.tea_allsum_
      0.00 MPI_Allreduce
    0.00 tea_leaf_kernel_cheby_module
```

```
0.00 machine Linux
  0.00 node jrc0280
    1.57 MPI Rank 0
    0.03 MPI Rank 1
```

```
0.00        4.87 (0.05%)     9594.86
0.00        4.85 (99.73%)    4.87
0.00        0.00 (0.00%)     4.85
```

```
metric:
 Time in MPI point-to-point receive operation waiting for a message [sec]
selected callpath:
 + 0.00 tea_leaf_baseline
  + 0.00 MAIN__
   + 0.00 diffuse_
    + 0.00 tea_leaf_module.tea_leaf_
     + 0.00 update_halo_module.update_halo_
      + 0.00 tea_module.tea_exchange_
       + 4.85 MPI_Waitall

 enter:      3.9308
 exit:       3.9323
 duration:   0.0015
 severity:   0.0014
 rank:            5
```

Shows instance details