Virtual Institute – High Productivity Supercomputing

scalasca

Performance analysis & tuning case studies

Brian Wylie & Markus Geimer
Jülich Supercomputing Centre
b.wylie@fz-juelich.de

September 2009

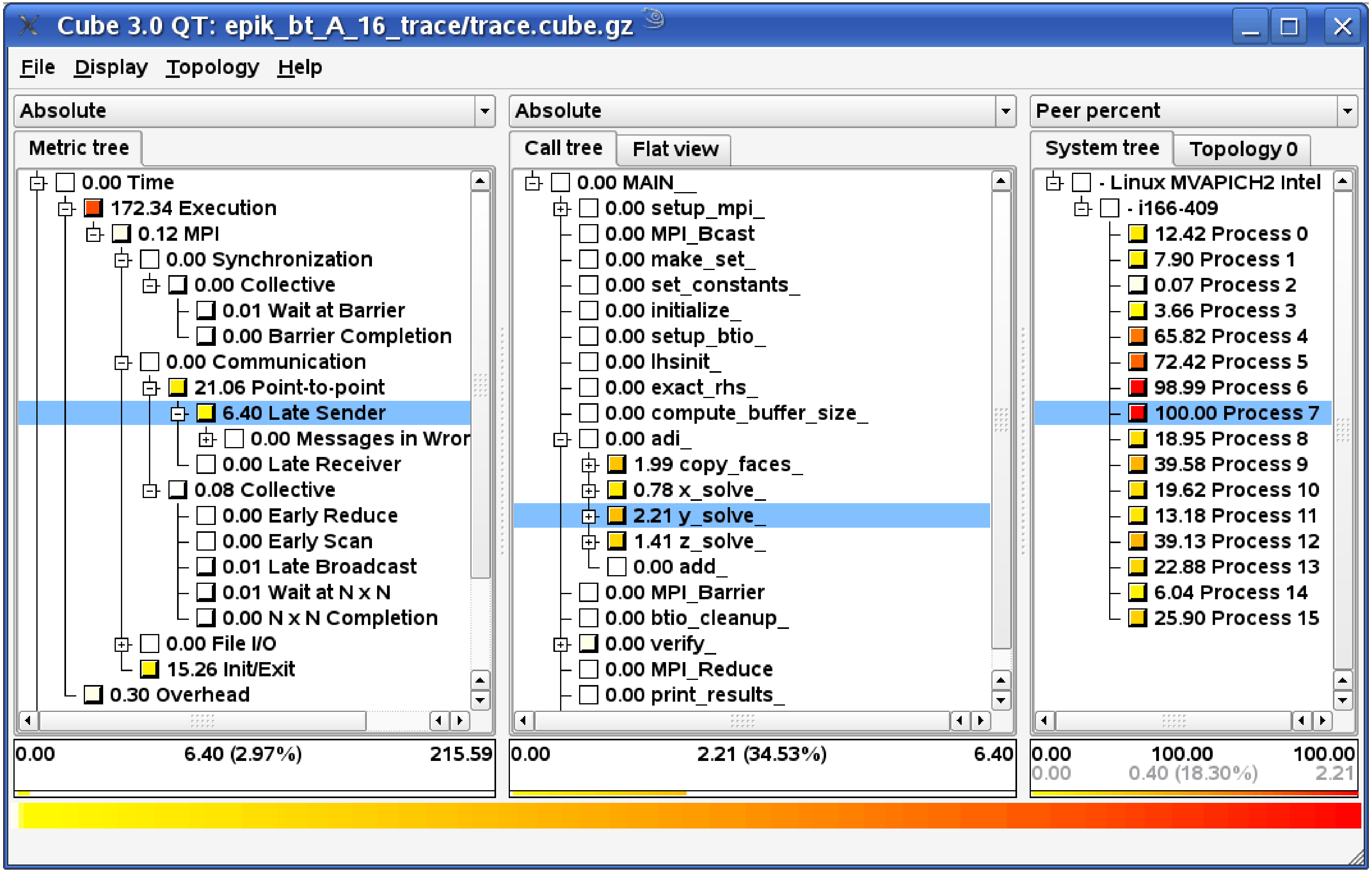
Additional Live-DVD example experiments

- Example experiment archives provided for examination:
  - jugene_sweep3d
    - 294,912 & 65,536 MPI processes on BG/P (trace)
  - jump_zeusmp2
    - 512 MPI processes on p690 cluster (summary & trace)
  - marenostrum_wrf-nmm
    - 1600 MPI processes on JS21 blade cluster, solver extract
    - summary analysis with 8 PowerPC hardware counters
    - trace analysis showing NxN completion problem on some blades
  - neptun_jacobi
    - 12 MPI processes, or 12 OpenMP threads, or 4x3 hybrid parallelizations implemented in C, C++ & Fortran on SGI Altix
  - ranger_smg2000
    - 12,288 MPI processes on Sun Constellation cluster, solve extract

Scalasca NPB-BT experiments

- Comparison of NPB-BT class A in various configurations run on a single dedicated 16-core cluster compute node
  - 16 MPI processes
    - optionally built using MPI File I/O (e.g., SUBTYPE=full)
    - optionally including PAPI counter metrics in measurement (e.g., EPK_METRICS=PAPI_FP_OPS:DISPATCH_STALLS)
  - 16 OpenMP threads
  - 4 MPI processes each with 4 OpenMP threads (MZ-MPI)
- NPB-BT-MZ class B on Cray XT5 (8-core compute nodes)
  - 32 MPI processes with OMP_NUM_THREADS=8
    - More threads created on some processes (and fewer on others) as application attempts to balance work distribution
- NPB-MPI-BT on BlueGene/P with 144k processes
  - 1536x1536x1536 gridpoints distributed on 384x384 processes

16-process trace analysis

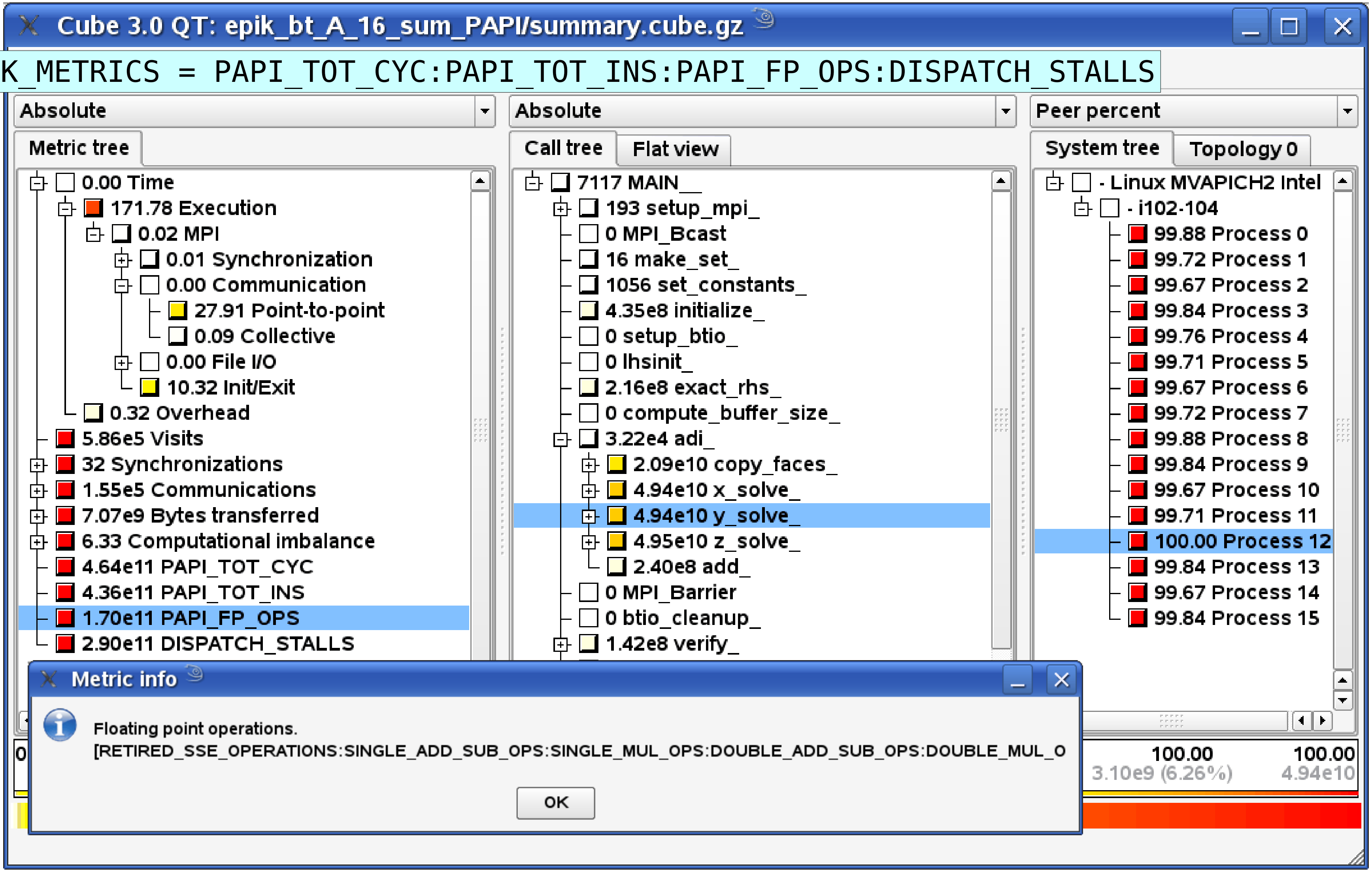


Absolute | Absolute | Peer percent

```
Metric tree                              Call tree                          System tree
□ 0.00 Time                              □ 0.00 MAIN__                      □ - Linux MVAPICH2 Intel
  ■ 172.34 Execution                       □ 0.00 setup_mpi_                  □ - i166-409
    ■ 0.12 MPI                             □ 0.00 MPI_Bcast                     ■ 12.42 Process 0
      □ 0.00 Synchronization               □ 0.00 make_set_                     ■ 7.90 Process 1
        □ 0.00 Collective                  □ 0.00 set_constants_                ■ 0.07 Process 2
          □ 0.01 Wait at Barrier           □ 0.00 initialize_                   ■ 3.66 Process 3
          □ 0.00 Barrier Completion        □ 0.00 setup_btio_                   ■ 65.82 Process 4
      □ 0.00 Communication                 □ 0.00 lhsinit_                      ■ 72.42 Process 5
        ■ 21.06 Point-to-point             □ 0.00 exact_rhs_                    ■ 98.99 Process 6
          ■ 6.40 Late Sender               □ 0.00 compute_buffer_size_          ■ 100.00 Process 7
            □ 0.00 Messages in Wror        □ 0.00 adi_                          ■ 18.95 Process 8
          □ 0.00 Late Receiver               ■ 1.99 copy_faces_                 ■ 39.58 Process 9
        □ 0.08 Collective                    ■ 0.78 x_solve_                    ■ 19.62 Process 10
          □ 0.00 Early Reduce                ■ 2.21 y_solve_                    ■ 13.18 Process 11
          □ 0.00 Early Scan                  ■ 1.41 z_solve_                    ■ 39.13 Process 12
          □ 0.01 Late Broadcast              □ 0.00 add_                        ■ 22.88 Process 13
          □ 0.01 Wait at N x N             □ 0.00 MPI_Barrier                   ■ 6.04 Process 14
          □ 0.00 N x N Completion          □ 0.00 btio_cleanup_                 ■ 25.90 Process 15
      □ 0.00 File I/O                      ■ 0.00 verify_
      ■ 15.26 Init/Exit                    □ 0.00 MPI_Reduce
  ■ 0.30 Overhead                          □ 0.00 print_results_
```

```
0.00    6.40 (2.97%)    215.59 | 0.00    2.21 (34.53%)    6.40 | 0.00    100.00           100.00
                               |                               | 0.00    0.40 (18.30%)    2.21
```

16-process summary analysis with HWC metrics



Cube 3.0 QT: epik_bt_A_16_sum_PAPI/summary.cube.gz

```
EPK_METRICS = PAPI_TOT_CYC:PAPI_TOT_INS:PAPI_FP_OPS:DISPATCH_STALLS
```

```
0.00 Time
  171.78 Execution
    0.02 MPI
      0.01 Synchronization
      0.00 Communication
        27.91 Point-to-point
        0.09 Collective
      0.00 File I/O
      10.32 Init/Exit
  0.32 Overhead
5.86e5 Visits
32 Synchronizations
1.55e5 Communications
7.07e9 Bytes transferred
6.33 Computational imbalance
4.64e11 PAPI_TOT_CYC
4.36e11 PAPI_TOT_INS
1.70e11 PAPI_FP_OPS
2.90e11 DISPATCH_STALLS
```

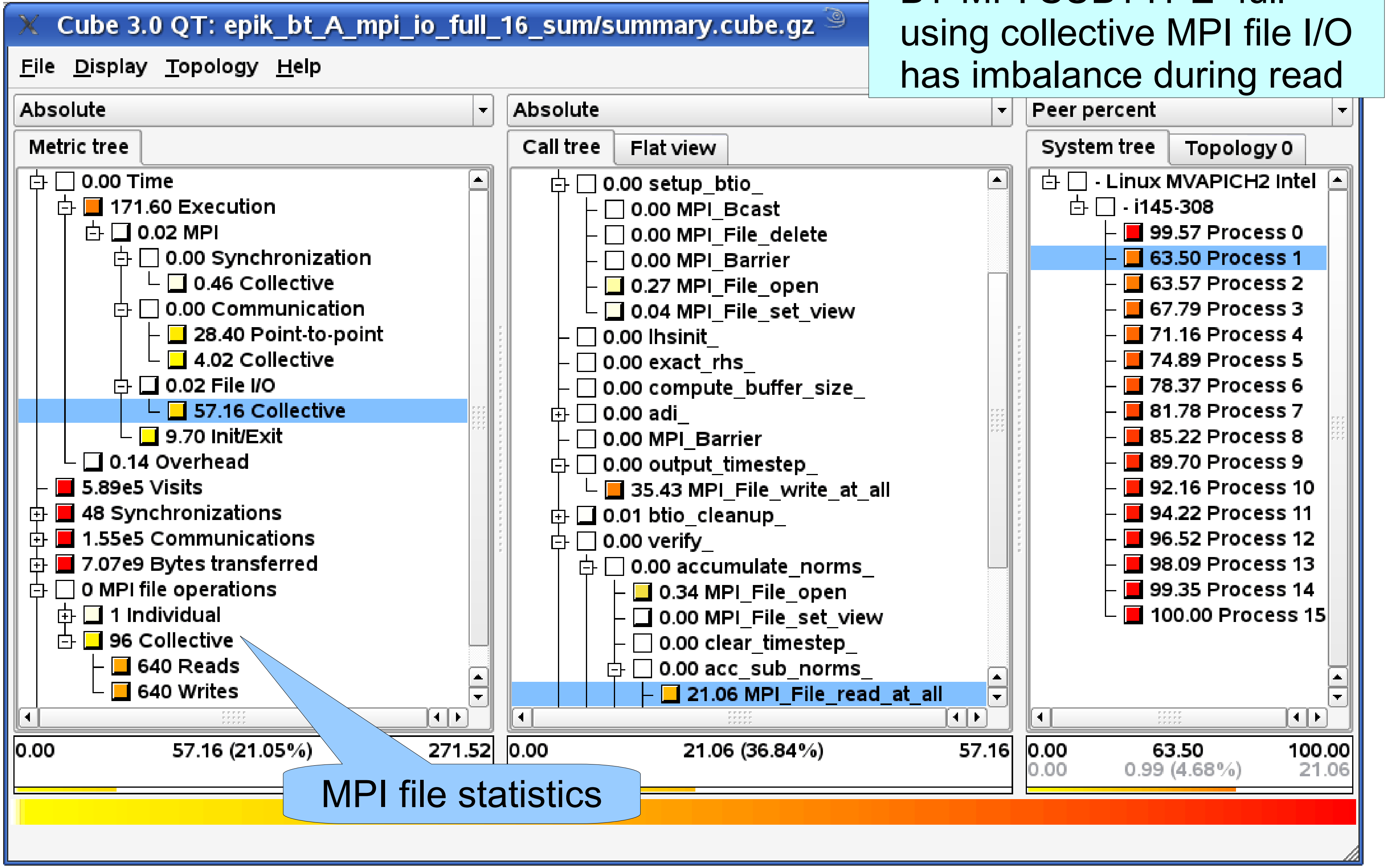
```
7117 MAIN__
  193 setup_mpi_
  0 MPI_Bcast
  16 make_set_
  1056 set_constants_
  4.35e8 initialize_
  0 setup_btio_
  0 lhsinit_
  2.16e8 exact_rhs_
  0 compute_buffer_size_
  3.22e4 adi_
    2.09e10 copy_faces_
    4.94e10 x_solve_
    4.94e10 y_solve_
    4.95e10 z_solve_
    2.40e8 add_
  0 MPI_Barrier
  0 btio_cleanup_
  1.42e8 verify_
```

```
- Linux MVAPICH2 Intel
  - i102-104
    99.88 Process 0
    99.72 Process 1
    99.67 Process 2
    99.84 Process 3
    99.76 Process 4
    99.71 Process 5
    99.67 Process 6
    99.72 Process 7
    99.88 Process 8
    99.84 Process 9
    99.67 Process 10
    99.71 Process 11
    100.00 Process 12
    99.84 Process 13
    99.67 Process 14
    99.84 Process 15
```

Metric info

Floating point operations.
[RETIRED_SSE_OPERATIONS:SINGLE_ADD_SUB_OPS:SINGLE_MUL_OPS:DOUBLE_ADD_SUB_OPS:DOUBLE_MUL_O

16-process summary analysis: MPI File I/O time



Cube 3.0 QT: epik_bt_A_mpi_io_full_16_sum/summary.cube.gz

File Display Topology Help

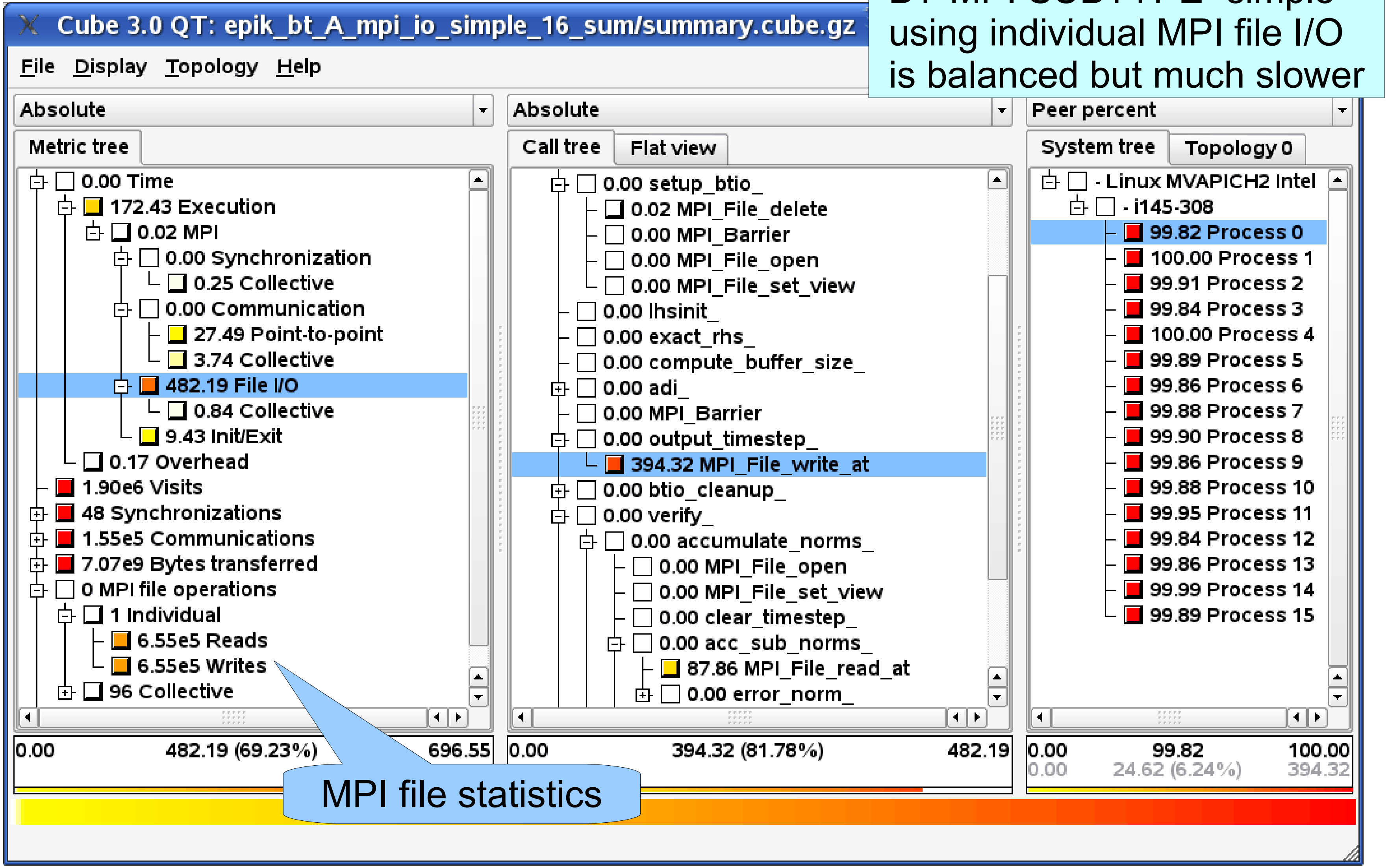
```
Absolute                              Absolute                                Peer percent
Metric tree                           Call tree | Flat view                   System tree | Topology 0

0.00 Time                             0.00 setup_btio_                        - Linux MVAPICH2 Intel
  171.60 Execution                      0.00 MPI_Bcast                          - i145-308
    0.02 MPI                            0.00 MPI_File_delete                      99.57 Process 0
      0.00 Synchronization              0.00 MPI_Barrier                          63.50 Process 1
        0.46 Collective                 0.27 MPI_File_open                        63.57 Process 2
      0.00 Communication                0.04 MPI_File_set_view                    67.79 Process 3
        28.40 Point-to-point          0.00 lhsinit_                               71.16 Process 4
        4.02 Collective               0.00 exact_rhs_                             74.89 Process 5
      0.02 File I/O                   0.00 compute_buffer_size_                   78.37 Process 6
        57.16 Collective              0.00 adi_                                   81.78 Process 7
      9.70 Init/Exit                  0.00 MPI_Barrier                            85.22 Process 8
    0.14 Overhead                     0.00 output_timestep_                       89.70 Process 9
  5.89e5 Visits                         35.43 MPI_File_write_at_all               92.16 Process 10
  48 Synchronizations                 0.01 btio_cleanup_                          94.22 Process 11
  1.55e5 Communications               0.00 verify_                                96.52 Process 12
  7.07e9 Bytes transferred              0.00 accumulate_norms_                    98.09 Process 13
  0 MPI file operations                   0.34 MPI_File_open                      99.35 Process 14
    1 Individual                          0.00 MPI_File_set_view                  100.00 Process 15
    96 Collective                         0.00 clear_timestep_
      640 Reads                           0.00 acc_sub_norms_
      640 Writes                            21.06 MPI_File_read_at_all

0.00     57.16 (21.05%)     271.52    0.00     21.06 (36.84%)     57.16      0.00     63.50          100.00
                                                                             0.00     0.99 (4.68%)   21.06
```

BT-MPI SUBTYPE=full using collective MPI file I/O has imbalance during read

MPI file statistics

16-process summary analysis: MPI File I/O time

BT-MPI SUBTYPE=simple using individual MPI file I/O is balanced but much slower



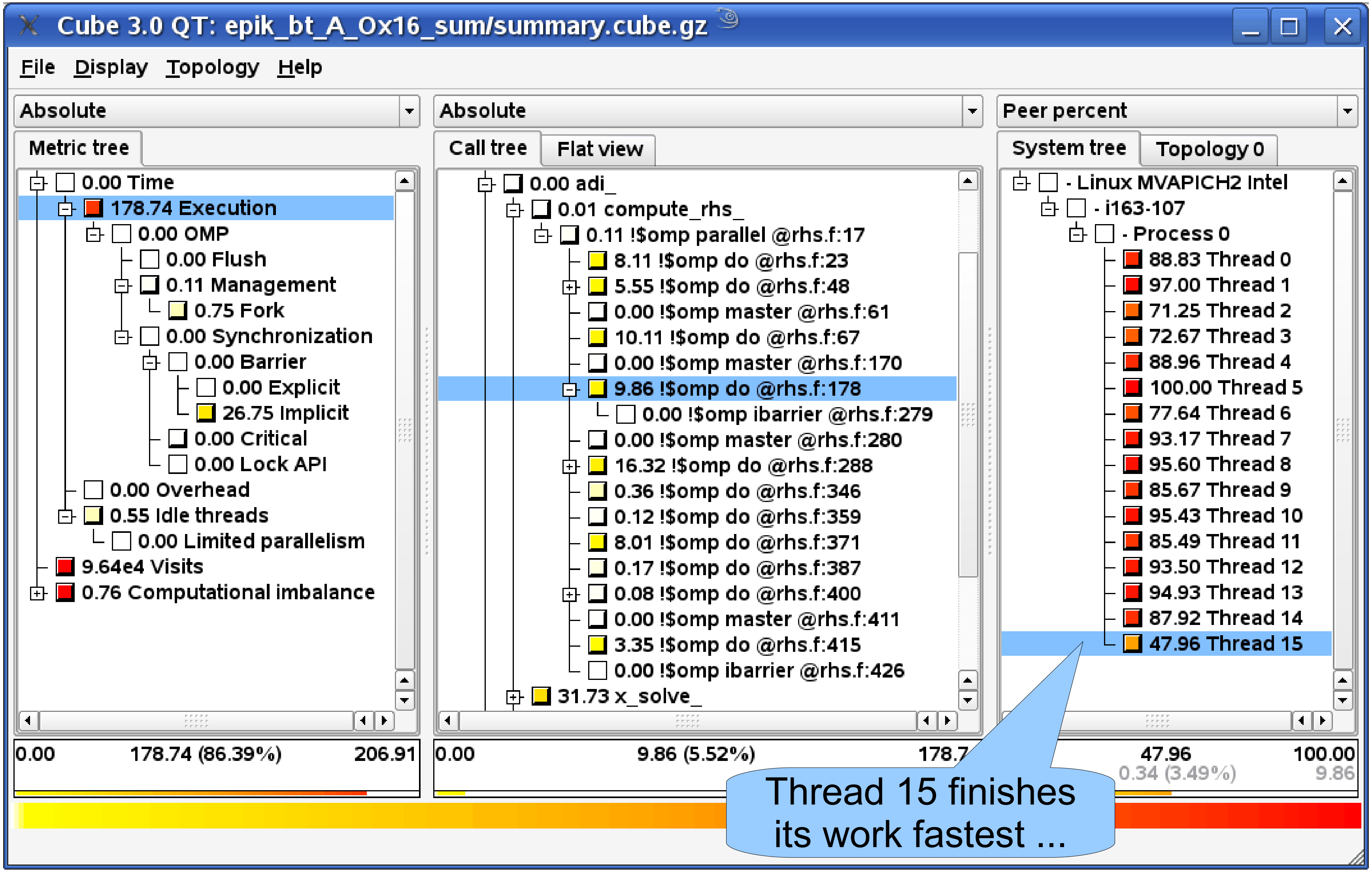
Cube 3.0 QT: epik_bt_A_mpi_io_simple_16_sum/summary.cube.gz

File Display Topology Help

```
Absolute                              Absolute                              Peer percent
Metric tree                           Call tree | Flat view                 System tree | Topology 0

□ 0.00 Time                           □ 0.00 setup_btio_                    □ - Linux MVAPICH2 Intel
  🟨 172.43 Execution                     □ 0.02 MPI_File_delete              □ - i145-308
    □ 0.02 MPI                            □ 0.00 MPI_Barrier                    🟥 99.82 Process 0
      □ 0.00 Synchronization              □ 0.00 MPI_File_open                  🟥 100.00 Process 1
        □ 0.25 Collective                 □ 0.00 MPI_File_set_view              🟥 99.91 Process 2
      □ 0.00 Communication            □ 0.00 lhsinit_                           🟥 99.84 Process 3
        🟨 27.49 Point-to-point        □ 0.00 exact_rhs_                         🟥 100.00 Process 4
        🟨 3.74 Collective             □ 0.00 compute_buffer_size_               🟥 99.89 Process 5
      🟧 482.19 File I/O              □ 0.00 adi_                               🟥 99.86 Process 6
        □ 0.84 Collective             □ 0.00 MPI_Barrier                        🟥 99.88 Process 7
      🟨 9.43 Init/Exit               □ 0.00 output_timestep_                   🟥 99.90 Process 8
  □ 0.17 Overhead                         🟧 394.32 MPI_File_write_at           🟥 99.86 Process 9
🟥 1.90e6 Visits                      □ 0.00 btio_cleanup_                      🟥 99.88 Process 10
🟥 48 Synchronizations                □ 0.00 verify_                            🟥 99.95 Process 11
🟥 1.55e5 Communications                  □ 0.00 accumulate_norms_              🟥 99.84 Process 12
🟥 7.07e9 Bytes transferred                 □ 0.00 MPI_File_open                🟥 99.86 Process 13
□ 0 MPI file operations                     □ 0.00 MPI_File_set_view            🟥 99.99 Process 14
  □ 1 Individual                            □ 0.00 clear_timestep_              🟥 99.89 Process 15
    🟧 6.55e5 Reads                         □ 0.00 acc_sub_norms_
    🟧 6.55e5 Writes                          🟨 87.86 MPI_File_read_at
  □ 96 Collective                             □ 0.00 error_norm_

0.00   482.19 (69.23%)   696.55      0.00   394.32 (81.78%)   482.19      0.00    99.82           100.00
                                                                          0.00    24.62 (6.24%)   394.32
```

MPI file statistics

16-thread summary analysis: Execution time



Cube 3.0 QT: epik_bt_A_Ox16_sum/summary.cube.gz

Absolute | Absolute | Peer percent

```
0.00 Time
  178.74 Execution
    0.00 OMP
      0.00 Flush
      0.11 Management
        0.75 Fork
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          26.75 Implicit
        0.00 Critical
        0.00 Lock API
    0.00 Overhead
    0.55 Idle threads
      0.00 Limited parallelism
9.64e4 Visits
0.76 Computational imbalance
```

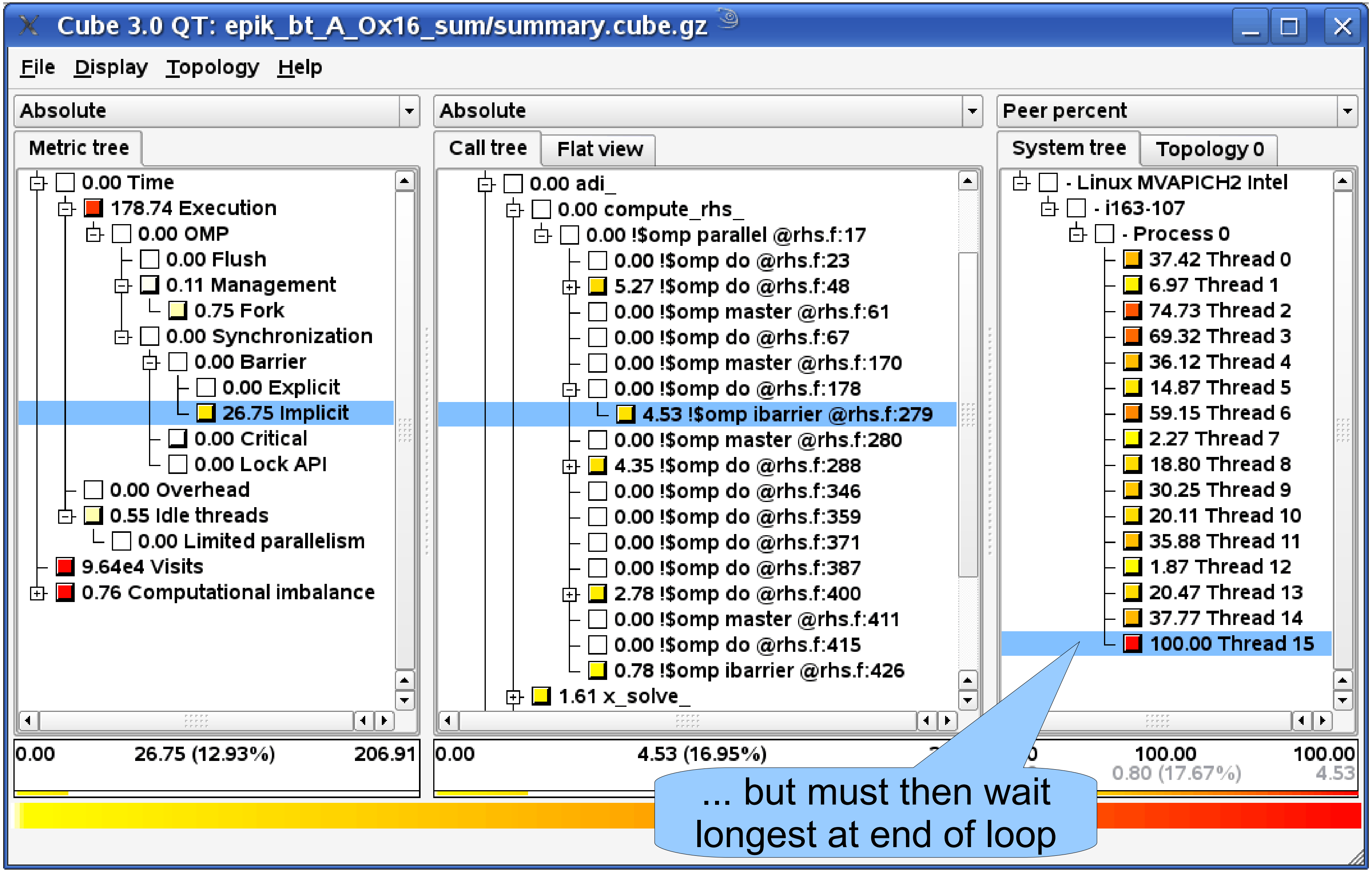
```
0.00 adi_
  0.01 compute_rhs_
    0.11 !$omp parallel @rhs.f:17
      8.11 !$omp do @rhs.f:23
      5.55 !$omp do @rhs.f:48
      0.00 !$omp master @rhs.f:61
      10.11 !$omp do @rhs.f:67
      0.00 !$omp master @rhs.f:170
      9.86 !$omp do @rhs.f:178
        0.00 !$omp ibarrier @rhs.f:279
      0.00 !$omp master @rhs.f:280
      16.32 !$omp do @rhs.f:288
      0.36 !$omp do @rhs.f:346
      0.12 !$omp do @rhs.f:359
      8.01 !$omp do @rhs.f:371
      0.17 !$omp do @rhs.f:387
      0.08 !$omp do @rhs.f:400
      0.00 !$omp master @rhs.f:411
      3.35 !$omp do @rhs.f:415
      0.00 !$omp ibarrier @rhs.f:426
  31.73 x_solve_
```

```
- Linux MVAPICH2 Intel
  - i163-107
    - Process 0
      88.83 Thread 0
      97.00 Thread 1
      71.25 Thread 2
      72.67 Thread 3
      88.96 Thread 4
      100.00 Thread 5
      77.64 Thread 6
      93.17 Thread 7
      95.60 Thread 8
      85.67 Thread 9
      95.43 Thread 10
      85.49 Thread 11
      93.50 Thread 12
      94.93 Thread 13
      87.92 Thread 14
      47.96 Thread 15
```

```
0.00    178.74 (86.39%)    206.91
0.00    9.86 (5.52%)    178.7
47.96    100.00
0.34 (3.49%)    9.86
```

Thread 15 finishes its work fastest ...

16-thread summary analysis: Implicit barrier time



Cube 3.0 QT: epik_bt_A_Ox16_sum/summary.cube.gz

```
⊟ □ 0.00 Time
   ⊟ ■ 178.74 Execution
      ⊟ □ 0.00 OMP
         □ 0.00 Flush
         ⊟ □ 0.11 Management
            ■ 0.75 Fork
         ⊟ □ 0.00 Synchronization
            ⊟ □ 0.00 Barrier
               □ 0.00 Explicit
               ■ 26.75 Implicit
            □ 0.00 Critical
            □ 0.00 Lock API
      □ 0.00 Overhead
      ⊟ ■ 0.55 Idle threads
         □ 0.00 Limited parallelism
■ 9.64e4 Visits
⊞ ■ 0.76 Computational imbalance
```

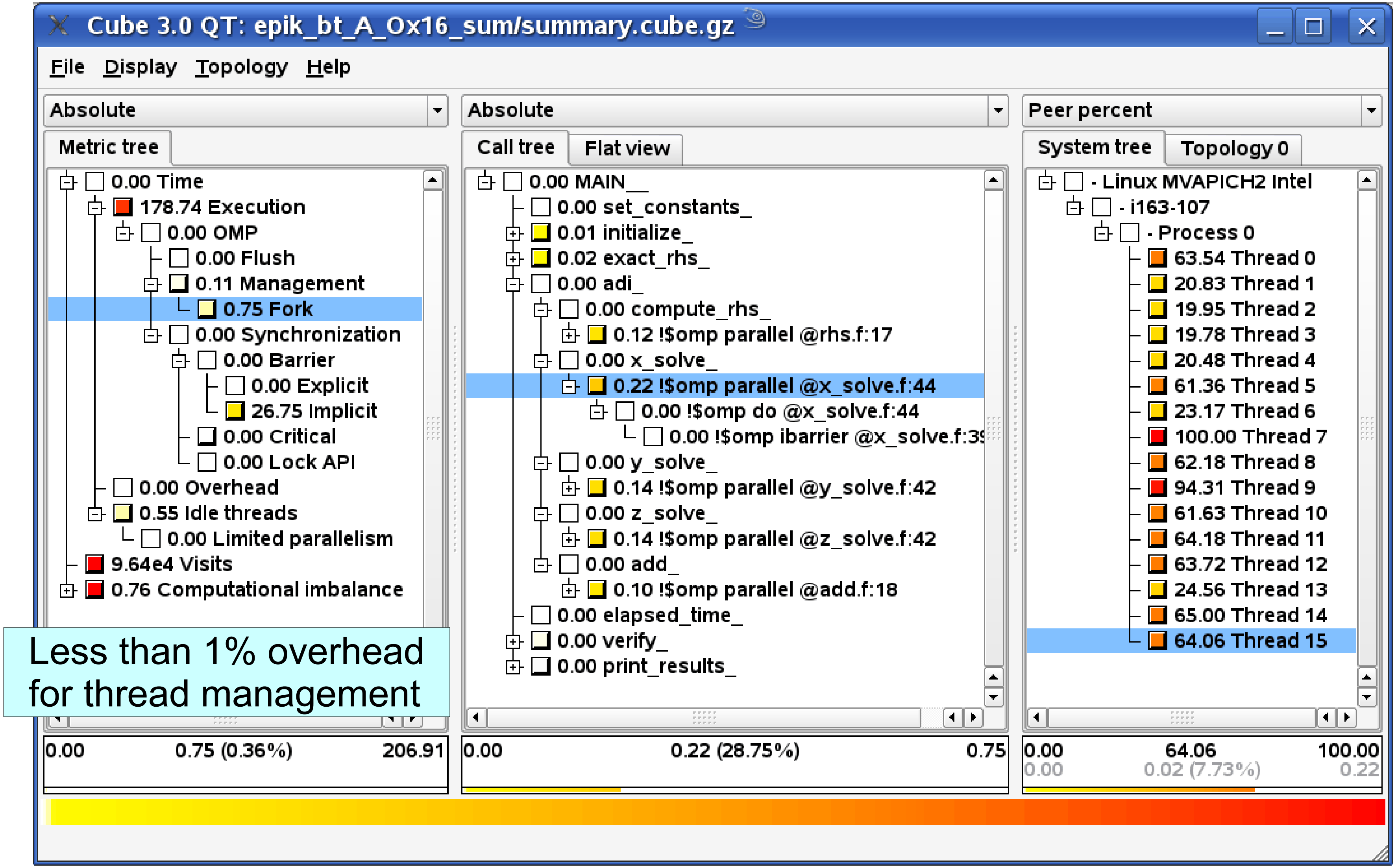
```
⊟ □ 0.00 adi_
   ⊟ □ 0.00 compute_rhs_
      ⊟ □ 0.00 !$omp parallel @rhs.f:17
         □ 0.00 !$omp do @rhs.f:23
         ⊞ ■ 5.27 !$omp do @rhs.f:48
         □ 0.00 !$omp master @rhs.f:61
         □ 0.00 !$omp do @rhs.f:67
         □ 0.00 !$omp master @rhs.f:170
         ⊟ □ 0.00 !$omp do @rhs.f:178
            ■ 4.53 !$omp ibarrier @rhs.f:279
         □ 0.00 !$omp master @rhs.f:280
         ⊞ ■ 4.35 !$omp do @rhs.f:288
         □ 0.00 !$omp do @rhs.f:346
         □ 0.00 !$omp do @rhs.f:359
         □ 0.00 !$omp do @rhs.f:371
         □ 0.00 !$omp do @rhs.f:387
         ⊞ ■ 2.78 !$omp do @rhs.f:400
         □ 0.00 !$omp master @rhs.f:411
         □ 0.00 !$omp do @rhs.f:415
         ■ 0.78 !$omp ibarrier @rhs.f:426
   ⊞ ■ 1.61 x_solve_
```

```
⊟ □ - Linux MVAPICH2 Intel
   ⊟ □ - i163-107
      ⊟ □ - Process 0
         ■ 37.42 Thread 0
         ■ 6.97 Thread 1
         ■ 74.73 Thread 2
         ■ 69.32 Thread 3
         ■ 36.12 Thread 4
         ■ 14.87 Thread 5
         ■ 59.15 Thread 6
         ■ 2.27 Thread 7
         ■ 18.80 Thread 8
         ■ 30.25 Thread 9
         ■ 20.11 Thread 10
         ■ 35.88 Thread 11
         ■ 1.87 Thread 12
         ■ 20.47 Thread 13
         ■ 37.77 Thread 14
         ■ 100.00 Thread 15
```

```
0.00      26.75 (12.93%)      206.91
0.00      4.53 (16.95%)      2
0      100.00      100.00
       0.80 (17.67%)    4.53
```

... but must then wait longest at end of loop

16-thread summary analysis: Thread fork time

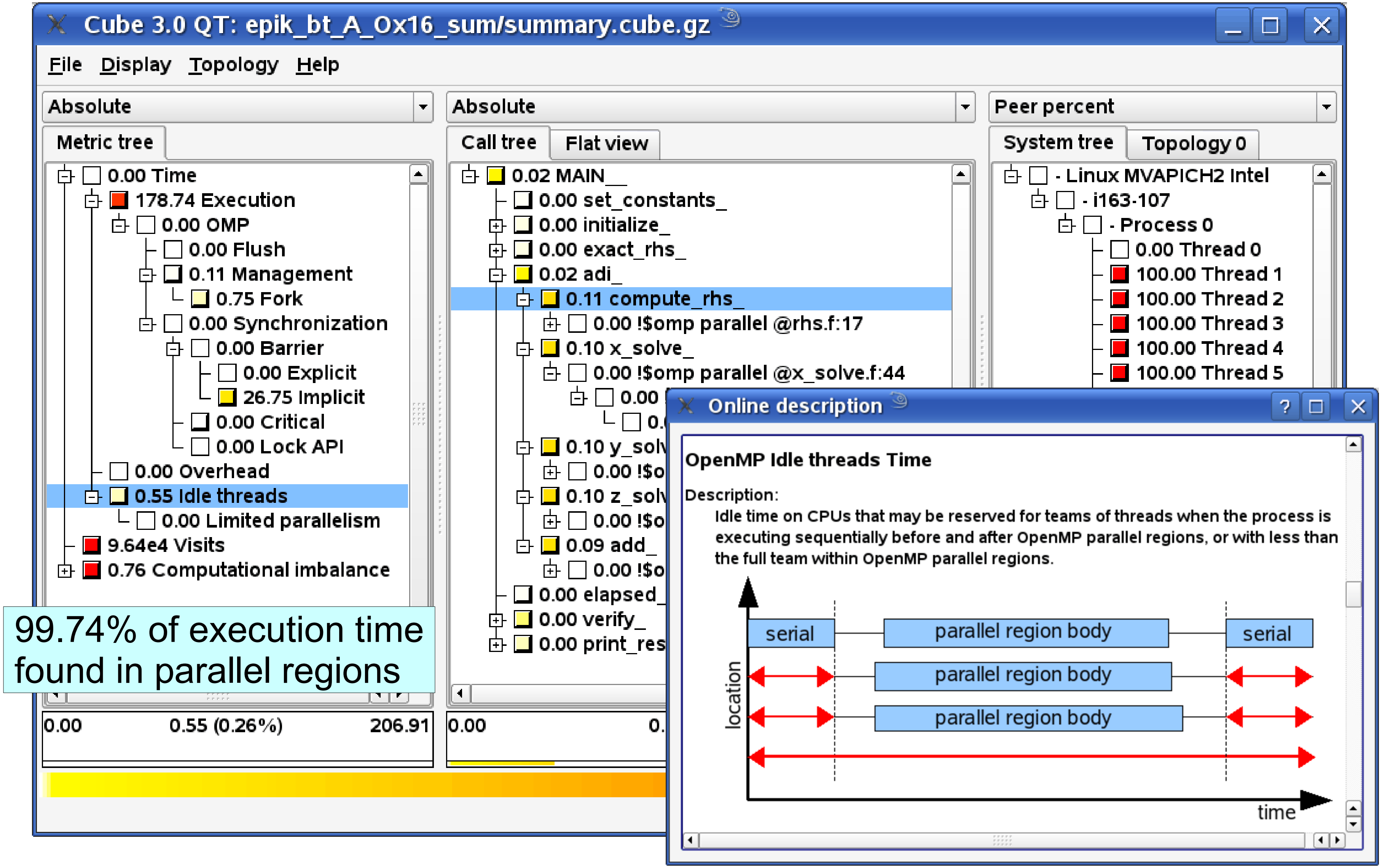


Cube 3.0 QT: epik_bt_A_Ox16_sum/summary.cube.gz

```
Absolute                          Absolute                                     Peer percent
Metric tree                       Call tree | Flat view                        System tree | Topology 0

0.00 Time                         0.00 MAIN__                                  - Linux MVAPICH2 Intel
  178.74 Execution                  0.00 set_constants_                          - i163-107
    0.00 OMP                        0.01 initialize_                               - Process 0
      0.00 Flush                    0.02 exact_rhs_                                  63.54 Thread 0
      0.11 Management               0.00 adi_                                        20.83 Thread 1
        0.75 Fork                     0.00 compute_rhs_                              19.95 Thread 2
      0.00 Synchronization              0.12 !$omp parallel @rhs.f:17                19.78 Thread 3
        0.00 Barrier                  0.00 x_solve_                                  20.48 Thread 4
          0.00 Explicit                 0.22 !$omp parallel @x_solve.f:44            61.36 Thread 5
          26.75 Implicit                  0.00 !$omp do @x_solve.f:44                23.17 Thread 6
        0.00 Critical                       0.00 !$omp ibarrier @x_solve.f:3         100.00 Thread 7
        0.00 Lock API                 0.00 y_solve_                                  62.18 Thread 8
    0.00 Overhead                       0.14 !$omp parallel @y_solve.f:42            94.31 Thread 9
    0.55 Idle threads                 0.00 z_solve_                                  61.63 Thread 10
      0.00 Limited parallelism          0.14 !$omp parallel @z_solve.f:42            64.18 Thread 11
  9.64e4 Visits                       0.00 add_                                      63.72 Thread 12
  0.76 Computational imbalance          0.10 !$omp parallel @add.f:18                24.56 Thread 13
                                    0.00 elapsed_time_                               65.00 Thread 14
                                    0.00 verify_                                     64.06 Thread 15
                                    0.00 print_results_

0.00     0.75 (0.36%)     206.91  0.00          0.22 (28.75%)          0.75    0.00       64.06        100.00
                                                                               0.00   0.02 (7.73%)      0.22
```

Less than 1% overhead for thread management

16-thread summary analysis: Idle threads time



Cube 3.0 QT: epik_bt_A_Ox16_sum/summary.cube.gz

Metric tree:

- ⊟ □ 0.00 Time
  - ⊟ 🟥 178.74 Execution
    - ⊟ □ 0.00 OMP
      - □ 0.00 Flush
      - ⊟ □ 0.11 Management
        - 🟨 0.75 Fork
      - ⊟ □ 0.00 Synchronization
        - ⊟ □ 0.00 Barrier
          - □ 0.00 Explicit
          - 🟨 26.75 Implicit
        - □ 0.00 Critical
        - □ 0.00 Lock API
    - □ 0.00 Overhead
    - ⊟ 🟨 0.55 Idle threads
      - □ 0.00 Limited parallelism
  - 🟥 9.64e4 Visits
  - ⊞ 🟥 0.76 Computational imbalance

Call tree:

- ⊟ 🟨 0.02 MAIN__
  - □ 0.00 set_constants_
  - ⊞ □ 0.00 initialize_
  - ⊞ □ 0.00 exact_rhs_
  - ⊟ 🟨 0.02 adi_
    - ⊟ 🟨 0.11 compute_rhs_
      - ⊞ □ 0.00 !$omp parallel @rhs.f:17
    - ⊟ 🟨 0.10 x_solve_
      - ⊟ □ 0.00 !$omp parallel @x_solve.f:44
        - ⊟ □ 0.00
          - □ 0.
    - ⊟ 🟨 0.10 y_solv
      - ⊞ □ 0.00 !$o
    - ⊟ 🟨 0.10 z_solv
      - ⊞ □ 0.00 !$o
    - ⊟ 🟨 0.09 add_
      - ⊞ □ 0.00 !$o
  - □ 0.00 elapsed_
  - ⊞ 🟨 0.00 verify_
  - ⊞ 🟨 0.00 print_res

System tree:

- ⊟ □ - Linux MVAPICH2 Intel
  - ⊟ □ - i163-107
    - ⊟ □ - Process 0
      - □ 0.00 Thread 0
      - 🟥 100.00 Thread 1
      - 🟥 100.00 Thread 2
      - 🟥 100.00 Thread 3
      - 🟥 100.00 Thread 4
      - 🟥 100.00 Thread 5

0.00 0.55 (0.26%) 206.91

0.00 0.

99.74% of execution time found in parallel regions

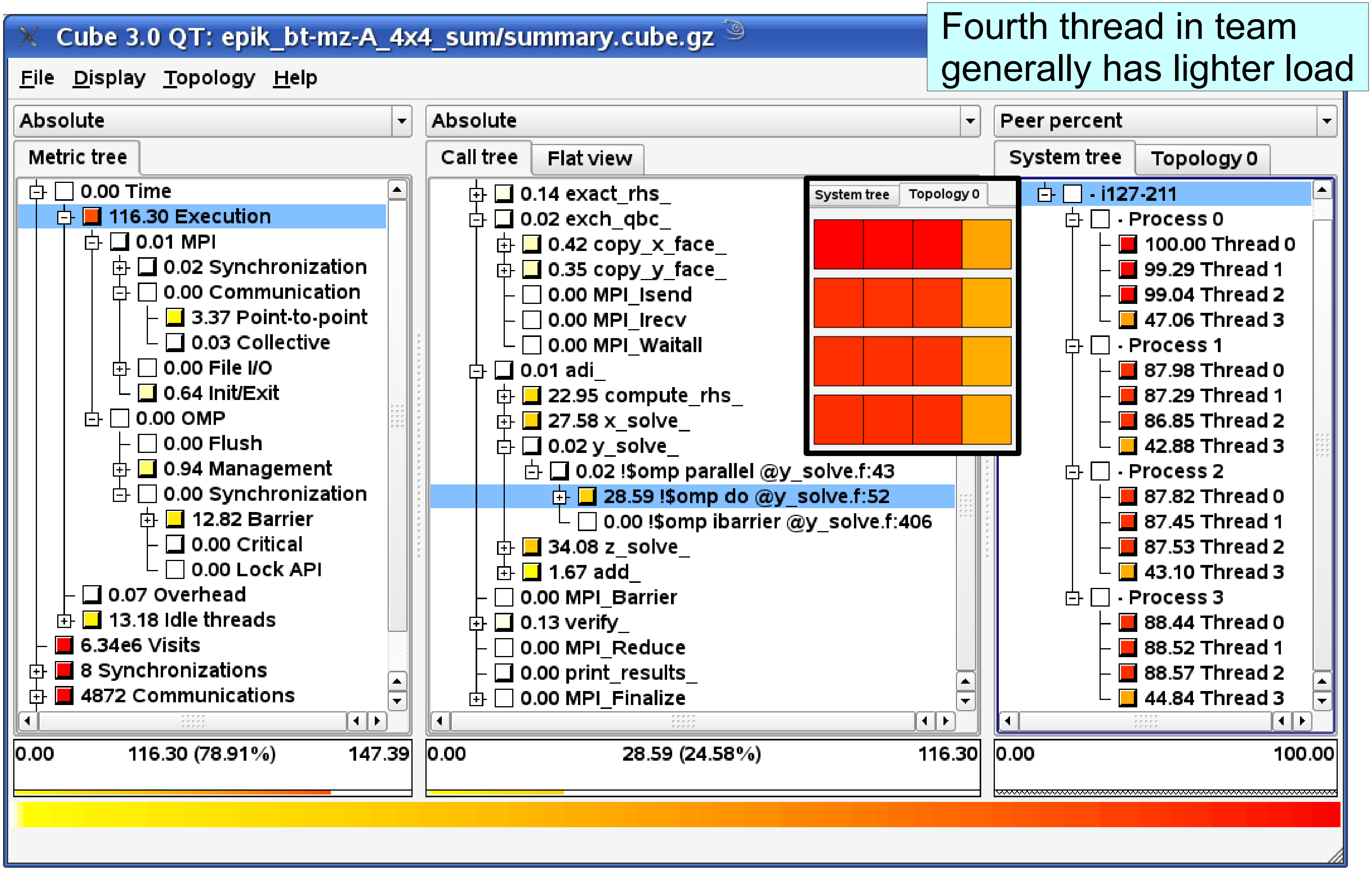
Online description

OpenMP Idle threads Time

Description:

Idle time on CPUs that may be reserved for teams of threads when the process is executing sequentially before and after OpenMP parallel regions, or with less than the full team within OpenMP parallel regions.

4x4 summary analysis: Execution time



Cube 3.0 QT: epik_bt-mz-A_4x4_sum/summary.cube.gz

Absolute | Absolute | Peer percent

```
⊟ ☐ 0.00 Time
  ⊟ 🟧 116.30 Execution
    ⊟ ☐ 0.01 MPI
      ⊞ ☐ 0.02 Synchronization
      ⊟ ☐ 0.00 Communication
        🟨 3.37 Point-to-point
        ☐ 0.03 Collective
      ⊞ ☐ 0.00 File I/O
      🟨 0.64 Init/Exit
    ⊟ ☐ 0.00 OMP
      ☐ 0.00 Flush
      ⊞ 🟨 0.94 Management
      ⊟ ☐ 0.00 Synchronization
        ⊞ 🟨 12.82 Barrier
        ☐ 0.00 Critical
        ☐ 0.00 Lock API
    ☐ 0.07 Overhead
    ⊞ 🟨 13.18 Idle threads
🟥 6.34e6 Visits
⊞ 🟥 8 Synchronizations
⊞ 🟥 4872 Communications
```

```
⊞ ☐ 0.14 exact_rhs_
⊟ ☐ 0.02 exch_qbc_
  ⊞ ☐ 0.42 copy_x_face_
  ⊞ ☐ 0.35 copy_y_face_
  ☐ 0.00 MPI_Isend
  ☐ 0.00 MPI_Irecv
  ☐ 0.00 MPI_Waitall
⊟ ☐ 0.01 adi_
  ⊞ 🟨 22.95 compute_rhs_
  ⊞ 🟨 27.58 x_solve_
  ⊟ ☐ 0.02 y_solve_
    ⊟ ☐ 0.02 !$omp parallel @y_solve.f:43
      ⊞ 🟨 28.59 !$omp do @y_solve.f:52
      ☐ 0.00 !$omp ibarrier @y_solve.f:406
  ⊞ 🟨 34.08 z_solve_
  ⊞ 🟨 1.67 add_
☐ 0.00 MPI_Barrier
⊞ ☐ 0.13 verify_
☐ 0.00 MPI_Reduce
☐ 0.00 print_results_
⊞ ☐ 0.00 MPI_Finalize
```

```
⊟ ☐ - i127-211
  ⊟ ☐ - Process 0
    🟥 100.00 Thread 0
    🟥 99.29 Thread 1
    🟥 99.04 Thread 2
    🟧 47.06 Thread 3
  ⊟ ☐ - Process 1
    🟥 87.98 Thread 0
    🟥 87.29 Thread 1
    🟥 86.85 Thread 2
    🟧 42.88 Thread 3
  ⊟ ☐ - Process 2
    🟥 87.82 Thread 0
    🟥 87.45 Thread 1
    🟥 87.53 Thread 2
    🟧 43.10 Thread 3
  ⊟ ☐ - Process 3
    🟥 88.44 Thread 0
    🟥 88.52 Thread 1
    🟥 88.57 Thread 2
    🟧 44.84 Thread 3
```

0.00 116.30 (78.91%) 147.39

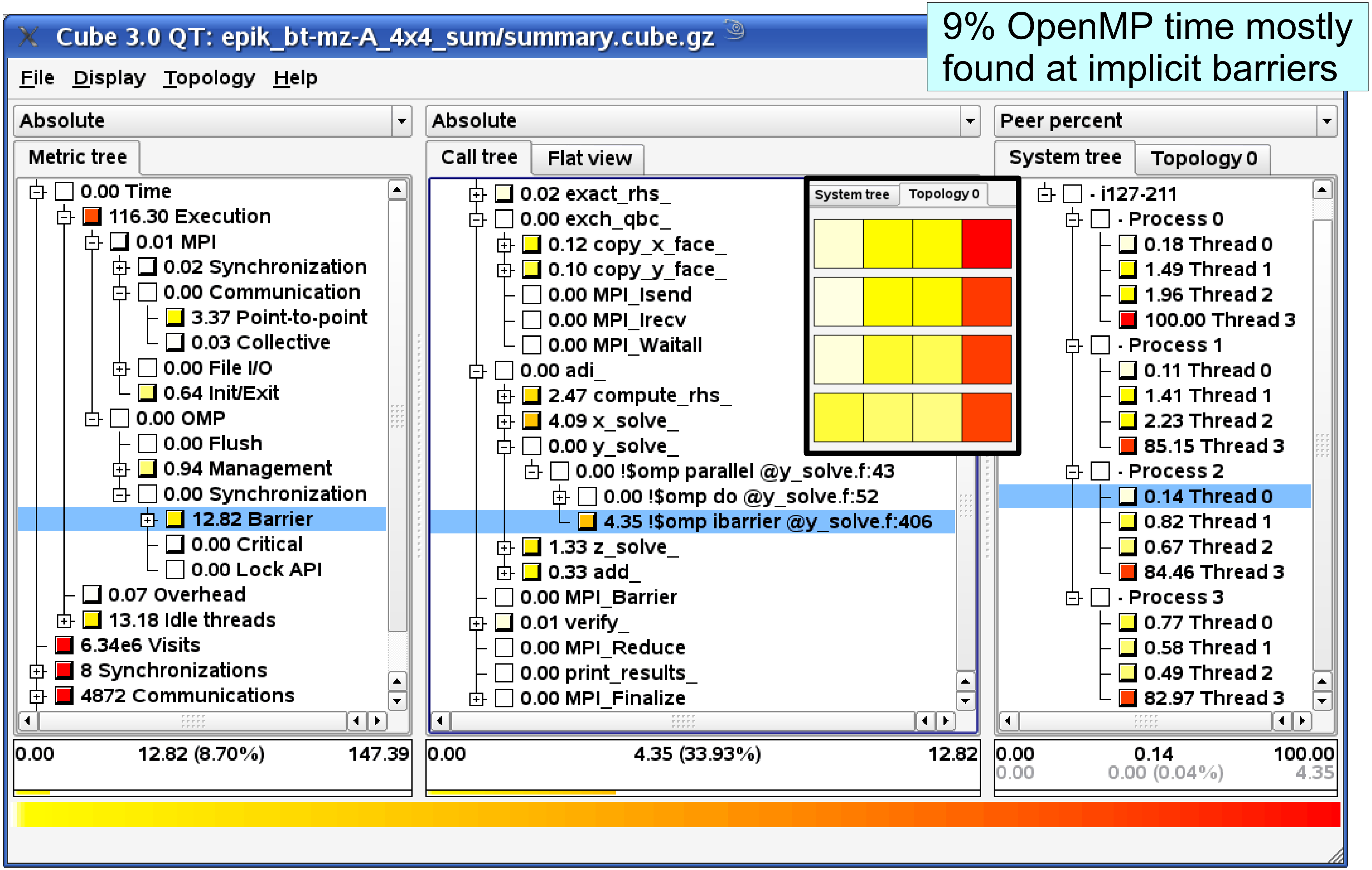
0.00 28.59 (24.58%) 116.30

0.00 100.00

Fourth thread in team generally has lighter load

4x4 summary analysis: OpenMP time

9% OpenMP time mostly found at implicit barriers



Cube 3.0 QT: epik_bt-mz-A_4x4_sum/summary.cube.gz

Absolute | Absolute | Peer percent

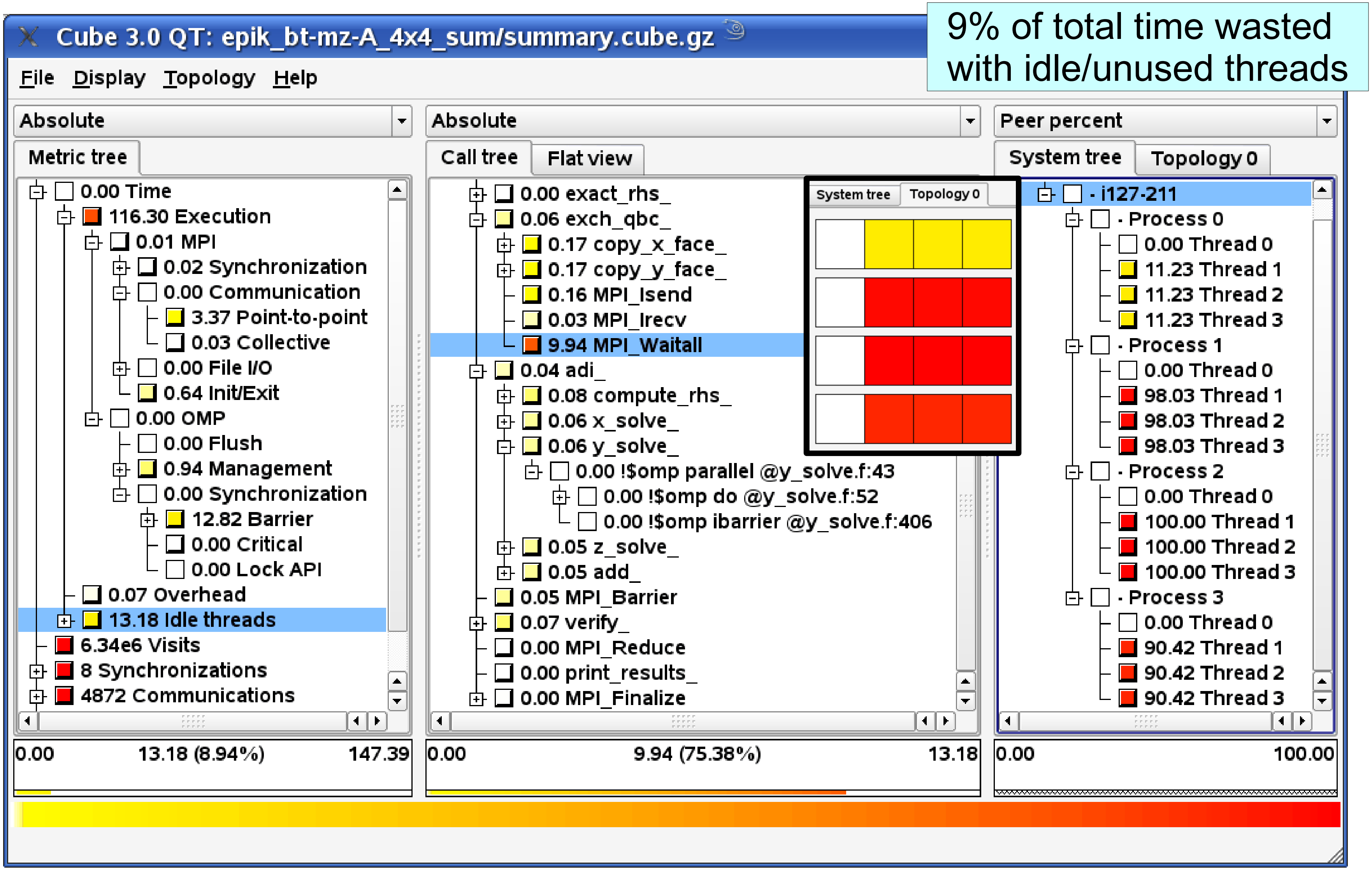
```
Metric tree
0.00 Time
  116.30 Execution
    0.01 MPI
      0.02 Synchronization
      0.00 Communication
        3.37 Point-to-point
        0.03 Collective
      0.00 File I/O
      0.64 Init/Exit
    0.00 OMP
      0.00 Flush
      0.94 Management
      0.00 Synchronization
        12.82 Barrier
        0.00 Critical
        0.00 Lock API
    0.07 Overhead
  13.18 Idle threads
6.34e6 Visits
8 Synchronizations
4872 Communications
```

```
Call tree
0.02 exact_rhs_
0.00 exch_qbc_
  0.12 copy_x_face_
  0.10 copy_y_face_
  0.00 MPI_Isend
  0.00 MPI_Irecv
  0.00 MPI_Waitall
0.00 adi_
  2.47 compute_rhs_
  4.09 x_solve_
  0.00 y_solve_
    0.00 !$omp parallel @y_solve.f:43
      0.00 !$omp do @y_solve.f:52
        4.35 !$omp ibarrier @y_solve.f:406
  1.33 z_solve_
  0.33 add_
0.00 MPI_Barrier
0.01 verify_
0.00 MPI_Reduce
0.00 print_results_
0.00 MPI_Finalize
```

```
System tree
- i127-211
  - Process 0
    0.18 Thread 0
    1.49 Thread 1
    1.96 Thread 2
    100.00 Thread 3
  - Process 1
    0.11 Thread 0
    1.41 Thread 1
    2.23 Thread 2
    85.15 Thread 3
  - Process 2
    0.14 Thread 0
    0.82 Thread 1
    0.67 Thread 2
    84.46 Thread 3
  - Process 3
    0.77 Thread 0
    0.58 Thread 1
    0.49 Thread 2
    82.97 Thread 3
```

```
0.00        12.82 (8.70%)       147.39
0.00        4.35 (33.93%)       12.82
0.00        0.14                100.00
0.00        0.00 (0.04%)        4.35
```

4x4 summary analysis: Idle threads time



Cube 3.0 QT: epik_bt-mz-A_4x4_sum/summary.cube.gz

```
0.00 Time
  116.30 Execution
    0.01 MPI
      0.02 Synchronization
      0.00 Communication
        3.37 Point-to-point
        0.03 Collective
      0.00 File I/O
      0.64 Init/Exit
    0.00 OMP
      0.00 Flush
      0.94 Management
      0.00 Synchronization
        12.82 Barrier
        0.00 Critical
        0.00 Lock API
  0.07 Overhead
  13.18 Idle threads
6.34e6 Visits
8 Synchronizations
4872 Communications
```

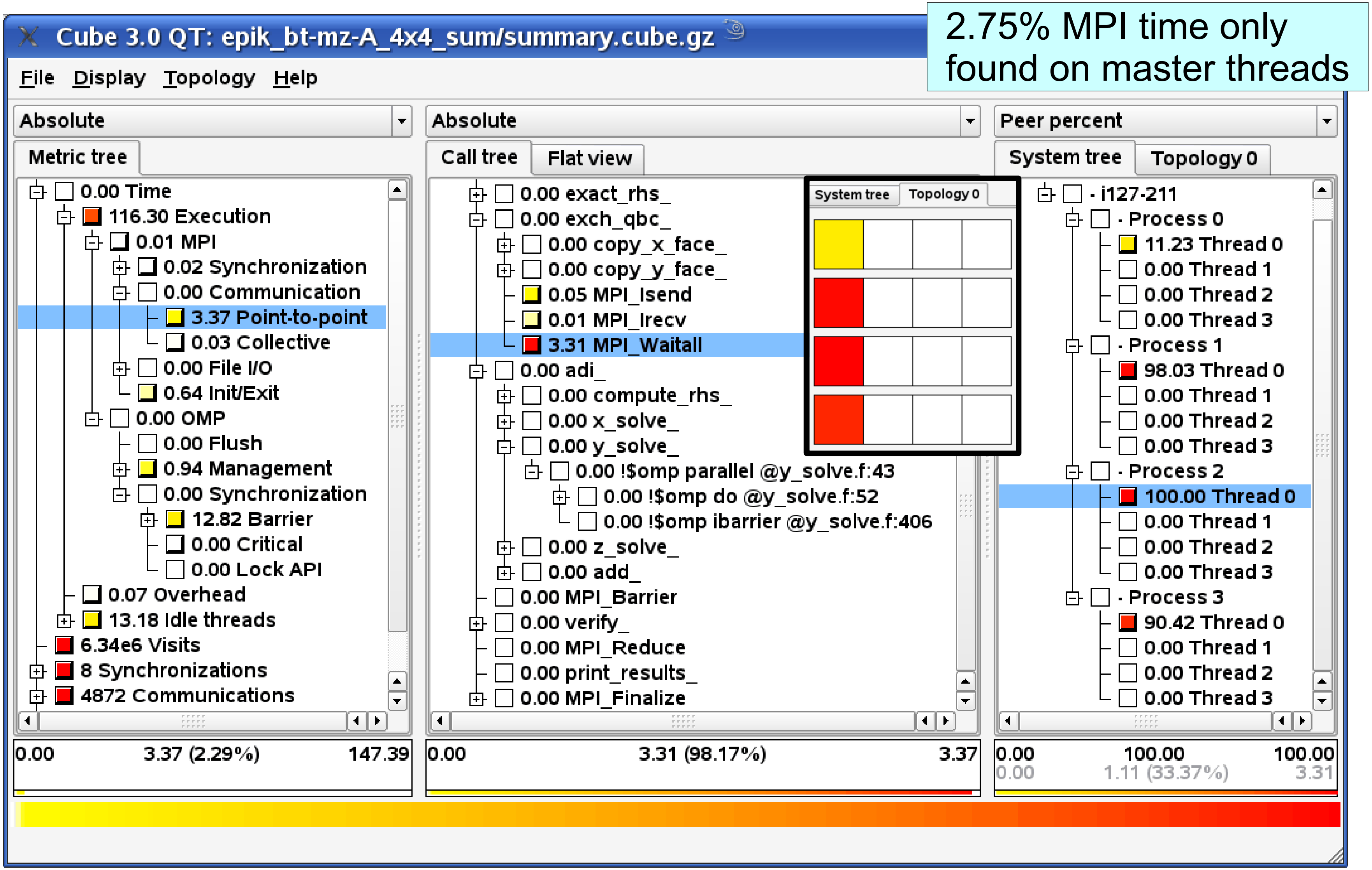
```
0.00 exact_rhs_
0.06 exch_qbc_
  0.17 copy_x_face_
  0.17 copy_y_face_
  0.16 MPI_Isend
  0.03 MPI_Irecv
  9.94 MPI_Waitall
0.04 adi_
  0.08 compute_rhs_
  0.06 x_solve_
  0.06 y_solve_
    0.00 !$omp parallel @y_solve.f:43
      0.00 !$omp do @y_solve.f:52
      0.00 !$omp ibarrier @y_solve.f:406
  0.05 z_solve_
  0.05 add_
0.05 MPI_Barrier
0.07 verify_
0.00 MPI_Reduce
0.00 print_results_
0.00 MPI_Finalize
```

```
- i127-211
  - Process 0
    0.00 Thread 0
    11.23 Thread 1
    11.23 Thread 2
    11.23 Thread 3
  - Process 1
    0.00 Thread 0
    98.03 Thread 1
    98.03 Thread 2
    98.03 Thread 3
  - Process 2
    0.00 Thread 0
    100.00 Thread 1
    100.00 Thread 2
    100.00 Thread 3
  - Process 3
    0.00 Thread 0
    90.42 Thread 1
    90.42 Thread 2
    90.42 Thread 3
```

```
0.00    13.18 (8.94%)    147.39
0.00    9.94 (75.38%)    13.18
0.00                     100.00
```

9% of total time wasted with idle/unused threads

4x4 summary analysis: MPI time



Cube 3.0 QT: epik_bt-mz-A_4x4_sum/summary.cube.gz

Absolute | Absolute | Peer percent

```
Metric tree:
□ 0.00 Time
  ■ 116.30 Execution
    □ 0.01 MPI
      □ 0.02 Synchronization
      □ 0.00 Communication
        ■ 3.37 Point-to-point
        □ 0.03 Collective
      □ 0.00 File I/O
      ■ 0.64 Init/Exit
    □ 0.00 OMP
      □ 0.00 Flush
      ■ 0.94 Management
      □ 0.00 Synchronization
        ■ 12.82 Barrier
        □ 0.00 Critical
        □ 0.00 Lock API
    □ 0.07 Overhead
  ■ 13.18 Idle threads
■ 6.34e6 Visits
■ 8 Synchronizations
■ 4872 Communications
```

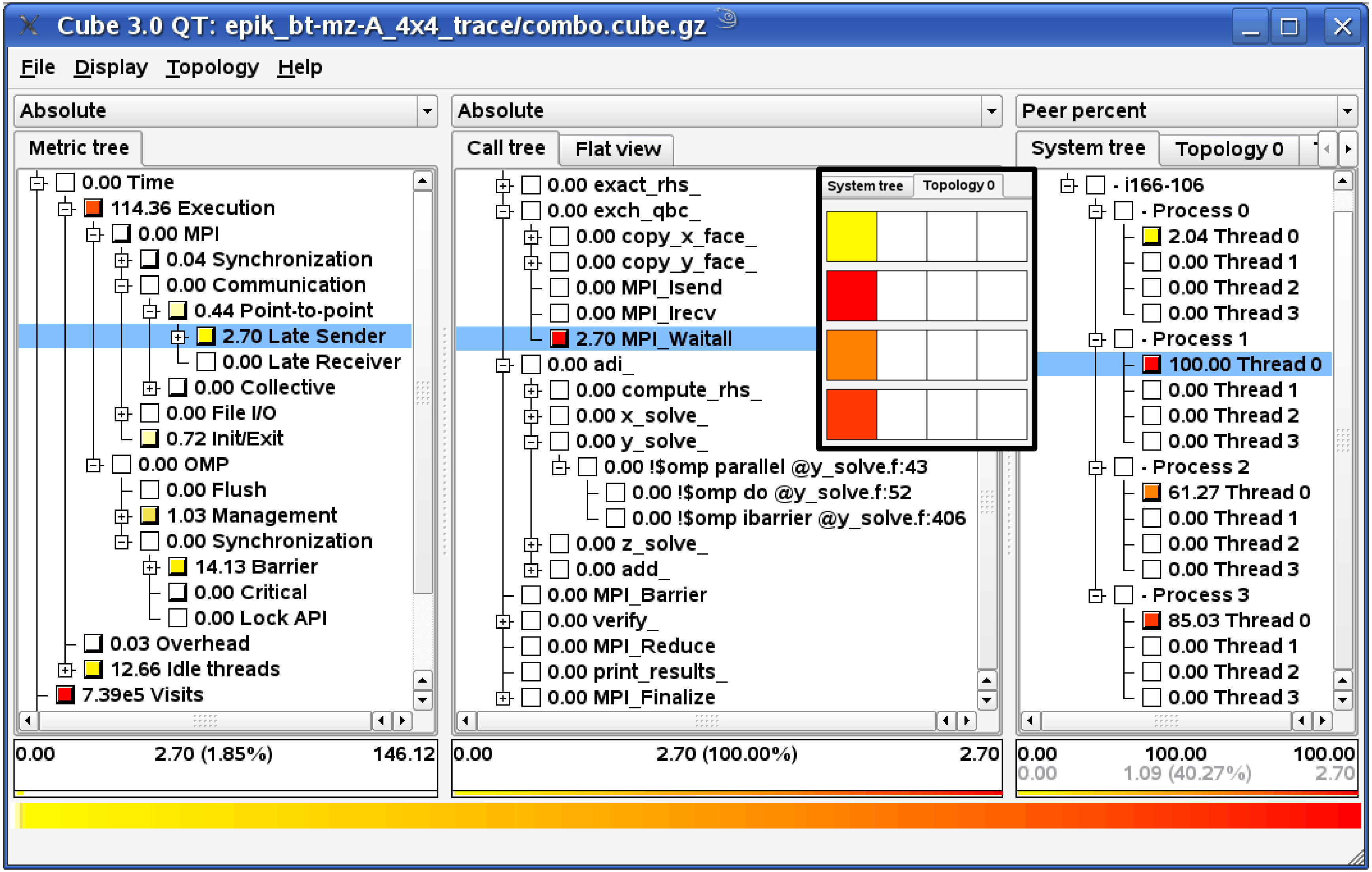
```
Call tree:
□ 0.00 exact_rhs_
□ 0.00 exch_qbc_
  □ 0.00 copy_x_face_
  □ 0.00 copy_y_face_
  ■ 0.05 MPI_Isend
  ■ 0.01 MPI_Irecv
  ■ 3.31 MPI_Waitall
□ 0.00 adi_
  □ 0.00 compute_rhs_
  □ 0.00 x_solve_
  □ 0.00 y_solve_
    □ 0.00 !$omp parallel @y_solve.f:43
      □ 0.00 !$omp do @y_solve.f:52
      □ 0.00 !$omp ibarrier @y_solve.f:406
  □ 0.00 z_solve_
  □ 0.00 add_
□ 0.00 MPI_Barrier
□ 0.00 verify_
□ 0.00 MPI_Reduce
□ 0.00 print_results_
□ 0.00 MPI_Finalize
```

```
System tree:
□ - i127-211
  □ - Process 0
    ■ 11.23 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
  □ - Process 1
    ■ 98.03 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
  □ - Process 2
    ■ 100.00 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
  □ - Process 3
    ■ 90.42 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
```

```
0.00    3.37 (2.29%)    147.39
0.00    3.31 (98.17%)   3.37
0.00    100.00          100.00
0.00    1.11 (33.37%)   3.31
```

2.75% MPI time only found on master threads

4x4 combined summary & trace analysis



Cube 3.0 QT: epik_bt-mz-A_4x4_trace/combo.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0

```
□ 0.00 Time
  🟥 114.36 Execution
    □ 0.00 MPI
      □ 0.04 Synchronization
      □ 0.00 Communication
        🟨 0.44 Point-to-point
          🟨 2.70 Late Sender
          □ 0.00 Late Receiver
        □ 0.00 Collective
      □ 0.00 File I/O
      🟨 0.72 Init/Exit
    □ 0.00 OMP
      □ 0.00 Flush
      🟨 1.03 Management
      □ 0.00 Synchronization
        🟨 14.13 Barrier
        □ 0.00 Critical
        □ 0.00 Lock API
    □ 0.03 Overhead
  🟨 12.66 Idle threads
🟥 7.39e5 Visits
```

```
□ 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
  🟥 2.70 MPI_Waitall
□ 0.00 adi_
  □ 0.00 compute_rhs_
  □ 0.00 x_solve_
  □ 0.00 y_solve_
    □ 0.00 !$omp parallel @y_solve.f:43
      □ 0.00 !$omp do @y_solve.f:52
      □ 0.00 !$omp ibarrier @y_solve.f:406
  □ 0.00 z_solve_
  □ 0.00 add_
□ 0.00 MPI_Barrier
□ 0.00 verify_
□ 0.00 MPI_Reduce
□ 0.00 print_results_
□ 0.00 MPI_Finalize
```

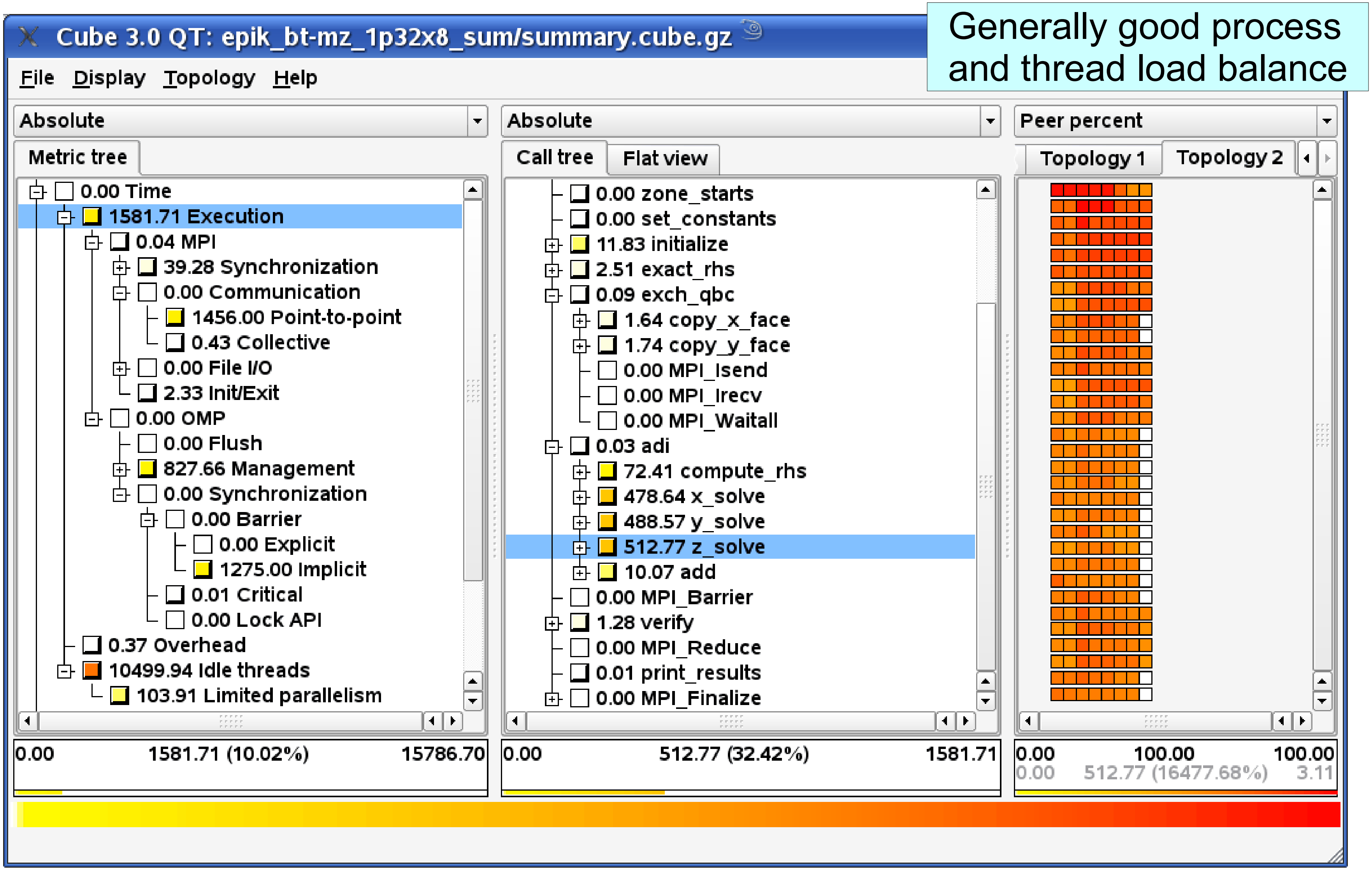
System tree | Topology 0

```
□ - i166-106
  □ - Process 0
    🟨 2.04 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
  □ - Process 1
    🟥 100.00 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
  □ - Process 2
    🟧 61.27 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
  □ - Process 3
    🟥 85.03 Thread 0
    □ 0.00 Thread 1
    □ 0.00 Thread 2
    □ 0.00 Thread 3
```

```
0.00     2.70 (1.85%)     146.12
0.00     2.70 (100.00%)   2.70
0.00     100.00           100.00
0.00     1.09 (40.27%)    2.70
```

32x8 summary analysis: Excl. execution time

Generally good process and thread load balance



Cube 3.0 QT: epik_bt-mz_1p32x8_sum/summary.cube.gz

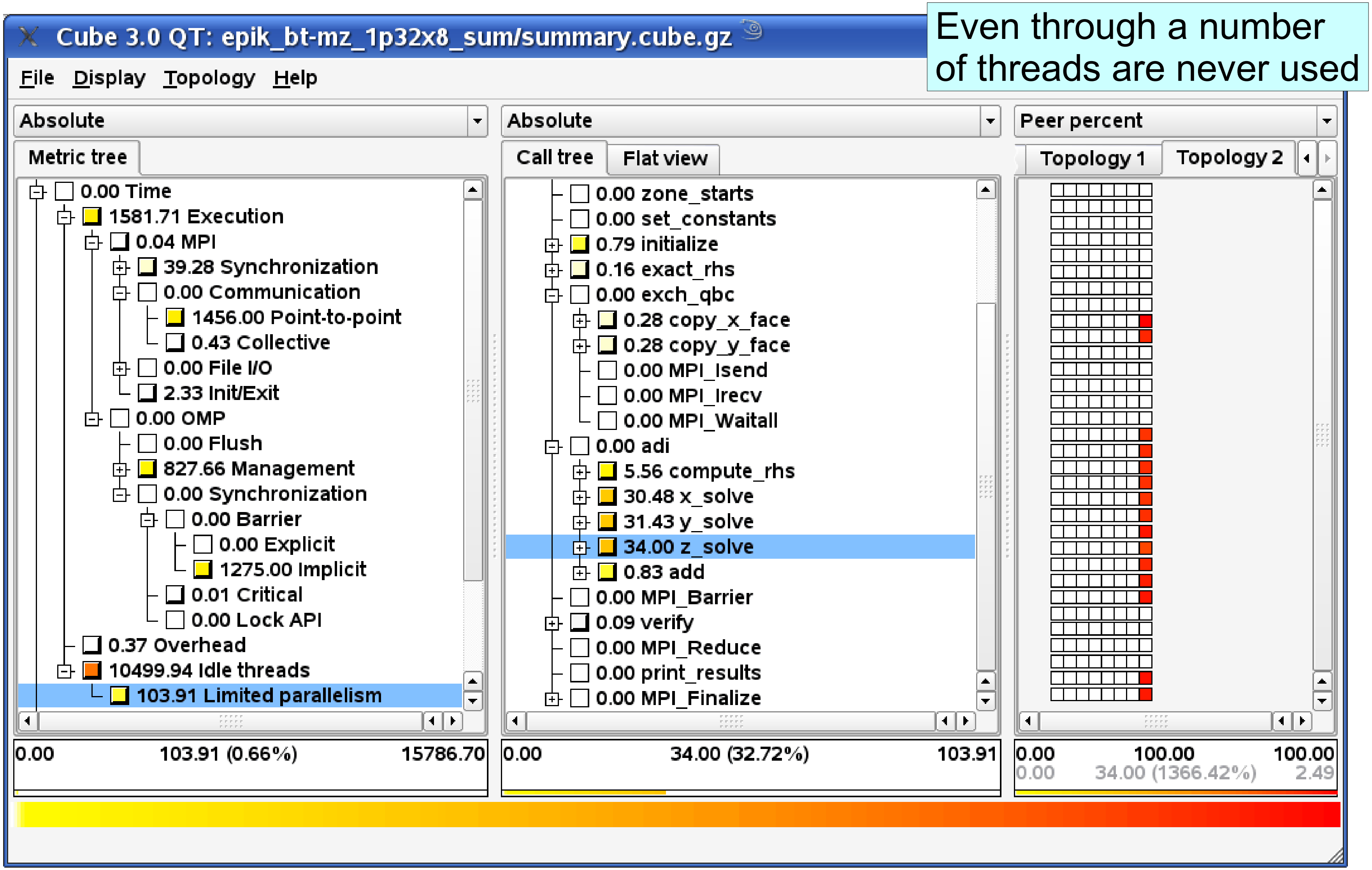
```
Metric tree
□ 0.00 Time
  □ 1581.71 Execution
    □ 0.04 MPI
      □ 39.28 Synchronization
      □ 0.00 Communication
        □ 1456.00 Point-to-point
        □ 0.43 Collective
      □ 0.00 File I/O
      □ 2.33 Init/Exit
    □ 0.00 OMP
      □ 0.00 Flush
      □ 827.66 Management
      □ 0.00 Synchronization
        □ 0.00 Barrier
          □ 0.00 Explicit
          □ 1275.00 Implicit
        □ 0.01 Critical
        □ 0.00 Lock API
  □ 0.37 Overhead
  □ 10499.94 Idle threads
    □ 103.91 Limited parallelism

0.00    1581.71 (10.02%)    15786.70
```

```
Call tree
  □ 0.00 zone_starts
  □ 0.00 set_constants
  □ 11.83 initialize
  □ 2.51 exact_rhs
  □ 0.09 exch_qbc
    □ 1.64 copy_x_face
    □ 1.74 copy_y_face
    □ 0.00 MPI_Isend
    □ 0.00 MPI_Irecv
    □ 0.00 MPI_Waitall
  □ 0.03 adi
    □ 72.41 compute_rhs
    □ 478.64 x_solve
    □ 488.57 y_solve
    □ 512.77 z_solve
    □ 10.07 add
  □ 0.00 MPI_Barrier
  □ 1.28 verify
  □ 0.00 MPI_Reduce
  □ 0.01 print_results
  □ 0.00 MPI_Finalize

0.00    512.77 (32.42%)    1581.71
```

```
0.00    100.00                100.00
0.00    512.77 (16477.68%)    3.11
```

32x8 summary analysis: Limited parallelism

Even through a number of threads are never used



Cube 3.0 QT: epik_bt-mz_1p32x8_sum/summary.cube.gz

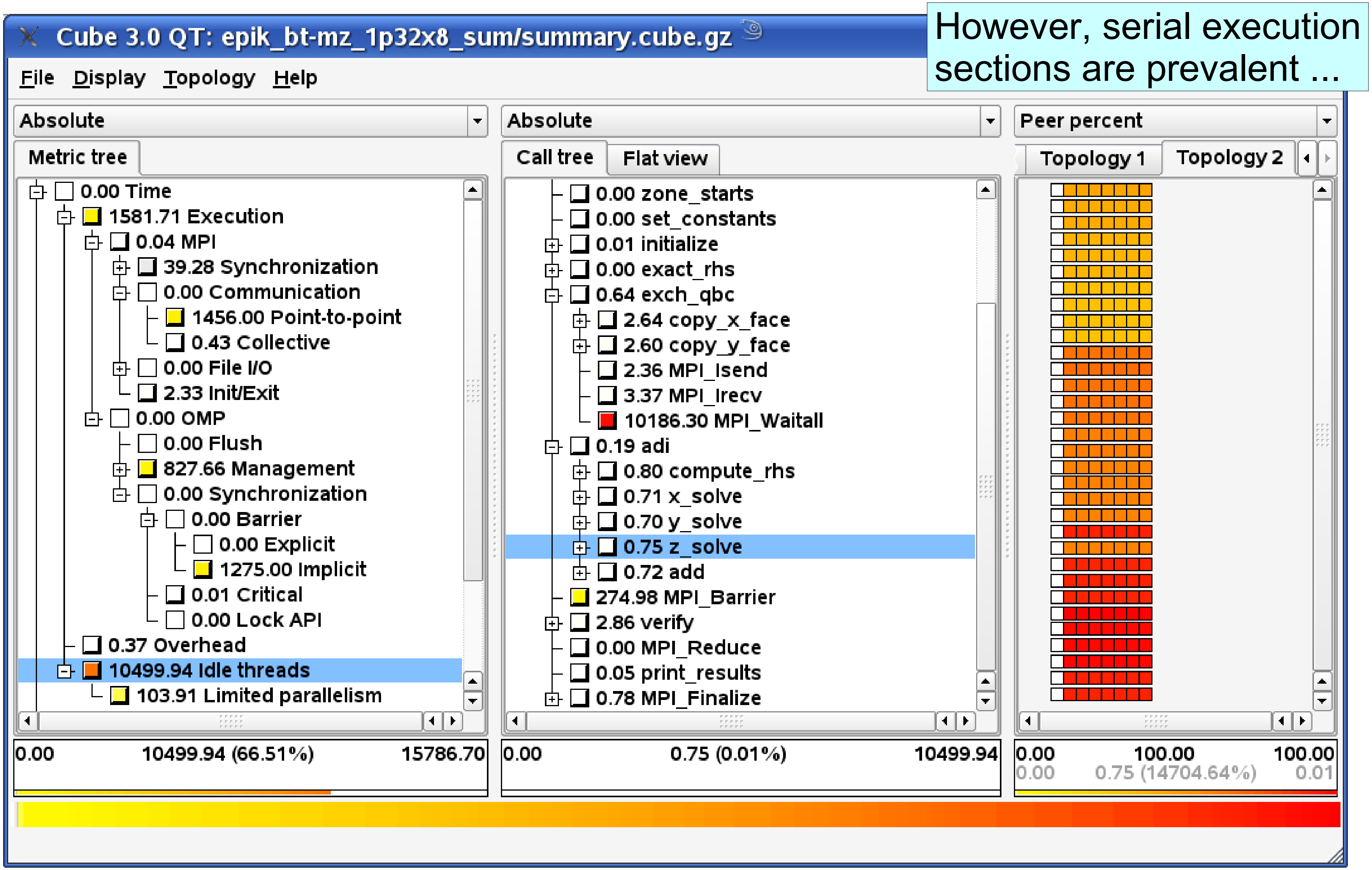
Absolute | Absolute | Peer percent

```
0.00 Time
  1581.71 Execution
    0.04 MPI
      39.28 Synchronization
      0.00 Communication
        1456.00 Point-to-point
        0.43 Collective
      0.00 File I/O
      2.33 Init/Exit
    0.00 OMP
      0.00 Flush
      827.66 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          1275.00 Implicit
        0.01 Critical
        0.00 Lock API
  0.37 Overhead
  10499.94 Idle threads
    103.91 Limited parallelism
```

```
0.00 zone_starts
0.00 set_constants
0.79 initialize
0.16 exact_rhs
0.00 exch_qbc
  0.28 copy_x_face
  0.28 copy_y_face
  0.00 MPI_Isend
  0.00 MPI_Irecv
  0.00 MPI_Waitall
0.00 adi
  5.56 compute_rhs
  30.48 x_solve
  31.43 y_solve
  34.00 z_solve
  0.83 add
0.00 MPI_Barrier
0.09 verify
0.00 MPI_Reduce
0.00 print_results
0.00 MPI_Finalize
```

```
0.00        103.91 (0.66%)      15786.70
0.00        34.00 (32.72%)      103.91
0.00        100.00              100.00
0.00        34.00 (1366.42%)    2.49
```

32x8 summary analysis: Idle threads time



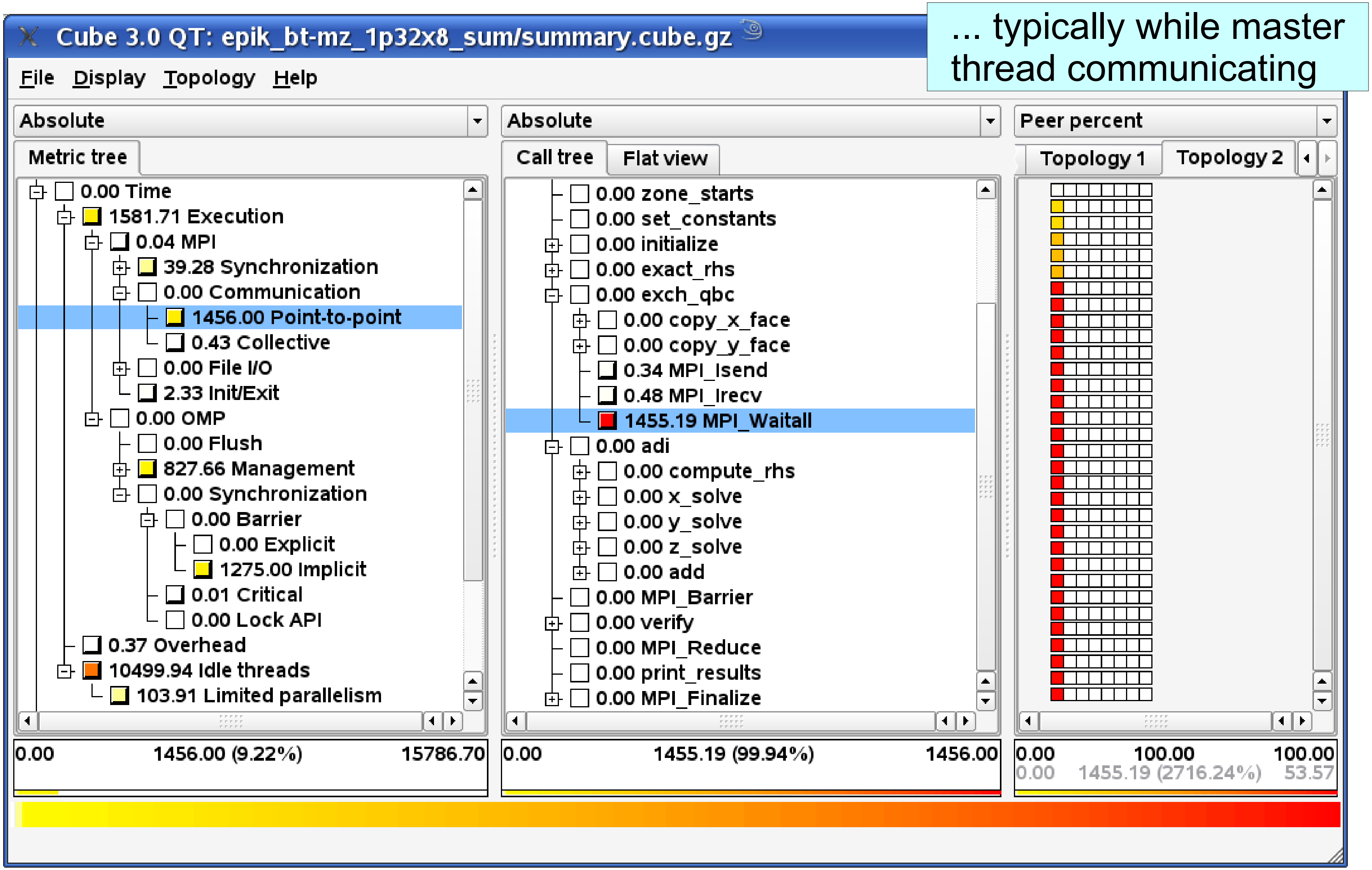
However, serial execution sections are prevalent ...

```
0.00 Time
  1581.71 Execution
    0.04 MPI
      39.28 Synchronization
      0.00 Communication
        1456.00 Point-to-point
        0.43 Collective
      0.00 File I/O
      2.33 Init/Exit
    0.00 OMP
      0.00 Flush
      827.66 Management
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          1275.00 Implicit
        0.01 Critical
        0.00 Lock API
  0.37 Overhead
  10499.94 Idle threads
    103.91 Limited parallelism
```

```
0.00 zone_starts
0.00 set_constants
0.01 initialize
0.00 exact_rhs
0.64 exch_qbc
  2.64 copy_x_face
  2.60 copy_y_face
  2.36 MPI_Isend
  3.37 MPI_Irecv
  10186.30 MPI_Waitall
0.19 adi
  0.80 compute_rhs
  0.71 x_solve
  0.70 y_solve
  0.75 z_solve
  0.72 add
274.98 MPI_Barrier
2.86 verify
0.00 MPI_Reduce
0.05 print_results
0.78 MPI_Finalize
```

```
0.00    10499.94 (66.51%)    15786.70
0.00    0.75 (0.01%)    10499.94
0.00    100.00    100.00
0.00    0.75 (14704.64%)    0.01
```

32x8 summary analysis: MPI communication time



Cube 3.0 QT: epik_bt-mz_1p32x8_sum/summary.cube.gz

... typically while master thread communicating

Metric tree:

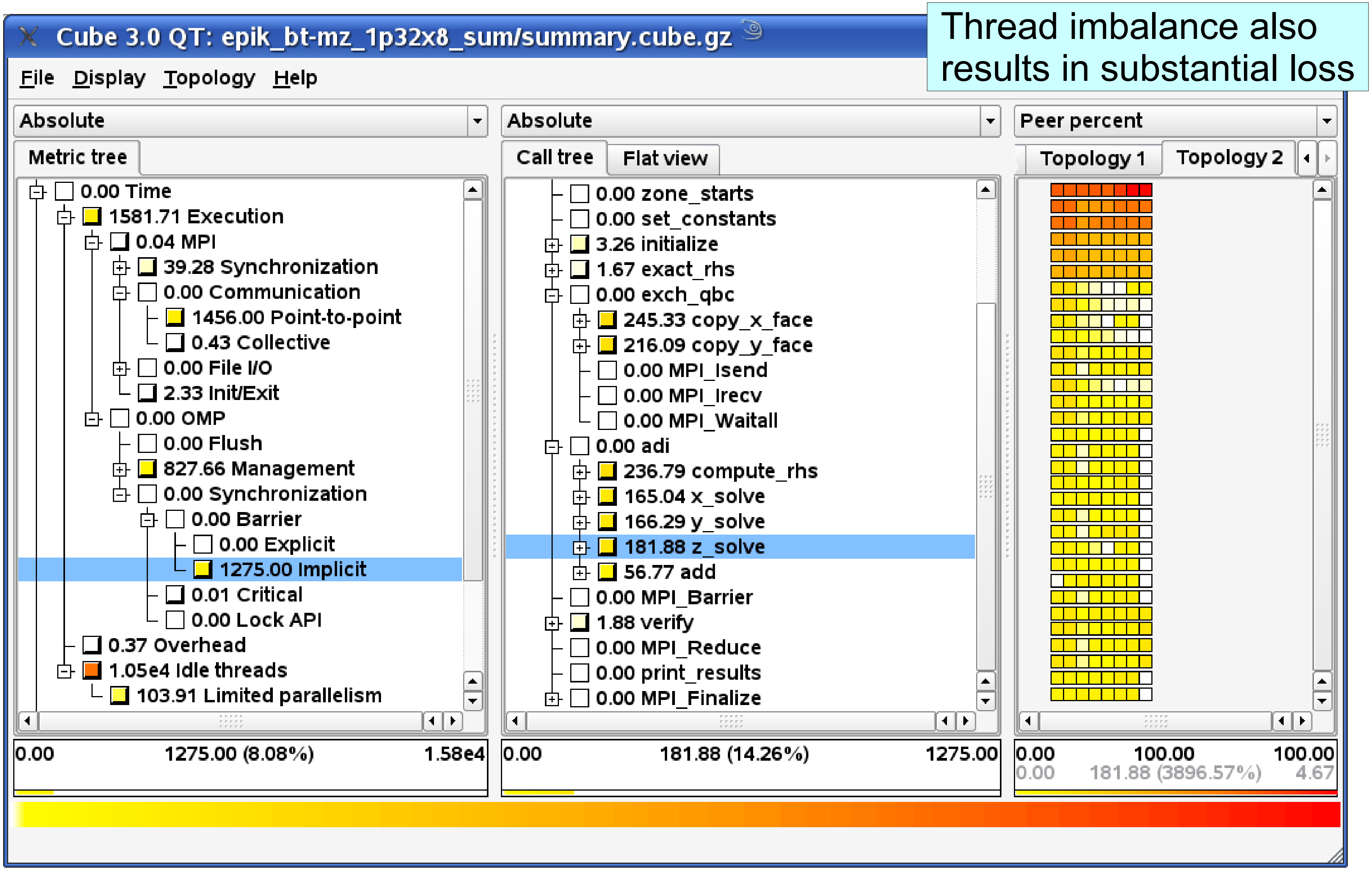
- □ 0.00 Time
  - 🟨 1581.71 Execution
    - □ 0.04 MPI
      - 🟨 39.28 Synchronization
      - □ 0.00 Communication
        - 🟨 1456.00 Point-to-point
        - □ 0.43 Collective
      - □ 0.00 File I/O
      - □ 2.33 Init/Exit
    - □ 0.00 OMP
      - □ 0.00 Flush
      - 🟨 827.66 Management
      - □ 0.00 Synchronization
        - □ 0.00 Barrier
          - □ 0.00 Explicit
          - 🟨 1275.00 Implicit
        - □ 0.01 Critical
        - □ 0.00 Lock API
  - □ 0.37 Overhead
  - 🟧 10499.94 Idle threads
    - 🟨 103.91 Limited parallelism

Call tree:

- □ 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
  - □ 0.34 MPI_Isend
  - □ 0.48 MPI_Irecv
  - 🟥 1455.19 MPI_Waitall
- □ 0.00 adi
  - □ 0.00 compute_rhs
  - □ 0.00 x_solve
  - □ 0.00 y_solve
  - □ 0.00 z_solve
  - □ 0.00 add
- □ 0.00 MPI_Barrier
- □ 0.00 verify
- □ 0.00 MPI_Reduce
- □ 0.00 print_results
- □ 0.00 MPI_Finalize

```
0.00        1456.00 (9.22%)        15786.70
0.00        1455.19 (99.94%)        1456.00
0.00             100.00              100.00
0.00        1455.19 (2716.24%)        53.57
```

32x8 summary analysis: Implicit barrier time



Cube 3.0 QT: epik_bt-mz_1p32x8_sum/summary.cube.gz

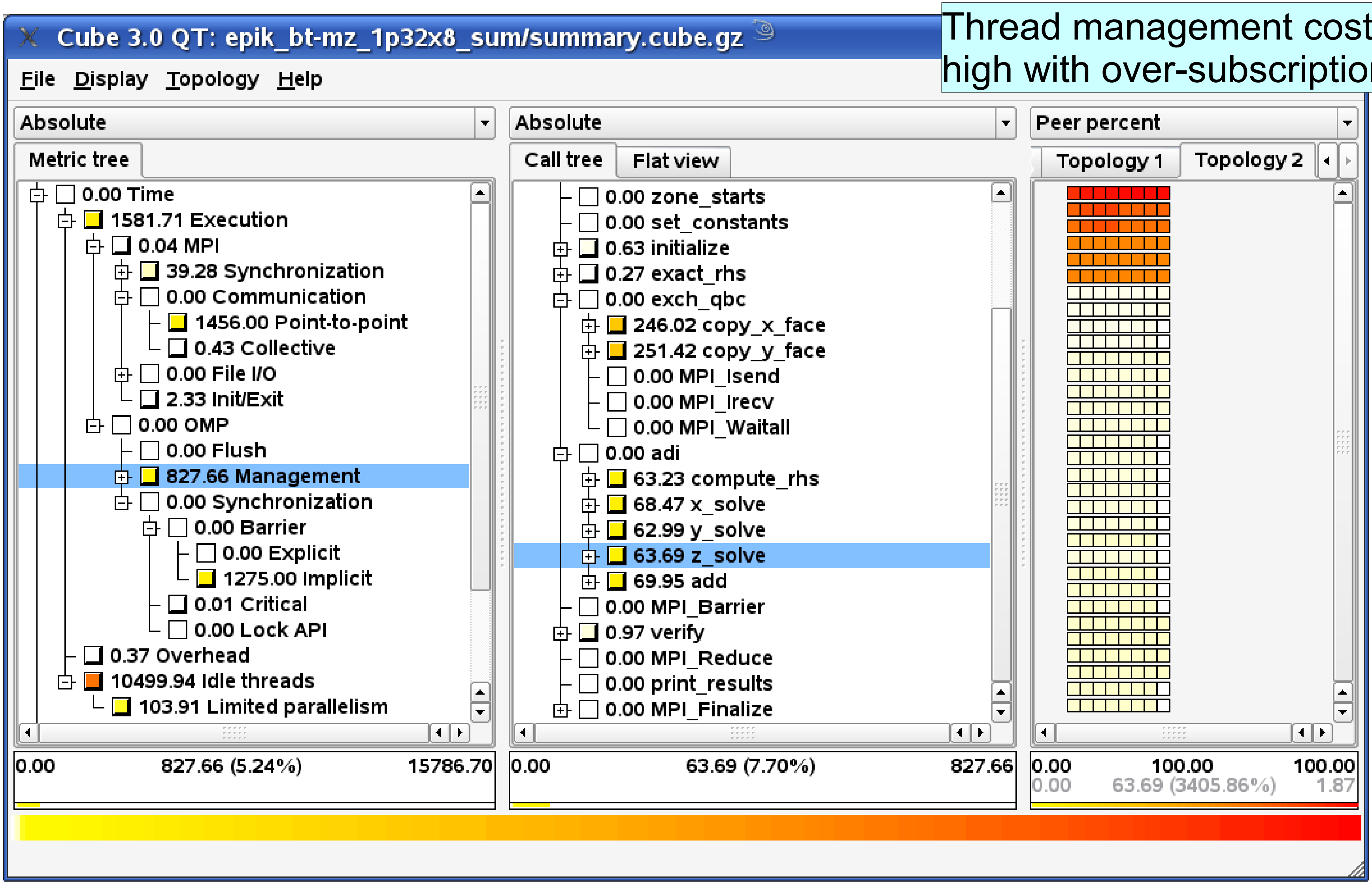
```
□ 0.00 Time
  ■ 1581.71 Execution
    □ 0.04 MPI
      ■ 39.28 Synchronization
      □ 0.00 Communication
        ■ 1456.00 Point-to-point
        □ 0.43 Collective
      □ 0.00 File I/O
      □ 2.33 Init/Exit
    □ 0.00 OMP
      □ 0.00 Flush
      ■ 827.66 Management
      □ 0.00 Synchronization
        □ 0.00 Barrier
          □ 0.00 Explicit
          ■ 1275.00 Implicit
        □ 0.01 Critical
        □ 0.00 Lock API
  □ 0.37 Overhead
  ■ 1.05e4 Idle threads
    ■ 103.91 Limited parallelism
```

```
  □ 0.00 zone_starts
  □ 0.00 set_constants
  ■ 3.26 initialize
  ■ 1.67 exact_rhs
  □ 0.00 exch_qbc
    ■ 245.33 copy_x_face
    ■ 216.09 copy_y_face
    □ 0.00 MPI_Isend
    □ 0.00 MPI_Irecv
    □ 0.00 MPI_Waitall
  □ 0.00 adi
    ■ 236.79 compute_rhs
    ■ 165.04 x_solve
    ■ 166.29 y_solve
    ■ 181.88 z_solve
    ■ 56.77 add
  □ 0.00 MPI_Barrier
  ■ 1.88 verify
  □ 0.00 MPI_Reduce
  □ 0.00 print_results
  □ 0.00 MPI_Finalize
```

```
0.00        1275.00 (8.08%)        1.58e4
0.00        181.88 (14.26%)        1275.00
0.00        100.00                 100.00
0.00        181.88 (3896.57%)      4.67
```

Thread imbalance also results in substantial loss

32x8 summary analysis: Thread management



Cube 3.0 QT: epik_bt-mz_1p32x8_sum/summary.cube.gz

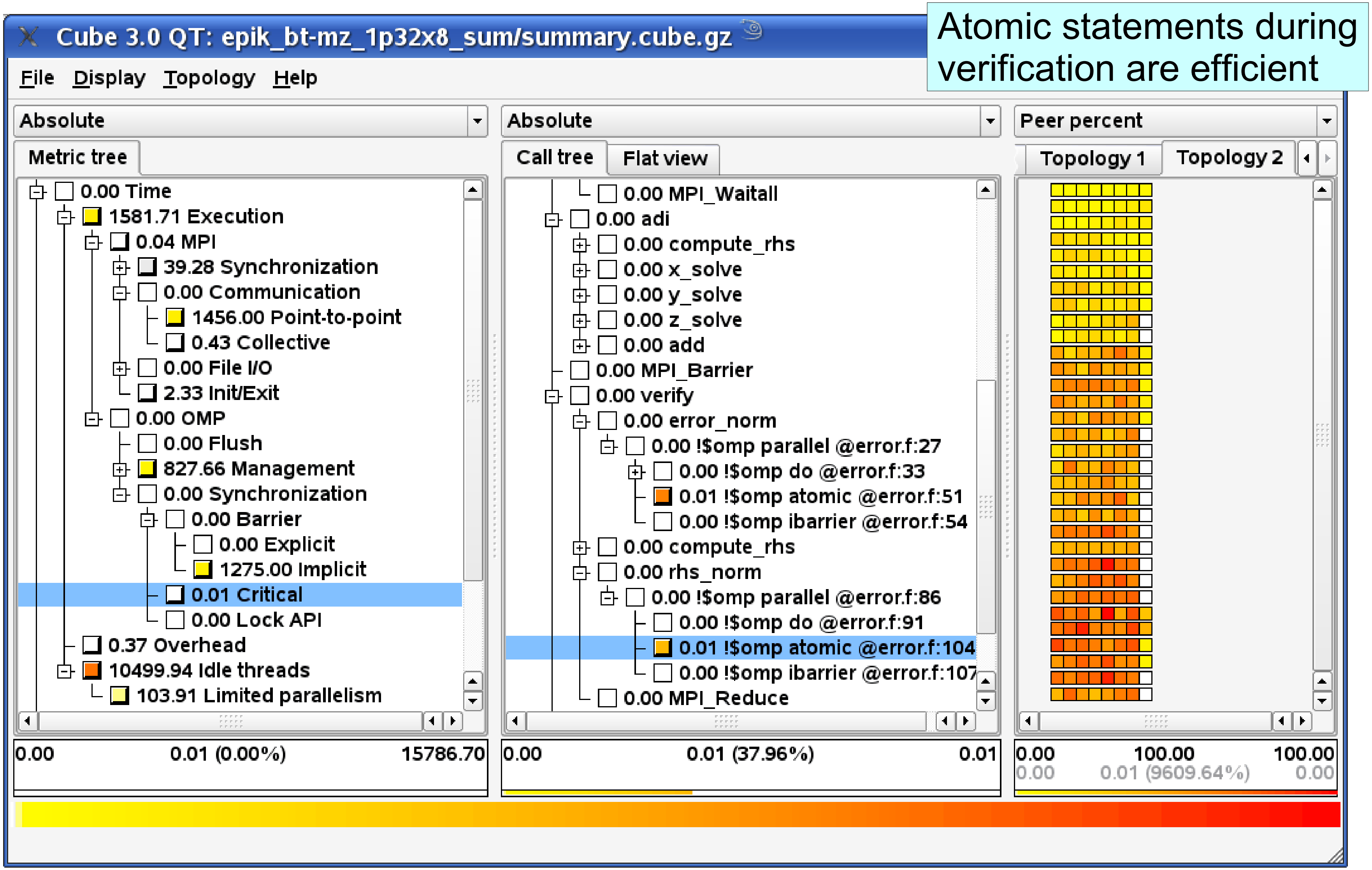
```
□ 0.00 Time
  ■ 1581.71 Execution
    □ 0.04 MPI
      ■ 39.28 Synchronization
      □ 0.00 Communication
        ■ 1456.00 Point-to-point
        □ 0.43 Collective
      □ 0.00 File I/O
      □ 2.33 Init/Exit
    □ 0.00 OMP
      □ 0.00 Flush
      ■ 827.66 Management
      □ 0.00 Synchronization
        □ 0.00 Barrier
          □ 0.00 Explicit
          ■ 1275.00 Implicit
        □ 0.01 Critical
        □ 0.00 Lock API
  □ 0.37 Overhead
  ■ 10499.94 Idle threads
    ■ 103.91 Limited parallelism
```

```
□ 0.00 zone_starts
□ 0.00 set_constants
□ 0.63 initialize
□ 0.27 exact_rhs
□ 0.00 exch_qbc
  ■ 246.02 copy_x_face
  ■ 251.42 copy_y_face
  □ 0.00 MPI_Isend
  □ 0.00 MPI_Irecv
  □ 0.00 MPI_Waitall
□ 0.00 adi
  ■ 63.23 compute_rhs
  ■ 68.47 x_solve
  ■ 62.99 y_solve
  ■ 63.69 z_solve
  ■ 69.95 add
□ 0.00 MPI_Barrier
□ 0.97 verify
□ 0.00 MPI_Reduce
□ 0.00 print_results
□ 0.00 MPI_Finalize
```

```
0.00    827.66 (5.24%)    15786.70
0.00    63.69 (7.70%)     827.66
0.00    100.00            100.00
0.00    63.69 (3405.86%)  1.87
```

Thread management cost high with over-subscription

32x8 summary analysis: Critical section time



Cube 3.0 QT: epik_bt-mz_1p32x8_sum/summary.cube.gz

```
□ 0.00 Time
  □ 1581.71 Execution
    □ 0.04 MPI
      □ 39.28 Synchronization
      □ 0.00 Communication
        □ 1456.00 Point-to-point
        □ 0.43 Collective
      □ 0.00 File I/O
      □ 2.33 Init/Exit
    □ 0.00 OMP
      □ 0.00 Flush
      □ 827.66 Management
      □ 0.00 Synchronization
        □ 0.00 Barrier
          □ 0.00 Explicit
          □ 1275.00 Implicit
        □ 0.01 Critical
        □ 0.00 Lock API
  □ 0.37 Overhead
  □ 10499.94 Idle threads
    □ 103.91 Limited parallelism
```

```
    □ 0.00 MPI_Waitall
  □ 0.00 adi
    □ 0.00 compute_rhs
    □ 0.00 x_solve
    □ 0.00 y_solve
    □ 0.00 z_solve
    □ 0.00 add
  □ 0.00 MPI_Barrier
  □ 0.00 verify
    □ 0.00 error_norm
      □ 0.00 !$omp parallel @error.f:27
        □ 0.00 !$omp do @error.f:33
        □ 0.01 !$omp atomic @error.f:51
        □ 0.00 !$omp ibarrier @error.f:54
    □ 0.00 compute_rhs
    □ 0.00 rhs_norm
      □ 0.00 !$omp parallel @error.f:86
        □ 0.00 !$omp do @error.f:91
        □ 0.01 !$omp atomic @error.f:104
        □ 0.00 !$omp ibarrier @error.f:107
    □ 0.00 MPI_Reduce
```

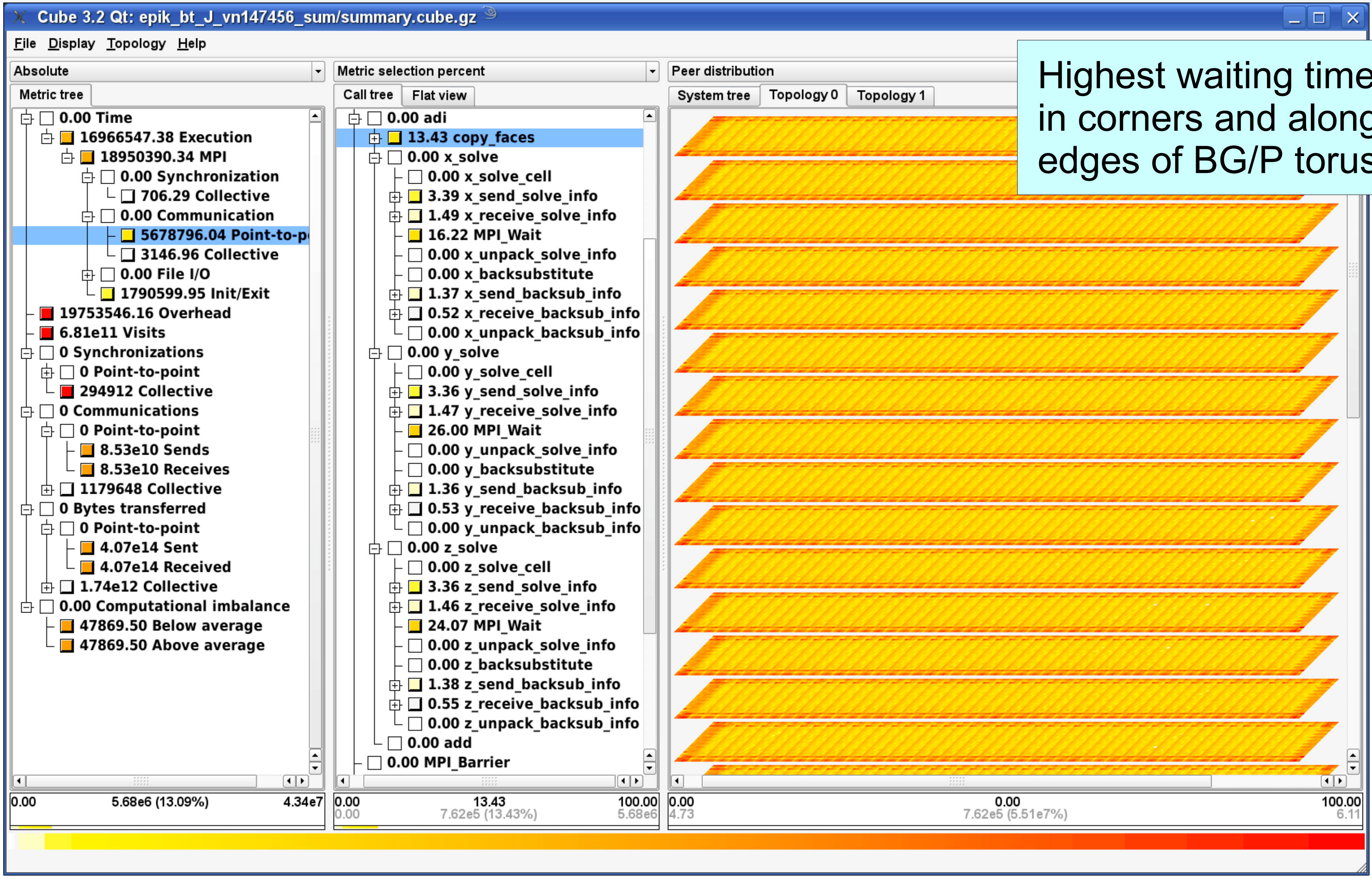
```
0.00        0.01 (0.00%)        15786.70
0.00        0.01 (37.96%)       0.01
0.00        100.00              100.00
0.00        0.01 (9609.64%)     0.00
```

Atomic statements during verification are efficient

NPB-MPI-BT on BlueGene/P case study

- 3D solution of unsteady, compressible Navier-Stokes eqs
  - NASA NAS parallel benchmark suite Block-Tridiagonal solver
  - series of ADI solve steps in X, Y & Z dimensions
  - ~9,500 lines (20 source modules), mostly Fortran77
- Run on IBM BlueGene/P in VN mode with 144k processes
  - Good scaling when problem size matched to architecture
    - 1536x1536x1536 gridpoints mapped onto 384x384 processes
  - Measurement collection took 53 minutes
  - 38% dilation for summarization measurement compared to uninstrumented execution (using 10 function filter)
  - MPI trace size would be 18.6TB
  - 25% of time in ADI is point-to-point communication time
    - 13% copy_faces, 23% x_solve, 33% y_solve, 31% z_solve
  - 128s for a single MPI_Comm_split during setup

NPB-MPI-BT on jugene@144k summary analysis



Cube 3.2 Qt: epik_bt_J_vn147456_sum/summary.cube.gz

Absolute | Metric selection percent | Peer distribution

```
0.00 Time
  16966547.38 Execution
    18950390.34 MPI
      0.00 Synchronization
        706.29 Collective
      0.00 Communication
        5678796.04 Point-to-p
        3146.96 Collective
      0.00 File I/O
      1790599.95 Init/Exit
19753546.16 Overhead
6.81e11 Visits
0 Synchronizations
  0 Point-to-point
  294912 Collective
0 Communications
  0 Point-to-point
    8.53e10 Sends
    8.53e10 Receives
  1179648 Collective
0 Bytes transferred
  0 Point-to-point
    4.07e14 Sent
    4.07e14 Received
  1.74e12 Collective
0.00 Computational imbalance
  47869.50 Below average
  47869.50 Above average
```

```
0.00 adi
  13.43 copy_faces
  0.00 x_solve
    0.00 x_solve_cell
    3.39 x_send_solve_info
    1.49 x_receive_solve_info
    16.22 MPI_Wait
    0.00 x_unpack_solve_info
    0.00 x_backsubstitute
    1.37 x_send_backsub_info
    0.52 x_receive_backsub_info
    0.00 x_unpack_backsub_info
  0.00 y_solve
    0.00 y_solve_cell
    3.36 y_send_solve_info
    1.47 y_receive_solve_info
    26.00 MPI_Wait
    0.00 y_unpack_solve_info
    0.00 y_backsubstitute
    1.36 y_send_backsub_info
    0.53 y_receive_backsub_info
    0.00 y_unpack_backsub_info
  0.00 z_solve
    0.00 z_solve_cell
    3.36 z_send_solve_info
    1.46 z_receive_solve_info
    24.07 MPI_Wait
    0.00 z_unpack_solve_info
    0.00 z_backsubstitute
    1.38 z_send_backsub_info
    0.55 z_receive_backsub_info
    0.00 z_unpack_backsub_info
  0.00 add
0.00 MPI_Barrier
```

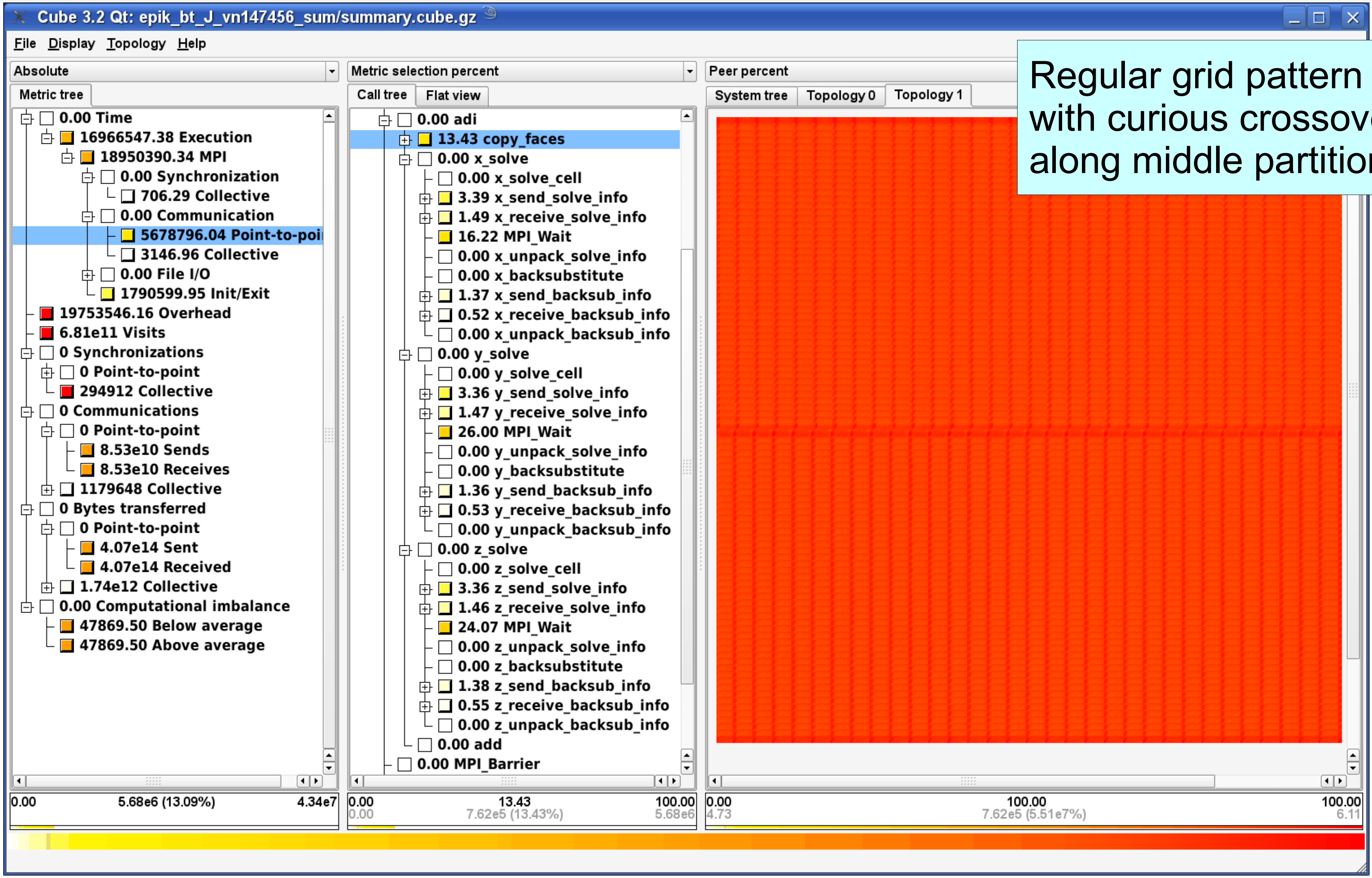
```
0.00        5.68e6 (13.09%)        4.34e7
```

```
0.00        13.43                 100.00
0.00        7.62e5 (13.43%)       5.68e6
```

```
0.00        0.00                  100.00
4.73        7.62e5 (5.51e7%)      6.11
```

Highest waiting times in corners and along edges of BG/P torus

NPB-MPI-BT on jugene@144k summary analysis



Cube 3.2 Qt: epik_bt_J_vn147456_sum/summary.cube.gz

Metric tree

- 0.00 Time
  - 16966547.38 Execution
    - 18950390.34 MPI
      - 0.00 Synchronization
        - 706.29 Collective
      - 0.00 Communication
        - 5678796.04 Point-to-poi
        - 3146.96 Collective
      - 0.00 File I/O
      - 1790599.95 Init/Exit
  - 19753546.16 Overhead
- 6.81e11 Visits
- 0 Synchronizations
  - 0 Point-to-point
  - 294912 Collective
- 0 Communications
  - 0 Point-to-point
    - 8.53e10 Sends
    - 8.53e10 Receives
  - 1179648 Collective
- 0 Bytes transferred
  - 0 Point-to-point
    - 4.07e14 Sent
    - 4.07e14 Received
  - 1.74e12 Collective
- 0.00 Computational imbalance
  - 47869.50 Below average
  - 47869.50 Above average

Call tree

- 0.00 adi
  - 13.43 copy_faces
  - 0.00 x_solve
    - 0.00 x_solve_cell
    - 3.39 x_send_solve_info
    - 1.49 x_receive_solve_info
    - 16.22 MPI_Wait
    - 0.00 x_unpack_solve_info
    - 0.00 x_backsubstitute
    - 1.37 x_send_backsub_info
    - 0.52 x_receive_backsub_info
    - 0.00 x_unpack_backsub_info
  - 0.00 y_solve
    - 0.00 y_solve_cell
    - 3.36 y_send_solve_info
    - 1.47 y_receive_solve_info
    - 26.00 MPI_Wait
    - 0.00 y_unpack_solve_info
    - 0.00 y_backsubstitute
    - 1.36 y_send_backsub_info
    - 0.53 y_receive_backsub_info
    - 0.00 y_unpack_backsub_info
  - 0.00 z_solve
    - 0.00 z_solve_cell
    - 3.36 z_send_solve_info
    - 1.46 z_receive_solve_info
    - 24.07 MPI_Wait
    - 0.00 z_unpack_solve_info
    - 0.00 z_backsubstitute
    - 1.38 z_send_backsub_info
    - 0.55 z_receive_backsub_info
    - 0.00 z_unpack_backsub_info
  - 0.00 add
- 0.00 MPI_Barrier

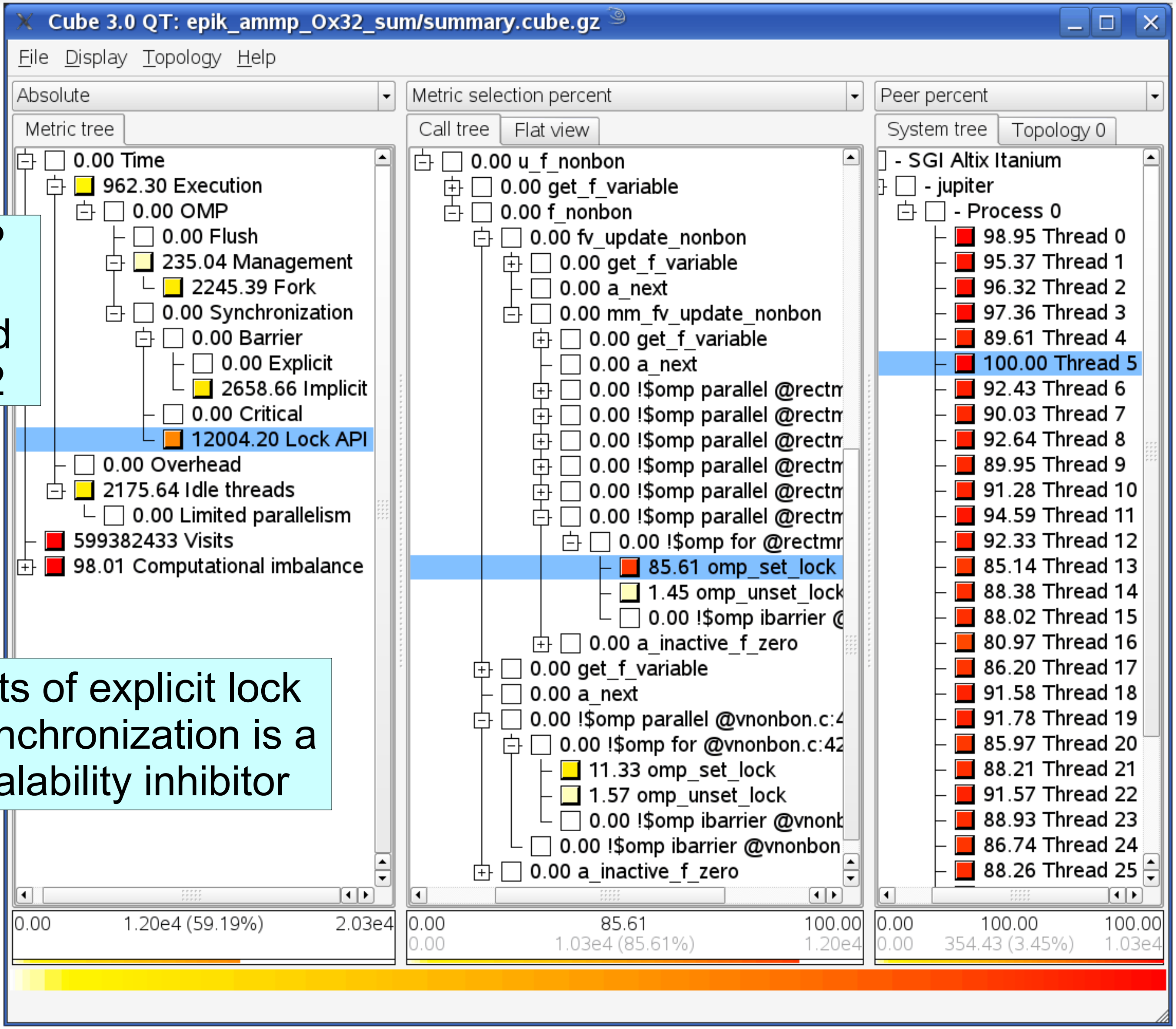
Panel	Min	Value	Max
Metric tree	0.00	5.68e6 (13.09%)	4.34e7
Call tree	0.00 0.00	13.43 7.62e5 (13.43%)	100.00 5.68e6
Topology	0.00 4.73	100.00 7.62e5 (5.51e7%)	100.00 6.11

Regular grid pattern with curious crossover along middle partition

AMMP on Altix case study

- Molecular mechanics simulation
  - original version developed by Robert W. Harrison
- SPEC OMP benchmark parallel version
  - ~14,000 lines (in 28 source modules): 100% C
- Run with 32 threads on SGI Altix 4700 at TUD-ZIH
  - Built with Intel compilers
  - 333 simulation timesteps for 9,582 atoms
- Scalasca summary measurement
  - Minimal measurement dilation
  - 60% of total time lost in synchronization with lock API
  - 12% thread management overhead

ammp on jupiter@32 OpenMP lock analysis



Cube 3.0 QT: epik_ammp_Ox32_sum/summary.cube.gz
File Display Topology Help
Metric tree
Call tree
Flat view
System tree
Topology 0

Absolute | Metric selection percent | Peer percent

```
Metric tree
□ 0.00 Time
  🟨 962.30 Execution
    □ 0.00 OMP
      □ 0.00 Flush
      🟨 235.04 Management
        🟨 2245.39 Fork
      □ 0.00 Synchronization
        □ 0.00 Barrier
          □ 0.00 Explicit
          🟨 2658.66 Implicit
        □ 0.00 Critical
        🟧 12004.20 Lock API
  □ 0.00 Overhead
  🟨 2175.64 Idle threads
    □ 0.00 Limited parallelism
🟥 599382433 Visits
🟥 98.01 Computational imbalance
```

```
Call tree
□ 0.00 u_f_nonbon
  □ 0.00 get_f_variable
  □ 0.00 f_nonbon
    □ 0.00 fv_update_nonbon
      □ 0.00 get_f_variable
      □ 0.00 a_next
      □ 0.00 mm_fv_update_nonbon
        □ 0.00 get_f_variable
        □ 0.00 a_next
        □ 0.00 !$omp parallel @rectm
        □ 0.00 !$omp parallel @rectm
        □ 0.00 !$omp parallel @rectm
        □ 0.00 !$omp parallel @rectm
        □ 0.00 !$omp parallel @rectm
        □ 0.00 !$omp parallel @rectm
          □ 0.00 !$omp for @rectmr
            🟥 85.61 omp_set_lock
            🟨 1.45 omp_unset_lock
            □ 0.00 !$omp ibarrier @
        □ 0.00 a_inactive_f_zero
    □ 0.00 get_f_variable
    □ 0.00 a_next
    □ 0.00 !$omp parallel @vnonbon.c:4
      □ 0.00 !$omp for @vnonbon.c:42
        🟨 11.33 omp_set_lock
        🟨 1.57 omp_unset_lock
        □ 0.00 !$omp ibarrier @vnonb
      □ 0.00 !$omp ibarrier @vnonbon
    □ 0.00 a_inactive_f_zero
```

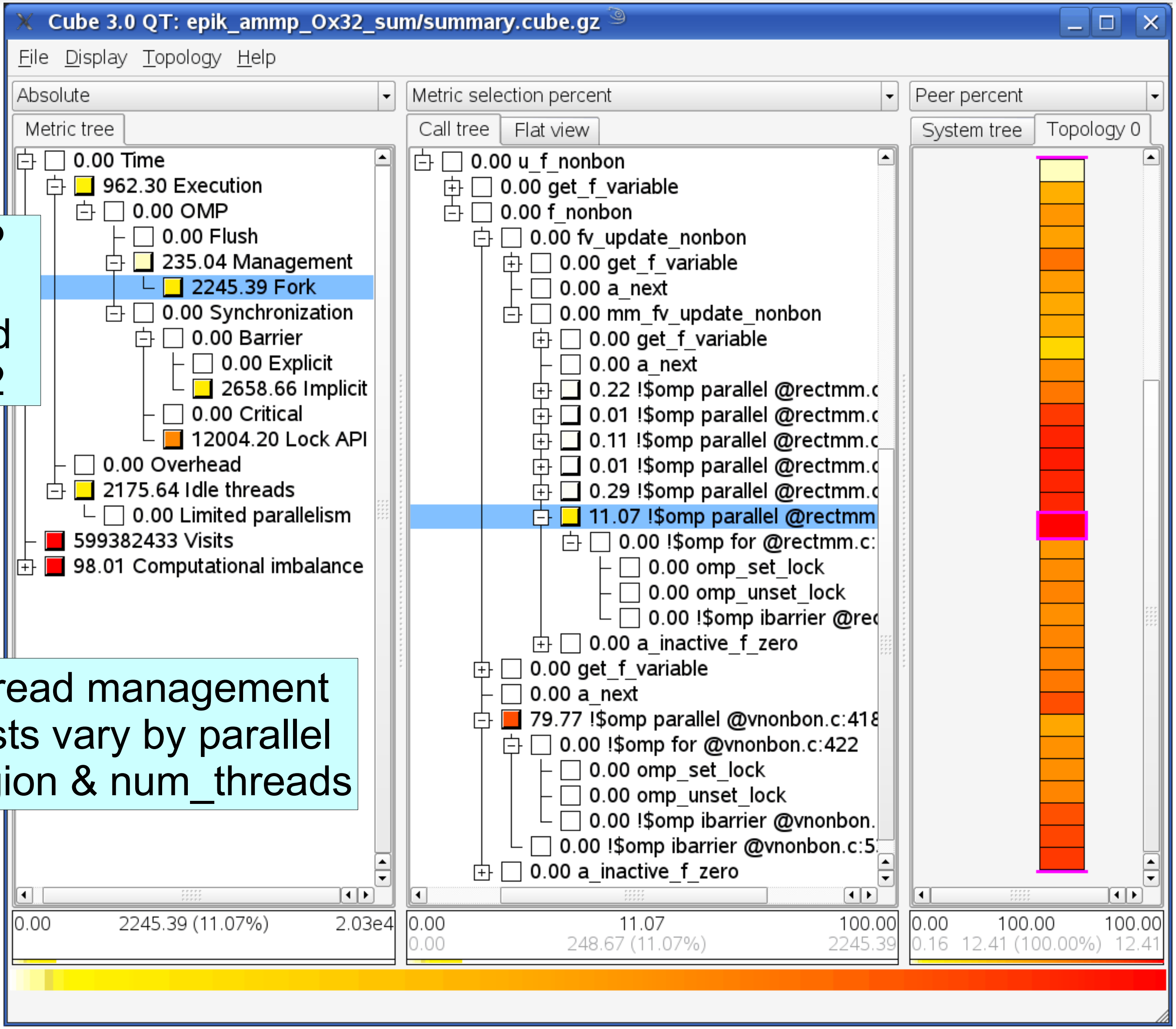
```
System tree
- SGI Altix Itanium
  □ - jupiter
    □ - Process 0
      🟥 98.95 Thread 0
      🟥 95.37 Thread 1
      🟥 96.32 Thread 2
      🟥 97.36 Thread 3
      🟥 89.61 Thread 4
      🟥 100.00 Thread 5
      🟥 92.43 Thread 6
      🟥 90.03 Thread 7
      🟥 92.64 Thread 8
      🟥 89.95 Thread 9
      🟥 91.28 Thread 10
      🟥 94.59 Thread 11
      🟥 92.33 Thread 12
      🟥 85.14 Thread 13
      🟥 88.38 Thread 14
      🟥 88.02 Thread 15
      🟥 80.97 Thread 16
      🟥 86.20 Thread 17
      🟥 91.58 Thread 18
      🟥 91.78 Thread 19
      🟥 85.97 Thread 20
      🟥 88.21 Thread 21
      🟥 91.57 Thread 22
      🟥 88.93 Thread 23
      🟥 86.74 Thread 24
      🟥 88.26 Thread 25
```

```
0.00    1.20e4 (59.19%)    2.03e4
0.00    85.61              100.00
0.00    1.03e4 (85.61%)    1.20e4
0.00    100.00             100.00
0.00    354.43 (3.45%)     1.03e4
```

OpenMP metrics reworked with v1.2

Lots of explicit lock synchronization is a scalability inhibitor

ammp on jupiter@32 OpenMP fork analysis



Absolute | Metric selection percent | Peer percent

```
0.00 Time
  962.30 Execution
    0.00 OMP
      0.00 Flush
      235.04 Management
        2245.39 Fork
      0.00 Synchronization
        0.00 Barrier
          0.00 Explicit
          2658.66 Implicit
        0.00 Critical
        12004.20 Lock API
  0.00 Overhead
  2175.64 Idle threads
    0.00 Limited parallelism
599382433 Visits
98.01 Computational imbalance
```

```
0.00 u_f_nonbon
  0.00 get_f_variable
  0.00 f_nonbon
    0.00 fv_update_nonbon
      0.00 get_f_variable
      0.00 a_next
      0.00 mm_fv_update_nonbon
        0.00 get_f_variable
        0.00 a_next
        0.22 !$omp parallel @rectmm.c
        0.01 !$omp parallel @rectmm.c
        0.11 !$omp parallel @rectmm.c
        0.01 !$omp parallel @rectmm.c
        0.29 !$omp parallel @rectmm.c
        11.07 !$omp parallel @rectmm
          0.00 !$omp for @rectmm.c:
            0.00 omp_set_lock
            0.00 omp_unset_lock
            0.00 !$omp ibarrier @rec
        0.00 a_inactive_f_zero
    0.00 get_f_variable
    0.00 a_next
    79.77 !$omp parallel @vnonbon.c:418
      0.00 !$omp for @vnonbon.c:422
        0.00 omp_set_lock
        0.00 omp_unset_lock
        0.00 !$omp ibarrier @vnonbon.
      0.00 !$omp ibarrier @vnonbon.c:5
    0.00 a_inactive_f_zero
```

```
0.00        2245.39 (11.07%)        2.03e4
```

```
0.00             11.07              100.00
0.00        248.67 (11.07%)        2245.39
```

```
0.00       100.00            100.00
0.16   12.41 (100.00%)        12.41
```

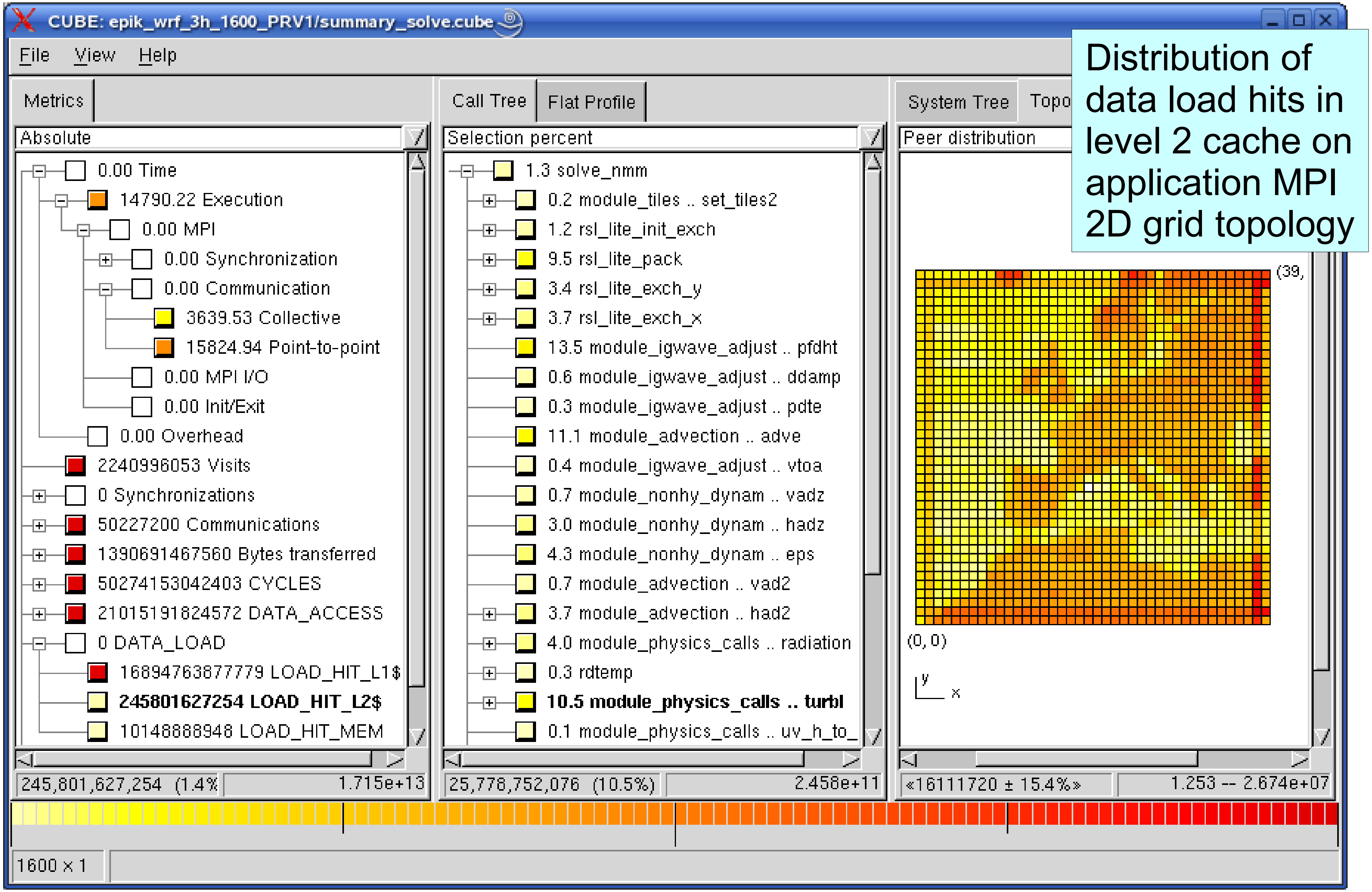
OpenMP metrics reworked with v1.2

Thread management costs vary by parallel region & num_threads

WRF/MareNostrum case study

- Numerical weather prediction
  - public domain code developed by US NOAA
  - flexible, state-of-the-art atmospheric simulation
  - Non-hydrostatic Mesoscale Model (NMM)
- MPI parallel version 2.1.2 (Jan-2006)
  - >315,000 lines (in 480 source modules): 75% Fortran, 25% C
- Eur-12km dataset configuration
  - 3-hour forecast (360 timesteps) with checkpointing disabled
- Run with 1600 processes on MareNostrum
  - IBM BladeCenter cluster at BSC
- Scalasca summary and trace measurements
  - 15% measurement dilation with 8 hardware counters
  - 23GB trace analysis in 5 mins

WRF on MareNostrum@1600 with HWC metrics



CUBE: epik_wrf_3h_1600_PRV1/summary_solve.cube

File View Help

Metrics | Call Tree | Flat Profile | System Tree | Topo

Absolute

```
0.00 Time
  14790.22 Execution
    0.00 MPI
      0.00 Synchronization
      0.00 Communication
        3639.53 Collective
        15824.94 Point-to-point
      0.00 MPI I/O
      0.00 Init/Exit
  0.00 Overhead
2240996053 Visits
0 Synchronizations
50227200 Communications
1390691467560 Bytes transferred
50274153042403 CYCLES
21015191824572 DATA_ACCESS
0 DATA_LOAD
  16894763877779 LOAD_HIT_L1$
  245801627254 LOAD_HIT_L2$
  10148888948 LOAD_HIT_MEM
```

245,801,627,254 (1.4% | 1.715e+13

Selection percent

```
1.3 solve_nmm
  0.2 module_tiles .. set_tiles2
  1.2 rsl_lite_init_exch
  9.5 rsl_lite_pack
  3.4 rsl_lite_exch_y
  3.7 rsl_lite_exch_x
  13.5 module_igwave_adjust .. pfdht
  0.6 module_igwave_adjust .. ddamp
  0.3 module_igwave_adjust .. pdte
  11.1 module_advection .. adve
  0.4 module_igwave_adjust .. vtoa
  0.7 module_nonhy_dynam .. vadz
  3.0 module_nonhy_dynam .. hadz
  4.3 module_nonhy_dynam .. eps
  0.7 module_advection .. vad2
  3.7 module_advection .. had2
  4.0 module_physics_calls .. radiation
  0.3 rdtemp
  10.5 module_physics_calls .. turbl
  0.1 module_physics_calls .. uv_h_to_
```

25,778,752,076 (10.5%) | 2.458e+11

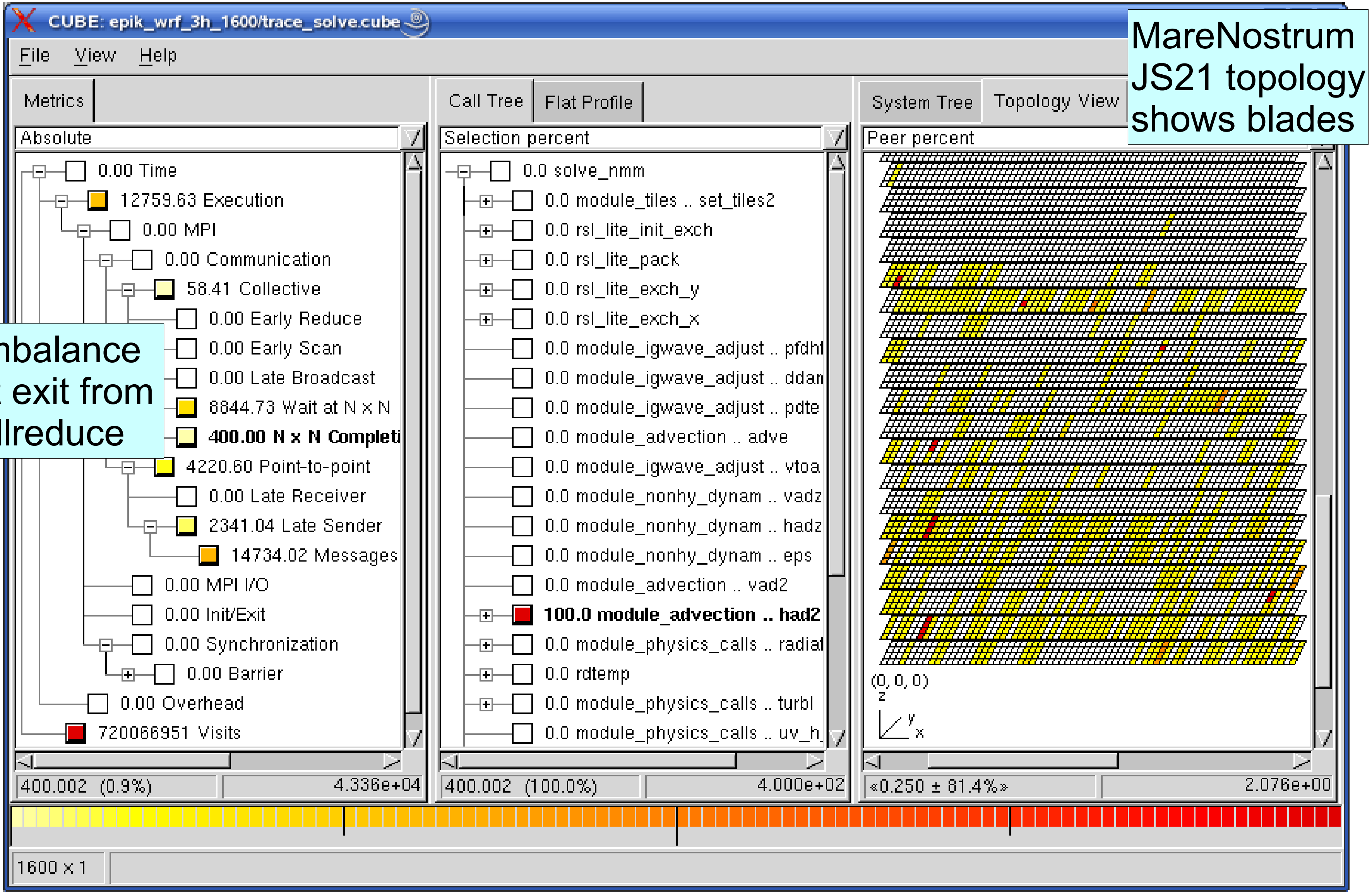
Peer distribution

«16111720 ± 15.4%» | 1.253 -- 2.674e+07

1600 x 1

Distribution of data load hits in level 2 cache on application MPI 2D grid topology

WRF on MareNostrum@1600 trace analysis



- □ 0.00 Time
  - 🟧 12759.63 Execution
    - □ 0.00 MPI
      - □ 0.00 Communication
        - 🟨 58.41 Collective
          - □ 0.00 Early Reduce
          - □ 0.00 Early Scan
          - □ 0.00 Late Broadcast
          - 🟨 8844.73 Wait at N x N
          - 🟨 **400.00 N x N Completi**
        - 🟨 4220.60 Point-to-point
          - □ 0.00 Late Receiver
          - 🟨 2341.04 Late Sender
            - 🟧 14734.02 Messages
      - □ 0.00 MPI I/O
      - □ 0.00 Init/Exit
      - □ 0.00 Synchronization
        - □ 0.00 Barrier
  - □ 0.00 Overhead
- 🟥 720066951 Visits

400.002 (0.9%) 4.336e+04

- □ 0.0 solve_nmm
  - □ 0.0 module_tiles .. set_tiles2
  - □ 0.0 rsl_lite_init_exch
  - □ 0.0 rsl_lite_pack
  - □ 0.0 rsl_lite_exch_y
  - □ 0.0 rsl_lite_exch_x
  - □ 0.0 module_igwave_adjust .. pfdht
  - □ 0.0 module_igwave_adjust .. ddan
  - □ 0.0 module_igwave_adjust .. pdte
  - □ 0.0 module_advection .. adve
  - □ 0.0 module_igwave_adjust .. vtoa
  - □ 0.0 module_nonhy_dynam .. vadz
  - □ 0.0 module_nonhy_dynam .. hadz
  - □ 0.0 module_nonhy_dynam .. eps
  - □ 0.0 module_advection .. vad2
  - 🟥 **100.0 module_advection .. had2**
  - □ 0.0 module_physics_calls .. radiat
  - □ 0.0 rdtemp
  - □ 0.0 module_physics_calls .. turbl
  - □ 0.0 module_physics_calls .. uv_h_

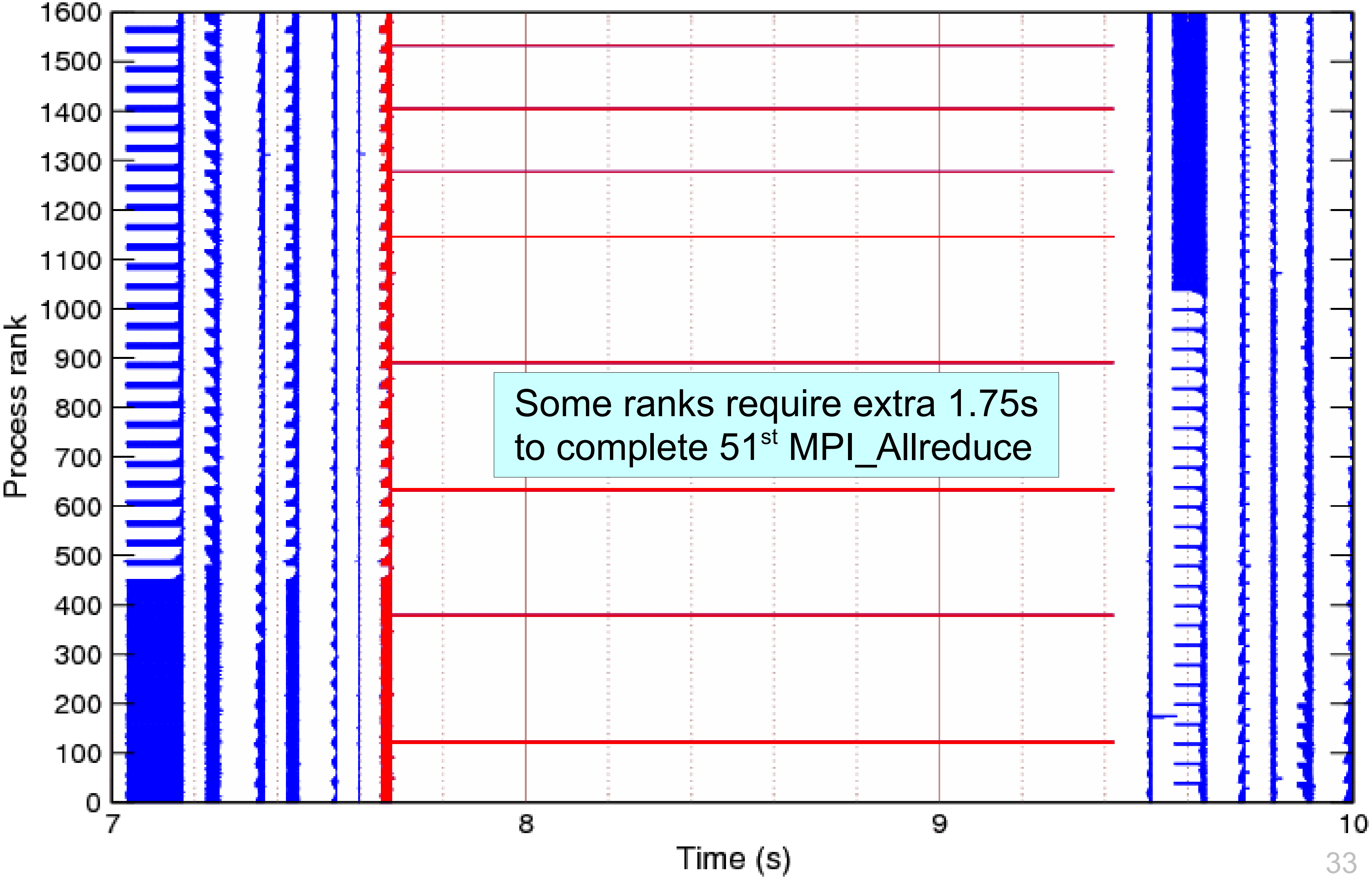
400.002 (100.0%) 4.000e+02

«0.250 ± 81.4%» 2.076e+00

MareNostrum JS21 topology shows blades

Imbalance at exit from Allreduce

WRF on MareNostrum@1600 time-line extract



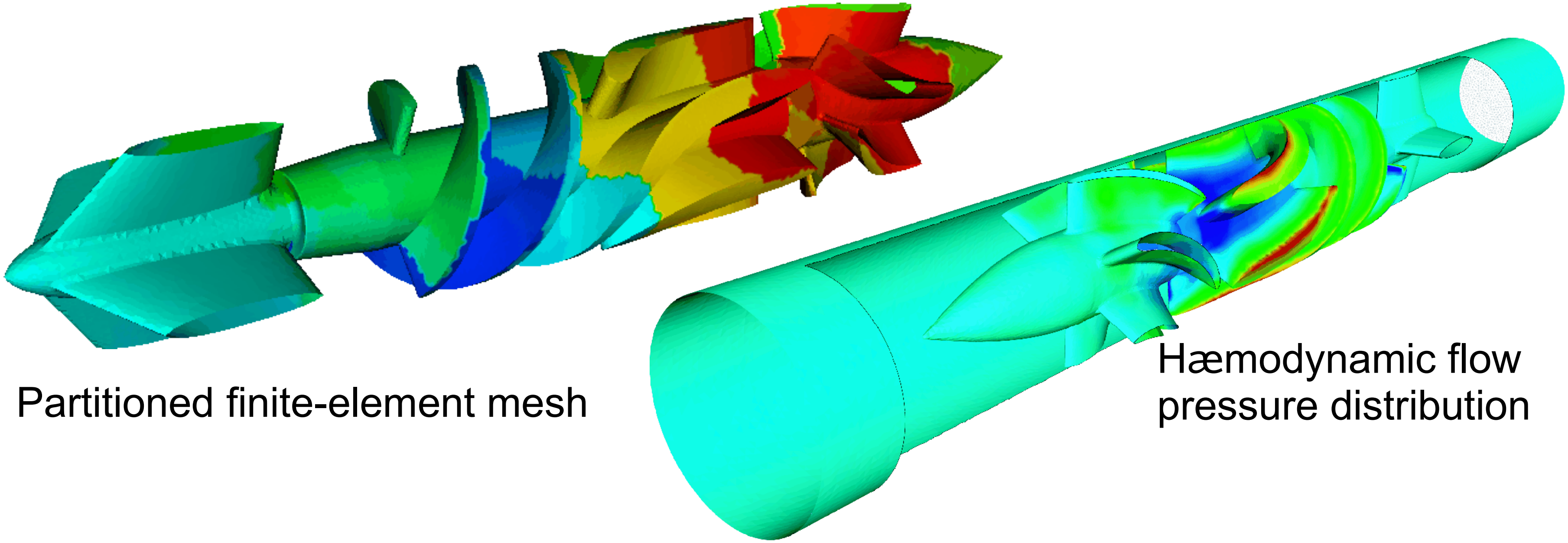
Some ranks require extra 1.75s to complete 51st MPI_Allreduce

WRF/MareNostrum experience

- Limited system I/O requires careful management
  - Selective instrumentation and measurement filtering
- PowerPC hardware counter metrics included in summary
  - Memory/cache data access hierarchy constructed
- Automated trace analysis quantified impact of imbalanced exit from MPI_Allreduce in “NxN completion time” metric
  - Intermittent but serious MPI library/system problem, that restricts application scalability
  - Only a few processes directly impacted, however, communication partners also quickly blocked
- Presentation using logical and physical topologies
  - MPI Cartesian topology provides application insight
  - Hardware topology helps localize system problems

XNS on BlueGene/L case study

- CFD simulation of unsteady flows
  - developed by RWTH CATS group of Marek Behr
  - exploits finite-element techniques, unstructured 3D meshes, iterative solution strategies
- MPI parallel version (Dec-2006)
  - >40,000 lines of Fortran & C
  - DeBakey blood-pump dataset (3,714,611 elements)

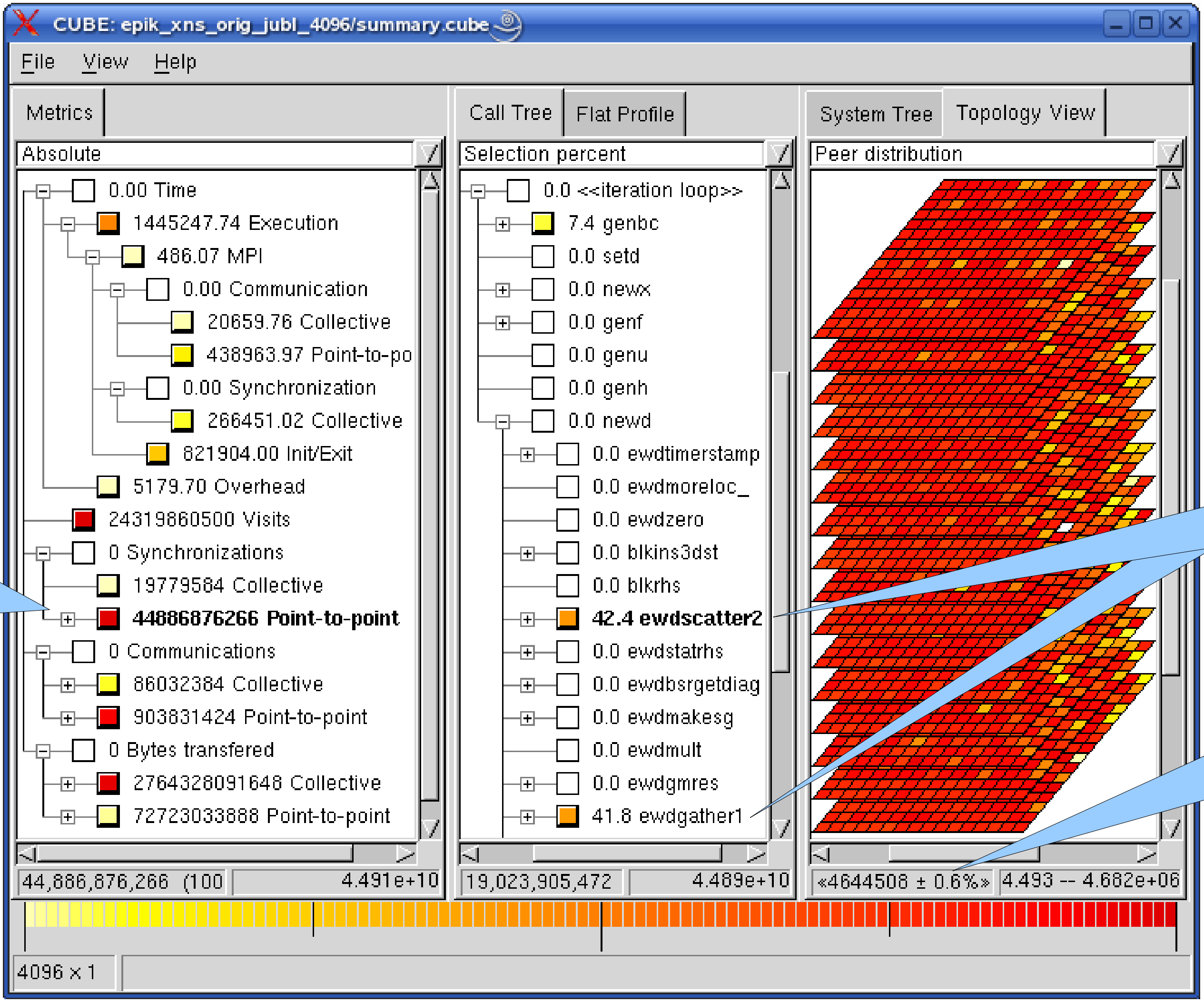


Partitioned finite-element mesh

Hæmodynamic flow pressure distribution

XNS-DeBakey on jubl@4096 summary analysis

Point-to-point msgs w/o data



```
Absolute
0.00 Time
    1445247.74 Execution
        486.07 MPI
            0.00 Communication
                20659.76 Collective
                438963.97 Point-to-po
            0.00 Synchronization
                266451.02 Collective
            821904.00 Init/Exit
    5179.70 Overhead
24319860500 Visits
0 Synchronizations
    19779584 Collective
    44886876266 Point-to-point
0 Communications
    86032384 Collective
    903831424 Point-to-point
0 Bytes transfered
    2764328091648 Collective
    72723033888 Point-to-point
44,886,876,266  (100    4.491e+10
```

```
Selection percent
0.0 <<iteration loop>>
    7.4 genbc
    0.0 setd
    0.0 newx
    0.0 genf
    0.0 genu
    0.0 genh
    0.0 newd
        0.0 ewdtimerstamp
        0.0 ewdmoreloc_
        0.0 ewdzero
        0.0 blkins3dst
        0.0 blkrhs
        42.4 ewdscatter2
        0.0 ewdstatrhs
        0.0 ewdbsrgetdiag
        0.0 ewdmakesg
        0.0 ewdmult
        0.0 ewdgmres
        41.8 ewdgather1
19,023,905,472    4.489e+10
```

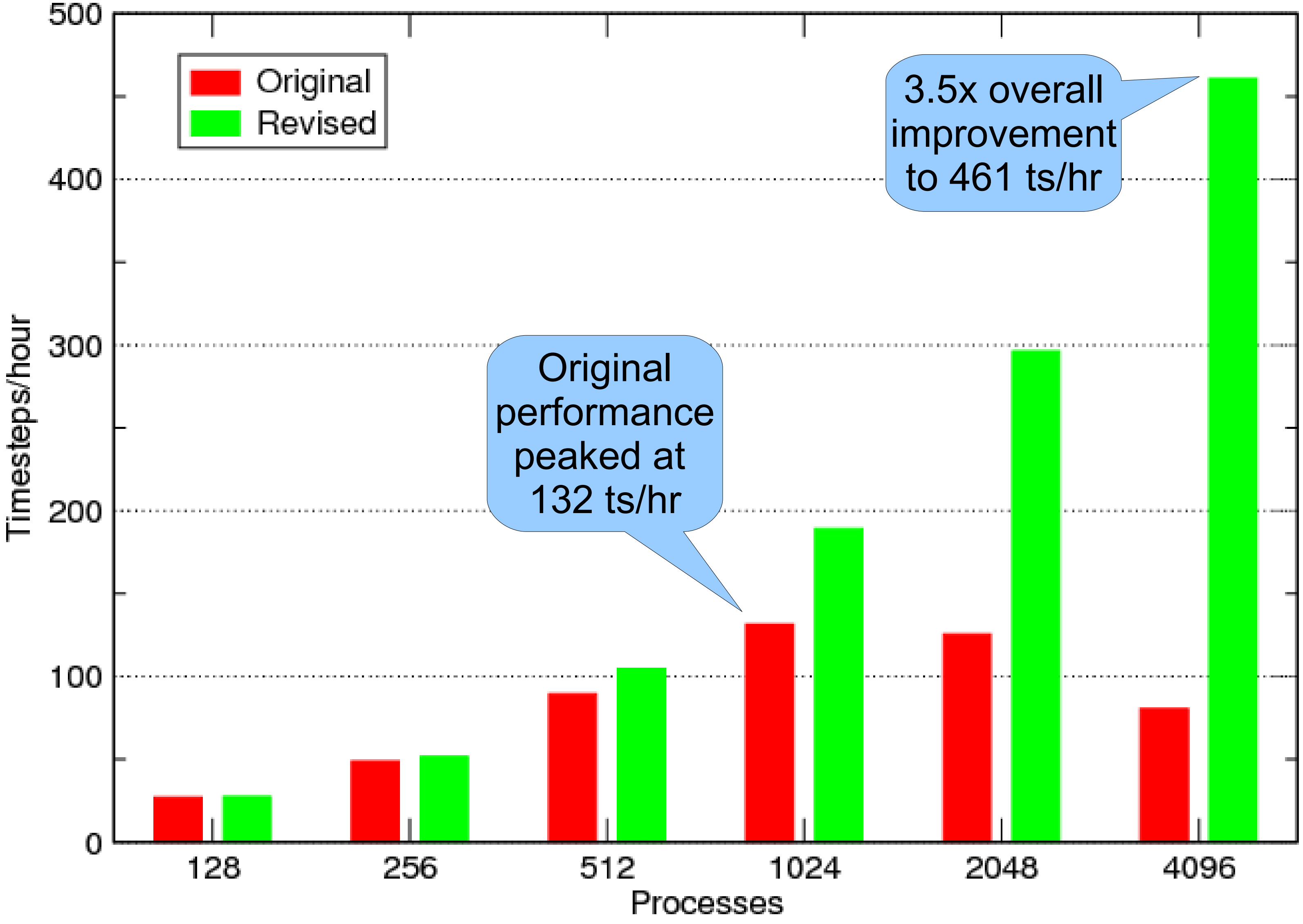
```
Peer distribution
«4644508 ± 0.6%»    4.493 -- 4.682e+06
4096 x 1
```

Masses of P2P synch operations

Primarily in scatter & gather

Processes all equally responsible

XNS-DeBakey scalability on BlueGene/L



Original performance peaked at 132 ts/hr

3.5x overall improvement to 461 ts/hr

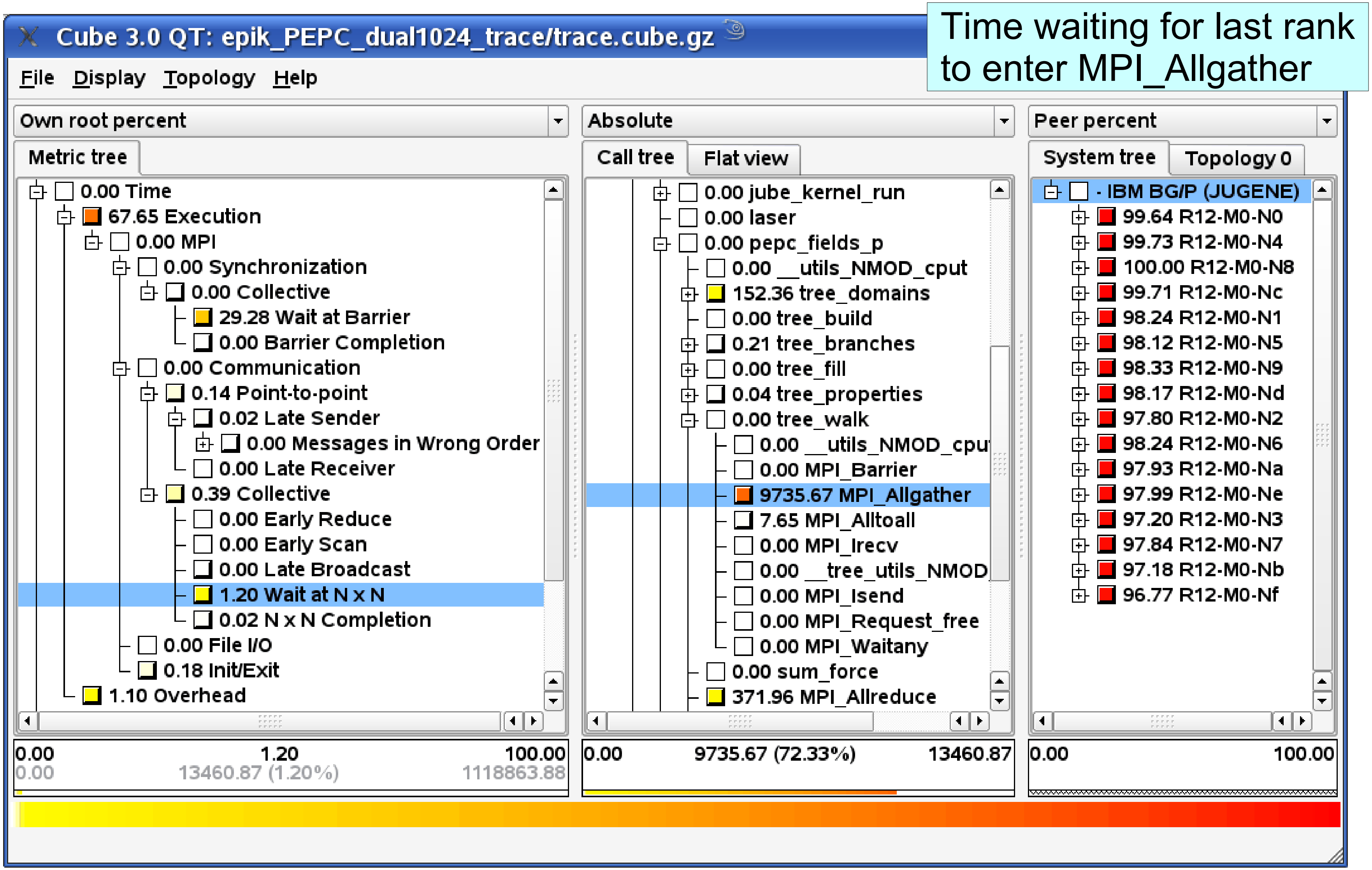
XNS on BlueGene/L experience

- Globally synchronized high-resolution clock facilitates efficient measurement & analysis
- Restricted compute node memory limits trace buffer size and analyzable trace size
- Summarization identified bottleneck due to unintended P2P synchronizations (messages with zero-sized payload)
- 4x solver speedup after replacing MPI_Sendrecv operations with size-dependant separate MPI_Send and MPI_Recv
- Significant communication imbalance remains due to mesh partitioning and mapping onto processors
- MPI_Scan implementation found to contain implicit barrier
  - responsible for 6% of total time with 4096 processes
  - decimated when substituted with simultaneous binomial tree

PEPC-B on BG/P & Cray XT case study

- Coulomb solver used for laser-plasma simulations
  - Developed by Paul Gibbon (JSC)
  - Tree-based particle storage with dynamic load-balancing
- MPI version
  - PRACE benchmark configuration, including file I/O
- Run on BlueGene/P in dual mode with 1024 processes
  - 2 processes per quad-core PowerPC node, 1100 seconds
  - IBM XL compilers, MPI library and torus/tree interconnect
- Run on Cray XT in VN (4p) mode with 1024 processes
  - 4 processes per quad-core Opteron node, 360 seconds
  - PGI compilers and Cray MPI, CNL, SeaStar interconnect

PEPC@1024 on BlueGene/P: Wait at NxN time



Time waiting for last rank to enter MPI_Allgather

Cube 3.0 QT: epik_PEPC_dual1024_trace/trace.cube.gz

File Display Topology Help

Own root percent | Absolute | Peer percent

Metric tree

```
⊟ □ 0.00 Time
  ⊟ 🟧 67.65 Execution
    ⊟ □ 0.00 MPI
      ⊟ □ 0.00 Synchronization
        ⊟ □ 0.00 Collective
            🟨 29.28 Wait at Barrier
            □ 0.00 Barrier Completion
      ⊟ □ 0.00 Communication
        ⊟ □ 0.14 Point-to-point
          ⊟ □ 0.02 Late Sender
            ⊞ □ 0.00 Messages in Wrong Order
            □ 0.00 Late Receiver
        ⊟ □ 0.39 Collective
            □ 0.00 Early Reduce
            □ 0.00 Early Scan
            □ 0.00 Late Broadcast
            🟨 1.20 Wait at N x N
            □ 0.02 N x N Completion
      □ 0.00 File I/O
      □ 0.18 Init/Exit
  🟨 1.10 Overhead
```

0.00 | 1.20 | 100.00
0.00 | 13460.87 (1.20%) | 1118863.88

Call tree | Flat view

```
⊞ □ 0.00 jube_kernel_run
  □ 0.00 laser
⊟ □ 0.00 pepc_fields_p
    □ 0.00 __utils_NMOD_cput
  ⊞ 🟨 152.36 tree_domains
    □ 0.00 tree_build
  ⊞ □ 0.21 tree_branches
  ⊞ □ 0.00 tree_fill
  ⊞ □ 0.04 tree_properties
  ⊟ □ 0.00 tree_walk
      □ 0.00 __utils_NMOD_cpu
      □ 0.00 MPI_Barrier
      🟧 9735.67 MPI_Allgather
      □ 7.65 MPI_Alltoall
      □ 0.00 MPI_Irecv
      □ 0.00 __tree_utils_NMOD
      □ 0.00 MPI_Isend
      □ 0.00 MPI_Request_free
      □ 0.00 MPI_Waitany
    □ 0.00 sum_force
    🟨 371.96 MPI_Allreduce
```

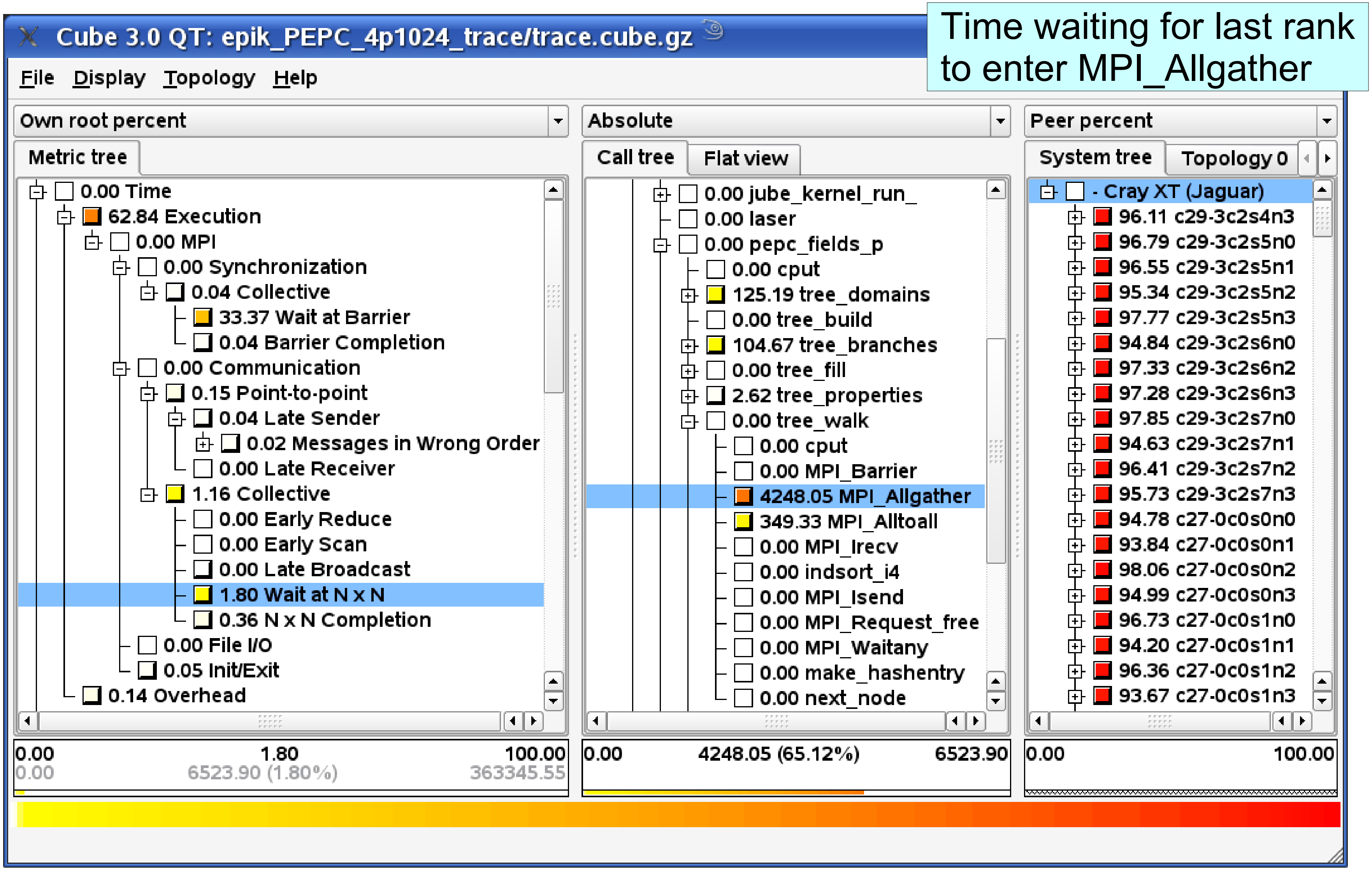
0.00 | 9735.67 (72.33%) | 13460.87

System tree | Topology 0

```
⊟ □ - IBM BG/P (JUGENE)
  ⊞ 🟥 99.64 R12-M0-N0
  ⊞ 🟥 99.73 R12-M0-N4
  ⊞ 🟥 100.00 R12-M0-N8
  ⊞ 🟥 99.71 R12-M0-Nc
  ⊞ 🟥 98.24 R12-M0-N1
  ⊞ 🟥 98.12 R12-M0-N5
  ⊞ 🟥 98.33 R12-M0-N9
  ⊞ 🟥 98.17 R12-M0-Nd
  ⊞ 🟥 97.80 R12-M0-N2
  ⊞ 🟥 98.24 R12-M0-N6
  ⊞ 🟥 97.93 R12-M0-Na
  ⊞ 🟥 97.99 R12-M0-Ne
  ⊞ 🟥 97.20 R12-M0-N3
  ⊞ 🟥 97.84 R12-M0-N7
  ⊞ 🟥 97.18 R12-M0-Nb
  ⊞ 🟥 96.77 R12-M0-Nf
```

0.00 | 100.00

PEPC@1024 on Cray XT4: Wait at NxN time



Cube 3.0 QT: epik_PEPC_4p1024_trace/trace.cube.gz

File Display Topology Help

Own root percent | Absolute | Peer percent

Metric tree

- 0.00 Time
  - 62.84 Execution
    - 0.00 MPI
      - 0.00 Synchronization
        - 0.04 Collective
          - 33.37 Wait at Barrier
          - 0.04 Barrier Completion
      - 0.00 Communication
        - 0.15 Point-to-point
          - 0.04 Late Sender
            - 0.02 Messages in Wrong Order
          - 0.00 Late Receiver
        - 1.16 Collective
          - 0.00 Early Reduce
          - 0.00 Early Scan
          - 0.00 Late Broadcast
          - 1.80 Wait at N x N
          - 0.36 N x N Completion
      - 0.00 File I/O
      - 0.05 Init/Exit
  - 0.14 Overhead

0.00 1.80 100.00
0.00 6523.90 (1.80%) 363345.55

Call tree | Flat view

- 0.00 jube_kernel_run_
- 0.00 laser
- 0.00 pepc_fields_p
  - 0.00 cput
  - 125.19 tree_domains
  - 0.00 tree_build
  - 104.67 tree_branches
  - 0.00 tree_fill
  - 2.62 tree_properties
  - 0.00 tree_walk
    - 0.00 cput
    - 0.00 MPI_Barrier
    - 4248.05 MPI_Allgather
    - 349.33 MPI_Alltoall
    - 0.00 MPI_Irecv
    - 0.00 indsort_i4
    - 0.00 MPI_Isend
    - 0.00 MPI_Request_free
    - 0.00 MPI_Waitany
    - 0.00 make_hashentry
    - 0.00 next_node

0.00 4248.05 (65.12%) 6523.90

System tree | Topology 0

- - Cray XT (Jaguar)
  - 96.11 c29-3c2s4n3
  - 96.79 c29-3c2s5n0
  - 96.55 c29-3c2s5n1
  - 95.34 c29-3c2s5n2
  - 97.77 c29-3c2s5n3
  - 94.84 c29-3c2s6n0
  - 97.33 c29-3c2s6n2
  - 97.28 c29-3c2s6n3
  - 97.85 c29-3c2s7n0
  - 94.63 c29-3c2s7n1
  - 96.41 c29-3c2s7n2
  - 95.73 c29-3c2s7n3
  - 94.78 c27-0c0s0n0
  - 93.84 c27-0c0s0n1
  - 98.06 c27-0c0s0n2
  - 94.99 c27-0c0s0n3
  - 96.73 c27-0c0s1n0
  - 94.20 c27-0c0s1n1
  - 96.36 c27-0c0s1n2
  - 93.67 c27-0c0s1n3

0.00 100.00

Time waiting for last rank to enter MPI_Allgather

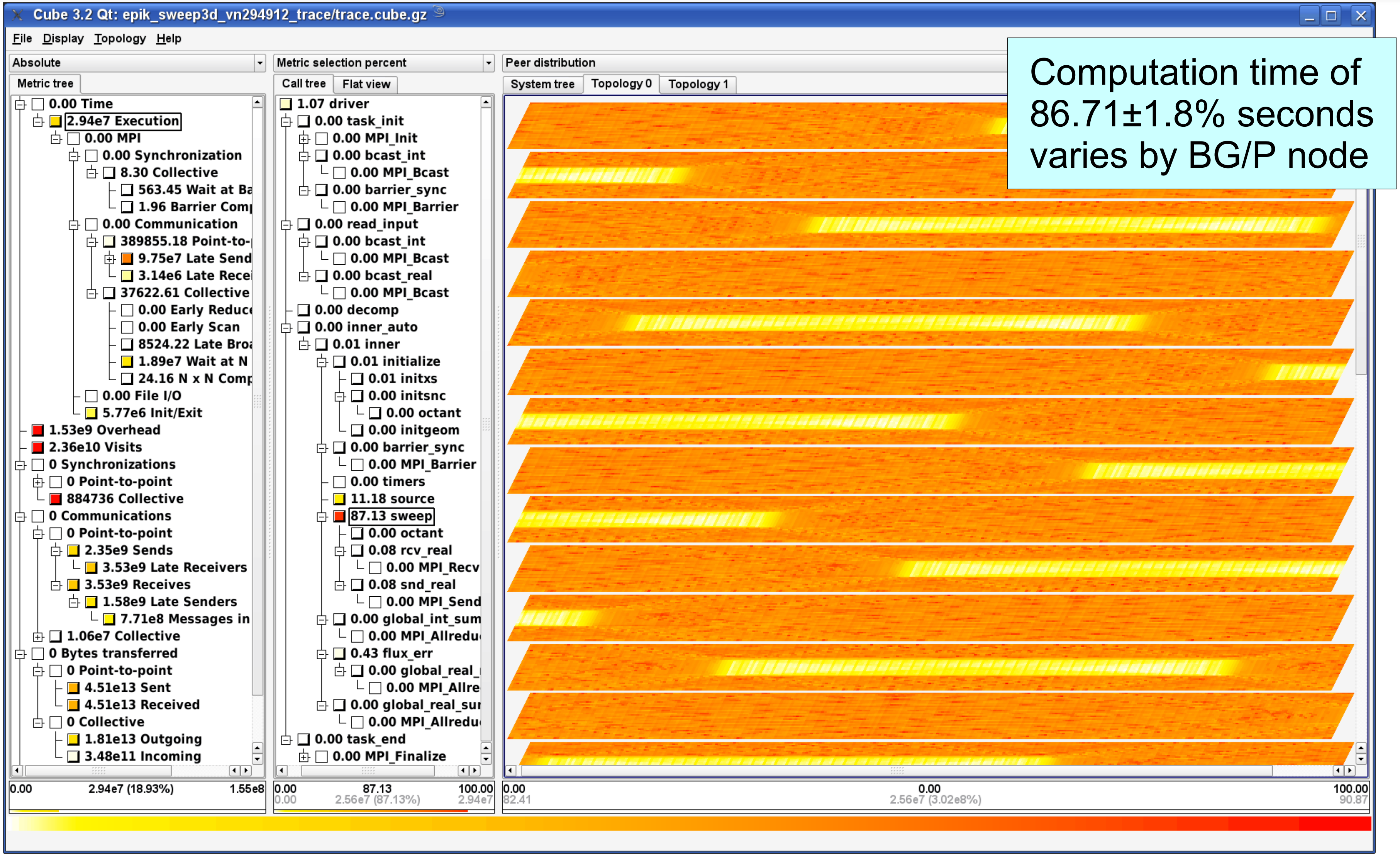
PEPC-B on BG/P & Cray XT experience

- Despite very different processor and network performance, measurements and analyses can be easily compared
  - different compilers affect function naming & in-lining
- Both spend roughly two-thirds of time in computation
  - tree_walk has expensive computation & communication
- Both waste 30% of time waiting to enter MPI_Barrier
  - not localized to particular processes, since particles are regularly redistributed
- Most of collective communication time is also time waiting for last ranks to enter MPI_Allgather & MPI_Alltoall
  - imbalance for MPI_Allgather twice as severe on BlueGene/P, however, almost 50x less for MPI_Alltoall
  - collective completion times also notably longer on Cray XT

Sweep3d on BlueGene/P case study

- 3D neutron transport simulation
  - ASC benchmark
  - direct order solve uses diagonal sweeps through grid cells
- MPI parallel version 2.2b using 2D domain decomposition
  - ~2,000 lines (12 source modules), mostly Fortran77
- Run on IBM BlueGene/P in VN mode with 288k processes
  - 790GB trace written in 47 minutes, analyzed in 7 minutes
    - plus 86 minutes just to create 294,912 files (one per MPI rank)
    - SIONlib being developed to address management of sets of files
  - Mapping of 576x512 grid of processes onto 3D physical torus results in regular pattern of performance artifacts

sweep3d on jugene@288k trace analysis



Cube 3.2 Qt: epik_sweep3d_vn294912_trace/trace.cube.gz

- □ 0.00 Time
  - □ 2.94e7 Execution
    - □ 0.00 MPI
      - □ 0.00 Synchronization
        - □ 8.30 Collective
          - □ 563.45 Wait at Ba
          - □ 1.96 Barrier Comp
      - □ 0.00 Communication
        - □ 389855.18 Point-to-
          - □ 9.75e7 Late Send
          - □ 3.14e6 Late Recei
        - □ 37622.61 Collective
          - □ 0.00 Early Reduc
          - □ 0.00 Early Scan
          - □ 8524.22 Late Broa
          - □ 1.89e7 Wait at N
          - □ 24.16 N x N Comp
      - □ 0.00 File I/O
      - □ 5.77e6 Init/Exit
- □ 1.53e9 Overhead
- □ 2.36e10 Visits
- □ 0 Synchronizations
  - □ 0 Point-to-point
  - □ 884736 Collective
- □ 0 Communications
  - □ 0 Point-to-point
    - □ 2.35e9 Sends
      - □ 3.53e9 Late Receivers
    - □ 3.53e9 Receives
      - □ 1.58e9 Late Senders
        - □ 7.71e8 Messages in
  - □ 1.06e7 Collective
- □ 0 Bytes transferred
  - □ 0 Point-to-point
    - □ 4.51e13 Sent
    - □ 4.51e13 Received
  - □ 0 Collective
    - □ 1.81e13 Outgoing
    - □ 3.48e11 Incoming

- □ 1.07 driver
  - □ 0.00 task_init
    - □ 0.00 MPI_Init
    - □ 0.00 bcast_int
      - □ 0.00 MPI_Bcast
    - □ 0.00 barrier_sync
      - □ 0.00 MPI_Barrier
  - □ 0.00 read_input
    - □ 0.00 bcast_int
      - □ 0.00 MPI_Bcast
    - □ 0.00 bcast_real
      - □ 0.00 MPI_Bcast
  - □ 0.00 decomp
  - □ 0.00 inner_auto
    - □ 0.01 inner
      - □ 0.01 initialize
        - □ 0.01 initxs
        - □ 0.00 initsnc
          - □ 0.00 octant
        - □ 0.00 initgeom
      - □ 0.00 barrier_sync
        - □ 0.00 MPI_Barrier
      - □ 0.00 timers
      - □ 11.18 source
      - □ 87.13 sweep
        - □ 0.00 octant
        - □ 0.08 rcv_real
          - □ 0.00 MPI_Recv
        - □ 0.08 snd_real
          - □ 0.00 MPI_Send
      - □ 0.00 global_int_sum
        - □ 0.00 MPI_Allredu
      - □ 0.43 flux_err
        - □ 0.00 global_real_
          - □ 0.00 MPI_Allre
      - □ 0.00 global_real_sur
        - □ 0.00 MPI_Allredu
  - □ 0.00 task_end
    - □ 0.00 MPI_Finalize

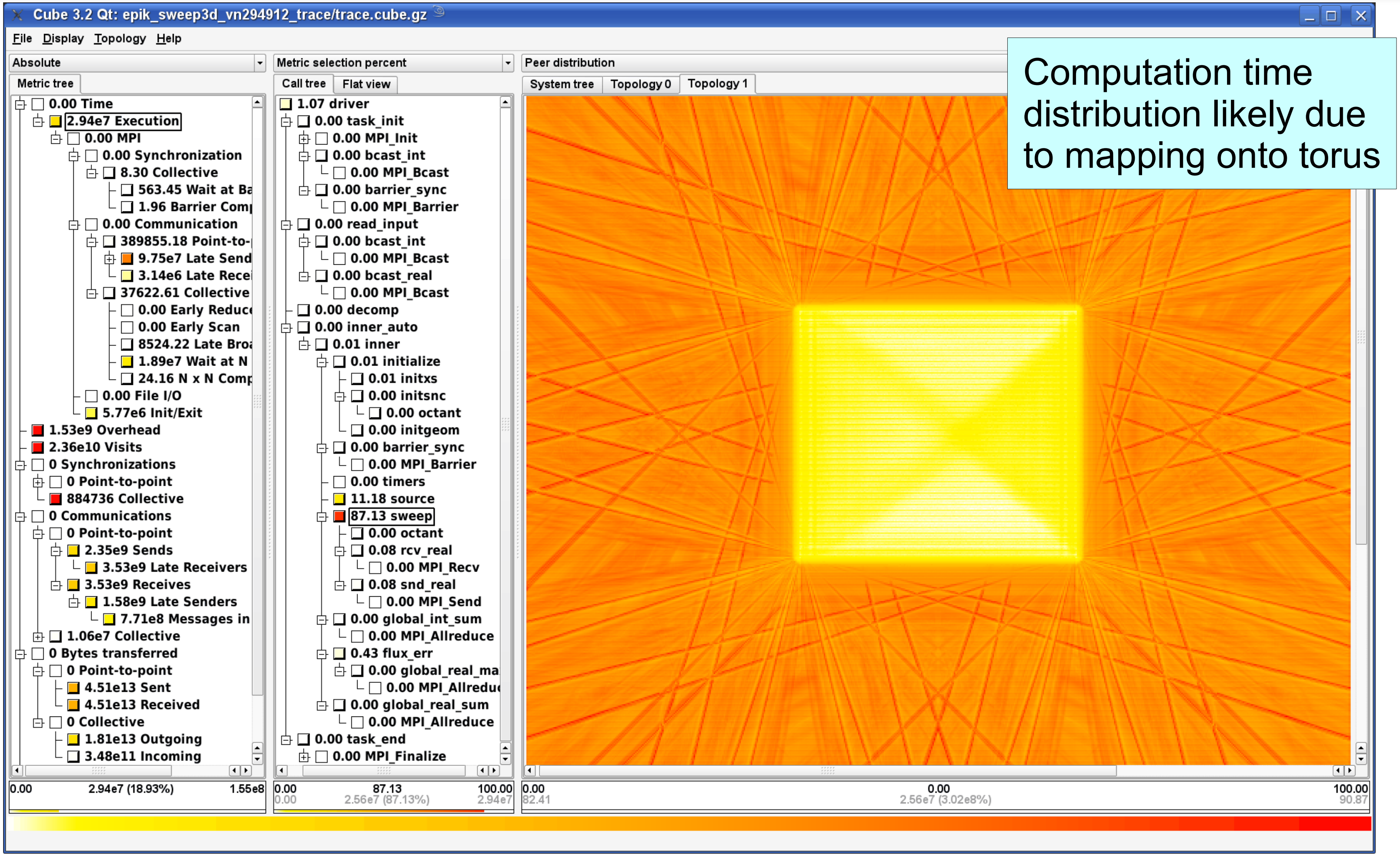
0.00 2.94e7 (18.93%) 1.55e8

0.00 87.13 100.00
0.00 2.56e7 (87.13%) 2.94e7

0.00 0.00 100.00
82.41 2.56e7 (3.02e8%) 90.87

Computation time of 86.71±1.8% seconds varies by BG/P node

sweep3d on jugene@288k trace analysis



Cube 3.2 Qt: epik_sweep3d_vn294912_trace/trace.cube.gz

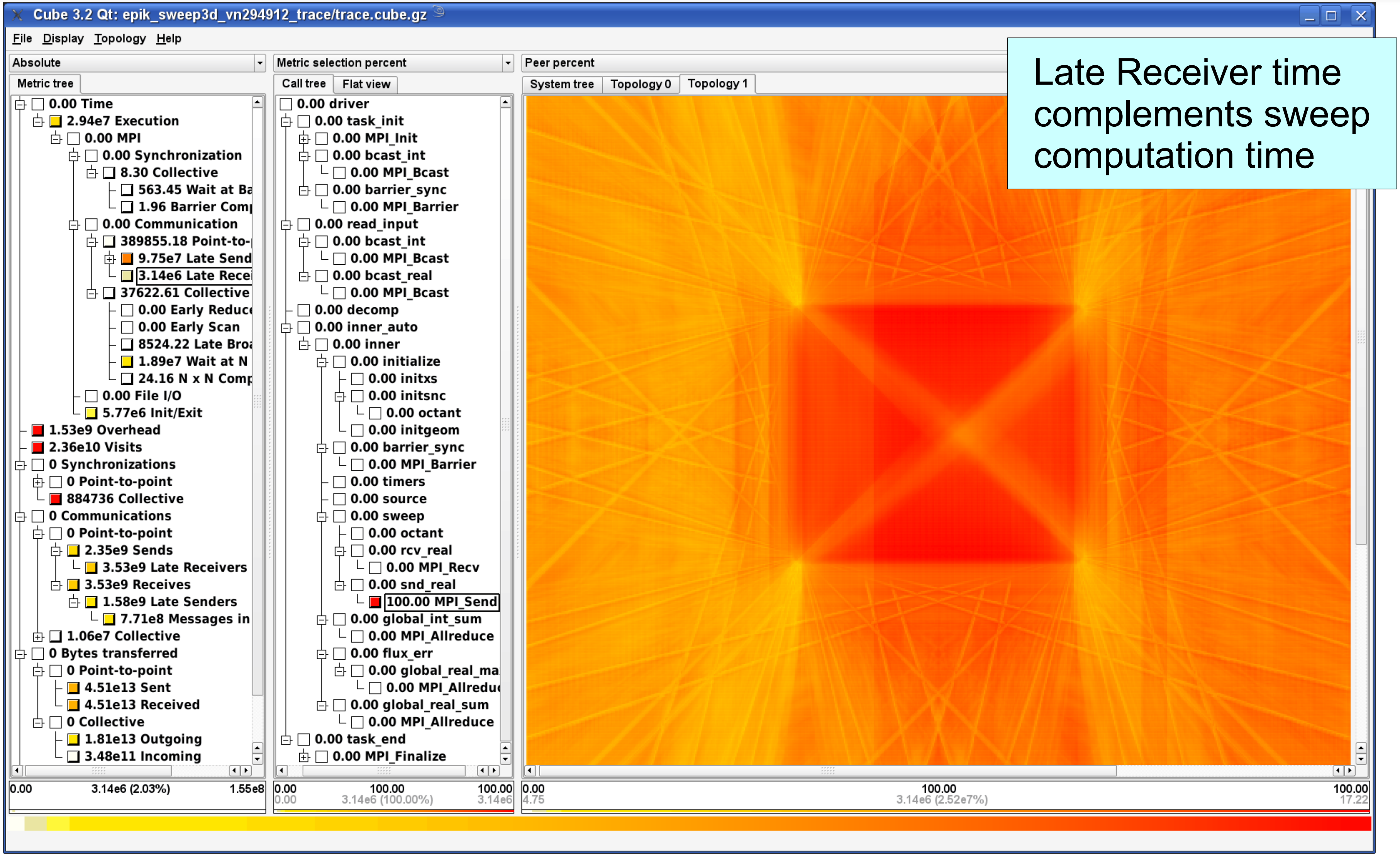
```
0.00 Time
  2.94e7 Execution
    0.00 MPI
      0.00 Synchronization
        8.30 Collective
          563.45 Wait at Ba
          1.96 Barrier Comp
      0.00 Communication
        389855.18 Point-to-
          9.75e7 Late Send
          3.14e6 Late Recei
        37622.61 Collective
          0.00 Early Reduc
          0.00 Early Scan
          8524.22 Late Broa
          1.89e7 Wait at N
          24.16 N x N Comp
      0.00 File I/O
      5.77e6 Init/Exit
1.53e9 Overhead
2.36e10 Visits
0 Synchronizations
  0 Point-to-point
  884736 Collective
0 Communications
  0 Point-to-point
    2.35e9 Sends
      3.53e9 Late Receivers
    3.53e9 Receives
      1.58e9 Late Senders
        7.71e8 Messages in
  1.06e7 Collective
0 Bytes transferred
  0 Point-to-point
    4.51e13 Sent
    4.51e13 Received
  0 Collective
    1.81e13 Outgoing
    3.48e11 Incoming
```

```
1.07 driver
  0.00 task_init
    0.00 MPI_Init
    0.00 bcast_int
      0.00 MPI_Bcast
    0.00 barrier_sync
      0.00 MPI_Barrier
  0.00 read_input
    0.00 bcast_int
      0.00 MPI_Bcast
    0.00 bcast_real
      0.00 MPI_Bcast
  0.00 decomp
  0.00 inner_auto
    0.01 inner
      0.01 initialize
        0.01 initxs
        0.00 initsnc
          0.00 octant
        0.00 initgeom
      0.00 barrier_sync
        0.00 MPI_Barrier
      0.00 timers
      11.18 source
      87.13 sweep
        0.00 octant
        0.08 rcv_real
          0.00 MPI_Recv
        0.08 snd_real
          0.00 MPI_Send
      0.00 global_int_sum
        0.00 MPI_Allreduce
      0.43 flux_err
        0.00 global_real_ma
          0.00 MPI_Allreduc
      0.00 global_real_sum
        0.00 MPI_Allreduce
  0.00 task_end
    0.00 MPI_Finalize
```

```
0.00        2.94e7 (18.93%)        1.55e8

0.00        87.13                  100.00
0.00        2.56e7 (87.13%)        2.94e7

0.00        0.00                   100.00
82.41       2.56e7 (3.02e8%)       90.87
```

Computation time distribution likely due to mapping onto torus

sweep3d on jugene@288k trace (wait) analysis



Cube 3.2 Qt: epik_sweep3d_vn294912_trace/trace.cube.gz

```
0.00 Time
  2.94e7 Execution
    0.00 MPI
      0.00 Synchronization
        8.30 Collective
          563.45 Wait at Ba
          1.96 Barrier Comp
      0.00 Communication
        389855.18 Point-to-
          9.75e7 Late Send
          3.14e6 Late Recei
        37622.61 Collective
          0.00 Early Reduc
          0.00 Early Scan
          8524.22 Late Broa
          1.89e7 Wait at N
          24.16 N x N Comp
      0.00 File I/O
      5.77e6 Init/Exit
1.53e9 Overhead
2.36e10 Visits
0 Synchronizations
  0 Point-to-point
  884736 Collective
0 Communications
  0 Point-to-point
    2.35e9 Sends
      3.53e9 Late Receivers
    3.53e9 Receives
      1.58e9 Late Senders
        7.71e8 Messages in
  1.06e7 Collective
0 Bytes transferred
  0 Point-to-point
    4.51e13 Sent
    4.51e13 Received
  0 Collective
    1.81e13 Outgoing
    3.48e11 Incoming
```

```
0.00 driver
  0.00 task_init
    0.00 MPI_Init
    0.00 bcast_int
      0.00 MPI_Bcast
    0.00 barrier_sync
      0.00 MPI_Barrier
  0.00 read_input
    0.00 bcast_int
      0.00 MPI_Bcast
    0.00 bcast_real
      0.00 MPI_Bcast
  0.00 decomp
  0.00 inner_auto
    0.00 inner
      0.00 initialize
        0.00 initxs
        0.00 initsnc
          0.00 octant
        0.00 initgeom
      0.00 barrier_sync
        0.00 MPI_Barrier
      0.00 timers
      0.00 source
      0.00 sweep
        0.00 octant
        0.00 rcv_real
          0.00 MPI_Recv
        0.00 snd_real
          100.00 MPI_Send
      0.00 global_int_sum
        0.00 MPI_Allreduce
      0.00 flux_err
        0.00 global_real_ma
          0.00 MPI_Allreduc
      0.00 global_real_sum
        0.00 MPI_Allreduce
  0.00 task_end
    0.00 MPI_Finalize
```

```
0.00        3.14e6 (2.03%)        1.55e8
```

```
0.00        100.00              100.00
0.00        3.14e6 (100.00%)    3.14e6
```

```
0.00        100.00              100.00
4.75        3.14e6 (2.52e7%)    17.22
```

Late Receiver time complements sweep computation time

Acknowledgements

- The Scalasca development team
- The application and benchmark developers who provided their codes and/or measurement archives
- The facilities who made their HPC resources available and associated support staff who helped us use them effectively
  - ALCF, BSC, CSC, CSCS, EPCC, JSC, HLRN, HLRS, ICL, LRZ, NCAR, NCCS, NICS, RWTH, RZG, TeraGrid/TACC, ZIH

Further information

Scalable performance **a**nalysis of **la**rge-**sca**le parallel **a**pplications

- 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