scalasca Performance analysis & tuning case studies

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

September 2011

Additional Live-DVD example experiments

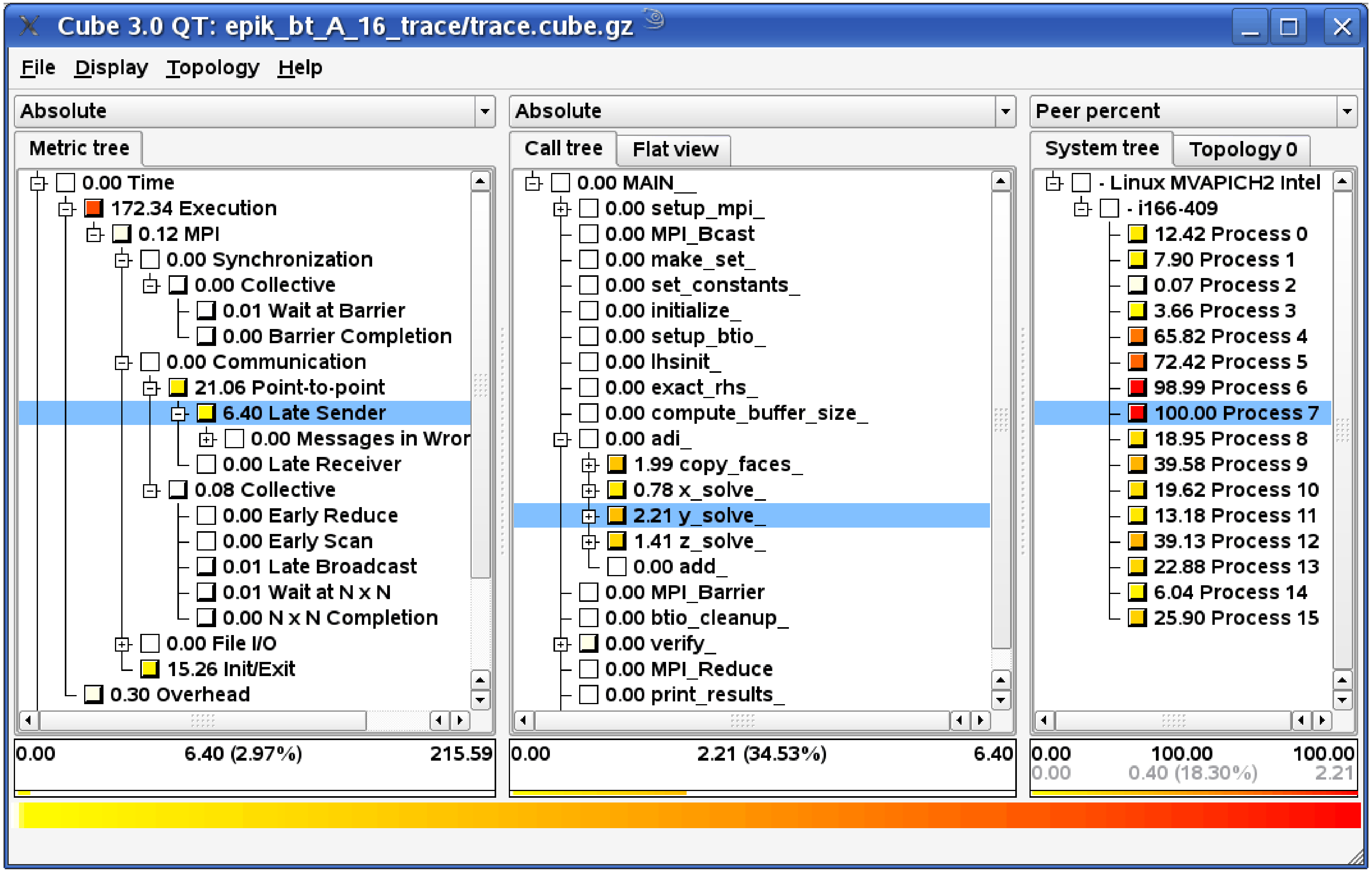
- Example experiment archives provided for examination:
  - jugene_sweep3d
    - 294,912 & 65,536 MPI processes on BG/P (trace)
  - jump_zeusmp2
    - 512 MPI processes on p690 cluster (summary & trace)
  - marenostrum_wrf-nmm
    - 1600 MPI processes on JS21 blade cluster, solver extract
    - summary 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

Concurrency & Computation: Practice & Experience 22(6):702-719 (2010)

16-process trace analysis



Cube 3.0 QT: epik_bt_A_16_trace/trace.cube.gz

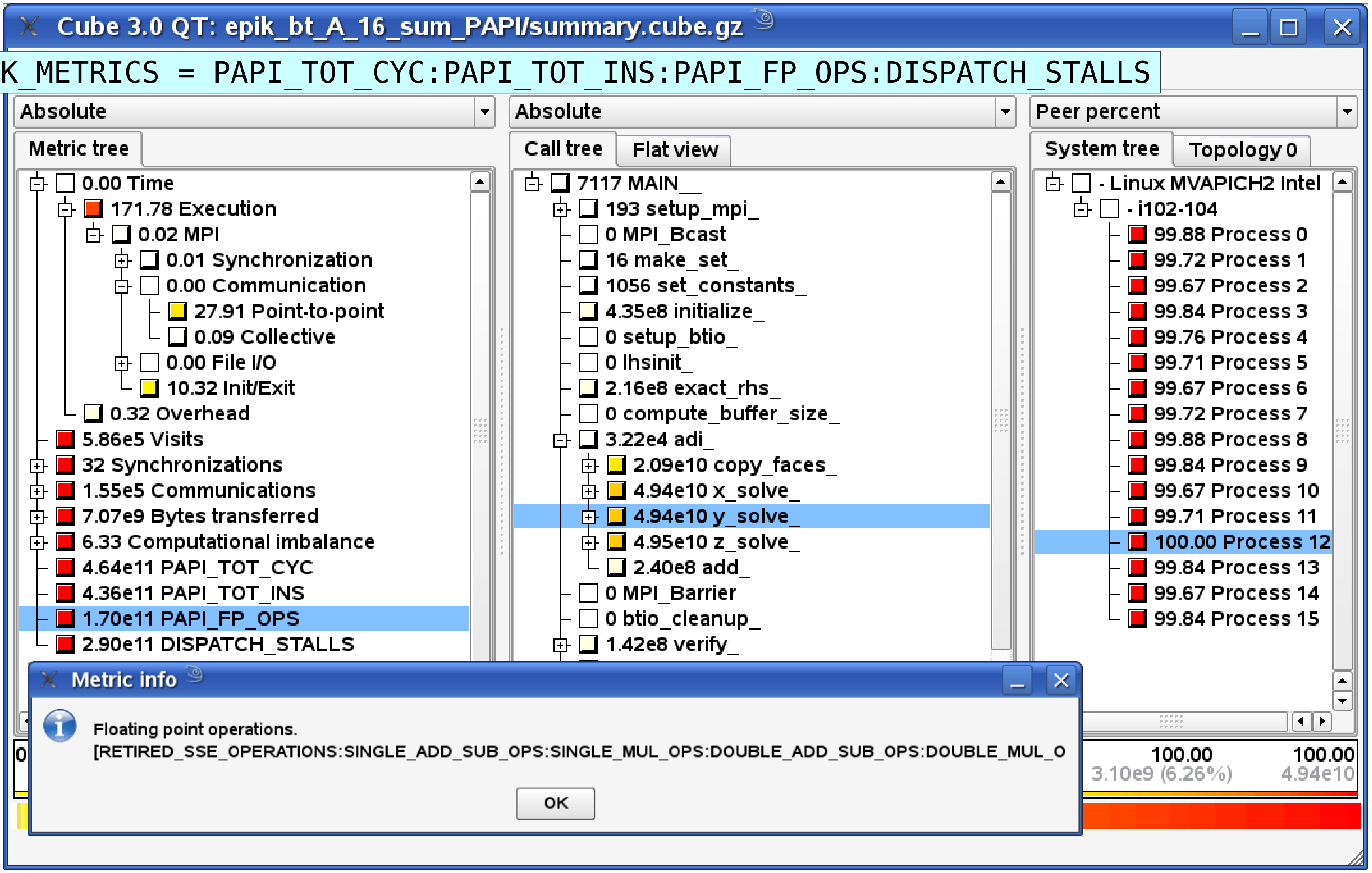
File Display Topology Help

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

0.00 Time                         0.00 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
```

Absolute | Absolute | Peer percent

Metric tree:

```
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
```

Call tree:

```
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_
```

System tree:

```
- 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
```

100.00 3.10e9 (6.26%) 100.00 4.94e10

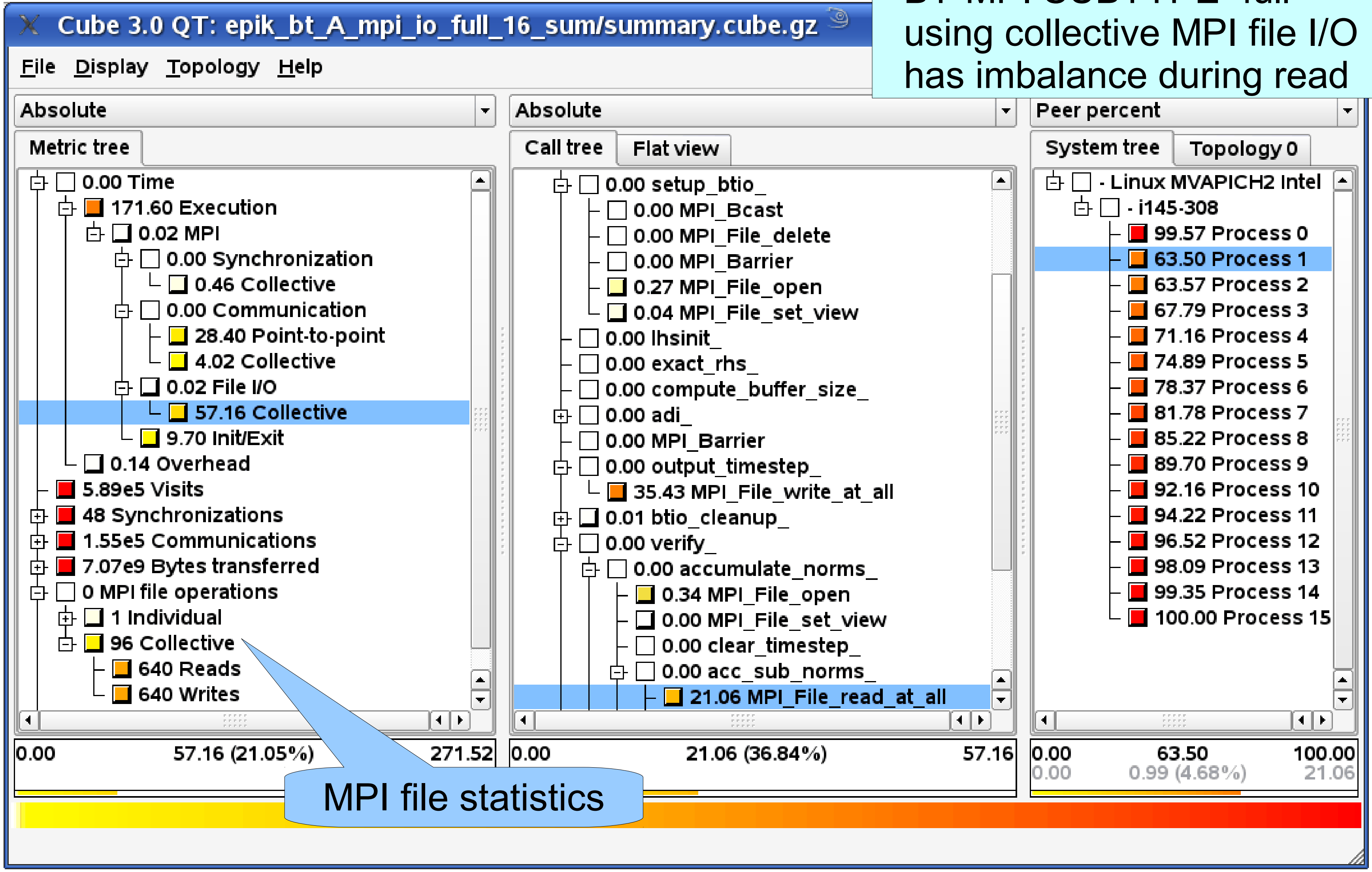
Metric info

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

OK

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

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



Metric tree:

```
0.00 Time
  171.60 Execution
    0.02 MPI
      0.00 Synchronization
        0.46 Collective
      0.00 Communication
        28.40 Point-to-point
        4.02 Collective
      0.02 File I/O
        57.16 Collective
      9.70 Init/Exit
    0.14 Overhead
5.89e5 Visits
48 Synchronizations
1.55e5 Communications
7.07e9 Bytes transferred
0 MPI file operations
  1 Individual
  96 Collective
    640 Reads
    640 Writes
```

```
0.00          57.16 (21.05%)          271.52
```

Call tree:

```
0.00 setup_btio_
  0.00 MPI_Bcast
  0.00 MPI_File_delete
  0.00 MPI_Barrier
  0.27 MPI_File_open
  0.04 MPI_File_set_view
0.00 lhsinit_
0.00 exact_rhs_
0.00 compute_buffer_size_
0.00 adi_
0.00 MPI_Barrier
0.00 output_timestep_
  35.43 MPI_File_write_at_all
0.01 btio_cleanup_
0.00 verify_
  0.00 accumulate_norms_
    0.34 MPI_File_open
    0.00 MPI_File_set_view
    0.00 clear_timestep_
    0.00 acc_sub_norms_
      21.06 MPI_File_read_at_all
```

```
0.00          21.06 (36.84%)          57.16
```

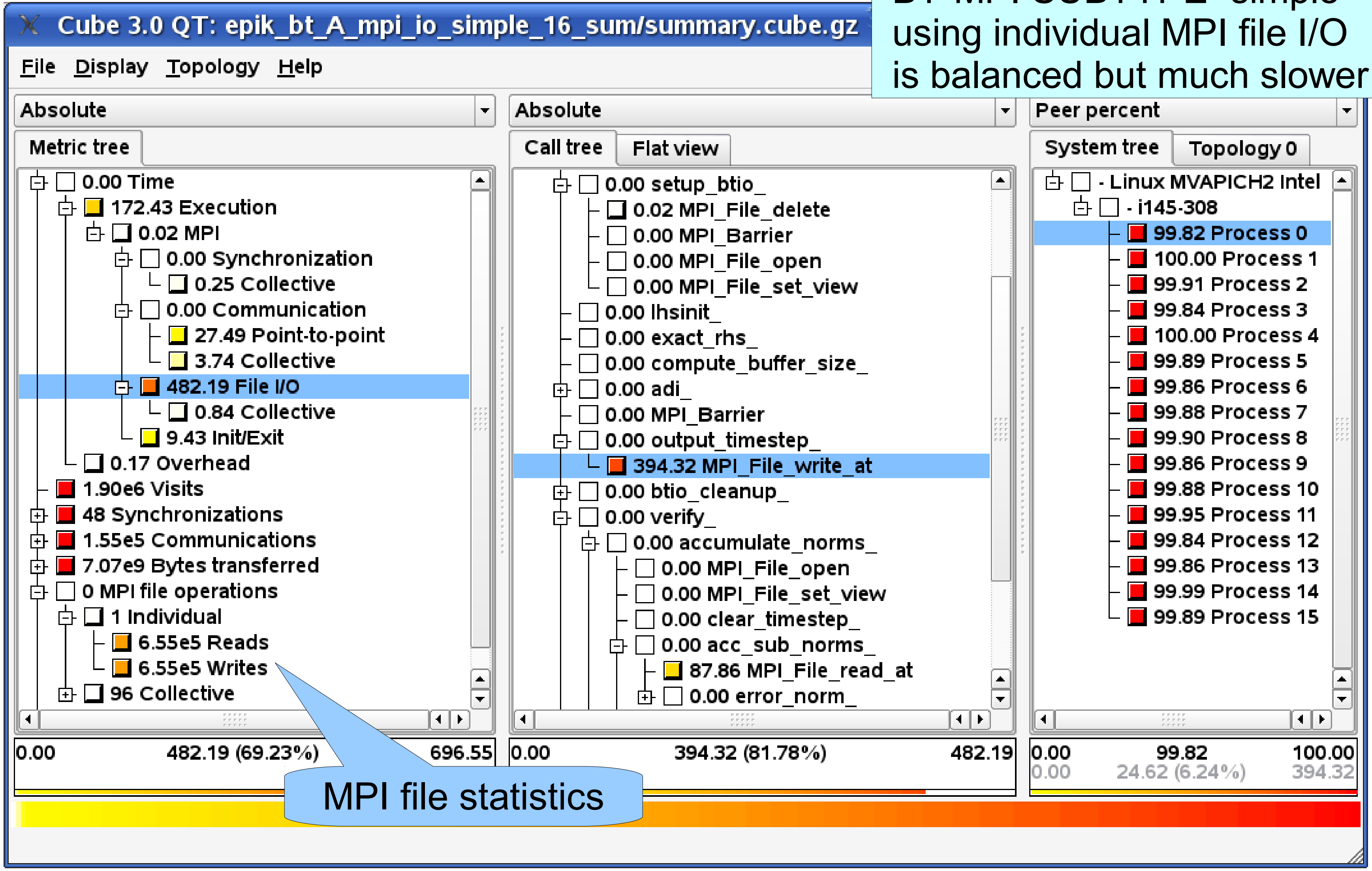
System tree:

```
- Linux MVAPICH2 Intel
  - i145-308
    99.57 Process 0
    63.50 Process 1
    63.57 Process 2
    67.79 Process 3
    71.16 Process 4
    74.89 Process 5
    78.37 Process 6
    81.78 Process 7
    85.22 Process 8
    89.70 Process 9
    92.16 Process 10
    94.22 Process 11
    96.52 Process 12
    98.09 Process 13
    99.35 Process 14
    100.00 Process 15
```

```
0.00          63.50          100.00
0.00     0.99 (4.68%)        21.06
```

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:

```
0.00 Time
  172.43 Execution
    0.02 MPI
      0.00 Synchronization
        0.25 Collective
      0.00 Communication
        27.49 Point-to-point
        3.74 Collective
      482.19 File I/O
        0.84 Collective
    9.43 Init/Exit
  0.17 Overhead
1.90e6 Visits
48 Synchronizations
1.55e5 Communications
7.07e9 Bytes transferred
0 MPI file operations
  1 Individual
    6.55e5 Reads
    6.55e5 Writes
  96 Collective
```

Call tree:

```
0.00 setup_btio_
  0.02 MPI_File_delete
  0.00 MPI_Barrier
  0.00 MPI_File_open
  0.00 MPI_File_set_view
0.00 lhsinit_
0.00 exact_rhs_
0.00 compute_buffer_size_
0.00 adi_
0.00 MPI_Barrier
0.00 output_timestep_
  394.32 MPI_File_write_at
0.00 btio_cleanup_
0.00 verify_
  0.00 accumulate_norms_
    0.00 MPI_File_open
    0.00 MPI_File_set_view
    0.00 clear_timestep_
    0.00 acc_sub_norms_
      87.86 MPI_File_read_at
    0.00 error_norm_
```

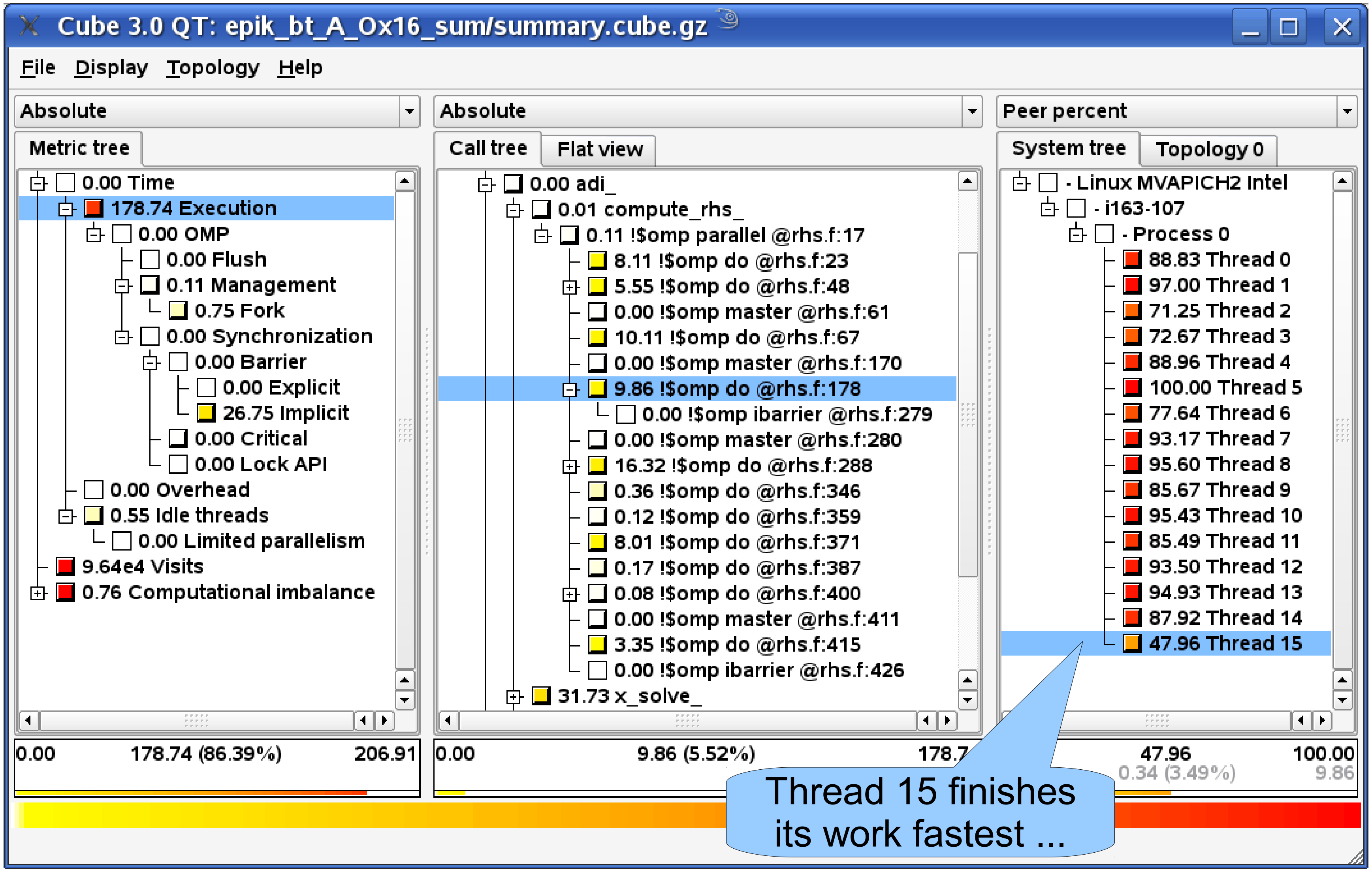
System tree:

```
- Linux MVAPICH2 Intel
  - i145-308
    99.82 Process 0
    100.00 Process 1
    99.91 Process 2
    99.84 Process 3
    100.00 Process 4
    99.89 Process 5
    99.86 Process 6
    99.88 Process 7
    99.90 Process 8
    99.86 Process 9
    99.88 Process 10
    99.95 Process 11
    99.84 Process 12
    99.86 Process 13
    99.99 Process 14
    99.89 Process 15
```

```
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

File Display Topology Help

Absolute | Absolute | Peer percent

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
```

0.00 178.74 (86.39%) 206.91

Call tree | Flat view

```
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_
```

0.00 9.86 (5.52%) 178.7

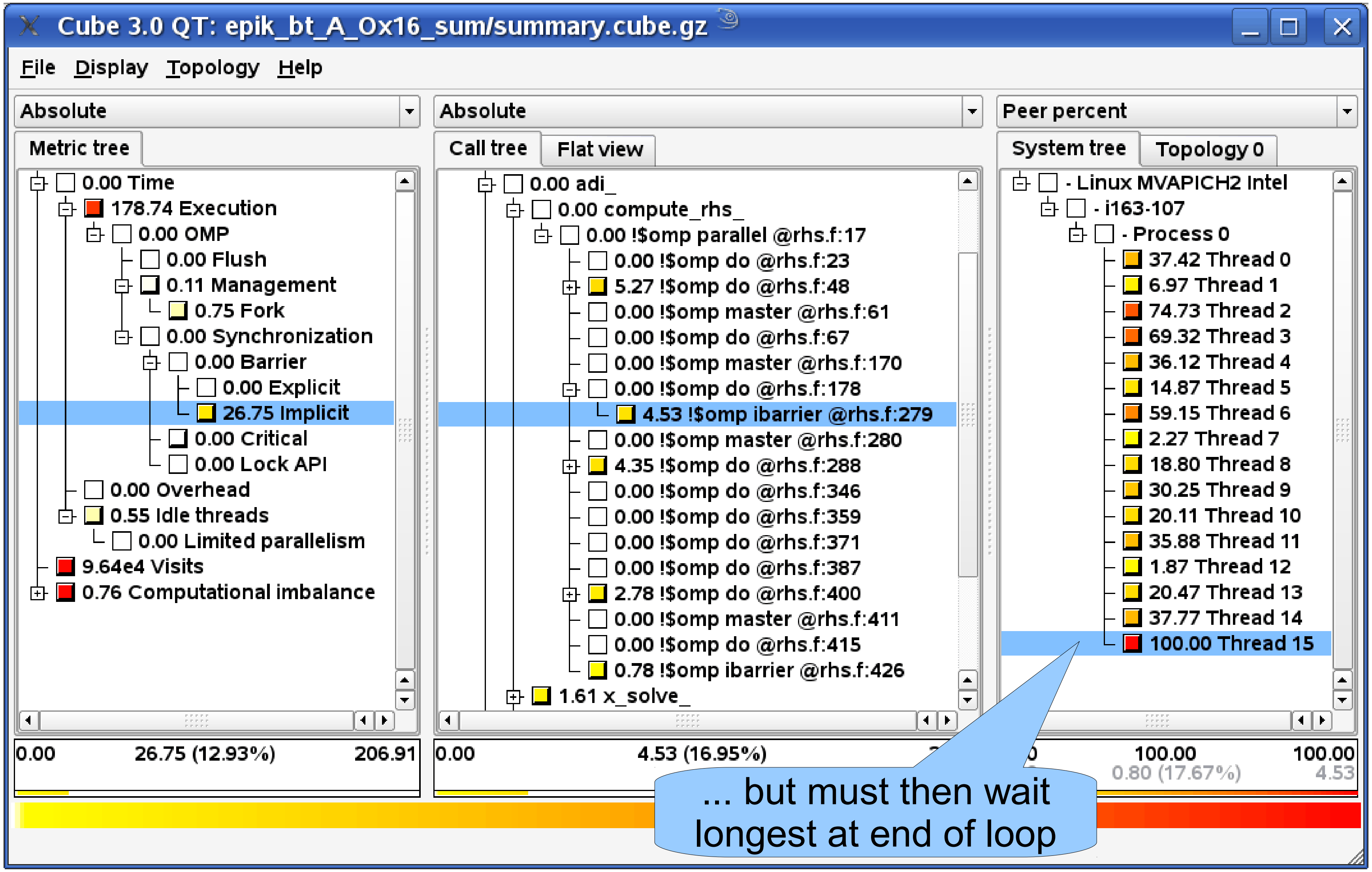
System tree | Topology 0

```
- 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
```

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

File Display Topology Help

Absolute | Absolute | Peer percent

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

```
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
```

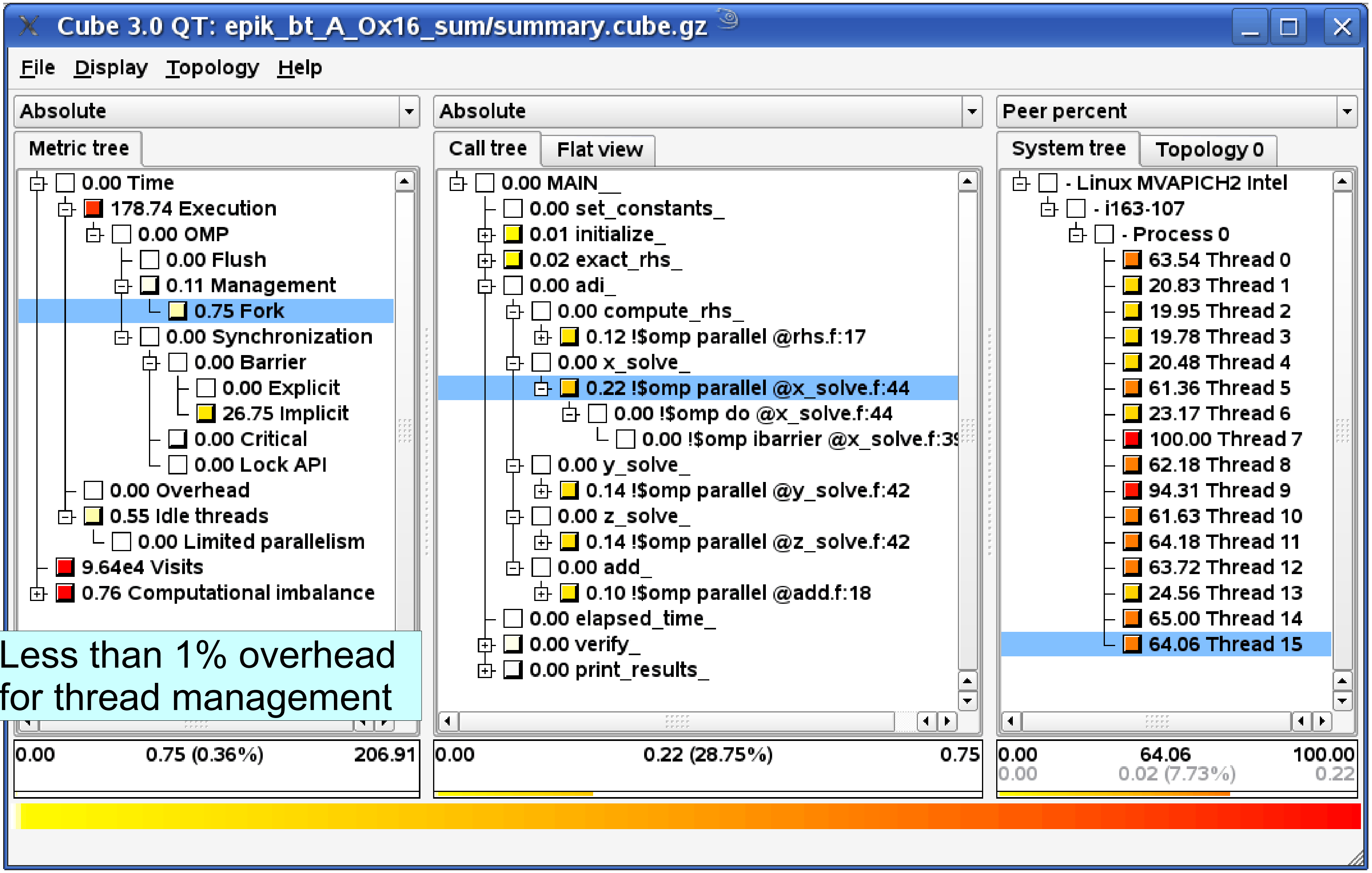
```
0.00 adi_
  0.00 compute_rhs_
    0.00 !$omp parallel @rhs.f:17
      0.00 !$omp do @rhs.f:23
      5.27 !$omp do @rhs.f:48
      0.00 !$omp master @rhs.f:61
      0.00 !$omp do @rhs.f:67
      0.00 !$omp master @rhs.f:170
      0.00 !$omp do @rhs.f:178
        4.53 !$omp ibarrier @rhs.f:279
      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%)
        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



```
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
```

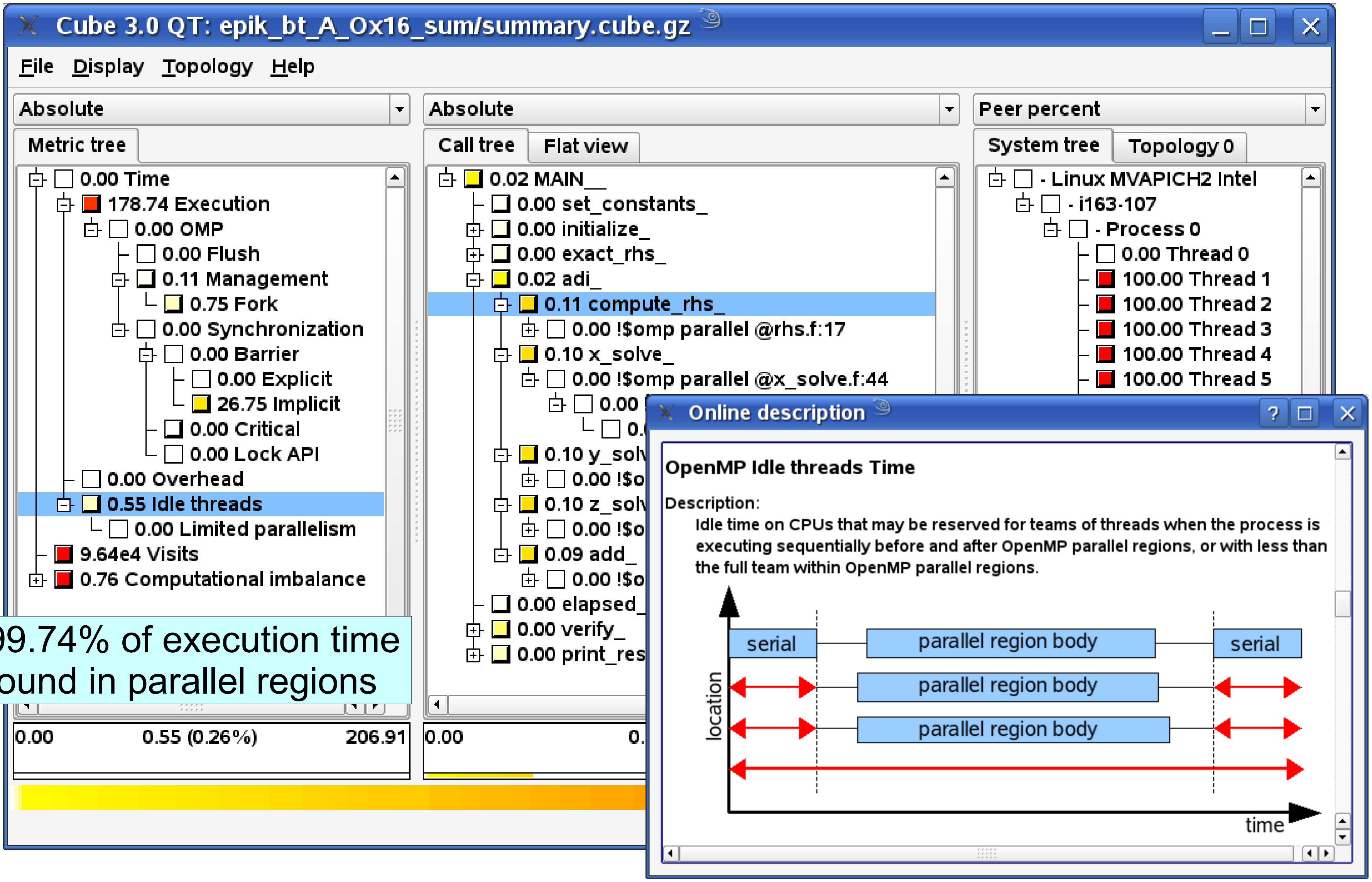
```
Call tree:
0.00 MAIN__
  0.00 set_constants_
  0.01 initialize_
  0.02 exact_rhs_
  0.00 adi_
    0.00 compute_rhs_
      0.12 !$omp parallel @rhs.f:17
    0.00 x_solve_
      0.22 !$omp parallel @x_solve.f:44
        0.00 !$omp do @x_solve.f:44
          0.00 !$omp ibarrier @x_solve.f:3
    0.00 y_solve_
      0.14 !$omp parallel @y_solve.f:42
    0.00 z_solve_
      0.14 !$omp parallel @z_solve.f:42
    0.00 add_
      0.10 !$omp parallel @add.f:18
  0.00 elapsed_time_
  0.00 verify_
  0.00 print_results_
```

```
System tree:
- Linux MVAPICH2 Intel
  - i163-107
    - Process 0
      63.54 Thread 0
      20.83 Thread 1
      19.95 Thread 2
      19.78 Thread 3
      20.48 Thread 4
      61.36 Thread 5
      23.17 Thread 6
      100.00 Thread 7
      62.18 Thread 8
      94.31 Thread 9
      61.63 Thread 10
      64.18 Thread 11
      63.72 Thread 12
      24.56 Thread 13
      65.00 Thread 14
      64.06 Thread 15
```

```
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

File Display Topology Help

Absolute | Absolute | Peer percent

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.

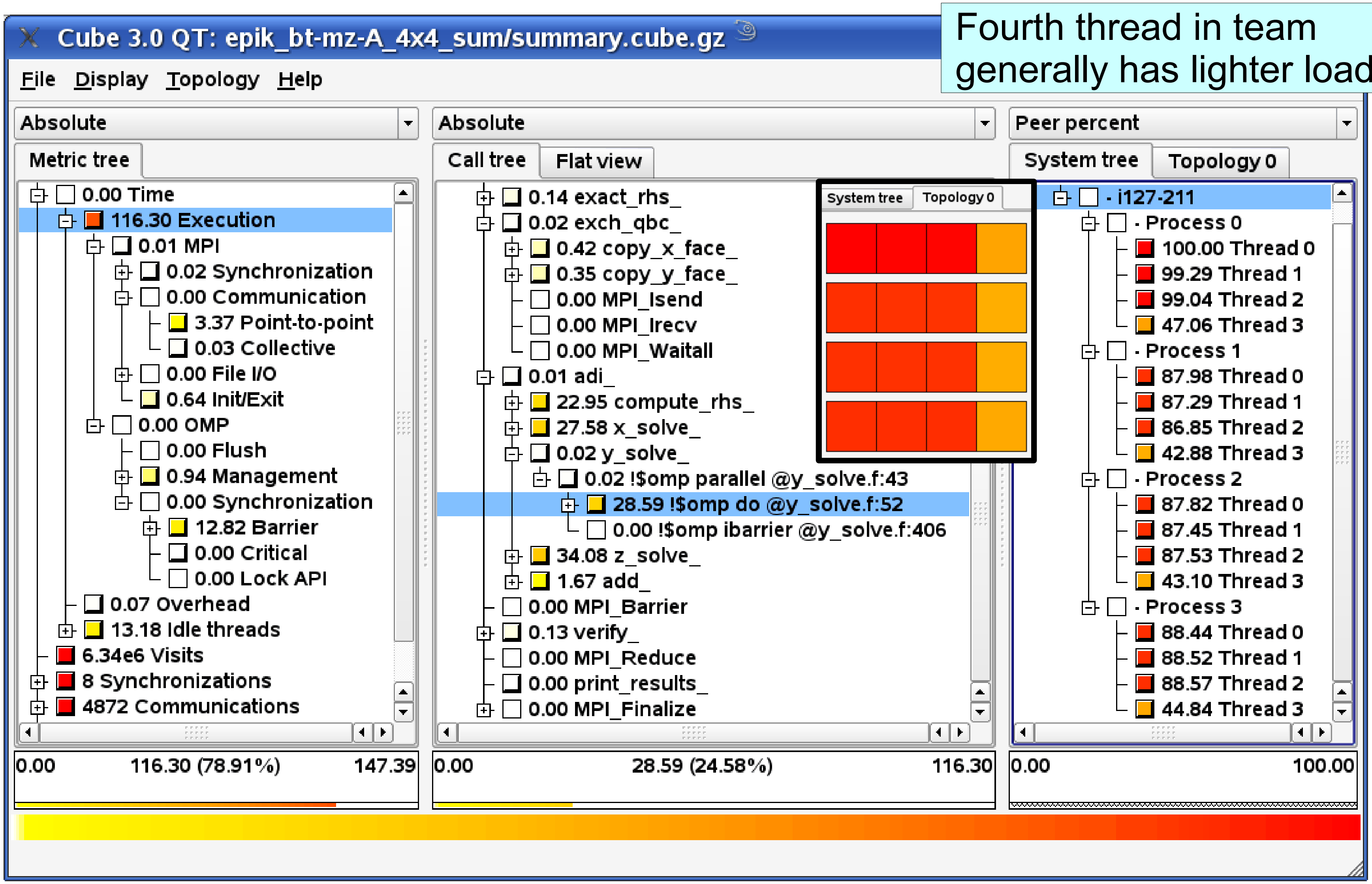
Online description

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.

99.74% of execution time found in parallel regions

4x4 summary analysis: Execution time



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

File Display Topology Help

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:

```
0.14 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
```

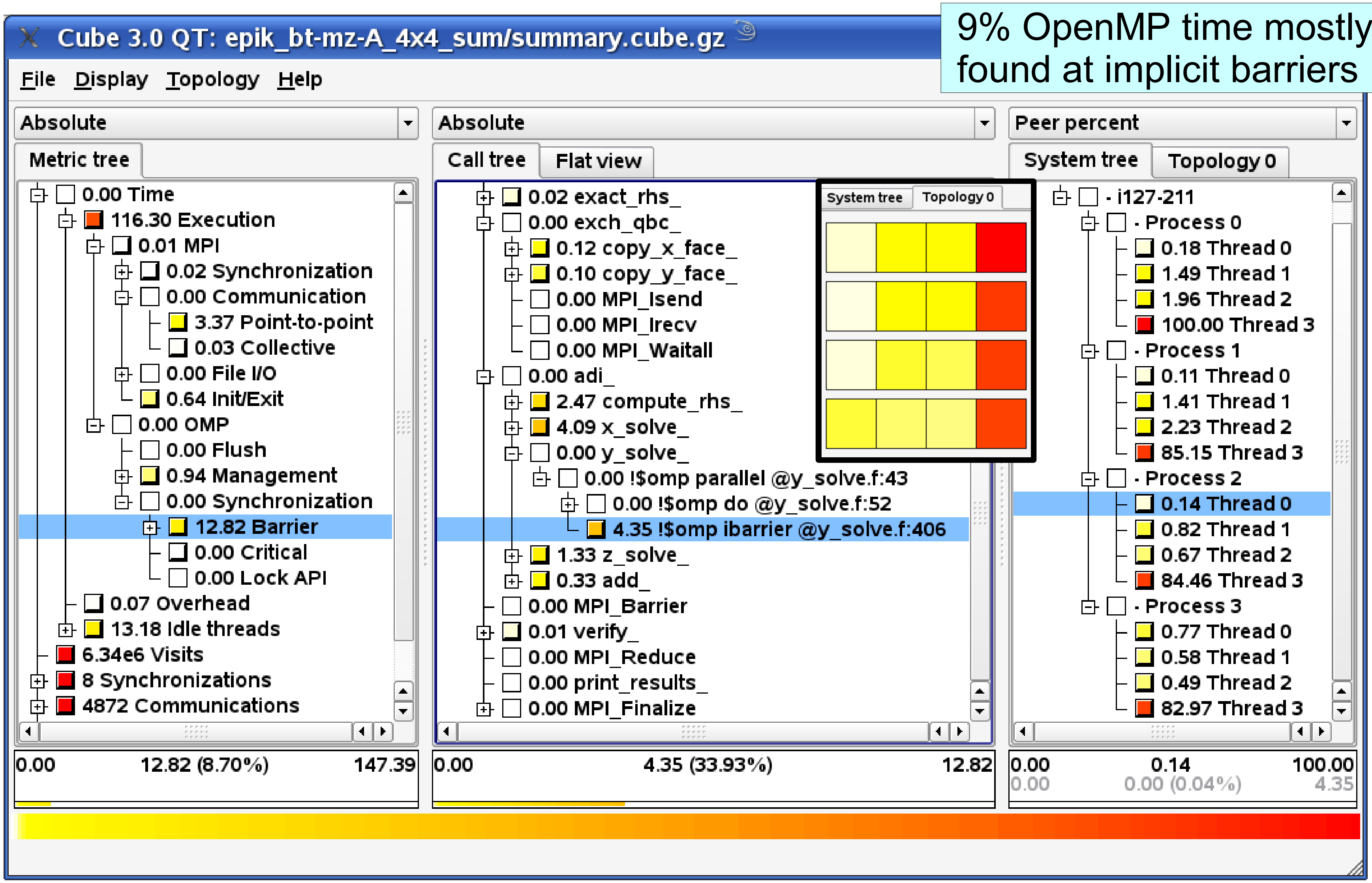
System tree:

```
- 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

File Display Topology Help

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
```

```
0.00        12.82 (8.70%)        147.39
```

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
```

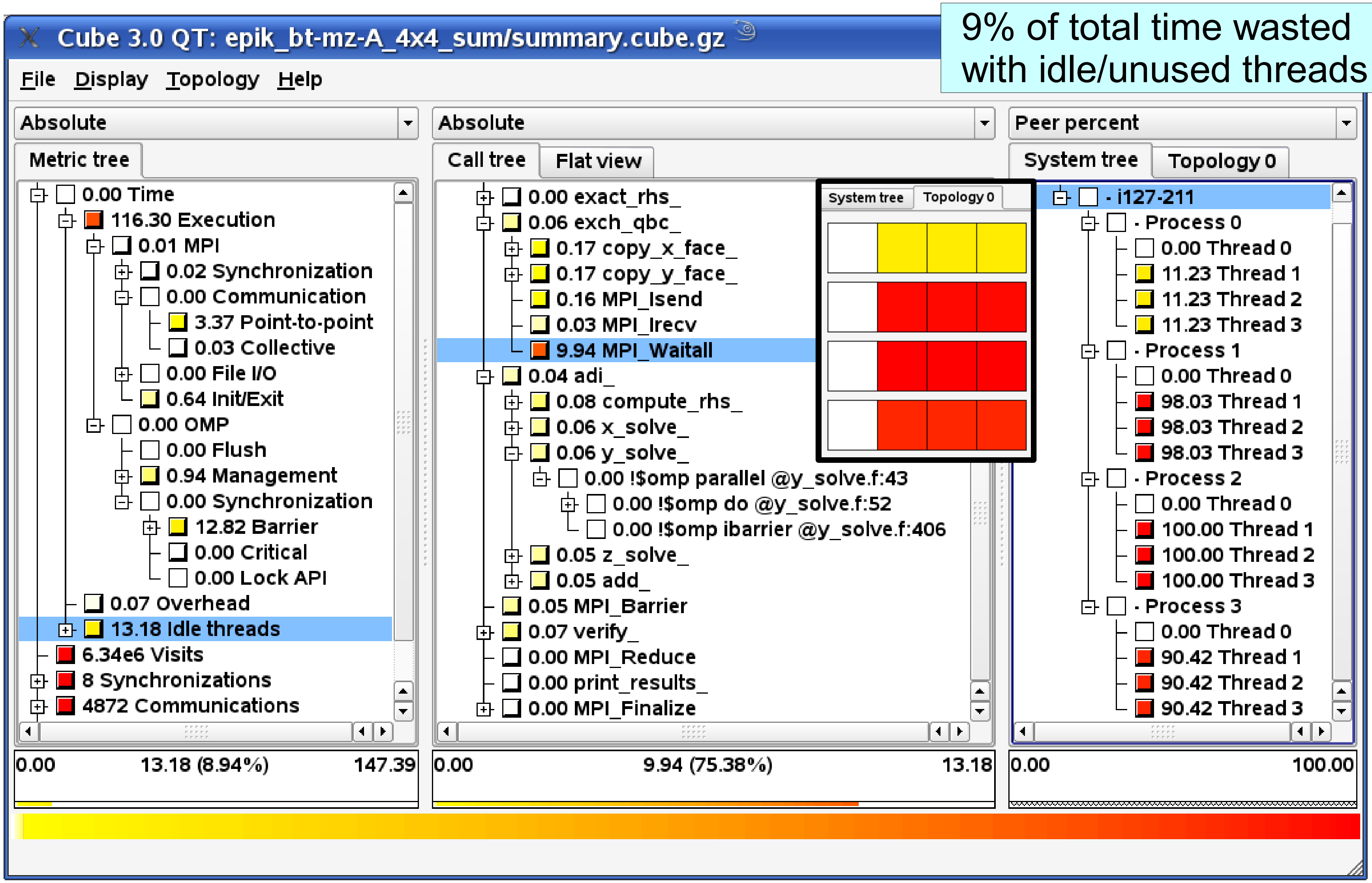
```
0.00        4.35 (33.93%)        12.82
```

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        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

File Display Topology Help

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 / Flat view

```
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
```

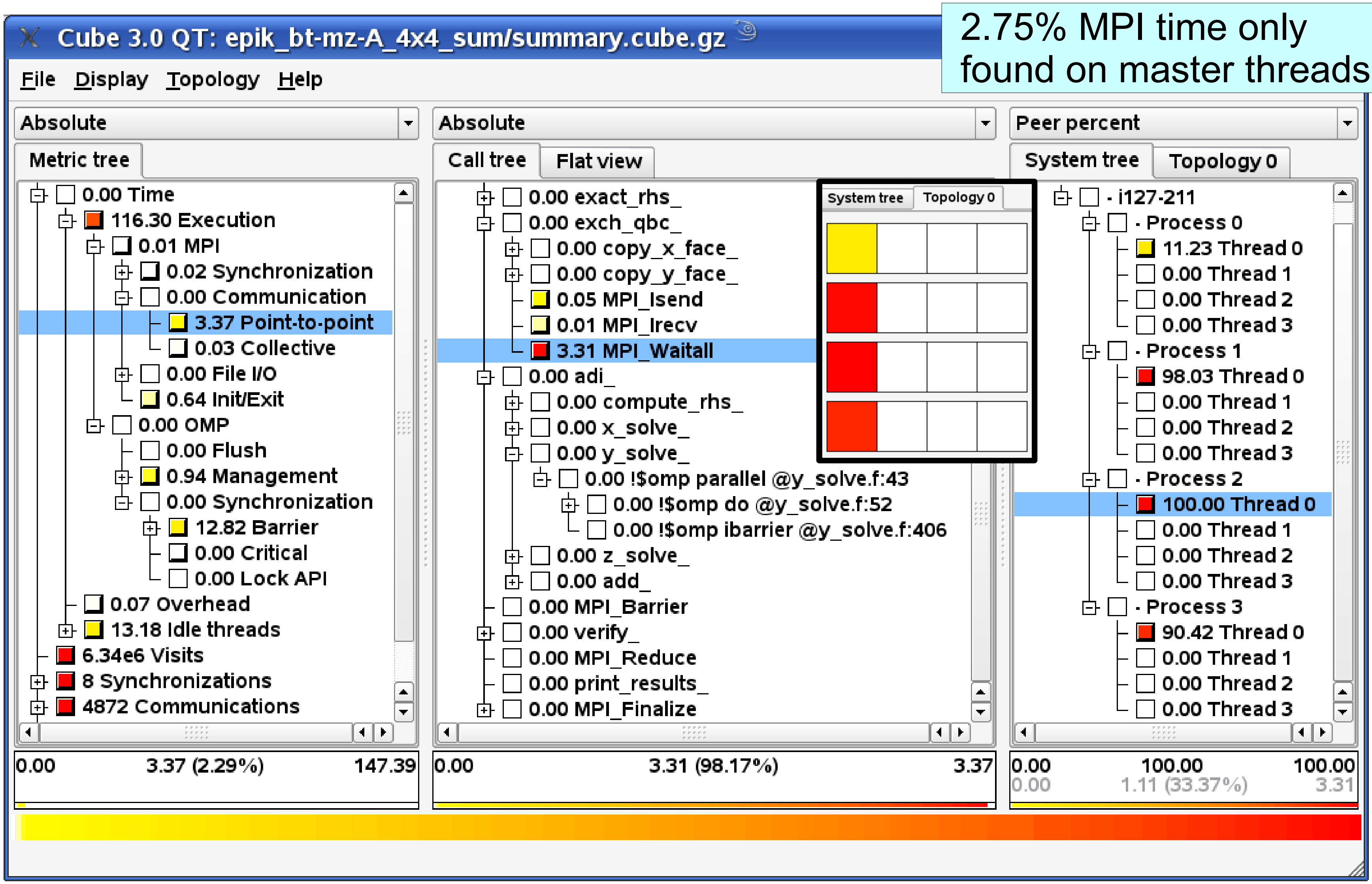
System tree / Topology 0

```
- 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



2.75% MPI time only found on master threads

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

File Display Topology Help

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:

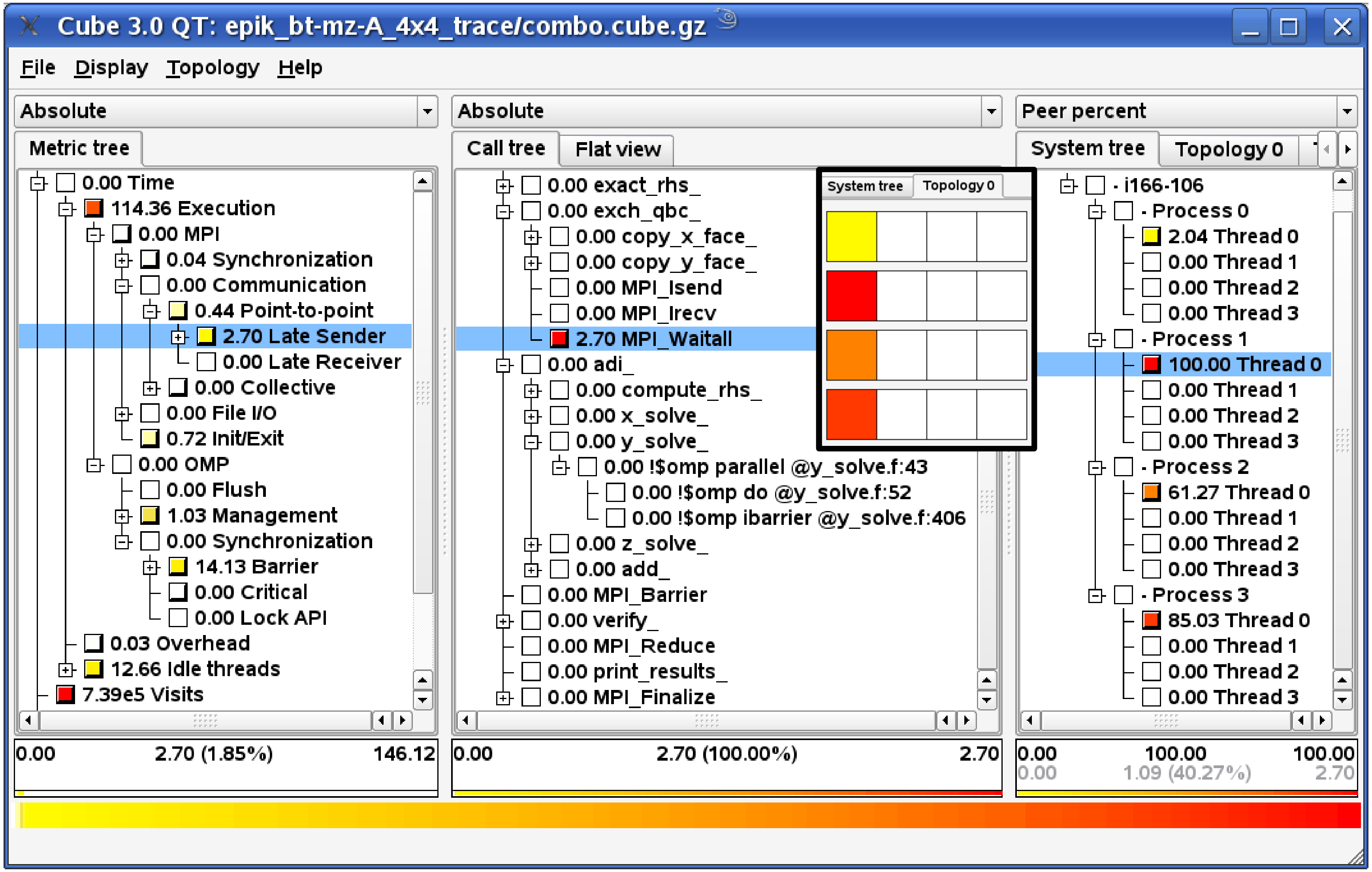
```
0.00 exact_rhs_
0.00 exch_qbc_
  0.00 copy_x_face_
  0.00 copy_y_face_
  0.05 MPI_Isend
  0.01 MPI_Irecv
  3.31 MPI_Waitall
0.00 adi_
  0.00 compute_rhs_
  0.00 x_solve_
  0.00 y_solve_
    0.00 !$omp parallel @y_solve.f:43
      0.00 !$omp do @y_solve.f:52
      0.00 !$omp ibarrier @y_solve.f:406
  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
```

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

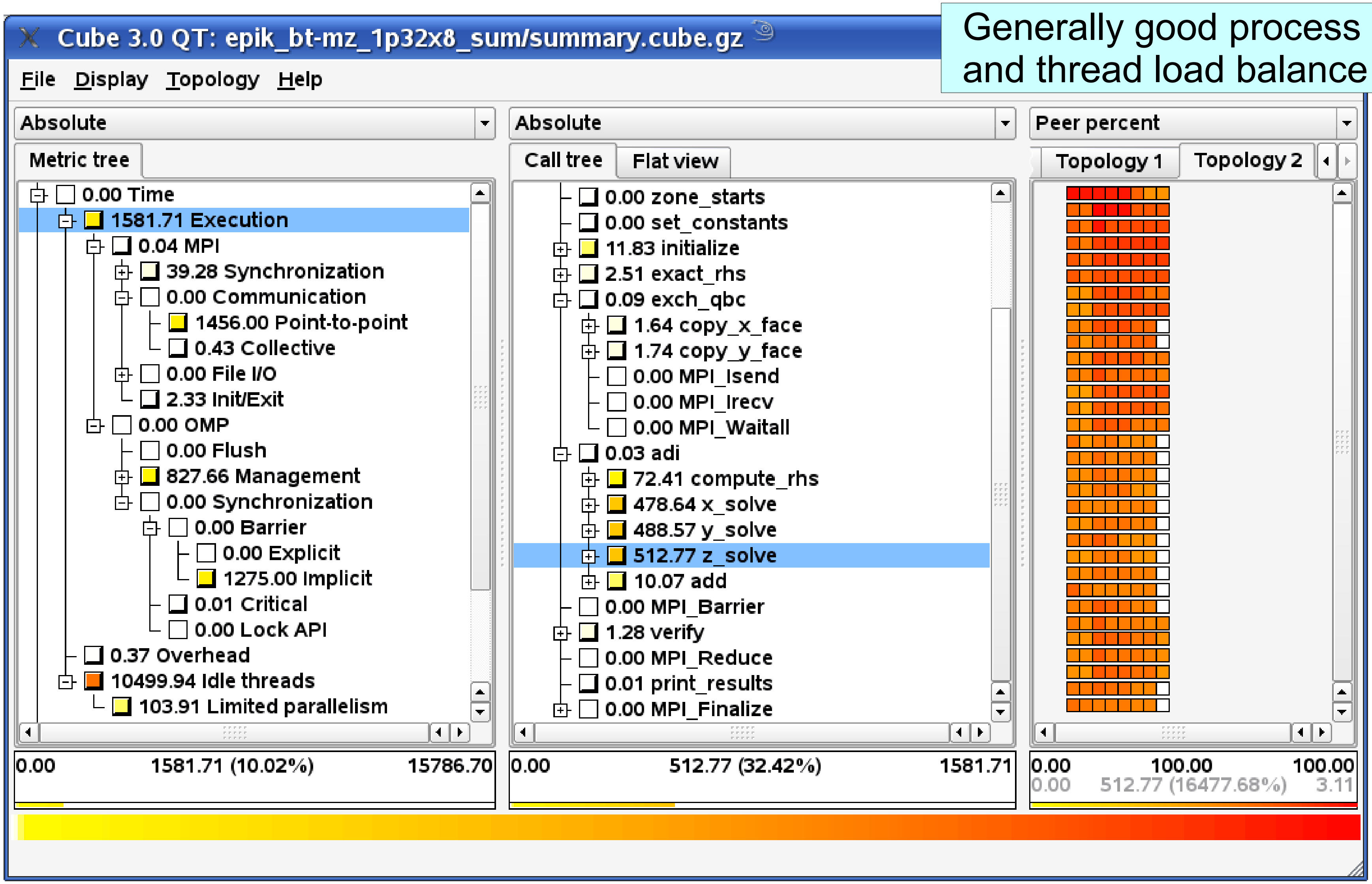
```
0.00 Time
  114.36 Execution
    0.00 MPI
      0.04 Synchronization
      0.00 Communication
        0.44 Point-to-point
          2.70 Late Sender
          0.00 Late Receiver
        0.00 Collective
      0.00 File I/O
      0.72 Init/Exit
    0.00 OMP
      0.00 Flush
      1.03 Management
      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
```

```
- 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



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

File Display Topology Help

Absolute | Absolute | Peer percent

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:

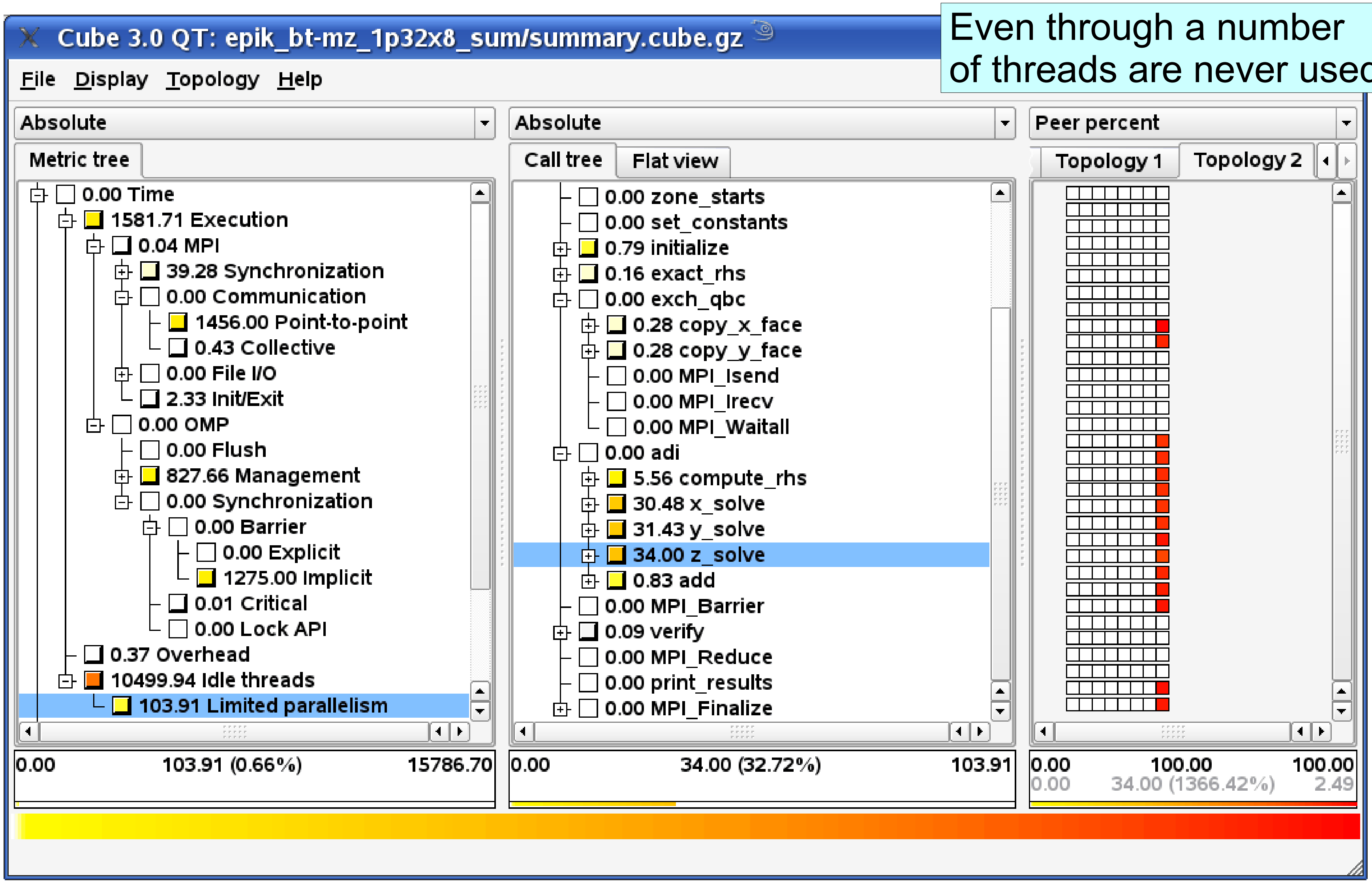
```
  0.00 zone_starts
  0.00 set_constants
  11.83 initialize
  2.51 exact_rhs
  0.09 exch_qbc
    1.64 copy_x_face
    1.74 copy_y_face
    0.00 MPI_Isend
    0.00 MPI_Irecv
    0.00 MPI_Waitall
  0.03 adi
    72.41 compute_rhs
    478.64 x_solve
    488.57 y_solve
    512.77 z_solve
    10.07 add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

Topology 1 | Topology 2

```
0.00        1581.71 (10.02%)        15786.70
0.00        512.77 (32.42%)         1581.71
0.00        100.00                  100.00
0.00        512.77 (16477.68%)      3.11
```

Generally good process and thread load balance

32x8 summary analysis: Limited parallelism



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
```

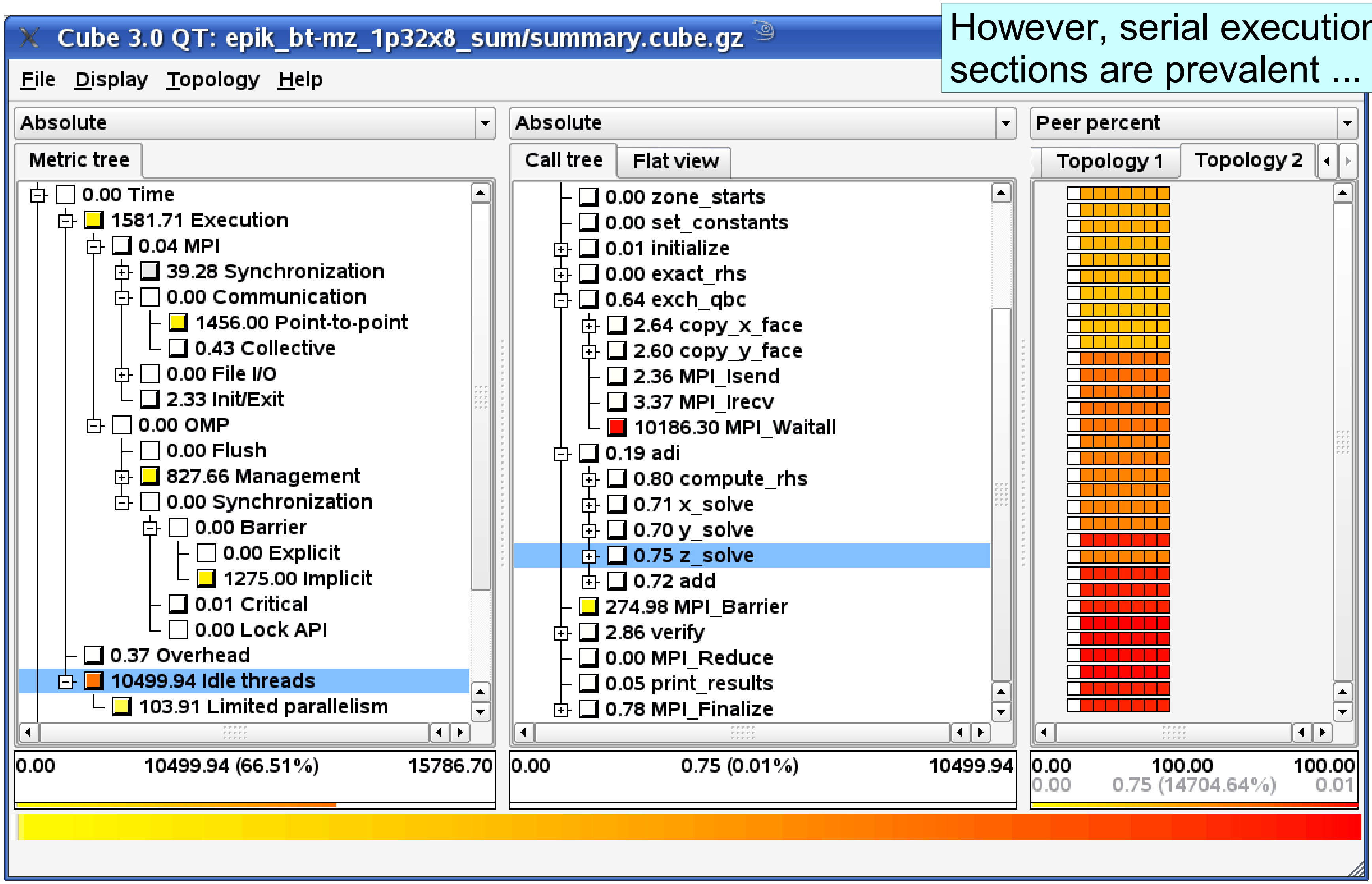
```
0.00 zone_starts
0.00 set_constants
0.79 initialize
0.16 exact_rhs
0.00 exch_qbc
  0.28 copy_x_face
  0.28 copy_y_face
  0.00 MPI_Isend
  0.00 MPI_Irecv
  0.00 MPI_Waitall
0.00 adi
  5.56 compute_rhs
  30.48 x_solve
  31.43 y_solve
  34.00 z_solve
  0.83 add
0.00 MPI_Barrier
0.09 verify
0.00 MPI_Reduce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

Even through a number of threads are never used

32x8 summary analysis: Idle threads time

However, serial execution sections are prevalent ...



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

File Display Topology Help

Absolute | Absolute | Peer percent

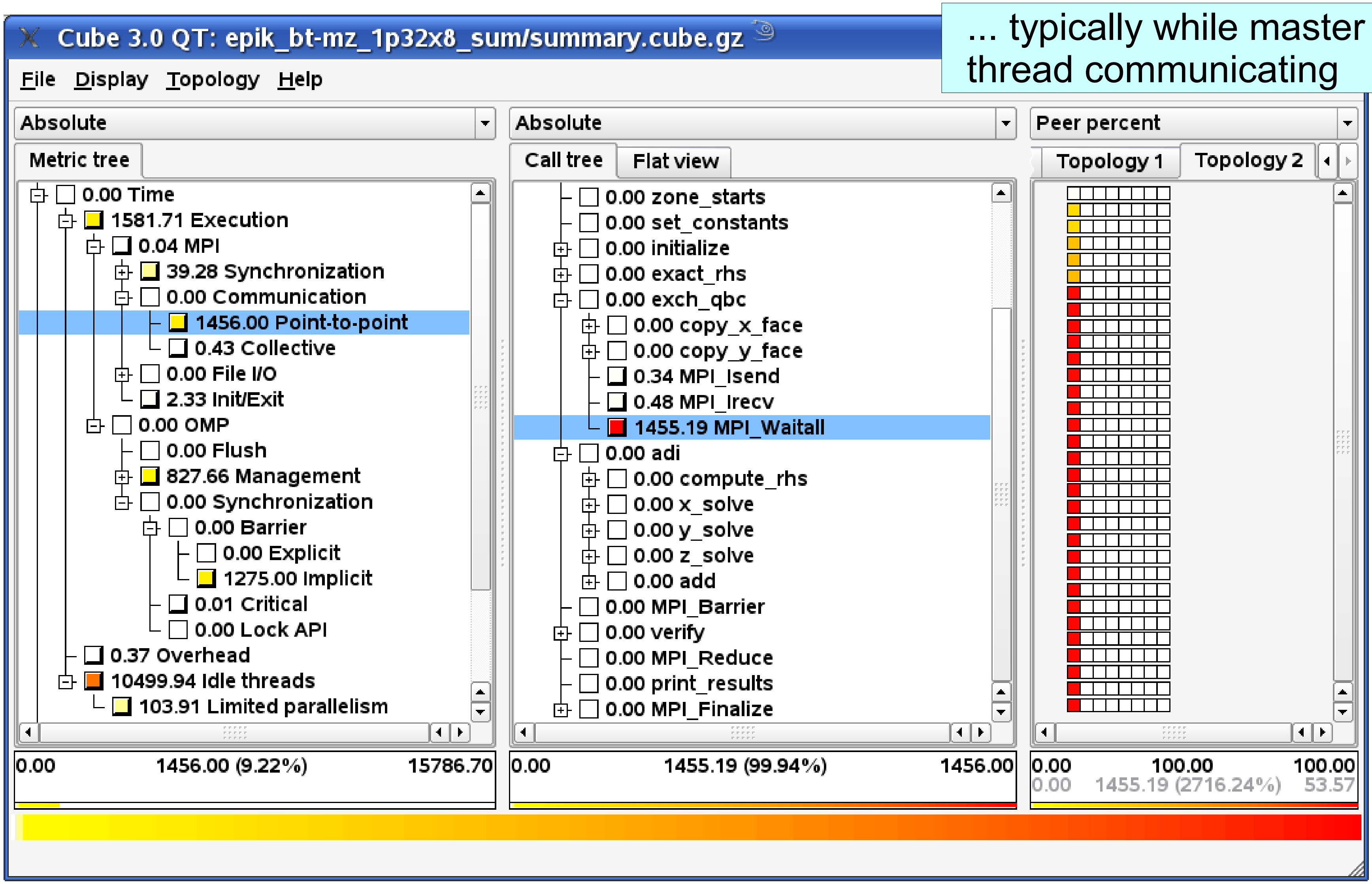
Metric tree | Call tree | Flat view | Topology 1 | Topology 2

```
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

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree | Call tree | Flat view | Topology 1 | Topology 2

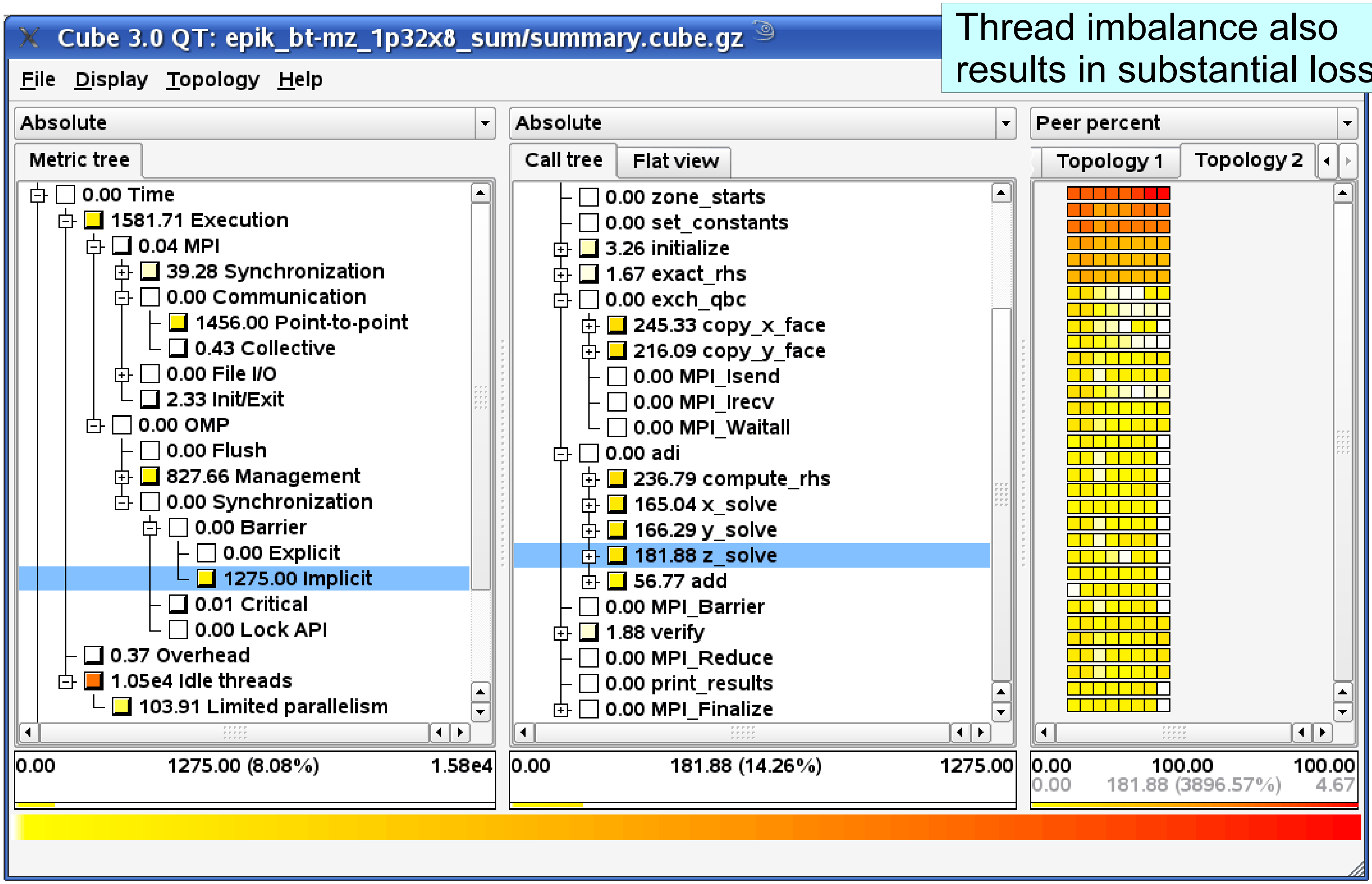
```
0.00 Time
  1581.71 Execution
    0.04 MPI
      39.28 Synchronization
      0.00 Communication
        1456.00 Point-to-point
        0.43 Collective
      0.00 File I/O
      2.33 Init/Exit
    0.00 OMP
      0.00 Flush
      827.66 Management
      0.00 Synchronization
        0.00 Barrier
          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
0.00 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
```

... typically while master thread communicating

32x8 summary analysis: Implicit barrier time



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
  1.05e4 Idle threads
    103.91 Limited parallelism
```

```
0.00          1275.00 (8.08%)          1.58e4
```

Call tree:

```
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
```

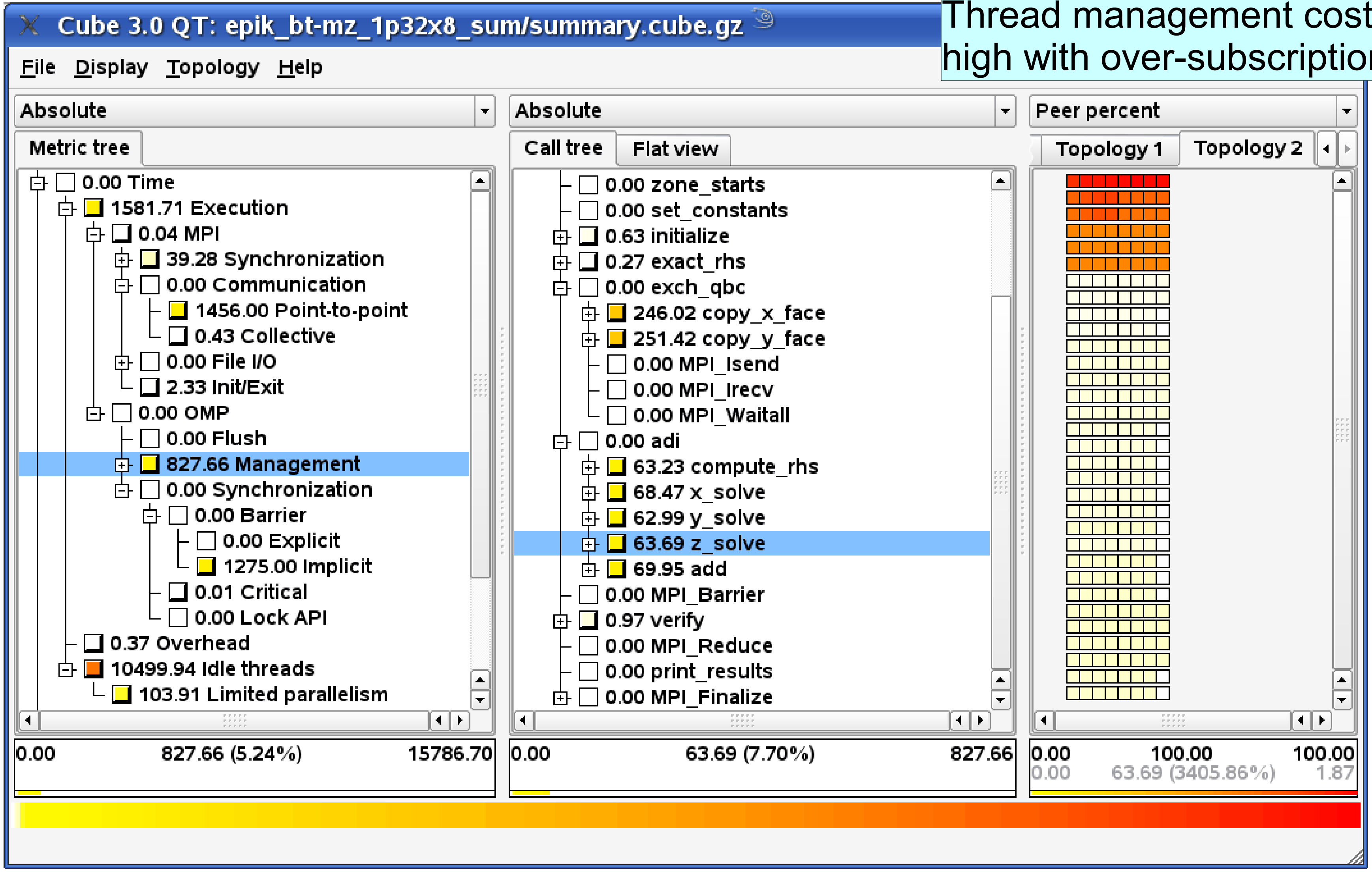
```
0.00          181.88 (14.26%)          1275.00
```

Topology:

```
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

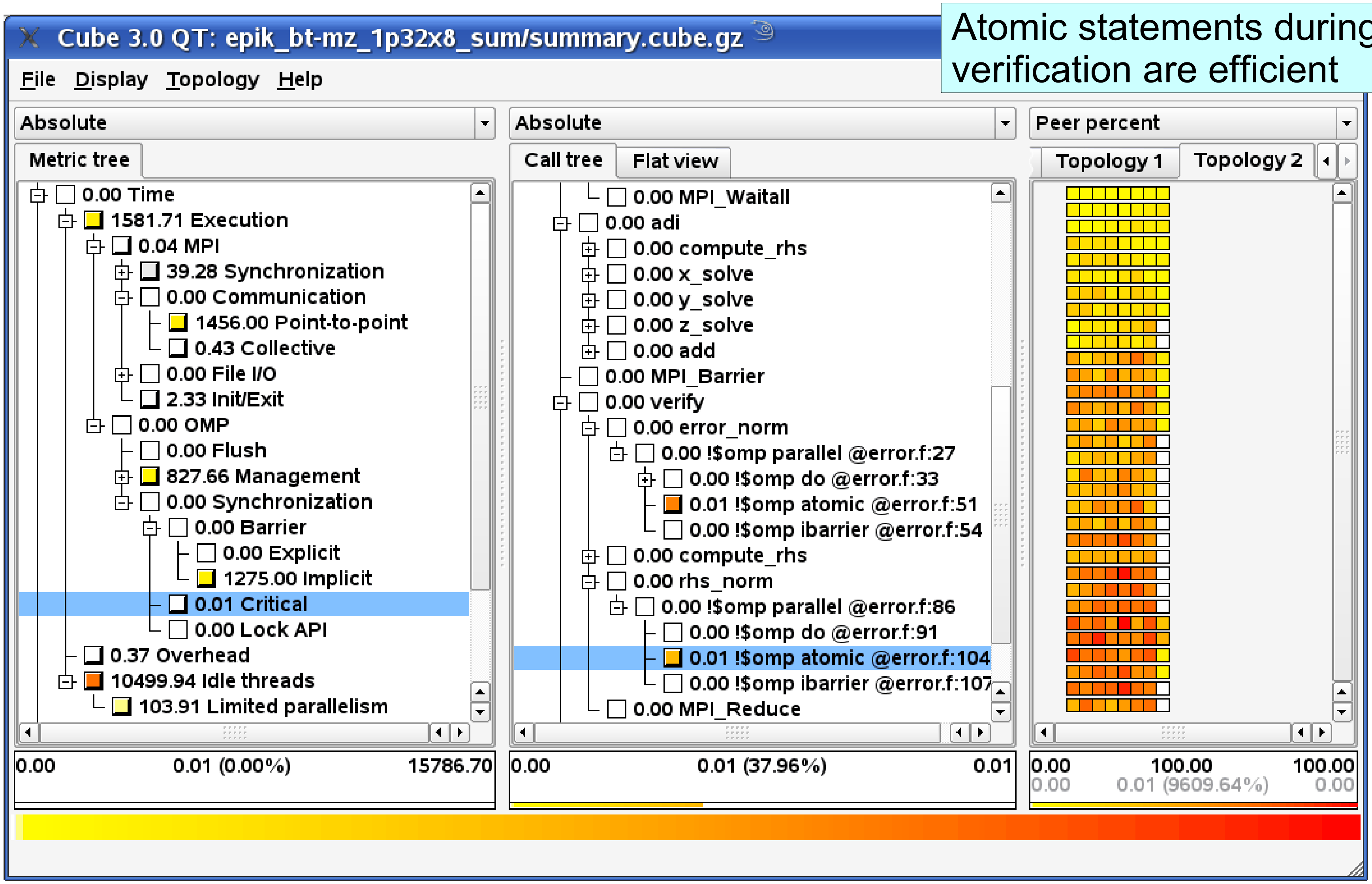
```
0.00 Time
  1581.71 Execution
    0.04 MPI
      39.28 Synchronization
      0.00 Communication
        1456.00 Point-to-point
        0.43 Collective
      0.00 File I/O
      2.33 Init/Exit
    0.00 OMP
      0.00 Flush
      827.66 Management
      0.00 Synchronization
        0.00 Barrier
          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



Atomic statements during verification are efficient

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

File Display Topology Help

Absolute | Absolute | Peer percent

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 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
```

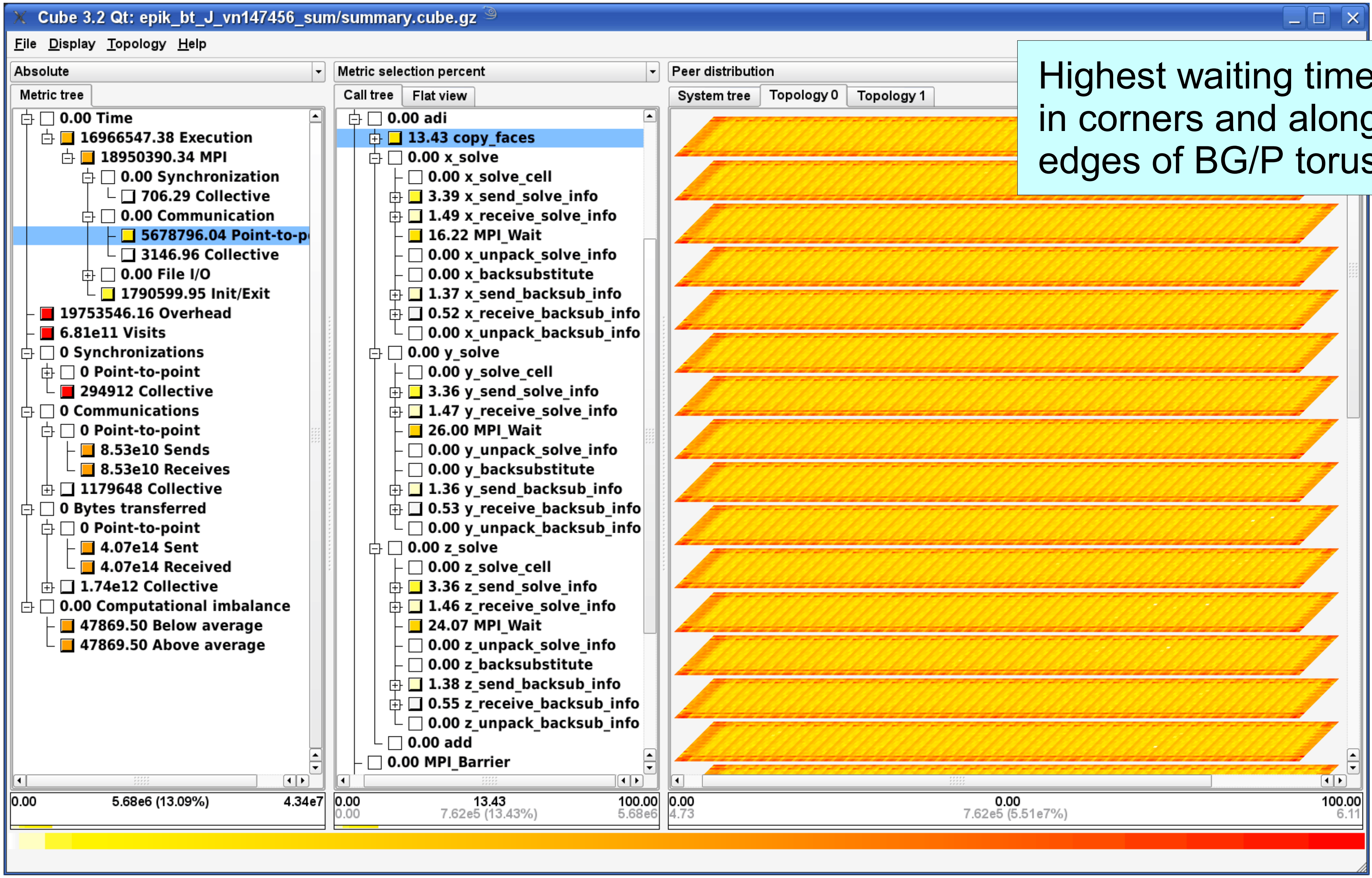
Topology 1 | Topology 2

```
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
```

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

File Display Topology Help

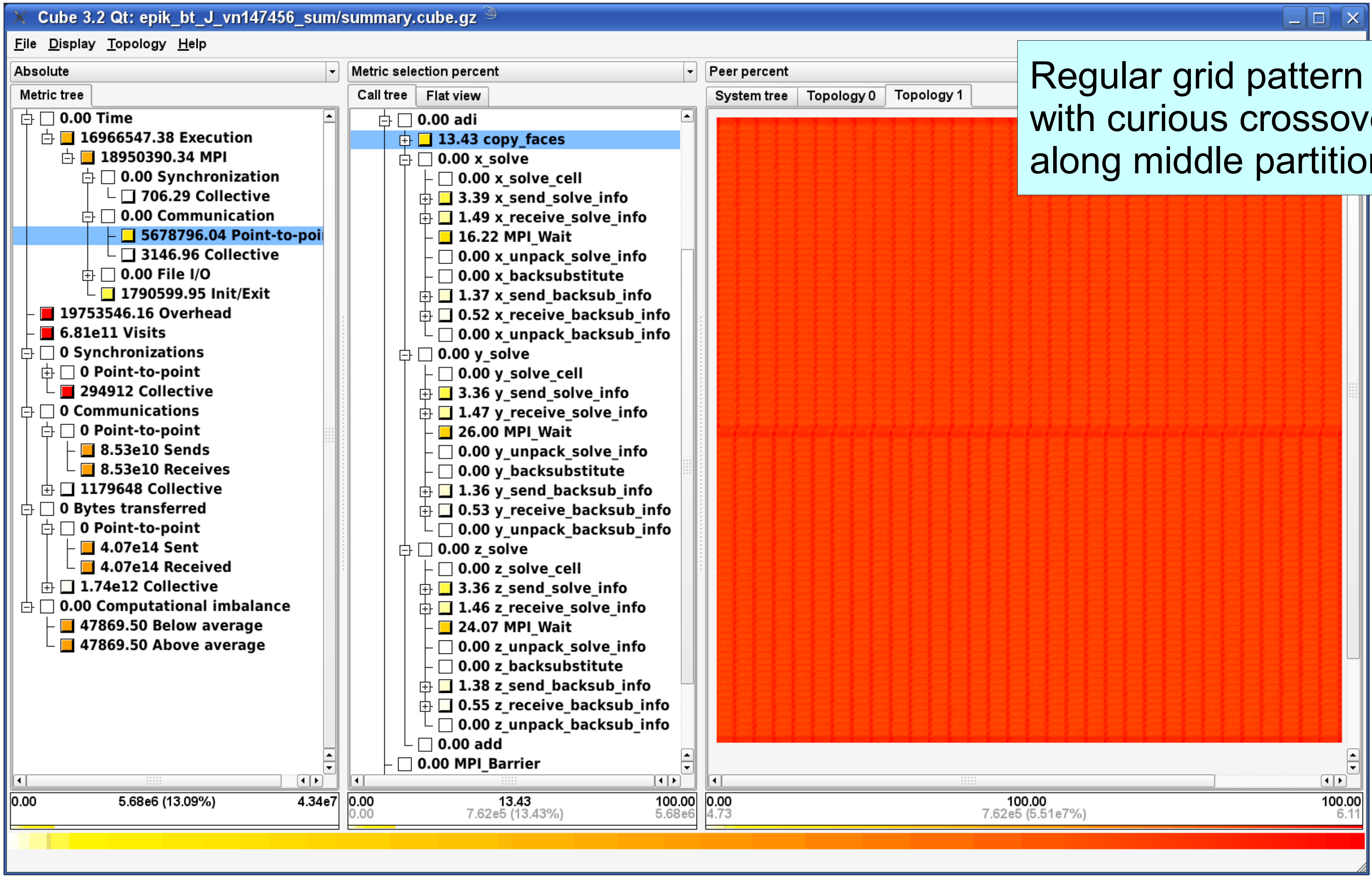
```
Absolute                              Metric selection percent              Peer distribution
Metric tree                           Call tree | Flat view                 System tree | Topology 0 | Topology 1

0.00 Time                             0.00 adi
  16966547.38 Execution                 13.43 copy_faces
    18950390.34 MPI                     0.00 x_solve
      0.00 Synchronization                0.00 x_solve_cell
        706.29 Collective                 3.39 x_send_solve_info
      0.00 Communication                  1.49 x_receive_solve_info
        5678796.04 Point-to-p             16.22 MPI_Wait
        3146.96 Collective                0.00 x_unpack_solve_info
      0.00 File I/O                       0.00 x_backsubstitute
      1790599.95 Init/Exit                1.37 x_send_backsub_info
19753546.16 Overhead                      0.52 x_receive_backsub_info
6.81e11 Visits                            0.00 x_unpack_backsub_info
0 Synchronizations                      0.00 y_solve
  0 Point-to-point                        0.00 y_solve_cell
  294912 Collective                       3.36 y_send_solve_info
0 Communications                          1.47 y_receive_solve_info
  0 Point-to-point                        26.00 MPI_Wait
    8.53e10 Sends                         0.00 y_unpack_solve_info
    8.53e10 Receives                      0.00 y_backsubstitute
  1179648 Collective                      1.36 y_send_backsub_info
0 Bytes transferred                       0.53 y_receive_backsub_info
  0 Point-to-point                        0.00 y_unpack_backsub_info
    4.07e14 Sent                        0.00 z_solve
    4.07e14 Received                      0.00 z_solve_cell
  1.74e12 Collective                      3.36 z_send_solve_info
0.00 Computational imbalance              1.46 z_receive_solve_info
  47869.50 Below average                  24.07 MPI_Wait
  47869.50 Above average                  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    0.00                100.00
                                      0.00    7.62e5 (13.43%)   5.68e6    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

File Display Topology Help

Absolute | Metric selection percent | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1

```
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
```

```
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
```

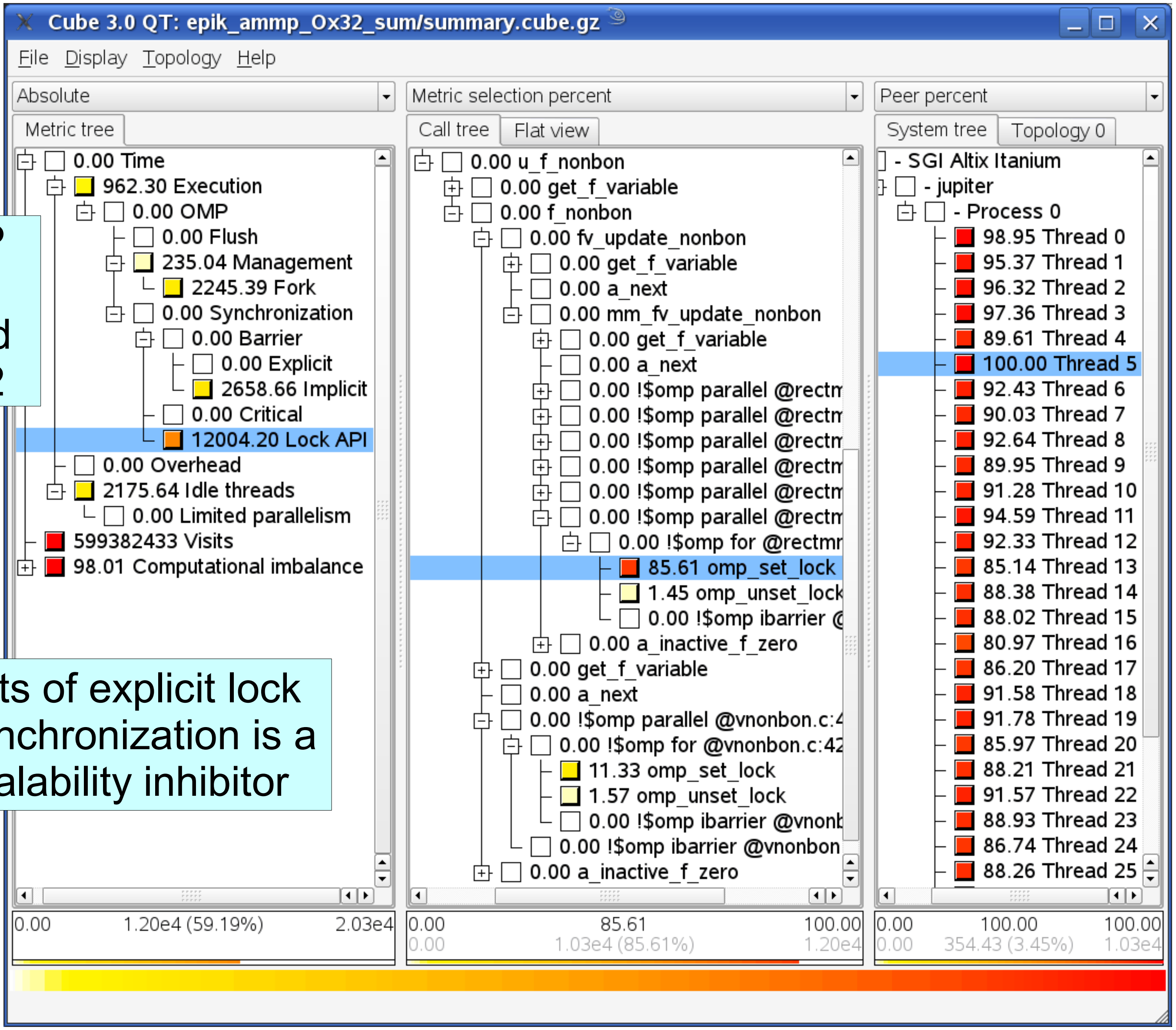
```
0.00          5.68e6 (13.09%)          4.34e7
0.00          13.43                    100.00
0.00          7.62e5 (13.43%)          5.68e6
0.00          100.00                   100.00
4.73          7.62e5 (5.51e7%)         6.11
```

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

Absolute | Metric selection percent | Peer percent

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

```
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

0.00    1.20e4 (59.19%)    2.03e4
```

```
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

0.00    85.61    100.00
0.00    1.03e4 (85.61%)    1.20e4
```

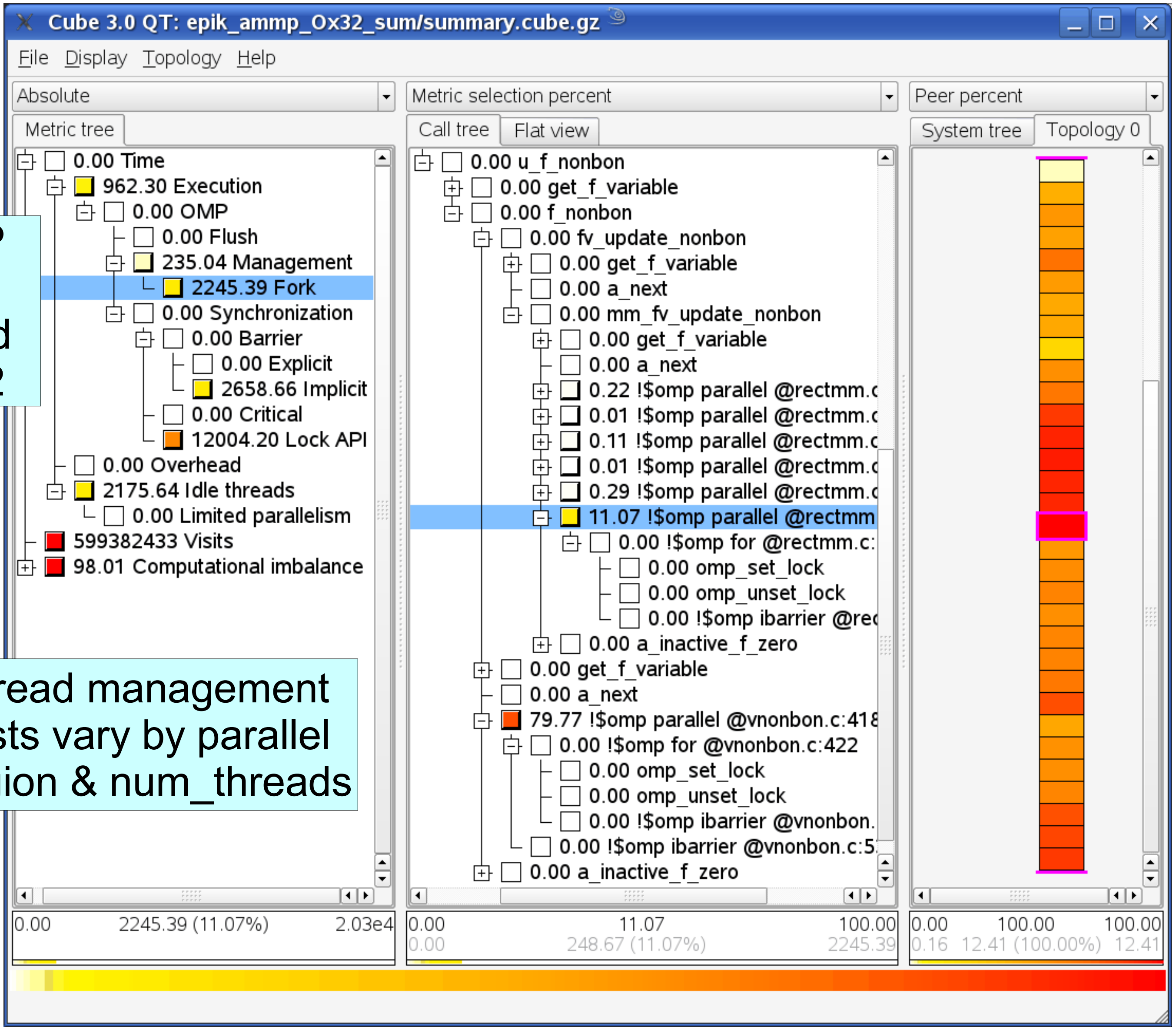
```
System tree:
- SGI Altix Itanium
  - jupiter
    - Process 0
      98.95 Thread 0
      95.37 Thread 1
      96.32 Thread 2
      97.36 Thread 3
      89.61 Thread 4
      100.00 Thread 5
      92.43 Thread 6
      90.03 Thread 7
      92.64 Thread 8
      89.95 Thread 9
      91.28 Thread 10
      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    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



Cube 3.0 QT: epik_ammp_Ox32_sum/summary.cube.gz

```
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

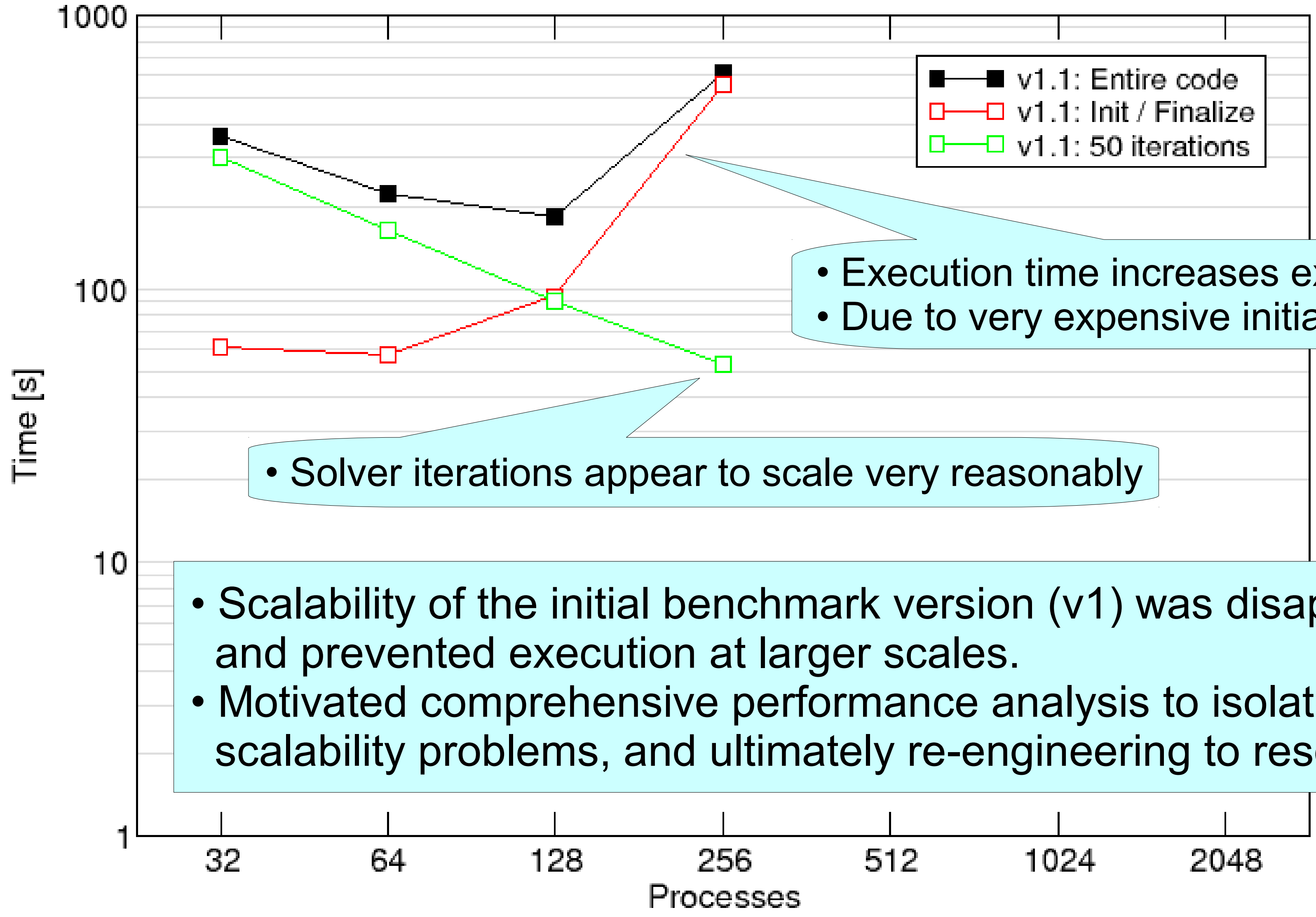
Thread management costs vary by parallel region & num_threads

GemsFDTD case study

- Computational electromagnetics solver
  - originates from KTH General ElectroMagnetics Solvers project
  - finite-difference time-domain method for Maxwell equations
- MPI parallel versions in SPEC MPI2007 benchmark suite
  - original ***v1.1*** (113.GemsFDTD) “medium” size
  - revised ***v2.0*** (145.lGemsFDTD) “large” size
  - built with PGI 9.0.4 Fortran90 compiler (21k lines of code)
    - typical benchmark optimization: `-fastsse -O3 -Mipa=fast,inline`
- Both run on 'hector' Cray XT4 at EPCC
  - using “ltrain” dataset from v2.0 benchmark (50 timesteps)
  - default Scalasca instrumentation for measurements
    - 9 of 90 application user-level source routines specified in filter determined by scoring initial summary experiment

Proc. 10th Int'l Workshop PARA (Reykjavík, Iceland, 2010)

GemsFDTD v1 scalability on Cray XT4



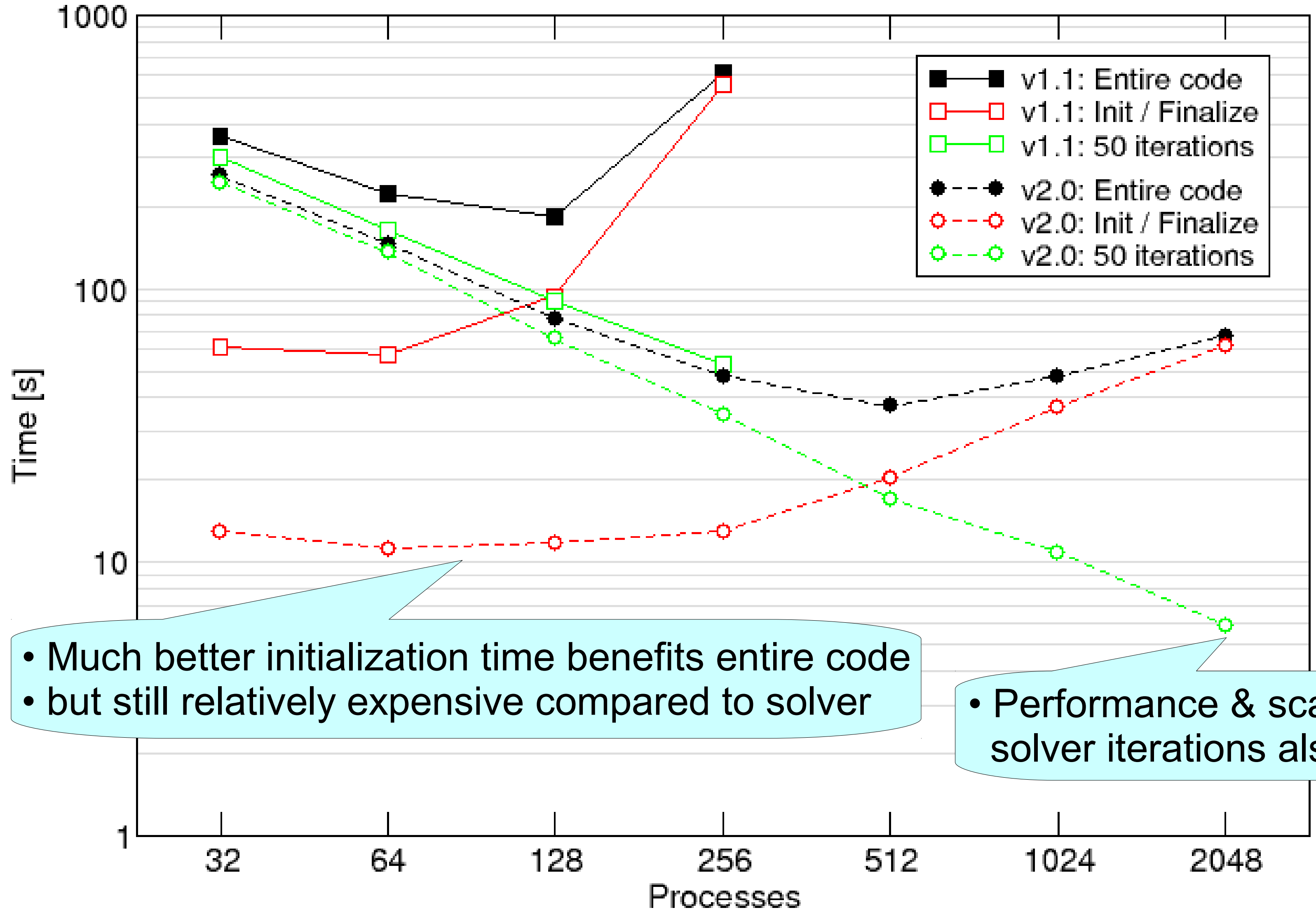
['ltrain' runs on CrayXT4 HECToR]

- Execution time increases exponentially
- Due to very expensive initialization

- Solver iterations appear to scale very reasonably

- Scalability of the initial benchmark version (v1) was disappointing and prevented execution at larger scales.
- Motivated comprehensive performance analysis to isolate scalability problems, and ultimately re-engineering to resolve them.

GemsFDTD v1 & v2 scalability on Cray XT4

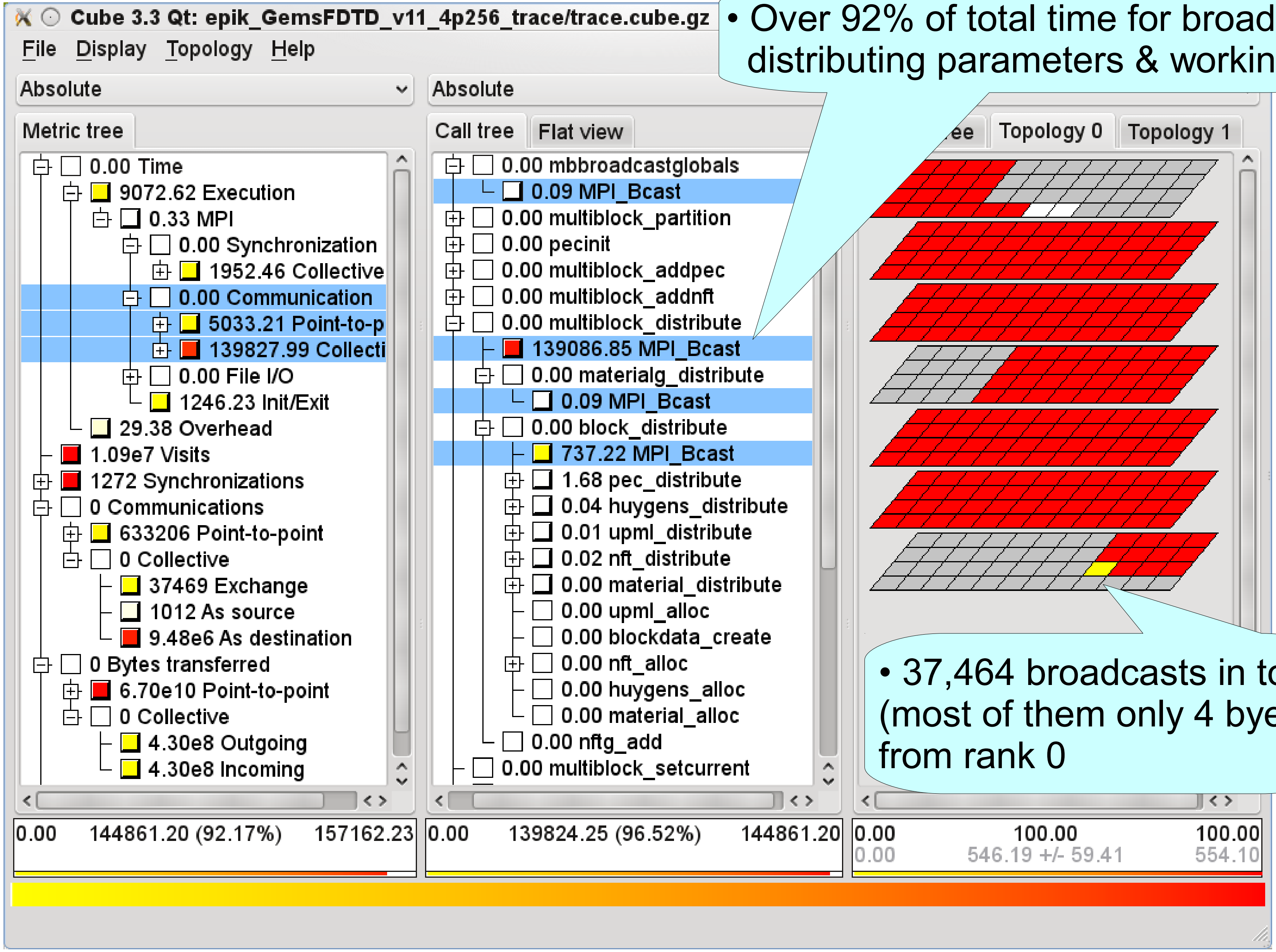


['ltrain' runs on CrayXT4 HECToR]

- Much better initialization time benefits entire code
- but still relatively expensive compared to solver

- Performance & scalability of solver iterations also improved

Time for initialization broadcasts (*v1.1*)



Cube 3.3 Qt: epik_GemsFDTD_v11_4p256_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute

Metric tree:

```
0.00 Time
  9072.62 Execution
    0.33 MPI
      0.00 Synchronization
        1952.46 Collective
      0.00 Communication
        5033.21 Point-to-p
        139827.99 Collecti
      0.00 File I/O
      1246.23 Init/Exit
    29.38 Overhead
1.09e7 Visits
1272 Synchronizations
0 Communications
  633206 Point-to-point
  0 Collective
    37469 Exchange
    1012 As source
    9.48e6 As destination
0 Bytes transferred
  6.70e10 Point-to-point
  0 Collective
    4.30e8 Outgoing
    4.30e8 Incoming
```

Call tree | Flat view:

```
0.00 mbbroadcastglobals
  0.09 MPI_Bcast
0.00 multiblock_partition
0.00 pecinit
0.00 multiblock_addpec
0.00 multiblock_addnft
0.00 multiblock_distribute
  139086.85 MPI_Bcast
  0.00 materialg_distribute
    0.09 MPI_Bcast
  0.00 block_distribute
    737.22 MPI_Bcast
    1.68 pec_distribute
    0.04 huygens_distribute
    0.01 upml_distribute
    0.02 nft_distribute
    0.00 material_distribute
    0.00 upml_alloc
    0.00 blockdata_create
    0.00 nft_alloc
    0.00 huygens_alloc
    0.00 material_alloc
  0.00 nftg_add
0.00 multiblock_setcurrent
```

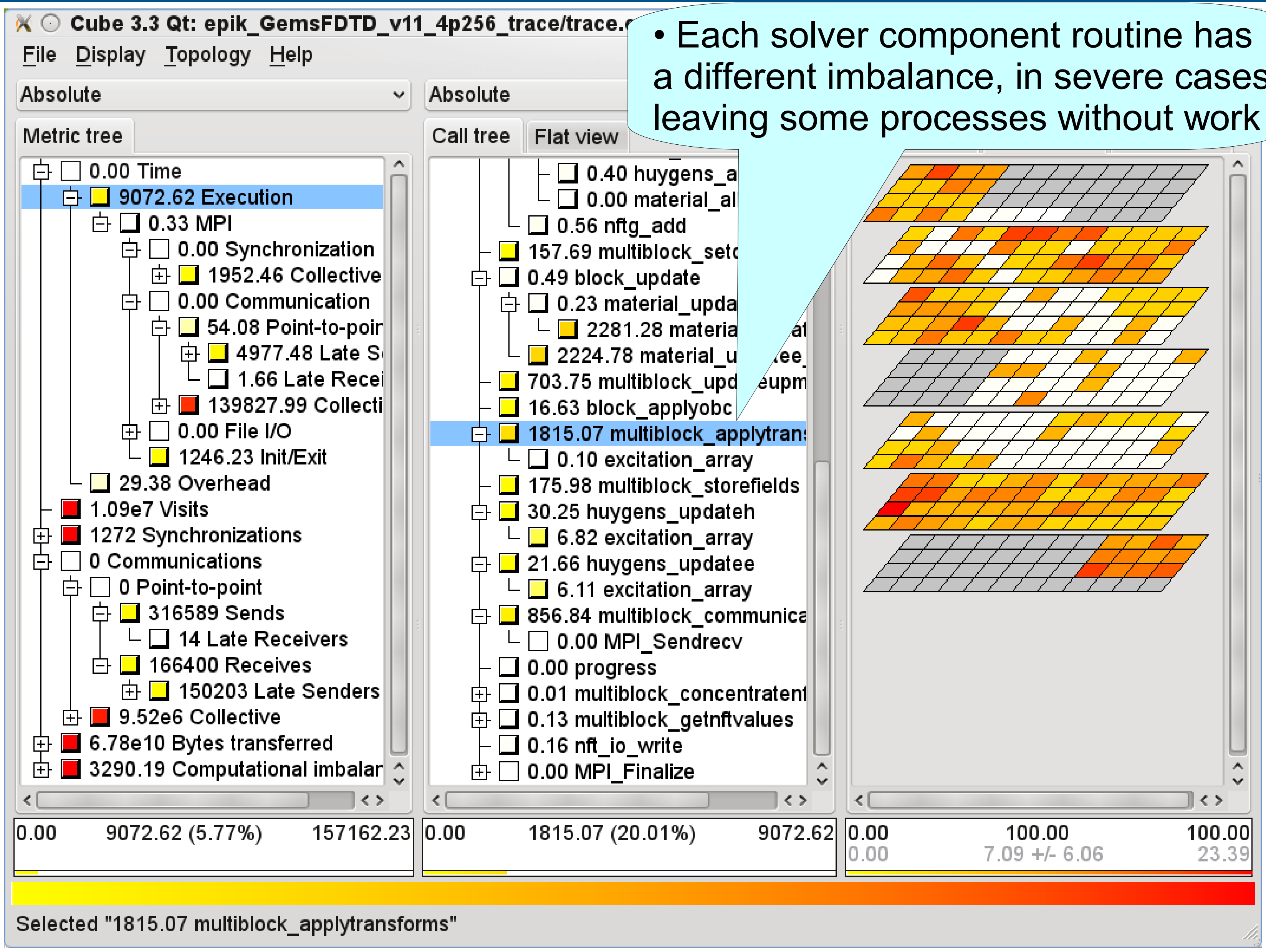
Topology 0 | Topology 1

```
0.00     144861.20 (92.17%)     157162.23
0.00     139824.25 (96.52%)     144861.20
0.00     100.00                 100.00
0.00     546.19 +/- 59.41       554.10
```

- Over 92% of total time for broadcasts distributing parameters & working set

- 37,464 broadcasts in total (most of them only 4 byes) from rank 0

Computation time in solver transforms (*v1.1*)



- Each solver component routine has a different imbalance, in severe cases leaving some processes without work

```
Cube 3.3 Qt: epik_GemsFDTD_v11_4p256_trace/trace.c
File  Display  Topology  Help
Absolute                         Absolute
Metric tree                      Call tree   Flat view
```

```
0.00 Time
  9072.62 Execution
    0.33 MPI
      0.00 Synchronization
        1952.46 Collective
      0.00 Communication
        54.08 Point-to-poir
          4977.48 Late S
          1.66 Late Recei
        139827.99 Collecti
      0.00 File I/O
      1246.23 Init/Exit
  29.38 Overhead
1.09e7 Visits
1272 Synchronizations
0 Communications
  0 Point-to-point
    316589 Sends
      14 Late Receivers
    166400 Receives
      150203 Late Senders
  9.52e6 Collective
6.78e10 Bytes transferred
3290.19 Computational imbalar
```

```
      0.40 huygens_a
      0.00 material_all
    0.56 nftg_add
  157.69 multiblock_setc
  0.49 block_update
    0.23 material_upda
      2281.28 materia
    2224.78 material_u
  703.75 multiblock_upd
  16.63 block_applyobc
  1815.07 multiblock_applytrans
    0.10 excitation_array
  175.98 multiblock_storefields
  30.25 huygens_updateh
    6.82 excitation_array
  21.66 huygens_updatee
    6.11 excitation_array
  856.84 multiblock_communica
    0.00 MPI_Sendrecv
  0.00 progress
  0.01 multiblock_concentratenf
  0.13 multiblock_getnftvalues
  0.16 nft_io_write
  0.00 MPI_Finalize
```

```
0.00   9072.62 (5.77%)   157162.23 | 0.00   1815.07 (20.01%)   9072.62 | 0.00   100.00        100.00
                                   |                                   | 0.00   7.09 +/- 6.06  23.39
```

Selected "1815.07 multiblock_applytransforms"

GemsFDTD case study

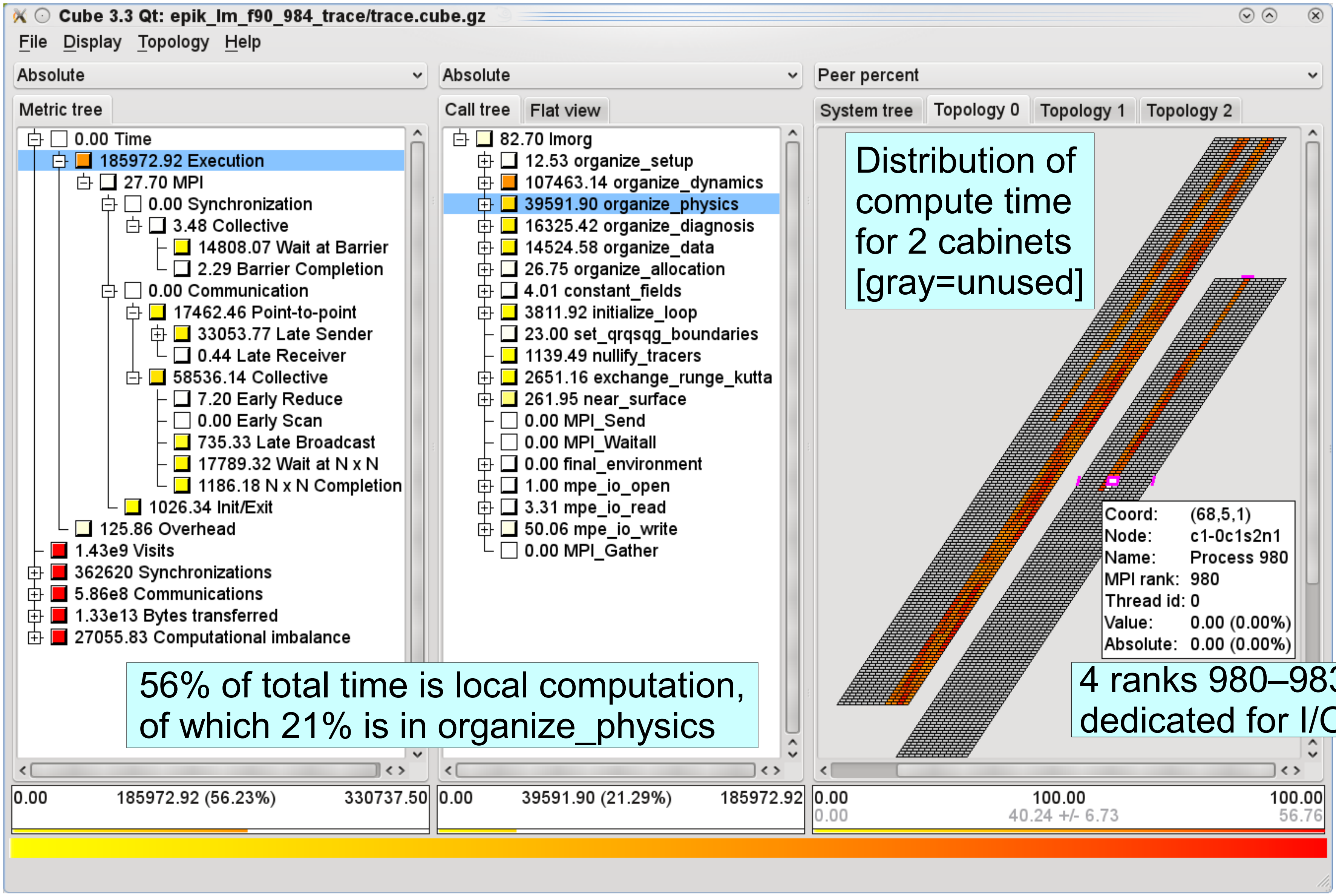
- Initialization originally dominated by numerous broadcasts and expensive serial multiblock partition by rank 0
  - Re-engineered implementation of scalable partition routine, aggregation of multiple data values into larger messages, and postpones allocations until all block information in broadcast
    - Initialization time reduced to less than 2% of total time
- Solver iterations using blocking communication manifests as *Late Sender* waiting originating from imbalance in local computation time (due to different computations)
  - Re-engineered implementation uses non-blocking comms and re-uses communication pattern used to exchange blocks (as well as 2 of 256 processes unintentionally idled throughout)
    - computation & communication time both improved more than 25%
- Scalabilty improved from 128 processes to more than 1024

COSMO-7/XE6 case study

- Regional climate and weather model
  - developed by Consortium for Small-scale Modeling (COSMO)
    - DWD, MeteoSwiss and others
  - non-hydrostatic limited-area atmospheric model (6.6km grid)
- MPI parallel version 4.12 (Jan-2011)
  - built with PGI 10.9 Fortran90 compiler (222k lines of code)
- MeteoSwiss operational 24-hour forecast of 06-Dec-2010
  - Western Europe 393x338x60 resolution, 1440 timesteps
- Run with 984 processes on 'palu' Cray XE6 at CSCS
  - 28x35 compute grid + 4 dedicated I/O processes
  - used 41 Opteron compute nodes each with 24 cores
  - Scalasca trace measurement with 19 of 178 routines filtered
  - 44GB trace written in 23s and analyzed in 82s

Courtesy of Oliver Fuhrer (MeteoSwiss) & CSCS

COSMO/XE6 physics computation time



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree

```
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
        3.48 Collective
          14808.07 Wait at Barrier
          2.29 Barrier Completion
      0.00 Communication
        17462.46 Point-to-point
          33053.77 Late Sender
          0.44 Late Receiver
        58536.14 Collective
          7.20 Early Reduce
          0.00 Early Scan
          735.33 Late Broadcast
          17789.32 Wait at N x N
          1186.18 N x N Completion
      1026.34 Init/Exit
    125.86 Overhead
1.43e9 Visits
362620 Synchronizations
5.86e8 Communications
1.33e13 Bytes transferred
27055.83 Computational imbalance
```

Call tree | Flat view

```
82.70 lmorg
  12.53 organize_setup
  107463.14 organize_dynamics
  39591.90 organize_physics
  16325.42 organize_diagnosis
  14524.58 organize_data
  26.75 organize_allocation
  4.01 constant_fields
  3811.92 initialize_loop
  23.00 set_qrqsqg_boundaries
  1139.49 nullify_tracers
  2651.16 exchange_runge_kutta
  261.95 near_surface
  0.00 MPI_Send
  0.00 MPI_Waitall
  0.00 final_environment
  1.00 mpe_io_open
  3.31 mpe_io_read
  50.06 mpe_io_write
  0.00 MPI_Gather
```

System tree | Topology 0 | Topology 1 | Topology 2

Distribution of compute time for 2 cabinets [gray=unused]

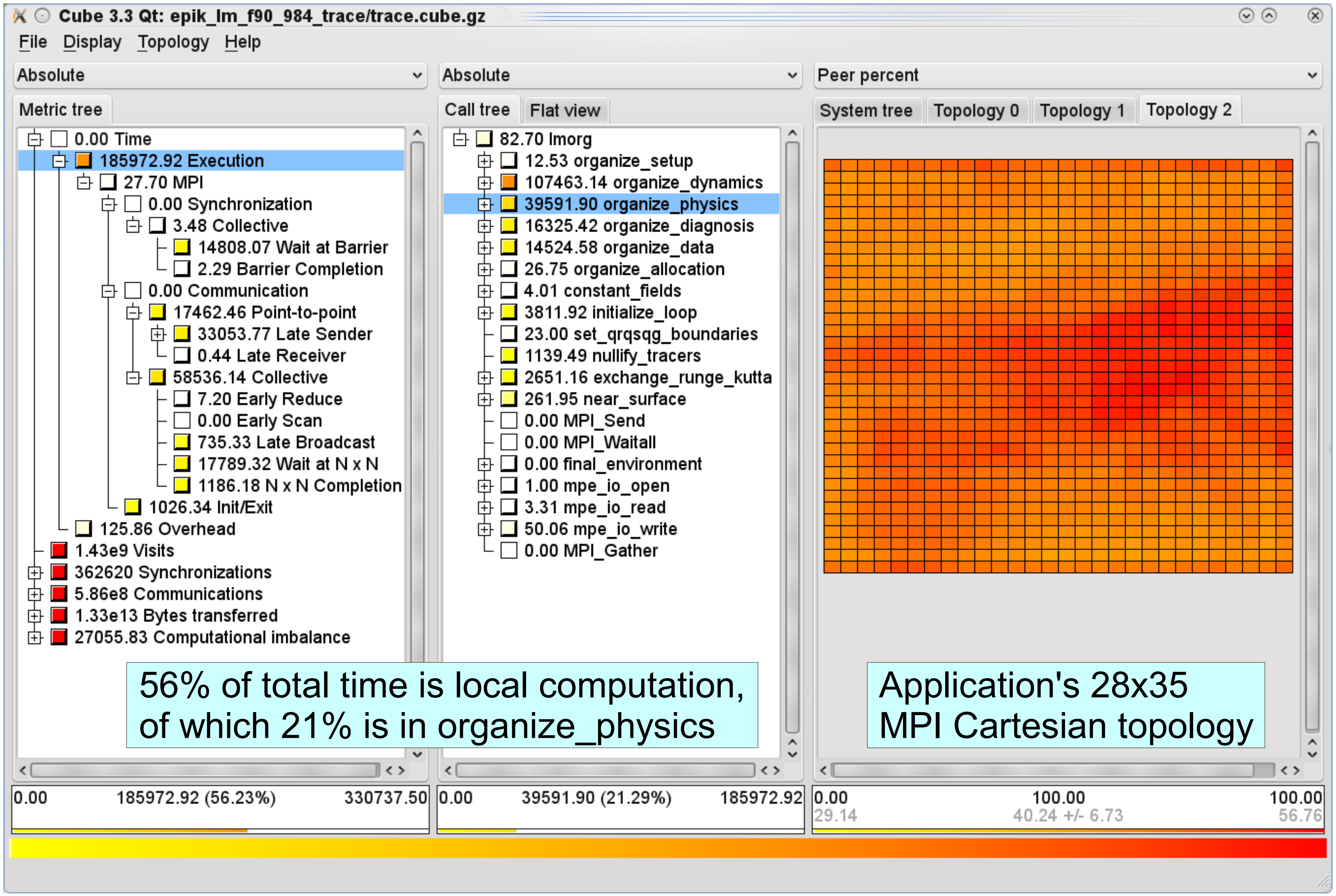
```
Coord:     (68,5,1)
Node:      c1-0c1s2n1
Name:      Process 980
MPI rank:  980
Thread id: 0
Value:     0.00 (0.00%)
Absolute:  0.00 (0.00%)
```

4 ranks 980–983 dedicated for I/O

56% of total time is local computation, of which 21% is in organize_physics

```
0.00   185972.92 (56.23%)   330737.50 | 0.00   39591.90 (21.29%)   185972.92 | 0.00   100.00          100.00
                                      |                                      | 0.00   40.24 +/- 6.73  56.76
```

COSMO/XE6 physics computation time



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree

```
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
        3.48 Collective
          14808.07 Wait at Barrier
          2.29 Barrier Completion
      0.00 Communication
        17462.46 Point-to-point
          33053.77 Late Sender
          0.44 Late Receiver
        58536.14 Collective
          7.20 Early Reduce
          0.00 Early Scan
          735.33 Late Broadcast
          17789.32 Wait at N x N
          1186.18 N x N Completion
      1026.34 Init/Exit
  125.86 Overhead
1.43e9 Visits
362620 Synchronizations
5.86e8 Communications
1.33e13 Bytes transferred
27055.83 Computational imbalance
```

Call tree | Flat view

```
82.70 lmorg
  12.53 organize_setup
  107463.14 organize_dynamics
  39591.90 organize_physics
  16325.42 organize_diagnosis
  14524.58 organize_data
  26.75 organize_allocation
  4.01 constant_fields
  3811.92 initialize_loop
  23.00 set_qrqsqg_boundaries
  1139.49 nullify_tracers
  2651.16 exchange_runge_kutta
  261.95 near_surface
  0.00 MPI_Send
  0.00 MPI_Waitall
  0.00 final_environment
  1.00 mpe_io_open
  3.31 mpe_io_read
  50.06 mpe_io_write
  0.00 MPI_Gather
```

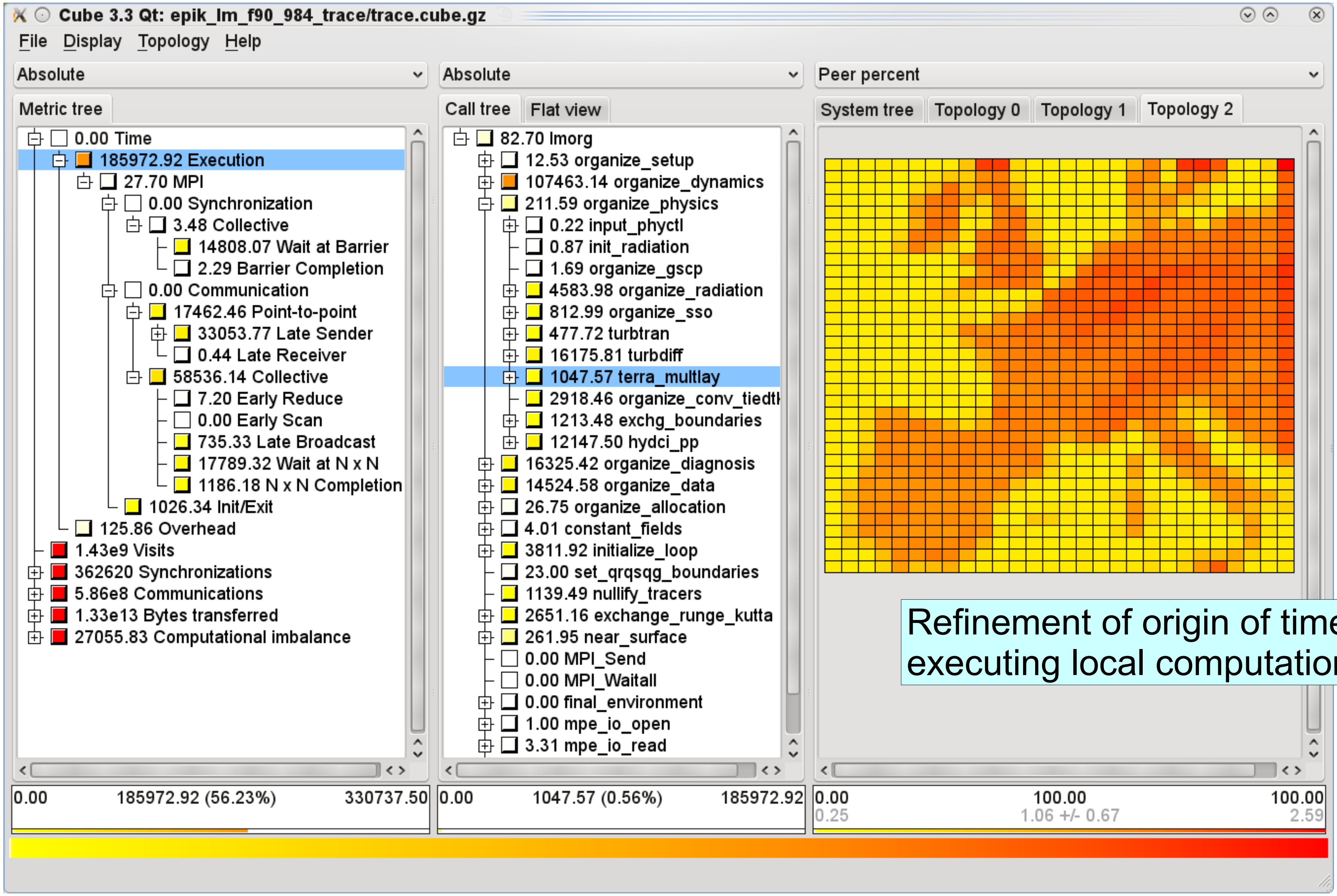
System tree | Topology 0 | Topology 1 | Topology 2

```
0.00     185972.92 (56.23%)     330737.50
0.00     39591.90 (21.29%)      185972.92
0.00          100.00              100.00
29.14      40.24 +/- 6.73          56.76
```

56% of total time is local computation, of which 21% is in organize_physics

Application's 28x35 MPI Cartesian topology

COSMO/XE6 physics computation imbalance


Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1 | Topology 2

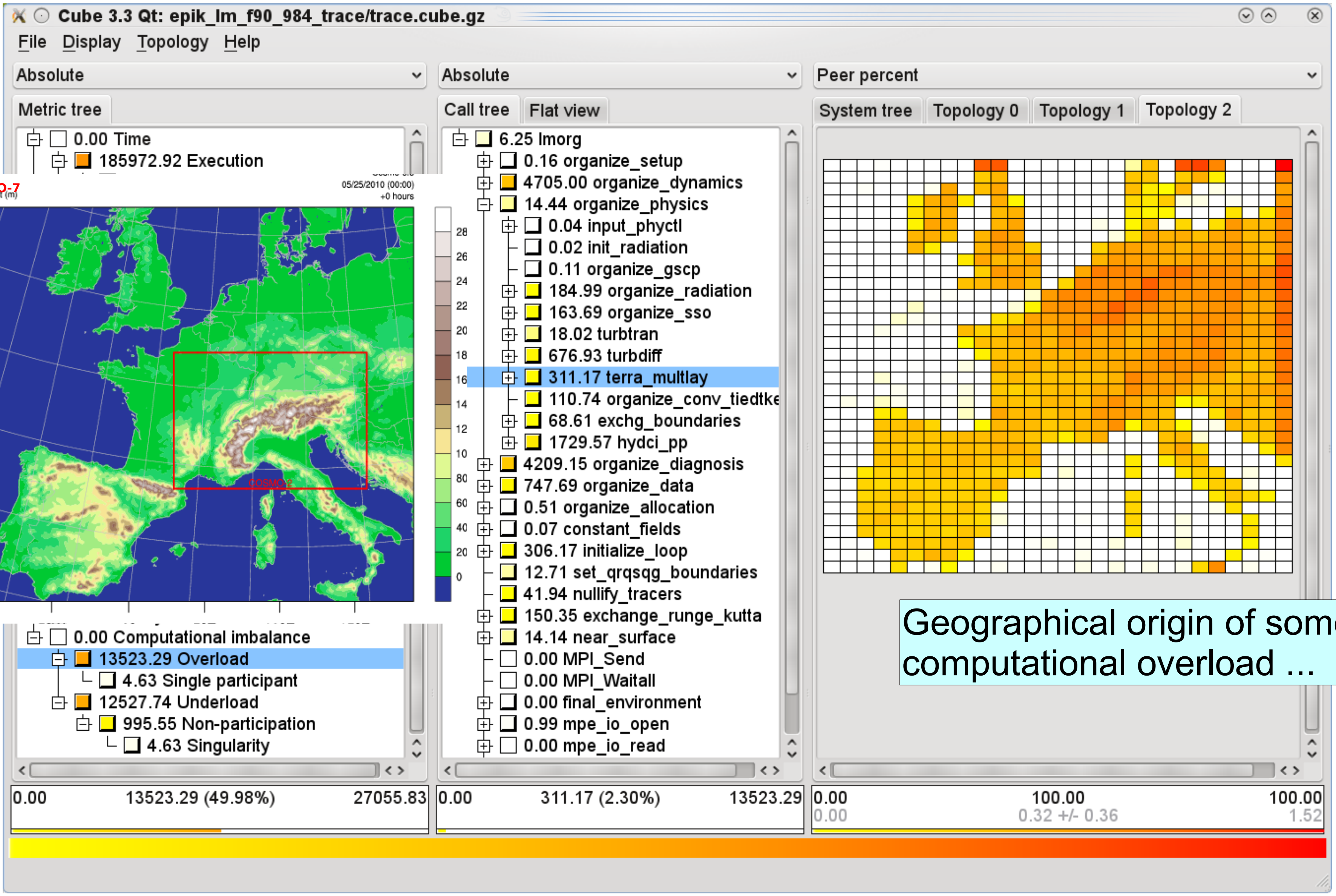
```
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
        3.48 Collective
          14808.07 Wait at Barrier
          2.29 Barrier Completion
      0.00 Communication
        17462.46 Point-to-point
          33053.77 Late Sender
          0.44 Late Receiver
        58536.14 Collective
          7.20 Early Reduce
          0.00 Early Scan
          735.33 Late Broadcast
          17789.32 Wait at N x N
          1186.18 N x N Completion
      1026.34 Init/Exit
  125.86 Overhead
1.43e9 Visits
362620 Synchronizations
5.86e8 Communications
1.33e13 Bytes transferred
27055.83 Computational imbalance
```

```
82.70 lmorg
  12.53 organize_setup
  107463.14 organize_dynamics
  211.59 organize_physics
    0.22 input_phyctl
    0.87 init_radiation
    1.69 organize_gscp
    4583.98 organize_radiation
    812.99 organize_sso
    477.72 turbtran
    16175.81 turbdiff
    1047.57 terra_multlay
    2918.46 organize_conv_tiedtk
    1213.48 exchg_boundaries
    12147.50 hydci_pp
  16325.42 organize_diagnosis
  14524.58 organize_data
  26.75 organize_allocation
  4.01 constant_fields
  3811.92 initialize_loop
  23.00 set_qrqsqg_boundaries
  1139.49 nullify_tracers
  2651.16 exchange_runge_kutta
  261.95 near_surface
  0.00 MPI_Send
  0.00 MPI_Waitall
  0.00 final_environment
  1.00 mpe_io_open
  3.31 mpe_io_read
```

```
0.00   185972.92 (56.23%)   330737.50
0.00   1047.57 (0.56%)   185972.92
0.00   100.00   100.00
0.25   1.06 +/- 0.67   2.59
```

Refinement of origin of time executing local computation

COSMO/XE6 computational overload (geo)



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

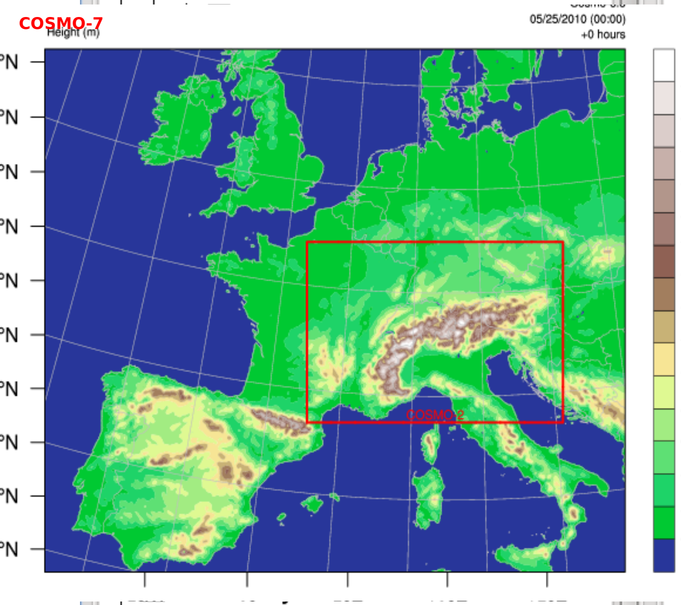
Absolute | Absolute | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1 | Topology 2

```
0.00 Time
  185972.92 Execution
...
0.00 Computational imbalance
  13523.29 Overload
    4.63 Single participant
  12527.74 Underload
    995.55 Non-participation
      4.63 Singularity
```

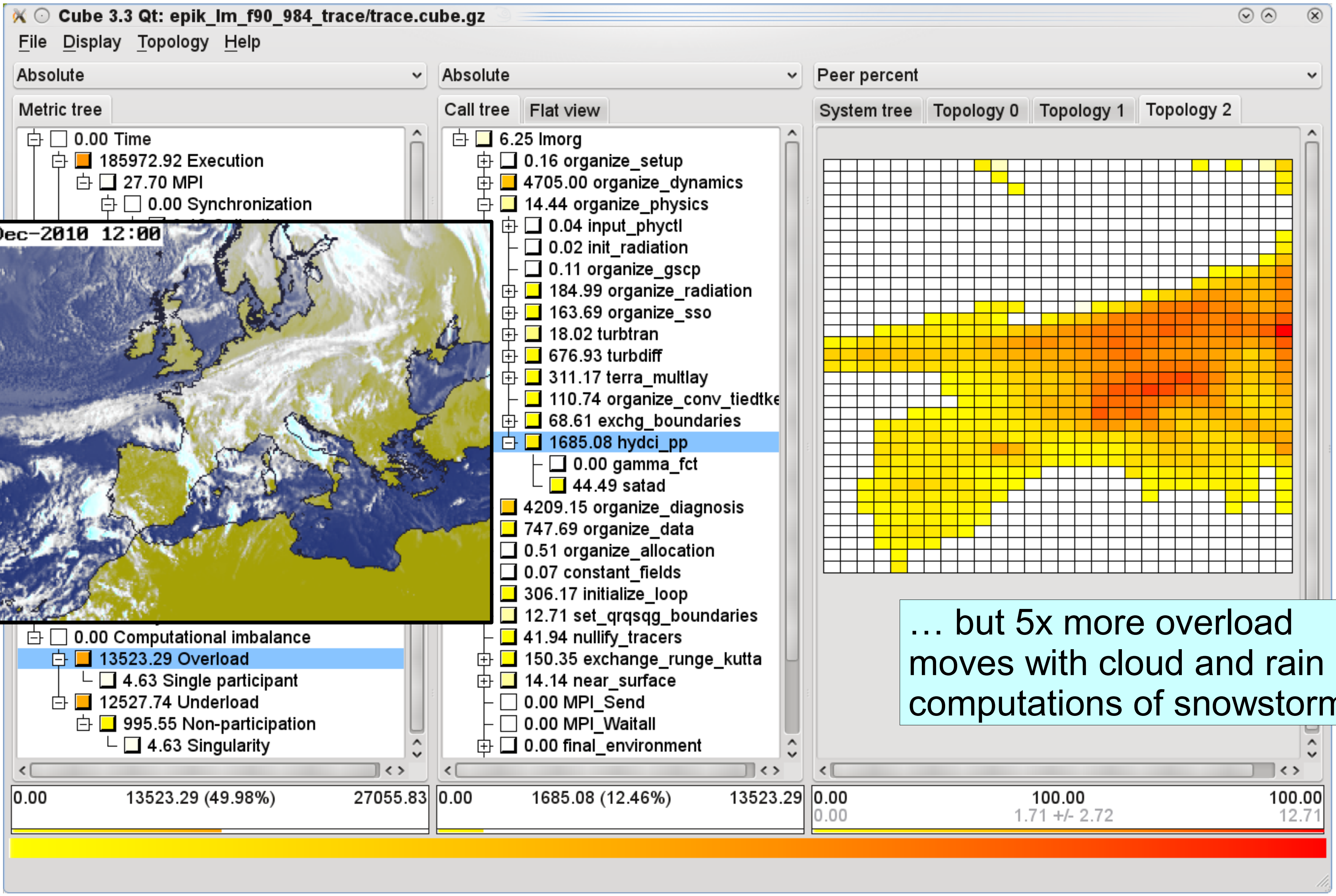
```
6.25 lmorg
  0.16 organize_setup
  4705.00 organize_dynamics
  14.44 organize_physics
    0.04 input_phyctl
    0.02 init_radiation
    0.11 organize_gscp
    184.99 organize_radiation
    163.69 organize_sso
    18.02 turbtran
    676.93 turbdiff
    311.17 terra_multlay
    110.74 organize_conv_tiedtke
    68.61 exchg_boundaries
    1729.57 hydci_pp
  4209.15 organize_diagnosis
  747.69 organize_data
  0.51 organize_allocation
  0.07 constant_fields
  306.17 initialize_loop
  12.71 set_qrqsqg_boundaries
  41.94 nullify_tracers
  150.35 exchange_runge_kutta
  14.14 near_surface
  0.00 MPI_Send
  0.00 MPI_Waitall
  0.00 final_environment
  0.99 mpe_io_open
  0.00 mpe_io_read
```

0.00	13523.29 (49.98%)	27055.83
0.00	311.17 (2.30%)	13523.29
0.00 / 0.00	100.00 / 0.32 +/- 0.36	100.00 / 1.52



Geographical origin of some computational overload ...

COSMO/XE6 computational overload (hydro)



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

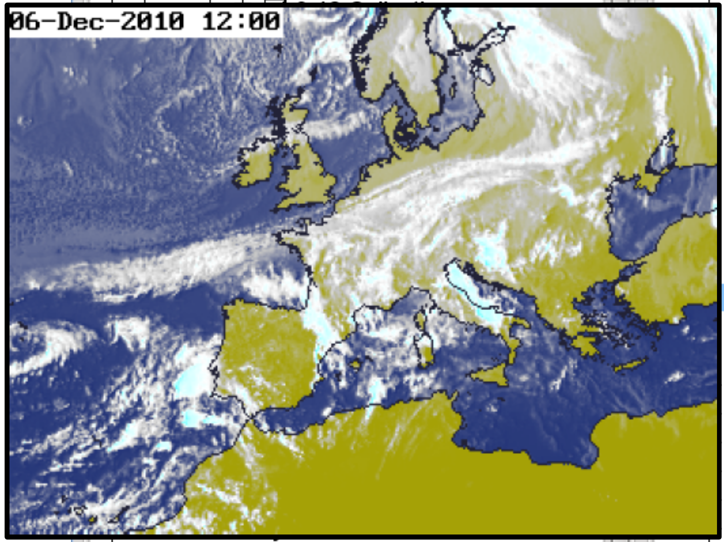
Absolute | Absolute | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1 | Topology 2

```
Metric tree
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
...
  0.00 Computational imbalance
    13523.29 Overload
      4.63 Single participant
    12527.74 Underload
      995.55 Non-participation
        4.63 Singularity
```

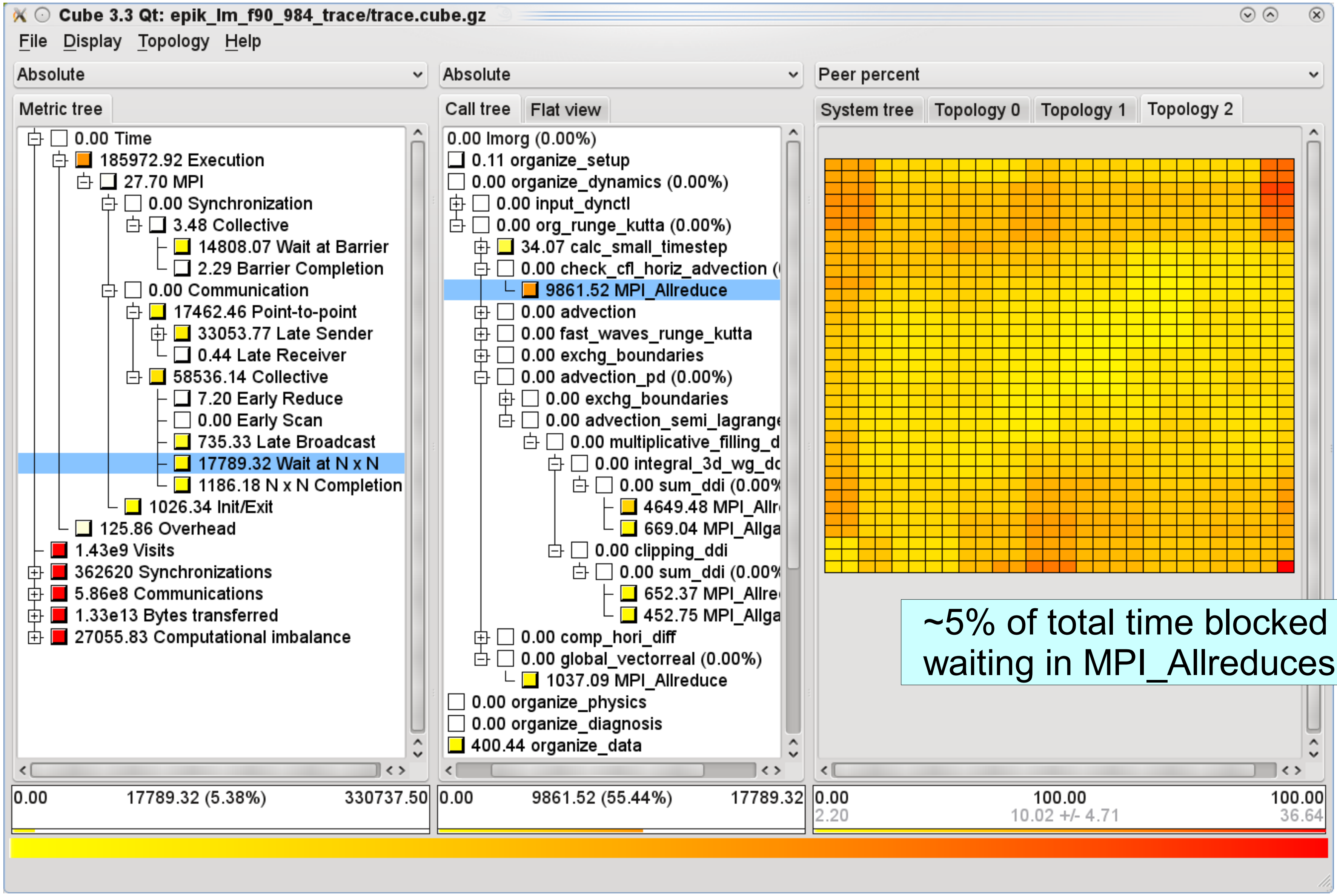
```
Call tree
6.25 lmorg
  0.16 organize_setup
  4705.00 organize_dynamics
  14.44 organize_physics
    0.04 input_phyctl
    0.02 init_radiation
    0.11 organize_gscp
    184.99 organize_radiation
    163.69 organize_sso
    18.02 turbtran
    676.93 turbdiff
    311.17 terra_multlay
    110.74 organize_conv_tiedtke
    68.61 exchg_boundaries
    1685.08 hydci_pp
      0.00 gamma_fct
      44.49 satad
  4209.15 organize_diagnosis
  747.69 organize_data
  0.51 organize_allocation
  0.07 constant_fields
  306.17 initialize_loop
  12.71 set_qrqsqg_boundaries
  41.94 nullify_tracers
  150.35 exchange_runge_kutta
  14.14 near_surface
  0.00 MPI_Send
  0.00 MPI_Waitall
  0.00 final_environment
```

```
0.00    13523.29 (49.98%)    27055.83
0.00    1685.08 (12.46%)     13523.29
0.00    100.00               100.00
0.00    1.71 +/- 2.72        12.71
```



… but 5x more overload moves with cloud and rain computations of snowstorm

COSMO/XE6 collective wait at N x N time



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree

```
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
        3.48 Collective
          14808.07 Wait at Barrier
          2.29 Barrier Completion
      0.00 Communication
        17462.46 Point-to-point
          33053.77 Late Sender
          0.44 Late Receiver
        58536.14 Collective
          7.20 Early Reduce
          0.00 Early Scan
          735.33 Late Broadcast
          17789.32 Wait at N x N
          1186.18 N x N Completion
      1026.34 Init/Exit
  125.86 Overhead
1.43e9 Visits
362620 Synchronizations
5.86e8 Communications
1.33e13 Bytes transferred
27055.83 Computational imbalance
```

Call tree | Flat view

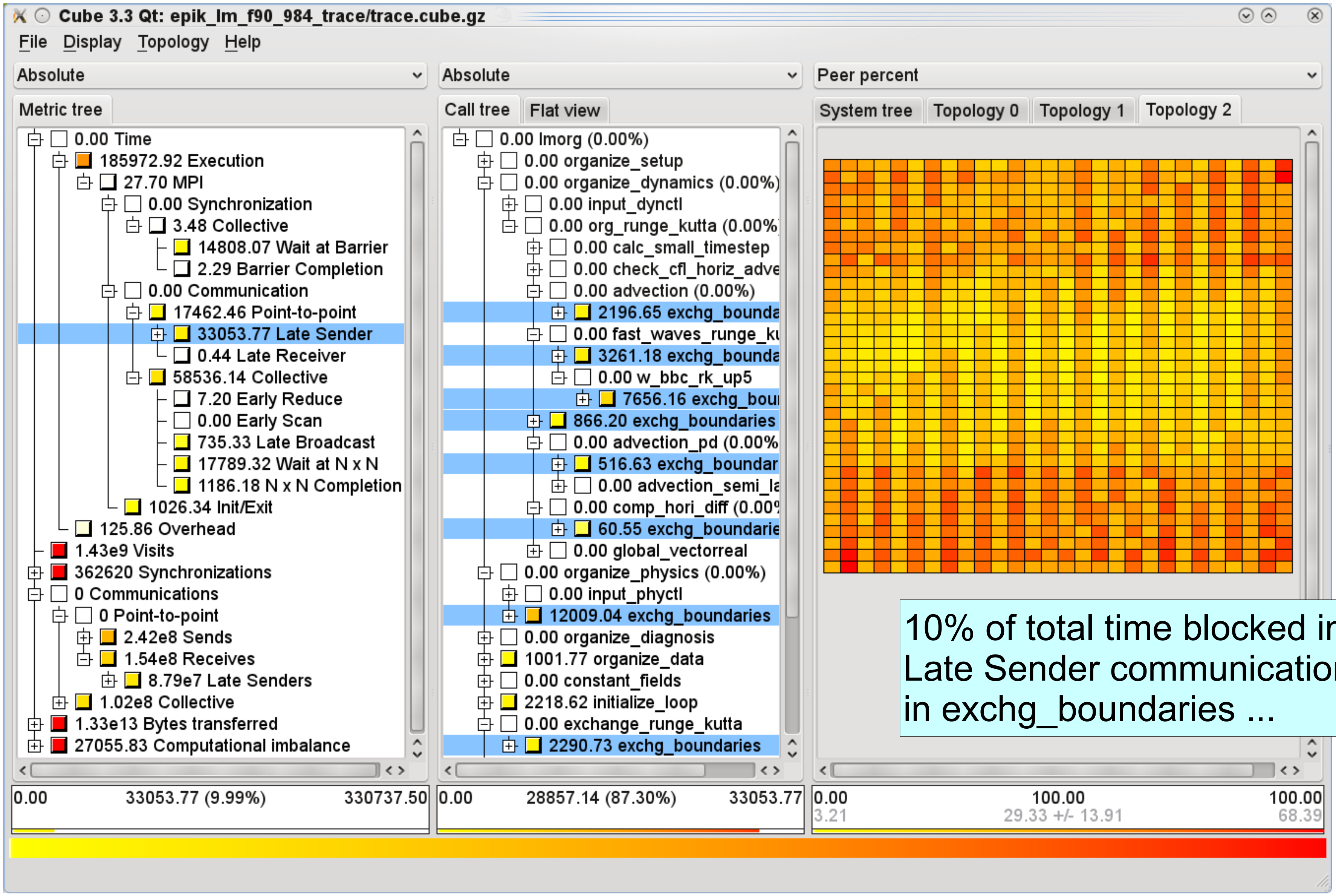
```
0.00 lmorg (0.00%)
0.11 organize_setup
0.00 organize_dynamics (0.00%)
  0.00 input_dynctl
  0.00 org_runge_kutta (0.00%)
    34.07 calc_small_timestep
    0.00 check_cfl_horiz_advection (
      9861.52 MPI_Allreduce
    0.00 advection
    0.00 fast_waves_runge_kutta
    0.00 exchg_boundaries
    0.00 advection_pd (0.00%)
      0.00 exchg_boundaries
      0.00 advection_semi_lagrange
        0.00 multiplicative_filling_d
          0.00 integral_3d_wg_dc
            0.00 sum_ddi (0.00%
              4649.48 MPI_Allr
              669.04 MPI_Allga
          0.00 clipping_ddi
            0.00 sum_ddi (0.00%
              652.37 MPI_Allre
              452.75 MPI_Allga
    0.00 comp_hori_diff
    0.00 global_vectorreal (0.00%)
      1037.09 MPI_Allreduce
0.00 organize_physics
0.00 organize_diagnosis
400.44 organize_data
```

System tree | Topology 0 | Topology 1 | Topology 2

```
0.00    17789.32 (5.38%)    330737.50
0.00    9861.52 (55.44%)    17789.32
0.00          100.00           100.00
2.20     10.02 +/- 4.71         36.64
```

~5% of total time blocked waiting in MPI_Allreduces

COSMO/XE6 late sender waiting time



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree

```
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
        3.48 Collective
          14808.07 Wait at Barrier
          2.29 Barrier Completion
      0.00 Communication
        17462.46 Point-to-point
          33053.77 Late Sender
          0.44 Late Receiver
        58536.14 Collective
          7.20 Early Reduce
          0.00 Early Scan
          735.33 Late Broadcast
          17789.32 Wait at N x N
          1186.18 N x N Completion
      1026.34 Init/Exit
    125.86 Overhead
1.43e9 Visits
362620 Synchronizations
0 Communications
  0 Point-to-point
    2.42e8 Sends
    1.54e8 Receives
      8.79e7 Late Senders
  1.02e8 Collective
1.33e13 Bytes transferred
27055.83 Computational imbalance
```

Call tree | Flat view

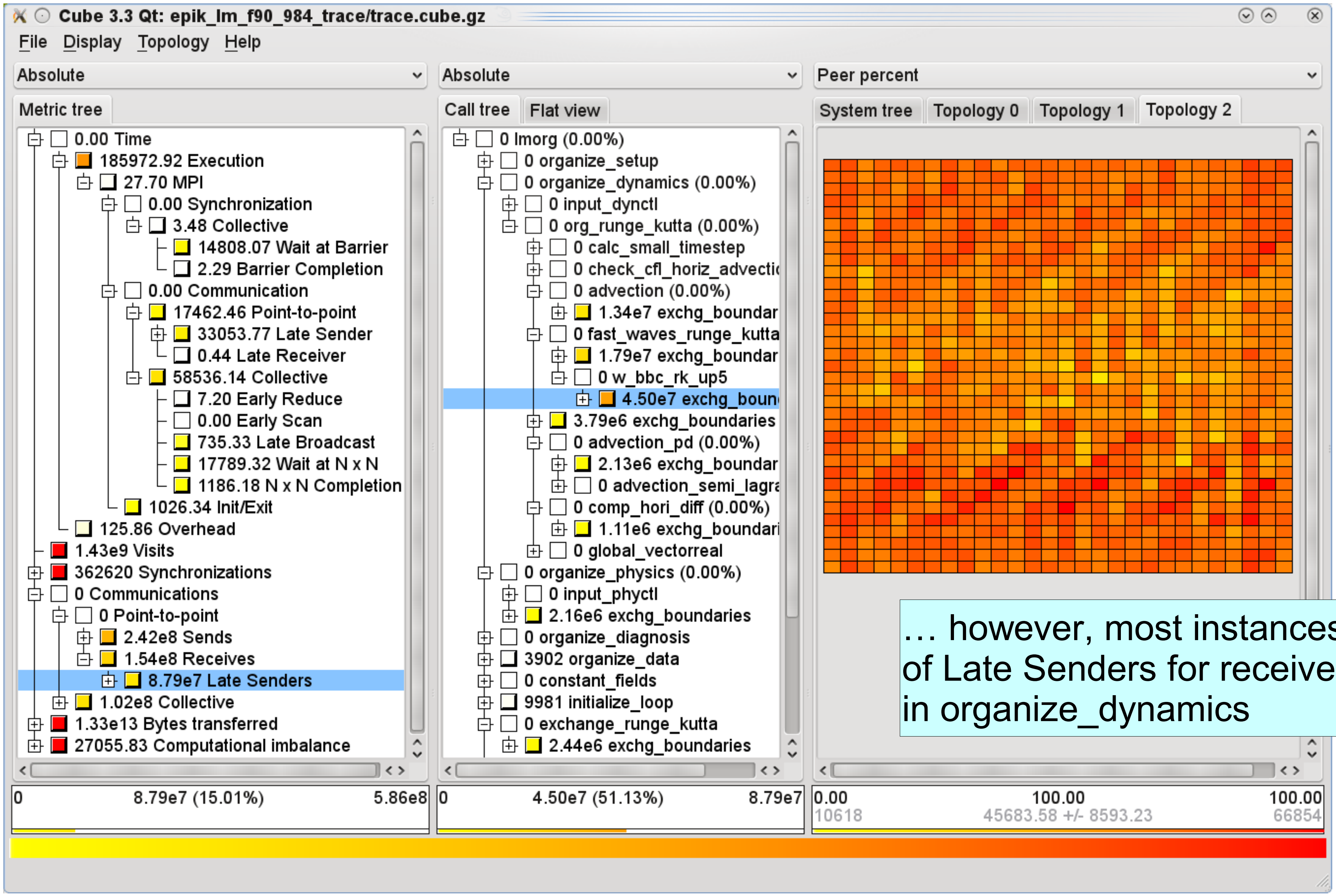
```
0.00 lmorg (0.00%)
  0.00 organize_setup
  0.00 organize_dynamics (0.00%)
    0.00 input_dynctl
    0.00 org_runge_kutta (0.00%)
      0.00 calc_small_timestep
      0.00 check_cfl_horiz_adve
      0.00 advection (0.00%)
        2196.65 exchg_bounda
      0.00 fast_waves_runge_ku
        3261.18 exchg_bounda
        0.00 w_bbc_rk_up5
          7656.16 exchg_bou
      866.20 exchg_boundaries
      0.00 advection_pd (0.00%
        516.63 exchg_boundar
        0.00 advection_semi_la
      0.00 comp_hori_diff (0.00%
        60.55 exchg_boundarie
      0.00 global_vectorreal
  0.00 organize_physics (0.00%)
    0.00 input_phyctl
    12009.04 exchg_boundaries
  0.00 organize_diagnosis
  1001.77 organize_data
  0.00 constant_fields
  2218.62 initialize_loop
  0.00 exchange_runge_kutta
    2290.73 exchg_boundaries
```

System tree | Topology 0 | Topology 1 | Topology 2

```
0.00        33053.77 (9.99%)     330737.50
0.00        28857.14 (87.30%)    33053.77
0.00        100.00               100.00
3.21        29.33 +/- 13.91      68.39
```

10% of total time blocked in Late Sender communication in exchg_boundaries ...

COSMO/XE6 late sender communications



Cube 3.3 Qt: epik_lm_f90_984_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1 | Topology 2

```
0.00 Time
  185972.92 Execution
    27.70 MPI
      0.00 Synchronization
        3.48 Collective
          14808.07 Wait at Barrier
          2.29 Barrier Completion
      0.00 Communication
        17462.46 Point-to-point
          33053.77 Late Sender
          0.44 Late Receiver
        58536.14 Collective
          7.20 Early Reduce
          0.00 Early Scan
          735.33 Late Broadcast
          17789.32 Wait at N x N
          1186.18 N x N Completion
      1026.34 Init/Exit
    125.86 Overhead
1.43e9 Visits
362620 Synchronizations
0 Communications
  0 Point-to-point
    2.42e8 Sends
    1.54e8 Receives
      8.79e7 Late Senders
  1.02e8 Collective
1.33e13 Bytes transferred
27055.83 Computational imbalance
```

```
0 lmorg (0.00%)
  0 organize_setup
  0 organize_dynamics (0.00%)
    0 input_dynctl
    0 org_runge_kutta (0.00%)
      0 calc_small_timestep
      0 check_cfl_horiz_advectic
      0 advection (0.00%)
        1.34e7 exchg_boundar
      0 fast_waves_runge_kutta
        1.79e7 exchg_boundar
        0 w_bbc_rk_up5
          4.50e7 exchg_boun
      3.79e6 exchg_boundaries
      0 advection_pd (0.00%)
        2.13e6 exchg_boundar
        0 advection_semi_lagra
      0 comp_hori_diff (0.00%)
        1.11e6 exchg_boundari
      0 global_vectorreal
  0 organize_physics (0.00%)
    0 input_phyctl
    2.16e6 exchg_boundaries
  0 organize_diagnosis
  3902 organize_data
  0 constant_fields
  9981 initialize_loop
  0 exchange_runge_kutta
    2.44e6 exchg_boundaries
```

0	8.79e7 (15.01%)	5.86e8

0	4.50e7 (51.13%)	8.79e7

0.00	100.00	100.00
10618	45683.58 +/- 8593.23	66854

… however, most instances of Late Senders for receives in organize_dynamics

COSMO-7/XE6 performance review

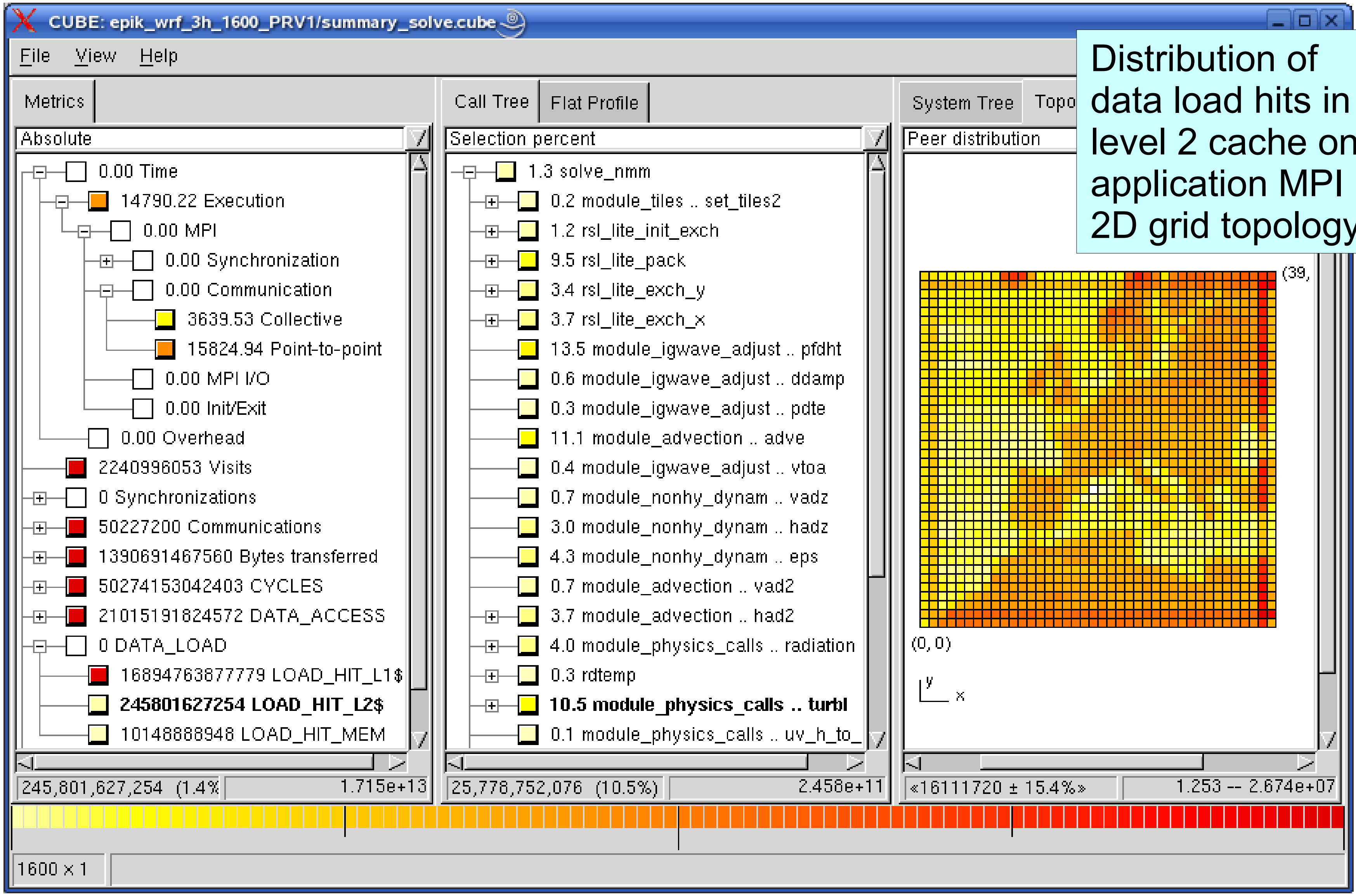
- 56% of total time in local computation
  - 32% in dynamics which is quite well balanced (11% std.dev)
  - 12% in physics is rather less well balanced (17% std.dev)
  - much of the imbalance is inherently physical/geographical
- 44% of total time in MPI
  - 5% collective synchronization (92% output_data)
  - 24% collective communication
    - 14% for MPI_Gather operations in output_data
    - 5% “Wait at NxN” mostly in dynamics check_cfl_horiz_advection
  - 15% point-to-point communication (91% exchg_boundaries)
    - 10% “Late Sender” time (44% dynamics, 36% physics)
    - 36% of receives are for “Late Senders” (95% in dynamics)
- Communication associated with file I/O was a major factor
  - the 4 dedicated I/O processes idle 95% of the time

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

Journal of Scientific Programming 16(2-3):167-181 (2008)

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 | Selection percent | Peer distribution

```
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
```

```
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_
```

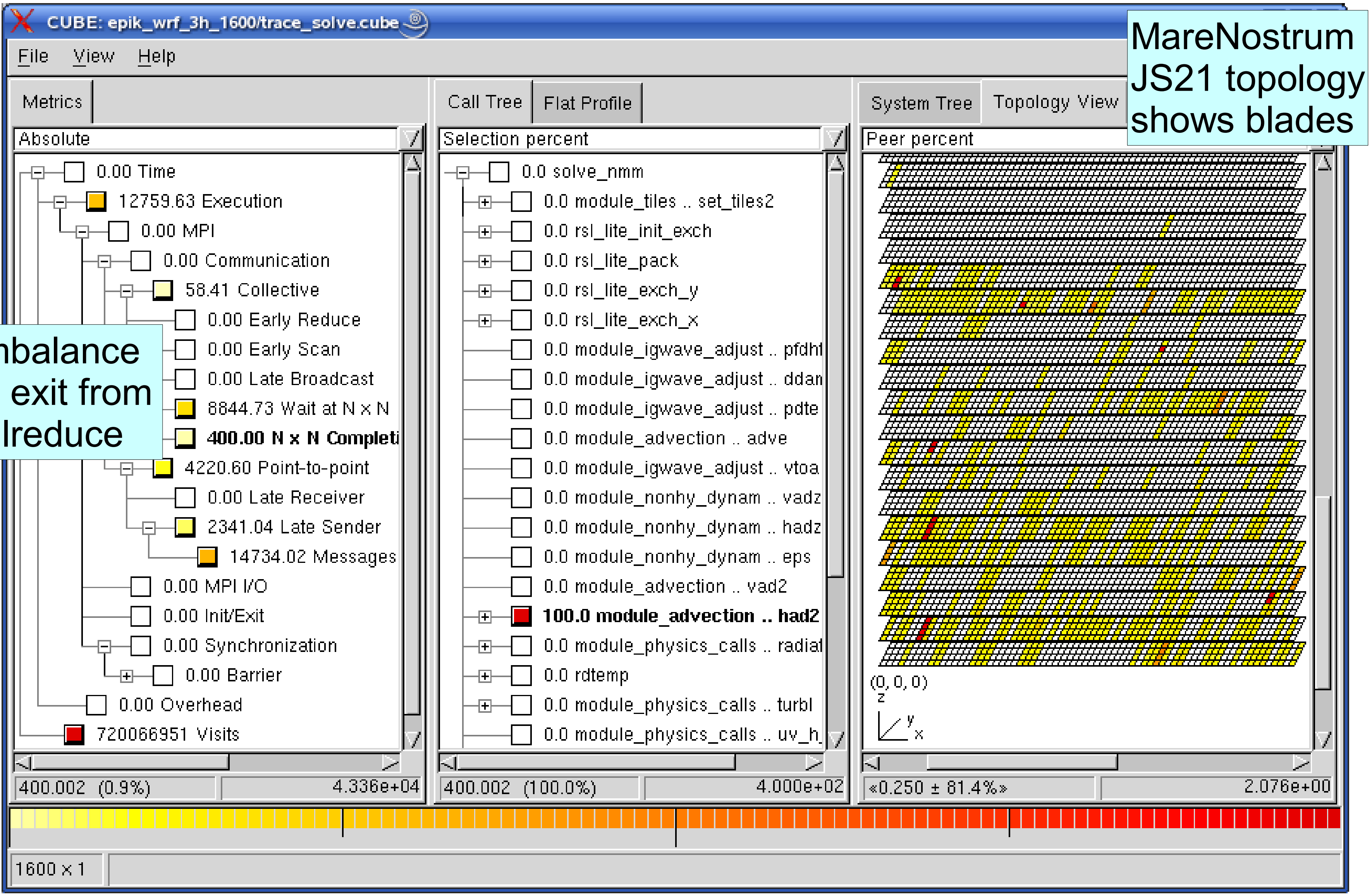
(39, | (0, 0) | y | x

245,801,627,254 (1.4% | 1.715e+13 | 25,778,752,076 (10.5%) | 2.458e+11 | «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



CUBE: epik_wrf_3h_1600/trace_solve.cube

File View Help

Metrics | Call Tree | Flat Profile | System Tree | Topology View

Absolute

```
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

Selection percent

```
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

Peer percent

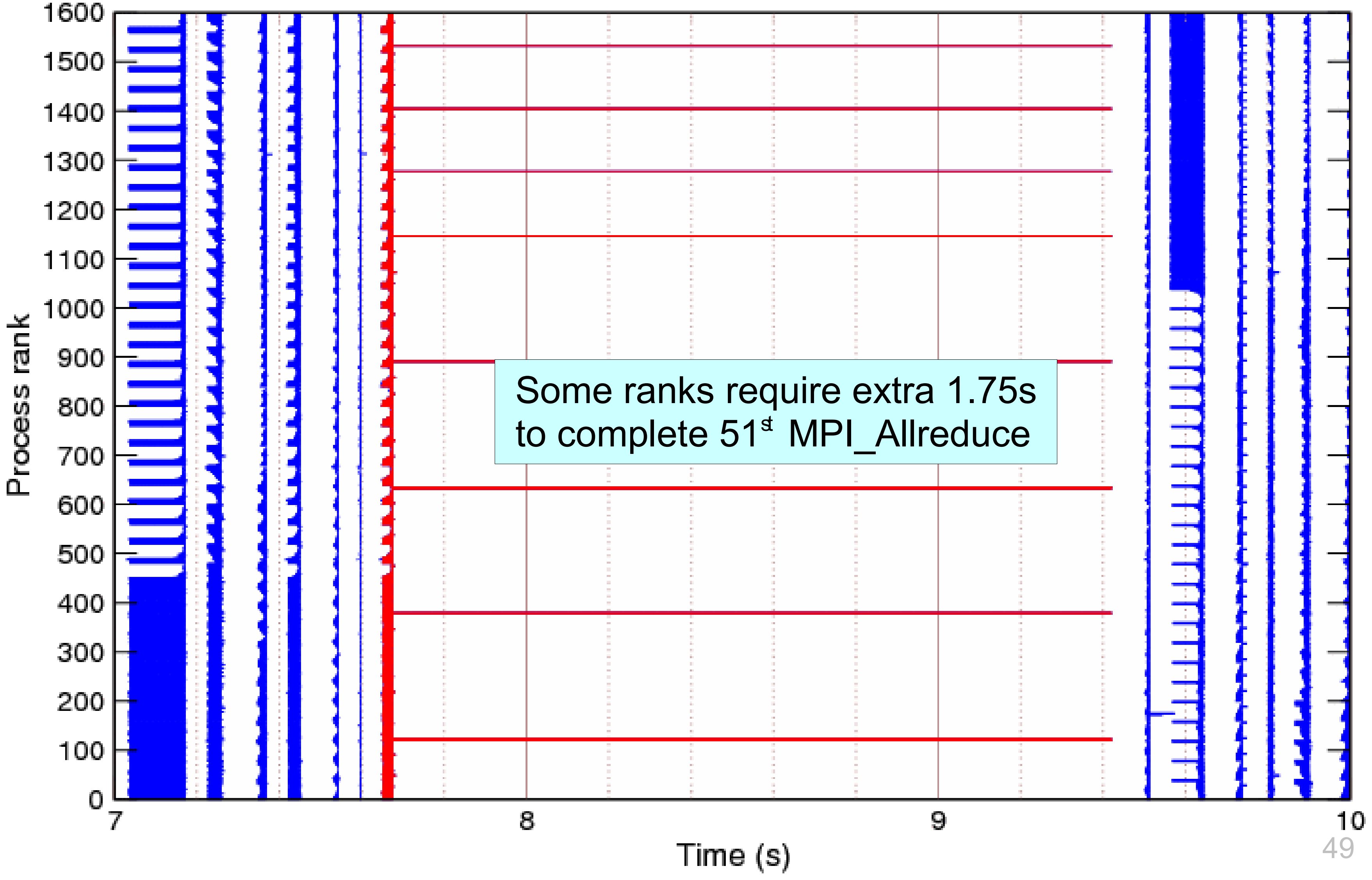
«0.250 ± 81.4%» 2.076e+00

1600 x 1

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

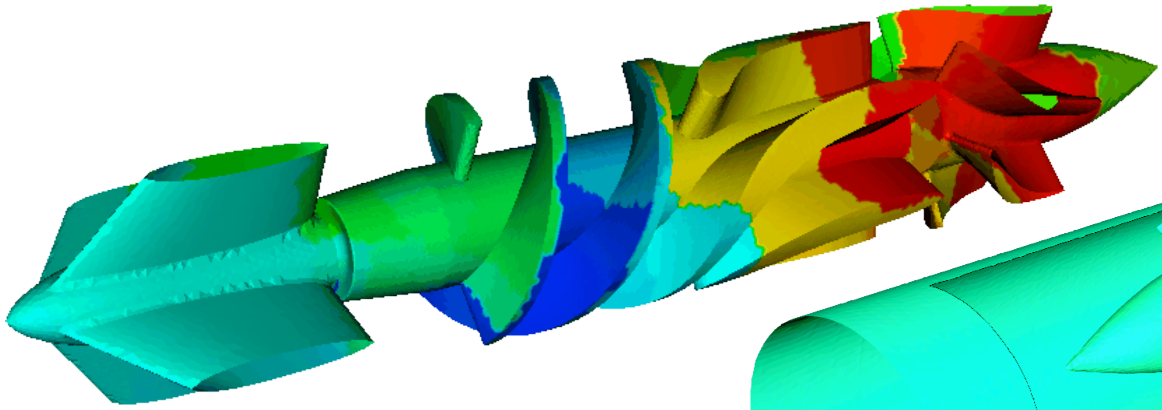
WRF/MareNostrum experience

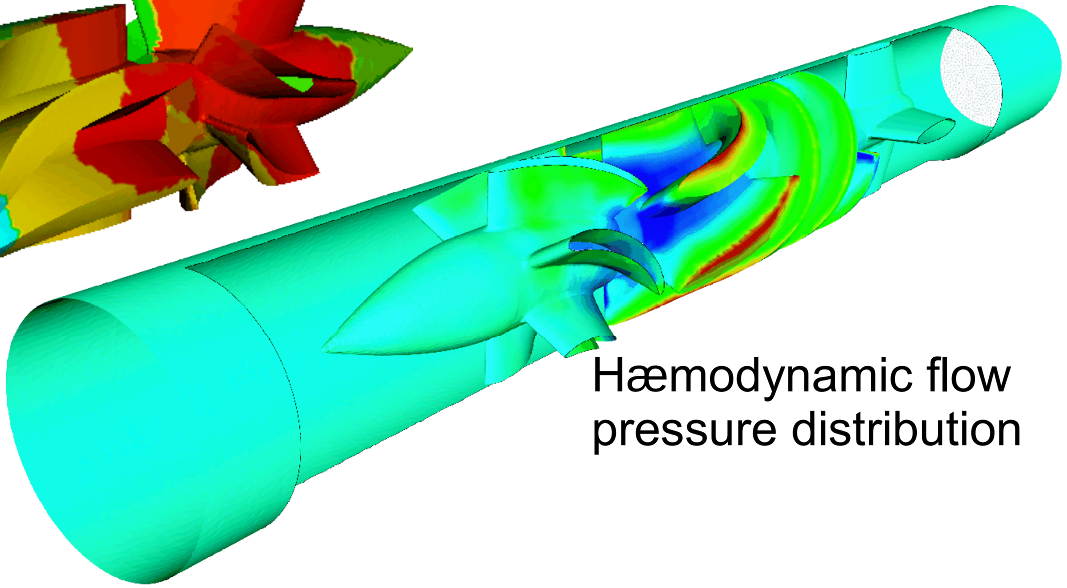
- Limited system I/O requires careful management
  - Selective instrumentation and measurement filtering
- PowerPC hardware counter metrics included in summary
  - Memory/cache data access hierarchy constructed
- Automated trace analysis quantified impact of imbalanced exit from MPI_Allreduce in “NxN 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

Journal of Scientific Programming 16(2-3):167-181 (2008)

XNS on BlueGene/L case study

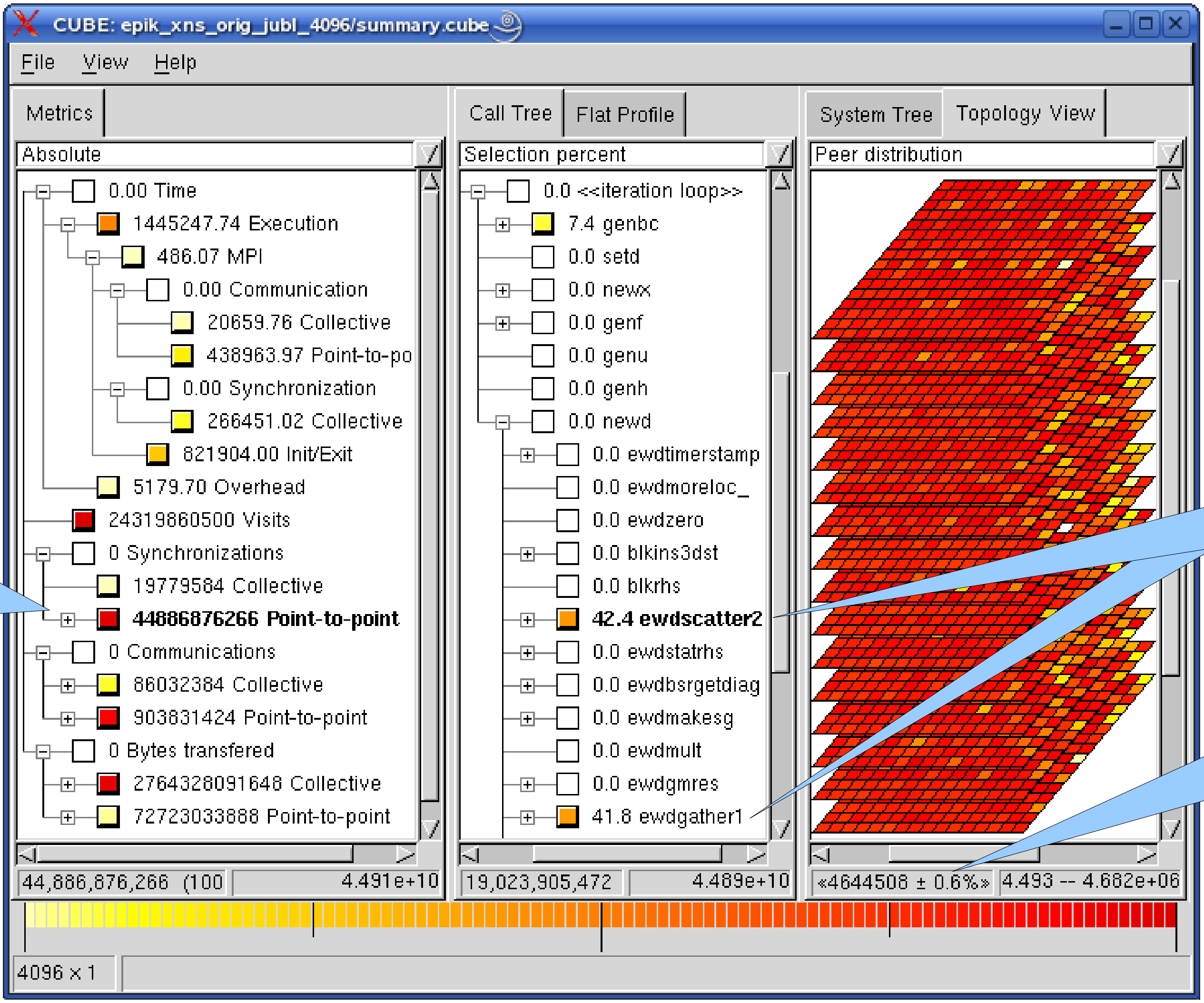
- CFD simulation of unsteady flows
  - developed by RWTH CATS group of Marek Behr
  - exploits finite-element techniques, unstructured 3D meshes, iterative solution strategies
- MPI parallel version (Dec-2006)
  - >40,000 lines of Fortran & C
  - DeBakey blood-pump dataset (3,714,611 elements)


Partitioned finite-element mesh


Hæmodynamic flow pressure distribution

XNS-DeBakey on jubl@4096 summary analysis

Point-to-point msgs w/o data



```
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
```

```
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
```

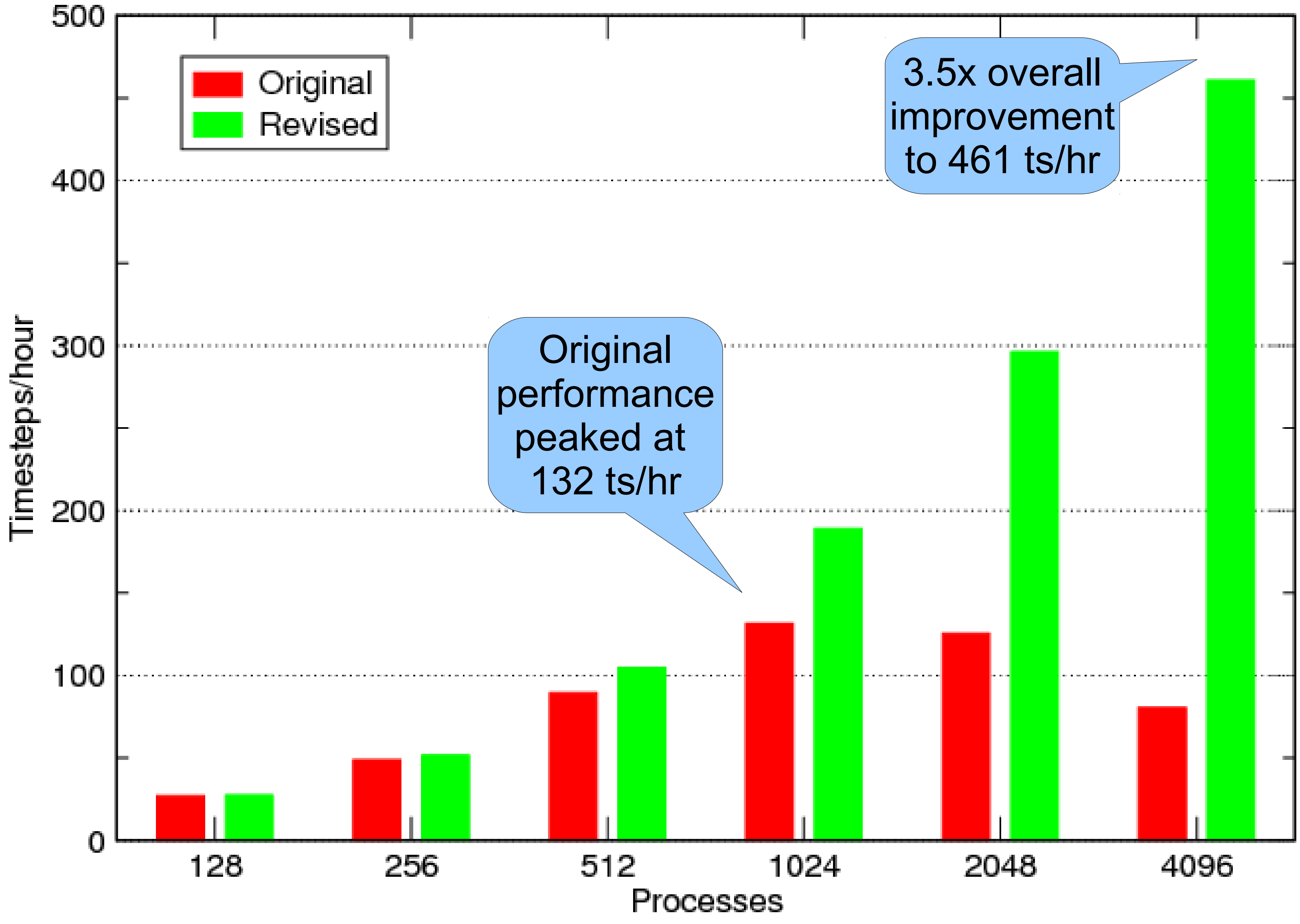
```
«4644508 ± 0.6%»   4.493 -- 4.682e+06
```

Masses of P2P synch operations

Primarily in scatter & gather

Processes all equally responsible

XNS-DeBakey scalability on BlueGene/L



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

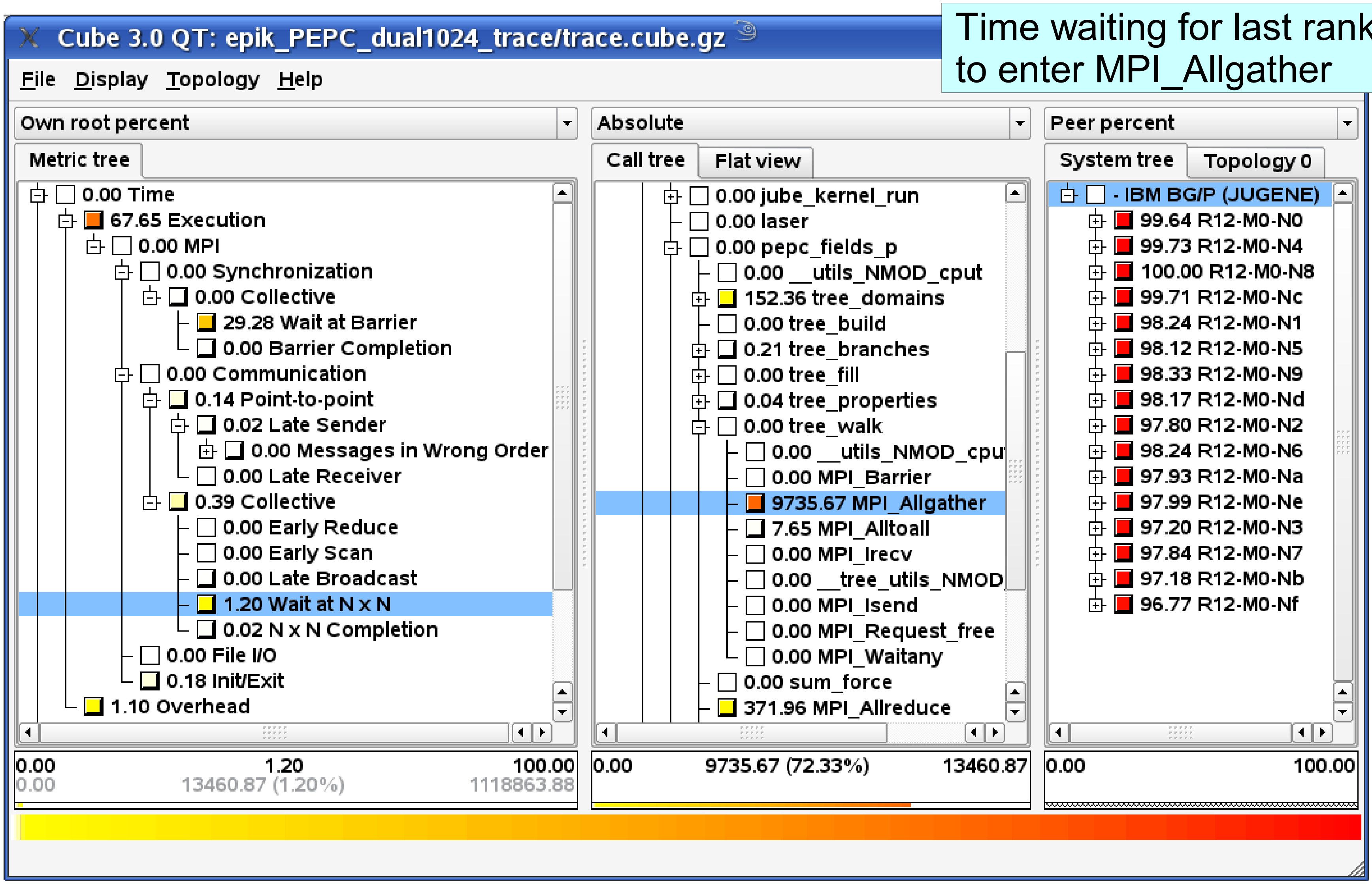
Proc. 14th EuroPVM/MPI, LNCS 4757 (2007)

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

Proc. 52nd Cray User Group (Edinburgh, 2010)

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



Cube 3.0 QT: epik_PEPC_dual1024_trace/trace.cube.gz

File Display Topology Help

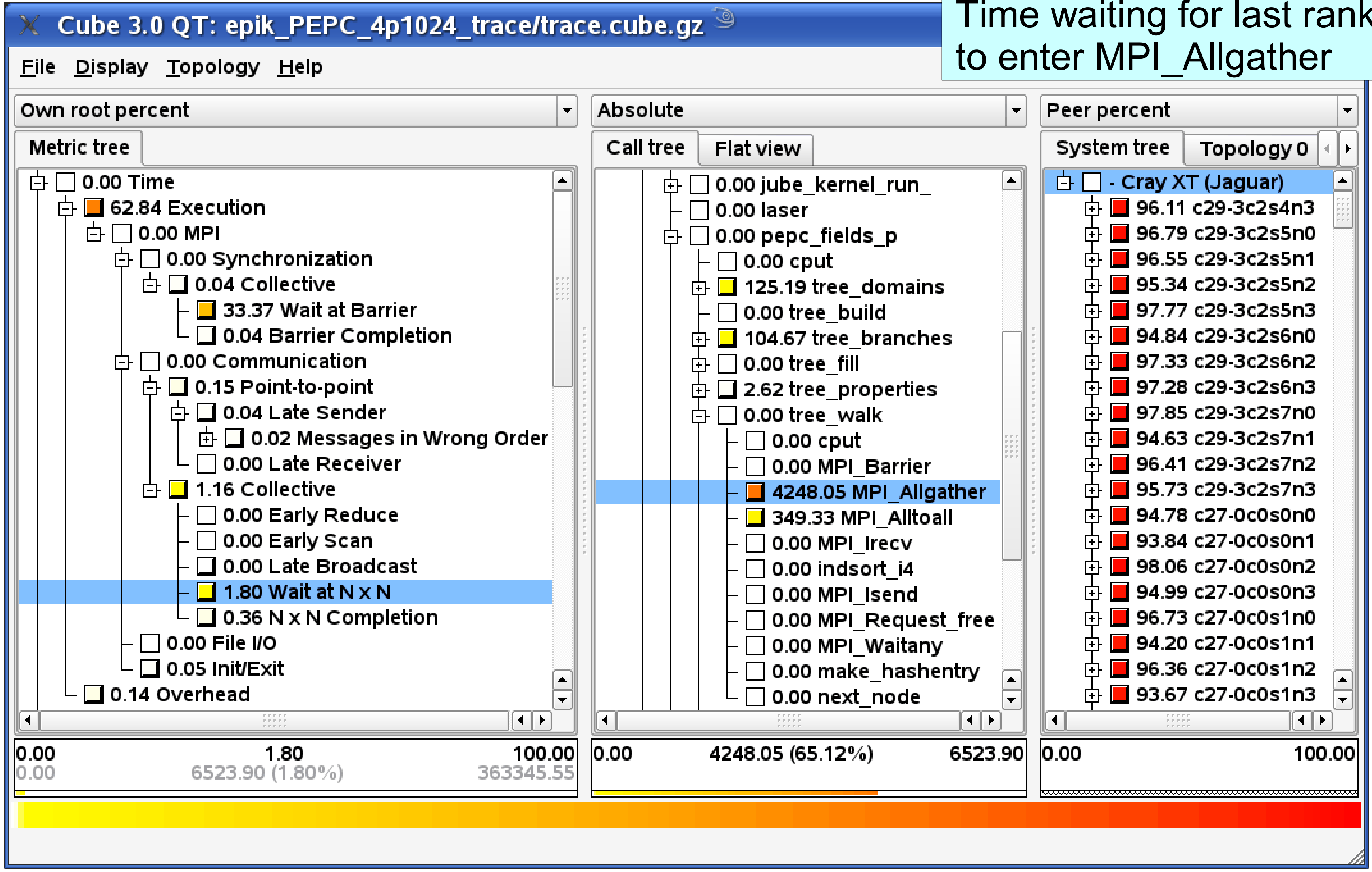
Time waiting for last rank to enter MPI_Allgather

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

0.00 Time                             0.00 jube_kernel_run                  - IBM BG/P (JUGENE)
  67.65 Execution                     0.00 laser                              99.64 R12-M0-N0
    0.00 MPI                          0.00 pepc_fields_p                      99.73 R12-M0-N4
      0.00 Synchronization              0.00 __utils_NMOD_cput                100.00 R12-M0-N8
        0.00 Collective                 152.36 tree_domains                   99.71 R12-M0-Nc
          29.28 Wait at Barrier         0.00 tree_build                       98.24 R12-M0-N1
          0.00 Barrier Completion       0.21 tree_branches                    98.12 R12-M0-N5
      0.00 Communication                0.00 tree_fill                        98.33 R12-M0-N9
        0.14 Point-to-point             0.04 tree_properties                  98.17 R12-M0-Nd
          0.02 Late Sender              0.00 tree_walk                        97.80 R12-M0-N2
            0.00 Messages in Wrong Order  0.00 __utils_NMOD_cpu               98.24 R12-M0-N6
          0.00 Late Receiver              0.00 MPI_Barrier                    97.93 R12-M0-Na
        0.39 Collective                   9735.67 MPI_Allgather               97.99 R12-M0-Ne
          0.00 Early Reduce               7.65 MPI_Alltoall                   97.20 R12-M0-N3
          0.00 Early Scan                 0.00 MPI_Irecv                      97.84 R12-M0-N7
          0.00 Late Broadcast             0.00 __tree_utils_NMOD              97.18 R12-M0-Nb
          1.20 Wait at N x N              0.00 MPI_Isend                      96.77 R12-M0-Nf
          0.02 N x N Completion           0.00 MPI_Request_free
      0.00 File I/O                       0.00 MPI_Waitany
      0.18 Init/Exit                    0.00 sum_force
  1.10 Overhead                         371.96 MPI_Allreduce

0.00        1.20             100.00   0.00   9735.67 (72.33%)   13460.87   0.00        100.00
0.00  13460.87 (1.20%)   1118863.88
```

PEPC@1024 on Cray XT4: Wait at NxN time



Time waiting for last rank to enter MPI_Allgather

Metric tree (Own root percent):

```
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 (Absolute):

```
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 (Peer percent):

```
- 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
```

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

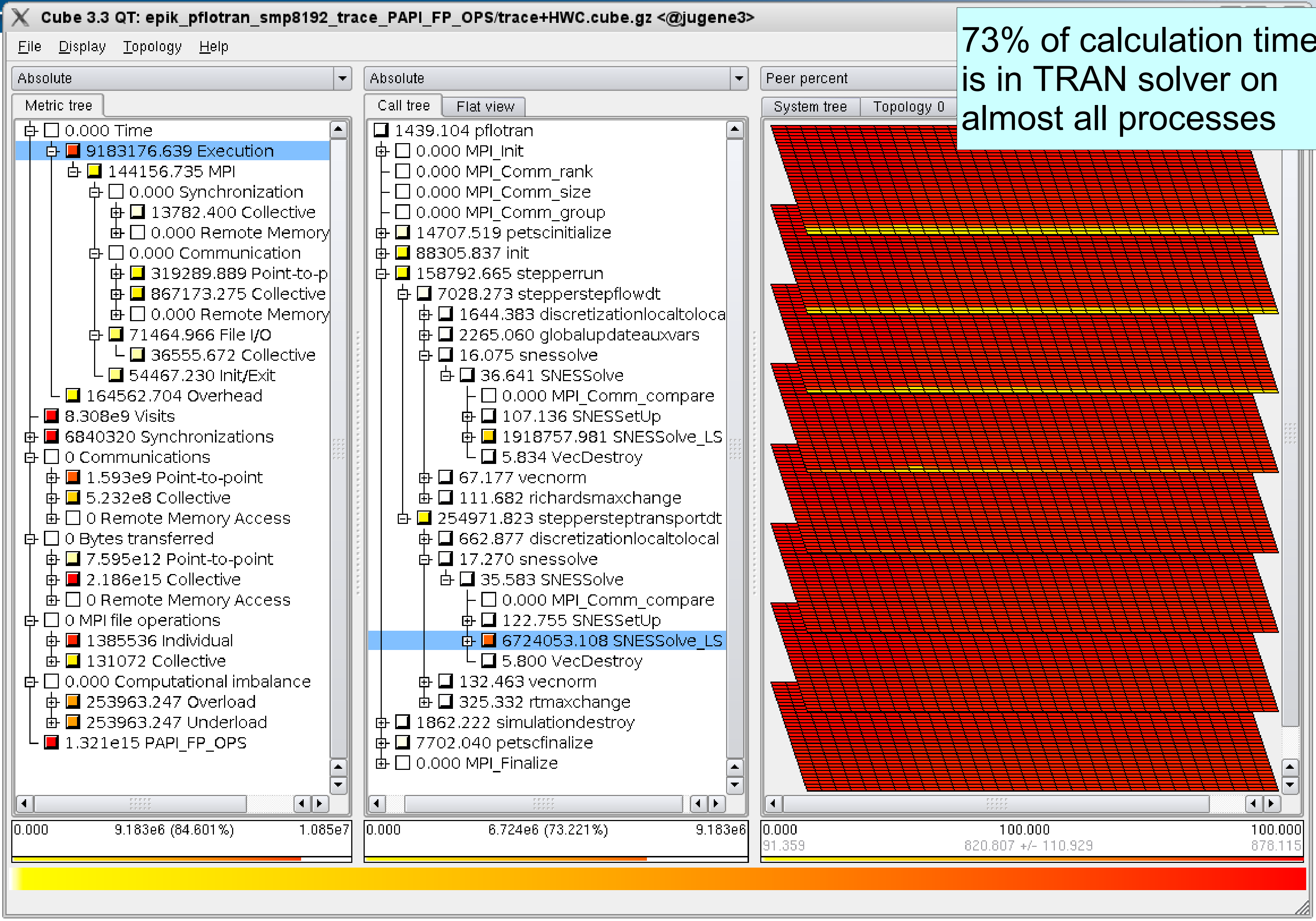
Proc. 52nd Cray User Group (Edinburgh, 2010)

PFLOTRAN on BlueGene/P case study

- 3D reservoir simulator combining alternating
  - PFLOW non-isothermal, multiphase groundwater flow
  - PTRAN reactive, multi-component contaminant transport
  - developed by LANL/ORNL/PNNL
- MPI with PETSc, LAPACK, BLAS & HDF5 I/O libraries
  - ~80,000 lines (97 source files) Fortran9X
  - PFLOTRAN & PETSc fully instrumented by IBM XL compilers
    - filter produced listing 856 USR routines (leaving 291 COM)
    - 1732 unique callpaths (399 in FLOW, 375 in TRAN)
    - 633 MPI callpaths (121 in FLOW, 114 in TRAN)
      - 29 distinct MPI routines recorded (excludes 15 misc. routines)
- Run on IBM BlueGene/P with '2B' input dataset (10 steps)
  - Scalasca summary & trace measurements (some with PAPI)
  - 22% dilation of FLOW, 10% dilation of TRAN [8k summary]

Proc. 53rd Cray User Group (Fairbanks, 2011)

PFLOTRAN jugene@smp8192 trace analysis



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

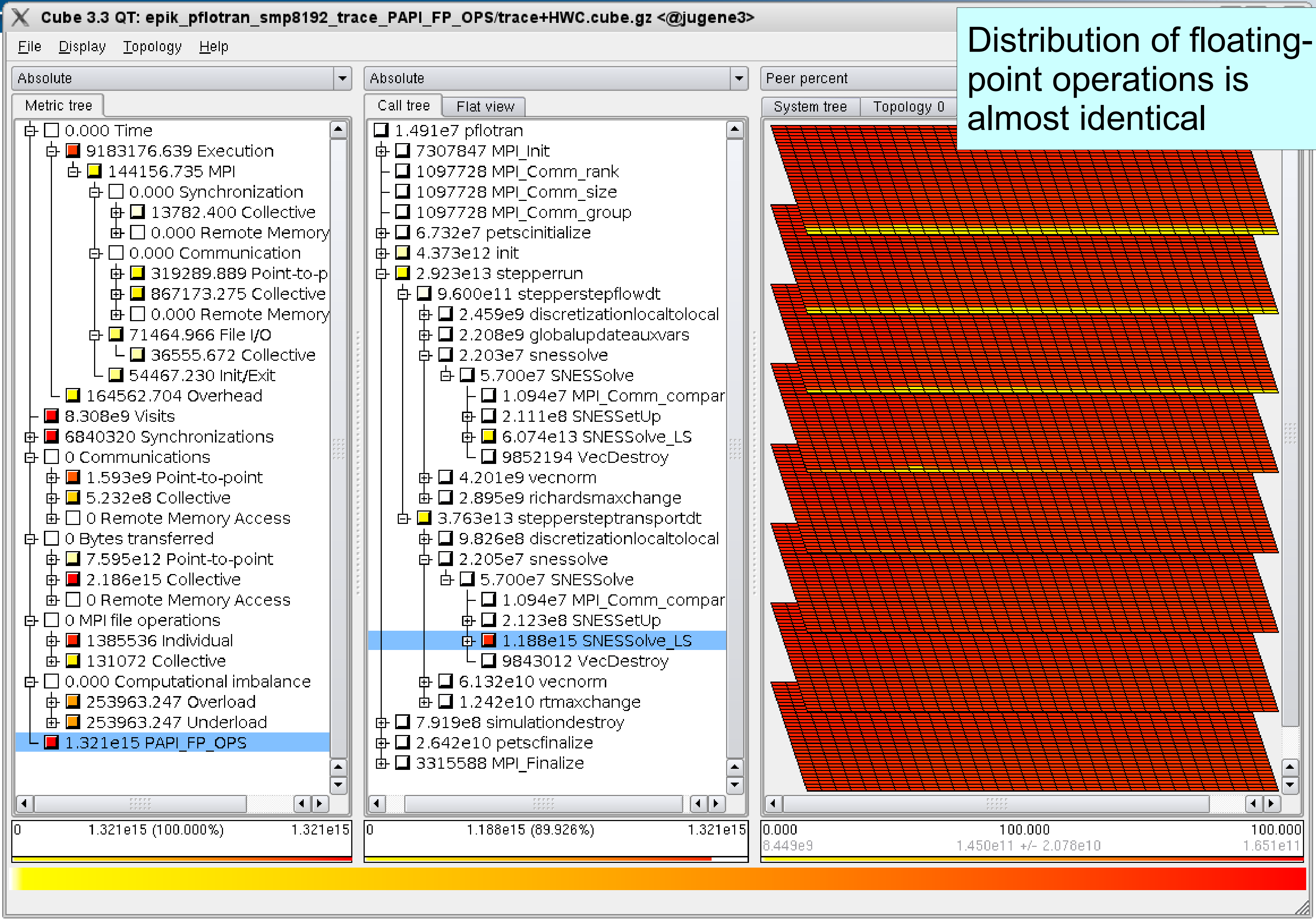
```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        13782.400 Collective
        0.000 Remote Memory
      0.000 Communication
        319289.889 Point-to-p
        867173.275 Collective
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
  164562.704 Overhead
8.308e9 Visits
6840320 Synchronizations
0 Communications
  1.593e9 Point-to-point
  5.232e8 Collective
  0 Remote Memory Access
0 Bytes transferred
  7.595e12 Point-to-point
  2.186e15 Collective
  0 Remote Memory Access
0 MPI file operations
  1385536 Individual
  131072 Collective
0.000 Computational imbalance
  253963.247 Overload
  253963.247 Underload
1.321e15 PAPI_FP_OPS
```

```
1439.104 pflotran
  0.000 MPI_Init
  0.000 MPI_Comm_rank
  0.000 MPI_Comm_size
  0.000 MPI_Comm_group
  14707.519 petscinitialize
  88305.837 init
  158792.665 stepperrun
    7028.273 stepperstepflowdt
      1644.383 discretizationlocaltoloca
      2265.060 globalupdateauxvars
      16.075 snessolve
        36.641 SNESSolve
          0.000 MPI_Comm_compare
          107.136 SNESSetUp
          1918757.981 SNESSolve_LS
          5.834 VecDestroy
      67.177 vecnorm
      111.682 richardsmaxchange
    254971.823 steppersteptransportdt
      662.877 discretizationlocaltolocal
      17.270 snessolve
        35.583 SNESSolve
          0.000 MPI_Comm_compare
          122.755 SNESSetUp
          6724053.108 SNESSolve_LS
          5.800 VecDestroy
      132.463 vecnorm
      325.332 rtmaxchange
  1862.222 simulationdestroy
  7702.040 petscfinalize
  0.000 MPI_Finalize
```

```
0.000        9.183e6 (84.601%)        1.085e7
0.000        6.724e6 (73.221%)        9.183e6
0.000              100.000             100.000
91.359     820.807 +/- 110.929         878.115
```

73% of calculation time is in TRAN solver on almost all processes

PFLOTRAN jugene@smp8192 trace analysis



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

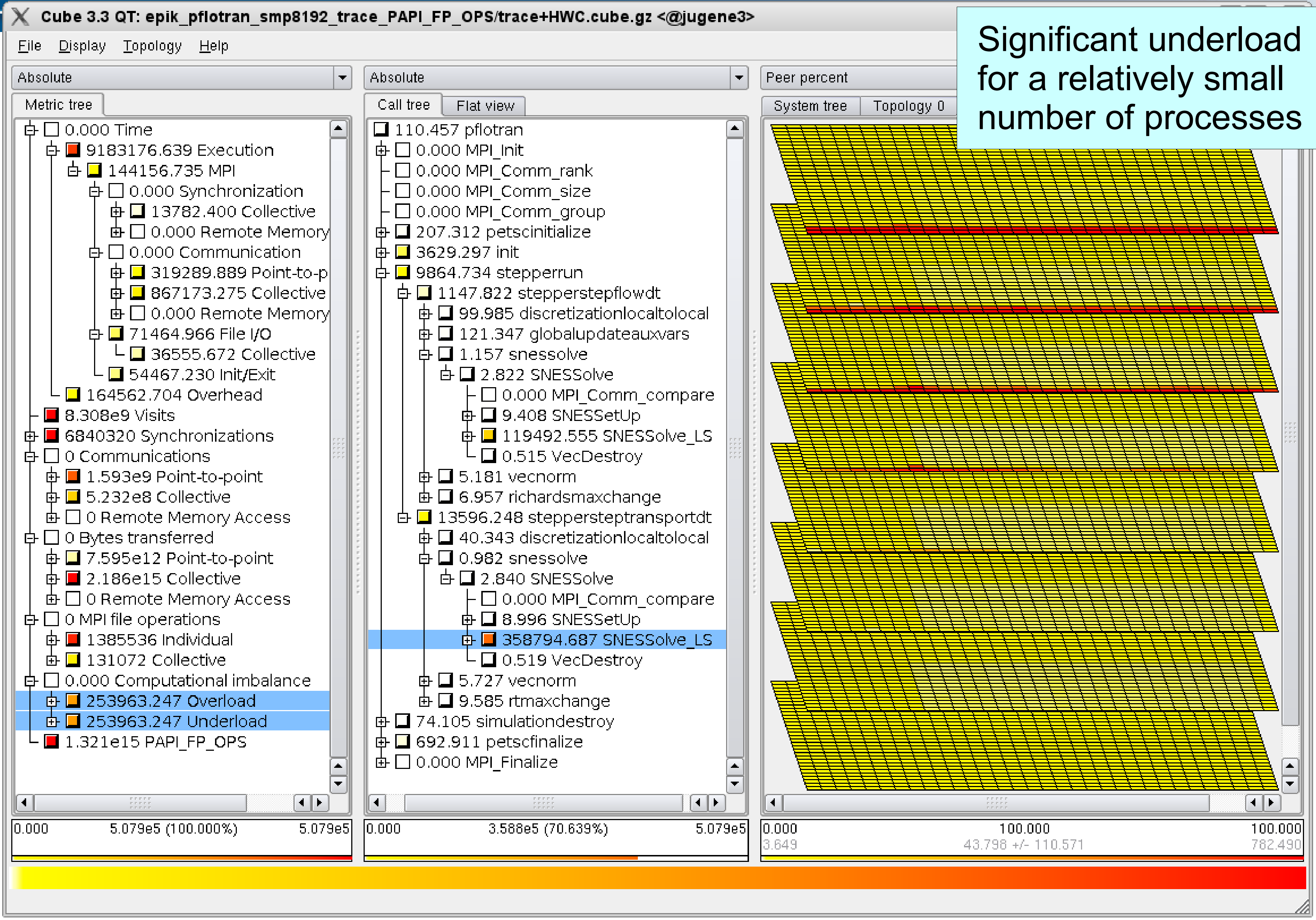
```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        13782.400 Collective
        0.000 Remote Memory
      0.000 Communication
        319289.889 Point-to-p
        867173.275 Collective
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
  164562.704 Overhead
8.308e9 Visits
6840320 Synchronizations
0 Communications
  1.593e9 Point-to-point
  5.232e8 Collective
  0 Remote Memory Access
0 Bytes transferred
  7.595e12 Point-to-point
  2.186e15 Collective
  0 Remote Memory Access
0 MPI file operations
  1385536 Individual
  131072 Collective
0.000 Computational imbalance
  253963.247 Overload
  253963.247 Underload
1.321e15 PAPI_FP_OPS
```

```
1.491e7 pflotran
  7307847 MPI_Init
  1097728 MPI_Comm_rank
  1097728 MPI_Comm_size
  1097728 MPI_Comm_group
  6.732e7 petscinitialize
  4.373e12 init
  2.923e13 stepperrun
    9.600e11 stepperstepflowdt
      2.459e9 discretizationlocaltolocal
      2.208e9 globalupdateauxvars
      2.203e7 snessolve
        5.700e7 SNESSolve
          1.094e7 MPI_Comm_compar
          2.111e8 SNESSetUp
          6.074e13 SNESSolve_LS
          9852194 VecDestroy
      4.201e9 vecnorm
      2.895e9 richardsmaxchange
    3.763e13 stepperstepransportdt
      9.826e8 discretizationlocaltolocal
      2.205e7 snessolve
        5.700e7 SNESSolve
          1.094e7 MPI_Comm_compar
          2.123e8 SNESSetUp
          1.188e15 SNESSolve_LS
          9843012 VecDestroy
      6.132e10 vecnorm
      1.242e10 rtmaxchange
  7.919e8 simulationdestroy
  2.642e10 petscfinalize
  3315588 MPI_Finalize
```

```
0        1.321e15 (100.000%)        1.321e15
0        1.188e15 (89.926%)         1.321e15
0.000         100.000              100.000
8.449e9   1.450e11 +/- 2.078e10   1.651e11
```

Distribution of floating-point operations is almost identical

PFLOTRAN jugene@smp8192 trace analysis



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

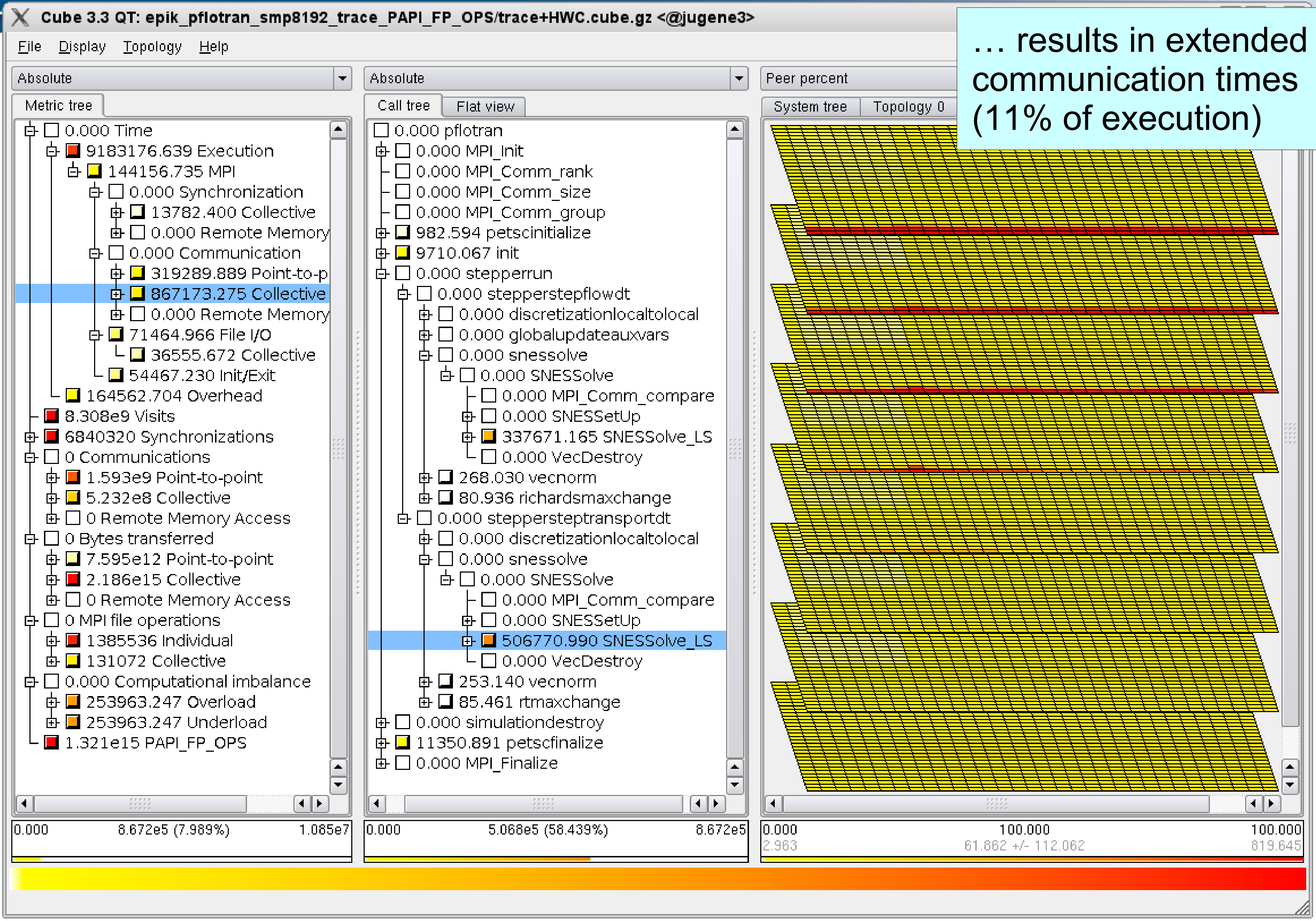
```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        13782.400 Collective
        0.000 Remote Memory
      0.000 Communication
        319289.889 Point-to-p
        867173.275 Collective
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
  164562.704 Overhead
8.308e9 Visits
6840320 Synchronizations
0 Communications
  1.593e9 Point-to-point
  5.232e8 Collective
  0 Remote Memory Access
0 Bytes transferred
  7.595e12 Point-to-point
  2.186e15 Collective
  0 Remote Memory Access
0 MPI file operations
  1385536 Individual
  131072 Collective
0.000 Computational imbalance
  253963.247 Overload
  253963.247 Underload
1.321e15 PAPI_FP_OPS
```

```
110.457 pflotran
  0.000 MPI_Init
  0.000 MPI_Comm_rank
  0.000 MPI_Comm_size
  0.000 MPI_Comm_group
  207.312 petscinitialize
  3629.297 init
  9864.734 stepperrun
    1147.822 stepperstepflowdt
      99.985 discretizationlocaltolocal
      121.347 globalupdateauxvars
      1.157 snessolve
        2.822 SNESSolve
          0.000 MPI_Comm_compare
          9.408 SNESSetUp
          119492.555 SNESSolve_LS
          0.515 VecDestroy
      5.181 vecnorm
      6.957 richardsmaxchange
    13596.248 steppersteptransportdt
      40.343 discretizationlocaltolocal
      0.982 snessolve
        2.840 SNESSolve
          0.000 MPI_Comm_compare
          8.996 SNESSetUp
          358794.687 SNESSolve_LS
          0.519 VecDestroy
      5.727 vecnorm
      9.585 rtmaxchange
  74.105 simulationdestroy
  692.911 petscfinalize
  0.000 MPI_Finalize
```

```
0.000      5.079e5 (100.000%)      5.079e5
0.000      3.588e5 (70.639%)       5.079e5
0.000      100.000                 100.000
3.649      43.798 +/- 110.571      782.490
```

Significant underload for a relatively small number of processes

PFLOTRAN jugene@smp8192 trace analysis



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree:

```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        13782.400 Collective
        0.000 Remote Memory
      0.000 Communication
        319289.889 Point-to-p
        867173.275 Collective
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
    164562.704 Overhead
8.308e9 Visits
6840320 Synchronizations
0 Communications
  1.593e9 Point-to-point
  5.232e8 Collective
  0 Remote Memory Access
0 Bytes transferred
  7.595e12 Point-to-point
  2.186e15 Collective
  0 Remote Memory Access
0 MPI file operations
  1385536 Individual
  131072 Collective
0.000 Computational imbalance
  253963.247 Overload
  253963.247 Underload
1.321e15 PAPI_FP_OPS
```

Call tree:

```
0.000 pflotran
  0.000 MPI_Init
  0.000 MPI_Comm_rank
  0.000 MPI_Comm_size
  0.000 MPI_Comm_group
  982.594 petscinitialize
  9710.067 init
  0.000 stepperrun
    0.000 stepperstepflowdt
      0.000 discretizationlocaltolocal
      0.000 globalupdateauxvars
      0.000 snessolve
        0.000 SNESSolve
          0.000 MPI_Comm_compare
          0.000 SNESSetUp
          337671.165 SNESSolve_LS
          0.000 VecDestroy
      268.030 vecnorm
      80.936 richardsmaxchange
    0.000 steppersteptransportdt
      0.000 discretizationlocaltolocal
      0.000 snessolve
        0.000 SNESSolve
          0.000 MPI_Comm_compare
          0.000 SNESSetUp
          506770.990 SNESSolve_LS
          0.000 VecDestroy
      253.140 vecnorm
      85.461 rtmaxchange
  0.000 simulationdestroy
  11350.891 petscfinalize
  0.000 MPI_Finalize
```

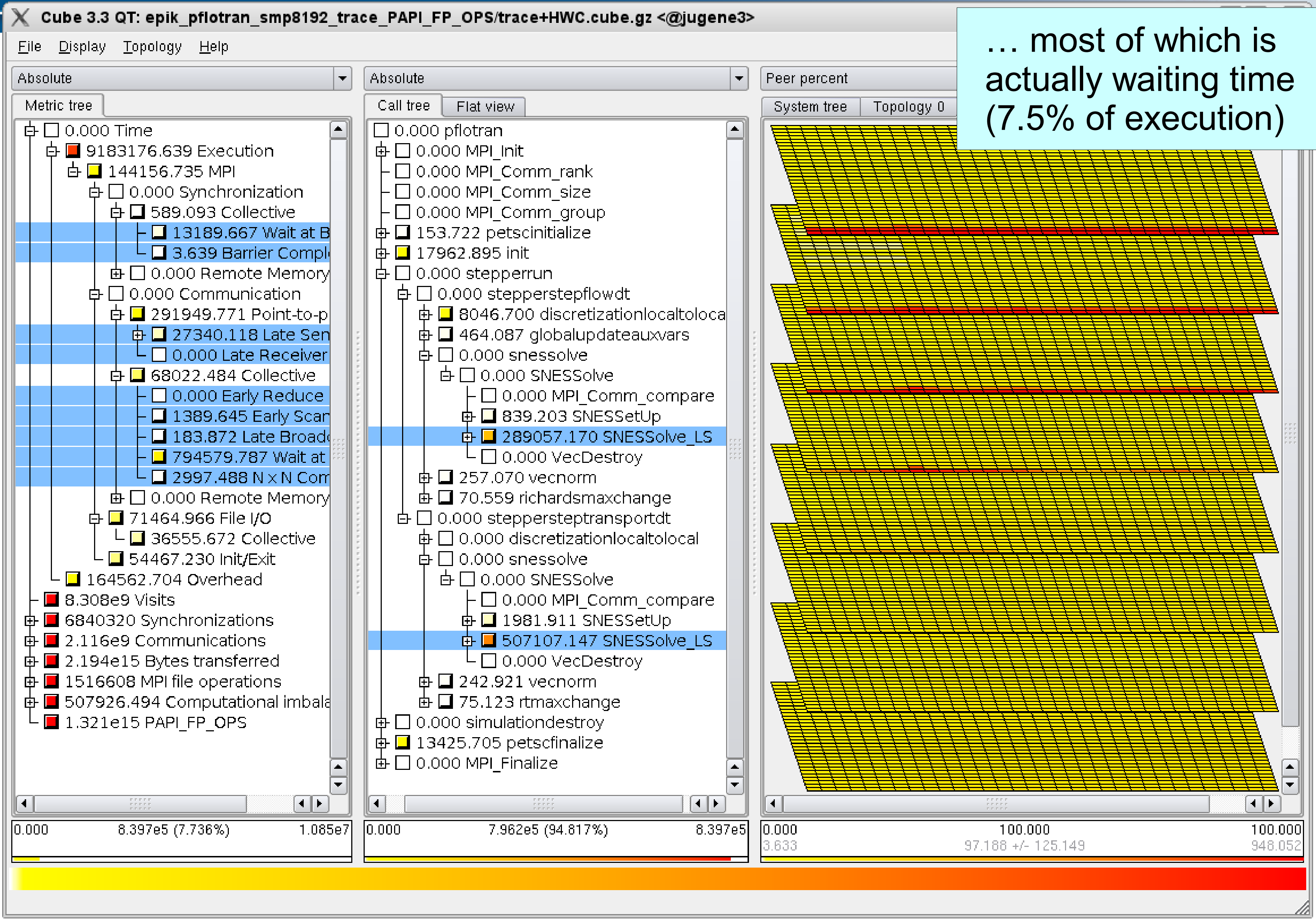
System tree | Topology 0

```
0.000    8.672e5 (7.989%)    1.085e7
0.000    5.068e5 (58.439%)   8.672e5
0.000    100.000             100.000
2.963    61.862 +/- 112.062  819.645
```

… results in extended communication times (11% of execution)

PFLOTRAN jugene@smp8192 trace analysis

VI-HPS



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

File Display Topology Help

Absolute | Absolute | Peer percent

Metric tree

```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        589.093 Collective
          13189.667 Wait at B
          3.639 Barrier Compl
        0.000 Remote Memory
      0.000 Communication
        291949.771 Point-to-p
          27340.118 Late Sen
          0.000 Late Receiver
        68022.484 Collective
          0.000 Early Reduce
          1389.645 Early Scan
          183.872 Late Broadc
          794579.787 Wait at
          2997.488 N x N Com
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
    164562.704 Overhead
  8.308e9 Visits
  6840320 Synchronizations
  2.116e9 Communications
  2.194e15 Bytes transferred
  1516608 MPI file operations
  507926.494 Computational imbala
  1.321e15 PAPI_FP_OPS
```

Call tree | Flat view

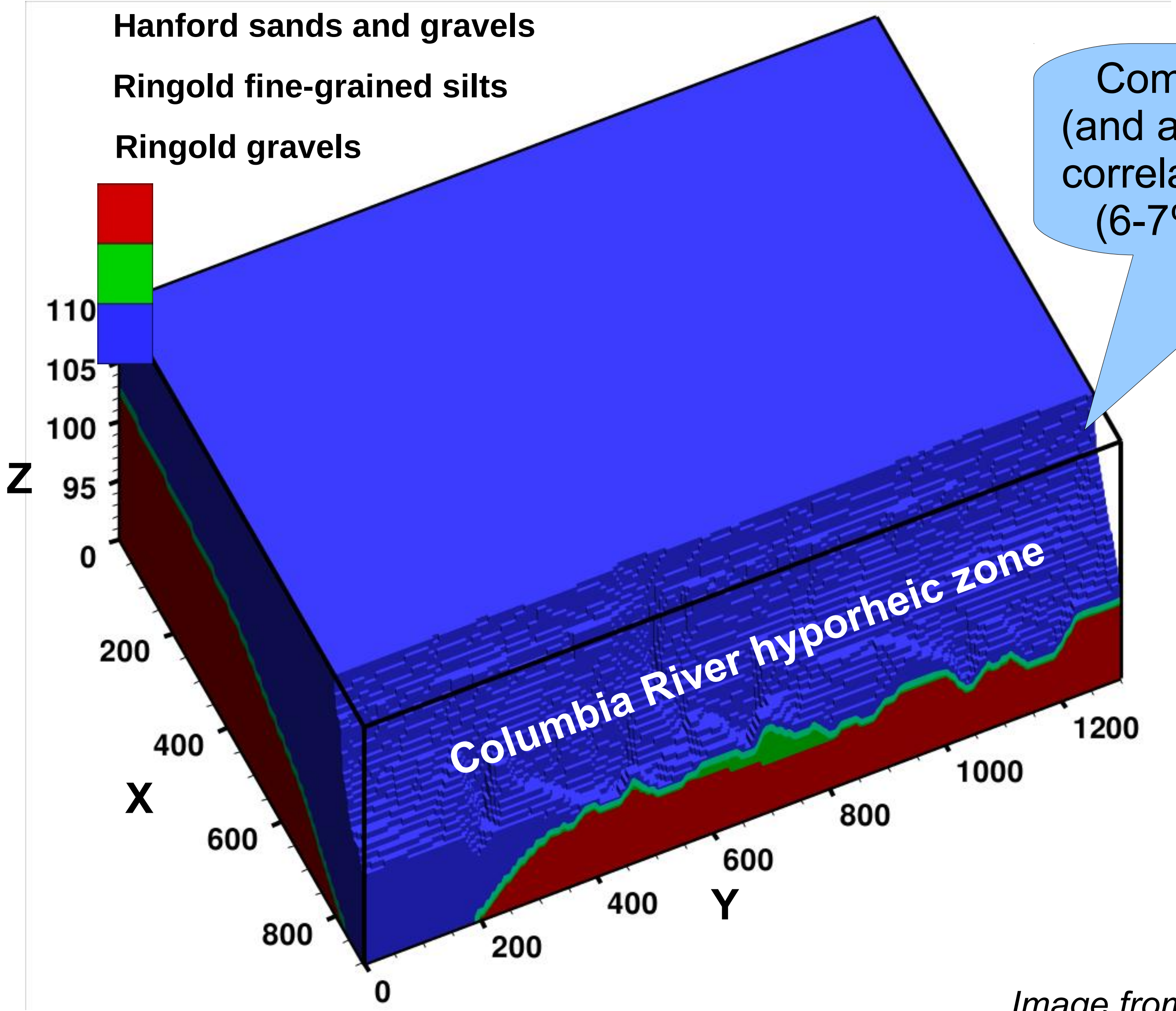
```
0.000 pflotran
  0.000 MPI_Init
  0.000 MPI_Comm_rank
  0.000 MPI_Comm_size
  0.000 MPI_Comm_group
  153.722 petscinitialize
  17962.895 init
  0.000 stepperrun
    0.000 stepperstepflowdt
      8046.700 discretizationlocaltoloca
      464.087 globalupdateauxvars
      0.000 snessolve
        0.000 SNESSolve
          0.000 MPI_Comm_compare
          839.203 SNESSetUp
          289057.170 SNESSolve_LS
          0.000 VecDestroy
      257.070 vecnorm
      70.559 richardsmaxchange
    0.000 stepperstepransportdt
      0.000 discretizationlocaltolocal
      0.000 snessolve
        0.000 SNESSolve
          0.000 MPI_Comm_compare
          1981.911 SNESSetUp
          507107.147 SNESSolve_LS
          0.000 VecDestroy
      242.921 vecnorm
      75.123 rtmaxchange
  0.000 simulationdestroy
  13425.705 petscfinalize
  0.000 MPI_Finalize
```

System tree | Topology 0

```
0.000        8.397e5 (7.736%)        1.085e7
0.000        7.962e5 (94.817%)       8.397e5
0.000        100.000                 100.000
3.633        97.188 +/- 125.149      948.052
```

… most of which is actually waiting time (7.5% of execution)

Geology of DOE Hanford 300 area (WA, USA)

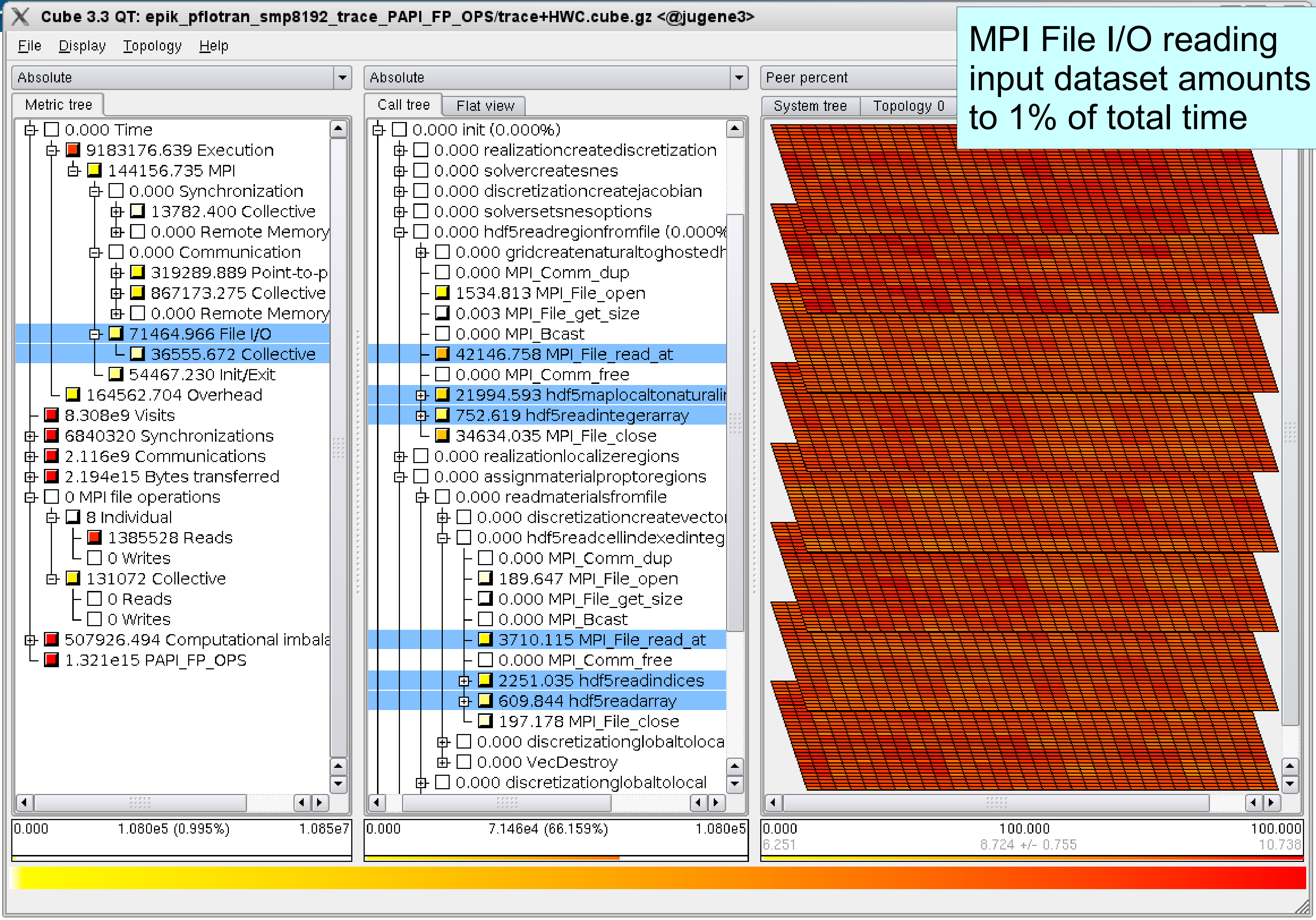


Hanford sands and gravels
Ringold fine-grained silts
Ringold gravels
Z
110
105
100
95
0
X
200
400
600
800
Y
0
200
400
600
800
1000
1200
Columbia River hyporheic zone

Computational imbalance (and associated waiting time) correlate to inactive grid cells (6-7% of total) within river

Image from Glenn Hammond (PNNL)

PFLOTRAN jugene@smp8192 trace analysis



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

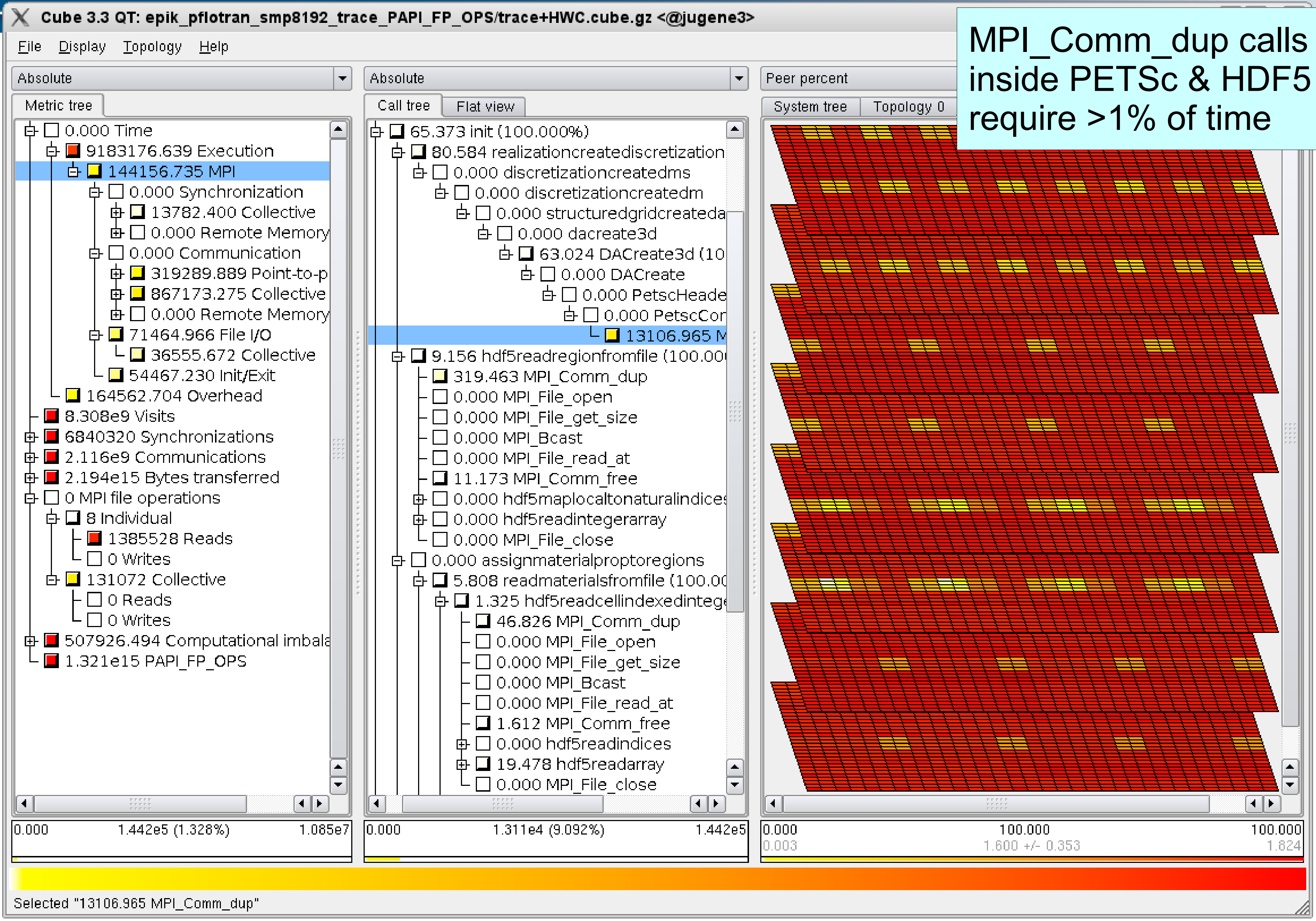
```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        13782.400 Collective
        0.000 Remote Memory
      0.000 Communication
        319289.889 Point-to-p
        867173.275 Collective
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
  164562.704 Overhead
8.308e9 Visits
6840320 Synchronizations
2.116e9 Communications
2.194e15 Bytes transferred
0 MPI file operations
  8 Individual
    1385528 Reads
    0 Writes
  131072 Collective
    0 Reads
    0 Writes
507926.494 Computational imbala
1.321e15 PAPI_FP_OPS
```

```
0.000 init (0.000%)
  0.000 realizationcreatediscretization
  0.000 solvercreatesnes
  0.000 discretizationcreatejacobian
  0.000 solversetsnesoptions
  0.000 hdf5readregionfromfile (0.000%
    0.000 gridcreatenaturaltoghostedh
    0.000 MPI_Comm_dup
    1534.813 MPI_File_open
    0.003 MPI_File_get_size
    0.000 MPI_Bcast
    42146.758 MPI_File_read_at
    0.000 MPI_Comm_free
    21994.593 hdf5maplocaltonaturali
    752.619 hdf5readintegerarray
    34634.035 MPI_File_close
  0.000 realizationlocalizeregions
  0.000 assignmaterialproptoregions
    0.000 readmaterialsfromfile
      0.000 discretizationcreatevecto
      0.000 hdf5readcellindexedinteg
        0.000 MPI_Comm_dup
        189.647 MPI_File_open
        0.000 MPI_File_get_size
        0.000 MPI_Bcast
        3710.115 MPI_File_read_at
        0.000 MPI_Comm_free
        2251.035 hdf5readindices
        609.844 hdf5readarray
        197.178 MPI_File_close
      0.000 discretizationglobaltoloca
      0.000 VecDestroy
    0.000 discretizationglobaltolocal
```

```
0.000        1.080e5 (0.995%)        1.085e7
0.000        7.146e4 (66.159%)       1.080e5
0.000        100.000                 100.000
6.251        8.724 +/- 0.755         10.738
```

MPI File I/O reading input dataset amounts to 1% of total time

PFLOTRAN jugene@smp8192 trace analysis



Cube 3.3 QT: epik_pflotran_smp8192_trace_PAPI_FP_OPS/trace+HWC.cube.gz <@jugene3>

```
0.000 Time
  9183176.639 Execution
    144156.735 MPI
      0.000 Synchronization
        13782.400 Collective
        0.000 Remote Memory
      0.000 Communication
        319289.889 Point-to-p
        867173.275 Collective
        0.000 Remote Memory
      71464.966 File I/O
        36555.672 Collective
      54467.230 Init/Exit
  164562.704 Overhead
8.308e9 Visits
6840320 Synchronizations
2.116e9 Communications
2.194e15 Bytes transferred
0 MPI file operations
  8 Individual
    1385528 Reads
    0 Writes
  131072 Collective
    0 Reads
    0 Writes
507926.494 Computational imbala
1.321e15 PAPI_FP_OPS
```

```
65.373 init (100.000%)
  80.584 realizationcreatediscretization
    0.000 discretizationcreatedms
      0.000 discretizationcreatedm
        0.000 structuredgridcreateda
          0.000 dacreate3d
            63.024 DACreate3d (10
              0.000 DACreate
                0.000 PetscHeade
                  0.000 PetscCor
                    13106.965 M
  9.156 hdf5readregionfromfile (100.00
    319.463 MPI_Comm_dup
    0.000 MPI_File_open
    0.000 MPI_File_get_size
    0.000 MPI_Bcast
    0.000 MPI_File_read_at
    11.173 MPI_Comm_free
    0.000 hdf5maplocaltonaturalindices
    0.000 hdf5readintegerarray
    0.000 MPI_File_close
  0.000 assignmaterialproptoregions
    5.808 readmaterialsfromfile (100.00
      1.325 hdf5readcellindexedintege
        46.826 MPI_Comm_dup
        0.000 MPI_File_open
        0.000 MPI_File_get_size
        0.000 MPI_Bcast
        0.000 MPI_File_read_at
        1.612 MPI_Comm_free
        0.000 hdf5readindices
        19.478 hdf5readarray
        0.000 MPI_File_close
```

```
0.000        1.442e5 (1.328%)        1.085e7
0.000        1.311e4 (9.092%)        1.442e5
0.000        100.000                 100.000
0.003        1.600 +/- 0.353         1.824
```

Selected "13106.965 MPI_Comm_dup"

MPI_Comm_dup calls inside PETSc & HDF5 require >1% of time

PFLOTRAN on BlueGene/P experience

- Initialization phase dominates at larger scales
  - 10% of total execution time spent duplicating communicators with 128k processes on Cray XT5
  - otherwise collective MPI File I/O relatively efficient
  - typically amortized in long simulation runs
- Solver scaled well to 64k processes before degrading
  - similar computation/communication patterns in FLOW & TRAN
    - callpath profiles distinguish costs
    - MPI_Allreduce collective communication becomes a bottleneck
    - communication overhead explodes for smaller FLOW problem
      - TRAN problem is 15x larger due to 15 chemical species
  - inactive processes induce clear computational imbalance
    - and are associated with large amounts of MPI waiting time
    - however, they constitute a relatively small minority

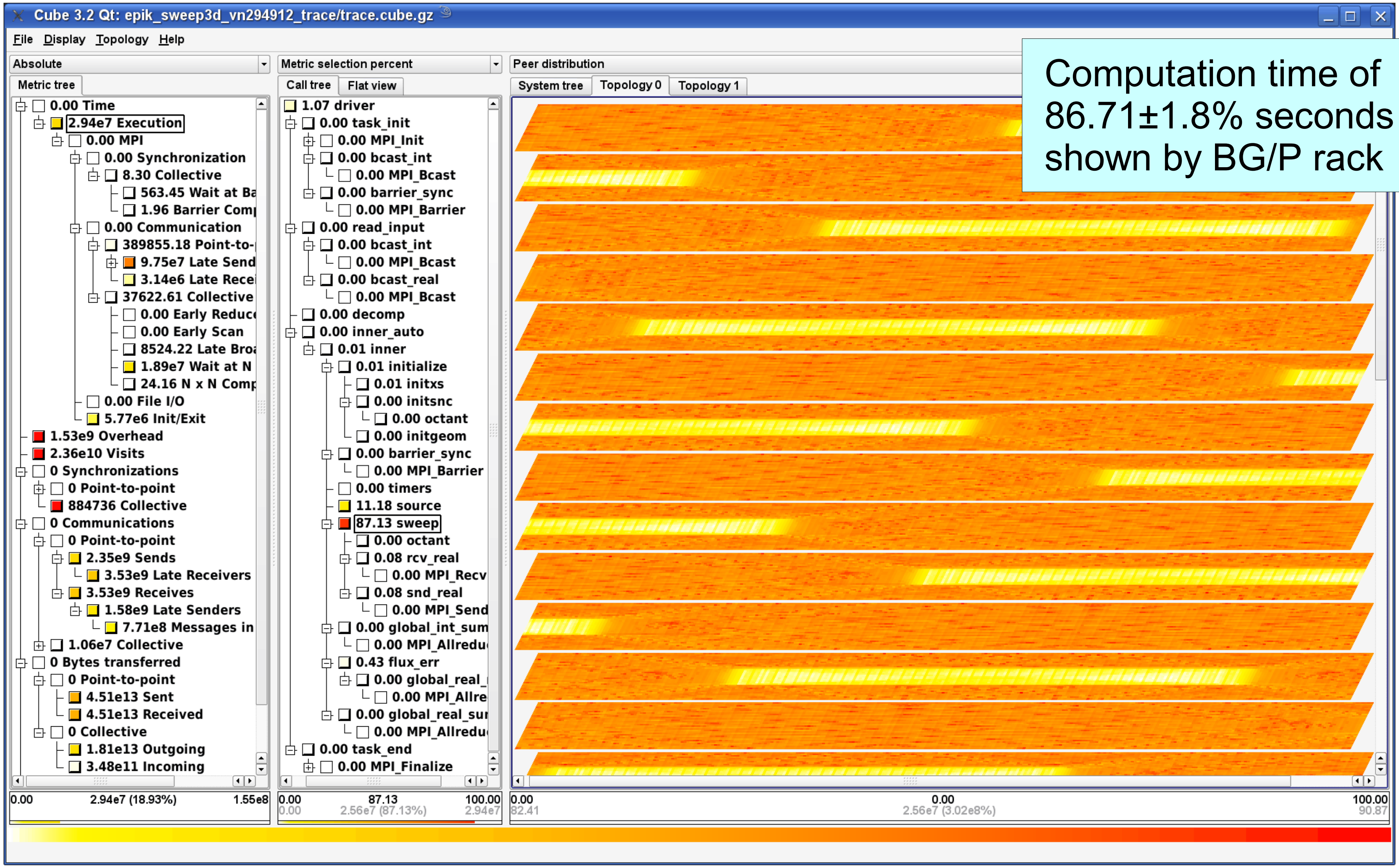
Proc. 53rd Cray User Group (Fairbanks, 2011)

Sweep3d on BlueGene/P case study

- 3D neutron transport simulation
  - ASC benchmark
  - direct order solve uses diagonal sweeps through grid cells
  - 'fixups' applied to correct unphysical (negative) fluxes
- 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
  - 7.6TB trace written in 17 minutes, analyzed in 10 minutes
    - of which 10 minutes for SIONlib open/create of 576 physical files
    - (compared to 86 minutes just to open/create a file per MPI rank)
  - Mapping of metrics onto application's 576x512 process grid reveals regular pattern of performance artifacts
    - computational imbalance originates from 'fixup' calculations
    - combined with diagonal wavefront sweeps amplifies waiting times

Proc. IPDPS Workshop on Large-Scale Parallel Processing (2010)

sweep3d on jugene@288k trace analysis



```
Cube 3.2 Qt: epik_sweep3d_vn294912_trace/trace.cube.gz
File  Display  Topology  Help
Absolute                 Metric selection percent        Peer distribution
Metric tree              Call tree | Flat view           System tree | Topology 0 | Topology 1
```

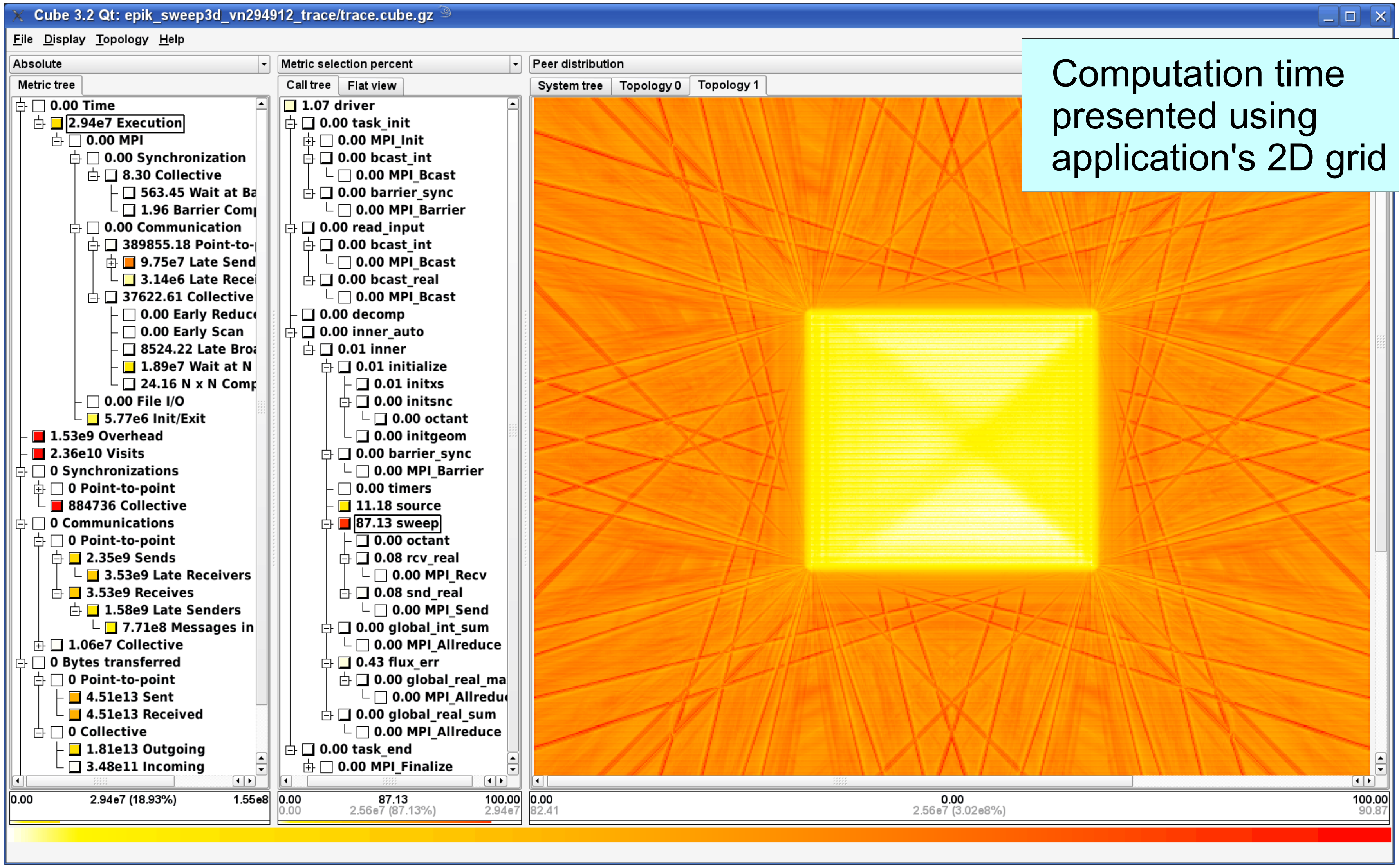
```
0.00 Time
  2.94e7 Execution
    0.00 MPI
      0.00 Synchronization
        8.30 Collective
          563.45 Wait at Ba
          1.96 Barrier Comp
      0.00 Communication
        389855.18 Point-to-
          9.75e7 Late Send
          3.14e6 Late Recei
        37622.61 Collective
          0.00 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                      0.00                        100.00
                                    | 0.00    2.56e7 (87.13%)   2.94e7 | 82.41              2.56e7 (3.02e8%)                   90.87
```

Computation time of 86.71±1.8% seconds shown by BG/P rack

sweep3d on jugene@288k trace analysis



Cube 3.2 Qt: epik_sweep3d_vn294912_trace/trace.cube.gz

File Display Topology Help

Absolute | Metric selection percent | Peer distribution

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1

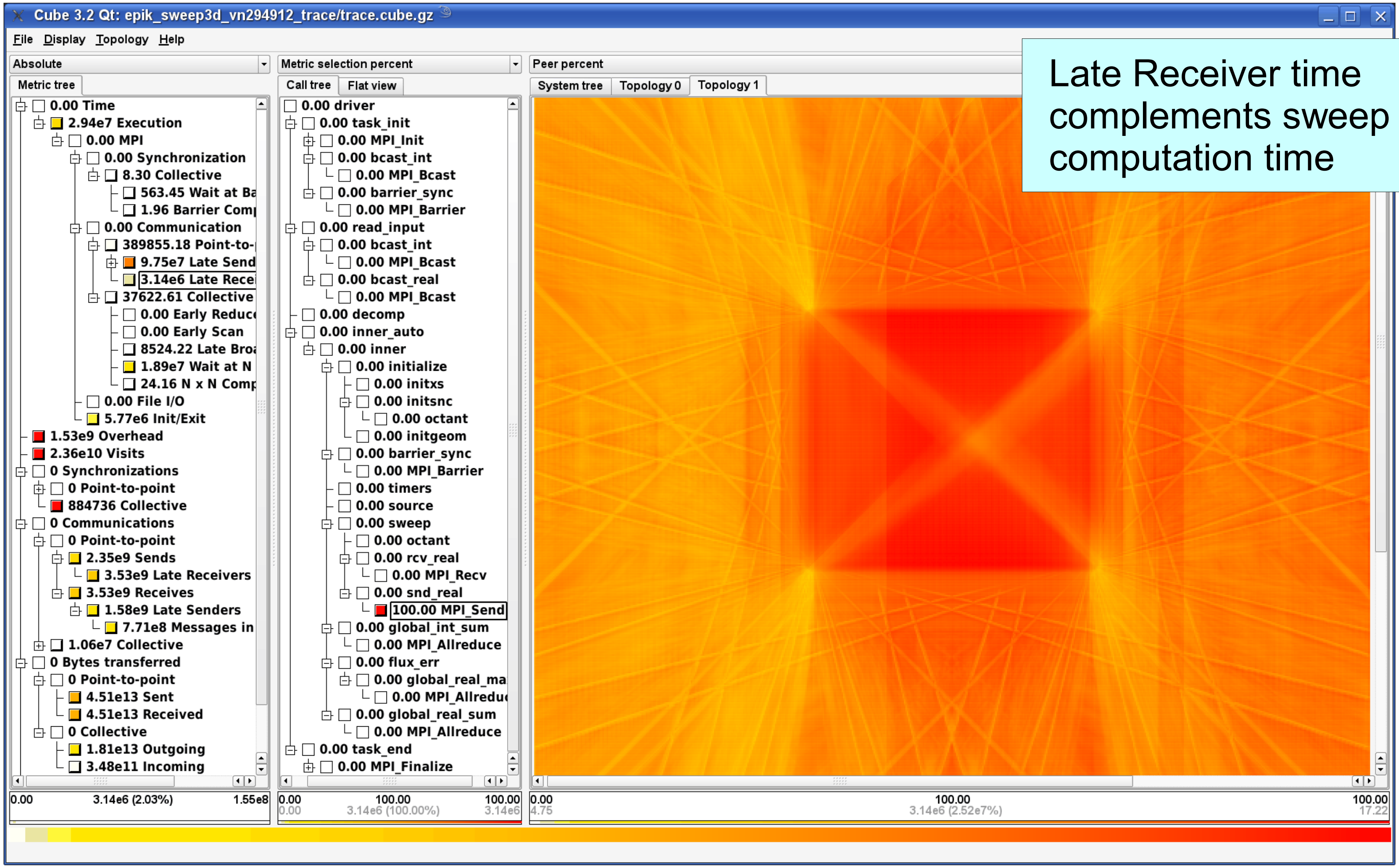
```
0.00 Time
  2.94e7 Execution
    0.00 MPI
      0.00 Synchronization
        8.30 Collective
          563.45 Wait at Ba
          1.96 Barrier Comp
      0.00 Communication
        389855.18 Point-to-
          9.75e7 Late Send
          3.14e6 Late Recei
        37622.61 Collective
          0.00 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 presented using application's 2D grid

sweep3d on jugene@288k trace (wait) analysis



Cube 3.2 Qt: epik_sweep3d_vn294912_trace/trace.cube.gz

File Display Topology Help

Absolute | Metric selection percent | Peer percent

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1

```
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 Rece
        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
```

Pane	Min	Selected	Max
Metric tree	0.00	3.14e6 (2.03%)	1.55e8
Call tree	0.00 / 0.00	100.00 / 3.14e6 (100.00%)	100.00 / 3.14e6
Topology	0.00 / 4.75	100.00 / 3.14e6 (2.52e7%)	100.00 / 17.22

Late Receiver time complements sweep computation time

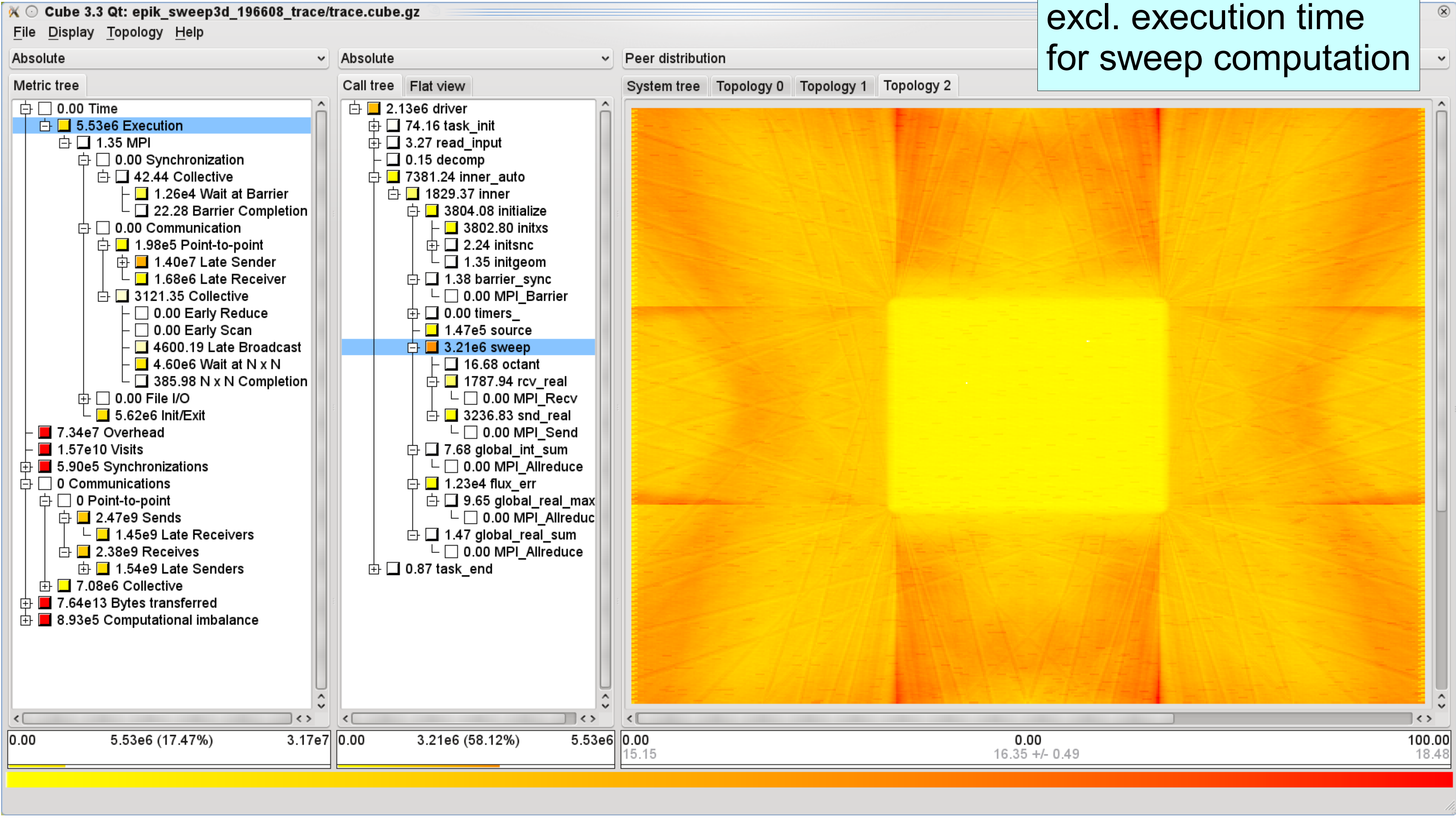
Sweep3d on Cray XT5 case study

- 3D neutron transport simulation
  - ASC benchmark
  - direct order solve uses diagonal sweeps through grid cells
  - 'fixups' applied to correct unphysical (negative) fluxes
- MPI parallel version 2.2b using 2D domain decomposition
  - ~2,000 lines (12 source modules), mostly Fortran77
- Run on Cray XT5 with 192k processes
  - 0.5TB trace written in 10 minutes, analyzed in 4 minutes
    - 6 minutes to open/create trace file for each rank
    - 25s for timestamp correction, 93s for parallel event replay
  - Mapping of metrics onto application's 512x384 process grid reveals regular pattern of performance artifacts
    - computational imbalance originates from 'fixup' calculations
    - combined with diagonal wavefront sweeps amplifies waiting times

Parallel Processing Letters 20(4):397-414 (2010)

sweep3d on jaguar@192k trace analysis

Regular imbalance in excl. execution time for sweep computation



Cube 3.3 Qt: epik_sweep3d_196608_trace/trace.cube.gz

File Display Topology Help

Absolute | Absolute | Peer distribution

Metric tree | Call tree | Flat view | System tree | Topology 0 | Topology 1 | Topology 2

```
0.00 Time
  5.53e6 Execution
    1.35 MPI
      0.00 Synchronization
        42.44 Collective
          1.26e4 Wait at Barrier
          22.28 Barrier Completion
      0.00 Communication
        1.98e5 Point-to-point
          1.40e7 Late Sender
          1.68e6 Late Receiver
        3121.35 Collective
          0.00 Early Reduce
          0.00 Early Scan
          4600.19 Late Broadcast
          4.60e6 Wait at N x N
          385.98 N x N Completion
      0.00 File I/O
    5.62e6 Init/Exit
7.34e7 Overhead
1.57e10 Visits
5.90e5 Synchronizations
0 Communications
  0 Point-to-point
    2.47e9 Sends
      1.45e9 Late Receivers
    2.38e9 Receives
      1.54e9 Late Senders
  7.08e6 Collective
7.64e13 Bytes transferred
8.93e5 Computational imbalance
```

```
2.13e6 driver
  74.16 task_init
  3.27 read_input
  0.15 decomp
  7381.24 inner_auto
    1829.37 inner
      3804.08 initialize
        3802.80 initxs
        2.24 initsnc
        1.35 initgeom
      1.38 barrier_sync
        0.00 MPI_Barrier
      0.00 timers_
      1.47e5 source
      3.21e6 sweep
        16.68 octant
        1787.94 rcv_real
          0.00 MPI_Recv
        3236.83 snd_real
          0.00 MPI_Send
      7.68 global_int_sum
        0.00 MPI_Allreduce
      1.23e4 flux_err
        9.65 global_real_max
          0.00 MPI_Allreduc
      1.47 global_real_sum
        0.00 MPI_Allreduce
  0.87 task_end
```

```
0.00        5.53e6 (17.47%)      3.17e7
0.00        3.21e6 (58.12%)      5.53e6
0.00              0.00            100.00
15.15        16.35 +/- 0.49        18.48
```

Acknowledgements

- The application and benchmark developers who generously 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, HLRN, HLRS, JSC, KSL, KTH, LRZ, NCAR, NCCS, NICS, RWTH, RZG, SARA, TACC, ZIH
    - Access & usage supported by European Union, German and other national funding organizations
- The Scalasca users for their comprehensive problem reports and improvement requests
  - as well as sharing reports of their analysis & tuning successes
- The Scalasca development team

Scalable performance **a**nalysis of **l**arge-**sc**ale parallel **a**pplications

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