GOT BOTTLENECKS?
SCORE-P AND FRIENDS TO THE RESCUE!

2025-03-26 I MARKUS GEIMER

ATML PARALLEL PERFORMANCE

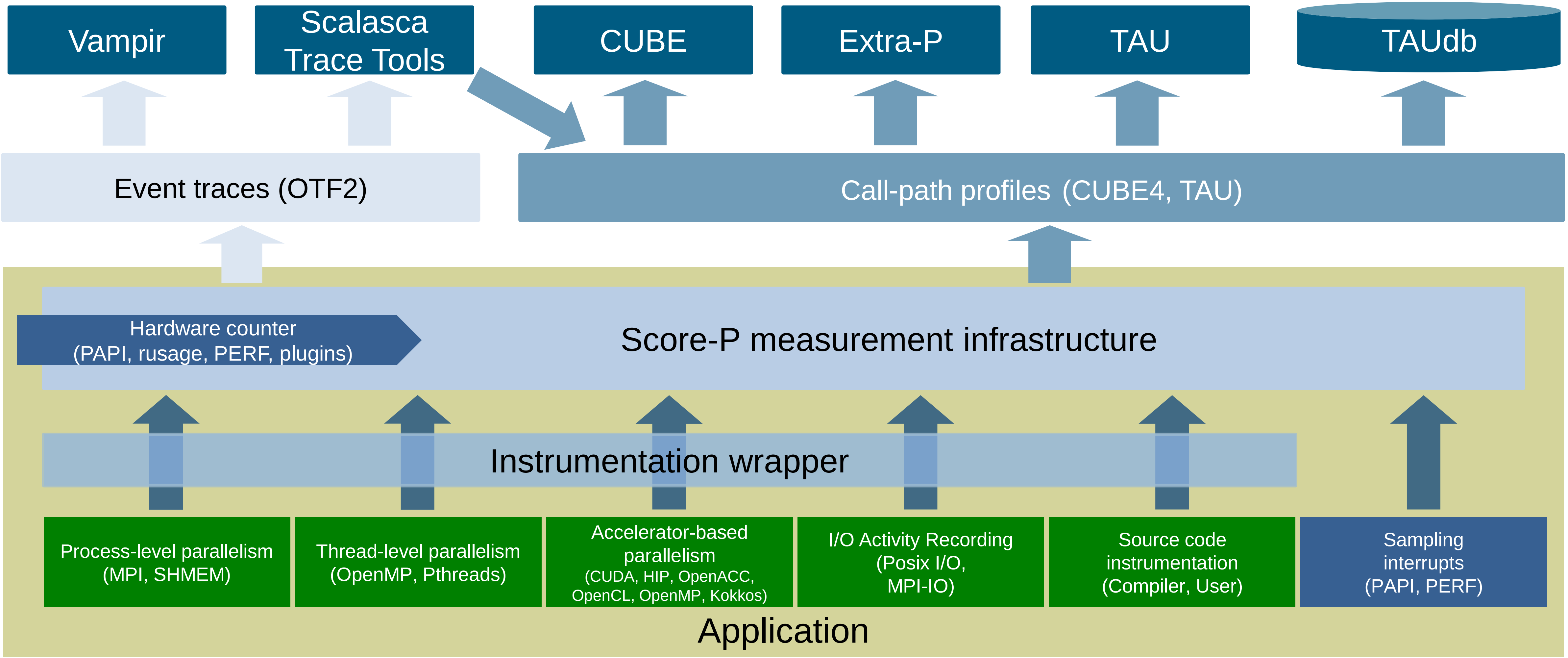
Mitglied der Helmholtz-Gemeinschaft

JÜLICH Forschungszentrum | JÜLICH SUPERCOMPUTING CENTRE

DOI 10.5281/zenodo.1240731

- Infrastructure for instrumentation and performance measurements of HPC applications
  - Open Source (3-clause BSD license)
  - Portable to all major HPC platforms
  - Proven scalability
  - Active, community-driven development by multiple partners
    - We maintain EasyBuild recipes… of course! (+ Spack & Fedora/OpenHPC pkgs)
- Instrumented applications can be used to produce different results:
  - Call-path profiles (runtime summaries)
  - Event traces
- Measurement configuration via environment variables
  - Instrument once, collect multiple measurements

SCORE-P TOOLS ECOSYSTEM



Vampir

Scalasca Trace Tools

CUBE

Extra-P

TAU

TAUdb

Event traces (OTF2)

Call-path profiles (CUBE4, TAU)

Hardware counter (PAPI, rusage, PERF, plugins)

Score-P measurement infrastructure

Instrumentation wrapper

Process-level parallelism (MPI, SHMEM)

Thread-level parallelism (OpenMP, Pthreads)

Accelerator-based parallelism (CUDA, HIP, OpenACC, OpenCL, OpenMP, Kokkos)

I/O Activity Recording (Posix I/O, MPI-IO)

Source code instrumentation (Compiler, User)

Sampling interrupts (PAPI, PERF)

Application

SCORE-P TOOLS ECOSYSTEM

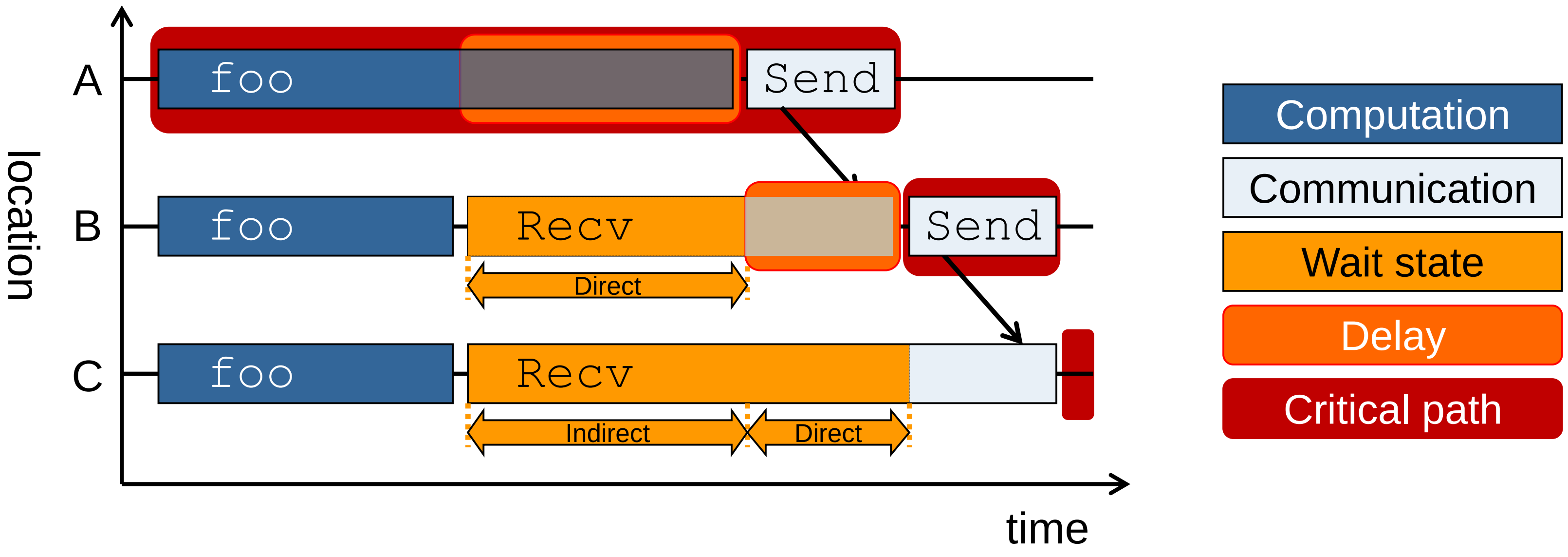
- Vampir (commercial)
  - Interactive trace data visualization & analysis
- Scalasca Trace Tools
  - Automatic event trace analysis
- Cube
  - Interactive performance report exploration
- Extra-P
  - Empirical performance modeling
- TAU / TAUdb
  - Interactive performance report exploration & data mining
  - Additional instrumentation/measurement capabilities (using Score-P as backend)

trace tools
scalasca

DOI 10.5281/zenodo.4103922

- Scalable, trace-based performance analysis toolset
  - Open Source (3-clause BSD license)
  - Portable to all major HPC platforms
  - Proven scalability
- Main features:
  - Automatic search for inefficiencies (i.e., wait states)
  - Identification of delays as their root causes
  - Identification of the critical-path profile
- Current focus: MPI, OpenMP (host only), POSIX threads (limited)
  - Analysis of accelerator device activities not yet supported, but still useful for host-side code (with some care)

SCALASCA TRACE ANALYSIS

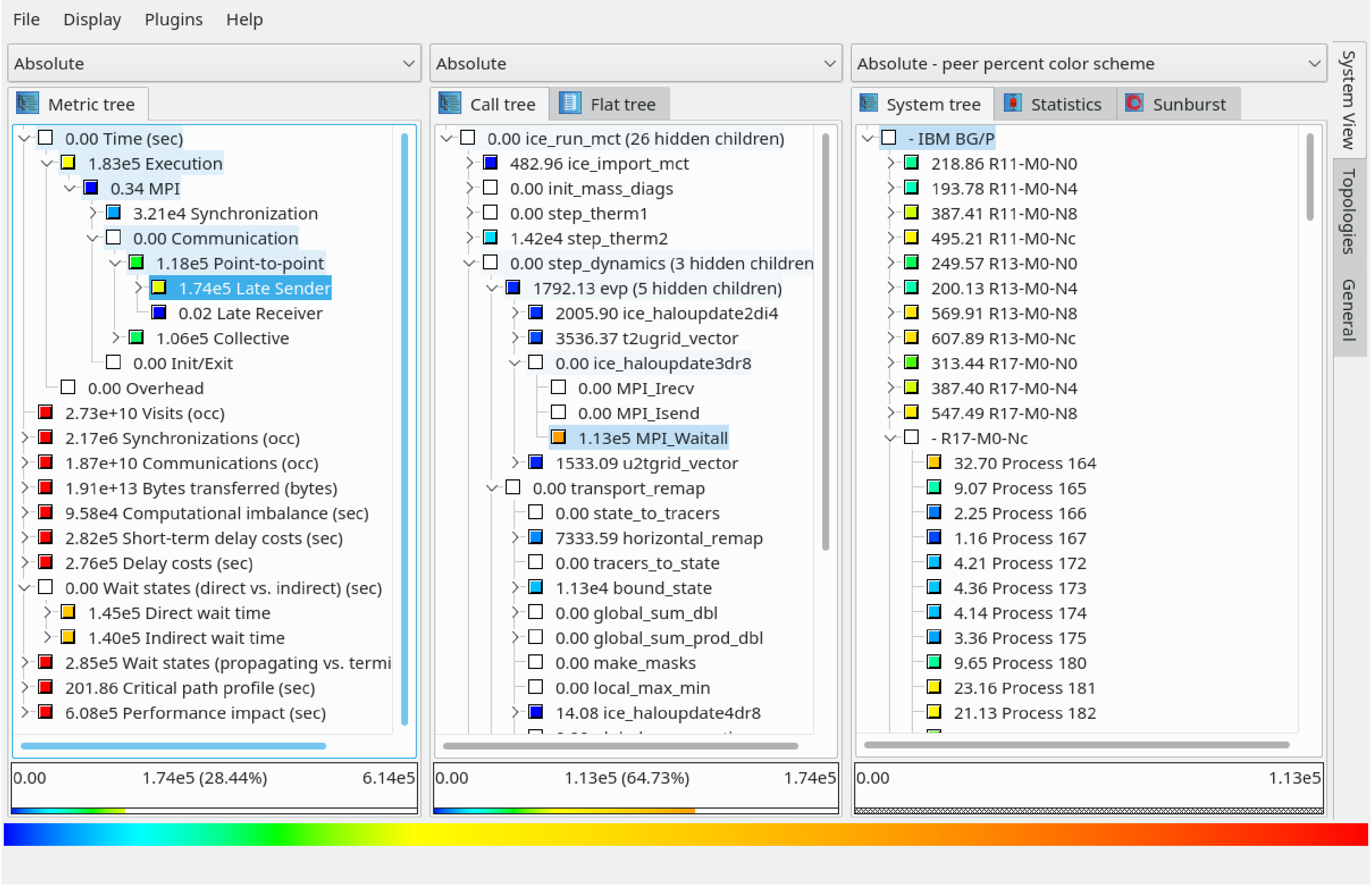


- Identifies wait states (here: 2 *“Late Sender”* situations)
- Classifies wait states into direct and indirect (i.e., caused by other wait states)
- Identifies *delays* (excess computation/communication) as root causes of wait states
- Attributes wait states as *delay costs*
- Calculates the *critical-path profile*

EXAMPLE: CESM SEA ICE MODULE

Late Sender Analysis

- Finds waiting time at `MPI_Waitall` inside ice boundary halo update
- Shows distribution of imbalance across system and ranks



Absolute | Absolute | Absolute - peer percent color scheme

```
0.00 Time (sec)
  1.83e5 Execution
    0.34 MPI
      3.21e4 Synchronization
      0.00 Communication
        1.18e5 Point-to-point
          1.74e5 Late Sender
          0.02 Late Receiver
        1.06e5 Collective
      0.00 Init/Exit
  0.00 Overhead
2.73e+10 Visits (occ)
2.17e6 Synchronizations (occ)
1.87e+10 Communications (occ)
1.91e+13 Bytes transferred (bytes)
9.58e4 Computational imbalance (sec)
2.82e5 Short-term delay costs (sec)
2.76e5 Delay costs (sec)
0.00 Wait states (direct vs. indirect) (sec)
  1.45e5 Direct wait time
  1.40e5 Indirect wait time
2.85e5 Wait states (propagating vs. termi
201.86 Critical path profile (sec)
6.08e5 Performance impact (sec)
```

0.00 1.74e5 (28.44%) 6.14e5

```
0.00 ice_run_mct (26 hidden children)
  482.96 ice_import_mct
  0.00 init_mass_diags
  0.00 step_therm1
  1.42e4 step_therm2
  0.00 step_dynamics (3 hidden children
    1792.13 evp (5 hidden children)
      2005.90 ice_haloupdate2di4
      3536.37 t2ugrid_vector
      0.00 ice_haloupdate3dr8
        0.00 MPI_Irecv
        0.00 MPI_Isend
        1.13e5 MPI_Waitall
      1533.09 u2tgrid_vector
    0.00 transport_remap
      0.00 state_to_tracers
      7333.59 horizontal_remap
      0.00 tracers_to_state
      1.13e4 bound_state
      0.00 global_sum_dbl
      0.00 global_sum_prod_dbl
      0.00 make_masks
      0.00 local_max_min
      14.08 ice_haloupdate4dr8
```

0.00 1.13e5 (64.73%) 1.74e5

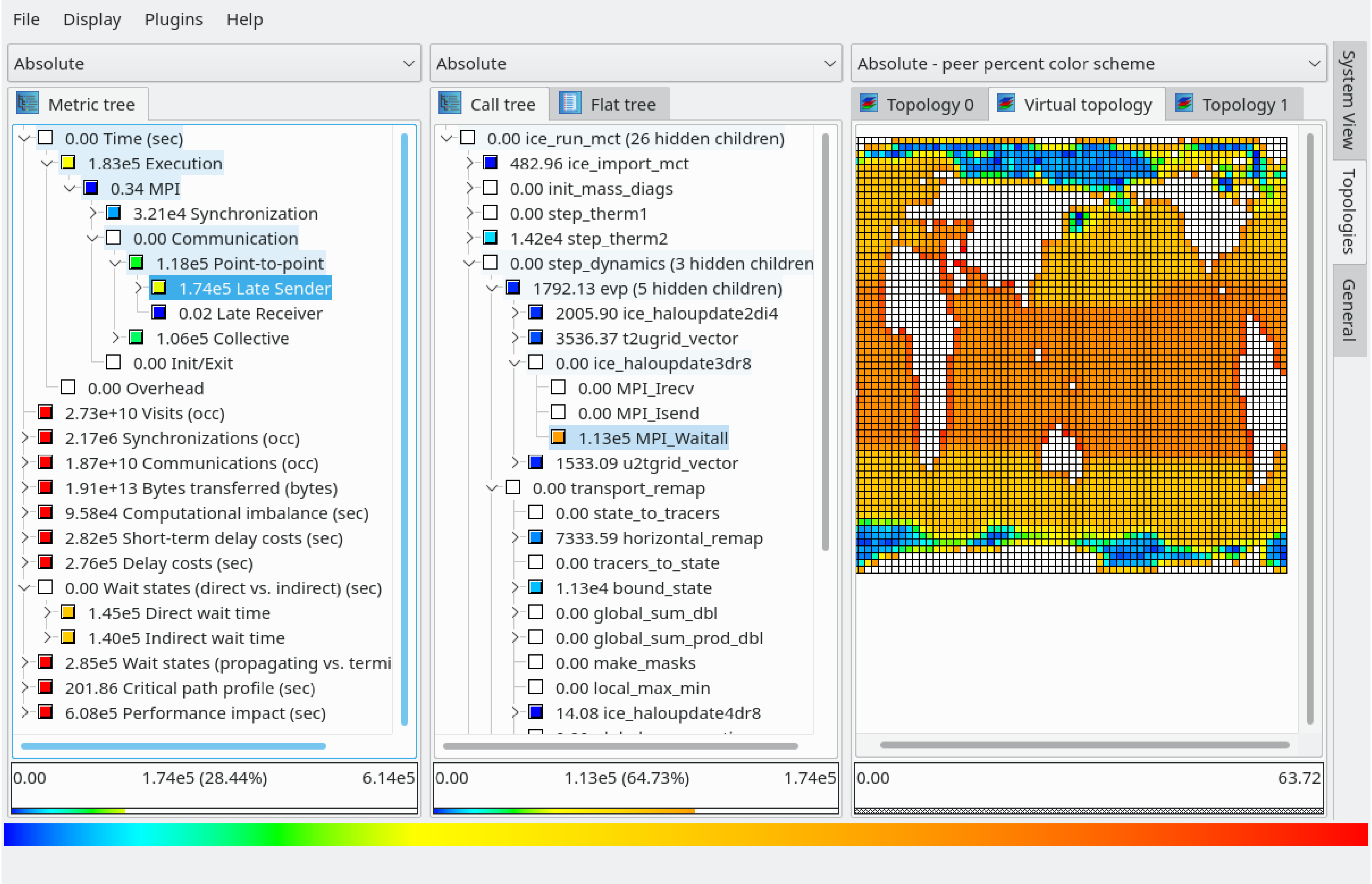
```
- IBM BG/P
  218.86 R11-M0-N0
  193.78 R11-M0-N4
  387.41 R11-M0-N8
  495.21 R11-M0-Nc
  249.57 R13-M0-N0
  200.13 R13-M0-N4
  569.91 R13-M0-N8
  607.89 R13-M0-Nc
  313.44 R17-M0-N0
  387.40 R17-M0-N4
  547.49 R17-M0-N8
  - R17-M0-Nc
    32.70 Process 164
    9.07 Process 165
    2.25 Process 166
    1.16 Process 167
    4.21 Process 172
    4.36 Process 173
    4.14 Process 174
    3.36 Process 175
    9.65 Process 180
    23.16 Process 181
    21.13 Process 182
```

0.00 1.13e5

EXAMPLE: CESM SEA ICE MODULE

Late Sender Analysis + Application Topology

- Shows distribution of imbalance over topology
- MPI topologies are automatically captured



Absolute | Absolute | Absolute - peer percent color scheme

- 0.00 Time (sec)
  - 1.83e5 Execution
    - 0.34 MPI
      - 3.21e4 Synchronization
      - 0.00 Communication
        - 1.18e5 Point-to-point
          - 1.74e5 Late Sender
          - 0.02 Late Receiver
        - 1.06e5 Collective
      - 0.00 Init/Exit
  - 0.00 Overhead
- 2.73e+10 Visits (occ)
- 2.17e6 Synchronizations (occ)
- 1.87e+10 Communications (occ)
- 1.91e+13 Bytes transferred (bytes)
- 9.58e4 Computational imbalance (sec)
- 2.82e5 Short-term delay costs (sec)
- 2.76e5 Delay costs (sec)
- 0.00 Wait states (direct vs. indirect) (sec)
  - 1.45e5 Direct wait time
  - 1.40e5 Indirect wait time
- 2.85e5 Wait states (propagating vs. termi
- 201.86 Critical path profile (sec)
- 6.08e5 Performance impact (sec)

0.00 1.74e5 (28.44%) 6.14e5

- 0.00 ice_run_mct (26 hidden children)
  - 482.96 ice_import_mct
  - 0.00 init_mass_diags
  - 0.00 step_therm1
  - 1.42e4 step_therm2
  - 0.00 step_dynamics (3 hidden children
    - 1792.13 evp (5 hidden children)
      - 2005.90 ice_haloupdate2di4
      - 3536.37 t2ugrid_vector
      - 0.00 ice_haloupdate3dr8
        - 0.00 MPI_Irecv
        - 0.00 MPI_Isend
        - 1.13e5 MPI_Waitall
      - 1533.09 u2tgrid_vector
    - 0.00 transport_remap
      - 0.00 state_to_tracers
      - 7333.59 horizontal_remap
      - 0.00 tracers_to_state
      - 1.13e4 bound_state
      - 0.00 global_sum_dbl
      - 0.00 global_sum_prod_dbl
      - 0.00 make_masks
      - 0.00 local_max_min
      - 14.08 ice_haloupdate4dr8

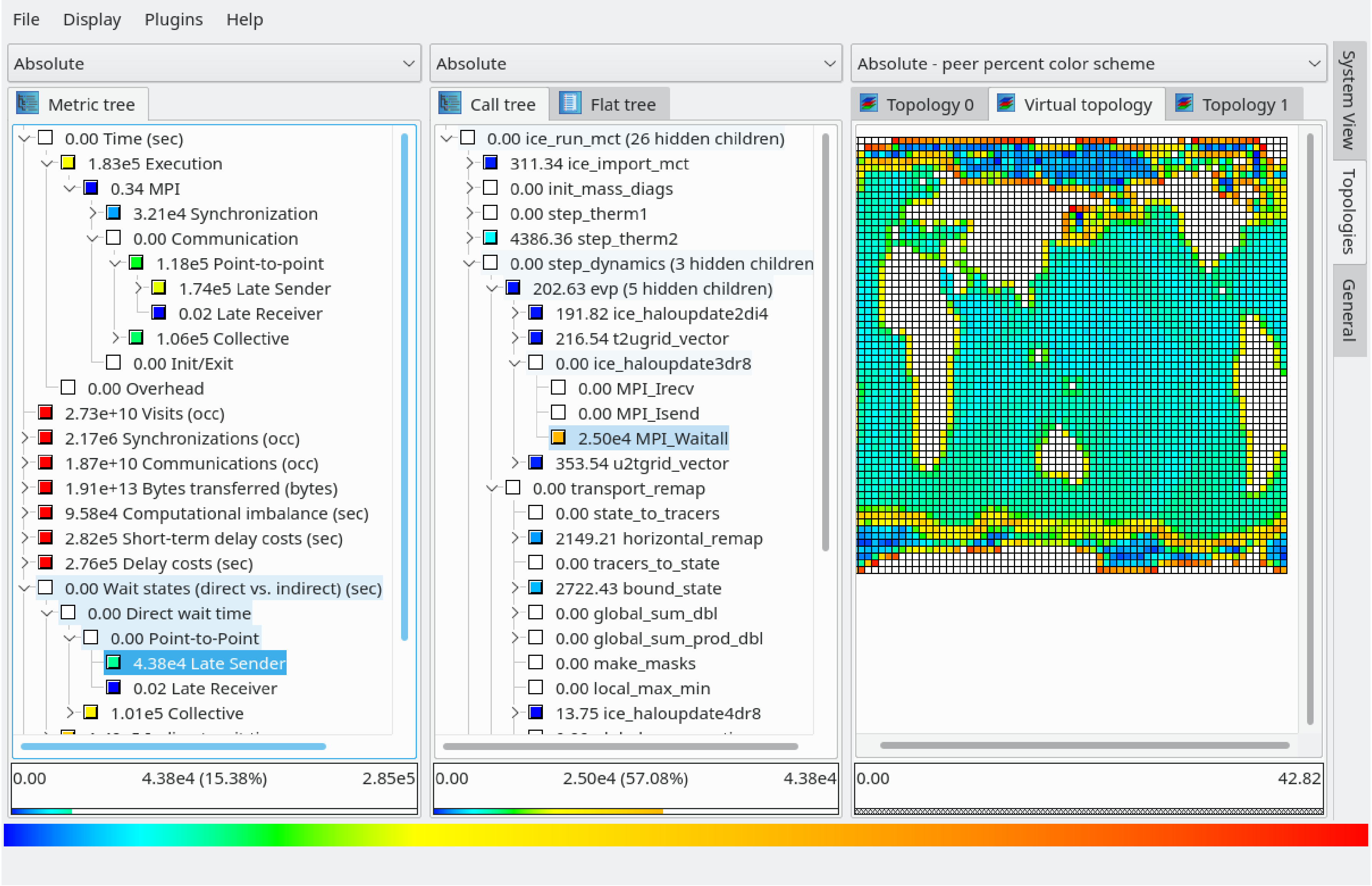
0.00 1.13e5 (64.73%) 1.74e5

0.00 63.72

EXAMPLE: CESM SEA ICE MODULE

Direct *Late Sender* Wait Time Analysis

- Direct wait states caused by ranks processing areas near the north and south ice borders



```
v □ 0.00 Time (sec)
  v 🟨 1.83e5 Execution
    v 🟦 0.34 MPI
      > 🟦 3.21e4 Synchronization
      v □ 0.00 Communication
        v 🟩 1.18e5 Point-to-point
          > 🟨 1.74e5 Late Sender
            🟦 0.02 Late Receiver
        > 🟩 1.06e5 Collective
      □ 0.00 Init/Exit
  □ 0.00 Overhead
  🟥 2.73e+10 Visits (occ)
> 🟥 2.17e6 Synchronizations (occ)
> 🟥 1.87e+10 Communications (occ)
> 🟥 1.91e+13 Bytes transferred (bytes)
> 🟥 9.58e4 Computational imbalance (sec)
> 🟥 2.82e5 Short-term delay costs (sec)
> 🟥 2.76e5 Delay costs (sec)
v □ 0.00 Wait states (direct vs. indirect) (sec)
  v □ 0.00 Direct wait time
    v □ 0.00 Point-to-Point
        🟩 4.38e4 Late Sender
        🟦 0.02 Late Receiver
    > 🟨 1.01e5 Collective
```

```
v □ 0.00 ice_run_mct (26 hidden children)
  > 🟦 311.34 ice_import_mct
  > □ 0.00 init_mass_diags
  > □ 0.00 step_therm1
  > 🟦 4386.36 step_therm2
  v □ 0.00 step_dynamics (3 hidden children
    v 🟦 202.63 evp (5 hidden children)
      > 🟦 191.82 ice_haloupdate2di4
      > 🟦 216.54 t2ugrid_vector
      v □ 0.00 ice_haloupdate3dr8
          □ 0.00 MPI_Irecv
          □ 0.00 MPI_Isend
          🟧 2.50e4 MPI_Waitall
      > 🟦 353.54 u2tgrid_vector
    v □ 0.00 transport_remap
        □ 0.00 state_to_tracers
      > 🟦 2149.21 horizontal_remap
        □ 0.00 tracers_to_state
      > 🟦 2722.43 bound_state
      > □ 0.00 global_sum_dbl
      > □ 0.00 global_sum_prod_dbl
        □ 0.00 make_masks
        □ 0.00 local_max_min
      > 🟦 13.75 ice_haloupdate4dr8
```

0.00 4.38e4 (15.38%) 2.85e5

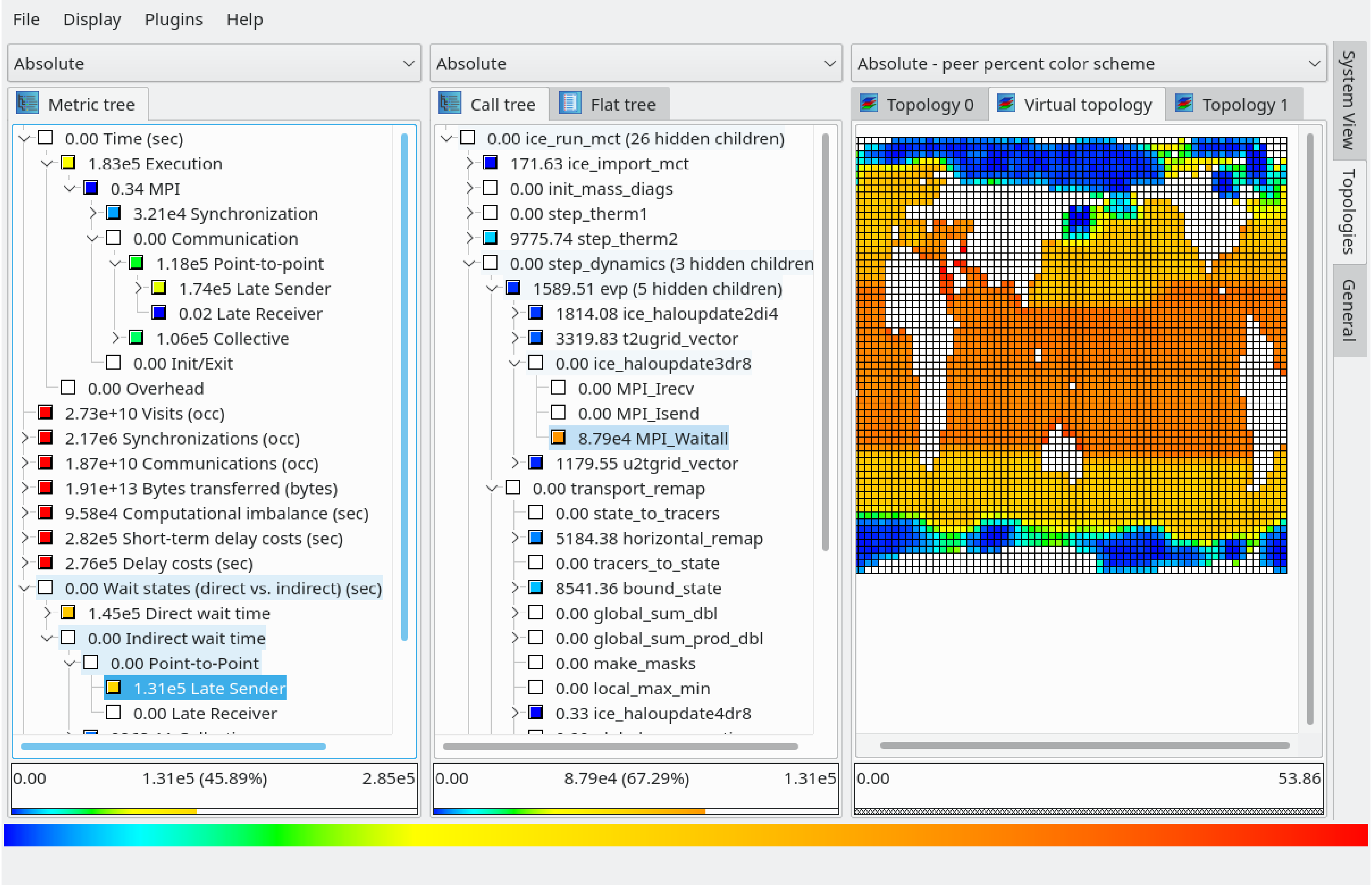
0.00 2.50e4 (57.08%) 4.38e4

0.00 42.82

EXAMPLE: CESM SEA ICE MODULE

Indirect *Late Sender* Wait Time Analysis

- Indirect wait states occur for ranks processing warmer areas



```
0.00 Time (sec)
  1.83e5 Execution
    0.34 MPI
      3.21e4 Synchronization
      0.00 Communication
        1.18e5 Point-to-point
          1.74e5 Late Sender
          0.02 Late Receiver
        1.06e5 Collective
      0.00 Init/Exit
  0.00 Overhead
2.73e+10 Visits (occ)
2.17e6 Synchronizations (occ)
1.87e+10 Communications (occ)
1.91e+13 Bytes transferred (bytes)
9.58e4 Computational imbalance (sec)
2.82e5 Short-term delay costs (sec)
2.76e5 Delay costs (sec)
0.00 Wait states (direct vs. indirect) (sec)
  1.45e5 Direct wait time
  0.00 Indirect wait time
    0.00 Point-to-Point
      1.31e5 Late Sender
      0.00 Late Receiver
```

```
0.00 ice_run_mct (26 hidden children)
  171.63 ice_import_mct
  0.00 init_mass_diags
  0.00 step_therm1
  9775.74 step_therm2
  0.00 step_dynamics (3 hidden children
    1589.51 evp (5 hidden children)
      1814.08 ice_haloupdate2di4
      3319.83 t2ugrid_vector
      0.00 ice_haloupdate3dr8
        0.00 MPI_Irecv
        0.00 MPI_Isend
        8.79e4 MPI_Waitall
      1179.55 u2tgrid_vector
    0.00 transport_remap
      0.00 state_to_tracers
      5184.38 horizontal_remap
      0.00 tracers_to_state
      8541.36 bound_state
      0.00 global_sum_dbl
      0.00 global_sum_prod_dbl
      0.00 make_masks
      0.00 local_max_min
      0.33 ice_haloupdate4dr8
```

0.00 1.31e5 (45.89%) 2.85e5

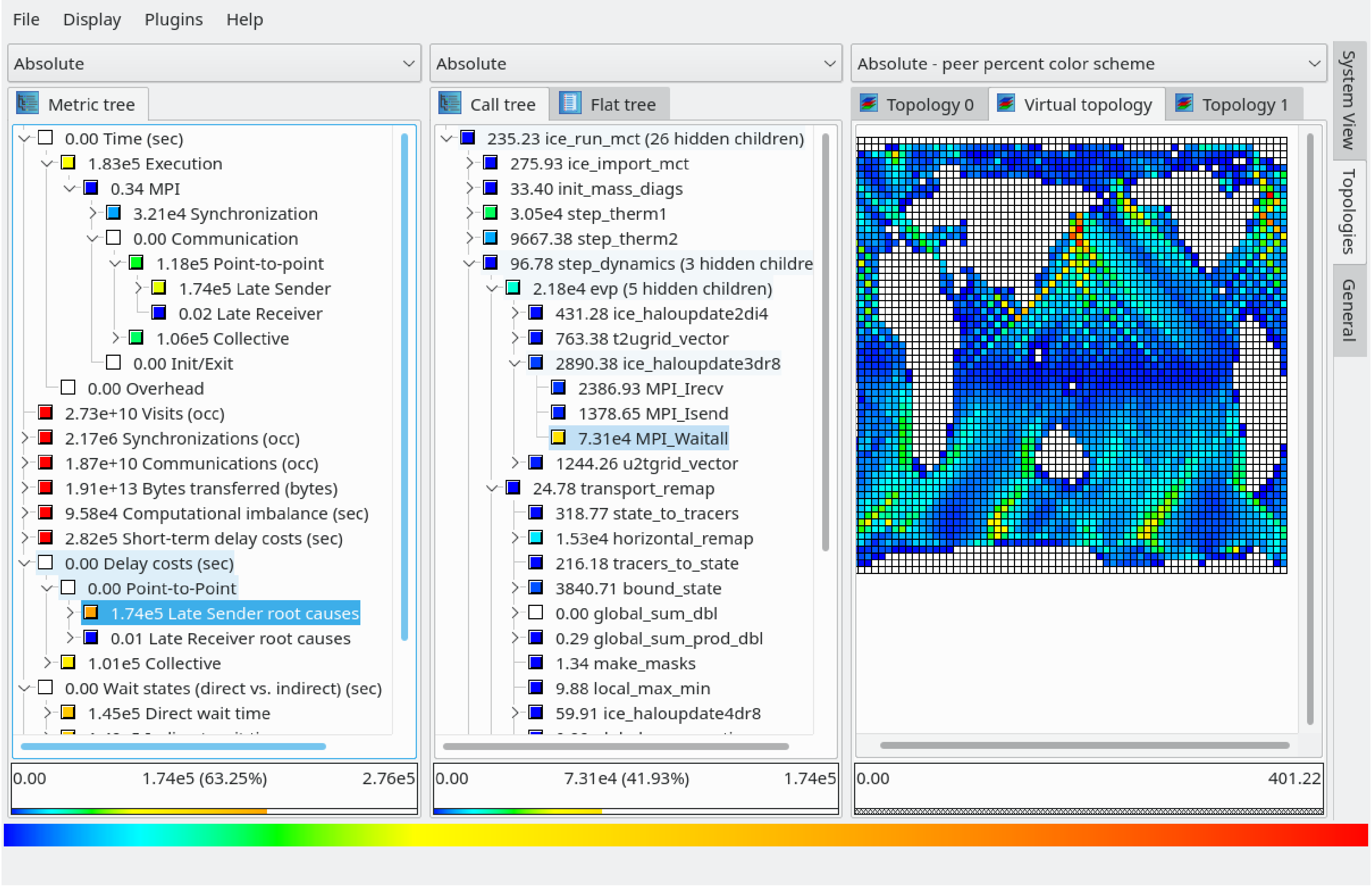
0.00 8.79e4 (67.29%) 1.31e5

0.00 53.86

EXAMPLE: CESM SEA ICE MODULE

Late Sender Delay Costs Analysis

- Delays **NOT** caused by ranks processing ice!



```
File   Display   Plugins   Help

Absolute                              Absolute                                   Absolute - peer percent color scheme
Metric tree                           Call tree   Flat tree                      Topology 0   Virtual topology   Topology 1

⌄ ☐ 0.00 Time (sec)
  ⌄ 🟨 1.83e5 Execution
    ⌄ 🟦 0.34 MPI
      › 🟦 3.21e4 Synchronization
      ⌄ ☐ 0.00 Communication
        ⌄ 🟩 1.18e5 Point-to-point
          › 🟨 1.74e5 Late Sender
            🟦 0.02 Late Receiver
        › 🟩 1.06e5 Collective
      ☐ 0.00 Init/Exit
  ☐ 0.00 Overhead
🟥 2.73e+10 Visits (occ)
› 🟥 2.17e6 Synchronizations (occ)
› 🟥 1.87e+10 Communications (occ)
› 🟥 1.91e+13 Bytes transferred (bytes)
› 🟥 9.58e4 Computational imbalance (sec)
› 🟥 2.82e5 Short-term delay costs (sec)
⌄ ☐ 0.00 Delay costs (sec)
  ⌄ ☐ 0.00 Point-to-Point
    › 🟧 1.74e5 Late Sender root causes
    › 🟦 0.01 Late Receiver root causes
  › 🟨 1.01e5 Collective
⌄ ☐ 0.00 Wait states (direct vs. indirect) (sec)
  › 🟨 1.45e5 Direct wait time

0.00          1.74e5 (63.25%)          2.76e5
```

```
⌄ 🟦 235.23 ice_run_mct (26 hidden children)
  › 🟦 275.93 ice_import_mct
  › 🟦 33.40 init_mass_diags
  › 🟩 3.05e4 step_therm1
  › 🟦 9667.38 step_therm2
  ⌄ 🟦 96.78 step_dynamics (3 hidden childre
    ⌄ 🟦 2.18e4 evp (5 hidden children)
      › 🟦 431.28 ice_haloupdate2di4
      › 🟦 763.38 t2ugrid_vector
      ⌄ 🟦 2890.38 ice_haloupdate3dr8
          🟦 2386.93 MPI_Irecv
          🟦 1378.65 MPI_Isend
          🟨 7.31e4 MPI_Waitall
      › 🟦 1244.26 u2tgrid_vector
    ⌄ 🟦 24.78 transport_remap
        🟦 318.77 state_to_tracers
      › 🟦 1.53e4 horizontal_remap
        🟦 216.18 tracers_to_state
      › 🟦 3840.71 bound_state
      › ☐ 0.00 global_sum_dbl
      › 🟦 0.29 global_sum_prod_dbl
        🟦 1.34 make_masks
        🟦 9.88 local_max_min
      › 🟦 59.91 ice_haloupdate4dr8

0.00          7.31e4 (41.93%)          1.74e5
```

```
0.00                                   401.22
```

QUESTIONS

?

Score-P

Scalable performance measurement infrastructure for parallel codes

- https://www.score-p.org
- support@score-p.org

trace tools

scalasca

- https://www.scalasca.org
- scalasca@fz-juelich.de