

International HPC Summer School 2019: Performance analysis and optimization

Case Studies

VI-HPS Team Ilya Zhukov – Jülich Supercomputing Centre



PENNANT

- PENNANT is an unstructured mesh physics mini-app designed for advanced architecture research
- contains mesh data structures and a few physics algorithms
- C++ application, supports MPI and OpenMP
- Developed by Los Alamos National Laboratory
- https://github.com/lanl/PENNANT

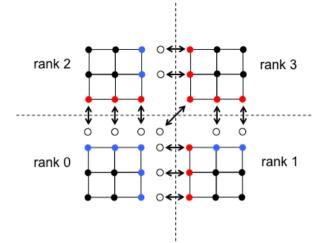
- Nohpoly testcase executed on Bridges (8 MPI x 4 OpenMP)
- Copy pennant_8x4_bridges.cubex to your livedvd

https://fz-juelich.sciebo.de/s/0qIEZzXpjS6uyte

Examine measurement with CUBE

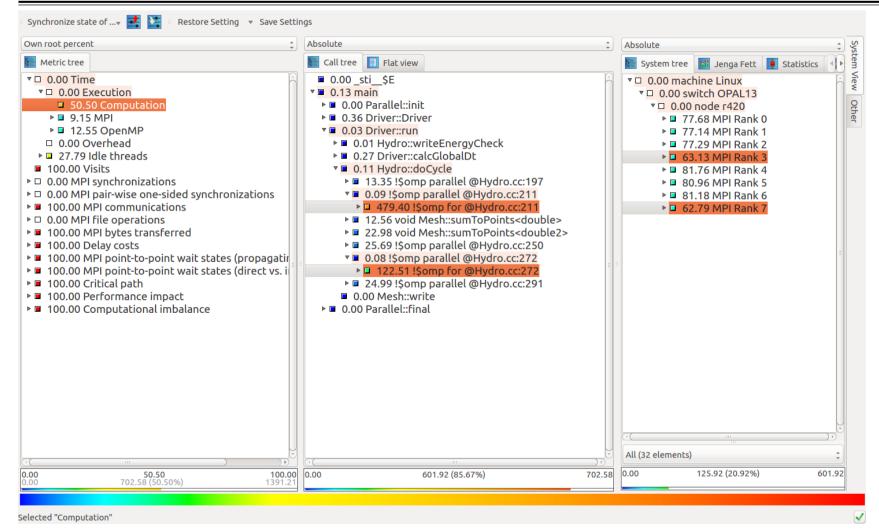
livedvd> cube pennant_8x4_bridges.cubex





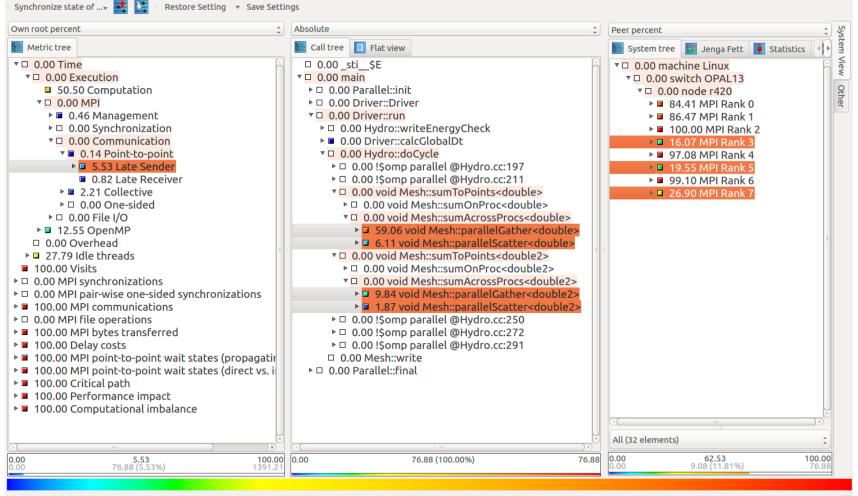
- In *Mesh::parallelGather*, slave (red) point values are assembled into messages, and sent to corresponding proxy points (white) on the same rank as their masters(blue).
- In *Mesh::parallelSum*, master points sum their own values and all proxy values, and store sum at master and all proxies.
- In *Mesh::parallelScatter*, the updated proxy point values are assembled into messages and sent back to their corresponding slave points.

PENNANT: execution breakdown



- Computation ~51%
- Communication ~9%
- OpenMP sync ~13%
- ~86% of computation spent in two OpenMP parallel regions
- Workload is not equally distributed

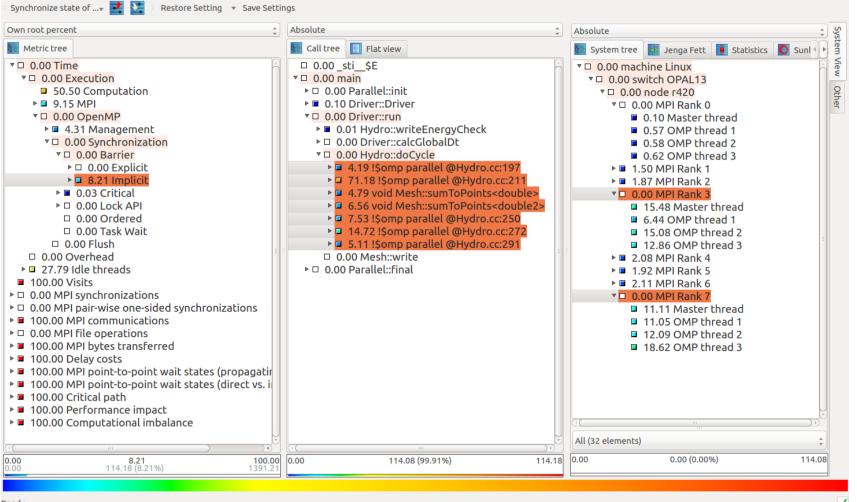
PENNANT: communication breakdown



- Late Sender ~6% in two routines parallelGather and parallelScatter
- Ranks 3, 5 and 7 have significantly smaller waiting time than others

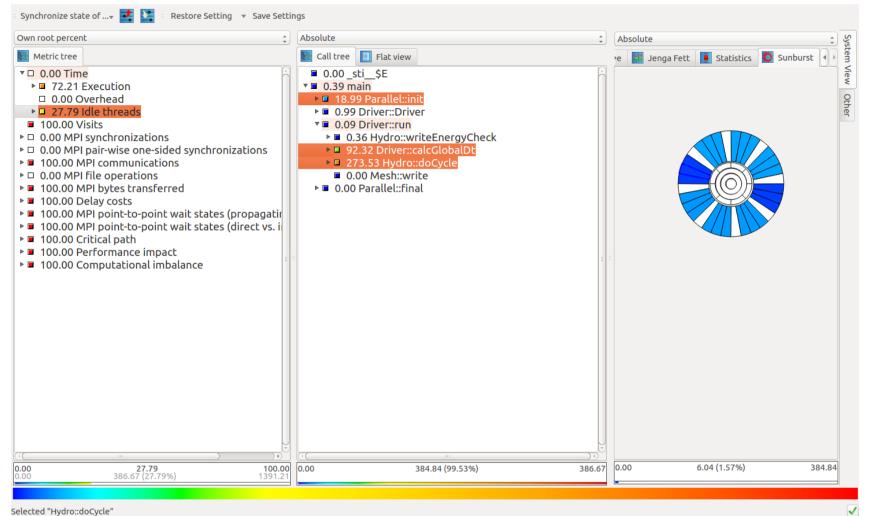
Selected "MPI Rank 5"

PENNANT: OpenMP sync



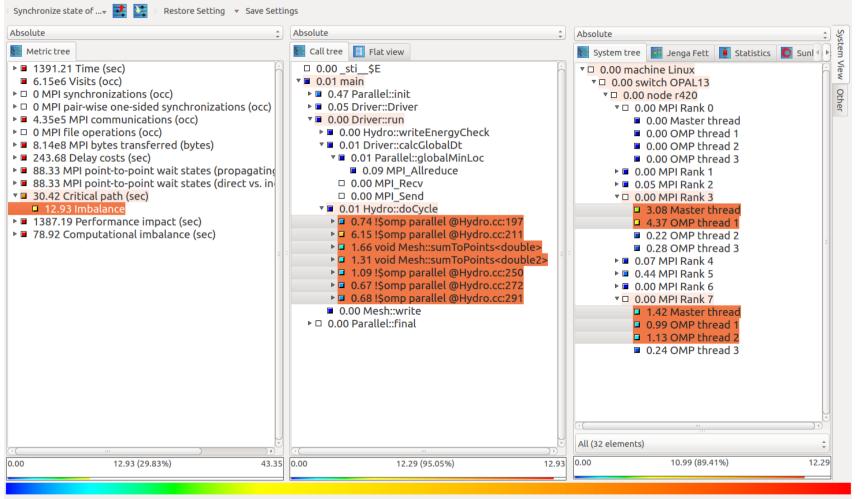
- OpenMP sync is spent in OpenMP implicit barriers
- Significant time on OpenMP barriers spent on rank 3 and 7
- ~86% of computation spent in two OpenMP parallel regions
- Workload is not equally distributed

PENNANT: idle threads



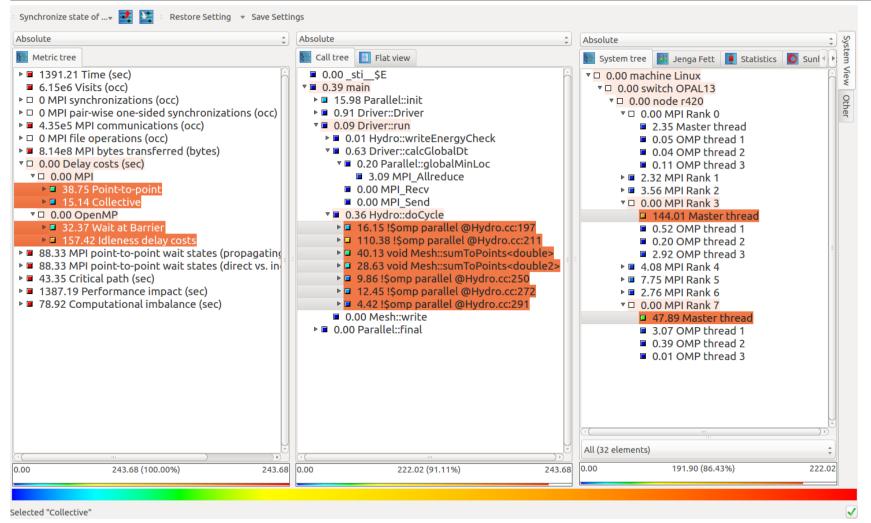
~28% of total runtime
 OpenMP threads are idling in
 three routines, i.e.
 Parallel:init,
 Driver::calcGlobalDt,
 Hydro::doCycle

PENNANT: critical path



 Selected routines of MPI ranks 3 and 5 show significant impact on critical path -> potential candidates for optimization

PENNANT: delay analysis



- Most of the delay caused by imbalanced ranks 3 and 5
- Consider decomposition scheme where load of ranks is balanced

VIRTUAL INSTITUTE - HIGH PRODUCTIVITY SUPERCOMPUTING

Terrestrial System Modeling Platform (TerrSysMP)

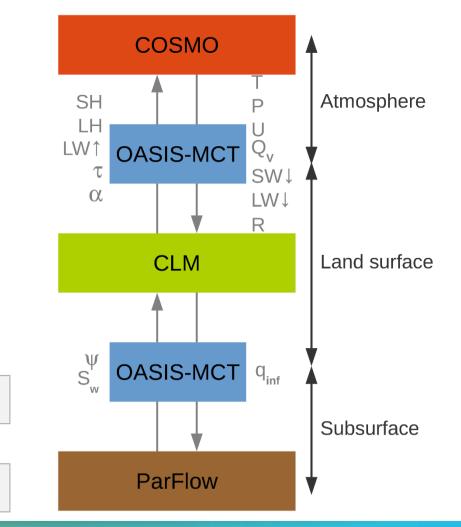
- TerrSysMP simulates the interaction between lateral flow processes in river basins with the lower atmospheric boundary layer
- MPMD: Multiple Program Multiple Data Execution Model
- Consists of three model components: COSMO, CLM , ParFlow and an external MPI-based coupler OASIS3 that drives the system
- Developed by Transregional Collaborative Research Center 32
- <u>http://www.terrsysmp.org</u>

- Testcase executed on JUQUEEN (512 MPI ranks)
- Copy terrsysmp_mpmd_juqueen.cubex to your livedvd

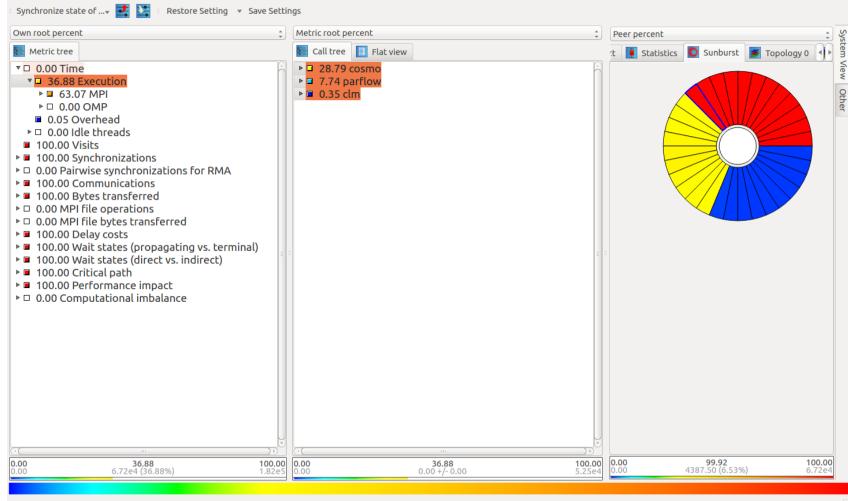
https://fz-juelich.sciebo.de/s/0qIEZzXpjS6uyte

Examine measurement with CUBE

livedvd> cube terrsysmp_mpmd_juqueen.cubex



TerrSysMP: execution breakdown

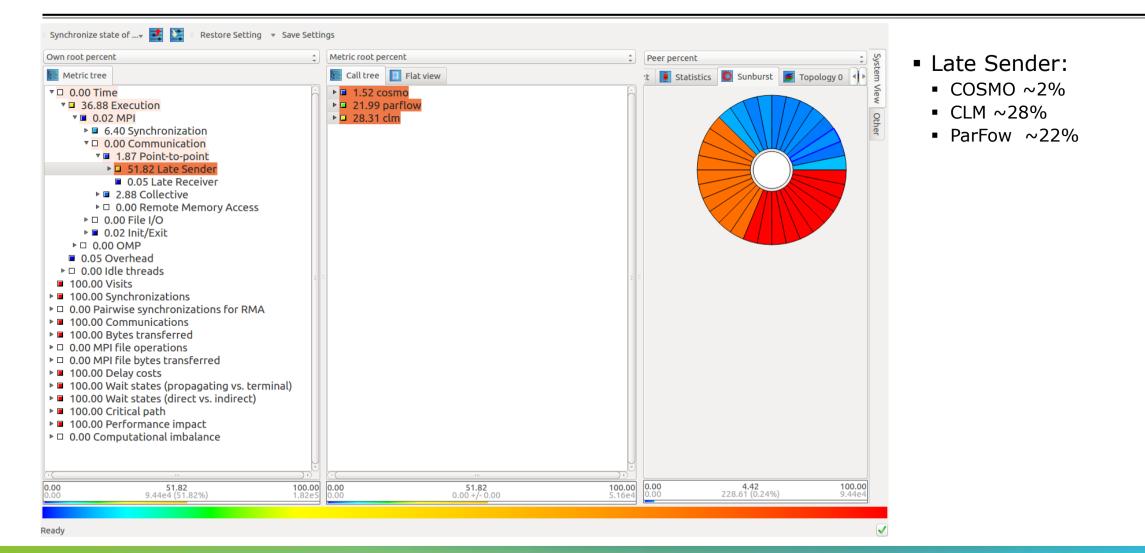


- COSMO 12 nodes (red)
- CLM 10 nodes (blue)
- ParFlow 10 nodes (yellow)
- ~37% in computation and ~63% in MPI

Selected "99.92 node R63-M0-N0a-J00 <0,1,0,1,1>

VICTOR CONFUSION CONTRACTOR CONFUSION CONTRACTOR CONFUSION CONTRACTOR CONFUSION CONFUSICO CONFUSICO

TerrSysMP: communication breakdown

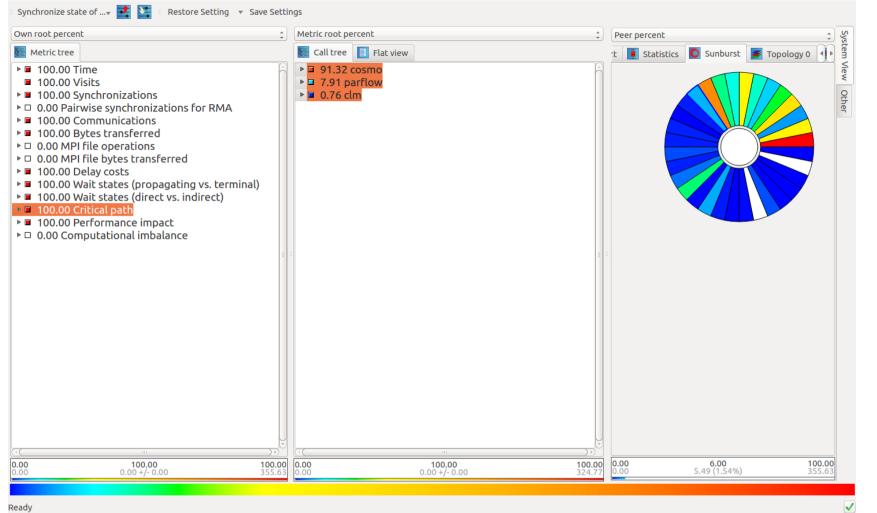


TerrSysMP: communication breakdown (cont)

Own root percent	Metric root percent	Peer percent
Metric tree	E Call tree E Flat view	🚺 System tree 🔢 Jenga Fett 🚺 Statistics 🜗
0.00 Time	↓ ■ 1.52 cosmo	A □ 0.00 node R63-M0-N0a-J09 < 0.0,1,0
36.88 Execution	0.00 parflow (0.00%)	▶ □ 0.00 node R63-M0-N0a-J0e <0,0,1,0
0.02 MPI	■ 0.00 Solve (0.00%)	▶ □ 0.00 node R63-M0-N0a-J05 <0,0,1,1
6.40 Synchronization	0.00 SolverRichards (0.00%)	▶ □ 0.00 node R63-M0-N0a-J02 <0,0,1,1
0.00 Communication	0.06 SetupRichards	▶ □ 0.00 node R63-M0-N0a-J0b <0,1,0,0
1.87 Point-to-point	0.00 AdvanceRichards (0.00%)	▶ □ 0.00 node R63-M0-N0a-J0c <0,1,0,0,
51.82 Late Sender	0.00 MPI_Wait	▶ □ 0.00 node R63-M0-N0a-J07 < 0,1,0,1
0.05 Late Receiver	0.04 MPI_Recv	▶ □ 0.00 node R63-M0-N0a-J00 <0,1,0,1
2.88 Collective	21.67 MPI_Waitall	76.74 node R63-M0-N0a-J0a <0,1,1,1
I 0.00 Remote Memory Access	0.04 MPI_Waitany	76.74 node R63-M0-N0a-J0d <0,1,1,
▶ □ 0.00 File I/O	0.02 FinalizeVectorUpdate	▶ ■ 76.75 node R63-M0-N0a-J06 <0,1,1,
Init/Exit	0.16 KinsolNonlinSolver	▶ ■ 76.74 node R63-M0-N0a-J01 <0,1,1,
▶ □ 0.00 OMP	▼ ■ 0.00 clm (100.00%)	▶ ■ 76.74 node R63-M0-N0a-J15 <1,0,0,
0.05 Overhead	0.01atmdrvmod_NMOD_atmdrv_init	▶ ■ 76.74 node R63-M0-N0a-J12 <1,0,0,
▷ □ 0.00 Idle threads	0.00atmdrvmod_NMOD_atmdrv	▶ ■ 76.74 node R63-M0-N0a-J19 <1,0,0,
100.00 Visits	0.00 receive_fld_2cos (0.00%)	▶ ■ 76.74 node R63-M0-N0a-J1e <1,0,0,
100.00 Synchronizations	0.00 oas_clm_rcv (0.00%)	▶ ■ 76.74 node R63-M0-N0a-J14 <1.0.1.
0.00 Pairwise synchronizations for RMA	28.23 MPI_Waitall	▶ ■ 76.74 node R63-M0-N0a-J13 <1,0,1,
100.00 Communications	0.01 send_fld_2cos	▶ ■ 99.98 node R63-M0-N0a-J18 <1.0.1.
100.00 Bytes transferred	0.00 send_fld_2pfl	▶ ■ 99.99 node R63-M0-N0a-J1f <1,0,1,1
0.00 MPI file operations	0.07perf_mod_NMOD_t_prf	▶ ■ 99.99 node R63-M0-N0a-J16 <1,1,0,
0.00 MPI file bytes transferred		▶ ■ 99.99 node R63-M0-N0a-J11 <1,1,0,
100.00 Delay costs		▶ ■ 99.99 node R63-M0-N0a-J1a <1,1,0,
100.00 Wait states (propagating vs. terminal)		▶ ■ 99.99 node R63-M0-N0a-J1d <1,1,0,
100.00 Wait states (direct vs. indirect)		▶ ■ 99.99 node R63-M0-N0a-J17 <1,1,1
100.00 Critical path		▶ ■ 100.00 node R63-M0-N0a-J10 <1,1,1
100.00 Performance impact		▶ ■ 99.99 node R63-M0-N0a-J1b <1,1,1
0.00 Computational imbalance		▶ ■ 99.99 node R63-M0-N0a-J1c <1,1,1,
	J	All (512 elements)
00 51.82 100.0 00 9.44e4 (51.82%) 1.82e	0 0.00 49.90	100.00 0.00 76.74 100.00 5.16e4 0.00 3947.96 (4.34%) 9.09e4

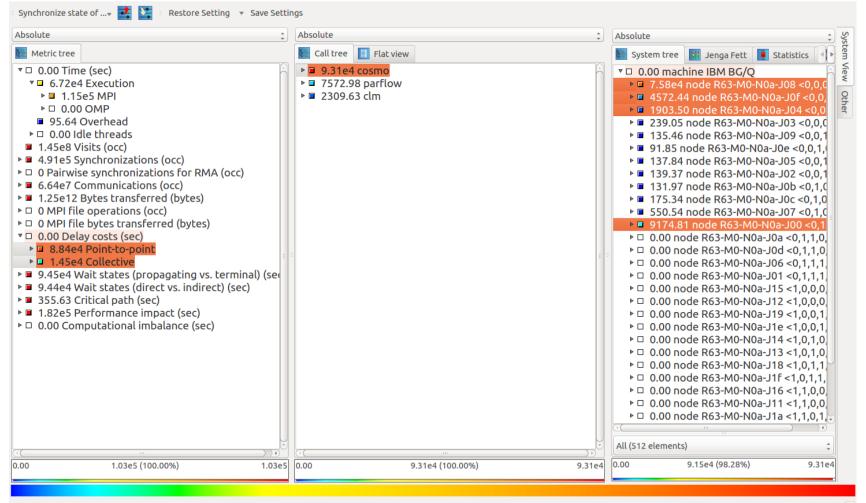
- Late Sender:
 - ~50% in MPI_Waitall

TerrSysMP: critical path



 The lion's share of the critical path is in COSMO V VIRTUAL INSTITUTE - HIGH PRODUCTIVITY SUPERCOMPUTING

TerrSysMP: delay analysis



 Most of the delays are caused by 4 nodes in COSMO

Selected "node R63-M0-N0a-J00 <0,1,0,1,1>"

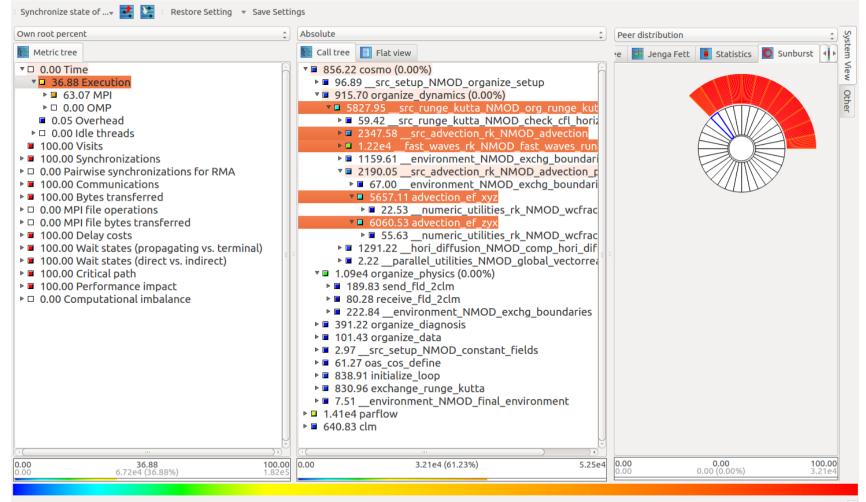
TerrSysMP: delay analysis (cont)

bsolute	Absolute \$	Absolute
Metric tree	Call tree 📋 Flat view	📕 System tree 🔢 Jenga Fett 🚺 Statistics 🖣
7 🗆 0.00 Time (sec)	■ 1345.33 cosmo	▼ □ 0.00 machine IBM BG/Q
6.72e4 Execution	232.07src_setup_NMOD_organize_setup	5.83e4 node R63-M0-N0a-J08 <0,0,0
1.15e5 MPI	1624.26 organize_dynamics	2998.31 node R63-M0-N0a-J0f <0,0,
▶ □ 0.00 OMP	0.61 input_dynctl	1424.04 node R63-M0-N0a-J04 <0,0
95.64 Overhead	8971.21src_runge_kutta_NMOD_org_runge_kut	98.10 node R63-M0-N0a-J03 <0,0,0,
▷ □ 0.00 Idle threads	0.59src_runge_kutta_NMOD_calc_small_time	73.79 node R63-M0-N0a-J09 <0,0,1,
1.45e8 Visits (occ)	I01.99src_runge_kutta_NMOD_check_cfl_hor	▶ ■ 44.76 node R63-M0-N0a-J0e <0,0,1,
4.91e5 Synchronizations (occ)	3153.46src_advection_rk_NMOD_advection	▶ ■ 79.39 node R63-M0-N0a-J05 <0,0,1,
• O Pairwise synchronizations for RMA (occ)	776.77environment_NMOD_exchg_bounda	▶ ■ 85.81 node R63-M0-N0a-J02 <0,0,1,
6.64e7 Communications (occ)	I.75e4fast_waves_rk_NMOD_fast_waves_run	75.01 node R63-M0-N0a-J0b <0,1,0,
1.25e12 Bytes transferred (bytes)	2149.39environment_NMOD_exchg_bound	73.90 node R63-M0-N0a-J0c <0,1,0,0
• O MPI file operations (occ)	2593.06environment_NMOD_exchg_boundari	270.53 node R63-M0-N0a-J07 <0,1,0
O MPI file bytes transferred (bytes)	3284.67src_advection_rk_NMOD_advection_p	6898.62 node R63-M0-N0a-J00 <0,1
0.00 Delay costs (sec)	▶ ■ 172.45environment_NMOD_exchg_bounda	0.00 node R63-M0-N0a-J0a <0,1,1,0
8.84e4 Point-to-point	8182.95 advection_ef_xyz	▶ □ 0.00 node R63-M0-N0a-J0d <0,1,1,0
1.45e4 Collective	▶ ■ 41.56numeric_utilities_rk_NMOD_wcfrac	▶ □ 0.00 node R63-M0-N0a-J06 <0,1,1,1
9.45e4 Wait states (propagating vs. terminal) (see	8987.77 advection_ef_zyx	▷ □ 0.00 node R63-M0-N0a-J01 < 0,1,1,1
9.44e4 Wait states (direct vs. indirect) (sec)	76.18numeric_utilities_rk_NMOD_wcfrac	▶ □ 0.00 node R63-M0-N0a-J15 <1,0,0,0
355.63 Critical path (sec)	2547.28hori_diffusion_NMOD_comp_hori_dif	▶ □ 0.00 node R63-M0-N0a-J12 <1,0,0,0
1.82e5 Performance impact (sec)	31.71parallel_utilities_NMOD_global_vectorre	▷ □ 0.00 node R63-M0-N0a-J19 <1,0,0,1
0.00 Computational imbalance (sec)	2.36e4 organize_physics	▶ □ 0.00 node R63-M0-N0a-J1e <1,0,0,1
	0.41 input_phyctl	▷ □ 0.00 node R63-M0-N0a-J14 <1,0,1,0
	▶ ■ 377.66 send_fld_2clm	▶ □ 0.00 node R63-M0-N0a-J13 <1,0,1,0
	29.97 receive_fld_2clm	▶ □ 0.00 node R63-M0-N0a-J18 <1,0,1,1
	91.16 oas_cos_rcv	▷ □ 0.00 node R63-M0-N0a-J1f <1,0,1,1,
	20.00 MPI_Barrier	▶ □ 0.00 node R63-M0-N0a-J16 <1,1,0,0
	537.27environment_NMOD_exchg_boundaries	▷ □ 0.00 node R63-M0-N0a-J11 <1,1,0,0
	1486.81 organize_diagnosis	▶ □ 0.00 node R63-M0-N0a-J1a <1,1,0,1
	▶ ■ 826.67 organize_data	
	Image: bold the second seco	All (512 elements)
00 1.03e5 (100.00%) 1.03e	5 0.00 7.04e4 (75.62%) 9.31e4	0.00 6.96e4 (98.86%) 7.04

 Most of the delays are caused by 6 routines

Selected "advection_ef_zyx"

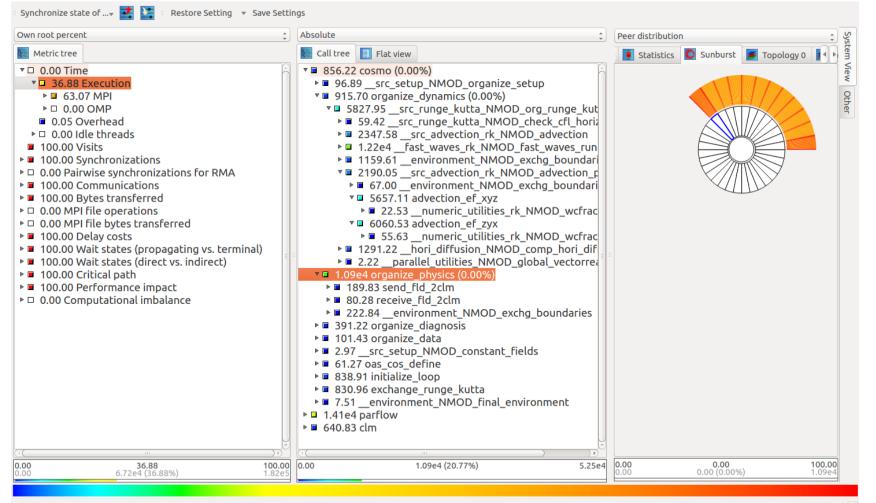
COSMO: computational load balance



 The first and the last MPI ranks in COSMO partition have less work in selected routines

Deselected "organize_physics"

COSMO: computational load balance



- The first and the last MPI ranks and nodes in COSMO partition have more work in selected routine
- Better load balance of COSMO component can reduce waiting time and improve overall performance

Selected "organize_physics"