www.bsc.es

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Towards the Optimization of the NMMB Model

George Markomanolis

Barcelona, 18 December 2013

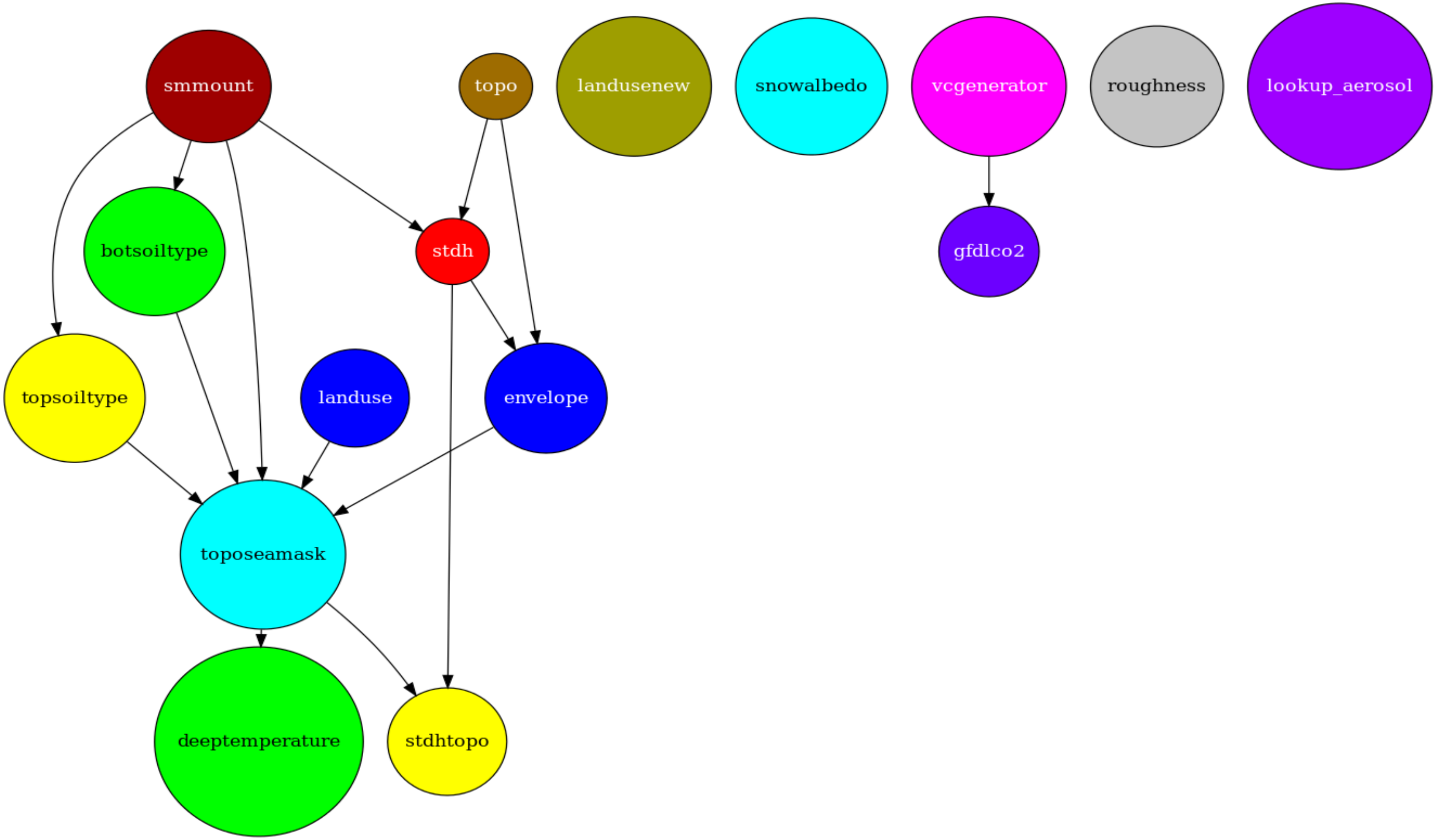
NMMB

- Nonhydrostatic Multiscale Meteorological Model on the B grid
  - Studies the air quality
  - We focus on dust module: On-line coupling of the dust model into the meteorological driver
  - There are other modules about other aerosols and chemical gas-phase
  - Pre-process
    - Fixed procedure is not executed often, it prepares the topology data
    - Variable procedure is executed quite often as its input is daily meteorological, surface sea temperatures etc.
  - Main application: Decomposition of the input
  - Post-process: Handling and visualizing the data

Why Performance Analysis?

- Is it needed?
  - Understand the behavior of an application
    - Can we execute an application faster?
    - Is a simulation finished on time?
  - Optimize an application
  - Do we use the resources optimally?
  - Could we decrease the electricity bills of a supercomputer?
  - Predict physical catastrophes on time?
  - Is needed for the next-generation of supercomputers
  - How do you know that your application can not be improved?

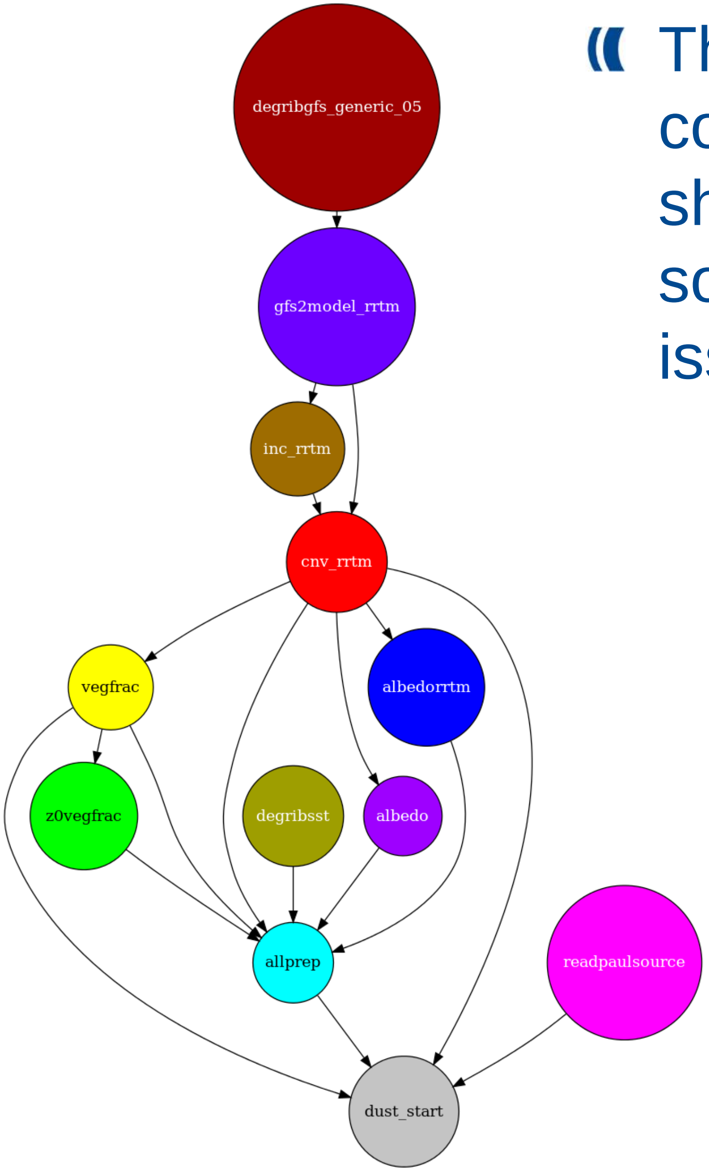
Fixed – COMPSs (Kim Serradell, Enric Tejedor)



Fixed

- We wanted to integrate the method in our bash script
- Adding Java code would improve the complexity of our approach
- Simple solution: Fortran/MPI application to execute Fixed method
- We used weights (execution time) per task in order to decide how many cores to use
  - Finally we used 5 cores and achieved a speedup of 2.7!
  - Not perfect, maybe we could optimize more some serial parts.

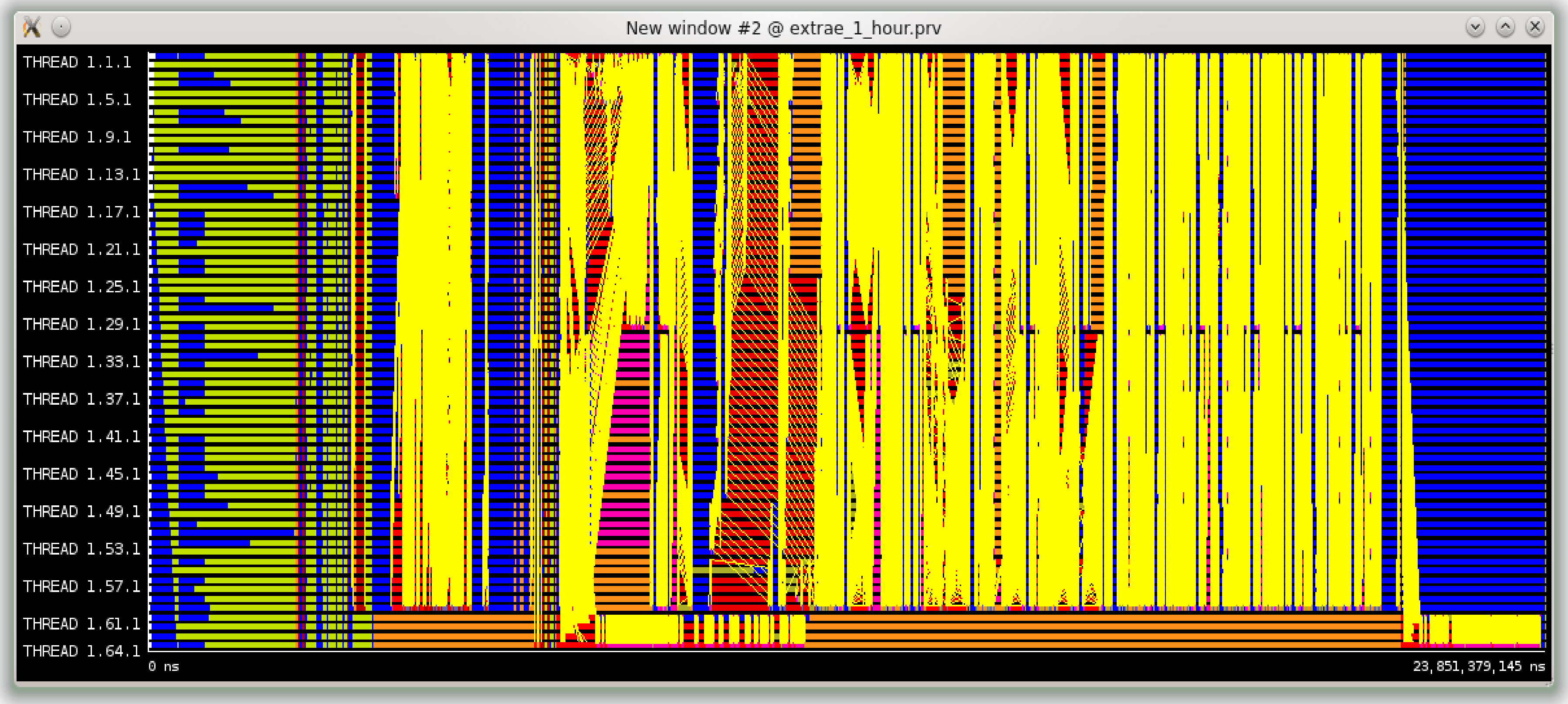
Variable – COMPSs (Kim Serradell, Enric Tejedor)



- The serial part *allprep* consumes a lot of time, we should investigate a hybrid solution because of memory issues

Paraver

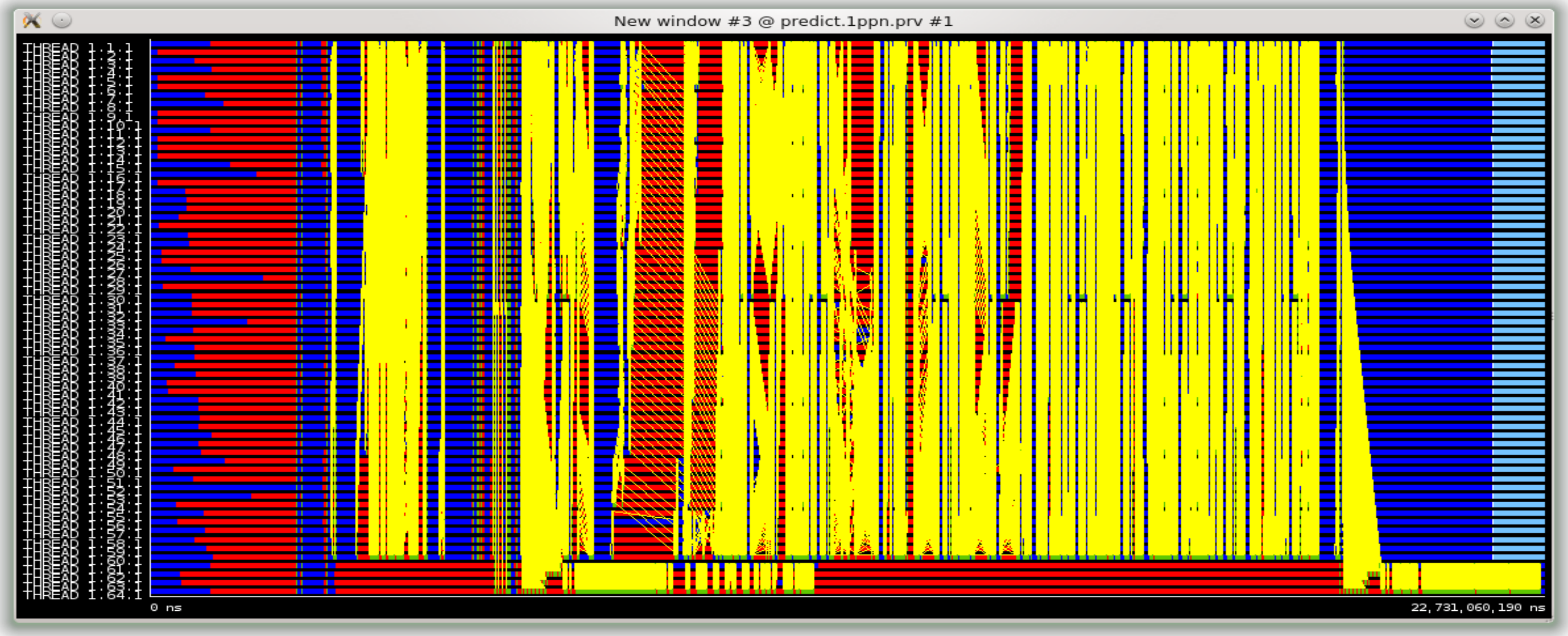
- One hour simulation of NMMB



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

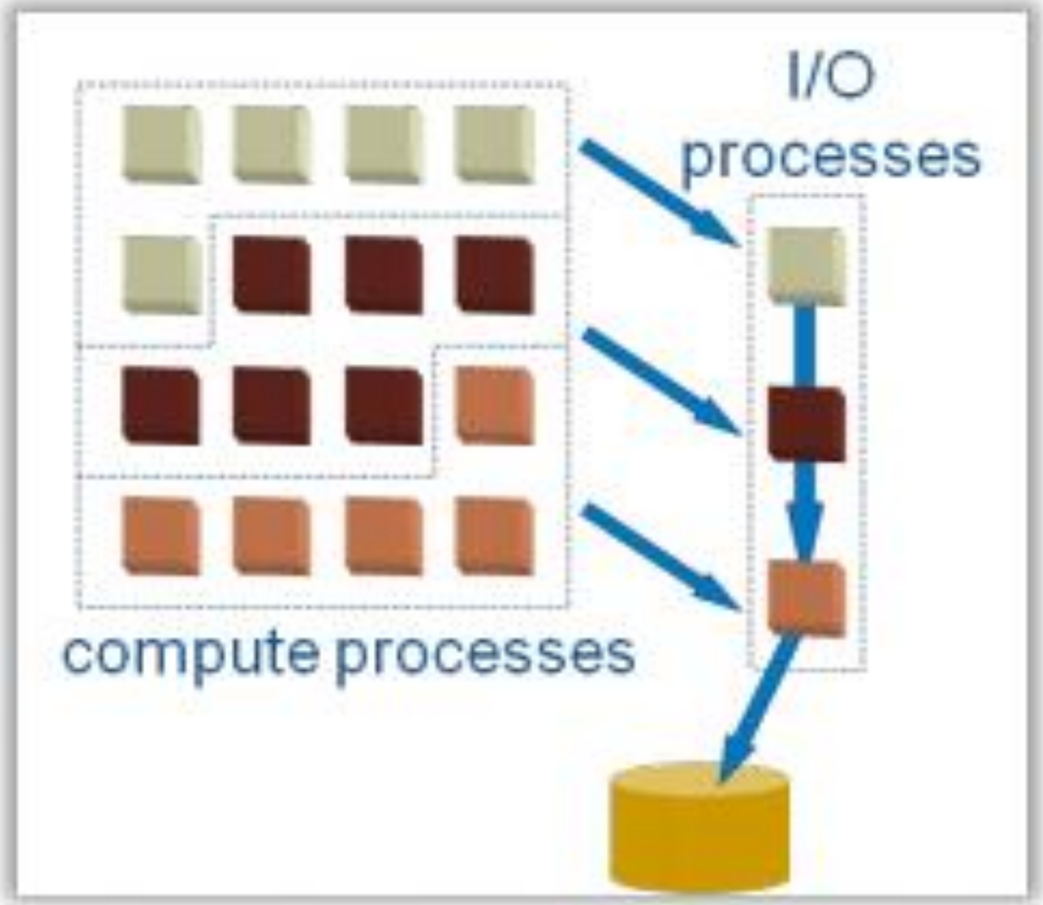
Paraver - Dimemas

- It seems that previously there was noise during the execution



Issue with I/O

- There is no parallel I/O implemented!



Issue with I/O

- Last binary is written with delay.
- Example regional 11km resolution

```
4778176548 Dec 15 09:25 nmmb_hst_01_bin_0000h_00m_00.00s
4778176548 Dec 15 09:28 nmmb_hst_01_bin_0001h_00m_00.00s
4778176548 Dec 15 09:31 nmmb_hst_01_bin_0002h_00m_00.00s
4778176548 Dec 15 09:34 nmmb_hst_01_bin_0003h_00m_00.00s
4778176548 Dec 15 09:38 nmmb_hst_01_bin_0004h_00m_00.00s
4778176548 Dec 15 09:41 nmmb_hst_01_bin_0005h_00m_00.00s
4778176548 Dec 15 10:42 nmmb_hst_01_bin_0006h_00m_00.00s
```

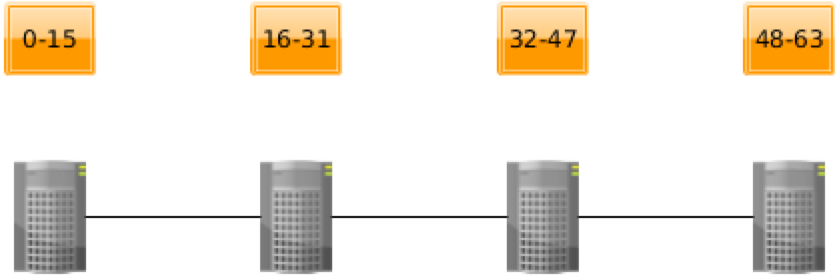
Issue with I/O – Instrumented execution

- The instrumented execution has no issue…

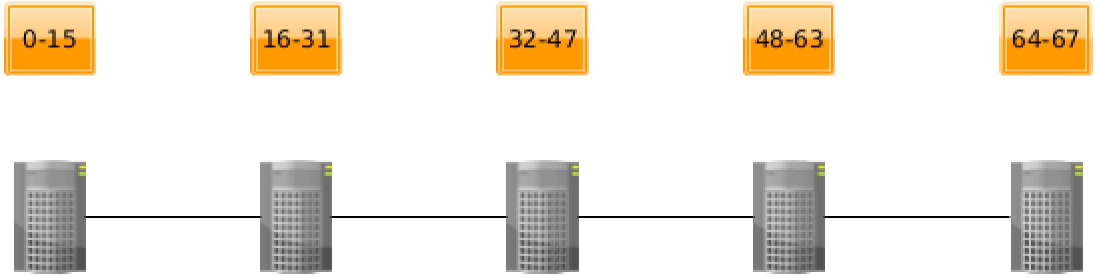
```
4778176548 Dec 15 11:14 nmmb_hst_01_bin_0000h_00m_00.00s
4778176548 Dec 15 11:17 nmmb_hst_01_bin_0001h_00m_00.00s
4778176548 Dec 15 11:21 nmmb_hst_01_bin_0002h_00m_00.00s
4778176548 Dec 15 11:24 nmmb_hst_01_bin_0003h_00m_00.00s
4778176548 Dec 15 11:27 nmmb_hst_01_bin_0004h_00m_00.00s
4778176548 Dec 15 11:30 nmmb_hst_01_bin_0005h_00m_00.00s
4715192924 Dec 15 11:33 nmmb_hst_01_bin_0006h_00m_00.00s
```

Issue with I/O – Mapping

- Initial mapping for an experiment with 64 cores where the last 4 ranks are the write tasks

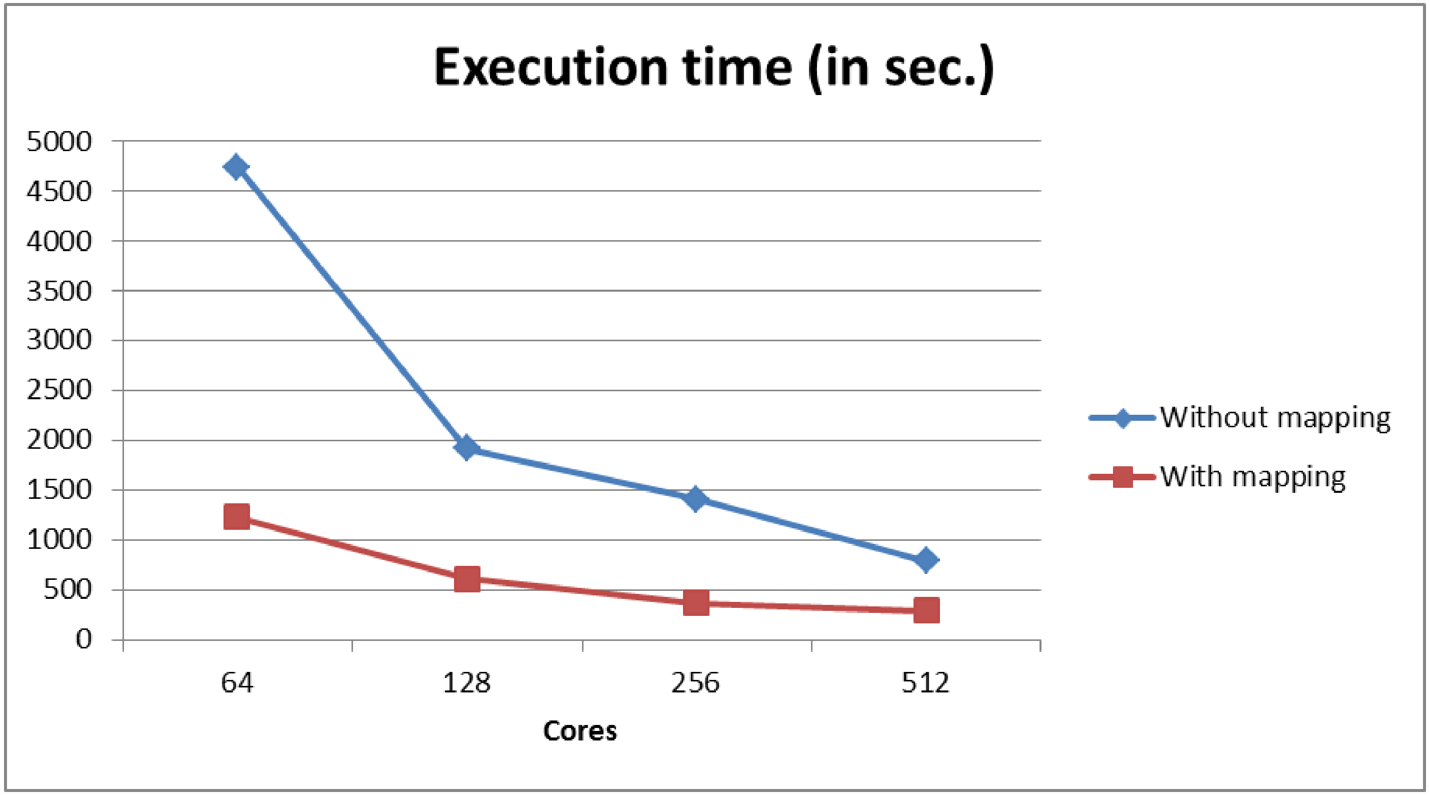


- Final mapping



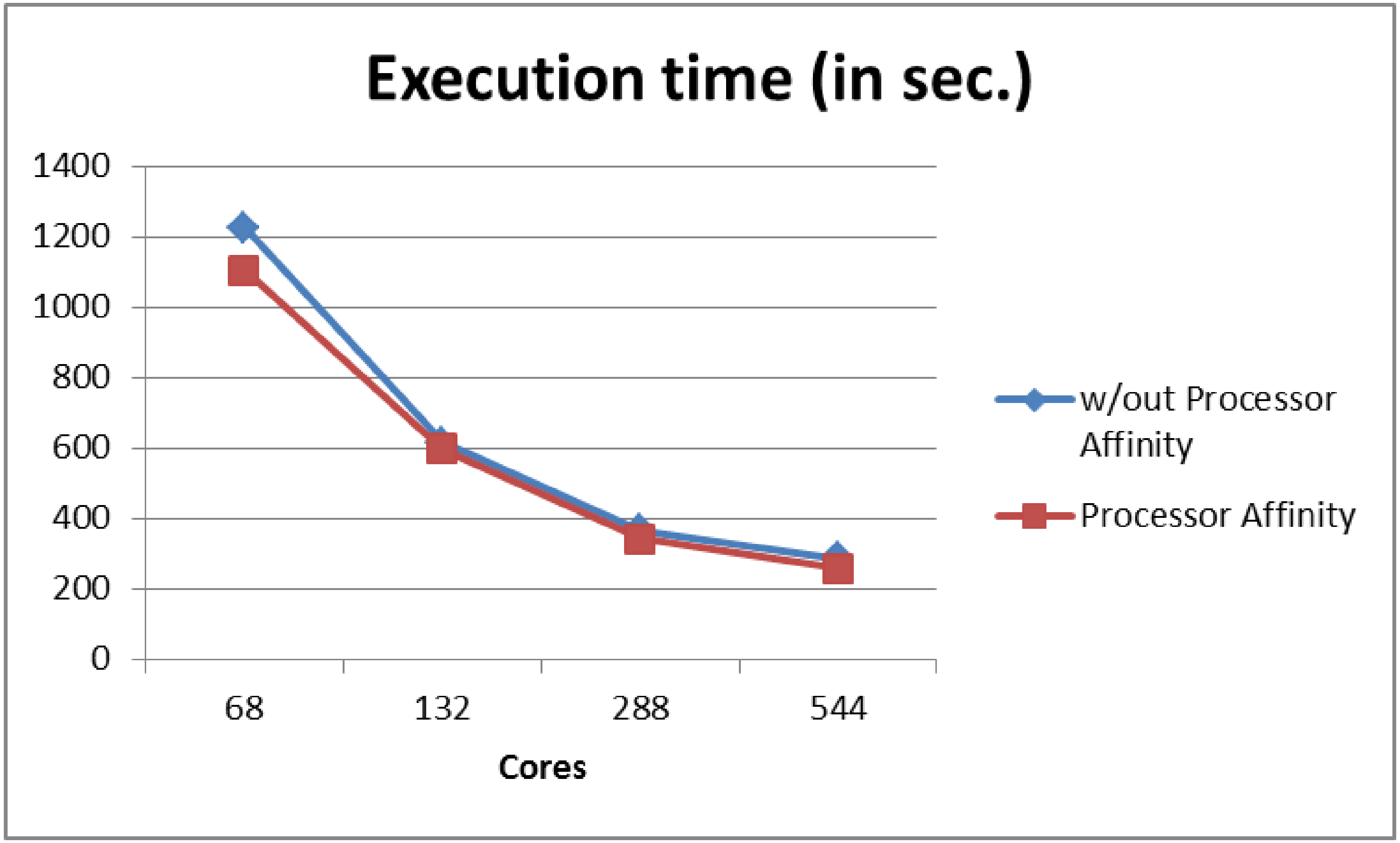
Performance of different mapping and more I/O servers

- The new mapping improved the execution time between 2.73 and 3.85 times



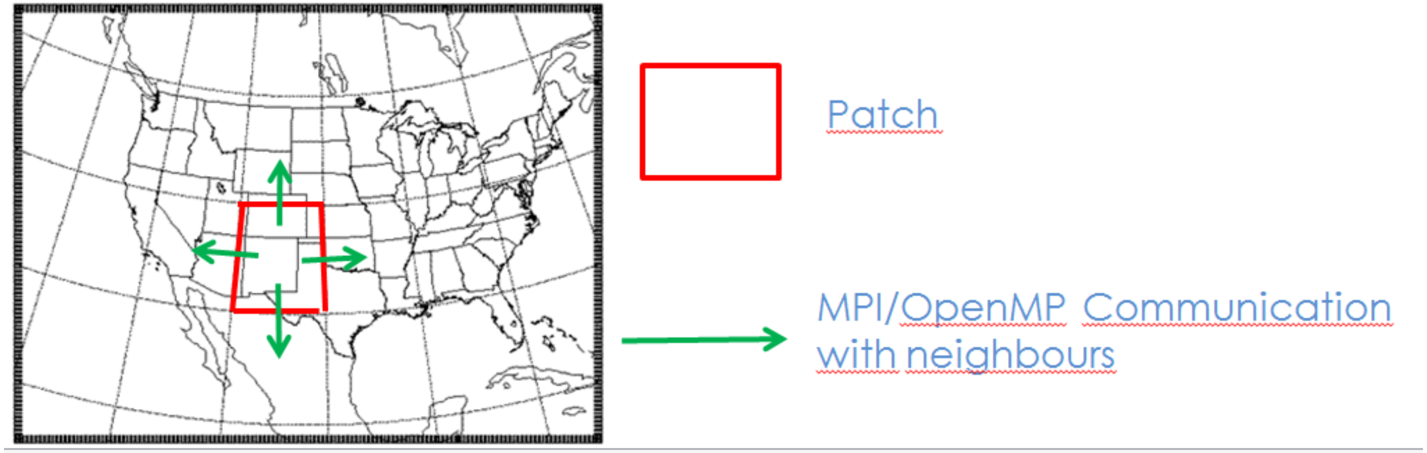
Processor Affinity

- Processor affinity improved the execution time between 2.8% and 10% (some colleagues reported 20% improvement)



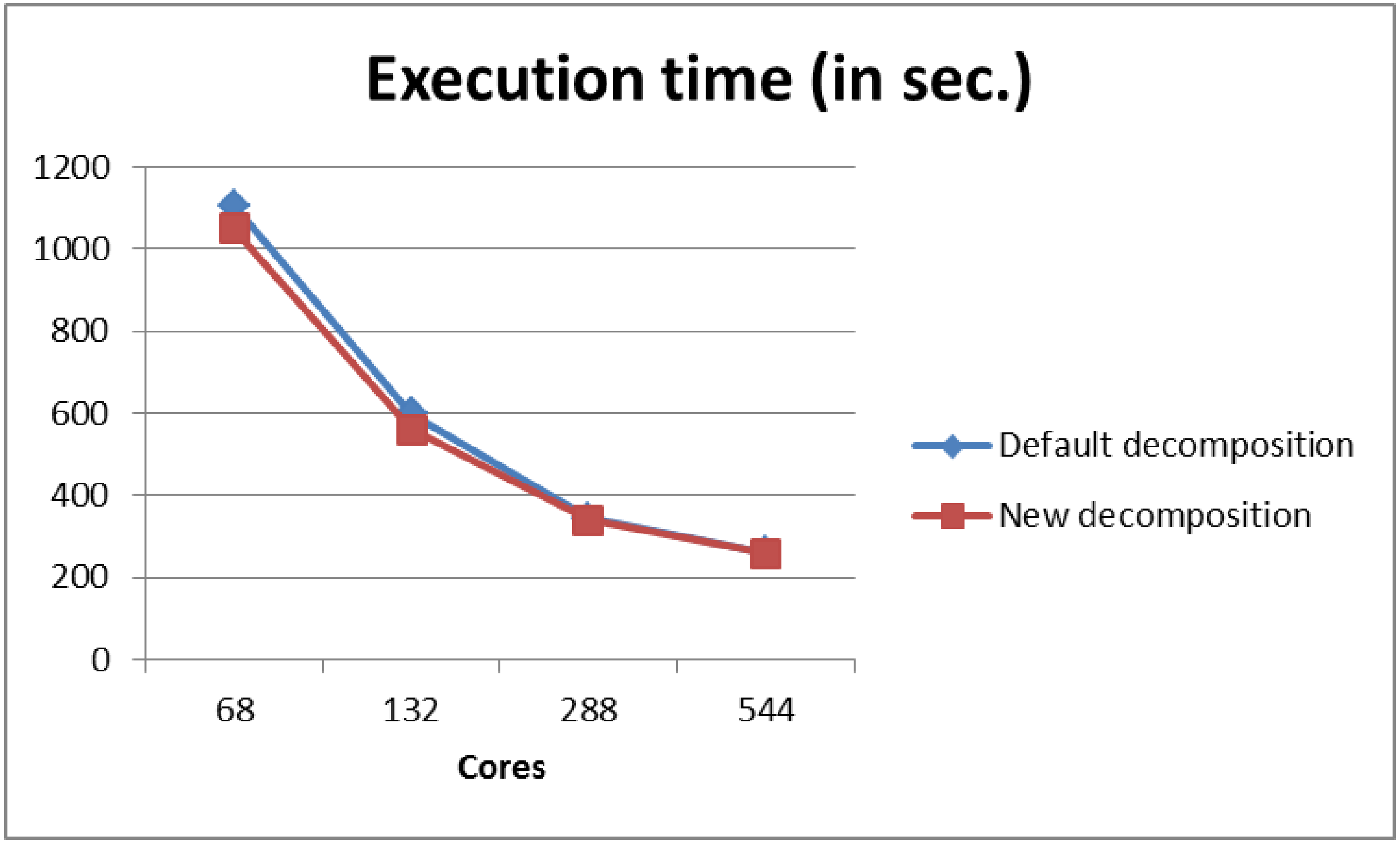
Decomposition (X,Y)

- Usually we use a square decomposition or something close to square.
- It is better to use values to a more rectangular decomposition (i.e. X<<Y). This leads to longer inner loops for better vector and register reuse, better cache blocking, and more efficient halo exchange communication pattern.



Decomposition

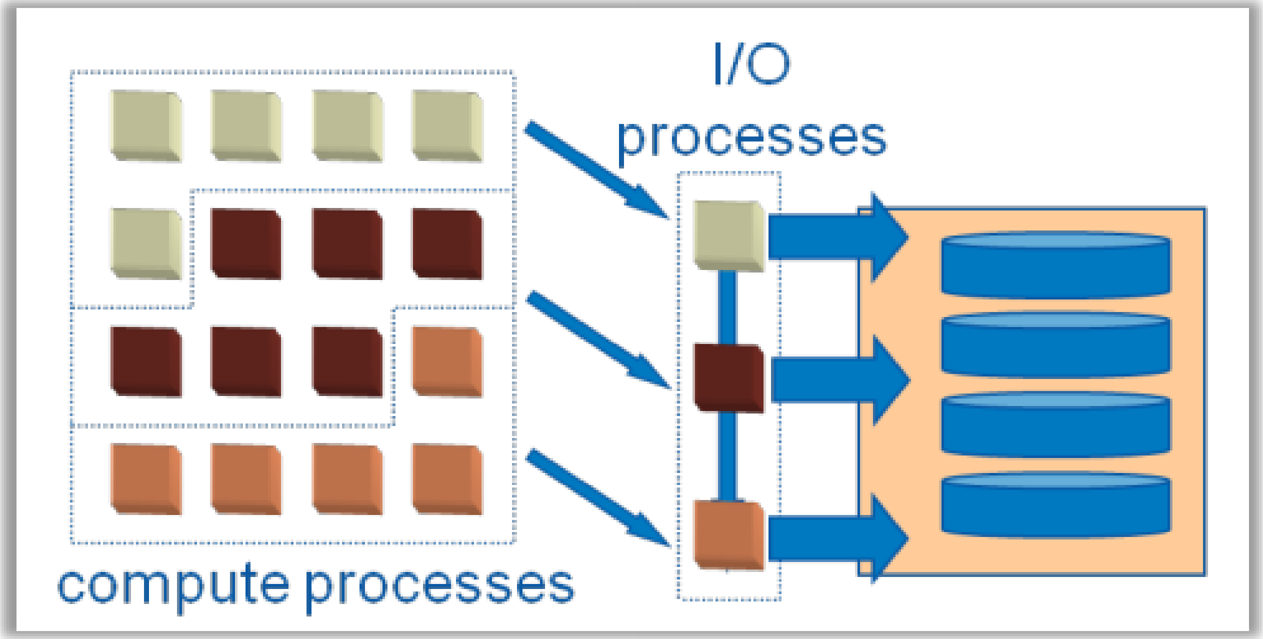
- New decomposition improved the execution time till 6.5%



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

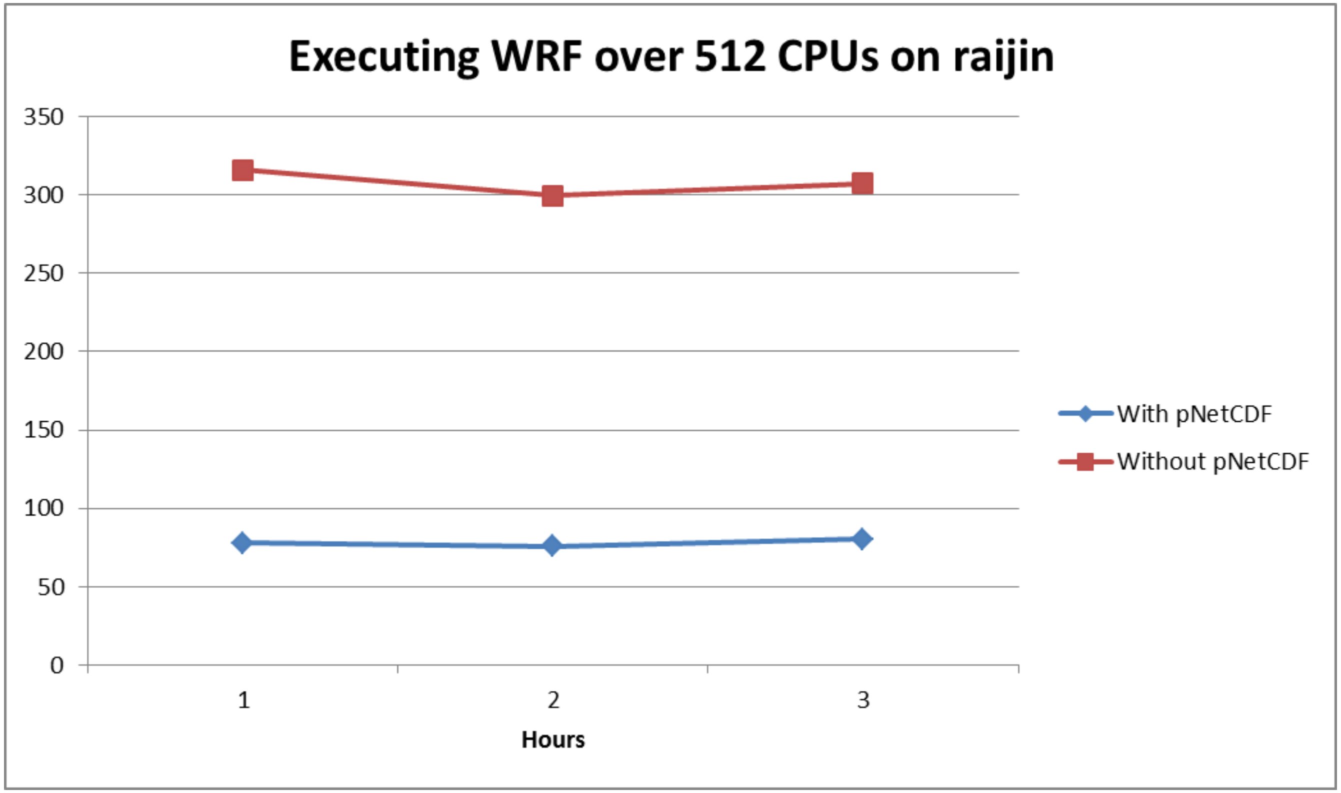
Improved I/O (future work)

- Parallel NetCDF written to single files by all MPI tasks.



Parallel NetCDF

- Results from ARC centre of excellence for Climate System Science.
- Parallel NetCDF for this case improves the performance by 3.8 to 4 times.



Roadmap to OmpSs

- NMMB is based on the Earth System Modelling Framework (ESMF)
- The current ESMF release (v3.1) is not supporting threads
- However, the development version of NMMB uses ESMF v6.3
- Post-process broke because of some other issues (is going to be fixed)
- The new version of NMMB with OmpSs support has been compiled by Julian Morillo (CS)
- Ready to apply and test OmpSs

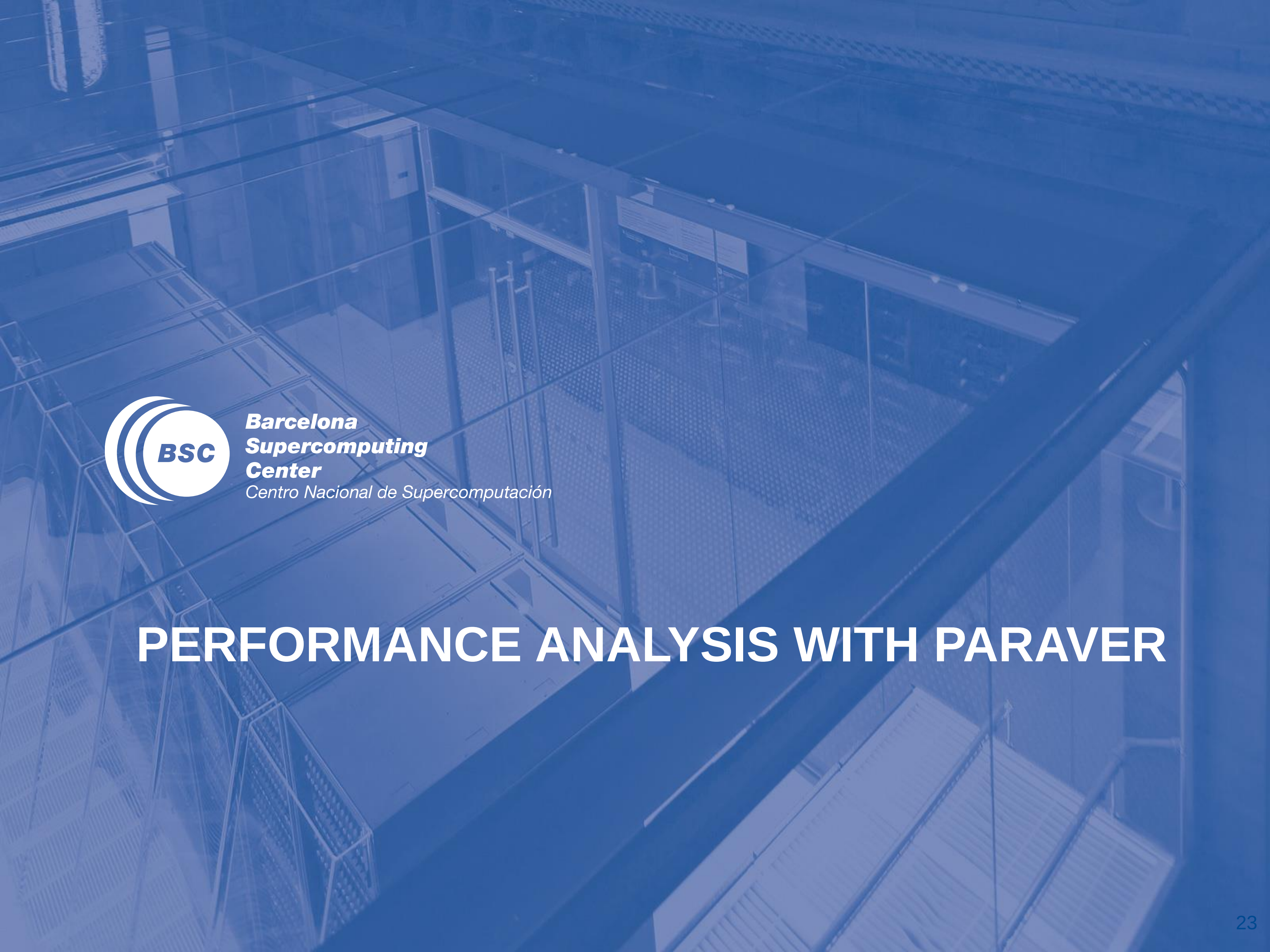
Future work

- Add parallel I/O (writing and reading)
- Use OmpSs programming model
  - Study GPU case
  - Explore Xeon Phi
- Improve performance and scale NMMB for thousands of cores
- Collaboration with the Computer Science department to prepare a submission to PRACE Scientific and Industrial Conference 2014.

Conclusions

- I/O can be a bottleneck
- Study your application and its configuration before start the operational execution
- Paraver can provide a lot of insight information about the behavior of an application
- Integrate new technologies

Questions?



Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

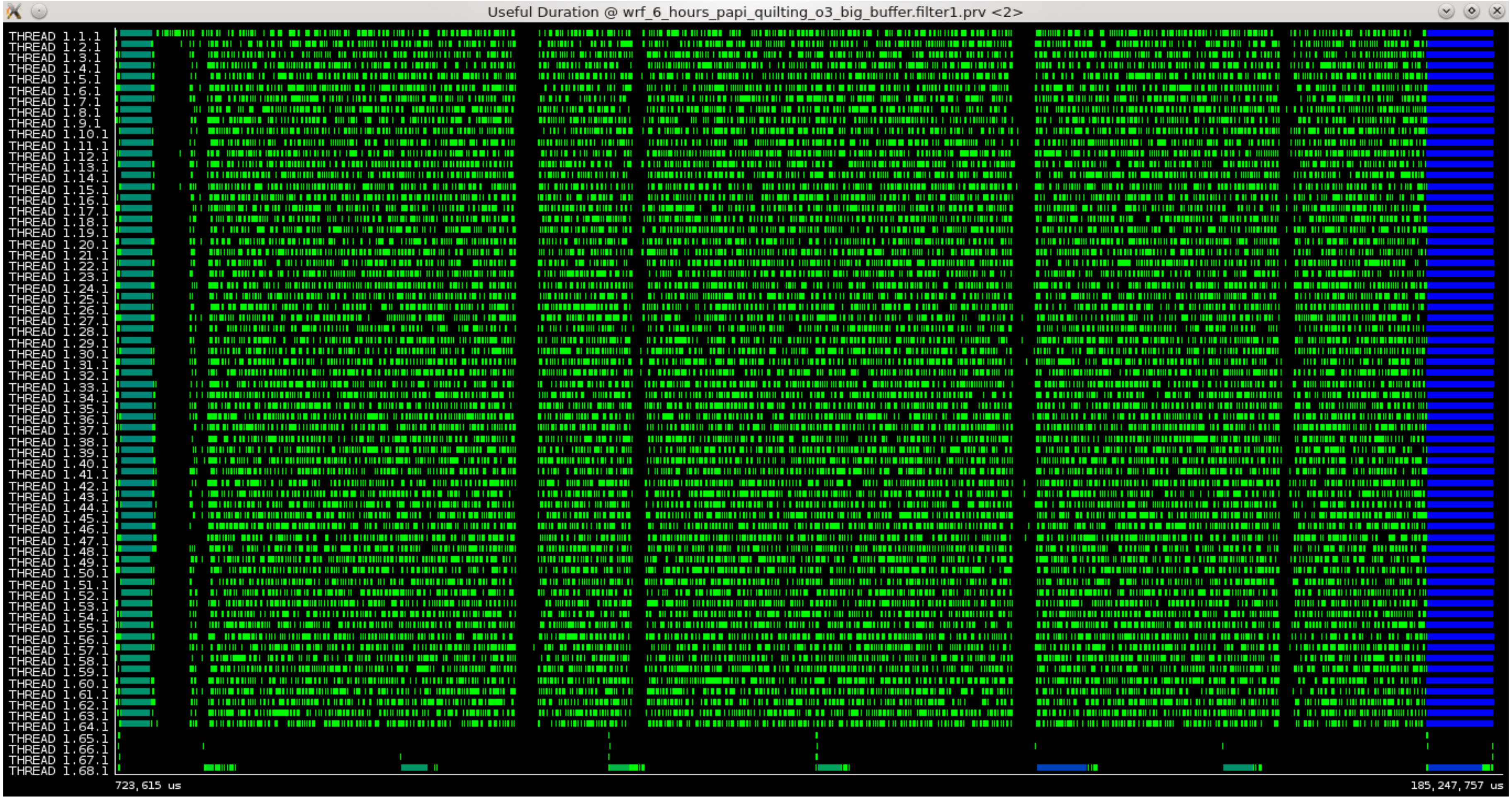
PERFORMANCE ANALYSIS WITH PARAVER

Paraver

- An application to analyze traces
- Discover bottlenecks
- Possible to do visual and statistical analysis of the various events
- Customizable semantics of the visualized information
- Provides views
- Information: http://www.bsc.es/paraver/

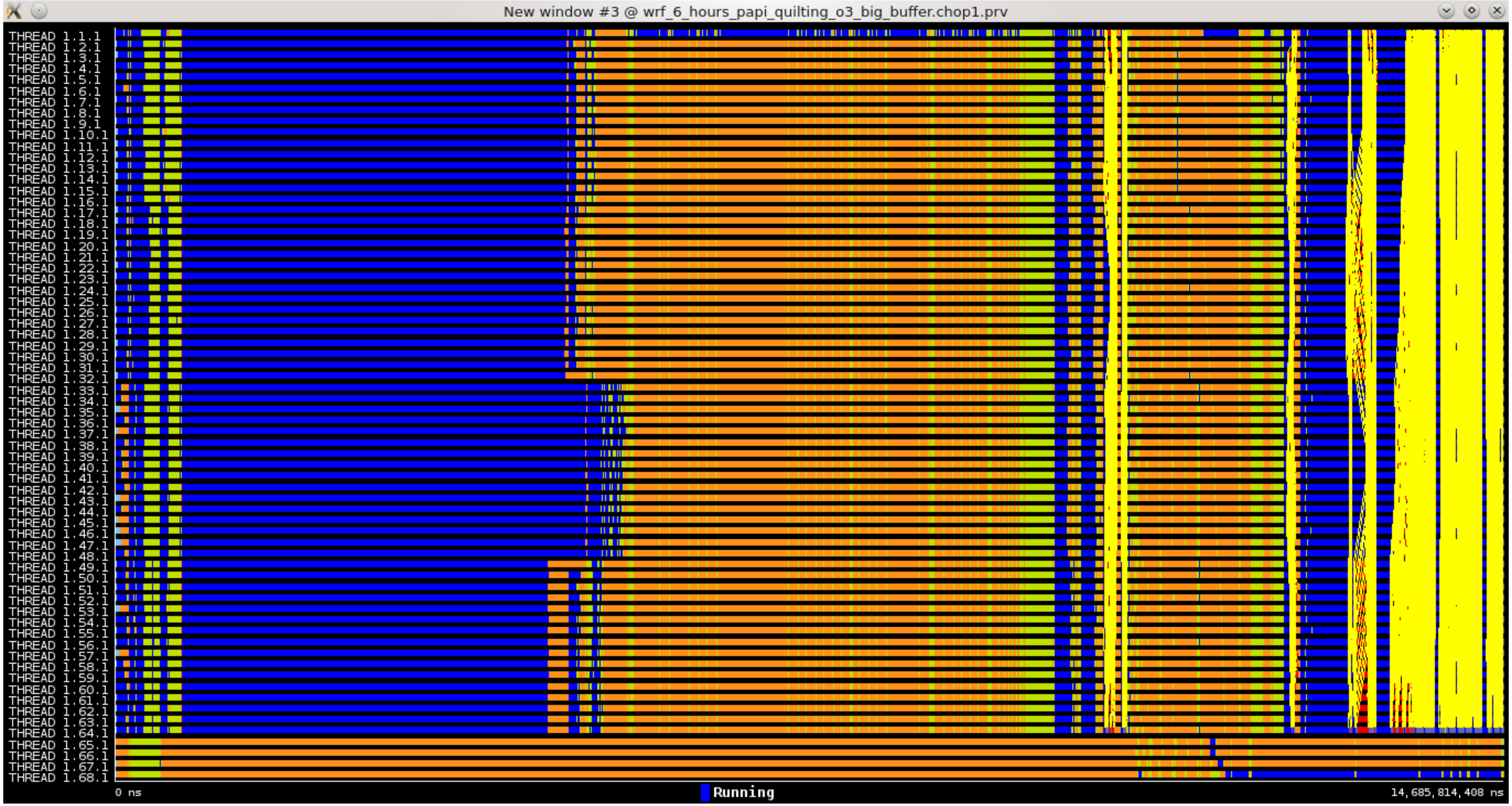
Visualizing the computation of a whole trace

- Visualizing computation duration of 6 hours simulation (5.2GB initial trace, 68 cores)



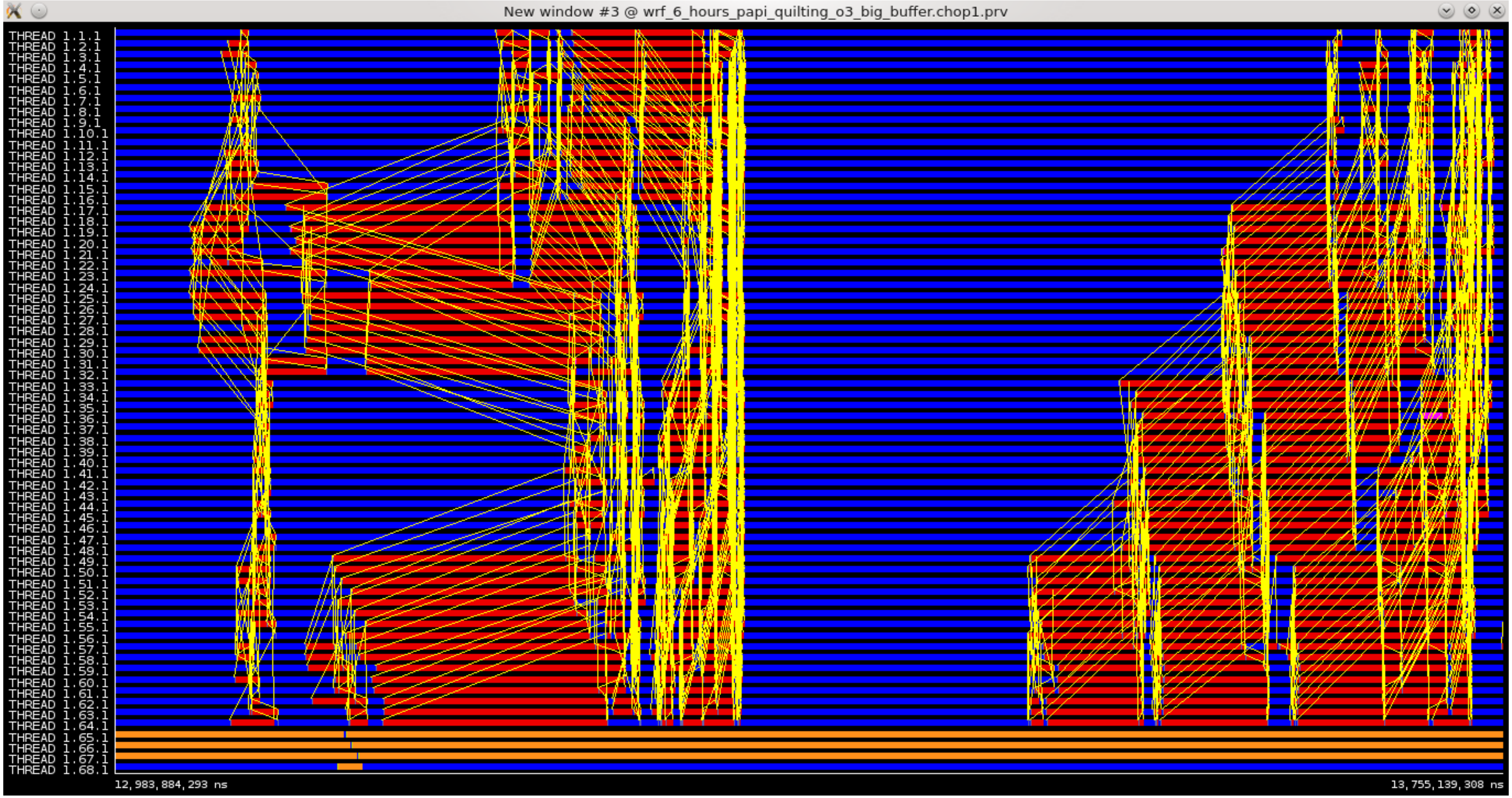
Trace Analysis – Beginning of the trace

- Beginning of the trace, 4 cores for I/O quilting
- Blue colour is running part, no communication, yellow colour is message transfer (send/recv etc.)



Trace Analysis – Beginning of the trace

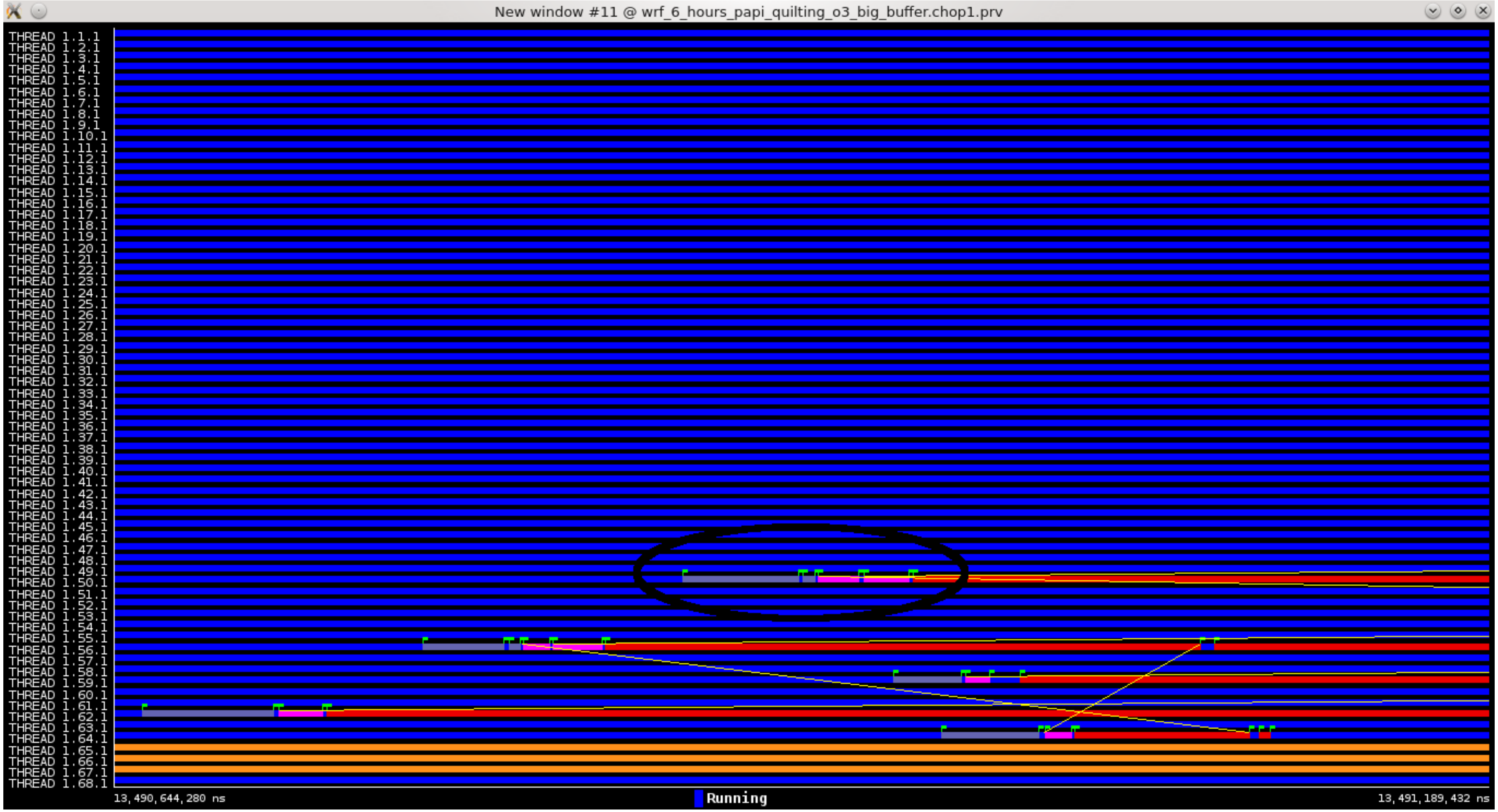
- Let's zoom a bit before the end of the previous visualization
- There are some long MPI_Wait calls (red colour)


New window #3 @ wrf_6_hours_papi_quilting_o3_big_buffer.chop1.prv
THREAD 1.1.1 … THREAD 1.68.1
12,983,884,293 ns
13,755,139,308 ns

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Trace Analysis – Beginning of the trace

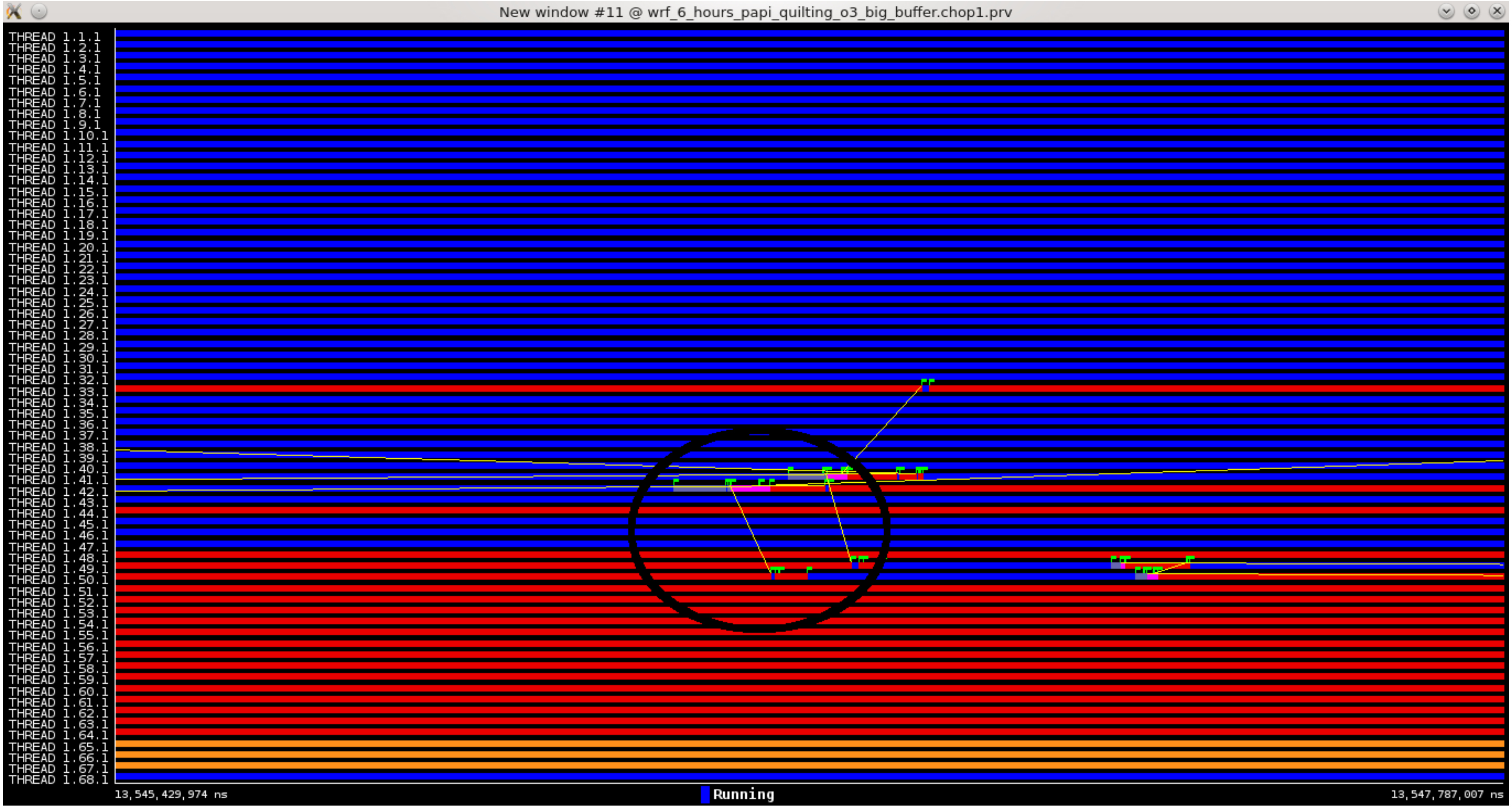
- We zoom at the beginning of the second half of the previous plot and we focus on rank 50
- There are two MPI_Irecv and MPI_Isend calls before the MPI_Wait call



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

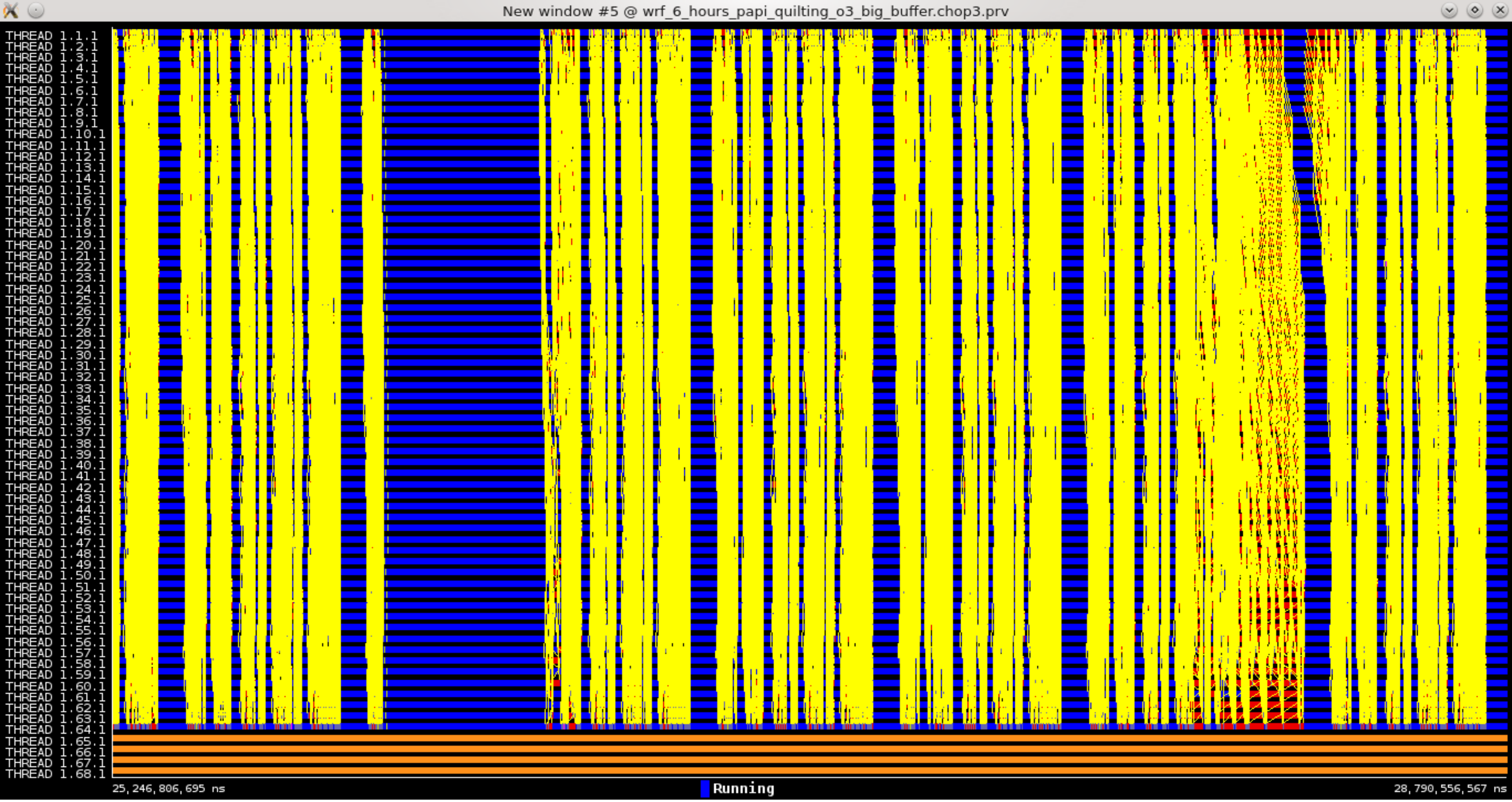
Trace Analysis – Beginning of the trace

- The corresponding MPI_Isend for the previous MPI_Irecv is called too late
- Possible solution move MPI_Wait of rank 50 after some computation phases



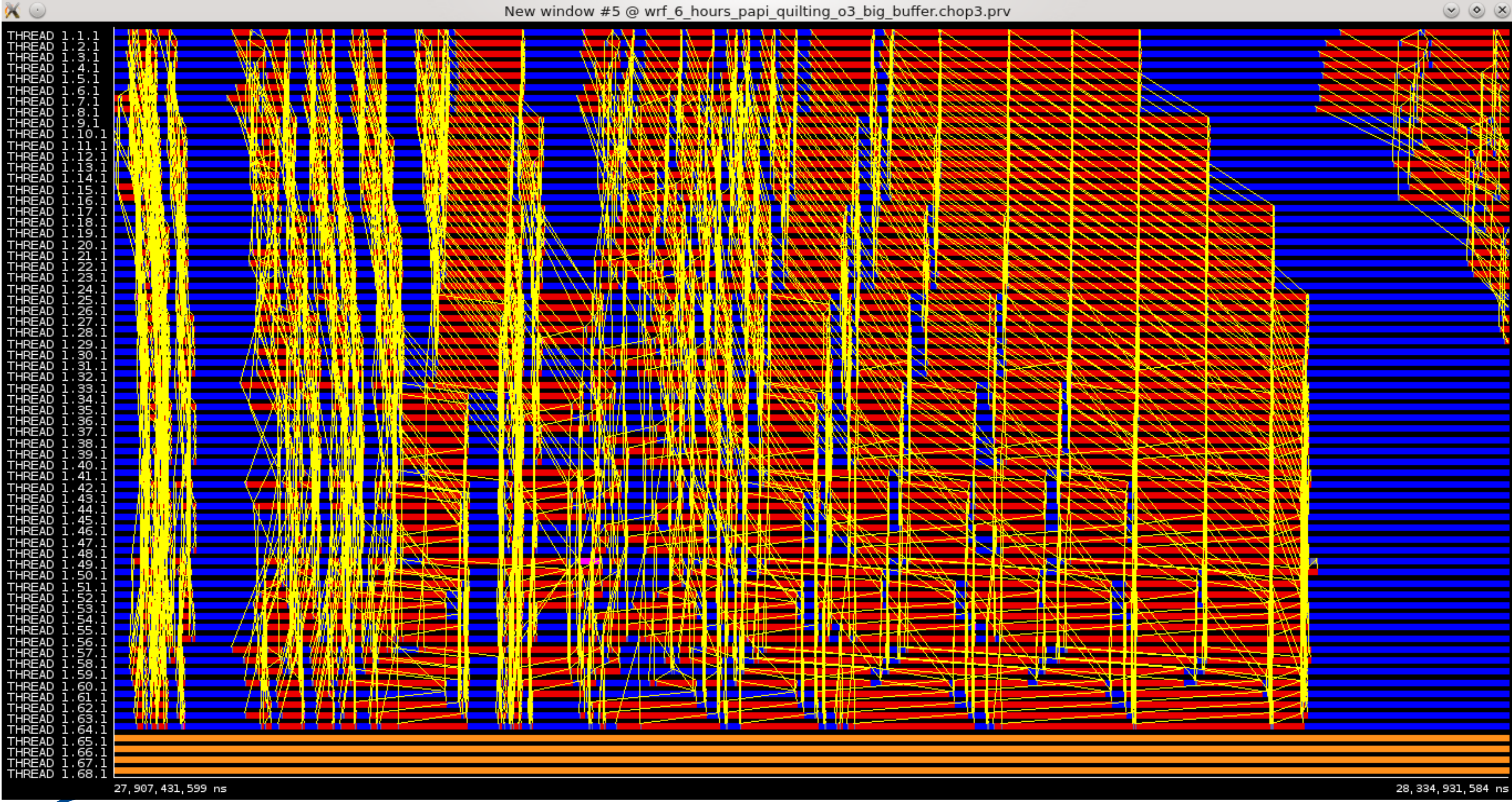
Trace Analysis

- We can observe some communications at the right that behave different than the rest ones



Trace Analysis

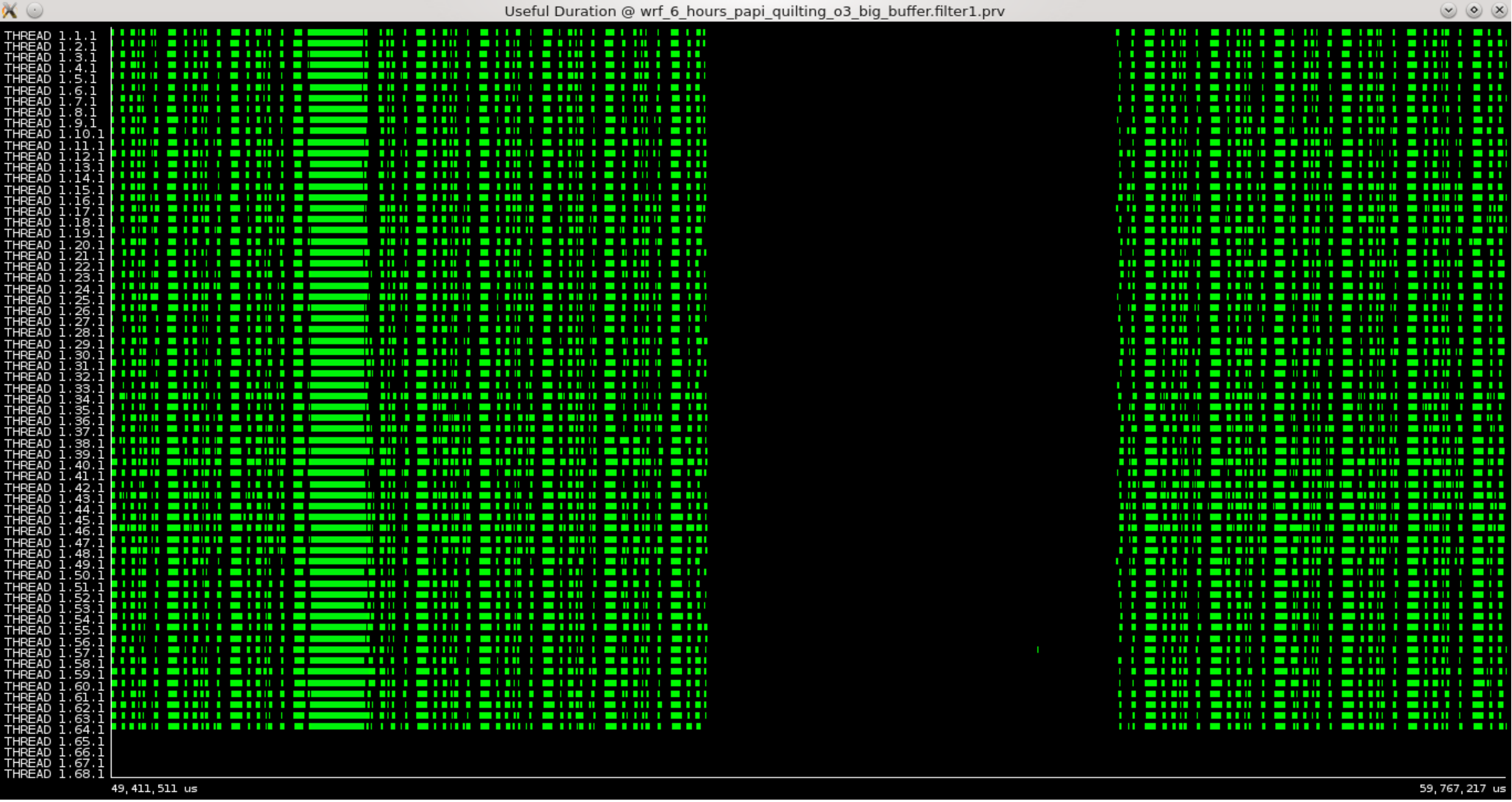
- If we zoom, we have the following
- Similar problems with some MPI_Wait calls



New window #5 @ wrf_6_hours_papi_quilting_o3_big_buffer.chop3.prv

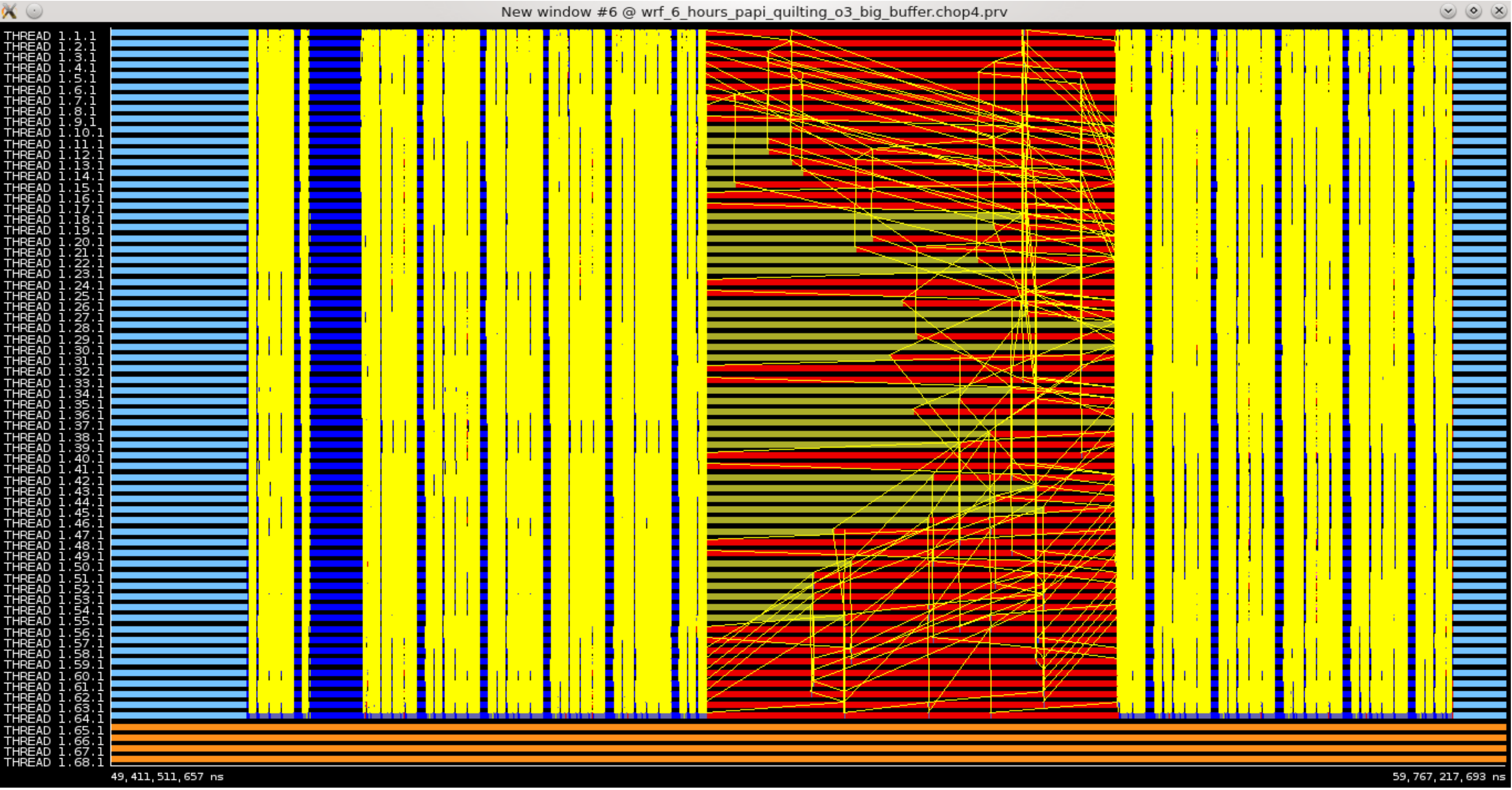
Trace Analysis

- During the visualization of the computation areas we can see a large black area



Trace Analysis

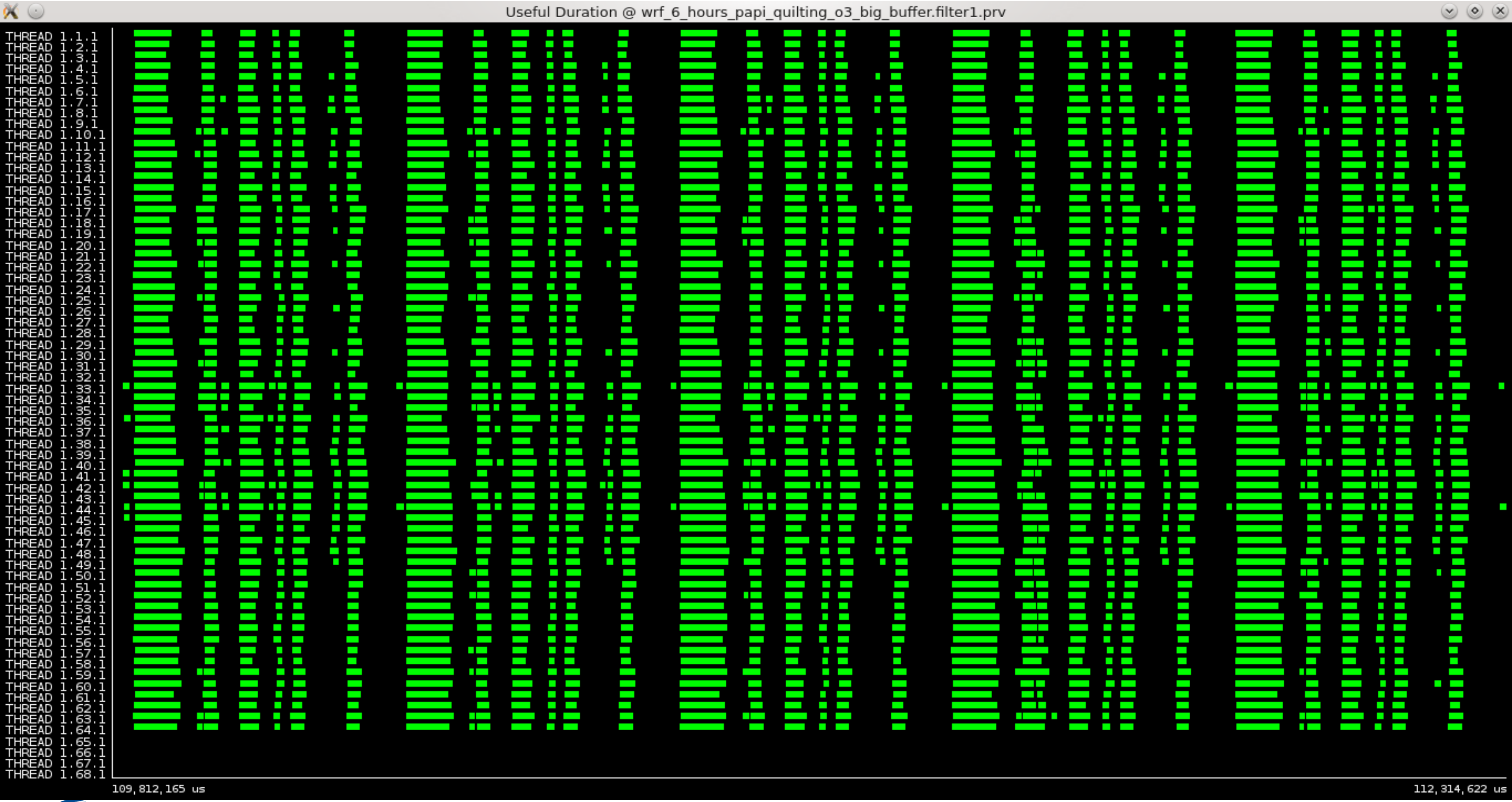
- The previous black area is caused by communication perturbation
- The brown area is the I/O caused from the flushing of the traces on the hard disk



New window #6 @ wrf_6_hours_papi_quilting_o3_big_buffer.chop4.prv

Trace Analysis

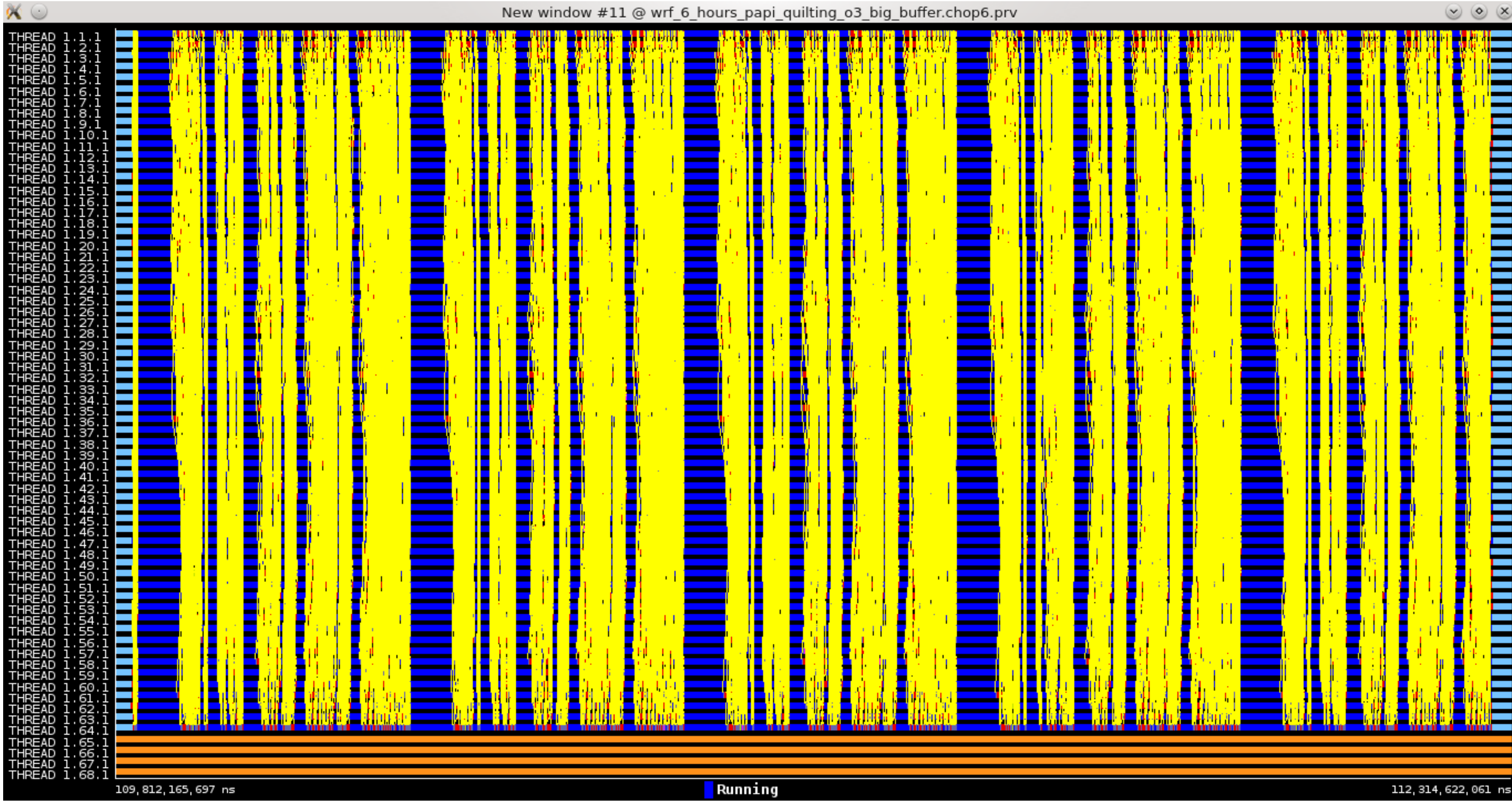
- Observing the patterns from the computation phases is a good approach to know where we should focus (we have 5 similar phases)



Useful Duration @ wrf_6_hours_papi_quilting_o3_big_buffer.filter1.prv

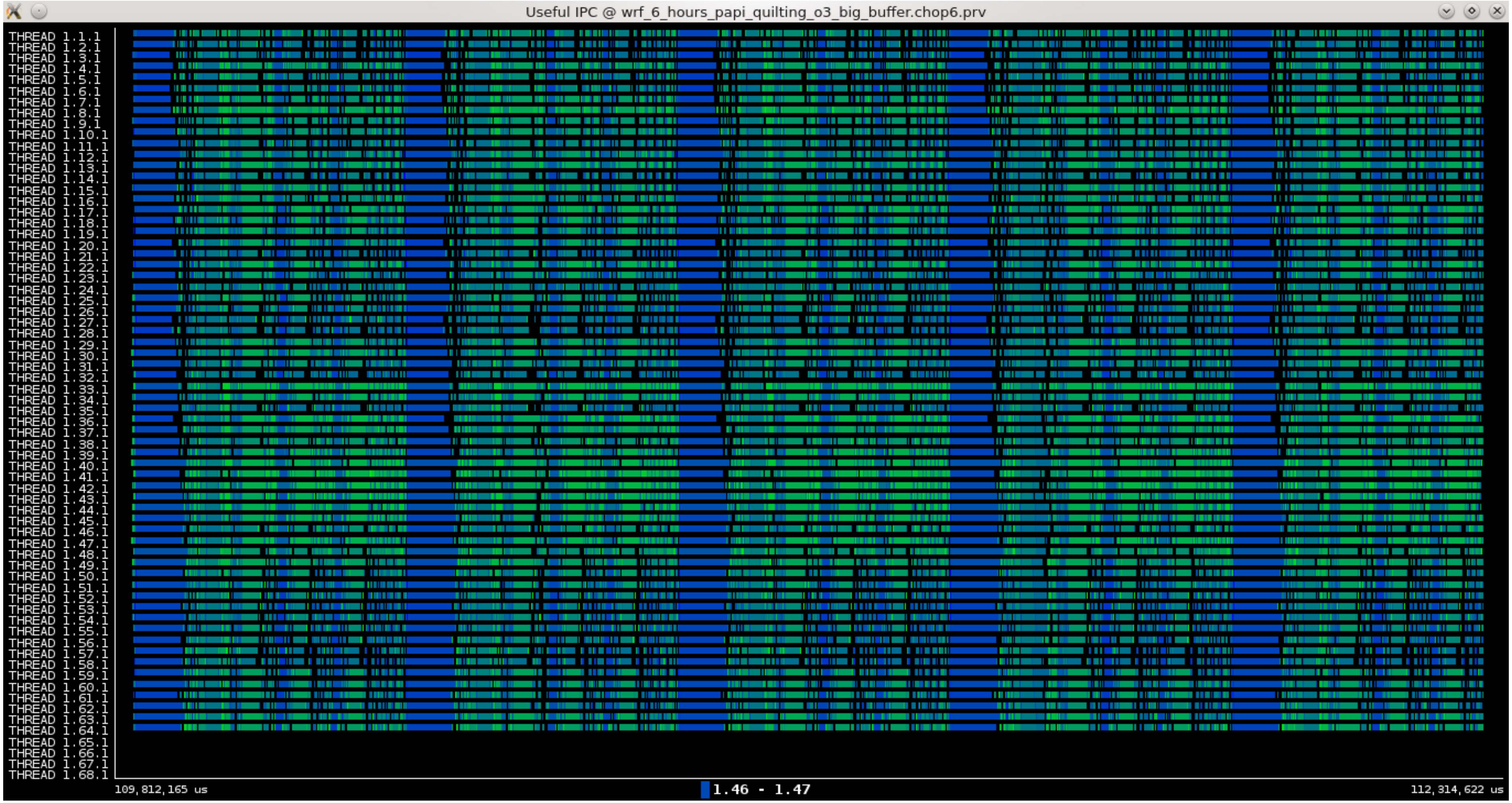
Trace Analysis

- The previous visualization with the communications and any extra metric



Trace Analysis

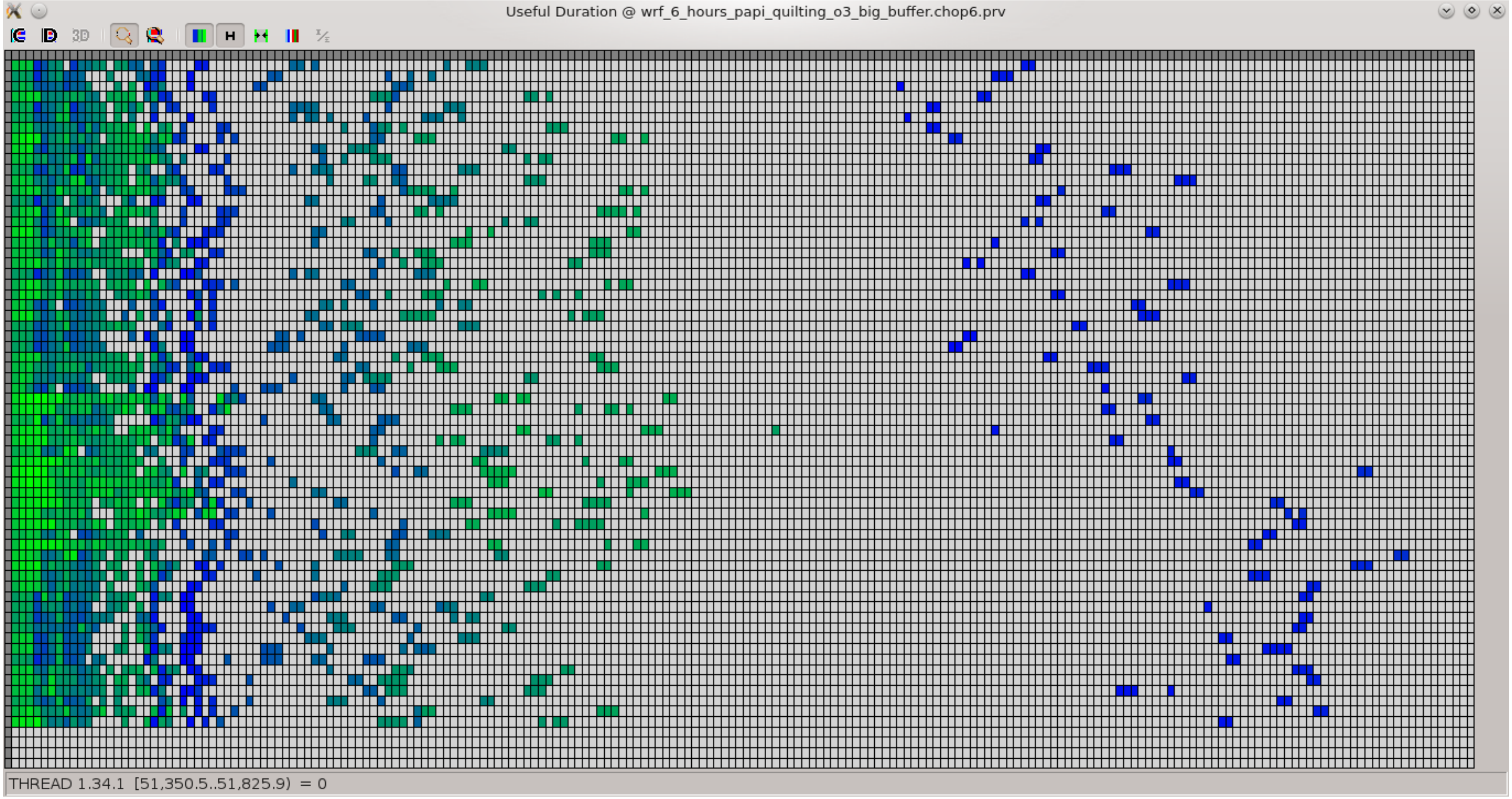
- Useful instructions per cycle. A value close to 2 is good. Much lower value means that the code should be improved



Useful IPC @ wrf_6_hours_papi_quilting_o3_big_buffer.chop6.prv

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Useful duration per process

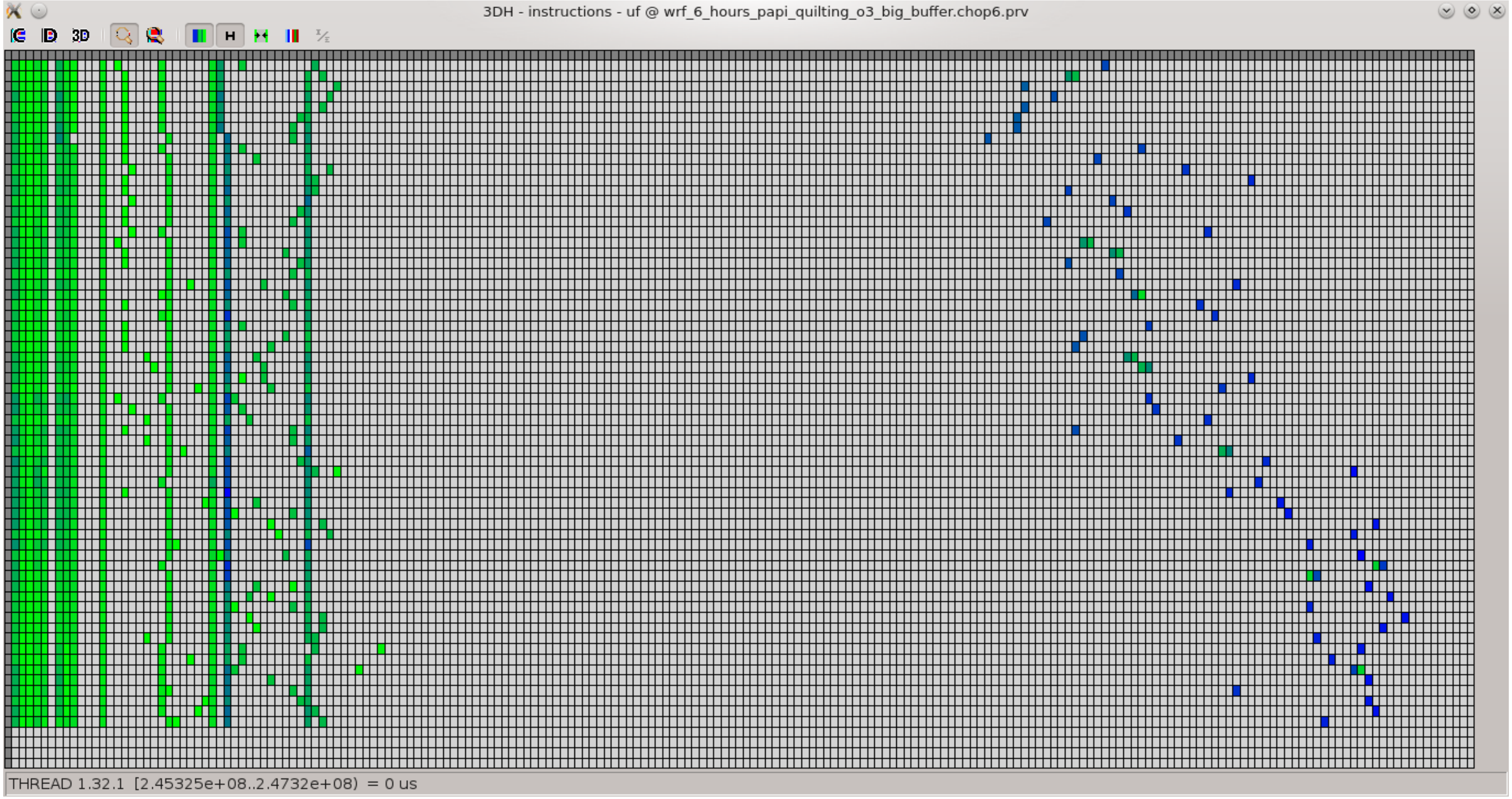


Useful Duration @ wrf_6_hours_papi_quilting_o3_big_buffer.chop6.prv

THREAD 1.34.1 [51,350.5..51,825.9) = 0

Trace Analysis

- Instructions per process
- In general we should have vertical lines



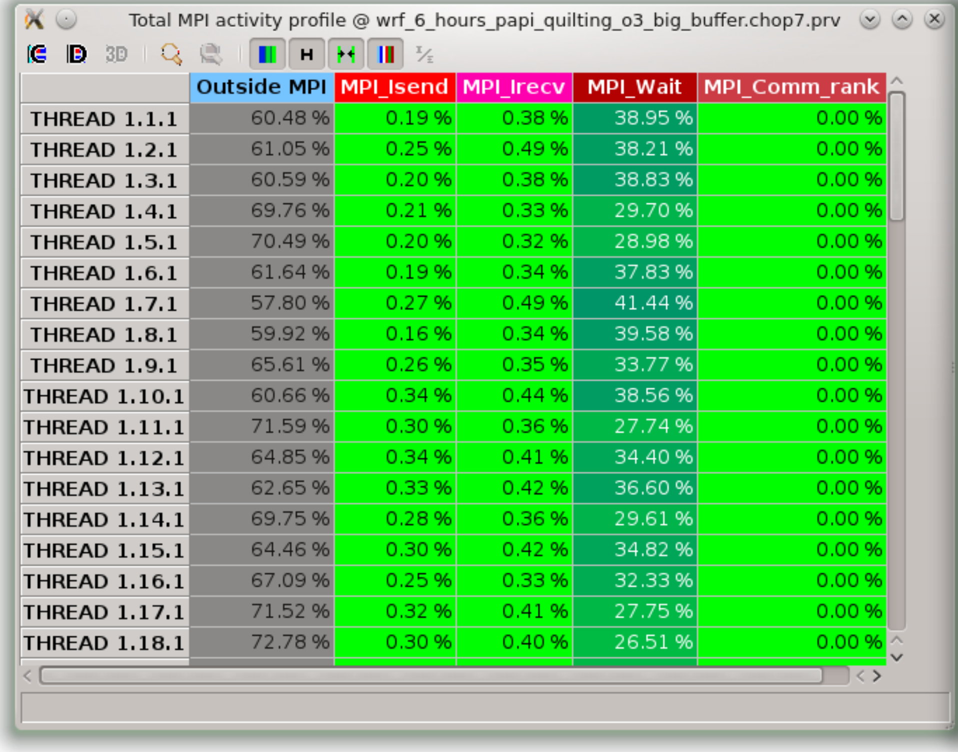
3DH - instructions - uf @ wrf_6_hours_papi_quilting_o3_big_buffer.chop6.prv
3D
H

THREAD 1.32.1 [2.45325e+08..2.4732e+08) = 0 us

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

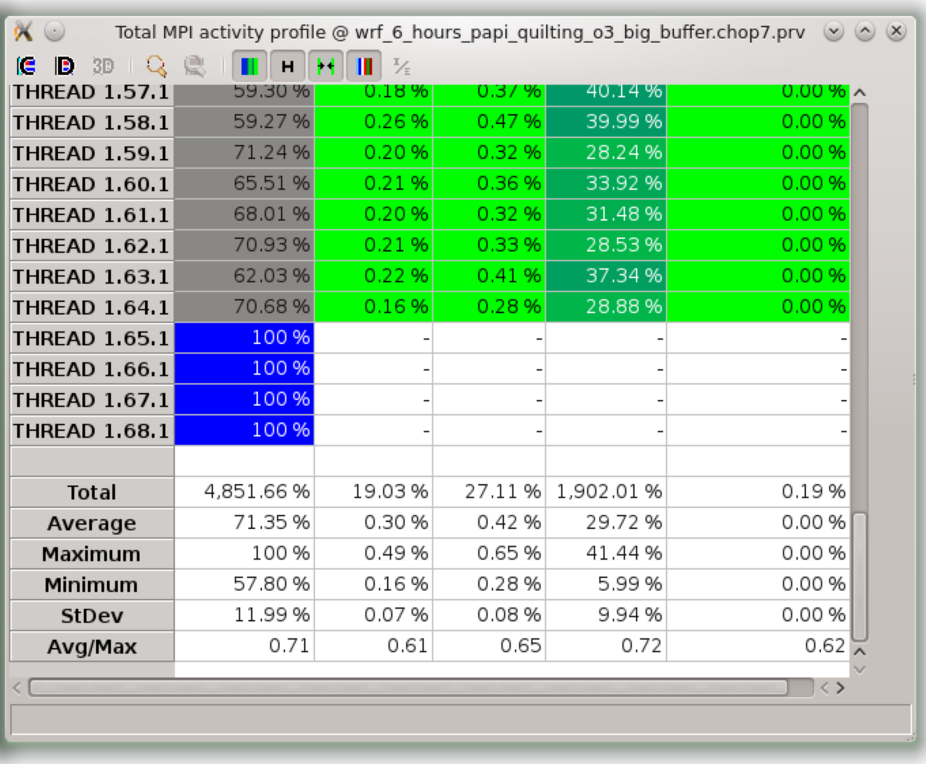
Trace Analysis

- MPI calls profiling



Total MPI activity profile @ wrf_6_hours_papi_quilting_o3_big_buffer.chop7.prv

	Outside MPI	MPI_Isend	MPI_Irecv	MPI_Wait	MPI_Comm_rank
THREAD 1.1.1	60.48 %	0.19 %	0.38 %	38.95 %	0.00 %
THREAD 1.2.1	61.05 %	0.25 %	0.49 %	38.21 %	0.00 %
THREAD 1.3.1	60.59 %	0.20 %	0.38 %	38.83 %	0.00 %
THREAD 1.4.1	69.76 %	0.21 %	0.33 %	29.70 %	0.00 %
THREAD 1.5.1	70.49 %	0.20 %	0.32 %	28.98 %	0.00 %
THREAD 1.6.1	61.64 %	0.19 %	0.34 %	37.83 %	0.00 %
THREAD 1.7.1	57.80 %	0.27 %	0.49 %	41.44 %	0.00 %
THREAD 1.8.1	59.92 %	0.16 %	0.34 %	39.58 %	0.00 %
THREAD 1.9.1	65.61 %	0.26 %	0.35 %	33.77 %	0.00 %
THREAD 1.10.1	60.66 %	0.34 %	0.44 %	38.56 %	0.00 %
THREAD 1.11.1	71.59 %	0.30 %	0.36 %	27.74 %	0.00 %
THREAD 1.12.1	64.85 %	0.34 %	0.41 %	34.40 %	0.00 %
THREAD 1.13.1	62.65 %	0.33 %	0.42 %	36.60 %	0.00 %
THREAD 1.14.1	69.75 %	0.28 %	0.36 %	29.61 %	0.00 %
THREAD 1.15.1	64.46 %	0.30 %	0.42 %	34.82 %	0.00 %
THREAD 1.16.1	67.09 %	0.25 %	0.33 %	32.33 %	0.00 %
THREAD 1.17.1	71.52 %	0.32 %	0.41 %	27.75 %	0.00 %
THREAD 1.18.1	72.78 %	0.30 %	0.40 %	26.51 %	0.00 %



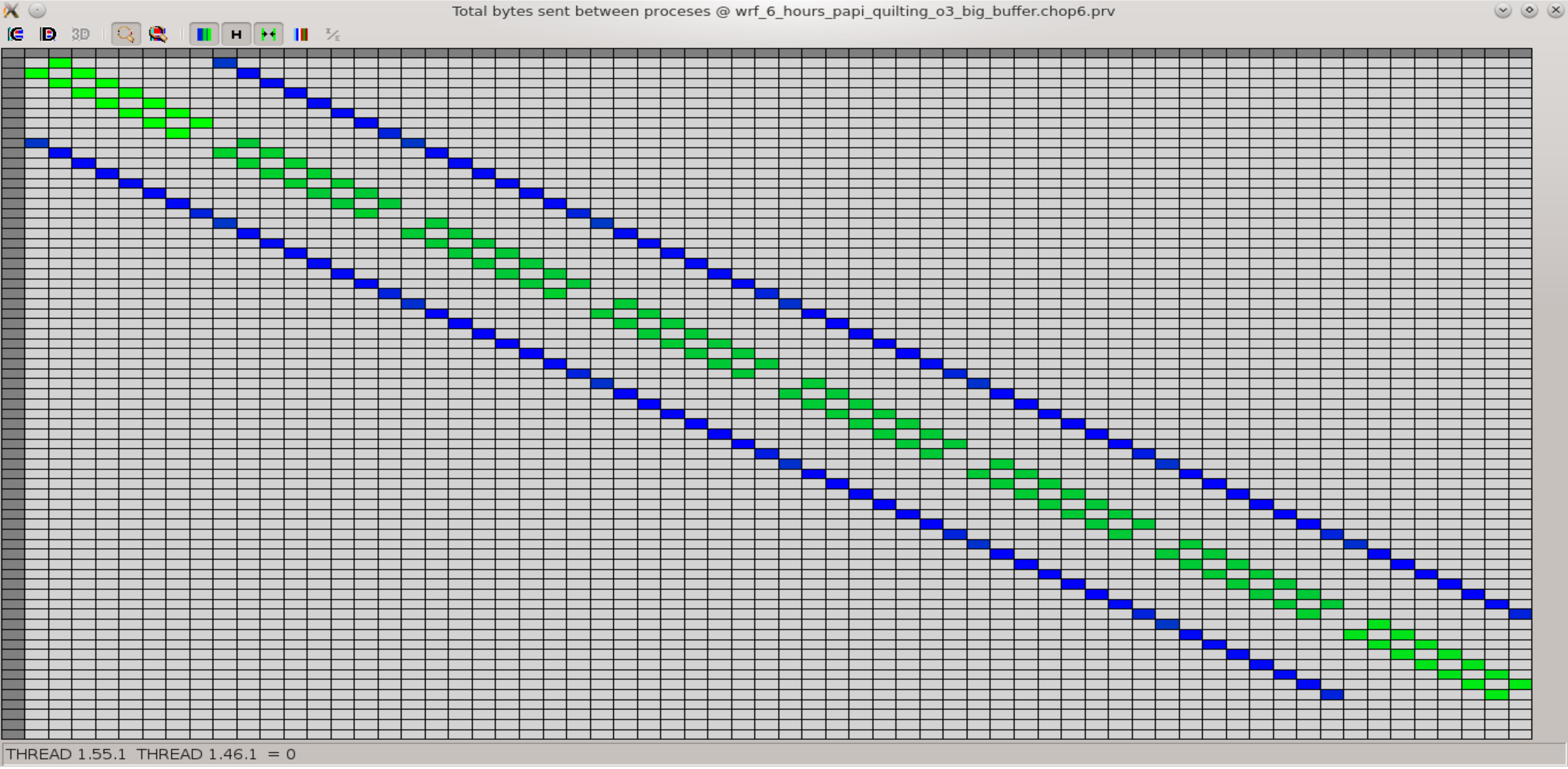
Total MPI activity profile @ wrf_6_hours_papi_quilting_o3_big_buffer.chop7.prv

THREAD 1.57.1	59.30 %	0.18 %	0.37 %	40.14 %	0.00 %
THREAD 1.58.1	59.27 %	0.26 %	0.47 %	39.99 %	0.00 %
THREAD 1.59.1	71.24 %	0.20 %	0.32 %	28.24 %	0.00 %
THREAD 1.60.1	65.51 %	0.21 %	0.36 %	33.92 %	0.00 %
THREAD 1.61.1	68.01 %	0.20 %	0.32 %	31.48 %	0.00 %
THREAD 1.62.1	70.93 %	0.21 %	0.33 %	28.53 %	0.00 %
THREAD 1.63.1	62.03 %	0.22 %	0.41 %	37.34 %	0.00 %
THREAD 1.64.1	70.68 %	0.16 %	0.28 %	28.88 %	0.00 %
THREAD 1.65.1	100 %	-	-	-	-
THREAD 1.66.1	100 %	-	-	-	-
THREAD 1.67.1	100 %	-	-	-	-
THREAD 1.68.1	100 %	-	-	-	-
Total	4,851.66 %	19.03 %	27.11 %	1,902.01 %	0.19 %
Average	71.35 %	0.30 %	0.42 %	29.72 %	0.00 %
Maximum	100 %	0.49 %	0.65 %	41.44 %	0.00 %
Minimum	57.80 %	0.16 %	0.28 %	5.99 %	0.00 %
StDev	11.99 %	0.07 %	0.08 %	9.94 %	0.00 %
Avg/Max	0.71	0.61	0.65	0.72	0.62

- For the study of the statistics we exclude the I/O processes (scripting)
- Maximum value: 93.21% (communication efficiency)
- Average value: 69.55% (parallel efficiency)
- Avg/max value: 74.6% (global load balance)
- Note: we study just a small part of the whole execution

Trace Analysis

- Communication matrix
- The previous mentioned mapping from Peter Johnsen is validated

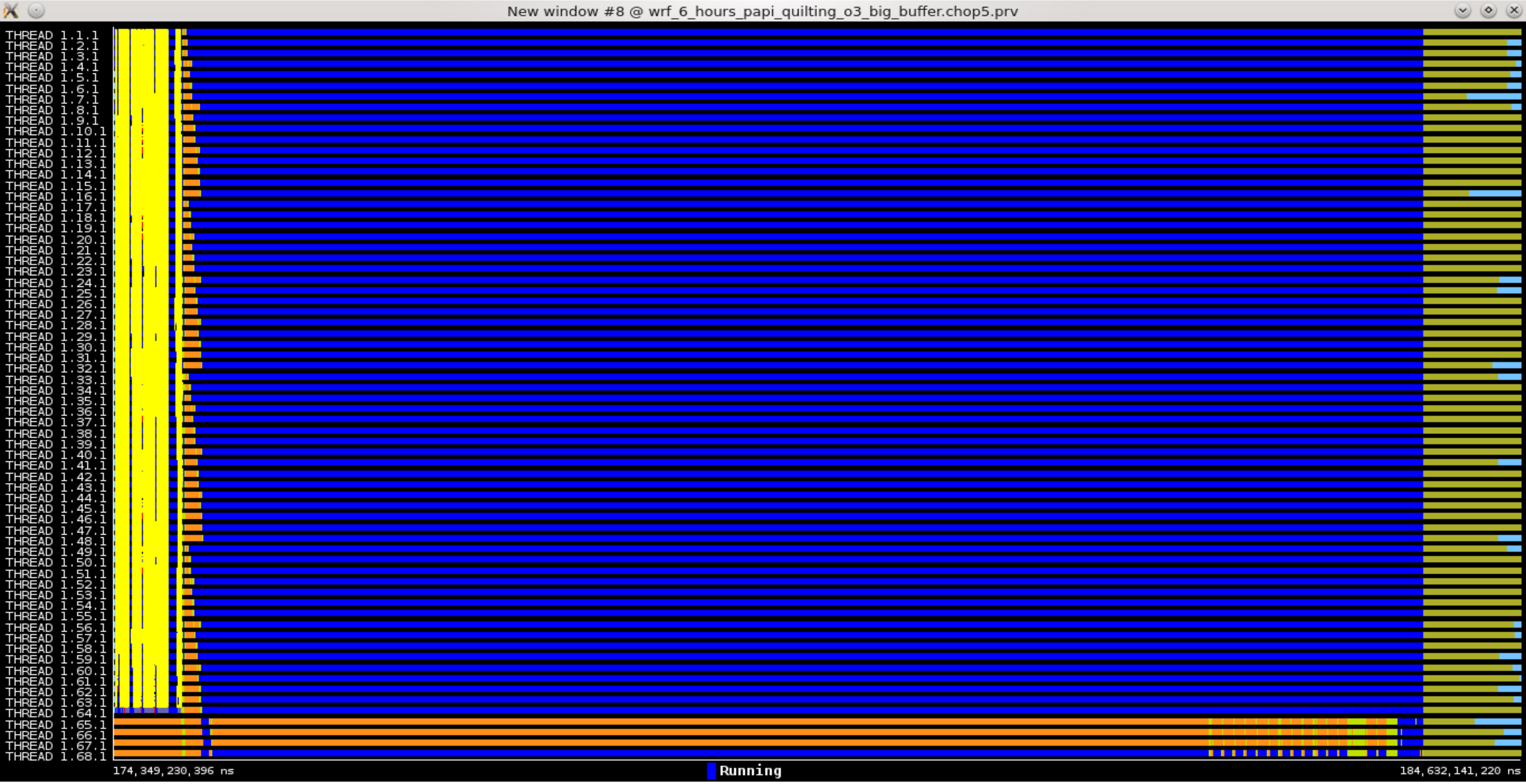


Total bytes sent between proceses @ wrf_6_hours_papi_quilting_o3_big_buffer.chop6.prv

THREAD 1.55.1 THREAD 1.46.1 = 0

Trace Analysis – End of the trace

- There is communication between the write tasks (last four).
- All the processes wait till the write tasks finish for the case of I/O quilting.



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Conclusions

- Optimize first your application through the provided options, you can be surprised

- Be careful about the combination of the optimization options

- Different number of processors and workload does not mean that they can be optimized with the same approach

- Paraver can provide a lot of insight information about the behavior of an earth science model

- Integrate new technologies