Ferramentas de Avaliação de Desempenho

Roteiro

Scalasca

https://www.scalasca.org

Scalasca

Exemplo: NAS Parallel Benchmarks (NPB)

```
scalasca/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_scalasca
       README
```

Scalasca

Exemplo: NAS Parallel Benchmarks (NPB)

```
scalasca/
    NPB3.3.1-MZ/
        NPB3.3-MZ-MPI/
        NPB3.3-MZ-OMP/
        NPB3.3-MZ-SER/
        Changes.log
        env_scalasca
        README
```

Scalasca

Preparando o ambiente

```
$ cat env_scalasca

module load openmpi/gnu/2.0.4.2
module load scalasca/2.4_openmpi_gnu
module load papi/5.5.1.0
module load papi-devel/5.5.1.0
```

Scalasca

Preparando o ambiente

```
$ source env_scalasca
scalasca 2.4 for GNU OpenMPI loaded
Compiled with openMPI 2.0.4.2 and GNU compilers Red Hat 4.8.5-36

$ scalasca
Scalasca 2.4
Toolset for scalable performance analysis of large-scale parallel applications
usage: scalasca [-v][-n] action
    1. prepare application objects and executable for measurement:
       scalasca -instrument <compile-or-link-command> # skin (using scorep)
    2. run application under control of measurement system:
       scalasca -analyze <application-launch-command> # scan
    3. interactively explore measurement analysis report:
       scalasca -examine <experiment-archive|report>  # square

Options:
   -c, --show-config     show configuration summary and exit
   -h, --help            show this help and exit
   -n, --dry-run         show actions without taking them
       --quickref        show quick reference guide and exit
       --remap-specfile  show path to remapper specification file and exit
   -v, --verbose         enable verbose commentary
   -V, --version         show version information and exit
```

Scalasca

scalasca -instrument (skin/scorep)

```
scalasca/
    NPB3.3.1-MZ/
        NPB3.3-MZ-MPI/
        NPB3.3-MZ-OMP/
        NPB3.3-MZ-SER/
        Changes.log
        env_scalasca
        README
```

Scalasca

scalasca -instrument (skin/scorep)

```
scalasca/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
          bin/
          BT-MZ/
          common/
          config/
          LU-MZ/
          SP-MZ/
          sys/
          Makefile
          README
          README.install
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_scalasca
       README
```

Scalasca

scalasca -instrument (skin/scorep)

```
scalasca/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
          bin/
          BT-MZ/
          common/
          config/
          LU-MZ/
          SP-MZ/
          sys/
          Makefile
          README
          README.install
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_scalasca
       README
```

Scalasca

scalasca -instrument (skin/scorep)

```
config/
   NAS.samples
   make.def -> make_scalasca.def
   make.def.template
   make_scalasca.def
   suite.def
   suite.def.template
```

Scalasca

scalasca -instrument (skin/scorep)

```
config/
   NAS.samples
   make.def -> make_scalasca.def
   make.def.template
   make_scalasca.def
   suite.def
   suite.def.template
```

Scalasca

scalasca -instrument (skin/scorep)

```
$ cat make_scalasca.def

#-------------------------------------------------------------------------
# This is the fortran compiler used for fortran programs
#-------------------------------------------------------------------------
#F77 = mpif77
F77 = scalasca -instrument mpif77
#F77 = scorep mpif77


#-------------------------------------------------------------------------
# This is the C compiler used for C programs
#-------------------------------------------------------------------------
#CC = mpicc
CC = scalasca -instrument mpicc
#CC = scorep mpicc
```

Scalasca

NPB: benchmark, classe e número de processos MPI

```
config/
   NAS.samples
   make.def -> make_scalasca.def
   make.def.template
   make_scalasca.def
   suite.def
   suite.def.template
```

Scalasca

NPB: benchmark, classe e número de processos MPI

```
config/
   NAS.samples
   make.def -> make_scalasca.def
   make.def.template
   make_scalasca.def
   suite.def
   suite.def.template
```

Estudo de caso

NPB: benchmark, classe e número de processos MPI

```
$ cat suite.def

# config/suite.def
# This file is used to build several benchmarks with a single command.
# Typing "make suite" in the main directory will build all the benchmarks
# specified in this file.
# Each line of this file contains a benchmark name, class, and number
# of nodes. The name is one of "sp-mz", "bt-mz", and "lu-mz".
# The class is one of "S", "W", and "A" through "F".
# No blank lines.
# The following example builds serial sample sizes of all benchmarks.
#sp-mz S 1
#lu-mz S 1
#bt-mz S 2
bt-mz    S       1
bt-mz    S       2
bt-mz    S       4
bt-mz    W       1
bt-mz    W       2
bt-mz    W       4
bt-mz    W       8
bt-mz    W       16
```

Estudo de caso

NPB: compilação

```
$ cd ..
$ make suite %compila o NPB
$ cd bin
```

Estudo de caso

NPB: compilação

```
$ ls -A1
bt-mz.S.1
bt-mz.S.2
bt-mz.S.4
bt-mz.W.1
bt-mz.W.2
bt-mz.W.4
BULL_srun_scan_prof.sh
BULL_srun_scan_trace.sh
BULL_srun_scan_trace_filt.sh
```

Scalasca

scalasca -analyze: coleta de dados da execução

```
$ ls -A1
bt-mz.S.1
bt-mz.S.2
bt-mz.S.4
bt-mz.W.1
bt-mz.W.2
bt-mz.W.4
BULL_srun_scan_prof.sh
BULL_srun_scan_trace.sh
BULL_srun_scan_trace_filt.sh
```

Scalasca

BULL_srun_scalasca.sh

```
#!/bin/bash

#SBATCH --nodes=1                          # here the number of nodes
#SBATCH --ntasks=1                         # here total number of mpi tasks
#SBATCH --cpus-per-task=1                  # number of cores per node
#SBATCH -p cpu_dev                         # target partition
#SBATCH --threads-per-core=1
#SBATCH -J NPB_BT-MZ                       # job name
#SBATCH --time=00:10:00                    # time limit
#SBATCH --exclusive                        # to have exclusive use of your nodes

echo "Cluster configuration:"
echo "==="
echo "Partition: " $SLURM_JOB_PARTITION
echo "Number of nodes: " $SLURM_NNODES
echo "Number of MPI processes: " $SLURM_NTASKS " (" $SLURM_NNODES " nodes)"
echo "Number of MPI processes per node: " $SLURM_NTASKS_PER_NODE
echo "Number of threads per MPI process: " $SLURM_CPUS_PER_TASK
echo "NPB Benchmark: " $1
echo "Bechmark class problem: " $2

#####################################
#            COMPILER               #
#####################################
module load openmpi/gnu/2.0.4.2
module load scalasca/2.4_openmpi_gnu
module load papi/5.5.1.0
module load papi-devel/5.5.1.0
```

Scalasca (cont.)

```
bench=${1}
class=${2}
executable="${bench}.${class}.$SLURM_NTASKS"

export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK
#export SCOREP_METRIC_PAPI='PAPI_TOT_INS,PAPI_TOT_CYC,PAPI_L2_DCA,PAPI_L2_DCM'

scalasca -analyze -s srun --resv-ports -n $SLURM_NTASKS ${executable}

scorepdirorig="scorep_${bench}_${SLURM_NTASKS}x${SLURM_CPUS_PER_TASK}_sum"
scorepdirdest="${bench}_${class}_MPI-${SLURM_NTASKS}_OMP-${SLURM_CPUS_PER_TASK}_JOBID-${SLURM_JOB
mv $scorepdirorig $scorepdirdest
mv slurm-${SLURM_JOBID}.out $scorepdirdest/

#pós-processamento contendo análise mais detalhada
scalasca -examine -s $scorepdirdest

scalascaresultdir=profiling/scalasca/NUMNODES-$SLURM_JOB_NUM_NODES
mkdir -p $scalascaresultdir

mv $scorepdirdest/ $scalascaresultdir/
```

Estudo de caso

NPB: submetendo job

```
$ sbatch BULL_srun_scan_prof.sh bt-mz S
Submitted batch job 437607
$ squeue -u $USER
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
437607 treinamen NPB_BT-M professo  R       0:02      1 sdumont3000
```

Estudo de caso

NPB: perfil de desempenho

```
$ ls -A1
bt-mz.S.1
bt-mz.S.2
bt-mz.S.4
bt-mz.W.1
bt-mz.W.2
bt-mz.W.4
BULL_srun_scan_prof.sh
BULL_srun_scan_trace.sh
BULL_srun_scan_trace_filt.sh
scorep_bt-mz_S_sum_MPI-1_OMP-1_JOBID-437607/
```

Estudo de caso

NPB: perfil de desempenho

```
scorep_bt-mz_S_sum_MPI-1_OMP-1_JOBID-437607/
   profile.cubex
   summary.cubex
   scorep.score
   scorep.cfg
   scorep.log
   slurm-437607.out
```

Estudo de caso

NPB: perfil de desempenho

```
scorep_bt-mz_S_sum_MPI-1_OMP-1_JOBID-437607/
   profile.cubex     --> análise básica, a partir de dados coletados durante execução
   summary.cubex     --> análise mais detalhada
   scorep.score      --> relatório formato texto com a análise
   scorep.cfg        --> configuração da coleta de dados
   scorep.log        --> output da aplicação
   slurm-437607.out --> output do SLURM
```

Estudo de caso

NPB: perfil de desempenho

```
scorep_bt-mz_S_sum_MPI-1_OMP-1_JOBID-437607/
   profile.cubex
   summary.cubex
   scorep.score
   scorep.cfg
   scorep.log
   slurm-437607.out
```

Estudo de caso

```
$ cat slurm-437607.out

Cluster configuration:
===
Partition:  treinamento
Number of nodes:  1
Number of MPI processes:  1  ( 1  nodes)
Number of MPI processes per node:  1
Number of threads per MPI process:  1
NPB Benchmark:  bt-mz
Bechmark class problem:  S
scalasca 2.4 for GNU OpenMPI loaded
Compiled with openMPI 2.0.4.2 and GNU compilers Red Hat 4.8.5-36
S=C=A=N: Scalasca 2.4 runtime summarization
S=C=A=N: ./scorep_bt-mz_1x1_sum experiment archive
S=C=A=N: Tue Jan 28 14:17:02 2020: Collect start
/usr/bin/srun --resv-ports -n 1 /scratch/treinamento/professor/MC1-I/tools/scalasca/NPB3.3.1-MZ/NP
[1580231823.240921] [sdumont5000:73441:0]         mxm.c:196  MXM  WARN  The 'ulimit -s' on the sys
[1580231823.242634] [sdumont5000:73441:0]         mxm.c:196  MXM  WARN  The 'ulimit -s' on the sys


 NAS Parallel Benchmarks (NPB3.3-MZ-MPI) - BT-MZ MPI+OpenMP Benchmark

 Number of zones:   2 x   2
 Iterations:  60    dt:   0.010000
 Number of active processes:     1
```

Estudo de caso (cont.)

```
Use the default load factors with threads
Total number of threads:      1   (  1.0 threads/process)

Calculated speedup =      1.00

Time step    1
Time step   20
Time step   40
Time step   60
Verification being performed for class S
accuracy setting for epsilon =  0.1000000000000E-07
Comparison of RMS-norms of residual
          1 0.1047687395830E+04 0.1047687395830E+04 0.1751386499571E-12
          2 0.9419911314792E+02 0.9419911314792E+02 0.1478425555772E-13
          3 0.2124737403068E+03 0.2124737403068E+03 0.9002435039286E-13
          4 0.1422173591794E+03 0.1422173591794E+03 0.3089634277625E-12
          5 0.1135441572375E+04 0.1135441572375E+04 0.3103895484466E-13
Comparison of RMS-norms of solution error
          1 0.1775416062982E+03 0.1775416062982E+03 0.1922618237923E-12
          2 0.1875540250835E+02 0.1875540250835E+02 0.1558955269742E-12
          3 0.3863334844506E+02 0.3863334844506E+02 0.1105356386074E-12
          4 0.2634713890362E+02 0.2634713890362E+02 0.3991337551951E-13
          5 0.1965566269675E+03 0.1965566269675E+03 0.2336704854379E-12
Verification Successful


BT-MZ Benchmark Completed.
Class           =                        S
Size            =             24x   24x  6
Iterations      =                       60
```

Estudo de caso (cont.)

```
 Time in seconds =                    0.35
 Total processes =                       1
 Total threads   =                       1
 Mop/s total     =                 1093.41
 Mop/s/thread    =                 1093.41
 Operation type  =           floating point
 Verification    =               SUCCESSFUL
 Version         =                    3.3.1
 Compile date    =              21 Jan 2020

 Compile options:
    F77           = scalasca -instrument mpif77
    FLINK         = $(F77)
    F_LIB         = (none)
    F_INC         = (none)
    FFLAGS        = -O3 -fopenmp
    FLINKFLAGS    = $(FFLAGS)
    RAND          = (none)


 Please send all errors/feedbacks to:

 NPB Development Team
 npb@nas.nasa.gov


S=C=A=N: Tue Jan 28 14:17:03 2020: Collect done (status=0) 1s
S=C=A=N: ./scorep_bt-mz_1x1_sum complete.
INFO: Post-processing runtime summarization report...
/opt/bullxde/utils/scalasca/openmpi-gnu/scorep/bin/scorep-score -r ./scorep_bt-mz_S_sum_MPI-1_OMP-
```

Estudo de caso (cont.)

```
INFO: Score report written to ./scorep_bt-mz_S_sum_MPI-1_OMP-1_JOBID-437607/scorep.score
```

Estudo de caso

NPB: submetendo job

```
$ sbatch BULL_srun_scan_prof.sh bt-mz W
Submitted batch job 437632
$ squeue -u $USER
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
% 437632 treinamen NPB_BT-M professo  R       0:02      1 sdumont3000
```

Estudo de caso

NPB: perfil de desempenho

```
scorep_bt-mz_W_sum_MPI-1_OMP-1_JOBID-437632/
   profile.cubex
   summary.cubex
   scorep.score
   scorep.cfg
   scorep.log
   slurm-437632.out
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=1 BULL_srun_scan_prof.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     1

 Use the default load factors with threads
 Total number of threads:      1   (  1.0 threads/process)

 Calculated speedup =       1.00

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                    12.59
 Total processes =                       1
 Total threads   =                       1
 Mop/s total     =                 1140.40
 Mop/s/thread    =                 1140.40
 Operation type  =           floating point
 Verification    =               SUCCESSFUL
 Version         =                   3.3.1

S=C=A=N: Tue Jan 28 14:57:17 2020: Collect done (status=0) 14s
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=2 BULL_srun_scan_prof.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     2

 Use the default load factors with threads
 Total number of threads:      2   (  1.0 threads/process)

 Calculated speedup =       1.98

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =              64x   64x  8
 Iterations      =                      200
 Time in seconds =                     6.46
 Total processes =                        2
 Total threads   =                        2
 Mop/s total     =                  2219.05
 Mop/s/thread    =                  1109.53
 Operation type  =            floating point
 Verification    =                SUCCESSFUL
 Version         =                     3.3.1

S=C=A=N: Tue Jan 28 15:01:09 2020: Collect done (status=0) 8s
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=4 BULL_srun_scan_prof.sh bt-mz W

 Number of zones:    4 x    4
 Iterations: 200    dt:    0.000800
 Number of active processes:     4

 Use the default load factors with threads
 Total number of threads:      4   (  1.0 threads/process)

 Calculated speedup =      3.95

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =              64x   64x  8
 Iterations      =                      200
 Time in seconds =                     3.40
 Total processes =                        4
 Total threads   =                        4
 Mop/s total     =                  4206.28
 Mop/s/thread    =                  1051.57
 Operation type  =            floating point
 Verification    =                SUCCESSFUL
 Version         =                     3.3.1

S=C=A=N: Tue Jan 28 15:01:34 2020: Collect done (status=0) 5s
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=8 BULL_srun_scan_prof.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     8

 Use the default load factors with threads
 Total number of threads:      8   (  1.0 threads/process)

 Calculated speedup =      4.87

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                     2.82
 Total processes =                        8
 Total threads   =                        8
 Mop/s total     =                  5086.41
 Mop/s/thread    =                   635.80
 Operation type  =           floating point
 Verification    =               SUCCESSFUL
 Version         =                    3.3.1

S=C=A=N: Wed Jan 29 10:51:43 2020: Collect done (status=0) 4s
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=16 BULL_srun_scan_prof.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:    16

 Use the default load factors with threads
 Total number of threads:     16   (  1.0 threads/process)

 Calculated speedup =       4.87

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                     2.84
 Total processes =                       16
 Total threads   =                       16
 Mop/s total     =                  5047.62
 Mop/s/thread    =                   315.48
 Operation type  =           floating point
 Verification    =               SUCCESSFUL
 Version         =                    3.3.1

S=C=A=N: Wed Jan 29 10:52:03 2020: Collect done (status=0) 6s
```

Visualizando: CubeGUI

- O CubeGUI pode ser baixado e instalado para visualizar os resultados obtidos com o Scalasca
- `https://www.scalasca.org/scalasca/software/cube-4.x/download.html`
- Binários prontos em "Supplementary packages for download (Comfort zone)"
- Resultados previamente obtidos no SDumont estão no arquivo **`profiling_scalasca_sequana.zip`** do repositório no GitHub:

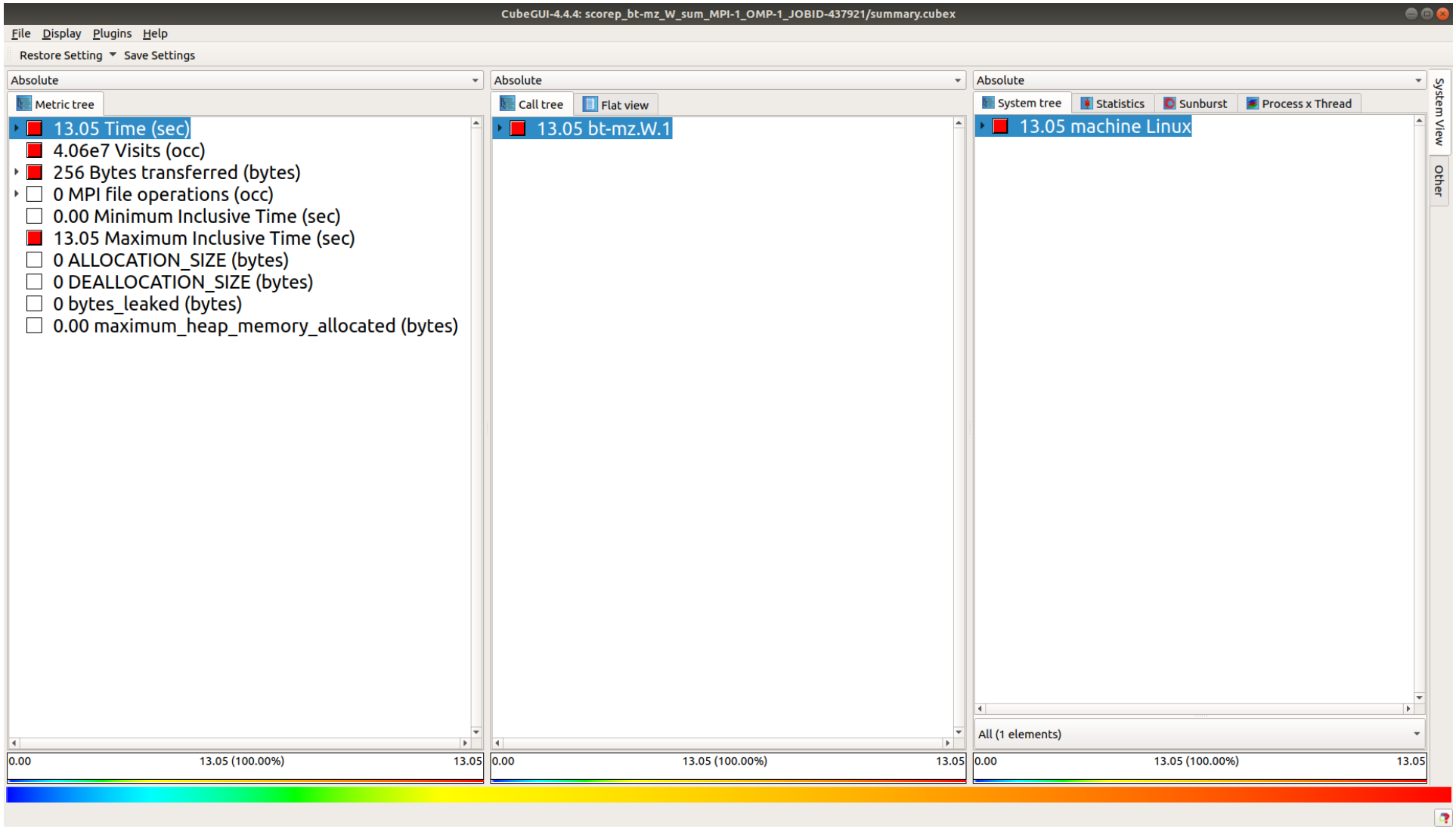
```
git clone https://github.com/robertopsouto/ESD2023.git
ESD2023/MC-SD01-I/scalasca/profiling_scalasca_sequana.zip
```

NPB: estudo de caso

```
$ cd profiling/NUMNODES-1/scorep_bt-mz_W_sum_MPI-1_OMP-1_JOBID-437921
$ cube summary.cubex
```

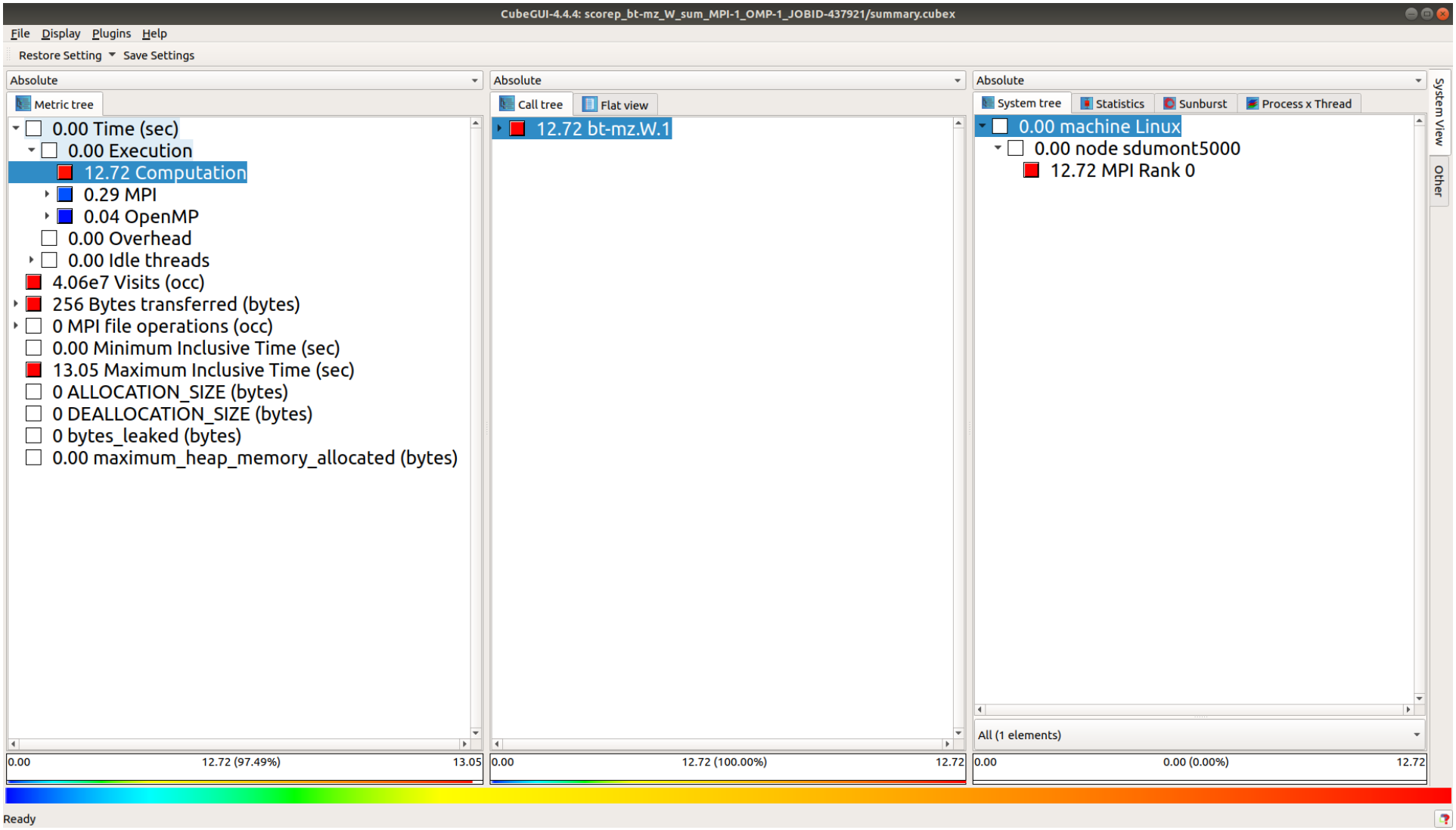
Visualizando: CubeGUI

-nodes=1 -ntasks=1



Visualizando: CubeGUI

-nodes=1 -ntasks=1



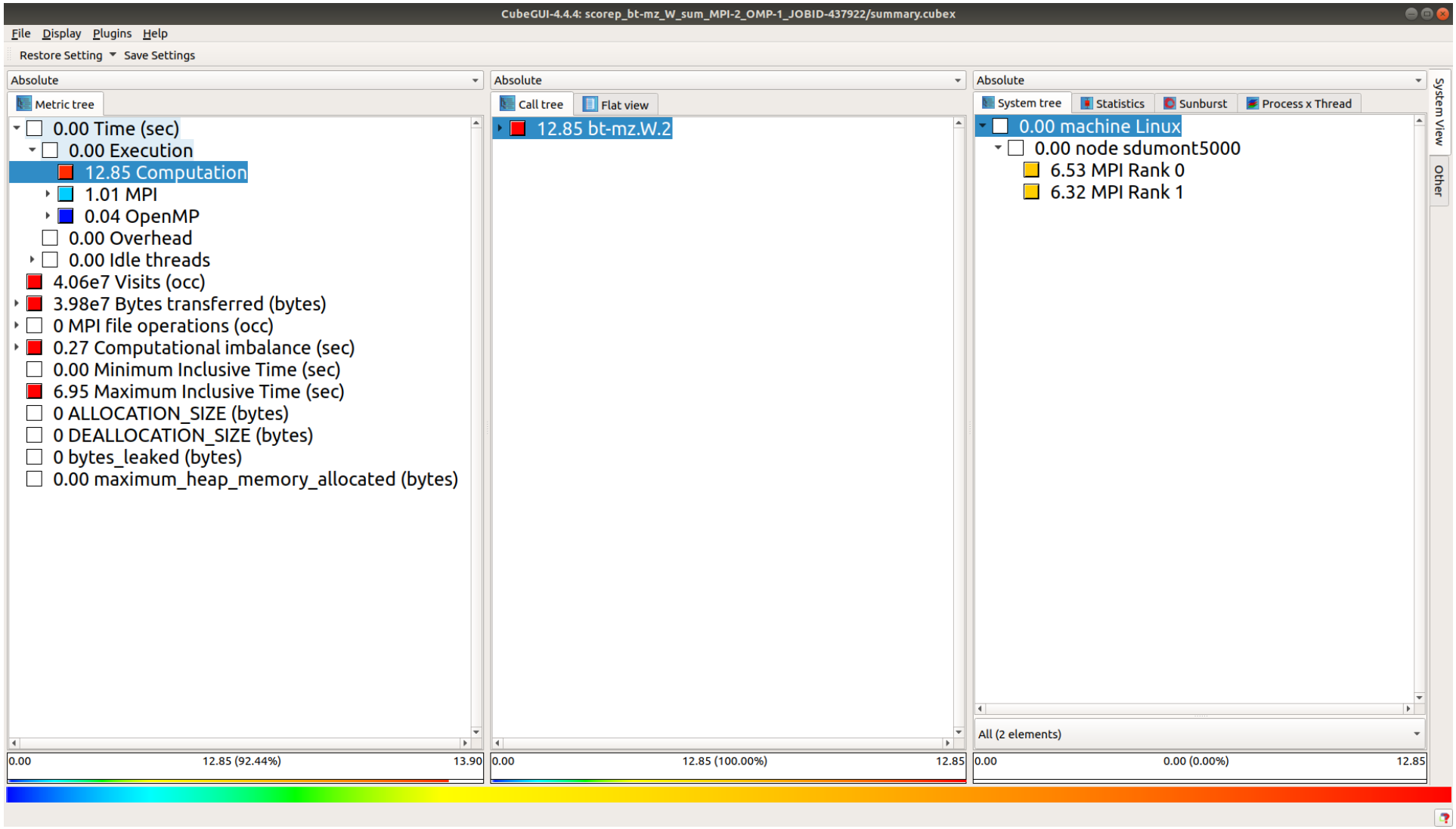
Visualizando: CubeGUI

divisão do tempo de processamento

- Tempo de computação: 12.72s
- Tempo de MPI: 0.29s
- Tempo de OpenMP: 0.04s

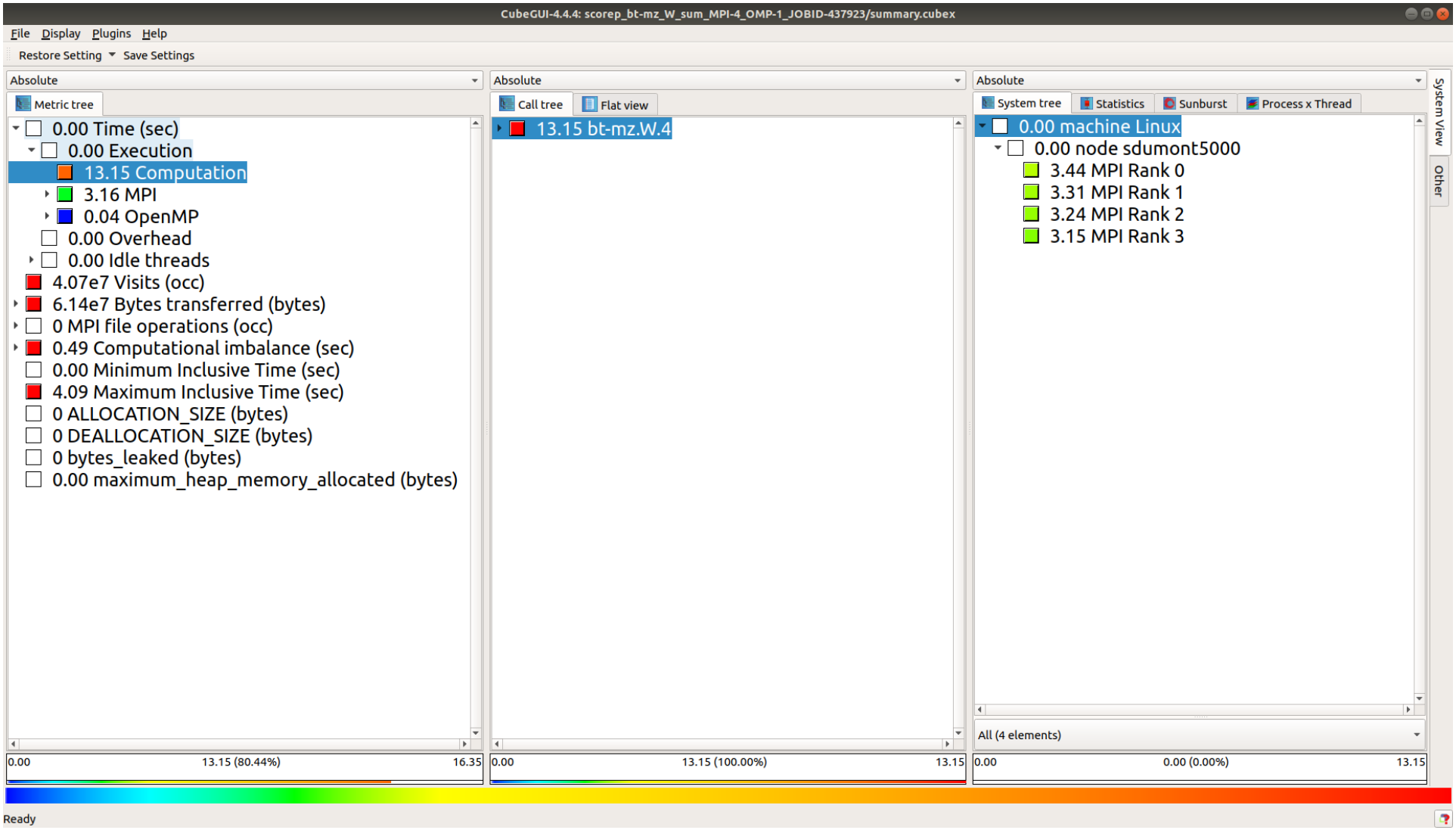
Visualizando: CubeGUI

-nodes=1 -ntasks=2



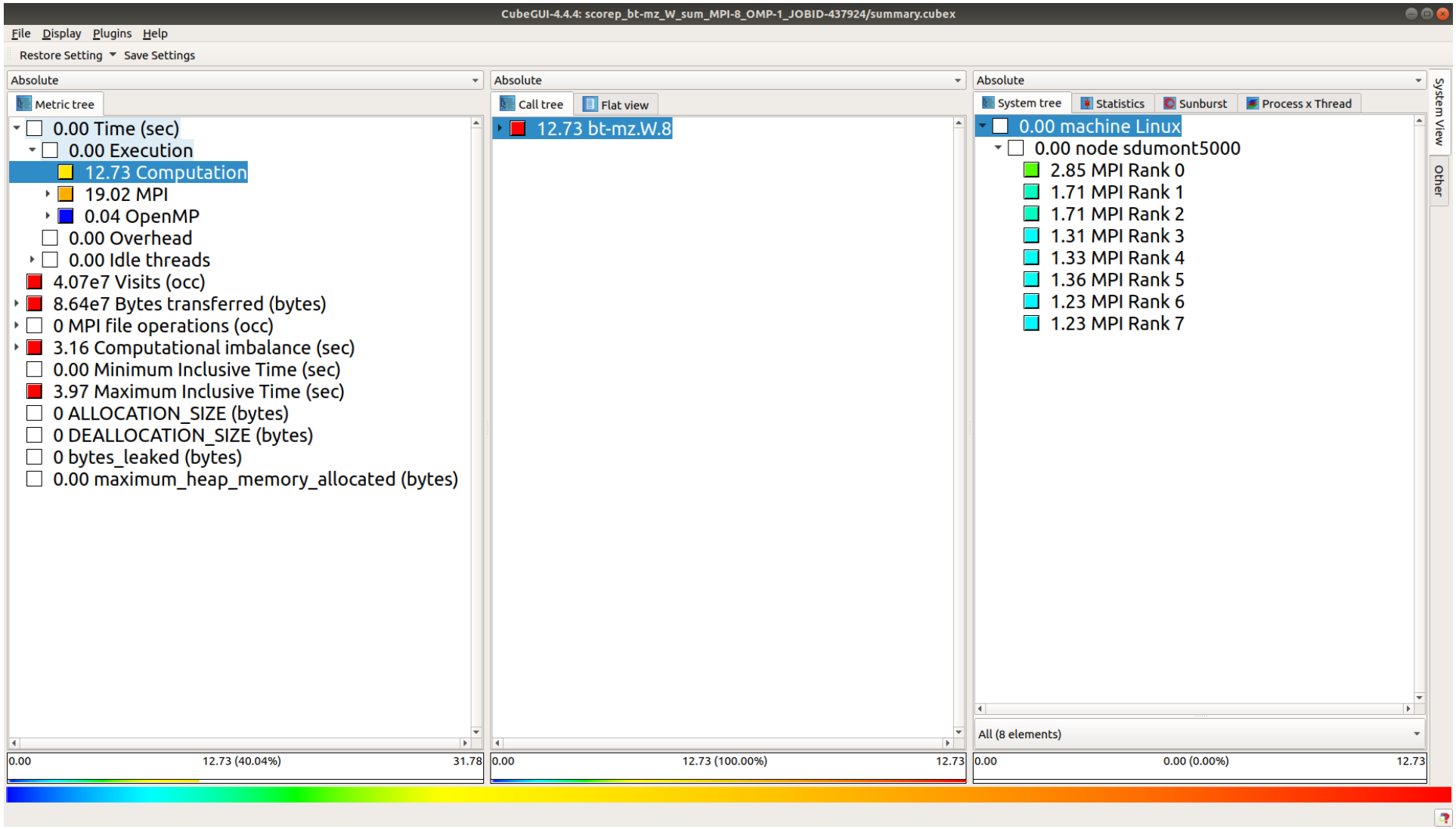
Visualizando: CubeGUI

-nodes=1 -ntasks=4



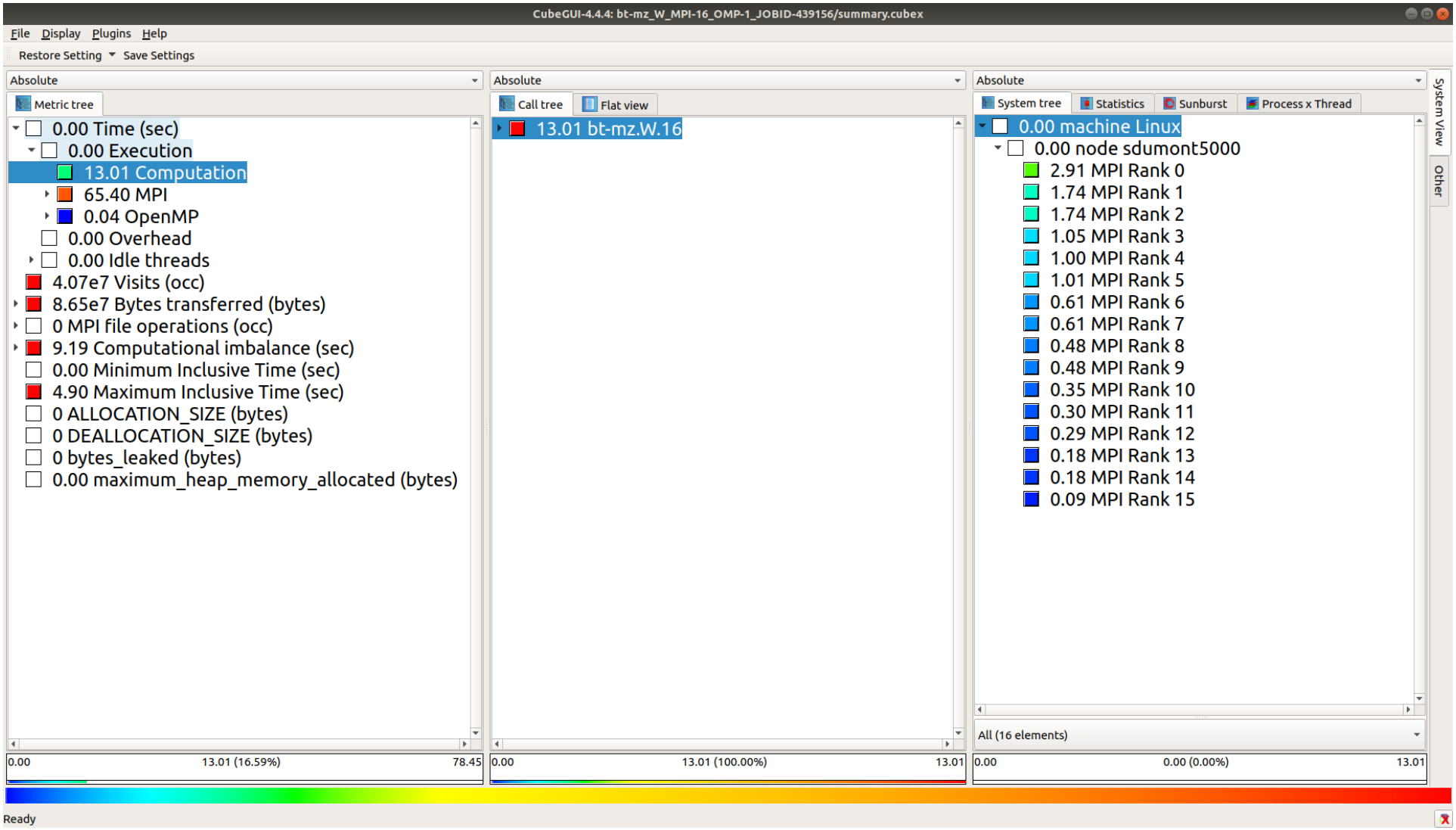
Visualizando: CubeGUI

-nodes=1 -ntasks=8



Visualizando: CubeGUI

-nodes=1 -ntasks=16



BT-MZ *benchmark*: divisão de domínio

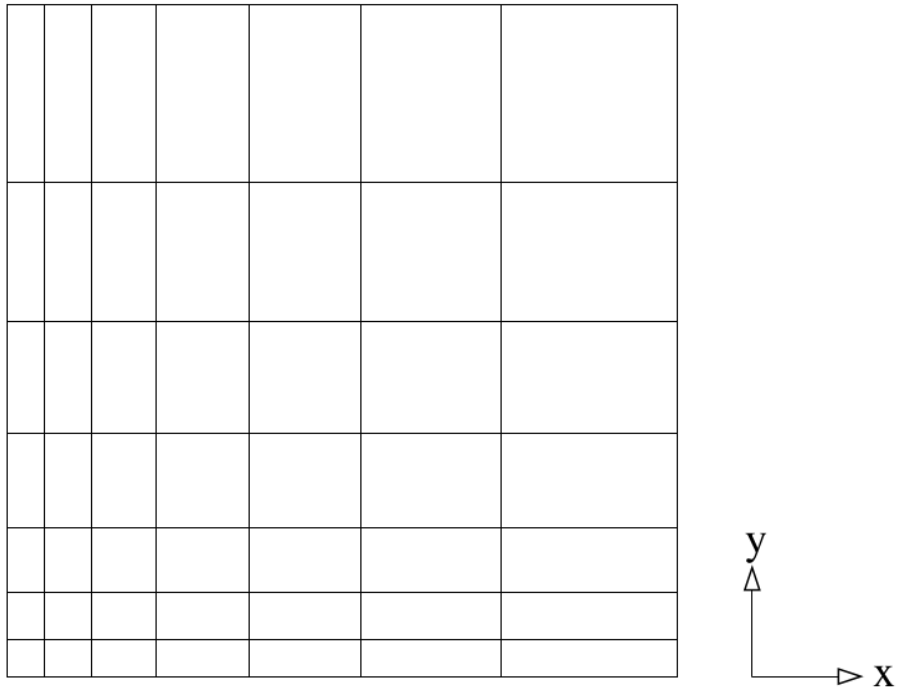


Figure 3: Example of uneven mesh tiling (horizontal cut through mesh system) for the BT-MZ benchmark.

Scalasca

Definição

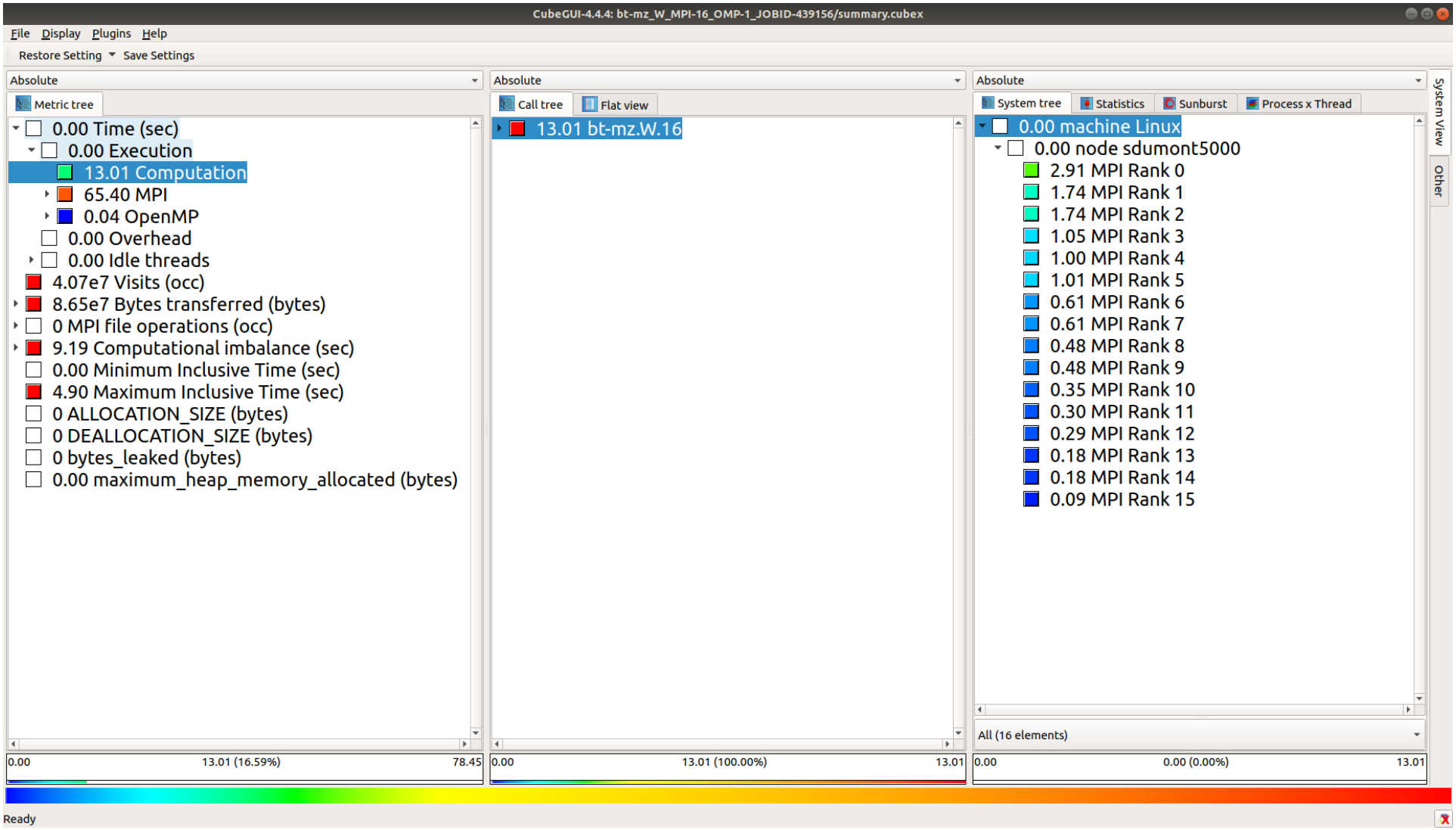
Balanço de carga de computação (LB):

$$LB = \frac{avg(tcomp)}{max(tcomp)}$$

FONTE: **`https://pop-coe.eu`**

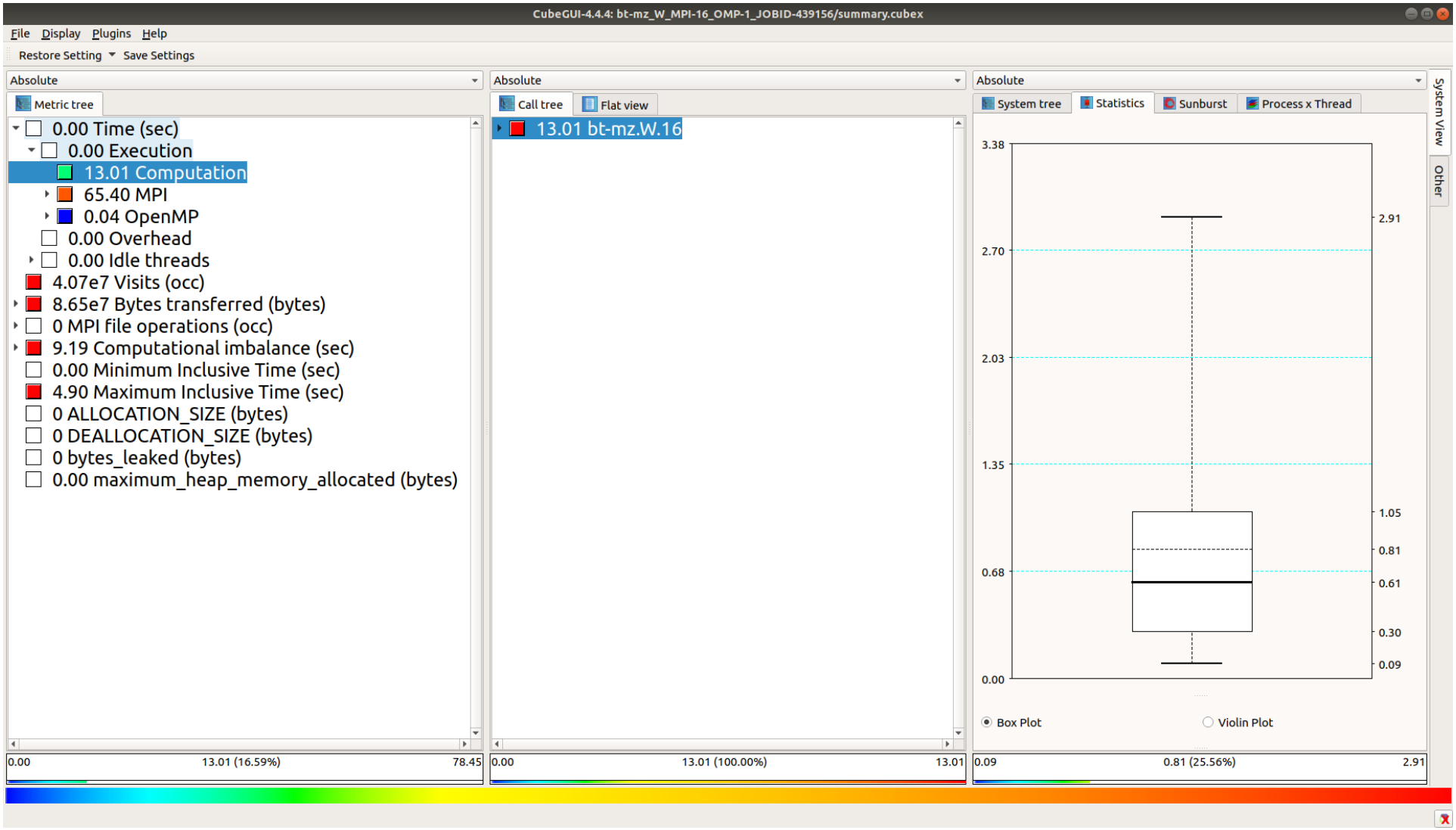
Visualizando: CubeGUI

-nodes=1 -ntasks=16



Visualizando: CubeGUI

-nodes=1 -ntasks=16



Scalasca

Cálculo

Balanço de carga de computação (LB)

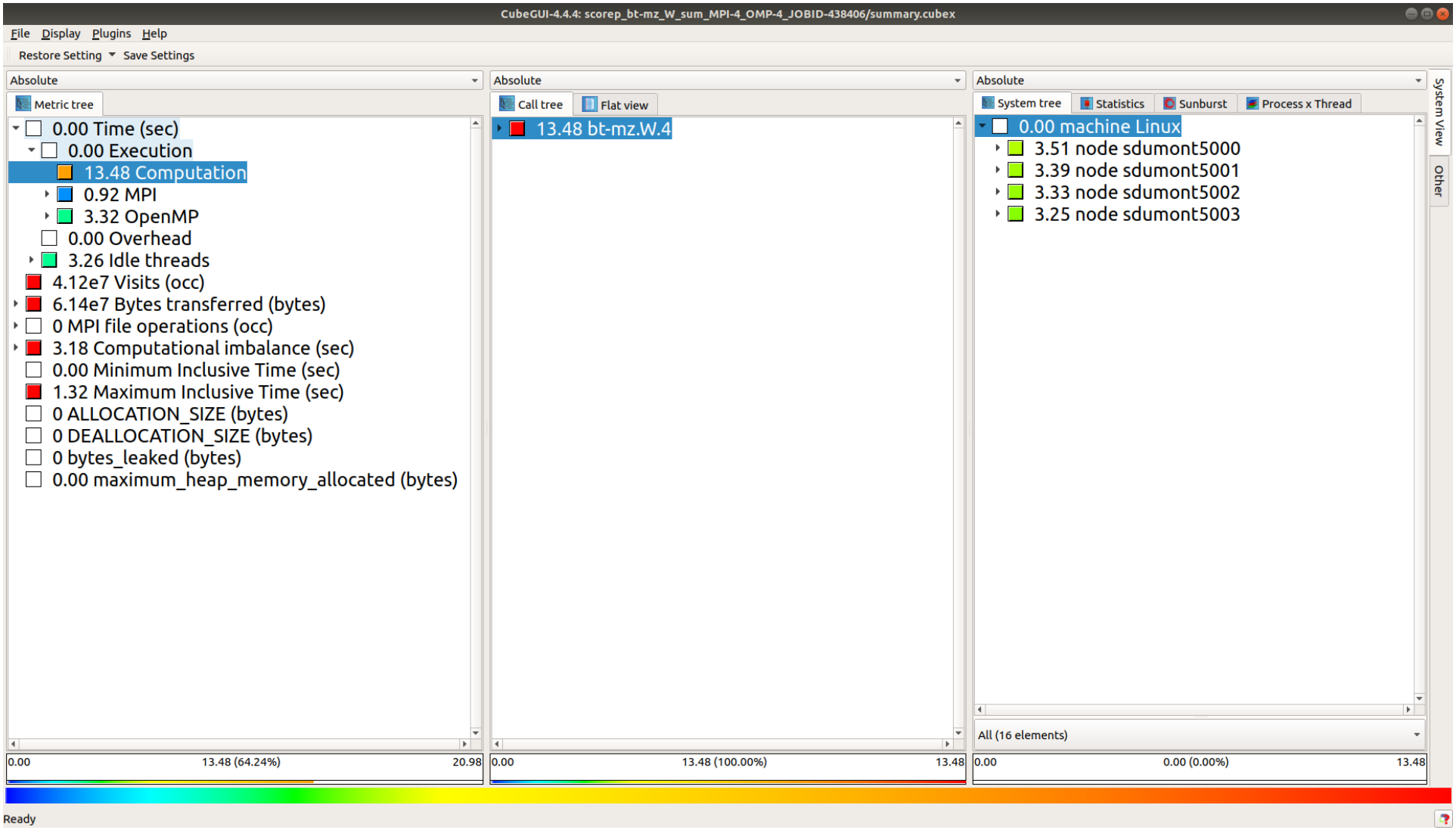
$$LB = \frac{avg(tcomp)}{max(tcomp)}$$

$$LB = \frac{0.81}{2.91}$$

$$LB = 0.28$$

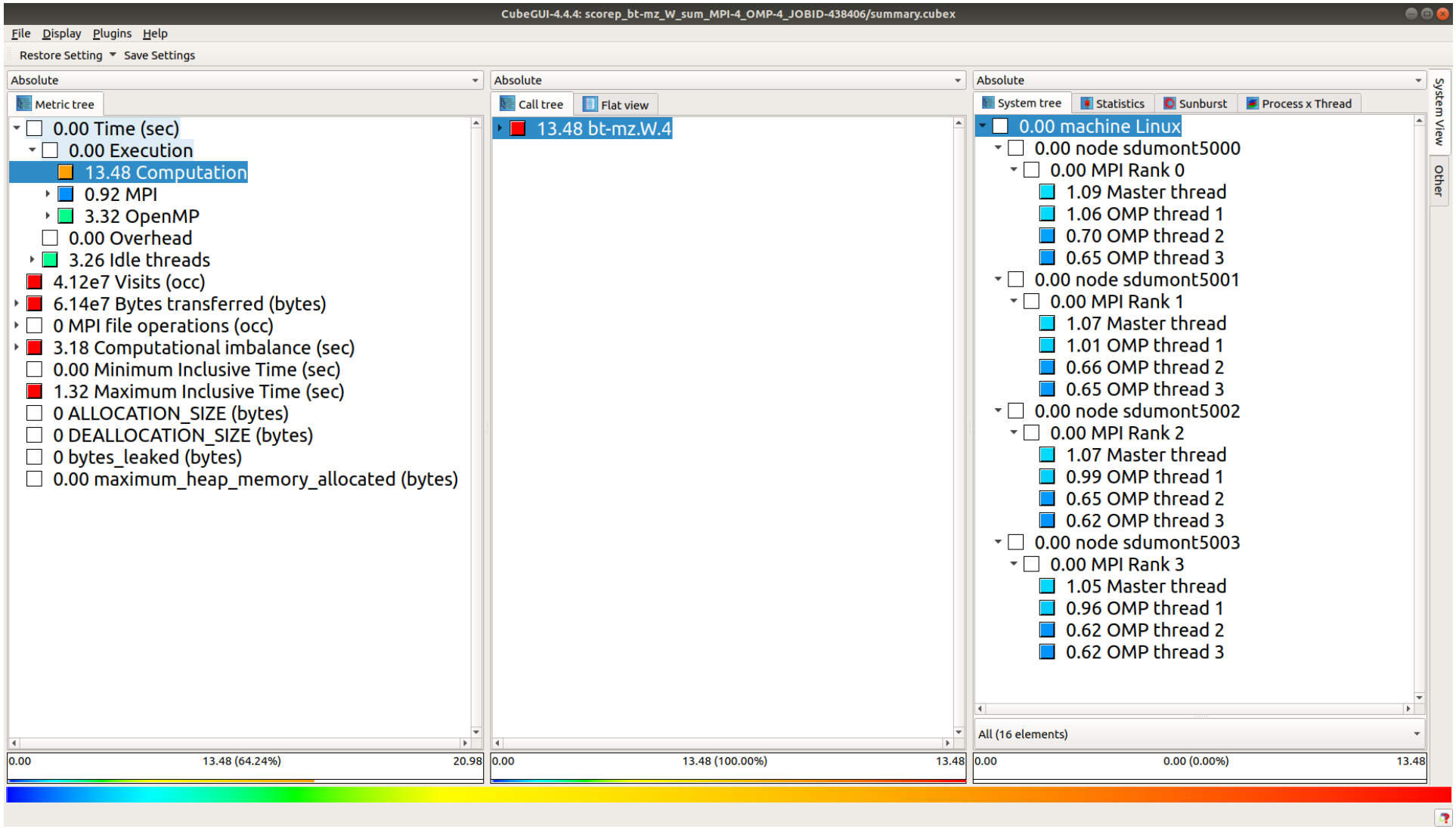
Visualizando: CubeGUI

-nodes=4 -ntasks=4 -cpus-per-task=4



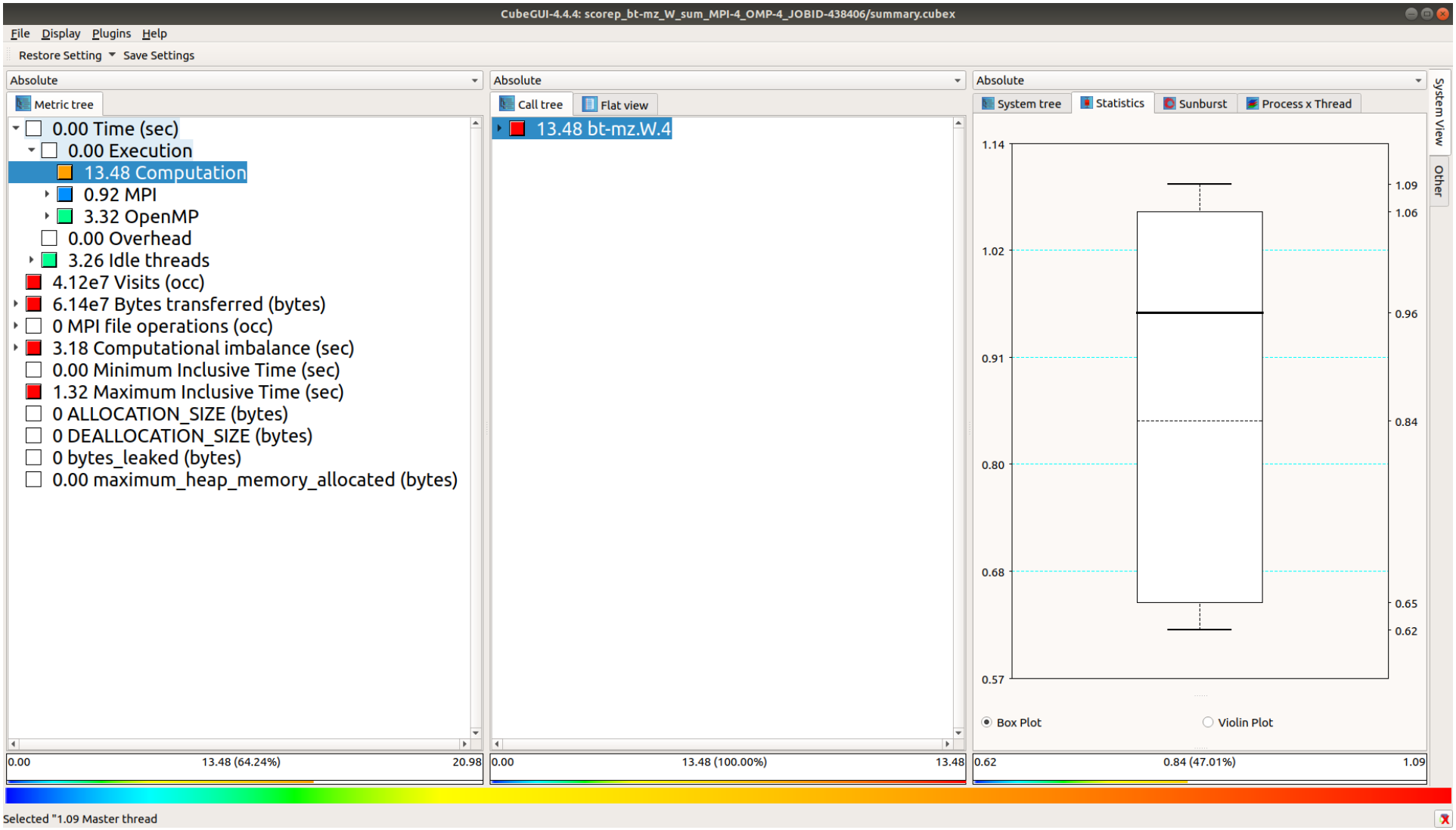
Visualizando: CubeGUI

-nodes=4 -ntasks=4 -cpus-per-task=4



Visualizando: CubeGUI

-nodes=4 -ntasks=4 -cpus-per-task=4



Scalasca

Cálculo

Balanço de carga de computação (LB)

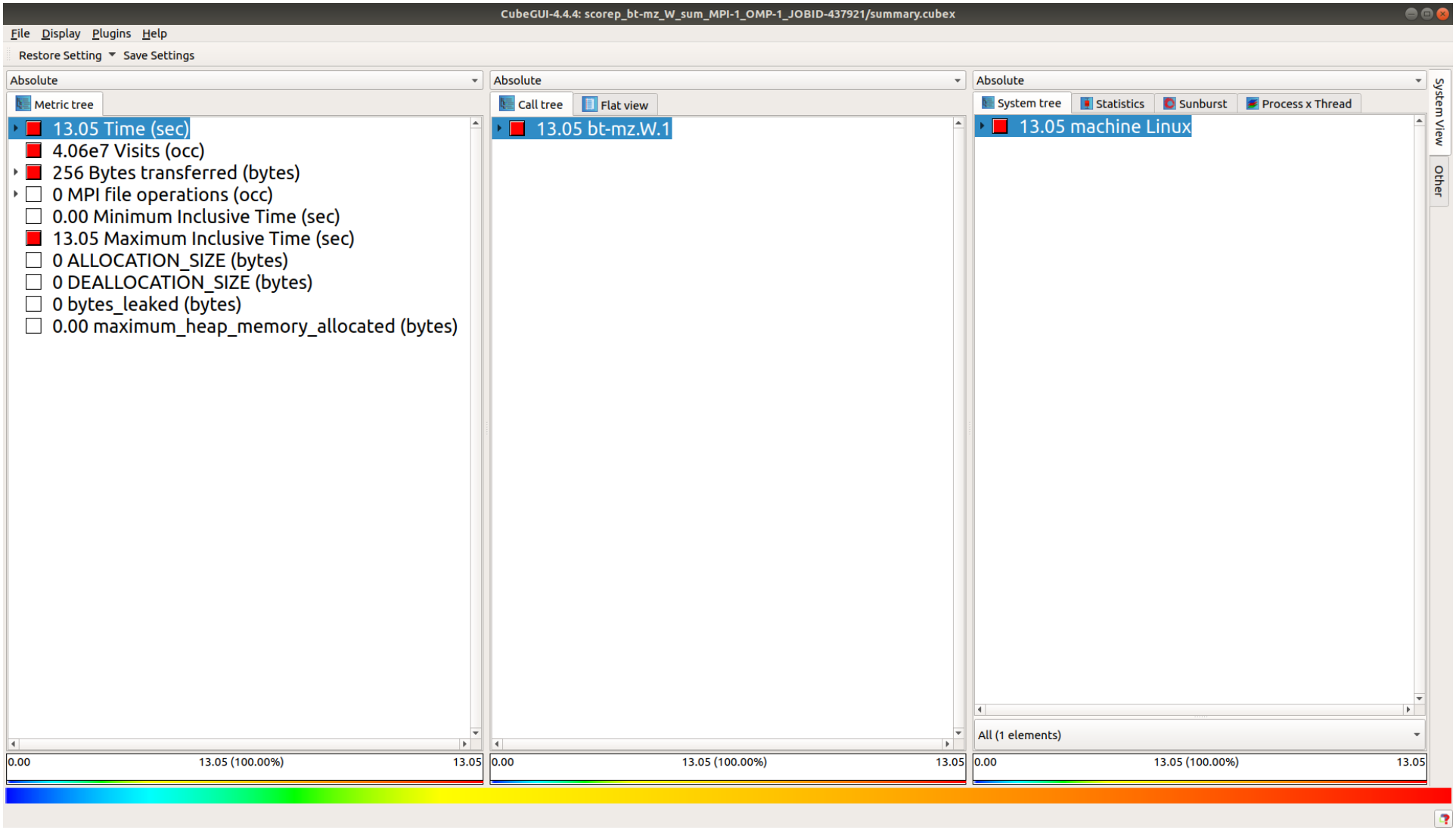
$$LB = \frac{avg(tcomp)}{max(tcomp)}$$

$$LB = \frac{0.84}{1.09}$$

$$LB = 0.77$$

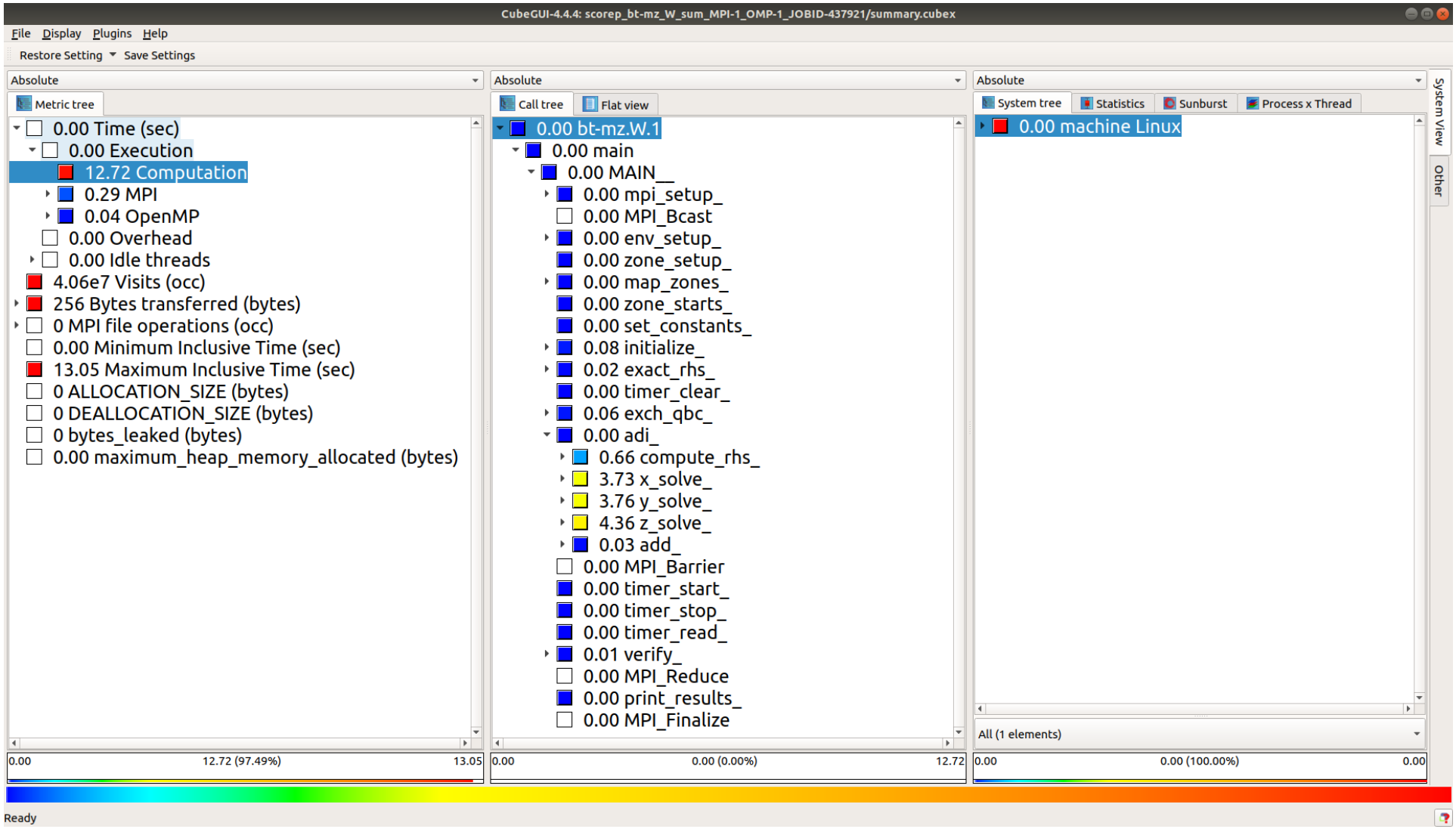
Visualizando: CubeGUI

`-nodes=1 -ntasks=1` / **Absolute**



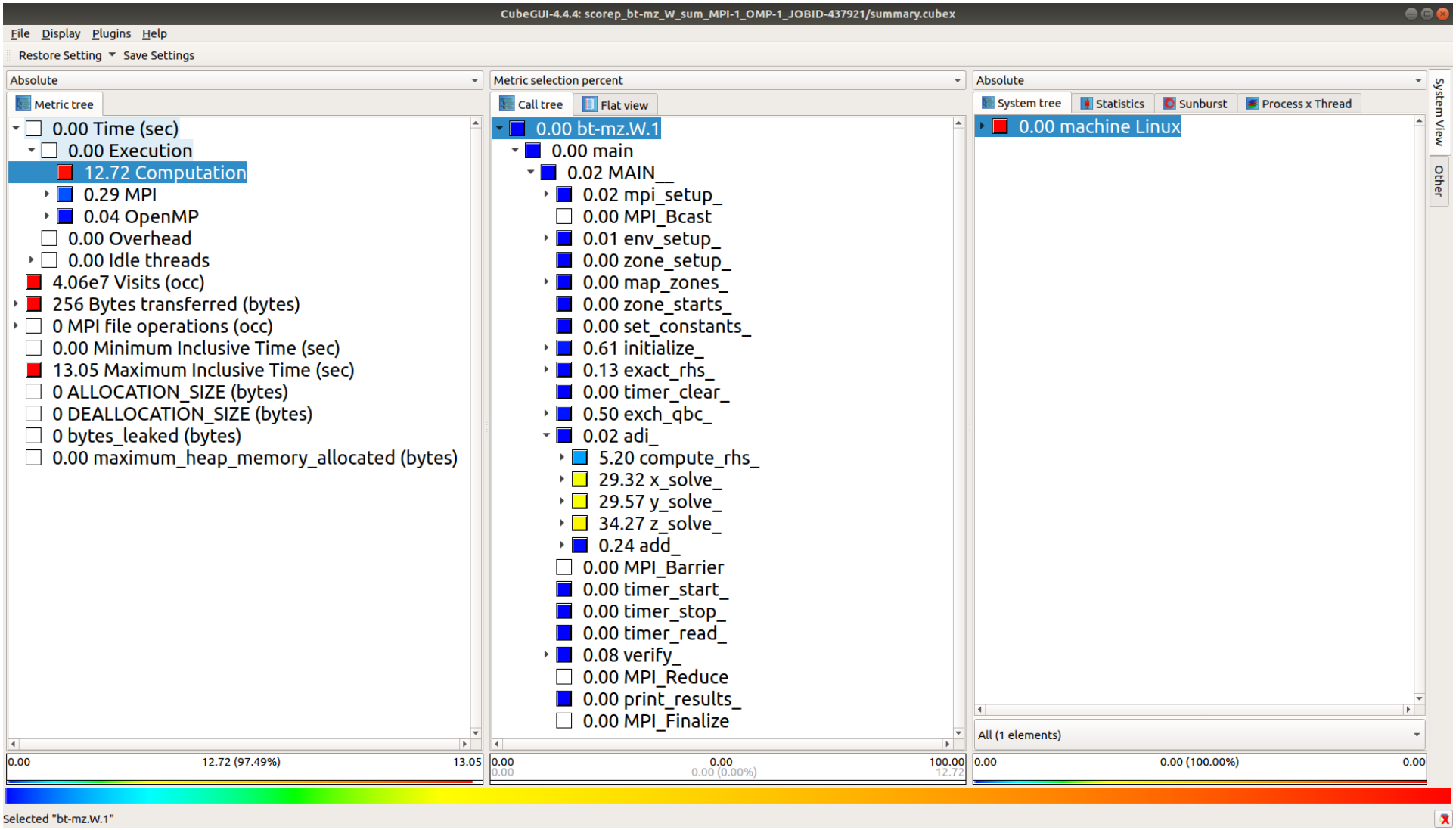
Visualizando: CubeGUI

-nodes=1 -ntasks=1 / Absolute



Visualizando: CubeGUI

-nodes=1 -ntasks=1 / Metric Own percent



Visualizando: CubeGUI

3 hotspots de computação

- **`x_solve`**: 3.73s (29.32%)
- **`y_solve`**: 3.76s (29.57%)
- **`z_solve`**: 4.36s (34.27%)

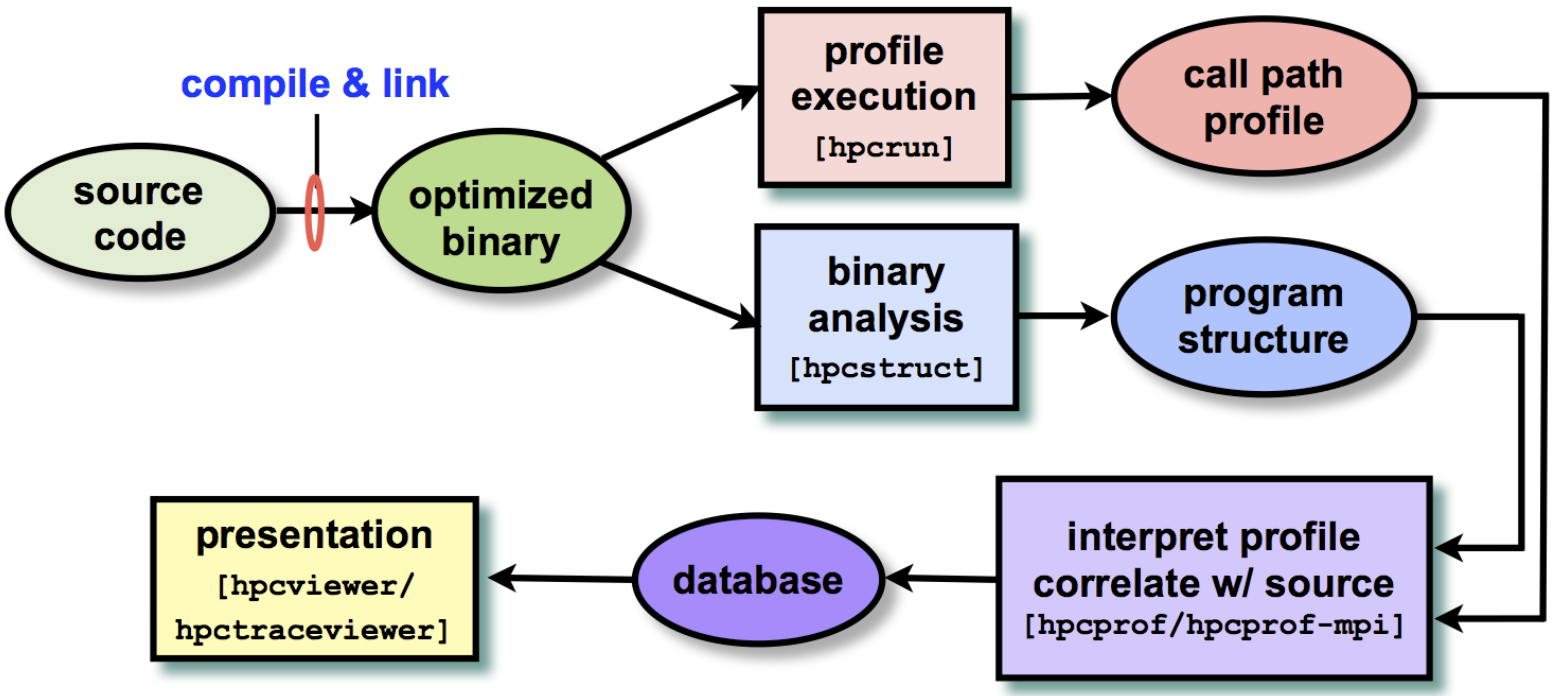
Roteiro

1. Scalasca, Score-p, Cube

2. hpctoolkit

HPCToolkit

http://hpctoolkit.org



HPCToolkit

Exemplo: NAS Parallel Benchmarks (NPB)

```
hpctoolkit/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_hpctoolkit
       README
```

HPCToolkit

Exemplo: NAS Parallel Benchmarks (NPB)

```
hpctoolkit/
    NPB3.3.1-MZ/
        NPB3.3-MZ-MPI/
        NPB3.3-MZ-OMP/
        NPB3.3-MZ-SER/
        Changes.log
        env_hpctoolkit
        README
```

HPCToolkit

Preparando o ambiente

```
$ cat env_hpctoolkit

module load openmpi/gnu/2.0.4.2
module load hpctoolkit/5.3.2_4712
module load papi/5.5.1.0
module load papi-devel/5.5.1.0
```

HPCToolkit

```
hpctoolkit/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_hpctoolkit
       README
```

HPCToolkit

```
hpctoolkit/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
          bin/
          BT-MZ/
          common/
          config/
          LU-MZ/
          SP-MZ/
          sys/
          Makefile
          README
          README.install
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_hpctoolkit
       README
```

HPCToolkit

```
hpctoolkit/
    NPB3.3.1-MZ/
       NPB3.3-MZ-MPI/
          bin/
          BT-MZ/
          common/
          config/
          LU-MZ/
          SP-MZ/
          sys/
          Makefile
          README
          README.install
       NPB3.3-MZ-OMP/
       NPB3.3-MZ-SER/
       Changes.log
       env_hpctoolkit
       README
```

HPCToolkit

```
config/
   NAS.samples
   make.def -> make_hpctoolkit.def
   make.def.template
   make_hpctoolkit.def
   suite.def
   suite.def.template
```

HPCToolkit

```
config/
    NAS.samples
    make.def -> make_hpctoolkit.def
    make.def.template
    make_hpctoolkit.def
    suite.def
    suite.def.template
```

HPCToolkit

```
$ cat make_hpctoolkit.def

#-------------------------------------------------------------------------
# This is the fortran compiler used for fortran programs
#-------------------------------------------------------------------------
F77 = mpif77
#F77 = scalasca -instrument mpif77
#F77 = scorep mpif77


#-------------------------------------------------------------------------
# This is the C compiler used for C programs
#-------------------------------------------------------------------------
CC = mpicc
#CC = scalasca -instrument mpicc
#CC = scorep mpicc
```

HPCToolkit

NPB: benchmark, classe e número de processos MPI

```
config/
   NAS.samples
   make.def -> make_hpctoolkit.def
   make.def.template
   make_hpctoolkit.def
   suite.def
   suite.def.template
```

HPCToolkit

NPB: benchmark, classe e número de processos MPI

```
config/
   NAS.samples
   make.def -> make_hpctoolkit.def
   make.def.template
   make_hpctoolkit.def
   suite.def
   suite.def.template
```

Estudo de caso

NPB: benchmark, classe e número de processos MPI

```
$ cat suite.def

# config/suite.def
# This file is used to build several benchmarks with a single command.
# Typing "make suite" in the main directory will build all the benchmarks
# specified in this file.
# Each line of this file contains a benchmark name, class, and number
# of nodes. The name is one of "sp-mz", "bt-mz", and "lu-mz".
# The class is one of "S", "W", and "A" through "F".
# No blank lines.
# The following example builds serial sample sizes of all benchmarks.
#sp-mz S 1
#lu-mz S 1
#bt-mz S 2
bt-mz   S       1
bt-mz   S       2
bt-mz   S       4
bt-mz   W       1
bt-mz   W       2
bt-mz   W       4
bt-mz   W       8
bt-mz   W       16
```

Estudo de caso

NPB: compilação

```
$ cd ..
$ make suite %compila o NPB
$ cd bin
```

Estudo de caso

NPB: compilação

```
$ ls -A1
bt-mz.S.1
bt-mz.S.2
bt-mz.S.4
bt-mz.W.1
bt-mz.W.2
bt-mz.W.4
bt-mz.W.8
bt-mz.W.16
BULL_srun_hpctoolkit.sh
```

HPCToolkit

BULL_srun_hpctoolkit.sh

```
#!/bin/bash

#SBATCH --nodes=1                          # here the number of nodes
#SBATCH --ntasks=1                         # here total number of mpi tasks
#SBATCH --cpus-per-task=1                  # number of cores per node
#SBATCH -p cpu_dev                         # target partition
#SBATCH --threads-per-core=1
#SBATCH -J NPB_BT-MZ                       # job name
#SBATCH --time=00:10:00                    # time limit
#SBATCH --exclusive                        # to have exclusive use of your nodes

echo "Cluster configuration:"
echo "==="
echo "Partition: " $SLURM_JOB_PARTITION
echo "Number of nodes: " $SLURM_NNODES
echo "Number of MPI processes: " $SLURM_NTASKS " (" $SLURM_NNODES " nodes)"
echo "Number of MPI processes per node: " $SLURM_NTASKS_PER_NODE
echo "Number of threads per MPI process: " $SLURM_CPUS_PER_TASK
echo "NPB Benchmark: " $1
echo "Bechmark class problem: " $2

#####################################
#            COMPILER               #
#####################################
module load openmpi/gnu/2.0.4.2
module load hpctoolkit/5.3.2_4712
module load papi/5.5.1.0
module load papi-devel/5.5.1.0
```

HPCToolkit (cont.)

```
bench=${1}
class=${2}
executable="${bench}.${class}.$SLURM_NTASKS"

export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK

srun --resv-ports -n $SLURM_NTASKS \
hpcrun -t -e WALLCLOCK@5000 \
./${executable}

hpcstruct ${executable}

hpcprof \
 -I ../BT-MZ/+ \
 -S ${executable}.hpcstruct hpctoolkit-${executable}-measurements-${SLURM_JOBID}

hpctoolkitresultdir=profiling/hpctoolkit/NUMNODES-$SLURM_JOB_NUM_NODES/${bench}_${class}_
                    MPI-${SLURM_NTASKS}_OMP-${SLURM_CPUS_PER_TASK}_JOBID-${SLURM_JOBID}

mkdir -p ${hpctoolkitresultdir}

mv slurm-${SLURM_JOBID}.out ${hpctoolkitresultdir}/
mv hpctoolkit-${executable}-database-${SLURM_JOBID} ${hpctoolkitresultdir}/
mv hpctoolkit-${executable}-measurements-${SLURM_JOBID} ${hpctoolkitresultdir}/
mv ${executable}.hpcstruct ${hpctoolkitresultdir}/
```

Estudo de caso

NPB: submetendo job

```
$ sbatch BULL_srun_hpctoolkit.sh bt-mz W
Submitted batch job 438988
$ squeue -u $USER
JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
438988 treinamen NPB_BT-M professo  R       0:02      1 sdumont3000
```

Estudo de caso

NPB: perfil de desempenho

```
$ ls -A1
bt-mz.S.1
bt-mz.S.2
bt-mz.S.4
bt-mz.W.1
bt-mz.W.2
bt-mz.W.4
bt-mz.W.8
bt-mz.W.16
BULL_srun_hpctoolkit.sh
hpctoolkit/
```

Estudo de caso

NPB: perfil de desempenho

```
hpctoolkit/
 NUMNODES-1
     bt-mz_W_MPI-1_OMP-1_JOBID-438988
        bt-mz.W.1.hpcstruct
        hpctoolkit-bt-mz.W.1-database-438988
        hpctoolkit-bt-mz.W.1-measurements-438988
        slurm-438988.out
```

Estudo de caso

NPB: perfil de desempenho

```
hpctoolkit/
 NUMNODES-1
     bt-mz_W_MPI-1_OMP-1_JOBID-438988
        bt-mz.W.1.hpcstruct
        hpctoolkit-bt-mz.W.1-database-438988
        hpctoolkit-bt-mz.W.1-measurements-438988
        slurm-438988.out
```

Estudo de caso

NPB: perfil de desempenho

```
hpctoolkit/
 NUMNODES-1
     bt-mz_W_MPI-1_OMP-1_JOBID-438988
        bt-mz.W.1.hpcstruct
        hpctoolkit-bt-mz.W.1-database-438988
        hpctoolkit-bt-mz.W.1-measurements-438988
        slurm-438988.out
```

Estudo de caso

```
$ cat slurm-438988.out

Cluster configuration:
===
Partition:  treinamento
Number of nodes:  1
Number of MPI processes:  1  ( 1  nodes)
Number of MPI processes per node:
Number of threads per MPI process:  1
NPB Benchmark:  bt-mz
Bechmark class problem:  W
[1580508594.340437] [sdumont5000:81015:0]          mxm.c:196  MXM  WARN  The 'ulimit -s' on the sys
[1580508594.342077] [sdumont5000:81015:0]          mxm.c:196  MXM  WARN  The 'ulimit -s' on the sys


 NAS Parallel Benchmarks (NPB3.3-MZ-MPI) - BT-MZ MPI+OpenMP Benchmark

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     1

 Use the default load factors with threads
 Total number of threads:      1  (  1.0 threads/process)

 Calculated speedup =      1.00

 Time step    1
 Time step   20
```

Estudo de caso (cont.)

```
Time step   40
Time step   60
Time step   80
Time step  100
Time step  120
Time step  140
Time step  160
Time step  180
Time step  200
Verification being performed for class W
accuracy setting for epsilon =  0.1000000000000E-07
Comparison of RMS-norms of residual
          1 0.5562611195402E+05 0.5562611195402E+05 0.2289019558898E-13
          2 0.5151404119932E+04 0.5151404119932E+04 0.3195605260010E-13
          3 0.1080453907954E+05 0.1080453907954E+05 0.4314917838667E-12
          4 0.6576058591929E+04 0.6576058591929E+04 0.2033067669511E-13
          5 0.4528609293561E+05 0.4528609293561E+05 0.3100863263992E-13
Comparison of RMS-norms of solution error
          1 0.7185154786403E+04 0.7185154786403E+04 0.4961924046085E-13
          2 0.7040472738068E+03 0.7040472738068E+03 0.3326408529931E-13
          3 0.1437035074443E+04 0.1437035074443E+04 0.1887614294376E-12
          4 0.8570666307849E+03 0.8570666307849E+03 0.3143720636440E-13
          5 0.5991235147368E+04 0.5991235147368E+04 0.6770467641700E-13
Verification Successful


BT-MZ Benchmark Completed.
Class           =                        W
Size            =             64x   64x  8
Iterations      =                      200
```

Estudo de caso (cont.)

```
 Time in seconds =                    5.58
 Total processes =                       1
 Total threads   =                       1
 Mop/s total     =                 2572.99
 Mop/s/thread    =                 2572.99
 Operation type  =           floating point
 Verification    =               SUCCESSFUL
 Version         =                    3.3.1
 Compile date    =              31 Jan 2020

 Compile options:
    F77           = mpif77
    FLINK         = $(F77)
    F_LIB         = (none)
    FFLAGS        = -O3 -fopenmp -g
    FLINKFLAGS    = $(FFLAGS)
    RAND          = (none)


 Please send all errors/feedbacks to:

 NPB Development Team
 npb@nas.nasa.gov


msg: STRUCTURE: /scratch/treinamento/professor/modulo1/MC1-I/tools/hpctoolkit/NPB3.3.1-MZ/NPB3.3-M
msg: Line map : /opt/bullxde/profilers/hpctoolkit/5.3.2_4712/lib/hpctoolkit/ext-libs/libmonitor.so
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/libmpi_mpifh.so.20.2.1
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/libmpi.so.20.0.4
msg: Line map : /usr/lib64/libgomp.so.1.0.0
```

Estudo de caso (cont.)

```
msg: Line map : /usr/lib64/libc-2.17.so
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/libopen-rte.so.20.1.2
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/libopen-pal.so.20.2.2
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_ess_pmi.so
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_pml_cm.so
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_pml_yalla.so
msg: Line map : /opt/mellanox/mxm/lib/libmxm.so.2.0.32
msg: Line map : /opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_mtl_mxm.so
msg: Populating Experiment database: /scratch/treinamento/professor/modulo1/MC1-I/tools/hpctoolkit
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=1 BULL_srun_hpctoolkit.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     1

 Use the default load factors with threads
 Total number of threads:      1   (  1.0 threads/process)

 Calculated speedup =       1.00

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                     5.58
 Total processes =                        1
 Total threads   =                        1
 Mop/s total     =                  2572.99
 Mop/s/thread    =                  2572.99
 Operation type  =            floating point
 Verification    =                SUCCESSFUL
 Version         =                     3.3.1
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=2 BULL_srun_hpctoolkit.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     2

 Use the default load factors with threads
 Total number of threads:      2   (  1.0 threads/process)

 Calculated speedup =       1.98

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                     2.84
 Total processes =                        2
 Total threads   =                        2
 Mop/s total     =                  5056.23
 Mop/s/thread    =                  2528.12
 Operation type  =            floating point
 Verification    =                SUCCESSFUL
 Version         =                     3.3.1
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=4 BULL_srun_hpctoolkit.sh bt-mz W

 Number of zones:    4 x    4
 Iterations: 200    dt:    0.000800
 Number of active processes:     4

 Use the default load factors with threads
 Total number of threads:      4   (  1.0 threads/process)

 Calculated speedup =      3.95

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =              64x   64x  8
 Iterations      =                      200
 Time in seconds =                     1.49
 Total processes =                        4
 Total threads   =                        4
 Mop/s total     =                  9610.43
 Mop/s/thread    =                  2402.61
 Operation type  =            floating point
 Verification    =                SUCCESSFUL
 Version         =                     3.3.1
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=8 BULL_srun_hpctoolkit.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:     8

 Use the default load factors with threads
 Total number of threads:      8   (  1.0 threads/process)

 Calculated speedup =       4.87

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                     1.29
 Total processes =                        8
 Total threads   =                        8
 Mop/s total     =                 11125.59
 Mop/s/thread    =                  1390.70
 Operation type  =            floating point
 Verification    =                SUCCESSFUL
 Version         =                     3.3.1
```

Estudo de caso

```
sbatch --nodes=1 --ntasks=16 BULL_srun_hpctoolkit.sh bt-mz W

 Number of zones:   4 x   4
 Iterations: 200    dt:   0.000800
 Number of active processes:    16

 Use the default load factors with threads
 Total number of threads:     16   (  1.0 threads/process)

 Calculated speedup =       4.87

 BT-MZ Benchmark Completed.
 Class           =                        W
 Size            =             64x   64x  8
 Iterations      =                      200
 Time in seconds =                     1.28
 Total processes =                       16
 Total threads   =                       16
 Mop/s total     =                 11181.49
 Mop/s/thread    =                   698.84
 Operation type  =           floating point
 Verification    =               SUCCESSFUL
 Version         =                    3.3.1
```

Visualizando no **hpcviewer**

- O **hpcviewer** pode ser baixado e instalado para visualizar os resultados obtidos com o HPCToolkit
- `http://hpctoolkit.org/download.html`
- Binários prontos em "Binary Releases of HPCToolkit's hpcviewer Graphical User Interface"
- Resultados previamente obtidos no SDumont estão no arquivo **profiling_hpctoolkit_sdbase.zip** do repositório no GitHub:

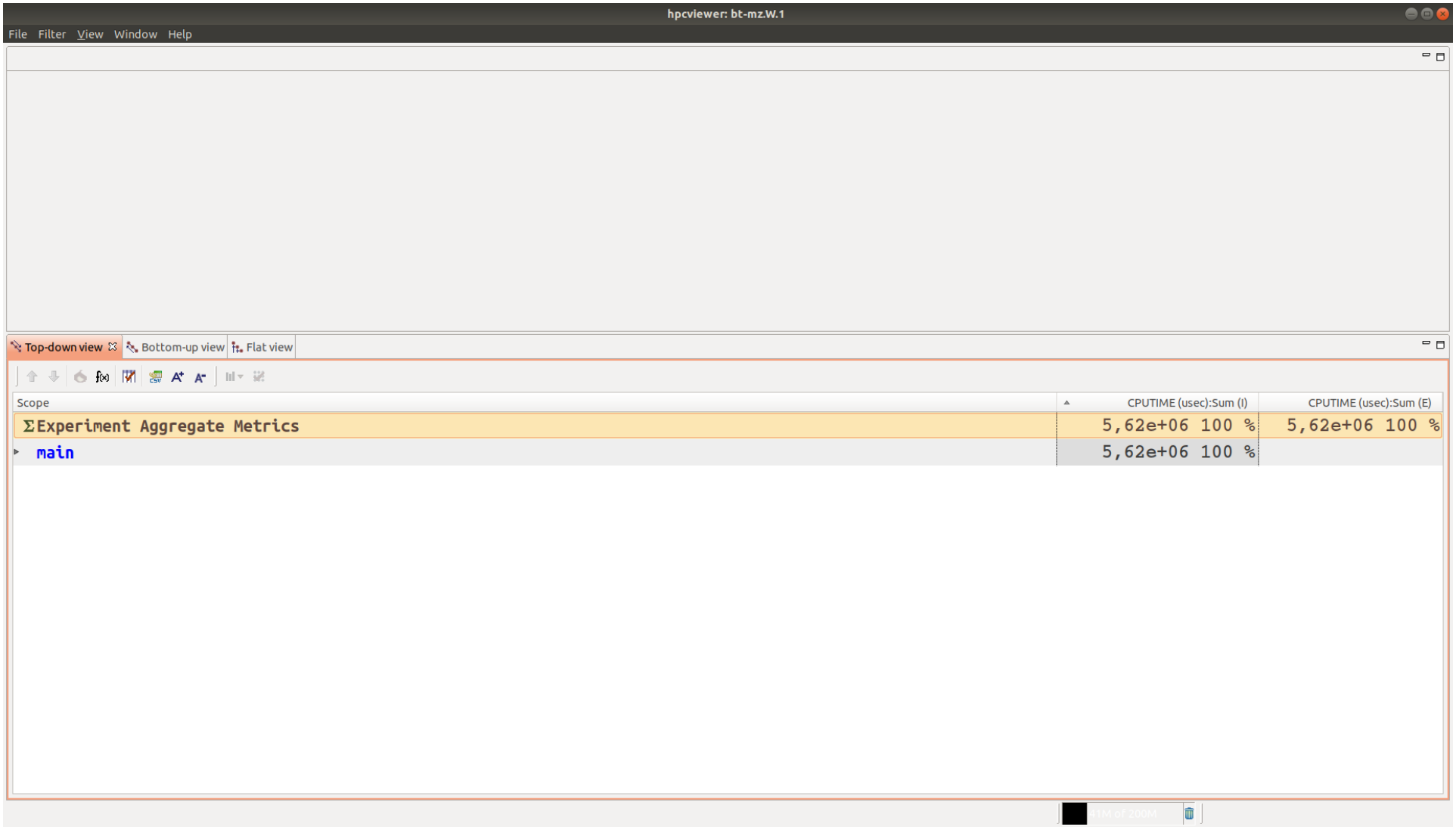
```
git clone https://github.com/robertopsouto/ESD2023.git
ESD2023/MC-SD01-I/hpctoolkit/profiling_hpctoolkit_sdbase.zip
```

NPB: estudo de caso

```
$ cd profiling/hpctoolkit/NUMNODES-1/bt-mz_W_MPI-1_OMP-1_JOBID-438988
$ hpcviewer hpctoolkit-bt-mz.W.1-database-438988/
```

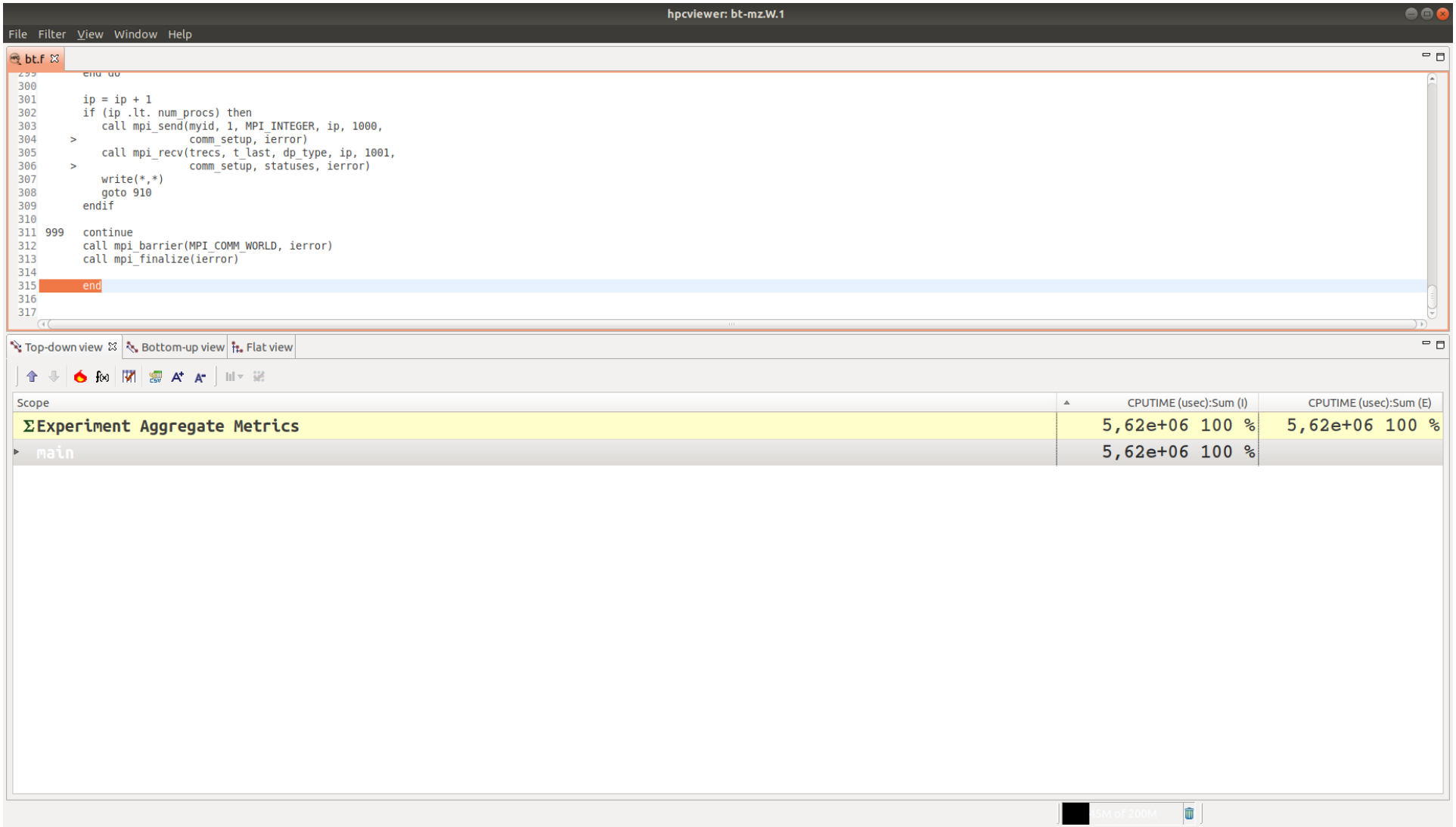
Visualizando no `hpcviewer`

`-nodes=1 -ntasks=1 / Top-down view`



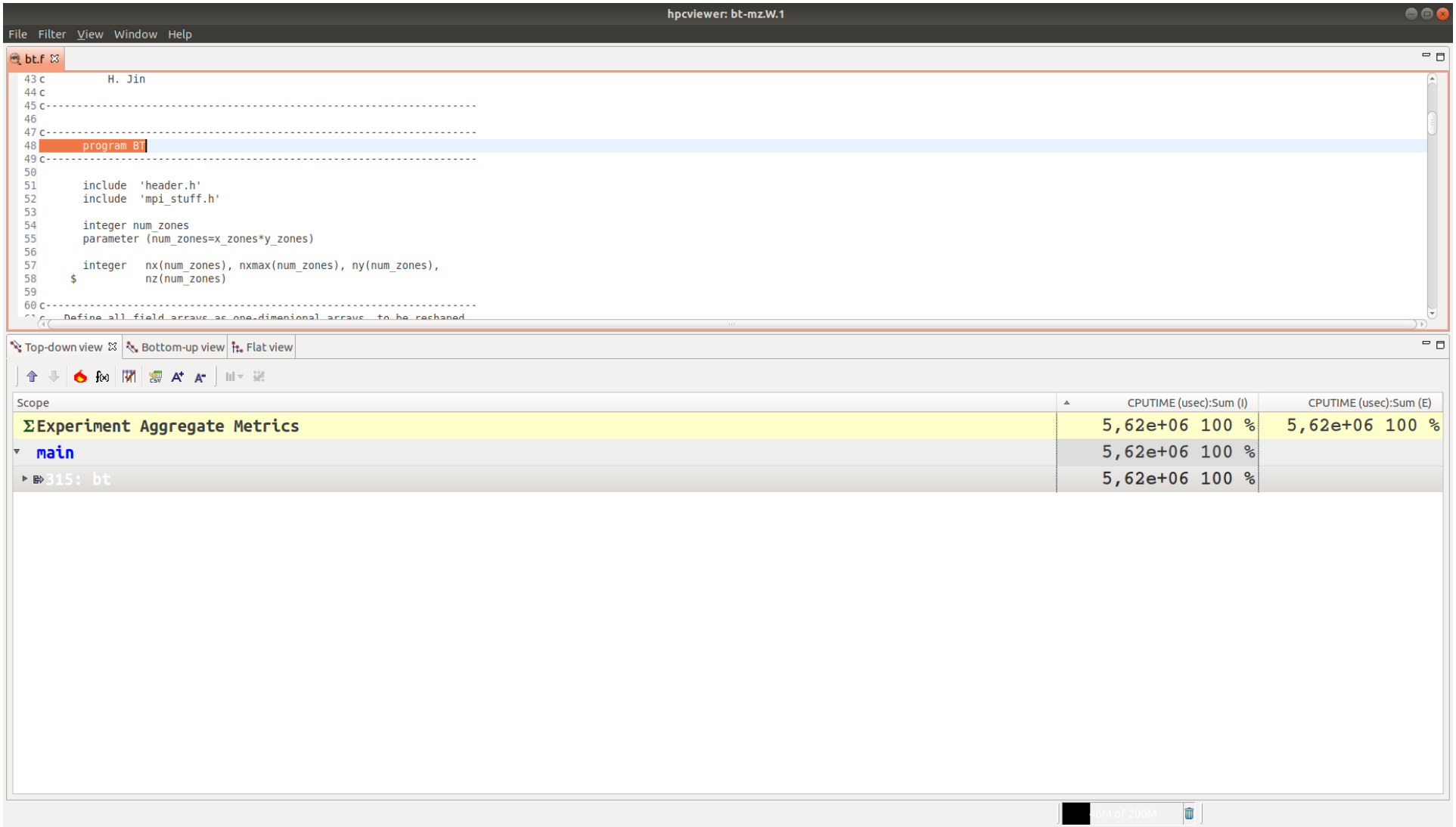
Visualizando no `hpcviewer`

`-nodes=1 -ntasks=1 / Top-down view`



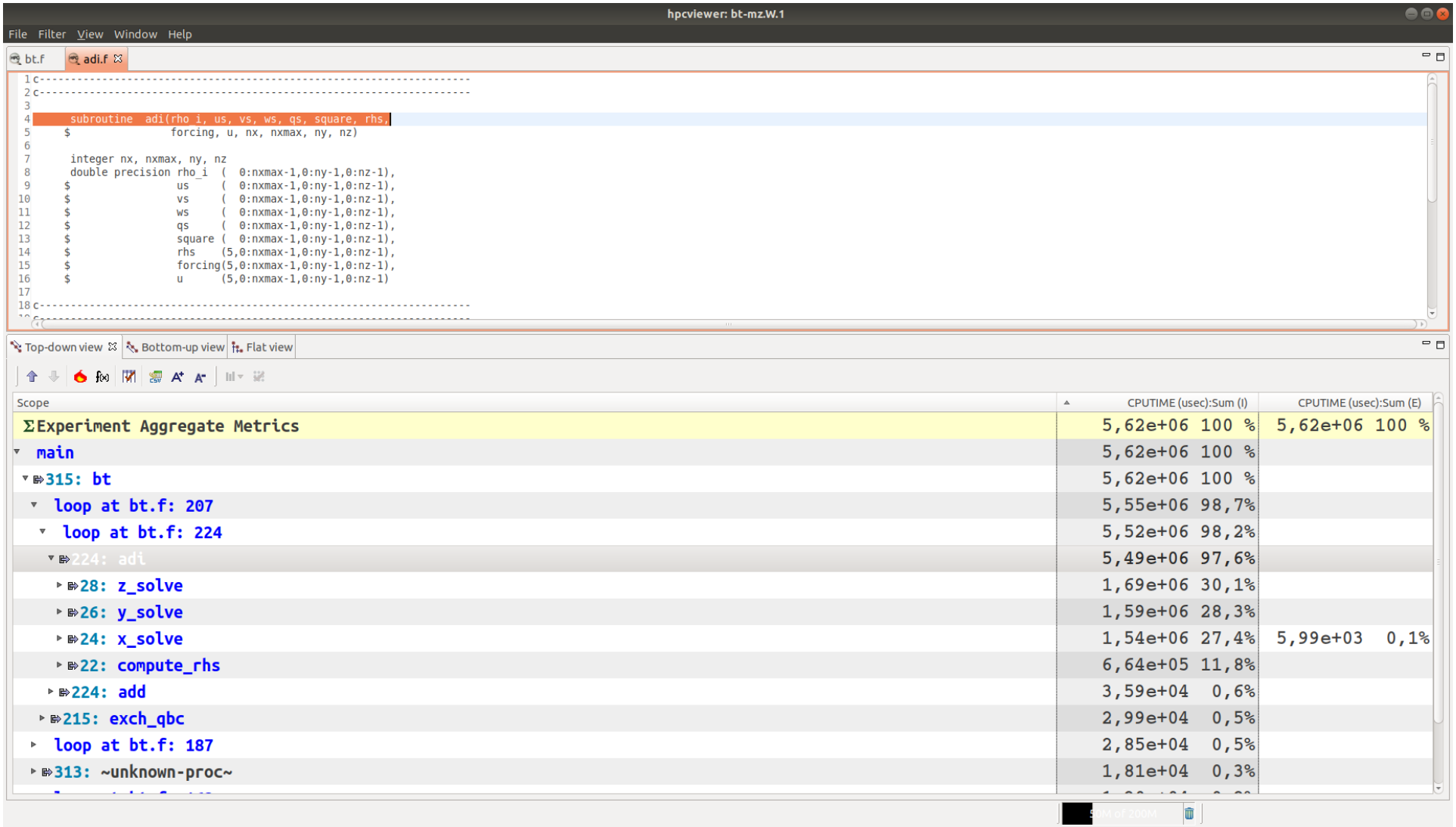
Visualizando no **hpcviewer**

-nodes=1 -ntasks=1 / Top-down view



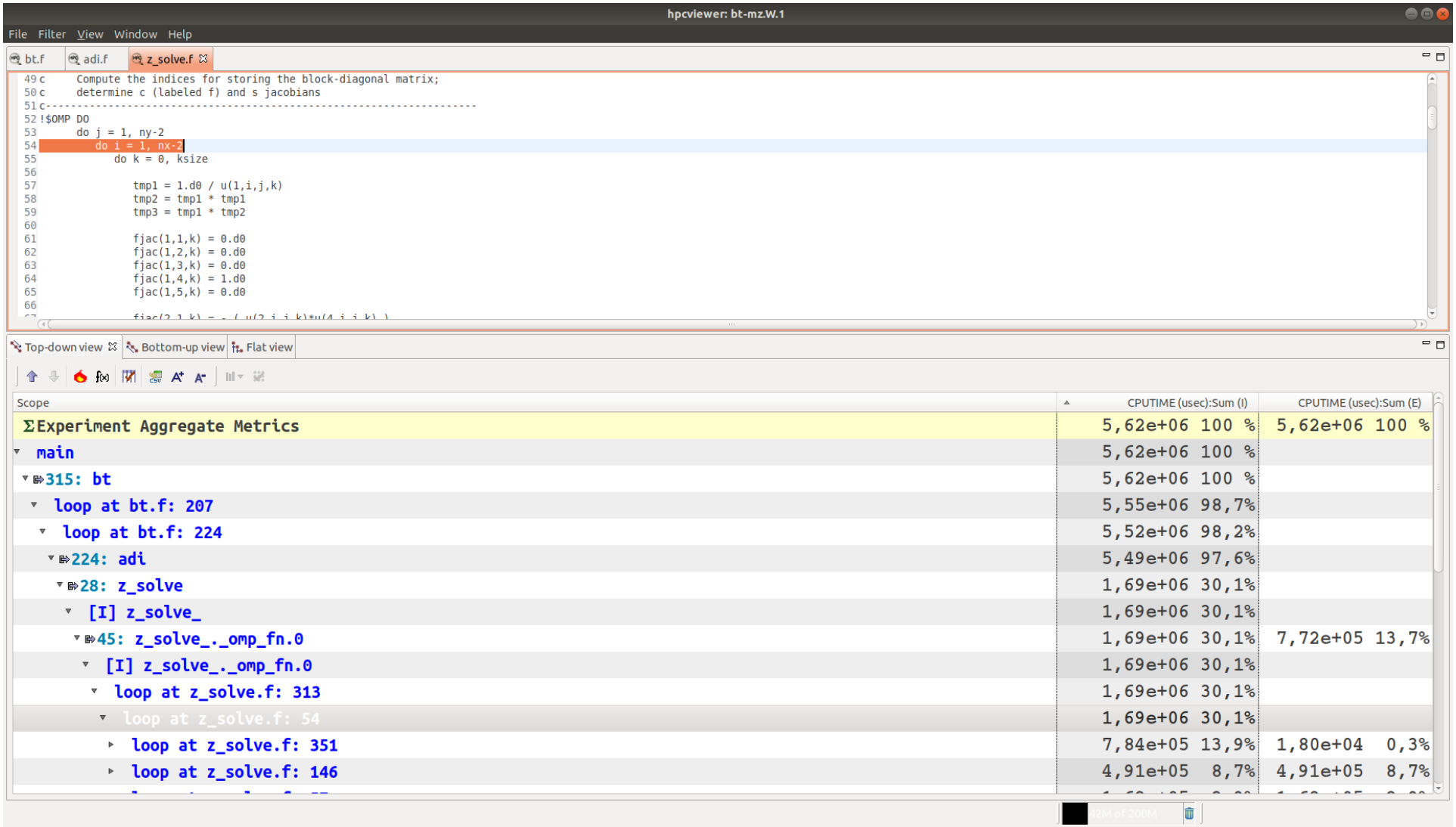
Visualizando no **hpcviewer**

-nodes=1 -ntasks=1 / Top-down view



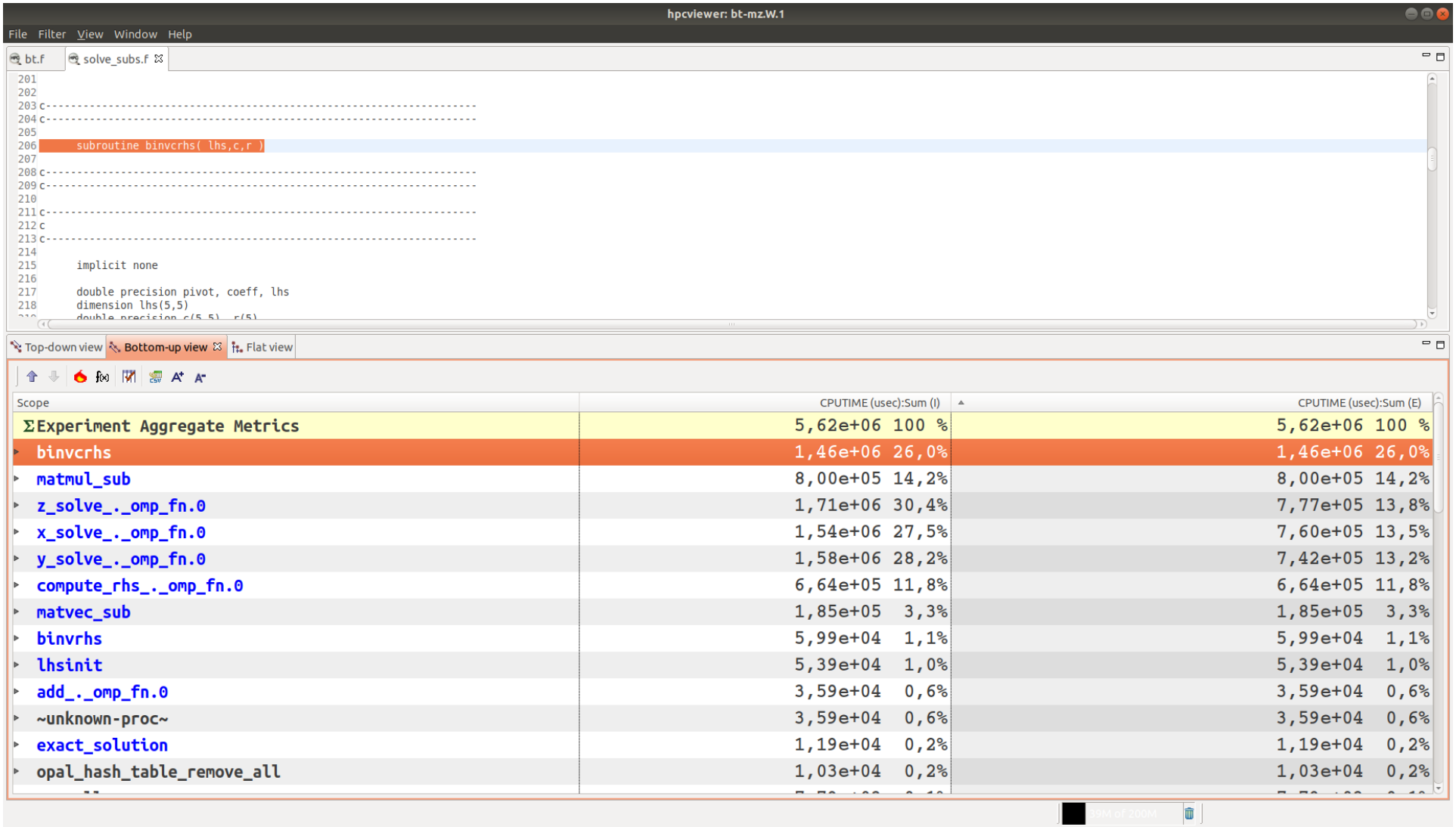
Visualizando no `hpcviewer`

`-nodes=1 -ntasks=1 / Top-down view`



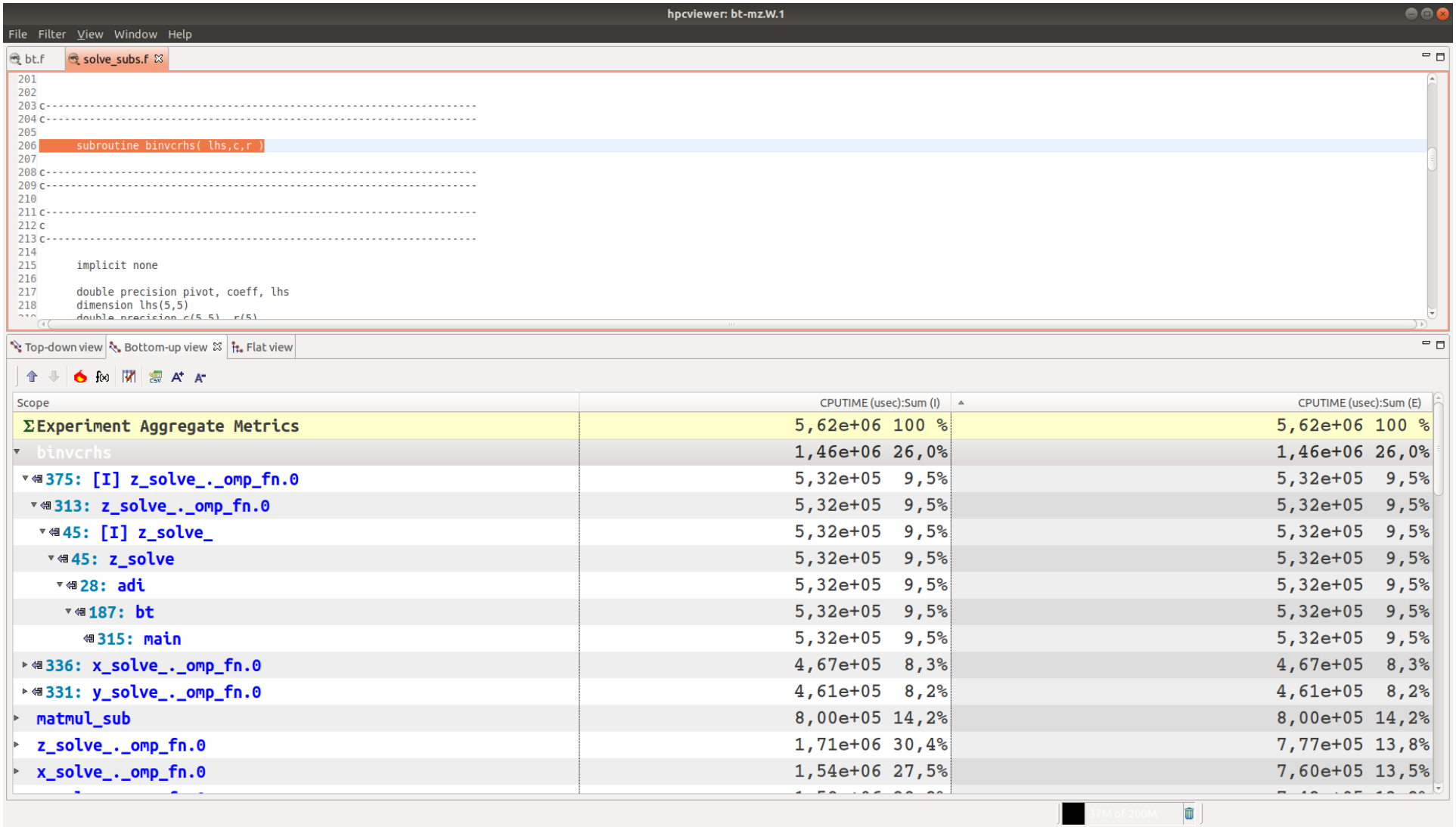
Visualizando no **hpcviewer**

-nodes=1 -ntasks=1 / Bottom-up view



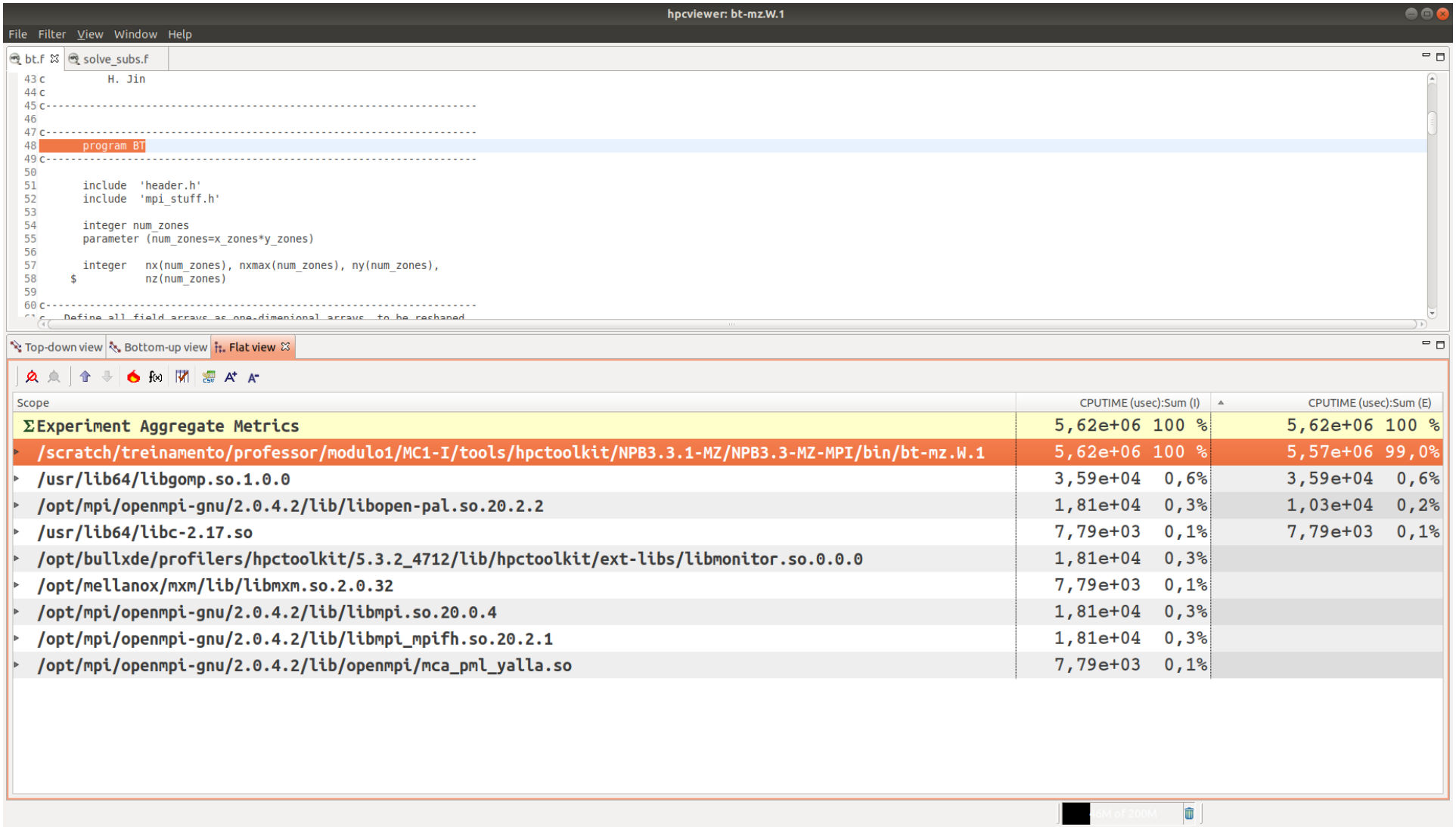
Visualizando no `hpcviewer`

`-nodes=1 -ntasks=1 / Bottom-up view`



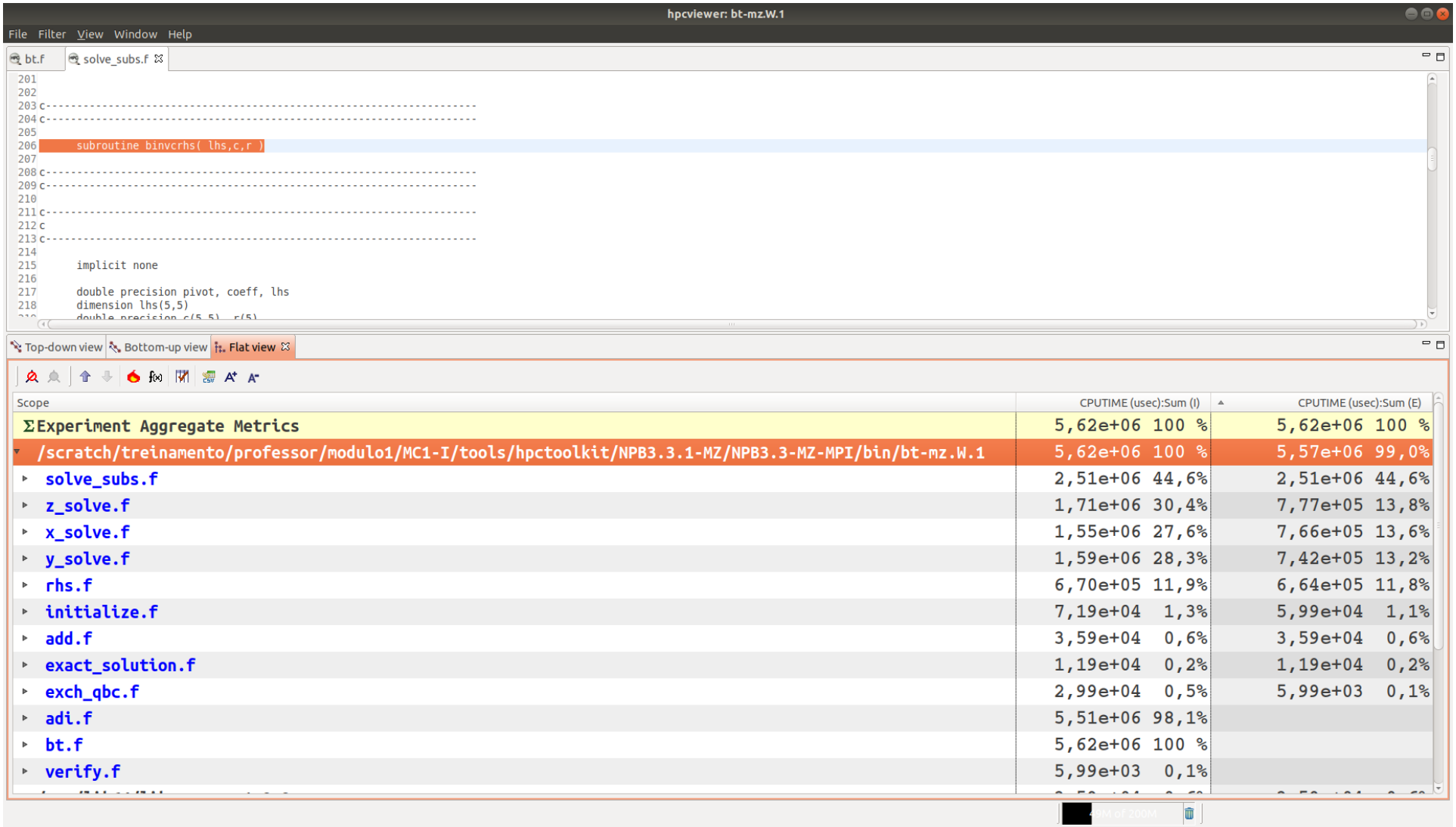
Visualizando no `hpcviewer`

`-nodes=1 -ntasks=1 / Flat view`



Visualizando no `hpcviewer`

`-nodes=1 -ntasks=1 / Flat view`



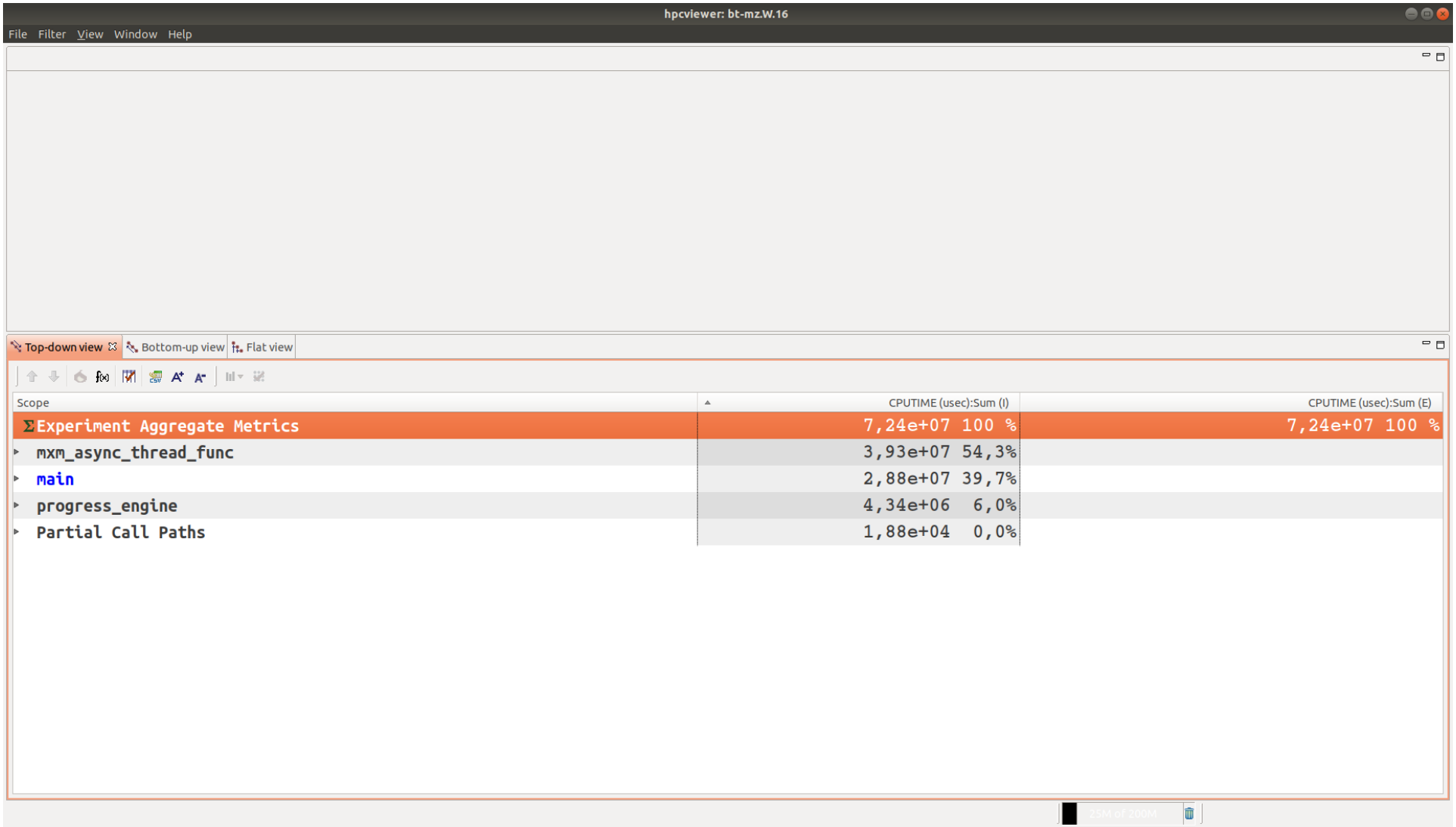
Visualizando no `hpcviewer`

NPB: estudo de caso

```
$ cd profiling/hpctoolkit/NUMNODES-1/bt-mz_W_MPI-16_OMP-1_JOBID-439032
$ hpcviewer hpctoolkit-bt-mz.W.16-database-439032
```

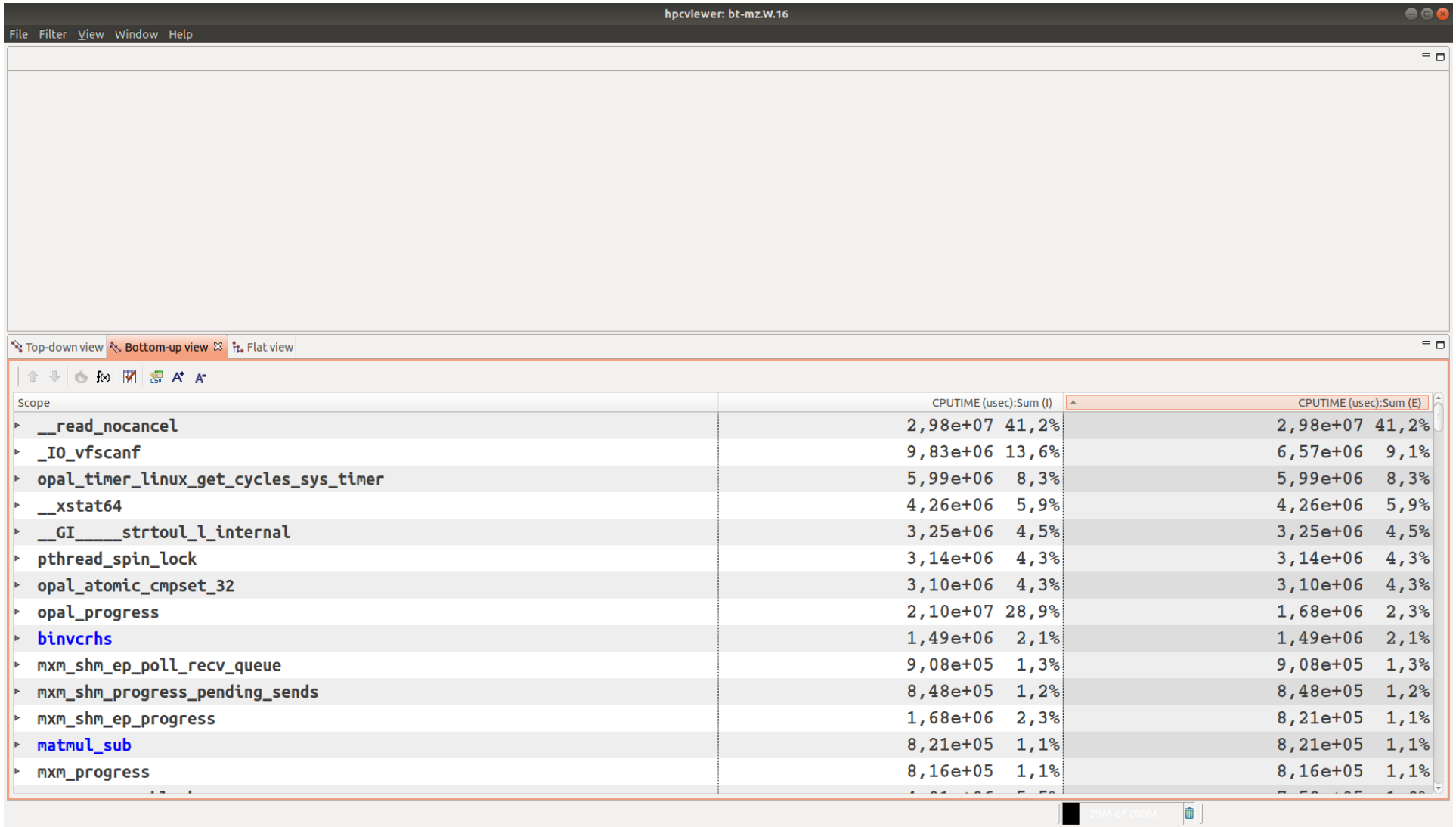
Visualizando no `hpcviewer`

`-nodes=1 -ntasks=16 / Top-down view`



Visualizando no **hpcviewer**

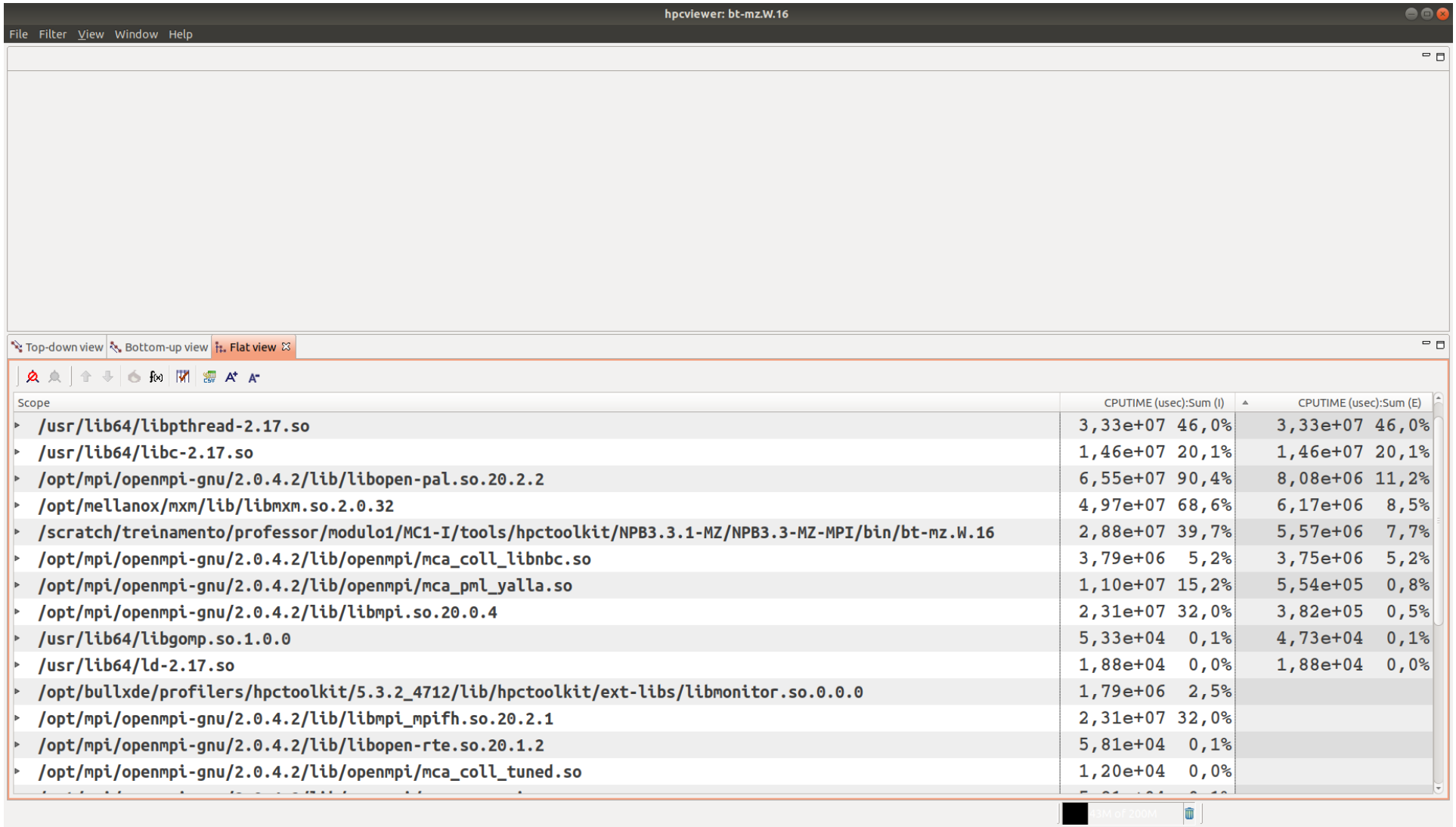
-nodes=1 -ntasks=16 / Bottom-up view



Scope	CPUTIME (usec):Sum (I)		CPUTIME (usec):Sum (E)	
__read_nocancel	2,98e+07	41,2%	2,98e+07	41,2%
_IO_vfscanf	9,83e+06	13,6%	6,57e+06	9,1%
opal_timer_linux_get_cycles_sys_timer	5,99e+06	8,3%	5,99e+06	8,3%
__xstat64	4,26e+06	5,9%	4,26e+06	5,9%
__GI_____strtoul_l_internal	3,25e+06	4,5%	3,25e+06	4,5%
pthread_spin_lock	3,14e+06	4,3%	3,14e+06	4,3%
opal_atomic_cmpset_32	3,10e+06	4,3%	3,10e+06	4,3%
opal_progress	2,10e+07	28,9%	1,68e+06	2,3%
binvcrhs	1,49e+06	2,1%	1,49e+06	2,1%
mxm_shm_ep_poll_recv_queue	9,08e+05	1,3%	9,08e+05	1,3%
mxm_shm_progress_pending_sends	8,48e+05	1,2%	8,48e+05	1,2%
mxm_shm_ep_progress	1,68e+06	2,3%	8,21e+05	1,1%
matmul_sub	8,21e+05	1,1%	8,21e+05	1,1%
mxm_progress	8,16e+05	1,1%	8,16e+05	1,1%

Visualizando no `hpcviewer`

`-nodes=1 -ntasks=16 / Flat view`



Scope	CPUTIME (usec):Sum (I)		CPUTIME (usec):Sum (E)	
/usr/lib64/libpthread-2.17.so	3,33e+07	46,0%	3,33e+07	46,0%
/usr/lib64/libc-2.17.so	1,46e+07	20,1%	1,46e+07	20,1%
/opt/mpi/openmpi-gnu/2.0.4.2/lib/libopen-pal.so.20.2.2	6,55e+07	90,4%	8,08e+06	11,2%
/opt/mellanox/mxm/lib/libmxm.so.2.0.32	4,97e+07	68,6%	6,17e+06	8,5%
/scratch/treinamento/professor/modulo1/MC1-I/tools/hpctoolkit/NPB3.3.1-MZ/NPB3.3-MZ-MPI/bin/bt-mz.W.16	2,88e+07	39,7%	5,57e+06	7,7%
/opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_coll_libnbc.so	3,79e+06	5,2%	3,75e+06	5,2%
/opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_pml_yalla.so	1,10e+07	15,2%	5,54e+05	0,8%
/opt/mpi/openmpi-gnu/2.0.4.2/lib/libmpi.so.20.0.4	2,31e+07	32,0%	3,82e+05	0,5%
/usr/lib64/libgomp.so.1.0.0	5,33e+04	0,1%	4,73e+04	0,1%
/usr/lib64/ld-2.17.so	1,88e+04	0,0%	1,88e+04	0,0%
/opt/bullxde/profilers/hpctoolkit/5.3.2_4712/lib/hpctoolkit/ext-libs/libmonitor.so.0.0.0	1,79e+06	2,5%		
/opt/mpi/openmpi-gnu/2.0.4.2/lib/libmpi_mpifh.so.20.2.1	2,31e+07	32,0%		
/opt/mpi/openmpi-gnu/2.0.4.2/lib/libopen-rte.so.20.1.2	5,81e+04	0,1%		
/opt/mpi/openmpi-gnu/2.0.4.2/lib/openmpi/mca_coll_tuned.so	1,20e+04	0,0%		

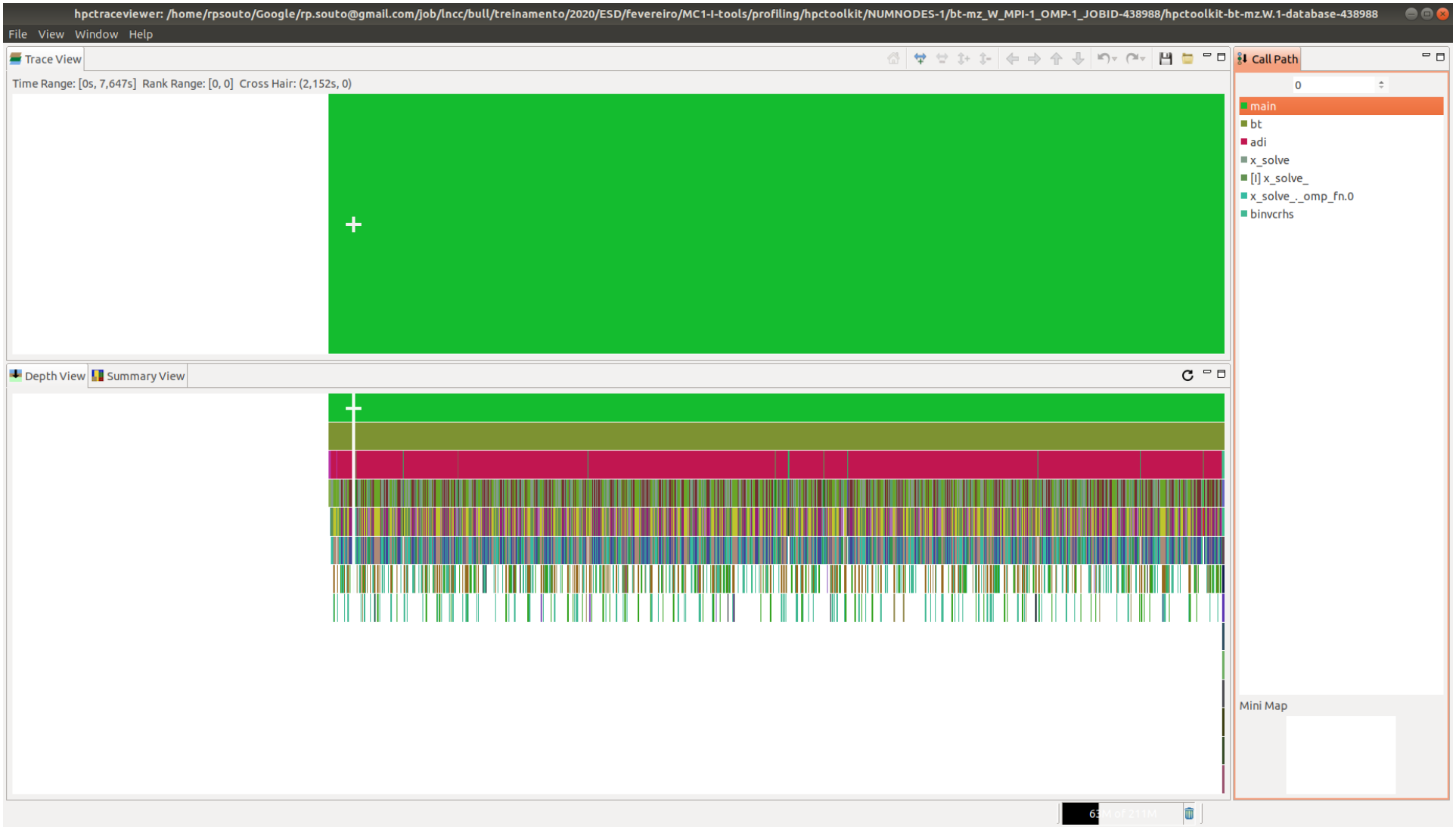
Visualizando no `hpctraceview`

NPB: estudo de caso

```
$ cd profiling/hpctoolkit/NUMNODES-1/bt-mz_W_MPI-1_OMP-1_JOBID-438988
$ hpctraceview hpctoolkit-bt-mz.W.1-database-438988/
```

Visualizando no `hpctraceview`

`-nodes=1 -ntasks=1`



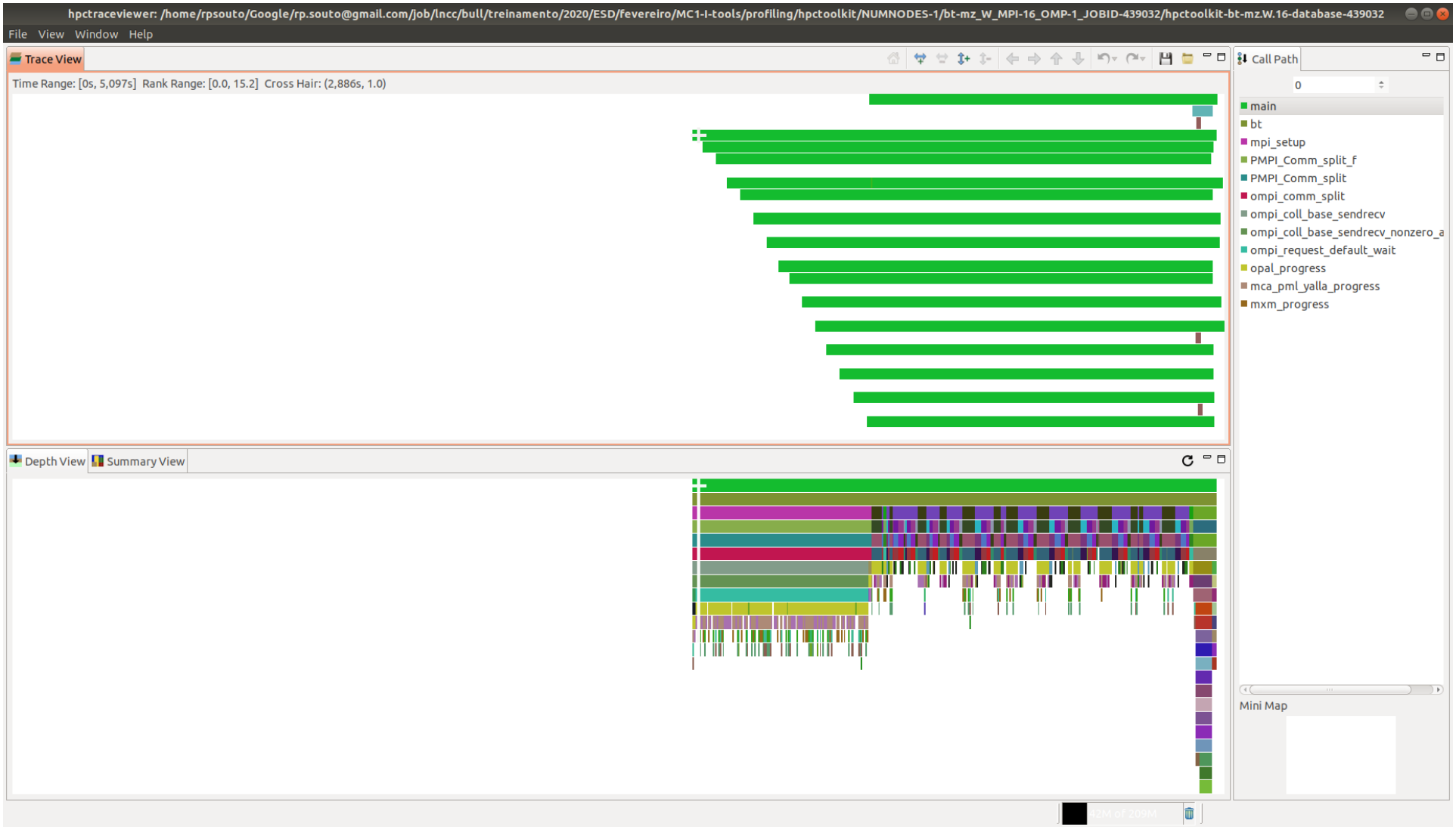
Visualizando no `hpctraceview`

NPB: estudo de caso

```
$ cd profiling/hpctoolkit/NUMNODES-1/bt-mz_W_MPI-16_OMP-1_JOBID-439032
$ hpctraceview hpctoolkit-bt-mz.W.16-database-439032
```

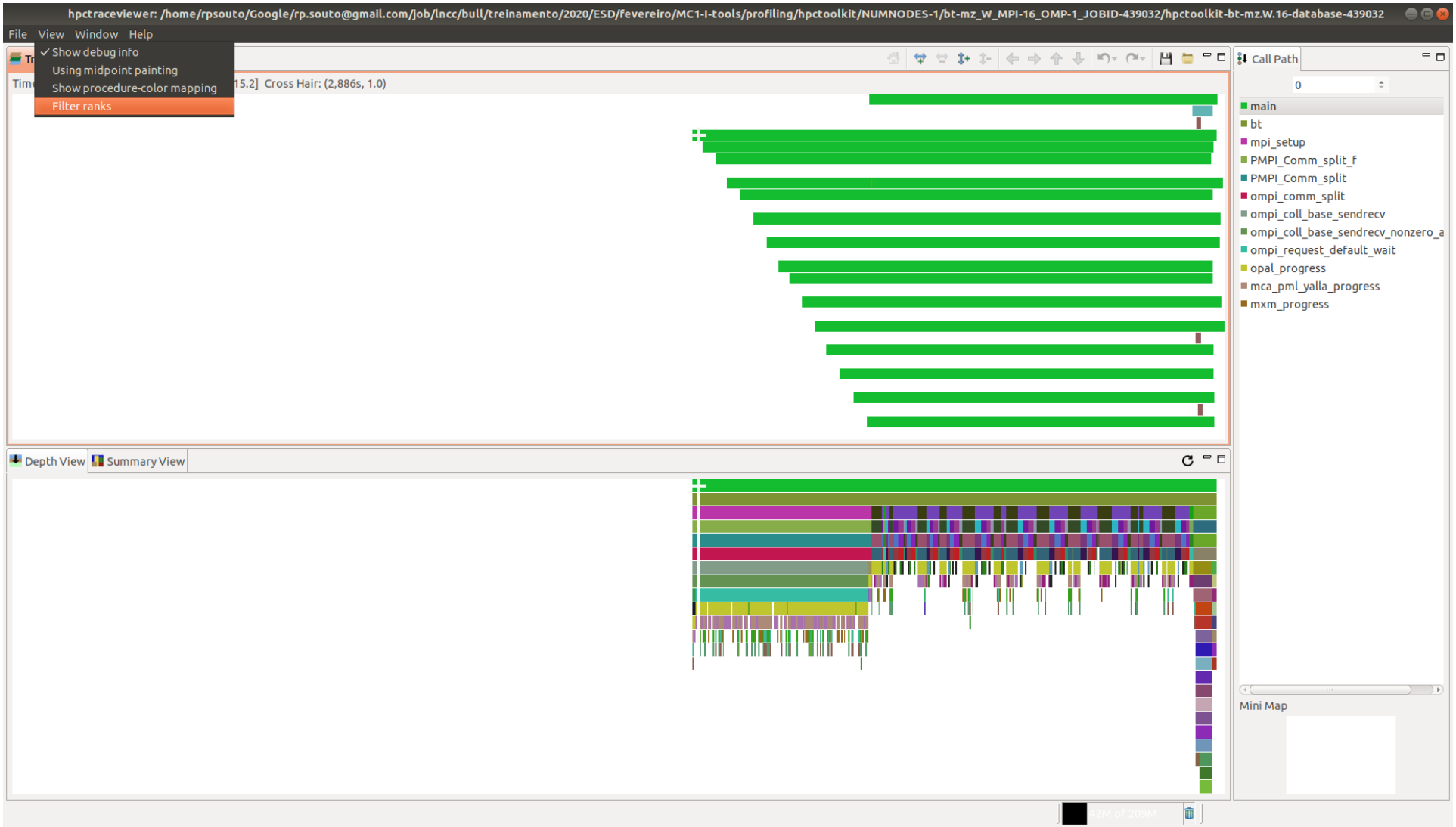
Visualizando no `hpctraceview`

`-nodes=1 -ntasks=16`



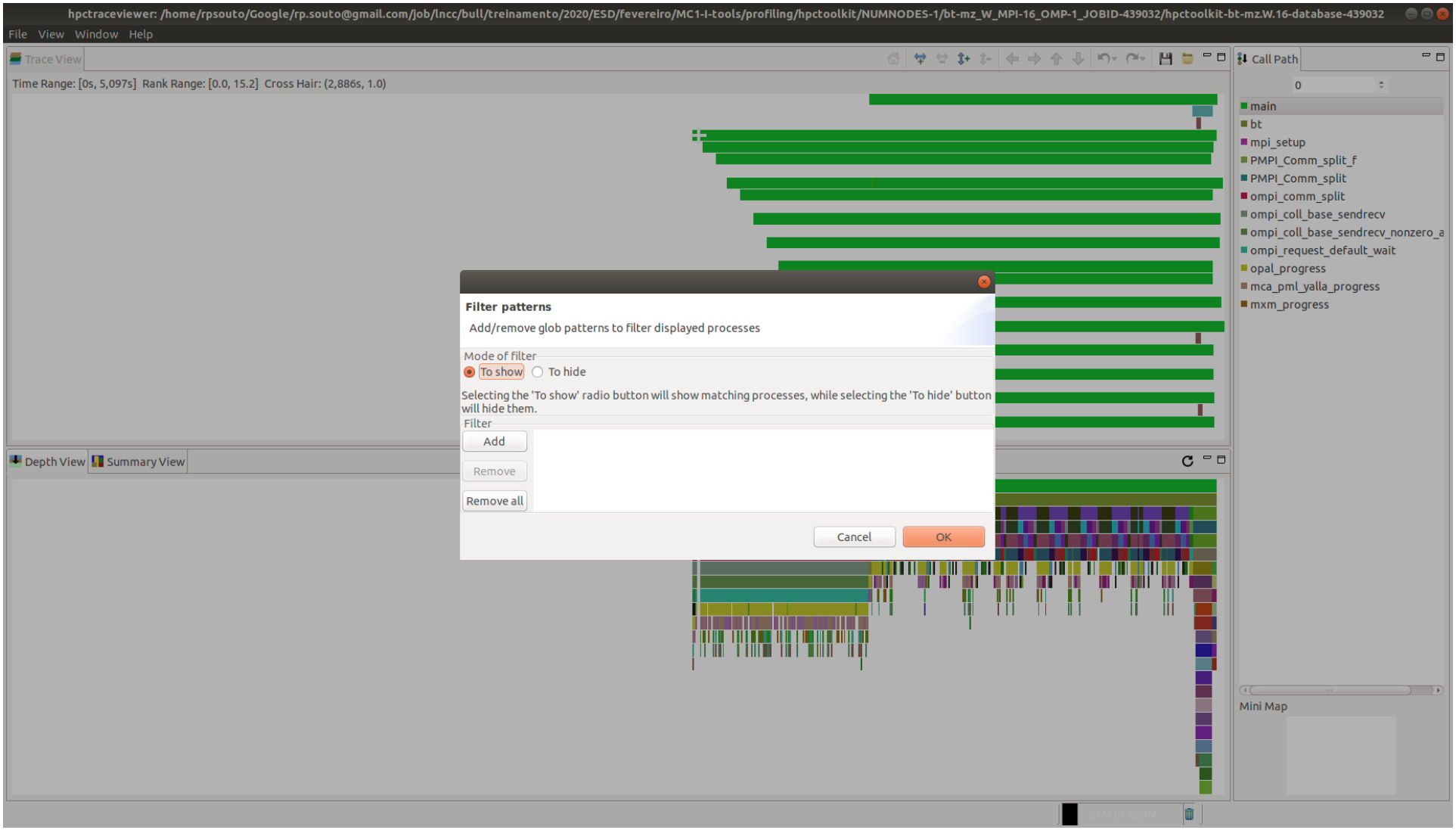
Visualizando no `hpctraceview`

`-nodes=1 -ntasks=16`



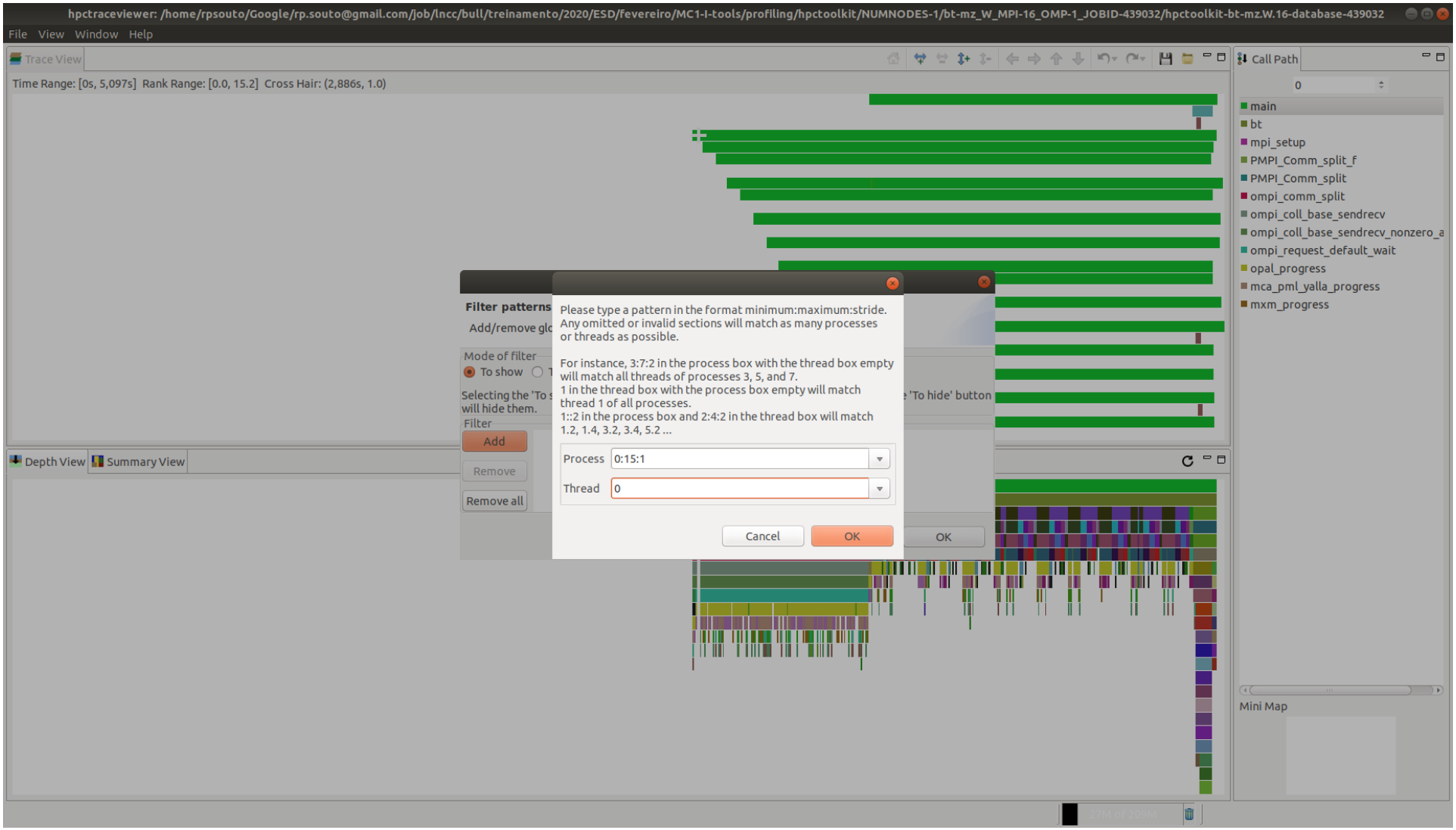
Visualizando no `hpctraceview`

`-nodes=1 -ntasks=16`



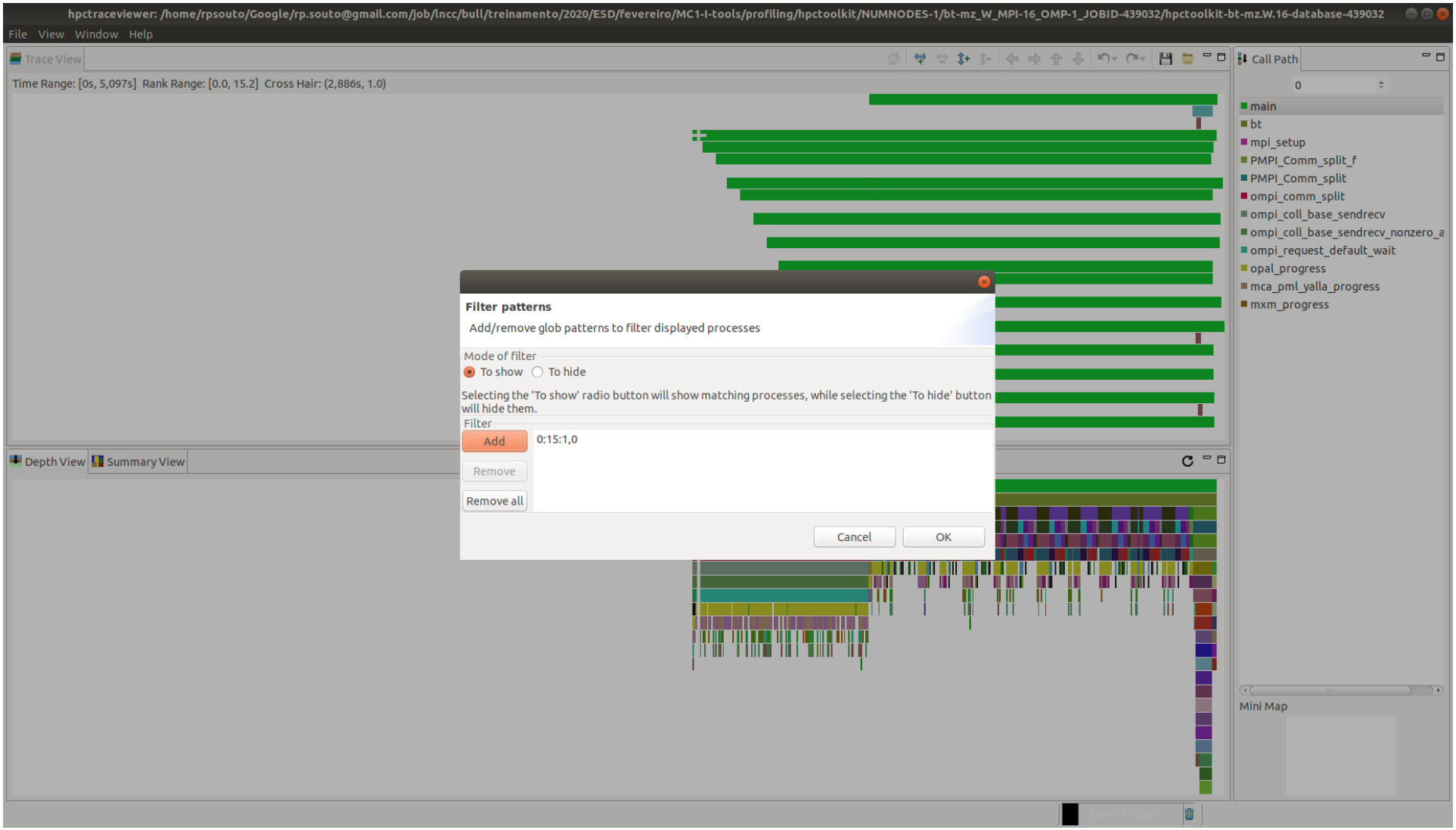
Visualizando no `hpctraceview`

`-nodes=1 -ntasks=16`



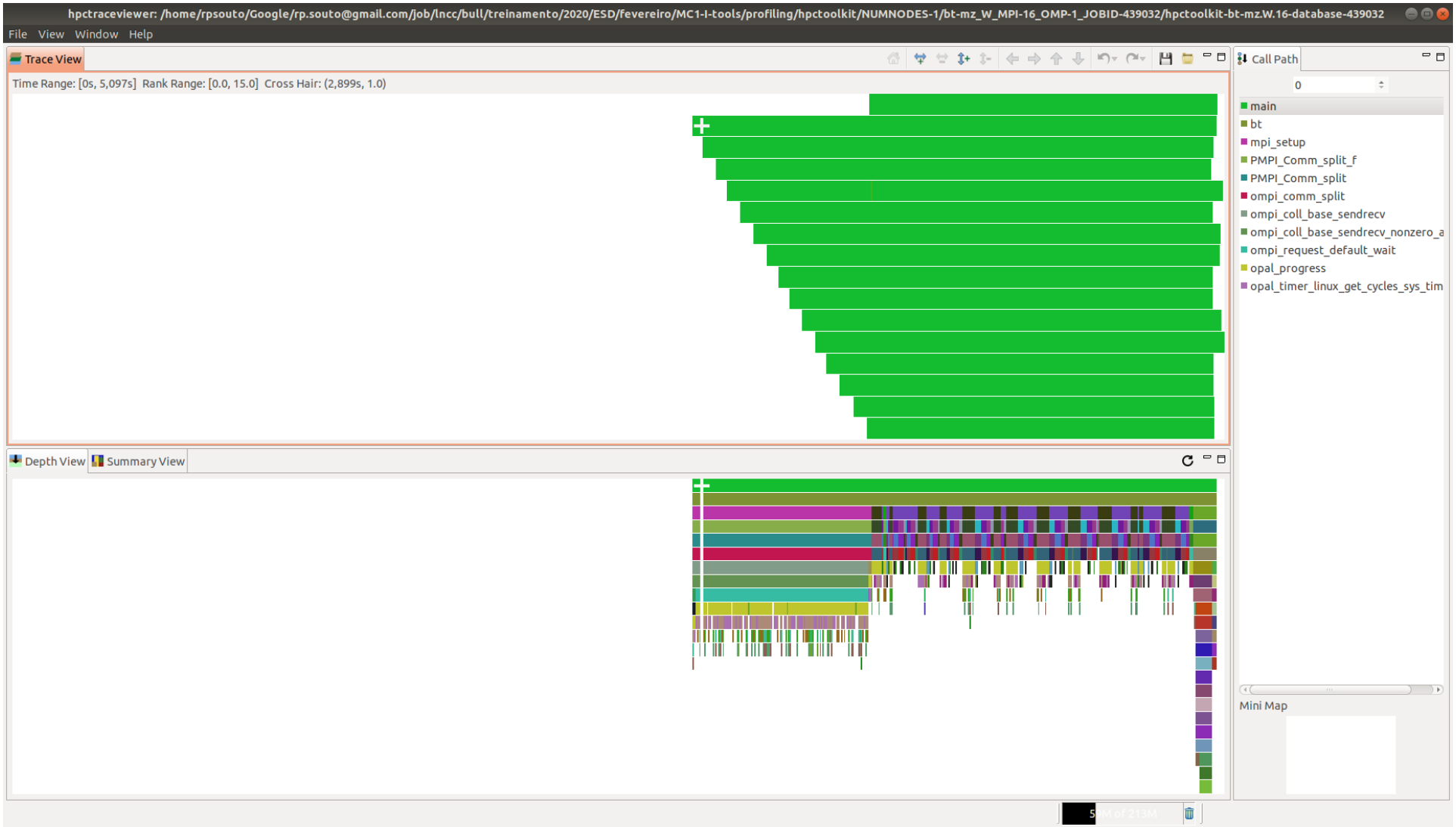
Visualizando no `hpctraceview`

`-nodes=1 -ntasks=16`



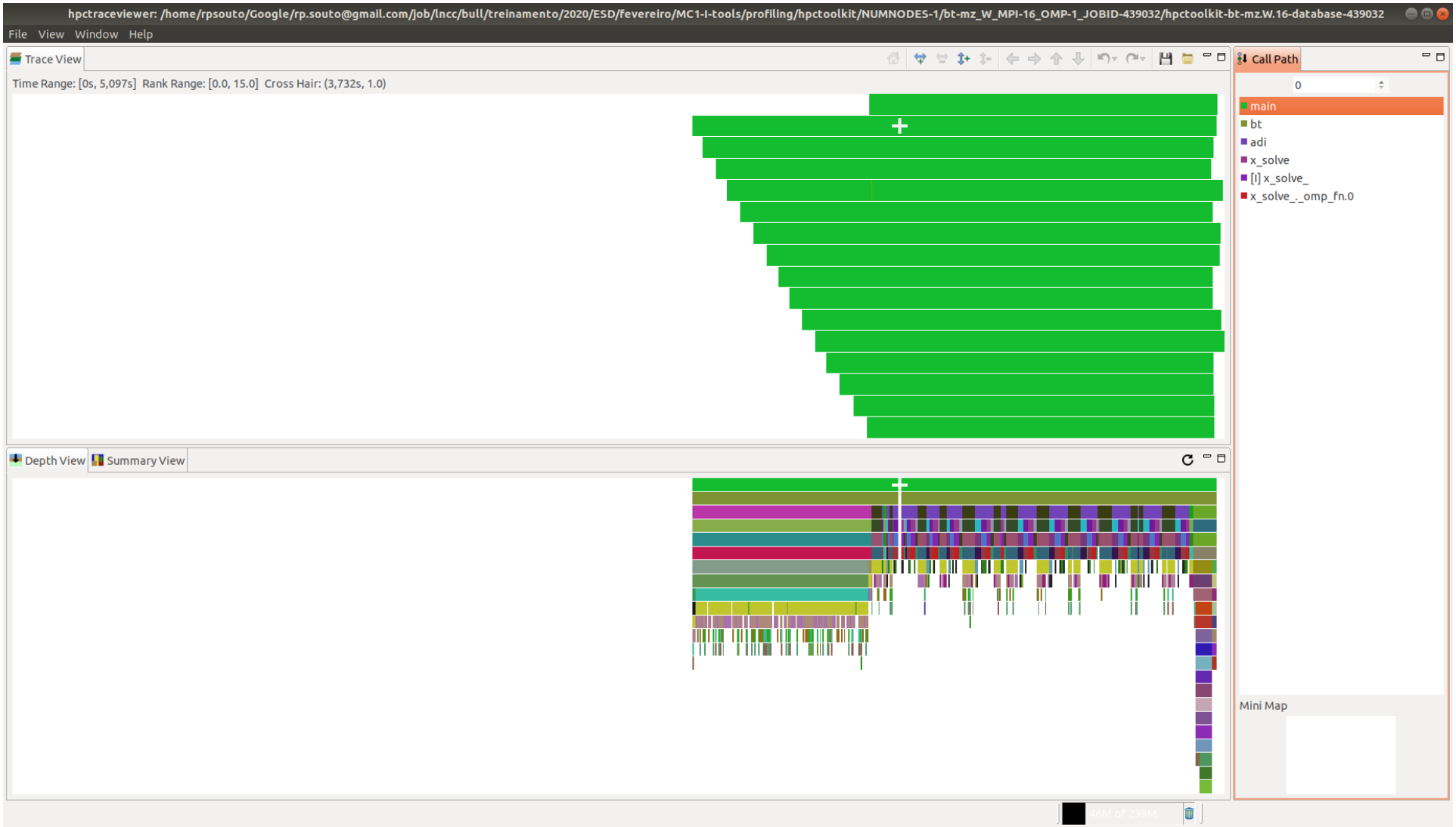
Visualizando no `hpctraceview`

`-nodes=1 -ntasks=16`



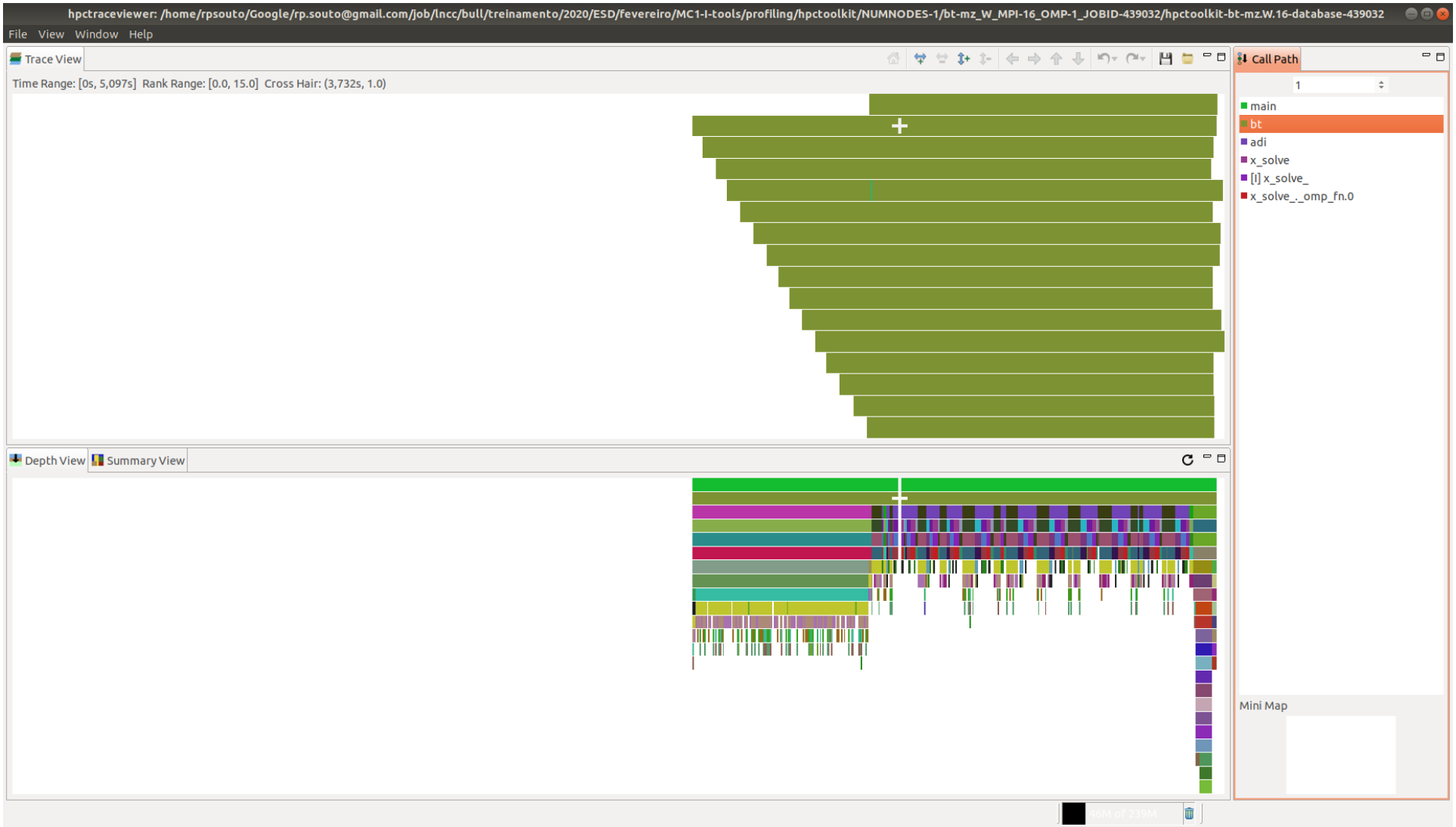
Visualizando no **hpctraceview**

-nodes=1 -ntasks=16 - função **main**



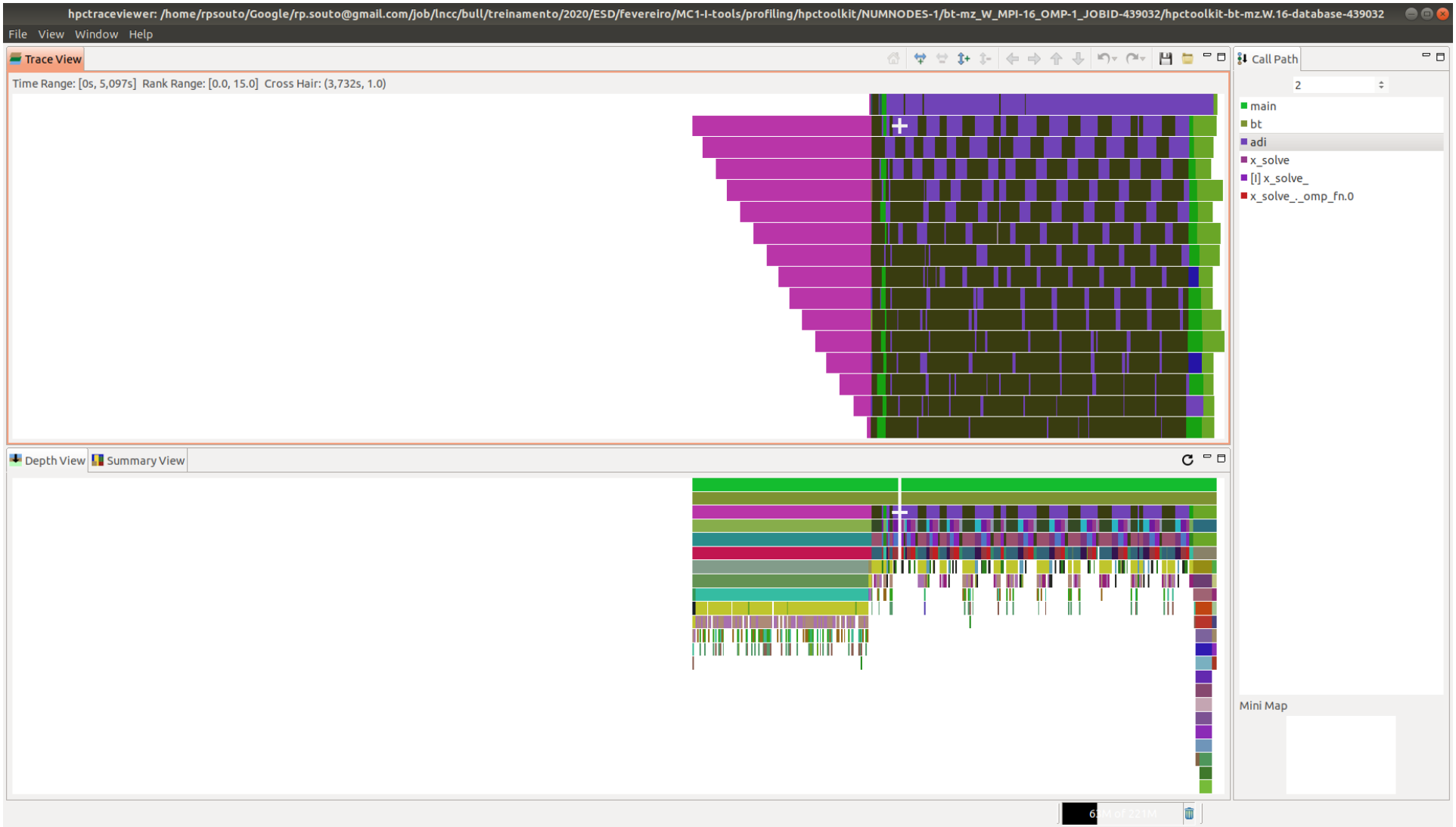
Visualizando no **hpctraceview**

-nodes=1 -ntasks=16 – função **bt**



Visualizando no **hpctraceview**

-nodes=1 -ntasks=16 – função **adi**



Visualizando no **hpctraceview**

-nodes=1 -ntasks=16 - função **xsolve**