Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

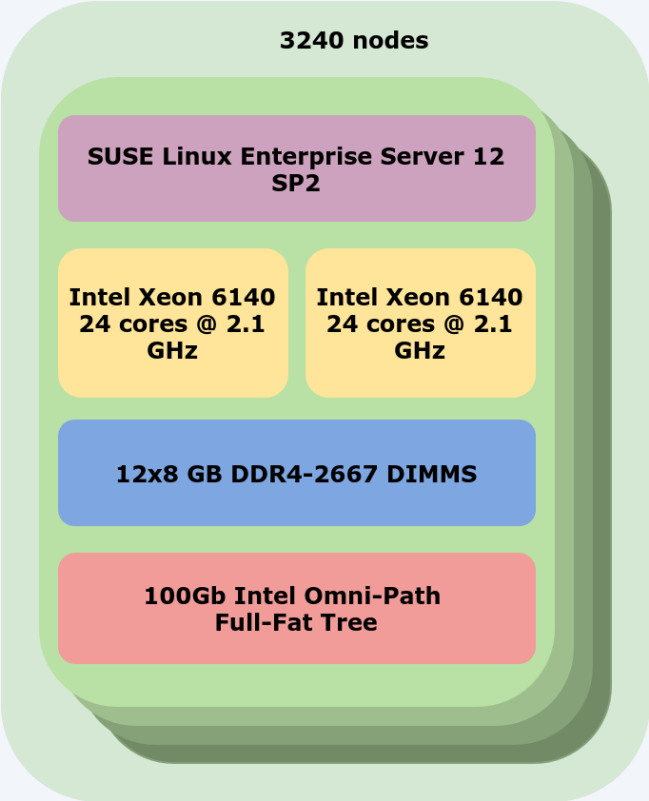
ERO2.0 Performance Assessment

Joan Vinyals and Marta Garcia

Environment

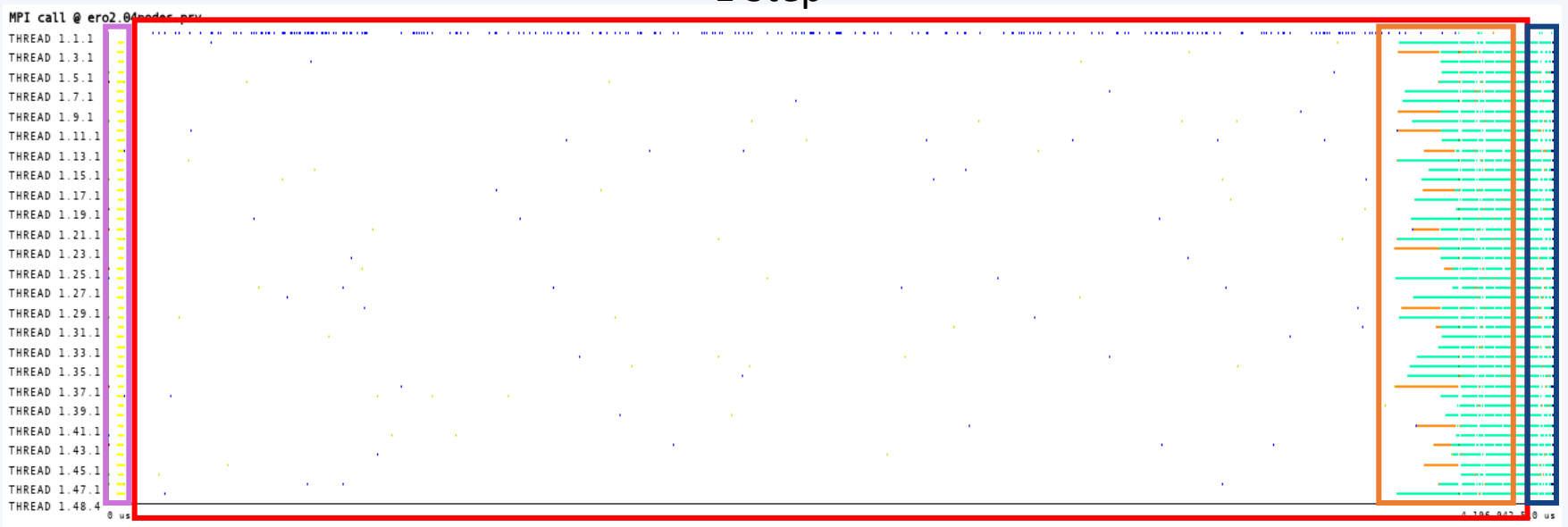
- Intel Compiler 2017.4
- MPI: IMPI 2017.4
- Libraries:
  - MKL 2017.4
  - HDF5 1.8.19
  - BOOST 175.0Z
- Input files from:
  - https://jugit.fz-juelich.de/ero/runs/jromazanov/jet/run03/seq01
  - randomSeed = false

MareNostrum 4



Structure

1 Step



Initialization

Computation (1 step)

Gather

Finalization

Scalability

4 nodes

8 nodes

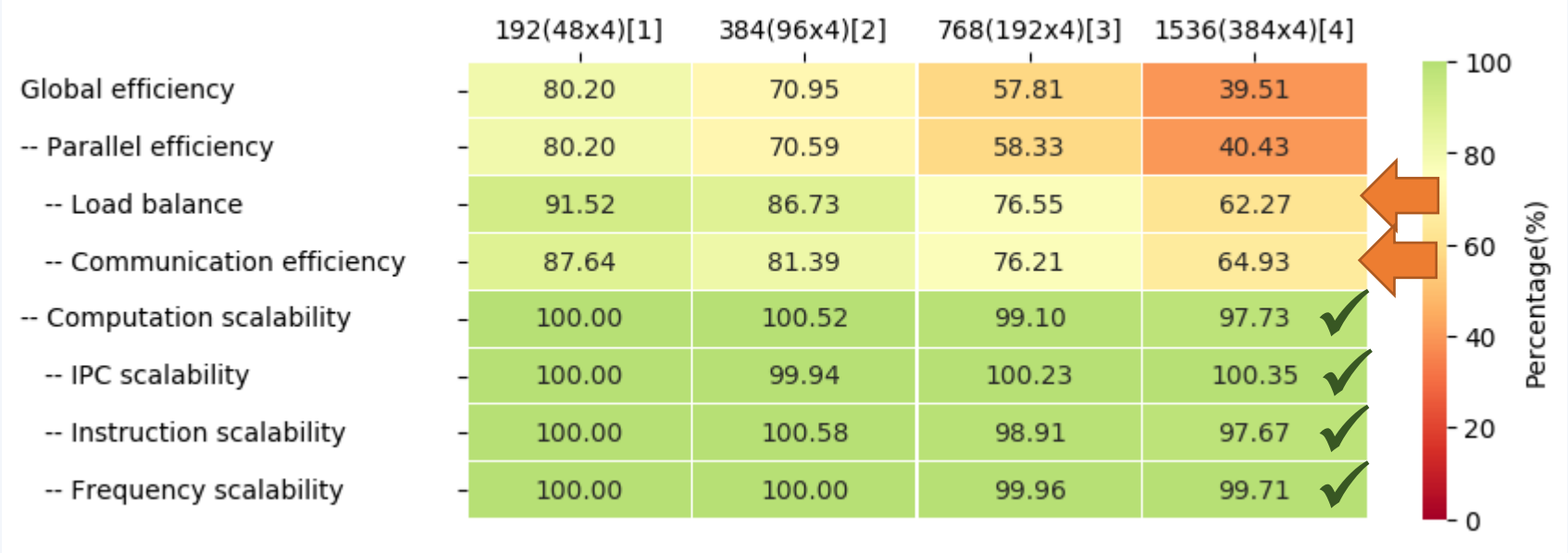
16 nodes

32 nodes

- Computation seems to scale well
- Gather phase does not scale

Efficiency metrics
MPI scaling

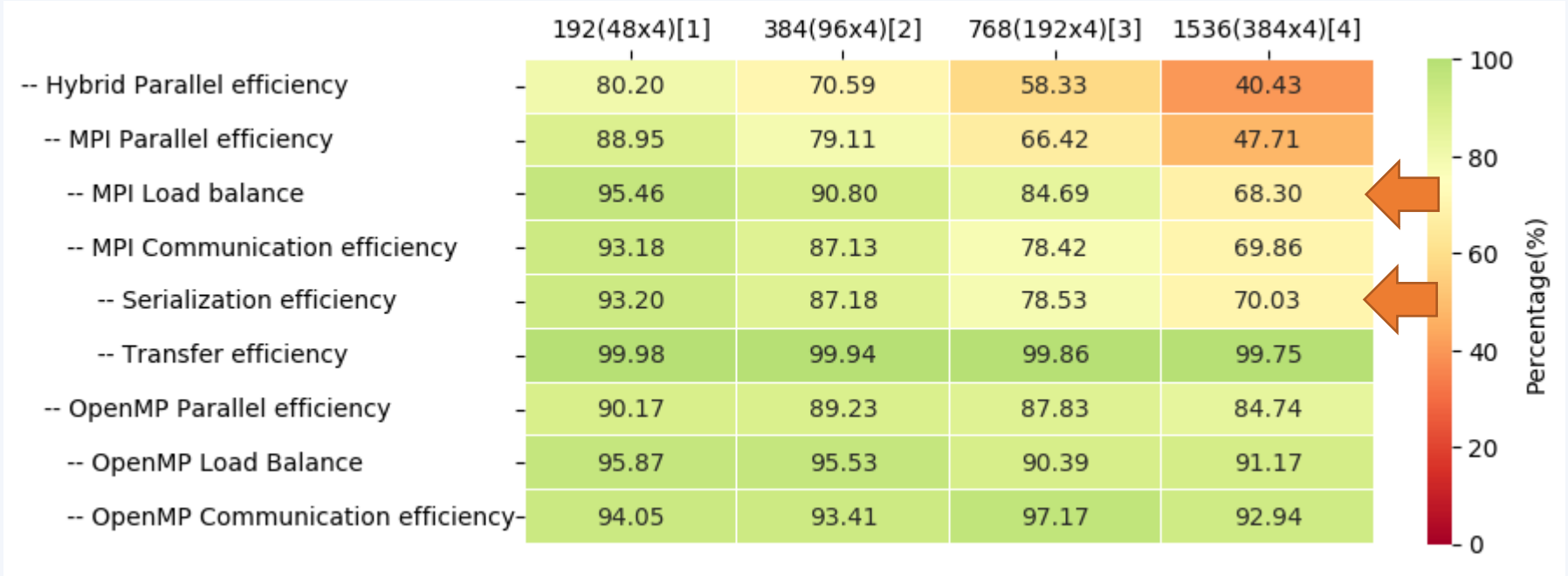
Nodes: 4 8 16 32



- Main scaling issues:
  - Load balance
  - Serialization
- Very good computation scalability

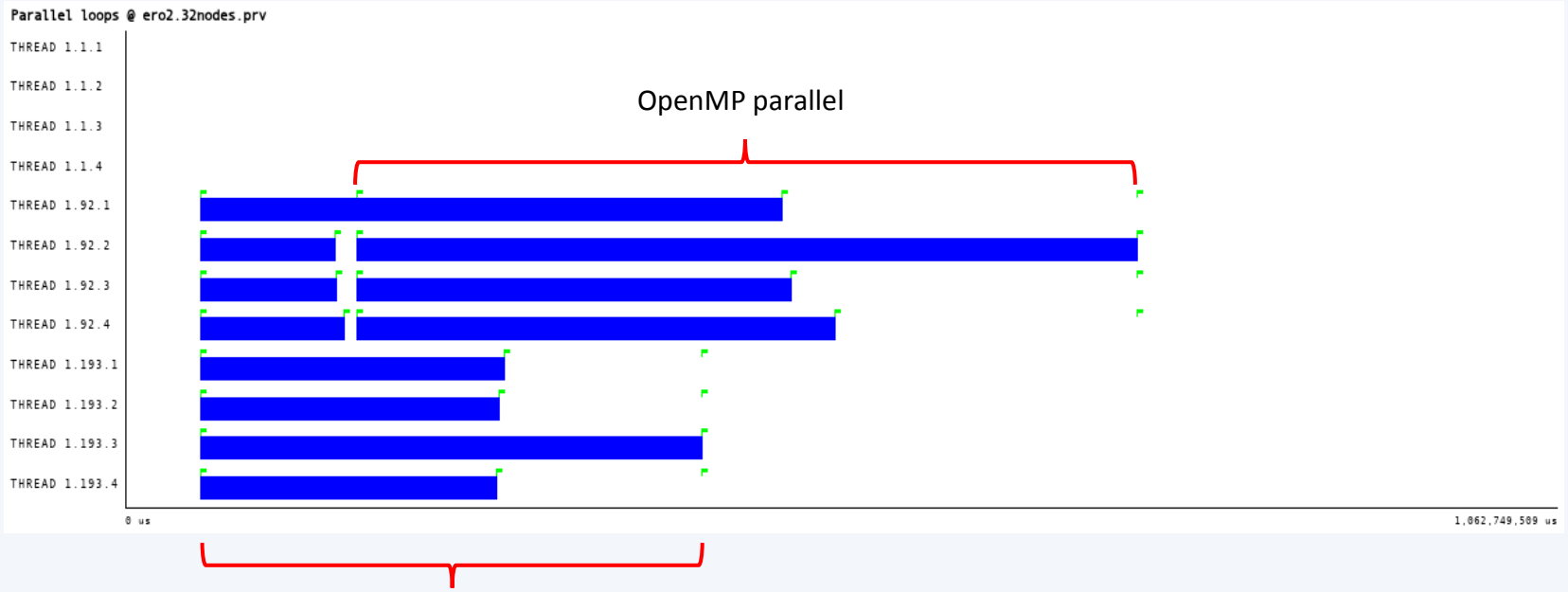
Hybrid Parallel Efficiency

Nodes: 4 8 16 32

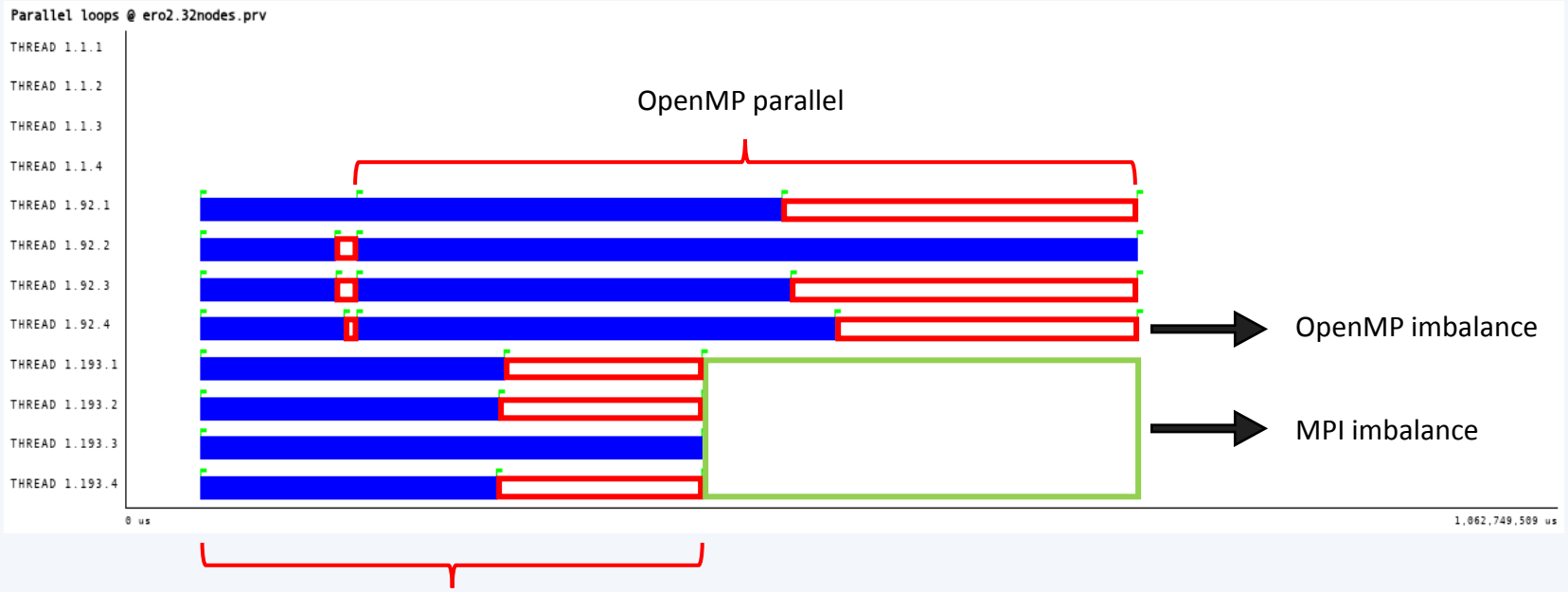


- Communication issues are due to Serialization

Load imbalance in detail



Load imbalance in detail

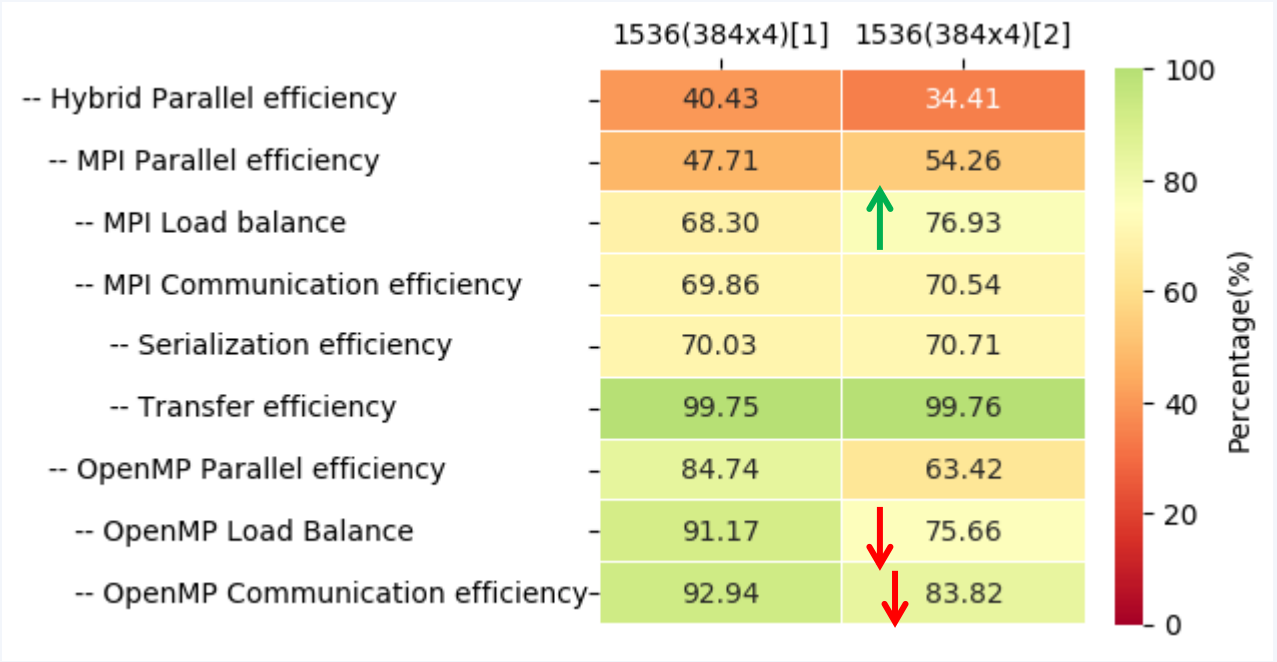


Load Balance Summary

- How **maxMpiChunkSize** affects load balance?
- OpenMP Load Imbalance
  - Suggestion: Taskify
- MPI Load Imbalance
  - Suggestion: Using Dynamic Load Balancing (DLB) library.

Load Imbalance – MPI grain

maxMpiChunkSize: 50 10



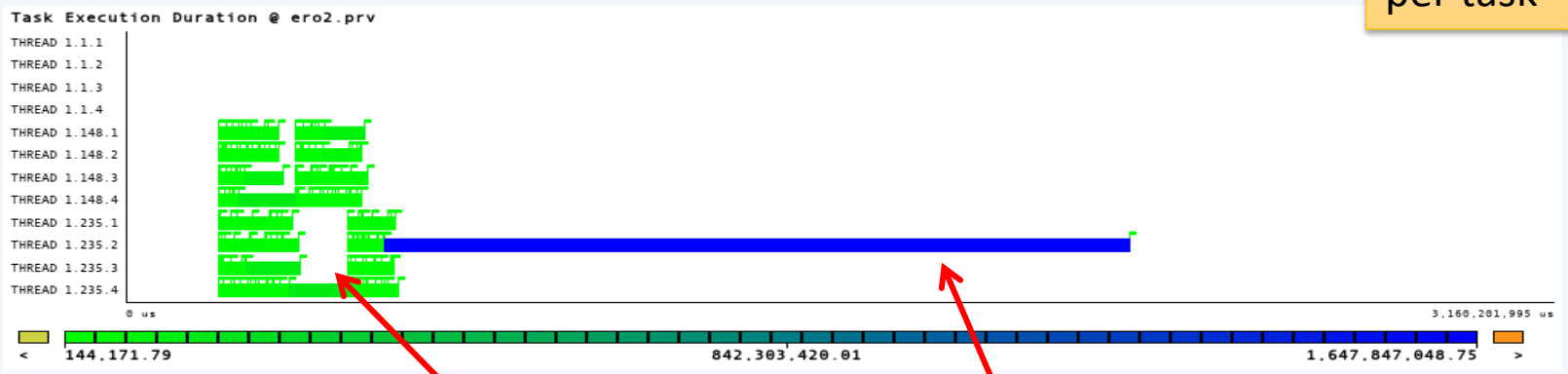
Overall Worst Load Balance

Taskification

```
#pragma omp parallel
{
   #pragma omp for
   for (int i=0; i<chunkSize; i++) {
      …
      transportParticleLoop (…);
      …
   }
}
```

```
#pragma omp parallel masked
{
   for (int i=0; i<chunkSize; i++) {
      #pragma omp task {
      …
      transportParticleLoop (…);
      …}
   }
}
```

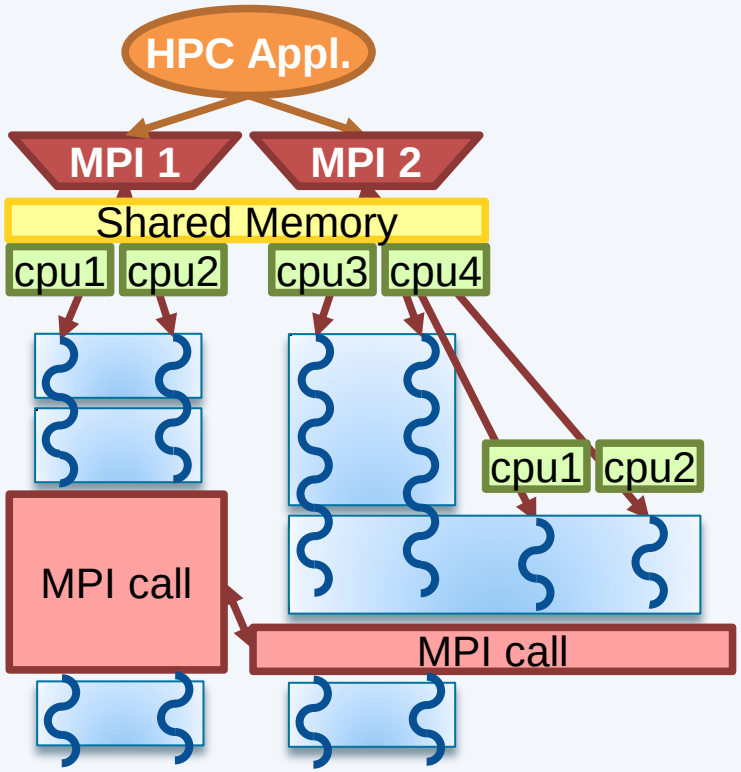
One particle per task



OpenMP Imbalance

Computing one particle

MPI Load Imbalance - DLB



- DLB is a runtime library transparent to the user
- Relays on OpenMP to adjust number of threads

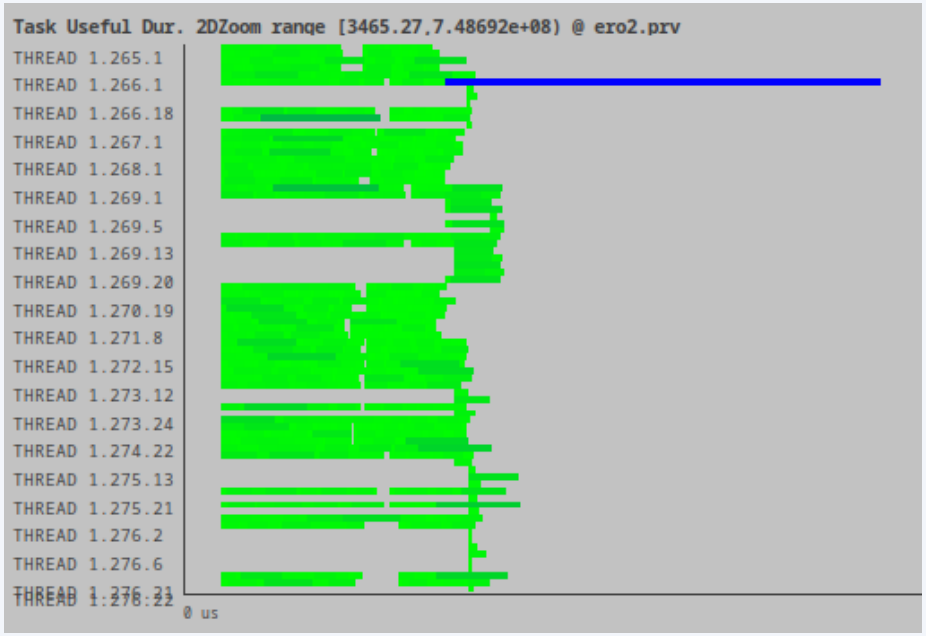
MPI Load Imbalance - DLB

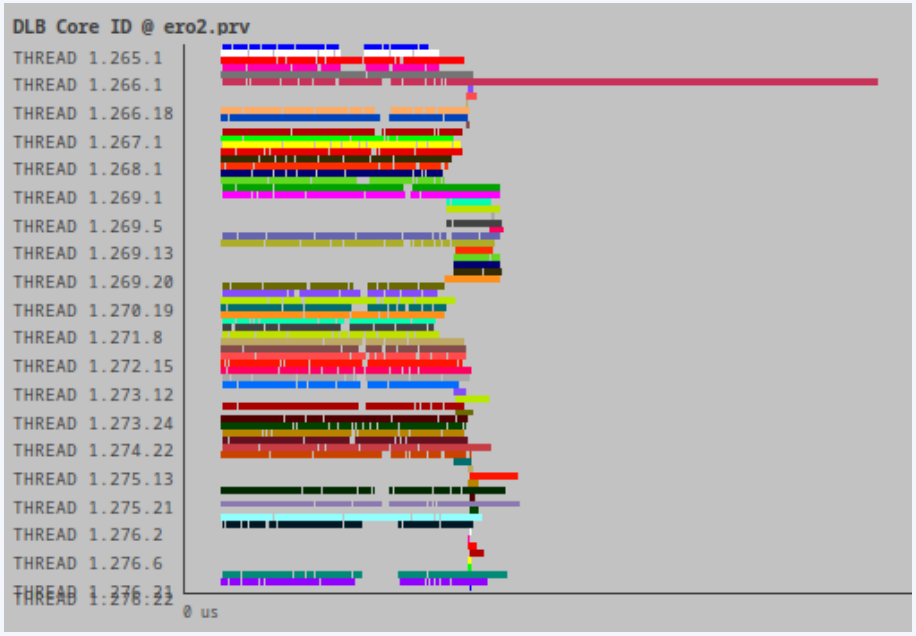
```
if (tag == QUIT) {
  MPI_Recv(...);
  break;
}
```

```
if (tag == QUIT) {
  MPI_Recv(...);
  DLB_Barrier();
  break;
}
```

```
export DLB_ARGS="--lewi --ompt --ompt-thread-manager=free-agents --lewi-ompt=mpi:borrow"
export OMP_WAIT_POLICY=passive
export OMP_NUM_THREADS=4
export KMP_FREE_AGENT_NUM_THREADS=20 # 4 + 20 = one socket
preload="$OMP_HOME/lib/libomp.so"
$DLB_HOME/bin/dlb_run --verbose env LD_PRELOAD="$LD_PRELOAD:$preload" ero2 -f ero2_config.xml
```

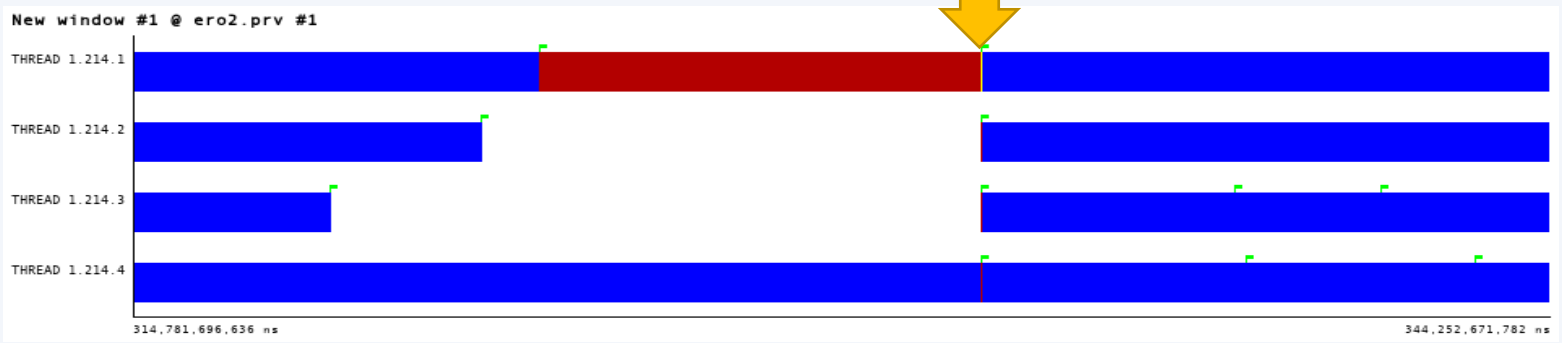
MPI Load Imbalance - DLB





OpenMP Load Imbalance - Advance communication

MPI Comm.

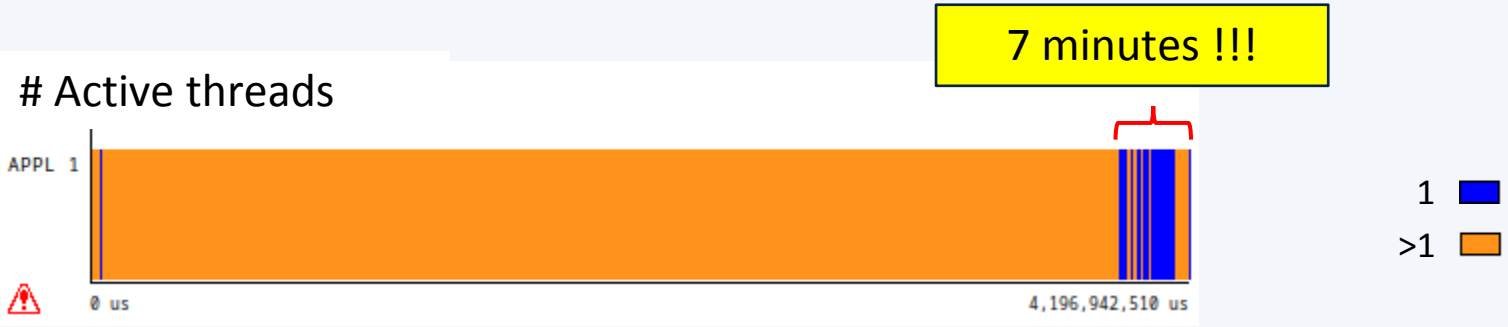


MPI Comm.

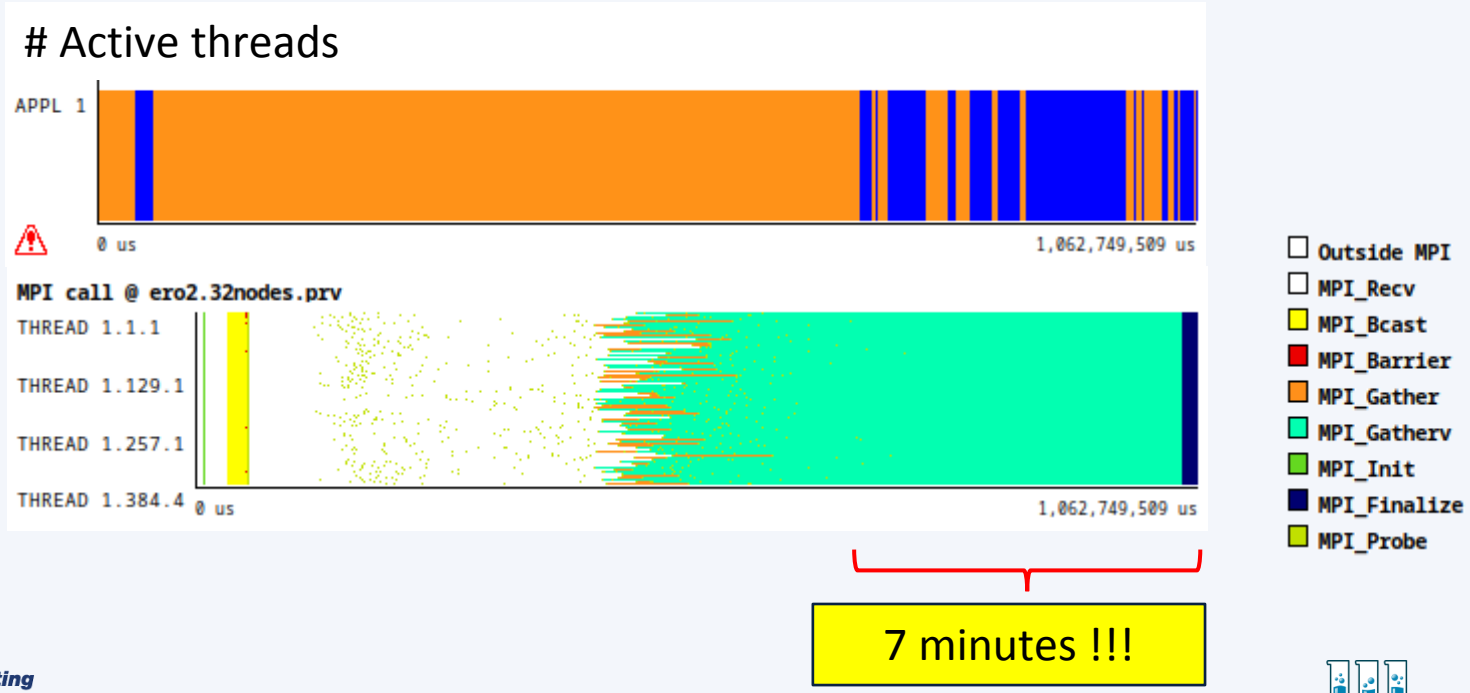
Testing idea in a mock-up code. Cannot visualize it yet.

Serialization Efficiency

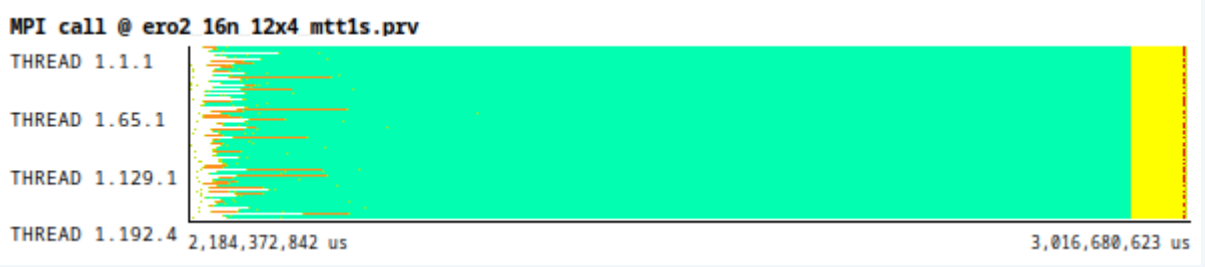
4 nodes

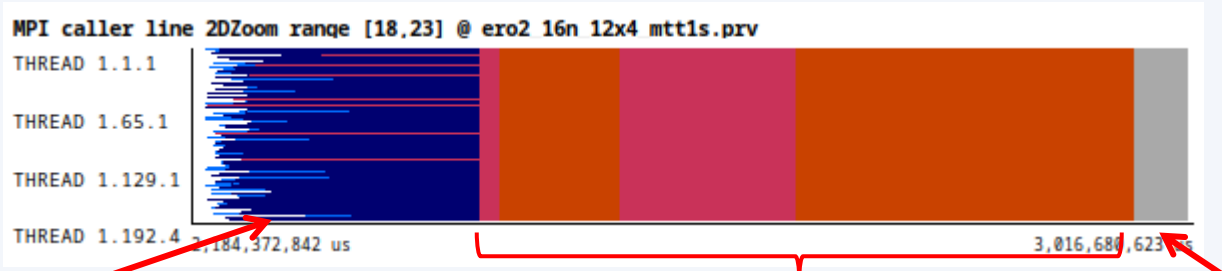


32 nodes



Serial functions



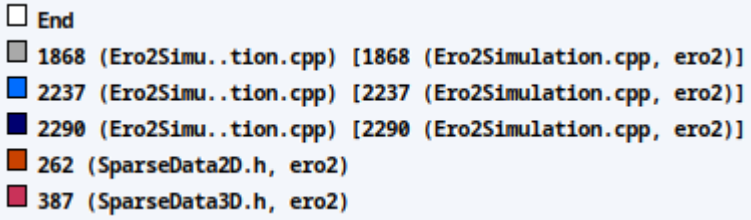


MPI Load Balance

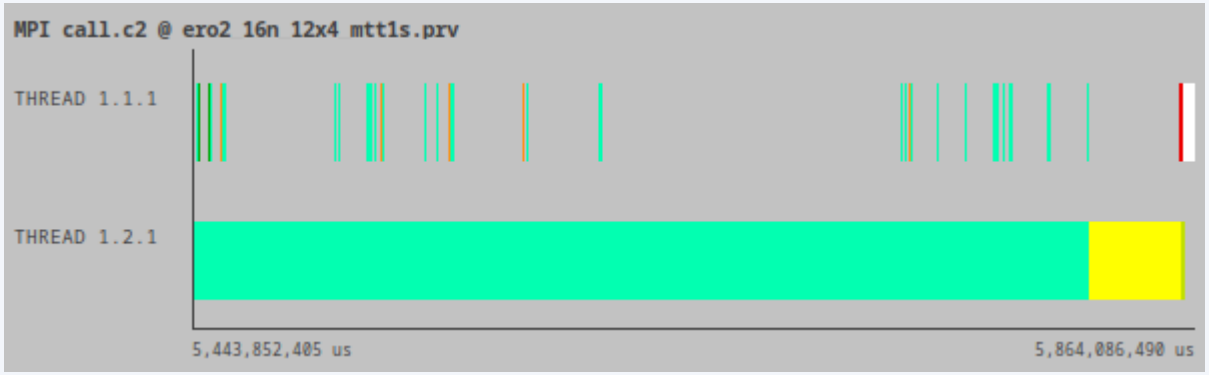
Waiting particles to finish

Serial functions

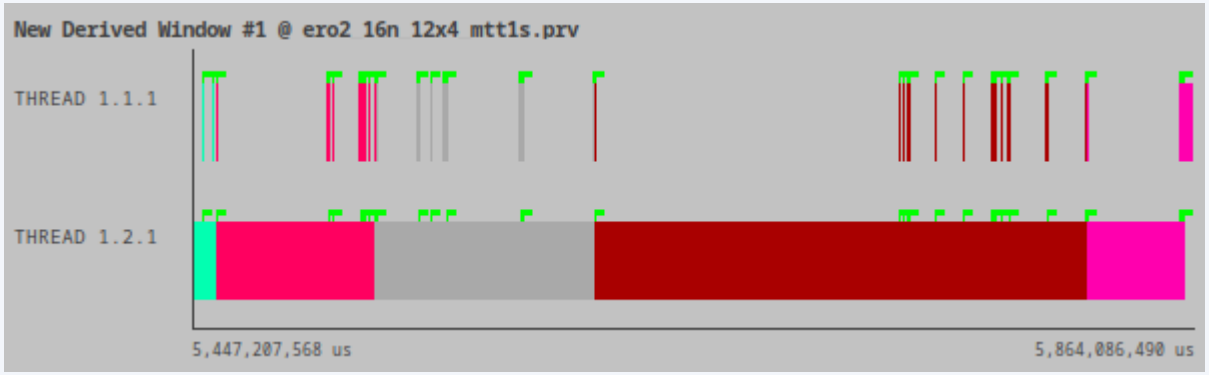
Work redistribution



Serial functions



MPI 2nd level caller



TODO/CONSIDER/Questions

- Can any part of *transportParticleLoop* be parallelised?
  - Yes: Use DLB
  - No: Can particles with extensive *transportParticleLoop* times be aborted, when blocking?
    - Is this what *maxTracingTime* does?
- OpenMP load balance:
  - Integrate communication inside the OpenMP parallelisation (advance communication)
  - Effect of the *maxMpiChunkSize*:
    - Bigger: Worst MPI Balance; better OMP Balance
    - Smaller: Better MPI Balance; worst OMP Balance
- Serialization:
  - Is gather phase possible to parallelise?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

marta.garcia@bsc.es

Joan.vinyals@bsc.es