



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



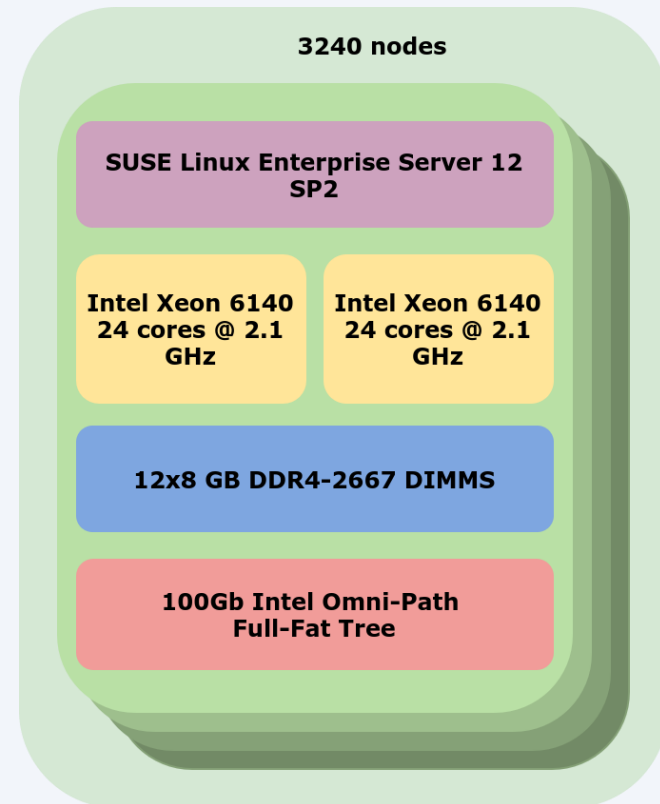
ERO2.0 Performance Assessment

Joan Vinyals and Marta Garcia

Environment

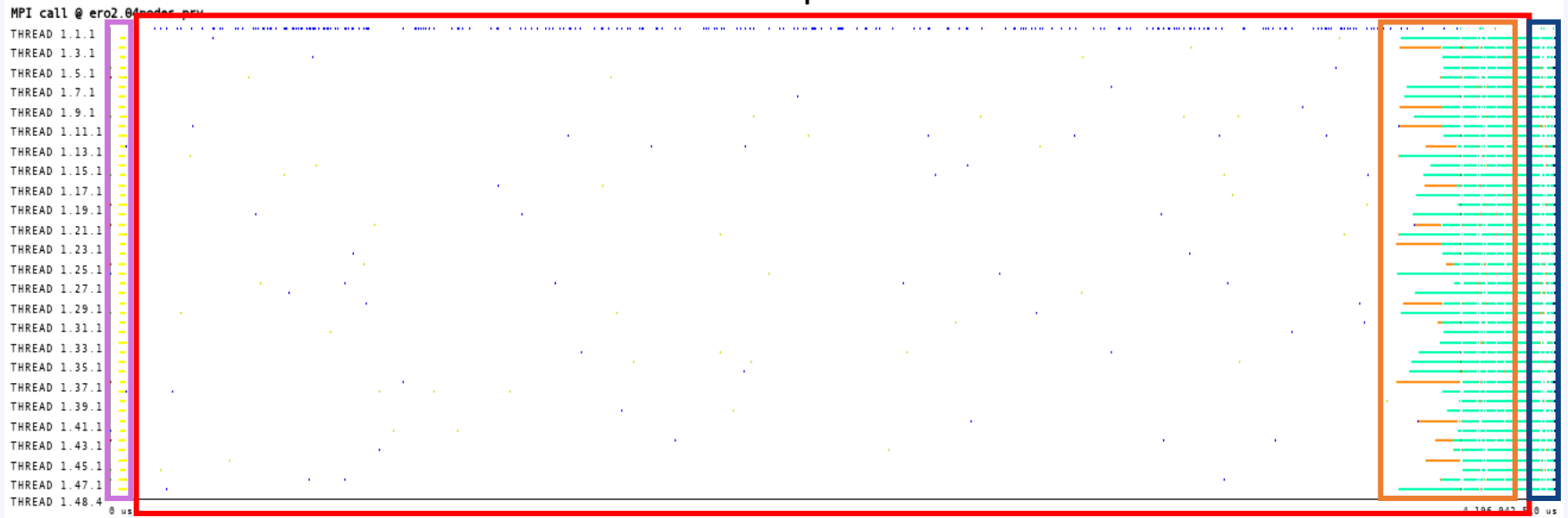
- Intel Compiler 2017.4
- MPI: IMPI 2017.4
- Libraries:
 - MKL 2017.4
 - HDF5 1.8.19
 - BOOST 175.0Z
- Input files from:
 - <https://jugit.fz-juelich.de/ero/runs/jromazanov/jet/run03/seq01>
 - randomSeed = false

MareNostrum 4



Structure

1 Step



Initialization

Computation
(1 step)

Gather

Finalization

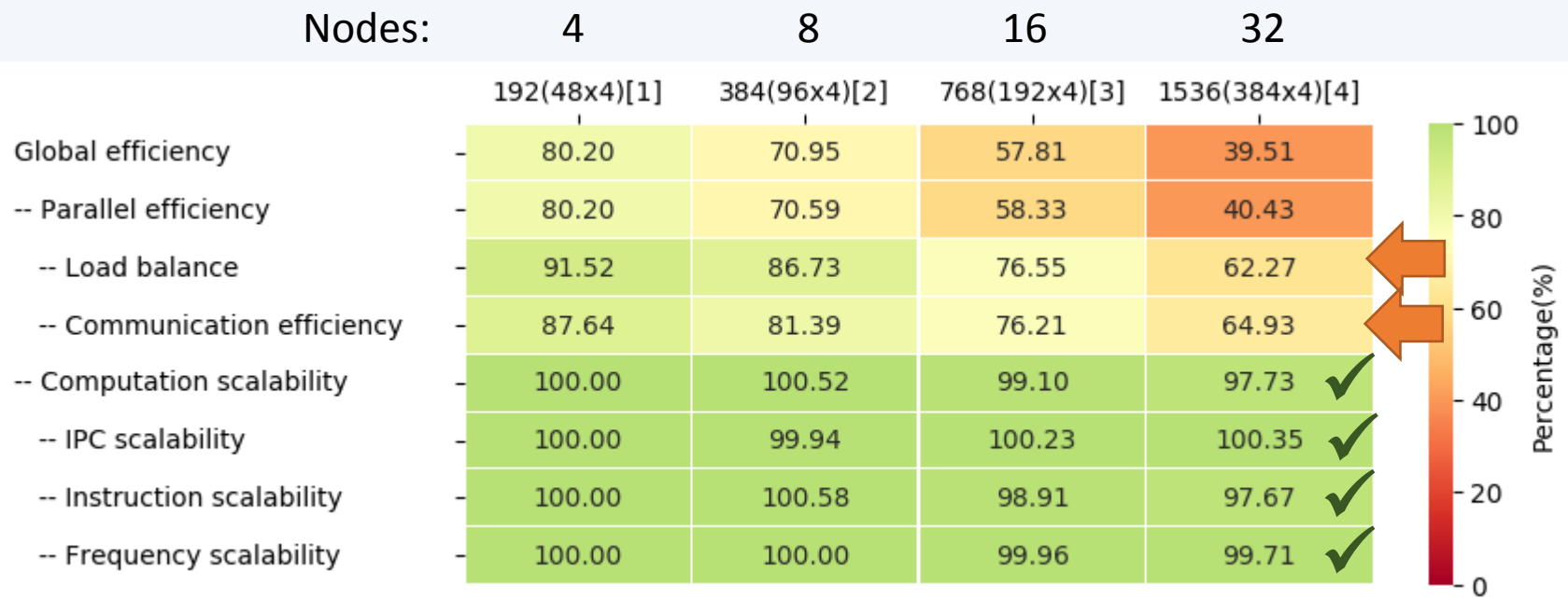
Scalability



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

Efficiency metrics

MPI scaling

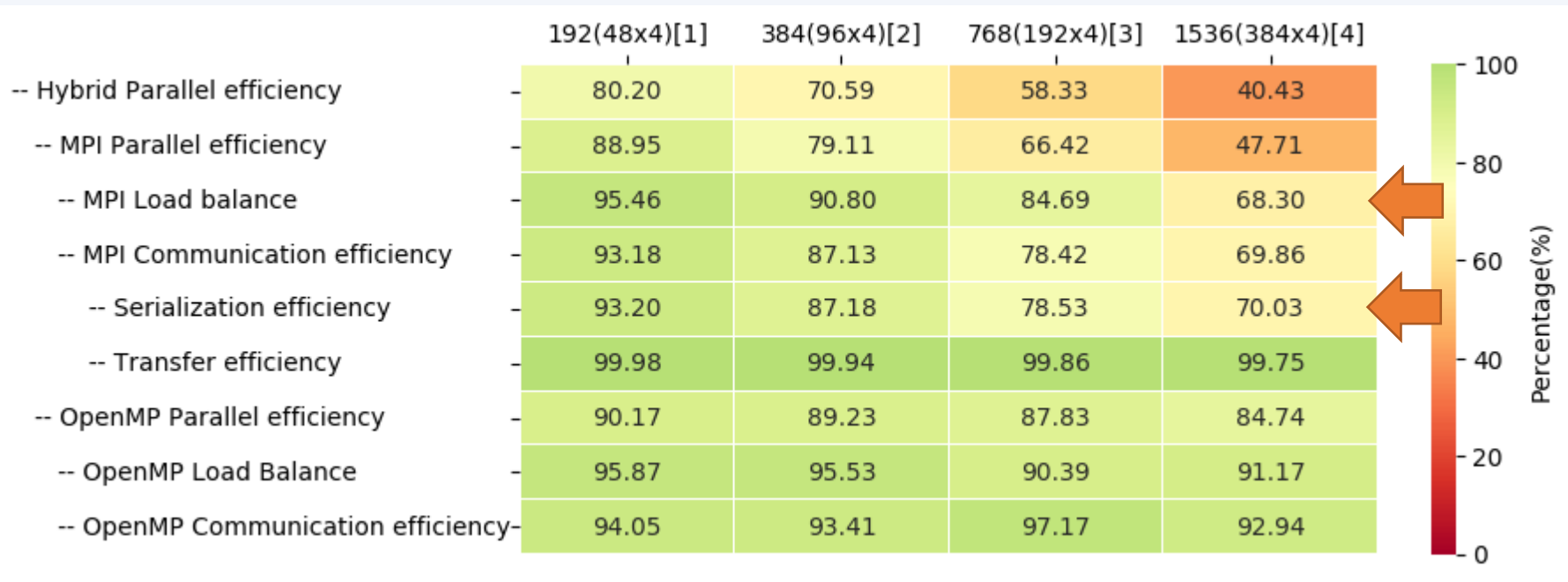


- Main scaling issues:
 - Load balance
 - Serialization
- Very good computation scalability



Hybrid Parallel Efficiency

Nodes: 4 8 16 32

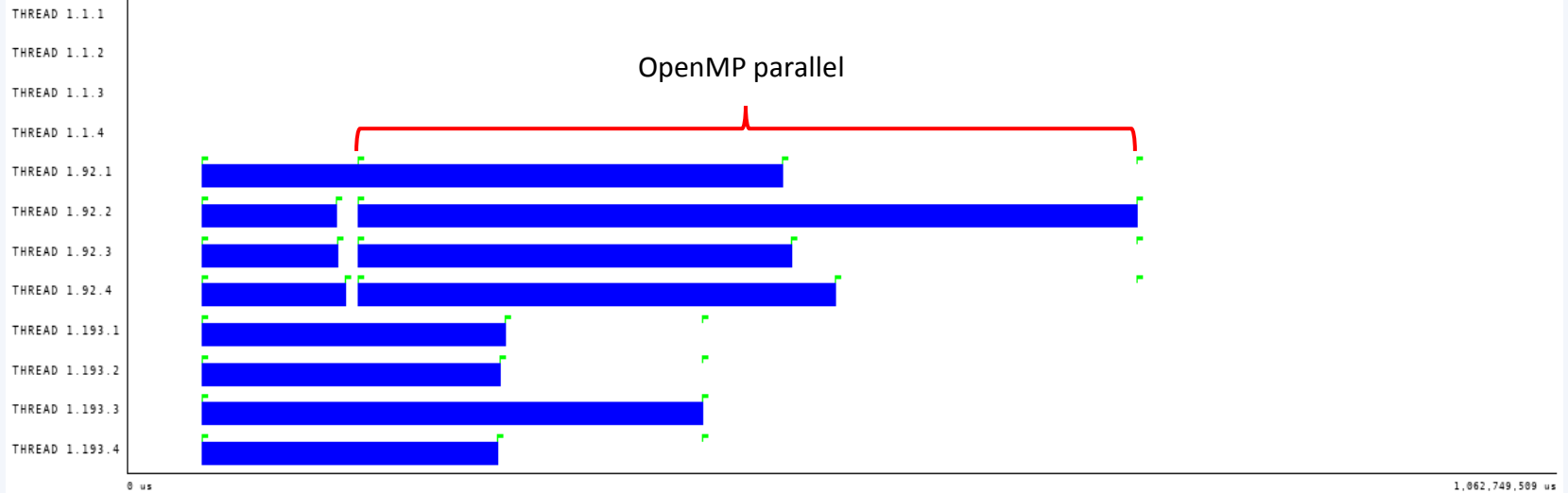


- Communication issues are due to Serialization



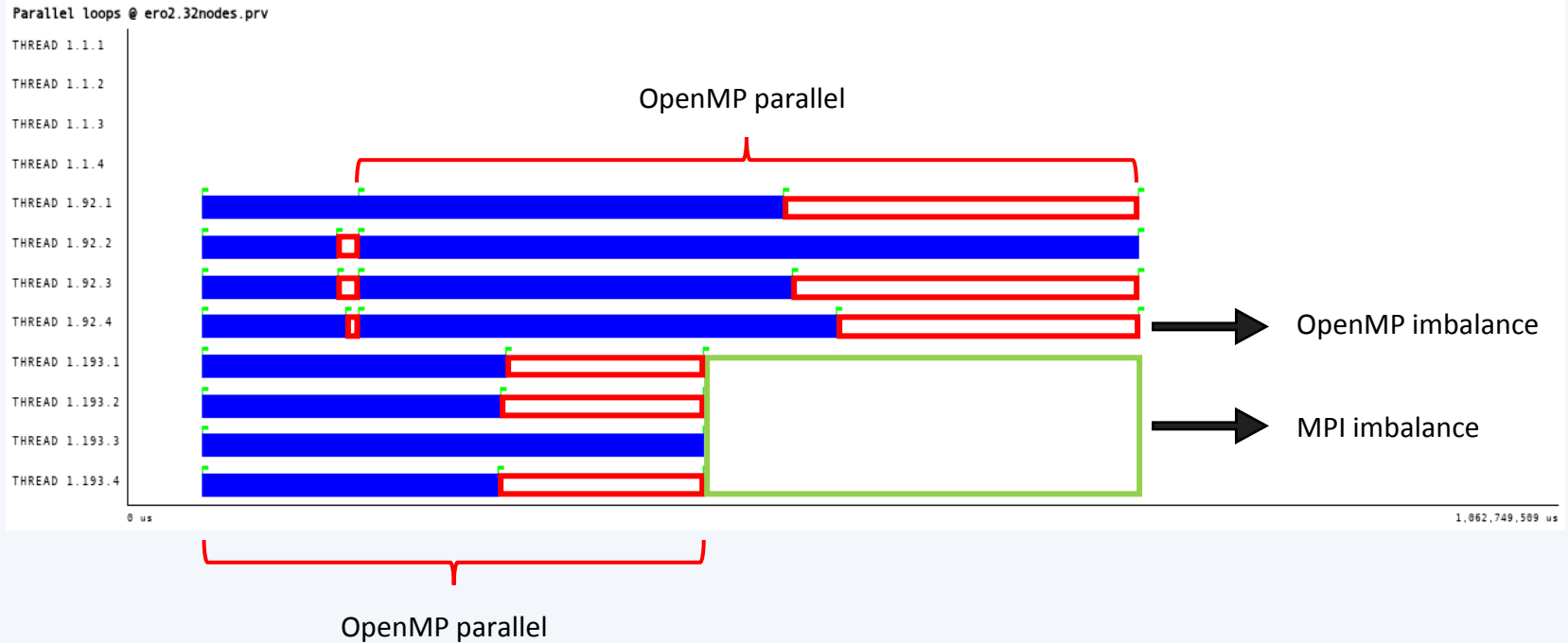
Load imbalance in detail

Parallel loops @ ero2.32nodes.prv



OpenMP parallel

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

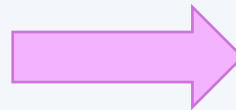
	1536(384x4)[1]	1536(384x4)[2]	
-- Hybrid Parallel efficiency	40.43	34.41	
-- MPI Parallel efficiency	47.71	54.26	
-- MPI Load balance	68.30	76.93	
-- MPI Communication efficiency	69.86	70.54	
-- Serialization efficiency	70.03	70.71	
-- Transfer efficiency	99.75	99.76	
-- OpenMP Parallel efficiency	84.74	63.42	
-- OpenMP Load Balance	91.17	75.66	
-- OpenMP Communication efficiency	92.94	83.82	
-- Load balance	62.27	58.21	

Overall Worst Load Balance



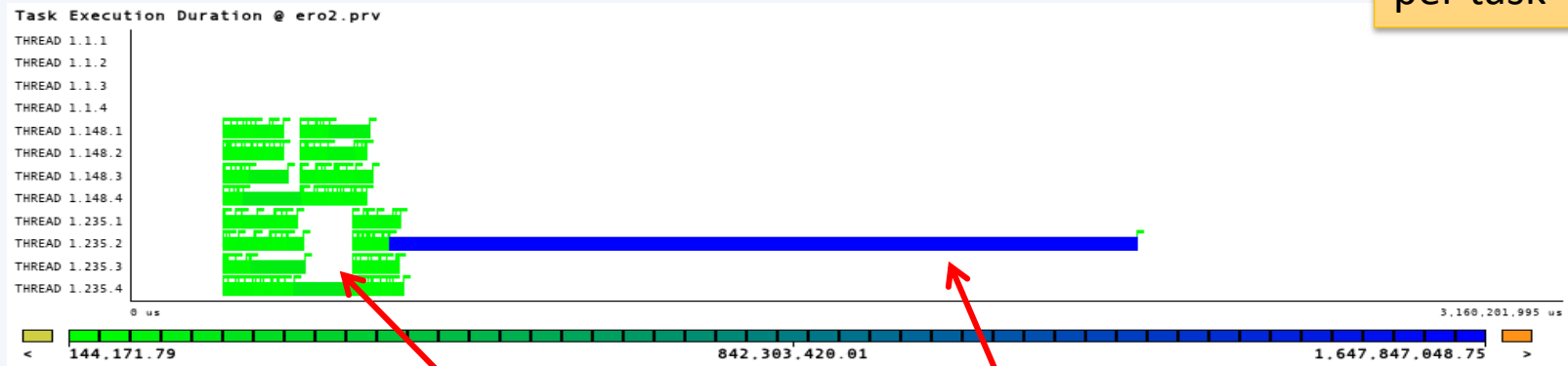
Taskification

```
#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

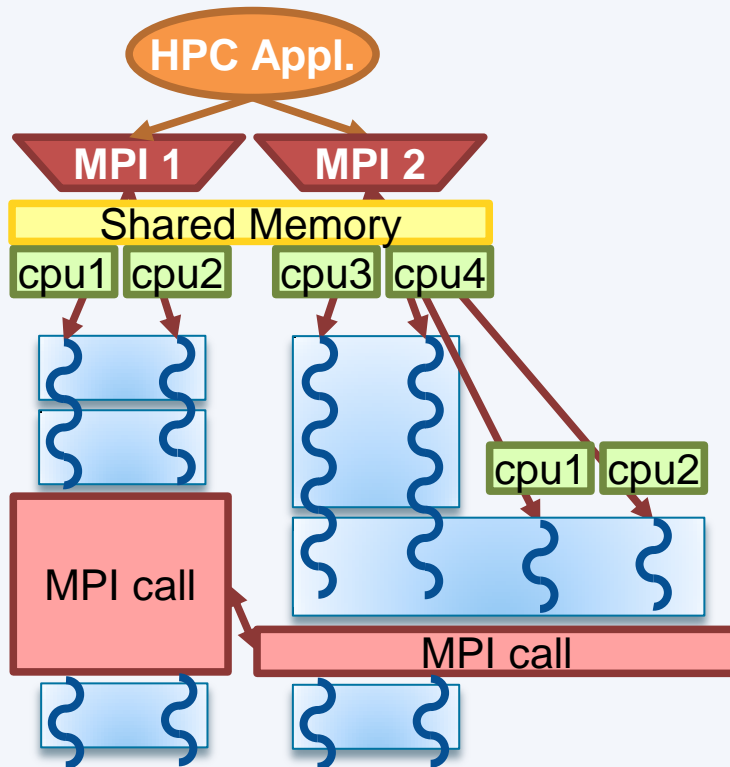


OpenMP
Imbalance

Computing
one particle



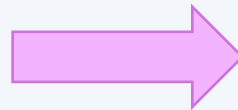
MPI Load Imbalance - DLB



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

MPI Load Imbalance - DLB

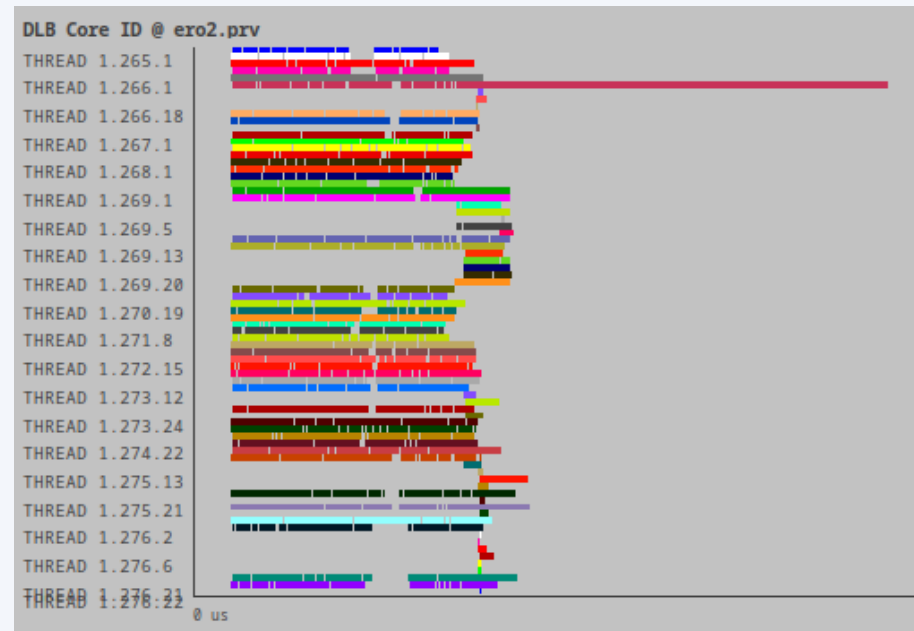
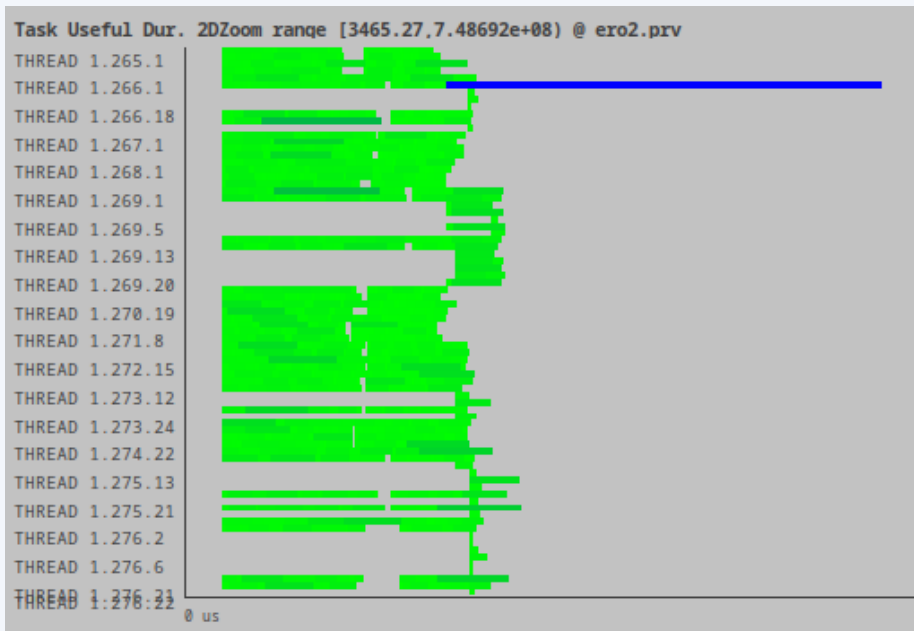
```
if (tag == QUIT) {  
    MPI_Recv(...);  
    break;  
}
```



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

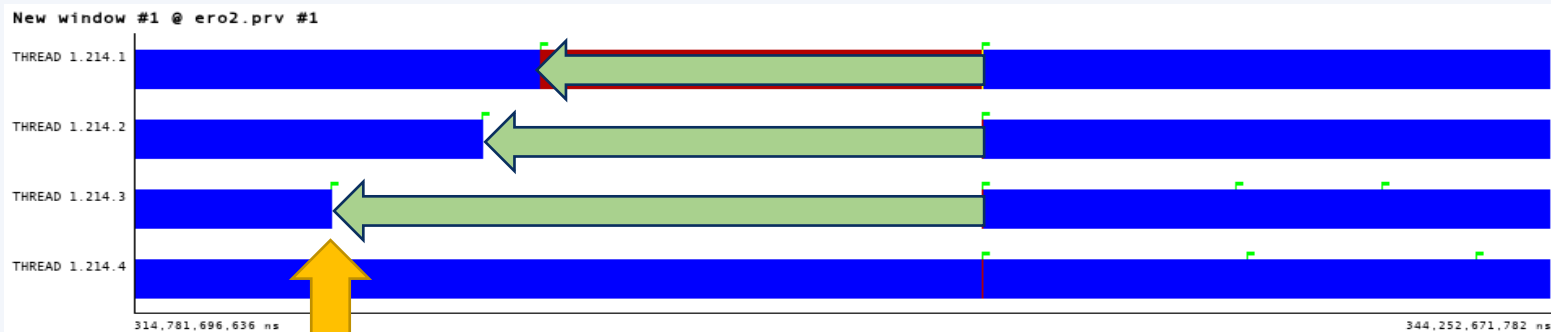
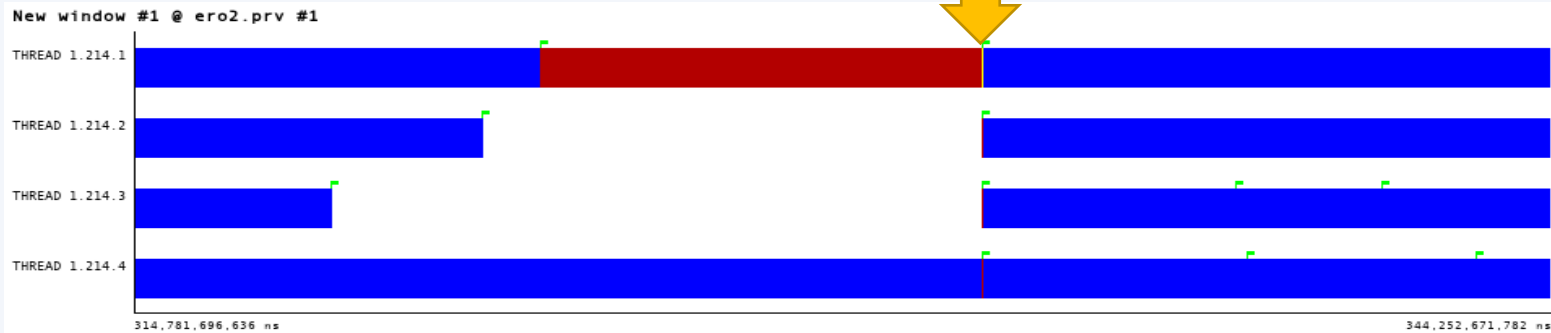
```
export DLB_ARGS="--lewi --ompt --ompt-thread-manager=free-agents --lewi-  
ompt=mpi:borrow"  
export OMP_WAIT_POLICY=passive  
export OMP_NUM_THREADS=4  
export KMP_FREE_AGENT_NUM_THREADS=20 # 4 + 20 = one socket  
preload="$OMP_HOME/lib/libomp.so"  
$DLB_HOME/bin/dlb_run --verbose env LD_PRELOAD="$LD_PRELOAD:$preload"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

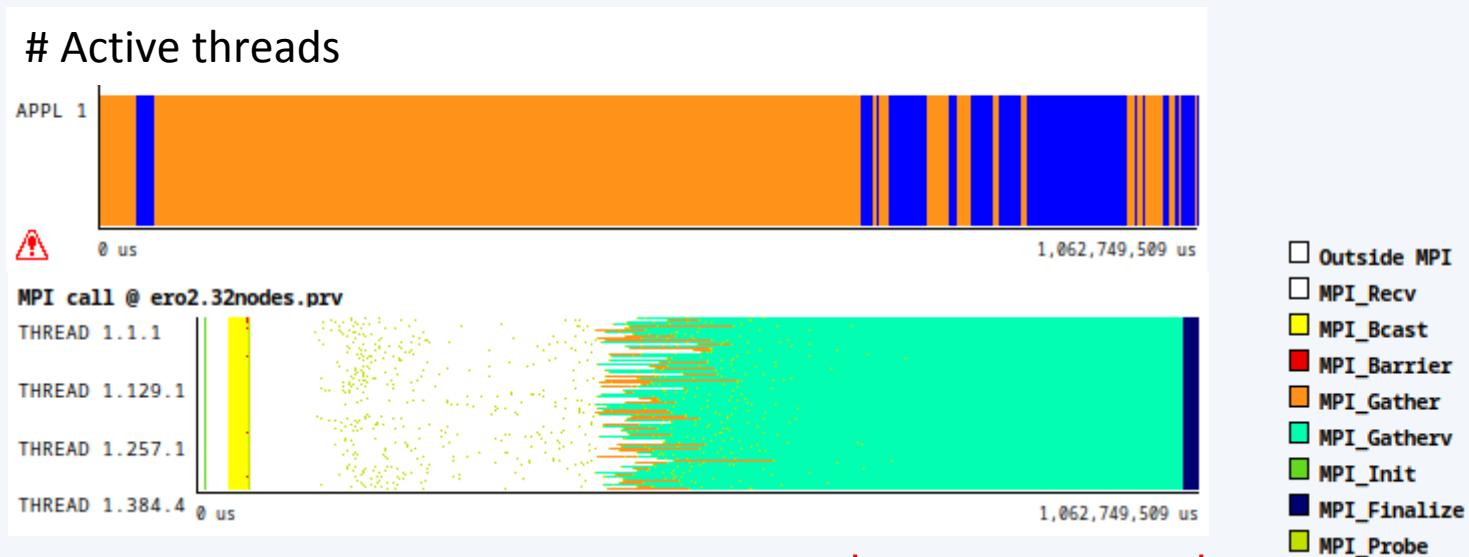


Serialization Efficiency

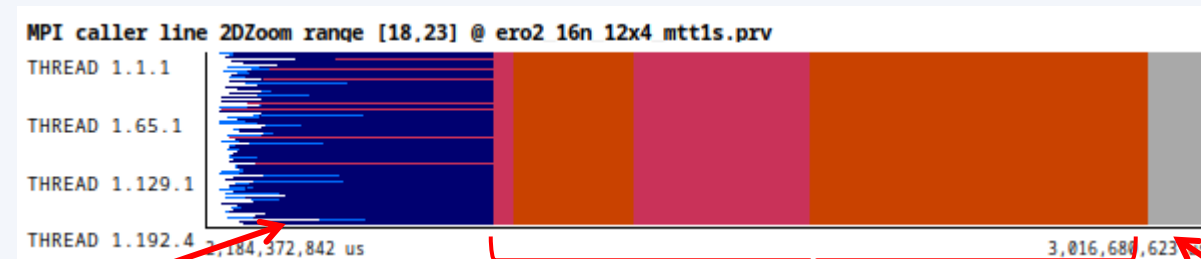
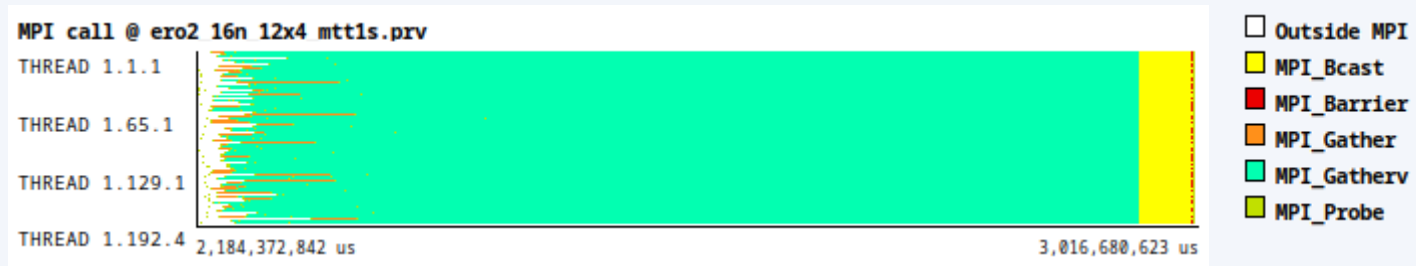
4 nodes



32 nodes



Serial functions



MPI Load Balance

Waiting particles to finish

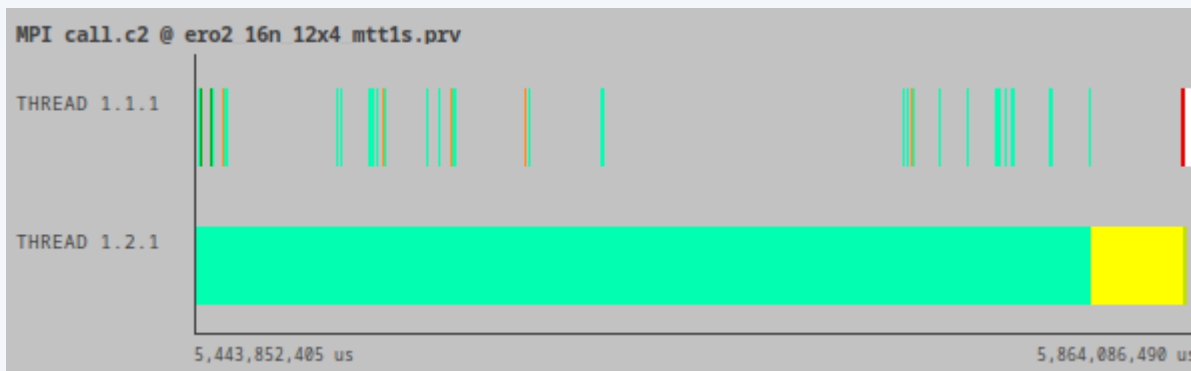
Serial functions

Work redistribution

- End
- 1868 (Ero2Simu..tion.cpp) [1868 (Ero2Simulation.cpp, ero2)]
- 2237 (Ero2Simu..tion.cpp) [2237 (Ero2Simulation.cpp, ero2)]
- 2290 (Ero2Simu..tion.cpp) [2290 (Ero2Simulation.cpp, ero2)]
- 262 (SparseData2D.h, ero2)
- 387 (SparseData3D.h, ero2)

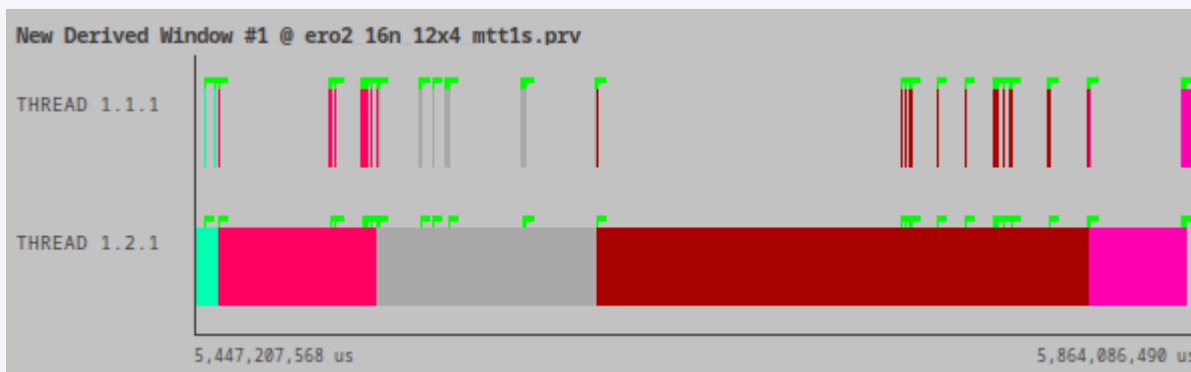


Serial functions



- Outside MPI
- MPI_Recv
- MPI_Bcast
- MPI_Barrier
- MPI_Reduce
- MPI_Gather
- MPI_Gatherv
- MPI_Probe

MPI 2nd level caller



- End
- ero2::Ero2Simulation::run [ero2::Ero2Simulation::run]
- ero2::DensityManager::gatherParticleDensities3D [ero2::DensityManager::gatherParticleDensities3D]
- ero2::DensityManager::gatherParticleDensities2D [ero2::DensityManager::gatherParticleDensities2D]
- ero2::DensityManager::gatherEmissionDensities3D [ero2::DensityManager::gatherEmissionDensities3D]
- ero2::DensityManager::gatherEmissionDensities2D [ero2::DensityManager::gatherEmissionDensities2D]



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



Thank you

marta.garcia@bsc.es

Joan.vinyals@bsc.es