

Message in a Bottle

Communication for the FMM

September 10, 2020 | Ivo Kabadshow, Mateusz Zych, Laura Morgenstern | Jülich Supercomputing Centre

Member of the Helmholtz Association



The Usual Way of Parallelization

The algorithm is sprinkled with parallelization blocks



- » MPI
- » Threading/Tasking
- » ILP/Unrolling
- » Vectorization

- » GPU-Offloading

Member of the Helmholtz Association

September 10, 2020

Slide 1



Layered Approach

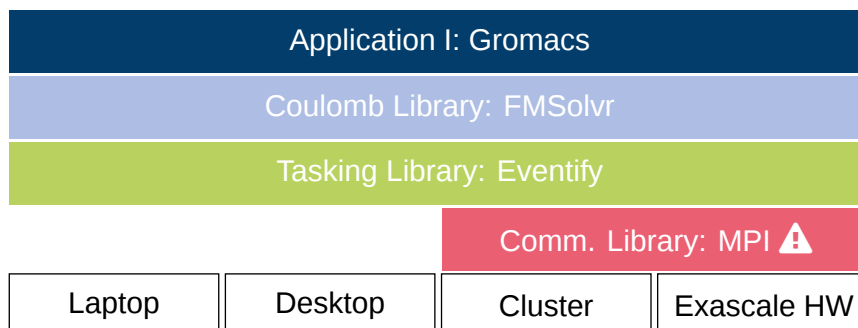
Parallelization is hidden in different layers



- » MPI
- » Threading/Tasking
- » ILP/Unrolling
- » Vectorization

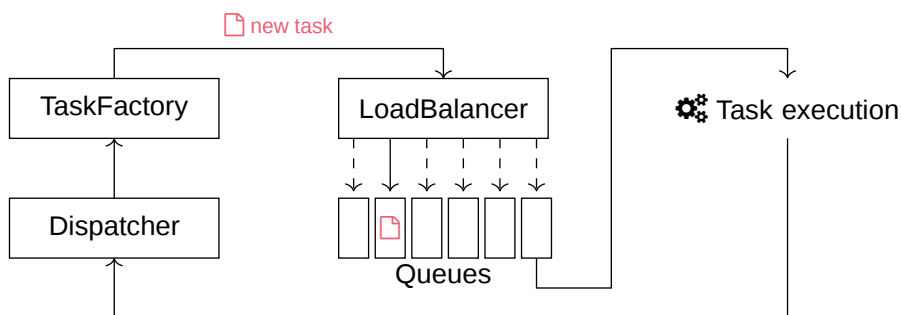
- » GPU-Offloading

Software Stack



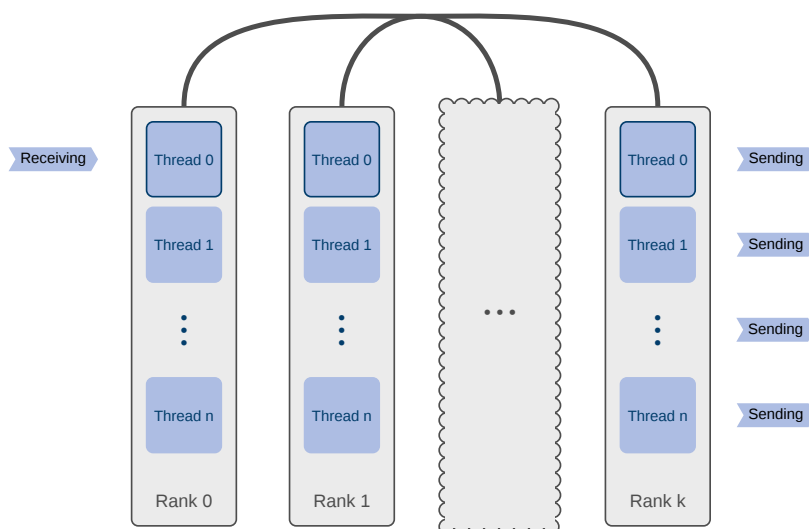
CPU Tasking Framework

Task life-cycle per thread



- Tasks can be computation or communication tasks
- Tasks can be prioritized by task type
- Only ready-to-execute tasks are stored in queue
- Workstealing from other threads is possible

Adding Inter-node Communication via MPI



Rationale: writing to data structure should not be concurrent → avoid critical sections

MPI Details

Code

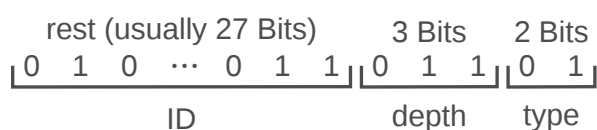
```
while (notFinished()){
  executeTask();
  Status = Communicator.Iprobe();
  if (Status.MessageNeeded()){
    Communicator.Irecv();
  } else {
    Communicator.Discard();
  }
  /*do something else */
  Communicator.Wait();
  /*use received data */
}
```

MPI Calls

- `Irecv` to all ranks
- `Iprobe` busy waiting for messages
- Call `Irecv` for messages in any case
- If message not needed, write data to dummy buffer
- Call `Wait` before using data

Distinguishing Incoming Messages

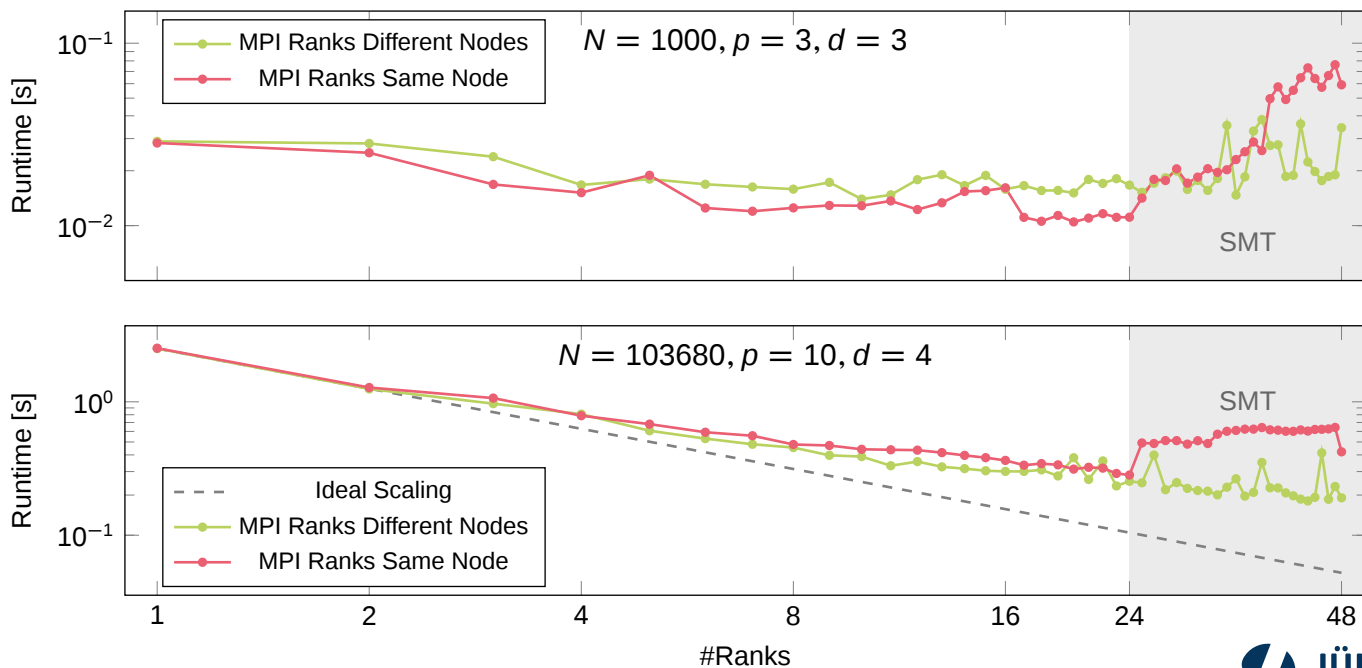
```
MPI_I[send|recv](buf,
                 count,
                 datatype,
                 [dest|source],
                 tag,
                 comm,
                 request)
```



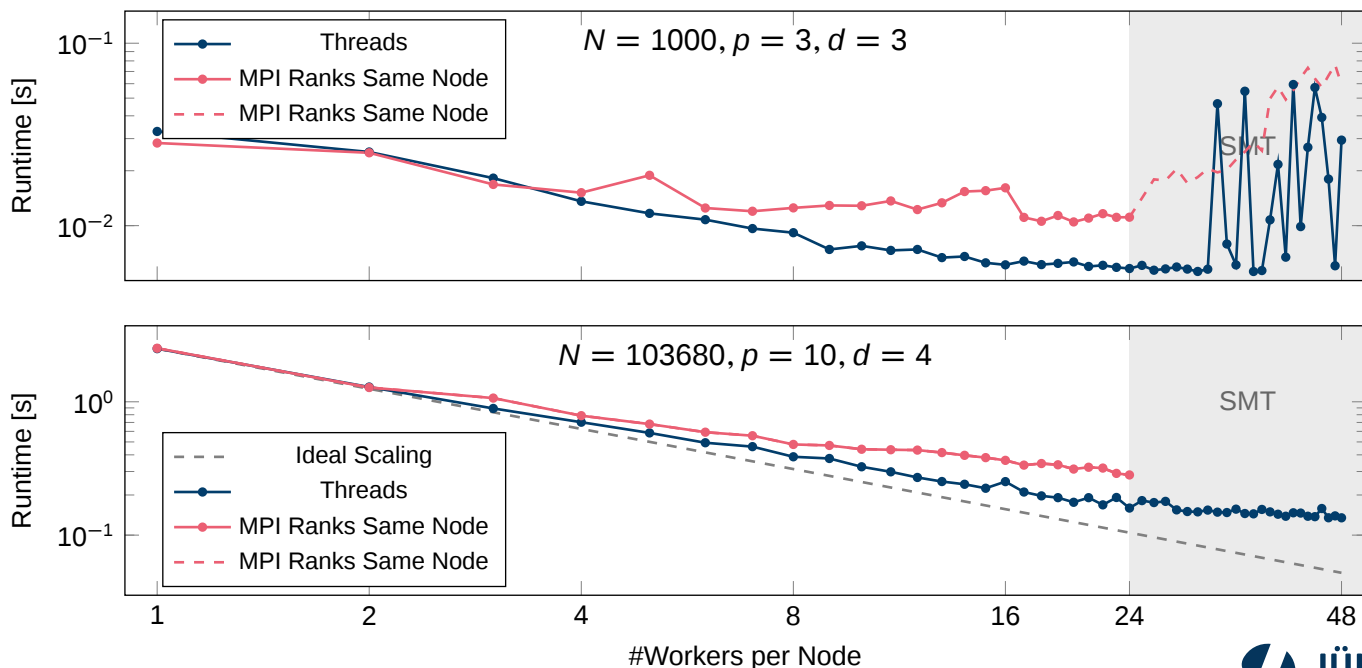
It's in the Tag

- MPI send / receive operations need a tag (integer)
- Information can be encoded in this tag
- Type of sent data (multipole, local moment, particle)
- Level of the corresponding box (2^d -level boxes)
- ID of box on this level
- This essentially mimics a matching probe / receive operation

Results from JURECA




Results from JURECA




Outlook

- Race condition in multithreaded MPI, no multithreading + MPI yet
- Handle message information more generally not via tag
- Restrict send operations to just some ranks
- Communicate data in larger chunks (defined by communication algorithm)

Questions?

 Please feel free to contact us via email if you have any questions.

 Ivo Kabadshow

 i.kabadshow@fz-juelich.de