Overlapping MPI communications with Intel TBB computation

Cassandra Rocha Barbosa, Pierre Lemarinier,
Marc Sergent, Guillaume Papauré, Marc Pérache

RADR 2020

05-18-2020

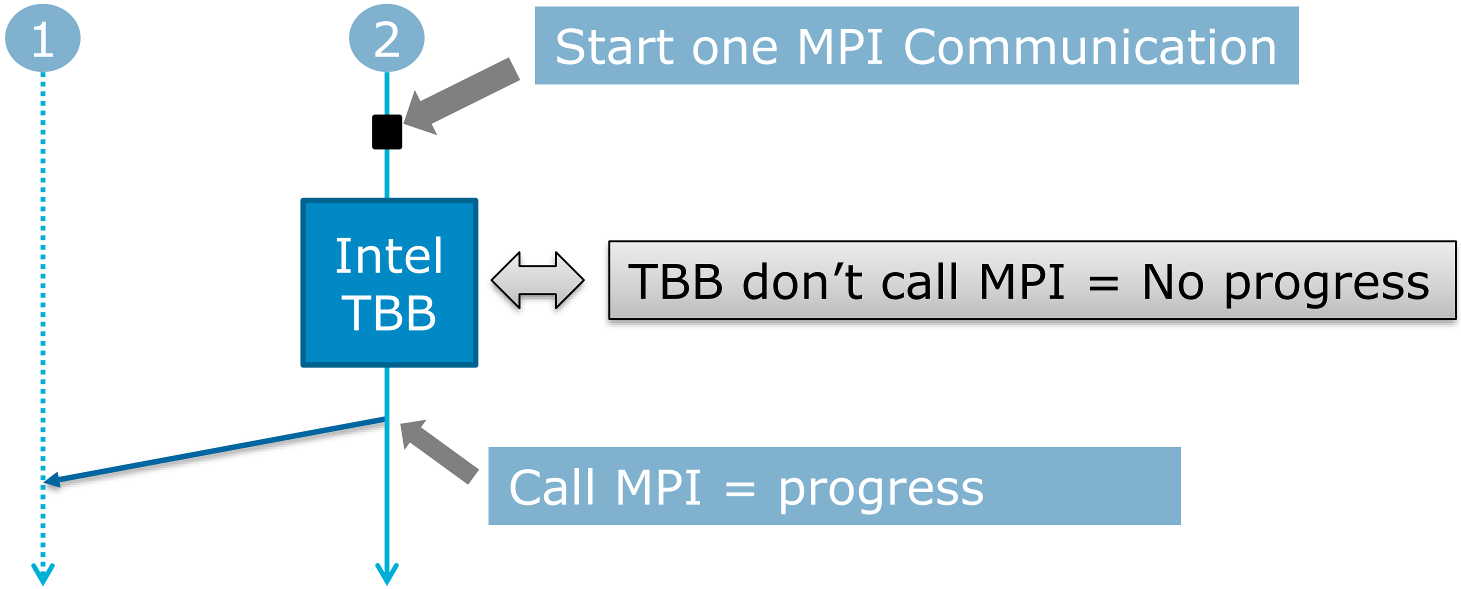
Introduction

- The MPI library is frequently used in applications to make communications between processes.

- Increasing the number of cores per node implies the use of other runtimes for managing parallelism locally. (OpenMP,Intel TBB,…)

- There are many applications called MPI+X which uses MPI and another runtime X.

Problem



Start one MPI Communication

TBB don’t call MPI = No progress

Call MPI = progress

Existing Method :

- Progress Thread
- Change hardware
- BHCO

Contributions

Specificity of Intel TBB ?

- Recursive task-based programming
- When TBB runs, its activity can be represented by a tree. Each node is an action to execute (user tasks).
- Intel TBB doesn’t use MPI library. There's no MPI communication progress.

How can we solve this problem ?

- We are looking to insert nodes whose action is to call MPI (progress task).
- We propose 3 methods to insert such tasks in the tree.

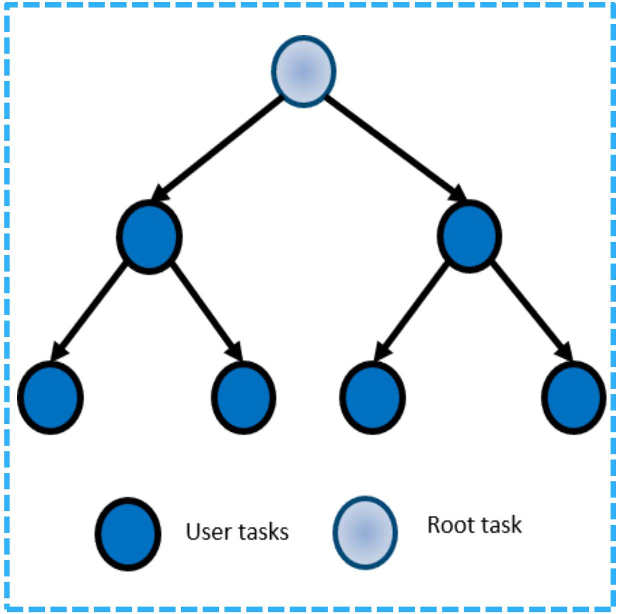


Figure 1 - example of a representation of an Intel TBB task graph.

Contributions

Root Method

How Insert progress tasks ?

1. **Local_spawn_and_root_wait :**

   local_spawn
   local_wait_for_all

2. **spawn :**

   local_spawn

   wait_for_all :

   local_wait_for_all

3. **Spawn_and_wait_for_all** :

   local_spawn
   local_wait_for_all

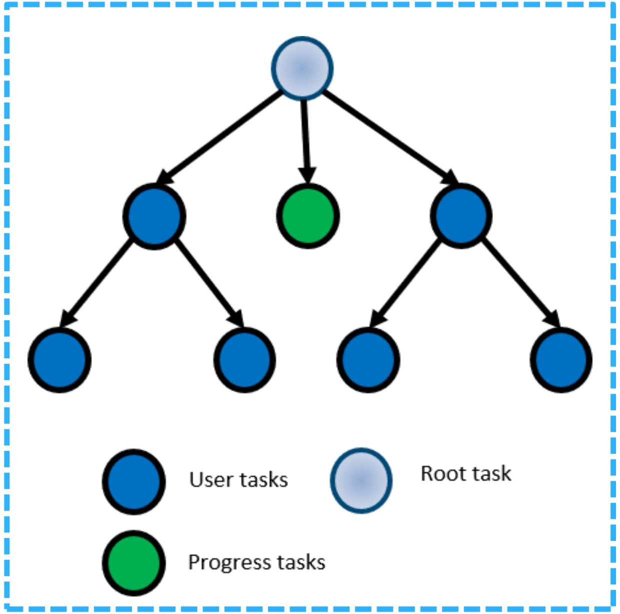


Figure 2 – example of a representation of Root method

Contributions

Non-leaves Method and Colored Method

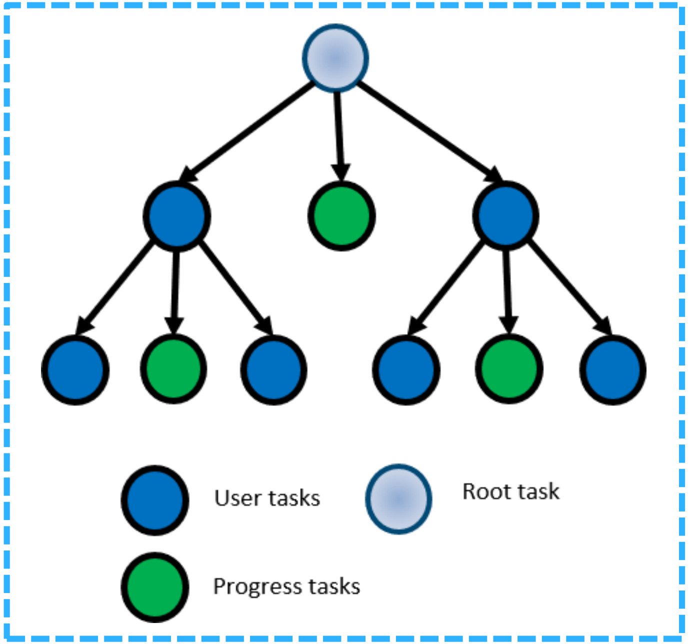


Figure 3 – example of a representation of Non-leaves method

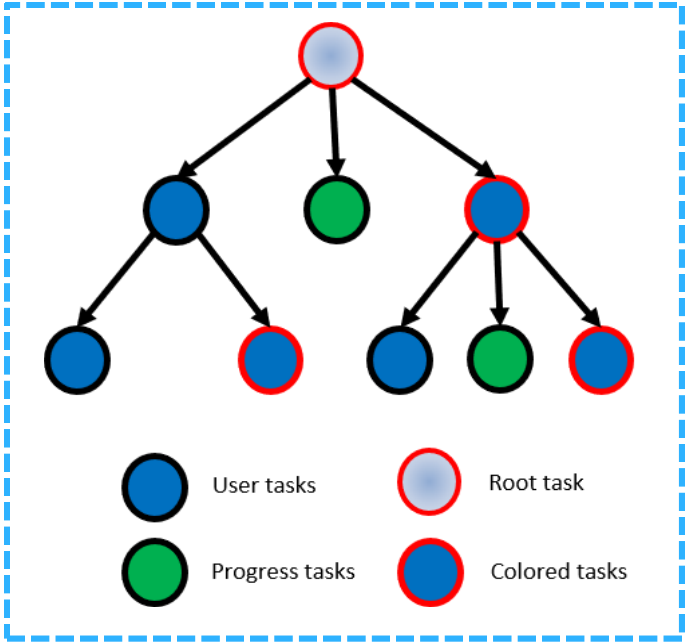


Figure 4 – example of a representation of Colored tasks method with *N=2*

Experimental results

Weak Scaling

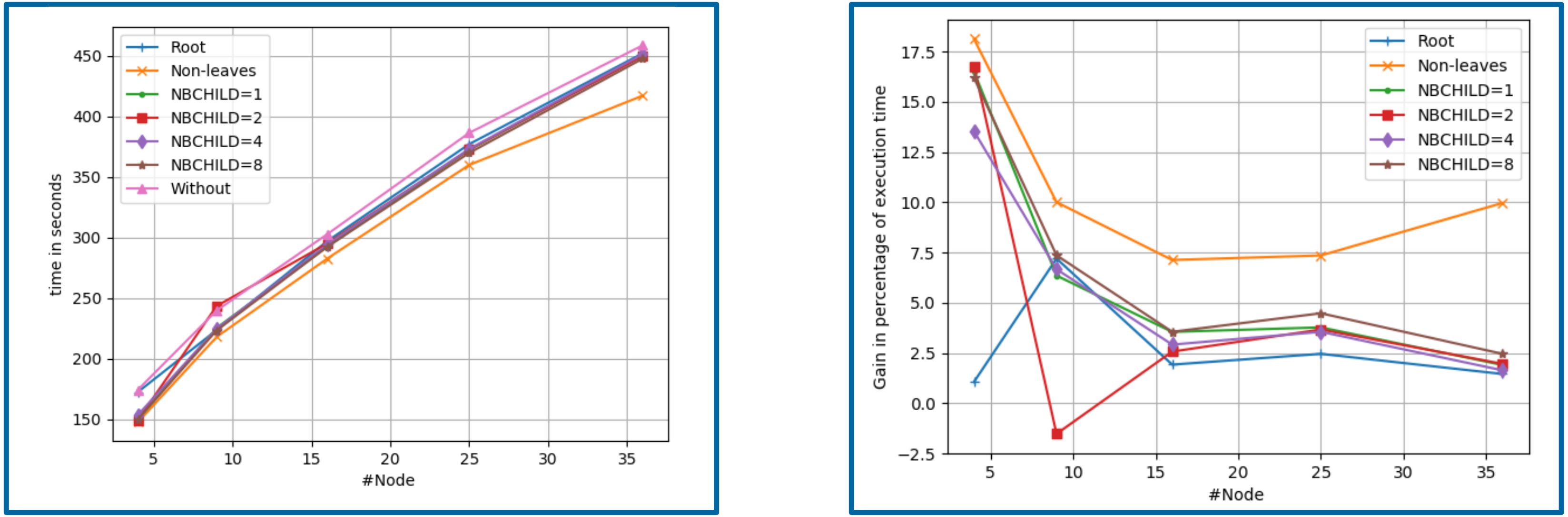


Figure 5 – Weak scaling

Experimental results

Constant number of nodes

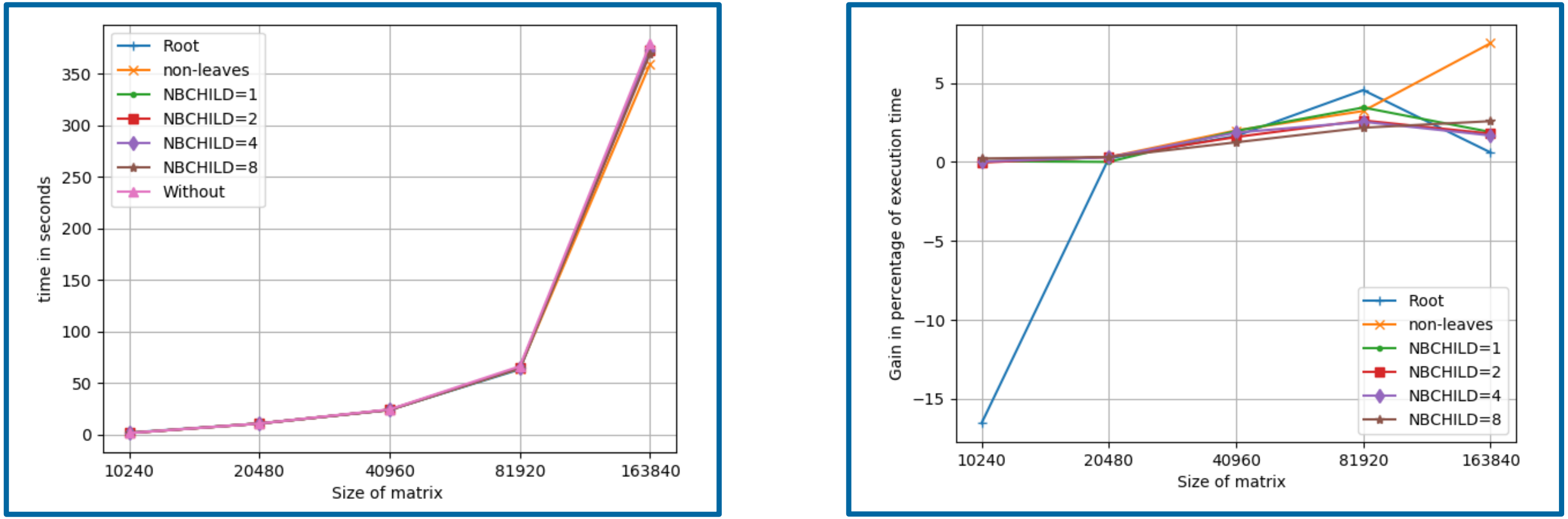


Figure 6 – Constant number of nodes scaling

Conclusion and Future Works

- Adding tasks can be profitable (up to 10% in our case).
- These methods are general enough to be adapted to other tasks based runtime (recursive or non-recursive)

- In future work, we will extend our work to deal with asynchronous progression in an MPI+X+accelerator context.

Thank you

For more information please contact:

cassandra.rochabarbosa@atos.net
pierre.lemarinier@atos.net
marc.sergent@atos.net
guillaume.papaure@atos.net
marc.perache@cea.fr