## Gromacs - Bug #1148

### switching to separate PME nodes with hybrid parallelization

02/12/2013 04:59 PM - Szilárd Páll

<table>
<thead>
<tr>
<th>Status:</th>
<th>Closed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority:</td>
<td>Low</td>
</tr>
<tr>
<td>Assignee:</td>
<td>mdrun</td>
</tr>
<tr>
<td>Category:</td>
<td>mdrun</td>
</tr>
<tr>
<td>Target version:</td>
<td></td>
</tr>
<tr>
<td>Affected version:</td>
<td>5.0</td>
</tr>
<tr>
<td>Affected version:</td>
<td>4.6.x</td>
</tr>
</tbody>
</table>

### Description

The switch to separate PME nodes happens based on the number of MPI ranks started. However, with hybrid (MPI+OpenMP) parallelization the switching should take into account the total number of resources (threads/cores) used and not only the number of ranks.

This improves performance substantially, for instance when running on 32 AMD cores with varying number of OpenMP threads per rank the improvement just from switching with 16, 8, and 4 ranks (2, 4, 8 threads/rank resp.) gives up to 15% performance improvement.

### Related issues:

- Related to Gromacs - Bug #1374: add warning that separate PME ranks are never...  Closed 10/30/2013

### History

**#1 - 04/29/2013 04:34 PM - Szilárd Páll**
- Target version changed from 4.6.1 to 4.6.2
- Affected version set to 4.6.1

**#2 - 04/30/2013 04:41 PM - Mark Abraham**

4.6.2 will go as soon as we’re done with FE fixes and PME-switch checks. Since I can see no patch yet, target 4.6.3 seems more likely to me.

**#3 - 05/22/2013 04:56 AM - Mark Abraham**
- Target version changed from 4.6.2 to 4.6.3

**#4 - 06/26/2013 12:43 AM - Mark Abraham**
- Target version changed from 4.6.3 to 4.6.x

**#5 - 06/09/2014 09:10 PM - Erik Lindahl**
- Tracker changed from Bug to Feature
- Priority changed from Normal to Low

Changing to a feature - better performance is nice, but it is not unintended behavior that causes incorrect results.

**#6 - 06/13/2014 02:34 PM - Berk Hess**

This is complicated to achieve, since the separate PME nodes decision is made before the OpenMP module is initialized (since the number of threads depends on the number of ranks). To make this work we would need partially initialize OpenMP earlier to determine how many hardware threads we have in over all nodes.

**#7 - 06/19/2014 12:10 PM - Rossen Apostolov**
- Target version deleted (4.6.x)
Erik Lindahl wrote:

Changing to a feature - better performance is nice, but it is not unintended behavior that causes incorrect results.

Based on the definition of "software bug" I know of any behavior that is incorrect or unintended counts as a bug. So far we have only marked the bugs that produce incorrect results as high or critical priority, but have not called the code that corrects obviously erroneous software behavior a "feature".

More concretely, in this case, having switched to hybrid parallelization, it is incorrect to switch to separate PME without taking into account that now we can have more than one thread per rank. So no, I don't agree, this is a bug and it is in 4.6.

Also, removing the target version without setting another one is totally unproductive and creates junk in redmine that will float around forever.

Erik Lindahl wrote:

@Szilard: We are already past the deadline we set for 5.0 patches, and given Berk's comment above it sounds highly unrealistic somebody is going to have a fix for this within 24h or so. I'll leave the target version to avoid a pissing contest, but this is obviously not a blocker.

I'm not sure how it creates less clutter to repeatedly set target versions we never hold but instead update to the next release, but that's a philosophical discussion I'll stay out of :-)

Erik Lindahl wrote:

PS: If the code in question had been a module with a proper unit test, the code would still pass every single test, since it behaves exactly as it was intended to behave when it was first written. I fully agree that the behavior you describe would be nice to have, but it does correspond to functionality that we did not have in mind when the code was written!

Erik Lindahl wrote:

I'm not sure how it creates less clutter to repeatedly set target versions we never hold but instead update to the next release, but that's a philosophical discussion I'll stay out of :-)

OK, I agree that it's too late for fixing, so it is better to not have 5.0 as target (although I would still prefer a less vague target e.g. 5.1). To facilitate fixing this, I'll look into the feasibility of calling the OpenMP initialisation early enough so that the PME rank decision can take this into account. This may not require too much code change; the number of threads is known if it is specified on the command line case in which we just need to read -ntomp and OMP_NUM_THREADS. Dealing with the case in which we automatically start threads to "fill" the machine is more tricky, though.
What I think creates clutter is setting empty or vague target versions which more or less indicates that we don't care enough about this issue to commit to addressing it. I myself am guilty at not addressing this issue earlier, too. I disagreed simply because I don't think it's a good practice to avoid commitment (by setting a vague target) nor turning an issue that describes incorrect behaviour into a feature request.

    PS: If the code in question had been a module with a proper unit test, the code would still pass every single test, since it behaves exactly as it was intended to behave when it was first written.

True, however, new functionality typically not only requires new unit tests, but also reevaluation of the assertions made by old tests. Asserting that if \#ranks < N => npme==0 is not correct anymore (from a performance point of view) - and that's the observation this report made.

#12 - 10/14/2015 06:20 PM - Mark Abraham
- Status changed from New to Closed
- Target version deleted (5.x)

5.1 mdrun does something like szilard originally wanted