Gromacs - Bug #1374
add warning that separate PME ranks are never used with GPUs

<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>Szilárd Páll</td>
</tr>
<tr>
<td>Category:</td>
<td>mdrun</td>
</tr>
<tr>
<td>Target version:</td>
<td>4.6.6</td>
</tr>
<tr>
<td>Affected version - extra info:</td>
<td>Affected version: 4.6.x</td>
</tr>
</tbody>
</table>

**Description**

With GPUs separate PME ranks are never used. This is partly because resource splitting between PP and PME ranks is too complex in GPU accelerated tunes to manage in an automated fashion. Secondly, as the CPU does PME and bonded overlapping with the GPU, offloading PME to separate ranks leaves very little computation on the CPU to overlap with the GPU. Additionally, the number if PP (or PP+PME) ranks is closely linked to the number of GPUs and therefore the user would anyway have to make sure that enough ranks are launched to accommodate PP ranks for all GPUs as well as separate PME ranks.

However, in practice, partly due to multi-threaded scaling, using separate PME ranks (which implicitly leads to less threads being used in both PP and PME) does improve performance. Most notably, on the Cray XK7 this is true already at moderate node counts.

Some related discussion can be found here: [https://gerrit.gromacs.org/#/c/2694](https://gerrit.gromacs.org/#/c/2694) (comments on patch set 15-17).

**Related issues:**

- Related to Gromacs - Bug #1148: switching to separate PME nodes with hybrid p... Closed

**Associated revisions**

**Revision 094302b3 - 06/17/2014 08:48 PM - Berk Hess**

Updated mdrun -npme documentation

The number of nodes at which PME nodes are used has increased.

Added note on PME nodes not being selected automatically with GPUs.

Fixes #1374.

Change-Id: le1de87abd3d1204d99af8b4f8e8809e7806f5c08

**Revision 31cc5ae9 - 06/05/2015 07:46 PM - Berk Hess**

Don't use PME ranks with GPUs and -npme=-1

The code disabling the automated PME rank choice with GPUs was accidentally moved after init_domain_decomposition. This caused PME ranks to be set up, but later a fatal_error occured for inconsistent PP rank and GPU counts.

Refs #1374.

Change-Id: i5f6bcc90f9ecac7f63b332b8f1acca7368b5f71bc
So there seem to be two regimes where PME nodes can help with GPUs:
1) Many physical nodes/MPI ranks. Here the user start MPI with a number of ranks and we have the number of GPUs given, so we can't really freely choose PME ranks, since #MPI-ranks==#GPUs.
2) A single node. But here we won't reach our current MPI rank limit, since we won't have more than 4 GPUs in a single node.
So although PME ranks might help, we can't automate for 1) and case 2) is very different from the current switching.

Berk Hess wrote:

So although PME ranks might help, we can't automate for 1) and case 2) is very different from the current switching.

I agree. Note that i ma not suggesting to automate the switching, but to warn the user that separate PME ranks may help, but he/she has to set it up manually. Some criteria on when to trigger the warning would be nice (based on the total number of cores/hardware threads used?) to avoid issuing spurious notes. However, as noted in the gerrit discussion, to be able to do this we'd need to solve #1148 first.

Warning about a limitation of the implementation is definitely not a feature.

Gerrit received a related patchset '1' for Issue #1374.
Uploader: Berk Hess (hess@kth.se)
Change-Id: le1de87abd3d1204d99a18b4f8e6809e780f5c08
Gerrit URL: https://gerrit.gromacs.org/3590

I don't think this requires a warning. I updated mdrun -h with a note on PME nodes not being selected automatically with GPUs. Since now mdrun does what it says, a warning is not necessary.
#8 - 06/17/2014 10:45 PM - Berk Hess
- Status changed from Fix uploaded to Resolved
- % Done changed from 0 to 100

Applied in changeset 094302b350cfce4e5ec755aeb069c72ac652a724e.

#9 - 06/18/2014 01:10 PM - Mark Abraham
- Status changed from Resolved to Closed

#10 - 06/19/2014 09:30 PM - Szilárd Páll
I missed the part that the recently merged change "fixes" this.

Most users expect that mdrun or mpiun -np N mdrun_mpi just works optimally. Don't you think that the user should be explicitly warned, if nothing else, at the rank count where the switching would automatically happen in CPU-only runs (unless #1148 gets fixed) that he/she should really consider separate PME ranks. On all machines where I ran so far, using separate PME ranks was always faster above 4-8 sockets.

#11 - 05/18/2015 03:05 AM - Szilárd Páll
Actually, this is not true, the fixing commit added an incorrect statement to the docs:

```
$ $mdrun -version 2>&1 | grep ' VERSION'
Gromacs version: VERSION 4.6.6-dev-20140522-d77dddb

$mdrun -ntmpi 32 -ntomp 1 -gpu_id $(streep 0 16)$(streep 1 16) -s ../topol.tpr
[...]
Will use 24 particle-particle and 8 PME only nodes
This is a guess, check the performance at the end of the log file
Using 32 MPI threads
Using 1 OpenMP thread per tMPI thread
Compiled acceleration: AVX_256 (Gromacs could use AVX_128_FMA on this machine, which is better)

2 GPUs detected:
#0: NVIDIA GeForce GTX TITAN, compute cap.: 3.5, ECC: no, stat: compatible
#1: NVIDIA GeForce GTX TITAN, compute cap.: 3.5, ECC: no, stat: compatible

2 GPUs user-selected for this run.
Mapping of GPUs to the 24 PP ranks in this node: #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #0, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1, #1

-------------------------------------------------------
Program mdrun, VERSION 4.6.8-dev-20150212-c060264
Source code file: /nethome/pszilard/projects/gromacs/gromacs-4.6/src/gmxlib/gmx_detect_hardware.c, line: 380
Fatal error:
Incorrect launch configuration: mismatching number of PP thread-MPI threads and GPUs.
mdrun was started with 24 PP thread-MPI threads, but you provided 32 GPUs.
For more information and tips for troubleshooting, please check the GROMACS website at http://www.gromacs.org/Documentation/Errors
```

11/19/2015
Gerrit received a related patchset '1' for Issue #1374.
Uploader: Berk Hess (hess@kth.se)
Change-Id: i5f6bcc90fecd763b332b8f1acca7368b5f71bc
Gerrit URL: https://gerrit.gromacs.org/4619