### GROMACS - Bug #1613

**Problem with 4.6.x MPI, thread affinity, slurm and node-uneven task spread**

10/02/2014 04:45 PM - Åke Sandgren

<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>5.0.5</td>
</tr>
<tr>
<td>Affected version:</td>
<td>4.6.x</td>
</tr>
<tr>
<td>Difficulty:</td>
<td>uncategorized</td>
</tr>
</tbody>
</table>

**Description**

Just managed to pin down a weird problem which is caused by uneven spread of tasks over nodes and thread affinity causing jobs to hang in `gmx_set_thread_affinity`.

This happens on our 48-core nodes using a 100 task job that when submitted through slurm (without specifying distribution manually) gets distributed over 3 nodes with 6+47+47 tasks.

We are also using cgroups to allow for multiple jobs per node, so the node with 6 tasks has an affinity mask set for only the 6 cores on a single NUMA. The nodes with 47 tasks have the whole node allocated and thus gets a full 48-core affinity mask.

(Actually due to a bug(/feature?) in slurm the tasks on the node with only 6 cores allocated actually get a single-core per task affinity, but that's not relevant here.)

Anyway, when the code gets to line 1629 in runner.c (this is 4.6.7) and the call to `gmx_check_thread_affinity_set` we start having problems.

The loop to set `bAllSet` ends up setting `bAllSet` to TRUE for the tasks on the two fully allocated nodes and FALSE on the tasks on the third node. This in turn changes `hw_opt->thread_affinity` to `threadaffOFF` on those 6 tasks, but leaves it at `threadaffAUTO` for the other 2x47 tasks.

`gmx_set_thread_affinity` then promptly returns for those poor 6 tasks and tries in vain to do a `MPI_Comm_split` with 6 tasks missing from the equation...

I suggest to gather the `bAllSet` result from all nodes in `gmx_check_thread_affinity_set` and make sure all tasks have the same view of the world...

**Suggested patch:**

```c
#include <mpi.h>

gmx_bool bAllSet_All;

int main(int argc, char **argv)
{
    MPI_Init(argc, argv);
    gmx_bool bAllSet = gmx_set_thread_affinity(0);
    bAllSet = true;
    MPI_Barrier(MPI_COMM_WORLD);
    MPI_Allreduce(bAllSet, &bAllSet_All, 1, MPI_INT, MPI_LAND, MPI_COMM_WORLD);
    bAllSet = bAllSet_All;
    MPI_Finalize();
}
```

**Related issues:**

Related to GROMACS - Bug #1614: thread-MPI has broken support for operations ... **Rejected**

**Associated revisions**

Revision 93a5a180 - 11/26/2014 11:22 AM - Åke Sandgren

Fix problem with mixed affinity mask on different nodes.

If task distribution (with slurm for instance) causes both fully allocated and not-fully allocated nodes to be assigned to the job then there may be tasks with an all-cores affinity mask and tasks with a not-all-cores affinity masks.

Fixes #1613

Change-Id: I71c0daa43a5dd42da57bfdf09037806ce1d9334b5

**History**

#1 - 10/02/2014 04:51 PM - Szilárd Páll

05/02/2020
Your suggested patch seems reasonable! FYI our code review system is open for anyone to submit patches, so you could upload the suggested three-liner it straight to gerrit.gromacs.org. (the 3 steps instructions are here: http://www.gromacs.org/Developer_Zone/Git/Gerrit#Getting_started)

#2 - 10/02/2014 05:22 PM - Åke Sandgren

Szilárd Páll wrote:

Your suggested patch seems reasonable! FYI our code review system is open for anyone to submit patches, so you could upload the suggested three-liner it straight to gerrit.gromacs.org. (the 3 steps instructions are here: http://www.gromacs.org/Developer_Zone/Git/Gerrit#Getting_started)

Ok, trying that then :-)

#3 - 10/02/2014 05:39 PM - Gerrit Code Review Bot

Gerrit received a related patchset '1' for Issue #1613.
Uploader: Åke Sandgren (ake.sandgren@hpc2n.umu.se)
Change-Id: I71c0daa43a5dd42da57bfd09037806ca1d9334b5
Gerrit URL: https://gerrit.gromacs.org/4116

#4 - 10/02/2014 05:40 PM - Åke Sandgren

Åke Sandgren wrote:

Szilárd Páll wrote:

Your suggested patch seems reasonable! FYI our code review system is open for anyone to submit patches, so you could upload the suggested three-liner it straight to gerrit.gromacs.org. (the 3 steps instructions are here: http://www.gromacs.org/Developer_Zone/Git/Gerrit#Getting_started)

Ok, trying that then :-)\n
I hope i got it right. Can you check?

#5 - 10/02/2014 10:28 PM - Mark Abraham

- Related to Bug #1614: thread-MPI has broken support for operations on MPI_INT added

#6 - 04/20/2015 02:23 PM - Mark Abraham

- Category set to mdrun
- Status changed from New to Resolved

#7 - 04/20/2015 02:41 PM - Mark Abraham

- Status changed from Resolved to Closed
- Target version set to 5.0.5