Process placement

2020

Trusted partner for your Digital Journey

Process placement

- Crucial to get right
  - MPI processes must be pinned to get a good performance
  - If it is not correct, the program is often several times slower

- SLURM and Intel MPI have good defaults
  - Choosing the number of MPI and OpenMP correctly is often good enough

What type of application?

- Pure MPI
  - Well-balanced code, scales fine.
  - Unbalanced code, usually multi-process codes (e.g. climate).

- Hybrid MPI-OpenMP
  - hyperthreading

- Options to verify process placement

Well-balanced MPI code

- Just use the default distribution, whether SLURM or Intel‵s mpirun
  - Each task on one core
  - Fill up the nodes core by core

Unbalanced code

- These codes often have a lot of MPI time (close to 50%)
- MPI time is indicative of load imbalance. No real communication of data, but waiting for comunication to start from the other side.
- As MPI time is mostly waiting time, the redistribution of MPI processes has little impact on MPI time.

- Strategy to distribute the tasks differently across and within nodes could help to move load from one NUMA domain across multiple NUMA domains.

- „Trial & error"

Unbalanced code
E.g. a job on 3 nodes

- srun -m cyclic
  - Rank 0 on node 0, rank 1 on node 1
  - Rank 3 on node 0, …
  - E.g. could benefit OCTOPUS

- srun -m plane=12 (try also 8,6,4,3,2)
  - Ranks 0-11 on node 0, ranks 12-23 on node 1
  - Ranks 36-47 on node 0

Unbalanced code
Process placement in a node

- srun --cpu-bind=map_cpu:0,1,2,3,24,25,26,27,48,49,50,51,72,73,74,75,4,5,…
  - Distribute the MPI tasks across NUMA domains within a node
  - Can be combined with the plane distribution from the last slide
  - E.g. could benefit fesom2

Hybrid MPI-OpenMP code

- Use at least 4 MPI processes on a CLX-AP node, 1 on each NUMA domain.
- OpenMP threads inside a NUMA domain
- Try what nr. of OpenMP threads works best.
- 96 cores gives you a lot of possibilities.

MPI/node	OpenMP	MPI*OpenMP
4	24	96
8	12	96
12	8	96
16	6	96
24	4	96
32	3	96
48	2	96

Hybrid MPI-OpenMP code hyperthreading

- „The proof of the pudding is in the eating‟

- Try with hyperthreads, use export OMP_WAIT_POLICY=passive
  - srun/sbatch: -c flag is equal to OMP_NUM_THREADS
  - mpirun distributes processes evenly.

- Try without hyperthreads
  - srun/sbatch: -c flag is 2x OMP_NUM_THREADS
  - mpirun distributes processes evenly.

Options to verify placement

- **`export I_MPI_DEBUG=4`**
  - Intel MPI prints the affinity and node for each process
- **`export KMP_AFFINITY=verbose`**
  - Intel OpenMP runtime prints the affinity for each thread

- `srun` **`--cpu-bind=verbose`** `app`
  - print the affinity of all processes
- **`srun -l hwloc-bind --get app`**
  - it prints the affinity of all processes, independent from srun

Terminology: bit masks

- --cpu-bind=verbose prints bit masks
  - hwloc-bind --get --pid as well
  - more tools use masks

- read from right to left
- each hexadecimal digit represents 4 logical CPUs, e.g.
  - 0x01 is a mask where the first logical CPU is on
  - 0x02 : second logical CPU is on
  - 0x03 : first and second logical CPU are on
  - 0xF0 : fifth to eighth logical CPUs are on

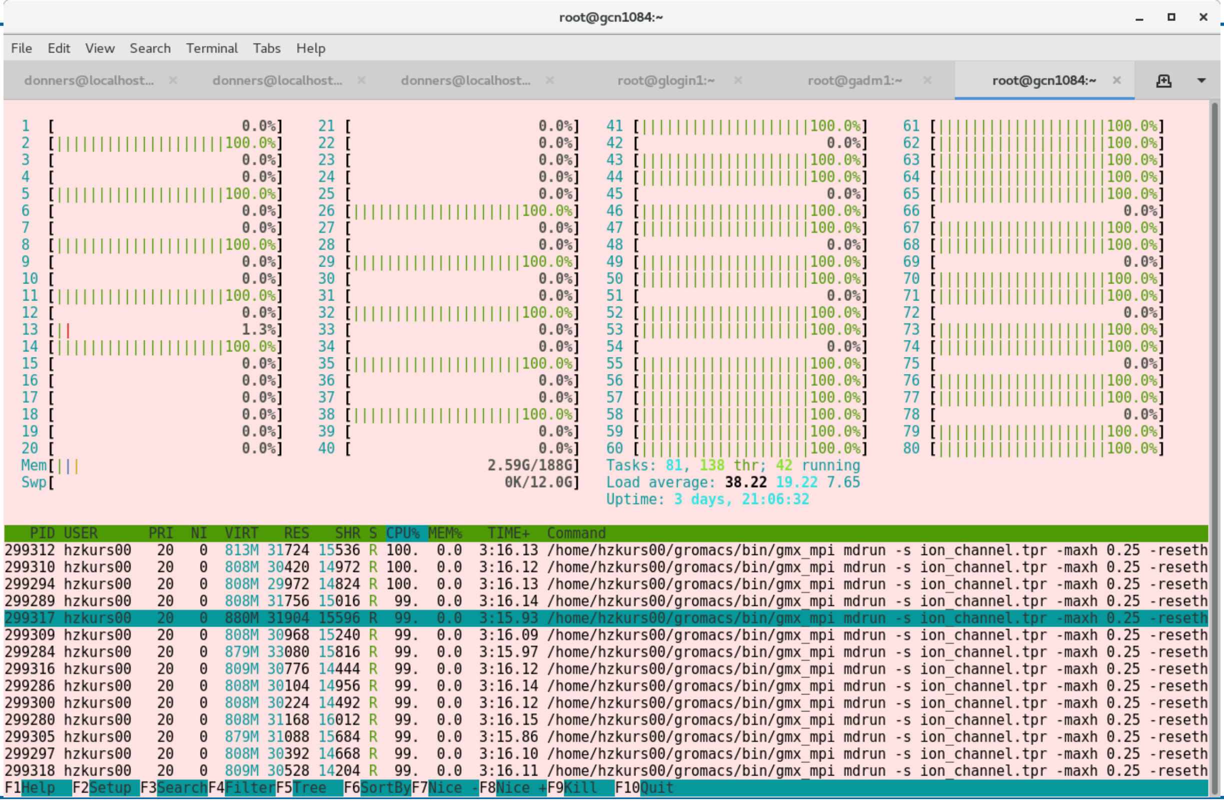
Options to verify placement (interactive)

- Login on compute node, then run `htop`

- All cores should be busy (green in htop)
  - Note that a core is shown as 2 logical cpus: core 1 is CPUs 0 & 96 in htop
  - Process should be busy on only one of the two logical CPUs.
  - However, if hyperthreading is used, both logical CPUs should be busy.

- Very little system time (red in htop)
  - A lot of system time usually points to a problem
  - Maybe I/O or task switching.

htop



```
  1  [                     0.0%]   21 [                     0.0%]   41 [|||||||||||||||||||100.0%]   61 [|||||||||||||||||||100.0%]
  2  [|||||||||||||||||||100.0%]   22 [                     0.0%]   42 [                     0.0%]   62 [|||||||||||||||||||100.0%]
  3  [                     0.0%]   23 [                     0.0%]   43 [|||||||||||||||||||100.0%]   63 [|||||||||||||||||||100.0%]
  4  [                     0.0%]   24 [                     0.0%]   44 [|||||||||||||||||||100.0%]   64 [|||||||||||||||||||100.0%]
  5  [|||||||||||||||||||100.0%]   25 [                     0.0%]   45 [                     0.0%]   65 [|||||||||||||||||||100.0%]
  6  [                     0.0%]   26 [|||||||||||||||||||100.0%]   46 [|||||||||||||||||||100.0%]   66 [                     0.0%]
  7  [                     0.0%]   27 [                     0.0%]   47 [|||||||||||||||||||100.0%]   67 [|||||||||||||||||||100.0%]
  8  [|||||||||||||||||||100.0%]   28 [                     0.0%]   48 [                     0.0%]   68 [|||||||||||||||||||100.0%]
  9  [                     0.0%]   29 [|||||||||||||||||||100.0%]   49 [|||||||||||||||||||100.0%]   69 [                     0.0%]
  10 [                     0.0%]   30 [                     0.0%]   50 [|||||||||||||||||||100.0%]   70 [|||||||||||||||||||100.0%]
  11 [|||||||||||||||||||100.0%]   31 [                     0.0%]   51 [                     0.0%]   71 [|||||||||||||||||||100.0%]
  12 [                     0.0%]   32 [|||||||||||||||||||100.0%]   52 [|||||||||||||||||||100.0%]   72 [                     0.0%]
  13 [||                   1.3%]   33 [                     0.0%]   53 [|||||||||||||||||||100.0%]   73 [|||||||||||||||||||100.0%]
  14 [|||||||||||||||||||100.0%]   34 [                     0.0%]   54 [                     0.0%]   74 [|||||||||||||||||||100.0%]
  15 [                     0.0%]   35 [|||||||||||||||||||100.0%]   55 [|||||||||||||||||||100.0%]   75 [                     0.0%]
  16 [                     0.0%]   36 [                     0.0%]   56 [|||||||||||||||||||100.0%]   76 [|||||||||||||||||||100.0%]
  17 [                     0.0%]   37 [                     0.0%]   57 [|||||||||||||||||||100.0%]   77 [|||||||||||||||||||100.0%]
  18 [                     0.0%]   38 [|||||||||||||||||||100.0%]   58 [|||||||||||||||||||100.0%]   78 [                     0.0%]
  19 [                     0.0%]   39 [                     0.0%]   59 [|||||||||||||||||||100.0%]   79 [|||||||||||||||||||100.0%]
  20 [                     0.0%]   40 [                     0.0%]   60 [|||||||||||||||||||100.0%]   80 [|||||||||||||||||||100.0%]
  Mem[|||                                               2.59G/188G]   Tasks: 81, 138 thr; 42 running
  Swp[                                                   0K/12.0G]   Load average: 38.22 19.22 7.65
                                                                     Uptime: 3 days, 21:06:32

   PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
299312 hzkurs00   20   0  813M 31724 15536 R 100.  0.0  3:16.13 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299310 hzkurs00   20   0  808M 30420 14972 R 100.  0.0  3:16.12 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299294 hzkurs00   20   0  808M 29972 14824 R 100.  0.0  3:16.13 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299289 hzkurs00   20   0  808M 31756 15016 R  99.  0.0  3:16.14 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299317 hzkurs00   20   0  880M 31904 15596 R  99.  0.0  3:15.93 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299309 hzkurs00   20   0  808M 30968 15240 R  99.  0.0  3:16.09 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299284 hzkurs00   20   0  879M 33080 15816 R  99.  0.0  3:15.97 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299316 hzkurs00   20   0  809M 30776 14444 R  99.  0.0  3:16.12 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299286 hzkurs00   20   0  808M 30104 14956 R  99.  0.0  3:16.14 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299300 hzkurs00   20   0  808M 30224 14492 R  99.  0.0  3:16.12 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299280 hzkurs00   20   0  808M 31168 16012 R  99.  0.0  3:16.15 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299305 hzkurs00   20   0  879M 31088 15684 R  99.  0.0  3:15.86 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299297 hzkurs00   20   0  808M 30392 14668 R  99.  0.0  3:16.10 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
299318 hzkurs00   20   0  809M 30528 14204 R  99.  0.0  3:16.11 /home/hzkurs00/gromacs/bin/gmx_mpi mdrun -s ion_channel.tpr -maxh 0.25 -reseth
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

Thanks for your attention

john.donners@atos.net