Tech Talk: From Mistral to Levante Q&A

Please type your name in front of each question/comment, such that we may contact you later if a question will not be answered here.

model forum

To help yourself in transition to Levante, we estabilshed a forum where all DKRZ users can share and discuss their thought and tricks on how models should be compiled and used:

https://redmine.dkrz.de/projects/model-forum/boards (https://redmine.dkrz.de/projects/model-forum/boards)

If a specific forum is missing, please send a message to support@dkrz.de (mailto:support@dkrz.de)

Q'n'A

- Miguel Andres (AWI): What does the `-mtune` compiler flag does, and why was not used for FESOM compilation?
    - (Thomas Jahns) -mtune is needed when an application is meant to support one set of architecture but optimize instruction use for a more recent derivative of that architecture. E.g. to build a program that supports any CPU since AVX is around (Intel Sandy Bridge) but still runs particularly well on Intel Haswell, one can choose -march=core-avx -mtune=core-avx2
- Mingyue Zhang:
- 1: What is the benefit of lowering the node distance?
    - (Flo) faster communication
- 2: How about CCLM on Levante? is there a running version?
    - (Ronny Petrik) Please ask the KSR group of the HEREON institute for compilation of CCLM on levante (ronny.petrik@hereon.de (mailto:ronny.petrik@hereon.de))
- Patrick Jöckel: Is `OMPI_MCA_coll="^ml,hcoll"` the same as `OMPI_MCA_coll="^hcoll,ml"` ?
    - (Dominik Zobel) yes any list item after ^ is negated (see also https://docs.open-mpi.org/en/v5.0.x/developers/frameworks.html (https://docs.open-mpi.org/en/v5.0.x/developers/frameworks.html))
- Patrick Jöckel: We recently detected that there is obviously a max. size of `mpmd.conf` ? Is this documented somewhere?
    - (Atos) this is known but unfortunately poorly documented; users will have to use wrapper scripts if the line length limitations are problematic
- Lukas Pilz: Could you provide some optimal compile flags for WRF also?
    - (Thomas Jahns) "Optimal" flags are only available for models that were part of the acceptance benchmark. If anyone provided a complete test case, we can provide assistance to find better flags. Please also use the model foum for WRF announced above to share your findings.

- Caroline Arnold: Is it possible to request heterogeneous resources across partitions (compute + gpu)?
    - (Hendryk Bockelmann) should be possible, we will try extensively once the GPU nodes are fully installed; and then document the results
- Brei Soliño: I remember encountering an issue compiling ICON with Intel compilers that, when consulted with DKRZ, mentioned it was related to a up-stream bug. However, it doesn't seem to be one of the "known issues". Is it something that has been solved already, or is there any update in general?
    - (DKRZ) specific informations for ICON on Levante are either shared in the model forum or via MPI-M and/or DWD internal wiki pages
- Adrien: Is there a list of software that you plan to install on Levante? I am looking to use PyFerret for example.
    - Patrick Jöckel: I can offer `module use -a /home/b/b302019/modules` and `module load pyferret` ... (it requires python3/2022.01-gcc-11.2.0)
- Erik: just a small note: the srun option --constraint=[a|b|...] (with the or operator) wont work from salloc or sbatch environment
    - Harald: --constraint="[a|b|...]" (look for "matching or" in manpages of sbatch salloc or srun) works for me, however there is no cell10 (yet). therefore it looks alike this: $salloc -p compute -N10 --constraint="[cell01|cell02|cell03|cell04|cell05|cell06|cell07|cell08|cell09|cell11]"
- Stefan Hagemann (Hereon): Are there any fortran constructs that work with ifort on Mistral, but not in Levante? Is the ifort more strict on Levante?
    - No. Behavior should be the same. It is like a change in compiler version. (Thanks)
- Dian Putrasahan (MPIM): Do we have an estimate when the GPUs would be available for use on Levante?
    - Perhaps end of the month. Thanks.
- Lukas Pilz: A couple of weeks ago, I tried to use --constraint=512G since this is documented in the levante configuration, but it didnt work. Are the 512G nodes not deployed yet?
    - (DKRZ) these nodes are deployed but maybe were configured in a false way; it's probably fixed by now. Thanks.
- (Thomas Jahns) Any news on jobs dying without creating the log file, another commonality is "jobs were cancelled by uid 0" and ran a few minutes (2-5 minutes)?
    - (Atos) no news so far
- (Ronny Petrik) What is with the 'Munge encode failed'?
    - (DKRZ) this was a mistake in updating SLURM on the fly - sorry for the inconvenience; it is fixed now
- Dian Putrasahan (MPIM): Mistral use to produce list of number of nodes used, length of job (time taken), etc at the end of the log file. This does not seem to be the case on Levante. Would it be brought back onto the epilog files on Levante?
    - Good to hear it is work in progress. Thanks!

- (Ronny Petrik) Nodes shared with other users need a lot of time to start running a job and often the job dies during execution. Could it be an allocation problem?
    - (DKRZ) we will add further nodes to the shared partition in order to reduce the waiting time
- (Ronny Petrik) Is it possible to overload a shared node by allocating too much resources?
    - (DKRZ) this should not be the case, but Levante is using a diffent configuration to limit the resources on nodes than Mistral; we will investigate if there are problems
- Patrick Jöckel: (referring to the spack netcdf example) Couldn't the same be found with `nc-config` and `nf-config` ?
    - (Hendryk Bockelmann) yes, but for other sw there might not be a 'foo-config'

Future DKRZ Tech talks

Please suggest topics that you would like to provide / see in future DKRZ Tech Talks

- Advanced SLURM options and scripting (Diego Jiménez de la Cuesta - MPI)