Questions for the course “Introduction to running R, Python, and Julia in HPC", 17-19 October 2023

TUESDAY, 17 October 2023, Python

Please note the following:

- Add your questions below, numbering them continuously.
- Be careful in case someone else is writing at the same time.
- Please **DO NOT** delete your questions even when they have been answered, as we are planning to use them to improve our material.
- Do not share any sensitive information as this document is accessible to anyone with the correct link.
- You can start a new line within a question by pressing <SHIFT> + <ENTER>.

1. General information / summary:
2. Hostnames for login nodes (the main course project is at UPPMAX):
    a. UPPMAX – rackham
        i. **SSH**: `rackham.uppmax.uu.se`
        ii. **ThinLinc**: `rackham-gui.uppmax.uu.se`
            1. 2FA may be needed, which can be handled by logging in with regular SSH, doing 2FA, logging out again, then there is a grace period of some minutes for you to login to ThinLinc. More here: https://nbisweden.github.io/workshop-ngsintro/2203-canvas/lab_connect.html
        iii. **From webbrowser**: https://rackham-gui.uppmax.uu.se/
    b. HPC2N - Kebnekaise
        i. **SSH**: `kebnekaise.hpc2n.umu.se`
        ii. **ThinLinc**: `kebnekaise-tl.hpc2n.umu.se`
        iii. **From webbrowser:** `https://kebnekaise-tl.hpc2n.umu.se:300/`
3. Course project at UPPMAX: NAISS2023-22-914
4. Storage area for the course project at UPPMAX: `/proj/naiss2023-22-914`
5. Course project at HPC2N: hpc2n2023-110
6. Storage area for the course project at HPC2N: `/proj/nobackup/hpc2n2023-110`
7. Create a directory in your project directory, at UPPMAX or HPC2N, and then you may get the course material from GitHub using: `git clone https://github.com/UPPMAX/R-python-julia-HPC.git`

Q/A Tuesday Python

8. My UPPMAX account hasn't been enabled yet, is there anything I can do about it?
    a. **A:** What is your name? You may be able to make it a private message in the chat, if you want. /Björn C!
9. In the project directory, what are the folders "nobackup" and "private" for?
    a. **A:** Not needed for the course, but "private" has less reading permissions for others and /nobackup is not "backed up". The /proj/nobackup/ is the file server on kebnekaise that the project storage resides on. Yes, it is just named that to remind people it is not backed up.
10. I am trying to connect with ThinLinc to Rackham, but the ThinLinc client hangs and does not connect (or it is taking very long to connect). Any ideas?
    a. **A:** Do you state "rackham-gui" instead of "rackham" ?
    b. **A:** You can try the web version as a start: `https://rackham-gui.uppmax.uu.se`
    c. There is a problem with 2-factor authentication using the ThinLinc desktop client. Connect first with SSH client and authenticate with 2FA, then connect with ThinLinc to use the grace period. See https://nbisweden.github.io/workshop-ngsintro/2203-canvas/lab_connect.html
11. INFO: use uquota command to see paths to project folder(s)
12. I still haven't managed to login... wrong username/password. When a new password is set throught naiss, is it propagated/linked immedately to the kebnekaise/rackham?
    I think that is what I "activated" at 9, and then it said it needs to be checked by someone before being accepted. When reloading the page after 5 min or so, it was activated, but I can still not login.
    a. From terminal: `ssh <username>@rackham.uppmax.uu.se`
    Will take a check again.
    So I had not requested the account from UPMAX, did that now. Can I use the hpc2n instead, that I should already have an account at?
        i. **A:** Use Kebnekaise!!
13. Not activated accounts: Is there a button or box you can click to activate it in SUPR?
    a. **A:** Did you follow all steps here: https://www.uppmax.uu.se/support/getting-started/applying-for-a-user-account/
14. Do you mind not saying our names on the open?
    a. **A:** I can probably cut out that part from the recordings when I process them
15. When trying to import package the XQuartz application opens, but no window is shown. If quit application I get error message in terminal. Should i have loaded python/3.11.4 before writing command 'pip list' and then 'import <package>?
    a. **A:** Yes, you should load any python modules before doing pip list or import package. They need the python module. You also need to start python before doing import package. But pip list you do outside python but after loading
16. Sorry, but what is the difference between load python and start python? Am I not starting python by the command module load python/3.11.4 ?

    a. **A:** Load python is the module load line. That activates important parts to python and its libraries. If not loading python you will get the system python!
    b. **A:** Then you can start python (terminal) and do your (interactive) work. You can also run python scripts with python <script>.
17. I managed to login with ssh from terminal but from the ThinLinc Client I got the wrong user/password message. Server name : kebnekaise-tl.hpc2n.umu.se, User adhs0001
    a. **A:** Did you put the password on the login box or wait for it to ask for password? It will not work unless you put the password in the login box: https://www.hpc2n.umu.se/documentation/guides/thinlinc
    - I followed the instruction, now I have a timed out message (more than three trials...)
18. NOTE: send an email to support@uppmax.uu.se if you have disabled accounts on UPPMAX
    a. Citation from web page: “After applying it might take up to 2 working days before you receive 2 emails with information on how to login to UPPMAX.”
    b. But I have asked my colleagues to make a fast lane. Unfortunately they are all on a meeting right now.
    c. We are sorry if you applied before Friday last week and still are waiting.
19. Sorry, why are we wasting so much time on this? There was too little timee before and now we sit in breakut rooms forever to jsut import one package? When are we coming back to lecture?
    a. Richel: I assume this is about the ‘pip’ part. The reason that this session took longer than needed, is that I could not predict how fast students are, where the previous time it did take 45 mins. I see many possible explanations for what caused this big difference.
20. LEARNINGS: The change of course structure means some tuning of timings! Sorry for that!
21. When do we want to only import packages and when do we want to install? Will you continue with this in next lecture?
    a. Yes, in next session! If the package is not found you may need to install it.
22. sorry, was the task just to import pandas?
    a. **A:** Yes, load python, see what packages exist (with pip list), and then start python and try import a library, for instance pandas
23. For future breakout rooms, will you bring us back to the ‘general’ room or will you give us a time when we should go back ourselves? This was a bit confusing this time.
    a. Richel: I agree that this was unclear! And this by deliberate choice: if all went well, I would not have liked to disturb the students. If students would need help, I would call them back. I assumed the schedule and me going through the Breakout rooms would clarify things. Maybe next time I will share my plans :-)
24. Regarding using ThinLinc on rackham with: `rackham-gui.uppmax.uu.se`
    a. 2FA may be needed, which can be handled by logging in with regular SSH (`rackham.uppmax.uu.se`), doing 2FA, logging out again, then there is a grace period of some minutes for you to login to ThinLinc. More here: https://nbisweden.github.io/workshop-ngsintro/2203-canvas/lab_connect.html

25. Will you go through the material on the isolated env on the https://uppmax.github.io/R-python-julia-HPC/python/isolated.html later? Or will there be a recording of this?
    a. Richel: I am unsure if this question is about venv or conda. If 'venv': I assume you've found the recording at the bottom part of the page. If this is about conda: we'll discuss that later
26. When will the recordings be up?
    a. Later today/tomorrow
27. The error that you get in the youtube video for venv should be addressed because now I dont know how to install an older version of a package if the environment has a newer version already installed.
    a. **A:** If you need a newer version, I know it works to use
       `pip install --upgrade <package>`
    b. **A:** If you need another/older different version you do:
       `pip install --upgrade package==version`
    c. **A:** If there are prerequisites, you need to do those first, or list all of them in order, e.g. if you install numba which depends on numpy and depends on scipy, you would do:
       `pip install --upgrade scipy numpy numba`
    d. `Richel: I agree that the video ideally did install something useful in the first demo. As 'Exercise 1' I felt it would be good enough. Also, when I scheduled to make the video, I expected the course materials to work. I was unexpectedly wrong. The material (and video) will likely be updated.`
28. so we need to create a new environment with the old version and then upgrade the others?
    a. **A:** You create a virtual environment with site-packages because there are many more than numpy in that. So use site-packages whenyou create the venv, and then upgrade the ones that are too old
29. After creating Example-gpu and installing numba, what are we supposed to do now? Is there more lecture coming?
    a. **A:** There will be "informal talk" at 11:45, then break at 12:00, then we continue with lectures at 13:00
30. I feel like the instructions part goes really fast and is a bit unclear - what we are supposed to do on our own can be clarified. The working in break out room time feels quite long. I'd like to suggest spending more time on instructions and making sure the task is clear. At this time I do not know if I have done what I should have done, or if I have done more or less than instructed. Edit: All parts, general thought.
    a. Richel: I am unsure about which part this is about. As it is about a long breakout session, I assume this is in my part. In my part, I went through all rooms one by one, except the Silent room, as it is a silent room. I go through all rooms *exactly* because I check if all learners what to do. We hence agree that all learners should be checked upon :-)
       Question back: So, I assume you were in the Silent room then? If yes, would you recommend visit the Silent room too? I felt I should not. Would you accept my

reasoning being a valid way to look at it? Maybe I should state that there are rooms for 1 person too

Answer: I fully understand the reasoning about not visiting the silent room. I think it is the time we are in the breakout rooms that throws at least me off, as it is quite a long time. Then it feels like we are supposed to do more than what was understood by instructions the first time around. A bit more structured instructions before entering breakout rooms I think would make this less confusing.

31. When updating numpy, scipy and numba – how do we know it is actually the newest versions that were updated/installed?
    a. You can specify the version for instace with “pip install numpy==1.15.4”
    b. But how do we know the version number?
    c. pip list will show the versions that are installed
    d. Sure that is correct, but if want to install a newer version – how now what version numbers one should specify? As the ones in pip list are the current ones right. Maybe I’m totally lost, I’ll make own research on this. Thanks.
32. If your run takes shorter amount of time than you allocated (e.g. 1 hour instead of 2) do you get charged with 1 hour?
    a. **A:** at HPC2N, the project is charged with the time the simulation takes not the allocated time
    b. **A:** if your job takes 2 hours and is scheduled for 1, it will stop after 1 hour and the scheduler will count 1 hour
    c. **A:** also for UPPMAX the project is charged with the time the simulation takes not the allocated time
33. Were we supposed to do the first part of the exercises, where virtual environment called analysis was to be created?
    a. Richel: I hope it got clear we did ‘Exercise 1’ from the slides, and than later create ‘GPU-example’, as I visited all Breakout Rooms twice, except for the Silent Room. If you were in the Silent Room, I hope you can at least understand why I let the Silent room be .. well .. quiet :-)
34. Can I also just type "conda activate <name>" in the batch script?
    a. Richel: Yes! I am unsure if this is recommended though ...
35. Can I forward parameters from “sbatch file.sh 2 3” to the python code file?
    a. Richel: Yes! Use the bash input variables, $1, $2, etc
36. How can I get cuda information (i.e. version) from Snowy? (need the pytorch package to match the cuda version).
    a. **A:** Start an interactive job on Snowy and use module spider to find a module version (general).
37. Can anybody create a project? As a student for example. Or we can only be added as a user?
    a. If you are a student, your teacher (as principal investigator) has to apply for the project
    b. https://uppmax.github.io/uppmax_intro/naiss.html#the-projects

c. application for projects are done in supr.nass.se There are national systems (Rackham, Dardel, Alvis, Tetralith) or local (HPC2N, Lunarc). If you are a PI that is a senior researcher you would be able to get medium and large projects
If not you could apply for small projects

38. Also can you stream output from a batch job (e.g. a progress bar)? So you can see how it is doing?
    a. One way is to create a log file (a line in the batch file) that is saved somewhere. Then you can read it while it is updating. I sugget to use the command tail to just see the last rows of the file
    b. More info about slurm at UPPMAX https://www.uppmax.uu.se/digitalAssets/560/c_560271-l_1-k_uppmax-slurm-2023-02.pdf and https://www.uppmax.uu.se/support/user-guides/slurm-user-guide/
    c. email:
    - --mail-type=BEGIN,END,FAIL,TIME_LIMIT_80
    - out/err redirection:
    - --output=slurm-%j.out and —error=slurm-%j.err
    …..by default, where %j will be replaced by the job ID
    - --output=my.output.file
    - --error=my.error.file
39. Just to confirm, so I add #SBATCH --gres=gpu:x in my script to activate GPU if needed?
    a. Yes
40. You (Birgitte) are in Python directory. Is this inside Exercises directory? Should we be in that as well I assume??
    Does it matter in which directory Exercises vs Python you are in while doing these batch jobs?
    a. The only thing that matters is that the exercises/programs that you want to run is in the directory you are in. Use the one you placed stuff in and change paths in virtual environments/batch scripts accordingly.
41. How to use conda with SLURM? Or SLURM with conda?
    a. Just load the conda module within the SLURM script and then do conda activate There will be something about conda later today
42. Does Thinlinc work with WSL2?
    a. At HPC2N, you can work with Thinlinc online: https://kebnekaise-tl.hpc2n.umu.se:300/ . Thus, you dont need to install it.
    b. I don't know if I would consider WSL2 in a thinlinc environment. You start thinlinc from windows not the linux terminal, I would say. There is no need to.
    c. Yes, there are ThinLinc clients for Linux, Windows, macOS. Nothing to do with WSL2. You just run the client suitable for your OS. I don't know if ThinLinc can be run from WSL2, maybe? I don't have access to a Windows computer to test right now. But no need to run it from WSL2 I guess.
43. Can I use dask for distributed computing on rackham?

    a. We haven't tried it fully yet but my plan is to test with slurm next week. I know a colleague has made it working a while ago. /Björn
    b. I have made it work within one node (20 cores). You may test from the material presented in the **links** in the end of the parallel material, using you own installed dask package! /Björn
    c. I can't find any dask code here: https://uppmax.github.io/R-python-julia-HPC/common/parallel.html#dask, or do you mean somewhere else?
        i. https://aaltoscicomp.github.io/python-for-scicomp/parallel/#dask-and-task-queues
        ii. https://enccs.github.io/hpda-python/dask/
    d. We will have another course for Python more dedicated to HPC in December. We plan to introduce DASK more thouroghly then. More info in a couple of weeks in the same info channels as you got the info about this course.
44. I tried out the instructions for jupyter (UPPMAX for rackham). How do I get out of the interactive session afterwards again? Or can I just stay in it for the other exercises too?
    a. $ exit
    b. you may want to test some more things! Then stay!
45. When I am using command "jupyter-notebook" I am getting following error. Bus::open: Can not get ibus-daemon's address.
    IBusInputContext::createInputContext: no connection to ibus-daemon
    I am using uppmax
46. my code is running but no jupyter notebook is showing off
    a. Which url did you copy? Should be the one with hostname. Are you inside ThinLinc (for HPC2N)?
    b. I'm not in ThinLinc, I will retry
    c. For HPC2N, you need to be within HPC2N domain, so easiest to do with ThinLinc
47. For running distributed code (parallel processes), do I need to be in interactive mode, or how do I combine that with SLURM?
    a. You can run with SLURM, you just need to ask for more cores and use srun for your programs. There are some examples on the Kebnekaise intro slides (near the end – download the PDF): https://github.com/hpc2n/intro-course/blob/master/slides/3.usage.pdf
    Also some examples here: https://www.hpc2n.umu.se/documentation/batchsystem/mpi-examples
48. https://uppmax.github.io/bianca_workshop/slurm-intro/
49. For distributed computing (multi-node) I am aware of mpi4py (which uses MPI) and Dask
50. How to use this code on the cluster? (Multiprocessing in parallel session)
    a. This is the code of a Python script. That means that it can be `sbatch`ed later. Probably using the GPUs in the batch scripts
    b. Do I need to sbatch it exactly with the number of processes as in the python script? with n=6?
    c. I predict: yes. I predict if you do wrong, it will work, yet less efficient

    d. you can allocate more cores I guess, but they will not be used. you will be accounted for the allocated cores though so it is more "expensive"
    e. note that your CPU usage counted will be the sum of all cores. Hence if you use 10x more CPUs for a 2x speed increase, you won't be in the front of the queue long anymore
    f. you may use <= number of allocated cores inside of your python script but not more. Otherwise, your script will create an overhead on the nodes
    g. Ok<y, and this will Work easily with n >> 20?
    h. if there are >20 cores in the node yes. Dardel for instance has 128 cores/node
    i. you can allocate more cores I guess, but they will not be used
    j. Note that your CPU usage counted will be the sum of all cores. Hence if you use 10x more CPUs for a 2x speed increase, you won't be in the front of the queue long anymore :-)
51. Question: I am using Bianca, for a script that generate figures, do I just add #SBATCH --gres=gpu:1 within my script? Do I need #SBATCH -M bianca as well?
52. When working in interactive mode, how does one know how many tasks/nodes to command?
53. Hoa about cross-node distributed computing?
    a. For distributed computing (multi-node) I am aware of mpi4py (which uses MPI) and Dask
54. (from chat) How do you load cuda, is that simply set up with certain python versions?
    a. There will be a session about it later, but in general it is included with some versions (though at HPC2N at least you mean need to load separarely)
55. HPC2N YouTube channel will have the recordings later today or tomorrow: https://www.youtube.com/user/HPC2N/videos
56. Direct link to the recordings playlist on HPC2N's YouTube: https://www.youtube.com/watch?v=aCSMEGglTG8&list=PL6jMHLEmPVLxdCllWrzseYkGoCCXjrukM
57. Evaluation survey for today: https://forms.office.com/e/NxtwCk7w5H