Q&A for New User Training June 12-13, 2024

- GDoc is used for Q&A (instead of Zoom chat)
  https://tinyurl.com/mpxxy4bs (Day 1 and Day 2)
  https://tinyurl.com/3ch38j7b (Day 1)

- Survey:
  - Day 1: https://forms.gle/tGgRARdKt8nikDq96
  - Day 2: https://forms.gle/V7DqYafjXgAM9XMBA

- Additional Hands-on exercises on Perlmutter
  % git clone https://github.com/NERSC/intro-NERSC-resources.git

  To access the reservation today, for CPU, until 11:45 am
  --reservation=new_users_June2024CPU -A ntrain6
  It is two short dashes in front of the "reservation=" above
  Outside of the reservation, just submit your batch job without the above two flags

- Slides and videos will be available on the NERSC New User Training Training Event page (https://www.nersc.gov/users/training/events/2024/new-user-trainingjune2024/)

- Zoom info for New user Training Office Hours on June 14, 2024:
  https://lbnl.zoom.us/j/96212457679?pwd=akNyK0ZFZGpieGNRNGVHZEdsdDdNQT09

==========
(Note: User and staff names will be removed from below before publishing on the event page.)

Please write your questions and names below.

Day 1 June 12

Q ([redacted]) What is the storage allocation in /home and /scratch for per user?
A (Helen): the default allocation for HOME is 40 GB, and the default for SCRATCH is 20 TB.

Q: ([redacted] how do I save my work if the "scratch" area is where the programming work is done (execution is on cluster), and it regularly gets wiped? I am aware of the other storage locations but not how to save files to those.
A (Helen): There are the CFS (Community File System) on the system and the long term storage (HPSS) that you can save your data to.
You could do a simple copy to CFS, and hsi/htar to HPSS. You could also use Globus for data transfer. There is an Data and IO talk next for more details.

A (Rebecca): Also keep in mind that files that have not been used are the ones that get purged. So if you are regularly accessing a file, for example, then that counts as using it and it will not be purged until it hasn't been used in any way for at least 8 weeks.

Q: ([redacted]) What are the key performance characteristics and benchmarks for the Perlmutter system, and how does it compare to other top supercomputers ?
A: (Helen) Here is the peak flops and memory table at https://docs.nersc.gov/systems/perlmutter/architecture/#system-performance
Perlmutter is currently #14 on the Top500 list. https://top500.org/lists/top500/list/2024/06/

Q: ([redacted]) I am not having my NERSC login credentials. Where I can get my login credentials to join the slack?
A (Helen): You will need to become a NERSC user. The easiest method is to ask an existing PI to ask you onto their project. Or you could apply for your own allocation. More details at https://docs.nersc.gov/accounts/#obtaining-a-user-account

Q: ([redacted]): My xterm crashes every time on ssh when I open Matlab. I have a Mac. I cannot visualize plots properly using xterm/xquartz as well. Can you help in any way? Is there any other way to visulize plots or Matlab?
A: (Helen). I suggest you to try NX, which supports accelerating X-forwarding. https://docs.nersc.gov/connect/nx/. Lipi also talked about this in her presentation.
About xterm crashes etc, please open a ticket with details at https://help.nersc.gov.

Q: ([redacted]) I applied for my NERSC account on Monday, still haven't got it. Is there a rush of new users? How long does vetting take if the "call us back if" condition is a week passing?
A: (Helen): new users need to be approved by DOE which can take some time. You would usually get approved within a week.
A: (Rebecca) Also, after your account is vetted, it needs to be approved by your PI. Check with them to make sure that they know your account is in their queue (they do get an email about it) and they need to click the "thumbs up" button to add your account to their project. Also, if you give us your username we can take a look and see what's holding it up.
Thanks for the offer. Username: nbidler
A: (Helen): I checked in Iris, and see your status is currently "vetted" and waiting for PI to approve.
Thanks much, I just emailed the PI.

Q:([redacted]) :I tried to delete files in my Global/home directory to get space, because it was almost out of space. However, even after deleting all them, the space usage didn't change. I suspect I need to empty the trash, but I can't cd to the trash path in Global/home/username/.local/share/Trash/files/ as usual . Any suggestions to clean or free my home space?
A: There is a lag sometimes. What is your username? I can help to check. My user name is: yinanhe. Your current quota (you can check with the "showquota" command) is 10 GB out of the 40GB quota, and inode usage is small too.

([redacted]) I can “cd /global/homes/y/yinanhe/.local/share” now. “cd /global/homes/y/yinanhe/.local/share” works, but I can not further cd trash (yinanhe)“There is a lag sometimes”? Does this mean I do not need to do anything? the trash would be cleaned automatically?

Q:([redacted]): How can I set up the MFA? I followed the instruction here https://docs.nersc.gov/connect/mfa/#configuring-and-using-an-mfa-token to log into IRIS but can not find the MFA tab in my account.
A (Helen): you talked to me during the break that you are using a training account today, right? An actual user account will have the MFA tab for you to setup. Training accounts can login with just the password.
([redacted]): Great, this worked! I used just my password and was able to log into perlmutter from my terminal because I am using a training account. Thank you so much!

Q: ([redacted]): How can I make Globus automatically start at login on Ubuntu? Additionally, how do I mount Global/home and Pscratch on NoMachine CentOS? as I do lots of image processing and need to check the processed images frequently, mounting Global/home and Pscratch would make my work easier.
A (Helen): On NoMachine, when you login to Perlmutter, global home and pscratch are mounted. You can set up personal Global endpoint on your local machine if that’s what you mean, then you can use Global transfer between your local machine and Perlmutter easily.
https://docs.nersc.gov/services/globus/
https://docs.nersc.gov/services/globus/#setting-up-a-personal-globus-endpoint
Q:([redacted]): What is the easiest way to have a double tab open into our shell? I generally login twice after opening a new tab, is there a better way to do it?
A (Helen): There isn't a built-in way within the shell itself to open multiple tabs. Use some shortcut to open a new tab is easiest. Are you probably concerned about login to Perlmutter repeatedly and enter password+MFA every time? If this is the case, we have the “sshproxy” utility that allows you to only enter password+MFA every 24 hrs, it works for ssh and any NERSC web sites that are password protected (such as Iris, mynersc, help.nersc.gov, etc.)
https://docs.nersc.gov/connect/mfa/#sshproxy

Q:([redacted]) How to access NVIDIA GPUs via PerlMutter or How to run CUDA code on PerlMutter? As I am working on a project that requires me to test my CUDA code on different CUDA versions and my local GPU is probably not compatible because it is throwing errors.
A (Helen): You will request GPU nodes via “#SBATCH -C gpu” flag in your Slurm batch script. You can find some info on building and running GPU applications at:
https://www.nersc.gov/assets/Uploads/Building-and-running-GPU-applications-on-Perlmutter.pdf
Also: https://docs.nersc.gov/development/programming-models/cuda/
https://docs.nersc.gov/systems/perlmutter/running-jobs/

Q [redacted]): Is the dvs_ro faster for jobs to use than /global? Why should we use that one?

A (Helen): Read-only is a lot faster than read-and-write permission for GPFS file systems ($HOME and $CFS) from compute nodes. The usage example is to use the dvs_ro for software and input files, and use the regular CFS directory or $SCRATCH for writing your output files.

Q. ([redacted]): in theory, the project i'm assigned to is just "update the old code with the new tool" and the actual testing for meaningful data has already been done. I am planning to find the "minimum size" dataset to run tests to make sure the before/after gives the same answers. Are there any other ways to minimize my time "taking up space" on the HPC cluster?
A. (Erik) Hi Nicholas, I think this depends a lot on the project you are working with. Is the dataset in shared space, such as CFS? If so, you could avoid creating your own local copy and instead copy over the test minimal data set from the CFS directory.
What are the sizes (roughly) of the data sets we're talking about?

Q ([redacted]) I do not see the current slides on the "presentation materials" section of https://www.nersc.gov/users/training/events/2024/new-user-trainingjune2024/ - is it somewhere else or is this covered under documentation?
A (Helen): today's slides have all been uploaded

Q.
A.

Questions below are copied from Zoom Q&A:

[redacted]) Will the video recording and/or slides be made available after this to review as well?
Helen He (You)
Yes, a few slides have already been uploaded to the event page.
Link to videos will be available in a few days after processing

Q ([redacted]) Can you please send the link to the event page?
A (Helen) https://www.nersc.gov/users/training/events/2024/new-user-trainingjune2024/

Q ([redacted]) I am not having my NERSC login credentials. Where I can get my login credentials to join the slack?
A (Helen) Copied your question over and nswered in Google Doc Q&A:
https://tinyurl.com/3ch38j7b

Q ([redacted]) Why is it that common is read only on compute nodes? Does this somehow yield better performance? If so how?
A (Helen) I have answered this same question from another user in the Google Doc Q&A:
https://tinyurl.com/3ch38j7b
Read-only is a lot faster than read-and-write permission for GPFS file systems ($HOME and $CFS) from compute nodes. The usage example is to use the dvs_ro for software and input files, and use the regular CFS directory or $SCRATCH for writing your output files.

Q ([redacted]) Do users get notified/warning about when files in SCRATCH will be deleted?
A (Helen): No, there will be no notification.

Day 2 June 13

Q ([redacted]): Should I be able to see project ntrain6 on Iris? I only see the project for my internship listed currently.
My NERSC login: hoill
A (Helen): what is your NERSC login name? I can check. We added all NERSC existing users to ntrain6 yesterday. I checked, and you are in ntrain6 both in Iris and on Perlmutter. You can check via the "iris" command on Perlmutter to see your projects and allocations.
([redacted]): I do see it there. Thank you.

Q (from before the start of the session): How do I apply for a non-training account at NERSC?
A (Rebecca): you will need to apply for a new user account at https://iris.nersc.gov/add-user . In the form, you will request to be added to your PI's project, which can be looked up either by PI name or project name. After you submit your request, your account will go through a vetting process, and then it will go to the PI to approve/reject joining their project.

Q ([redacted]) My question may not be specific to this talk. I commonly use VASP. How do I optimize my jobs such that I can calculate that this job will need xx amount of compute hours? This may help me to plan ahead of time what jobs to run and what not. I have different projects which are chargeable to NERSC.
A (Helen): we are having a VASP training later in the summer, and general information on using VASP on Perlmutter including performance will be provided, For now, it is best for you to submit a ticket at https://help.nersc.gov, with your specific input file, and we can direct you to our VASP support staff. Also please do the experiments for your setup as Rebecca suggested below.
A (Rebecca): You will probably need to experiment somewhat with different processor counts – there's no way of knowing how long a program will take to run without doing some representative runs ahead of time. I think VASP tends to do some iterations of a particular computation, so one thing you could try is to run the problem for 15 minutes on different processor counts. Then look at the output and see how many iterations it completed in that amount of time. Be sure to take into account the startup (initialization) time, but after that, you should be able to know about how long it takes per iteration. Once you know that, if you know how many iterations it's going to take to solve the problem, you should be able to know how long you will need to run with that particular processor count. Does this make sense?

Q ([redacted]): Do batch jobs have a lower priority than interactive jobs, even if they are requesting the same resources? For instance, if I have a job that requires a single gpu node for 15 mins, would you recommend using an interactive or batch job? Thanks!

A (Helen): we have set aside some nodes just for interactive qos. So you either get the interactive nodes (requested via salloc, not available via sbatch) within 6 minutes, or your job will exit with a message that nodes are not available at the moment. Interactive usually works great for small jobs, up to 4 nodes, 4 hours.
For batch jobs, you can also use the debug qos for small short jobs, up to 8 nodes and 30 min. Debug jobs have higher priority than regular qos jobs. There is also the premium qos that is of higher priority than the regular qos. Your PI needs to add you to use premium qos in Iris.

Q ([redacted]) Can we specify the number of cpus?
A (Helen): yes, you can request via the -N <value> -C cpu.

Q ([redacted]) Is there a problem of jobs getting preempted?
A (Helen): For jobs to take advantage of the preempt qos, it is best that your application has the checkpoint/restart capability. You will need to use special flags in your batch script, please see details at https://docs.nersc.gov/jobs/examples/#preemptible-jobs

Q ([redacted]) Can we directly submit a Makefile to SLURM?
A (Helen): You can include the Makefile command to build your application inside the batch script before the srun commands to run the executable.
A (Rebecca): But, keep in mind that valuable compute time is ticking down while your application is being built. If possible, it's best to build your application before submitting a job.

Q ([redacted]): The shared queue mentioned a maximum of ½ node. Can we control how much of the node is given to us in the shared qos? Or is it just allocated as the job progresses?
A (Helen): yes you can request up to half of the node, and you can specify the portion of the node via the -c xx or --mem flags. Please see details at
https://docs.nersc.gov/jobs/examples/#shared

Q ([redacted]): so if I am only updating code to use a new tool without changing the "answers," I should be using "-q debug" or "-q qebug" to repeat short runs of code to make sure it runs/matches pre-change outputs?
A (Helen) yes, debug queue is specifically for the purpose of doing short debug/test runs before you submit longer production runs to -q regular!

Q ([redacted]) So when I do scancel, does it cancel the most recent job submitted or all of my job currently running on my account?
A (Helen): "scancel -u <user_name>" will cancel all your jobs. You can use "scancel -j <jobid>" to cancel a specific job

Q ([redacted]): Is backfilling automatic? Around how short do jobs need to be backfilled? thanks
A (Rebecca): Yes! Every few minutes, Slurm goes through the priority-ordered list of jobs twice. On the first pass, it plans its schedule for the next 48 hours or so. First it looks at the highest priority job and places it in its schedule (if possible). Then it looks at the next job and places it in the schedule (if possible), going down the list down to a certain priority level. Next, it goes

through the entire list to do what we're referring to as backfill – it's checking whether, without making any changes to the schedule it just planned in the first pass, a given job could be started RIGHT NOW. As you can imagine, there are a lot of holes in the schedule that Slurm creates in that first pass. The holes tend to be pretty short in duration, generally 2 hours or less. Interestingly, they can be very broad in terms of number of nodes. Up to about half the size of the partition the size of your job with a similarly short walltime doesn't matter (i.e., 1 node or 1000 nodes, you still have a good chance of getting your job started in the backfill pass of the scheduler). The wait time is strongly correlated with the walltime and weakly correlated with the number of nodes requested. TL;DR: jobs less than 2 hours walltime (1 or less is even better) are more likely to fit into the gaps in the schedule and therefore run as backfill; node count doesn't matter as much as walltime.

Q: Where can we find the presentation recordings and slides?
A: Slides for Day1 have been published on the event page at https://www.nersc.gov/users/training/events/2024/new-user-trainingjune2024/
Day 2 slides will be added today. Videos will be published on the NERSC YouTube channel after being processed (split + trim), and the link to the playlist will be added to the event page as well. We will also order professional captions and add them to the recordings.

Q ([redacted]): How can we open a jupyter notebook to edit code without starting a node? Say I expect to be using the debug queue for a long time to check individual functions in a program.
A (Rebecca): Just request a Login Node if you're not going to be doing any computing. There is no charge to your allocation for using a login node.

Q. After loging on to JupyterHub, which node should we default start with again?
A (Helen): After you login to https://jupyter.nersc.gov, there is no default option, and you could pick any of the 5 options: login node, shared CPU node, exclusive CPU node, exclusive GPU node, and configurable job.

Q ([redacted]): regarding "Additional Hands-on exercises on Perlmutter - % git clone https://github.com/NERSC/intro-NERSC-resources.git" - into what location should we clone the files? Our $HOME directory on the cluster or our local terminal, or some other thing?
A (Helen): I recommend after login to Perlmutter, "cd $SCRATCH", then "git clone", and do your exercises there. Please keep in mind that $SCRATCH is purged, so you should save important files in $HOME, $CFS/<your_project>/<your_login>, or HPSS.

Q ([redacted])- Hey guys, I am not sure if I am just overlooking this or not, but I was tasked with opening a jupyter file with a conda environment. Is there a detailed step by step guide in regards to this available via NERSC materials?