Q&A - Crash Course in SuperComputing, Friday, June 28, 2024

- GDoc is used for Q&A (instead of Zoom chat)
  https://tinyurl.com/4fvkzeud

- Apply for a training account if no NERSC account at the time of registration or if your MFA to log into NERSC systems is not yet set up: https://iris.nersc.gov/train and use the 4-letter code ~~bk8X (is full now)~~, can use code e71A for today too

- Hands-on exercises on Perlmutter
  ```
  % cd $SCRATCH
  % git clone
  https://github.com/NERSC/crash-course-supercomputing.git
  ```

- Slides and videos will be available on both the NERSC Training Event page and CSASP CS Summer Program page
  Crash Course in Supercomputing, June 28, 2024
  Computing Sciences Summer Program 2024

  Course slides:
  Already posted on the event page: **Crash Course in Supercomputing, June 28, 2024**

- Please help us with answering a short survey afterward
  https://tinyurl.com/562bvv62

=======

Example commands with salloc:

yunhe@perlmutter:login36:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite>
salloc -N 1 -C cpu -t 30:00 --reservation=crash_course -A ntrain3 -q interactive
salloc: Granted job allocation 27305372
salloc: Waiting for resource configuration
salloc: Nodes nid004175 are ready for job
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite> lsc
c fortran runall_batch.sh runall.sh
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite> cd c
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> cc -o
darts darts.c
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> ./darts

Computing pi in serial:
For 1000000 trials, pi = 3.141648

```
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> cc -o
darts-mpi darts-mpi.c
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> cc
-fopenmp -o darts-omp darts-omp.c
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c>
./darts-omp

        Computing pi using OpenMP:
        For 1000000 trials, pi = 3.141648

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_NUM_THREADS=8
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c>
./darts-omp

        Computing pi using OpenMP:
        For 1000000 trials, pi = 3.141648

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c>
OMP_NUM_THREADS=32 ./darts-omp

        Computing pi using OpenMP:
        For 1000000 trials, pi = 3.141648

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> cc
-fopenmp -o darts-hybrid darts-hybrid.c
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408
```

```
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_NUM_THREADS=16
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_NUM_THREADS=32
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_NUM_THREADS=64
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_NUM_THREADS=128
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_PROC_BIND=spread
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> export
OMP_PLACES=threads
yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4
-c 64 --cpu-bind=cores ./darts-mpi

        Computing pi using six basic MPI functions:
        For 1000000 trials, pi = 3.141408
```

yunhe@nid004175:/pscratch/sd/y/yunhe/crash-course-supercomputing/darts-suite/c> srun -n 4 -c 128 --cpu-bind=cores ./darts-mpi
srun: error: Unable to create step for job 27305372: More processors requested than permitted
(this is expected, since you do not have 4*128 logical cores on a single node)

Q ([redacted]) where did you want us to put our names? In the question document?
A (Lipi): If you are in person in the room, you can join the Zoom and just make sure your Zoom name is your actual name and not “computer_1939” or something.

Q ([redacted]): Will we touch on CUDA in this seminar?
A (Lipi): Not in today’s session (I believe) but Helen or Charles can confirm. We did have a recent session about MPI-aware CUDA (the video of the session will be on youtube soon!)
A (Helen): we will host an intro to CUDA Programming training soon, and can let you all know when it is scheduled.
A ([redacted]): Ok, thank you both for clarifying!

Q (Name): Is there a ssh address we should connect to, to access our training accounts?
A (Helen): ssh your_login_name@perlmutter.nersc.gov. See Slide 6 of the morning slides deck.

Q (Name): I signed up for a training account a few days ago but I don’t think I got any email message back, how do we figure out the “OTP” part of the password + otp login prompt @perlmutter.nersc.gov?
A (Helen): Training accounts do not need OPT (or MFA), you should have got a password on your screen the time you got your training account. If you do not remember that, please go ahead and apply for a new training account at https://iris.nersc.gov/train, use 4-letter code bk8X

Q ([redacted]): I am getting the following error when trying to ssh to perlmutter
(base) NERSC-training $ ssh <user-name>@perlmutter.nersc.gov
ssh: connect to host perlmutter.nersc.gov port 22: Operation timed out
Is this a problem on my end? Do I need to specify a port? Thank you.
A(Name): You should not need to specify a port. Can you copy and paste exactly what you are putting on the command line (into this document so I can see it)
Follow up: *(base) NERSC-training $ ssh <user-name>@perlmutter.nersc.gov*
Where username is the one I got after the registration and I’m using bash shell on a mac.

Q (Name): I tried to apply for a new training account at https://iris.nersc.gov/train. But it says that it Could not find an available training account.
A (Helen): use 4-letter code bk8X, do not put any other characters before or after that 4-letter code

A (Rebecca): update: all the allocated training accounts have been taken. Use e71A instead. Accounts from e71A will expire on July 5.

Q ([redacted]): How can I check if I was properly added to the ntrain3 project? Could someone verify I am on the project? My account was approved between registration and the event.
Username: nryanl
A (Lisa): In general you can check in iris.nersc.gov what projects you belong to. I don't see the project there. We will check that you can be added to the project. Yun He Can you add nryanl to the project?

Q ([redacted]): I received my NERSC account *after* I registered for the crash course. Should I go ahead and create a training account?
Q ([redacted]): Does it matter that I would not have been added to the ntrain3 project?
Q ([redacted]): My NERSC username is[redacted]. I have only set a password. How would I go about setting up MFA? I've just started reading docs.nersc.gov
A (Name): If you have a NERSC account, please use that. You will need to set up MFA in order to use your NERSC account - you can still use it regardless of when you registered for this course.
A - no it does not matter, you are automatically part of a default project and you will be able to submit jobs for computation using that project's hours.
A (Helen): please let me know your NERSC account name and I can add you to ntrain3 now for accessing node reservation for today. Or you can go ahead and appy for a training account, then no MFA is needed, and you will be in ntrain3 too. Added jaytau to ntrain3 in Iris (account management side), it may take up to an hour to be effective from the Perlmutter side

Q (Name): how do we compute number of cpu hours for a task? Will we get to this today?
A (Lipi): This is an interesting question - it is often that you will need to do some testing to figure out how long your job needs to run in order to complete. You may be able to use the debug queue to get a sense of how long part of your code runs and then estimate the full requirement from there. The speakers may discuss this today :)

Q (Name): Is there a way to request the number of cpus on a compute node, with salloc I always get the entire node?
A (Lipi): You can use the shared queue to request up to half of a node if you do not need more than 64 cores on a node (1 CPU). Salloc is the method for getting an interactive session on our system, but you can specify which "queue" you want to use, which will decide how much of the system you can use during the interactive session. Let me know if you need more help with this!

Q ([redacted]): I applied for a training account and received a password about a week ago, the password doesn't log in and I am getting the error stating "Could not find an available training account." with the correct 4 letter code, what should I do?
A (Rebecca): looks like there are actually no more accounts available under that code. Use e71A
Instead. Those accounts will only be good through July 5 instead of July 10.

Thanks! It worked. ✅
A (Helen): Rebecca is right. bk8X is full. e71A code we used for the New User Training also works for today.

Q (Name): What's the difference between cluster architecture vs massively parallel processing?
A (Name): MPP strictly is not an architecture, but it is a way of describing the cluster architecture with how it is used.
Okay, thank you!

Q ([redacted]): When I ssh into *$ ssh <user-name>@perlmutter.nersc.gov, how do I make sure that I am on the training account and not on my other account?*
A (Name): If you have a NERSC account, you can just go ahead and use that instead of the training account. Do that by using your NERSC account user name (probably something related to your name).
Q ([redacted]): Thank you! However, I have been added to ntrain3 and I wanted to use that allocation… please advise
A (Lipi): your account has been added to the project, so once you login and begin running some kind of computation you can specify the project using the -A flag in your submit scripts or salloc command. Thus you will use hours in the ntrain3 project. You could log in once and switch between ntrain3 and any other project for your computation; it is not tied to how you log in.
Q ([redacted]): Thank you, finally, what exactly is the 'res_name' in the following command: salloc -A ntrainN -N 1 -t 00:30:00 --reservation=res_name. Thank You!
`--reservation=crash_course`
=ls

Q (Name): what is meant by "compute power", more specifically?
A (Lipi): I think Charles means "number of available resources that can work together" (roughly). For example: Perlmutter has more compute power than a laptop. :)
A (Helen): roughly, the computing capability of a machine

Q ([redacted]): In this training, are we actually using Perlmutter, or a small part of it like ODO? ODO is a model of Frontier.
A (Lipi): Not sure what ODO is :) Our only public system is Perlmutter so YES! You will be using Perlmutter :) No using Frontier- this is a NERSC training :) Though we do sometimes co-host trainings with OLCF, but getting an account on Frontier is much harder so likely this is why they use a model for their trainings. At NERSC getting accounts (and training accounts) is a faster process so you get to use the actual system!
A ([redacted]): Great!

Q ([redacted]): I have a NERSC account already but I am yet to be added to the ntrain3 project.
A (Name): Helen can you add this user (kazeem). Kazeem please provide your NERSC username.
A (Helen) added you to ntrain3. It may take up to an hour for you to run jobs via -A ntrain3

Q.([redacted]): My username is kazeem
A.(Name): added you to ntrain3
Thank you!

Q (Name): I also have a NERSC account but am not in the ntrain3 project. My username is joywang
A (Lisa): Helen, Can you add joywang to the ntrain3 project?
A (Helen): added joywang to ntrain3
Thank you!

Q ([redacted]): I also have a NERSC account but have not been added to ntrain3, as far as I can tell. I don't see ntrain3 in Iris and I cannot read /global/cfs/cdirs/ntrain3 … my username is [redacted]
A (Lipi): provide your NERSC user name so we can add you :)
A (Helen): I have added you to ntrain3
A (Jordan) Thank you!

Q ([redacted]): Why would you use one MPI implementation over another? I understand for intel mpi you get speed ups on intel architecture but what about all the other implementations?
A (Charles Lively): Different MPI implementations can perform better on different systems depending on the system network speed, chip cache/memory, and shared memory. Some implementations have different algorithms for communication intensive operations, such as collectives.
A(Helen): For example on Perlmutter, the recommended MPI implementation is cray-mpich that is an HPE customized MPI implementation based on MPICH2 and is optimized for the Perlmutter high speed interconnect Slingshot

Q ([redacted]): I have a training account under e913 ending on July 4. Will the end date be extended to July 10?
A (Name): what is your training account name?
A (Tony Se): train451
A (Helen): train451 in ntrain1? I added you to ntrain3 for node reservation for today. The account train451 is valid for ntrain1 until 7/4
A ([redacted]): I'm not entirely sure about the project name, but it was for Julia.
A (Helen): yes, julia used ntrain1. Please use code [redacted] to apply for another training account after 7/4 if you still need Perlmutter access, it expires 7/14.
A ([redacted]): Will do, thank you!

Q ([redacted]): Are there specific advantages/disadvantages that Fortran and C/C++ have over each other for scientific computing, if so what are they?
A (Charles Lively): Fortran and C/C++ are compiled-time languages that are low-level, so they are efficient at quick scientific tasks and formula translations. The use of one over the other can depend on various factors. Fortran is often preferred for pure numerical and scientific computing tasks due to its performance, simplicity, and legacy code base. While C/C++ are chosen for

their flexibility, extensive ecosystem, and support for more complex and diverse programming paradigms.

A ([redacted]): An important difference is the use of pointers in Fortran is required and aliasing is always assumed - whereas in C you can choose to pass by copy to avoid a pointer indirection cost in situations where that matters, and you can also choose whether or not pointer arguments can alias. For the most part, C++ adds a level of structure to C/Fortran that you are almost always required to pay for - even if you don't use it, which will slow it down. The one exception is the use for template meta-programming, which allows for more aggressive optimization by the compiler that isn't possible in C/Fortran. The Eigen C++ linear algebra library is a good example of this.

Thanks!

Q ([redacted]): What's the recommended way to learn Fortran? Is there something like the Rust Book but for Fortran? I already know C.
A (Charles Lively): https://fortran-lang.org/learn/

Q ([redacted]): What are the potential consequences for not calling mpi_finalize? Such as when using exit() or throwing an uncaught error?
A (Charles Lively): There can be a lot of bad consequences. Not calling MPI_Finalize can lead to resource leaks, incomplete communication, inconsistent states, and undefined behavior.

This can result in memory and communication buffer leaks, data loss, deadlocks, and unpredictable program outcomes. Ensuring MPI_Finalize is called in all exit paths mitigates these issues, promoting proper resource release and consistent program termination.

Q (Name): What is the reasoning behind only using one or two CPUs on each node (e.g. in slide 54) instead of just one node using six CPUs?
A (Lipi): Our nodes on Perlmutter only have 2 CPUs per node, so to use more CPUs you must use more nodes. The number of CPUs per node is a function of the system architecture.

Q ([redacted]): If mpi_comm_world is an input variable to mpi_send, it must be able to be something else–meaning it must be possible to have more than just mpi_comm. Why do we need more than one mpi_comm?
A (Name): You can create sub communicators (such as the MPI ranks used in a component of a simulation) other than the MPI_COMM_WORLD (which contains all MPI ranks), In Climate simulation, it is a common practice to have a sub communicator for all MPI ranks on the atmosphere component, another one for all MPI ranks on the ocean component, etc.
([redacted]) thanks

Q ([redacted]): What data types can be sent through MPI? It sounds like the *buf is a pointer to a general array of whatever you want, not necessarily just ints, floats, doubles, etc. Would you be able to send structs or instances of a typedef?

A (Name): Yes actually you can, mpi accepts custom datatypes as well.
https://www.open-mpi.org/doc/v1.5/man3/MPI_Type_create_struct.3.php

Q ([redacted]): How are GPUs and CPUs connected together on Perlmutter, also what about CPU-CPU and GPU-GPU?
A (Name): You can see the architecture of the nodes here:
https://docs.nersc.gov/systems/perlmutter/architecture/
Thank you!

Q ([redacted]): For GPU based MPI are ranks assigned first on CPU then communicated to GPU or is there some other way?
A (Lisa): This is a good article to read to understand how MPI interacts with cuda https://developer.nvidia.com/mpi-solutions-gpus Please let me know if that does not answer your question

Q ([redacted]): Is there any non-blocking MPI receive function?
A (Helen): MPI_IRECV is non-blocking

Q ([redacted]): Do Perlmutter login nodes also double as compute nodes? Or are there dedicated nodes which are just used for logins?
A (Lisa): Login nodes are dedicated notes. When you login to perlmutter you are first on a login node. You can submit a job script from the login node that will be executed on a compute node.

Q ([redacted]): It seems I have not been added to ntrain3 account, even though I see it on IRIS
>$ sacctmgr list user $USER
      User   Def Acct   Admin
---------- ---------- ---------
[redacted]

>$ salloc -A ntrain3 -N 1 -t 00:30:00 --reservation=crash_course
salloc: error: Job request does not match any supported policy.
salloc: error: Job submit/allocate failed: Unspecified error
A (Rebecca): sometimes it takes a while for this to sync to Perlmutter. Please try after 11 am (I believe it is hourly on the hour)
A (helen): It will take up to an hour for the Iris record to propagate to Perlmutter. And when you see ntrain3 appears in your iris command output as above, please also add -C cpu flag in your salloc command to ask for CPU nodes

Q (Name): When already signed up and received login info for a NERSC account , what is the link that I can log back in?
A (Lisa): Are you asking about login to Perlmutter?

Q (Name): For mpi with gpu, does rank mean the entire GPU or is there a way to choose something like streaming multiprocessor?
A (Lisa Claus): A lot of codes are written such that one MPI rank as access to 1 GPU, however you can allow multiple MPI ranks to have access to the same GPUs. When running on Perlmutter you can specify that with flags in your job script, for example --gpus-per-task=1 means 1 GPU is assigned to 1 GPU.

Q ([redacted]): I'm attempting https://github.com/NERSC/crash-course-supercomputing. I have cloned it into `$SCRATCH`. I ran `salloc -A ntrain3 -N 1 -t 00:30:00 --reservation=crash_course`
I then ran `sbatch runall_batch.sh`. However, I see a bunch of errors for each program.
```
srun: error: Job request does not match any supported policy.
srun: error: Unable to allocate resources: Unspecified error
```
A (Rebecca): you'll need to customize runall_batch.sh so that it will work with Perlmutter. And you would run it as a batch job, not within an interactive job (salloc gives you an interactive job on compute nodes that you can use interactively, but sbatch is something where you just put the job into the queue and walk away and then come back later when it's finished).

Q ([redacted]): Let's say I submit a MPI job that needs two nodes at Perlmutter. Each node has two cpus, each cpus has 128 cores. In this case, is the number of MPI processes to be 512? Or just 4?
A (Rebecca): You may notice in the course I say "MPI process" and not "processor". That is because you can run multiple processes on a single processor. You can choose to run up to 512 MPI processes across your two Perlmutter nodes. Or, you could run a total of 2 processes, one on each node. Or, anything in between. If your code needs a lot of memory for each MPI process, you might choose to run fewer MPI processes per node.

Q ([redacted]: In the deadlocking example, do we have to manually divide the processes into 2 groups, 3 groups, 4 groups, …? Or would MPI_Isend automatically assign groups to group the processes in?
OK, thank you!
A (Name): Answered live

Q (Name): what is MPI_COMM_WORLD? It seems like an environmental variable–how do I pass this into my code?
A (Name): it is automatically defined in MPI and no action is required on your part. It is defined in the file mpi.h (which we include at the top of our code file)

Q (Name): ([redacted]) Is anyone else getting an auth error "Authentication failed for 'https://github.com/NERSC/crash-course-superc/'
train580@perlmutter:login38:/pscratch/sd/t/train580>"
A (Suzan): Yes i have issue with Google Auth
A (Rebecca): the line is cut off, should end with crash-course-supercomputing.git

git clone https://github.com/NERSC/crash-course-supercomputing.git
Thank you

Q ([redacted]): I just joined the training recently. I have a NERSC account already. When logging in, it is asking for my password + OTP. Where can I find the OTP as a first-time user? I don't think I have, definitely have not used auth for NERSC before. Understood, doing it now. Thank you for your help!
A (Name): you need to set up your one-time password (OTP) token in Iris (iris.nersc.gov). Have you done that? You will need a phone and the google authenticator app. The instructions are here: https://docs.nersc.gov/connect/mfa/#configuring-and-using-an-mfa-token
A ([redacted]): Successfully logged in, thank you once again!

Q (Name): I have a NERSC account and my MFA. However, when I entered in my password + OTP I was still unable to login. Should I contact IT about this or could this be an error associated with the fact that I have never logged in before? It worked, thank you for your help!
A (Name): Joy, you need to set up MFA. Looking at your account you don't have it set up. Please see the link above on how to do that.
(https://docs.nersc.gov/connect/mfa/#configuring-and-using-an-mfa-token)

Q ([redacted]): Is the solution (to the first exercise) stored somewhere so I can check myself? Thank you!
A (Helen): **dart-mpi.c** or dart-mpi.f90 is the solution. Thanks!

Q ([redacted]): Maybe I missed it, but I can't get the dart-mpi example to compile. In a default Perlmutter terminal, trying to get doing `make darts-mpi` in the Fortan folder:

```
ftn -g -Wall -Wextra -c darts-mpi.f90
Error invoking pkg-config!
Package mpichf90 was not found in the pkg-config search path.
Perhaps you should add the directory containing `mpichf90.pc'
to the PKG_CONFIG_PATH environment variable
No package 'mpichf90' found
make: *** [Makefile:41: darts-mpi.o] Error 1
```

I think I have the right modules?

```
craype-x86-milan
libfabric/1.15.2.0
craype-network-ofi
xpmem/2.6.2-2.5_2.38__gd067c3f.shasta
PrgEnv-gnu/8.5.0
cray-dsmml/0.2.2
craype/2.7.30
```

```
perftools-base/23.12.0
cpe/23.12
cudatoolkit/12.2
craype-accel-nvidia80
gpu/1.0
gcc/12.2.0
```

A (Rebecca): I do not see that you have the cray-mpich module loaded (which is necessary). Here is what I have loaded (the defaults – I have no customizations in my account to enable me to better diagnose user issues):
Currently Loaded Modules:

1) craype-x86-milan	6) cray-dsmml/0.2.2	11) perftools-base/23.12.0
2) libfabric/1.15.2.0	7) cray-libsci/23.12.5	12) cpe/23.12
3) craype-network-ofi	8) cray-mpich/8.1.28	13) cudatoolkit/12.2
4) xpmem/2.6.2-2.5_2.38__gd067c3f.shasta	9) craype/2.7.30	14) craype-accel-nvidia80
5) PrgEnv-gnu/8.5.0	10) gcc-native/12.3	15) gpu/1.0

If you don't have that and don't know how to get to this point, just log out and log back in, and as long as there's nothing going on in your .bashrc file that loads/unloads modules, you should be back to the defaults.

Q ([redacted]): My parallel script seems to be running in serial despite my MPI adjustments, only seeing 1 rank. Do I need to copy mpi.h somewhere or is something else to blame?
A ([redacted]): Fixed by allocating compute node and passing # ranks to srun -n

Q ([redacted]): what is displs in gather?
A ([redacted].): It says in the slides: "entry i in displs array specifies displacement relative to recvbuf[0] at which to place data from corresponding process number"

Q ([redacted]): Is there a cut-and-paste list somewhere of recommended compiler commands / submission commands for using MPI for this course that we are supposed to access for those of us who haven't used Perlmutter (or MPI) before today? The course slides talk about using "--reservation=hpc_course -A ntrain3 -C cpu " for sbatch or salloc sessions. In the questions above, people are using "salloc -A ntrain3 -N 1 -t 00:30:00 --reservation=crash_course". I've figured out how to compile (I think) but don't understand what the commands are to actually submit a job]using the compiled code. It's not included in the slide-deck, that I can see. Thanks!
A (S[redacted]): Partially answering my own question [**but thanks so much for your help, Lisa!**], in that I think there was an expectation attendees had gone through the New User training. I was able to find slide-decks on compilation (04-PE-and-Compilation-2023-v2-1.pdf (nersc.gov)) and on submitting jobs (05-Running-Jobs-2023.pdf (nersc.gov)). Below is what I used to compile and run on a compute node.
To compile, I used the command: cc -o darts-mpi.x darts-mpi.c

To submit the job: salloc --reservation=crash_course -A ntrain3 -C cpu -N 1
Then I used: ./darts-mpi.x
Once it finished I typed exit to get back to the login node.

A (Lisa): After you compile your code you either submit a job through a job script or an interactive session.
The compilation line above looks correct to me.
salloc just gives you access to the compute node. Let me find the documentation page. https://docs.nersc.gov/jobs/#salloc that explains the salloc command you need to type srun after you are on a compute node.
As an alternative you can submit a job from the login node to run on a compute node with the sbatch command https://docs.nersc.gov/jobs/#sbatch

So one would need to create darts-mpi.sh and then run sbatch darts-mpi.sh YES!
Haha thank you.. And to make the sh it is literally a make command (I don't know what you mean by that) Here is an example for a .sh script that one would submit with sbatch: https://docs.nersc.gov/systems/perlmutter/running-jobs/#1-node-1-task-1-gpu

[redacted] again - I think what this answer shows is that it would be useful for the instructors to actually walk the class through a compilation and submission process since not everyone is familiar with it. Really appreciate your help, Lisa!

Yes, that makes sense, this will come later, I believe after the OMP introduction.

Thank you!!

Q ([redacted]): When I print out the comm size @ np it always gives 1, it does change it from 0 so it sees it but it only sees 1 subprocess. How do I ensure that multiple run, is there something I input with command line args/Makefile? I also have a similar issue w/ comm size when running the compiled darts-mpi.o code

```
int main(int argc, char **argv){
        // start mpi
        MPI_Status status;
        MPI_Init(&argc, &argv);
        int rec, me, np, q, sendto = 0;
        // get mpi size & rank
        MPI_Comm_size(MPI_COMM_WORLD, &np);
        MPI_Comm_rank(MPI_COMM_WORLD, &me);
        printf("comm size %d\n",np);
        printf("use %d\n",me);
```

A (Lisa): How did you submit your job?
([redacted]) I just ran the code using ./darts_out.o, running srun states that it is unable to allocate resources

(Lisa I need some more information to be able to help you figure it out. Did you allocate a compute node with salloc? If so what is the exact command that you used (sometthing like *salloc -N 1 -q interactive -t 00:30:00 -C cpu)* ?

A (Rebecca): in order to run on multiple processes you need to use “srun” when you invoke your executable (e.g., srun -n 2 ./a.out). In order to be able to run srun, you need to be on a compute node. There are two ways to get access to a compute node:

1. Through submitting a batch script to slurm, which then runs your job while you go do other things. (sbatch myscript.sl)
2. Through an interactive job, which you invoke with the “salloc” command (e.g., salloc -q interactive -C cpu -N 1 -t 30:00 -A ntrain3), which finds you some nodes that you can run on. Once those nodes are ready, then you can run commands like “srun” for the period of time that you have access to those compute nodes.

Thanks! I got it to work using salloc

Q (Name): Will we get the recorded video?
A (Name): All resources/videos are available/will be available on the Event Web Page: https://www.nersc.gov/users/training/events/2024/hpc-crash-course-jun2024/

Q ([redacted]): my user name is[redacted]. Can you please add me to the training?
A (Name): looks like you have been added.

Q ([redacted]): if parallel for goes over n iterations, will it spawn n threads by default?
A (Rebecca): the default number of threads is based on some default settings of environment variables on your compute system. Probably, the default number of threads is 1, unless you set it to be something else. You can use the environment variable OMP_NUM_THREADS to tell your code how many threads to spawn.
Alfred Tang: Thanks I think I remember now. What is OMP_NUM_THREADS if there are multiple loops?
A (Rebecca): for every loop, unless you change it internally (and you can with a function called omp_set_num_threads(...) but I would not recommend doing that), then whatever you set via the OMP_NUM_THREADS variable at runtime is the number of threads that will be forked for each loop that is parallelized with OpenMP in your code.
Alfred: Thanks. It makes sense. OMP_NUM_THREADS should probably be set by the limitation of hardware and not software.
A (Rebecca): that’s right. It should be based on what is possible with the hardware that is being used to run your job.
Alfred: Great. Thanks.

Q ([redacted]): There are some differences between presenter slides and slides posted on the event page (for Helen and Charles)
A (Rebecca): we were making some tweaks up until the presentation. We’ll update the uploads after the event. Thanks for pointing this out!

Q ([redacted]): I thought the idea of openMP is to spawn threads in GPUs. Why does Helen's example refers to CPU only?
A (Name): OpenMP can be used for offload of work to GPUs, but its original purpose was for multicore CPUs. We are just focused on CPUs today.
Alfred: Got it.

Q (Name): For "srun -4 -c 64 –cpu-bind=cores", we have 4 mpi processes, and each mpi process uses 64 cores, am I right?
A (Name): yes, that is basically right. I would say 64 "logical cores" for 100% correctness. The way I like to think of it is like sitting in a movie theater. There are 256 seats in the movie theater (but actually in order to be comfortable, each person needs 2 seats). You can think of this as 4 families that want to go to the movie theater and each family has a section of the theater where their family members can sit (because they don't want to be crowded in, they want to have enough space to spread out, and they don't want to sit next to strangers).

Q ([redacted]): Since we have the reservation only for today, but our training accounts are active until July x, does that mean after today we can/should still use commands like:
salloc -N 1 -C cpu -t 30:00 -A ntrain3 -q interactive
A (Name): Looks like we should, according to the training chat

Q ([redacted]): When using these parallel paradigms for scientific computing, its typical that other libraries like BLAS, LAPACK and ScaLAPCK are also used. Does anyone have any good recommendations for tutorials, courses, or books that would assist in understanding those libraries and how to use them?
A (Name):

Q ([redacted]): Does Helen's book delve into hybrid programming?
A (Name): Hybrid MPI/OpenMP is not covered in the book