Research Computing University of Colorado Boulder
1 Courses using RC Resources 3
2 Acknowledging RC 5
Documentation covering the use of Research Computing resources.

Here are some quick links into the documentation to get you started.

- Logging In
- Research Computing Filesystems
- Compiling Software
- Batch Jobs
- The Module System
- Frequently Asked Questions (FAQ)

More information is available at https://www.colorado.edu/rc.

If you have any questions, please contact rc-help@colorado.edu.
Courses using RC Resources

Students are welcome to use RC resources on their own for class projects and can request access as a regular UCB affiliate via the link off the RC homepage at: https://www.colorado.edu/rc. When requesting help please indicate that the work is for a class project and any deadlines. If students are to be required to use RC resources for a class, see below.

Instructors who wish to lead a class using RC resources must contact us at rc-help@colorado.edu before the class begins. This is to insure that our resources can meet your needs and if adequate resources and support are available. Early in the process we will need to know details about the proposed class usage such as:

- Number of students
- Software needed, and if it will be installed by instructor/TA
- Typical computational work (number of jobs or sessions, length, number of CPUs)
- Date if 1st usage in class/lab
- Class roster including TAs and auditors.
Use of University of Colorado Research Computing resources, including (but not limited to) the Janus and Summit supercomputers, the Blanca Condo Cluster, and the PetaLibrary data storage service must be acknowledged in any and all publications.

**Acknowledging Summit:** “This work utilized the Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University.”

**Acknowledging PetaLibrary:** “Data storage supported by the University of Colorado Boulder ‘PetaLibrary’”

### 2.1 Frequently Asked Questions

See our documentation [homepage](#) for information about our most common topics.

1. *I have a new phone. How do I move my Duo onto it?*
2. *How do I check how full my Summit directories are?*
3. *When will my job start?*
4. *How can I get system metrics?*
5. *How much memory did my job use?*
6. *Where is my current fair share priority level at?*
7. *Why is my job pending with reason ‘ReqNodeNotAvail’?*
9. *How can I check what allocations I belong to?*
10. Why do I get the following ‘LMOD’ error when I try to load slurm/summit?: Lmod has detected the following error: The following module(s) are unknown: "slurm/summit"

11. How do I install my own python library?

2.1.1 I have a new phone. How do I move my Duo onto it?

You can add a new device to your duo account by visiting https://duo.colorado.edu. After a CU authorization page you will be directed to a Duo authentication page. Ignore the Duo Push prompt and instead click “Settings”:

In this settings side bar click “Add a new device”: 
Duo will then try to authenticate your account by push notification to verify your identity. Cancel this push notification...

... and click on “Enter a Passcode”, or “Call Me”.

- If you select “Call Me” the simply receive the call and press 1.
- If you select “Enter a passcode” then click “Text me new codes” and you will be sent a list of one time passwords. Type in any one of the codes and you will be authenticated. Once you have verified your identity, follow the instructions provided by Duo to add your device.

If you cannot authenticate your account, contact rc-help@colorado.edu for further assistance.

2.1.2 How do I check how full my Summit directories are?

You have three directories allocated to your username ($USER). These include /home/$USER (2 G), /projects/$USER (250 G) and /scratch/summit/$USER (10 T). To see how much space you’ve used in each, from a Summit ‘scompile’ node, type curc-quota as follows:

```bash
[user@shas0100 ~]$ curc-quota
```

(continues on next page)
You can also check the amount of space being used by any directory with the `du -h` command:

```
[janedoe@shas0136 ~]$ du -h /scratch/summit/janedoe/WRF
698M  WRF/run
698M   WRF
```

### 2.1.3 When will my job start?

You can pull up information on your job’s start time using the `squeue` command:

```
squeue --user=your_rc-username --start
```

Note that Slurm’s estimated start time can be a bit inaccurate. This is because Slurm calculates this estimation off the jobs that are currently running or queued in the system. Any job that is added in later with a higher priority may delay your job.

For more information on the `squeue` command, take a look at our Useful Slurm Commands tutorial. Or visit the Slurm page on `squeue`.

Note that you can also see system level wait times and how they change through time by visiting the CURC metrics portal at https://xdmod.rc.colorado.edu

### 2.1.4 How can I get metrics about CURC systems such as how busy they are, wait times, and account usage?

Visit the CURC metrics portal at https://xdmod.rc.colorado.edu

### 2.1.5 How much memory did my job use?

You can check how much memory your job used by using the `sacct` command. Simply replace `YYYY-MM-DD` with the date you ran the job:

```
sacct --starttime=YYYY-MM-DD --jobs=your_job-id --format=User,JobName,JobId,MaxRSS
```

If you’d like to monitor memory usage on jobs that are currently running, use the `sstat` command:

```
sstat --jobs=your_job-id --format=User,JobName,JobId,MaxRSS
```

For more information on `sstat` or `sacct` commands, take a look at our Useful Slurm Commands tutorial. Or visit the Slurm reference pages on `sstat` and `sacct`.

### 2.1.6 How can I see my current FairShare priority?

You can check your current fair share priority level using the `sshare` command:
The `sshare` command will print out a table of information regarding your usage and priority on all allocations. The `-U` flag will specify the current user and the `-l` flag will print out more details in the table. The field we are looking for is the `LevelFS`. The LevelFS holds a number from 0 to infinity that describes the fair share of an association in relation to its other siblings in an account. Over serviced accounts will have a LevelFS that’s between 0 and 1. Under serviced accounts will have a LevelFS that’s greater than 1. Accounts that haven’t run any jobs will have a LevelFS of infinity (inf).

For more information on fair share the `sshare` command, take a look at Slurm’s documentation on fair share or check out the Slurm reference page on sshare.

### 2.1.7 Why is my job pending with reason ‘ReqNodeNotAvail’?

The ‘ReqNodeNotAvail’ message usually means that your node has been reserved for maintenance during the period you have requested within your job script. This message often occurs in the days leading up to our regularly scheduled maintenance, which is performed the first Wednesday of every month. So, for example, if you run a job with a 72 hour wall clock request on the first Monday of the month, you will receive the ‘ReqNodeNotAvail’ error because the node is reserved for maintenance within that 72-hour window. You can confirm whether the requested node has a reservation by typing `scontrol show reservation` to list all active reservations.

If you receive this message, the following solutions are available: 1) run a shorter job that does not intersect the maintenance window; or 2) wait until after maintenance.

### 2.1.8 Why do I get an ‘Invalid Partition’ error when I try to run a job?

This error usually means users do not have an allocation that would provide the service units (SUs) required to run a job. This can occur if a user has no valid allocation, specifies an invalid allocation, or specifies an invalid partition. Think of SUs as “HPC currency”: you need an allocation of SUs to use the system. Allocations are free. New CU users should automatically get added to a ‘ucb-general’ allocation upon account creation which will provide a modest allocation of SUs for running small jobs and testing/benchmarking codes. However, if this allocation expires and you do not have a new one you will see this error. ‘ucb-general’ allocations are intended for benchmarking and testing, and it is expected that users will move to a project allocation. To request a Project and apply for a Project Allocation visit our allocation site.

### 2.1.9 How can I check what allocations I belong to?

You can check the allocations you belong to with the `sacctmgr` command. Simply type:

```
sacctmgr -p show associations user=$USER
```

...from a login or compile node. This will print out an assortment of information including allocations and QoS available to you. For more information on sacctmgr, check out the Slurm’s documentation.

### 2.1.10 Why do I get an ‘LMOD’ error when I try to load Slurm?

The slurm/summit module environment can not be loaded from compile or compute nodes. It should only be loaded from login nodes when attempting to switch between Blanca and Summit environments. This error can be disregarded, as no harm is done.

---

# 2.1. Frequently Asked Questions
2.1.11 How do I install my own python library?

Although Research Computing provides commonly used Python libraries as module, you may need to install individual python libraries for your research. This is best handled by utilizing Research Computing’s Anaconda installation to set up a local Conda environment.

Find out more about using Python with Anaconda here.

2.2 Logging In

Research Computing offers a variety of resources for researchers to use in their own projects. To get started with Research Computing resources you need the following:

- A Research Computing account
- Duo 2 factor authentication
- The PuTTY application (if you are a Windows user)

Users accessing RC’s resources will be connected to a login node. A login node is an outward facing node within the Research Computing environment that users can connect to from their local machines. Once on a login node, users can perform a limited number of tasks:

- Edit files
- Transfer Data
- Running Jobs
- Access storage resources

Note that the login nodes should not be used for resource-intensive tasks such as running code. For all other tasks, users should run batch jobs, interactive jobs, or use the compile nodes.

- For CSU users, please refer to the CSU login guide.
- For RMACC users, please refer to RMACC Access to Summit.

2.2.1 Getting an account

Although Research Computing resources are free and available to the CU Boulder research community, they are also managed resources with large value and high demand. Therefore, Research Computing mandates that each user obtain a Research Computing account. Please note that Research Computing accounts are separate from your CU accounts. You must sign up for an individual Research Computing account if you wish to access Research Computing resources.

A Research Computing account can be secured quickly and easily by filling out the form here. Once you’ve received an account with Research Computing, you will automatically be sent an invitation for Duo enrollment via email. After you have accepted the Duo invitation, you will be able to authenticate via Duo and log into the system.

2.2.2 Logging in from a Windows Machine

Logging in from a Windows machine requires the additional step of installing the PuTTY ssh client onto your local machine. This application allows users to connect to remote servers with the ssh protocol. Note that there are other ssh clients that allow Windows machines to connect to remote ssh servers; Research Computing recommends PuTTY for reliability and simplicity.
1. Open the PuTTY application on your computer
   - Under “Host Name (or IP address)”, enter login.rc.colorado.edu. Select “SSH” as the connection type. Click on “Open”.
2. Enter your Identikiy in response to the “login as” prompt
3. When prompted to enter your password:
   - If you are logging in using Duo Push, simply type your Identikiy password. You will then receive an authentication request from the Duo app on your phone. Approve the request.
   - If you are using Duo SMS, Phone Call, or Token login methods, instructions can be found here.
   - Note that as a security feature, PuTTY does not display any text while you type your password

### 2.2.3 Logging in from a Mac

Logging in with a Mac requires no extra installation on your local machine. Simply utilize the terminal application that is pre-installed with your operating system to access Research Computing resources.

1. Under “File”, open a new finder window. Navigate to the “Applications” folder, then the “Utilities” folder. Open a terminal window and type `ssh username@login.rc.colorado.edu`, where `username` is your assigned username. Press enter.
2. Enter your password:
   - If you are logging in using Duo Push, type your Identikiy password. You will then receive an authentication request on the Duo app on your phone. Approve the request.
   - If you are using Duo SMS, Phone Call, or Token login methods, instructions can be found here.

### 2.2.4 Logging in from Linux

Much like with Macs, Linux machines require no additional setup to access Research Computing resources. Simply utilize the your Linux terminal to access Research Computing resources.

1. Open a terminal window from your application menu and type `ssh username@login.rc.colorado.edu`, where `username` is your research computing username.
2. Enter your password:
   - If you are logging in using Duo Push, simply type your Identikiy password. You will then receive an authentication request on the Duo app on your phone. Approve the request.
   - If you are using Duo SMS, Phone Call, or Token login methods, instructions can be found here.

### 2.2.5 SSH host keys

The first time you log into an RC login node you will be asked to verify the host key. You can refer to the keys published here to confirm that you are connecting to a valid RC login node.

Note that each login node may support more than one type of key, but only one is used (or displayed) by your client at any given time.
2.3 Duo 2-factor Authentication

Research Computing utilizes a 2-factor authentication utility called **Duo** that must be utilized in order to log into RC resources. **This document covers installing and logging in with Duo for CU Boulder users only.**

- For CSU Users please see CSU’s documentation on Duo 2-factor authentication
- For RMACC Users please see XSEDE’s documentation on Duo 2-factor authentication
2.3.1 Setting up Duo

1. You will receive a Duo invitation when your RC account is created. Open your Duo invitation email (from Duo security or duo.com) on a desktop/laptop computer.
   - Within one business day you should receive a duo invitation. If you didn’t receive an invitation or your invitation has expired, then please contact rc-help@colorado.edu for a new invitation.

2. Login with your CU identikey and follow the instructions to setup
   - Research Computing recommends users utilize Duo Push as Duo push, install the Duo Mobile app and link your account with your device. Duo

3. Login to RC Resources via ssh as described below.

Common Issues

A few common issues users will come across when using Duo include:
- A Duo Invite email may be sent to your Spam folder.
- Requesting a phone call if you want to use the Push app for authentication.
- A Duo account purged after remaining unused for 6-9 months.
- Having a new device and want to move Duo onto it.
   - (Check out our FAQ for detailed instructions on accomplishing this.)[https://curc.readthedocs.io/en/latest/faq.html#i-have-a-new-phone-how-do-i-move-my-duo-onto-it]

Please contact us at rc-help@colorado.edu if you encounter these or any other issues regarding Duo.

2.3.2 Logging in with Duo

Duo offers a variety of methods to log into your account. Depending on what you select when setting up your Duo account, you should have several different methods of 2-factor authentication when logging into RC Resources.

Using Duo Push

Duo Push is Research Computing’s recommended method of 2-factor authentication. Because Duo Push is tied to your physical smart device instead of a phone number or account, Duo Push is provides a more secure method of 2-factor authentication than either SMS or phone call.

1. Type: `ssh <username>@login.rc.colorado.edu` into the command line. For example, if my username is jodo2018 I would type `ssh jodo2018@login.rc.colorado.edu`

2. Enter your identikey password when prompted.

3. Wait for a push to your phone.

Note: Duo push is tied to your device so **so you will need to add or remove your device if you get a new phone.**

Using Duo SMS

If you prefer to not use the Duo app or if you don’t have a smart device, then Duo offers an SMS method of 2-factor authentication:
1. `ssh username@login.rc.colorado.edu`. For example, if my username is jodo2018 I would type `ssh jodo2018@login.rc.colorado.edu`

2. Enter your password when prompted, as `mypasswd,sms`. For example, if my password is Ih3artdu0 I would type: `Ih3artdu0,sms`

3. A list of one-time-passwords (OTPs) will be texted to you via SMS. Another login prompt will appear on your screen. Using the first OTP in the list, login with your password as `mypasswd,OTP`. For example, if my password is Ih3artdu0 and my OTP is 330456 I would type (without quotes): `Ih3artdu0,330456`.

4. Note that the next time you login, you can either request a new list of OTPs using Step 2 and then enter the first OTP via Step 3, or you can just use the next OTP in the list, and skip directly to Step 3.

**Using Duo Phone Call**

Duo also provides a phone call solution for 2-factor authorization if you only have a land line, or prefer to not use Push or SMS:

1. Type `ssh username@login.rc.colorado.edu` into the command line. For example, if my username is jodo2018 I would type `ssh jodo2018@login.rc.colorado.edu`

2. Enter your password when prompted, as `mypasswd,phone`. For example, if my password is Ih3artdu0 I would type: `Ih3artdu0,phone`

3. Wait for a phone call. Answer the call, select option #1, and you will automatically be logged in.

**Using the Duo Token**

If you prefer a physical device, similar to the Vasco OTP. You have to arrange with RC staff to acquire a token. There is a small fee for this device. Email rc-help@colorado.edu for assistance.

1. `ssh username@login.rc.colorado.edu`. For example, if my username is jodo2018 I would type `ssh jodo2018@login.rc.colorado.edu`

2. Enter your password when prompted, as `mypasswd,6-digit-number`. The 6-digit number is given to you on the Duo token. For example, if my password is Ih3artdu0, and I press the button on the token and it is 123456, I would type: `Ih3artdu0,123456`

**2.3.3 Managing Duo Devices**

Users can manage their own Duo devices by visiting https://duo.colorado.edu. This can allow users to add or remove activated devices at their discretion.

After a CU authorization page, you will be directed to a Duo authentication page. Do not respond to the Push notification and instead click the “Settings” button:
This will bring up a menu that provides several options on device management and general help.
Clicking “Add a new device” will allow you to add a new smart phone, tablet, or land-line. Simply select the option you wish to add and follow the steps provided by Duo complete setting up your new device.

Selecting “My Settings & Devices” provides a more detailed list of all devices you have registered. From here you can also add a new device, set your default device, and change your default authentication method when you attempt to log in.

2.4 RMACC Access to Summit

If you are at an RMACC institution other than CU Boulder or CSU and would like to use Summit, please see the below steps to gain access.

2.4.1 Getting an XSEDE account

Visit the XSEDE User Portal and use the “Create Account” button. Complete the form and follow the instructions to create an XSEDE account. As part of the process, you will select an XSEDE username, which will be used to access the XSEDE User Portal and the XSEDE SSO Hub.
2.4.2 Configuring Duo

Once your XSEDE account has been created, follow the Multi-Factor Authentication setup instructions on the XSEDE website. Multi-factor authentication with Duo is required for access to the XSEDE SSO Hub, which provides access to RMACC Summit.

2.4.3 Getting authorization for RMACC Summit

Send an email from your institutional email address to rc-help@colorado.edu requesting access to RMACC Summit via XSEDE. Include the full name of your home institution, specific department within the institution, a short summary
of why you require access to RMACC Summit, and your XSEDE username. You will receive a response indicating whether your account has been added to the rmacc-summit.colorado.xsede.org project.

### 2.4.4 Logging in to CU Boulder RC using the XSEDE SSO Hub

Use an ssh client to connect to the XSEDE SSO Hub using your XSEDE credentials. You will be prompted to authenticate both using your XSEDE password and via Duo.

```bash
ssh -l <your-xsede-username> login.xsede.org
```

From there, use gsissh (available in the SSO Hub environment) to log into rmacc-summit, which serves as an alias for the CU Boulder RC login environment.

```bash
gsissh rmacc-summit
```

### 2.5 Blanca

CU Research Computing operates a shared “condo” compute cluster, named Blanca, which consists of nodes owned by individual research groups or departments. Condo partners get significantly prioritized access on nodes that they own and can run jobs on any nodes that are not currently in use by other partners.

An allocation of CPU time is not needed in order to run on Blanca.

If you would like to purchase a Blanca node, please visit the Research Computing website for more details.

#### 2.5.1 Blanca Quick-Start

1. If your group is a Blanca partner, ask your PI or PoC to send an email to rc-help@colorado.edu requesting access for you to their high-priority queue.
2. From a login node, run “module load slurm/blanca” to access the Slurm job scheduler instance for Blanca.
3. Consult the Table and the Examples section below to learn how to direct your jobs to the appropriate compute nodes.
4. If needed, compile your application on the appropriate compute node type.

#### 2.5.2 Job Scheduling

All jobs are run through a batch/queue system. Interactive jobs on compute nodes are allowed but these must be initiated through the scheduler. Each partner group has its own high-priority QoS (analogous to a queue) for jobs that will run on nodes that it has purchased. High-priority jobs move to the top of the queue and are thus guaranteed to start running within a few minutes, unless other high-priority jobs are already queued or running ahead of them. High-priority jobs can run for a maximum wall time of 7 days. All partners also have access to a low-priority preemptable QoS that can run on any Blanca nodes that are not already in use by their owners. Low-priority jobs have a maximum wall time of 24 hours.

Blanca uses a separate instance of the Slurm scheduling system from the other RC compute resources. You can use Blanca’s Slurm instance by loading a special module on a login node: “module load slurm/blanca”.

More details about how to use Slurm can be found here.
2.5.3 QoS

Slurm on Blanca uses “Quality of Service”, or QoS, to classify jobs for scheduling. A QoS in this case is analogous to a “queue” in other scheduling systems. Each partner group has its own high-priority QoS called \texttt{blanca-<group identifier>} and can also use the condo-wide low-priority QoS, which is called \texttt{preemptable}.

If you are a new Blanca user, ask your PI or Point of Contact person to request access for you to your group’s high-priority QoS; requests should be made via email to \texttt{rc-help@colorado.edu}. You are only allowed to use a high-priority QoS if you have been added as a member of it, and you can only use the low-priority preemptable QoS if you are also a member of a high-priority QoS. Your PI may also be able to point you to group-specific documentation regarding Blanca.

2.5.4 Node-QoS-Features

Since not all Blanca nodes are identical, you can include node features in your job requests to help the scheduler determine which nodes are appropriate for your jobs to run on when you are using the preemptable QoS.

To determine which nodes exist on the system, type \texttt{scontrol show nodes} to get a list.

2.5.5 Node Features Tables

Blanca Core

2.5.6 Description of features

- \texttt{westmere-ex}: Intel processor generation
- \texttt{sandybridge}: Intel processor generation
- \texttt{ivybridge}: Intel processor generation
- \texttt{haswell}: Intel processor generation
- \texttt{broadwell}: Intel processor generation
- \texttt{skylake}: Intel processor generation
- \texttt{cascade}: Intel processor generation
- \texttt{avx}: AVX processor instruction set
- \texttt{avx2}: AVX2 processor instruction set
- \texttt{fdr}: InfiniBand network generation
- \texttt{edr}: InfiniBand network generation
- \texttt{Quadro}: NVIDIA GPU generation
- \texttt{Tesla}: NVIDIA GPU generation
- \texttt{k2000}: NVIDIA K2000 GPU
- \texttt{P100}: NVIDIA P100 GPU
- \texttt{localraid}: large, fast RAID disk storage in node
- \texttt{rhel7}: RedHat Enterprise Linux version 7 operating system
2.5.7 Examples

Here are examples of Slurm directives that can be used in your batch scripts in order to meet certain job requirements. Note that the “constraint” directive constrains a job to run only on nodes with the corresponding feature.

1. To run a 32-core job for 36 hours on a single blanca-ics node:

```
#SBATCH --qos=blanca-ics
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=32
#SBATCH --time=36:00:00
```

1. To run a 56-core job across two blanca-sha nodes for seven days:

```
#SBATCH --qos=blanca-sha
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=28
#SBATCH --time=7-00:00:00
#SBATCH --export=NONE
```

1. To run an 8-core job in the low-priority QoS on any node that has broadwell processors and uses the RHEL 7 operating system:

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=8
#SBATCH --time=4:00:00
#SBATCH --export=NONE
#SBATCH --constraint="broadwell&rhel7"
```

1. To run an 8-core job in the low-priority QoS on any node that has either the AVX or AVX2 instruction set:

```
#SBATCH --qos=preemptable
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=8
#SBATCH --time=4:00:00
#SBATCH --export=NONE
#SBATCH --constraint="avx|avx2"
```

1. To start a 2-hr interactive job on one core on a blanca-ceae node, run this at the command line:

```
sinteractive --qos=blanca-ceae --export=NONE --time=02:00:00
```

Note that the interactive job won’t start until the resources that it needs are available, so you may not get a command prompt on your assigned compute node immediately.

2.5.8 Important notes

1. To see what modules are available, start an interactive job on a compute node and use `module avail` or `module spider` on it.

2. `/home`, `/projects`, and `/pl/active (PetaLibrary Active)` are available on all Blanca nodes. Scratch I/O can be written to `/rc_scratch`, which should offer much better performance than `/projects`. All Blanca nodes also have at least 850 GB of scratch space on a local disk, `/local/scratch`. For more info on the different RC storage spaces, please see our page on storage.

2.5. Blanca
3. There are no dedicated Blanca compile nodes. To build software that will run on Blanca, start an interactive job on a node like the one on which you expect your jobs to run, and compile your software there. Do not compile on the login nodes!

4. Multi-node MPI jobs that do a lot of inter-process communication do not run well on most standard Blanca nodes. Nodes equipped with specialty fabrics like Blanca CCN or any node on Blanca HPC can run MPI application much more efficiently.

2.5.9 Blanca Preemptable QOS

*(effective 2018-03-01)* Each partner group has its own high-priority QoS (“blanca-”) for jobs that will run on nodes that it has contributed. High-priority jobs can run for up to 7 days. All partners also have access to a low-priority QoS (“preemptable”) that can run on any Blanca nodes that are not already in use by the partners who contributed them. Low-priority jobs will have a maximum time limit of 24 hours, and can be preempted at any time by high-priority jobs that request the same compute resources being used by the low-priority job. The preemption process will terminate the low-priority job with a grace period of up to 120-seconds. Preempted low-priority jobs will then be requeued by default. Additional details follow.

**Usage**

To specify the preemptable QoS in a job script:

```
#SBATCH --QoS=preemptable
```

To specify the preemptable QoS for an interactive job:

```
$ sinteractive --qos=preemptable <other_arguments>
```

Batch jobs that are preempted will automatically requeue if the exit code is non-zero. (It will be non-zero in most cases.) If you would prefer that jobs not requeue, specify:

```
#SBATCH --no-requeue
```

Interactive jobs will not requeue if preempted.

**Best practices**

Checkpointing: Given that preemptable jobs can request wall times up to 24 hours in duration, there is the possibility that users may lose results if they do not checkpoint. Checkpointing is the practice of incrementally saving computed results such that – if a job is preempted, killed, canceled or crashes – a given software package or model can continue from the most recent checkpoint in a subsequent job, rather than starting over from the beginning. For example, if a user implements hourly checkpointing and their 24 hour simulation job is preempted after 22.5 hours, they will be able to continue their simulation from the most recent checkpoint data that was written out at 22 hours, rather than starting over. Checkpointing is an application-dependent process, not something that can be automated on the system end; many popular software packages have checkpointing built in (e.g., ‘restart’ files). In summary, users of the preemptable QoS should implement checkpointing if at all possible to ensure they can pick up where they left off in the event their job is preempted.

Requeuing: Users running jobs that do not require requeuing if preempted should specify the `--no-requeue` flag noted above to avoid unnecessary use of compute resources.
Example Job Scripts

Run a 6-hour preemptable python job on 32 cores without specifying a partition (job will run on any available compute partitions on Blanca, regardless of features, so long as they have at least 16 cores each).

```
#!/bin/bash
#SBATCH --time=06:00:00
#SBATCH --qos=preemptable
#SBATCH --job-name=test
#SBATCH --nodes=2
#SBATCH --ntasks=32
#SBATCH --output=test.%j.out

module purge
module load python

python myscript.py
```

Same as Example 1, but specify a specific partition (‘blanca-ccn’) (job will only run on blanca-ccn nodes)

```
#!/bin/bash
#SBATCH --time=06:00:00
#SBATCH --qos=preemptable
#SBATCH --partition=blanca-ccn
#SBATCH --job-name=test
#SBATCH --nodes=2
#SBATCH --ntasks=32
#SBATCH --output=test.%j.out

module purge
module load python

python myscript.py
```

Same as Example 1, but specify desired node features, in this case the avx2 instruction set and Redhat V7 OS (job will run on any node meeting these feature requirements, and which has at least 16 cores per node).

```
#!/bin/bash
#SBATCH --time=06:00:00
#SBATCH --qos=preemptable
#SBATCH --constraint="avx2&rhel7"
#SBATCH --job-name=test
#SBATCH --nodes=2
#SBATCH --ntasks=32
#SBATCH --output=test.%j.out

module purge
module load python

python myscript.py
```

Other considerations

Grace period upon preemption: When jobs are preempted, a 120 second grace period is available to enable users to save and exit their jobs should they have the ability to do so. The preempted job is immediately sent SIGCONT and SIGTERM signals by Slurm in order to provide notification of its imminent termination. This is followed by the SIGCONT, SIGTERM and SIGKILL signal sequence upon reaching the end of the 120 second grace period. Users
wishing to do so can monitor the job for the SIGTERM signal and, when detected, take advantage of this 120 second grace period to save and exit their jobs.

The ‘blanca’ QoS: Note that as of 1 March, 2018, the “preemptable” qos replaces the previous low-priority QoS, “blanca”, which is no longer active.

2.6 Node types

Research Computing has several node types available on our resources. Each node type is meant for certain tasks. These node types are relatively common for other HPC centers. We will discuss each node type and its intended use below.

2.6.1 Login nodes

• Four virtual machines
• This is where you are when you log in
• No computation, compiling code, interactive jobs, or long running processes
• Script or code editing
• Running Jobs

2.6.2 Compile nodes

• Where you compile code, such as Fortran, C, C++
• No heavy computation
• Running Jobs
• Access these nodes by typing ssh scompile from a login node

2.6.3 Compute nodes

This is where jobs are executed after being passed to the scheduler.

• Intended for heavy computation
• When run an interactive job will be performing tasks directly on the compute nodes

2.7 Filesystems

All users are allocated space on the /home and /projects filesystems. In addition, separate scratch directories are visible from Summit and Blanca. These scratch directories are hosted on separate, high-performance filesystems designed to support intensive, parallel I/O operations.

Please note that the use of /home or /projects for high-performance I/O may negatively affect the environment for all users. As a result, all compute jobs should write to the appropriate scratch filesystem. Users performing intensive I/O on the /home or /projects filesystems will have their jobs terminated and may have their accounts temporarily disabled.
2.7.1 The Home Filesystem

Every user is allocated 2 GB of space on the /home filesystem in a subdirectory corresponding to their user name (e.g., /home/janedoe). Home directories are backed up frequently and are intended for the use of their owner only; sharing the contents of home directories with other users is strongly discouraged. Your /home directory is a good place to store source code, small compiled programs, and job scripts.

2.7.2 The Projects Filesystem

Each user has access to a 250 GB of space in their subdirectory of /projects (e.g., /projects/janedoe). As with the /home system, these directories are visible from all Research Computing nodes and are regularly backed up. The projects directory is intended to store software builds and smaller data sets.

2.7.3 Scratch Filesystems

Summit users are provided a subdirectory on /scratch/summit, the high-performance parallel scratch filesystem meant for I/O from jobs running on that system (e.g., /scratch/summit/janedoe). By default, each user is limited to a quota of 10 TB worth of storage space and 20M files and directories. Email rc-help@colorado.edu if you need these limits increased. Blanca users should write to /rc_scratch/janedoe instead of /scratch/summit.

Scratch space should be used for all compute jobs run on Summit or Blanca. These high-performance scratch directories are not backed up, and are not appropriate for long-term storage. Data may be purged at any time subject to overall system needs. Files are automatically removed 90 days after their initial creation.

Users requiring longer-term retention of their files should perform regular backups to their local machine if they have not purchased space on the PetaLibrary. Inappropriate use of scratch storage, including attempts to circumvent the automatic file purge policy, may result in loss of access to Research Computing resources.

2.7.4 Monitoring Disk Usage

Disk usage may be checked using the curc-quota command. When run from a Summit compile node, you will see output similar to:

```
[janedoe@shas0136 ~]$ curc-quota
------------------------------------------------------------------------
<table>
<thead>
<tr>
<th>Used</th>
<th>Avail</th>
<th>Quota Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>/home/janedoe</td>
<td>1.7G</td>
<td>339M</td>
</tr>
<tr>
<td>/projects/janedoe</td>
<td>67G</td>
<td>184G</td>
</tr>
<tr>
<td>/scratch/summit</td>
<td>29G</td>
<td>10211G</td>
</tr>
</tbody>
</table>
------------------------------------------------------------------------
```

If the command is run from a login node, information concerning /scratch/summit will be omitted.

Note that the space occupied by a particular directory and its subdirectories can be obtained via the du -h command:

```
[janedoe@shas0136 ~]$ du -h /scratch/summit/janedoe/WRF
698M WRF/run
698M WRF
```

2.7. Filesystems
2.7.5 Backups

Regular backups are performed for all /home and /projects directories and at a range of cadences. Low-cadence backups are retained for longer periods of time than high-cadence backups. A summary of the backup schedule is provided in the table below.

If disaster strikes and you need access to a previous version of your /home or /projects directories, change to that directory and look through the .snapshot hidden subdirectory. You will see a subdirectory associated with each snapshot of your /home or /projects directory, named using the time-stamp associated with the snapshot.

2.7.6 Workspace Sharing

All users have complete control over their personal directory permissions. While we encourage you to share your /projects and /scratch directories with collaborators as appropriate, we strongly discourage sharing of your /home directory due to the limited space and potentially sensitive information stored there.

Directories may be shared with all Research Computing users or with only a subset of our users. In the latter case, a system administrator will need to add your chosen collaborators to your Linux group. Please email rc-help@colorado.edu if you would like to add users to your Linux group.

In the example that follows, we make our /projects directory open to all users and then create subdirectories with select read/write permissions for all users and our chosen collaborators.

First, we make our /projects directory world-readable:

```
[janedoe@shas0136 ~]$ chmod a+rx /projects/janedoe
```

Next, we create a subdirectory that is visible to all users and which is read-only:

```
[janedoe@shas0136 ~]$ cd /projects/janedoe
[janedoe@shas0136 ~]$ mkdir world_read
[janedoe@shas0136 ~]$ chmod a+rx world_read
```

For our collaborators, we may want a writeable directory in addition to a read-only directory:

```
[janedoe@shas0136 ~]$ cd /projects/janedoe
[janedoe@shas0136 ~]$ mkdir group_read
[janedoe@shas0136 ~]$ chmod g+rx group_read
[janedoe@shas0136 ~]$ mkdir group_read_write
[janedoe@shas0136 ~]$ chmod g+rwx group_read_write
```

A similar methodology will need to be followed for all subdirectories you wish to share. If you make a mistake or change your mind, use the – symbol in lieu of + to remove privileges. Note that the x is necessary if you want other users to be able to cd into your directory.

2.8 The modules system

Research Computing uses a module system to load most software into a user’s environment. Most software is not accessible by default and must be loaded in. This allows Research Computing to provide multiple versions of the software concurrently and enables users to easily switch between different versions.
2.8.1 The `module` Command

**Modules should be loaded in job scripts, interactive jobs, or on compile nodes only. They should not be loaded when on one of the login nodes.** The `login node` will restrict the loading of modules, so you won’t be able to access software unless you do so through a job or a compile node.

To see what modules are available to load, ssh into a compile node by typing `ssh scompile` from a login node, and type:

```
module avail
```

This will return a list of modules available to load into the environment. **Please note if you run this command on a login node you will not receive a list of modules present on the system.**

To load your chosen modules into the environment type:

```
module load some_module
```

You can specify the version of the software by appending a `/` with the version number:

```
module load some_module/version
```

```
# example: "module load python/3.5.1"
```

The Lmod hierarchical module system provides five layers to support programs built with compiler and library consistency requirements. A module’s dependencies must be loaded before the module can be loaded.

The Layers include:

- Independent programs
- Compilers
- Compiler dependent programs
- MPI implementations
- MPI dependent programs

If you cannot load a module because of dependencies, you can use the `module spider` to find what dependencies you need to load the module.

```
module spider some_module
```

```
# example: "module spider openmpi"
```

2.8.2 Loading Modules in a Job Script

Loading a module will enable access to the modules described software package. Additionally, modules will set or modify a user’s environment variables.

Modules in a job script can be loaded after your `#SBATCH` directives and before your actual executable is called. A sample job script that loads Python into the environment is provided below:

```bash
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --time=00:01:00
```

(continues on next page)
#SBATCH --ntasks=1
#SBATCH --job-name=test-job
#SBATCH --output=test-job.%j.out

module purge
module load python/3.5.1

python3 test-program.py

2.8.3 Subcommands

The module command has a variety of subcommands, outlined in the table below. You may shorten the command to ml, but the shortened command may require specialized syntax.

2.9 Data transfer

Research Computing supports several methods of file transfers onto the CURC system. File transfers from a local machine can be done through a web-based application called Globus or through command line tools such as secure copy (scp), secure ftp (sftp) and rsync. Please note that use of some command line tools are limited to Mac and Linux Operating Systems.

Data transfers using SSH protocols can be done via the CURC login nodes, or through the CURC data transfer nodes (DTNs). Transfers via the login nodes work well for small and/or infrequent file transfers, and do not require users to be connected to the CU network. Transfers via the DTNs work well for all types of transfers, including large and/or frequent file transfers and automated (passwordless) transfers. Users must be connected to the CU network to use the DTNs. Details on all types of transfers are provided below.

2.9.1 Globus

On Globus, file transfers are handled through an interactive web application. Globus addresses deficiencies in secure copy requests by automating large data transfers, resuming failed transfers, and simplifying the implementation of high performance transfers between computing centers. Globus is available on Mac, Linux, and Windows operating systems and is RC’s recommended way of transferring data.

Sign into Globus Connect by selecting “University of Colorado at Boulder” from the dropdown menu and by logging in using your CU IdentiKey and password.

- If you’re with an institution outside of the University of Colorado at Boulder that is registered with Globus, sign in with your appropriate credentials.
- If your institution is not registered with Globus, you will need to make an account with Globus.
Files can easily be transferred from Summit to your local computer with Globus.

- Research Computing resources are installed with a Globus endpoint. You can connect to this endpoint by clicking the “endpoint” field and searching for the endpoint: CU Boulder Research Computing. Log into the end point by using your Research computing credentials.

- Your local computer must also have an endpoint. You can easily set up a Globus endpoint by installing Globus Connect Personal on your local machine.

Using the web app, connect your local workstation endpoint with the Research Computing endpoint and transfer files easily using the Globus GUI.
2.9.2 Globus Shared Endpoints

With Globus you can share files with users outside of your organization by creating shared endpoints. You can share any file/folder that you have access to. The user you are sharing with has to have a Globus account.

To learn how to setup a shared endpoint: Globus Shared Endpoint Documentation

2.9.3 Secure Copy scp

The Secure Copy protocol or scp allows users to send and receive data to the server remotely via a terminal command. The command appears as:

```
# Command to copy files from a local workstation to Research Computing resources
# Replace <path-to-file> with the path of the file you wish to copy
# Replace <username> with your Research Computing username
# Replace <target-path> with the full path to the directory you would like to send the file to.
scp <path-to-file> <username>@login.rc.colorado.edu:<target-path> # using a login node
scp <path-to-file> <username>@dtn.rc.int.colorado.edu:<target-path> # using a dtn node on campus network
scp <path-to-file> <username>@dtn.rc.colorado.edu:<target-path> # using a dtn node outside campus network
```

Windows users can access scp through PowerShell or a GUI application like WinSCP.

For more information on secure copy take a look at some of our listed resources or consult the scp manual page.

2.9.4 Using rsync on Summit

Another popular file transfer utility that can be used is the rsync command. While similar in function to scp, the major differences between rsync and scp are how the commands approach data transfer. Scp will bindly copy files from one server to another. Rsync aims to synchronize 2 files/directories to be the same. Because of this approach, rsync only copies files that are different from the source and target directories. This can be very useful in reducing the amount of copies you may perform when synchronizing two datasets. On a local machine, the command is called as follows:

```
# Command to synchronizing from a local machine to Research Computing resources
```

(continues on next page)
Windows users cannot access rsync by default and must install external software to access the command or through the Windows Subsystem for Linux (WSL).

For more information on rsync check out some of our listed resources or consult the rsync manual page.

### 2.9.5 Secure File Transfer Protocol: sftp

The Secure File Transfer Protocol is an interactive terminal solution to transfer data to and from Research Computing resources. SFTP works much like moving files in a terminal, we use a series of file system commands to navigate, move, remove, and copy information from a client’s personal machine to research computing resources. To invoke sftp type the command:

```
sftp <username>@login.rc.colorado.edu
```

```
sftp <username>@dtn.rc.int.colorado.edu
```

```
sftp <username>@dtn.rc.colorado.edu
```

We can then use various commands to traverse and manipulate both file systems. A list of commands are listed below:

Windows users can access sftp through PowerShell or a GUI application like WinSCP.

For more information on sftp check out some of our listed resources or consult the sftp manual page.

### 2.9.6 Passwordless scp and rsync

The scp and rsync commands both allow a user to transfer files without needing to reenter a password. All that is required is a few simple set up procedures to prepare your local machine.

**Note:** Passwordless data transfers are only available for Mac and Linux users. You must be logged into the campus VPN to perform passwordless data transfers to CURC.
Generate an ssh keypair on your local laptop/desktop

You only need to perform this step once. From a local terminal run:

```
ssh-keygen -t ed25519
```

This will create `~/.ssh/id_ed25519` and `~/.ssh/id_ed25519.pub` on your local machine (note: the “~” denotes your home directory).

Copy the public key to `~/.ssh/authorized_keys` on a CURC login node

You only need to perform this step once. From a local terminal run:

```
cat ~/.ssh/id_ed25519.pub | ssh <your-username>@login.rc.colorado.edu -T "cat >> ~/.ssh/authorized_keys"
```

... where you should substitute your CURC username for `<your-username>`; you will be required to enter your password and accept a Duo push in order to transfer the key.

Note: If you have trouble running the command above, you can also just login to a CURC login node, open `~/.ssh/authorized_keys` and paste the text from `~/.ssh/id_ed25519.pub` that resides on your local machine.

Use `rsync` or `scp` to transfer files without a password

Now you are ready to transfer files. Passwordless transfers must be through RC Data Transfer Nodes. Make sure you are within the CU network and transfer with scp/rsync to a DTN node. Examples:

User “ralphie” employs `rsync` to transfer `myfile.txt` to `/projects/ralphie` on CURC:

```
rsync -av ./myfile.txt dtn.rc.int.colorado.edu:/projects/ralphie/myfile.txt
```

User “ralphie” employs `scp` to transfer `myfile.txt` from Ralphie’s local machine to a PetaLibrary allocation called “crdds” that Ralphie has access to:

```
scp ./myfile.txt dtn.rc.int.colorado.edu:/pl/active/crdds/myfile.txt
```

2.9.7 Other Options

OIT also offers a file transfer service with a web interface which provides a good way to transfer files to collaborators. Files are uploaded to a server and a link to download the file can be emailed to an on, or off-campus, user: http://oit.colorado.edu/safe-transfer

2.9.8 DTN SSH Host Keys (as of 9 February 2021)

The first time you use the data transfer nodes you will be asked to verify the host key. You can refer to the keys published here to confirm that you are connecting to a valid data transfer node.

Note that each data transfer node may support more than one type of key, but only one is used (or displayed) by your client at any given time.

Fingerprint: 256 SHA256:c8362Adxws21Si5dyqngBpleCfvo7m/3cjT+gG6Nln4 no comment (ECDSA)
2.9.9 More reading

- Indiana University Tutorial on SFTP
- Linux Academy’s Tutorial on SSH and SCP
- ssh.com’s Tutorial on SCP and SFTP
- Linxize’s Tutorial on Rsync
- Ubuntu’s Documentation on Rsync

2.10 Compiling and linking

Before compiling in the RC environment, first ssh to one of the Summit compile nodes via `ssh scompile`. Next, load those modules corresponding to the compiler, MPI version (if needed), and third-party libraries required by your application. The load order should always be compiler first, MPI second, and third-party libraries last.

For example, suppose your application requires MPI and the HDF5 library. To compile using the Intel compiler and Intel MPI, the sequence of `module` commands would be:

```
module purge
module load intel
module load impi
module load hdf5
```

Supporting library-modules will be loaded as needed, and your environment will be updated so that the appropriate library directories are prepended to your `$PATH` and `$LD_LIBRARY_PATH`. The standard compiler variables `FC`, `CC` and `CXX` are set as appropriate for your compiler/MPI combination. These environment variables reference to the Fortran, C, and C++ compilers respectively.
In addition, several environment variables are set that may be useful during the compilation process. These variables possess the prefix CURC and may easily be found by searching your environment for CURC via `env | grep CURC`. This will yield output similar to:

```
[johndoe@shas0137 ~]$ env | grep CURC
CURC_INTEL_BIN=/curc/sw/intel/17.4/bin
CURC_INTEL_INC=/curc/sw/intel/17.4/include
CURC_INTEL_ROOT=/curc/sw/intel/17.4
CURC_INTEL_LIB=/curc/sw/intel/17.4/lib
CURC_HDF5_ROOT=/curc/sw/hdf5/1.10.1/impi/17.3/intel/17.4
CURC_HDF5_INC=/curc/sw/hdf5/1.10.1/impi/17.3/intel/17.4/include
CURC_HDF5_BIN=/curc/sw/hdf5/1.10.1/impi/17.3/intel/17.4/bin
CURC_HDF5_LIB=/curc/sw/hdf5/1.10.1/impi/17.3/intel/17.4/lib
[...]
```

Once the relevant modules are loaded, you are ready to compile. For our HDF5 example, a compilation command that uses the environment variables set by the module system may look like:

```
$FC my_program.f90 -I$CURC_HDF5_INC -L$CURC_HDF5_LIB -lhdf5_fortran -o my_program
```

**Note:** Your run-time environment should reflect your compilation environment. Be sure to include the same sequence of `module` commands in your job script as that used at compile time.

### 2.10.1 Navigating the Software Stack

The RC module system is hierarchical in nature, and available software libraries become visible to the user only after the compiler and MPI implementations that they depend on have been loaded. As noted above, modules should be loaded in the order: compiler, MPI, third-party software. At each stage of the load, executing `module avail` will reveal a list of newly available modules. The `module purge` command can be used to unload all currently loaded modules.

For example, before choosing a compiler, we can view the available compilers with

```
[janedoe@shas0136 ~]$ module purge
[janedoe@shas0136 ~]$ module avail
```

This will yield output similar to

```
----------------------------------- Compilers --------------------------------------
gcc/6.1.0  intel/16.0.3 (m)  intel/17.0.0 (m)  intel/17.4 (m,D)  pgi/16.5 ,
```

Several compiler-independent modules will also be displayed. Those modules (e.g., the Julia module) can be loaded at any time, irrespective of the compiler or MPI version in use.

If multiple versions of a package are available, a D is used to indicate the default version. When the version number is omitted during the `module load` command, the default version will be used. Considering the output above, the following two commands are equivalent:

```
[janedoe@shas0136 ~]$ module load intel
[janedoe@shas0136 ~]$ module load intel/17.4
```

Once the compiler is loaded, MPI-implementations and third-party serial libraries that depend on that compiler appear in the available module list until MPI Implementations and Compiler Dependent Applications:
Choosing an MPI implementation will similarly reveal MPI-dependent software under the header MPI Dependent Applications:

```
[janedoe@shas0136 ~]$ module load impi
[janedoe@shas0136 ~]$ module avail
```

You may search for a particular software package using the module spider command. This is typically a two-stage process. First search on the general software name without including any version information. If the software exists on our system, a list of available versions will appear:

```
[janedoe@shas0136 ~]$ module spider hdf5
```

Finally, to see which modules must be loaded to make your desired version available, run the module spider command again with the version information included:

```
[janedoe@shas0136 ~]$ module spider hdf5/1.10.1
```

2.10. Compiling and linking
2.10.2 Compiler and Optimization Recommendations

The Summit and Blanca clusters run on Intel-designed hardware. As such, we strongly recommend using the Intel compiler along with Intel’s MPI library when compiling software. For production, we suggest compiling with the -O2 or -O3 optimization flags along with the vectorization flags appropriate for the node you plan to run on. For Haswell nodes, this means compiling with the -xCORE-AVX2 flag. For the Xeon-Phi and Skylake nodes, use -xCORE-AVX512.

Compilation commands for a typical Summit Haswell node should resemble:

```
$FC -O3 -xCORE-AVX2 my_program.f90 -o my_program.out
$CC -O3 -xCORE-AVX2 my_program.c -o my_program.out
$CXX -O3 -xCORE-AVX2 my_program.cpp -o my_program.out
```

For the Phi and Skylake nodes, the appropriate commands would be:

```
$FC -O3 -xCORE-AVX512 my_program.f90 -o my_program.out
$CC -O3 -xCORE-AVX512 my_program.c -o my_program.out
$CXX -O3 -xCORE-AVX512 my_program.cpp -o my_program.out
```

2.10.3 Linking to the Math Kernel Library (MKL)

The Intel Math Kernel Library (MKL) provides optimized routines for a number of common mathematical operations. Notably, it provides interfaces to the LAPack and BLAS linear algebra libraries as well as the FFTW Fourier transform package.

If you wish to link MKL to your Intel-compiled application, use the -mkl flag:

```
$CXX -O3 -xCORE-AVX2 my_program.cpp -o my_program.out -mkl
```

If your application uses FFTW, you will also need to include MKL’s FFTW directory in your compilation command:

```
$CXX -O3 -xCORE-AVX2 -I$CURC_MKL_INC/fftw my_program.cpp -o my_program.out -mkl
```

For the GNU and PGI compilers, the link syntax becomes more complex. The Intel Link Advisor can be used to generate the appropriate linking syntax based on your application’s needs.

For the GNU compiler, linking against sequential MKL libraries, the appropriate Fortran linking syntax is:

```
$FC my_program.f90 -m64 -I$CURC_MKL_INC -o my_program.out -L$CURC_MKL_LIB -Wl,--no-as-needed -lmkl_gf_lp64 -lmkl_sequential -lmkl_core -lpthread -lm -ldl
```

The comparable c/c++ syntax would be:

```
$FC my_program.cpp -m64 -I$CURC_MKL_INC -o my_program.out -L$CURC_MKL_LIB -Wl,--no-as-needed -lmkl_intel_lp64 -lmkl_sequential -lmkl_core -lpthread -lm -ldl
```

Note that if your application uses FFTW, you will must use the FFTW include flag just as with the Intel compiler. See the link advisor or contact rc-help@colorado.edu if you have additional questions about how to link MKL to your application.
2.11 The PetaLibrary

2.11.1 Description of service

The University of Colorado Boulder Research Computing PetaLibrary research data storage service is available to any researcher affiliated with the University of Colorado Boulder. Storage space is pre-purchased, with a minimum purchase of 1 TB for 1 year.

Access

External access to PetaLibrary storage is provided primarily via the Research Computing data transfer service which supports Globus, SFTP, and (by request only) anonymous FTP data transfer protocols.

Service classes

Two primary classes of storage are available:

Active
- Appropriate for data that is frequently written or read
- Stores data in a parity-protected RAID array or similar
- Directly accessible (read+write) from Research Computing compute resources

Archive
- Appropriate for data that is infrequently accessed
- Stores data on tapes in a robotic tape library, with all data written to at least two tapes

2.11.2 Allocations

PetaLibrary storage is presented as a file system directory with an administratively-defined and automatically-enforced size limit. When an allocation is created the full, supported path is defined and provided, based on an agreed allocation name.

Allocation contacts

Each PetaLibrary allocation must define an allocation owner. This contact is often, but is not required to be, the Principal Investigator for a related research effort. Among other responsibilities, the allocation owner is required to ensure compliance with PetaLibrary Terms of Service, for themself and their authorized allocation users.

An optional technical contact and billing contact may also be defined. These contacts are treated as delegates of the allocation owner for normal or regular operation.

Additional allocation users may also be granted access to the allocation.

Access control

Access to a PetaLibrary allocation is granted using an access group. This group may be an existing group in the Research Computing environment or a new group created specifically for the purpose of managing access to the allocation. Allocation users are made members of this access group using their existing Research Computing accounts by request of the allocation owner or delegate contact to rc-help@colorado.edu. Changes to the access group membership
must be authorized by the allocation owner or delegate contact. The allocation owner or delegate contact should audit access group membership at least annually.

The PetaLibrary relies on the overall Research Computing identity management directory and the CU Boulder identity management and “IdentiKey” authentication infrastructure. All PetaLibrary users must have a valid Research Computing user account.

All PetaLibrary allocations are also eligible to use the Globus Sharing service, with which Globus Online can be used to provide read and/or write access to your allocation to anyone with a Globus account, including individuals who do not have a Research Computing user account. Research Computing will sponsor one Globus Plus account (for share administration) per PetaLibrary allocation, typically via the allocation’s technical contact.

 Fees

PetaLibrary storage is provided for an annual fee based on accessible capacity. Actual fees are evaluated internally and subject to change annually. Current rates are published at https://www.colorado.edu/rc/resources/petalibrary/storageandrates.

We do not provide refunds or pro-rate the PetaLibrary fee for any time lost due to repairs, maintenance events (planned or otherwise), or any other temporary down time.

 Reports

A report indicating each allocation’s capacity, load, and access is automatically generated and distributed to each allocation owner and delegate contact.

 Expiration and deletion

An allocation that has not been funded for its upcoming period is considered expired. When an allocation expires a notification is sent to the allocation owner and all allocation contacts notifying them of the allocation expiry. If the allocation remains unfunded for 30 days it may be made “read-only.” If the allocation remains unfunded for 60 days the allocation may be deleted, including all data.

2.11.3 Reliability

The PetaLibrary is designed for low-cost, reliable, remote-accessible research data storage. Data stored in the PetaLibrary is protected from hardware failure by redundancies (either redundant copies or parity) but is not “backed up.” The PetaLibrary may serve as a backup for data stored elsewhere. Optionally, replication of data between the active and archive storage classes may be sufficient for some needs.

The infrastructure supporting the PetaLibrary is funded through the Office of Information Technology. Operation of the PetaLibrary is expected to continue with regular funding through “renewal and replacement” contributions. In the event that Research Computing is unable or otherwise ceases to provide the PetaLibrary or any comparable service, at least 60 days advance notice will be given to all allocation owners and delegate contacts. It will be the responsibility of the client to transfer their data to other storage resources within that time window.

2.11.4 Support

Authorized PetaLibrary users may request support and assistance via rc-help@colorado.edu. While Research Computing will make every effort to address each support request, those support requests which require domain-specific knowledge or expertise may not be able to be handled by Research Computing alone. In these cases, the support request may be forwarded to the allocation’s technical contact.
The PetaLibrary infrastructure is operated as a “best effort” service, with regular business hours 08:00 - 17:00, Monday - Friday. There is no guarantee of assistance or incident response after-hours, on weekends, or during campus holidays. Support requests, including, but not limited to, access control and account creation, are addressed during regular business hours.

While every reasonable effort will be made to ensure the reliability and availability of the PetaLibrary and of the data stored on it, access to data in the PetaLibrary may be affected by circumstances outside of the control of Research Computing.

### 2.11.5 Maintenance

Planned maintenance of the PetaLibrary is usually scheduled on the first Wednesday of a given month. Upcoming maintenance, both planned and emergency, is announced at https://curc.statuspage.io.

### 2.11.6 Performance

Research Computing will publish expected performance capabilities of each PetaLibrary class at https://www.colorado.edu/rc/resources/petalibrary; however, the PetaLibrary is a shared infrastructure, and actual, individual performance will vary depending on each individual workload and competing workloads from other clients.

The PetaLibrary service is designed for file storage and retrieval, and is not an ideal backend for highly transactional workloads (e.g., relational databases).

Use the PetaLibrary to store, archive and share research data.

- **Minimum project size:** 2 TB
- **2 classes of storage:** active and archive
  - 5 different storage options within these classes
  - See our website for pricing information
- **New customers are limited to either**
  - 75 TB in Active Storage, or
  - 100 TB in Archive Storage
  - Email rc-help@colorado.edu to request more space if necessary

### 2.11.7 Getting Started

Request access to the PetaLibrary by:

- **Reading the PetaLibrary Memorandum of Understanding**
- **Filling out the attached End User Agreement and Order Form** (found in the MOU), and emailing those documents to rc-help@colorado.edu

Each person who accesses the PetaLibrary is required to have a Research Computing account and two-factor authentication.

Your PetaLibrary allocation will have a name that is selected by you. For example, Jane Doe might name her lab’s allocation jdoe_lab.

- **active storage allocations will have the path** `/pl/active/<your_allocation_name>` (e.g., `/pl/active/jdoe_lab`)
• **archive** storage allocations will have the path `/pl/archive/<your_allocation_name>` (e.g., `/pl/archive/jdoe_lab`)

Note that access via the login nodes is not recommended for frequent or large read/writes of archived data.

### 2.11.8 Video tutorial

### 2.11.9 More information

- [https://www.colorado.edu/rc/resources/petalibrary](https://www.colorado.edu/rc/resources/petalibrary)

### 2.12 Backing up local data to PetaLibrary

This guide details the process of making a backup for local data on your laptop or lab server to a PetaLibrary allocation. The method employs the software **rclone**, which is a command line application that is available for many architectures. While the following tutorial is tailored for a MacOS user, the general steps to follow are the same and are relevant for Windows and Linux users too.

#### 2.12.1 Generate public/private keypair on login.rc.colorado.edu

In this step you will login to an RC login node (`ssh login.rc.colorado.edu`) and generate an ssh keypair. You don’t need to complete this step on a login node if you have ssh-keygen installed on your system. Be sure to generate a key in PEM format. Here is the command to generate a key in PEM format with a uniquename, after you have logged in:

```bash
$ ssh-keygen -f ~/.ssh/rclone_ssh_key -m PEM
```

You will be prompted to enter a passphrase to protect the key. For automated backups, you will either want a key with no passphrase, or to set a passphrase and use something like ssh-agent to hold the key in memory. The easiest option is to not set a passphrase. Here is sample output from running ssh-keygen:

```bash
$ ssh-keygen -f ~/.ssh/rclone_ssh_key -m PEM
Generating public/private rsa key pair.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/jesse/.ssh/rclone_ssh_key.
Your public key has been saved in /home/jesse/.ssh/rclone_ssh_key.pub.
The key fingerprint is:
SHA256:ftryJYQ2LRnNQ/fSzm+iB0Tg8/LkePqVyEDky2z3JvA jesse@login10
The key’s randomart image is:
|   o.  |   o.  |
|  * o  | * o   |
| +.+.  | S + . |
| o + Xoo + |
| . + B** o |
| . + B.+Eo.|
| . o+o+o  +. |
```

38 Chapter 2. Acknowledging RC
Your home directory should now have the files `~/.ssh/rclone_ssh_key` and `~/.ssh/rclone_ssh_key.pub`. The `rclone_ssh_key` is your private key, be careful with it as anyone with this key can access resources that trust the public key.

### 2.12.2 Copy public key to your authorized_keys file

Copy the newly generated public key `rclone_ssh_key.pub` to your `authorized_keys` file:

```
$ cat ~/.ssh/rclone_ssh_key.pub >> ~/.ssh/authorized_keys
```

This will append the contents of `rclone_ssh_key.pub` to the file `authorized_keys` (`authorized_keys` will be created if it does not exist). Now anyone with your private key (`rclone_ssh_key`) can login to hosts that mount home directories from RC core storage.

### 2.12.3 Copy private key to the backup source host

This next step can’t be documented exactly as every client system will be different, but use a secure method (such as scp or sftp) to copy the private key in `~/.ssh/rclone_ssh_key` to the system that you want to back up to PetaLibrary (e.g., your laptop or lab server).

**Windows:** Windows users should copy client (e.g. Globus, WinSCP, WSL2, etc) as Admin. Change directories to `rclone` file location and run:

### 2.12.4 Install rclone on the backup source host

The rclone application is available download here for a variety of architectures. Download `rclone` here and follow Rclone’s provided instructions to setup the application.

*Note: You must have administrative privileges on your laptop or lab server in order to install software. If you do not, you’ll need to ask your system administrator.*

**Windows:** Download rclone and manually unzip the compressed files to your desired install location. Take note of this install location since you will need to manually access these binaries to setup and run rclone.

### 2.12.5 Configure rclone

The rclone application will require you to configure endpoints. Once your endpoints are configured, you can copy data to/from your local system to configured endpoints. Please note that rclone should only ever be connected to an RC Data Transfer Node (DTN). Because of this, we will configure an sftp endpoint in rclone that points to RC’s DTN hosts. You must be connected to CU’s Network for this connection to work. For more information on DTN nodes, check out our documentation on data transfers.

In this example we use rclone to create an sftp endpoint with the following settings:

```ini
name: cu_rc_dtn
type: sftp
host: dtn-data.rc.int.colorado.edu
key_file = /Users/jesse/.ssh/rclone_ssh_key
user = jesse
```

The rclone application is interactive and will prompt you for all of the above information. Here is the ouput of an example interactive session when creating an endpoint with the above settings:

```
$ rclone config
```
Windows: From the Command Prompt, navigate to rclone file location and run: 

```
$ .\rclone.exe config
```

2020/05/18 15:00:00 NOTICE: Config file "/Users/jesse/.config/rclone/rclone.conf" not found - using defaults

No remotes found - make a new one
n) New remote
s) Set configuration password
q) Quit config
n/s/q> n
name> cu_rc_dtn
Type of storage to configure.
Enter a string value. Press Enter for the default ("").
Choose a number from below, or type in your own value

1 / 1Fichier
   "fichier"
2 / Alias for an existing remote
   "alias"
3 / Amazon Drive
   "amazon cloud drive"
4 / Amazon S3 Compliant Storage Provider (AWS, Alibaba, Ceph, Digital Ocean, Dreamhost, IBM COS, Minio, etc)
   "s3"
5 / Backblaze B2
   "b2"
6 / Box
   "box"
7 / Cache a remote
   "cache"
8 / Citrix Sharefile
   "sharefile"
9 / Dropbox
   "Dropbox"
10 / Encrypt/Decrypt a remote
    "crypt"
11 / FTP Connection
    "ftp"
12 / Google Cloud Storage (this is not Google Drive)
    "google cloud storage"
13 / Google Drive
    "drive"
14 / Google Photos
    "google photos"
15 / Hubic
    "hubic"
16 / In memory object storage system.
    "memory"
17 / JottaCloud
    "jottacloud"
18 / Koofr
    "koofr"
19 / Local Disk
    "local"
20 / Mail.ru Cloud
    "mailru"
21 / Mega
    "mega"
22 / Microsoft Azure Blob Storage
    "azureblob"

(continues on next page)
23 / Microsoft OneDrive
  \ "onedrive"
24 / OpenDrive
  \ "opendrive"
25 / Openstack Swift (Rackspace Cloud Files, Memset Memstore, OVH)
  \ "swift"
26 / Pcloud
  \ "pcloud"
27 / Put.io
  \ "putio"
28 / QingCloud Object Storage
  \ "qingstor"
29 / SSH/SFTP Connection
  \ "sftp"
30 / Sugarsync
  \ "sugarsync"
31 / Transparently chunk/split large files
  \ "chunker"
32 / Union merges the contents of several remotes
  \ "union"
33 / Webdav
  \ "webdav"
34 / Yandex Disk
  \ "yandex"
35 / http Connection
  \ "http"
36 / premiumize.me
  \ "premiumizeme"

Storage> 29
** See help for sftp backend at: https://rclone.org/sftp/ **

SSH host to connect to
Enter a string value. Press Enter for the default ("").
Choose a number from below, or type in your own value
1 / Connect to example.com
  \ "example.com"
host> dtn-data.rc.int.colorado.edu
SSH username, leave blank for current username, jesse
Enter a string value. Press Enter for the default ("").
user>
SSH port, leave blank to use default (22)
Enter a string value. Press Enter for the default ("").
port>
SSH password, leave blank to use ssh-agent.
y) Yes type in my own password
g) Generate random password
n) No leave this optional password blank (default)
y/g/n>
Path to PEM-encoded private key file, leave blank or set key-use-agent to use ssh-agent.
Enter a string value. Press Enter for the default ("").
key_file> /Users/jesse/.ssh/rclone_ssh_key
The passphrase to decrypt the PEM-encoded private key file.

Only PEM encrypted key files (old OpenSSH format) are supported. Encrypted keys in the new OpenSSH format can't be used.
y) Yes type in my own password
g) Generate random password
n) No leave this optional password blank (default)
y/g/n>
When set forces the usage of the ssh-agent.

When key-file is also set, the ".pub" file of the specified key-file is read and only → the associated key is requested from the ssh-agent. This allows to avoid "Too many authentication failures for *username*" errors when the ssh-agent contains many keys.
Enter a boolean value (true or false). Press Enter for the default ("false").
key_use_agent>
Enable the use of insecure ciphers and key exchange methods.

This enables the use of the the following insecure ciphers and key exchange methods:

- aes128-cbc
- aes192-cbc
- aes256-cbc
- 3des-cbc
- diffie-hellman-group-exchange-sha256
- diffie-hellman-group-exchange-shal

Those algorithms are insecure and may allow plaintext data to be recovered by an → attacker.
Enter a boolean value (true or false). Press Enter for the default ("false").
Choose a number from below, or type in your own value
1 / Use default Cipher list.
  \ "false"
2 / Enables the use of the aes128-cbc cipher and diffie-hellman-group-exchange—sha256, diffie-hellman-group-exchange-shal key exchange.
  \ "true"
use_insecure_cipher>
Disable the execution of SSH commands to determine if remote file hashing is → available.
Leave blank or set to false to enable hashing (recommended), set to true to disable → hashing.
Enter a boolean value (true or false). Press Enter for the default ("false").
disable_hashcheck>
Edit advanced config? (y/n)
y) Yes
n) No (default)
y/n>
Remote config
--------------------
[cu_rc_dtn]
type = sftp
host = dtn-data.rc.int.colorado.edu
key_file = /Users/jesse/.ssh/rclone_ssh_key
--------------------
y) Yes this is OK (default)
e) Edit this remote
d) Delete this remote
y/e/d>
Current remotes:

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
</table>

(continues on next page)
2.12.6 Verify rclone config

You can verify your settings by running ‘rclone config show’. The results from the example above looked like this after running through the initial configuration:

```
$ rclone config show
Windows: Windows host users $ .\rclone.exe config
```

```
[cu_rc_dtn]
type = sftp
host = dtn-data.rc.int.colorado.edu
key_file = /Users/jesse/.ssh/rclone_ssh_key
user = jesse
md5sum_command = none
sha1sum_command = none
```

2.12.7 Test rclone

**Example:** The syntax for using rclone to synchronize a local directory `/tmp/local_backup_dir` to a directory `pl_backup_dir` in a PetaLibrary allocation named `pl_allocation`, the command (executed from your laptop or lab server) would be:

```
$ rclone sync /tmp/local_backup_dir cu_rc_dtn:/pl/active/pl_allocation/pl_backup_dir
```

This should synchronize the data in the two directories (note that `rclone` does not report on the transfer so it will look like your terminal is “frozen” until the transfer is complete). The source directory will not be modified, and `pl_backup_dir` will have files added/removed to match the contents of `local_backup_dir`.

2.12.8 Run rclone on a schedule

On Mac or Linux hosts, you can set up a cron job to run the rclone sync job regularly. To create a crontab entry type:

```
$ crontab -e
```

And then add the line for the automated job you want to run, e.g.:

```
0 4 * * 1 rclone sync /tmp/local_backup_dir cu_rc_dtn:/pl/active/pl_allocation/pl_backup_dir
```
This would run the rclone sync command every Monday at 4am. The syntax for a crontab entry is:

```
* * * * * command
* - minute (0-59)
* - hour (0-23)
* - day of the month (1-31)
* - month (1-12)
* - day of the week (0-6, 0 is Sunday)
command - command to execute
(from left-to-right)
```

Windows systems can set up scheduled tasks to run rclone automatically.

**Windows:** Windows host users, if you do not have WSL2 installed use taskschd.msc (Task Scheduler) as Admin and run Action=>Create Basic Task

### 2.13 Running applications with Jobs

Because Summit is shared among many researchers, Research Computing manages usage of the system through jobs. **Jobs** are simply an allotment of resources that can be used to execute processes. Research Computing uses a program named the *Simple Linux Utility for Resource Management*, or **Slurm**, to create and manage jobs.

In order to run a program on Summit, you must request resources from Slurm to generate a job. Resources can be requested from a login node or a compile node. You must then provide commands to run your program on those requested resources. Where you provide you commands depends on whether you are running a **batch job** or an interactive job.

When you run a batch job or an interactive job, it will be placed in a queue until resources are available. A detailed guide on the Slurm queue and accounting tools can be found here.

#### 2.13.1 Batch Jobs

The primary method of running applications on Research Computing resources is through a batch job. A **batch job** is a job that runs on a compute node with little or no interaction with the users. You should use batch jobs for:

- Any computationally expensive application that could take hours or days to run
- Any application that requires little or no user input
- Applications that you do not need to monitor extensively

Unlike running an application on your personal machine, you do not call the application you wish to run directly. Instead you create a **job script** that includes a call to your application. Job scripts are simply a set of resource requests and commands. When a job script is run, all the commands in the job script are executed on a compute node.

Once created, you can run your job script by passing it to the Slurm queue with the `sbatch` command followed by your job script name:

```
sbatch <your-jobsxript-name>
```

If no job script is provided then `sbatch` will take whatever commands follow as standard input.

A detailed guide on constructing and running Job scripts can be found here.
2.13.2 Interactive Jobs

Another method of running applications on Research Computing resources is through an interactive job. As the name would imply, an interactive job is a job that allows users to interact with requested resources in real time. Users can run applications, execute scripts, or run other commands directly on a compute node. Interactive jobs should be used for:

- Debugging applications or workflows
- Any application that requires user input at runtime
- Any application with a GUI (Graphical User Interface)

You can request an interactive job by using the `sinteractive` command. Unlike the `sbatch`, resources must be requested via the command line through the use of flags. Though running `sinteractive` without any flags is possible, this will result in default values being used for your jobs. Research Computing highly recommends you provide a `qos` and a `time` parameter to avoid long queue times or accidental overuse of your priority.

```
sinteractive --qos=interactive --time=00:10:00
```

A list of `sinteractive` parameters can be found here. The example above will run an interactive job that will run a terminal session on one core of one node with the interactive quality of service (QoS) for ten minutes. Once the interactive session has started you can run any interactive terminal application you may need on the command line.

More details on running Interactive Jobs can be found here.

2.14 Batch Jobs and Job Scripting

Batch jobs are by far the most common type of job on Summit. Batch jobs are resource provisions that run applications on nodes away from the user and do not require supervision or interaction. Batch jobs are commonly used for applications that run for long periods of time or require little to no user input.

Batch jobs are created from a job script which provide resource requirements and commands for the job.

2.14.1 Job Scripts

Even though it is possible to run jobs completely from the command line, it is often overly tedious and unorganized to do so. Instead, Research Computing recommends constructing a job script for your batch jobs. A job script is set of Linux commands paired with a set of resource requirements that can be passed to the Slurm job scheduler. Slurm will then generate a job according to the parameters set in the job script. Any commands that are included with the job script will be run within the job.

2.14.2 Running a Job Script

Running a job script can be done with the `sbatch` command:

```
sbatch <your-job-script-name>
```

Because job scripts specify the desired resources for your job, you won’t need to specify any resources on the command line. You can, however, overwrite or add any job parameter by providing the specific resource as a flag within `sbatch` command:
#SBATCH --partition=sgpu <your-job-script>

Running this command would force your job to run on the sgpu partition no matter what your job script specified.

## 2.14.3 Making a Job Script

Although Research Computing provides a variety of different sample scripts users can utilize when running their own jobs, knowing how to draft a job script can be quite handy if you need to debug any errors in your jobs or you need to make substantial changes to a script.

A job script looks something like this:

```bash
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --ntasks=1
#SBATCH --time=00:10:00
#SBATCH --partition=shas-testing
#SBATCH --output=sample-%j.out
module purge
module load intel
module load mkl
echo "== This is the scripting step! =="
sleep 30
./executable.exe
echo "== End of Job =="
```

Normally job scripts are divided into 3 primary parts: directives, loading software, and user scripting. Directives give the terminal and the Slurm daemon instructions on setting up the job. Loading software involves cleaning out the environment and loading specific pieces of software you need for your job. User scripting is simply the commands you wish to be executed in your job.

### 1. Directives

A directive is a comment that is included at the top of a job script that tells the shell information about the script. The first directive, the shebang directive, is always on the first line of any script. The directive indicates which shell you want running commands in your job. Most users employ bash as their shell, so we will specify bash by typing:

```bash
#!/bin/bash
```

The next directives that must be included with your job script are `SBATCH` directives. These directives specify resource requirements to Slurm for a batch job. These directives must come after the shebang directive and before any commands are issued in the job script. Each directive contains a flag that requests a resource the job would need to complete execution. An sbatch directive is written as such:

```
#SBATCH --<resource>=<amount>
```

For example if you wanted to request 2 nodes with an sbatch directive, you would write:

```
#SBATCH --nodes=2
```
A list of some useful sbatch directives can be found here. A full list of commands can be found in Slurm’s documentation for sbatch.

## Software

Because jobs run on a different node than from where you ran, any shared software that is needed must be loaded via the job script. Software can be loaded in a job script just like it would be on the command line. First we will purge all software that may be left behind from your working environment on a compile node by running the command:

```
module purge
```

After this you can load whatever software you need by running the following command:

```
module load <software>
```

More information about software modules can be found here.

## User Scripting

The last part of a job script is the actual user scripting that will execute when the job is executing. This part of the job script includes all user commands that are needed to set up and execute the desired task. Any Linux command can be utilized in this step. Scripting can range from highly complex loops iterating over thousands of files to a simple call to an executable. Below is a simple example of some user scripting:

```
echo "== This is the scripting step! =="
touch tempFile1.in
touch tempFile2.in
sleep 30
./executable.exe tempfile1.in tempfile2.in
echo "== End of Job =="
```

### 2.14.4 Examples

Job script to run a 5 minute long, 1 node, 1 core C++ Job:

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --time=00:05:00
#SBATCH --partition=shas-testing
#SBATCH --ntasks=1
#SBATCH --job-name=cpp-job
#SBATCH --output=cpp-job.%j.out
module purge
module load gcc
./example_cpp.exe
```

Job script to run a 7 minute long, 1 node, 4 core C++ OpenMP Job:
Job script to run a 10 minute long, 2 node, 24 core C++ MPI Job:

```bash
#!/bin/bash
module purge
module load intel
module load impi
mpirun -np 24 ./example_mpi.exe
```

### 2.14.5 Job Flags

The `sbatch` command supports many optional flags. To review all the options, please visit the Slurm `sbatch` page. Below are a few flags you may want to consider when running your job via `sbatch`.

### 2.15 Interactive jobs

Interactive jobs allow a user to interact with applications in real time within an HPC environment. With interactive jobs, users request time and resources to work on a compute node directly. Users can then run graphical user interface (GUI) applications, execute scripts, or run other commands directly on a compute node. Common reasons for running interactive jobs include debugging, designing workflows, or preference in using the GUI interface of an application.

#### 2.15.1 General Interactive Jobs

To run an interactive job on Research Computing resources, request an interactive session by utilizing the `sinteractive` command. The `sinteractive` command creates a job with parameters provided through flags run with the command. After moving through the Slurm queue the interactive job will put the user onto the command line of a compute node to interactively use their resource allotment.
Any resource that could be specified in a job script or with `sbatch` can also be used with `sinteractive`. Check out this page for a list of Slurm directives that can be used with interactive jobs.

The primary flags we recommend users specify are the `partition` flag and the `time` flag. These flags will specify partition and amount of time for your job respectively. The `sinteractive` command is run as follows:

```
sinteractive --partition=shas-interactive --time=00:10:00
```

This will run an interactive job to the Slurm queue that will start a terminal session that will run on one core of one node on the interactive shas partition for ten minutes. Once the session has started you can run any application or script you may need from the command line. For example, if you type `python` you will open an interactive python shell on a compute node (rather than the login nodes, which is forbidden).

### 2.15.2 Interactive GUI Applications

To run an interactive GUI application on Summit, we must install an X windows server application and enable X11 forwarding on our personal computer.

**Windows setup**

On Windows we must first install an X windows server application to allow Summit to forward the GUI information to your local system. For Windows, we will use an application called Xming to accomplish this. Download the Xming here.

Next we must enable X11 forwarding on the PuTTY application. Download and install the PuTTY application here if you have not done so already.
Expand the SSH tab on the left side of the application and click X11.
In the X11 Menu check the “Enable X11 Forwarding” checkbox and type “localhost:0” in the X display location field. Clicking open will open a terminal window where you can login.

**macOS setup**

Using macOS, we will also need to install an X windows server application to allow Summit to forward GUI information to your local system. For Mac, we will use an application called XQuartz to accomplish this. Download and install XQuartz here.

Opening the application will bring up a terminal window. In this window, you will ssh to login.rc.colorado.edu as you normally would except you’ll include the “-X” flag:

```
ssh -X your_rc-username@login.rc.colorado.edu
```

**Running GUI Applications**

Once you have logged into Summit with X11 Forwarding enabled, you will be able to initialize a GUI application by starting an interactive job and running your selected application. The X-window server application installed on your local system will display the window generated on the cluster on your local machine.

If you plan on running interactive job from a compile node, you must also enable x11 forwarding when you ssh into scompile:

```
ssh -X scompile
```

From here you will be able to run your interactive job like normal and X11 forwarding will carry through to the job.
2.16 Useful Slurm commands

Slurm provides a variety of tools that allow a user to manage and understand their jobs. This tutorial will introduce these tools, as well as provide details on how to use them.

2.16.1 Finding queuing information with squeue

The `squeue` command is a tool we use to pull up information about the jobs in queue. By default, the `squeue` command will print out the job ID, partition, username, job status, number of nodes, and name of nodes for all jobs queued or running within Slurm. Usually you wouldn’t need information for all jobs that were queued in the system, so we can specify jobs that only you are running with the `--user` flag:

```
$ squeue --user=your_rc-username
```

We can output non-abbreviated information with the `--long` flag. This flag will print out the non-abbreviated default information with the addition of a timelimit field:

```
$ squeue --user=your_rc-username --long
```

The `squeue` command also provides users with a means to calculate a job’s estimated start time by adding the `--start` flag to our command. This will append Slurm’s estimated start time for each job in our output information. (Note: The start time provided by this command can be inaccurate. This is because the time calculated is based on jobs queued or running in the system. If a job with a higher priority is queued after the command is run, your job may be delayed.)

```
$ squeue --user=your_rc-username --start
```

When checking the status of a job, you may want to repeatedly call the `squeue` command to check for updates. We can accomplish this by adding the `--iterate` flag to our `squeue` command. This will run `squeue` every n seconds, allowing for a frequent, continuous update of queue information without needing to repeatedly call `squeue`:

```
$ squeue --user=your_rc-username --start --iterate=n_seconds
```

Press `ctrl-c` to stop the command from looping and bring you back to the terminal.

For more information on `squeue`, visit the Slurm page on `squeue`

2.16.2 Stopping jobs with scancel

Sometimes you may need to stop a job entirely while it’s running. The best way to accomplish this is with the `scancel` command. The `scancel` command allows you to cancel jobs you are running on Research Computing resources using the job's ID. The command looks like this:

```
$ scancel your_job-id
```

To cancel multiple jobs, you can use a comma-separated list of job IDs:

```
$ scancel your_job-id1, your_job-id2, your_job-id3
```

For more information, visit the Slurm manual on `scancel`
2.16.3 Learning status information with `sstat`

The `sstat` command allows users to easily pull up status information about their currently running jobs. This includes information about **CPU usage, task information, node information, resident set size (RSS),** and **virtual memory (VM)**. We can invoke the `sstat` command as such:

```
$ sstat --jobs=your_job-id
```

By default, `sstat` will pull up significantly more information than what would be needed in the commands default output. To remedy this, we can use the `--format` flag to choose what we want in our output. The format flag takes a list of comma separated variables which specify output data:

```
$ sstat --jobs=your_job-id --format=var_1,var_2, ... , var_N
```

A chart of some these variables are listed in the table below:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>jobid</td>
<td>Job ID</td>
</tr>
<tr>
<td>cputime</td>
<td>CPU Time</td>
</tr>
<tr>
<td>maxrss</td>
<td>Max RSS</td>
</tr>
<tr>
<td>ntasks</td>
<td>Number of Tasks</td>
</tr>
</tbody>
</table>

For an example, let’s print out a job’s average job id, cpu time, max rss, and number of tasks. We can do this by typing out the command:

```
sstat --jobs=your_job-id --format=jobid,cputime,maxrss,ntasks
```

A full list of variables that specify data handled by `sstat` can be found with the `--helpformat` flag or by visiting the slurm page on `sstat`.

2.16.4 Analyzing past jobs with `sacct`

The `sacct` command allows users to pull up status information about past jobs. This command is very similar to `sstat`, but is used on jobs that have been previously run on the system instead of currently running jobs. We can use a job’s id...

```
$ sacct --jobs=your_job-id
```

...or your Research Computing username...

```
$ sacct --user=your_rc-username
```

...to pull up accounting information on jobs run at an earlier time.

By default, `sacct` will only pull up jobs that were run on the current day. We can use the `--starttime` flag to tell the command to look beyond its short-term cache of jobs.

```
$ sacct --jobs=your_job-id --starttime=YYYY-MM-DD
```

To see a non-abbreviated version of `sacct` output, use the `--long` flag:

```
$ sacct --jobs=your_job-id --starttime=YYYY-MM-DD --long
```

**Formatting `sacct` output**

Like `sstat`, the standard output of `sacct` may not provide the information we want. To remedy this, we can use the `--format` flag to choose what we want in our output. Similarly, the format flag is handled by a list of comma separated variables which specify output data:

```
$ sacct --user=your_rc-username --format=var_1,var_2, ... ,var_N
```
A chart of some variables is provided below:

As an example, suppose you want to find information about jobs that were run on March 12, 2018. You want to show information regarding the job name, the number of nodes used in the job, the number of cpus, the maxrss, and the elapsed time. Your command would look like this:

```
$ sacct --jobs=your_job-id --starttime=2018-03-12 --format=jobname,nnodes,ncpus,maxrss,elapsed
```

As another example, suppose you would like to pull up information on jobs that were run on February 21, 2018. You would like information on job ID, job name, QoS, Number of Nodes used, Number of CPUs used, Maximum RSS, CPU time, Average CPU time, and elapsed time. Your command would look like this:

```
$ sacct --jobs=your_job-id --starttime=2018-02-21 --format=jobid,jobname,qos,nnodes,ncpu,maxrss,cputime,avecpu,elapsed
```

A full list of variables that specify data handled by sacct can be found with the `--helpformat` flag or by visiting the slurm page on sacct.

### 2.16.5 Controlling queued and running jobs using `scontrol`

The `scontrol` command provides users extended control of their jobs run through Slurm. This includes actions like suspending a job, holding a job from running, or pulling extensive status information on jobs.

To suspend a job that is currently running on the system, we can use scontrol with the `suspend` command. This will stop a running job on its current step that can be resumed at a later time. We can suspend a job by typing the command:

```
$ scontrol suspend job_id
```

To resume a paused job, we use scontrol with the `resume` command:

```
$ scontrol resume job_id
```

Slurm also provides a utility to hold jobs that are queued in the system. Holding a job will place the job in the lowest priority, effectively “holding” the job from being run. A job can only be held if it’s waiting on the system to be run. We use the `hold` command to place a job into a held state:

```
$ scontrol hold job_id
```

We can then release a held job using the `release` command:

```
$ scontrol release job_id
```

`scontrol` can also provide information on jobs using the `show job` command. The information provided from this command is quite extensive and detailed, so be sure to either clear your terminal window, grep certain information from the command, or pipe the output to a separate text file:

```
# Output to console
$ scontrol show job job_id

# Streaming output to a textfile
$ scontrol show job job_id > outputfile.txt

# Piping output to Grep and find lines containing the word "Time"
$ scontrol show job job_id | grep Time
```
For a full primer on grep and regular expressions, visit GNU’s page on Grep
For more information on scontrol, visit the Slurm page on scontrol

### 2.17 Slurm Flags, Partitions, and QoS

Slurm allows the use of flags to specify resources needed for a job. Below is a table describing some of the most common Slurm resource flags, followed by tables describing available Summit partitions and Quality of Service (QoS) options.

#### 2.17.1 Slurm Resource Flags

Job scripts, the `sbatch` command, and the `sinteractive` command support many different resource requests in the form of flags. These flags are available to all forms of jobs. To review all possible flags for these commands, please visit the Slurm page on sbatch. Below, we have listed some useful flags to consider when running your job script.

#### 2.17.2 Partitions

On Summit, nodes with the same hardware configuration are grouped into partitions. You will need to specify a partition using `--partition` in your job script in order for your job to run on the appropriate type of node.

These are the partitions available on Summit.

1. The `smem` partition is limited to 96 cores (2 entire nodes) across all running `smem` jobs. For example, you can run one 96-core job or up to two 48-core jobs, four 24-core jobs, ninty-six 1-core jobs, etc. If you need more memory or cores, please contact rc-help@colorado.edu.

2. Skylake nodes are seperated into 2 partitions and consists of 5 public nodes in the `ssky` partition and 15 private nodes that can be used by anyone with the `ssky-preemptable`. Jobs running on the `ssky-preemptable` partition will be pushed off the node if the condo owner of the node runs a job to their private partition. More information on Skylake nodes can be found here.

In addition to these partitions, Research Computing also provides specialized partitions for interactive and test jobs. These partitions allow quick access to a reserved set of cores provided for testing and interactive use.

3. The `shas-testing` partition is limited to 24 cores total. These cores can come from up to 2 nodes but a user is limited to maximum of 24 cores per job.

To run a job longer than 24 hours on the `shas`, `sgpu`, or `sknl` partitions, use the `long QOS`.

More details about each type of node can be found here.

#### 2.17.3 Quality of Service

On Summit, Quality of Service or QoS is used to constrain or modify the characteristics that a job can have. This could come in the form of specifying a QoS to request for a longer run time or a high priority queue for condo owned nodes. For example, by selecting the `long QoS`, a user can place the job in a lower priority queue with a max wall time increased from 24 hours to 7 days.

Normally, this slurm directive does not need to be set for most jobs. Only set a QoS when requesting a long or condo job.

The available QoS’s for Summit are:
2.18 squeue status and reason codes

The `squeue` command details a variety of information on an active job’s status with state and reason codes. *Job state codes* describe a job’s current state in queue (e.g. pending, completed). *Job reason codes* describe the reason why the job is in its current state.

The following tables outline a variety of job state and reason codes you may encounter when using squeue to check on your jobs.

2.18.1 Job State Codes

A full list of these Job State codes can be found in Slurm’s documentation.

2.18.2 Job Reason Codes

A full list of these Job Reason Codes can be found in Slurm’s documentation.

2.19 Using the Summit `ssky` Condo Partition

Summit has 20 Intel Skylake nodes that were provisioned through contributions by individual research groups. These Skylake “condo” nodes are members of the `ssky` partition. Condo contributors have preemptive access to their contributed nodes (collectively 15-of-the-20 nodes). All Summit users have access to a general-access pool of `ssky` nodes (5-of-the-20 nodes contributed by CU and CSU), and preemptable access to all `ssky` nodes.

Each Skylake node has 24 cores, 174.6 GB of available memory, and employs the AVX-512 instruction set. By comparison the Summit Haswell nodes have 24 cores, 113 GB available memory, and employ the AVX2 instruction set.

2.19.1 Access for condo contributors

Summit condo contributors have access to a summit condo contributor account with the format:

```
ucb-summit-{group} or csu-summit-{group}
```

For example if the RC operations group had contributed nodes then their condo account would be: `ucb-summit-rcops` (note: the `rcops` group is just used here for the sake of example). You can determine what accounts you are affiliated with using the command:

```
sacctmgr list associations cluster=summit user=$USER format=Account -p
```

Contributed condo nodes have been placed into discrete partitions, with a name based on the contributor account name and node type; for example, `ssky-ucb-rcops` for RC operations contributed Skylake compute nodes. You can see the list of all ssky partitions using the command:

```
scontrol show partition | grep PartitionName=ssky
```

To run a job (in this case assuming your group is `rcops`), specify the account and partition in your job script or interactive job command (e.g., `--account=ucb-summit-rcops` and `--partition=ssky-ucb-rcops`). Jobs run with condo accounts/partitions will immediately preempt jobs running on a given condo partition that were executing via the `ssky-preemptable` partition (see below).
Tip: If it is appropriate for your jobs, you can use an ordered list of preferred partitions when running your jobs to ssky. E.g., `--partition=ssky-ucb-rcops,ssky` or (if you are willing to run preemptably) `--partition=ssky-ucb-rcops,ssky,ssky-preemptable`. If your first choice of partition is not available, the next partition in the list will be attempted, and so on.

2.19.2 General and preemptable access for all Summit users

All Summit users, including condo contributors, can run jobs to either general-access (e.g., ssky) partitions or preemptable (e.g., ssky-preemptable) partitions. Jobs using the ssky partition will not be preempted, but are limited to the five general-access nodes. Jobs using the ssky-preemptable partition may run on any ssky node, but are subject to being preempted by a job from a condo contributor, as described above.

To run a job to ssky, users would simply change the “partition” flag in their job script or interactive job command to either `--partition=ssky` or `--partition=ssky-preemptable`. Other flags in their job script would remain the same, though users may want to tweak flags or recompile code (optimized for AVX-512) per the specifications of ssky nodes (see above).

2.19.3 Summary

2.20 JupyterHub (Python and R notebooks on CURC)

Jupyter notebooks are an excellent resource for interactive development and data analysis using Python, R, and other languages. Jupyter notebooks can contain live code, equations, visualizations, and explanatory text which provide an integrated environment to use, learn, and teach interactive data analysis.

CU Research Computing (CURC) operates a JupyterHub server that enables users to run Jupyter notebooks on Summit or Blanca for serial (single core) and shared-memory parallel (single node) workflows. The CURC JupyterHub uses the next-generation JupyterLab user interface. The CURC JupyterHub runs atop of Anaconda. Additional documentation on the CURC Anaconda distribution is available and may be a good pre-requisite for the following documentation outlining use of the CURC JupyterHub.

2.20.1 Step 1: Log in to CURC JupyterHub

CURC JupyterHub is available at https://jupyter.rc.colorado.edu. To log in use your RC credentials. If you do not have an RC account, please request an account before continuing.

2.20.2 Step 2: Start a notebook server

To start a notebook server, select one of the available options in the Select job profile menu under Spawner Options and click Spawn. Available options are:

- **Summit interactive (1 core, 12hr, instant access)** (a 12-hour, 1 core job on a Summit “shas” node)
- **Summit Haswell (12 cores, 4 hours)** (a 4-hour, 12 core job on a Summit “shas” node)
- **Blanca (1 core, 12hr)** (A 12-hour, 1 core job on your default Blanca partition; only available to Blanca users)
- **Blanca (12 cores, 4 hr)** (A 4-hour, 12 core job on your default Blanca partition; only available to Blanca users)
- **Blanca CSDMS (12hr)** (A 12-hour, 1 core job on the Blanca CSDMS partition; only available to Blanca CSDMS users)
- **Summit interactive Rstudio (1 core, 12hr)** (a 12-hour, 1 core Rstudio job on a Summit “shas” node)
Note: The “Summit interactive (1 core, 12hr, instant access)” option spawns a 1-core job to a partition on Summit called “shas-interactive”. This partition is intended to provide “instant” access to computing resources for Jupyter-Hub users. The caveat is that 1) users may only run one “shas-interactive” job at a time, and 2) “shas-interactive” jobs only have 1 core and 4 GB of memory allocated to them. Therefore, this option works well for light work such as interactive code development and small processing tasks, but jobs may crash if large files are ingested or memory-intensive computing is conducted. If this is your case, please consider running your workflow via a batch job on Summit, or try the “Summit Haswell (12 cores, 4 hours)” option (queue waits will be longer for this option). Dask users should either run their workflows via a batch job on Summit, or use the “Summit Haswell (12 cores, 4 hours)” option because this provides 12-cores to the Dask array. Using “shas-interactive” for Dask jobs would only provide one core to the Dask array, negating its utility.

The server will take a few moments to start. When it does, you will be taken to the Jupyter home screen, which will show the contents of your CURC /home directory in the left menu bar. In the main work area on the right hand side you will see the “Launcher” and any other tabs you may have open from previous sessions.

2.20.3 Step 3: Navigating the JupyterLab Interface

The following features are available in the JupyterLab Interface:

• **Left sidebar:** Click on a tab to change what you see in the left menu bar. Options include the file browser, a list of running kernels and terminals, a command palette, a notebook cell tools inspector, and a tabs list.

• **Left menu bar:**
  - The file browser will be active when you log in.
    * You can navigate to your other CURC directories by clicking the folder next to /home/username. Your other CURC file systems are available too: /projects/username, /pl/active (for users with PetaLibrary allocations), /scratch/username (Summit only), and /rc_scratch/username (Blanca only).
    * To open an existing notebook, just click on the notebook name in the file browser (e.g., mynotebook.ipynb).
    * Above your working directory contents are buttons to add a new Launcher, create a new folder, upload files from your local computer, and refresh the working directory.

• **Main Work Area:** Your workspaces will be in this large area on the right hand side. Under the “Launcher” tab you can:
  - Open a new notebook with any of the kernels listed:
    * **Python 3 (idp):** Python3 notebook (Intel Python distribution)
    * **Bash:** BASH notebook
    * **R:** R notebook
    * . . . and any other custom kernels you add on your own (see the section below on creating your own custom kernels).
  - Open a new console (command line) for any of the kernels.
  - Open other functions; the “Terminal” function is particularly useful, as it enables you to access the command line on the Summit or Blanca node your Jupyterhub job is currently running on.

• See Jupyter’s documentation on the JupyterLab Interface for additional information.
Tip for finding the packages available to you within a notebook

The Python 3 (idp) notebook kernels have many preinstalled packages. To query a list of available packages from a python notebook, you can use the following nomenclature:

```python
from pip._internal import main as pipmain
pipmain(['freeze'])
```

If the packages you need are not available, you can create your own custom environment and Jupyter kernel.

For users who prefer the “old school” classic Jupyter interface in favor of JupyterLab

You can access the Jupyter classic view by going to the address bar at the top of your browser and changing “lab” to “tree” in the URL. For example, if your session URL is https://jupyter.rc.colorado.edu/user/janedoe/lab, you can change this to https://jupyter.rc.colorado.edu/user/janedoe/tree.

2.20.4 Step 4: Shut down a Notebook Server

Go to the “File” menu at the top and choose “Hub Control Panel”. Use the Stop My Server button in the Control Panel to shut down the Jupyter notebook server when finished (this cancels the job you are running on Summit or Blanca). You also have the option to restart a server if desired (for example, if you want to change from a “Summit Interactive” to a “Summit Haswell” server).

Alternately, you can use the Quit button from the Jupyter home page to shut down the Jupyter notebook server.

Using the Logout button will log you out of CURC JupyterHub. It will not shut down your notebook server if one happens to be running.

2.20.5 Additional Documentation

Creating your own custom Jupyter kernels

The CURC JupyterHub runs on top of the CURC Anaconda distribution. Anaconda is an open-source python and R distribution that uses the conda package manager to easily install software and packages. Software and associated Jupyter kernels other than python and R can also be installed using conda. The following steps describe how to create your own custom Anaconda environments and associated Jupyter kernels for use on RC JupyterHub.

Follow these steps from a terminal session. You can get a new terminal session directly from Jupyter using New->Terminal.

1. Configure your conda settings

Follow our Anaconda documentation for steps on configuring your conda settings via ~/.condarc.

2. Activate the CURC Anaconda environment

```
[johndoe@shas0137 ~]$ source /curc/sw/anaconda3/latest
```

You will know that you have properly activated the environment because you should see (base) in front of your prompt. E.g.:
3. Create a new custom environment.

Follow our Anaconda documentation for steps on creating your own custom conda environment.

4. Activate your new environment

(Note: we assume here that you’ve named your environment mycustomenv; please replace mycustomenv with whatever name you gave your environment!)

5. Create your own custom kernel, which will enable you to use this environment in CURC Jupyter-hub:

For a python kernel

The first command will install the ipykernel package if not installed already. The second command will create a python kernel with the name mycustomenv with the Jupyter display name mycustomenv (note that the name and display-name are not required to match the environment name – call them anything you want). By specifying the --user flag, the kernel will be installed in /home/$USER/.local/share/jupyter/kernels (a directory that is in the default JUPYTER_PATH) and will ensure your new kernel is available to you the next time you use CURC JupyterHub.

For an R kernel

The first command will install the irkernel package if not installed already. The second command will start R. The third command, executed from within R, will create an R kernel with the name mycustomenv with the Jupyter display name mycustomenv (note that the name and display-name are not required to match the environment name – call them anything you want). The kernel will be installed in /home/$USER/.local/share/jupyter/kernels (a directory that is in the default JUPYTER_PATH) and will ensure your new kernel is available to you the next time you use CURC JupyterHub.

Notes:

- If you have already installed your own version of Anaconda or Miniconda, it is possible to create Jupyter kernels for your preexisting environments by following Step 4 above from within the active environment.
- If you need to use custom kernels that are in a location other than /home/$USER/.local/share/jupyter (for example, if your research team has a group installation of Anaconda environments located in /pl/active/<some_env>), you can create a file in your home directory named ~/.jupyterrc containing the following line:
export JUPYTER_PATH=/pl/active/<some_env>/share/jupyter

If you need assistance creating or installing environments or Jupyter kernels, contact us at rc-help@colorado.edu.

### Using Dask to spawn multi-core jobs from CURC JupyterHub

*Dask* is a flexible library for parallel computing in Python. Documentation for using Dask on RC JupyterHub is forthcoming. In the meantime, if you need help integrating Dask with Slurm so that you can run multicore jobs on the CURC JupyterHub, please contact us at rc-help@colorado.edu.

#### 2.20.6 Troubleshooting

Jupyter notebook servers spawned on RC compute resources log to `~/.jupyterhub-spawner.log`. Watching the contents of this file provides useful information regarding any problems encountered during notebook startup or execution.

#### 2.20.7 See Also

- CURC Anaconda distribution
- RC JupyterHub CHANGELOG
- JupyterLab homepage

#### 2.21 EnginFrame *(visualization, GUIs on CURC)*

NICE EnginFrame provides a 3d-accelerated remote desktop environment on an Nvidia GPU-equipped compute node. Coupled with the proprietary Desktop Cloud Visualization (DCV) VNC server, the EnginFrame service supports the use of common visualization applications in a typical desktop environment via your web browser.

##### 2.21.1 Step 1: Login to EnginFrame

**NOTE 1:** To use EnginFrame you’ll need to be on the CU Network, or if working remotely, the CU Virtual Private Network (VPN). If you already have a VPN client on your machine you can connect to `vpn.colorado.edu` through that (login credentials are your CU IdentiKey and password). If you don’t have a VPN client on your machine, you can download one from the CU Office of Information Technology.

**NOTE 2:** If you registered for a CURC account in 2021 or later you should have automatically received access to EnginFrame at the time the account was provisioned. However, if you find you cannot login per the documentation below, or if you registered for a CURC account prior to 2021, please email rc-help@colorado.edu to request access.

In your web browser, navigate to the CURC EnginFrame instance at `https://viz.rc.colorado.edu`.

From the welcome page, select “Views” from the available interfaces (or use this direct link).
Provide your RC login credentials at the login prompt. You will be prompted to use a second authentication factor (e.g., the Duo mobile app) to log in.
2.21.2 Step 2: Starting a remote desktop

After logging in, you’ll see some “Remote Desktop” options in the list of services in the left sidebar. Choose the option for the “Nvidia Quadro RTX 8000” (there may be two NVIDIA options; either will work fine).

When starting a Remote Desktop session you may customize the resources allocated to the session and other characteristics of the dispatched Slurm job. In most cases the defaults will be adequate (i.e., no need to specify an account, nodelist or reservation).

Once the session has started, an indicator of the running session appears in the Sessions list. EnginFrame will attempt to open the session automatically, but may be blocked by the browser. In that case, simply select the session from the list, or use the “click here” link in the notification text.
2.21.3 Step 3: Use graphical software

With the Remote Desktop session running and open, you should be able to run standard Linux desktop applications that have a graphical user interface (GUI). A commonly-used example would be Matlab. To start the Matlab GUI, do the following:

1. Open a terminal by clicking Applications->System Tools->MATE Terminal
2. Load the Matlab software and start Matlab by typing:

```
module load matlab/2019b
matlab
```
2.21.4 Tips and Tricks

• Desktop sessions in EnginFrame have default screensaver and screen locking functions that activate after a short period of inactivity. When this happens, the user will need to login again with their CURC credentials and accept a Duo push. To lengthen the period of inactivity that will cause the screen to lock, in the desktop session navigate to System->Control Center->Screensaver Preferences and increase the time for the “Regard the computer as idle after” category. To ensure unauthorized users cannot access the CURC system, please use this feature with caution and lock your screen manually if you will be leaving the session unattended.

2.21.5 Additional Resources

• https://www.nice-software.com/products/enginframe
• https://www.nice-software.com/products/dcv

2.22 XDMoD (CURC system metrics)

A portal for viewing metrics at the system-, partition- and user-levels.

Overview: Would you like to know average queue wait times? Do you need to better understand your and historical resource utilization, or utilization of your project account by user? The XDMoD (XD Metrics on Demand) web-based tool provides users with the ability to easily obtain detailed metrics for high performance computing resources. This open-source tool was developed by the University at Buffalo Center for Computational Research (CCR). CU Boulder Research Computing runs its own instance, CURC XDMoD that enables users to query metrics for the RMACC Summit and Blanca computing resources.

2.22.1 Getting started with XDMoD

All CURC users have access to XDMoD (CU Boulder, CSU and RMACC). At this time, login is only supported for CU Boulder users. Non-CU Boulder users may still query all of the statistics available to CU Boulder users, they just won’t have the ability to personalize metrics.

Step 1: Navigate to the CURC XDMoD instance

In your browser navigate to https://xdmod.rc.colorado.edu. Upon reaching there you will see a summary screen similar to the following image.
This screen provides some “quick stats” and summary plots that address some of the most common user questions, such as average wait times and recent resource usage by system (Summit or Blanca) and partition. These metrics may be all you need. If you want to personalize metrics you can login with your CURC username and password (currently supported for CU Boulder users only).

**Step 2: Login (CU Boulder users only)**

Choose the **Sign In** option near the upper left of the screen. This will initiate a pop-up window that gives you the option to “Sign in with CU Boulder Research Computing” or “Sign in with a local XDMoD account”.

Choose the option for “Sign in with CU Boulder Research Computing” and enter your CURC username and password. The portal uses 2-factor authentication, so you will need to accept the Duo push to your phone to complete login.
Step 3: Familiarize yourself with XDMoD

Whether or not you login, you’ll start on the “Summary” screen.

The following tabs will be available, depending on whether you are logged in:

- **Summary** (the screen you are on when you login)
- **Usage** (provides access to an expansive set of resource-wide metrics)
- **Metrics Explorer** (similar to the Usage tab, but with additional functionality)
- **Data Export** (enables raw data to be output in csv or json format for use in other apps)
- **Report Generator** (facilitates the creation of reports that can be saved and shared)
- **Job Viewer** (enables users to search for and view jobs that meet specified criteria)
- **About** (provide general information on the XDMoD software)

**- only available to users who are logged in.

Step 4: Become a pro!

XDMoD can query a seemingly endless number of metrics, more than could ever be described in this documentation. To learn how to query specific metrics, customize your views, etc., please refer to the XDMoD documentation:
https://xdmod.rc.colorado.edu/user_manual/index.php

2.22.2 Example use case

Let’s say you want to see how many core hours you project account has used over time, including the usage by user.

- Go to the Usage tab.
• In the “Metrics and Options” menu, choose CPU Hours: Total to create a graph of total CPU hours consumed over a default period.

• Click anywhere on the blue line in the graph to expose the “Drill Down” menu:

![Graph with Drilling Options]

• Choose the “PI” option. In XDMoD syntax a “PI” is a project account (e.g., ucb-general or ucb124_summit1).

• This will revise the graph to show CPU usage for different “PIs” (accounts), showing only the accounts with the greatest usage. Your account may not be shown. To find it click the Filter tab at the top and search for your project (e.g., ucb-general).

• You will now see a graph showing only core hours used by your account. To see core hours used for each user of the account, click anywhere on the line to expose the “Drill Down” menu and choose the User option.

• This will revise the graph to show CPU usage by user. If you don’t see your user of interest, you can use the Filter tab at the top to find them.

• You can change the time range of the x-axis by specifying the dates in the “Start” and “End” boxes near the top of the screen.

2.23 Load Balancer

The CU Research Computing Load Balancer is an effective tool for optimally utilizing multiple processors and nodes on the Summit HPC resource, without the need to learn OpenMP or MPI. This document assumes user knowledge of Slurm jobs, shell scripting, and some python.
2.23.1 Why Use the Load Balancer?

Suppose you have a very simple serial program that crops a photo, and you need to apply it to crop several million photos. You could rewrite the serial program into a parallel program that would utilize multiple processors to more quickly run the program over the entire set of photos (compared to doing one-at-a-time), but this would require some knowledge of parallel programming. Even worse, if your code is in a language that has limited parallelization capabilities, so this may not be an option. The easiest solution for this problem is to utilize the Load Balancer.

2.23.2 Using the Load Balancer

The Load Balancer is a tool provided by CU Boulder Research Computing that allows shell commands (for example, calls to serial programs) to be distributed amongst nodes and cores on Summit. This means code doesn’t need to be explicitly parallelized for MPI or OpenMP. Additionally, code can be written in any language that can be run from a Linux shell.

Let’s create a simple ‘Hello World’ serial python script to demonstrate the Load Balancer tool. We will call the script `hello_World.py` and it will print “Hello World from process: ” followed by a command line argument:

```python
import sys
print "Hello World from process: ", sys.argv[1]
```

Now we will create a list of calls to the python script that will be distributed to multiple cores. (Each compute node has one or more discrete compute processor; most modern processors are made up of multiple compute “cores”, each of which can operate independently and simultaneously.)

Instead of slowly typing out commands one-at-a-time, we will use a bash shell script to create our commands. In a text editor, create a bash shell script called `create_hello.sh`, that has the following text:

```bash
#!/bin/bash
for i in {1..4}
do
  echo "python hello_World.py $i;" >> lb_cmd_file
done
```

Next run the bash script by first changing permissions of the script to be executable by typing: `chmod +x create_hello.sh` and then by typing: `./create_hello.sh` at the terminal prompt. It will create a file called `lb_cmd_file` that contains 4 calls to our `hello_World.py` script:

```bash
python hello_World.py 1;
python hello_World.py 2;
python hello_World.py 3;
python hello_World.py 4;
```

Now create a job script called `run_hello.sh` that will run all instances of your python script in `lb_cmd_file` with the Load Balancer. Within the script, before using Load Balancer, we need to load the Python, and the Load Balancer utility itself. Your job script should look something like this:

```bash
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --time 00:02:00
#SBATCH --partition shas-testing
#SBATCH --ntasks=4
#SBATCH --job-name lbPythonDemo
```

(continues on next page)
Running this script via sbatch will run the commands we stored in lb_cmd_file in parallel. A successful job will result in output that looks something like this:

<table>
<thead>
<tr>
<th>Hello World from process: 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello World from process: 1</td>
</tr>
<tr>
<td>Hello World from process: 4</td>
</tr>
<tr>
<td>Hello World from process: 3</td>
</tr>
</tbody>
</table>

### 2.23.3 Additional Resources

- https://github.com/ResearchComputing/lb
- https://github.com/t-brown/lb

### 2.24 Gaussian

**Important:** Gaussian is available on Summit only to members of universities that have purchased Gaussian licenses. It cannot be run by other Summit users. Please note and abide by the licensing, rights, and citation information shown at the top of your Gaussian output files.

This document describes how to run G16 jobs efficiently on Summit. It does not attempt to teach how to use Gaussian for solving science/engineering questions.

Good general instructions can be found at here; however some minor modifications are needed when running on Summit.

#### 2.24.1 Environment

To set up your shell environment to use G16, load a Gaussian software module (e.g. `module load gaussian/16_avx2`). Nearly all necessary environment variables are configured for you via the module. You do not need to source g16.login or g16.profile if running single-node jobs, but if you are running multi-node parallel jobs you will need to add source `$g16root/g16/bsd/g16.login` (tcsh shell) or source `$g16root/g16/bsd/g16.profile` (bash shell) to your job script after you load the Gaussian module.

However, it is important to specify `GAUSS_SCRDIR` to tell G16 where to put its large scratch files. These should always be on a scratch storage system (`/scratch/summit/$USER` on Summit or `rc_scratch/$USER` on Blanca.) If `GAUSS_SCRDIR` is not set, then the scratch files will be created in whatever directory G16 is run from; if this directory is in `/projects` or `/home, then your job’s performance will be dramatically reduced.
2.24.2 Running G16

If you create a Gaussian input file named h2o_dft.com then you can execute it simply via `g16 h2o_dft`. Output from the computation will go to a file called h2o_dft.log.

2.24.3 Memory

The default dynamic memory request in G16 is frequently too small to support the amount of memory that needs to be allocated to efficiently support computations on even modest-sized molecules. If too little memory is requested, the job can crash. Thus, use the `-m` flag in your `g16` command line (e.g. `-m=48gb`) to specify at least 50% of the amount of memory your Slurm job has requested.

2.24.4 Parallel jobs

Single-node (SMP) parallelism

Many G16 functions scale well to 8 or more cores on the same node. You can specify how many cores to use via the `-p` flag to `g16` (e.g. `-p=24`). This value should correspond to the number of cores that your Slurm job has requested. You should test your G16 computations with several different core counts to see how well they scale, as there may be diminishing returns beyond a certain number of cores.

Example SMP BASH script:

```
#!/bin/bash

#SBATCH --job-name=g16-test
#SBATCH --partition=shas
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=24
#SBATCH --time=00:50:00
#SBATCH --output=g16-test.%j.out

module load gaussian/16_avx2

# Always specify a scratch directory on a fast storage space (not /home or /projects!)
export GAUSS_SCRDIR=/scratch/summit/$USER/$SLURM_JOBID
# or export GAUSS_SCRDIR=$SLURM_SCRATCH

# alternatively, to use the local SSD; max 159GB available

# the next line prevents OpenMP parallelism from conflicting with Gaussian's internal
# SMP parallelization
export OMP_NUM_THREADS=1

mkdir $GAUSS_SCRDIR  # only needed if using /scratch/summit

date  # put a date stamp in the output file for timing/scaling testing if desired
g16 -m=50gb -p=24 my_input.com

date
```

Multi-node (Linda) parallelism

In order to run on more than 24 cores in the “shas” partition on Summit, your job will need to span multiple nodes using the Linda network parallel communication model. We advise using one Linda worker per node, with multiple SMP cores per node. For example, your `g16` flags might include
which tells G16 to use 24 cores on each of three shas nodes, and to set up the connections between nodes using ssh.

Since you don’t know in advance what nodes your job will be assigned to, you will have to determine the arguments for `-w` at runtime via commands in your Slurm batch script. See the batch script example below.

Not all G16 computations scale efficiently beyond a single node! According to the G16 documentation: “HF, CIS=Direct, and DFT calculations are Linda parallel, including energies, optimizations, and frequencies. TDDFT energies and gradients and MP2 energies and gradients are also Linda parallel. Portions of MP2 frequency and CCSD calculations are Linda parallel.” As with SMP parallelism, testing the scaling of your Linda parallel computation is very important.

**Linda Parallel**

```bash
#!/bin/bash
#SBATCH --job-name=g16-test
#SBATCH --partition=shas
#SBATCH --nodes=2
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=24
#SBATCH --time=00:50:00
#SBATCH --output=g16-test.%j.out
module load gaussian/16_avx2
source $g16root/g16/bsd/g16.profile
for n in `scontrol show hostname | sort -u`; do
    echo $n -opa
done | paste -s -d, > tsnet.nodes.$SLURM_JOBID

# Always specify a scratch directory on a fast storage space (not /home or /projects!)
export GAUSS_SCRDIR=/scratch/summit/$USER/$SLURM_JOBID

# the next line prevents OpenMP parallelism from conflicting with Gaussian's internal parallelization
export OMP_NUM_THREADS=1

# the next line increases the verbosity of Linda output messages
export GAUSS_LFLAGS="-v"

mkdir $GAUSS_SCRDIR # only needed if using /scratch/summit
date # put a date stamp in the output file for timing/scaling testing
g16 -m=20gb -p=24 -w=`cat tsnet.nodes.$SLURM_JOBID` my_input.com
date
rm tsnet.nodes.$SLURM_JOBID
```

**G16 on GPU**

Please read [http://gaussian.com/running/?tabid=5](http://gaussian.com/running/?tabid=5) carefully to determine whether the K80 GPUs in Summit’s “sgpu” partition will be effective for your calculations. In many cases, SMP parallelization across all of the cores in a shas node will provide better speedup than offloading computational work to a GPU.
G16 on Knight’s Landing

We do not recommend running Gaussian16 on RMACC Summit’s “sknl” partition.

2.24.5 Sample input file

Here’s an input file that can be used for both SMP and Linda parallel testing:

```plaintext
#P b3lyp/6-31g* test stable=(opt,qonly)

Gaussian Test Job 135:
Fe=O perpendicular to ethene, in triplet state.

0 3
X
Fe X RXFe
C1 X R XC Fe 90.
C2 X R XC Fe 90. C1 180.
O X RXO C1 90. Fe 0.
H1 C1 R CH C2 CCH Fe Angle1
H2 C1 R CH C2 CCH Fe -Angle1
H3 C2 R CH C1 CCH Fe Angle2
H4 C2 R CH C1 CCH Fe -Angle2
RXFe 1.7118
RXC 0.7560
RXO 3.1306
RCH 1.1000
Angle1 110.54
Angle2 110.53
CCH 117.81
```

2.25 Matlab

Research Computing (RC) provides a large suite of software on RC resources. In this tutorial we will learn how to run Matlab on these resources. The tutorial assumes you are familiar with Matlab and basic Linux terminal commands.

There are two basic ways to run Matlab (or many other kinds of software) on RC resources. The first is through an interactive job, and the second is through a batch job. An interactive job allows one to work in real-time with Matlab. Two reasons you may want to do this would be if you are actively debugging your code, or if you would like to use the GUI (in this instance, the Matlab Desktop). However, there might be other reasons you would like to work interactively with Matlab.

The second way to run Matlab on RC resources through a batch job. This allows the job to run in the background when resources become available. You may choose to use this method if you have a large job that may wait in the queue for awhile, or if you are not debugging or in need of a GUI. Both ways to work with Matlab are below.

2.25.1 Running Matlab Interactive Jobs

Running Matlab interactive jobs on RC resources is both a simple and easy task to accomplish. In this section we will learn how to launch Matlab as an interactive job. For more information on launching interactive jobs check out our interactive jobs tutorial
Begin by launching an interactive job by loading slurm/summit into your environment and running the `sinteractive` command.

```bash
module load slurm/summit
sinteractive
```

From here you will load the Matlab module into your environment.

```bash
module load matlab
```

Lastly we will run Matlab from the terminal.

```bash
matlab -nosplash
```

By default Matlab will load an interactive terminal session. If you would like to access the Matlab GUI then simply run Matlab with X11 forwarding enabled.

To find out how you enable X11 forwarding in your terminal session, check out our X11 forwarding tutorial here.

### 2.25.2 Running Matlab Batch Jobs

Here, we will learn how to run a Matlab script in a non-interactive batch job. For more general information on batch job scripts on Summit, please see our tutorial on batch jobs

Let’s begin by constructing a small Matlab script that prints ‘hello world’ to the user. The Matlab script we will use for the purposes of this tutorial is called `hello_world.m` and contains only one line, the Matlab command:

```matlab
fprintf('Hello world
')
```

Which simply prints “Hello world” when called.

Next, we will construct our batch script that will enable us to run this job. The batch script organizes the variety of flags slurm needs to run a job and specifies the software commands we want to execute. An advantage of batch scripts is that they are easily reusable and adaptable for other similar tasks.

We will run this job using a bash script titled: `slurm_hello.sh`, which contains the following lines:

```bash
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --time=0:01:00
#SBATCH --partition=shas-testing
#SBATCH --ntasks=1
#SBATCH --job-name=Matlab_Hello_World
#SBATCH --output=Matlab_Hello_World.out
module purge
module load matlab
matlab -nodesktop -nodisplay -r 'clear; hello_world,'
```

This file has a few basic parts:

1. The first line specifies that it is a bash shell script, and ensures the rest of the lines will be interpreted in the correct shell.
2. The lines beginning with `#SBATCH` specify the Slurm parameters that will be used for this job. These lines are viewed as comments by bash, but will be read by Slurm. Of particular note is the `--output` parameter which
specifies the file where stderr and stdout (including the output from our Matlab script) will be written. For a
description of the Slurm parameters, please see the general Slurm documentation here.

3. The lines beginning with `module purge` remove any unneeded software and ensure that the appropriate
Matlab module is loaded on the compute node.

4. The final line calls Matlab and instructs it to run our script. This entire line includes commands that are specific
to Matlab; the `nodesktop` and `nodisplay` flags ensure that the Matlab Desktop will not open, and the `r` flag will run the script `hello_world`. The `clear` command forces Matlab to clear any existing variables,
and is simply included as good coding practice.

You have now completed your batch script. After saving the script and exiting your text editor, run the job as follows:

```
module purge
```

Once the job has run, the output of the Matlab script, “Hello world” will be shown in `Matlab_Hello_World.out`.

### 2.25.3 Parallel Matlab on Summit

To fully utilize the multi-core capabilities of Summit to speed up jobs, most code must first be parallelized. Matlab
has many built in tools to accomplish this task. In this tutorial we will parallelize our “Hello World” program.

Let’s begin with the Matlab script we created above called `hello_world.m`. First we will modify the `fprintf` line
so that it includes a variable ‘`i`’ that will print out the iteration of the parallel loop.

```
fprintf("Hello World from process \%i", i)
```

Next, we need to encapsulate the print statement in a parallel ‘for’ loop. Matlab uses the construct `parfor` to separate
the task into multiple threads. In order to utilize the `parfor` command one must ensure that the Parallel Computing
Toolbox is available as part of the Matlab software package. RC has this available and thus no additional action is
required on your part if you are utilizing RC resources.

The order of runtime in the loop is not guaranteed, so the output may not be in sequential order. The loop is formatted
as such:

```
parfor (i = initial_Value:final_Value, maximum_amount_of_threads)
```

For example, let’s use `parfor` to implement an 5-iteration loop with a maximum of 4 processors in our script (new lines
are highlighted here in blue):

```
parfor (i = 1:5, 4)
   fprintf("Hello, World from process \%i", i)
end
```

Now all we have left to do is modify our batch script to specify that we want to run 4 tasks on the node (we can use up
to 24 cores on each ‘shas’ node on Summit). We can also change the name of the job and the output file if we choose.

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --time=0:01:00
#SBATCH --partition=shas-testing
#SBATCH --ntasks=4
#SBATCH --job-name=Matlab_Parallel_Hello
#SBATCH --output=Matlab_Parallel_Hello.out
module purge
```

(continues on next page)
module load matlab
matlab -nodesktop -nodisplay -r 'clear; hello_world;'

Now we run the job using the `sbatch` command shown above, and our output in `Matlab_Parallel_Hello.out` will be as follows (the process order may be different in your output):

```
Hello World from process 4
Hello World from process 1
Hello World from process 2
Hello World from process 3
```

RC Matlab currently does not support parallelization across nodes, only across cores on one node.

## 2.26 Python and R with Anaconda

To support the diverse `python` and `R` workflows and high levels of customization Research Computing users require, Anaconda is installed on the CURC system. Anaconda uses the `conda` package manager to easily install software and associated packages, and supports `python`, `R`, and many other applications. The following documentation describes how to activate the CURC Anaconda distribution and our default environments, as well as how to create and activate your own custom Anaconda environments. Additional documentation on the CURC JupyterHub is available for users desiring to interact with their custom environments via Jupyter notebooks.

*Note: CURC also hosts several legacy python and R modules for those users who prefer modules over Anaconda. Type `module spider python` or `module spider R` for a list of available versions. We recommend using Anaconda.*

### 2.26.1 Configuring conda with `.condarc`

The conda package manager allows modification of default settings to be done through a text file known as the `.condarc`. This file exists within a user's `~/home/$USER/` directory and can be quickly be accessed using the file's full path at `~/.condarc`.

Your `~/home/$USER` directory is small – only 2 GB. By default, conda puts all package source code and environments in your `~/home/$USER` directory, and it will quickly become full. The steps here modify the conda configuration file to change the default locations of packages and environments to your larger `/projects/$USER` directory.

Open your `.condarc` file in your favorite text editor (e.g., `nano`): *(note: this file may not exist yet – if not, just create a new file with this name)*

```
[johndoe@shas0137]$ nano ~/.condarc
```

...and paste the following four lines:

```
pkgs_dirs:
  - /projects/$USER/.conda_pkgs
envs_dirs:
  - /projects/$USER/software/anaconda/envs
```

...then save and exit the file. You won’t need to perform this step again – it’s permanent unless you modify `.condarc` later.
The .condarc file provides a variety of settings that can be detailed to speed up your workflows. For more information on .condarc, check out the Anaconda documentation.

2.26.2 Using the CURC Anaconda environment

Follow these steps from a Research Computing terminal session on a Summit scompile node or within a Summit/Blanca batch or interactive job.

Activate the CURC Anaconda environment:

Loading the CURC Anaconda environment is a little different compared to other software on Summit. Run the following command to initialize the base Anaconda software:

```
[johndoe@shas0137 ~]$ source /curc/sw/anaconda3/latest
```

Note: The command above activates the base environment for python3, which as of 2020 is the only supported python standard. For users requiring legacy python2, you can still use conda to create a custom environment with the python2.X version of your choice (we provide an example of how to do this below).

You will know that you have properly activated the environment because you should see (base) in front of your prompt. E.g.:

```
(base) [johndoe@shas0137 ~]$
```

Using Conda:

Now that you have activated the base conda environment, you can use conda for python and R! There are two ways forward, depending on your needs. You can:

1. **Use one of CURC’s pre-installed environments.**
   - **Pros:** You can begin using one of these immediately, and they contain many of the most widely used python and R packages.
   - **Cons:** These are root-owned environments, so you can’t add additional packages.

or

2. **Create your own custom environment(s).**
   - **Pros:** You own these, so you can add packages as needed, control package versions, etc.
   - **Cons:** There really aren’t any cons, other than the time needed to create a custom environment (usually 5-30 minutes depending on the number of packages you install).

Both options are discussed below.

**Using one of CURC’s pre-installed environments:**

To use the CURC Intel Python distribution or “idp” (python v3.6.8), run the following command with Anaconda initialized:

```
(base) [johndoe@shas0137 ~]$ conda activate idp
```
You will know that you have properly activated the environment because you should see `(idp)` replace the `(base)` in front of your prompt. Now the Intel Python distribution is loaded into your environment and can be accessed with the `python` command.

To see the python packages available in the idp environment, you can type `conda list`. Similarly to use the CURC R distribution (R v3.6.0), run the following command with Anaconda initialized:

```
(base) [johndoe@shas0137 ~]$ conda activate rstudio
```

You will know that you have properly activated the environment because you should see `(rstudio)` in front of your prompt. To see the R packages available in the environment, you can type `conda list`. Now, you can use R as you normally would.

Because interactive development is more easily done locally, most CURC R users exclusively run R code within batch jobs. Should you need to use rstudio on top of R for interactive development on Summit, you can login to our system with X11-forwarding (`ssh -X`) and initiate an rstudio session from within an interactive job.

**Create your own custom environment:**

*Note: In the examples below the environment is created in `/projects/$USER/software/anaconda/envs`, which is specified under `envs_dirs` in your `.condarc` file. Environments can be installed in any user-writable location the user chooses; just add the path to ~/.condarc.*

1. **Initialize Anaconda if you haven't already done so:**

```
[johndoe@shas0137 ~]$ source /curc/sw/anaconda3/latest
(base) [johndoe@shas0137 ~]$
```

2. **Create a custom environment:**

Here we create a new environment called `mycustomenv` (you can call it anything you want!)

```
(base) [johndoe@shas0137 ~]$ conda create -n mycustomenv
```

If you want a specific version of python or R, you can modify the above command as follows (e.g.):

- **Python v2.7.16:**

  ```
  (base) [johndoe@shas0137 ~]$ conda create -n mycustomenv python==2.7.16
  ```

- **Python v3.6.8:**

  ```
  (base) [johndoe@shas0137 ~]$ conda create -n mycustomenv python==3.6.8
  ```

- **Latest version of R:**

  ```
  (base) [johndoe@shas0137 ~]$ conda create -n mycustomenv r-base
  ```

3. **Activate your new environment:**

```
(base) [johndoe@shas0137 ~]$ conda activate mycustomenv
```

If successful, your prompt will now be preceded with `(mycustomenv)`.  

4. **Install needed packages in your new environment:**

The best way to do this for python packages is to install everything you need with one command, because it forces `conda` to resolve package conflicts. E.g.:
For R packages, it is easiest to start an R session and install the packages as you normally would with “install.packages”. E.g.:

```
(mycustomenv) [johndoe@shas0137 ~]$ R
>install.packages("ggplot2")
```

For more information on managing conda environments, check out Anaconda’s documentation here.

**Basic conda commands to get you started:**

**Troubleshooting**

If you are having trouble loading a package, you can use `conda list` or `pip freeze` to list the available packages and their version numbers in your current conda environment. Use `conda install <package>` to add a new package or `conda install <package==version>` for a specific version; e.g., `conda install numpy=1.16.2`.

Sometimes conda environments can “break” if two packages in the environment require different versions of the same shared library. In these cases you try a couple of things.

- Reinstall the packages all within the same install command (e.g., `conda install <package1> <package2>`). This forces conda to attempt to resolve shared library conflicts.
- Create a new environment and reinstall the packages you need (preferably installing all with the same `conda install` command, rather than one-at-a-time, in order to resolve the conflicts).

**See Also**

- CURC JupyterHub

## 2.27 GNU Parallel

GNU Parallel is an effective tool for optimally using multiple cores and nodes on RMACC Summit to run lots of independent tasks without the need to learn OpenMP or MPI. This tutorial assumes user knowledge of Slurm jobs, shell scripting, and some Python.

### 2.27.1 Why Use GNU Parallel?

Suppose you have a very simple serial program that crops a photo, and you need to apply it to crop several million photos. You could rewrite the serial program into a parallel program that would use multiple processors to more quickly run the program over the entire set of photos (compared to doing one-at-a-time), but this would require some knowledge of parallel programming. If your code is in a language that has limited parallelization capabilities, this may not even be an option. The easiest solution for this problem is to use GNU Parallel.

### 2.27.2 Using GNU Parallel

GNU Parallel is provided as a software module on RMACC Summit. It allows shell commands (for example, calls to serial programs) to be distributed amongst nodes and cores on RMACC Summit. This means code doesn’t need to be
explicitly parallelized for MPI or OpenMP. Additionally, code can be written in any language that can be run from a Linux shell.

Let’s create a simple ‘Hello World’ serial python script to demonstrate the GNU Parallel tool. We will call the script `hello_World.py` and it will print “Hello World from task: ” followed by a command line argument:

```python
import sys

print "Hello World from task: ", sys.argv[1]
```

Now create a job script called `run_hello.sh` that will use GNU Parallel to run as many instances of your python script as you want. Before running GNU Parallel in our script, we need to load the Python and GNU Parallel modules. Your job script should look something like this:

```bash
#!/bin/bash

#SBATCH --time 00:02:00
#SBATCH --partition shas-testing
#SBATCH --ntasks=4
#SBATCH --job-name gpPythonDemo
#SBATCH --output gnuparallel.out

module purge
module load python
module load gnu_parallel

my_parallel="parallel --delay .2 -j $SLURM_NTASKS"
my_srun="srun --export=all --exclusive -n1 --cpus-per-task=1 --cpu-bind=cores"
${my_parallel} "${my_srun} python hello_World.py" ::: {1..20}
```

Note the last three lines of the script. We customize the GNU Parallel `parallel` command by creating a variable called `$my_parallel` that delays the execution of each task by 0.2 seconds (`--delay 0.2`) which mitigates bottlenecks for tasks that have heavy I/O when they start, and which specifies the number of tasks to run simultaneously (`-j $SLURM_NTASKS`). The environment variable `$SLURM_NTASKS` is set by Slurm at runtime and contains the number of `--ntasks` (cores) requested in the `#SBATCH` directives near the top of the job script (in this case the value is 4). We then customize the `srun` command so that it properly allocates the GNU parallel tasks to the allocated cores (`--export=all --exclusive -n1 --cpus-per-task=1 --cpu-bind=cores`). Note that the use of `srun` will also ensure that GNU parallel runs properly for cases where we request cores across multiple nodes (e.g., if we request `--ntasks=100`). Finally, we invoke GNU Parallel to run our python script 20 times using the customized `parallel` and `srun` commands we just created, `$my_parallel` and `$my_srun` respectively. Running this script via `sbatch` will run the commands. A successful job will result in output that looks something like this:

```
Hello World from task: 1
Hello World from task: 2
Hello World from task: 3
Hello World from task: 4
Hello World from task: 5
Hello World from task: 6
Hello World from task: 7
Hello World from task: 8
Hello World from task: 9
Hello World from task: 10
Hello World from task: 11
Hello World from task: 12
Hello World from task: 13
Hello World from task: 14
```

(continues on next page)
In this example the 20 invocations of your python script will run across the 4 cores requested; as each core finishes one task, the next remaining task will be executed on that core until all 20 have finished. The printed output above may or may not be in order depending on how quickly each task completes.

Tip: For sufficiently-large workflows one can add the `--joblog` and `--resume` flags in `$my_srun`. These flags will enable GNU Parallel to keep track of tasks it has run successfully and, if needed, rerun tasks that failed or were not executed. Additional details can be found in the links below.

### 2.27.3 Additional Resources

- [https://www.gnu.org/software/parallel/](https://www.gnu.org/software/parallel/)
- [https://github.com/ResearchComputing/HTC_Short_Course_Spring_2019](https://github.com/ResearchComputing/HTC_Short_Course_Spring_2019)
- [https://rcc.uchicago.edu/docs/tutorials/kicp-tutorials/running-jobs.html](https://rcc.uchicago.edu/docs/tutorials/kicp-tutorials/running-jobs.html)

### 2.28 Containerization on Summit

When installing software, you may come across applications that have complex chains of dependencies that are challenging to compile and install. Some software may require very specific versions of libraries that may not be available on Summit or conflict with libraries needed for other applications. You may also need to move between several workstations or HPC platforms, which often requires reinstalling your software on each system. Containers are a good way to tackle all of these issues and more.

#### 2.28.1 Containerization Fundamentals

Containers build upon an idea that has long existed within computing: hardware can be emulated through software. **Virtualization** simulates some or all components of computation through a software application. Virtual machines use this concept to generate an entire operating system as an application on a host system. Containers follow the same idea, but at a much smaller scale and contained within a system’s kernel.

**Containers** are portable compartmentalizations of some or all of the following: An operating system, software, libraries, data, and workflows. Containers offer:

- Portability: containers can run on any system equipped with its specified container manager.
- Reproducibility: because containers are instances of prebuilt isolated software, software will always execute the same every time.

Containers distinguish themselves through their low computational overhead and their ability to utilize all of a host system’s resources. Building containers is a relatively simple process that starts with a container engine.
2.28.2 Docker

Docker is by far the most popular container engine, and can be used on any system where you have administrative privileges. Because of this need for administrative privileges, Docker containers cannot be built or run directly on Research Computing resources. To utilize a Docker container on Research Computing resources please build a singularity image using a Docker image as a base.

See the documentation on Singularity (below) if you wish to run a Docker container on RMACC Summit or Blanca.

2.28.3 Singularity

Singularity is a containerization software package that does not require users to have administrative privileges when running containers, and can thus be safely used on Research Computing resources such as RMACC Summit and Blanca. Singularity is preinstalled on Research Computing resources, so all that is needed to run Singularity containers is to load the Singularity module on a compute node on RMACC Summit or Blanca:

```
module load singularity/3.0.2
```

Much like Docker, Singularity is a containerization software designed around compartmentalization of applications, libraries, and workflows. This is done through the creation of Singularity images which can be run as ephemeral Singularity containers. Unlike Docker, however, Singularity does not manage images, containers, or volumes through a central application. Instead, Singularity generates saved image files that can either be mutable or immutable based on compression.

Singularity Hub

Singularity Hub is a container registry that allows users to pull images from a server and into a system with Singularity installed. Singularity Hub uses Github to host image recipes, builds images in the cloud from these recipes, and places the resulting images in the Singularity Hub registry.

Note: You do not need an account with Github if you only wish to pull Singularity images.

https://singularity-hub.org/

Singularity Hub has a variety of useful prebuilt images for different software packages and workflows so be sure to check if the software you need is already available.

Note: As of 2019, there are presently two Singularity container registries. The former is Singularity Hub, described above, which is managed by Stanford University and Lawrence Berkeley National Laboratory. The latter is the Sylabs Singularity Container Library, which was created in late 2018 when Singularity was spun off into the private company Sylabs. Below we provide documentation on how to pull images from either repository, and on how to build images on Singularity Hub via Github, and in the Sylabs Singularity Container Library using their “Remote Builder” functionality.

Pulling Singularity Images

Because we cannot build our own Singularity images on HPC systems, we must instead bring our images over from another location. Pulling images from public repositories is often the easiest solution to this problem.

We can use the `singularity pull` command to remotely download our chosen image file. The command requires the container registry we would like to use, followed by the repository’s name:

```
singularity pull <container-registry>://<repository-name>
```
A container registry is simply a server that manages uploaded containers. Some examples of these container registries include Docker Hub, Singularity Hub, and the Singularity Container Library.

Pull from Docker Hub:

```
singularity pull docker://another:example
```

Pull from Singularity Hub:

```
singularity pull shub://example:repo
```

Pull from Singularity Container Library (Singularity version 3.0 and greater):

```
singularity pull library://example:repo
```

Lastly we can rename the Singularity image file pulled from a repository by utilizing the \( -n/\--name \) flag.

```
singularity pull \-n ExampleContainer.sif shub://example:tag
```

Example:

Pulling the Docker image of the latest tag of ubuntu can be done with the following command:

```
singularity pull docker://ubuntu:latest
```

### Running a Singularity image as a container

Singularity images can be run as containers much like Docker images. Singularity commands, however, follow a bit more nuanced syntax depending on what you’d like to do. After pulling your image from either Docker Hub or Singularity Hub, you can run the image by using the `singularity run` command. Type:

```
singularity run <image-name>
```

Running a Singularity container will execute the container’s default program that is specified in container definition file. To execute specific programs in your container, we can use the `singularity exec` command, and then specify the program:

```
singularity exec <image-name> <program>
```

Much like specifying an application in Docker, this will allow a user to execute any program that is installed within your container. Unlike Docker however, you do not need to specify a shell application to shell into the container. We can simply use the `singularity shell` command:

```
singularity shell <image-name>
```

Example:

Say we have a Singularity image that contains python 3.7 as the default software, and we want to run python from the container. We can do this with the command:

```
singularity run python-cont.img
```

If the default application for the image is not python we could run python as follows:

```
singularity exec python-cont.img python
```
File Access

By default most user-owned files and directories are available to any container that is run on RMACC Summit and Blanca (this includes files in /home/$USER, /projects/$USER, /scratch/summit/$USER and /rc_scratch/$USER). This means that normally a user will not need to bind any folders to the container’s directory tree. Furthermore, a container will also have access to the files in the same folder where it was initialized.

Sometimes, however, certain folders that are not bound by default may be necessary to run your application. To bind any additional folders or files to your Singularity container, you can utilize the -B flag in your singularity run, exec, and shell commands. To bind an additional folder to your Singularity container, type:

```
singularity run -B /source/directory:/target/directory sample-image.img
```

Additionally you can bind directories by utilizing the SINGULARITY_BINDPATH environment variable. Simply export a list of directory pairs you would like to bind to your container:

```
export SINGULARITY_BINDPATH=/source/directory1:/target/directory1,\
/source/directory2:/target/directory2
```

Then run, execute, or shell into the container as normal.

Building a Singularity image

Important: You cannot build Singularity images directly on Summit. If you cannot build an image on your local machine you will need to build it on Singularity Hub or Sylabs Remote Builder.

Singularity Build

Just like Docker, Singularity allows a user to build images using a definition file. The file is saved with the name “Singularity” and contains instructions on how to prepare a Singularity image file. Just like a Dockerfile, this file has a variety of directives that allow for the customization of your image. A sample image would look something like this:

```
Bootstrap: shub
From: ubuntu
%help
   I am help text!
%setup
   apt-get update
   apt-get install nano
   apt-get install gcc
%runscript
   echo "hello! I am a container!"
```

Once you have written your Singularity recipe, you can build the application either remotely (see below) or locally with the singularity build command. To build a Singularity image locally, type:

```
sudo singularity build <img-name.img> <recipe-name.def>
```

Again, it is important to note that if you build an image locally as described above, you must build your image on a computer that you have administrative privileges on. If you do not have administrative privileges you will not be able to build the container in this manner. Fortunately, there are other ways to build containers remotely, which are discussed next.
Building Images Remotely with Singularity Hub

To build images with Singularity Hub, you must first create a Github account at https://github.com/join if you do not have one already. After completing this step log into your github account and create an empty repository.

After creating your repository, upload a Singularity definition file named “Singularity” to the repository. This is all we need to generate our Singularity image.

Now, log into Singularity Hub with your Github credentials and navigate to “My Container Collections” and click the link “Add a Collection.” From here a list of Github repositories you contribute to will be listed. Simply click the button on the repository you wish to add to Singularity Hub.

Your container should build automatically if you have a recipe file named “Singularity” within your repository. By default Singularity Hub will attempt to build any time something is pushed to the github repository. This can be changed in the settings tab in the containers build page. If the build fails the first time, revise the Singularity recipe and the build will initiate again.

More on building containers: https://www.sylabs.io/guides/3.0/user-guide/build_a_container.html

Building Images Remotely with the Singularity Remote Builder

With Singularity 3.0, users have the ability to build containers remotely through Sylabs remote builder. Unlike Singularity Hub though, the Singularity remote builder can be utilized directly on the command line from RMACC Summit or Blanca without needing to upload to a repository.

To begin using Singularity Remote Builder, navigate to your home directory and run the commands:

```
mkdir .singularity
cd .singularity
```

Now on your local machine, navigate to: https://cloud.sylabs.io/auth

... and log into Sylabs with your Google, Github, Gitlab, or Microsoft account. Once you have logged into Sylabs, provide a label for your token under the field “Create A New Access Token” and click “Create a new Token.” This will generate a large string that will be read by Singularity on RMACC Summit or Blanca.

Now on RMACC Summit or Blanca run the command:

```
echo "<your-token>" > ~/.singularity/sylabs-token
```

After this you can now build containers through the Sylabs remote builder on RMACC Summit or Blanca. Simply load Singularity 3.0.2 into your module stack and run the command:

```
singularity build --remote <desired-image-name> <your-recipe>
```

Building MPI-enabled Singularity images

MPI-enabled Singularity containers can be deployed on RMACC Summit with the caveat that the MPI software within the container stays consistent with MPI software available on the system. This requirement diminishes the portability of MPI-enabled containers, as they may not run on other systems without compatible MPI software. Regardless, MPI-enabled containers can still be a very useful option in many cases.

Here we provide an example of using a gcc compiler with OpenMPI. RMACC Summit uses an Omni-Path interconnect (a low latency network fabric that enables MPI to be efficiently implemented across nodes). In order to use a Singularity container with OpenMPI (or any MPI) on Summit, there are two requirements:
Singularity container needs to have Omni-Path libraries installed inside. OpenMPI needs to be installed both inside and outside of the Singularity container. More specifically, the SAME version of OpenMPI needs to be installed inside and outside (at least very similar, you can sometimes get away with two different minor versions, ex: 2.1 and 2.0).

The following Singularity recipe ensures that OpenMPI 2.0.1 is installed in the image, which matches the openmpi/2.0.1 module that is available on RMACC Summit. This recipe can be used as a template to build your own MPI-enabled container images for RMACC Summit and can be found here.

Once you’ve built the container with one of the methods outlined above, you can place it on RMACC Summit and run it on a compute node. The following is an example of running a gcc/OpenMPI container with Singularity on RMACC Summit. The syntax is a normal MPI run where multiple instances of a Singularity image are run. The following example runs mpi_hello_world with MPI from a container.

```bash
ml gcc/6.1.0
ml openmpi/2.0.1
ml singularity/3.0.2
mpirun -np 4 singularity exec openmpi.sif mpi_hello_world"
```

Note that it is also possible to build intel/IMPI containers for use on RMACC Summit, which are likely to have enhanced performance on Summit’s intel architecture compared to gcc/OpenMPI containers. If you would like assistance building MPI-enabled containers contact rc-help@colorado.edu.

### 2.29 Coding best practices

Programmers employ numerous tactics to ensure readable and organized code. These include:

1. using naming conventions for variables;
2. placing whitespaces, indentations and tabs within code;
3. adding comments throughout to aid in interpretation.

In this tutorial we will examine these concepts.

#### 2.29.1 Variable Naming Conventions

Variable naming is an important aspect in making your code readable. Naming variables follow a simple idea: Create variables that describe their function and which follow a consistent theme throughout your code. Let’s take a look at some naming conventions.

**Multiword Delimited**

This convention is to separate words in a variable name without the use of whitespace. Whitespace within variables is usually difficult for programming languages to interpret. Because of this variables must be delimated in some way. Here are several delimiting conventions commonly used in code:

**Snakecase:** Words are delimited by an underscore.

```plaintext
variable_one
variable_two
```

**Pascalcase:** Words are delimited by capital letters.
VariableOne
VariableTwo

**Camelcase:** Words are delimited by capital letters, except the initial word.

variableOne
variableTwo

**Hungarian Notation:** This notation describes the variable type or purpose at the start of the variable name, followed by a descriptor that indicates the variable’s function. The Camelcase notation is used to delimit words.

arrDistributeGroup //Array called “Distribute Group”
sUserName //String called “User Name”
iRandomSeed //Integer called “Random Seed”

These conventions are by no means binding, but instead examples of how many programmers format their code. Consistency and readability are key ideas that should be utilized in the naming of variables. Regardless of how you choose to name your variables, always ensure that your naming conventions are consistent throughout the code. Consistency allows others to more easily understand your code.

### 2.29.2 Function and Class Naming conventions

Much like variable naming conventions, functions and classes should also follow a similar structure of descriptive titles delimited with the conventions described above. An important aspect of naming is to ensure your classes, functions, and variables can be distinguished from each other. For example, one could use Camelcase and Pascalcase for functions and classes respectively, while reserving Snakecase or Hungarian notation for variable names. Distinguishing functions, classes, and variables with different naming conventions can greatly aid other users of your code, and can eliminate the need for large sections of comments that would otherwise be needed.

### 2.29.3 Whitespace and Tabbing

Whitespace and tabbing are critical for organizing code. Whitespace is any space in your code that is not taken up by physical characters. Tabbing is one way to create whitespace in consistent units using the ‘tab’ key. While some languages ignore whitespace and tabbing all together, others entirely rely on the concept. Because of this, it is important to ensure your code is consistently organized. Whitespace and tabbing are often used to clarify nested loops and logical statements. Let’s look at some examples of C code that demonstrate the effect of whitespace.

#### Minimal Whitespace:

```c
#include <stdio.h>
int main(int argc, char const *argv[]) { int loop_Sum = 0; for(int i = 0; i < 50; i++) loop_Sum += 1; printf("%d\n", loop_Sum); return 0; }
```

#### Liberal use of Whitespace:

```c
#include <stdio.h>

int main(int argc, char** argv){
    int loop_Sum = 0;
    for(int i = 0; i < 50; i++){
        loop_Sum += 1;
    }
}
```

(continues on next page)
printf("%d
", loop_Sum);
    return 0;
}

It is important to note that like variable naming, whitespace can be utilized in various different styles and approaches. Just remember to use whitespace and tabbing in a consistent, readable style throughout your code.

2.29.4 Commenting your code

Commenting may be the most important way to organize and segment code. Comments are sections of code that the compiler ignores, which are useful to label code and segment code. For example, one can label loops, scopes, functions, and other code snippets with the expected function of the code. Lines with comments in code are preceded by a symbol that tells the compiler/interpreter to ignore that line, for example “#”, “!”, or “;” (the symbol depends on the programming language). Let’s look at some C++ code that uses no comments vs. comments.

No Comments:

```cpp
#include <stdio.h>
#include <vector>
using namespace std;

int main(int argc, char** argv){
    vector<int> multiples;
    for(int i = 0; i < 50; i++){
        if(i % 5 == 0){
            multiples.push_back(i);
        }
    }
    for(int i; i < multiples.size(); i++){
        printf("%d is a multiple of 5
", multiples[i]);
    }
    return 0;
}
```

Comments:

```cpp
#include <stdio.h>
#include <vector>
using namespace std;

int main(int argc, char** argv){
    //Declare a vector to store values
    vector<int> multiples;
    //Iterate from 0 to 50
    for(int i = 0; i <= 50; i++){
        //for(int i = 0; i < 50; i++){
        //    printf("%d is a multiple of 5\n", multiples[i]);
        
        printf("%d is a multiple of 5\n", multiples[i]);
    }
    return 0;
}
```
Note that it is entirely possible to “over comment” code. Code should be designed in an efficient, consistent and intuitive manner such that comments enhance user understanding but are not needed to describe the entire code.

2.30 Fundamentals of parallel programming

Parallel computation can often be a bit more complex compared to standard serial applications. This page will explore these differences and describe how parallel programs work in general. We will also assess two parallel programming solutions that utilize the multiprocessor environment of a supercomputer.

Useful Links:
https://computing.llnl.gov/tutorials/parallel_comp/#Whatis

2.30.1 Why Parallel?

Say you are attempting to assemble a 10,000-piece jigsaw puzzle* on a rainy weekend. The number of pieces is staggering, and instead of a weekend it takes you several weeks to finish the puzzle. Now assume you have a team of friends helping with the puzzle. It progresses much faster, and you are able to finish the puzzle within desired weekend. This principle is the central idea behind parallel computation. You can dramatically cut down on computation by splitting one large task into smaller tasks that multiple processors can perform all at once. With parallel processes a task that would normally take several weeks can potentially be reduced to several hours.

- Puzzle analogy for describing parallel computing adopted from Henry Neeman’s Supercomputing in Plain English tutorial series.

2.30.2 Serial and Parallel Processes

A serial process is simply a process that is run entirely by one core of one processor. This means tasks are run one after another as they appear in code. This is analogous to you doing the jigsaw puzzle on your own.

A parallel process is a process that is divided among multiple cores in a processor or set of processors. Each sub process can have its own set of memory as well as share memory with other processes. This is analogous to doing the puzzle with the help of friends. Because a supercomputer has a large network of nodes with many cores, we must implement parallelization strategies with our applications to fully utilize a supercomputing resource.

2.30.3 How parallel computation works

Parallel computation connects multiple processors to memory that is either pooled or connected via high speed networks. Here are three different types of parallel computation.

```c
// If iterator is a multiple of 5 add it to the vector
if (i % 5 == 0){
    multiples.push_back(i);
}
}

// Print all items that are a multiple of 5
for (int i; i < multiples.size(); i++){
    printf("%d is a multiple of 5\n", multiples[i]);
}
return 0;
```
Shared Memory Model:
In a shared memory model all processors to have access to a pool of common memory that they can freely use.

(Image courtesy of LLNL https://computing.llnl.gov/tutorials/parallel_comp/)

Distributed Memory Model:
In a distributed memory model a separate segment of memory is available to each processor. Because memory isn’t shared inherently, information that must be shared between processes is sent over a network.

(Image courtesy of LLNL https://computing.llnl.gov/tutorials/parallel_comp/)

Distributed/Shared Model:
A split distributed/shared model is a hybrid between a shared and distributed model and has the properties of both. Each separate set of processors sharing a set of common memory is called a node.

(Image courtesy of LLNL https://computing.llnl.gov/tutorials/parallel_comp/)

Summit utilizes a hybrid distributed/shared model: there are 380 nodes, each having 24 cores.

2.30.4 Tools for Parallel Programming

Two common solutions for creating parallel code are OpenMP and MPI. Both solutions are limited to the C++ or Fortran programming languages. (Though other languages may be extended with C++ or Fortran code to utilize OpenMP or MPI.)

OpenMP

OpenMP (“Open Multi-Processing”) is a compiler-side application programming interface (API) for creating code that can run on a system of threads. No external libraries are required in order to parallelize your code. OpenMP is often considered more user friendly with thread safe methods and parallel sections of code that can be set with simple scoping. OpenMP is, however, limited to the amount of threads available on a node – in other words, it follows a shared memory model. On Summit, this means that no more than 24 processors can be utilized with programs parallelized using OpenMP.

MPI

MPI (“Message Passing Interface”) is a library standard for handling parallel processing. Unlike OpenMP, MPI has much more flexibility in how individual processes handle memory. MPI is also compatible with multi-node structures, allowing for very large, multi-node applications (i.e, distributed memory models). MPI is, however, often considered less accessible and more difficult to learn. Regardless, learning the library provides a user with the ability to maximize processing ability. MPI is a library standard, meaning there are several libraries based on MPI that you can use to develop parallel code. Two solutions available on Summit are OpenMPI and Intel MPI.

2.31 MPI Best practices

MPI or Message Passing Interface is a powerful library standard that allows for the parallel execution of applications across multiple processors on a system. It differs from other parallel execution libraries like OpenMP by also allowing
a user to run their applications across multiple nodes. Unfortunately it can sometimes be a bit tricky to run a compiled
MPI application within an HPC resource. The following page outlines best practices in running your MPI applications
across Summit and Blanca resources.

Please note that this page does not go over compiling or optimization of MPI applications.

### 2.31.1 MPI Compatible Compilers and Libraries

#### Loading the new Module Stack

Currently, only CURC’s older modules are visible within our base module stack. Luckily newer modules do exist
within Summit and simply needs to be activated with a few commands. To see more newer versions of compilers and
MPI libraries, simply run the following commands on a compute or compile node:

```bash
source /curc/sw/opt/spack/linux-rhel7-haswell/gcc-4.8.5/lmod-8.3-
...pwxkksysumgym34z7b7cq52uny77cfx41/lmod/lmod/init/bashexport
MODULEPATH=/curc/sw/modules/spack/spring2020/linux-rhel7-x86_64/Core
```

To return to the base module stack, simply exit the current node and return to a compile or compute node.

#### Selecting your Compiler and MPI

The two families of compilers that are available to users of the system are Intel and GCC with Intel MPI and OpenMPI
available respectively. To load Intel with Intel MPI or GCC with OpenMPI, run one the following commands from a
job script or compile node.

**Intel**

```bash
module load intel intel-mpi
```

or...

**GCC**

```bash
module load gcc openmpi
```

It is important to note that GCC and OpenMPI should be paired with the `SLURM_EXPORT_ENV=ALL` environment
variable when loading the pair into a job script. This will ensure the job can function when run on a login node!

In most situations you will want to try to compile and run your applications utilizing the Intel set of compilers and MPI
libraries. Most cores on Summit and Blanca are of Intel architecture, so utilizing Intel will ensure the highest level of
optimization comes from your compiler. GCC should only be utilized when your application cannot be compiled on
intel software or if compiler specific optimizations exist within your code.

### 2.31.2 Commands to Run MPI Applications

Regardless of compiler or MPI distribution, there are 3 “wrapper” commands that will run MPI applications: `mpirun`,
`mpiexec`, and `srun`. These “wrapper” commands should be used after loading in your desired compiler and MPI
distribution and simply prepend whatever application you wish to run. Each command offers their own pros and cons
alongside nuance as to how they function.
mpirun is probably the most direct method to run MPI applications with the command being tied to the distribution. This means distribution dependent flags can be passed into the command as well as the command being the most reliable to work with:

```
mpirun -np <core-count> ./<your-application>
```

mpiexec is a standardized MPI command execution command that allows for more general MPI flags to be passed. This means that commands you use of one MPI distribution can be used on another MPI distribution.

```
mpiexec -np <core-count> ./<your-application>
```

The final command srun is probably the most abstracted away from a specific implementation. This command lets Slurm figure out specific MPI features that are available in your environment and handles running the process as a job. This command is usually a little less efficient and may have some issues in reliability.

```
srun -n <core-count> ./<your-application>
```

RC usually recommends mpirun and mpiexec for simplicity and reliability when running MPI applications. srun should be used sparing to avoid issues with execution.

### 2.31.3 Running MPI on Summit

Because Summit exists as a mostly homogeneous compute cluster, running MPI applications across nodes isn’t usually too troublesome.

Simply select your Compiler/MPI and MPI wrapper command you wish to use, and place them all in a job script. Below is an example of what this can look like. In this example we run a 48 core, 4 hour job with the Intel compiler and Intel distribution of MPI:

```bash
#!/bin/bash
#SBATCH --nodes=2
#SBATCH --time=04:00:00
#SBATCH --partition=shas
#SBATCH --ntasks=48
#SBATCH --job-name=mpi-job
#SBATCH --output=mpi-job.%j.out

source /curc/sw/opt/spack/linux-rhel7-haswell/gcc-4.8.5/1mod-8.3-...pvwkxsyumgym34z7b7cq52uny77cfx41/lmod/lmod/init/bashexport MODULEPATH=/curc/sw/...modules/spack/spring2020/linux-rhel7-x86_64/Core

module purge
module load intel intel-mpi

#Run a 48 core job across 2 nodes:
mpirun -np $SLURM_NTASKS /path/to/mycode.exe

#Note: $SLURM_NTASKS has a value of the amount of cores you requested
```

### 2.31.4 Running MPI on Blanca

Unlike Summit, Blanca is often a bit more complicated because of the diverse variety of nodes it is composed of. In general, there are 3 types of nodes on Blanca that can all run single node multi-core MPI processes that may require additional flags and parameters to achieve cross node parallelism.
General Blanca Nodes

General Blanca nodes are not intended to run multi-node processes but this can still be achieved through the manipulation of some network fabric settings. In order to achieve cross node parallelism we must force MPI to utilize ethernet instead of our normal high speed network fabric. We can enforce this with various `mpirun` flags for each respective compiler.

**Intel**

```bash
mpirun -genv I_MPI_FABRICS shm:tcp
```

**OpenMPI**

```bash
mpirun --mca btl tcp
```

Please note that this does not ensure high speed communications in message passing, but it will allow for basic parallelization across nodes.

ROCE Enabled Nodes

Some Blanca nodes are equipped with high speed network fabrics that are a bit more suited for cross node MPI processes. These nodes are labeled as *RoCE enabled* and require applications to be compiled with GCC 8.4.0 and OpenMPI_UCX 4.0.0 in order to function. Users should compile their applications using this compiler and MPI distribution that is found on Summit’s base module stack.

If you are unsure if your node supports this feature then you can check by using the `scontrol` command on your node.

```bash
scontrol show node <your-bnode>
```

You will be presented a block information that details all the nodes features. The key feature you should look for is `fdr`. If your Blanca node lacks this feature then it is not ROCE Enabled.

Blanca HPC

Blanca HPC come equipped with high speed interconnects that would normally allow for high speed communication between nodes. These nodes currently only support Intel and Intel MPI compiler/MPI combo. Unfortunately there also exists a few issues regarding fabrics available to each node group so a bit of nuance must be had when running your applications.

Blanca HPC nodes can easily be distinguished from other Blanca nodes with the node’s name in the cluster. Nodes will clearly be distinguished with the `bhpc` prefix.

There are currently 2 fabrics that can really be utilized within Blanca HPC depending on the node. The most efficient of these is the `ofi` fabric. Regardless of node, users should run a test job with this fabric to validate if it is supported on their nodes. You can do this with the following flag:

**Intel**

```bash
mpirun -genv I_MPI_FABRICS shm:ofi
```

If this fabric returns an error then your jobs should function by loading the `ofa` fabric instead. This fabric can be loaded similarly with:

**Intel**

```bash
mpirun -genv I_MPI_FABRICS shm:ofa
```
2.32 Using MPI with C

Parallel programs enable users to fully utilize the multi-node structure of supercomputing clusters. Message Passing Interface (MPI) is a standard used to allow several different processors on a cluster to communicate with each other. In this tutorial we will be using the Intel C++ Compiler, GCC, IntelMPI, and OpenMPI to create a multiprocessor ‘hello world’ program in C++. This tutorial assumes the user has experience in both the Linux terminal and C++.

Resources:

- http://www.dartmouth.edu/~rc/classes/intro_mpi/intro_mpi_overview.html
- http://mpitutorial.com/tutorials/
- http://condor.cc.ku.edu/~grobe/docs/intro-MPI-C.shtml
- https://computing.llnl.gov/tutorials/mpi/

2.32.1 Setup and “Hello, World”

Begin by logging into the cluster and using ssh to log in to a compile node. This can be done with the command:

```bash
ssh scompile
```

Next we must load MPI into our environment. Begin by loading in your choice of C++ compiler and its corresponding MPI library. Use the following commands if using the GNU C++ compiler:

**GNU C++ Compiler**

```bash
module load gcc
module load openmpi
```

Or, use the following commands if you prefer to use the Intel C++ compiler:

**Intel C++ Compiler**

```bash
module load intel
module load impi
```

This should prepare your environment with all the necessary tools to compile and run your MPI code. Let’s now begin to construct our C++ file. In this tutorial, we will name our code file: **hello_world_mpi.cpp**

Open **hello_world_mpi.cpp** and begin by including the C standard library <stdio.h> and the MPI library <mpi.h>, and by constructing the main function of the C++ code:

```cpp
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    return 0;
}
```

Now let’s set up several MPI directives to parallelize our code. In this ‘Hello World’ tutorial we’ll be utilizing the following four directives:

**MPI_Init()**: This function initializes the MPI environment. It takes in the addresses of the C++ command line arguments argc and argv.

**MPI_Comm_size()**: ...
This function returns the total size of the environment via quantity of processes. The function takes in the MPI environment, and the memory address of an integer variable.

*MPI_Comm_rank():*

This function returns the process id of the processor that called the function. The function takes in the MPI environment, and the memory address of an integer variable.

*MPI_Finalize():*

This function cleans up the MPI environment and ends MPI communications.

These four directives should be enough to get our parallel ‘hello world’ running. We will begin by creating two variables, *process_Rank* and *size_Of_Cluster*, to store an identifier for each of the parallel processes and the number of processes running in the cluster respectively. We will also implement the *MPI_Init* function which will initialize the mpi communicator:

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster
    MPI_Init(&argc, &argv);
    return 0;
}
```

Let’s now obtain some information about our cluster of processors and print the information out for the user. We will use the functions *MPI_Comm_size()* and *MPI_Comm_rank()* to obtain the count of processes and the rank of a process respectively:

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster;
    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);
    printf("Hello World from process %d of %d\n", process_Rank, size_Of_Cluster);
    return 0;
}
```

Lastly let’s close the environment using *MPI_Finalize()*:

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster;
    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);
    MPI_Finalize();
}
```
printf("Hello World from process %d of %d\n", process_rank, size_Of_Cluster);
    MPI_Finalize();
    return 0;
}

Now the code is complete and ready to be compiled. Because this is an MPI program, we have to use a specialized compiler. Be sure to use the correct command based off of what compiler you have loaded.

**OpenMPI**

```bash
mpic++ hello_world_mpi.cpp -o hello_world_mpi.exe
```

**Intel MPI**

```bash
mpiicc hello_world_mpi.cpp -o hello_world_mpi.exe
```

This will produce an executable we can pass to Summit as a job. In order to execute MPI compiled code, a special command must be used:

```bash
mpirun -np 4 ./hello_world_mpi.exe
```

The flag `-np` specifies the number of processor that are to be utilized in execution of the program.

In your job script, load the same compiler and OpenMPI choices you used above to compile the program, and run the job with slurm to execute the application. Your job script should look something like this:

**OpenMPI**

```bash
#!/bin/bash
#SBATCH -N 1
#SBATCH --ntasks 4
#SBATCH --job-name parallel_hello
#SBATCH --partition shas-testing
#SBATCH --time 0:01:00
#SBATCH --output parallel_hello_world.out

module purge
module load gcc
module load openmpi

mpirun -np 4 ./hello_world_mpi.exe
```

**Intel MPI**

```bash
#!/bin/bash
#SBATCH -N 1
#SBATCH --ntasks 4
#SBATCH --job-name parallel_hello
#SBATCH --partition shas-testing
#SBATCH --time 0:01:00
#SBATCH --output parallel_hello_world.out

module purge
module load intel
```

(continues on next page)
module load impi

mpirun -np 4 ./hello_world_mpi.exe

It is important to note that on Summit, there is a total of 24 cores per node. For applications that require more than 24 processes, you will need to request multiple nodes in your job. Our output file should look something like this:

| Hello World from process 3 of 4 |
| Hello World from process 2 of 4 |
| Hello World from process 1 of 4 |
| Hello World from process 0 of 4 |

Ref: http://www.dartmouth.edu/~rc/classes/intro_mpi/hello_world_ex.html

### 2.3.2.2 MPI Barriers and Synchronization

Like many other parallel programming utilities, synchronization is an essential tool in thread safety and ensuring certain sections of code are handled at certain points. `MPI_Barrier` is a process lock that holds each process at a certain line of code until all processes have reached that line in code. `MPI_Barrier` can be called as such:

```c
MPI_Barrier(MPI_Comm comm);
```

To get a handle on barriers, let's modify our “Hello World” program so that it prints out each process in order of thread id. Starting with our “Hello World” code from the previous section, begin by nesting our print statement in a loop:

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster;
    MPI_Init(&argc, &argv);  
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);  
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);
    for (int i = 0, i < size_Of_Cluster, i++)
    {
        printf("Hello World from process %d of %d\n", process_Rank, size_Of_Cluster);
    }
    MPI_Finalize();
    return 0;
}
```

Next, let’s implement a conditional statement in the loop to print only when the loop iteration matches the process rank.

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster;
    MPI_Init(&argc, &argv);  
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);  
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);
    for (int i = 0, i < size_Of_Cluster, i++)
    {
        printf("Hello World from process %d of %d\n", process_Rank, size_Of_Cluster);
    }
    MPI_Finalize();
    return 0;
}
```

(continues on next page)
Lastly, implement the barrier function in the loop. This will ensure that all processes are synchronized when passing through the loop.

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster;
    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);
    for(int i = 0, i < size_Of_Cluster, i++){
        if(i == process_Rank){
            printf("Hello World from process %d of %d\n", process_Rank, size_Of_Cluster);
        }
    }
    MPI_Finalize();
    return 0;
}
```

Compiling and running this code will result in this output:

```
Hello World from process 0 of 4
Hello World from process 1 of 4
Hello World from process 2 of 4
Hello World from process 3 of 4
```

### 2.32.3 Message Passing

Message passing is the primary utility in the MPI application interface that allows for processes to communicate with each other. In this tutorial, we will learn the basics of message passing between 2 processes.

Message passing in MPI is handled by the corresponding functions and their arguments:

```c
MPI_Send(void* message, int count, MPI_Datatype datatype, int dest, int tag, MPI_Comm, communicator);
MPI_Recv(void* data, int count, MPI_Datatype datatype, int from, int tag, MPI_Comm, communicator);
```

The arguments are as follows:
Let’s implement message passing in an example:

Example

We will create a two-process process that will pass the number 42 from one process to another. We will use our “Hello World” program as a starting point for this program. Let’s begin by creating a variable to store some information.

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster, message_Item;

    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);

    MPI_Finalize();
    return 0;
}
```

Now create if and else if conditionals that specify appropriate process to call `MPI_Send()` and `MPI_Recv()` functions. In this example we want process 1 to send out a message containing the integer 42 to process 2.

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Cluster, message_Item;

    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);

    if(process_Rank == 0){
        message_Item = 42;
        printf("Sending message containing: %d\n", message_Item)
    }
    else if(process_Rank == 1){
```
Lastly we must call `MPI_Send()` and `MPI_Recv()`. We will pass the following parameters into the functions:

```c
MPI_Send(&message_Item, 1, MPI_INT, 1, 1, MPI_COMM_WORLD);
MPI_Recv(&message_Item, 1, MPI_INT, 0, 1, MPI_COMM_WORLD, MPI_STATUS_IGNORE);
```

Let's implement these functions in our code:

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv) {
    int process_Rank, size_Of_Cluster, message_Item;
    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Cluster);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);
    if(process_Rank == 0){
        message_Item = 42;
        MPI_Send(&message_Item, 1, MPI_INT, 1, 1, MPI_COMM_WORLD);
        printf("Message Sent: %d
", message_Item);
    }
    else if(process_Rank == 1){
        MPI_Recv(&message_Item, 1, MPI_INT, 0, 1, MPI_COMM_WORLD, MPI_STATUS_IGNORE);
        printf("Message Received: %d
", message_Item);
    }
    MPI_Finalize();
    return 0;
}
```

Compiling and running our code with 2 processes will result in the following output:
2.32.4 Group Operators: Scatter and Gather

Group operators are very useful for MPI. They allow for swaths of data to be distributed from a root process to all other available processes, or data from all processes can be collected at one process. These operators can eliminate the need for a surprising amount of boilerplate code via the use of two functions:

**MPI_Scatter**:
```c
void* send_Var; //Address of the variable that will be scattered.
int send_Count; //Number of elements that will be scattered.
MPI_Datatype send_Type; //MPI Datatype of the data that is scattered.
void* recv_Var; //Address of the variable that will store the scattered data.
int recv_Count; //Number of data elements that will be received per process.
MPI_Datatype recv_Type; //MPI Datatype of the data that will be received.
int root_Process; //The rank of the process that will scatter the information.
MPI_Comm comm; //The MPI_Communicator.
```

**MPI_Gather**:
```c
void* send_Var; //Address of the variable that will be sent.
int send_Count; //Number of data elements that will sent.
MPI_Datatype send_Type; //MPI Datatype of the data that is sent.
void* recv_Var; //Address of the variable that will store the received data.
int recv_Count; //Number of data elements per process that will be received.
MPI_Datatype recv_Type; //MPI Datatype of the data that will be received.
int root_Process; //The rank of the process rank that will gather the information.
MPI_Comm comm; //The MPI_Communicator.
```

In order to get a better grasp on these functions, let’s go ahead and create a program that will utilize the scatter function. Note that the gather function (not shown in the example) works similarly, and is essentially the converse of the scatter function. Further examples which utilize the gather function can be found in the MPI tutorials listed as resources at the beginning of this document.

**Example**

We will create a program that scatters one element of a data array to each process. Specifically, this code will scatter the four elements of an array to four different processes. We will start with a basic C++ main function along with variables to store process rank and number of processes.

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv){
    int process_Rank, size_Of_Comm;
    return 0;
}
```

Now let’s setup the MPI environment using `MPI_Init`, `MPI_Comm_size`, `MPI_Comm_rank`, and `MPI_Finalize`.
```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv)
{
    int process_Rank, size_Of_Comm;

    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Comm);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);

    MPI_Finalize();
    return 0;
}
```

Next let’s generate an array named `distro_Array` to store four numbers. We will also create a variable called `scattered_Data` that we shall scatter the data to.

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv)
{
    int process_Rank, size_Of_Comm;
    int distro_Array[4] = {39, 72, 129, 42};
    int scattered_Data;

    MPI_Init(&argc, &argv);
    MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Comm);
    MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);

    MPI_Finalize();
    return 0;
}
```

Now we will begin the use of group operators. We will use the operator scatter to distribute `distro_Array` into `scattered_Data`. Let’s take a look at the parameters we will use in this function:

```c
MPI_Scatter(
    &distro_Array,  //Address of array we are scattering from.
    1,              //Number of items we are sending each processor
    MPI_INT,        //MPI Datatype of scattering array.
    &scattered_Data, //Address of array we are receiving scattered data.
    1,              //Amount of data each process will receive.
    MPI_INT,        //MPI Datatype of receiver array.
    0,              //Process ID that will distribute the data.
    MPI_COMM_WORLD  //MPI Communicator.
)
```

Let’s see this implemented in code. We will also write a print statement following the scatter call:

```c
#include <stdio.h>
#include <mpi.h>

int main(int argc, char** argv)
{
    int process_Rank, size_Of_Comm;
    int distro_Array[4] = {39, 72, 129, 42};
    int scattered_Data;
```

(continues on next page)
MPI_Init(&argc, &argv);
MPI_Comm_size(MPI_COMM_WORLD, &size_Of_Comm);
MPI_Comm_rank(MPI_COMM_WORLD, &process_Rank);

MPI_Scatter(&distro_Array, 1, MPI_INT, &scattered_Data, 1, MPI_INT, 0, MPI_COMM_WORLD);

printf("Process has received: %d \n", scattered_Data);
MPI_Finalize();
return 0;
}

Running this code will print out the four numbers in the distro array as four separate numbers each from different processors (note the order of ranks isn’t necessarily sequential):

| Process has received: 39 |
| Process has received: 72 |
| Process has received: 129 |
| Process has received: 42 |

### 2.33 Using MPI with Fortran

Parallel programs enable users to fully utilize the multi-node structure of supercomputing clusters. Message Passing Interface (MPI) is a standard used to allow different nodes on a cluster to communicate with each other. In this tutorial we will be using the Intel Fortran Compiler, GCC, IntelMPI, and OpenMPI to create a multiprocessor programs in Fortran. This tutorial assumes the user has experience in both the Linux terminal and Fortran.

**Helpful MPI tutorials:**
- http://www.dartmouth.edu/~rc/classes/intro_mpi/intro_mpi_overview.html
- http://condor.cc.ku.edu/~grobe/docs/intro-MPI.shtml
- https://computing.llnl.gov/tutorials/mpi/

#### 2.33.1 Setup and “Hello World”

Begin by logging into the cluster and using ssh to log in to a Summit compile node. This can be done with the command:

```bash
ssh scompile
```

Next we must load MPI into our environment. Begin by loading in the Fortran compiler and OpenMPI. Use the following commands if using the GNU Fortran compiler:

**GNU Fortran Compiler**

```bash
module load gcc
module load openmpi
```

Or, use the following commands if you prefer to use the Intel Fortran compiler:

**Intel Fortran Compiler**

```bash
```
module load intel
module load impi

This should prepare your environment with all the necessary tools to compile and run your MPI code. Let’s now begin to construct our Fortran program. In this tutorial, we will name our program file: `hello_world_mpi.f90`

Open `hello_world_mpi.f90` and begin by including the mpi library `'mpif.h'`, and titling the program `hello_world_mpi`

```fortran
PROGRAM hello_world_mpi
include 'mpif.h'
```

Now let’s set up several MPI directives to parallelize our code. In this ‘Hello World’ tutorial we will be calling the following four functions from the MPI library:

- **`MPI_INIT()`**:  
  This function initializes the MPI environment. It takes in the an error handling variable.

- **`MPI_COMM_SIZE()`**:  
  This function returns the total size of the environment in terms of the quantity of processes. The function takes in the MPI environment, an integer to hold the commsize, and an error handling variable.

- **`MPI_COMM_RANK()`**:  
  This function returns the process id of the process that called the function. The function takes in the MPI environment, an integer to hold the comm rank, and an error handling variable.

- **`MPI_FINALIZE()`**:  
  This function cleans up the MPI environment and ends MPI communications.

These four directives are enough to get our parallel ‘hello world’ program running. We will begin by creating three integer variables, `process_Rank`, `size_Of_Cluster`, and `ierror` to store an identifier for each of the parallel processes, store the number of processes running in the cluster, and handle error codes respectively. We will also implement the `MPI_Init` function which will initialize the mpi communicator:

```fortran
PROGRAM hello_world_mpi
include 'mpif.h'
integer process_Rank, size_Of_Cluster, ierror
call MPI_INIT(ierr)
```

Let’s now obtain some information about our cluster of processors and print the information out for the user. We will use the functions `MPI_Comm_size()` and `MPI_Comm_rank()` to obtain the count of processes and the rank of a given process respectively:

```fortran
PROGRAM hello_world_mpi
include 'mpif.h'
integer process_Rank, size_Of_Cluster, ierror
call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierror)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierror)
print *, 'Hello World from process: ', process_Rank, 'of ', size_Of_Cluster
```

Lastly let’s close the environment using `MPI_Finalize()`:
PROGRAM hello_world_mpi
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierr, tag

call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierr)
print *, 'Hello World from process: ', process_Rank, 'of ', size_Of_Cluster
call MPI_FINALIZE(ierr)
END PROGRAM

Now the code is complete and ready to be compiled. Because this is an MPI program, we have to use a specialized compiler. The compilation command will be one of the following:

GNU Fortran Compiler

mpif90 hello_world_mpi.f90 -o hello_world_mpi.exe

Intel Fortran Compiler

mpiifort hello_world_mpi.f90 -o hello_world_mpi.exe

This will produce an executable we can pass to Summit as a job. In order to execute MPI compiled code, a special command must be used:

mpirun -np 4 ./hello_world_mpi.exe

The flag -np specifies the number of processor that are to be utilized in execution of the program. In your job script, load the same compiler and OpenMPI choices you used above to create and compile the program, and run the job to execute the application. Your job script should look something like this:

GNU Fortran Compiler

#!/bin/bash
#SBATCH -N 1
#SBATCH --ntasks 4
#SBATCH --job-name parallel_hello
#SBATCH --partition shas-testing
#SBATCH --time 0:01:00
#SBATCH --output parallel_hello_world.out
module purge
module load gcc
module load openmpi
mpirun -np 4 ./hello_world_mpi.exe

Intel Fortran Compiler

#!/bin/bash
#SBATCH -N 1
#SBATCH --ntasks 4
#SBATCH --job-name parallel_hello

(continues on next page)
It is important to note that on Summit, there are 24 cores per node. For applications that require more than 24 processes, you will need to request multiple nodes in your job (i.e., "-N ").

Our output file should look something like this (note the order of ranks isn’t necessarily sequential):

<table>
<thead>
<tr>
<th>Hello World from process</th>
<th>3 of 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello World from process</td>
<td>2 of 4</td>
</tr>
<tr>
<td>Hello World from process</td>
<td>1 of 4</td>
</tr>
<tr>
<td>Hello World from process</td>
<td>0 of 4</td>
</tr>
</tbody>
</table>

Ref: http://www.dartmouth.edu/~rc/classes/intro_mpi/hello_world_ex.html

### 2.33.2 MPI Barriers and Synchronization

Like many other parallel programming utilities, synchronization is an essential tool in thread safety and ensuring certain sections of code are handled at certain points. MPI_BARRIER is a process lock that holds each process at a certain line of code until all processes have reached that line. MPI_BARRIER can be called as such:

```
call MPI_BARRIER(MPI_com comm, integer ierror);
```

To get a handle on barriers, let’s modify our “Hello World” program so that it prints out each process in order of thread id. Starting with our “Hello World” code from the previous section, begin by putting our print statement in a loop:

```
PROGRAM hello_world_mpi
include 'mpif.h'
integer process_Rank, size_Of_Cluster, ierror

call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierr)

DO i = 0, 3, 1
   print *, 'Hello World from process: ', process_Rank, ' of ', size_Of_Cluster
END DO

call MPI_FINALIZE(ierr)
END PROGRAM
```

Next, let’s implement a conditional statement in the loop to print only when the loop iteration matches the process rank.

```
PROGRAM hello_world_mpi
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierror

call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierr)

DO i = 0, 3, 1
   if (i == process_Rank) then
      print *, 'Hello World from process: ', process_Rank, ' of ', size_Of_Cluster
   end if
END DO

call MPI_FINALIZE(ierr)
END PROGRAM
```
integer process_Rank, size_Of_Cluster, ierror

call MPI_INIT(ierrerror)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierror)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierror)

DO i = 0, 3, 1
  IF(i == process_Rank) THEN
    print *, 'Hello World from process: ', process_Rank, 'of ', size_Of_Cluster
  END IF
END DO

call MPI_FINALIZE(ierrerror)
END PROGRAM

Lastly, implement the barrier function in the loop. This will ensure that all processes are synchronized when passing through the loop.

PROGRAM hello_world_mpi
include 'mpif.h'
integer process_Rank, size_Of_Cluster, ierror

call MPI_INIT(ierrerror)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierror)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierror)

DO i = 0, 3, 1
  IF(i == process_Rank) THEN
    print *, 'Hello World from process: ', process_Rank, 'of ', size_Of_Cluster
  END IF
  call MPI_BARRIER( MPI_COMM_WORLD, i_error)
END DO

call MPI_FINALIZE(ierrerror)
END PROGRAM

Compiling and running this code will result in the following output (note the ranks are now sequential):

Hello World from process 0 of 4
Hello World from process 1 of 4
Hello World from process 2 of 4
Hello World from process 3 of 4

2.33.3 Message Passing

Message passing is the primary utility in the MPI application interface that allows for processes to communicate with each other. Next, we will learn the basics of message passing between two processes. Message passing in MPI is handled by the corresponding functions and their arguments:

call MPI_SEND(integer message, integer count, MPI_Datatype datatype, integer dest,
integer tag, MPI Comm comm, integer ierror);
call MPI_RECV(integer data, integer count, MPI_Datatype datatype, integer from,
integer tag, MPI Comm comm, MPI_Status* status, integer ierror);

2.33. Using MPI with Fortran
The arguments are as follows:

**MPI_SEND:**

- `integer message` !Variable storing message you are sending.
- `integer count` !Number of elements being sent through the array.
- `MPI_Datatype datatype` !The MPI-specific data type being passed through the array.
- `integer dest` !Process rank of destination process.
- `integer tag` !Message tag.
- `MPI_Comm comm` !The MPI Communicator handle.
- `integer ierror` !An error handling variable.

**MPI_RECV:**

- `integer message:` !Variable storing message you are receiving.
- `integer count:` !Number of elements being sent through the array.
- `MPI_Datatype datatype:` !The MPI-specific data type being passed through the array.
- `integer from:` !Process rank of sending process.
- `integer tag:` !Message tag.
- `MPI_Comm comm:` !The MPI Communicator handle.
- `MPI_Status* status:` !Status object.
- `integer ierror` !An error handling variable.

Let's implement message passing in an example:

### 2.33.4 Example

We will pass the number 42 from one process to another. We will use our “Hello World” program as a starting point for this program. Let’s begin by renaming our program and creating a variable to store some information.

```fortran
PROGRAM send_recv_mpi
include 'mpif.h'
integer process_Rank, size_Of_Cluster, ierror, message_Item
call MPI_INIT(ierrerr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierrerr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierrerr)
call MPI_FINALIZE(ierrerr)
END PROGRAM
```

Now create ‘if’ and ‘else if’ conditionals that specify the appropriate processes to call `MPI_SEND()` and `MPI_RECV()` functions. In this example we want process 1 to send out a message containing the integer 42 to process 2.

```fortran
PROGRAM send_recv_mpi
include 'mpif.h'
integer process_Rank, size_Of_Cluster, ierror, message_Item

call MPI_INIT(ierrerr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierrerr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierrerr)
call MPI_FINALIZE(ierrerr)

IF(process_Rank == 0) THEN
   message_Item = 42
ENDIF
```

(continues on next page)
Lastly we must call MPI_SEND() and MPI_RECV(). We will pass in the following parameters into the functions:

**MPI_SEND**
```
message_Item, !Variable storing the message we are sending.
1, !Number of elements handled by the array.
MPI_INT, !MPI_TYPE of the message we are sending.
1, !Rank of receiving process
1, !Message Tag
MPI_COMM_WORLD !MPI Communicator
ierror !Error Handling Variable
```

**MPI_RECV**
```
message_Item, !Variable storing the message we are receiving.
1, !Number of elements handled by the array.
MPI_INT, !MPI_TYPE of the message we are sending.
0, !Rank of sending process
1, !Message Tag
MPI_COMM_WORLD !MPI Communicator
MPI_STATUS_IGNORE !MPI Status Object
ierror !Error Handling Variable
```

Lets implement these functions in our code:

**PROGRAM** send_recv_mpi
```
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierror, message_Item

call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierr)

IF (process_Rank == 0) THEN
  message_Item = 42
  call MPI_SEND(message_Item, 1, MPI_INT, 1, 1, MPI_COMM_WORLD, ierr)
  print *, "Sending message containing: ", message_Item
ELSE IF (process_Rank == 1) THEN
  call MPI_RECV(message_Item, 1, MPI_INT, 0, 1, MPI_COMM_WORLD, MPI_STATUS_IGNORE, ierr)
  print *, "Received message containing: ", message_Item
END IF

call MPI_FINALIZE(ierr)
END PROGRAM
```

Compiling and running a batch job with our code that requests 2 processes (-ntasks 2) will result in the following output:

2.33. Using MPI with Fortran
2.33.5 Group Operators: Scatter and Gather

Group operators are very useful for MPI. They allow for swaths of data to be distributed from a root process to all other available processes, or data from all processes can be collected at one process. These operators can eliminate the need for a surprising amount of boilerplate code via two functions:

**MPI_Scatter:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>integer send_Var</code></td>
<td>Variable storing the values that will be scattered.</td>
</tr>
<tr>
<td><code>integer send_Count</code></td>
<td>Number of elements that will be scattered.</td>
</tr>
<tr>
<td><code>MPI_Datatype send_Type</code></td>
<td>MPI Datatype of the data that is scattered.</td>
</tr>
<tr>
<td><code>integer recv_Var</code></td>
<td>Variable that will store the scattered data.</td>
</tr>
<tr>
<td><code>integer recv_Count</code></td>
<td>Number of data elements that will be received per process.</td>
</tr>
<tr>
<td><code>MPI_Datatype recv_Type</code></td>
<td>MPI Datatype of the data that will be received.</td>
</tr>
<tr>
<td><code>integer root_Process</code></td>
<td>The rank of the process that will scatter the information.</td>
</tr>
<tr>
<td><code>MPI_Comm comm</code></td>
<td>The MPI_Communicator.</td>
</tr>
<tr>
<td><code>integer ierror</code></td>
<td>An error handling variable.</td>
</tr>
</tbody>
</table>

**MPI_Gather:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>integer send_Var</code></td>
<td>Variable storing the value that will be sent.</td>
</tr>
<tr>
<td><code>integer send_Count</code></td>
<td>Number of data elements that will sent.</td>
</tr>
<tr>
<td><code>MPI_Datatype send_Type</code></td>
<td>MPI Datatype of the data that is sent.</td>
</tr>
<tr>
<td><code>integer recv_Var</code></td>
<td>Variable that will store the gathered data.</td>
</tr>
<tr>
<td><code>integer recv_Count</code></td>
<td>Number of data elements per process that will be received.</td>
</tr>
<tr>
<td><code>MPI_Datatype recv_Type</code></td>
<td>MPI Datatype of the data that will be received.</td>
</tr>
<tr>
<td><code>integer root_Process</code></td>
<td>The rank of the process rank that will gather the information.</td>
</tr>
<tr>
<td><code>MPI_Comm comm</code></td>
<td>The MPI_Communicator.</td>
</tr>
<tr>
<td><code>integer ierror</code></td>
<td>An error handling variable.</td>
</tr>
</tbody>
</table>

In order to get a better grasp on these functions, let’s go ahead and create a program that will utilize the scatter function. Note that the gather function (not shown in the example) works similarly, and is essentially the converse of the scatter function. Further examples which utilize the gather function can be found in the MPI tutorials listed as resources at the beginning of this document.

**Example**

We will create a new program that scatters one element of a data array to each process. Specifically, this code will scatter the four elements of a vector array to four different processes. We will start with a Fortran header along with variables to store process rank and number of processes.

```fortran
PROGRAM scatter_mpi
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierror, message_Item
END PROGRAM
```
Now let’s setup the MPI environment using MPI_Init, MPI_Comm_size, MPI_Comm_rank, and MPI_Finalize:

```fortran
PROGRAM scatter_mpi
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierror, message_Item

call MPI_INIT(ierror)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierror)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierror)
call MPI_FINALIZE(ierror)
END PROGRAM
```

Next let’s generate an array named distro_Array to store four numbers. We will also create a variable called scattered_Data to which we will scatter the data.

```fortran
PROGRAM scatter_mpi
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierror, message_Item
integer scattered_Data
integer, dimension(4) :: distro_Array
distro_Array = (/39, 72, 129, 42/)
call MPI_INIT(ierror)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_Of_Cluster, ierror)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierror)
call MPI_FINALIZE(ierror)
END PROGRAM
```

Now we will begin the use of group operators. We will use the operator scatter to distribute distro_Array into scattered_Data. Let’s take a look at the parameters we will use in this function:

```fortran
MPI_Scatter(
  distro_Array, !Array we are scattering from.
  1, !Number of items we are sending each processor
  MPI_INT, !MPI Datatype of scattering array.
  scattered_Data, !Variable to which are receiving scattered data.
  1, !Amount of data each process will receive.
  MPI_INT, !MPI Datatype of receiver array.
  0, !Process ID that will distribute the data.
  MPI_COMM_WORLD !MPI Communicator.
  ierr !Error Handling Variable
)
```

Let’s implement this in the code. We will also write a print statement following the scatter call:

```fortran
PROGRAM scatter_mpi
include 'mpif.h'

integer process_Rank, size_Of_Cluster, ierror, message_Item
integer scattered_Data
integer, dimension(4) :: distro_Array
distro_Array = (/39, 72, 129, 42/)
```

(continues on next page)
call MPI_INIT(ierr)
call MPI_COMM_SIZE(MPI_COMM_WORLD, size_of_Cluster, ierr)
call MPI_COMM_RANK(MPI_COMM_WORLD, process_Rank, ierr)
call MPI_Scatter(distro_Array, 1, MPI_INT, scattered_Data, 1, MPI_INT, 0, MPI_COMM_WORLD, ierr);

print *, "Process ", process_Rank, " received: ", scattered_Data
call MPI_FINALIZE(ierr)

END PROGRAM

Running this code will print out the four numbers in the distro array as four separate numbers each from different processes (note the order of ranks isn’t necessarily sequential):

<table>
<thead>
<tr>
<th>Process</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>0</td>
<td>72</td>
</tr>
<tr>
<td>3</td>
<td>129</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
</tr>
</tbody>
</table>

2.34 Using OpenMP with C

Because Summit is a cluster of CPUs, the most effective way to utilize these resources involves parallel programming. Probably the simplest way to begin parallel programming involves the utilization of OpenMP. OpenMP is a Compiler-side solution for creating code that runs on multiple cores/threads. Because OpenMP is built into a compiler, no external libraries need to be installed in order to compile this code. These tutorials will provide basic instructions on utilizing OpenMP on both the GNU C++ Compiler and the Intel C++ Compiler.

This guide assumes you have basic knowledge of the command line and the C++ Language.

Resources:

Much more in depth OpenMP and MPI C++ tutorial:

• https://computing.llnl.gov/tutorials/openMP/

2.34.1 Parallel “Hello, World” Program

In this section we will learn how to make a simple parallel hello world program in C++. Let’s begin with the creation of a program titled: parallel_hello_world.cpp. From the command line run the command:

nano parallel_hello_world.cpp

We will begin with include statements we want running at the top of the program:

#include <stdio.h>
#include <omp.h>

These flags allow us to utilize the stdio and omp libraries in our program. The <omp.h> header file will provide openmp functionality. The <stdio.h> header file will provide us with print functionality.

Let’s now begin our program by constructing the main function of the program. We will use omp_get_thread_num() to obtain the thread id of the process. This will let us identify each of our threads using that unique id number.
Let’s compile our code and see what happens. We must first load the compiler module we want into our environment. We can do so as such:

GCC:

module load gcc

Or

Intel:

module load intel

From the command line, where your code is located, run the command:

GCC:

g++ parallel_hello_world.cpp -o parallel_hello_world.exe -fopenmp

Or

Intel:

icc parallel_hello_world.cpp -o parallel_hello_world.exe -qopenmp

This will give us an executable we can run as a job to Summit. Simply run the job specifying slurm to run the executable. Your job script should look something like this:

```
#!/bin/bash
#SBATCH --nodes=1
#SBATCH --time=0:01:00
#SBATCH --partition=shas-testing
#SBATCH --ntasks=4
#SBATCH --job-name=CPP_Hello_World
#SBATCH --output=CPP_Hello_World.out
./parallel_hello_world.exe
```

Our output file should look like this:

```
Hello from process: 0
```

As you may have noticed, we only get one thread giving us a Hello statement. How do we parallelize the print statement? We parallelize it with pragma ! The `#pragma omp parallel { ... }` directive creates a section of code that will be run in parallel by multiple threads. Let’s implement it in our code:

```
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    printf("Hello from process: %d\n", omp_get_thread_num());
    return 0;
}
```

(continues on next page)
int main(int argc, char** argv){
    #pragma omp parallel
    {
        printf("Hello from process: %d\n", omp_get_thread_num());
    }
    return 0;
}

We must do one more thing before achieving parallelization. To set the amount of threads we want OpenMP to run on, we must set an Linux environment variable to be specify how many threads we wish to use. The environment variable: OMP_NUM_THREADS will store this information. Changing this variable does not require recompilation of the the program, so this command can be placed in either the command line or on your job script:

```bash
export OMP_NUM_THREADS=4
```

Important to note: this environment variable will need to be set every time you exit your shell. If you would like to make this change permanent you will need to add these lines to your .bash_profile file in your home directory:

```bash
OMP_NUM_THREADS=4;
export OMP_NUM_THREADS
```

Now let's re-compile the code and run it to see what happens:

**GCC**

```bash
g++ parallel_hello_world.cpp -o parallel_hello_world.exe -fopenmp
```

Or

**Intel**

```bash
icc parallel_hello_world.cpp -o parallel_hello_world.exe -qopenmp
```

Running our job script and we should end with an output file similar to this one:

```
Hello from process: 3
Hello from process: 0
Hello from process: 2
Hello from process: 1
```

Don’t worry about order of processes that printed, the threads will print out at varying times.

### 2.34.2 Private vs. Shared Variables

Memory management is a quintessential component of any parallel program that involves data manipulation. In this section, we will learn about the different variable types in OpenMP as well as a simple implementation of these types into the program we made in the previous section.

OpenMP has a variety of tools that can be utilized to properly describe how the parallel program should handle variables. These tools come in the forms of shared and private variable type classifiers.

- Private types create a copy of a variable for each process in the parallel system.
- Shared types hold one instance of a variable for all processes to share.

To indicate private or shared memory, declare the variable before your parallel section and annotate the pragma omp directive as such:
Variables that are created and assigned inside of a parallel section of code will be inherently private, and variables created outside of parallel sections will be inherently public.

Example

Let’s adapt our ‘Hello World’ code to utilize private variables as an example. Starting with the code we left off with, let’s create a variable to store the thread id of each process.

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int thread_id;
    #pragma omp parallel
    {
        printf("Hello from process: %d\n", omp_get_thread_num());
    }
    return 0;
}
```

Now let’s define `thread_id` as a private variable. Because we want each task to have a unique thread id, using the `private(thread_id)` will create a separate instance of `thread_id` for each task.

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int thread_id;
    #pragma omp parallel private(thread_id)
    {
        printf("Hello from process: %d\n", omp_get_thread_num());
    }
}
```

Lastly, let’s assign the thread id to our private variable and print out the variable instead of the `omp_get_thread_num()` function call:

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int thread_id;
    #pragma omp parallel private(thread_id)
    {
        thread_id = omp_get_thread_num();
        printf("Hello from process: %d\n", thread_id );
    }
    return 0;
}
```

Compiling and running our code will result in a similar result to our original hello world:
Hello from process: 3
Hello from process: 0
Hello from process: 2
Hello from process: 1

2.34.3 Barrier and Critical Directives

OpenMP has a variety of tools for managing processes. One of the more prominent forms of control comes with the 
barrier:

```
#pragma omp barrier
```

... and the critical directives:

```
#pragma omp critical { ... }
```

The barrier directive stops all processes for proceeding to the next line of code until all processes have reached the 
barrier. This allows a programmer to synchronize sequences in the parallel process.

A critical directive ensures that a line of code is only run by one process at a time, ensuring thread safety in the body 
of code.

Example

Let’s implement an OpenMP barrier by making our ‘Hello World’ program print its processes in order. Beginning 
with the code we created in the previous section, let’s nest our print statement in a loop which will iterate from 0 to the 
max thread count. We will retrieve the max thread count using the OpenMP function: `omp_get_max_threads()`

Our ‘Hello World’ program will now look like:

```
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    //define loop iterator variable outside parallel region
    int i;
    int thread_id;
    #pragma omp parallel
    {
        thread_id = omp_get_thread_num();
        //create the loop to have each thread print hello.
        for(i = 0; i < omp_get_max_threads(); i++)
        {
            printf("Hello from process: %d\n", thread_id);
        }
    }
    return 0;
}
```

Now that the loop has been created, let’s create a conditional that requires the loop to be on the proper iteration to print 
its thread number:
Lastly, to ensure one process doesn’t get ahead of another, we need to add a barrier directive in the code. Let’s implement one in our loop:

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int i;
    int thread_id;

    #pragma omp parallel
    {
        thread_id = omp_get_thread_num();

        for(i = 0; i < omp_get_max_threads(); i++)
        {
            if(i == thread_ID){
                printf("Hello from process: %d\n", thread_id);
            }
        }
    }
    return 0;
}
```

Compiling and running our code should order our print statements as such:

```
Hello from process: 0
Hello from process: 1
Hello from process: 2
Hello from process: 3
```

### 2.34.4 Work Sharing Directive: omp for

OpenMP’s power comes from easily splitting a larger task into multiple smaller tasks. Work-sharing directives allow for simple and effective splitting of normally serial tasks into fast parallel sections of code. In this section we will learn how to implement omp for directive.
The directive `omp for` divides a normally serial for loop into a parallel task. We can implement this directive as such:

```c
#pragma omp for { ... }
```

### Example

Let’s write a program to add all the numbers between 1 and 1000. Begin with a main function and the stdio and omp headers:

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    return 0;
}
```

Now let’s go ahead and setup variables for our parallel code. Let's first create variables `partial_Sum` and `total_Sum` to hold each thread’s partial summation and to hold the total sum of all threads respectively.

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int partial_Sum, total_Sum;
    return 0;
}
```

Next let’s begin our parallel section with `pragma omp parallel`. We will also set `partial_Sum` to be a private variable and `total_Sum` to be a shared variable. We shall initialize each variable in the parallel section.

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int partial_Sum, total_Sum;
    #pragma omp parallel private(partial_Sum) shared(total_Sum)
    {
        partial_Sum = 0;
        total_Sum = 0;
    }
    return 0;
}
```

Let’s now set up our work sharing directive. We will use the `#pragma omp for` to declare the loop as to be work sharing, followed by the actual C++ loop. Because we want to add all number from 1 to 1000, we will initialize our loop at one and end at 1000.

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv){
    int partial_Sum, total_Sum;
    #pragma omp for shared(total_Sum) reduction(+:total_Sum)
    for (int i = 1; i <= 1000; i++)
        total_Sum += i;
    return 0;
}
```
Now we must join our threads. To do this we must use a critical directive to create a thread safe section of code. We do this with `#pragma omp critical` directive. Lastly we add partial sum to total sum and print out the result outside the parallel section of code.

```c
#include <stdio.h>
#include <omp.h>

int main(int argc, char** argv)
{
    int partial_Sum, total_Sum;

    #pragma omp parallel private(partial_Sum) shared(total_Sum)
    { 
        partial_Sum = 0;
        total_Sum = 0;

        #pragma omp for
        { 
            for(int i = 1; i <= 1000; i++)
            {
                partial_Sum += i;
            }
        }

        //Create thread safe region.
        #pragma omp critical
        { 
            //add each threads partial sum to the total sum
            total_Sum += partial_Sum;
        }
        
        printf("Total Sum: %d\n", total_Sum);
        return 0;
    }
}
```

This will complete our parallel summation. Compiling and running our code will result in this output:

```
Total Sum: 500500
```
2.35 Using OpenMP with Fortran

Because Summit is a cluster of CPUs, parallel programming is the most effective way to utilize these resources. Probably the simplest way to begin parallel programming is utilization of OpenMP. OpenMP is a Compiler-side solution for creating code that runs on multiple cores/threads. Because OpenMP is built into a compiler, no external libraries need to be installed in order to compile this code. These tutorials provide basic instructions on utilizing OpenMP on both the GNU Fortran Compiler and the Intel Fortran Compiler.

This guide assumes you have basic knowledge of the command line and the Fortran Language.

Resources:

Much more in depth OpenMP tutorial:

- https://computing.llnl.gov/tutorials/openMP/

2.35.1 Parallel “Hello, World” Program

In this section we will learn how to make a simple parallel hello world program in Fortran. Let’s begin with creation of a program titled: parallel_hello_world.f90. From the command line run the command:

```
nano parallel_hello_world.f90
```

We will begin with the program title and the use statement at the top of the program:

```
PROGRAM Parallel_Hello_World
USE OMP_LIB
```

These flags allow us to utilize the omp library in our program. The ‘USE OMP_LIB’ line of code will provide openmp functionality.

Let’s now begin our program by constructing the main body of the program. We will use `OMP_GET_THREAD_NUM()` to obtain the thread id of the process. This will let us identify each of our threads using that unique id number.

```
PROGRAM Parallel_Hello_World
USE OMP_LIB

PRINT *, “Hello from process: ”, OMP_GET_THREAD_NUM()

END
```

Let’s compile our code and see what happens. We must first load the compiler module we want into our environment. We can do so as such:

**GNU Fortran**

```
module load gcc
```

Or

**Intel Fortran**

```
module load intel
```

From the command line, where your code is located, run the command:

**GNU Fortran**
Research Computing University of Colorado Boulder

gfortran parallel_hello_world.f90 -o parallel_hello_world.exe -fopenmp

Or

Intel Fortran

ifort parallel_hello_world.f90 -o parallel_hello_world.exe -qopenmp

This will give us an executable we can run as a job on Summit. Simply run the job specifying slurm to run the executable. Your job script should look something like this:

```bash
#!/bin/bash

#SBATCH --nodes=1
#SBATCH --time=0:01:00
#SBATCH --partition=shas-testing
#SBATCH --ntasks=4
#SBATCH --job-name=Fortran_Hello_World
#SBATCH --output=Fortran_Hello_World.out

./parallel_hello_world.exe
```

Our output file should look like this:

```
Hello from process: 0
```

As you may have noticed, we only get one thread giving us a Hello statement.

How do we parallelize the print statement? We parallelize it with `omp parallel`!

The `!$OMP PARALLEL` and `!$OMP END PARALLEL` directives creates a section of code that is run from all available threads.

```fortran
PROGRAM Parallel_Hello_World
USE OMP_LIB

!$OMP PARALLEL
    PRINT *, "Hello from process: ", OMP_GET_THREAD_NUM()
!$OMP END PARALLEL
END
```

We must do one more thing before achieving parallelization. To set the amount of threads we want OpenMP to run on, we must set an Linux environment variable to be specify how many threads we wish to use. The environment variable: `OMP_NUM_THREADS` will store this information. Changing this variable does not require recompilation of the the program, so this command can be placed in either the command line or on your job script:

```
export OMP_NUM_THREADS=4
```

**Important to note: this environment variable will need to be set every time you exit your shell.** If you would like to make this change permanent you will need to add these lines to your `.bash_profile` file in your home directory:

```
OMP_NUM_THREADS=4;
export OMP_NUM_THREADS
```

Now let’s re-compile the code and run it to see what happens:

2.35. **Using OpenMP with Fortran**
2.35.2 Private vs. Shared Variables

Memory management is a quintessential component of any parallel program that involves data manipulation. In this section, we will learn about the different variable types in OpenMP as well as a simple implementation of these types into the program we made in the previous section.

OpenMP has a variety of tools that can be utilized to properly indicate how the parallel program should handle variables. These tools come in the forms of shared and private variable classifiers.

- Private classifiers create a copy of a variable for each process in the parallel system.
- Shared classifiers hold one instance of a variable for all processes to share.

To indicate private or shared variables, declare the variable before your parallel section and annotate the omp directive as such:

```fortran
!$OMP PARALLEL SHARED(shar_Var1) PRIVATE(priv_Var1, priv_Var2)
```

Variables that are created and assigned inside of a parallel section of code will be inherently be private, and variables created outside of parallel sections will be inherently public.

Example:

Let’s adapt our ‘Hello World’ code to utilize private variables as an example. Starting with the code we left off with, let’s create a variable to store the thread id of each process. We will also change the name of the program as good coding practice.

```fortran
PROGRAM Parallel_Stored_Hello
USE OMP_LIB

INTEGER :: thread_id

!$OMP PARALLEL

    PRINT *, "Hello from process: ", OMP_GET_THREAD_NUM()

!$OMP END PARALLEL

END
```

Now let’s define `thread_id` as a private variable. Because we want each task to have a unique thread id, using the `private(thread_id)` will create a separate instance of `thread_id` for each task.
Lastly, let's assign the thread id to our private variable and print out the variable instead of the `OMP_GET_THREAD_NUM()` function call:

```fortran
PROGRAM Parallel_Stored_Hello
USE OMP_LIB
INTEGER :: thread_id
!$OMP PARALLEL PRIVATE(thread_id)
    thread_id = OMP_GET_THREAD_NUM()
    PRINT *, "Hello from process: ", thread_id
!$OMP END PARALLEL
END
```

Compiling and running our code will result in a similar result to our original hello world:

```
Hello from process: 3
Hello from process: 0
Hello from process: 2
Hello from process: 1
```

### 2.35.3 Barrier and Critical Directives

OpenMP has a variety of tools for managing processes. One of the more prominent forms of control comes with the `barrier`:  

```fortran
!$OMP BARRIER
```

... and the `critical` directives:

```fortran
!$OMP CRITICAL
...
!$OMP END CRITICAL
```

The barrier directive stops all processes for proceeding to the next line of code until all processes have reached the barrier. This allows a programmer to synchronize processes in the parallel program.

A critical directive ensures that a line of code is only run by one process at a time, ensuring thread safety in the body of code.
Example

Let’s implement an OpenMP barrier by making our ‘Hello World’ program print its processes in order. Beginning with the code we created in the previous section, let’s nest our print statement in a loop which will iterate from 0 to the max thread count. We will retrieve the max thread count using the OpenMP function:

```
OMP_GET_MAX_THREADS()
```

Our ‘Hello World’ program will now look like:

```
PROGRAM Parallel_Ordered_Hello
USE OMP_LIB

INTEGER :: thread_id

!$OMP PARALLEL PRIVATE(thread_id)
    thread_id = OMP_GET_THREAD_NUM()
    DO i=0,OMP_GET_MAX_THREADS()
        PRINT *, "Hello from process: ", thread_id
    END DO

!$OMP END PARALLEL
END
```

Now that the loop has been created, let’s create a conditional that will stop a process from printing its thread number until the loop iteration matches its thread number:

```
PROGRAM Parallel_Ordered_Hello
USE OMP_LIB

INTEGER :: thread_id

!$OMP PARALLEL PRIVATE(thread_id)
    thread_id = OMP_GET_THREAD_NUM()

    DO i=0,OMP_GET_MAX_THREADS()
        IF (i == thread_id) THEN
            PRINT *, "Hello from process: ", thread_id
        END IF
    END DO

!$OMP END PARALLEL
END
```

Lastly, to ensure one process doesn’t get ahead of another, we need to add a barrier directive in the code. Let’s implement one in our loop.

```
PROGRAM Parallel_Ordered_Hello
USE OMP_LIB

INTEGER :: thread_id

!$OMP PARALLEL PRIVATE(thread_id)
    thread_id = OMP_GET_THREAD_NUM()
```

(continues on next page)
DO i=0,OMP_GET_MAX_THREADS()
  IF (i == thread_id) THEN
    PRINT *, "Hello from process: ", thread_id
  END IF
  !$OMP BARRIER
END DO
!$OMP END PARALLEL

Compiling and running our code should order our print statements as such:

Hello from process: 0
Hello from process: 1
Hello from process: 2
Hello from process: 3

### 2.35.4 Work Sharing Directive: `omp do`

OpenMP’s power comes from easily splitting a larger task into multiple smaller tasks. Work-sharing directives allow for simple and effective splitting of normally serial tasks into fast parallel sections of code. In this section we will learn how to implement `omp do` directive. The directive `omp do` divides a normally serial for loop into a parallel task. We can implement this directive as such:

```
!$OMP DO
  ...
!$OMP END DO
```

**Example**

Let’s write a program to add all the numbers between 1 and 1000. Begin with a program title and the `omp_lib` header:

```
PROGRAM Parallel_Do
USE OMP_LIB
END
```

Now let’s go ahead and setup variables for our parallel code. Let’s first create variables `partial_Sum` and `total_Sum` to hold each thread’s partial summation and to hold the total sum of all threads respectively.

```
PROGRAM Parallel_Hello_World
USE OMP_LIB

INTEGER :: partial_Sum, total_Sum

END
```

Next let’s begin our parallel section with `!$OMP PARALLEL`. We will also set `partial_Sum` to be a private variable and `total_Sum` to be a shared variable. We shall initialize each variable in the parallel section.
Let’s now set up our work sharing directive. We will use the !$OMP DO to declare the loop to be work sharing, followed by the actual Fortran loop. Because we want to add all number from 1 to 1000, we will initialize out loop at one and end at 1000.

Now we must join our threads. To do this we must use a critical directive to create a thread safe section of code. We do this with the !$OMP CRITICAL directive. Lastly we add partial sum to total sum and print out the result outside the parallel section of code.
This will complete our parallel summation. Compiling and running our code will result in this output:

```
Total Sum: 500500
```

## 2.36 University of Utah videos

These videos were made by the Center for High Performance Computing (CHPC) at the University of Utah. These videos include information that is useful for CU Research Computing users, but some of the information is specific to resources at the University of Utah. If you are a CU Boulder user with questions about these videos, please contact rc-help@colorado.edu.

### 2.36.1 Introduction to Linux

This video provides an introduction to using Linux on a high-performance computing resource.

### 2.36.2 Introduction to CHPC Modules

This video demonstrates the use of modules on a CHPC resource. CHPC uses the Lmod system, which CU Research Computing also uses.

### 2.36.3 Slurm Basics

This video provides a description of the clusters, partitions and types of accounts available at CHPC. It also demonstrates use of the “squeue” and “srun” commands, as well as the use of the “sbatch” command to run a batch job.

### 2.36.4 Slurm Batch Scripting

This video provides information about the use of modules and the use of “sbatch” commands in Slurm Batch Scripting.

### 2.36.5 Slurm Interactive Jobs

This video demonstrates how to run interactive jobs and provides advice on how to determine whether an interactive job best meets your needs.
2.37 Blanca: Memorandum of Understanding

The “Blanca” compute cluster is a service provided by Research Computing at University of Colorado Boulder. Blanca aggregates compute nodes contributed by individual researchers into a heterogeneous compute cluster managed by Research Computing.

- Contributions remain the property of the contributors, and may be fitted with the contributor’s asset tag.
- Contributions may be removed from the Blanca environment by the contributor should the Blanca environment not sufficiently meet the contributor’s needs. Removal also removes access to the benefits associated with the Blanca environment.
- Administrative (e.g., “root”) access to the Blanca compute environment, including as deployed on contributed resources, is reserved for Research Computing operational staff and its delegates.
- Contributions are installed in a CU data center at the discretion of Research Computing.
- Contributed compute nodes are provisioned with a Linux-based operating system and environment at the discretion of Research Computing.
- Contributed compute nodes are provisioned with access to Research Computing Core services and the PetaLibrary to the extent such access is technically feasible.
- Research Computing reserves the right to determine what contributions are accepted into Blanca, with recommended example configurations published on the Research Computing website.
- Contributed Blanca compute nodes are configured as nodes in a Slurm batch compute cluster.
- All Blanca compute nodes accept preemptable jobs from all Blanca contributors (and their appointees). These jobs may be preempted when contributors (and their appointees) run jobs to their contributed nodes.
- Contributors agree to give Research Computing “right of refusal” before surplussing or otherwise disposing of any Blanca contributions. Such contributions are transferred to Research Computing’s custodianship at no cost to Research Computing, and continue to be operated as part of the aggregate Blanca environment for use by all active Blanca contributors (and their appointees). This agreement does not restrict the contributor’s right to remove their Blanca contributions for re-use outside of Blanca.
- Research Computing reserves the right to update this MOU over time. Any change to the MOU that is found unacceptable by a contributor will not restrict the contributor’s right to remove their Blanca contributions for re-use outside of Blanca.

2.38 Facilities, equipment, and other resources

2.38.1 Expertise

Research Computing at CU Boulder consists of a small group of computational scientists, high-performance computing specialists, and system administrators with the mission to provide leadership in developing, deploying, and operating an integrated cyberinfrastructure. This cyberinfrastructure consists of high-performance computing, storage, and high speed networking that supports research, collaboration and discovery. Research Computing contributes to the educational mission of the university by providing training workshops and consultation services for cyberinfrastructure related topics.

2.38.2 Compute

- Research Computing operates the joint RMACC Summit supercomputer, funded by NSF under Grant No. AC-1532236. The system has peak performance of over 400 TFLOPS. The 472 general compute nodes each have
24 cores aboard Intel Haswell CPUs, 128 GB of RAM and a local SSD. Additionally, the system has 11 GPU nodes with two NVIDIA K80 GPUs each, 5 high-memory nodes with ~2TiB of main memory, and 20 Xeon Phi (“Knight’s Landing”) nodes each with 68 real cores supporting 272 threads. All nodes are connected through a high-performance network based on Intel Omni-Path with a bandwidth of 100 Gb/s and a latency of 0.4 microseconds. A 1.2 PB high-performance IBM GPFS file system is provided. This system is available to CU-Boulder and Colorado State University researchers and collaborators, and 10% of cycles are provided to members of the Rocky Mountain Advanced Computing Consortium.

- The Research Computing Condo Computing service offers researchers the opportunity to purchase and own compute nodes that are operated as part of a cluster, named “Blanca.” The aggregate cluster is made available to all condo partners while maintaining priority for the owner of each node.
- Research Computing provides a 3d-accelerated virtual desktop environment for real-time visualization and rendering using EnginFrame and DCV. This environment is powered by two visualization nodes, each equipped with 2x AMD EPYC 7402 24-core processors, 256GiB memory, and 2x Nvidia Quadro RTX 8000 GPU accelerators. Each accelerator is itself equipped with 48 GiB of high-speed GDDR6 memory.

2.38.3 Networking

The current CU Boulder network is a 40 Gbps fiber core with Cat 5 or higher wiring throughout campus. Research Computing has created an 80 Gbps Science-DMZ to connect the RMACC Summit supercomputer to storage and to bring individual dedicated 10 Gbps circuits to various locations as needed. CU Boulder participates in I2 (the Internet 2 higher education, government, and vendor research computing consortium) and is an active member of the Front-Range gigapop and other networks. Research Computing has begun to provide campus researchers with a leading-edge network that meets their needs and facilitates collaboration, high performance data exchange, access to co-location facilities, remote mounts to storage, and real-time communications.

2.38.4 File Transfer

For moving large volumes of data Research Computing has several nodes dedicated to GridFTP file transfer. The CU Office of Information Technology also offers a file transfer service with a web interface, which provides an ideal way to transfer large files to collaborators. Files are uploaded to a server and a link to download the file is emailed to an on- or off-campus user.

2.38.5 Storage

Each Research Computing user has a 2 GB home directory and a 250 GB projects directory, each of which are backed up regularly. Each RMACC Summit user has a 10 TB scratch directory.

2.38.6 PetaLibrary Storage Services

The PetaLibrary is a CU Research Computing service supporting the storage, archival, and sharing of research data. It is available at a subsidized cost to any researcher affiliated with the University of Colorado Boulder. The two main categories of service offered to customers of the PetaLibrary are Active storage for data requiring frequent access, and Archive storage for data that is accessed infrequently. Active data is stored on spinning disk and is accessible to researchers on compute resources managed by Research Computing. Archive storage consists of a two-level hierarchical storage management (HSM) solution, with disk storage for data that is more likely to be accessed and tape storage for data that is less likely to be accessed frequently. The cost for CU researchers is $45/TB/year for Active and $20/TB/year for Archive.
Through a collaboration with the CU Libraries, the PetaLibrary can also host the publication and long-term archival of large datasets. The datasets are assigned unique digital object identifiers (DOIs) that are searchable and accessible via the “CU Scholar” institutional repository.

### 2.38.7 JupyterHub

JupyterHub is a multi-user server for Jupyter notebooks. It provides a web service enabling users to create and share documents that contain live code, equations, visualizations and explanatory text. The CU Research Computing JupyterHub deploys into the RMACC Summit supercomputer and Blanca condo cluster and includes support for parallel computation.

### 2.38.8 EnginFrame

NICE EnginFrame provides a 3D-accelerated remote desktop environment on Nvidia GPU-equipped compute nodes. Coupled with the proprietary Desktop Cloud Visualization (DCV) VNC server, the Research Computing EnginFrame supports the use of common visualization applications in a typical desktop environment using only a modern web browser.

### 2.38.9 Center for Research Data and Digital Scholarship (CRDDS)

The Center for Research Data & Digital Scholarship (CRDDS) is a collaboration between Research Computing and University Libraries, offering a full range of data services for both university and community members. The aim of CRDDS is to provide support to community members on areas related to data intensive research. CRDDS fulfills this mission by providing education and support on such issues as data discovery, reuse, access, publication, storage, visualization, curation, cleaning, and preservation, as well as digital scholarship. CRDDS is located in Norlin Library on Main Campus at CU Boulder.

CRDDS offers many opportunities to students working with data. The expert staff work hand-in-hand with researchers via weekly office hours, one-on-one consultations, and group trainings in programming, data visualization and more. CRDDS serves as a resource for data management, manipulation and publication for trainees working through undergraduate and graduate coursework.

Examples of workshops/trainings CRDDS has offered include:

- High performance computing
- Programming in R
- Programming in Python
- Containerization
- Data mining

### 2.39 CHANGelog

#### 2.39.1 Wednesday, 17 October 2018

- Configured Slurm to reject invalid jobs during queueing, rather than accept and pend forever
2.39.2 Wednesday, 16 May 2018

- Shut down legacy login nodes

2.39.3 Wednesday, 9 May 2018

- Moved login.rc.colorado.edu to a new set of login nodes

2.39.4 Tuesday, 27 February 2018

- Reduced memory limits on Summit

2.39.5 Tuesday, 20 February 2018

- Installed version control clients on Summit in response to a request for Mercurial

2.39.6 Monday, 19 February 2018

- Reduced default per-core memory requests on Summit

2.39.7 25 September 2017

- /work/ mounts are now supported for JupyterHub Virtual Notebooks.
- JupyterHub Notebooks can now be spawned on Summit, and the Crestone spawn option has been deprecated.
- JupyterHub ipyparallel profiles have been updated for Summit. To remove stale Janus/Crestone profiles, please follow the procedure described in the JupyterHub User Guide
- JupyterHub Cluster notebooks now use lmod instead of virtualenv for software dependencies.