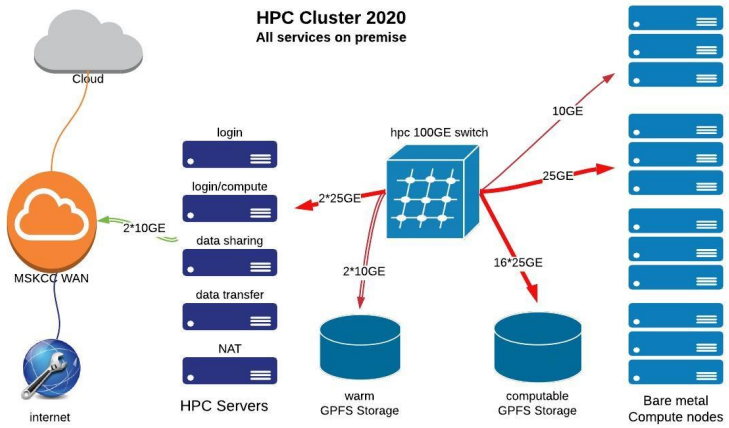
Welcome FAQ

Answers to frequently asked questions for new users

What is High Performance Computing (HPC)?

A HPC cluster is a high-performance, parallel computing infrastructure which consists of three key components: compute, network, and storage. On a cluster one can take advantage of multiple cores by running several instances of a program at once or using a parallelized version of the program. The lilac cluster is appropriate for all types of workloads including Genomic Analysis, Artificial Intelligence and Machine Learning. It is available to all users at MSKCC. The juno cluster is dedicated to processing Genomic pipelines and analysis and access is limited. You can find more information about the two clusters at http://hpc.mskcc.org/compute-accounts/

All access to the clusters is via SSH to the login nodes. From the login node you can view files and dispatch jobs to compute nodes on the private network. IBM Spectrum Scale LSF is the job scheduler we use to manage these jobs. All nodes on a cluster mount a shared GPFS filesystem. Each node also has a local 1TB /scratch drive for temporary data. We also provide special data transfer servers with optimized network connections for moving large data sets to and from the clusters.



How do job schedulers work?

On an HPC cluster, the scheduler manages which jobs run where and when. On our clusters, you control your jobs using a job scheduling system called IBM Spectrum Scale LSF that allocates and manages compute resources for you. You can submit your jobs in one of two ways. For testing and small jobs you may want to run a job interactively. This way you can directly interact with the compute node(s) in real time. The other way, which is the preferred way for multiple jobs or long-running jobs, involves writing your job commands in a script and submitting that to the job scheduler. We will be updating or LDF documentation soon.

Where can I find a basic linux tutorial?

The course materials for the Tri-I Bioinformatics and Computational Workshops UNIX course are available at

https://chagall.med.cornell.edu/UNIXcourse/

There are also many linux tutorials and cheat sheets available online such as:

http://www.ee.surrey.ac.uk/Teaching/Unix/index.html

https://linuxjourney.com/

http://mywiki.wooledge.org/BashGuide

What do I do once my account is created?

All access to the clusters is via SSH keys. We recommend that you use authentication forwarding. If you have trouble connecting please read the SSH page at

http://mskcchpc.org/display/CLUS/Secure+Shell+SSH

Storage and quotas

Your home directory has a 100G quota. High performance GFPS Please use your lab's data directory /data/labname for datasets and analysis. Each compute node has >1T of local scratch disk. We also have /warm storage which is lower performance and not computable. This storage is not backed up. Information about our storage offerings is at http://hpc.mskcc.org/data-storage/

How do I transfer data to and from the cluster?

Each cluster has a special data transfer (xfer) server with a faster network connection optimized for transferring data called lilac-xfer01-mkscc.org and juno-xfer01.mskcc.org. To use them just SSH to them and start your data transfers from there.

Data can be transferred from other linux or MacOS using rsync over SSH.

Data can be transferred from Windows or SAMBA shares using smbclient. Please make sure to use the fully qualified domain name of the samba server in the smbclient command. The correct syntax for smbclient is:

> *smbclient //servername.mskcc.org/sharename -W mskcc -U username*

This will prompt for your MSKCC password and connect. Here is an example:

```
lilac-xfer01:~> smbclient //skimcs.mskcc.org/HPC/ -W mskcc -U edingtoj
Enter MSKCC\edingtoj's password:
Try "help" to get a list of possible commands.
smb: \>
```

What software is available?

Available Software

Installing Software

How do I be a good citizen on the cluster?

Don't run compute jobs on the login nodes.

Do not use /tmp as a scratch space and clean up any scratch data you have generated when your job finishes.

Use the data transfer servers for transferring data on and off of the clusters.