# Obtaining your data from PNCC: important facts

PNCC processes data at and distributes it from the EMSL (https://www.emsl.pnl.gov) user facility at Pacific Northwest National Laboratory in Richland, Washington. The EMSL computing facility features parallel HPC systems, petabytes of data storage, and dedicated high-speed data transfer capability.

PNCC data sets are very large, with some of them comprising multiple terabytes. Distance matters a great deal when transferring such large quantities of data across geographical distances on a wide area network. The speed of network transfers has to be taken into account, as they can cause transfer times to vary by factors of 1000!

To work around some of the inherent problems of transferring large quantities of data across long distances, PNCC recommends that you:

1. Ensure that machines at your institution are well-configured for long-distance data transfer
2. Choose a file transfer tool that can perform well
3. Test your transfers end-to-end from EMSL to troubleshoot any problems prior to taking data at PNCC

## Configure and check your machines

In order to get good data transfer rates from PNCC, your receiving machine(s) must have:

1. A reliable network connection that does not experience errors or dropped packets
2. A fast wired network interface (10 gigabit/second is recommended). Do not use wireless for these data transfers!
3. Tuning of the operating system (Linux, Windows, or MacOS) that dedicates extra system memory to data transfer
4. Fast disk or SSD storage (capable of at least 100 megabyte/second) that is large enough to hold your data set(s)

The farther you are from the Pacific Northwest United States, the more important #1, #2, and #3 are. See http://fasterdata.es.net/host-tuning for instructions on #3.

*Background info: The speed of light in a cable has significant effects at long distances, and essentially many megabytes of data can be “in the wire” between computers that are far apart. The computers at either end must be tuned to dedicate at least a “cable full” of their memory to each active data transfer, or they will drop data that will then have to be re-sent. This problem is noticeable at distances of about 500 miles, and can slow down transfers by a factor of 1000 on cross-country connections. Almost no computer is tuned for this “out of the box,” but fortunately tuning guidelines are available at https://fasterdata.es.net/host-tuning.*

##

## Choose your transfer tool(s)

The three recommended ways to retrieve PNCC data from EMSL are:

* Globus
* aria2
* MyEMSL data portal pages

They are listed in order of performance from fastest to slowest. As you might expect, the higher performance options require more setup. None of these tools will perform well if you are distant from the Pacific Northwest United States and have not addressed issues described in the prior section of this document.

If you received an email message entitled “MyEMSL Notification - Data Uploaded for Project XXXXX,” that message contains direct links to your data in the EMSL data portal and a metalink that may be used to download your data with aria2.

### Globus

#### What is Globus? Why are we using it?

Globus is widely used in the scientific research community for sharing and transferring large quantities of data. It is designed to provide secure, reliable, high performance data access across multiple sites in a single interface. Globus is used day in and day out to transfer PNCC data from OHSU to EMSL, and moves multiple terabytes of data per day.

Introductory information about Globus is available at the Globus web site: https://www.globus.org/data-transfer.

#### How to obtain a globus ID

In order to use Globus, you will need a Globus ID. The Globus ID identifies you to Globus, and allows you to access Globus endpoints (see below), including the EMSL Data Transfer Node (emsl#dtn) that hosts PNCC user data. See https://docs.globus.org/how-to/get-started/

#### Minimum technical requirements for running globus

The normal Globus mode of operations is to use *endpoints* at either end of a data transfer. An endpoint is usually a dedicated data storage system with high performance network interfaces and disk storage, tuned for long distance data transfers. It is also possible to set up a personal computer as a Globus endpoint by installing Globus Connect Personal software on it, but this may not perform as well as a dedicated, tuned endpoint. If you have a local server-class endpoint at your institution, it will likely be better to transfer data to that endpoint, then make a local copy down to your own machine.

### Expected Globus transfer times (caveats with distance & file size, etc.)

The emsl#msc Globus endpoint available to OHSU users at PNNL has been observed sending data via Globus to Arizona as fast as 275 megabytes/second. At that rate, one terabyte can be transferred in about an hour. A poor cross-country connection has been seen to perform as badly as 250 kilobytes/second, in which case that same terabyte would take a month and a half(!) to transfer.

### aria2

aria2 is a multi-protocol data download tool that is available at https://aria2.github.io. Aria2 can use the metalink that is included in the It has several useful features, including the ability to transfer multiple data streams simultaneously to improve performance. It can use the metalink to your data that is in the “MyEMSL Notification - Data Uploaded for Project XXXXX” email message. It can also resume an interrupted data transfer without having to restart.

### MyEMSL data portal

The simplest but least performant way to download data is from the EMSL User Portal at https://eus.emsl.pnnl.gov/Portal. Log in to the portal with your EMSL-assigned email address and password. Click on the “Get Data” tab to see a list of instruments that have data available to you. You can navigate to the data set of interest, select the files and folders that you want to download, and then click on the “Queue Selected Files for Download.” The portal will then prepare a downloadable set of files for you, which may take several minutes. You may leave the page and return to it later; the portal will continue preparing your data.

Notes:

* The “Not Released” banner across the corner of the page indicates that the data set has not yet been released to the public. This means it is available only to approved members of the project and authorized EMSL staff
* If you do not see any instruments or data sets on the portal pages, make sure you go to https://eus.emsl.pnnl.gov/Portal and log in, then reload the pages
* Once a downloadable file set (a “cart”) has been prepared you may download it with your web browser, or use a tool like wget, curl, or aria2.

## Test your transfers

It is an excellent idea to try transferring data from EMSL before you begin to copy your PNCC data. When a new Globus Shared Endpoint is created for your project, a large file named 'testdata' containing random data will be placed in it. You should download that file to test transfer rates to your institution.

If you want to test with known high-performance servers, or compare the speed of distant vs. local transfers, The Energy Sciences Network (ESNet) maintains a set of machines for such testing. They use the GridFTP protocol (which Globus uses) for maximum performance. See https://fasterdata.es.net/performance-testing/DTNs.