



# Data transfer with GridFTP

H. Heller, J. Laitinen, F. Scheiner, G. Pringle, D. Girou(eds.)

(c) 2004 - 2010 DEISA



DEISA is funded by the European Commission in FP7 under grant agreement RI-222919



## Table of contents

1 Data transfer.....	1
2 GridFTP and file transfer.....	3
2.1 How can GridFTP be used for file transfer?.....	3
2.2 globus-url-copy syntax.....	3
2.3 Command line options for globus-url-copy.....	4
3 Access from a DEISA machine .....	7
4 Access from a machine outside of the DEISA environment .....	9
5 Data transfer with globus-url-copy.....	11
5.1 Copying data between a local workstation and the DEISA infrastructure .....	11
5.2 Copying data inside the DEISA infrastructure .....	11
5.3 Copying files and directories with globus-url-copy.....	12



# 1 Data transfer

This document describes the deployment and usage of the data transfer facility *GridFTP*, a component of the *Globus middleware suite*[1]. It allows to transfer data efficiently between DEISA sites as well as between a non-DEISA site and a DEISA site.

Large scientific applications typically running on the DEISA HPC infrastructure both analyse and generate huge amounts of data. These data sets cannot be expected to be permanently stored within the DEISA multi-site shared filesystem. In addition, an alternative site-to-site file transfer mechanism is required in case that one partner site can not be connected to the DEISA shared filesystem for technical reasons. To enable users to stage data in and out of DEISA HPC systems without common shared filesystems, DEISA sites provides an efficient file transfer service based on GridFTP.

GridFTP is a protocol for high-performance, secure and reliable data transfer for high-bandwidth WANs such as the one employed as the DEISA internal network, and DEISA users need to move data between the DEISA infrastructure and their local file systems. Thus, for this purpose, some sites also provide GridFTP servers that can be reached via the public internet. These so-called door nodes are listed in the Table 4 in section 4 Access from a machine outside of the DEISA environment[2]

DEISA uses the GridFTP server implementation that comes with the *Globus toolkit 4*[3].

- 
1. <http://www.globus.org/>
  2. <http://www.deisa.eu/usersupport/user-documentation/data-transfer-with-GridFTP/access-from-a-machine-outside-of-the-deisa-environment>
  3. <http://www.globus.org/toolkit>



## 2 GridFTP and file transfer

### 2.1 How can GridFTP be used for file transfer?

In order to use GridFTP for file transfer, one needs a GridFTP client program that provides the interface between the user and a remote GridFTP server. There are several clients available for GridFTP, one of which is *globus-url-copy*, a command line tool which can transfer files using the GridFTP protocol as well as other protocols such as http and ftp. *globus-url-copy* is distributed with the Globus Toolkit and usually available on machines that have the Globus Toolkit installed.

### 2.2 globus-url-copy syntax

The basic syntax of the `globus-url-copy` command is:

```
globus-url-copy [options] sourceURL destinationURL
```

where the arguments are described in the following table.

Aurgument	Description
[options]	The optional command line switches as described in 2.3 Command line options for <code>globus-url-copy</code> [1]
sourceURL	The URL of the file(s) to be copied. If it is a directory, it must end with a slash (/), and all files within that directory will be copied.
destURL	The URL to which to copy the file(s). To copy several files to one destination URL, <code>destURL</code> must be a directory and be terminated with a slash (/)

Table 1: Arguments of *globus-url-copy*

`globus-url-copy` supports multiple protocols, so the format of the source and destination URLs can be either

```
file://path
```

when you refer to a *local* file or directory or

```
protocol://host[:port]/path
```

1. <http://www.deisa.eu/usersupport/user-documentation/data-transfer-with-GridFTP/gridftp-and-file-transfer#doc-2>

## GridFTP and file transfer

when you refer to a *remote* file or directory. While globus-url-copy is supporting other protocols such as http, https and ftp as well, in the DEISA infrastructure it is only possible to use the GridFTP protocol: gsiftp://

The port number can be omitted if the GridFTP server's listens on the default port 2811.

The path

- must be an absolute path for file://
- for gsiftp:// the path can be relative to the user's home directory, in which case it must start with ~
- must be terminated with a slash (/), if it refers to a directory.

To transfer data with globus-url-copy using the gsiftp:// protocol, the user must have valid credentials, as will be described below. Normally you will use file:// for addressing a local and gsiftp:// for addressing a remote file or directory. However, note that the GridFTP protocol supports so-called third party-transfers where you can transfer data between two remote servers. In this case you have to use gsiftp:// both for the source and the destination URL.

### 2.3 Command line options for globus-url-copy

We present the most important command line options. For a much more comprehensive description of available options, see the documentation on the Globus website <http://www.globus.org/>.

When you use the optional parameters given in the table below, you will get additional *information*:

Option	Description
-help	Prints usage information for the globus-url-copy program.
-version	Prints the version of the globus-url-copy program.
-vb	During the transfer, displays: (1) number of bytes transferred (2) performance since the last update (every 5 seconds) (3) average performance for the whole transfer

Table 2: *Optional parameters*

## GridFTP and file transfer

The following table lists parameters which you can set to *optimize the performance* of your data transfer:

Option	Description
-tcp-bs <size>	Specifies the size (in bytes) of the TCP buffer to be used by the underlying GridFTP data channels.
-p <number of paral- lel streams>	Specifies the number of parallel streams to be used in the GridFTP transfer.
-stripe	Use this parameter to initiate a “striped” GridFTP transfer that uses more than one node at the source and destination. As multiple nodes contribute to the transfer, each using its own network interface, a larger amount of the network bandwidth can be consumed than with a single system. Thus, at least for “big” (> 100 MB) files, striping can considerably improve performance

Table 3: *Parameters to optimize performance*

How to choose values for these parameters?

Concerning the first two parameters – TCP buffer size and parallelism –, generally the optimal values depend on factors such as the latency between the source and destination sites, the available bandwidth, network traffic etc. Some of the parameters are fixed (for instance you can measure the latency yourself using ping), others such as the limiting bandwidth are only known to the network administrators at the various DEISA sites. However, as a rule of thumb we recommend to use the following values:

- four parallel streams should be enough.
- for the typical latencies that occur in the DEISA network use 4MB for the TCP buffer size.

If you plan a lot of transfers of big files, it might be advisable to vary the value to see how it influences performance. For instance, a higher TCP buffer size than the recommended one could give you more performance between sites with a larger latency, however, more memory is used, which may affect transfer.

With regard to striping, currently the following DEISA sites are supporting multi-striping: CINECA, IDRIS, RZG and SARA.

In order to make GridFTP usage easier for the DEISA users, we deployed on all DEISA sites a wrapper script, called `gscp`, around the `globus-url-copy` command. Reading static parameters such as server names, port numbers, optimal TCP buffer size, etc. from a configuration file, this tool will pre-set most of the values for you, but will give you the freedom to overwrite them. It is possible to use site names used in `deisa_service` script. For example to copy a file (`source.txt`) from the current directory to the home directory at SARA's GridFTP server

```
gscp source.txt sara:target.txt
```

To copy a file (`source.txt`) from SARA to current directory the command would be:

```
gscp sara:source.txt target.txt
```

## GridFTP and file transfer

For more options please see

```
gscp -help
```

### 3 Access from a DEISA machine

First, make sure the deisa and globus modules have been loaded via the 'module load deisa' and 'module load globus' commands. This sets the \$DEISA\_HOME and \$DEISA\_DATA environment variables and make the Globus client commands like globus-url-copy available.

```
module load deisa globus
```

```
echo $DEISA_HOME  
/deisa/lrz/home/lrz00001/lrz015ab
```

```
echo $DEISA_DATA  
/deisa/lrz/data/lrz00001/lrz015ab
```

Then copy the file myfile to \$DEISA\_HOME and \$DEISA\_DATA directories mounted on the GridFTP server at RZG:

```
globus-url-copy file://`pwd`/myfile gsiftp://`deisa_service` -i -f rzg`/$DEISA_HOME/myfile
```

```
globus-url-copy file://`pwd`/myfile gsiftp://`deisa_service` -i -f rzg`/$DEISA_DATA/myfile
```



## 4 Access from a machine outside of the DEISA environment

To transfer files from machines which are not on the DEISA network, i.e. your workstation, you can use the machines mentioned in the following table. The address `gridftp.deisa.eu` refers to CINECA or LRZ. If a server is found to be unavailable, then try one of the alternative addresses. Both servers are using the default GridFTP port. Please note that to use the LRZ site, you first have to register your IP address. This can be done by submitting a request to the DEISA Helpdesk service[1].

Site	Address	Port
LRZ or CINECA	<code>gridftp.deisa.eu</code>	2811
CINECA	<code>gridftp1.deisa.eu</code>	2811
LRZ	<code>gridftp2.deisa.eu</code>	2811

*Table 4: GridFTP nodes available for Internet and DEISA network*

---

1. <https://tts.deisa.eu/UserSupport/>



## 5 Data transfer with globus-url-copy

In this subsection we describe how you can use globus-url-copy to

- copy data between your local workstation and the DEISA infrastructure
- copy data from one DEISA platform to another DEISA platform

After that we give some concrete examples that show how to use the globus-url-copy command.

### 5.1 Copying data between a local workstation and the DEISA infrastructure

To transfer files from your local workstation to DEISA, you have to have Globus installation available and to use one of the DEISA GridFTP door nodes listed in the Table 4[1].

On the GridFTP door node server

- Globus toolkit has been installed,
- connections to the DEISA network and thus to the GridFTP servers at every DEISA site
- the machine can be accessed from the public internet,
- the machine has mounted all the DEISA GPFS file systems.

Three GridFTP servers can be accessed currently from the public internet:

Use your DEISA\_HOME or DEISA\_DATA directories for uploading and downloading data, as these are transparently available at nearly all the DEISA sites. On most of the DEISA HPC system data can be managed using the Global Parallel File System GPFS. GridFTP can be used in case that data has to be moved into a non-GPFS storage system.

### 5.2 Copying data inside the DEISA infrastructure

This requires access to DEISA systems. The user can access these machines in two ways:

- From a machine that has GSISsh-Term installed, as described in the Interactive access User Guide section 4.5 Using GSISsh-Term.
- From a machine with a Globus installation that provides the gsissh client.

---

1. <http://www.deisa.eu/usersupport/user-documentation/data-transfer-with-GridFTP/access-from-a-machine-outside-of-the-deisa-environment>

## Data transfer with globus-url-copy

In the second case, after generating a proxy credential with `grid-proxy-init`, the user will connect to one of the "door nodes" by using the `gsissh` command:

```
gsissh -p 2222 a01.hlrb2.lrz-muenchen.de
```

After successful login to the DEISA machine with either `GSISsh-Term` or `gsissh`, the user is mapped to a DEISA user account, whose name depends on the user's Home Site.

For example, a user whose Home site is LRZ will be mapped to an account whose name starts with "lrz":

```
whoami
lrz015ab
```

In the next subsections we give some examples that show how to use `globus-url-copy` for different purposes.

### 5.3 Copying files and directories with globus-url-copy

We show how to copy files and entire directories.

Before using `globus-url-copy` you have to generate a proxy credential based on your credentials (that means your permanent public/private key pair) with `grid-proxy-init`. What you have to do with your credentials and the trusted CA certificates has also been described in DEISA Certificates FAQ[2].

```
grid-proxy-init
Your identity:
    /C=DE/O=GridGermany/OU=Leibniz-Rechenzentrum/OU=HLS/CN=Gabriel Mateescu
Enter GRID pass phrase for this identity:
Creating proxy ..... Done
Your proxy is valid until: Fri Mar 10 05:09:41 2008
```

#### Copy a file

Assume that the user has stored a large file "myfile" in the current working directory of his local workstation. He wants to use it as input file for a calculation on some DEISA production system. To upload it to DEISA using GridFTP, he has to use either CINECA's or LRZ's

```
globus-url-copy file://`pwd`/myfile gsiftp://gridftp.deisa.eu/<DEISA_HOME>/myfile
```

For <DEISA\_HOME> he has to enter the absolute path to his DEISA\_HOME directory. DEISA\_HOME and DEISA\_DATA follow a specific pattern (please refer also to section 4.1 of the DEISA Primer[3]):

```
DEISA_HOME=/deisa/SITE_NAME/home/DEISA_GROUP_NAME/DEISA_USER_NAME
DEISA_DATA=/deisa/SITE_NAME/data/DEISA_GROUP_NAME/DEISA_USER_NAME
```

For example, the values for the DEISA user `lrz015ab` with home site LRZ are:

- 
2. <http://www.deisa.eu/usersupport/user-documentation/faq/CertificatesFAQ>
  3. <http://www.deisa.eu/usersupport/primer/file-systems-and-data-management#doc-0>

## Data transfer with globus-url-copy

```
id
uid=1054392(lrz015ab)
gid=1000001(lrz00001)
groups=1000001(lrz00001)
```

```
module load deisa
```

```
echo $DEISA_HOME /deisa/lrz/home/lrz00001/lrz015ab
```

```
echo $DEISA_DATA /deisa/lrz/data/lrz00001/lrz015a
```

Thus, the user can construct the values of DEISA\_HOME and DEISA\_DATA, once he knows the DEISA user ID and group. However, this requires the user to find out the DEISA user and group IDs and store these values on the client and reuse them each time she want to access DEISA\_HOME and DEISA\_DATA using the pattern shown above.

### Copy a directory

We will copy the subdirectory "mydirectory" of the current directory to the user's remote DEISA\_HOME directory:

```
globus-url-copy -cd -r file://`pwd`/mydirectory/
gsiftp://gridftp.deisa.eu/<DEISA_HOME>/mydirectory/
```

where the -cd option stands for "create directory" and its purpose is to create the directory "mydirectory" on the GridFTP server as a subdirectory of the remote DEISA\_HOME directory. To include subdirectories is used recursive copy option -r. Note that we terminate the URLs with a / to indicate that we refer to a directory. As 2811 is the GridFTP default port we can omit it here.