Archiving data on an external drive

pst-laurent@vims.edu

May 26, 2019

1 Introduction

This is a tutorial on how to archive data from VIMS onto an external drive. The external drive serves as a long-term backup of your data. The lifetime of the drive will depend on how often you use it and how well you treat it (shocks or extreme cold/heat/humidity will shorten the lifetime).

The tutorial assumes that the drive's filesystem is NTFS. NTFS is the factory default for the 'Seagate PC Expansion drive'. Windows computers can read and modify the files on a NTFS drive. Linux computers can read and modify the files on a NTFS drive. Apple computers can read the files on a NTFS drive but cannot modify them (nor add new files). This should not be a problem since the drive's sole purpose is that of a backup, i.e. we do not intend to modify the files once they are on the drive. See Appendix A for more information on filesystems and the transfer speeds that can be expected with each of them.

There are two difficulties when archiving large datasets. First, we need a mechanism that monitors the file transfer and verifies that the copied files are identical to the original files. We will rely on Globus online for this. The second difficulty is the time required to copy large datasets onto a drive (sometimes it takes multiple days). We thus need a computer that is constantly connected to the VIMS servers. We will rely on the desktop station '**Vlab6**' of the VizLab for this purpose. **Vlab6 is the last computer on the right and it is the closest to the projector screen**.

2 Creating a Globus account

You will need a 'Globus' account before attempting a data archive. Skip this section if you already have a Globus account.

- Log onto 'Vlab6' using your VIMS username and password.
- Click 'Applications' in the upper-left corner of the screen.
- Select 'Internet' and 'FireFox Web Browser'.
- Go to `https://www.globusid.org/create`
- Follow the instructions on how to create your account.

3 Installing the Globus client on Vlab6

You will need to install 'Globus' on Vlab6 before attempting a data archive. Skip this section if you have already installed the globus client on Vlab6.

- Log onto 'Vlab6' using your VIMS username and password.
- Click 'Applications' in the upper-left corner of the screen.
- Select 'Internet' and 'FireFox Web Browser'.
- Go to `https://docs.globus.org/how-to/globus-connect-personal-linux/`
- In the section 'Installation', select 'Click here to create a Globus Connect Personal endpoint'.
- Click on 'Globus ID to sign in' and enter your Globus username and password.
- The next step is to create a Globus 'endpoint' on Vlab6. We will call it 'my_globus_endpoint_on_vlab6'. After entering this name, click on 'Generate Setup Key'.
- Click on 'copy' to save the Setup Key inside the clipboard. Then, in the upper-left corner of the screen, click on 'Applications', 'Accessories', and 'gedit'. Put the mouse cursor over the newly opened window and then click the middle button (wheel) of the mouse to paste the setup key into the gedit window.
- For safety, write down the setup key on a piece of paper too. Make sure to distinguish between 'zero' and the letter 'o', between number one and the letter 'l', etc.
- Back in the browser, click on 'for Linux' to download the Globus client onto the computer ('Save File', 'Ok').
- FireFox has now saved the Globus client somewhere. We will find it using a terminal window. Click 'Applications' in the upper-left corner of the screen. Select 'Utilities' and then 'Terminal'. You should now see a window with a prompt like `[pierre@vlab6 ~]$`
- Move the client to your home directory by typing
  `mv Downloads/globusconnectpersonal-latest.tgz .`
  (don't forget the period at the end. Then, press 'Enter'.)
- Decompress the client by typing
  `tar xzf globusconnectpersonal-latest.tgz`
- You should now have a new directory named `globusconnectpersonal-2.3.4`. To verify, type `ls` in the terminal window. The command `ls` lists the content of your home.
- Go inside the new directory by typing
  `cd ~/globusconnectpersonal-2.3.4`
- Complete the installation by typing
  `./globusconnectpersonal -setup that_long_setup_key_from_before`
  Here, `that_long_setup_key_from_before` should be that long setup key that you pasted into the gedit window. Remember that in Linux/Unix you can copy/paste text by highlighting it with the mouse ('copy') and then by clicking the middle mouse button ('paste'). Alternatively, you can just type the key manually.
- If you did it right, it will say 'Done!'.

4 Connecting the external drive to Vlab6

The next step is to connect the drive to the computer.

- First, connect the drive to a power socket on the wall. Make sure it's well plugged in.
- Then, connect the drive to one of the USB ports at the back of Vlab6. Make sure it's well plugged in on both ends.
- Linux should have automatically mounted the drive at this point. Let's find out where exactly the drive was mounted using a terminal window. Click 'Applications' in the upper-left corner of the screen, select 'Utilities' and then 'Terminal'. You should see a prompt like `[pierre@vlab6 ~]$`
- Type `ls /run/media/your_username/` and then Enter.
  You should see a directory named `Seagate Expansion Drive`
- The last step is to create a symbolic link inside your home directory and have it pointing toward our drive (Note: You only need to create this link once in your life; Skip this step if the link already exists). Type:
  `ln -s /run/media/your_username/Seagate\ Expansion\ Drive/ ~/my_link_to_the_drive`

5 Connecting the Globus client to the VIMS servers

Now that the drive is connected, we will initiate a Globus connection between Vlab6 and the VIMS servers.

- If you don't already have a terminal window, click 'Applications' in the upper-left corner of the screen, select 'Utilities' and then 'Terminal'. You should see a prompt like `[pierre@vlab6 ~]$`.
- Go to the directory containing the Globus client by typing
  `cd ~/globusconnectpersonal-2.3.4`
- Start the client by typing
  `./globusconnect &`
- In the new window that just popped up, click on 'File' and then on 'Preferences'.
- Click on '+' and then enter the path:
  `/local/home/your_username_on_vlab6/my_link_to_the_drive/`
- Check the box 'Write' and then click on 'Save'.
- Click 'Connect'.
- In the upper-left corner of the screen, click on 'Applications'. Select 'Internet' and 'FireFox Web Browser'.
- Go to: `https://www.globus.org/app/transfer`
- Click on 'Globus ID to sign in', then enter your Globus username and password.

- You will be directed to an interface with two boxes. One box represents **where the files should come from**, the other box is (obviously) **where the files should be transferred to**. These two are called 'end-points' in Globus lingo. In our case, the ROMS files are on the chesapeake cluster, and you are trying to transfer them onto the external drive. So the first end-point to enter is `wmhpc#chesapeake` and the second end-point is `my_globus_endpoint_on_vlab6`. Each time you enter the name of a end-point you will be asked to enter your username and password corresponding to this end-point.
- Assuming you went through all this successfully, the next step is to indicate the path to your files at both end-points. For chesapeake it will be something like:
  `/ches/data10/your_username_on_chesapeake/your_directory_containing_the_roms_files/`
  (you can either type the path in the box 'Path' or you can navigate the folders using the mouse.)
- For the second end-point/box, the path should be:
  `/~/my_link_to_the_drive/`
- Next step is to select the files (or folders) to be transferred. Just click on the files (or folders) on the chesapeake end-point to select them. As usual, you can hold the 'Control' key of the keyboard to select multiple files/folders. You can also hold the 'Shift' key of the keyboard to select a range of files/folders.
- Next, make sure that the box 'verify file integrity after transfer' is truly checked at the bottom of the screen.
- Finally, click on one of the giant blue arrow (at the top) to transfer the file from the chesapeake end-point to the end-point representing the drive. You should then see a message saying 'Transfer request submitted successfully'.
- At this point the transfer is *ongoing*. Large file transfers take a long time so there is nothing else to do at this point but logout from the webpage and wait for the transfer to complete. To log-out of the Globus webpage, click on 'Account' in the upper-right corner of the Globus webpage. Then, click on 'Logout'. Finally, close the browser. Make sure that the window 'Globus Connect Personal' is still there (we need it for the transfer to continue). See the next next section on how to safely preserve your session on Vlab6 while the transfer is ongoing.

6 Preserving the file transfer over long periods of time

In many cases the transfer will require multiple hours if not days. The following instructions allow your transfer to continue during your absence.

- Click on the 'Power' symbol (circle with tiny vertical bar) in the upper-right corner of the screen.
- Click on the 'Lock' symbol. This will 'lock' the screen, i.e. it preserves your session on Vlab6 but also allow other people to log on Vlab6 and use it while your transfer is ongoing. Once the screen is locked, other users just have to hit a key on the keyboard and then select 'Log in as another user' to use Vlab6. It will not harm your transfer.
- You can monitor the file transfer from home or any other place. Just log onto Globus from your web browser
  `https://www.globus.org/app/transfer`
  and click on 'Activity'. It will show the status of your transfer(s), how many files were transferred so far, how many are left, the rate of transfer, etc.
  Note that Globus will automatically attempt to restart your transfer if anything happens (power outage, interrupted connection, etc).

- Once the transfer is completed, Globus will send you an email saying 'Succeeded!'. The email address is obviously the one you entered when you created your Globus account.
- At any time, you can resume your session on Vlab6 by hitting a key on the keyboard (to bring up the log-in window), making sure your username is selected, entering your password, and finally clicking 'Unlock'.

7 Disconnecting the external drive from Vlab6

Once you are completely done transferring files, you must follow these steps to disconnect the drive from Vlab6.

- In the window 'Globus Connect Personal', click 'Disconnect', then 'File' and 'Quit'.
- Next we must 'unmount' the drive. Click 'Applications' in the upper-left corner of the screen, select 'Accessories' and then 'Files'.
- This will have opened a new window. There should be a list on the left that somewhere says 'Seagate Expansion Drive'. Next to 'Seagate Expansion Drive' there is an 'Eject' symbol (triangle). Click on that symbol to unmount the drive.
- Unmounting a drive can take anywhere between 2 seconds or a few minutes if the transfer ended recently. To verify that the drive is unmounted, type the following line in a terminal window:
  `ls ~/my_link_to_the_drive/`
  If you get `ls: cannot access` then the drive is definitely unmounted.
- Unplug the USB cable from the computer.
- Unplug the drive from the power socket on the wall.

8 Logging out of Vlab6

Once the drive is disconnected, you are ready to log out of Vlab6:

- In the upper-right corner of the screen, click on the 'Power' symbol (circle with bar).
- Click on your user name, e.g., 'Pierre St-Laurent'.
- Select 'Log Out'.

A Appendix: Filesystems and transfer speeds

Most computer clusters use Linux and its filesystem `ext4`, while the default filesystem for the Seagate 'Expansion drive' is Microsoft's `NTFS`. Therefore, each time you transfer data between Linux and the external drive, some sort of conversion takes place in the background (each byte of information must transit through the computer's

RAM and CPU). In most cases, this expensive conversion is the primary bottleneck, *i.e.*, it is this conversion that ultimately determines the transfer speed.

Processes affecting the transfer speed (in decreasing order of importance):

$$\texttt{NTFS} \leftrightarrow \texttt{ext4} \gg \text{drive speed} \gg \text{USB connection} \tag{1}$$

'drive speed' is a penalty associated with reading/writing on the source and destination drives. As a general rule, solid state drives (SSDs) are considerably faster at reading/writing data than a conventional drive such as the Seagate 'Expansion' drive. For example, I experience substantially faster transfer speeds to the external drive if the source drive is a SSD.

'USB connection' is a penalty associated with the USB protocol itself. The 'Expansion' drive supports the USB 3 standard, the latter being associated with maximum transfer speeds on the order of 500 MB/s. This is a theoretical maximum speed that is never approached in practice because of the processes ('bottlenecks') mentioned above.

The present tutorial was written for cases where the archiving is not constrained by time, and where it is acceptable for the transfer to take multiple days. However, if time is of the importance, the primary bottleneck (the conversion `NTFS`↔`ext4`) can be removed by re-formatting the Seagate 'Expansion' drive to the `ext4` format:

1. Power the external drive and hook its USB cable to the computer.
2. Make sure the external drive isn't mounted (un-mount it if the computer automatically mounted it).
3. Type `lsblk` to determine which device is associated with the external drive. The next steps assume the device is `/dev/sdc`
4. Type `sudo gparted /dev/sdc`
5. You should see two partitions, a small one containing Seagate files and a larger one with a `NTFS` format (`sdc2` in this example).
6. Right-click on `sdc2` and 'Format' as a 'ext4' filesystem.
7. Edit → Apply all operations
8. Right-click on `sdc2` and 'Manage Flags'. Remove the existing flag `msftdata` (it is now irrelevant).
9. Right-click on `sdc2` and 'Label file system' to give it a meaningful name, *e.g.*, `seagate_external`
10. Edit → Apply all operations (if needed).
11. GParted → Quit
12. Un-hook the USB cable.
13. Re-hook the USB cable, and verify that Linux automatically recognizes+mounts the new `ext4` partition.
14. The new partition will have a default ownership of `root`, *i.e.* only the system administrator can write on it. To change ownership:
    `chown -R pierre:pierre /media/pierre/seagate_external`