# Programming for Bioinformatics | BIOL7200

## Week 2 Exercise

August 31, 2021

The goal of these exercises is to get you used to working with some basic UNIX commands and their options. Try and think about what each command does, and where it can be used. Reading the man page for each of these is **recommended**.

We’re going to cover the following commands:

|  |  |
| --- | --- |
| Command | Explanation |
| who and finger | Look up information about users |
| top | See the status of running system processes |
| free | See the amount of free and used memory |
| wget | Download files |
| curl | Download files |
| du | See the amount of used storage space |
| df | See the amount of free storage space |
| ps | Lookup information for specific processes |
| vim | Another text editor |
| fg | Foreground |
| which | Find location of a binary on your PATH |
| ln | Create link |
|  |  |

### Instructions for submission

* Prepare two solution sheets for this exercise - one for submission and another for your own reference. **Only submit the submission sheet on Canvas**. The solution sheet you create for your reference should help you going forward - make it as detailed/brief as you'd like for your own learning style. For the submission sheet, copy the question and write the correct answers below the question.
* Name the submission sheet: gtusername.txt , gtusername.pdf or gtusername.docx

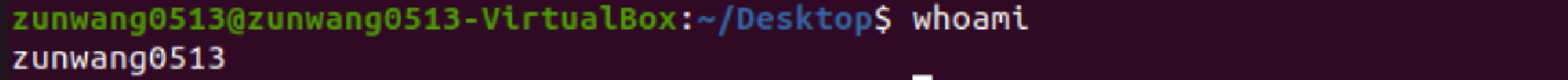
### Grading Rubric

The assignment will be graded out of 100.

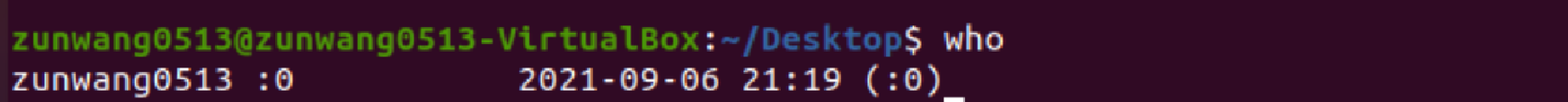
* 15 points for correctly naming and submitting your sheets.
* Questions 1-7 carry 5 points each (35 points)
* Questions 8-12 carry 10 points each (50 points)

### Exercises

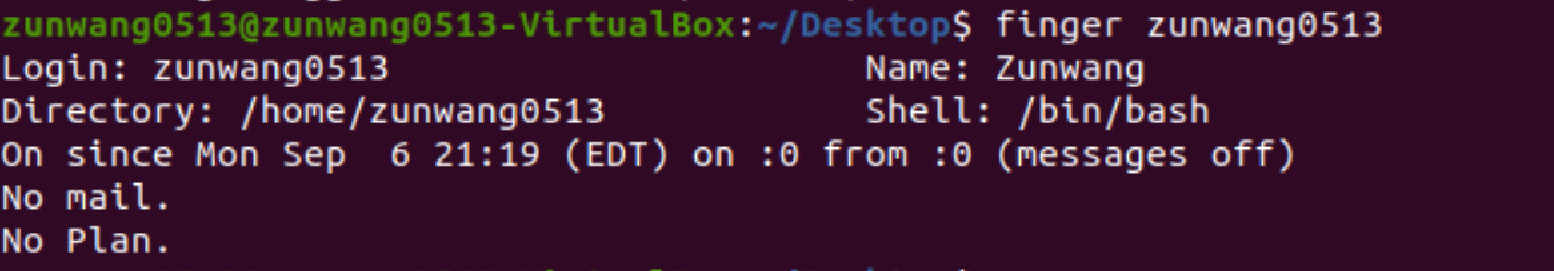
1. Using who and finger
   1. Use who to see yourself logged into the machine



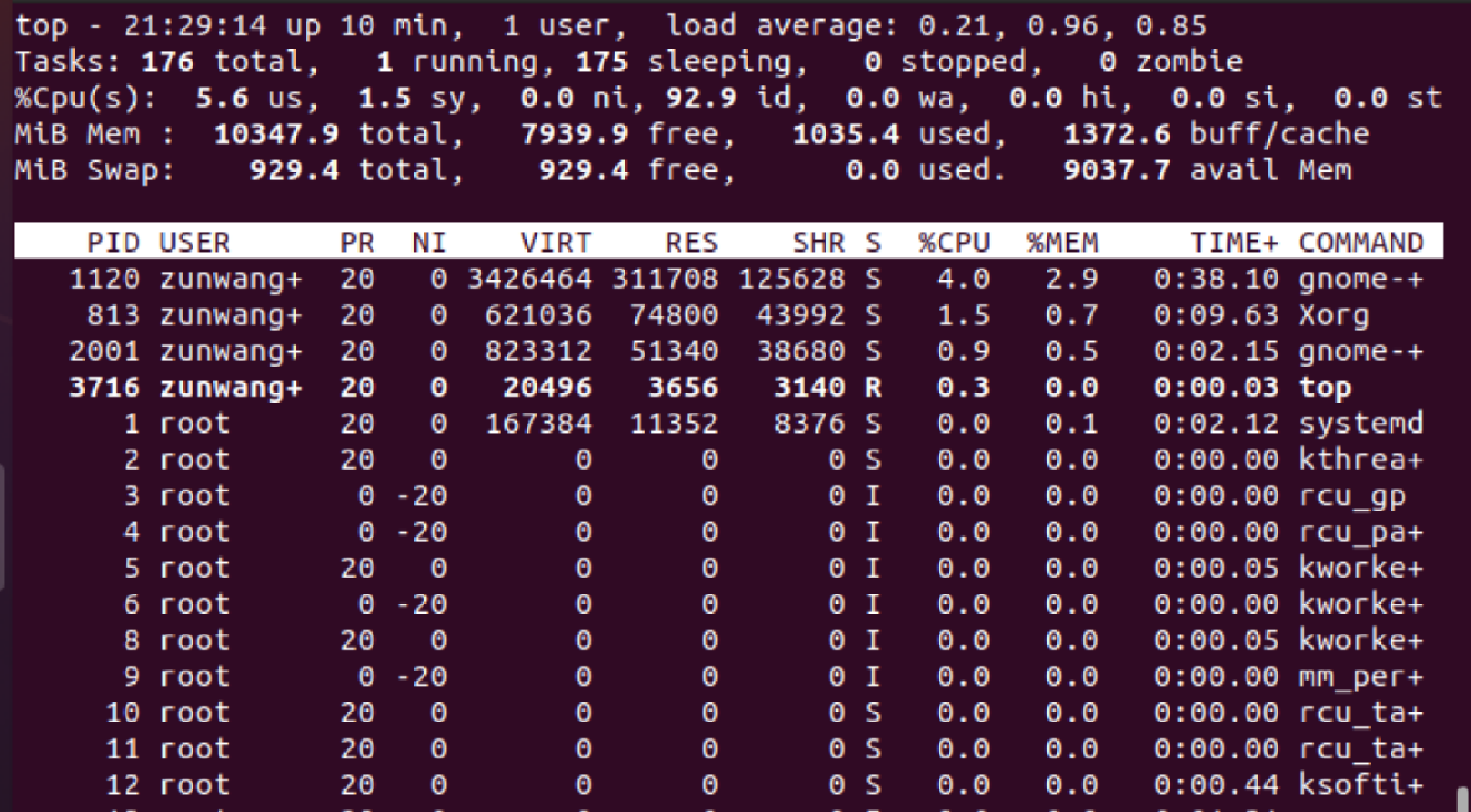
* 1. Find ***all*** the logins on the machine



* 1. Use finger to find information about yourself on your Linux VM/whatever you are using



1. Monitoring system usage with top and free
   1. Use top to look at what’s running on the computer

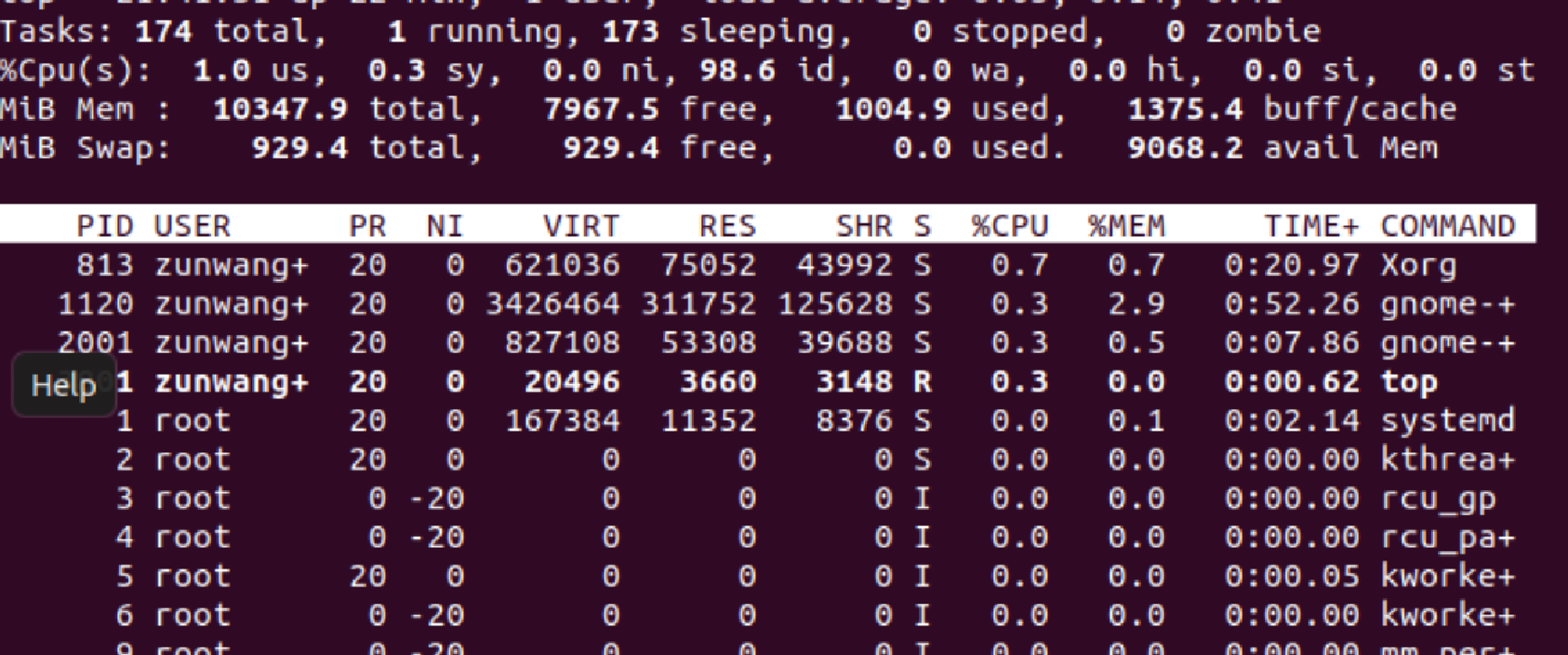


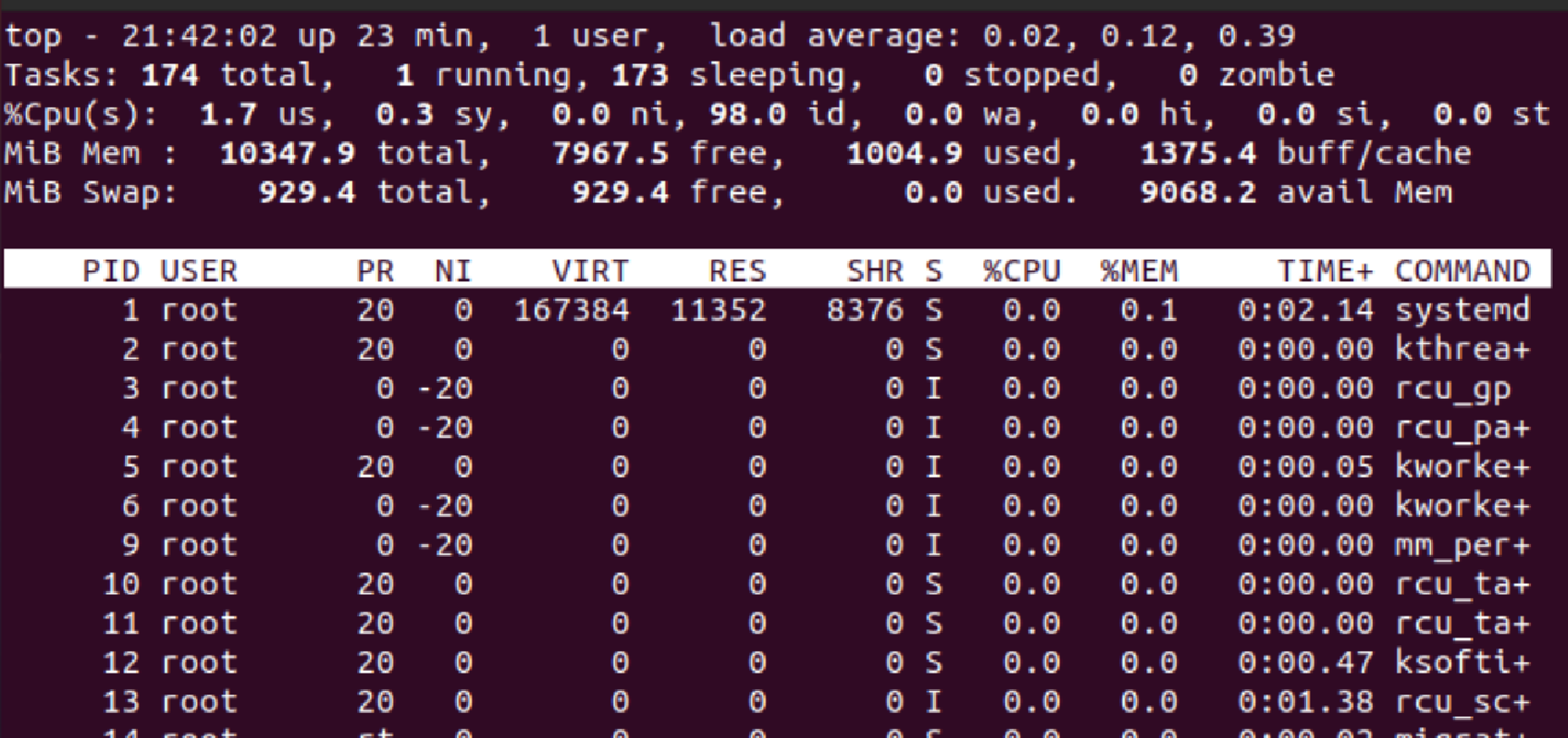
* 1. Find the CPU and memory usage of a program; how much memory is left?

For the top program zunwang+, 1035Mb is used and 7939.9Mb is left.

* 1. Order the processes by CPU usage, ascending and descending

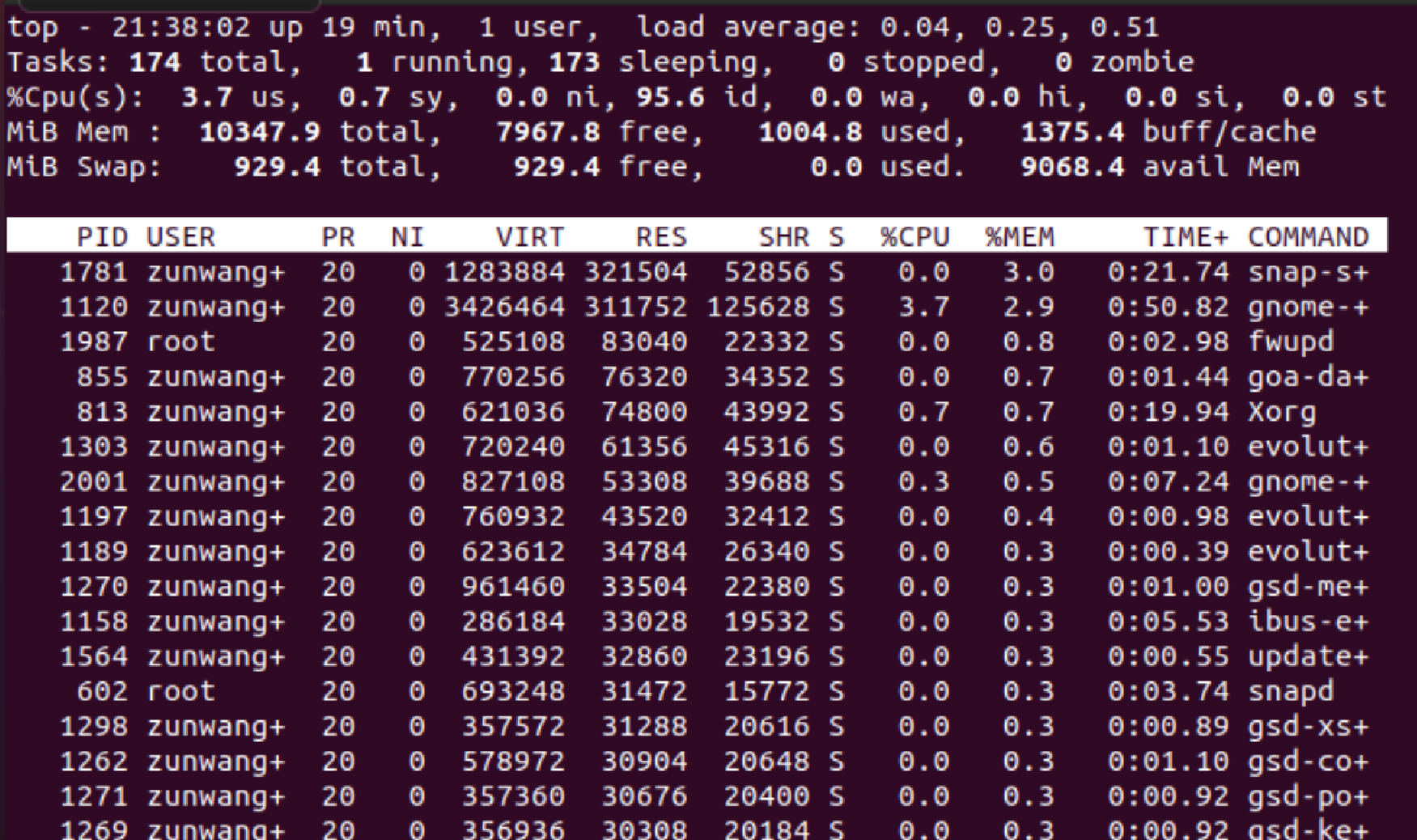
Press shift+p, to make it descending, them press shift+R

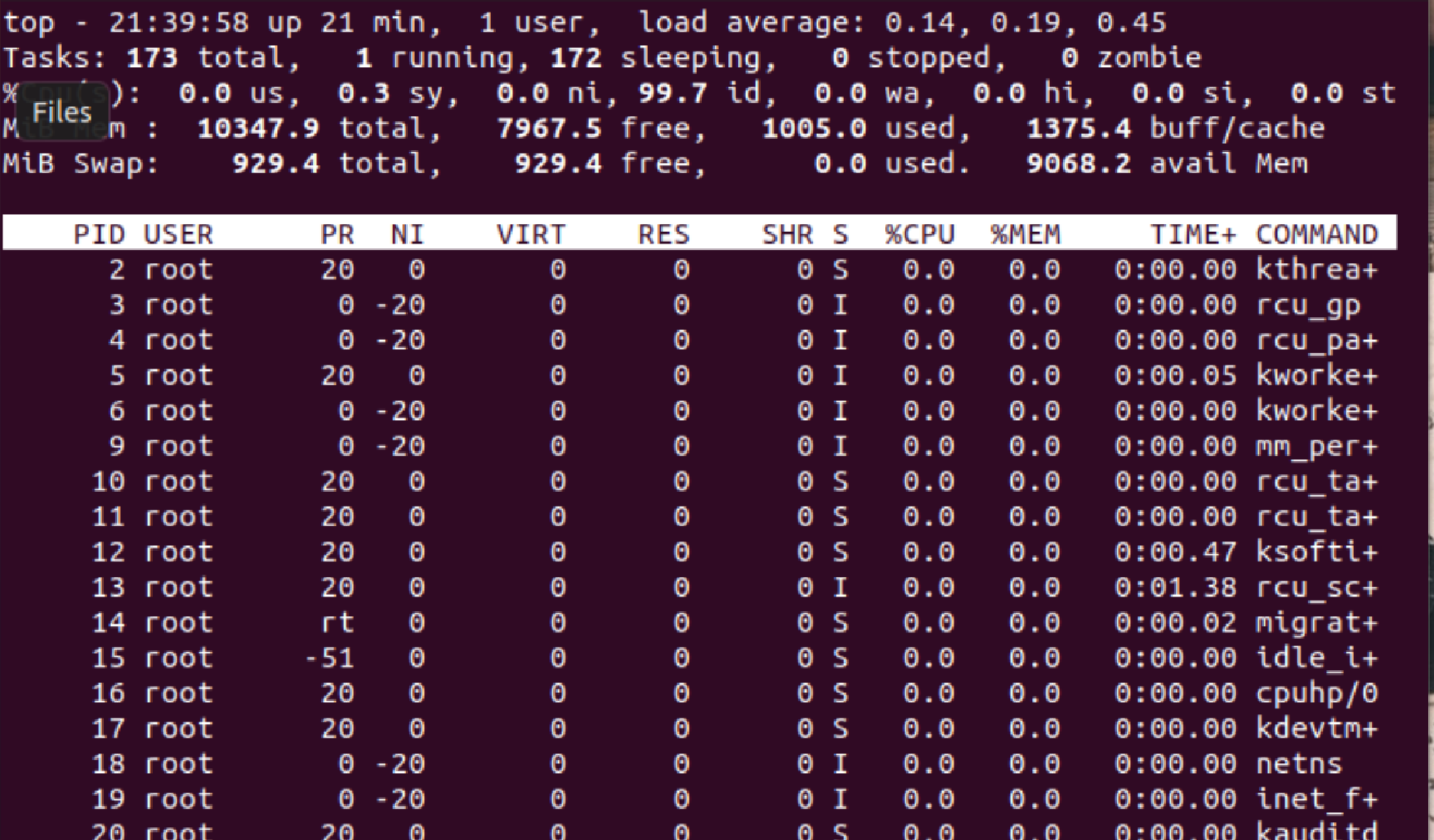




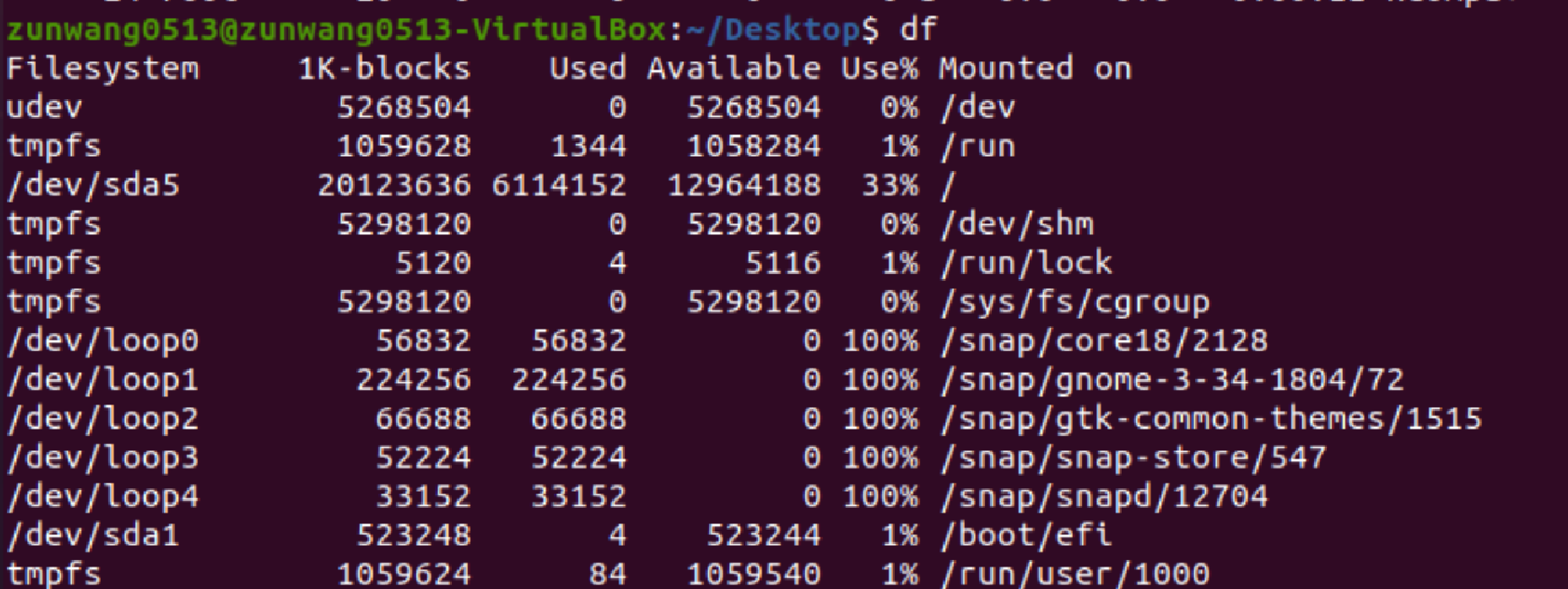
* 1. Do the same for memory

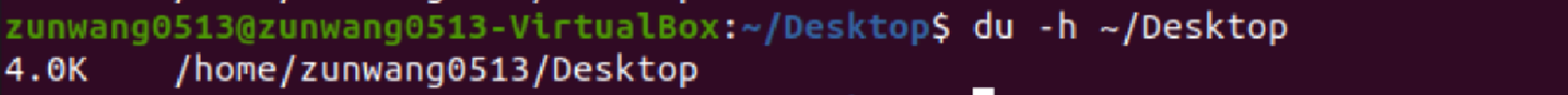
After opening top, press shift+M, to make it descending, then press shift+R





1. Looking at hard disk space
   1. Use df to find the free space on your machine’s hard disk

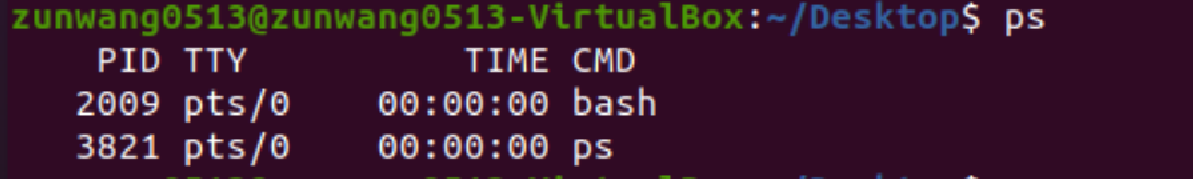


* 1. Use du to find the total size of some directory; make the output easy to read, e.g. ‘315M’

Use du -h

The size of the directory Desktop is 4k.

1. Finding running processes
   1. Find all the processes that you have running using ps

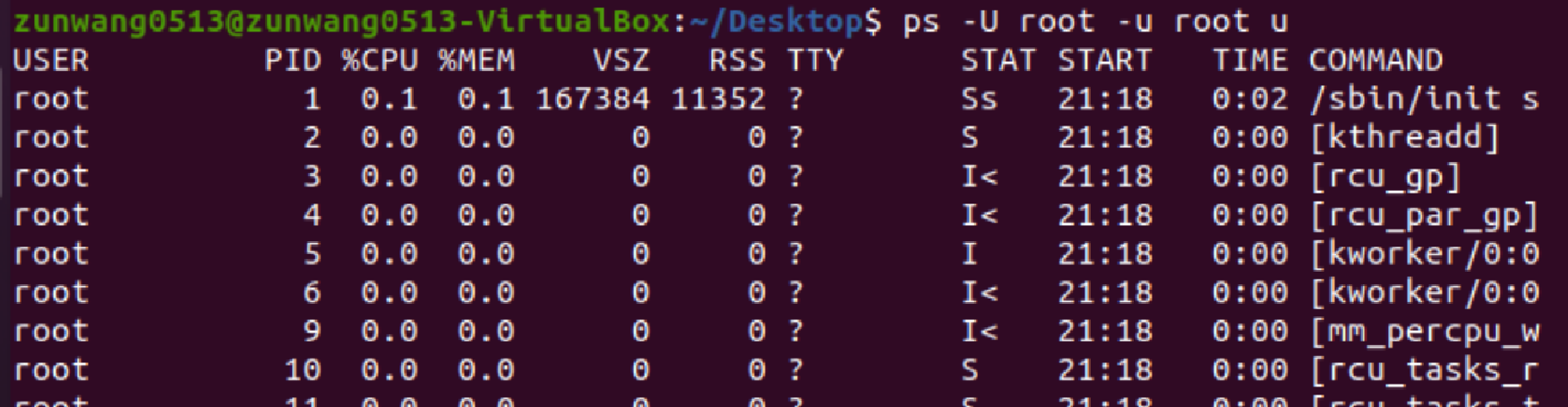


* 1. Find **all** processes that are running on the machine



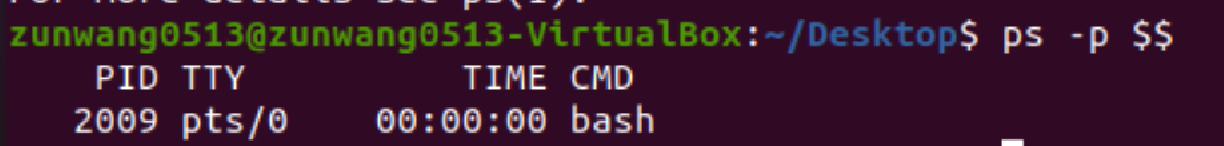
* 1. Find all of the processes being run by the user ‘root’ in user format

ps -U root -u root u



1. Murder most foul
   1. Find the PID of your login shell using ps

PID is 2009

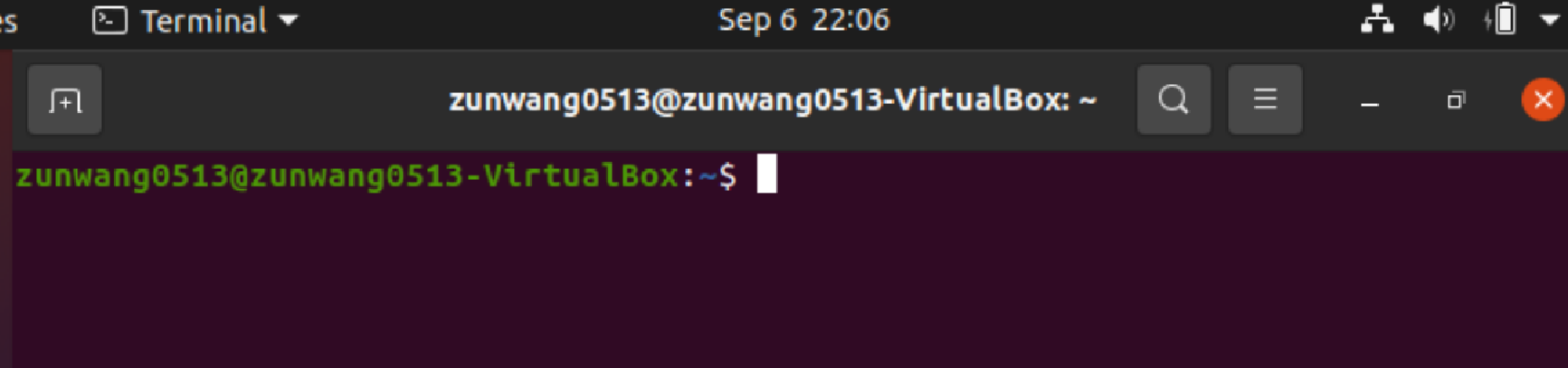


* 1. Kill your login shell with a command

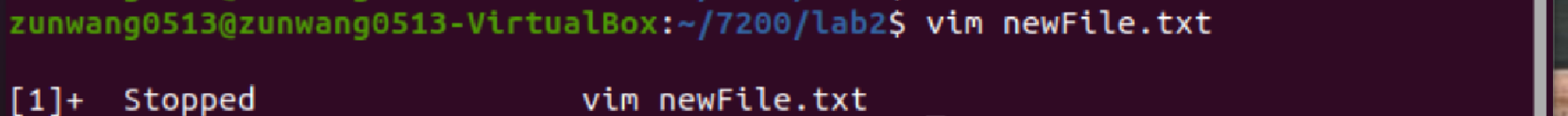
kill -HUP 2009

The shell closed

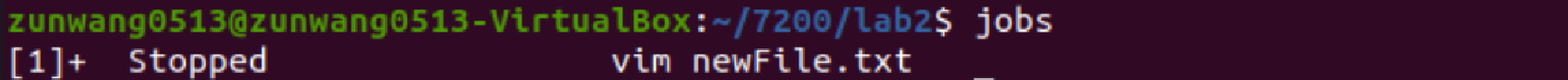
* 1. Open a new shell



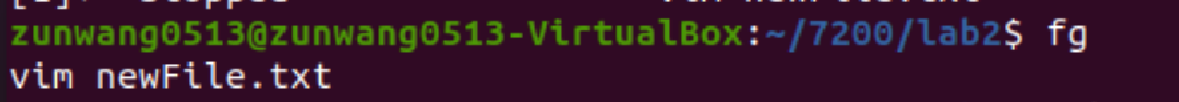
1. Managing jobs
   1. Start editing a file with vim, then put it in the background using Ctrl + Z



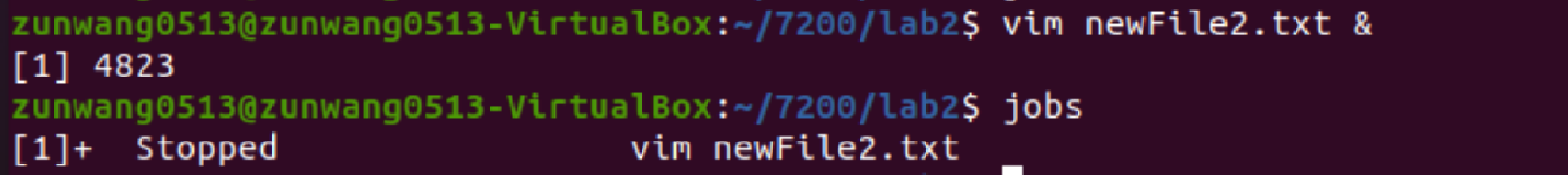
* 1. Find it using the jobs command



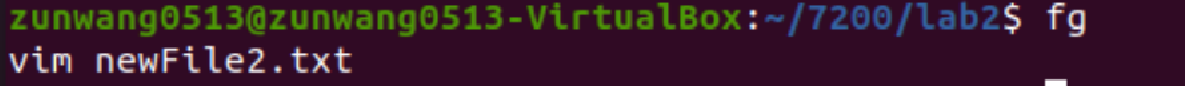
* 1. Restore it to the foreground using fg



* 1. Start a new job in the background using & after the command



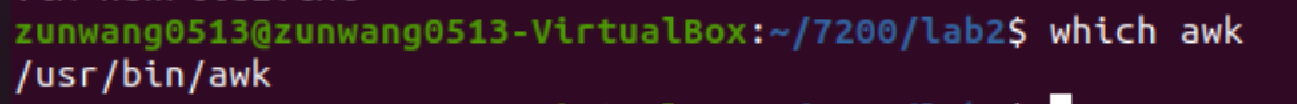
* 1. Move the new job to the foreground using fg



* 1. What is the difference between the commands Ctrl + c and Ctrl + z

Ctrl+c is to kill a process, and ctrl+z is to pause a process.

1. Finding the location of binaries, linking them and alias
   1. Find out where the awk command is located using which

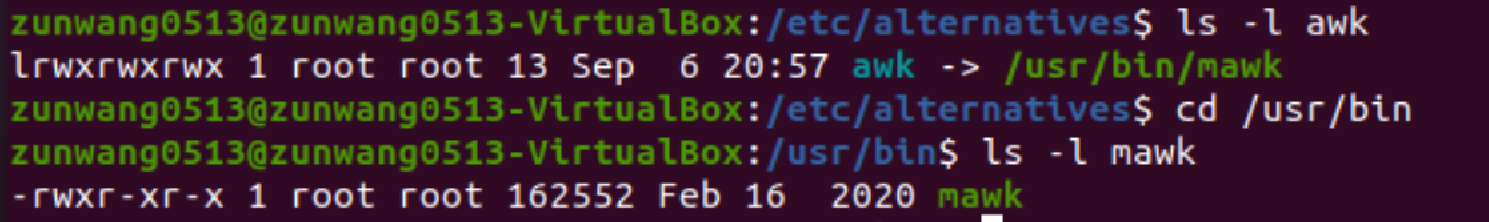


* 1. Find out if it is a link or a real binary. Follow this through until you have found the real binary.

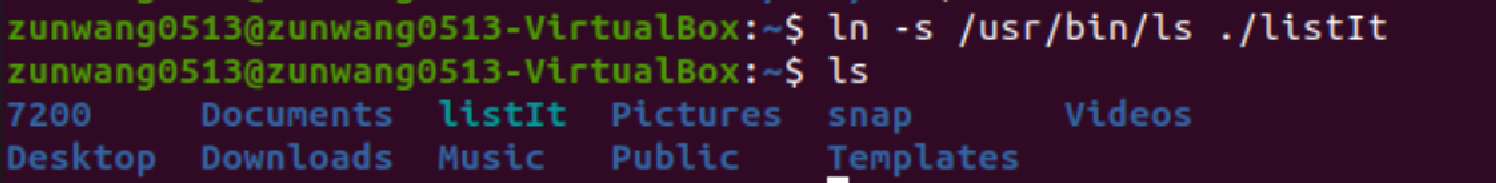
It is a link.

Then go to /etc/alternatives/

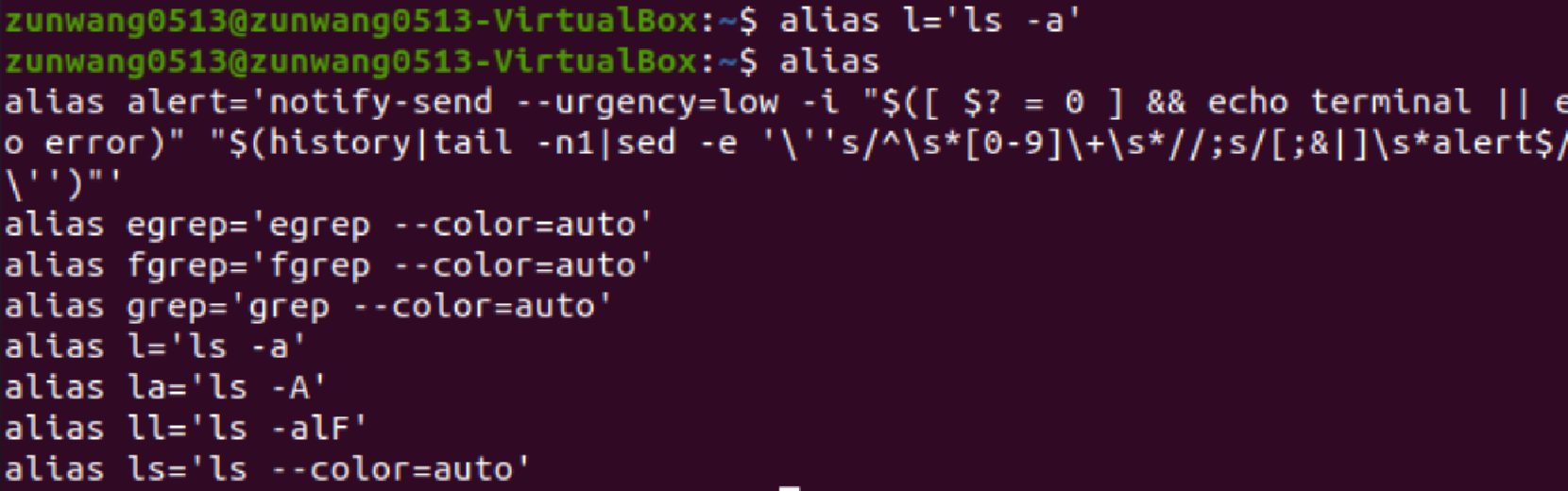
The awk there is still a link.

The real binary lies in usr/bin/mawk

* 1. Create a symbolic link to the ls command in your home directory called listIt using the ln command



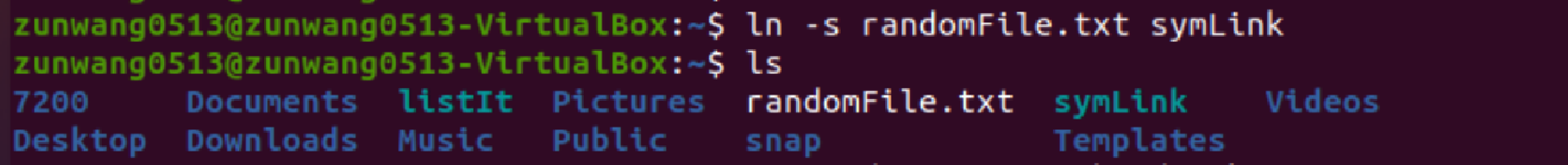
* 1. Create an alias for ls -a



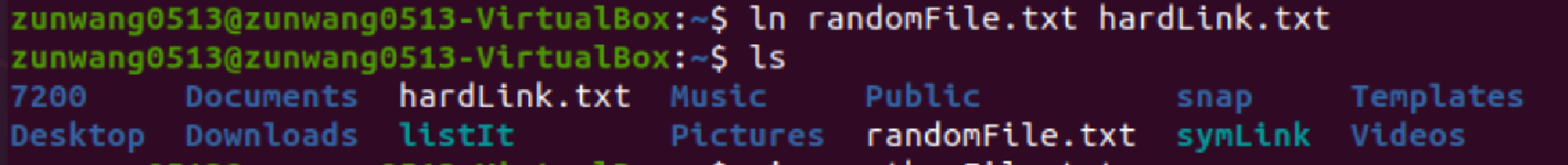
1. Hard vs Symbolic links
   1. Create a file randomFile.txt and type the following in it:  
      This is some text.  
      This is some more text.  
      And some more…



* 1. Now create a symbolic link for randomFile.txt; name the link symLink.txt



* 1. Now create a hard link for randomFile.txt; name the link hardLink.txt



* 1. Create another file anotherFile.txt. Put this text in it:  
     1234  
     5678  
     9101



* 1. Now delete randomFile.txt; which of the two links still works?

The softlink still works. The hardlink does not work anymore.

* 1. Rename anotherFile.txt to randomFile.txt. Which of the two still works? If both work, is there a difference between the two? If only one works, why?

mv anotherFile.txt randomFile.txt

The softlink still works. The hardlink does not work. Because harlink is the exact copy and does not dynamically link to that file.

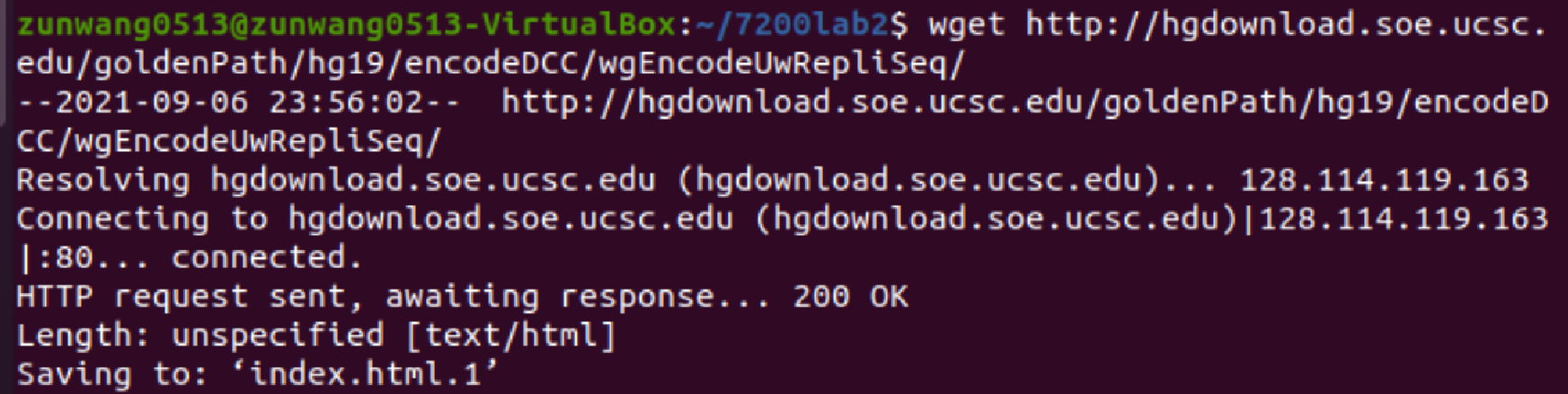
1. Downloading UCSC Genome Browser files with wget Toy exercise – the objective here is only to learn how to use wget. You don’t need to wait for each file to finish downloading; you can quit the download if you have determined that you are using the correct command. Once you see a file getting downloaded, kill the download process if you want. Read wget manual for helpful options.  
   1. What is the UCSC Genome Browser? What information does it store? What organisms does it focus on? What is the FTP site for accessing the databases?

UCSC genome browser is a platform to view the sequences of various species. It stores genetic information generally DNA sequences. It focuses more on vertebrates. https://hgdownload.soe.ucsc.edu/downloads.html

* 1. Navigate to the ENCODE repository on the UCSC FTP site.
     1. What is an ENCODE repository?

ENCODE is the encyclopedia of DNA elements. It contains DNA files.

* + 1. Download a file (pick a small one of your choice) from the ENCODE repository on the UCSC Genome Browser with wget and briefly explain what data it contains.



* 1. Download all of the Pol2b binding data available in one line with wget

Wget -m -nd -A “\*Pol2b\*” http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC

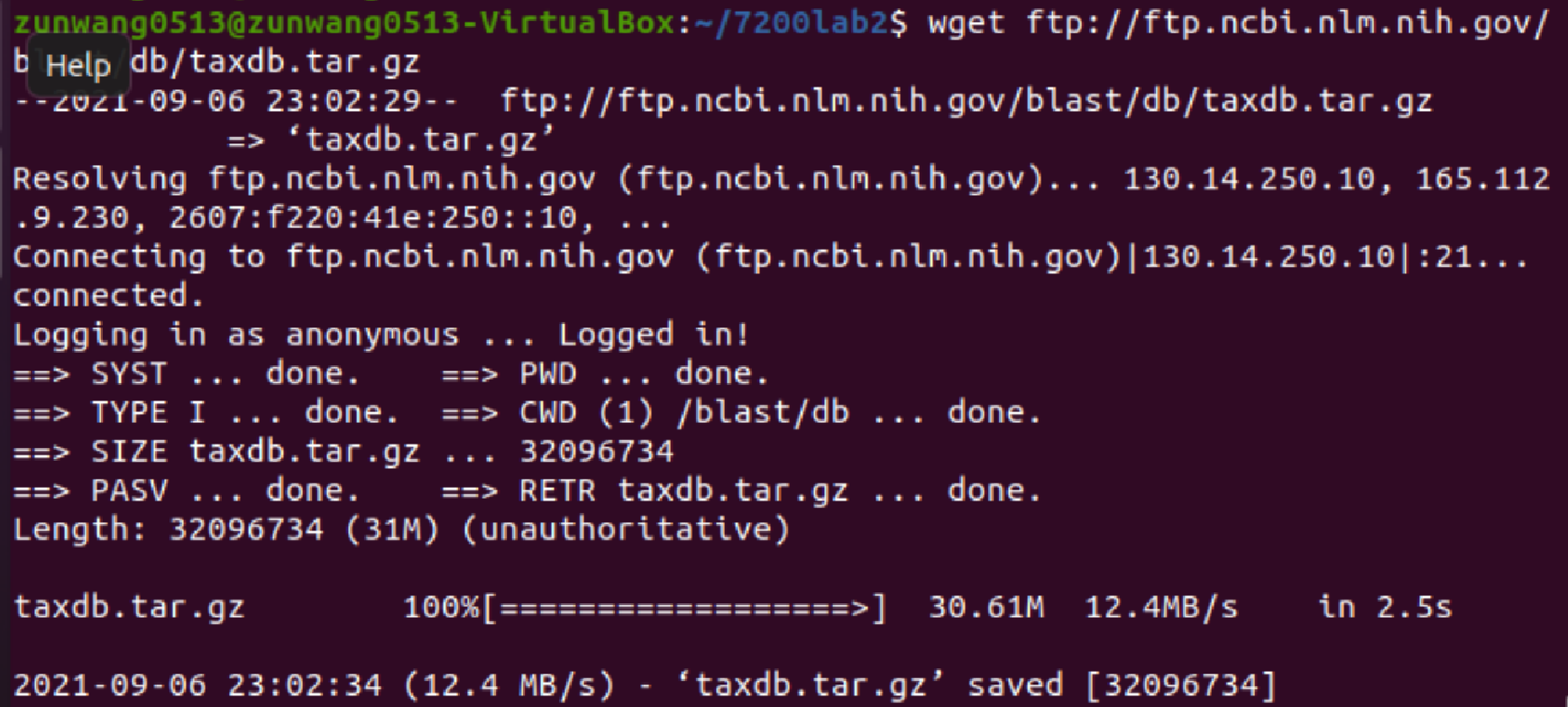
* 1. The database folder located in hg19 in goldenPath is an important folder. We are going to fetch four files: knownGene.txt.gz, knownGene.sql, kgXref.txt.gz and kgXref.sql. Copy their link addresses and paste in a file. Now use wget to download the files by reading the URLs from the file you just created. In other words, this time the input URL comes from a file rather than from you giving it in the command, and there are multiple URLs. **Hold on to these files; we will use them later during the course.**

Wget <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/knownGene.txt.gz>

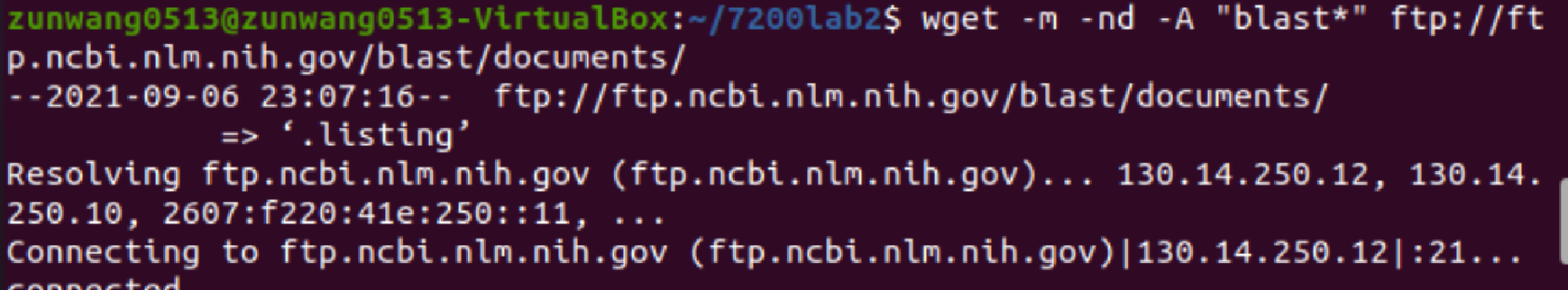
* 1. Test if the URL http://www.ncbi.nlm.nih.gov/SNP/snp\_ref.cgi?searchType=adhoc\_search&type=rs&rs=rs12345 is correct using wget. [Hint: Have you seen error 404 while accessing a website? See if HTTP Status codes can help].

It does not work

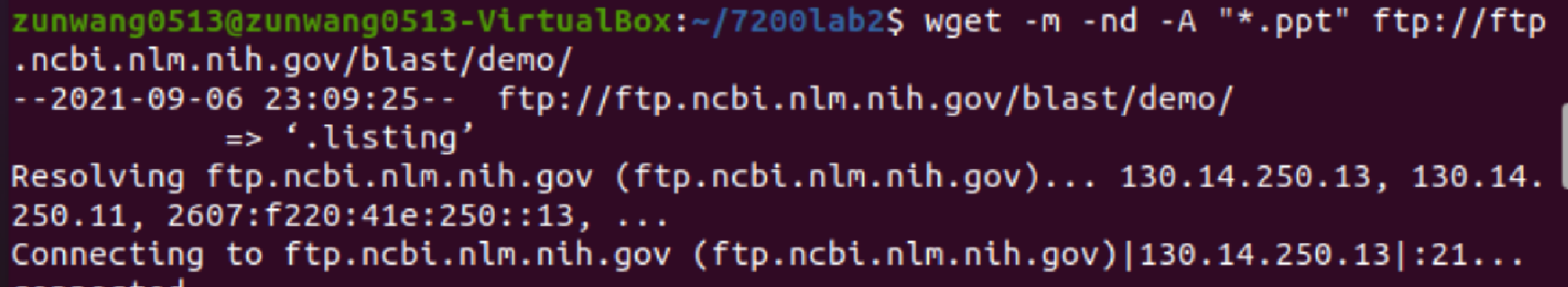
1. Write one-liners to perform the following actions with wget or curl
   1. Download the file taxdb.tar.gz from ftp://ftp.ncbi.nlm.nih.gov/blast/db/



* 1. Download all files that start with “blast” from ftp://ftp.ncbi.nlm.nih.gov/blast/documents/



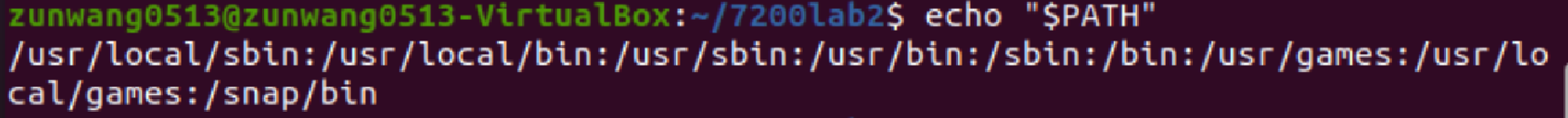
* 1. Download all “ppt” files from ftp://ftp.ncbi.nlm.nih.gov/blast/demo/



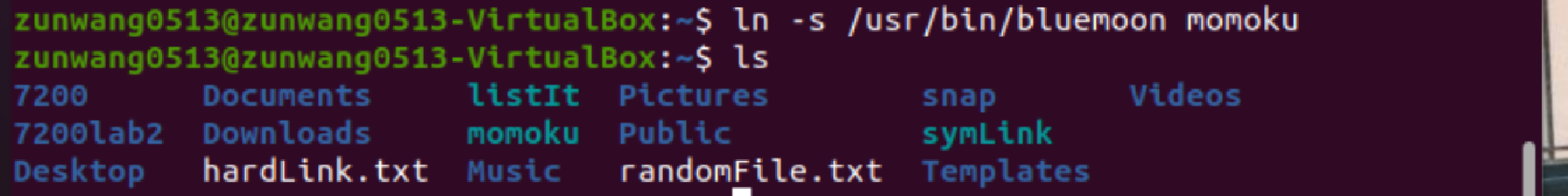
1. The PATH to enlightenment

The PATH environmental variable tells the shell where to look for executables. When you use the ls command, it works because the shell knows to look for it on the PATH .

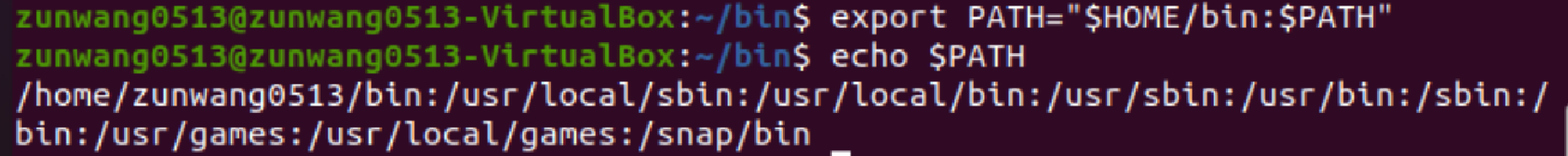
* 1. Print your PATH environmental variable to the screen. What are the directories in it? Some of them should look familiar at this point.



* 1. Make a directory call bin in your home directory and put a symbolic link to some already existing executable in the newly created bin directory. Call the link something different from the actual executable or this is all for naught.



* 1. Add you bin directory to your PATH environmental variable. Keep in mind what character separates the different directories in the PATH environmental variable. If you mess up your PATH , then just leave the shell and make a new one.



* 1. Try running the executable through the symbolic link. Did it work? Use which to see the alias.



* 1. Add your new bin directory to your PATH in your login file. Make sure you can still access all of the programs on your PATH.

Directly export bin to path

* 1. Congratulations! You now have a centralized folder that will be on your PATH every time you log in. You can install executables here if you don’t have permission to install things.
  2. If you’re on a system with multiple users, what might be an advantage of install an executable in one of the directories on the default PATH ?

So that every users have the permission to call the executable in the path.

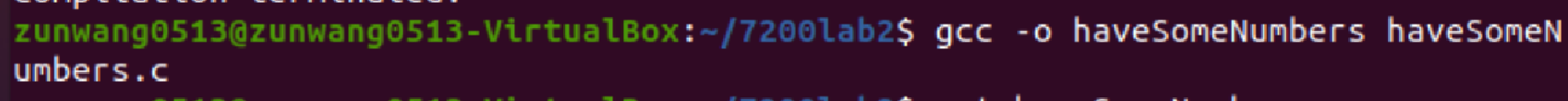
1. The C compilers

There are various C compilers available today. They pretty much do the same thing: turn C code into a program, but they do it with widely varying efficiency.

1. Find out what your cc command points to. In other words, what is your default C compiler?

Gcc is the default compiler.

1. Compile the haveSomeNumbers.c file from Canvas and run the result (you’re going to need the -o flag. Read about it before using it). Add the results to your PATH and run it that way.





1. C is beyond the scope of this course, but try to figure out what the haveSomeNumbers.c file does anyway. This program is very simple. Many, many bioinformatics programs are written in C.

In the for loop, i increased by 1 every loop and is printed out from 1 to 100.

1. Look up comparisons of gcc and icc online. What are the advantages and disadvantages of each? What other C compilers exist?

Generally, icc uses less time, uses better optimization. But gcc is more convenient to use and has more user freedom. Other c compilers include clang and tiny c compiler.