SHIVGAMI

Simplifying tHe titanIc blastx process using aVailable GAthering of coMputational unIts

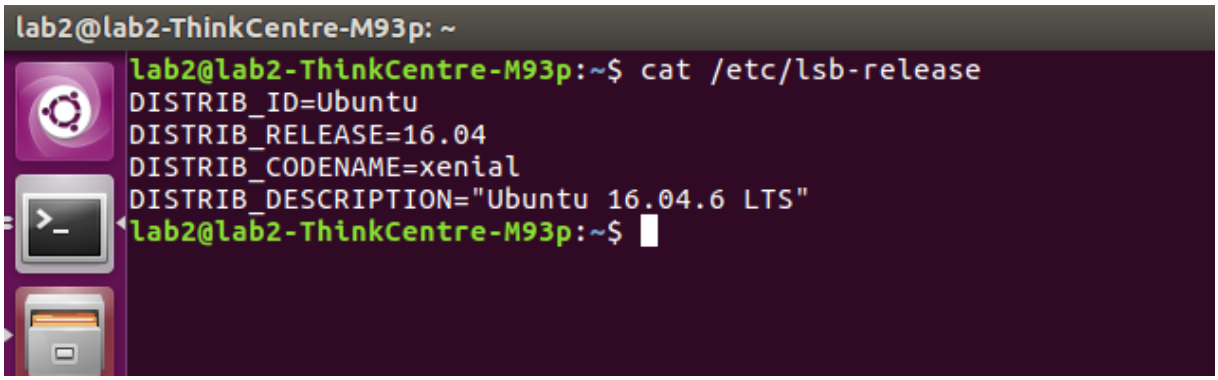
About SHIVGAMI

blastx is widely used program to annotate unknown nucleotide sequences. SHIVGAMI assemble the available CPU cores from LINUX machines and utilizes them to run blastx process. SHIVGAMI motivates those researchers/labs who have small computational units instead of high end computing facilities. SHIVGAMI assemble those small units, divide the blastx input queries amongst them and finally combine those results into a single output file. Additionally, the well-equipped labs, whose servers are loaded can also combined their other units with available CPU cores using SHIVGAMI.

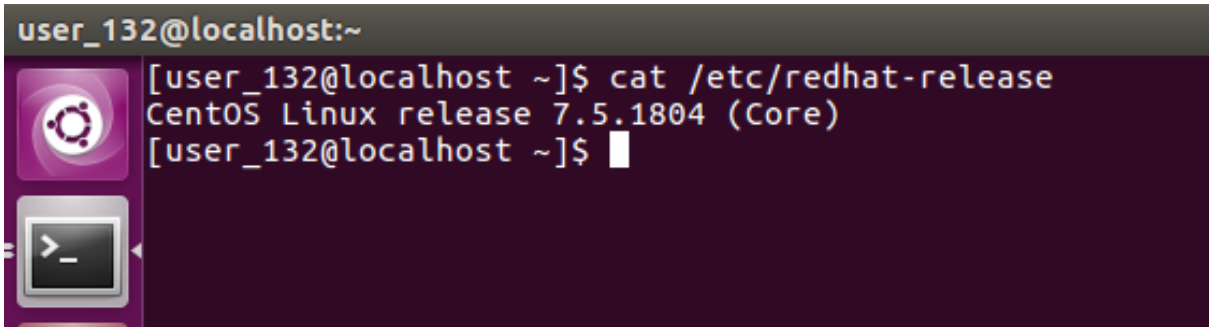
1. Gathering of Linux computational units

Collecting information for those computational units which are having Linux Operating Systems(OS) User can include all types of Linux Machines like Normal PCs, High-config PCs and Servers. We have not tested SHIVGAMI on laptops, but user can include, if the laptop is LAN connected.
We have tested SHIVGAMI on Ubuntu and CentOS

Ubuntu



CentOS

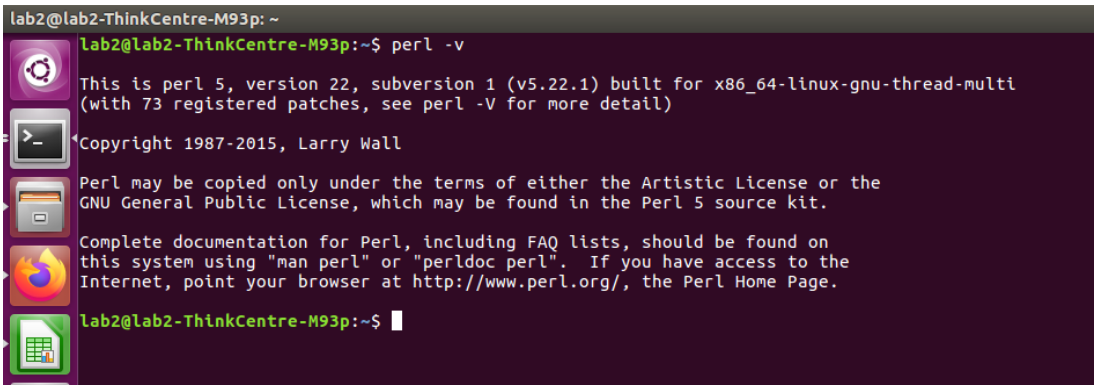


2. Preparation

(I) Prerequisites

i. Perl

Below command will show you, which version of perl is installed perl –v



If, perl is not installed then, run the following commands:

Debian/Ubuntu

sudo apt-get install perl

RedHat/CentOS

yum install perl (run this as a root user)

ii. sshpasss

Debian/Ubuntu

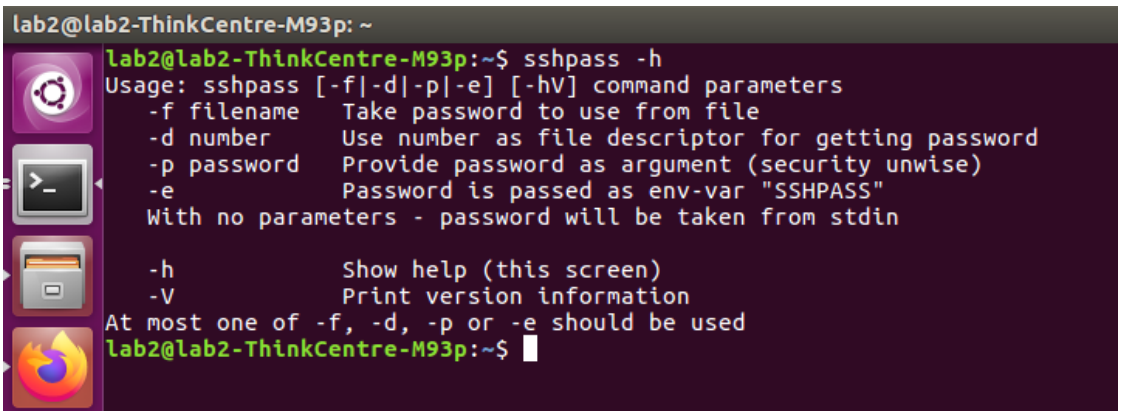
sudo apt-get install sshpass

RedHat/CentOS

yum install sshpass (run this as a root user)

user can re-check the installation using command:

sshpass –h



iii. OpenSSH Server

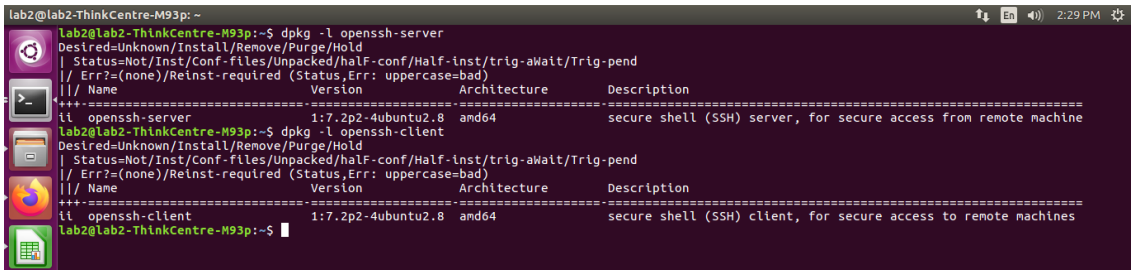
iv. OpenSSH Client

Ubuntu OS may have openssh-client, if not, it can be retrieved using commnd:

Debian/Ubuntu

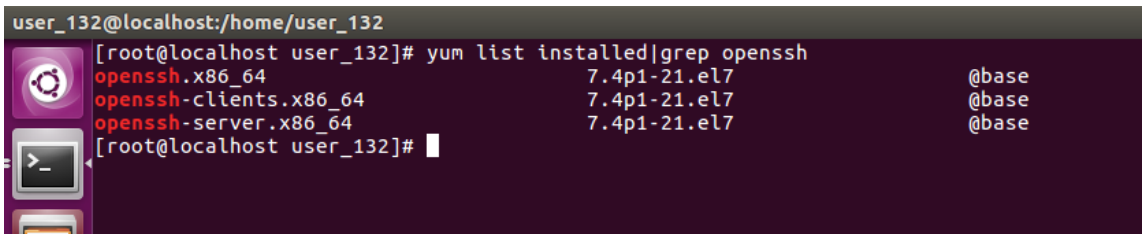
sudo apt-get install openssh-server openssh-client

The installation can be verified using commands:
sudo dpkg -l openssh-server
sudo dpkg -l openssh-client



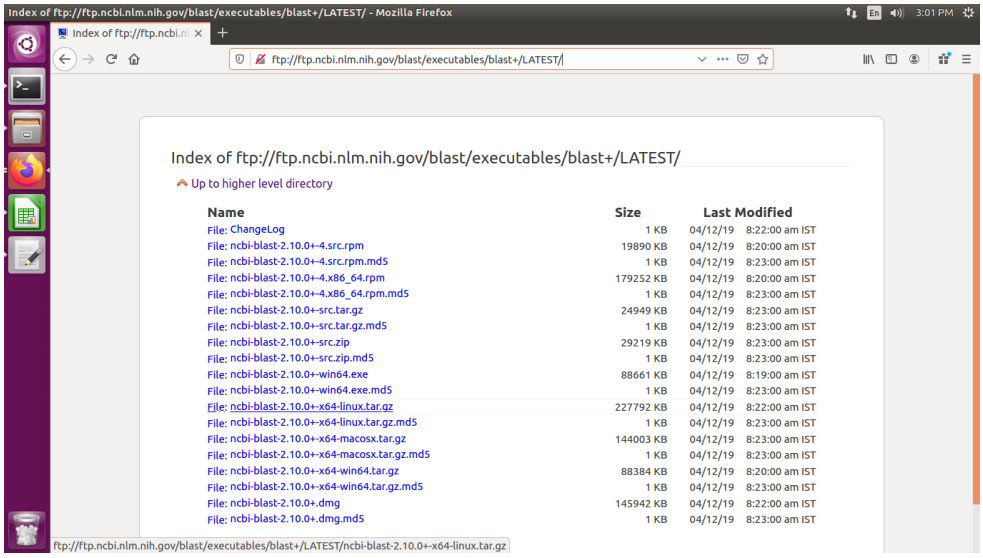
RedHat/CentOS

sudo apt-get install openssh-server openssh-client



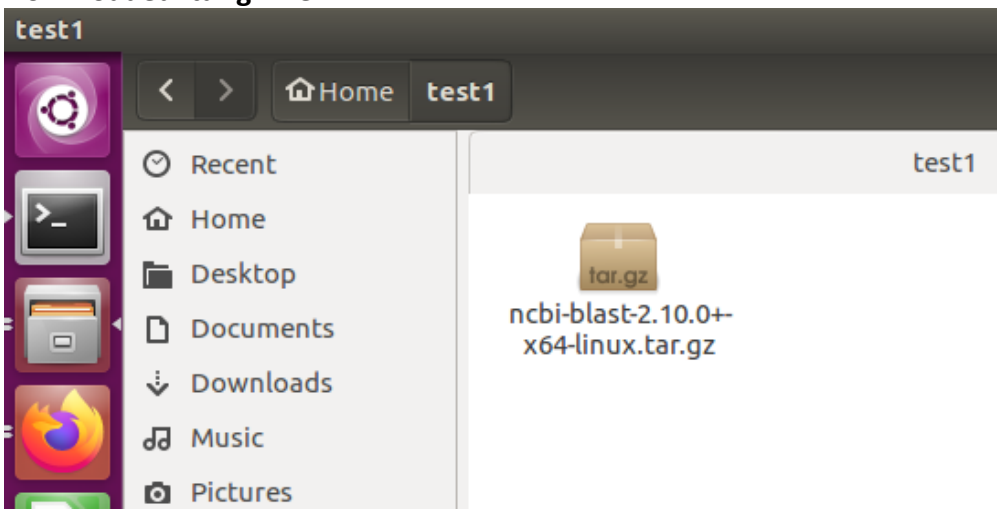
v. Blast+

The latest Blast+ program can be obtained from NCBI ftp site, below are the steps
ftp://ftp.ncbi.nlm.nih.gov/
|
v
click on blast
|
v
click on executables
|
v
click on blast+
|
v
click on LATEST
(final link is: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST)

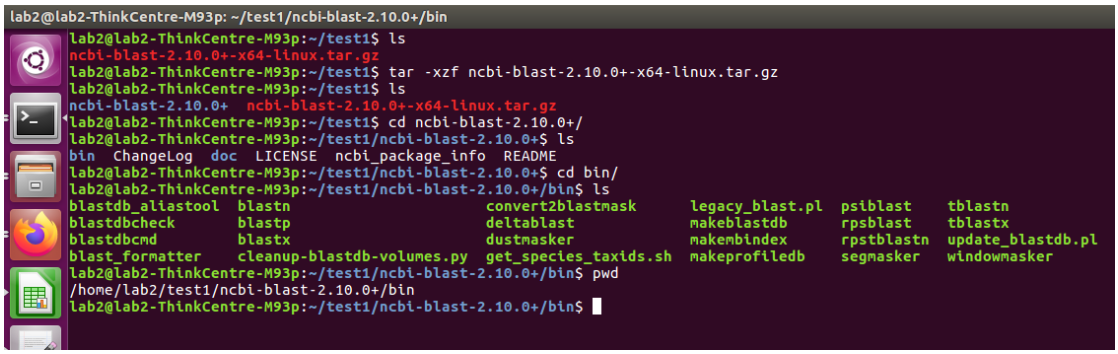


There are various file types, but linux.tar.gz is an easy way to go!
Below is an illustration:

Downloaded .tar.gz file



tar command to extract:



vi. Protein database

A user can make a customized protein database as per the requirement, But to use it for blastx, the database must be formatted as per the NCBI blast+ package.

Suppose a user has a protein database fasta file called - protein.fasta
The command to format protein.fasta is:

```
/home/lab2/test1/ncbi-blast-2.10.0+/bin/makeblastdb -in
protein.fasta -input_type fasta -dbtype prot
```

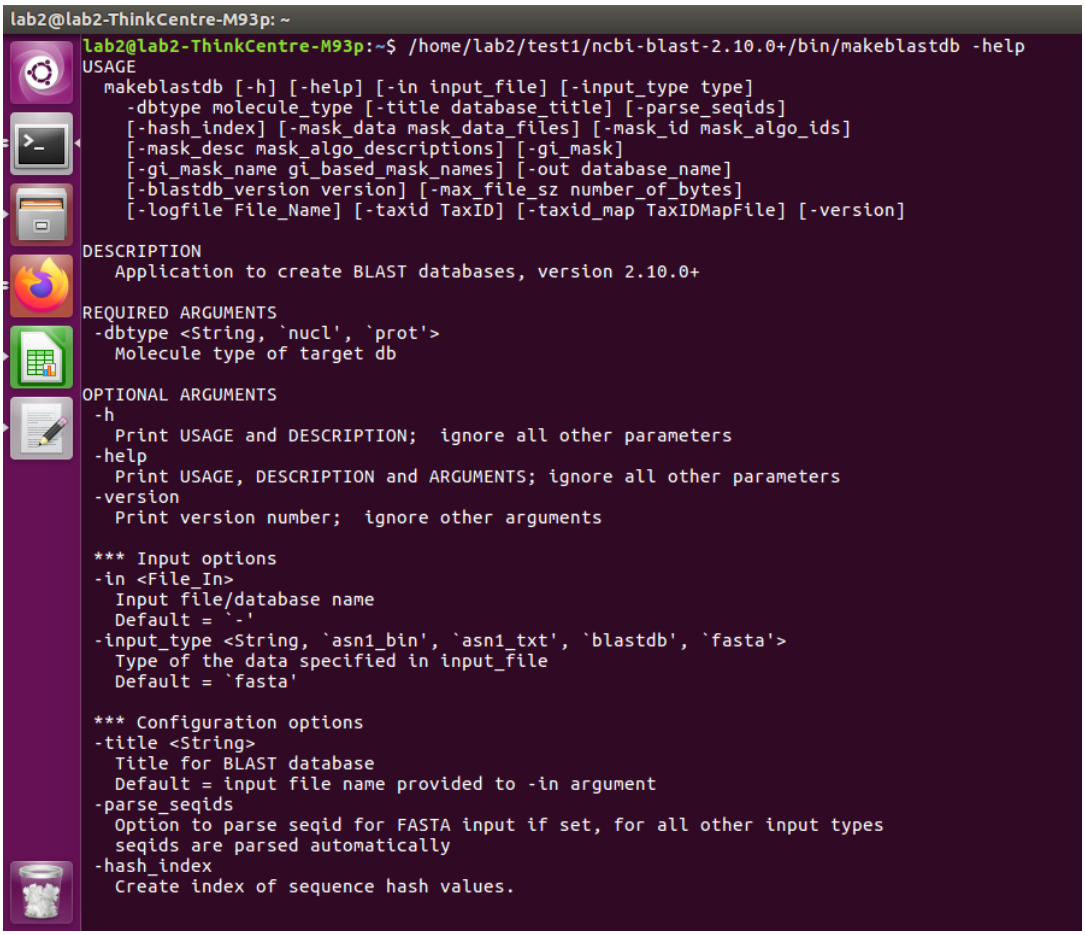
Where,

`/home/lab2/test1/ncbi-blast-2.10.0+/bin/makeblastdb`:
is a fullpath to the makeblastdb executable

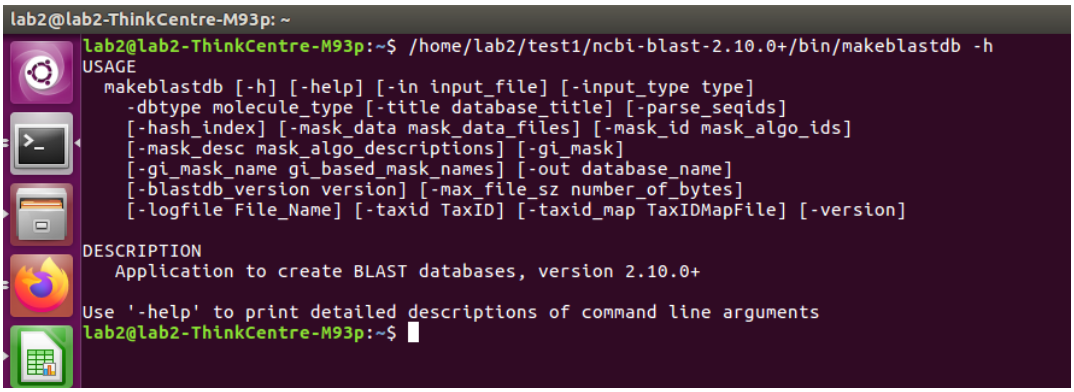
-in : input file
-input_type : input file type
-dbtype : molecule type (String OR nucl(nucleotide) OR prot(protein))

The same can be seen in detail with makeblastdb help

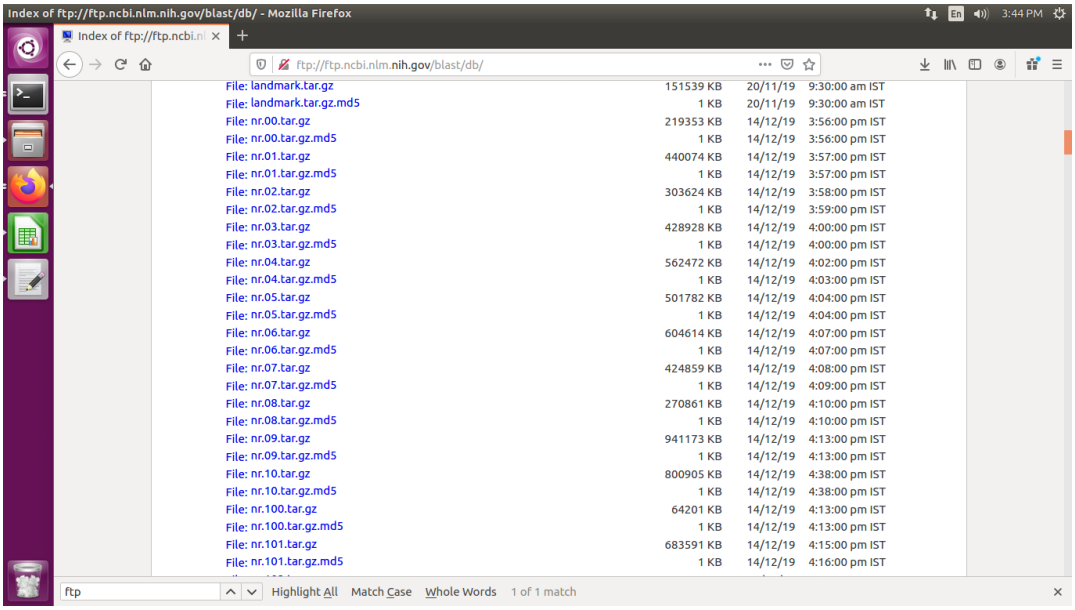
Long Help



Short Help



If a user wanted to run blastx against NCBI-NR database, then the advantage is, it can be obtained as preformatted and can be downloaded from: ftp://ftp.ncbi.nlm.nih.gov/blast/db/



A user must donwload all files with prefix "nr."
(nr.00 to nr.142 as on 13-Jan-2020 3:43 IST)

vii. SSH connection

SSH (Secure Shell) connection is must to run this software.

(II) Master node

A master node is a computational unit with Linux OS, which distributes the process amongst the child nodes, but do not take part in the blastx process.

Master node can be a simple computer or a server or a high-end pc.

(III) Child node system information

Child nodes are all computational units with Linux OS, except master node.

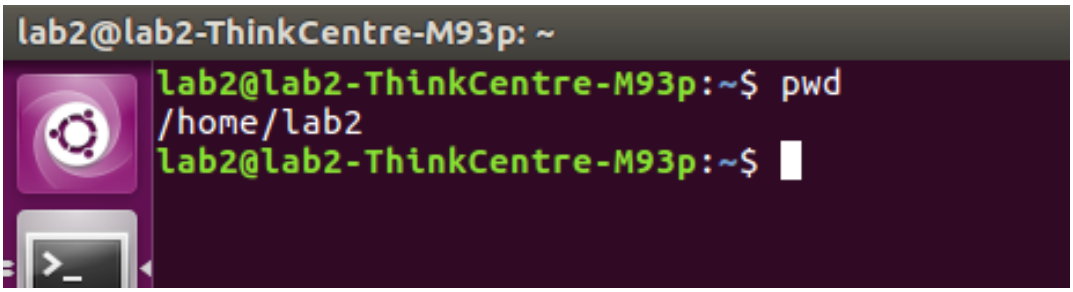
A text file "child_node_system_info.txt" (user can rename it, if required) creation is required, which must has all child nodes' information with below mentioned 7 entities:

i. Serial no.

This is just a serial number like 1,2,3..

ii. User

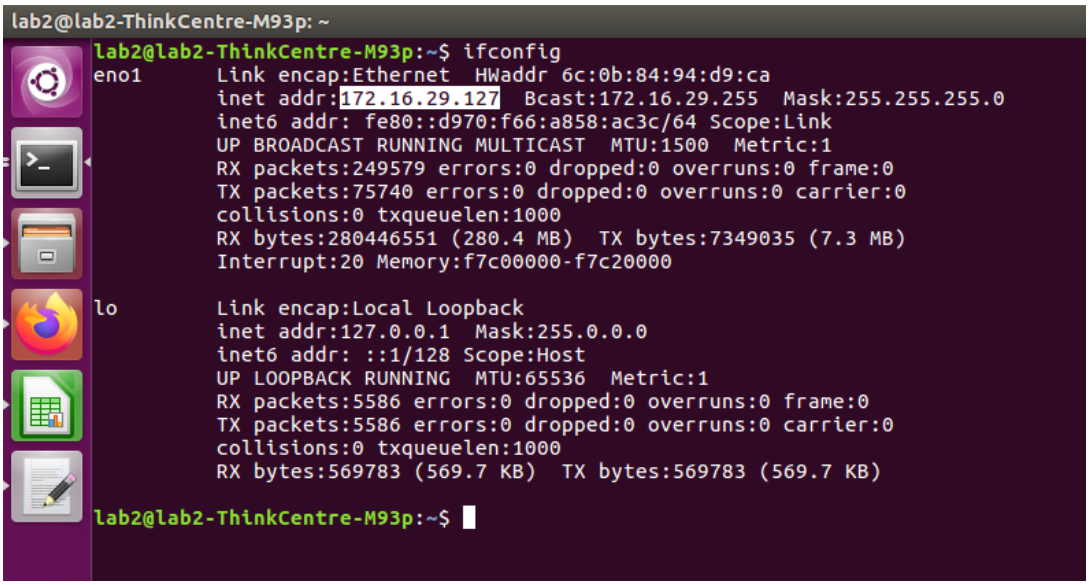
Username of a current login
Ex. ram.
Usually, it's parent directory is /home
Below, the user is "lab2"



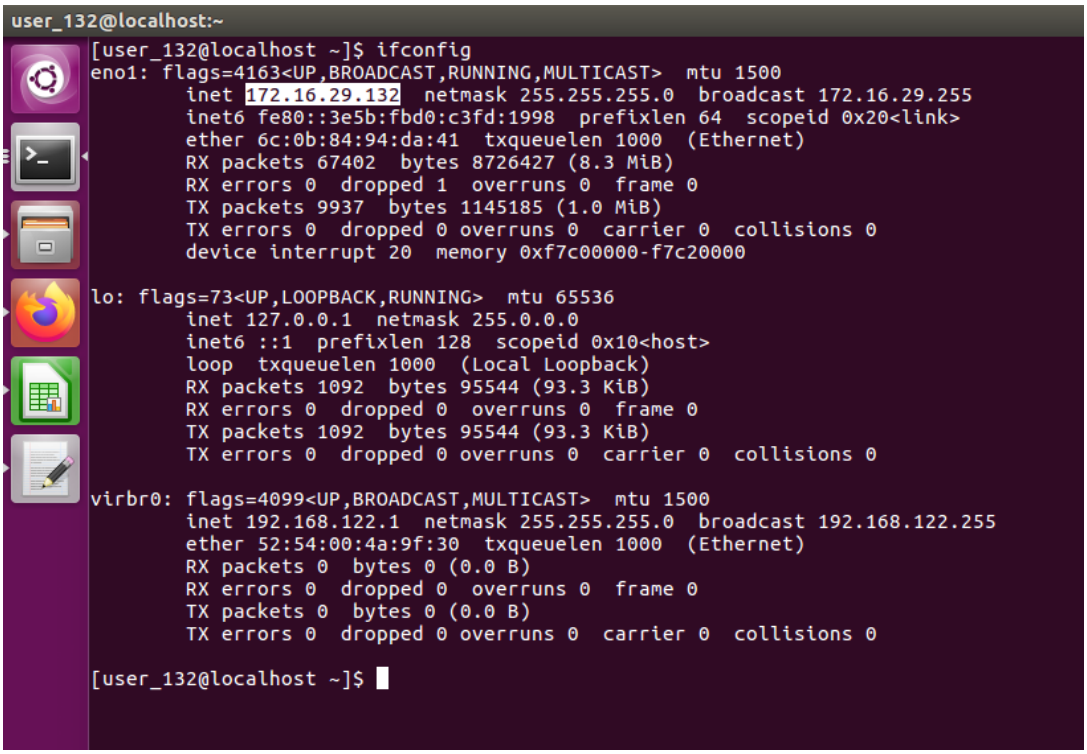
iii. IP Address

It is an Internet Protocol address, can be displayed using 'ifconfig' command

Ubuntu



CentOS

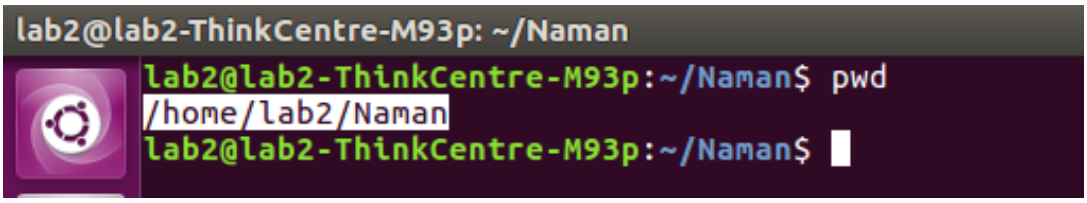


iv. Password

A password for the child node system
Ex. lab2_123

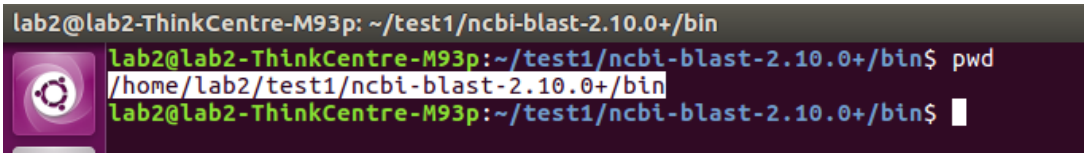
v. Process path

A path or location where the process will run
Ex: /home/lab2/Naman



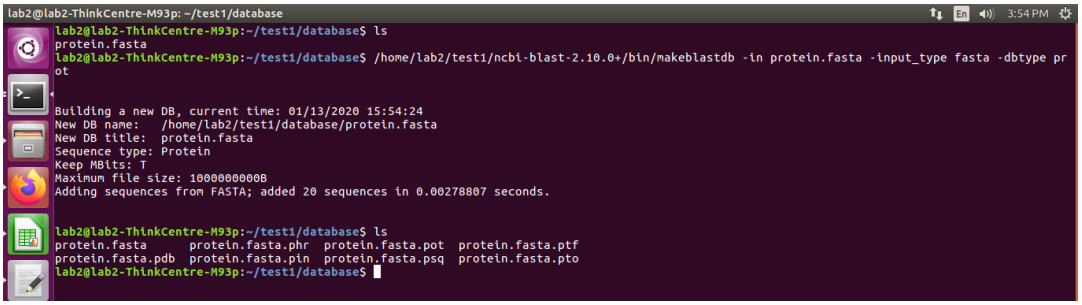
vi. Blast path

A path where blast+ executable is located
Ex: /home/lab2/test1/ncbi-blast-2.10.0+/bin

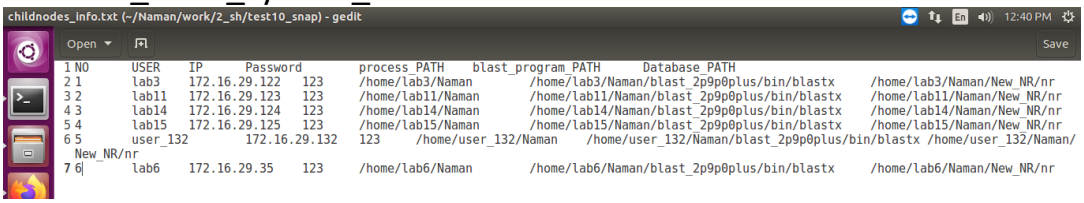


vii. Database path

A path where protein database is located
Ex: /home/lab2/test1/database



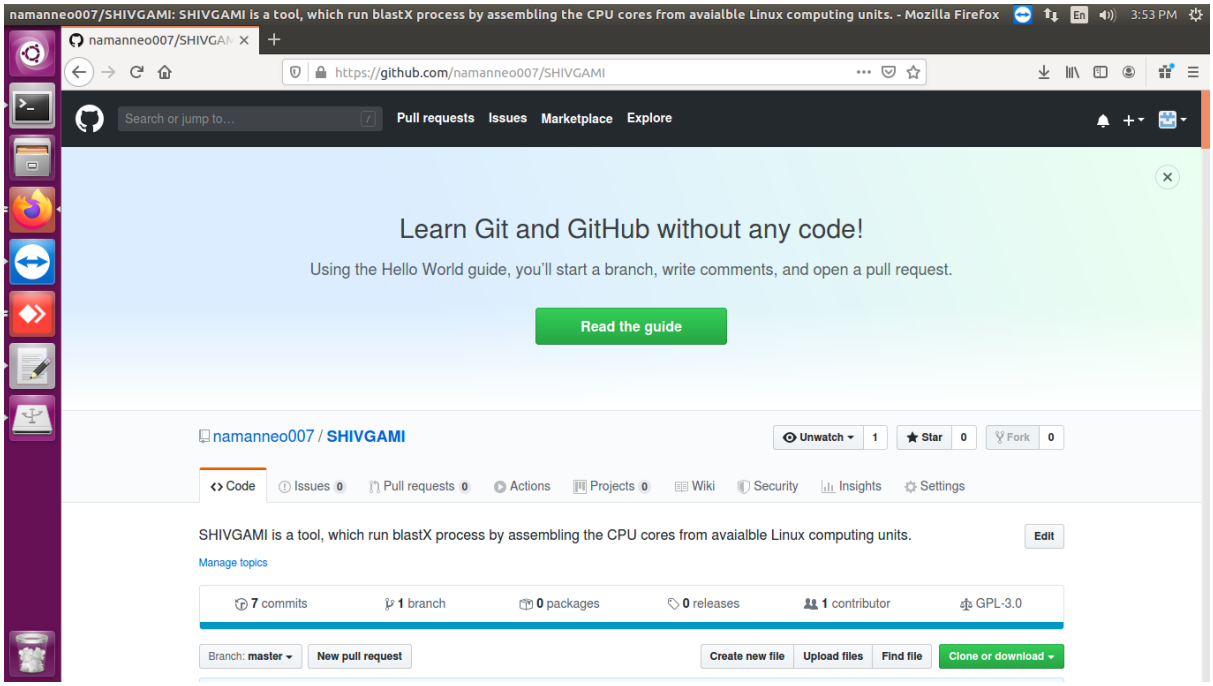
The child_node_system_info.txt will look like:

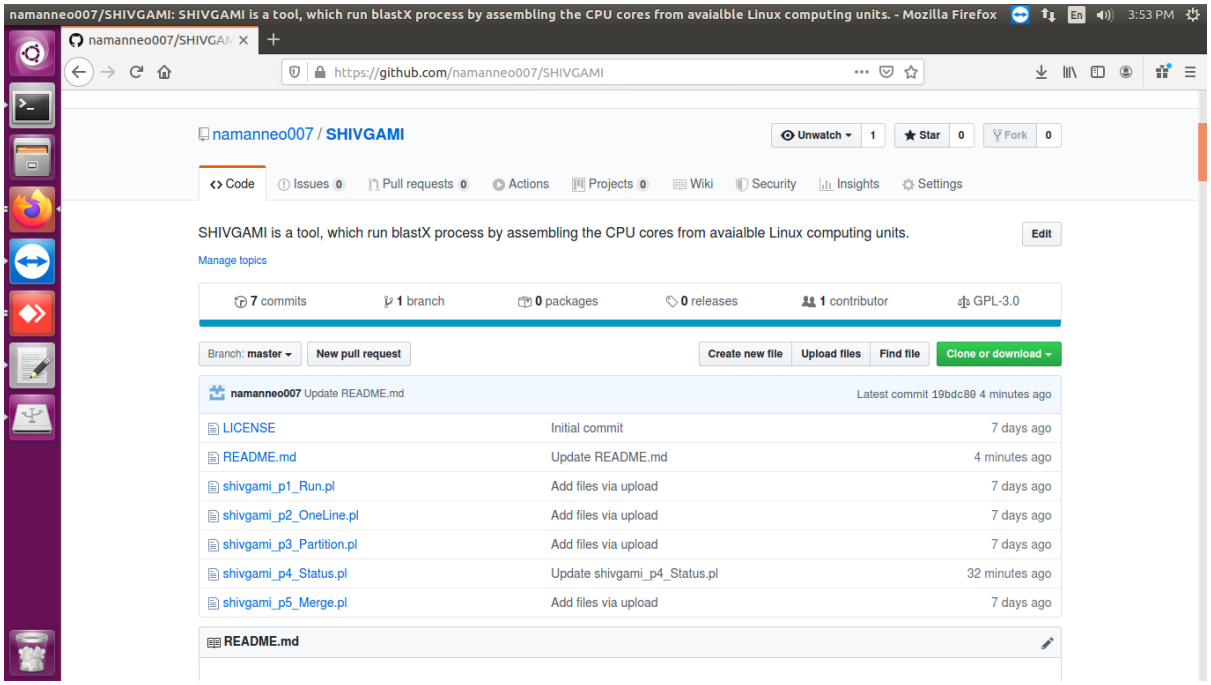


3. Download

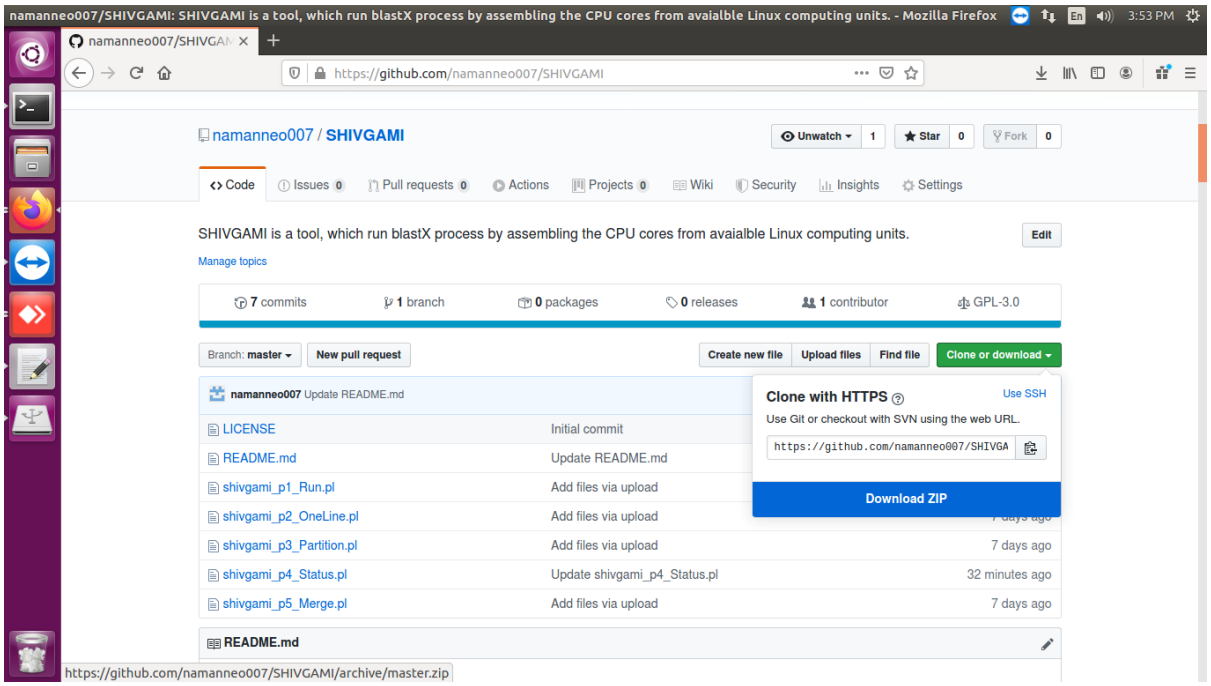
User can download SHIVGAMI from website www.shivgami.net and Github.
Here, we are giving illustration for github

Open a link : https://github.com/namanneo007/SHIVGAMI

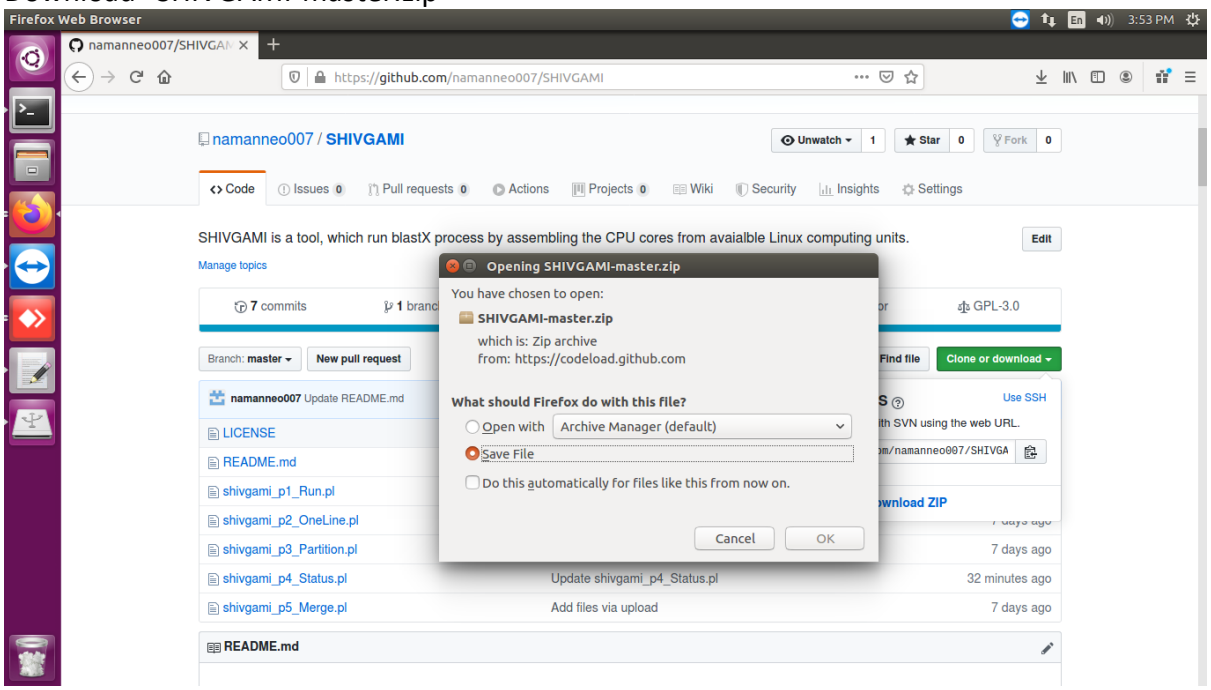




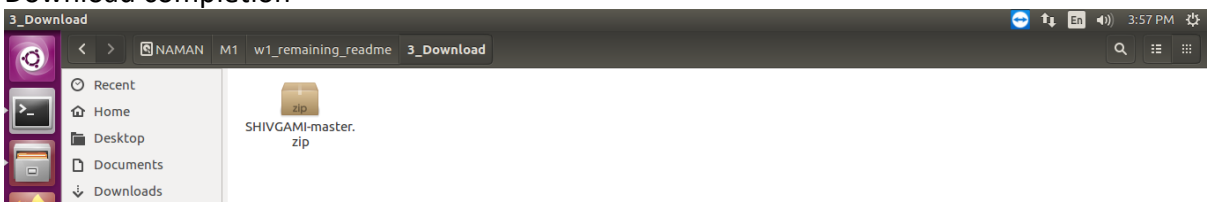
Click on "Clone or download"



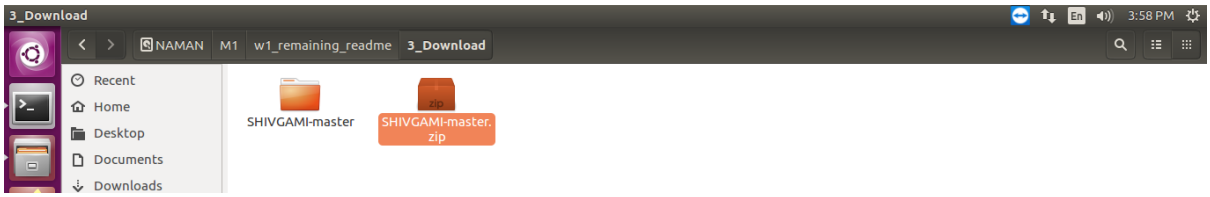
Download "SHIVGAMI-master.zip"



Download completion



Extraction



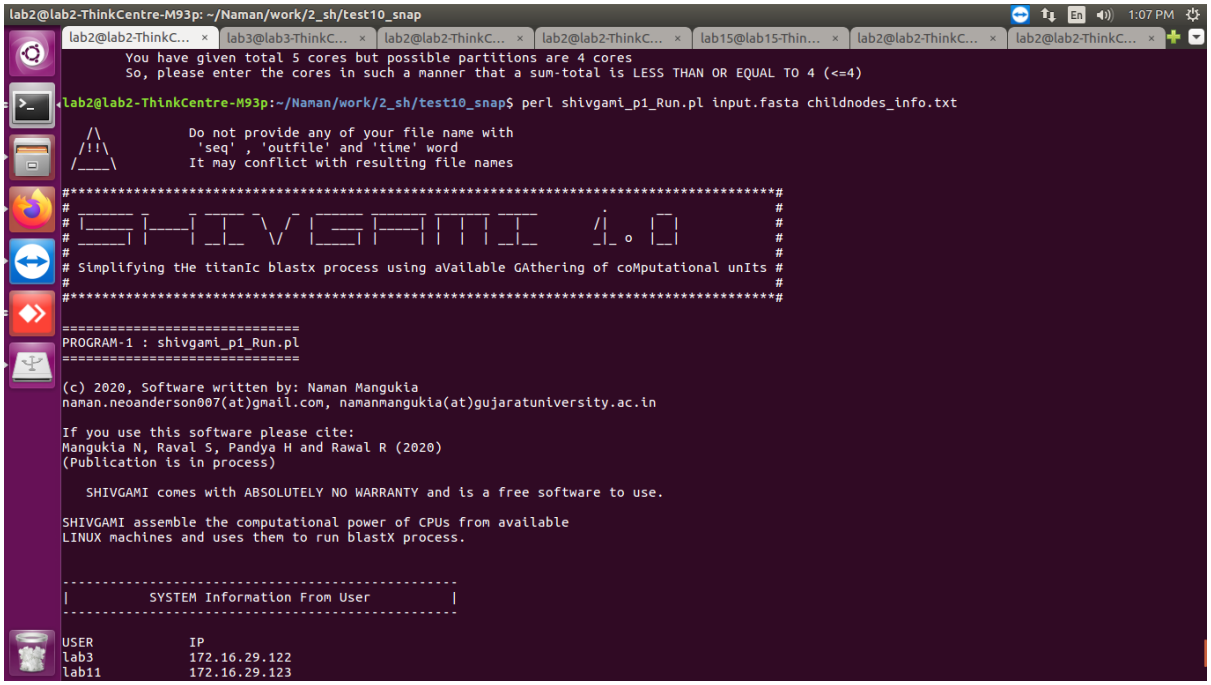
Exploring SHIVGAMI folder

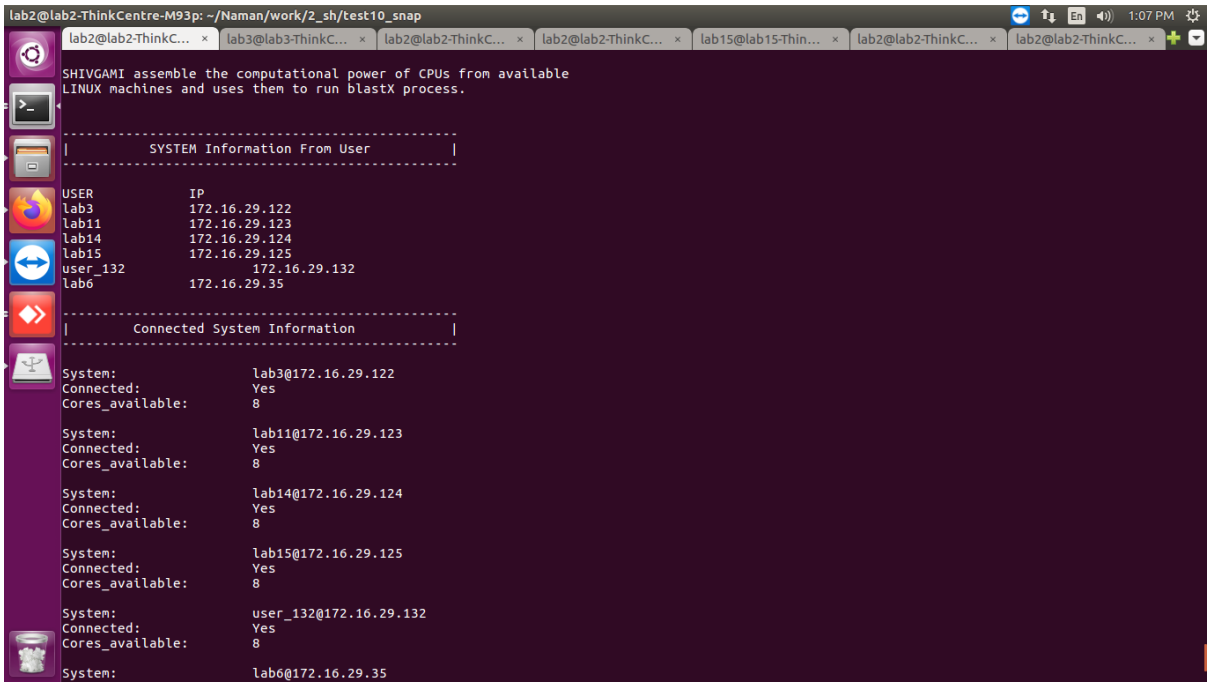
steps:

Put all these programs in a folder of a Master-node, from where you want to initiate the process.

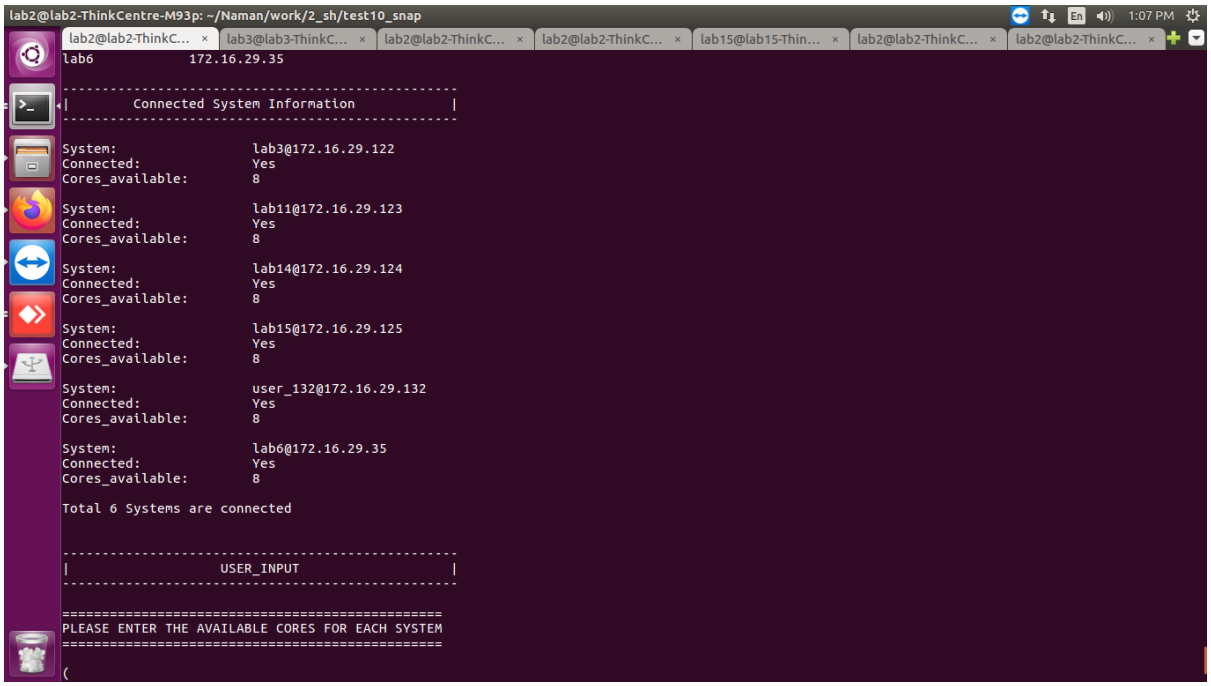
4. Run

Initiate SHIVGAMI by running program-1:shivgami_p1_Run.pl and providing 2 arguments <input-fasta-file> and <child-nodes-information> respectively. Here, we have taken a sample input fasta file comprising 6 nucleotide sequences to search against NR database (Alias created-05/15/2019 23:58:27, NSEQ 203058027, LENGTH 73894506806) using blastx program (ncbi-blast-2.9.0+).

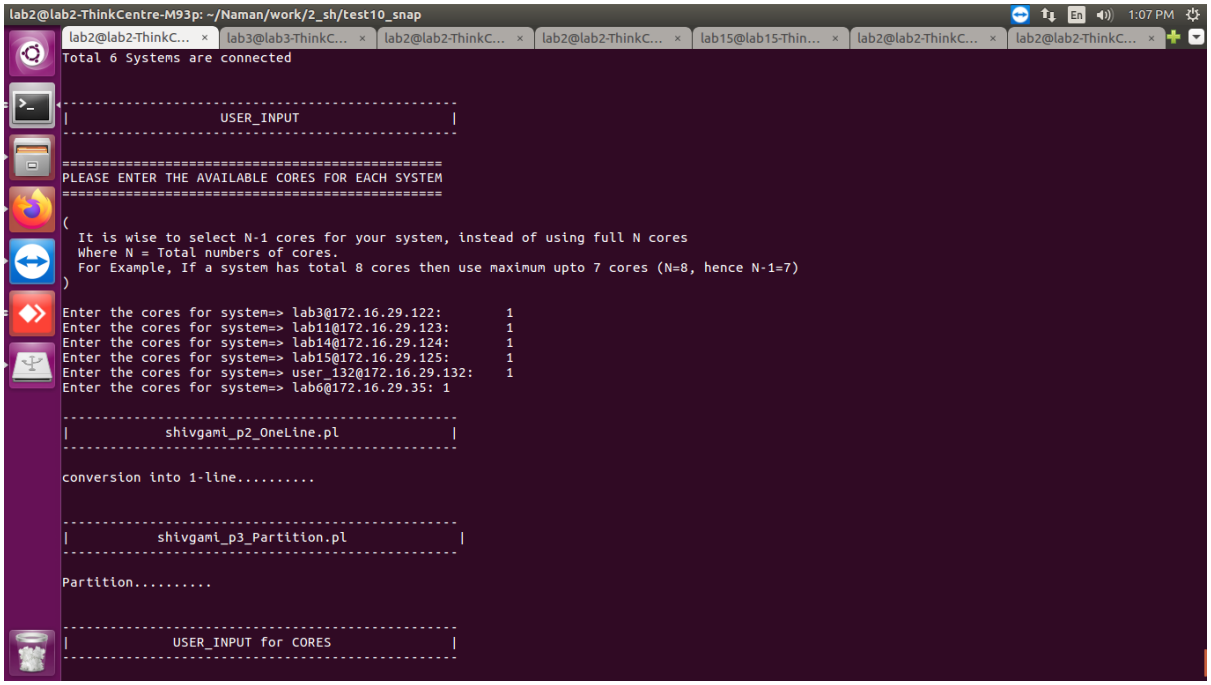




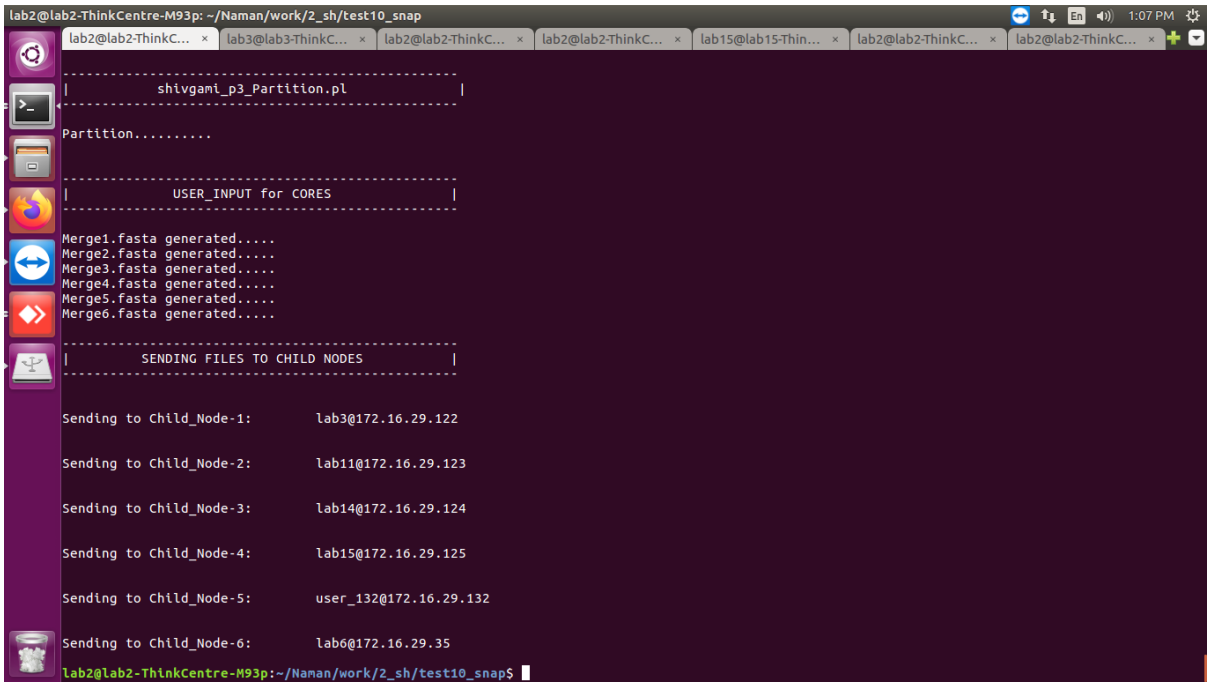
Then program shows the TOTAL NUMBER of CORES comprised by the connected system:



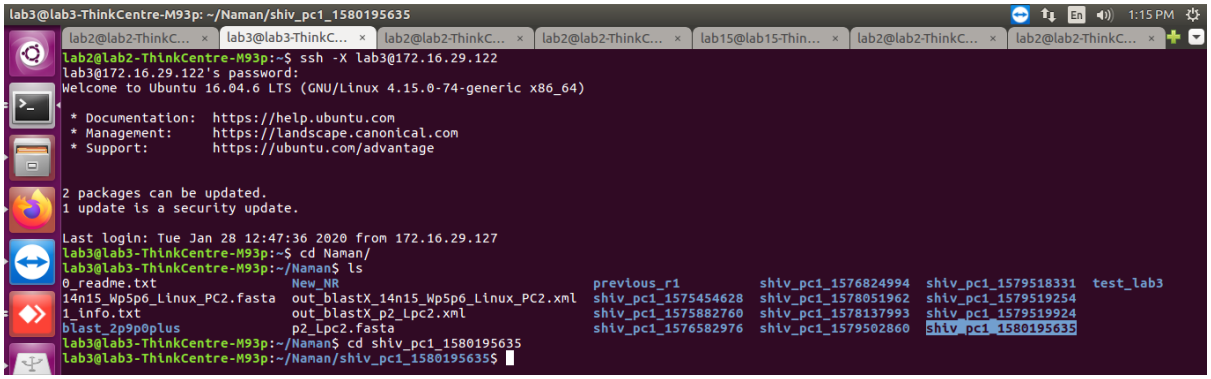
Now, program asks for USER-INPUT, Like how many cores user want to use. Here, in example, 1 core is used for all 6 systems.

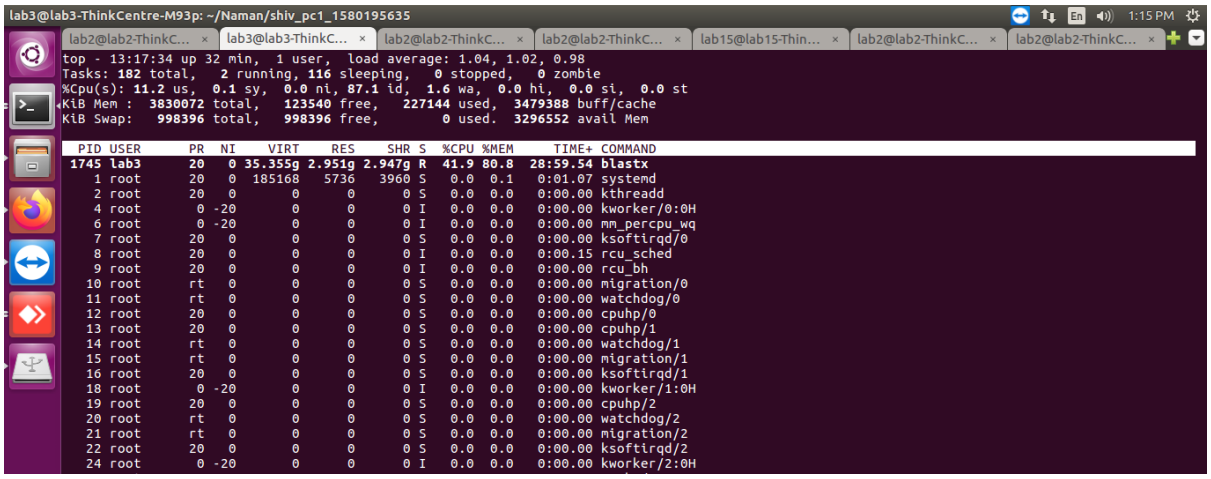


Then, program send partitioned sequences to the respective child-nodes.

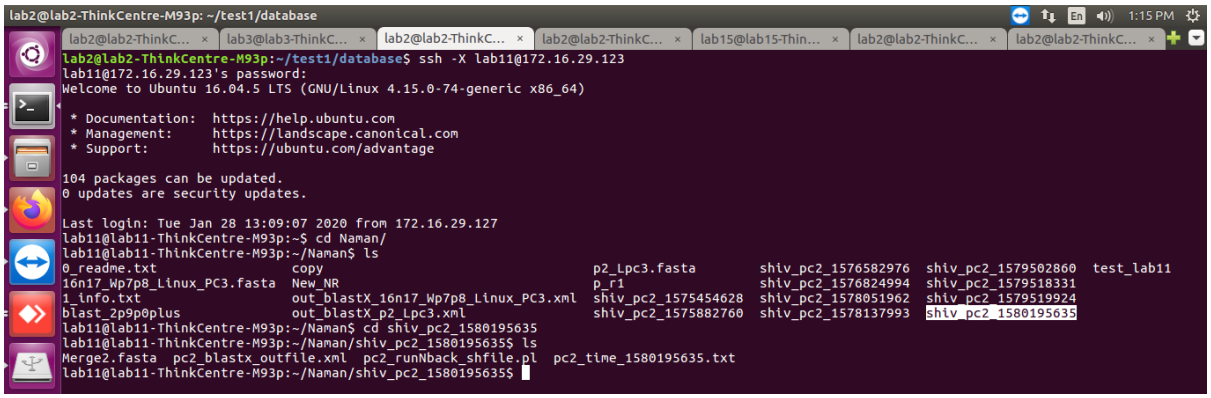


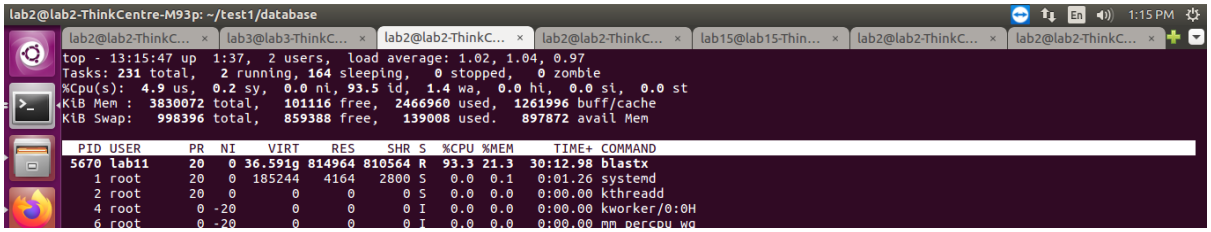
Checking the blastx on PC-1:



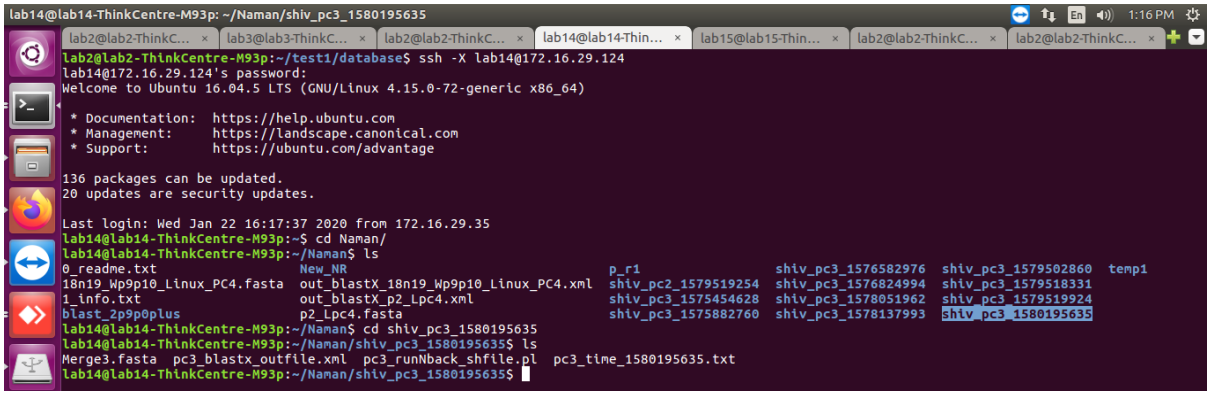


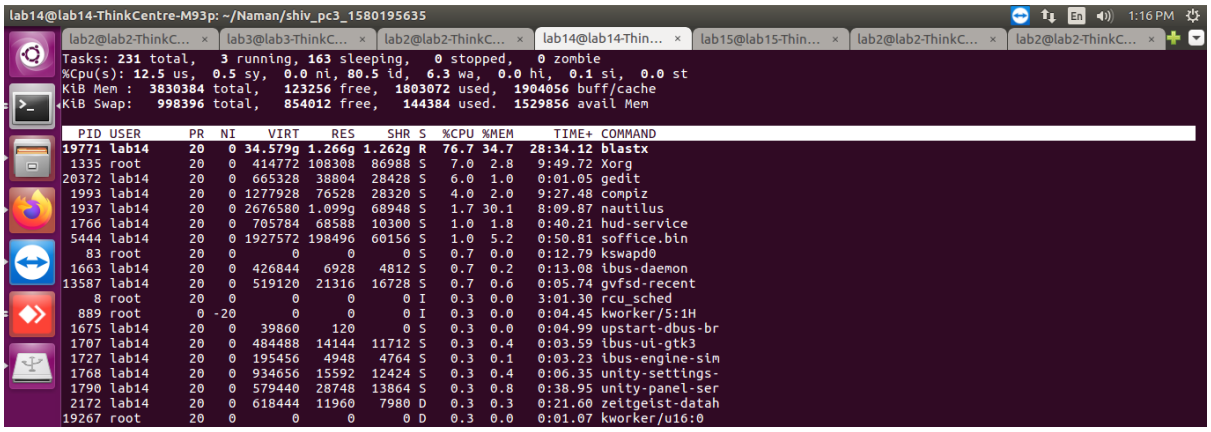
Checking the blastx on PC-2:



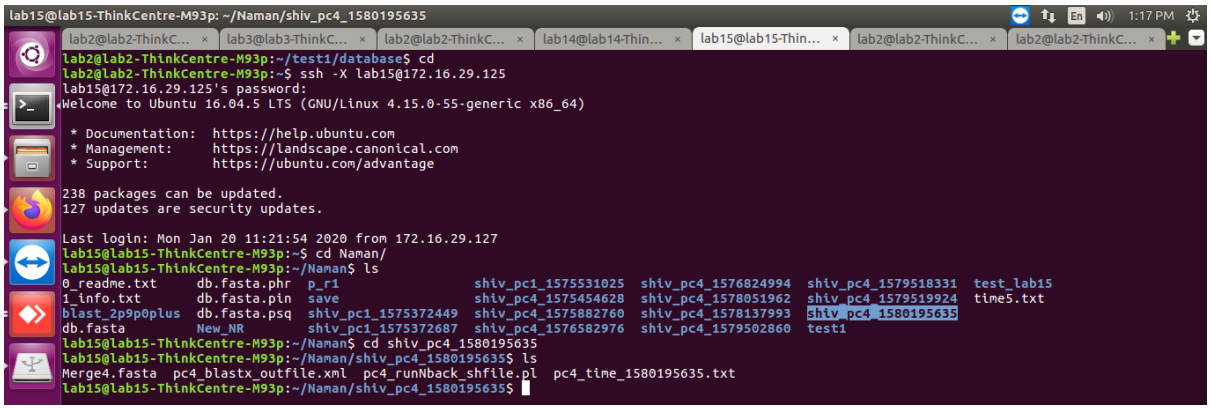


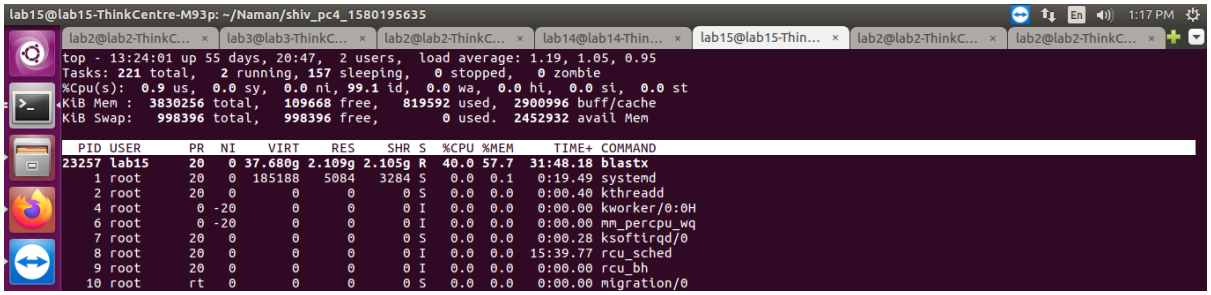
Checking the blastx on PC-3:



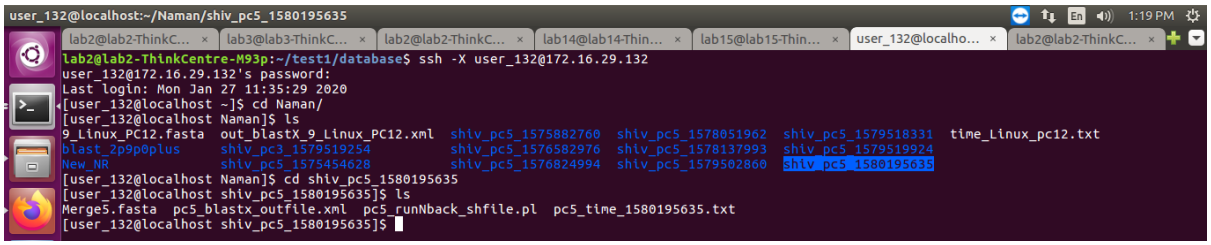


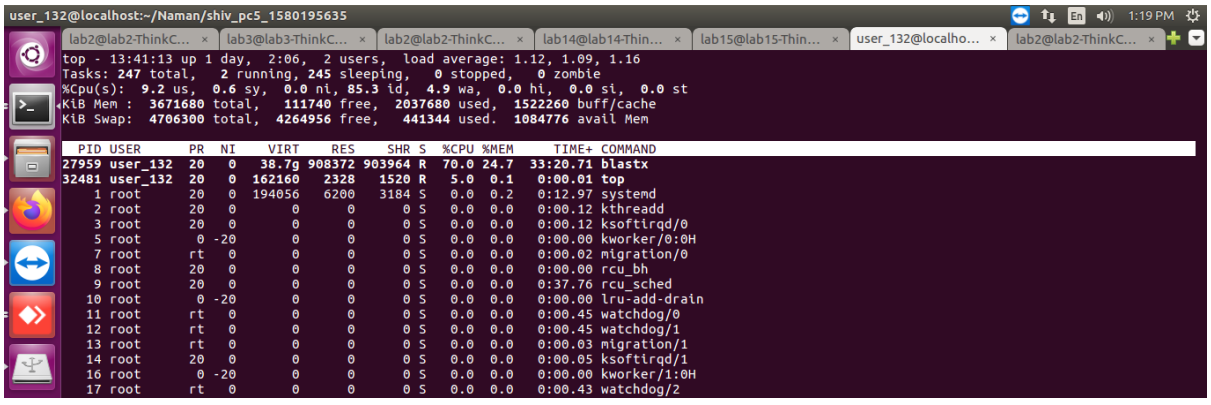
Checking the blastx on PC-4:



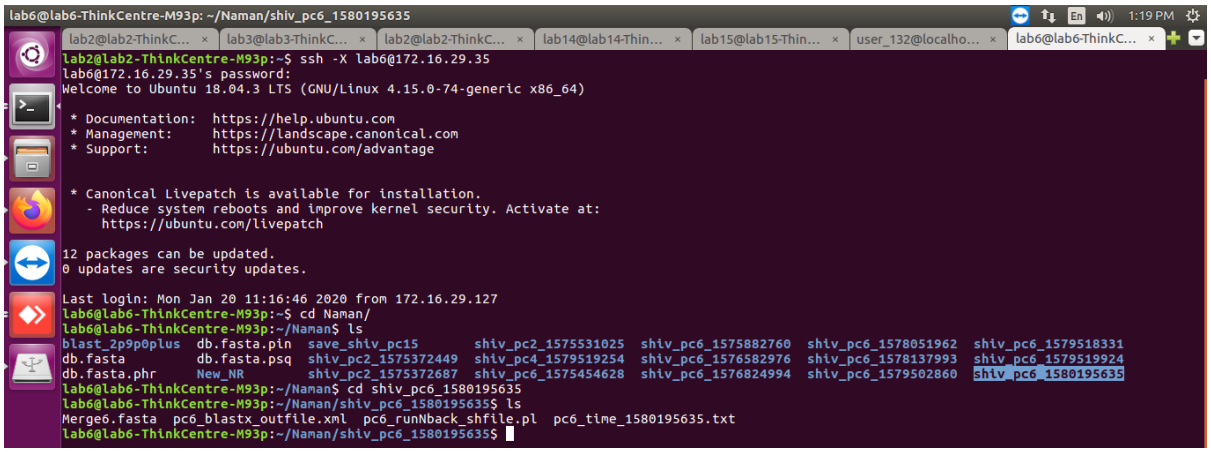


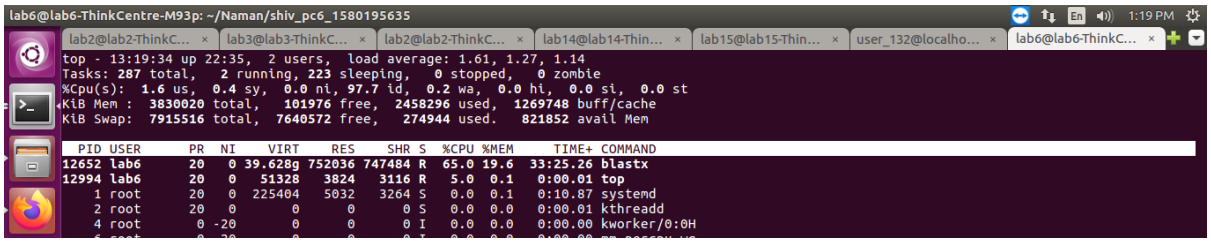
Checking the blastx on PC-5:





Checking the blastx on PC-6:

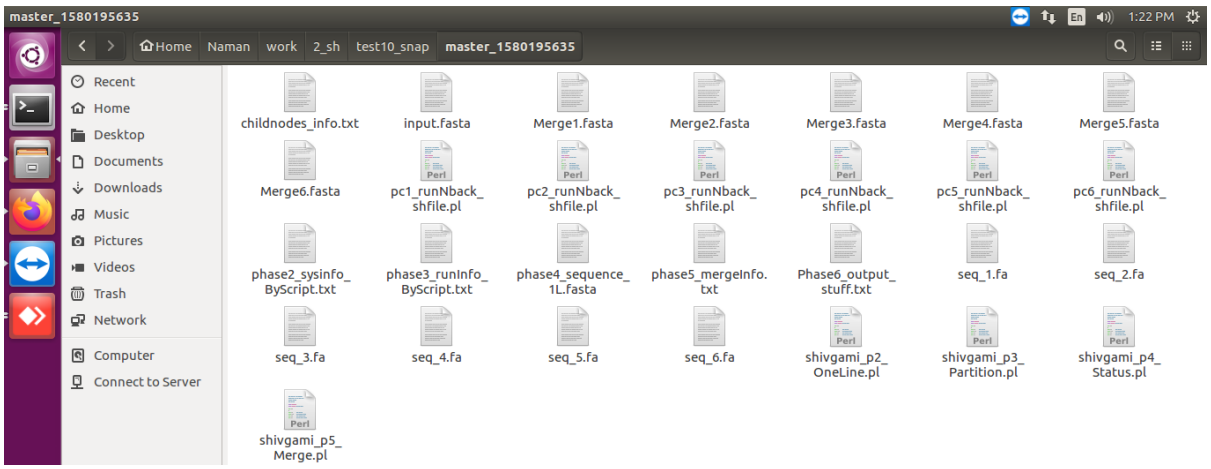




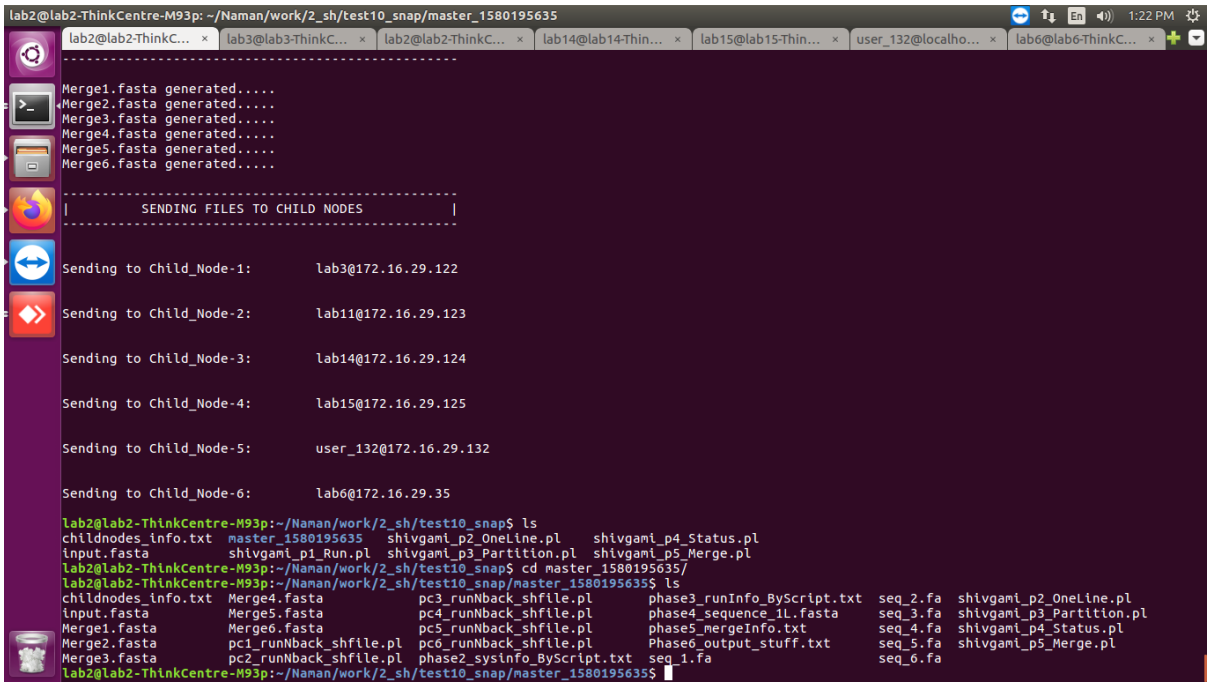
The SHIVGAMI process generate a result folder named master_xxxxxxxxxx, where xxxxxxxxxx=digits created by time function to generate each time a unique result folder. In this example the result folder is: master_1580195635.

This folder contains partitioned files (seqN.fasta, where N=1, 2, 3..), combined partitioned files for child nodes (mergeN.fasta, where N=1,2,3..), some temp files and other SHIVGAMI program files.

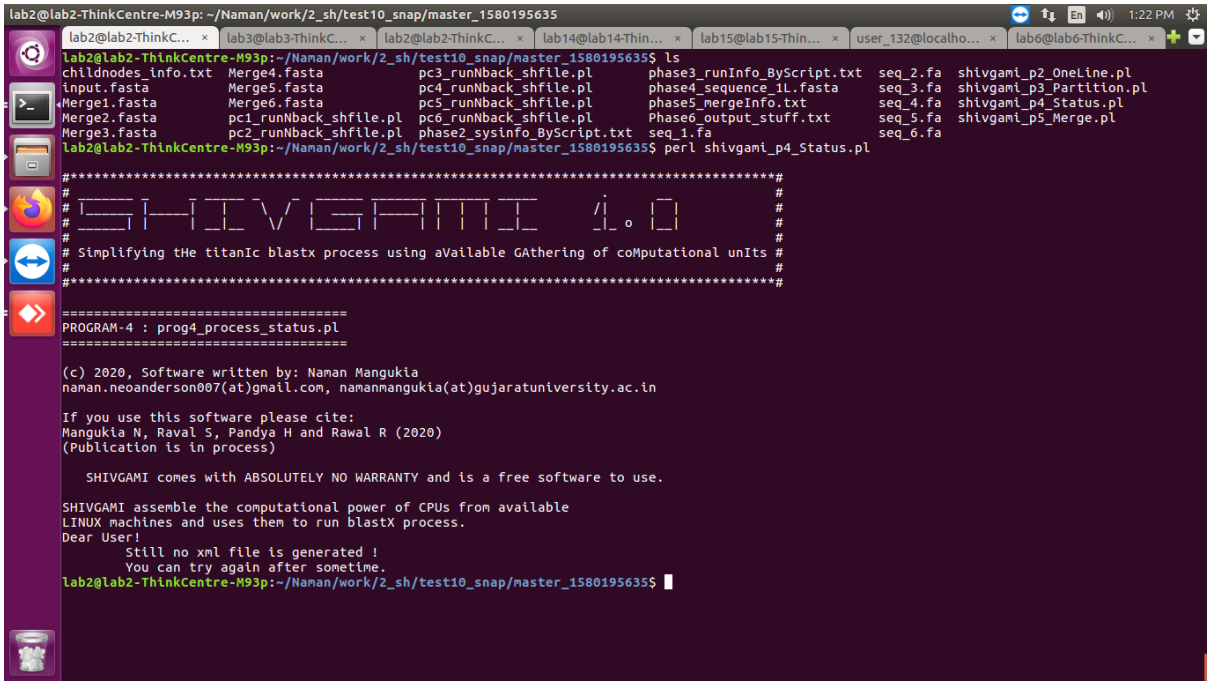
Folder view:



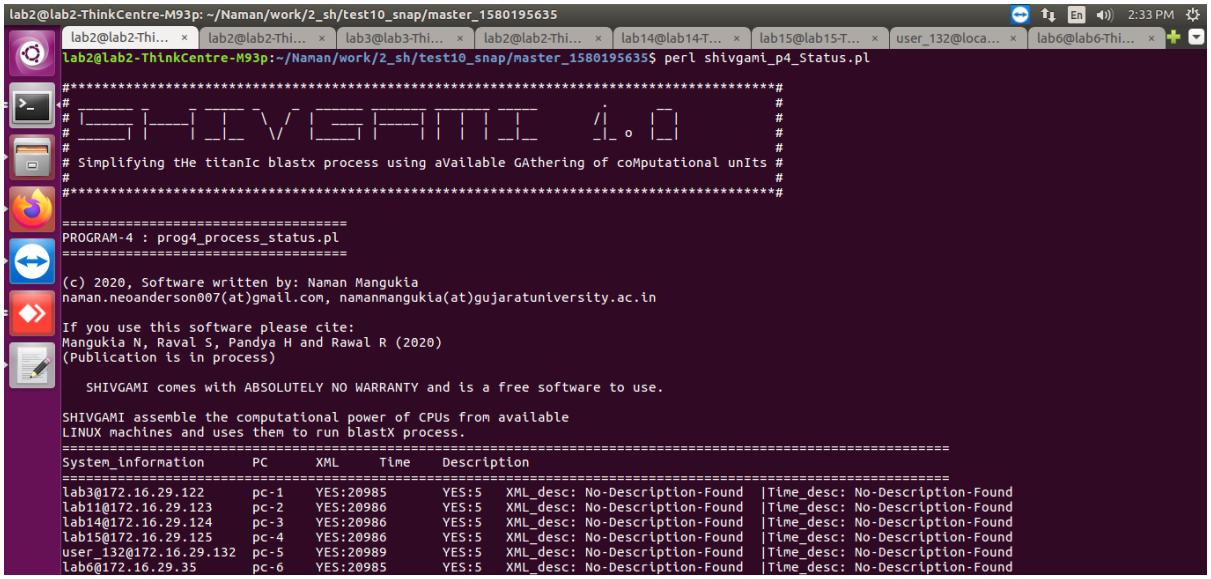
Command line view:

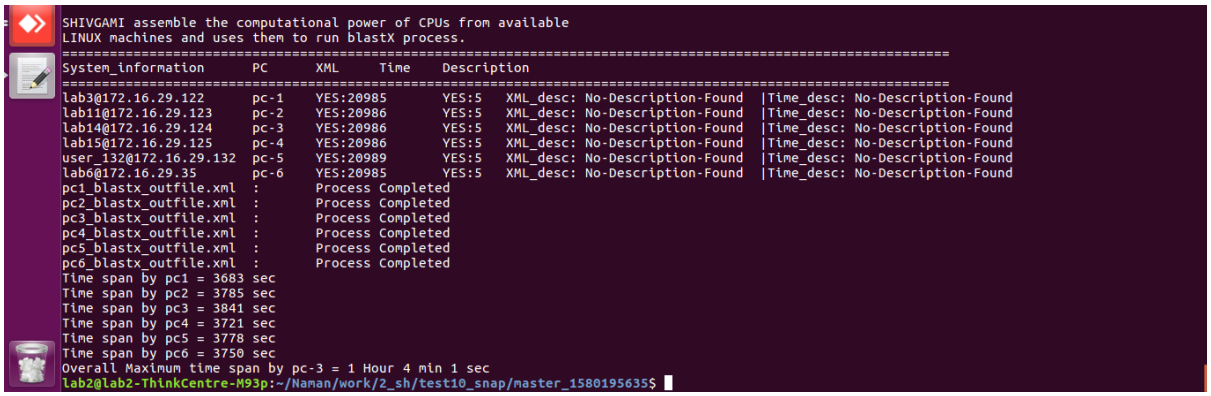


Now, Program-4 -> shivgami_p4_Status.pl will display the status of blastx, here we have shown the status just after the process initiated. As no xml file is generated, the program is not showing any status and telling to run it after sometime!



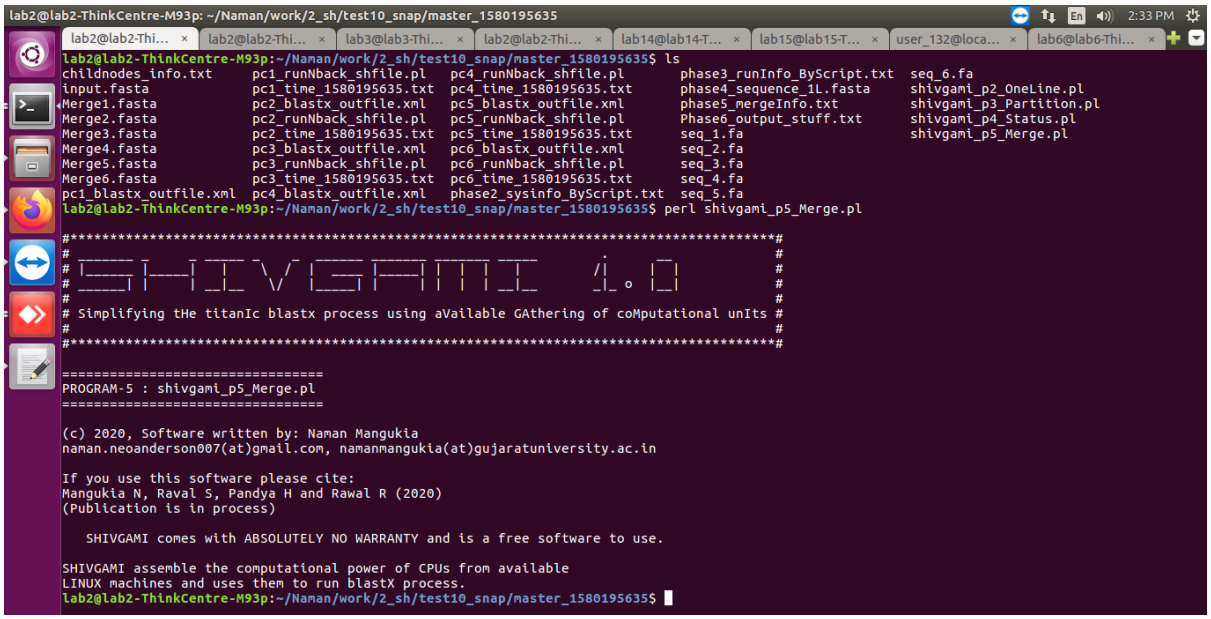
Once the process got completed for all child node, the Program-4 -> shivgami_p4_Status.pl will display the status along with the time consumed by all the child-nodes along with the child-node name which completed the blastx with maximum time span.





5. Merge

After, the completion of blastx process on all child-nodes, user should run Program-5 -> shivgami_p5_Merge.pl, which will merge all XML-result files from respective child nodes into a single xml file.



The file named as "Final_result.xml" is the merged file and can be used for the further downstream analysis.

From below illustration of result count, it can be seen that, the merged result file Final_result.xml - is having all 6 sequences' result.