Blast on the lani cluster

What do I need to connect to the cluster?

- Need to have SSH software installed
  - Windows: SSH 3.2.9 from ITS
    - http://www.hawaii.edu/askus/778
  - Mac: Built in for Mac OS X or newer
    - SSH: via terminal on the Mac
      - ssh <username>@<IP Address>
    - GUI for SFTP/SCP: fugu, filezilla
      - http://rsug.itd.umich.edu/software/fugu/
      - http://filezilla-project.org/
  - Linux: Normally pre-installed
    - openSSH via terminal
    - File transfer via scp or filezilla
- This guide will focus on Windows and SSH 3.2.9

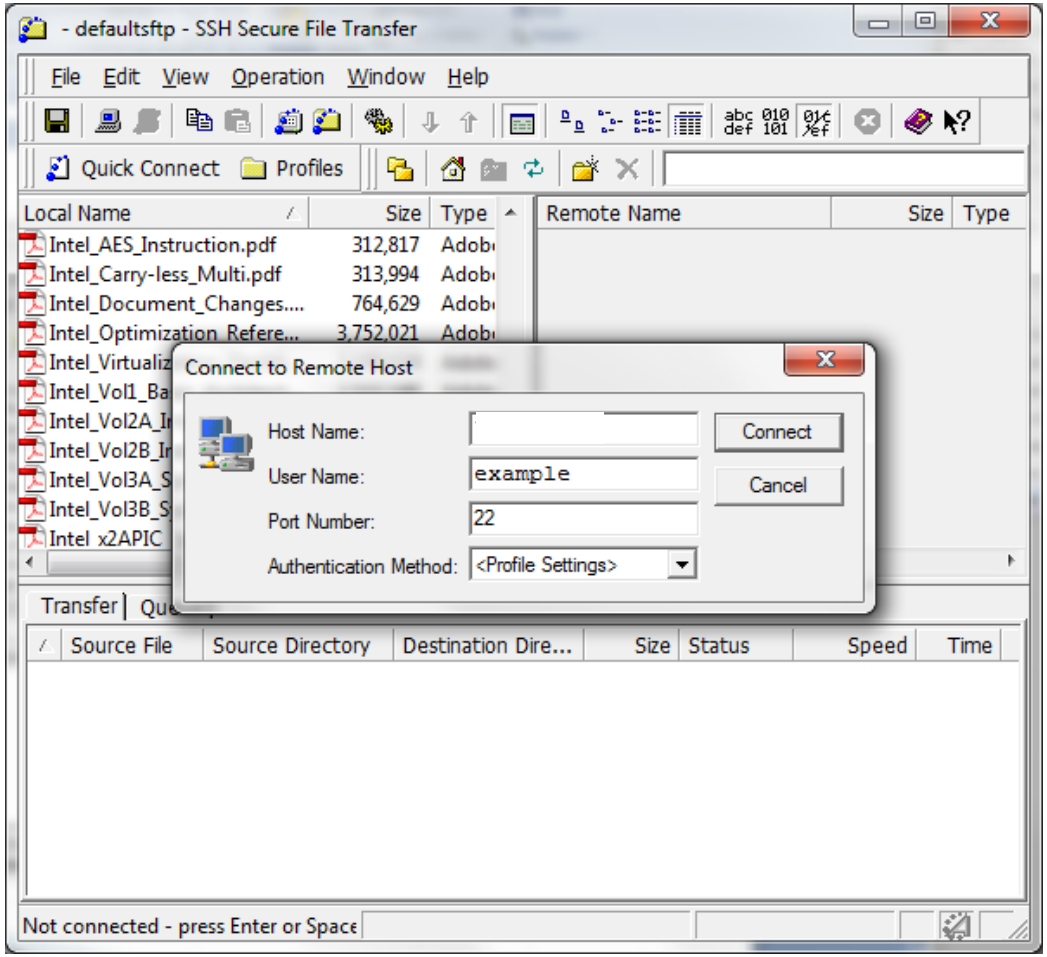
How do I connect to the cluster?

- Cluster is named lani
  - IP Address: <ip given during account creation process >
  - Consists of 55 machines for processing
  - Is shared amongst multiple users/labs

- Account information
  - Name: <Assigned upon request>
  - Password should be changed on first login using passwd
    - Password requirements:
      - 8 characters, using 3 out of 4 following groups
        - upper case, lower case, digits, symbols
      - Upper case as the first character or a digit as the last character does not fulfill the group requirements

How do I connect to the cluster?

Enter Lani's IP address and your lab user name and press connect



How do I upload data to process?

- Once you have connected using the SSH Secure File Transfer Client you are able to upload files.

- Uploading files is as simple as dragging files from your local computer onto to the SSH client under where it says “Remote name”

How do I submit a blast job?

- Blast jobs should be submitted using
  - biohadoop_job_submit
- Non-graphical question/answer process
  - Answer in the form of a number or plain text
- Terminology:
  - **Home directory**
    - The directory you are in first in when connecting to lani
  - Working directory
    - The directory your terminal is currently in.
- Example

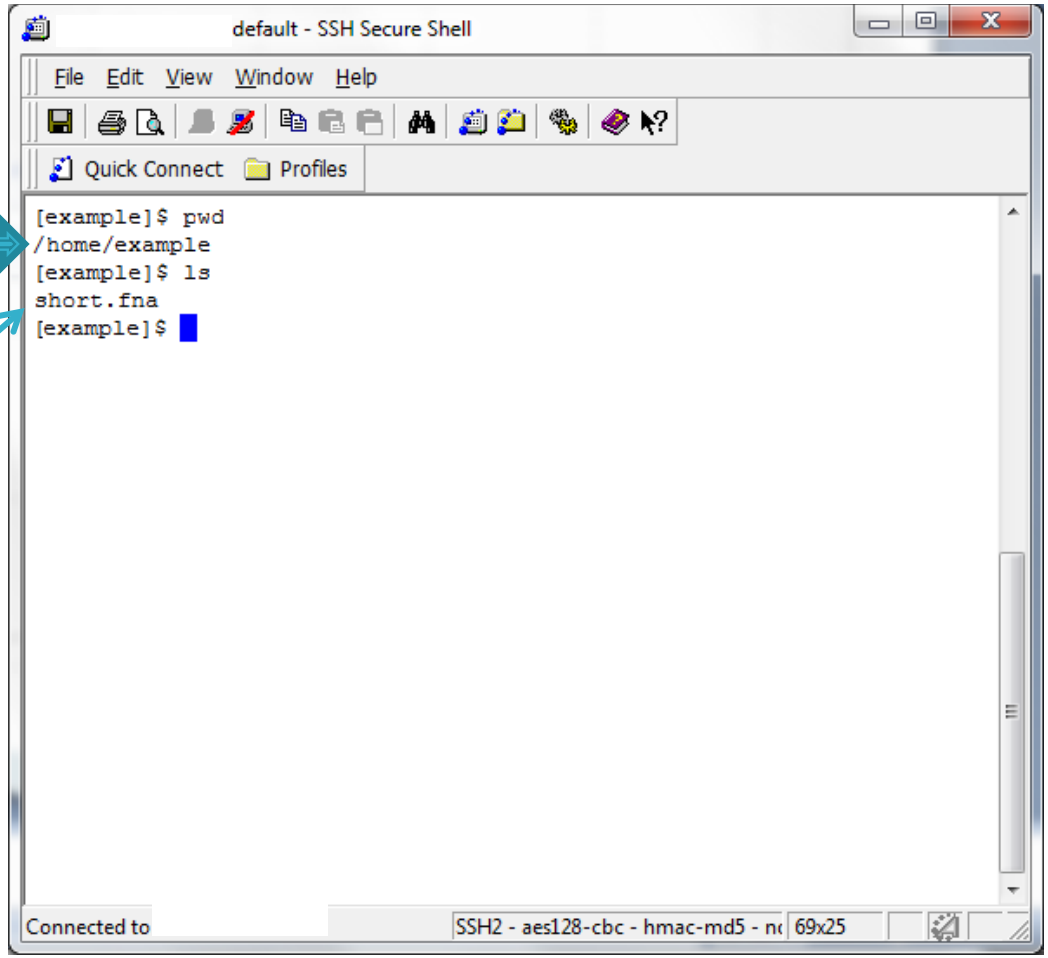
How do I submit a blast job?

- Terminal commands of interest
  - General commands
    - **cd** – Change Directory
    - **ls** – List Directory
    - **mkdir** – Make directory
    - **passwd** – Change password
    - **pwd** – Print the working directory
    - [Control/Command-c] **Ctrl** + **c** – Aborts a program running in the current terminal
  - Lani specific
    - **showq** – Shows currently queued and running jobs
      - Shows jobs for all users on Lani
    - **biohadoop_job_submit** – Create and submit a job
    - **biohadoop_job_status** – View known information about your jobs
    - **biohadoop_kill_job** – List your jobs and provides a way to cancel a job
    - **biohadoop_blastdb** – Create or remove custom blast databases

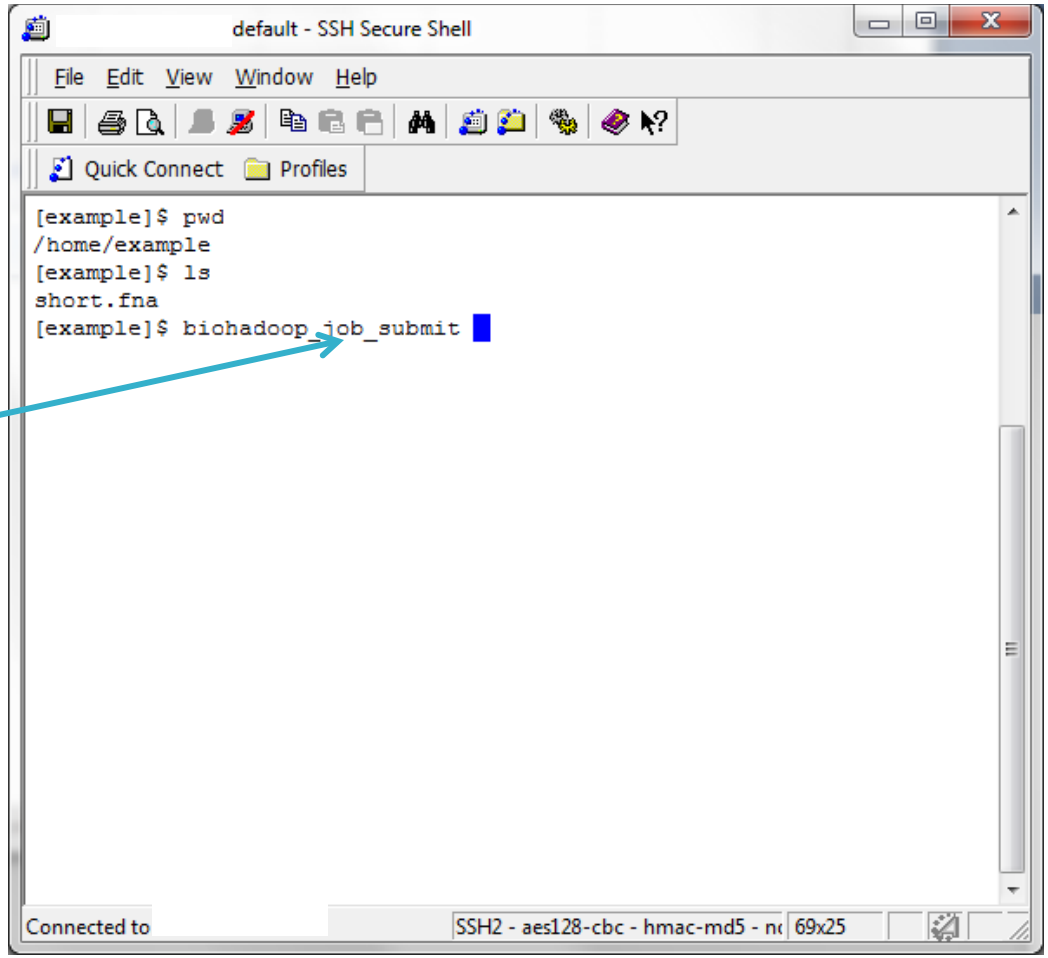
How do I submit a blast job?

Working directory

Fasta file in the working directory



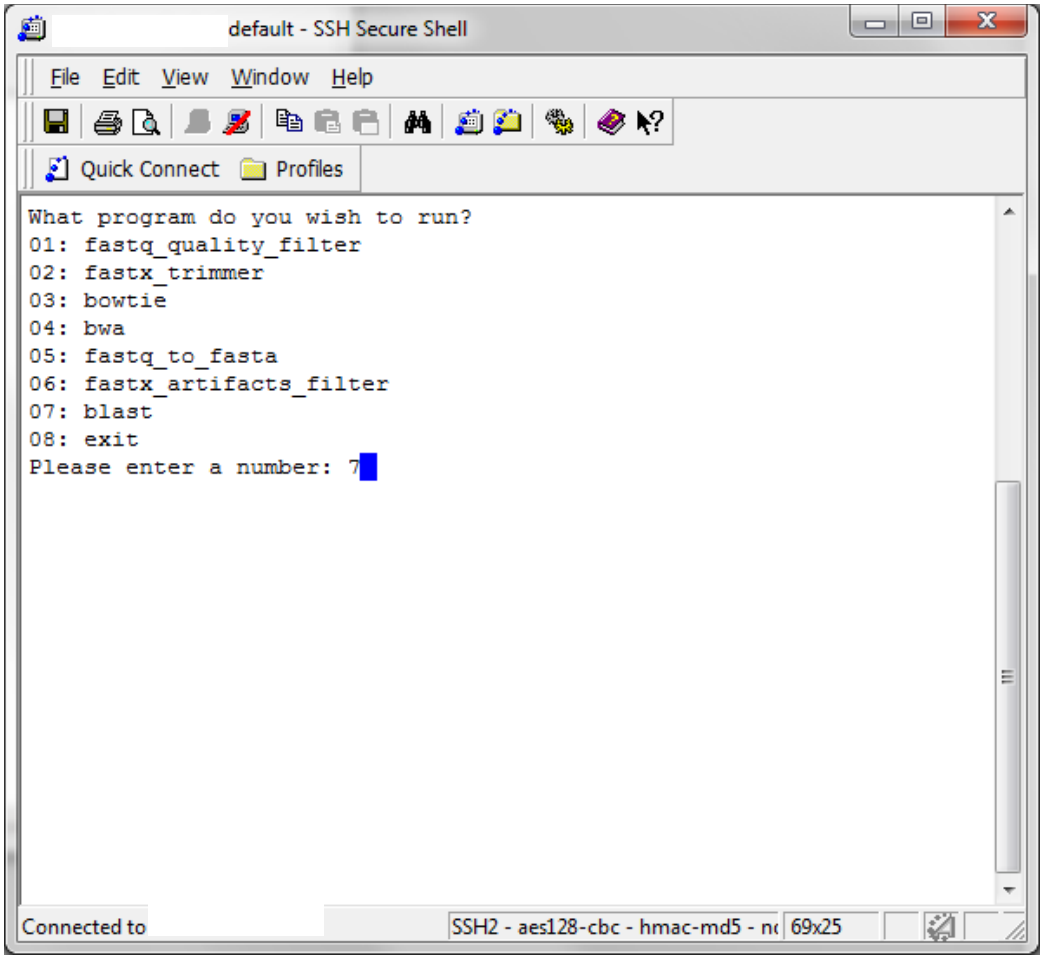
How do I submit a blast job?



Job submission command

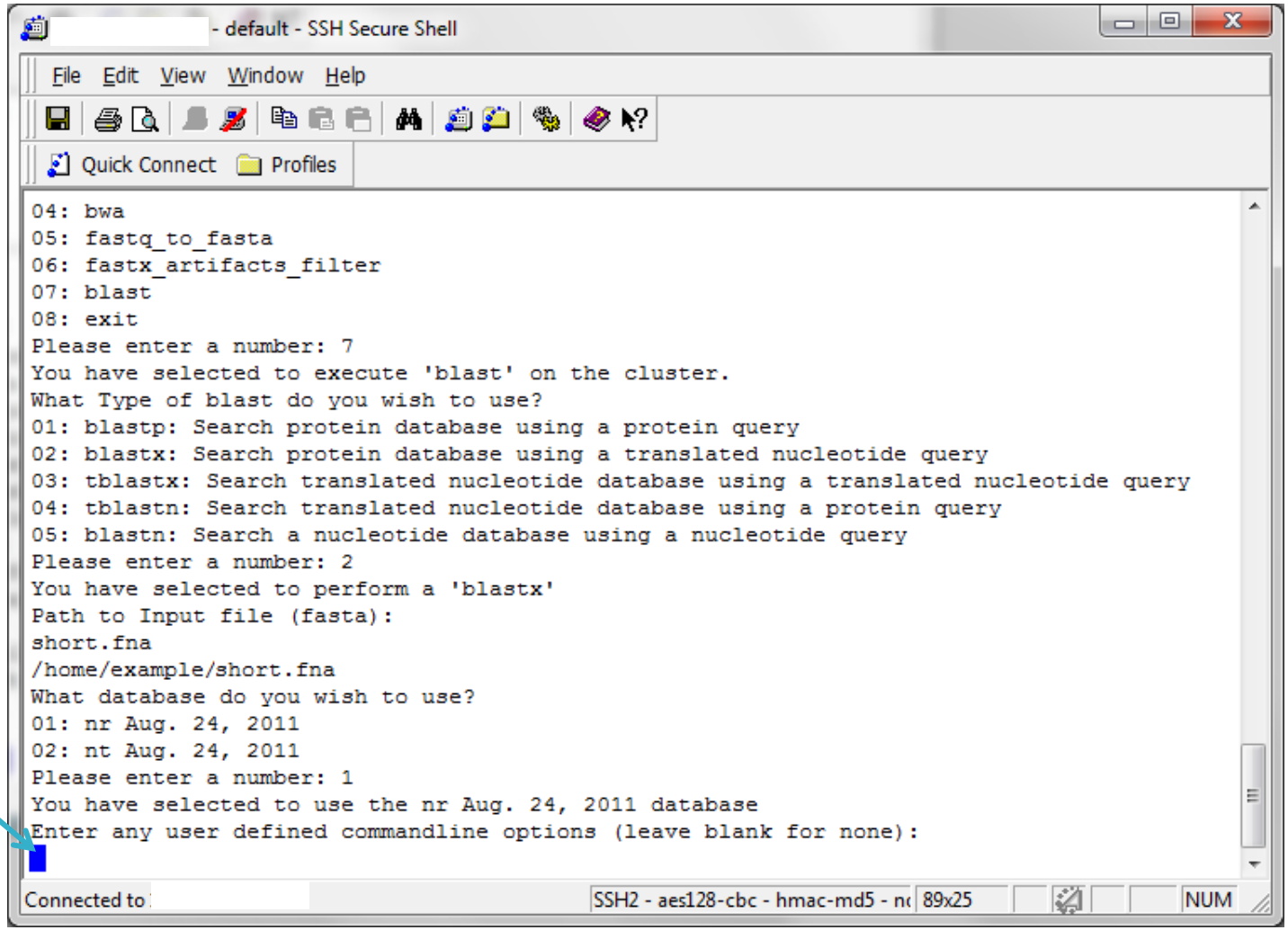
Execute this from the same directory as your input to simplify future steps

How do I submit a blast job?



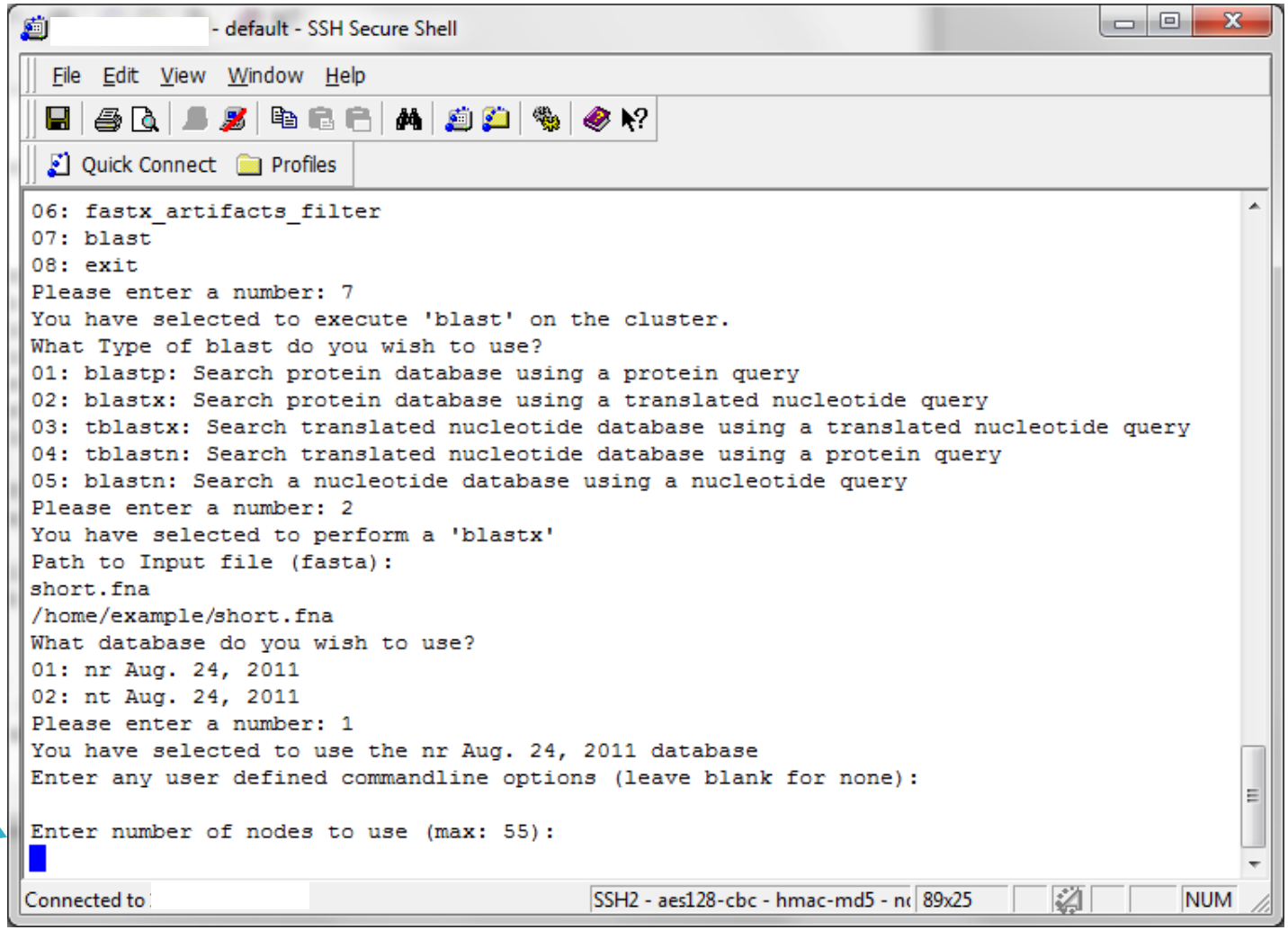
How do I submit a blast job?

Able to add additional options that are available to the application



How do I submit a blast job?

This depends
on how many records
you plan to blast



How do I submit a blast job?

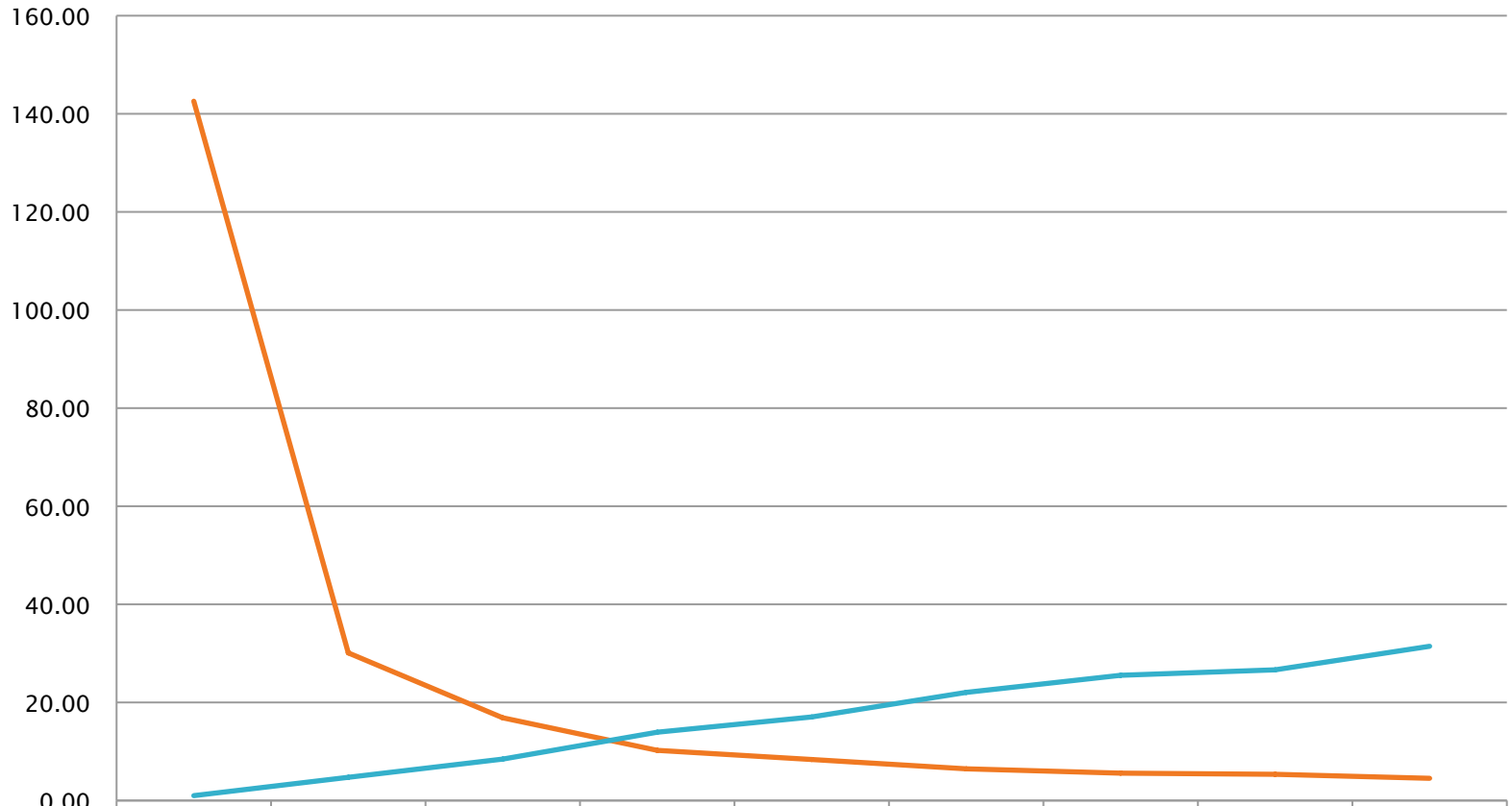
- How many nodes do I need for my job?
  - This depends largely on how big your input is.
- The cluster is designed for large datasets
  - Large is defined as at least 5,000 sequences
  - It will work on smaller datasets but may not be optimal
- The cluster is a shared resource
  - The amount of nodes requested should reflect this
  - Recommended max is about ½ the total nodes per job
    - More nodes maybe used if the job is short lived (1-2 days)

How do I submit a blast job?

- Runtime with a real input
  - Blastx
  - Database: nr
  - e-value cut off: 1e-05
  - Input Size: 247,586 sequences
  - Nodes: 35 nodes

- Results
  - Execution time: 2 days 11 hrs and 44 min
  - 69.08 sequences per minute
  - 1.97 sequences per minute per node

Runtime vs. # of nodes for blastx

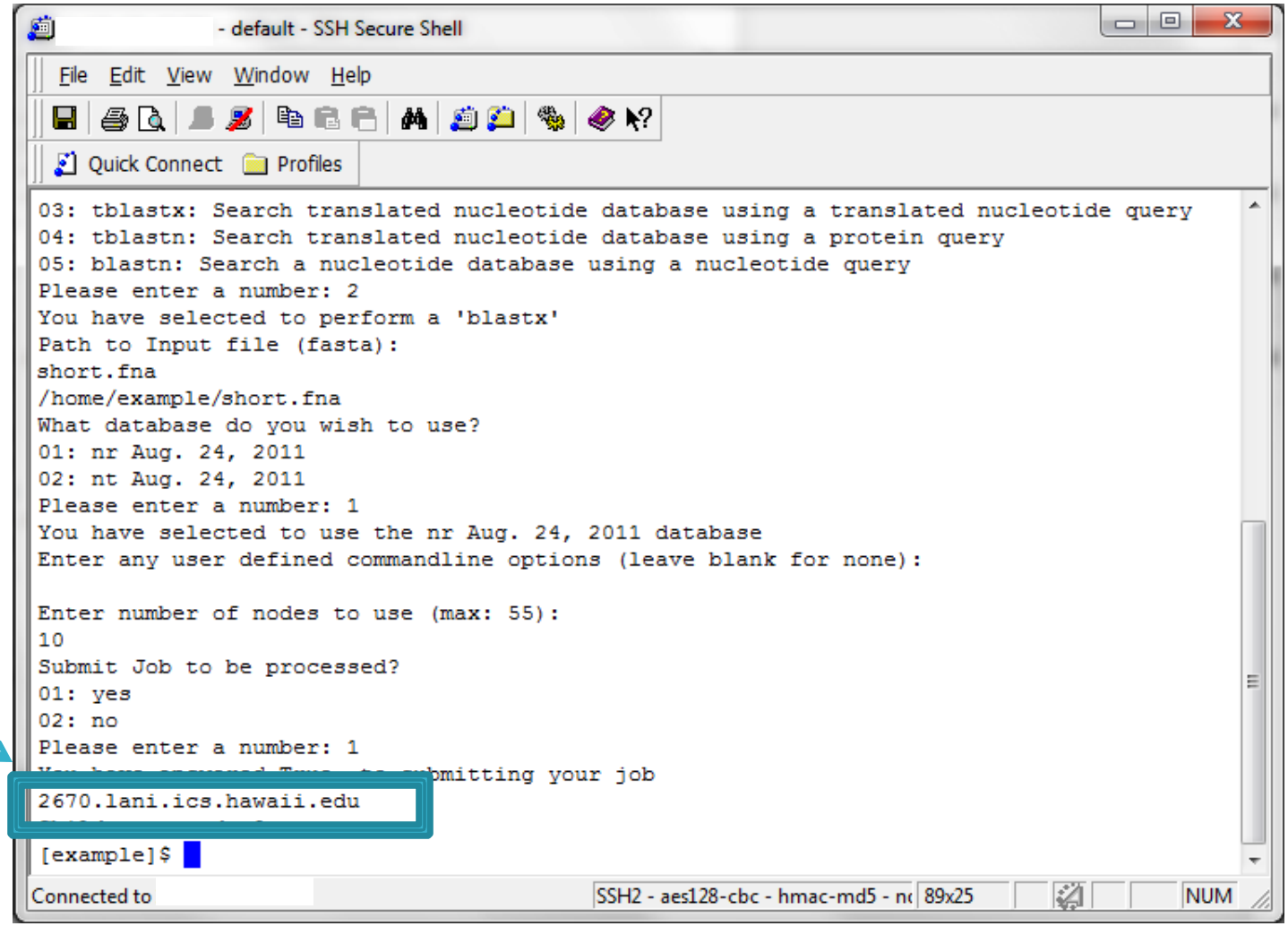


	1	5	10	15	20	25	30	35	40
Runtime (hrs)	142.57	30.10	16.87	10.23	8.37	6.47	5.58	5.35	4.53
Speed up	1.00	4.74	8.45	13.93	17.04	22.05	25.53	26.65	31.45

Blastx vs. nr , e-value= 1e-05
using 4318 sequences as input

How do I submit a blast job?

Copy this line down.
This is your Job ID



What happens after I submit?

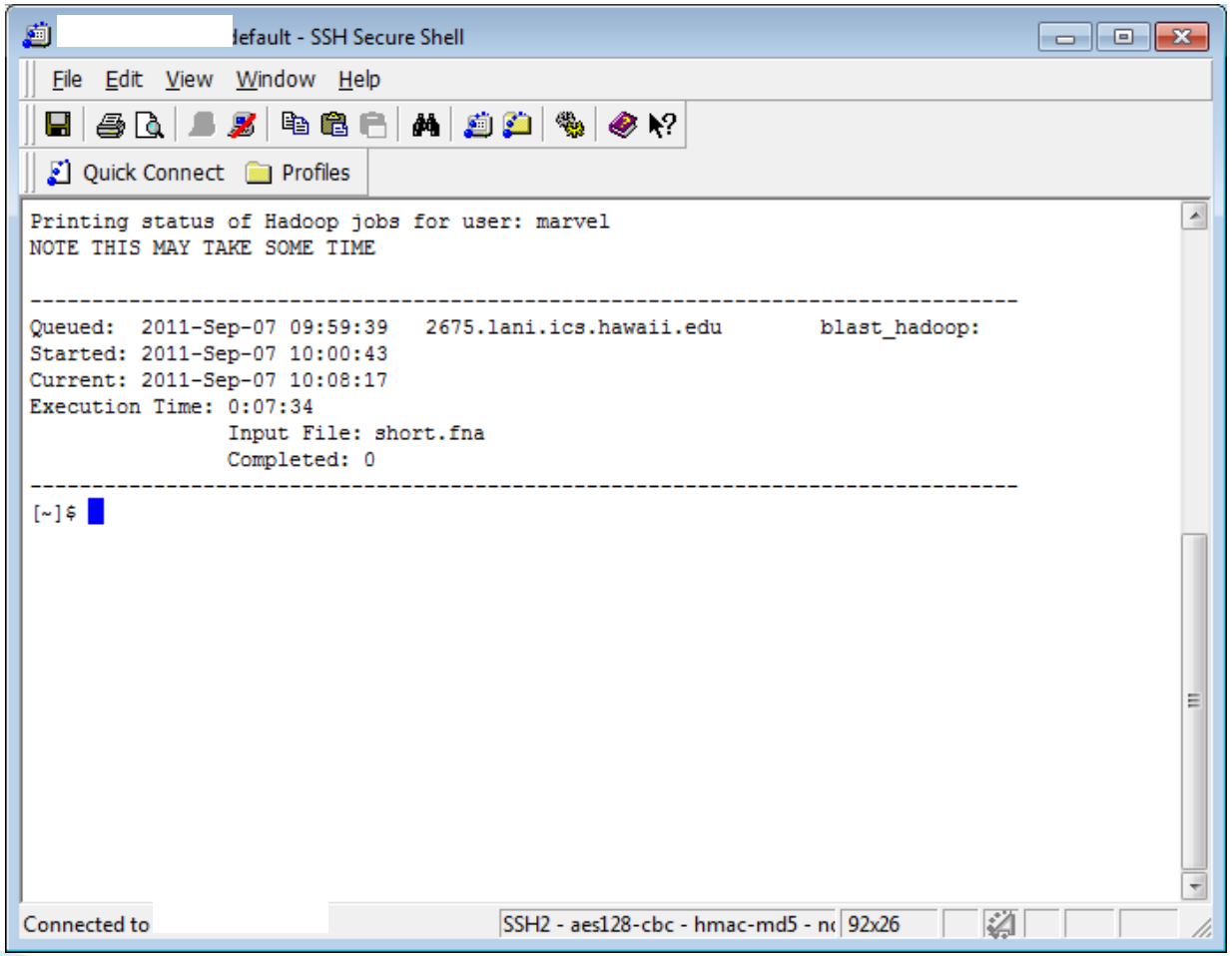
- Two files are created in the working directory
  - *.submit – Specifies the work flow of the job
  - *.yaml – Information/parameters specific to the job
  - Both files are moved into the job folder once the job becomes active

- A new folder is made in the working directory when the job starts to run
  - The new folder is named after your Job ID (known as the job folder)
    - Contains important job information during execution
    - Contains the results upon the completion of the job
      - *.pbsout – Information returned about the job
      - error – The error output from the job
      - output – The output or results from the job
      - input – Contains the name of the original input

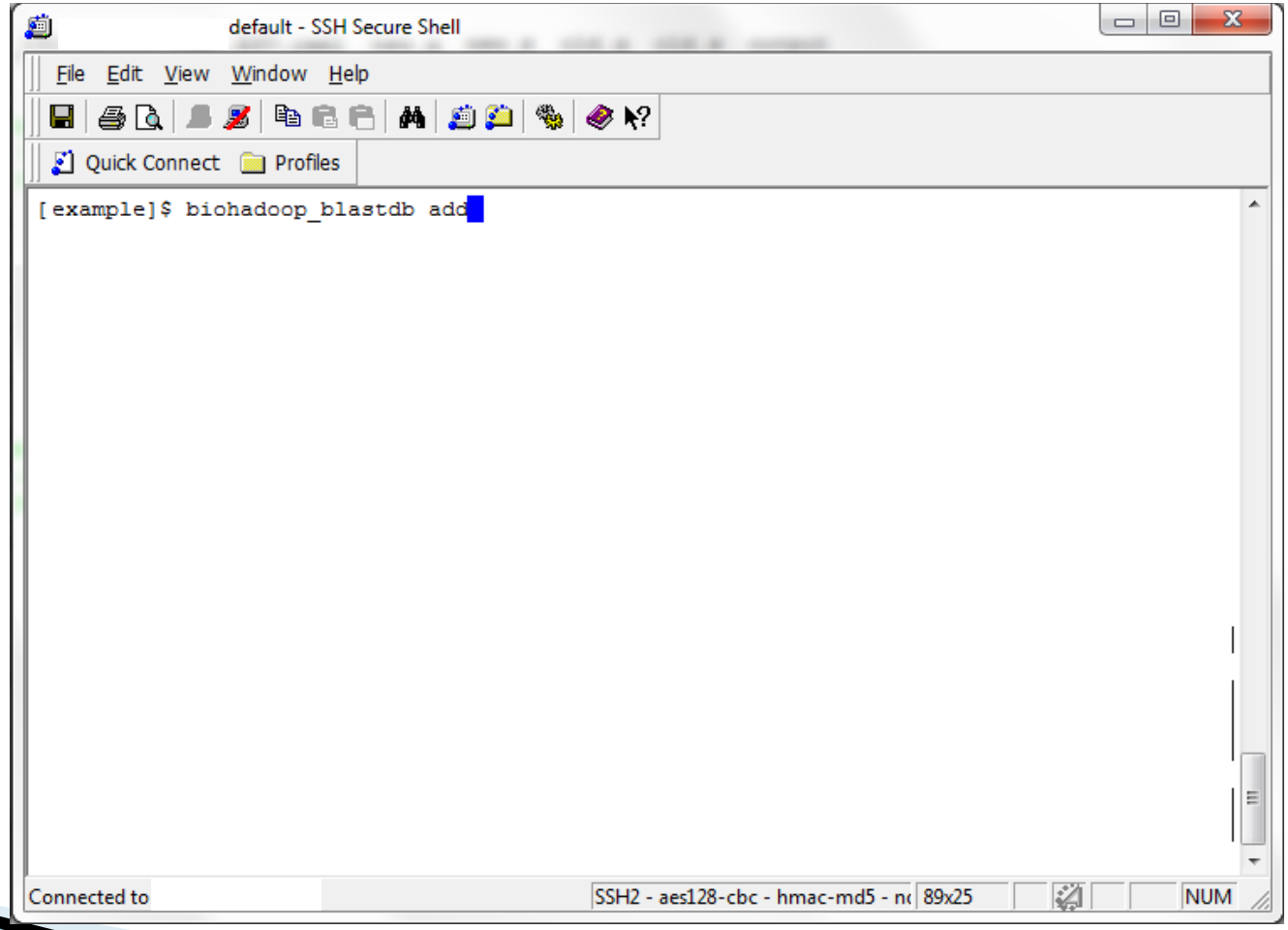
- While a job is running:
  - Do not:
    - Move/delete the input file
    - Move/delete the job folder
    - Move/delete the two files created in your working directory

How do I verify the status of my jobs?

Run the command biohadoop_job_status



How do I make a database?



How do I make a database?

default - SSH Secure Shell

File Edit View Window Help

Quick Connect Profiles

```
Path to Input file (fasta):
short.fna
/home/example/ short.fna
What Type of sequences is your fasta file?
01: Protein
02: Nucleotide
Please enter a number: 2
You have selected to format using the 'F' option
Database name (current type [Nuc, Prot], date and time are automatically appended)
Please only use alpha numberic and space characters only:
Database_example
```

Connected to SSH2 - aes128-cbc - hmac-md5 - n 89x25 NUM

How do I make a database?