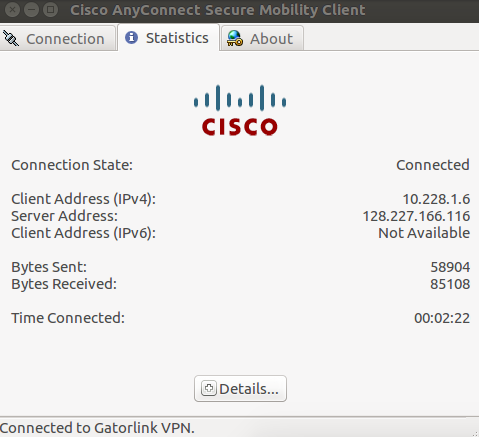
We firstly need to use ‘cisco anyconnect’ to get access to connect to our cluster if we are not on campus, since our spark is deployed at some servers of university.



Then we can use ssh to connect to our remote server. Our cluster for development contains three servers: test01, tes02 and test03. For each server, we can use ssh to get into another server directly. Here is an example that I login to test02 and go to test01.



We have downloaded some raw data and quality-controlled data from s3, and put them in both test02 and test01, since test03 is master node.