Lab #11: Application with FPGA Accelerator

06/10/2018

4190.309A: Hardware System Design
(Spring 2018)

Lab 11 WARNING

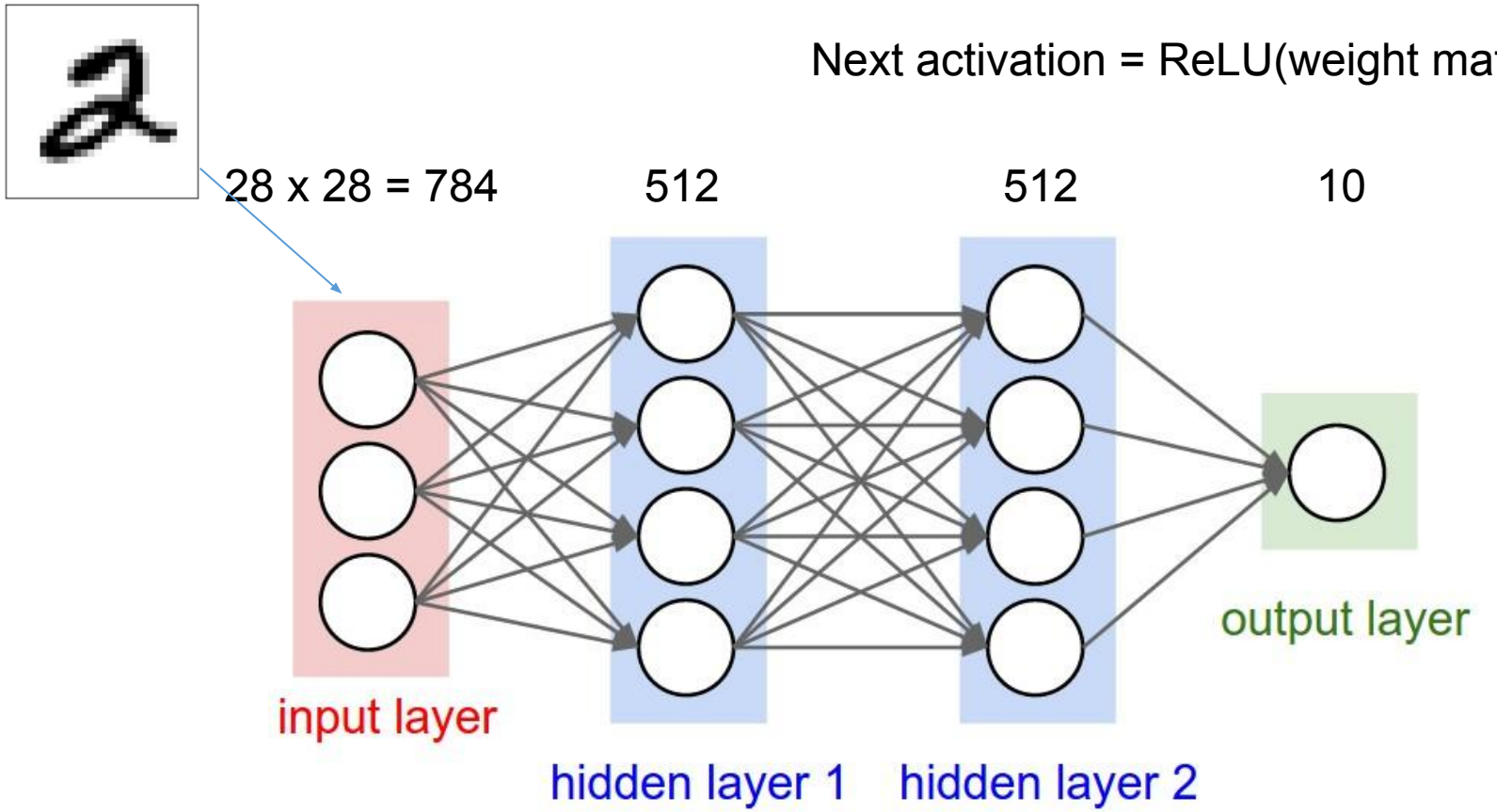
- **Never enter** “sudo poweroff” when you leave from minicom.
  - Must enter “**ctrl+a q**” instead.
  - We have history log of your commands,
    so if any problem happens, we will check your logs.
  - The server will stop if you enter it.
- Must leave **10 minutes before** your finish time.
  - Otherwise, next team will not be able to access minicom due to lock issue.
  - ex. If your team time is 14:00~15:00 p.m,
    you should enter “ctrl+a q” and leave before 14:50 p.m.
- *If your team violates above rules, your Lab11 score will get **-50**.*

Lab 11: Overview

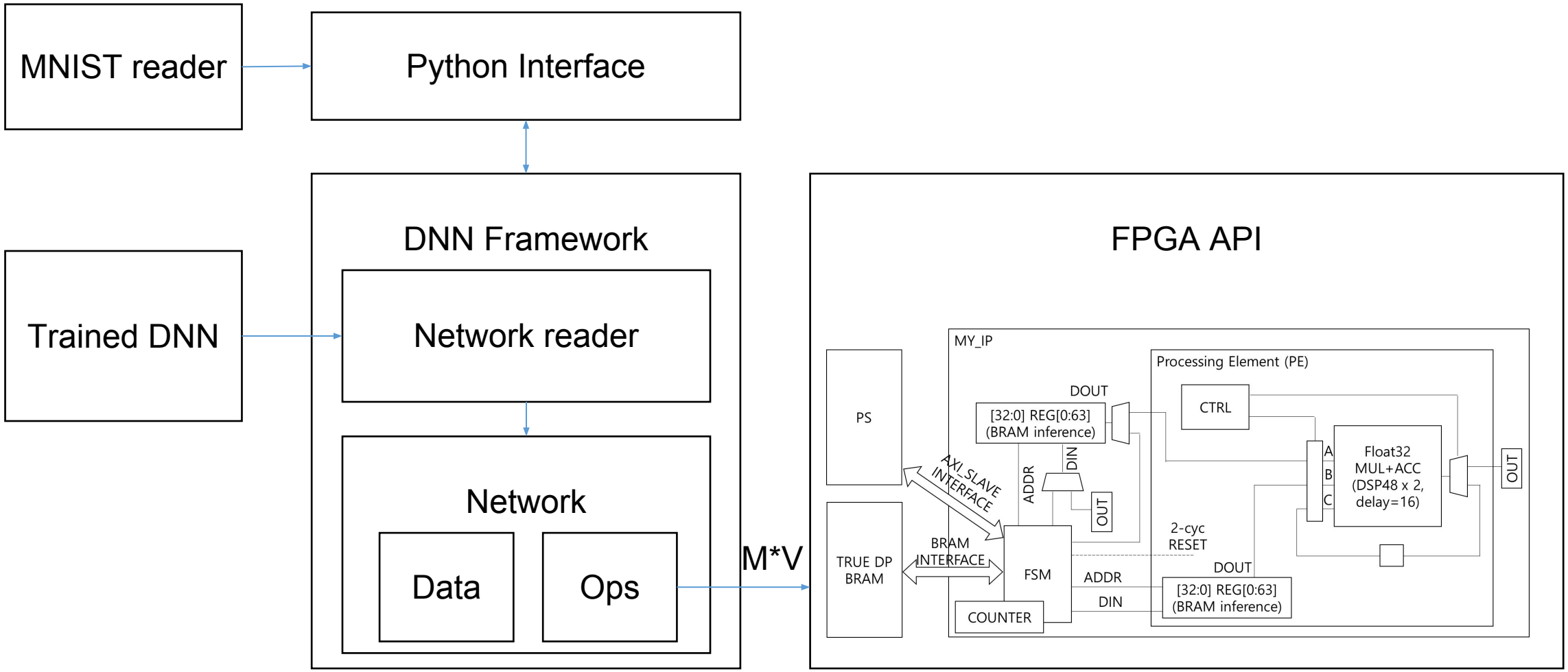
- Run the DNN to recognize the digit image
  - DNN Framework operates based on FPGA M*V API
- Analyze the effect of DNN Optimization

Review : DNN for MNIST

- 3 Layer MLP for MNIST dataset

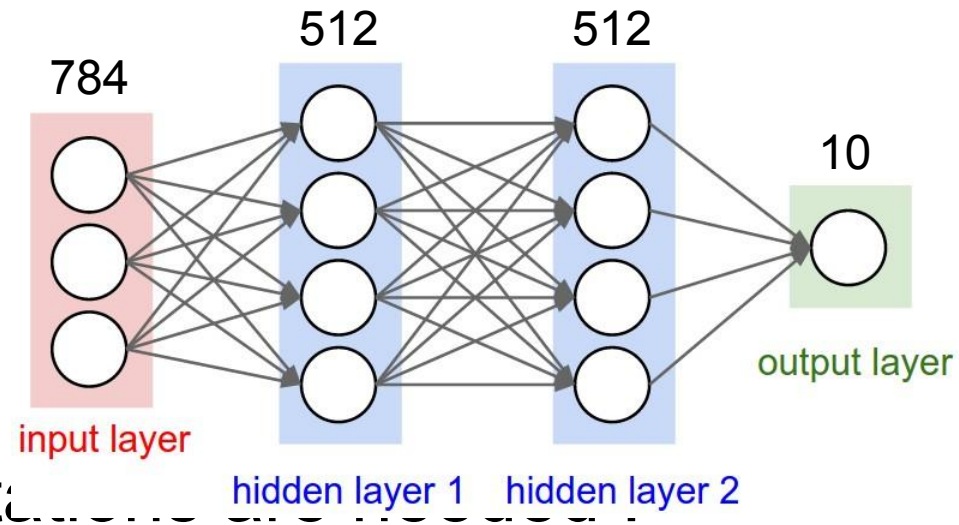


Review : DNN Framework



Computations in DNN

- Each layer of MLP is a matrix-vector multiplication
  - (N by M matrix) * (M-element vector) -> (N-element vector)



- So following computations are needed :
  - 512 by 784 matrix * 512-element vector
  - 512 by 512 matrix * 512-element vector
  - 10 by 512 matrix * 10-element vector

Computations in DNN

- Whole matrix-vector multiplication

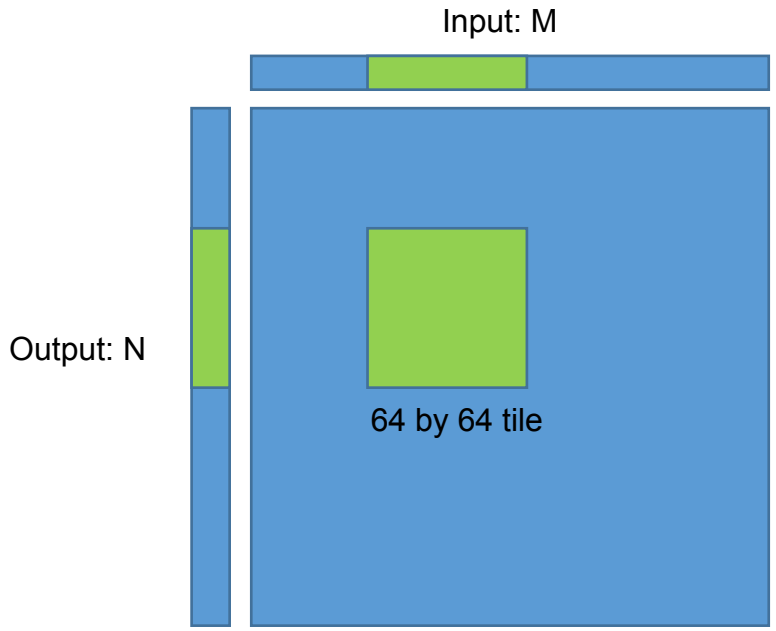
= matrix-vector multiplication on each tile + accumulation

- By tiling, a N by M matrix is split into `ceil(N/64)*ceil(M/64)` tiles

  where `ceil(x)` is min. integer equal to or greater than `x`

- matrix-vector multiplication on each tile is done by a single HW IP call

-> we need to call HW IP `ceil(N/64)*ceil(M/64)` times to complete the multiplication



Computations in DNN

```
zed@debian-zynq:/sdcard/HSD_LAB11$ sudo python mnist.py
read dataset...
create network...
run test...
image 0
HW IP was called 104 times
HW IP was called 64 times
HW IP was called 8 times
real number: 7
prediction result: 7
```

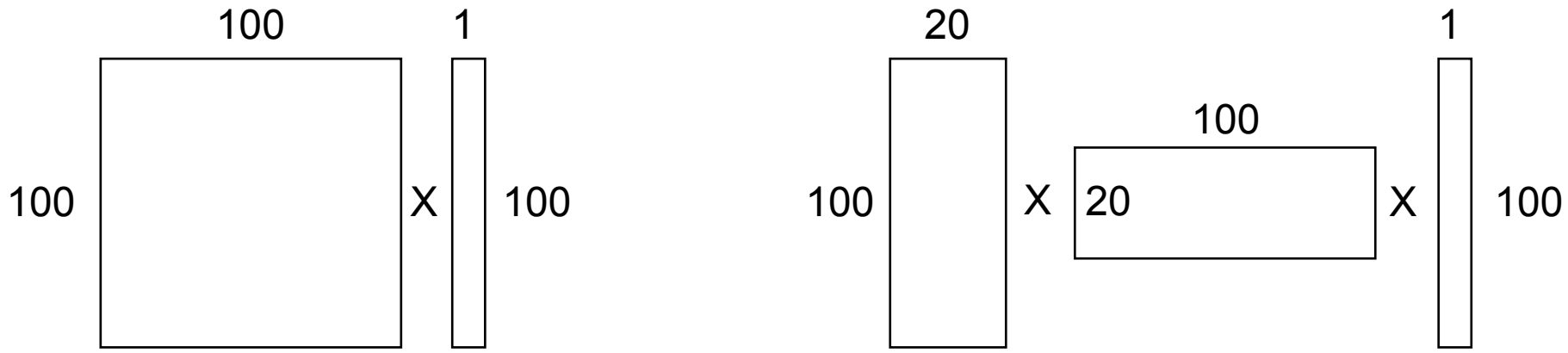
- 512 by 784 matrix * 512-element vector
  -> `ceil(512/64)*ceil(784/64)= 8*13 = 104` calls
- 512 by 512 matrix * 512-element vector
  -> `ceil(512/64)*ceil(512/64)= 8*8 = 64` calls
- 10 by 512 matrix * 10-element vector
  -> `ceil(10/64)*ceil(512/64)= 1*8 = 8` calls

- You can also compare “real number” and “prediction result” to check that DNN is working right with HW IP.

Optimization Example : SVD

- (N by M matrix) * (M-element vector) -> (N-element vector) can be approximated with following sequential steps below :
  - 1. (L by M matrix) * (M-element vector) -> (L-element vector)
  - 2. (N by L matrix) * (L-element vector above) -> (N-element vector)
- Through this optimization, we can reduce # of multiplications from NM to LM + NL (when L is small enough)
- Search for Singular Value Decomposition for additional details

Optimization Example : SVD



- For example, if we apply SVD with 20 on (100 by 100) * (100), (100 by 100) * (100) -> (100 by 20) * (20 by 100) * (100)
- # of multiplications : 10000 -> 4000 (60% reduction)

Optimization Example : SVD

```
zed@debian-zynq:/sdcard/HSD_LAB11$ sudo python mnist_16.py
read dataset...
create network...
run test...
image 0
HW IP was called 13 times
HW IP was called 8 times
HW IP was called 8 times
HW IP was called 8 times
HW IP was called 8 times
real number: 7
prediction result: 7
```

- mnist_x.py applies SVD on the DNN so that L(on the previous slide) = x.

- 16 by 784 matrix * 784-element vector
  -> `ceil(16/64)*ceil(784/64)= 13` calls
- 512 by 16 matrix * 16-element vector
  -> `ceil(512/64)*ceil(16/64)= 8` calls

- 16 by 512 matrix * 512-element vector
  -> `ceil(16/64)*ceil(512/64)= 8` calls
- 512 by 16 matrix * 16-element vector
  -> `ceil(512/64)*ceil(16/64)= 8` calls

- 10 by 512 matrix * 512-element vector
  -> `ceil(10/64)*ceil(512/64)= 8` calls

Optimization Tradeoffs

- Aggressive optimizations deprives accuracy of the result !
- No optimization may lead to unsatisfying performance, but too much optimization may lead to wrong result.
- Q1 : Among the five options for SVD (16, 32, 64, 128, 256), what do you think is the best optimization option? Why?
- Q2 : What if the tile size for HW IP is increased to from 64 to 128 ?

Goal

- Grading policy
  - Run mnist_16.py ~ mnist_256.py and get how many times the IP is called for each option (50 points)
  - Briefly answer Q1 on the previous slide (30 points)
  - Briefly answer Q2 on the previous slide (20 points)

- Submit “L11.**pdf**” within 1 page on eTL
  - Due : 6/17 (Sun) 11:59 PM

Todo

1. Reserve the time you want to access the server.
2. Access the server.
3. Access minicom. (`$> sudo minicom -D /dev/ttyACM0`)
4. Go to the repository for this lab. (`$> cd /sdcard/HSD_LAB11`)
5. Run DNN for MNIST via python scripts :

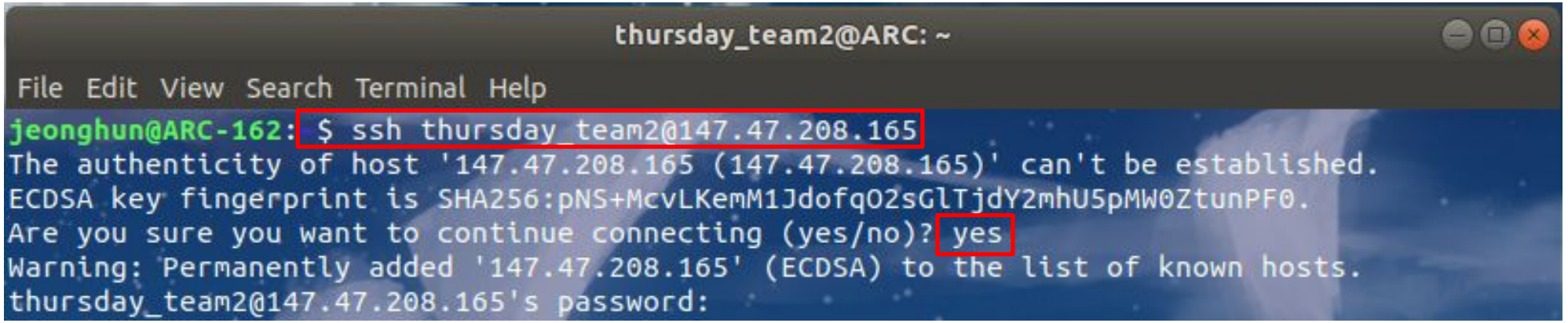
   mnist.py, mnist_16.py, … , mnist_256.py
6. Check how many times the IP is called for each case.
7. Exit minicom without “sudo poweroff”.
8. Finish ten minutes earlier for the next team.

Server Reservation

- Server running time : 6/10(Sun) 9:00 AM ~ 6/15(Fri) 11:59 PM
- To avoid multiple accesses at a time, please write down your team number and class at link below before you connect.
  - https://docs.google.com/spreadsheets/d/1IoOxgpYF1Fr-jsu55gj9mkVkjdaVBOLiaTn_tFY8g0w/edit?usp=sharing
- If you are asked for an access permission, send us the permission request.

Server Access

- Connect to server via secure shell(SSH)
  - IP address: 147.47.208.165
  - ID : thursday_team** / friday_team**
    - Example : Thursday team 12 : thursday_team12
  - Password : student number of your team(earlier one)
    - Example : 1234-12345



Lab 11 WARNING REMINDER

- **Never enter** “sudo poweroff” when you leave from minicom.
  - Must enter “**ctrl+a q**” instead.
  - We have history log of your commands,
    so if any problem happens, we will check your logs.
  - The server will stop if you enter it.
- Must leave **10 minutes before** your finish time.
  - Otherwise, next team will not be able to access minicom due to lock issue.
  - ex. If your team time is 14:00~15:00 p.m,
    you should enter “ctrl+a q” and leave before 14:50 p.m.
- *If your team violates above rules, your Lab11 score will get **-50**.*