Georgia Institute of Technology

ECE 8803: Hardware-Software Co-Design for Machine Learning Systems (HML)

Spring 2025

Lab 4A

Due: Friday, April 4, 2025 @ 11:59 pm EST

Created by: Changhai Man (cman8@gatech.edu)
Last Revised: Mar 15, 2025

Instructions

Please read the following instructions carefully.

- The lab is distributed in 3 parts (Each worth 2 points).
- You will need to modify the code in various files, generate output dot file and submit them in a pack.
- Rename the zip according to the format: **LastName_FirstName_ECE_8803_HML_sp25_lab4A.zip**
- **DO NOT MODIFY the submitted dot files.**
- It is encouraged for you to discuss homework problems amongst each other, but any copying is strictly prohibited and will be subject to Georgia Tech Honor Code.
- Late homework is not accepted unless arranged otherwise and in advance.
- Comment on your codes.
- For all problems, please post queries on piazza. If you add a comment to an answered
- query, make sure to change the comment to “Unresolved”.

Lab Layout

Part A-1: Implementing Activation Recomputation - 2 points
Part A-2: Modifying Backward Pass to Utilize Recomputed Activations - 3 points
Part A.3: Adding Control Dependencies of Activation Recomputation - 1 points

Lab Description

As you may have experienced in Lab 3, Activation Recomputation is a commonly used technique in large-scale LLM training. It is a powerful technique to save the memory footprint of each NPU when training the model, which allows it to train even larger models under constrained resources.

Your goal in Lab 4 is to implement the Activation Recomputation technique. You will be asked to modify the compute graph (like torch.fx which we covered in Lab 3A) to inject Activation Recomputation into the original computational graph.

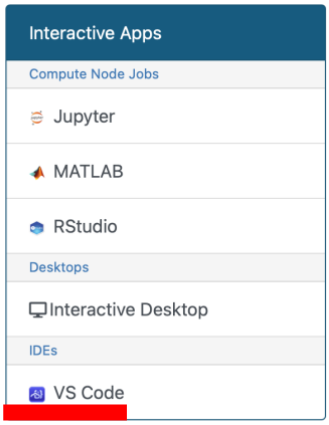
Lab Environment Setup

In this lab, you will follow the instructions and update the code in a Jupyter Notebook file named "lab4a.ipynb", similar to Lab 3. You may download this Jupyter Notebook file locally and work on it, however, please note that our reference machine is PACE-ICE and we'll grade your answer based on the PACE-ICE setup. Therefore, we highly recommend you use the PACE-ICE machine when working on this lab. Please follow the instructions below to launch this Jupyter Notebook file on the PACE-ICE OnDemand session:

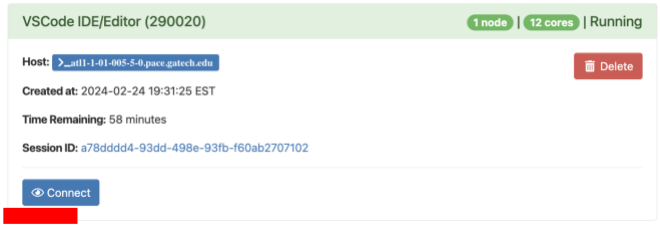
Pace OnDemand Setup

1. Configure GlobalProtect VPN. Refer to the following link.
   https://vpn.gatech.edu/
2. Use the following link to access to OnDemand ICE cluster.
   Link:
   https://ondemand-ice.pace.gatech.edu/pun/sys/dashboard
3. Go to “My Interactive Sessions” on the top

4. Select “VSCode IDE/Editor version” under IDE.

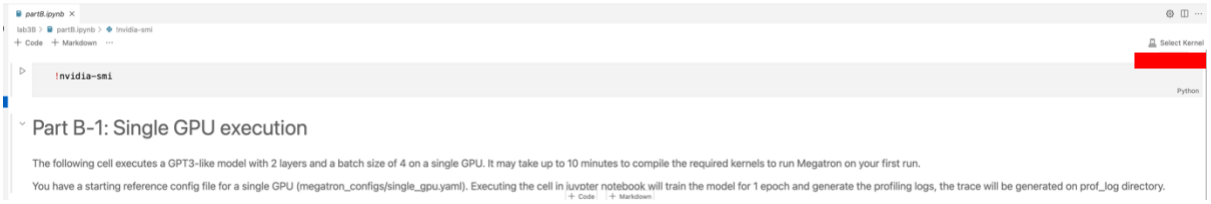


5. Configure the setting as follows:
    - **Modules:** Custom Pre-load
    - **Custom Pre-load Commands:**
      module load anaconda3
      bash /storage/ice-shared/ece8803hml/lab4A_setup.sh
      Quality of Service: Default (none)
    - **Node Type:** Intel CPU
    - **Cores:** 4
    - **Memory (GB):** 32
    - **Number of hours:** 4 # Or whichever hours you need. Please be respectful to others; delete instances immediately after completing the lab.
6. Once configured launch the session. Your session will start soon (This can take some time if there is heavy traffic).
7. Click “connect” to open the session.



VSCode Setup

1. Click “open folder” on the left, and open:
   /home/hice1/$YOURGATECHID/ece8803_hml_lab4
    - If the folder does not appear, verify your configuration in step 5.
2. Open “lab4A.ipynb” in directory
3. Click “Select Kernel” on top-right



4. Select Kernel → Python Environment → Create Python Environment → venv → /bin/python
5. Wait for the creation of the venv.
6. Now you're ready to work on the actual lab assignments. Please follow the instructions in the Jupyter Notebook, as well as the guidance below.

Lab Breakdown

You should finish the lab by going through the jupyter notebook to find detailed steps, here is a summary of lab breakdown.

Part A-0: Read through the Jupyter Notebook [0 points].

(You need to start from the beginning to understand how to finish this lab)

Part A-1: Implementing Activation Recomputation [3 points].

Part A-2: Modifying Backward Pass to Utilize Recomputed Activations [2 points].

Part A-3: Adding Control Dependencies of Activation Recomputation [1 points].

Lab Submission

- Submission due: Friday, Apr 4, 2025, by 11:59 pm EDT.
- Submission format: 1 zip with following contents
    - lab4a.ipynb with finished codes
    - llm_mlp.dot
    - llm_mlp_ar.dot