EE193 – Emerging Memory Technologies
Fall 2021
Prof. Marco Donato

Lab assignment #4

For this assignment we are going to evaluate the impact of different memory architectures on the performance of a DNN hardware accelerator. Our experiments are going to be based on the Nvidia Deep Learning Architecture (NVDLA) open-source model, and NVSim simulation results.

The NVDLA performance model is a convenient tool for DNN hardware rapid design space exploration, and can be downloaded here. The spreadsheet allows to customize many design specifications such as data type, operating frequency, and memory architecture. The hardware architecture is evaluated using three convolutional deep neural networks (AlexNet, GoogleNet, and ResNet 50). Based on the input configuration settings, the model outputs a collection of performance metrics. You can find more details in the first sheet (Readme).

For our experiments we are going to focus on AlexNet and extend the model to include energy estimates and eNVM storage.

NOTE: You may have to manually re-calculate the spreadsheet in order to apply your changes. In order to do so, go to the 'Formulas' tab and select 'Calculate Now'. You can also automate the process by selecting 'Automatic' in the 'Calculate Options' menu.

Part I: Getting started

The sheet describing the computational model for AlexNet can be divided in 5 regions:

- The top left region reports a summary of the design parameters derived from the Configuration Input sheet. In addition, the Compression rate parameter allows to vary the model sparsity;
- The block on the left (green) summarizes the DNN architecture and computes the required memory and computational parameters;
- The block in the center (pink) computes intermediate results based on the accelerator and memory configuration for an architecture using a 4MB on-chip memory;
- The block on the right (blue) computes intermediate results based on the accelerator and memory configuration for an architecture without on-chip memory;
- The block on the bottom (orange) shows the final performance metric results;

Answer the following questions based on the default configuration reported in the spreadsheet:

1) What are the total memory requirements for the AlexNet model?

2) What factors determine the overall data traffic for both off- and on-chip memory?

3) Compare the performance results for the two proposed design choices (with and without on-chip SRAM) and comment on any difference between the two. What layers have the highest impact on performance?

Part II: Evaluating the memory system energy

The NVDLA examples we have used in Part I provide performance estimates but do not include any energy metric. For this part of the assignment we are going to extend the model to include energy estimates for the memory architecture. Note that it is also possible to compute the MAC energy for better model accuracy, but since we are not going to make any changes to the datapath implementation we are going to ignore its contribution.

1) Modify the spreadsheet to include a more accurate memory model for both SRAM and DRAM. For the SRAM model, include the NVSim estimates for read and write bandwidth, area, and read and write energy. For the DRAM model you can assume the same bandwidth, and use a total power of 200 mW. Use these parameters to compute the **energy per layer**, and **total energy per inference**.

2) Repeat the steps in 1) replacing the SRAM memory with SLC RRAM. Run experiments for both **iso-capacity** and **iso-area** cases. Comment on your results highlighting the advantages and/or disadvantages of these design choices both in terms of energy and performance.

Part II: Memory system optimization

Based on the results from Part I and II, we are now going to optimize the system by implementing a hybrid memory in which both SRAM and SLC RRAM are used. Your goal is to achieve the best performance and energy with minimal area overhead.

1) Based on NVSim results create a fourth design case in which both SRAM, eNVM and DRAM can be used. You should evaluate different target optimizations in NVSim to get the best implementation for your design.

2) Repeat the experiment from 1) by evaluating different model compression options. You can select the compression rate by editing the value in cell F2. Report your results for compression rates between 20% and 80%.

In your report, state all the assumptions you have made for creating your model and describe the NVSim design settings you have used for your experiments. The deliverables should include a copy of your edited NVDLA spreadsheet.