MAESTRO Hands-on Exercises

Prasanth Chatarasi, PhD student, Habanero Research Group (Prof. Vivek Sarkar), Georgia Institute of Technology
MAESTRO: Analytical Cost/Benefit Model

DNN Layer Sizes

HW Resources

Mapping (Dataflow)

Abstract HW Model

Data Reuse Analysis

Communication Analysis

Computation Analysis

• Size Requirement
• Access Count (Energy)

Buffer Analysis

• BW Requirement
• NoC Activity Count

NoC Analysis

• Roofline Throughput
• Expected Runtime

Runtime Analysis
Layer Description

Data Dimensions (Loop Variables)
- $K/C$: Input/output Channel
- $Y/X$: Input Height/Width
- $R/S$: Filter Height/Width
- $N$: Batch
Dataflow Description

• MAESTRO directives
  - Temporal Map: Temporal_Map (size, Offset) \textit{dim}
  - Spatial Map: Spatial_Map (size, Offset) \textit{dim}
  - Cluster: Cluster (size)

Unroll: A syntactic sugar for Temporal_Map(size(\textit{dim}), *) \textit{dim}
Launching MAESTRO

“run.sh” has all the command line arguments of MAESTRO

```bash
./maestro --dataflow_file='data/dataflow/ws.m' \  
   --layer_file='data/layer/re_vgg16_conv2.m' \  
   --noc_bw=4 \  
   --num_pes=16
```
Exercises

• **Exercise #1 (Varying hardware parameters for VGG CONV2)**
  1. With NoC bandwidth = 256 elements, sweep total cycles as a function of the number of PEs 1-256 (powers of 2)
     • At what point do you see diminishing returns? Why?
  2. With NoC bandwidth = 256 elements, sweep total cycles as a function of the number of PEs 1-256 (powers of 2)
     • Do you see diminishing returns? If not, why?

• **Exercise #2 (Varying dataflow for VGG CONV2)**
  1. With NoC bandwidth = 256 elements and PE’s as 16, sweep size and offset of Temporal_Map (3,1) Y in increments of 2, e.g., (5,3), (7,5)?
     • Do you see total cycles going down? If so, why?
Exercises

• Exercise #3 (Varying layers)
  1. With NoC bandwidth = 256 elements and PE’s as 16, Replace X map with Temporal_Map (3,1) and replace C with Spatial_Map(1,1), and sweep size and offset of C in powers of 2
   • Is it better than without the original without replacement? If so what is the reason?

• Challenge (Varying all – layers, dataflow, and hardware parameters)
  1. Design a new dataflow with the following properties:
     • No more than 16 PEs, bandwidth 4, L2 to L1 BW < 0.04 per cycle, 1 Spatial_Map and 5 Temporal_Maps
     • Better than all provided dataflows on late layers, but can be worse on early layers