Power/Performance analysis and optimization for deep learning on CPU-GPU platform

Ahmet Fatih Inci

Ting-Wu (Rudy) Chin

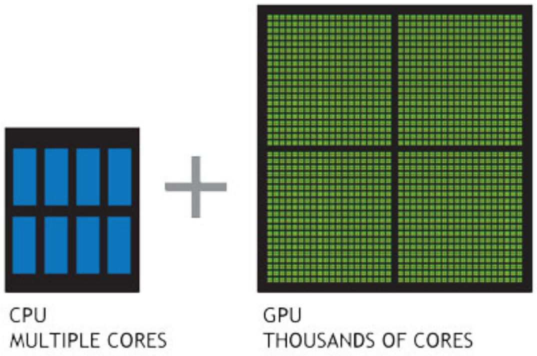
18-743 Energy-Aware Computing

Outline

- » Motivation
- » Introduction
- » Methodology
- » Objectives & Deliverables
- » Milestones

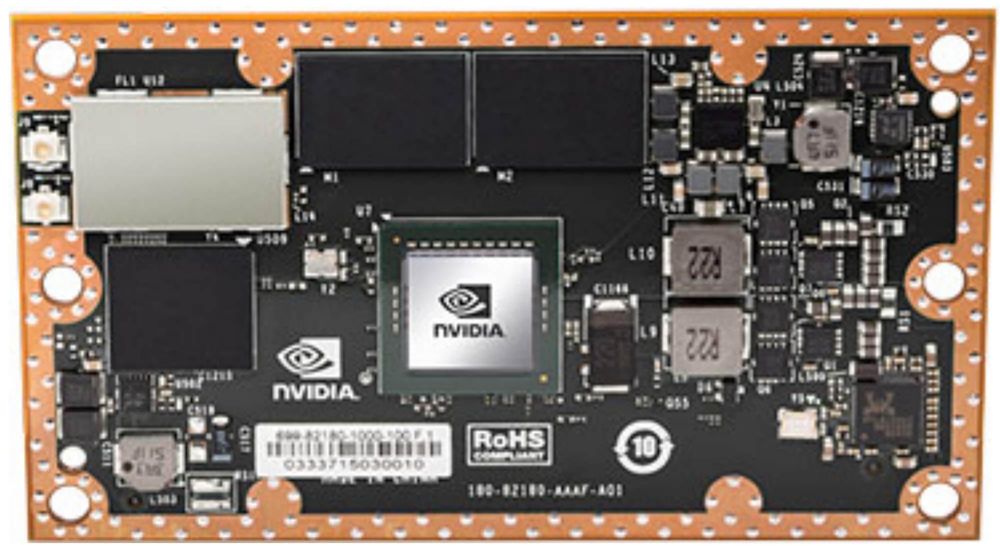
Motivation

- focus on GPU to run DNNs faster
  - data parallelism of DNNs
- what is left to be done on CPU?
  - share same power budget
  - close to each other on SoC
  - underutilized CPU
- characterize the optimum system performance



Introduction

» Profiling power/performance of embedded platform (TX1) while inferencing DNN on GPU and running SPLASH on CPU.



Nvidia Jetson TX1

Methodology

- GPU
  - various DNN architectures
  - various frequency
- CPU
  - various SPLASH benchmarks
  - various frequency
- characterize power/performance individually and jointly

Methodology

- Power
  - current sensors in TX1
- Performance
  - DNN execution time (Caffe framework)
  - CPU utilization (stats)
  - IPC (performance counters)
- Temperature
  - thermal sensors in TX1
  - take off heat sink to simulate embedded platforms

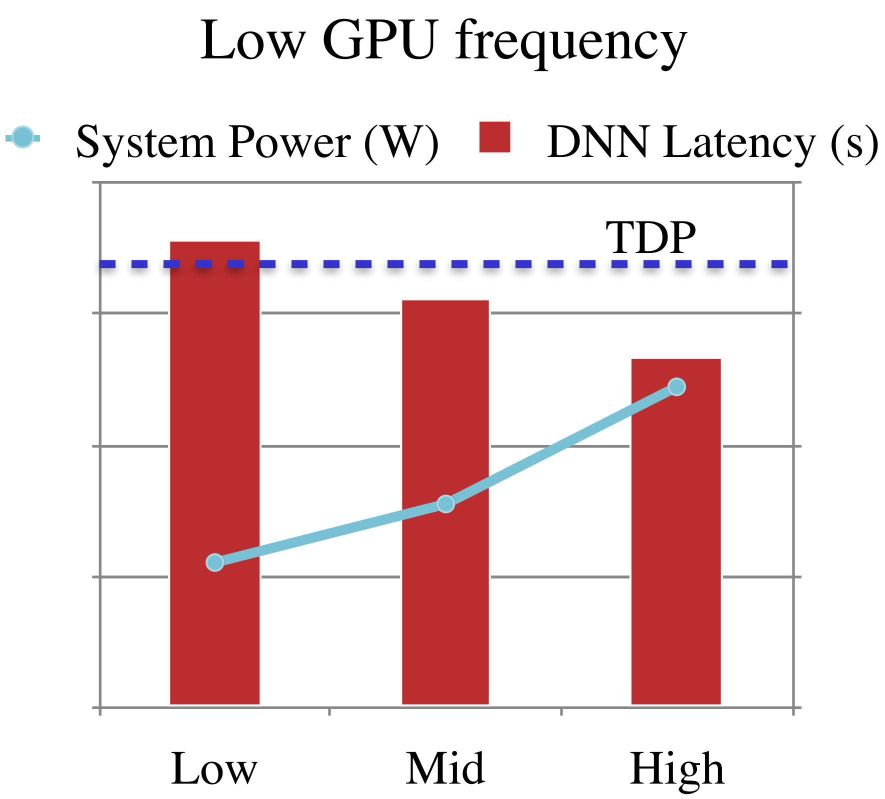
Objectives & Deliverables

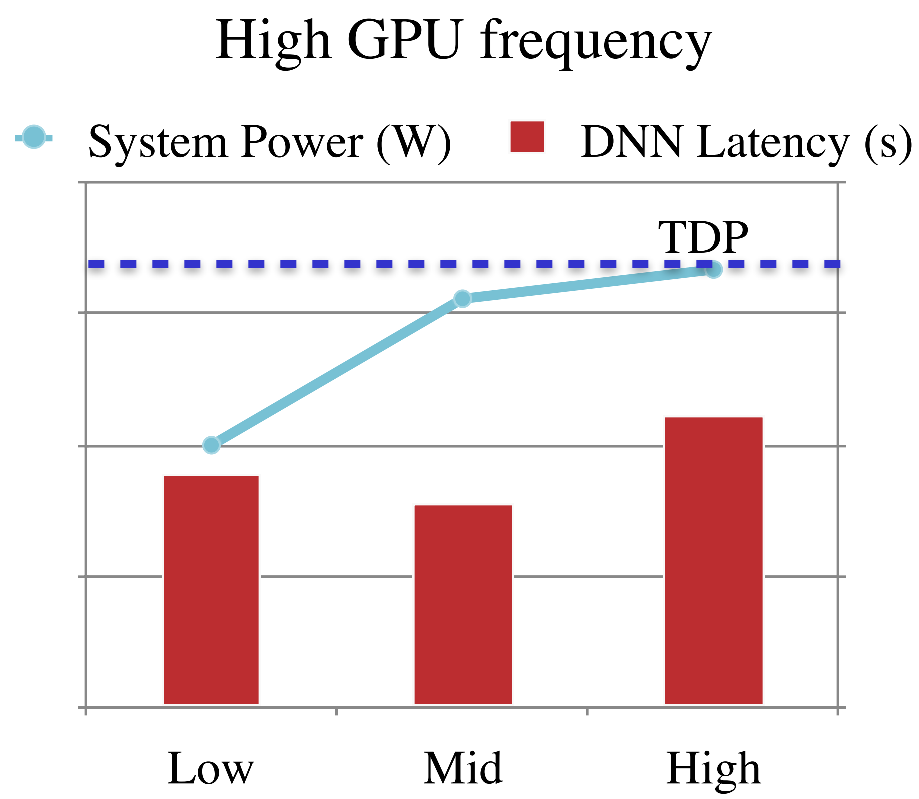
» how different CPU frequency and workloads affect GPU performance and system power under TDP constraint.

» how CPU-GPU workloads affect temperature

» what type of CPU workloads (memory and compute intensive) has more IPC

» e.g.





» I/O operations on CPU

» Thermal Design Power (TDP) constraint

Milestones

- M1
  - come up with CPU-GPU benchmarks
  - run CPU-GPU workloads individually (baseline)
- M2
  - run CPU-GPU together by changing CPU-GPU frequency
  - analyze results

Q&A