An Open-Source Power Monitoring Framework for Real-Time Energy-Aware GPU Scheduling Research

Mohsen Karimi, Yidi Wang, and Hyoseung Kim

University of California, Riverside

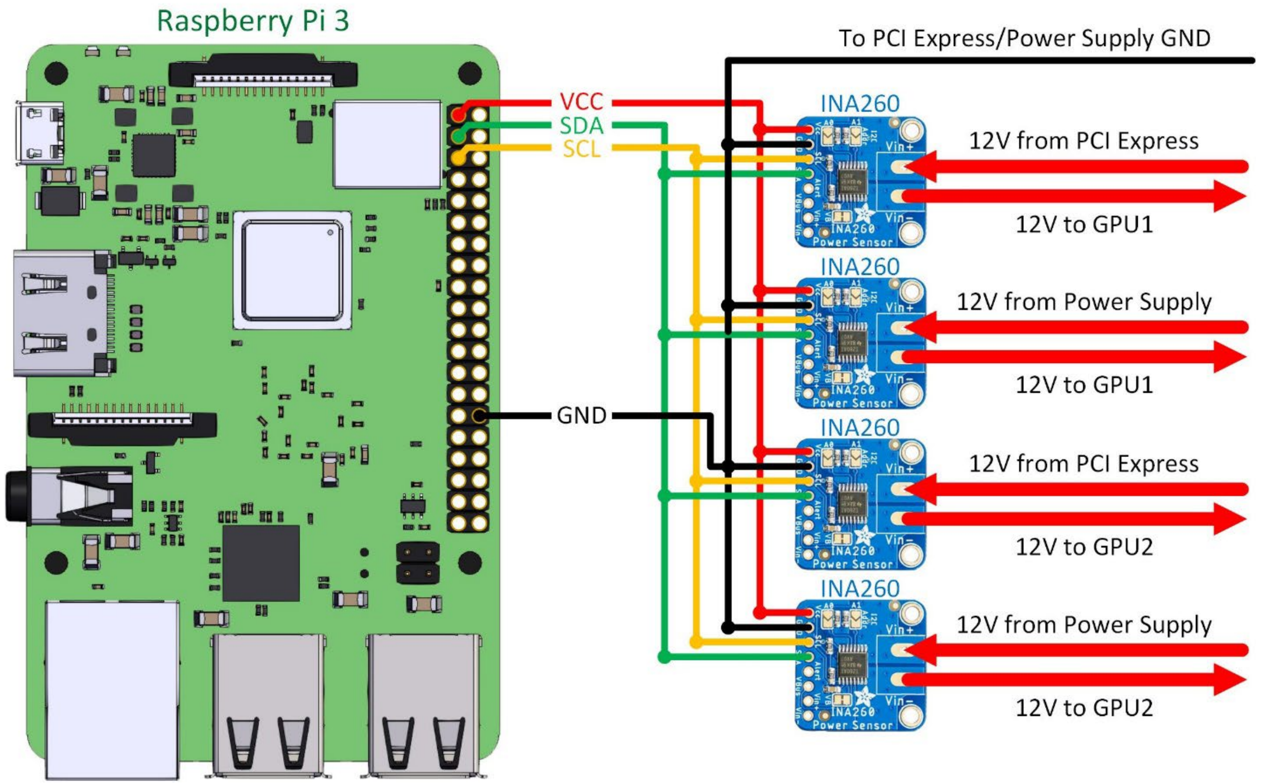
Motivation

- Power consumption of GPUs is concerning in real-time systems with stringent power constraints such as automobiles
- Analytical study of real-time systems with power consumption constraint on GPUs requires comprehensive knowledge about GPU architecture
- The detailed architecture of COTS GPUs is not publicly open
- Onboard sensors and APIs, e.g. the one provided by NVIDIA, are slow and imprecise
    - The rate is reported as 50Hz [1]

[1] Bridges, Robert A., Neena Imam, and Tiffany M. Mintz. “Understanding GPU Power.” ACM Computing Surveys 49, no. 3 (December 13, 2016): 1–27

Demo Description

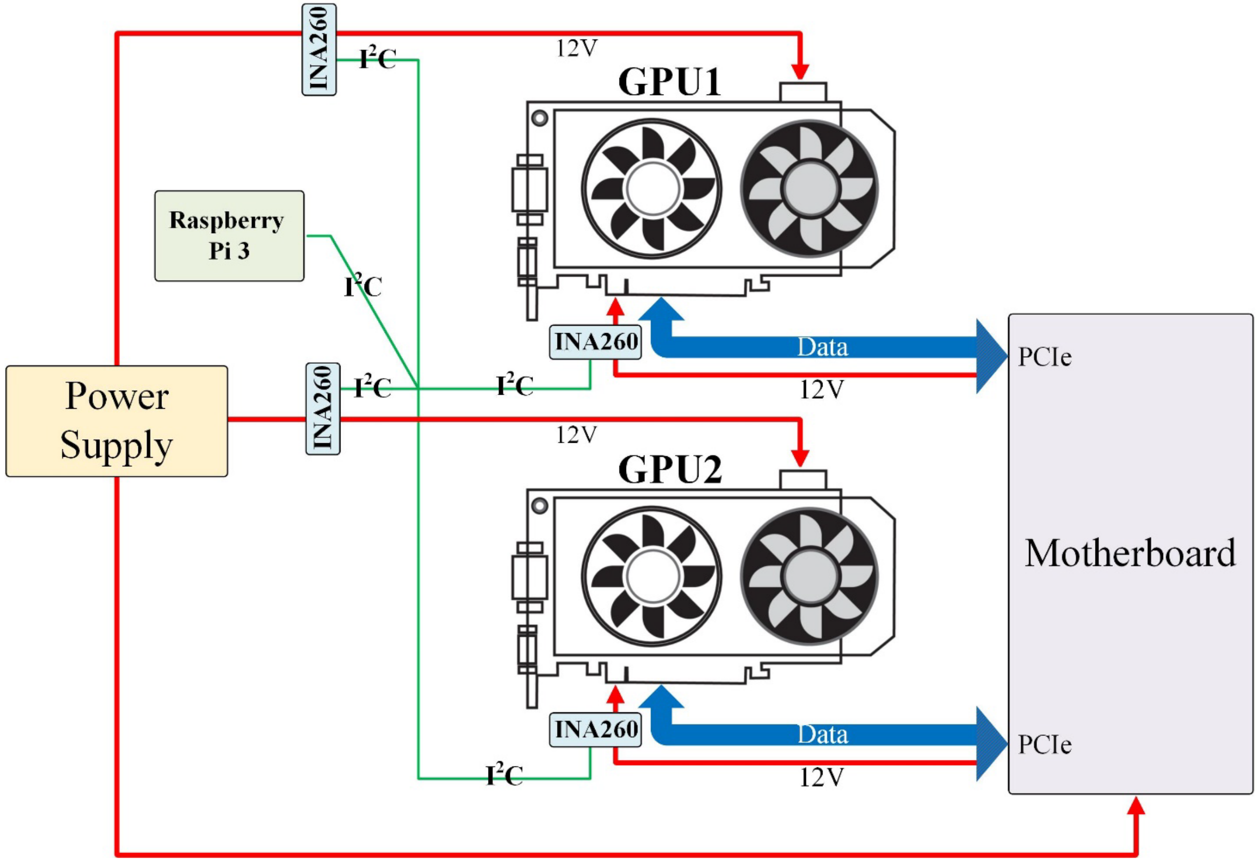
- INA260 sensors to measure voltage and current
  - 140us sampling rate
  - 1.5 mA resolution
- Raspberry Pi to collect and send data
  - Each voltage and current data is stored with time stamp (in microsecond resolution)
  - The data is sent over WiFi
- An open-source library for high-speed sensor data collection [2]
  - C language

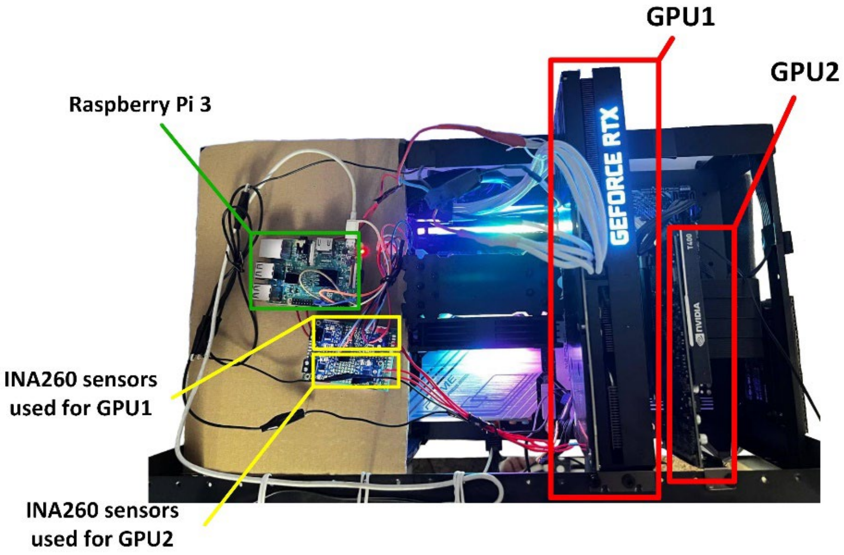


[2] GPU Power Monitoring System. https://github.com/rtenlab/gpu_power_monitoring

Implementation

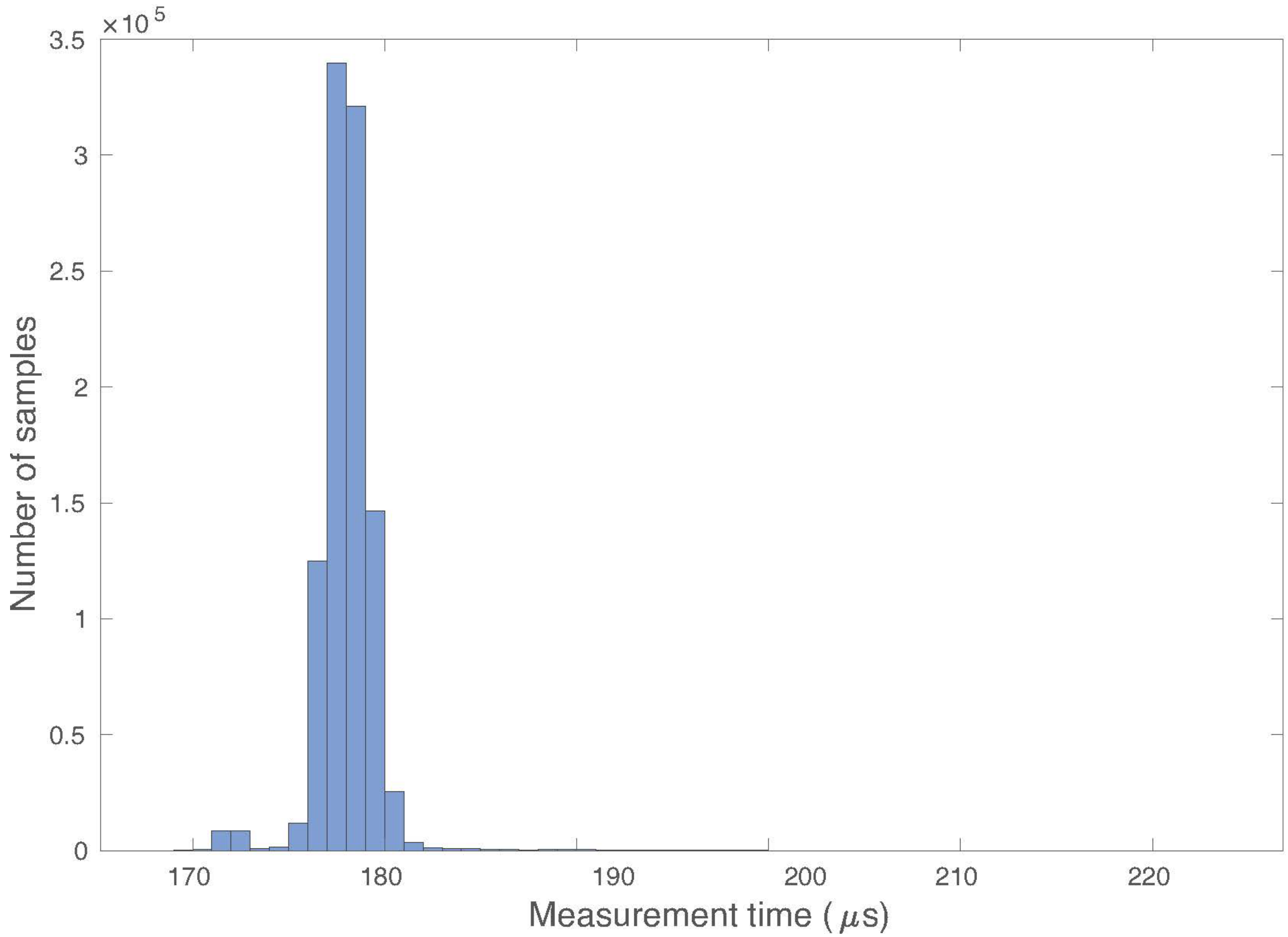
- Two GPUs
  - NVIDIA RTX 3070
  - NVIDIA T400
- Linux OS
  - *Real-time* priority is given to the task to reduce the delay caused by other tasks running on OS
  - Unnecessary services such as GUI are disabled





Results

- 1 Million samples are collected
- More than 5KHz sampling rate
  - Minimum measurement time: 168 µs
  - Average measurement time: 177 µs
  - Maximum measurement time: 224 µs
  - 99th percentile: 181 µs



Thank You

https://github.com/rtenlab/gpu_power_monitoring