Small-scale GPUs for Power-limited Applications in Computer Vision and Mobile Robots

Vic Castillo

Lawrence Livermore National Laboratory

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.
LLNL-POST-670572

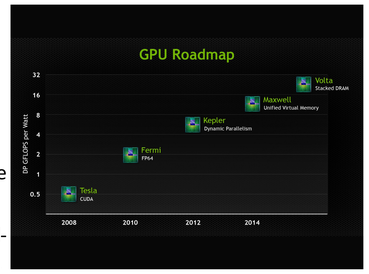
Abstract

Graphics processing units (GPUs) are well-suited for a variety of computational tasks in computer vision, robotics, and machine learning. In situations with limited power, small GPU devices can provide the needed computational performance. In this short talk, I will introduce the NIVIDIA Jetson TK1 embedded development kit that includes a 4-Plus-1 ARM CPU and Kepler architecture GPU with 192 CUDA cores and runs on only 5-10 watts. This development kit is great for hobbyists, researchers, and educators. Its performance and energy efficiency will be compared against other commodity options. I also hope to start a conversation about other possible applications.

GPU Computing

GPU Computing provides power-efficient computing at many scales. The embedded GPU computer such as the NVIDIA Jetson , the typical gaming laptop, the Titan graphics cards, and the next generation , ultra-high performance supercomputer, Sierra … part of the CORAL Collaboration of Oak Ridge National Laboratory (ORNL), Argonne National Laboratory (ONL) and Lawrence Livermore National Laboratory (LLNL) collective.

NVIDIA has recently release a roadmap for their next-generation processor with a focus on increasing the double precision computational performance per Watt. Currently, the LLNL LC visualization machines use the Kepler-class GPUs. Commodity "gaming graphics" cards are now available with the more-efficient Maxwell-class processors. The DOE next generation supercomputers will use the Volta series.



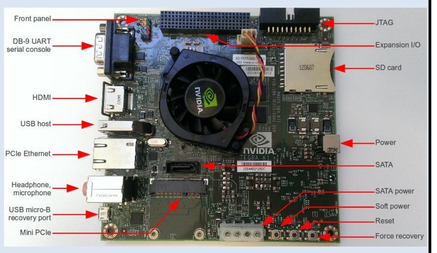
Machine Learning and Computer Vision

Applications in Machine Learning and Computer Vision have greatly benefitted from the advances in GPU capability. The highly parallel and high throughput is well suited for these algorithms. The NVIDIA CUDA-accelerated libraries allow access to the GPU. The GPU "device" is controlled by the CPU "host". Multiple GPU's can be effectively controlled in this fashion. The typical limitation is the GPU memory and the host-device data transfer.

Jetson TK1 Dev Board

The NVIDIA Jetson TK1 Development kit is a mobile embedded system designed to harness the power of the GPU. Powered 192 CUDA cores based on the Kepler architecture, the Jetson delivers over 300 GFLOPs of computational performance suitable for robotics, computer vision, and machine learning applications.

OS: Linux for Tegra (L4T)
modified Ubuntu 14.04
CUDA 6.5
Torch cuDNN + Lua
OpenCV4Tegra (CPU + GPU)



This device is an example of low-power, mobile GPU processors.

Torch + Lua

Torch7 is actively develop and use by big companies (Facebook, Google, Twitter). Definitely it's one of the best frameworks for Machine Learning and numerical analysis.

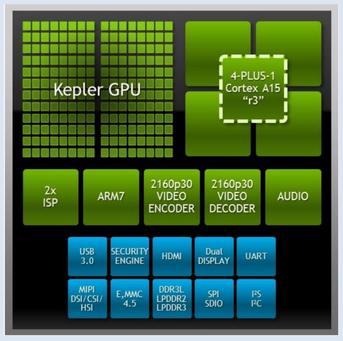
Lua is a powerful, fast, lightweight, embeddable scripting language… the leading scripting language in games. Several benchmarks show Lua as the fastest language in the realm of interpreted scripting languages. Lua is fast not only in fine-tuned benchmark programs, but in real life too. Lua is distributed in a small package and builds out-of-the-box in all platforms that have a standard C compiler. Also, it is easy to extend Lua with libraries written in other languages.

A fundamental concept in the design of Lua is to provide meta-mechanisms for implementing features, instead of providing a host of features directly in the language. For example, although Lua is not a pure object-oriented language, it does provide meta-mechanisms for implementing classes and inheritance. Lua's meta-mechanisms bring an economy of concepts and keep the language small, while allowing the semantics to be extended in unconventional ways.

Also, Lua interpreter built with all standard Lua libraries takes 241K!

Jetson Base Board

- Tegra K1 SOC
- NVIDIA Kepler GPU with 192 CUDA cores
- NVIDIA 4-Plus-1™ quad-core ARM® Cortex-A15 CPU2 GB memory
- 16 GB eMMC
- Gigabit Ethernet
- USB 3.0
- SD/MMC
- miniPCIe
- HDMI 1.4
- SATA
- Line out/Mic in
- RS232 serial port
- Expansion ports for additional display, GPIOs, and high-bandwidth camera interface



OpenCV image matching demo

This is a demo of 2D feature detection for OpenCV that uses both CPU and CUDA implementations.

Spatial convolution Deep Neural Network

This uses Torch7 demo

Performance Benchmarking