TR-2014-17

An Overview of NVIDIA Tegra K1 Architecture

Ang Li, Radu Serban, Dan Negrut

November 20, 2014

Abstract

This paperwork gives an overview of NVIDIA's Jetson TK1 Development Kit and its Tegra K1 architecture (32-bit version). We also compare some critical metrics between the Kepler GPU in Tegra K1 and that used in high-end systems, and highlighed that Tegra K1 is more power efficient. Furthermore, we conducted an experiment which shows how Tegra K1 performed compared with Tesla K40C in a specific application. We found that Tegra K1 can achieve better performance per Watt in around half of the benchmarks.

Contents

1 Introduction of Jetson TK1 Development Kit

Embedded Computing is the next frontier where GPUs can help accelerate the pace of innovation and deliver significant benefits in the fields of computer vision, robotics, automotive, image signal processing, network security, medicine, and many others. The Jetson TK1 Development Kit is specifically designed to enable rapid development of GPU-accelerated embedded applications, bringing significant parallel processing performance and exceptional power efficiency to embedded applications. [5]

The Jetson TK1 Development Kit is designed around the revolutionary 192-core NVIDIA Tegra K1 mobile processor, along with 2GB of RAM, 16GB of on-board storage and numerous peripherals and I/O ports. Tegra K1 is based on the same NVIDIA Kepler [3] GPU architecture used in supercomputers and High Performance Computing systems around the world. The Jetson Development Kit delivers a fully functional NVIDIA CUDA platform and includes the Board Support Package, CUDA 6, OpenGL 4.4, and the NVIDIA VisionWorks toolkit. With a complete suite of development and profiling tools, plus out-of-the-box support for cameras and other peripherals, the NVIDIA Jetson TK1 Development Kit is the ideal development platform to shape a brand new future for Embedded Computing.

2 NVIDIA Tegra K1 Architecture

There are two versions of Tegra K1 mobile processors: 32-bit version and 64-bit version, which are pin compatible[1]. The 32-bit version uses a quad-core Cortex-A15 CPU, which runs at clock rates up to 2.3GHz, is 3-way SuperScalar, and has 32KB L1 Instruction Cache and 32KB L1 Data Cache. The 64-bit version uses a custom dual-core Denver CPU, which runs at clock rates up to 2.5GHz, is 7-way SuperScalar, and has 128KB L1 Instruction Cache and 64KB L1 Data Cache. As NVIDIA has not published paperworks for the 64-bit dual core version yet, only the 32-bit quad core version will be overviewed.

Figure 1 shows the high-level architecture of the 32-bit version of NVIDIA Tegra K1 mobile processor. Some of the key features of the Tegra K1 SoC architecture are:

- **4-PLUS-1 Cortex A15 "r3" CPU architecture** that delivers higher performance and is more power efficient than the previous generation Tegra 4.
- **Kepler GPU architecture** that utilizes 192 CUDA cores to deliver advanced graphics capabilities, GPU computing with NVIDIA CUDA 6 support, breakthrough power efficiency and performance for the next generation of gaming and GPU-accelerated computing applications.
- **Dual Image Stream Processing (ISP) Core** that delivers 1.2 Giga Pixels per second of raw processing power supporting camera sensors up to 100 Megapixels.

[1] In electronics, a pin-compatible device is one that has the same functions assigned to the same particular pins.

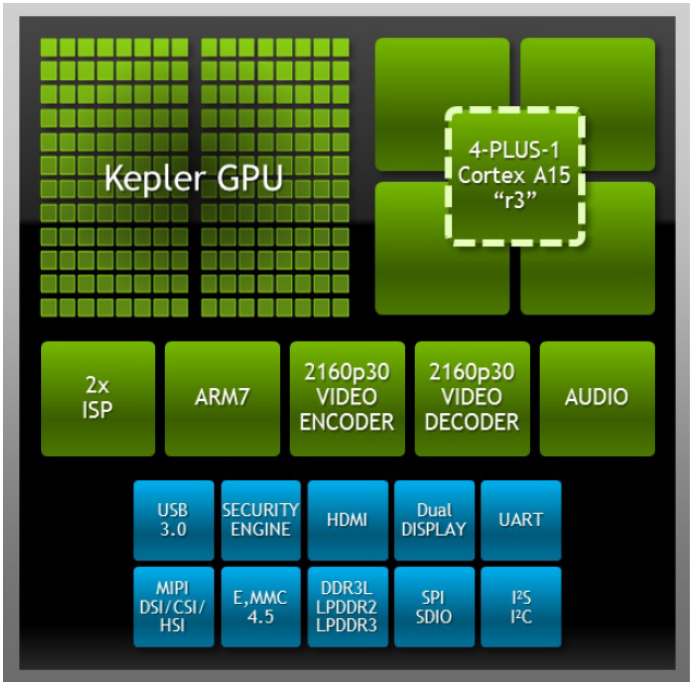

Figure 1: NVIDIA Tegra K1 mobile processor (32-bit version)

- **Advanced Display Engine** that is capable of simultaneously driving both the 4K local display and a 4K external monitors via HDMI
- Built on the **TSMC 28 nm High Performance Mobile process** to deliver excellent performance and power efficiency.

Besides these key features, the **Universal Asynchronous Receiver/Transmitter (UART)** translates data between parallel and serial forms; **MIPI DSI/CSI/HSI** stand for the Display/Camera/High-speed Synchronous Serial Interface of Mobile Industry Processor Interface; **Secure Digital Input/Output (SDIO)** is a type of secure digital card interface, which is a synchronous serial data link that operates in full duplex mode.

2.1 CPU

Tegra 4 was NVIDIA's first attempt to introduce the Cortex A15 quad core architecture into Tegra series, which runs at clock rates up to 1.9GHz. Tegra K1 sticks with the same CPU architecture, but NVIDIA makes improvements such that now it runs at clock rates up to 2.3GHz. Tegra K1 is claimed to facilitate up to 40% more performance at the same power level, or use only 45% of the power at a specific performance point in SPECint2000 benchmark suite [1].

As stated at the beginning of this section, the 32-bit version of Tegra K1 uses a quad-core Cortex-A15 CPU, which runs at clock rates up to 2.3GHz, is 3-way SuperScalar, and has

32KB L1 Instruction Cache and 32KB L1 Data Cache. The four Cortex A15 cores share a 2MB, 16-way set associative L2 Cache (shown in Fig. 2). There is an extra optimized "battery saver" A15 CPU core. This fact enables the flexibility to switch this special core to handle low performance tasks and thus extend battery life.

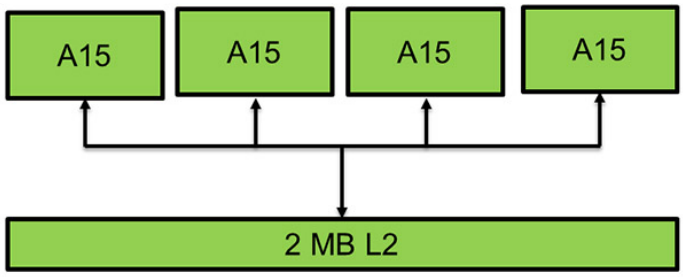

Figure 2: Nvidia's four Cortex-A15 cores share a 2MB, 16-way set associative L2 Cache

2.2 GPU

The architecture of the Kepler GPU in Tegra K1 is virtually identical to the Kepler GPU architecture used in high-end systems. Table 1 summarizes some feature metrics of NVIDIA's Tesla K40C and Tegra K1. Though all of [3, 5, 6] claim that due to the power optimizations of Tegra K1, Tegra K1 consumes significantly lower power compared with a Kepler GPU in high-end systems, and can achieve 1.5 times of performance/Watt compared with both Sony Xpheria Z Ultra and iPhone 5S, there is no performance comparison on certain benchmarks.

	K40C	Tegra K1
Memory Clock Rate [Hz]	3G	950M
Peak single FP perf. [FLOPS]	4.29T	364G
Memory Bandwidth [GB/s]	288	17
Memory Size [GB]	12	1.7
♯ of SMX unit	15	1
♯ of CUDA cores	2880	192
Power consumption [W]	~ 100	sub 2

Table 1: Metrics comparison between NVIDIA's Tesla K40C and Tegra K1. The "power consumption" measures average power on GPU power rail while playing a collection of popular mobile games.

3 Experiments

We ran a CPU/GPU hybrid implementation of Cuthill-Mckee (`CM`) [4] for 116 real application matrices, from both Florida Matrix Collection [7] and Simulation Based Engineering Lab [2], on a Jetson Tegra K1 card. For comparison, we ran the same implementation on a combination of NVIDIA Tesla's K40C card and Intel Nehalem Xeon E5520 2.26GHz processor. A slowdown threshold was set at 50 to achieve the same performance per Watt for the two configurations[2].

Fig. 3 shows the slowdown of all 116 matrices. The figure shows that Jetson TK1 is able to achieve better performance per Watt for around half of the matrice.

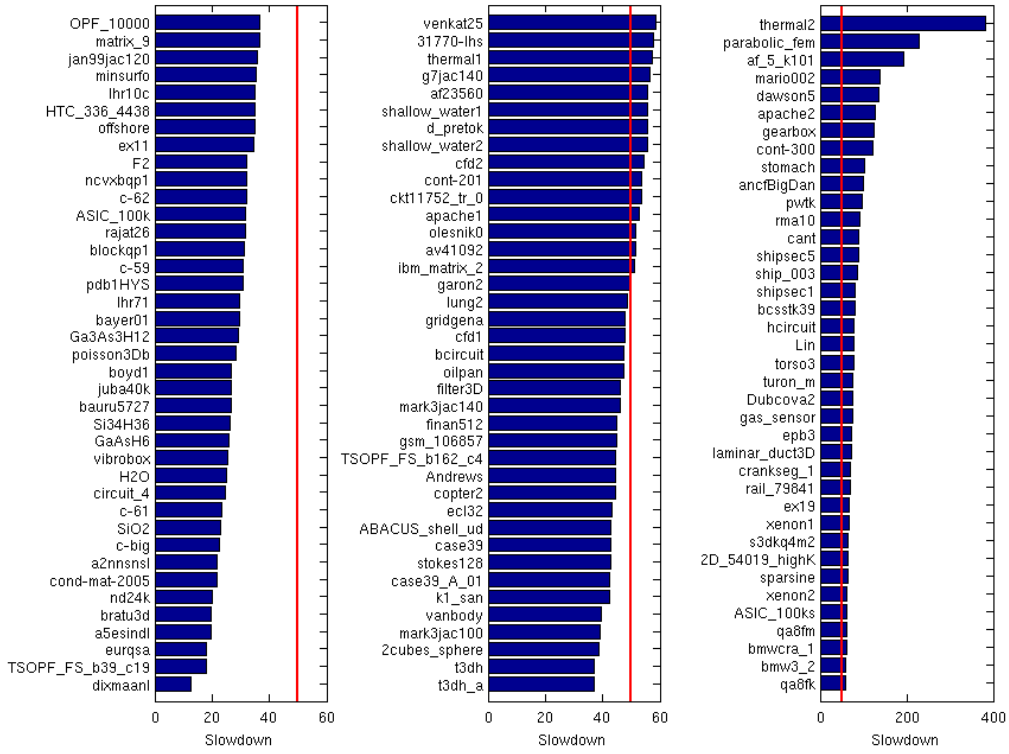

Figure 3: Execution time on Jetson TK1 card over execution time on Tesla's K40C card. The "red line" shows the threshold of a slowdown of 50 times.

[2]This conclusion (possibly problematically) assumes that the K40C and Intel Xeon combination consumes 50 times of power as Jetson TK1 does when running the `CM` application.

4 Summary

This paperwork gives an overview of NVIDIA's Jetson TK1 Development Kit and its Tegra K1 architecture (32-bit version). In the overview, three key points are as follows.

- There is a 32-bit version and a 64-bit version of Tegra K1 chip which are pin compatible, and we focused on the 32-bit version.
- We compared how Kepler architecture in Tegra K1 is different from that in Tesla K40C, and highlighted that Tegra K1 is power efficient. We also mentioned that we were still interested to see how Tegra K1 will perform compared with GPUs used in high-end systems.
- We conducted an experiment which shows how Tegra K1 performed compared with Tesla K40C in a specific application hybrid `CM` with 116 benchmarks. We concluded that Tegra K1 can achieve better performance per Watt in around half of the benchmarks, under the assumption that Tegra K1 consumes less than one 50th as Tesla K40C does.

References

[1] CINT2000 (Integer Component of SPEC CPU2000). http://www.spec.org/cpu2000/CINT2000, 2003.

[2] Simulation based engineering lab. http://sbel.wisc.edu, 2006.

[3] NVIDIAs Next Generation CUDA Compute Architecture: Kepler GK110. http://www.nvidia.com/content/PDF/kepler/ NVIDIA-Kepler-GK110-Architecture-Whitepaper.pdf, 2012.

[4] A Hybrid GPU-CPU Parallel CM Reordering Algorithm for Bandwidth Reduction of Large Sparse Matrices. http://sbel.wisc.edu/documents/TR-2014-12.pdf, 2014.

[5] NVIDIA Jetson TK1 Development Kit. http://developer.download.nvidia.com/embedded/jetson/TK1/docs/Jetson_platform_brief_May2014.pdf, 2014.

[6] Nvidia Tegra K1 In-Depth: The Power Of An Xbox In A Mobile SoC? http://www.tomshardware.com/reviews/tegra-k1-kepler-project-denver,3718.html, 2014.

[7] T. A. Davis and Y. Hu. The university of florida sparse matrix collection. *ACM Transactions on Mathematical Software (TOMS)*, 38(1):1, 2011.