CATEGORY: **SUPERCOMPUTING**

POSTER
SC04

CONTACT NAME
Matthias Lee: matthias.a.lee+jhu@gmail.com

GPU TECHNOLOGY CONFERENCE

CUDA on ARM: Tegra3 Based Low-Power GPU Compute Node

Institute for Data Intensive Engineering and Sciences
Johns Hopkins University

Matthias A Lee

Advisors: Tamás Budavári & Alex Szalay

The Johns Hopkins University

CUDA™ RESEARCH CENTER

Introduction

As scientific and enterprise fields have become more reliant on massive data sets, our need for quickly and efficiently processing this data has grown proportionately. Much of today's new research in fields such as Astronomy, Physics, Chemistry and Genomics rely on large amounts of computation, projects such as NCSA's Blue Waters and ORNL's Titan supercomputer make this need for this computation very evident. With the advent of the modern, generally-programmable GPU, GPUs were introduced into the data processing pipeline of modern supercomputers to gain performance at a small power premium. In general GPUs can deliver much higher Floating Point Operations per Second (FLOPS) per Watt than CPUs can. This has been a great advantage for power strapped data-centers looking to squeeze more performance out of their power resources.

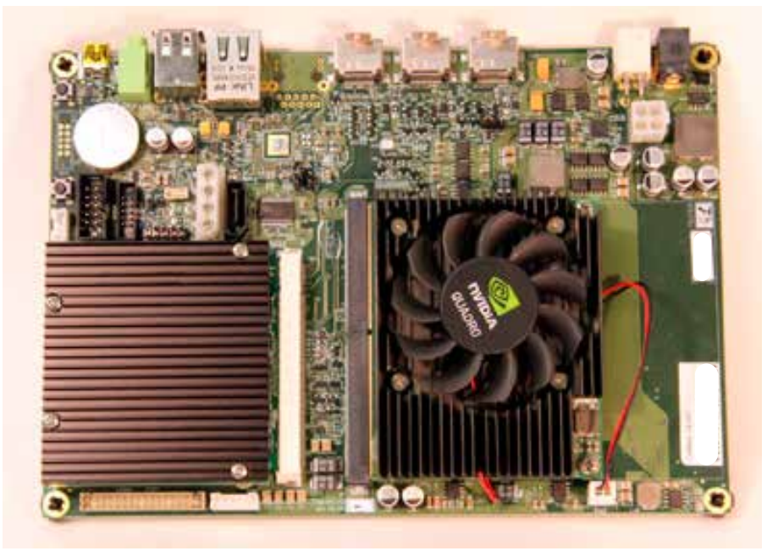


Figure 1: SECO CARMA DevKit, Quad-Core Tegra3 CPU and Quadro 1000M GPU

Simultaneously, as a byproduct of the smartphone and tablet market explosion, many companies, including Nvidia, have made investments in advances towards higher-performance power-efficient ARM processors. Recently a flood of diverse ARM-based boards have been released, the $35 RaspberryPi, the Arndale board, the BeagleBone, the PandaBoard and Dell's Copper, just to name a few. All of these are fantastic ARM-based computers, but very few of them are serious contenders when it comes to computational power. SECO's CARMA DevKit is one of the front runners in this crusade of combining low-power consumption with high performance computation and I/O. It combines a power efficient ARM-based Tegra3 processor with a powerful Quadro 1000M GPU, providing a low-power host for a very capable co-processor.

Previous Work

In 2010 Alex Szalay et al. [1] proposed an alternative approach to the traditional high-power clusters of machines found in today's data-centers and supercomputers. Szalay proposed an architecture comprised of a larger number of energy-efficient compute nodes coupled with solid state disks to achieve high I/O throughput and high compute performance per watt. Szalay's approach settled on the Intel Atom330-powered Zotac IONITX-A-U with an onboard Nvidia ION GPU. Intel's Atom processors and their AMD's counterparts provide great energy-efficiency in comparison to their high-end siblings, but x86 based processors pale in comparison to the power efficiency of ARM based processors. This is where SECO's CARMA DevKit comes in. The DevKit sports 1x/2x GigE port, 1x SATA port, a Quad-Core Nvidia Tegra3 processor and most importantly a CUDA enabled 96-Core Nvidia Quadro 1000M graphics processor, this set of features is a rare sight when it comes to ARM boards.

We aim to evaluate the DevKit's performance per watt characteristics and compare them to the same Intel Atom board Szalay found to be the best energy-efficient board of his comparison. To compare the performance we have setup a power-monitored stress test, which assumes that the majority of the workload is done by the GPU.

Benchmarking

Our goal with this benchmarking is to highlight the combination of an energy-efficient base system and a powerful GPU. Our hardware setup is as follows: the SECO CARMA DevKit runs a vanilla Ubuntu 11.04 install with CUDA 4.2 provided by SECO. Attached to the DevKit is an OCZ Vertex-2 120GB solid state drive containing the test user's home directory. The Zotac board's root disk is also an OCZ Vertex-2 120GB solid state drive, with a vanilla Ubuntu 11.04 server install configured with CUDA 4.2. (see: Figure 3 & 5) Important to note is that both boards are powered by a fan-less "brick" power-supply. Similar boards powered by traditional ATX power supplies consume ~ 5-10 Watts more, likely due to the fan. During initial testing we found the GPU fan of the DevKit and the CPU/GPU fan on the Zotac board to draw approximately 1 Watt each.

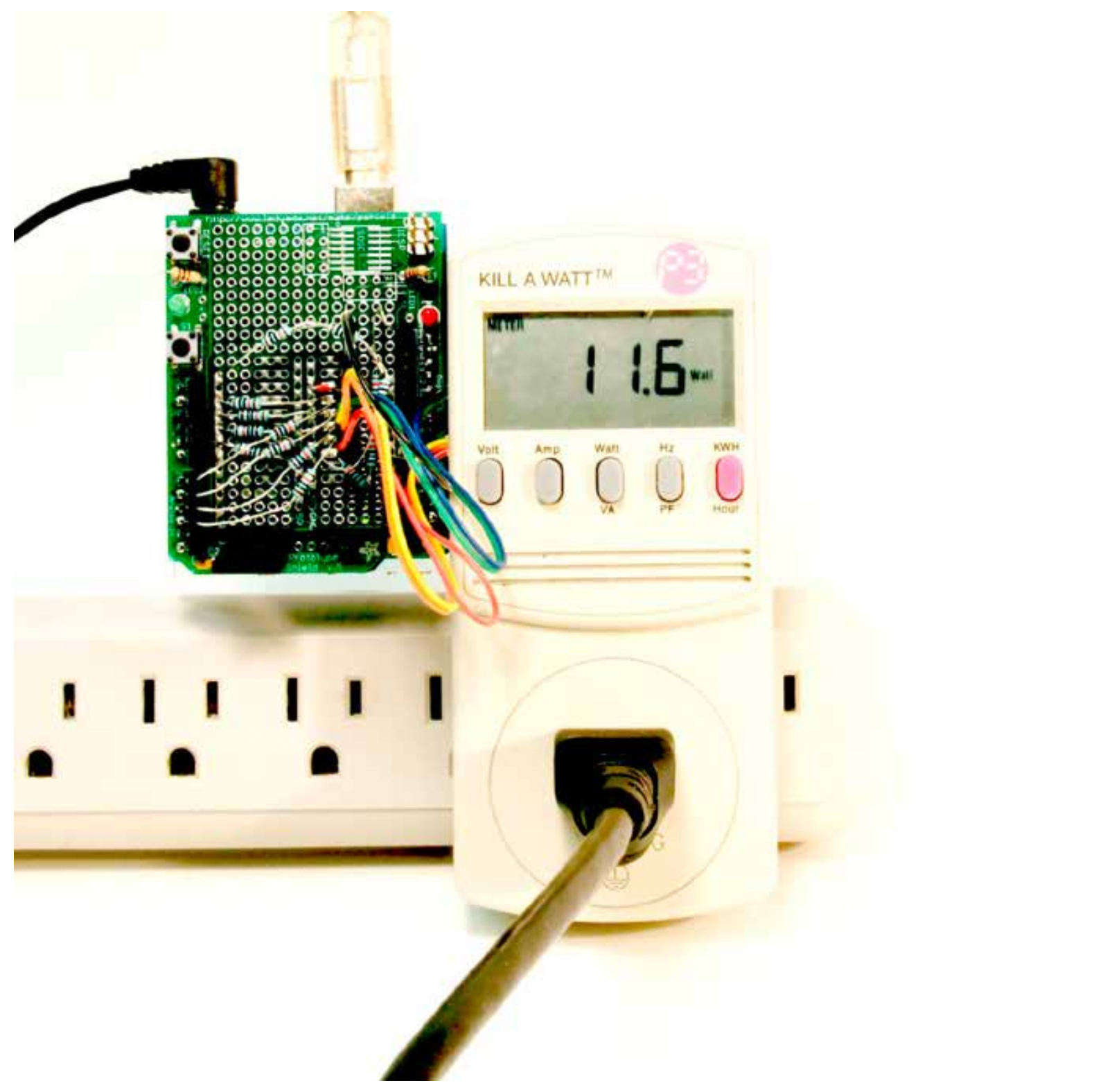


Figure 2: *modified Kill-A-Watt connected over FTDI/Serial to test machine. For more details on the Kill-A-Watt project, see http://github.com/madmaze/serialKAW*

To adequately stress each of the boards GPU performance, we formulated a simple GPU vector calculation kernel, containing a loop of *I* iterations with 16 useful Floating Point Operations. Every GPU thread executes multiple vector operations on 1 element of each of the two input array, then stores the solution into a result array. To ensure NVCC does not optimize out any of the above noted 16 Floating Point Operations, we examine the generated PTX code (see: Figure 4).

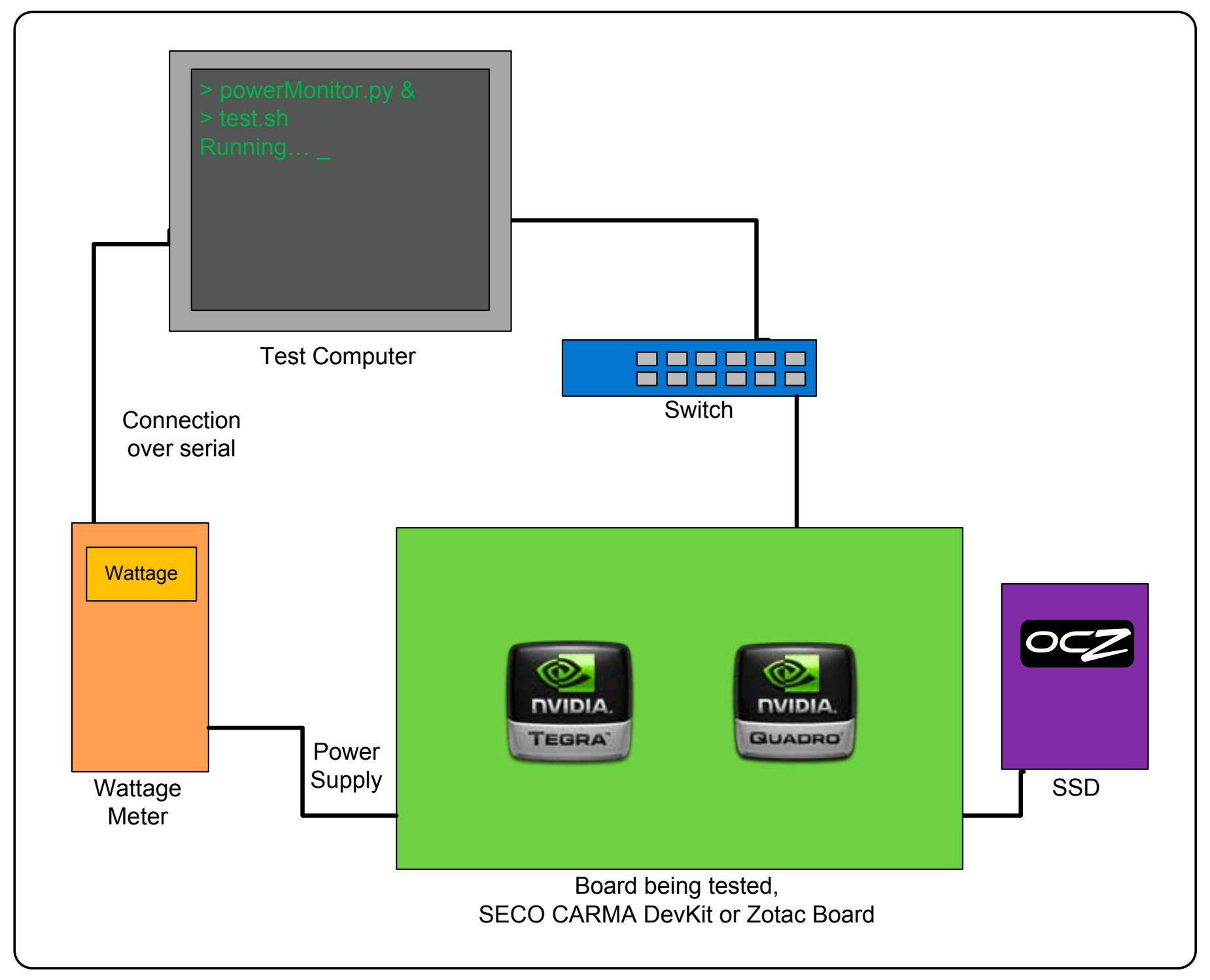
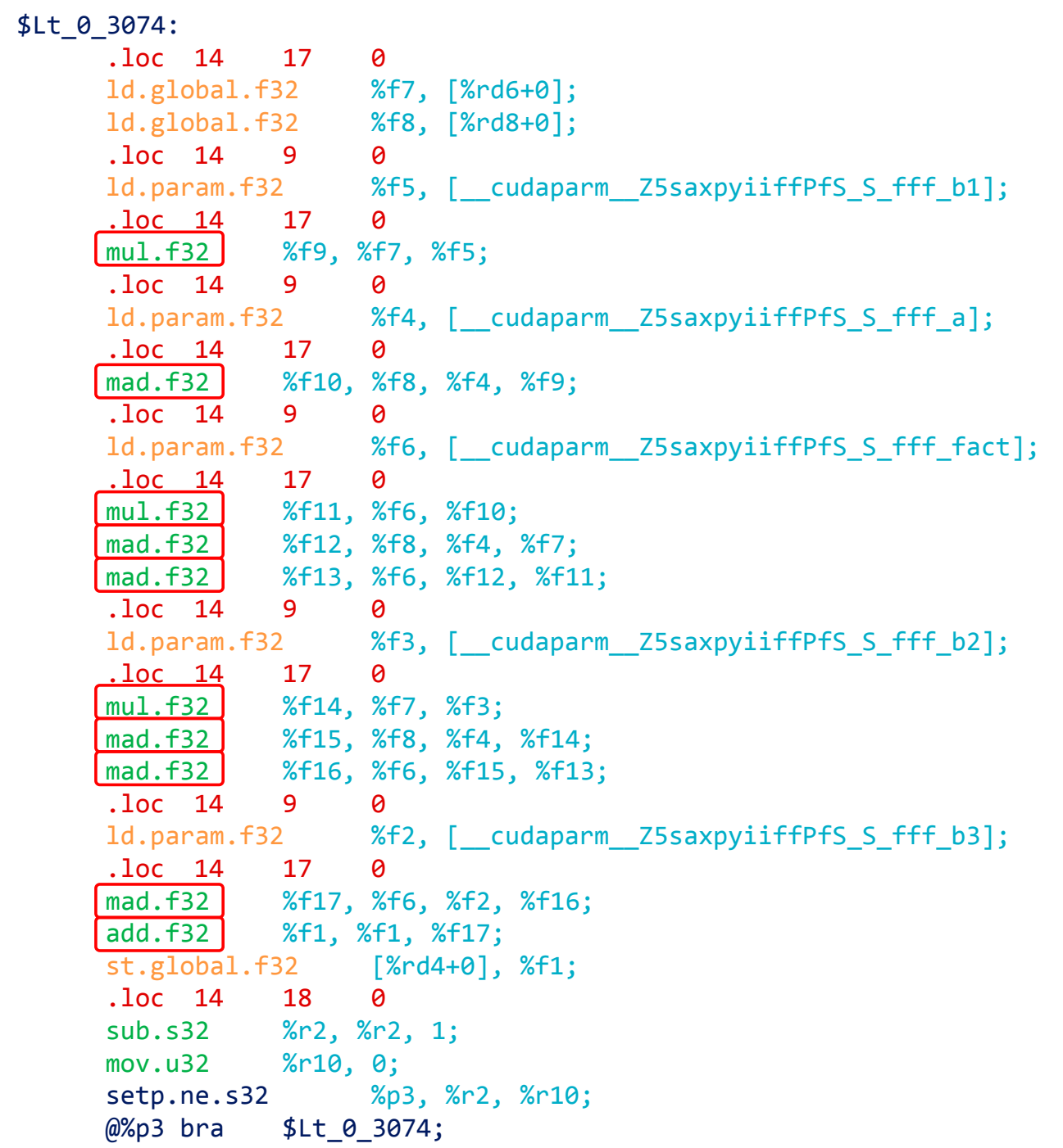


Figure 3: Testing setup

We initialize our input and output arrays with 2^24 elements and initialize *I* to 2^14. This gives us a grand total of 2^42 (2^24 x 2^14 x 2^4) Floating Point Operations. For power monitoring we use a P3 Kill-A-Watt P4400 which has been modified by adding 2 sense lines to the output of the P4400's main op-amp, which are read out over serial from an Arduino Duemilanove's Analog-Digital Converter(ADC). This setup allows us to automatically read, monitor and log power consumption during our test runs. During a test run, the test computer drives the test execution via SSH on the target board while also recording power consumption.



```
$Lt_0_3074:
        .loc   14    17    0
        ld.global.f32     %f7, [%rd6+0];
        ld.global.f32     %f8, [%rd8+0];
        .loc   14    9     0
        ld.param.f32      %f5, [__cudaparm__Z5saxpyiiffPfS_S_fff_b1];
        .loc   14    17    0
        mul.f32      %f9, %f7, %f5;
        .loc   14    9     0
        ld.param.f32      %f4, [__cudaparm__Z5saxpyiiffPfS_S_fff_a];
        .loc   14    17    0
        mad.f32      %f10, %f8, %f4, %f9;
        .loc   14    9     0
        ld.param.f32      %f6, [__cudaparm__Z5saxpyiiffPfS_S_fff_fact];
        .loc   14    17    0
        mul.f32      %f11, %f6, %f10;
        mad.f32      %f12, %f8, %f4, %f7;
        mad.f32      %f13, %f6, %f12, %f11;
        .loc   14    9     0
        ld.param.f32      %f3, [__cudaparm__Z5saxpyiiffPfS_S_fff_b2];
        .loc   14    17    0
        mul.f32      %f14, %f7, %f3;
        mad.f32      %f15, %f8, %f4, %f14;
        mad.f32      %f16, %f6, %f15, %f13;
        .loc   14    9     0
        ld.param.f32      %f2, [__cudaparm__Z5saxpyiiffPfS_S_fff_b3];
        .loc   14    17    0
        mad.f32      %f17, %f6, %f2, %f16;
        add.f32      %f1, %f1, %f17;
        st.global.f32     [%rd4+0], %f1;
        .loc   14    18    0
        sub.s32      %r2, %r2, 1;
        mov.u32      %r10, 0;
        setp.ne.s32       %p3, %r2, %r10;
        @%p3 bra     $Lt_0_3074;
$Lt_0_2562:
```

Figure 4: *PTX code for vector calculation kernel. Note the 10 highlighted floating point operations. 6x mad(2 FL-OP each), 3x mul(1 FL-OP each), 1x add(1 FL-OP each). The 1x sub at the end is not counted as it is the loop iterator and not a useful FL-OP*

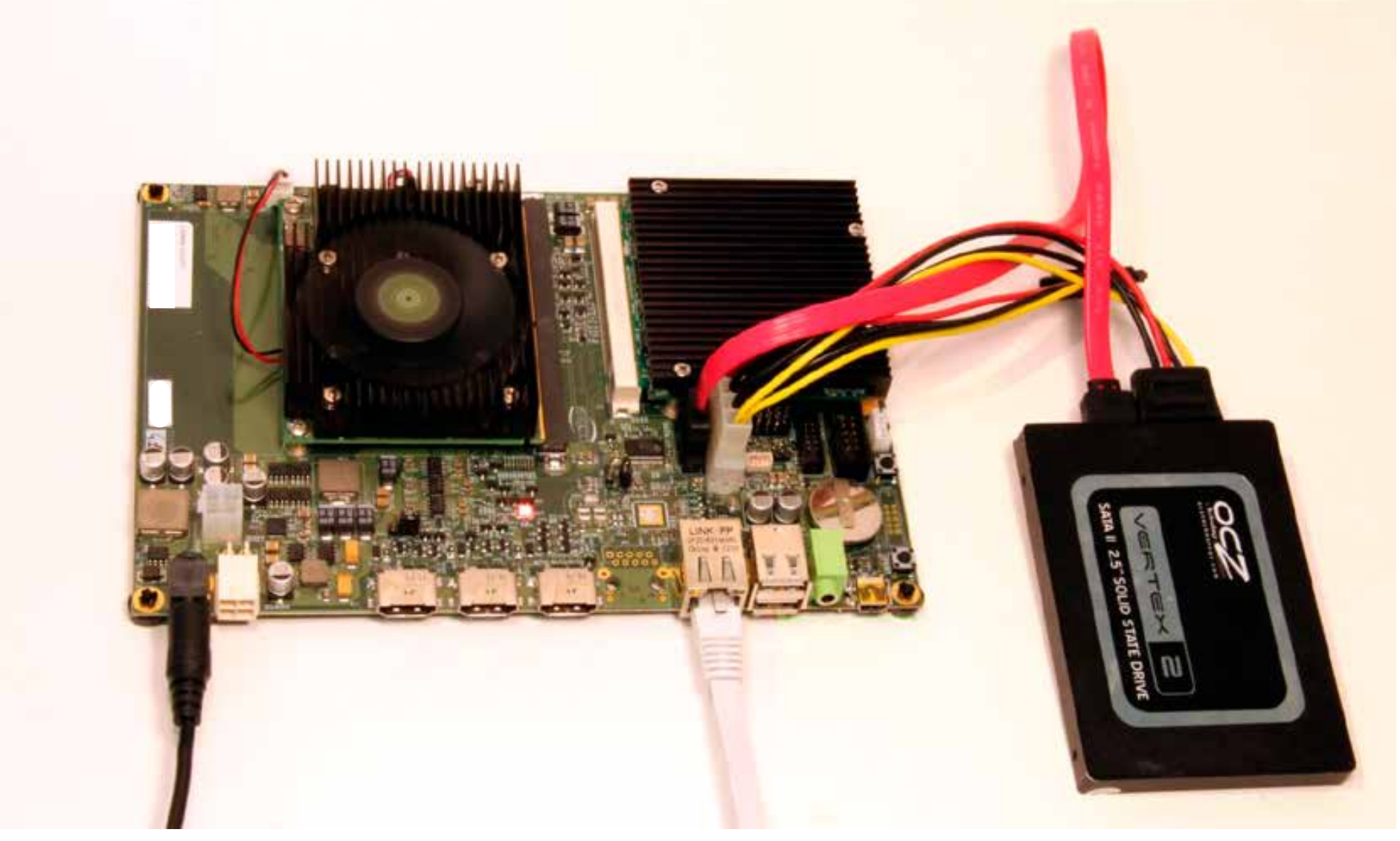


Figure 5: SECO CARMA DevKit setup during testing.

Results

Our test results indicate that SECO's CARMA DevKit provides a more than 10x speedup and an 8.65x increase in performance per watt. The DevKit idles at approximately 53% of the Zotac board's idle wattage. This gives an almost 47% increase in energy-efficiency. There is also a significant difference in the efficiency of the GPUs, the 1000M produces approximately 3.42 GFLOPS/Watt and the ION approximately 0.87 GFLOPS/Watt. It is important to note that these benchmark performance results are based on only one type of work load and these GPUs may be optimized differently, hence delivering different performance. It is apparent that the SECO's CARMA DevKit outshines the Zotac board in our benchmark, but some of this is to be expected as we are comparing boards and GPUs of different generations.

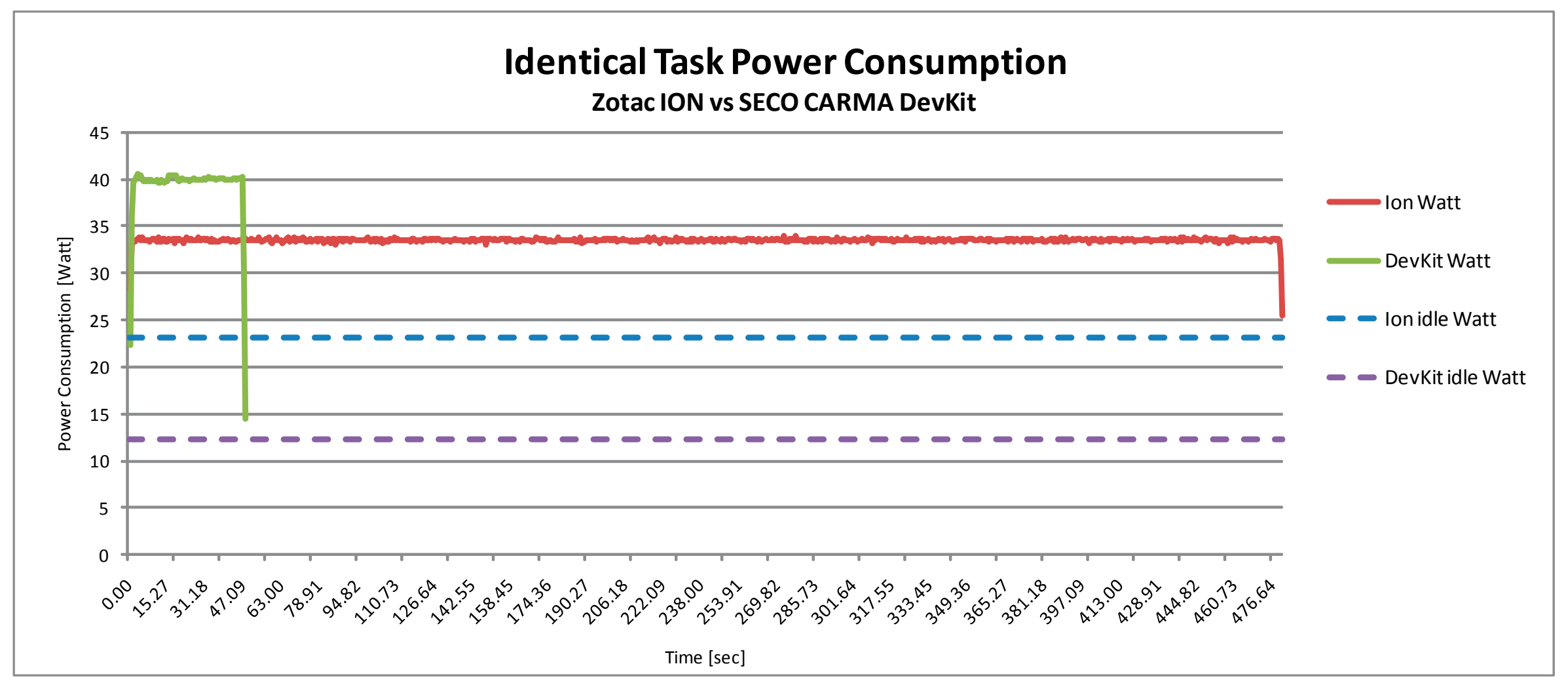


Figure 6: Power Consumption of the benchmark kernel, Note the difference in run length between the two boards.

Runtime and Performance of Benchmarks

Board	GPU	Total FLOPs	Time[s]	GFLOPS	Speedup
DevKit	1000M	2^42	47.65 sec	92.31	10.13x
Zotac	ION	2^42	482.81 sec	9.11	1.00x

Power Consumption before and during Benchmark

Board	Idle Watt	Peak Watt	Ave Load Watt	GFLOPS/Watt	Efficiency Factor
DevKit	12.24 W	40.55 W	39.26 W	2.35	8.65x
Zotac	23.05 W	33.99 W	33.50 W	0.27	1.00x

Estimated GPU Power Consumption: *[Ave(LoadWatt-IdleWatt)]*

Board	Workload Watt	GFLOPS/Watt	Efficency Factor
DevKit	27.02 W	3.42 W	3.92x
Zotac	10.45 W	0.87 W	1.00x

All of our test results and measurements were collected over a series of test runs with low relative standard deviations (RSD). For the measurement of kernel runtime, we observed a 0.0029% RSD for the SECO CARMA DevKit and 0.0304% RSD for the Zotac board. For the average power consumption, we observed a 5.97% RSD for the DevKit and a 1.125% RSD for the Zotac board.

References

[1] Szalay, A. S., Bell, G. C., Huang, H. H., Terzis, A., & White, A. (2010). Low-power amdahl-balanced blades for data intensive computing. ACM SIGOPS Operating Systems Review, 44(1), 71-75.

Contact:

Matthias A Lee
MatthiasLee@jhu.edu

Department of Computer Science
The Johns Hopkins University
3400 N Charles Street
Baltimore, MD 21238