A Sub 2 Watt 64-core 100 GFLOPS Accelerator Programmable in C/C++ or openCL

Adapteva, Inc.
1666 Massachusetts Ave, Lexington, MA, USA
Email: {andreas, roman, oleg, yaniv}@adapteva.com

Abstract

This demo will demonstrate Adapteva's latest 28nm floating point accelerator chip. The C/C++/OCL programmable chip contains 64 high performance RISC cores and operates at up to 800 MHz with a peak performance of over 100 GFLOPS. Demo applications will include an openCL based matrix multiplication example, an openCV based face detection application, compiled C-code signal processing kernels that reach 50-70% of peak, and a distributed 2D FFT application.

Epiphany Architecture Overview

The multi-core architecture in this work was designed to accelerate signal processing kernels requiring floating point math, such as large FFTs and matrix inversions. Examples of embedded applications requiring floating point math include: synthetic aperture radar, ultra sound, cellular antenna beam forming, and graphics processing. Figure 1 shows a block diagram of the architecture, with an example 16 processor tiles arranged as a 4 x 4 array.

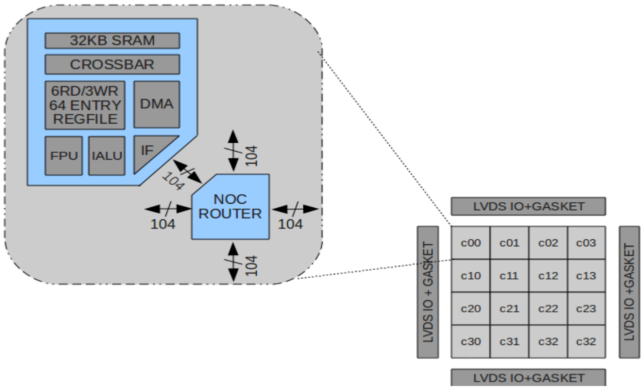

Figure 1: Multi-core Floating Point Accelerator Architecture

The tiles are connected through a 2D mesh network. Each processor tile contains a full routing cross-bar, a custom dual issue floating point RISC CPU, a DMA engine, and 32KB of multi-bank SRAM. All cores are ANSI-C programmable and share a single unified 32 bit flat address map. The processor cores can be programmed and run completely independently of each other or can work together to solve larger problems. An important architectural decision was the replacement of the traditional power hungry cache hierarchy with a distributed flat memory model that offers a total memory bandwidth of 32GB/s per processor core. The high bandwidth memory architecture and the flat unprotected 32 bit memory map lets up to 4096 cores communicate with each other directly with zero startup communication cost.

Chip Implementation

A chip product based on the above proposed architecture was implemented in a 28nm triple-Vt low power CMOS process, and contains 200 million transistors, staggered pad-ring wire bonding, and is packaged in a 324 ball 15x15mm BGA package. A chip photograph illustrating the chip packaging is shown in Figure 2.

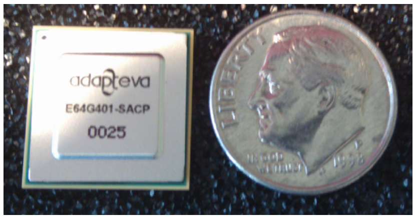

Figure 2: Photograph Illustrating Size (15 x 15mm)

Figure 3 shows the silicon evaluation platform for the 28nm chip. The platform was identical to a previously used 65nm platform allowing for characterization within a week of receiving the initial samples. The evaluation platform hooks up to a standard GNU debugger based tool chain running on a Linux distribution. The chip daughter card is located at the right side of the picture with two external power supplies, a 1V core supply and a 1.8V IO supply.

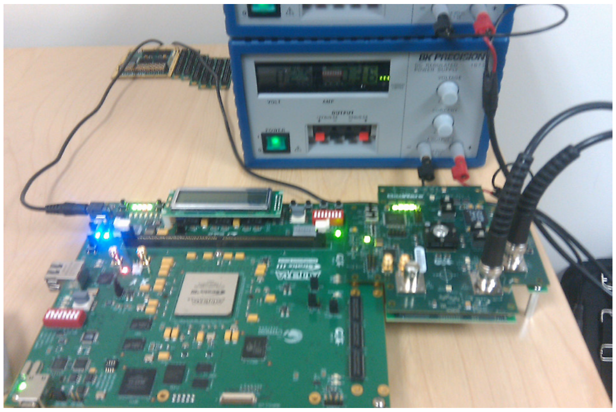

Figure 3: 28nm Silicon Evaluation Platform

Chip Performance Measurements

The following table shows the measured performance of the 64-core Epiphany chip. Power and performance numbers will demonstrated in real time during the demo using benchtop power supplies.

Metric	Measurement
Functional CPU cores	64
Peak Frequency	800MHz
Peak Performance	102.4 GFLOPS
Effective CPU Frequency	51.2 GHz
Core Efficiency (w/o IO)	72 GFLOPS/Watt, measured at 500MHz
Chip Max Power (with IO)	1.9W
Chip Standby Power	80mW