Inventing the Future of Computing

An Alternative to GPU Acceleration For Mobile Platforms

Andreas Olofsson
andreas@adapteva.com

50th DAC
June 5th, Austin, TX

adapteva
Adapteva Achieves 3 “World Firsts”

1. First commercial chip to reach 50 GFLOPS/W

2. First mobile processor with an open source OpenCL™ SDK

3. First semiconductor company to successfully crowd-source project
What is Adapteva

Company History:
• Fabless semiconductor company founded in 2008
• 16-core 65nm Epiphany-III chip product sampling since May 2011
• 64-core 28nm Epiphany-IV chip product sampling since July 2012
• Parallella open computing platform launched in October 2012

Notable Achievements:
• #1 in microprocessor energy efficiency
• 4 chips on $2.5M in raised capital
• $2M in total revenue to date
• 5K customers, 6,300 boards pre-sold
• 18 Patents pending
Our guiding light

Parallel

Efficient

Heterogeneous

Robust
Any Reason to Think the Future of Computing is NOT Parallel?

- No Computing
  - No Electronic Computing -1943

- Parallel Computing
  - “Von Neumann Age” Serial Computing 1943-2013?
  - Parallel Computing 2013-??
A Practical Start: True Heterogeneous Computing

SYSTEM-ON-CHIP

- BIG CPU
- Graphics
- Weird Math
- 1000’s of small RISC CPUs

O/S Application

- O/S
- BIG CPU
- Math

Copyright © Adapteva. All rights reserved.
The Accelerator Challenge

Status Quo Approach (~1.3X speedup)

1. Application
2. Move Data
3. Context Switch
4. Something Else
5. Context Switch
6. Move Data
7. Application
8. Move Data
9. Context Switch
10. Something Else
11. Context Switch
12. Move Data
13. Application

Smart Coprocessor (>10X speedup?)

1. Application
2. Move Data
3. Context Switch
4. Something Else
5. Context Switch
6. Move Data
7. Application
The Epiphany Coprocessor

- 1 GHz RISC Core
- Local Memory
- Multicore Framework
- Router
- 32-128KB Local Memory
- 1.6 GFLOPS Per Core @ ~25mW

Packet Based Network-On-Chip With 100GB/s Bisection BW

Coprocessor for ARM/x86 Host

<20pJ / FLOP!

MIMD/Task-Parallel Accelerator

Copyright © Adapteva. All rights reserved.
Epiphany-IV -- GLOBALFOUNDRIES 28SLP IP

- 64 CPUs
- IEEE Floating Point (SP)
- 800 MHz Max Frequency
- 100 GFLOPS Performance
- 6.4 GB/s IO BW
- 200 GB/s peak NOC BW
- 1.6 TB/sec on chip memory BW
- 25 Billion Messages/sec
- 2MB on chip memory
- 10 mm² total silicon area in 28nm
- 2 Watt total chip power
- 324 ball 15x15mm BGA
- Sampling since July, 2012
## Epiphany ANSI-C Benchmarks

<table>
<thead>
<tr>
<th>(Cycles)</th>
<th>Naïve C</th>
<th>Optimal C</th>
<th>Theoretical</th>
<th>C-Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>8x8 Matrix Multiplication</td>
<td>2852</td>
<td>773</td>
<td>512</td>
<td>66%</td>
</tr>
<tr>
<td>16 Tap FIR Filter (32 points)</td>
<td>1562</td>
<td>620</td>
<td>512</td>
<td>82%</td>
</tr>
<tr>
<td>Bi-quad IRR Filter (32 points)</td>
<td>n/a</td>
<td>991</td>
<td>768</td>
<td>77%</td>
</tr>
<tr>
<td>Dot-product (256 point)</td>
<td>800</td>
<td>557</td>
<td>256</td>
<td>49%</td>
</tr>
</tbody>
</table>

### Comparison to Servers

<table>
<thead>
<tr>
<th></th>
<th>Adapteva E64 800 MHz</th>
<th>Tilera GX36 1.4GHz</th>
<th>Intel Xeon L5640 2.2GHz</th>
<th>Nvidia Tegra-2 1GHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoreMark TM Score</td>
<td>77,912</td>
<td>165,276</td>
<td>118,571</td>
<td>5,866</td>
</tr>
<tr>
<td># Cores</td>
<td>64</td>
<td>36</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Power</td>
<td>2W</td>
<td>~30-50W</td>
<td>~50-100W</td>
<td>~1-2W</td>
</tr>
<tr>
<td>1024-Core Chip</td>
<td>2,493,184</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

---

*1 day per benchmark (compare to GPUs?)*

*Server Level Performance at 2Watts!!*
### Architecture Comparison

<table>
<thead>
<tr>
<th>Technology</th>
<th>FPGA</th>
<th>DSP</th>
<th>GPU</th>
<th>CPU</th>
<th>Epiphany</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>28nm</td>
<td>40nm</td>
<td>28nm</td>
<td>32nm</td>
<td>28nm</td>
</tr>
<tr>
<td>Programming</td>
<td>VHDL</td>
<td>OCL/C++/C</td>
<td>CUDA/OCL</td>
<td>OCL/C/C++</td>
<td>OCL/C/C++</td>
</tr>
<tr>
<td>Area (mm^2)</td>
<td>590</td>
<td>108</td>
<td>294</td>
<td>216</td>
<td>10</td>
</tr>
<tr>
<td>Chip Power (W)</td>
<td>40</td>
<td>22</td>
<td>135</td>
<td>130</td>
<td>2</td>
</tr>
<tr>
<td>“CPUs”</td>
<td>n/a</td>
<td>8</td>
<td>32</td>
<td>4</td>
<td>64</td>
</tr>
<tr>
<td>Max GFLOPS</td>
<td>1500</td>
<td>160</td>
<td>3000</td>
<td>115</td>
<td>102</td>
</tr>
<tr>
<td>GHz * Cores</td>
<td>n/a</td>
<td>12</td>
<td>32</td>
<td>14.4</td>
<td>51.2</td>
</tr>
<tr>
<td>Compile Time</td>
<td>Hours</td>
<td>Minutes</td>
<td>Minutes</td>
<td>Minutes</td>
<td>Minutes</td>
</tr>
<tr>
<td>L1 Memory</td>
<td>6MB</td>
<td>512KB</td>
<td>2.5MB</td>
<td>256KB</td>
<td>2MB</td>
</tr>
</tbody>
</table>

- Efficiency is everything
- Peak performance means very little
- No magic bullet!
Epiphany: A Truly Scalable Architecture

A Single Unified Instruction Set Architecture!

# Epiphany Cores

- 16
- 64
- 256
- 1,024
- 4,096
- 16,384

0.35W

1.4W

5.7W

23W

92W

G F L O P S

Copyright © Adapteva. All rights reserved.
How the @#$%^ Do We Program This Thing?
Epiphany Programming Models

**MODEL #1**
**DATA PARALLEL MODEL**
- openCL programmable
- Easy integration with C/C++
- openMP/MPI roadmap

**MODEL #2**
**WORKER BEE MODEL**
- Great for up to 2GFLOPS
- Supports standard C/C++
- “Cloud on a chip”

![Diagram of X86/ARM/FPGA Host with multiple MINI CPUs and interconnections](image-url)
Parallel Programming Frameworks

- Erlang
- SystemC
- Intel TBB
- Co-Fortran
- Lisp
- Janus
- Scala
- Haskell
- Pragmas
- Fortress
- Hadoop
- Linda
- Smalltalk
- CUDA
- Clojure
- UPC
- PVM
- Alef
- Julia
- OpenCL
- Go
- X10
- Posix
- XC
- Occam
- OpenHMPP
- ParaSail
- APL
- Simulink
- Charm++
- Occam-pi
- OpenMP
- Ada
- Labview
- Ptolemy
- StreamIt
- Verilog
- OpenACC
- C++ Amp
- Rust
- Sisal
- Star-P
- VHDL
- Cilk
- Chapel
- MPI
- MCAPI
- ??????????
Stupid Hurdles That Hinder Collaboration

- Proprietary SDKs and programming frameworks
- Lack of datasheets/documents
- Closed source drivers
- Expensive lock-in hardware
- NDA requirements
- Exclusive access

Open HW is now following the same successful path as open SW!
Parallella: Our “Secret Weapon”

• A $99 single board "parallel" computer that runs Linux
• Open source (SDK, board files, drivers) (github.com/parallella)
• Open documentation (adapteva.com/all-documents)
• Open to all (forums.parallella.org)
The Parallella Board

- Zynq dual core ARM- A9 (with FPGA Logic)
- 16-core Epiphany Coprocessor
- 1GB SDRAM
- Gigabit Ethernet
- uUSB
- uHDMI
- uSD
- 5V DC
- 1GB SDRAM

Copyright © Adapteva. All rights reserved.
Parallella Kickstarter Campaign

• 5,000 customers
• 6,300 boards "pre-sold" in 4 weeks
• 67 countries, all 50 US states
• 50-75% of backers are developers
• 12,000 more signups since Jan 1st
• Backer Application Interest:
  • Software Defined Radio
  • Ray tracing/rendering
  • Image processing
  • Robotics
  • Gaming

  • Cryptography
  • Parallel computing research
  • Distributed Computing
  • Machine Learning
  • HPC
Epiphany IP Conclusions

• #1 in processor energy efficiency at 70 GFLOPS/Watt (core)
• Silicon proven in GLOBALFOUNDRIES 28SLP node
• Only multicore IP that is scalable to 1000’s of cores on chip
• Easier to use than GPGPUs