On the Efficacy of a Fused CPU+GPU Processor (or APU) for Parallel Computing

Mayank Daga, Ashwin M. Aji, and Wu-chun Feng
Dept. of Computer Science
“Sampling” of fields that use GPUs

Mac OS X

Cosmology

Molecular Dynamics and Modeling

Computational Fluid Dynamics
### GPUs in HPC

<table>
<thead>
<tr>
<th>Rank</th>
<th>Computer</th>
<th>$R_{\text{max}}$</th>
<th>$R_{\text{peak}}$</th>
<th>%age ($R_{\text{max}}/R_{\text{peak}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>K computer</strong> – SPARC64 VIIIfx 2.0 GHz, Tofu Interconnect</td>
<td>8162.00</td>
<td>8773.00</td>
<td><strong>93.03 %</strong></td>
</tr>
<tr>
<td>2</td>
<td>Tianhe-1A - NUDT TH MPP, X5670 2.93Ghz 6C, NVIDIA GPU FT-1000 8C</td>
<td>2566.00</td>
<td>4701.00</td>
<td><strong>54.6 %</strong></td>
</tr>
<tr>
<td>3</td>
<td>Jaguar – Cray XT5-HE Opteron 6-core 2.6Ghz</td>
<td>1759.00</td>
<td>2331.00</td>
<td><strong>75.5 %</strong></td>
</tr>
<tr>
<td>4</td>
<td>Nebulae – Dawning TC6300 Blade, Intel X5650, NVIDIA Tesla C2050 GPU</td>
<td>1271.00</td>
<td>2984.30</td>
<td><strong>42.6 %</strong></td>
</tr>
<tr>
<td>5</td>
<td>TSUBAME 2.0 – HP ProLiant SL390s G7 Xeon 6C X5670, NVIDIA GPU</td>
<td>1192.00</td>
<td>2287.63</td>
<td><strong>52.1 %</strong></td>
</tr>
</tbody>
</table>

*Systems with GPUs achieve only ~50 % of $R_{\text{peak}}$*

*Systems without GPUs achieve ~84 % of $R_{\text{peak}}$*
Architecture of Discrete GPUs

SIMD Engines (~500 Gflop/s)

Thread Execution Control

Thread Processors

Thread Processors

Thread Processors

... 

... 

Device Memory

System Memory (Host)

X86 CPU Cores

DMA/PCIe
A Reason for Poor Efficiency

Symmetric Multi-Core (N-cores)

Overhead

Sequential Processor

Data Transfer Overhead

Accelerator-based System
A Reason for Poor Efficiency

Symmetric Multi-Core (N-cores)

FMAD

Single-core CPU

Multi-core CPU (4 cores)

Discrete GPU

Time (ms)

Serial Time  Parallel Time  Overhead

Data Transfer Overhead

Overhead

p/N

p/N

p/N

s

s

s

FMAD

Symmetric Multi-Core (N-cores)
Ideal Efficiency Scenario

Symmetric Multi-Core (N-cores)

Overhead

Sequential Processor

Data Transfer Overhead
Ideal Efficiency Scenario

Symmetric Multi-Core (N-cores)

Sequential Processor

Accelerator-based System

Overhead
Ideal Placement of CPU and GPU Cores

SIMD Engines

Thread Execution Control

Thread Processors

Thread Processors

Thread Processors

System Memory (Host)

X86 CPU Cores

Device Memory

DMA/PCIe
Towards a “fused” CPU+GPU…
Outline

• Motivation
• AMD Fusion APU – A Fused CPU+GPU
• Revisiting Amdahl’s Law
• Experimental Analysis
  – Application Benchmarks
  – Results and Discussion
• Conclusions and Future Work
AMD Fusion APU – A Fused CPU+GPU

- **High Performance Bus and Memory Controller**
  - X86 CPU Cores
  - System Memory
  - Platform Interfaces
  - SIMD Engines
    - Thread Execution Control
    - Thread Processors
      - Thread Processors
      - Thread Processors
      - ...
State of the Data Transfer

- Discrete GPU

- AMD Fusion APU (1\textsuperscript{st} Generation)

AMD provides high speed block transfer engines that move data between the x86 and SIMD memory partitions.

*(AMD, “AMD Fusion Family of APUs: Enabling a Superior, Immersive PC Experience”)*
Outline

• Motivation
• AMD Fusion APU – A Fused CPU+GPU
• Revisiting Amdahl’s Law
• Experimental Analysis
  – Application Benchmarks
  – Results and Discussion
• Conclusions and Future Work
Revisiting Amdahl’s Law

(M. Hill and M. Marty, “Amdahl’s Law in the Multi-core Era”)

Symmetric Multi-core

Asymmetric Multi-core

Speedup values for different serial fractions

Higher Efficiency of Asymmetric Chips
Revisiting Amdahl’s Law

\[ S = \frac{1}{s + \frac{p}{N}} \]

\[ S' = \frac{1}{s + p' + o} \]

Symmetric Multi-Core (N-cores)

Sequential Processor

Accelerator-based System
Revisiting Amdahl’s Law

Sequential Processor

Accelerator-based System

\[ S' = \frac{1}{s + p' + o} \]

- \( o_{\text{DiscreteGPU}} \) vs. \( o_{\text{Fusion}} \)
  - Fusion is expected to be better than discrete GPUs
- \( p'_{\text{DiscreteGPU}} \) vs. \( p'_{\text{Fusion}} \)
  - Depends on several factors, like algorithmic mapping, memory bandwidth, number of compute units, etc.
Implications

• Asymmetric chips always offer better efficiency than symmetric chips…
  – …if researchers continue to address scheduling and overhead challenges
• Fusing CPU and GPU cores reduce data transfer overheads to a great extent
• AMD Fusion, Intel Knights Ferry, and NVIDIA Tegra are all steps in the right direction.
  – Our focus today: AMD Fusion
Outline

• Motivation
• AMD Fusion APU – A Fused CPU+GPU
• Revisiting Amdahl’s Law
• Experimental Analysis
  – Application Benchmarks
  – Results and Discussion
• Conclusions and Future Work
Experimental Analysis

- Systems
  - AMD Zacate APU
    - Engineering sample of AMD Fusion
    - *Dual CPU cores + 80 GPU cores*
  - AMD Radeon HD 5870
    - High-powered discrete GPU
    - *1600 GPU cores*
  - AMD Radeon HD 5450
    - Low-powered discrete GPU
    - *80 GPU cores*
## Experimental Setup

<table>
<thead>
<tr>
<th>Platform</th>
<th>AMD Zacate APU</th>
<th>AMD Radeon HD 5870</th>
<th>AMD Radeon HD 5450</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stream Processors</td>
<td>80</td>
<td>1600</td>
<td>80</td>
</tr>
<tr>
<td>Compute Units</td>
<td>2</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Memory Bus Type</td>
<td>NA</td>
<td>GDDR5</td>
<td>DDR3</td>
</tr>
<tr>
<td>Device Memory</td>
<td>192 MB</td>
<td>1024 MB</td>
<td>512 MB</td>
</tr>
<tr>
<td>Local Memory</td>
<td>32 KB</td>
<td>32 KB</td>
<td>32 KB</td>
</tr>
<tr>
<td>Max. Workgroup Size</td>
<td>256 Threads</td>
<td>256 Threads</td>
<td>128 Threads</td>
</tr>
<tr>
<td>Peak Core Clock Freq.</td>
<td>492 MHz</td>
<td>850 MHz</td>
<td>675 MHz</td>
</tr>
<tr>
<td>Peak FLOPS</td>
<td>80 GFlop/s</td>
<td>2720 GFlop/s</td>
<td>104 GFlop/s</td>
</tr>
<tr>
<td><strong>Host:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Processor</td>
<td>AMD Engg. Sample @1.6 GHz</td>
<td>Intel Xeon E5405 @2.0 GHz</td>
<td>Intel Celeron 430 @1.8 GHz</td>
</tr>
<tr>
<td>System Memory</td>
<td>2 GB (NA)</td>
<td>2 GB DDR2</td>
<td>2 GB DDR2</td>
</tr>
<tr>
<td>Kernel</td>
<td>Ubuntu 2.6.35.22</td>
<td>Ubuntu 2.6.28.19</td>
<td>Ubuntu 2.6.32.24</td>
</tr>
</tbody>
</table>
Experimental Analysis

• Application Benchmarks
  – Bandwidth Test
    o Measures PCIe bandwidth for discrete GPU
    o Measures memory bandwidth for APU
  – FFT
    o Measures performance of a 2-D Fast Fourier Transform
    o Computes multiple FFTs of size 512 in parallel
  – MD
    o Measures performance of pairwise calculation of Lennard-Jones potential
  – Scan
    o Measures performance of the parallel prefix sum algorithm on a large array of floating point data
  – Reduction
    o Measures performance of a sum reduction operation using floating point data
Bandwidth Test

Host to Device

Device to Host

Bandwidth (GB/s)

Size (KB)

Zacate APU
Radeon HD 5870
Radeon HD 5450

Bandwidth (GB/s)

Size (KB)

Zacate APU
Radeon HD 5870
Radeon HD 5450
Fast Fourier Transform (FFT)

- APU reduces data transfer times for all problem sizes.
- Kernel Execution time is more for APU because of its lower memory bandwidth.

Problem Size (MB)

<table>
<thead>
<tr>
<th>Problem Size</th>
<th>AMD Zacate APU</th>
<th>AMD HD5450</th>
<th>AMD Radeon HD5870</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Time (ms)

- Data Transfer
- Kernel Execution
APU reduces data transfer times for all problem sizes.
The kernel executes fastest on discrete AMD 5870 due to more and faster GPU cores. The fused Zacate APU is next fastest.
Total execution time is equal for discrete and fused GPUs.
This is stunning given that discrete GPUs have 20-times more cores.
These cores are computationally more powerful as well.
Reduction

- Total execution time is 3-times better for the APU
- The efficacy of the APU increases as the problem size increases

![Graph showing the comparison between AMD Zacate APU and AMD Radeon HD5870 for different vector sizes](image)

- Data Transfer
- Kernel Execution

Vector Size (MB): 64, 32, 16, 8, 4

Time (ms): 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100

I/O-bound
Reduction

**Transfer Time**

- **Vector Size (MB)**
  - Time (ms)
  - AMD Fusion
  - AMD Radeon HD 5870

**Kernel Execution Time**

- **Vector Size (MB)**
  - Time (ms)
  - AMD Fusion
  - AMD Radeon HD 5870

**Total Execution Time**

- **Vector Size (MB)**
  - Time (ms)
  - 3x
  - AMD Fusion
  - AMD Radeon HD 5870
Outline

• Motivation
• AMD Fusion APU – A Fused CPU+GPU
• Revisiting Amdahl’s Law
• Experimental Analysis
  – Application Benchmarks
  – Results and Discussion
• Conclusions and Future Work
Future Work

• A more robust model also capturing the computational differences between fused and discrete GPUs

• Power modeling based on AMD’s Power Gating technology
Conclusions

• Fused CPU+GPU is a step in the right direction for efficient supercomputers
  – Data transfer overhead is largely mitigated (up to 6x)
  – Application execution time can be largely sped up (up to 3x in some cases)
  – No change is needed in the programming model

• But this is still not a panacea
  – GPU cores on the APU are not yet as powerful or as plentiful in number as the discrete GPUs
  – Device memory bandwidth does not yet match that of discrete GPUs

Questions?

Contacts
• Mayank daga (mdaga@cs.vt.edu)
• Ashwin M. Aji (aaji@cs.vt.edu)
• Dr. Wu-chun Feng (feng@cs.vt.edu)