Characterizing the Impact of Memory-Access Techniques on AMD Fusion

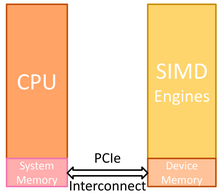
Kenneth Lee, Heshan Lin, Wu-chun Feng ({klee1,hlin2, feng}@cs.vt.edu) | Department of Computer Science

VirginiaTech

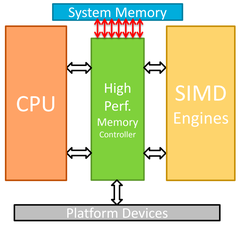
SyNeRGy http://synergy.cs.vt.edu

1. Introduction & Motivation

Discrete GPU + CPU



Fused GPU + CPU



	AMD Llano	AMD HD 5450
GPU Core Frequency	660 MHz	650 MHz
Device Memory Speed	DDR3 - 1066	DDR3 - 800
Data Theoretical Maximum Speed	>> 4GB/s	4GB/s
GPU Cores	80	80

Research Questions

1. **How can we leverage Fusion's high data transfer speed?**
2. **When does fused outperform discrete architecture?**
3. **How does data reuse impact performance of data accesses?**

Approach

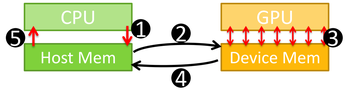
- Use micro-benchmarks to understand performance of different memory movement techniques
- Apply the data from the micro-benchmarks to improve memory performance for discrete and fused GPUs

Applications

- **Vector Add**
  - Computes the sum C = A + B for vectors A, B, and C.
- **Scan**
  - Calculates the Parallel Prefix Sum for a given vector V.
- **Reduce**
  - Performs a sum reduction on a vector of floating-point numbers
- **Cyclic Redundancy Check**
  - Computes the remainder of a polynomial division A/B, where A and B are arrays of bits.

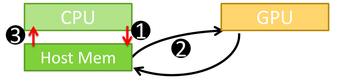
2. Memory Movement Techniques

Default



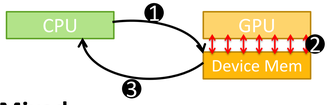
1.CPU writes data to host memory.
2.Data is copied to device memory
3.Kernel execution
4.Data is copied back to host
5.CPU reads data from host memory

CPU-Resident



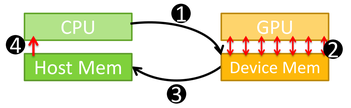
1.CPU writes data to host memory.
2.Kernel execution (data is read and written over the interconnect).
3. CPU reads data from host memory

GPU-Resident



1.CPU writes data directly to device memory.
2.Kernel execution
3.CPU reads data from device memory, via the interconnect.

Mixed



1.CPU writes data directly to device memory.
2.Kernel execution
3.Data is copied from device to host memory.
4.CPU reads host memory.

3. Micro-Benchmark Results

	AMD Llano	AMD HD 5450
Host Buffer → Device Buffer	2.61 GB/s	3.04 GB/s
Host Buffer ← Device Buffer	3.17 GB/s	3.30 GB/s
CPU ← Host Buffer (Read)	5.67 GB/s	5.67 GB/s
CPU → Host Buffer (Write)	5.46 GB/s	5.46 GB/s
GPU← Host Buffer (Read)	16.26 GB/s	5.03 GB/s
GPU→ Host Buffer (Write)	4.96 GB/s	5.25 GB/s
CPU ← Device Buffer (Read)	**0.01** GB/s	**0.01** GB/s
CPU → Device Buffer (Write)	7.49 GB/s	5.45 GB/s
GPU← Device Buffer (Read)	17.54 GB/s	5.38 GB/s
GPU→ Device Buffer (Write)	14.31 GB/s	5.34 GB/s

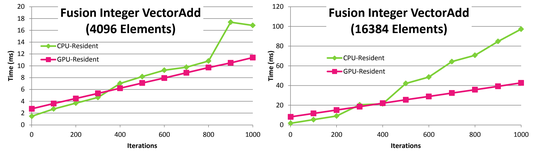
Results taken with BufferBandwidth, included in the AMD OpenCL SDK

4. Application Results

Data Reuse on CPU & GPU Resident Memory

Goal

- Understand impact of data reuse for CPU and GPU Resident memories
- Increase the data reuse by increase the number of kernel iterations

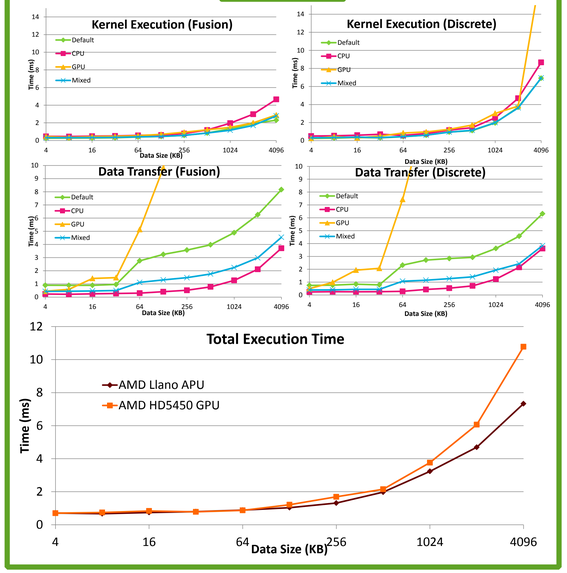


Conclusions

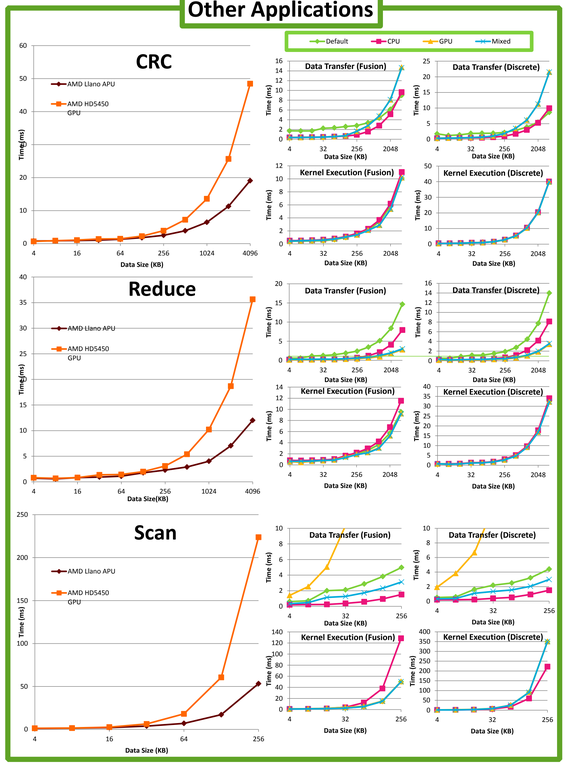
- Low Data Reuse favors CPU-Resident memory
- Small Device to Host data transfer favors GPU-Resident memory

Comparison of Fused vs. Discrete Architectures

Vector Add



Other Applications



Conclusions

- Fusion is capable of achieving better performance than a discrete GPU when the data transfer to computation ratio is high.
- Decreasing device-to-host data transfers improves Fusion's performance over discrete GPUs.

5. Future Work

- Construct a model to predict performance based on data movement model.
- Construct models to compare performance of applications run on the Discrete CPU/GPU vs. APU.