

UirginiaTech



MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE

VIGNESH ADHINARAYANAN, INDRANI PAUL, JOSEPH L. GREATHOUSE, WEI HUANG, ASHUTOSH PATTNAIK, WU-CHUN FENG

- Power and energy have become a first-class design constraint in all areas of computing
 - Phones and laptops must optimize for energy as they are battery operated
 - Desktops require loud cooling solutions
 - Supercomputers and other large data centers are constrained by power



POWER AND ENERGY ARE FIRST-CLASS DESIGN CONSTRAINTS

- Power and energy have become a first-class design constraint in all areas of computing
 - Phones and laptops must optimize for energy as they are battery operated
 - Desktops require loud cooling solutions
 - Supercomputers and other large data centers are constrained by power



Modified from <u>Image by Vernon Chan</u> <u>CC BY 2.0</u>



POWER AND ENERGY ARE FIRST-CLASS DESIGN CONSTRAINTS

- Power and energy have become a first-class design constraint in all areas of computing
 - Phones and laptops must optimize for energy as they are battery operated
 - Desktops require loud cooling solutions
 - Supercomputers and other large data centers are constrained by power



Modified from <u>Image by Vernon Chan</u> CC BY 2.0



Image by Glenn Batuyong <u>CC BY 2.0</u>





4 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

POWER AND ENERGY ARE FIRST-CLASS DESIGN CONSTRAINTS

- Power and energy have become a first-class design constraint in all areas of computing
 - Phones and laptops must optimize for energy as they are battery operated
 - Desktops require loud cooling solutions
 - Supercomputers and other large data centers are constrained by power



Modified from <u>Image by Vernon Chan</u> CC BY 2.0







Image by Patrick Strandberg <u>CC BY-SA 2.0</u>



5 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

DATA MOVEMENT: A MAJOR SOURCE OF POWER CONSUMPTION AMD

Data movement through the memory hierarchy thought to be a major source of power consumption



Based on: Shalf et al., "Exascale Computing Technology Challenges," VECPAR 2010

Measurements on real system lacking

- Current estimates based on simulation studies
- Real-world measurements are coarse-grained and do not give break down for data movement



THIS TALK IN ONE SENTENCE

A new methodology to measure data-movement power on real hardware



MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016 7 |



8 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

















G. Kestor et al., "Quantifying the energy cost of data movement in scientific applications," IISWC 2013

 \blacksquare E_{L2} = Energy consumed by L2 microbenchmark





G. Kestor et al., "Quantifying the energy cost of data movement in scientific applications," IISWC 2013

E_{L2} = Energy consumed by L2 microbenchmark





- \blacksquare E_{L2} = Energy consumed by L2 microbenchmark
- E_{L1} = Energy consumed by L1 microbenchmark





G. Kestor et al., "Quantifying the energy cost of data movement in scientific applications," IISWC 2013

- E_{L2} = Energy consumed by L2 microbenchmark
- E_{L1} = Energy consumed by L1 microbenchmark
- Energy cost of moving data from L2 to L1 = $E_{L2} E_{L1}$





G. Kestor et al., "Quantifying the energy cost of data movement in scientific applications," IISWC 2013

- E_{L2} = Energy consumed by L2 microbenchmark
- E_{L1} = Energy consumed by L1 microbenchmark
- Energy cost of moving data from L2 to L1 = $E_{L2} E_{L1}$
- ▲ Issue: Over-estimation of data-movement energy



STATE-OF-THE-ART MEASUREMENT APPROACH: LIMITATION

ComputeComputeL1L1InterconnectInterconnectL2L2

L2 microbenchmark

L1 microbenchmark

Issue: Data-movement power also includes L2access power

17 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

UrginiaTech



OUR PROPOSED APPROACH 4



Representative block diagram of AMD FirePro[™] W9100 GPU (Previously Code-named "Hawaii")





Representative block diagram of AMD FirePro[™] W9100 GPU (Previously Code-named "Hawaii")

Distance between shader engines and L2 banks differ





Representative block diagram of AMD FirePro[™] W9100 GPU (Previously Code-named "Hawaii")

Distance between shader engines and L2 banks differ





Representative block diagram of AMD FirePro[™] W9100 GPU (Previously Code-named "Hawaii")

Distance between shader engines and L2 banks differ





Representative block diagram of AMD FirePro[™] W9100 GPU (Previously Code-named "Hawaii")

Physical distance traversed by data thought to affect data movement power



UirginiaTech

OUR PROPOSED APPROACH





Short-path microbenchmark



Long-path microbenchmark

Design microbenchmarks based on data-movement distance to properly isolate the interconnect





CHALLENGES





UirginiaTech



OpenCL[™] lacks native support to pin threads to programmer-specified compute units



OpenCL[™] lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption



OpenCL[™] lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe







UirginiaTech

Only a small difference between used-L2 and L1 cache size which can make it challenging to write L2-only microbenchmarks





Only a small difference between used-L2 and L1 cache size which can make it challenging to write L2-only microbenchmarks

> Ensuring that the same amount of work is done by the two microbenchmarks can be challenging due to NUCA effects





OpenCL[™] lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe



Addressing the challenges PINNING THREADS TO CORES WITH OPENCL[™]



(a	ι)	Initial	0	penCL	code	snippet
----	----	---------	---	-------	------	---------

s_cbranch_scc0 label_0011 // 00000000030: BF840004	<pre>s_min_u32</pre>	<pre>s0, s0, 0x0000ffff s0, s16, s0 s0, s0, s1 v0, vcc, s0, v0 s0, s16, s2 s0, -1 label_0011</pre>	 	00000000014: 0000000001C: 00000000020: 00000000024: 00000000028: 00000000022C: 00000000030:	8380FF00 9300010 81000100 4A000000 81000210 BF02C100 BF840004	0000FFFF
--	----------------------	--	------------------------	---	--	----------

(b) Equivalent assembly code

00	$\mathbf{F}\mathbf{F}$	80	83	$\mathbf{F}\mathbf{F}$	$\mathbf{F}\mathbf{F}$	00	00
10	00	00	93	00	01	00	81
00	00	00	4A	10	02	00	81
00	C1	02	BF	04	00	84	BF
00	FF	04	BF	81	00	00	00
C1	80	01	85	01	00	82	BF

(c) Equivalent binary (in hex)

 00
 FF
 80
 83
 FF
 FF
 00
 00

 10
 00
 00
 93
 00
 01
 00
 81

 00
 00
 00
 4A
 04
 32
 00
 B9

 00
 C1
 02
 BF
 04
 00
 84
 BF

 00
 FF
 04
 BF
 81
 00
 00
 00

 C1
 80
 01
 85
 01
 00
 82
 BF

(d) Modified binary (in hex)

(e) Equivalent OpenCL code

Can be accomplished with some binary hacking

UrginiaTech



ADDRESSING THE CHALLENGES PINNING THREADS TO CORES WITH OPENCL[™]



(a) Initial OpenCL code snippet

s_min_u32 s_mu1_i32 s_add_i32 v_add_i32 s_add_i32	<pre>s0, s0, 0x0000ffff s0, s16, s0 s0, s0, s1 v0, vcc, s0, v0 s0, s16, s2</pre>	<pre>// 00000000014: // 0000000001C: // 00000000020: // 00000000024: // 00000000028:</pre>	8380FF00 93000010 81000100 4A000000 81000210	0000FFFF
s_cmp_gt_i32	s0, -1	// 0000000002C:	BF02C100	
s_cbranch_scc(0 label_0011	// 000000000030:	BF840004	

(b) Equivalent assembly code

00	$\mathbf{F}\mathbf{F}$	80	83	$\mathbf{F}\mathbf{F}$	$\mathbf{F}\mathbf{F}$	00	00	
10	00	00	93	00	01	00	81	
00	00	00	4A	10	02	00	81	
00	C1	02	BF	04	00	84	BF	
00	FF	04	BF	81	00	00	00	
C1	80	01	85	01	00	82	BF	

(c) Equivalent binary (in hex)

00 FF 80 83 FF FF 00 00 10 00 00 93 00 01 00 81 00 00 00 4A 04 32 00 B9 00 C1 02 BF 04 00 84 BF 00 FF 04 BF 81 00 00 00 C1 80 01 85 01 00 82 BF

(d) Modified binary (in hex)

kernel void 12 read(global float *data, global float *output) { int gid = get global id(0); int cu id = get cu id(0); if (cu id >= 0 && cu id <= 10) { // Read data from L2

(e) Equivalent OpenCL code

1. Use workgroup ID as a placeholder for CU ID

🛄 Virginia lech Invent the Future



Addressing the challenges PINNING THREADS TO CORES WITH OPENCL[™]



(a) Initial OpenCL code snippet

// 00000000014. 8380FF00 0000FF
// 000000001C: 93000010
// 00000000020: 81000100
// 00000000024: 4A000000
// 0000000028: 81000210
// 0000000002C: BF02C100
(/ 000000000000000000000000000000000000

(b) Equivalent assembly code

kernel void 12 read(global float *data,

00	$\mathbf{F}\mathbf{F}$	80	83	$\mathbf{F}\mathbf{F}$	$\mathbf{F}\mathbf{F}$	00	00
10	00	00	93	00	01	00	81
00	00	00	4A	10	02	00	81
00	C1	02	BF	04	00	84	BF
00	FF	04	BF	81	00	00	00
C1	80	01	85	01	00	82	BF

(c) Equivalent binary (in hex)

(d) Modified binary (in hex)

00 FF 80 83 FF FF 00 00

10 00 00 93 00 01 00 81

00 00 00 4A 04 32 00 B9

00 C1 02 BF 04 00 84 BF

00 FF 04 BF 81 00 00 00 c1 80 01 85 01 00 82 BF

(e) Equivalent OpenCL code

2. Identify instruction that writes workgroup ID to the register that gets checked in the conditional

36 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

UirginiaTech



global float *output) {


Addressing the challenges PINNING THREADS TO CORES WITH OPENCL[™]



(a) Initial OpenCL code snippet

		(/ 0000000014-	0.20077700	000000000
s_min_u3z	SU, SU, UXUUUUIIII	// 0000000014:	83801100	00001111
s_mul_i32	s0, s16, s0	// 0000000001C:	93000010	
s add i32	s0, s0, s1	// 00000000020:	81000100	
v_add_i32	v0, vcc, s0, v0	// 00000000024:	4A000000	
s_add_i32	s0, s16, s2	// 00000000028:	81000210	
s_cmp_gt_i32	s0, -1	// 0000000002C:	BF02C100	
s_cbranch_scc	0 label_0011	// 00000000030:	BF840004	

(b) Equivalent assembly code

00	$\mathbf{F}\mathbf{F}$	80	83	$\mathbf{F}\mathbf{F}$	$\mathbf{F}\mathbf{F}$	00	00
10	00	00	93	00	01	00	81
00	00	00	4A	10	02	00	81
00	C1	02	BF	04	00	84	BF
00	FF	04	BF	81	00	00	00
C1	80	01	85	01	00	82	BF

(c) Equivalent binary (in hex)

(d) Modified binary (in hex)

00 FF 80 83 FF FF 00 00

10 00 00 93 00 01 00 81

00 00 00 4A 04 32 00 B9

00 C1 02 BF 04 00 84 BF

00 FF 04 BF 81 00 00 00 c1 80 01 85 01 00 82 BF (e) Equivalent OpenCL code

3. Identify instruction that writes workgroup ID to the register that gets checked in the conditional in binary



Addressing the challenges PINNING THREADS TO CORES WITH OPENCL[™]



(a) Initial	OpenCL	code	snippe	t
---	---	-----------	--------	------	--------	---

s_min_u32 s	s0, s0, 0x0000ffff	- / /	000000000014:	8380FF00	0000FF
s_mul_i32 s	s0, s16, s0	//	0000000001C:	93000010	
s add i32 s	s0, s0, s1	//	000000000020:	81000100	
v_add_i32 v	70, vcc, s0, v0	//	00000000024:	4A000000	
s_add_i32 s	:0, s16, s2	11	00000000028:	81000210	
s cmp gt i32 s	s0, -1	//	0000000002C:	BF02C100	
s_cbranch_scc0	label 0011	//	00000000030:	BF840004	
	—				

(b) Equivalent assembly code

00	$\mathbf{F}\mathbf{F}$	80	83	FF	$\mathbf{F}\mathbf{F}$	00	00
10	00	00	93	00	01	00	81
00	00	00	4A	10	02	00	81
00	C1	02	BF	04	00	84	BF
00	FF	04	BF	81	00	00	00
C1	80	01	85	01	00	82	BF

(c) Equivalent binary (in hex)

(d) Modified binary (in hex)

00 FF 80 83 FF FF 00 00

10 00 00 93 00 01 00 81

00 00 00 4A 04 32 00 B9

00 C1 02 BF 04 00 84 BF

00 FF 04 BF 81 00 00 00 C1 80 01 85 01 00 82 BF (e) Equivalent OpenCL code

4. Replace workgroup ID with CU ID (read from readonly HW register) in the correct register in the binary

38 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

UirginiaTech



Addressing the challenges PINNING THREADS TO CORES WITH OPENCL[™]



(a) Initial OpenCL code snippet

s_min_u32	s0, s0, 0x0000ffff	// 00000000014: 8380FF	00 0000FFFF
s mul i32	s0, s16, s0	/ 0000000001C: 930000	10
s add i32	s0, s0, s1	/ 00000000020: 810001	0.0
v_add_i32	v0, vcc, s0, v0	/ 00000000024: 4A0000	0 0
s add i32	s0, s16, s2	/ 00000000028: 810002	10
s cmp gt i32	s0, -1	/ 0000000002C: BF02C1	00
s_cbranch_scc	0 label 0011	/ 00000000030: BF8400	04
	—		

(b) Equivalent assembly code

00	$\mathbf{F}\mathbf{F}$	80	83	FF	$\mathbf{F}\mathbf{F}$	00	00
10	00	00	93	00	01	00	81
00	00	00	4A	10	02	00	81
00	C1	02	BF	04	00	84	BF
00	FF	04	BF	81	00	00	00
C1	80	01	85	01	00	82	BF

(c) Equivalent binary (in hex)

 00
 FF
 80
 83
 FF
 FF
 00
 00

 10
 00
 00
 93
 00
 01
 00
 81

 00
 00
 00
 4A
 04
 32
 00
 B9

 00
 C1
 02
 BF
 04
 00
 84
 BF

 00
 FF
 04
 BF
 81
 00
 00
 00

 C1
 80
 01
 85
 01
 00
 82
 BF

(d) Modified binary (in hex)

(e) Equivalent OpenCL code

5. Get functionally equivalent OpenCL snippet

UrginiaTech





OpenCL[™] lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe



ADDRESSING THE CHALLENGES

ELIMINATING TEMPERATURE EFFECTS





ADDRESSING THE CHALLENGES

ELIMINATING TEMPERATURE EFFECTS



Solution 1: Run GPU fans at very high speed to limit temperature difference between runs

42 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016



ADDRESSING THE CHALLENGES

ELIMINATING TEMPERATURE EFFECTS



Solution 2: Model idle power separately and subtract from measured power

UrginiaTech



OpenCL[™] lacks native support to pin threads to programmer-specified compute units

Temperature of device during tests will affect the power consumption

Power difference between the two sets of microbenchmarks can be hard to observe





- Power difference between microbenchmarks can be too low to be reliably observed
 - Unless the amount of data going through the interconnect increases





- Power difference between microbenchmarks can be too low to be reliably observed
 - Unless the amount of data going through the interconnect increases
- Difficult to saturate interconnect bandwidth without increasing the number of wavefronts
 - Fewer wavefronts can result in stalling exposing latency difference issues





- Power difference between microbenchmarks can be too low to be reliably observed
 - Unless the amount of data going through the interconnect increases
- Difficult to saturate interconnect bandwidth without increasing the number of wavefronts
 - Fewer wavefronts can result in stalling exposing latency difference issues
- More wavefronts can lead to one of the following issues
 - More L1 hits when accesses per thread is low
 - Register pressure and memory spills if access per thread is high





- Power difference between microbenchmarks can be too low to be reliably observed
 - Unless the amount of data going through the interconnect increases
- Difficult to saturate interconnect bandwidth without increasing the number of wavefronts
 - Fewer wavefronts can result in stalling exposing latency difference issues
- More wavefronts can lead to one of the following issues
 - More L1 hits when accesses per thread is low
 - Register pressure and memory spills if access per thread is high
- Solution: Modify firmware to artificially shrink L1 cache size and then increase number of wavefronts

CHALLENGES

Only a small difference between used-L2 and L1 cache size which can make it challenging to write L2-only microbenchmarks

> Ensuring that the same amount of work is done by the two microbenchmarks can be challenging due to NUCA effects

Additional details in the paper

49 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016



CHARACTERIZATION STUDIES 4



51 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016









Average interconnect distance

Data toggle rate



53 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

Average interconnect distance

Data toggle rate

Bandwidth observed on the interconnect



54 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

Average interconnect distance

Data toggle rate

Bandwidth observed on the interconnect

Interconnect's voltage and frequency

55 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

UrginiaTech

INTERCONNECT POWER VS DISTANCE

56



Our experiments confirm that the distance traversed by data affects power consumption



UirginiaTech

INTERCONNECT POWER VS DISTANCE

57





Within 15% of industrial estimates for energy/bit/mm





INTERCONNECT POWER VS DISTANCE



Linear relationship observed between datamovement distance and interconnect power





INTERCONNECT POWER VS TOGGLE RATE



Linear relationship observed between toggle rate and interconnect power



INTERCONNECT POWER VS TOGGLE RATE





UrginiaTech

INTERCONNECT POWER VS TOGGLE RATE



Transmitting 0s slightly more expensive than 1s



INTERCONNECT POWER VS VOLTAGE AND FREQUENCY



Impact of voltage and frequency as expected



PUTTING IT ALL TOGETHER

- Interconnect Power = Energy/bit/mm * avg. distance * avg. bits/sec * scaled voltage² x scaled frequency * avg. toggle rate
- ▲ For real applications,
 - Avg. bits per second is calculated from performance counters (e.g. L1 and L2 accesses)
 - The other parameters are pre-computed or pre-measured (e.g. Distance is measured from layout; energy/bit/mm is computed to be 110 fJ/bit/mm)

EVALUATION OF APPLICATIONS







Up to 14% dynamic power spent on interconnects in today's chips

66 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

UrginiaTech





L2-MC (7nm) ■ L1-L2 (7nm) ■ Reg-L1 (7nm) ■ L2-MC (28nm) ■ L1-L2 (28nm) ■ Reg-L1 (28nm)

Even non-memory bound applications can show high interconnect power (but at different hierarchy)

67 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016

PennState

🛄 Virginia lect



L2-MC (7nm) L1-L2 (7nm) Reg-L1 (7nm) L2-MC (28nm) L1-L2 (28nm) Reg-L1 (28nm)

Data-movement power problem exacerbated in future technology nodes



UirginiaTech RennState

OPTIMIZATIONS



Layout-Based Optimizations

- We explore how much impact layout-based optimization can have on real world applications
- We examine two layouts one reduces distance between L1 and L2 and the other reduces distance between L2 and memory controller

Cache Resizing (details in the paper)

- Increasing cache size decreases average data movement distance (most data is fetched from nearer memories)
- We quantify the magnitude of difference when we increase the cache size 4 times



INTERCONNECT POWER OPTIMIZATIONS – LAYOUT OPTIMIZATION **AMD**





INTERCONNECT POWER OPTIMIZATIONS – LAYOUT OPTIMIZATION **AMD**



72 | MEASURING AND MODELING ON-CHIP INTERCONNECT POWER ON REAL HARDWARE | SEPTEMBER 26, 2016


INTERCONNECT POWER OPTIMIZATIONS – LAYOUT OPTIMIZATION **AMD**



Interconnect power reduces by 48% on average when we use an L1-L2 distance optimized layout



- Distance-based microbenchmarking is a promising approach to measure data movement power
- Over 14% of dynamic power can go towards on-chip data movement in today's chips
 - Lesser than past estimates as we separated out data access power from data movement power
 - Can increase to 22% by 7nm technology
- Optimizing on-chip interconnects to reduce the distance of frequently accessed portions can reduce on-chip data movement power by 48%



The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2016 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, AMD FirePro, and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. OpenCL is a trademark of Apple, Inc. used by permission by Khronos. Other names are for informational purposes only and may be trademarks of their respective owners.

