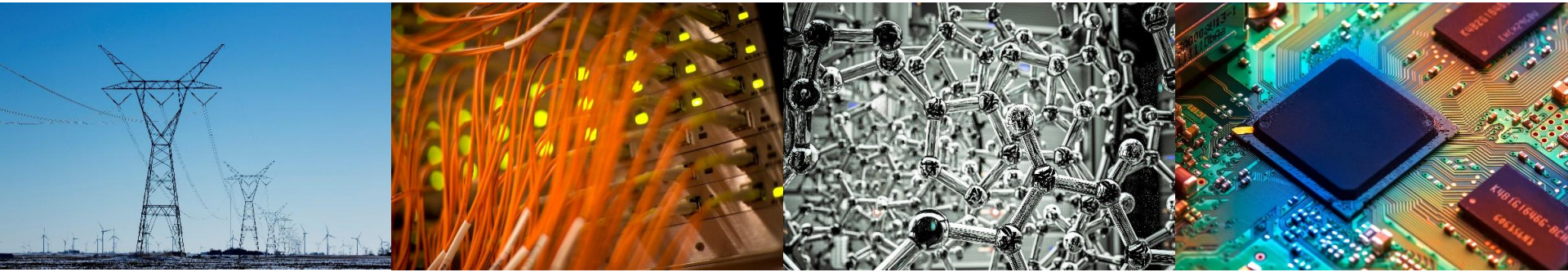
Bigger GPUs and Bigger Nodes

Carl Pearson (pearson@illinois.edu)
PhD Candidate, advised by Professor Wen-Mei Hwu



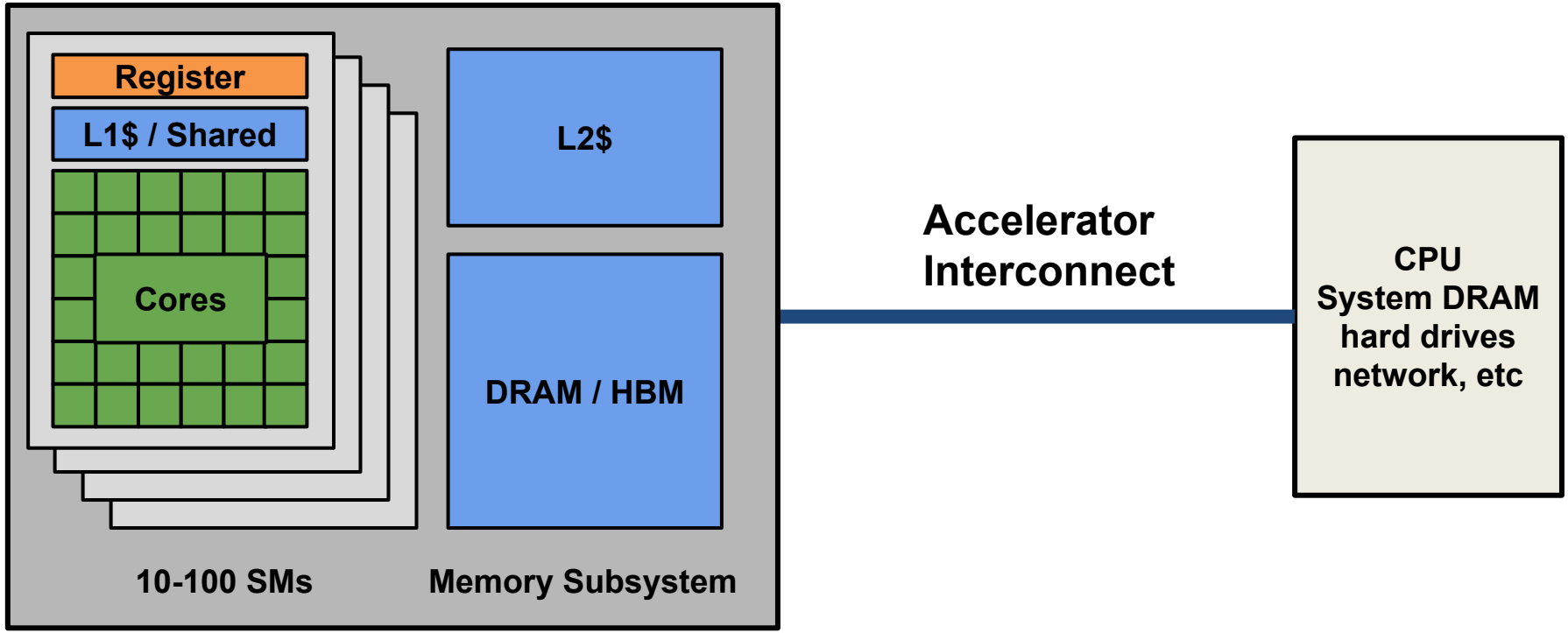
ILLINOIS
Electrical & Computer Engineering
COLLEGE OF ENGINEERING

Outline

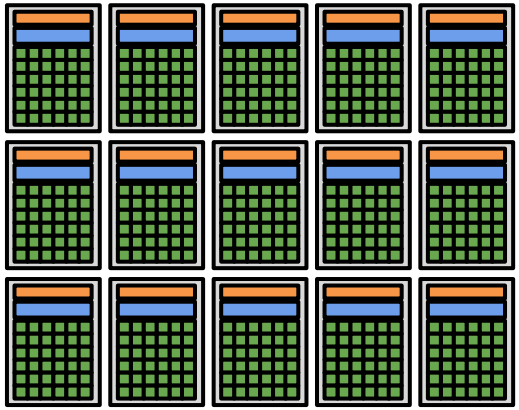
Experiences from working with domain experts to develop GPU codes on Blue Waters

- Kepler and Volta GPUs
- HPC Kepler to Volta Speedup
- Blue Waters, Summit, Sierra
- Intra-Node Communication Performance

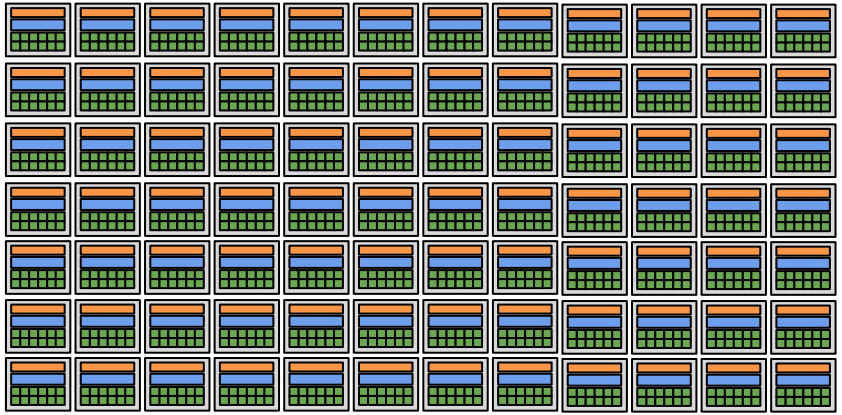
GPU Architecture Bird's Eye View (not to scale)



Kepler

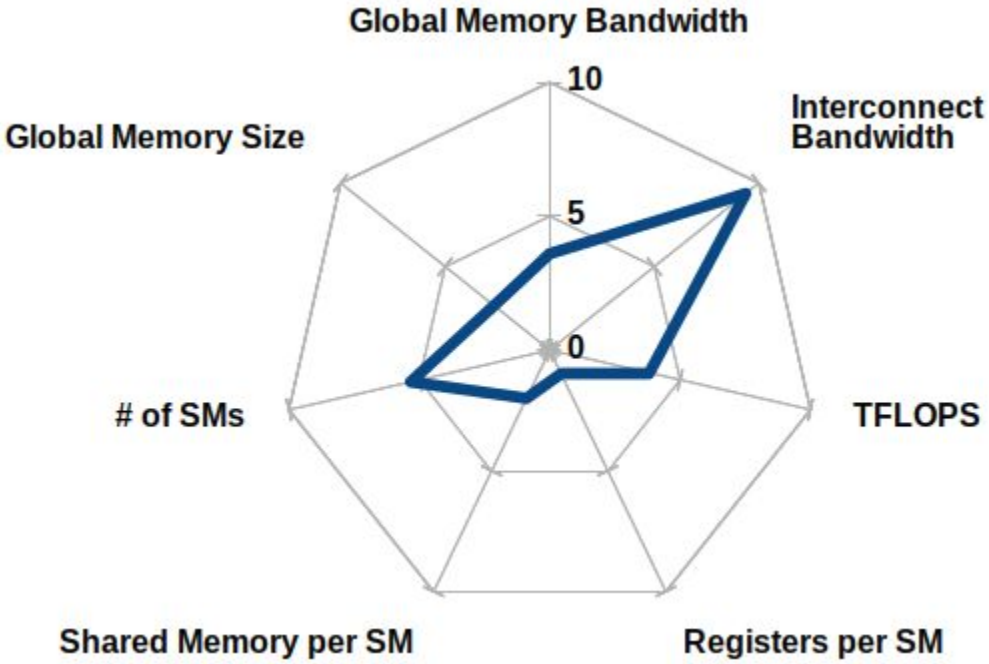


Volta



	Number of SMs	Maximum Blocks / SM	Shared Memory / SM	Registers / SM	Single Precision Rate	Global Memory Bandwidth
K20X (Kepler)	15	16	48 KB	64 K	3.94 TFLOPS	250 GB/s
V100 (Volta)	80	32	96 KB	64 K	15 TFLOPS	900 GB/s

K20x to V100: Architectural Parameters



HPC Case Studies

AWP-ODC	ChaNGa
Tom Jordan, Yifeng Cui	Tom Quinn
Southern California Earthquake Center	University of Washington
University of Southern California	
Anelastic Wave propagation	Charm N-body Gravity Solver
Solves a velocity-stress formulation of the 3D wave equation	Collisionless N-body simulations

AWP and ChaNGa V100 Speedup

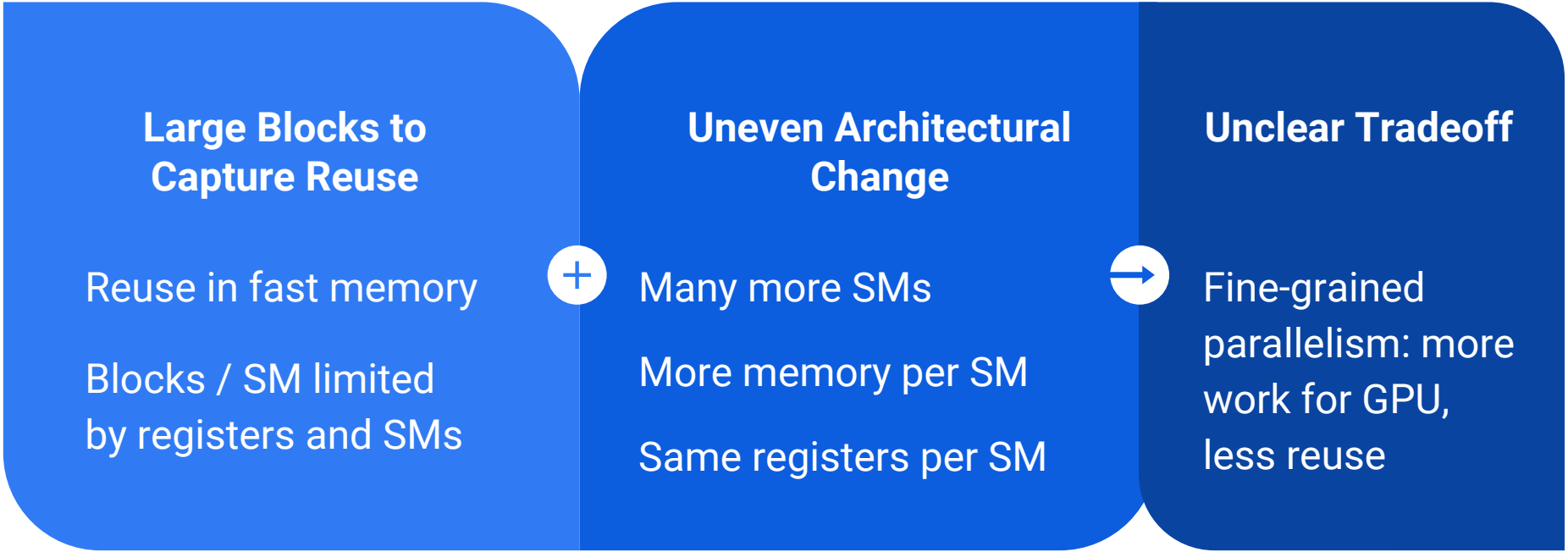
	Vs. P100	Vs. K20x (Blue Waters)
ChaNGa	3.28	4.73
AWP	1.71	5.19

AWP Detail

SP over p100	SP over K20X
1.711	5.188

	K20x		V100	
	Kernel 1	**Kernel 2**	**Kernel 1**	**Kernel 2**
GPU Time	72.4 %	27.5 %	70.1 %	29.3 %
Mem BW	145.7 GB/s	136.1 GB/s	726.7 GB/s	600.2 GB/s
	Latency-Limited		Bandwidth-Limited	

AWP Optimizations



Takeaways

Laissez-faire Approach:

- 3-5x kernel speedup over optimized Kepler
- 3-5x interconnect speedup over optimized Kepler
- Larger problem to fill GPU

Redesign/Rewrite Approach:

- Finer-grained parallelism to fill GPU
- Harder to capture reuse (key to performance)

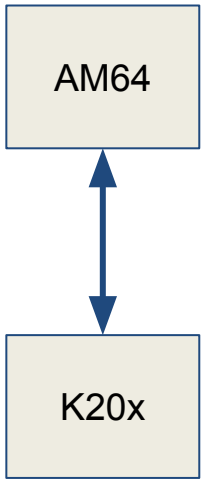
Nodes are Getting Bigger

	BW	Summit[1] (ORNL)					
CPU	1x AMD64 32 threads 16 FP	POWER9 88 threads 22 FP			POWER9 88 threads 22 FP		
GPU	K20X 6 GB 4 TF	V100 16 GB 15 TF	V100 16 GB 15 TF	V100 16 GB 15 TF	V100 16 GB 15 TF	V100 16 GB 15 TF	V100 16 GB 15 TF
Accelerator Interconnect (unidirectional)	PCIe 2x16 8 GB/s	NVLink 2.0 x2 50 GB/s					
Memory	32GB	512 GB					

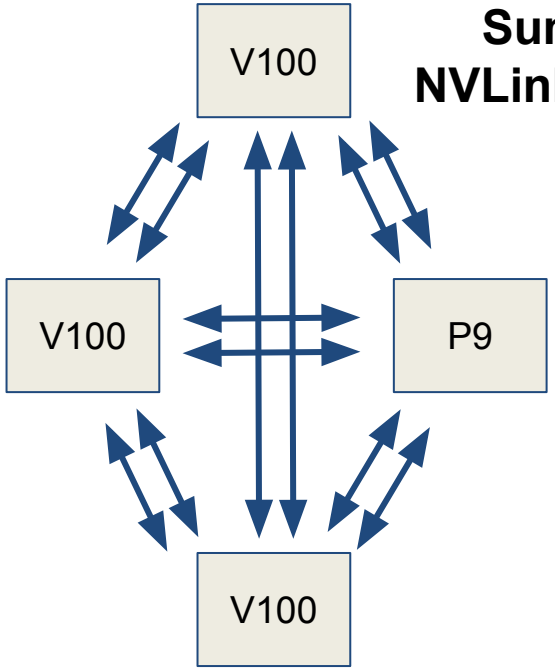
1: https://www.olcf.ornl.gov/for-users/system-user-guides/summit/system-overview/

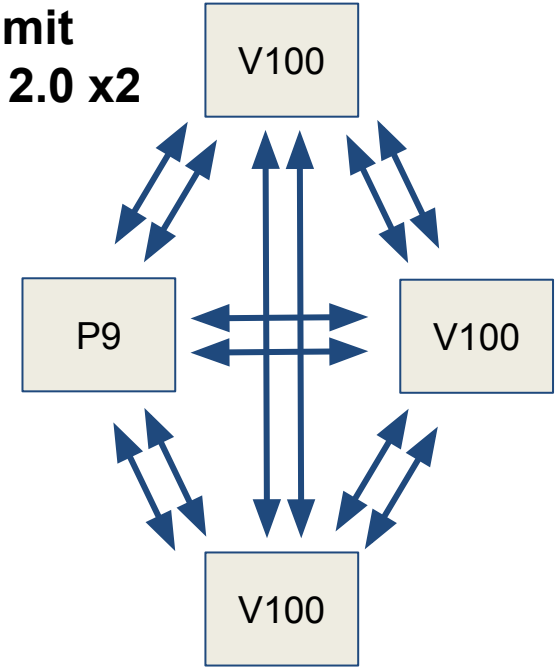
Blue Waters XK and Summit Intra-Node Interconnects

Blue Waters
PCIe 2.0 x16



Summit
NVLink 2.0 x2





System Performance Research

CUDA

- Microbench: https://github.com/rai-project/microbench

Neural Networks

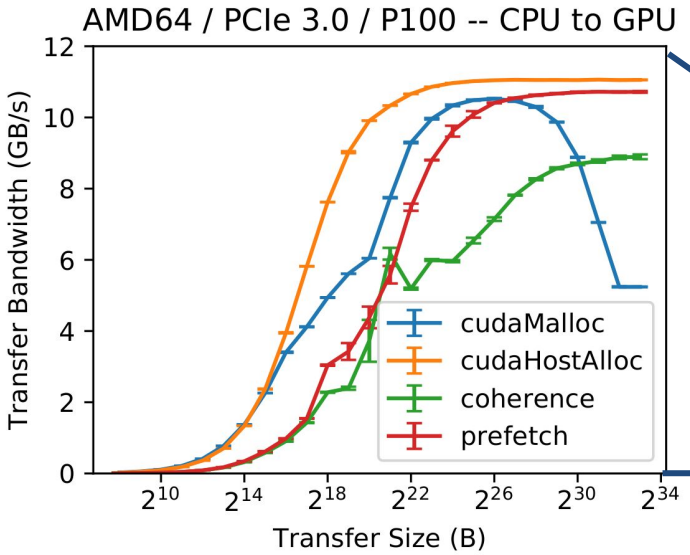
- MLModelScope: http://ml-arc-minsky.netlify.com/

Future Directions:

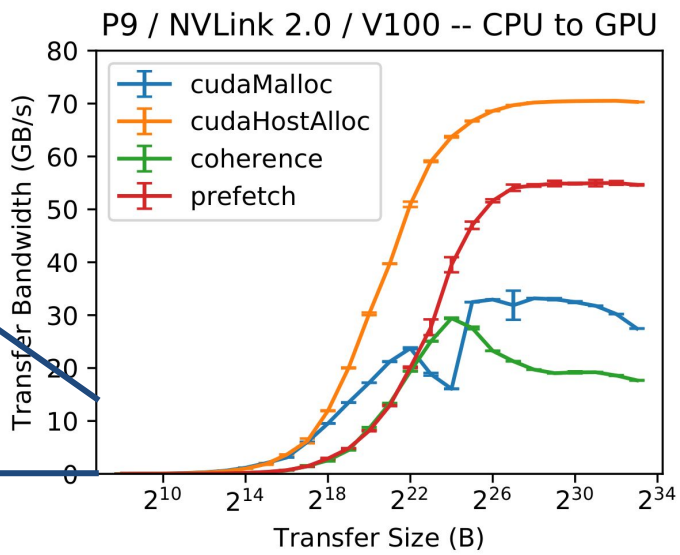
- Quick application-driven architecture design
- Performance modeling of neural networks

Faster Interconnects

PCIe 3.0 x16 (2x BW)
15.8 GB/s



NVLink 2.0 x3 (1.5x Summit)
75 GB/s



github.com/rai-project/microbench

Unified Memory

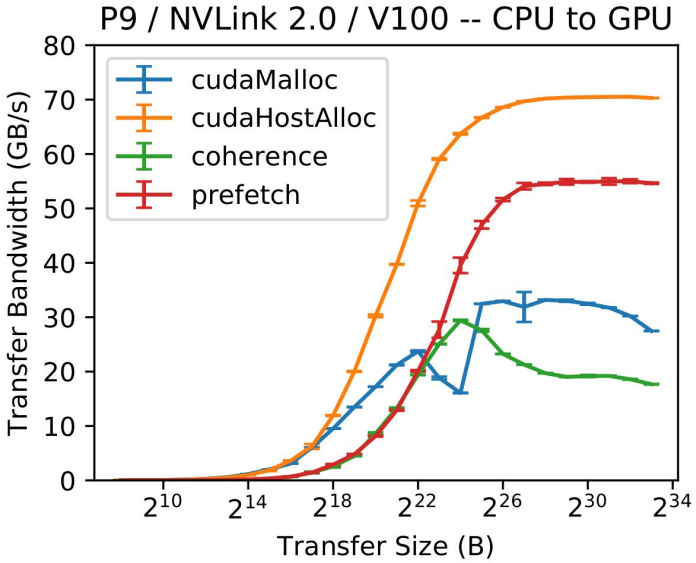
Allocations accessible from CPU and GPU
Implicit data transfer (no cudaMemcpy)

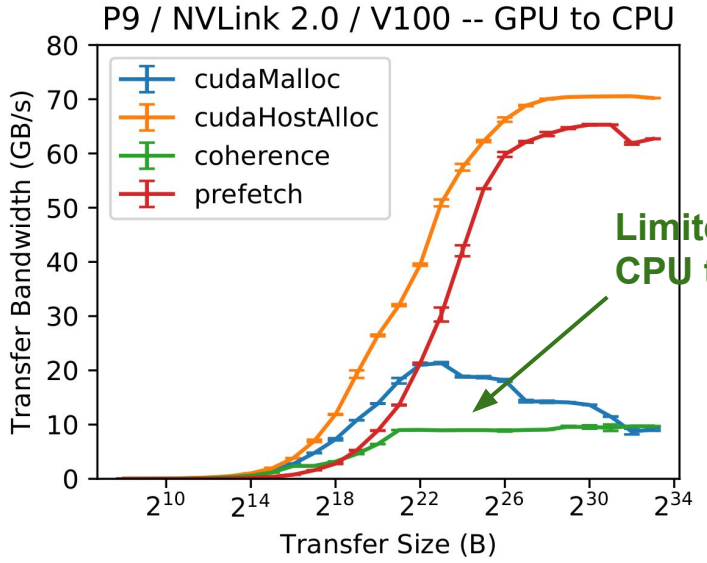
	GPU 0	GPU 1	CPU	
`cudaSetDevice(0);` `cudaMallocManaged(&a,...);`				
`a[page0] = 0; // gpu0`				
`a[page1] = 1; // gpu1`				Page fault and migration
`a[page2] = 2; // cpu`				Page fault and migration

`cudaMemAdvise(a, gpu1,`
`cudaMemAdviseSetPreferredLocation);`
`a[page1] = 1; // cpu`
				Write served over NVLink

`cudaMemPrefetcAsync(a, gpu1);`				Bulk page migration

P9 Unified Memory Performance

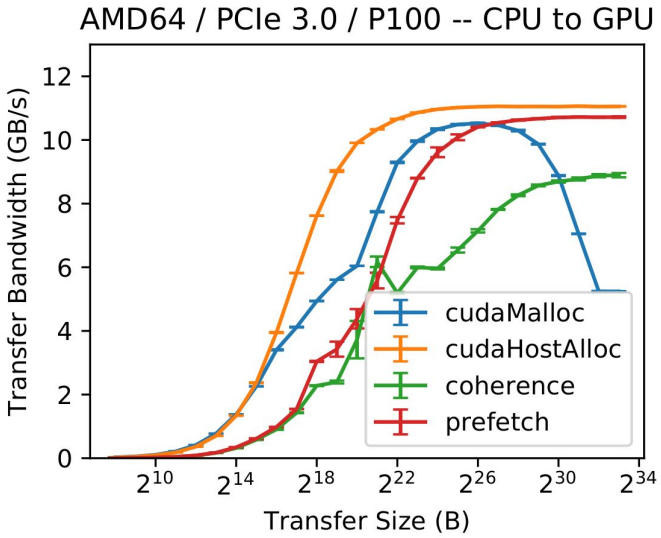


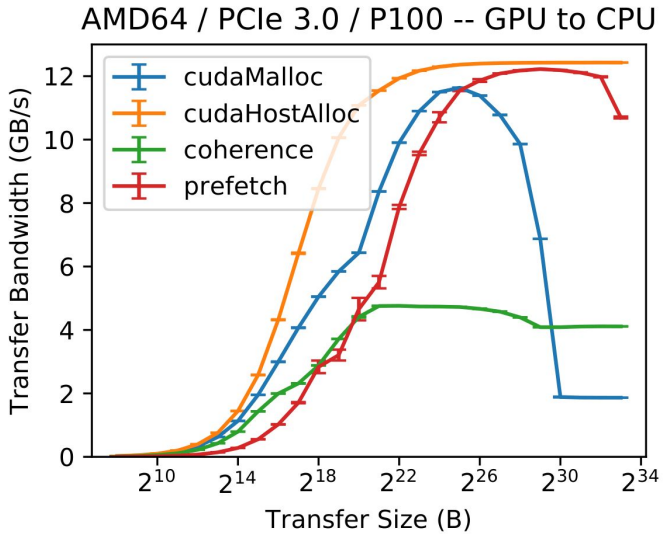


Coherence: 30% of explicit management
Prefetch: 50-80% of explicit

github.com/rai-project/microbench

AMD64 Unified Memory Performance

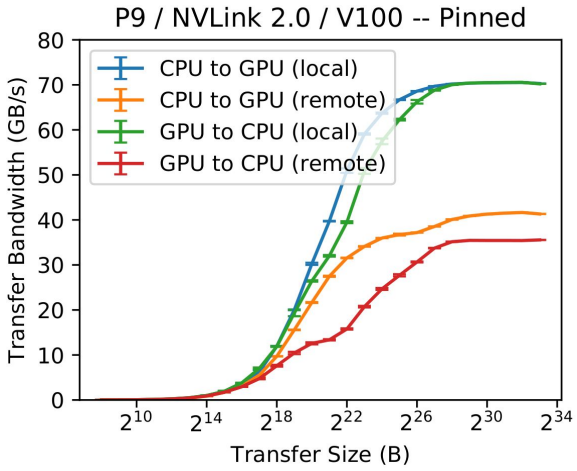


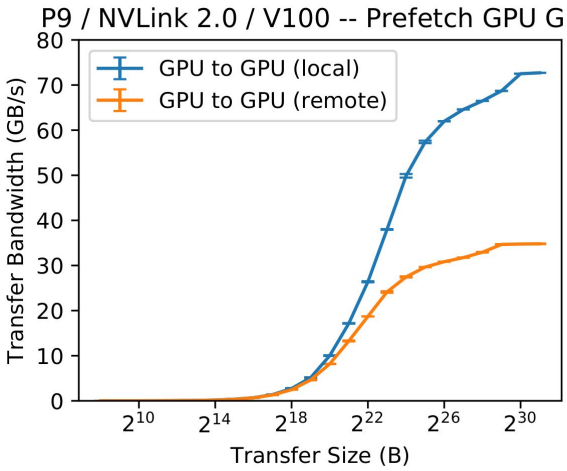


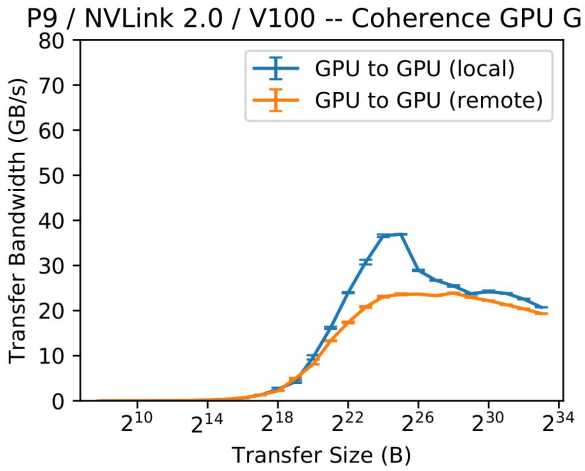
Coherence: 30-70% of explicit management
Prefetch: 50-95% of explicit

github.com/rai-project/microbench

Device Affinity







Data placement on big nodes can have a dramatic communication impact

github.com/rai-project/microbench

MLModelScope: Neural Network Performance Data

http://ml-arc-minsky.netlify.com

`(model -- machine -- framework)` triples

- `( AlexNet -- Jetson TX-1 -- Tensorflow )`
- `( VGG19 -- AWS P2 X-large -- MxNet )`

Neural-network performance primitive benchmarks

Thank You

https://cwpearson.github.io
pearson@illinois.edu

Special thanks to

- Professor Wen-Mei Hwu
- John Larson, Simon Garcia de Gonzalo, Zaid Qureshi, Mert Hidayetoglu, Abdul Dakkak and Cheng Li (University of Illinois)
- Isaac Gelado (NVIDIA)
- Jinjun Xiong and I-Hsin Chung (IBM)
- The IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR) - a research collaboration as part of the IBM Cognitive Horizon Network.