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

VirginiaTech
Invent the Future

SyNeRG
synergy.cs.vt.edu

“Sampling” of fields that use GPUs

Mac OS X

Cosmology

Molecular Dynamics and Modeling

Computational Fluid Dynamics

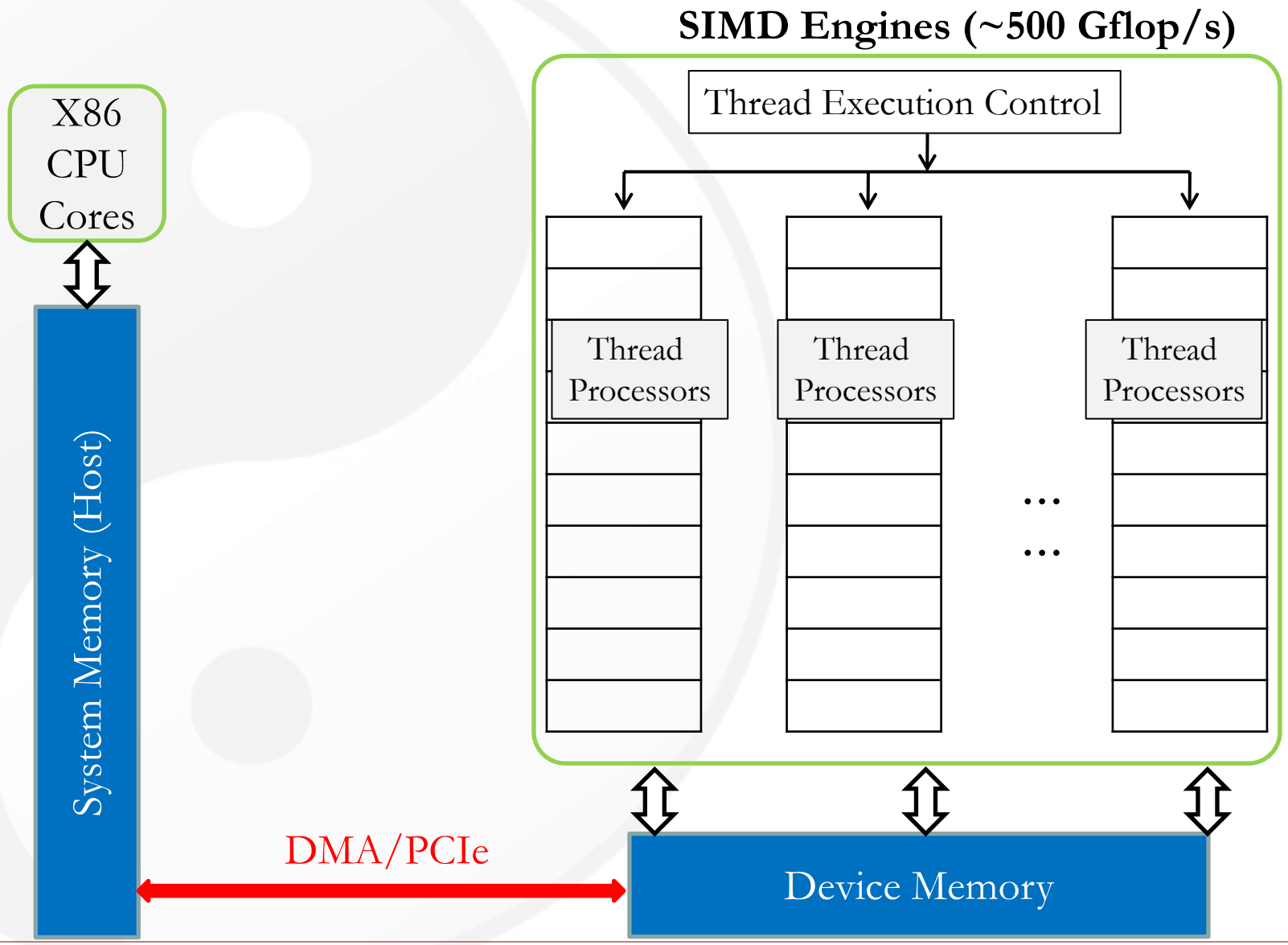
GPUs in HPC

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

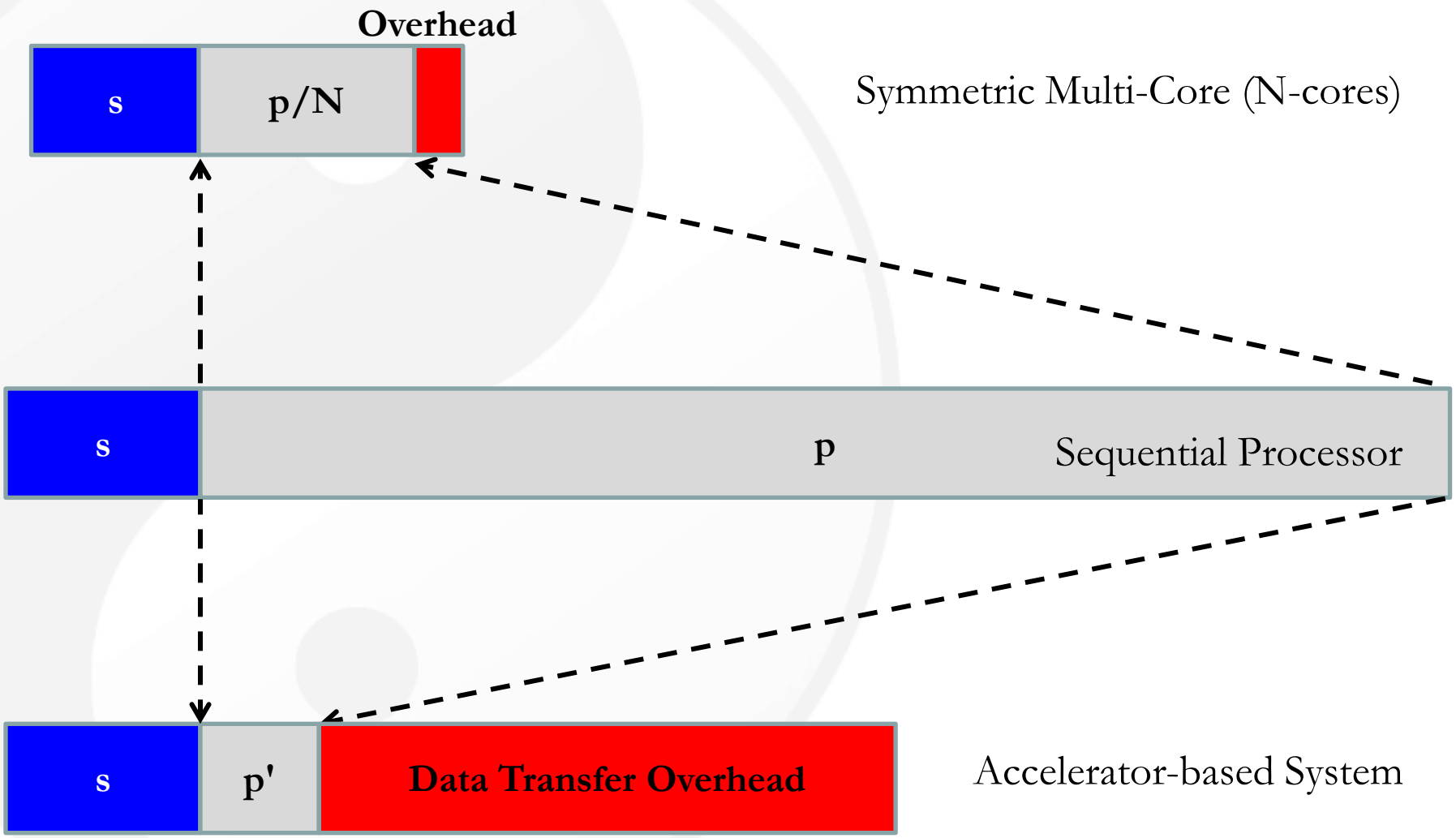
Systems *with* GPUs achieve only ~50 % of R_{peak}
Systems *without* GPUs achieve ~84 % of R_{peak}

http://www.top500.org

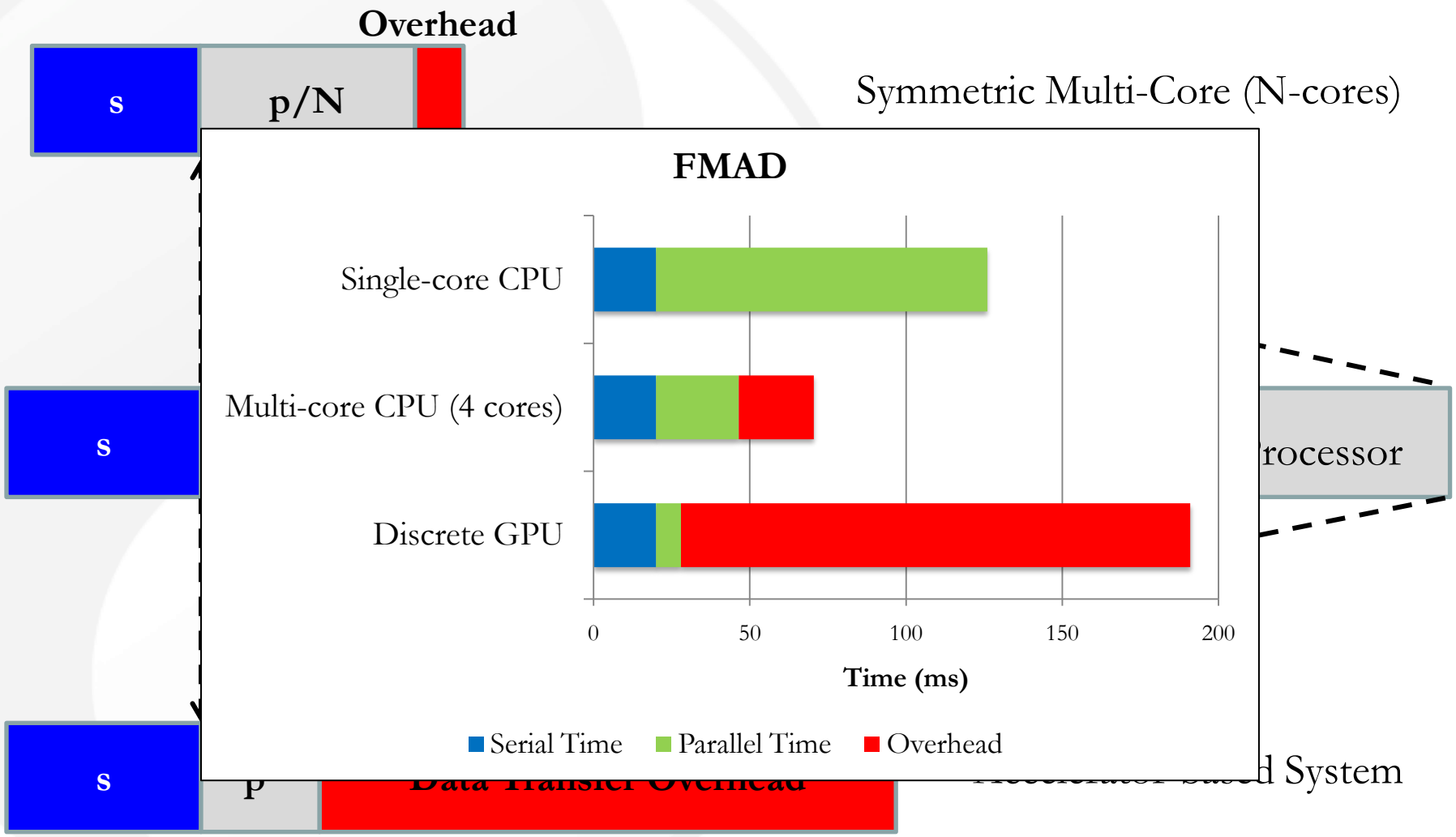
Architecture of Discrete GPUs



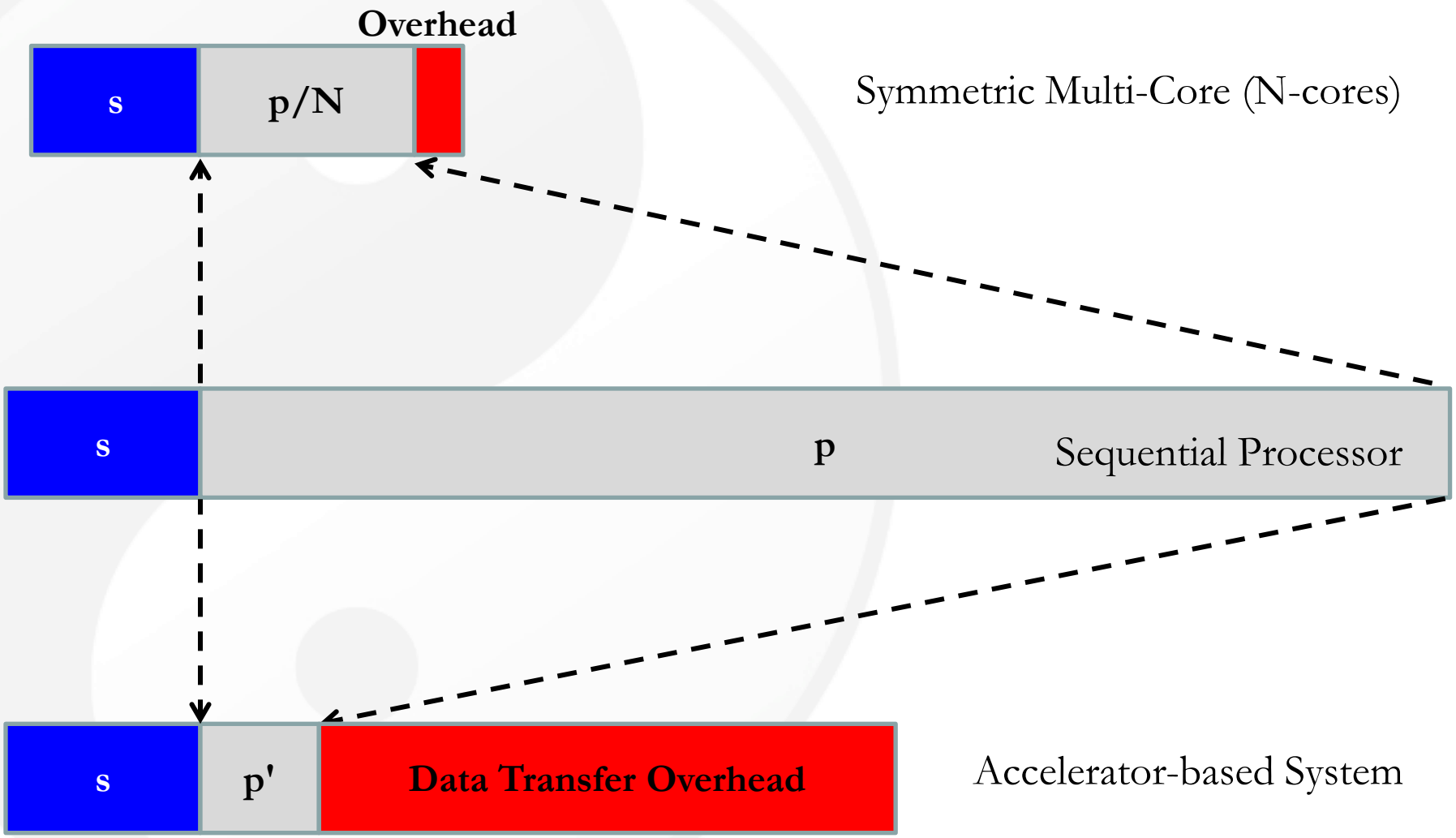
A Reason for Poor Efficiency



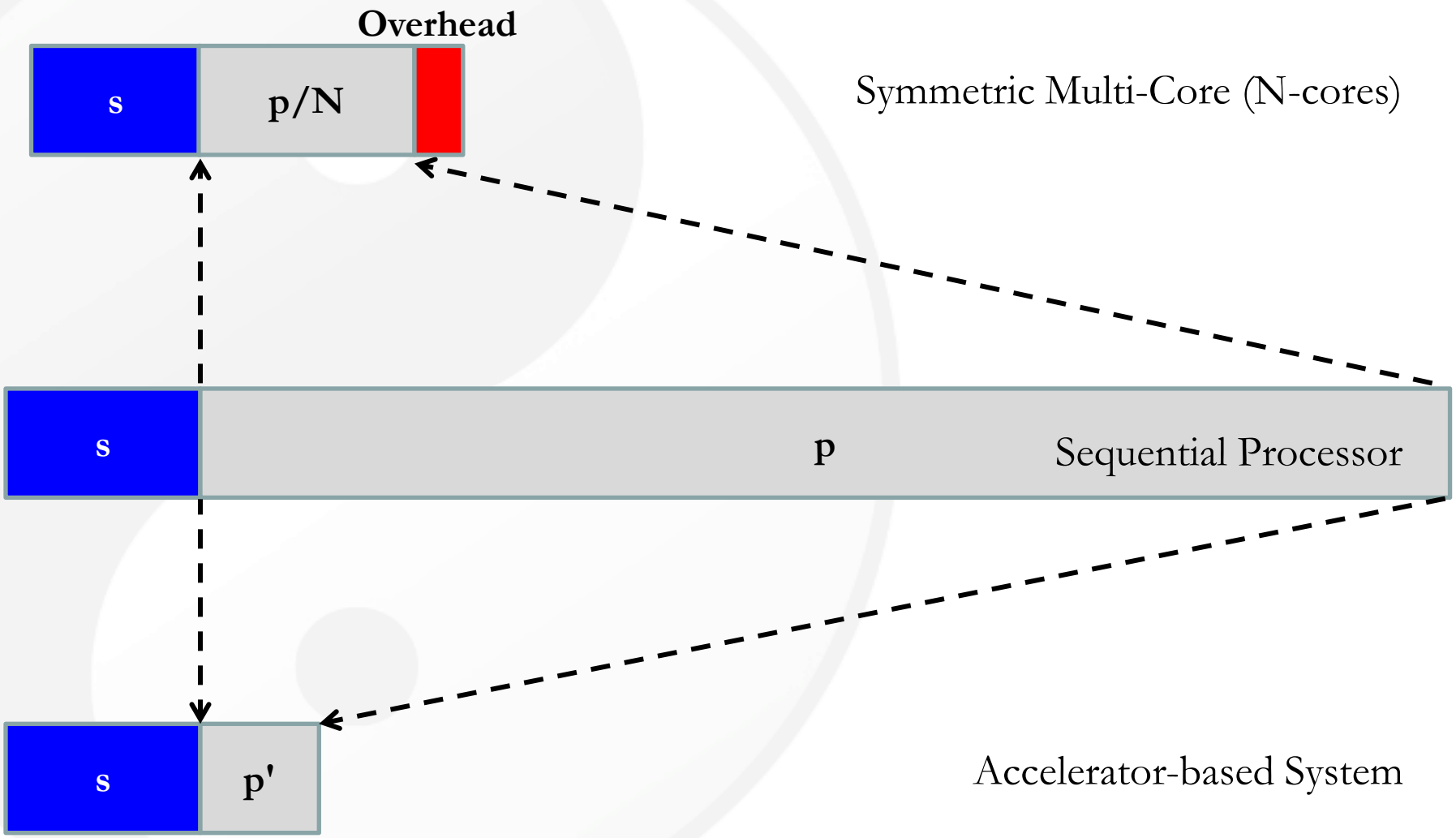
A Reason for Poor Efficiency



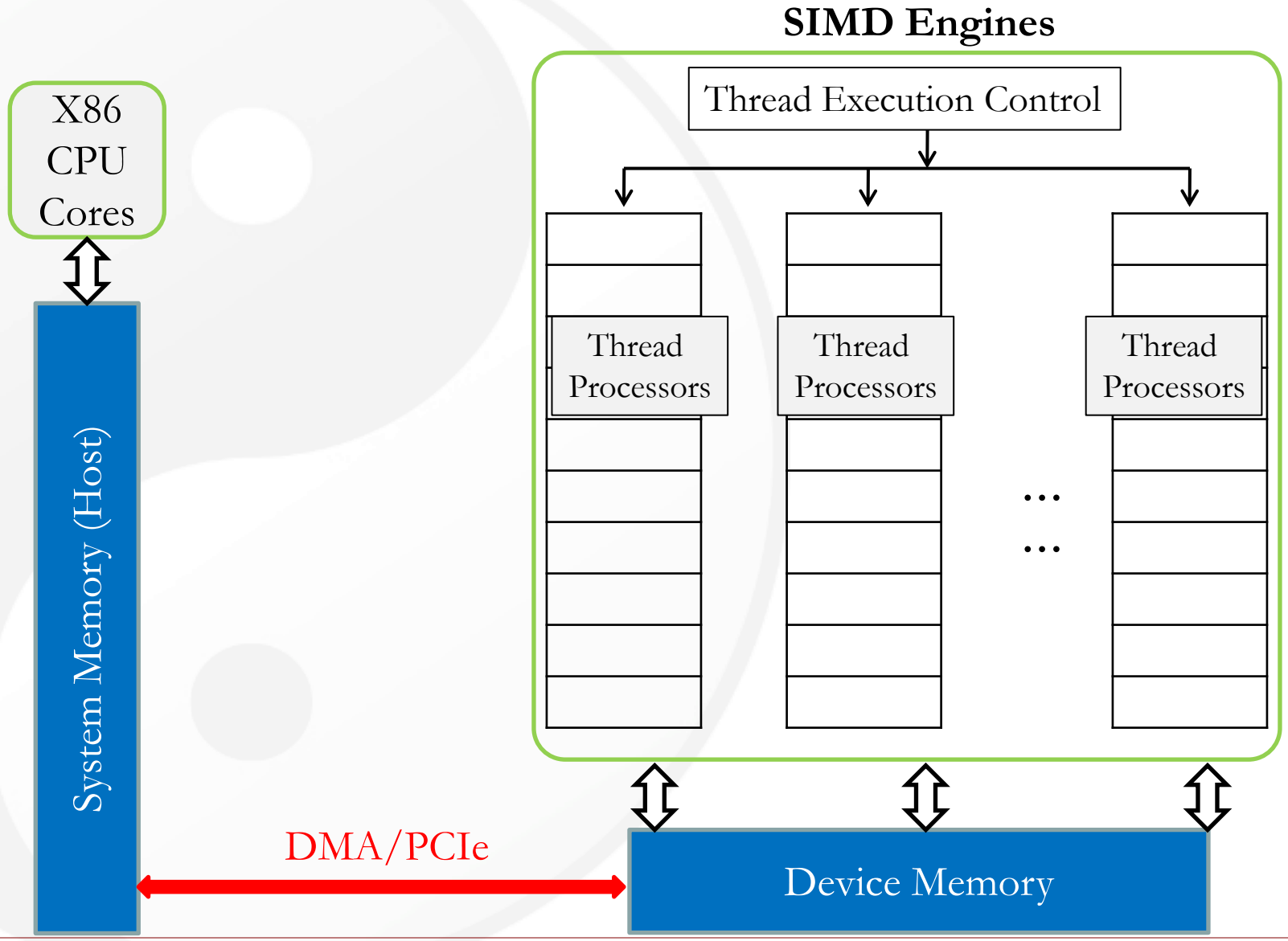
Ideal Efficiency Scenario



Ideal Efficiency Scenario

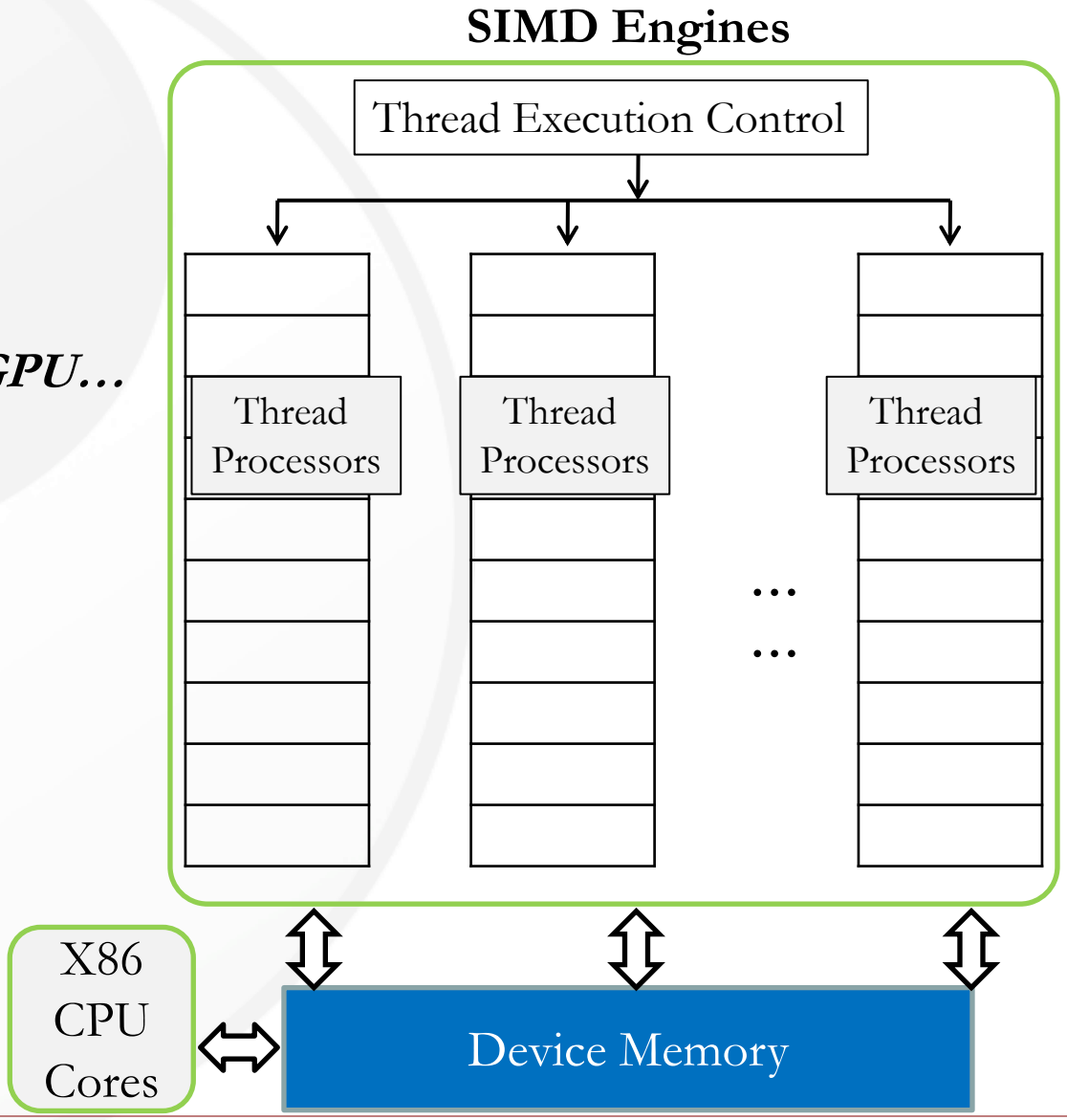


Ideal Placement of CPU and GPU Cores



Ideal Placement of CPU and GPU Cores

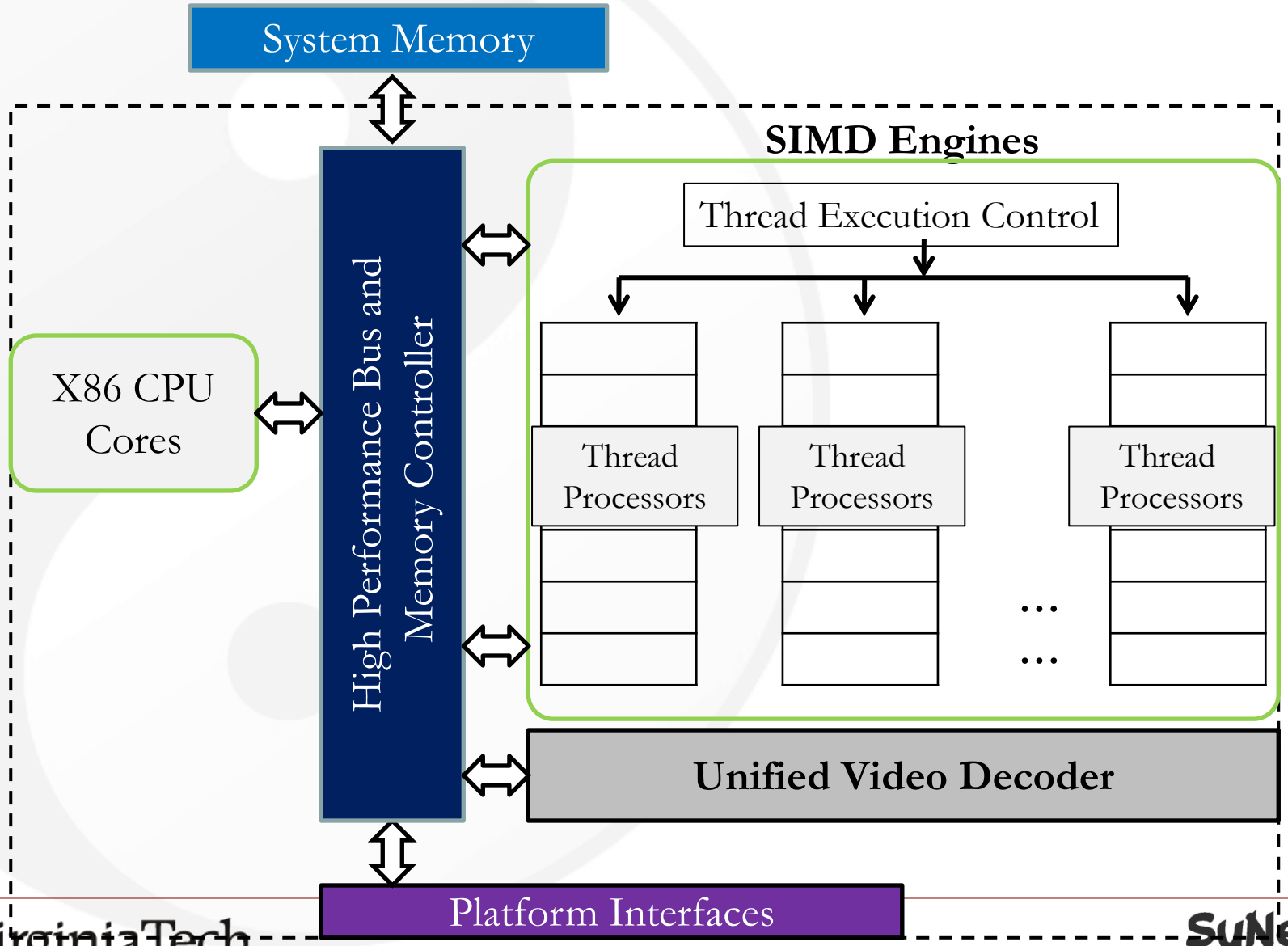
Towards a "fused" CPU+GPU…



Outline

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

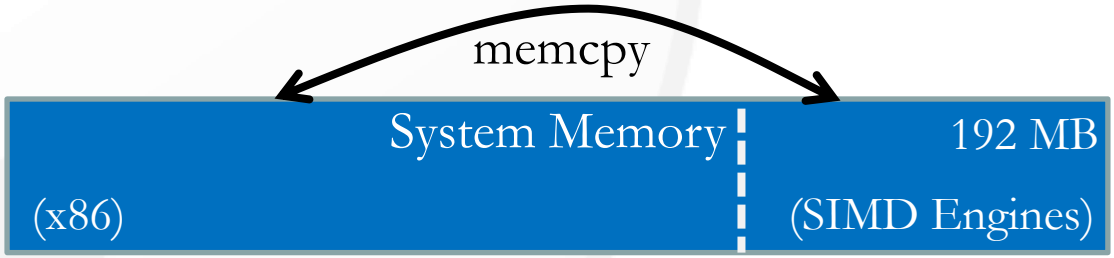
AMD Fusion APU – A Fused CPU+GPU



State of the Data Transfer

- Discrete GPU

- AMD Fusion APU (1st 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"*)

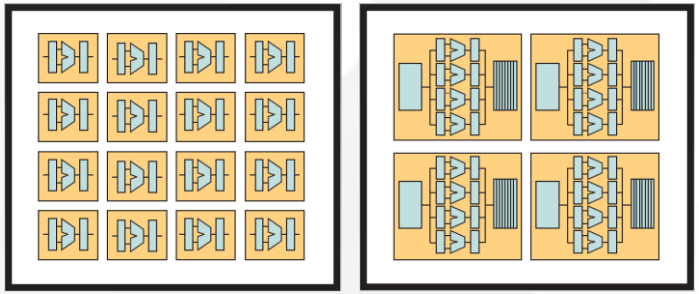
Outline

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

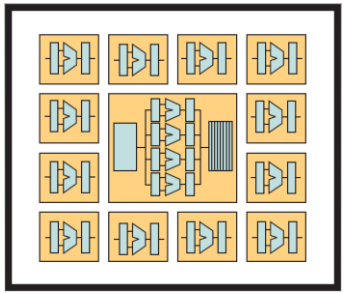
Revisiting Amdahl's Law

(M. Hill and M. Marty, *"Amdahl's Law in the Multi-core Era"*)

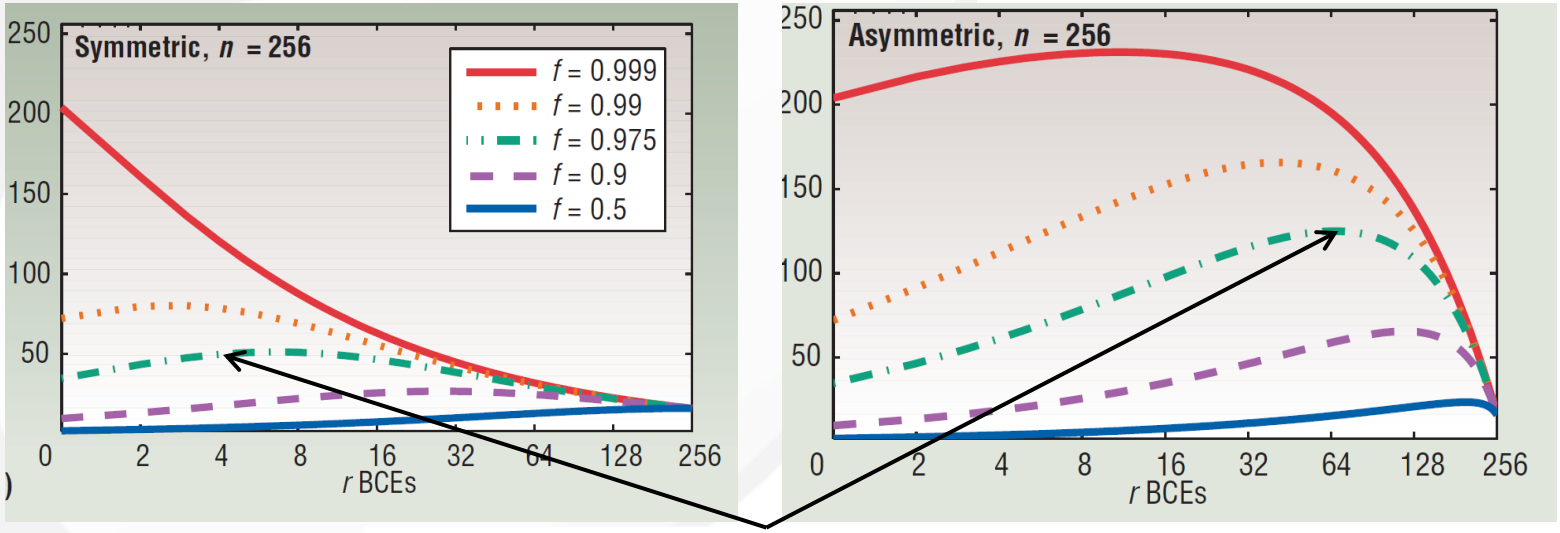
Symmetric Multi-core



Asymmetric Multi-core

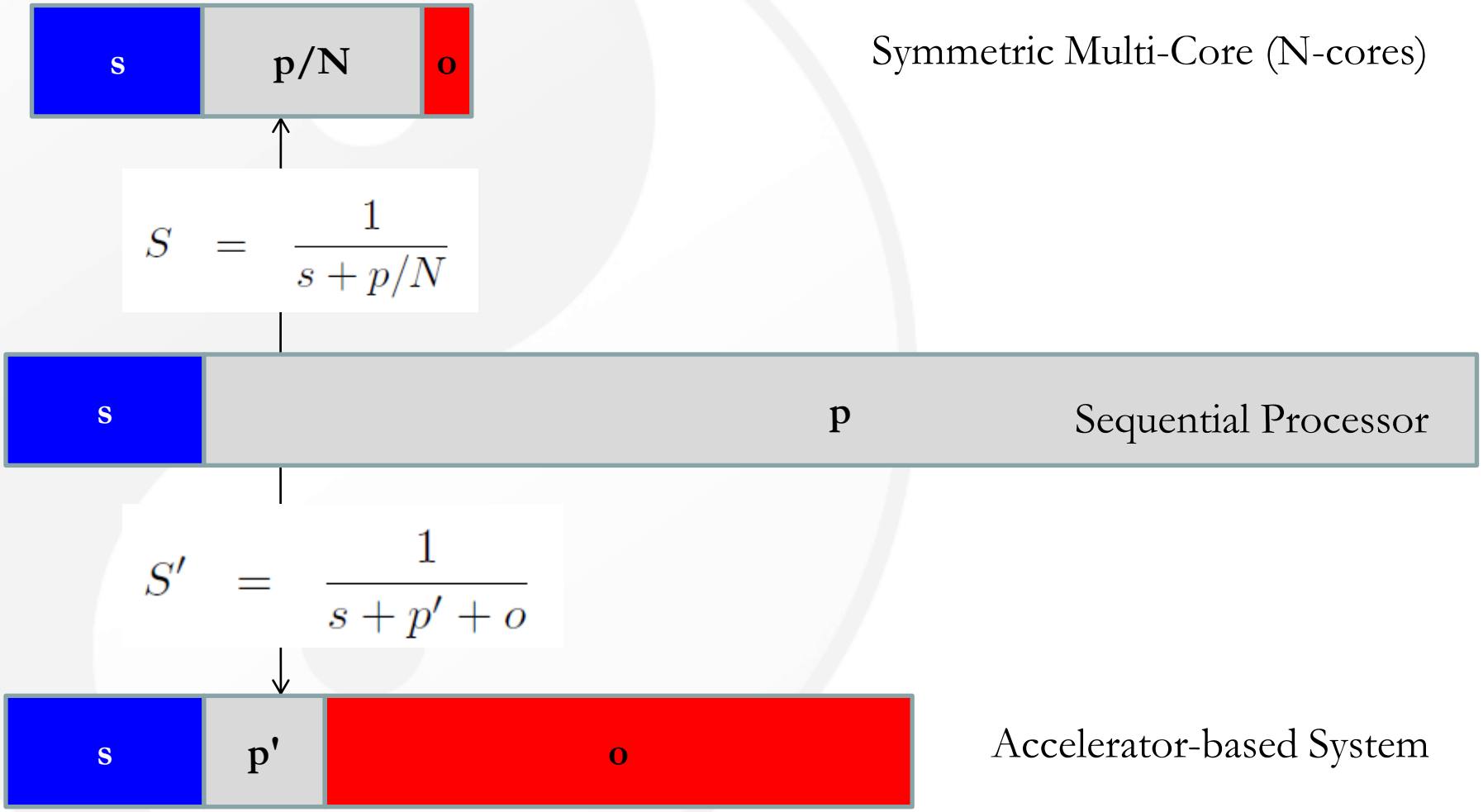


Speedup values for different serial fractions

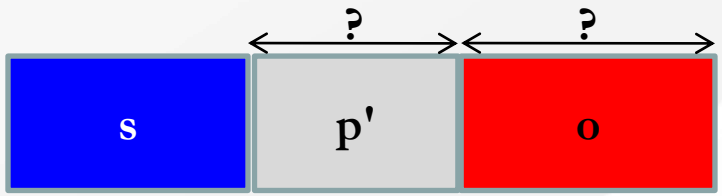


Higher Efficiency of Asymmetric Chips

Revisiting Amdahl's Law



Revisiting Amdahl's Law



Accelerator-based System

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

- $o_{DiscreteGPU}$ vs. o_{Fusion}
  - Fusion is expected to be better than discrete GPUs
- $p'_{DiscreteGPU}$ vs. p'_{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

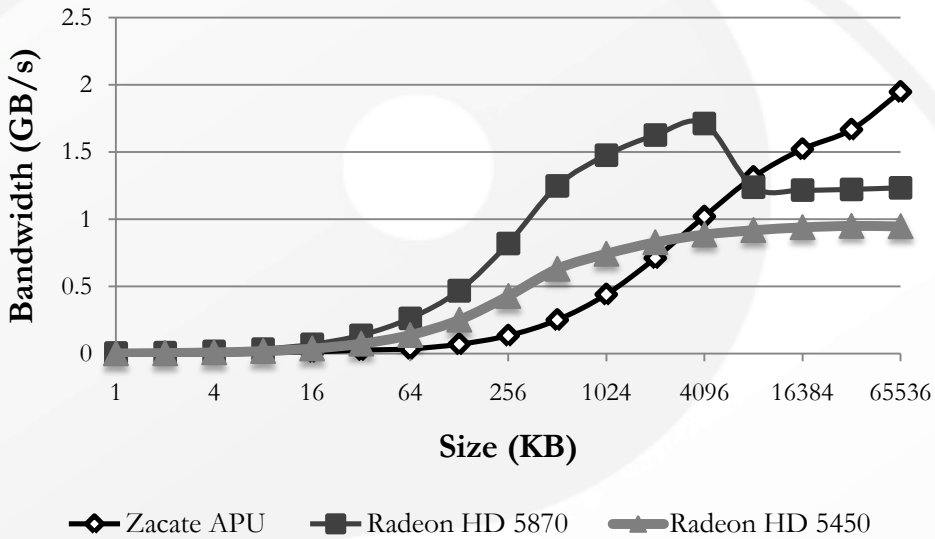
Platform	AMD Zacate APU	AMD Radeon HD 5870	AMD Radeon HD 5450
Stream Processors	80	1600	80
Compute Units	2	20	2
Memory Bus Type	NA	GDDR5	DDR3
Device Memory	192 MB	1024 MB	512 MB
Local Memory	32 KB	32 KB	32 KB
Max. Workgroup Size	256 Threads	256 Threads	128 Threads
Peak Core Clock Freq.	492 MHz	850 MHz	675 MHz
Peak FLOPS	80 GFlop/s	2720 GFlop/s	104 GFlop/s
Host:			
Processor	AMD Engg. Sample @1.6 GHz	Intel Xeon E5405 @2.0 GHz	Intel Celeron 430 @1.8 GHz
System Memory	2 GB (NA)	2 GB DDR2	2 GB DDR2
Cache	L1: 32K, L2: 512K	L1: 32K, L2: 6M	L1: 32K, L2: 512K
Kernel	Ubuntu 2.6.35.22	Ubuntu 2.6.28.19	Ubuntu 2.6.32.24

Experimental Analysis

SHOC Benchmark

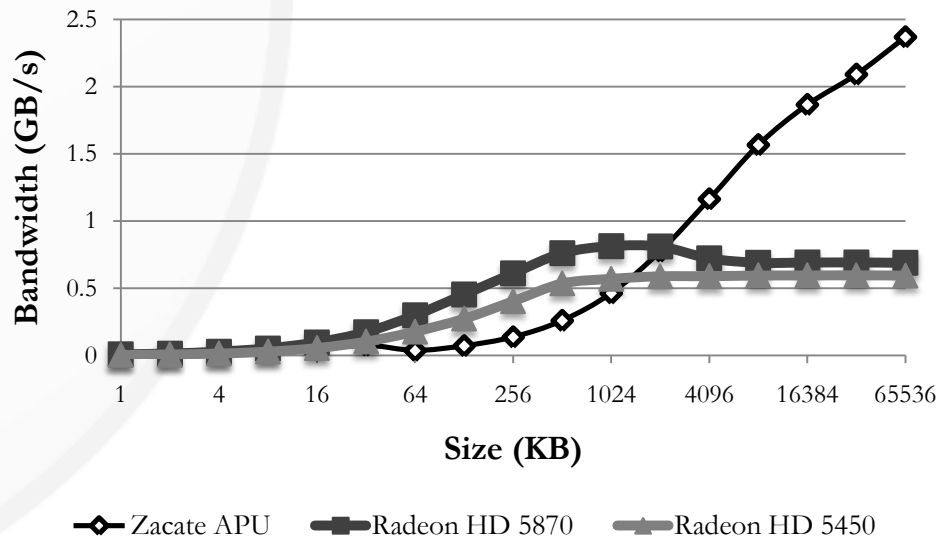
- Application Benchmarks
  - Bandwidth Test
    - Measures PCIe bandwidth for discrete GPU
    - Measures memory bandwidth for APU
  - FFT
    - Measures performance of a 2-D Fast Fourier Transform
    - Computes multiple FFTs of size 512 in parallel
  - MD
    - Measures performance of pairwise calculation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

Bandwidth Test

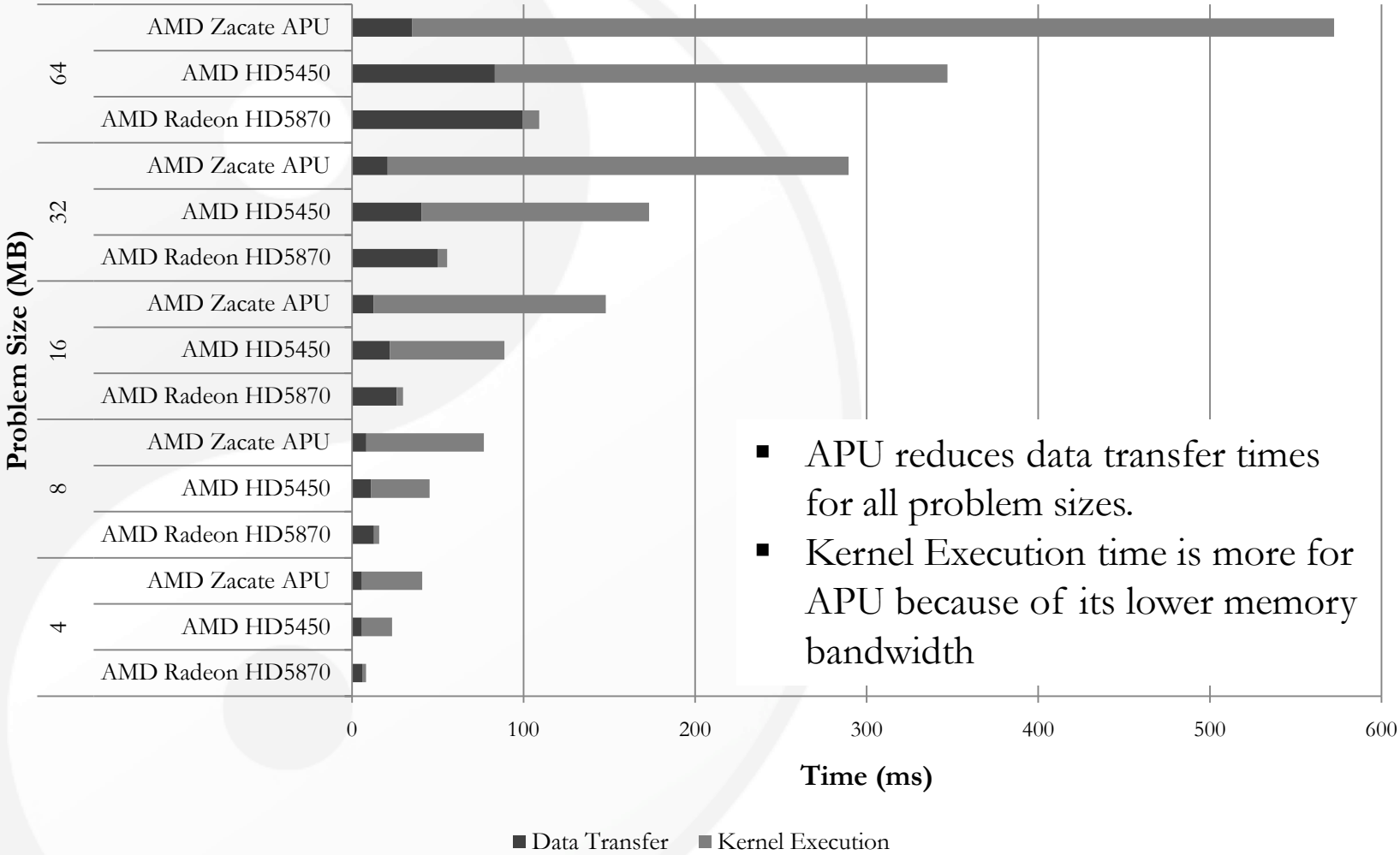


Host to Device

Device to Host

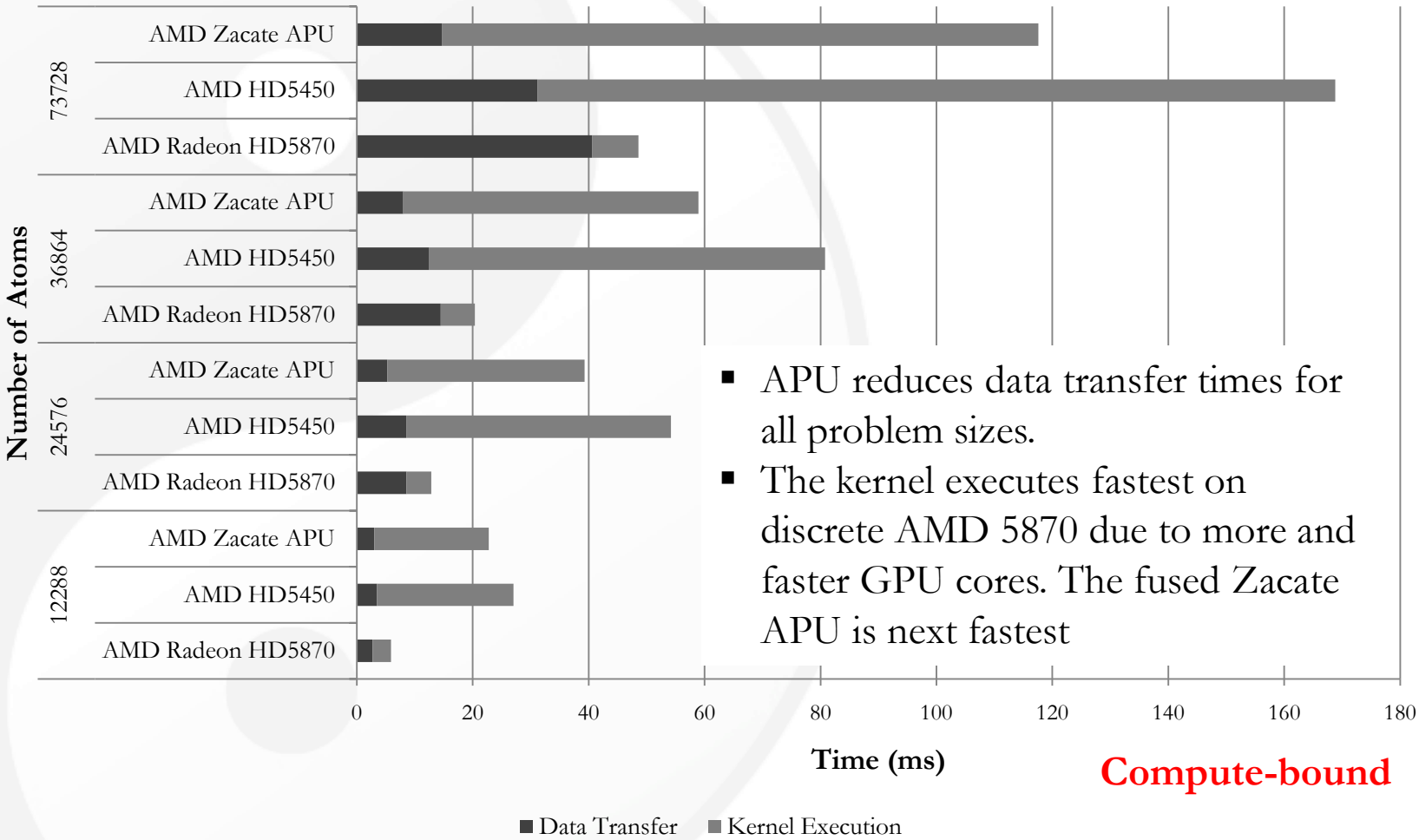


Fast Fourier Transform (FFT)



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

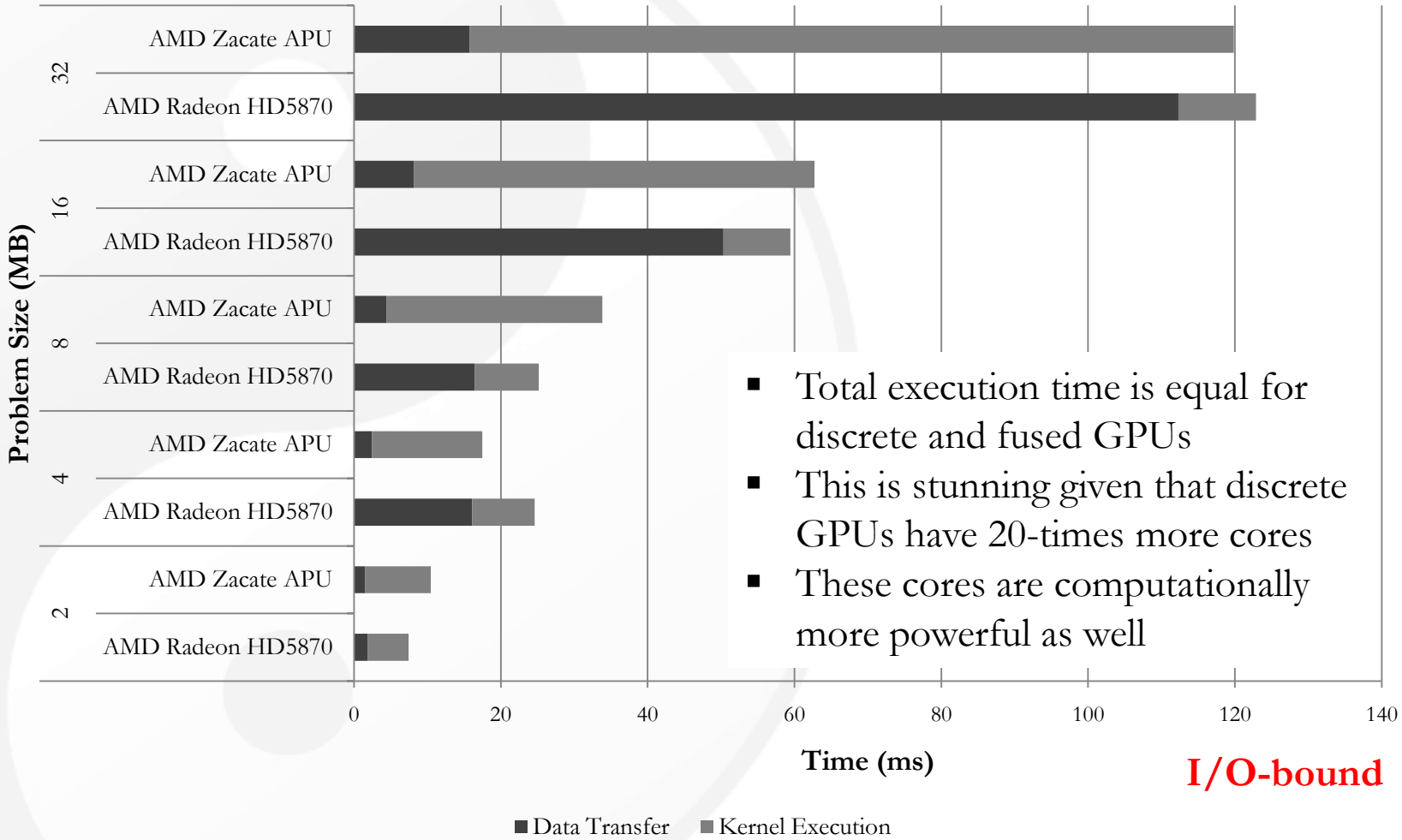
Molecular Dynamics



- 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

Compute-bound

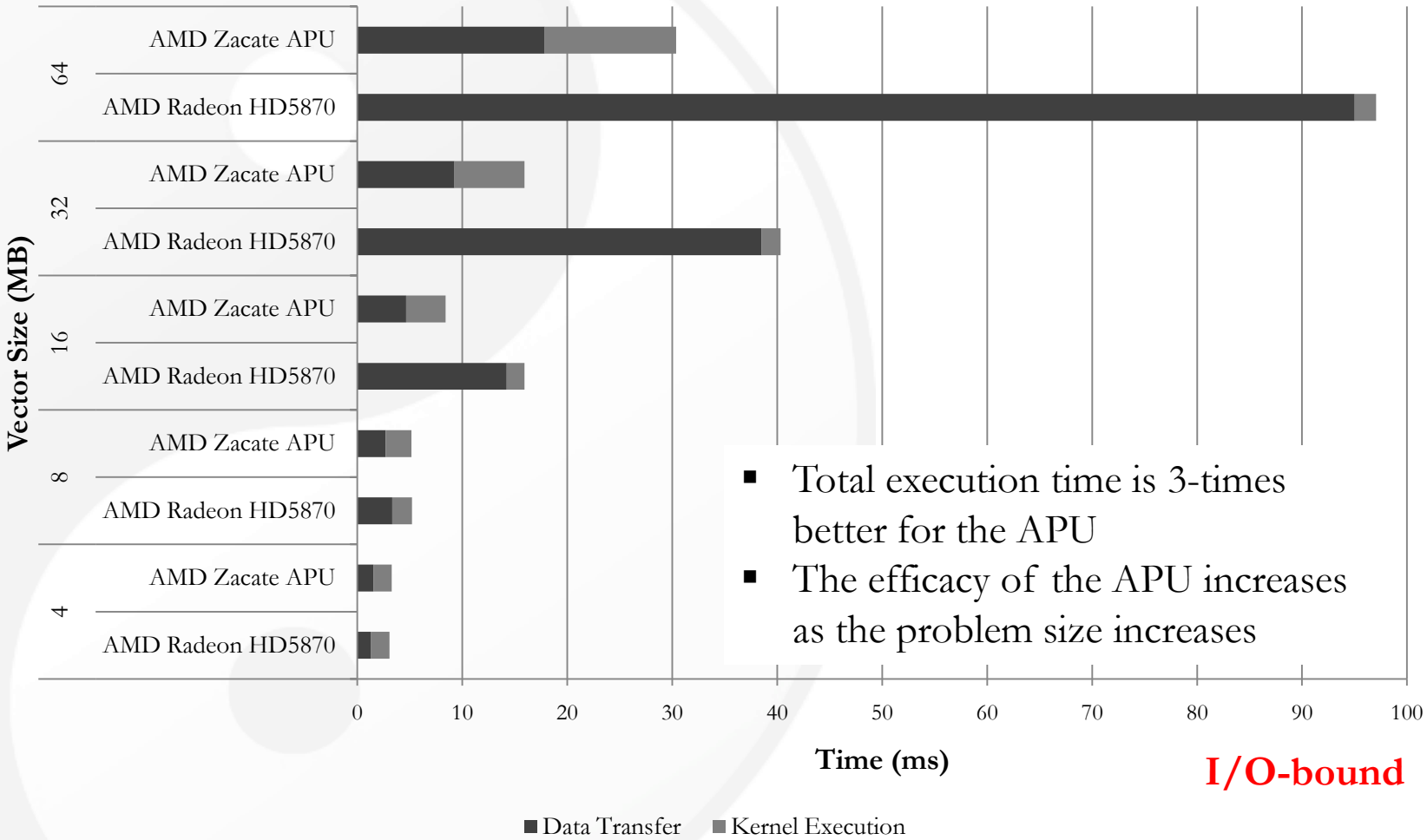
Scan



- 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

I/O-bound

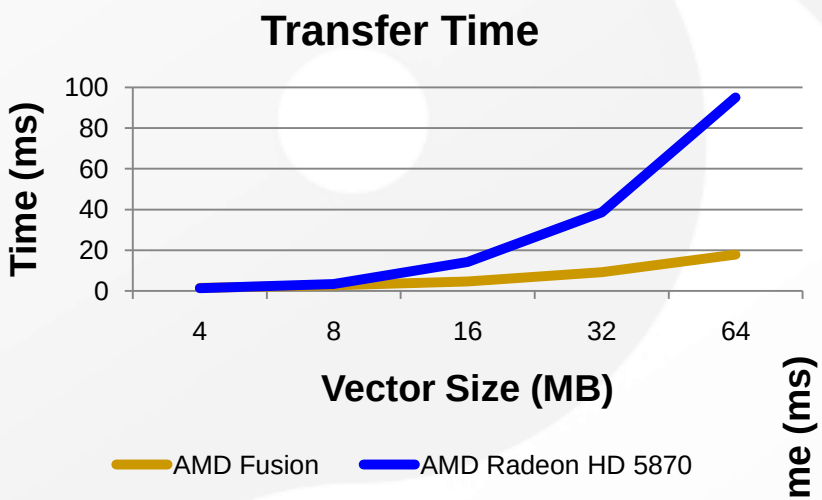
Reduction

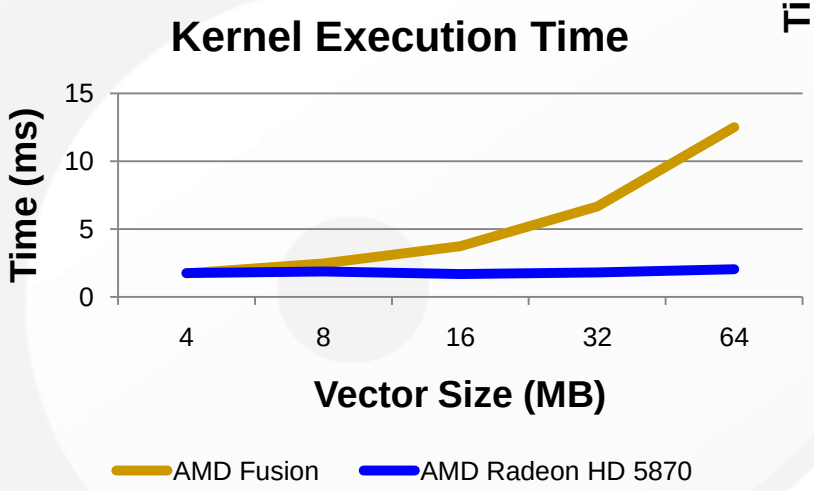


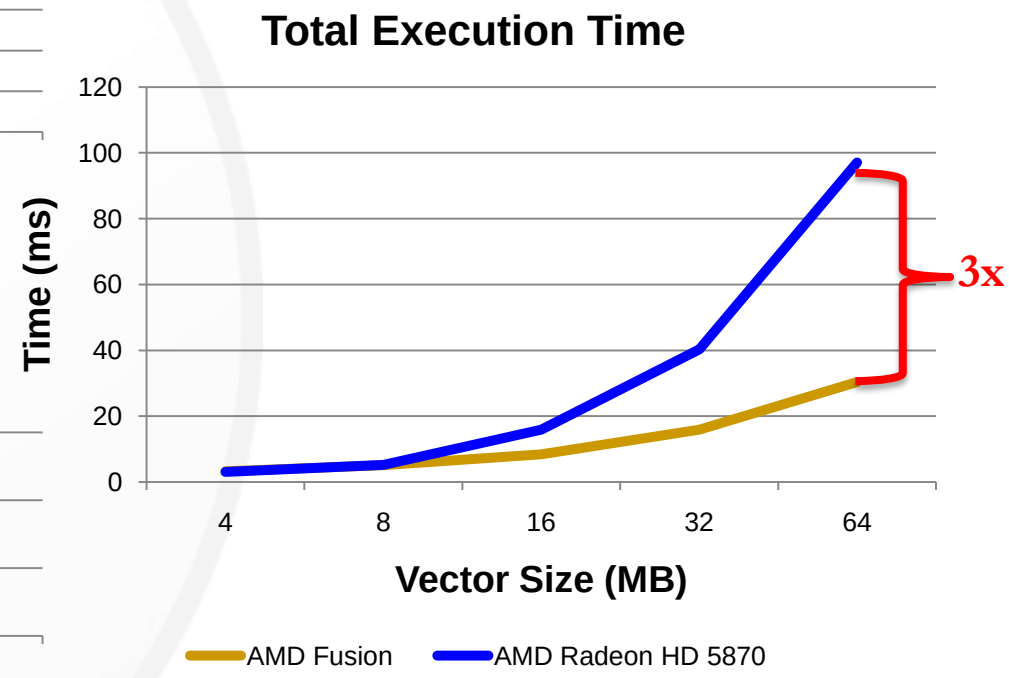
- Total execution time is 3-times better for the APU
- The efficacy of the APU increases as the problem size increases

I/O-bound

Reduction







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

Contacts

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

Questions?