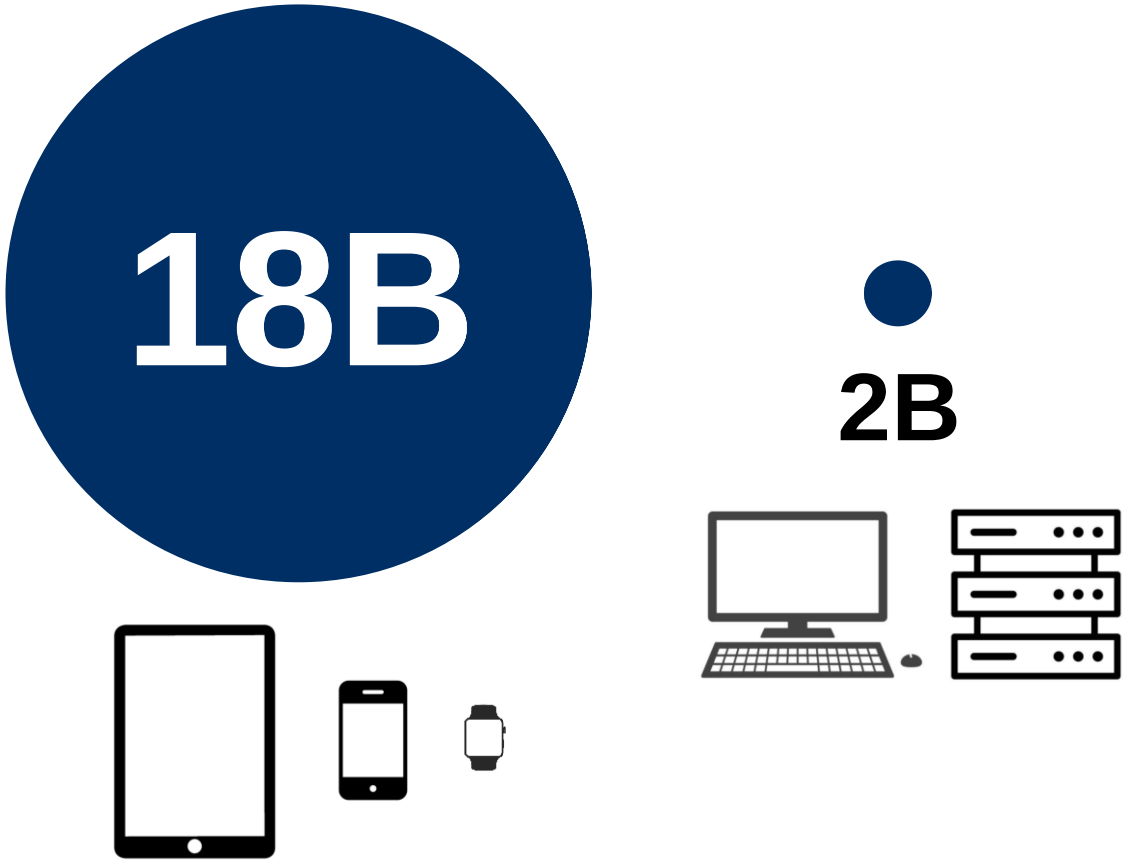
Workload Characterization of Commercial Mobile Benchmark Suites

Victor Kariofillis, Natalie Enright Jerger

UNIVERSITY OF TORONTO

Do we focus enough on mobile devices?

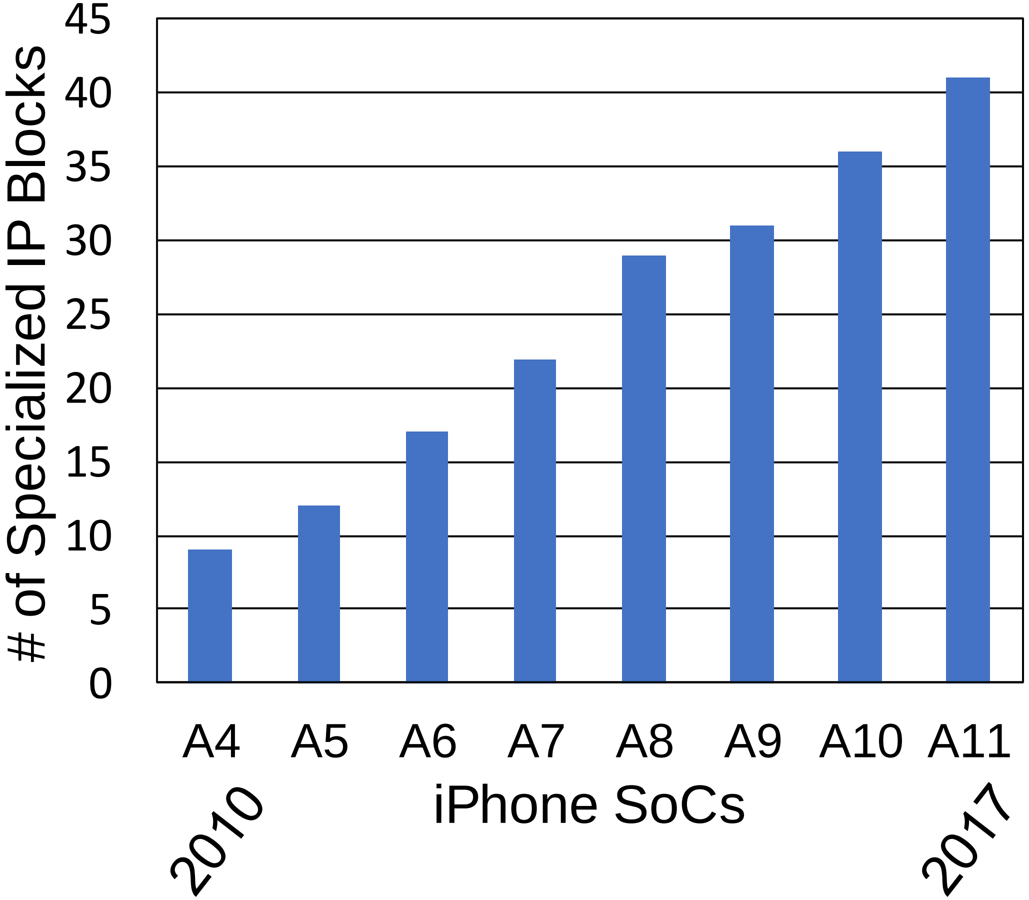
- Over 18 billion mobile devices
- Around 2 billion computers
- 1% of top tier publications focus on mobile computing (2018 study) [1]
- Mobile SoCs are distinct
  - Tight integration of hardware components
  - Significant heterogeneity
  - Rapid evolution



[1] V. J. Reddi, H. Yoon, and A. Knies, “Two Billion Devices and Counting,” IEEE Micro, vol. 38, no. 1, pp. 6–21, 2018.

Do we focus enough on mobile devices?

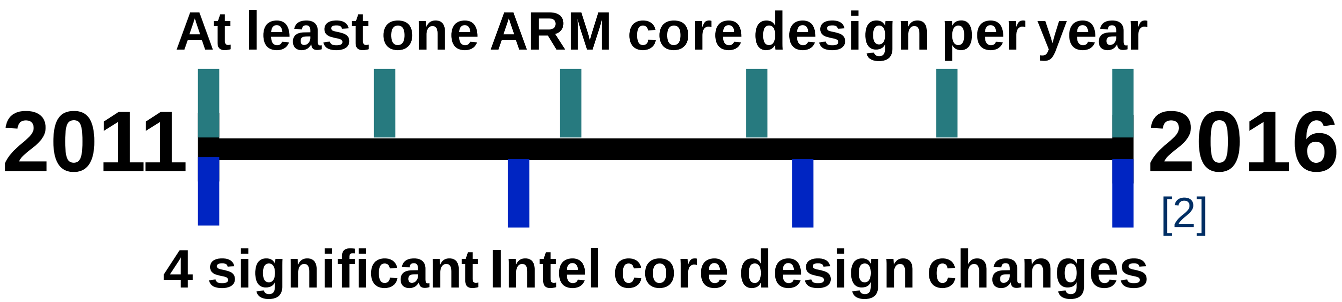
- Over 18 billion mobile devices
- Around 2 billion computers
- 1% of top tier publications focus on mobile computing (2018 study) [1]
- Mobile SoCs are distinct
  - Tight integration of hardware components
  - Significant heterogeneity
  - Rapid evolution



[1] V. J. Reddi, H. Yoon, and A. Knies, "Two Billion Devices and Counting," IEEE Micro, vol. 38, no. 1, pp. 6–21, 2018.

Do we focus enough on mobile devices?

- Over 18 billion mobile devices
- Around 2 billion computers
- 1% of top tier publications focus on mobile computing (2018 study) [1]
- Mobile SoCs are distinct
  - Tight integration of hardware components
  - Significant heterogeneity
  - Rapid evolution


At least one ARM core design per year

[2]

4 significant Intel core design changes

Mobile SoCs are different

[2] M. Halpern, Y. Zhu, and V. J. Reddi, "Mobile CPU's rise to power: Quantifying the impact of generational mobile CPU design trends on performance, energy, and user satisfaction," in 2016 IEEE HPCA

Which benchmarks can we use?

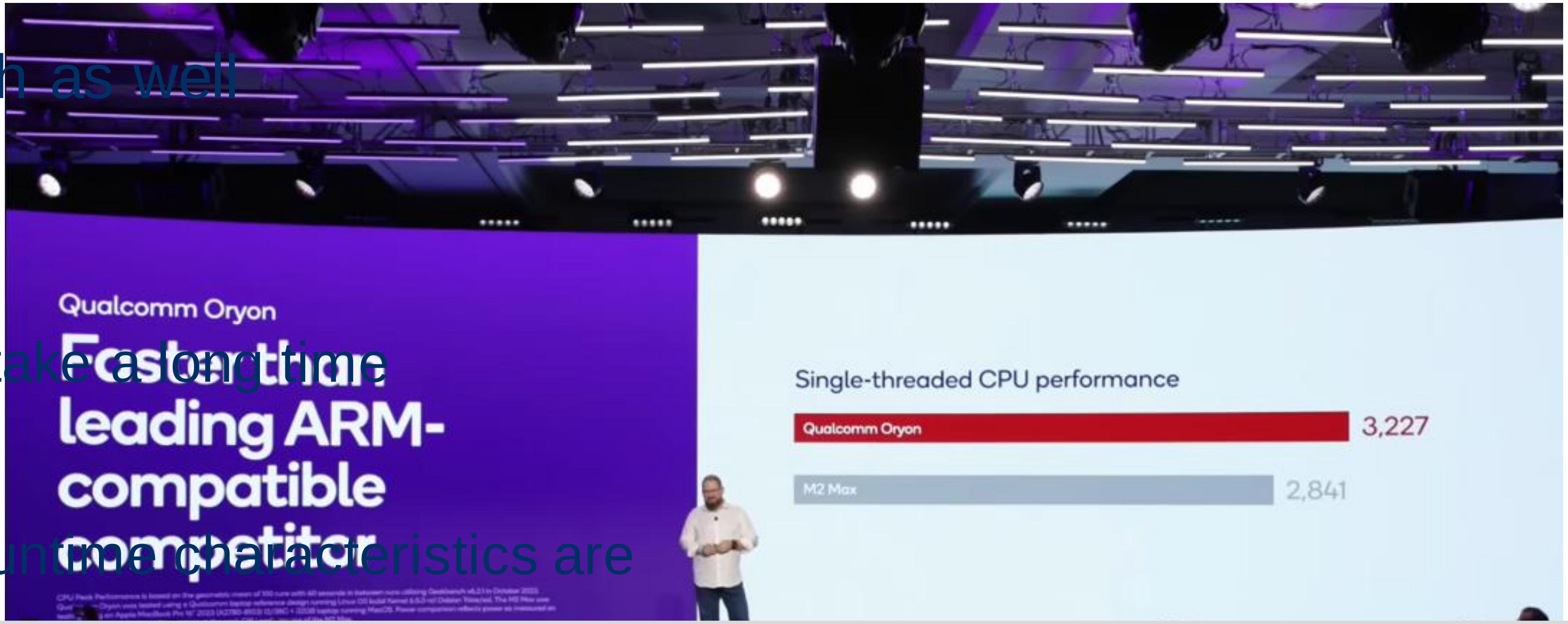
- Popular benchmarks (e.g., SPEC CPU, PARSEC) are not representative of mobile workloads [3]
- How about academic mobile benchmarks?
  - Narrowly focused on specific domains, thus limiting their utility
    - e.g., BBench for web browsing
    - e.g., ARBench for augmented reality
  - Difficult to keep them up-to-date and sometimes even working

How about commercial mobile benchmarks?

[3] M. Hayenga et al., "Accurate System-Level Performance Modeling and Workload Characterization for Mobile Internet Devices," in MEDEA 2008

Commercial mobile benchmarks

- Widely used by industry
- Used in academic research as well
- A lot of options
  - Difficult to choose
  - Running all of them would take a long time
- There’s one problem
  - We don’t know what their runtime characteristics are



Here’s where our work comes in

Our contributions

- Analysis of the runtime performance characteristics of commercial mobile benchmark suites
- Provide researchers with in-depth insights into the behavioural patterns
  - Judiciously select benchmarks aligned with their specific requirements
- Propose a representative benchmark set
  - Reduces execution time by 75%

Benchmarks

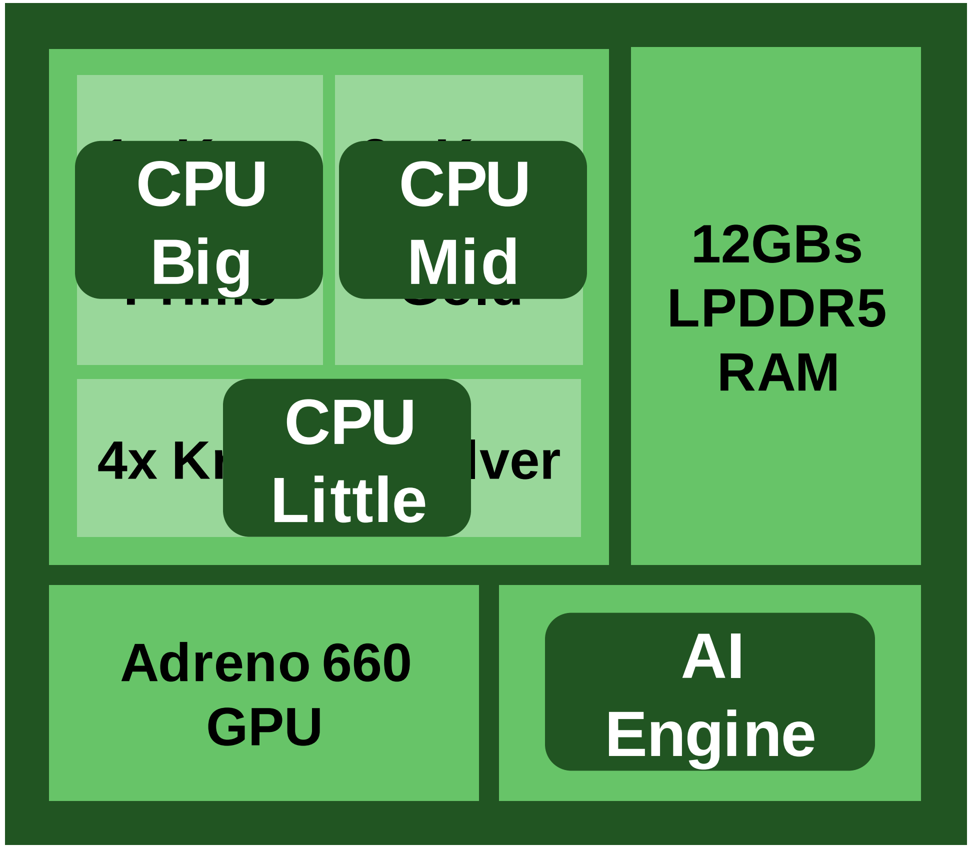
- 7 benchmark suites – 18 individual benchmarks

Benchmark Suite	Benchmark Name
3DMark	Slingshot
	Slingshot Extreme
	Wild Life
	Wild Life Extreme
Antutu	CPU
	GPU
	Mem
	UX
Aitutu	-

Benchmark Suite	Benchmark Name
Geekbench v5	CPU
	Compute
Geekbench v6	CPU
	Compute
GFXBench	High Level
	Low Level
	Special Tests
PCMark	Storage
	Work

Methodology

- Qualcomm Snapdragon 888 Board
  - Android 11
- Qualcomm Snapdragon Profiler
  - Over 190 metrics – CPU, GPU, AI Engine, Memory, Temperature



CPU Big

CPU Mid

12GBs LPDDR5 RAM

4x Kr

CPU Little

lver

Adreno 660 GPU

AI Engine

Analysis consists of 3 parts

Temporal Behaviour

Examine the values of the metrics across the entire benchmark's runtime

CPU Heterogeneity

Check the usage levels of the three CPU core clusters

Similarity & Redundancy

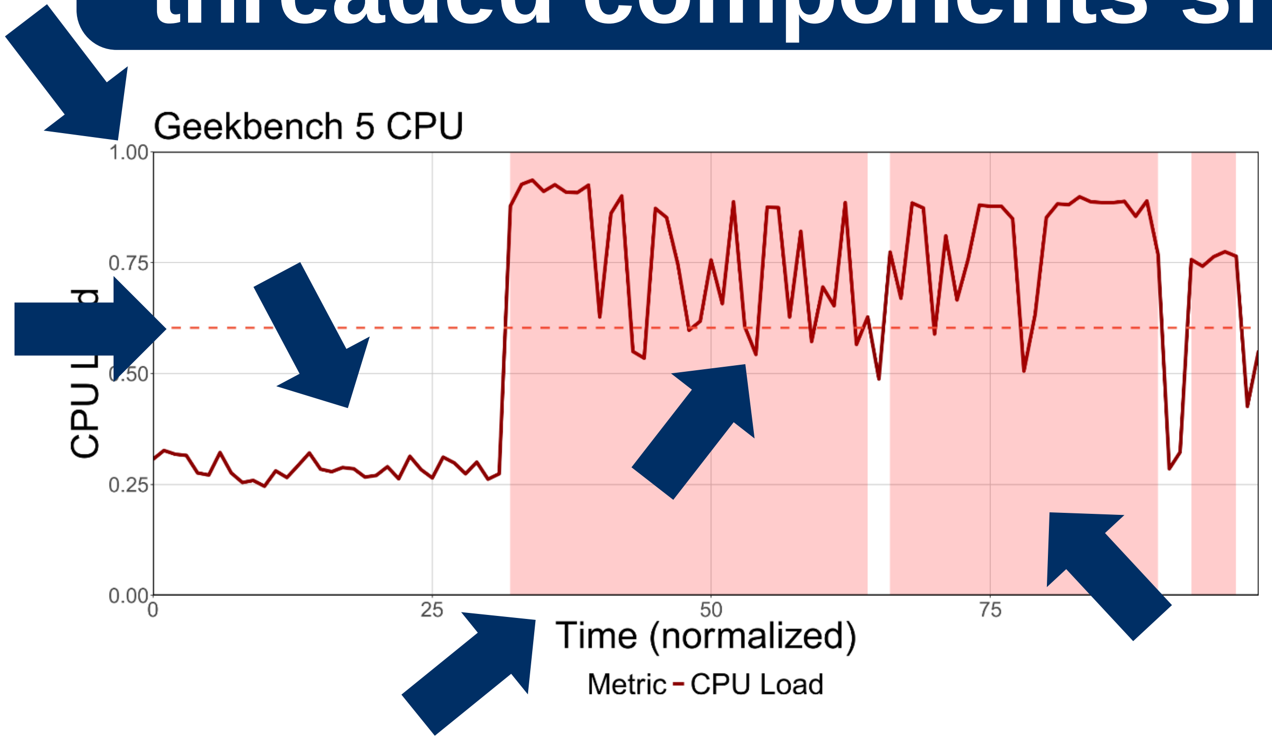
Evaluate how similar various the benchmarks are

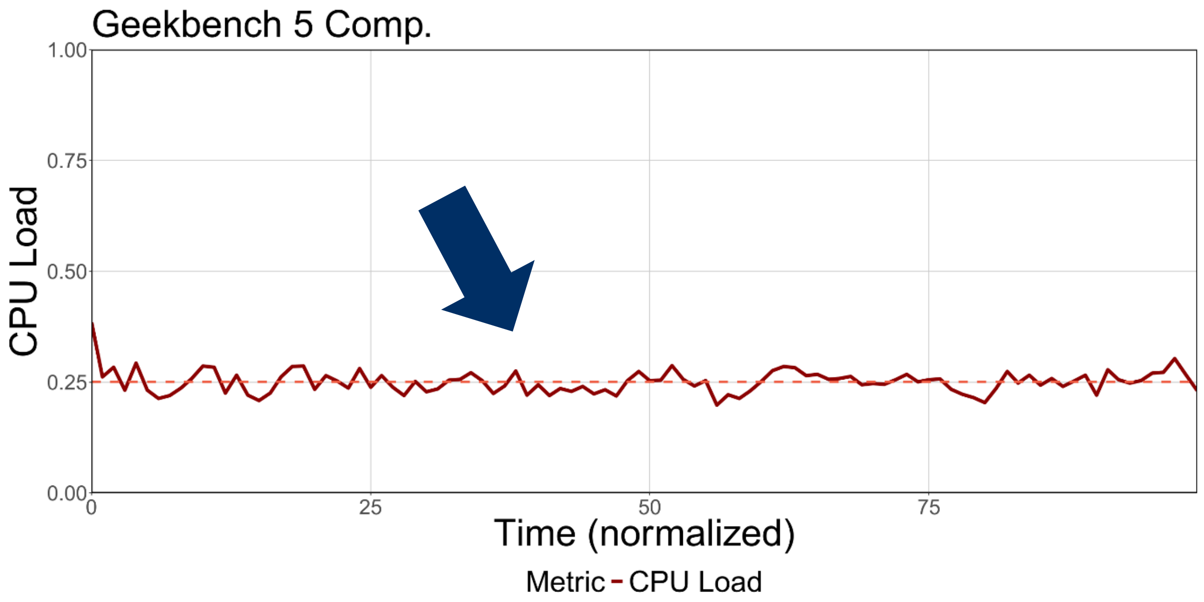
Let's look at some of the observations we made

Temporal Behaviour

Observation #1

Benchmarks that include multi-core or multi-threaded components show high CPU load levels

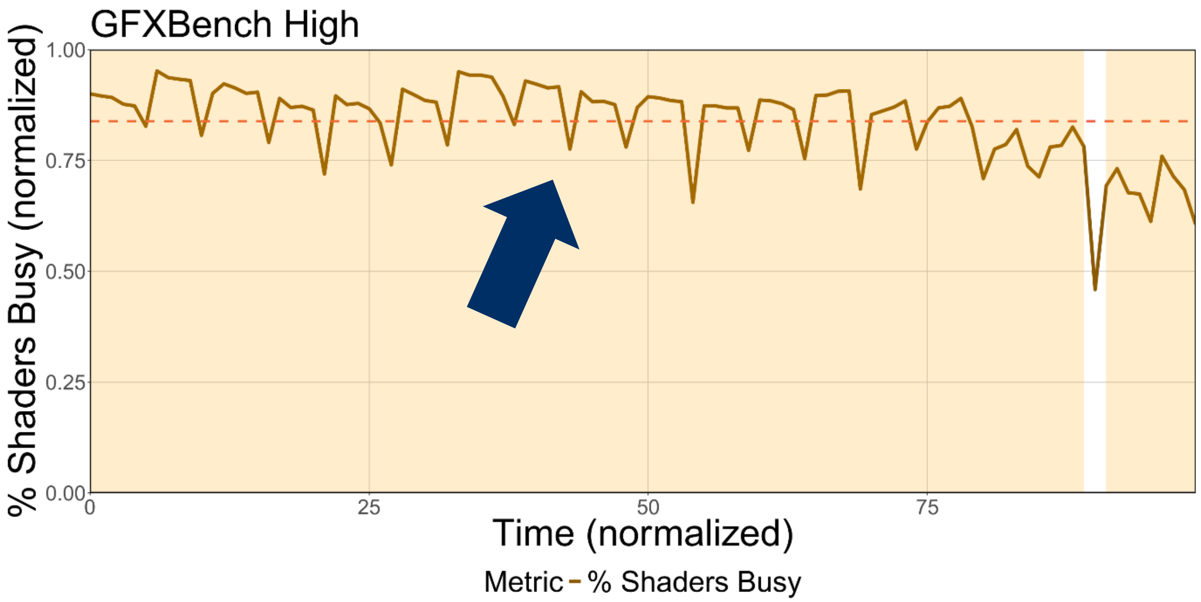


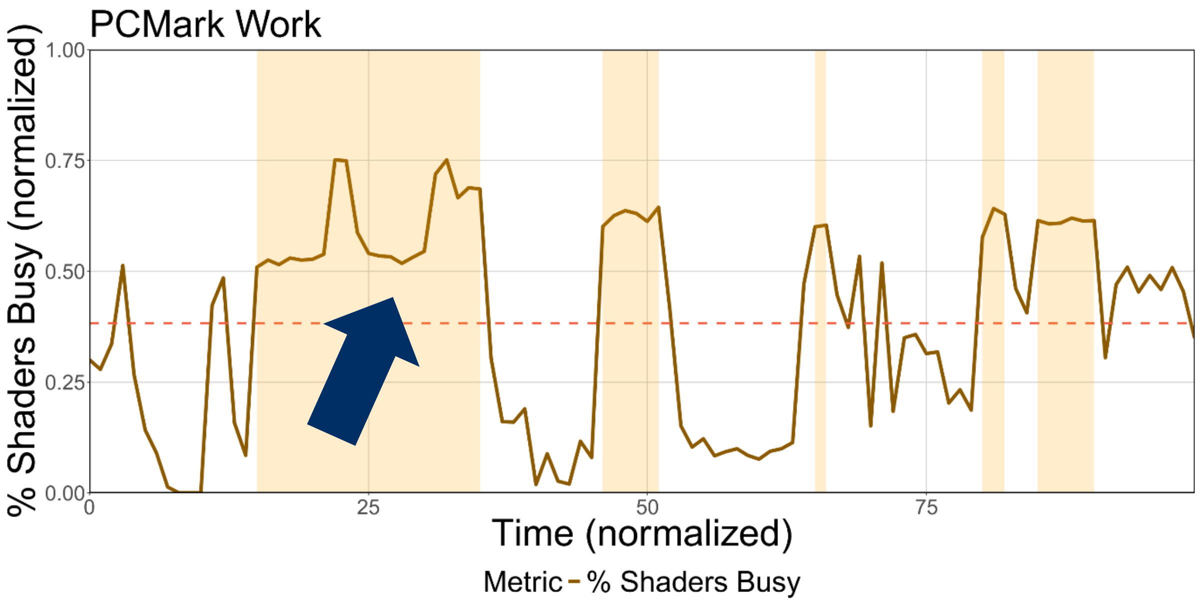


$$CPU\ Load = CPU\ Frequency * CPU\ Utilization$$

Observation #2

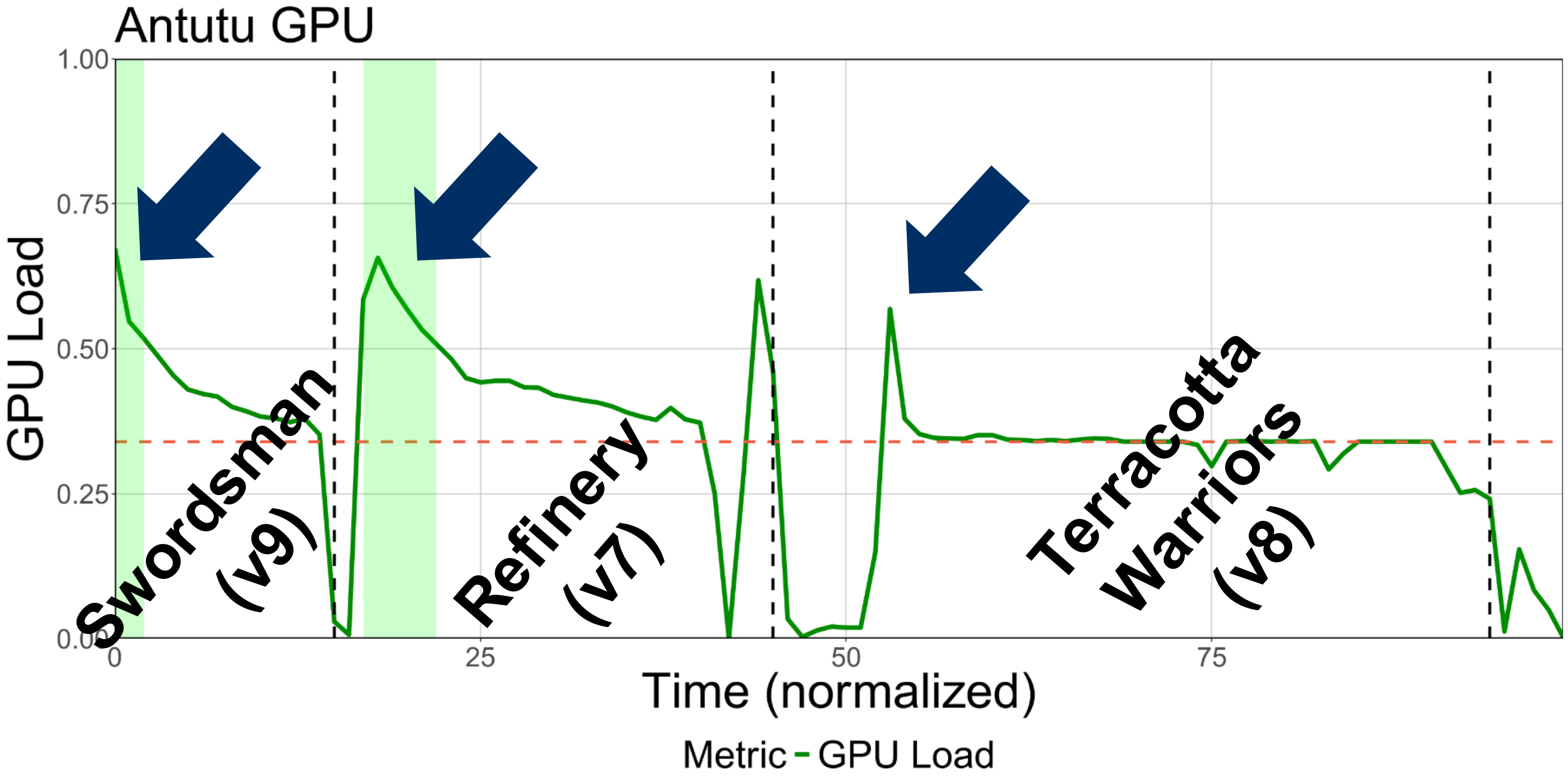
Usage of GPU resources is not limited to GPU-related benchmarks


GFXBench High


PCMark Work

Observation #3

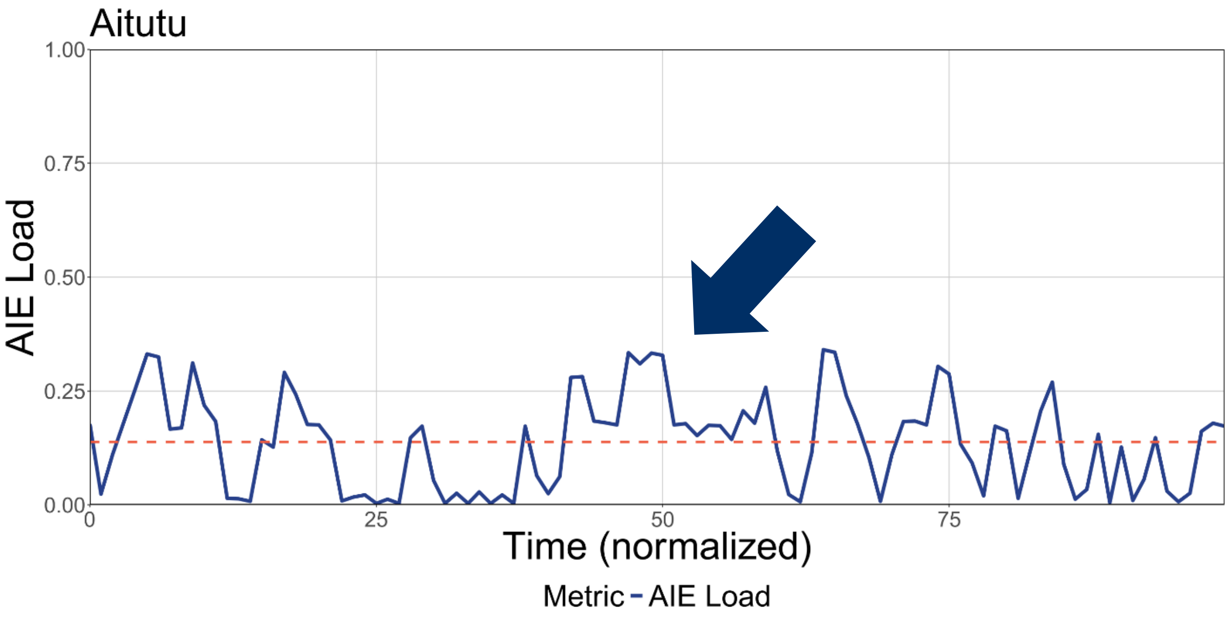
Newer benchmarks are not always more computationally intensive

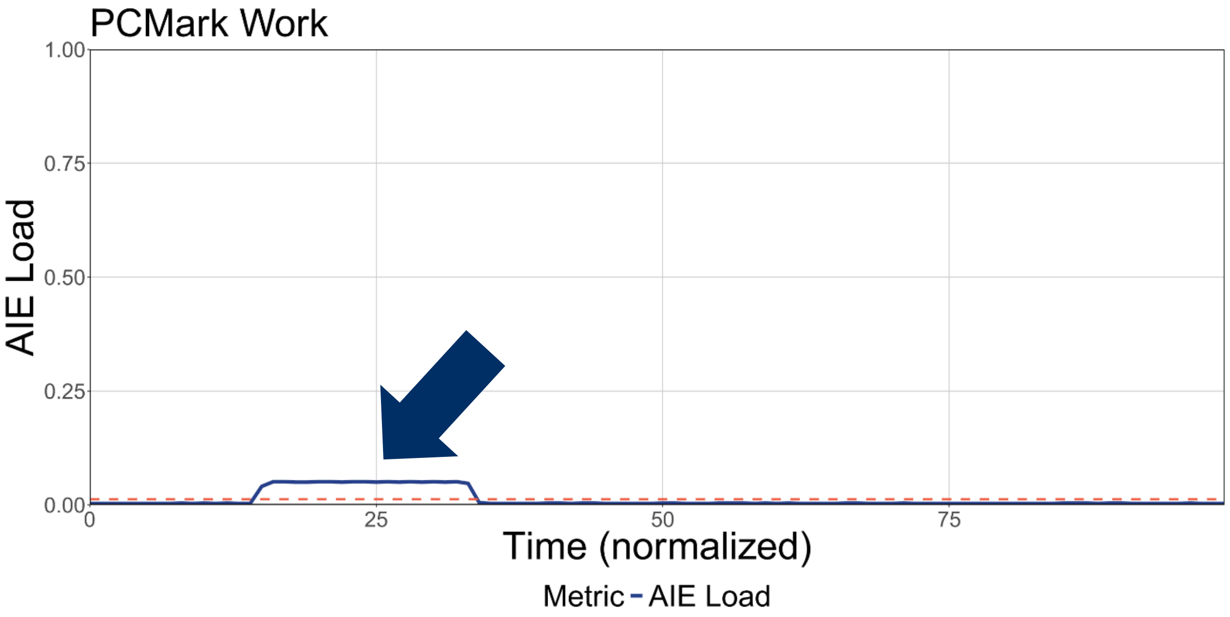


GPU Load = *GPU Frequency* ∗ *GPU Utilization*

Observation #4

Benchmarks make little use of the AI engine

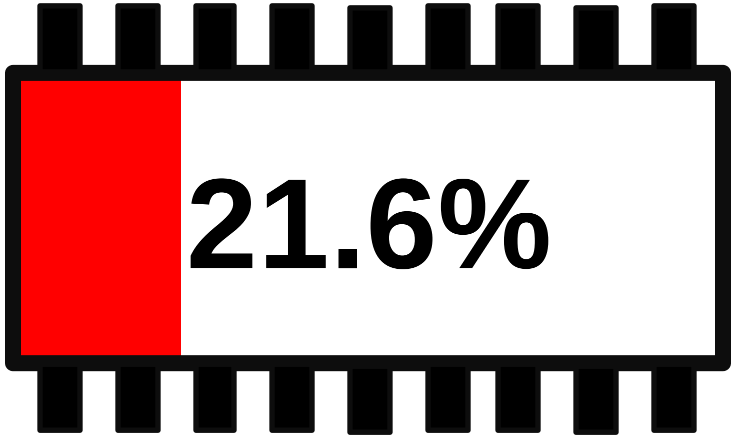




$$AIE\ Load = AIE\ Frequency * AIE\ Utilization$$

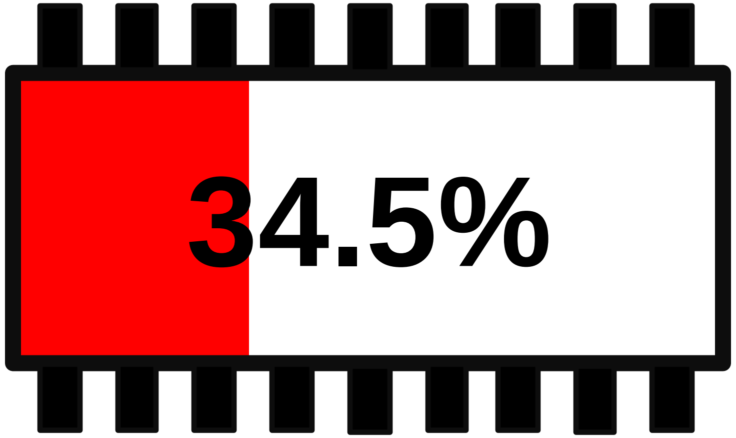
Observation #5

The memory footprint of benchmarks is moderate



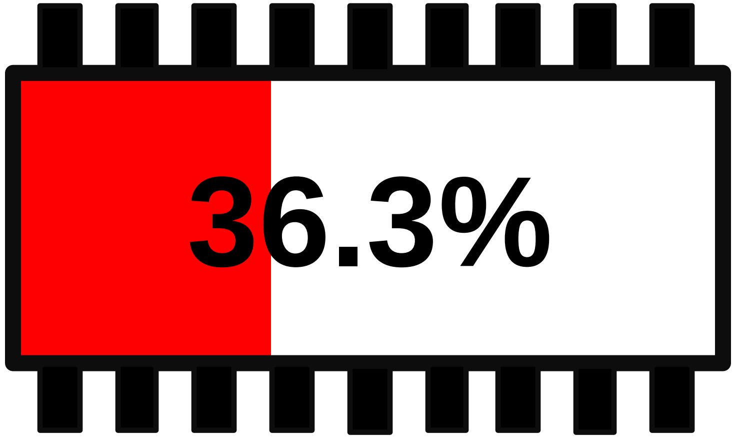
21.6%

Average System Memory Used



34.5%

Highest Average (3DMark Wild Life Extreme)



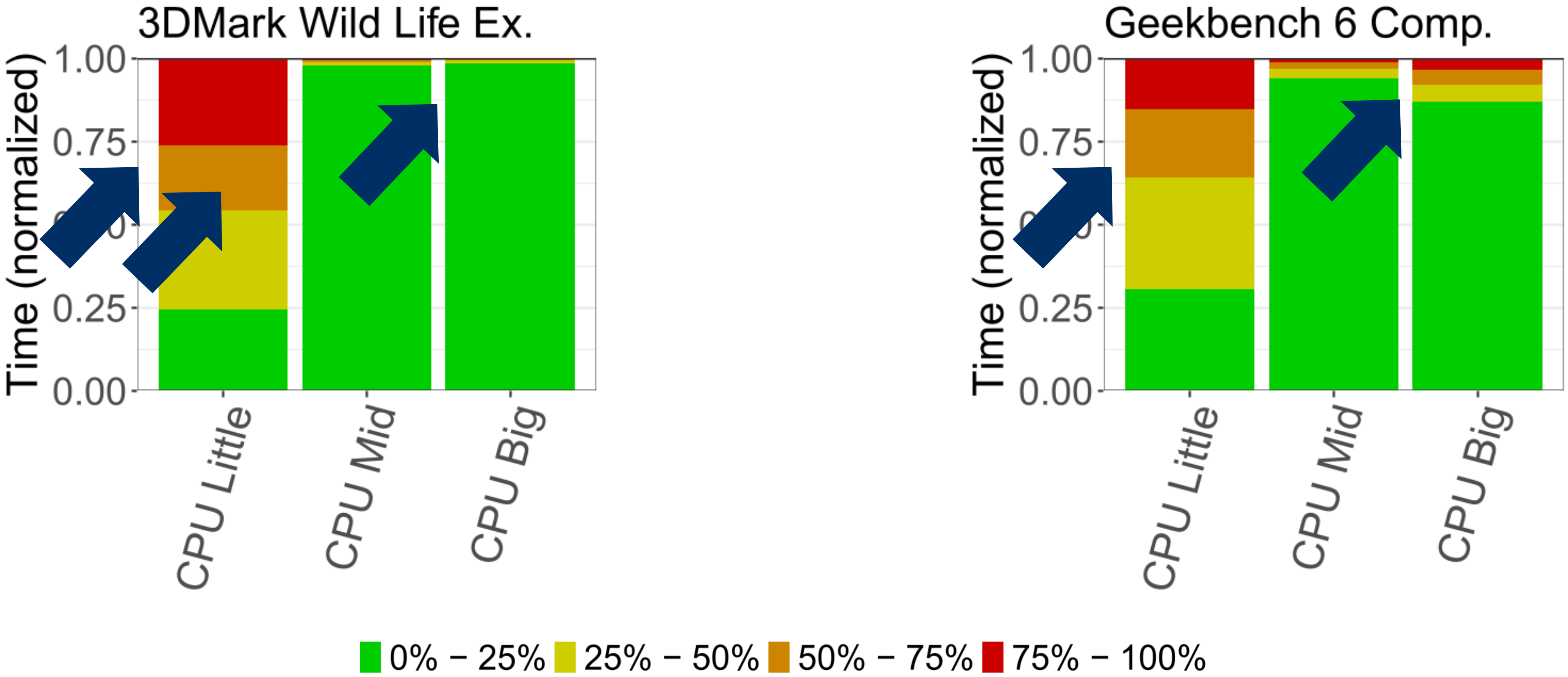
36.3%

Highest Memory Usage (Single Point)

CPU Heterogeneity

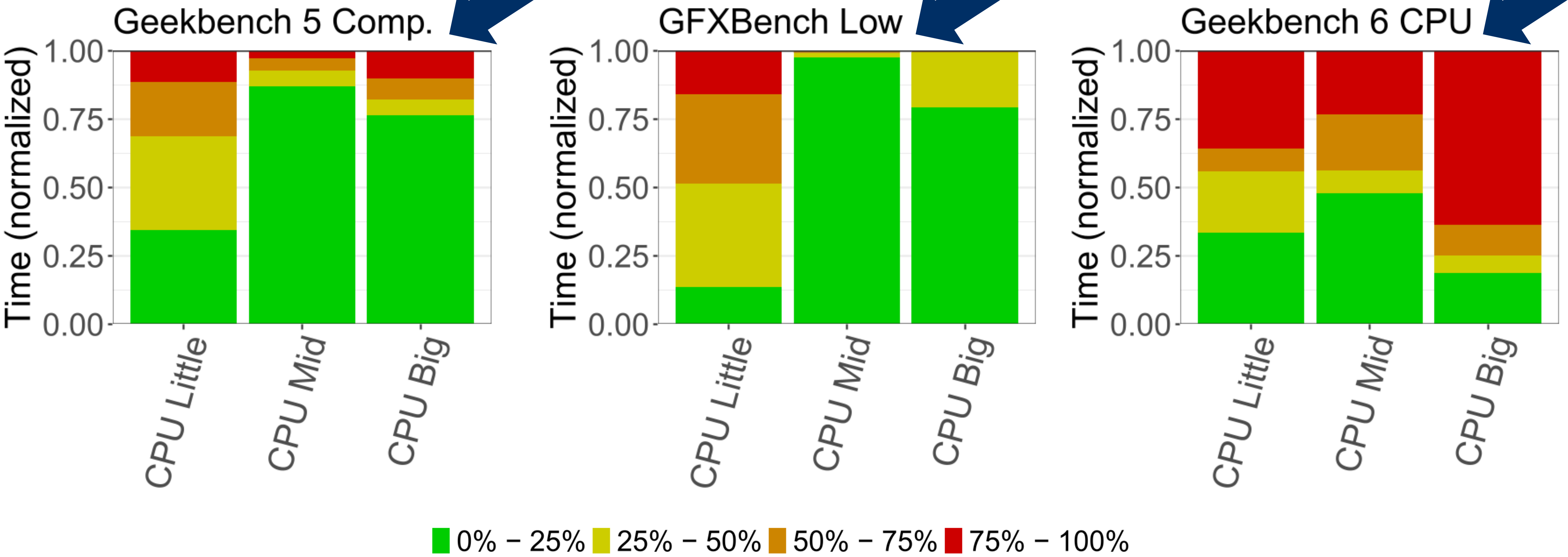
Observation #6

GPU tests tend to use only the energy-efficient cores



Observation #7

Workloads tend not to exploit more than one type of core concurrently



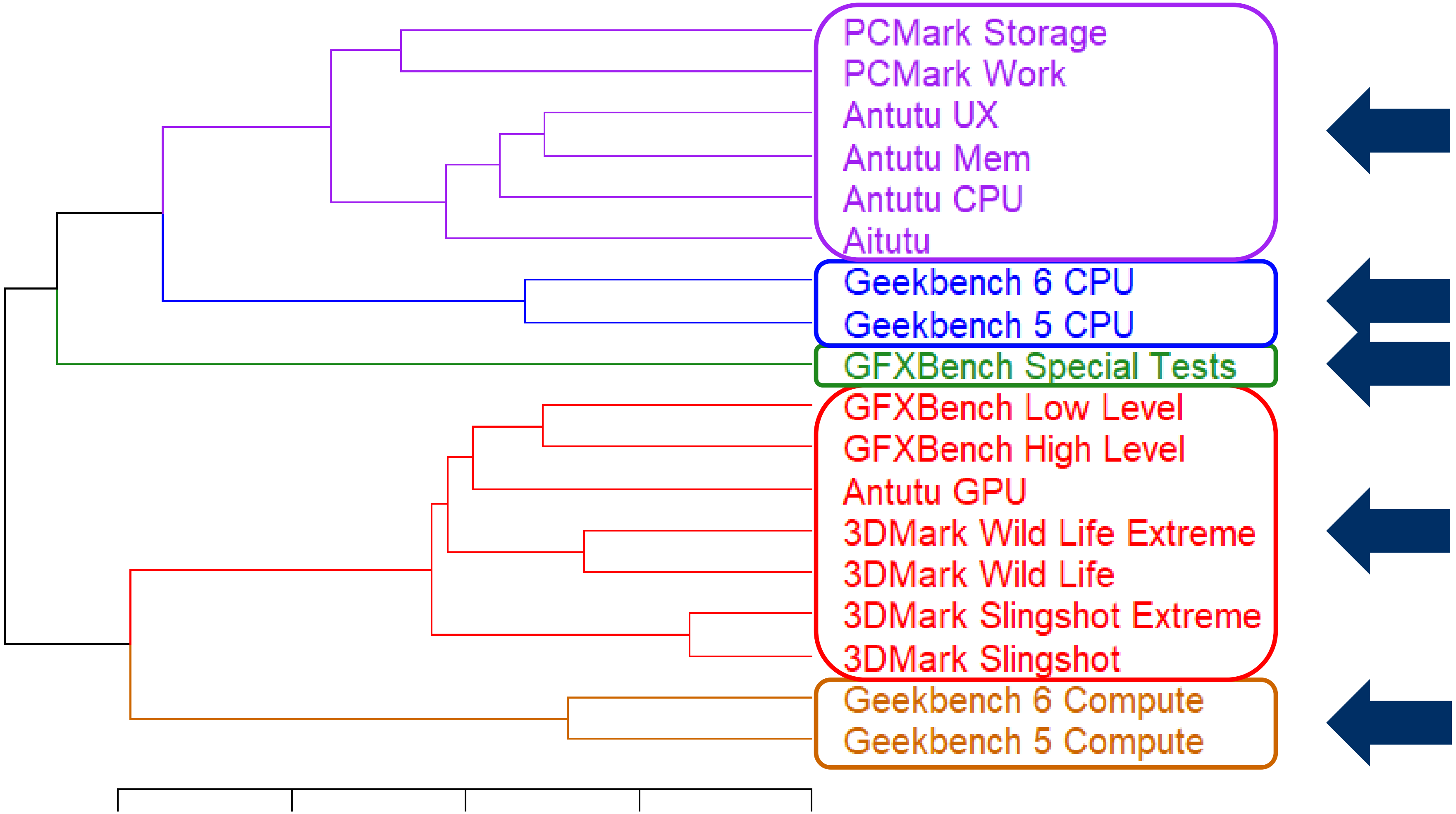
Similarity & Redundancy

Benchmark Similarity

- 18 benchmarks with many more sub-benchmarks
  - Over 110 minutes of runtime on a real device
  - Using simulators take a lot longer [4]
  - Finding similarity is a prerequisite to find redundancy
- 3 clustering algorithms
  - K-means
  - Partitioning Around Medoids (PAM)
  - Agglomerative Hierarchical Clustering
- How do we know the right number of clusters?
- **5 clusters** is the sweet spot

[4] A. Sandberg et al., “Full speed ahead: Detailed architectural simulation at near-native speed,” in 2015 IEEE International Symposium on Workload Characterization

Benchmark Similarity



Benchmark Redundancy

- All these benchmarks take a lot of time to execute
- Select a representative subset
  - Antutu – Covers all areas
  - GFXBench Special Tests – Highest AI engine load
  - Geekbench 5 CPU – Highest CPU load while stressing all CPU clusters
  - Geekbench 6 Compute – Highest GPU load

	Original Set	Reduced Set
Running Time (sec)	4429.5	1108.36
Running Time Reduction	-	75%

Conclusion

- We thoroughly explored commercial mobile benchmark suites
- Our analysis offers important insights for the computer architecture community
- The proposed representative benchmark set reduces execution time by 75%

I’ll be happy to answer your questions

Victor Kariofillis, Natalie Enright Jerger – viktor.karyofyllis@mail.utoronto.ca

UNIVERSITY OF TORONTO