Generational Mobile CPU Design Trends and Its Impact on Performance, Energy, and User Satisfaction *

Matthew Halpern Yuhao Zhu Vijay Janapa Reddi

The University of Texas at Austin, Department of Electrical and Computer Engineering
{matthalp, yzhu}@utexas.edu, vj@ece.utexas.edu

1. Introduction

Over the past decade, mobile CPUs have evolved from embedded processors to desktop-like single-chip multiprocessors. Yet, mobile CPUs, like any embedded processor, operate under a stricter set of power, thermal and energy constraints than their desktop counterparts. Therefore, the rapid design innovation, pervasiveness in society, and power-constrained nature of mobile hardware *necessitate the need to understand the implications of their current design trends on future designs*.

Our work quantitatively analyzes mobile CPU design trends and how they have impacted the end-user, hardware design, and the holistic mobile device. To that end, we conduct a measurement-based study across ten cutting-edge mobile CPU designs released throughout the past seven years. These mobile CPU designs span seven different microarchitectures, five major process technology nodes, different cache configurations and memory technologies, and also multicore designs.

To the best of our knowledge, our study is the *first* of its kind to rigorously evaluate mobile CPU design trends. In doing so, we make four key observations. Each observation is *quantitatively reinforced* – valuable in and of itself – regardless of whether it aligns with conventional wisdom or is surprising:

- **Desktop-like Scaling:** Mobile CPUs have adopted many of the high-performance mechanisms found in desktop CPUs. User satisfaction is latency-sensitive, which emphasizes the need for single-threaded performance improvements. However, the "low hanging fruit" (i.e. low-power) performance-oriented techniques are already being used in mobile CPUs. *Future hardware and software will need to understand how to identify and efficiently mitigate user-critical bottlenecks.*
- **Multicore CPUs:** Multicore CPUs are often characterized to be under-utilized in mobile applications [1]. However, we show they serve important roles to deliver end-user satisfaction. *User critical application functionalities are often multithreaded to leverage multicores speedups, and also background processes can interfere with user-facing application processes when there are not enough cores available.*
- **CPU Criticality:** Mobile applications are developed in general-purpose programming languages that primarily target the mobile CPU. Even for applications that utilize other SoC components, such as the GPU and image decoder, we show user satisfaction is dependent on CPU performance. Therefore, *general-purpose CPU design remains relevant as acceleration and heterogeneous execution catch on.*
- **Power Wall:** In contrast to other smartphone components, such as the display and radio, mobile CPU power consumption has risen excessively over time. Single-core power consumption has hit a power wall, and multicore power consumption can significantly surpass SoC-level TDPs without considering the rest of the SoC. SoC and device designers must *pay closer attention to CPU power consumption and understand the role of the CPU in the accelerator rich era.*

The long-term impact of our work is twofold. First, we, for the first time, quantitatively establish the onset of the "dark silicon" phenomenon [2,3] for mobile hardware. Our measurements reinforce the trend towards hardware specialization, but more importantly the general paradigm shift will change the role of the CPU in the "era of specialization". Our measurements show the system-level issues future mobile devices, and other "always on, always constrained" computer systems, such as autonomous vehicles, drones, robotics, and wearable devices that reuse many of the same smartphone SoC components.

2. Measuring Performance, Power, and Energy

Through systematic, quantitative measurement, our key takeaway is that mobile CPU designs are becoming increasingly indistinguishable from their desktop counterparts. By incorporating two decades of desktop CPU design techniques in a seven-year span, mobile CPU performance has dramatically improved at the expense of excessive power consumption. Most notably, *mobile CPUs have begun to hit a power wall,* leaving little room for scale up approaches in future designs.

2.1. Methodology

An important aspect of conducting any generational study is selecting the right "samples" to study. Our work focuses on ten ARM-based mobile CPUs released between 2009 and 2015 as found in *top-selling smartphones* for each year. We focus our study on ARM-based designs because they were (and continue to be) both the most popular and well-established line of mobile CPUs when we conducted our study [4]. For the smartphones released between 2012 through 2014, we study both stock (e.g. Samsung Exynos) and customized (e.g. Qualcomm Krait) ARM IP, which are two competing mobile CPU design strategies commonly employed today.

These smartphones we study, starting from the oldest, are the Motorola Droid (`D`), Samsung Galaxy S (`S`), Samsung Galaxy Nexus (`N`), and Samsung Galaxy S3, S4, S5, and S5 (`S3`, `S4`, `S5`, `S6`). We study the CPUs found within these devices holistically, meaning that we consider all of the processing subsystems that support general purpose compute (i.e. core and memory) together. Both the core and memory subsystems across these mobile CPUs have changed dramatically

*The original article appears under the title "Mobile CPU's Rise to Power: Quantifying the Impact of Generational Mobile CPU Design Trends on Performance, Energy, and User Satisfaction," published in the 22nd IEEE Symposium on. High Performance Computer Architecture, March 2015.

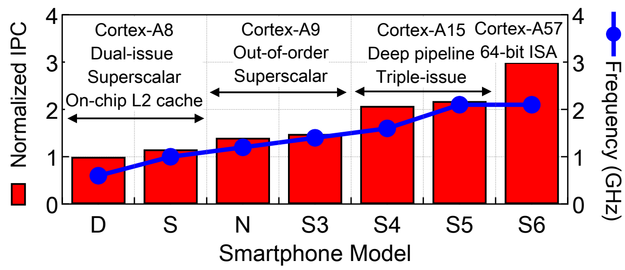

Figure 1: **Microarchitecture IPC and Clock Frequency Trends.**

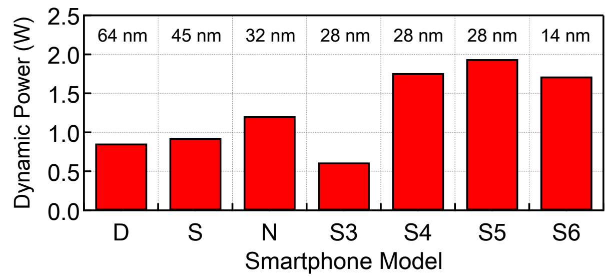

Figure 2: **Single-core Power Consumption Trends.**

improved over time, so we study their evolution in tandem.

We use a variety of benchmarking applications to extract our mobile CPU performance data. We construct our benchmarking suite based on reviewing prior work [5] and industry white papers [6], as well as conversations with industry architects [7]. The workloads we study include the EEMBC *Coremark*, a synthetic benchmark popular within industry, *Geekbench*, a popular mobile benchmarking suite, *Sunspider*, a long-standing JavaScript benchmark, *Stream*, a popular memory bandwidth benchmark, and a subset of the *SPEC* benchmark suite. Here, we show a subset of the data.

2.2. Mobile CPU Performance Trends

Overall, mobile CPU performance has dramatically improved throughout the mobile CPUs we study. Mobile CPU single-core performance has improved by 32% on average from generation-to-generation – resulting in a *10X* speedup over this seven-year period. We systematically attribute these performance improvements to two factors: frequency and microarchitecture. First, frequency scaling alone has fostered significant performance improvements across the mobile CPU generations. Fig. 1 shows the peak clock frequencies (as a line on the right *y*-axis) across the stock mobile CPU designs we study. The custom IP cores exhibit similar trends. The peak clock frequencies across the devices we study increase 3.5X, from the `D` (i.e. 600 MHz) to the `S6` (i.e. 2.1 GHz).

In addition to clock frequency, microarchitecture enhancements were also crucial to the observed performance improvements. This is demonstrated in Fig. 1 which shows the normalized IPC as bars on the left *y*-axis for each of the mobile CPU designs. Each bar is the the geometric mean of the workloads. The IPC increases by 3X across the CPUs. The IPC improvement is a result of microarchitecture designers progressively incorporating desktop CPU design techniques into mobile CPUs. In a seven-year span, the mobile CPUs we study encompass the transition from in-order to out-of-order pipelines, increases in pipeline depths and instruction issue widths, introduction of more aggressive speculation mechanisms (i.e. branch prediction and memory prefetching), and augmenting the memory hierarchy both on- and off-chip.

In the paper, we identify the most critical sources of microarchitectural performance improvements. We also show the importance of their continued improvements.

2.3. Mobile CPU Power and Energy Trends

Smartphones do not provide (or openly disclose) mechanisms to directly measure CPU power consumption. *We perform battery-level power measurements* to extract the mobile CPUs' power and energy trends based on the well-established differential power measurement methodologies [5].

Our measurements demonstrate that mobile CPU performance improvements have come at the expense of excessively high power consumption. Fig. 2 shows the average single-core power consumption across the workloads for each CPU. Most notably, we quantitatively observe that the *mobile CPU's single-core TDP has saturated at around 1.5 W*, similar to the 100 W power ceiling in mobile desktops. Process technology has significantly improved across the mobile CPUs we study (from 64 *nm* to 14 *nm*). However, as the technology scaling is projected to stop in just a few generations without a practical alternative, future mobile CPU designs will have to come up innovative ways for power-efficiency.

3. Bridging CPU Design and User Satisfaction

Despite the user-facing nature of mobile devices, current mobile CPU architecture research exclusively focuses on the hardware-software interface, largely ignoring the end-user. We believe this is because a rigorous methodology for quantifying end-user satisfaction *at scale* does not currently exist.

Our paper contributes a new *crowdsourcing*-based methodology for conducting large-scale user satisfaction study. Specifically, it reached to over *over 25,000* participants, orders of magnitude greater than in any prior work. Such a methodology allows us to bridge the mobile CPU design trends and their impact on end-user experience, a critical topic that was not well-understood in both industry and academia.

While emulating the different mobile CPUs we study, our participants evaluate a wide variety of applications that exhibit different types of user interaction and computational characteristics common to current (e.g. Angry Birds), and likely future (e.g. augmented reality), mobile applications. This methodology allows us to what mobile CPU design techniques are providing the most benefit to the end-user's quality of user experience. Additional details about the applications used and our user study methodology can be found in the original paper.

While the study resulted in several interesting insights, we would like to highlight two. First, the introduction of multicore CPUs has enabled a better end-user experience. While multicore CPUs are often under-utilized in mobile devices [1], there are user-critical application segements that can benefit from multi-threading. Second, mobile CPU performance remains user-critical even amongst applications that utilize the GPU and other fixed function accelerators. Therefore, the CPU remains a critical processing element to optimize for future (accelerator-rich) mobile systems.

4. Long-term Impact

Mobile computing has reached critical mass in recent years, having replacing traditional desktop-based personal computing. Therefore, mobile is no longer the exception, but the norm. However, mobile computing has been historically underrepresented in computer architecture research. Our work addresses this void by highlighting the challenges within this domain, establishing new research methodologies, and generating a wealth of valuable data for other researchers to utilize.

4.1. Mobile CPU Role in Emerging Computing Domains

Our measurements quantitatively demonstrate the onset of the "dark silicon" phenomenon [2, 3] in mobile hardware. Table 1 shows the manufacturer-estimated peak power consumption for SoC components. *While the SoCs are designed to sustain a 3.5 watt TDP, the underlying processing systems are capable of peak power consumption upwards of 2X this limit.* An emerging class of embedded computing systems, which include robots and drones, will further push the design requirements of mobile hardware. These domains require *sustained, high performance computation* while operating under the same thermal and energy constraints as mobile devices. Our study conveys the challenge for mobile SoCs, which are already "unsustainable by design" to be designed for these systems.

Our insights apply more broadly beyond smartphones to other mobile devices, wearables, and other embedding computing domains. For example, the Samsung Galaxy Tab 12.6 tablet utilizes the same SoC as the `S5` and Google Glass uses the same SoC as `N`. Further, these devices employ similar workloads and operate under the same constraints.

Therefore, we expect future SoCs to increasingly utilize hardware specialization to drive the performance and power efficiency improvements necessitated in future mobile systems. For instance, the peak power consumption the A15 CPU cluster in each SoC can easily exceed the SoC TDP budget on its own. Therefore, the "era of specialization" will need to redefine how CPUs are used in future mobile systems.

In our study, we posit what these roles will be so to guide future work to prepare mobile CPUs, and the rest of the mobile SoC, for when this paradigm is fully realized. Generally speaking, that the proliferation of specialized processing units will reduce the amount of compute performed by the CPU. The CPU will be invoked most frequently for irregular codes that do not map well to accelerators and to manage the system complexity and other resources as a whole. Nonetheless, the CPU will continue to be a long-standing target for code portability and backwards compatibility for other applications.

Table 1: Reported SoC TDP Numbers

Smartphone System-on-Chip	Samsung Galaxy S5 Exynos 5422	Samsung Galaxy S6 Exynos 7420
A7 Cluster	2.77 W	2.30 W
A15 Cluster	6.27 W	4.428 W
GPU	0.80 W	0.89 W
Other	0.50 W	0.50 W
Total	**10.34 W**	**8.12 W**
TDP Budget	**3.50 W**	**3.50 W**

4.2. Open Datasets to Foster New Research

Trend-based studies, specifically using real systems, help identify impactful research opportunities [2, 5, 8]. Providing data from these studies for open-use help future researchers identify impending bottlenecks and issues that may otherwise go unnoticed until it is too late. To aid this effort, we release two datasets from our study for other researchers to use:

Power, performance, and energy [9] We release our power, performance, and energy measurements. This data can provide value for computer architecture and mobile systems researchers whom wish to create back-of-the-envelope power modeling and for creating projecting of future trends. Additionally, this data can be used to quantitatively establish power, thermal, and energy targets for future systems and optimizations to adhere within.

User Satisfaction [10] We have also released the user satisfaction study data. While not our original intention, we are humbled to see that this dataset has already resulted in publishable work in other fields. A recent paper published in the Conference on Human Computation and Crowdsourcing (HCOMP) used our user satisfaction dataset to develop data sanitation techniques for subjective crowdsourcing tasks [11].

4.3. Citation from 2027

The authors present a historical analysis of mobile CPU design trends and their impact on performance, energy, and user satisfaction. Their measurements quantitatively demonstrated the onset of the dark silicon phenomenon in mobile computer architecture, establishing that the era of specialization is here for edge devices. Additionally, their work was the first to heavily emphasize methodologies to incorporate the end-user into mobile computer architecture evaluation.

References

[1] C. Gao, A. Gutierrez, M. Ranaj, R. Dreslinski, T. Mudge, and C.-J. Wu, "A study of mobile device utilization," in *Proc. of ISPASS*, 2015.

[2] H. Esmaeilzadeh, E. Blem, R. St Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Prof. of ISCA*, pp. 365–376, IEEE, 2011.

[3] G. Venkatesh, J. Sampson, N. Goulding, S. Garcia, V. Bryksin, and M. B. Taylor, "Conservation cores: Reducing the energy of mature computations," in *ACM SIGARCH Computer Architecture News*, 2010.

[4] Intel watches ARM as low-powered computing thrives.

[5] E. Blem, J. Menon, and K. Sankaralingam, "Power struggles: Revisiting the risc vs. cisc debate on contemporary arm and x86 architectures," in *High Performance Computer Architecture (HPCA2013), 2013 IEEE 19th International Symposium on*, pp. 1–12, IEEE, 2013.

[6] CoreMark Benchmarking for ARM Cortex Processors.

[7] R. Peri. Personal (email) communication, Nov. 23 2014. Intel.

[8] H. Esmaeilzadeh, T. Cao, Y. Xi, S. M. Blackburn, and K. S. McKinley, "Looking back on the language and hardware revolutions: measured power, performance, and scaling," in *ACM SIGARCH Computer Architecture News*, vol. 39, pp. 319–332, ACM, 2011.

[9] Mobile CPU Performance, Power, and Energy Data. https://github.com/Matthalp/mobile-cpu-perf-power-data.

[10] Mobile CPU User Satisfaction Data. https://github.com/Matthalp/mobile-cpu-user-sat-data.

[11] A. T. Nguyen, M. Halpern, B. C. Wallace, and M. Lease, "Probabilistic modeling for crowdsourcing partially-subjective ratings," in *Proceedings of The Conference on Human Computation and Crowdsourcing (HCOMP)*, 2016.