The Interplay of Transistors and Accelerators

Adi Fuchs
Princeton University
adif@princeton.edu

David Wentzlaff
Princeton University
wentzlaf@princeton.edu

ABSTRACT

The slowdown and the nearing end of transistor scaling play a vital role in the rise of domain-specific architectures. Domain-specific architectures rely on "accelerators" that are specialized chips designed to efficiently execute dominating computation patterns and improve performance under a given transistor budget.

Unfortunately, accelerators like FPGAs and ASIC chips are also implemented using transistors, and therefore abide by the same transistor limitations that motivated their deployment. The widespread of accelerators necessitates new ways to evaluate the impact of specialization in modern chips. We present a methodology that decouples specialization-driven benefits from transistor-driven benefits, and show that in popular accelerator applications specialization gains are less significant than transistor-driven gains.

CCS CONCEPTS

• **Computer systems organization** → **Architectures**; • **Hardware** → *Application specific processors*;

ACM Reference Format:
Adi Fuchs and David Wentzlaff. 2018. The Interplay of Transistors and Accelerators. In *MICRO-51 ACM Student Research Competition (ACM SRC'18).* ACM, New York, NY, USA, 2 pages.

1 BACKGROUND AND MOTIVATION

Domain-specific architecture design is one of the driving principles of the "new golden era for computer architecture" envisioned in the 2018 Turing award lecture [5], and it is motivated by three notable trends. (i) *New systems:* the widespread adoption of battery-limited mobile devices and cloud services that run on power-hungry warehouse-scale data centers and serve computations from billions of users [7, 8]. (ii) *Emerging applications:* recent popularity of compute-intensive workloads, such as machine learning, that can be efficiently mapped to specialized hardware [3]. (iii) *CMOS scaling limits:* while transistor density keeps increasing, transistor power stopped dropping at similar rates due to the failure of classical "Dennardian" CMOS scaling [12]. Consequently, chip power density rates keep increasing, and since the chip power is bounded by thermal design power (TDP) budget, the maximal number of active chip transistors is fixed, while the remaining transistors are un-utilized, forming a "dark silicon" regime [4]. Power became the limiting factor of the chip's transistor budget, which will stop improving following the end CMOS scaling, projected in 2021 [1, 6].

Accelerators In The Dark Silicon Era: The non-improving chip transistor budget necessitates means for efficient processing. Therefore, domain-specific architectures that rely on specialized "accelerator" chips became a prominent solution to ameliorate the gap between growing computation demands and stagnating processing capabilities [10]. The benefits of accelerator-centric architectures come from computation-specific chip specialization that enables the co-designing of software and hardware to support the minimal computational functionality and improve processing capabilities by orders-of-magnitude under a given transistor budget.

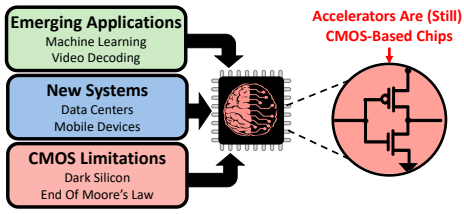


Figure 1: Accelerator Motivations and CMOS-Dependence.

Accelerators In The Post-Moore's Law Era: As Figure 1 illustrates, accelerators are also implemented using transistors, and abide by the same limitations posed by the end of CMOS scaling that motivated their development. Since there is a finite number of ways to map a given computation problem to hardware, the benefits of chip specialization under a given physical chip budget are also limited and must be adequately evaluated.

Evaluating Accelerators: An accelerator-ubiquitous future calls for new methods and processor evaluation metrics. We develop an accelerator-centric evaluation methodology that decouples specialization-driven benefits from transistor-driven benefits. We construct an application-independent physical chip model using contemporary CMOS scaling equations and datasheets from thousands of chips. Using our model, we construct the scaling equations for the gain functions an accelerator strives to maximize, e.g., performance or performance-per-power. We define the *"Chip Specialization Returns" (CSR)* metric that quantifies the gains of an accelerator under the physical chip budget, for a given application and gain function. Using the CSR metric, we analyze accelerators and answer the following question: *"how much did an accelerator improve the target function compared to its silicon potential?".*

2 PHYSICAL CHIP MODEL

Our physical chip model provides an estimate for the gain of a chip (e.g., energy efficiency) given its physical properties: CMOS process, thermal design power (TDP), chip frequency, and transistor count. We constructed the model in the following phases:

(i) Device-Level Scaling: We use contemporary CMOS studies [9] and projections for 5nm CMOS from the recent IRDS report [6] and construct a device scaling model, shown in Figure 2.

(ii) Transistor Count and Thermal Budget Modeling: As accelerator transistor count is not always disclosed, we approximate the number of chip transistors. We use datasheets from thousands of commercial GPUs and CPUs [2, 11], and construct the model of transistor count as a function of chip transistor density, (Die Area/CMOS Node^2) as shown in Figure 3a. To account for the number of active transistors under a TDP, we model the relation of TDP, transistor count and operation frequency, shown in Figure 3b.

(iii) Chip-Level Gain Modeling: We integrate the transistor count, TDP, and device power and speed, to model chip-level gains. The gain is the target function a chip strives to maximize. Figure 4 shows the chip-level model for performance and energy-efficiency. Our methodology can be used for other gain functions as well.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ACM SRC'18, October 2018, Fukuoka, Japan
© 2018 Copyright held by the owner/author(s).

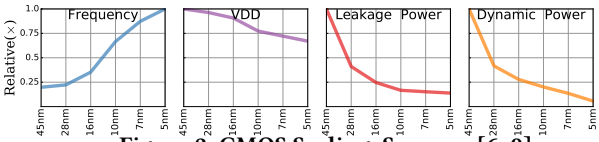
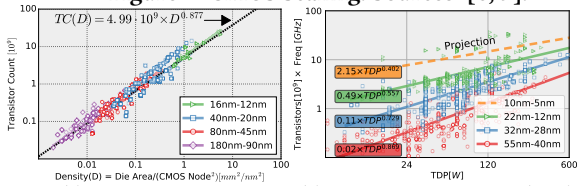
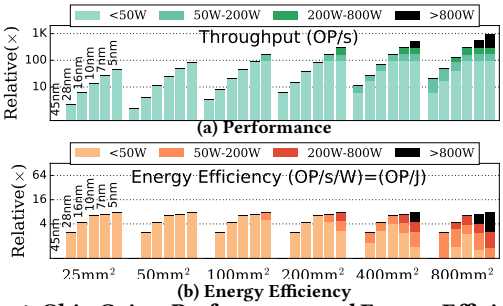


Figure 2: CMOS Scaling. Sources: [6, 9].



(a) Transistor Count (b) Thermal Design Power (TDP)

Figure 3: Chip Transistor Count and Power Budget Models.



(a) Performance

(b) Energy Efficiency

Figure 4: Chip Gains: Performance and Energy Efficiency for Different CMOS Nodes, TDPs, and Die Sizes for: $f = 1GHz$.

3 QUANTIFYING SPECIALIZATION GAINS

Using the chip budget described in Section 2, we can estimate the transistor-driven gains, and decouple them from an accelerator's specialization gains. We define the "Chip Specialization Returns" (CSR) metric in Equation 1. The CSR quantifies the transistor-independent gain of an accelerator:

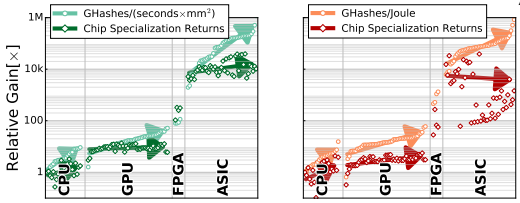
$$Gain = \underbrace{\frac{Gain}{Gain(Physical)}}_{CSR} \times Gain(Physical) \quad (1)$$

By quantifying gains under a given CMOS budget, CSR measures the benefits of chip specialization for an application and gain function. Equation 2 shows the comparative approach we use on sets of accelerators. Given an accelerator, A, and a baseline accelerator, B, we compare their obtained gains and physical gain potentials, and estimate whether gains are transistor-driven, i.e., $Gain(Phy.)$ dominated or specialization-driven, i.e., CSR dominated:

$$\frac{Gain_A}{Gain_B} = \frac{\frac{Gain_A}{Gain(Phy._A)} \times Gain(Phy._A)}{\frac{Gain_B}{Gain(Phy._B)} \times Gain(Phy._B)} = \frac{CSR_A}{CSR_B} \times \frac{Gain(Phy._A)}{Gain(Phy._B)} \quad (2)$$

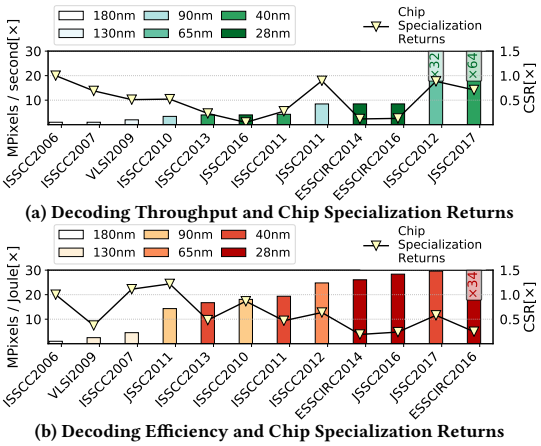
We conduct a chip specialization study on popular FPGA and ASIC applications. For each application, we quantify its transistor-driven gains and decouple them from its transistor-independent gains (i.e., its CSR). We present the subset of our study in Figures 5+6. Figure 5 shows how transitioning from CPUs to ASICs improved Bitcoin mining gains orders-of-magnitude, but the CSR trends hint that for each specialization architecture (e.g., within ASICs) specialization does not improve, and gains are mainly achieved due to better physical capabilities (e.g., new and efficient CMOS nodes). The same trends are shown in Figure 6 for video decoding ASICs.

The empirical gains for the accelerators evaluated were mainly achieved due to the improvement in CMOS technology, and gains under given chip budget, reflected by the CSR, were orders-of-magnitude less significant. These trends put in question the role of accelerators as a long term remedy for the end of Moore's law.



(a) Performance Scaling (b) Energy Efficiency Scaling

Figure 5: CPU, GPU, FPGA, and ASIC Bitcoin Miner Trends.



(a) Decoding Throughput and Chip Specialization Returns

(b) Decoding Efficiency and Chip Specialization Returns

Figure 6: Academic Video Decoding ASICs Scaling Trends.

ACKNOWLEDGMENTS

This material is based on research sponsored by the NSF under Grants No. CNS-1823222 and CCF-1453112, Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under agreement No. FA8650-18-2-7846 and FA8650-18-2-7852. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA), the NSF, or the U.S. Government.

REFERENCES

[1] R. Courtland. 2016. Transistors could stop shrinking in 2021. *IEEE Spectrum.*
[2] Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, and Mark Horowitz. 2012. CPU DB: Recording Microprocessor History. *Queue* 10, 4.
[3] Jeff Dean, David A. Patterson, and Cliff Young. 2018. A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution. *IEEE Micro* 38, 2.
[4] Hadi Esmaeilzadeh, Emily Blem, Renee St. Amant, Karthikeyan Sankaralingam, and Doug Burger. 2011. Dark Silicon and the End of Multicore Scaling. In *Intl. Symp. on Computer Architecture (ISCA '11).*
[5] John L. Hennessy and David A. Patterson. 2018. A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. In *Intl. Symp. on Computer Architecture (ISCA '18).*
[6] IRDS. 2017. International Roadmap for Devices and Systems 2017 Edition.
[7] Svilen Kanev, Juan Pablo Darago, Kim Hazelwood, Parthasarathy Ranganathan, Tipp Moseley, Gu-Yeon Wei, and David Brooks. 2015. Profiling a Warehouse-scale Computer. In *Intl. Symp. on Computer Architecture (ISCA '15).*
[8] V. J. Reddi, H. Yoon, and A. Knies. 2018. Two Billion Devices and Counting. *IEEE Micro* 38, 1.
[9] Aaron Stillmaker and Bevan Baas. 2017. Scaling equations for the accurate prediction of CMOS device performance from 180 nm to 7 nm. *VLSIJ-58.*
[10] Michael B. Taylor. 2012. Is Dark Silicon Useful?: Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse. In *Proceedings of the 49th Annual Design Automation Conference (DAC '12).*
[11] TechPowerUp. 2018. GPU and CPU Database. https://www.techpowerup.com.
[12] Ganesh Venkatesh, Jack Sampson, Nathan Goulding, Saturnino Garcia, Vladyslav Bryksin, Jose Lugo-Martinez, Steven Swanson, and Michael Bedford Taylor. 2010. Conservation Cores: Reducing the Energy of Mature Computations. In *Intl. Conf. on Arch. Support for Programming Languages & Operating Systems (ASPLOS '10).*