Modeling of Novel Transistors, Manufacturing Technologies, and Architectures to Preserve Digital Computing Performance Scaling

George Michelogiannakis, *Member, IEEE,* Dave Donofrio, *Member, IEEE,* and John Shalf, *Member, IEEE*

Abstract—The approaching end of traditional MOSFET technology scaling creates the need to identify novel devices, manufacturing technologies, memories, and architectures to preserve digital computing performance scaling. To this end, we argue for the need to generate circuit-level models and integrate them into existing simulation and digital design infrastructure for rapid architectural design exploration. Using CHISEL, we can generate behavioral and circuit models of novel technologies.

1 Introduction

Recent years have brought us closer to the end of traditional MOSFET technology scaling. Traditional CMOS is now predicted to initially slow down and eventually cease scaling by the beginning or middle of the next decade [13], with industry forecasting that technologies beyond 2nm or 3nm may be infeasible or impractical [9], [10]. This realization does not mean the end of performance scaling for digital computing, but rather an invitation to preserve performance scaling by adopting novel CMOS devices, manufacturing technologies, memories, and architectures. The approaching end of lithographic scaling threatens decades of DOE investment in hardware and software, as well as threatens to hinder advancement in numerous scientific and society challenges that depend and will continue to depend on digital computing [1].

To create a tangible strategy, each of novel technology should be evaluated in the architectural and eventually the system levels. This will allow answering such questions as the feasibility of the increased parallelism that tunnel FETs (TFETs) [7] require to improve performance, how to alleviate reliability challenges by novel devices using architectural techniques, how to 3D stack logic and memory layers to reduce data movement and heat density, what is the impact of TB-level nonvolatile memory on top of processing logic to programming models and power management, as well as many other similar questions. Answering such questions requires architectural-level modelling and evaluation and helps make best use of emerging technologies as well as guide each technology's future progress. Whats more, evaluating novel devices in the architectural level allows studying the potential of architectural techniques such as specialization (use of accelerators), adopting non-Von Neumann architectures, and the impact they have to the software layer such as programming models.

To enable this crucial design space exploration, existing architectural- and system-level simulators need to be updated with circuit-level models for the aforementioned novel technologies. This effort includes generating those circuit-level models from low-level models such as voltage–current curves for transistors, and must include important characteristics such as reliability in addition to performance and energy. In this position paper, we make the case for extending a hardware description language (HDL) such as CHISEL [5], to generate both software and hardware architectural models that include new technologies.

• *The authors are with the Computer Architecture Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94702.*
E-mail: {mihelog,ddonofrio,jshalf}@lbl.gov

PMES Workshop, Salt Lake City, 14 Nov 2016. `http://j.mp/pmes2016`

2 Background and Related Work

Several novel devices (transistors) have recently been fabricated and demonstrated, and are promising candidates to replace MOSFETs. Carbon nanotube transistors (CNFETs) have demonstrated a 1000× improvement of the energy-delay product (EDP) for memory-bound applications, 10× EDP improvement for compute-bound applications, and 30× for mixed workloads [3]. In addition, TFETs and negative capacitance FETs operate at a lower voltage than equivalent MOSFET transistors [4], [11]. Each new device introduces different tradeoffs such as performance at low and high voltages, reliability, energy, and others, which need to be carefully evaluated in the architectural level to properly assess impact. At the same time, new manufacturing technologies such as 3D stacking [16] of multiple logic and memory layers are quickly becoming feasible, but their higher-level impact and potential are not readily apparent. These options, combined with new memory technologies each with a different set of tradeoffs such as magnetic RAM [8], resistive RAM [2], create a vast landscape of options to preserve digital computing performance scaling.

Numerous alternatives exist for architectural-level software simulation, such as Gem5 [6]. Recently, HDLs such as PyMTL [12], Bluespec [14], and CHISEL [5] were proposed that can generate both behavioral (software) and circuit-level (hardware) models from a single code base. Software models can execute autonomously, much like a software simulator, while hardware models have to go through synthesis and placement to produce silicon or be placed on an FPGA. These new HDLs provide powerful means for rapid design exploration of large-scale architectures, but currently lack support for novel technologies. Past work has taken the first step towards evaluating new technologies [15], but only evaluated new devices and only for a 32-bit adder instead of representative future architectures. Similar infrastructure is being developed for alternative computation models such as neuromorphic and quantum, but the focus of this paper is modelling new technologies for digital computing.

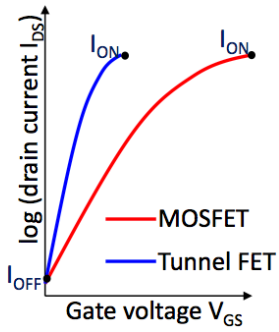

Fig. 1. Based on current–voltage curves of new devices such as the one shown for TFETs, we need to generate circuit-level models suitable for architectural simulation.

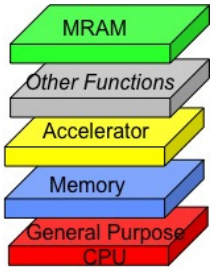

Fig. 2. Different combinations of memory and logic layers change distances and available bandwidth between layers.

3 Circuit-Level Models of Novel Technologies

Generating circuit-level architectural models of new devices requires deriving energy and delay per operation from low-level voltage and current curves (Figure 1), and a given set of assumptions such as operating temperature. Moreover, these models should include error rate, variability, and other aspects important to adopting new devices. The same is true for modelling new memory technologies. Different memories have different access times and energy per access. Those may depend on where data is located in the memory array as well other active requests to the memory. In addition to different error rates, some memory technologies are also non-volatile. This is a critical property of new memories that also needs to be understood in the architectural and system levels. Such models enable us to better gauge the impact to the memory hierarchy and power management of having large amounts of non-volatile memory near or on top of future processors. This contradicts current programming assumptions that non-volatile memory is distant and expensive to access.

3D integration of multiple kinds of layers also affects distances and relevant performance–cost tradeoffs as shown in Figure 2. In that picture, the choice of the kind of logic (e.g., accelerators or general-purpose) and memory for each layer affects the latency, energy, and available bandwidth for any two blocks to communicate. In addition, some combinations of layers may not be feasible due to high heat density, and some combinations may be a poor choice if they stress the limited bandwidth available between layers. 3D integration should be a parameter in the architectural models of our infrastructure in a way that each memory or logic block can be quickly placed in different layers, and the energy cost and available bandwidth between layers can be readily calculated based on technology models.

Finally, specialized architectures such as accelerators, GPUs, or fixed-function blocks should also be easily modelled. This can be done either by implementing the specialized architecture in an HDL, or more easily by creating abstract blocks with configurable delays and energy costs to perform a certain action. The action can be specified in terms of result even in high-level functional language, without having to describe the detailed hardware to generate the result. Essentially this allows simulation of a circuit where different blocks are modelled in different levels of detail. This implies that different blocks must have different levels of models, and that the level of abstraction of each component must be selected carefully. This will enable a quick exploration of potential new specialized architectures before fully implementing them.

4 Integration to Existing Infrastructure

We consider that extending a HDL such as CHISEL is a promising avenue to realizing the aforementioned goals. CHISEL generated both hardware and software models from a single code base, both of which are necessary for a complete study. Currently, CHISEL uses a backbone where the code description is transformed into a graph, and then converted to the desired output. CHISEL's backbone can be extended to allow a choice of which devices and other technologies to use, and the performance, energy, and reliability models for each. The same is true for specialized architectures which can be defined as black boxes with associated cost and performance models.

The performance and cost models can be incorporated into CHISEL's software (simulation) models such that performance models affect timing and delays during the simulation, and cost models also affect the reported power consumption after all events during the simulation are recorded. For hardware models (e.g., Verilog) generated by CHISEL, performance and cost models will both affect timing and placement during synthesis. Both hardware and software models can be used to estimate heat density and system-level error rates, based on relevant circuit-level models.

As a next step, we can develop a system-on-chip using CHISEL that includes cores as well as a complete memory hierarchy (including caches). This serves as a testbed for experiments on programming models, compilers, and algorithms. This is a necessary step to capture implications that novel technologies have on programming models, stemming from reliability, an increased need for parallelism, non-volatile nearby memories, and other potential aspects on top of typical performance–cost tradeoffs.

5 Conclusion

To respond to the approaching end of MOSFET technology scaling and preserve performance scaling of digital computing, we need to create reliable circuit-level performance and cost models of emerging technologies to use in architectural and system studies. This will allow us to evaluate new technologies in the architectural scale which will better guide technology development as well as motivate changes in the architecture and software. In this position paper, we argue for the need to develop such models for novel devices, memories, 3D integration, and specialized architectures. In addition, we briefly describe how to integrate these models into existing infrastructure to perform the necessary architectural- and system-level evaluations and gauge the impact of new technologies to the architecture and software.

Acknowledgments

This work was supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE- AC02-05CH11231.

REFERENCES

[1] S. Aaronson, *Computer Science – Theory and Applications: Second International Symposium on Computer Science in Russia, CSR 2007, Ekaterinburg, Russia, September 3-7, 2007. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ch. The Limits of Quantum Computers, pp. 4–4.

[2] H. Akinaga and H. Shima, "Resistive random access memory (ReRAM) based on metal oxides," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2237–2251, Dec 2010.

[3] M. M. S. Aly, M. Gao *et al.*, "Energy-efficient abundant-data computing: The N3XT 1,000x," *Computer*, vol. 48, no. 12, pp. 24–33, Dec 2015.

[4] U. E. Avci, D. H. Morris, and I. A. Young, "Tunnel field-effect transistors: Prospects and challenges," *IEEE Journal of the Electron Devices Society*, vol. 3, no. 3, pp. 88–95, May 2015.

[5] J. Bachrach, H. Vo *et al.*, "Chisel: Constructing hardware in a scala embedded language," in *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, June 2012, pp. 1212–1221.

[6] N. Binkert, B. Beckmann *et al.*, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, Aug. 2011.

[7] K. Boucart and A. M. Ionescu, "Threshold voltage in tunnel FETs: physical definition, extraction, scaling and impact on IC design," in *ESSDERC 2007 - 37th European Solid State Device Research Conference*, Sept 2007, pp. 299–302.

[8] I. B. M. Dason, V. R. Kumar, and A. A. Kirubaraj, "Realization of magnetic RAM using magnetic tunneling junction in atomic level," in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, vol. 4, April 2011, pp. 397–401.

[9] S. Dhar, M. Pattanaik, and P. Rajaram, "Advancement in nanoscale CMOS device design en route to ultra-low-power applications," *VLSI Des.*, vol. 2011, pp. 2:1–2:19, Jan. 2011. [Online]. Available: http://dx.doi.org/10.1155/2011/178516

[10] N. Z. Haron and S. Hamdioui, "Why is CMOS scaling coming to an END?" in *2008 3rd International Design and Test Workshop*, Dec 2008, pp. 98–103.

[11] M. H. Lee, J. C. Lin *et al.*, "Ferroelectric negative capacitance hetero-tunnel field-effect-transistors with internal voltage amplification," in *2013 IEEE International Electron Devices Meeting*, Dec 2013, pp. 4.5.1–4.5.4.

[12] D. Lockhart, G. Zibrat, and C. Batten, "PyMTL: A unified framework for vertically integrated computer architecture research," in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture*, Dec 2014, pp. 280–292.

[13] C. A. Mack, "Fifty years of moore's law," *IEEE Transactions on Semiconductor Manufacturing*, vol. 24, no. 2, pp. 202–207, 2011.

[14] R. Nikhil, "Bluespec system verilog: efficient, correct RTL from high level specifications," in *Formal Methods and Models for Co-Design, 2004. MEMOCODE '04. Proceedings. Second ACM and IEEE International Conference on*, June 2004, pp. 69–70.

[15] D. E. Nikonov and I. A. Young, "Overview of Beyond-CMOS devices and a uniform methodology for their benchmarking," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec 2013.

[16] Q. Zhu, B. Akin *et al.*, "A 3D-stacked logic-in-memory accelerator for application-specific data intensive computing," in *3D Systems Integration Conference (3DIC), 2013 IEEE International*, Oct 2013, pp. 1–7.