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 threats 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 non-volatile 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. What 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.

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.
in an HDL, or more easily by creating abstract blocks with

can be done either by implementing the specialized architecture

between layers can be readily calculated based on technology
different layers, and the energy cost and available bandwidth

a way that each memory or logic block can be quickly placed in

parameters in the architectural models of our infrastructure in

bandwidth available between layers. 3D integration should be

combinations may be a poor choice if they stress the limited

two blocks to communicate. In addition, some combinations of

affects the latency, energy, and available bandwidth for any

accelerators or general-purpose) and memory for each layer

Figure 2. In that picture, the choice of the kind of logic (e.g.,
tances and relevant performance–cost tradeoffs as shown in

migration of multiple kinds of layers also affects dis-
tances 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 as-

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 dis-
tances 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

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 dis-
tances and available bandwidth between layers.

4 INTEGRATION TO EXISTING INFRASTRUCTURE

We consider that extending a HDL such as CHISEL is a promis-
ing 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 per-
formance, 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 algo-
rithms. 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 inte-
gration, and specialized architectures. In addition, we briefly
describe how to integrate these models into existing infrastruc-
ture 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


