2017 IEEE 35th International Conference on Computer Design

Very Low Voltage (VLV) Design

Ramon Bertran*, Pradip Bose*, David Brooks[+], Jeff Burns*, Alper Buyuktosunoglu*, Nandhini Chandramoorthy*, Eric Cheng[$], Martin Cochet*, Schuyler Eldridge*, Daniel Friedman*, Hans Jacobson*, Rajiv Joshi*, Subhasish Mitra[$], Robert Montoye*, Arun Paidimarri*, Pritish Parida*, Kevin Skadron[#], Mircea Stan[#], Karthik Swaminathan*, Augusto Vega*, Swagath Venkataramani*, Christos Vezyrtzis*, Gu-Yeon Wei[+], John-David Wellman*, Matthew Ziegler*

[*]IBM T. J. Watson Research Center, Yorktown Heights, NY

[+]Harvard University, Cambridge, MA

[$]Stanford University, Palo Alto, CA

[#]University of Virginia, Charlottesville, VA

***Abstract*—This paper is a tutorial-style introduction to a special session on: Effective Voltage Scaling in the Late CMOS Era. It covers the fundamental challenges and associated solution strategies in pursuing very low voltage (VLV) designs. We discuss the performance and system reliability constraints that are key impediments to VLV. The associated trade-offs across power, performance and reliability are helpful in inferring the optimal operational voltage-frequency point. This work was performed under the auspices of an ongoing DARPA program (named PERFECT) that is focused on maximizing system-level energy efficiency.**

Keywords—low power design, voltage scaling, frequency scaling, real-time performance, energy efficiency, system reliability.

I. Introduction

In the late CMOS era, traditional Dennard scaling rules of transistor sizes and supply voltage levels do not apply. In particular, supply voltage scaling has slowed drastically, and clock frequency growth has all but stalled. With nominal voltage operating points hovering around the 0.9 to 1.0 volt range, past research in near-threshold voltage (NTV) design [1] has largely remained confined to academic research. However, in recent years, with the rapid emergence of the Internet-of-Things (IoT) revolution, as well as cognitive edge computing (and associated accelerators), there has been a renewed demand for very low voltage (VLV) designs. In the higher end server space, wide operating range (WOR) designs that can offer robust performance and functionality across a wide voltage range (e.g. 0.28 V to 1.2 V as in [2]) has been pursued in industrial design – with the objective of utilizing a single design across a wide range of computing products (from low-cost embedded to high performance, server-class processors). Even within a given application domain, a WOR design enables the processor to achieve the best possible energy efficiency, while satisfying varying performance demands across the target applications.

However, one of the key obstacles to robust functionality at very low voltages is the difficulty in achieving high yield and robust functionality of storage elements (e.g. SRAM caches and buffers as well as latches) – which constitute a very significant proportion (typically ~50% or in some cases even more) of the chip area. Also, excessive clock frequency degradation due to aggressive voltage scaling may not be acceptable for application phases that require high single-thread performance (e.g. to meet real-time deadlines in embedded systems).

In this article, we provide a comprehensive summary of the latest research in circuit, micro-architecture and tool innovations that are paving the way for VLV and WOR functionality. Particular innovations in voltage noise mitigation and adaptive guard banding are summarized here to emphasize the circuit-microarchitecture co-design techniques that are key in this context. Also, recent and prior innovations in very low V_{min} SRAM design are referred to.

The work reported in this lead article, along with the companion papers [13-15] in this special session, represents research pursued under the auspices of an ongoing DARPA program (named PERFECT [27]). The objective in this program is to maximize system-level energy efficiency. The target domain is embedded processor systems geared toward applications like real-time-constrained image and video analytics. The latter application domain is representative of embedded processing requirements in unmanned aerial vehicles (UAVs) engaged in defense operations. The developed technologies are of course immediately applicable to commercial markets as well, e.g. in sectors like autonomous (self-driving) land vehicles (cars, buses, trucks and trains).

II. VLV Impediments

As pointed out in Section I, robust functionality and yield of SRAM arrays at NTV or VLV are one key impediments to driving down the operating voltage. This challenge is posed due to variability and functional yield difficulties at low voltages. Another challenge is soft error rate (SER) sensitivity to voltage scaling [10, 11]. The critical charge required to flip a stored storage bit datum (e.g. from a 1 to a 0 or vice versa) is referred to as Q_{crit} and it is extremely sensitive to the voltage level. As Q_{crit} decreases sharply with voltage, it becomes easier for an incident charged particle (e.g. cosmic neutrons or package-sourced alpha particles) to cause a bit-flip in an SRAM or latch element. A third impediment to lowering the voltage is the degradation in operating clock frequency, because of the increase in circuit delay (in logic) or access latency (for storage elements). Voltage noise effects in logic circuits as well as storage elements pose another challenge to minimizing operating voltage for a given target frequency. In

1063-6404/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCD.2017.105

a system architectural design that pursues VLV features for efficiency, the above are assumed to be the main impediments – as far as our work under the DARPA PERFECT program is concerned.

III. Reducing SRAM Vmin

Our recent work at IBM Research (Joshi et al. [8]) uses dynamic voltage boosting techniques to enable extreme low voltage operations in 14nm 8T SOI FinFET SRAM. The technique exploits the unique capacitive coupling effect in a FinFET device to dynamically boost the virtual macro supply voltage during active mode. This reduces the effective access latency, while enabling very low voltage retention states. Write assist techniques are also incorporated in the design to further improve yield. Hardware measurements show a 2.5 – 3x access time improvement at lower voltages and a functional V_{min} as low as 0.3V. Alternate methods that combine capacitive and inductive boosting for the first time, have also been experimented with more recently [7]. Across multiple test chip experiments using different boost mechanisms, effective functional V_{min} levels down to 0.26 V have been demonstrated in the laboratory by Joshi et al. At this time, our team is working on implementing a test chip (Eldridge et al. [12]) that uses this low-V_{min} SRAM technology to demonstrate its use in an efficient, yet accurate machine learning accelerator. In this context, it would be good to also note that pioneering work on low-V_{min} 8T SRAM design (without boost techniques) was pursued at IBM Research a decade ago [33]. (See also [9] for a very good survey of pioneering low power strategies during that period).

IV. Controlled Undervolting

Once the circuit-level V_{min} has been reduced by focusing on the storage element designs (i.e. latches, buffers, register files and SRAM caches), the actual lowest operating voltage point ($V_{min\text{-}act}$) in a WOR processor is still limited by uncertainties due to voltage noise, device aging effects and so on. This means that for the minimum operating clock frequency (f_{min}) to provide reliable functionality, the actual minimum voltage $V_{min\text{-}act}$ would have to be set as follows:

$$V_{min\text{-}act} = V_{min} + V_g$$

Here V_g is a voltage guard-band that is built in to the supply voltage to account for worst-case voltage noise, long-term aging effects, etc. Adaptive under-volting refers to techniques where the supply voltage is dropped below $V_{min\text{-}act}$ (thereby cutting into the guard-band). This provides efficient, yet reliable operation for most of an application run. In rare time segments where the voltage noise spikes to threaten circuit timing integrity, the voltage could be boosted up (on-demand), or the clock frequency could be dialed down [4, 5, 6]; or, some sort of micro-architectural throttling could be invoked to control the voltage spiking [3, 6]. Our Minerva DNN-accelerator research [22, 23] uses a particularly novel set of adaptive under-volting heuristics to achieve impressive low-voltage functionality without the use of any special low-V_{min} SRAM macros.

Such sophisticated, rapid-response voltage-frequency control loops are beginning to appear in real products; but we are pursuing advanced research in our project to achieve more aggressive, application-directed under-volting without compromising overall system resilience. These features will be demonstrated when the VELOUR chip [12] moves to the ASIC silicon implementation phase.

V. Software Support to VLV Design

The software support to VLV design consists of: (a) pre- and post-silicon software tools to help define, model and synthesize such chips; and (b) system software (e.g. compilers and operating systems) to enhance power-performance efficiency and/or system resilience. Under the first category of software support, we have developed the SHIVA and CLEAR toolset, as depicted in Figure 1.

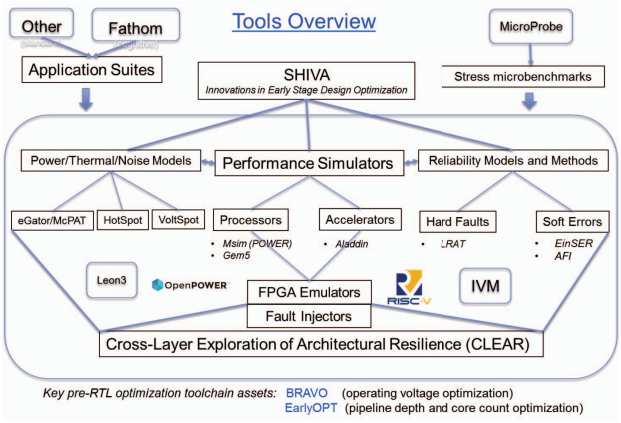

Figure 1. Tools developed in support of resilient VLV design

The central module within the SHIVA framework is the pair of cycle-accurate simulators that is provided to the user for pre-RTL processor definition. One of these is a POWER architecture simulator, patterned after IBM's POWER8 processor core [16]. We also provide access to a generic open-source cycle-accurate simulator (gem5) [17] with multi-ISA support. Also, application-specific accelerator simulation capability is available through Aladdin [18]. Modular integration of power, thermal, voltage noise and reliability models are available. The RTL-calibrated eGator reference POWER processor power model from IBM [19] along with an accuracy-enhanced open-source McPAT power model [20] is available as part of the SHIVA package. The HotSpot (temperature) and VoltSpot (voltage noise) models from University of Virginia are described in more detail in the companion paper by Roelke et al. [13]. Within the category of reliability models, the LRAT module provides analysis capability for lifetime reliability with regard to hard fault incidence; and, EinSER is focused on soft error rate (SER) models associated with high energy particle strikes (single event upset scenarios).

A representative use case that illustrates the worth of an integrated toolset like SHIVA is that of determining the optimal operational voltage point for a given processor in the context of an input application workload. The BRAVO methodology [32] uses the features within SHIVA to determine the point of voltage optimality. In doing so, fundamental trade-offs in balancing performance against

energy efficiency and reliability are presented. Similarly, the EarlyOpt methodology represents another key use case, where the optimal pipeline depth and optimal core count are inferred for use by the design team in concept-phase definition work. This latter methodology represents work that is still in progress; it is targeted for completion before the final release of the SHIVA toolset in early 2018.

Within the CLEAR [14] capability, FPGA emulation and associated high-speed statistical fault injection characterization capability is available across a range of architectures (e.g. Leon3, ARM, POWER, RISC-V and Illinois Verilog Machine or IVM). The CLEAR (cross-layer exploration of architecting resilience) framework is referred to again in the next section and is separately described in some detail in the accompanying paper by Cheng et al. [14]. This is a key toolset for pre-silicon design definition and cross-layer optimization thereof.

Within the SHIVA framework, the MicroProbe toolset [25, 26] provides a unique facility for testing out the resilience corners of the target system being modeled. It is an automatic micro-benchmark generation software toolkit, which is easily retargetable across different instruction set architectures (ISAs) and microarchitectures. Using MicroProbe, one can generate stress-tests: e.g. maximum power and maximum voltage noise micro-benchmarks. These capabilities make MicroProbe an indispensable aid in designing power-efficient (and yet reliable) processors. This technology has been transitioned for use in real product development within IBM [3, 26]. In addition, performance verification test cases can be generated by MicroProbe to diagnose performance bugs in the system.

Generation and characterization of application workloads of relevance to the PERFECT program is an integral part of our design methodology. As shown in Figure 1, in addition to the synthetic stress micro-benchmarks generated by MicroProbe, we also rely on systematically generated application benchmark suites. The Fathom suite [28] represents our team's timely contribution in this regard; because this is the first benchmark suite that represents the domain of machine learning. In specific areas like UAV-driven image/video analytics, the work reported in [29] describes a representative vision analytics application that we construct, test and characterize to quantify the limits of resilience under approximations and error incidence.

Under the second category of software support, we include compiler and operating systems (OS) level enhancements to facilitate the objective of achieving energy efficiency via voltage scaling, without compromising system-level resilience. The work by Kanev et al. [21] provides a systematic analysis of the effects of compiler optimization on voltage noise. The compiler as a system-level derating control knob is also described in the work by Gupta et al. [24]; in this latter case, the focus of attention is high-energy particle induced soft error incidence. In both scenarios (i.e. voltage noise and SER), we observe that increased levels of compiler optimization generally result in greater vulnerability to errors (i.e. reduced system reliability). Ideas about how to choose the right-level of code optimization that balances processor performance against reliability are touched upon in these works.

Thread scheduling heuristics, incorporated at the OS-level to improve power-performance efficiency is also an area of research in our project. In particular, the work by Vega et al. [30, 31] is focused on the use of thread consolidation and scheduling techniques that make near-optimal use of power management knobs like per-core dynamic voltage-frequency scaling (DVFS) and per-core power gating (PG).

VI. Cross-Layer Optimization

One of the fundamental tenets of our VLV (low power) resilient processor design paradigm is the idea of cross-layer optimization. Figure 2 depicts an exemplary cross-layer system stack that is the target of our integrated optimization methodology. The basic idea is that a layer-by-layer (localized) optimization to achieve resilience-constrained power-performance optimization results in over-design; which means that at the system level, one would get a product that is inefficient, and over-provisioned in terms of the reliability (or resilience) dimension.

In contrast, a holistic, cross-layer optimization model, one can achieve a software-hardware co-designed system, where only the appropriate levels of resilience are factored into each layer, taking into account the large degrees of masking (or "derating") that are available in each layer. The integrated resilience stack referred to in Figure 2, denotes a cross-layer resilience-optimized view of the system. In this view, the masking afforded by the software (application) is balanced against error detection coverage provisioned in the other software-hardware layers of the system stack.

The systematic methodology used in achieving this integrated resilience stack is provided by the CLEAR framework which is covered in the accompanying paper by Cheng et al. [14]. The cross-layer resilience-optimized processor is then offered for evaluation to the SHIVA integrated modeling framework. The SHIVA-CLEAR iterative flow may be repeated a few times to achieve an acceptable balance across reliability and power-performance efficiency.

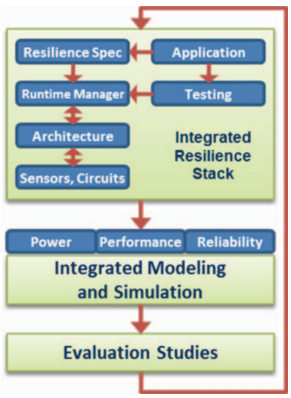

Figure 2. Cross-layer optimization system stack

VII. Conclusion

Very low voltage (VLV) design is a core capability within the field of energy-efficient embedded computing technologies. Even in high-end, server-class multi-core processors, VLV technologies are needed to enable wide

operating range (WOR) functionality in order to achieve targeted efficiencies. In this lead article, we provide a tutorial-style introduction to a set of papers [13-15] that describe the work we pursue in an IBM-led, DARPA-sponsored project within the PERFECT program [27]. Stanford University, Harvard University and University of Virginia are IBM's collaborating sub-contractors in this project. Pradip Bose (IBM) is the Principal Investigator (his contact email being pbose@us.ibm.com).

ACKNOWLEDGMENT

This work was performed with funding provided by the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. government.

REFERENCES

[1] R. G. Dreslinski, M. Wieckowski, D. Blauuw, D. Sylvester, T. Mudge, "Near-threshold computing: reclaiming Moore's Law through energy efficient integrated circuits," Proc. of the IEEE, vol. 98, no. 2, Feb. 2010.

[2] S. Jain et al., "A 280mV-to-1.2V wide-operating range IA-32 processor in 32nm CMOS," Int'l. Solid-State Circuits Conference (ISSCC), Feb. 2017.

[3] T. Webel et al., "Robust power management in the IBM z13," IBM Journ. Res & Develop., vol. 59, no. 4/5, paper 16, July/September 2015, pp. 16:1 – 16:12.

[4] P. Chuang et al., "Power supply noise in a 22nm z13 microprocessor," Int'l. Solid-State Circuits Conference (ISSCC), Feb. 2017.

[5] C. Lefurgy et al., "Active management of timing guardband to save energy in POWER7," Int'l. Symp. on Microarchitecture (MICRO), pp. 1-11, Dec. 2011.

[6] J. Leng, A. Buyuktosunoglu, R. Bertran, P. Bose, V. Reddi, "Safe limits on voltage reduction efficiency in GPUs: a direct measurement approach," Int'l. Symp. on Microarchitecture (MICRO), pp. 294-307, Dec. 2015.

[7] R. Joshi, M. Ziegler, H. Wetter, "A Low Voltage SRAM Using Resonant Supply Boosting," J. Solid-State Circuits 52(3): 634-644 (2017).

[8] R. Joshi, M. Ziegler, H. Wetter, C. Wandel, H. Ainspan, "14nm FinFET based supply voltage boosting techniques for extreme low Vmin operation," Symposium on VLSI Circuits, June 2015.

[9] L. Chang, D. Frank, R. Montoye, S. Koester, B. Ji, P. Coteus, R. Dennard, W. Haensch, "Practical strategies for power-efficient computing technologies," Proc. of the IEEE 98(2): 215-236 (2010).

[10] D. Heidel, K. Rodbell, E. Cannon, C. Cabral Jr., M. Gordon, P. Oldiges, H. Tang, "Alpha-particle-induced upsets in advanced CMOS circuits and technology, IBM Journ. of Res. & Develop. 52(3): 225-232 (2008).

[11] S. Mukherjee, Architecture Design for Soft Errors, Morgan-Kaufmann, March 2008.

[12] S. Eldridge, K. Swaminathan, N. Chandramoorthy, A. Buyuktosunoglu, A. Roelke, X. Guo, V. Verma, R. Joshi, M. Stan, P. Bose, "A low voltage RISC-V heterogeneous system: boosted SRAMs, machine learning, and fault injection in VELOUR," presented at the CARRV workshop at the 50th Ann. Int'l. Symp. on Microarchitecture (MICRO), October 2017.

[13] A. Roelke et al., "Pre-RTL voltage and power optimization for low-cost, thermally challenged multicore chips," Int'l. Conf. on Computer Design (ICCD), Oct. 2017.

[14] E. Cheng et al., "Cross-layer resilience in low-voltage digital systems: key insights," Int'l. Conf. on Computer Design (ICCD), Oct. 2017.

[15] S. Kodali, P. Hansen, N. Mulholland, P. Whatmough, D. Brooks, G-Y. Wei, "Applications of deep neural networks for ultra low-power IoT," Int'l. Conf. on Computer Design (ICCD), Oct. 2017.

[16] SIM-PPC POWER architecture simulator(s): https://developer.ibm.com/linuxonpower/sdk-packages/ .

[17] gem5 open-source simulators: http://www.gem5.org/ .

[18] Y. Shao, B. Reagen, G-Y. Wei. D. Brooks, "Aladdin: a pre-RTL, power-performance accelerator simulator enabling large design space exploration of customized architectures," Int'l. Symp. on Computer Arch. (ISCA), pp. 97-108, June 2014.

[19] H. Jacobson, A. Buyuktosunoglu, P. Bose, E. Acar, R. Eickemeyer, "Abstraction and microarchitecture scaling in early stage power modeling," Int'l. Symp. on High Performance Computer Architecture (HPCA), pp. 394-405, Feb. 2011.

[20] S. Xi, H. Jacobson, P. Bose, G-Y. Wei, D. Brooks, "Quantifying sources of error in McPAT and potential impacts on architectural studies," Int'l. Symp. on High Performance Computer Architecture (HPCA), pp. 577-589, Feb. 2015.

[21] S. Kanev, T. Jones, G-Y. Wei, D. Brooks, V. Reddi, "Measuring code optimization impact on voltage noise," Workshop on Silicon Errors in Logic – System Effects (SELSE), April 2013.

[22] B. Reagen, P. Whatmough, R. Adolf, S. Rama, H. Lee, S-K. Lee, J. Hernández-Lobato, "Minerva: enabling low-power, highly-accurate deep neural network accelerators," Int'l. Symp. on Computer Arch. (ISCA), pp. 267-278, June 2016.

[23] P. Whatmough, S-K. Lee, H. Lee, S. Rama, D. Brooks, G-Y. Wei, "A 28nm SoC with a 1.2GHz 568 nj/prediction sparse deep neural network engine with >0.1 timing error rate tolerance for IoT applications," Int'l. Solid-State Circuits Conference (ISSCC), Feb. 2017.

[24] M. Gupta, J. Rivers, L. Wang, P. Bose, "Cross-layer system resilience at affordable power," Int'l. Reliability Physics Symp. (IRPS), June 2014.

[25] R. Bertran, A. Buyuktosunoglu, M. Gupta, M. Gonzalez, P. Bose, "Systematic energy characterization of CMP/SMT processor systems via automated micro-benchmarks," Int'l. Symp. on Microarchitecture (MICRO), pp. 199-211, Dec. 2012.

[26] R. Bertran, A. Buyuktosunoglu, P. Bose, T. Slegel, G. Salem, S. Carey, R. Rizzolo, T. Strach, "Voltage noise in muti-core processors: empirical characterization and optimization opportunities," Int'l. Symp. on Microarchitecture (MICRO), pp. 368-380, Dec. 2014.

[27] Power Efficiency Revolution for Embedded Computing Technologies (PERFECT), Andreas Olofsson (program manager): https://www.darpa.mil/program/power-efficiency-revolution-for-embedded-computing-technologies.

[28] R. Adolf, S. Rama, B. Reagen, G-Y. Wei, D. Brooks, "Fathom: reference workloads for deep learning methods," Int'l. Symp. on Workload Characterization (IISWC), Sept. 2016.

[29] R. Venkatagiri, K. Swaminathan, C-C. Lin, L. Wang, A. Buyuktosunoglu, P. Bose, S. Adve, "Resilience characterization of a vision analytics application under varying degrees of approximation," Int'l. Symp. on Workload Characterization (IISWC), Sept. 2016.

[30] A. Vega, A. Buyuktosunoglu, P. Bose, "SMT-centric power-aware thread placement in chip multiprocessors," Int'l. Symp. on Parallel Architectures and Compilation Techniques (PACT), pp. 167-176, 2013.

[31] A. Vega, A. Buyuktosunoglu, H. Hanson, P. Bose, S. Ramani, "Crank it up or dial it down: coordinated multiprocessor frequency and folding control," Int'l. Symp. on Microarchitecture (MICRO), pp. 210-221, Dec. 2013.

[32] K. Swaminathan, N. Chandramoorthy, C. Cher, R. Bertran, A. Buyuktosunoglu, P. Bose, "BRAVO: balanced, reliability-aware voltage optimization,", Int'l Symp. on High-Performance Computer Architecture (HPCA), pp. 97-108, Feb. 2017.

[33] R. Joshi et al., "6.6 GHz Low Vmin, read and half select disturb-free 1.2 Mb SRAM," VLSI Circuits Symposium, pp. 250-251, 2007.
