Reusing Leakage Current for Improved Energy Efficiency of Multi-Voltage Systems

MD Shazzad Hossain and Ioannis Savidis
Drexel University, Philadelphia, USA

***Abstract*—A novel method for the delivery of power to sub-threshold (sub-V_t) circuits is proposed. The unused leakage current during idle-mode operation of super-threshold (super-V_t) circuits is used to supply sub-V_t circuits. Super-V_t and sub-V_t circuits are simulated in a 45 nm CMOS technology, where the super-V_t circuits operate at 1.2 V and generate a sub-V_t voltage of 380 mV. The proposed technique is compared with two conventional methods, one that uses separate power distribution networks for super-V_t and sub-V_t circuits (baseline) and the second that implements voltage stacking. The implementation of the proposed leakage reuse (LR) technique on the `s27` ISCAS89 benchmark circuit results in a reduction of the total, static, and peak power consumption to, respectively, 0.41×, 0.207×, and 0.7× that of the baseline technique. The leakage reuse technique also reduces the peak voltage noise on V_{SS} and the V_{SS} settling time to, respectively, 0.68× and 0.44× that of the baseline at a cost of a 1.24× increase in the FO4 delay. In addition, the LR technique implemented on the `s208` ISCAS89 benchmark circuit reduces the peak voltage noise on the virtual ground (*VGND*) and the *VGND* settling time to, respectively, 0.28× and 0.23× that of the voltage stacking technique.**

I. Introduction

The contribution of leakage current to the total on-chip power consumption of microprocessors and system on chips (SoCs) has been increasing due to a reduction in the gate length, the threshold voltage of the transistors, and the gate oxide thickness, as well as the increase in transistor density per unit area [1]. In addition, the long idle periods of most battery-operated mobile devices makes leakage current a dominant component of the total power consumption [2]. A technique to reuse leakage current, therefore, significantly reduces the total energy dissipation of an integrated circuit.

Current heterogeneous multi-core systems integrate high-performance super-V_t cores with energy-efficient cores operating at near- and sub-V_t voltages [3–10]. State-of-the-art microprocessors operating at super-V_t supply voltages consume a significant portion of the total power as leakage, which provides no computational benefit [11]. In contrast, the computation in sub-V_t circuits is performed by consuming only leakage current.

The leakage power consumption of a Xeon Tulsa processor operating at 3.4 GHz and with a 1.25 V supply is 44.1 W, which is 31% of the total power consumption of 150 W [12]. Similarly, the leakage power consumption of the UltraSPARC T1 Niagara processor is 16 W (26% of the total power of 63 W), when operating at 1.2 GHz and with a 1.2 V supply [13]. Prior work implementing sub-V_t processors has shown ultra low power operation and very low energy dissipation per instruction. A sub-V_t processor was implemented in a 130 nm technology for sensory network applications with a 130 mV supply voltage, consuming 11 nW of power [14]. A sub-V_t Phoenix processor was implemented in a 180 nm technology for sensing applications with a 500 mV supply voltage while consuming 35.4 pW and 226 nW in, respectively, standby and active mode [15]. In addition, a sub-V_t SoC was implemented in a 130 nm technology for wireless electrocardiogram (ECG) monitoring with a 280 mV supply voltage, consuming 2.6 μW of power [16]. A tiny fraction of the total leakage power dissipation from the processors operating at a nominal supply voltage is, therefore, sufficient to drive an entire sub-V_t processor.

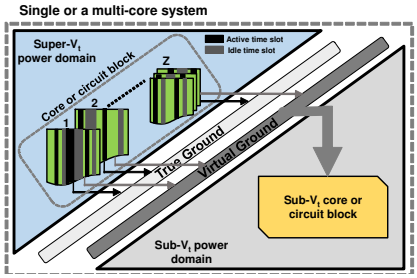


Fig. 1: System level model of reusing the leakage current of super-V_t circuits to drive sub-V_t circuits.

In this paper, the leakage current of the idle cores or circuit blocks operating at a nominal supply voltage is used to drive the circuits of a sub-V_t core. The proposed system level model that accounts for the reuse of the leakage current is shown in Fig. 1. Due to the advent of modern low power methods and enhanced power management techniques, there are enough idle circuit blocks and cores within any system to reuse unused leakage energy [17–19]. Therefore, for this paper, at least one super-V_t core (or circuit block) is assumed to operate in idle mode at any given time. The system level model is applicable to a) a single core with circuit blocks operating at both super-V_t and sub-V_t and b) a multi-core system with cores operating at both super-V_t and sub-V_t supply voltages.

The primary contributions of this paper are: 1) a novel method of delivering power to sub-V_t circuits using the leakage current of super-V_t circuits operating in standby mode, 2) the delivery of current to sub-V_t circuits without using separate voltage regulators, which are required for conventional sub-V_t circuits [20,21], and 3) no boosted supply voltage is required as needed for conventional voltage stacking techniques [22–24].

The rest of the paper is organized as follows. The circuit model of the proposed leakage reuse technique is provided in Section II. Simulation and analysis of the proposed technique are described in Section III. Some concluding remarks are provided in Section IV.

II. Circuit Model of the Leakage Reuse Technique

The circuit model developed to analyze the leakage reuse technique is shown in Fig. 2. Four functional blocks within a core are implemented, which are supplied by a conventional hierarchical power delivery system. The ground network is, however, modified to leverage the leakage current from the super-V_t CMOS circuits. The functional blocks *B* and *D* are assumed to be performing a high activity workload and are also highly sensitive to ground bounce. Blocks *B* and *D* are, therefore, directly connected to true ground. The functional blocks *A* and *C* are, however, assumed to be performing low

978-1-7281-0397-6/19/$31.00 ©2019 IEEE

activity tasks and are less sensitive to ground bounce noise. In this case, the leakage current from blocks A and C is used to power the sub-V_t core. The total width of all transistors within a functional block provides an estimate of the total leakage current contributed during standby mode [25]. A continuous power supply to the sub-V_t core exists, since at any given time at least one of the functional blocks is operating in idle mode. Both the A and C blocks are grounded through two switches: one PMOS and one NMOS transistor (A through S^A_{VG} and S^A_G, and C through S^C_{VG} and S^C_G). Transistor S^A_{VG} (S^C_{VG}) sinks current from block A (C) to the sub-V_t power network through a virtual ground during idle mode operation, while transistor S^A_G (S^C_G) sinks current from block A (C) to the true ground while in the active mode.

The circuit implementation of the switching mechanism is shown on the left side of Fig. 2. The NMOS transistor connects the virtual ground to the true ground when either circuit block A or C is in active mode. The PMOS transistor connects the virtual ground to the power network of the sub-V_t core when either A or C is idle. The gates of transistors S^A_{VG}, S^A_G, S^C_{VG}, and S^C_G are connected to control signals Φ_A and Φ_C as shown in Fig. 2. The value of Φ_A and Φ_C during idle and active mode is, respectively, 0 and 1. The transistors S^A_{VG} and S^C_{VG} must be large enough to quickly drain the leakage current through the virtual ground without causing significant delay and resistive drop, while the threshold voltages of transistors S^A_G and S^C_G must be large enough to prevent leakage loss during the idle mode of operation of the super-V_t cores.

There are three primary advantages of the proposed circuit technique: 1) a reduction in the total leakage power of the super-V_t cores, 2) a reduction in the total power consumption of the system by leveraging the leakage power of the super-V_t cores to supply current to circuits in the sub-V_t cores, and 3) no separate voltage source is required for the sub-V_t cores.

III. Simulation and Analysis

The proposed technique to reuse leakage power is evaluated through SPICE simulation in a 45 nm CMOS process. The super-V_t supply voltage is set to 1.2 V, while the sub-V_t supply voltage is set to 380 mV. The 45 nm CMOS process includes multi-threshold transistors, where low-threshold (low-V_t), nominal threshold (nominal-V_t), and high-threshold (high-V_t) devices are provided. Low-V_t transistors are used for the circuit blocks operating with a super-V_t supply voltage to improve performance, while nominal-V_t transistors are used for the circuit blocks operating with a sub-V_t supply voltage to reduce power consumption. In addition, high-V_t NMOS transistors are used for the sleep transistors S^A_G and S^C_G to reduce the power consumption during standby mode, while low-V_t PMOS transistors are used for the switches connecting to the sub-V_t core. Note that the threshold voltage for a nominal-V_t, low-V_t, and high-V_t NMOS transistor is, respectively, 410 mV, 322 mV, and 608 mV.

A. Modeling Methodology

The leakage reuse technique is evaluated using a chain of inverters (`COI`) and two ISCAS89 benchmarks circuit (`s27` and `s208`). The ISCAS89 benchmark circuits are used to represent super-V_t cores, while the `COI` is used to represent both super-V_t and sub-V_t cores. The super-V_t chain consists of six equally sized inverters, whereas the sub-V_t chain consists of six tapered inverters that are optimized for sub-V_t operation. The ground networks of both the `COI` and the ISCAS89 benchmark circuits are modified to enable a connection with either true ground or the sub-V_t power network through the virtual ground.

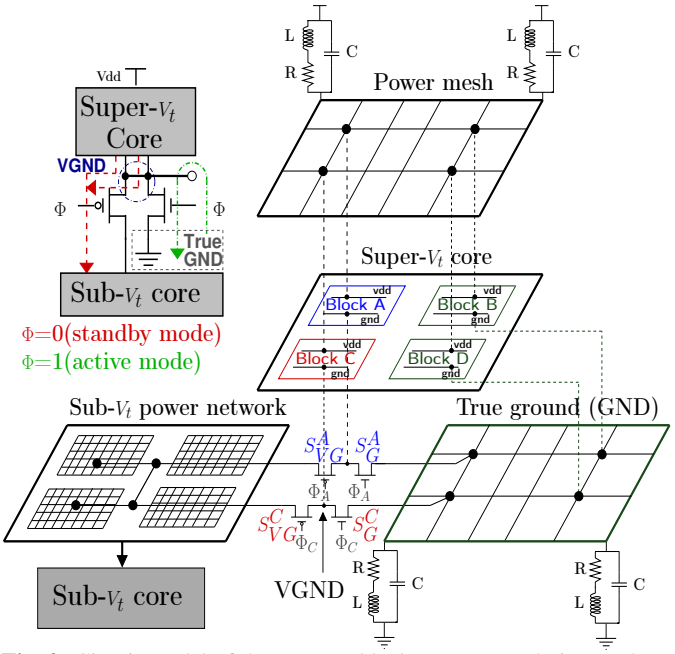


Fig. 2: Circuit model of the proposed leakage reuse technique, where the ground distribution network is modified to allow for voltage stacking during idle mode operation of the super-V_t core.

Three circuit topologies are considered for simulation:

Topology 1 (Baseline): Two isolated super-V_t circuit blocks representing two individual super-V_t cores and one isolated sub-V_t circuit block (chain of tapered inverters) representing an individual sub-V_t core. Note that there is no connection between the super-V_t and sub-V_t circuit blocks, with each block supplied current independently, which is the conventional practice.

Topology 2 - Proposed leakage reuse (L.R.): Two isolated super-V_t circuit blocks representing two individual super-V_t cores. The ground network of both super-V_t circuit blocks is connected to either true ground or the power network of a sub-V_t circuit block consisting of a chain of tapered inverters through the switching circuit shown in Fig. 2. The leakage current of the super-V_t circuit blocks supply the current required by the sub-V_t circuit block.

Topology 3 - Voltage stacking (V.S.): Two isolated super-V_t circuit blocks representing two individual super-V_t cores, where the ground network of both super-V_t circuit blocks is directly connected with the power network of a sub-V_t circuit block consisting of a chain of tapered inverters. The voltage stacking technique proposed in [26] to generate near-V_t supply voltages has been modified in this paper to generate sub-V_t supply voltages.

Similar input signals and capacitive loads are used for all three topologies, which ensures that the active and idle mode transitions are equally applied to the three configurations. The power supply and ground networks of all three circuit topologies are represented with equivalent electrical parameters obtained from the V_{cc} and V_{ss} pins of a DIP-40 package

TABLE I: The active area of the chain of inverters (COI), s27, and s208 circuits for the three topologies analyzed.

		Area of super-V_t circuit block (μm^2)	Switch area (μm^2)	Area of sub-V_t circuit block (μm^2)
COI	Baseline	2.331	N/A	7.56
	Leakage reuse	2.331	0.35 (15% of Super-V_t COI) S_{VG}:S_G of 1.5 µm: 2 µm	7.56
	Voltage stacking	2.331	N/A	7.56
s27	Baseline	6.73	N/A	21
	Leakage reuse	6.73	0.31 (4.6% of Super-V_t s27) S_{VG}:S_G of 3 µm: 0.1 µm	21
	Voltage stacking	6.73	N/A	21
s208	Baseline	21.9	N/A	21
	Leakage reuse	21.9	0.6 (2.7% of Super-V_t s208) S_{VG}:S_G of 5 µm: 1 µm	21
	Voltage stacking	21.9	N/A	21

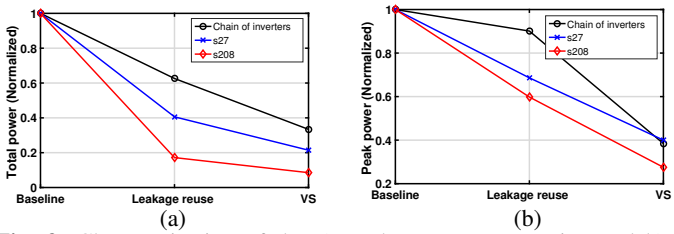
model to analyze the transient noise induced on the power and ground networks. The R, L, and C for the model are, respectively, 0.217 Ω, 8.18 nH, and 5.32 pF.

The three circuit topologies are implemented using a COI and the s27 and s208 ISCAS89 benchmark circuits, all operating at 5 MHz. The total active area for the super-V_t circuit blocks, sub-V_t circuit blocks, and switches is listed in Table I for a COI, s27, and s208. For each topology, a similar area is used, where the only exception is the additional area required for the switches when implementing the leakage reuse technique. Therefore, iso-frequency and iso-area analysis is performed on the three topologies.

The voltage on the virtual ground, which is the supply voltage of the sub-V_t block generated by the leakage current of the super-V_t cores, is dependent upon several circuit parameters. The parameters include 1) the total leakage current of the super-V_t circuit blocks, 2) the contribution of leakage current from each super-V_t circuit block, 3) the on and off current through the switches, which are partly dictated by the threshold voltage, gate voltage, and the size of the switches, and 4) the total current required by the sub-V_t circuit block.

B. Characterization of Power Consumption

The total and peak power consumption of all three circuit topologies is shown in, respectively, Figs. 3(a) and 3(b). The total power (and peak power) of the leakage reuse technique is reduced to 0.63× (0.9×), 0.41× (0.7×), and 0.17× (0.6×) that of the baseline for, respectively, a COI, s27, and s208. The results indicate an increasing benefit of implementing the leakage reuse technique as the size of a circuit increases. However, the leakage reuse technique consumes more power than the voltage stacking technique as the super-V_t and sub-V_t circuits are stacked for the entire duration of circuit operation when voltage stacking is implemented. The leakage reuse technique applied to s208 exhibits a total and peak power consumption of, respectively, 2.02× and 2.17× that of the voltage stacking technique.

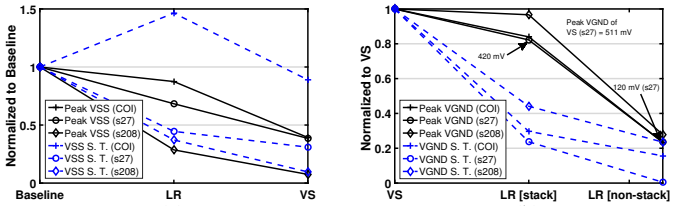


(a) (b)

Fig. 3: Characterization of the a) total power consumption and b) peak power consumption for COI, s27, and s208.

C. Characterization of Noise on True and Virtual Ground

The voltage drop on the super-V_t power network is within ±5% of the nominal voltage of 1.2 V for all circuit topologies.



(a) (b)

Fig. 4: Characterization of a) the peak V_{SS} bounce and the settling time of the peak V_{SS}, and b) the peak *VGND* bounce and the settling time of the *VGND* node, where switches for the LR technique are sized for 10% noise on the stacked and non-stacked VGND.

However, the voltage noise on the ground network exceeds ±5% of the ideal ground voltage of 0 V. The characterization of the peak ground voltage bounce on (V_{SS}) and the settling time of the V_{SS} node (the time required to settle within ±5% of the ideal ground voltage) are shown in Fig. 4 (a), where the results obtained for the leakage reuse and voltage stacking techniques are normalized to the results obtained for the baseline. The leakage reuse (LR) technique exhibits reduced peak noise on V_{SS} and a shortened settling time of V_{SS} as compared to the baseline for the chain of inverters (COI), s27, and s208 circuits. The implementation of the leakage reuse technique on s27 (s208) reduces the peak noise on V_{SS} and the V_{SS} settling time to, respectively, 0.68× (0.29×) and 0.44× (0.37×) that of the baseline.

Unlike the voltage stacking technique, the leakage reuse technique enables the stacking of idle super-V_t circuit blocks with the sub-V_t block, while the active super-V_t circuit blocks are connected to true ground. However, the noise transients propagate between the true ground of the active circuit blocks and the virtual ground of the idle circuit blocks. For the leakage reuse technique, the super-V_t active and idle circuit blocks are described as, respectively, non-stacked and stacked.

Similar to true ground, the *VGND* voltage bounce is analyzed by characterizing both the peak voltage noise at the *VGND* node and the settling time of the *VGND* node (the time required to settle within ±5% of the steady state *VGND*) for all three circuit topologies. The results of the characterization of the voltage noise are shown in Fig. 4(b) for both the leakage reuse (stacked and non-stacked) and voltage stacking techniques. The voltage of the *VGND* node at steady state is the same as the supply voltage of the sub-V_t circuit, which is 380 mV. The peak *VGND* bounce and *VGND* settling time of the active circuit blocks when implementing the leakage reuse technique (non-stacked) are less than both the voltage stacking technique and the idle circuit blocks implemented with the leakage reuse technique (stacked), while ensuring that the voltage bounce on *VGND* for both the active and idle circuit block does not exceed 10% of the rail-to-rail voltage of, respectively, 1.2 V (super-V_t V_{dd}) and 380 mV (sub-V_t V_{dd}). The peak voltage noise on *VGND* (*VGND* settling time) for the leakage reuse technique implemented on the s208 circuit is reduced to 0.97× (0.4×) and 0.28× (0.23×) in, respectively, the stacked and non-stacked mode, as compared and normalized to the voltage stacking technique. The noise on the *VGND* node (non-stacked) is characterized for 15%, 10%, and 5% allowed voltage bounce, which results in an increase in the power consumption due to the larger switches necessary to reduce the noise on *VGND*, as shown from results listed in Table II. There

TABLE II: Trade-off between the noise on *VGND* and the total power consumption when in the non-stack mode. Power consumption is normalized to the baseline (independent voltage sources for super-V_t and sub-V_t cores) for each circuit type.

Total power normalized to baseline	Noise in *VGND* [non-stack] as compared to ideal ground		
	5%	10%	15%
COI	0.642	0.635	0.616
s27	0.410	0.409	0.408
s208	0.174	0.168	0.167

is an average increase in the power consumption of 1.03× when the *VGND* bounce for the active circuit is constrained to 5% instead of 15% of the rail-to-rail voltage of V_{dd}.

D. Trade-off Between V. S. and Proposed L. R. Technique

The transient analysis of the leakage reuse and voltage stacking techniques is shown in Fig. 5, where the virtual ground bounce and the stability of the output voltages of the super-V_t and sub-V_t cores are compared. Similar clock and input switching patterns are applied to both implementations of the circuits.

The simulation is performed using two super-V_t cores each consisting of an ISCAS s27 benchmark circuit and one sub-V_t core consisting of a chain of six inverters. The clock, S_{VG}, S_G, *VGND*, and output signals of one core are shown in Fig. 5. In addition, clock-gating is applied to core 1 when inactive. In the case of voltage stacking, although core 1 is idle, the switching transients ($L \cdot di/dt$) of the super-V_t core 2 directly affects the *VGND* node, where charging and discharging of internal nodes causes significant voltage noise due to large internal capacitance. The noise induced on the *VGND* directly impacts the logical output of core 2, where the voltage for a logic 1 varies from 0.17 V to 0.41 V. The result is unwanted delay and logical failure for the circuits connected to the logical outputs of the voltage-stacked super-V_t cores. In addition, the noise induced on the *VGND* node of the voltage stacked super-V_t core generates a noisy V_{DD} for the sub-V_t circuits. Therefore, additional voltage regulators are required to produce a steady supply voltage for the bottom core in the stack when voltage stacking is implemented, which results in a significant overhead in energy consumption [23,24,26].

In contrast, for the leakage reuse technique, the *VGND* node of core 1 is not affected by the switching of core 2. A stable voltage is, therefore, found at the *VGND* node that is used to supply current to the sub-V_t core. The logical outputs of core 2 are also full-swing voltage signals, even as core 1 is stacked with sub-V_t circuits. In addition, a stable virtual *VGND* voltage provides a stable sub-V_t V_{dd} and, therefore, a stable sub-V_t output signal as shown in Fig. 5.

The voltage stacking technique reduces the leakage current for the entire duration of operation through the stacking effect, which increases the effective resistance in a given charging or discharging path. Therefore, the total power, peak power, and peak V_{ss} are reduced as compared to the baseline and leakage reuse technique at a cost of increased *VGND* bounce, as indicated by results shown in Figs. 3 and 4. For s27 the peak *VGND* noise of the voltage stacking technique is 2.07× greater than the leakage reuse (non-stacked) technique, which imposes severe limitations for circuits operating in super-V_t as the drive strength of the NMOS transistors is significantly reduced.

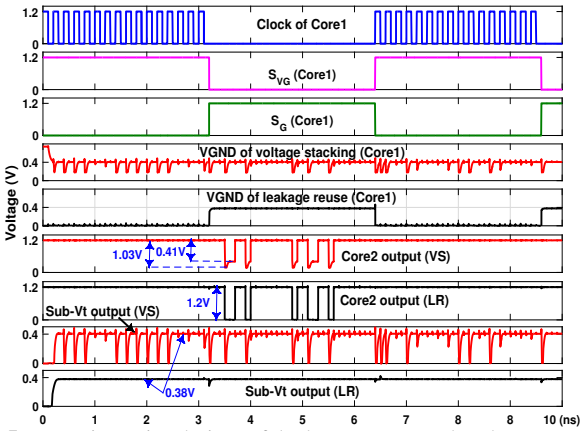


Fig. 5: Transient simulation of leakage reuse and voltage stacking techniques comparing the noise on the *VGND* node and the voltage of the output signals of the super-V_t and sub-V_t circuits.

Therefore, the propagation delay of the circuits connected to the output of the super-V_t circuit block increases. The fanout-of-four (FO4) delay is characterized at 1.2 V in a 45 nm CMOS technology with ideal ground and for three peak *VGND* values of 120 mV, 420 mV, and 511 mV (see Fig. 4 (b)). The implementation of the voltage stacking and the leakage reuse techniques results in an increase of the FO4 delay by, respectively, 4.17× (*VGND* of 511 mV) and 1.24× (*VGND* of 120 mV) that of the baseline technique (ideal ground).

In addition, the noise margins of the circuits connected to the output of s27 are significantly degraded as *VGND* increases. The noise margins of a CMOS inverter (PMOS width of 3.6 µm and NMOS width of 1.2 µm) are characterized at a 5 MHz operating frequency and with 5 fF of capacitive load. The noise margin low of the inverter is 483 mV, which is less than the 511 mV *VGND* when implementing the voltage stacking technique. Therefore, a logic low is not discernible with the voltage stacking technique if an inverter is connected to the output of s27. In contrast, implementing the leakage reuse technique on s27 provides a discernible logic low as *VGND* is 120 mV, while also reducing the total and peak power to, respectively, 0.41× and 0.7× that of the baseline. However, the delay increases by 1.24×. The peak voltage of 420 mV on *VGND* during the stacking mode of the leakage reuse technique does not effect the circuit delay and noise margin as the stacked super-V_t circuit blocks are in idle mode.

IV. Conclusions

A novel method of delivering power to sub-V_t circuits is proposed, where the unused leakage current from super-V_t circuits is used by sub-V_t circuits. An additional voltage source is, therefore, not required for the sub-V_t circuits, while the leakage current of the super-V_t circuits is also reduced. The proposed leakage reuse technique implemented on the s27 ISCAS89 benchmark circuit reduces the total, static, and peak power consumption to, respectively, 0.41×, 0.207×, and 0.7× that of the baseline. In addition, the peak voltage noise on V_{SS} and the V_{SS} settling time are reduced to, respectively, 0.68× and 0.44× that of the baseline at a cost of a 1.24× increase in the FO4 delay. The leakage reuse technique implemented on s208 exhibits a total (peak) power consumption of 2.02× (2.17×), while reducing the peak *VGND* (*VGND* settling time) to 0.28× (0.23×) that of the voltage stacking technique.

References

[1] H. Singh, K. Agarwal, D. Sylvester, and K. J. Nowka, "Enhanced Leakage Reduction Techniques using Intermediate Strength Power Gating," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 15, No. 11, pp. 1215–1224, November 2007.

[2] S. Kim, S. V. Kosonocky, D. R. Knebel, and K. Stawiasz, "Experimental Measurement of a Novel Power Gating Structure with Intermediate Power Saving Mode," *Proceedings of the International Symposium on Low Power Electronics and Design*, pp. 20–25, August 2004.

[3] M. S. Hossain and I. Savidis, "Robust Near-threshold Inverter with Improved Performance for Ultra-Low Power Applications," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 738–741, May 2016.

[4] M. S. Hossain and I. Savidis, "Bi-directional Input/Output Circuits with Integrated Level Shifters for Near-threshold Computing," *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems*, pp. 1240–1243, August 2017.

[5] E. Kultursay, K. Swaminathan, V. Saripalli, V. Narayanan, M. T. Kandemir, and S. Datta, "Performance Enhancement Under Power Constraints using Heterogeneous CMOS-TFET Multicores," *Proceedings of the IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, pp. 245–254, October 2012.

[6] Saber Moradi, Ning Qiao, Fabio Stefanini, and Giacomo Indiveri, "A Scalable Multicore Architecture with Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)," *IEEE Transactions on Biomedical Circuits and Systems*, Vol. 12, No. 1, pp. 106–122, August 2017.

[7] H. Mair, G. Gammie, A. Wang, S. Gururajarao, I. Lin, H. Chen, W. Kuo, A. Rajagopalan, W. Ge, and R. Lagerquist, "A Highly Integrated Smartphone SoC Featuring a 2.5 GHz Octa-Core CPU with Advanced High-Performance and Low-Power Techniques," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 1–3, March 2015.

[8] Z. Wang, Y. Liu, Y. Sun, Y. Li, D. Zhang, and H. Yang, "An Energy-Efficient Heterogeneous Dual-Core Processor for Internet of Things," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 2301–2304, July 2015.

[9] M. S. Hossain and I. Savidis, "Noise Constrained Optimum Selection of Supply Voltage for IoT Applications," *Proceedings of the IEEE International Symposium on Circuits and Systems*, pp. 1–5, May 2018.

[10] M. S. Hossain and I. Savidis, "Multi-Voltage Domain Power Distribution Network for Optimized Ultra-Low Voltage Clock Delivery," *Proceedings of the IEEE International Green and Sustainable Computing Conference*, pp. 244–251, October 2018.

[11] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pp. 469–480, December 2009.

[12] S. Rusu, S. Tam, H. Muljono, *et al.*, "A Dual-Core Multi-Threaded Xeon Processor with 16MB L3 Cache," *Proceedings of the IEEE International Solid State Circuits Conference*, Vol. 5, pp. 102–103, February 2006.

[13] A. S. Leon, K. W. Tam, J. L. Shin, D. Weisner, and F. Schumacher, "A Power-Efficient High-Throughput 32-Thread SPARC Processor," *IEEE Journal of Solid-State Circuits*, Vol. 42, No. 1, pp. 7–16, January 2007.

[14] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, *et al.*, "Performance and Variability Optimization Strategies in a Sub-200mV, 3.5 pJ/Inst, 11nW Subthreshold Processor," *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 152–153, June 2007.

[15] Y. Pu, J. P. Gyvez, H. Corporaal, and Y. Ha, "An Ultra-Low-Energy/Frame Multi-Standard JPEG Co-Processor in 65nm CMOS with Sub/Near-Threshold Power Supply," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 146–147, February 2009.

[16] S. C. Jocke, J. F. Bolus, S. N. Wooters, A. D. Jurik, A. C. Weaver, T. N. Blalock, and B. H. Calhoun, "A 2.6-μW Sub-Threshold Mixed-Signal ECG SoC," *Proceedings of the IEEE Symposium on VLSI Circuits*, pp. 60–61, June 2009.

[17] J. Leverich, M. Monchiero, V. Talwar, P. Ranganathan, and C. Kozyrakis, "Power Management of Datacenter Workloads using Per-Core Power Gating," *IEEE Computer Architecture Letters*, Vol. 8, No. 2, pp. 48–51, August 2009.

[18] M. Arora, S. Manne, I. Paul, N. Jayasena, and D. M. Tullsen, "Understanding Idle Behavior and Power Gating Mechanisms in the Context of Modern Benchmarks on CPU-GPU Integrated Systems," *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, pp. 366–377, March 2015.

[19] Q. Diao and J. Song, "Prediction of CPU Idle-Busy Activity Pattern," *Proceedings of the IEEE International Symposium on High Performance Computer Architecture*, pp. 27–36, October 2008.

[20] P. Meinerzhagen, C. Tokunaga, A. Malavasi, V. Vaidya, A. Mendon, D. Mathaikutty, J. Kulkarni, C. Augustine, M. Cho, S. Kim, *et al.*, "An Energy-Efficient Graphics Processor Featuring Fine-Grain DVFS with Integrated Voltage Regulators, Execution-Unit Turbo, and Retentive Sleep in 14nm Tri-Gate CMOS," *Proceedings of the IEEE International Solid-State Circuits Conference*, pp. 38–40, February 2018.

[21] T. Tong, S. K. Lee, X. Zhang, D. Brooks, and G. Wei, "A Fully Integrated Reconfigurable Switched-Capacitor DC-DC Converter with Four Stacked Output Channels for Voltage Stacking Applications," *IEEE Journal of Solid-State Circuits*, Vol. 51, No. 9, pp. 2142–2152, July 2016.

[22] R. T. Possignolo, E. Ebrahimi, E. K. Ardestani, A. Sankaranarayanan, J. L. Briz, and J. Renau, "GPU NTC Process Variation Compensation with Voltage Stacking," *IEEE Transactions on Very Large Scale Integration Systems*, Vol. 26, No. 9, pp. 1713–1726, September 2018.

[23] K. Blutman, A. Kapoor, J. G. Martinez, H. Fatemi, and J. P. Gyvez, "Lower Power by Voltage Stacking: A Fine-Grained System Design Approach," *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 78:1–78:5, June 2016.

[24] R. Zhang, K. Mazumdar, B. H. Meyer, K. Wang, K. Skadron, and M. R. Stan, "A Cross-Layer Design Exploration of Charge-Recycled Power-Delivery in Many-Layer 3D-IC," *Proceedings of the ACM/IEEE Design Automation Conference*, pp. 1–6, June 2015.

[25] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, Vol. 91, No. 2, pp. 305–327, February 2003.

[26] K. Mazumdar and M. R. Stan, "Charge Recycling On-Chip DC-DC Conversion for Near-Threshold Operation," *Proceedings of the IEEE Subthreshold Microelectronics Conference*, pp. 1–3, October 2012.