Microelectronics Journal 101 (2020) 104782

Contents lists available at ScienceDirect

Microelectronics Journal

journal homepage: www.elsevier.com/locate/mejo

# Recycling of unused leakage current for energy efficient multi-voltage systems

MD Shazzad Hossain *, Ioannis Savidis

*Drexel University, Philadelphia, PA, USA*

ARTICLE INFO

*Keywords:*
Leakage reuse system-on-chips (SoCs)
multi-voltage systems energy-efficient
computing voltage stacking sub-threshold
computing

ABSTRACT

A novel method for the delivery of power to sub-threshold (sub-$V_t$) circuits is proposed. The unused leakage current during the idle-mode operation of super-threshold (super-$V_t$) circuits is used to supply current to the sub-$V_t$ circuits. Super-$V_t$ and sub-$V_t$ circuits are characterized 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 energy break-even point of the leakage reuse technique is analyzed both analytically and through simulation. The proposed technique is compared with two conventional methods, one that implements separate power distribution networks for the super-$V_t$ and sub-$V_t$ circuits (baseline) and the second that applies voltage stacking. The proposed leakage reuse technique implemented on the `s27` ISCAS89 benchmark circuit reduces the average and peak power consumption to, respectively, 0.41× and 0.7× that of the baseline technique, while also reducing the peak voltage noise on the true ground node VSS and the settling time of the true ground voltage $V_{SS}$ to, respectively, 0.68× and 0.44× at a cost of a 1.24× increase in the FO4 delay. The leakage reuse technique implemented on the `s208` ISCAS89 benchmark circuit resulted in a reduction of the peak voltage noise on the virtual ground node VGND and a reduction in the settling time of the virtual ground voltage $V_{GND}$ to, respectively, 0.28× and 0.23× that of the voltage stacking technique. In addition, the regulation of the sub-$V_t$ supply voltage is evaluated for a variable workload executing on a 32-bit RISC-V core operating at 380 mV.

## 1. Introduction

The contribution of leakage current to the total on-chip power consumption of microprocessors and system on chips (SoCs) continues to increase due to reductions in the physical dimensions of the transistor, as well as an increase in the transistor density per unit area [1]. In addition, the long idle periods of most battery-operated mobile devices results in leakage current becoming a dominant component of the total power consumption [2], which is given by (1). The primary components of the leakage current of CMOS devices include 1) gate leakage, 2) sub-$V_t$ leakage, and 3) junction leakage, as given by (2) [3]. The gate leakage as a percent of the total leakage power has significantly increased as the gate length and oxide thickness are reduced, where a 30× increase in gate leakage has been reported when advancing each scaled fabrication node [1,4]. In addition, despite the reduced leakage current as compared to CMOS devices, FinFET devices also consume a significant amount of leakage power as a percentage of the total power [5]. In [5], the ISCAS85 benchmark circuits are implemented with shorted-gate and low-power mode FinFET logic cells, and on average, 13% of the total power consumption is attributed to leakage current. Therefore, as the transistor is scaled and the number of transistors in a circuit increases, a greater amount of leakage current is lost to the ground and substrate nodes of the circuit. The energy loss due to leakage current is more significant in state-of-the-art multi-core systems as cores idle for longer periods of time [6,7]. A technique to reuse (or recycle) leakage current of idle core(s) or circuit block(s), consequently, significantly reduces the total energy dissipation of an integrated circuit.

$$P_{total} = \alpha \cdot V^2 \cdot C_L \cdot f + V \cdot I_{Leakage} \tag{1}$$

$$I_{Leakage} = I_{gate} + I_{sub-V_t} + I_{junction} \tag{2}$$

As the performance requirements of executing applications varies, the use of both low-performance energy-efficient cores and high-performance cores within a multi-core system is needed to improve the overall energy-efficiency of the SoC. Current heterogeneous multi-core systems integrate high-performance super-$V_t$ cores with energy-efficient cores operating at near- and sub-$V_t$ voltages [8–12]. In addition, different voltage and frequency scaling techniques are implemented to

* Corresponding author.
*E-mail address:* msh89@drexel.edu (M.S. Hossain).
*URL:* http://ice.ece.drexel.edu/.

https://doi.org/10.1016/j.mejo.2020.104782
Received 11 December 2019; Received in revised form 9 March 2020; Accepted 6 April 2020
Available online 12 April 2020

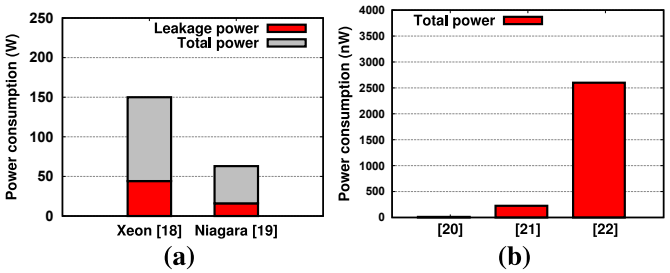


**Fig. 1.** Comparison of (a) total and leakage power consumption of super-$V_t$ cores, and (b) total power consumption of sub-$V_t$ cores.

improve the energy efficiency of the circuit including adaptive voltage scaling (AVS) [13,14], dynamic frequency scaling (DFS) [14], dynamic voltage scaling (DVS) [15,16], dynamic voltage and frequency scaling (DVFS) [12,15], and dynamic voltage and threshold scaling (DVTS) [17] to operate circuits from the super-$V_t$ to sub-$V_t$ region based on the performance requirements of the executing applications. Furthermore, microprocessors and processing elements that include a globally asynchronous and locally synchronous (GALS) clocking system achieve greater energy efficiency by operating sub-modules in multiple clock and voltage domains [18,19]. Note that the leakage current of state-of-the-art microprocessors operating at super-$V_t$ supply voltages provides no computational and storage benefit. In contrast, the computation and storage in sub-$V_t$ circuits is performed by consuming only leakage current.

The total and leakage power consumption of cores and SoCs implemented for both super-$V_t$ and sub-$V_t$ operation are shown in Fig. 1(a) and (b), respectively. The leakage power consumption of a Xeon Tulsa processor operating at 3.4 GHz and with a power supply voltage of 1.25 V is 44.1 W, which is equivalent to 31% of the total power consumption of 150 W, as shown in Fig. 1 [20]. 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 voltage [21]. Prior work implementing sub-$V_t$ processors indicate ultra low power operation and very low energy dissipation per instruction. A sub-$V_t$ processor implemented in a 130 nm technology for sensory network applications operates at 66 KHz and at a 160 mV supply voltage, resulting in a power consumption of 11 nW [22]. A sub-$V_t$ Phoenix processor implemented in a 180 nm technology for sensing applications operates at 106 KHz and at a 500 mV supply voltage, resulting in a power consumption of 35.4 pW and 226 nW in, respectively, idle and active mode [23]. In addition, a sub-$V_t$ SoC implemented in a 130 nm technology for wireless electrocardiogram (ECG) monitoring, operating at 475 KHz and at a 280 mV supply voltage, consumes 2.6 μW of power [24]. 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. Note that the terms nominal and super-$V_t$ are used interchangeably.

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. A methodology to reuse the leakage current from the nominal cores is developed [25]. A simple representation of a conventional power delivery network (PDN) with two independent voltage domains [22,26] and the proposed PDN implementing the leakage reuse technique is shown in Fig. 2, where the current from the super-$V_t$ core reused by the sub-$V_t$ core is regulated by the *Control circuit* block. In addition, the operating flow diagram of the proposed leakage reuse technique is shown in Fig. 3, where the super-$V_t$ cores are switched to either normal or leakage reuse mode depending on the workload activity of the super-$V_t$ cores. The developed technique assigns an active super-$V_t$ core to normal operating mode. An idle super-$V_t$ core is, however, switched to leakage reuse mode to supply current to an energy efficient sub-$V_t$ core. Given the growing interest in efficient power management techniques for multi-core systems, the proposed leakage reuse technique provides improved overall energy efficiency of multi-voltage domain SoCs consisting of both super-$V_t$ (high performance) and sub-$V_t$ (low performance) cores.

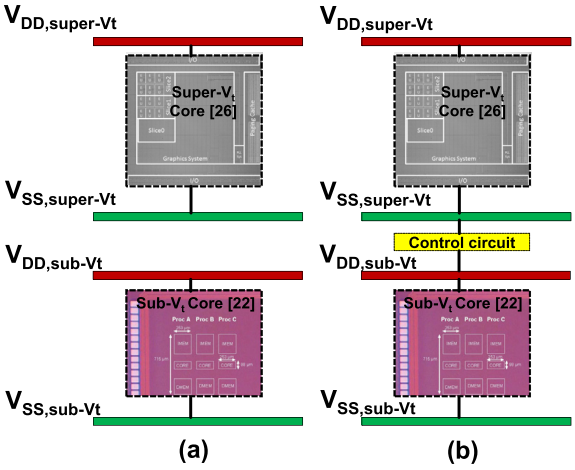


**Fig. 2.** Comparison of a) conventional PDNs with two independent voltage domains, and b) the proposed PDN implementing the leakage reuse technique.

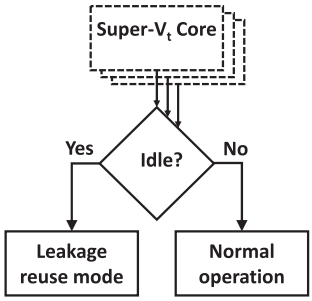


**Fig. 3.** Operating flow diagram of the proposed leakage reuse technique.

The proposed method is applicable to any heterogeneous multi-core system implementing dynamic voltage scaling as well as for circuits that integrate disparate technologies with multiple voltage domains including 3-D integrated systems, multi-core systems composed of hybrid technologies, neuromorphic systems, system-on-chips for deep neural networks, and processing elements implementing multiple clock and voltage domains through globally asynchronous and locally synchronous technique [18,28]. The primary contributions of this paper include:

1) A novel leakage reuse method that reduces the total power consumption of a multi-voltage system by utilizing the leakage current of idle super-$V_t$ cores for computation and storage in sub-$V_t$ cores. Separate voltage regulators and power delivery networks (PDNs) are, therefore, not required for the sub-$V_t$ circuits, unlike conventional power delivery to sub-$V_t$ circuits that require an independent and dedicated PDN and expensive voltage regulators [26,27], and
2) a reduction in the total leakage current of the super-$V_t$ circuits due to power network stacking during idle mode operation without significantly affecting the performance of the super-$V_t$ core(s) operating in active mode.

The rest of the paper is organized as follows. The feasibility of the leakage reuse technique is described in Section 2 through discussion of prior research on core idleness and voltage stacking. The system and circuit level model of the proposed leakage reuse technique is provided in Section 3. The energy overhead and the energy break-even point of the leakage reuse technique are discussed in Section 4. Simulation and

analysis of the proposed technique are described in Section 5. Some concluding remarks are provided in Section 6.

## 2. Feasibility of leakage reuse technique

Idle core(s) or circuit block(s) are continuously available as varying workloads executing on a multi-core system do not use all hardware resources at any given time. A discussion of the availability of idle cores is provided in Section 2.1. In addition, the proposed technique recycles the leakage current from idle super-$V_t$ core(s) through voltage stacking, which is described in Section 2.2.

### 2.1. Availability of idle Core(s) and circuit Block(s)

Most modern low power circuit and power management techniques reduce the amount of energy consumed during idle mode operation of the circuit [7,29]. Circuit blocks within a core or cores within a multi-core system are assigned to different operating modes (C-states) based on idle activity patterns to improve the performance per watt [7,30]. Power and clock gating are the most common techniques implemented to reduce the power consumption during idle mode operation of the circuit [31–34]. In [30], a machine learning based prediction method in conjunction with an OS power management policy is implemented in a state-of-the-art multi-core processor to assign cores to a C-state based on the idleness of a given core, where the C-4 state places the core in power gated mode. For the analysis of system idleness, a 3 GHz quad-core processor ran for 10K cycles. Within the 10K cycles, there was not a single instance when all cores were simultaneously active. In addition, the SPECWeb benchmark suite was executed on a dual-core processor, and similarly, there was no instance when both cores were concurrently active [30].

Per core power gating (PCPG) has been proposed as an effective power management option for multi-core processors along with dynamic voltage and frequency scaling [31]. In [31], core utilization traces are simulated to analyze the use of PCPG on a 2.5 GHz AMD Phenom X4 9850, which includes four cores implemented in a 65 nm technology. The utilization traces for a commercial application server (PHARMA04), two web servers (HCOM10, ECOM3), and a desktop computer (DESKTOP) are used to monitor the activity of the four cores, and again, there is no instance when all four cores are simultaneously active [31]. In addition, the idleness behavior of state-of-the-art processors is characterized using both consumer and CPU-GPU benchmarks including *DirectX9*, *KMeans*, and *Gaussian* from the PCMark and Rodinia suites [7]. The results from the simulation of the benchmark workloads demonstrate that a minimum of 110 multi-cycle idle events per second occurred for a broad range of applications executing on a 16 nm FinFET technology [7]. In this paper, to characterize the CPU utilization of individual cores, simulation of a system running Red Hat Linux and with 48 Intel Xeon CPU cores each operating at 3 GHz is used to execute different applications. At any given time, at least one core remains idle, while all 48 cores are idle for more than 90% of the runtime. Therefore, there are enough idle circuit blocks and cores within any system to allow for the reuse of unused leakage current [7,30,31].

### 2.2. Voltage stacking during idle mode

Voltage stacking is a technique where the power distribution networks of two circuit blocks within a core or two cores within a multi-core system are vertically stacked, sharing a common path from the supply ($V_{DD}$) to ground ($V_{SS}$). The conventional method of voltage stacking is shown in Fig. 4(a), where the circuit block set to a lower absolute voltage is vertically stacked with a circuit block set to a higher absolute voltage regardless of the executing workloads of the higher voltage circuit block. In other words, the two circuit blocks are vertically stacked for the entire operation of the device [35–38].

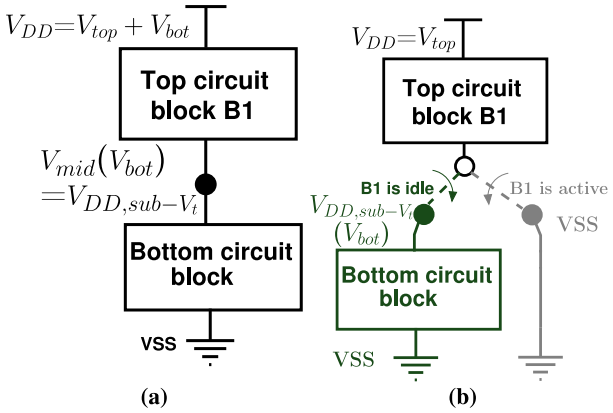


**Fig. 4.** Delivering current to a sub-threshold circuit block through a) voltage stacking, where two circuit blocks within a core are vertically stacked regardless of the circuit activity of the super-$V_t$ (top) circuit block, and b) the proposed technique of delivering current to sub-$V_t$ (bottom) circuit blocks, where *only* an idle super-$V_t$ circuit block is stacked.

Voltage stacking has been recently used for logic and memory circuits in 2-D and 3-D integrated circuits to 1) minimize the total power consumption and the peak rush current [37–42], which reduces inductive noise ($L{\cdot}di/dt$), and 2) limit interconnect wear-out due to electromigration (EM) as the current density is reduced [43]. Prior work has shown a 60% reduction in transient noise when voltage stacking is applied to a 3-D IC as compared to a non-stacked 2-D IC [42]. In addition, the stacking of SRAM banks during idle mode, as proposed in [39], reduces the leakage power by 93%. However, there are limitations of implementing the voltage stacking technique as circuit blocks are stacked during both active and idle mode operation. Limitations include 1) the need for regulation of the mid-node voltage ($V_{mid}$ in Fig. 4) due to variations in load current [42], 2) variation in voltage due to workload and current imbalances [42], 3) the need for a boosted supply voltage, which increases the power overhead [38,42], and 4) a reduction in the energy efficiency of voltage stacked circuit blocks as the imbalance in current load between the stacked voltage domains increases [40].

In this research, however, voltage stacking is applied to only idle super-$V_t$ cores or circuit blocks. The proposed method increases the effective resistance of the stacked path only during idle mode operation of the super-$V_t$ core and, therefore, reduces the total leakage current through the stacked path. Due to the stacking of only idle super-$V_t$ cores, the proposed approach is not adversely effected like conventional voltage stacking techniques, where careful regulation is required to compensate for variations in the workload and current imbalances between stacked voltage domains. During the idle mode operation of the super-$V_t$ circuit, the leakage current from the super-$V_t$ circuit block is delivered to the sub-$V_t$ circuit block for computation and storage as shown in Fig. 4 (b). Therefore, the super-$V_t$ circuit block remains unaffected during active mode operation while the leakage current from an unused super-$V_t$ circuit block is recycled to deliver power to the sub-$V_t$ circuit block.

## 3. Reusing leakage current of Super-$V_t$ Cores to drive Sub-$V_t$ Cores

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

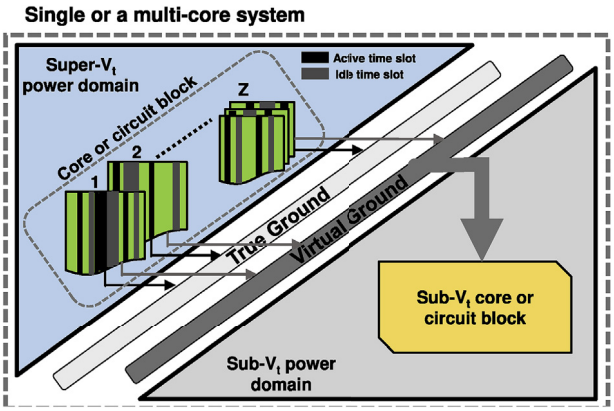


**Fig. 5.** System level model of reusing the leakage current of super-threshold circuits to drive sub-threshold circuits.

3.1 and 3.2, respectively. In addition, the control circuit of a super-$V_t$ core consisting of two switches is described in Section 3.3.

### 3.1. System level model of leakage reuse technique

The proposed system level model that accounts for the reuse of the leakage current from idle super-$V_t$ cores or circuit blocks to supply current to circuits operating in a sub-$V_t$ voltage domain is shown in Fig. 5. Due to significant idle states in existing state-of-the-art circuits, as discussed in Section 2.1, at any given time, at least one super-$V_t$ core (or circuit block) is assumed to operate in idle mode.

The system level model is applicable to a) a single core that includes both a super-$V_t$ and sub-$V_t$ voltage domain such as a system that implements globally asynchronous and locally synchronous clocking and b) a multi-core system with cores operating at both super-$V_t$ and sub-$V_t$ voltages. Therefore, the proposed leakage reuse technique is categorized as either 1) inter-core, where two cores that operate at a super-$V_t$ supply voltage drive a core operating at a sub-$V_t$ supply voltage, or 2) intra-core, where two circuit blocks within a core operate at a super-$V_t$ supply voltage and drive a circuit block operating at a sub-$V_t$ voltage.

### 3.2. Circuit model accounting for leakage reuse

Conventionally, the circuits within a core operating in a single voltage domain receive current through a hierarchical power delivery system. In general, a 10% activity factor implies that 10% of all gates switch at any given time. Therefore, both dynamic and leakage current is consumed for 10% of the gates of a core, while 90% of the gates consume only leakage current.

The circuit model used to analyze the leakage reuse technique is shown in Fig. 6. The super-$V_t$ core consists of four functional blocks that are supplied by a conventional hierarchical power delivery system. The ground network is, however, modified to allow for the reuse of the leakage current from the super-$V_t$ circuits. The functional blocks $B$ and $D$ are assumed to be executing a high activity workload and are highly sensitive to ground bounce and are, therefore, directly connected to true ground. The functional blocks $A$ and $C$ are, however, assumed to be performing low activity tasks and are less sensitive to ground bounce noise. In this case, the leakage current from blocks $A$ and $C$ is used to supply the sub-$V_t$ core. The cumulative total width of all transistors within a functional block provides an estimate of the total leakage current of the circuit during idle mode [3]. A continuous power supply to the sub-$V_t$ core exists, since at any given time at least one of the functional blocks (either $A$ or $C$) is operating in idle mode. Both super-$V_t$ blocks $A$ and $C$ are connected to ground through either of two switches, implemented as one PMOS and one NMOS transistor, where $A$ is connected through $S^A_{VG}$ and $S^A_G$, and $C$ through $S^C_{VG}$ and $S^C_G$ as shown in Fig. 6. Transistor $S^A_{VG}$ ($S^C_{VG}$) sinks current from block $A$ ($C$) to the sub-$V_t$ power network through a virtual ground (VGND) during idle mode operation of the blocks, while transistor $S^A_G$ ($S^C_G$) sinks current from block $A$ ($C$) to true ground during active mode.

### 3.3. Control circuit block applying the leakage reuse technique

The control circuit block that implements the proposed leakage reuse technique behaves similar to sleep transistors used for power gating as the two footer MOS transistors switch the operating mode of a super-$V_t$ core between normal activity and leakage reuse [1,31,43–45]. Note that during idle mode operation of the circuit, the ground node of the super-$V_t$ cores becomes a virtual ground.

The circuit implementation of the MOS switches is shown in Fig. 7. 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 $S^A_{VG}$ and $S^A_G$ in Fig. 6 are connected to the control signal $\Phi_A$, while $S^C_{VG}$ and $S^C_G$ are connected to $\Phi_C$. $\Phi_A$ and $\Phi_C$ are each set to 0 or 1 when the corresponding super-$V_t$ circuit is operating in either idle mode or active mode, respectively. The transistors $S^A_{VG}$ and $S^C_{VG}$ must be large enough to sufficiently conduct the leakage current through the virtual ground (sub-$V_t$ power network) within a given clock period and without causing significant 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 operation of the super-$V_t$ cores.

The proposed leakage reuse technique does not replace power gating. Simultaneous implementation of both power gating and leakage reuse is, therefore, possible. A multi-core system that implements both power gating and the leakage reuse technique is more energy efficient than a system that only applies power gating as current is delivered to an energy efficient sub-$V_t$ core without requiring a dedicated PDN with expensive voltage regulators.

## 4. Energy overhead and challenges of implementing leakage reuse

The PMOS and NMOS transistors shown in Figs. 6 and 7 and the circuitry that provides the control signal $\Phi$ consume additional energy. An analysis of the energy break-even point (BEP) is, therefore, performed, which is defined as the number of clock cycles at which the total energy savings obtained from the implementation of the leakage reuse technique equals the total energy consumed by the PMOS/NMOS switches and the control circuit [46]. The leakage current model changes significantly from one technology to another [47,48]. Instead of performing a technology dependent analysis, heuristic based analytical expressions are derived using standard circuit and device parameters, which are listed in Table 1, that are applicable to any CMOS technology. In order to reduce the complexity of the analysis, a few assumptions are made: 1) The total energy required to operate a sub-$V_t$ core is supplied by one super-$V_t$ core and, therefore, analytical expressions are derived for *Core*1 and $S_{Core}1$ only; a similar analysis can be performed for more complex systems, and 2) separate ground networks are implemented for the active super-$V_t$ core and sub-$V_t$ core to prevent noise coupling.

A schematic representation of an implementation of the leakage reuse technique is shown in Fig. 8, where $i$ number of super-$V_t$ cores supply unused leakage current to $m$ number of sub-threshold cores. The two footer transistors $S_{VG}$ (PMOS) and $S_G$ (NMOS) are utilized as control switches for each super-$V_t$ core, where additional switches are added for a more distributed supply of current to the sub-$V_t$ core. The control signal $\Phi$ is set to *logic low* when an onset of a long idle period of *Core*1 is detected, which results in $S_{VG}$ turning on and $S_G$ turning off as shown in Fig. 9. The control signal $\Phi$ applied to the gate of $S_G$ transitions to a *logic high* when *Core*1 switches to active mode.

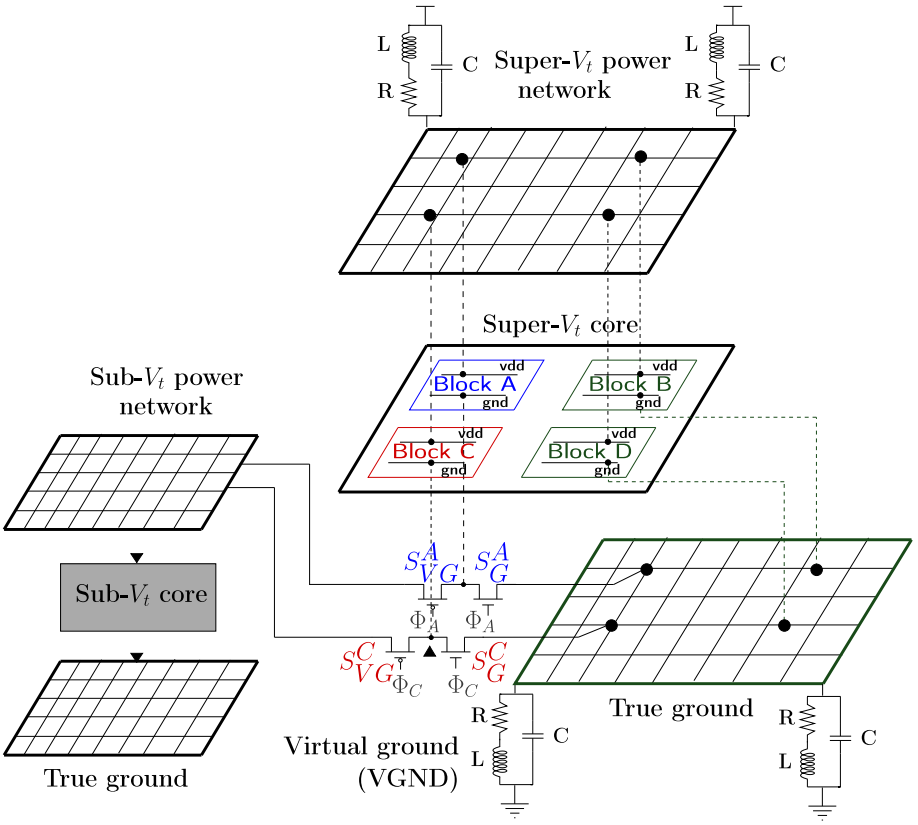


**Fig. 6.** Circuit model of the proposed leakage reuse technique, where the ground distribution network is modified to allow for voltage stacking during idle mode operation of the super-$V_t$ core(s). Blocks A and C supply leakage current to one sub-$V_t$ core.

**Table 1**
Circuit parameters used in the analysis of the energy break-even point of the leakage reuse technique.

| *Parameter* | *Definition* |
|---|---|
| $V_{DD,super-V_t}$ | Supply voltage of super-$V_t$ circuit |
| $V_{DD,sub-V_t}$ | Supply voltage of sub-$V_t$ circuit |
| $V_{GND}$ | Virtual ground voltage |
| $S_G$ | NMOS transistor connected to real ground |
| $S_{VG}$ | PMOS transistor connected to sub-$V_t$ core |
| $\alpha_{super-V_t}$ | Switching activity of super-$V_t$ core |
| $\alpha_{sub-V_t}$ | Switching activity of sub-$V_t$ core |
| $C_{S,super-V_t}$ | Total switching capacitance of super-$V_t$ core including drain, source, and gate capacitance as well as interconnects |
| $C_{int,super-V_t}$ | Total source capacitance of all transistors connected to virtual ground |
| $C_{S,sub-V_t}$ | Total switching capacitance of sub-$V_t$ core including drain, source, and gate capacitance of all transistors |
| $C_{VGND}$ | Total capacitance of virtual ground due to $S_G$ and $S_{VG}$ |
| $W_r$ | Ratio of switch size (sum of $S_G$ and $S_{VG}$) to the size of super-$V_t$ core |

The onset of the switching transitions of $S_{VG}$ and $S_G$ occur at times $T0$ and $T2$ as indicated by the transient waveforms shown in Fig. 9, while $T1$ and $T3$ represent the beginning of a period of steady-state voltage for the sub-$V_t$ core. Assume that the idle activity of *Core*1 begins at time $t = T0$ and $\Phi$, which is shown in Fig. 8, is set to logic 0. As transistor $S_G$ begins to turn off, *Core*1 begins to supply unused leakage current to the sub-$V_t$ core. In addition, at $t = T0$, $C_{V_{GND}}$ and $C_{int,super-V_t}$ are charged as *Core 1* transitions to an idle state. Note that the virtual ground node includes capacitances $C_{V_{GND}}$, $C_{int,super-V_t}$, and $C_{S,super-V_t}$, where half of the internal nodes of the super-$V_t$ core contribute to the capacitance of the virtual ground node as described in [48]. At $t = T1$, the capacitance of the virtual ground node $C_{V_{GND}}$ is fully charged to the target $V_{GND}$ voltage, and the leakage current through $S_{VG}$ saturates. The voltage drop across the $S_{VG}$ transistor is negligible, and $V_{GND}$ is approximately equal to the sub-$V_t$ supply voltage $V_{DD,sub-Vt}$ of 380 mV.

As the potential of the VGND node reaches a steady-state of 380 mV, the overall leakage current of *Core*1 is significantly reduced due to 1) an increase in the threshold voltage attributed to the reverse body-bias effect and 2) an increase in the effective resistance when stacked with the sub-$V_t$ core. However, there is an additional energy loss $E_{loss1}$ during the time interval $t = T1 - T0$ due to the switching of both the $S_{VG}$ and $S_G$ transistors, where $E_{loss1}$ per cycle is given by

$$E_{loss1} = \frac{1}{2}\alpha_{super-V_t} V_{DD}^2 \left(C_G + C_{VG} + C_{int,super-V_t} + \frac{1}{2}C_{S,super-V_t}\right). \quad (3)$$

Similarly, the energy overhead per cycle during the time interval $T3 - T2$ is given by

$$E_{loss2} = \frac{1}{2}\alpha_{super-V_t} V_{DD}^2 \left(C_G + C_{VG} + C_{int,super-V_t} + \frac{1}{2}C_{S,super-V_t}\right), \quad (4)$$

where at time $t = T2$, *Core*1 switches into an active mode and $S_G$ and

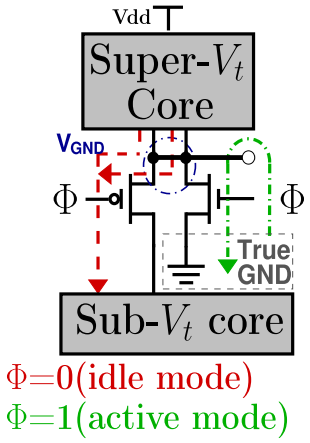


**Fig. 7.** MOS switches to connect the super-$V_t$ core to true ground when active or to the sub-$V_t$ core when idle.

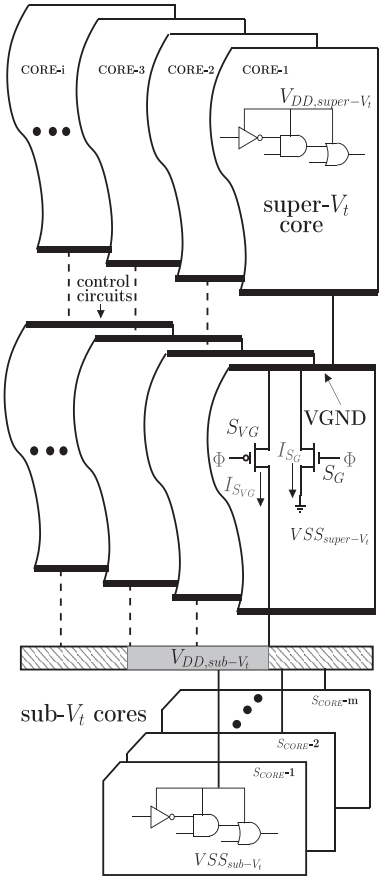


**Fig. 8.** Schematic representation of the proposed technique to reuse the leakage current from $i$ number of super-$V_t$ cores supplying current to one sub-$V_t$ core.

$S_{VG}$ are turned on and off, respectively, which results in the discharge of the VGND node through $S_G$. Between $t = T2$ and $t = T1$, the leakage energy of *Core*1 is used for computation by the sub-$V_t$ core. At $t = T3$, the VGND node of *Core*1 discharges to true ground as *Core*1 begins to operate in active mode. The total energy overhead per cycle between time interval $T0$ and $T3$ due to the implementation of the leakage reuse technique is given by

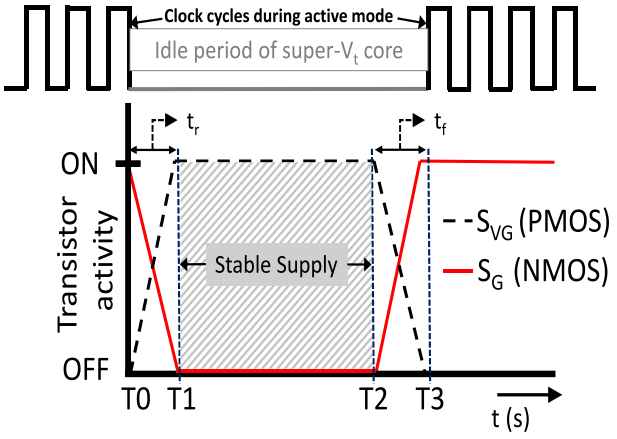


**Fig. 9.** Switching transitions of $S_{VG}$ and $S_G$ corresponding to the time intervals when leakage reuse is active ($\Phi = 0$).

$$E_{loss,total} = E_{loss1} + E_{loss2}$$
$$= \alpha_{super-V_t} V_{DD}^2 \left( C_G + C_{VG} + C_{int,super-V_t} + \frac{1}{2} C_{S,super-V_t} \right). \quad (5)$$

The reduction in the energy consumption due to the utilization of the leakage reuse technique over $N$ number of cycles is given by (6). The $E_L^{T1}$ term describes the total energy savings between time interval $T0$ and $T1$ and $E_{sub-V_t}^{cyc}$ defines the total energy dissipation per cycle of the sub-$V_t$ core. In addition, the average leakage energy dissipation per cycle when *Core*1 is in active mode and idle mode is given by $E_{L,active}^{cyc}$ and $E_{L,idle}^{cyc}$, respectively.

$$E_{saved}^{total} = E_L^{T1} + \sum_{i=0}^{N} \left( E_{sub-V_t}^{cyc} + E_{L,active}^{cyc} - E_{L,idle}^{cyc} \right) \quad (6)$$

The total energy dissipation of the sub-$V_t$ core per cycle, as given by (7), is calculated as the sum of the static and dynamic energy dissipated per cycle, where $I_{S_{VG}}$ is the current through transistor $S_{VG}$ and $T$ is the clock period. The energy consumption of the circuit is reduced due to the implementation of the footer transistors $S_{VG}$ and $S_G$, which behave similar to a switch in power-gated circuits. The energy savings between the onset of an idle event ($T0$) and the time when $V_{GND}$ begins to maintain a steady state ($T1$) is given by (8) and (9) [46,48,49]. The number of completed clock cycles during the time interval $T0$ to $T1$ is given as $N_x$ and is empirically formulated as (10), where $f$ is the operating frequency of the super-$V_t$ core, $W_r$ is the ratio of the size of the switch to the size of the super-$V_t$ core, and $a$ is the wake up latency, which is 5.27 ns for a 1 GHz clock frequency [50].

$$E_{sub-V_t}^{cyc} = \alpha_{sub-V_t} C_{sub-V_t} V_{DD,sub-V_t}^2 + I_{S_{VG}} V_{DD,sub-V_t} T \quad (7)$$

$$E_L^{T1} = \frac{F_{DIBL}}{nV_T} N_x T E_L^{cyc} \quad (8)$$

$$= \frac{F_{DIBL}}{nV_T} \frac{\alpha_{super-V_t} \delta V_{DD}}{4\left(\frac{1}{2} + \frac{C_{V_{GND}}}{C_{S,super-V_t}}\right)} N_x T \quad (9)$$

$$N_x = \frac{fa}{W_r} \quad (10)$$

The average leakage energy for time interval $T0$ to $T1$, given by $E_L^{T1}$, is dependent on the drain-induced barrier lowering factor $F_{DIBL}$ ($\approx 0.1$), the sub-threshold slope factor $n$ ($\approx 1.3$), and the thermal voltage $V_T$ ($= kT/q \approx 25$mV). The leakage factor $\delta$, described in [48], is given as

the ratio of the average leakage energy dissipation to the switching energy dissipation per cycle ($\delta = E_L^{cyc}/E_S^{cyc}$).

The switching energy dissipation per cycle of the super-$V_t$ core is given by

$$E_S^{cyc} = \frac{1}{2}\alpha_{super-V_t} C_{S,super-V_t} V_{DD}^2, \tag{11}$$

where $C_{S,super-V_t}$ represents the total switching capacitance of the core that includes the drain, source, and gate capacitances of the transistors as well as the capacitance due to the interconnects. The average leakage energy dissipated per cycle when *Core*1 is in active mode (not stacked) is given by

$$\begin{aligned} E_{L,active}^{cyc} &= TI_L V_{DD} \\ &= \delta E_S^{cyc} \\ &= \frac{1}{2}\alpha_{super-V_t}\delta C_{S,super-V_t} V_{DD}^2, \end{aligned} \tag{12}$$

where $I_L$ is the average leakage current through all leakage paths in *Core*1 [48]. The leakage current through the PMOS transistor $S_{VG}$ is negligible since the *drain* of $S_{VG}$ is connected to $V_{DD,sub-V_t}$. Therefore, all of $I_L$ is assumed to pass through $S_G$ during the active mode operation of *Core*1.

The average leakage energy dissipated per cycle when *Core*1 is in idle mode (stacked) is given by

$$E_{L,idle}^{cyc} = TI_{S_G} V_{GND}, \tag{13}$$

where $I_{S_G}$ is the current through $S_G$ when operating in the idle mode. If an average leakage energy dissipation per cycle is assumed, the leakage energy dissipated over $N$ cycles is given by

$$\sum_{i=0}^{N} E_{L,idle}^{cyc} = NTI_{S_G} V_{GND}. \tag{14}$$

At the energy break-even point, the reduction in the total energy consumption equals the energy loss due to implementing the leakage reuse technique. Therefore, (5) and (6) are modified to, respectively,

$$E_{saved}^{total} = E_{loss,total}, \text{and} \tag{15}$$

$$\begin{aligned} &E_L^{T1} + \sum_{i=0}^{N}\left(E_{sub-V_t}^{cyc} + E_{L,active}^{cyc} - E_{L,idle}^{cyc}\right) \\ &= \sum_{i=0}^{N} \alpha_{super-V_t} V_{DD}^2 \left(C_G + C_{VG} + C_{int,super-V_t} + \frac{1}{2}C_{S,super-V_t}\right). \end{aligned} \tag{16}$$

The energy consumption per cycle remains relatively constant except for a few clock cycles after a power down or power up event. For the provided analysis, a constant energy consumption is assumed for each cycle, where (7) is simplified by assuming $E_{sub-V_t}^0 \approx E_{sub-V_t}^1 \approx E_{sub-V_t}^2 \approx \ldots \approx E_{sub-V_t}^N$ for $N$ cycles. A similar assumption is made for $E_{L,active}^{cyc}$ and $E_{L,idle}^{cyc}$. Substituting for $E_L^{T1}$, $E_{L,active}^{cyc}$, $E_{L,idle}^{cyc}$, and $E_{sub-V_t}^{cyc}$ in (16) results in

$$\begin{aligned} &\left(E_L^{T1} + NE_{sub-V_t}^{cyc} + NE_{L,active}^{cyc} - NE_{L,idle}^{cyc}\right) = \left(C_G + \right. \\ &\left. C_{VG} + C_{int,super-V_t} + \frac{1}{2}C_{S,super-V_t}\right) N\alpha_{super-V_t} V_{DD}^2, \end{aligned} \tag{17}$$

where

$$\begin{aligned} N_{break-even} &= |N| \\ &= \frac{E_L^{T1}}{\left(\left(C_G + C_{VG} + C_{int,super-V_t} + \frac{1}{2}C_{S,super-V_t}\right)\alpha_{super-V_t} V_{DD}^2 - E_{sub-V_t}^{cyc} + E_{L,active}^{cyc} - E_{L,idle}^{cyc}\right)}. \end{aligned} \tag{18}$$

Advanced power management and task scheduling algorithms are applied to identify the idle intervals in a multi-core system [6]. The reuse of the leakage current of the super-$V_t$ core is activated through the control signal $\Phi$ only when the number of idle cycles in the super-$V_t$ core exceeds $N_{break-even}$ and the voltage on the VGND node meets the target sub-$V_t$ supply voltage. Therefore, the activation of the control signal $\Phi$ is formulated as

$$\forall_i, \Phi_{core,i} = \begin{cases} 0 \text{ (LR on)}, & \text{when} N > N_{break-even} \ \& \\ & |V_{k,i} - V_{DD,sub-V_t}| \le 0.1 V_{DD,sub-V_t}^{target}, \\ 1 \text{ (LR off)}, & \text{otherwise} \end{cases} \tag{19}$$

where $i$ is the super-$V_t$ core identifier for $n$ cores $= \{1, 2, 3, \ldots n\}$. The voltage at the VGND node after $k$ number of cycles is given as $V_{k,i}$. The target supply voltage of the sub-$V_t$ core is represented by $V_{DD,sub-V_t}^{target}$. The minimum number of cycles $N_{\min}$ that the super-$V_t$ core must be idle is given by (20), which ensures an overall gain in the energy efficiency of the system and a supply voltage for the sub-threshold core that is within 10% of $V_{DD,sub-V_t}^{target}$. The number of cycles required to settle within 10% of $V_{DD,sub-V_t}^{target}$ is given by $N_{DD,sub-V_t}^{target}$.

$$N_{min} = \max\left\{N_{break-even}, N_{DD,sub-V_t}^{target}\right\} \tag{20}$$

Simulation is performed to determine the total energy consumption of a multi-core system where the required energy to operate the sub-$V_t$ core is supplied by one super-$V_t$ core. The results of the characterization of the number of cycles obtained from simulation are compared with the number of cycles determined through the analytical expressions. The super-$V_t$*Core*1 and the sub-$V_t$ core are each implemented with a chain of six inverters. To analytically determine the $N_{break-even}$, the technology parameters listed in Table 1 are applied to (18) for both the super-$V_t$ and the sub-$V_t$ cores, where $C_{V_{GND}}/C_{S,super-V_t} = 0.1$, $\delta = 0.1$, $V_{GND} \approx V_{DD,sub-V_t} = 0.38$ V, $T = 1$ ns, $\alpha_{super-V_t} = 1$, $\alpha_{sub-V_t} = 0.05$, $I_{S_G} = I_{S_{VG}} = 0.15 I_{total}$, $I_{total} = 10$ μA, $W_r = 0.125$, and $C_{S,super-V_t} = 70$ fF. The total area of both the super-$V_t$ and sub-$V_t$ chain of inverters is listed in Table 2. The analytical and simulated results are plotted in Fig. 10, where the number of idle clock cycles corresponding to the break-even point in energy consumption is evaluated for 50 MHz to 1 GHz operating frequencies of the super-$V_t$ core. For example, the number of cycles $N$ when operating the super-$V_t$ core at a frequency of 1 GHz that results in the energy break-even point ($N_{break-even}$) is 125 and 133 for the analytical and simulated analysis, respectively. The $N_{\min}$ that satisfies (20) is taken as 133 cycles. *Core*1 must, therefore, be stacked for at least 133 ns when operating at 1 GHz for the leakage reuse technique to provide an improvement in the overall energy efficiency of the circuit. The percent difference between the analytical and simulated results in Fig. 10 is calculated as $\frac{|Analytical - Simulation|}{Simulation} \times 100$. The maximum and minimum percent error in the calculated number of idle cycles needed for the energy break-even point using the analytical expressions is, respectively, 11% (at 50 MHz) and 6% (at 1 GHz) as compared to simulated results. Note that the percent error decreases as the operating frequency increases since, for a given difference in the number of cycles between the analytical and simulated results, the relative difference is smaller at higher frequencies due to the larger number of cycles.

The total energy consumed by the super-$V_t$ and sub-$V_t$ cores per cycle and the energy consumed by the switches ($S_{VG}$ and $S_G$) per cycle are characterized for 1 GHz operation of the super-$V_t$ core starting from

**Table 2**
Characterization of active area and peak noise of the chain of inverters (COI), s27, and s208 circuits for the three analyzed topologies.

| | | Area of two super-$V_t$ circuit blocks (μm$^2$) | Switch area (μm$^2$) | Area of sub-$V_t$ circuit block (μm$^2$) | Peak noise on $V_{SS}$ as a percent of rail-to-rail voltage (1.2 V) |
|---|---|---|---|---|---|
| **COI** | Baseline | 2.331 | N/A | 7.56 | <5% |
| | Leakage reuse | 2.331 | 0.35 (15% of Super-$V_t$ COI) $S_{VG}$ : $S_G$ of 1.5 μm: 2 μm | 7.56 | <5% |
| | Voltage stacking | 2.331 | N/A | 7.56 | <5% |
| **s27** | Baseline | 6.73 | N/A | 21 | <5% |
| | Leakage reuse | 6.73 | 0.31 (4.6% of Super-$V_t$ s27)$S_{VG}$ : $S_G$ of 3 μm: 0.1 μm | 21 | <5% |
| | Voltage stacking | 6.73 | N/A | 21 | <5% |
| **s208** | Baseline | 21.9 | N/A | 21 | <5% |
| | Leakage reuse | 21.9 | 0.6 (2.7% of Super-$V_t$ s208) $S_{VG}$ : $S_G$ of 5 μm: 1 μm | 21 | <5% |
| | Voltage stacking | 21.9 | N/A | 21 | <5% |

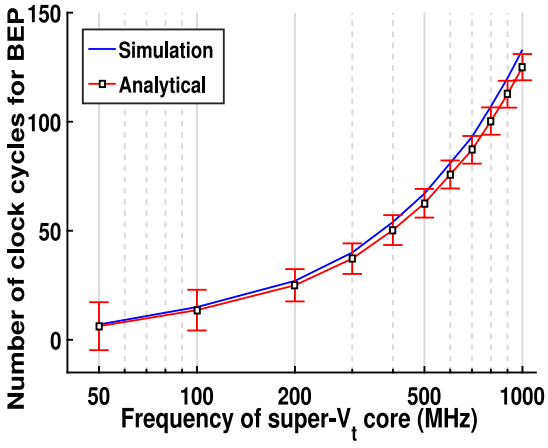


**Fig. 10.** Comparison between analytical and simulation results when determining the number of idle clock cycles of the super-$V_t$ core corresponding to the break-even point in energy consumption for frequencies of 50 MHz to 1 GHz.

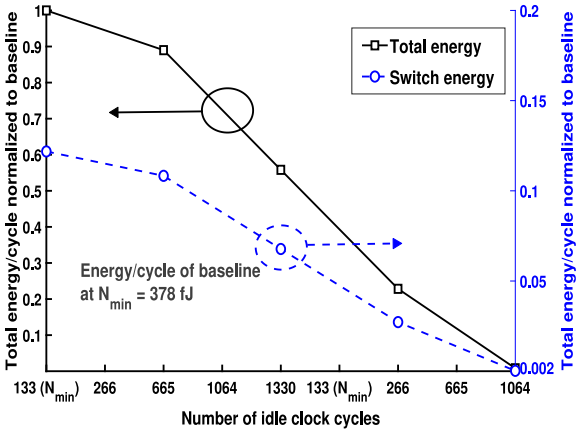


**Fig. 11.** Characterization of the energy per cycle normalized to the baseline for a range of idle cycles (133–1330) when the super-$V_t$ core is operating at 1 GHz and for 1.33 μs. The left Y-axis represents the energy consumed by both the super-$V_t$ and sub-$V_t$ core per cycle, while the right Y-axis represents energy consumed by switches $S_G$ and $S_{VG}$ per cycle.

$N_{min}$ (133 idle cycles at 1 GHz) to 10×$N_{min}$ (1330 idle cycles), with results shown in Fig. 11. The analysis of the leakage reuse technique is normalized to the baseline, where independent power delivery networks are implemented for the super-$V_t$ and sub-$V_t$ cores as described by Topology 1 in Section 5.1. For the leakage reuse technique, the size of the super-$V_t$ core is increased by 1.2× as compared to the baseline to account for the overhead due to the supporting circuits required to generate and propagate the control signal Φ. The overhead in area due to the supporting circuits as a percentage of the total area is smaller as the size of the super-$V_t$ core increases since the area of the supporting circuits remains about constant. Similar switching activity is applied to both the leakage reuse and the baseline implementations of the circuit. The savings in the energy per cycle due to the use of the leakage reuse technique increases as the leakage current from the super-$V_t$ core is supplied for an increasing number of idle cycles. For example, the total energy per cycle is 0.23× the baseline for 1064 idle cycles, while consuming 0.89× the baseline when a core is idle for 266 cycles. In addition, as the number of continuous idle cycles of the super-$V_t$ core increases, the fewer the number of triggered switching events needed to switch to an idle super-$V_t$ core, which reduces the energy overhead due to the switches as a percentage of the total energy consumed. As an example, the energy per cycle of two switches is 0.12× the total energy when charge is reused for 133 cycles, while the switches consume 0.027× of the total energy per cycle when charge is recycled for 1064 cycles.

## 5. Characterization and analysis

The proposed leakage reuse technique 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 target sub-$V_t$ supply voltage is 380 mV. The 45 nm fabrication process includes low-threshold (low-$V_t$), nominal threshold (nominal-$V_t$), and high-threshold devices (high-$V_t$). Low-$V_t$ transistors are used for the circuit blocks operating at a super-$V_t$ supply voltage to improve performance, while nominal-$V_t$ transistors are used for the circuit blocks operating at a sub-$V_t$ supply voltage to reduce power consumption. In addition, high-$V_t$ NMOS devices are used as sleep transistors $S_G^A$ and $S_G^C$ to reduce the idle power consumption, while low-$V_t$ PMOS transistors are used for the switches connecting to the sub-$V_t$ core. Note that the threshold voltage of a nominal-$V_t$, low-$V_t$, and high-$V_t$ NMOS transistor in a 45 nm process is, respectively, 410 mV, 322 mV, and 608 mV.

### 5.1. Model and simulation methodology

The leakage reuse technique is evaluated using a chain of inverters (COI) and two ISCAS89 benchmarks circuit (s27 with 19 gates and s208 with 112 gates). The ISCAS89 benchmark circuits represent super-$V_t$ cores, while separate COIs are used to represent both the 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 of the super-$V_t$ core.

Three circuit topologies are simulated, which are an implementation of the system level model shown in Fig. 5 applicable to both inter- and intra-core current reuse. The circuit topologies include:

**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 as each includes an independent power supply, 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 the super-$V_t$ circuit blocks is connected through the switching circuit shown in Fig. 6 to either true ground or the power network of a sub-$V_t$ circuit block consisting of a chain of tapered inverters. The leakage current of the super-$V_t$ circuit blocks is used as the current source for 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 the 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. In [40], stacked voltage domains within the same core are implemented for implicit down conversion of the voltage of the power supply of the near-threshold circuits, which are placed at the bottom of the stacked topology. In this work, the technique proposed in [40] is extended to deliver power to sub-$V_t$ circuits as shown in Fig. 4(a), which is then compared with the proposed leakage reuse technique.

Similar input signals and output capacitive loads are applied to 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 model of the DIP-40 package to analyze the transient noise induced on the power and ground networks. The resistance $R$, inductance $L$, and capacitance $C$ of the pins are, respectively, 0.217 Ω, 8.18 nH, and 5.32 pF.

The voltage stacking technique is not directly comparable to the leakage reuse technique as the circuit blocks in a voltage stacked system are continuously stacked for the entire duration of the operation of the circuit. In contrast, the leakage reuse technique stacks two circuit blocks within a core or in two different cores within a multi-core system only during the idle mode operation of the circuits operating at a super-$V_t$ supply voltage.

The three power delivery topologies are implemented on a `COI` and the `s27` and `s208` benchmark circuits operating at 5 MHz, where the active area of the super-$V_t$ and sub-$V_t$ circuit blocks is optimized for each of the three benchmarks circuits to obtain a sub-$V_t$ voltage of 380 mV. The total active area for the super-$V_t$ circuit blocks, sub-$V_t$ circuit blocks, and switches is listed in Table 2. The same circuit area is occupied by each topology, except for the leakage reuse technique, where an additional area is required for the switches. Therefore, iso-frequency and iso-area analysis is performed across the three topologies.

Due to the large reduction in the drive strength of the transistors when in sub-$V_t$ operation, the transistor widths are increased to provide sufficient drive current to charge and discharge loads, where the sub-$V_t$ circuit block is 3.24× and 3.12× the total area of the two super-$V_t$ circuit blocks for, respectively, a chain of inverters and the s27 benchmark circuit. However, the area of the sub-$V_t$ circuit block is 0.96× the size of the two s208 circuit blocks as the increased area of s208 (21.9 $\mu m^2$) provides sufficient current to drive the sub-$V_t$ circuit block (21 $\mu m^2$) without further increasing the size of the transistors.

The voltage on the virtual ground node $V_{GND}$, which is the supply voltage of the sub-$V_t$ block generated by the leakage current of the super-$V_t$ core, is dependent upon several circuit parameters including 1) the ratio of the area of the super-$V_t$ circuit block to the area of the sub-$V_t$ circuit block, which is directly correlated to the current ratio, 2) the total leakage current of the super-$V_t$ circuit block, 3) the on and off current through the switches, which are partly dependent on the threshold voltage, gate voltage, and the dimensions of the transistors, and 4) the total current required by the sub-$V_t$ circuit block. For a set size of the super-$V_t$ and sub-$V_t$ cores, the width of the switches acts as a controlling knob that allows for a trade-off between the voltage level at the VGND node and the total area and power consumption of the circuit. An analysis of the effect of switch size is, therefore, performed when the super-$V_t$ core is active, as described in Section 5.4.

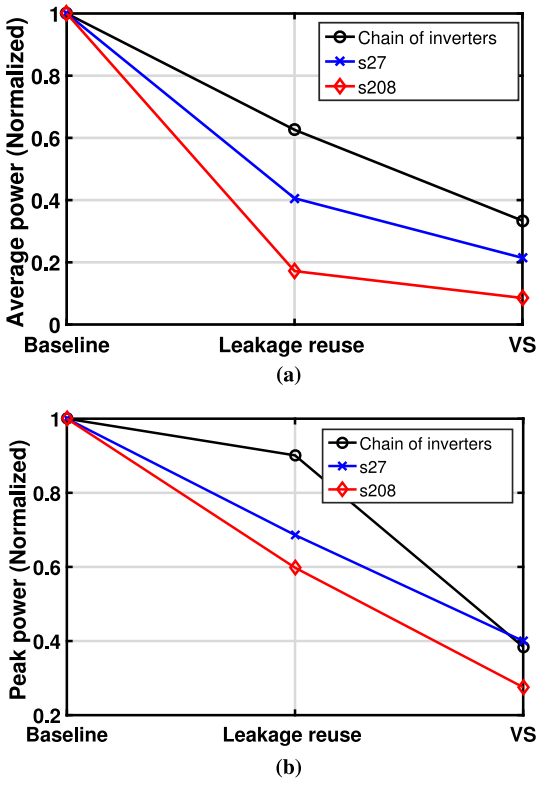


**Fig. 12.** Characterization of a) average power consumption per cycle, and b) peak power consumption of the `COI`, `s27`, and `s208` benchmark circuits at 5 MHz.

### 5.2. Characterization of power consumption

The average power consumption per cycle and peak instantaneous power consumption of the multi-voltage system, which includes the super-$V_t$ and sub-$V_t$ circuit blocks, for the three circuit topologies described in Section 5.1 are shown in, respectively, Fig. 12(a) and (b). The average (instantaneous peak) power consumption of the baseline is 25.24 μW (8.69 mW), 15.47 μW (4.154 mW), and 112.4 μW (7.636 mW) for, respectively, the `COI`, `s27`, and `s208` benchmark circuits. The average (peak) power consumption of the leakage reuse technique is 0.63× (0.9×), 0.41× (0.7×), and 0.17× (0.6×) that of the baseline for, respectively, `COI`, `s27`, and `s208`. The reduced power consumption due to the implementation of the leakage reuse technique is, therefore, greater as the circuit size 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` consumes an average and peak power of, respectively, 2.02× and 2.17× that of the voltage stacking technique.

The chain of inverters operating at a super-$V_t$ supply voltage of 1.2 V are characterized for both average and peak power consumption at operating frequencies between 5 MHz and 2 GHz using the baseline, leakage reuse, and voltage stacking techniques. The results from simulation of the average and peak power consumption at different operating

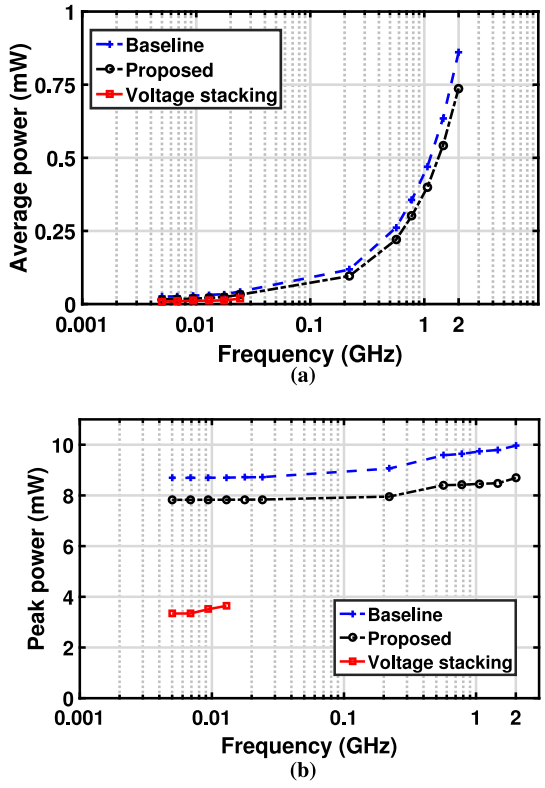


**Fig. 13.** Characterization of a) average power consumption per cycle and b) peak power consumption for 5 MHz to 2 GHz operation of the super-$V_t$ circuit blocks.

frequencies is shown in, respectively, Fig. 13(a) and (b). At all frequencies, the average and peak power consumption of the leakage reuse technique is less than the average and peak power consumption of the baseline. In addition, the leakage reuse technique provides a greater reduction in the average and peak power consumption at higher frequencies as compared to the baseline technique. Therefore, a sufficient amount of leakage current is supplied to the sub-$V_t$ core(s) during the idle mode operation of the nominal cores for a range of operating frequencies. Although voltage stacking consumes less power as compared to both leakage reuse and the baseline, the voltage stacking technique is limited to frequencies no greater than 25 MHz as the super-$V_t$ circuit does not provide correct logical output at higher frequencies due to the significant noise on the VGND node.

### 5.3. 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 the three circuit topologies described in Section 5.1. However, the voltage noise on the ground network exceeds ±5% of the ideal ground voltage of 0 V. The results from the characterization of the peak ground voltage bounce on the VSS node and the settling time of $V_{SS}$ (the time required to settle within ±5% of the ideal ground voltage) are shown in Fig. 14, where the results obtained for the leakage reuse and voltage stacking techniques are normalized to the results obtained for the baseline. The leakage reuse technique exhibits a reduced peak noise on the VSS node and a reduced 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 VSS and the $V_{SS}$ settling time to, respectively, 0.68× (0.29×) and 0.44× (0.37×) that of the baseline. The peak voltage noise (settling time of $V_{SS}$) on the true ground node VSS for the baseline is 30.07 mV (56 ns), 17.66 mV (110 ns), and 85.82 mV (56.5 ns) for, respectively, `COI`, `s27`, and `s208`.

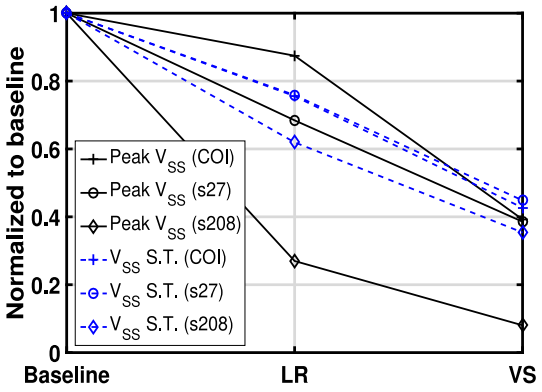


**Fig. 14.** Characterization of the peak $V_{SS}$ bounce and the settling time (S.T.) of the peak $V_{SS}$.

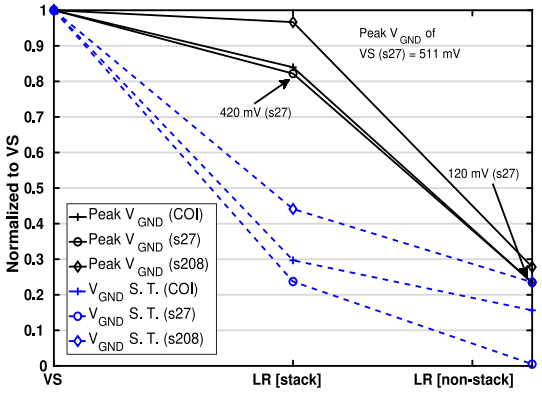


**Fig. 15.** Characterization of the peak $V_{GND}$ bounce and settling time of the $V_{GND}$ voltage, where switches for LR are sized for 10% noise on the stacked and non-stacked VGND node.

Unlike the voltage stacking method, the leakage reuse technique allows the stacking of only idle super-$V_t$ circuit blocks with the sub-$V_t$ circuit block, while connecting the active super-$V_t$ circuit blocks 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. The voltage bounce on the virtual ground node of the super-$V_t$ core is analyzed by characterizing the peak voltage noise and settling time (the time required to settle within ±5% of the steady state $V_{GND}$) of the `COI`, `s27`, and `s208` circuits each implemented with the leakage reuse (stacked and non-stacked) and voltage stacking techniques. The results from characterization of the peak $V_{GND}$ noise and the $V_{GND}$ settling time are shown in Fig. 15, where the leakage reuse technique is normalized to the voltage stacking technique. The peak $V_{GND}$ ($V_{GND}$ settling time) of the voltage stacking technique is 513.3 mV (64 ns), 511 mV (105 ns), and 431.4 mV (34 ns) for, respectively, the `COI`, `s27`, and `s208` circuits.

The steady-state $V_{GND}$ voltage on the VGND node is the supply voltage of the sub-$V_t$ circuit, which is set to 380 mV. The peak $V_{GND}$ bounce and $V_{GND}$ 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). In all cases, the voltage bounce on the VGND node for both the active and idle circuit blocks does not exceed 10% of the rail-to-rail voltage of 1.2 V for the super-$V_t$ block and 380 mV for the sub-$V_t$ block. The peak voltage noise on VGND ($V_{GND}$ 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,

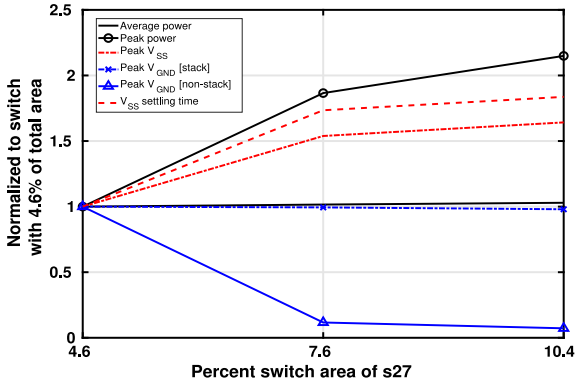


**Fig. 16.** Variation in peak power, average power, and settling time for s27 as a function of switch size.

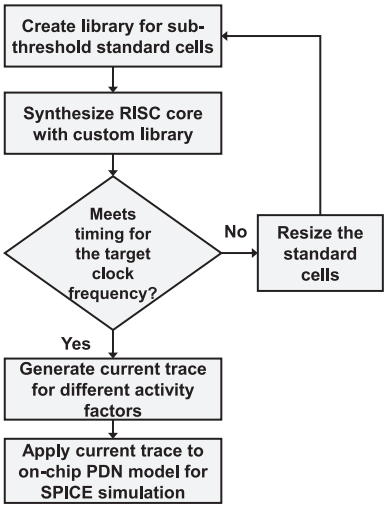


**Fig. 17.** Methodology to characterize workload variation of a sub-threshold core with SPICE simulation.

the stacked and non-stack mode, as compared and normalized to the voltage stacking technique.

The peak voltage noise on the virtual ground node VGND, on the true ground node VSS, and on the $V_{DD,sub-V_t}$ node are characterized in the *tt*, *ff*, and *ss* process corners for both the leakage reuse and voltage stacking techniques. The worst-case process variation, taken as the larger of the two values of $\frac{|tt-ff|}{tt}\times 100\%$ and $\frac{|tt-ss|}{tt}\times 100\%$, is determined for the `s208` benchmark circuit. The leakage reuse technique exhibits a worst-case variation in the peak $V_{GND}$ voltage noise for the stacked topology, peak $V_{GND}$ voltage noise for the non-stacked topology, peak $V_{SS}$ voltage noise, and sub-$V_t$ $V_{DD}$ noise of, respectively, 0.82%, 5.82%, 5.05%, and 1.54%. The voltage stacking technique exhibits a worst-case variation of 9.6%, 56.33%, and 2.05% in, respectively, the peak $V_{GND}$ noise, the peak $V_{SS}$ noise, and the peak noise on the sub-$V_t$ $V_{DD}$ node.

### 5.4. Characterization of switch size used for implementation of the leakage reuse technique

A circuit implementing the leakage reuse technique is sensitive to the dimensions of the transistors $S_G$ and $S_{VG}$ used as switches. The size of the switches impacts both the virtual and true ground nodes and, therefore, the power consumption and performance as a large instantaneous transient current is induced when the switches are turned on. The leakage reuse technique implemented on the s27 benchmark circuit is characterized for three different switch sizes, with results as shown in Fig. 16. The three implemented switches occupy 4.6%, 7.6%, and 10.4% of the total area of the two s27 circuit blocks operating at a super-$V_t$ supply voltage. As indicated from the results of the analysis, both the total and peak power consumption increase with larger switches. The switch occupying 10.4% of the area consumes an average and peak power of, respectively, 1.03× and 2.15× that of the switch occupying 4.6% of the total area. However, switches occupying 10.4% of the total area reduce the peak noise on the virtual ground node VGND to 0.07× and 0.98× in, respectively, the non-stacked and stacked mode, as compared to switches occupying 4.6% of the total area.

Due to the increased influx current (initial current) through the switches, the peak bounce on the true ground node VSS and the settling time of the $V_{SS}$ voltage for the switches occupying 10.4% of the total area increase by, respectively, 1.64× and 1.83× that of switches occupying 4.6% of the total area. Therefore, there is a trade-off between the power consumption and ground bounce when sizing the switches, where for the non-stacked mode, the peak noise on the VGND node is reduced to 0.07× at a cost of increasing the total power consumption and the switch area by, respectively, 1.03× and 2.26× when increasing the percentage of the total area occupied by the switches from 4.6% to 10.4%.

In addition, the noise on the VGND node (non-stacked), when the super-$V_t$ circuits are operating in active mode, is characterized for 5%, 10%, and 15% allowed voltage bounce from the nominal value of $V_{SS}$ (0 V). The reduction in the noise on the VGND node (non-stacked) results in an increase in the area and power consumption due to the required larger switches, as shown from results listed in Table 3. There is an average increase in the power consumption of 1.03× when the voltage noise on the VGND node is constrained to 5% rather than 15% of the rail-to-rail voltage of $V_{DD}$ (1.2 V for the super-$V_t$ core).

### 5.5. Regulation of sub-threshold power supply voltage considering workload variation in the sub-threshold core

The super-threshold cores or circuit blocks are idle when applying the leakage reuse technique. Therefore, there are no executing workloads on the super-threshold cores that directly affect the generated sub-threshold voltage. However, the variation in the workload of the sub-threshold core does affect the generated sub-threshold power supply

**Table 3**
Trade-off between the percentage noise permitted on the VGND node, the total power consumption, and the area in the non-stacked mode. Power consumption and area are normalized to the baseline topology (independent power delivery networks for the super and sub-$V_t$ cores) for the `COI`, `s27`, `s208` circuits.

| | | Permitted noise on VGND [non-stack] as compared to ideal ground | | |
|---|---|---|---|---|
| | | 5% | 10% | 15% |
| Total power normalized to baseline | COI | 0.642 | 0.635 | 0.616 |
| | s27 | 0.410 | 0.409 | 0.408 |
| | s208 | 0.174 | 0.168 | 0.167 |
| Total area normalized to baseline | COI | 1.0480 | 1.0261 | 1.0182 |
| | s27 | 1.0083 | 1.0079 | 1.0072 |
| | s208 | 1.0096 | 1.0081 | 1.0075 |

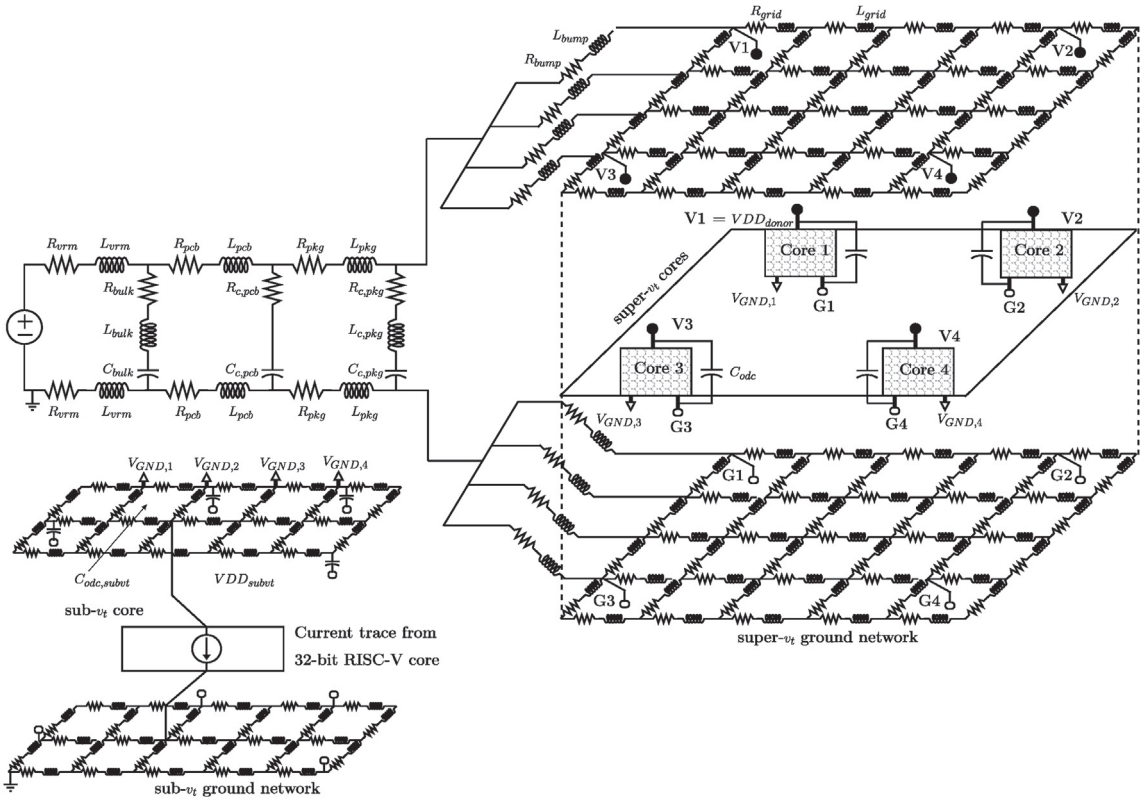


**Fig. 18.** Circuit model of the off-chip and on-chip PDN used to evaluate the voltage noise due to variation in the workload of a 32-bit RISC-V core operating at a sub-threshold voltage of 380 mV.

voltage. Specifically, the generated voltage at the VGND node increases when a sub-threshold core transitions from a high activity state to a low activity state and decreases when transitioning from a low to high activity state. For the transition from a high activity state to a low activity state and for a given sequence of tasks of an executing application *A1*, there is a reduction in the current demand of the sub-threshold core. Therefore, the recycled leakage current from the idle super-threshold core is greater than the required current of the sub-threshold core when executing the application *A1*. The effects of workload variation of a sub-threshold core on the generated sub-threshold supply voltage is, therefore, analyzed. A 32-bit RISC-V core is synthesized in a 45 nm CMOS technology with an operating voltage and frequency of respectively, 380 mV and 1 MHz. Custom logic cells are designed for a 380 mV supply voltage as standard cells and libraries are not properly optimized for sub-threshold operation in standard process development kits (PDKs) [51]. The methodology to characterize the workload variation of a synthesized sub-threshold 32-bit RISC-V core is shown in Fig. 17. The custom cells are used to estimate the total current of the RISC-V core, which is obtained for different activity factors that emulate variations in the workload of the core. The sensitivity of the sub-$V_t$ supply voltage to variation in the workloads of the sub-$V_t$ core is described in Section 5.5.1. In addition, a dynamic workload-aware idle core assignment that provides a stable sub-$V_t$ supply voltage is discussed in Section 5.5.2.

### 5.5.1. Sensitivity of sub-$V_t$ supply to workload variations

The circuit used to evaluate the voltage noise on the VGND node due to variations in the workload of the RISC-V core is shown in Fig. 18. Similar to Fig. 6, four super-threshold cores supply current to a sub-threshold core through the proposed leakage reuse technique. The super-threshold cores operate at 1.2 V and with a clock frequency of 1 GHz, while a 380 mV supply voltage is generated for the RISC-V core operating at a frequency of 1 MHz. A current source is placed between the sub-threshold power and ground grids to emulate the total current load of the synthesized RISC-V core for different activity factors.

To accurately characterize the voltage noise on the sub-threshold supply voltage with variations in the workload of a RISC-V core operating at 380 mV, the four super-threshold cores are supplied current through a distributed power delivery network (PDN) modeled with both off-chip and on-chip power and ground networks as shown in Fig. 18. The electrical parameters of the off-chip and on-chip power and ground grids are listed in Table 5. A 6×5 on-chip grid is implemented for simulation to accurately model the impedance of the PDN, where the electrical parameters of the on-chip grid ($R_{grid}$ and $L_{grid}$) and the C-4 bumps ($R_{bump}$ and $L_{bump}$) are based on the impedance of a PDN of a four core SoC described in [52]. The electrical impedance of the off-chip voltage regulator module (VRM), board (PCB), and package are determined based on the PDN models presented in [53,54]. The output of the VRM is maintained at 1.2 V, which emulates the closed-loop response of the VRM.

The PDN is optimized to maintain the voltage ripple to within 10% of the $V_{DD}$ voltage and $V_{SS}$ voltage. On-chip decoupling capacitors $C_{odc}$ are placed close to each load to compensate for the voltage ripples on the PDN, where a total of 4 nF of decoupling capacitance is included for the four super-threshold cores. In addition, four on-chip decoupling capacitors, $C_{odc,subvt}$, each providing a capacitance of 1.5 nF are placed between the power ($V_{GND,1}$, $V_{GND,2}$, $V_{GND,3}$, and $V_{GND,4}$) and ground nodes of the sub-threshold core to minimize the voltage ripple on the power distribution network of the sub-threshold core during the highest circuit activity.

Each of the four super-threshold cores is composed of a s27 ISCAS89 benchmark circuit and a chain of buffers. The chain of buffers are added to accurately model the ratio of the logic area of the super-threshold core to the sub-threshold core. The circuit model shown in Fig. 18 is used for SPICE simulation to characterize the current profile of workloads executing on a synthesized 32-bit RISC-V core, where the circuit

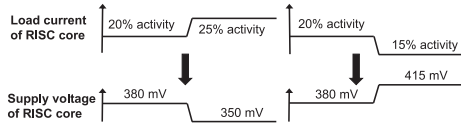


**Fig. 19.** A 380 mV supply voltage is generated for the RISC core through the leakage reuse technique, where the target activity factor is 20%. The variation in the supply voltage of the RISC core is characterized for circuit activity factors of 15% to 25%.

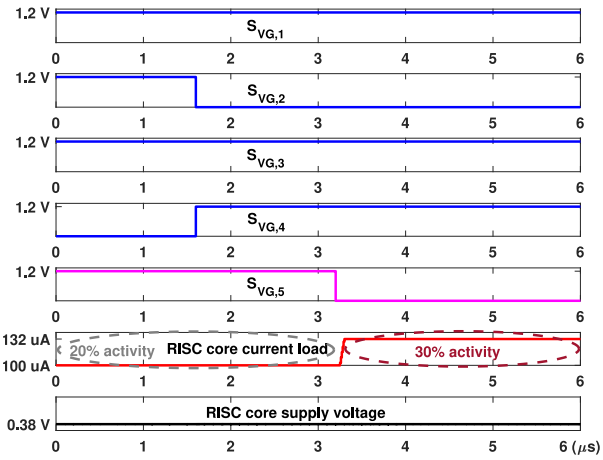


**Fig. 20.** SPICE simulation results indicating a stable sub-threshold supply voltage when the activity of the RISC-V core increases from 20% to 30%.

activity is restricted to between 15% and 25% of all logical gates. The total logical area of each super-$V_t$ core is 1 mm$^2$, while the total logical area of the RISC-V core is 10× larger (10 mm$^2$) than each super-threshold core. The 1 to 10 ratio of the area of the super-threshold to sub-threshold cores results in a 380 mV supply voltage for the sub-threshold RISC-V core executing workloads that activate 20% of all devices. In addition, for the given area ratio, the supply voltage is maintained within ±10% of 380 mV when the activity factor of the RISC-V core varies between 15% and 25% as indicated by results shown in Fig. 19. Therefore, a stable supply with a bounded variation in voltage is available to the sub-threshold RISC-V core as long as the executing workloads are bound to a utilization of 15% to 25% of all transistors at any given time. However, for the case when the activity of the RISC-V core exceeds 25% or drops below 15%, a sub-threshold workload-aware idle core assignment technique is implemented to maintain a stable sub-threshold supply voltage.

### 5.5.2. Dynamic super-threshold idle core assignment based on sub-threshold workloads

In the case when the circuit activity factor of the RISC-V core exceeds 25%, a novel sub-threshold workload aware idle core assignment methodology is applied. Depending on the executing workloads and, therefore, activity factor of the sub-threshold RISC-V core, the number of super-threshold idle cores or circuit blocks providing current to the RISC-V core is either increased or decreased at run time to maintain the required supply voltage at the VGND node. Therefore, close-loop voltage regulation of the power network is not required.

The proposed workload-aware idle core assignment is evaluated through SPICE simulation with the results shown in Fig. 20, where the noise on the grid of the sub-threshold power supply is characterized for a change in the circuit activity of the RISC-V core from 20% to 30% transistor utilization. The total current of the RISC-V core operating at a clock frequency of 1 MHz with a 380 mV supply voltage is 82 μA, 100 μA, 115 μA, and 132 μA for, respectively, 15%, 20%, 25%, and 30% activity factors. The execution of the leakage reuse technique on the four super-threshold cores is managed by four control signals, $S_{VG,1}$, $S_{VG,2}$, $S_{VG,3}$, and $S_{VG,4}$, which are applied to respectively, core 1, core 2, core 3, and core 4 as shown in Fig. 20. The circuit model shown in Fig. 18 is used to obtain the results shown in Fig. 20 through SPICE simulation. The current load of the sub-threshold RISC-V core is changed from 100 μA (20% activity) to 132 μA (30% activity) to evaluate the voltage noise generated on the VGND node when leakage reuse is applied. A sub-block of core 4 is added to the existing circuit configuration through a fifth control signal $S_{VG,5}$ to provide the additional charge needed to maintain the power supply voltage of the sub-threshold core within ±10% of 380 mV. The sub-block is sized such that the provided leakage current is sufficient to meet the additional current demand of the RISC-V core as the load increases to 132 μA from 100 μA. Therefore, a stable supply voltage of 380 mV is maintained on the power supply network of the sub-threshold core despite the increase in the circuit activity of the RISC-V core from 20% to 30%.

### 5.6. Trade-off between voltage stacking and leakage reuse technique

The transient analysis of both the leakage reuse and voltage stacking techniques is shown in Fig. 21, where the voltage bounce on the VGND node and the voltage levels of the logical output from the super-$V_t$ and sub-$V_t$ cores are compared. Similar clock and input switching patterns are applied to both implementations of the circuit.

The transient 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, $S_{VG}$, $S_G$, $V_{GND}$, and output signals of one core are shown in Fig. 21. In addition, clock-gating is applied when *Core*1 is 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 couple into the virtual ground node, where the charging and discharging of internal nodes results in significant voltage noise due to a large amount of internal capacitance. The noise induced on the VGND node directly impacts the logical output of *Core* 2, where the voltage on VGND varies from 0.17 V to 0.41 V for a logic 1, which causes unwanted delay and logical failure for the circuits connected to the logical outputs of the super-$V_t$ cores. In addition, the noise induced on the virtual ground node of the voltage stacked super-$V_t$ core directly couples to the supply node of the sub-$V_t$ circuits as the VGND node and sub-$V_t$ supply node are shorted when voltage stacking is implemented. Therefore, additional voltage regulators are required to generate a sta-

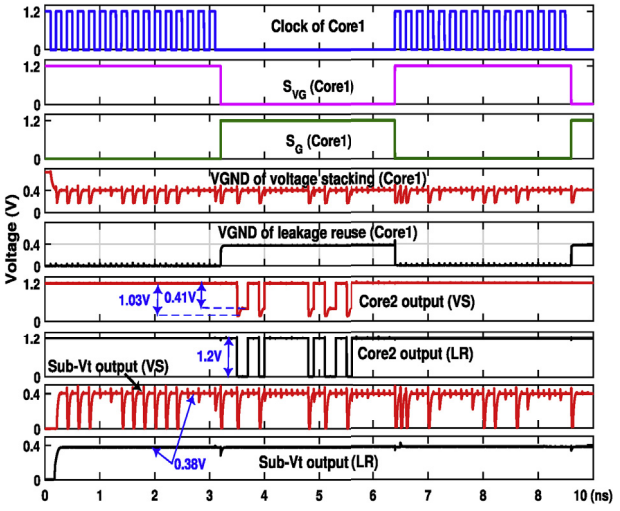


**Fig. 21.** Transient simulation of the leakage reuse and voltage stacking techniques comparing the noise on the virtual ground node and the voltage of the output signals of the super-$V_t$ and sub-$V_t$ circuits.

**Table 4**
Characterization of power consumption and noise through SPICE simulation of s27 implemented with the baseline, leakage reuse, and voltage stacking techniques.

| | Total power (μW) | Peak power (mW) | Peak $V_{dd}$ (V) | Peak $V_{ss}$ (mV) | Peak $V_{GND}$ [non-stack] (mV) | Peak $V_{GND}$ [stack] (mV) | Switching speed | Super-$V_t$ transistors | Switch transistors | Sub-$V_t$ transistors |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | 15.47 | 4.154 | 1.219 | 17.66 | N/A | N/A | 4× clock speed | Low $V_t$ | N/A | Nominal $V_t$ |
| Leakage reuse | 6.448 | 3.822 | 1.211 | 13.41 | 120 | 420 | 4× clock speed | Low $V_t$ | NMOS: 0.9 μm (high $V_t$), PMOS: 3 μm (low $V_t$) | Nominal $V_t$ |
| Voltage stacking | 3.315 | 1.662 | 1.206 | 6.803 | N/A | 511 | 4× clock speed | Low $V_t$ | N/A | Nominal $V_t$ |

**Table 5**
Off-chip and on-chip electrical parameters of the power delivery network [52–54].

| Resistance | Value (Ω) | Inductance | Value (H) | Capacitance | Value (F) |
|---|---|---|---|---|---|
| $R_{vrm}$ | 100 u | $L_{vrm}$ | 5 n | $C_{bulk}$ | 132 u |
| $R_{pcb}$ | 485 u | $L_{bulk}$ | 510 p | $C_{pcb}$ | 10 u |
| $R_{bulk}$ | 4.365 m | $L_{pcb}$ | 48 p | $C_{pkg}$ | 440 n |
| $R_{ppcb}$ | 10 m | $L_{pkg}$ | 19 p | $C_{odc}$ | 1 n |
| $R_{pkg}$ | 123 u | $L_{ppkg}$ | 104 p | $C_{odc,subvt}$ | 1.5 n |
| $R_{ppkg}$ | 15.72 m | $L_{bump}$ | 38 p | | |
| $R_{bump}$ | 400 u | $L_{grid}$ | 5.6 f | | |
| $R_{grid}$ | 50 m | | | | |

ble supply voltage for the bottom core in the stack when voltage stacking is implemented, which results in a significant overhead in energy consumption [38,40,41].

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, present at the VGND node used to supply current to the sub-$V_t$ core. The logical outputs of *Core*2 are full-swing voltage signals, even as *Core*1 is stacked with the sub-$V_t$ circuits. In addition, as the $V_{GND}$ voltage is stable, the output signals from the sub-$V_t$ core are also stable, as shown in Fig. 21.

Simulated results of the power consumption and peak voltage on the VGND and true ground nodes obtained for the s27 ISCAS benchmark circuit, which represents each super-$V_t$ circuit block, and a chain of inverters, which represents the sub-$V_t$ circuit block, are listed in Table 4. The super-$V_t$ and sub-$V_t$ circuit blocks are sized to generate a sub-$V_t$ supply voltage of 380 mV.

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 voltage bounce on the VGND node, as indicated by results shown in Figs. 12, 14 and 15. For `s27`, the resulting peak noise on VGND due to the implementation 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. 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 for ideal ground and for voltages of 120 mV, 420 mV, and 511 mV on the VGND node obtained from the results shown in Fig. 15 for, respectively, the non-stacked LR, the stacked LR, and the VS technique. The implementation of the voltage stacking and the leakage reuse techniques result in an increase in the FO4 delay of, respectively, 4.17× ($V_{GND}$ voltage of 511 mV) and 1.24× ($V_{GND}$ voltage of 120 mV) as compared to the baseline technique (ideal ground).

In addition, the noise margins of the circuits connected to the output of `s27` are significantly degraded as the voltage on the VGND node increases. The noise margins of a CMOS inverter with a PMOS width of 3.6 μm and NMOS width of 1.2 μm are characterized for an operating frequency of 5 MHz and with a 5 fF capacitive load. The noise margin low of the CMOS inverter is 483 mV, which is less than the 511 mV voltage on the VGND node when implementing the voltage stacking technique. Therefore, logic low is not discernible with the voltage stacking technique if an inverter is connected to the output of `s27`. In contrast, the leakage reuse technique implemented on `s27` provides a discerning logic low as the voltage on the VGND node is 120 mV, while also reducing the total and peak power consumption to, respectively, 0.41× and 0.7× that of the baseline. However, the FO4 delay increases by 1.24× as compared to the baseline. The peak voltage of 420 mV on the VGND node 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 operating in idle mode.

## 6. Conclusions

A novel leakage reuse methodology of delivering power to sub-$V_t$ circuits is proposed, where the unused leakage current from the super-$V_t$ circuits supply the sub-$V_t$ circuit blocks. An additional voltage source is, therefore, not required for the sub-$V_t$ circuits, while the leakage overhead of the super-$V_t$ circuits is also reduced. An analysis of the energy break-even point is presented, which provides the minimum idle number of cycles of a super-$V_t$ core required to achieve an overall improvement in the energy-efficiency of the multi-core system. The proposed leakage reuse technique implemented on the `s27` ISCAS89 benchmark circuit reduces the average and peak power consumption to, respectively, 0.41× and 0.7× that of the baseline. In addition, the peak voltage noise on VSS 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 average (peak) power consumption of 2.02× (2.17×), while reducing the peak $V_{GND}$ ($V_{GND}$ settling time) to 0.28× (0.23×) that of the voltage stacking technique. In addition, a novel method of dynamically assigning idle super-$V_t$ cores or circuit blocks for leakage reuse is implemented to maintain a stable sub-$V_t$ supply voltage for workloads executing on sub-$V_t$ circuits with variable current load demands.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mejo.2020.104782.

## References

[1] H. Singh, K. Agarwal, D. Sylvester, K.J. Nowka, Enhanced leakage reduction techniques using intermediate strength power gating, IEEE Transactions on Very Large Scale Integrated Systems, 15 (11) (November 2007) 1215–1224.

[2] S. Kim, S.V. Kosonocky, D.R. Knebel, K. Stawiasz, Experimental measurement of a novel power gating structure with intermediate power saving mode, in: Proceedings of the International Symposium on Low Power Electronics and Design, August 2004, pp. 20–25.
[3] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits, Proceedings of the IEEE, 91 (2) (February 2003) 305–327.
[4] S. Jain, S. Khare, S. Yada, V. Ambili, P. Salihundam, S. Ramani, S. Muthukumar, M. Srinivasan, A. Kumar, S.K. Gb, et al., A 280mV-to-1.2 V wide-operating-range IA-32 processor in 32nm CMOS, in: Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers, February 2012, pp. 66–68.
[5] A. Muttreja, N. Agarwal, N.K. Jha, CMOS logic design with independent-gate FinFETs, in: Proceedings of the International Conference on Computer Design, August 2007, pp. 560–567.
[6] A. Esmaili, M. Nazemi, M. Pedram, Modeling processor idle times in MPSoC platforms to enable integrated DPM, DVFS, and task scheduling subject to a hard deadline, in: Proceedings of the Asia and South Pacific Design Automation Conference, January 2019, pp. 532–537.
[7] M. Arora, S. Manne, I. Paul, N. Jayasena, D.M. Tullsen, Understanding idle behavior and power gating mechanisms in the context of modern benchmarks on CPU-GPU integrated systems, in: Proceedings of the IEEE International Symposium on High Performance Computer Architecture, March 2015, pp. 366–377.
[8] M.S. Hossain, I. Savidis, Bi-directional input/output circuits with integrated level shifters for near-threshold computing, in: Proceedings of the IEEE International Midwest Symposium on Circuits and Systems, August 2017, pp. 1240–1243.
[9] E. Kultursay, K. Swaminathan, V. Saripalli, V. Narayanan, M.T. Kandemir, S. Datta, Performance enhancement under power constraints using heterogeneous CMOS-TFET multicores, in: Proceedings of the IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis, October 2012, pp. 245–254.
[10] S. Moradi, N. Qiao, F. Stefanini, G. Indiveri, A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs), IEEE Transactions on Biomedical Circuits and Systems, 12 (1) (August 2017) 106–122.
[11] H. Mair, G. Gammie, A. Wang, S. Gururajarao, I. Lin, H. Chen, W. Kuo, A. Rajagopalan, W. Ge, R. Lagerquist, A highly integrated smartphone SoC featuring a 2.5 GHz octa-core CPU with advanced high-performance and low-power techniques, in: Proceedings of the IEEE International Solid-State Circuits Conference, March 2015, pp. 1–3.
[12] Z. Wang, Y. Liu, Y. Sun, Y. Li, D. Zhang, H. Yang, An energy-efficient heterogeneous dual-core processor for internet of things, in: Proceedings of the IEEE International Symposium on Circuits and Systems, July 2015, pp. 2301–2304.
[13] B. Keller, M. Cochet, B. Zimmer, J. Kwak, A. Puggelli, Y. Lee, M. Blagojević, S. Bailey, P. Chiu, P. Dabbelt, A RISC-V processor SoC with integrated power management at submicrosecond timescales in 28 nm FD-SOI, IEEE Journal of Solid-State Circuits, 52 (7) (July 2017) 1863–1875.
[14] M. Igarashi, T. Uemura, R. Mori, H. Kishibe, M. Nagayama, M. Taniguchi, K. Wakahara, T. Saito, M. Fujigaya, K. Fukuoka, A 28 nm high-K/MG heterogeneous multi-core mobile application processor with 2 GHz cores and low-power 1 GHz cores, IEEE Journal of Solid-State Circuits, 50 (1) (January 2015) 92–101.
[15] M. Konijnenburg, Y. Cho, M. Ashouei, T. Gemmeke, C. Kim, J. Hulzink, J. Stuyt, M. Jung, J. Huisken, S. Ryu, Reliable and energy efficient 1 MHz 0.4 V dynamically reconfigurable SoC for ExG applications in 40 nm LP CMOS, in: Proceedings of the IEEE International Solid-State Circuits Conference, February 2013, pp. 430–431.
[16] K. Craig, Y. Shakhsheer, S. Arrabi, S. Khanna, J. Lach, B.H. Calhoun, A 32b 90 nm processor implementing panoptic DVS achieving energy efficient operation from sub-threshold to high performance, IEEE Journal of Solid-State Circuits, 49 (2) (February 2014) 545–552.
[17] N. Mehta, B. Amrutur, Dynamic supply and threshold voltage scaling for CMOS digital circuits using in-situ power monitor, IEEE Transactions on Very Large Scale Integration Systems, 20 (5) (May 2012) 892–901.
[18] K. Chong, K. Chang, B. Gwee, J.S. Chang, Synchronous-logic and globally-asynchronous-locally-synchronous (GALS) acoustic digital signal processors, IEEE Journal of Solid-State Circuits, 47 (3) (March 2012) 769–780.
[19] A. Iyer, D. Marculescu, Power efficiency of voltage scaling in multiple clock, multiple voltage cores, in: Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, November 2002, pp. 379–386.
[20] S. Rusu, S. Tam, H. Muljono, et al., A dual-core multi-threaded Xeon processor with 16 MB L3 cache, in: Proceedings of the IEEE International Solid State Circuits Conference, February 2006, pp. 315–324.
[21] A.S. Leon, K.W. Tam, J.L. Shin, D. Weisner, F. Schumacher, A power-efficient high-throughput 32-thread SPARC processor, IEEE Journal of Solid State Circuits, 42 (1) (January 2007) 7–16.
[22] 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-200 mV, 3.5 pJ/inst, 11 nW subthreshold processor, in: Proceedings of the IEEE Symposium on VLSI Circuits, June 2007, pp. 152–153.
[23] S. Hanson, M. Seok, Y. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, D. Blaauw, A low voltage processor for sensing applications with picowatt standby mode, IEEE Journal of Solid State Circuits, 44 (4) (April 2009) 1145–1155.
[24] S.C. Jocke, J.F. Bolus, S.N. Wooters, A.D. Jurik, A.C. Weaver, T.N. Blalock, B.H. Calhoun, A 2.6-$\mu$W sub-threshold mixed-signal ECG SoC, in: Proceedings of the IEEE Symposium on VLSI Circuits, June 2009, pp. 60–61.
[25] M.S. Hossain, I. Savidis, Reusing leakage current for improved energy efficiency of multi-voltage systems, in: Proceedings of the IEEE International Symposium on Circuits and Systems, May 2019, pp. 1–5.
[26] 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 14 nm tri-gate CMOS, in: Proceedings of the IEEE International Solid-State Circuits Conference, February 2018, pp. 38–40.
[27] T. Tong, S.K. Lee, X. Zhang, D. Brooks, 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, 51 (9) (July 2016) 2142–2152.
[28] M.S. Hossain, I. Savidis, Multi-voltage domain power distribution network for optimized ultra-low voltage clock delivery, in: Proceedings of the IEEE International Green and Sustainable Computing Conference, October 2018, pp. 1–8.
[29] B. Pandey, J. Yadav, M. Pattanaik, N. Rajoria, Clock gating based energy efficient ALU design and implementation on FPGA, in: Proceedings of the International Conference on Energy Efficient Technologies for Sustainability, June 2013, pp. 93–97.
[30] Q. Diao, J. Song, Prediction of CPU idle-busy activity pattern, in: Proceedings of the IEEE International Symposium on High Performance Computer Architecture, October 2008, pp. 27–36.
[31] J. Leverich, M. Monchiero, V. Talwar, P. Ranganathan, C. Kozyrakis, Power management of datacenter workloads using per-core power gating, IEEE Computer Architecture Letters, 8 (2) (August 2009) 48–51.
[32] R. Ye, Q. Xu, Learning-based power management for multicore processors via idle period manipulation, IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, 33 (7) (July 2014) 1043–1055.
[33] L. Benini, A. Bogliolo, G.D. Micheli, A survey of design techniques for system-level dynamic power management, IEEE Transactions on Very Large Scale Integration Systems, 8 (3) (June 2000) 299–316.
[34] A. Silva, M. Liu, M. Moghaddam, Power-management techniques for wireless sensor networks and similar low-power communication devices based on nonrechargeable batteries, Journal of Computer Networks and Communications, 2012 (August 2012) 1–10.
[35] A. Zou, J. Leng, X. He, Y. Zu, C.D. Gill, V.J. Reddi, X. Zhang, Voltage-stacked power delivery systems: reliability, efficiency, and power management, in: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, January 2020, pp. 1–13.
[36] J. E. Myers, “Unregulated Voltage Stacked Memory,” August 2019, US Patent 10,395,724.
[37] R.T. Possignolo, E. Ebrahimi, E.K. Ardestani, A. Sankaranarayanan, J.L. Briz, J. Renau, GPU NTC process variation compensation with voltage stacking, IEEE Transactions on Very Large Scale Integration Systems, 26 (9) (September 2018) 1713–1726.
[38] K. Blutman, A. Kapoor, J.G. Martinez, H. Fatemi, J.P. Gyvez, Lower power by voltage stacking: a fine-grained system design approach, in: Proceedings of the ACM/IEEE Annual Design Automation Conference, June 2016, pp. 78:1–78:5.
[39] A.C. Cabe, Z. Qi, M.R. Stan, Stacking SRAM banks for ultra low power standby mode operation, in: Proceedings of the ACM/IEEE Design Automation Conference, June 2010, pp. 699–704.
[40] K. Mazumdar, M.R. Stan, Charge recycling on-chip DC-DC conversion for near-threshold operation, in: Proceedings of the IEEE Subthreshold Microelectronics Conference, October 2012, pp. 1–3.
[41] R. Zhang, K. Mazumdar, B.H. Meyer, K. Wang, K. Skadron, M.R. Stan, A cross-layer design exploration of charge-recycled power-delivery in many-layer 3D-IC, in: Proceedings of the ACM/IEEE Design Automation Conference, June 2015, pp. 1–6.
[42] R. Zhang, K. Mazumdar, B.H. Meyer, K. Wang, K. Skadron, M.R. Stan, Transient voltage noise in charge-recycled power delivery networks for many-layer 3D-IC, in: Proceedings of the IEEE/ACM International Symposium on Low Power Electronics and Design, September 2015, pp. 152–158.
[43] E. Pakbaznia, F. Fallah, M. Pedram, Charge recycling in power-gated CMOS circuits, IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, 27 (10) (October 2008) 1798–1811.
[44] H. Jiang, M. Marek-Sadowska, S.R. Nassif, Benefits and costs of power-gating technique, in: Proceedings of the IEEE International Conference on Computer Design, October 2005, pp. 559–566.
[45] S. Roy, N. Ranganathan, S. Katkoori, A framework for power-gating functional units in embedded microprocessors, IEEE Transactions on Very Large Scale Integration Systems, 17 (11) (March 2009) 1640–1649.
[46] A. Lungu, P. Bose, A. Buyuktosunoglu, D.J. Sorin, Dynamic power gating with quality guarantees, in: Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design, August 2009, pp. 377–382.
[47] Q. Xie, X. Lin, Y. Wang, S. Chen, M.J. Dousti, M. Pedram, Performance comparisons between 7-nm FinFET and conventional bulk CMOS standard cell libraries, IEEE Transactions of Circuits and Systems II: Express Briefs, 62 (8) (August 2015) 761–765.
[48] Z. Hu, A. Buyuktosunoglu, V. Srinivasan, V. Zyuban, H. Jacobson, P. Bose, Microarchitectural techniques for power gating of execution units, in: Proceedings of the International Symposium on Low Power Electronics and Design, August 2004, pp. 32–37.

[49] K. Usami, N. Ohkubo, A design approach for fine-grained run-time power gating using locally extracted sleep signals, in: Proceedings of the IEEE International Conference on Computer Design, October 2007, pp. 155–161.
[50] Y. Çakmak, W. Toms, J. Navaridas, M. Luján, Cyclic power-gating as an alternative to voltage and frequency scaling, IEEE Comput. Architect. Lett. 15 (2) (September 2015) 77–80.
[51] M. Li, C. Ieong, M. Law, P. Mak, M. Vai, R.P. Martins, Sub-threshold standard cell library design for ultra-low power biomedical applications, in: Proceedings of the IEEE International Conference on Engineering in Medicine and Biology Society, July 2013, pp. 1454–1457.
[52] M.S. Gupta, J.L. Oatley, R. Joseph, G. Wei, D.M. Brooks, Understanding voltage variations in chip multiprocessors using a distributed power-delivery network, in: Proceedings of the Conference on Design, Automation and Test in Europe, April 2007, pp. 624–629.
[53] L.D. Smith, E. Bogatin, Principles of Power Integrity for PDN Design: Robust and Cost Effective Design for High Speed Digital Products, Prentice Hall, 2017.
[54] D. Pathak, I. Savidis, On-chip power supply noise suppression through hyperabrupt junction varactors, IEEE Transactions on Very Large Scale Integration Systems, 26 (11) (November 2018) 2230–2240.