Near-Threshold Voltage (NTV) Computing

Vivek De, Sriram Vangal, and Ram Krishnamurthy
Intel Corporation

Editor's note:
Near-threshold computing has emerged as an attractive paradigm for energy efficiency. This article discusses challenges and opportunities for designing complex system on chips that can operate in the near-threshold voltage range. Evaluation for 32- and 22-nm test chips is presented.
—Muhammad Shafique, Vienna University of Technology

NEAR-THRESHOLD VOLTAGE (NTV) computing is about computing at the point of peak energy efficiency. As you scale supply voltage, energy consumed per operation goes down. Performance also reduces. As the voltage approaches the threshold voltage of the transistor, energy efficiency peaks. As you go down further in voltage, energy efficiency actually becomes worse. The minimum operating energy is achieved at the point where the switching and leakage components are in balance. At voltages larger than the NTV operating point, switching energy dominates. Below the NTV operating point, leakage energy dominates since the cycle time becomes very large.

Dennard scaling ended a decade ago since continued scaling of the supply voltage while maintaining adequate transistor performance would require proportional scaling of transistor threshold voltage, causing an exponential increase in transistor leakage. As a result, chip performance has been limited by the hotspot power density over the last few technology generations. Even as higher levels of integration have become increasingly affordable with continued process technology scaling, stringent power limits preclude full utilization of the available silicon capability, leading to advent of the "dark silicon" era. These challenges can be mitigated to some extent by using highly parallel NTV computing for throughput-oriented applications.

In this article, we present challenges and opportunities for realizing NTV computing in complex system-on-chip (SoC) designs in scaled CMOS process nodes. Examples of NTV designs in 32-nm high-K/metal-gate and 22-nm trigate CMOS process are highlighted.

Digital Object Identifier 10.1109/MDAT.2016.2573593
Date of publication: 26 May 2016; date of current version: 23 February 2017.

NTV IA processor

NTV design challenges

The most common limit to voltage scaling is failure of SRAM and logic circuits. SRAM cells fail at low voltage because device mismatches degrade stability of the cell for read or write or data retention. SRAM cells typically use the smallest transistors. Also, they are the most abundant among all circuit types on a die. Therefore, the minimum operating voltage (Vmin) of the SRAM cell array limits Vmin of the entire chip. Logic circuits, clocking, and sequentials fail at low voltage because of noises, process variations, etc. Alpha and cosmic ray-induced soft errors cause transient failure of memory, sequentials, and logic at low voltage. Frequency starts degrading exponentially as voltage approaches transistor threshold. This sets a limit on Vmin. This limit can be alleviated to some extent by trigate or FinFET transistors. Since they have a steeper subthreshold swing, they can provide a lower threshold voltage (Vt) for the same leakage current target. Aging degradations

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

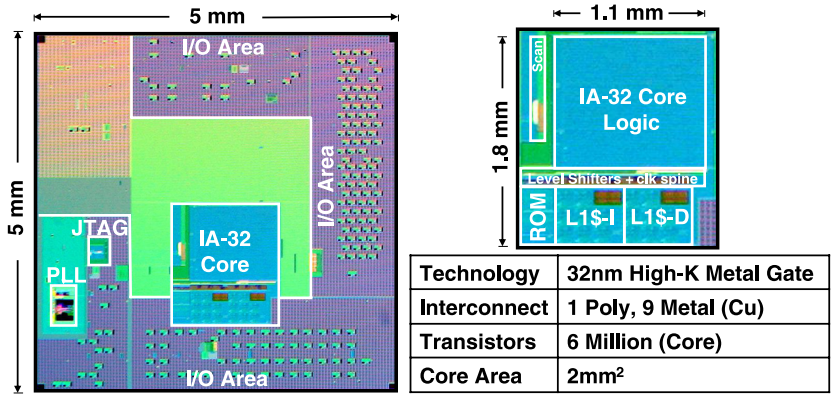
cause failure of SRAM cells at low voltages since different transistors in the cell undergo different amounts of Vt shift under voltage–temperature stress and thus worsen device mismatches in the cell. All of these effects degrade and limit Vmin.

NTV design techniques

We designed an IA processor with the capability to operate at NTV [1] (Figure 1). It is implemented in a 32-nm high-K/metal-gate CMOS process. The Pentium core occupies 2-mm^2 area and contains 6M transistors. The logic core is fully synthesized. The first level cache arrays are custom designed. The logic and arrays are on separate voltage domains to ensure that the Vmin of the array does not limit the Vmin of the logic core which has higher activity and consumes a larger portion of the total power. In general, for a complex SoC containing many different circuit types, it is difficult to scale all circuits to the same low voltage without incurring too much area overhead. Thus, we need to adopt a "divide-and-conquer" strategy via multivoltage design. The SoC can be partitioned into multiple independent voltage domains, where circuits of the same type (such as SRAM, register file, flip-flops, logic gates, etc.) and similar population are grouped in the same voltage domain. Circuit groups that are located far apart on the die are placed in different voltage domains to allow mitigation of within-die Vmin variations. This approach enables scaling of the voltage of the entire die without being limited by the Vmin of a specific circuit type, and thus helps reduce overall energy consumption of the SoC without too much overhead.

As voltage is reduced, noise margin degrades and circuits become more prone to functional failure due to noise. Circuits with large fan-in are more vulnerable to failure at low voltages due to worse ON-current to OFF-current (Ion/Ioff) ratio compared to low fan-in circuits. So, fan-in needs to be limited to two ultralow-voltage operation (Figure 2). A 4:1 mux is implemented as cascade of 2:1 muxes to improve voltage scalability. Similarly, wide logic gates such as NAND or NOR should also be split into cascades of 2-fanin gates.

At low voltages, Ion of stacked transistors degrades faster than a single transistor. Taller stacks degrade more severely due to lower Vg's and body-effect-induced Vt increases. This can lead to functional failures for reduced Ion/Ioff or more severe frequency rolloff at low voltages. For ultralow-voltage designs, max stack height should be limited to 2.



Technology	32nm High-K Metal Gate
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	6 Million (Core)
Core Area	2mm^2

Figure 1. NTV IA processor in 32-nm CMOS.

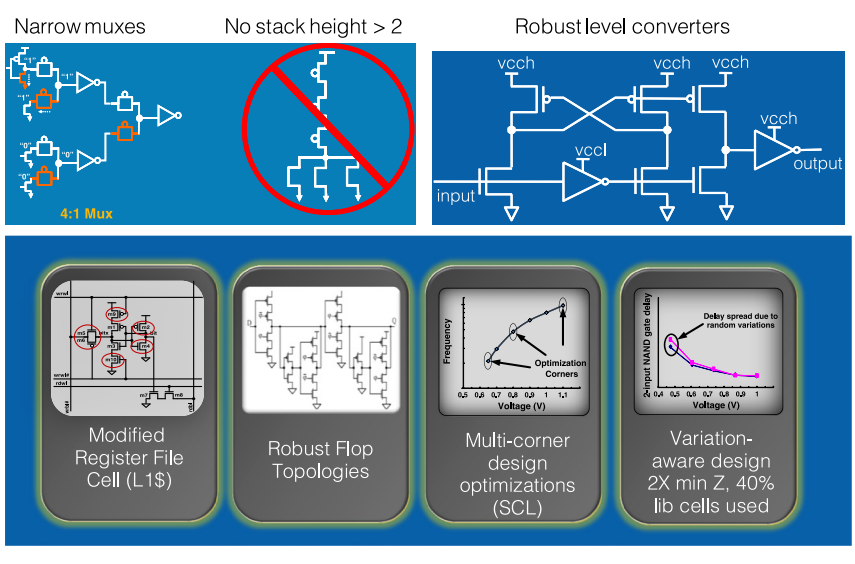


Figure 2. NTV design techniques.

Low- to high-voltage level converters can slow down dramatically and fail functionally if the input voltage goes below certain limits. In cascode-voltage-switch-logic (CVSL) converters, the key limiter for input voltage is the contention from the pull-up PMOS. As input voltage (Vin) reduces, the pull-down NMOS becomes weaker but the pull-up PMOS remains strong. Also, direct loading of the cross-coupled node by output capacitance can degrade speed severely. The split-output topology shown in Figure 2 significantly improves output drive while allowing better rationing of the pull-up PMOS and pull-down NMOS in the cross-coupled circuit topology for more robust ultralow-voltage operation.

We use fully interrupted single-ended write with transmission gate write access for register file cells to eliminate all contentions during a write operation and thus better scalability to low voltages. The flip-flops are designed to be robust at very low voltages by inserting an inverter between the master latch and the slave latch that prevents faulty write-back from the slave to the master at low voltages.

We developed a multivoltage performance verification (PV) methodology that accounts for the wide dynamic operating range in voltage and frequency. If the timing and transistor sizing is done only at the nominal operating voltage, then several problems arise for low-voltage performance. First, transistor delays dominate at low voltage over interconnect RC delays. Second, the gate delay dependence on device parameters such as threshold voltage is extremely nonlinear at low voltages since the transistor drive current in the vicinity of threshold is exponentially dependent on threshold voltage. Therefore, device parameter variations impact critical path delays very differently at low voltages. While some averaging of delay variations along a path due to random parameter variations happens at nominal voltage, parameter variation of a single gate in the path can dominate path delay at low voltage and serve as a "choke" point. These effects are accounted for carefully to fix critical paths across the entire voltage range, especially when considering within-die parameter variations. In addition, clock edges and hold times of sequentials degrade severely at low voltages. The degree of clock skews caused by process variations are also much larger at low voltages. This is because gate delays are more sensitive to device parameter variations at lower voltages. Therefore, probability of min-delay failure due to race conditions is dramatically higher at low voltages and can limit Vmin. The design methodology carefully accounts for the aggravated vulnerability to min-delay failures at NTV. As an additional precaution, usage of minimum width transistors is avoided across the entire design to limit the magnitude of device parameter variations whose impact is amplified exponentially at low voltage.

All the Vmin improvement techniques cost 10%–20% in area and cause switched capacitance (Cdyn) to increase. This impacts maximum power and max frequency at nominal voltage to some extent. The NTV design techniques need to be used judiciously to carefully balance the overheads of voltage scalability against the adverse impacts on maximum power performance across a wide range.

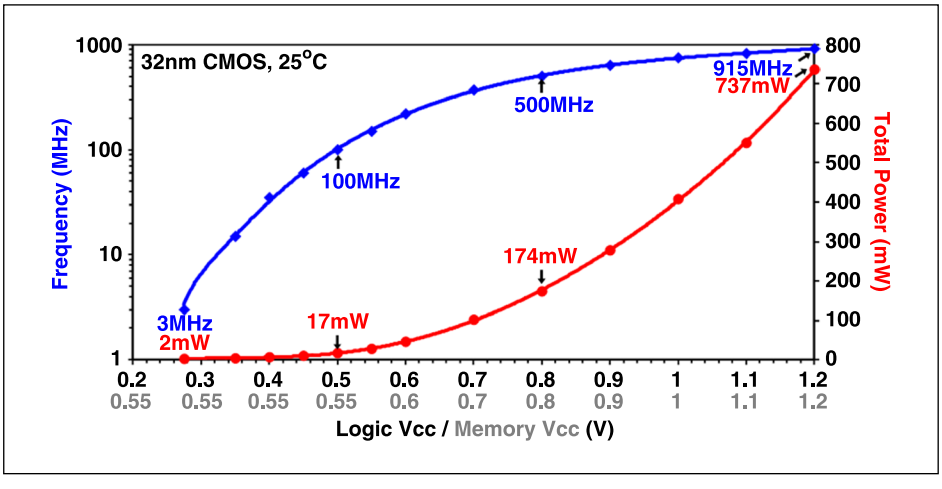


Figure 3. Power performance of the NTV IA processor in 32-nm CMOS.

Chip measurements

The chip operates from 3 to 915 MHz (Figure 3). The power ranges from 2 to 737 mW. The level 1 cache array cannot go below 0.55 V, but the logic core operates all the way down to 280 mV. This is just for a typical die. The minimum operating voltages will change if a large number of dies are measured. Note that by separating the cache and

core logic voltage domains we are able to achieve much better energy efficiency and a wider power-performance range. If they share the same voltage domain, we cannot go below 0.55 V for the whole design.

The minimum operating energy is achieved at a voltage where switching and leakage components are in balance (Figure 4). Energy per operation at the NTV operating point of 450 mV is five times better, and the frequency is ten times slower.

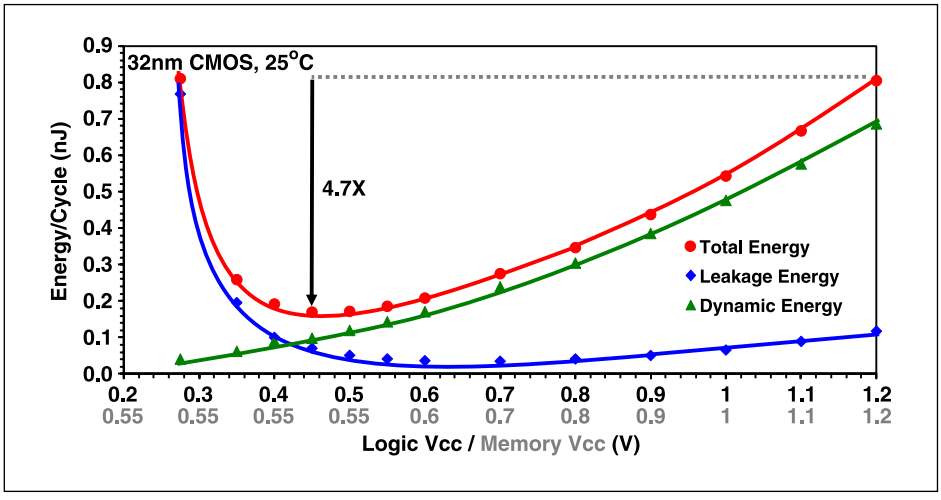


Figure 4. NTV operation of the IA processor in 32-nm CMOS.

This power-performance range can be used in many different ways depending on workload. For performance-critical, latency-sensitive workloads, the higher voltage/frequency operating points should be used. We then consume whatever energy is needed to meet the performance goal. The high voltage/frequency operating point is important not only for providing fast response on demand, but also for maximizing overall energy efficiency by sometimes completing a job quickly and then shutting down completely until the next job arrives. This "race-to-halt" mechanism is sometimes more energy efficient than running slower at low voltage for the same overall active/idle duration. When the performance need is well below the NTV operating point, then we should operate at the minimum energy point and then shut down completely and wake up when the next workload arrives. When we are simply waiting for the next workload to arrive but do not have enough time to shut off completely and wake up on time, we can park in the minimum-energy or NTV operating point, then quickly ramp up to the desired operating point when we need to restart computing. For highly parallel, throughput-oriented workloads, we can use multiple units in parallel operating at the minimum energy point to deliver the same throughput as a single unit operating at the highest voltage/frequency. This costs us die area, but allows us to achieve the best energy efficiency for a target throughput for parallel workloads, limited primarily by the energy consumed by inter-unit communications. We can thus exploit the advantages of continued Moore's law scaling which enables more affordable transistors in smaller die areas for major improvements in energy efficiency for throughput-oriented parallel workloads. A design that supports a wide dynamic voltage/frequency range including NTV allows us the flexibility to deliver the performance on demand with the highest possible energy efficiency.

NTV and variability

The NTV operating point depends on a variety of factors as shown for different process skews—slow, typical, and fast—for the same workload (Figure 5).

The slowest skew achieves the lowest voltage and smallest energy. But the max frequency is 30% slower at the NTV operating point. So, we need more units in parallel to achieve the target throughput.

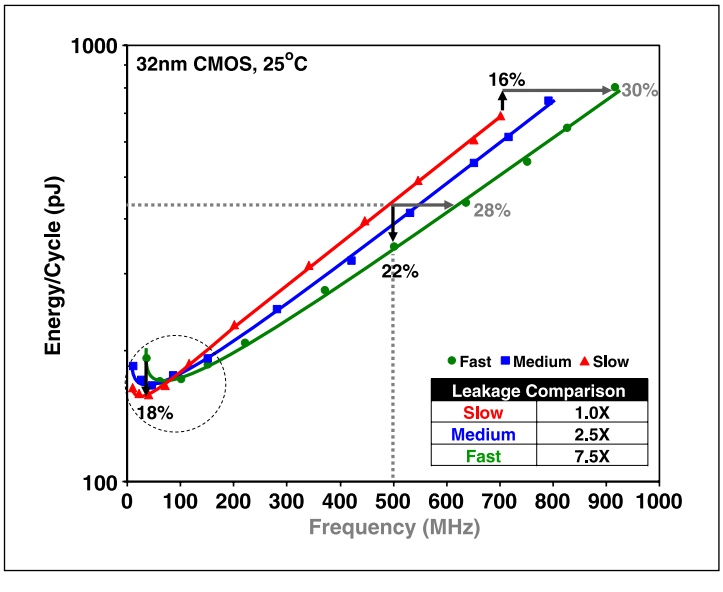


Figure 5. NTV operation for different process skews.

Process	22nm CMOS
Nominal Vcc	0.9V
Permute Xbar	256b byte-wise any-to-any
Register File	32x256b 3R/1W
Die Area	0.048mm^2
Pad Count	30

Figure 6. NTV SIMD engine in 22-nm trigate CMOS.

Thus, the process target must be chosen carefully to maximize energy efficiency while not consuming too much die area to meet the throughput targets. We need to account for variation of NTV operating points across units within the same die, and also from die to die. We need to develop core mapping techniques, testing, and die binning techniques to address these variations. The effective switching activity factor also impacts the minimum energy operating point. So, it depends on the workload and should be chosen properly at runtime depending on workload. Lower activity factor shifts NTV operating point to the right as leakage starts dominating earlier. It is critical to turn off all components that are not active to achieve this minimum energy operation. Inactive components contribute leakage power and shift the minimum energy operating point to the right, and degrade energy efficiency at the same time. So, fine-grain power management is absolutely critical for achieving minimum energy operation. Finally, in considering the right deployment of Vmin improvement techniques, we must consider the range of minimum energy operating voltages across dies and workloads. We need not push Vmin for most of the design below the minimum of this range, except for those parts of the design that need to be in "parked" mode for a fixed time for quickly waking up the rest of the design. Pushing down Vmin for the majority of the design too far costs additional area and switched capacitance and compromises the high-performance operating regime.

NTV SIMD engine

We have designed a reconfigurable single-instruction–multiple-data (SIMD) vector permutation engine with 2-D shuffle for multimedia, graphics, and signal processing workloadsin22-nmtrigateCMOS[2] (Figure 6). Since trigate or FinFET devices offer steeper subthreshold swing and better short-channel effects, they offer better variability and energy efficiency at NTV. The engine performs byte-wise any-to-any 2-D shuffle across a 32 × 32 matrix. The register file with three read ports and one write port is used for vertical shuffle. The permute crossbar is used for horizontal shuffle.

Low-voltage register files, flip-flops, and logic

The register file uses dual-ended write with transmission gates for access devices (Figure 7). This provides cell-level "symmetry" and transistor-level "redundancy"

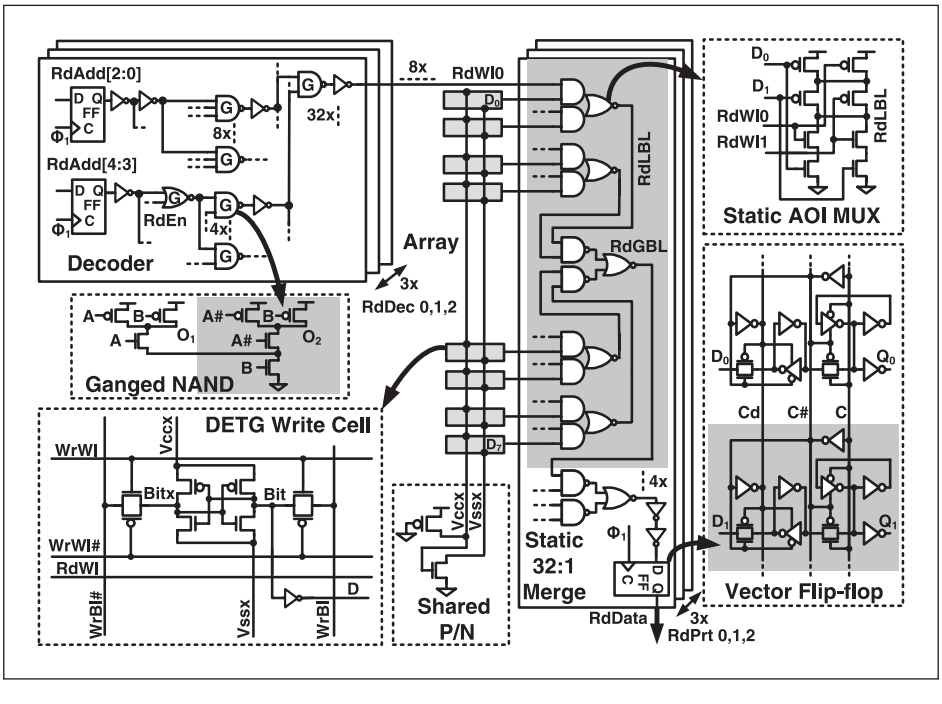


Figure 7. Low voltage register file and flip-flops.

that help mitigate impacts of device parameter variations and improve minimum write voltage. Also, we use a PMOS in cell supply and a NMOS in the ground for a group of cells in the same row. During write, the extra current in the supply/ground induces a local voltage droop in the cell. This weakens the cross-coupled inverters and helps reduce contention with the transmission gate. This eases the write operation at low voltages. The read operation is fully static, and thus more robust at low voltages than pre-charge/evaluate read.

The clock edge degrades severely at low voltages, especially since the local clock drivers in the flip-flops are very small and prone to process variations. This can cause hold time degradations and min delay failures. We use vector flip-flops where the clock outputs of local drivers are shared across the flip-flops. This helps average the impact of device parameter variations.

In the permute crossbar (Figure 8), we use fully interrupted voltage level translators to allow operation at very low voltages. Also, devices are shared across multiple gates to reduce impacts of parameter variations.

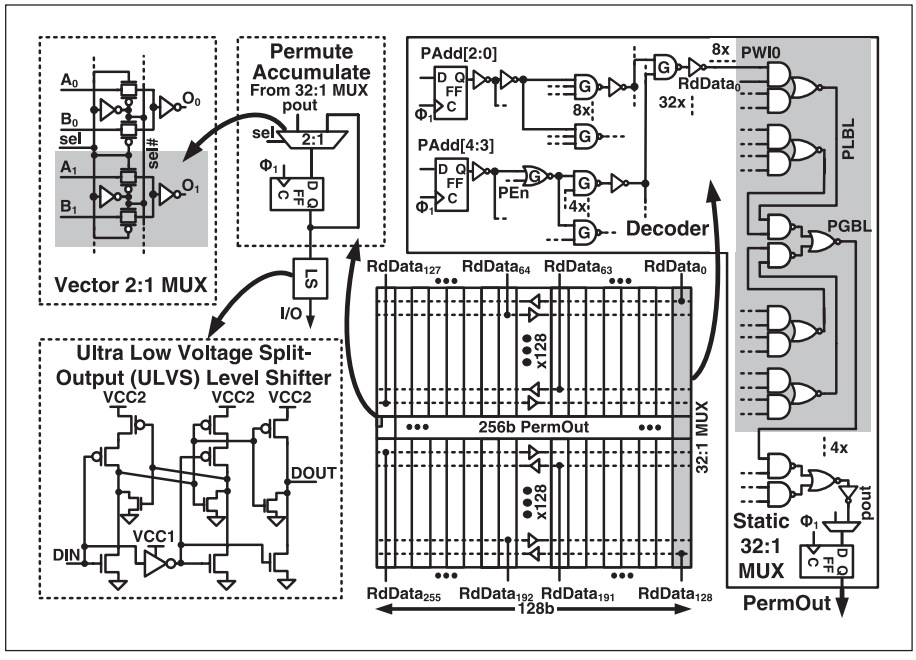


Figure 8. Low-voltage crossbar logic.

Simulations and measurements

Read and write Vmin improvements for the register file across a range of device parameter variations are shown in Figure 9. For 6 sigma, static read improves read Vmin by 200 mV. The DETG cell with shared P/N improves write Vmin by 275 mV. The combination of these techniques improves the overall Vmin of the register file by 250 mV. Originally it was limited by write, and now it is limited by read. The logic Vmin improves by 150 mV, using fully interrupted level shifters, shared gates, and vector flip-flops.

Chip measurements show that the register file and crossbar operate from 3 GHz down to 10 MHz (Figure 10). The energy efficiency in GOPS/W peaks at 280 mV, nine times higher than the efficiency at nominal voltage.

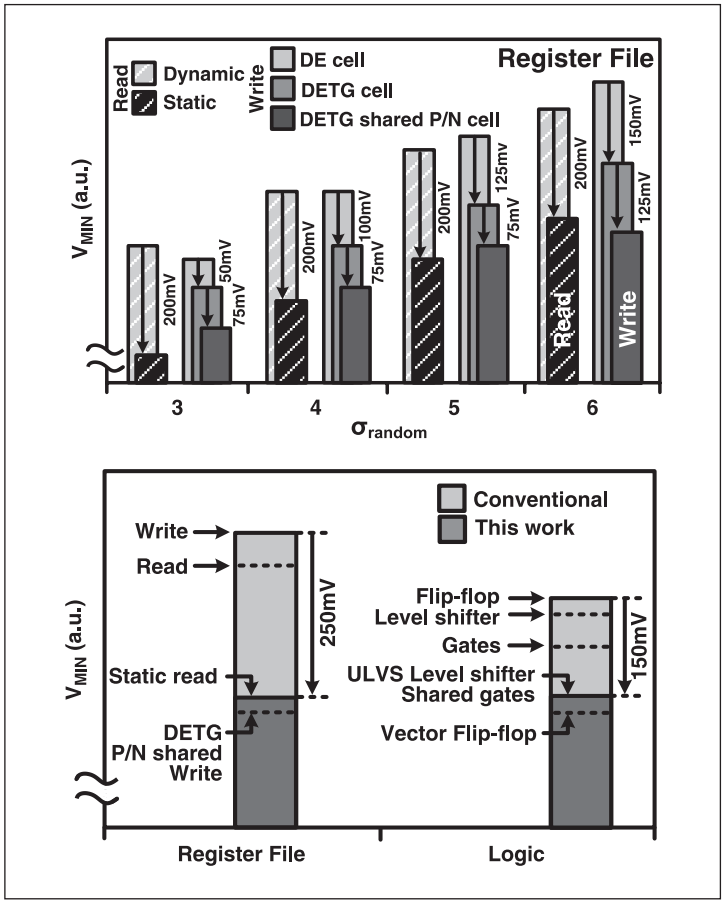


Figure 9. Vmin improvements of register files, flip-flops, and logic.

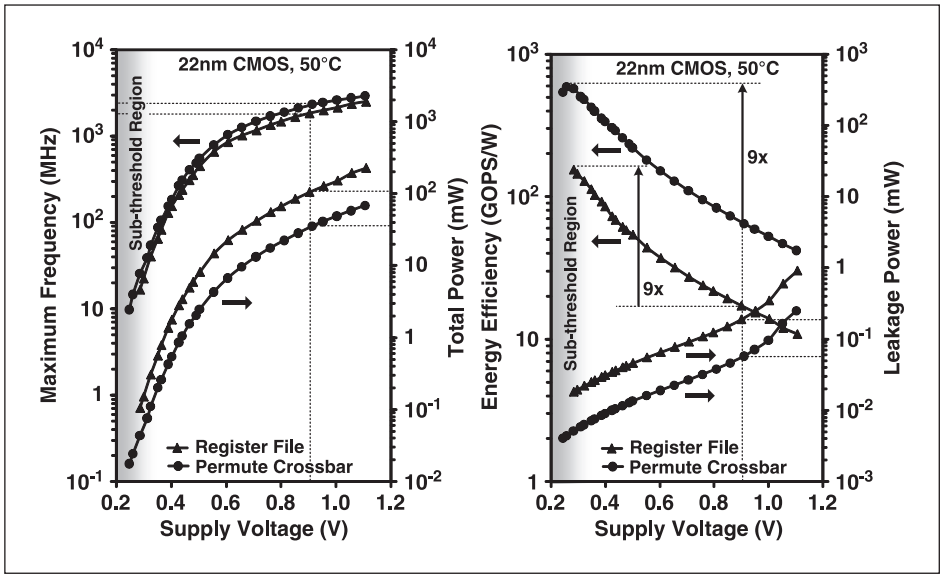


Figure 10. NTV SIMD engine power and performance.

NTV COMPUTING with wide dynamic range provides the flexibility to provide the performance on demand for a variety of workloads while minimizing energy consumption. It allows us to exploit the advantages of continued Moore's law to provide highest energy efficiency for throughput-oriented parallel workloads without compromising performance. The overheads of NTV design techniques in complex SoCs must be carefully balanced against impacts on power performance at the higher end of the operating regime. Novel circuit techniques, multivoltage designs, and variation-aware design methodologies are essential for realizing robust NTV SoCs in scaled CMOS process nodes. ■

References

[1] S. Jain et al., "A 280 mV-to-1.2 V wide-operating-range IA-32 processor in 32 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.* 2012, pp. 66–67.

[2] S. Hsu et al., "A 280 mV-1.1 V reconfigurable SIMD vector permutation engine with 2-dimesnsional shuffle in 22 nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.* 2012, pp. 178–180.

Vivek De is an Intel Fellow and Director of Circuit Technology Research at Intel Labs, Hillsboro, OR, USA. He is responsible for providing strategic technical directions for long-term research in future circuit technologies and leading energy-efficiency research across the hardware stack. De has a PhD in electrical engineering from Rensselaer Polytechnic Institute, Troy, NY, USA. He is a Fellow of the IEEE.

Sriram Vangal is a Principal Engineer at Intel Labs, Hillsboro, OR, USA. He leads energy-efficient computing research. Vangal has a PhD in electrical engineering from Linkoping University, Linkoping, Sweden. He is a Senior Member of the IEEE.

Ram Krishnamurthy is a Senior Principal Engineer at Intel Labs, Hillsboro, OR, USA. He leads high-performance circuits research. Krishnamurthy has a PhD in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA. He is a Fellow of the IEEE.

■ Direct questions and comments about this article to Vivek De, Intel Corporation, Hillsboro, OR 97124 USA; vivek.de@intel.com