A 9-mm^2 Ultra-Low-Power Highly Integrated 28-nm CMOS SoC for Internet of Things

Yu Pu, Chunlei Shi, *Member, IEEE,* Giby Samson, Dongkyu Park, Ken Easton, Rudy Beraha, Adam Newham, Mark Lin, Venkat Rangan, Karam Chatha, Danny Butterfield, *Member, IEEE,* and Rashid Attar

***Abstract*— This paper gives an overview of the Blackghost 1.0 system-on-chip (SoC) from Qualcomm Research, which was our first test chip that paved the way toward the commercialization of Qualcomm's most recent ultra-low-power Blackghost SoC family. Specifically designed for battery powered Internet of Things, sensor fusion, wearables, and e-medical applications, this highly integrated SoC delivers high-power efficiency through low-power innovations in architecture and circuit domains. It integrates a small footprint sensor control processor based on ARM Cortex-M0, a vision classifier processor, a streaming DSP hardware accelerator, an ultra-low-power analog-front-end, and an on-die power management unit with direct Li-Ion battery attach capability. The die size of this prototype chip is 3 x 3 mm^2 in a 28LP CMOS process technology. To date, this SoC family has successfully progressed to the mass production of near-threshold computing. The logic computation operates at near-threshold voltages (<0.6 V) at frequencies up to 50 MHz and draws less than 9μ A/MHz from the directly attached battery.**

***Index Terms*— Accelerator, Internet of Things (IoT), near threshold, power management, sensor fusion, system on chip (SoC), ultra-low power.**

I. Introduction

INTERNET OF THINGS (IoT), sensor fusion, and e-healthcare have been largely fueling the wave of semiconductor innovations. It is predicted that by the year of 2025 there will be about 27 billion connected devices potentially generating up to U.S. $3 trillion in overall aggregate revenue. While the demand for low-cost microcontroller (MCU) solutions remains strong, the embedded system-on-chip (SoC) market trend is continuously shifting from "sense making" to "decision making," which requires devices with augmented intelligence and computation capability. A highly integrated and power-efficient SoC, which contains sensor-hub controller, DSP hardware accelerator(s), analog-front-end (AFE) for context awareness, and embedded power management unit (PMU), is therefore desired.

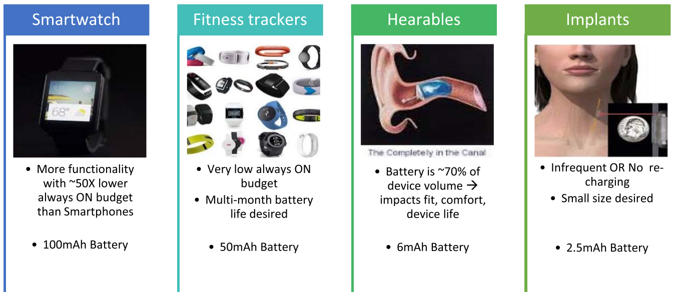

Smartwatch
Fitness trackers
Hearables
Implants
68°
• More functionality with ~50X lower always ON budget than Smartphones
• 100mAh Battery
• Very low always ON budget
• Multi-month battery life desired
• 50mAh Battery
The Completely in the Canal
• Battery is ~70% of device volume → impacts fit, comfort, device life
• 6mAh Battery
• Infrequent OR No re-charging
• Small size desired
• 2.5mAh Battery

Fig. 1. Low-power use cases of *Blackghost* SoC family.

To offer rich functionalities under a stringent power budget, in Qualcomm Research, we launched the research of *Blackghost* SoC in 2014. The code name *Blackghost* comes from the name of the world's most energy-efficient fish that lives in Amazonia Rivers. This fish emits weak "always-ON (AON)" electric field to sense objects in the deep dark water using its "electric headlamp." We envision this SoC improving the battery life of "AON" emerging applications, which are categorized by the typical battery volumes in Fig. 1. To achieve this goal, meticulous low-power considerations were given to system, architecture, circuit, process technology, and design methodology. To date, we have taped-out a series of test chips and have successfully progressed to the mass production phase. This paper gives an overview of the *Blackghost* 1.0, which was the first test chip that marked the milestone for many proofs of concept.

We take the opportunity of writing this paper to revisit the road we have traveled. We hope this reflection would incite more research in highly integrated ultra-low-power systems of such kind. The rest of this paper is organized as follows. First, we will briefly describe the architecture and the key building blocks of the SoC. Second, we will outline the effort in enabling NTC, with emphasis on logic circuit and PMU with direct Li-Ion battery attach. Third, we will show the custom low-power SRAM (LP-SRAM) techniques. After which, we will present some measurement results. Finally, we will draw conclusions from this low-power platform.

II. Architecture

Fig. 2 shows the overall SoC architecture. The main building blocks include a sensor control processor, a streaming DSP hardware accelerator for applications like voice keyword detection, a vision classifier hardware accelerator for

Manuscript received August 1, 2017; revised October 12, 2017, November 12, 2017, and December 1, 2017; accepted December 1, 2017. Date of publication January 1, 2018; date of current version February 21, 2018. This paper was approved by Associate Editor Marian Verhelst. *(Corresponding author: Yu Pu.)*

The authors are with Qualcomm Research, San Diego, CA 92121 USA (e-mail: ypu@qti.qualcomm.com; cshi@qti.qualcomm.com; gsamson@qti.qualcomm.com; dongkyup@qti.qualcomm.com; keaston@qti.qualcomm.com; rberaha@qti.qualcomm.com; anewham@qti.qualcomm.com; mlin@qti.qualcomm.com; vrangan@qti.qualcomm.com; kchatha@qti.qualcomm.com; dkb@qti.qualcomm.com; rattar@qti.qualcomm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2017.2783680

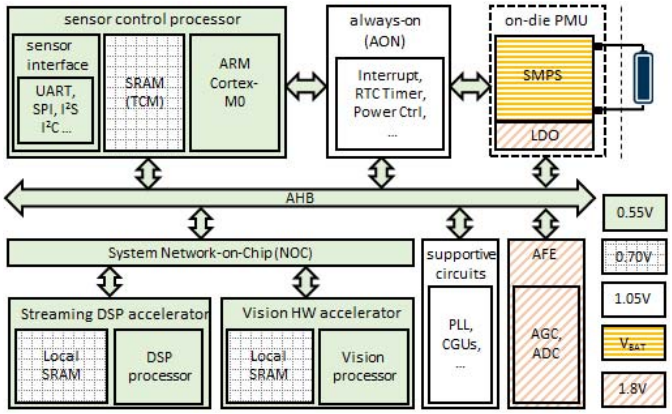


Fig. 2. SoC diagram overview.

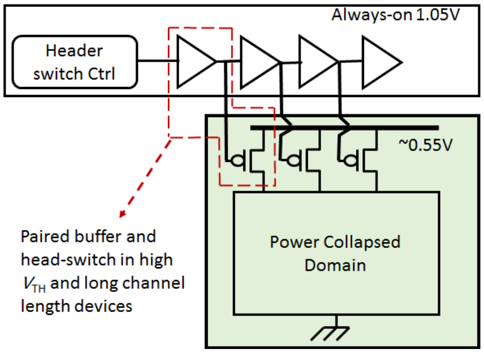

Always-on 1.05V
Header switch Ctrl
~0.55V
Paired buffer and head-switch in high V_TH and long channel length devices
Power Collapsed Domain

Fig. 3. "Super-cutoff" power gating.

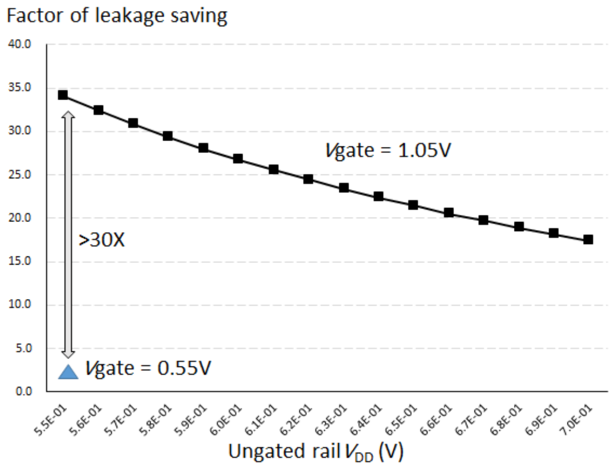


Fig. 4. Simulated leakage saving through "super-cutoff" power gating.

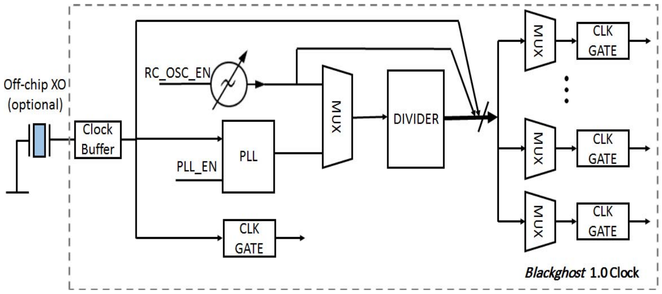


Fig. 5. Simplified clock diagram.

applications like face recognition and motion tracking, AFE, a high efficiency PMU and other supportive components such as phase-locked loop (PLL) and clock generation units. The sensor control processor mainly comprises of the sensor interface (i.e., FIFOs and hardened ASICs that implement sensor algorithms), an ARM Cortex-M0 processor and its tightly coupled memory (TCM). The incoming data from external sensors is buffered in the FIFOs. The FIFOs wake up the sensor control processor when the amount of stored data reaches a software-defined threshold. The sensor control processor then (pre-) processes the data and decides whether it is necessary to wake up the hardware accelerators. This "event-driven" wakeup sequence enables the power gating of most parts of the chip in ambient mode (ambient mode is a mode often used in high-end Android Wear AON apps. In contrast to interactive mode, ambient mode refers to the energy saver mode by only using co-processor, while the main App CPU is not activated). The SoC has multiple SPI, I^2C and UART peripheral ports and abundant GPIOs to support various sensor interfaces.

A low-frequency (32 kHz–50 MHz), tiny (a few hundred logic gates), and 1.05-V AON block handles mission-critical functions like interrupt, wakeup and system power state machine. To minimize the leakage, the AON domain is implemented with logic gates in high V_{TH} and long channel length devices. Used as sensor peripherals, the AFE includes: 1) a low-noise sense amplifier buffer that has a sufficiently high and tunable amplification gain and2) a 12-bit low-power fully differential successive approximation ADC. The AFE reduces the bill of material for applications like bio-signal monitoring, acoustic sensing and electrical motor control. The entire AFE is implemented with thick gate-oxide transistors and squeezed in less than 0.15-mm^2 silicon area. It consumes less than 1 μA in total leakage current.

Other than the PMU, all the power domains in the SoC can be power gated. The "super-cutoff" scheme is deployed, where a negative Vgs bias is applied to the pMOS-type header-switches during power gating. As illustrated in Fig. 3, the header-switch control signals are propagated through buffer chains in the 1.05-V AON domain. The buffer and header-switch pairs are uniformly distributed in the power gated domains. Fig. 4 shows the reduction in leakage when the gate voltage of header-switch is over-driven to 1.05-V voltage, while sweeping the ungated power rail from 0.7 to 0.55 V, as compared to the conventional power-gating scheme where $Vg = Vs = 0.55$ V. As shown, the super-cutoff scheme is $>30\times$ more effective in reducing the leakage current. This ultra-low-leakage feature is especially crucial at high temperatures, where leakage currents skyrocket. Furthermore, the paired buffers and header-switches are implemented with long channel length transistors for leakage minimization.

A simplified clock diagram is shown in Fig. 5. The SoC provides three different clock sources that are software selectable for users.

1) An internal *RC* relaxation oscillator. Based on the results from the on-die process sensor and temperature sensor, it uses programmable codes to tune the R and C

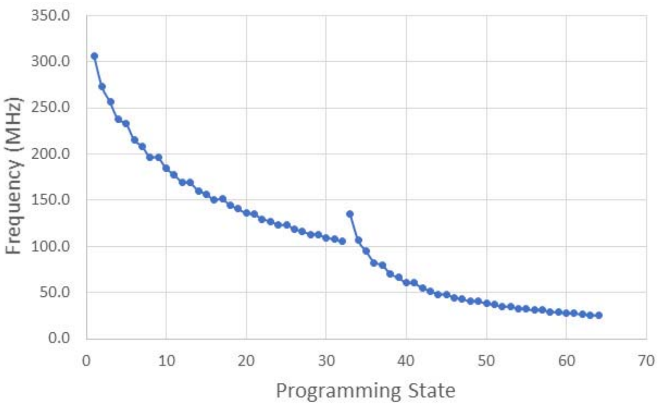


Fig. 6. *RC* oscillator: frequency at different program states.

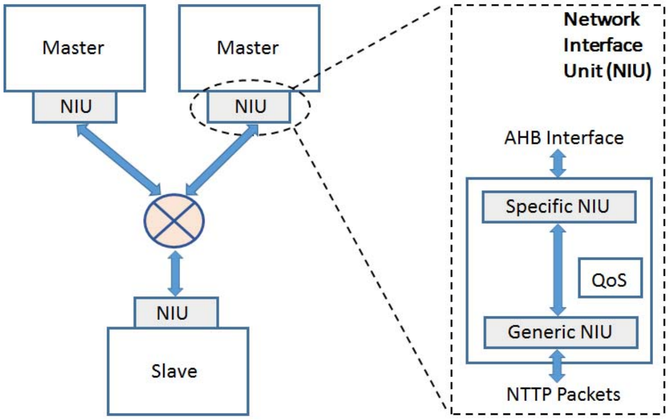


Fig. 7. Simplified NoC diagram.

against corner and temperature through the close-loop calibration at system level. As shown in Fig. 6, the oscillator covers a frequency range from 25 MHz up to 300 MHz. If fast wake-up time or ultra-low-power is more preferred than clock accuracy, using *RC* oscillator is preferred. At 50 MHz of clock frequency, it consumes about 20 μW of power.

2) An optional off-chip crystal oscillator (XO) for low-speed operation.
3) An on-chip PLL, if high frequency or high accuracy is required.

The on-chip communication and data transfer between the building blocks are through advanced high-performance bus (AHB) and Network-on-Chip (NoC). A simplified NoC diagram is illustrated in Fig. 7. The network interface unit converts AHB transactions to *NoC Transaction and Transport Protocol* packets, which are then ferried to different sections of the chip through a series of switches and links. Different than busses that fan out the wires to all the peripherals, the links between NoC units are point-to-point connections, which reduce the total number and distance of global interconnections, thereby lowering the associated capacitance load switched per transaction. Also, the power wasted in bus retry cycles do not exist in the NoC. Fair comparisons between the bus and NoC require actual layout experiments, but our experience is that dynamic power consumption of a large SoC is often lower for the NoC than for the busses at equivalent system performance. More importantly, the adoption of NoC reserves high flexibility and scalability for the future expansion of the SoC platform. It also relieves the stress from the timing closure of the synchronous ultra-low-voltage building blocks, as will be discussed later in this paper.

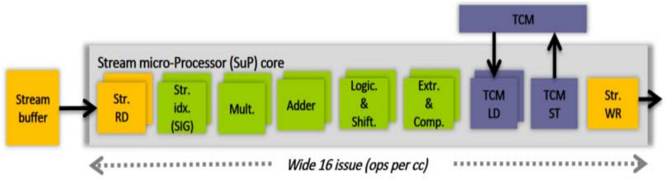


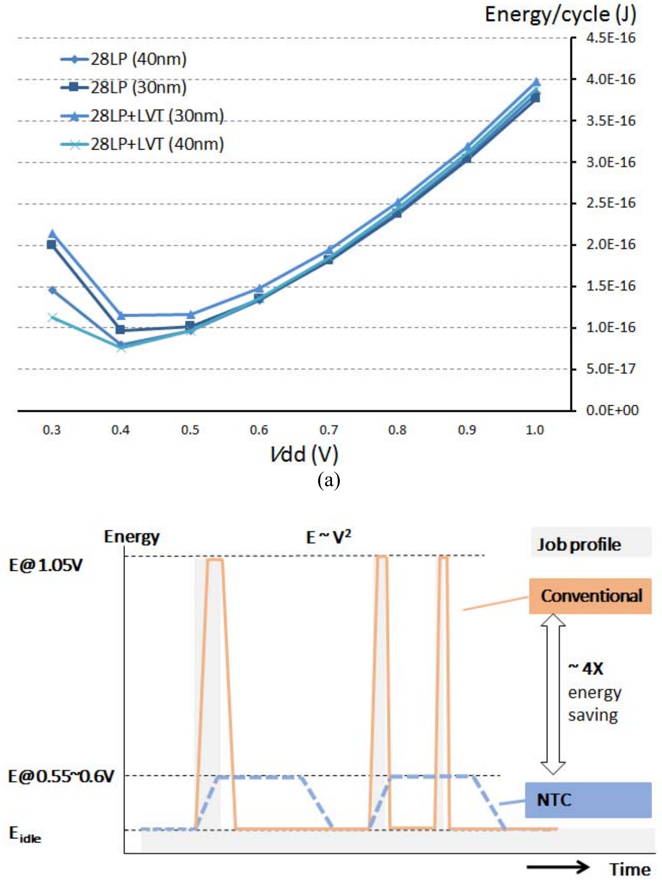
Fig. 8. Wide issue of 16 instructions/cycle of *SuP*.

The DSP hardware accelerator is the *Stream Micro Processor* (*SuP*), which is an in-house architecture developed in Qualcomm Research. *SuP* accelerates DSP processing for streaming applications. It is a 32-b dataflow accelerator architecture that supports 8-/16-/32-b instructions and concurrent execution of up to 16 instructions. Different from classical instruction processor architectures that follow fetch-decode-execute instruction cycle, *SuP*: 1) loads a signal processing kernel once with all its instructions into the local instruction store; 2) the kernel is executed many times, which is typical for many signal processing workloads; and 3) on completion the next kernel is loaded on the *SuP*. The *SuP* has a rich instruction set and can execute the full range of DSP kernels including (and not limited to) fast fourier transform (FFT), Gaussian mixture model, and Semi Markov model. The wide-issue (of up to 16 instructions/cycle) of *SuP* is illustrated in Fig. 8. As the *SuP* is an accelerator architecture, it is lower power than a comparable classical instruction processor, and an effective high-performance off-load engine for the Cortex-M0 control processor. Thus, heavy DSP tasks are mainly done using *SuP*, whereas the control/glue code runs on Cortex-M0. Existing MCU + DSP-based solutions, such as ARM Cortex-M4 [1] and Samsung Bio-Processor [2], deliver 1.25 DMIPS/MHz and 3.4 Coremarks/MHz. *SuP* provides more than four DMIPS/MHz and seven Coremarks/MHz. For control code energy efficiency, the Cortex-M4 and Cortex-M0 are very similar, but for signal processing code energy efficiency *SuP* outperforms Cortex-M4. This implies that: 1) if we run at the same frequency, *SuP* halves the active duty cycle as compared to Cortex-M4 and 2) *SuP* allows more aggressive voltage scaling to meet the same throughput. Similarly, the vision processor accelerates vector operations for energy-efficient object detection. For use cases like voice keyword detection and vision object detection, the combination of *SuP* and vision processors leads to more than 2× lower power over existing solutions.

III. Enablement of Near-Threshold Computing

After careful evaluation, finally TSMC 28LP CMOS process technology was selected for the SoC implementation, mainly because of the following.

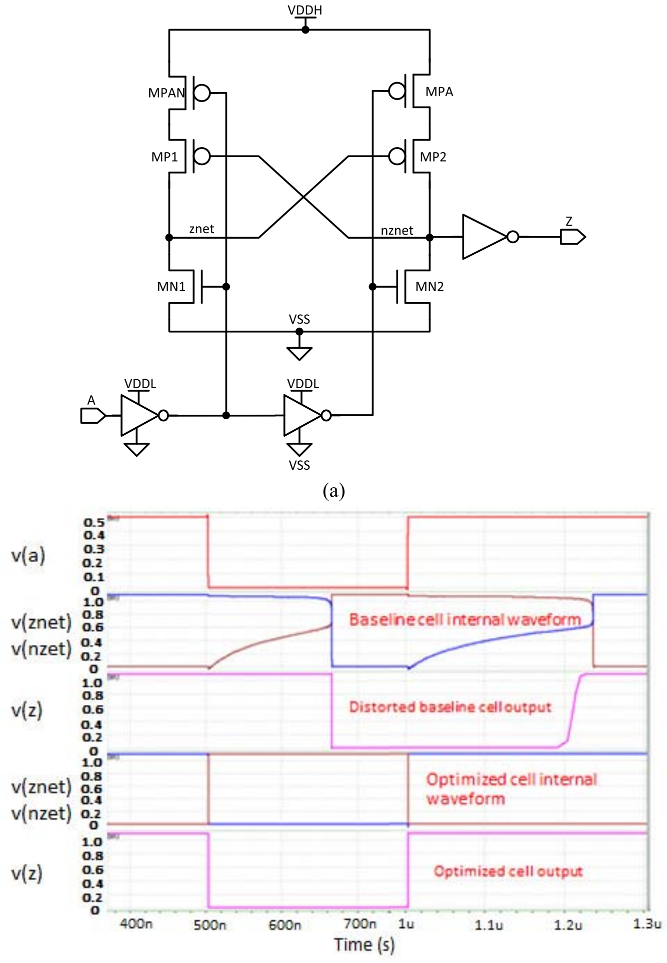
1) It gives good balance in performance, active, and leakage power for sensor applications. Advanced process nodes,



(a)

(b)

Fig. 9. (a) Optimal energy point for benchmark circuit with different V_{TH} flavors and channel lengths. (b) Illustration of NTC and turbo active modes.



(a)

(b)

Fig. 10. Optimized level shifter for wide voltage conversion from NTC power domains to AON power domain.

such as 14-/10-nm FinFET processes, could largely improve performance and reduce dynamic power, but their excessively high leakage power is fatal to our use cases.

2) A 28-nm node is a mature process node with well-controlled yield window. Such low process variation is highly appreciated for ultra-low-voltage circuits and systems.

3) Triple-well process with deep n-well (DNW) is a must for power transistor isolation inside the PMU.

4) It holds the possibility of integrating a rich set of cost effective RF connectivity IPs [e.g., Bluetooth Low-Energy (BLE), Wi-Fi] and embedded non-volatile memory (e.g., eFlash and MRAM). Recently, foundries have made significant progress in advancing ultra-low-voltage process technologies, such as ULP Bulk-CMOS and FD-SOI. For the future generations of such ultra-low-power platform, we must evaluate the potential benefits of the new process offerings.

Sensor applications typically have low duty cycles. They exhibit "burst" characteristics—most of the time they only require medium or low frequency, while some of them infrequently require high performance. We took a small digital block as benchmark circuit and implemented it with various device types, i.e., combinations of different channel lengths and V_{TH} flavors. Our simulation shows that, at typical–typical (TT) process and 25 °C temperature, with the low duty cycles of our interest and regardless of the device types, a V_{DD} around 0.4 V yields the optimal energy per operation [see Fig. 9(a)]. However, the 0.4-V sub-threshold operation is too slow (<1 MHz) to complete tasks "in time" for the use cases. To benefit from aggressive V_{DD} scaling while meeting job processing deadlines, a near-threshold computing (NTC) operation mode is introduced, with transistors operating at a V_{DD} that is close to V_{TH}. The NTC mode is specified as 50-MHz computing frequency at around 0.55-V V_{DD}. In the turbo mode, the SoC reaches up to 200-MHz frequency at the nominal 1.05-V V_{DD}, at the cost of 4× energy consumption, as illustrated in Fig. 9(b). Between the NTC mode and the turbo mode, there are several active modes specified at intermediate frequencies and voltages. In addition to the active modes, the SoC has an active standby mode (clock gating), a sleep mode (partly power gating), a deep sleep mode (all power gating) and a retention mode (i.e., memory data and logic states are retained at about 0.40-V voltage, while the rest of the core is power gated). All the operation modes are easily programmed by software, so the end-users are welcome to explore the

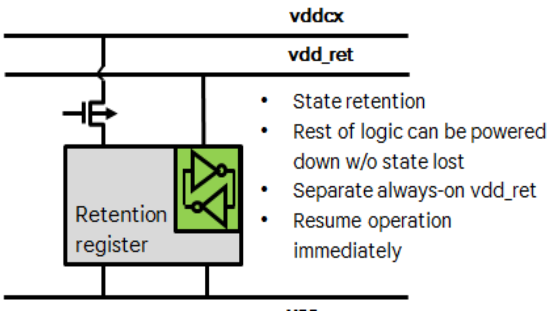

vddcx
vdd_ret
Retention register
State retention
Rest of logic can be powered down w/o state lost
Separate always-on vdd_ret
Resume operation immediately

Fig. 11. Concept of retention flip-flop.

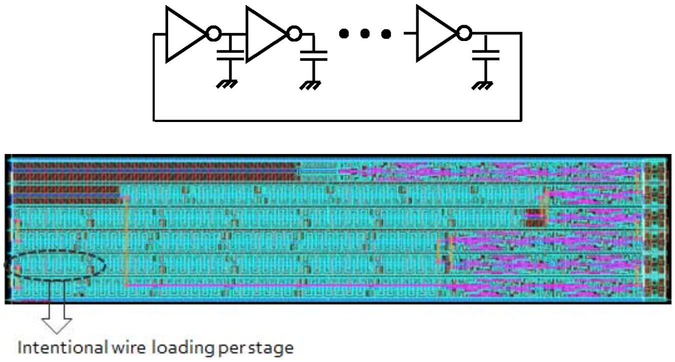


Fig. 12. Ring oscillator test structure with intentional wire loading

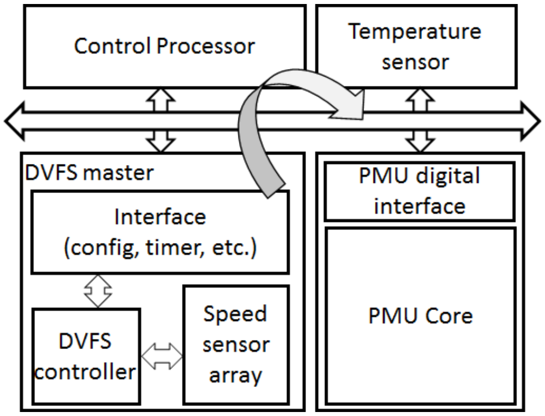


Fig. 13. High-level diagram of DVFS control.

tradeoff sweet spot between power and performance based upon their own job profiles.

Near/sub-threshold circuit has been the subject of a great deal of academic research and the results are promising (see [3]–[9]). However, it remains challenging for high volume production, due to excessively high yield loss caused by susceptibility to noise, sensitivity to temperature and variation from process drift. The SoC has carried out the following countermeasures to mitigate yield risks and design overhead.

1) Test structures, such as single transistors (with different V_{TH} flavors, channel lengths, aspect ratios and layout orientations), logic gates, oscillators and passive devices, were fabricated on design-of-experiments (DOE) wafers and characterized at ultra-low-voltages across a wide temperature range. The SPICE model was correlated and validated with the characterized wafer data.

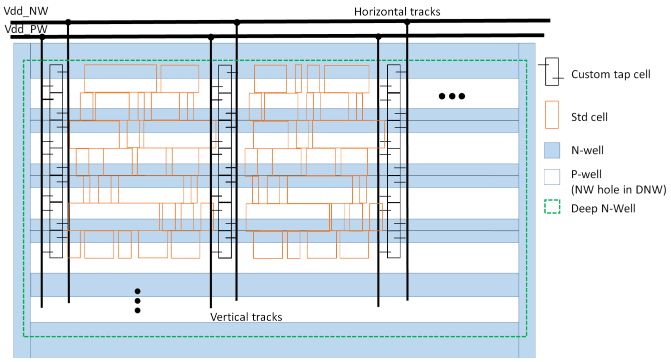
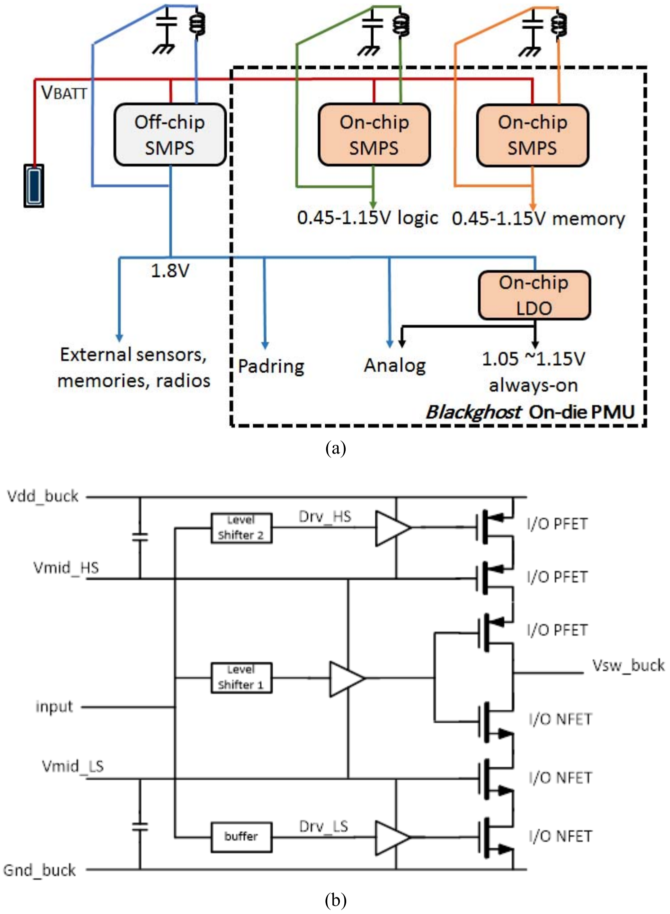


Fig. 14. Improved logic area utilization in a body-bias domain.

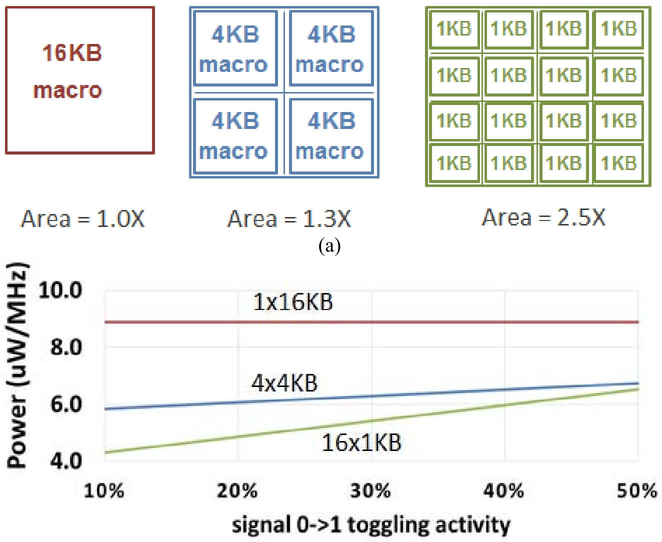


(a)

(b)

Fig. 15. (a) SoC power tree. (b) Triple-stacking SMPS output stage with interleaved timing control for direct battery attach.

2) A high-density ultra-low-voltage friendly standard cell library was developed to operate reliably at 0.5 V across all PVT corners. First, basic cell architectures (base pMPS/nMOS beta-ratio, routing grid and number of tracks) were considered. The nine-track architecture was selected, because it renders high density, low power and sufficient drive strength to meet our performance target. Second, cell library profiling was performed on the nine-track conventional standard cell library. During this step, the standard cells that show overly degraded slew and delay in their timing arcs were pruned. Restricting the usage of high V_{TH}, complex or minimum-sized logic



(a)

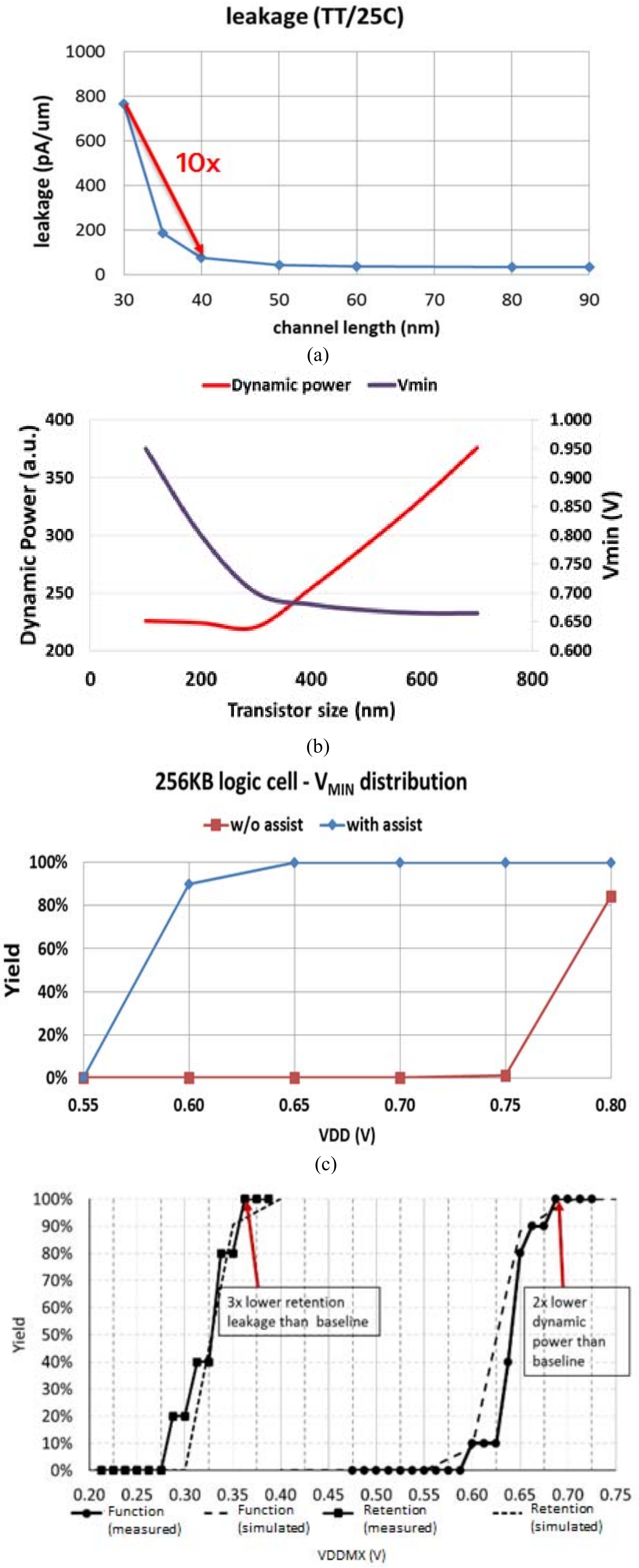
(b)

Fig. 16. Re-organizing 16-KB SRAM macro into smaller macros. (a) Concept and area penalty. (b) Switching power when data toggle.

gates (see [10], [11]) are crucial, due to their higher sensitivity to process variability. Close to 1000 cells were selected. Third, special cells like level-shifters, clamps, retention flops, de-cap cells were added. Finally, trial synthesis was performed to make sure the pruned library does not hamper design closure and optimization at both the nominal voltage and the NTC voltage. From the logic synthesis experiments, as compared to the design which is optimized with the conventional cell library at the nominal supply voltage, the design optimized with the ultra-low-voltage cell library for near-threshold operation introduces 5%–10% of area efficiency degradation.

3) For the targeted 50-MHz operation, the total power is dominated by the dynamic power. Based on the estimation, the alternative standard V_{TH} (SVT) implementation requires about 0.7-V V_{DD} in TT process corner and about 0.77-V V_{DD} in SS process corner, which imply $>50\%$ dynamic power overhead compared to the low V_{TH} (LVT) implementation. To reduce the leakage power overhead associated with LVT devices, long channel LVT devices are used for fixing hold time violation as well as leakage recovery in the functional paths. SVT devices are allowed in DFT paths. While it is commonly assumed that near-threshold design can save leakage power, contrarily, our design leaks 5–10 times more, simply because the implementation needs a much larger amount of LVT devices to meet the performance target. The increment in leakage current will be particularly significant at high temperatures, at which the exponentially increased leakage power will dominate the total power consumption. Therefore, we concluded that it was imperative to have the 0.55-V NTC domains be fully power gated using the "super-cutoff" scheme, as discussed earlier.

4) Level-shifter was customized for wide voltage conversion from 0.5-V VDDL to 1.2-V VDDH, with a few nanoseconds of A-to-Z delay. The level-shifter schematic is shown in Fig. 10(a). It has a strengthened pull-down path and a weakened pull-up path through



(a)

(b)

(c)

(d)

Fig. 17. LP-SRAM (a) bitcell leakage and channel length tradeoff, (b) dynamic power, $V_{\min}$ and bitcell transistor width tradeoff, (c) low-voltage yield improvement through word-line boosting circuit technique, and (d) simulated and measured yield at low voltages.

TABLE I

SRAM KEY METRIC COMPARISON

	Macro Size	Area	Retention Vdd (V)	64KB Retention Power (µW)	Vmin for 50MHz (V)	Macro dynamic power (µW/MHz)
Foundry	4KB	1.0X	0.70	~ 7.9	0.945	~5.5
LPSRAM	4KB	2.4X	0.40	~ 2.0	0.700	~ 3.4

transistor stacking, thereby reducing the inherent contention between the pull-up devices (MP1/MP2) and the corresponding pull-down devices (MN1/MN2). The simulation results for the level-shifter in operation are shown in Fig. 10(b).

5) Retention flip-flops were used in the power gated logic domains to store the logic state of registers, as illustrated in Fig. 11. This allows resuming operation immediately upon power domain wake-up. In the retention flip-flop, the cross-couple inverters form an "AON" latch. They are powered through the ungated power rail, implemented with low leakage devices and properly sized to achieve the lowest possible retention voltage. More than 1-M sample-based Monte-Carlo simulations provides high-sigma confidence. Mismatch factor in the statistical model is adjusted according to the process control monitoring data that is collected at the last stage of wafer fabrication process. At TT/FF/SS process conditions, the minimal retention voltage is less than 300 mV, whereas at FS/SF asymmetric process corners the retention voltage needs to be increased to about 400 mV. Then, tens of millivolt of V_{DD} margin is added on top to cover NBTI effect. Silicon tests were carried on a high volume of parts, to extract and correlate measured failure probability to statistical simulation and to identify if there are any anomalies in the retention failures. At cell level, a retention flip-flop is 40%–50% bigger than the regular flip-flop of the same drive strength, because of the extra AON latch to store the value. The SoC contains more than 1000 retention flip-flops, which is a small portion of the total number of flip-flops. For the entire chip area (including memory, I/O, and analog), the extra overhead from using retention flip-flop is less than 1%.

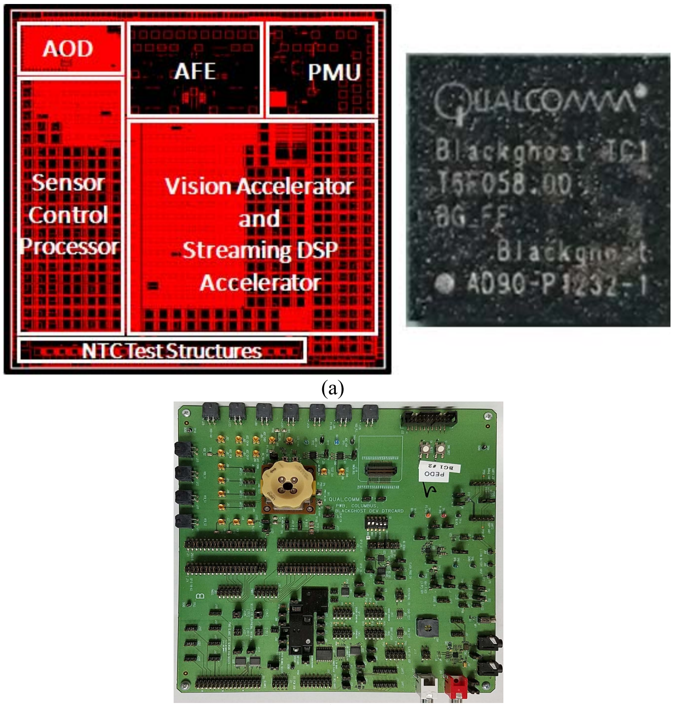
6) Guard-bands that are padded in design parameters like setup time, hold time, skew and jitter of clock tree and OCV, were inflated to account for uncertainties. On top of the flat margin and derating factors generated from Monte-Carlo analysis, at NTC voltages extra margin was adjusted to take account into account silicon miss-correlation uncertainties resulted from back-end-of-line variation, layout parameter extraction uncertainties, temperature/voltage gradients, etc. This extra margin was derived from results captured from wafer-probing test structures on the DOE wafers. The worst discrepancies were from structures with heavy metal wire loadings (an example is shown in Fig. 12). At the nominal voltage, the mismatch is typically <2%, but at NTC voltages additional it is increased by a few percent due to the combined effects.

7) Use asynchronous NoCs. Our physical design trial runs showed that, when EDA tools strived for the design closure of a fully synchronous clock domain in the NTC mode, the number of clock buffers in the clock tree and data buffers (that are needed to fix transition slew, setup time and hold time violations) went up super-linearly with the size of the clock domain. Beyond a certain size, the design encountered severe placement and route congestion followed by timing closure failure. These negative effects were further exacerbated as we had padded a little extra design margins. To ease timing sign-off, resolve routing congestion and reduce design overhead, asynchronous NoC was adopted, which turned the SoC to global asynchronous local synchronous, i.e., the clock within a regime is a synchronous clock but it is truly asynchronous to all other regimes. FIFOs and the associated control logic are instantiated in NoC for asynchronous clock crossing. The FIFOs occupy about 10%–15% of the NoC module, while control logic occupies about 9%. From area perspective, the adoption of NoC is not the best choice, but it is a reliable and scalable approach for a large-scale NTC design.

8) V_{DD} droop is especially detrimental to ultra-low-V_{DD} operation. To lower static IR drop, the power distribution network was enhanced using thick metal layers and the header-switch placement pattern was customized. De-cap cells and edge-cap macros were deployed in a higher density than conventional designs, to lower dynamic IR drop. In our simulation, the V_{DD} profile in the NTC mode is evenly distributed across the entire core, with a maximal 30-mV total droop at the worst hot spots.

9) On-die process, temperature and performance monitor sensors were embedded in the NTC power domains. The sensor readouts are used for run-time close-loop dynamic-voltage-frequency-scaling (DVFS) control. A high-level system diagram is shown in Fig. 13. The results generated by the speed sensor array is read out and re-formatted by the DVFS controller. The difference between actual and expected values determines the PMU voltage recommendation. Meanwhile, the control processor also periodically reads the temperature sensor. If the temperature goes lower than the safe boundary, the PMU will disable NTC mode and move to 0.7 V, regardless of the voltage recommendation from the DVFS master. Without activating body bias, the floor voltage is set as 0.55 V. If DVFS block suggests any voltage less than the floor voltage (i.e., for the FF samples), the PMU will not entertain the request, because in the design specification 0.55 V is specified as the lowest voltage corner that guarantees sufficient hold time margin in mass production.

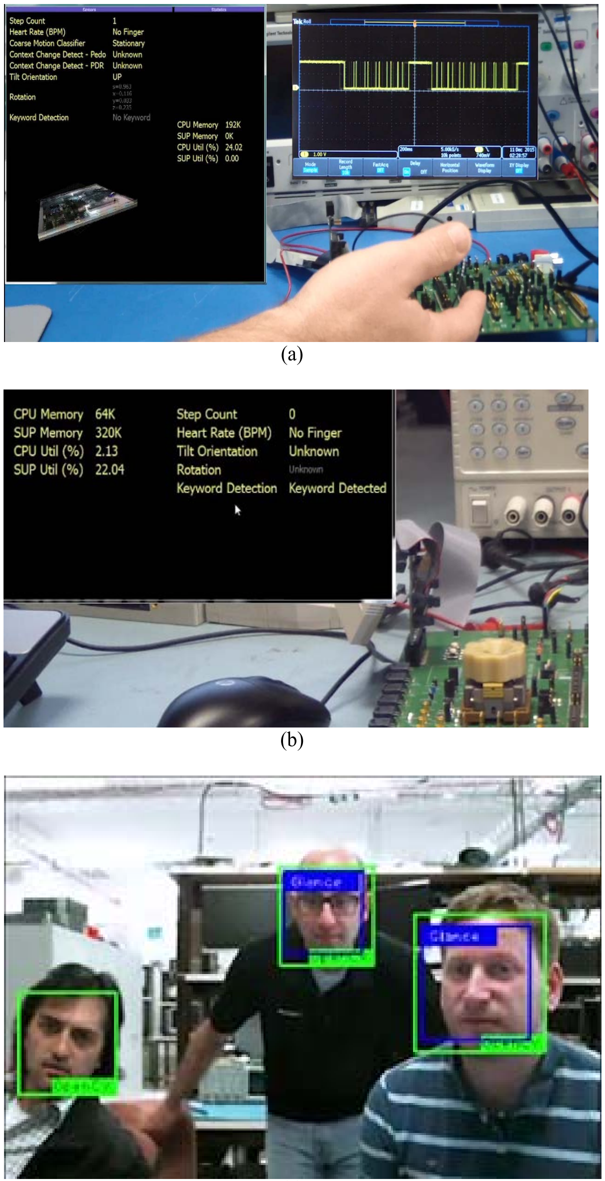
10) As an optional feature, V_{TH} tuning using body-bias has been applied to the NTC domains. The body bias control state machine resides in the AON domain, which issues body-bias enable and bypass instructions to each body-bias power domain. To implement power gating and body-bias in the same power domain, a blank keep-



(a)

(b)

Fig. 18. (a) SoC layout floorplan and the packaged die. (b) Evaluation board.


Sensors Statistics
Step Count 1
Heart Rate (BPM) No Finger
Coarse Motion Classifier Stationary
Context Change Detect - Pedo Unknown
Context Change Detect - PDR Unknown
Tilt Orientation UP
Rotation s=0.963 x=-0.116 y=0.033 z=-0.235
Keyword Detection No Keyword
CPU Memory 192K
SUP Memory 0K
CPU Util (%) 24.02
SUP Util (%) 0.00

(a)

CPU Memory 64K
SUP Memory 320K
CPU Util (%) 2.13
SUP Util (%) 22.04
Step Count 0
Heart Rate (BPM) No Finger
Tilt Orientation Unknown
Rotation Unknown
Keyword Detection Keyword Detected

(b)

(c)

Fig. 19. Snapshots from (a) tilt motion demo, (b) voice keyword detection demo, and (c) 320 × 240 low-resolution face recognition demo.

out space between header-switch rows and logic gate rows must be reserved to honor the DNW to OD/NW physical spacing rule. Fig. 14 illustrates the physical implementation that improves the logic area utilization and the extra power meshes in a body-bias power domain. The minimal area penalty we obtained is about 20% as compared to the conventional implementation that does not require body bias. To prevent excessive leakage current from turning on p-n junctions, a maximal of 0.3-V forward body-bias voltage was allowed. In this test chip, the body-bias voltages are generated off-chip.

In parallel to NTC logic circuit enablement, we must address an equally important question: internally how do we generate the ultra-low-voltages efficiently for the SoC? The on-die PMU must provide low-power voltage and current references, overcurrent and reverse current protections, clocks and high efficiency supplies for the memory, digital circuits and analog blocks. The prior art [12] uses on-die low-dropout (LDO) converters. Clearly, in the presence of a large gap between input and output voltages, LDO converter is not the best choice, because of the fundamental linear efficiency loss.

The SoC uses "hybrid" PMU, which includes an LDO regulator and two switch-mode-power-supply (SMPS) buck regulators. The power tree of the chip is depicted in Fig. 15(a). The inductors are off-chip passive components. As seen, both SMPSs have programmable output voltages ranging from 0.45 to 1.15 V. The LSB programmable step of each SMPS is 12.5 mV, with +/−2.5% dc accuracy error that is limited by the trimming accuracy as well as the variation of the SMPS regulator itself plus the on-die bandgap reference. To enable the feature of direct lithium-ion battery attach, the SMPSs must handle input voltages up to 4.5 V. However, the maximal I/O device voltage in this 28-nm CMOS process is only 1.8 V, in contrast to 3.3-V I/O devices in 65-nm process and 6-V I/O devices in 180-nm process. To ensure all transistors are operated within the maximal voltage rating, our SMPSs use a triple-stacking output stage with interleaved timing control scheme, as shown in Fig. 15(b). Besides, the buck SMPS regulators have a unique ultra-low-power pulse-frequency-modulation mode, in which the quiescent current is well below 1 μA, allowing high power efficiencies (>80%) even under very light loads (<10 μA). This feature is extremely important for ultra-low-power applications. An off-chip SMPS converts battery voltage to 1.8-V system power rail (Vsys), to supply the pad-rings, analog modules and off-chip sensors. An on-die LDO converts the 1.8-V Vsys voltage to the 1.05-V default voltage for the AON domain. Because the AON current is usually small, using on-die LDO saves area and external component. The LDO output voltage is programmable in the range from 0.8 to 1.15 V, with 10-mV programmable LSB.

IV. Low-Power SRAM Design

All the memories in the SoC, such as the local TCMs and stream buffers, are implemented with SRAMs. TCM refers to small size, fully predictable and dedicated low-latency memory that is accessed by the processors every cycle. If not treated carefully, they would consume a heavy portion of the total chip power. The SRAM power is reduced through: 1) SRAM macro-re-organization and 2) low-voltage SRAM circuit techniques.

A. SRAM Macro-Re-Organization

The sizes of memories in the SoC range from 16 to 320 KB (i.e., 20 chunks of 16 KB). Instead of using a 16-KB SRAM macro directly, each 16-KB memory is re-organized into smaller sub-macros, as illustrated in Fig. 16(a). The usage of smaller SRAM sub-macros (that are hierarchically addressable) enables finer-grain leakage management. It also segments the word-lines and bit-lines in the otherwise 16-KB macro, hence lowering total switching capacitance and switching power, as shown in Fig. 16(b). Such power advantage is particularly predominant at low toggling rates. In the SoC, 4-KB small macro was chosen from power and area tradeoff, at the price of 1.3× silicon area.

B. Low-Voltage SRAM Circuit Techniques

The foundry SRAM array has the highest density because the feature size of bitcell is blessed by foundry push-rule (pushing of design rules). This benefits performance-driven applications which do not have much headroom for V_{DD} scaling, but it is not power optimized for low and medium speed applications. On one hand, a smaller bitcell size means smaller capacitances, which tend to reduce power. On the other hand, a smaller bitcell has a wider process variation distribution, resulting in higher functional $V_{\min}$ and retention voltage V_{ret}. This tends to increase power. Therefore, the optimal bitcell area and power tradeoff in the context of $V_{\min}$ scaling should be analyzed.

The SRAMs in the SoC are the custom dual-rail LP-SRAM arrays that use logic transistor-based 6-T bitcells. The peripheral circuit shares the NTC logic power rail. In Fig. 17(a), LP-SRAM bitcell chose 40-nm channel length devices, as it is a good tradeoff between leakage and area. In Fig. 17(b), the optimal transistor width and $V_{\min}$ were determined. Then, word-line boosting technique (see [13], [14]) was applied, to improve performance at low voltages. The above 6-sigma Monte-Carlo simulation shows that, a 100-mV world-line boosting can reduce LP-SRAM $V_{\min}$ by about 200 mV. This effectiveness is shown in Fig. 17(c). Overall, LP-SRAM lowered the bit-cell array's operating $V_{\min}$ to 0.7 V (which is the lowest voltage that meets the 50-MHz operation) and retention voltage to 0.4 V. Compared with the foundry SRAM (which is the baseline), the area of LP-SRAM is 2× larger, but it gives 2× lower dynamic power and 3× lower retention power thanks to the low-voltage operations. The key metrics are summarized and compared in Table I. In Fig. 17(d), the functional yield and retention yield curves of the LP-SRAM (both simulation and silicon measurement results) are shown. As seen, the silicon characterization results and the simulation data match with each other quite well.

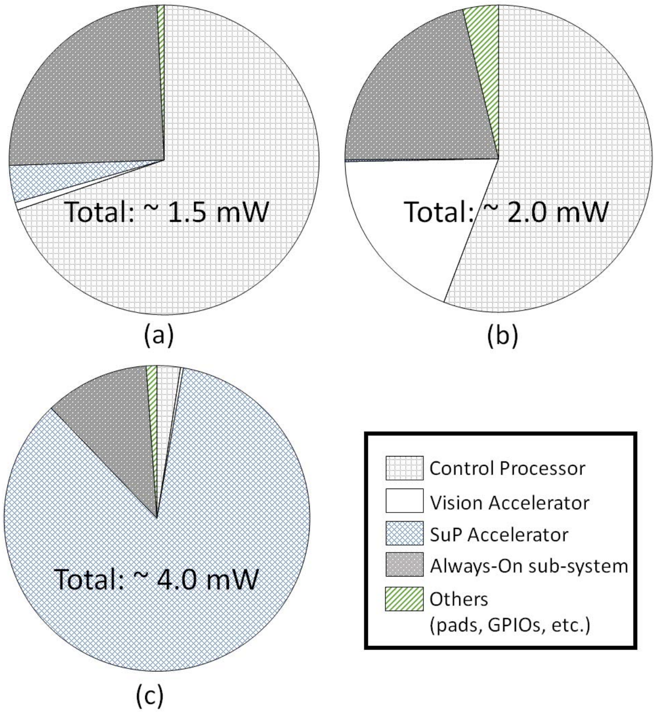


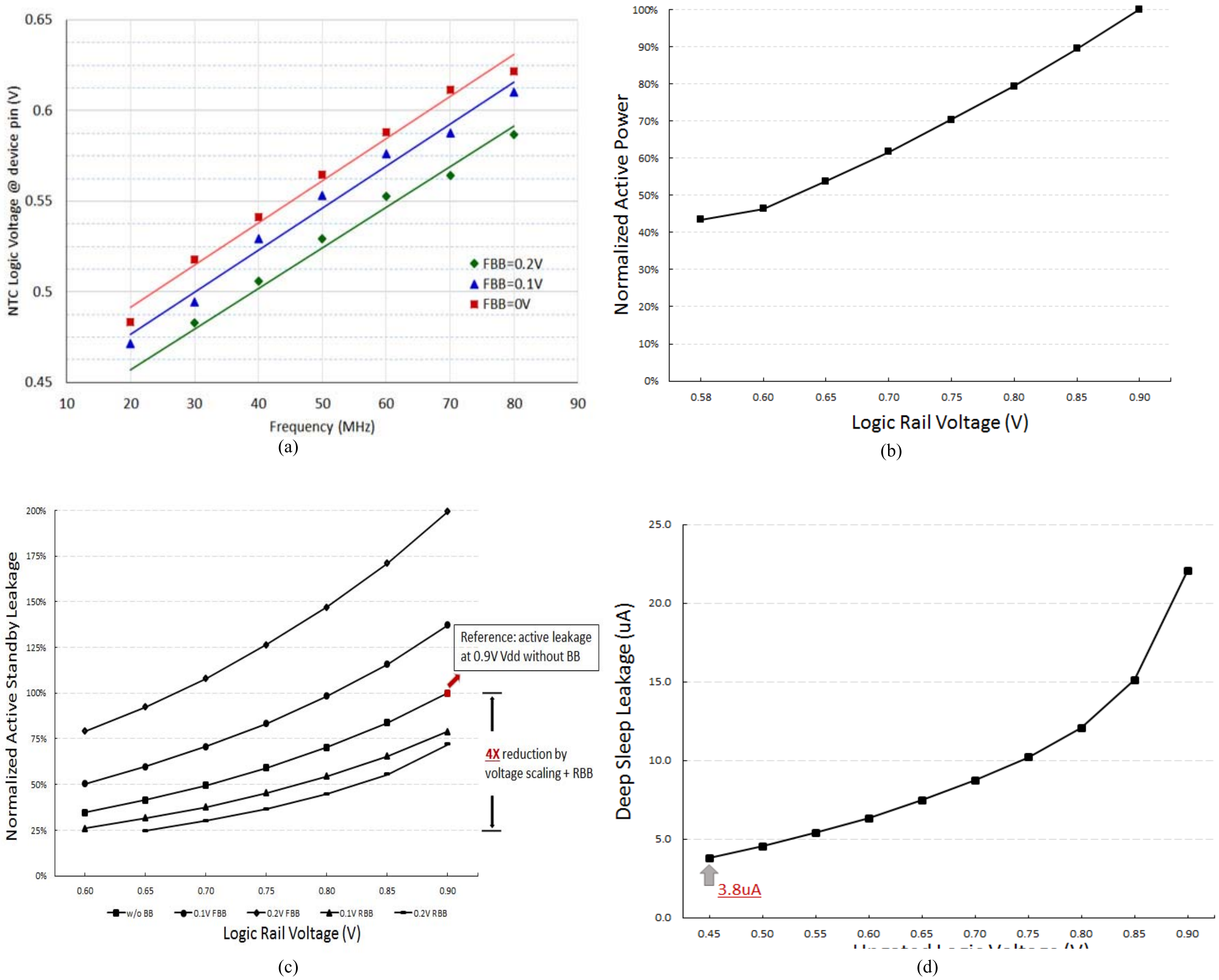
Fig. 20. Total power and breakdown of key IPs, for three active use cases.

V. Measurement Results

Fig. 18(a) shows the chip layout, where the key building blocks are highlighted. Also shown is a photograph of a packaged die. The total die size is about 3×3 mm^2. Because it was the first test chip, the layout compactness and package form factor were not particularly optimized. Some test-structures to characterize NTC circuit and device behaviors were carried on the vehicle, as annotated at the bottom of the chip layout picture. Fig. 18(b) shows the evaluation daughter card.

We had over one hundred test chip parts from each split corner lots. The samples from TT wafers started to pass tests for the expected 50-MHz operating frequency at 0.55-V logic V_{DD} at 25 °C room temperature. At 0.58-V V_{DD}, all the TT samples passed tests. The active current drawn by the logic part of the sensor control processor from a 3.8-V battery is less than 9 μA/MHz (or 35 pJ/cycle) when executing the *while(1)* benchmark test. However, μA/MHz and energy/cycle do not take the performance and the architecture (like 8/16/32 bit, MCU/DSP) into account. Although scaling voltage into the deep sub-threshold regime (see [3]) could lead to additional several times energy/cycle improvement, in our case it is the 50-MHz performance target that does not allow further voltage scaling. For the silicon samples from slow–slow (SS) wafers to pass tests, the DVFS controller had to raise the V_{DD} by 50–60 mV (i.e., 4–5 additional SMPS programmable steps). In addition, two PMU steps (25 mV) increase is needed to bound low-temperature effect.

Fig. 19 shows three snapshots from: 1) a tilt motion demo, where sloping up movement was detected [see Fig. 19(a)];



(a)

(b)

(c)

(d)

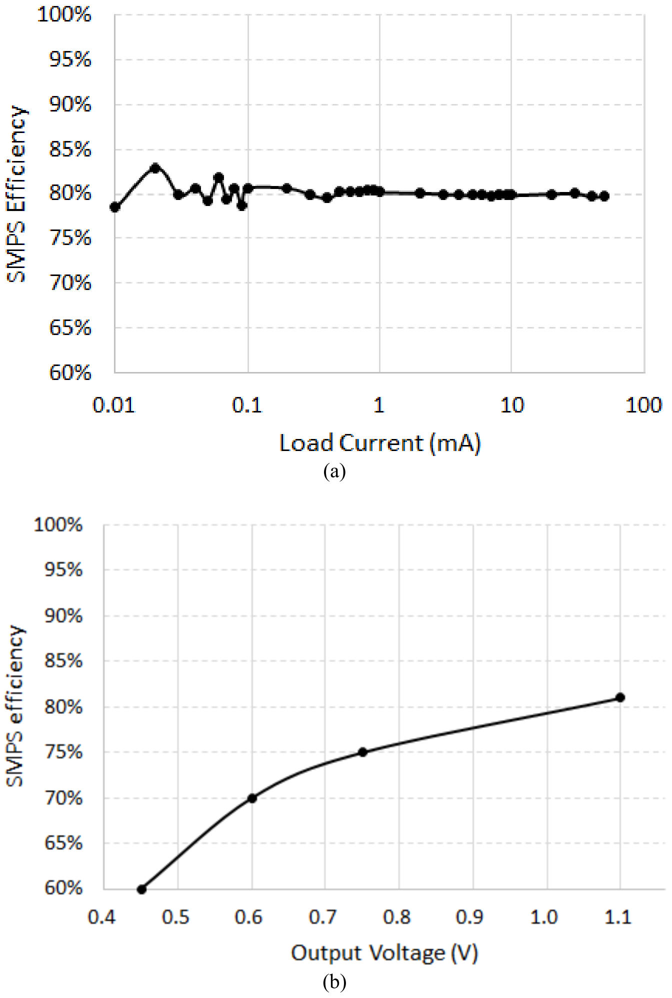
Fig. 21. (a) Performance with forward-body bias. (b) Normalized active power versus the logic rail voltage V_{DD}. (c) Active standby leakage at different VDD and body-bias voltage conditions. (d) Logic leakage power consumption in deep sleep mode.

2) a voice keyword detection demo, where the keyword was captured from a speech [see Fig. 19(b)]; and 3) an AON face low-resolution recognition demo [see Fig. 19(c)], where 320 × 240 pixels were from off-chip optic/analog image sensor. In all the demos, the sensor control processor and the accelerators were running in the NTC mode.

Fig. 20 shows the total power and power breakdown of the key IPs for three active use cases: 1) all key IPs are power gated, except the control processor which executes a *while(1)* loop; 2) the vision accelerator does filtering application, while the control processor executes control codes; and 3) the control processor first programs *SuP* to do FFT processing and it then goes to deep sleep with retention.

In Fig. 21(a), the measurement shows that, optionally, with a forward body bias of 100–200 mV in the NTC mode, the lowest operating voltage reduced to below 0.55 V. Conversely, when the voltage was maintained at 0.58 V, the design could operate at 70 MHz (which is 40% above the design target). Fig. 21(b) plots the normalized active power as a function of the logic rail voltage V_{DD}. Fig. 21(c) shows the active standby leakage current trend, as a function of V_{DD} and body-bias voltage. Through the combination of V_{DD} scaling and reverse body-bias, 4× active standby leakage reduction was achieved. In the deep sleep mode, the leakage current drawn from the ungated logic power rail is 3.8 μA at the 0.45-V SMPS output voltage, as seen from Fig. 21(d). When running SMPS at 0.45-V deep sleep mode with retention, the power mainly comprises the 1.05-V AON power and the retention power from retention registers and memories. Excluding power from components such as pads, DFT modules, ECO-fillers, the AON domain roughly takes about 30-μA running from the LDO.

The measured conversion efficiency of the SMPS in the embedded PMU is 80% for current loadings from 0.1 to 100 mA at 1-V output [see Fig. 22(a)] and 65%–70% for output voltages between 0.55 and 0.6 V [see Fig. 22(b)]. In the retention mode, to ensure the pin voltage of IPs to be above 0.45 V, the PMU output voltage is recommended to be still around 0.5 V. The efficiency is close to 60% and curve



(a)

(b)

Fig. 22. Measured SMPS efficiency. (a) Efficiency versus load current ($V_{\text{in}} =$ 3.8 V and $V_{\text{out}} = 1.1$ V). (b) Efficiency versus V_{out}.

remains flat over the load current from 10 μ A to 10 mA. At the same $V_{\text{in}}/V_{\text{out}}$, theoretically LDO efficiency is only less than 15%. As V_{DD} goes higher than the targeted 0.55-V V_{DD}, the core consumes more power, but the SMPS efficiency also increases. The net effect is that, the power-versus-V_{DD} curve remains reasonably flat in the vicinity of the 0.55-V V_{DD} point. This implies that the V_{DD} of the NTC mode could be increased slightly, without a significant increment in the total power consumption. For example, raising V_{DD} by 30 mV to 0.58 V could receive more than 90% of the low power benefits compared with the targeted 0.55-V V_{DD} operation, but doing so would allow the chips in the typical process condition to safely pass the 50-MHz frequency target, without the assistance of forward-body bias or DVFS. To improve the yield of this NTC dominated SoC, it is, therefore, suggested to redefine the NTC mode voltage. It is worth mentioning that the efficiency was measured using a non-optimized socket card. With optimized solder-down cards, the efficiency would be a few percentage higher than shown.

VI. Conclusion

This paper introduced the *Blackghost* 1.0 test chip. Driven by low-power innovations in SoC architecture and circuit, this highly integrated SoC features NTC logic operated at about 0.55-V V_{DD}, an embedded PMU with over 80% power conversion efficiency, a vision classifier processor and a streaming DSP hardware accelerator. The SoC offers rich functionalities and good power efficiencies for battery-powered IoT applications. Looking forward, we will integrate MRAM as soon as this high-density non-volatile memory technology matures in the 28-nm process, to save memory leakage power and silicon area. We will also integrate RF connectivity modules, such as BLE and Wi-Fi, to enable "power-connectivity-computation" all-in-one platforms.

Acknowledgment

The authors would like to thank the ASIC, hardware, software, and testing teams in Qualcomm, San Diego, CA, USA, for the effort toward the chip tape out and silicon characterization.

References

[1] ARM. [Online]. Available: https://developer.arm.com/products/processors/cortex-m/cortex-m4

[2] Samsung. [Online]. Available: http://www.samsung.com/semiconductor/products/bio-processor/bio-processor/

[3] Y. Pu, J. P. de Gyvez, H. Corporaal, and Y. Ha, "An ultra-low-energy multi-standard JPEG co-processor in 65 nm CMOS with sub/near threshold supply voltage," *IEEE J. Solid-State Circuits*, vol. 45, no. 3, pp. 668–680, Mar. 2010.

[4] J. Kwong *et al.*, "A 65 nm sub-Vt microcontroller with integrated SRAM and switched-capacitor DC-DC converter," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2008, pp. 318–616.

[5] S. Hanson *et al.*, "Exploring variability and performance in a sub-200-mV processor," *IEEE J. Solid-State Circuits*, vol. 43, no. 3, pp. 881–891, Apr. 2008.

[6] M. Seok *et al.*, "The Phoenix processor: A 30 pW platform for sensor applications," in *Proc. IEEE Symp. VLSI Circuits (VLSI)*, Jun. 2008, pp. 188–189.

[7] B. Zhai *et al.*, "Energy-efficient subthreshold processor design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 8, pp. 1127–1137, Aug. 2009.

[8] N. Couniot, G. de Streel, F. Botman, A. KutiLusala, D. Flandre, and D. Bol, "A 65 nm 0.5 V DPS CMOS image sensor with 17 pJ/frame.pixel and 42 dB dynamic range for ultra-low-power SoCs," *IEEE J. Solid-State Circuits*, vol. 50, no. 10, pp. 2419–2430, Oct. 2015.

[9] H. Reyserhove and W. Dehaene, "A differential transmission gate design flow for minimum energy sub-10-pJ/cycle ARM cortex-M0 MCUs," *IEEE J. Solid-State Circuits*, vol. 52, no. 7, pp. 1904–1914, Jul. 2017.

[10] S. Jain *et al.*, "A 280 mV-to-1.2 V wide-operating-range IA-32 processor in 32 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2012, pp. 66–67.

[11] N. Ickes *et al.*, "A 28 nm 0.6 V low power DSP for mobile applications," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 35–46, Jan. 2012.

[12] M. Saint-Laurent *et al.*, "A 28 nm DSP powered by an on-chip LDO for high-performance and energy-efficient mobile applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 81–91, Jan. 2015.

[13] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, Mar. 2007.

[14] B. Zimmer *et al.*, "SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 59, no. 12, pp. 853–857, Dec. 2012.

Yu Pu received the B.S. degree from Zhejiang University, Hangzhou, China, in 2004, and the Ph.D. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in association with the NXP Research, Eindhoven, and the National University of Singapore, Singapore, in 2009.

From 2009 to 2011, he was a Research Associate with Sakurai Lab, the University of Tokyo, Tokyo, Japan. From 2011 to 2012, he was a Research Scientist with the Accelerator Team, IBM Research Zurich, Rüschlikon, Switzerland. From 2012 to 2013, he was a Principal Scientist with NXP Research, where he led research in ultra-low-power MCUs. Since 2014, he has been with Qualcomm Research, San Diego, San Diego, CA, USA. He has authored or co-authored over 30 scientific publications and hold ten patents. His current research interests include (ultra-) low-power mixed-signal circuits and systems.

Dr. Pu is currently the Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS (TCAS-I).

Chunlei Shi (M'01) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1994 and 1997, respectively, and the Ph.D. degree in electrical engineering from Ohio State University, Columbus, OH, USA, in 2001.

Since 2001, he has been with Qualcomm Inc, San Diego, CA, USA. He has authored or co-authored over 20 scientific publications and holds 15 patents with six more pending. His current research interests include analog/mixed-signal/power-management IC design and low-power techniques.

Dr. Shi has been serving as a TPC Member of the IEEE CICC since 2014.

Giby Samson received the B.Tech. degree in electronics and communication engineering from Regional Engineering College, Calicut, India, in 1999, and the M.S and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, AZ, USA, in 2004 and 2008, respectively.

From 2008 to 2012, he was a Senior Design Engineer at Advanced Micro Devices, Sunnyvale, CA, USA, where he was involved in the architecture and analysis of CPU core clock distribution and timing margining methodologies. He is currently a Senior Staff Engineer with the Central Engineering Team, Qualcomm Technologies, Inc., San Diego, CA, USA, focusing on standard cell development for SOC-level power management. His current research interests include ultra-low-power circuits, process monitors and adaptive compensation techniques, and standard cell topologies for chip-level Vccmin reduction.

Dongkyu Park received the B.S. degree in electrical engineering from Korea University, Seoul, South Korea, in 2004, and the M.S. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2006.

In 2007, he joined Qualcomm Inc., San Diego, CA, USA, where he has been involved in the research and development of embedded memories and low-power circuits for mobile SoCs.

Ken Easton received the B.S. degree from Harvey Mudd College, Claremont, CA, USA, in 1989, and the M.S. degree from Stanford University, Stanford, CA, USA, in 1990.

He has been with Qualcomm, San Diego, CA, USA, since 1990, where he is currently a Senior Director with Technology in Corporate Research and Development.

Rudy Beraha was the Sr. Director of engineering at Central Research and Development Organization, Qualcomm, San Diego, CA, USA. He has more than 24 years of experience in electronics Research and Development, product development, and program management. He has been with Qualcomm Research and Development since 2005, where he has led the research team responsible for development of several key technologies use in Qualcomm's system-on-chip designs. He involved in Qualcomm's first-generation OFDMA cellular system, network-on-chip architectures, and heterogeneous computing platforms, and domain specific compute engines.

He held various HW engineering and management positions at Sun Microsystems, Santa Clara, CA, USA, Agilent Technologies, and Hewlett-Packard Company, Palo Alto, CA, USA. He is currently a VP Engineer with Atlazo Inc., La Jolla, CA, USA.

Adam Newham was with Nextwave Wireless, San Diego, CA, USA, Ensemble Communications, and Hughes Network Systems, Germantown, MD, USA, and Motorola, Chicago, IL, USA. He has over 25 years of experience in industry. In 2008, he joined Qualcomm, San Deigo, CA, USA, as part of their Research and Development division, where he was involved in multiple projects including ultra-low-power ASIC design, wearables, WCDMA, and personal and body area networks. As part of his role on ultra-low-power ASIC design, he is the Project Engineer of the Blackghost family of ASICs, CR&D Medical Implant and Internet-of-Things Security projects. As part of his work on ultra-low-power ASIC design, he continues to drive initiatives including the use of pMRAM, near-threshold logic, and low-power HW/SW architectures. He holds 19 issued patents with 24 pending patents.

Mark Lin received the B.S. degree in electrical and computer engineering from the University of California, San Diego, CA, USA, in 1992, and the M.S. degree in electrical engineering from the University of Southern California, in 1996.

Since 2004, he has been with Qualcomm, San Deigo, CA, USA, where he is currently a Senior Manager with the Corporate Research and Development ASIC Group. His current research interests include architecture and design techniques for ultra-low-power SoC.

Venkat Rangan is currently a Director Engineering with Corporate Research and Development, Qualcomm, San Diego, CA, USA.

Karam Chatha received the Ph.D. degree from the University of Cincinnati, Cincinnati, OH, USA, in 2001.

He was a tenured Associate Professor at Arizona State University, Tempe, AZ, USA. He has been with Qualcomm, San Diego, CA, USA, where he is currently a Senior Director, Engineering with Corporate Research and Development, and leads the SoC Architecture Research Team. He has authored over 60 conference and journal publications.

Dr. Chatha has been a General Chair of NOCS'13 and ESWEEK'14. He was a recipient of three best paper awards and the NSF Career Award.

Danny Butterfield (M'97) received the B.S.E.E. and M.S.E.E. degrees from the University of California, San Diego, CA, USA, in 1986 and 1994, respectively, with emphases in Communications Theory and Systems, Analog, Digital and RF integrated circuits, and then VLSI for communications systems, respectively.

In 1987, he joined Qualcomm, Inc., San Diego, CA, USA, where he was involved in the design and test of RF hardware for the OmniTRACS Ku-Band mobile satellite system for which he developed the Q3x36 PLL frequency synthesizer ICs from 1989 to 1993. From 1994 to 2002, he was involved in the design of baseband analog circuits including integrated analog filters, ADCs, RF LO, and clock PLL frequency synthesizers for multi-mode CDMA, GSM, and UMTS transceivers, audio codecs, and multimedia interfaces. From 2003 to 2007, he led the development of embedded analog IC cores for SOCs including PLLs, DACs, and timing circuits for wireless baseband and multimedia interfaces. In 2007, he joined the Corporate Research and Development Division, Qualcomm, where he focused on ultra-low-power mixed signal and circuits and systems for UWB IR transceivers, sensor analog front-ends and ADCs, RFDAC transmitter for MICS band implant devices, and voiceband audio codec. He is currently focused on mixed signal high-speed I/O.

Rashid Attar joined Qualcomm, San Deigo, CA, USA, and has involved in various aspects CDMA wireless data (EV-DO) and voice systems (IS-95, 1x-Advanced) in 1996, where he was the Project Engineer of CDMA2000-advanced from 2009 to 2013 and CDMA Modem Systems Lead at QCT from 20 through 2013. From 2014 to mid-2016, he led the ultra-low-power ASIC platform project. He is currently a Vice President Engineering with Corporate Research and Development, Qualcomm. He leads the ASIC and Hardware Department in Qualcomm Research. The Qualcomm Research portfolio consists of Communications (5G, Cellular V2X, Satellite Communications, Wi-Fi, and Industrial Internet of Things), ASIC and HW Research and Development, and Embedded IoE systems (Always ON computer vision, Autonomous Driving, Robotics, and AR/VR). The ASIC and Hardware Group Research and Development portfolio consists of 5G (RFICs, PAs, Interfaces, Packaging), processors (CPUs, Programmable deep learning accelerators), ultra-low-power platform (processor, communications, memory, machine learning accelerators, power management, wireless charging), core CMOS Research and Development (3-DIC and Thermal-aware designs), and Antenna Design. He holds approximately 160 granted U.S. patents.