Abstract—Embedded energy monitoring of critical system components can be used to enable better power management by capturing runtime system conditions such as temperature and application load. In this work, an energy sensing circuit that provides digitally represented absolute energy per operation of a 128kbit SRAM is presented. Designed in a 65nm low-power CMOS process, SRAMs can operate down to 370 mV. Energy sensing circuit consumes 16.7 µW during sensing at 1.2V (only 0.28% of SRAM active power at the same voltage). For improved performance, SRAMs utilize body-biased PMOS-input strong-arm type sense amplifiers that can achieve 45% tighter input offset distribution for only ~3.5% of total SRAM area overhead.

I. INTRODUCTION

With continuous reduction of transistor feature sizes, addressing the challenges of complex systems is becoming more difficult. Those complex systems usually need to meet multiple constraints such as a throughput requirement and a power budget. Furthermore, they often need to be optimized for conflicting goals. Additionally, those systems must perform efficiently under dynamically changing application loads and environmental changes. Self-aware systems can be used to achieve better system optimization by utilizing run-time monitoring of critical system metrics and adapting to changes [1]. In this context, on-fly energy monitoring of various system components can be used for better allocation of system resources.

On-chip memories are responsible for a large portion of the total energy consumption of modern processors [2]. Moreover, technology scaling is fueling integration of larger on-chip caches (50MB reported in [3]). Thus, an energy sensing circuit that is capable of generating a digital representation of absolute energy per operation (EOP) consumed by the caches would provide useful information to the system. However, energy sensing operation should be minimally intrusive to SRAM operation and need to introduce minimum overhead in terms of area and power.

On-chip caches not only dominate system power consumption but also limit system performance. Power consumption can be decreased by using voltage-scalability. However, scaling SRAM voltage degrades both SRAM performance and functionality. Thus, non-traditional bit-cell topologies and peripheral assist circuits are often used. Secondly, sense amplifiers are critical components of SRAMs as their input offset limit SRAM performance. Hence, reducing sense-amplifier input offset with minimum area penalty is significant to achieve better system performance.

In this work, we implemented an on-chip energy-sensing circuit for a 128kbit 8T bit-cell based SRAM design. This circuit provides a digital representation of absolute EOP of SRAM blocks. Furthermore, we propose an offset compensation technique using body-biasing to reduce input-referred offset of sense-amplifiers.

The rest of the paper is structured as follows: Section 2 presents on-chip energy sensing circuit architecture and its implementation. Then, in Section 3, 8T low-voltage SRAM design with offset compensated sense-amplifiers using body-biasing are discussed. Section 4 provides measurement results for the prototype test chip. Finally, section 5 concludes this paper.

II. ENERGY-SENSING CIRCUIT

The block diagram of our 128kbit SRAM with embedded energy sensing circuit is given in Fig. 1-a. When energy sensing is activated, an off-chip storage capacitor (Csto) powers up SRAMs. Filtering capacitor of a power converter
Figure 2. Energy sensing circuit (a) block diagram and (b) oscilloscope output of critical signals.

Figure 3. On-chip SRAM organization and simulated transient waveforms for the two-stage sensing scheme.

The clamp can be used as $C_{sto}$. During the energy sensing period, the switch which is connected to the power supply node (AVDD) stays OFF. When the voltage over $C_{sto}$ ($V_{sto}$) drops by $\Delta V$ from $V_1$ to $V_2$ in $N$ cycles (Fig. 1-c), $EOP$ can be calculated as:

$$EOP = \frac{1}{N} \times C_{sto} \times (V_1^2 - V_2^2) = \frac{C_{sto} \times (V_1 \times \Delta V)}{N}$$

Previous work presents an energy metering circuit designed specifically for systems that use pulse frequency modulated switching regulators by observing the switching frequency variation of the regulator output [4]. Similarly, the authors in [5] introduced dynamically tracking the minimum energy point of an FIR by generating relative $EOP$ numbers and without restricting the voltage drop of the supply voltage.

Absolute $EOP$ numbers might be required in a system that operates with multiple energy sensing circuits. Furthermore, for SRAMs, $\Delta V$ needs to be precisely controlled in order not to degrade stability during energy sensing mainly at low voltages. As it can be seen from Fig. 1-b, a 50 mV voltage drop at AVDD = 400 mV changes hold margin $\mu/\sigma$ from 4.9 to 2.5 which would result into a drastic degradation of robustness due to the inability of many bit-cells to retain their states.

Fig. 2-a shows the energy sensing circuit block diagram. AVDDdig and $\Delta V$dig are 7-bit inputs. $\Delta V$dig can be set with 10 mV step sizes. It can be selected sufficiently large to capture an average $EOP$ across many cycles or can be set small enough in order not to degrade SRAM operation especially at low voltages.

Multiplier and divider circuits are designed to be sequential to minimize the area overhead. The low-offset comparator is carefully sized and can be digitally controlled to ensure an input offset voltage of less than 1mV.

Fig. 2-b demonstrates the operation of the energy sensing circuit by showing the oscilloscope outputs for its critical signals. When energy sensing is initiated by ESstart, custom-designed 7-bit capacitive DAC generates $V_{ref} = AVDD - \Delta V$. Every cycle, a comparator checks if $V_{sto}$ is larger than $V_{ref}$ and comparator output (disch) keeps a 10-bit synchronous counter enabled during this period to generate $N$. When $V_{sto}$ drops below $V_{ref}$, a custom-design multiplier and fixed-point divider are used to calculate absolute $EOP$ provided that value of $C_{sto}$ is known.

III. SRAM WITH OFFSET COMPENSATED BBSA

Fig. 3 shows the SRAM architecture used in this work. Test chip features four SRAM blocks, each structured as 256 rows and 128 columns. An 8T bit-cell is used to enable low-voltage operation. Write word-line (WWL) is powered up by a higher voltage (AVDDH) to improve write-ability below 600mV. SRAMs are measured to be operational down to 370mV.

To improve performance, this work implements a two-stage sensing scheme. The first-level sensing circuits (FLS) are placed at every 32 rows of each block. FLS is performed with two static inverters. These inverters are designed to have shifted VTC characteristics by 100mV in order to result in faster tripping. Second-level sensing (SLS) utilizes body-biased PMOS-input strong-arm type sense-amplifiers (BBSA) that are proposed in this work.

The ability of a sense-amplifier to sense small input differences is limited by its input-referred offset. The authors
in [6] and [7] utilize tunable SAs which select REF among multiple different voltage potentials to compensate for their offset. Authors in [8] use SA redundancy and select better performing sense amplifier during the calibration process. This scheme requires replicating the sense amplifiers as well as a calibration scheme at start-up.

In Fig. 4-a, the BBSA circuit implementation proposed in this work is shown. BBSA utilizes body biasing to reduce input-referred offset voltage of its input transistors (M1 and M2) by controlling their body voltages (VP1 and VP2). M3 to M6 PMOS transistors drive VP1 and VP2 to either AVDD or AVDD depending on the outcome of the calibration process.

A calibration process has to be performed once at the startup. The timing diagram of the calibration signals are shown in Fig.5-a. At the beginning, calRst is kept high to make VP1 and VP2 of all BBSA circuits equal to AVDD. Then, snsEn is asserted while L-BL and REF are 0V. After sensing, DATA_OUT carries the information about positive or negative input referred offset of each BBSA. With callLatch signal assertion, this information is stored into a latch and body voltages are assigned to oppose the offsets. As shown in Fig. 5-b and Fig. 5-c, measured offset distribution of 512 sense-amplifiers can be made 45% tighter since the distribution drops from 190mV to 100mV by using AVDDB = AVDD + 150mV.

Fig. 4-b compares the statistical simulation results of read access time of this work and a conventional SRAM design. The latter has 256 cells per bit-line and uses small-signal strong-arm type SAs without offset compensation. This work brings up to 5.5x read time improvement due to two-level sensing and offset reduction using BBSA.

IV. MEASUREMENT RESULTS OF THE TEST-CHIP

Body-biased sense amplifiers for offset-compensation and on-fly fine grained energy monitoring circuit are implemented in a 128kbit 65nm SRAM test-chip. Fig 6 shows the die photo and the specifications of the test chip. Die size is 1.4mm x 1.4mm. SRAMs are measured to achieve functionality down to 0.37V with no errors.

Fig. 7 shows the EOP numbers using three different methods:
1- Reading the digital EOP output of the energy monitoring circuit.
2- Measuring energy consumption values of the test-chip.
3- Using transistor-level extracted simulation.

As it can be seen from the figure, EOP values are reduced with voltage scaling. EOPs calculated by the energy sensing circuit are accurate and always within 9.95% of the measured values.

Fig. 8 shows measured EOP calculated by the on chip energy sensing circuit across different temperature and by varying total read and write operation ratios. Read accesses result in larger energy consumption and the ratio of total read and write operations can change significantly from one application to the other. Moreover, higher temperature results in larger leakage and larger overall EOP. Depending on these conditions, EOP can change by more than 2x, so systems can greatly benefit from run time energy sensing calculations.
V. CONCLUSION

In this paper, a 128kbit 8T SRAM design that can work down to 0.37V is presented. SRAMs utilize body-biased sense-amplifiers for offset-compensation to achieve 1.9x tighter input offset distribution. In addition, an energy metering circuit that can measure run time absolute energy per operation numbers of the SRAM blocks is shown. Energy sensing circuit and DAC introduce 16.3% area compared to total SRAM area but this overhead will be smaller for larger SRAM sizes. Active power consumed by energy sensing circuit and DAC is 0.28% of the SRAM active power. The small area and power overhead of the energy sensing circuit makes it possible to be used for system-level power management.

ACKNOWLEDGMENT

This work was funded by the U.S. Government under the DARPA UHPC program. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S.

The authors thank TSMC University Shuttle Program for chip fabrication.

REFERENCES


