An Ultra-Low-Power Memory With a Subthreshold Power Supply Voltage

Jinhui Chen, *Student Member, IEEE*, Lawrence T. Clark, *Senior Member, IEEE*, and Tai-Hua Chen, *Student Member, IEEE*

***Abstract*—A 512 × 13 bit ultra-low-power subthreshold memory is fabricated on a 130-nm process technology. The fabricated memory is fully functional for read operation with a 190-mV power supply at 28 kHz, and 216 mV for write operation. Single bits are measured to read and write properly with V_{DD} as low as 103 mV and 129 mV, respectively. The memory operates at a 1-MHz clock rate with a 310-mV power supply. This operating point has 1.197 μW power consumption, of which 0.366 μW is due to leakage and 0.831 μW is due to dynamic power dissipation. Analysis of the available fan-out or fan-in that can be supported at a given voltage is summarized. A number of circuit techniques are presented to overcome the substantially reduced on-to-off current ratios and the poor drive strength of transistors operating in subthreshold. These include a gated feedback memory cell, and hierarchical read and decode circuits. The memory is dynamic, with pseudo-static operation provided via self-timed control of the keeper transistors to mitigate increased variability manifested in subthreshold operation.**

***Index Terms*—High fan-in/out, on-to-off current ratio, subthreshold memory, ultra-low power.**

I. Introduction

REDUCING the operating voltage is the most effective way to reduce integrated circuit power consumption. Recently, interest in operating CMOS circuits with power supply voltages below the transistor threshold voltage has been increasing [1]–[4]. Research into the limits of this technique, i.e., how low voltage can be effectively scaled, dates back to the early 1970s [5]. Besides the quadratic dynamic power savings, very low supply voltages promise greatly reduced leakage power [6], [7]. For example, a 1-V reduction in the supply voltage can reduce the transistor leakage current, I_{OFF}, by over one decade due to the drain-induced barrier-lowering (DIBL) effect. The gate oxide leakage can also be reduced more than 100× with the same 1-V drop in supply voltage [8], [9]. Consequently, subthreshold circuits can allow ultra-low-power designs to be fabricated on modern processes.

Subthreshold operation is applicable to a wide range of applications, ranging from wireless "motes" [10], wristwatch computation [11], to biomedical applications such as hearing aids and pacemakers [12], [13], as well as spacecraft applications [14]. Today, battery-powered handheld systems have been proliferating faster than other integrated circuit applications. Examples such as cell phones, MP3 players, and portable games abound. All benefit from increased battery life, which lowering integrated circuit (IC) power dissipation can provide. In modern system-on-chip (SoC) devices, some components, such as digital signal processors and microprocessors, must operate at high frequencies, at least intermittently. Many components do not need to run as fast, but must be integrated on the same high-performance silicon die. Additionally, some applications require a small subset of the circuits to operate continuously, e.g., real-time clocks and wakeup circuitry. Operating a subset of circuits with subthreshold supply voltages may offer the best solution to otherwise difficult power-versus-performance compromises in process selection.

Despite the power advantages of subthreshold operation, significant limitations exist. First, circuit speed is diminished due to the low transistor drive current. This suggests reduced gates per pipeline stage to provide as much speed as possible. Second, high fan-in/out circuits present difficult circuit design challenges due to the reduced on-to-off current ratio, I_{ON}/I_{OFF} [15]. Memory read bitlines (RBLs) and write bitlines (WBLs) are examples of circuits with high fan-in and fan-out, respectively. In this paper, we refer to fan-in and fan-out by the amount of transistor diffusion loading, which creates leakage paths that diminish the circuit I_{ON}/I_{OFF} ratio, rather than the conventional capacitive fan-out. The reduced I_{ON}/I_{OFF} ratio in subthreshold operation results in vanishing noise margins, eventually leading to circuit failure, and is the primary limiter for subthreshold memory design.

A. Reduced I_{ON}/I_{OFF} Ratio in Subthreshold Operation

MOS transistors operate as transconductors. The current on a 130-nm process technology is shown in Fig. 1. The leakage current I_{OFF} is conventionally defined at $V_{GS} = 0$ and $V_{DS} = V_{DD}$ for an nMOS transistor. When V_{DD} is above threshold, I_{ON}/I_{OFF} is nearly a constant over a wide range of V_{DD}. However, I_{DS} drops exponentially with V_{GS} and follows the subthreshold slope factor S when $V_{GS} < V_t$. The drain current I_{DS} is dominated by diffusion current for subthreshold operation. I_{DS} varies exponentially with the controlling gate input voltage V_{GS} and supply voltage applying V_{DS} as [16], [17]

$$I_{DS} = I_0 \left(\frac{W}{L}\right)\left(1 - \exp\left(-\frac{V_{DS}}{V_{th}}\right)\right)\exp\left(\frac{V_{GS} - V_t}{nV_{th}}\right) \quad (1)$$

Manuscript received January 5, 2006; revised March 10, 2006. This work was supported by Intel Corporation under Grant 20487, by AFRL/VSSE in Albuquerque, NM, under Contract F29601-00-D0244 0020 as well by the National Science Foundation's State/Industry/University Cooperative Research Centers (NSF-S/IUCRC) Center for Low Power Electronics (CLPE). CLPE is supported by the National Science Foundation under Grant #EEC-9523338, the State of Arizona, and an industrial consortium.

The authors are with the Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287 USA (e-mail: Jinhui.Chen@asu.edu; Lawrence.Clark@asu.edu; Tai-Hua.Chen@asu.edu).

Digital Object Identifier 10.1109/JSSC.2006.881549

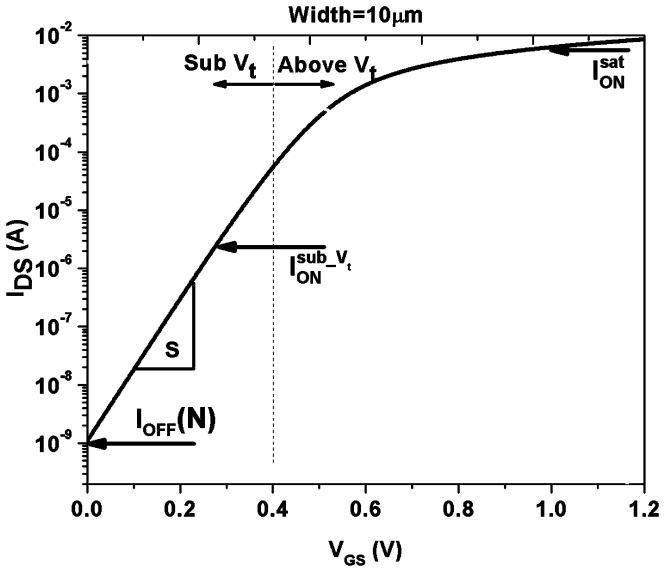

Fig. 1. MOS transistor $I_{\rm DS}$ versus $V_{\rm GS}$.

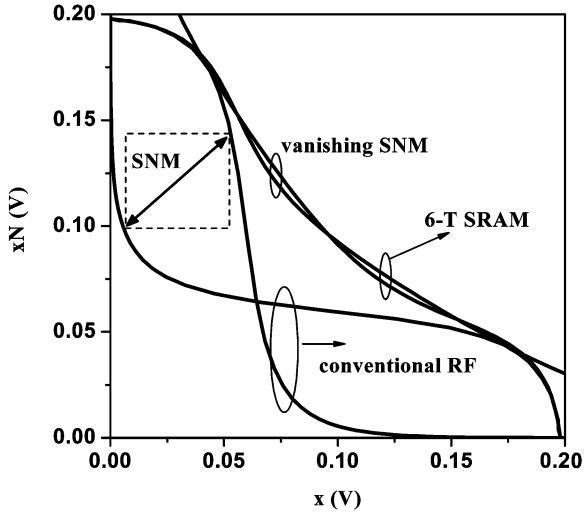

Fig. 2. SNM comparison of SRAM and RF at $V_{\rm DD} = 200$ mV.

where

$$I_0 = \mu C_{\rm ox}\left(1 + \frac{C_{\rm depl}}{C_{\rm ox}}\right) V_{\rm th}^2 \exp(1.8) \qquad (2)$$

and $C_{\rm ox}$ and $C_{\rm depl}$ are gate oxide capacitance and depletion capacitance, respectively. $V_{\rm th}$ is the thermal voltage and μ is the electron mobility. The technology constant n characterizes the process subthreshold swing, and is usually between 1 and 2 [18]. As $V_{\rm DD}$ is scaled below $V_{\rm t}$, the transistor $I_{\rm ON}$ is diminished exponentially, approaching the transistor $I_{\rm OFF}$. In fact, in subthreshold the $I_{\rm ON}$ is just $I_{\rm OFF}$ multiplied by $10^{V_{\rm DD}/s}$, as evident from Fig. 1. Hence, the ratio of $I_{\rm ON}/I_{\rm OFF}$ may be many orders of magnitude lower than for above threshold operation.

B. Challenges in Subthreshold Memory Design

Logic gates, such as NAND, NOR, or other combinational logic gates, work normally in subthreshold except that their propagation delays increase exponentially with the reduction of supply voltage. However, subthreshold memory design is problematic due to the reduced $I_{\rm ON}/I_{\rm OFF}$ [15]. It is well known that a traditional static random access memory (SRAM) cell has vanishing read stability at low voltage since the read current can raise the logic low storage voltage to the trip-point of the cell [19]. This can be alleviated by the use of a register file type of design, which provides a read current path independent of the storage node. Adding two transistors to keep the read current from affecting the cell value, such as in the conventional register file read-out circuit, increases the read stability at low $V_{\rm DD}$ dramatically compared to a traditional SRAM cell. Simulation results for both a traditional six-transistor SRAM and a register file (RF) cell are shown in Fig. 2 with $V_{\rm DD} = 200$ mV. A RF cell maintains static noise margin (SNM) in the storage cell when reading, as evident in the figure. Therefore, in the design described here, we use a RF type of cell rather than a SRAM cell. Of course, a larger cell area (due to more transistors) is required to obtain this better SNM at lower operating voltage. The goal here is to design a reasonably dense memory with the lowest possible operating voltage ($V_{\rm min}$). Consequently, the highest achievable density is not the primary objective and is explicitly traded off to obtain a lower $V_{\rm min}$.

Besides SNM, there are additional difficulties for subthreshold memory design. The high fan-in/out of the WBL and RBL, even for relatively small numbers of cells, e.g., 16 or 32 can result in circuit failure. The key for design is determination of when the circuit fails as $V_{\rm DD}$ scales down, i.e., what is the minimum operation voltage, $V_{\rm min}$, at a given fan-in/out. Alternatively, determination of the maximum fan-in/out at a given operation voltage, can be used to guide the memory design.

A latch with no gating transistor in the feedback path, i.e., a jam latch is generally used for the storage in RF cells to minimize the cell size [20]–[23]. This can reduce the write-ability of the cell at low voltages, since the driving transistor $I_{\rm ON}$ can be comparable to the feedback transistor $I_{\rm ON}$ at process corners. Variability effects are larger when operating in subthreshold, so these effects must be comprehended. It is also important to avoid circuit races, which are more likely to fail in subthreshold due to increased circuit variability at process corners and due to random variation.

II. Theoretical Basis

A. Failure of High Fan-In/Out Circuits in Subthreshold

As mentioned, the low $I_{\rm ON}/I_{\rm OFF}$ ratio presents difficulties primarily for circuit nodes that have a very high or low P-to-N ratio. The most common cases occur on the memory bitlines, where a large number of nMOS drain connections are driven by a single device. For conventional CMOS circuits, the inverter output is expected to be $V_{\rm DD}$ when the input is low. However, the driver output deviates from $V_{\rm DD}$ when the high fan-out nMOS leakage current, $I_{\rm OFF}$, is non-negligible with respect to $I_{\rm ON}$ of the driving pMOS. The resulting low $I_{\rm ON}/I_{\rm OFF}$ ratio degrades the high output level ($V_{\rm OH}$) for high nMOS fan-in/out, and similarly, the low output level ($V_{\rm OL}$) cannot reach $V_{\rm SS}$ for high pMOS fan-in/out. This degradation of the circuit noise margins is increased by process variation [15]. An analytical model is

used here to provide guidelines for robust subthreshold memory design.

B. Analytical Model

As the fan-out increases or $V_{\rm DD}$ decreases, the circuit SNM degrades until the circuit fails. A gate with fan-out driving another gate without fan-out, referred to here as the driver and receiver, respectively, is used as a model circuit. Focus here is on high nMOS fan-out circuits since the high fan-in circuit is equivalent, and since pMOS and nMOS have symmetric characteristics. Additionally, common circuits such as RF and SRAM have fan-in/out dominated by nMOS transistors. The leakage current in high nMOS fan-out helps provide $V_{\rm OL}^{\rm driver} \leq V_{\rm IL}^{\rm receiver}$. Focus is thus on the driver $V_{\rm OH}$ and receiver $V_{\rm IH}$ values as they are affected by operating voltage and fan-out.

In [15], the driver output and receiver input values were derived as

$$V_{\rm OH}^{\rm driver} = V_{\rm th} \cdot \ln\left(\frac{(K_1-1)+\sqrt{(K_1-1)^2+4K_1\alpha^{-1}}}{2K_1\alpha^{-1}}\right) \quad (3)$$

$$V_{\rm IH}^{\rm receiver} = \frac{n}{2} \cdot V_{\rm th} \cdot \ln\left(\beta(V_{\rm OL}^{\rm receiver})\right) \quad (4)$$

$$V_{\rm OL}^{\rm receiver} = V_{\rm th} \cdot \ln\left(\frac{n+1+\alpha}{n+2} - \sqrt{\left(\frac{n+1+\alpha}{n+2}\right)^2 - \alpha}\right). \quad (5)$$

Here,

$$\beta(V_O) = K_0 \cdot \frac{\left(1-\exp\left(\frac{V_O - V_{\rm DD}}{V_{\rm th}}\right)\right)}{\left(1-\exp\left(-\frac{V_O}{V_{\rm th}}\right)\right)} \quad (6)$$

$$K_0 = \frac{I_0^P \cdot \left(\frac{W}{L}\right)_P}{I_0^N \cdot \left(\frac{W}{L}\right)_N} \cdot \exp\left(\frac{V_{\rm DD} + V_{\rm tn} - V_{tp}}{nV_{\rm th}}\right) \quad (7)$$

and $\alpha = \exp(V_{\rm DD}/V_{\rm th})$, $K_1 = K_0/(N+1)$. N is the fan-out at the circuit node in question, which describes the total width of diffusion-connected gates normalized to the width of the driver. When operating with $V_{\rm DD}$ above threshold, the term $4K_1\alpha^{-1}$ is negligible since α^{-1} is vanishing, giving $V_{\rm OH}^{\rm driver}$ a linear relationship with $V_{\rm DD}$. However, $V_{\rm OH}^{\rm driver}$ is nonlinear with respect to $V_{\rm DD}$ in subthreshold operation. $V_{\rm IH}^{\rm receiver}$ has a linear relationship with $V_{\rm DD}$ in both sub- and above-threshold operation.

The analytical model is computationally efficient since the formulations are closed-form and require no iteration for solution. The parameters are extracted directly from transistor characteristics. The model is verified by comparison with circuit simulation on the target 130-nm process. Fig. 3 shows the comparison of receiver $V_{\rm IL}$ and $V_{\rm IH}$ between the analytical model and circuit simulation. The error is less than 10% of $V_{\rm DD}$. $V_{\rm OL}$ and $V_{\rm OH}$ comparisons are also shown in the figure, and the results agree to within 4% of $V_{\rm DD}$.

For correct operation, $V_{\rm OH}^{\rm driver}$ and $V_{\rm IH}^{\rm receiver}$ must satisfy

$$V_{\rm OH}^{\rm driver} \geq V_{\rm IH}^{\rm receiver} \quad (8)$$

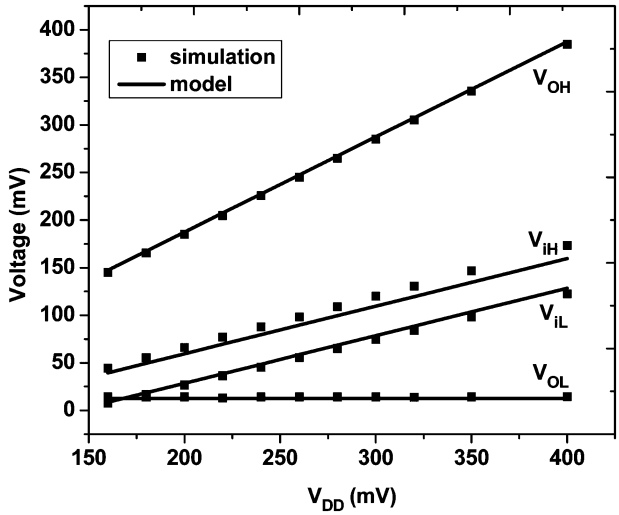

Fig. 3. Comparison between simulation and analytical model for receiver.

to provide positive SNM. Substituting $V_{\rm OH}^{\rm driver}$ and $V_{\rm IH}^{\rm receiver}$ in (8) yields the maximum fan-out $N_{\rm max}$ at a given operating voltage as

$$N_{\rm max} = \frac{\left(C_1^2-1\right)K_0}{2C_1+2-4\alpha^{-1}} - 1 \quad (9)$$

where

$$C_1 = 2\alpha^{-1}\beta\left(V_{\rm OL}^{\rm receiver}\right)^{n/2} - 1. \quad (10)$$

Solving (8) numerically yields the minimum operating supply voltage $V_{\rm min}$ at a given fan-out N. Fig. 4 shows the onset of correct operation with positive SNM as $V_{\rm DD}$ increases at different fan-out N. As expected from (3), the driver $V_{\rm OH}$ varies nonlinearly with $V_{\rm DD}$. The results also demonstrate that the minimum supply voltage for correct operation $V_{\rm min}$ increases as the fan-out goes up. Fig. 4 is for the typical process. The fast-N slow-P corner of operation raises $V_{\rm min}$. Consequently, minimizing the fan-in/out, while achieving acceptable array efficiency, is imperative for subthreshold memory design.

III. Circuit Design and Operation

The 512 × 13 bit memory is implemented hierarchically, with sub-banks each having 32 rows of 13-bit words (12-bit words plus one parity bit). There are four memory banks, each comprised of four sub-banks arranged horizontally. The physical layout is described in detail in Section IV. Hence, each bank has 32 global wordlines (WLs), each driving local WL drivers. Each sub-bank WL driver is locally gated. The local bitlines (BLs) (those within each sub-bank) have 16 cells above and 16 cells below the read and write circuits. The maximum circuit write bitline (WBL) fan-out is thus 16, to minimize the required operating voltage, while still using reasonably conventional circuits. The read bitline (RBL) fan-in is 8 to limit the read dynamic power consumption and fan-in as described below. The local RBLs drive the global RBLs, which run the height of the array. The array is organized with four sub-banks per global BL, with their outputs multiplexed at the bottom of the array.

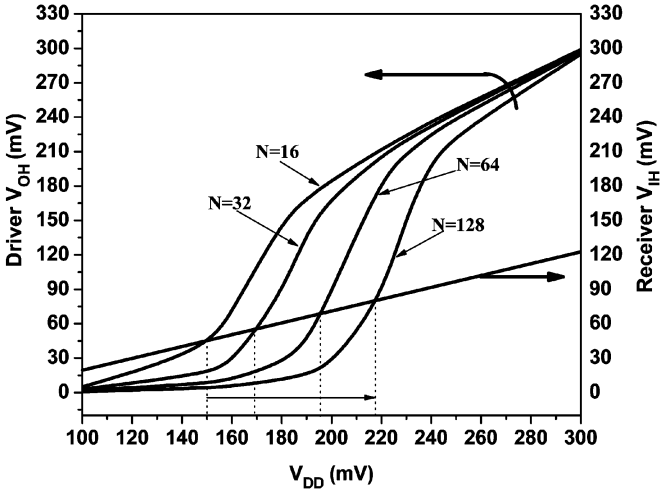

Fig. 4. Different fan-out circuits SNM versus $V_{\rm DD}$. The minimum operation voltage $V_{\rm min}$ is the $V_{\rm DD}$ where the driver $V_{\rm OH}$ exceeds the receiver $V_{\rm IH}$.

The chip core circuits operate at a nominal $V_{\rm DD}$ of 1.2 V while the chip level I/O circuits operate at 1.5 to 2.5 V. The former utilize conventional level shifting circuits [24]–[26]. The subthreshold memory circuit logic levels are obviously incompatible with the core voltage. A specially developed level shifter is used to convert the memory output data to the chip core circuit voltage. This level shifter differs from the conventional design in its ability to convert signals from subthreshold to above threshold with a reasonable nMOS-to-pMOS size ratio [27].

A. Memory Cell

The traditional RF cell (see Fig. 5) writes by pulling down on one side or the other when the write wordline (WWL) is asserted high, depending on the write bitline (WBL) state. The node to be written is a ratioed circuit, where the input pull-down transistors must overcome the pMOS feedback device to write. In the conventional RF cell design, transistors N3, N4, and N5 are sized to provide adequate strength to overcome the pMOS devices across process, voltage, and temperature (PVT) corners. In subthreshold, the variation effects are larger, and sizing may be inadequate.

Ideally, the latch feedback (FB) loop is open during the write operation, and closed when the write is complete. This is accomplished by gating one of the feedback path transistors, Pgate, as shown in Fig. 5, which improves the write margin in subthreshold. The WWL controls the single additional pMOS transistor. The feedback inverter pMOS helps to raise the voltage at node XN when that node is being written to a logic "1". A simulated write operation is shown in Fig. 6 with $V_{\rm DD} = 180$ mV. At the rising edge of the WWL, the write transistor drives the output storage node. The input node transitions slowly due to the open FB loop. When the loop is closed by the WWL, the cell value is quickly updated.

Fig. 5 has the same read-out circuit as a conventional RF cell, i.e., a two transistor stack as mentioned in Section I-B. A typical conventional register file has 16 or more pull-down nMOS transistors connected to the RBL [20], [22]. This fan-in limits the memory $V_{\rm min}$. To reduce the RBL fan-in, the outputs of two storage cells drive a single-stage complex CMOS gate to share

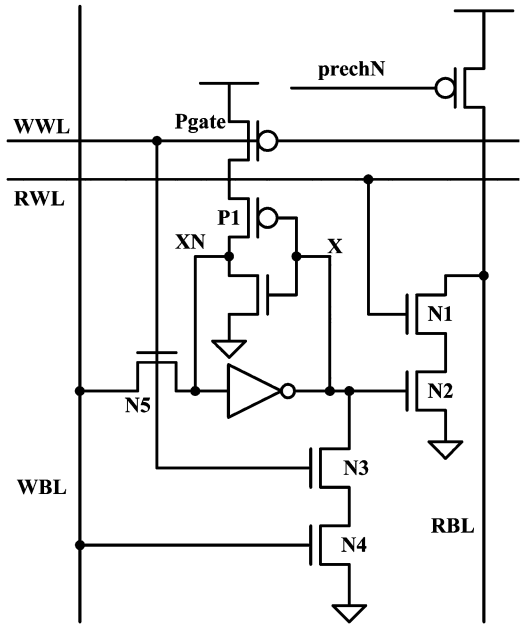

Fig. 5. RF cell with conventional read circuit. An additional transistor, Pgate, is added to aid write margin in subthreshold.

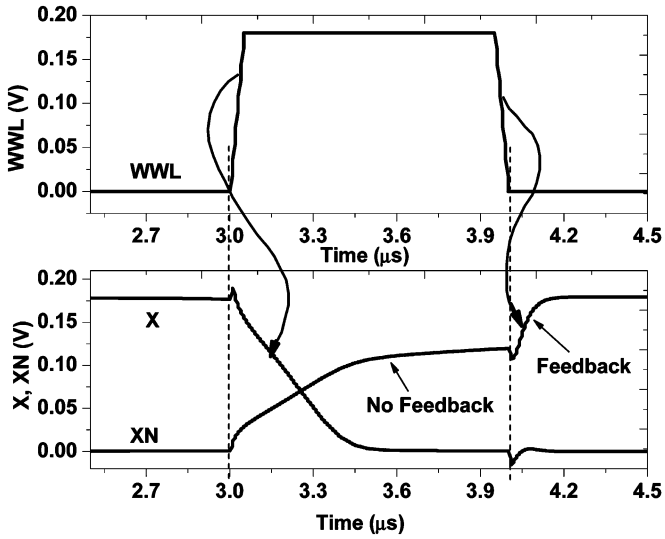

Fig. 6. Simulated write operation waveform.

one pull-down nMOS transistor as shown in Fig. 7. This scheme is similar to the single pull-down per RF cell used in [28]. The single pull-down transistor on the RBL reduces the diffusion capacitance as well as the circuit fan-out. Since there is no stack, the transistors can be half as wide as in a conventional design. When combined with the reduced fan-in, the overall RBL capacitance savings over the conventional design approaches 4×, reducing both delay and RBL contribution to the dynamic power dissipation by that amount. The capacitive load of the RWL is also reduced, which improves speed. The single-stage complex merge gate is implemented using minimum-sized transistors, which adds a small delay, but improves the speed by a greater amount due to the reduction in the RWL capacitive load. In a conventional RF array, the RWL driver is not located close to the RBL pull-down network, i.e., this domino pre-charged circuit

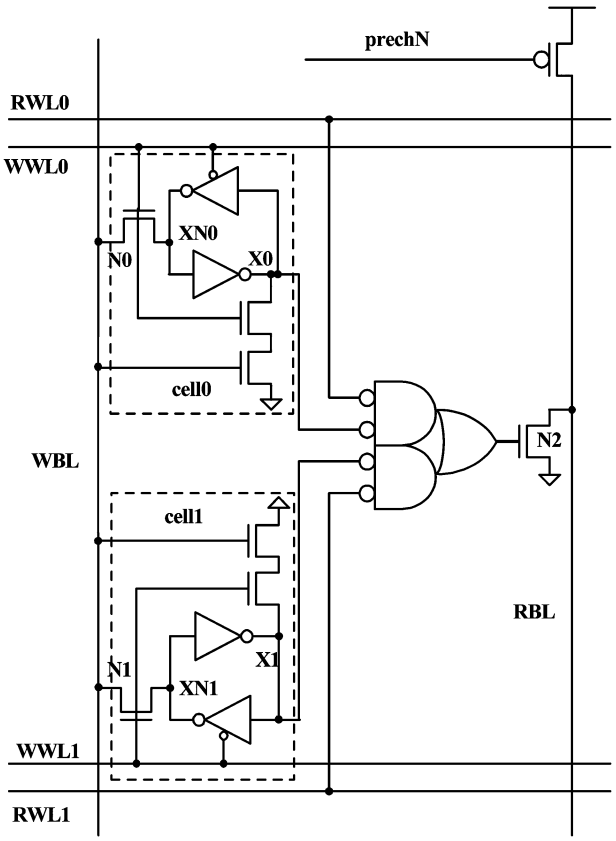

Fig. 7. Controlled feedback RF cell.

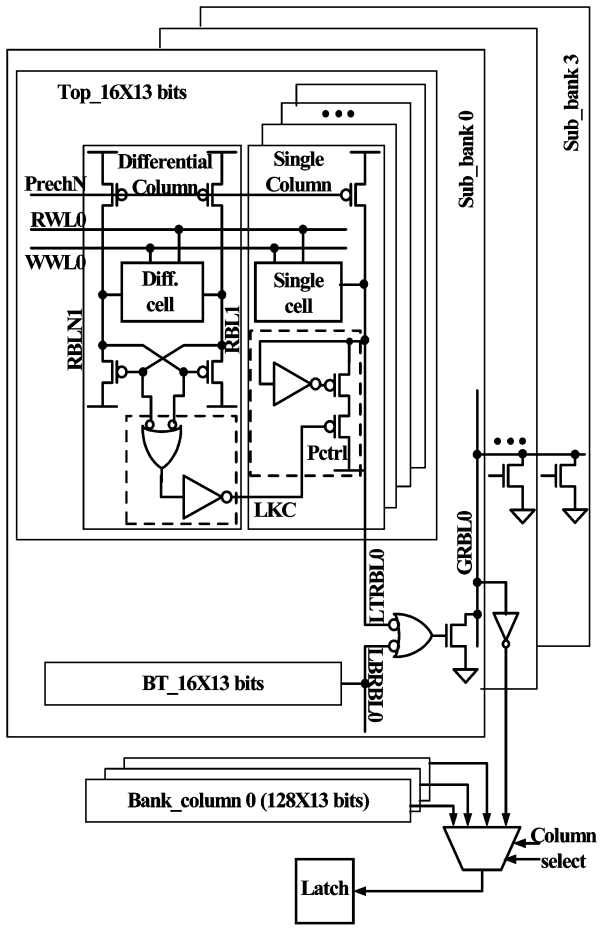

Fig. 8. Subthreshold memory structure.

has a nonlocal ground, which can increase noise at the domino (input) read transistors. In the cell used here, the static gate provides a local ground, increasing noise immunity as well as combining the cell outputs.

B. Read Circuits

The RBL is a precharge/discharge domino circuit. A pMOS keeper is usually used to prevent RBL discharge due to leakage, in the case when the read-out bit is logic one. In the conventional domino keeper design, the RBL has high fan-in, which produces a large leakage current, I_N. A strong keeper can tolerate greater leakage current but can make the RBL unwritable by the cell, since the RBL is a ratioed circuit until the keeper is turned off. The situation is analogous to the write-ability issues outlined for the cell above. The ideal keeper size should be variable, and this problem has been addressed extensively [29], [30]. The specific solution used here is to gate the keeper off during the read, but to turn it on once the read is complete. This provides no contention during the read, provides pseudo-static operation, and robustness against leakage discharging the RBL.

Replica timing was ruled out, based on high circuit variability at low voltages. Instead, a self-timed circuit, which relies on cells attached to the same RWL to generate a timing signal to control the keeper is used. This control signal turns the keeper on or off at a time determined by the RBL status during the evaluation phase. An additional pMOS transistor, Pctrl, is used to gate the keeper transistor as shown in Fig. 8. The keeper is turned on or off depending on both the RBL status and the keeper control signal. One bit cell is read differentially, and the other 12 bits have single-ended readout. The differential cell has both RBL and its complement output RBLN. Both RBL and RBLN are high during the precharge phase. During the evaluation phase, one is discharged low depending on the value read out.

A differential cell does not require a keeper control, since each of the cross-coupled keepers is activated by the complimentary bitline. The local keeper control signal LKC (or global GKC) is generated by a NAND gate that detects either RBL or RBLN transitioning low (see Fig. 8). Since the differential and single-ended RBLs have nominally identical delays, the single ended RBL keepers are turned on two gate delays after a read is completed. Every sub-bank has its own differential cells, so that variation is localized. This scheme is extended to the pMOS keepers on the global RBLs as well. The simulated read and keeper operation is illustrated in Fig. 9. The differential local RBL's labeled LRBL1 and LRBLN1, activate the local keeper control signal LKC as shown. The low assertion turns on the gating pMOS transistor Pctrl in Fig. 8. The keeper delayed from the self-timing signals allows buffering time and provides timing margin.

C. Address Decode

The poor drive current of transistors in subthreshold mandates multiple levels of decode to balance the driven capaci-

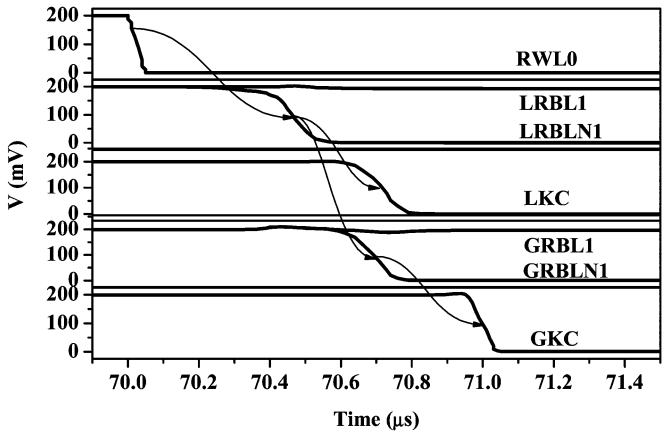

Fig. 9. Simulated keeper control signal waveform.

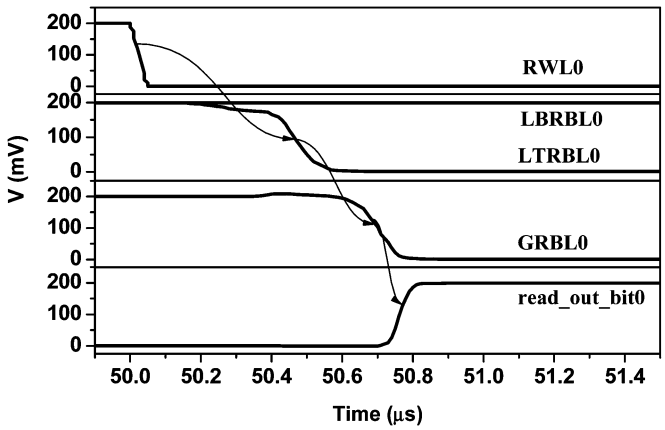

Fig. 10. Simulated read operation waveform at $V_{\rm DD} = 200$ mV.

tances across multiple gate stages. Each sub-bank has a local WL driver located at its center. Only the WL's of the selected sub-bank are ever active, limiting total active power dissipation. The global decoder is located at the center of the memory array. The GWL generated by the global decoder asserts the bank local wordline drivers. The global decoder and GWL drivers are kept small by reducing the load driven, through the hierarchical WLs and optimizing the logic levels in the decoder. The decoders use static CMOS logic exclusively. The decoder setup time is comparable to that required to read out the array after the clock activates the GWL.

D. Operation for Low Power

As mentioned, only one sub-bank is active in any single operation to minimize active power dissipation. The unselected sub-bank RBL's remain in the precharged state, which also keeps them from activating the hierarchical global read BLs. The active RBL remains high if the data read is "0", otherwise, it is discharged. The top and bottom local RBLs are multiplexed by logically NANDing them so no multiplexer select signal is required. The simulated read operation is shown in Fig. 10. The local upper RBL (LURBL) is discharged after the RWL is enabled. The local I/O circuit receives the RBL value and, depending on the value, asserts the gate of a pull-down nMOS transistor (see Fig. 8) to discharge the global RBL (GRBL).

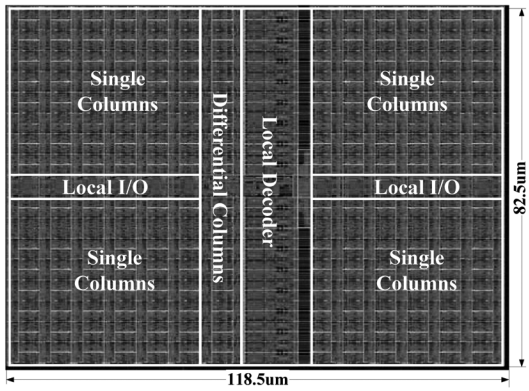

Fig. 11. Sub-bank layout.

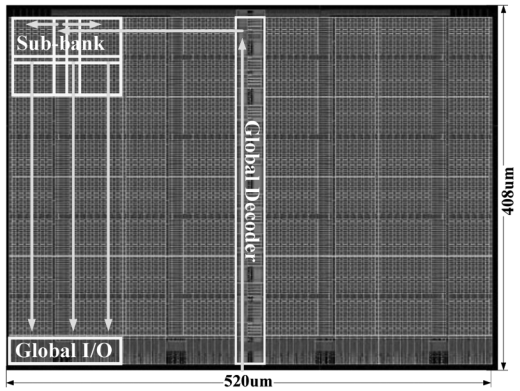

Fig. 12. 512 × 13 bit subthreshold memory layout.

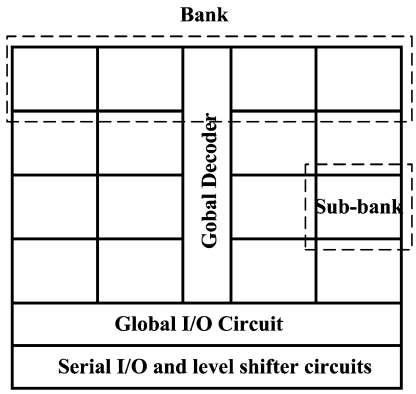

Fig. 13. Block diagram of memory structure.

IV. Physical Design

The design is implemented in a 130-nm process technology. To minimize the time delay due to wire capacitance and overall power consumption, minimum area was a goal of the design. The differential and single ended cells have the same height of 4.74 μm as all the cells in one row share both WWL and RWL signals. Fig. 11 shows the sub-bank floor plan. The local decoder is located at the center. The differential column is located next to the decoder, to minimize the distance and hence maximize the matching to the single-ended read-out columns. It was constructed by adding a second read circuit to the other side of the single-ended cell. The local I/O circuit is placed at the center, to minimize the RBL capacitance by splitting 16 cells

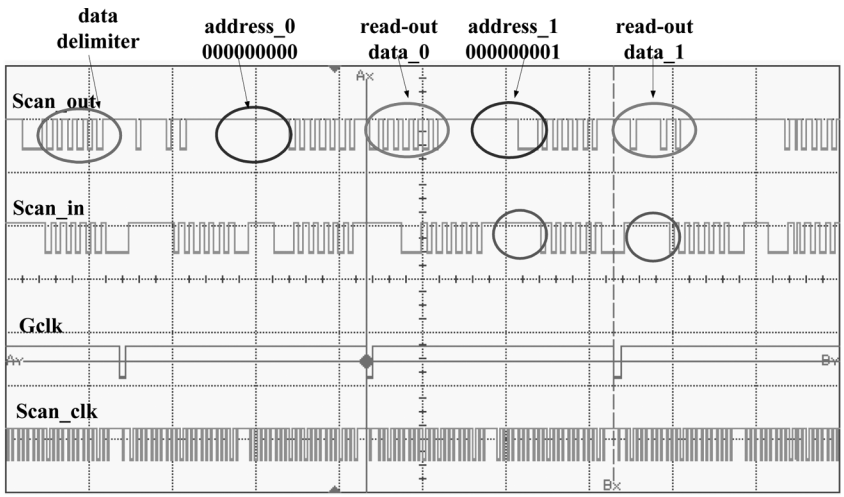

Fig. 14. Functionality test.

above and 16 below. The sub-bank is $118.5 \times 82.5\ \mu\text{m}^2$. The overall memory layout is shown in Fig. 12. The total layout area is $520 \times 480\ \mu\text{m}^2$. Access to the memory is via scan chains to limit the required I/O count. A block diagram of the memory partitioning into banks and sub-banks is shown in Fig. 13.

V. Measured Results

The fabricated memory was measured at room temperature at different power supply voltages and operating frequencies. Functionality, maximum speed versus voltage, and power consumption at different frequencies and V_{DD} were determined. In the experimental setup, a Perl program is used to drive a field programmable gate array (FPGA) based test board. The FPGA generates stimulus to drive the test chip. The output can be either observed by a scope directly or returned to the test board FPGA and transferred to a PC through an RS-232 port. In this manner, slow functional testing can be completely driven by software, without reprogramming the FPGA. A separate pin is used to supply power to the memory core, allowing the power consumption to be measured directly and without contribution from the test access and scan circuits.

A. Functional Test

For functional test, the target address and data are loaded into the scan chains and written into the memory. After loading the read address into the scan chain, it is transferred to the decoder and enables the mapped word in the selected sub-bank. The read-out bits are sent to the memory I/O circuits in parallel. At this point, the subthreshold logic values are level shifted to the core voltage. The scan chain then shifts out the data in serial fashion. A few pins are brought out directly to facilitate speed testing. Representative measured operational waveforms comprise Fig. 14. Here, Gclk is the memory global clock. The scan chain clock (Scan_clk) is asserted to capture the values and shift the data in and out. The shift register operation, i.e., parallel load or serial shift, is controlled by the scan enable signal. The scan clock and serial data both entering and leaving the IC pins are evident.

B. Minimum Operational Voltage

The subthreshold memory has a minimum theoretical operational voltage (V_{min}) with the given fan-out/in of 16 and 8 for the WBL and RBL, respectively, as described in Section II. Above this point, the memory has positive SNM for V_{DD} down to about 180 mV based on the model. To determine V_{min} for data read operation, data was written into the memory cells in the memory at high voltage, and then read out repeatedly, while adjusting the memory voltage from high to low. When the read-out data differs from that written, the resulting V_{min} was recorded. The sub-bank failure voltage bitmap for read operation is shown in Fig. 15(a). The failure voltage is mostly distributed randomly, presumably due to random process variations, but some row and column effects are evident. The highest V_{min} is 190 mV and the lowest cell failure operation voltage is 103 mV. The highest V_{min} is somewhat higher than the simulation result of 160 mV, but V_{min}'s are distributed around that value.

For write operation, data is written into cells at different voltages from high to low. The data is then read out at a high voltage. When the read-out data is different from that written, the write operation voltage is recorded as the writing failure voltage for the cell. The measured results are shown in Fig. 15(b). The highest write operation failure voltage is 216 mV, and the lowest is 129 mV. Both are higher than those of the read operation. This was expected, since the WBL has fan-out of 16, which is higher than the fan-in of 8 on the RBL.

The measured distributions of write failure voltages have a higher deviation than that of the read operation. The most likely failure voltage for write operation is 160 mV, which is higher than the median of 150 mV for the read operation. Table I shows the measured write and read minimum operational supply voltages. In no case was the measured minimum data retention voltage limiting—the cells retain state at lower power supply voltages than required for reads and writes. The effects of systematic and random variations on the theoretical V_{min} are described in more detail in [15]. In a product operating in subthreshold, the necessary supply voltage guard band would depend on the specific process variability, yield, and applications.

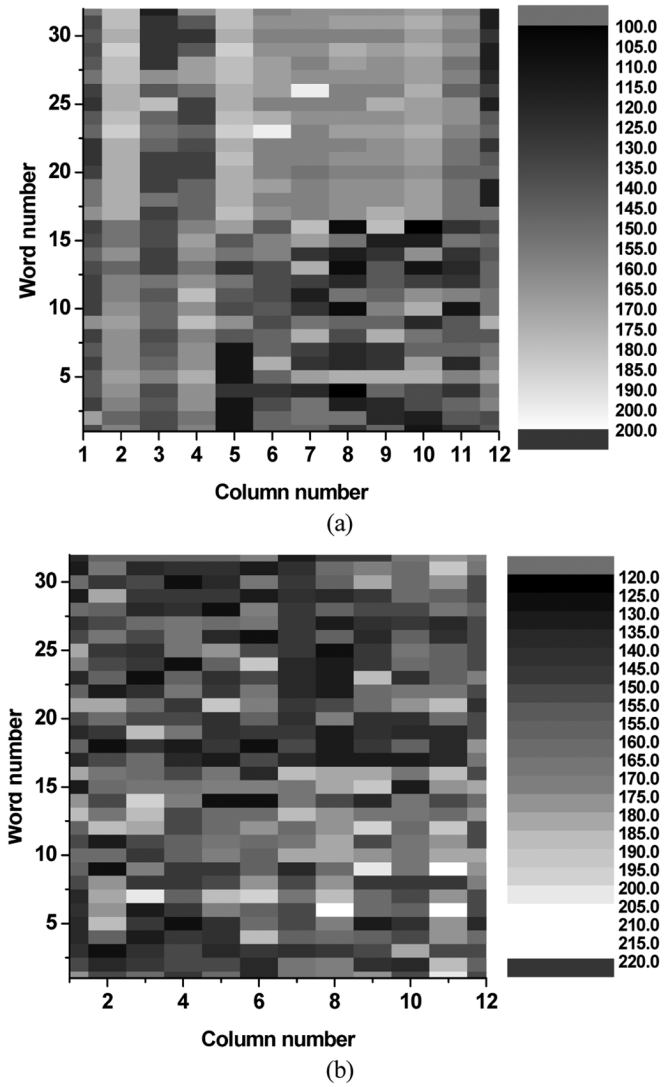

Fig. 15. Measured sub-bank failure voltage bitmap of (a) read operation and (b) writing operation. The values to the right are in mV.

TABLE I
COMPARISON OF MEASURED RESULTS OF READ AND WRITING OPERATION

	Fan-in/out	Highest failure voltage (mV)	Lowest failure voltage (mV)	Most likely failure voltage (mV)
Write operation	16	216	129	160
Read operation	8	190	103	150

C. Performance Measurement

The memory achieved a maximum operating frequency of 28 kHz at $V_{\rm DD} = 190$ mV measured at room temperature as shown in Fig. 16. The maximum operating frequency $F_{\rm max}$ was 2 MHz at $V_{\rm DD} = 325$ mV. As expected, the measured results show speed reduces exponentially when the power supply voltage scales down in the subthreshold mode. Of course, the memory is capable of operating above threshold, and was tested to be functional up to 1.2 V. The measured critical path

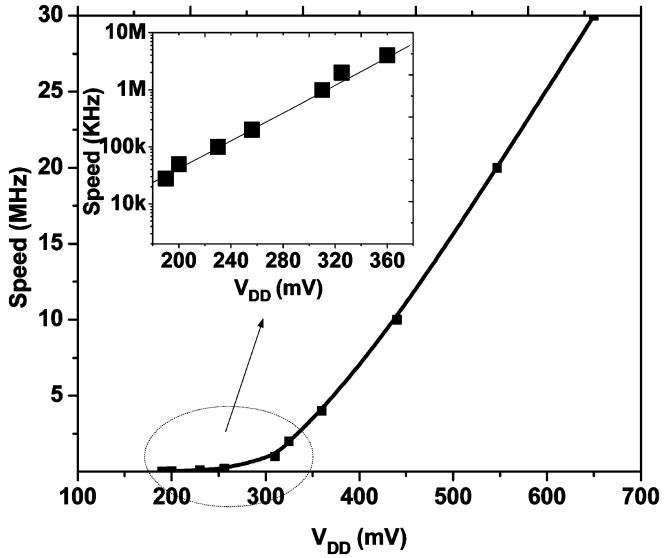

Fig. 16. Measured test chip speed versus $V_{\rm DD}$.

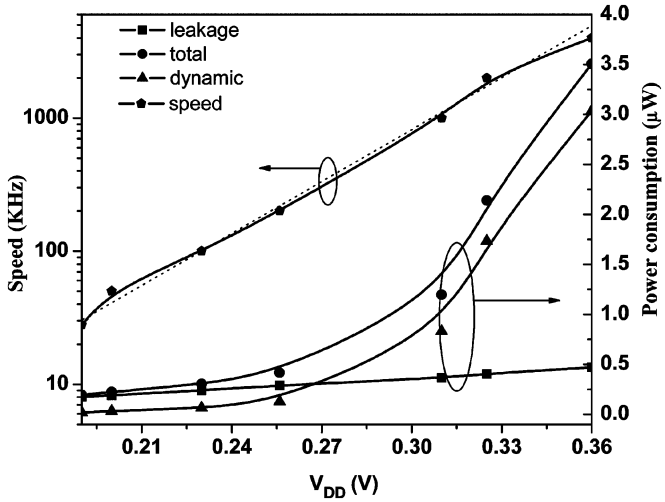

Fig. 17. Measured test chip power consumption in subthreshold mode.

is through the decoder, which limits the maximum operating frequency for a balanced clock duty cycle.

D. Measured Power

The total power includes both static power and dynamic power as

$$P_{\rm total} = P_{\rm static} + P_{\rm dynamic} = P_{\rm static} + \alpha C V_{\rm DD}^2 f \qquad (11)$$

where f is the operating frequency and α is the circuit activity factor. The static power is due to transistor leakage, which is reduced as $V_{\rm DD}$ scales down. Fig. 17 shows the measured test chip speed at different supply voltages in the subthreshold region of operation. The power consumption, including the leakage, dynamic, and total power components, was measured with the memory clocked at its maximum operational frequencies for each supply voltage, as determined by the failure point. The passing points are shown. The test chip consumes 1.197 μW of power, which includes 0.366 μW of leakage power and 0.831 μW of dynamic power consumption

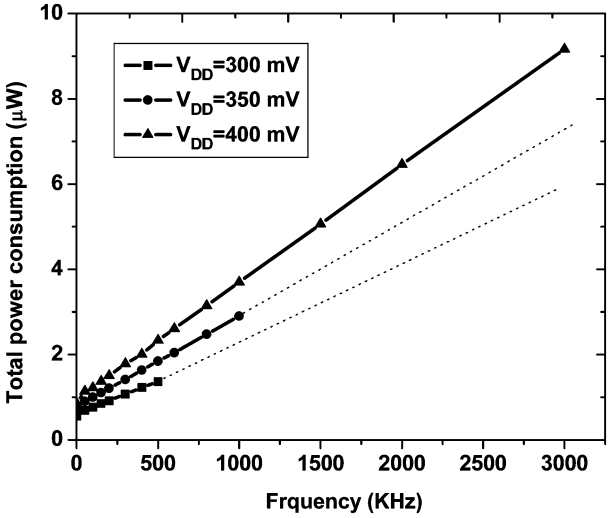

Fig. 18. Measured total power consumption versus operation frequency.

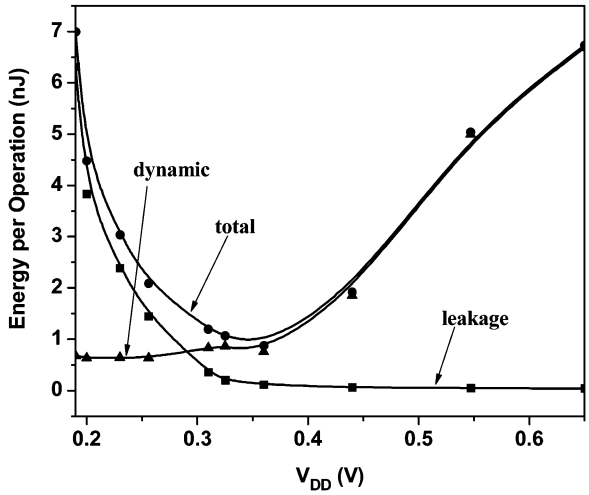

Fig. 19. Measured test chip energy consumption per operation.

at 1-MHz clock rate and 310-mV power supply. At higher supply voltages, the memory circuits are faster and the total power consumption is increasingly dominated by the dynamic component. The foundry process used for fabrication exhibits relatively low leakage and has low DIBL, as evident. It also has negligible gate leakage current. This leakage component will be greatly reduced by the low $V_{\rm DD}$ for processes with high gate leakage, e.g., sub-130-nm processes. At $V_{\rm DD} < 270$ mV the leakage power exceeds the dynamic power. The measured dynamic power exhibits the expected quadratic relationship with the power supply voltage. Fig. 18 shows the power is linear with frequency at a fixed voltage, also as expected. The measurements, taken at three $V_{\rm DD}$ values, have different Y axis intercepts due to the voltage dependence of the leakage currents.

E. Energy per Operation

A circuit's efficiency is usually defined by the energy consumed for each operation, i.e.,

$$E = \frac{P_{\rm total}}{f}. \tag{12}$$

This figure of merit is interesting as it can define the point where the most computation can be performed at the least total energy, assuming that time to complete the computation is not a constraint. This minimum energy operating point is thus important for systems that must maximize battery lifetime in lieu of other constraints. Fig. 19 shows the subthreshold memory energy per operation at a number of voltages. Again, the memory is operated at the $f_{\rm max}$ for each voltage.

The total energy consumption per operation includes both the leakage and dynamic components. The figure shows that the leakage component is exponentially related to the supply voltage. The dynamic component is $\alpha C V_{\rm DD}^2$. Of course, α is very low in the memory, given that one in 128 GWLs is asserted and one in eight GBLs is discharged per cycle on average. This makes the leakage component more important in memories than in logic. The minimum $V_{\rm DD}$ for the summation of those two opposite, monotonic curves describing the leakage and active power components, is below 350 mV.

VI. Conclusion

Operating with the supply voltage below the threshold voltage is the most effective method to produce circuits for ultra-low-power applications. However, such low voltages create difficulties, particularly for memory design, since the ratio of $I_{\rm ON}/I_{\rm OFF}$ is greatly reduced in high fan-in/out circuits such as bitlines. The increased sensitivity to PVT variations that subthreshold circuits exhibit will require substantial design margin to obtain high yields. Here, lowering the minimum operation voltage as much as possible by design is required to maximize the design margin while still operating at very low $V_{\rm DD}$. An analytical model to determine the onset of positive noise margin with respect to $V_{\rm DD}$ and circuit fan-out was outlined and used as the basis for the memory design.

Subthreshold memory design requires unconventional design approaches. A number of applicable circuit and micro-architecture level techniques have been described here to reduce fan-in/out and address the poor drive currents afforded in subthreshold. These include hierarchical memory organization, reduced fan-in by combining cell outputs, and self-timed keeper controls. The self-timed scheme allows extensive use of dynamic circuits while allowing safe, pseudo-static operation at extremely low operating voltages. The techniques are suitable for above threshold supply voltages where circuits exhibit high leakage, such as very high performance or high operating temperature circuits. This has allowed a memory using dynamic read and relatively high density as opposed to previous single power supply subthreshold approaches [2]. The use of multiple memory supply voltages has been proposed to limit leakage power with a subthreshold voltage, while avoiding stability compromise by reading at higher voltages, using conventional six-transistor SRAM cells [31] and has been shown to achieve higher density [32].

A 512×13 bit subthreshold memory fabricated on a 130-nm process technology was tested to be fully functional at 190 mV with 28-kHz clock frequency. The speed and array efficiency is much improved over that of the subthreshold memory design presented in [2]. Single bits can work as low as 129 mV. The memory achieves a 1-MHz clock rate with a 310-mV power supply, and consumes 1.196 μW at that voltage. The memory

consumes 1 nJ of energy per operation in laboratory measurements at room temperature, or less than 77 fJ of energy per bit, at a 345-mV supply voltage.

ACKNOWLEDGMENT

The authors would like to thank AFRL for test chip fabrication and Arizona State University Flexible Display Center for test chip measurement assistance. They also thank P. Eaton at Micro-RDC for assistance with the FPGA test board and programming.

REFERENCES

[1] H. Soeleman, K. Roy, and B. Paul, "Robust subthreshold logic for ultra-Low power operation," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 1, pp. 90–99, Feb. 2001.

[2] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[3] J. B. Burr and A. M. Peterson, "Ultra low power CMOS technology," in *Proc. NASA VLSI Design Symp.*, 1991, pp. 4.2.1–4.2.13.

[4] A. Bryant *et al.*, "Low-power CMOS at Vdd = 4kT/q," in *Proc. Device Research Conf.*, 2001, pp. 22–23.

[5] R. Swanson and J. Meindl, " Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. SC-7, no. 2, pp. 146–153, Apr. 1972.

[6] A. Chandrakasan, S. Sheng, and R. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 473–84, Apr. 1992.

[7] A. Chandrakasan and R. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE*, vol. 83, no. 4, pp. 498–523, Apr. 1995.

[8] B. H. Calhoun and A. Chandrakasan, "Characterizing and modeling minimum energy operation for subthreshold circuits," in *Proc. Int. Symp. Low-Power Electronics and Design*, Aug. 2004, pp. 90–95.

[9] D. Lee, D. Blaauw, and D. Sylvester, "Gate oxide leakage current analysis and reduction for VLSI circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 2, pp. 155–166, Feb. 2004.

[10] A. Ricadela, "Sensors everywhere—Tiny, wireless sensors may be able to track anything, anytime, anywhere," *InformationWeek.com*, Jan. 27, 2005.

[11] T. Starner, "Human-powered wearable computing," *IBM Sys. J.*, vol. 35, pp. 618–629, 1996.

[12] L. Geddes, "Historical highlights in cardiac pacing," *IEEE Eng. Medicine Biol.*, vol. 9, no. 2, pp. 12–18, Jun. 1990.

[13] P. Pentland *et al.*, " The digital doctor: an experiment in wearable telemedicine," in *Proc. 1st Int. Symp. Wearable Computers*, 1997, pp. 173–174.

[14] H. Benz *et al.*, "Low power radiation tolerant VLSI for advanced spacecraft," in *Proc. IEEE Aerospace Conf.*, 2002, vol. 5, pp. 5–2401–5–2406.

[15] J. Chen, L. Clark, and Y. Cao, "Maximum fan-in/out: Ultra-low voltage circuit design in the presence of variations," *IEEE Circuits Devices Mag.*, vol. 21, no. 1, pp. 12–20, Jan./Feb. 2005.

[16] Y. Taur and T. Ning, *Fundamentals of Modern VLSI Devices*. New York: Cambridge Univ. Press, 1998.

[17] S. Sze, *Modern Semiconductor Device Physics*, 3 ed. New York: Wiley, 1998.

[18] N. Weste and D. Harris, *CMOS VLSI Design: A Circuit and System Perspective*, 3 ed. Reading, MA: Addison Wesley, 2005.

[19] E. Seevinck, F. List, and J. Lohstroh, "Static noise margin analysis of MOS SRAM cells," *IEEE J. Solid-State Circuits*, vol. SC-22, no. 5, pp. 748–754, Oct. 1987.

[20] R. Krishnamurthy *et al.*, "A 130-nm 6-GHz 256 × 32 bit leakage-tolerant register file," *IEEE J. Solid-State Circuits*, vol. 37, no. 5, pp. 624–632, May 2002.

[21] A. Chandrakasan, W. Bowhill, and F. Fox, *Design of High-Performance Microprocessor Circuits*. New York: IEEE Press, 1998.

[22] C. Kim *et al.*, "A process variation compensating techniques for sub-90 nm dynamic circuits," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2003, pp. 205–206.

[23] S. Hsu *et al.*, "A 90 nm 6.5 GHz 256 × 64 b dual supply register file with split decoder scheme," in *Symp. VLSI Circuits Dig. Tech. Papers*, 2003, pp. 237–238.

[24] L. T. Clark, "A high-voltage output buffer fabricated on a 2 V CMOS technology," in *Symp. VLSI Circuits Dig. Tech. Papers*, 1999, pp. 61–62.

[25] S. Imai, N. Nakanishi, Y. Suzuki, and K. Umeda, "Low-power consumption level-shifter used clamping circuit technique and LTPS technology for TFT-LCD," in *Proc. Int. Symp. Intelligent Signal Processing and Communication Systems*, Nov. 2004, pp. 792–795.

[26] W. Wang, M. Ker, M. Chiang, and C. Chen, "Level shifter for high-speed 1-V to 3.3-V interfaces in a 0.13-μm CU-interconnection/low-K CMOS technology," in *Proc. Symp. VLSI Technology, System and Applications*, 2001, pp. 307–310.

[27] T. Chen, J. Chen, and L. Clark, "Subthreshold to above threshold level shifter design," *ASP J. Low Power Electron.*, submitted for publication.

[28] L. Clark and F. Ricci, "Low standby power state storage for sub-130-nm technologies," *IEEE J. Solid-State Circuits*, vol. 40, no. 2, pp. 498–506, Feb. 2005.

[29] A. Alvandpour *et al.*, "A sub-130-nm conditional keeper technique," *IEEE J. Solid-State Circuits*, vol. 37, no. 5, pp. 633–638, May 2002.

[30] V. Kursun and E. Friedman, "Domino logic with variable threshold voltage keeper," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 6, pp. 1080–1093, Dec. 2003.

[31] K. Flautner *et al.*, "Drowsy caches: Simple techniques for reducing leakage power," in *Proc. Int. Symp. Computer Architecture*, 2002, pp. 148–157.

[32] M. Yamaoka *et al.*, "0.4-V logic-library-friendly SRAM array using rectangular-diffusion cell and delta-boosted-array voltage scheme," *IEEE J. Solid-State Circuits*, vol. 39, no. 6, pp. 934–940, Jun. 2004.

Jinhui Chen (S'03) received the B.Sc. degree from the Department of Electrical Engineering, Hefei University of Technology, China, in 1996, and the M.S. degree in electrical engineering from the University of New Mexico, Albuquerque, in 2004. He is currently working toward the Ph.D. degree in electrical engineering at Arizona State University, Tempe.

His research interests include ultra-low-voltage and ultra-low-power VLSI design for handheld devices, VLSI architecture and variations, and transient EM modeling for high-speed integrated circuits.

Lawrence T. Clark (M'90–SM'01) received the B.S. degree in computer science from Northern Arizona University, Flagstaff, in 1984, and the M.S. and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, in 1987 and 1992, respectively.

He worked at Intel Corporation in 1982 and 1984–1985 in product and test engineering and at VLSI Inc. from 1990 to 1992 in chipset design. From 1992 to 2003, he worked at Intel Corporation in various capacities including microprocessor design, participating in Pentium, Itanium, and XScale processor designs, compact modeling for circuit simulation, and CMOS imager design. Most recently, he was a Principal Engineer and Circuit Design Manager for XScale Microprocessors. In 2003, he joined the Department of Electrical and Computer Engineering, University of New Mexico, as an Associate Professor. In 2004, he joined the Department of Electrical and Computer Engineering, Arizona State University. He has been awarded over 50 patents and has approximately 15 pending. His research interests are circuits, architectures, computer-aided design, and radiation hardening for high-performance and low-power VLSI systems.

Tai-Hua Chen (S'02) received the B.S. degree in electrical engineering from Yuan-Ze University, Taiwan, in 1999 and the M.S. degree in electronics engineering from National Yunlin University of Science and Technology, Taiwan, in 2001. He is currently working toward the Ph.D. degree at Arizona State University, Tempe.

His research interests include low-power VLSI design techniques, radiation hardened by design circuits, and analog circuit design.