Low-Power SRAM Design Using Half-Swing Pulse-Mode Techniques

Kenneth W. Mai, Toshihiko Mori, Bharadwaj S. Amrutur, Ron Ho, Bennett Wilburn,
Mark A. Horowitz, Isao Fukushi, Tetsuo Izawa, and Shin Mitarai

***Abstract*—This paper describes a half-swing pulse-mode gate family that uses reduced input signal swing without sacrificing performance. These gates are well suited for decreasing the power in SRAM decoders and write circuits by reducing the signal swing on high-capacitance predecode lines, write bus lines, and bit lines. Charge recycling between positive and negative half-swing pulses further reduces the power dissipation. These techniques are demonstrated in a 2-K × 16-b SRAM fabricated in a 0.25-μm dual-V_t CMOS technology that dissipates 0.9 mW operating at 1 V, 100 MHz, and room temperature. On-chip voltage samplers were used to probe internal nodes.**

***Index Terms*—Low power, low voltage, memory architecture, self-timing, SRAM chips.**

I. Introduction

CONSIDERABLE attention has been paid to the design of low-power, high-performance SRAM's since they are a critical component in both hand-held devices and high-performance processors. Current SRAM's routinely apply a number of low-power techniques [1] and have achieved power dissipations in the milliwatt range [2]–[4]. This paper extends these methods to include the use of half-swing signals and shows that applying this reduced swing signalling to the decoder and write circuitry can further reduce the power of an SRAM significantly [5].

Low-power, high-performance SRAM's require optimization of the architecture, circuits, and technology. At the architectural level, the key goals are localizing signals to reduce the capacitance that switches, reducing signal swings, and eliminating any dc currents. Partitioned memory arrays and hierarchical word lines reduce the total capacitance that is switched per access [6]. Using clocked sense amplifiers eliminates the sense amplifiers as sources of dc currents. Pulsed word lines keep the swing of the bit lines during reads to the minimum needed for sensing [7].

At the circuit level, designers use pulse-mode circuits to improve performance and generate the pulses that are needed to satisfy the architectural demands. A pulse-mode, self-resetting gate can be made faster than a normal static CMOS gate, since the forward path can be optimized for a single transition (like a dynamic gate) and the reset transition can be handled by the separate self-reset signal path [8]. In function blocks like decoders, where very few of the total gates transition per cycle, self-resetting gates dissipate much less power than precharged logic families, since they do not need a global clock for resetting.

Technology optimization is the final tool that a designer can use to produce low-power, high-performance SRAM's. One of the most effective techniques to reduce power dissipation is to reduce the supply voltage. However, to maintain reasonable performance, the transistor threshold voltage must also be lowered, which causes the subthreshold cell leakage current to become a significant source of power dissipation. Two proposed techniques to combat this problem are multiple-threshold CMOS (MT-CMOS) [9] and variable-threshold CMOS (VT-CMOS) [10]. MT-CMOS uses different thresholds for the devices used in the cells and those used in the decode and peripheral logic. The high-V_t devices are used in the cells to prevent significant leakage currents, while the low-V_t devices are used in the decode and peripheral logic to provide good performance. High-V_t devices can be selectively used outside of the cell array to reduce the leakage currents further by acting as power-supply switches that are turned off during standby mode. VT-CMOS controls the transistor thresholds by varying the bias of the well(s) and/or substrate. The threshold is increased when the part enters standby mode to reduce the leakage current.

While these techniques have greatly reduced the power needed to read data from the memory, the write power and decoder power have not been reduced as quickly. The rest of the paper will show how using half-swing techniques can reduce this power. Section II introduces the half-swing pulse-mode gate that combines level-conversion and logical AND functionality. Section III briefly reviews decoder design and shows how the new gate can be used to reduce decoder power while maintaining high performance. Section IV describes the write path in an SRAM and again shows how power can be saved by using half-swing signals. Much of the power savings comes from operating the bit lines from $V_{dd}/2$ rather than V_{dd}. This section also discusses noise-margin issues for the memory cell. To demonstrate the feasibility of these techniques, a small prototype memory was built and is described in Section V. That section will also include a brief description of the on-chip samplers that were used to simplify measurement of on-chip waveforms.

Manuscript received April 7, 1998; revised June 8, 1998. This work was supported by the Advanced Research Projects Agency under Contract J-FBI-92-194 and by Fujitsu, Ltd.

K. W. Mai, B. S. Amrutur, R. Ho, B. Wilburn, and M. A. Horowitz are with the Center for Integrated Systems, Stanford University, Stanford, CA 94305 USA (e-mail: demon@chroma.stanford.edu).

T. Mori is with the Center for Integrated Systems, Stanford University, Stanford, CA 94305 USA and Fujitsu Laboratories, Ltd., Kanagawa Japan.

I. Fukushi is with Fujitsu Laboratories, Ltd., Kanagawa 211-8588 Japan.

T. Izawa and S. Mitarai are with Fujitsu, Ltd., Kanagawa 211-8588 Japan.

Publisher Item Identifier S 0018-9200(98)07046-2.

0018–9200/98$10.00 © 1998 IEEE

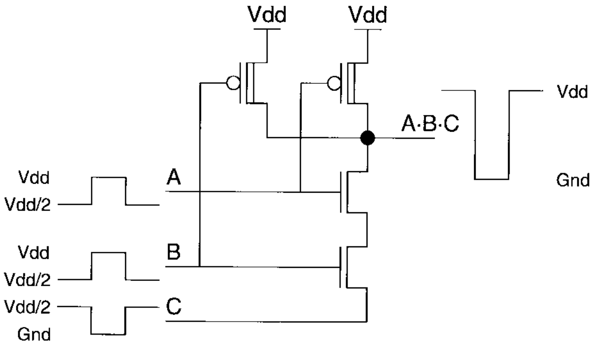


Fig. 1. Half-swing pulse-mode AND gate.

II. A Half-Swing Pulse-Mode Gate Family

The typical disadvantage of using reduced swing signals is the need for level-conversion and/or reduced gate overdrive at the receiving gates [11]–[16], which causes a loss of performance. However, if positive half-swing (swinging from the rest state of $V_{dd}/2$ to V_{dd} and back to $V_{dd}/2$) and negative half-swing (swinging from the rest state of $V_{dd}/2$ to Gnd and back to $V_{dd}/2$) pulses are combined with the receiver-gate logic style shown in Fig. 1, all of the forward transition driving transistors see a full gate overdrive, and the effect of the low swing inputs on the receiver performance is negligible. Combined with self-resetting techniques, this provides an interesting opportunity to use a half-swing pulsed signalling scheme. We assume the existence of a $V_{dd}/2$ supply voltage, whose generation is discussed in Section V.

The gate in Fig. 1 merges the voltage-level conversion (from half-swing to full-swing) [17] with a logical AND operation. It performs the logical AND of two positive pulse signals and one negative pulse signal (asserted when the signal is at Gnd). Its structure is similar to a standard two-input static-CMOS NAND gate using low-V_t PMOS and high-V_t NMOS, except that there is a third input that is sent to the source of the lower NMOS transistor. The two gate-input signals are positive pulses, while the source-input signal is a negative pulse.

In the select case, when all of the pulses are asserted (A, $B = V_{dd}$ and $C = 0$), the NMOS transistors have a full V_{dd} across their gate to source, the PMOS transistors are off, and so the output is pulled low. There is negligible performance loss due to the use of half-swing pulsed inputs, since the NMOS transistors have a full $V_{dd} - V_t$ gate overdrive. When none of the pulses are asserted (A, B, $C = V_{dd}/2$), the pulldown stack is off, the PMOS transistors are on (assuming that the $|V_t|$ of the PMOS devices is less than $V_{dd}/2$), and so the output is pulled high. In the half-select cases when some, but not all, of the input pulses are asserted, the gate output also remains high. The output then is a full-swing negative pulse and implements the AND of the three inputs.

The two disadvantages of this gate are the reduced noise margin and the degraded speed of the output reset transition. The noise margin is reduced in the worst case half-select cases shown in Fig. 2. Fig. 2(a) shows the case in which both positive pulses are asserted (A, $B = V_{dd}$) but the negative pulse is not ($C = V_{dd}/2$). Both of the PMOS devices are off, and if $V_{dd}/2$ is greater than the high-V_t of the NMOS, then the pulldown network is conducting, and the output will eventually settle to a value determined by the leakage currents of the NMOS and PMOS networks. This case is mitigated somewhat by the V_t of lower NMOS being boosted by the body effect.

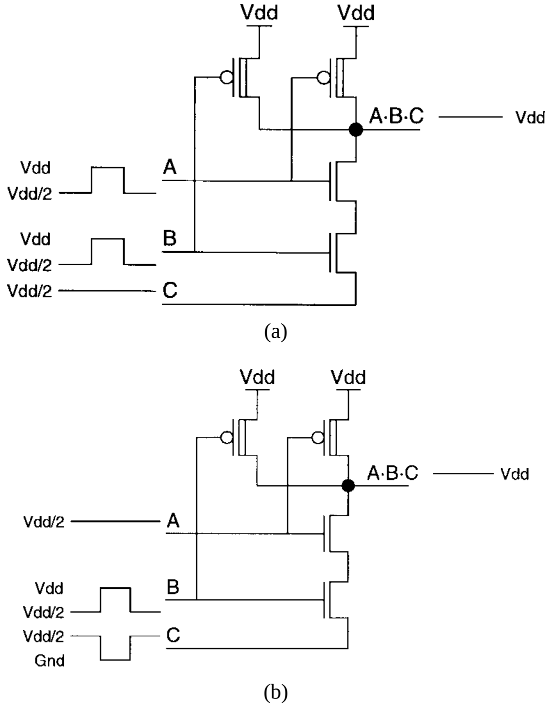


Fig. 2. Half-select cases. (a) No negative pulse. (b) No positive pulse.

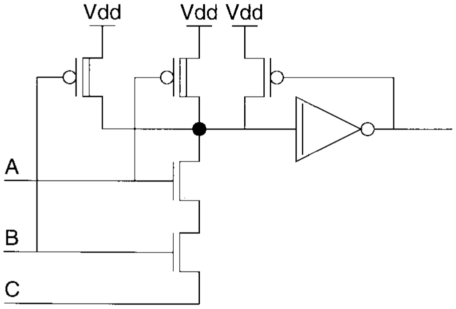


Fig. 3. Half-swing pulse-mode gate with PMOS leaker.

Fig. 2(b) shows the other worst case half-select case, where only one of the positive pulses is asserted ($A = V_{dd}/2$, $B = V_{dd}$) and the negative pulse is asserted ($C = 0$). The lower NMOS has a full $V_{dd} - V_t$ gate overdrive and is fully on. Assuming that the upper NMOS's source node has been pulled to Gnd, it has a $V_{dd}/2 - V_t$ gate overdrive and will be weakly on. However, one of the low-V_t PMOS devices is also on, since one of the positive pulses is not asserted, and this fights any leakage current in the pulldown network.

To increase the noise margins, a small high-V_t PMOS leaker feedback transistor is added (see Fig. 3). This PMOS leaker device increases noise margins by fighting leakage currents in the pulldown network, just as such a device does in precharged logic families. Even with the leaker device, this gate has smaller noise margins than a conventional gate, and noise coupling into its inputs must be carefully managed.

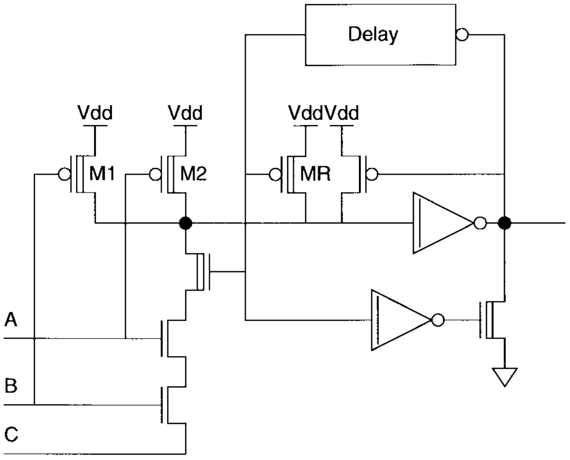


Fig. 4. Self-resetting half-swing pulse-mode gate.

In addition to degrading the noise margin, leakage currents in the half-select cases will contribute to the power dissipation. However, the leakage only exists during the brief time that the pulses are asserted. Also, only the gates that experience one of the half-select cases can potentially cause leakage. This additional power is small when V_{dd} is around two times high-V_t, since $V_{dd}/2$ is then very close to high-V_t. Even for larger V_{dd}, the added half-select leakage power is not a significant fraction of the total power for the test chip (see Section V).

The second disadvantage of this gate is the slow output reset transition. The transistor sizes in the gate are skewed to accelerate the forward output assert transition. The high-V_t NMOS transistors in the gate are large, while the low-V_t PMOS are small. The transistor sizes in the low-V_t inverter are also skewed to accelerate the output assertion edge. However, this sizing slows the output reset transition. The half-swing signalling scheme further slows the output reset, since the low-V_t PMOS only have $V_{dd}/2$ across their gate to source when none of the input pulses is asserted.

To speed up the output reset transition, a self-resetting technique is employed, as is shown in Fig. 4. A fixed delay after the output assertion edge, the low-V_t PMOS reset device (MR) is turned on to restore the output of the half-swing pulse-mode gate to a high value. MR sees a full gate overdrive and can be sized to be relatively large since it does not strongly affect the forward output assertion transition speed. When the gate is in the rest state, waiting for input pulses to arrive, MR is off. Thus, only the extra diffusion capacitance of MR's drain affects the forward output assertion speed, but MR greatly accelerates the reset transition. The PMOS devices in the gate (M1 and M2) are only used to hold the output high after the reset pulse has been deasserted. An optional low-V_t NMOS device can be added to the pulldown stack to ensure that there is no fighting if the leading edge of the reset pulse arrives before the input pulse has been deasserted.

The pulldown network of this half-swing pulse-mode logic style can be arbitrarily complex, as long as all of the branches terminate at the one negative pulsed source input. There cannot be more than one negative pulse input to a source, since this could produce a case where there would be fighting between the $V_{dd}/2$ and Gnd supplies. The possible functions that this type of gate is capable of generating are thus restricted to those with a final outer AND term with the negative pulse input.

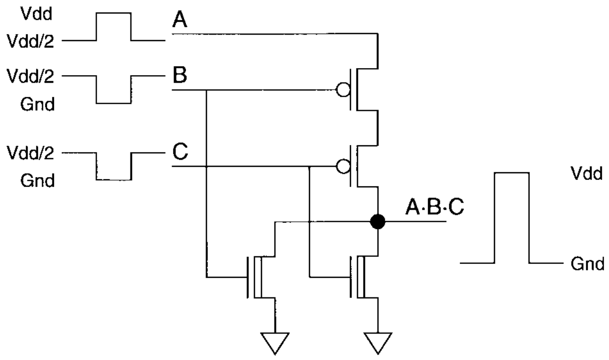


Fig. 5. PMOS-style half-swing pulse-mode gate.

Similar to the gate shown in Figs. 2 and 3, a gate where a positive pulse is sent into the source of a high-V_t PMOS and all the other inputs are negative pulses going into the gates of high-V_t PMOS devices and low-V_t NMOS devices could be built (Fig. 5). The disadvantage of such a gate is that the PMOS devices are inherently inferior to the NMOS ones, and so from a performance standpoint, it is preferable to use the NMOS-type gate. The output of a PMOS-style gate is a full-swing positive pulse, and in some cases, such as the write amplifier described in Section IV, this type of output is needed.

In addition to reducing the signal swing on the input lines, using both positive and negative half-swing pulses can further reduce power dissipation by taking advantage of charge recycling [12], [17]. The charge used to produce the assert transition of a positive pulse can also be used to produce the reset transition of a negative pulse. If the capacitances of the positive and negative pulses match, then no current would be drawn from the $V_{dd}/2$ supply. In practice, the capacitances of the pulsed lines could be designed to nominally match, and the $V_{dd}/2$ supply would compensate for any mismatches. Recycling charge between the two types of pulses reduces the current drawn from the $V_{dd}/2$ supply, and thus the efficiency of the $V_{dd}/2$ supply generation is not critical to the overall power dissipation. Generation of the $V_{dd}/2$ supply is addressed in Section V.

If the half-swing input lines are high-capacitance, high-activity lines, then the power savings can be significant. The potential power savings can be quantified using a simple example consisting of two wire-load-dominated, high-capacitance lines, each with a capacitance of C. For a conventional pulsed design, both lines pulse full-rail each cycle. The power dissipation to drive the lines is then $2*C*V_{dd}*V_{dd}*f$. If the lines are reduced to half-swing, but the charge to swing the lines is still drawn from the V_{dd} supply, the power dissipation is $2*C*V_{dd}*(V_{dd}/2)*f$, a 50% power savings. But if the charge is also recycled between the two lines (one a positive half-swing pulse and one a negative half-swing pulse), then the power dissipation is $C*V_{dd}*(V_{dd}/2)*f$, since the charge to reset the negative pulse is essentially free, because it is the recycled charge from the positive pulse line resetting. Thus, the theoretical power savings is 75%. The overall power savings

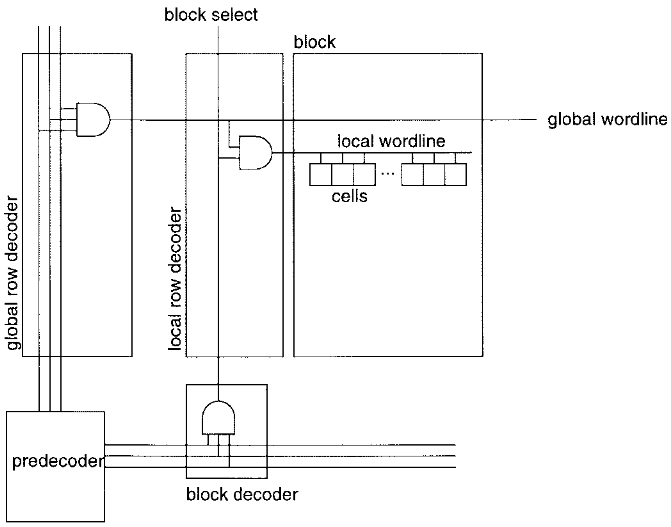


Fig. 6. Decoder structure.

for a block of logic depends on the percentage of the total capacitance that transitions that can be converted to half-swing pulses.

III. Decoders Using Half-Swing Pulse-Mode Gates

The decoder selects the row of cells in the array to access according to the address input. To reduce the area and improve performance, the decode is done in multiple stages, sharing gates that generate common terms. Fig. 6 shows the decoder structure for a typical partitioned memory array with hierarchical word lines. The predecoder outputs are shared among the global row decoders and block decoders throughout the array, and hence the predecoder outputs are long, high-capacitance lines. By using half-swing pulse-mode gates in the global row decoders and block decoders, the predecoder outputs can be reduced to half-swing signals, thus providing a significant power savings with negligible performance loss. Additionally, the global word lines themselves are long, high-capacitance lines. So using half-swing pulse-mode gates in the local row decoders allows the global word lines to be half-swing signals also, thus saving even more power.

As noted in Section I, low-power, high-performance SRAM's typically use pulse-mode self-resetting gates in the decoder. Standard pulse-mode, self-resetting gates can be easily converted to generate positive and negative pulse outputs by changing the voltage supplies that drive the final inverter. For a gate that generates a positive pulse, the final inverter is driven from V_{dd} and $V_{dd}/2$ instead of from V_{dd} and Gnd. For a gate that generates a negative pulse, the final inverter is driven from $V_{dd}/2$ and Gnd. This is illustrated for the negative pulse case in Fig. 7. Since the source of M1 is at $V_{dd}/2$, it only has $V_{dd}/2$ across its gate to source when it is on. To obtain a reasonably fast output reset edge, M1 must be made large, but this would slow the forward output assert transition. Adding an explicit reset device (M2) driven by the self-resetting path circumvents this problem and allows M1 to remain small while maintaining both a fast assert and reset edge on the output.

On the receiving end, the half-swing pulse-mode gates have an inherently lower noise margin than standard logic families, as detailed in Section II. Thus, cross talk and noise injection are potential problems for the long, high-capacitance input lines, since they run in parallel to other lines for a long distance. We propose two simple layout techniques (Fig. 8) for mitigating these problems. First, if an equal number of positive pulse and negative pulse lines are available, they can be interleaved [Fig. 8(a)] to reduce the detrimental noise coupling. Thus, for any given line, the two adjacent lines can only inject noise in the opposite direction from the direction that the given line transitions. For example, on an inactive positive pulse line, if it has noise injected into it from adjacent negative pulse lines, this is not a problem, since this noise can only be in the negative direction and only positive direction noise can cause the subsequent receiver gate to fire incorrectly. The disadvantage of this technique is that in the worst case, where a negative and a positive pulse are asserted in adjacent lines, the bus speed is degraded.

Another technique is to twist the lines in a one-hot bus to minimize the worst case noise coupling. If a large number of lines of the same pulse type are laid out in parallel, then each wire couples very strongly to its two adjacent neighbors. However, if the lines are twisted multiple times in the layout in such a way that they each capacitively couples to each of the other lines in the bus by the same amount, then the worst case coupling between any two lines is reduced significantly. Twisting requires the use of an extra metal layer for the jumps over/under the other lines. A twisting scheme for an eight-wire bus is shown in Fig. 8(b). In general, the twisting can reduce the worst case coupling capacitance of one line to another by a factor of $n/2$ for an n-bit-wide bus using n twists.

For the decoder to take advantage of charge recycling, the capacitance of negative and positive pulse lines must be as closely matched as possible. In general, this is not too difficult, since there are a large number of predecoder outputs that drive roughly the same capacitance. In the prototype design, the charge recycling balancing was done in a directed trial-and-error manner. From simulation data, we noted that the half-swing pulse-mode gates operate slightly faster if the negative pulse arrives ahead of the positive pulses. This is due to the sizing of the devices (small low-V_t PMOS and large high-V_t NMOS) and the body-effected V_t of the bottom NMOS device. So the signals that were generated early tended to be chosen to be negative pulses.

In addition to matching the actual capacitance value of the positive and negative signals, the types of parasitic capacitances (wire, diffusion, and gate) that make up the two capacitances should also be matched as closely as possible, since process skews can be negatively correlated across the various types of capacitances. For example, if a line that has mostly wire capacitance is matched with another line with mostly diffusion capacitance, a process variation may upset the charge balancing by increasing the diffusion capacitance while not affecting (or even decreasing) the wire capacitance.

The global row decoders demonstrate that half-swing signalling can be used throughout a multistage logic operation where long, high-capacitance lines separate the stages. The global row decoders take in half-swing signals and in turn generate the negative half-swing global word lines. The global

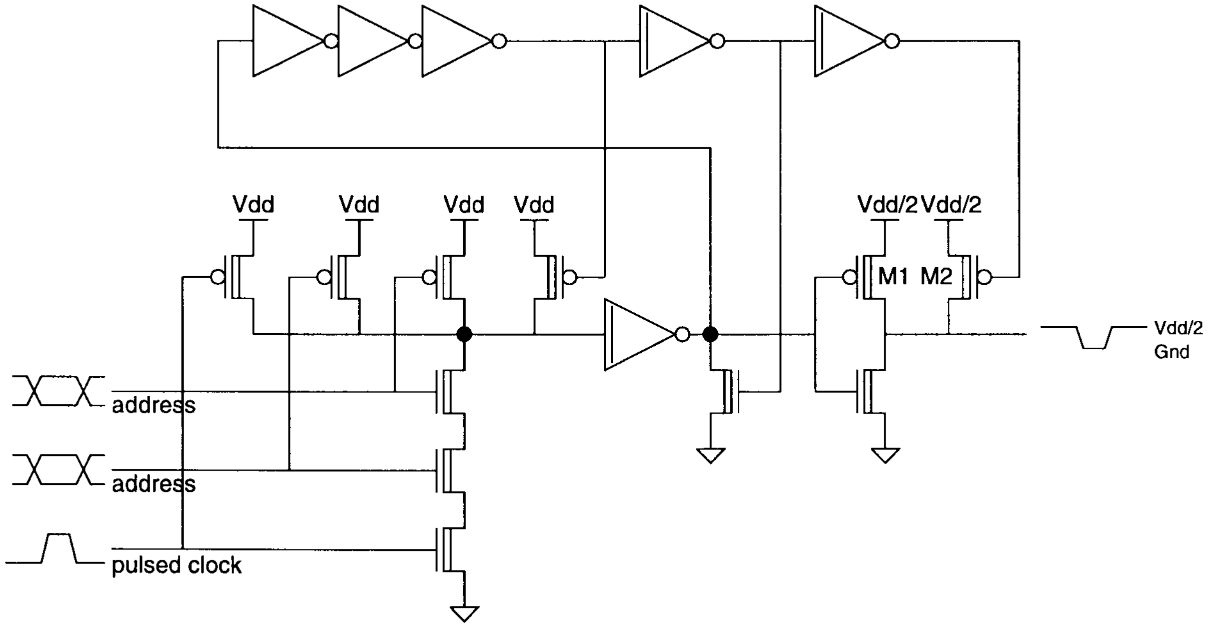
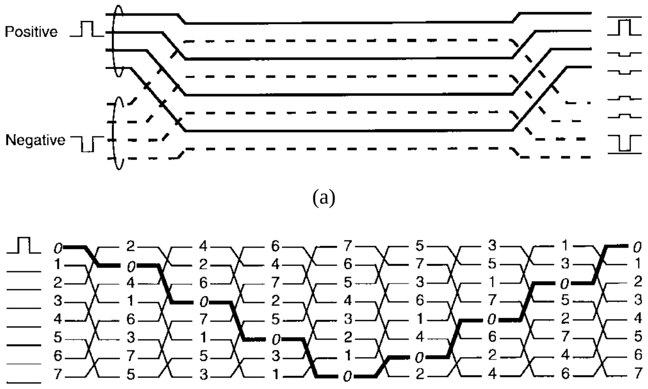


Fig. 7. Predecoder gate with negative half-swing pulse output.



(a)

(b)

Fig. 8. Layout techniques for lowering noise. (a) Interleaving. (b) Twisting.

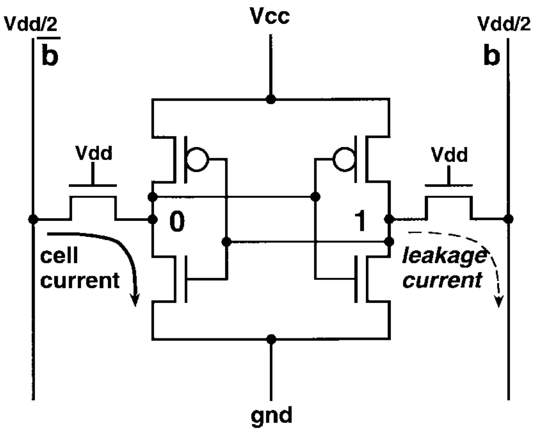


Fig. 9. Cell read with $V_{dd}/2$ bit lines.

row decoders in the test chip are not self-resetting and instead rely on the input pulses to reset the output via the low-V_t PMOS transistors in the gate. In simulation, the global word-line reset transition was adequately fast, but as can be seen in Fig. 19(a), this proved to be untrue in the actual silicon.

The local row decoders combine the negative half-swing global word-line signals with the full-swing block select signal to generate the local word lines. The local word-line pulse width is controlled to obtain the minimum bit-line swings needed for proper sensing by controlling the pulse width of the block select through a replica bit line [7]. Since the local word-line reset transition should come from the block select to properly control the pulse width, the local row decoders are not self-resetting. Making the block select a full-swing signal accelerates the local word lines reset transition, since the low-V_t PMOS in the local row decoders have full gate overdrive when the block select is low. The block decoders are self-resetting, using the replica bit line as the delay element to set the block select (and thus the local word line) pulse width.

IV. Reducing the Write Power

In low-power embedded SRAM's, with large access widths, the write power can be significantly larger than the read power since the bit lines are referenced to V_{dd}, and during writes, they are discharged almost to ground. Thus, write power can be reduced by decreasing the bit-line swings during writes. Alowersson [18] proposed using a low reference voltage for the bit lines, and reducing the word-line voltage during reads to prevent cell instability. However, this technique slows the access time, because the read cell current is reduced. Instead, we propose a bit-line reference of $V_{dd}/2$, which enables us to reduce the bit-line swing during writes by half of the conventional technique.

However, using a $V_{dd}/2$ reference for bit lines can potentially lead to cell instability during reads. Leakage current from the high node degrades the high voltage and hence the drive strength of the driver device connected to the low node (Fig. 9). This problem can be solved by using a larger cell voltage (V_{cc}), but a potential disadvantage of any boosted

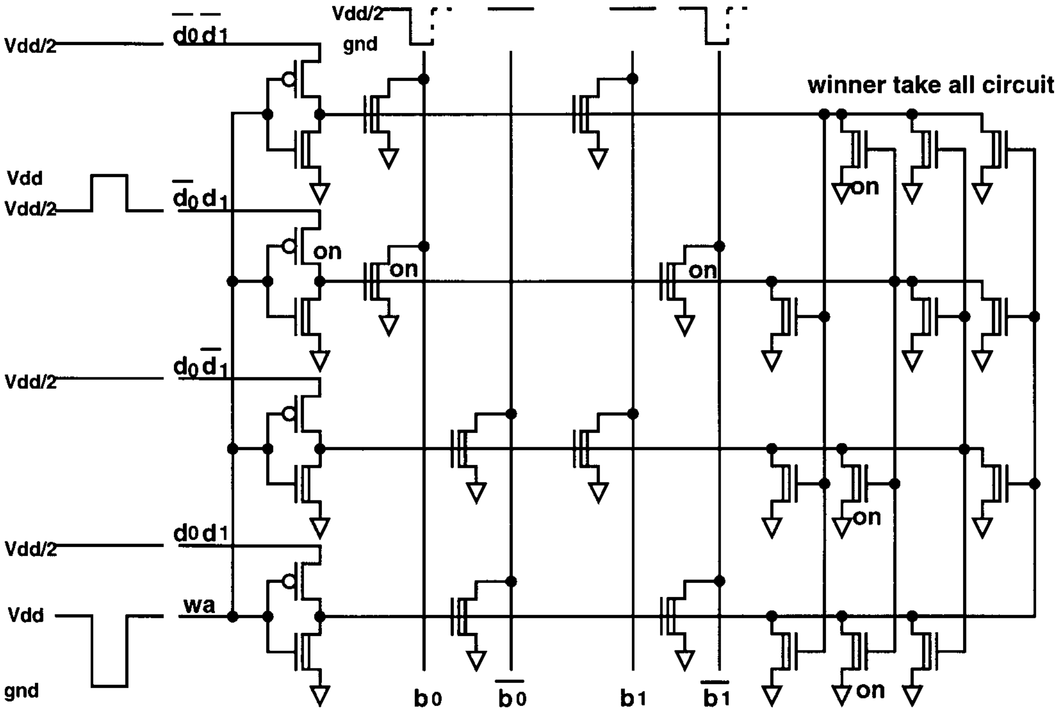


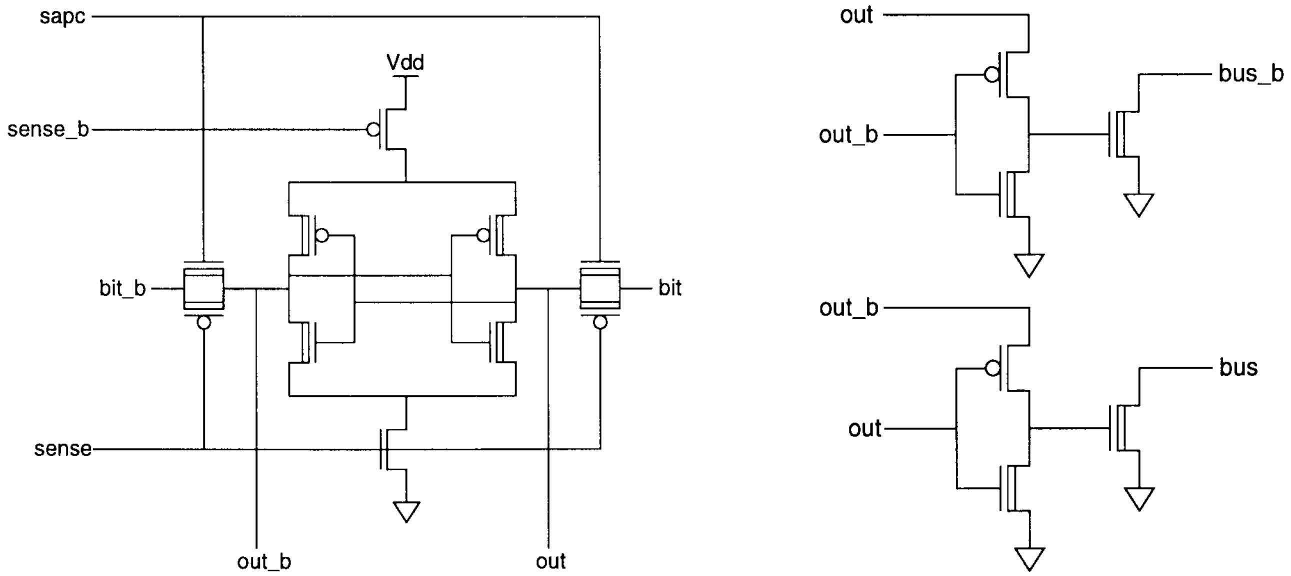
Fig. 10. Write amplifier.

cell voltage design is the degraded write margin under certain process skews. In a process skew where the pullup PMOS becomes stronger and the access NMOS becomes weaker, the high node might not be pulled down sufficiently to achieve a write. This problem can be solved by weakening the pullup PMOS. As a further safeguard, we also lower the cell voltage of the accessed row during writes, which incurs an area penalty of about 10% for the extra logic in the local row decoders needed to switch the cell supply voltage dynamically. Thus, the write margins can be maintained without altering the NMOS transistor sizes in the cell, and hence the cell area is not affected by the use of the $V_{dd}/2$ bit lines. Experimental results indicate no errors in the test-chip operation, even when $V_{cc} = V_{dd}$. If negatively correlated process skews for NMOS and PMOS are well controlled, then this dynamic V_{cc} scheme for writes can be eliminated in future designs.

Further power reduction is possible by observing that the write bit-line swings are like negative half-swing pulses and hence can be recycled with suitably matched positive half-swing pulses. Matching the capacitance of the bit lines with the write data bus can theoretically reduce the write power of the bit lines by 75% of the full-swing version. In our implementation, the write data-bus capacitance was about twice as large as the bit-line capacitance. To achieve capacitance matching, we employed the two-to-four encoding technique proposed in [19]. Every two bits of data is encoded as a one-hot-out-of-four signal. Compared to an unencoded differential signalling scheme, the activity factor is reduced by a factor of two, thus allowing for good matching between the bit-line capacitance and the write data-bus capacitance.

At the accessed block, the write data are decoded in the write amplifier shown in Fig. 10. A set of PMOS-style half-swing pulse-mode gates and an NMOS pulldown network decode the data and pull down the appropriate bit lines. To perform the write data decoding without any performance penalty, the encoded positive-pulse write data are combined with the full-swing negative-pulse signal "write amplify" (*wa*). The *wa* signal is generated locally at each block and is not a high-capacitance line, so the power savings for making it a half-swing signal would not be significant. Additionally, the full-swing *wa* signal prevents the write amplifiers in unselected blocks from having any leakage current, since all *wa* signals except the selected one are at V_{dd}. PMOS-style gates are used since a full-swing positive-pulse output is needed to directly drive the NMOS transistors that pull down the bit lines. The other part of the write amplifier is a winner-take-all circuit that suppresses any spurious transitions on the unselected lines. In combination with the full-swing *wa* signal, this ensures robust noise margins for the write amplifier.

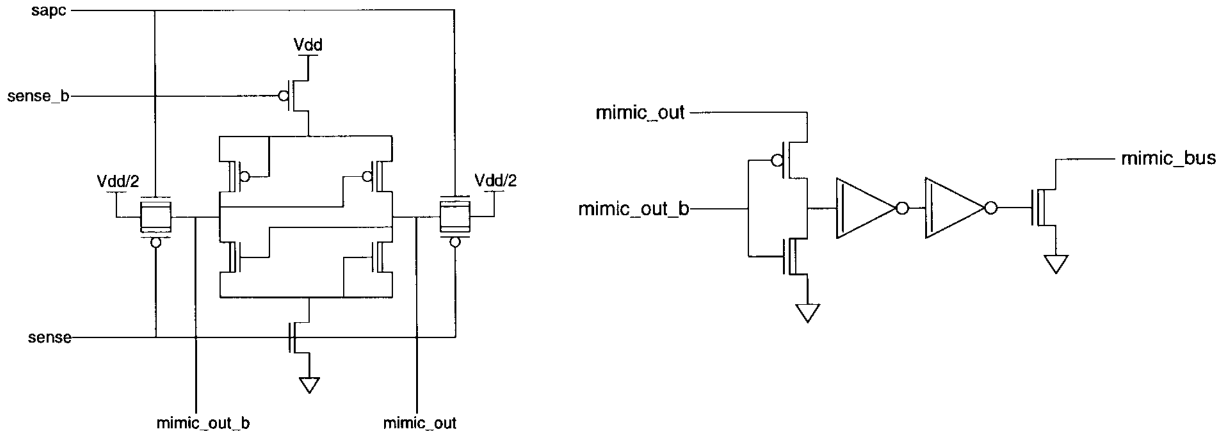
To properly sense the $V_{dd}/2$ referenced bit lines during reads, a latch-style sense amplifier [Fig. 11(a)] with both V_{dd} and Gnd cutoff devices is used. The latched data are transmitted from the block to the output via a differential, low-swing, precharged-before-use bus. The bus reference voltage is the supply V_{bus}, which is held at 250 mV independent of V_{dd}. The bus driver that immediately follows the sense amplifier is shown in Fig. 11(b). The low-swing signals are sent to a set of global clocked latch-style sense amplifiers. The sense enable signal for the global sense amplifiers is a full-swing timing pulse that is generated locally at the block and sent along with the data. This signal is generated by a sense-amplifier mimic circuit [Fig. 12(a)] and a bus-driver mimic circuit [Fig. 12(b)]. The sense-amplifier mimic is a sense-amplifier that is wired to fire always in the same direction.



(a)

(b)

Fig. 11. Read I/O circuits. (a) Sense amplifier. (b) Bus driver.



(a)

(b)

Fig. 12. Read mimic. (a) Sense-amplifier mimic. (b) Mimic bus driver.

V. Fabrication and Measured Results

A prototype 2-K × 16-b SRAM was designed and fabricated in a 0.25-μm dual-V_t CMOS process. The process and SRAM features are summarized in Table I. The chip was packaged in a 100-pin ceramic pin grid array. A die photo is shown in Fig. 13. The prototype is partitioned into four quadrants, each containing eight blocks, arranged as 64 rows × 16 columns of high-V_t 6T SRAM cells (Fig. 14).

Four different supply levels are required for our design, as indicated in Fig. 15. The external supply (V_{dd}) drives the full-swing circuits such as the decoders and the peripherals. It could also be used to generate three other levels on-chip, a boosted supply for the cell array (V_{cc}) via a charge pump, a $V_{dd}/2$ level ($V_{\rm mid}$) via a voltage regulator, and the low-swing read bus reference voltage ($V_{\rm bus}$) via a dc–dc voltage converter [20]. For the test chip, these voltages were supplied externally.

The measured read, write, and standby currents are shown in Table II for $V_{dd} = 1$ V, $V_{cc} = 1$ V, $V_{\rm mid} = 0.5$ V, and

TABLE I
PROCESS AND SRAM FEATURES

Technology	0.25µm dual-Vt CMOS, 3-metal, CMP planarized
Organization	2K x 16b
Supply voltage	1.0V
Clock rate	100MHz
Active power	0.9mW read 0.9mW write
Core size	2.0mm^2
Cell size	21.6µm^2

Transistor	Tox = 5.5nm, Co-Salicide
	low-Vtn = 0.1V
	high-Vtn = 0.45V
	low-Vtp = -0.1V
	high-Vtp = -0.45V

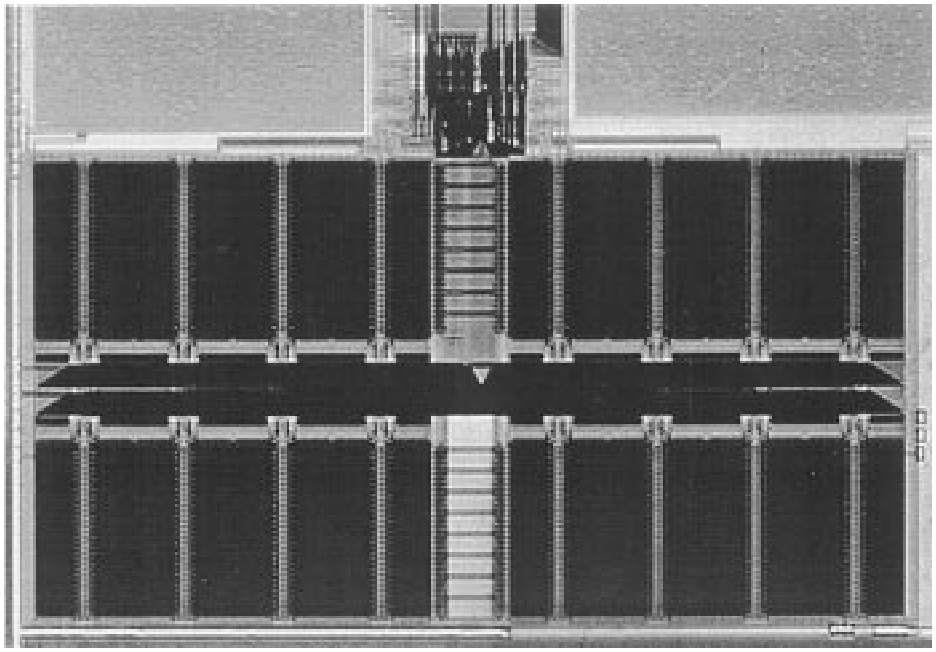


Fig. 13. Die photo.

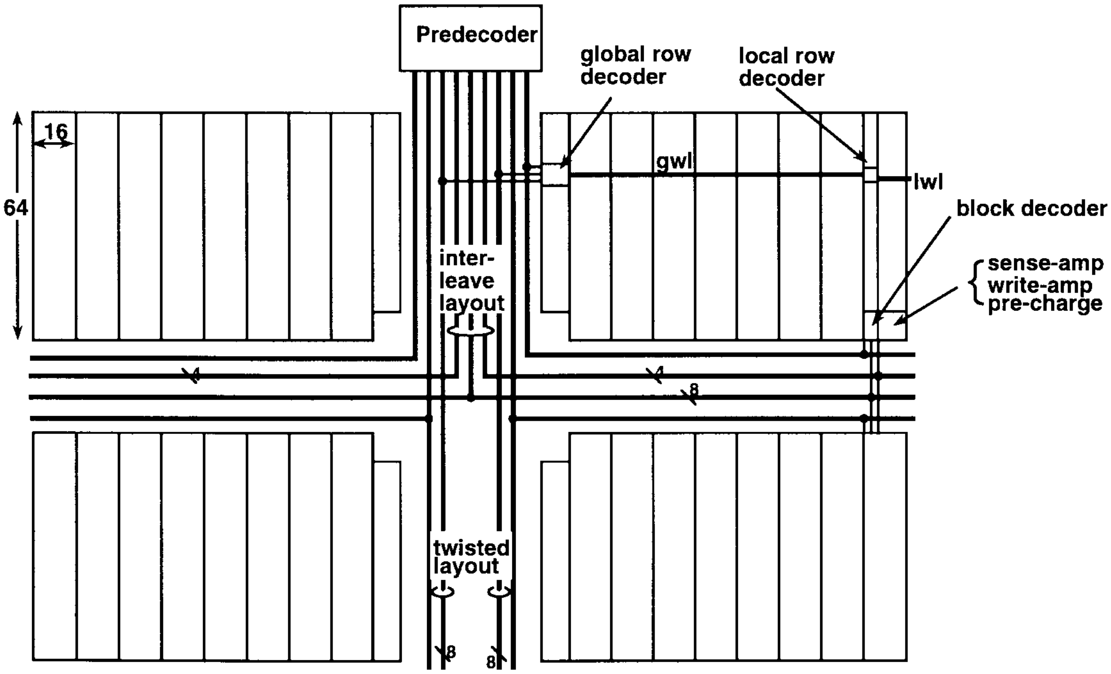


Fig. 14. Test-chip block diagram.

$V_{\rm bus} = 0.25$ V for 100-MHz operation. The write current is almost equal to the read current, indicating the effectiveness of the $V_{dd}/2$ bit lines and half-swing write bus in reducing the power. The small currents from $V_{\rm mid}$ indicate that the charge recycling mechanism is effective in generating the internal $V_{dd}/2$ voltage. The standby currents can be further reduced by implementing high-V_t cutoff devices.

The chip is functional from 0.9 to 2.4 V, demonstrating robust design of the half-swing signalling path. The measured total power is plotted in Fig. 16. The efficiencies of the $V_{\rm mid}$ supply (voltage regulator) and the $V_{\rm bus}$ supply (dc–dc converter) are assumed to be 40 and 80%, respectively. Since the currents drawn from these supplies are low, the efficiencies of the supply generators do not strongly affect the overall power dissipation.

Failure below 0.9 V is due to an insufficiently wide sense-enable pulse generated by the block decoder self-reset path and the replica bit line. The deviation from ideal V_{dd}^2 scaling at a V_{dd} of 2.4 V is only 12%, indicating that the leakage current in the decoder receiver gates during the half-select cases does not pose a significant power problem, even when $V_{dd}/2$ is over two times high-V_t.

TABLE II
SUPPLY CURRENTS AT 1 V, 100 MHz (mA)

	I(Vdd)	I(Vmid)	I(Vcc)	I(Vbus)
Read	762	9.0	< 0.25	149
Write	877	-80	2	< 0.25
1:1 Read/Write	826	-33.1	1	71.7
Standby	14.1	20.8	< 0.25	< 0.25

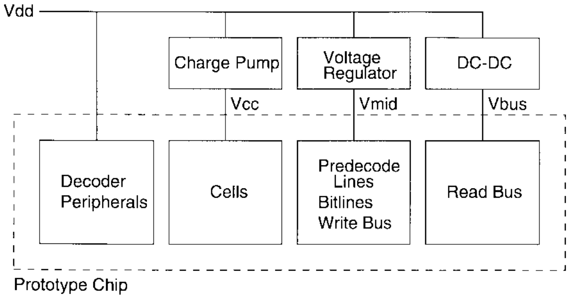


Fig. 15. Power supplies.

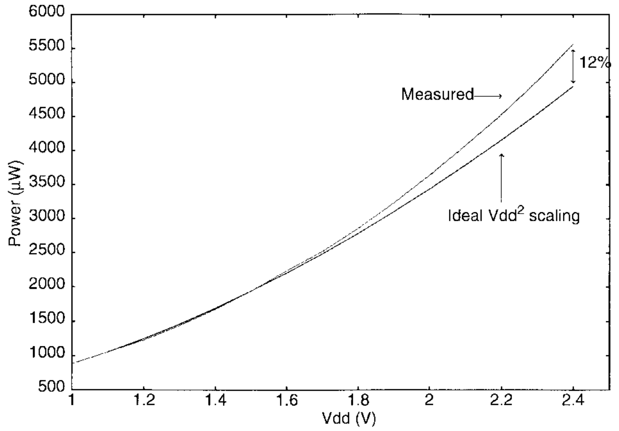


Fig. 16. Power versus V_{dd}.

Running at 100 MHz at 1-V V_{dd}, the 2-K × 16-b test chip dissipates 0.9 mW for a 1 : 1 mix of reads and writes. For comparison, a recently demonstrated 512 × 16-b SRAM running at 100 MHz at 0.9-V V_{dd} in a similar 0.25-μm dual-V_t technology dissipates 1.5 mW [2]. Note that the access widths are the same, and our test chip has four times the capacity of [2], yet dissipates only 60% as much power.

Fig. 17 compares the estimated power breakdown of the half-swing prototype with a full-swing design. A savings of 18% in read power and 46% in write power is obtained. The power savings in the decoder is 34%, while in the write bus/bit lines it is 64%, which is close to the 75% maximum savings. The higher power savings of the write bus/bit lines is due to a larger fraction of the capacitance in these units' being converted to half swing, while there was still a large amount of capacitance in the decoder that switched full-rail.

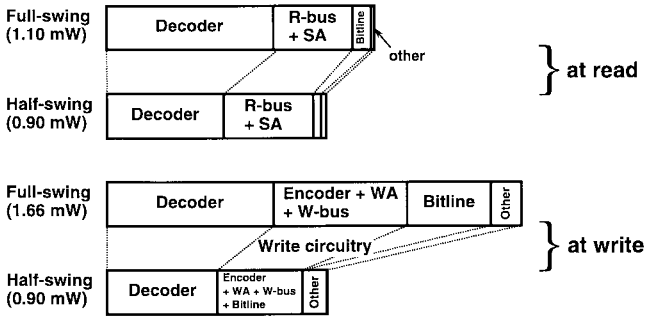


Fig. 17. Power breakdown.

We anticipated that testing the SRAM, with its low-voltage swings, would be greatly facilitated by the ability to display the real-time behavior of critical signals like the bit lines. However, since probing the real-time behavior of on-chip nets is difficult and expensive, we used a simple on-chip sampling circuit to display the analog waveforms of high-bandwidth signals on an inexpensive laboratory oscilloscope [21]. It is based on the subsampling of periodic signals as described in [22]. The sampler exploits the high bandwidth of MOS transmission gates by using one to periodically sample the analog voltage on the capacitance of an internal node. This voltage is then converted to a current that is driven off-chip into an oscilloscope. By making the internal waveform repetitive and then sampling it only once per period, we can allow the bandwidth of the output current to be significantly lower than that of the internal signal being measured. Furthermore, by sampling at a period $T + \Delta t$ that is slightly different from the chip period T, we can capture the entire waveform over the course of $T/\Delta t$ samples. The oscilloscope thus will display a time-expanded version of the on-chip waveform, running at the beat frequency of the chip and sampler clocks. Several samplers were used to measure different signals (see Fig. 18), and each was individually calibrated to avoid process variation inaccuracies between samplers.

Fig. 19 plots the on-chip sampled waveforms for the global word line, a predecoder output, sense enable, bit lines, and the read bus. The SPICE simulated waveforms for these nodes are also shown. The measured results are 13% faster than the simulated results for the signal assertion edges. As can be seen from Fig. 19(a), the global word-line recovery is very slow. The global row decoders are not self-resetting, and the small PMOS devices in the global row decoder, which only

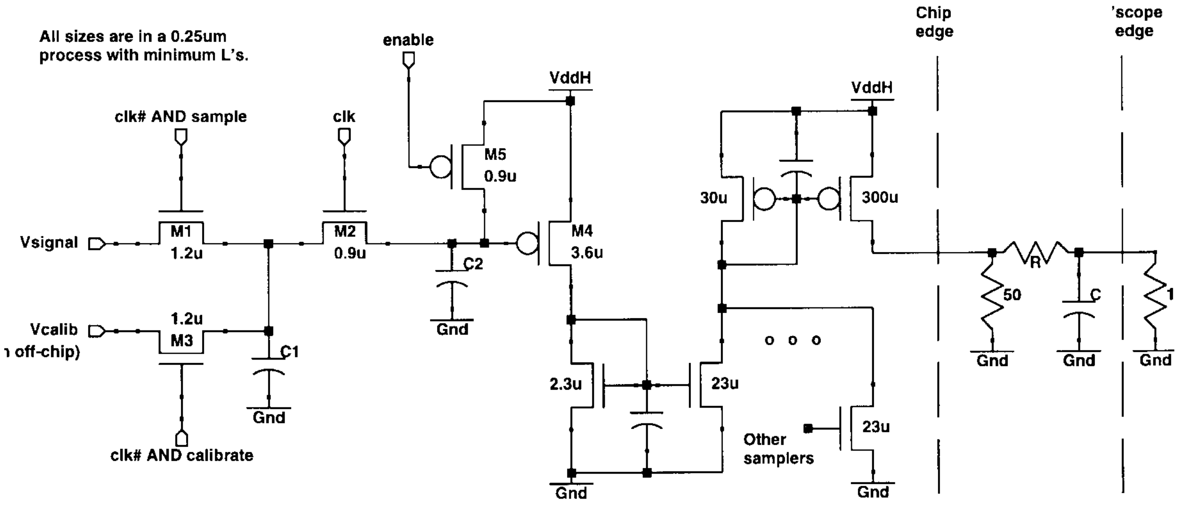
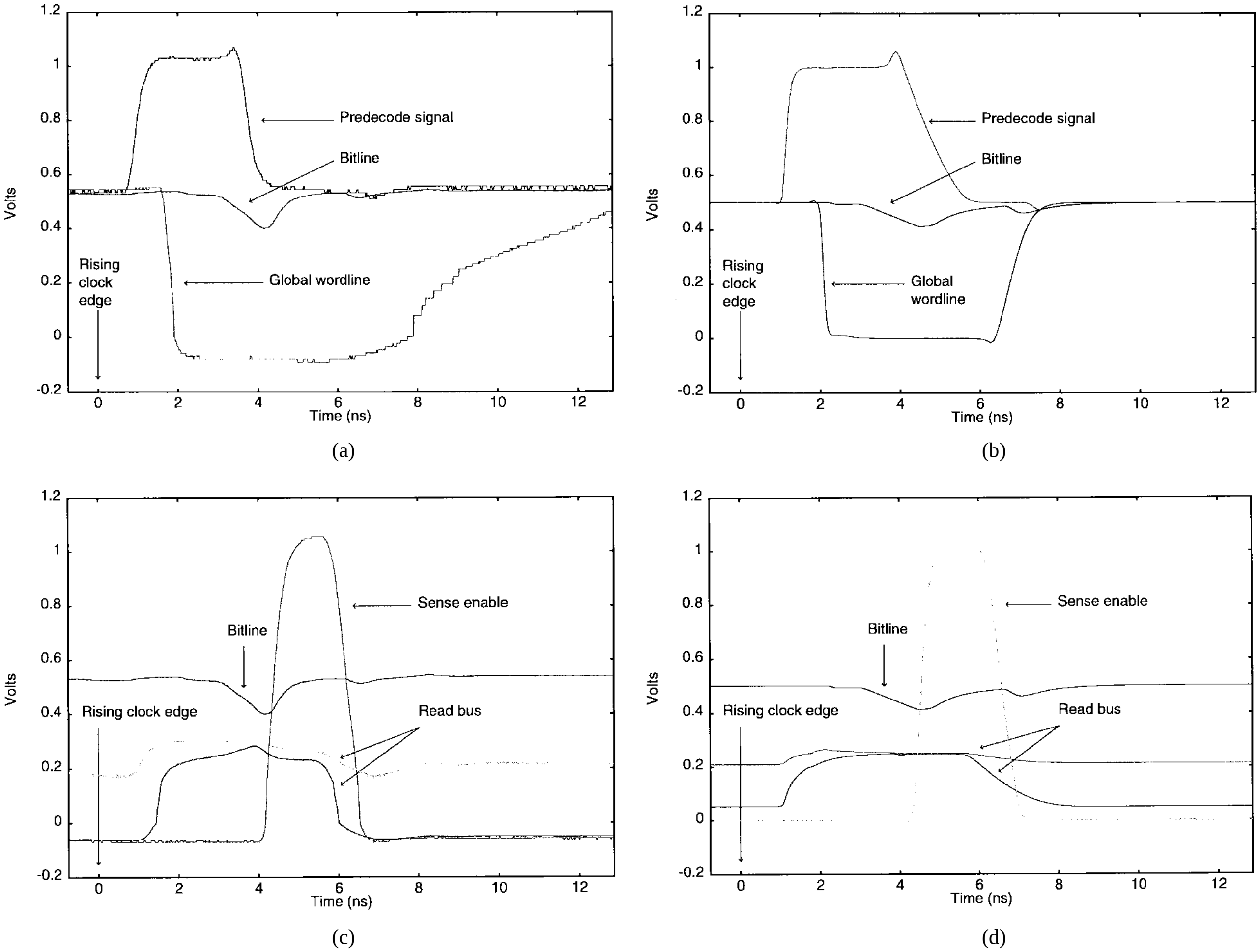


Fig. 18. Sampler architecture.



(a)

(b)

(c)

(d)

Fig. 19. Waveforms. (a) Sampled decoder. (b) Simulated decoder. (c) Sampled read. (d) Simulated read.

have $V_{dd}/2$ across their gate-to-source terminals, must reset the gate. Additionally, a process skew increased the absolute value of the low-V_{tp}, further weakening the device. The global word line is the cycle-time-limiting signal.

Since we did not place a sampler on the SRAM outputs, we estimate the access time by adding the simulated global sense delay to the measured delay up to the inputs of the global sense amplifiers (Fig. 20). At 1-V V_{dd}, the access time was 7.3 ns. A simulation of a version of the test-chip design using all full-swing signals (predecoder outputs, global word lines, bit lines, and write bus) and all low-V_t NMOS transistors in the decoder had an access time 13% faster than the simulation

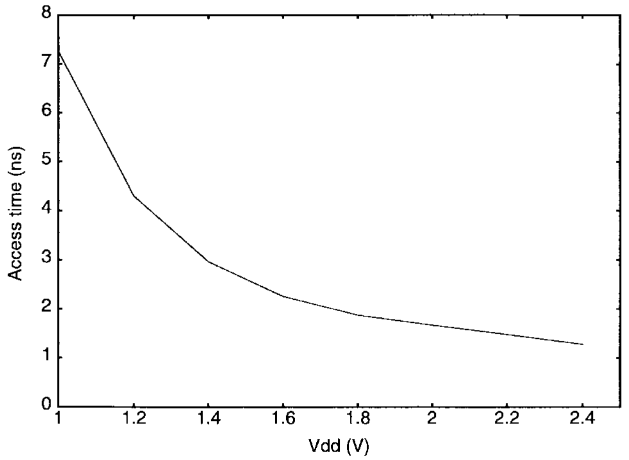


Fig. 20. Access time versus V_{dd}.

of the half-swing design. Thus, the speed degradation due to the use of half-swing signalling is small.

The samplers were also used to measure the noise injection due to capacitive coupling on the predecoder outputs. The worst case interwire noise coupling in a twisted 8-bit bus was measured to be 40 mV for a 0.5-V pulse attacker. The wires are 1 mm long, 0.45 μm wide, and with 0.6 μm spacing.

The accuracy of the charge recycling capacitance matching can be obtained by not connecting the $V_{\rm mid}$ supply and allowing the chip to free-run. If no external supply is connected to $V_{\rm mid}$, then the supply line will asymptotically approach the value $(C_p/(C_p + C_n)) * V_{dd}$, where C_p is the positive pulse capacitance and C_n is the negative pulse capacitance. The $V_{\rm mid}$ free-run voltage is frequency dependent at low frequencies due to the leakage current of the $V_{\rm mid}$ supply line but asymptotically approaches the settling value for higher frequencies as the leakage current becomes small compared to the signal switching currents. When the prototype was allowed to free-run at 100 MHz and 1-V V_{dd}, $V_{\rm mid}$ settled at 0.49 V for reads, 0.58 V for writes, and 0.53 V for a 1 : 1 mix of reads and writes. So for reads, when the only charge recycling is occurring in the decoder, C_n is 4% larger than C_p. For writes, when there is charge recycling in the decoder and between the write bus and bit lines, C_p is 38% larger than C_n. The less efficient charge recycling for the bit lines and write bus is due to the two sets of signals being matched having differing capacitance makeups and hence being more sensitive to process skews. While in the decoder, the signals being matched all had roughly the same capacitance makeup. The test chip operated properly during the free-run testing, further indicating robustness in the half-swing signalling paths.

VI. Conclusions

This paper has introduced a half-swing pulse-mode gate family that when combined with self-resetting techniques has shown that significant power savings can be achieved without affecting performance in certain applications. The reduced noise-margin problems are surmountable by using a PMOS leaker device and careful layout. The gates operate robustly even at high supply voltages. By using this technique, the 2-K × 16-b SRAM prototype dissipates 0.9 mW at 100 MHz using a 1-V V_{dd}. The half-swing bit lines contribute to the significant reduction of the write power, and correct operation is observed even when $V_{cc} = V_{dd}$. Additionally, highly efficient charge recycling is shown to be possible with careful design and simulation.

For larger SRAM designs, or for any other application where there are numerous high-capacitance lines between logic blocks, the power savings for using the half-swing pulse-mode gates would be significant. The large reduction in write power using the $V_{dd}/2$ referenced bit lines would be especially useful in embedded SRAM's with large word widths and frequent writes. While the reduced noise margin will limit the use of these techniques to circuits with regular wiring, in these situations they look promising for reducing power with a minimum performance penalty.

Acknowledgment

The authors would like to thank G. Wei and C.-K. K. Yang of Stanford University for helpful discussions; K. Hashimoto, K. Fujita, and the Mie Factory of Fujitsu, Ltd., for fabrication; and S. Kawashima, S. Hijiya, and T. Nakamura of Fujitsu Laboratories for encouragement.

References

[1] K. Itoh, K. Sasaki, and Y. Nakagome, "Trends in low-power RAM circuit technologies," in *Dig. Tech. Papers, 1994 Symp. Low Power Electronics,* 1994, pp. 84–87.

[2] M. Izumikawa, H. Igura, K. Furuta, H. Ito, H. Wakabayashi, K. Nakajima, T. Mogami, T. Horiuchi, and M. Yamashina, "A 0.25 μm CMOS 0.9 V 100-MHz DSP core," *IEEE J. Solid-State Circuits,* vol. 32, pp. 52–61, Jan. 1997.

[3] W. Lee, P. Landman, B. Barton, S. Abiko, H. Takahashi, H. Mizuno, S. Muramatsu, K. Tashiro, M. Fusumada, L. Pham, F. Boutaud, E. Ego, G. Gallo, H. Tran, C. Lemonds, A. Shih, M. Nandakumar, R. Eklund, and I. Chen, "A 1-V programmable DSP for wireless communications," *IEEE J. Solid-State Circuits,* vol. 32, pp. 1766–1776, Nov. 1997.

[4] T. Iwata, H. Yamauchi, H. Akamatsu, Y. Terada, and A. Matsuzawa, "Gate-over-driving CMOS architectures for 0.5 V single-power-supply-operated devices," in *ISSCC Dig. Tech. Papers,* Feb. 1997, pp. 290–291.

[5] T. Mori, B. Amrutur, K. Mai, M. Horowitz, I. Fukushi, T. Izawa, and S. Mitarai, "A 1 V 0.9 mW at 100 MHz 2K×16b SRAM utilizing a half-swing pulsed-decoder and write-bus architecture in 0.25μm dual-V_t CMOS," in *ISSCC Dig. Tech. Papers,* Feb. 1998, pp. 354–355.

[6] M. Yoshimoto, K. Anami, H. Shinohara, T. Yoshihara, H. Takagi, S. Nagao, S. Kayano, and T. Nakano, "A divided word-line structure in the static RAM and its application to a 64 K full CMOS RAM," *IEEE J. Solid-State Circuits,* vol. 18, pp. 479–484, Oct. 1983.

[7] B. Amrutur and M. Horowitz, "Techniques to reduce power in fast wide memories," in *Dig. Tech. Papers 1994 Symp. Low Power Electronics,* 1994, pp. 92–93.

[8] T. Chappell, B. Chappell, S. Schuster, J. Allan, S. Klepner, R. Joshi, and R. Franch, "A 2-ns cycle, 3.8-ns access 512 kb CMOS ECL SRAM with a fully pipelined architecture," *IEEE J. Solid-State Circuits,* vol. 26, pp. 1577–1584, Nov. 1991.

[9] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, and J. Yamada, "1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS," *IEEE J. Solid-State Circuits,* vol. 30, pp. 847–853, Aug. 1995.

[10] T. Kuroda, T. Fujita, S. Mita, T. Nagamatsu, S. Yoshioka, K. Suzuki, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9-V, 150-MHz, 10 mW, 4-mm^2, 2-D discrete cosine transform core processor with variable threshold-voltage (VT) scheme," *IEEE J. Solid-State Circuits,* vol. 31, pp. 1770–1777, Nov. 1996.

[11] Y. Nakagome, K. Itoh, M. Isoda, K. Takeuchi, and M. Aoki, "Sub-1-V swing internal bus architecture for future low-power ULSI's," *IEEE J. Solid-State Circuits,* vol. 28, pp. 414–419, Apr. 1993.

[12] H. Yamauchi, H. Akamatsu, and T. Fujita, "An asymptotically zero power charge-recycling bus architecture for battery-operated ultrahigh

data rate ULSI's," *IEEE J. Solid-State Circuits*, vol. 30, pp. 423–431, Apr. 1995.

[13] M. Hiraki, H. Kojima, H. Misawa, T. Akazawa, and Y. Hatano, "Data-dependent logic swing internal bus architecture for ultralow-power LSI's," *IEEE J. Solid-State Circuits*, vol. 30, pp. 397–402, Apr. 1995.

[14] H. Kojima, S. Tanaka, and K. Sasaki, "Half-swing clocking scheme for 75% power saving in clocking circuitry," *IEEE J. Solid-State Circuits*, vol. 30, pp. 432–435, Apr. 1995.

[15] H. Kawaguchi and T. Sakurai, "A reduced clock-swing flip-flop (RCSFF) for 63% power reduction," *IEEE J. Solid-State Circuits*, vol. 33, pp. 807–811, May 1998.

[16] N. Kushiyama, C. Tan, R. Clark, J. Lin, F. Perner, L. Martin, M. Leonard, G. Coussens, and K. Cham, "An experimental 295 MHz 4 K × 256 SRAM using bidirectional read/write shared sense amps and self-timed pulsed word-line drivers," *IEEE J. Solid-State Circuits*, vol. 30, pp. 1286–1290, Nov. 1995.

[17] B. Kong, J. Choi, S. Lee, and K. Lee, "Charge recycling differential logic (CRDL) for low power application," *IEEE J. Solid-State Circuits*, vol. 31, pp. 1267–1276, Sept. 1996.

[18] J. Alowersson and P. Andersson, "SRAM cells for low-power write in buffer memories," in *Dig. Tech. Papers 1995 Int. Symp. Low Power Electronics and Design*, 1995, pp. 60–61.

[19] S. Kawashima, T. Mori, R. Sasagawa, M. Hamaminato, S. Wakayama, K. Sukegawa, and I. Fukushi, "A charge transfer amplifier and an encoded bus architecture for low power SRAM," in *Symp. VLSI Circuits Dig. Tech. Papers*, June 1997, pp. 77–78.

[20] A. J. Stratakos, S. Sanders, and R. Brodersen, "A low voltage CMOS dc–dc converter for a portable battery-operated system," in *Proc. IEEE Power Electronics Specialists Conf.*, June 1994, vol. 1, pp. 619–626.

[21] R. Ho, B. Amrutur, K. Mai, B. Wilburn, T. Mori, and M. Horowitz, "Applications of on-chip samplers for test and measurement of integrated circuits," in *Symp. VLSI Circuits Dig. Tech. Papers*, June 1998, p. 139.

[22] P. Larsson and C. Svensson, "Measuring high-bandwidth signals in CMOS circuits," *Electron. Lett.*, vol. 29, pp. 1761–1762, Sept. 1993.

Kenneth W. Mai received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, in 1993 and 1997, respectively, and currently is pursuing the Ph.D. degree in electrical engineering there.

His research interests include low-power and high-performance circuit design.

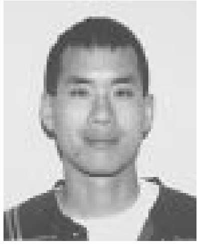
Mr. Mai is a member of Tau Beta Pi and Phi Beta Kappa.

Toshihiko Mori was born in Kyoto, Japan, in 1959. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Osaka University, Osaka, Japan, in 1983, 1985, and 1996, respectively.

He joined Fujitsu Laboratories, Ltd., Kawasaki, Japan, in 1985. He researched resonant tunneling hot electron transistors until 1996. Since then, he has been researching low-power SRAM's. Currently, he is a Visiting Scholar at Stanford University, Stanford, CA.

Bharadwaj S. Amrutur received the B.Tech. degree in computer science and engineering from the Indian Institute of Technology, Bombay, in 1990. He received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1994, where he currently is pursuing the Ph.D. degree.

His research interests are in low-power and high-performance circuit design.



Ron Ho received the B.S. degree in electrical engineering and the A.B. degree in science, technology, and society from Stanford University, Stanford, CA, in 1992. He received the M.S. degree in electrical engineering from Stanford in 1993 and currently is pursuing the Ph.D. degree in electrical engineering there.

In 1993, he joined Intel Corp., where he has worked on microprocessor design and design methodologies. His research interests are in high-performance circuit design.

Mr. Ho is a member of Tau Beta Pi and Phi Beta Kappa. He was the 1993–1994 IEEE Fortescue Scholar.

Bennett Wilburn received the B.S. and M.S. degrees in electrical engineering from Stanford University, Stanford, CA, in 1993, where he currently is pursuing the Ph.D. degree.

He was with Hewlett Packard, Fort Collins, CO, for two years, working mostly on the PA-8000 microprocessor. His research interests are VLSI design, image-based rendering, and machine vision.

Mark A. Horowitz received the B.S. and M.S. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, in 1978 and the Ph.D. degree from Stanford University, Stanford, CA, in 1984.

He is the Yahoo Founders Professor of Electrical Engineering and Computer Science at Stanford. His research area is in digital system design. He has led a number of processor designs, including MIPS-X, one of the first processors to include an on-chip instruction cache; TORCH, a statically scheduled, superscalar processor; and FLASH, a flexible DSM machine. He has also worked in a number of other chip design areas, including high-speed memory design, high-bandwidth interfaces, and fast floating point. In 1990, he took leave from Stanford to help start Rambus, Inc., a company designing high-bandwidth memory interface technology. His current research includes multiprocessor design, low-power circuits, memory design, and high-speed links.

Dr. Horowitz received the 1985 Presidential Young Investigator Award, the IBM Faculty Development Award, and the 1993 Best Paper Award at the International Solid State Circuits Conference.

Isao Fukushi was born in Miyagi, Japan, in 1957. He received the B.S. degree in applied physics from Tohoku University, Sendai, Japan, in 1982.

He joined Fujitsu, Ltd., Kawasaki, Japan, in 1982, where he developed high-speed ECL/BiCMOS SRAM's for supercomputers. In 1994, he joined Fujitsu Laboratories, Ltd., Kawasaki, Japan, where he has researched low-power, low-voltage LSI design. Currently, his main area of research activity is low-power embedded SRAM's.

Tetsuo Izawa was born in Ibaraki, Japan, on February 5, 1960. He received the B.S.(Eng.) degree in applied physics from the University of Tsukuba, Ibaraki, Japan, in 1983.

He joined Fujitsu, Ltd., Kawashima, Japan, in 1983, and has been engaged in research and development of MOS transistors for high-speed LSI's.

Mr. Izawa is a member of the Japan Association of Applied Physics.

Shin Mitarai was born in Mie, Japan, on February 3, 1958. He received the B.E. degree in electrical engineering from Waseda University, Tokyo, Japan, in 1980.

He joined Fujitsu, Ltd., Kawasaki, Japan, in 1980. He was engaged in the development of MCU's and MPU's, and he has been engaged in the development of CMOS technology. He currently is the Project Manager of the Advanced CMOS Department, Technology Development Division.