A Replica Technique for Wordline and Sense Control in Low-Power SRAM's

Bharadwaj S. Amrutur and Mark A. Horowitz, *Senior Member, IEEE*

***Abstract*—With the migration toward low supply voltages in low-power SRAM designs, threshold and supply voltage fluctuations will begin to have larger impacts on the speed and power specifications of SRAM's. We present techniques based on replica circuits which minimize the effect of operating conditions' variability on the speed and power. Replica memory cells and bitlines are used to create a reference signal whose delay tracks that of the bitlines. This signal is used to generate the sense clock with minimal slack time and control wordline pulsewidths to limit bitline swings. We implemented the circuits for two variants of the technique, one using bitline capacitance ratioing in a 1.2-μm 8-kbyte SRAM, and the other using cell current ratioing in a 0.35-μm 2-kbyte SRAM. Both the RAM's were measured to operate over a wide range of supply voltages, with the latter dissipating 3.6 mW at 150 MHz at 1 V and 5.2 μW at 980 kHz at 0.4 V.**

***Index Terms*—Low power, low swing bus, low voltage, pulsed decoder, replica technique, self-timing, sense clock control, SRAM's, threshold variation, wordline pulsing.**

I. Introduction

LOW-POWER circuit designers have been continually pushing down supply voltages to minimize the energy consumption of chips for portable applications [1]–[3]. The same trend has also applied to low-power SRAM's in the past few years [4]–[6]. While the supply voltages are scaling down at a rapid rate, to control subthreshold leakage, the threshold voltages have not scaled down as fast, which has resulted in a corresponding reduction of the gate overdrive for the transistors. With the fluctuations in the threshold voltages also not expected to decrease in future submicron devices [7], [8], the delay variability of low-power circuits across process corners will increase in the future [9]. The large delay spreads across process corners will necessitate bigger margins in the design of the bitline path in an SRAM, and also will result in larger bitline power dissipation and loss of speed. This problem can be mitigated by using a self-timed approach to designing the bitline path, based on delay generators which track the bitline delays across operating conditions.

Traditionally, the bitline swings during a read access have been limited by using active loads of either diode-connected nMOS or resistive pMOS [10], [11], but these clamp the bitline swing at the expense of a steady bitline current. A more power-efficient way of limiting the bitline swings is to use high-impedance bitline loads and pulse the wordlines [12]–[15]. Bitline power can be further minimized by controlling the wordline pulsewidth to be just wide enough to guarantee the minimum bitline swing development. This type of bitline swing control can be achieved by a precise pulse generator that can match the bitline delay. Low-power SRAM's also use clocked sense amplifiers to limit the sense power. These are either the current mirror type [16], [17] or cross-coupled latch type [18], [19] designs. In the former, the sense clock turns on the amplifier sometime before the sensing, to set up the amplifier in the high-gain region. To reduce power, the amount of time the amplifier is ON should be minimized. In the latch-type amplifiers, the sense clock starts the amplification, and hence the sense clock needs to track the bitline delay to ensure correct and fast operation.

Fundamentally, the clock path needs to match the data path to ensure fast and low-power operation. The data path starts from the local block select and/or global wordline, and goes through the wordline driver, memory cell, and bitline to the input of the sense amps. The clock path often starts from the local block select or some clock phase, and goes through a buffer chain to generate the sense clock. The delay variations in the former are dominated by the bitline delay since the memory cells are made out of minimum sized devices and are more vulnerable to process variations. Therefore, the delays of the two paths do not track each other very well over all process and environment conditions. Enough delay margin has to be provided to the sense clock path for worst case conditions, which reduces the average case performance. The rest of this paper describes methods of using replica circuits, which mimic the delay of the bitline path over all conditions to create the clocks, and gives experimental results from using these techniques. The next section presents simulation data comparing the matching of bitline delay with inverter chain delay and replica circuit delay under different operating conditions. The following two sections describe different methods of building replica circuits. Section III presents a clock circuit which uses a dummy memory cell that drives bitlines with reduced capacitance, and Section IV describes a circuit which uses a full bitline load. Results from two prototype chips which implement the two different replica techniques are presented in Section V.

II. Clock Matching

The prevalent technique to generate the timing signals within the array core essentially uses an inverter chain. This can take one of two forms—the first kind relies on a clock

Manuscript received August 30, 1997; revised March 4, 1998. This work was supported by the Advanced Research Projects Agency under Contract J-FBI-92-194 and by Fujitsu Ltd.

The authors are with the Center for Integrated Systems, Stanford University, Stanford, CA 94305-4070 USA (e-mail: amrutur@chroma.stanford.edu).

Publisher Item Identifier S 0018-9200(98)05522-X.

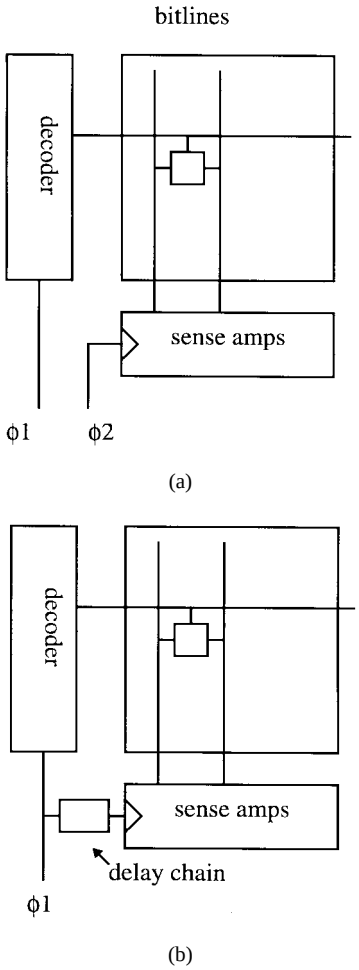

Fig. 1. Common sense clock generation techniques.

phase to do the timing [Fig. 1(a)] [20], and the second kind uses a delay chain within the accessed block, and is triggered by the block select signal [Fig. 1(b)] or a local wordline [21]. The main problem in these approaches is that the inverter delay does not track the delay of the memory cell over all process and environment conditions. The tracking issue becomes more severe for low-power SRAM's operating at low voltages due to enhanced impact of threshold and supply voltage fluctuations on delays as described by

$$\frac{\sigma_T^2}{T^2} \propto \frac{\sigma_{Vdd}^2 + \sigma_{Vt}^2}{(Vdd - Vt)^2} \tag{1}$$

which shows that delay variations are inversely proportional to the gate overdrive. Fig. 2 plots the ratio of bitline delay to obtain a bitline swing of 120 mV from a 1.2-V supply and the delay of two different delay elements for various operating conditions. One delay element is based on an inverter chain with a fan-out of four loading (diamonds), and the other is based on a replica structure consisting of a replica memory cell and a dummy bitline. The process and temperature are encoded as XYZ where X represents the nMOS type (S = slow, F = fast, T = typical), Y represents the pMOS type (one of S, F, T), and Z is the temperature (H for 115 °C and C for 25 °C). The S and F transistors have a 2-sigma threshold variation unless suffixed by a 3, in which case they represent 3-sigma threshold variations. The process used is a typical 0.25-μm CMOS process, and simulations are done for a bitline spanning 64 rows. We can observe that the bitline delay to inverter delay ratio can vary by a factor of two over these conditions, the primary reason being that, while the memory cell delay is mainly affected by the nMOS thresholds, the inverter chain delay is affected by both nMOS and pMOS thresholds. The worst case matching for the inverter delay chain occurs for process corners where the nMOS and pMOS thresholds move in the opposite direction. In the above simulations, it is assumed that they move independently, while in reality, there will be some correlation between them which would make the mismatch for the inverter delay chain less pronounced, but still worse than that of the replica element.

The delay element is designed to match the delay of a nominal memory cell in a block. But in an actual block of cells, there will be variations in the cell currents across the cells in the block. Fig. 3 displays the ratio of delays for the bitline and the delay elements for varying amounts of threshold mismatch in the access device of the memory cell compared to the nominal cell. The graph is shown only for the case of the accessed cell being weaker than the nominal cell as this would result in a lower bitline swing. The curves for the inverter chain delay element (hatched) and the replica delay element (solid) are shown with error bars for the worst case fluctuations across process corners. The variation of the delay ratio across process corners in the case of the inverter chain delay element is large even with zero offset in the accessed cell, and grows further as the offsets increase. In the case of the replica delay element, the variation across the process corners is negligible at zero offsets, and starts growing with increasing offsets in the accessed cell. This is mainly due to the adverse impact of the higher nMOS thresholds in the accessed cell under slow nMOS conditions. It can be noted that the tracking of the replica delay element is better than that of the inverter chain delay element across process corners, even with offsets in the accessed memory cell.

There are two more sources of variations that are not included in the graphs above and make the inverter matching even worse. The minimum sized transistors used in memory cells are more vulnerable to *delta–W* variations than the nonminimum sized devices used typically in the delay chain. Furthermore, accurate characterization of the bitline capacitance is also required to enable a proper delay chain design. These two sources of variations would make the matching even worse for the inverter chain delay element.

All of the sources of variations have to be taken into account in determining the speed and power specifications for the part. To guarantee functionality, the delay chain has to be designed for worst case conditions, which means that the clock circuit must be padded in the nominal case, degrading performance. Replica-based delay elements, by virtue of their good tracking, offer the possibility of designing SRAM's with tight specifications across all process corners [22]. Two ways of creating and using these replica structures are explained in the following sections.

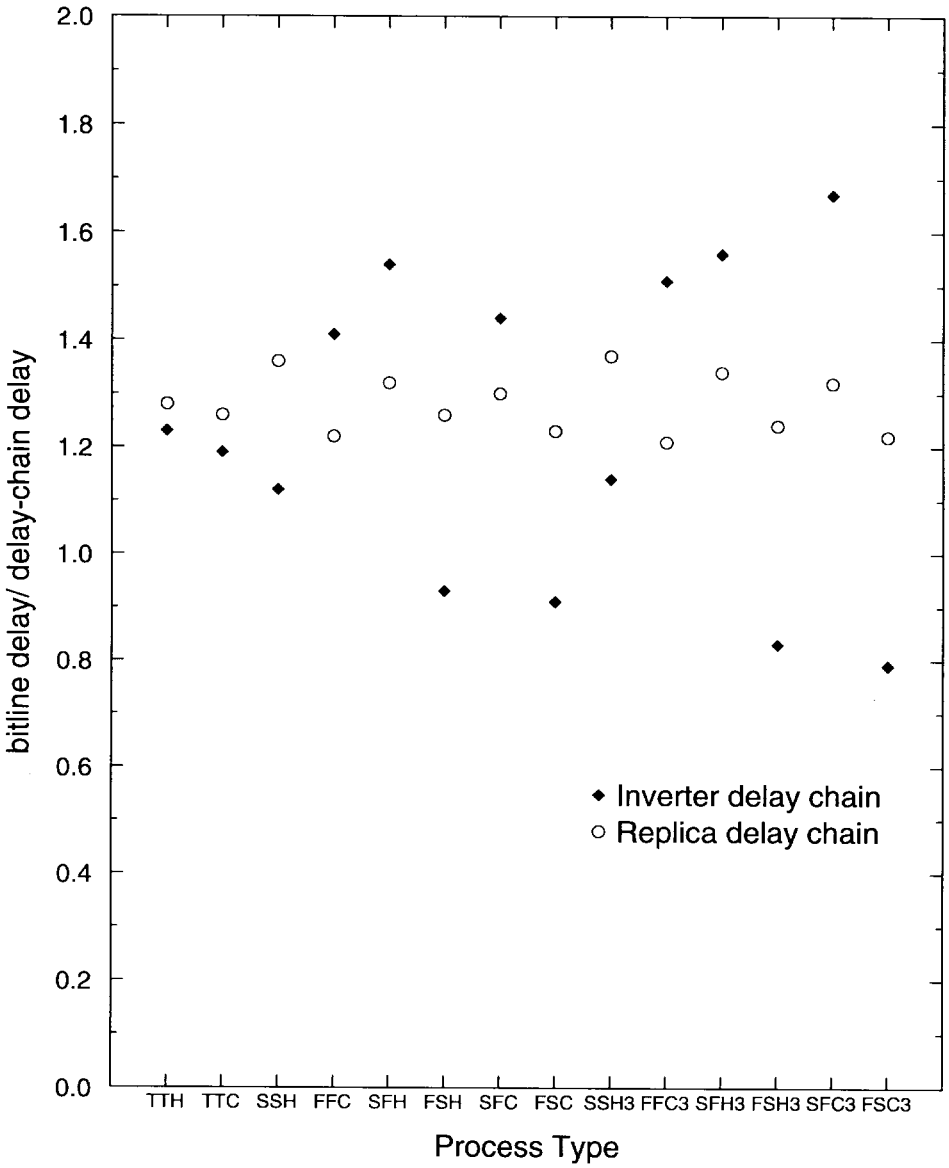

Fig. 2. Delay matching between the bitline delay to generate 120 mV and two delay elements, one based on an inverter chain and the other on a replica cell–bitline combination.

III. Feedback Based on Capacitance Ratioing

The replica delay stage is made up of a memory cell connected to a dummy bitline whose capacitance is set to be a fraction of the main bitline capacitance. The value of the fraction is determined by the required bitline swing for proper sensing. For the clocked voltage sense amplifiers we use (Fig. 4), the minimum bitline swing for correct sensing is around a tenth of the supply. An extra column in each memory block is converted into the dummy column by cutting its bitline pair to obtain a segment whose capacitance is the desired fraction of the main bitline (Fig. 5). The replica bitline has a similar structure to the main bitlines in terms of the wire and diode parasitic capacitances. Hence, its capacitance ratio to the main bitlines is set purely by the ratio of the geometric lengths r/h. The replica memory cell is programmed to always store a zero so that, when activated, it discharges the replica bitline. The delay from the activation of the replica cell to the 50% discharge of the replica bitline tracks that of the main bitline very well (see Fig. 2—circles). The delays can be made equal by fine tuning of the replica bitline height using simulations. The replica structure takes up only one additional column per block, and hence has very little area overhead.

The circuits to control the sense clock and wordline pulsewidths are shown in Fig. 6. The block decoder activates the replica delay cell (node fwl). The output of the replica delay cell is fed to a buffer chain to start the local sensing, and is also fed back to the block decoder to reset the block select signal. Since the block select pulse is ANDed with the global wordline signal to generate the local wordline pulse, the latter's pulsewidth is set by the width of block select signal. It is assumed that the block select signal does not arrive earlier than the global wordline. The delay of the buffer chain to drive the sense clock is compensated by activating the replica delay cell with the unbuffered block select signal.

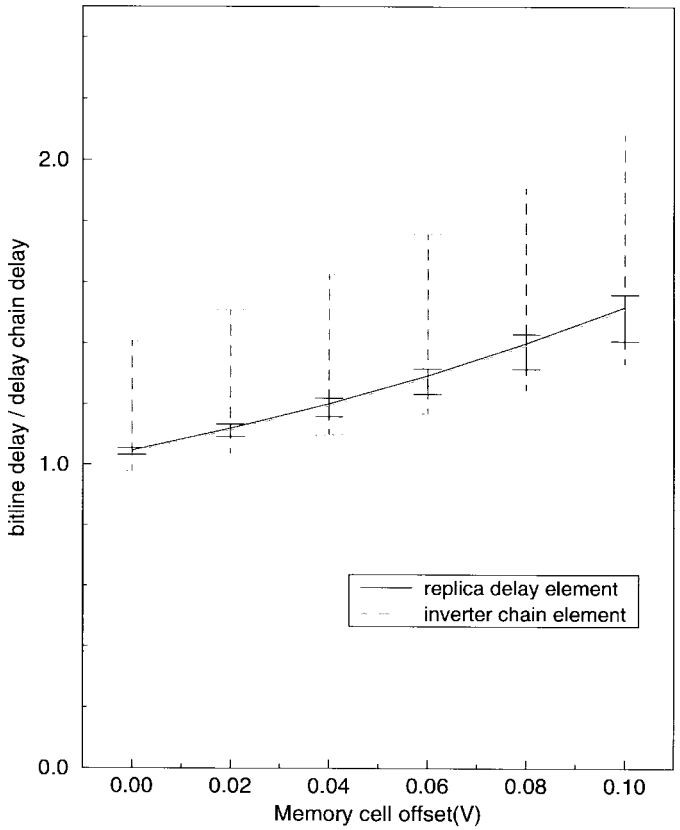

Fig. 3. Matching of the bitline delay with the inverter chain delay and the replica cell–bitline delay across process fluctuations over varying threshold offsets for the accessed memory cell.

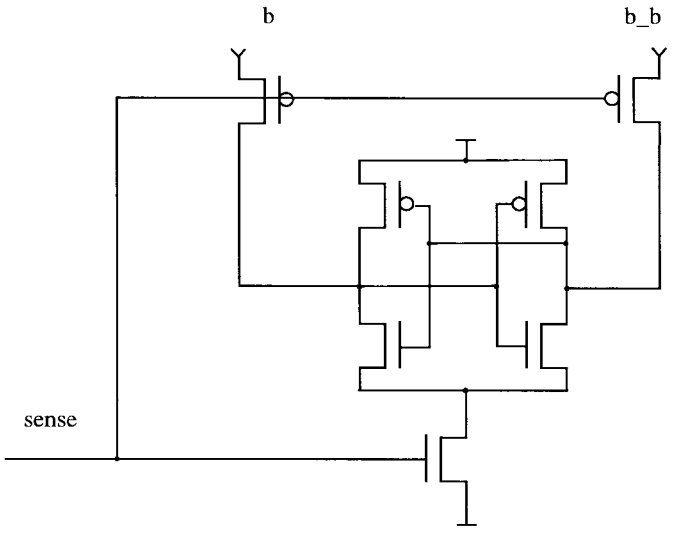

Fig. 4. Latch-type sense amplifier.

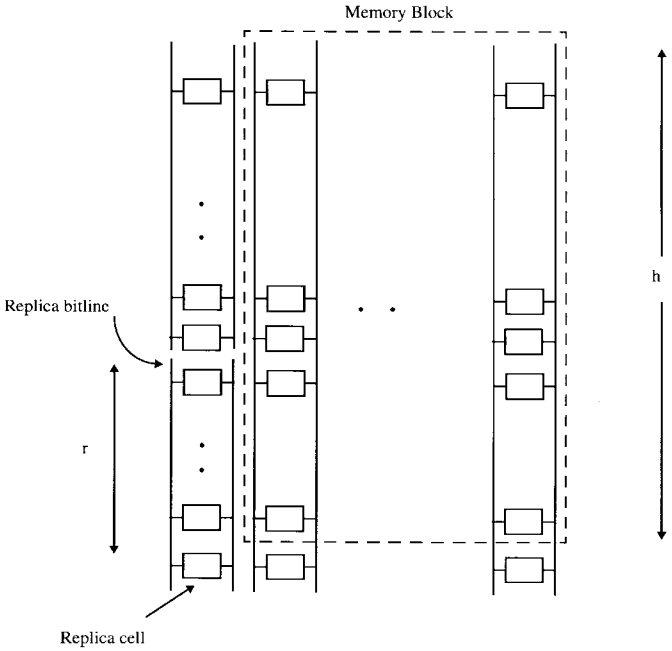

Fig. 5. Design of the replica bitline column.

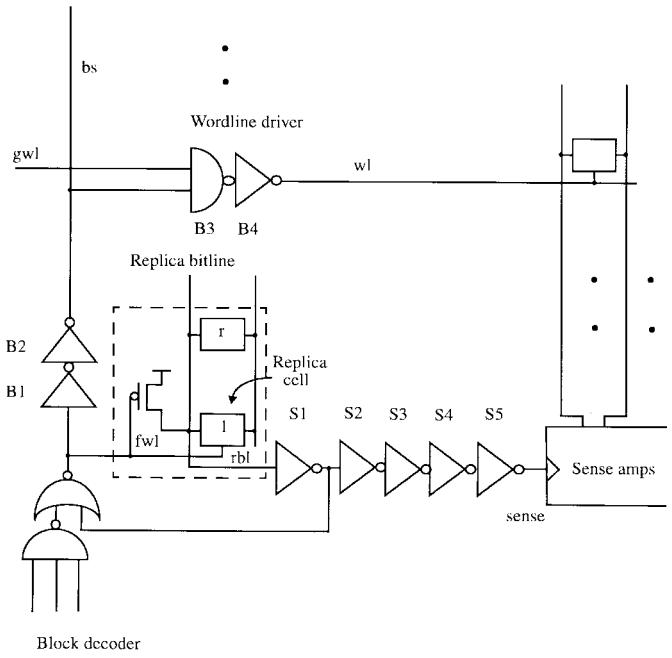

Fig. 6. Control circuits for sense clock activation and wordline pulse control.

The delay of the five inverters in the buffer chain, $S1$–$S5$, is set to match the delay of the four stages, $B1$–$B4$, of the block select to local wordline path (the sense clock needs to be a rising edge). The problem of delay matching has now been pushed from having to match bitline and inverter chain delay to having to match the delay of one inverter chain to a chain of inverters and an AND gate. The latter is easier to tackle, especially since the matching for only one pair of edges needs to be done. A simple heuristic for matching the delay of a rising edge of the five-long chain, $S1$–$S5$, to the rising delay of the four-long chain, $B1$–$B4$, is to ensure that the sums of falling delays in the two chains are equal, as well as the sum of rising delays [23] (Fig. 7). The S chain has three rising delays and two falling delays, while the B chain has two rising and falling delays. This simple sizing technique ensures that the rising and falling delays in the two chains are close to each other, giving good delay tracking between the two chains over all process corners. The delay from fwl (see Fig. 6) to minimum bitline swing is $t_{B\text{chain}} + t_{\text{bitline}}$

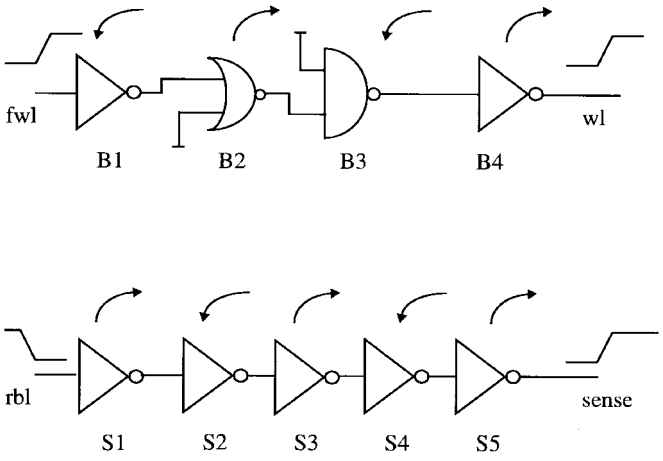

Fig. 7. Delay matching of two buffer chains.

TABLE I
BLOCK PARAMETERS: 256 ROWS, 64 COLUMNS

	Replica Delay Element (replica bitline height = 29)			Inverter Chain Element			
Process, Supply(V)	Bitline Swing (mV)	Max Bitline Swing (mV)	tSlack (as fraction of fo4del)	Bitline Swing (mV)	Max Bitline Swing (mV)	tSlack (as fraction of fo4del) a.u.	Relative Max Bitline swing (%)
TTH, 1.2	128	136	1.67	139	158	2.24	16.4
FF3, 1.2	114	133	0.64	181	187	4.2	41
SS3, 1.2	136	138	2.37	126	147	1.63	6
SF3, 1.2	97	102	-0.3	101	110	0	7
FS3, 1.2	167	176	2.17	240	251	5.25	43
TTH, 1.08	121	127	1.55	114	134	0.93	5.6
TTH, 1.32	135	145	1.72	172	187	3.64	30

and the delay to the senseclock is $t_{\text{replica}} + t_{S\text{chain}}$ delay. If t_{bitline} equals t_{replica} and t_{Bchain} equals $t_{S\text{chain}}$, then the sense clock fires exactly when the minimum bitline swings have developed.

We next look at two design examples, one for a block of 256 rows and 64 columns, and the other for a block with 64 rows and 32 columns. The number of rows in these blocks is typical of large and small SRAM's, respectively. For each block, the replica-based implementation is compared to an inverter-chain-based one. Table I summarizes the simulated performance of the 256 block design over various process corners. Five process corners are considered along with a 10% supply voltage variation at the TT corner. The delay elements are designed to yield a bitline swing of around 100 mV when the sense clock fires, under all conditions, with the weakest corner being the $SF3$ corner with slow nMOS and fast pMOS (since the memory cell is weak and inverters are relatively faster). For each type of delay element, the table provides the bitline swing when the sense clock fires and the maximum bitline swing after the wordline is shut off. An additional column notes the "slack" time of the sense clock with respect to an ideal sense clock as a fraction of a fan-out of 4 gate delay. This time indicates the time lost in the turning ON of the sense amp compared to an ideal sense clock generator

TABLE II
BLOCK PARAMETERS: 64 ROWS, 32 COLUMNS

	Replica Delay Element (replica bitline height = 6)			Inverter Chain Element			
Process, Supply(V)	Bitline Swing (mV)	Max Bitline Swing (mV)	tSlack (as fraction of fo4del)	Bitline Swing (mV)	Max Bitline Swing (mV)	tSlack (as fraction of fo4del) a.u.	Relative Max Bitline swing (%)
TTH, 1.2	101	198	0.24	123	192	0.38	-3
FF3, 1.2	100	210	0	162	207	0.85	-1
SS3, 1.2	125	193	0.42	102	177	0	-8
SF3, 1.2	101	155	0	100	145	0	-7.5
FS3, 1.2	110	211	0.1	194	211	1.1	0
TTH, 1.08	113	176	0.24	88	162	-0.24	-8
TTH, 1.32	122	228	0.37	162	223	0.83	-2

which would magically fire under all conditions exactly when the bitline swings are 100 mV, and directly adds to the critical path delay for the SRAM. The last column shows the excess swing of the bitlines for the inverter chain case relative to the replica case as a percentage and represents the excess bitline power for the former over the latter. The last two rows show the performance at ±10% of the nominal supply of 1.2 V under typical conditions. Considering all of the rows of the table, we note that the slack time for the replica case is within 2.4 gate delays, while that of the inverter case goes up to 5.25 gate delays, indicating that the latter approach will lead to a speed specification which is at least 3 gate delays slower than the former. The bitline power overhead in the inverter-based approach can be up to 40% more than the replica-based approach. If we were to consider only correlated threshold fluctuations for the nMOS and the pMOS, then the delay spread for both of the approaches is lower by one gate delay, but the relative performance difference still exists. The main reason for the variation in the slack time for the replica approach is the mismatch in the delays of the buffer chains across the operating conditions. This comes about mainly due to the variation of the falling edge rate of the replica bitline. In the case of the inverter-based design, the spread in slack time comes about due to the lack of tracking between the bitline delay and the inverter chain delay, as discussed in the earlier section. To study the scalability of the replica technique, designs for a 64-row block are compared in Table II. The small bitline delay for short bitlines is easy to match even with an inverter chain delay element, and there is only a slight advantage for the replica design in terms of delay spread, while there is not much difference in the maximum bitline swings. The maximum bitline swings are considerably larger than the previous case, mainly due to the smaller bitline capacitance.

This technique can be modified for clocked current mirror sense amplifiers, where the exact time of arrival of the sense clock is not critical as long as it arrives sufficiently ahead to set up the amplifiers in the high-gain stage by the time the bitline signal starts developing. Delaying the sense clock to be as late as safely possible minimizes the amplifier static power dissipation. This can be achieved in the above scheme

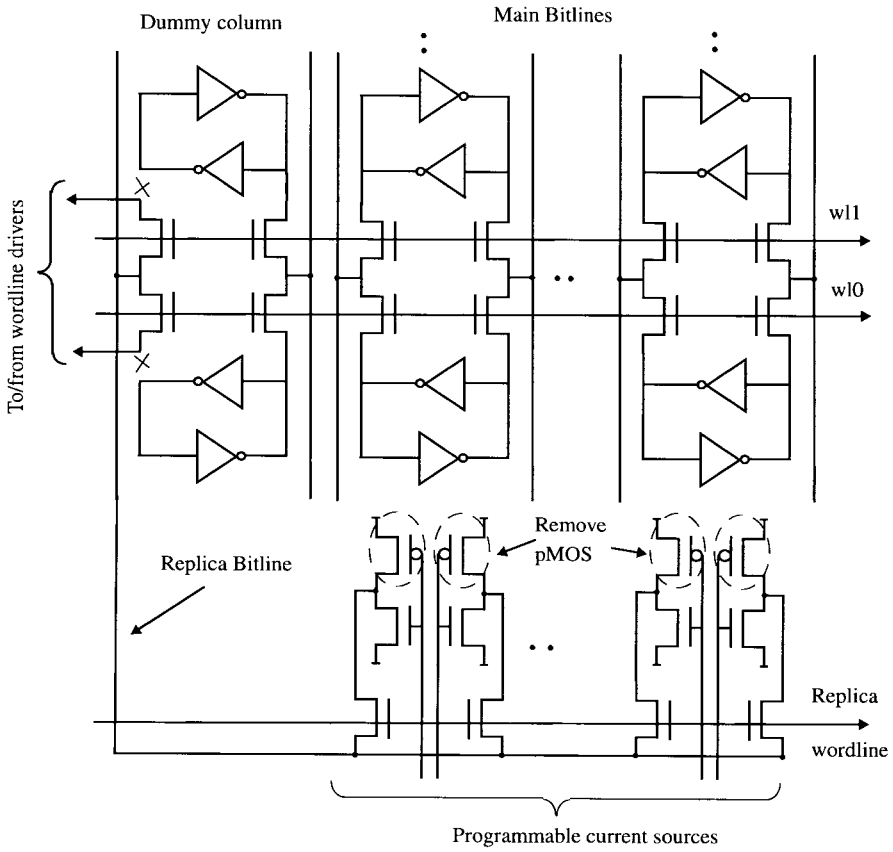

Fig. 8. Current-ratio-based replica structure.

by merely trimming the delay of the S chain with respect to the B chain to ensure that the sense clock turns on a fixed number of gate delays before the bitlines differentiate.

IV. Feedback Based on Cell-Current Ratioing

While the above technique works well, it can be modified further to improve the access time. If the reset timing signal for the wordline can be generated locally, then the wordline driver can be skewed to speed up the propagation of the rising block select transition, reducing the access time, with the falling wordline transition being triggered off the local reset signal, similar to the postcharge gates discussed in [24].

An extra row and column containing replica memory cells can be used to provide local resetting timing information for the wordline drivers. The extra row contains memory cells whose pMOS devices are eliminated to act as current sources, with currents equal to that of an accessed memory cell (Fig. 8). All of their outputs are tied together, and they simultaneously discharge the replica bitline. This enables a multiple of memory cell current to discharge the replica bitline. The current sources are activated by the replica wordline, which is turned on during each access of the block. The replica bitline is identical in structure to the main bitlines, with dummy memory cells providing the same amount of drain parasitic loading as the regular cells. By connecting n current sources to the replica bitline, the replica bitline slew rate can be made to be n times that of the main bitline slew rate, achieving the same effect as bitline capacitance ratioing described earlier.

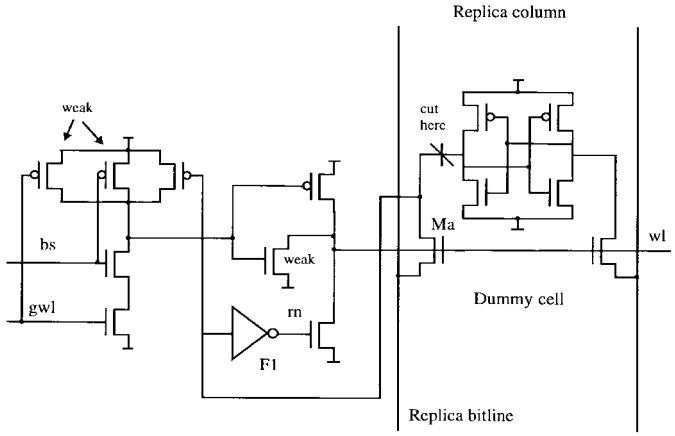

Fig. 9. Skewed wordline driver.

The local wordline drivers are skewed to speed up the rising transition, and they are reset by the replica bitline as shown in Fig. 9. The replica bitline signal is forwarded into the wordline driver through the dummy cell access transistor Ma. This occurs only in the activated row since the access transistor of the dummy cell is controlled by the row wordline wl, minimizing the impact of the extra loading of $F1$ on the replica bitline. The forward path of the wordline driver can be optimized for speed, independent of the resetting of the block-select or global wordline by skewing the transistor sizes.

The control circuits to activate the replica bitline and the sense clock are shown in Fig. 10. The dummy wordline driver

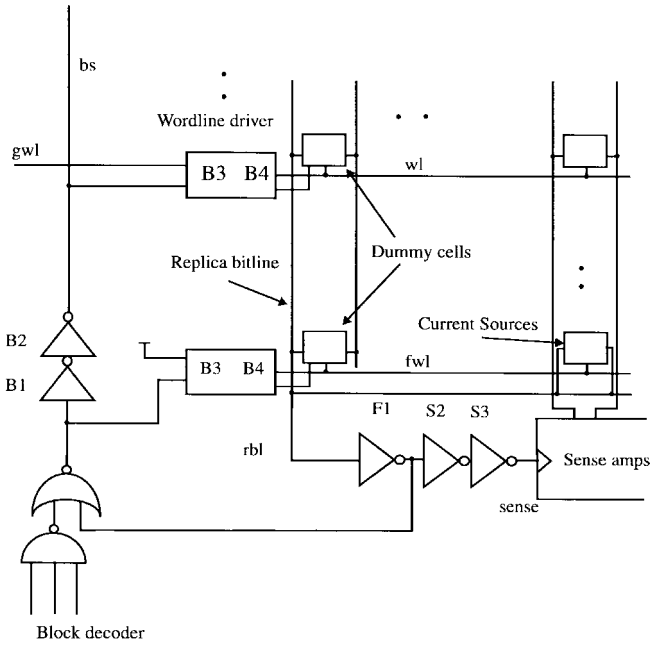

Fig. 10. Control circuits for current-ratio-based replica structure.

TABLE III
BLOCK PARAMETERS: 256 ROWS, 64 COLUMNS

	Replica Delay Element (replica current sources = 8)		
Process, Supply(V)	Bit-line Swing (mV)	Max Bitline Swing (mV)	tSlack (as fraction of fo4del)
TTH, 1.2	120	155	1.2
FF3, 1.2	106	170	0.3
SS3, 1.2	124	155	1.5
SF3, 1.2	94	153	-0.5
FS3, 1.2	144	155	1.53
TTH, 1.08	115	144	1.1
TTH, 1.32	124	172	1.2

is activated by the unbuffered block select fwl. The replica bitline is detected by $F1$, and buffered to drive the sense clock signal. If the delay of the replica bitline is matched with the bitline delay and the delay of $F1$, $S2$, $S3$ is made equal to $B1$, $B2$, then the sense clock fires when the bitline swing is the desired amount. Also, if the delay of $B1$, $B2$ is equal to the delay of generating rn (Fig. 9) from the replica bitline, then the wordline pulsewidth will be the minimum needed to generate the required bitline swing. The performance for a 256-row block with 64 columns implementing this replica technique is summarized in Table III. The slack in activating the sense clock is less than 1.5 gate delays, and the maximum bitline swing is within 170 mV. When compared with the implementation based on capacitance ratioing discussed in the previous section, this design is faster by about two-thirds of a gate delay due to the skewing of the wordline driver.

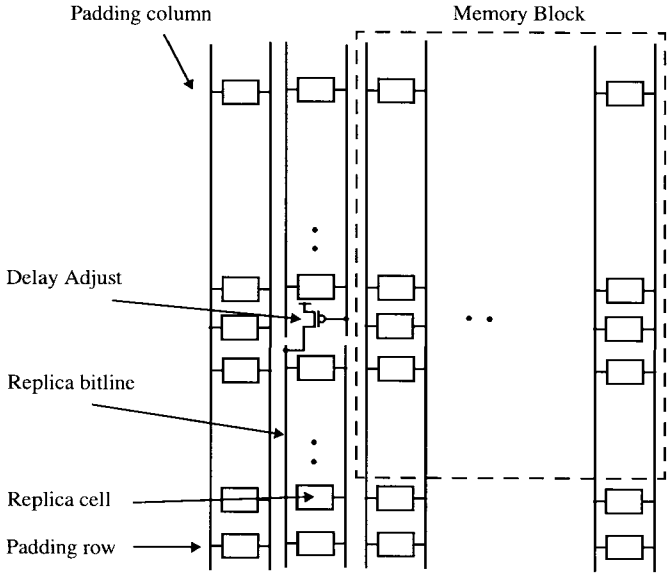

Fig. 11. Layout of a block with replica cell and column.

The power dissipation overhead of this approach is the switching power for the replica bitline which has the same capacitance as the main bitline. The power overhead becomes small for large access width SRAM's. The area overhead consists of one extra column and row, and the extra area required for the layout of the skewed local wordline drivers.

V. MEASURED RESULTS

A. SRAM Test Chip with Replica Feedback Based on Capacitance Ratioing

The replica feedback technique based on capacitive ratioing was implemented in a 1.2-μm process as part of a 2K $\times$ 32 SRAM [25]. The SRAM array was partitioned into eight blocks, each with 256 rows and 32 columns. Making the block width equal to the access width ensures that the bitline power is minimized since only the desired number of bitline columns swing in any access. Two extra columns near the wordline drivers and two extra rows near the sense amps are provided for each block, with the replica column being the second one from the wordline drivers and the replica cell being the second cell in the column (Fig. 11). The extra row and column surrounding the replica cell contain dummy cells for padding purposes, so that the replica cell does not suffer from any processing variations at the array edge. The bitlines in the replica column are cut at a height of 26 rows, yielding a capacitance for the replica bitline which is one-tenth that of the main bitline. Further control for the replica delay is provided for testing purposes by utilizing part of the replica structure above the cut to provide an adjustable current source. This consists of a pMOS transistor whose drain is tied to the replica bitline and whose gate is controlled from outside the block. Since the replica bitline runs only part way along the column, the bitline wiring track in the rest of the column can be used for running the gate control for this transistor. By varying the gate voltage, we can subtract from the replica cell

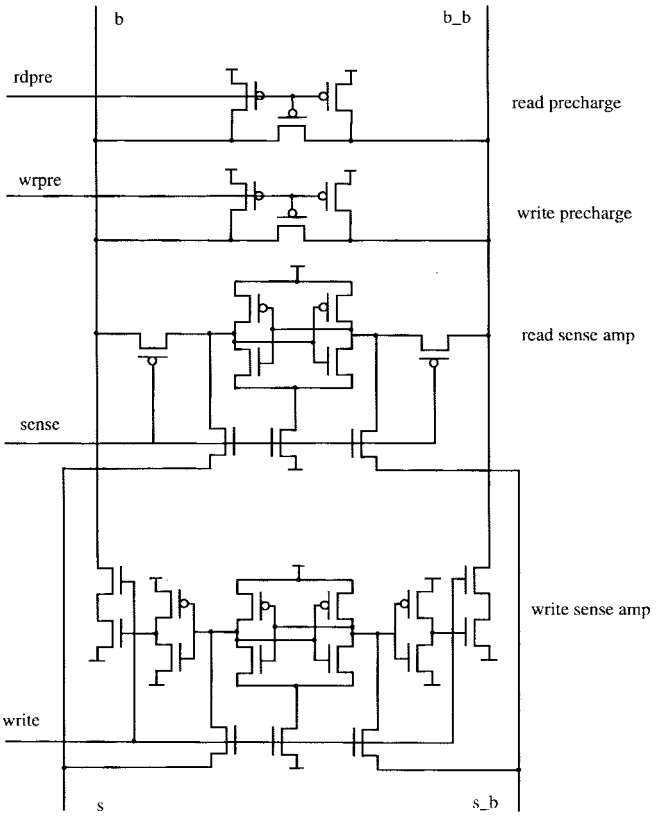

Fig. 12. Column IO circuits.

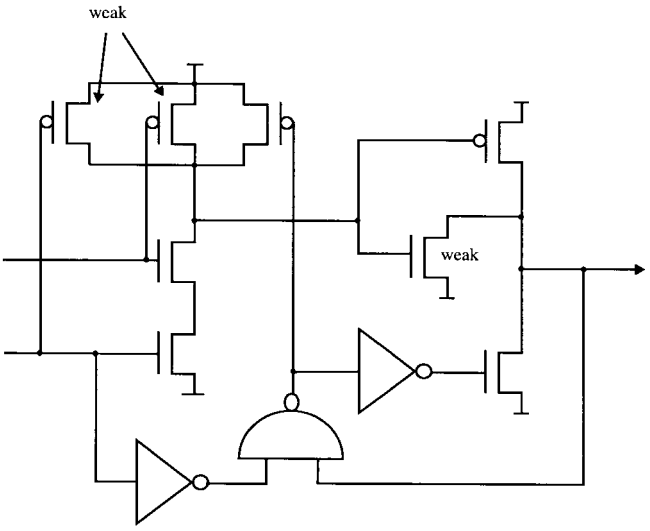

Fig. 13. Pulsed global wordline driver.

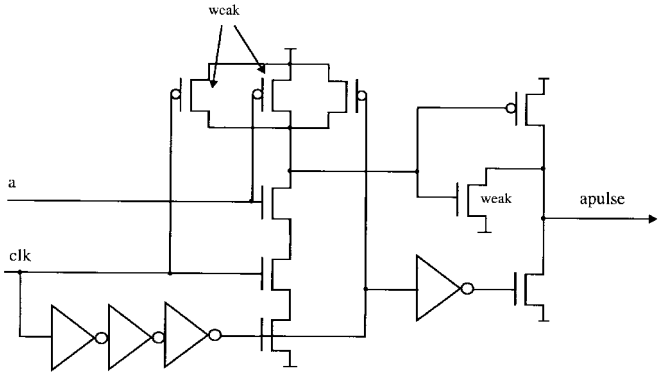

Fig. 14. Level-to-pulse converter.

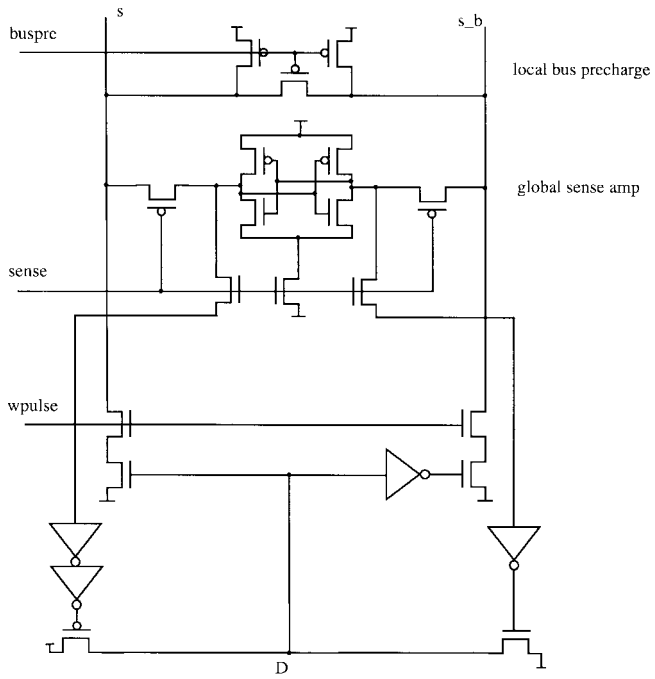

Fig. 15. Global IO circuits.

pulldown current, thus slowing the replica delay element if needed. The inverters $S1$, $S2$, and $F1$ (Fig. 6) are laid close to the replica cell and to each other to minimize wire loading.

Clocked voltage sense amplifiers, precharge devices and write buffers are all laid on one side of the block as shown in Fig. 12. Since the bitline swings during a read are significantly less than that for writes, the precharge devices are partitioned in two different groups, with one activated after every access and the other activated only after a write, to reduce power in driving the precharge signal. Having the block width equal to access width requires that all of this peripheral circuitry be laid out in a bit pitch. The precharge and write control are locally derived from the block select, write enable, and the replica feedback signals.

An 8- to 256-row decoder is implemented in three stages of two-input AND gates. The transistor sizes in the gates are skewed to favor one transition as described in [24] to speed up the clock to wordline rising delay. The input and output signals for the gates are in the form of pulses. In that design, the gates are activated by only one edge of the input pulse, and the resetting is done by a chain of inverters to create a pulse at the output. While this approach can lead to a cycle time faster than the access time, careful attention has to be paid to ensure sufficient overlap of pulses under all conditions. This is not a problem for pulses within the row decoder as all of the internal convergent paths are symmetric. But for the local wordline driver inputs, the global wordlines (the outputs of the row decoder) and the block select signal (output of the block decoder) converge from two different paths. To ensure

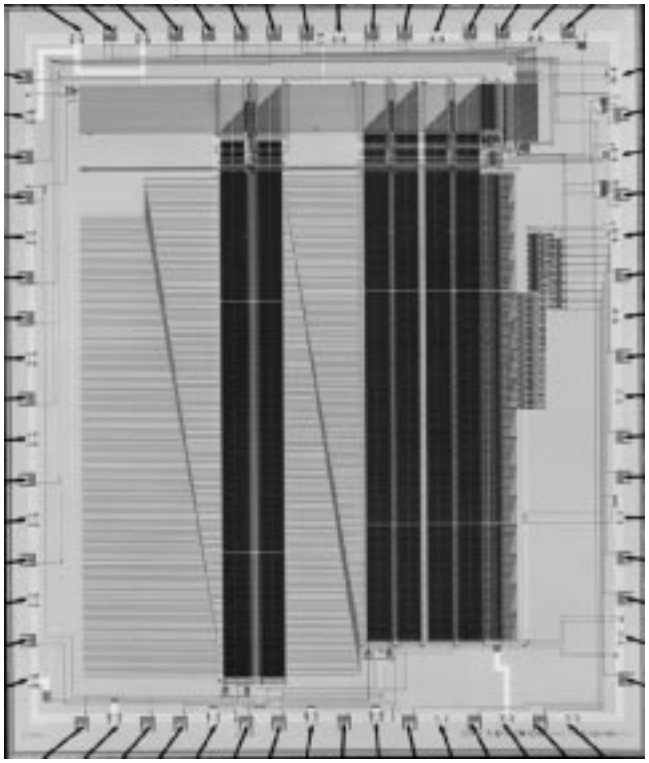

Fig. 16. Die photo of a 1.2-μm prototype chip.

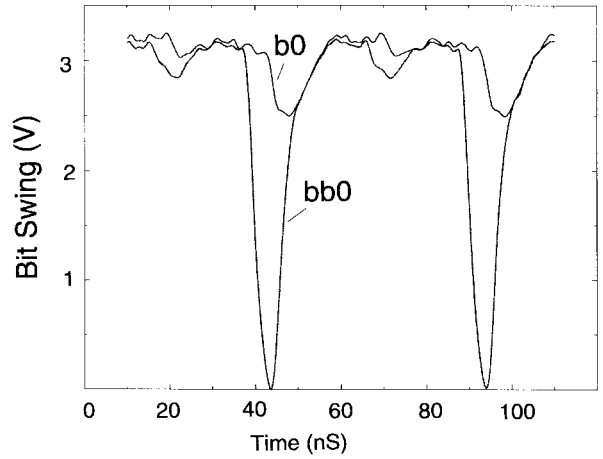

Fig. 17. On-chip probed bitline waveform.

sufficient margins for their overlap, the global wordline signal is prolonged by the modifying the global wordline driver as shown in Fig. 13. Here, the reset signal is generated by ANDing the gate's output and one of the address inputs to increase the output pulsewidth. Clearly, for the gate to have any speed advantage over a static gate, the extra loading of the inverter on one of the inputs must be negligibly small compared to the pMOS devices in a full NAND gate. The input pulses to the row decoder are created from level address inputs by a "chopper" circuit shown in Fig. 14, which could potentially be merged with a 2-bit decode function.

The data bus in the SRAM, which connects the blocks to the IO ports, is implemented as a low-swing, shared differential

TABLE IV
MEASUREMENT DATA FOR THE 1.2-μm PROTOTYPE CHIP

Supply (V)	ΔIO^w (%)	ΔIO^r (%)	ΔB^r (%)	Tacc(nS)	Power (mW)
1.5	16	17	16	60.0	2.88@10MHz
2.5	12	16	12	21.1	31.1@37MHz
3.5	11	19	11	12.8	105@40MHz

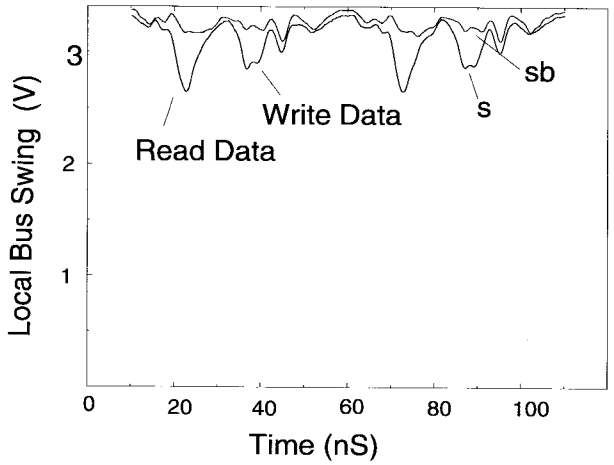

Fig. 18. On-chip probed databus waveforms.

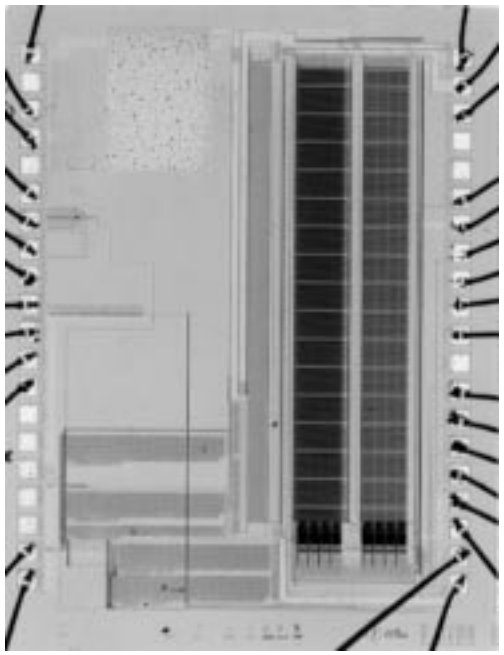

Fig. 19. Die photo of a prototype chip in 0.25-μm technology.

bus. The voltage swings for reads and writes are limited by using a pulsing technique similar to that described in Section IV. During reads, the bitline data are amplified by the sense amps and transferred onto the local data bus through devices $M1$ and $M2$ (see Fig. 12). The swing on the local data bus is limited by pulsing the sense clock. The sense clock pulse width is set by a delay element consisting of stacked

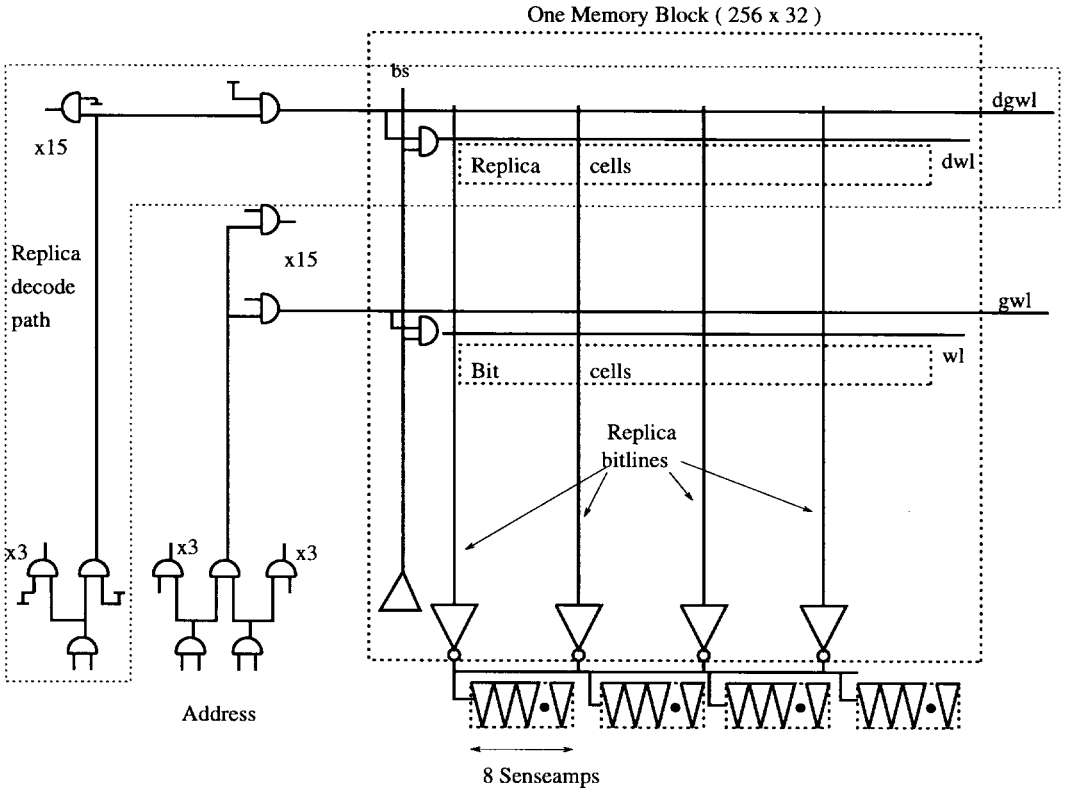

Fig. 20. Architecture of the 0.25-μm prototype SRAM.

pulldown devices connected to a global line which mimics the capacitance of the databus. The data bus mimic line also serves as a timing signal for the global sense amps at the end of the data bus, and has the same capacitance as that of the worst case data bit. The global sense amps, write drivers, and the data bus precharge are shown in Fig. 15. The same bus is also used to communicate the write data from the IO ports to the memory blocks during writes. The write data are gated with a pulse, whose width is controlled to create low swings on the databus. The low-swing write data are then amplified by the write amplifiers in the selected block and driven onto the bitlines (Fig. 12).

Fig. 16 displays the die photo of the chip. Only two blocks are implemented in the prototype chip to reduce the test-chip die area, but extra wire loading on some of the global wordlines and IO lines is provided to emulate the loading in a full SRAM chip. Accesses to these rows and bits are used to measure the power and speed. The bitline waveforms were probed for different supply voltages to measure the bitline swing. Fig. 17 displays the on-chip measured waveform of a bitline pair at 3.5 V supply. The bitline swings are limited to be about 11% of the supply at 3.5 V. Table IV gives the measured speed, power, bitline swing and IO line swing for the chip at three different supply voltages of 1.5, 2.5, and 3.5 V. The on-chip probed waveforms for the databus are shown in Fig. 18 for a consecutive read and write operation. The IO bus swings for writes are limited to 11% of the supply at 3.5 V, but are 19% of the supply for reads. The rather large read swings are due to improper matching of the sense-amp mimic circuit with the read sense amps.

B. SRAM Test Chip with Replica Feedback Based on Cell-Current Ratioing

A cell-current ratioing-based replica technique was implemented in a 0.35-μm (0.25-μm drawn gate length) process from Texas Instruments. A 2-kbyte RAM is partitioned into two blocks of 256 rows by 32 columns. Two extra rows are added to the top of each block, with the topmost row serving as a padding for the array and the next row containing the replica cells.

The prototype chip (Fig. 19) has two blocks, each 256 rows $\times$ 32 bits, to form a 8-kbit memory. The premise in this chip was that the delay relationship between the global wordline and the block select is unknown, i.e., either one of them could be the late arriving signal, and hence the local wordline could be triggered by either. This implies that the replica bitline path too needs to be activated by either the local block select or a replica of the global wordline. The replica of the global wordline is created by mirroring the critical path of the row decoders as shown in Fig. 20. This involves creating replica predecoders, global word drivers, replica predecode, and global wordlines with loading identical to the main predecode and global wordlines [26]. The replica global wordline and the block select signal in each block generate the replica wordline which discharges the replica bitline as described in the previous section. Thus, the delay from the address inputs to the bitline is mirrored in the delay to the replica bitlines. But the additional delay of buffering this replica bitline to generate the sense-amp signal cannot be cancelled in this approach. Hence, the only alternative to

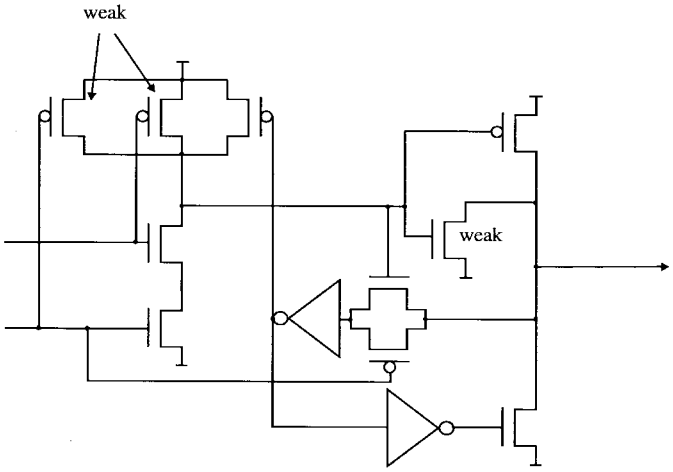

Fig. 21. Modified global wordline driver.

TABLE V
MEASUREMENT DATA FOR THE 0.35-μm PROTOTYPE CHIP

Supply (V)	Tcyc	Power
1.0	6.6nS	3.64mW
0.75	14.1nS	0.99mW
0.5	138nS	50uW
0.45	350nS	17.15uW
0.4	1020nS	5.24uW

minimize this overhead in our prototype was to distribute the replica bitline structure throughout the array. The test chip has one replica bitline structure per eight bitline columns, with just one sense clock buffer stage (Fig. 20). The large swing replica bitline is shielded from the low swing bitlines by cell ground lines running parallel to it. The coupling is further reduced by twisting the bitlines. The row decoder is implemented in the postcharge style for high speeds. The global wordline drivers shown are modified from Fig. 13 to have two fewer transistors and yet achieve the same functionality of ensuring a fast rising edge with a long pulse at the output (Fig. 21).

Table V shows the measured results for the chip. The test chip operates down to 0.4 V at 980 kHz and 5.2 μW dissipation. At 1 V, the chip was measured to dissipate 3.6 mW at 150 MHz. In an SRAM, except for the bitlines and the sense amps, all of the other circuits are completely digital and hence are expected to have a large operating range. The use of latch-type sense amps (Fig. 3) coupled with replica-based generation of their sense clock allows for the wide operating range for the whole SRAM path, as we observe in the table. The replica predecode structure was simulated to consume 15% of the total power in the test chip, and the replica bitline paths (four per block) were simulated to consume 9% of the chip power. By putting the constraint that the block select cannot arrive earlier than the global wordline, the replica decode path and the multiple replica bitline paths can be eliminated to achieve a savings of 20% in power dissipation. Extrapolating from test-chip measurements, we estimate that a typical cache RAM of 16 kbytes will dissipate 7 mW at 150 MHz at 1 V and 10.4 uW at 980 kHz at 0.4 V.

VI. SUMMARY

We presented a replica-based delay circuit which is used to generate the sense clock and control wordline pulse widths to limit bitline swings in low-power SRAM's. The circuit helps to minimize the variations in the speed and power of SRAM's across varying operating conditions, and is especially suitable for SRAM's with large bitline loads. Two flavors of the circuit, one which uses bitline capacitance ratioing and the other which uses cell current ratioing, were discussed. The former yields an implementation with the smallest area overhead and no other changes required in the rest of the SRAM circuitry. The latter allows for implementing skewed wordline driver designs to achieve a slightly faster operation. Prototype chips in 1.2 and 0.35 μm with both of the implementations were measured to have a wide operating range, with deep subvolt operation for the 0.35-μm implementation.

ACKNOWLEDGMENT

The authors thank C.-K. K. Yang and D. Weinlader for help with the design and testing of the chips, C. Lemonds and Texas Instruments for providing the opportunity to fabricate the second chip, Rambus Inc. for providing access to their lasers, and T. Mori of Fujitsu Ltd. for helpful discussions.

REFERENCES

[1] A. P. Chandrakasan *et al.*, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, Apr. 1992.
[2] W. Lee *et al.*, "A 1 V DSP for wireless communications," in *1997 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 92–93.
[3] M. Izumikawa *et al.*, "A 0.25-μm CMOS 0.9-V 100-MHz DSP core," *IEEE J. Solid-State Circuits*, vol. 32, pp. 52–61, Jan. 1997.
[4] K. Ishibashi *et al.*, "A 1 V TFT-load SRAM using a two-step word voltage method," in *1992 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 206–207.
[5] H. Yamauchi *et al.*, "A 0.5V/100 MHz over-Vcc grounded data storage (OVGS) SRAM cell architecture with boosted bit-line and offset source over-driving schemes," in *1996 IEEE Int. Symp. Low Power Electron. Design, Dig. Tech. Papers*, pp. 49–54.
[6] K. Itoh, A. R. Fridi, A. Bellaouar, and M. I. Elmasry, "A deep sub-V, single power-supply SRAM cell with multi-Vt, boosted storage node and dynamic load," in *1996 Symp. VLSI Circuits, Dig. Tech. Papers*, pp. 132–133.
[7] J. D. Meindl *et al.*, "The impact of stochastic dopant and interconnect distributions on gigascale integration," in *1997 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 232–233.
[8] T. Mizuno *et al.*, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFET's," *IEEE Trans. Electron Devices*, vol. 41, pp. 2216–2221, Nov. 1994.
[9] M. Eisele *et al.*, "The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits," in *1996 IEEE Int. Symp. Low Power Electron. Design, Dig. Tech. Papers*, pp. 237–242.
[10] S. Murukami *et al.*, "A 21-mW 4-Mb CMOS SRAM for battery operation," *IEEE J. Solid-State Circuits*, vol. 26, pp. 1563–1569, Nov. 1991.
[11] M. Matsumiya *et al.*, "A 15-ns 16-Mb CMOS SRAM with interdigitated bit-line architecture," *IEEE J. Solid-State Circuits*, vol. 27, pp. 1497–1502, Nov. 1992.
[12] O. Minato *et al.*, "A 20ns 64K CMOS SRAM," in *1984 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 222–223.
[13] S. Yamamoto *et al.*, "256k CMOS SRAM with variable impedance data-line loads," *IEEE J. Solid-State Circuits*, vol. 20, pp. 924–928, Oct. 1985.

[14] K. J. Schultz *et al.*, "Low-supply-noise low-power embedded modular SRAM for mixed analog-digital IC's," in *Proc. 1992 IEEE Custom Integrated Circuits Conf.*, pp. 7.1/1–7.1/4.
[15] P. Reed *et al.*, "A 250MHz 5W RISC microprocessor with on-chip L2 cache controller," in *1997 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 412–413.
[16] T. Hirose *et al.*, "A 20nS 4Mb CMOS SRAM with hierarchical word decoding architecture," in *1990 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 132–133.
[17] S. Flannagan *et al.*, "8ns CMOS 64k $\times$ 4 and 256k $\times$ 1 SRAM's," in *1990 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 134–135.
[18] J. S. Caravella, "A low voltage SRAM for embedded applications," *IEEE J. Solid-State Circuits*, vol. 32, pp. 428–432, Mar. 1997.
[19] A. R. Pelella *et al.*, "A 2ns Access, 500MHz 288Kb SRAM MACRO," in *1996 Symp. VLSI Circuits, Dig. Tech. Papers*, pp. 128–129.
[20] S. Tachibana *et al.*, "A 2.6-ns wave-pipelined CMOS SRAM with dual-sensing-latch circuits," *IEEE J. Solid-State Circuits*, vol. 30, pp. 487–490, Apr. 1995.
[21] S. E. Schuster *et al.*, "A 15-ns CMOS 64K RAM," *IEEE J. Solid-State Circuits*, vol. 21, pp. 704–711, Oct. 1986.
[22] H. Nambu *et al.*, "A 1.8ns access, 550MHz 4.5Mb CMOS SRAM," in *1998 IEEE Int. Solid-State Circuits Conf., Dig. Tech. Papers*, pp. 360–361.
[23] M. Shoji, "Elimination of process-dependent clock skew in CMOS VLSI," *IEEE J. Solid-State Circuits*, vol. 21, pp. 875–880, Oct. 1986.
[24] T. Chappell *et al.*, "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. 1577–1585, Nov. 1991.
[25] B. Amrutur and M. Horowitz, "Techniques to reduce power in fast wide memories," in *1994 IEEE Symp. Low Power Electron., Dig. Tech. Papers*, pp. 92–93.
[26] A. L. Silburt *et al.*, "A 180-MHz 0.8-μm BiCMOS modular memory family of DRAM and multiport SRAM," *IEEE J. Solid State Circuits*, vol. 28, pp. 222–231, Mar. 1993.

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

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

Mark A. Horowitz (S'77–M'78–SM'95) received the B.S. and M.S. degrees in electrical engineering from Massachusetts Institute of Technology, Cambridge, 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 University. In 1990, he took leave from Stanford to help start Rambus Inc., a company that designs high-bandwidth memory interface technology. His research area is in digital system design, and 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. His current research includes multiprocessor design, low-power circuits, memory design, and high-speed links.

Dr. Horowitz is the recipient of a 1985 Presidential Young Investigator Award and an IBM Faculty Development Award, as well as the 1993 Best Paper Award at the International Solid-State Circuits Conference.