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.

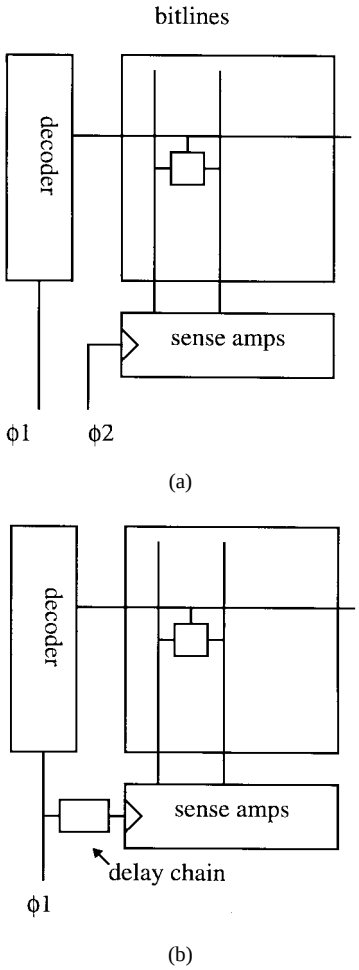

Fig. 1. Common sense clock generation techniques.

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

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

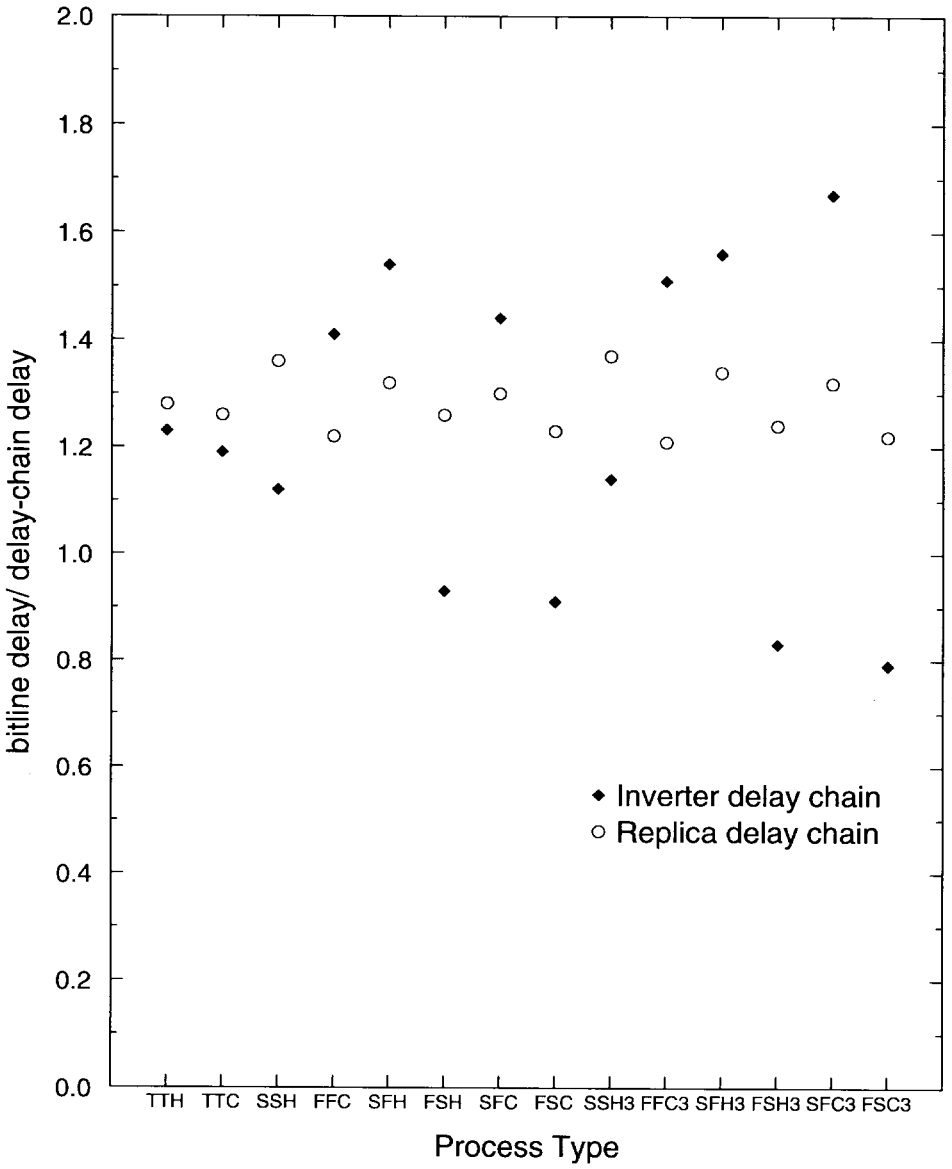

Fig. 2. Delay matching between the bitline delay to generate 120 mV and two delay elements, one based on an inverter chain and the other on a replica cell–bitline combination.

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. 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

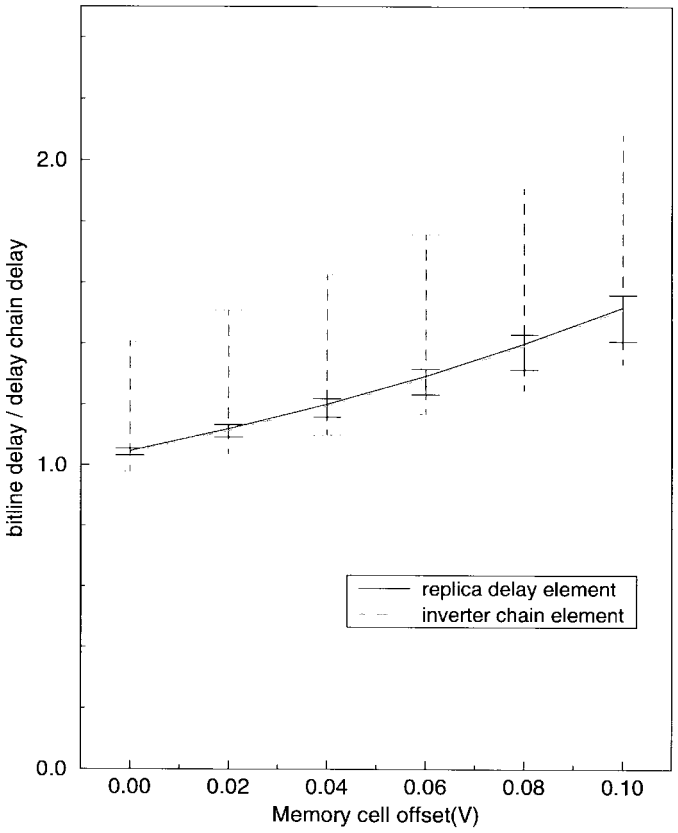

Fig. 3. Matching of the bitline delay with the inverter chain delay and the replica cell–bitline delay across process fluctuations over varying threshold offsets for the accessed memory cell.

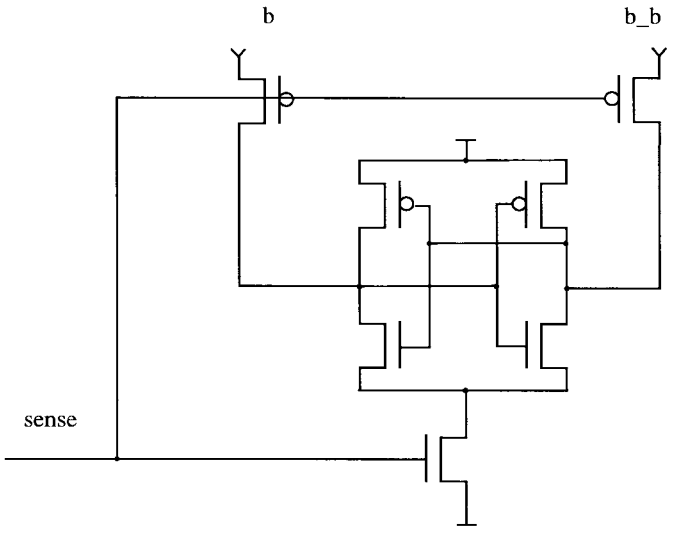

Fig. 4. Latch-type sense amplifier.

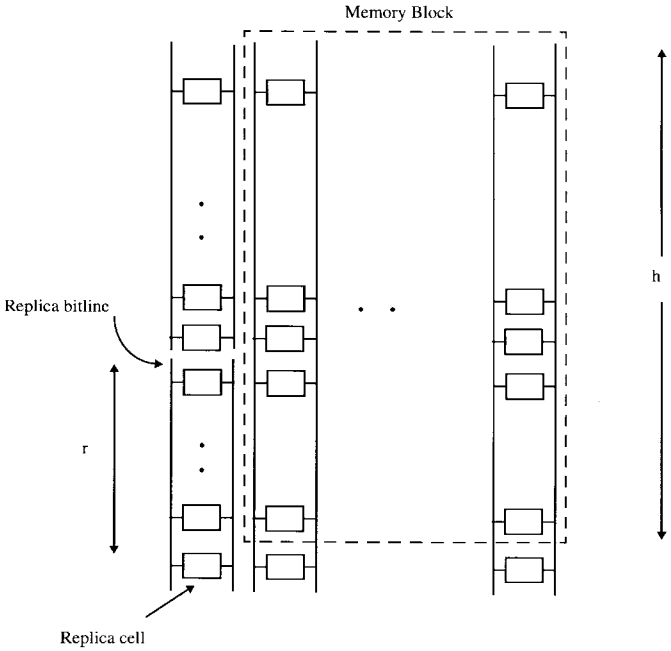

Fig. 5. Design of the replica bitline column.

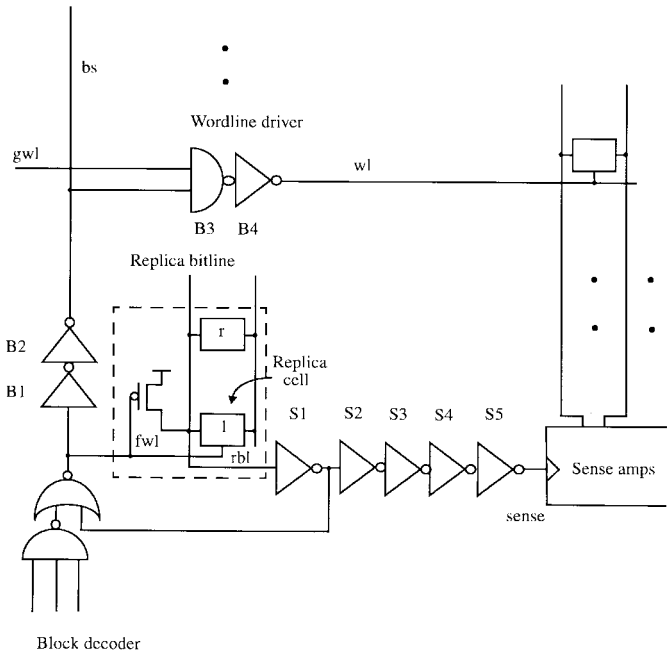

Fig. 6. Control circuits for sense clock activation and wordline pulse control.

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 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

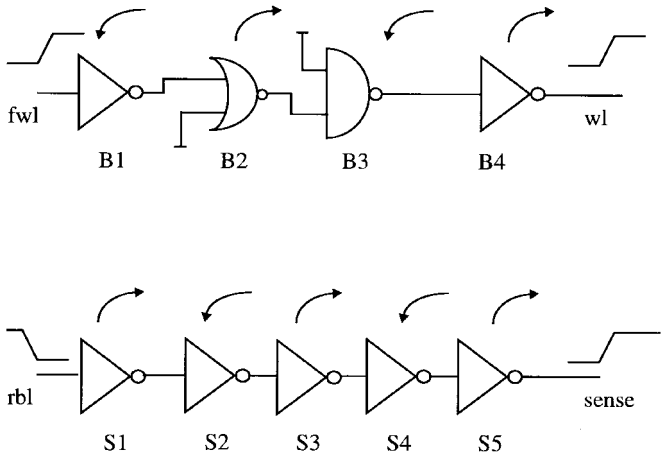

Fig. 7. Delay matching of two buffer chains.

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 $t_{\text{replica}} + t_{Schain}$ delay. If t_{bitline} equals t_{replica} and t_{Bchain} equals t_{Schain}, 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

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

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.