Pipelined Non-Strobed Sensing Scheme for Lowering BL Swing in Nano-scale Memories

Sudhanshu Khanna*
Microcontroller Silicon Development
Texas Instruments
Dallas, TX, USA
email: skhanna@ti.com

Satyanand V. Nalam and Benton H. Calhoun
Department of Electrical and Computer Engineering
University of Virginia
Charlottesville, VA, USA
email: svn2u@virginia.edu, bcalhoun@virginia.edu

***Abstract*— Conventional strobed sense amplifiers (SA) have a fixed offset that dictates the minimum BL droop required during a memory read. BL droop is the major component of memory read delay and energy. In this paper we propose a novel non-strobed sensing scheme that can tradeoff BL droop with SA delay, allowing a memory to operate with lower BL droop, and thus lower energy. We demonstrate the sensing scheme on an 16KB SRAM. Lower BL swing results in 15% lower energy per read operation. The performance penalty due to higher SA delay is avoided by pipelining the SA delay into the next clock cycle. Thus, in addition to lower energy per read, a pipelined non-strobed SRAM is 52% faster than a conventional strobed SRAM and 26% faster than a pipelined strobed SRAM. This is the first work that demonstrates how BL droop can be traded-off with SA delay, enabling lower energy operation. We show the concept, circuit implementation, and simulation results in a commercial 45nm technology node.**

Keywords-SRAM; pipelining; sense amplifier; non-strobed sensing;

I. INTRODUCTION

On-chip SRAMs are a ubiquitous component of modern digital systems ranging from high end server processors to micro-sensor nodes. SRAMs dominate the silicon area of digital processing systems, often accounting more than 50% of the die. Thus lowering the bit-cell footprint can substantially reduce die area and cost. Technology scaling helps by lowering bit-cell area. However, as technology scales, local variations in transistor parameters also increase. The small transistors in SRAM bit-cells are worst affected by this variation. With small bit-cells driving the large bit-line (BL) capacitance, BL droop development is usually the longest phase in an SRAM read operation. BL swing is also the major component of energy per read operation. In conventional SRAMs, the minimum BL droop is dictated by the input referred offset of the strobed SA [1]. SRAM timing is designed such that the WL is turned on for just the amount of time that's needed for the BL droop to equal the SA offset. Self timing, replica BL and replica bitcell techniques help reduce the pessimism in SA enable timing, but ultimately, BL droop is limited by the SA offset.

Non-strobed SAs [2-4] improve SRAM performance by reducing the impact of variation due to internal compensation and by eliminating the SA enable signal. With no strobe, no self timed or replica read scheme is needed, and thus pessimistic SA enable timing margins can be eliminated. We use a novel non-strobed sensing scheme to reduce BL droop. Strobed SAs need a minimum offset below which sensing can fail. Non-strobed SAs do not have such an offset. A higher BL droop makes a non-strobed SA faster, but a low BL droop doesn't cause the sensing to fail; the SA just gets slower. We can use this crucial observation to construct memories with lower BL droop. Lower BL droop not only lowers energy per read operation, and also opens possibilities for higher performance at the same time. In this paper we demonstrate the application of this property on a 16KB SRAM implemented in a commercial 45nm technology node.

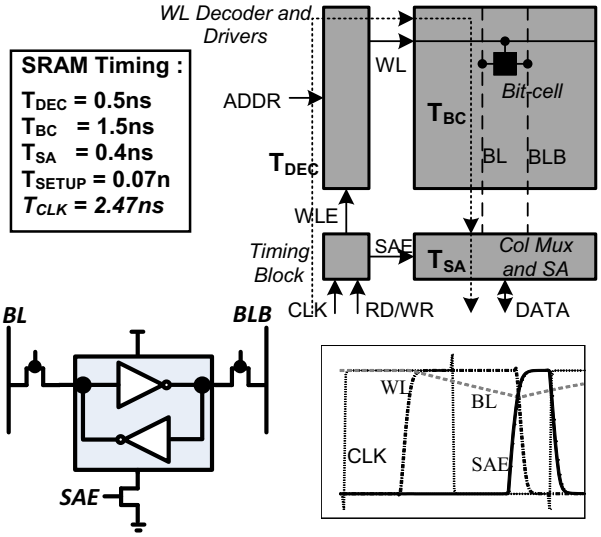

Figure 1: Conventional strobed SRAM with timing arches (dotted). Timing numbers for worst case corner across 20K MC sims.

*The author was part of University of Virginia at the time this work was done.

1063-9667/14 $31.00 © 2014 IEEE
DOI 10.1109/VLSID.2014.31

Drawing from our designs and analysis, this paper makes the following key contributions:

- Shows for the first time how the unique characteristics of a non-strobed SA can be used to drastically reduce BL droop by truncating the WL at the cost of higher SA delay (Section 2).
- Applies truncated WL non-strobed sensing to an SRAM. The lower performance due to higher SA delay is overcome by pipelining the SRAM. Circuits and techniques to pipeline the SRAM and truncated WL non-strobed SAs are described. The final SRAM has 15% lower energy per read than a conventional strobed SRAM. It is also 52% faster than a conventional strobed SRAM and 26% faster than a strobed SRAM with a similar 2-stage pipeline (Section 3).

II. NON-STROBED SENSING WITH TRUNCATED WL PULSE

Sense amplifiers most commonly used in SRAMs are based on a latch with an enable signal or strobe. Figure 1 shows a common circuit configuration for such a SA [1]. Figure 1 also shows the block diagram of a strobed SRAM, its critical timing arches and the associated waveforms. A delay T_{DEC} after the rising edge of the clock the word-line (WL) pulse is generated. The delay T_{DEC} is set by the row decoder and WL driver delay. Upon the arrival of the WL pulse, the bit-cell starts developing droop on the BL (in case of read-0). Also, for the duration of the WL pulse, the BLs are connected to the sense amplifiers (SA) by the column multiplexor. A delay T_{BC} after the rising edge of the WL pulse, the sense amplifier enable (SAE) pulse arrives. T_{BC} is set such that the weakest bit-cell has enough time to pull the BL below the SA offset. Some timing margin to account of differences in timing circuit and bit-cell-BL path also needs to be added to the WL pulse width. BL droop value is limited by the SA offset and timing margin. The SA has a finite but small resolution time, and this sets the width of the SAE pulse, T_{SA}.

Non-strobed sense amplifiers [2-4] have lower SA delay and variation due to internal offset compensation. Since there is no strobe, and the SA is self timed, the timing margin that must be added to a regular SA timing path is also not required. Non-strobed SAs can be single ended, and thus are also useful for single ended 8T SRAM bitcells and other single ended non-volatile memories. However, non-stobed SAs have higher power consumption and need coupling capacitors. In this section we first describe operation of a popular non-strobed SA [2]. Then we describe how it can be used with a novel truncated WL scheme to reduce BL swing.

A. Non-Strobed Sensing Overview

In this work we use a single ended non-strobed regenerative sense amplifier (NSR-SA) [2], however the concepts described in our work can be applied to other AC-coupled non-strobed sense amplifiers like [3] as well. NSR-SA circuit and timing are shown in Figure 2. Like the strobed sensing scheme in Figure 1, the first phase of the read operation is WL decode (of duration T_{DEC}). BLs are also pre-charged during WL decode. At the end of the WL decode phase, the WL of the selected row is pulsed high.

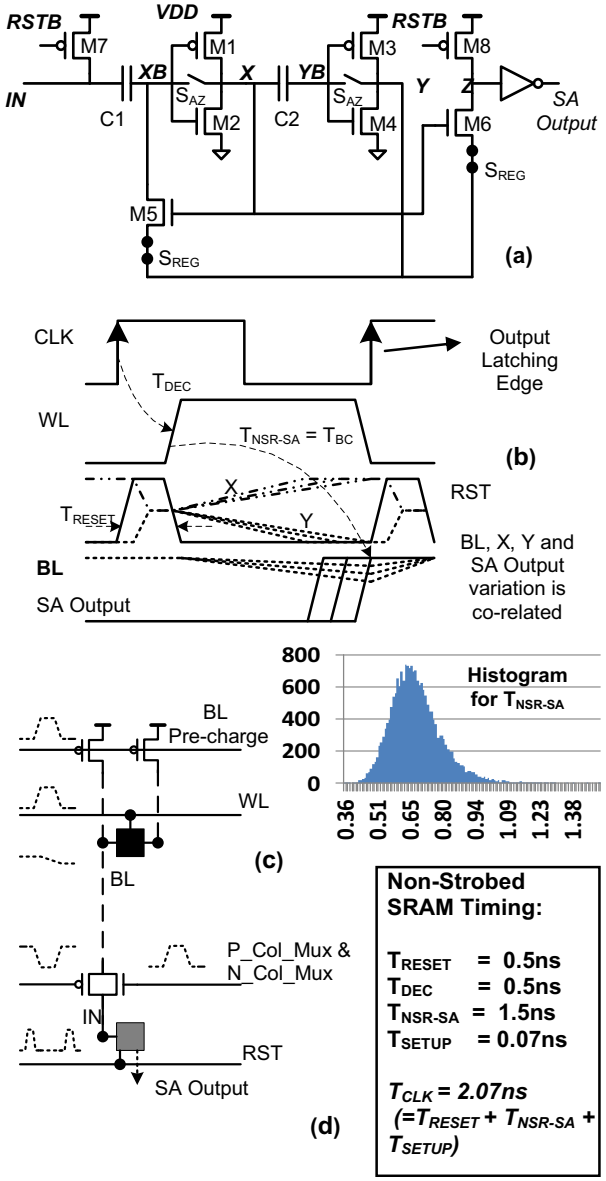

Figure 2: NSR-SA (a) circuit and (b) timing; BL signal shows variation due to variation in bit-cell. SA Output is dependent on bit-cell variation and SA variation (c) Circuits on a single column of a non-strobed SRAM with NSR-SA (d) Timing for a non-strobed SRAM. Timing numbers are at worst corner across 20K MC simulations.

The NSR-SA has an AC-coupled capacitor (C1) at its input followed by an inverter (M1-M2). This capacitor-inverter pair is then repeated (C2, M3-M4) in a cascade. The important difference from strobed timing is that during WL decode, the switches S_{AZ} are turned on (by signal RST), thus resetting inverters M1-M2 and M3-M4 at their respective switching thresholds. Note that reset happens in parallel with WL decode operation. Thus, X and XB are biased at V_M (switching threshold) of M1-M2 and Y and YB at V_M of M3-M4. The pulse RST controls switches S_{AZ} and RSTB (inverted RST) controls switches S_{REG} as shown in Figure 2b. Pre-charge transistor M7 controlled by RSTB (inverted RST) pre-charges the NSR-SA input (IN), and node Z during reset. So, at the end

of the reset, one terminal of C1 is at VDD, and the other terminal is at the V_M of M1-M2. The reset pulse ends just before the WL pulse arrives. For the duration of the WL pulse, the NSR-SA is coupled to the BL by the column multiplexor, as shown in Figure 2c. As soon as the WL pulse goes high the BL starts drooping (for read-0). Thus, BL droop gets coupled to the inverter M1-M2 input XB through the AC coupled input capacitor, and then ripples through C2 and M3-M4. Thus, X goes high and Y goes low. This turns on the regeneration transistor M5, further accelerating the process. Y reaches a low much faster than the BL because of inverter gain and the regenerative feedback provided by M5, thereby realizing sense amplifier behavior. Eventually M6 turns on, discharging the pre-charged node Z. The final output of the NSR-SA, is the inverted version of the logic stored in the bit-cell.

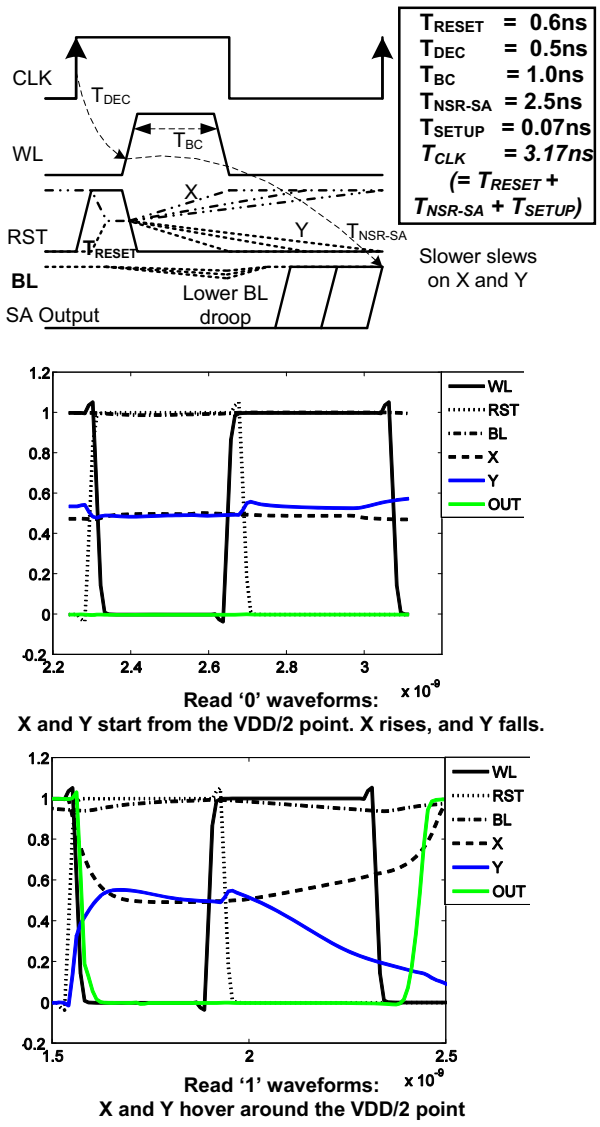

Figure 3: NSR-SA timing with truncated WL for a single read operation; Shorter WL pulse causes lesser BL droop, and thus makes X, Y & SA Output slower than in Fig 2(b).

Thus, a delay T_{NSR-SA} from the rising edge of the WL pulse, a valid output is available. The WL pulse remains high till a valid output develops on the NSR-SA, and thus T_{NSR-SA} is set by the worst case bit-cell & NSR-SA combination. Finally the output of the NSR-SA is latched at the rising edge of the clock. In the case of reading a 1, X and Y stay close to V_M thus M6 never turns on, leaving the SA output low. This circuit and timing description is available in further detail in [2]. The worse case SA Output delay sets T_{NSR-SA}, which is nothing but the WL pulse width.

B. Non-Strobed Sensing with Truncated WL Pulse Width

Now we show how the NSR-SA can be used with a novel timing scheme to lower BL droop at the cost of higher SA delay. Figure 3 shows the timing waveforms of an NSR-SA with a truncated WL pulse. A smaller WL pulse means less BL droop. This means that XB, X, YB, and Y have slower slews than shown in Figure 2. Figure 3 also shows internal waveforms of the SA during a read 0 and 1.

Figure 4 shows a timing and waveform comparison between a regular and truncated WL. With a long WL pulse, XB tends to go low both due to coupling from the drooping BL and due to the regeneration transistor M5. We now truncate the WL pulse such that BL goes low only till X and Y are different enough to turn on M5, but after that point further change in X, XB, Y and YB happens only because of M5 and not because of coupling from the BL.

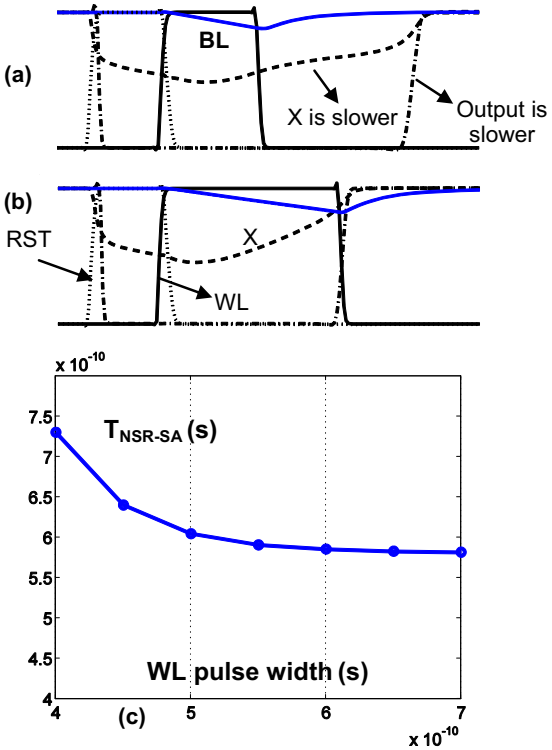

Figure 4: Shows the timing waveforms for (a) truncated and (b) normal WL cases. As WL pulse width is reduced, internal nodes like X and Y becomes slower, and as a result the SA output arrives later. (c) As WL increases, SA delay reduces, but then saturates as WL pulse width exceed the SA delay itself. Simulations at the nominal corner.

In other words, truncating WL pulse simply increases the NSR-SA delay, but doesn't cause the SA to fail. As the WL pulse width increases, SA delay decreases and then flattens out as shown in Figure 4c.

In this manner, we can trade-off the WL pulse width (T_{BC}) with NSR-SA delay (T_{NSR-SA}). To turn off the WL pulse we do not wait for the NSR-SA to develop an output. Rather we turn off the WL pulse lot earlier, with the cost being higher NSR-SA delay (2.5ns as compared to 1.5ns in Figure 2). The key observation in truncating the WL pulse width is that the bit-cell access delay T_{NSR-SA} is dependent on the WL pulse width (T_{BC}). If the WL pulse width is large, the BL droop given to NSR-SA is high, and it evaluates very fast. If the WL pulse width is small, the BL droop given is low, and the NSR-SA evaluates slowly, but correctly. This critical pulse width is much smaller than the WL pulse width of Figure 2.

While truncating the WL reduces the BL droop, it decreases the overall performance because the SA output arrives later. In some applications where reducing the BL droop is critical for energy reduction, this is may a good tradeoff since we can reduce energy without lower the supply voltage or using assist circuits. Also, for memories where the bitcell is slow and BL droop develop dominates the overall cycle time by a large percentage, a truncated WL may still result in higher overall performance. In the SRAM that we look at, truncating the WL increases overall cycle time. Thus, in the next section we use pipelining to overcome the performance penalty while retaining the energy benefit of lower BL swing.

III. NON-STROBED PIPELINED SENSING SCHEME

Truncating the WL pulse reduces the BL droop and energy but reduces overall performance for our SRAM configuration. Figure 2, with long WL pulse had clock period of 2.07ns, while Figure 3, with truncated WL pulse has clock period of 3.17ns. To overcome the performance penalty of truncating the WL pulse we pipeline the SRAM read operation into two cycles. Looking at the waveforms in Figure 3, the read operation of the non-strobed SRAM with truncated WL can be split into 3 phases: NSR-SA reset (T_{RESET}), the truncated WL pulse width (T_{BC}), and the portion of the NSR-SA resolution time after the WL has gone low ($T_{NSR-SA} - T_{BC}$). By controlling the T_{BC} vs. T_{NSR-SA} trade-off, we can set the timing such that first two phases together have duration ($T_{RESET} + T_{BC}$) close to the last phase ($T_{NSR-SA} - T_{BC} + T_{SETUP}$*). We call this pipelining scheme non-strobed pipelined sensing (NSPS). (*T_{INV} and T_{MUX} will be added to this as mentioned later in this section)

The waveforms in Figure 5a show the NSPS timing scheme. The column mux (Figure 5b) at the boundary of bit-cell and SA acts as a dynamic transmission gate latch. Figure 5a shows the SRAM with two back to back read operations. After the first WL pulse goes down, the next cycle (and read access) starts immediately. During this cycle, the next address is accessed, and in parallel, the NSR-SA is evaluating the previous read. BL pre-charge occurs in the beginning of each cycle in parallel with the WL decode. Figure 5b shows the circuits associated with a single column of the SRAM. When the WL goes high, the NSR-SA is coupled to the BL. When the WL goes down, the NSR-SA is decoupled from the BL, but it input remains at the same level. We optimize T_{BC} such that first two phases together have duration ($T_{RESET} + T_{BC}$) close to the last phase ($T_{NSR-SA} - T_{BC} + T_{SETUP} + T_{INV} + T_{MUX}$). Thus the clock period of the 2-stage pipeline is set at 1.625ns (=$T_{NSR-SA} - T_{BC} + T_{SETUP} + T_{INV} + T_{MUX}$). By using NSPS, we make the SRAM clock cycle 1.5X faster (=2.47ns/1.625ns) faster than conventional strobed SRAM. T_{BC} and T_{NSR-SA} are co-related. A certain value of T_{BC} is selected such that T_{NSR-SA} (after montecarlo simulations) is within the clock period requirement. Both bitcell and SA are in the montecarlo simulations so variations in transistors of both circuits are comprehended.

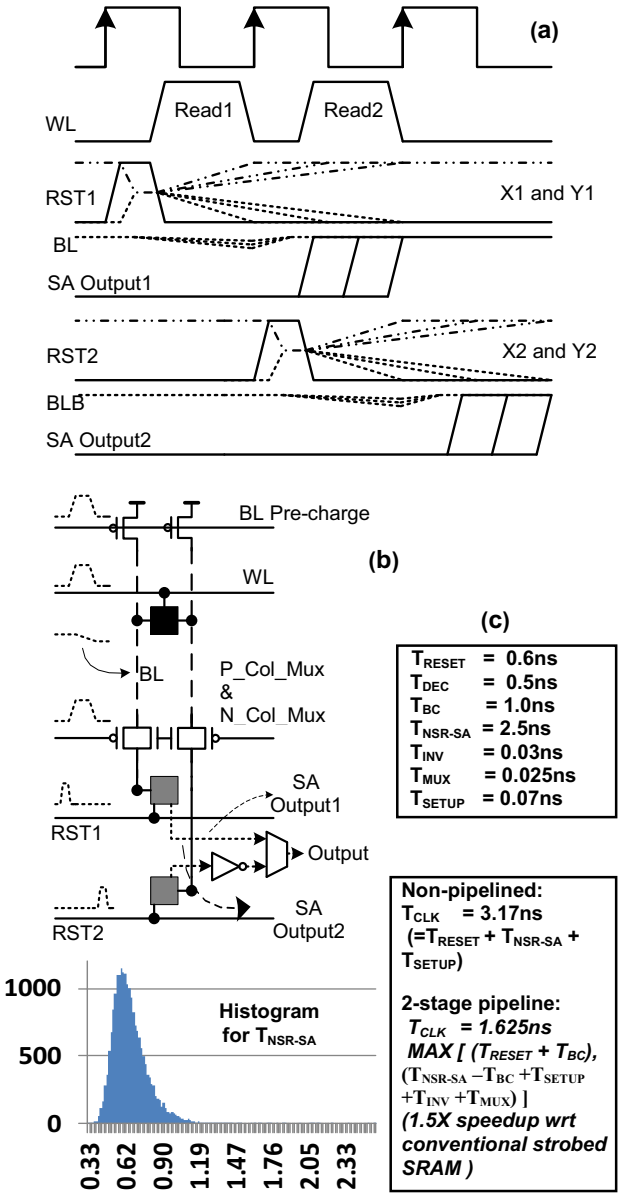

Figure 5: (a) NSPS timing for 2 back to back read operations; two sets of SA signals correspond to two NSR-SAs used in alternate cycles (b) Circuits in a single column of the NSPS SRAM (c) Pipelined timing for NSPS SRAM with the regular read operation split into 2 cycles; The reset (T_{RESET}) and BL droop development (T_{BC}) are balanced with the *portion* of sensing time *after* WL goes low ($T_{NSR-SA} - T_{BC} + T_{SETUP} + T_{INV} + T_{MUX}$).

Note that with a strobed SRAM, pipelining can be accomplished by doing decode and BL droop in the first clock cycle and sensing in the next [5]. However, BL droop development phase is a long unbalanced pipeline stage that limits the benefit of pipelining in strobed SRAMs. In Figure 1, a strobed SRAM can be pipelined by putting the SA in a second clock cycle, but the clock period is set by $T_{DEC} + T_{BC}$ to 2.04ns. The NSPS SRAM is 1.26X faster than a pipelined strobed SRAM as well while retaining energy benefit of ~15% per read operation.

A circuit issue in implementing NSPS is the timing of the reset pulse. Let us take the example of the back to back read operations in Figure 5b. During cycle 2, the NSR-SA is resolving the BL signal from the read access of cycle 1. However, during cycle 2, the NSR-SA also needs to be reset and connected to the BL for sensing the read in the current read access of cycle 2. This is a structural hazard in the SRAM pipeline. To resolve this structural hazard, we use a Sense Amplifier Alternation (SAA) scheme. The scheme involves having two NSR-SAs per column and alternating between them during consecutive read accesses. One NSR-SA is connected to BL, and the other to BLB. This way, both NSR-SAs get an entire cycle to develop the output. In every cycle one NSR-SA is in reset and BL droop development mode, and the other is in signal resolving mode. In the design of the NSPS scheme, it is important to ensure that once the WL goes low and the NSR-SA gets decoupled from the BL, the node IN remains at the drooped level where it was when the WL went low. For this, we need to ensure that the charge injection and clock feed-through noise injected into the node IN by the column multiplexor is low. To lower the noise, we use transmission gate based column multiplexors. Note that with the addition of inverter and multiplexor, the second phase delay becomes ($T_{NSR\text{-}SA} - T_{BC} + T_{SETUP} + T_{INV} + T_{MUX}$). The design was also optimized by adding a transmission gate mux and inverter into a dual NSR-SA custom circuit block.

Paper	Work	Energy Per Read Op	Read Cycle Time	BL Droop at Nom
[1]	Conventional Strobed SRAM	3.5fJ	2.47ns	255mV
[5]	Pipelined Strobed SRAM	3.5fJ	2.04ns	255mV
[2]	Conventional Non-Strobed SRAM	4.2fJ	2.07ns	270mV
This Work	**NSPS: Pipelined Non-Strobed SRAM**	**3.0fJ**	**1.625ns**	**180mV**

Table 1: Results summary. NSPS is 1.5X faster than a conventional strobed SRAM and 1.26X faster than a pipelined strobed SRAM. At the same time, it is has 15% lower energy per read operation. Novel use of non-strobed sensing with truncated WL pulse helps lower BL droop, lower energy and provide speedup using efficient pipelining for SRAMs.

Figure 1 compares the Non-Strobed Pipelined Sensing (NSPS) SRAM with a conventional strobed SRAM, pipelined strobed SRAM and conventional non-strobed SRAM. For all simulations, appropriate parasitics are used on BLs, WLs and control signals. Major contributors to energy numbers are BL pre-charge, SA and control signal switching. The energy per read operation of a regular strobed SRAM is 3.5pJ for 32-bitwidth, 8 to 1 column mux-ing, and 512 rows. For NSPS the energy per operation is 3.0pJ. NSPS has energy overhead due to use of the non-strobed SA which burns static power during its reset phase and also while evaluating a '1'. However, lower droop on the BLs due to the truncated WL results in massive energy saving that compensates for the use of the more power hungry NSR-SA as compared to a strobed SA. Overall, the NSPS scheme has 15% lower energy per read as compared to the regular strobed SRAM. In [2], NSR-SA area and conventional strobed SA area was reported as $19um^2$ and $12um^2$ respectively. Since we use two NSR-SAs, we can pessimistically estimate our SA area to be about 2.5X the NSR-SA area. Taking bit-cell area to be $0.25um^2$ and a 512x256, 32-bit output SRAM with 8 to 1 column muxing, and 75% array efficiency, on the SRAM level, there is a 2.9% overhead of using NSPS because of our dual SA scheme.

Pipelining has been previously used in SRAMs to a limited extent. The work in [5] presents a pipelined cache architecture based on splitting the SRAM read into three phases shown in Figure 1. The authors highlight the long unbalanced BL droop development phase and propose a technique to reduce its duration thereby balancing the three stages. To lower the BL droop development time the authors split the SRAM into smaller banks, thereby reducing BL length (thus capacitance). However, splitting an SRAM into banks results in significant area overhead as the peripheral circuits have to be replicated for each bank. The second reason which restricts pipelining in an SRAM is the system level impact. Though pipelining an SRAM results in higher throughput, pipelining has the drawback that the address must be provided to the SRAM by the processor one cycle in advance. This is easily accomplished in the case of instruction memories, where most accesses are sequential in nature. The next address is predictable except in the case when the processor encounters a branch instruction that is taken. In case of a branch instruction that is taken, a processor using a pipelined memory is likely to have one extra stall as compared to the usual number of stalls associated with a taken conditional branch. Thus, unless the speedup due to pipelining is large, the impact of the extra stall restricts the applicability of pipelining an SRAM. The long unbalanced BL droop development phase has till now limited the speedup provided by pipelining a strobed SRAM, but our technique can achieve a much higher speedup that results in an overall win even after accounting for the extra stall.

Across a wide range of benchmarks, the average number of branch instructions is reported to be around 20%, and out of these branches 65% are taken on an average [7]. This means that 13% (=20*0.65) of the instructions are branches that are taken. Assuming the worst case scenario that a processor has no stalls for a branch while using a normal instruction cache and one stall for a conditional branch while using the pipelined instruction cache, the CPI while using a pipelined instruction cache would be 1.13 (=1*0.87+2*0.13). Table 1 shows that the speedup provided by the NSPS SRAM is 1.5X. When used in a processor as an instruction cache, this SRAM would have an effective speedup of around 1.32X (=1.5X/1.13).

IV. CONCLUSION

In this paper we show how non-strobed sensing can be used with a truncated WL to reduce BL swing and thus memory read energy at the cost of higher SA delay. This can act as a practical energy reduction technique for SRAMs or other memories because it doesn't change the bitcell voltages. We apply this technique to a 16KB SRAM where we recover the lost performance due to higher SA delay by pipelining the SA into the next clock cycle. The resulting NSPS SRAM is 15% lower in read energy and 52% faster than a strobed SRAM. As compared to a 2-stage pipelined strobed SRAM, NSPS is 26% faster while having the same 15% energy benefit. The area penalty estimate is less than 3%.

REFERENCES

[1] Yeung, J.; et al; "Robust Sense Amplifier Design under Random Dopant Fluctuations in Nano-Scale CMOS Technologies," SOC Conference, 2006

[2] Verma, N.; et al; "A High-Density 45nm SRAM Using Small-Signal Non-Strobed Regenerative Sensing," ISSCC 2008.

[3] Qazi, M.; et al; "A 512kb 8T SRAM macro operating down to 0.57V with an AC-coupled sense amplifier and embedded data-retention-voltage sensor in 45nm SOI CMOS," ISSCC 2010

[4] Ryan, J.; et al; "An Analytical Model for Performance Yield of Nanoscale SRAM Accounting for the Sense Amplifier Strobe Signal, " ISLPED 2011

[5] Shakhsheer, Y.; et al; "A 90nm data flow processor demonstrating fine grained DVS for energy efficient operation from 0.25V to 1.2V," CICC 2011

[6] A. Agarwal, et al; "Exploring High Bandwidth Pipelined Cache Architecture for Scaled Technology," Design, Automation and Test in Europe, 2003.

[7] Lalja, D. J.; "Reducing the branch penalty in pipelined processors," Computer , Jul 1988