A Power-Aware SWDR Cell for Reducing Cache Write Power

Yen-Jen Chang
CSIE Department
National Taiwan University
Taipei, Taiwan 106
886-2-33667569
ychang@nets.csie.ntu.edu.tw

Chia-Lin Yang
CSIE Department
National Taiwan University
Taipei, Taiwan 106
886-2-23625336 ext. 411
yangc@csie.ntu.edu.tw

Feipei Lai
CSIE & EE Department
National Taiwan University
Taipei, Taiwan 106
886-2-33665001
flai@cc.ee.ntu.edu.tw

ABSTRACT
Low power caches have become a critical component of both hand-held devices and high-performance processors. Based on the observation that an overwhelming majority of the data written to the cache are '0', in this paper we propose a power-aware SRAM cell with one single-bitline write port and one differential-bitlines read port, called *SWDR* cell, to minimize the cache power consumption in writing '0'. The SWDR cell uses a circuit-level technique, which is software independent and orthogonal to other low power techniques at architecture-level. Compared to the conventional SRAM cell, the experimental results show that without compromise of both performance and stability, the SWDR cell can result in 73%~92% reduction in average cache write power dissipated in bitlines.

Categories and Subject Descriptors
B.3.1 [**Memory Structures**]: Semiconductor Memories–*Static memory (SRAM)*

General Terms
Design

Keywords
Low Power, Cache, SRAM, Circuit-Level, Write Power

1. INTRODUCTION
Most microprocessors employ the caches to bridge the performance gap between the processor and main memory. However, the cache accesses usually contribute significantly to the total power consumption of the chip. By examining the write data of the benchmarks, we first observe an overwhelming majority of the cache write bits are '0', and then propose a novel power-aware SRAM cell that can reduce the cache power dissipated in writing '0' drastically. Because the proposed cell consists of one write port with single-bitline and one read port with differential-bitlines, it is referred to as *SWDR* cell throughout this paper. The contributions of the proposed SWDR cell are as follows. (1) Unlike the conventional SRAM cell where the power dissipated in both writing '0' and '1' are the same, the SWDR cell can prevent the single write bitline from being discharged if the written value is '0'. Therefore, the write '0' power is far less than the write '1' power in the SWDR cell. (2) Writing cell state from low to high is considerably difficult in single-bitline configuration because it presents conditions similar to that of the read mode. Instead of the traditional boosted wordline technique [1], the SWDR cell uses a tail transistor to disconnect the pull-down path, such that writing cell state to high is easy to be achieved.

We evaluate the 0/1 distribution of the write data from the *SPEC2000* benchmarks, and all of the power consumption data are obtained from the *HSPICE* simulation of the extracted layout in TSMC 0.35μm technology with a 3.3V supply. The results show that by minimizing the power dissipated in writing '0', the SWDR cell can reduce the average cache write power dissipated in bitlines up to 92% without impairing the cache stability and performance, but with *5.8%* cell area increase.

The rest of this paper is organized as follows. Section 2 presents our motivation and approach. In Section 3, we describe the circuitry of the proposed SWDR cell, and then the impacts of the SWDR cell on stability, access delay, cell area and write power consumption are provided in Section 4. Experimental results are given in Section 5, and Section 6 offers some brief conclusions.

ISLPED'03, August 25–27, 2003, Seoul, Korea.

2. PRELIMINARY
The power consumption of cache read can be reduced significantly by using a pulsed-wordline technique [2] to turn off the wordline when a sufficient voltage differential has developed on the bitlines. Compared to the cache read, in order to flip the cell state correctly, the cache write typically consumes considerably large power due to the full voltage swing on the bitlines. Although the frequency of cache writes is less than that of cache reads, due to the large power consumption, the impact of cache write on the total cache power consumption cannot be ignored, especially for the data caches or the instruction caches with a high miss ratio.

2.1 0/1 Distribution of the Write Data
Fig. 1 shows the proportion of '0' bits to the total cache write bits (referred to as *write-zero rate*) examined from the execution of the SPEC2000 benchmarks. From this figure, around 85% of the instruction write bits are '0', and over 90% of the data write bits are '0'. Motivated by the extremely asymmetric distribution of '0' and '1' bits in the write data, we propose a novel power-

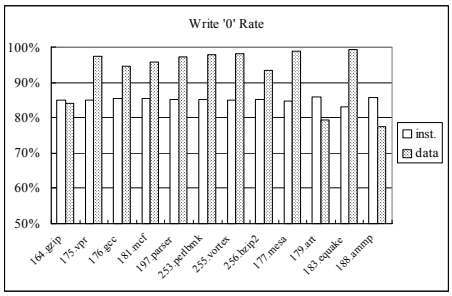


Figure 1. Write-zero rates of instruction and data caches for *SPEC2000*.

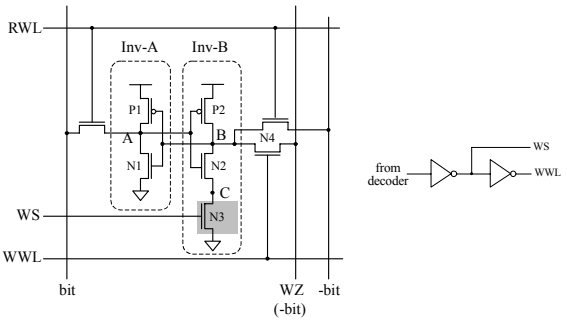
aware SWDR cell, in which the power dissipated in writing '0' is much less than the power dissipated in writing '1'. By exploiting the prevalence of '0' bits in the write data, the proposed cell can effectively reduce the cache power consumption during a write.

2.2 Related Work

The half-swing pulse-mode technique [3] was used to reduce the bitlines swing during cache writes by half of the conventional technique. However, using a $V_{DD}/2$ reference for bitlines potentially lead to cell instability during the cache read. In [4], a dynamic zero compression scheme was proposed to reduce the energy required for cache accesses by only writing and reading a single bit for every zero-value byte. The major disadvantage of the dynamic zero compression is that the power reduction is limited by the cluster of '0' bits. This is especially unfavorable for instruction due to the instruction format.

3. POWER-AWARE SWDR CELL

Fig. 2 shows the schematic of the proposed SWDR cell and its relative signals, where write select (*WS*), write wordline (*WWL*) are used to select a cell for writing, and the data line (*WZ*) is used for signaling whether the current operation is writing '0' or not.



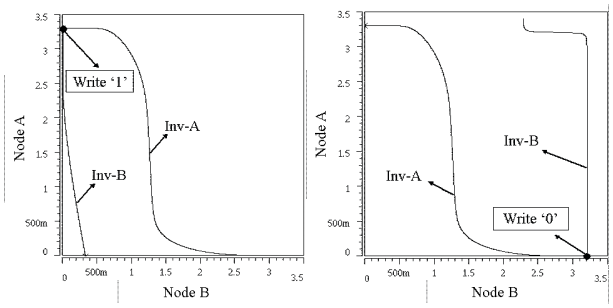
(a) **(b)**

Figure 2. (a) Power-aware SWDR cell. (b) The generation of write select *(WS)* and write wordline *(WWL)* signals.

Read mode: In the read mode, *WWL* is held to 0 and the tail transistor *N3* is turned on to activate Inv-B. Because we consider the cell with split one read port and one write port, the read port has read wordline (*RWL*) for cell selection, which is different from the write wordline (*WWL*) of the write port. Therefore, the read operation of the SWDR cell is the same as that of the conventional cell.

Write '1' mode: In the write '1' mode, node *B* must be written to low that is done by setting *WZ* to 0 and asserting *WWL*. The first possible case is writing the cell state from '1' to '1' (1->1). Because both node *B* and *WZ* are 0, no state transition arises in this case. Another possible case is 0->1. In this case because access transistor *N4* has much larger conductance than *P2*, it is easy to flip the cell state from '0' to '1' by discharging node *B* through *N4*. The electrical characteristics of the inverters in the SWDR cell during the write '1' mode are shown in Fig. 3(a).

Write '0' mode: In the write '0' mode, node *B* must be written to high that is done by setting *WZ* to V_{DD} and asserting *WWL*. The first possible write pattern is 0->0. Because both node *B* and *WZ* are high, no state transition arises in this case. Another possible write pattern is 1->0, which is considerably difficult in single-bitline configuration because it presents conditions similar to that of the read mode. Boosted wordline technique [1] is a traditional solution to this problem, but it potentially induces the unreliable read and hardware overheads. Instead of the boosted wordline technique, the SWDR cell uses a tail transistor *N3* to facilitate writing node *B* from low to high. In this case, because *N3* is turned off by *WS* before asserting *WWL*, the pull-down path through driver transistor *N2* is disconnected. Therefore, it is easy to flip the cell state from '1' to '0' by charging node *B* through *N4*. The electrical characteristics of the inverters in the SWDR cell during the write '0' mode are shown in Fig. 3(b).



(a) **(b)**

Figure 3. Electrical characteristics of the inverters in the SWDR cell during (a) the write '1' mode and (b) the write '0' mode.

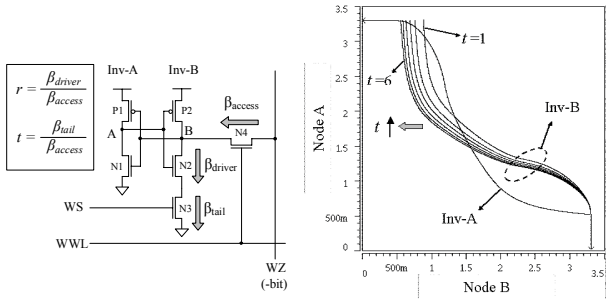
4. STABILITY, ACCESS DELAY AND WRITE POWER REDUCTION

In this section, we first estimate the impacts of the SWDR cell on the stability and performance (i.e., access delay). With the same stability and performance as the conventional SRAM cell, the write power reduction of the SWDR cell is provided.

4.1 Stability

In general, the static noise margin (SNM) is an important parameter in determining the cell stability. The SNM of SRAM cell is defined as the maximum value of noise that can be tolerated by the cross-coupled inverters before altering state. A basic understanding of the SNM is obtained by drawing and mirroring the inverter characteristics, and then finding the maximum possible square between them.

As shown in Fig. 4(a), the major difference between the conventional cell and the SWDR cell is the Inv-B, in which the additional tail transistor *N3* results in an asymmetrical inverter



(a) (b)

Figure 4. (a) The SNM of the SWDR cell (SNM_{SWDR}) is determined by the *cell ratio r* and *tail ratio t*. (b) Graphical representation of the SNM_{SWDR}. It increases with the *tail ratio t* if the cell ratio is fixed (*r=1* in this case).

pair that potentially degrades the stability. According to [5], the SNM of the traditional SRAM cell (SNM_{Conv}) increases with the *cell ratio r*, defined by $r=\beta_{driver}/\beta_{access}$. β_{driver} and β_{access} are the *W/L* ratios of driver transistor (*N2*) and access transistor (*N4*), respectively. In the SWDR cell, because the tail transistor *N3* is on the critical path in driving node *B* to low, besides the cell ratio, the SNM of the SWDR cell (SNM_{SWDR}) is also determined by the ratio of β_{tail} to β_{access}, referred to as *tail ratio* $t=\beta_{tail}/\beta_{access}$, in which β_{tail} is the *W/L* ratio of tail transistor *N3*. Fig. 4(b) shows how the SNM_{SWDR} varies with the tail ratio. The SNM_{SWDR} would increase with the tail ratio if the cell ratio is fixed. Keeping the SNM_{SWDR} the same as the SNM_{Conv} can be achieved by appropriate choice of *r* and *t*. Fig. 5 shows the SNM_{SWDR} in different combinations of *r* and *t*. The key observation is that when the cell ratio is *3* and the tail ratio is *5*, the SNM_{Conv} and SNM_{SWDR} are almost the same value *654mV*.

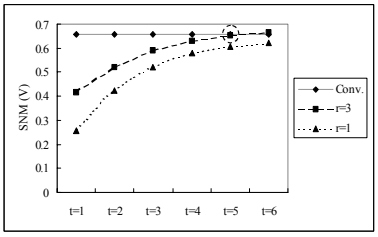


Figure 5. The SNM_{SWDR} in different combinations of cell ratio *r* and tail ratio *t*.

4.2 Access Delay

Read delay: We define the read delay as the elapsed time from asserting *RWL* to the sufficient bitline swing for correct data sensing.

(1) In the case of read '0', the *bit* line would be discharged to low through the driver transistor *N1* of Inv-A. This path is identical to the conventional cell in reading '0'. Thus, the read '0' delays are of the same *1.2385ns* for both the conventional and SWDR cells.

(2) In the case of read '1', the *–bit* line would be discharged to low through the driver transistor *N2* and tail transistor *N3* of Inv-B. Because *N3* is always turned on in the read mode, similar to SNM, the read '1' delay also depends on both the cell and tail ratios. For a better SNM, the cell ratio is fixed to be 3 and Fig. 6 shows how the read '1' delay varies with the tail ratio. It is clear

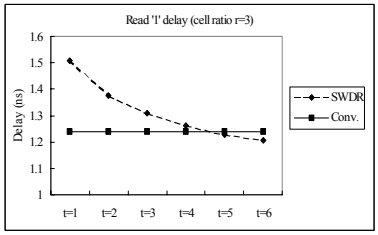


Figure 6. The read '1' delay varies with the tail ratio if cell ratio is fixed to be 3.

from this figure that when the tail ratio is 5, the read '1' delays of both the conventional and SWDR cells are almost the same *1.23ns*.

Write delay: The write delay is defined as the elapsed time from asserting *WWL* to the states of both nodes *A* and *B* become steady. There are four cases in write operation: writing the cell state from '0' to '0' (0->0), '0' to '1' (0->1), '1' to '0' (1->0) and '1' to '1' (1->1). Due to no state transition in cases of 0->0 and 1->1, we only consider the write delay in cases of 0->1 and 1->0.

(1) In the case of 0->1, by setting *WZ* to 0 and then asserting *WWL*, node *B* with initial high state would be discharged to low. Compared to the traditional write port with differential bitlines, because in the SWDR cell the state transition is driven by only one path, the 0->1 write delay of SWDR cell is slightly larger than that of the conventional cell, as shown in Table 1. In determining write cycle, this minor difference can be ignored.

(2) In the case of 1->0, by setting *WZ* to V_{DD} and then asserting *WWL*, node *B* with initial low state would be driven to high to flip the state of node *A*. As shown in Table 1, because *N3* is turned off in the write mode, the 1->0 write delay of the SWDR cell is even smaller than that of the conventional cell.

Table. 1. Write delay summary.

Write Delay (ns)	0->1 Write Delay	1->0 Write Delay
Conv.	0.7531	0.7533
SWDR	0.7587	0.7512

4.3 Write Power Reduction

Based on the analyses described above, we conclude that the SWDR cell does not compromise either stability or access delay when the cell ratio is *3* and tail ratio is *5*. In the SWDR cell, *WS* signal is used to guarantee the correct write operation. Because it shares the load capacity of *WWL*, the additional *WS* does not induce any power penalty. Table 2 shows the column power consumption for various write patterns. In the conventional cell, regardless of write pattern, the column power consumptions are the same. Compared to the conventional cell, in the 1->0 write pattern, the SWDR cell reduces the column power consumption by *97.21%*. Due to no state transition and bitline discharge, even *98.82%* the column power reduction can be achieved in the 0->0

Table. 2. Summary of write power dissipated in one column.

Column Power (mW)	Conv.	SWDR	***Reducction***
1->0	4.65E-01	1.30E-02	*97.21%*
0->0	4.37E-01	5.17E-03	*98.82%*
1->1	4.35E-01	4.20E-01	*3.50%*
0->1	4.92E-01	4.64E-01	*5.70%*

write pattern. Consequently, in writing '0' (1->0 or 0->0), the SWDR cell consumes far less power than the conventional cell.

4.4 Cell Area

Both the conventional and SWDR cells have 8 transistors for one read port and one write port. As described in analysis of SNM, to compensate the stability loss due to the asymmetrical inverter pair in the SWDR cell, we have to enlarge the cell ratio and tail ratio. Compared to the conventional cell, the SWDR cell area is increased from $114.21\mu m^2$ to $120.85\mu m^2$. Most area overhead is introduced by the large driver transistor *N2* and tail transistor *N3* in Inv-B that imposes around a *5.8%* cell area overhead.

5. EXPERIMENTAL RESULTS

5.1 Benchmarks and Baseline Cache

To investigate the impact of the SWDR cell on cache write power, we use *SimpleScalar* to evaluate the 0/1 distribution of the write data for *SPEC2000* benchmarks. We use a baseline with split instruction and data caches, which are a 32KB, 2-way instruction cache (IC) and a 32KB 4-way data cache (DC), respectively. To avoid an explosion in the number of simulations, the block size for both caches is fixed to be 32 bytes.

5.2 Results and Discussions

A cache consists of tag and data arrays, which are used to store the tag and actual data, respectively. The tag is the high order bits of the address for determining whether the access is a hit or miss. Because the program size is usually an insignificant fraction of the entire address space, most tag bits are '0'.

Table. 3. Write pattern distribution of both tag and data arrays for IC and DC

		0->0	1->0	0->1	1->1
IC (32K 2-way)	tag	75.57%	6.19%	8.01%	10.22%
	data	77.36%	7.72%	9.81%	5.10%
DC (32K 4-way)	tag	58.87%	9.53%	9.56%	22.04%
	data	71.52%	21.26%	3.45%	3.77%

Write Pattern Distribution: Table 3 shows the write pattern distribution of both tag and data arrays for IC and DC. Because the difference between integer and floating-point programs is hardly noticeable, we do not present these two benchmarks separately. From this table, except for DC tag, the percentage of the 0->0 write pattern is over 70% for all other cases. This write characteristic is particularly beneficial to our SWDR cell, which hardly consumes power in the 0->0 write pattern.

Average Column Write Power (ACWP): We define the *average column write power (ACWP)* as the power dissipated in one column during each write. Because there are four write patterns, by definition, the ACWP is given by:

$$ACWP=(CP_{0->0}\times R_{0->0})+(CP_{1->0}\times R_{1->0}) +(CP_{0->1}\times R_{0->1})+(CP_{1->1}\times R_{1->1}) \quad (1)$$

$CP_{0->0}$ is the power dissipated in one column for the 0->0 write pattern, and $R_{0->0}$ is the ratio of the 0->0 write pattern to all write operations. Depending on cache configuration, the power dissipated in one column for various write patterns are listed in Table 4. We assume the tag array is implemented with the same SRAM cell as the data array. Applying the data shown in Tables 3 and 4 to Equation (1), the ACWP for each configuration are obtained and listed in Table 5. The results show that the ACWP of the conventional cell is almost equal to any column write power. This is because the column write power of the conventional cell is independent of write pattern. In contrast, by minimizing the power dissipated in writing '0' (including 0->0 and 1->0), the SWDR cell can reduce the ACWP by about 80% for IC tag and 83% for IC data. For DC data, the SWDR even reduces the ACWP by 92.71%.

Table. 4. The power dissipated in one column for various write patterns.

Column Power (mW)		$CP_{0->0}$	$CP_{1->0}$	$CP_{0->1}$	$CP_{1->1}$
Conv.	IC	4.37E-01	4.65E-01	4.92E-01	4.35E-01
	DC	2.62E-01	2.65E-01	2.56E-01	2.61E-01
SWDR	IC	5.17E-03	1.30E-02	4.64E-01	4.20E-01
	DC	2.58E-03	6.48E-03	2.32E-01	2.10E-01

Table. 5. The impact of SWDR cell on the ACWP for both IC and DC.

ACWP (mW)		Conv.	SWDR	***Reducction***
IC (32K 2-way)	tag	4.43E-01	8.48E-02	*80.86%*
	data	4.44E-01	7.19E-02	*83.81%*
DC (32K 4-way)	tag	2.62E-01	7.06E-02	*73.02%*
	data	2.62E-01	1.91E-02	*92.71%*

6. CONCLUSIONS

Most low power SRAM techniques only reduce read power, but generally write power is larger than read power. In this paper we concentrate on reducing the cache write power. Based on over 85% and 90% of the values written to the instruction cache and data cache are '0', we propose a novel SWDR cell to minimize the cache power dissipated in writing '0'. While exploiting the prevalence of '0' to reduce the average write power, the SWDR cell can retain the same stability and performance as the conventional cell with a cell area increase of 5.8%. The experimental results show that the SWDR cell can reduce the average cache write power dissipated in bitlines up to 92%.

7. REFERENCES

[1] M. Ukita et al., "A Single-Bit-Line Cross-Point Cell Activation (SCPA) Architecture for Ultra-Low-Power SRAM's," IEEE Journal of Solid-State Circuits, Vol. 28, No. 11, Nov. 1993, pp. 1114-1118.

[2] B. Amrutur and M. Horowitz, "Techniques to Reduce Power in Fast Wide Memories," in Proc. of Symposium on Low Power Electronics, Oct. 1994, pp. 92-93.

[3] K. W. Mai et al., "Low-Power SRAM Design Using Half-Swing Pulse-Mode Techniques," IEEE Journal of Solid-State Circuits, Vol. 33, No. 11, Nov. 1998, pp. 1659-1671.

[4] L. Villa, M. Zhang and K. Asanovic, "Dynamic Zero Compression for Cache Energy Reduction," in Proc. of 33rd International Symposium on Microarchitecture Micro-33, 2000, pp. 214-220.

[5] E. Seevinck, F. J. List and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," IEEE Journal of Solid-State Circuits, Vol. SC-22, No. 5, Oct. 1987, pp. 748-754.