A 0.8-V 128-kb Four-Way Set-Associative Two-Level CMOS Cache Memory Using Two-Stage Wordline/Bitline-Oriented Tag-Compare (WLOTC/BLOTC) Scheme

Perng-Fei Lin, *Member, IEEE*, and James B. Kuo, *Fellow, IEEE*

***Abstract*—This paper reports a 0.8-V 128-kb four-way set-associative two-level CMOS cache memory using a novel two-stage wordline/bitline-oriented tag-compare (WLOTC/BLOTC) and sense wordline/bitline (SWL/SBL) tag-sense amplifiers with an eight-transistor (8-T) tag cell in Level 2 (L2) and a 10-T shrunk logic swing (SLS) memory cell with the ground/floating (G/F) data sense amplifier in Level 1 (L1) for high-speed operation for low-voltage low-power VLSI system applications. Owing to the reduced loading at the SWL in the new 11-T tag cell using the WLOTC scheme, the 10-T SLS memory cell with G/F sense amplifier in L1, and the split comparison of the index signal in the 8-T tag cells with SWL/SBL tag sense amplifiers in L2, this 0.8-V cache memory implemented in a 1.8-V 0.18-μm CMOS technology has a measured L1/L2 hit time of 11.6/20.5 ns at the average dissipation of 0.77 mW at 50 MHz.**

I. Introduction

LOW-POWER low-voltage cache memory has become an indispensable component in VLSI systems for computer and communication-related applications [1], [2]. Cache memory chips are usually implemented using the bitline-oriented tag-compare (BLOTC) structure [3], [4], where for each bitline a sense amplifier (sense amp) is required. However, the signal from the output of the sense amp must be compared with the index to produce the hit/miss signal, and the large parasitic capacitances of the bitlines and the two-step hit/miss signal generation slows down the speed performance, especially at a low power-supply voltage. Recently, a 1-V four-way set-associative CMOS cache memory using the wordline-oriented tag-compare (WLOTC) structure been reported [5], [6]. Using the WLOTC structure with the content-addressable memory (CAM) 10-transistor (10-T) tag cell (TC) and the one-step hit/miss signal generation, a high-speed hit access at a low power consumption has been obtained. For the cache memory using a lower power-supply voltage of 0.8 V, the architecture adopted in the 1-V cache memory [5], [6] may not be sufficient due to reduced gate-overdrive voltage. In this paper, overcoming the drawback from the low gate-overdrive voltage, we describe a 0.8-V 128-kb four-way set-associative two-level CMOS cache memory using a novel two-stage wordline/bitline-oriented tag-compare (WLOTC/BLOTC) and sense wordline/bitline (SWL/SBL) tag sense amps with an 8-T TC in level 2 and a 10-T shrunk logic swing (SLS) memory cell with the ground/floating (G/F) data sense amp in level 1 for high-speed operation for low-voltage low-power VLSI system application. It will be shown that owing to the reduced loading at the SWL in the new 11-T TC using the WLOTC scheme, the 10-T SLS memory cell with G/F sense amp in L1, and the split comparison of the index signal in the 8-T TCs with SWL/SBL tag sense amps in L2, this 0.8-V cache memory implemented in a 1.8-V 0.18-μm CMOS technology has an L1/L2 hit time of 11.6/20.5 ns at an average dissipation of 0.77 mW at 50 MHz. In Section II, the architecture of this two-level cache memory is described, followed by the timing chart in Section III, the measured results in Section IV, and the conclusion in Section V.

Manuscript received January 8, 2002; revised June 21, 2002. A brief summary of this paper was presented at the European Solid State Circuits Conference, Villach, Austria, Sept. 2001.

P.-F. Lin is with Goyatek Technology Inc., Hsinchu 300, Taiwan, R.O.C.

J. B. Kuo is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: jbkuo@uwaterloo.ca).

Publisher Item Identifier 10.1109/JSSC.2002.803023.

II. Two-Level Cache Memory Architecture

Fig. 1 shows the block diagram of the 0.8-V 128-kb four-way set-associative two-level CMOS cache memory. As shown in the figure, in order to implement the two-stage WLOTC/BLOTC tag-compare scheme in L2 with an 8-T TC and a 10-T SLS memory cell with the G/F data sense amp structure, this cache memory is designed to have the memory portions, the tag portions, the tag and the data sense amps, the predecoder, and the multiplexers in two levels with 22-bit index data, 8-bit address, 8-bit write data, and 8-bit data out. The two-level hierarchical approach has been adopted in this cache memory. Level 1 is associated with the five bits of the 8-bit input address and Level 2 is referred to all eight bits of the input address. Therefore, the size of the tag-cell array and the memory-cell array in L1 is smaller. In each level, there are four memory portions and four tag portions. In L1, each tag portion has 128×22 TCs and 128×32 memory cells. In L2, each tag portion contains 22×256 TCs and 128×256 memory cells. By adopting the two-level BLOTC/WLOTC scheme, this proposed cache memory has a reduced power consumption as described in Sections III and IV. In Sections II-A and B, the tag portion and the memory portion are described in detail.

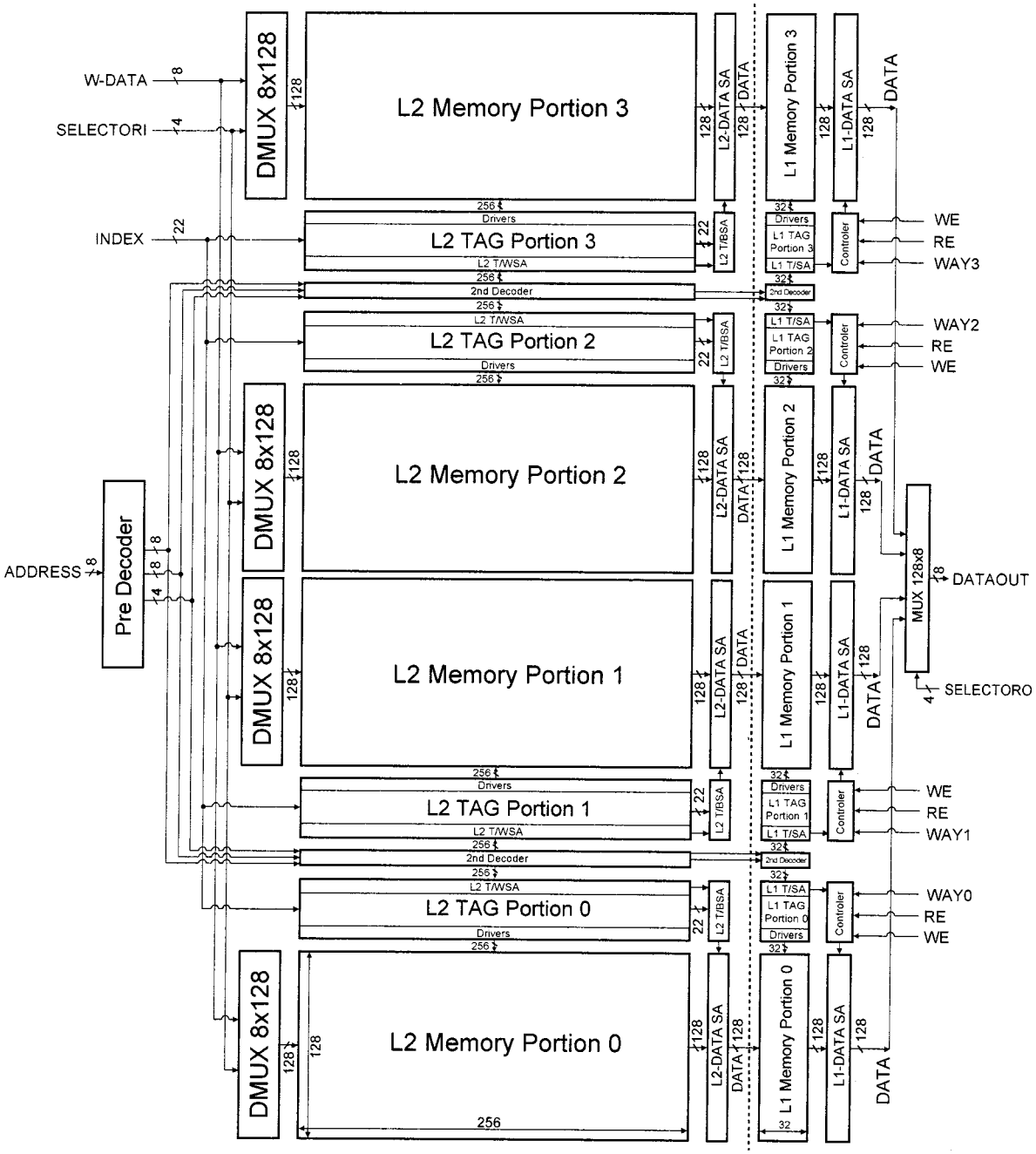


Fig. 1. Block diagram of the 0.8-V 128-kb four-way set-associative two-level CMOS cache memory.

A. Tag Portion

Fig. 2 shows the schematic of each of the four tag portions in this 0.8-V 128-kb four-way set-associative two-level CMOS cache memory. As shown in the figure, in the tag portion, there are two groups, Level 1 and Level 2. In Level 1, in each tag portion 22×32 11-T TCs with the WLOTC structure have been used. In Level 2, both WLOTC and BLOTC structures have been adopted to organize each tag portion of 22×256 8-T TCs. For each column of 22 TCs, all the tag sense wordlines (TSWLs) are vertically connected to an SWL tag sense amp at top. For each row of 256 TCs, all the SBLs are horizontally connected to an SBL tag sense amp at right. In Level 1, the WLOTC tag-compare scheme as described in [5] and [6] have been adopted. In each L1 TC column, there is a tag write wordline L1-TWWL, a tag read wordline $\overline{\text{L1-TRWL}}$, and a tag sense wordline L1-TSWL connected vertically to the WLOTC tag sense amp. In L1, there are 32 WLOTC tag sense amps in total. As for Level 2, in order to reduce power consumption, increase speed performance, and save layout area, the two-stage WLOTC/BLOTC tag-compare scheme based on the index data has been used. In each L2 TC column, a tag read wordline $\overline{\text{L2-TRWL}}$, a tag write wordline L2-TWWL, and a tag sense wordline L2-TSWL are connected vertically

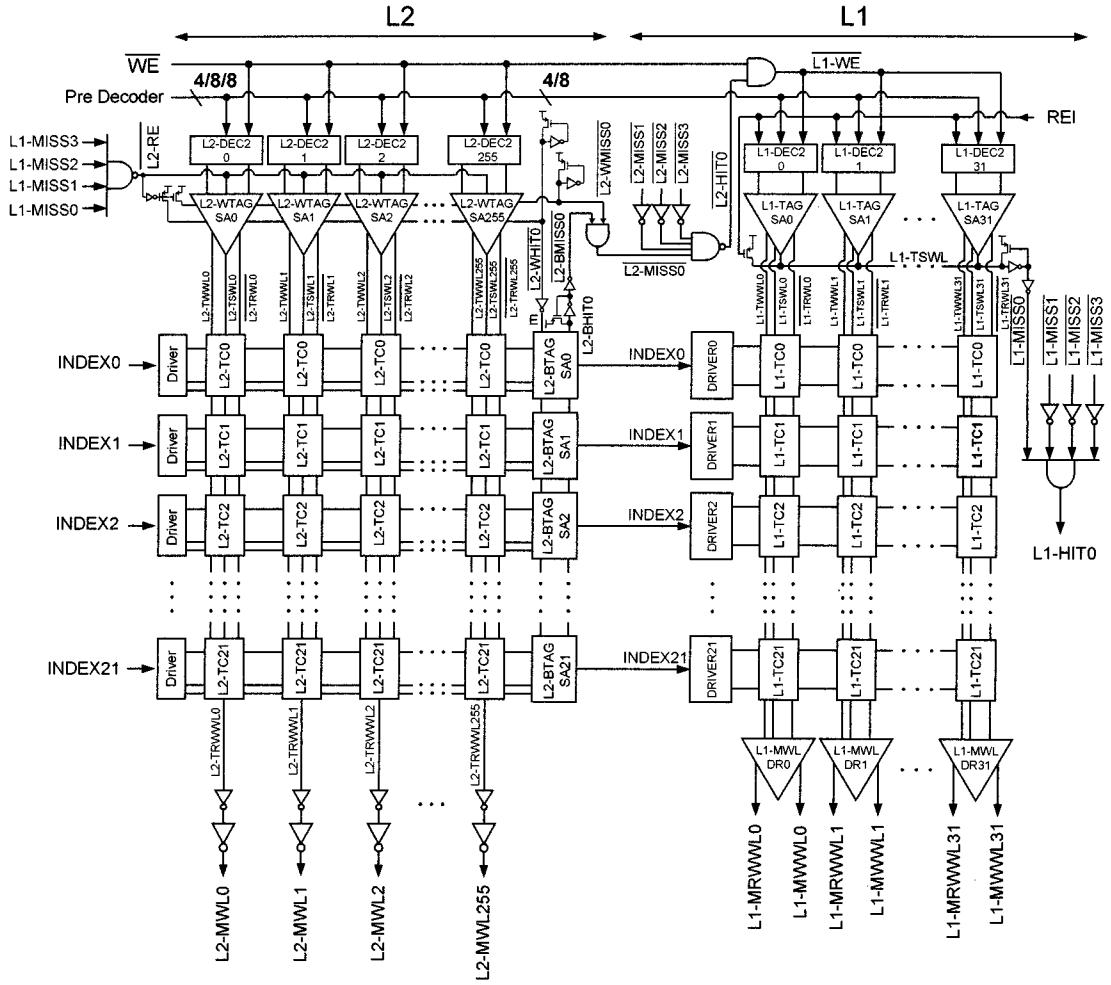


Fig. 2. Schematic of the tag portion in this two-level CMOS cache memory.

to a WLOTC tag sense amp. In L2, there are 256 WLOTC tag sense amps in total. During the TC access procedure, in a specific TC column selected by the 8-bit input data address, the TCs with their respective index data bits of logic-0 are connected to the WLOTC tag sense amp at top via SWL during the WLOTC sending period. After the sensing period of the WLOTC tag sense amp, if there is a TSWL-related hit signal for this column, the follow-up BLOTC tag sensing will be initiated. The remaining TCs in that specific column with their respective index data bits of logic-1 are connected to the SBL tag sense amps via the horizontal SBL for further sensing. If a further SBL-related hit signal can be obtained, a hit signal is generated to trigger the readout of the data from the specified memory cell.

In order to facilitate the two-level two-stage WLOTC/BLOTC tag-compare scheme in L2 and WLOTC tag-compare scheme in L1, innovative designs of L1/L2 TCs are used in the tag portion. These L1/L2 tag-portion-related circuits, including the second-level decoder and the tag sense amp, are described in the following.

1) L1/L2 Tag Cells: In order to implement the WLOTC scheme, Fig. 3(a) shows an innovative 11-T TC circuit used in L1, which is derived from a 10-T TC circuit as shown in Fig. 3(b). In this 11-T TC, the SRAM cell portion is similar to that of the 10-T one. Unlike in the 10-T one, in this 10-T TC, only one pair of bitlines are used in both the SRAM cell and the tag-compare portions. (Note that in the 10-T one, both write and read bitlines (W-BL/R-BL) are adopted.) Similar to the 10-T one, for the SRAM cell portion, a write wordline (W-WL) controls a pair of bitlines connected to the internal storage nodes. In the tag-compare portion, instead of four nMOS devices connected between the sense wordline as in the 10-T one, three nMOS and two pMOS devices (MP2–4, MN4–5) have been used to determine the logic state of the L1 TSWL. When the tag read wordline $\overline{\text{L1-TRWL}}$ is high, MN4 is on and the MN5 is off. The tag sense wordline L1-TSWL is disconnected from this TC and maintains its precharged high state. When the tag read wordline $\overline{\text{L1-TRWL}}$ is low, MP4 is on. Under this situation, whether or not the tag sense wordline L1-TSWL is pulled low depends on the internal storage data (BIT) and the bitline. If the internal storage data is different from the bitline (say, $\text{BIT} = 0, \text{BL} = 1$), then either MP2 or MP3 is on (MP2 is on). Therefore, PS4 is high and the L1-TSWL is pulled low, indicating a miss. If the BIT is the same as the bitline, then both MP2 and MP3 are off and MN5 is off. Thus, the L1-TSWL stays precharged high, indicating a hit. The adoption of three pMOS devices and two nMOS devices in the tag-compare portion of the 11-T TC instead of four nMOS

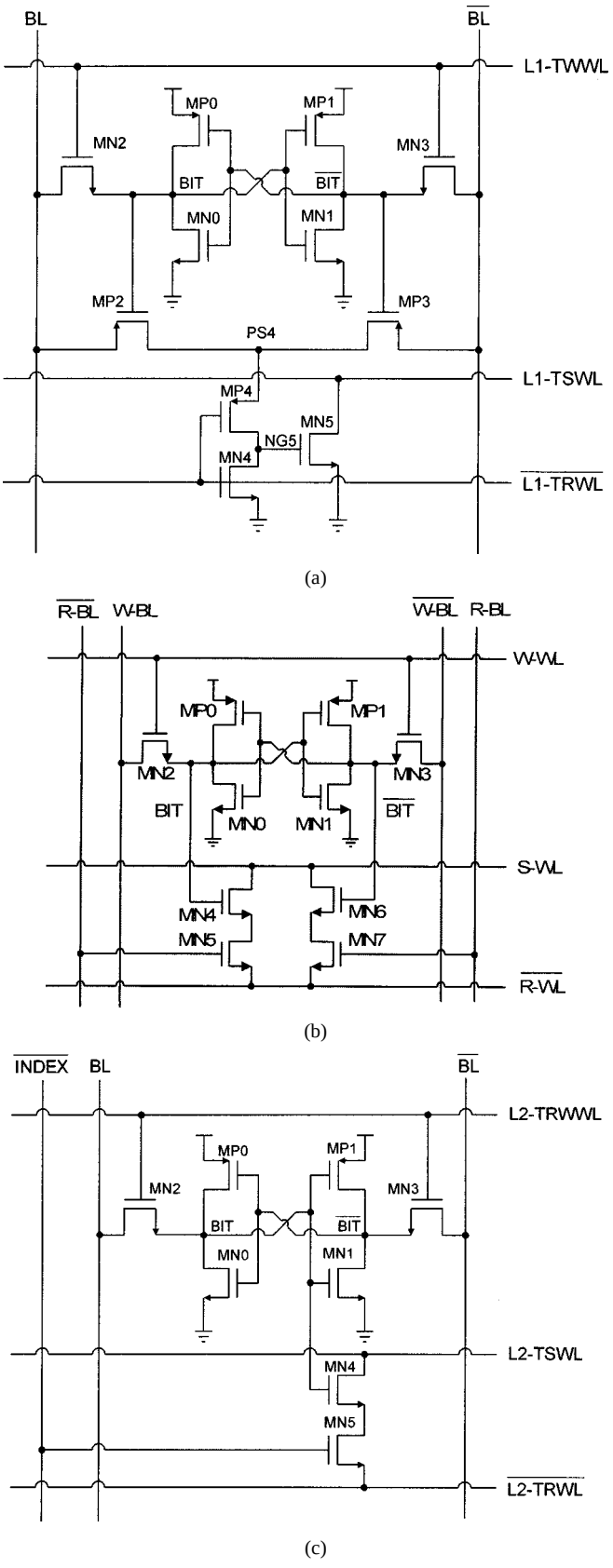


Fig. 3. (a) L1 11-T TC circuit. (b) 10-T TC circuit. (c) L2 8-T TC circuit.

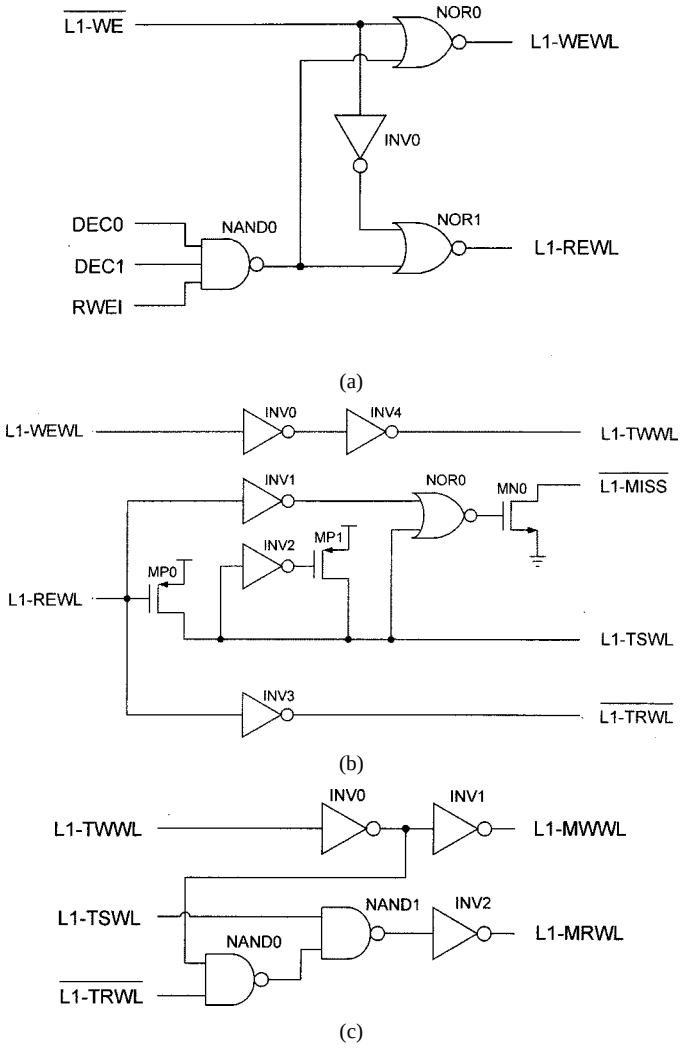


Fig. 4. Other L1 tag portion related circuits. (a) Second-level decoder. (b) WLOTC tag sense amp. (c) Memory driver.

devices as in the 10-T one, provides advantages. In the entire 11-T TC, there are five pMOS and six nMOS devices, which provide a more balanced count between nMOS and pMOS such that the layout density can be higher. In addition, at the tag sense wordline, only one nMOS device (MN5) is connected, instead of two nMOS devices as in the 10-T case shown in Fig. 3(b). Therefore, smaller parasitic capacitance connected to the tag sense wordline (L1-TSWL) and, hence, a higher speed can be obtained. In addition, since there are no more stacked nMOS devices between the L1-TSWL and ground as in the 10-T case, a smaller on resistance allows a higher speed during the pulldown process.

Fig. 3(c) shows the 8-T TC circuit used in L2. As shown in the figure, this 8-T TC circuit is derived from the 10-T TC as shown in Fig. 3(b), except that in the tag-compare portion only two nMOS devices instead of four connect the tag sense wordline (L2-TSWL) to the the tag read wordline ($\overline{\text{L2-TRWL}}$). The two nMOS devices (MN4 and MN5) between the L2-TSWL and the $\overline{\text{L2-TRWL}}$ are controlled by the internal storage node BIT and the index signal $\overline{\text{INDEX}}$. The other two nMOS devices controlled by the other related internal storage node $\overline{\text{BIT}}$ and the other related index signal INDEX have been removed, because in the L2 WLOTC tag-compare operation, only the TCs with their corresponding index data bit of logic-1 status are connected to the WLOTC tag sense

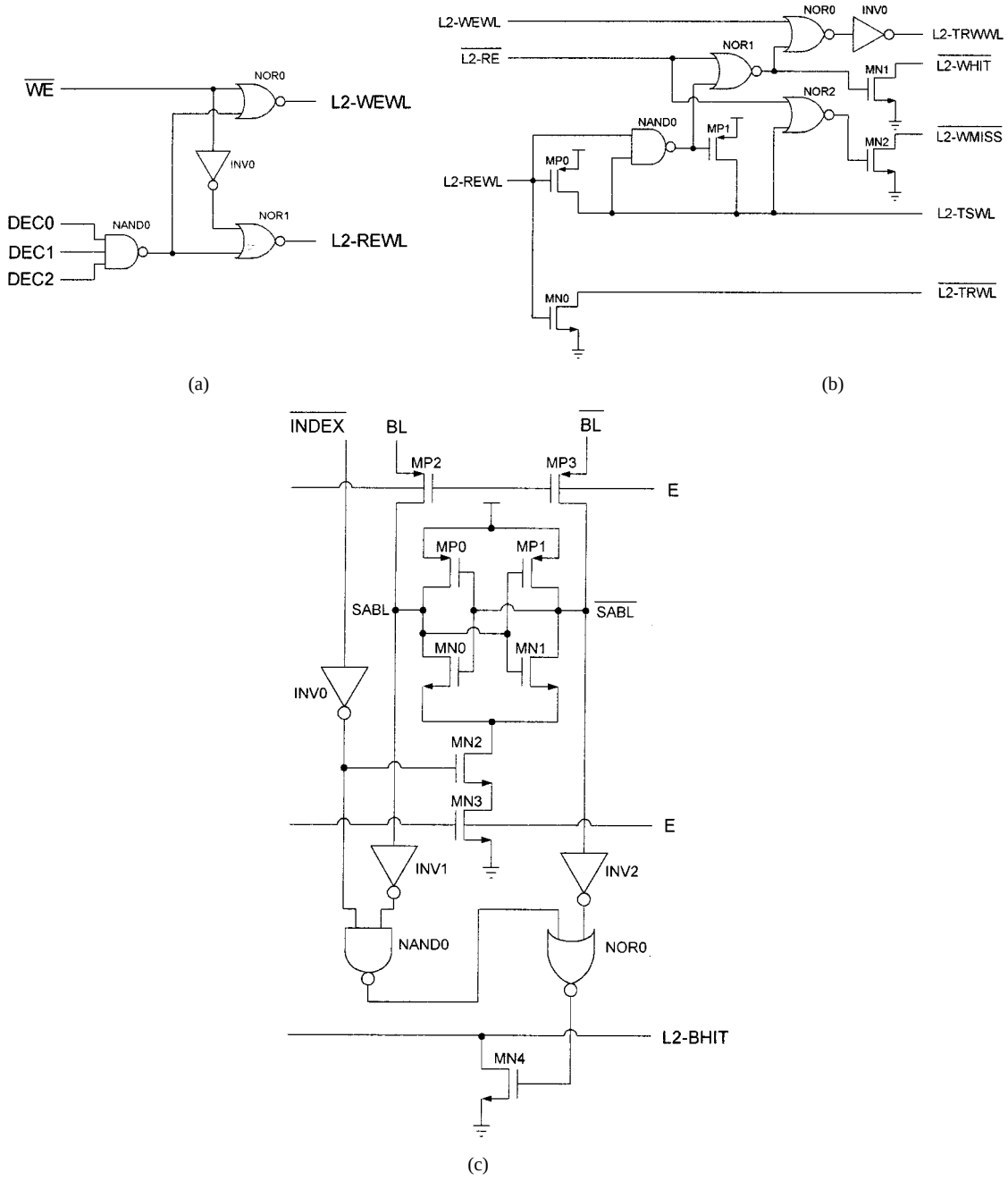


Fig. 5. Other L2 tag portion related circuits. (a) Second-level decoder. (b) WLOTC tag sense amp. (c) BLOTC tag sense amp.

amp of the column. Consequently, the layout of the innovative L2 TC is smaller as compared with the 10-T TC. Therefore, the parasitic capacitance associated with each L2 tag sense wordline is reduced.

2) L1 Tag-Portion-Related Circuits: Fig. 4 shows the other L1 tag-portion-related circuits. In each column of 22 TCs, there is a second-level decoder at the top as shown in Fig. 4(a). The second-level decoder used in each TC column is used to generate the write-enable wordline (L1-WEWL) and read-enable wordline (L1-REWL) based on the outputs from the predecoder as shown in Fig. 1, where eight bits of read address (ADDRESS) are divided into three groups of 3, 3, and 2 bits to produce three predecoder outputs (DEC) of 8, 8, and 4 bits, respectively. Each of the two groups of 8- and 4-bit predecoder outputs donates a bit of its output to form a set of 2-bit inputs (DEC0, DEC1) to the L1 second-level decoder as shown in Fig. 4(a) for producing 32 L1-REWLs and the L1-WEWLs for 32 L1 TC columns.

In addition to the second decoder circuit, the other important L1 tag-portion-related circuit is the WLOTC tag sense amp as shown in Fig. 4(b), where it is used to provide the miss signal $\overline{\text{L1-MISS}}$ and the $\overline{\text{L1-TRWL}}$ based on the L1-REWL and the L1-TSWL and the three predecoder outputs (DEC0–2). Each of the three groups of 8-, 8-, and 4-bit predecoder outputs donates a bit of its output to form a set of three-bit inputs (DEC0–2) to the L2 second-level decoder, as shown in the figure, for producing 256 L2-REWLs and the L2-WEWLs for 256 L2 TC columns. In order to simplify the design, a straightforward random-logic approach was used to implement the L1 tag sense

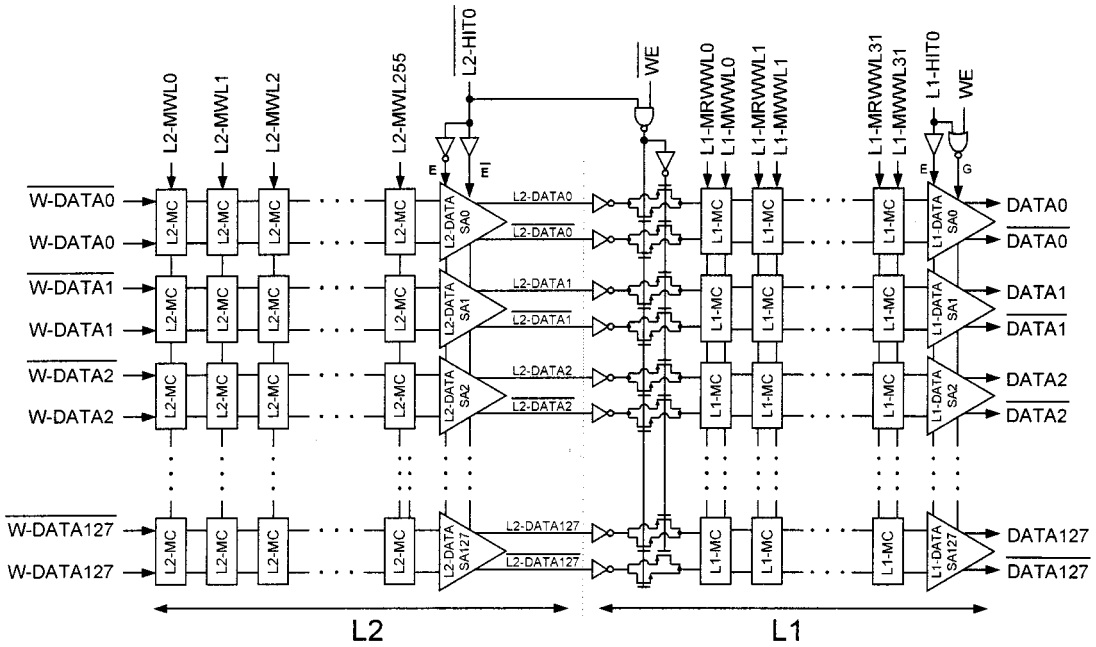


Fig. 6. Schematic of the memory portion in this two-level CMOS cache memory.

amp. In each of the L1 TC columns, there is also a memory driver as shown in Fig. 4(c). Each memory-driver circuit is used to provide a memory-write wordline (L1-MWWL) and a memory-read wordline (L1-MRWL) for their corresponding memory cell based on the L1-TWWL, the L1-TSWL, and the $\overline{\text{L1-TRWL}}$ if a hit signal has been generated or a write access is on.

3) L2 Tag-Portion-Related Circuits: Fig. 5 shows other L2 tag-portion-related circuits including a second-level decoder, a WLOTC tag sense amp, and a BLOTC tag sense amp. As shown in Fig. 5(a), the L2 second-level decoder, which is similar the L1 one described above, is used to generate the L2-WEWL and the L2-REWL based on the predecoder outputs (DEC0–2) and the write-enable signal $\overline{\text{WE}}$. As shown in Fig. 5(b) and (c), there are two kinds of tag sense amps in each column of L2 TC, an L2 WLOTC tag sense amp and an L2 BLOTC tag sense amp, to facilitate the two-stage WLOTC/BLOTC tag sensing procedure. As shown in Fig. 5(b), the L2 WLOTC tag sense amp, which is similar to the L1 WLOTC tag sense amp described previously, is used to generate the wordline hit signal $\overline{\text{L2-WHIT}}$, the wordline miss signal $\overline{\text{L2-WMISS}}$, and the tag read write wordline signal (L2-TRWWL) based on the L2-WEWL, the L2-REWL, the read-enable wordline $\overline{\text{L2-RE}}$, and the L2-TSWL.

As shown in Fig. 5(c), in the BLOTC tag sense amp, the latch-up sense amp (MN0–1, MP0–1) is the center core. When the index signal $\overline{\text{INDEX}}$ is low and the vertical hit signal has been generated, MN2 and MN3 turn on to switch on the latch-type sense amp. For the TCs associated with the index bits $\overline{\text{INDEX}}$ of logic-0 in a specific TC column, after the vertical hit signal has been generated by the WLOTC tag sense amp, their associated bitlines are individually connected to their own BLOTC tag sense amps. Note that for each TC column accessed, only one WLOTC tag sense amp is on and more than one BLOTC tag sense amps may turn on depending on the number of bits of the associated index signal $\overline{\text{INDEX}}$ with a logic value of zero.

B. Memory Portion

In this two-level CMOS cache memory, in addition to the four tag portions, there are also four memory portions for four-way outputs. Fig. 6 shows the schematic of a memory portion in this two-level CMOS cache memory. As shown in the figure, each memory portion is divided into two levels, L1 and L2. In the L1 memory portion, there are 32 columns of memory cells (MCs). Each column contains 128 MCs. Each column is controlled by a memory read/write wordline (L1-MRWWL) and a memory write wordline (L1-MWWL), which are high if an L1 hit has been obtained and if this column has been specified by the address. In each row of 32 MCs, there are a pair of bitlines connected to a data sense amp at right. In the data sense amp column, there are a total of 128 data sense amps to provide 128 output data, of which eight output data are selected via a multiplexer to become the final data-out signals.

Fig. 7 shows the 10-T SLS memory cell in L1 and the memory cell in L2. As shown in Fig. 7(a), the L1 10-T SLS memory cell is targeted for G/F logic operation. Usually, the logic operation of a digital circuit is based on the two-logic value ground/V_{dd}. In order to reduce the switching time, the SLS memory cell is designed for operation with logic states of ground and floating. As shown in the figure, the pass transistor MN2 controlled by the L1-MWWL is on only during the write access. During the read access, MN2 is off and the data in the memory cell is read out via MN4 and MN6. Unlike the conventional approach, one of the two logic states, ground or floating, is available to be detected by the data sense amp. During the read access, if the internal storage node $\overline{\text{BIT}}$ has a logic-1, then MN6 is on. Therefore, during the readout access, the L1-MRWWL is high, and the bitline is raised since MN4 is on. If the $\overline{\text{BIT}}$ is low, then MN6 is off. Since MN2 is also off during the readout operation, the bitline is connected to a floating node. In contrast, in the L2 memory cell as shown in Fig. 7(b), which is based on a standard 6-T SRAM cell, both bitlines are connected to the

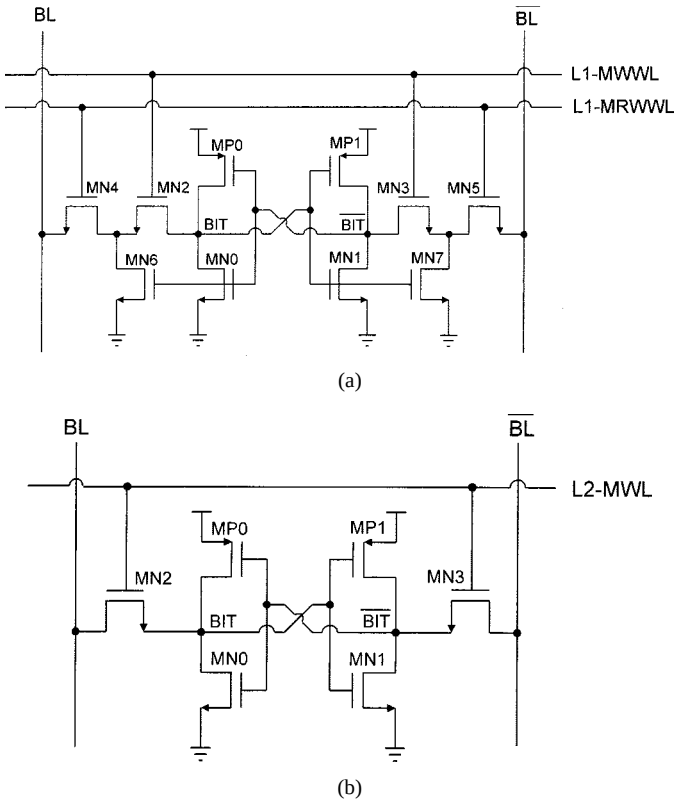


Fig. 7. (a) Shrunk logic swing MC in L1. (b) MC in L2.

internal storage nodes controlled by the L2-MWL for read-out and write-in accesses.

Fig. 8 shows the data sense amps in L1 and L2. As shown in Fig. 8(a), in the L1 data sense amp, a latch-type sense amp based on the G/F scheme has been adopted. When the signal G is low, which indicates an L1 hit for the data sensing period, MN4–6 are off. In addition, the bitlines are connected to two sides of the latch-type sense amp (SABL) via MN2/MN3 for amplifying the signals on the bitlines such that they can become the output data signals. When the signal G is high, which indicates the standby period, MN4–6 are on and MP2–3 are on since the signal E is low $(E = \overline{G})$ and both sides of the BL are grounded such that a steady state can be reached for the next sensing period. As shown in Fig. 8(b), a latch-type sense amp (MP0–1, MN0–1) has also been used in the L2 data sense amp. When the signal E is high, which indicates an L2 hit, MN2 is on, which turns on the latch-type sense amp. Under this situation, the bitlines are connected to both sides of the SABL for amplification of signals such that they can become the output data signals. The adoption of the SLS memory cell with the G/F data sense amp in L1 provides advantages in reduced power consumption and enhanced speed. For the conventional latch sense amp, during the sense operation, the connected bitlines need to be precharged high first. When the wordline turns on, one bitline discharges. The latch sense amp functions only when there is a substantial amount of voltage difference occurring between the bitlines. The power consumption of the latch sense amp during the sense operation is from precharge and discharge of the bitlines. In contrast, for the G/F sense amp, during the sense operation, when the wordline turns on, the bitline originally grounded

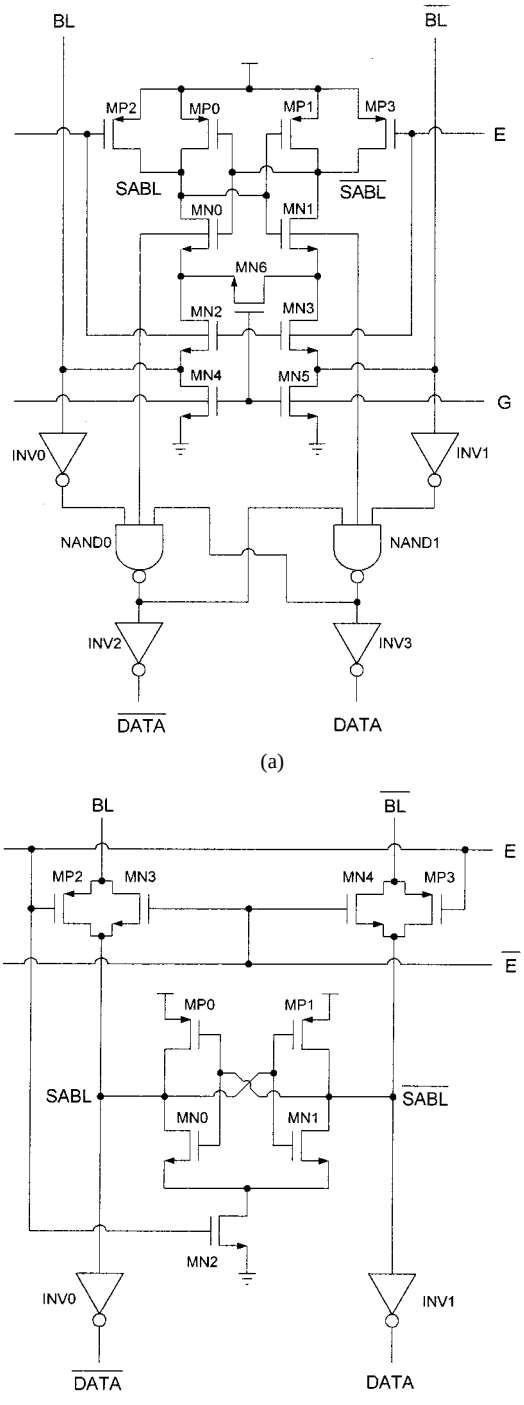


Fig. 8. (a) Data sense amp using the G/F scheme in L1. (b) Data sense amp in L2.

is separated from ground. Therefore, one bitline is grounded and the other one is floating, which is sensed by the G/F sense amp as high. Since two bitlines are not required to precharge or discharge during the sense period, the power consumption of the G/F sense amp has a reduced power consumption. For the latch sense amp, it does not function until a substantial amount of voltage difference existing between the bitlines. In contrast, for the G/F sense amp, it can function immediately without waiting for the voltage difference to develop. Thus, the G/F sense amp has an enhanced speed performance.

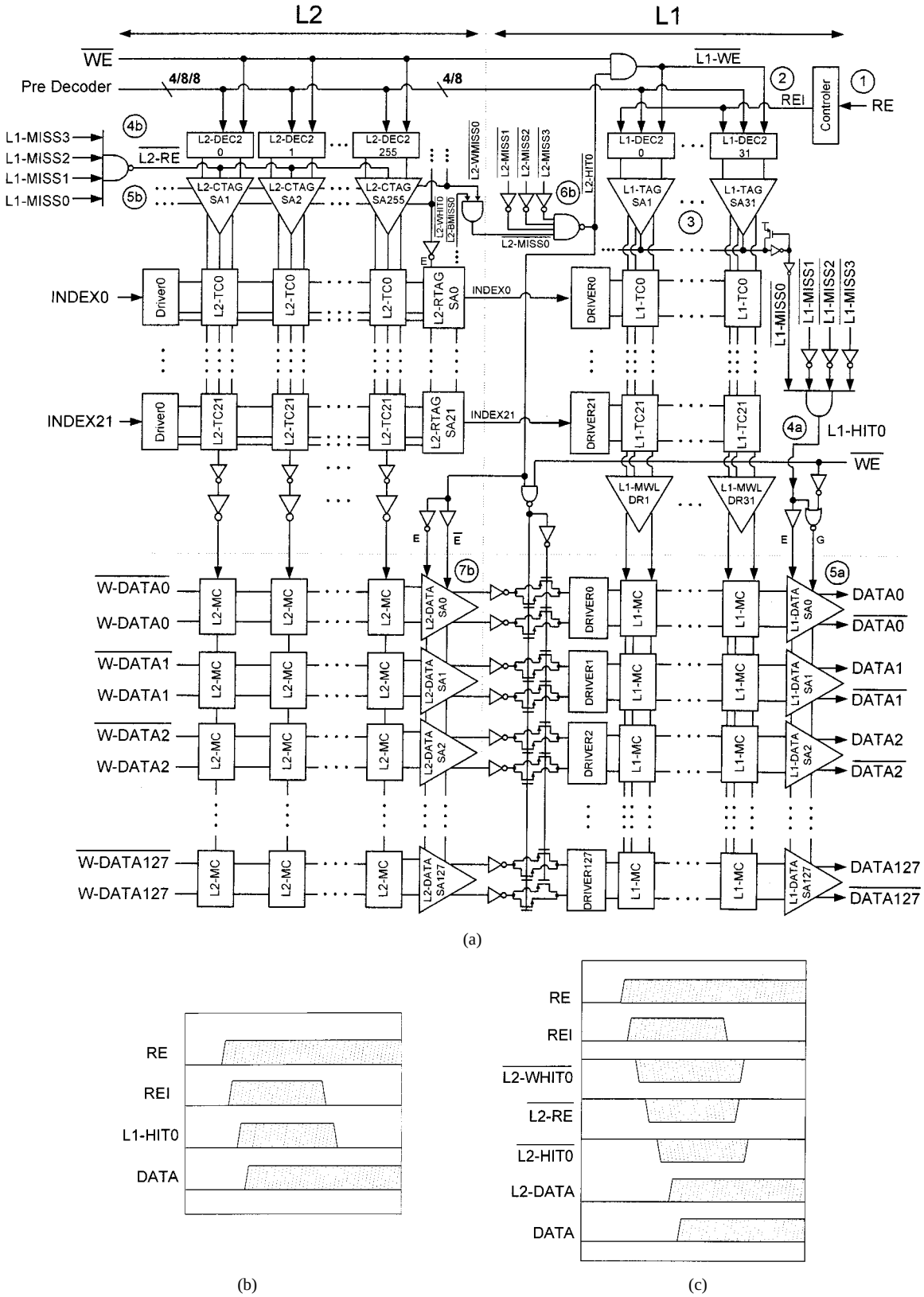


Fig. 9. (a) Critical path from read enable (RE) to data out (DATA) during the L1/L2 hit access of this two-level CMOS cache memory. (b), (c) Timing charts for the read-out operation of (b) L1 and (c) L2 hits.

III. Timing Chart

Fig. 9 shows the critical path from read enable (RE) to data out (DATA) during the L1/L2 hit access of the new two-level CMOS cache memory, and the timing chart for the read-out operation of L1 and L2 hits for this cache memory. As shown in the figure, for the L1 hit case the readout operation is initiated by the RE signal, which is provided externally. Then an internal

TABLE I
DEVICE PARAMETERS OF THE 0.180-μm CMOS TECHNOLOGY

Device Parameters	
Technology	0.18um, 1.8V, 1P6M
t_{ox}	4.08nm
V_{th} (N/P)	0.513V/0.566V
I_{dsat} (N/P)	600/260 uA/um
6T Memory cell	2.62um x 2.935um
10T Memory cell	2.62um x 4.7um
8T Tag cell	3.22um x 2.935um
11T Tag cell	3.22um x 4.7um
Active area	1.30mm x 2.07mm
Total area	1.90mm x 2.31mm

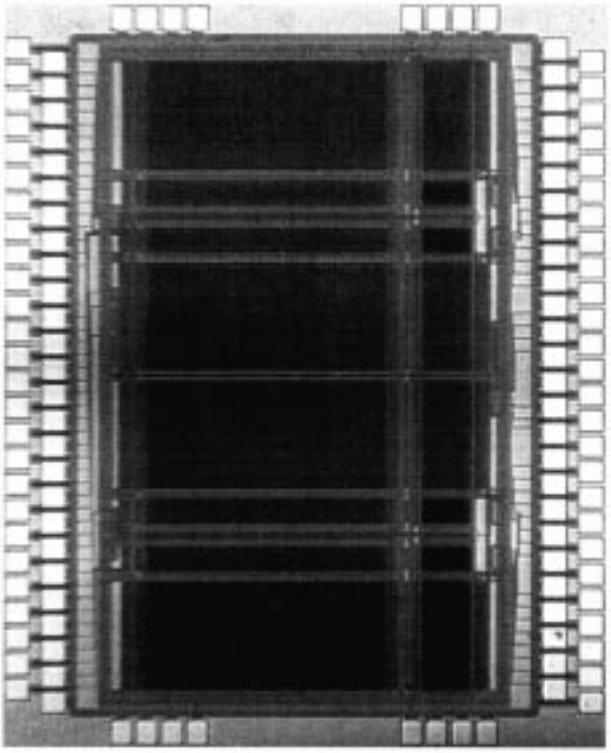


Fig. 10. Die photo of the 0.8-V 128-kb four-way set-associative two-level CMOS cache memory. The die area is 1.3 mm × 2.07 mm.

read enable signal (REI) with a fixed pulsewidth is generated by a controller. As controlled by the REI signal, the L1 WLOTC tag sense amp generates an L1-HIT signal, which triggers readout of the data stored in an associated memory cell. Compared to the L1 hit case, the readout operation of the L2 hit case seems more complicated. After the REI is generated, the WLOTC hit or miss signal (L2-WHIT) is first generated by the WLOTC tag sense amp. Then the read enable signal (L2-RE) is generated. Via the BLOTC tag sense amp, the final hit signal (L2-HIT) is produced, followed by the readout of the data in the specified L2 memory cells (L2-DATA). Before becoming available at the final output data (DATA), the L2 output data (L2-DATA) is written to the corresponding L1 memory cells. For the L2 hit operation, the tag-compare procedure is composed of two stages,

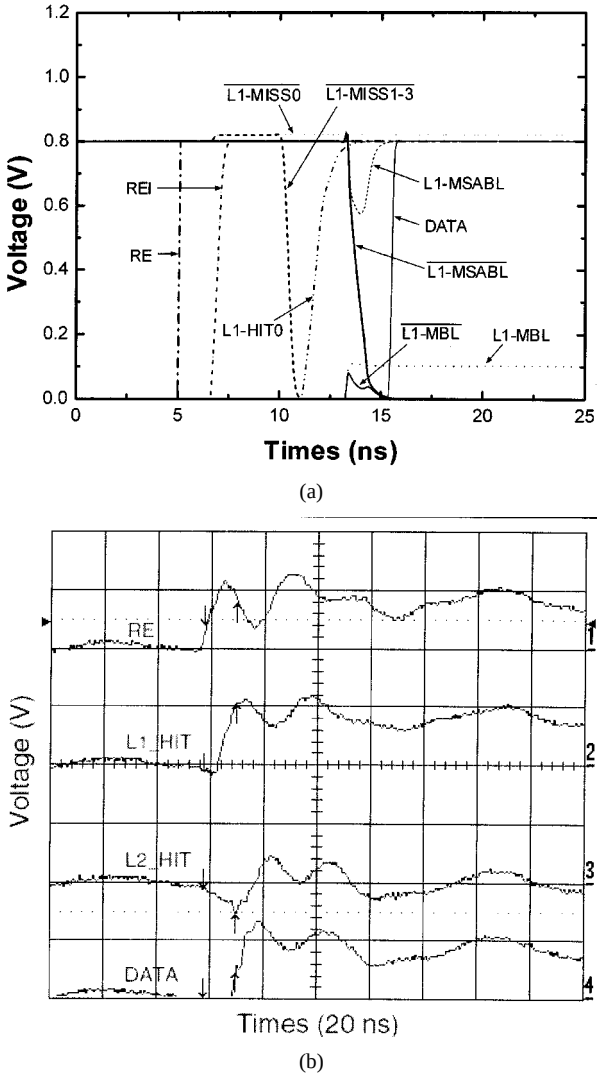


Fig. 11. (a) SPICE-simulated transient waveforms and (b) the measured results of the 0.8-V 128-kb four-way set-associative two-level CMOS cache memory during the L1 hit access.

the WLOTC tag-compare stage and the BLOTC tag-compare stage. This two-stage tag-compare procedure reduces parasitic capacitances associated with the SWLs and the SBLs to enhance speed performance. In addition, power consumption can be reduced. By adopting the 8-T TC in the L2 tag portion, a further reduction in the parasitic capacitance associated with the SWL and SBL also enhances the speed performance.

IV. MEASURED RESULTS

In order to assess the performance of this new two-level CMOS cache memory, a test chip has been designed and tested using a 1.8-V 0.18-μm CMOS technology with one polysilicon layer and six metal layers. As shown in Table I, for the 0.18-μm CMOS devices, the gate oxide is 4.08 nm and the threshold voltage for the nMOS/pMOS device is 0.513 V/0.566 V. The layout area of the 6-T/10-T memory cell is $2.62 \times 2.93\,\mu$m/$2.62 \times 4.7\,\mu$m. For the 8-T/11-T TC, it is $3.22 \times 2.935\,\mu$m/$3.22 \times 4.97\,\mu$m. Fig. 10 shows the die photo of this 0.8-V 128-kb four-way set-associative two-level CMOS cache memory. The active die area is 1.3 mm × 2.07 mm.

TABLE II
COMPARISON OF THE POWER CONSUMPTION DISTRIBUTION BETWEEN THIS TWO-LEVEL CMOS CACHE MEMORY USING WLOTC/BLOTC SCHEME AND THAT USING THE CONVENTIONAL BLOTC SCHEME DURING L1 AND L2 HIT ACCESS

Power Consumption Distribution during L1 hit

	Proposed Cache	Conventional BLOTC Cache
Decoder & Peripheral Logic	0.150mW(51.72%)	0.240mW(14.94%)
L1 Tag Portion	0.018mW(6.21%)	0.557mW(34.66%)
L1 Memory Portion	0.106mW(36.55%)	0.810mW(50.40%)
L2 Tag Portion WLOTC	0.016mW(5.52%)	NA
Total Power Consumption	0.290mW	1.607mW

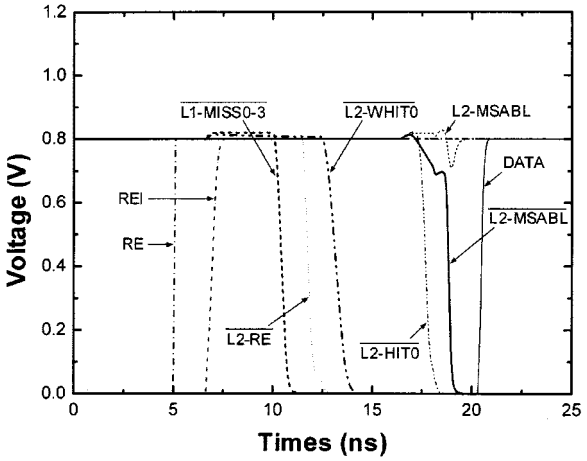
Power Consumption Distribution during L2 hit

	Proposed Cache	Conventional BLOTC Cache
Decoder & Peripheral Logic	0.897mW(17.70%)	0.666mW(9.52%)
L1 Tag Portion	0.021mW(0.41%)	0.560mW(8.00%)
L2 Tag Portion WLOTC	0.016mW(0.32%)	NA
L2 Tag Portion BLOTC	0.713mW(14.07%)	2.352mW(33.60%)
L2 Memory Portion	3.421mW(67.50%)	3.421mW(48.88%)
Total Power Consumption	5.068mW	6.999mW

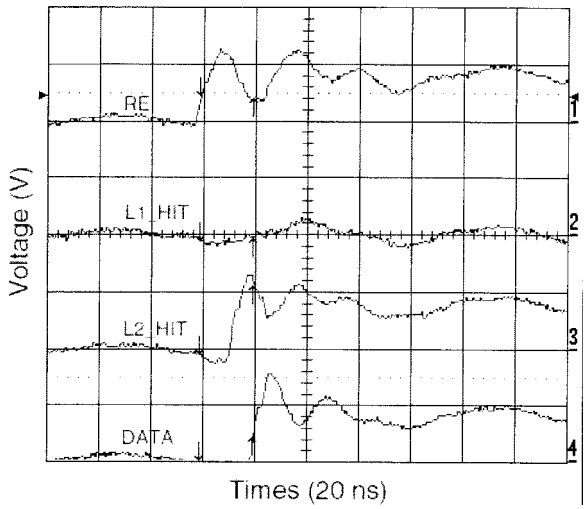
There are 208 staggered I/O pads surrounding a die area of 3.5 mm $\times$ 3.5 mm.

Fig. 11 shows SPICE simulated transient waveforms and the measured results of this two-level CMOS cache memory during the L1 hit access. As shown in the figure, from the transition of the RE signal, the REI signal is generated, followed by the L1-HIT signal, and the signals at the bitlines of the memory cell (L1-MBL). Finally, via the signals at the both sides of the data sense amp (L1-MSABL), the final output data signal is formed. The L1 hit access time is 6 ns at power consumption of 0.29 mW, which is 5.5 times smaller than the 16-kb L1/128-kb L2 cache memory using the BLOTC structure. The small power consumption is attributed to the distributed tag sense-amp in the WLOTC structure.

Fig. 12 shows SPICE simulated transient waveforms and the measured results of this two-level CMOS cache memory during the L2 hit access. As shown in the figure, the L2 hit detection time is 11/14 ns at power consumption of 5.07 mW, which is 28% smaller than the standard cache memory using the BLOTC structure. Based on the 10% L1 hit and 90% L2 hit statistics, the average power dissipation is 0.77 mW at 50 MHz. The small power consumption is attributed to the distributed tag sense-amp in the L1 WLOTC and L2 WLOTC/BLOTC structure. The high-speed performance is due to the two-level two-stage WLOTC/BLOTC tag compare scheme in L2 and WLOTC tag-compare scheme in L1 and innovative designs of the L1/L2 TCs, the second-level decoder, and the tag sense amp as the L1/L2 tag-portion-related circuits. Table II shows the comparison of the power consumption distribution between this two-level CMOS cache memory using WLOTC/BLOTC scheme and the one using the conventional BLOTC scheme during L1 hit access and L2 hit access. For the proposed CMOS cache memory during the L1 hit access, the total power consumption is 0.29 mW, which is 82% less than the conventional one using the BLOTC scheme, due to the WLOTC scheme in the L1. During the L2 hit access, the total power consumption is 5.07 mW, which is 27% smaller due to the WLOTC/BLOTC structure used for the proposed scheme.



(a)



(b)

Fig. 12. (a) SPICE-simulated transient waveforms and (b) the measured results of the 0.8-V 128-kb four-way set-associative two-level CMOS cache memory during the L2 hit access.

V. Conclusion

In this paper, a 0.8-V 128-kb four-way set-associative two-level CMOS cache memory using a novel two-stage WLOTC/BLOTC and SWL/SBL tag sense amps with an 8-T TC in level 2 and a 10-T SLS memory cell with a G/F data sense amp in Level 1 for high-speed operation for low-voltage low-power VLSI system applications has been obtained. Due to the reduced loading at the sense wordline in the new 11-T TC using the WLOTC scheme and the 10-T SLS memory cell with G/F sense amp in L1 and the split comparison of the index signal in the 8-T TCs with SWL/SBL tag sense amps in L2, this 0.8-V cache memory implemented in a 1.8-V 0.18-μm CMOS technology has a measured L1/L2 hit time of 11.6/20.5 ns at the average dissipation of 0.77 mW at 50 MHz.

Acknowledgment

The authors would like to thank Taiwan Semiconductor Manufacturing Company (TSMC) for implementing the test chip.

References

[1] J. B. Kuo and J. H. Lou, *Low-Voltage CMOS VLSI Circuits*. New York: Wiley, 1999.

[2] J. B. Kuo and S. C. Lin, *Low-Voltage SOI CMOS VLSI Devices and Circuits*. New York: Wiley, 2001.

[3] H. Mizuno, N. Matsuzaki, K. Osada, T. Shinbo, N. Ohki, H. Ishida, K. Ishibashi, and T. Kure, "A 1-V 100-MHz 10-mW cache using a separated bitline memory hierarchy architecture and domino tag comparators," *IEEE J. Solid-State Circuits*, vol. 31, pp. 1618–1624, Nov. 1996.

[4] K. Osada, H. Higuchi, K. Ishibashi, N. Hashimoto, and K. Shiozawa, "A 2-ns access 285-MHz two-port cache macro using double global bitline pairs," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, 1997, pp. 402–403.

[5] J. B. Kuo, P. F. Lin, F. Wang, H. H. Chang, W. T. Wang, and C. H. Chen, "A 1-V 3.44-ns 4.1-mW at 50-MHz 128-kb four-way set-associative CMOS cache memory implemented by 1.8-V 0.18-μm foundry CMOS technology for low-voltage low-power VLSI system applications," in *26th Eur. Solid-State Circuits Conf. (ESSCIRC) Dig. Tech. Papers*, Sept. 2000, pp. 308–311.

[6] P. F. Lin and J. B. Kuo, "A 1-V 128-kb four-way set-associative CMOS cache memory using worldline-oriented tag-compare (WLOTC) structure with the content-addressable-memory (CAM) 10-transistor tag cell," *IEEE J. Solid-State Circuits*, vol. 36, pp. 666–675, Apr. 2001.

[7] H. Kadota, J. Miyake, Y. Nishimichi, H. Kudoh, and K. Kagawa, "An 8-kb content-addressable and reentrant memory," *IEEE J. Solid-State Circuits*, vol. SC-20, pp. 951–957, Oct. 1985.

[8] A. K. Goksel, R. H. Krambeck, P. P. Thomas, M.-S. Tsay, C. Y. Chen, D. G. Clemons, F. D. LaRocca, and L.-P. Mai, "A content addressable memory management unit with on-chip data cache," *IEEE J. Solid-State Circuits*, vol. 24, pp. 592–596, Mar. 1989.

Perng-Fei Lin (M'01) was born in Yu-lin, Taiwan, R.O.C., on April 10, 1967. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1992, 1996, and 2001, respectively.

From 1996 to 2001, he was a Principal Engineer with the DSD Department, Taiwan Semiconductor Manufacturing Company, Hsin-chu, Taiwan. Since 2001, he has been with Goyatek Technology Inc., Hsin-chu, as a Technical Director of the Silicon Component Design Department. His research interest is in low-voltage CMOS digital circuits.

James B. Kuo (M'79–SM'92–F'00) received the B.S.E.E. degree from National Taiwan University, Taipei, Taiwan, in 1977, the M.S.E.E. degree from The Ohio State University, Columbus, in 1978, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, in 1985.

He was with Penril Data Communications and Racal Vadic during 1978–1981 as a Research Engineer working on integrating telecommunication modem chips using CMOS technology. He was an Engineering Research Associate with the Integrated Circuits Laboratory, Stanford University, during 1985–1987, working on BiCMOS devices. In 1987, he joined National Taiwan University as an Associate Professor, where he became a Professor in 1990. In 2000, he joined the University of Waterloo, Waterloo, ON, Canada, as a tenured full Professor, on leave from National Taiwan University. He has published 250 technical papers including 46 IEEE journal papers. He holds 16 invention patents including seven U.S. patents on low-voltage CMOS VLSI circuits, and is the author of nine books. He has graduated 48 M.S. and Ph.D. students specialized in CMOS circuit designs and device modeling, currently working in leading U.S. and Taiwan microelectronics companies. His research interest is in the field of low-voltage CMOS VLSI circuits and SPICE compact modeling of deep-submicron bulk and SOI CMOS and BiCMOS VLSI devices.

Dr. Kuo was awarded the prestigious Canada Research Chair Professor by the Canadian Government in 2001. He serves as an Associate Editor for the *IEEE Circuits and Devices Magazine* and the Membership Committee Chair for the IEEE Electron Devices Society. He is an IEEE Distinguished Lecturer.