A compact PE memory for vision chips

A compact PE memory for vision chips*

Shi Cong(石匆)^{1,2}, Chen Zhe(陈哲)^{1}, Yang Jie(杨杰)^{1}, Wu Nanjian(吴南健)^{1,†}, and Wang Zhihua(王志华)^{2,3}

[1] State Key Laboratory of Superlattices and Microstructures, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China
[2] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
[3] Institute of Microelectronics, Tsinghua University, Beijing 100084, China

Abstract: This paper presents a novel compact memory in the processing element (PE) for single-instruction multiple-data (SIMD) vision chips. The PE memory is constructed with 8 × 8 register cells, where one latch in the slave stage is shared by eight latches in the master stage. The memory supports simultaneous read and write on the same address in one clock cycle. Its compact area of 14.33 μm^2/bit promises a higher integration level of the processor. A prototype chip with a 64 × 64 PE array is fabricated in a UMC 0.18 μm CMOS technology. Five types of the PE memory cell structure are designed and compared. The testing results demonstrate that the proposed PE memory architecture well satisfies the requirement of the vision chip in high-speed real-time vision applications, such as 1000 fps edge extraction.

Key words: vision chip; PE; memory architecture; SIMD; edge extraction
DOI: 10.1088/1674-4926/35/9/095002 **EEACC:** 2570

1. Introduction

Recently, the vision chip has gained more and more attention from worldwide researchers[1−8]. The chip integrates a high-speed image sensor and parallel processors onto a single silicon die, and eliminates the serial processing bottleneck in the traditional vision systems. Many vision chips adopt a pixel-parallel processing element (PE) array processor to speed up low-level image processing in a single-instruction multiple-data (SIMD) fashion[2, 3, 6−8]. The number of PE units in the PE array processor determines the image resolution and processing performance. Under the strict and fixed vision chip area limitation, a smaller PE area can support more PE units. One PE unit mainly consists of a simple arithmetic-logic unit (ALU) and a local memory with small storage capability. Since the PE memories are distributed all over the PE array processor and occupy a major part of the PE area, a compact PE memory structure is expected. A large chip area would be consumed by sense amplifiers if the memories are automatically compiled by the design tools[8]. Some early vision chips use registers as the PE memory cells[2]. However, the memory capacity is strictly limited by the large area consumption of these registers, thus the PE cannot handle complex algorithms. The authors in Ref. [3] turned to use latches as the memory cell to reduce the area consumption, but it could not support simultaneous read and write on the same address in one clock cycle. This degrades the processing speed and makes programming difficult. Some recent vision chips employ fully customized latches to further reduce the area consumption[6, 7]. They also lack the ability to support the simultaneous read and write on the same address in one clock cycle.

In this paper, a novel structure of the PE memory is proposed. It contains 8 × 8 cells arranged in a two-dimensional plane. The proposed PE memory cell is based on a two-stage register. In the register, the master-stage is a static latch, while the slave-stage is a dynamic latch. Moreover, one slave-stage latch is shared among eight master-stage latches in eight registers. So the area consumption can be significantly reduced. This PE memory can realize the optimization of the tradeoff among processing performance, area consumption and robustness. It behaves as registers that enables simultaneous read and write on the same address in one clock cycle. On the other hand, it consumes as little area as the memory based on fully-customized latches. The memory is fabricated in a UMC 0.18 μm 1P6M CMOS technology. One memory cell occupies only 14.33 μm^2, thus a total 64-bit PE memory consisting of 8 × 8 such cells only consumes 916.9 μm^2 chip area. This permits the integration of a larger 128 × 128 PE array on a relatively small 30 mm^2 chip.

2. PE array architecture

The PE array architecture is shown in Fig. 1. It consists of 64 × 64 PE units and an I/O interface. The PE array is 4-connected in a two-dimensional (2D) topology. In the PE array processor, all PE elements run with the same single instruction but different data (SIMD). The PE array processor performs pixel-parallel low-level image processing, such as background subtraction, image filtering and thresholding, to enhance the image details of interest. The 2D SIMD PE array processor speeds up the low-level processing by a factor of O (N × N). In order to realize the pixel-parallel processing, each PE must

* Project supported by the National Natural Science Foundation of China (Nos. 60976023, 61234003) and the Special Funds for Major State Basic Research Project of China (No. 2011CB932902).

† Corresponding author. Email: nanjian@semi.ac.cn

Received 25 January 2014, revised manuscript received 10 March 2014

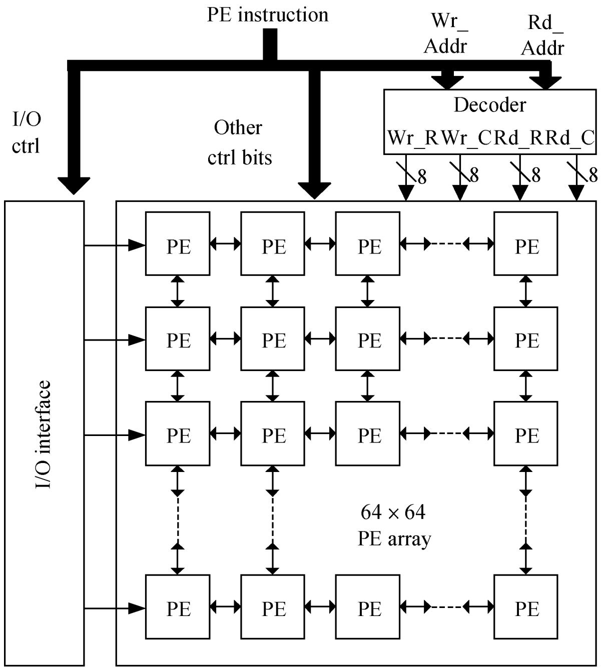


Fig. 1. The PE array architecture.

simultaneously and independently access its own pixel data, so a local memory is needed in each PE.

Figure 2 shows the PE block diagram. Each PE unit mainly contains a 1-bit arithmetic-logic unit (ALU), an 8-bit data buffer and a 64-bit local memory with a 1-bit width. The 1-bit ALU performs single-bit operations of ADD, INV, AND and OR. It can also perform multiple-bit operations in a bit-serial manner. The ALU executes one-cycle read-modify-write operations on the memory data from itself, or from one of its four nearest neighboring PEs on the east (E), south (S), west (W) and north (N). The processed data can be written back either into the memory or into the 1-bit temporal register (T-reg). The 8-bit data buffer is inserted into the PE cell to support simultaneous global data I/O flow and SIMD PE processing. The buffer is implemented in a shift-register structure, which is based on fully customized transmission gates to save area. The local memory consists of 8 × 8 memory cells arranged in a 2D plane. The ALU operand selection signals, the ALU operation codes, and the memory read and write addresses are all from the PE instructions stored in an external memory. One memory address is decoded and split into two 8-bit one-hot signals to activate only one row line and only one column line. So the memory cell on the intersection of the activated row and column lines can be accessed. For example, the write address (Wr_Addr) in Fig. 1 is decoded into Wr_R [7 : 0] to select the cell row, and Wr_C [7 : 0] to select the cell column.

3. Circuit design

3.1. Static latch and dynamic latch

In this paper, a 7-T static latch and a 5-T dynamic latch are designed as two basic circuits used in memory cell. The static latch has the advantage in keeping the signal level, while the dynamic latch occupies less area.

As shown in the schematic of the static latch in Fig. 3(a), it consists of two consecutive inverters and a feedback PMOS transistor. The control signal at the PMOS transistor decides whether the feedback loop is open or closed. When the feedback loop is closed, the signal level is stable, and when the feedback loop is open, the stored 1-bit data is able to be refreshed. Outside the feedback loop, we designed two separate NMOS transistors serving as data transmission control logic to save area. The control signal on the input data path is exactly the control signal at the PMOS transistor so that when the WR NMOS is on, the feedback PMOS is off to accomplish the data write process, and when the WR NMOS is off, the feedback PMOS is on for keeping the signal state stable. The WR and RD NMOS transistors should not be on simultaneously in case the static latch is transparent from input to output. The timing diagram of the dynamic latch is given in Fig. 3(b).

The 7-T static latch circuit has been simulated by the commercial simulation tools and its basic read and write functions are successfully achieved. Meanwhile, two non-idealistic characteristics of the circuit need to be mentioned. First, the WR NMOS transmission transistor on the input data path may cause a threshold voltage drop when transmitting a high signal level at node N1. Secondly, the rapid change from 0 to 1 of the control signal at the RD NMOS may cause a slight voltage fluctuation at inner node N3. However, the closed feedback loop helps recover the original signal level in a very short time.

Compared to a 7-T static latch, the chip area can be further reduced by employing a 5-T dynamic latch with fewer transistors. As the schematic of the dynamic latch shows in Fig. 4(a), it simply utilizes an inverter and a weak PMOS transistor to keep the signal level. It should be emphasized that the feedback PMOS transistor needs to be weak enough to allow the input data 0 to be written. When the input data is 1, the signal level at N1 is high, and the signal level at N2 is low. In this case, the PMOS transistor is open, and N1 recovers to VDD. However, when the input data is 0, the signal level at N1 is low, and the signal level at N2 is high. In this case, the PMOS transistor is closed, and N1 keeps its signal level dynamically by keeping in charge of the parasitic capacitance at N1. The timing diagram of the dynamic latch is given in Fig. 4(b).

The function of the 5-T dynamic latch is successfully verified by the simulation results. The transient waveform shows that the WR NMOS transmission transistor does not cause a threshold voltage drop at N1, because the PMOS transistor recovers the signal level to VDD soon after N1 turns from signal level low to high. It should be mentioned that the output signal of the dynamic latch is inverted because the dynamic latch has only one inverter inside. The interference from the output node to N2 through RD NMOS is also weakened compared to the 7-T design because there is a lack of feedback path in the 5-T dynamic latch.

3.2. Five types of memory cell

The PE memory can be implemented with various types of cell. Different types of memory cell exhibit different advantages and disadvantages. We have designed five types of memory cell in this subsection.

The first type of memory cell consists of the designed 7-T static latch, as shown in Fig. 5(a). In order to reduce the routing difficulty, we present a novel memory structure supporting

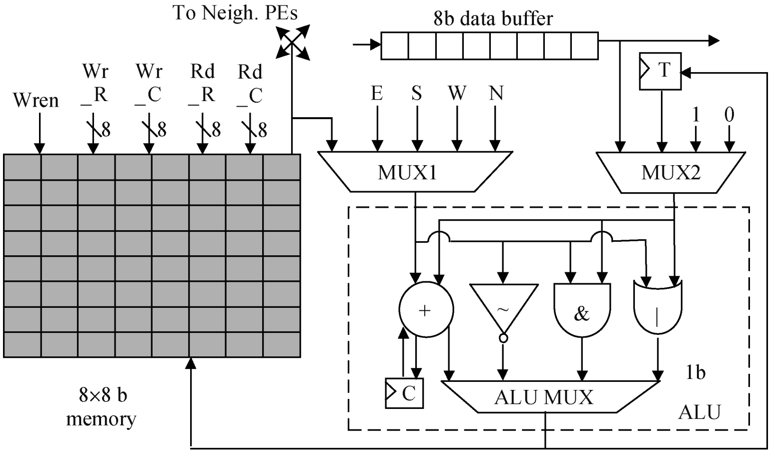


Fig. 2. The PE block diagram.

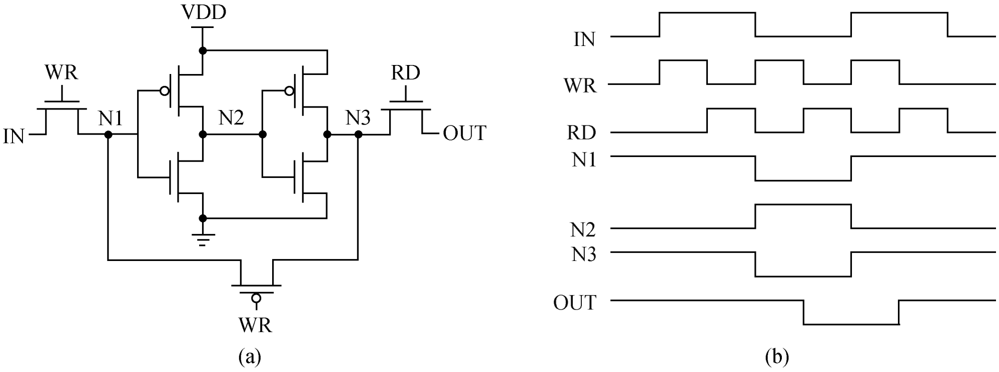


Fig. 3. The 7-T static latch. (a) Circuit schematic. (b) Timing diagram.

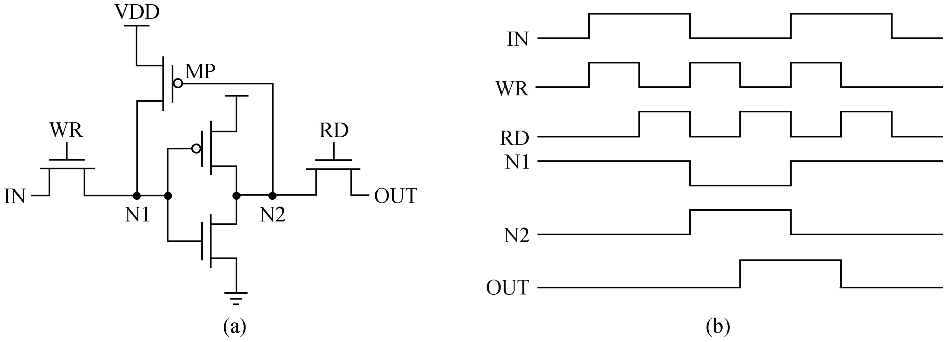


Fig. 4. The 5-T dynamic latch. (a) Circuit schematic. (b) Timing diagram.

addressing from two perpendicular directions. The designed memory cell contains an 8 × 8 latch cell. For this static-latch type memory cell, eight static latches are connected in parallel. For each 8-bit-latch structure, separate NMOS transmission transistors are added in both of the input and output data paths. These transmission transistors serve as an 8-to-1 multiplexer to realize random access addressing in a most compact way.

Based on the static-latch type memory cell, a slightly different design method is used to replace the 7-T static latch with a 5-T dynamic latch, as shown in Fig. 5(b). In this way, the memory cell benefits in circuit area from the reduced transistor number. On the other hand, the dynamic-latch type memory cell has a weaker capability in storing data compared to the former design. However, this lack of stableness is to some extent compensated by the high data refreshing rate in the massively parallel processor.

Although the two memory cells have the advantage in circuit area, its random access function is limited because reading and writing the same address memory cell is not allowed at the same clock rising edge. To further conquer this drawback, we present another two types of memory cell architecture with separate master and slave stages.

In the third type of memory cell architecture shown in Fig. 5(c), eight static latches connected in parallel serve as the master stage of eight memory cells. In each static latch unit, an additional inverter is included to suppress the interference

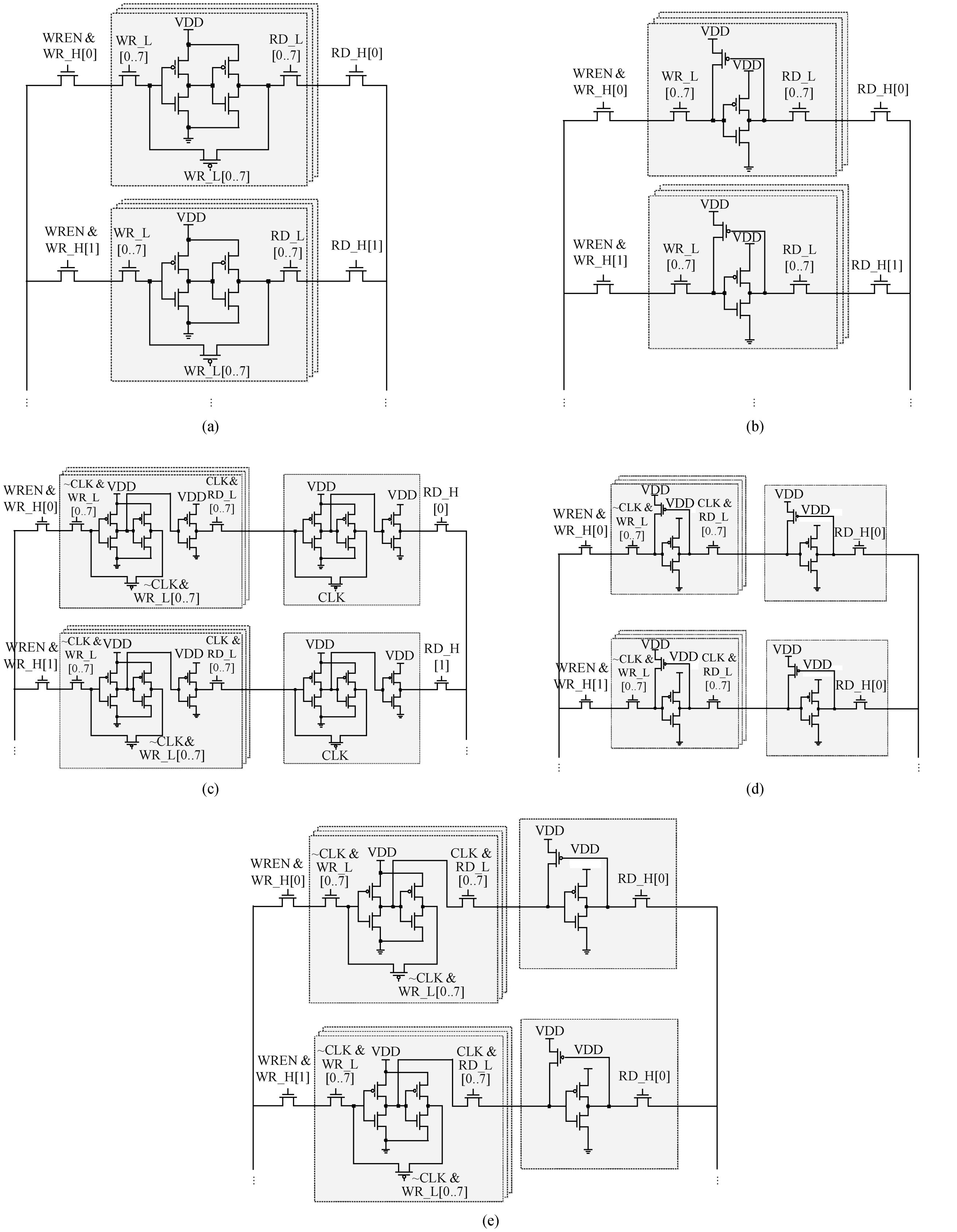


Fig. 5. Five types of PE memory cell. (a) The static-latch cell. (b) The static-latch cell. (c) The static-master static-slave (SMSS) cell. (d) The dynamic-master dynamic-slave (DMDS) cell. (e) The static-master dynamic-slave (SMDS) cell.

Table 1. Summary of PE memory cell types.

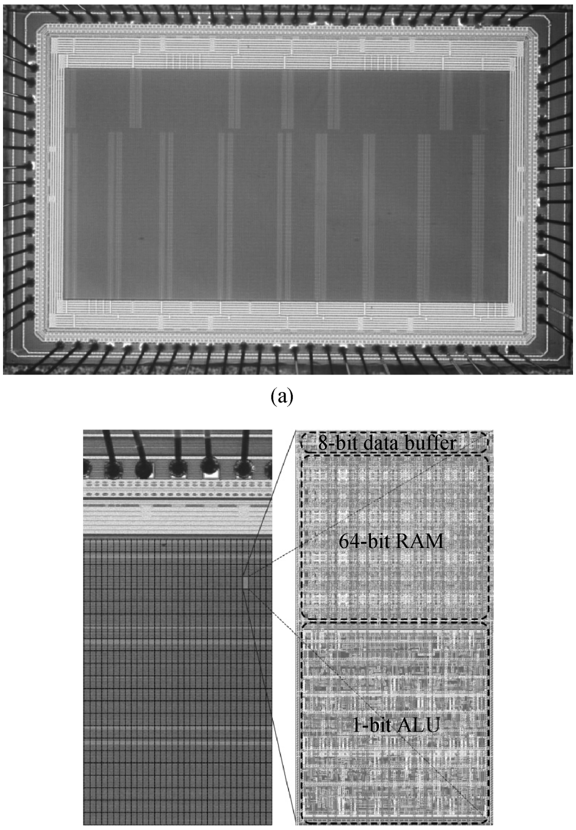
Memory cell structure	PE array	Area/64-bit (μm^2)	Simul. Wr & Rd	Robustness
Static latch	32 × 16	897.5	No	Poor
Dynamic latch	32 × 16	783.5	No	Very poor
SMSS register	64 × 16	1229.9	Yes	Very strong
DMDS register	32 × 32	783.9	Yes	Poor
SMDS register	32 × 32	916.9	Yes	Strong

from the output node to the inside structure in the master stage. The slave stage serves as an output buffer made up of a similar static latch with the additional inverter, and its transmission transistor in the input data path is omitted. The separated master and slave memory cell architecture allows new data to be written into the master stage while keeping current data to be read in the slave stage, like a 1-bit flip–flop register. In this way, the memory cell supports writing and reading the same address at the same clock rising edge. In the master stage, the control signal of the NMOS transmission transistor on the input data path and the feedback PMOS transistor is the logic AND of the inverted clock signal and the decoded writing address, and the control signal of the NMOS transmission transistor on the output data path is the logic AND of the clock signal and the decoded reading address. In the slave stage, the control signal of the feedback PMOS transistor is the clock signal. In this way, when the rising edge of the clock signal comes, the master stage is in the sustaining period while the slave stage is in the evaluation period. On the other hand, when the falling edge of the clock signal comes, the master is in the evaluation period and the slave is in the sustaining period. The advantage of the static-master static-slave (SMSS) architecture is that it is stable, but the cost of the robust feature is that it has nearly 40% more transistors than the static-latch type architecture.

By utilizing dynamic latches to replace the static latches in the third type of memory cell, the dynamic-master dynamic-slave (DMDS) architecture is realized, as shown in Fig. 5(d). The advantage of this design is that the memory circuit area is nearly the same compared to that in the second type. However, it still suffers from a relatively weak capability of keeping stored signal levels.

By combining the third and the fourth type of memory cells, we finally came up with a novel memory cell architecture, which has the static-master and the dynamic-slave (SMDS) stages, as shown in Fig. 5(e). Since the shared slave stage refreshes its signal level in each clock period, it does not need to keep the same signal level for longer than tens of nanoseconds, so the signal level in the slave stage would be well reserved. In the master stage, the internal inverter for interference signal isolation is canceled to save the circuit area. Compared to the SMSS architecture, the transistor number is reduced by more than 30%, and the circuit area is also decreased by about 25%.

To sum up, five types of memory cells are presented in this paper, each of which employs a different memory architecture. Among these types of memory cells, the SMDS type shows the best overall performance. First, it can read and write the same address at the same clock rising edge. Secondly, the design of a dynamic slave stage lets the architecture meet a balance between the compact circuit area and the robust capability of keeping signal levels.



(a)

(b)

Fig. 6. (a) The chip microphotograph. (b) PE layout.

4. Chip implementation

A prototype chip of the PE array processor based on the five types of memory cell was fabricated in a UMC 0.18 μm 1P6M CMOS technology. The 64 × 64 PE array is partitioned into five regions. Each region uses one of the five proposed types of PE memory cell, as listed in Table 1. The maximum clock frequency of the chip reaches 50 MHz. However, we were not able to measure the maximum clock frequency for the PE memory alone, because it is tightly coupled with the PE logic circuits on this prototype chip without an independent testing port. Even so, the post-layout simulation results showed that the maximum clock frequency of the PE memory alone can reach as high as 100 MHz. The PE memory will not produce a speed bottleneck for the PE array processor. Figure 6(a) shows the chip microphotograph. Its area is 4.7 × 3.1 mm^2. The layout of one PE unit with the SMDS memory cell structure is shown in Fig. 6(b). The 1-bit ALU and the 64-bit PE memory consume nearly an equal area in one PE. Among the five memory types, the SMDS register-based memory cell achieves the

Table 2. Comparison with a synthesized design and other vision chips.

Reference	This work	Synthesized design[8]	SPE[2]	ASPA[6]
Technology (μm)	0.18	0.18	0.35	0.35
PE array	64 × 64	64 × 64	64 × 64	19 × 22
PE cell size (μm^2)	66.3 × 33.4	~ 6000	67.4 × 67.4	100 × 117
# of Transistor/PE	864	N/A	N/A	460
PE memory	64-bit register	64-bit SRAM	24-bit register	64-bit DRAM
Memory cell size (μm^2)	14.33	78.1	N/A	3.6 × 8.7

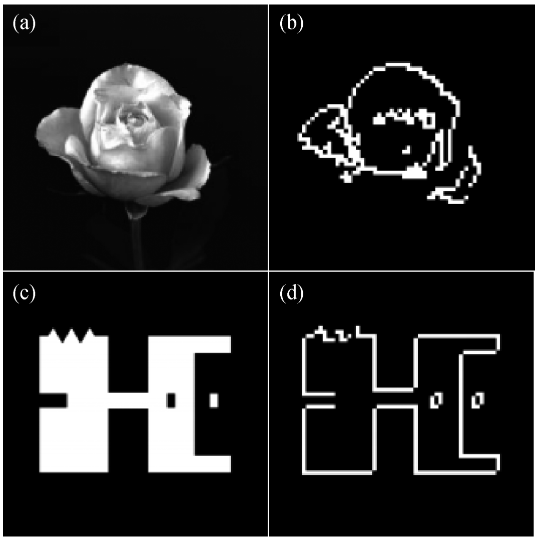


Fig. 7. Image edge extraction on the prototype chip. (a), (c) are original images, and (b), (d) are corresponding results.

compact area of 14.33 μm^2/bit, thus a total 64-bit PE memory based on this cell only consumes a 916.9 μm^2 chip area. This permits an integration of a larger 128 × 128 PE array on a relatively small 30 mm^2 chip.

An evaluation system was developed to test the performance of the prototype chip. In this evaluation system, the chip to be tested is mounted on a PCB board, which is connected to an external FPGA board. The FPGA board is finally linked to a PC host. We have also developed the necessary software for testing the chip, including a PE code compiler and the processing results monitor. We wrote the PE program in a C-like language, and then used the compiler to generate binary PE instruction codes. These codes and the test images were first downloaded to the FPGA board. Then the instructions were stored in the memory of the FPGA, and the image data was shifted into the prototype chip's PE array column by column under the control of the FPGA board. After that, the PE array began to perform various low- and mid-level image processing algorithms following the PE instructions stored in the FPGA. Finally, the processed results were outputted to the PC host through the FPGA board.

Figure 7 gives the processing results of edge extraction on the prototype chip. A 64 × 64 image is inputted into the PE array on the chip. Then the Laplacian filtering is performed on this image to compute its gratitude image. For each pixel $f_{\rm I}(x, y)$ in the original image, its corresponding pixel value $f_{\rm P}(x, y)$ in the filtered image is calculated according to Eq. (1):

$$f_{\rm P}(x,y) = f_{\rm I}(x-1,y) + f_{\rm I}(x+1,y) + f_{\rm I}(x,y-1) + f_{\rm I}(x,y+1) - 4f_{\rm I}(x,y). \tag{1}$$

Finally, the $f_{\rm P}(x, y)$ image is thresholded into a binary image $f_{\rm O}(x, y)$, which represents the extracted edge. The thresholding is formulated by Eq. (2):

$$f_{\rm O}(x,y) = \begin{cases} 1, & |f_{\rm P}(x,y)| \geqslant T, \\ 0, & |f_{\rm P}(x,y)| < T, \end{cases} \tag{2}$$

where constant T is the threshold. The edge extraction is relatively complex, and involves almost all possible operations of the PE. It can be finished at a very high speed of above 1000 fps on the 50 MHz prototype chip.

To further validate the robustness of the proposed PE memory, the chip was operated to perform the algorithm on a large set of digital images at 50 MHz under normal conditions. Then the processing results of the chip were compared with the corresponding software-based results using the same algorithm. No differences of the corresponding results were observed during this comparison. Thus the robustness of the PE memory can be asserted.

The prototype chip is compared with a synthesized chip layout[8] under the same architecture and similar silicon technology. In this comparison, the PE array in our prototype chip is assumed to uniformly adopt the SMDS register-based memory cell structure, which exhibits the best tradeoff between performance and area consumption. In this work, the SMDS register-based 64 × 1-bit PE memory occupies only a 916.9 μm^2 chip area, as shown in Table 1. The synthesized PE logic circuits occupy a 1297 μm^2 chip area. So the PE area is totally 2214 μm^2. On the other hand, in the previously reported vision chip[8], the synthesized PE logic circuits occupy ~1000 μm^2 chip area, but its compiled PE memory with the same 64 × 1-bit capacitance occupies 5000 μm^2. As a result, the PE area in the previous chip consumes as large as a ~6000 μm^2 area. Such a comparison indicates that our full custom design with the novel PE memory structure leads to a significant area reduction by as much as 63%. We have also compared our design with other state-of-the-art vision chips. All the comparison results are listed in Table 2.

5. Conclusion

This paper proposes a novel PE memory for the SIMD vision chip. Five types of PE memory cell were designed.

Among them, the static-master-dynamic-slave register-based cell structure exhibits the best tradeoff between performance and area consumption. It realizes simultaneous read and write at the same address in one clock cycle, and reduces the chip area by sharing one slave-stage dynamic latch among eight master-stage static latches. It consumes only 14.33 μm^2/bit, which is 63% smaller than the conventional compiled memory structure with comparable robustness. A prototype with a 64×64 PE array is fabricated with a 0.18 μm CMOS technology, and was successfully applied in 1000 fps image edge extraction. The experimental results suggest that the vision chip with the proposed novel PE memory structure can be used in widespread high-speed vision applications. One drawback of this work is the lack of independent testing ports for the PE memory alone. We plan to add the independent testing circuits and testing ports for the PE memory in our following work on vision chips.

References

[1] Ishikawa M, Ogawa K, Komoro T. A CMOS vision chip with SIMD processing element array for 1 ms image processing. IEEE Int Solid State Circuits Conf (ISSCC), San Francisco, CA, 1999: 206

[2] Komuro T, Kagami S, Ishikawa M. A dynamically reconfigurable SIMD processor for a vision chip. IEEE J Solid-State Circuits, 2004, 39(1): 265

[3] Miao W, Lin Q, Zhang W, et al. A programmable SIMD vision chip for real-time vision applications. IEEE J Solid-State Circuits, 2008, 43(6): 1470

[4] Yamashita H, Sodini C. A CMOS imager with a programmable bit-serial column-parallel SIMD/MIMD processor. IEEE Trans Electron Devices, 2009, 56: 2534

[5] Cheng C, Lin C, Li C, et al. iVisual: an intelligent visual sensor SoC with 2790 fps CMOS image sensor and 205 GOPS/W vision processor. IEEE J Solid-State Circuits, 2009, 44(1): 127

[6] Lopich A, Dudek P. A SIMD cellular processor array vision chip with asynchronous processing capabilities. IEEE Trans Circuits Syst I: Regular Papers, 2011, 58(10): 13

[7] Zhang W, Fu Q, Wu N. A programmable vision chip based on multiple levels of parallel processors. IEEE J Solid-State Circuits, 2011, 46(9): 2132

[8] Shi C, Yang J, Han Y, et al. A 1000 fps vision chip based on a dynamically reconfigurable hybrid architecture comprising a PE array and self-organizing map neural network. IEEE Int Solid State Circuits Conf (ISSCC), San Francisco, CA, 2014: 128