Embedded SRAM Technology for High-End Processors

● Hiroshi Nakadai ● Gaku Ito ● Toshiyuki Uetake

Fujitsu is the only company in Japan that develops its own processors for use in server products that support the social infrastructure. Its processor development strategy is to collaborate with the internal semiconductor group and simultaneously develop processor and semiconductor technology. This paper introduces SRAM development technology, which is a complex technology combining both semiconductor manufacturing and circuit systems. It fully meets conflicting server processor requirements such as high performance, small area and low power consumption. It is a technology that is essential for starting up new technology and having it fully operational at the same time. Primary cache SRAM speed determines the processor clock rate, while the access frequency of external memory, which is the data processing bottleneck, is determined by the RAM capacity of the high-density secondary cache SRAM. Thus SRAM is a key technology for server processors. As finer semiconductor technologies progress, various problems arise and the variability of the memory cells in SRAM gets bigger. Consequently, development of SRAM that meets the server processor requirements is getting critical. This paper describes our SRAM development methodology.

1. Introduction

Fujitsu is the only company in Japan that develops its own processors for use in server products that support the social infrastructure. While high performance, high density and low power consumption are required for server processors, it is primary cache SRAM that determines the rate of clock frequency limit. However, the RAM capacity of secondary cache SRAM, which determines the access frequency of external memory, is a data processing bottleneck. Therefore, SRAM is considered to be a key component of processors, and it needs to be optimized at a high level to satisfy conflicting requirements just the same as with processors. SRAM is comprised of memory cells that serve as memory devices and a peripheral circuit used to control them. SRAM has been optimized by reviewing this configuration on a continuous basis.

While semiconductor technology has been getting smaller and smaller, as predicted by Moore's Law, the performance and stability of SRAM memory cells required as memory has been getting steadily worse due to inevitable quality dispersion in the manufacturing process accompanying this miniaturization.

In such a situation, it has become more and more difficult to develop SRAM that can satisfy the requirements for server processors. In this paper, we will describe the specific points of this technical challenge and also Fujitsu's efforts for finding a solution.

2. Technical challenges in SRAM development

Semiconductor manufacturers strive to take the initiative in the competition to miniaturize chips by using their advanced technologies. With each new technology, the area of an SRAM memory cell that symbolizes the process technology is halved. However, downsizing of the memory cell will increase the inconsistency of device characteristics. This is due to inevitable physical phenomena such as diffusion and fluctuation of impurities necessary for manufacturing transistor devices composing a memory cell or non-uniform geometry. As a consequence, designing SRAM has become extremely difficult.

Generally speaking, the performance and stability of an SRAM memory cell are in a trade-off relationship. Lowering the threshold value for transistors composing a memory cell will improve the memory cell performance, but this leads to poor stability. With increased dispersion of threshold values accompanying semiconductor miniaturization, the incidence of poor stability of memory cells will increase. Therefore, to ensure yield, the threshold value of a device should be set conservatively (with the tradeoff of lower performance) in general. Accordingly, downsizing a memory cell in accordance with the trend of semiconductor technology leads to a relative degradation of memory cell performance. This, in turn, means it is hard to make SRAM that satisfies the requirements in server processor applications, where high-speed operations are sought after.

3. Approaches in SRAM development

In such circumstances, the authors' approaches to develop SRAM with high-performance, high-density and low power consumption characteristics for server processors focus roughly on the following three points:

1) Development of a memory cell optimal for server processors

The smallest memory cells in each generation are not always the best ones. Area and performance may need to be optimized depending on the processor-related requirements and SRAM circuit type. From the early stage of technological development, the authors have worked in collaboration with Fujitsu's Semiconductor Division in an attempt to jointly develop an optimal memory cell for server processors.

2) Development of circuit technologies that serve as solutions for miniaturization-related issue

SRAM is comprised of memory cells and peripheral circuits to control them. The authors have formulated a technology roadmap for SRAM memory cells to control them. This was done by researching technical trends and undertaking unique R&D activities to address the above-mentioned technical challenges. Based on this roadmap, our team is committed to establishing a new circuit technology by repeating prototype tests for next-generation devices. At the same time, it is committed to following a product macro-design scheme.

3) Improvement of simulation technology

Fewer tests using actual prototypes are required while assuring quality in the design stage by calculating the statistical worst case model for a memory cell. This is done while giving consideration to the dispersion of device characteristics and reflecting the data in circuit simulations of the whole SRAM. This approach makes it possible to establish semiconductor technology and full operations of processors.

These three approaches are explained in more detail in the next section.

4. Development of optimal memory cell

First, we will detail the issues caused by miniaturization of semiconductors. High density SRAM memory cells of the 45 nm process generation adopted in SPARC64 VIIIfx processor[1] and its equivalent circuit are shown in **Figure 1**. As can be seen from the illustration, 6 transistors are arranged on a field less than 1 μm^2. Several

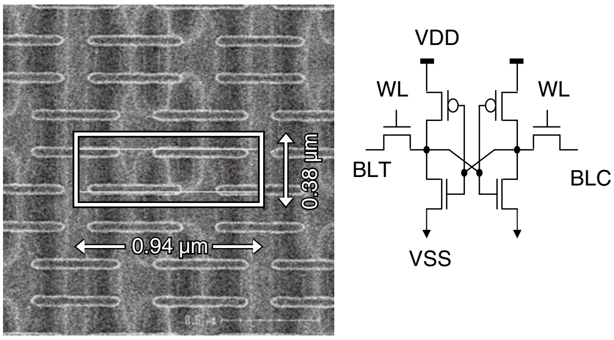
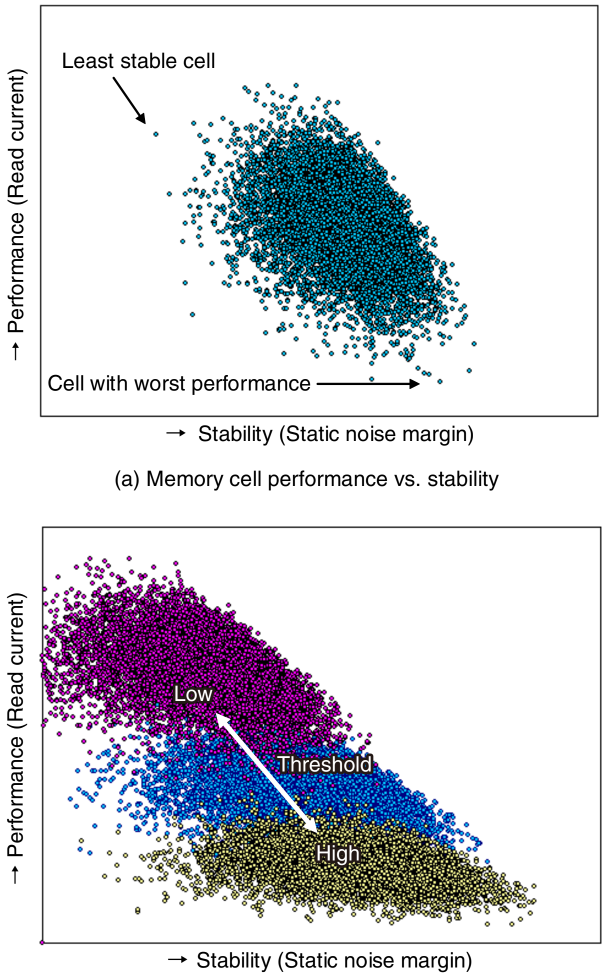


Figure 1
SEM image of SRAM memory cell and equivalent circuit.



(a) Memory cell performance vs. stability

(b) Fluctuation of distribution by threshold adjustment for memory cell

Figure 2
Characteristic distribution of memory cell.

ten million circuits like this are used to compose secondary cache memory.

Using a large amount of such minute transistors will make the memory susceptible to the impact of manufacturing dispersions. **Figure 2 (a)** indicates the results of a simulation showing the distribution of memory cell characteristics when the manufacturing dispersion of individual transistors is considered. The vertical axis represents the read current, an indicator of the performance of a memory cell, while the horizontal axis represents the static noise margin (SNM), an indicator of its stability. One dot in the graph corresponds to one memory cell. While 10 000 cells are plotted in this simulation, an actual processor has several thousand times this number. Therefore, the area of dispersion will become wider in reality. The performance of a processor is determined by the worst memory cell among the many memory cells integrated in it, and the cell with the worst stability determines the yield. A greater dispersion means further expansion of the distribution of points in the graph.

If the greater dispersion results in an SNM less than 0 in the memory cell with the worst stability, it will be impossible to maintain data any more. To prevent this, SNM can be enhanced by elevating the threshold value of the device. However, as indicated in **Figure 2 (b)**, this is only possible at the expense of performance. This is the problem in memory cell development accompanying its miniaturization.

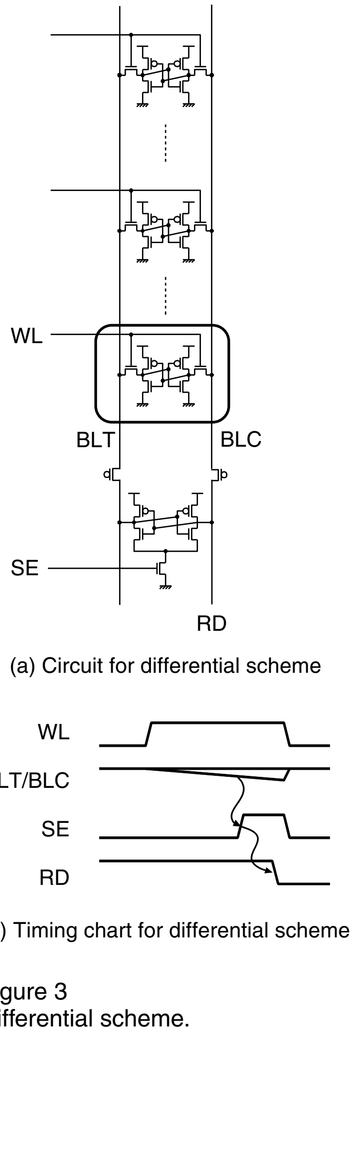
In each generation, semiconductor manufacturers have halved the area of SRAM memory cells through strategic use of advanced technologies. However, the memory cell with the minimum size in each technology is not always the optimal one for server processors. SRAM is comprised of memory cells and peripheral circuits to control them. These two closely related elements require optimization depending on the requirements of SRAM in terms of its performance, area and power consumption. For instance, to realize high-speed operations by using the smallest memory cell, a sense amplifier available for reading out even the fine amplitude by a dropped read current is necessary. In general, when input amplitude becomes smaller, the space occupied by the sense amplifier that amplifies the input will be larger. Because of this, the SRAM size becomes larger even if the memory cell is small. Further, to drive a larger sense amplifier, the power consumption gets larger. Accordingly, to ensure a read current that satisfies the performance requirements while minimizing the area occupied by SRAM, it is necessary to adopt a somewhat larger memory cell. Thus, the optimal area of a memory cell will change depending on the requirements of the SRAM. The authors optimized primary cache SRAM that needs to operate at high speed and secondary cache SRAM for which reduction of space is a priority. As a consequence, memory cells with different areas were adopted. Further, by changing the threshold of transistors, the authors aimed to optimize the characteristics of these SRAM devices.

The optimal type of cell depends not only on the requirements of the SRAM but also on its circuit configuration. Therefore, the authors work in collaboration with the Semiconductor Division from the early stage of technological development.

5. Development of circuit technology

First of all, we will describe the read operation of SRAM based on the conventional differential scheme. A circuit diagram and timing chart of differential SRAM is shown in **Figure 3**. The bit lines indicated by true bit line (BLT) and compliment bit line (BLC) in the diagram are connected to many memory cells. The memory cells hold data by interconnecting the input and output of two inverters. By giving the value of "1" to a word line (WL), the memory cell to be read is selected and the held data is transferred to bit lines. However, memory cells are composed of extremely small transistors and cannot fully drive bit lines connected to many memory cells. Therefore, a voltage difference between BLT and BLC has less amplitude than the amplitude of the voltage difference between VDD and GND. Therefore, to amplify this small amplitude, a sense amplifier is activated by an enable signal SE and it is transferred to the output signal RD. Read operation is completed in this way.

Next, we will describe the circuit technology that we introduced. Various arguments have been made in places such as academic meetings on how to avoid the degradation of memory cell stability accompanying miniaturization. In our approach, we tried to reduce the load on bit lines during the memory cell drive. To achieve this, we

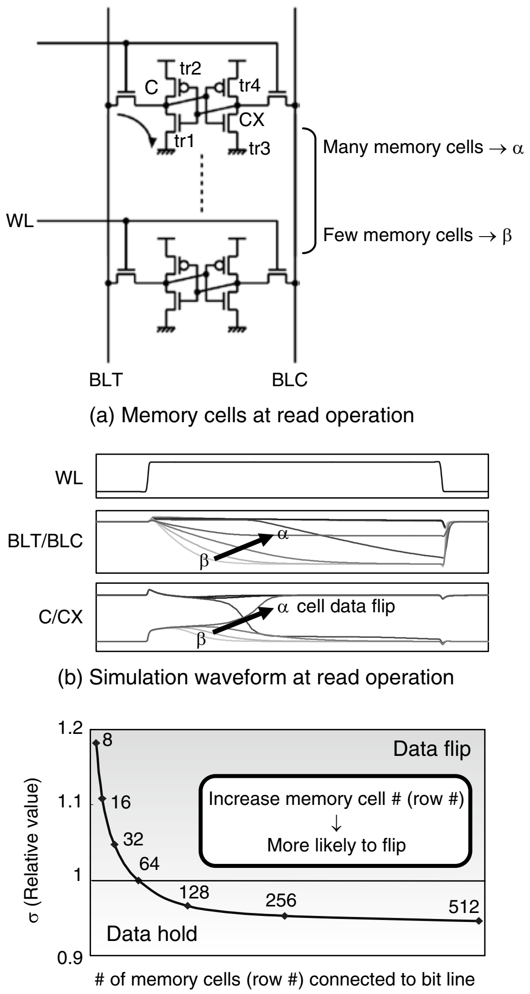


(a) Circuit for differential scheme

(b) Timing chart for differential scheme

Figure 3
Differential scheme.

reduced the number of memory cells connected to a bit line and ensured a drastic discharge of bit line at the read operation. We noticed that data flip of a memory cell can be prevented by this approach even if the memory cell is unstable. We considered that it is possible to downsize a memory cell with this approach without compromising its performance. A circuit diagram of a memory cell is indicated in **Figure 4 (a)**. **Figure 4 (b)** indicates simulation waveforms while changing the load on the bit lines for the memory cells which are likely to flip with a significant level of dispersion in manufacturing. If a read operation occurs when the load on the bit line is large (α mode) like this, the voltage of Node C will elevate. This is because the read current flows in from the bit lines. This process triggers a response and switching of the inverter comprised of tr3 and tr4, resulting in a destruction of the held data.



(a) Memory cells at read operation

(b) Simulation waveform at read operation

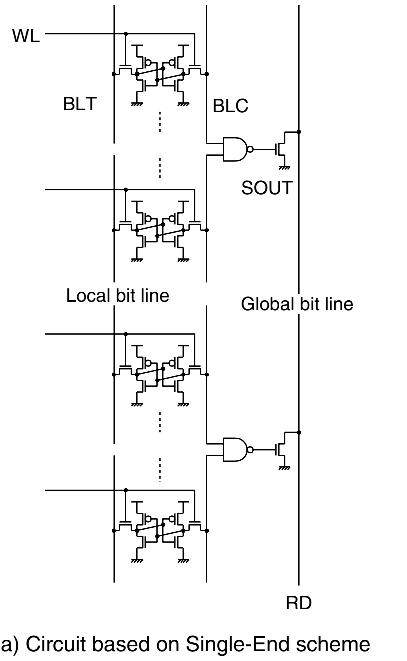
(c) Bit line load and cell stability limit

Figure 4
Memory cell stability vs. cell array structure.

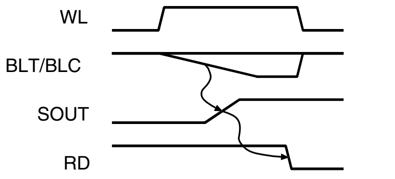
On the other hand, if the load on the bit lines is reduced (β mode), a drastic drop of potential on the bit line makes it possible to complete a read before the above-mentioned switching of the inverter. Thus, the data flip can be prevented. **Figure 4 (c)** indicates the results of a simulation. Limit dispersion to trigger switching of a maintained value is simulated while changing the load on the bit lines by changing the number of memory cells connected to the bit lines. This graph shows a plot of the switching limit, when the number of memory cells is changed. The graph is based on the assumption that the number of memory cells connected to the bit lines is 64 and the dispersion level σ at the switching limit is 1. As demonstrated by this simulation, switching is less likely to occur when there are fewer memory cells connected to the bit lines even if the level of dispersion is significant.

Further, when coming up with a concept of an SRAM circuit using this effect, it is possible to handle the process as a digital signal by sufficiently reducing the number of memory cells and enlarging the amplitude of the bit line. In this way, SRAM based on the Single-End scheme can be constructed.

Figures 5 (a) and **(b)** indicate a Single-End circuit diagram and its timing chart. Compared with the conventional scheme, the load is reduced by dividing a bit line in the Single-End scheme, making it possible to cause a full swing of the bit line if the WL is open at a read operation. This mechanism makes it possible to read using the ordinary logic gates without a differential sense amplifier. This divided bit line is called a local bit line. To consolidate the data divided with the division of the bit lines, a global bit line is used. Output to the output signal RD is carried out based on two-step read (local/global).



(a) Circuit based on Single-End scheme



(b) Timing chart based on Single-End scheme

Figure 5
Single-End scheme.

This approach offers advantages over the conventional scheme also in terms of performance, area and power consumption.

First, we will describe the advantage in terms of performance. In the Single-End scheme, the discharge time of a memory cell can be drastically reduced by making the bit line length 1/*N*. *N* units of control circuits for local bit lines are necessary and the delay overhead is generated for data integration. However, the ratio of cell discharge time versus total delay is smaller in comparison with a conventional SRAM. As a consequence, the impact on performance can be mitigated even if there are some slow memory cells due to dispersion.

Next, in terms of area, the conventional differential scheme did not lead to downsizing of the sense amplifier due to increased dispersion accompanying the miniaturization. In the Single-End scheme, however, an ordinary Logic Gate can be used while the number of control circuits is multiplied by *N* times. Therefore, it is possible to downsize a memory cell based on the normal area reduction rate accompanying semiconductor miniaturization.

Then, in terms of power consumption, the electric charge necessary for read/write operation (dynamic current) is 1/*N* in the Single-End scheme. Further, because the load on a bit line itself becomes 1/*N*, the driver to drive the line can be reduced to 1/*N*. These factors work in synergy to reduce the dynamic current.

6. Improvement of simulation technologies

The performance of memory cells has a significant influence on the overall characteristics of SRAM. Therefore, device dispersion should be taken into account in the SRAM design phase so that the memory cell with the worst characteristics to be actually generated can be calculated and modeled in advance with high accuracy. In this way, the memory cell with the worst characteristics should be reflected in the SRAM simulation.

The Monte Carlo method is a widely known method for estimating dispersion. The characteristic dispersion of memory cells in Figure 2 is the result of a simulation based on the Monte Carlo method with regard to performance and stability of 10 000 memory cells. However, in an actual processor, several ten million memory cells are installed only for secondary cache memory. To calculate the memory cell with the worst characteristics (i.e., the worst cell) to be generated in this cohort with high accuracy with an error rate within 1%, more than several billion simulations are necessary. This is practically impossible due to limited computer resources and time.

This problem was solved by applying the SRAM analysis system developed by Fujitsu Laboratories in calculating the worst cell.

In this analysis system, the dispersion coefficient is allocated, and then the worst cell is searched for while reducing the margin. In the next step, weighted sampling is carried out based on the Importance Sampling Monte Carlo (ISMC) method[2] by generating a random number while centering on the vicinity of the explored cells. In this process, Latin Hypercube Sampling[3] is used to determine a multi-dimensional random number to be generated so that the configuration of sampling points in each dimension is equivalent.

This approach made it possible to calculate the worst cell with high accuracy with simulations of just several million times. Compared with the conventional Monte Carlo method, the calculation time could be reduced to more than one-millionth with our method.

By integrating the worst-cell model calculated by the above-mentioned system and conducting highly accurate SRAM simulations, we are committed to improving design quality while reducing the number of prototypes produced.

7. Conclusion

In this paper, we reported the technical challenge in developing SRAM for server processors and our three approaches to addressing this issue.

Through these approaches, we adopted Single-End scheme to primary cache memory for SPARC64 VIIIfx processor in the 45 nm process generation. We plan to deploy this scheme also in secondary cache memory.

We want to continue working to overcome challenges associated with the miniaturization of semiconductors. We hope to help improve the performance of server processors through developing SRAM with high speed, high density and low power consumption characteristics.

References

1) T. Maruyama: SPARC64™ VIIIfx: Fujitsu's New Generation Octo Core Processor for PETA Scale Computing. Hot Chips 21, 2009.

2) R. Kanj et al.: Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events. DAC 2006, pp. 69–72.

3) A. Olsson et al.: On Latin hypercube sampling for structural reliability analysis. *Structural Safety*, Vol.25, Issue 1, pp. 47–68 (2003).

Hiroshi Nakadai
Fujitsu Ltd.
Mr. Nakadai is engaged in the development of SRAM for server processors.

Gaku Ito
Fujitsu Ltd.
Mr. Ito is engaged in the development of high-integration SRAM for secondary cache in server processors.

Toshiyuki Uetake
Fujitsu Ltd.
Mr. Uetake is engaged in the development of high-speed SRAM for primary cache in server processors.