*Invited Paper*

# Design Philosophy of a High Performance BiCMOS Microprocessor

YASUHIRO NAKATSUKA*, TAKASHI HOTTA*, SHIGEYA TANAKA*, TADAAKI BANDOH*, TETSUO NAKANO*†, ATSUO HOTTA*†, TAKASHI MORIYAMA**, SHIGEMI ADACHI** and SHOJI IWAMOTO**

The world's first 32-bit single-chip BiCMOS microprocessor has been developed. Its typical operating frequency is 70 MHz, which is 1.5 to 2.0 times faster than today's CMOS microprocessors.

To achieve such a high frequency, the delay times of critical paths had to be shortened. Three design strategies are used for this. First, to reduce the delay time of the logic circuits, bipolar transistors are used effectively by limiting their use to driving a load capacitance and sensing low-level signals. Second, to accelerate memory access, the number of inter-chip communications in the critical paths is reduced. Last, to speed up the control logic, functions that may be executed slowly are eliminated from critical paths.

Although the bipolar transistor is inherently larger than an MOS transistor, it had to be designed for a small chip area. This was done by using an optimal number of bipolar transistors and BiCMOS logic gates. The proportion of bipolar transistors used in this chip is only 1.5%, and the performance of the microprocessor depends on this low proportion.

## 1. Introduction

The performance of single-chip 32-bit microprocessors has been improving rapidly through progress in Very Large Scale Integration (VLSI) technology and computer architecture. In particular, the new Reduced Instruction Set Computer (RISC) architecture has greatly accelerated the improvement in performance. RISC microprocessors are made by using Complimentary Metal Oxide Semiconductor (CMOS) technology with a 1.5- to 0.8-$\mu$m process to achieve high integration density.

A newer technology, BiCMOS (Bipolar CMOS), is also now available. Its logic circuits offer a switching speed about twice as fast as that of the same-generation CMOS logic circuits. However, the BiCMOS integration density and power dissipation are similar to those of the CMOS.

Although there are many applications of BiCMOS technology, such as Static Random Access Memories (SRAMs) [1,2] and gate arrays [3], it has not been previously applied to a microprocessor. We have developed the world's first 32-bit BiCMOS microprocessor, with a typical operating frequency of 70 MHz.

*Hitachi Research Laboratory, Hitachi, Ltd., 4026 Kuji-cho, Hitachi-shi, Ibaraki-ken 319-12, Japan.
*†Device Development Center, Hitachi, Ltd., Ome-shi, Tokyo 198, Japan.
**Asahi Works Hitachi, Ltd., Owariasahi-shi, Aichi-ken 488, Japan.



In this paper, the design philosophy and the speed-up methodologies of the BiCMOS microprocessor are discussed. The design philosophy is described in Section 2, and the basic concept of the BiCMOS elementary circuits is explained in Section 3. Sections 4 and 5 describe the methodologies of the pipeline architecture and the memory system configuration, respectively. The VLSI implementation is described in Section 6.

## 2. Design Philosophy

Although the RISC architecture gives a very good performance improvement, we adopted the CISC architecture, so as to maintain software compatibility in Hitachi's computer family. Therefore, to provide a performance superior to that of RISC microprocessors, we set the operating frequency at 70 MHz.

The maximum switching speed of the BiCMOS logic circuit is up to twice as fast as that of the CMOS one. But it may slow down to the CMOS speed, depending on the implementation. Since the BiCMOS logic circuit is larger than the CMOS one, in order to obtain high performance without an increase in size, only critical paths in the microprocessor need to be shortened.

There are four typical critical paths (1) the Control Storage (CS) paths (2) the Arithmetic and Logic Unit (ALU) paths (3) the stage control paths and (4) the memory access path (Fig. 1). The CS stores microinstructions. The CS path consists of a CS access path and the next CS address generation. The ALU

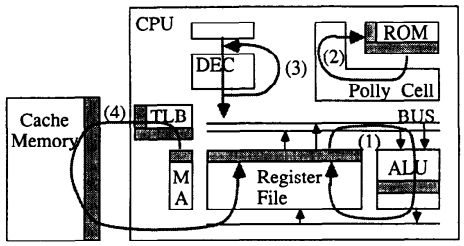



Fig. 1 Typical critical paths.

path consists of a Register File (RF) read access path, an ALU operation path, and an RF write access path. The stage control path consists of an instruction decoder and a sequence controller path. The memory access path consists of address translation and cache memory access.

The strategies we adopted to reduce the critical paths are as follows:

1. To reduce the delay time of critical paths by effectively using BiCMOS circuits.
2. To reduce the number of inter-chip communications.
3. To simplify logic forming critical paths by eliminating functions that may be executed slowly.

The critical paths consist of random and regular logic, and inter-chip communication signal lines. To speed up random and regular logic, we developed BiMOS basic cells and macrocells. The basic cells are used mainly for constructing control logic. Their functions are basic operations such as NAND, NOR, and inversion. The macrocells, which are for memories, ALU, and so en, use the smallest number of bipolar transistors among CMOS transistors now available. They realize high performance with a small increase in size. By using BiCMOS basic cells and macrocells,the delay time in the critical paths can be reduced. The next point we considered was the I/O driver.

The BiCMOS I/O driver supports a TTL signal level, which is often used by devices interfacing with the microprocessor. It is characterized by low power dissipation, and has almost the same speed as a CMOS driver. Inter-chip communications at the TTL signal level takes more than 10 ns, so it is important to reduce their number in the memory access critical path as shown in strategy 2.

Although the delay times are shortened by strategies 1 and 2 it is not enough to shorten the critical paths such as the CS path and the stage control path.

The numbers of instructions and addressing modes for the BiCMOS microprocessor are about 400 and 50, respectively. Besides these instructions, various types of checks and error routines are implemented by microinstructions for high reliability. Therefore, by adopting strategy 3, as discussed in Sections 4 and 5, the following defects of a CISC microprocessor must be overcome:

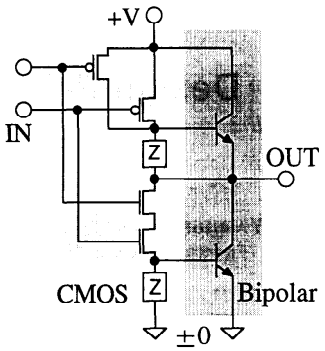



Fig. 2 BiCMOS Driver.

(i) The delay time of the decoder is long, Since it requires a very complicated instruction decoder that decodes many instructions and addressing modes.

(ii) The large amount of microprogram memory necessary to implement many high-level instructions cannot be integrated in a single VLSI chip. If they are stored in the external memory, the number of inter-chip communications is doubled and the delay time of the CS path is lengthened.

(iii) The hardware to implement the CISC architecture occupies a large area of the chip and, as a result, a sufficient amount of cache memory cannot be integrated in the chip. If there are no cache memories on the chip, the number of inter-chip communications in the memory access critical path is more than doubled.

## 3. High-speed BiCMOS Circuits

### 3.1 BiCMOS Elementary Circuits

To achieve strategy 1, we developed two types of BiCMOS circuits and applied them to macrocells.

(A) BiCMOS driver

Figure 2 shows a rough sketch of a 2-NAND BiCMOS driver. It consists of a CMOS logic circuit and a pair of bipolar transistors operating as a booster. As the bipolar transistors have high driving capability, they efficiently drive a large load capacitance. However, since both of these two transistors are not active simultaneously, the power dissipation of the driver is relatively low. Although the BiCMOS drivers are about four times larger than CMOS ones, they are needed only in the critical path, and the total integration density can be high [3].

The delay time of 2-NAND drivers in relation to their load capacitance is shown in Fig. 3. The thicker line shows the shortest delay time. The switching speed of the BiCMOS driver is 1.5 times faster than that of the CMOS driver at 1.0 pF load capacitance.

(B) BiCMOS sense circuit

Figure 4 shows a BiCMOS sense circuit. It consists of an N-type Metal Oxide Semiconductor (NMOS) logic

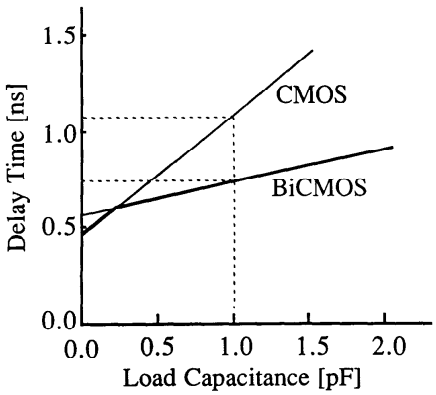



Fig. 3 Delay Time of Drivers.

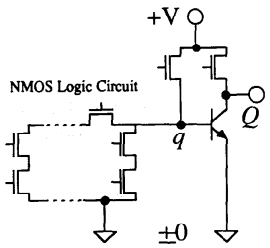



Fig. 4 BiCMOS Sense Circuit.

circuit and one bipolar transistor. The NMOS logic circuit operates as a current switch that controls the base current of the bipolar transistor. As the bipolar transistor suppresses the output voltage of the NMOS logic circuit within 0.8 V, it can operate very fast. The bipolar transistor, which has a very high cut-off frequency, can amplify this low-level and high-speed signal to a high-level signal with a short delay time. The BiCMOS sense circuit can operate twice as fast as the CMOS one by using this structure. It can be used as the sensing amplifier of ROMs, RAMs, and RFs, and also for the carry propagation circuit in the ALU [4-6].

Figure 5 shows the ratio of the delay time of CMOS/BiCMOS circuits. Cases (a), (b), and (C) show the access time of an ROM the carry generation time of the ALU, and the access time of a register file, respectively [7].

### 3.2 Application to the Macrocell

Macrocells use BiCMOS drivers and sense circuits. Figure 6 shows an example of a ROM circuit. The address decoder drives the word line, which has a large load capacitance. The memory cells and bipolar transistors make up the BiCMOS sense circuit. The ratios of bipolar transistors used are only 0.8% for the ROM and 1.7% for the data structure macrocell, which contains

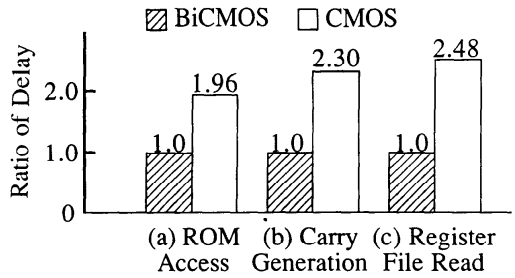



Fig. 5 Delay Time Ratio of Sense Circuits.

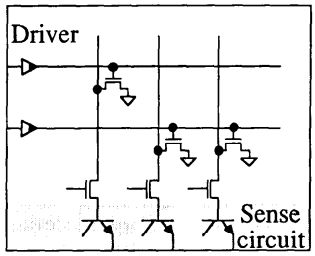



Fig. 6 ROM circuit.

ALUs, RF, and dynamic data buses. Therefore, macrocells can be accelerated by using BiCMOS technologies without increasing the area.

## 4. Cache Memory Configuration

### 4.1 External Cache Memory

To accelerate the memory access, microprocessors often have a cache memory. However, the integration density possible with today's technology (1.0 $\mu$m-0.8 $\mu$m CMOS/BiCMOS process) is insufficient to integrate a large amount of cache memory. We must realize a higher-performance memory system with a smaller number of inter-chip communications, according to strategy 2. Having a small cache memory, however, would not be effective because its miss ratio is not negligible. For example, for a 1-KB cache memory with a block size of 16 B, the miss ratio is nearly 15% [8]. If the penalty for a miss is 20 cycles, the mean number of memory access cycles is

$$0.85*1+0.15*20=3.85 \text{ cycles}$$

To reduce this penalty, the microprocessor has a 64-KB external cache memory, so the miss ratio is less than 2% [8], Even if it takes two cycles to access the external cache memory, the mean number of access cycles is reduced to

$$0.98*2+0.02*20=2.36 \text{ cycles}$$

The delay time of the cache memory is caused mainly

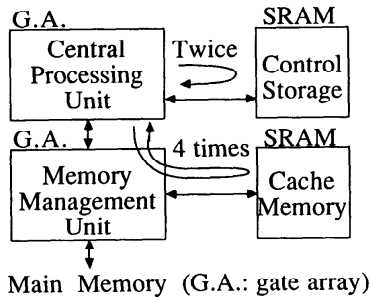



Fig. 7(a) Structure of previous system.

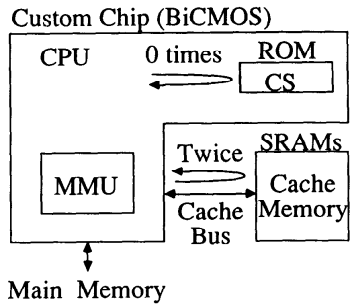



Fig. 7(b) System with standard SRAM.

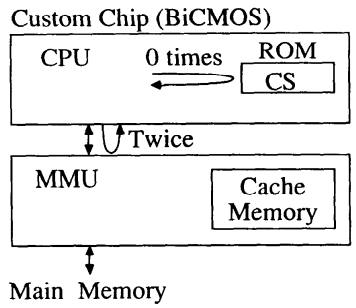



Fig. 7(c) System with custom MMU.

by the delay time of the inter-chip communication. It depends on the memory system structure.

### 4.2 Structure of Cache Memory

The memory access critical path contains a path for TLB access. The TLB is often integrated in the (Memory Management Unit) MMU chip for an ordinary microprocessor. The MMU also integrates the control logic for the cache memory. But with this system structure, there are four inter-chip communications on the cache memory access, as shown in Fig. 7(a).

To reduce the number of inter-chip communications, two system structures are evaluated, as shown in Figs. 7(b) and 7(c). Figure 7(b) shows the structure with standard SRAMs. The TLB and the control logic for the cache memory are integrated in the microprocessor. Figure 7(c) shows the structure with a custom MMU integrates the TLB and cache memory and their controller, the amount of cache memory is small; 8 KB, for example. Therefore, to realize 64 KB of cache memory, we must use eight MMU chips. To shorten the access cycles of the cache memory, SRAMs/MMUs are directly connected to the microprocessor, independent of the system bus connected tot the main memory. The delay times of the memory access path of both system structures are estimated as 32.1 ns and 33.9 ns, respectively. We adopted the standard SRAM method. Although both give similar performance, the former can be realized at a cheaper cost and with simpler logic.

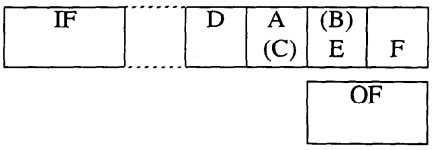



Fig. 8 Pipeline stages.

## 5. Architectural Consideration

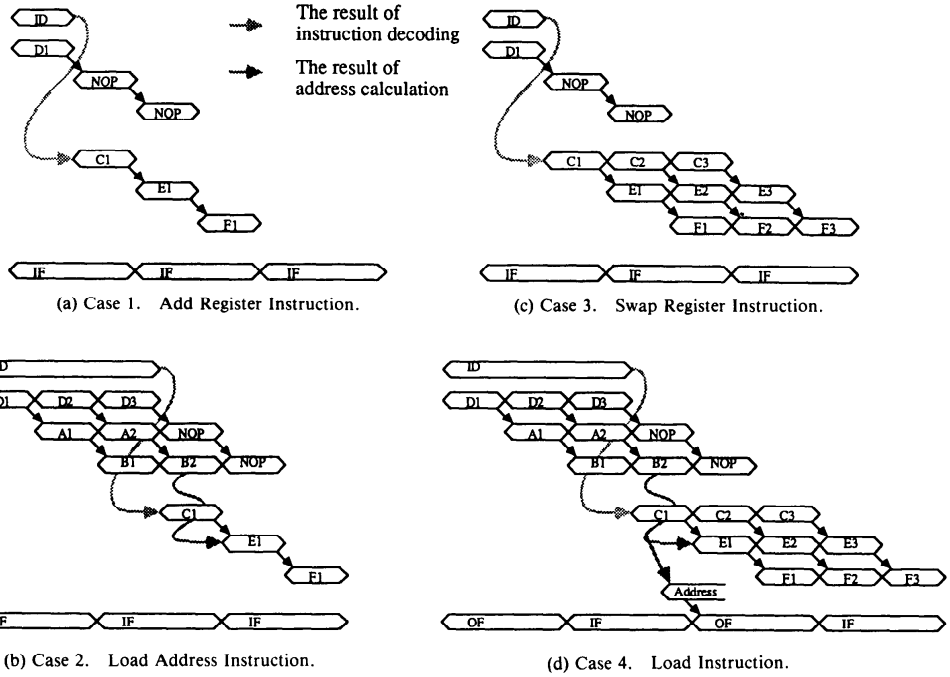
### 5.1 Pipeline Structure

To overcome the three defects of the CISC microprocessor described in Section 2, we must consider the pipeline architecture. The BiCMOS microprocessor has a five-stage pipeline structure, namely, the decode (D), address calculation (A), first execution (E), second execution (F), and instruction/operand fetch (IF/OF) stages (Fig. 8). Since two cycles are needed for the ALU path, there are two execution stages, E and F. These stages are controlled by signals generated in the previous cycle, which we call the C stage. As in the execution stages, a second address calculation stage is needed, which we call the B stage. The A and B stages are controlled by signals generated in the D stage.

Fig. 9 outlines the pipeline stages of the microprocessor.

Case 1 show the pipeline flow of an add register instruction, which adds a pair of data in the RF together and stores the result in the RF. This is a typical register-register instruction. The ID stage is the stage of instruction decoding. As these instructions do not need the address calculation, they can be executed in one cycle.

Case 2 shows the execution of a load address instruction, which stores an address in the RF. This instruction needs an address calculation, but there is only one execution cycle. We assumed the address calculation to be the addition of the contents of the base and index registers and the displacement. The first step calculation (denoted 1) adds the base register to the displacement, and the second step (denoted 2) adds the result to the in-




(a) Case 1. Add Register Instruction.

(c) Case 3. Swap Register Instruction.

(b) Case 2. Load Address Instruction.

(d) Case 4. Load Instruction.

Fig. 9 Pipeline Structure of BiCMOS Microprocessor.

dex register. The third step is used for generating control information. Thus, the address calculation takes three cycles.

Case 3 shows the swap register instruction, which swaps the contents of two registers. This instruction does not need an address calculation, but it takes three cycles to execute.

Case 4 shows the load instruction, which fetches data from memory and stores them in the RF. This is a typical memory-register instruction. Since the OF and IF stages operate every two cycles, four cycles are needed to execute the load instruction.

### 5.2 Two Level Decoder

We adopted a logic array for the decoder, to get logical flexibility. Although it is composed of a macrocell, the delay time of the logic array will be long if the number of the product terms, which is determined by the number of instructions to be decoded, is large. To reduce this number, we classified instructions into level-1 instructions, which are decoded within a cycle, and level-2 insructions, which are not. Register-register instructions are included in level 1. Since the results need not be decoded immediating for level-2 instructions whose address calculation sequences take more than three cycles, the delay time of the decoder for these instructions can be long. Figure 10 shows the structure of two decoders. A result of the level-2 decoder is available only when the level-1 decoder cannot decode

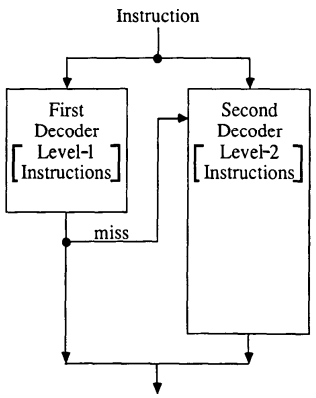



Fig. 10 Two-level instruction decoder.

an instruction.

In our architecture, the number of level-1 instructions is less than a fifth of that of all instructions. Thus, the decoder delay time for level-1 instructions, which determines the delay time of the D stage, is small.

### 5.3 External Microprogram

In ordinary microprocessors, a CS is made from a ROM on the chip. Since our microprocessor has a large number of microprograms nor everything can be integrated in the chip. On the other hand, if the CS is out-

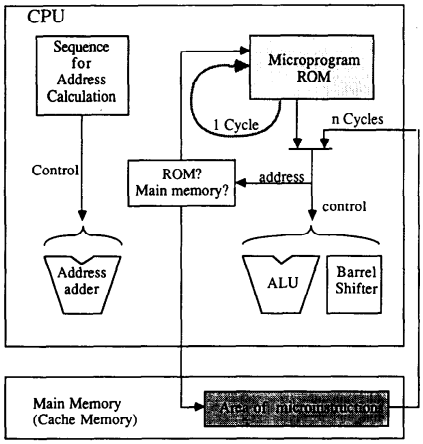



Fig. 11 External Microprogram.

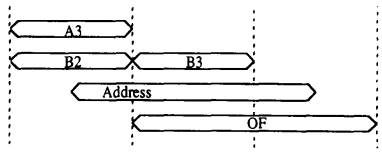



Fig. 12 Memory address pipeline.

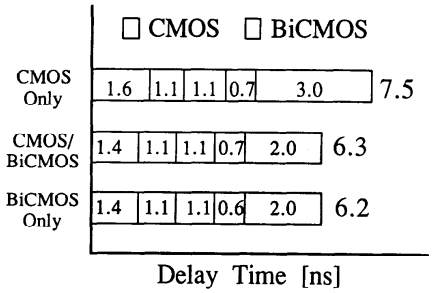



Fig. 13 Delay Time of Basic Cells.

side the chip, the delay time of the CS path includes the time for inter-chip communications. To shorten the delay time of the CS path, we divided the microinstructions into two parts (Fig. 11). The first part is a set of microinstructions that are used frequently and located in the ROM on the chip. The microinstructions in the ROM can be fetched quickly by strategy 1. The access time of the ROM determines the delay time of the CS path, and thus the C stage. The second part is a set of microinstructions located in RAMs outside the chip. These microinstructions are typically used for error routines or interruption-handling routines. As they are not used frequently, the total penalty is very small, even if it takes more than five cycles to get a microinstruction from them. In this way, a CS with high frequency and large capacity is realized at the same time.

### 5.4 Memory Address Pipeline and First Operand Fetch

It takes a long time to access cache memory because it cannot be integrated in the chip. To shorten the delay time of memory access, we introduced a memory address pipeline in which the new memory address can be sent to the cache memory before the data of the previous access are received. To realize this structure, the memory address must have been determined before the IF/OF stage. This is easy for the IF stage, because the latter operates asynchronously to other stages by instruction buffering. However, it is not easy for the or stage, especially when the data are used in the first microinstruction. We call this case (First Operand Fetch) OFO. Since the address calculation is finished early in the B stage, the result is directly sent to the memory address register. Therefore, it can be sent to cache memory a half cycle before the OF stage begins. But how does the microprocessor know the FOF instruction exists,so that it can control the memory address register and cache memory? It does so by decoding the information as to wheter the instruction needs the FOF or not. The information consists of a yes/no answer and the access length. In the case of FOF, the request to a memory access controller is issued not from the C stage, but from the A stage (Fig. 12).

## 6. VLSI Implementation and Chip Specifications

### Chip Size Constraint

The microprocessor consists of macrocells, I/O drivers, and random logic parts. The size of the macrocells and I/O drivers are determined by the function and system structure of the microprocessor. As the total chip area is restricted, the random logic part must be in the remaining area. Since the ratio of the area increase of BiCMOS/CMOS basic cells is not negligible, the number of BiCMOS basic cells in the chip should be limited. On the other hand, their effect tends to reach saturation point. Figure 13 shows the delay time of a sample path consisting of five basic cells in the random logic part. If two of them with a large load capacitance are replaced by BiCMOS basic cells, the delay time is reduced to 84%. But even if all of them are replaced, the delay time of the path is almost the same. Since the basic cells that belong to the critical paths are only a part of the random logic, the number of BiCMOS basic cells can be reduced by using them optimally.

Experience shows that The number of basic cells used in a microprocessor with our architecture should be over six thousand. The area and number have a tradeoff relationship, as shown in Fig. 14. The proportion of BiCMOS basic cells is less than 20% of the total number of cells. This indicates that BiCMOS basic cells

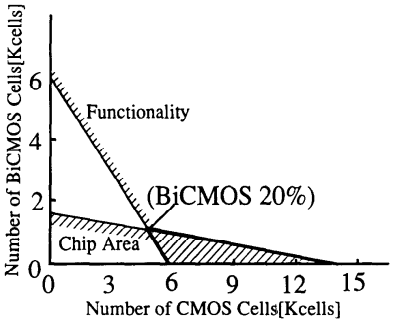



Fig. 14 Usage of BiCMOS Basic Cells.

Table 1. Chip Specifications.

| Item | Value |
|---|---|
| Operating Frequency | 70 MHz |
| Peak Performance | 70 MIPS |
| Power Consumption | 2.1 W(40 MHz) |
| LSI Technology | 1.0 $\mu$m Hi-BiCMOS |
| Die Size | 12.98 mm × 12.98 mm |
| Number of Transistors | 529 K Tr(Bipolar 8 K Tr) |
| Number of Basic Cells | 5.99 K Cells(BiCMOS 18%) |

must be used only to drive a line with large load capacitance caused by an inter-block signal or many fanouts.

### 6.2 Chip Specifications

Table 1 lists the specifications of the BiCMOS microprocessor. The chip size is 12.98 mm and has 529 K transistors integrated in it. The power dissipation is about 2.1 W at 40 MHz. Since the ratio of bipolar transistors used us only 1.5%, it has characteristics of high density, high speed, and low power dissipation. In other words, the performance depends on the 1.5% of the transistor logic, and we were able to improve it using BiCMOS technology.

Figure 15 shows a photograph of the microprocessor chip. It integrates a ROM, a TLB, and a 32-bit data structure that includes the RF, and ALUs.

## 7. Conclusions

We have developed the world's first BiCMOS 32-bit microprocessor. Three strategies were adopted as a design philosophy for accelerating operating frequency. The first was the effective use of bipolar transistors to reduce the delay time of logic circuits. We introduced BiCMOS macrocells that operate twice as fast as CMOS ones. The second strategy was the reduction of the number of inter-chip communications to accelerate memory access. The microprocessor has a cache memory system structure using integrated TLB and standard SRAMs. The third one was eliminating functions

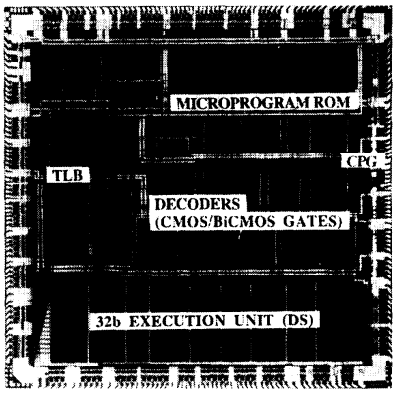



Fig. 15 Photograph of the BiCMOS Microprocessor.

that may be executed slowly from critical paths, to speed up the control logic. To implement the microprocessor in a single VLSI, we reduced the number of BiCMOS basic cells.

The operating frequency of 70 MHz is about 1.5 to 2 times higher than that of ordinary CMOS microprocessors, and the integration density of 529 K transistors per 12.98 mm 2 chip and the power dissipation of 2.1 W are almost the same as those of the CMOS. We believe that all high-performance microprocessors will be made from the BiCMOS VLSI in the near future.

**References**

1. NISHIO, Y. et al. 0.45 ns 7 k Hi-BiCMOS gate array with configurable 3-port 4.6 k SRAM, IEEE Custom Integrated Circuits Conference, May (1987), 203-204.
2. KATO, Y. et al. A 16 ns 16 K bipolar RAM, 1983 *IEEE International Solid State Circuits Conference*, 106-107.
3. NISHIO, Y. et al. A subnanosecond low power advanced bipolar-CMOS gate array, *in proc. ICCD*, Oct. (1984), 428-433.
4. HOTTA, T. et al. CMOS/bipolar circuits for 60 MHz digital processing, *IEEE J. Solid-State Circuits*, SC-**21** (Oct. 1986), 808-813 also in ISSCC Dig. Tech. Papers, Feb. (1986), 190-191.
5. HOTTA, T. et al. Hi-BiCMOS 32-bit execuiton unit, *Trans. Inst. Electron. Inform. Commun. Eng.*, J**70**-C, 4 (Apr. 1987) 469-478.
6. KURITA, K. et al. Very high speed ROM using bipolar/MCMOS technology, *Trans. Inst. Electron. Inform. Commun. Eng.*, J**70**-C, 6 (June 1987).
7. HOTTA, T. et al. 1.3 $\mu$m CMOS/bipolar Macrocell Library for the VLSI Computer, *IEEE J. Solid-State Circuits*, SC-**23** 2 (Apr. 1988) (500-506).
8. AGARWAL, A. et al. An analytical cache model, *ACM Transactions on Computer Systems*, **7**, 2 (May 1989), 184-215.
9. HOTTA, T. et al. A 70 MHz 32b microprocessor with 1.0 $\mu$m BiCMOS macrocell library, *IEEE International Solid State Circuits Conference*, THAM 9.7 (Feb. 1989) 124-125.
10. TANAKA, S. et al. A BiCMOS 32-bit execution unit for a 70 MHz VLSI computer, *IEEE Custom Integrated Circuits Conference* (May 1989), 10.8-10.4.
11. NAKATSUKA, Y. et al. A high performance BiCMOS 32-bit microprocessor, *IEEE International Conference of Computer Design*, to be published in October 1989.

(Received September 18, 1989)