ICOT Technical Memorandum: TM-0856

TM-0856

An Intelligent Cache memory for Inference Machine

by
K. Furutani, K. Yasuda (Mitsubishi)

February, 1989

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03) 456-3191~5
Telex ICOT J32964

Institute for New Generation Computer Technology

An Intelligent Cache memory for Inference Machine

K.Furutani, K.Yasuda, A.Maeda, H.Nakashima* and Y.Takeda*

LSI Research and Development Lab., Mitsubishi Electric Corp.
4-1 Mizuhara, Itami 664 Japan
*Information Systems & Electronics Development Lab.,Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura 247 Japan

1. Introduction

In the Japanese Fifth Generation Computer Project, parallel inference machine (PIM) systems are being developed. For processor elements of PIM/m we have been developed a VLSI intelligent cache memory suitable for the execution of logic programming languages. This device includes 1K words instruction cache memory, 1K tags for accessing the off-chip 4K word data cache memory, and DRAM controller. This cache chip, processor chip, network control chip and memory devices realize compact processor element as shown in Fig.1.

In this report, the novel architectures in device configuration, performance, testability and cell-based design are described.

2. Device configuration

Fig.2 shows the block diagram of the cache chip. The device features are summarized in Table I. The instruction and data cache memories employ the physical address caching scheme with on-chip TLB's[1]. The unique features of this device are described in the following.

2.1. Embedded program counter

Recently the Harvard architecture is gaining ground as a style of high-end microprocessors[2]. It is suitable for the processor which pipelines the decode and execution of the instructions because the processor unit can fetch the instruction and operand simultaneously. However the requirement of dedicated address and data buses for both instruction and operand accesses makes its implementation difficult when the processing unit and the cache unit are separated on different chips because the increased number of pin count affects the chip size and packaging costs.

To solve this problem, the cache chip has a program counter that is the copy of the program counter in the processor chip as shown in Fig.2. The program counter is set via the instruction data bus and incremented after the fetch operation. When there are no branch operations, the processor unit need not feed the instruction address. Thus the address and data signals for the instruction cache can share the bus to reduce the pin count by 32 with little penalty.

2.2. Support for logical Inference

In the execution of logic programming languages, a variable assignment by unification should be reset when the inference falls into contradiction. The trail buffer memory in Fig.2 supports this process. When the processor unit unifies a variable, the address of the variable is pushed into the trail buffer memory. When the processor unit invalidates the unification, the trail buffer memory pops the address of the variable, and the data tag which stands for the un-unified state is written into the variable. Since the unification is one of the most important processes in the logic program execution, its hardware support is effective.

3. Device performance

The proper choice of the TLB and cache memory configuration is important in order to reduce the hardware while keeping high hit-ratio. We justified the configuration of this chip by simulation. The device behavior during the execution of the Prolog compiler is simulated. Since the Prolog compiler is a large and practical program, the simulated result seems to be a good measure of the device performance.

Fig.3 and Fig.4 show the simulated hit-ratio versus the TLB size and cache size, respectively. The hit-ratio of 99.83 % and 99.66 % are obtained by the 32-entry 2-way set associative TLB's for the instruction and operand address translation, respectively. The 94.4 % and 99.2 % hit-ratio are obtained by the 1K words instruction and 4K words operand direct mapped cache memories, respectively. Therefore this cache chip can successfully compensate the speed gap between the processor chip and DRAM's.

As to the cache coherency, the write back method executes the main memory access in only 15 % of miss-hit write operations according to the simulation, while the write through method, in general, requires the main memory access in every write operation. Since this chip realizes high hit-ratio, the less chance of the main memory access pays for the hardware of the write back method.

The worst condition (Vcc=4.5V,Ta=75°C) access times are estimated by simulation to be 50 nsec from clock to instruction, 56 nsec from clock to instruction cache ready signal and 22 nsec from clock to data cache ready signal, realizing 16.7MHz system clock for the processor element shown in Fig.1.

4. Testability

The control circuits are tested using the scan test method. Its pattern is automatically generated by the in-house software "MULTES". There is a single scan path with 420 scan resisters. This method increases chip area by only 2 %. As to the test of 80-Kbit on-chip memory cells, the usual scan test requires lengthy serial test pattern. Thus we employed the special commands to access every RAM through 40-bit data bus in order to reduce the test time. With the special commands, the test time is reduced to 12 % of that with the conventional scan test.

5. Layout data preparation

This device is laid out by software with automatically generated RAM's, PLA's, and standard cells, integrating 610K transistors in 14.47 X 14.84 mm² device area as shown in Fig.5.

In the layout of large chips, the control of signal propagation delay is important. The chip is laid out hierarchically to minimize the wire length. However it is inevitable that a small number of signals are routed with unexpectedly long paths. Re-execution of automatic routing takes time and often ends in failure. Therefore we introduced the adjustable standard cell. At first, the fan out number decides the drive capacity of standard cells. Standard cells with different drive capacity have different width with the same height. After the automatic place and route, the actual delay is estimated from the parasitic capacitance extracted from the layout data. The standard cell which drives a signal line with unexpected delay is replaced by a greater drive capacity cell with same width and different height as shown in Fig.7. Due to the double layer aluminium process, the replacement is carried out without changing the original routing.

6. Conclusion

The instruction/data cache and DRAM controller are integrated on a chip to compose high performance inference machines. The embedded program counter reduces pin count and the trail buffer memory facilitates execution of logic programming language. The cache configuration is optimized by simulation so as to achieve high hit-ratio and realize 1.28M LIPS (logical inference per second) for inference machine.

References
[1] J.Cho et al.,"A 40K Cache Memory and Memory Management Unit," ISSCC, pp.50-51, 1986.
[2] Les Kohn et al., "Introducing the Intel i860 64-Bit Microprocessor," IEEE Micro, vol.9, no.4, pp.15-30, 1989.

Table. I Features

Instruction Cache	
TLB	: 32 entry, 2 way set associative
Tag	: 22 bit X 256 entry, direct mapping
Cache	: 40 bit X 1024 word
Data Cache	
TLB	: 32 entry, 2 way set associative
Tag	: 22 bit X 1024 entry, direct mapping
Cache (Off-chip)	: 40 bit X 4096 word
Main memory interface	
Write back for data cache coherency	
SEC-DED Hamming code ECC	
4 word block read/write for line replacement	
DRAM refresh control	
Access time (Vcc=4.5V, Ta=75°C)	
Instruction cache ready from clock	56 ns
Data cache ready from clock	22 ns
Instruction cache data from clock	56 ns
Supply voltage	5 volt
Power	2.5 watt
Package	361 pin PGA
Chip size	14.47 X 14.84 mm²
Transistor Count	610 K
Technology	Psub Twin well 1 μm CMOS single poly-Si & double Al
Memory cell	CMOS 6T SRAM Cell

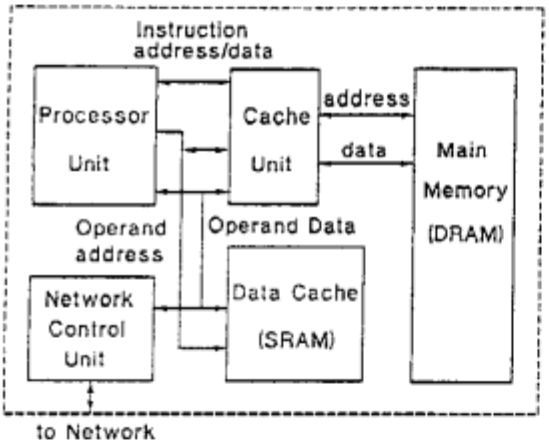

Fig.1 Configuration of the processor element.

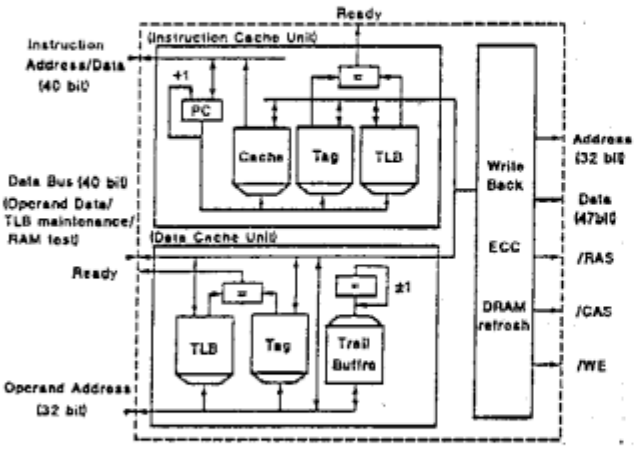

Fig.2 Block diagram of the cache chip.

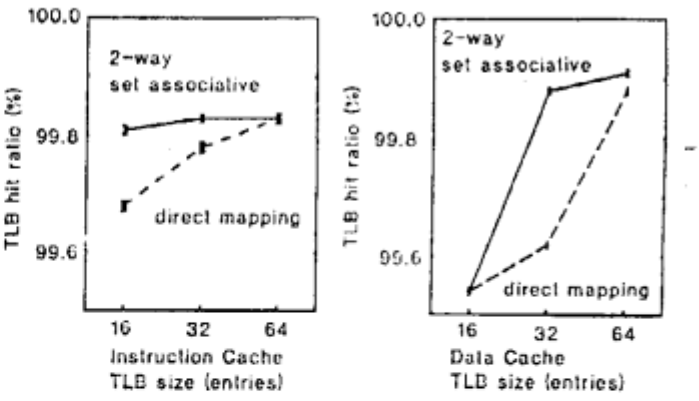

Fig.3 Estimation of TLB hit ratio.

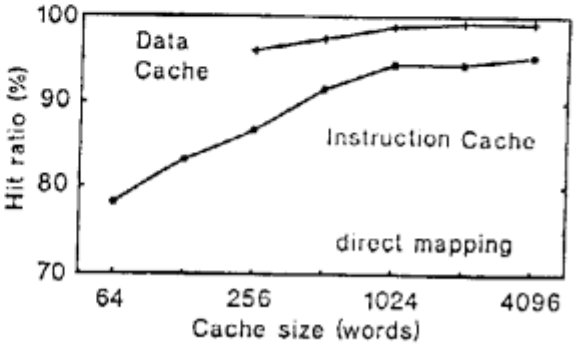

Fig.4 Estimation of Cache hit ratio.

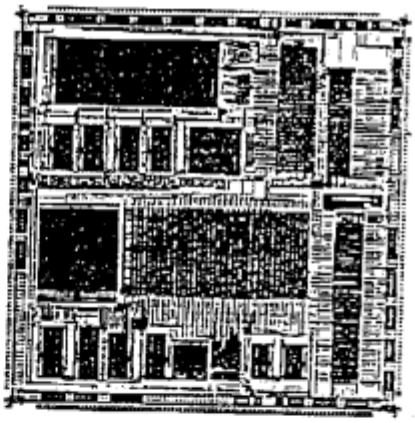

Fig.5 Photomicrograph of the cache chip.

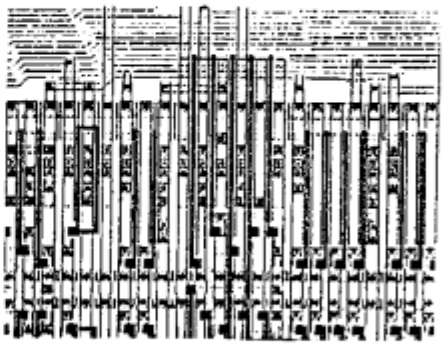

Fig.6 Layout data with adjustable standard cell.