TR-634

A 1.5 MLIPS 40-Bit AI Processor

by

H. Machida, H. Ando, C. Ikenaga, H. Nakashima,
A. Maeda & M. Nakaya (Mitsubishi)

March, 1991

ICOT

Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan

(03)3456-3191∼5
Telex ICOT J32964

Institute for New Generation Computer Technology

A 1.5 MLIPS 40-BIT AI PROCESSOR

Hirohisa MACHIDA, Hideki ANDO, Chikako IKENAGA, Hiroshi NAKASHIMA,*
Atsushi MAEDA, and Masao NAKAYA

Mitsubishi Electric Corporation, LSI R & D Laboratory
4-1 Mizuhara, Itami, Hyogo, 664 Japan

**Information Systems and Electronics Development Laboratory, Ofuna, Japan*

Abstract

A high performance 40-bit AI processor with a capability of 1.5 MLIPS (Mega Logical Inference Per Second) in *append* has been developed. The performance of this processor is achieved by the combination of novel architectures of pipelined *data typing* and *dereference*, a 0.8-µm CMOS technology, and a clock scheme.

1. INTRODUCTION

Application specific processors for symbolic processing languages such as Lisp and Prolog played a great part in research and development of AI. It is not suitable for general purpose processors to execute Lisp and Prolog, because additional specific operations are necessary for these respective languages. These operations are *data typing* which means checking the data type by the tag, *dereference* which involves searching the reference data, and so on. General processors are not good at these operations. Hence several AI processors such as Ivory [1], Pegasus [2], and PSI-II [3] have been developed.

To attain higher performance, it is effective to make a lot of processors operate in parallel and to reduce time per machine cycle *(Tm)* of the processors. We have, therefore, developed a 40-bit AI processor called PU. The PU chip is a key component of the PIM/m (Parallel Inference Machine) and AI workstation. PIM/m is being developed in the Japanese Fifth Generation Computer Systems Project [4]. The PU chip is able to execute two different type logic programming languages, KL1 [5] for PIM/m and ESP [6] for the AI workstation. KL1 is a parallel logic programming language and ESP is an object oriented logic programming language.

For *data typing* and *dereference*, the PU chip has powerful mechanisms to manipulate tagged data. Especially, the pipelined *data typing* and *dereference* are the most unique features of the PU chip and they contribute to reducing *Tm* and execution steps. The PU chip has been fabricated in a 0.8-µm CMOS technology and this technology also contributes to reducing *Tm*. When the PU chip is integrated into the AI workstation, the performance of a 1.5 MLIPS in *append* is achieved.

This paper describes the novel architectures of pipelined *data typing* and *dereference*, a 0.8-µm CMOS technology, and a clock scheme for achieving the high performance of the PU chip.

2. HARDWARE ARCHITECTURE

The PU chip is a 40-bit pipelined microprocessor under the control of a microprogram stored in a 32K-word WCS [7]. The hardware architecture is described in this section.

2.1 Tag Architecture

The PU chip supports fast execution of the logic programming languages through a tagged 40-bit architecture. A 40-bit data includes an 8-bit tag which indicates a data type. In general, the logic programming languages have no data type declaration of variables. But operation varies with data types such as integer, character, and pointer. The tag is, therefore, required besides an operand value, and *data typing* to check the data type by the tag is an important operation.

There is also a specific and frequent operation called *dereference* in the logic programming languages. *Dereference* means that a CPU examines a chain of data which has a tag of reference pointer and processes the last data having a non-reference pointer as shown in Figure 1. *Data typing* and *dereference* are very important for an efficient implementation of the logic programming languages.

2.2 Pipeline Organization

The PU chip consists of five pipeline stages which are an instruction decode (D-stage), an operand address calculation (A-stage), an operand data read (R-stage), an operand data set up(S-stage), and an execution (E-stage) as shown in Figure 2.

There is a RAM table (OPT) for instruction decode in the D-stage. Each entry of the table contains a start address of a microprogram routine and a nano-code to control the following stages. This RAM decoder makes it easy to develop the microprogram and makes it possible to execute several languages such as ESP, KL1, and Lisp.

In the A-stage, an operand address is calculated from two of following resources according to the nano-code. They are:
(1)an operand field of the instruction,
(2)a program counter,
(3)a register file (ARF) of the A-stage, and
(4)two address registers.
ARF is a copy of a register file (RF) of the E-stage. The A-stage is also used to control instruction fetch, including conditional and unconditional branch operation.

An operand is fetched from a data cache in the R-stage, if necessary.

The S-stage is used to select operands from the following resources and to transfer them to the E-stage according to the nano-code. They are:
(1)an operand field of instruction,
(2)the operand fetched by the R-stage and its address,
(3)a register file (RF), and
(4)a working register (WR).
The S-stage is an additional special stage of the PU chip. This stage is required for the pipelined *data typing* and *dereference*, as discussed later.

There are two pipelined phases controlled by microprograms in the E-stage. The first phase contains RF, WR, and special registers. This phase is shared by the S-stage and the E-stage for the operand set up. The second phase has two temporary registers, two memory address registers, and two memory data registers. Two operands of those registers are computed by ALU, and the result is written into registers in the first and/or second phase.

2.3 Pipelined *Data Typing* and *Dereference*

Both *data typing* and *dereference* are performed by checking the tag of data and changing the control flow according to the result. The PU chip has powerful mechanisms, including the pipelined *data typing* and *dereference*, for these operations. The pipelined *data typing* and *dereference*, which are the most unique features, mainly depend on the S-stage.

There are the following three functions for *data typing* in the S-stage. They are:
(1)modifying the microprogram entry address by comparing the tag of the operand fetched by the R-stage with an immediate value,
(2)setting up the offset of a multi-way jump by the tag of the operand which was fetched by the R-stage, and
(3)setting up the two-way jump condition by comparing the tag of an operand transferred to the E-stage with an immediate value.
The first two functions require the special stage between the R-stage and the E-stage.

The S-stage also performs *dereference*. There are two cases for *dereference*; *case-1* is *dereference* from data of RF and *case-2* of the memory. In the R-stage, the operand is fetched if the data of RF contains reference pointer for *case-1*, and always fetched for *case-2*. In both cases, the tag of the data is examined and the operand is fetched repeatedly from the memory in the S-stage until non-reference data is obtained.

3. LAYOUT DESIGN FEATURES

The PU chip photomicrograph is shown in Figure 3 and the device features are summarized in Table 1. A cell-based design method is adopted to reduce the layout design time. The chip consists of fifty kinds of standard cells and macro cells such as RAMs and PLAs. Their macro cells are generated by the in-house module generator. The layout design and a verification are completed for only two weeks. But the clocking scheme is a problem in large-die and high performance VLSIs which are designed by using cell-based design method.

3.1 Problem of Clock Distribution

A hierarchical and tree clock distribution method is often useful for automatic layout design [8][9]. But it is necessary to control load balance for each buffer, control line length, and adjust the buffer size after the layout on the hierarchical clock distribution method. Their adjustments are difficult for a commercial automatic place and route program. When 12,000 standard cells are routed automatically, a clock skew is increased between any two registers' active clock edges. We, therefore, adopt the following methods to solve the above problem.

3.2 Clock Strategy

In order to reduce clock skew and delay, big clock buffers of two steps are utilized and clock distribution lines are placed as shown in Figure 4. Two phase non-overlapping clocks control many gates of flip-flops. These two phase clocks must, therefore, drive heavy loads. Several other phase clocks are necessary for.the control of RAMs and PLAs. These clocks drive light loads. The first buffer drives the second buffers, light loads, and dummy loads. The second buffers drive heavy loads as shown in Figure 5. A difference of delays between the light loads and heavy loads is minimized by inserting dummy loads.

In order to reduce the clock skew of heavy loads, the outputs of the second buffers are connected in horizontal channels of a standard cell area. The width of the vertical lines is wide to reduce the resistance, that is 10 μm. And that of the horizontal lines is 1.4 μm which is the minimum width to reduce load capacitance. The channel width of the second buffers are 3200 μm for p-ch transistor, 1600 μm for n-ch transistor.

4. 0.8-μm PROCESS TECHNOLOGY

The PU chip has been integrated in a 0.8-μm CMOS technology with single-polysilicon and double-metal. The 16.3-mm x 13.6-mm PU chip contains 384,000 transistors and is packaged in a 361-pin PGA.

The main features of this technology are summarized in Table 2. Gate length is 0.8μm for n-ch and p-ch transistor. A "gate/n- overlapped LDD" [10] structure is used in n-ch MOS transistor. This structure is formed using the rotational oblique n- ion implantation. Conventional LDD structures are used in p ch MOS transistor which improves current drivability. Sufficient punch-through tolerance is achieved by optimized doping profile for channel region. For the advanced structure with multi-level interconnections, TEOS (Tetra-Ethyl-Ortho-Silicate) and O_3 (Ozone) are used as an inter-layer dielectric.

5. PRACTICAL RESULTS AND APPLICATIONS

5.1 Performance

The realizations of 1-nsec clock skew and 2-nsec clock delay are measured by an EB tester. Figure 6 (a) shows the skew and delay of heavy loads, and (b) shows the delay of light loads. These skews and delays result in a 30-MHz chip operation at room temperature with a 5-V power supply. Figure 7 shows a typical schmoo plot of time per machine cycle *(Tm)* versus a supply voltage.

The PU chip has been integrated into an AI workstation and the capability of 1.5 MLIPS in *append*, which is quite typical in the logic programming languages, is achieved at a 16.7-MHz operation (worst condition). The typical power consumption is 2.5 W with a 5-V power supply at a 16.7-MHz operation.

The performance of the PU chip has been improved by a factor of 3.78 as compared with the previous processor, PSI-II [3], in the ESP execution. Details of the factor are shown in Table 3. The factor of 1.67 is due to the 0.8-μm double-metal CMOS technology, and the factor of 2.27 is due to the novel architecture. The novel architecture reduces both *Tm* and execution steps. The improvement ratios of performance are

1.67 for *Tm* and 1.36 for the execution steps.

5.2 Parallel Inference Machine

The PU chip is also a key component of PIM/m (Parallel Inference Machine). Up to 256 processor elements are connected to form a two-dimensional mesh network in PIM/m. The processor element has the processing unit which includes the PU chip, the cache unit, the main memory, the floating point processor, and the network control unit as shown in Figure 8.

PIM/m has four times more processor elements than the multi-PSI which is a multi computer system using PSI-II. And the performance of the processor element will be improved five times more in the KL1 execution. The first PIM/m system will be completed in the first quarter of 1991.

6. CONCLUSION

A high performance 40-bit AI processor has been successfully fabricated. The high performance with a capability of 1.5 MLIPS in *append* is due to the combination of the novel architecture, the 0.8-μm CMOS technology, and the clock scheme. First silicon has been fully functional and has booted the SIMPOS operating system. The AI workstation, PSI-UX, which carries this processor has been completed already and has produced since January in 1991. The AI processor makes it possible to establish a high performance multi computer system in a small size.

ACKNOWLEDGMENT

The authors wish to thank Dr. H. Komiya, Dr. T. Nakano, and Dr. Y. Horiba for their encouragement and support of this research program, and also thank Dr. S. Uchida and researchers in the ICOT for giving us the opportunity to conduct this research.

We also wish to acknowledge all the people in the design, process, and test groups who have contributed to the development of this device.

REFERENCES

[1] C. Baker, D. Chan, J. Cherry, A. Corry, G. Efland, B. Edwards, M. Matson, et al., "The Symbolics Ivory Processor: A 40 Bit Tagged Architecture Lisp Microprocessor," in *Proc. of Intl. Conf. on Computer Design*, pp. 512-515, 1987.

[2] K. Seo and T. Yokota, "Design and Fabrication of Pegasus Prolog Processor," in *Proc. of Intl. Conf. on VLSI*, 1989.

[3] H. Nakashima and K. Nakajima, "Hardware Architecture of The Sequential Inference Machine: PSI-II," in *Proc. of 4th Symp. on Logic Programming*, pp. 104-113, 1987.

[4] S. Uchida, K. Taki, K. Nakajima, A. Goto, and T. Chikayama, "Research and Development of the Parallel Inference System in the Intermediate Stage of the FGCS Project," in *Proc. of Intl. Conf. on Fifth Generation Computer Systems 1988*, pp. 16-36, 1988.

[5] T. Chikayama, H. Sato, and T. Miyazaki, "Overview of the Parallel Inference Machine Operating System(PIMOS)," in *Proc. of Intl. Conf. on Fifth Generation Computer Systems 1988*, pp. 208-229, 1988.

[6] T. Chikayama, "Unique Features of ESP," in *Proc. of Intl. Conf. on Fifth Generation Computer Systems 1984*, pp. 292-298, 1984.

[7] H. Nakashima, Y. Takeda, K. Nakajima, H. Andou, and K. Furutani, "A Pipelined Microprocessor for Logic Programming Languages," in *Proc. of Intl. Conf. on Computer Design*, pp. 355-359, 1990.

[8] T. Tokumaru, E. Masuda, C. Hori, K. Usami, M. Miyata, and J. Iwamura, "Design of A 32bit Microprocessor, TX1," in *Symp. on VLSI Circuits Dig. Tech. Papers*, pp. 33-34, 1988.

[9] S. Boon, S. Butler, R. Byrne, and B. Setering, "High Performance Clock Distribution for CMOS ASICs," in *Proc. of Custom Integrated Circuits Conf.*, pp. 15.4.1-15.4.5, 1989.

[10] M. Inuishi, K. Mitsui, S. Komori, M. Shimizu, H. Oda, J. Mitsuhashi, and K. Tsukamoto, "Optimum Design of Gate/n- Overlapped LDD Transistor," in *Symp. on VLSI Technology Dig. Tech. Papers*, pp. 33-34, 1989.

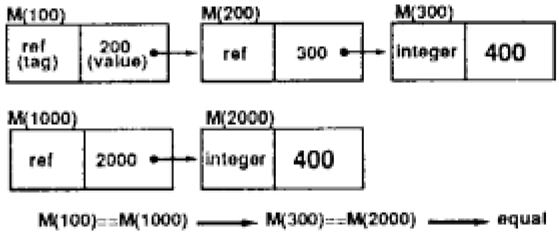


Figure 1 Unification with Dereference

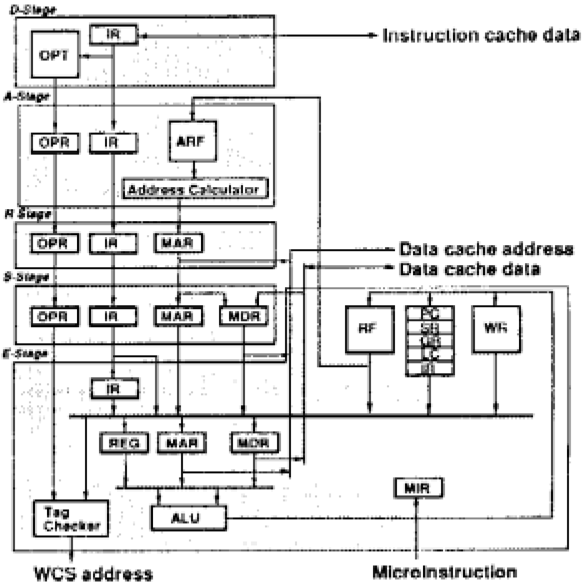


Figure 2 Configuration of AI Processor

Table 1 Chip Features

Chip size	*16.3 mm x 13.6 mm*
Transistor count	*384k*
random logics	*110k*
RAMs	*270k*
PLAs	*4k*
Package	*361-pin PGA*
Data width	*40bit*
ALU	*32bit*
Pipeline	*5 stage*
Frequency (worst condition)	*16.7 MHz*
Power supply	*5 V*
Power dissipation (at 16.7MHz)	*2.5 W*
System performance	*1.51 MLIPS*

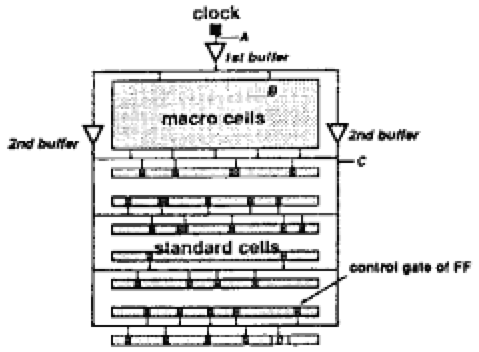


Figure 4 Schematic Diagram of Clock Distribution

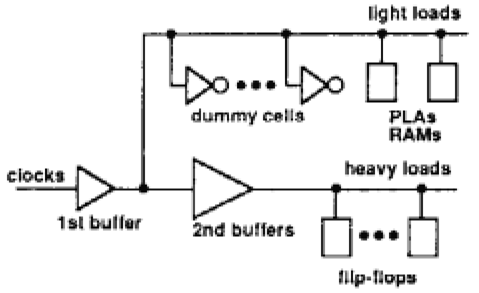
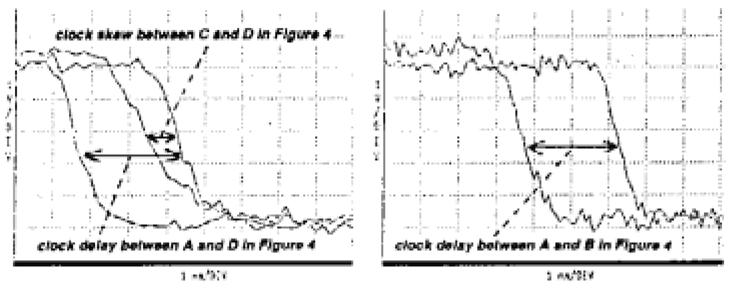


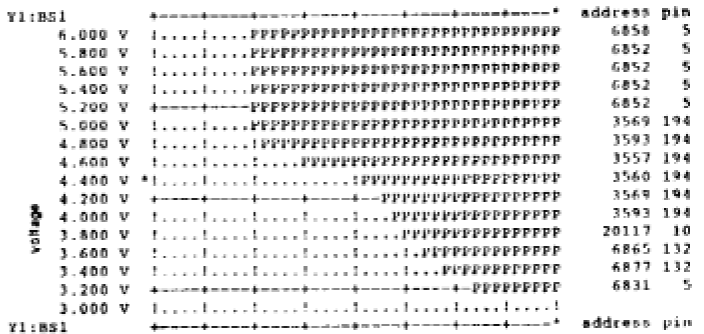
Figure 5 Clock Distribution of PU



(a) clock delay and skew by 1st and 2nd buffer

(b) clock delay by 1st buffer

Figure 6 Clock Waveforms Measured by EB Tester



```
Y1:BS1          +----+----+----+----+----+----+----+----*  address pin
     6.000 V    !....!....PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     6858    5
     5.800 V    !....!....PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     6852    5
     5.600 V    !....!....PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     6852    5
     5.400 V    !....!....PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     6852    5
     5.200 V    +----+----PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     6852    5
     5.000 V    !....!....PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     3569  194
     4.800 V    !....!....!PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP     3593  194
     4.600 V    !....!....!....PPPPPPPPPPPPPPPPPPPPPPPPPPP     3557  194
     4.400 V   *!....!....!..........!PPPPPPPPPPPPPPPPPPPP     3560  194
     4.200 V    +----+----+----+----+--PPPPPPPPPPPPPPPPPPP     3569  194
     4.000 V    !....!....!....!....!...PPPPPPPPPPPPPPPPPP     3593  194
     3.800 V    !....!....!....!....!....PPPPPPPPPPPPPPPPP    20117   10
     3.600 V    !....!....!....!....!....!.PPPPPPPPPPPPPPP     6865  132
     3.400 V    !....!....!....!....!....!...PPPPPPPPPPPPP     6877  132
     3.200 V    +----+----+----+----+----+----+-PPPPPPPPPP     6831    5
     3.000 V    !....!....!....!....!....!....!....!....!
Y1:BS1          +----+----+----+----+----+----+----+----*  address pin
X1:RATE         20.00nS    30.00nS    40.00nS    50.00nS    60.00nS
```

Figure 7 Schmoo Plot (machine cycle vs. supply voltage)

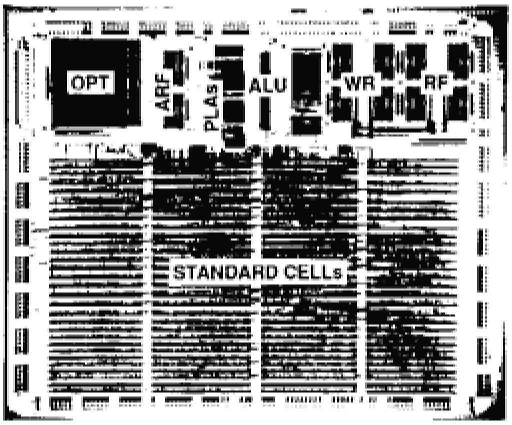


Figure 3 Chip Photomicrograph

Table 2 Technology Features

0.8μm Twin-well CMOS	
gate length (N / P)	*0.8μm*
gate oxide thickness	*18nm*
well depth (P-well)	*4.0μm*
(N-well)	*2.5μm*
diffusion space (P+)	*1.4μm*
(N+)	*1.0μm*
interlevel dielectric thickness	
(under 1st-metal)	*0.7μm*
(1st-metal / 2nd-metal)	*0.8μm*
metal pitch (1st-metal)	*2.4μm*
(2nd-metal)	*3.2μm*
contact hole size	*0.8μmx0.8μm*
via hole size	*0.8μmx1.0μm*

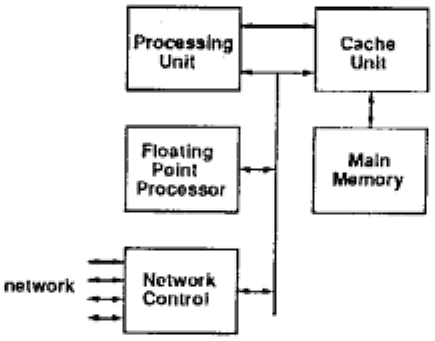


Figure 8 System Configuration of the Processor Element of PIM/m

Table 3 Performance Improvement by PU

	time per machine cycle	append operation in ESP
PSI-I (previous processor)	167nsec	15steps
0.8-μm PSI-II	100nsec	15steps
PU	60nsec	11steps