TR-0799

A VLSI Chip Set for a Large Scale Parallel Inference Machine: PIM/m

by
H. Machida, H. Ando, K. Yasuda, K. Furutani,
Y. Yamashita, Y. Takeda, K. Nakajima,
M. Sakao, M. Makaya (Mitsubishi)
& H. Nakashima (Kyoto Univ.)

August, 1992

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 VLSI Chip Set for a Large Scale Parallel Inference Machine: PIM/m

Hirohisa Machida, Hideki Ando, Kenichi Yasuda, Kiyohiro Furutani, Yukihiro Yamashita,
Hiroshi Nakashima*, Yasutaka Takeda*, Katsuto Nakajima*,
Masayoshi Sakao**, and Masao Nakaya

Return Address
Hirohisa MACHIDA
LSI Laboratory, Mitsubishi Electric Corporation
4-1 Mizuhara, Itami, 664 Japan
Phone : +81-727-84-7347 Fax: +81-727-84-7439

*Computer & Information Systems Laboratory, Kamakura, Japan
**Computer Works, Kamakura, Japan

Abstract

This paper presents three VLSI chips which are a processor (PU) chip, a cache memory (CU) chip, and a network control (NU) chip for a large scale parallel inference machine. The PU chip has been designed to be adapted to logic programming languages. The CU chip implements a hardware support called "Trail Buffer" which is suitable for the execution of the Prolog-like languages. The NU chip makes it possible to connect 256 processing elements in a mesh network. The parallel inference machine (PIM/m) runs a Prolog-like network-based operating system called PIMOS as well as many applications and has a peak performance of 128 MLIPS(Mega Logical Inference per Second). The PU chip containing 384,000 transistors is fabricated in a 0.8-μm double-metal CMOS technology. The CU chip and the NU chip contain 610,000 and 329,000 transistors, respectively. They are fabricated in a 1.0-μm double-metal CMOS technology. A cell-based design method is used to reduce the layout design time.

I. INTRODUCTION

A sophisticated machine has been required for a massive knowledge information processing. It is not suitable for general purpose computers to make an inference from huge amounts of data. In order to solve this problem, expert systems which execute logic programming languages in high speed have been developed. To attain higher performance, it is effective to make a lot of processing elements operate in parallel and to reduce the machine cycle time of processing element. A VLSI chip is the key to realize these requirements.

A large scale Parallel Inference Machine (PIM/m) has been developed for a massive knowledge information processing in the Japanese Fifth Generation Computer Systems Project [1]. The maximum configuration of PIM/m is shown in Figure 1. Up to 256 processing elements are connected to form a two dimensional mesh network. The processing element (PE) consists of the processing unit, the cache unit, the local memory, the floating point processor, and the network control unit as shown in Figure 1. We have developed a processor (PU) chip, a cache memory (CU) chip, and a network control (NU) chip required for the key devices of PIM/m [2].

This paper describes characteristics of a VLSI chip set for a large scale parallel inference machine (PIM/m). In the next section , we introduce the architecture and layout design of the PU chip [3]. In section III, the CU chip is explained in brief as the detail description is given in [4]. And we present the NU chip which is important in a loosely coupled highly parallel computer system in section IV. In section V, the performances of the AI workstation and the parallel inference machine are shown. Finally, some conclusions are drawn.

II. PROCESSOR CHIP

The PU chip is key component of PIM/m and an AI workstation and it is a 5-stage 40-bit pipelined microprocessor under the control of a microprogram stored in a 32K-word WCS [3][5]. 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 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 process the last data having a non-reference pointer as shown Figure 2. Based on analysis [6] of the logic programming language execution mechanism, data typing and dereference provide an efficient way to perform operation.

A Hardware Architecture

The PU chip has capability to execute two different type logic programming languages, KL1[7] for PIM/m and ESP[8] for the AI workstation. KL1 is a parallel logic programming language and is very powerful to represent parallel process communicating. ESP is the language in which Prolog and object oriented language features are combined.

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), and operand data set up(S-stage), and execution (E-stage) as shown in Figure 3. 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 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 operand from 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 two pipelined phases controled by microprograms in the E-stage. The first stage contains RF, WR, and special registers. This phase is shared by 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 operand of those registers are computed by ALU , and the result is written into registers in the first and/or second phase.

B. 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 mainly depend on the S-stage.

There are the following three functions for data typing in the S-stage. They are :

(1) modifying microprogram entry address by comparing tag of operand which was 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 operand transferred to the E-stage with an immediate value.

The first two functions required 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 for reference pointer for case-1, and always for case-2. In both cases, the tag of the data examineed and the operand is fetched repeatedly from the memory in the S-stage until non-reference data is obtained.

C. Layout design features

The PU chip has been fabricated in a 0.8-μm CMOS technology. A cell-based design method is adopted to reduce the layout design time. The PU chip consists of fifty kinds of standard cells and macro cells such as RAMs and PLAs. Their macro cells are generated by in-house module generator. The layout design and 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.

A hierarchical and tree clock distribution method is often useful for automatic layout design [9][10]. 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 and they cause design time increase. Owing to design cost and chip area reduction, we approached non-hierarchical design. When 12,000 standard cells are routed automatically, a clock skew is increased between any two registers' active clock edge. We, therefore, adopt the following methods to solve the above problem.

In order to reduce clock skew and delay, large clock buffers of two stages 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. The first buffer drives the second buffers, and the second buffers drive the heavy loads. In order to reduce clock skew , the outputs of the second buffers are connected in horizontal channels of a standard cell area. The width of vertical lines is wide to reduce the resistance, that is 10 μm . And that of the horizontal lines is 1.4μm 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.

The realization of 1-nsec clock skew and 2-nsec clock delay are measured by an EB tester. Figure 5 shows the skew and delay time of heavy loads. These skew and delay result in a 33-MHz chip operation at room temperature with 5-V power supply.

III. Cache Memory Chip

The CU chip is a VLSI intelligent cache memory which is also key component as the PIM/m and the AI workstation. This CU chip contains 610 K transistors including 80 K bits memory cells. The chip measures 14.47 mm x 14.84 mm and fabricated by using 1.0μm CMOS double-metal technology. The features of this device were descrived in [4].

A. Hardware

Figure 6 shows the block diagram of the CU chip, which contains of an instruction cache block, a data cache block, and a main memory interface block as a DRAM controller. This chip includes the "Trail Buffer" with 16-word x 32-bit RAM which accelerates the execution of Prolog-like languages. The detailed cache features are listed in Table 1. To prevent an increase in the number of pins, we embedded a program counter and common bus for the instruction cache block shared by the address and data. In the execution of the Prolog-like language, a variable assignment by unification should be reset when the inference falls into contradiction. In this process, the trail buffer is useful for reseting data quickly.

The instruction and data cache blocks employ the physical address caching scheme with on-chip translation lookaside buffers (TLB's) for high hit-ratio. TLB replacement method is a least recently used algorithm. We have determined the cache memory size by practical simulation taking the relationship between the chip size and hit-ratio of the cache memory into configuration. Hit ratios of 99.83% and 99.88 % are obtained by the 32-entry, 2-way set associative TLB's for the instruction and data address translations. And hit-ratios of 94.4% and 99.2 % are obtained with the cache memories, respectively.

There are total of 80 Kbits memory cells in this device. Normally a scan test requires a lengthy serial test pattern, however, in order to reduce the test time, we employed special commands to access every RAM's cell, through the 40-bit data bus shown in Figure 6. With some special commands, the test time was reduced to 12% of the conventional scan test. Furthermore, the commands allow us to perform the RAM pause test and voltage bumping test.

B. Support for Logical Inference

To accelerate logical inference, the CU chip has hardware called "Trail Buffer". In Prolog-like languages, there is an operation called "Backtrack". When backtrack operation occurs, a variable assignment by unification should be reset. In this operation, when a processor unifies a variable, the address of the variable should be stored in "Trail Stack". The trail buffer, which can hold up to 16 addresses of assigned variables, is a cache of the trail stack. The functions of the trail buffer are to store the addresses of assigned variables, to supply address data of the variable to be reset, and to remember the number of reset times.

It is estimated by simulation, that the trail buffer improves the speed of the logical inference by 12.1% for quick-sort program.

IV. Network Control Chip

In a loosely coupled highly parallel computer system, it is important to develop network circuits which realize high speed communication and reduce loads of processors[12]. The NU chip has four bidirectional channels to connect adjacent four PEs and two buffers for message packets. The packet transmission and buffering are automatically performed without any interruption of the execution of the PU and the CU chip.

A. Hardware

Figure 7 shows the block diagram of the NU chip. It consists of a network control unit, a trace memory, a system timer, a maintenance control unit, and an external bus control unit. The network control unit has five pairs of channels which are used for four adjacent PEs and its own processor chip. The channels which are connected to interface bus have 1024 x 9-bit read buffer memory (RB) and 1024 x 9-bit write buffer memory (WB). Each of four transmission channels has 64 x 10-bit FIFO buffer memory to reduce the possibility of network choking. Packet data are transferred asynchronously in 10-bit parallel including a parity bit. The external bus control unit has a function which arbitrates among the NU chip, FPP, and SCSI.

B. Switch circuits

The NU chip supports a message-passing communication between PEs. Switch circuits are shown in Figure 8. Four adjacent PEs and its own processor unit are connected by 5 x 4 switch circuits. Each receiving channel has its own path table (PT) to determine the channel to transmit a packet. Channel controller looks up this table using the destination address of the packet, and connects receiving and transmission channels according to the PT data. Packets are, therefore, switched in parallel regardless of the processor operation. The index of the PT is the coordinate of the destination point which is represented by a 1024 x 2 bit map. Each entry of the PT contains the channel number to which packets are transmitted. The network control scheme is summarized in Table 2.

C. Message Handling in a PE

The read/write operation of the RB and WB are executed by microcode. A low level microcoded routine in the PE is responsible for handling message to and from the NU chip. The microcoded routine performs decomposition and composition of message packets. The interrupt signal of informing a message arrival invokes a micro routine to decode it just before starting the next reduction. This timing is suitable for breaking the current execution sequence, because processor has the minimum context at this timing.

The micro routine decodes all the messages arrived before starting a next reduction. Before decoding each message, the micro routine examines the size of the empty space in the RB. If it is below the fixed value, the micro routine moves the messages, instead of decoding, from the RB to a large memory area, called Read Packet Buffer. It allows to take in further incoming messages and avoid traffic disturbance in the network. After the movement of the messages, the micro routine resumes decoding from the Read Packet Buffer.

On sending a message, if there is not enough area to put the whole message in the WB, the micro routine will wait until area becomes available.

V. PERFOEMANCE EVALUATION

A. AI Workstation Performance

The PU and CU chip have been assembled in an AI workstation and the capability of 1.5 MLIPS is achieved at a 16.7-MHz operation (typical condition) in append benchmark program, which is quite typical operation in the logic programming languages .

The performance of the PU chip has been improved by a factor of 3.51 as compared with the previous processor, PSI-II[11], in the ESP operation. Exact factors are shown in Table 3. The factor of 1.55 is due to the 0.8μm double-metal CMOS technology, and the factor of 2.27 is due to the novel architecture. The novel architecture reduces both the machine cycle time and execution steps. The improvement ratios of performance are 1.67 for the machine cycle time and 1.36 for the execution steps.

B. Parallel Machine Performance

These three chips have been successfully constructed. PIM/m which is utilized by these chips has passed system level tests, and has correctly executed the KL1 at 16.7 MHz under typical condition. Table 4 shows the single processor performance of PIM/m for four benchmarks [13]. The

table also includes the performance of proto type parallel machine (multi-PSI/v2) and the ratio of PIM/m and Multi-PSI/v2 to show the effort of architectural improvement and VLSI implementation.

The performance of PIM/m is 4.8 times as high as that of Multi-PSI/v2 in append. The other three benchmarks besides append are search programs with various parallel algorithms and load distribution strategies. Best-path finds out the shortest path between two vertices. Pentomino makes OR-parallel exhaustive search to solve a packing piece puzzle problem. 15-puzzle solves a well-known puzzle problem. Although these programs are not practical, the algorithms and load distribution strategies should be generally adopted to various application programs of parallel processing.

System performance is strongly related with load distribution strategy and communication cost. The speedup factor is shown by dividing execution time for single processor by that for n processors. Figure 9 shows the speedup of PIM/m for best-path and pentomino.

VI. Conclusions

Three VLSI chips, which are the processor chip, the cache memory chip, and the network control chip, for a high performance large scale parallel inference machine have been developed. The processor chip and the cache memory chip are also used in the AI workstation. The processor chip has high logical inference performance, which is achieved by the combination of novel architectures of pipelined data typing and dereference and the clock scheme. In the cache memory chip, the embedded program counter reduces the pin count and Trail Buffer memory facilitates execution of the logic programming languages. The network control chip achieves automatically data transfer without the processor operation.

All device features are summarized in Table 5. All chips operate by a single 5-V power supply. Total power consumption of three chips is about 7W at 16.7-MHz operation. All photomicrographs of the three chips are shown in Figure 10 and a photograph of a 32.5-cm x 31.0-cm PIM/m processing element board is shown in Figure 11. PIM/m keeps the same size and the power

consumption increase of it is only 30% owing to the VLSI chip set, though the number of PE of PIM/m is 4 times as many as that of the multi-PSI/v2.

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 ICOT for giving us the opportunity to conduct this research.

We also greatly indebted to the device processing staff members who have contributed to the development of these devices, and to Mr. C. Tsunetomo and Mr. H. Kakiuchi for the testing.

REFERENCES

[1] Uchida, S., Taki, K., Nakajima, K., Goto, A., and Chikayama, T.,"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.

[2] Machida, H., Ando,H., Ikenaga,C.,Yasuda, K., Furutani, K., , Nakashima, H., Takeda, Y., Nakajima, K., and Nakaya, M., "A VLSI Chip Set for a Large Scale Parallel Inference Machine :PIM/m," in Proc. of the IEEE 1992 Custom Integrated Circuits Conf., pp.30.1.1- 30.1.4, 1991.

[3] Machida, H., Ando,H., Ikenaga,C., Maeda,A., Nakashima, H., and Nakaya, M., "A 1.5 MLIPS 40-bit AI processor," in Proc. of the IEEE 1991 Custom Integrated Circuits Conf., pp.15.3.1-15.3.4, 1991.

[4] Yasuda, K., Furutani, K., Maeda, A., Wakano, S., Nakashima, H., Takeda, Y., and Yamada, M.," An Intelligent Cache Memory Chip suitable for Logical Inference," IEICE Tran. on Electronics, Vol. E 74, No. 11, pp. 3796-3802, 1991.

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

[6] Touati, H. and Despain, A., "An Empirical Study of the Warren Abstract Machine," Proc. of Intl. Symp. on Logic Programming, pp. 114-124, August 1987.

[7] Chikayama, T., Sato, H., and Miyazaki, T., "Overview of the Parallel Inference Machine Operating System(PIMOS)," in Proc. of Intl. Conf. on Fifth Generation Computer Systems 1988, pp.16-36, 1988.

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

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

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

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

[12] Nakajima, K., Inamura, Y., Ichiyoshi, N., Rokusawa, K., and Chikayama, T., " Distributed Implementation of KL1 on the Multi-PSI/v2," in Proc. of the sixth Intl. Conf. on Logic Programming, 1989.

[13] Nakashima, H., Nakajima, K., and Takeda, Y. "Architecture and Implementation of PIM/m," in Proc. of Intl. Conf. on Fifth Generation Computer Systems 1992, pp. 292-298, 1992.

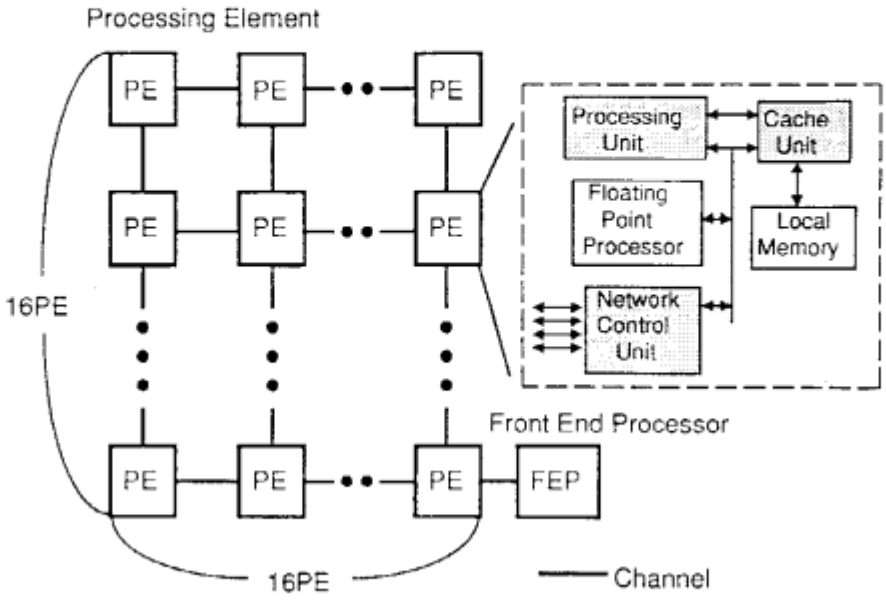


Figure 1 Configuration of PIM/m

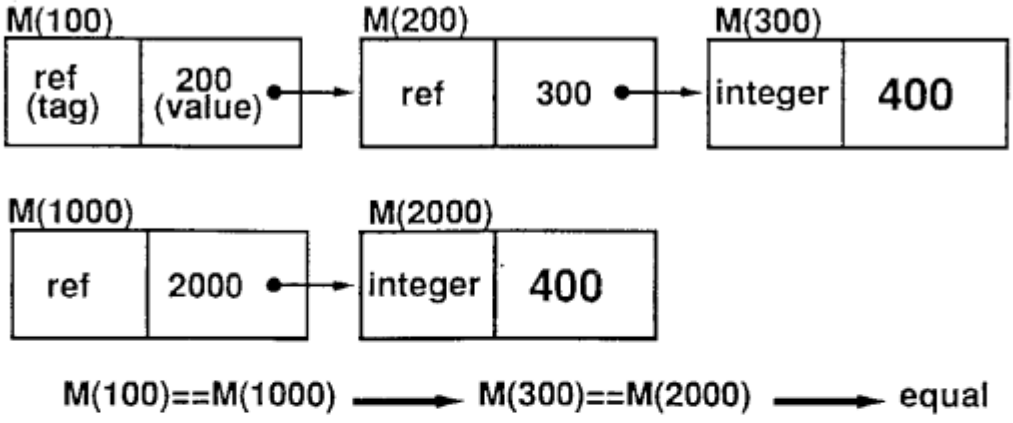


Figure 2 Unification with Dereference

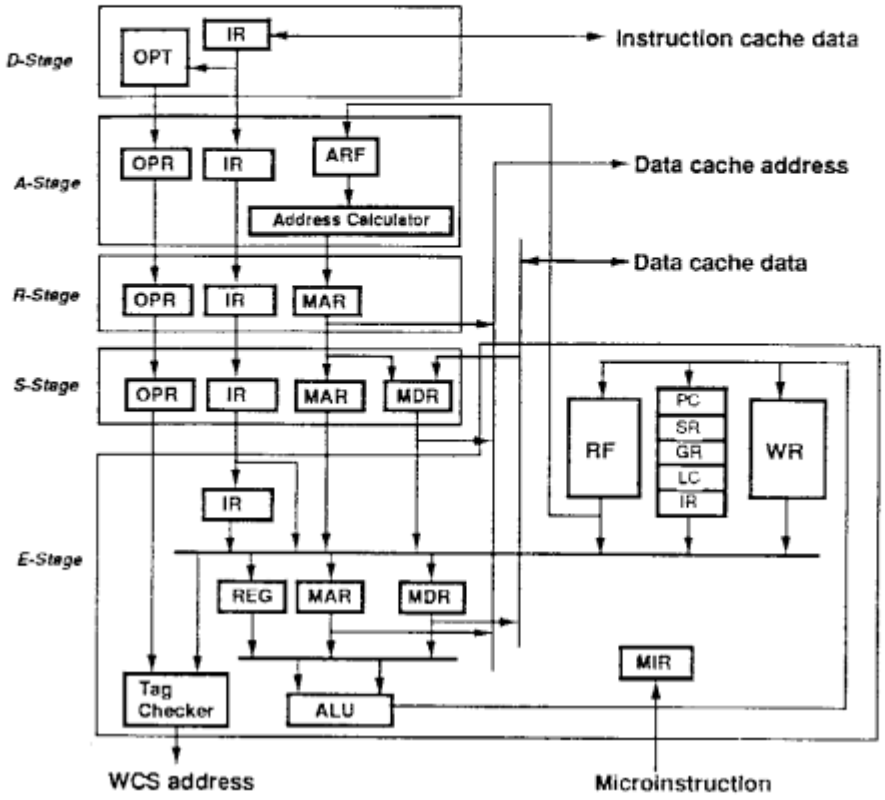


Figure 3 Block Diagram of the PU Chip

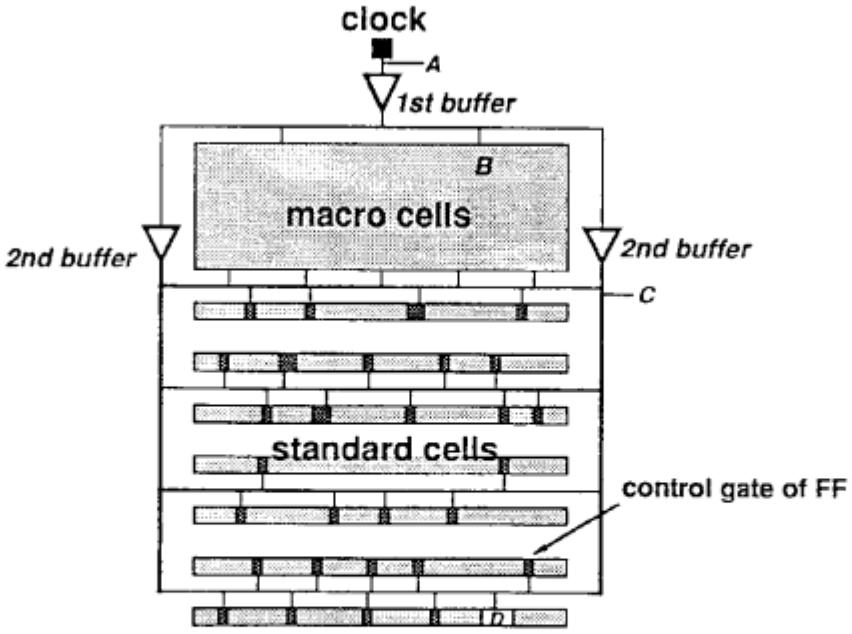


Figure 4 Clock Distribution Scheme of the PU Chip

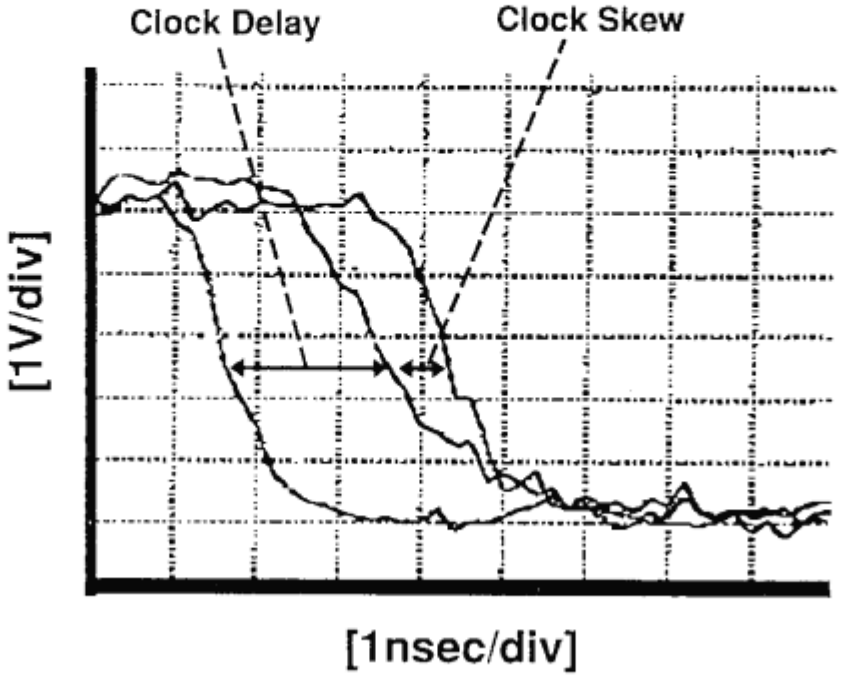


Figure 5 Clock Delay and Skew Waveforms of the PU Chip

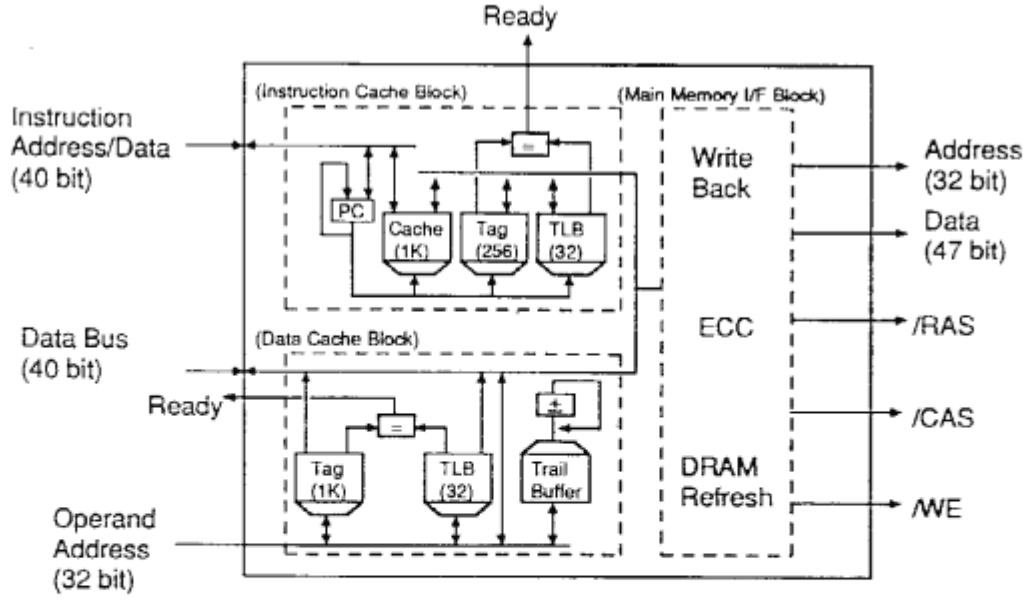


Figure 6 Block Diagram of the CU Chip

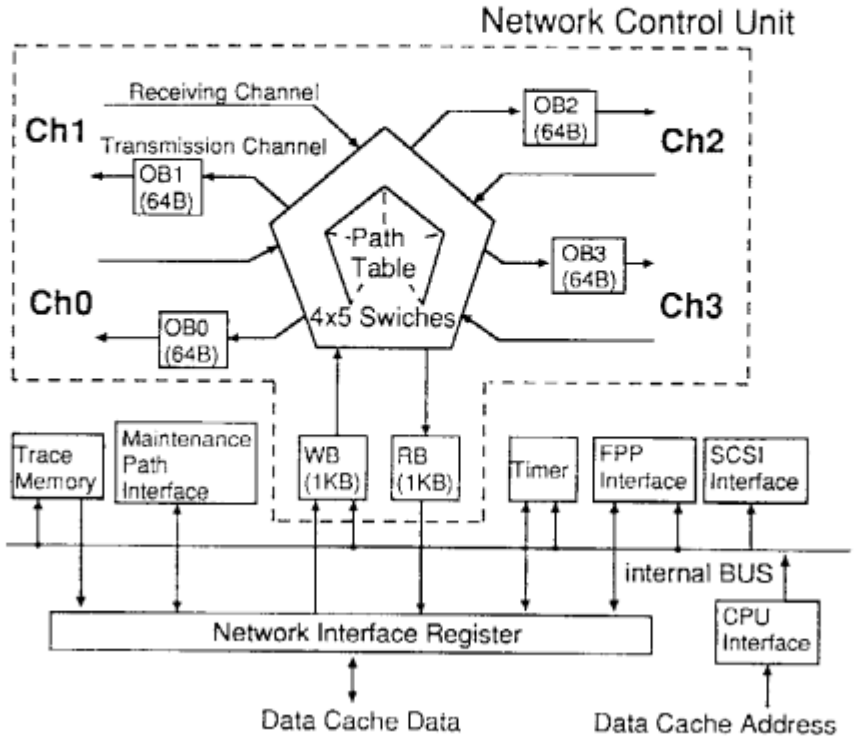


Figure 7 Block Diagram of the NU Chip

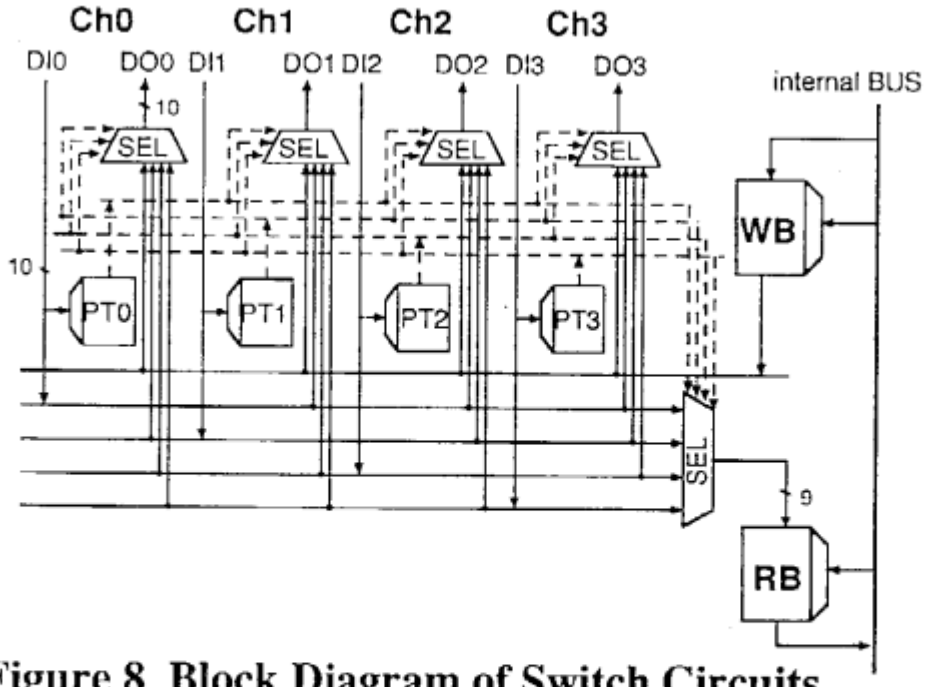


Figure 8 Block Diagram of Switch Circuits

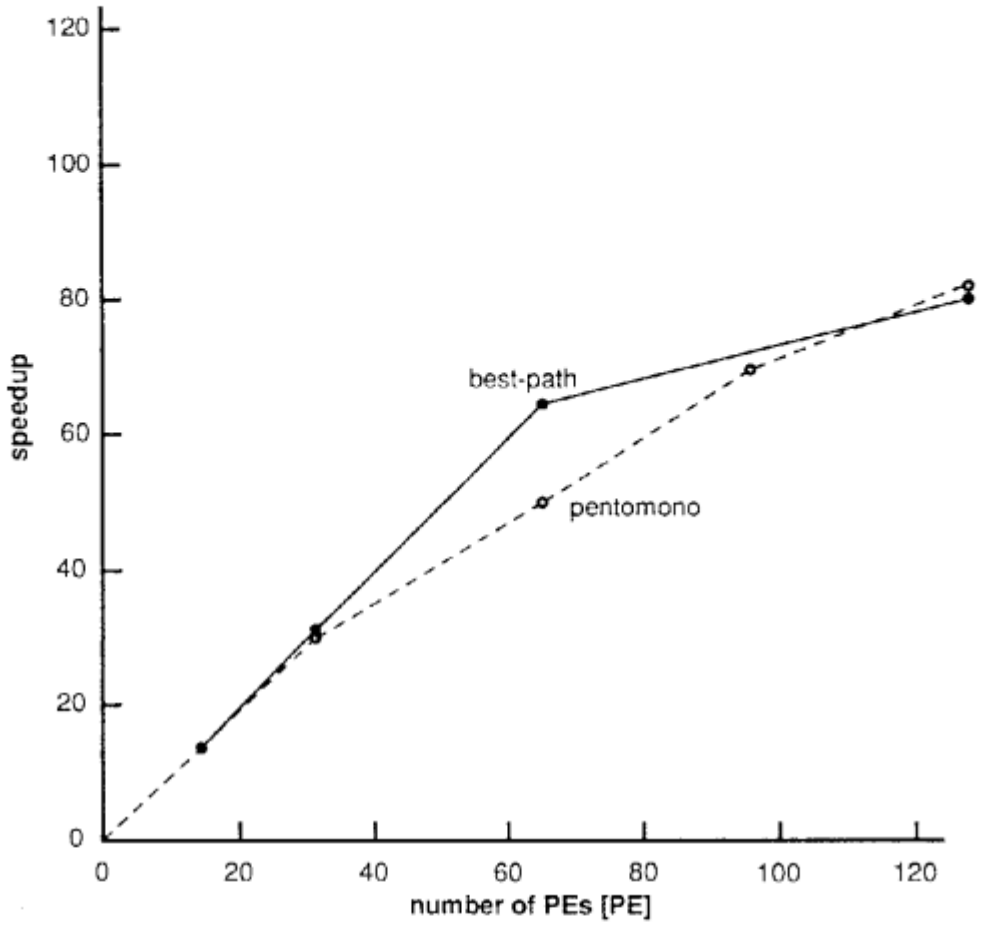


Figure 9 Speedups for best-path and pentomino

Table 1 Cache Features

	Instruction cache	Data cache
Address	Physical	Physical
Associativity	Direct map	Direct map
Cache size	5KBytes	20KBytes (off chip)
Tag entry	256	1024
TLB	32entry 2-way	32entry 2-way
TLB replacement	LRU	LRU
Parity		SECDED
Coherency		write back
		DRAM refresh

Table 2 Characteristics of NU Chip

Inter-PE Communications	
Operation Mode	Asynchronous
Control Strategy	Distributed Control
Switching Methodology	Packet Switching
Network Topology	Mesh Connection
Number of Channels	5 (4 ; adjacent, 1 ; own)
Width of Channel	10 bits (9 bits ; data, 1 bit ; parity)
Transmission Buffer	640 bits
Read Buffer	9 Kbits
Write Buffer	9 Kbits
Bandwidth	40 Mbits/sec

Table 3 Performance Improvement by AI workstation

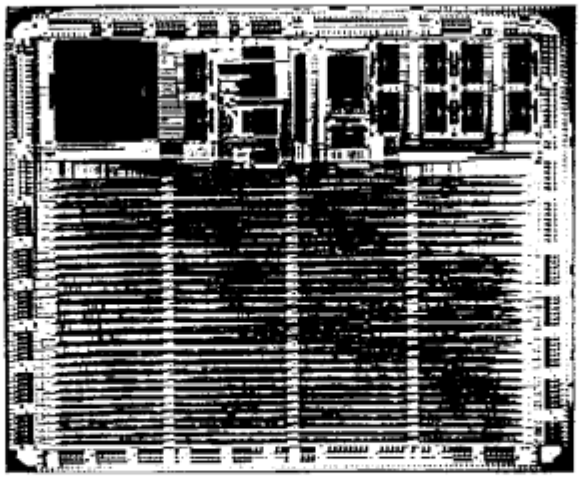
	cycle time	append operation in ESP	MLIPS
PSI-II (previous processor)	155nsec	15steps	0.43
	↓ **1.55x**		
0.8-μm PSI-II	100nsec	15steps	↓ **3.51x**
	↓ **1.67x**	↓ **1.36x**	
PU	**60nsec**	**11steps**	**1.4**

Table 4 Performance Improvement by PIM/m

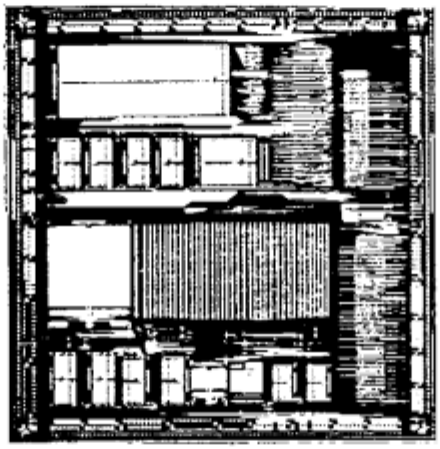
benchmark	condition	PIM/m	Multi-PSI	ratio
append	1,000 elements	1.63 msec	7.80 msec	4.8
best-path	90,000 nodes	142 sec	213 sec	1.5
pentomino	8 x 5 box	107 sec	240 sec	2.2
15-puzzle	5,885K nodes	9,283 sec	21,660 sec	2.3

Table 5 Device Features

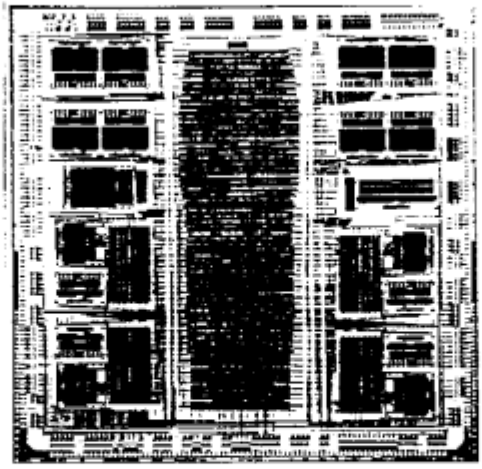
	Processor chip	Cache memory chip	Network control chip
Chip size	16.3 x 13.6 mm	14.5x14.8mm	14.2x14.0mm
Transistor count	384K	610K	329K
RAMs, PLAs	274K	545K	261K
Logics	110K	65K	68K
Scan path stage	409	420	268
Pad count	352	347	325
Package	361 pin PGA	361 pin PGA	361 pin PGA
Chip cycle time	30nsec	33nsec	28nsec
Power supply	5V	5V	5V
Power consumption	2.5W (at 16.7MHz)	2.3W (at 16.7MHz)	2.2W (at 16.7MHz)
Process technology	0.8μm double-metal CMOS	1.0μm double-metal CMOS	1.0μm double-metal CMOS



(a) PU Chip



(b) CU Chip



(c) NU Chip

Figure 10 Chip Photomicrographs

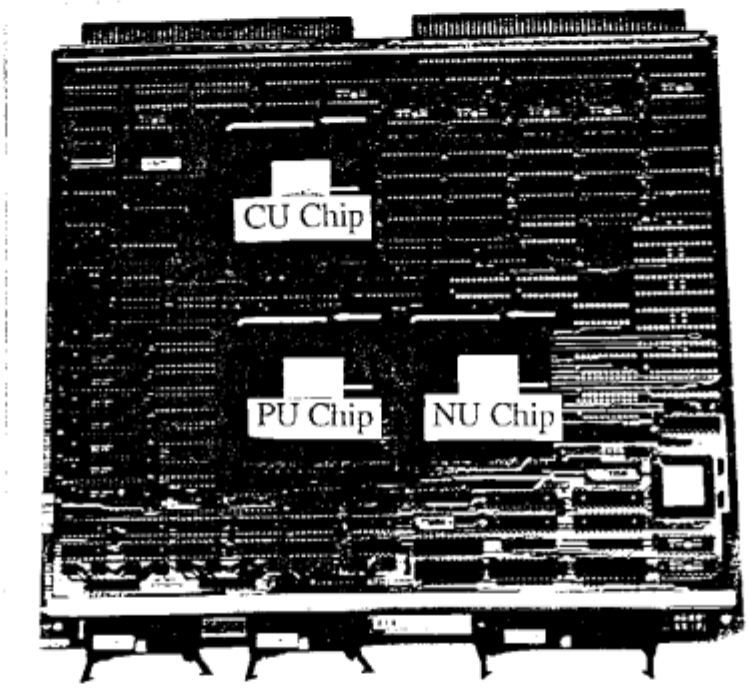


Figure 11 Photograph of Processing Element Board