Hardware Technology of the SX-9 (1) - Main System -

NAKAZATO Satoshi, TAGAYA Satoru, NAKAGOMI Norihito, WATAI Takayuki, SAWAMURA Akihiro

Abstract

The CPU of the SX-9 uses 8 sets of vector pipelines to achieve a high performance over 100GFLOPS with a single unit. The memory with a maximum 1T-byte capacity can be shared by up to 16 CPUs, the data transfer between the CPU and MMU has a maximum of 4T bytes/sec and that between the I/O unit and the MMU has a maximum of 64G bytes/sec. × 2.

This paper describes the CPU, memory and I/O processing unit of the SX-9 system featuring such high performances.

Keywords

supercomputer, processor, vector processing, shared memory, MMU, I/O processing

1. Introduction

The SX-9 inherits the highly effective performance and the convenient in-node shared memory that has been previously proven with the SX-4 to SX-8 systems. In order to meet the ever growing needs of S&T computations it also features enhancement of the overall system performance, achievement of high computation performance over 100GFLOPS with a single processor and an increased data transfer capability of the main memory.

In the following sections, we will introduce the CPU (Central Processing Unit), MMU (Main Memory Unit) and the I/O processing unit by focusing on particular features of each unit.

2. Processor

The CPU of the SX-9 uses the same SX architecture as before while further enhancing performance and functions and succeeding thus in achieving the performance of a node of the previous SX-8 using a single CPU.

2.1 Processor Configuration

Roughly speaking, the CPU is composed of a scalar unit and a vector unit. These are connected to the MMU through the processor-memory network. The SX-9 features the addition of an ADB (Assignable Data Buffer) as a new function in the processor-memory network. **Fig. 1** shows the configuration of the CPU.

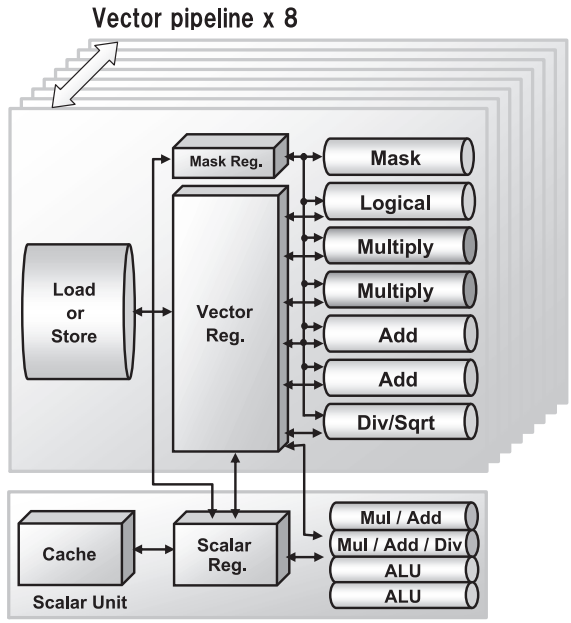

Fig. 1 CPU configuration.

2.2 Scalar Unit

The scalar unit of the SX-9 employs the 64-bit RISC architecture that is compatible with the SX series products and incorporates 128 general-purpose registers of 64-bit width and two L1 caches with 32K bytes for instructions and data. The features of the scalar unit are as described below.

(1)6-way Super-scalar Construction

The SX-9 uses a super-scalar configuration with a total of 6

execution pipelines including two for the floating-point arithmetic operations, two for integer arithmetic operations and two for memory operations. This configuration improves the performance in handling applications with high degrees of parallelism.

(2) Instruction Execution Control

The out-of-order execution mechanism contributes to the high performance by executing instructions that become executable, regardless of the program order. However, in order to improve the performance further, it is essential to search and retrieve the executable instructions over a large domain across multiple branch instructions. For that reason the SX-9 enables search and retrieval across up to 8 branch instructions. It additionally adopts a speculative instruction execution mechanism that performs tentative execution of succeeding instructions before the actual execution of a branch instruction and restarts an instruction from the correct condition in a case when the branch prediction fails. In order to expand the instruction window in the out-of-order execution mechanism the SX-9 has the reorder buffer entries enhanced to 64 entries.

2.3 Vector Unit

The vector unit is composed of the vector operation block and the vector control block.

(1) Vector Operation Block

The vector operation block has 8 sets of vector pipelines per processor, each of which is composed of the basic operation pipelines with concurrent operation capabilities including the Logical, Multiplication, Add/Shift and Div pipelines (the Multiplication and Add/Shift use 2 pipelines respectively) as well as the Mask and Load/Store pipelines, 16 mask registers with a 64-bit capacity per register and 72 vector registers with 512 bytes capacity per register. With the total capacity of the vector registers adding up to 144k bytes, this block can execute powerful vector operations by running the 48 basic operation pipelines concurrently.

1) Vector Operation Pipelines

Every operation pipeline supports the IEEE double-precision floating point and IEEE single-precision floating point data formats, which can be toddled on a per-instruction basis.

When the result of a preceding operation is used in a succeeding operation, the result of the preceding operation is usually used by means of a register. However, this vector unit has enhanced the function for transferring the proceeding operation result data directly to the operator of the succeeding operation. This strategy has increased the speed of processing instructions with short vector lengths compared to previous models.

The data transfer between vector pipes is also as high as 100G bytes/sec., which has enabled an increase in the speed of vector data compression and decompression by means of vector data transfer between vector pipes and mask bits.

2) Vector Register

In order to enable effective utilization of functional resources data can be supplied to and saved in vector operation pipelines and Load/Store pipelines simultaneously.

3) Vector Mask Register

Mask bits are generated concurrently using the Logical pipeline. The performance has also been improved by reinforcing the mechanism for chaining other operations using mask bits with inter-vector mask register operations and vice versa.

(2) Vector Control Block

The number of instruction issue stages, which was 2 with the SX-8, is increased to 24 with the SX-9 to enable instruction issuing by skipping up to 24 instructions as well as for the simultaneous issuing of 7 instructions. In addition, the vector control registers for use in the mask control during execution are multiplexed so that out-of-order issue for subsequent mask update instructions becomes possible without waiting for the completion of execution of a proceeding mask read instruction. The out-of-order issue of a mask read instruction after a subsequent mask update instruction also becomes possible. This procedure contributes to a significant reduction of the time loss in waiting for mask bit establishment in instructions and particularly of short vector lengths.

2.4 ADB (Assignable Data Buffer)

The ADB is located in the processor-memory network and functions mainly in selective buffering of vector data. The data transfer between the ADB and the vector unit is capable of achieving a higher performance than that between the processor and the memory with a shorter latency. As a result the effective performance can be improved further by storing frequently-used data in the ADB. This additionally makes it possible to reduce the memory traffic and avoid competition of memory banks, thereby reducing the degradation of multiprocessor performance due to concurrent processing.

The data stored in the ADB can be set to vector data only,

scalar data only or both vector and scalar data, so the optimum usage can be selected according to the executed application. In the vector data storage control, whether or not the ADB is used for each vector load/store instruction can be selected so that it is possible to prevent the necessary data from being unintentionally expelled from the ADB. In addition, in order to facilitate positive use of the ADB, a prefetch instruction has been added newly to support the prefetch function based on software control.

3. Main Memory Unit

The MMU adopts the shared memory system with which each CPU can access the memory evenly and at a high speed. The memory elements used in offering the large-capacity memory are the DDR-SDRAM (Double Data Rate-SDRAM) devices.

3.1 MMU Specifications

In the maximum configuration, the MMU is composed of 512 MMU cards. The memory capacity per MMU card is 2G bytes, and the system supports the capacity from 256G up to 1T bytes (see **Table**).

The MMU cards are capable of simultaneous, concurrent operations of memory access requests from the CPU. It features an extremely high data transfer capability with a memory throughput performance of 8G bytes/sec. per MMU card, or a maximum of 4T bytes/sec. for the system.

In addition, maximum of 32,768-way interleaving is set to the memory so that DDR3-SDRAM working rate increases and the effective performance improves.

3.2 Features of the MMU

The MMU features enhancement of the control LSI functions in order to achieve a theoretical peak performance as well as to achieve a highly effective overall performance.

The MMU incorporates a memory bank cache for each CPU. When a cache hit is found with a load request from a CPU, the data is read from the cache, so the memory access waiting time of succeeding requests can be reduced and the throughput performance can therefore be improved.

The MMU cards are required to be highly reliable because a large number of cards are used per system. In order to ensure high reliability, the SX-9 adopts measures as described in the following.

The timing budget between the control LSI and memory device is becoming more severe each year due to the increased data transfer speeds of the memory devices. In order to deal with this issue, the DDR3-SDRAM is provided with two additional calibration functions, which are the Write Leveling and Multi-Purpose Register (MPR) functions. These two functions are utilized in optimizing the memory device interface that is increasing the speed, thus extending the timing margin and further improving the reliability.

Moreover, the SX-9 adopts the system packaging of one high-speed CMOS LSI and 24 BGA (Ball Grid Array)-sealed DDR3-SDRAMs on a single circuit board. The wiring lengths of the LSI and RAM are decreased to improve the reliability of the high-speed signal transfer.

In addition to the above measures, the MMU adopts the ECC (Error Correcting Code) to improve the reliability and checks the timing, parity and dual circuitry to achieve a high failure detection rate. It also improves the RAS (Reliability, Availability, Serviceability) functions by incorporating the circuit check diagnostic function generating simulated faults and the built-in diagnostic function identifying the error point immediately from the error details.

Table MMU specifications.

Item	Specifications
Storage capacity	256G -1T bytes
Interleaving	8192 -32768 ways
Memory devices	1G-bitDDR3-SDRAM
Logic devices	CMOS LSI
Data supply capacity	1T -4Tbytes/sec.

4. I/O Processing Unit

The I/O processing unit of the SX-9 continues to be supported by the direct I/O system that has been adopted since the SX-8.

4.1 Features of the I/O Processing Unit

(1) All of the processors in the system are capable of equal-basis access to all of the I/O devices. Effective CPU utilization is made possible by allowing the software to set the CPU that the termination notification interrupts into as desired and allocating the CPUs with low application execution loads, to

the I/O control.
(2) The I/O control method adopted with the SX-9 is the direct I/O method, with which the memory in the HBA (Host Bus Adapter) is accessed directly according to the I/O access instructions from the CPU.
(3) The I/O interface can be selected according to the types of packaged channel cards. This facilitates expansion because a hardware setting is not necessary in the case of an extension, deletion or type change of the channel cards.

4.2 I/O Processing Unit Configuration

Fig. 2 shows the configuration of the I/O processing unit. The I/O processing unit of the SX-9 is composed of the IOFs (IO Features: Host Bridge Units) and PCI Express control units.

Up to 16 IOFs can be mounted per system, and two PCI Express control units can be connected to each IOF. As each IOF is virtually represented as a single channel device, all of the functions incorporated in the IOFs can be set and modified at the desired timing from the software using the same access method as a general purpose channel card.

The PCI Express control unit is a controller incorporating a PCI Express (× 8) slot, and implements a data transfer performance of max. 2G bytes/sec. × 2 (bidirectional) per unit. As a result, the total maximum transfer performance per system adds up to 64G bytes/sec. × 2.

Each IOF and each PCI Express control unit incorporates an independent data buffer and manages the data stream on a per-channel basis. This structure makes it possible to eliminate interference of transfer processing due to data waiting between channels and to achieve high throughputs permanently in all of the channels.

In addition, the IOF has a separate fault processing path from the normal request path so that it is capable of initializing or setting the IOF and the PCI Express control unit under it as well as collecting their fault logs, even when the normal request path is faulty.

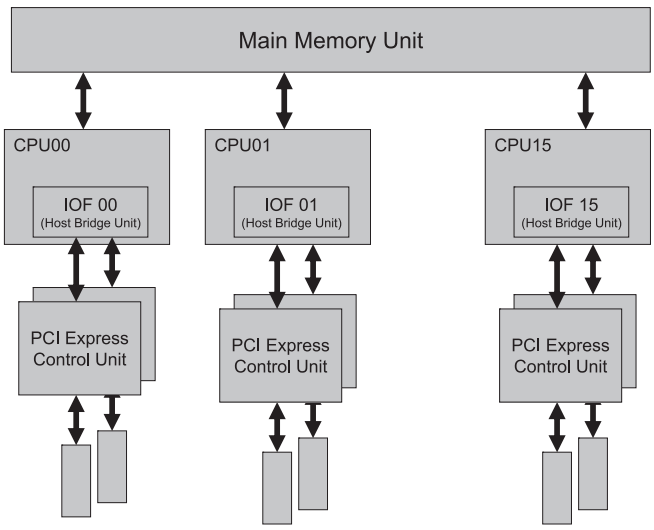

Fig. 2 I/O control unit configuration.

5. Conclusion

In the above, we introduced the SX-9 by focusing on its processor, main memory unit and I/O processing unit. While inheriting the SX architecture approved for highly effective performance, the SX-9 has been developed as a supercomputer product with higher execution and cost efficiencies. It also features enhancements in the instruction execution control and memory throughput performances.

In the future, we intend to develop even better supercomputer products by incorporating user needs and taking full command of the most advanced technologies.

Authors' Profiles

NAKAZATO Satoshi
Manager,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

TAGAYA Satoru
Manager,
Computers Division,
1st Computers Operations Unit,
NEC Corporation

NAKAGOMI Norihito
Manager,
2nd Computer Technology Dept.,
NEC Computertechno, Ltd.

WATAI Takayuki
Assistant Manager,
2nd Computer Technology Dept.,
NEC Computertechno, Ltd.

SAWAMURA Akihiro
Manager,
Computers Division,
1st Computers Operations Unit,
NEC Corporation