Title	
Author(s)	,
Citation	
Issue Date	1999-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1285
Rights	
Description	Supervisor: , ,

Japan Advanced Institute of Science and Technology

Potential of the high frequency operation
of the pipelined cache memory

Ukai Kazutoshi

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 1999

Keywords: pipeline, cache, memory-cell array, high frequency operation.

1   Introduction

This paper describes a design and performance evaluation of pipelined cache memory and its potential of the high frequency operation is discussed.

Great efforts have been made to shorten of a machine cycle, and RISC (Reduced Instruction Set Computer) has developed. However, improvement of processor performance in speed is saturated by wiring delayes. In order to overcome the situation, the advanced pipeline processing is intoroduced to processor enthusiastically so that it improves throughput. The Multithreaded Ultra Pipelined Processor which processes a number of threads in parallel with time sharing has been proposed and designed at our laboratory. This processor can shorten the clock cycle time of one stage by dividing the stages, and the number of pipeline stages is inevitably increased. In order to operate this processor efficiently, the pipelined cache which supplies instructions and data by the high frequency is indispensable.

When introducing the pipelining to a cache mechanism, increasing the number of pipeline stages enables to shorten a clock cycle time. However, the clock cycle time depends on the latest stage such as a memory-cell array. This research examines how high frequency operation becomes possible by pipelining. For this objective, the circuit are designed for every stage in detail, and evaluated.

2 Pipelining of a cache mechanism

2.1 Pipelining of a cache mechanism

Pipeline processing is a method of raising throughput by dividing one instruction into several stage and carrying out in parallel. In this research, this pipeline processing applied to a cache mechanism, so that it improves throughput of a cache memory.

2.2 Specifications of the cache

The specifications of the cache designed are given as follows so that it might be suitable as a primary cache for the Multithreaded Processors.

- The cache capacity of thread common is the 128K word data capacity for 16 threads. It is 8K word per one thread.

- It is 4 way set associative cache, and line size is 4 words.

- The write-in system of the cache is a write back system.

- The total capacity of the cache is 4608Kb for data section 4Mb, tag section 480Kb, and effective bit section 32Kb.

2.3 The hierarchical decode method

The hierarchical decode method is introduced to perform a decode by dividing a decoder into 3 stage.

2.4 Division of a memory-cell array

The decoder can be pipelined by inserting latches. However, read-out operation of a memory-cell array cannot be pipelined. Therefore, physical size of the memory-cell array must be made small as to shorten delay.

3 A design of a pipeline cache

The $0.25\mu m$ technology is introduced to design the detailed circuit for every stage. The simulation and the design are related closely. Therefore they are repeated so that it is designed the optimum. Especially, in order to find the part to insert a latch, the simulation was repeated, and it has been endeavored so that a latch as small as possible might be inserted in the optimum position.

Moreover, in order to derive the area of the circuit increased by pipelining, the layout of various basic circuits which constitute the pipelined cache is introduced.

4 SPICE simulation

The MOSFET model used in this simulation is introduced, and the property of pMOS and nMOS are investigated. As a result of the simulation for every stage, the number of stages and the frequency of operation of the pipelined cache are clarified.

5 Consideration

5.1 Consideration about division of a memory-cell array

When pipelining a cache, it becomes important how size a memory-cell array is made. As to the size of one memory-cell array with $64 \times 64bit$, $128 \times 128bit$ and $256 \times 256bit$, the simulation which changed the size is performed.

If the memory-cell array is divided small, the rate of the wiring delay occupied to all delay time will become smaller than the rate of gate delay. Therefore, if division is advanced, the rate of improvement in the speed becomes small. In $0.25\mu m$ technology, it was derived that the size of $128 \times 128bit$ was suitable.

5.2 Consideration about a precharge circuit

In $128 \times 128bit$ data memory, the processing time in the case of having a precharge circuit and not having are compared. Consequently, it was found that a small memory-cell array with $128 \times 128bit$ in special composition called a pipeline cache is not accelerated by the precharge circuit.

6 Conclusion

A Multithreaded Processor desires a high throughput more than a low latency to a cache. When the cache is used for such use, this research examines how high frequency operation becomes possible by pipelining.

When pipelining a cache, it becomes important how size a memory-cell array is made. The rate of improvement in the speed by division of a memory-cell array and the effect by existence of a precharge circuit was clarified.

Moreover, in this research, the design of the cache with the data capacity of 4Mb is designed in detail. And, the minimum clock cycle time of each stage is examined.

Consequently, in $0.25\mu m$ technology, it was derived that the pipelined cache could operate on the frequency of 3GHz as 9 stages. Moreover, in $0.10\mu m$ technology, it could operate on the frequency of 7.7GHz as 9 stages of same stage composition as $0.25\mu m$. And further, by dividing a stage, it could operate on the frequency of 6.5GHz.