

When the addressed word in a Read operation is not in the cache, a read miss occurs. The block of words that contains the requested word is copied from the main memory into the cache. After the entire block is loaded into the cache, the particular word requested is forwarded to the processor. Alternatively, this word may be sent to the processor as soon as it is read from the main memory. The latter approach, which is called load-through, or early restart, reduces the processor's waiting period somewhat, but at the expense of more complex circuitry.

During a Write operation, if the addressed word is not in the cache, a write miss occurs. Then, if the write-through protocol is used, the information is written directly into the main memory. In the case of the write-back protocol, the block containing the addressed word is first brought into the cache, and then the desired word in the cache is overwritten with the new information.

MAPPING FUNCTION

OR

6(a) Explain basic concepts of pipeling in details . Also explain the hazards **PIPELING**

BASIC CONCEPTS

The speed of execution of programs is influenced by many factors. One way to improve performance is to use faster circuit technology to build the processor and the main memory. Another possibility is to arrange the hardware so that more than one operation can be performed at the same time. In this way, the number of operations performed per second is increased even though the elapsed time needed to perform any one operation is not changed.

(a) Sequential execution

(b) Hardware organization

[10]

 $CO5$ L₂

(c) Pipelined execution

Figure 8.1 Bosic idea of instruction pipelining.

The computer is controlled by a clock whose period is such that the fetch and execute steps of any instruction can each be completed in one clock cycle. Operation of the computer proceeds as in Figure 8.1c. In the first clock cycle, the fetch unit fetches an instruction I_1 (step F_1) and stores it in buffer B1 at the end of the clock cycle. In the second clock cycle, the instruction fetch unit proceeds with the fetch operation for instruction I_2 (step F_2). Meanwhile, the execution unit performs the operation specified by instruction I_1 , which is available to it in buffer B1 (step E_1). By the end of the

second clock cycle, the execution of instruction I_1 is completed and instruction I_2 is available. Instruction I_2 is stored in B1, replacing I_1 , which is no longer needed. Step E_2 is performed by the execution unit during the third clock cycle, while instruction I_3 is being fetched by the fetch unit. In this manner, both the fetch and execute units are kept busy all the time. If the pattern in Figure 8.1c can be sustained for a long time, the completion rate of instruction execution will be twice that achievable by the sequential operation depicted in Figure 8.1a.

In summary, the fetch and execute units in Figure 8.1b constitute a two-stage pipeline in which each stage performs one step in processing an instruction. An interstage storage buffer, B1, is needed to hold the information being passed from one stage to the next. New information is loaded into this buffer at the end of each clock cycle.

The processing of an instruction need not be divided into only two steps. For example, a pipelined processor may process each instruction in four steps, as follows:

- F Fetch: read the instruction from the memory.
- D Decode: decode the instruction and fetch the source operand(s).
- Е Execute: perform the operation specified by the instruction.
- W Write: store the result in the destination location.

ROLE OF CACHE MEMORY

Each stage in a pipeline is expected to complete its operation in one clock cycle. Hence, the clock period should be sufficiently long to complete the task being performed in any stage. If different units require different amounts of time, the clock period must allow the longest task to be completed. A unit that completes its task early is idle for the remainder of the clock period. Hence, pipelining is most effective in improving

performance if the tasks being performed in different stages require about the same amount of time.

This consideration is particularly important for the instruction fetch step, which is assigned one clock period in Figure 8.2a. The clock cycle has to be equal to or greater than the time needed to complete a fetch operation. However, the access time of the main memory may be as much as ten times greater than the time needed to perform basic pipeline stage operations inside the processor, such as adding two numbers. Thus, if each instruction fetch required access to the main memory, pipelining would be of little value.

(a) Instruction execution divided into four steps

PIPELINE PERFORMANCE

For a variety of reasons, one of the pipeline stages may not be able to complete its processing task for a given instruction in the time allotted. For example, stage E in the four-stage pipeline of Figure 8.2b is responsible for arithmetic and logic operations, and one clock cycle is assigned for this task. Although this may be sufficient for most operations, some operations, such as divide, may require more time to complete. Figure 8.3 shows an example in which the operation specified in instruction I_2 requires three cycles to complete, from cycle 4 through cycle 6. Thus, in cycles 5 and 6, the Write stage must be told to do nothing, because it has no data to work with. Meanwhile, the information in buffer B2 must remain intact until the Execute stage has completed its operation. This means that stage 2 and, in turn, stage 1 are blocked from accepting new instructions because the information in B1 cannot be overwritten. Thus, steps D_4 and F_5 must be postponed as shown.

Figure 8.3 Effect of an execution operation taking more than one clock cycle.

BASIC CONCEPT 8. T

Pipelined operation in Figure 8.3 is said to have been *stalled* for two clock cycles Normal pipelined operation resumes in cycle 7. Any condition that causes the pipeline to stall is called a hazard. We have just seen an example of a data hazard. A data hazard is any condition in which either the source or the destination operands of an instruction are not available at the time expected in the pipeline. As a result some operation has to be delayed, and the pipeline stalls.

The pipeline may also be stalled because of a delay in the availability of an instruction. For example, this may be a result of a miss in the cache, requiring the instruction to be fetched from the main memory. Such hazards are often called control hazards of *instruction hazards*. The effect of a cache miss on pipelined operation is illustrated in Figure 8.4. Instruction I_1 is fetched from the cache in cycle 1, and its execution proceeds normally. However, the fetch operation for instruction I_2 , which is started in cycle 2 results in a cache miss. The instruction fetch unit must now suspend any further fetch requests and wait for I_2 to arrive. We assume that instruction I_2 is received and loaded into buffer B1 at the end of cycle 5. The pipeline resumes its normal operation at that point

(a) Instruction execution steps in successive clock cycles

(b) Function performed by each processor stage in successive clock cycles

Figure 8.4 Pipeline stall caused by a cache miss in F2.

PART D

7(a) Explain three types of mapping functions for cache memory. (Direct mapping -5M, Associative Mapping - 5M, Set -associative mapping - 5M)

DIRECT MAPPING

This technique is easy to implement but not very flexible.

Block j of the main memory maps onto j modulo 128 of the cache. For example, whenever one of the main memory blocks 0, 128, 256, Is loaded in the cache, it is stored in cache block Q_{ni} Main memory blocks 1, 129, 257_{min} are stored in cache block 1 (one at a time), and so on. Contention may occur for a single cache block required by multiple memory blocks. E_{z} g when for program execution both memory block 1 and 129 are required but cache block 1 can only store one memory block. To resolve this, new blocks are allowed to overwrite the currently resident block.

From example,

4096 memory blocks need to be mapped to 128 cache blocks. i.e., each cache block identified 32 memory blocks(4096/128).

Main memory address is divided into three parts:

Tag (5 bits): identify which memory block (out of 32 in this case) is currently resident in the cache

Block (7 bits): cache block position where the new memory block must be stored

Word (4 bits): selects one of the words of the memory block (out of 16 words per block in this case)

[15]

CO3

 $L2$

- It is more flexible than direct mapping technique but more expensive. Main memory block can be placed into any cache block position.
- Memory address is divided into two fields: \bullet - Low order 4 bits identify the memory word within a block. - High order 12 bits or tag bits identify a memory block when residing in the cache.
- Flexible, and uses cache space efficiently. \bullet
- Replacement algorithms can be used to replace an existing block in the cache when the cache is full.
- Cost is higher than direct-mapped cache because of the need to search all 128 patterns to determine whether a given block is in the cache.

are 7.15 An example of microinstructions for Figure 7.6.

ORGANIZATION OF MICROPROGRAMMED CONTROL UNIT TO SUPPORT CONDI **BRANCHING**

• Drawback of previous Microprogram control:

> It cannot handle the situation when the control unit is required to check the status of condition codes or external inputs to choose between alternative courses of action.

Solution:

> Use conditional branch microinstruction.

. In case of conditional branching, microinstructions specify which of the external inputs, c codes should be checked as a condition for branching to take place.

• Starting and Branch Address Generator Block loads a new address into µPC microinstruction instructs it to do so (Figure 7.18).

. To allow implementation of a conditional branch, inputs to this block consist of

 \rightarrow external inputs and condition-codes &

 \rightarrow contents of IR.

• µPC is incremented every time a new microinstruction is fetched from microprogram memor in following situations:

1) When a new instruction is loaded into IR, µPC is loaded with starting-address of mici for that instruction.

2) When a Branch microinstruction is encountered and branch condition is satisfied loaded with branch-address.

3) When an End microinstruction is encountered, µPC is loaded with address of firs microroutine for instruction fetch cycle.

Figure 7.18 Organization of the control unit to allow conditional branching in the microprogram.

Figure 7.17 Microroutine for the instruction Branch < 0.

MICROINSTRUCTIONS

• A simple way to structure microinstructions is to assign one bit position to each con required in the CPU.

. There are 42 signals and hence each microinstruction will have 42 bits.

• Drawbacks of microprogrammed control:

- 1) Assigning individual bits to each control-signal results in long microinstructions bec the number of required signals is usually large.
- 2) Available bit-space is poorly used because
	- only a few bits are set to 1 in any given microinstruction.
- · Solution: Signals can be grouped because
	- 1) Most signals are not needed simultaneously.
	- 2) Many signals are mutually exclusive. E.g. only 1 function of ALU can be activated at For ex: Gating signals: IN and OUT signals (Figure 7.19).
		- Control-signals: Read, Write.
		- ALU signals: Add, Sub, Mul, Div, Mod.
- · Grouping control-signals into fields requires a little more hardware because
- decoding-circuits must be used to decode bit patterns of each field into individual contro • Advantage: This method results in a smaller control-store (only 20 bits are needed to patterns for the 42 signals).

MICROINSTRUCTIONS

• A simple way to structure microinstructions is to assign one bit position to each cont required in the CPU.

. There are 42 signals and hence each microinstruction will have 42 bits.

• Drawbacks of microprogrammed control:

- 1) Assigning individual bits to each control-signal results in long microinstructions beca the number of required signals is usually large.
- 2) Available bit-space is poorly used because
	- only a few bits are set to 1 in any given microinstruction.
- · Solution: Signals can be grouped because
	- 1) Most signals are not needed simultaneously.
	- 2) Many signals are mutually exclusive. E.g. only 1 function of ALU can be activated at
		- For ex: Gating signals: IN and OUT signals (Figure 7.19).
			- Control-signals: Read, Write.
			- ALU signals: Add, Sub, Mul, Div, Mod.
- · Grouping control-signals into fields requires a little more hardware because

decoding-circuits must be used to decode bit patterns of each field into individual control-• Advantage: This method results in a smaller control-store (only 20 bits are needed to patterns for the 42 signals).

(b) Explain with a neat diagram hardwired control. [8] $\sqrt{\frac{COS}{COS}}$ L₂

Figure 7.11 Separation of the decoding and encoding functions.

- Hardwired control is a method of control unit design (Figure 7.11).
- . The control-signals are generated by using logic circuits such as gates, flip-flops, decoders etc.
- Decoder/Encoder Block is a combinational-circuit that generates required control-outputs depending on state of all its inputs.

• Instruction Decoder

- > It decodes the instruction loaded in the IR.
- > If IR is an 8 bit register, then instruction decoder generates $2^8(256$ lines); one for each instruction.
- \triangleright It consists of a separate output-lines INS₁ through INS_m for each machine instruction.
- > According to code in the IR, one of the output-lines INS1 through INS_m is set to 1, and all other lines are set to 0.
- . Step-Decoder provides a separate signal line for each step in the control sequence.
- Encoder
	- > It gets the input from instruction decoder, step decoder, external inputs and condition codes.
	- \triangleright It uses all these inputs to generate individual control-signals: Y_{in}, PC_{out}, Add, End and so on.
	- > For example (Figure 7.12), $Z_{in} = T_1 + T_6$.ADD+T₄.BR

; This signal is asserted during time-slot T_1 for all instructions.

during T_6 for an Add instruction.

during T₄ for unconditional branch instruction

- . When RUN=1, counter is incremented by 1 at the end of every clock cycle. When RUN=0, counter stops counting.
- . After execution of each instruction, end signal is generated. End signal resets step counter.
- Sequence of operations carried out by this machine is determined by wiring of logic circuits, hence the name "hardwired".
- Advantage: Can operate at high speed.
- Disadvantages:
	- 1) Since no. of instructions/control-lines is often in hundreds, the complexity of control unit is very high.
	- 2) It is costly and difficult to design.
	- 3) The control unit is inflexible because it is difficult to change the design.