Chapter 07: Instruction–Level Parallelism– VLIW, Vector, Array and Multithreaded Processors …

Lesson 06:

Multithreaded Processors

Objective

- To learn meaning of thread
- To understand multithreaded processors, hyper-threading and simultaneous multithreading technology
- To understand operations in Multithreaded processors

Thread

Process

- Program can be considered to made up of processes
- A process can be considered to made up of **threads**
- Each process consists of at least one thread
- Processes can be made up of multiple threads
- Each of these threads can have its own local context in addition to the process's context, which is shared by all the threads in a process

Thread

- The program or process (set of instructions), which actually shares RAM with all of the other currently running threads
- The program that has to wait its turn for a slice or scheduling of CPU time in order to execute, just like all of the other programs on the system

Context

- Means CPU registers including program counter and stack pointer of a thread

Thread

- The CPU or CPU pipelines can *execute* only one of the threads at a time
- The operating system (OS) maintains the illusion of concurrency by switching between running threads either at a fixed interval, called a **time slice or as per priority**
- The OS schedules different threads in different intervals

Running of a thread

- When allotted time interval of thread is over, its context is saved to memory so that when the thread is scheduled next by the OS, the context can be restored to the exact same state that it was in when its execution stopped
- The thread runs again and its context has been restored exactly as it was when it left off last

Context Switch from one thread to another in multithreaded program

- Hardware supports the context switching
- A context switch takes certain number of CPU cycles
- Multithreaded processors does fast context switches between the threads

Symetric Multithreaded Processor

Symmetric multithreaded processor (SMP)

- Instead of instruction-level parallelism, there is thread-level parallelism in multithreaded processors
- SMP is term used when there is symmetric multiprocessing (SMP) on processor cores
- A single-threaded SMP means one thread running on one processor core and another on another

Super threading in Processor

Superthreading

- Time-sliced multithreading
- A superthreading processor called a multithreaded processor
- There is restriction that all the instructions issued by the instruction issue logic each clock cycle be from the same thread
- Such processors are capable of executing more than one thread in the pipeline stages, as two or more threads can execute by issuing each thread instructions in successive clock cycles

Superthreading

- Each processor pipeline stage is issued instructions for one and only one thread in a clock cycle, so that the instructions from each thread move in lockstep through the processor in a multithreaded processor

Simultaneous multi-threading (SMT) [Hyper-Threading]

Hyper-Threading Technology

- Thread-level-parallelism (TLP) on each processor resulting in increased utilization of processor execution resources
- As a result, resource utilization yields higher processing throughput

Hyper-Threading Technology

- Multiple threads of an application program can be run simultaneously on one processor or processor-core
- Hyper-threading can be said to be next level of superthreading
- There is no restriction in simultaneous multi-threading
- All the instructions issued by the instruction issue logic each clock be from the same thread
- Several threads can share a set of resources during processor execution

Simultaneous multi-threading technology (SMT)

- Hyper-Threading Technology is a form of simultaneous multi-threading technology (SMT) where multiple threads of software applications can be run simultaneously on one processor

Simultaneous Multithreading

- Allows multiple threads to issue instructions each cycle
- Simultaneous multithreading enables multithreaded applications to execute threads in parallel on a single processor or single multi-core processor instead of processing threads in a linear fashion on multiple parallel pipelines

Pipelines and Pipeline Stages in Simultaneous Multi-threading Multithreaded Processor

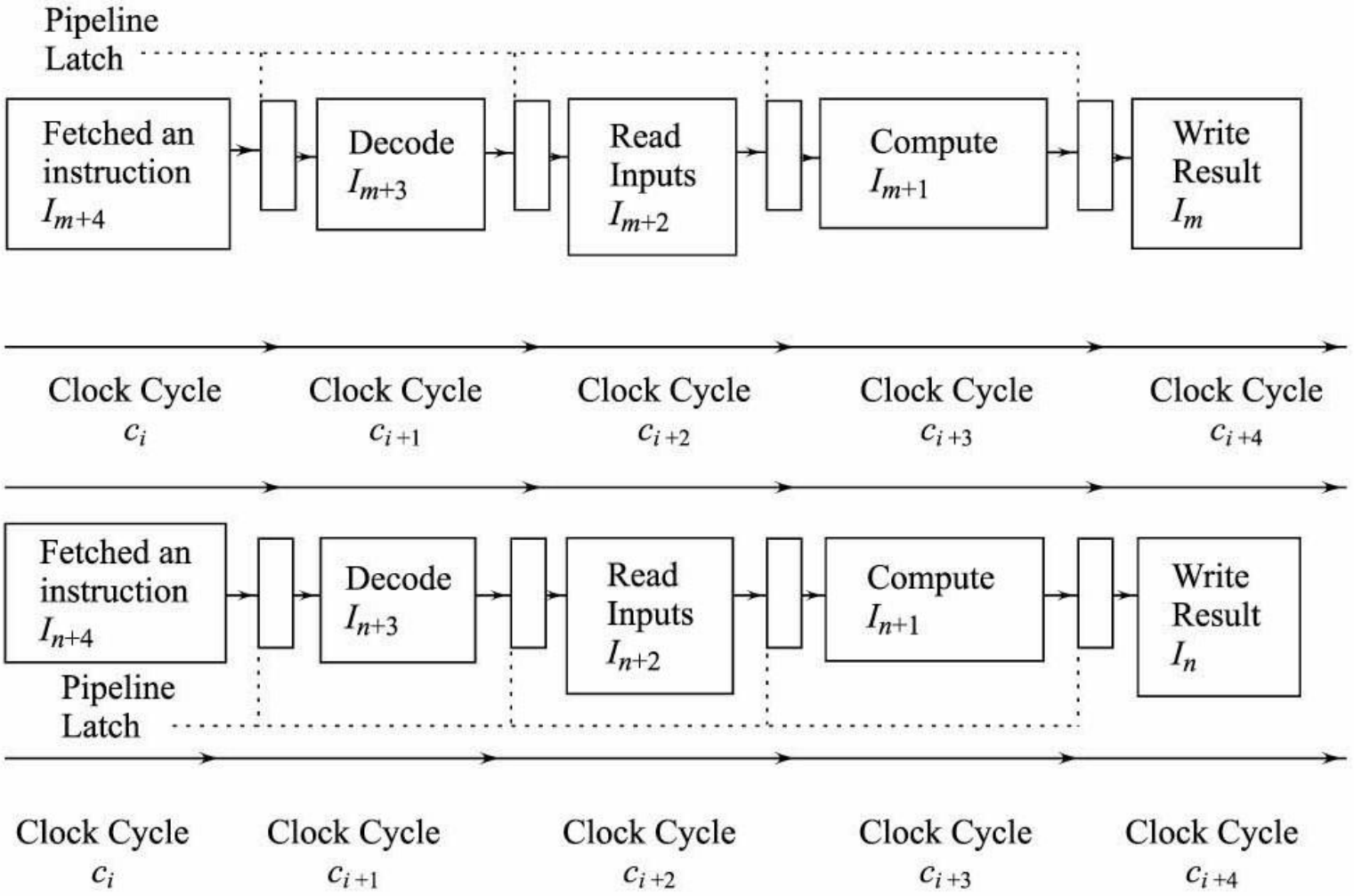
Set of pipelines or pipeline stages

- A can be assigned to one thread and another set of pipelines or pipeline stages to another
- Two or more threads are sequentially executing in each pipeline
- Usually, a pipeline stalls for a number of cycles due to data unavailability and the thread and instruction issue logic enforce the context switch to another thread

Set of pipelines or pipeline stages

- A multithreaded processor feature is that it reduces the total number of context switches as the number of threads run on a single processor or processor core
- As and when a pipeline stall is detected due to need of some message, the instruction issue logic issues the instructions of another thread

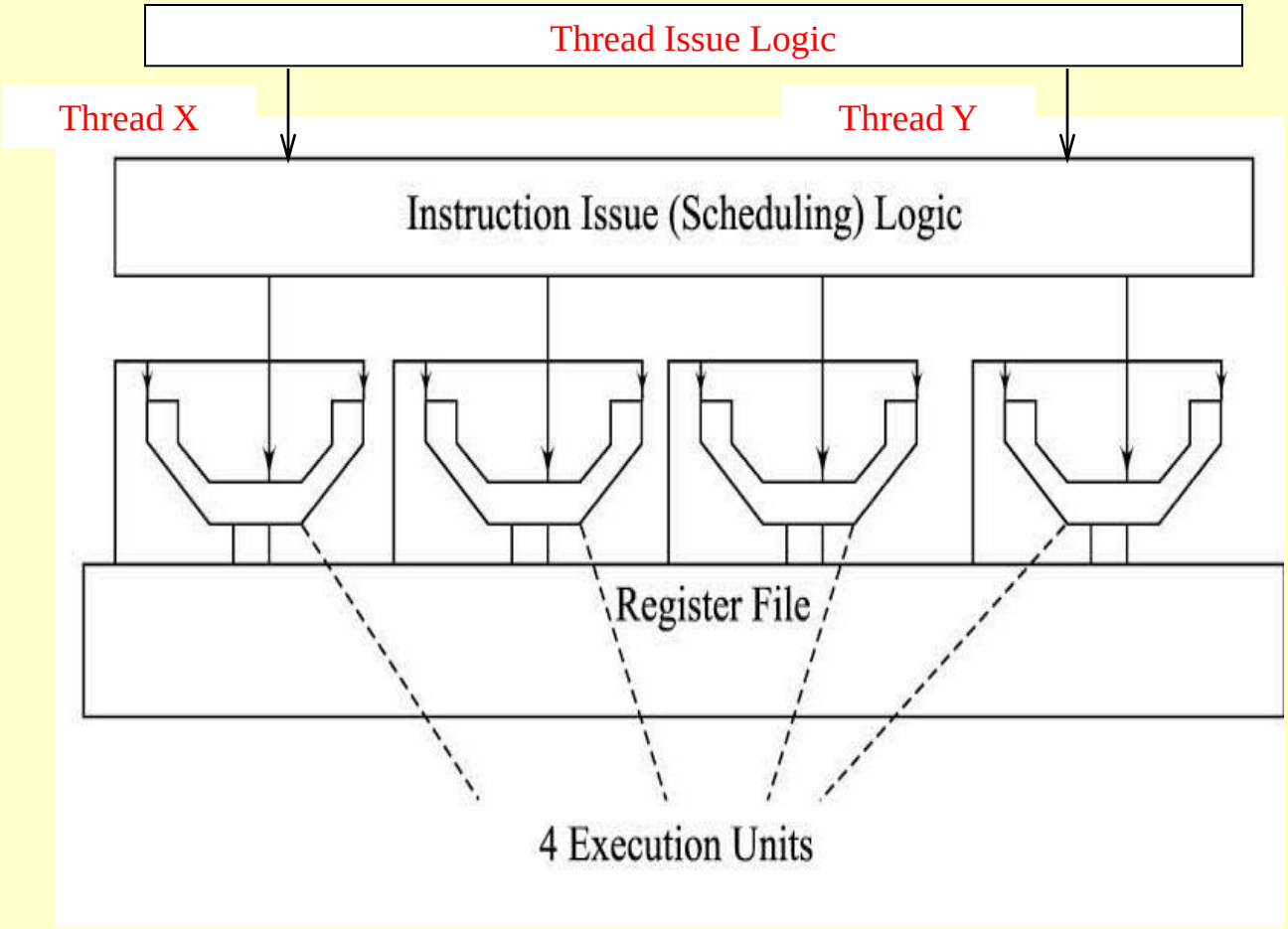
Two Pipelines and scheduling of Two threads on them in multithreaded processor using hyper- threading technology



Use of thread issue logic

- Use of thread issue logic for issuing instructions of multiple threads to instruction issue logic and instruction issue logic in simultaneous multithreaded processor issuing instructions to multiple pipeline stages in the pipelines in each clock cycle

Multiple threads to instruction issue logic and instruction issue logic in simultaneous multithreaded processor



Context Switch from one thread to another in multithreaded program

- Hardware supports the fast context switching for each pipeline so that the thread context saves and new retrieves and new thread run in that pipeline
- A context switch takes certain number of CPU cycles in a pipeline
- Multithreaded processors does fast context switches between the threads in a given pipeline

Performance improvement in multithreaded processor

- Performs a context switch at every instruction stage
- Multithreaded pipelining eliminates delays due to data and control hazards, and masks the effects of memory latency
- Memory latency means that data read of operands are delayed
- Processors by allowing independent instructions to execute simultaneously (called instruction-level parallelism)

Thread in a Pipeline

- A pipeline be assigned to one process (thread)
- The thread is sequentially executing on that in case of hyper-threading technology

Thread in a Pipeline

- A multithreaded processor feature is that as and when a stall is there due to time interval over for a thread, the processor starts executing the instructions of another thread

A multithreaded processor feature

- As and when a stall on time interval over is detected, the processor starts executing the instructions of another thread after saving the previously executing thread context and retrieving the new thread context

Thread in a pipeline

- Let us assume that a thread *i* consists of instructions, I_n, I_n+1, I_n+2, I_n+3, and I_n+4 .., in pipeline X and another thread *j* consists of instructions I_m, I_m+1, I_m+2, I_m+3, and I_m+4 … in pipeline Y.

Thread Stall on time interval over in Pipeline X

- The pipeline executing the thread then switches to another thread instructions
- Suppose a stall occurred on time interval over then a new thread instruction will be issued in the pipeline after saving the thread context and retrieving new thread context

Summary

We Learnt

- A program or process can be considered as consisting of multiple threads
- The OS schedules the threads in the pipelines
- When there is stall due to time slice over, the other thread starts execution after saving the previous thread context and loading the new context

End of Lesson 06 on

Multithreaded Processors