Pipelining (II)

Pipelining Calculation

- **Speedup** is used to measure the performance of pipelining. It’s calculated as follow:

  \[ \tau = \max_i(\tau_i) + d = \tau_m + d \]

  \( \tau_i \): the logic delay for stage \( i \)

  \( \tau_m \): the maximum logic delay for any stage

  \( d \): latch delay

- The speedup is calculated as follow

  \[ \text{Speedup} = \frac{\text{Time to execute N instructions on the single-stage CPU}}{\text{Time to execute N instructions on K stages}} = \frac{N T_1}{[K + (N - 1)] \tau} \]

  \( T_1 \): the single-stage logic delay plus the latch delay

Example

- Assume there is a four-stage CPU. Each stage completes one step of the four steps of instruction execution: fetch, decode, execute, and write back. The logic delay for each step is 5 ns, and the latch delay is 1 ns. Answer the following questions.

  - What is the time to execute one instruction on the 4-stage CPU?
    \((5 + 1) \times 4 = 24\text{ ns}\)

  - Assume there is a single-stage CPU that completes all four steps in one stage. What is the time to execute one instruction on the single-stage CPU?
    \(5 \times 4 + 1 = 21\text{ ns}\)

  - What is the time to execute five instructions on the single-stage CPU?
    \(5 \times 21 = 105\text{ ns}\)

  - What is the time to execute five instructions on the four-stage CPU?
    \((4 + 5 - 1) \times (5 + 1) = 48\text{ ns}\)

  - What is the speedup for 5 instructions using the four-stage CPU?
    \(105/48 = 2.2\)
Real Computer Pipeline
- In real computers, the general four steps of instruction execution, Fetch, Decode, Execute, and Write-back, are often divided further to better balance the length of each step.

- MIPS pipeline
  - MIPS is a RISC developed by MIPS technologies, a classic RISC pipeline
  - Five steps to execute an instruction
    - Fetch instruction from memory
    - Read registers while decoding the instruction. The regular format of MIPS instructions allows reading and decoding to occur simultaneously.
    - Access an operand in data memory
    - Execute the operation or calculate an address
    - Write the result into a register

- Modern Pipeline
  - E.g. Pentium / PowerPC
  - Seven steps to execute an instruction
    - Instruction fetch unit: get instructions
    - Decode the instructions
    - Identify where the operands are
    - Fetch the operands to the right place
    - Execute the operation
    - Store the operands
    - Operation Commit Unit: update the architectural state of the CPU

- In this course, we use the four steps for simplicity without losing the generality of this topic.

Hazards
- Pipelining can improve the throughput and speed up the instruction execution. However, there are some situations that prevent the next instruction from starting in the next cycle. We call these situations “hazards.”
- Hazards can generally be grouped into three types: structural hazards, data hazards, and control hazards.
Structural Hazards
- Structural hazards occur when a required resource is busy. For example, in dirty laundry example, if the load B starts using the washer once the load A is done with washing, the dryer is still drying the load A and is unavailable to the load B when it is done with washing.
- To address the structural hazards, more stages are usually helpful. More stages need more control hardware, but it also lowers the bound on cycle time of each stage. Modern CPU divides the four stages we discussed above into even smaller steps, so the time spent on each stage will be more balanced.
- Another way to address the structural hazards is waiting. Wait till the required resource becomes available. Actually, all hazards can be resolved by waiting. We call the waiting “stall.”

Data Hazards
- Data hazards occur in executing arithmetic or data transfer operations, when the next instruction needs to wait for the previous instruction to complete its data read/write.

- For example, if we have the following two instructions:
  
  LOAD R1 A # A => R1
  SUB R0 R1 R2 # R0 - R1 => R2

  • R1 used by SUB won’t be ready till after executing LOAD R1 A.
  • So, we have to stall one stage for the SUB instruction. The pipelining for these two instructions look like
    
    load F D E W
    
    sub F D E W

  • We also call this type of data hazard “load-use data hazard,” which is a specific form of data hazard, in which the date being loaded by a load instruction has not yet become available when it is needed by another instruction.