Lecture 29-30: April 30, 2021
Computer Architecture and Organization-I
Biplab K Sikdar

0.6 Reservation Table

Reservation table describes occupancy of pipeline stages in different clock cycles.

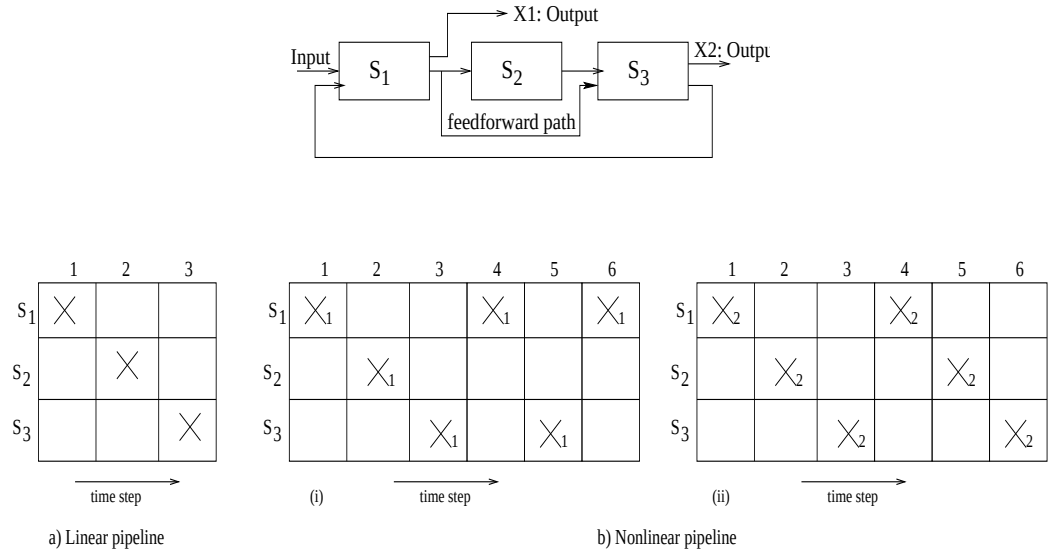


Figure 22: Reservation table a) Linear pipeline b) Nonlinear pipeline

Row: represents resource/pipeline stage. Column: time-slice/pipeline cycle.

Reservation tables of Figure 22 represents a 3-stage (S_1, S_2, and S_3) pipeline system.

Linear function X (Figure 22(a)) utilizes each of stages S_1, S_2, and S_3 only once.

Multiple entries in a row denote : feedback paths.

A pipeline configuration (hardware structure) can be utilized to realize multiple reservation tables (each corresponds to a pipeline function).

However, a reservation table corresponds to only one pipeline function.

Collision can occur in a nonlinear pipeline to realize a function X.

Collision: When two initiations of X try to use same resource at the same time.

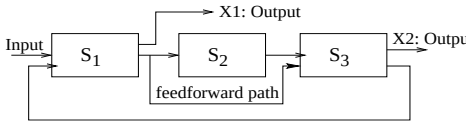


Figure 23 shows collision.

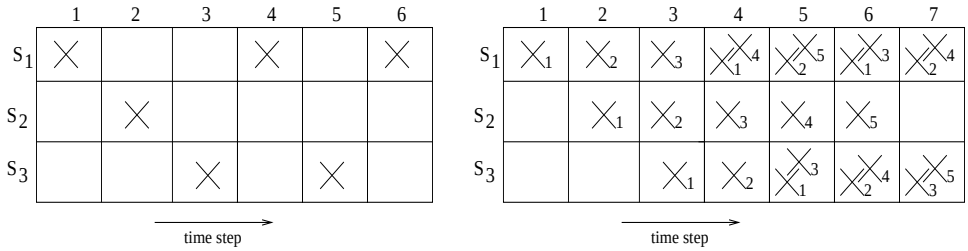


Figure 23: Collision in nonlinear pipeline

In linear pipeline, there is no collision (Figure 24).

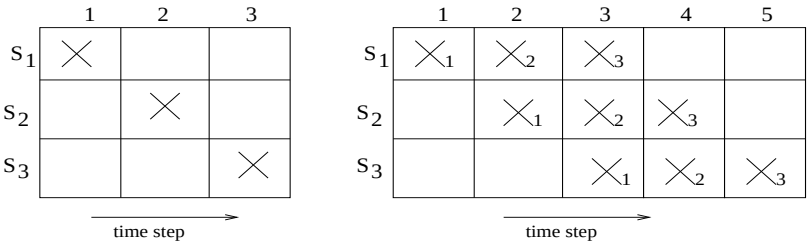


Figure 24: No collision in linear pipeline

0.6.1 Latency analysis

Latency: Number of time units between two initiations of an evaluation.

Forbidden latency: Latency that creates collisions.

Latency sequence: A sequence of permissible non-forbidden latencies between successive task initiations.

Latency cycle: A sequence that repeats itself. Example: latency cycle (1,5) represents infinite latency sequence 1, 5, 1, 5, 1, 5, ...

Latency cycle represents schedule of initiations avoiding collision.

Average latency of a cycle: Sum of latencies in the cycle divided by number of latencies. Example : average latency for (1,5) is 3. It is 3 for (3) also.

Demands for scheduling of initiations:

1. Collision free scheduling
2. Achieve the shortest average latency

0.6.2 Collision free scheduling

Target is to obtain minimal average latency (MAL). Steps taken to find MAL.

- Step 1. Identify forbidden latencies and set of permissible latencies (can be obtained from reservation table).

  In Figure 25, forbidden latencies : 2/3/5. 1 and 4 are permissible latencies.

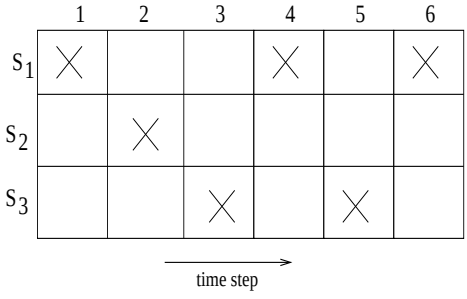


Figure 25: Nonlinear pipeline

- Step 2. Find collision vector C = ($C_d C_{d-1} C_{d-2} \cdots C_2 C_1$), $d \leq (n-1)$

  Where n is compute time (number of columns) of reservation table.

  C_i = 1, if latency i causes a collision (i is forbidden latency),

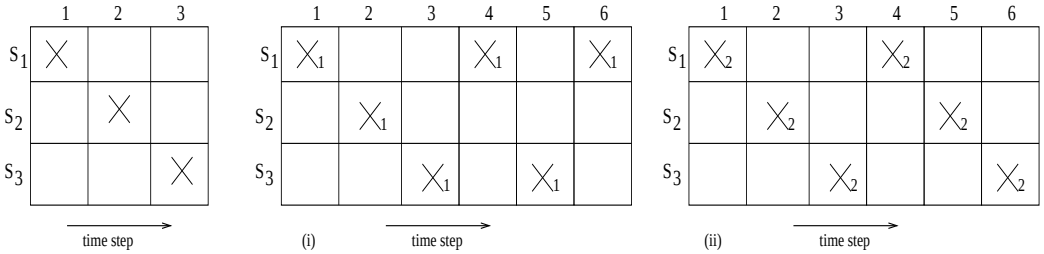
  C_i = 0, if latency i is permissible.

  Most significant (d^{th}) bit of C is always '1'.

  For Figure 25, C = 10110 (d = 5 = n-1 = 6 - 1).

  Maximum value of d can be the maximum value of forbidden latency.

  C corresponding to reservation table of Figure 22(b)(ii) is 100.



a) Linear pipeline b) Nonlinear pipeline

- Step 3. Contruct state diagram - permissible transitions - C as initial state.

  a. For each permissible latency p, compute new state for transition with p,

  (i) perform p-bit right shift of current state to get CS and then,

  (ii) perform OR of CS and C.

  c. Process (b) is continued for all possible states.

  CS defines latencies forbidden for I_{i+1} due to presence of I_{i-1}, I_{i-21}, ...

  For Figure 25, state diagram is Figure 26.

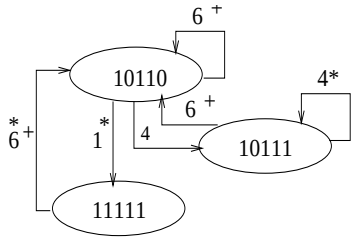


Figure 26: State diagram

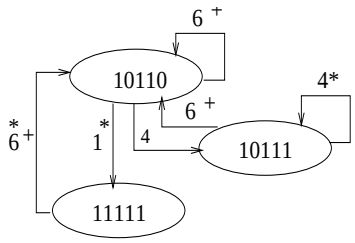
After p = 1 (permissible latency) the next state is

$$01011 + 10110 = 11111.$$

All d+1 latencies are permissible latencies.

When number of shifts $\geq$ d+1, all transitions are redirected to initial state.

'*' denotes minimum latency edge. In Figure 26, there are three such latencies.



- Step 4. Determine optimal latency cycles - minimal average latency (MAL).

  a) Find legitimate cycles. For Figure 26, the legitimate cycles are (1,6), (6), (4), (4,6), (6,1,6), (4,6,1,6) $\cdots$.

  b) Find simple cycles (a latency cycle in which a state appears only once). For this example, simple cycles are (1,6), (4), (6), (4,6).

  c) Find greedy cycles (whose edges are made with minimum latencies) from simple cycles. For this example, greedy cycles are (1,6) and (4).

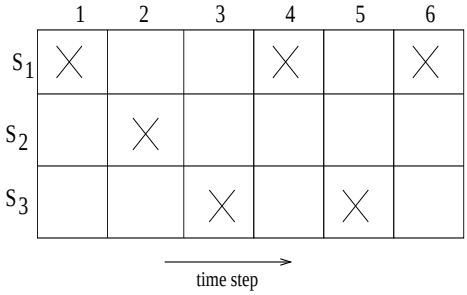
  d) Compute average latencies of greedy cycles. Here, these are 3.5 and 4.

  e) Find greedy cycles with MAL. For this example, MAL is 3.5.

However, greedy cycle is not sufficient to guarantee the optimality of MAL.

Lower bound of MAL - maximum number of checkmarks in any row of RT.

For this example, lower bound of MAL is 3 (Figure 25).



0.6.3 Scheduling optimization

Insert noncompute delay D in RT (keeping pipeline function intact) to yield an optimal greedy cycle for MAL (Figure 27(a)). Choice of delay: no specific algorithm.

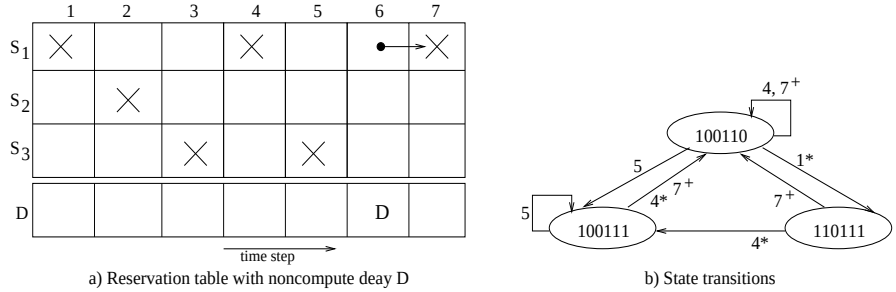


Figure 27: Insertion of noncompute delay

Now, forbidden latencies are 2, 3 and 6. Collision vector is C = 100110.

New state transition diagram is shown in Figure 27(b).

There is only one greedy cycle (1, 4, 4). MAL = $\frac{1+4+4}{3}$ = 3.

Configuration of new pipeline is in in Figure 28(b). Original one is in Figure 28(a).

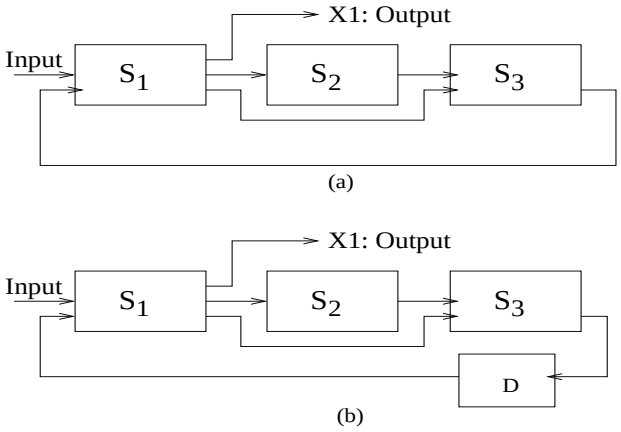


Figure 28: Pipeline architecture