EECS 583 – Class 11
Instruction Scheduling

University of Michigan

February 19, 2024

Announcements & Reading Material

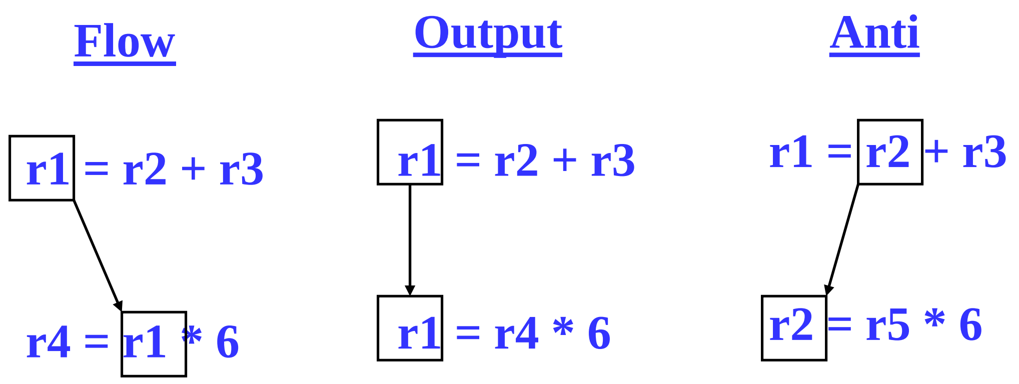
- HW 2 – Due Wed at midnight!
  - » See piazza for answered questions, Talk to Aditya/Yunjie for help
- Project discussion meetings (Mar 11-15)
  - » Project proposal meeting signup next next week – Signup on Google Calendar
    - Next week is spring break!
  - » Each group meets 10 mins with Aditya, Yunjie, and I
  - » Action items
    - Need to identify group members
    - Use piazza to recruit additional group members or express your availability
    - Think about general project areas that you want to work on
- Today’s class
  - » “The Importance of Prepass Code Scheduling for Superscalar and Superpipelined Processors,” P. Chang et al., IEEE Transactions on Computers, 1995, pp. 353-370.
- Next class
  - » “Iterative Modulo Scheduling: An Algorithm for Software Pipelining Loops”, B. Rau, MICRO-27, 1994, pp. 63-74.

From Last Time: Code Generation

- Map optimized "machine-independent" assembly to final assembly code
- Input code
  - Classical optimizations
  - ILP optimizations
  - Formed regions (sbs, hbs), applied if-conversion (if appropriate)
- Virtual → physical binding
  - 2 big steps
  - 1. Scheduling
    - Determine when every operation executions
    - Create MultiOps (for VLIW) or reorder instructions (for superscalar)
  - 2. Register allocation
    - Map virtual → physical registers
    - Spill to memory if necessary

Data Dependences

- Data dependences
  - If 2 operations access the same register, they are dependent
  - However, only keep dependences to most recent producer/consumer as other edges are transitively redundant
  - Types of data dependences



More Dependences

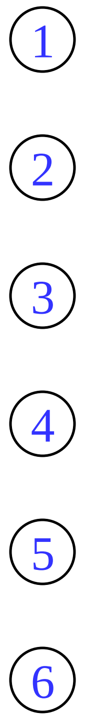
- Memory dependences
  - Similar as register, but through memory
  - Memory dependences may be certain or maybe
- Control dependences
  - We discussed this earlier
  - Branch determines whether an operation is executed or not
  - Operation must execute after/before a branch

$$\begin{tikzcd}
\underline{\textbf{Mem-flow}} & \underline{\textbf{Mem-output}} & \underline{\textbf{Mem-anti}} & \underline{\textbf{Control}} \\
\textbf{store (r1, r2)} \arrow[d] & \textbf{store (r1, r2)} \arrow[d] & \textbf{r2 = load(r1)} \arrow[d] & \textbf{if (r1 != 0)} \arrow[d] \\
\textbf{r3 = load(r1)} & \textbf{store (r1, r3)} & \textbf{store (r1, r3)} & \textbf{r2 = load(r1)}
\end{tikzcd}$$

Dependence Graph

- Represent dependences between operations in a block via a DAG
  - Nodes = operations/instructions
  - Edges = dependences
- Single-pass traversal required to insert dependences
- Example

```
        1: r1 = load(r2)
        2: r2 = r1 + r4
        3: store (r4, r2)
        4: p1 = cmpp (r2 < 0)
        5: branch if p1 to BB3
        6: store (r1, r2)
BB3:
```



Dependence Graph - Solution

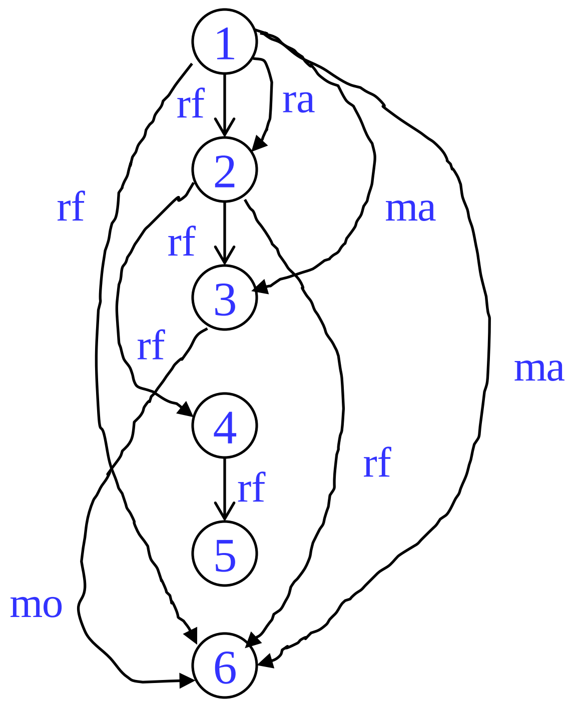
- Example

1: r1 = load(r2)
2: r2 = r1 + r4
3: store (r4, r2)
4: p1 = cmpp (r2 < 0)
5: branch if p1 to BB3
6: store (r1, r2)

BB3:

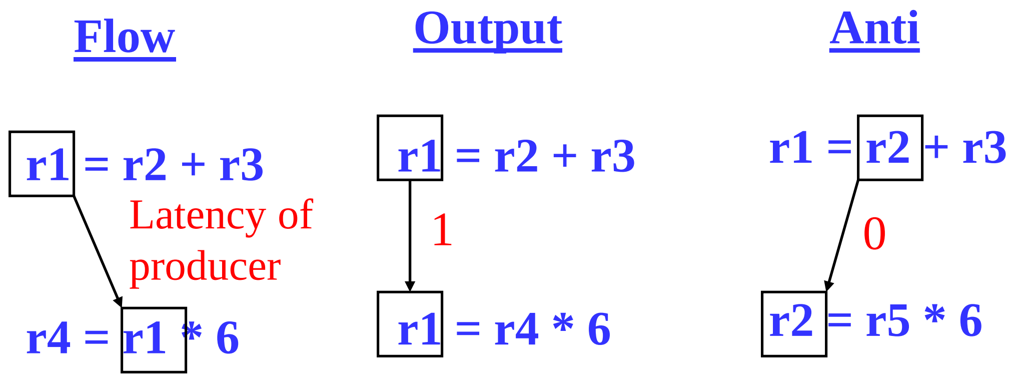
Instructions 1-4 have
control dependence to instruction 5

5→6 control dependence



Dependence Edge Latencies

- Edge latency = minimum number of cycles necessary between initiation of the predecessor and successor in order to satisfy the dependence
- Is negative latency possible?
  - Yes, means successor can start before predecessor
  - We will only deal with latency >= 0



Dependence Edge Latencies (2)

- Memory dependences
  - » Ordering memory instructions to ensure correct memory state
- Control dependences
  - » branch → b
    - Instructions inside then/else paths dependent on branch
  - » a → branch
    - Op a must be issued before the branch completes

Mem-flow

$$\begin{tikzcd} \text{store (r1, r2)} \arrow[d, "1"] \\ \text{r3 = load(r1)} \end{tikzcd}$$

Mem-output

$$\begin{tikzcd} \text{store (r1, r2)} \arrow[d, "1"] \\ \text{store (r1, r3)} \end{tikzcd}$$

Mem-anti

$$\begin{tikzcd} \text{r2 = load(r1)} \arrow[d, "1"] \\ \text{store (r1, r3)} \end{tikzcd}$$

Control

$$\begin{tikzcd} \text{r3 = r4 + r5} \arrow[d, "0"] \\ \text{if (r1 != 0)} \arrow[d, "1"] \\ \text{r2 = load(r1)} \end{tikzcd}$$

Class Problem – Add Latencies to Dependence Edges

latencies

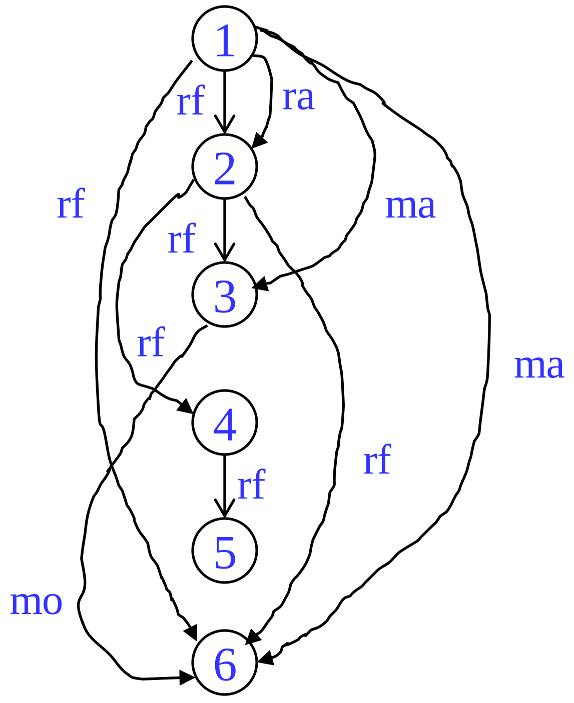
add:	1
cmpp:	1
load:	2
store:	1

❖ Example

1: r1 = load(r2)
2: r2 = r1 + r4
3: store (r4, r2)
4: p1 = cmpp (r2 < 0)
5: branch if p1 to BB3
6: store (r1, r2)
BB3:

Instructions 1-4 have control dependence to instruction 5

5→6 control dependence



Homework Problem 1

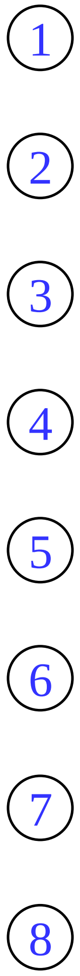
machine model

latencies

add:	1
mpy:	3
load:	2
store:	1

1. Draw dependence graph
2. Label edges with type and latencies

```
1. r1 = load(r2)
2. r2 = r2 + 1
3. store (r8, r2)
4. r3 = load(r2)
5. r4 = r1 * r3
6. r5 = r5 + r4
7. r2 = r6 + 4
8. store (r2, r5)
```



Homework Problem 1: Answer

machine model

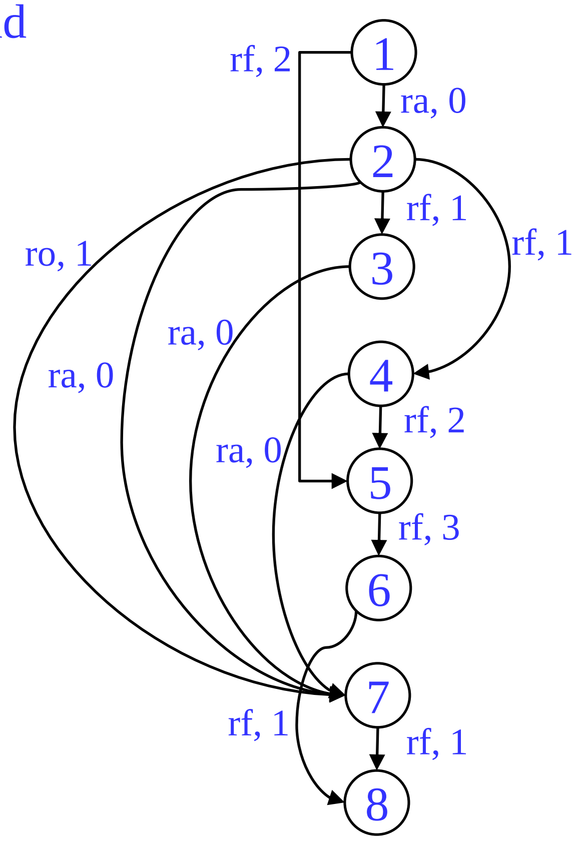
latencies

add:	1
mpy:	3
load:	2
store:	1

Store format (addr, data)

1. Draw dependence graph
2. Label edges with type and latencies

```
1. r1 = load(r2)
2. r2 = r2 + 1
3. store (r8, r2)
4. r3 = load(r2)
5. r4 = r1 * r3
6. r5 = r5 + r4
7. r2 = r6 + 4
8. store (r2, r5)
```

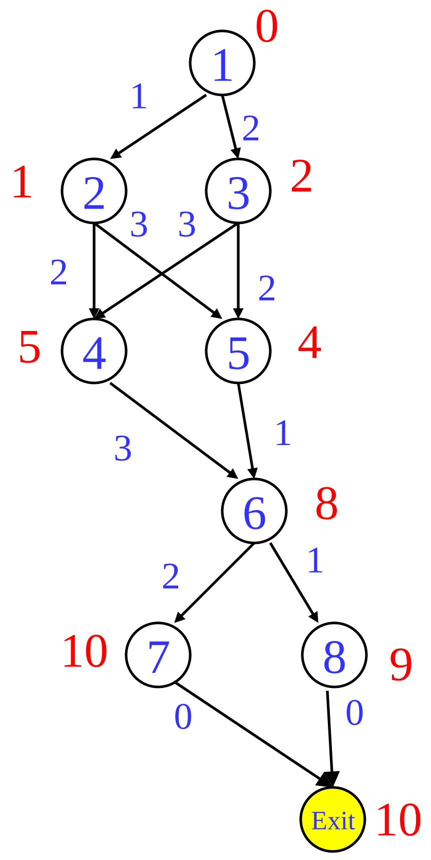


Memory deps all with latency =1: 1→3 (ma), 1→8 (ma), 3→4 (mf), 3→8 (mo), 4→8 (ma)

No control dependences

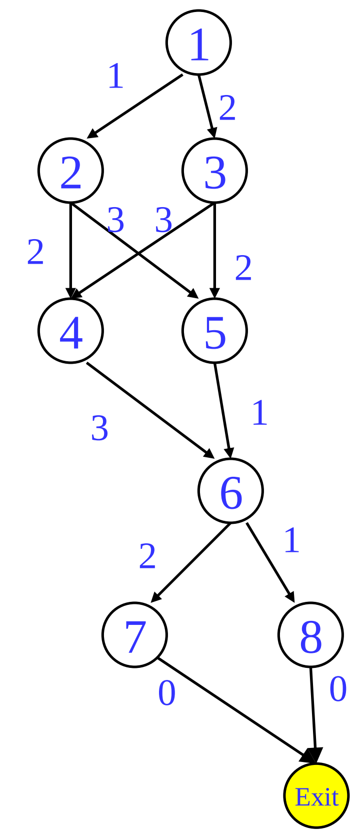
Dependence Graph Properties - Estart

- Estart = earliest start time, (as soon as possible - ASAP)
  - Schedule length with infinite resources (dependence height)
  - Estart = 0 if node has no predecessors
  - Estart = MAX(Estart(pred) + latency) for each predecessor node
  - Example



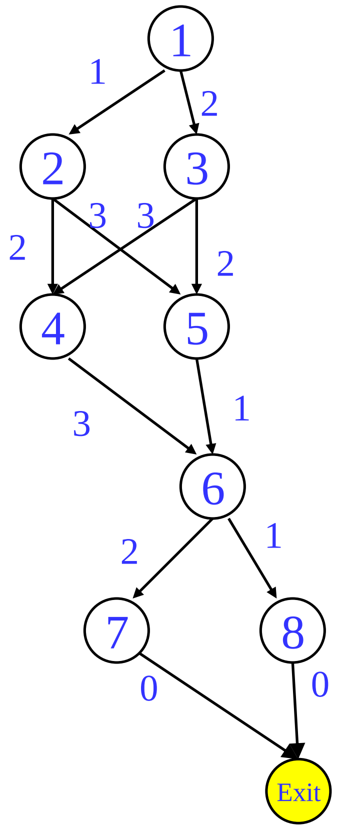
Lstart

- Lstart = latest start time, ALAP
  - Latest time a node can be scheduled s.t. sched length not increased beyond infinite resource schedule length
  - Lstart = Estart if node has no successors
  - Lstart = MIN(Lstart(succ) - latency) for each successor node
  - Example



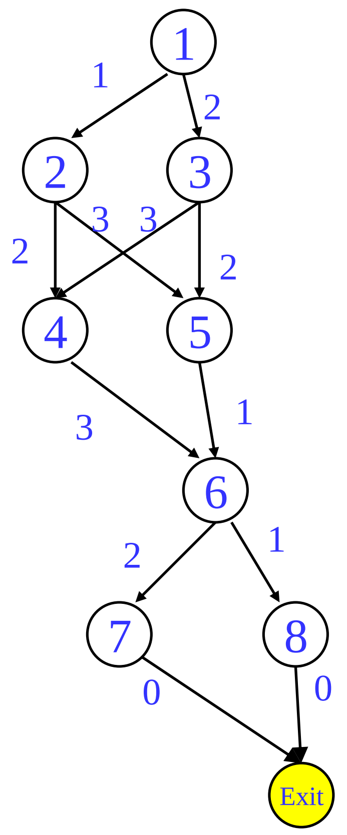
Slack

- Slack = measure of the scheduling freedom
  - Slack = Lstart – Estart for each node
  - Larger slack means more mobility
  - Example

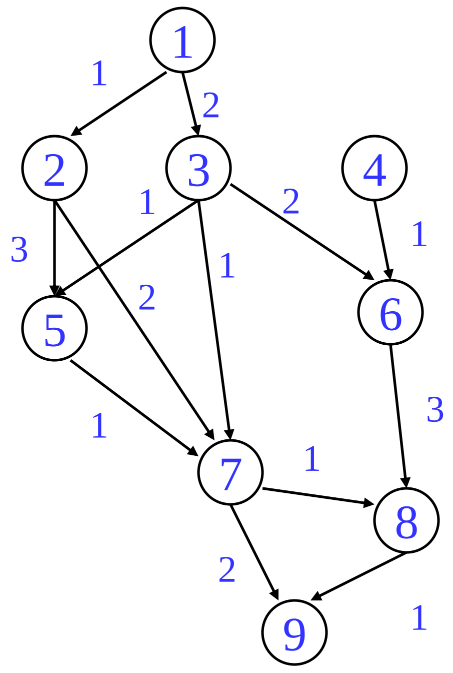


Critical Path

- Critical operations = Operations with slack = 0
  - No mobility, cannot be delayed without extending the schedule length of the block
  - Critical path = sequence of critical operations from node with no predecessors to exit node, can be multiple crit paths



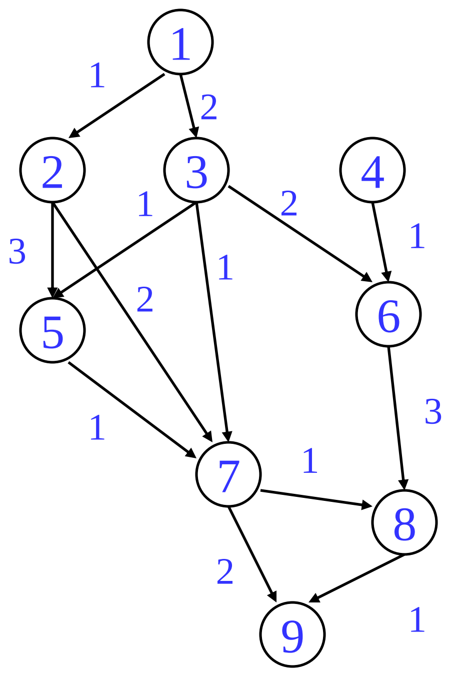
Homework Problem 2



Node	Estart	Lstart	Slack
1			
2			
3			
4			
5			
6			
7			
8			
9			

Critical path(s) =

Homework Problem 2 - Answer



Node	Estart	Lstart	Slack
1	0	0	0
2	1	2	2
3	2	2	0
4	0	3	3
5	4	5	1
6	4	4	0
7	5	6	1
8	7	7	0
9	8	8	0

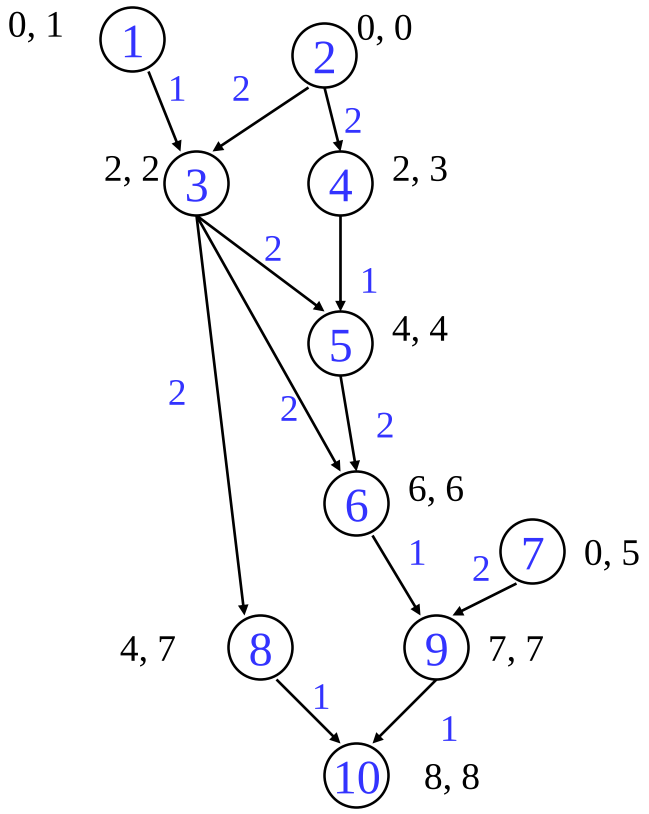
Critical path(s) = 1,3,6,8,9

Operation Priority

- Priority – Need a mechanism to decide which ops to schedule first (when you have multiple choices)
- Common priority functions
  - Height – Distance from exit node
    - Give priority to amount of work left to do
  - Slackness – inversely proportional to slack
    - Give priority to ops on the critical path
  - Register use – priority to nodes with more source operands and fewer destination operands
    - Reduces number of live registers
  - Uncover – high priority to nodes with many children
    - Frees up more nodes
  - Original order – when all else fails

Height-Based Priority

- Height-based is the most common
  - » priority(op) = MaxLstart – Lstart(op) + 1



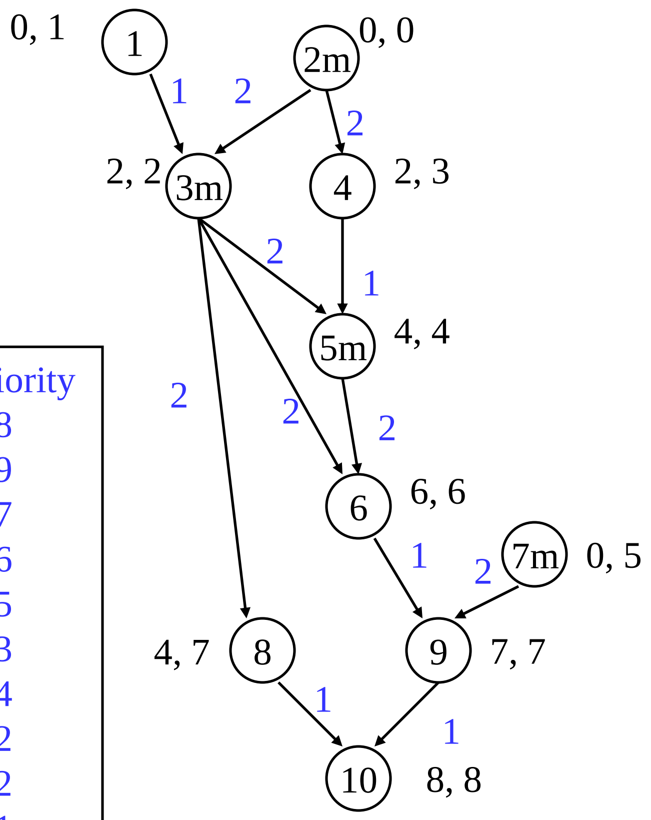
op	priority
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

List Scheduling (aka Cycle Scheduler)

- Build dependence graph, calculate priority
- Add all ops to UNSCHEDULED set
- time = -1
- while (UNSCHEDULED is not empty)
  - time++
  - READY = UNSCHEDULED ops whose incoming dependences have been satisfied
  - Sort READY using priority function
  - For each op in READY (highest to lowest priority)
    - op can be scheduled at current time? (are the resources free?)
      - Yes, schedule it, op.issue_time = time
        - Mark resources busy in RU_map relative to issue time
        - Remove op from UNSCHEDULED/READY sets
      - No, continue

Cycle Scheduling Example

Processor: 2 issue, 1 memory port, 1 ALU
Memory port = 2 cycles, pipelined
ALU = 1 cycle



op	priority
1	8
2	9
3	7
4	6
5	5
6	3
7	4
8	2
9	2
10	1

RU_map

time	ALU	MEM
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		

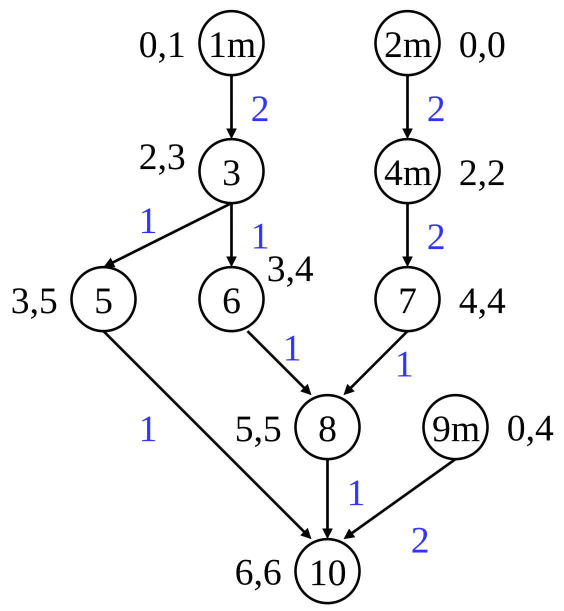
Schedule

time	Instructions
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

Time =
Ready =

Homework Problem 3

Processor: 2 issue, 1 memory port, 1 ALU
Memory port = 2 cycles, pipelined
ALU = 1 cycle



RU_map time	ALU	MEM
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		

Schedule time	Instructions
0	
1	
2	
3	
4	
5	
6	
7	
8	
9	

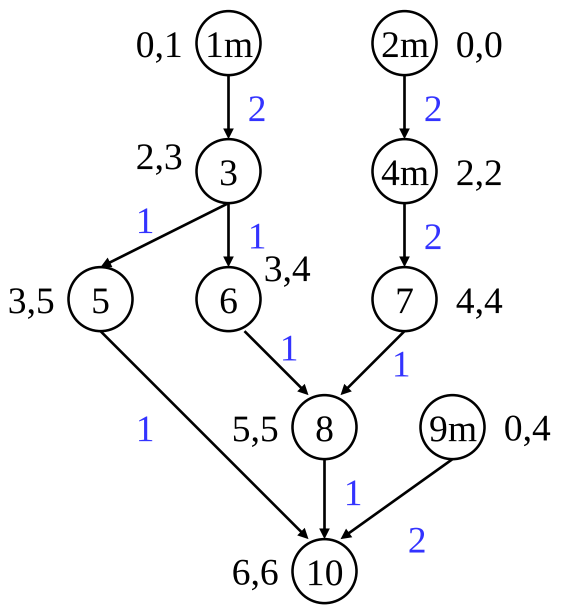
Time =
Ready =

1. Calculate height-based priorities
2. Schedule using cycle scheduler

Homework Problem 3 – Answer

Processor: 2 issue, 1 memory port, 1 ALU
Memory port = 2 cycles, pipelined
ALU = 1 cycle

Op	priority
1	6
2	7
3	4
4	5
5	2
6	3
7	3
8	2
9	3
10	1



RU_map

time	ALU	MEM
0		X
1		X
2		X
3	X	X
4	X	
5	X	
6	X	
7	X	
8	X	

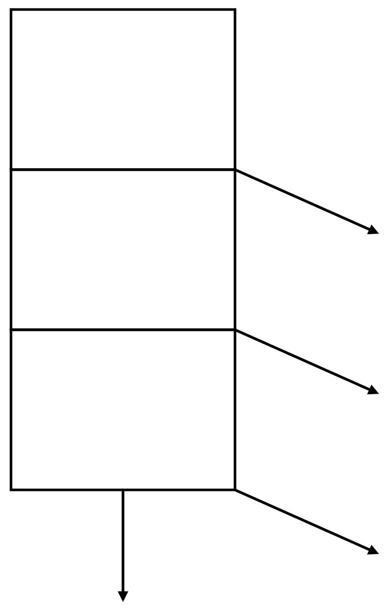
Schedule

Time	Instructions
0	2
1	1
2	4
3	3, 9
4	6
5	7
6	5
7	8
8	10

1. Calculate height-based priorities
2. Schedule using Operation scheduler

Generalize Beyond a Basic Block

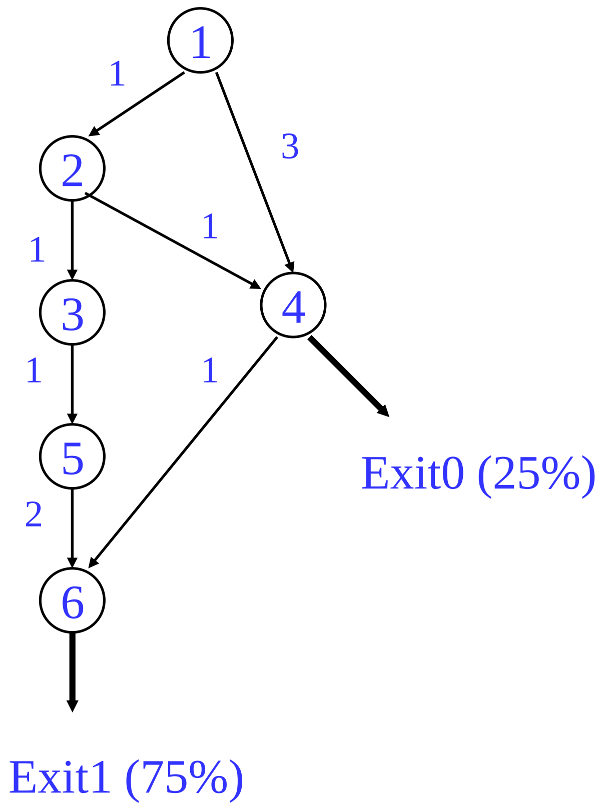
- Superblock
  - Single entry
  - Multiple exits (side exits)
  - No side entries
- Schedule just like a BB
  - Priority calculations needs change
  - Dealing with control deps



Lstart in a Superblock

- Not a single Lstart any more
  - 1 per exit branch (Lstart is a vector!)
  - Exit branches have probabilities

op	Estart	Lstart0	Lstart1
1			
2			
3			
4			
5			
6			



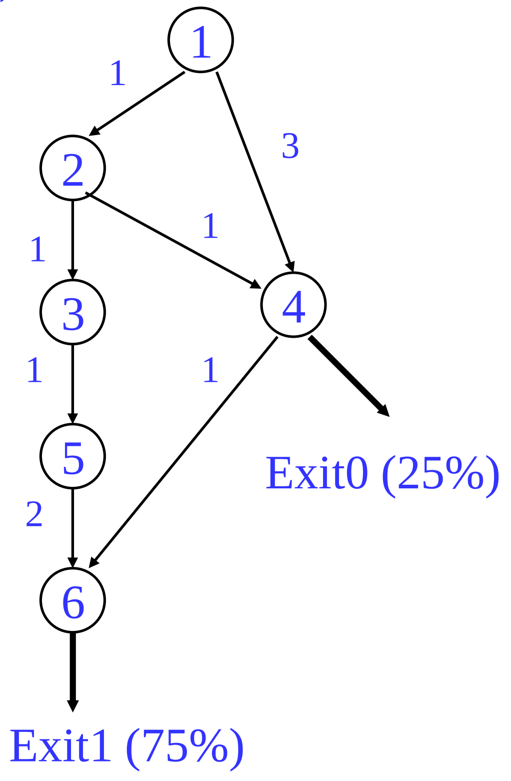
Operation Priority in a Superblock

- Priority – Dependence height and speculative yield
  - Height from op to exit * probability of exit
  - Sum up across all exits in the superblock

Priority(op) = SUM(Probi * (MAX_Lstart – Lstarti(op) + 1))

valid late times for op

op	Lstart0	Lstart1	Priority
1			
2			
3			
4			
5			
6			



To Be Continued