Lab 3: Improving the Quality of Schedules

Comp 412

Material from the Performance Tutorial
Correctness

Before we worry about performance, the scheduler must produce correct code.

• As before, correct code produces the same output as the original code

• With the lab3 simulator, you must be careful about the –s (stall) flag
  – Original code runs with the –s 3 flag
  – Scheduled code runs with the –s 1 flag

If you don’t specify –s 1, you will think the code is correct when it is not

Your scheduler must insert NOPs to fill idle cycles.

• If the ready queue is empty in some cycle, emit a NOP

If you are getting the right answers with –s 3 and the wrong answers with -s 1, check the logic for moving items onto the ready queue and for emitting those NOPs.

As always, you can compare the values at each step by running the simulator with the trace option (-t) on both the original code and the scheduled code.
Performance: What does it mean?

In Lab 3, 50% of the code grade is based on the speed of the code that your scheduler produces

• We will measure performance using the **Lab 3 ILOC Simulator**
  
  – Take the original code & run it, -s 3
  – Take the scheduled code & run it, -s 1
  – Make sure that the answers are identical
  – Difference between the two runs is the improvement due to scheduling

• The goal of Lab 3 is to maximize the improvement due to scheduling
  
  – The 50% credit for performance is based on how your scheduler does relative to the schedulers of other students (& relative to the reference implementation — which can be beaten)
  – Attention to detail can improve your scheduler
  – **But**, the problem is NP-Complete, so you shouldn’t expect to win every time

---

1 The Lab 3 simulator differs from the Lab 2 simulator. Be sure to use the Lab 3 simulator and to measure the code’s speed with the correct interlock setting, “-s 1”
What Affects Schedule Quality?

What goes into creating a good schedule?

• A good priority function
  – Priorities determine which operation is chosen when > 1 is ready
  – There are many possibilities

• Tie-breaking
  – how does scheduler choose between two equal-priority ops?
  – Multiple levels of priority function, applied in a carefully chosen order or a weighted linear combination of multiple priority functions
  – Do not write a report that says you use random tie breaking as spin for “I did nothing about tie breaking and took whatever was first.”

• An accurate dependence graph
  – Need to represent every necessary dependence
  – Should avoid dependences that are not necessary

• Must it be complicated?
  – No. The reference scheduler uses a fairly simple set of heuristics.
What Affects Schedule Quality?

Priority Functions

• Priorities determine which operation is chosen when > 1 is ready

• Many priority functions are recommended in the literature
  – Latency-weighted depth
  – Most descendants in the graph
  – Breadth-first numbering
  – Depth-first numbering
  – Boost the priority of operations that are restricted to one functional unit
  – Boost the priority of loads over stores

• Cannot make up your mind?
  – Try a linear combination of two priority functions  \((\alpha \times \text{prio}_1 + \beta \times \text{prio}_2)\)
  – At that point, if you have time, you can try tuning \(\alpha\) & \(\beta\)
Tie Breaking

Picking an operation from the *Ready* queue is critically important

Among the possible considerations:

- Operations restricted to one or the other of the functional units
  - Should it give priority to a restricted operation, or not?
- Overall progress toward the goal of scheduling all operations
- The impact of preferring one operation over another

Time tested (& sometimes contradictory) ideas

- Boost priority of multi-cycle operations (multiply, load & store)
- Boost priority of operations that have more successors
- Balance priority along all paths toward the leaves
  - Breadth first may create more ILP and, hence, fill more empty slots
- Boost priority of operations along the critical path
  - The only thing that matters is the critical path through the block

Tie Breaking

Picking an operation from the *Ready* queue is critically important

Among the possible considerations:

- Operations restricted to one or the other of the functional units
  - Should it give priority to a restricted operation, or not?
- Overall progress toward the goal of scheduling all operations
- The impact of preferring one operation over another

Encode tie breaking directly into your priority function

- Go back to the $\alpha$-$\beta$ tuning idea
  - Fix $\alpha + \beta = 1$ and try a parameter sweep with increments of 0.1 or 0.05.
- For tie breaking, you could bias strongly toward one metric
  - e.g., $\alpha = 1000$ and $\beta = 1$
  - Result is a number that breaks ties in $\text{prio}_1$ with values from $\text{prio}_2$
  - Simple arithmetic replacing complicated nests of *if-then-else* statements
Tie Breaking

Picking an operation from the *Ready* queue is critically important

Among the possible considerations:

- Operations restricted to one or the other of the functional units
  - Should it give priority to a restricted operation, or not?
- Overall progress toward the goal of scheduling all operations
- The impact of preferring one operation over another

Encode tie breaking directly into your priority function

- Go back to the priority function
  - Fix $\alpha + \beta = 1$ and try a parameter sweep with increments of 0.1 or 0.05.
  - For tie breaking, you could bias strongly toward one metric
    - e.g., $\alpha = 1000$ and $\beta = 1$
    - Result is a number that breaks ties in $\text{prio}_1$ with values from $\text{prio}_2$
    - Simple arithmetic replacing complicated nests of *if-then-else* statements

There is no single answer.
The problem is NP-Complete

COMP 412, Fall 2019
What Affects Schedule Quality?

What goes into creating a good schedule?

• A good priority function
  – Priorities determine which operation is chosen when > 1 is ready
  – There are many possibilities

• Tie-breaking
  – how does scheduler choose between two equal-priority ops?
  – Multiple levels of priority function, applied in a carefully chosen order or a weighted linear combination of multiple priority functions

• An accurate dependence graph
  – Need to represent every necessary dependence
  – Should avoid dependences that are not necessary

• Must it be complicated?
  – No. The reference scheduler uses a fairly simple set of heuristics.
2. Build a dependence graph to capture critical relationships in the code

Explanatory Notes

1. ‘o’ refers to both the operation in the original block and the node that represents it in the graph. The meaning should be clear by context.

2. At a minimum, in the absence of other information:
   - load & output need an edge to the most recent store (full latency)
   - output needs an edge to the most recent output (serialization)
   - store needs an edge to the most recent store & output, as well as each previous load (serialization)

If your lab tries to simplify the graph, it may need more edges for store & output nodes

Building the Graph

Create an empty map, $M$
walk the block, top to bottom
at each operation $o$ that defines $VR_i$:
  - create a node for $o$
  - set $M(VR_i)$ to $o$
for each $VR_j$ used in $o$, add an edge from $o$ to the node in $M(VR_j)$
if $o$ is a load, store, or output operation, add edges to ensure serialization of memory ops

$O(n + m^2)$ where $m$ is $|\text{memory ops}|$
Dependence Among Memory Operations

The key point is simple:  

*The original code is the specification for the computation*

- The order of operations in the original code is correct
- Any legal schedule must produce equivalent results

Consequences

- Loads & outputs that precede a store in the original code must precede it in the scheduled code
  - Unless the scheduler can **prove** that the operations are disjoint
    → *Store never touches the same memory location as the load or output*
- Stores that precede a load or output in the original code must precede them in the scheduled code
  - Unless the scheduler can **prove** that the operations are disjoint
- Outputs must be strictly ordered, according to the original code
  - The order of execution of outputs is critical to correctness
The key point is simple:  

*The original code is the specification for the computation*

- The order of operations in the original code is correct
- Any legal schedule must produce equivalent results

Consequences

- To preserve these relationships, you will add a lot of edges to the graph
  - Those edges constrain the schedule, of course ...
  - If they constrain the schedule needlessly, that hurts performance
- Many students implement some kind of graph simplification phase
  - Limits on what you can do (arbitrary, dictatorial limits)
  - Limits on what the analysis can detect
Dependence Among Memory Operations

Example

1. loadI 8 => vr0
2. loadI 12 => vr4
3. load vr0 => vr3
4. load vr4 => vr5
5. add vr0, vr3 => vr2
6. store vr2 => vr4
7. output 8
8. sub vr2, vr3 => vr1
9. store vr1 => vr0
10. output 8

Original Code

Dependence Graph

COMP 412, Fall 2019
Dependence graph for the original code

Dependence Among Memory Operations

Example

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

Original Code

Dependence Graph
Dependence Among Memory Operations

**Example**

Interlock settings: memory registers branches

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

**Original Code**

Original code, executed with full (-s 3) interlocks

Executed 10 instructions and 10 operations in 17 cycles.


-s 3 can tell that this output is independent of the prior store because r2 (12) ≠ 8

Output generates => 20

Output generates => 8
Dependence Among Memory Operations

Example

Interlock settings: memory registers branches

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

Original Code

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 (addr: 8) => r3 (20)
4. load r2 (addr: 12) => r4 (30)
5. stall
6. stall
7. add r1 (8), r3 (20) => r5 (28)
8. store r5 (28)
9. output 8 (20)
10. sub r5 (28), r3 (20) => r6 (8)
11. store r6 (8) => r1 (addr: 8)
12. stall
13. stall
14. stall
15. stall
16. output 8 (8)

Original code, executed with full (-s 3) interlocks


-executed 10 instructions and 10 operations in 17 cycles.

-s 3 can tell that this output is dependent on the prior store because r1 (8) = 8

output generates => 20
output generates => 8
Dependence Among Memory Operations

Example

Interlock settings: memory registers branches

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

Original Code

loadI 8 => r1
loadI 12 => r2
load r1 => r3
load r2 => r4
add r1, r3 => r5
store r5 => r2
output 8
sub r5, r3 => r6
store r6 => r1
output 8

Executed 10 instructions and 10 operations in 17 cycles.
Dependence Among Memory Operations

**Example**

1. `loadI 8 => r1`
2. `loadI 12 => r2`
3. `load r1 => r3`
4. `load r2 => r4`
5. `add r1, r3 => r5`
6. `store r5 => r2`
7. `output 8`
8. `sub r5, r3 => r6`
9. `store r6 => r1`
10. `output 8`

**Original Code**

```
diana% ..//412sched -n -s Lab3Perf-2.i
[loadI 8 => r0; loadI 12 => r4]
[load r0 => r3; nop]
[load r4 => r5; nop]
[nop; nop]
[nop; nop]
[nop; nop]
[add r0, r3 => r2; nop]
[store r2 => r4; sub r2, r3 => r1]
[nop; nop]
[nop; nop]
[nop; nop]
[output 8; nop]
[store r1 => r0; nop]
[nop; nop]
[nop; nop]
[nop; nop]
[nop; nop]
[output 8; nop]
diana%
```

Scheduler cannot tell whether or not `r4 = 8`, so it must insert `nops`. 
Scheduled code, no graph simplification

Dependence Among Memory Operations

Example

```
diana% ..../412sched -n -s Lab3Perf-2.i
[loadI 8 => r0; loadI 12 => r4]
[load r0 => r3; nop]
[load r4 => r5; nop]
[add r0, r3 => r2; nop]
[store r2 => r4; sub r2, r3 => r1]
[store r1 => r0; nop]
[output 8; nop]
[output 8; nop]
[output 8; nop]
[output 8; nop]

r0 = 8, whether or not the scheduler knows, so it must insert nops.
```

Execution takes 19 cycles rather than 17 cycles

(lab3_ref added 4 nops, but scheduled 2 ops in cycles 0 and 7)
Dependence Among Memory Operations

Example

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

Original Code

Dependence Graph

Anti-dependences or serialization edges
Dependence Among Memory Operations

**Example**

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

**Original Code**

**Dependence Graph**

Notice the edge from 7 to 6
What can a simplifier do?

Dependence Among Memory Operations

Example

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

Dependence Graph

Graph After Simplification

The edge (7,6) is gone, so nodes moved.

Original Code

COMP 412, Fall 2019
What can a simplifier do?

Dependence Among Memory Operations

Example

1. loadI 8 => vr0
2. loadI 12 => vr4
3. load vr0 => vr3
4. load vr4 => vr5
5. add vr0, vr3 => vr2
6. store vr2 => vr4
7. output 8
8. sub vr2, vr3 => vr1
9. store vr1 => vr0
10. output 8

Original Code

Dependence Graph

True dependences that -s 3 proved false — same two edges
What can a simplifier do?

Dependence Among Memory Operations

Example

1. loadI 8 => r1
2. loadI 12 => r2
3. load r1 => r3
4. load r2 => r4
5. add r1, r3 => r5
6. store r5 => r2
7. output 8
8. sub r5, r3 => r6
9. store r6 => r1
10. output 8

Original Code

In this example, the removed anti-dependence edges did not improve the speed of the code. In some cases, they will.
What can a simplifier do?

Dependence Among Memory Operations

Example

Interlock settings: branches

1  load 8 => r1
2  load 12 => r2
3  load r1 => r3
4  load r2 => r4
5  add  r1, r3 => r5
6  store r5 => r2
7  output 8
8  sub  r5, r3 => r6
9  store r6 => r1
10 output 8

Original Code

Executed 14 instructions and 28 operations in 14 cycles.
**Warning:** Your scheduler must not perform optimizations other than instruction scheduling and register renaming. It must not remove useless (dead) instructions; it must not fold constants. Any instruction left in the code by the optimizer is presumed to have a purpose. Assume that the optimizer had a firm basis for not performing the optimization. Your lab should schedule the instructions, not search for other opportunities for improvement. *This lab is an investigation of instruction scheduling, not a project in general optimization algorithms.*

Additionally, your scheduler must not rely on input constants specified in comments. When the TAs grade your scheduler, they are not required to use the input specified in the comments. Your scheduler is expected to generate correct code regardless of whether or not the TAs use the input specified in the comments.

If the result of an op is not used, you cannot delete the op.
Simplifying the Dependence Graph

A simplifier does not need to be complex

for each IOEdge $E = <src, sink>$
  if $op(src)$ is independent of $op(sink)$
  then delete edge $E$

The Reference Scheduler’s Algorithm

Proving independence is the hard part

• Simple comparison of $loadI$ gets many of the test cases
• You may do arithmetic on the values from $loadIs$
• You may do algebra on expressions involving $loadIs$
  – e.g., $@a + 4 \neq @a + 12$
• You may not rely on unknown values
  – $MEM[12]$ may or may not be identical to $MEM[32]$
• You may not fold constants in the code (see prev slide)

Algebra Example

<table>
<thead>
<tr>
<th>Operation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>loadI 1024</td>
<td>$r0$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>addI $r0,4$</td>
<td>$r1$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>store $r0$</td>
<td>$r1$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>addI $r0,12$</td>
<td>$r2$</td>
</tr>
<tr>
<td>load $r2$</td>
<td>$r3$</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

Are the load and the store independent?