Review

- **Controller**
  - Tells universal datapath how to execute each instruction

- **Instruction timing**
  - Set by instruction complexity, architecture, technology
  - Pipelining increases clock frequency, “instructions per second”
    - But does not reduce time to complete instruction

- **Performance measures**
  - Different measures depending on objective
    - Response time
    - Jobs / second
    - Energy Efficiency (joules/operation, joules/instruction)
Processor

Processor-Memory Interface

Processor

Control

Datapath

PC

Registers

Arithmetic & Logic Unit (ALU)

Memory

Enable?
Read/Write

Address
Write Data

Read Data

Program

Bytes

Data
Pipelining Overview (Review)

- Pipelining doesn’t help \textit{latency} of single task, it helps \textit{throughput} of entire workload.
- Multiple tasks operating simultaneously using different resources.
- Potential speedup = Number pipe stages.
- Time to “fill” pipeline and time to “drain” it reduces speedup: 2.3X v. 4X in this example.
  - With lots of laundry, approaches 4X.

\textit{"wash, dry, fold, and put away"}
Agenda

- RISC-V Pipeline
- Pipeline Control
- Hazards
  - Structural
  - Data
    - R-type instructions
    - Load
- Control
- Superscalar processors
Recap: Pipelining with RISC-V

Instruction sequence:
- add t0, t1, t2
- or t3, t4, t5
- sll t6, t0, t3

**Single Cycle**

<table>
<thead>
<tr>
<th>Timing</th>
<th>$t_{step} = 100 \ldots 200$ ps</th>
<th>$t_{cycle} = 200$ ps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register access</td>
<td>only 100 ps</td>
<td>All cycles same length</td>
</tr>
<tr>
<td>Instruction time</td>
<td>$t_{instruction} = t_{cycle} = 800$ ps</td>
<td>$1000$ ps</td>
</tr>
<tr>
<td>Clock rate, $f_s$</td>
<td>$1/800$ ps = $1.25$ GHz</td>
<td>$1/200$ ps = $5$ GHz</td>
</tr>
<tr>
<td>Relative speed</td>
<td>$1$ x</td>
<td>$4$ x</td>
</tr>
</tbody>
</table>
RISC-V Pipeline

instruction sequence

add t0, t1, t2
or t3, t4, t5
slt t6, t0, t3
sw t0, 4(t3)
lw t0, 8(t3)
addi t2, t2, 1

t_{cycle} = 200 \text{ ps}

t_{instruction} = 1000 \text{ ps}

Resource use of instruction over time

Resource use in a particular time slot
Single-Cycle RISC-V RV32I Datapath
Pipelining RISC-V RV32I Datapath

Instruction Fetch (F)

Instruction Decode/Register Read (D)

ALU Execute (X)

Memory Access (M)

Write Back (W)
Pipelined RISC-V RV32I Datapath

Recalculate PC+4 in M stage to avoid sending both PC and PC+4 down pipeline

Must pipeline instruction along with data, so control operates correctly in each stage
Each stage operates on different instruction

Pipeline registers separate stages, hold data for each instruction in flight
Agenda

• RISC-V Pipeline
• **Pipeline Control**

• Hazards
  • Structural
  • Data
    • R-type instructions
    • Load

• Control

• Superscalar processors
Pipelined Control

- Control signals derived from instruction
  - As in single-cycle implementation
  - Information is stored in pipeline registers for use by later stages
Question:

Pipelining the single-cycle processor can increase processor performance by:

\[
\frac{\text{time}}{\text{program}} = \frac{\text{instructions}}{\text{program}} \cdot \frac{\text{cycles}}{\text{instruction}} \cdot \frac{\text{time}}{\text{cycle}}
\]

<table>
<thead>
<tr>
<th></th>
<th>Instructions /program</th>
<th>Cycles/instruction</th>
<th>Time/cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>decrease</td>
<td>decrease</td>
<td>same</td>
</tr>
<tr>
<td>B</td>
<td>same</td>
<td>increase</td>
<td>decrease</td>
</tr>
<tr>
<td>C</td>
<td>same</td>
<td>same</td>
<td>decrease</td>
</tr>
<tr>
<td>D</td>
<td>increase</td>
<td>decrease</td>
<td>increase</td>
</tr>
</tbody>
</table>
Hazards Ahead
Agenda

- RISC-V Pipeline
- Pipeline Control
- Hazards
  - **Structural**
- Data
  - R-type instructions
  - Load
- Control
- Superscalar processors
Structural Hazard

- **Problem:** Two or more instructions in the pipeline compete for access to a single physical resource

- **Solution 1:** Instructions take turns to use resource, some instructions have to stall

- **Solution 2:** Add more hardware to machine

- *Can always solve a structural hazard by adding more hardware*
Regfile Structural Hazards

- Each instruction:
  - can read up to two operands in decode stage
  - can write one value in writeback stage
  - *therefore two different instructions might be accessing the register file on the same cycle!*

- Avoid structural hazard by having separate “ports”
  - two independent read ports and one independent write port

- *Reads from one instruction and writes from another can happen simultaneously*
Structural Hazard: Memory Access

- Instruction and data memory used simultaneously
  - Use two separate memories

```
add t0, t1, t2
sw t0, 4(t5)
slt t6, t0, t3
or t3, t4, t5
lw t0, 8(t3)
```
Instruction and Data Caches

Caches: relatively small and fast “buffer” memories
Structural Hazards – Summary

- Conflict for use of a resource
- In RISC-V pipeline with a single memory
  - Load/store requires data access
  - Without separate memories, instruction fetch would have to stall for that cycle
    - All other operations in pipeline would have to wait
- Pipelined datapaths require separate instruction/data memories
  - Or at least separate instruction/data caches
- Multi-ported register file
- RISC ISAs (including RISC-V) designed to avoid structural hazards
  - e.g. at most one memory access/instruction
  - limited operands per instruction
Agenda

• RISC-V Pipeline
• Pipeline Control
• Hazards
  • Structural
  • Data
    • R-type instructions
    • Load
• Control
• Superscalar processors
Data Hazard: Register Access

- Separate ports, but what if write to same value as read?
- Does `sw` in the example fetch the old or new value?

```
add t0, t1, t2
or t3, t4, t5
slt t6, t4, t3
sw t0, 4(t3)
lw t0, 8(t3)
```
Register Access Policy

- Exploit high speed of register file (100 ps)
  1) WB updates value
  2) ID reads new value
- Indicated in diagram by shading

Might not always be possible to write then read in same cycle, especially in high-frequency designs.
Data Hazard: ALU Result

s0 holds “5” then add instr changes s0 to “9”

Value of s0

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>5/9</th>
<th>9</th>
<th>9</th>
<th>9</th>
<th>9</th>
</tr>
</thead>
</table>

- add s0, t0, t1
- sub t2, s0, t0
- or t6, s0, t3
- xor t5, t1, s0
- sw s0, 8(t3)

Without some fix, sub and or will calculate wrong result!
Solution 1: Stalling

- Problem: Instruction depends on result from previous instruction
  - `add s0, t0, t1`
  - `sub t2, s0, t3`

- Bubble:
  - stall the dependent instruction
  - effectively NOP: affected pipeline stages do “nothing”
Stalls and Performance

- Stalls reduce performance
- But stalls might be required to get correct results
- Compiler could try to arrange code to avoid hazards and stalls
- Requires knowledge of the pipeline structure
Solution 2: Forwarding

Value of $t_0$
- add $t_0$, $t_1$, $t_2$
- or $t_3$, $t_0$, $t_5$
- sub $t_6$, $t_0$, $t_3$
- xor $t_5$, $t_1$, $t_0$
- sw $t_0$, 8(t3)

Forwarding: grab operand from pipeline stage, rather than register file
Forwarding (aka Bypassing)

- Access result before it is stored in a register
- Requires extra connections in the datapath
1) Detect Need for Forwarding (example)

- Add $t0$, $t1$, $t2$
- Or $t3$, $t0$, $t5$
- Sub $t6$, $t0$, $t3$

Compare destination of older instructions in pipeline with sources of new instruction in decode stage.
Must ignore writes to $x0$!
Example Forwarding Path

Forwarding Control Logic

Same idea extends to rs2, and to instruction inst_D, inst_M pairing
Agenda

• RISC-V Pipeline
• Pipeline Control
• Hazards
  • Structural
  • Data
    • R-type instructions
    • Load
• Control
• Superscalar processors
Load Data Hazard

1 cycle stall unavoidable
forward
unaffected
Stall Pipeline

Program execution order (in instructions)

- `lw $2, 20($1)`
- and **becomes** `nop`
- `and $4, $2, $5`
- `or $8, $2, $6`
- `add $9, $4, $2`

Stall

Repeat `and` and forward
Data Hazard

- Slot after a load is called a load delay slot
- If that instruction uses the result of the load, then the hardware will stall for one cycle
- Equivalent to inserting an explicit \texttt{nop} in the slot
  - except the latter uses more code space
- Performance loss!

Idea:

- Put unrelated instruction into load delay slot
- No performance loss!
Code Scheduling to Avoid Stalls

- Reorder code to avoid use of load result in the next instruction!
- RISC-V code for \( D=A+B; \ E=A+C; \)

Original Order:

\[
\begin{align*}
\text{lw} & \quad t1, 0(t0) \\
\text{lw} & \quad t2, 4(t0) \\
\text{add} & \quad t3, t1, t2 \\
\text{sw} & \quad t3, 12(t0) \\
\text{lw} & \quad t4, 8(t0) \\
\text{add} & \quad t5, t1, t4 \\
\text{sw} & \quad t5, 16(t0)
\end{align*}
\]

Alternative:

\[
\begin{align*}
\text{lw} & \quad t1, 0(t0) \\
\text{lw} & \quad t2, 4(t0) \\
\text{lw} & \quad t4, 8(t0) \\
\text{add} & \quad t3, t1, t2 \\
\text{sw} & \quad t3, 12(t0) \\
\text{add} & \quad t5, t1, t4 \\
\text{sw} & \quad t5, 16(t0)
\end{align*}
\]

13 cycles

11 cycles
Cat Break: From Thomas Thurston
Agenda

• RISC-V Pipeline
• Pipeline Control
• Hazards
  • Structural
  • Data
    • R-type instructions
    • Load
  • Control
• Superscalar processors
Control Hazards

beq t0, t1, label
sub t2, s0, t5
or t6, s0, t3
xor t5, t1, s0
sw s0, 8(t3)

executed regardless of branch outcome!
executed regardless of branch outcome!!!
PC updated reflecting branch outcome
Observation

• If branch not taken, then instructions fetched sequentially after branch are correct

• If branch or jump taken, then need to flush incorrect instructions from pipeline by converting to NOPs
Kill Instructions after Branch if Taken

```
beq t0, t1, label
sub t2, s0, t5
or t6, s0, t3
label: xxxxxx
```

- **Taken branch**
- **Convert to NOP**
- **PC updated reflecting branch outcome**
Reducing Branch Penalties

- Every taken branch in simple pipeline costs 2 dead cycles
- To improve performance, use “branch prediction” to guess which way branch will go earlier in pipeline
- Only flush pipeline if branch prediction was incorrect
Branch Prediction

beq t0, t1, label

label: .....  

.....

Taken branch

Guess next PC!

Check guess correct
Implementing Branch Prediction...

• This is a CS152 topic, but some ideas:
  • Branch prediction is critical for performance on deeper pipelines/superscalar as the "Misprediction penalty" is vastly higher

• Keep a branch prediction buffer/cache: Small memory addressed by the lowest bits of PC
  • During instruction decode, if branch: Look up whether branch was taken last time?
    • If yes, compute PC + offset and fetch that (or store actual branch target address from last time)
    • If no, stick with PC + 4
  • If branch hasn't been seen before
    • Assume forward branches are not taken, backward branches are taken
  • Update state on predictor with results of branch when it is finally calculated
More on the Branch Predictor...

- What to do on the first time you see a branch?
  - RISC-V: Assume forward is not-taken, backwards will be taken
    - Forward == IF statement, backwards == loop
  - Compiler can then use this as a hint when structuring code

- Jumps are never mispredicted.... But JALR can be!
  - Maintain a similar cache for JALR instructions
  - One for JALR RA RX
    - Used to predict object-oriented function calls
  - One for JALR X0 RA as a stack
    - Used to handle function returns
Agenda

• RISC-V Pipeline
• Pipeline Control
• Hazards
  • Structural
  • Data
    • R-type instructions
    • Load
  • Control
• Superscalar processors
Increasing Single Processor Core Performance

1. Clock rate
   • Limited by technology and power dissipation
2. Pipelining
   • “Overlap” instruction execution
   • Deeper pipeline: 5 => 10 => 15 stages
     • Less work per stage → shorter clock cycle
     • But more potential for hazards (CPI > 1)
     • Depth limited by max clock rate and power dissipation
3. Multi-issue “superscalar” processor
Superscalar Processor

- Multiple issue “superscalar”
  - Replicate pipeline stages → multiple pipelines
  - Start multiple instructions per clock cycle
  - CPI < 1, so use Instructions Per Cycle (IPC)
  - E.g., 4GHz 4-way multiple-issue
    - 16 BIPS, peak CPI = 0.25, peak IPC = 4
  - Dependencies reduce this in practice

- “Out-of-Order” execution
  - Reorder instructions dynamically in hardware to reduce impact of hazards:
    EG, memory/cache misses

- CS152 discusses these techniques!
Out Of Order Superscalar Processor

Instruction fetch and decode unit

Reservation station  Reservation station  ...  Reservation station  Reservation station
Integer  Integer  ...  Floating point  Load-store

Functional units

Out-of-order execute

Commit unit

In-order issue

In-order commit

P&H p. 340
Benchmark: CPI of an older Intel Core i7

CPI = 1

CPI of Intel Core i7 920 running SPEC2006 integer benchmarks.
And That Is A Beast...

- 6 separate functional units
- 3x ALU
- 3 for memory operations
- 20-24 stage pipeline
- Aggressive branch prediction and other optimizations
  - Massive out-of-order capability: Can reorder up to 128 micro-operation instructions!
- And yet it still barely averages a 1 on CPI!
And then Multicore...

• Don't just have one super-scaler processor core, have multiple ones!
• Past approach: Take your best single-core and replicate it...
  • Problem: Those superscalar beasts consume a huge amount of energy
• New approach: Big/Little
  • Some of your processor cores are the single-core beasts...
  • Some are a more energy-optimized design
    • Far less energy/instruction, but fewer instructions/second
  • When absolute performance isn't needed, run tasks on the energy-efficient cores
• Started in cellular phones but now desktop & laptop processors (Intel "Alder Lake", Apple M1) adopt this approach
Pipelining and ISA Design

• RISC-V ISA designed for pipelining
  • All instructions are 32-bits in the RV-32 ISA
    • Easy to fetch and decode in one cycle
      • Variant additions add 16b and 64b instructions, but can tell by looking at just the first bits what type it is
    • Versus x86: 1- to 15-byte instructions
      • Requires additional pipeline stages for decoding instructions
  • Few and regular instruction formats
    • Decode and read registers in one step
  • Load/store addressing
    • Calculate address in 3rd stage, access memory in 4th stage
  • Alignment of memory operands
    • Memory access takes only one cycle
And Apple M1

- The following is informed speculation...
  - But the M1 seems to have delivered an >25% performance boost over comparable x86s

- The M1 is ARM rather than Intel x86
  - Arm is best described as "Mostly-RISC"
  - 32 general purpose registers, a load/store architecture, and an easy instruction decoding
    - Probably reduces branch mispredict penalty by a couple of clock cycles
  - But a fair bit of ISA bloat
    - Patterson & Hennessy ARM edition uses a subset of the ISA

- Where is the performance from?
  - Wide decode: 8 instructions/cycle
    - Latest Intel is only 6
  - Monstrous L1 caches, flatter cache hierarchy
    - L1: 192/128kB vs Intel’s 32/48kB ("Alder Lake")
  - Huge reordering: 630 deep out-of-order, 100+ memory loads/stores in process
    - Oh, and 32 general purpose registers to make the compiler happy!
In Conclusion

- Pipelining increases throughput by overlapping execution of multiple instructions
- All pipeline stages have same duration
  - Choose partition that accommodates this constraint
- Hazards potentially limit performance
  - Maximizing performance requires programmer/compiler assistance
- Superscalar processors use multiple execution units for additional instruction level parallelism
  - Performance benefit highly code dependent