CENG 3420
Computer Organization & Design

Lecture 16: Instruction-Level Parallelism

Bei Yu
CSE Department, CUHK
byu@cse.cuhk.edu.hk

(Textbook: Chapter 6)

Spring 2023

Overview

Introduction

Extracting Yet More Performance

Superpipelining

Increase the depth of the pipeline to increase the clock rate

The more stages in the pipeline, the more forwarding/hazard hardware needed and the more pipeline latch overhead (i.e., the pipeline latch accounts for a larger and larger percentage of the clock cycle time)

Multiple-Issue

Fetch (and execute) more than one instructions at one time (expand every pipeline stage to accommodate multiple instructions)

Example on Multiple-Issue

- The instruction execution rate, CPI, will be less than 1, so instead we use IPC: instructions per clock cycle
- E.g., a 3 GHz, four-way multiple-issue processor can execute at a peak rate of 12 billion instructions per second with a best case CPI of 0.25 or a best case IPC of 4

Question: If the datapath has a five stage pipeline, how many instructions are active in the pipeline at any given time?

ILP & Machine Parallelism

Instruction-level parallelism (ILP)

A measure of the average number of instructions in a program that a processor might be able to execute at the same time

Mostly determined by the number of true (data) dependencies and procedural (control) dependencies in relation to the number of other instructions

Machine Parallelism

A measure of the ability of the processor to take advantage of the ILP of the program

Determined by the number of instructions that can be fetched and executed at the same time.

To achieve high performance, need both ILP and Machine Parallelism

Multiple-Issue Processor Styles

Static multiple-issue processors (aka VLIW)

- Decisions on which instructions to execute simultaneously are being made statically (at compile time by the compiler)

Example: Intel Itanium and Itanium 2 for the IA-64 ISA

- EPIC (Explicit Parallel Instruction Computer)
- 128-bit "bundles" containing three instructions, each 41-bits plus a 5-bit template field (which specifies which FU each instruction needs)
- Five functional units (IntALU, Mmedia, Dmem, FPALU, Branch)
- Extensive support for speculation and predication

Dynamic multiple-issue processors (aka superscalar)

- Decisions on which instructions to execute simultaneously (in the range of 2 to 8) are being made dynamically (at run time by the hardware)

IBM Power series, Pentium 4, MIPS R10K, AMD Barcelona

Static v.s. Dynamic

Static: "let's make our compiler take care of this"

- Fast runtime
- Limited performance (variable values available when is running)

Dynamic: "let's build some hardware that takes care of this"

- Hardware penalty
- Complete knowledge on the program

Dependencies

Dependencies

Structural Hazards – Resource conflicts

- A SS/VLIW processor has a much larger number of potential resource conflicts
- Functional units may have to arbitrate for result buses and register-file write ports
- Resource conflicts can be eliminated by duplicating the resource or by pipelining the resource

Data Hazards – Storage (data) dependencies

Limitation more severe in a SS/VLIW processor due to (usually) low ILP

Control Hazards – Procedural dependencies

- Ditto, but even more severe
- Use dynamic branch prediction to help resolve the ILP issue

Resolved through combination of hardware and software.

Data Hazards

```
R3 := R3 * R5
R4 := R3 + 1
R3 := R5 + 1
```

True data dependency (RAW)
Antidependency (WAR)
Output dependency (WAW)

True dependency (RAW)

Later instruction using a value (not yet) produced by an earlier instruction.

Antidependencies (WAR)

Later instruction (that executes earlier) produces a data value that destroys a data value used as a source in an earlier instruction (that executes later).

Output dependency (WAW)

Two instructions write the same register or memory location.

Question: Find all data dependences in this instruction sequence.

```
I1: ADD R1, R2, R1
I2: LW  R2, 0(R1)
I3: LW  R1, 4(R1)
I4: OR  R3, R1, R2
```

Data Hazards

```
R3 := R3 * R5
R4 := R3 + 1
R3 := R5 + 1
```

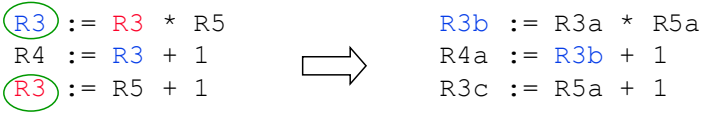
True data dependency (RAW)
Antidependency (WAR)
Output dependency (WAW)

- True dependencies (RAW) represent the flow of data and information through a program
- Antidependencies (WAR) and output dependencies (WAW) arise because the limited number of registers, i.e., programmers reuse registers for different computations leading to storage conflicts
- Storage conflicts can be reduced (or eliminated) by
  - Increasing or duplicating the troublesome resource
  - Providing additional registers that are used to re-establish the correspondence between registers and values
  - Allocated dynamically by the hardware in SS processors

Resolve Storage Conflicts

Register Renaming

The processor renames the original register identifier in the instruction to a new register (one not in the visible register set)



- The hardware that does renaming assigns a "replacement" register from a pool of free registers
- Releases it back to the pool when its value is superseded and there are no outstanding references to it

Resolve Control Dependency

Speculation

Allow execution of future instr's that (may) depend on the speculated instruction:

- Speculate on the outcome of a conditional branch (branch prediction)
- Speculate that a store (for which we don't yet know the address) that precedes a load does not refer to the same address, allowing the load to be scheduled before the store (load speculation)

Must have (hardware and/or software) mechanisms for

- Checking to see if the guess was correct
- Recovering from the effects of the instructions that were executed speculatively if the guess was incorrect

Ignore and/or buffer exceptions created by speculatively executed instructions until it is clear that they should really occur

VLIW

Static Multiple Issue Machines (VLIW)

Static multiple-issue processors (aka VLIW) use the compiler (at compile-time) to statically decide which instructions to issue and execute simultaneously

- Issue packet – the set of instructions that are bundled together and issued in one clock cycle – think of it as one large instruction with multiple operations
- The mix of instructions in the packet (bundle) is usually restricted – a single "instruction" with several predefined fields
- The compiler does static branch prediction and code scheduling to reduce (control) or eliminate (data) hazards

VLIW has

- Multiple functional units
- Multi-ported register files
- Wide program bus

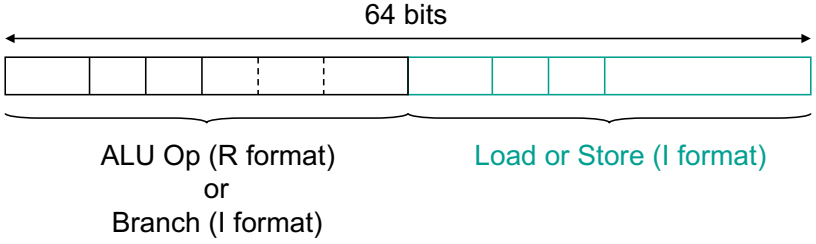
An Example: A VLIW RISC-V

The ALU and data transfer instructions are issued at the same time.

Instruction type	Pipe stages							
ALU or branch instruction	IF	ID	EX	MEM	WB			
Load or store instruction	IF	ID	EX	MEM	WB			
ALU or branch instruction		IF	ID	EX	MEM	WB		
Load or store instruction		IF	ID	EX	MEM	WB		
ALU or branch instruction			IF	ID	EX	MEM	WB	
Load or store instruction			IF	ID	EX	MEM	WB	
ALU or branch instruction				IF	ID	EX	MEM	WB
Load or store instruction				IF	ID	EX	MEM	WB

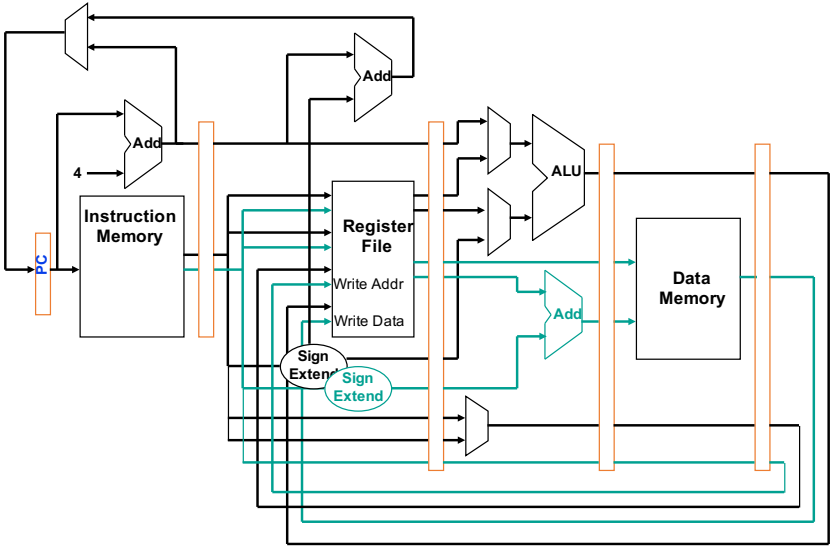
An Example: A VLIW RISC-V

Consider a 2-issue RISC-V with a 2 instr bundle



- Instructions are always fetched, decoded, and issued in pairs
- If one instr of the pair can not be used, it is replaced with a noop
- Need 4 read ports and 2 write ports and a separate memory address adder

A RISC-V VLIW (2-issue) Datapath



No hazard hardware (so no load use allowed)

Compiler Techniques for Exposing ILP

1. Instruction Scheduling
2. Loop Unrolling

Instruction Scheduling Example

Consider the following loop code

```
lp: lw   $t0, 0($s1)      # $t0=array element
    addu $t0, $t0, $s2    # add scalar in $s2
    sw   $t0, 0($s1)      # store result
    addi $s1, $s1, -4     # decrement pointer
    bne  $s1, $0, lp      # branch if $s1 != 0
```

Must "schedule" the instructions to avoid pipeline stalls

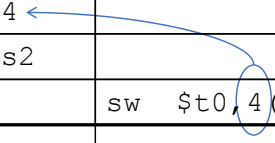
- Instructions in one bundle must be independent
- Must separate load use instructions from their loads by one cycle
- Notice that the first two instructions have a load use dependency, the next two and last two have data dependencies
- Assume branches are perfectly predicted by the hardware

The Scheduled Instruction (Not Unrolled)

	ALU or branch	Data transfer	CC
			1
			2
			3
			4
			5

The Scheduled Instruction (Not Unrolled)

	ALU or branch	Data transfer	CC
lp:		`lw $t0,0($s1)`	1
	`addi $s1,$s1,-4`		2
	`addu $t0,$t0,$s2`		3
	`bne $s1,$0,lp`	`sw $t0,4($s1)`	4
			5



- Four clock cycles to execute 5 instructions
- CPI of 0.8 (versus the best case of 0.5)
- IPC of 1.25 (versus the best case of 2.0)
- `noop`s don't count towards performance

Loop Unrolling

- Loop unrolling – multiple copies of the loop body are made and instructions from different iterations are scheduled together as a way to increase ILP
- Apply loop unrolling (4 times for our example) and then schedule the resulting code
  - Eliminate unnecessary loop overhead instructions
  - Schedule so as to avoid load use hazards
- During unrolling the compiler applies register renaming to eliminate all data dependencies that are not true data dependencies

Unrolled Code Example

```
lp: lw   $t0,0($s1)    # $t0=array element
    lw   $t1,-4($s1)   # $t1=array element
    lw   $t2,-8($s1)   # $t2=array element
    lw   $t3,-12($s1)  # $t3=array element
    addu $t0,$t0,$s2   # add scalar in $s2
    addu $t1,$t1,$s2   # add scalar in $s2
    addu $t2,$t2,$s2   # add scalar in $s2
    addu $t3,$t3,$s2   # add scalar in $s2
    sw   $t0,0($s1)    # store result
    sw   $t1,-4($s1)   # store result
    sw   $t2,-8($s1)   # store result
    sw   $t3,-12($s1)  # store result
    addi $s1,$s1,-16   # decrement pointer
    bne  $s1,$0,lp     # branch if s1 != 0
```

The Scheduled Code (Unrolled)

	ALU or branch	Data transfer	CC
			1
			2
			3
			4
			5
			6
			7
			8

The Scheduled Code (Unrolled)

	ALU or branch	Data transfer	CC
lp:	addi $s1,$s1,-16	lw $t0,0($s1)	1
		lw $t1,12($s1)	2
	addu $t0,$t0,$s2	lw $t2,8($s1)	3
	addu $t1,$t1,$s2	lw $t3,4($s1)	4
	addu $t2,$t2,$s2	sw $t0,16($s1)	5
	addu $t3,$t3,$s2	sw $t1,12($s1)	6
		sw $t2,8($s1)	7
	bne $s1,$0,lp	sw $t3,4($s1)	8

- Eight clock cycles to execute 14 instructions
- CPI of 0.57 (versus the best case of 0.5)
- IPC of 1.8 (versus the best case of 2.0)

Compiler Support for VLIW Processors

- The compiler packs groups of independent instructions into the bundle – Done by code re-ordering (trace scheduling)
- The compiler uses loop unrolling to expose more ILP
- The compiler uses register renaming to solve name dependencies and ensures no load use hazards occur
- While superscalars use dynamic prediction, VLIW's primarily depend on the compiler for branch prediction
  - Loop unrolling reduces the number of conditional branches
  - Predication eliminates if-the-else branch structures by replacing them with predicated instructions
- The compiler predicts memory bank references to help minimize memory bank conflicts

SuperScalar (SS)

SuperScaler – Dynamic multiple-issue processors

Use hardware at run-time to dynamically decide which instructions to issue and execute simultaneously

- Instruction-fetch and issue – fetch instructions, decode them, and issue them to a FU to await execution
- Defines the Instruction lookahead capability – fetch, decode and issue instructions beyond the current instruction
- Instruction-execution – as soon as the source operands and the FU are ready, the result can be calculated
- Defines the processor lookahead capability – complete execution of issued instructions beyond the current instruction
- Instruction-commit – when it is safe to, write back results to the RegFile or D$ (i.e., change the machine state)

In-Order

Instruction Fetch and Decode Units

are required to issue instructions in-order so that dependencies can be tracked

Commit Unit

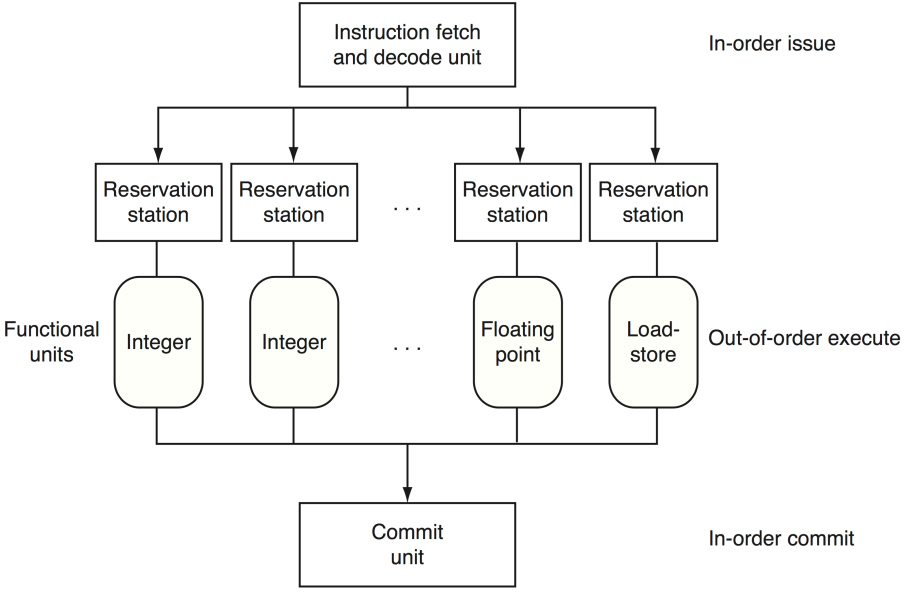
is required to write results to registers and memory in program fetch order so that

- If exceptions occur the only registers updated will be those written by instructions before the one causing the exception
- If branches are mispredicted, those instructions executed after the mispredicted branch don't change the machine state (i.e., we use the commit unit to correct incorrect speculation)

Out-of-Order

- Although the front end (fetch, decode, and issue) and back end (commit) of the pipeline run in-order
- FUs are free to initiate execution whenever the data they need is available – out-of-(program) order execution
- Allowing out-of-order execution increases the amount of ILP

In-Order v.s. Out-of-Order



Out-of-Order Execution

With out-of-order execution, a later instruction may execute before a previous instruction so the hardware needs to resolve both write after read (WAR) and write after write (WAW) data hazards.

```
lw      $t0, 0($s1)
addu    $t0, $t1, $s2
. . .
sub     $t2, $t0, $s2
```

- If the `lw` write to `$t0` occurs after the addu write, then the sub gets an incorrect value for `$t0`
- The addu has an output dependency on the `lw` – write after write (WAW)
- The issuing of the `addu` might have to be stalled if its result could later be overwritten by an previous instruction that takes longer to complete

Kernel-memory-leaking Intel Processor Design Flaw

Kernel-memory-leaking Intel processor design flaw forces Linux, Windows redesign

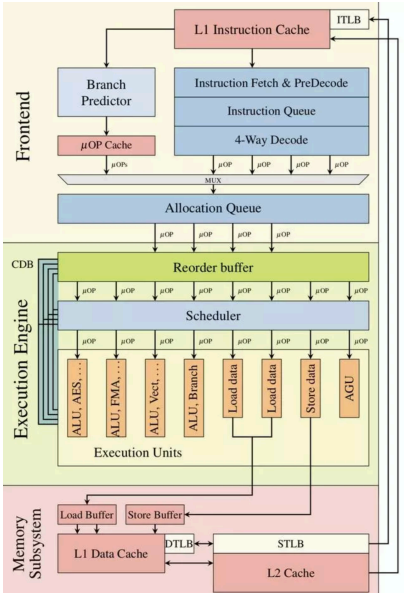
Speed hits loom, other OSes need fixes

By John Leyden and Chris Williams 2 Jan 2018 at 19:29 450 SHARE ▼

Final update A fundamental design flaw in Intel's processor chips has forced a significant redesign of the Linux and Windows kernels to defang the chip-level security bug.

Programmers are scrambling to overhaul the open-source Linux kernel's virtual memory system. Meanwhile, Microsoft is expected to publicly introduce the necessary changes to its Windows operating system in an upcoming Patch Tuesday: these changes were seeded to beta testers running fast-ring Windows Insider builds in November and December.

Crucially, these updates to both Linux and Windows will incur a performance hit on Intel products. The effects are still being benchmarked, however we're looking at a ballpark figure of five to 30 per cent slow down, depending on the task and the processor model. More recent Intel chips have features – such as PCID – to reduce the performance hit. Your mileage may vary.



http://www.theregister.co.uk/2018/01/02/intel_cpu_design_flaw/

Summary

Does ILP Work?

- Yes, but not as much as we'd like
- Programs have real dependencies that limit ILP
- Some dependencies are hard to eliminate, e.g., pointer aliasing
- Some parallelism is hard to expose, e.g., limited window size during instruction issue
- Memory delays and limited bandwidth: hard to keep pipelines full
- Speculation can help if done well

Does ILP Work?

- Yes, but not as much as we'd like
- Programs have real dependencies that limit ILP
- Some dependencies are hard to eliminate, e.g., pointer aliasing
- Some parallelism is hard to expose, e.g., limited window size during instruction issue
- Memory delays and limited bandwidth: hard to keep pipelines full
- Speculation can help if done well (Security Issue!)

Summary: Extracting More Performance

To achieve high performance, need both machine parallelism and instruction level parallelism (ILP) by

- Superpipelining
- Static multiple-issue (VLIW)
- Dynamic multiple-issue (superscalar)

- A processor's instruction issue and execution policies impact the available ILP
- Register renaming can solve these storage dependencies

CISC vs RISC vs SS vs VLIW

	CISC	RISC	Superscalar	VLIW
Instr size	variable size	fixed size	fixed size	fixed size (but large)
Instr format	variable format	fixed format	fixed format	fixed format
Registers	few, some special Limited # of ports	Many GP Limited # of ports	GP and rename (RUU) Many ports	many, many GP Many ports
Memory reference	embedded in many instr's	load/store	load/store	load/store
Key Issues	decode complexity	data forwarding, hazards	hardware dependency resolution	(compiler) code scheduling

Evolution of Pipelined, SS Processors

	Year	Clock Rate	# Pipe Stages	Issue Width	OOO?	Cores /Chip	Power
Intel 486	1989	25 MHz	5	1	No	1	5 W
Intel Pentium	1993	66 MHz	5	2	No	1	10 W
Intel Pentium Pro	1997	200 MHz	10	3	Yes	1	29 W
Intel Pentium 4 Willamette	2001	2000 MHz	22	3	Yes	1	75 W
Intel Pentium 4 Prescott	2004	3600 MHz	31	3	Yes	1	103 W
Intel Core	2006	2930 MHz	14	4	Yes	2	75 W
Sun USPARC III	2003	1950 MHz	14	4	No	1	90 W
Sun T1 (Niagara)	2005	1200 MHz	6	1	No	8	70 W

Power Efficiency

- Complexity of dynamic scheduling and speculations requires power
- Multiple simpler cores may be better (next lecture)