
CENG/CSCI 3420

Computer Organization and Design

Spring 2014

Lecture 02: Performance and ISA

XU, Qiang 徐強

[Adapted from UC Berkeley's D. Patterson's and
from PSU's Mary J. Irwin's slides with additional credits to Y. Xie]

Performance Metrics

❑ Purchasing perspective

- given a collection of machines, which has the
 - best performance ?
 - least cost ?
 - best cost/performance?

❑ Design perspective

- faced with design options, which has the
 - best performance improvement ?
 - least cost ?
 - best cost/performance?

❑ Both require

- basis for comparison
- metric for evaluation

❑ Our goal is to understand what factors in the architecture contribute to overall system performance and the relative importance (and cost) of these factors

Throughput versus Response Time

- ❑ Response time (execution time) – the time between the start and the completion of a task
 - Important to individual users

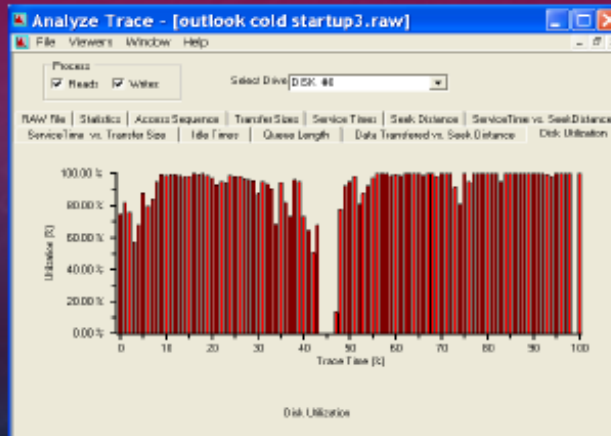
- ❑ Throughput (bandwidth) – the total amount of work done in a given time
 - Important to data center managers

- ❑ Will need different performance metrics as well as a different set of applications to benchmark **embedded** and **desktop** computers, which are more focused on response time, versus **servers**, which are more focused on throughput

Response Time Matters

It's the Hard Disk, Stupid!

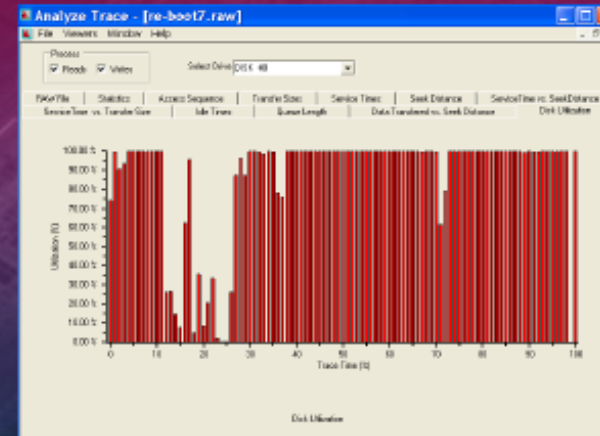
Re-Boot/Startup
on Home PC



Elapsed Time 105.213536, s
Disk Busy Time 91.368480, s
Average Data Rate 6.60669, MB/s

86% BUSY

Starting Outlook



Elapsed Time 45.700667, s
Disk Busy Time 41.056997, s
Average Data Rate 1.37389, MB/s

89% BUSY



Defining (Speed) Performance

- ❑ To maximize performance, need to **minimize** execution time

$$\text{performance}_x = 1 / \text{execution_time}_x$$

If X is n times faster than Y, then

$$\frac{\text{performance}_x}{\text{performance}_y} = \frac{\text{execution_time}_y}{\text{execution_time}_x} = n$$

- ❑ Decreasing response time almost always improves throughput

A Relative Performance Example

- ❑ If computer A runs a program in 10 seconds and computer B runs the same program in 15 seconds, how much faster is A than B?

Performance Factors

- ❑ CPU execution time (CPU time) – time the CPU spends working on a task
 - Does not include time waiting for I/O or running other programs

$$\text{CPU execution time for a program} = \# \text{ CPU clock cycles for a program} \times \text{clock cycle time}$$

or

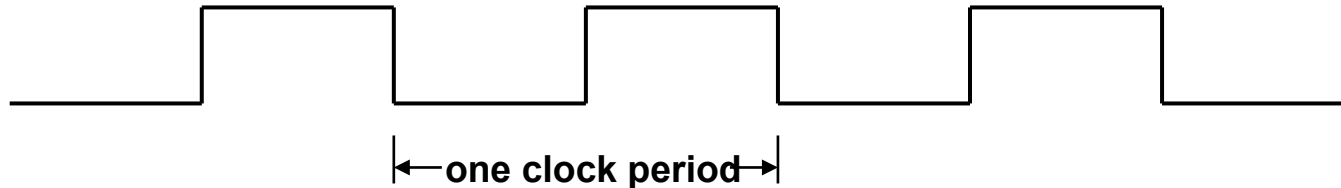
$$\text{CPU execution time for a program} = \frac{\# \text{ CPU clock cycles for a program}}{\text{clock rate}}$$

- ❑ Can improve performance by reducing either the **length of the clock cycle** or the **number of clock cycles required for a program**

Review: Machine Clock Rate

- Clock rate (clock cycles per second in MHz or GHz) is inverse of clock cycle time (clock period)

$$CC = 1 / CR$$



10 nsec clock cycle => 100 MHz clock rate

5 nsec clock cycle => 200 MHz clock rate

2 nsec clock cycle => 500 MHz clock rate

1 nsec (10^{-9}) clock cycle => 1 GHz (10^9) clock rate

500 psec clock cycle => 2 GHz clock rate

250 psec clock cycle => 4 GHz clock rate

200 psec clock cycle => 5 GHz clock rate

Improving Performance Example

- ❑ A program runs on computer A with a 2 GHz clock in 10 seconds. What clock rate must a computer B run at to run this program in 6 seconds? Unfortunately, to accomplish this, computer B will require 1.2 times as many clock cycles as computer A to run the program.

Clock Cycles per Instruction

- ❑ Not all instructions take the same amount of time to execute
 - One way to think about execution time is that it equals the number of instructions executed multiplied by the average time per instruction

$$\begin{array}{l} \# \text{ CPU clock cycles} \\ \text{for a program} \end{array} = \begin{array}{l} \# \text{ Instructions} \\ \text{for a program} \end{array} \times \begin{array}{l} \text{Average clock cycles} \\ \text{per instruction} \end{array}$$

- ❑ **Clock cycles per instruction (CPI)** – the average number of clock cycles each instruction takes to execute
 - A way to compare two different implementations of the same ISA

	CPI for this instruction class		
	A	B	C
CPI	1	2	3

Using the Performance Equation

- ❑ Computers A and B implement the same ISA. Computer A has a clock cycle time of 250 ps and an effective CPI of 2.0 for some program and computer B has a clock cycle time of 500 ps and an effective CPI of 1.2 for the same program. Which computer is faster and by how much?

Effective (Average) CPI

- ❑ Computing the overall effective CPI is done by looking at the different types of instructions and their individual cycle counts and averaging

$$\text{Overall effective CPI} = \sum_{i=1}^n CPI_i \times IC_i$$

- Where IC_i is the count (percentage) of the number of instructions of class i executed
 - CPI_i is the (average) number of clock cycles per instruction for that instruction class
 - n is the number of instruction classes
- ❑ The overall effective CPI varies by instruction mix – a measure of the dynamic frequency of instructions across one or many programs

THE Performance Equation

- Our basic performance equation is then

$$\text{CPU time} = \text{Instruction_count} \times \text{CPI} \times \text{clock_cycle}$$

or

$$\text{CPU time} = \frac{\text{Instruction_count} \times \text{CPI}}{\text{clock_rate}}$$

- These equations separate the **three key** factors that affect performance
 - Can measure the CPU execution time by running the program
 - The clock rate is usually given
 - Can measure overall instruction count by using profilers/ simulators without knowing all of the implementation details
 - CPI varies by instruction type and ISA implementation for which we must know the implementation details

Determinates of CPU Performance

$$\text{CPU time} = \text{Instruction_count} \times \text{CPI} \times \text{clock_cycle}$$

	Instruction_count	CPI	clock_cycle
Algorithm			
Programming language			
Compiler			
ISA			
Core organization			
Technology			

A Simple Example

Op	Freq	CPI _i	Freq x CPI _i
ALU	50%	1	.
Load	20%	5	
Store	10%	3	
Branch	20%	2	
			$\Sigma =$

- ❑ How much faster would the machine be if a better data cache reduced the average load time to 2 cycles?
- ❑ How does this compare with using branch prediction to shave a cycle off the branch time?
- ❑ What if two ALU instructions could be executed at once?

A Simple Example

Op	Freq	CPI _i	Freq x CPI _i				
ALU	50%	1	.5	.5	.5	.25	
Load	20%	5	1.0	.4	1.0	1.0	
Store	10%	3	.3	.3	.3	.3	
Branch	20%	2	.4	.4	.2	.4	
			$\Sigma =$	2.2	1.6	2.0	1.95

- How much faster would the machine be if a better data cache reduced the average load time to 2 cycles?

CPU time new = 1.6 x IC x CC so 2.2/1.6 means 37.5% faster

- How does this compare with using branch prediction to shave a cycle off the branch time?

CPU time new = 2.0 x IC x CC so 2.2/2.0 means 10% faster

- What if two ALU instructions could be executed at once?

CPU time new = 1.95 x IC x CC so 2.2/1.95 means 12.8% faster

Workloads and Benchmarks

- ❑ Benchmarks – a set of programs that form a “workload” specifically chosen to measure performance
- ❑ SPEC (System Performance Evaluation Cooperative) creates standard sets of benchmarks starting with SPEC89. The latest is SPEC CPU2006 which consists of 12 integer benchmarks (CINT2006) and 17 floating-point benchmarks (CFP2006).

www.spec.org

- ❑ There are also benchmark collections for power workloads (SPECpower_ssj2008), for mail workloads (SPECmail2008), for multimedia workloads (mediabench), ...

Old SPEC Benchmarks

Integer benchmarks		FP benchmarks	
gzip	compression	wupwise	Quantum chromodynamics
vpr	FPGA place & route	swim	Shallow water model
gcc	GNU C compiler	mgrid	Multigrid solver in 3D fields
mcf	Combinatorial optimization	applu	Parabolic/elliptic pde
crafty	Chess program	mesa	3D graphics library
parser	Word processing program	galgel	Computational fluid dynamics
eon	Computer visualization	art	Image recognition (NN)
perlbnk	perl application	equake	Seismic wave propagation simulation
gap	Group theory interpreter	facerec	Facial image recognition
vortex	Object oriented database	ammp	Computational chemistry
bzip2	compression	lucas	Primality testing
twolf	Circuit place & route	fma3d	Crash simulation fem
		sixtrack	Nuclear physics accel
		apsi	Pollutant distribution

SPEC CINT2006 on Barcelona (CC = 0.4 x 10⁹)

Name	ICx10 ⁹	CPI	ExTime	RefTime	SPEC ratio
perl	2,1118	0.75	637	9,770	15.3
bzip2	2,389	0.85	817	9,650	11.8
gcc	1,050	1.72	724	8,050	11.1
mcf	336	10.00	1,345	9,120	6.8
go	1,658	1.09	721	10,490	14.6
hmmer	2,783	0.80	890	9,330	10.5
sjeng	2,176	0.96	837	12,100	14.5
libquantum	1,623	1.61	1,047	20,720	19.8
h264avc	3,102	0.80	993	22,130	22.3
omnetpp	587	2.94	690	6,250	9.1
astar	1,082	1.79	773	7,020	9.1
xalancbmk	1,058	2.70	1,143	6,900	6.0
Geometric Mean					11.7

Comparing and Summarizing Performance

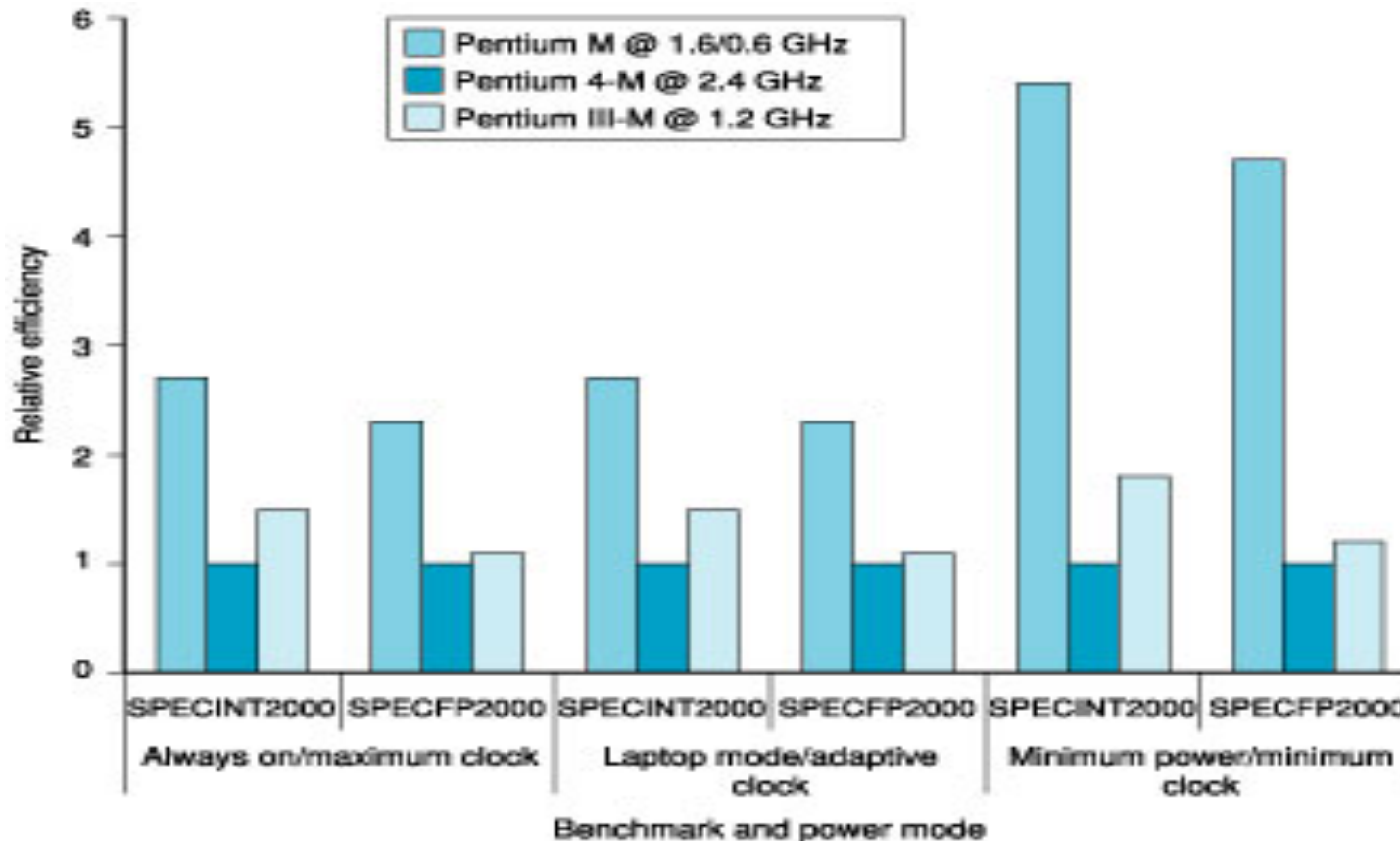
- How do we summarize the performance for benchmark set with a **single** number?
 - First the execution times are normalized giving the “SPEC ratio” (bigger is faster, i.e., SPEC ratio is the inverse of execution time)
 - The SPEC ratios are then “averaged” using the **geometric mean** (GM)

$$GM = \sqrt[n]{\prod_{i=1}^n \text{SPEC ratio}_i}$$

- Guiding principle in reporting performance measurements is **reproducibility** – list everything another experimenter would need to duplicate the experiment (version of the operating system, compiler settings, input set used, specific computer configuration (clock rate, cache sizes and speed, memory size and speed, etc.))

Other Performance Metrics

- ❑ Power consumption – especially in the embedded market where battery life is important
 - For power-limited applications, the most important metric is energy efficiency



Summary: Evaluating ISAs

❑ Design-time metrics:

- Can it be implemented? With what performance, at what costs (design, fabrication, test, packaging), with what power, with what reliability?
- Can it be programmed? Ease of compilation?

❑ Static Metrics:

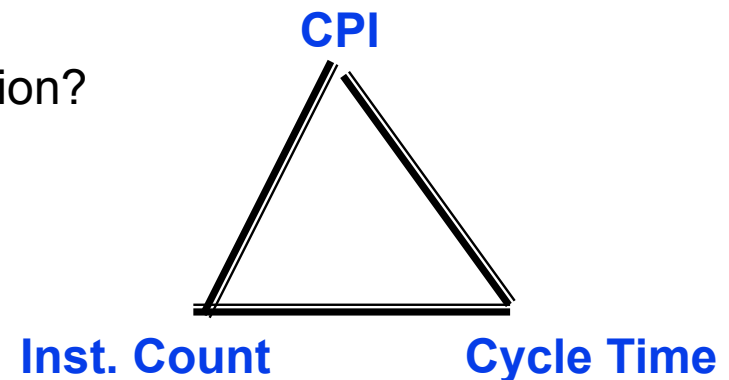
- How many bytes does the program occupy in memory?

❑ Dynamic Metrics:

- How many instructions are executed? How many bytes does the processor fetch to execute the program?
- How many clocks are required per instruction?
- How "lean" a clock is practical?

Best Metric: Time to execute the program!

depends on the instructions set, the processor organization, and compilation techniques.



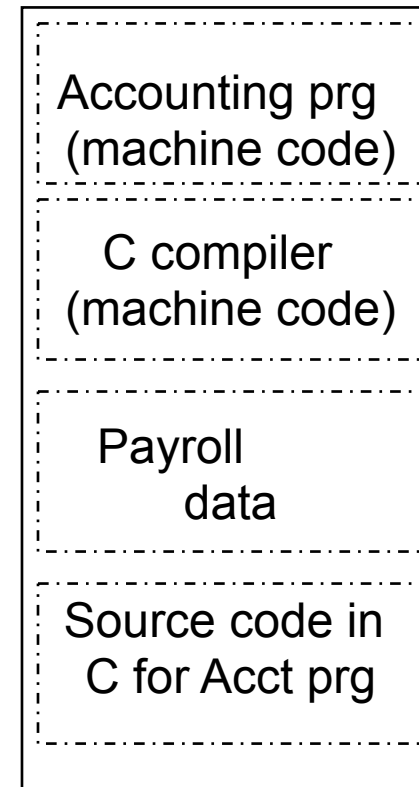
Two Key Principles of Machine Design

1. Instructions are represented as numbers and, as such, are indistinguishable from data
2. Programs are stored in alterable memory (that can be read or written to) just like data

□ Stored-program concept

- Programs can be shipped as files of binary numbers – **binary compatibility**
- Computers can inherit ready-made software provided they are compatible with an existing ISA – leads industry to align around a small number of ISAs

Memory

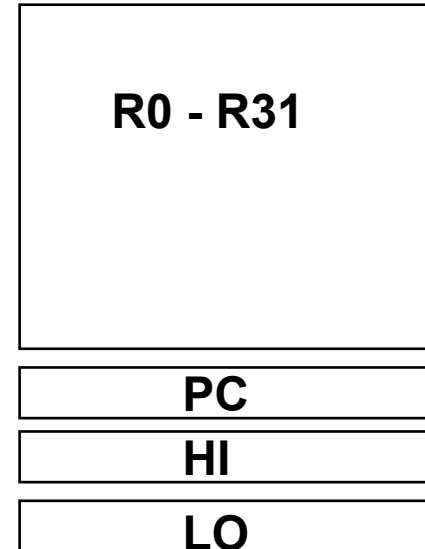


MIPS-32 ISA

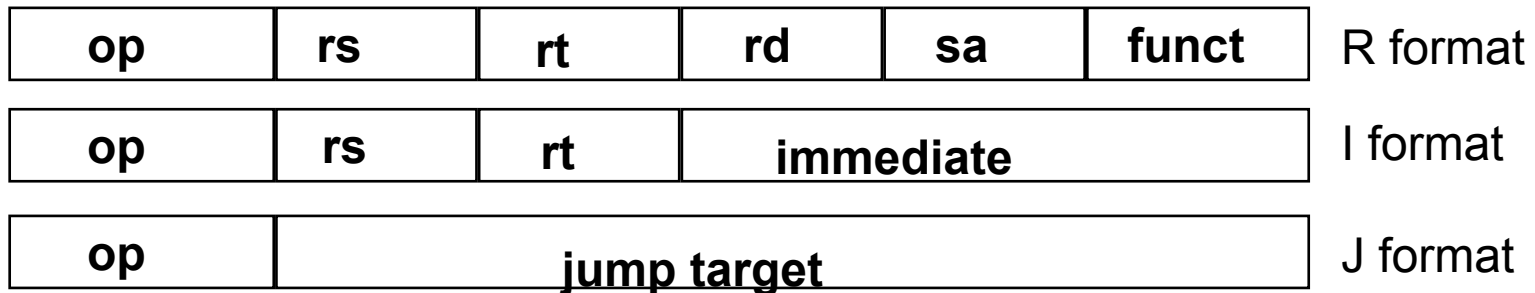
□ Instruction Categories

- Computational
- Load/Store
- Jump and Branch
- Floating Point
 - coprocessor
- Memory Management
- Special

Registers



3 Instruction Formats: **all 32 bits wide**



MIPS (RISC) Design Principles

- ❑ Simplicity favors regularity
- ❑ Smaller is faster
- ❑ Make the common case fast
- ❑ Good design demands good compromises

MIPS (RISC) Design Principles

□ Simplicity favors regularity

- fixed size instructions
- small number of instruction formats
- opcode always the first 6 bits

□ Smaller is faster

- limited instruction set
- limited number of registers in register file
- limited number of addressing modes

□ Make the common case fast

- arithmetic operands from the register file (load-store machine)
- allow instructions to contain immediate operands

□ Good design demands good compromises

- three instruction formats

MIPS Arithmetic Instructions

- ❑ MIPS assembly language arithmetic statement

```
add $t0, $s1, $s2
```

```
sub $t0, $s1, $s2
```

- ❑ Each arithmetic instruction performs **one** operation
- ❑ Each specifies exactly **three** operands that are all contained in the datapath's register file ($\$t0, \$s1, \$s2$)

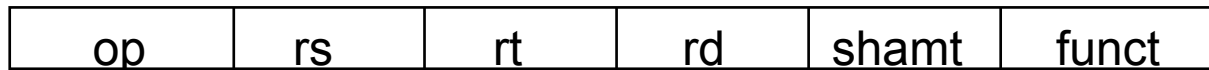
destination = source1 **op** source2

- ❑ Instruction Format (**R** format)

0	17	18	8	0	0x22
---	----	----	---	---	------

MIPS Instruction Fields

- ❑ MIPS fields are given names to make them easier to refer to



op	6-bits	opcode that specifies the operation
rs	5-bits	register file address of the first source operand
rt	5-bits	register file address of the second source operand
rd	5-bits	register file address of the result's destination
shamt	5-bits	shift amount (for shift instructions)
funct	6-bits	function code augmenting the opcode

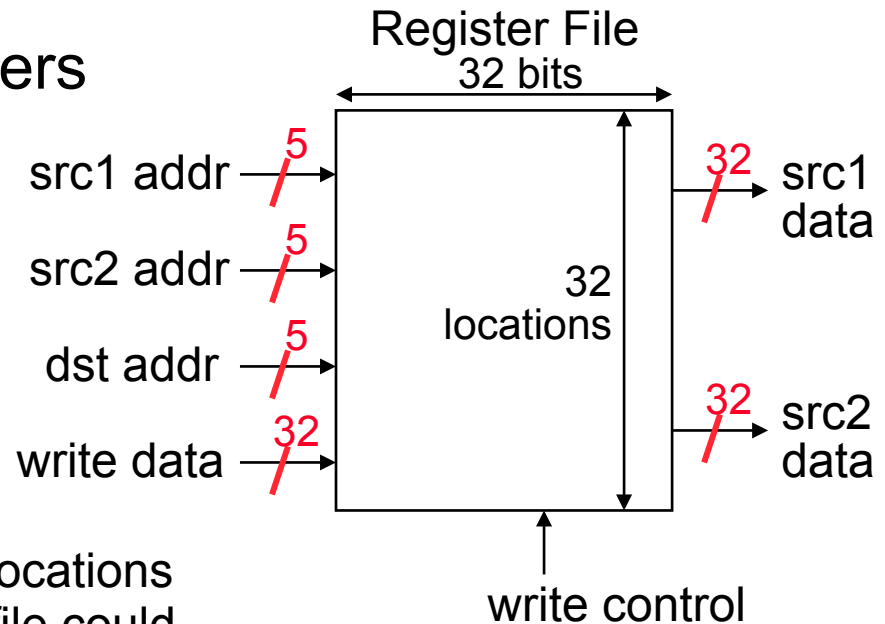
MIPS Register File

- ❑ Holds thirty-two 32-bit registers

- Two read ports and
- One write port

- ❑ Registers are

- **Faster** than main memory
 - But register files with more locations are slower (e.g., a 64 word file could be as much as 50% slower than a 32 word file)
 - Read/write port increase impacts speed quadratically
- Easier for a compiler to use
 - e.g., $(A*B) - (C*D) - (E*F)$ can do multiplies in any order vs. stack
- Can hold variables so that
 - code density improves (since register are named with fewer bits than a memory location)



Aside: MIPS Register Convention

Name	Register Number	Usage	Preserve on call?
\$zero	0	constant 0 (hardware)	n.a.
\$at	1	reserved for assembler	n.a.
\$v0 - \$v1	2-3	returned values	no
\$a0 - \$a3	4-7	arguments	yes
\$t0 - \$t7	8-15	temporaries	no
\$s0 - \$s7	16-23	saved values	yes
\$t8 - \$t9	24-25	temporaries	no
\$gp	28	global pointer	yes
\$sp	29	stack pointer	yes
\$fp	30	frame pointer	yes
\$ra	31	return addr (hardware)	yes

MIPS Memory Access Instructions

- ❑ MIPS has two basic **data transfer** instructions for accessing memory

```
lw    $t0, 4($s3)    #load word from memory
```

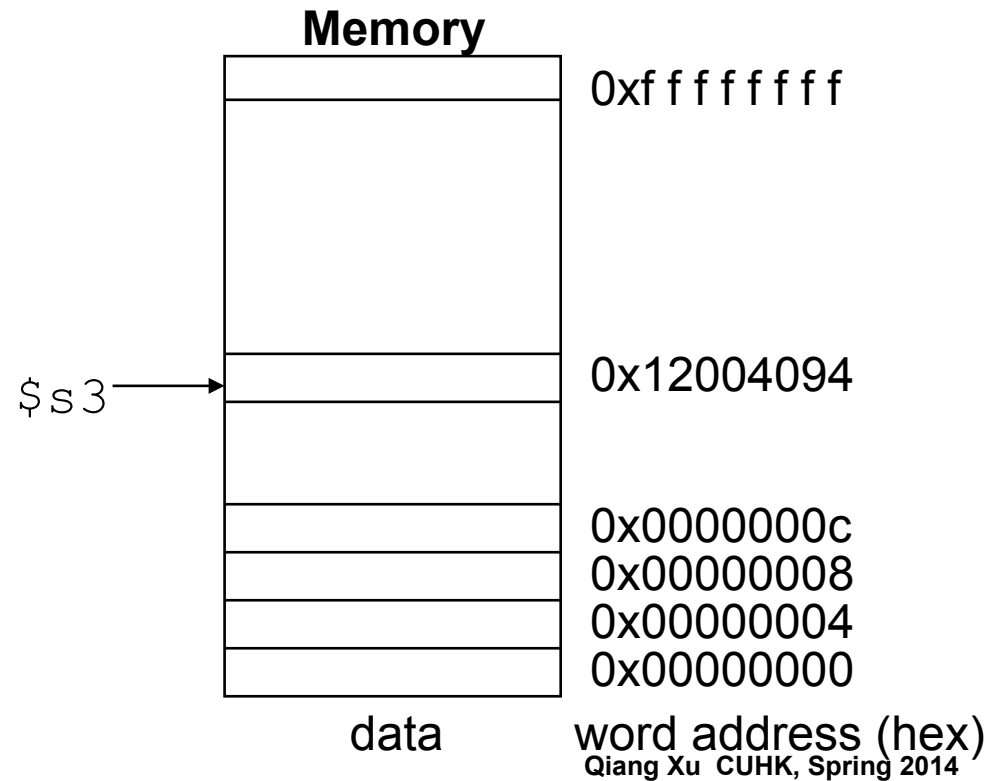
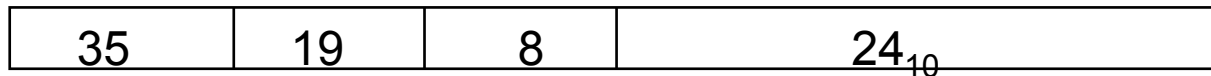
```
sw    $t0, 8($s3)    #store word to memory
```

- ❑ The data is loaded into (lw) or stored from (sw) a register in the register file – a 5 bit address
- ❑ The memory address – a 32 bit address – is formed by adding the contents of the **base address register** to the **offset** value
 - A 16-bit field meaning access is limited to memory locations within a region of $\pm 2^{13}$ or 8,192 words ($\pm 2^{15}$ or 32,768 bytes) of the address in the base register

Machine Language - Load Instruction

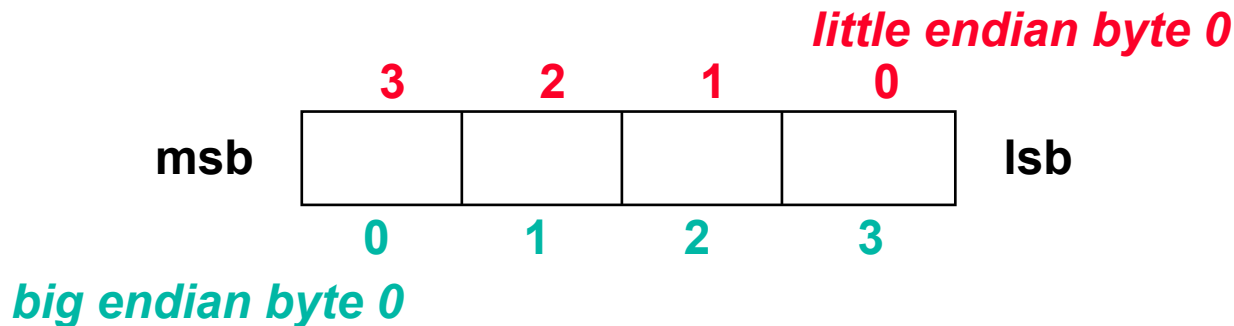
- Load/Store Instruction Format (I format):

`lw $t0, 24($s3)`



Byte Addresses

- ❑ Since 8-bit bytes are so useful, most architectures address individual **bytes** in memory
 - **Alignment restriction** - the memory address of a **word** must be on natural word boundaries (a multiple of 4 in MIPS-32)
- ❑ **Big Endian:** leftmost byte is word address
IBM 360/370, Motorola 68k, **MIPS**, Sparc, HP PA
- ❑ **Little Endian:** rightmost byte is word address
Intel 80x86, DEC Vax, DEC Alpha (Windows NT)

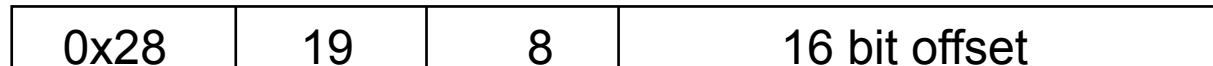


Aside: Loading and Storing Bytes

- ❑ MIPS provides special instructions to move bytes

```
lb    $t0, 1($s3)    #load byte from memory
```

```
sb    $t0, 6($s3)    #store byte to memory
```



- ❑ What 8 bits get loaded and stored?

- load byte places the byte from memory in the rightmost 8 bits of the destination register
 - what happens to the other bits in the register?
- store byte takes the byte from the rightmost 8 bits of a register and writes it to a byte in memory
 - what happens to the other bits in the memory word?

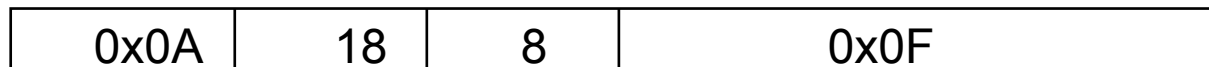
MIPS Immediate Instructions

- ❑ Small constants are used often in typical code
- ❑ Possible approaches?
 - put “typical constants” in memory and load them
 - create hard-wired registers (like \$zero) for constants like 1
 - **have special instructions that contain constants !**

```
addi $sp, $sp, 4      # $sp = $sp + 4
```

```
slti $t0, $s2, 15     # $t0 = 1 if $s2 < 15
```

- ❑ Machine format (I format):

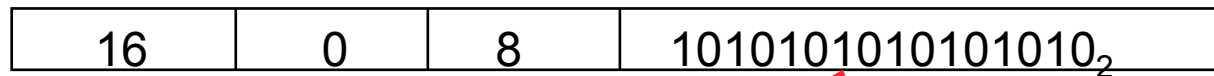


- ❑ The constant is kept **inside** the instruction itself!
 - Immediate format **limits** values to the range $+2^{15}-1$ to -2^{15}

Aside: How About Larger Constants?

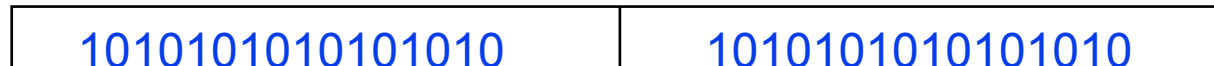
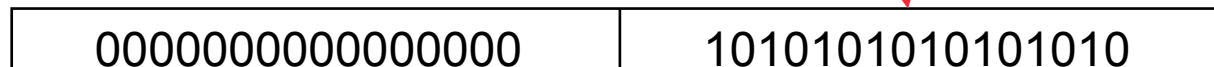
- ❑ We'd also like to be able to load a 32 bit constant into a register, for this we must use two instructions
- ❑ a new "load upper immediate" instruction

```
lui $t0, 1010101010101010
```



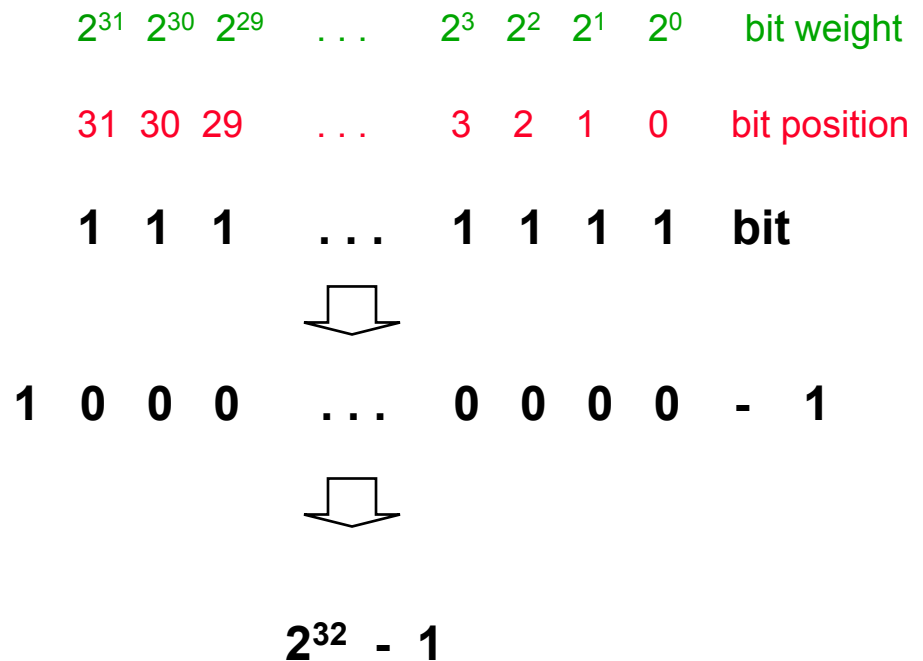
- ❑ Then must get the lower order bits right, use

```
ori $t0, $t0, 1010101010101010
```



Review: Unsigned Binary Representation

Hex	Binary	Decimal
0x00000000	0...0000	0
0x00000001	0...0001	1
0x00000002	0...0010	2
0x00000003	0...0011	3
0x00000004	0...0100	4
0x00000005	0...0101	5
0x00000006	0...0110	6
0x00000007	0...0111	7
0x00000008	0...1000	8
0x00000009	0...1001	9
	...	
0xFFFFFFFFC	1...1100	$2^{32} - 4$
0xFFFFFFFDD	1...1101	$2^{32} - 3$
0xFFFFFFFDE	1...1110	$2^{32} - 2$
0xFFFFFFFFF	1...1111	$2^{32} - 1$



Review: Signed Binary Representation

		2'sc binary	decimal
	$-2^3 =$	1000	-8
	$-(2^3 - 1) =$	1001	-7
		1010	-6
		1011	-5
		1100	-4
		1101	-3
		1110	-2
		1111	-1
		0000	0
		0001	1
		0010	2
		0011	3
		0100	4
		0101	5
		0110	6
		0111	7

complement all the bits

0101

1011

and add a 1

and add a 1

0110

1010

complement all the bits

$2^3 - 1 =$

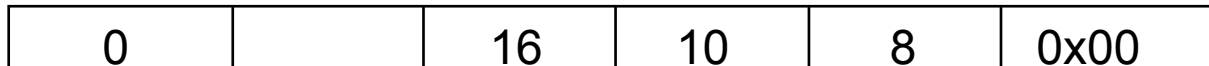
MIPS Shift Operations

- ❑ Need operations to **pack** and **unpack** 8-bit characters into 32-bit words
- ❑ Shifts move all the bits in a word left or right

```
sll $t2, $s0, 8    # $t2 = $s0 << 8 bits
```

```
srl $t2, $s0, 8    # $t2 = $s0 >> 8 bits
```

- ❑ Instruction Format (**R** format)



- ❑ Such shifts are called **logical** because they fill with **zeros**
 - Notice that a 5-bit shamt field is enough to shift a 32-bit value $2^5 - 1$ or **31 bit positions**

MIPS Logical Operations

- There are a number of **bit-wise** logical operations in the MIPS ISA

and \$t0, \$t1, \$t2 # \$t0 = \$t1 & \$t2

or \$t0, \$t1, \$t2 # \$t0 = \$t1 | \$t2

nor \$t0, \$t1, \$t2 # \$t0 = not (\$t1 | \$t2)

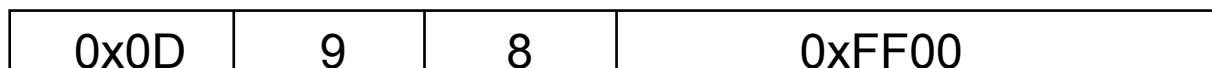
- Instruction Format (**R** format)



andi \$t0, \$t1, 0xFF00 # \$t0 = \$t1 & ff00

ori \$t0, \$t1, 0xFF00 # \$t0 = \$t1 | ff00

- Instruction Format (**I** format)



MIPS Control Flow Instructions

❑ MIPS conditional branch instructions:

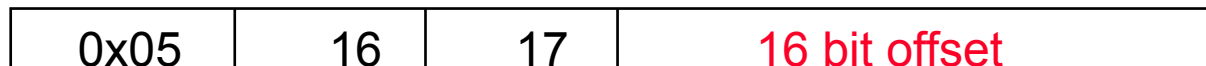
```
bne $s0, $s1, Lbl #go to Lbl if $s0  $\neq$  $s1  
beq $s0, $s1, Lbl #go to Lbl if $s0=$s1
```

● Ex: if (i==j) h = i + j;

```
      bne $s0, $s1, Lbl1  
      add $s3, $s0, $s1
```

```
Lbl1:       ...
```

❑ Instruction Format (I format):

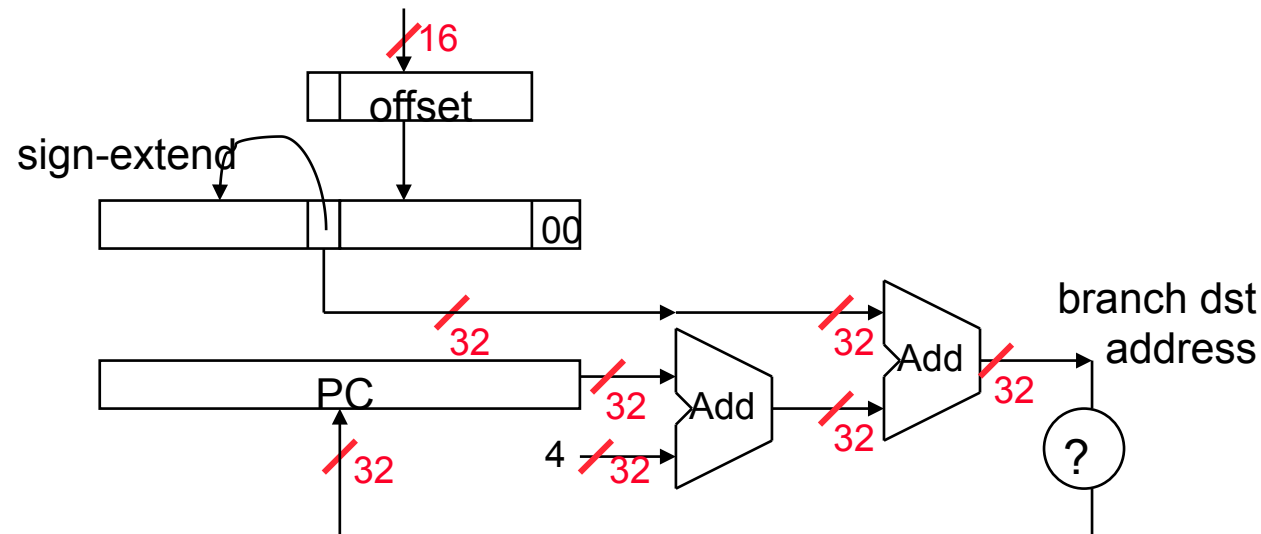


❑ How is the branch destination address specified?

Specifying Branch Destinations

- Use a register (like in lw and sw) added to the 16-bit offset
 - which register? Instruction Address Register (the PC)
 - its use is automatically implied by instruction
 - PC gets updated (PC+4) during the fetch cycle so that it holds the address of the next instruction
 - limits the branch distance to -2^{15} to $+2^{15}-1$ (word) instructions from the (instruction after the) branch instruction, but most branches are local anyway

from the low order 16 bits of the branch instruction



In Support of Branch Instructions

- ❑ We have `beq`, `bne`, but what about other kinds of branches (e.g., branch-if-less-than)? For this, we need yet another instruction, `slt`

- ❑ Set on less than instruction:

```
slt $t0, $s0, $s1    # if $s0 < $s1    then
                    # $t0 = 1          else
                    # $t0 = 0
```

- ❑ Instruction format (**R** format):



- ❑ Alternate versions of `slt`

```
slti $t0, $s0, 25    # if $s0 < 25 then $t0=1 ...
```

```
sltu $t0, $s0, $s1   # if $s0 < $s1 then $t0=1 ...
```

```
sltiu $t0, $s0, 25   # if $s0 < 25 then $t0=1 ...
```

Aside: More Branch Instructions

- ❑ Can use `slt`, `beq`, `bne`, and the fixed value of 0 in register `$zero` to **create** other conditions

- less than `blt $s1, $s2, Label`

```
slt  $at, $s1, $s2    # $at set to 1 if
bne  $at, $zero, Label # $s1 < $s2
```

- less than or equal to `ble $s1, $s2, Label`
- greater than `bgt $s1, $s2, Label`
- great than or equal to `bge $s1, $s2, Label`

- ❑ Such branches are included in the instruction set as pseudo instructions - recognized (and expanded) by the assembler
 - Its why the assembler needs a reserved register (`$at`)

Bounds Check Shortcut

- ❑ Treating signed numbers as if they were unsigned gives a low cost way of checking if $0 \leq x < y$ (index out of bounds for arrays)

```
sltu $t0, $s1, $t2      # $t0 = 0 if
                        # $s1 > $t2
(max)                  # or $s1 < 0
(min)
beq $t0, $zero, IOOB   # go to IOOB if
                        # $t0 = 0
```

- ❑ The key is that negative integers in two's complement look like large numbers in unsigned notation. Thus, an unsigned comparison of $x < y$ also checks if x is negative as well as if x is less than y .

Aside: Branching Far Away

- ❑ What if the branch destination is further away than can be captured in 16 bits?

Instructions for Accessing Procedures

- ❑ MIPS **procedure call** instruction:

```
jal ProcedureAddress #jump and link
```

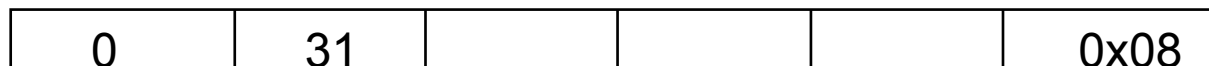
- ❑ Saves PC+4 in register \$ra to have a link to the next instruction for the procedure return
- ❑ Machine format (**J** format):



- ❑ Then can do procedure **return** with a

```
jr $ra #return
```

- ❑ Instruction format (**R** format):

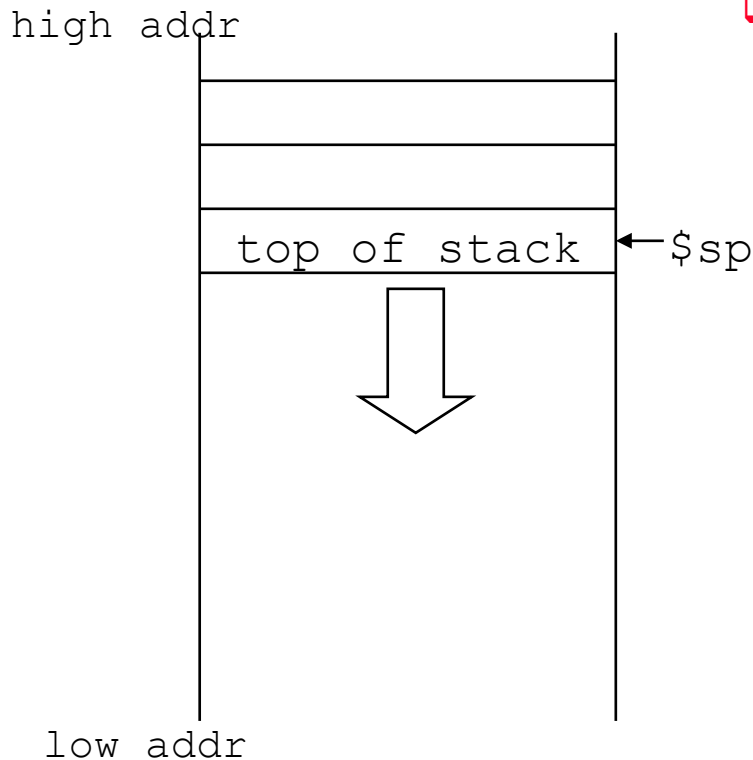


Six Steps in Execution of a Procedure

1. Main routine (**caller**) places parameters in a place where the procedure (**callee**) can access them
 - `$a0 - $a3`: four **argument** registers
2. **Caller** transfers control to the **callee**
3. **Callee** acquires the storage resources needed
4. **Callee** performs the desired task
5. **Callee** places the result value in a place where the **caller** can access it
 - `$v0 - $v1`: two **value** registers for result values
6. **Callee** returns control to the **caller**
 - `$ra`: one **return address** register to return to the point of origin

Aside: Spilling Registers

- What if the **callee** needs to use more registers than allocated to argument and return values?
 - **callee** uses a **stack** – a last-in-first-out queue



- One of the general registers, $\$sp$ ($\$29$), is used to address the stack (which “grows” from high address to low address)

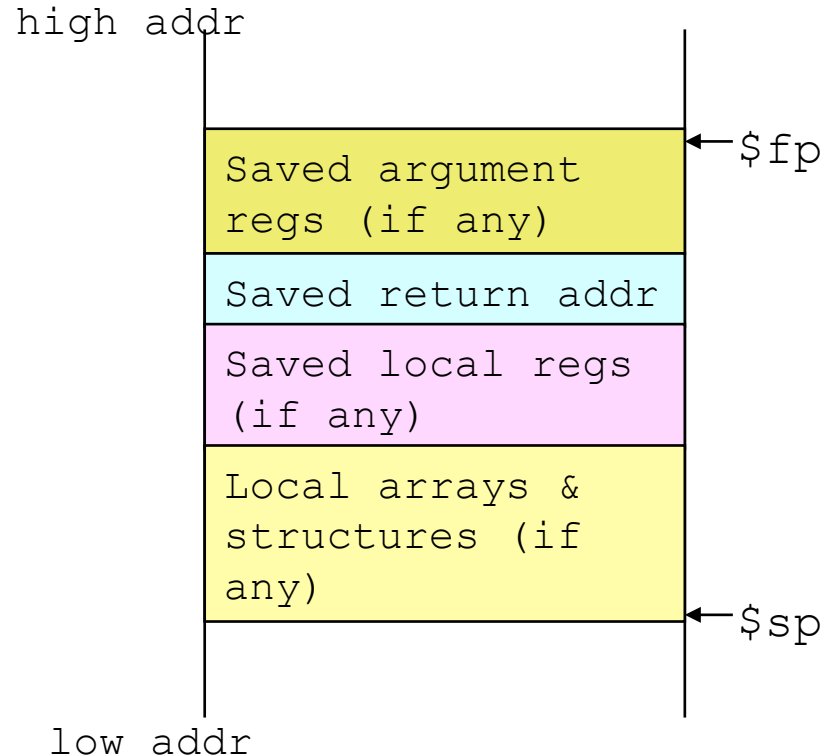
- add data onto the stack – **push**

$\$sp = \$sp - 4$
data **on** stack at new $\$sp$

- remove data from the stack – **pop**

data **from** stack at $\$sp$
 $\$sp = \$sp + 4$

Aside: Allocating Space on the Stack

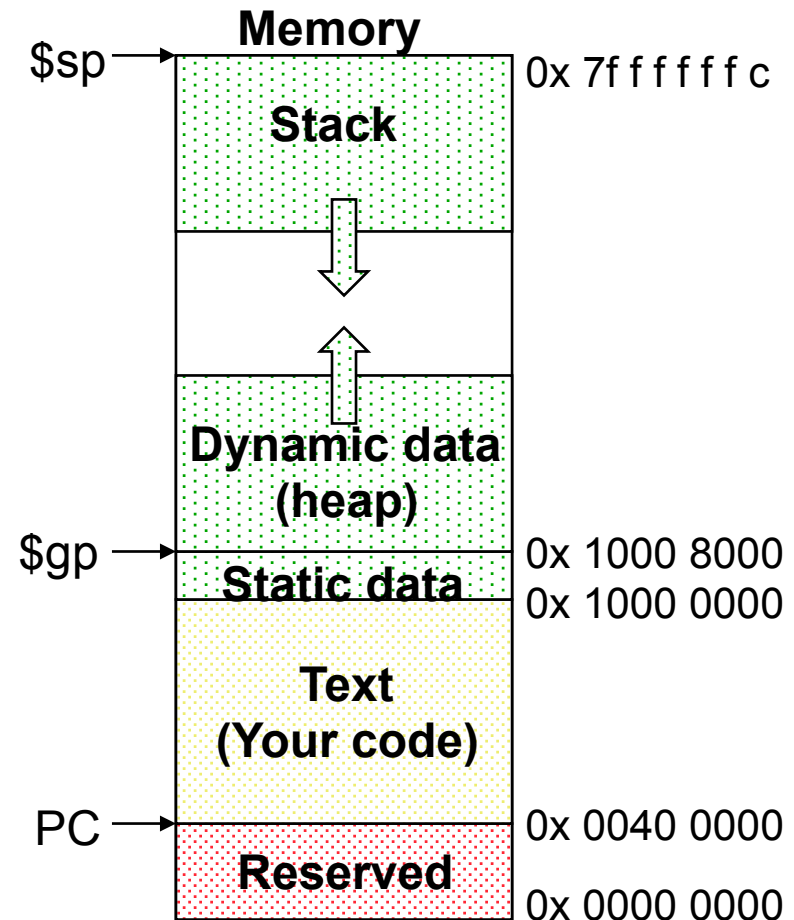


□ The segment of the stack containing a procedure's saved registers and local variables is its **procedure frame** (aka **activation record**)

- The frame pointer ($\$fp$) points to the first word of the frame of a procedure – providing a stable “base” register for the procedure
 - $\$fp$ is initialized using $\$sp$ on a call and $\$sp$ is restored using $\$fp$ on a return

Aside: Allocating Space on the Heap

- ❑ Static data segment for constants and other static variables (e.g., arrays)
- ❑ Dynamic data segment (aka **heap**) for structures that grow and shrink (e.g., linked lists)
 - Allocate space on the heap with `malloc()` and free it with `free()` in C



MIPS Instruction Classes Distribution

- Frequency of MIPS instruction classes for SPEC2006

Instruction Class	Frequency	
	Integer	Ft. Pt.
Arithmetic	16%	48%
Data transfer	35%	36%
Logical	12%	4%
Cond. Branch	34%	8%
Jump	2%	0%

Atomic Exchange Support

- ❑ Need hardware support for synchronization mechanisms to avoid **data races** where the results of the program can change depending on how events happen to occur
 - Two memory accesses from different threads to the same location, and at least one is a write

- ❑ **Atomic exchange** (atomic swap) – interchanges a value in a register for a value in memory atomically, i.e., as one operation (instruction)
 - Implementing an atomic exchange would require both a memory read and a memory write in a single, uninterruptable instruction. An alternative is to have a pair of specially configured instructions

```
ll    $t1, 0($s1)           #load linked
sc    $t0, 0($s1)           #store
conditional
```

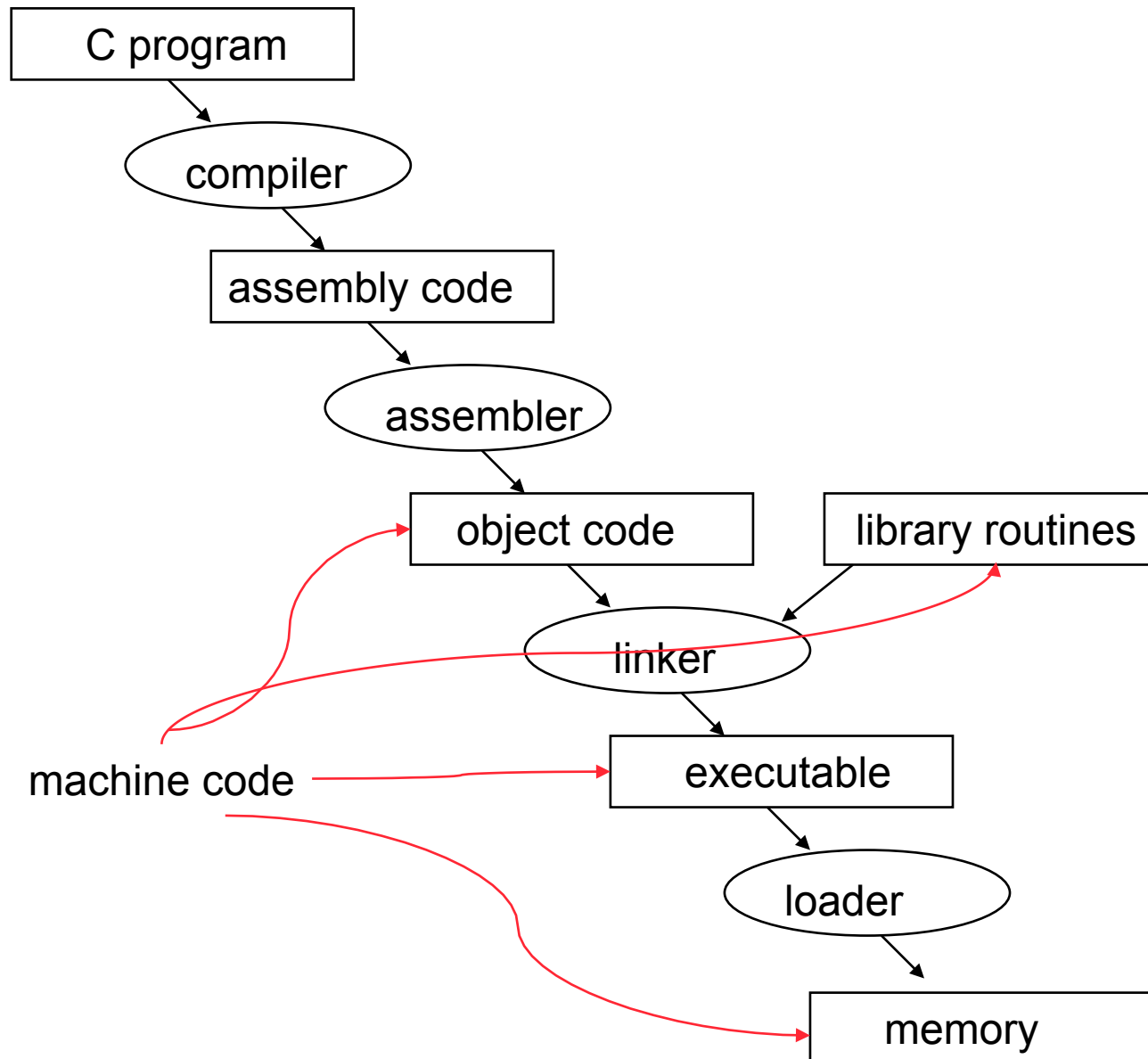
Atomic Exchange with ll and sc

- ❑ If the contents of the memory location specified by the ll are changed before the sc to the same address occurs, the sc fails (returns a zero)

```
try:  add $t0, $zero, $s4      #$t0=$s4 (exchange value)
      ll  $t1, 0($s1)         #load memory value to $t1
      sc  $t0, 0($s1)         #try to store exchange
                                #value to memory,
                                #$t0 will be 0
if fail
      beq $t0, $zero, try     #try again on failure
      add $s4, $zero, $t1     #load value in $s4
```

- ❑ If the value in memory between the ll and the sc instructions changes, then sc returns a 0 in \$t0 causing the code sequence to try again.

The C Code Translation Hierarchy



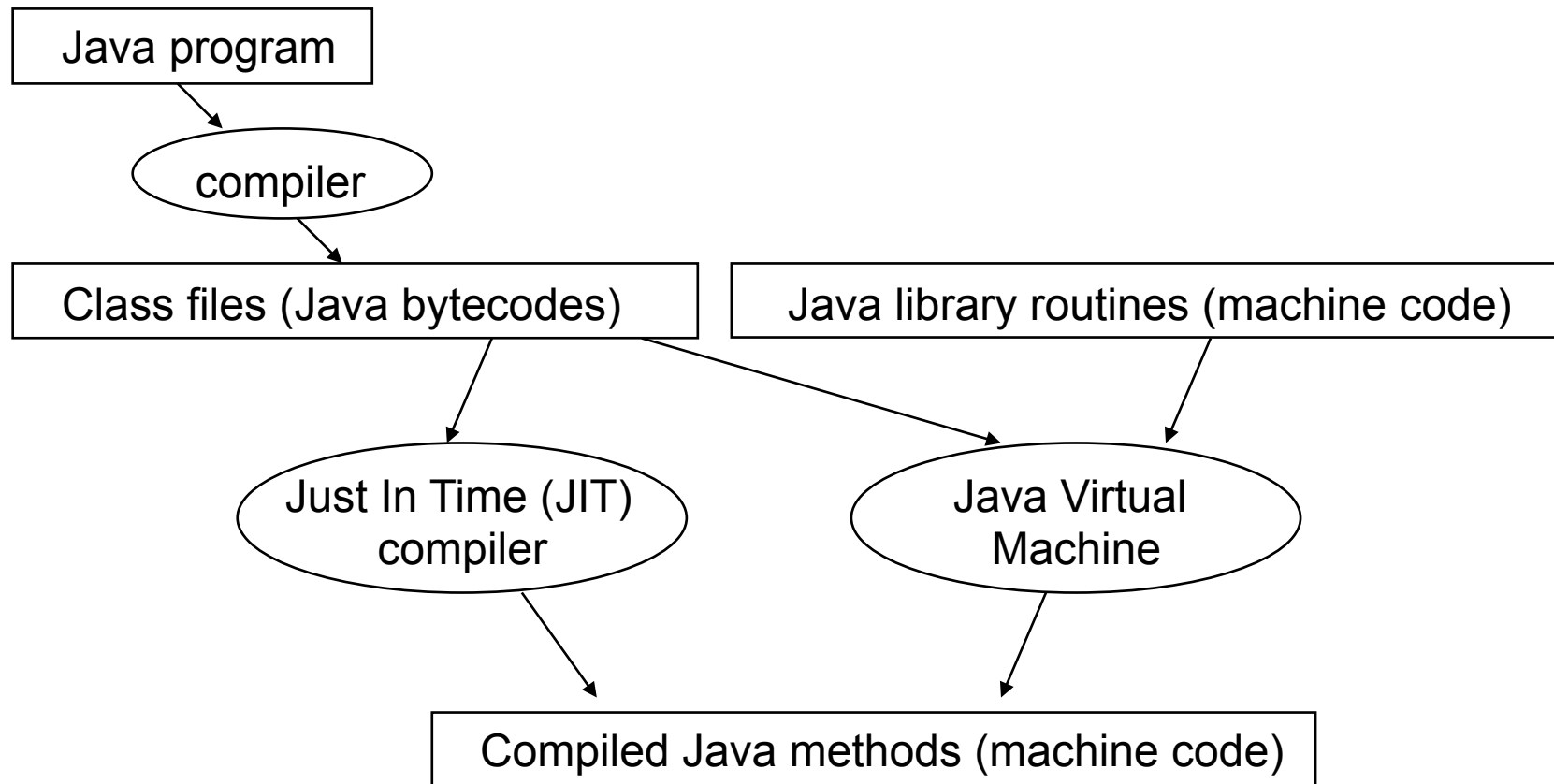
Compiler Benefits

- ❑ Comparing performance for bubble (exchange) sort
 - To sort 100,000 words with the array initialized to random values on a Pentium 4 with a 3.06 clock rate, a 533 MHz system bus, with 2 GB of DDR SDRAM, using Linux version 2.4.20

gcc opt	Relative performance	Clock cycles (M)	Instr count (M)	CPI
None	1.00	158,615	114,938	1.38
O1 (medium)	2.37	66,990	37,470	1.79
O2 (full)	2.38	66,521	39,993	1.66
O3 (proc mig)	2.41	65,747	44,993	1.46

- ❑ The unoptimized code has the best CPI, the O1 version has the lowest instruction count, but the O3 version is the fastest. Why?

The Java Code Translation Hierarchy



Sorting in C versus Java

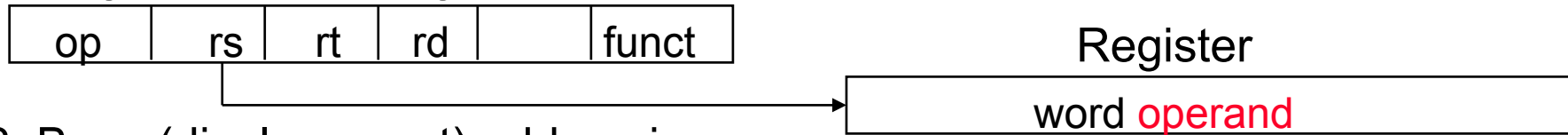
- ❑ Comparing performance for two sort algorithms in C and Java
 - The JVM/JIT is Sun/Hotspot version 1.3.1/1.3.1

	Method	Opt	Bubble	Quick	Speedup quick vs bubble
			Relative performance		
C	Compiler	None	1.00	1.00	2468
C	Compiler	O1	2.37	1.50	1562
C	Compiler	O2	2.38	1.50	1555
C	Compiler	O3	2.41	1.91	1955
Java	Interpreted		0.12	0.05	1050
Java	JIT compiler		2.13	0.29	338

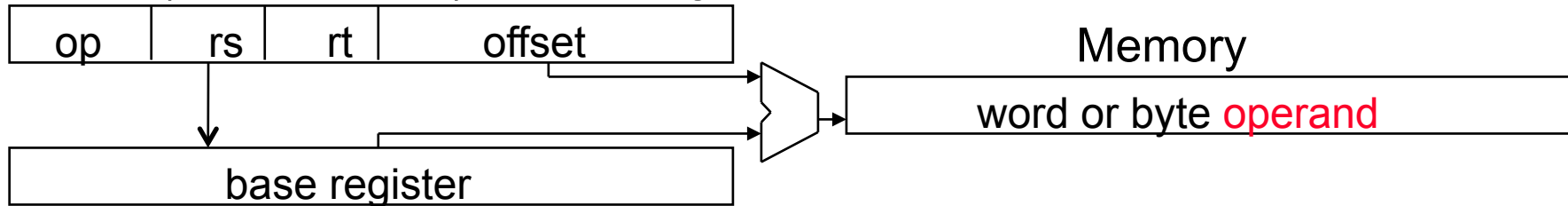
- ❑ Observations?

Addressing Modes Illustrated

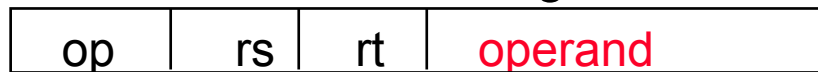
1. Register addressing



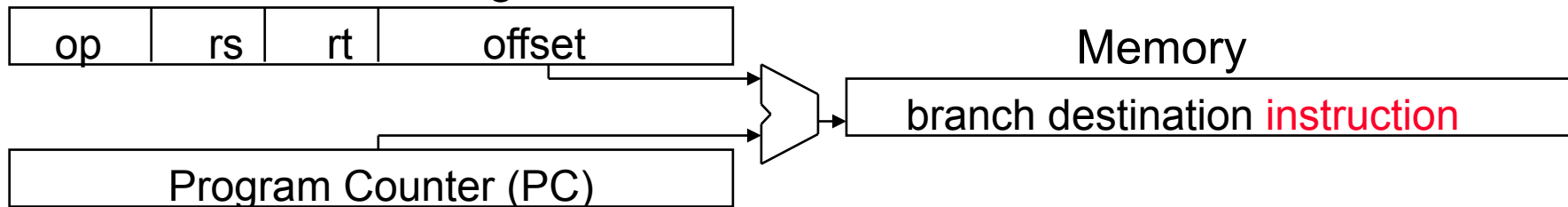
2. Base (displacement) addressing



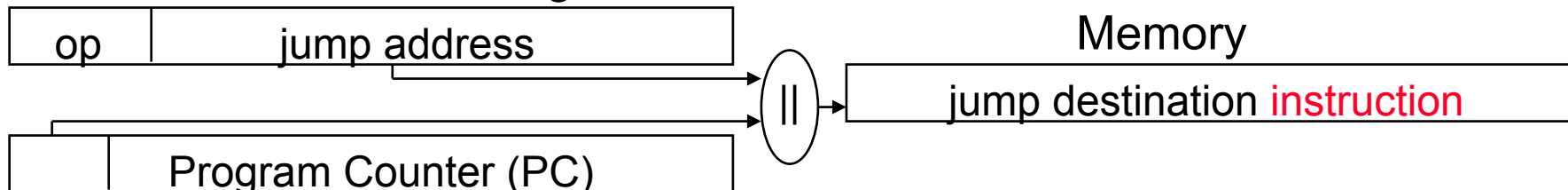
3. Immediate addressing



4. PC-relative addressing



5. Pseudo-direct addressing



MIPS Organization So Far

