北京大学
PEKING UNIVERSITY

AI ASIC: Design and Practice (ADaP)

Fall 2024

CPU Organization & Architecture

燕博南

Instruction Set Architecture (ISA)

- The contract between software and hardware
- Typically described by giving all the programmer-visible state (registers + memory) plus the semantics of the instructions that operate on that state
- IBM 360 was first line of machines to separate ISA from implementation (aka. *microarchitecture*)
- Many implementations possible for a given ISA
  - E.g., Soviets built code-compatible clones of the IBM360, as did Amdahl after he left IBM.
  - E.g.2., AMD, Intel, VIA processors run the AMD64 ISA
  - E.g.3: many cellphones use the ARM ISA with implementations from many different companies including Apple, Qualcomm, Samsung, Huawei, etc.
- We use RISC-V as standard ISA in class (www.riscv.org)
  - Many companies and open-source projects build RISC-V implementations

ISA to Microarchitecture Mapping

- ISA often designed with particular microarchitectural style in mind, e.g.,
  Accumulator ⇒ hardwired, unpipelined
  CISC ⇒ microcoded
  RISC ⇒ hardwired, pipelined
  VLIW ⇒ fixed-latency in-order parallel pipelines
  JVM ⇒ software interpretation
- But can be implemented with any microarchitectural style
  - Intel Ivy Bridge: hardwired pipelined CISC (x86) machine (with some microcode support)
  - Apple M1 (native ARM ISA, emulates x86 in software)
  - Spike: Software-interpreted RISC-V machine
  - ARM Jazelle: A hardware JVM processor
  - **This lecture: a microcoded RISC-V machine**

Control versus Datapath

- Processor designs can be split between
  - ***datapath***, where numbers are stored and arithmetic operations computed, and
  - ***control***, which sequences operations on datapath

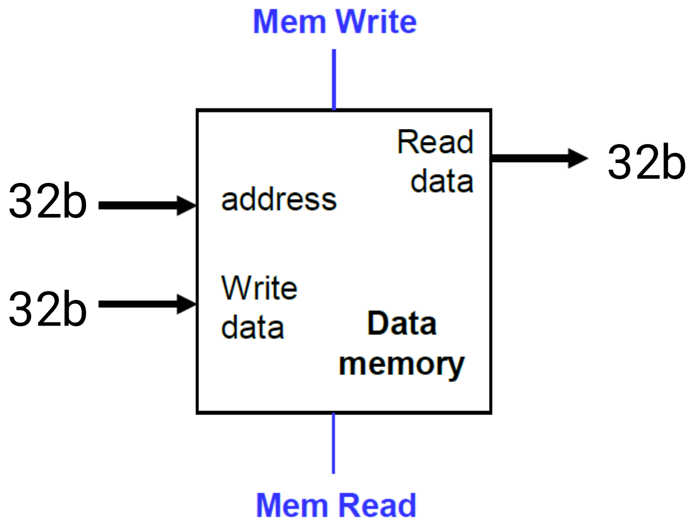
A computer is just a big fancy
state machine.

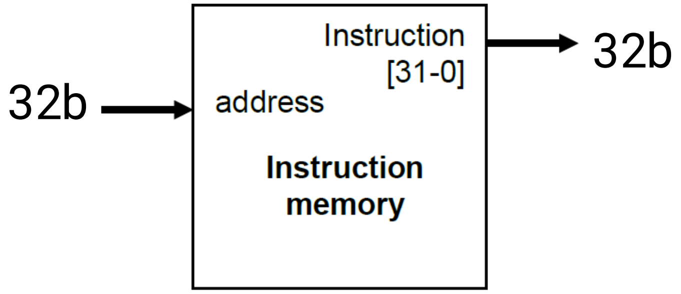
John von Neumann

- In the old days, “programming” involved actually changing a machine’s physical configuration by flipping switches or connecting wires.
  - A computer could run just one program at a time.
  - Memory only stored data that was being operated on.
- Then around 1944, John von Neumann and others got the idea to **encode instructions in a format that could be stored in memory just like data.**
  - The processor interprets and executes instructions from memory.
  - One machine could perform many different tasks, just by loading different programs into memory.
  - The“stored program” design is often called a Von Neumann machine.

Memories

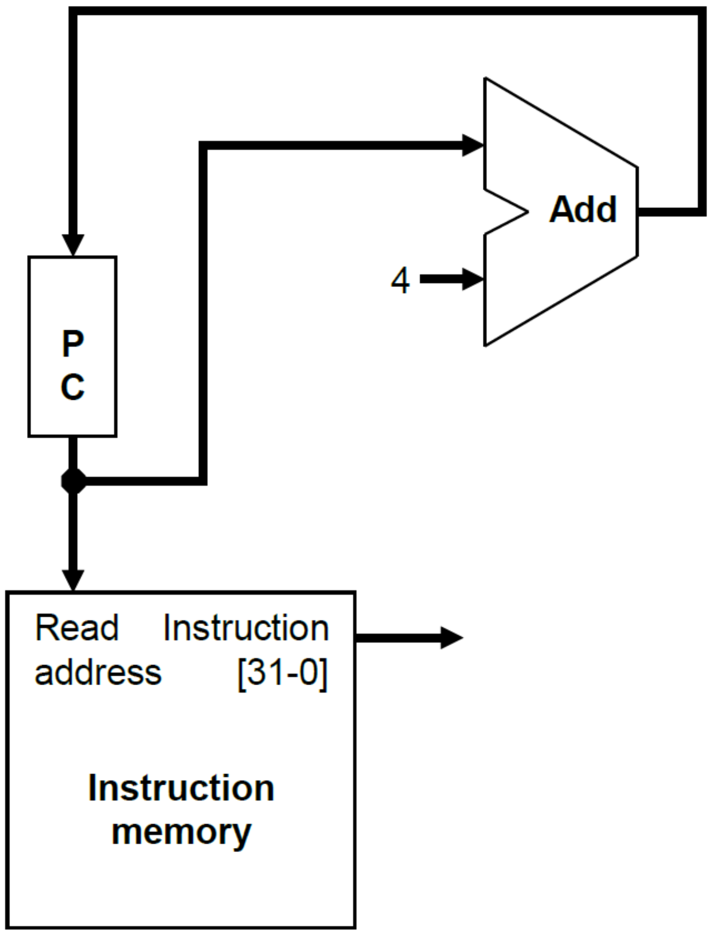
- Harvard architecture :
  - programs and data stored in separate memories.
- Blue lines represent control signals. MemRead and MemWrite should be set to 1 if the data memory is to be read or written respectively, and 0 otherwise.
- When a control signal does something when it is set to 1, we call it active high(vs. active low) because 1 is usually a higher voltage than 0.
- Pretend it's already loaded with a program, which doesn't change while it's running.





Instruction Fetching

- The CPU is in a infinite loop
- The program counter or PC register holds the address of the current instruction
- Given our instruction is 4 byte (32b) long
  - >> PC = PC + 4 after obtaining an instruction



Encoding R-type instructions

- Register-to-register arithmetic instructions use the R-type format.
  - op is the instruction opcode, and func specifies a particular arithmetic operation
  - rs, rt and rd are source and destination registers.

op	rs	rt	rd	shamt	func
6 bits	5 bits	5 bits	5 bits	5 bits	6 bits

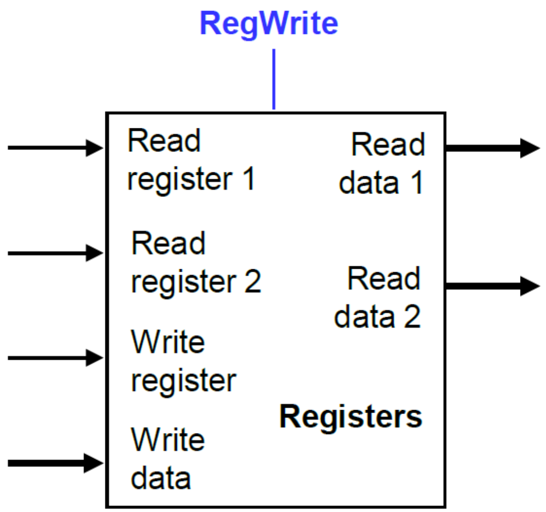
- Example

Now pretend you know assembly!

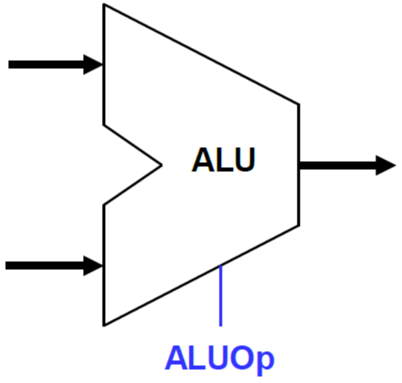
add \$s4, \$t1, \$t2	000000	01001	01010	10100	00000	1000000

Register File & ALU

- R-type instructions must access registers and an ALU
- Our register file stores thirty-two 32-bit values.
  - Each register specifier is 5 bits long.
  - You can read from two registers at a time (2 ports).
  - RegWrite is 1 if a register should be written.
- Here's a simple ALU with five operations, selected by a 3-bit control signal ALUOp.

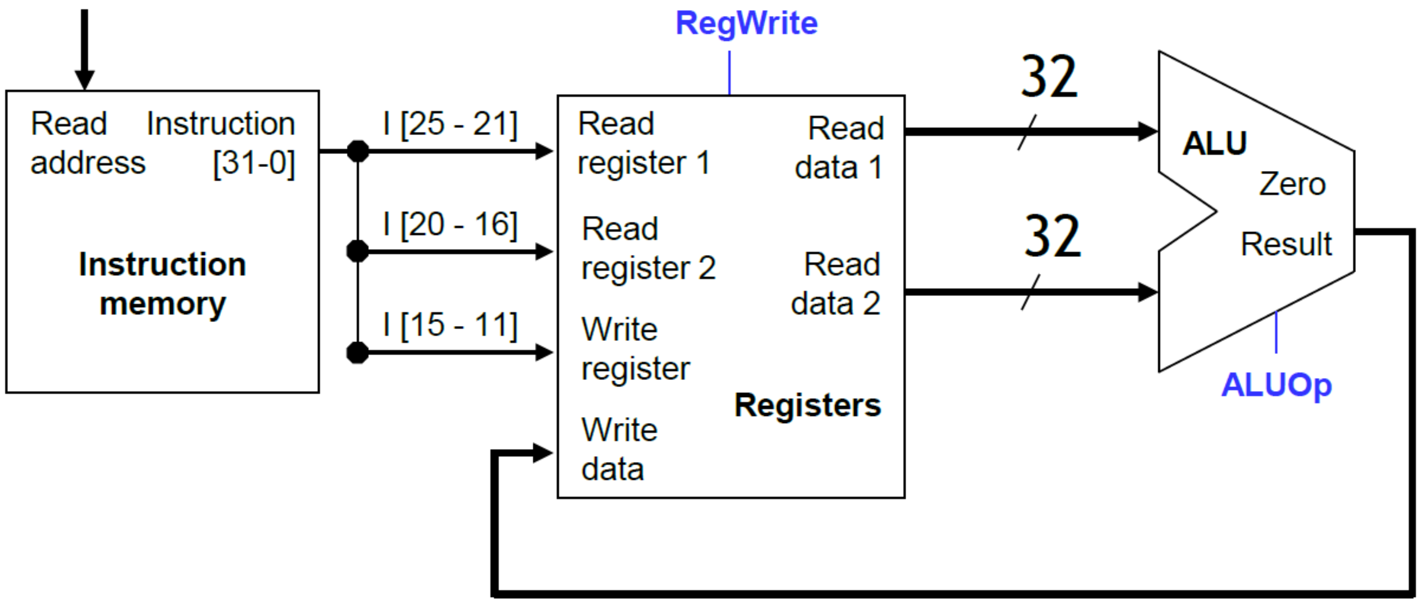


ALUOp	Function
000	and
001	or
010	add
110	subtract
111	slt



Executing an R-type instruction

PEKING UNIVERSITY



op	rs	rt	rd	shamt	func
31 26	25 21	20 16	15 11	10 6	5 0

- **Fetch an instruction from "instruction memory"**
- **Fetch data from registers rs & rt**
- **ALU does computation**
- **Put results into rd**

Encoding I-type instructions

- Immediate number instructions (I-type)
  - Rt is the destination for lw, but a source for beq and sw
  - Address is a 16-bit signed constant

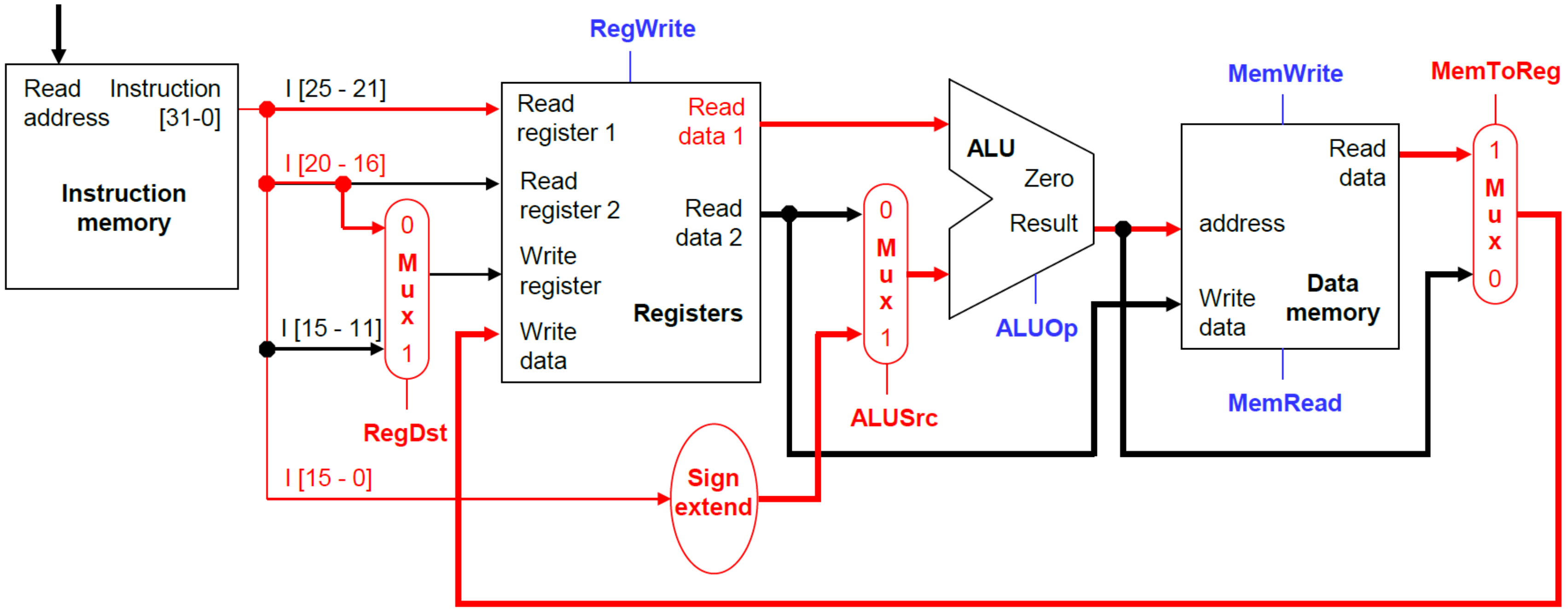
immediate	rs1	funct3	rd	opcode
12 bits	5 bits	3 bits	5 bits	7 bits

- Example

```
ld x9, 64(x22) // Temporary reg x9 gets A[8]
```

31 — 20	19 — 15	14 — 12	11 — 7	6 — 0
imm[11:0]	rs1	funct3	rd	opcode
12	5	3	5	7
offset[11:0]	base	width	dest	LOAD

Accessing Data Memory




```
ld x9, 64(x22) // Temporary reg x9 gets A[8]
```

Data memory Address: (the content in x22)+64
Operation: load the data in the “data memory” into x9

Branches

- For branch instructions, next PC should be obtained in the

```
        beq   $at, $0, L
        add   $v1, $v0, $0
        add   $v1, $v1, $v1
        j     somewhere
L:      add   $v1, $v0, $v0
```

31	30 25	24 20	19 15	14 12	11 8	7	6 0
imm[12]	imm[10:5]	rs2	rs1	funct3	imm[4:1]	imm[11]	opcode
1	6	5	5	3	4	1	7
offset[12|10:5]		src2	src1	BEQ/BNE	offset[11|4:1]		BRANCH
offset[12|10:5]		src2	src1	BLT[U]	offset[11|4:1]		BRANCH
offset[12|10:5]		src2	src1	BGE[U]	offset[11|4:1]		BRANCH

BEQ: if rs1==rs2, then to go to the current PC+offset

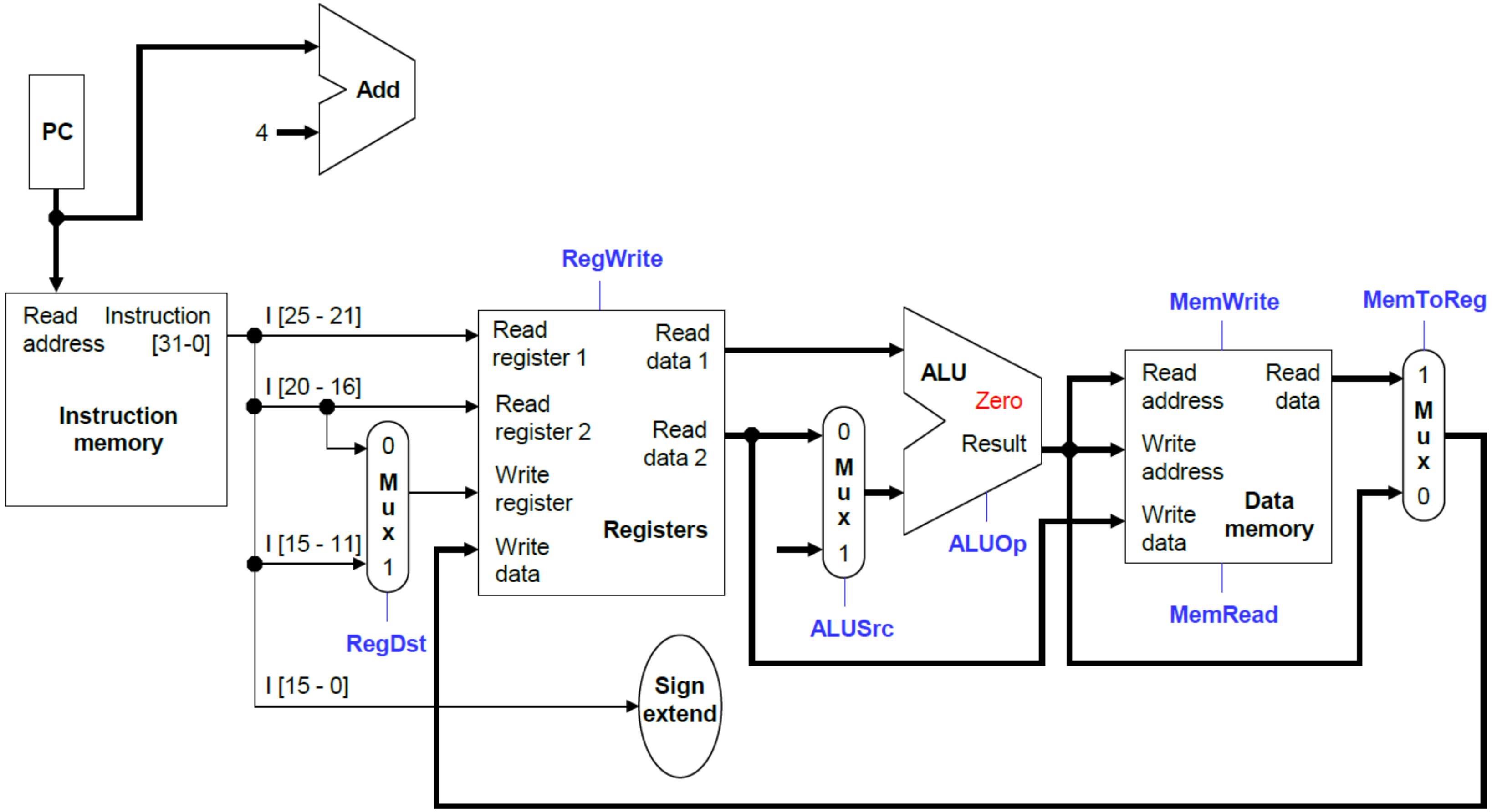
So Execute BEQ should be:

1. Fetch the instruction, like beq $at, $0, offset, from memory.

2. Compare $at and $0

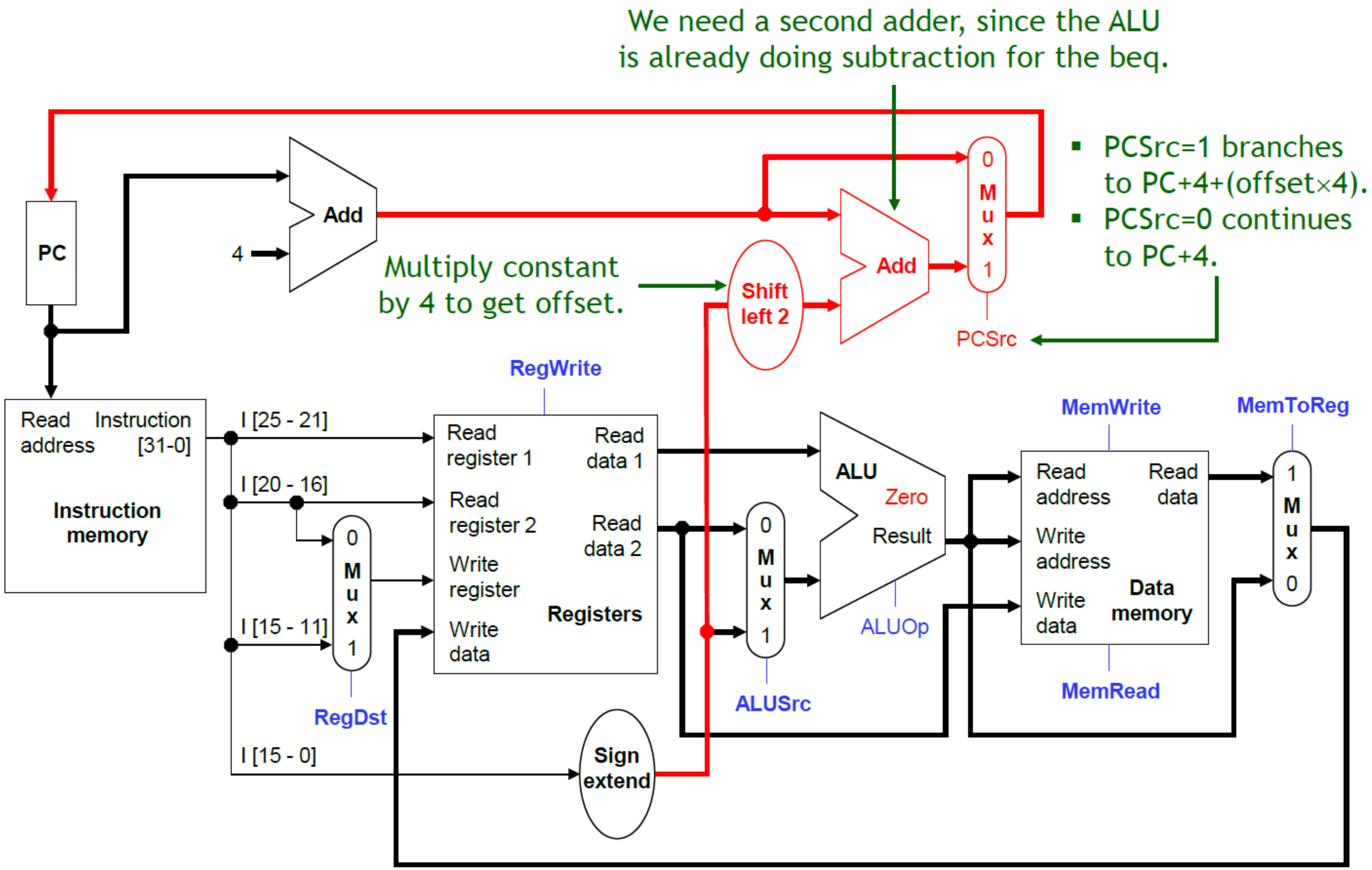
3. If yes, next PC = PC + offset * 4 byte/instruction

Hardware



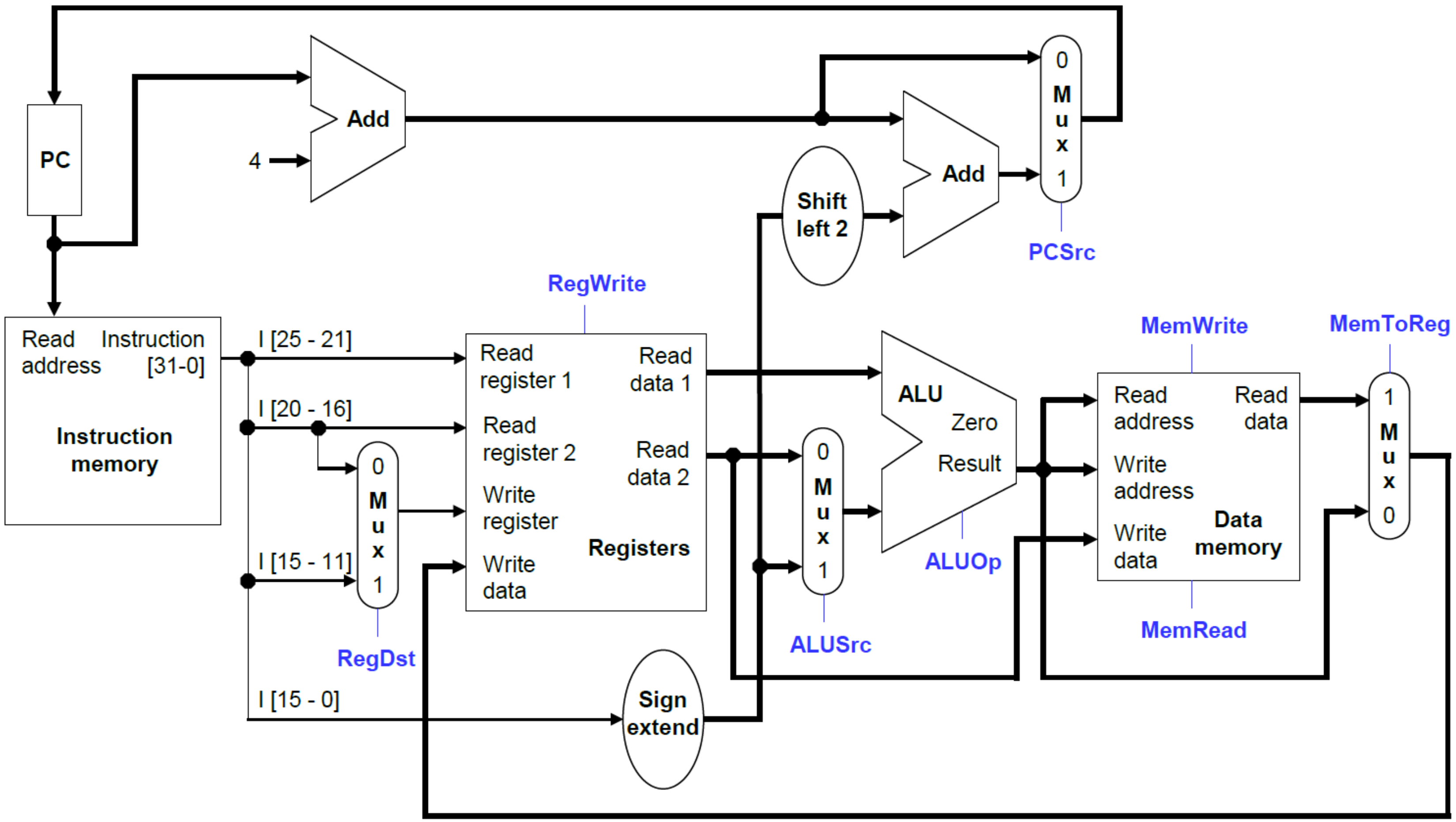
PC
Add
4
Read address
Instruction [31-0]
Instruction memory
I [25 - 21]
I [20 - 16]
I [15 - 11]
I [15 - 0]
RegDst
0
M
u
x
1
RegWrite
Read register 1
Read register 2
Write register
Write data
Read data 1
Read data 2
Registers
Sign extend
ALUSrc
ALU
Zero
Result
ALUOp
MemWrite
MemRead
Read address
Write address
Write data
Read data
Data memory
MemToReg

Hardware



We need a second adder, since the ALU
is already doing subtraction for the beq.
PCSrc=1 branches to PC+4+(offset×4).
PCSrc=0 continues to PC+4.
Multiply constant by 4 to get offset.
PC
Add
4
Shift left 2
Add
0
Mux
1
PCSrc
RegWrite
Read address
Instruction [31-0]
Instruction memory
I [25 - 21]
I [20 - 16]
I [15 - 11]
I [15 - 0]
0
Mux
1
RegDst
Read register 1
Read register 2
Write register
Write data
Read data 1
Read data 2
Registers
Sign extend
0
Mux
1
ALUSrc
ALU
Zero
Result
ALUOp
MemWrite
Read address
Write address
Write data
Read data
Data memory
MemRead
MemToReg
1
Mux
0

Final Hardware



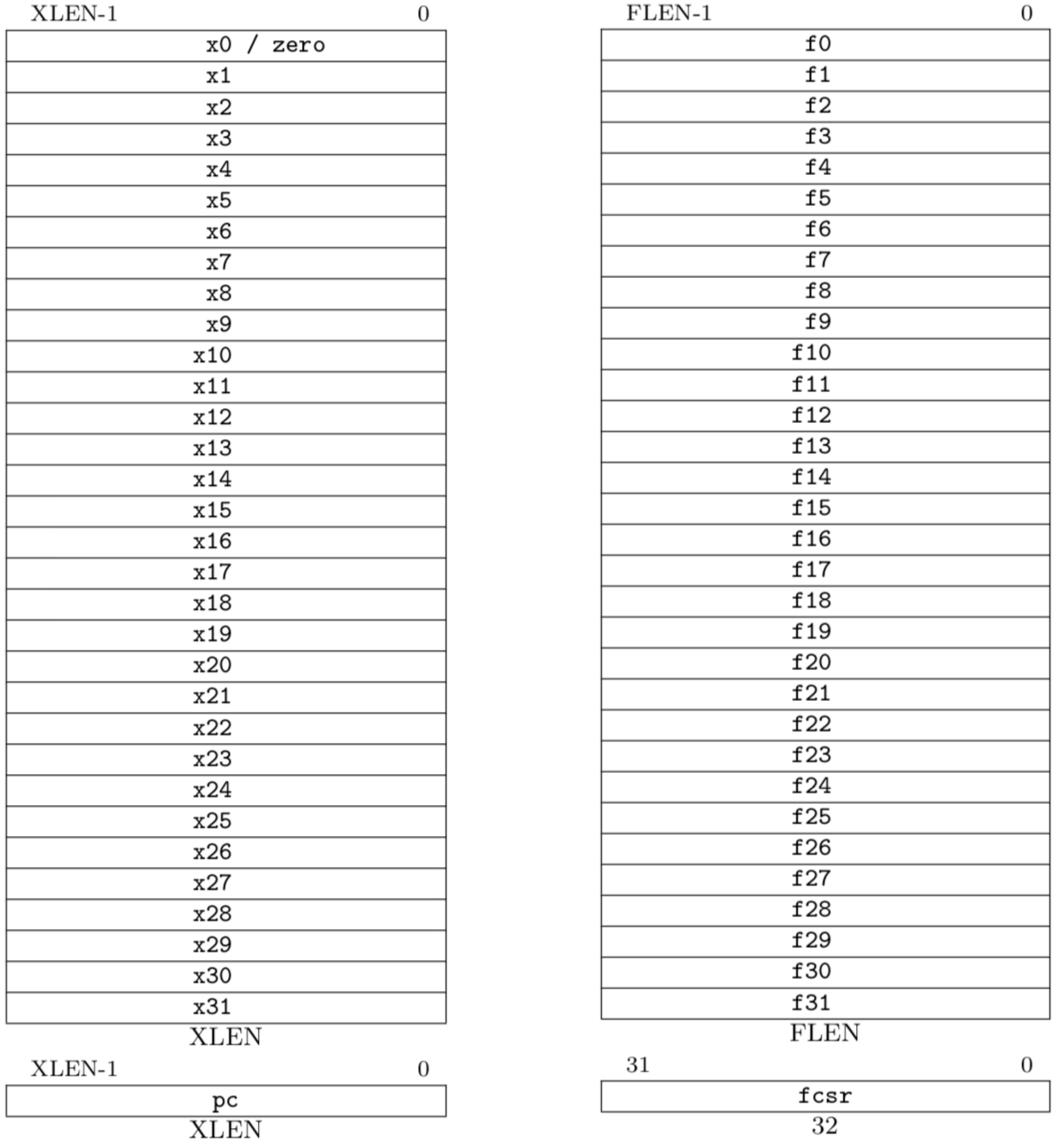
Review: RV32I Processor State

Program counter (**pc**)

32x32-bit integer registers (**x0-x31**)
- **x0** always contains a 0

32 floating-point (FP) registers (**f0-f31**)
- each can contain a single- or double-precision FP value (32-bit or 64-bit IEEE FP)

FP status register (**fcsr**), used for FP rounding mode & exception reporting



RISC-V Instruction Encoding

		xxxxxxxxxxxxxxaa	16-bit (aa ≠ 11)
	xxxxxxxxxxxxxxxx	xxxxxxxxxxxbbb11	32-bit (bbb ≠ 111)
· · ·xxxx	xxxxxxxxxxxxxxxx	xxxxxxxxxx011111	48-bit
· · ·xxxx	xxxxxxxxxxxxxxxx	xxxxxxxxx0111111	64-bit
· · ·xxxx	xxxxxxxxxxxxxxxx	xnnnxxxxx1111111	(80+16*nnn)-bit, nnn≠111
· · ·xxxx	xxxxxxxxxxxxxxxx	x111xxxxx1111111	Reserved for ≥192-bits
Byte Address: base+4	base+2	base	

- Can support variable-length instructions.
- Base instruction set (RV32) always has fixed 32-bit instructions lowest two bits = 11_2
- All branches and jumps have targets at 16-bit granularity (even in base ISA where all instructions are fixed 32 bits)

RISC-V Instruction Formats

Additional opcode bits/immediate

Reg. Source 2

Reg. Source 1

Destination Reg.

7-bit opcode field (but low 2 bits $=11_2$)

31	30 … 25	24 … 21	20	19 … 15	14 … 12	11 … 8	7	6 … 0	
funct7		rs2		rs1	funct3	rd		opcode	R-type
imm[11:0]				rs1	funct3	rd		opcode	I-type
imm[11:5]		rs2		rs1	funct3	imm[4:0]		opcode	S-type
imm[12]	imm[10:5]	rs2		rs1	funct3	imm[4:1]	imm[11]	opcode	B-type
imm[31:12]						rd		opcode	U-type
imm[20]	imm[10:1]		imm[11]	imm[19:12]		rd		opcode	J-type

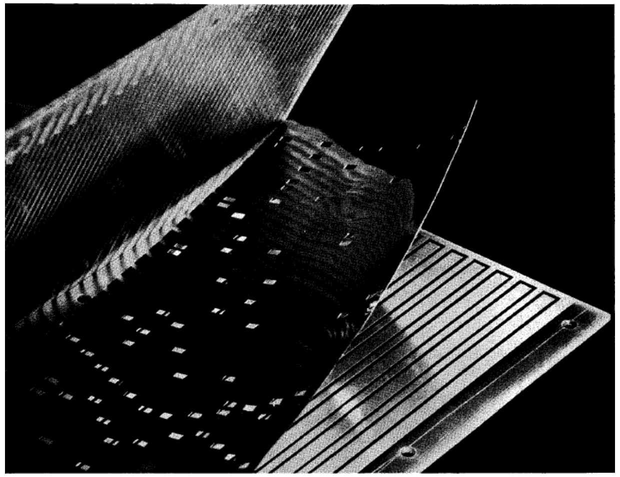
Inside Instruction Memory

Address				Data	
μPC	Opcode	Cond?	Busy?	Control Lines	Next μPC
fetch0	X	X	X	MA,A:=PC	fetch1
fetch1	X	X	1		fetch1
fetch1	X	X	0	IR:=Mem	fetch2
fetch2	ALU	X	X	PC:=A+4	ALU0
fetch2	ALUI	X	X	PC:=A+4	ALUI0
fetch2	LW	X	X	PC:=A+4	LW0
….					
ALU0	X	X	X	A:=Reg[rs1]	ALU1
ALU1	X	X	X	B:=Reg[rs2]	ALU2
ALU2	X	X	X	Reg[rd]:=ALUOp(A,B)	fetch0

Back into History

PEKING UNIVERSITY 北京大学

IBM 360



雷军像诗一样的代码：汇编（1994年）

```
GoINT1C: db 0eah
oldInt1C_addr dw 0, 0
newINT1C:
test cs:Status, SKbit
jnz GoINT1C
cmp cs:StopFlag, 0
jz @@0
;
; Mr. Lei 2/8/1993
; Problem: if WPS quit and reenter, old RI cann't control keyboard. ;
push ds
push ax
xor ax, ax
mov ds, ax
mov ax, ds:[94]
cmp ax, offset NewInt9
pop ax
pop ds
jnz GoINT1C
mov cs:StopFlag, 0

@@0: push ax
push ds
push es
xor ax, ax
mov ds, ax
mov es, ds:[94+2]
cmp word ptr es:[101h], 'IE' ; 'LEI'
jz @@1
cli
mov cs:StopFlag, 1
mov ax, ds:[94]
mov cs:oldINT9_addr2, ax
mov ax, ds:[94+2]
mov cs:oldINT9_addr2[2], ax
mov ds:[94], offset newINT9_2
mov ds:[94+2], cs
sti
@@1: pop es
pop ds
pop ax
jmp GoINT1C
```

```
; ------------ KERNEL PROGRAM ---------------------------------------------
RemoveTSR:
pop ax
cli ; Set stack
mov sp, cs
mov ss, sp
mov sp, 100h
sti
push ax

cmp cs:Power, True
jnz @@1
call Init8259
@@1:
push cs
pop ds
@@_0:
mov ax,ds:NextDataSeg
cmp ax, -1
jz @@_1
mov cs:PrevDataSeg, ds
mov ds, ax
jmp @@_0
@@_1: mov si,ds:DataBegin
mov cs:WorkSeg, ds
lodsw
cmp ax, 'XX'
jz @@_2
call Beep
ret
@@_2:
call RestoreEnvStr
call RestoreMCB ; restore current mcb
call CloseFiles
call RestorePort
call RestoreLEDs
call RestoreVecList ; Restore vectors list
call RestoreFloppyParam
cmp cs:Power, True
jnz @@2
call RestoreCVTchain ; Restore cvt chain
call RestoreMemoryManager
@@2:
call RestoreBiosData
call Enable8259
mov ah, 1 int 16h
```

IBM 360

ISA Compatible Computers

	M30	M40	M50	M65
Datapath width (bits)	8	16	32	64
µinst width (bits)	50	52	85	87
µcode size (K µinsts)	4	4	2.75	2.75
µstore cycle (ns)	750	625	500	200
memory cycle (ns)	1500	2500	2000	750
Rental fee ($K/month)	4	7	15	35

- Only the fastest models (75 and 95) were hardwired

Assembly Language Snap Tutorial

Registers

北京大学 PEKING UNIVERSITY

Register	ABI Name	Description	Saver
x0	zero	Hard-wired zero	—
x1	ra	Return address	Caller
x2	sp	Stack pointer	Callee
x3	gp	Global pointer	—
x4	tp	Thread pointer	—
x5	t0	Temporary/alternate link register	Caller
x6–7	t1–2	Temporaries	Caller
x8	s0/fp	Saved register/frame pointer	Callee
x9	s1	Saved register	Callee
x10–11	a0–1	Function arguments/return values	Caller
x12–17	a2–7	Function arguments	Caller
x18–27	s2–11	Saved registers	Callee
x28–31	t3–6	[illegible]	Caller
f0–7	ft0–7	FP temporaries	[illegible]
f8–9	fs0–1	FP saved registers	Callee
f10–11	fa0–1	FP arguments/return values	Caller
f12–17	fa2–7	FP arguments	Caller
f18–27	fs2–11	FP saved registers	Callee
f28–31	ft8–11	FP temporaries	[illegible]

Temporaries

Basic (Integer) Commends

Instruction Example	Description
lb t0, 8(sp)	Loads (dereferences) from memory address (sp + 8) into register t0. lb = load byte, lh = load halfword, lw = load word, ld = load doubleword.
sb t0, 8(sp)	Stores (dereferences) from register t0 into memory address (sp + 8). sb = store byte, sh = store halfword, sw = store word, sd = store doubleword.
add a0, t0, t1	Adds value of t0 to the value of t1 and stores the sum into a0.
addi a0, t0, -10	Adds value of t0 to the value -10 and stores the sum into a0.
sub a0, t0, t1	Subtracts value of t1 from value of t0 and stores the difference in a0.
mul a0, t0, t1	Multiplies the value of t0 to the value of t1 and stores the product in a0.
div a1, s3, t3	Dividies the value of t3 (denominator) from the value of s3 (numerator) and stores the quotient into the register a1.
rem a1, s3, t3	Divides the value of t3 (denominator) from the value of s3 (numerator) and stores the remainder into the register a1.

and a3, t3, s3	Performs logical AND on operands t3 and s3 and stores the result into the register a3.
or a3, t3, s3	Performs logical OR on operands t3 and s3 and stores the result into the register a3.
xor a3, t3, s3	Performs logical XOR on operands t3 and s3 and stores the result into the register a3.

sub a0, zero, a1

Translate: a0 = 0 – a1

Floating-Point Assembly

```
# Load a double-precision value
flw     ft0, 0(sp)
# ft0 now contains whatever we loaded from memory + 0
flw     ft1, 4(sp)
# ft1 now contains whatever we loaded from memory + 4
fadd.s  ft2, ft0, ft1
# ft2 is now ft0 + ft1
```

RISC-V supports floating-point

In fact, RISC-V has many modules:

Base	Version	Status
RVWMO	2.0	**Ratified**
RV32I	**2.1**	**Ratified**
RV64I	**2.1**	**Ratified**
RV32E	*1.9*	*Draft*
RV128I	*1.7*	*Draft*
Extension	Version	Status
M	**2.0**	**Ratified**
A	**2.1**	**Ratified**
F	**2.2**	**Ratified**
D	**2.2**	**Ratified**
Q	**2.2**	**Ratified**
C	**2.0**	**Ratified**
Counters	*2.0*	*Draft*
L	*0.0*	*Draft*
B	*0.0*	*Draft*
J	*0.0*	*Draft*
T	*0.0*	*Draft*
P	*0.2*	*Draft*
V	*0.7*	*Draft*
Zicsr	**2.0**	**Ratified**
Zifencei	**2.0**	**Ratified**
Zam	*0.1*	*Draft*
Ztso	*0.1*	*Frozen*

Branching

```c
for (int i = 0;i < 10;i++) {
    // Repeated code goes here.
}
```

```asm
# t0 = 0
li      t0, 0
li      t2, 10
loop_head:
bge     t0, t2, loop_end
# Repeated code goes here
addi    t0, t0, 1
j       loop_head
loop_end:
```

Example: Use the Stack

sp is a special register that is a stack

Statck: last in first out (LIFO)

```
addi    sp, sp, -8
sd      ra, 0(sp)
call    printf
ld      ra, 0(sp)
addi    sp, sp, 8
ret
```

C Function

```c
void my_function();
```

```asm
my_function:
    # Prologue
    addi    sp, sp, -32
    sd      ra, 0(sp)
    sd      a0, 8(sp)
    sd      s0, 16(sp)
    sd      s1, 24(sp)

    # Epilogue
    ld      ra, 0(sp)
    ld      a0, 8(sp)
    ld      s0, 16(sp)
    ld      s1, 24(sp)
    addi    sp, sp, 32
    ret
```

Good News is …

- We have compiler that can convert C code into assembly
- 搭建RISC-V编译环境与运行环境 - 知乎 (zhihu.com)
- riscv-collab/riscv-gnu-toolchain: GNU toolchain for RISC-V, including GCC (github.com)
- Compiler Explorer (godbolt.org)
- RISC-V Interpreter (cornell.edu)
- riscv-assembler (riscvassembler.org)