Merging datapaths: (`add`,`lw`, `sw`)

The datapaths that we saw last lecture considered each instruction in isolation. I drew only those elements that were needed for each instruction. Of course, the various wires and inputs are hardware and thus they are there in every instruction. In this lecture we merge the various datapaths. We do so using multiplexors to select from different possible inputs for circuit elements, for example, to select which values are to be written into registers or Memory on each clock cycle. We also need to say more about control signals, for example, the selector signals for the multiplexors.

The figure on the following page shows how the `add`, `lw`, `sw` datapaths from last lecture are merged into a single circuit. I have not draw a single "control unit", but instead I have just shown the various control lines (in green) where they are needed. Let's discuss these control signals one by one. I'll begin by discussing the controls of the three multiplexors, shown in blue.

- RegDst: Recall the R and I format instructions have the form:

bits	26-31	21-25	16-20	11-15	6-10	0-5
R-format	op	*rs*	*rt*	*rd*	shamt	funct
numbits	6	5	5	5	5	6

I-format	op	*rs*	*rt*	immediate
numbits	6	5	5	16

  For either R or I format instructions, at least one register is *read*. For this reason, one of the register slots in the instruction format is always used for a read. This is the *rs* (source) slot, namely bits [21-25]. A 5 bit line is always fed into the ReadReg1 input and specifies which register we read from.

  It is not true, however, that an instruction always *writes* to one of the 32 registers. For example, `sw` does not write to a register. The *rt* slot, namely bits [16-20], holds a second register number which `add` and `sw` read from, and which `lw` writes to.

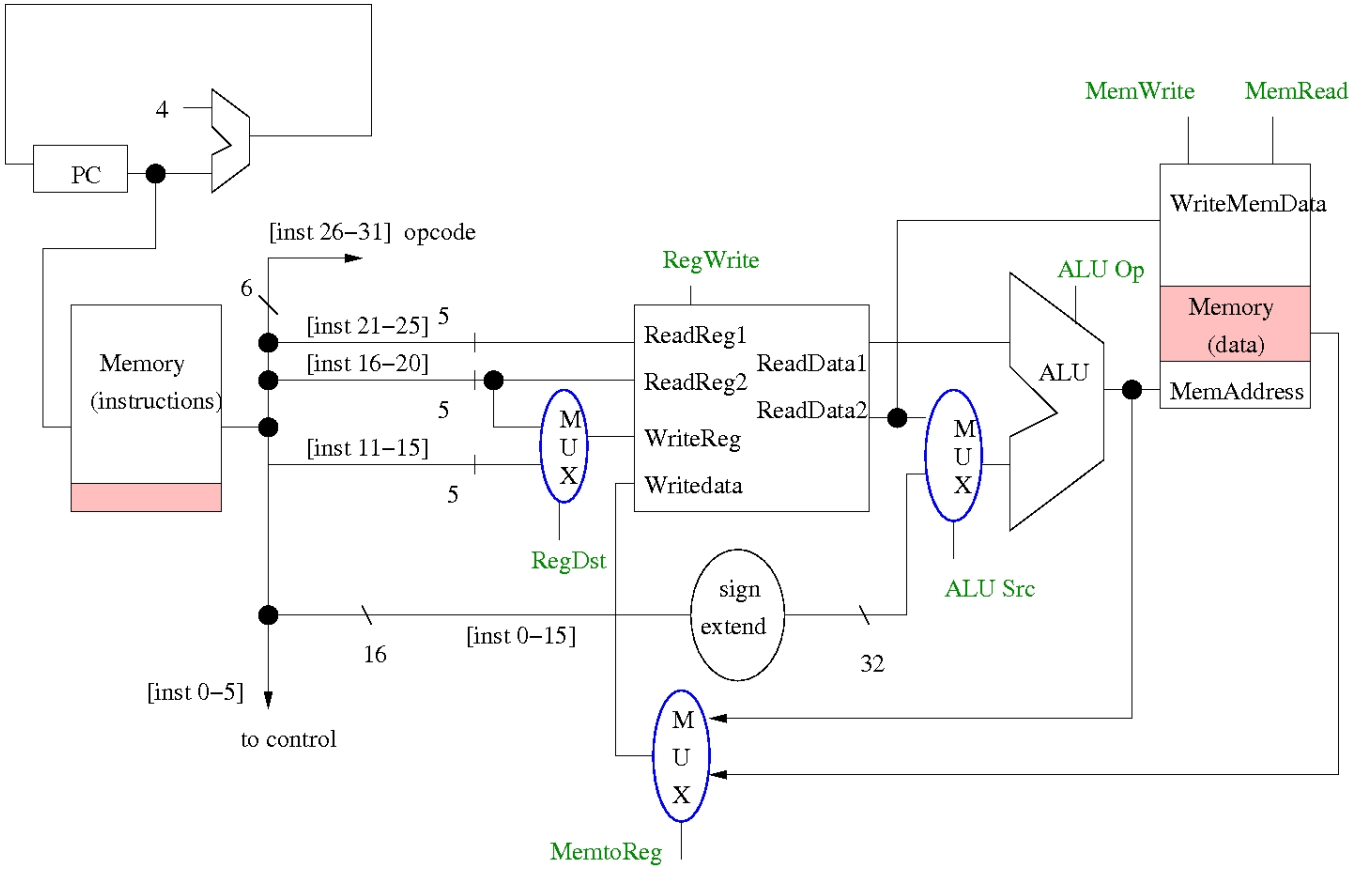
  If an R-format instruction uses three registers (`add`), then two of them are read from and one of them is written to. The *rd* register (R format) is the one that is written to.

  To summarize, *rs* is always specifies a read register ("s" standing for source), *rd* always specifies a write register ("d" standing for destination) and *rt* is a register that is either read from or written to, depending on the instruction. (I am not sure that "t" stands for anything.)

  If you look at the merged datapath, you see that the three 5-bit register number lines are routed to the register array and that a multiplexor is used to decide which line goes to the WriteReg input. The selector signal for this multiplexor is called RegDst (register destination); the value is 1 means *rd* is selected (e.g. `add`) and the value is 0 means *rt* is selected (e.g. `lw`).

- ALUSrc: One input of the ALU always comes from a register. So one of the ReadData registers (namely ReadData1) is hardwired to one of the ALU inputs. The other ALU input comes from either a register (in the case of an `add`) or from the signed-extended 16 bit immediate field (in the case of a `lw` or `sw`). Thus a multiplexor is needed to choose from among these two possible inputs to the second ALU input. The selector for this multiplexor is ALUSrc.

- MemtoReg: The third multiplexor shown in the figure selects where the register write value comes from. It might come from the ALU (R-format) or it might come from Memory `lw`.



Other controls are shown in the merged data path. RegWrite specifies whether or not we are writing to a register. MemWrite specifes whether or not we are writing to Memory. MemRead specifies whether or not we are reading from Memory. Note that the RegWrite signal was called "WriteEnable" in lecture 6 page 5. The MemWrite signal was shown in lecture 6 p. 6,7.

	input	output (of control, i.e. control variables)					
	opcode	RegWrite	MemWrite	MemRead	MemToReg	RegDst	ALUsrc
any R-format	000000	1	0	0	0	1	1
lw	100011	1	0	1	1	0	0
sw	101011	0	1	0	X	X	0
bne	000100	0	0	0	0	X	1

The ALUOp control variable needs more explanation. This control variable specifies which operation the ALU should perform. For example, `add, sub, slt, and, or` all require different operations to be performed by the ALU. Recall that the operation is determined by the instructions function field, since all R-format instructions have the same opcode, so the function field will also be needed as a input to the circuit that computes the control signals.

Recall how the ALU circuit works. It uses full adders, along with AND and OR gates so that the values `A` and `B` computes are either summed, anded, or ored. To compute subtraction, the `B` variable is inverted and a Binvert control was used for this. So, there are really two controls that need to be coded here: the Operation control and the Binvert control.

input			output (ALUOp)	
inst	opcode	funct	Binvert	Operation
`add`	000000	100000	0	10
`sub`	000000	100010	1	10
`and`	000000	100100	0	00
`or`	000000	100101	0	01
`slt`	000000	101010	1	11
:	:	:	:	:
`lw`	100011	xxxxxx	0	10
`sw`	101011	xxxxxx	0	10
`bne`	000100	xxxxxx	1	10
:	:	:	:	:

As you recall from Lecture 4, the ALU circuit has two control inputs. There is a Binvert which is used for subtraction and that takes the twos complement of the B variable, and there is the operation (either AND, OR, or +). For the latter, there are three operations to be coded, so we need two bits. *Note that this 2-bit Operation control is different from from the 6-bit "opcode" field which is part of the 32 bit instruction.*

Let's suppose that the Operation control for the ALU uses 10 for adding the two inputs. The `add` and `sub` instructions also use the 10. (The difference between `add` and `sub` comes from the Binvert control, not the Operation control.) The `lw` and `sw` instructions also perform an addition and so they have the same Operation control. However, the `and` and `or` instructions use different op controls: `and` has Operation control is 00, and `or` has operation control 01. These 2 bit values are not meaningful: they just as selector signals within the circuit.

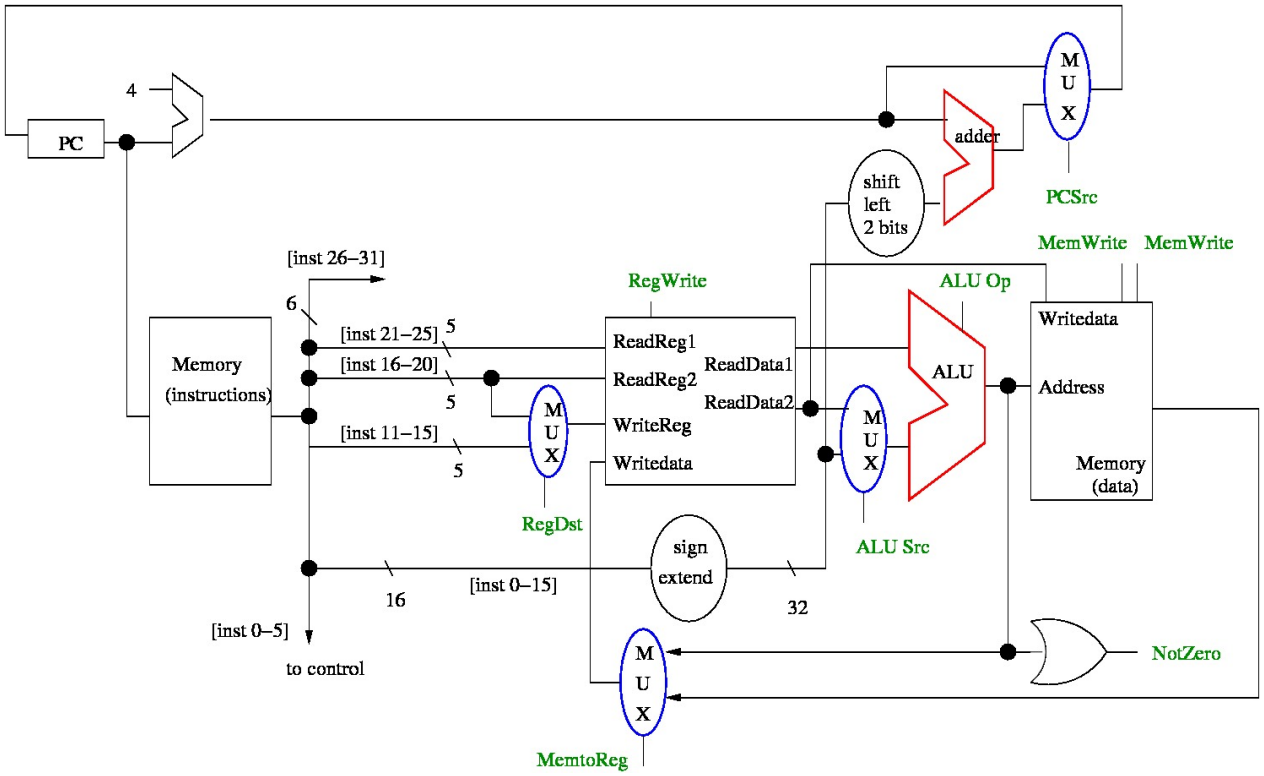
There is one more op value that could be used: 11. In fact, this is used for `slt`, "set if less than". This instruction performs a subtraction (A - B), and if the result of the subtract is negative, then the result of the instruction is defined to be 1, and otherwise it is 0. An extra circuit that is required for this. (See Exercises 5.)

[ASIDE: In fact, in MIPS, the ALU control circuit that implements the above truth table uses two combinational circuits: one that handles the case of opcode 000000 and uses the funct field input, and one that handles the case where the opcode is different from 000000. I am omitting the details this year. If you look it up e.g. in Patterson and Hennessey's text, you'll find that the controls are labelled differently from what I am giving here. Conceptually though, there is no difference.]

Merging the datapaths for (`bne`)

Let's consider how we can merge in the `bne` datapath that we saw last class. We need a multiplexor to select between PC+4 and PC+4+offset. Call this selector signal PCSrc. In general, the value of PCSrc depends on the instruction (`bne` vs. `add`, etc) as well as on the value of other variables such as NotZero (in the case of `bne`). There are other possibilities to update the PC, for example, if there is an unconditional jump instruction (see below). So we need to build a small control unit that determines PCSrc. This control unit would need the opcode [inst 26-31] and other inputs (for the case of conditional branches).

Notice that we use a separate adder circuit to compute the branch address for the `bne` instruction, rather than the ALU. The reason that we need an adder here is that the ALU is already being used in this instruction, namely to check whether the two arguments have the same value.

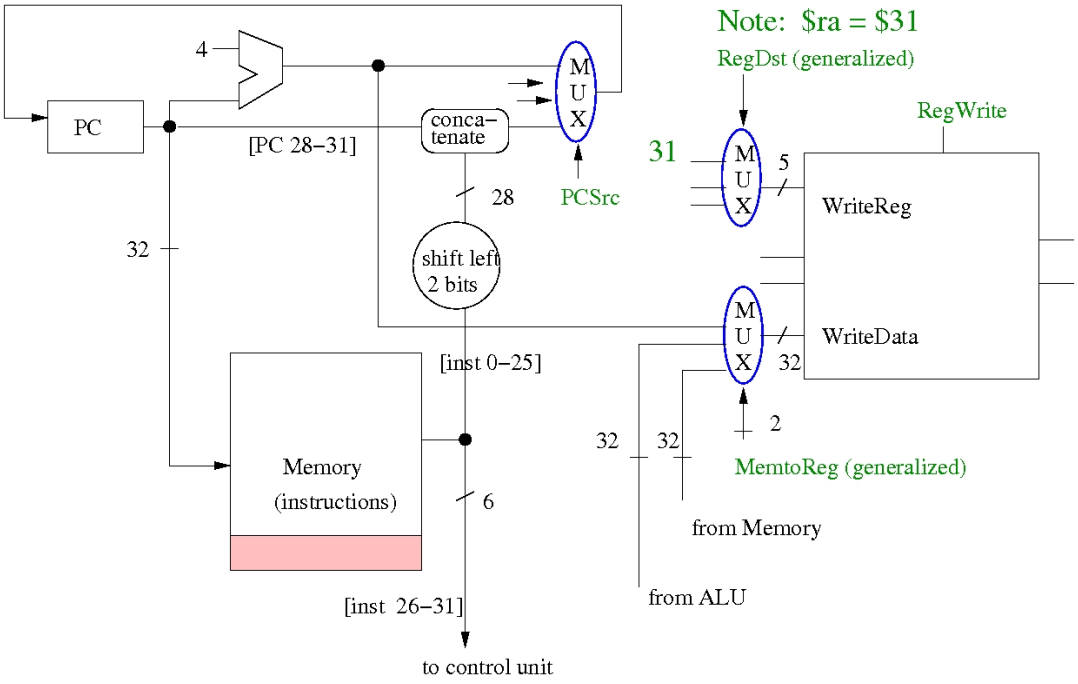


Datapath for jal (jump and link)

Last lecture we looked at the jump instruction `j`. Let's look at a related instruction: jump and link `jal`. This instruction also has J format:

field	op	address
numbits	6	26

Like the jump instruction `j`, the jump and link instruction jumps to an address that is determined by a concatenation of the upper 4 bits of the PC, the lower 26 bits of the instruction, and the bits 00 which go into the two lowest order address bits (because instructions are word aligned).



The difference between `jal` and `j` is that `jal` also writes PC+4 into register $ra, which is $31. I have written this 5 bit values as 31 in the figure. I have also added a line which shows how PC+4 gets written into register $31.

Notice in the figure that the two controls WriteDataSrc and MemtoReg are now "generalized" All I mean by that is that previously these controls were 1 bit because they had to choose between two possibilities, but now there are three possibilities and so they need to be two bits. (I could have changed the names of these controls to something more general, but we won't need them further so there is no need for that.) bother.)

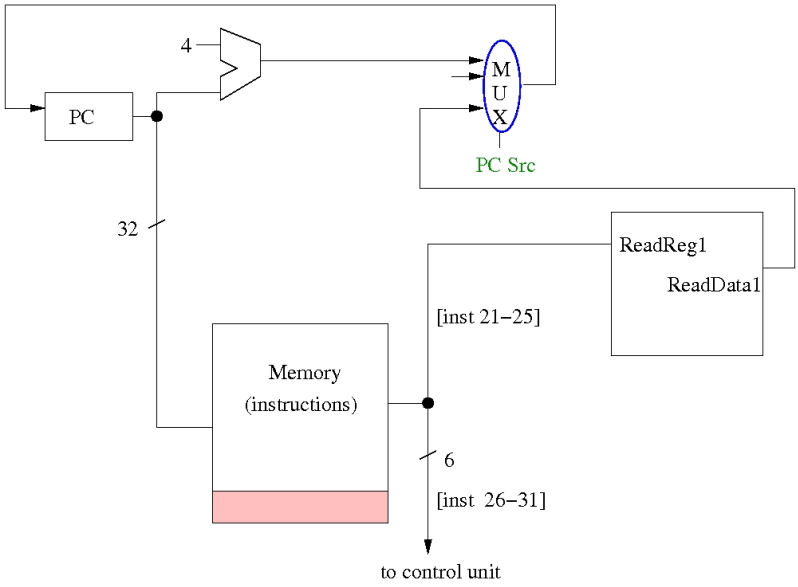
Datapath for jr (jump register, R format)

You may be surprised to discover that the `jr` instruction has R format. From our arguments at the beginning of this lecture, you can see why. The instruction `jr $r*` (typically `$ra`) is encoded by inputting the register number (e.g. `$ra` = 31) into ReadReg1. The address to jump to is read from this register. Usually this is the return address of a function. This return address is read out from the register and selected as the value to be written into the PC, at the end of the clock cycle.

Since we are reading from a register, the register number is in the *rs* field. The *rt*, *rd*, and shamt fields contain junk for this instruction. The funct field does not contain junk though. Since this is an R-format instruction, the opcode is the same as for the `add` and other instructions, and so `jr` is only distinguished from these other R-format instructions by the funct field).

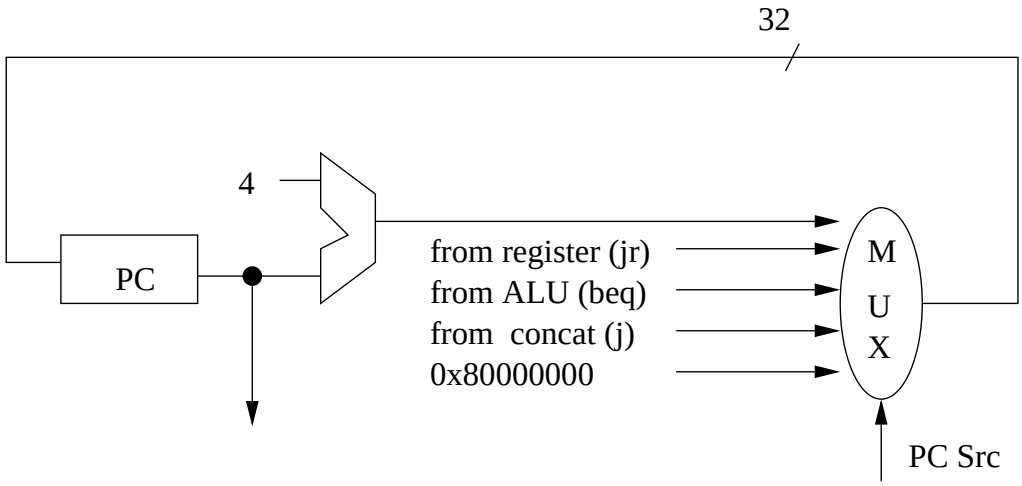
Exceptions - jumping to the kernel [ADDED March 2]

We have concentrated on single MIPS programs running in the *user* part of MIPS memory, that is, below address 0x80000000. We have also discussed how errors can occur when the programs



is running, such as integer division by zero, or a bad address is used for an instruction or for a load from or store to Memory. When such errors occur, the computer stops executing the program and jumps to the kernel. The program counter PC must be updated accordingly. In particular, the program jumps to the first instruction address in the kernel, namely 0x80000000. This is the first instruction of a kernel program called the *exception handler*. Here we consider briefly the mechanism of how this is done.

The PC is updated by selecting from one of many inputs to a multiplexor, namely a multiplexor whose control is is PCSrc. Now we add another possible address that can go there and this is shown in the figure below. (0x80000000).



There can be many types of exceptions. For example, a syscall call is another type. And there are other types of exceptions called *interrupts* which we will discuss later. For now, I just want to emphasize that the PC can take many different values on the next clock cycle and that the

value the PC takes is determined by a selector PCSrc. This control variable can depend on various combinations of other control variables which describe the present state of the program.

Summary of Single Cycle implementation

What we have described is a *single cycle implementation*. The entire instruction (including the instruction fetch, the update of the PC, the ALU operation, the read or write from memory) is done in one clock cycle. This is possible because the datapaths for these instructions never involve involve more than one write into register or memory. Careful here: you can write into the PC and into a register within one instruction, but you are using two different paths within that instruction. The combinational circuits within these two paths run in parallel.

As mentioned earlier, in a single cycle implementation, the interval between clock pulses must be long enough to make sure that all the signals stablize by the end of the clock cycle (e.g. the 32 bit ALU requires all the Carryin signals to ripple through...) If a word is read from Memory, the value must be available on the output wires from the Memory. Since the clock pulses occur at a regular rate, we must design for the worst case.

What is the worst case? For the instructions we discussed above, the worst case is the `lw` instruction. We need to read from registers, to do an addition, read from Memory, and then to write the data item into a register. For the `add` instruction, you don't need to use the Memory at all, so this is much faster.

Designing for the worst case is a big problem, in terms of efficiency. And in fact, real MIPS computers do not use a single cycle implementation. Rather they use multiple cycles. Next lecture, I will introduce some ideas of how a multiple cycle implementation could work.

The idea of using multiple clock cycles should not be entirely new to you. We have already discussed how integer multiplication and division require a sequences of shifts and either additions or subtractions, and that floating point operations require multiple shifts, compares (of exponents), adds and subtracts, etc. Next lecture, I will not discuss integer multiplication and division, and floating point operations, however. Instead, I will discuss CPU instructions such as we have been discussing this lecture and the last one. I will discuss how multiple clock cycles are used for these instructions.