Computer Architecture and Programming

ICS312

Machine-level and Systems Programming

Henri Casanova (henric@hawaii.edu)

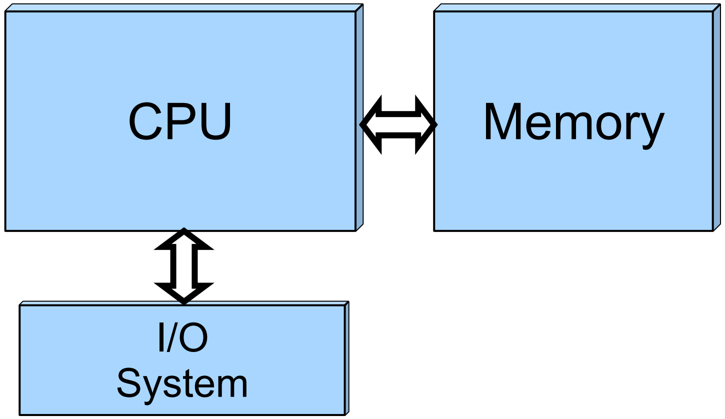
“Computer Architecture”?

- The field of Computer Architecture is about the fundamental structure of computer systems
  - What are the components?
  - How do they interact with each other?
  - How fast does the whole system operate?
  - How much power does it consume?
  - How much does it cost to mass-produce?
  - How to achieve desired speed/power/cost trade-offs?
- The conceptual model for computer architecture, that hasn’t fundamentally changed since 1945: the Von-Neumann architecture

Von-Neumann

- In 1944, John von Neumann joined ENIAC
- He wrote a memo about computer architecture, formalizing ENIAC ideas
  - Eckert and Mauchly have pretty much been forgotten (they were in the trenches)
- These ideas became the Von Neumann architecture model
  - A processor that performs operations and controls all that happens
  - A memory that contains **code and data**
  - I/O of some kind

Von-Neumann Model



- Amazingly, it's still possible to think of the computer this way at a conceptual level (model from ~80 years ago!!!)
- Even though a computer today doesn't look quite like this
  - But these are just "details" really
  - You can just think of the above picture, with tons of (performance enhancing) bells and whistles

Data Stored in Memory

- All "information" in the computer is in binary form
  - Since Claude Shannon's M.S. thesis in the 1930's
  - 0: zero voltage, 1: positive voltage (e.g., 5V)
  - bit: the smallest unit of information (0 or 1)
- The basic unit of memory is a byte
  - 1 Byte = 8 bits, e.g., "0101 1101"
  - 1 KiB = 2^{10} byte = 1,024 bytes
  - 1 MiB = 2^{10} KiB = 2^{20} bytes (~ 1 Million)
  - 1 GiB = 2^{10} MiB = 2^{30} bytes (~ 1 Billion)
  - 1 TiB = 2^{10} GiB = 2^{40} bytes (~ 1 Trillion)
  - 1 PiB = 2^{10} TiB = 2^{50} bytes (~ 1000 Trillion)
  - 1 EiB = 2^{10} PiB = 2^{60} bytes (~ 1 Million Trillion)
  - ...
- Note the "i" in the notations above: means "power of 2"
  - Unlike in a kilometer (km), where k means 1,000 (not 1,024)

Data Stored in Memory

- Each byte in memory is labeled by a unique address
- An address is a number that identifies the memory location of each byte in memory
  - e.g., the byte at address 3 is 00010010
  - e.g., the byte at address 241 is 10110101
- Typically, we write addresses in binary as well
  - e.g., the byte at address 00000011 is 00010010
  - e.g., the byte at address 11110001 is 10110101
- All addresses in RAM have the same number of bits
  - e.g., 8-bit addresses
- The processor has instructions that say “Read the byte at address X and give me its value” and “Write some value into the byte at address X”

Example Byte-Addressable RAM with 16-bit addresses

address	content
0000 0000 0000 0000	0110 1110
0000 0000 0000 0001	1111 0100
0000 0000 0000 0010	0000 0000
0000 0000 0000 0011	0000 0000
0000 0000 0000 0100	0101 1110
0000 0000 0000 0101	1010 1101
0000 0000 0000 0110	0000 0001
0000 0000 0000 0111	0100 0000
0000 0000 0000 1000	1111 0101
...	...

Example Byte-Addressable RAM with 16-bit addresses

address	content
0000 0000 0000 0000	0110 1110
0000 0000 0000 0001	1111 0100
0000 0000 0000 0010	**0000 0000**
0000 0000 0000 0011	0000 0000
0000 0000 0000 0100	0101 1110
000	
000	
000	
0000 0000 0000 1000	1111 0101
...	...

At address 0000 0000 0000 0010 the content is 0000 0000

Example Byte-Addressable RAM with 16-bit addresses

address	content
0000 0000 0000 0000	0110 1110
0000 0000 0000 0001	1111 0100
0000 0000 0000 0010	0000 0000
0000 0000 0000 0011	0000 0000
0000 0000 0000 0100	0101 1110
000	
000	
000	
0000 0000 0000 1000	1111 0101
...	...

At address 0000 0000 0000 0100 the content is 0101 1110

Address Space

- With d-bit long addresses we can address 2^d different "things"
- Example:
  - 2-bit addresses
    - 00, 01, 10, 11
    - 4 "things"
  - 3-bit addresses
    - 000, 001, 010, 011, 100, 101, 110, 111
    - 8 "things"
- In our case, these things are "bytes" because we consider byte-addressable RAM
  - One cannot address anything smaller than a byte

- How big is my RAM if my addresses are 13-bit? (poll)

Address Space

- With d-bit long addresses we can address 2^d different "things"
- Example:
  - 2-bit addresses
    - 00, 01, 10, 11
    - 4 "things"
  - 3-bit addresses
    - 000, 001, 010, 011, 100, 101, 110, 111
    - 8 "things"
- In our case, these things are "bytes" because we consider byte-addressable RAM
  - One cannot address anything smaller than a byte

- How big my RAM if my addresses are 13-bit?
  2^{13} bytes = $2^3 * 2^{10}$ = 8 KiB!

Both Code and Data in Memory

- Once a program is loaded in memory, its address space contains **both code and data**
- To the CPU, code and data are just bytes, but the programmer (hopefully) knows which bytes are data an which bytes are code
  - Conveniently hidden from you if you've never written assembly
  - But we'll have to keep code/data straight in these lecture notes

Example Address Space

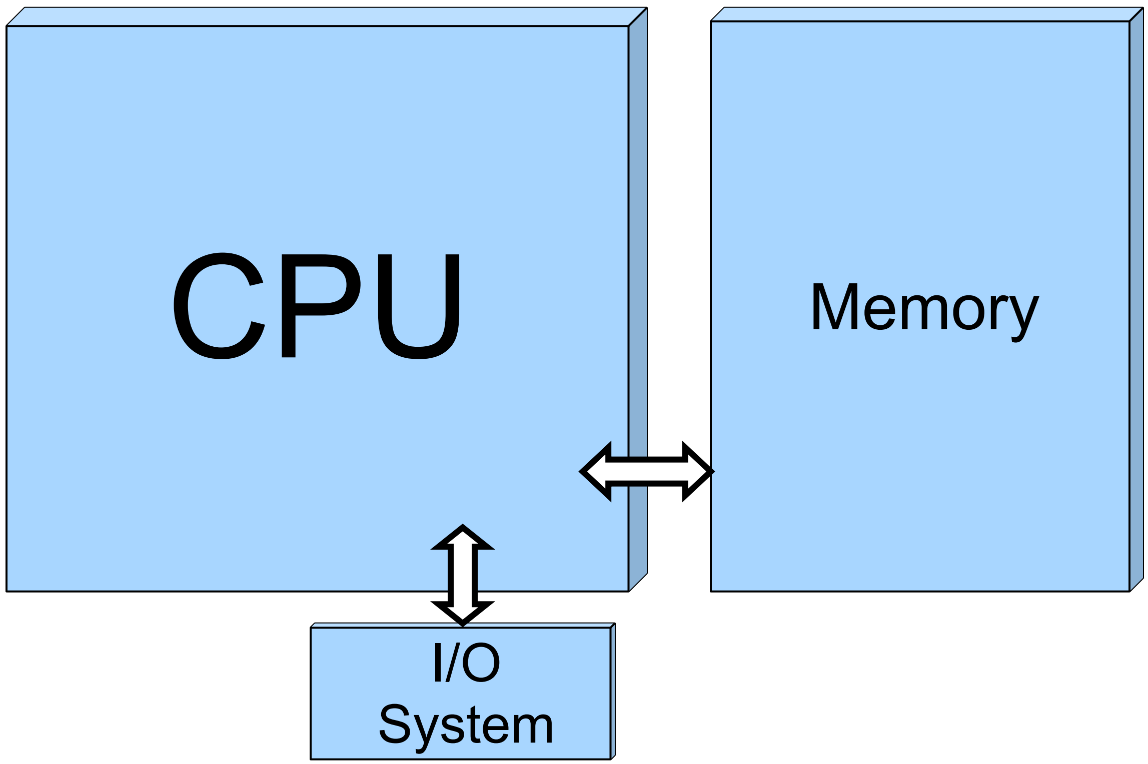
	Address	**Value**
Code	0000 1100	0110 1011
Code	0000 1101	1111 0010
Code	0000 1110	0010 0001
	…	…
Data	1000 0000	1111 0000
Data	…	…
Data	1111 0010	0101 1111

Memory

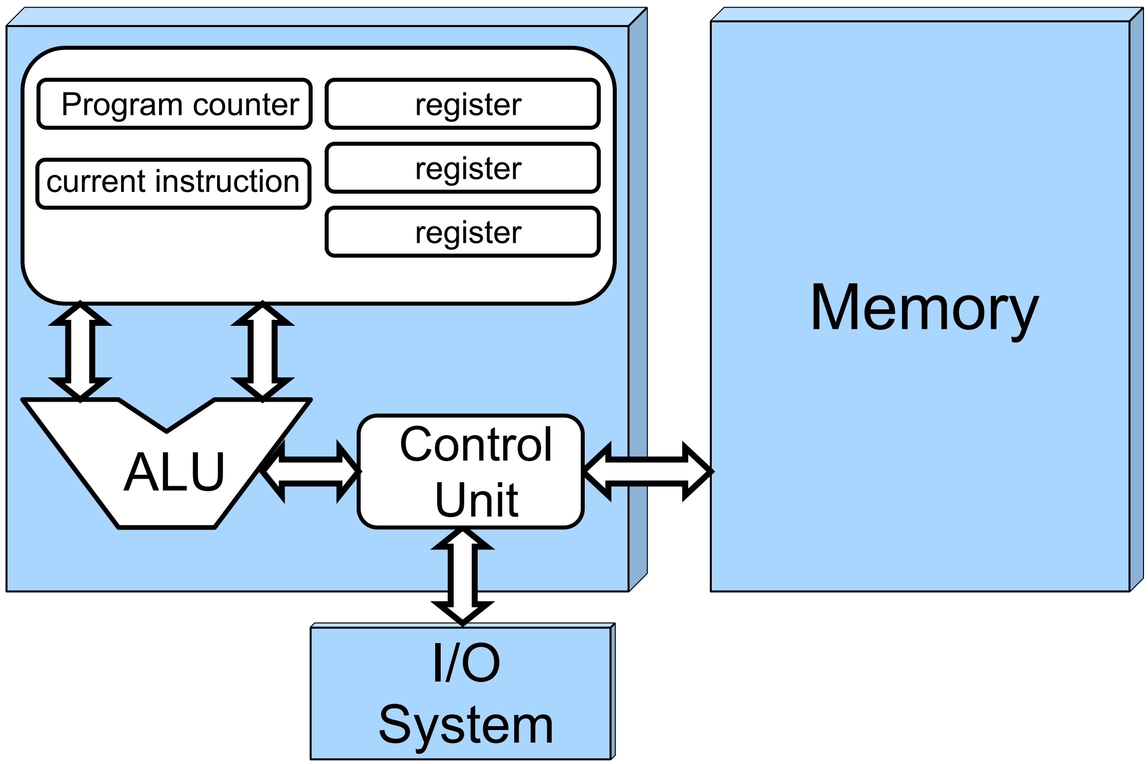
The CPU is a Memory Modifier

- So now we have a memory in which we can store/retrieve bytes at precise locations
- These bytes presumably have some useful meaning to us
  - e.g., integers, ASCII codes of characters, floating points numbers, RGB values
  - e.g., instructions that specify what to do with the data; when you buy a processor, the vendor defines the instruction set (e.g., instruction “0010 1101” means “increment some useful counter”)
- The CPU is the piece of hardware that modifies the content of memory
  - In fact, one can really think of the CPU as a device that takes us from one memory state (i.e, all the stored content) to another memory state (some new, desired stored content)

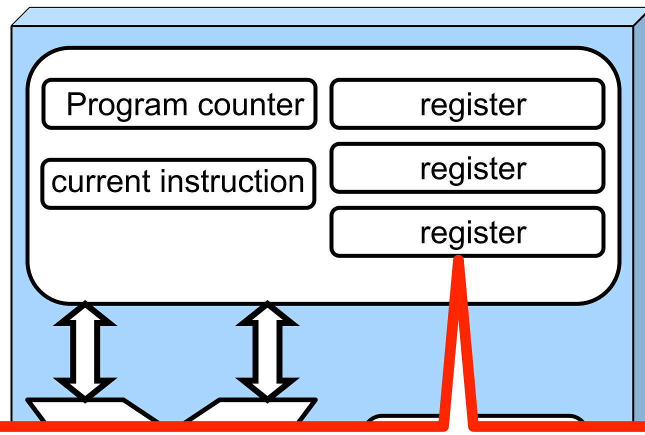
What's in the CPU?



What's in the CPU?



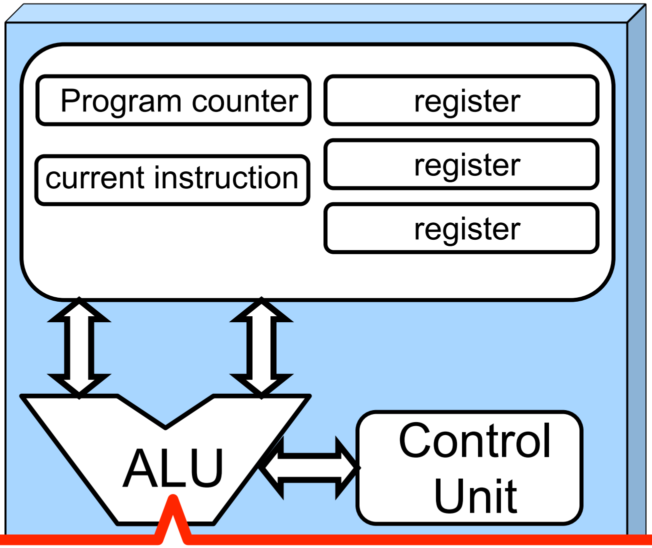
What’s in the CPU?



Registers: values that hardware instructions work with

Data can be loaded from memory into a register
Data can be stored from a register back into memory
Operands and results of computations are ALL in registers
Accessing a register is really fast
There is a limited number of registers (which will make our life a bit difficult)

What's in the CPU?

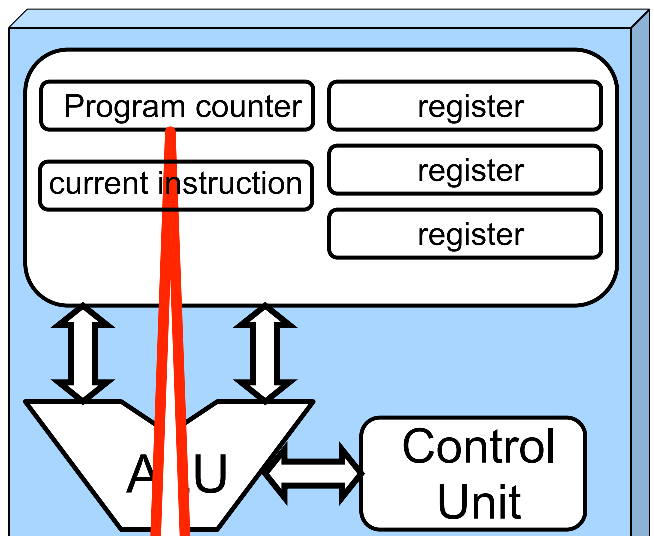


Arithmetic and Logic Unit: what you do computation with

used to compute a value based on current register values and store the result back into a register

+, *, /, -, OR, AND, XOR, etc.

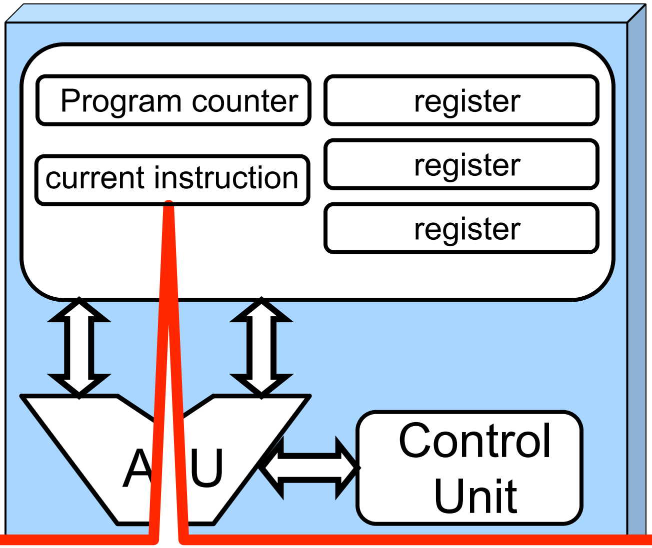
What's in the CPU?



Program Counter: Points to the next instruction

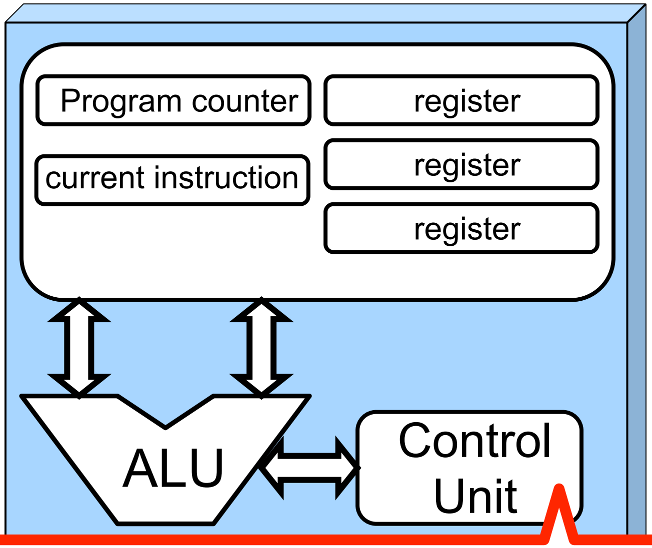
Special register that contains the address in memory of the next instruction that should be executed
(gets incremented after each instruction, or can be set to whatever value whenever there is a change of control flow)

What’s in the CPU?



Current Instruction: Holds the instruction that’s currently being executed

What's in the CPU?



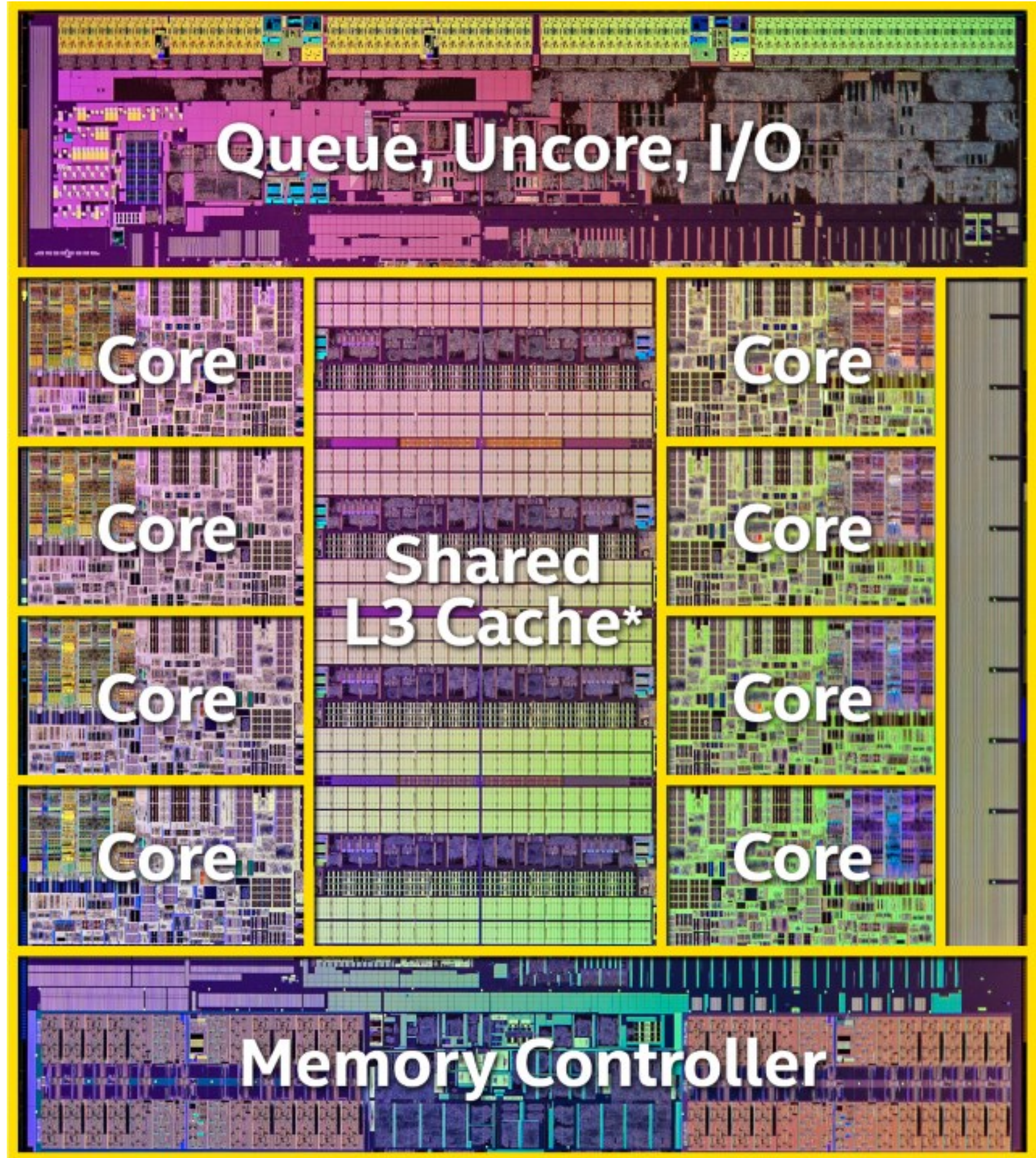
Control Unit: Decodes instructions and make them happen

Logic hardware that decodes instructions (i.e., based on their bits) and sends the appropriate (electrical) signals to hardware components in the CPU

A CPU in its "Glory": Intel Haswell

Each Core contains a "CPU" (as in in previous slides) plus more stuff (e.g., L1 and L2 Caches)

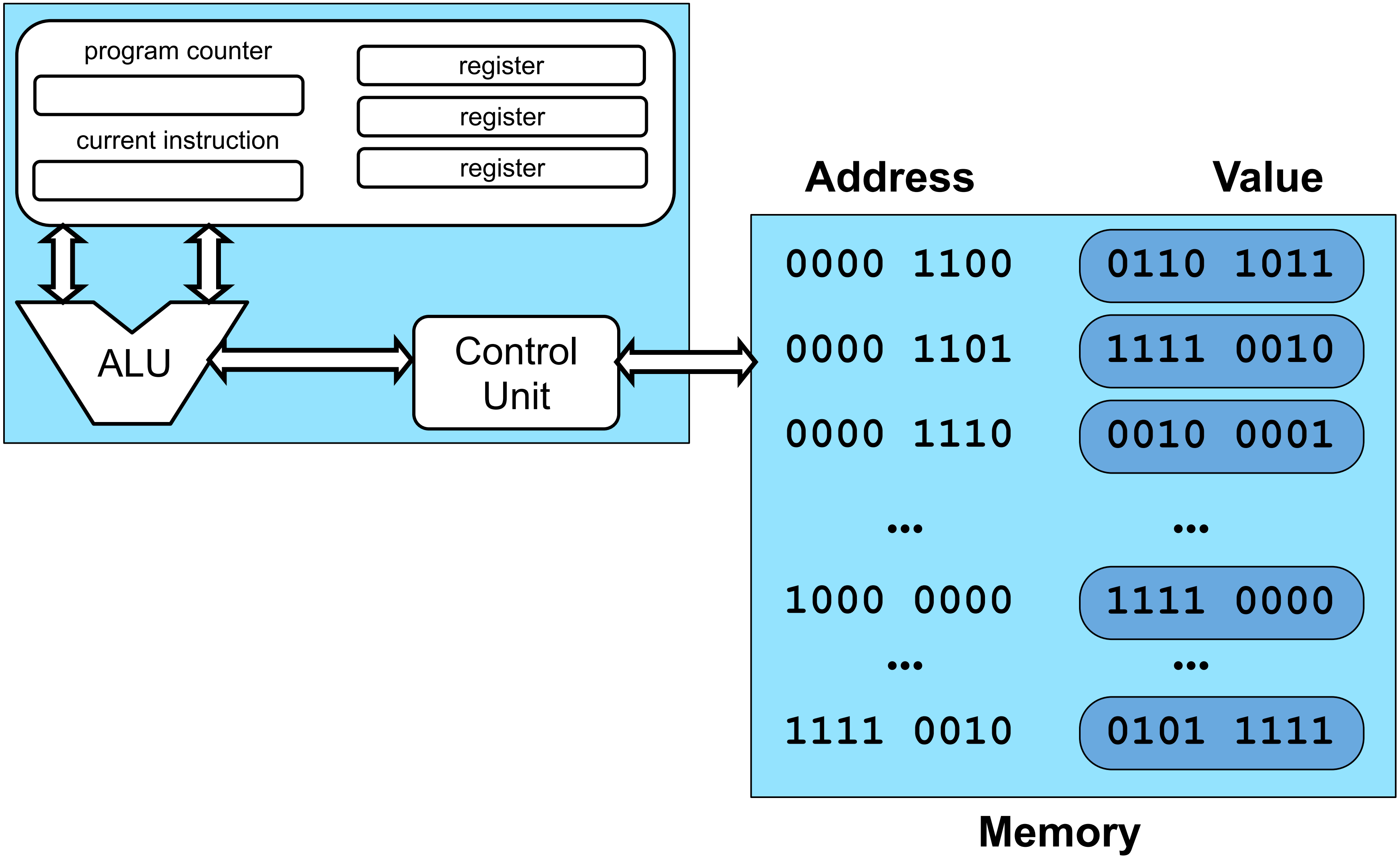
In this course we assume a single, very simple core (which we call the CPU)



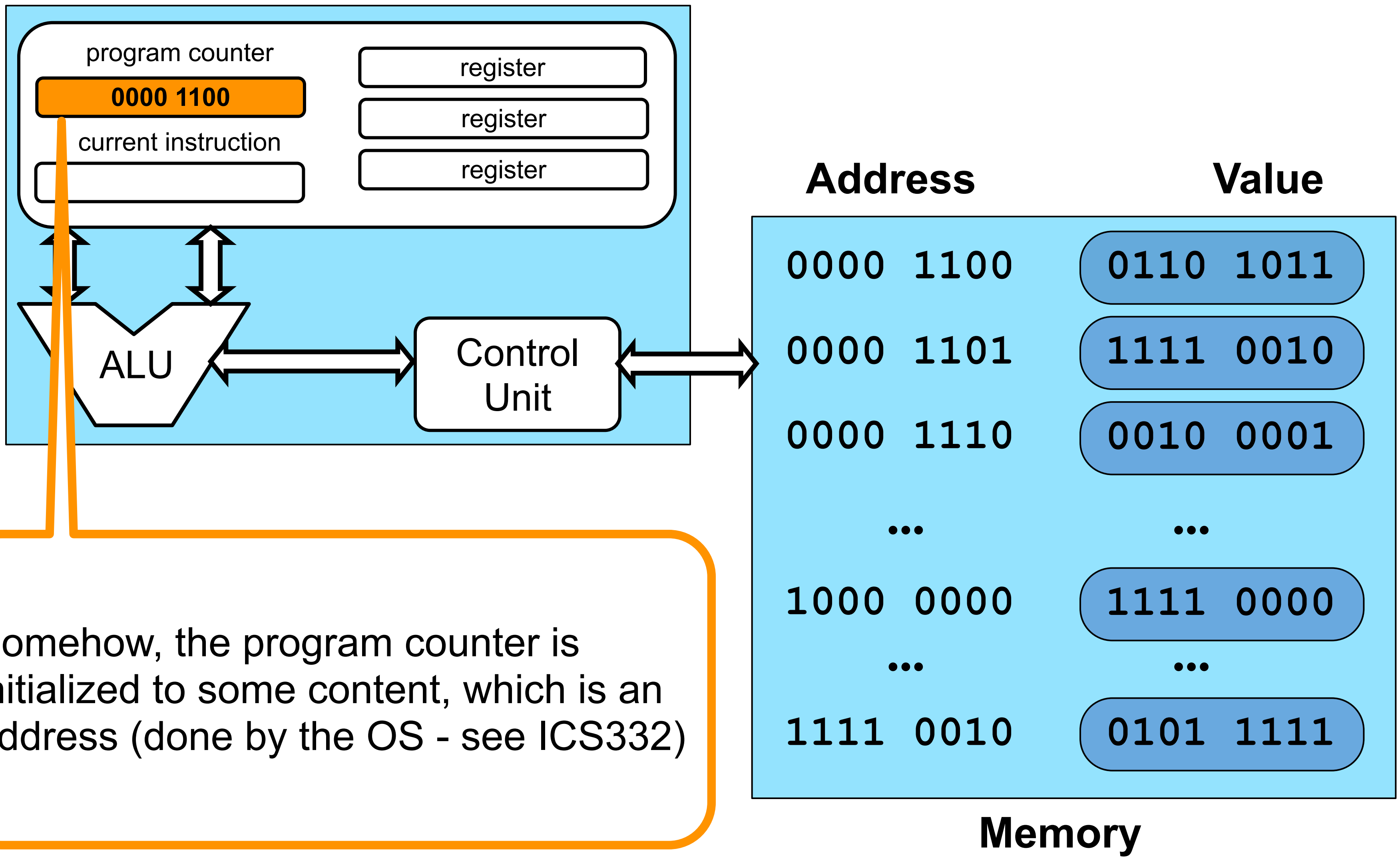
Fetch-Decode-Execute Cycle

- The Fetch-Decode-Execute cycle
  - The control unit **fetches** the next program instruction from memory
    - Using the program counter to figure out where that instruction is located in the memory
  - The control unit **decodes** the instruction and signals are sent to hardware components
    - e.g., is the instruction loading something from memory? is it adding two register values together?
  - The instruction is **executed**
    - Operands are fetched from memory and put in registers, if needed
    - The ALU executes computation, if any, and stores the computed results in the registers
    - Register values are stored back to memory, if needed
  - Repeat
- Computers today implement MANY variations on this model
- But one can still program with the above model in mind
  - But then without understanding performance issues

Fetch-Decode-Execute

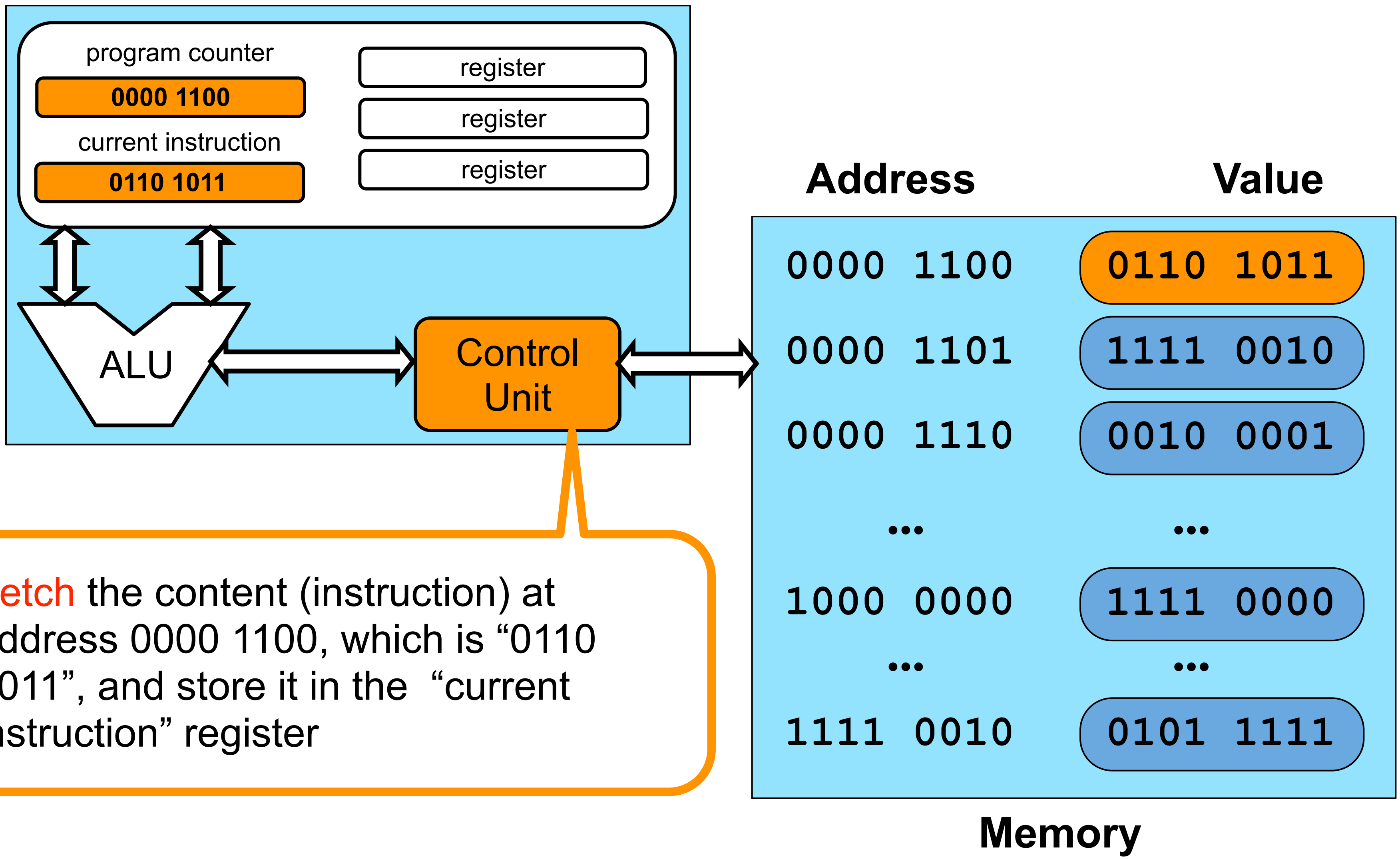


Fetch-Decode-Execute



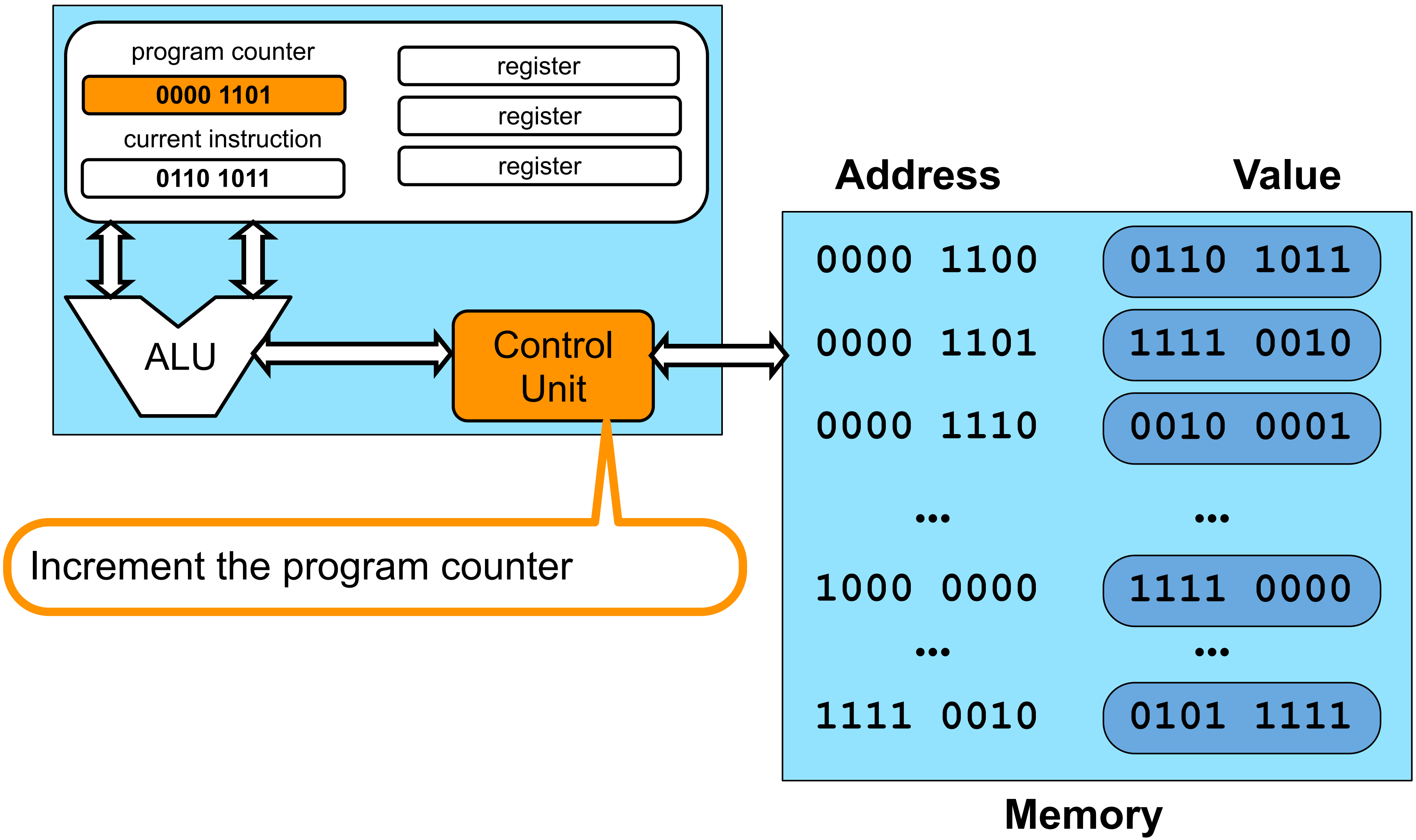
Somehow, the program counter is initialized to some content, which is an address (done by the OS - see ICS332)

Fetch-Decode-Execute

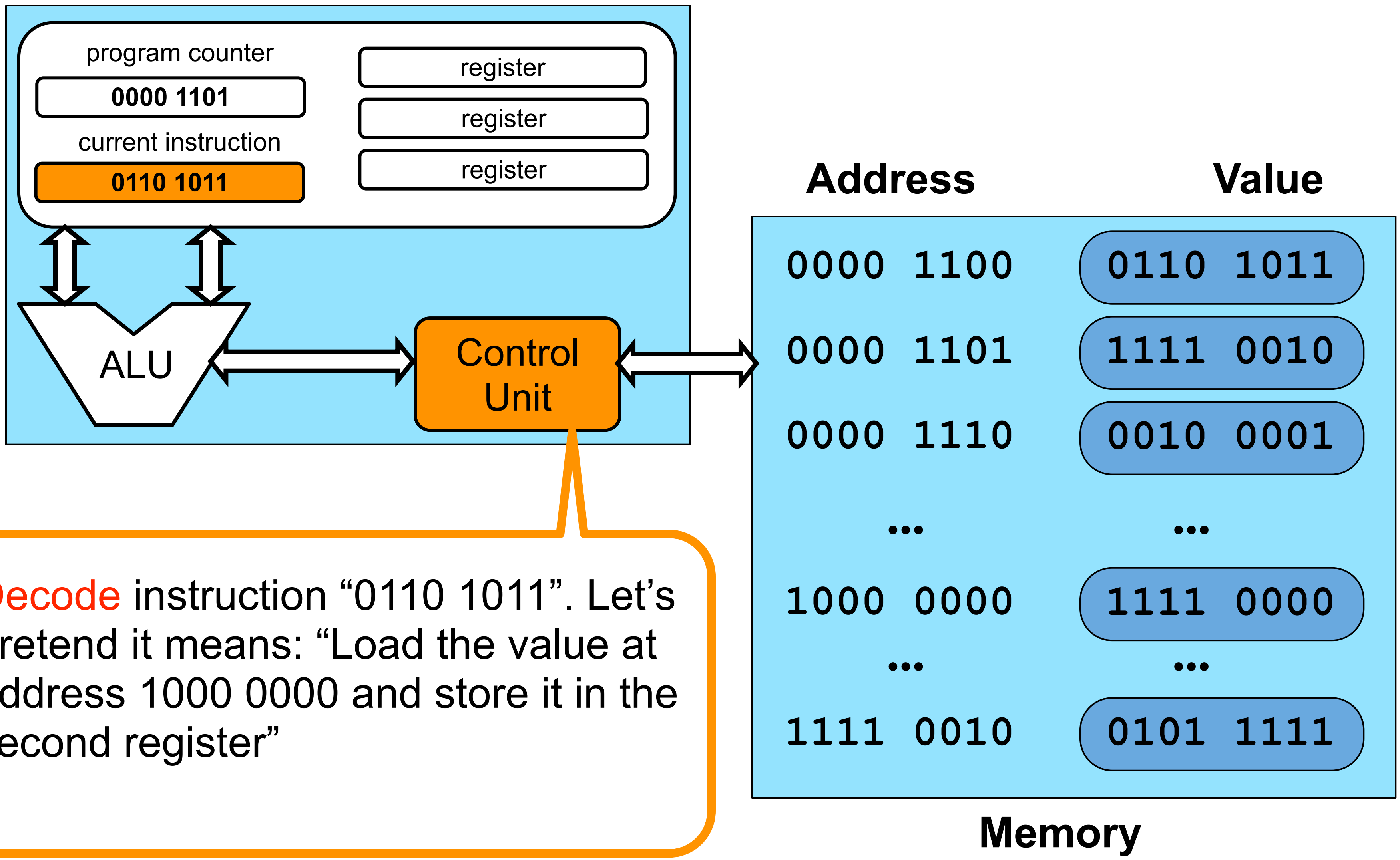


Fetch the content (instruction) at address 0000 1100, which is “0110 1011”, and store it in the “current instruction” register

Fetch-Decode-Execute

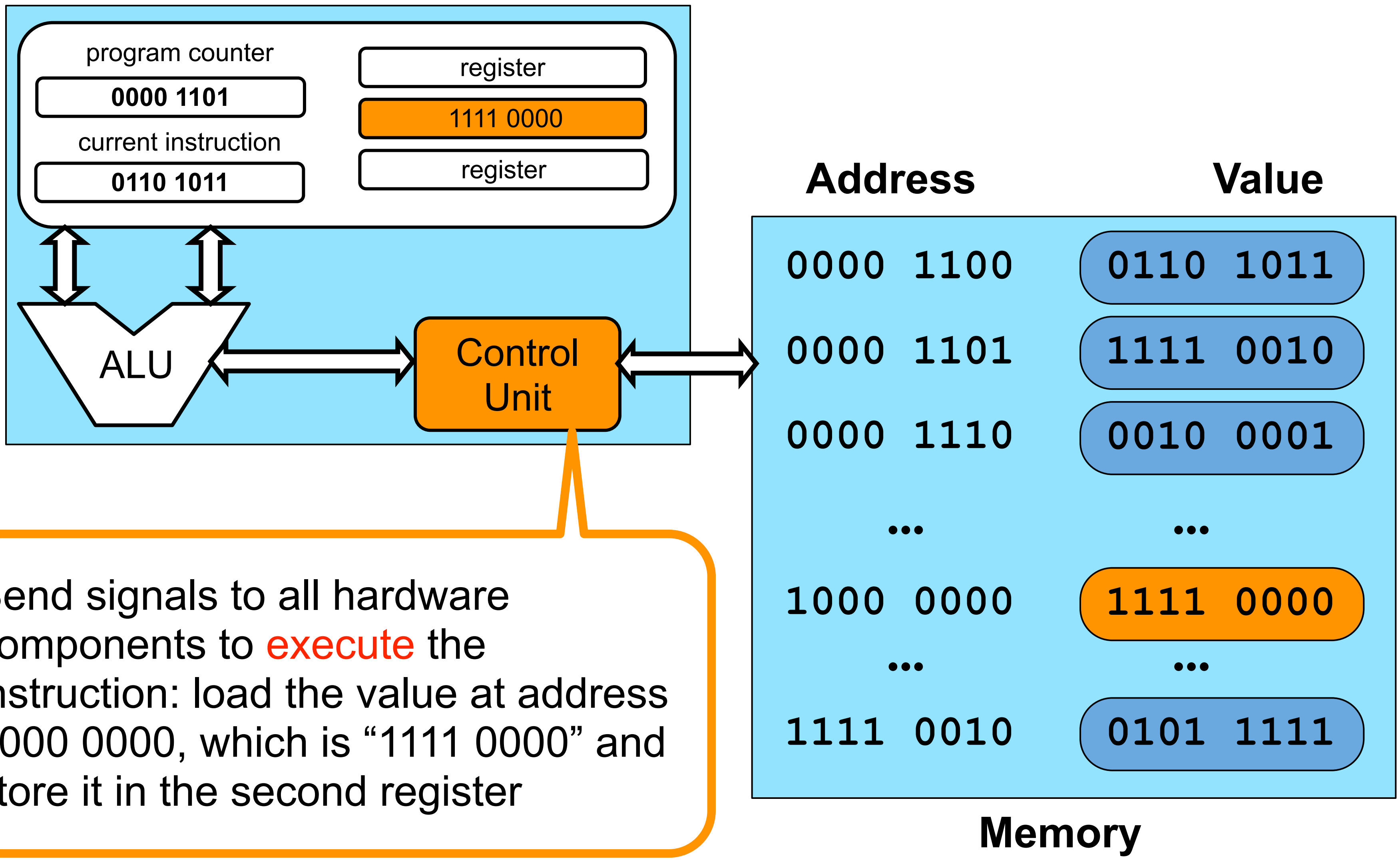


Fetch-Decode-Execute



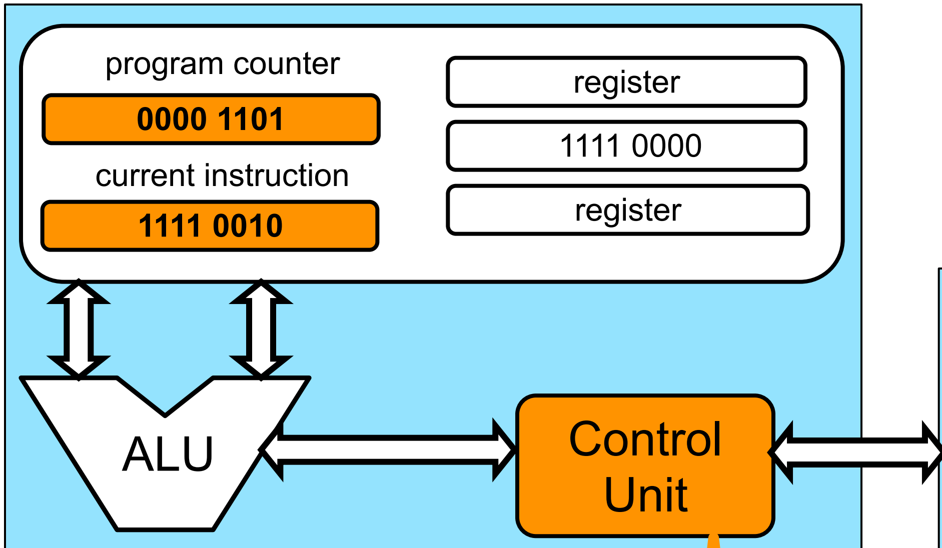
Decode instruction "0110 1011". Let's pretend it means: "Load the value at address 1000 0000 and store it in the second register"

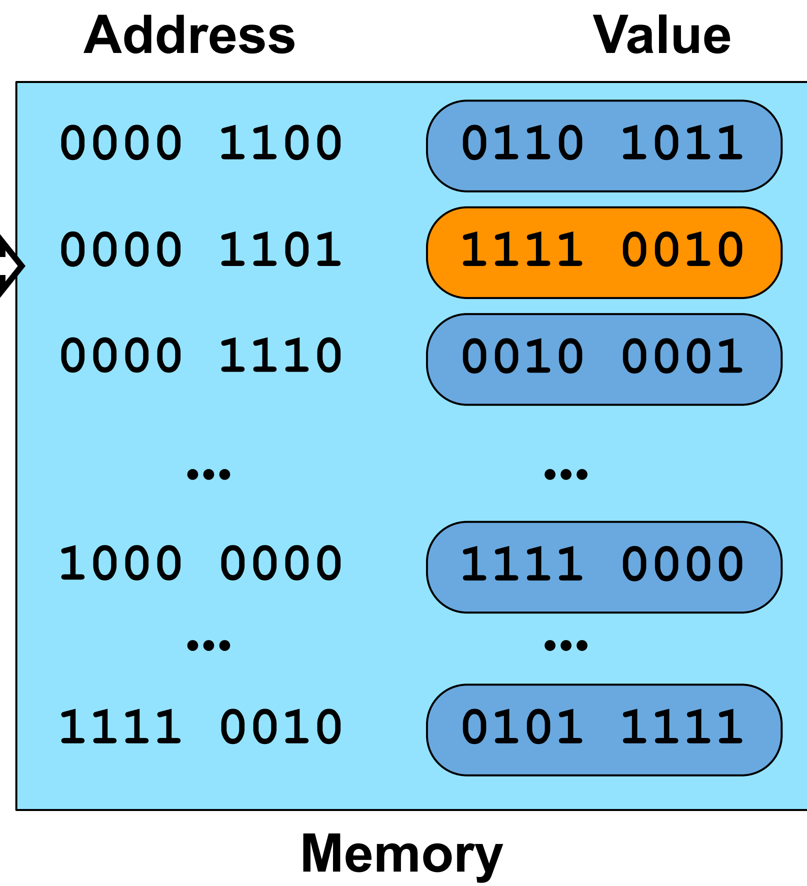
Fetch-Decode-Execute



Send signals to all hardware components to execute the instruction: load the value at address 1000 0000, which is "1111 0000" and store it in the second register

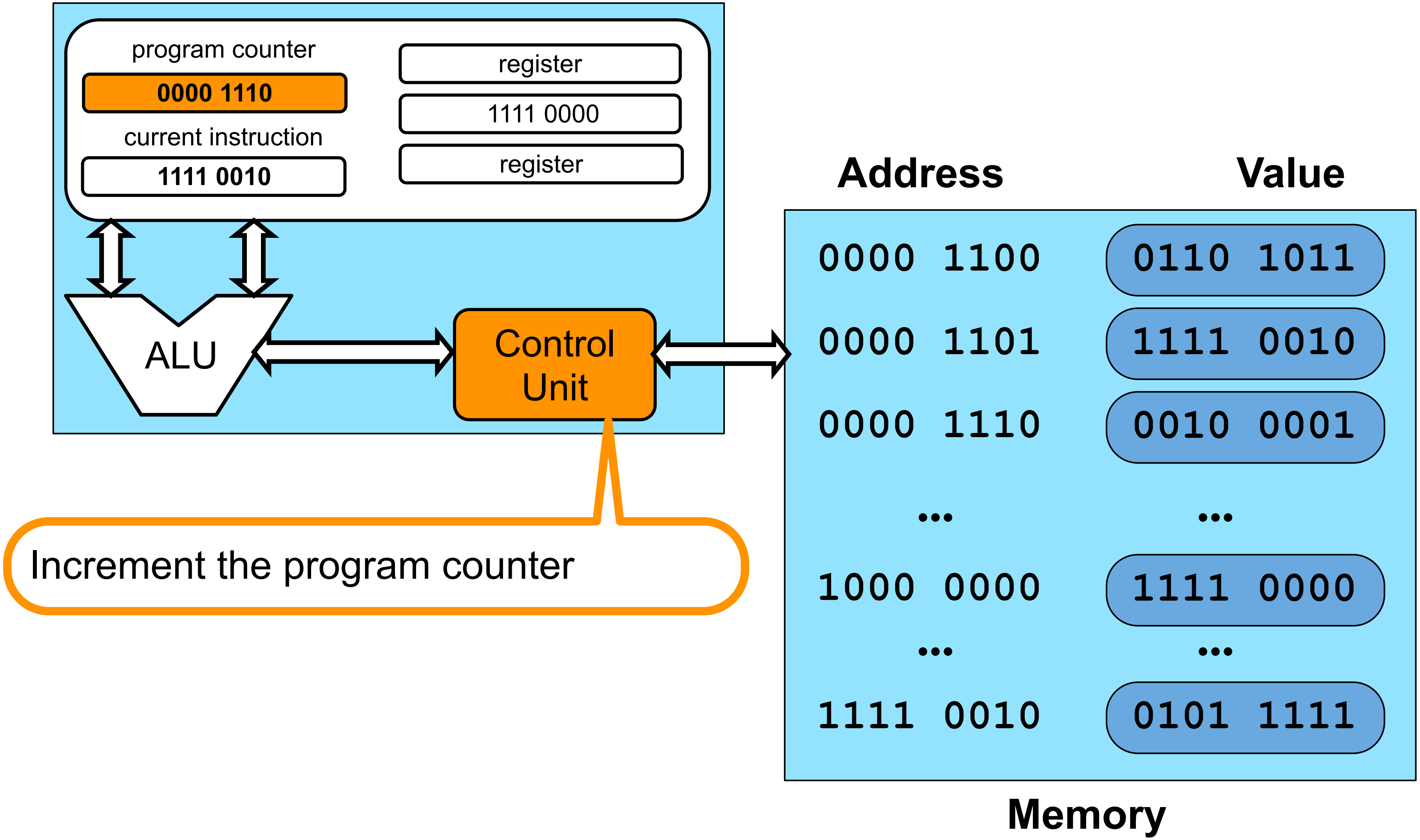
Fetch-Decode-Execute



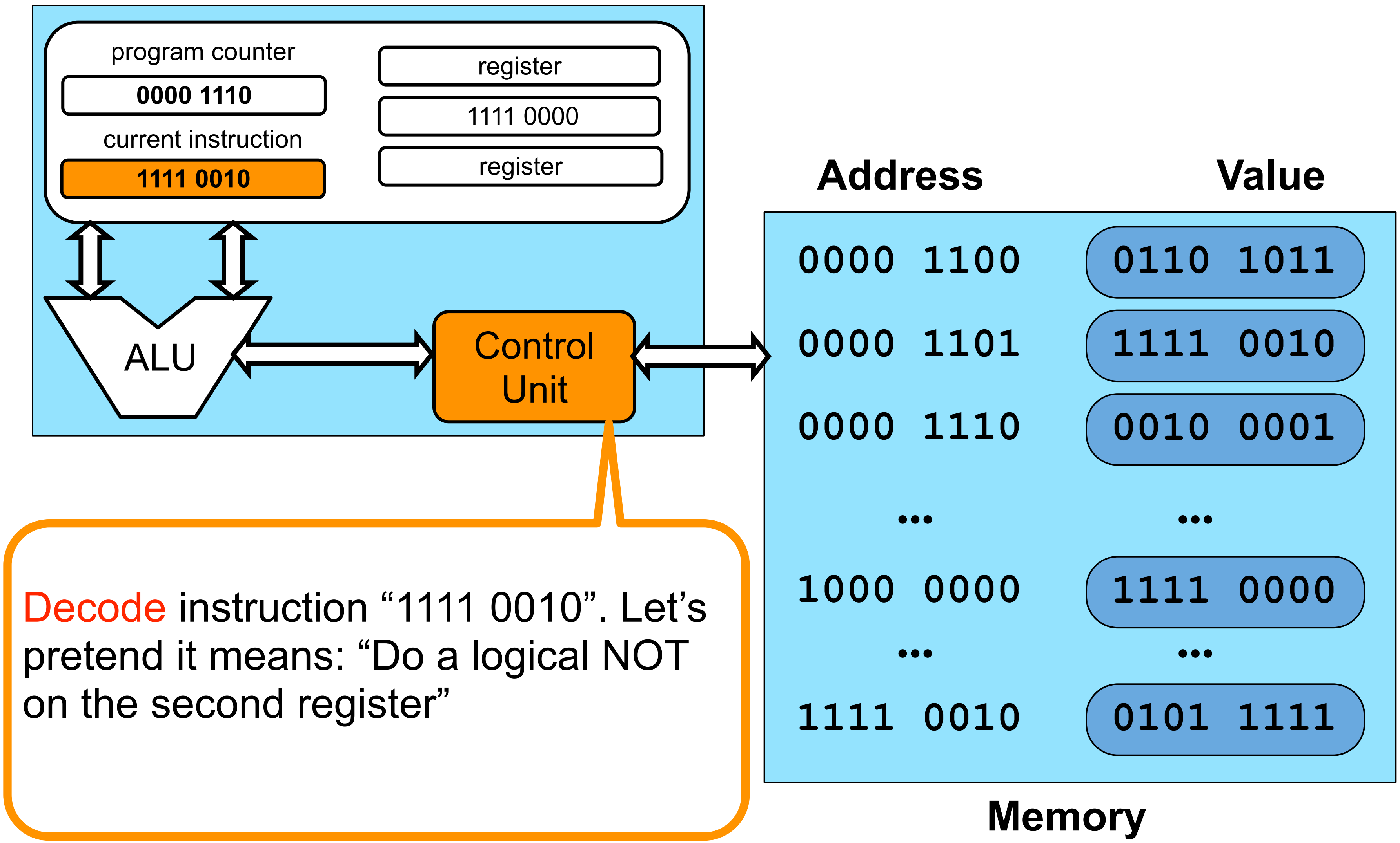


Fetch the content (instruction) at address 0000 1101, which is "1111 0010", and store it in the "current instruction" register

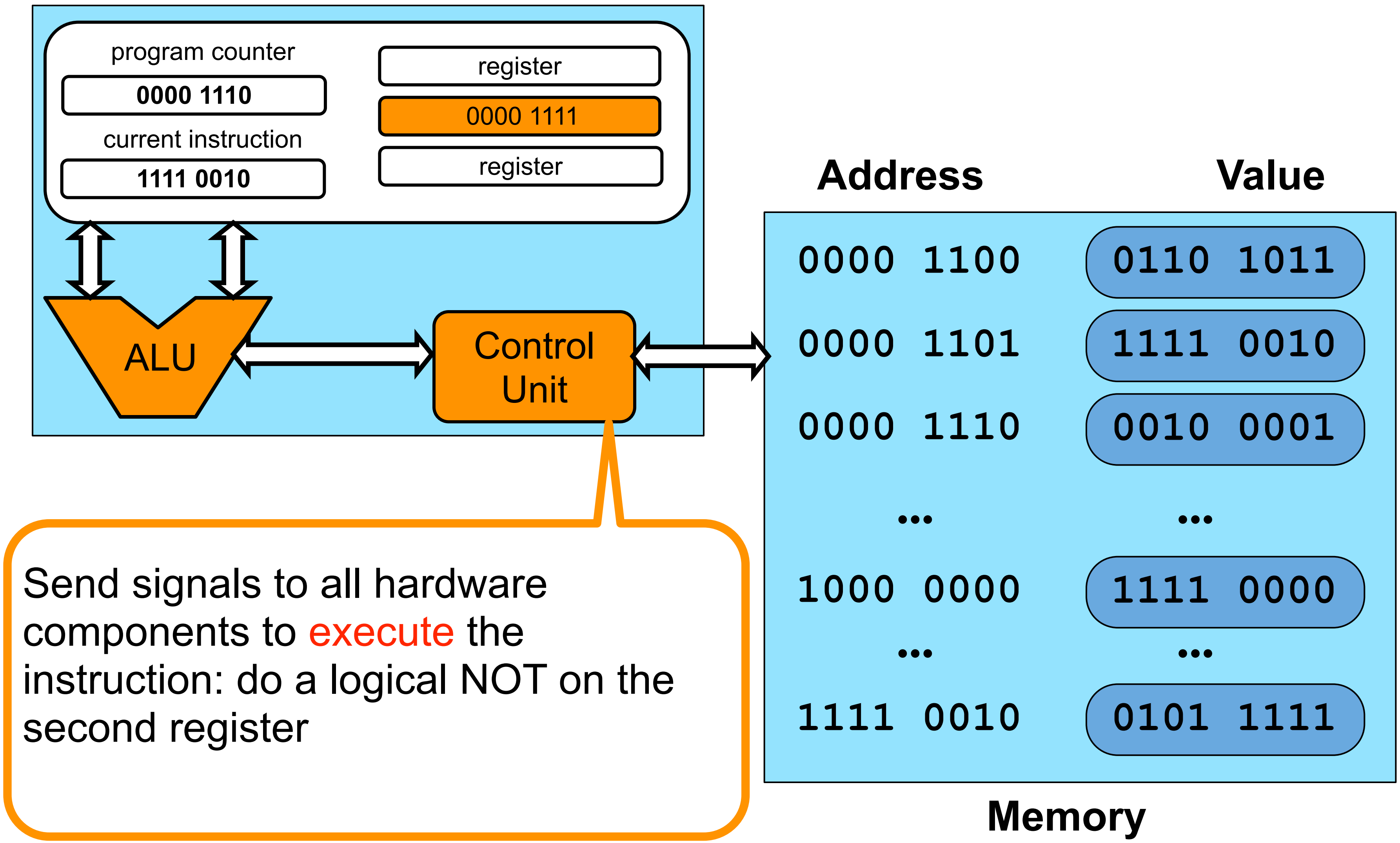
Fetch-Decode-Execute



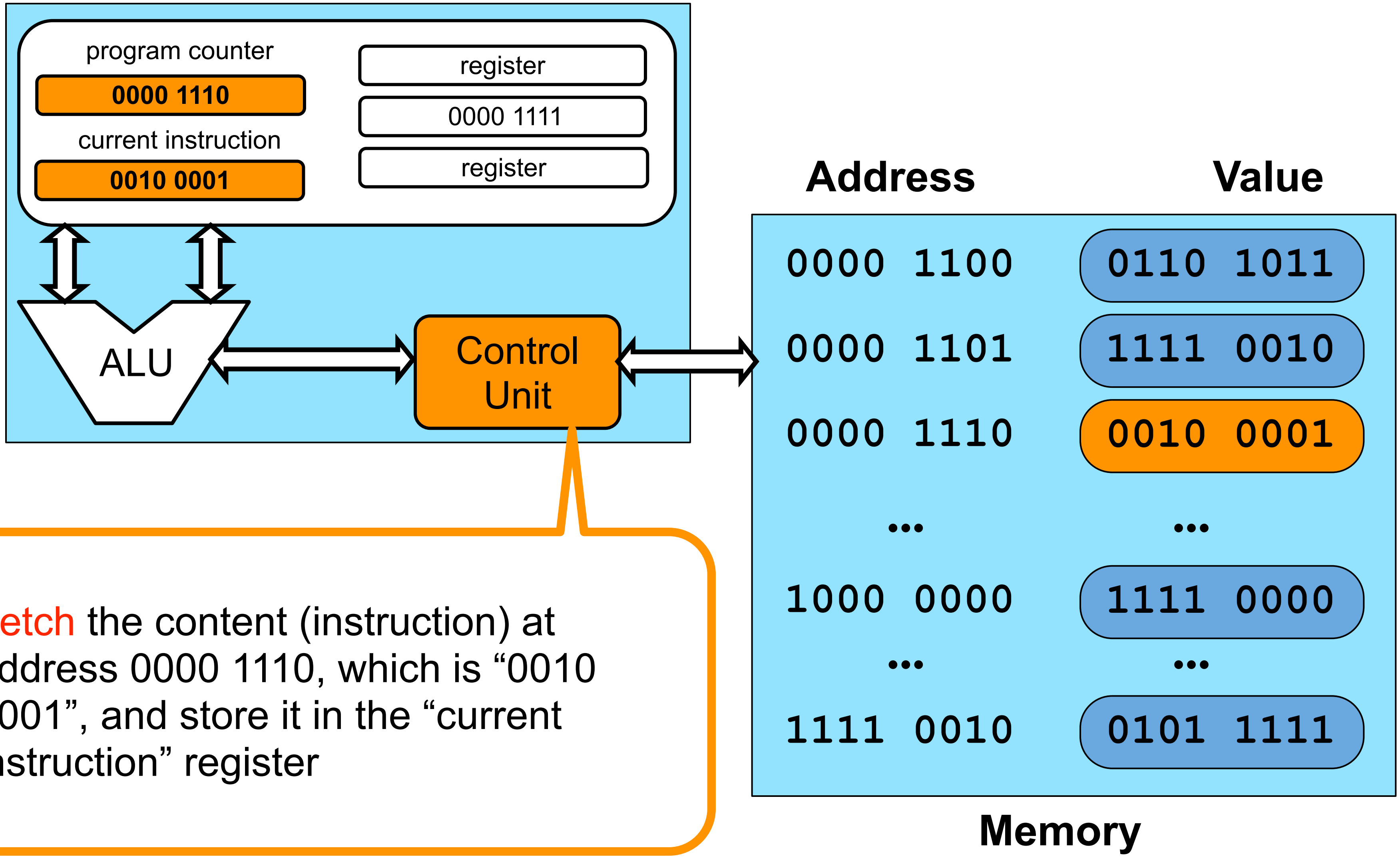
Fetch-Decode-Execute



Fetch-Decode-Execute

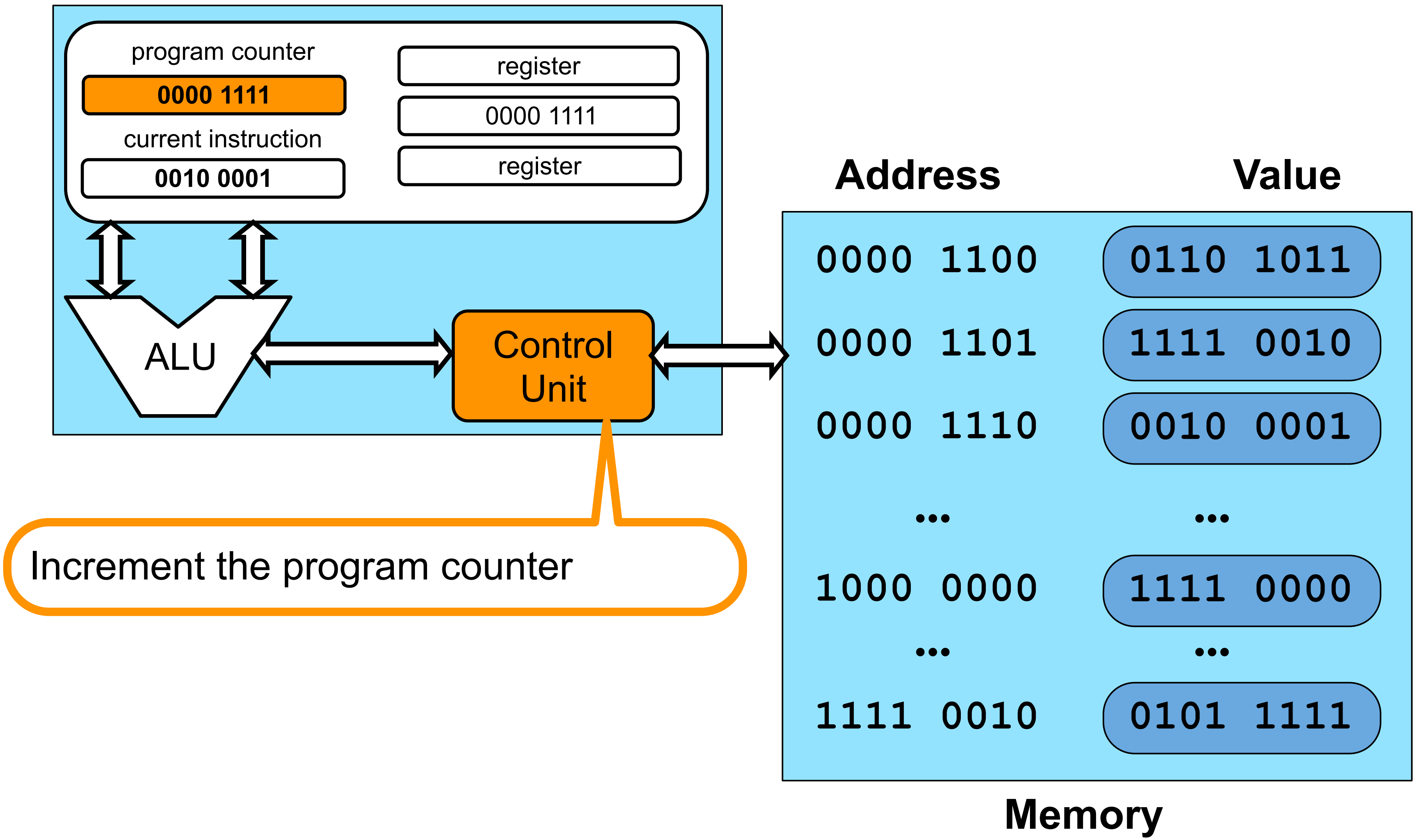


Fetch-Decode-Execute

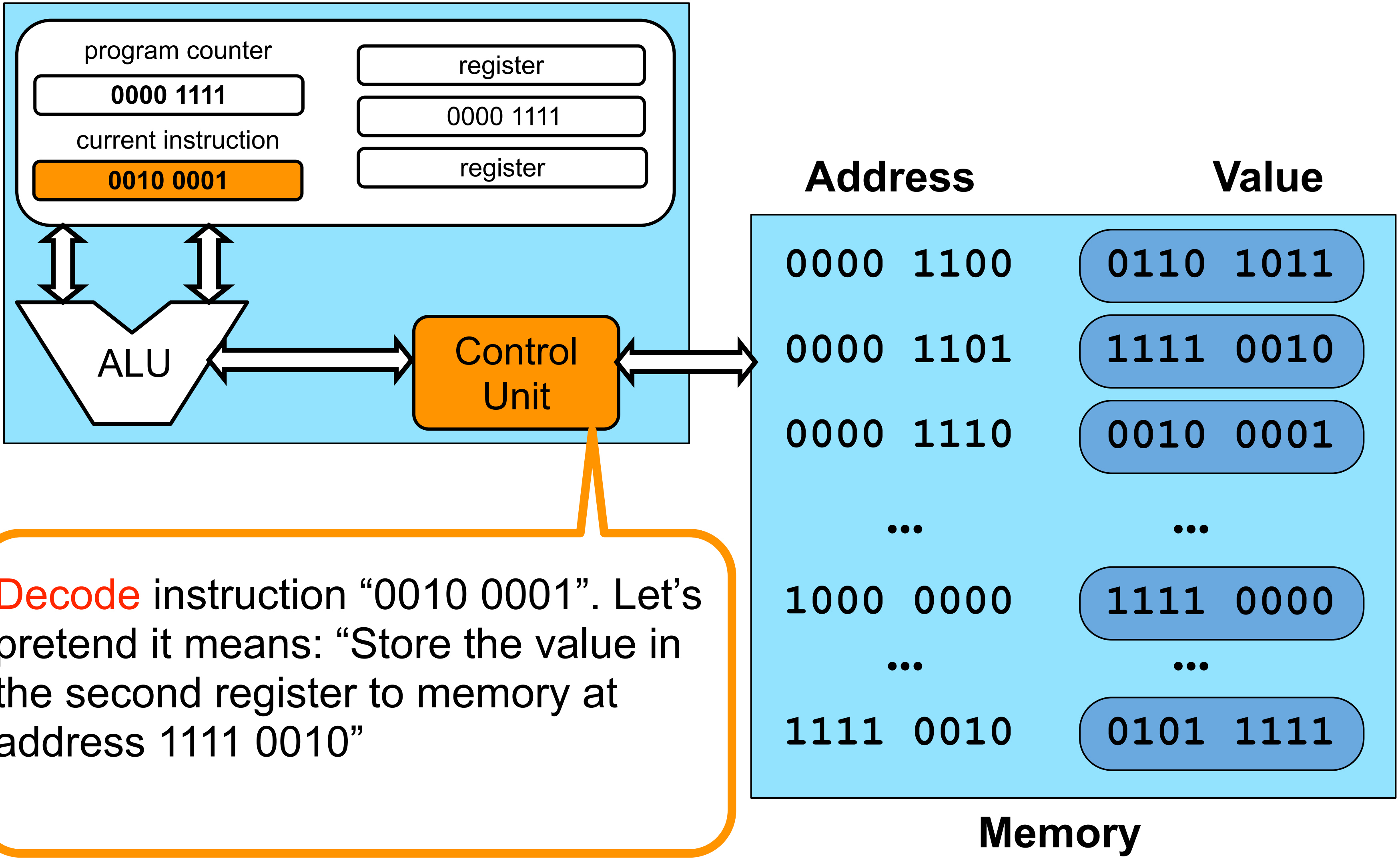


Fetch the content (instruction) at address 0000 1110, which is "0010 0001", and store it in the "current instruction" register

Fetch-Decode-Execute

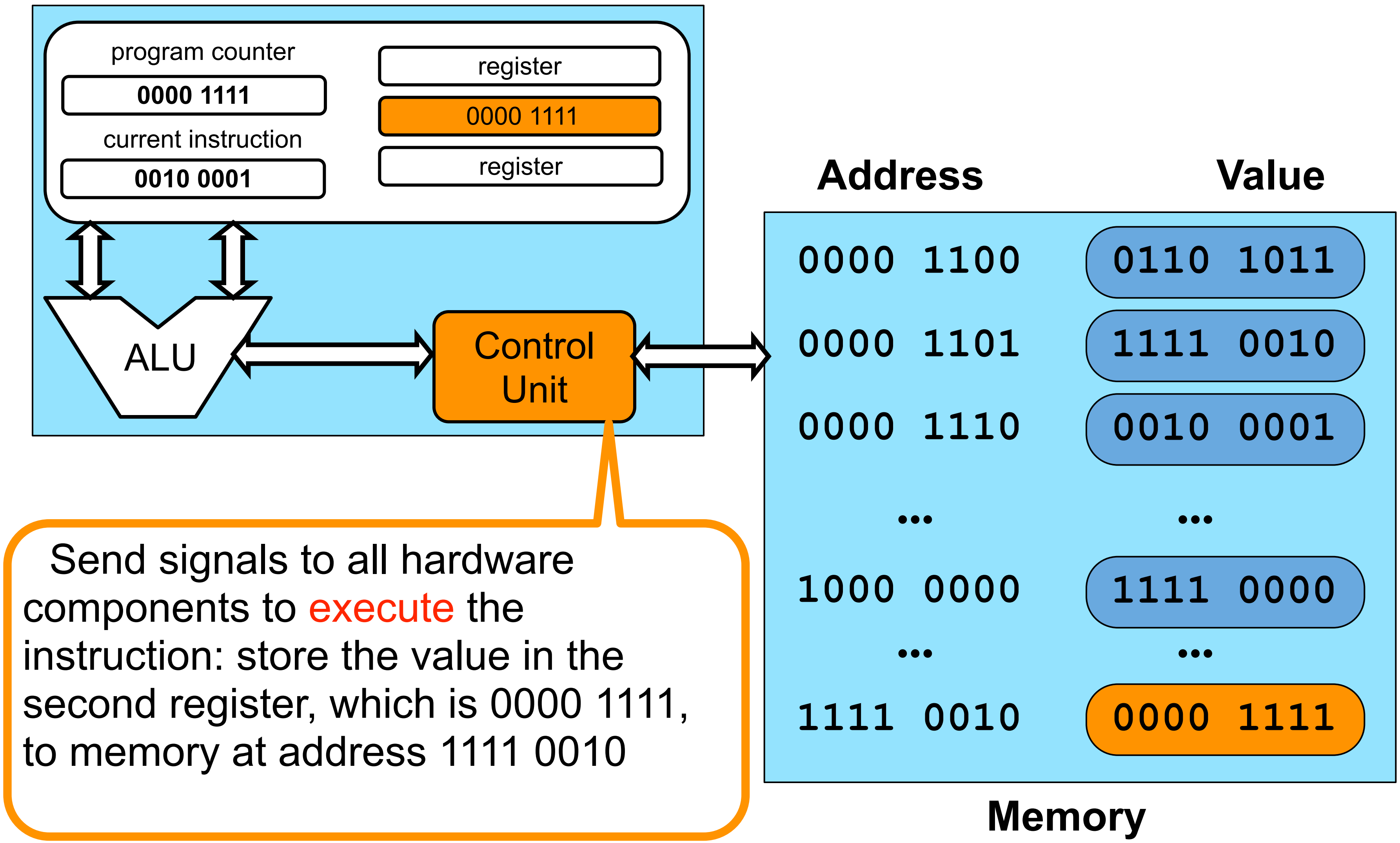


Fetch-Decode-Execute



Decode instruction "0010 0001". Let's pretend it means: "Store the value in the second register to memory at address 1111 0010"

Fetch-Decode-Execute



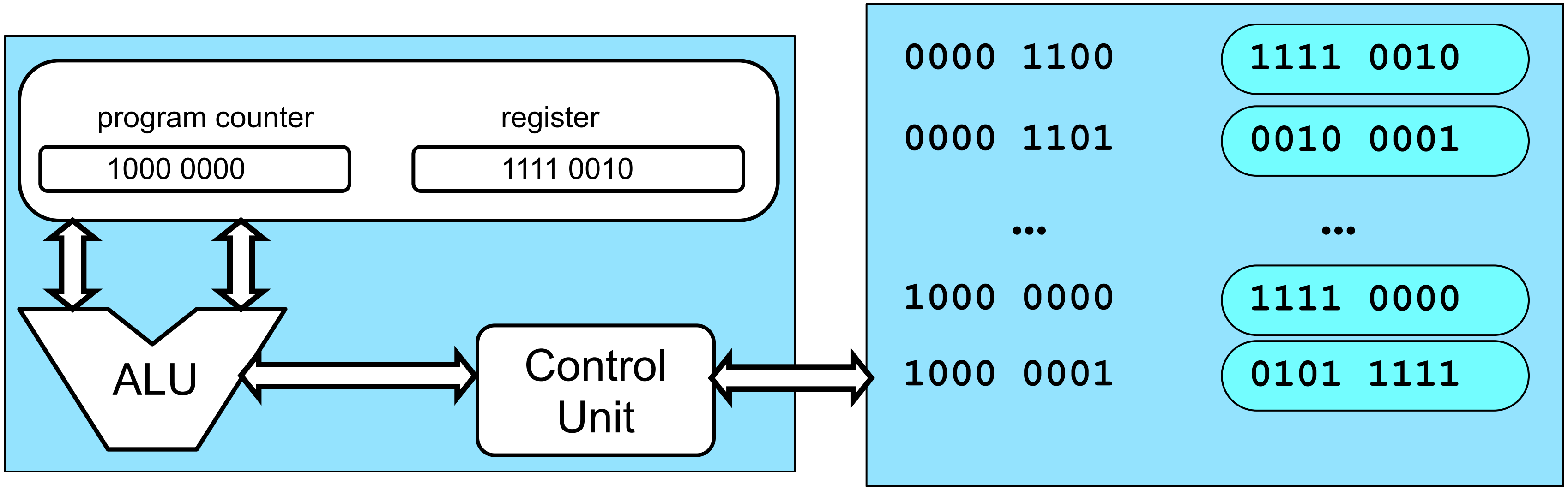
Fetch-Decode-Execute

- This is only a simplified view of the way things work
- The “control unit” is not a single thing
  - Control and data paths are implemented by several complex hardware components
- There are multiple ALUs, there are caches, there are multiple CPUs in fact (“cores”)
- Execution is pipelined: e.g., while one instruction is fetched, another one is being executed
- Decades of computer architecture research have gone into improving performance, thus often leading to staggering hardware complexity
  - Doing smart things in hardware requires more logic gates and wires, thus increasing processor cost
- But conceptually, fetch-decode-execute is it

In-Class Exercise

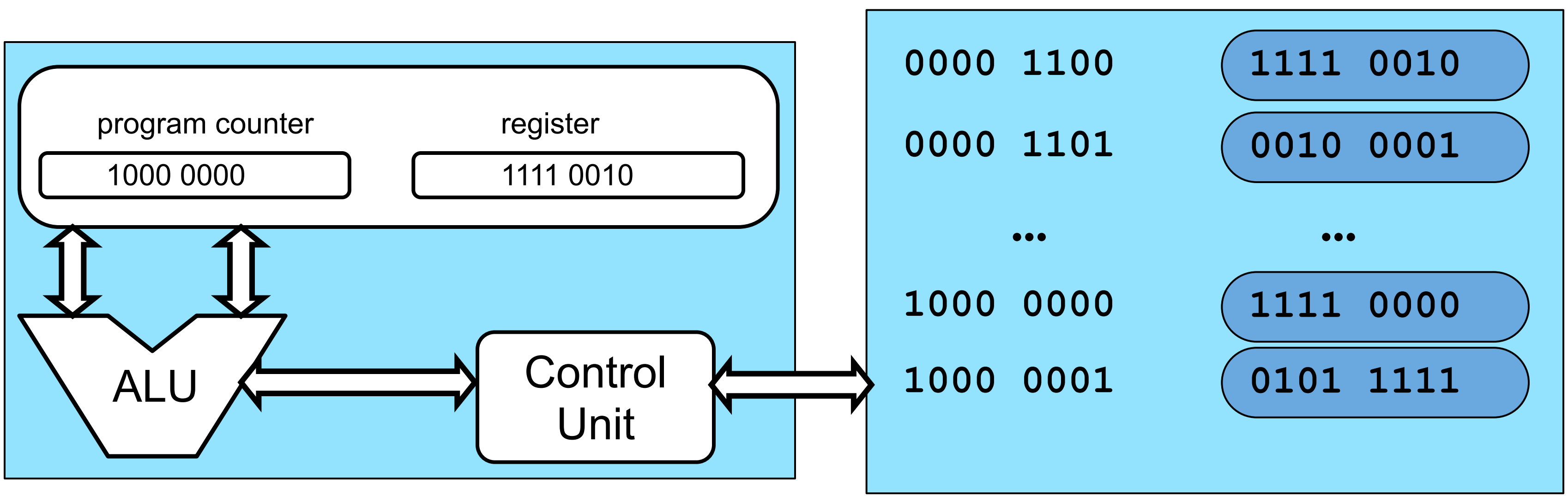
- With the following instruction set definition and machine state, what is the new memory state after execution completes?

code	operation
1111 0000	Increment the register
1111 0010	Decrement the register
0101 1111	Store register to address NOT(register)



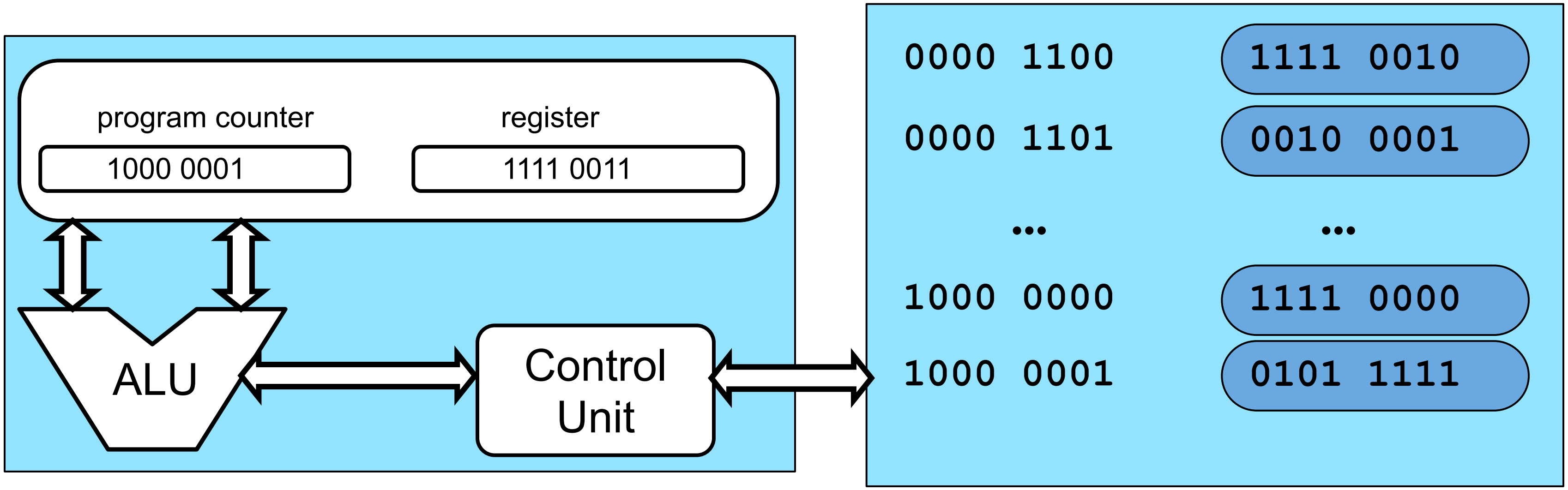
- Fetch the instruction: “1111 0000”

code	operation
1111 0000	Increment the register
1111 0010	Decrement the register
0101 1111	Store register to address NOT(register)



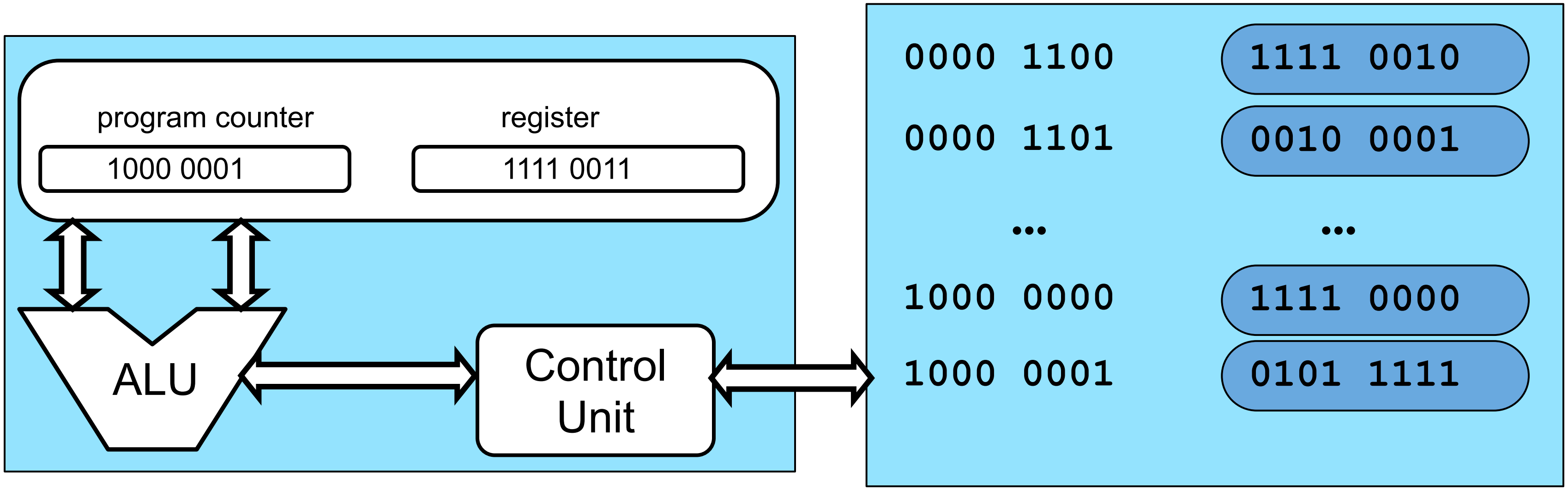
- Fetch the instruction: “1111 0000”
- Execute it: increment register to value “1111 0011”

code	operation
1111 0000	Increment the register
1111 0010	Decrement the register
0101 1111	Store register to address NOT(register)



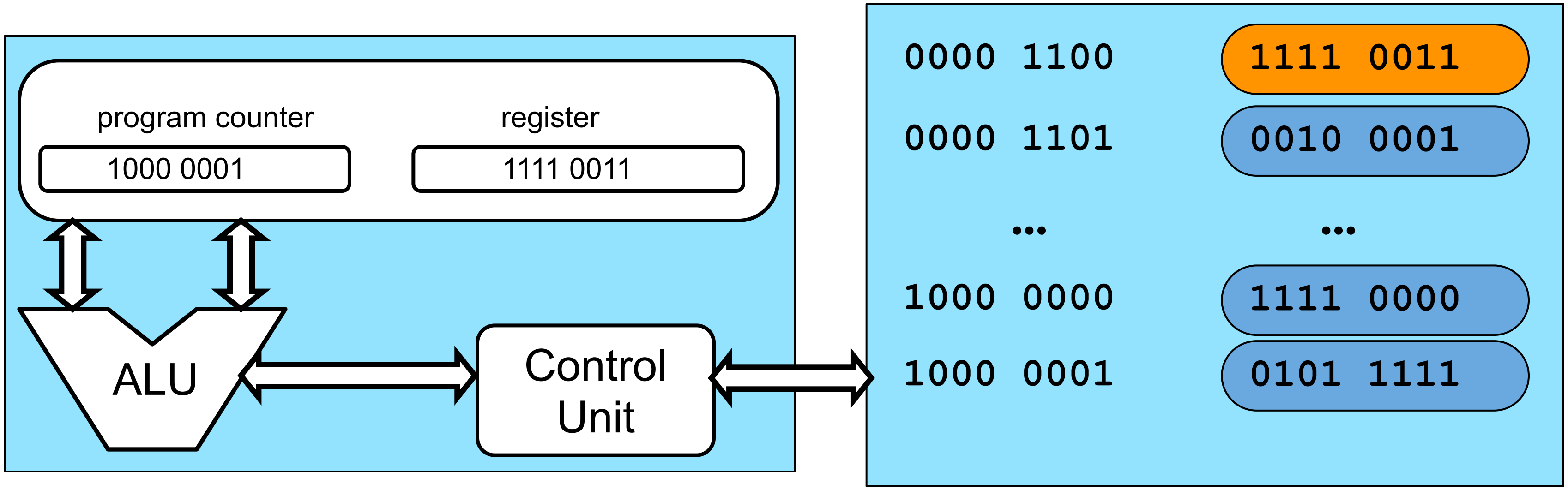
- Fetch the instruction: “1111 0000”
- Execute it: increment register to value “1111 0011”
- Fetch the next instruction: “0101 1111”
-

code	operation
1111 0000	Increment the register
1111 0010	Decrement the register
0101 1111	Store register value to address NOT(register)



- Fetch the instruction: “1111 0000”
- Execute it: increment register to value “1111 0011”
- Fetch the next instruction: “0101 1111”
- Execute it: store value “1111 0011” to address “0000 1100”

code	operation
1111 0000	Increment the register
1111 0010	Decrement the register
0101 1111	Save register to address NOT(register)



The Clock

- Every computer maintains an internal clock that regulates how quickly instructions can be executed, and is used to synchronize system components
  - Just like a metronome
- In the previous example, each “event” happens at a different “tick” of the clock
- The frequency of the clock is called the clock rate
- The time in between two clock ticks is called a clock cycle or cycle for short
- Clock cycle = 1 / Clock Rate
  - Clock rate = 2.4 GHz
  - Clock cycle = 1 / (2.4*1000*1000*1000)

    = 0.416 e^{-9} sec

    = 0.416 ns (nanosec)

Instructions

- Instructions are encoded in binary machine code
  - e.g.: 01000110101101 may mean “perform an addition of two registers and store the results in another register”
- The CPU is built using gates (OR, AND, etc.) which themselves use transistors
- These gates implement instruction decoding
  - Based on the bits of the instruction code, signals are sent to different electronic components, which then perform useful tasks
- Typically, an instruction consists of two parts
  - The opcode: what the instruction should do
  - The operands: the input to the computation

opcode	operands
0 1 0 0 0	1 1 0 1 0 1 1 0 1

Instruction Set Architecture (ISA)

- When designing a CPU, one must define the set of all the instructions it understands
  - This is one thing that Intel engineers do
- This is called the ISA: Instruction Set Architecture
- Typical ISA include instructions for
  - Performing arithmetic operations on register values
  - Load values from memory into registers
  - Store values from registers into memory
  - Test register values to decide what instruction to execute next
  - ...
- Envision a loooong specification manual that lists all the possible instructions...

ISA specification Example: x86

Let’s look at the Web site http://ref.x86asm.net/

pf	0F	po	so	o	proc	st	m	rl	x	mnemonic	op1	op2	op3	op4	iext	tested f	modif f	def f	undef f	f values	description, notes
		00		r					L	ADD	**r/m8**	r8					o..szapc	o..szapc			Add
		01		r					L	ADD	**r/m16/32**	r16/32					o..szapc	o..szapc			Add
		02		r						ADD	**r8**	r/m8					o..szapc	o..szapc			Add
		03		r						ADD	**r16/32**	r/m16/32					o..szapc	o..szapc			Add
		04								ADD	**AL**	imm8					o..szapc	o..szapc			Add
		05								ADD	**eAX**	imm16/32					o..szapc	o..szapc			Add
		06								PUSH	ES										Push Word, Doubleword or Quadword Onto the Stac
		07								POP	**ES**										Pop a Value from the Stack
		08		r					L	OR	**r/m8**	r8					o..szapc	o..sz.pc	.....a..	o......c	Logical Inclusive OR
		09		r					L	OR	**r/m16/32**	r16/32					o..szapc	o..sz.pc	.....a..	o......c	Logical Inclusive OR
		0A		r						OR	**r8**	r/m8					o..szapc	o..sz.pc	.....a..	o......c	Logical Inclusive OR
		0B		r						OR	**r16/32**	r/m16/32					o..szapc	o..sz.pc	.....a..	o......c	Logical Inclusive OR
		0C								OR	**AL**	imm8					o..szapc	o..sz.pc	.....a..	o......c	Logical Inclusive OR
		0D								OR	**eAX**	imm16/32					o..szapc	o..sz.pc	.....a..	o......c	Logical Inclusive OR
		0E								PUSH	CS										Push Word, Doubleword or Quadword Onto the Stac

operands

- pf Prefix
- 0F 0F Prefix
- po Primary Opcode
- so Secondary Opcode

opcode in HEX

Human-readable Mnemonic (assembly)

what it does

- x Lock Prefix/FPU Push/FPU Pop
- mnemonic Instruction Mnemonic
- op1, op2, ... Instruction Operands
- iext Instruction Extension Group
- grp1, grp2, grp3 Main Group, Sub-group, Sub-sub-group
- tested f, modif f, def f, undef f Tested, Modified, Defined, and Undefined Flags
- f values Flags Values
- description, notes

Assembly language

- It’s really difficult for humans to read/remember binary instruction encodings
  - But people used to do it!
  - One would typically use hexadecimal encoding, but still it seems impossible to memorize all this in today’s world
- Therefore, it is typical to use a set of *mnemonics*
- We call these mnemonics the assembly language
  - It is often said that the CPU understands assembly language
  - This is not technically true: the CPU understands machine code, which we, as humans, choose the represent using assembly language
- An assembler is a computer program that transform assembly code into machine code (i.e., from a human-readable format into a binary CPU-readable format)

Assembly Language

- It used to be that *all* computer programmers did all day was write assembly code
- This was difficult for many reasons
  - Difficult to read and maintain (in spite of using the mnemonics)
  - Difficult to debug
  - Different from one computer to another!
- The exclusive use of assembly language for all programming prevented the (sustainable) development of large software project with more than a few programmers
- This was the main motivation for developing high-level languages
  - FORTRAN, Cobol, C, etc.

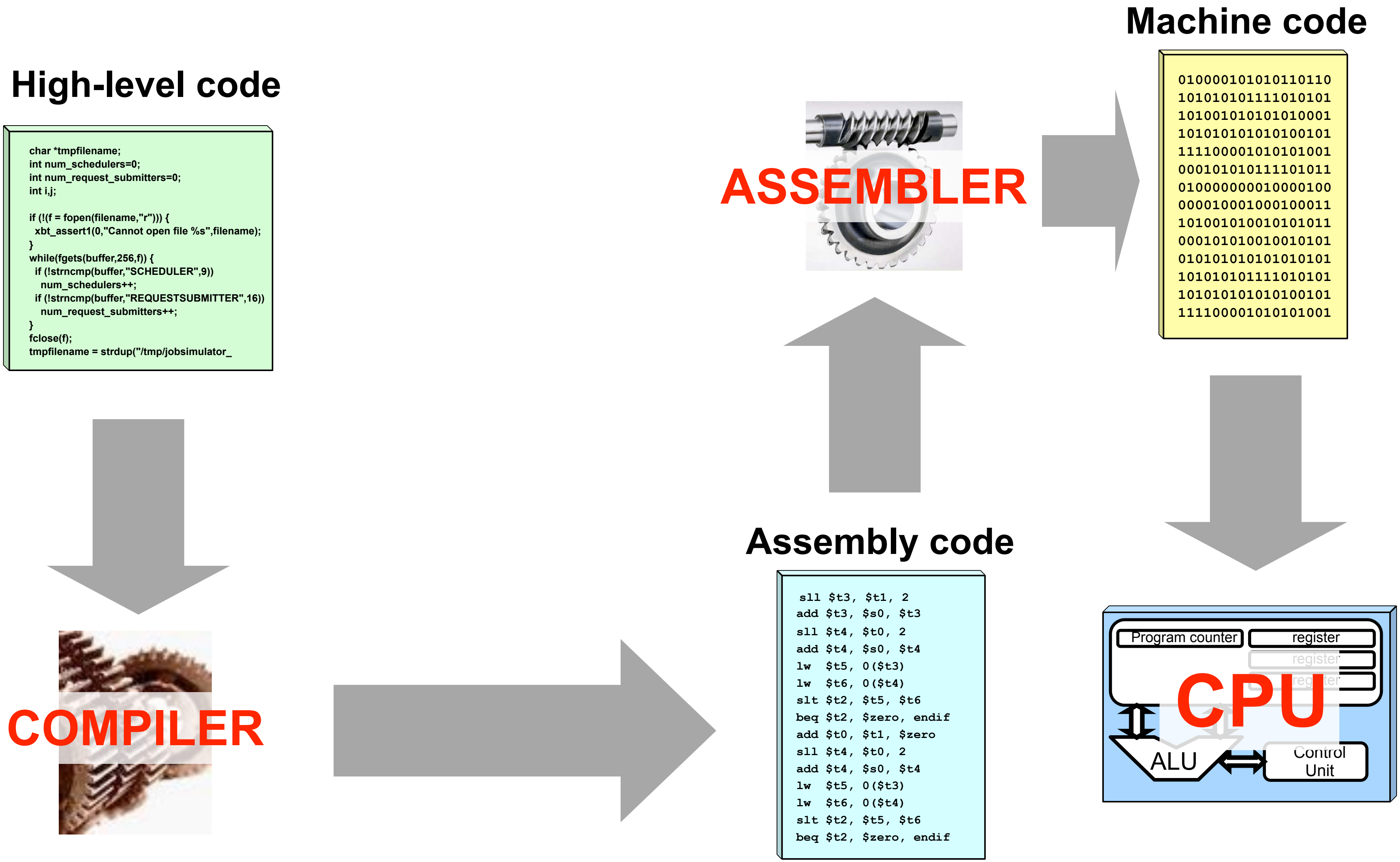
High-level Languages

- The first successful high-level language was FORTRAN
  - Developed by IBM in 1954 to run on they 704 series
  - Used for scientific computing
- The introduction of FORTRAN led people to believe that there would never be bugs again because it made programming so easy!
  - But high-level languages led to larger and more complex software systems, hence leading to bugs
- Another early programming language was COBOL
  - Developed in 1960, strongly supported by DoD
  - Used for business applications
- In the early 60s IBM had a simple marketing strategy
  - On the IBM 7090 you used FORTRAN to do science
  - On the IBM 7080 you used COBOL to do business
- Many high-level languages have been developed since then, and they are what most programmers use
  - Fascinating history (see ICS 313)

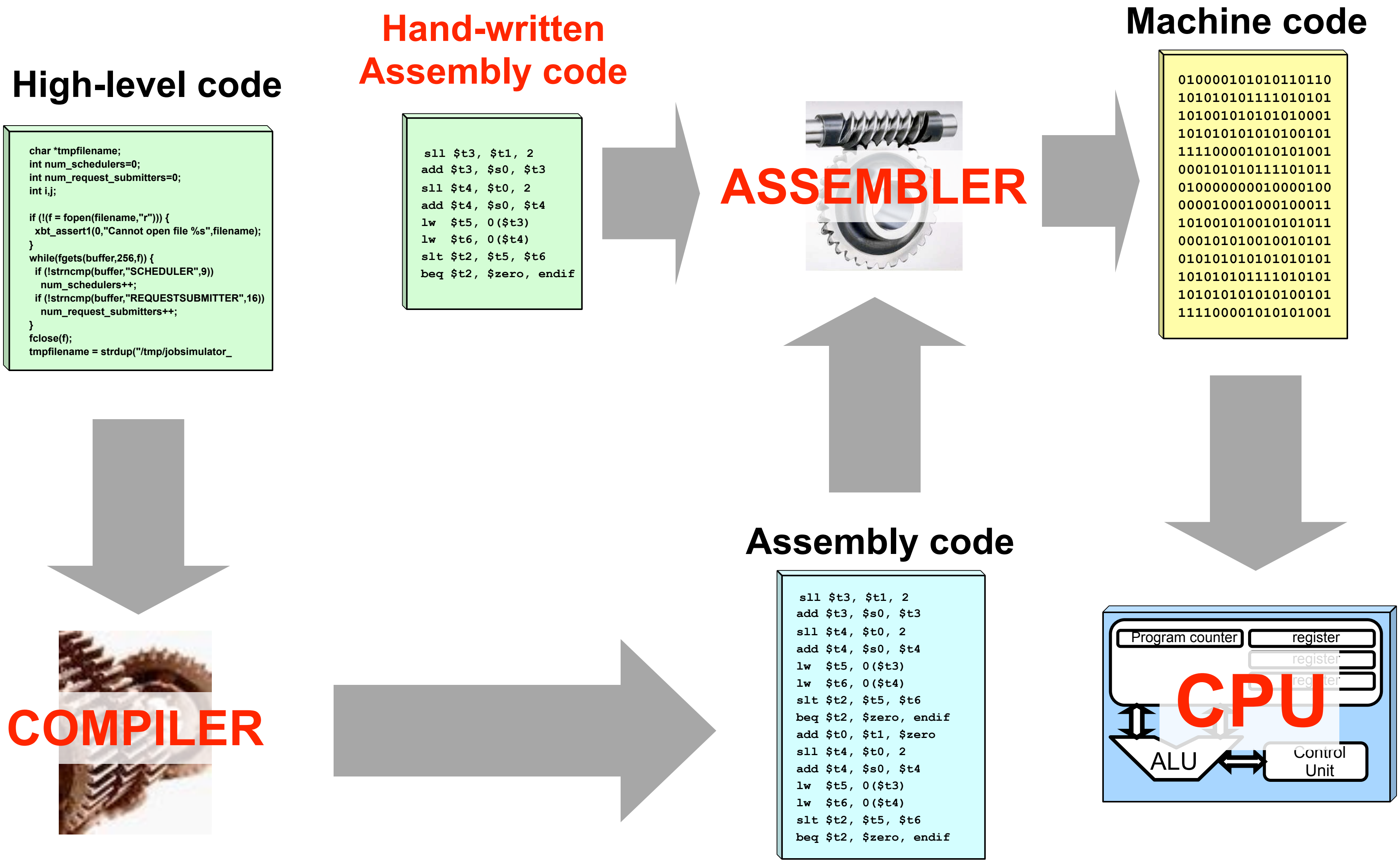
High-Level Languages

- Having high-level languages is good, but CPUs do not understand them
  - As we saw, they only understand very basic instructions to manipulate registers, etc.
- Therefore, there needs to be a translation from a high-level language to machine code
- The translation is done by a compiler
- Let’s see this on a picture....

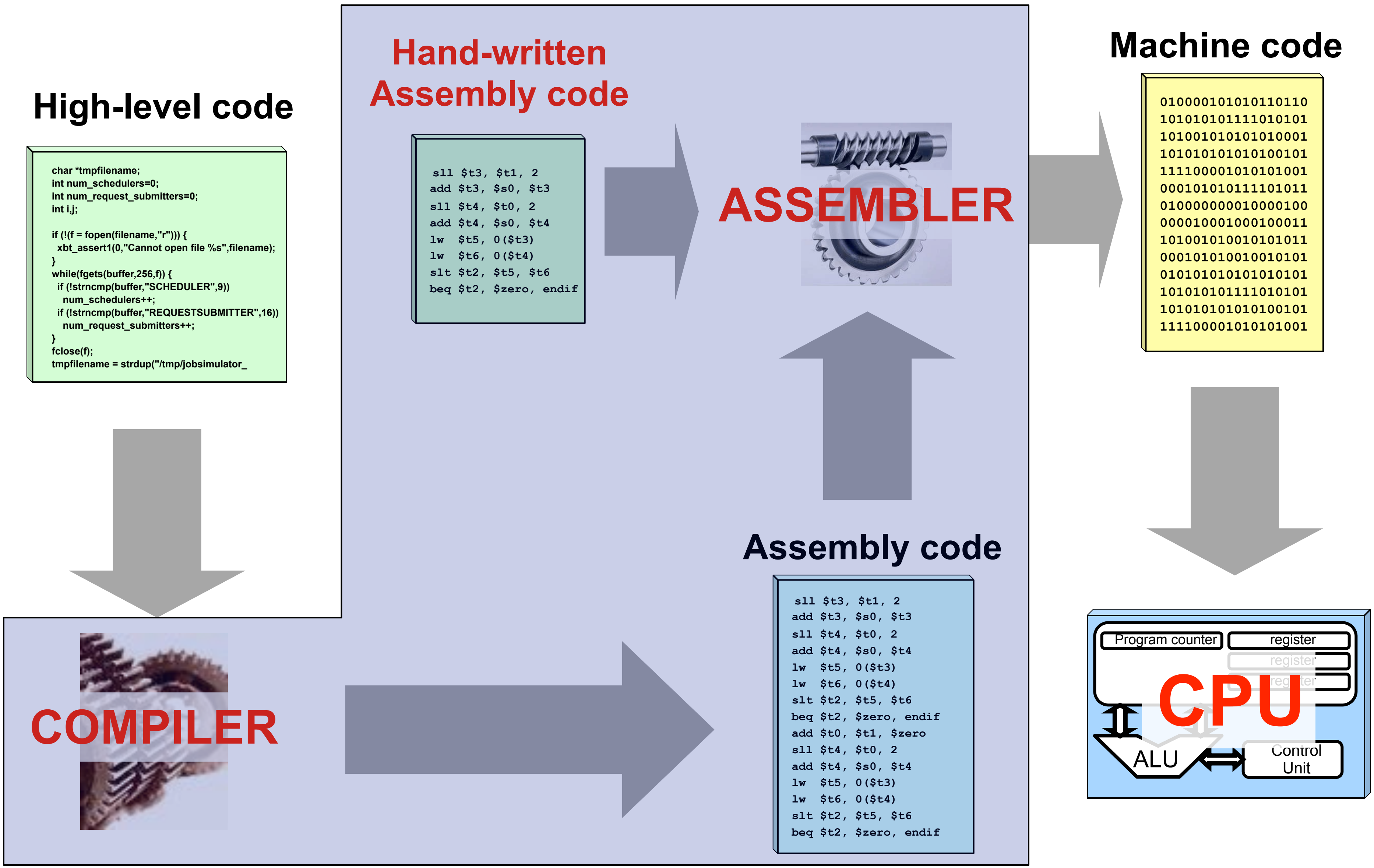
The Big (Simplified) Picture



The Big (Simplified) Picture



This course's topics:



What we do in this course

- First part of the semester (bulk of the course)
  - Learn how to write assembly code
    - For the x86 architecture
  - Use an assembler to generate binary code from our assembly code and then run it
- Second part of the semester (shorter, but absolutely fundamental)
  - Learn about important tools tools
    - loader, linker, **compiler**, debugger, etc.

Why should we learn all this?

- Why should we learn how to write assembly code?
  - Students: “We won’t write assembly code for a living!”
- **Reason #1:** Many of you will have to read/write *some* assembly code at some point
  - Write small piece of assembly for performance optimization as part of larger software projects
  - Write assembly code for embedded devices
  - Read generated assembly to better understand what’s going on (once in a blue-moon, but a total life saver)
- **Reason #2:** Learning assembly makes you a better programmer in high-level languages
  - Makes you keenly aware of what happens under the cover, which allows for easier debugging
  - Makes you understand “performance bugs”
  - Allows you to write more efficient high-level code

Why should we learn all this?

- Why should we learn how compilers work?
  - Students: “We won’t develop compilers for a living!”

- **Reason #1:** Many of you will develop “some” compilers
  - Some of you may develop a compiler for a programming language
  - But often one has to write “compilers” for things that one doesn’t always think of as programming languages
    - e.g., to convert files from one format to another!

- **Reason #2:** Knowing how a compiler works makes you a better programmer
  - You now understand the connection between high-level code and generated assembly code (see previous slide)
  - You understand what some high-level language constructs really entail under the cover, and thus understand their performance implications
  - You can better understand what a compiler can and cannot do for you

Why should we learn all this?

- Meta-reason: this course should go a long way in giving you a holistic understanding of how a program goes from just a text file to a running code
  - You should be able to describe in low-level details how you go from “I wrote a piece of C code that calls a function that adds 2 and 2 together and prints the result” to “the computer prints 4”
    - In its full glory after you’ve taken ICS332
  - The complexity of such a simple thing is actually quite stunning
  - There should be something satisfying in knowing how things work from top to bottom!
- 99% of students come into ICS312 thinking “why do we have to do this???”, and ~75% leave thinking “now I understand so much more and I feel like a computer scientist!!”
  - This is based on my own discussion/e-mail with students/alumni, not an official study :)

Conclusion

- If you want to know more
  - Take a computer architecture course
  - Classic Textbook: Computer Organization and Design, Fourth Edition: The Hardware/ Software Interface (Patterson and Hennessy, Morgan Kaufmann)
- Next lecture we'll go through a practice (ungraded) quiz in class