**Computer Architecture**

**Course Description:**

This course explores how computers are built and how they work. We will cover three main areas:

1. **Instruction Set Architecture (ISA):** is the set of commands that a computer's CPU can understand and execute. It serves as a bridge between software and hardware. Think of ISA as the language of the computer, with its own vocabulary (the instructions) and grammar (the rules for writing those instructions). For example, an ISA might include commands like LOAD, ADD, and STORE, which tell the computer to move data, perform arithmetic, or save results. Each type of computer, such as Intel or ARM, has its own ISA, which means software written for one type may not work on another without modification. ISA is crucial for programmers because it defines how they can communicate with the computer. Understanding ISA allows developers to write efficient programs that can be executed by the hardware. Overall, ISA plays a fundamental role in the design and functionality of computer systems.
2. **Micro-architecture:** When the computer hardware (CPU, memory, buses) receives instructions defined by its **Instruction Set Architecture (ISA)**, the micro-architecture explain how those instructions will be executed by the hardware. It focuses on how different parts of the computer, like the CPU (the brain), memory (where data is stored), and other connections (buses), work together to perform tasks. For example, when the computer needs to add two numbers, micro-architecture decides how to fetch those numbers, perform the addition, and save the result.
3. **System Architecture:** System Architecture is about how all the parts of a computer are organized and how they work together. It looks at the main components, like the CPU (the brain), memory (where data is stored), hard drives (for long-term storage), and devices like keyboards and printers. You can think of it like a city where different buildings (components) are connected by roads (the ways they communicate). For example, when you save a file, the CPU tells the memory where to keep it, and system architecture decides how that information travels between them.

**Course Objectives:**

1. **Discuss Representation of Data and Algorithms:**

We will learn how data is represented in computers (like using 0s and 1s) and how different algorithms (step-by-step instructions) are used to perform tasks on this data. For example, how numbers are stored in memory or how text is encoded.

1. **Demonstrate Different Operations in Terms of Micro-operations:**

**Micro-operations** are the tiny steps that a computer takes to complete a single instruction. When you give the computer a command, it breaks that command down into smaller actions to make it easier to handle. For example, if you want the computer to add two numbers, the steps might include:

1. **Getting the first number** from memory and putting it in a temporary storage area (called a register).
2. **Getting the second number** and putting it in another register.
3. **Adding the two numbers** together using a special part of the computer called the arithmetic logic unit (ALU).
4. **Saving the result** back in memory or another register.

Each of these steps is a micro-operation. By breaking down tasks this way, the computer can manage complex operations more effectively. Understanding micro-operations helps us see how computers perform calculations and process data efficiently. Overall, they are the building blocks that allow computers to carry out larger instructions.

1. **Explain Architecture of Basic Computer and Micro-programmed Control Unit:**
   1. We will look at the basic building blocks of a computer (like the CPU, memory, and input/output systems) and how these parts work together. A micro-programmed control unit is like the manager of these parts, directing them on what to do based on the commands it receives.
2. **Understand Memory and I/O Organization of a Typical Computer System:**

In a typical computer, **memory** is divided into different types, mainly **RAM** (Random Access Memory) and **cache**.

* **RAM** is where the computer keeps the data and programs it is currently using, making it easy for the CPU (the brain of the computer) to access this information quickly.
* **Cache memory** is a smaller and faster type of memory that sits close to the CPU. It stores frequently used data, allowing even faster access when the CPU needs it.

**Input/Output (I/O) devices** are the tools we use to interact with the computer. For example, keyboards allow us to input information, mice help us navigate, and printers output documents. These devices connect to the computer to send and receive data, helping us communicate with the system.

Data moves in and out of the computer through **buses**, which are like pathways that carry information between the CPU, memory, and I/O devices. For instance, when you type on a keyboard, that information goes to the CPU, which processes it and then may send the result to the screen or a printer. Overall, understanding how memory and I/O devices work together helps us see how a computer handles information and interacts with users.

1. **Demonstrate Benefits of Pipelined Systems:**
   1. Pipelining is a technique that allows a computer to work on multiple instructions at the same time, similar to an assembly line in a factory. We will study how this improves performance and speeds up the execution of programs by overlapping different stages of instruction processing.

**Computer Architecture** is the study of how computers are built and work. It covers three main areas:

* **Instruction Set Architecture (ISA):** This is the language computers understand. It defines the commands (instructions) a CPU can execute.
* **Micro-architecture:** This explains how the hardware (CPU, memory, buses) carries out the instructions defined by the ISA.
* **System Architecture:** This focuses on how all the parts of a computer are organized and work together.

The course objectives include:

* Understanding how data is represented in computers and how algorithms are used to process it.
* Learning about micro-operations, the small steps a computer takes to execute instructions.
* Exploring the basic components of a computer and how they work together.
* Understanding memory and input/output (I/O) organization.
* Demonstrating the benefits of pipelining, a technique that allows computers to work on multiple instructions at once.

### Real-Life Example of Overlapping Register Windows

**Imagine a restaurant kitchen with multiple chefs working on different dishes.** Each chef has their own individual workspace, but they also have a shared pantry where they can store ingredients.

**Traditional Approach:**

* **Separate Workspaces:** Each chef has their own workspace with a limited number of ingredients.
* **Storing and Retrieving:** When a chef needs an ingredient that isn't in their workspace, they have to go to the pantry, retrieve it, bring it back to their workspace, and store it there. This takes time and effort.

**Using a Shared Workspace:**

* **Centralized Storage:** Instead of each chef having their own pantry, there's a central shared pantry.
* **Direct Access:** Chefs can directly access ingredients from the shared pantry without having to store them in their individual workspaces.
* **Reduced Overhead:** This eliminates the extra steps of storing and retrieving ingredients, making the chefs' work more efficient and faster.

**Relating this to Register Windows**

* **Individual Workspaces:** Each register window is like a separate workspace for a function.
* **Shared Pantry:** **The overlapping region** is the shared pantry between adjacent windows.
* **Data (arguments or return values)** can be placed directly in the overlapping region without needing to be saved to memory.
* **When a function is called**, the window pointer moves to the next window, but the overlapping region remains accessible to both windows.
* **Avoiding Extra Steps:** By using the overlapping region, functions can directly access and share data without needing to save and restore it in their own private registers. This avoids the overhead of saving and restoring data, making function calls more efficient.

**Efficient Function Calls**

Imagine you're at a restaurant where the chef is cooking multiple courses for a meal. Normally, after finishing one course, the chef would **completely clean and reset the kitchen** before starting to cook the next course. This would take a lot of time.

Instead, the chef decides to **keep some tools and ingredients** from the previous course already set up and ready to use. For example, if the knife and cutting board are still needed, they are left in place rather than being put away and taken out again.

By doing this, the chef can **start the next dish much faster** because everything isn’t being reset completely each time. Similarly, overlapping register windows in a computer allow for **efficient switching between tasks** without having to completely "clean up" after each one, making the whole process faster.

Flynn's Classification: A Real-World Analogy

**Imagine a busy restaurant kitchen.**

* Chef => Instruction stream
* Tasks => Data stream

1. SISD (Single Instruction, Single Data)
2. SIMD (Single Instruction, Multiple Data)
3. MISD (Multiple Instruction, Single Data)
4. MIMD (Multiple Instruction, Multiple Data)

**SISD (Single Instruction, Single Data)**

* **Chef:** A single chef preparing a single dish (One dish, one step at a time)
* **Tasks:** The ingredients for that particular dish.
* **Example:** A chef preparing a pasta dish, boiling the pasta, adding sauce, and then grating cheese.

**SIMD (Single Instruction, Multiple Data):**

* **Chef:** A single chef using a commercial oven to bake multiple cookies.
* **Tasks:** Multiple items being processed simultaneously with the same instruction.
* **Example:** A pastry chef baking a tray of cookies. All cookies are baked in the same oven at the same time.

**MISD (Multiple Instruction, Single Data):**

* **Chefs:** Multiple chefs working on the same dish, each with a different task.
* **Tasks:** A single item being processed by multiple instructions.
* **Example:** A team of chefs preparing a roast. One chef might sear the meat, another might prepare the gravy, and another might roast the vegetables.

**MIMD (Multiple Instruction, Multiple Data):**

* **Chefs:** Multiple chefs preparing different dishes, each following their own recipes
* **Tasks:** Multiple items being processed simultaneously with different instructions.
* **Example:** A busy kitchen with multiple chefs preparing a variety of dishes, each working independently.

**SISD** is like a single chef preparing a single dish.

**SIMD** is like a chef using a commercial oven to bake multiple cookies at once.

**MISD** is like multiple chefs working together on a single dish.

**MIMD** is like a busy kitchen with multiple chefs preparing different dishes.

A diagram of a simd

Description automatically generatedA diagram of a pool

Description automatically generated

**Instruction Pool (green box) =>** These are the program instructions that tell the computer what operations to perform.

**Data Pool (blue box) =>** This contains the data that the instructions will operate on.

**PU (Processing Unit, shown in pink) =>** This is the central component that:

* Fetches instructions from the instruction pool
* Retrieves data from the data pool
* Executes the instructions on the data

**=========================================================**

**Instruction Pool (top):** There is a single instruction stream that broadcasts the same instruction to all Processing Units. This means all PUs will perform the same operation simultaneously.

**Data Pool (left):** Contains different data elements that need to be processed.

**Multiple Processing Units (PUs in pink):** There are multiple processing units (4 shown in this diagram) that:

* All receive the same instruction at once
* Each works on different data elements
* Execute in parallel

**Vector Unit (blue box containing the PUs):** This is the parallel processing component that contains all the PUs working together.

A diagram of a data pool

Description automatically generatedA diagram of a data pool

Description automatically generated

**Instruction Pool (green box):** Contains multiple different instructions. Unlike SISD or SIMD, each Processing Unit receives different instructions to execute.

**Data Pool (blue box):** Contains a single stream of data that is shared between all Processing Units.

**Processing Units (PUs in pink):** Multiple PUs each:

* Receive different instructions
* Work on the same data
* Operate independently but on the same data stream

**=========================================================**

**Instruction Pool (green box):** Contains multiple different instructions. Each Processing Unit can execute different instructions independently.

**Data Pool (blue box):** Contains multiple data streams. Each Processing Unit can work on different data independently.

**Processing Units (PUs in pink):** Multiple PUs (8 shown in this diagram) where each:

* Can execute different instructions
* Can work on different data
* Operates independently of other PUs
* Can coordinate with other PUs when needed

Real-Life Examples of Pipeline Hazards

**Structural Hazards: A Busy Restaurant**

* **Scenario:** Imagine a small restaurant with a single chef. If multiple customers order complex dishes simultaneously, the chef might be overwhelmed and unable to handle all orders at once.
* **Pipeline Hazard:** The limited resources (the single chef) can't handle the workload (multiple complex orders) simultaneously.
* **Solution 1: Hire more chefs:** Increase the number of chefs to handle multiple orders simultaneously.

**Data Hazards: Waiting for a Friend**

* **Scenario:** You're meeting a friend for lunch. You need to wait for your friend to arrive before ordering.
* **Pipeline Hazard:** The action of ordering (equivalent to an instruction) depends on the result of your friend arriving (a previous event). If your friend hasn't arrived yet, ordering would be premature.
* **Solution 2: Bring your own food:** If you're unsure about your friend's arrival, be prepared with your own food.

**Control Hazards: Changing Plans**

* **Scenario:** You're planning a trip to the beach. If the weather forecast predicts rain, you might decide to change your plans and go to a museum instead.
* **Imagine you're driving a car.** Control hazards are like coming to a fork in the road. You need to decide which way to go based on the signs or your navigation system. If you make the wrong decision, you might have to backtrack or take a detour.
* **Pipeline Hazard:** The decision to change plans (equivalent to a branch instruction) affects the subsequent actions (visiting the museum instead of the beach).
* **Solution 1: Check the weather forecast regularly:** Stay updated on weather conditions to make informed decisions.

Mitigation Techniques

Let's explore real-life scenarios that correspond to the mitigation techniques used to address pipeline hazards:

### Forwarding: Bypassing the Data

* **Scenario:** You're making a sandwich. You need to spread mayonnaise on the bread before adding cheese.
* **Mitigation:** Instead of waiting for the mayonnaise to fully spread, you start adding cheese while the mayonnaise is still being spread. This avoids a delay in the sandwich-making process.

### Branch Prediction: Predicting the Outcome

* **Scenario:** You're driving on a highway and approaching an exit. Based on your previous experience or navigation system, you predict that you'll take the exit.
* **Mitigation:** You start slowing down and preparing to exit even before you reach the exit sign. This avoids sudden braking or lane changes if your prediction is correct.

### Delayed Branches: Inserting a Nop

* **Scenario:** You're planning a trip and need to decide whether to go to the beach or the mountains. You know that checking the weather forecast will take a few minutes.
* **Mitigation:** You decide to wait a few minutes before making a decision to ensure you have the most accurate weather information.

### Pipeline Scheduling: Rearranging Instructions

* **Scenario:** You're assembling a piece of furniture. You realize that you can install the legs before attaching the top panel, which will save time.
* **Mitigation:** You rearrange the assembly steps to optimize the process and avoid unnecessary delays.

Understanding the Store-Program Concept: A Real-Life Analogy

**Imagine a cookbook.**

* **Instructions:** The recipes in the cookbook are like instructions for a computer.
* **Memory:** The cookbook itself is like the computer's memory. It stores the instructions (recipes).
* **Program Counter:** As the cook, you follow the steps in the recipe in order. Similarly, in a computer, a special register called the program counter keeps track of which instruction to execute next.
* **Fetch-Execute Cycle:** When you cook, you:
  + - **Fetch:** Look at the next step in the recipe.
    - **Decode:** Understand what the step means (e.g., "add 1 cup of flour").
    - **Execute:** Perform the action (add the flour).
    - **Increment PC:** Once you've completed a step, you move on to the next one.

**Real-Life Example:**

Let's say you want to bake a cake.

1. **Instructions:** You open your cookbook to the cake recipe.
2. **Fetch:** You read the first step: "Preheat oven to 350°F."
3. **Decode:** You understand that you need to set the oven temperature.
4. **Execute:** You go to the oven and adjust the temperature.
5. **Increment PC:** You move on to the next step in the recipe.

**In essence, a computer's store-program concept is like following a recipe, where the recipe is stored in memory and the computer follows the instructions step by step.** This flexibility allows computers to perform a wide range of tasks, from simple calculations to complex simulations, by simply changing the instructions (recipes) they are given.

**Understanding the Key Register in Associative Memory**

**Conventional Memory (Address-Based)**

Imagine you’re at a parking lot. If you want to find your car, you need to remember the exact parking spot number where you left it (say, spot A5). Without knowing that number, finding your car would be hard because you have to check every spot.

**Associative Memory (Content-Based)**

Now, think of a different kind of parking lot where you don’t need to remember the exact spot. Instead, each car is tagged with things like its color, model, or license plate number. If you’re looking for a red car, you can just tell the parking attendant to find all the red cars. You don’t need to know the exact spot; you just use what you know about the car itself.

**The Difference**

* **Conventional Memory:** You find things by their exact location (like remembering a parking spot number).
* **Associative Memory:** You find things based on what they are or what you know about them (like searching for a red car).

**Components:**

1. Argument Register
2. Key Register
3. Associative Memory Array
4. Matching Process
5. Match Register
6. Results

**Scenario: Searching for books by the author "J.K. Rowling."**

**Argument Register:** You enter "J.K. Rowling" into the argument register. This is the exact name of the author you're searching for.

**Key Register:** You set the key register to 11111111, which indicates that the entire name "J.K. Rowling" should be used for the search. In this case, all bits are set to 1 because you want to match the complete author’s name.

**Associative Memory Array:** This is the library's database containing information about all the books. The system will compare the entire "J.K. Rowling" string against each entry in the database, which might look like this:

* "J.K. Rowling", "Harry Potter", ISBN12345
* "John Rowe", "History of Art", ISBN67890
* "Jane Doe", "Gardening Tips", ISBN54321

**Matching Process:** During the matching process, the associative memory system compares the argument register ("J.K. Rowling") with each entry in the associative memory array. Because the key register is set to 11111111, it will look for an exact match for the full name "J.K. Rowling."

* **First Entry:** "J.K. Rowling", "Harry Potter", ISBN12345 matches exactly with the argument register.
* **Second Entry:** "John Rowe", "History of Art", ISBN67890 does NOT match because "John Rowe" is different from "J.K. Rowling."
* **Third Entry:** "Jane Doe", "Gardening Tips", ISBN54321 does NOT match because "Jane Doe" is different from "J.K. Rowling."

**Match Register:** The match register will only set the corresponding bit to 1 for the first entry, indicating that "J.K. Rowling" was found in the database.

**Results:** The results will show that only the book by "J.K. Rowling" matched the search criteria. The system will display "J.K. Rowling", "Harry Potter", ISBN12345 as the result.

**Scenario: Searching for Authors Whose Name Contains "Row"**

**Argument Register:** You enter "Row" into the argument register. This indicates you’re searching for any author whose name contains "Row."

**Key Register:** You set the key register to 11100000 (assuming an 8-bit system), where the first three bits are set to 1 because you want to search only for the first three letters, "Row". The remaining five bits are 0, indicating that those parts of the names will be ignored in the search.

**Associative Memory Array:** This is the library's database containing information about all the books. The system will compare the "Row" substring against each entry in the database:

* + "J.K. Rowling", "Harry Potter", ISBN12345
  + "John Rowe", "History of Art", ISBN67890
  + "Jane Doe", "Gardening Tips", ISBN54321

**Matching Process:** During the matching process, the associative memory system compares the first three letters of the argument register ("Row") with each entry in the associative memory array:

* + **First Entry:** "J.K. Rowling", "Harry Potter", ISBN12345 matches because the first three letters "Row" match with "Row" in "Rowling."
  + **Second Entry:** "John Rowe", "History of Art", ISBN67890 also matches because "Row" matches with "Row" in "Rowe."
  + **Third Entry:** "Jane Doe", "Gardening Tips", ISBN54321 does NOT match because "Doe" does not start with "Row".

**Match Register:** The match register will set the corresponding bits to 1 for the first two entries, indicating that both "J.K. Rowling" and "John Rowe" matched the criteria.

**Results:** The results will show that the search found two matches:

* + "J.K. Rowling", "Harry Potter", ISBN12345
  + "John Rowe", "History of Art", ISBN67890

These two books are displayed because their authors' names contain the substring "Row" in the first three letters, satisfying the search criteria.

**How does DMA controller work? Give an example of DMA data transfer**

**Traditional Approach (CPU-Driven)**

1. **Initiation**:
   * When you insert the USB drive and select a file to copy, the operating system (OS) sends a request to the CPU to begin the transfer process.
2. **Byte-by-Byte Transfer**:
   * The CPU starts reading data from the USB drive. It does this byte by byte:
     + It sends a command to the USB controller to read the first byte of data.
     + The USB controller retrieves the byte from the USB drive and sends it back to the CPU.
     + The CPU then writes that byte to the destination location on the hard drive.
     + CPU have to do perform for each byte fetch decode execute instruction
   * This process repeats for every byte in the file, which can be very time-consuming, especially for large files.
3. **Resource Consumption**:
   * During this transfer, the CPU is heavily involved in managing the data transfer. It has to keep track of each byte, manage interrupts from the USB and hard drive controllers, and ensure that data is correctly written to memory.
   * This can significantly slow down the system, especially if you are running multiple applications, as the CPU's resources are dedicated to this file transfer.

**DMA Approach**

1. **Request**:
   * When a file transfer is initiated (for example, copying a file from a USB drive to a hard drive), the operating system informs the CPU of the request.
   * The OS then sends a command to the DMA controller, specifying the source (USB drive), destination (hard drive), and the amount of data to be transferred.
2. **Grant**:
   * **CPU Authorization**: The DMA controller needs to gain permission from the CPU to take control of the system bus.
   * The CPU acknowledges the request from the DMA controller, allowing it to proceed. Once permission is granted, the CPU release control of the bus to the DMA controller for the duration of the data transfer.
3. **Transfer**:
   * **Data Transfer**: After being granted access to the bus, the DMA controller takes charge of the data transfer:
     + It communicates directly with the USB controller to instruct it to send a block of data to the hard drive.
     + The DMA controller manages this process by moving multiple bytes at once, rather than handling each byte individually as the CPU would. This chunk-based approach significantly accelerates the transfer process.
4. **Notification**:
   * **Completion Signal**: Once the data transfer is complete, the DMA controller sends an interrupt signal to the CPU to notify it that the transfer has finished.
   * The CPU can then check the status of the transfer, process any error conditions if necessary, and continue with its other tasks, having been free to work on other processes while the data transfer occurred.

**Example of DMA Data Transfer**

**Scenario**: Copying a large file from a USB flash drive to the computer's memory (RAM).

**Process**:

1. **Initiation**: You insert a USB flash drive into your computer and select a large file to copy.
2. **Request**: The operating system (OS) sends a request to the DMA controller to begin the transfer, providing details about the source (USB drive), destination (RAM), and the size of the file.
3. **Grant**: The CPU acknowledges this request and gives permission for the DMA controller to access the system bus.
4. **Transfer**:
   * The DMA controller takes control and instructs the USB controller to send a block of data directly to RAM.
   * Instead of moving data byte by byte, the DMA controller transfers large blocks of data at once, significantly speeding up the process.
5. **Notification**: Once the transfer is complete, the DMA controller sends an interrupt signal to the CPU, notifying it that the operation is finished. The CPU can then check for errors and proceed with other tasks.

**Differentiate between hardwired control unit and a micro programmed control unit.**

Real-Life Example: A Vending Machine

**Hardwired Control Unit:**

Imagine a simple vending machine that only dispenses soda. Its control unit would be hardwired to perform a specific sequence of actions:

1. **Accept coin:** Check if the inserted coin is valid.
2. **Select product:** Determine the selected product based on the button pressed.
3. **Dispense product:** Release the soda can.
4. **Return change:** Calculate and dispense any remaining change.

This vending machine has a **fixed way of working**. If you wanted to add different products (like snacks or other drinks) or accept different types of coins, you would have to change the machine's hardware (its physical parts), which would be difficult.

**Microprogrammed Control Unit:**

Now imagine a fancier vending machine that sells all kinds of items, like snacks, drinks, and candy. It uses a microprogrammed control unit, which means its actions are controlled by small software programs (microinstructions) stored in memory.

1. **Microinstructions:** The machine’s software contains different steps for handling various tasks, like accepting coins, selecting products, and dispensing items. The microprogram can be easily updated to add new instructions.
2. **Flexibility:** This machine can be **easily updated** to sell new products or accept different currencies. You can change the software instructions to adjust the machine’s behavior without touching the hardware.

**Key Differences:**

**Hardwired Control Unit**: Like a basic machine with a set way of working. If you want to change something, you need to physically change the machine’s parts. It’s not easy to modify.

**Microprogrammed Control Unit**: Like a versatile machine that can be adjusted by updating the software. You can easily change how it works without changing the hardware, making it much more flexible.

### In Summary

A **hardwired control unit** is like a basic vending machine that’s designed to do a few specific tasks, while a **microprogrammed control unit** is like a more advanced machine that can be easily updated to do different things by changing its software. The hardwired one is less flexible, while the microprogrammed one can adapt to different tasks more easily.