

Advanced Computer Architectures - Notes -  
v0.3.0

260236

March 2025

## Preface

Every theory section in these notes has been taken from two sources:

- Computer Architecture: A Quantitative Approach. [1]
- Pipelining slides. [2]
- Course slides. [3]

About:

 [GitHub repository](#)



These notes are an unofficial resource and shouldn't replace the course material or any other book on advanced computer architectures. It is not made for commercial purposes. I've made the following notes to help me improve my knowledge and maybe it can be helpful for everyone.

As I have highlighted, a student should choose the teacher's material or a book on the topic. These notes can only be a helpful material.

# Contents

<b>1</b>	<b>Pipelining</b>	<b>4</b>
1.1	Basic Concepts . . . . .	4
1.2	RISC-V Pipelining . . . . .	9
1.2.1	Pipelined execution of instructions . . . . .	12
1.2.2	Pipeline Implementation . . . . .	14
1.3	Problem of Pipeline Hazards . . . . .	16
1.3.1	RISC-V Optimized Pipeline . . . . .	18
1.3.2	Solutions to RAW Hazards . . . . .	21
1.4	Performance evaluation . . . . .	23
<b>2</b>	<b>Control Hazards and Branch Prediction</b>	<b>26</b>
2.1	Conditional Branch Instructions . . . . .	26
2.2	Control Hazards . . . . .	28
2.3	Naïve Solutions to Control Hazards . . . . .	30
2.4	Intro to Branch Prediction . . . . .	33
2.5	Static Branch Prediction . . . . .	34
2.5.1	Branch Always Not Taken . . . . .	35
2.5.2	Branch Always Taken . . . . .	37
2.5.3	Backward Taken Forward Not Taken (BTFNT) . . . . .	40
2.5.4	Profile-Driven Prediction . . . . .	41
2.5.5	Delayed Branch . . . . .	42
2.5.5.1	From Before . . . . .	44
2.5.5.2	From Target . . . . .	45
2.5.5.3	From Fall-Through . . . . .	47
2.5.5.4	From After . . . . .	49
2.6	Dynamic Branch Prediction . . . . .	50
2.6.1	1-bit Branch History Table . . . . .	52
2.6.2	2-bit Branch History Table . . . . .	54
2.6.3	Branch Target Buffer . . . . .	56
	<b>Index</b>	<b>59</b>

# 1 Pipelining

## 1.1 Basic Concepts

**Pipelining** is a fundamental **technique** in computer architecture aimed at **improving instruction throughput by overlapping the execution of multiple instructions**. The main idea behind pipelining is to **divide the execution of an instruction into distinct stages and process different instructions simultaneously in these stages**. This approach significantly increases the efficiency of instruction execution in modern processors.

### ✂ Understanding the RISC-V instruction set

Before delving into pipelining, it is essential to understand the **basic instruction set** of the RISC-V architecture. The instruction set consists of three major categories:

- **ALU Instructions (Arithmetic and Logic Operations)**

- Performs **addition between registers**:

```
1 add rd, rs1, rs2
```

Performs the addition between the values in registers `rs1` and `rs2` and stores the result in register `rd`.

$$rd \leftarrow rs1 + rs2$$

- Performs an **addition between a constant and a register**:

```
1 addi rd, rs1, 4
```

Performs the addition between the value in register `rs1` and the value 4 and stores the result in register `rd`.

$$rd \leftarrow rs1 + 4$$

- **Load/Store Instructions (Memory Operations)**

- **Loads** data from memory:

```
1 ld rd, offset(rs1)
```

Load data into register `rd` from an address formed by adding `rs1` to a signed `offset`.

$$rd \leftarrow M[rs1 + offset]$$

- **Stores** data in memory:

```
1 sd rs2, offset(rs1)
```

Store data from register `rs2` to an address formed by adding `rs1` to a signed `offset`.

$$M[rs1 + offset] \leftarrow rs2$$

- **Branching Instructions (Control Flow Management)**

- **Conditional Branches**

- \* Branch on equal:

```
1 beq rs1, rs2, L1
```

Branch to the label L1 if the value in register `rs1` is equal to the value in register `rs2`.

$$rs1 = rs2 \xRightarrow{\text{go to}} L1$$

- \* Branch on not equal:

```
1 bne rs1, rs2, L1
```

Branch to the label L1 if the value in register `rs1` is not equal to the value in register `rs2`.

$$rs1 \neq rs2 \xRightarrow{\text{go to}} L1$$

- **Unconditional Jumps**

- \* Jump to the label (jump):

```
1 j L1
```

Jump directly to the L1 label.

- \* Jump to the address stored in a register (jump register):

```
1 jr ra
```

Take the value in register `ra` and use it as the address to jump to. So it is assumed that `ra` contains an address.

These basic instructions will be used throughout the course.

### ≡ Execution phases in RISC-V

1. **IF (Instruction Fetch)**: The instruction is **fetched** from memory.
2. **ID (Instruction Decode)**: The instruction is **decoded**, and the **required registers are read**.
3. **EX (Execution)**: The instruction is **executed**, typically involving ALU operations.
4. **ME (Memory Access)**: For *load/store* instructions, this stage **reads from** or **writes to** memory.
5. **WB (Write Back)**: The **result is written back** to the destination register.

These five stages form the basis of the RISC-V pipeline.

### ✂ Implementation of the RISC-V Data Path

The **RISC-V Data Path** is a fundamental component of the processor's architecture, responsible for **executing instructions efficiently by coordinating various hardware units**. It defines how instructions flow through different stages of execution, interacting with memory, registers, and the Arithmetic Logic Unit (ALU).



Figure 1: Generic implementation of the RISC-V Data Path.



Figure 2: Specific implementation of the RISC-V Data Path.

Its fundamental components include:

- **Instruction Memory and Data Memory Separation.** RISC-V adopts a Harvard Architecture style, where the **Instruction Memory (IM)** and **Data Memory (DM)** are **separate**. This **prevents structural hazards** where instruction fetch and memory access could conflict in a single-memory design (this topic will be addressed later).
- **General-Purpose Register File (RF).** It consists of **32 registers**, each **32-bit** wide. The register file has **two read ports** and **one write port** to **support simultaneous read and write operations**. This setup allows faster register access, which is crucial for pipelined execution.
- **Program Counter (PC).** It holds the **address of the next instruction to be fetched**. Automatically increments during execution, typically by 4 bytes (for 32-bit instructions).

- **Arithmetic Logic Unit (ALU)**. Performs arithmetic and logical operations required by instructions. Inputs to the ALU come from registers or immediate values decoded from the instruction.

Other components that we can see in the general implementation of the RISC-V data path are:

- **Register File**. Stores temporary values used by instructions. Contains read ports (two registers can be read simultaneously for ALU operations) and write port (one register can be updated per clock cycle). The register file ensures high-speed execution of operations by reducing memory accesses.
- **Instruction Fetch (IF)**. The PC (Program Counter) retrieves the next instruction from Instruction Memory. The PC is incremented using an adder ( $PC + 4$ ), ensuring sequential instruction flow.
- **Instruction Decode (ID)**. Extracts opcode (determines the instruction type), source and destination registers, immediate values (if present). It reads values from the Register File based on instruction requirements.
- **Execution (EX)**. The ALU performs arithmetic and logical operations. A multiplexer (MUX) selects the second operand: a register value (for R-type instructions) or an immediate value (for I-type instructions like `addi`). The ALU result is forwarded to the next stage.
- **Memory Access (ME)**. Load (`ld`) and Store (`sd`) instructions interact with data memory. Data is either loaded from memory into a register or stored from a register into memory.
- **Write Back (WB)**. The result from ALU or memory is written back to the Register File.

**Example 1: Data Path Execution Example**

Let's consider a simple RISC-V **load instruction** (`ld x10, 40(x1)`) passing through the data path:

1. **IF Stage:** Instruction Fetch
  - PC  $\rightarrow$  Instruction Memory  $\rightarrow$  `ld x10, 40(x1)` fetched
  - PC updated to PC + 4
2. **ID Stage:** Instruction Decode
  - Registers read: `x1` (base register for memory access)
  - Immediate value extracted: 40
3. **EX Stage:** Execution
  - ALU calculates memory address: `x1 + 40`
4. **ME Stage:** Memory Access
  - Data is loaded from `M[x1 + 40]`
5. **WB Stage:** Write Back
  - Data stored in `x10`



## 1.2 RISC-V Pipelining

Pipelining is analogous to an assembly line in a factory. Instead of waiting for one instruction to complete before starting the next, **different instructions are executed simultaneously in different stages**.

If we consider a **non-pipelined execution**:

- Each instruction completes all five stages sequentially before the next instruction starts.
- If each instruction stage (IF stage, ID stage, etc.) takes, say, 2 nanoseconds, executing all stages of an instruction (IF, ID, EX, MEM, WB) takes 5 times 2 nanoseconds, then 10 nanoseconds. If we also want to execute 5 instructions, we need 10 nanoseconds times 5, then 50 nanoseconds!

Now, we consider a **pipelined execution**:

- Once the first instruction moves to the second stage, the next instruction starts in the first stage.
- The **pipeline becomes fully utilized** after the first few cycles, significantly **improving throughput**.

In an **ideal scenario**, a 5-stage pipeline should provide a speedup of  $5\times$  reducing execution time to:

$$(5 + 4) \times 2 \text{ ns} = 18 \text{ ns}$$

Where 5 are the steps of the first instruction, 5 are the steps of the last instruction, minus 1 because one step is already counted in the first instruction, so 4. Therefore, 9 is multiplied by 2 nanoseconds, the time taken by each stage. The result, 18 nanoseconds, is the time it takes the pipeline to execute 5 instructions in an ideal scenario.



Figure 3: Sequential vs. Pipelining execution.

### 🔧 Pipeline Performance and Speedup

The ideal performance improvement from pipelining is derived from the fact that **once the pipeline is filled, a new instruction completes every cycle**. The key performance metrics include:

- **Latency (Execution Time):** The total time to complete a single instruction does not change (sequential or pipeline).
- **Throughput (Instructions per Unit Time):** The number of completed instruction per unit time significantly increases.
- **Speedup Calculation**
  - A non-pipelined CPU with 5 execution cycles of 2 ns would take 10 ns per instruction.
  - A pipelined CPU with 5 stages of 2 ns results in 1 instruction completing every 2 ns.
  - This gives a theoretical speedup of  $5\times$  (ideal case).

Unfortunately, real-world implementations are subject to **pipeline hazards** that reduce efficiency.

### Understanding Pipelining Performance

Pipelining **improves instruction throughput** by allowing multiple instructions to be processed simultaneously in different stages. The **execution of an instruction is divided into 5 pipeline stages**:

1. IF (Instruction Fetch)
2. ID (Instruction Decode)
3. EX (Execution)
4. MEM (Memory Access)
5. WB (Write Back)

**Each stage takes 2 ns** (a *pipeline cycle*), meaning that an **instruction moves from one stage to the next every 2 ns**. Now, let's analyze the timeline of instruction execution:

Clock Cycle	IF	ID	EX	MEM	WB
1st (0-2 ns)	I1				
2nd (2-4 ns)	I2	I1			
3rd (4-6 ns)	I3	I2	I1		
4th (6-8 ns)	I4	I3	I2	I1	
5th (8-10 ns)	I5	I4	I3	I2	I1
6th (10-12 ns)	I6	I5	I4	I3	I2
7th (12-14 ns)	I7	I6	I5	I4	I3
8th (14-16 ns)	I8	I7	I6	I5	I4
9th (16-18 ns)	I9	I8	I7	I6	I5

Table 1: Pipelining timeline execution in an ideal case.

- The first instruction I1 takes 5 (clock) cycles to complete, i.e., 10 ns.
- However, starting from cycle 5, a new instruction finishes every cycle (every 2 ns).
- In a non-pipelined system, each instruction would take 10 ns (5 stages  $\times$  2 ns each).
- In a pipelined system, once the pipeline is full, an instruction completes every cycle (every 2 ns), achieving a  $5\times$  speedup compared to the non-pipelined execution.

Thus, after an initial “fill” time (1st, 2nd, 3rd, 4th), **a new instruction completes every 2 ns** (from 5th to 6th, I1 is finished; from 6th to 7th, I2 is finished; from 7th to 8th, I3 is finished), which is the duration of a single pipeline stage.

### 1.2.1 Pipelined execution of instructions

Each RISC-V instruction follows the five pipeline stages, but their interactions with the pipeline vary depending on the instruction type.

- **ALU Instructions** (e.g., `op $x, $y, $z`)

These are register-based operations that do not require memory access. Since there is no memory operation, the instruction **bypasses the ME stage**.

Stage	Description
<b>IF</b>	Fetch instruction from memory
<b>ID</b>	Decode instruction, read registers <code>\$y</code> and <code>\$z</code>
<b>EX</b>	Perform ALU operation ( <code>\$x = \$y + \$z</code> )
<b>ME</b>	No memory access (skipped)
<b>WB</b>	Write the ALU result to <code>\$x</code>

- **Load Instructions** (e.g., `lw $x, offset($y)`)

These instructions retrieve data from memory and store it in a register. The **memory access stage (ME)** is **crucial** here since the instruction must fetch data from memory.

Stage	Description
<b>IF</b>	Fetch instruction from memory
<b>ID</b>	Decode instruction, read base register <code>\$y</code>
<b>EX</b>	Compute memory address ( <code>\$y + offset</code> )
<b>ME</b>	Read data from memory
<b>WB</b>	Write data into destination register <code>\$x</code>

- **Store Instructions** (e.g., `sw $x, offset($y)`)

These instructions write data from a register into memory. Unlike `lw`, **store instructions do not require the WB stage**, as data is written directly into memory.

Stage	Description
<b>IF</b>	Fetch instruction from memory
<b>ID</b>	Decode instruction, read base register <code>\$y</code> and source register <code>\$x</code>
<b>EX</b>	Compute memory address ( <code>\$y + offset</code> )
<b>ME</b>	Write <code>\$x</code> into memory at the computed address
<b>WB</b>	No write-back stage (skipped)

- **Conditional Branches** (e.g., `beq $x, $y, offset`)

Branching introduces control hazards, as the pipeline needs to determine whether the branch is taken or not. Branches can introduce **stalls** due to dependencies on comparison results. This issue is typically mitigated using branch prediction.

Stage	Description
<b>IF</b>	Fetch instruction from memory
<b>ID</b>	Decode instruction, read registers <code>\$x</code> and <code>\$y</code>
<b>EX</b>	Compare <code>\$x</code> and <code>\$y</code> , compute target address
<b>ME</b>	No memory access (skipped)
<b>WB</b>	Update PC if branch is taken

This section breaks down how **different types of instructions behave in the pipeline**:

- ALU Instructions complete in the EX stage and do not use memory.
- Load Instructions require a memory access in the ME stage.
- Store Instructions write to memory instead of registers.
- Branch Instructions introduce control hazards because they may change the PC.

This means that **not all instructions behave the same** in the pipeline. Some instructions **skip certain stages** (e.g., stores do not have WB, ALU instructions skip ME), and some instructions **introduce potential problems** (e.g., branches can cause delays).

In conclusion, this section sets the stage for understanding pipeline stalls, forwarding, and hazard resolution techniques that are essential for designing high-performance processors.

### 1.2.2 Pipeline Implementation

The **RISC-V pipeline implementation** is designed to efficiently execute multiple instructions simultaneously, following the classical five-stage pipeline model:

1. IF (Instruction Fetch)
2. ID (Instruction Decode)
3. EX (Execution)
4. MEM (Memory Access)
5. WB (Write Back)

Each clock cycle, a new instruction enters the pipeline while previous instructions move to the next stage, allowing **five different instructions to be in execution at the same time**.



Figure 4: Structure of RISC-V pipeline.

### ✂ Execution Stages and Pipeline Modules

Each stage of the pipeline corresponds to a specific hardware module in the CPU. The RISC-V pipeline is composed of five primary hardware modules:

- **Instruction Fetch (IF) Module:** Fetches instructions from instruction memory and updates the PC.
- **Instruction Decode (ID) Module:** Decodes the fetched instruction and reads register values.
- **Execution (EX) Module:** Performs arithmetic/logical operations in the ALU or computes memory addresses.
- **Memory Access (MEM) Module:** Reads from or writes data to memory.

- **Write Back (WB) Module:** Writes the computed result back into the register file.

Each module is responsible for a specific **stage of execution**, and together they allow overlapping execution of multiple instructions.

### Pipeline Registers

To maintain separation between stages, **pipeline registers** are used (see Figure 4, page 14). These registers **store intermediate results and ensure proper communication between stages**:

- **IF/ID Register:** Holds fetched instruction and updated PC.
- **ID/EX Register:** Stores decoded instruction, read register values, and control signals.
- **EX/MEM Register:** Holds ALU results, destination register, and memory access information.
- **MEM/WB Register:** Stores memory data or ALU result to be written back to registers.

These pipeline registers **eliminate the need for re-fetching or re-decoding instructions** at each cycle, thus maintaining pipeline efficiency.

### 1.3 Problem of Pipeline Hazards

#### ⚠ Assumptions Made

Until now, our discussion on the RISC-V pipeline implementation has relied on several key assumptions to simplify the analysis and focus on fundamental concepts. These **assumptions help in understanding the ideal case of pipelining** before introducing complexities like hazards and optimizations.

1. All instructions are independent, so there are no dependencies between them.
2. No branches or jumps that change execution flow.

This is a theoretical idealization, because in real-world scenarios, **hazards** (structural, data, and control) **interfere with smooth execution**. Also, our second assumption ignores **branch instructions** (`beq`, `bne`, `j`, `jr`), which **cause control hazards** that require branch prediction or pipeline flushing.

#### ❓ What is a Pipeline Hazard?

Now that we have understood the ideal execution of a RISC-V pipeline, we must discuss pipeline hazards, which are obstacles that prevent the pipeline from operating at maximum efficiency.

A **Hazard** (or conflict) is a phenomenon that occurs when the **overlapping execution of instructions in the pipeline changes the expected order of instruction execution**. This can lead to incorrect results or the **need to insert stalls** (*pipeline bubbles*), reducing performance.

In other words, **hazards cause the next instruction in the pipeline to be delayed, which reduces the ideal throughput of 1 instruction per cycle**. Thus, hazards disrupt the smooth flow of instructions and require techniques to resolve them.

#### ≡ Classes of Pipeline Hazards

- **Structural Hazards:** Attempt to use the same resource from different instructions simultaneously.  
❓ **Example:** Single memory for both instruction and data access.
- **Data Hazards:** Attempt to use a result before it is ready.  
❓ **Example:** Instruction depending on a result of a previous instruction still in the pipeline.
- **Control Hazards:** Try to make a decision about the next statement to execute before the condition is evaluated.  
❓ **Example:** Conditional branch execution.



### ✓ Structural Hazards

A **structural hazard** occurs when multiple pipeline stages need to use the same hardware resource at the same time.

✓ **Structural Hazard cannot be applied to RISC-V**. This is a great thing, because thanks to the Harvard Architecture, **RISC-V uses separate instruction and data memory**, and this adoption avoids structural hazards.

### ? Control Hazards

A **control hazard** occurs when the pipeline does not know which instruction to fetch next, usually due to a branch or jump instruction. It is discussed in the following sections.

### ? Data Hazards

A **data hazard** occurs when an instruction depends on the result of a previous instruction that is still in the pipeline.

There are several types of data hazards:

- **RAW (Read After Write)**. An instruction tries to read a register before a previous instruction writes to it.

#### ? Example:

```
1 lw x2, 0(x1)
2 add x3, x2, x4
```

The add instruction needs x2, but x2 is still being fetched from memory in the MEM stage. Without hazard resolution, the processor would get the wrong value for x2.

- **WAR (Write After Read)**. A later instruction writes to a register before an earlier instruction reads it (rare in RISC).
- **WAW (Write After Write)**. Two instructions try to write to the same register in the wrong order.

### 1.3.1 RISC-V Optimized Pipeline

The **RISC-V optimized pipeline** introduces refinements that **reduce stalls, improve data access, and enhance instruction throughput**. The key optimizations include:

- ✓ **Efficient Register File Access.** In the standard RISC-V pipeline, register accesses **happen in two stages**:

- ID (Instruction Decode) → Reads register values.
- WB (Write Back) → Writes computed values back to registers.

#### 🔧 Optimization: Read and Write in the Same Cycle

- In the optimized pipeline:
  - Register **writing** happens in the **first half** of the **clock cycle**;
  - While register **reading** happens in the **second half** of the **clock cycle**.
- This means an instruction can write its result to a register in WB, and the **next instruction can immediately read** that value in ID during the **same cycle**.

This optimization **removes unnecessary stalls** when an instruction immediately depends on a result written in the previous cycle.



Figure 5: Visual **example** of an optimized pipeline; here the result (WB stage) of I1 is written in the first half of the clock cycle and the read (ID stage) of I4 is done in the second half of the clock cycle. So there is no hazards!

- ✓ **Forwarding (Bypassing) to Reduce Stalls.** **Forwarding** (also called **bypassing**) is a hardware technique that **eliminates stalls by providing ALU results directly to dependent instructions without waiting for the WB stage**. It is a possible solution for Data Hazards.

🔧 **Forwarding Paths:** To support forwarding, the **pipeline includes extra paths** that allow instructions to fetch values from intermediate pipeline registers instead of waiting for WB.

- **EX/EX Path.** Allows ALU results to be forwarded from **EX stage output to the next EX stage input**. Used when an **instruction depends on an arithmetic result of the previous instruction**.

#### Example 2: EX/EX Forwarding

```

1 sub x2, x1, x3    # Compute x2 = x1 - x3
2 and x12, x2, x5   # Use x2 immediately

```

Cycle	sub x2, x1, x3	and x12, x2, x5
1	IF	
2	ID	IF
3	EX	ID
4	MEM	<i>Stall</i>
5	WB	<i>Stall</i>
6		EX
7		MEM
8		WB

The **and** instruction **must wait until WB writes x2 to the register file**. Two stall cycles are introduced and this wastes execution time.

Instead of waiting for WB, we forward the ALU result from the EX stage of **sub** directly to the EX stage of **and**.

Cycle	sub x2, x1, x3	and x12, x2, x5
1	IF	
2	ID	IF
3	EX	ID
4	MEM	EX (forwarded x2 from EX)
5	WB	MEM
6		WB

In cycle 4, **and x12, x2, x5** gets the forwarded x2 from the EX stage of **sub**, **removing stalls**.

This is EX/EX forwarding, taking ALU results from one EX stage directly into the next EX stage.

- **MEM/EX Path.** Forwards the ALU result from MEM stage to EX stage. Used when an instruction depends on an ALU operation two cycles before.



Figure 6: Example of MEM/EX path.

- **MEM/MEM Path.** Forwarding directly between two memory operations in the MEM stage. It removes stalls in Load/Store dependencies.



Figure 7: Example of MEM/MEM path.



Figure 8: Implementation of RISC-V with Forwarding Unit.

### 1.3.2 Solutions to RAW Hazards

To handle RAW hazards, we can use both **static (compile-time)** and **dynamic (hardware-based)** techniques. These include:

- **Static (compile-time):**
  - ✓ **nop insertion:** compiler adds empty instructions to delay execution.
  - ✓ **Instruction Scheduling:** compiler reorders instructions to avoid conflicts.
- **Dynamic (hardware-based):**
  - ✓ **Pipeline Stalling (bubbles):** inserts delay cycles when necessary.
  - ✓ **Forwarding (bypassing):** uses intermediate values from the pipeline instead of waiting.

#### ✓ *Static (compile-time) solution: Inserting nops (naïve)*

One simple way to handle RAW hazards is to **insert nop instructions manually between dependent instructions**. This gives the pipeline time to complete the write-back of the needed value.

Key takeaway of inserting nops:

- ✗ **Simple**, but inefficient because it wastes clock cycles. It should be the very last solution considered.
- ✗ Instead of using useful instructions, the **processor waits**, reducing performance.

#### Example 3: nop insertion

```
1 sub x2, x1, x3
2 nop           # Delay slot (bubble)
3 and x12, x2, x5 # Now x2 is ready
```

#### ✓ *Static (compile-time) solution: Instruction Scheduling*

A more efficient technique is **instruction reordering**, also known as **compiler scheduling**. The **compiler reorders instructions to avoid data hazards without inserting nops**.

Key takeaway of instruction scheduling:

- Instruction reordering is a **compiler optimization**.
- ✓ It works well **if independent instructions are available**.
- ✗ In some cases, no independent instructions exist, so **stalling or forwarding is needed**.

### Example 4: Instruction Scheduling

```

1 sub x2, x1, x3
2 # Independent instruction
3 # (can execute while sub is completing)
4 add x4, x10, x11
5 and x12, x2, x5 # Now x2 is ready

```

Instead of a `nop`, we insert `add x4, x10, x11`, which does not depend on `x2`. This keeps the pipeline utilized while avoiding RAW hazards.

### ✓ *Dynamic (hardware-based): Pipeline Stalling (Bubble Insertion)*

When no independent instructions can be scheduled, the **hardware must stall the pipeline** by inserting a **bubble (stall cycle)**.

Key takeaway of pipeline stalling:

- ✗ Stalling is simple but **reduces performance** (pipeline sits idle).
- ✓ We **prefer forwarding** (next solution) instead of stalling.



Figure 9: Example of inserting stalls.

### ✓ *Dynamic (hardware-based): Forwarding (Bypassing)*

Forwarding is an optimized hardware technique that avoids pipeline stalls by **directly passing results between pipeline registers**. The entire implementation has already been explained on page 18.

Key takeaway of forwarding:

- ✓ **Forwarding is the best solution** because it eliminates stalls and maximizes performance.
- It **requires extra hardware** (MUX and control logic), but it significantly improves throughput.

## 1.4 Performance evaluation

Evaluating the performance of a pipelined processor is essential to understanding the impact of stalls, hazards, and instruction throughput. In an **ideal scenario**, a **pipeline achieves one instruction per cycle** ( $CPI = 1$ ), but real-world execution includes pipeline stalls, which degrade performance.

Several **key metrics** are used to evaluate the efficiency of pipelining:

- **Instruction Count (IC)**. Represents the **total number of instructions executed**. Used as a basis for performance calculations.
- **Clocks Per Instruction (CPI)**. CPI measures the **average number of clock cycles required to execute one instruction**. **Ideal CPI for a pipelined processor is 1**, but hazards and stalls increase CPI.

$$CPI = \frac{\text{Total Clock Cycles}}{\text{Instruction Count (IC)}} \quad (1)$$

Where the total clock cycles is:

$$\text{Total Clock Cycles} = IC + 4 + \text{Stall Cycles} \quad (2)$$

Where  $+4$  is the **fill time of the first instruction**. The  $+4$  represents the initial pipeline fill time required before the pipeline reaches full execution throughput.

### Example 5: Why is the Pipeline Startup Overhead $+4$ ?

A 5-stage pipeline (IF, ID, EX, MEM, WB) requires 4 extra cycles before the first instruction completes. Consider the following scenario:

Clock Cycle	IF	ID	EX	MEM	WB
1	I1				
2	I2	I1			
3	I3	I2	I1		
4	I4	I3	I2	I1	
5	I5	I4	I3	I2	I1
6	I6	I5	I4	I3	I2
7	I7	I6	I5	I4	I3

The first instruction (I1) requires 5 cycles to complete. The next instruction (I2) completes in cycle 6, and so on. After the first 5 cycles, the **pipeline reaches steady state**, completing 1 instruction per cycle (ideal scenario, no hazards).

- **Instruction Per Clock (IPC)**. IPC is the inverse of CPI:

$$IPC = \frac{1}{CPI} \quad (3)$$

Measures **how many instructions complete per clock cycle**.

- **Millions of Instructions Per Second (MIPS)**. Evaluates processor speed in terms of **millions of instructions executed per second**:

$$\text{MIPS} = \frac{f_{\text{clock}}}{\text{CPI} \times 10^6} \quad (4)$$

Higher **clock frequency** ( $f_{\text{clock}}$ ) and lower **CPI** result in better MIPS.

#### Example 6: Performance Calculation

Given:

- Instruction Count (IC) = 5
- Stall Cycles = 2
- Clock Frequency = 500MHz

Metrics:

- Total Clock Cycles:

$$\text{Clock Cycles} = \text{IC} + \text{Stall Cycles} + 4 = 5 + 2 + 4 = 11$$

- CPI Calculation:

$$\text{CPI} = \frac{11}{5} = 2.2$$

- MIPS Calculation:

$$\text{MIPS} = \frac{500 \text{ MHz}}{2.2 \times 10^6} = 227$$

Without stalls, CPI would be 1 (ideal pipeline). But stalls increase CPI, reducing MIPS and overall efficiency.



## 🔗 Performance in Loops and Asymptotic Analysis

When evaluating **loops** or **long-running programs**, we use asymptotic performance metrics.

For a loop with:

- $m$  **instructions** per iteration.
- $k$  **stall** cycles per iteration.
- $n$  **iterations**.

We have:

- **Clock Cycles per Iteration:**

$$\text{Clock Cycles per Iteration} = m + k + 4 \quad (5)$$

- **CPI per Iteration:**

$$\text{CPI}_{\text{iter}} = \frac{(m + k + 4)}{m} \quad (6)$$

- **MIPS per Iteration:**

$$\text{MIPS}_{\text{iter}} = \frac{f_{\text{clock}}}{\text{CPI}_{\text{iter}} \times 10^6} \quad (7)$$

For **large**  $n$ , the impact of pipeline startup delay (+4 cycles) is reduced:

- **CPI per Iteration:**

$$\begin{aligned} \text{CPI}_{\text{AS}} &= \lim_{n \rightarrow \infty} \frac{(\text{IC}_{\text{AS}} + \text{Stall Cycles}_{\text{AS}} + 4)}{\text{IC}_{\text{AS}}} \\ &= \lim_{n \rightarrow \infty} \frac{(m \times n + k \times n + 4)}{(m \times n)} \\ &= \frac{(m + k)}{m} \end{aligned} \quad (8)$$

- **Millions of Instructions Per Second (MIPS):**

$$\text{MIPS}_{\text{AS}} = \frac{f_{\text{clock}}}{\text{CPI}_{\text{AS}} \times 10^6} \quad (9)$$

For **large programs**, startup stalls become negligible, and **performance depends mainly on stall cycles per iteration**. **Minimizing**  $k$  (**stalls per iteration**) is crucial to achieving high efficiency.

## 🔗 Why CPI is Greater than 1 in Real Pipelines

Even with an **optimized pipeline**, **real-world execution is affected by hazards**. Thus, actual CPI is always greater than 1, even in well-optimized designs.

## 2 Control Hazards and Branch Prediction

### 2.1 Conditional Branch Instructions

In pipelined processor architectures, control flow is not always linear, and **decisions about the next instruction to execute are often dependent on certain conditions**. This introduces the necessity for **conditional branch instructions**, particularly relevant in RISC-V architectures, where typical instructions include:

- **beq** (branch if equal): `beq rs1, rs2, L1`  
Transfers execution to the label L1 if the contents of registers `rs1` and `rs2` are equal.
- **bne** (branch if not equal): `bne rs1, rs2, L1`  
Transfers control to L1 if `rs1` and `rs2` hold different values.

These branch instructions are essential in implementing control structures such as loops, conditionals (`if/else`), and function returns.

At the hardware level, the **Branch Target Address (BTA)** plays a central role. This **address** represents **where the processor should continue execution if the branch is taken** (i.e., if the condition specified by the branch instruction evaluates as true). When the condition is satisfied, the processor **updates the Program Counter (PC)** with the BTA, thus redirecting the flow of instruction fetch.

Conversely, if the **condition is not satisfied**, the **branch is not taken**, and the processor continues **sequential execution**. In RISC-V, since instructions are generally 32 bits (4 bytes) long, the next instruction is fetched from `PC + 4`.

Understanding whether a branch is taken or not is crucial for instruction fetch in pipelined architectures. **Mispredicting** this can introduce **performance penalties**, which are addressed in detail through the study of control hazards and branch prediction techniques in later sections.

### ✂ Execution of Branches in Pipelined Architectures

When executing a **branch instruction**, such as `beq rx, ry, L1`, the processor must **compare two registers** (`rx`, `ry`) to determine whether the branch should be **taken** (i.e., jump to label L1) or **not taken** (continue sequentially). In **RISC-V**, the **Branch Outcome** (Taken/Not Taken) and **Branch Target Address (BTA)** are **calculated during the EX stage** (when the ALU performs arithmetic and logical operations):

- EX Stage:
  - Compare `rx` and `ry` using the ALU.
  - Compute `PC + offset` to obtain the BTA.

- ME Stage:
  - Based on the comparison, update the PC to either  $PC + 4$  (if not taken) or  $PC + \text{offset}$  (if taken).

MIPS follows a similar structure but emphasizes that **branch decisions are finalized at the end of the EX stage**, with the **PC update happening at the ME stage**. This introduces a **delay in resolving the branch**, which becomes critical for understanding control hazards.

#### Implication for IF

Since new instructions are fetched every clock cycle, the **processor needs to decide early which instruction to fetch next**. However, with branches, this decision is **not immediately clear because the branch condition hasn't yet been evaluated**. The Program Counter (PC) **cannot be updated correctly until the branch outcome is known**, leading to potential pipeline stalls or incorrect instruction fetches.

## 2.2 Control Hazards

In a pipelined architecture, one of the primary challenges in achieving high performance is dealing with **Control Hazards**, also known as **branch hazards**. These **arise due to the presence of conditional branch instructions**, where the **processor must decide the next instruction to fetch before knowing whether the branch will be taken**.

### ? What really causes Control Hazards?

To sustain the pipeline and **avoid idle stages**, a processor needs to **fetch one instruction per clock cycle**. However, with branch instructions, this becomes problematic because the **branch decision** (branch outcome) and the **branch target address** (BTA) are **not immediately available**. Specifically, **during the IF stage**, when the next instruction is fetched, the **processor still does not know whether the branch will be taken or not**, because this information typically becomes available later in the pipeline.

This leads to uncertainty:

- ? Which instruction should be fetched after a branch?
- ? Should it be sequential instruction ( $PC + 4$ ) or the instruction at the BTA?

If the processor fetches the wrong instruction, it might need to discard or “flush” it later, **wasting valuable cycles**. Alternatively, the **processor may stall**, delaying the fetching of any instruction until the branch decision is known, which also hurts performance.

The key issue is this: in a pipeline, the **instruction stream needs to continue**, but the **correct path is unclear** until the branch condition is evaluated. Thus:

1. Either **wrong-path instructions are fetched** (requiring flushing later)
2. Or **pipeline stalls** are introduced (causing delay and loss of ideal speedup)

### Key Takeaways: Control Hazards

- **Definition:** Pipeline hazard due to **uncertainty in branch outcome** during instruction fetch.
- **Cause:** The **branch condition is unresolved** when the next instruction must be fetched (IF stage).
- **Instructions Involved:** Conditional branches (beq, bne) and jumps, all **instructions modifying the PC**.
- **Pipeline Timing Conflict:** BO and BTA known only in EX or later, but instruction fetch **must occur every cycle**.

- **Main Problem:** Cannot decide whether to fetch next sequential instruction ( $PC + 4$ ) or BTA.
- **Possible Outcomes:**
  - Stall: Delay fetch until branch resolved.
  - Fetch wrong instruction  $\rightarrow$  flush.
- **Performance Impact:** Loss of ideal pipelining speedup; reduced throughput due to stalls or wasted fetches.
- **Goal of Solutions:** Mitigate stalls and improve fetch accuracy through early evaluation or prediction.

## 2.3 Naïve Solutions to Control Hazards

To manage control hazards in a simple and reliable way, one of the **earliest approaches** developed was the **conservative solution** of introducing **branch stalls**. The idea is straightforward: when a branch instruction enters the pipeline, the **processor stalls the pipeline until the branch decision is known** and the correct next instruction can be safely fetched.

### ? How does the conservative solution work?

In the typical 5-stage pipeline, the Branch Outcome (i.e., whether the branch is taken or not) and the Branch Target Address (BTA) are usually resolved **at the end of the EX stage**. However, the **Program Counter (PC)** is actually **updated at the end of the ME stage**. This introduces a **delay of multiple cycles** between the branch instruction entering the pipeline and the point when the next instruction can be fetched with certainty.

To avoid fetching an incorrect instruction, the **processor simply pauses instruction fetch for 3 clock cycles** after the branch instruction enters the pipeline. These are called **stalls**, essentially empty cycles where no new instructions enter the pipeline. Once the PC is correctly updated based on the branch outcome, instruction fetch resumes.

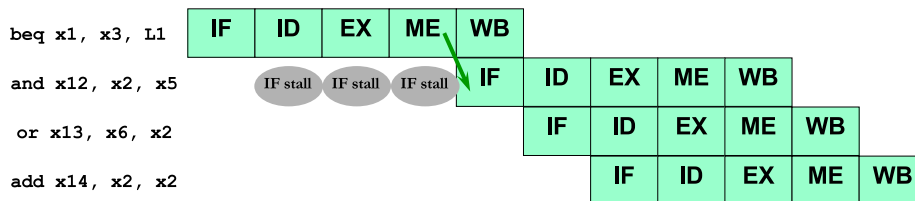


Figure 10: Example of stalls inserted in the pipeline to read the correct value after a branch condition.

### 🔊 Performance Impact

It's pretty obvious that if **each branch introduces a penalty of 3 cycles**, it will **significantly degrade performance**, especially in programs with frequent branching. Since pipelining aims to maximize instruction throughput, **this solution sacrifices speed for correctness**. In fact, it is **called conservative because it does not attempt to guess or speculate about the branch outcome**. Instead, it **waits for certainty, favoring reliability over efficiency**.

### ? Can optimized evaluation at the ID stage improve performance?

Although the conservative solution degrades performance, it can be relatively optimized thanks to hardware optimization. Processor designers have introduced **hardware optimizations** that allow the **branch outcome (BO)** and the **branch target address (BTA)** to be **computed earlier** in the pipeline. Specifically, these computations can be **moved from the EX stage to the**

**ID stage** (during the decoding phase). This optimization is often referred to as **Early Branch Evaluation**.

### ✂ How Early Branch Evaluation Works

To achieve this, the **Instruction Decode (ID)** stage must be **enhanced with additional hardware logic** that allows it to:

1. **Compare register values** (rx and ry) to determine the branch condition.
2. **Compute the BTA** using the **sign-extended offset** from the instruction and the current PC value.
3. **Update the PC** as soon as BO and BTA are known.

By doing this, **both BO and BTA are available at the end of the ID stage**, allowing the processor to update the PC immediately and fetch the correct instruction in the following cycle.

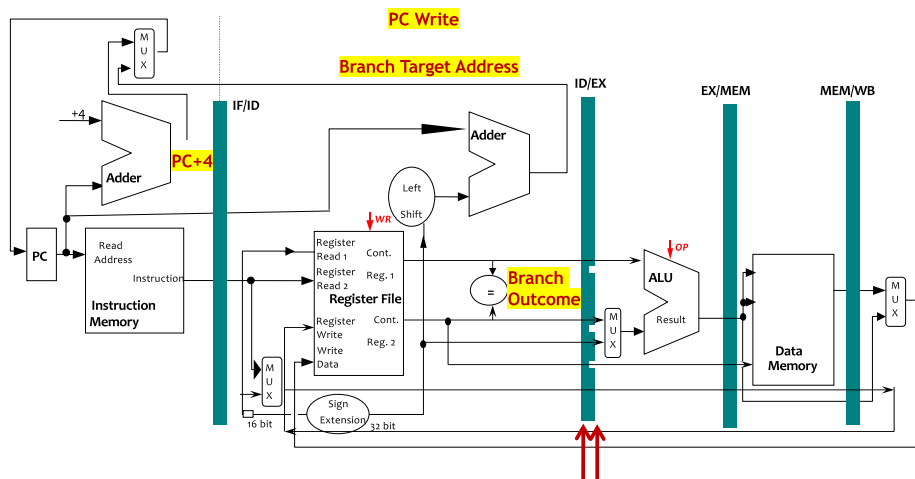


Figure 11: Hardware features modifications to allow early branch evaluation.

### ✂ Hardware Overhead: Complexity vs. Performance

This optimization requires **more complex hardware**, as the ID stage now includes:

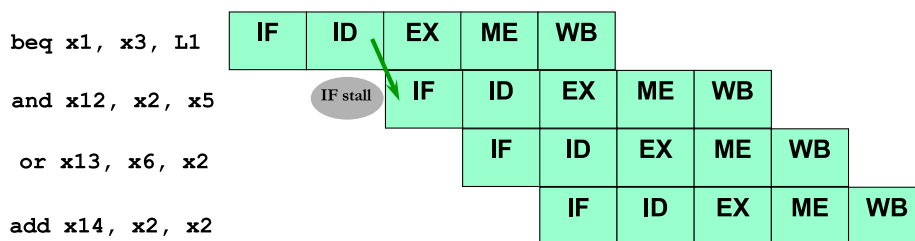
- **ALU logic** for comparison and addition.
- Additional **multiplexer and control signals** to direct the PC update.
- Expanded **data paths** for handling the offset and register values.

### ✔ Effect on Pipeline Execution

Let's consider an example. In a pipeline using early evaluation:

- The processor **only stalls for 1 cycle** after a branch instruction, as opposed to the 3 cycles required by the conservative approach.
- This **one-cycle stall** allows the processor to fetch the correct instruction **immediately after the branch** is resolved.

The following diagram illustrates that the instruction fetch after the branch is delayed by only one stall cycle, resulting in a **smaller performance hit**.



### 🔗 Conclusion

In summary, by anticipating branch evaluation at the ID stage, the processor reduces the branch penalty to 1 cycle per branch. This is a significant improvement over the 3-cycle stall of conservative stalling. While it introduces **hardware overhead**, it offers a **better balance between performance and correctness** and serves as a stepping stone toward even more advanced techniques, such as branch prediction.



## 2.4 Intro to Branch Prediction

In modern computer architectures, achieving high performance requires efficient instruction-level parallelism (ILP)<sup>1</sup>. However, one of the **major obstacles to ILP is the occurrence of branch hazards**, which happen when the processor encounters a branch instruction (e.g., `if`, `for`, `while`) and cannot immediately determine which instruction to execute next. To mitigate the performance loss caused by these hazards, branch prediction is employed.

**Branch Prediction** is essentially a **speculative execution technique** where the processor *guesses the outcome of a branch instruction*, whether the branch will be taken (control jumps to the branch target) or not taken (execution continues sequentially), **before the actual result is known**. Instead of stalling the pipeline and waiting for the branch condition to be resolved, the processor proceeds based on the predicted outcome. If the prediction:

- ✓ Is **correct**, performance is preserved.
- ✗ Is **wrong** (a **misprediction**), the incorrectly fetched instructions are flushed, and execution restarts at the correct address, causing a performance penalty.

### ≡ Branch Prediction categories

Branch prediction techniques are generally classified into two main categories:

- **Static Branch Prediction Techniques**. In this method, the **branch direction** (taken/untaken) is **decided at compile time and remains fixed during the program's execution**. Static prediction often relies on compiler heuristics or profiling data to guess likely outcomes. Since the behavior doesn't adapt to runtime changes, this **technique works best** when **branch outcomes are highly predictable and consistent** across executions.
- **Dynamic Branch Prediction Techniques**. Unlike static methods, dynamic prediction uses **hardware mechanisms to observe past branch behavior at runtime** and make predictions accordingly. The **prediction adapts to actual program execution**, making it more effective for applications with **complex or data-dependent control flow**. This method can dynamically switch its guess depending on the *branch history*.

It's important to note that in both static and dynamic techniques, the **processor must avoid updating its internal state** (registers, memory, etc.) **until the branch outcome is known with certainty**. This ensures speculative execution doesn't cause side effects in case of misprediction.

Additionally, **hybrid approaches** are possible, where static and dynamic predictions are combined to optimize performance further.

<sup>1</sup>**Instruction-Level Parallelism (ILP)**: A measure of **how many of the operations in a computer program can be performed simultaneously**. High ILP enables multiple instructions to be executed in parallel within a single processor cycle, exploiting the parallelism inherent in sequential instruction streams through techniques like pipelining, superscalar execution, and out-of-order execution.

## 2.5 Static Branch Prediction

Static branch prediction represents one of the **simplest approaches** to handling branch hazards. In this technique, the **prediction** regarding whether a branch will be taken or not is **made at compile time** and **remains unchanged throughout program execution**. This method **relies heavily on heuristics or compiler-generated hints**, which estimate the likely behavior of each branch without any consideration of the program's actual runtime behavior.

### ✓ When does static branch prediction work well?

This approach is particularly effective in scenarios where the **branch behavior is stable and highly predictable**, such as in embedded or domain-specific applications. In such cases, the overhead and complexity of dynamic prediction mechanisms may not justify the potential benefits, making static prediction a practical alternative.

### ⚠ RISC-V assumption

A key architectural note here is the assumption that we are working with a **RISC-V processor**, which is **optimized for early branch evaluation during the Instruction Decode (ID) stage** (see more in section 2.3, page 31). This means that in RISC-V, the decision to predict a branch direction occurs early in the pipeline, minimizing the potential for instruction fetch delays if the prediction is accurate.

### 2.5.1 Branch Always Not Taken

The **Branch Always Not Taken** strategy is the simplest form of static branch prediction. It operates under the **assumption that the branch condition will never be satisfied**, i.e., the control flow of the program will **continue sequentially as if the branch is not taken**. As a result, instructions immediately following the branch in program order are fetched and executed **without any need to determine or access the Branch Target Address (BTA)**.

#### ✔ When is Branch Always Not Taken effective?

This approach is especially effective for **certain control flow patterns**, such as **if-then-else** structures where the **then** clause is more likely to be executed than the **else** clause. For example:

```

1 z = x + y;
2 if (z > 0)
3     w = x;
4 else
5     w = y;
```

Assuming **z** is typically positive, the branch is not taken because execution proceeds sequentially to **w = x**. This makes predict-not-taken a suitable and effective default strategy for such cases.

#### ✂ Implementation Details

The prediction is made **at the end of the Instruction Fetch (IF) stage, without calculating or knowing the BTA** (since the branch is always not taken and the next instruction to execute is the **PC + 4**, as always). This makes the approach **lightweight and efficient**.

#### ⚠ Misprediction Case

If the actual **branch outcome (BO)** evaluated during the Instruction Decode (ID) stage is **not taken**, then the **prediction is correct**, and **no penalty cycles** are incurred. The pipeline proceeds as planned.

Otherwise, if the actual **branch outcome (BO)** turns out to be **taken**, then the **prediction was incorrect**, leading to a **misprediction penalty**. The processor must:

1. **Flush** the fetched instruction(s) after the branch (turned into NOPs).
2. **Fetch the instruction** at the Branch Target Address (**BTA**) and **restart execution from there**.

This results in a **one-cycle branch penalty**, which is minimal but still affects performance.

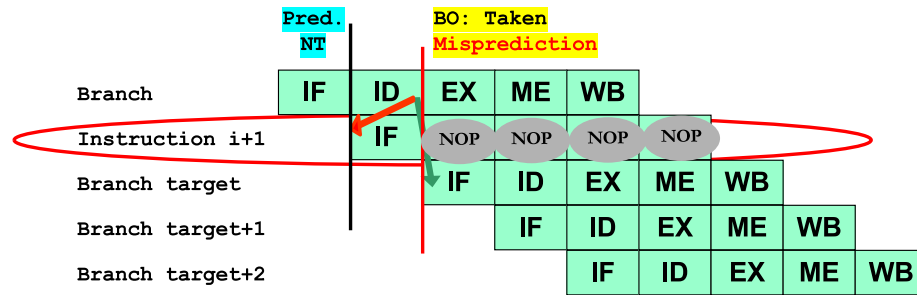


Figure 12: Branch Always Not Taken techniques failed and the processor must flush instruction i+1 and restart execution from the BTA.

### 2.5.2 Branch Always Taken

This approach represents the dual case of the previous technique (page 35): it assumes that **every branch will be taken**, meaning the **control flow will jump to the branch target address** rather than continue sequentially. This method is especially **useful for backward branches**, which occur in **loops** such as **for**, **while**, and **do-while**, since these branches are typically **taken repeatedly during loop iterations**.

#### ❓ Implementation Challenge

Unlike the not-taken strategy, where the processor simply continues to  $PC + 4$ , the **taken strategy requires knowledge of the Branch Target Address (BTA)** at the Instruction Fetch (IF) stage. This is **non-trivial** because:

- ❓ The **BTA depends on the branch instruction's target**, which typically requires decoding.
- ✓ To solve this, we introduce a **Branch Target Buffer (BTB)**, a special hardware structure.

#### ❓ What is BTB and why do we need it?

The **Branch Target Buffer (BTB)** is a **specialized cache** in the processor **designed to predict the target address** of a taken branch instruction **before the branch condition is actually resolved**.

In Branch Always Taken, we assume that the program will jump to a new address (the Branch Target Address, BTA). However, this **BTA is not immediately known during Instruction Fetch (IF)** because it typically requires decoding the branch instruction (Instruction Decode, ID). **To avoid delays**, the **BTB remembers past branch target addresses**, allowing the processor to quickly predict where to jump when encountering a branch.

#### ✂ How does BTB work? Quickest explanation

- **BTB Structure**, it is a kind of lookup table or cache where:
  - **Key**: **address of the branch instruction** (the PC value where the branch resides)
  - **Value**: Predicted Target Address (PTA), i.e., where to jump if the branch is taken
- **BTB Lookup**: when fetching a branch instruction, the processor simultaneously queries the BTB via the branch PC.
  - ✓ If a **match is found** (*cache hit*), the **BTB immediately provides the Predicted Target Address (PTA)**, and the **processor starts fetching from that address**, before knowing if the branch is actually taken.

- ✗ If a **no match** (*cache miss*), the processor might **default to sequential execution** ( $PC + 4$ ) or **wait for the BTA** to be calculated, which causes delay.

#### Example 1: BTB and Branch Always Taken technique

Let's say:

- A loop branch at address  $0x100$  typically jumps to  $0x80$ .
- The BTB stores:  $0x100 \rightarrow 0x80$  (key  $\rightarrow$  value).

When the branch at  $0x100$  is fetched again:

- The BTB predicts the next instruction will be at  $0x80$  (taken).
- The processor **starts fetching from  $0x80$ , without waiting** to evaluate the branch condition.

If it turns out the **branch was not taken**, the processor **flushes the incorrect fetch** from  $0x80$  and resumes at  $0x104$ .

#### ✔ Correct Prediction Path

If the **branch** is indeed **taken**, and the **BTB** correctly supplies the **BTA**, the processor proceeds **without penalty**. Execution continues from the **target address** just as expected.

#### ✗ Misprediction Case

If the **actual outcome** is **not taken**, the **prediction** is **incorrect**:

1. The Instruction Fetched (IF) from the **target address** is **flushed** (NOP).
2. The **processor must fetch the sequential instruction** at  $PC + 4$ .
3. **One-cycle penalty** incurred, similar to the not-taken misprediction case.

#### 🔗 When is this technique effective?

This method is **well-suited for loop constructs**, where **branches typically go backward** and are **taken with high probability**. For example:

- In a **do-while** loop, the branch is taken almost every time except the last iteration.
- Conversely, in forward branches like **if-then-else**, the branch is less likely to be taken, making this technique less effective.

This underscores that **branch direction** (forward or backward) **can influence the effectiveness of prediction strategies**.

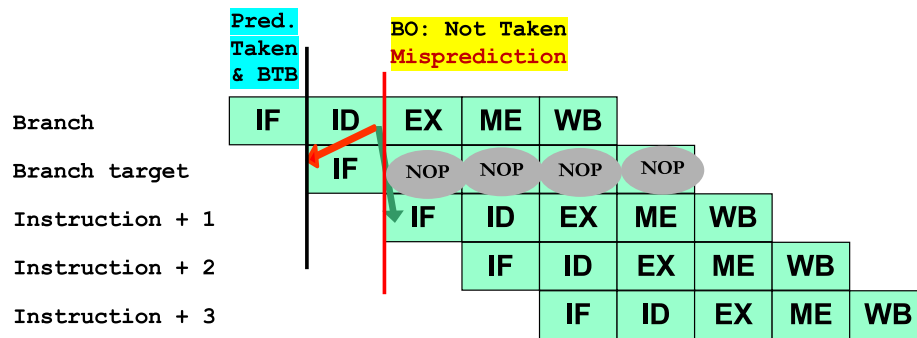


Figure 13: Branch Always Taken techniques failed and the processor must flush instruction  $i+1$  and restart execution from the BTA.

### 2.5.3 Backward Taken Forward Not Taken (BTFNT)

The **Backward Taken Forward Not Taken (BTFNT)** strategy represents a **refinement of static prediction** that uses a simple yet effective heuristic: the **direction of the branch**, whether it **jumps backward or forward** in memory, can be **used to predict its outcome**.

#### 📖 Prediction Rule

- **Backward-going branches** (i.e., branches where the target address is lower than the current PC) are predicted as **taken**.
  - These branches **often occur in loops**, where execution loops back to an earlier instruction (e.g., in **for**, **while**, or **do-while** constructs).
- **Forward-going branches** (i.e., target address is greater than the current PC) are predicted as **not taken**.
  - These branches typically correspond to **if-then-else constructs**, where the **else** path is **less probable** and **control usually proceeds sequentially**.

#### ❓ Why does this work?

The rationale behind BTFNT lies in **empirical observations**:

- **Loops** tend to execute **multiple times**, hence **backward branches are mostly taken**.
- **Conditional statements** often have **rarely taken else paths**, hence **forward branches are mostly not taken**.

#### ✅ Pros and ❌ Cons

- ✓ Simple to implement because BTFNT requires just a comparison of the **target address vs the current PC**:
  - $\text{target address} < \text{PC} \Rightarrow \text{predict taken}$
  - $\text{target address} > \text{PC} \Rightarrow \text{predict not taken}$

Also, better accuracy than uniform always-taken or always-not-taken, especially for mixed codebases.

- ❌ Not adaptive; fails for atypical control flows where direction doesn't align with expected behavior.



### 2.5.4 Profile-Driven Prediction

**Profile-Driven Prediction** is a static prediction technique that **uses empirical data from previous program executions** to guide the prediction of branch outcomes. Rather than relying solely on heuristics or branch direction, this method **leverages profiling to derive probabilistic insights** about how branches behave under typical conditions.

#### ✂ How does it work?

1. The **target application is executed multiple times beforehand, using diverse data sets** to simulate realistic execution scenarios.
2. During these early runs, the **behavior of each branch instruction is recorded**. Specifically, how often it was taken or not taken.
3. This **profiling produces statistics** for each branch, e.g., a pattern like:

T T T T T T T T NT NT NT

“Taken” is most probable.

4. Once the profiling is complete, the **compiler encodes a hint** directly into each branch instruction (e.g., in a dedicated **hint bit** in the instruction format):
  - 1: if the branch is **usually taken**.
  - 0: if the branch is **usually not taken**.

This enables the **processor to consult the hint during execution** and predict accordingly, without requiring runtime monitoring.

#### ✓ Advantages

- ✓ Offers **higher accuracy than heuristics alone**, especially for **applications with stable branch behavior**.
- ✓ **No runtime hardware cost**, since prediction decisions are guided by **static hints**.

#### ✂ Limitations

- ✂ **Static nature**: Predictions don’t adapt to runtime variability; changes in data patterns may invalidate the profile.
- ✂ Requires **extra compilation effort**: profiling and hint encoding add complexity to the build process.
- ✂ **Less effective** for programs with **input-dependent control flow**.

### 2.5.5 Delayed Branch

The **Delayed Branch** technique is a static scheduling approach where the **compiler** plays a central role in mitigating branch penalties. Unlike traditional branch prediction, which involves guessing the outcome of a branch, delayed branching **reorders instructions so that useful work is done regardless of the branch direction**.

#### Core Concept

When a **branch instruction** is executed, it typically introduces a **delay before the processor can determine where to fetch the next instruction**. During this delay (known as the **branch delay slot**), rather than letting the pipeline sit idle or fetch incorrect instructions, the **compiler schedules an independent instruction to execute in that slot**.

- The instruction in the **branch delay slot** is always executed, regardless of whether the branch is taken or not.
- This allows useful work to be completed during what would otherwise be a stall or pipeline bubble.

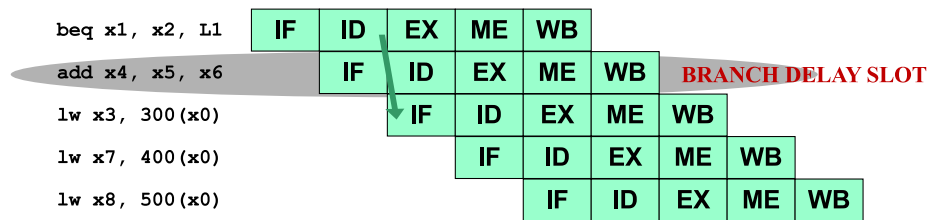


Figure 14: Example Scenario. In this case, the **add** instruction is scheduled after the branch, always executed, and does not affect the branch condition or outcomes.

#### Compiler Responsibility

A **critical task for the compiler** is to find a **valid and useful instruction** to place in the branch delay slot. The instruction must be:

- **Independent** from the branch decision.
- **Safe to execute** whether the branch is taken or not.

To guide this, the compiler can choose an instruction:

1. Section 2.5.5.1, page 44 - From **before** the branch.
2. Section 2.5.5.2, page 45 - From the **target** of the branch.
3. Section 2.5.5.3, page 47 - From the **fall-through path** (i.e., the sequential next instruction).
4. Section 2.5.5.4, page 49 - From **after** the branch.

We'll explore each of these four scheduling strategies step-by-step next.

Technique	Prediction Source	Complexity
<b>Always Not Taken</b>	Assume PC + 4	Very Low
<b>Always Taken</b>	Assume jump to BTA	Moderate (BTB)
<b>BTFNT</b>	Direction-based	Low
<b>Profile-Driven</b>	Prior run data	High (compile time)
<b>Delayed Branch</b>	Compiler scheduled	High (compiler)

---

Technique	Risk	Best for
<b>Always Not Taken</b>	Mispredict backward branches	<b>if-then-else</b>
<b>Always Taken</b>	Mispredict forward branches	Loops (backward branches)
<b>BTFNT</b>	Errors in irregular control flow	Mixed code (simple logic)
<b>Profile-Driven</b>	Profile mismatch	Stable behavior, performance-tuned
<b>Delayed Branch</b>	Wasted work if not efficient	RISC pipelines, e.g., MIPS processors

Table 2: Quick Comparison Table.

### 2.5.5.1 From Before

In the **“From Before”** strategy of delayed branch scheduling, the **compiler selects an instruction that appears *before* the branch** in program order and **moves it into the branch delay slot**. The selected **instruction must be independent of the branch decision** and safe to execute regardless of whether the branch is taken.

#### Key Characteristics

- The **instruction in the branch delay slot is always executed**.
- This instruction **will never be flushed**, since it is **guaranteed to execute** irrespective of the branch’s outcome.
- **After the delay slot, execution continues normally**, either to the branch target or the fall-through instruction, depending on whether the branch is taken.

#### Example 2: From Before

Original code:

```
1 add x1, x2, x3
2 beq x2, x4, L1
3 [delay slot, stall]
```

After scheduling:

```
1 beq x1, x2, x3
2 add x1, x2, x3      # delay slot filled
```

Here, the `add` is originally before the `beq` and has **no dependency** on the branch condition. It is safely moved into the delay slot.

#### ? Pipeline Behavior

- **Branch Not Taken**
  - The instruction in the delay slot is executed.
  - Execution continues sequentially with the next instruction after the branch.
- **Branch Taken**
  - The delay slot instruction still executes.
  - Execution jumps to the branch target after the delay slot.

The **instruction moved to the delay slot is always executed**.

#### ✓ Advantages

- ✓ **No need for instruction flushing**: the delay slot instruction is always valid.
- ✓ **Efficiency**: reuses existing instructions from earlier in the program to **hide the branch delay**.

### 2.5.5.2 From Target

In the **“From Target”** strategy, the **compiler schedules an instruction from the *branch target* into the delay slot**. This technique is **useful when the branch is likely to be taken**, as the delay slot instruction corresponds to what would naturally execute next in that control path.

#### Key Characteristics

- The delay slot contains an instruction from the branch target path.
- This strategy is typically **used when the branch is taken with high probability**, such as in **loops**.
- **Challenge**: If the branch is **not taken**, this instruction may be **invalid** and might have **to be flushed** (if mispredicted), or it must be **safe to execute even if not needed**.

#### Example 3: From Target

Original code:

```
1 sub x4, x5, x6      # target instruction
2 add x1, x2, x3      # branch instruction
3 if x1 == 0 then     # branch condition
4 [delay slot, stall]
```

After scheduling:

```
1 add x1, x2, x3      # if branch taken, go here!
2 if x1 == 0 then     # branch condition
3 sub x4, x5, x6      # delay slot filled
```

Here, the `sub` instruction from the branch target is moved into the delay slot. If the branch is taken, execution proceeds smoothly. If not, we either flush `sub` or ensure it causes no side effects.

#### Pipeline Behavior

- **Branch Taken**
  - The **delay slot** instruction is **part of the intended control flow**.
  - Execution continues with the **next instruction in the target path**.
- **Branch Not Taken**
  - Delay slot **may need to be flushed** (as it's not part of the sequential path), or must be **safe to execute anyway** (**no side effects or wasted computation**).

### ⚠ Instruction Duplication

When we **move an instruction from the branch target into the delay slot**, we **still need to keep it at the target location** because **other parts of the code might also jump there**.

Let's take an example to illustrate the problem. Let's say:

- We have **two branches that can jump** to label L1.
- Instruction X is the first instruction at L1.
- We move Instruction X into the delay slot of one branch, **but the other branch still needs to find instruction X at L1**.

Here's what happens:

```
1 Branch A → L1
2 Branch B → L1
3
4 L1:
5     Instruction X
6     Instruction Y
```

If Branch A decides to move Instruction X inside its delay slot, Branch B cannot see Instruction X anymore! Therefore, we have to keep Instruction X in two places:

1. In Branch A's delay slot
2. At Label L1 for Branch B

*Okay, and that should be a problem?* For three reasons:

- We've **duplicated** Instruction X.
- **More code = more memory = larger executable**.
- **Harder to maintain**: if we change Instruction X in one place, we might **forget to update the duplicate**.

### ✓ Best Use Case

**Loops**, particularly **do-while** constructs, where **backward branches are taken most of the time**.

### 2.5.5.3 From Fall-Through

In the **“From Fall-Through”** strategy, the **compiler selects an instruction that comes after the branch** in program order (i.e., from the fall-through path) and **moves it into the branch delay slot**. This method is **suitable when the branch is unlikely to be taken**, as execution will naturally continue sequentially.

#### Key Characteristics

- The fall-through path is taken when the *branch is not taken*.
- The **delay slot instruction** comes from this path, meaning it is **executed anyway if the branch is not taken**.
- If the **branch is taken**, the delay slot instruction is either:
  - **Flushed** (discarded), or
  - Must be **safe to execute** (no side effects), even though it becomes useless work.

#### Example 4: From Fall-Through

Original code:

```

1 add x1, x2, x3
2 if x1 == 0 then      # branch condition
3 [delay slot, stall]
4 or x7, x8, x9        # execute if branch is not taken
5 sub x4, x5, x6        # execute if branch is taken

```

After scheduling:

```

1 add x1, x2, x3
2 if x1 == 0 then      # branch condition
3 or x7, x8, x9        # delay slot filled
4 sub x4, x5, x6        # execute if branch is taken

```

Here, `or x7, x8, x9` is **moved into the delay slot** from the instruction that would **normally execute next** if the branch is **not taken**.

#### Pipeline Behavior

- **Branch Not Taken** (Mist Likely)
  - **Delay slot instruction** is correctly **executed**.
  - Execution proceeds sequentially.
- **Branch Taken**
  - **Delay slot instruction** is **not needed**.
  - It must be **flushed**, or **safe to execute** even though its result is discarded.

✔ When is this strategy used?

- ✔ When the **branch is not likely to be taken**.
- ✔ **Common in forward branches**, such as **if-then-else**, where **else** is rare.

Strategy	Delay Slot Instruction	Executed when branch
From Fall-Through	Instruction at <b>PC + 4</b> (next in sequence)	<b>Not Taken</b> (common case)
From Target	Instruction at BTA (label target)	<b>Taken</b> (common case)

Table 3: Comparison between “From Target” and “From Fall-Through”.



#### 2.5.5.4 From After

The “From After” scheduling technique is **rarely used because it is too complex to be practical**. However, in the **“From After”** strategy, the **instruction** scheduled in the **branch delay slot** is **taken** from a later point in the code, specifically, from **after the fall-through instruction**.

Let’s number the instructions to make it easy:

```

1 Instr A      # Before branch
2 Branch       # Branch condition
3 Instr B      # PC + 4 (fall-through)
4 Instr C      # After fall-through ← from after
5 Instr D

```

In “from after”, the compiler **moves Instr C into the delay slot**, even though Instr B (PC + 4) should come right after the branch in normal execution (if not taken).

#### ⚠ Why is this hard?

To safely move Instr C up into the delay slot, the **compiler must guarantee**:

##### 1. No Data Dependency Conflicts

- Instr C must not use or modify data that depends on Instr A, B, or the branch outcome. For example, if Instr C uses a value computed in Instr B, moving it before Instr B causes incorrect results.

##### 2. Safe if Executed Early

- Even if the branch is taken and Instr C should never execute, now it always executes in the delay slot.
- So Instr C must be safe to execute even when it’s not needed.
- We call this a *speculatively safe instruction*.

##### 3. No Control Flow Violation

- If Instr C should only run after a condition is met, moving it earlier might break program logic.

## 2.6 Dynamic Branch Prediction

While static branch prediction relies on fixed rules or compile-time knowledge, **dynamic branch prediction** aims to **learn and adapt during program execution**. It uses **hardware mechanisms to observe past branch behavior and predict future outcomes at runtime**.

### ♥ Core Idea

Dynamic prediction is based on a key assumption: **if a branch behaved a certain way in the past, it's likely to behave the same way again**. Therefore, instead of guessing statically, the **processor monitors each branch at runtime** and **uses past outcomes to inform future predictions**.

### 🔧 Hardware Components

Dynamic prediction relies on **two tightly-coupled hardware blocks**, both situated in the **Instruction Fetch (IF)** stage:

1. **Branch Outcome Predictor (BOP):**
  - **Predicts branch direction:** Taken (T) or Not Taken (NT).
  - **Based on runtime history** (past outcomes of this or other branches).
2. **Branch Target Buffer (BTB):**
  - **Predicts the target address** to jump to if the **branch is predicted taken**.
  - Returns the **Predicted Target Address (PTA)**<sup>2</sup>.
  - **Useful only when BOP predicts Taken**; irrelevant if Not Taken.

### ✂ How it works





During instruction fetch:

1. BOP predicts T/NT.
2. If Taken (T), BTB provides the PTA.
3. The processor fetches the next instruction accordingly.

---

<sup>2</sup>**Predicted Target Address (PTA):** The memory address that the processor predicts as the destination for a taken branch. If the branch is predicted taken, the Branch Target Buffer (BTB) provides the PTA so that instruction fetch can continue from this address without waiting for the branch condition to be resolved.

### ❓ Execution Scenarios

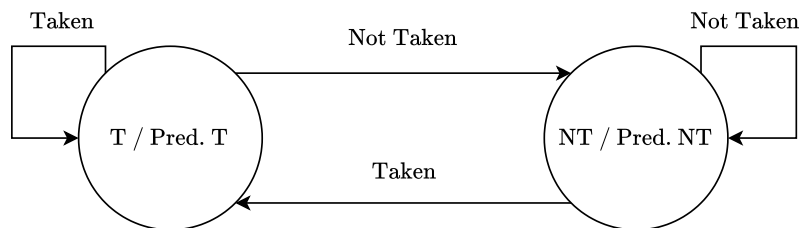
- **Prediction: Not Taken (PC + 4)**
  - If Branch Outcome = Not Taken  $\Rightarrow$   Correct prediction  $\Rightarrow$  No penalty.
  - If Branch Outcome = Taken  $\Rightarrow$   Misprediction:
    1. Flush next instruction (NOP)
    2. Fetch from BTA (to understand where to jump)
    3. One-cycle penalty
- **Prediction: Taken (BTB used)**
  - If Branch Outcome = Taken  $\Rightarrow$   Correct prediction  $\Rightarrow$  No penalty.
  - If Branch Outcome = Not Taken  $\Rightarrow$   Misprediction:
    1. Flush fetched target instruction (NOP) provided by BTB
    2. Fetch PC + 4 (next instruction sequentially)
    3. One-cycle penalty

Unlike static prediction, **dynamic prediction is adaptive**. If a branch changes its behavior at runtime, future predictions adjust accordingly.

### 2.6.1 1-bit Branch History Table

In general, the **Branch History Table (BHT)**, or **Branch Prediction Buffer**, is a **dynamic hardware structure** that **predicts branch outcomes based on recent behavior**. The **simplest version uses 1 bit per branch** to remember whether the branch was **recently taken or not taken**. For this reason, it is called a **1-bit Branch History Table (1-bit BHT)**. It operates at runtime and uses a **Final State Machine (FSM) with 1-bit history**:

- If last outcome was Taken  $\Rightarrow$  predict Taken.
- If last outcome was Not Taken  $\Rightarrow$  predict Not Taken.



#### ✂ How it works

- Each branch instruction's address (or part of it) indexes a table entry.
- That entry holds a single bit (T/NT) representing the last observed outcome.
- On next encounter:
  - Use the **bit to predict the outcome**.
  - After actual branch resolution:
    - \* **Correct prediction**  $\Rightarrow$  keep bit unchanged
    - \* **Incorrect prediction**  $\Rightarrow$  flip bit

#### 📖 Indexing the Table

- Use  $k$  lower bits (right side) of the branch's address as the index.

**⚠ No tags:** any branch with the same low-order bits shares the entry (can cause interference).

- $2^k$  entries total.

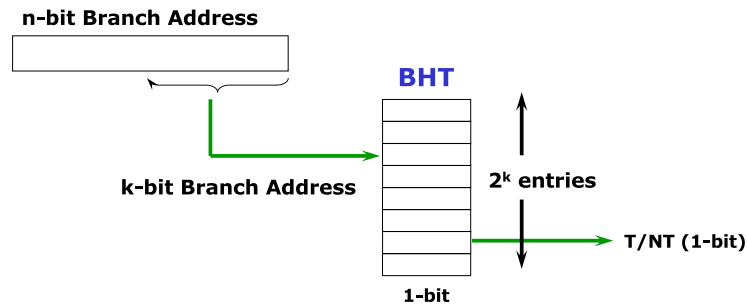


Figure 15: Visual representation of the 1-bit Branch History Table.

**Example 5: Accuracy Issue**

Consider a loop that executes 10 iterations. The expected behavior of the branch is:

T T T T T T T T T NT

Where the last is Not Taken because the code must exist from the loop and must continue (and not jump).

There are **two mispredictions**:

- **At the end:** Since iteration 9 is marked as Taken, the 10th iteration is predicted to be Taken, since the BHT contains Taken. This throws a misprediction because the branch result is Not Taken.
- **On the next time loop starts:** since the last iteration is stored in the BHT as Not Taken, the BHT has to flip the bit on the next time loop starts, and a misprediction occurs.

As a result, with 100% of 10 iterations, the BHT only catches 8 out of 10 iterations, the accuracy is 80%.

**✖ Shortcomings**

- ✖ **Flipping prediction after 1 misprediction causes instability**, especially in loops.
- ✖ **Conflict problem:** two different branches with same index overwrite each other's bit.

**✔ Partial Solutions**

To **reduce interference**:

- Increase table size (more  $k$  bits).
- Use hashing to mix address bits better.

### 2.6.2 2-bit Branch History Table

**2-bit Branch History Table (BHT)** is an **improvement over 1-bit BHT** designed to increase prediction stability, especially for loops, and reduce mispredictions.

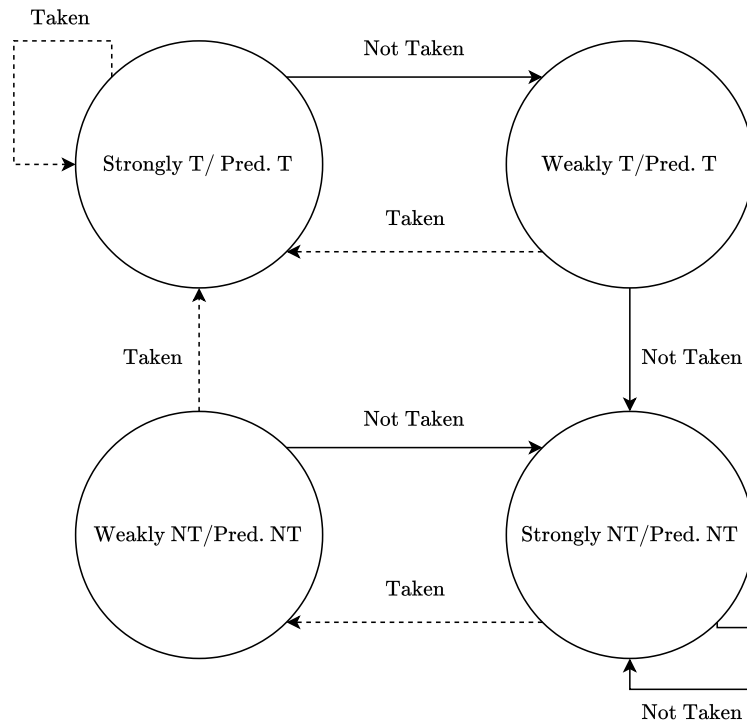
#### ✔ Why 2-bit? The Problem with 1-bit BHT

In loops, 1-bit BHT suffers from **flip-flopping**: **1 misprediction is enough to change the prediction**. This **causes two misprediction per loop**:

- Exiting the loop (NT mispredicted as T).
- Re-entering the loop (T mispredicted as NT).

The 2-bit BHT introduces a **4-state FSM** using **2 bits per entry**. It requires **2 consecutive mispredictions to change the predicted outcome**, thus adding stability. The FSM states are:

1. **Strongly Taken (ST)** → Predict Taken
2. **Weakly Taken (WT)** → Predict Taken
3. **Weakly Not Taken (WNT)** → Predict Not Taken
4. **Strongly Not Taken (SNT)** → Predict Not Taken



### ✔ Effect on Loops

Assume a loop with:

T T T T T T T T NT

- Exit NT causes 1 misprediction, but FSM moves from Strongly Taken to Weakly Taken. So the prediction remains Taken.
- Re-enter on the loop causes a Branch Outcome Taken, and the 2-bit BHT predicts correctly because it is on the WT state.

**Only 1 misprediction per loop**, improving accuracy to 90% (from 80%).

### ✔ Benefits

- **Improved accuracy** in loops and repetitive patterns.
- **Reduces misprediction penalty** in typical branch-heavy code.
- Balances **prediction stability** and **adaptability**.

### 2.6.3 Branch Target Buffer

The **Branch Target Buffer (BTB)** is a **specialized cache used to store target address of taken branches**. The stored Predicted Target Address (PTA) allows the processor to fetch instructions from the target without delay when a branch is predicted taken. The PTA is typically stored in PC-relative format (offset from current PC).

#### Core Idea

While the Branch History Table (BHT) predicts *whether* a branch will be taken, the Branch Target Buffer (BTB) predicts *where the program should go if the branch is taken*. The **BTB stores Predicted Target Addresses (PTAs)** for previously encountered branches and **enables fast redirection of control flow**.

#### How Is the BTB Structured?

- The BTB is designed as a **direct-mapped cache**:
  - The **address of the branch instruction is used to index** the BTB.
  - **Tags** are **used** for associative lookup to **confirm correctness** (i.e., ensure the indexed entry really belongs to the current branch)
- Components per **entry**:
  - **Tags**: Identifies the branch instruction.
  - **PTA**: The Predicted Target Address.
  - Often combined with **T/NT bits** from a **Branch History Table** (1-bit or 2-bit) for branch outcome prediction.

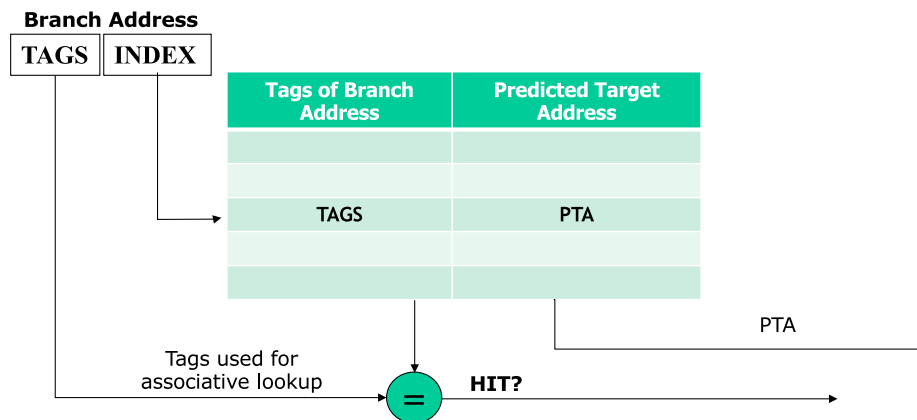


Figure 16: Branch Target Buffer without Branch Outcome Predictor.



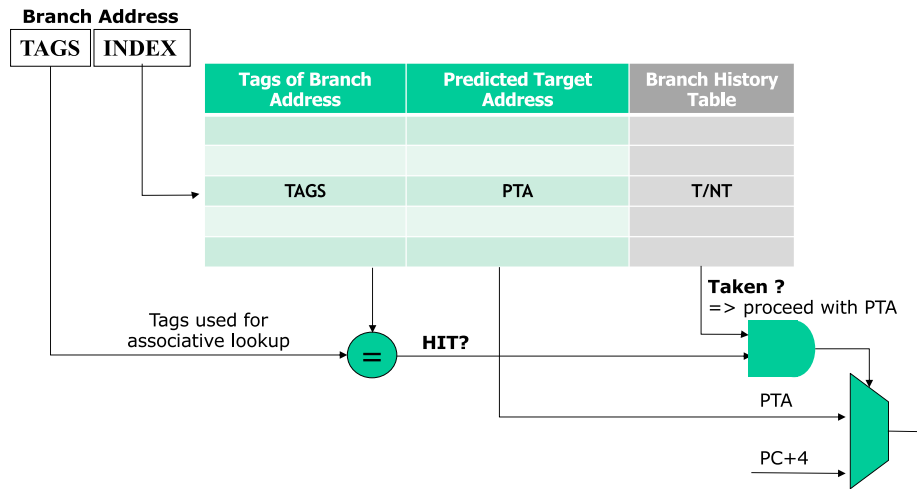


Figure 17: Branch Target Buffer with Branch Outcome Predictor.

### BTB in the Pipeline

It is placed in the Instruction Fetch (IF) phase. During fetch:

- The BTB is queried using the current PC (branch address).
- If hit and BHT predicts Taken, **fetch from PTA**.
- If miss or predict Not Taken, **continue** at PC + 4.

Prediction	BTB use	Action
Predict Not Taken	BTB <b>not used</b>	Fetch from PC + 4
Predict Taken (BTB hit)	BTB <b>used</b>	Fetch from PTA stored in BTB
Predict Taken (BTB miss)	BTB <b>miss</b>	Stall or default to PC + 4, then calculate BTA

Table 4: Summary of Behavior.

### Advantages

- ✓ **Eliminates delay** from calculating the **branch target address** manually.
- ✓ Enables speculative instruction fetch from correct target, **improving pipeline efficiency**.

## References

- [1] J.L. Hennessy and D.A. Patterson. *Computer Architecture: A Quantitative Approach*. ISSN. Elsevier Science, 2017.
- [2] Cristina Silvano. Lesson 1, pipelining. Slides from the HPC-E master's degree course on Politecnico di Milano, 2024.
- [3] Cristina Silvano. Advanced computer architecture. Slides from the HPC-E master's degree course on Politecnico di Milano, 2024-2025.

## Index

### Symbols

1-bit Branch History Table (1-bit BHT)	52
2-bit Branch History Table (BHT)	54

### B

Backward Taken Forward Not Taken (BTFNT)	40
Branch Always Not Taken	35
Branch Delay Slot	42
Branch History Table (BHT)	52
Branch Outcome	26
Branch Outcome Predictor (BOP)	50
Branch Prediction	33
Branch Prediction Buffer	52
Branch Target Address (BTA)	26
Branch Target Buffer (BTB)	37, 50, 56

### C

Clocks Per Instruction (CPI)	23
Compiler Scheduling	21
Control Hazards	16, 17, 28

### D

Data Hazards	16, 17
Delayed Branch	42
Delayed Branch Scheduling: From After	49
Delayed Branch Scheduling: From Before	44
Delayed Branch Scheduling: From Fall-Through	47
Delayed Branch Scheduling: From Target	45
Dynamic Branch Prediction Techniques	33

### E

Early Branch Evaluation	31
EX (Execution)	5
EX/EX Path	19

### F

Forwarding	18
------------	----

### H

Hazard	16
--------	----

### I

ID (Instruction Decode)	5
IF (Instruction Fetch)	5
Instruction Count (IC)	23
Instruction Per Clock (IPC)	23
Instruction-Level Parallelism (ILP)	33

### M

ME (Memory Access)	5
--------------------	---

MEM/EX Path	20
MEM/MEM Path	20
Millions of Instructions Per Second (MIPS)	24
Misprediction	33
<b>P</b>	
Pipeline Registers	15
Pipelining	4
Predicted Target Address (PTA)	50
Profile-Driven Prediction	41
<b>R</b>	
RAW (Read After Write)	17
RISC-V Data Path	6
<b>S</b>	
Static Branch Prediction Techniques	33
Strongly Not Taken (SNT)	54
Strongly Taken (ST)	54
Structural Hazards	16, 17
<b>W</b>	
WAR (Write After Read)	17
WAW (Write After Write)	17
WB (Write Back)	5
Weakly Not Taken (WNT)	54
Weakly Taken (WT)	54