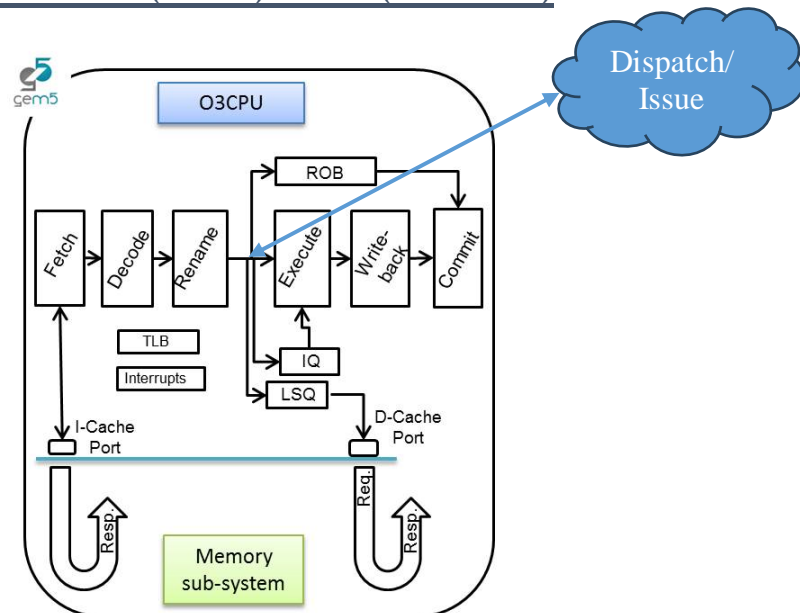


Expected delivery of **lab\_4.zip** must include:

- each configuration of the custom architecture (riscv\_o3\_custom.py) that you modify.
- This document with all the field compiled and in PDF form.

## Introduction and Background

### Simulating an Out-of-Order (OoO) CPU (O3CPU)



In this laboratory, you will be able to configure an OoO CPU by using a script called `riscv_o3_custom.py`. In a few words, the script configures an Out-of-Order (O3) processor based on the *DerivO3CPU*, a superscalar processor with a reduced number of features.

### Pipeline

The processor pipeline stages can be summarized as:

- **Fetch stage:** instructions are fetched from the instruction cache. The `fetchWidth` parameter sets the number of fetched instructions. This stage does branch prediction and branch target prediction.
- **Decode stage:** This stage decodes instructions and handles the execution of unconditional branches. The `decodeWidth` parameter sets the maximum number of instructions processed per clock cycle.
- **Rename stage:** As suggested by the name, registers are renamed, and the instruction is pushed to the IEW (Issue/Execute/Write Back) stage. It checks that the *Instruction Queue (IQ)*/*Load and Store Queue (LSQ)* can hold the new instruction. The maximum number of instructions processed per clock cycle is set by the `renameWidth` parameter.

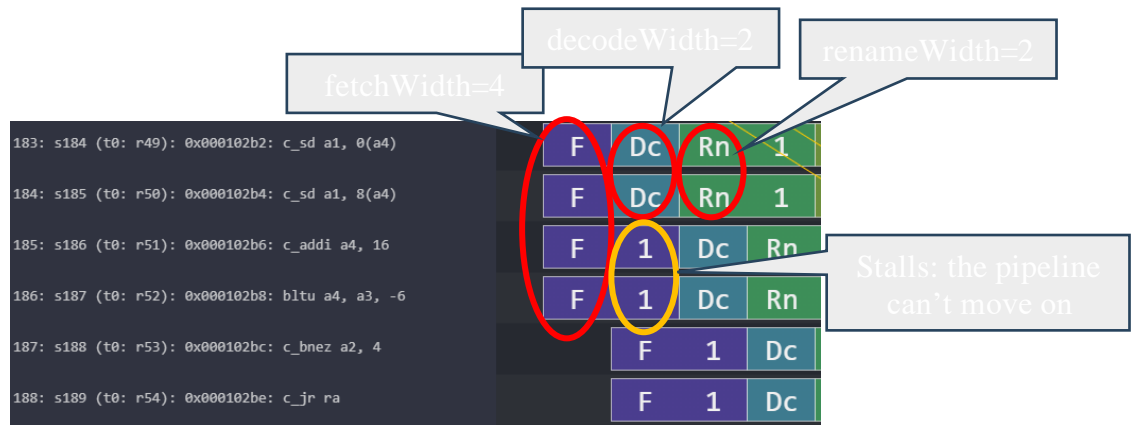


Figure 1: Understanding configurable OoO CPU parameters.

- **Dispatch stage:** instructions whose renamed operands are available are dispatched to functional units (FU). For loads and stores, they are dispatched to the Load/Store Queue (LSQ). The maximum number of instructions processed per clock cycle is set by the `dispatchWidth` parameter.
- **Issue stage:** The simulated processor has a single instruction queue from which all instructions are issued. Ordinarily, instructions are taken in-order from this queue. An instruction is issued if it does not have any dependency.
- **Execute stage:** the functional unit (FU) processes their instruction. Each functional unit can be configured with a different latency. Conditional branch mispredictions are identified here. The maximum number of instructions processed per clock cycle depends on the different functional units configured and their latencies.
- **Writeback stage:** it sends the result of the instruction to the reorder buffer (ROB). The maximum number of instructions processed per clock cycle is set by the `wbWidth` parameter.
- **Commit stage:** it processes the reorder buffer, freeing up reorder buffer entries. The maximum number of instructions processed per clock cycle is set by the `commitWidth` parameter. Commit is done in order.

In the event of a **branch misprediction**, trap, or other speculative execution event, "squashing" can occur at all stages of this pipeline. When a pending instruction is squashed, it is removed from the instruction queues, reorder buffers, requests to the instruction cache, etc.

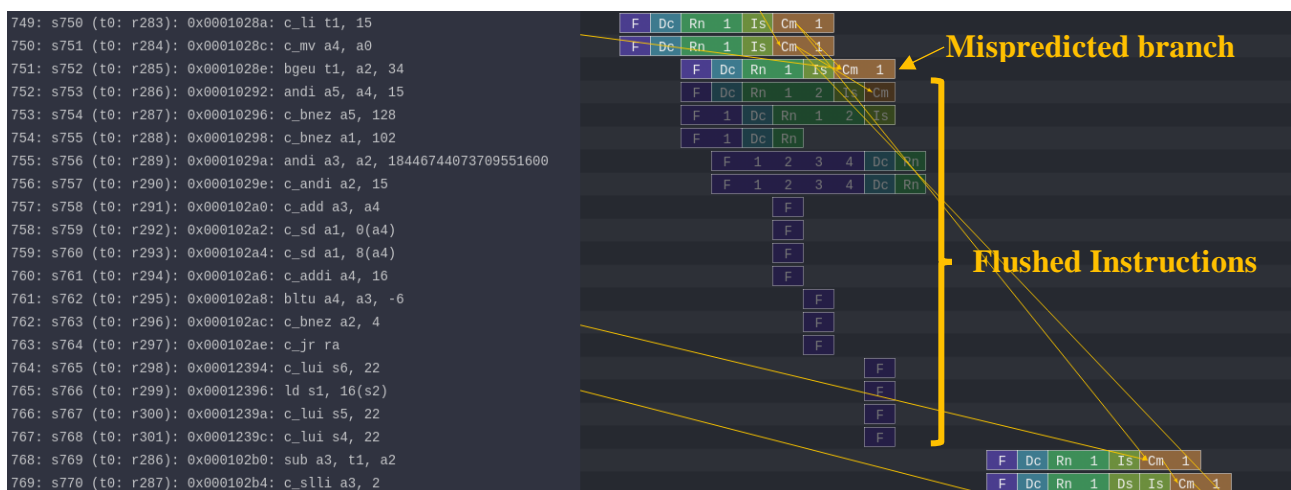


Figure 2: Example of a branch *misprediction* (transparent rows)

## Pipeline Resources

Additionally, it has the following structures:

- Branch predictor (BP)
  - Allows for selection between several branch predictors, including a local predictor, a global predictor, and a tournament predictor. Also has a branch target buffer (BTB) and a return address stack (RAS).
- Reorder buffer (ROB)
  - Holds instructions that have reached the back end. Handles squashing instructions and keep instructions in program order.
- Instruction queue (IQ)
  - Handles dependencies between instructions and scheduling ready instructions. Uses the **memory dependence predictor** to tell when memory operations are ready.
- Load-store queue (LSQ)
  - Holds loads and stores that have reached the back end. It hooks up to the d-cache and initiates accesses to the memory system once memory operations have been issued and executed. Also handles forwarding from stores to loads, replaying memory operations if the memory system is blocked, and detecting memory ordering violations.
- Functional units (FU)
  - Provides timing for instruction execution. Used to determine the latency of an instruction executing, as well as what instructions can issue each cycle.
  - **Floating point units, floating point registers**, and respective instructions are supported.

|  |   |    |    |    |    |    |    |    |    |   |   |
|--|---|----|----|----|----|----|----|----|----|---|---|
| 560: s561 (t0: r160): 0x00010106: fmv_w_x fa5, zero  | F | Dc | Rn | 1  | Is | 1  | 2  | 3  | Cm | 1 |   |
| 561: s562 (t0: r161): 0x0001010a: c_addi16sp sp, -64 | F | Dc | Rn | 1  | Is | Cm | 1  | 2  | 3  | 4 |   |
| 562: s563 (t0: r162): 0x0001010c: c_fsdsp fs0, 8(sp) | F | 1  | Dc | Rn | 1  | Is | Mc | 1  | 2  | 3 | 4 |
| 563: s564 (t0: r163): 0x0001010e: c_fsdsp fs1, 0(sp) | F | 1  | Dc | Rn | 1  | 2  | 3  | Is | Mc | 1 | 2 |

Figure 3: Pipeline example of FP instructions and FP registers

## Laboratory: hands-on

All the needed resources are at a GitHub repository:

[https://github.com/cad-polito-it/ase\\_riscv\\_gem5\\_sim](https://github.com/cad-polito-it/ase_riscv_gem5_sim)

To create your simulation environment:

For HTTPS clone:

```
~/my_gem5Dir$ git clone https://github.com/cad-polito-it/ase_riscv_gem5_sim.git
```

For SSH:

```
~/my_gem5Dir$ git clone git@github.com:cad-polito-it/ase_riscv_gem5_sim.git
```

The environment is configured to be executed on the **LABINF MACHINES**.

Follow the HOWTO instructions available on the GitHub Repository for simulating a program.

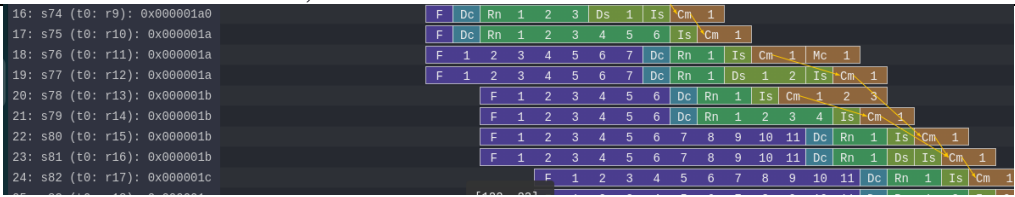
### Exercise 1:

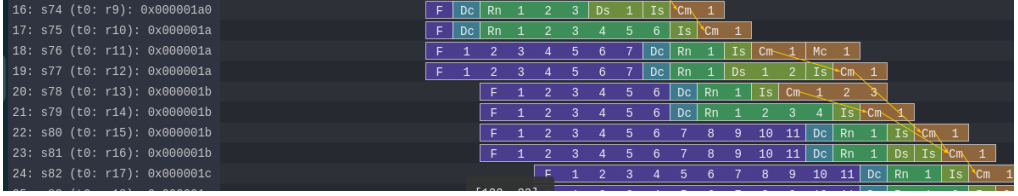
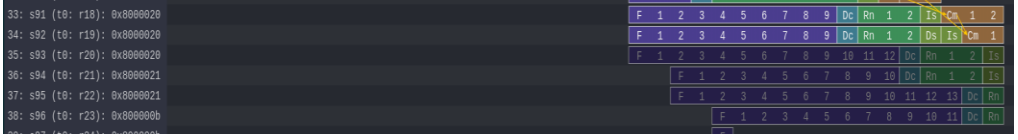
Simulate the benchmark *my\_c\_benchmark* (*main.c*) by using the gem5 simulator to obtain the *trace.out* file. Then, you can visualize the pipeline (i.e., load the *trace.out* file on Konata).

Based on the CPU architecture described in *riscv\_o3\_custom.py*, visualize the Konata's pipeline to find out the conditions:

1. Out-of-order execution (issue), in-order commit (commit)
2. Two commits in the same clock cycle
3. Flush of the pipeline.

For every condition, fill the following tables.

| Condition   | Out-of-order execution, in-order commit  |
|---|--|
| <b>Screenshot from Konata</b>                                       |    |
| <b>Explain the reason behind the condition</b>                      | Instruction 19-20<br>Perche istruzione 19 è dipendente dal risultato dell'istruzione 18, mentre l'istruzione 20 no, quindi entra in fase di issue prima  |
| <b>Briefly explain the advantages of the OoO execution in a CPU</b> | <ul style="list-style-type: none"><li>- Maggiore utilizzo delle risorse della CPU: OoO consente alla cpu di mantenere occupate le sue unità di esecuzione fornendo loro istruzioni pronte per essere elaborate</li><li>- Aumento delle performance</li><li>- Se una certa istruzione sta aspettando un dato, la cpu può continuare ad eseguire altre istruzioni nel frattempo, anziché stallare</li><li>- Branch prediction migliore: consente al processore di effettuare un'esecuzione speculativa, riducendo il costo di una previsione sbagliata</li></ul> |
| <b>Condition</b>  | Two or more commits in the same clock cycle  |

|  |   |
|--|---|
| <b>Screenshot from Konata</b>                  |   |
| <b>Explain the reason behind the condition</b> | <p>Instruction 19-20</p> <p>Le istruzioni terminano nello stesso clock cycle, perchè l'issue può essere fatto out of order ma il commit va fatto in order.</p> <p>In questo caso il commit viene fatto, per le due istruzioni, nello stesso clock cycle, come possiamo notare l'istruzione 19 stalla 1 volta prima di committare e l'istruzione 20 stalla 3 volte, committando nello stesso clock cycle</p> |
| <b>Briefly explain the Commit functioning</b>  | La fase di commit è la scrittura del risultato nei registri o in memoria  |
| <b>Condition</b>                               | Flush of the pipeline   |
| <b>Screenshot from Konata</b>                  |    |
| <b>Explain the reason behind the condition</b> | Avviene quando ad esempio un salto non viene presto, quindi si vanno a perdere molti colpi di clock, anche perché carichiamo 4 istruzioni alla volta  |

## Exercise 2:

Given your benchmark (*main.c* in *my\_c\_benchmark*), optimize the CPU architecture (i.e., modify the *riscv\_o3\_custom.py* file) and write down the improvements in terms of CPI and speedup.

- To optimize the CPU architecture, open the configuration file of the CPU (i.e., the *riscv\_o3\_custom.py*), and tune specific hardware-related parameters.

You have to change specific values in **one or more** stages of the pipeline:

- # - FETCH STAGE
  - Tune parameters such as the *fetchWidth*, *fetchBufferSize* and so on, and see the effects on your system.
- # - DECODE STAGE
- # - RENAME STAGE

- Try changing some values, but don't touch the "Phys" ones.
- # - DISPATCH/ISSUE STAGE
- # - EXECUTE STAGE
  - Here you can optimize the Functional units of your CPU like the INT ALU, the FP ALU, the FP Multiplier/Divider and so on.
  - Tune the number of units (*count*) that you have in the system, as well as their latency (*opLat*) to see how this affects the execution of your program.
- You can create a different branch predictor. They are defined in *create\_predictor.py*)
- You can also try to change the parameters of the L1 Cache. Look for the "class L1Cache" in the *riscv\_o3\_custom.py* file. The L1 cache, also referred to as the primary cache, is the smallest and fastest level of memory. It is located directly on the processor, and it is used to store frequently accessed data by the CPU. In this way, the CPU saves time with respect to the normal access to the main memory.

**HINT:** To implement the best hardware optimization, and understand how to change the parameters, the best option consists in analysing the *stats.txt* file (in **ase\_riscv\_gem5\_sim/results/my\_c\_benchmark**).

Find information regarding the workload profiling. In other words, look for lines such as "system.cpu.commitStats0.committedInstType::IntAlu", and the following ones to understand which kind of instructions are executed the most. In this way, you can target a specific functional unit and modify its specifications.

Fill the following Tables with the CPI that you obtain with the old and the new architectures. Compute also the equivalent speedup that you obtain.

HINT: You can get the CPI and other useful information from the *stats.txt* file.

| Parameters                              | Configura<br>tion 1         | Configurati<br>on 2    | Configurati<br>on 3    | Configurati<br>on 4      | Configuration 5                                   | Configuration 6             |
|---|-----------------------------|------------------------|------------------------|--------------------------|---|-----------------------------|
| <b>First<br/>changed<br/>parameter</b>  | the_cpu.fetchWidth = 0xc1a0 | the_cpu.issueWidth = 2 | CPU_IntALU<br>Count= 1 | the_cpu.decodeWidth = 12 | In CPU_FP_ALU<br>"FloatAdd",<br>opLat=2 da 4      | the_cpu.decodeWidth = 12    |
| <b>Second<br/>changed<br/>parameter</b> | the_cpu.dispatchWidth = 1   | None                   |                        |                          | "FloatCvt",<br>opLat=2 da 4                       | the_cpu.numROBEntries = 128 |
| <b>Third<br/>changed<br/>parameter</b>  |                             | None                   |                        |                          | IN CPU_FP_MultDiv<br>"FloatMult",<br>opLat=2 da 4 | the_cpu.numIQEntries = 6    |
| <b>Fourth<br/>changed<br/>parameter</b> |                             |                        |                        |                          | "FloatDiv",<br>opLat=2da 4                        | the_cpu.renameWidth = 4     |
| <b>Fifth<br/>changed<br/>parameter</b>  |                             |                        |                        |                          |   | the_cpu.numRobs = 4         |

Original CPI (no hardware optimization):

|   | Configura<br>tion 1   | Configurati<br>on 2 | Configurati<br>on 3 | Configurati<br>on 4 | Configurati<br>on 5 | Configuration 6 |
|---|---|---------------------|---------------------|---------------------|---------------------|-----------------|
| <b>CPI</b>                                    | Mi da<br>errore sulla<br>memory<br>usage, ma<br>anche se<br>mettiamo<br>10 mi<br>fetcha<br>sempre 4<br>istruzioni | 2.083105            | 2.083105            | 1.960070            | 2.017095            | 1.372598        |
| <b>Speedup<br/>(wrt<br/>Original<br/>CPI)</b> |   | 1                   | 1                   | 1.06                | 1.03                | 1.51            |

Which is the best optimization in terms of CPI and speedup, why?

Your answer:

La miglior configurazione che ho trovato è stata la configurazione n6.

Prima di tutto ho incrementato la decodeWidth da 2 a 12, siccome determina quante istruzioni per ciclo la cpu può decodificare, quindi ora può decodificare 6 volte più istruzioni alla volta per ciclo.

Poi ho raddoppiato il ROB da 64 a 128, consentendo al processore di tenere traccia di più istruzioni fuori ordine, aumentando così il parallelismo e riducendo gli stalli dovuti a dipendenze

numIQEntries da 3 a 6, IQ sarebbe la instruction queue, che consente alla cpu di conservare un maggior numero di istruzioni pronte per essere eseguite non appena le risorse necessarie sono disponibili, aumentandone il valore la latenza si riduce e le prestazioni migliorano.

cpu.renameWidth da 2 a 4, anche il renaming register è stato raddoppiato, siccome aumentandone l'ampiezza è possibile rinominare più registri in ogni clock cycle. Rinominando registri si riescono a risolvere le dipendenze dei registri, quindi aumentandone il valore si andranno sicuramente a ridurre stalli, dipendenze.

In ultimo, cpu.numRobs sempre da 2 a 4, aumentando le istanze del ROB viene consentito l'elaborazione fuori ordine di un maggior numero di istruzioni, aumentando ancora una volta la parallelizzazione di esse.

