



**POLITECNICO
DI TORINO**

Recap Computer Architectures (02LSEOV)

Jacopo Nasi
Computer Engineer
Politecnico di Torino

I Period - 2017/2018

20 novembre 2017

Indice

1	Introduction Computer Design	4
1.1	Computer Evolution	4
1.2	Designing	5
2	Instruction Set Principles	8
2.1	Introduction	8
2.2	Memory	9
3	Pipelining	10
3.1	Introduction	10
3.2	Behaviour	11
3.3	Hazards	12
3.4	Exception	15
3.5	Multicycle Operations	17
4	Instruction Level Parallelism	18

License

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

You are free:

- **to Share:** to copy, distribute and transmit the work
- **to Remix:** to adapt the work

Under the following conditions:

- **Attribution:** you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work)
- **Noncommercial:** you may not use this work for commercial purposes.
- **Share Alike:** if you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

More information on the Creative Commons website (<http://creativecommons.org>).



Acknowledgments

Questo breve riepilogo non ha alcuno scopo se non quello di agevolare lo studio di me stesso, se vi fosse di aiuto siete liberi di usarlo.

Le fonti su cui mi sono basato sono quelle relative al corso offerto (**Computer Architectures (02LSEOV)**) dal Politecnico di Torino durante l'anno accademico 2017/2018.

Non mi assumo nessuna responsabilità in merito ad errori o qualsiasi altra cosa. Fatene buon uso!

1 Introduction Computer Design

1.1 Computer Evolution

The first general-purpose computer was created in the late 40s. What now we can buy for 500\$ is equivalent (performance) to what could be bought for about \$1M in 85'.

During the years the performance growth was not linear, as you can see in figure 1, during the first 10 years the annual increase was around 25-30%/year, from the late 80s to the 2000 the growth is increased around 50%/year and, in the last few years it decrease to the 22%. Why this change during the increase?

The manufacturers have found a lot of physical problem related to the creation

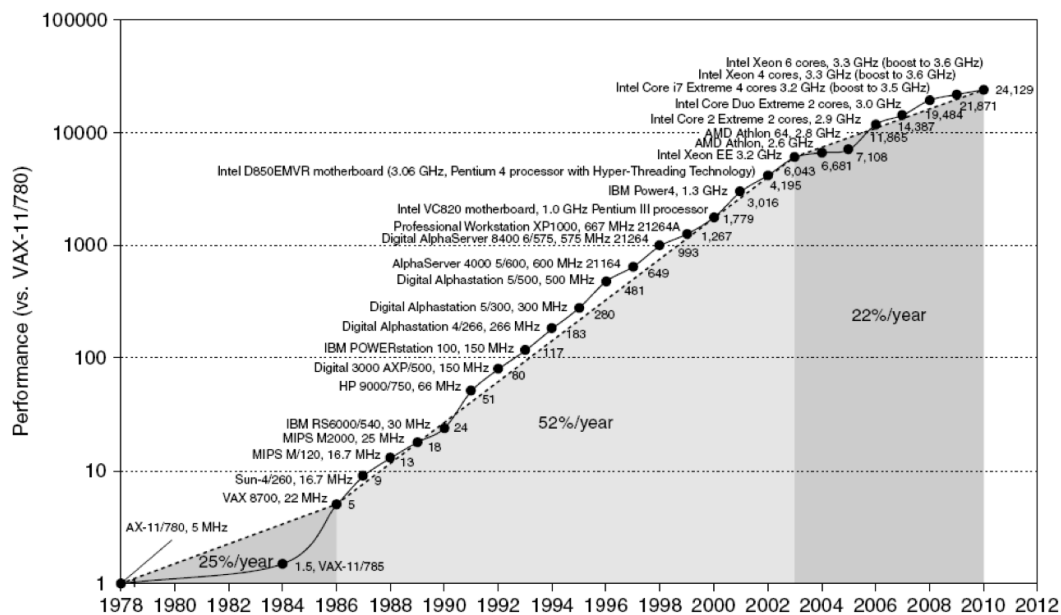


Figure 1: CPUs Growth over years

of new products, this problem are mainly related to:

- Power-Issue.
- Lower instruction-level parallelism.
- Unchanged memory latency.

in fact, since 2004, the major industry have changed the conceptual ways to desing processors, switching from single to multi-core architectures. We can say that, in anytime, this growth is incredible an is due to improvements in technology, microprocessor architecture and software development. Since the multi core introduction the major prefers to investe on multicore system rather than develop faster CPU.

1.2 Designing

There are 5 main market areas:

- **Personal Mobile Device (PMD):** Smartphone, tablet. They are focused in energy efficiency and real-time app.
- **Desktop Computing:** From PC to workstation and the main purpose is optimize the price-performance ratio.
- **Server:** Larger-scale and more reliable computing services.
- **Cluster - WAS:** Emphasis on availability, price-performance and power consumption.
- **Embedded Computers:** Fastest growing portion of PCs market. All special-purpose computer-based application, from cheap to high-end processors.

There are two **Classes of Parallelism:**

- **Data-Level (DLP):** Many data items that can be operated on at the same.
- **Task-Level (TaskLP):** Many tasks of a work can operate independently.

The first solution allows the processor to split the data operation over multiple cores, you can for example divide an array of n elements over 4 cores and, if T is the computational time needed for the entire array, the final time will be $T/4$ plus a little time for the reunion of the data. It splits the one task on different data.

The TLP instead is able to manage multiple tasks over the same data, this is the common behaviour of the actual system (pipelining techniques).

There are different Parallel Architectures:

- **Instruction-level (ILP):** This modestly uses the data-level parallelism.
- **Vector and GPU:** It exploits DLP.
- **Thread-level (TLP):** It exploits DLP and TaskLP.
- **Request-level (RLP):** Exploits parallelism among decoupled (not-related) tasks.

The designing of a new computer involves important analysis to the main purpose of it, you need to study which attributes are important for the new machine and you need to maximize performance and matching cost and power constraints. During the last decade the PC design took advantage of architectural and technology improvements, the performance increase is more than a

factor of 15 on what would have been obtained by relying solely on technology.

The **Moore's Law** says that: *The number of transistors that can be integrated into a single chip doubles every 18/24 months.* Until now the law has worked, as you can see in the figure 2 and, probably, it will work for other time.

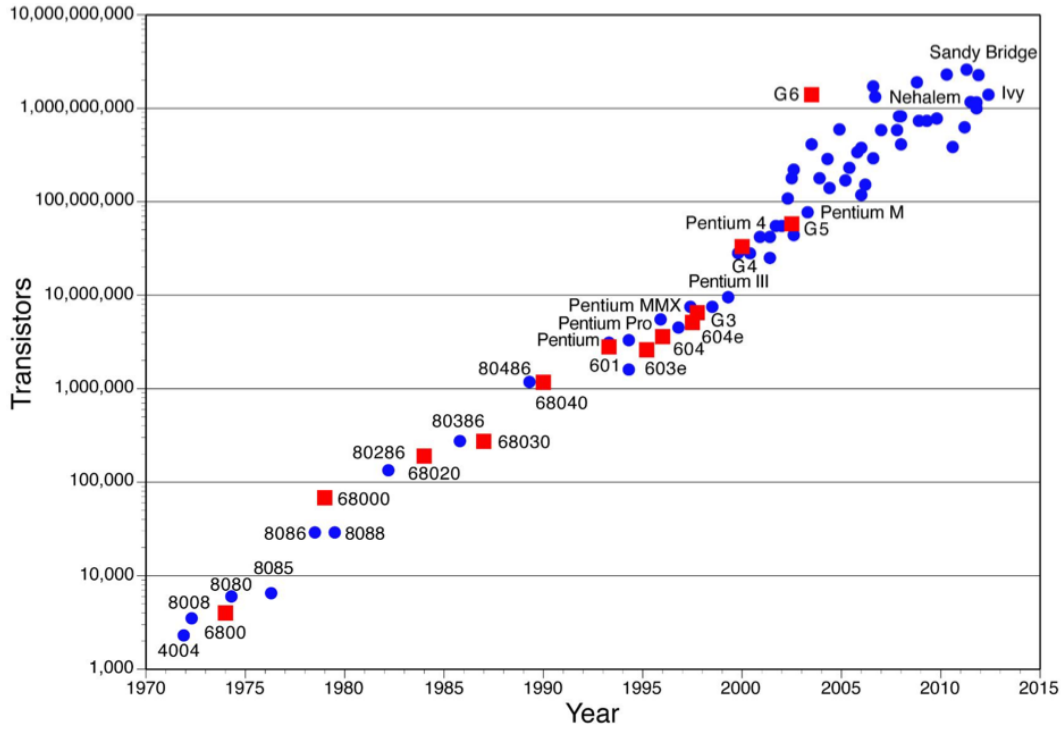


Figure 2: Transistors growth on CPUs

Cost During the evaluation of the IC manufacturing cost it is important not to forget the impact of yield, the percentage of products that pass the test phase. The production process for every product undergoes an evolution which normally leads to an improvement in yield (learning curve). The cost decrease is due to yield increase. Probably the most important part of the manufacturing cost is related to the validation and testing procedures.

Other designing problem The continuous increase of the system complexity and device integration leads to problem with power consumption, for the Power and for the Energy (mainly for portable devices). Until now the greater power consumption contribution is due to the transistor switching phase, related to the physical formulas the solution applied is trying to reduce the voltage.

Another important factor is the **Dependability** that is the quality of the

system to deliver a correct service, is traditionally very high, but it can be lowered by software or hardware bugs both during the production or in designing phase. During this years the safety-critical areas where the microprocessor have relevant importance are increased, initially only space, avionics and nuclear, but now we have rail-road, automotive, biomedical, telecom, ecc...

The dependability is often measured using:

- **Mean Time To Failure (MTTF)**: 1 failure in one billion hours.
- **Mean Time Between Failures (MTBF)**
- **Mean Time To Repair (MTTR)**

This three measures are related by the formula: $MTBF = MTTF + MTTR$

Performance they be analyzed under multiple point of view, *Time between start and completion of an operation*, *Total amount of work done in time unit*, ecc.... The UNIX system provide 4 different values:

- Elapsed time
- CPU Time
 - User
 - System

The evaluation of often performed letting work a computer and observing its behavior. Unfortunately, the choice of the application severely affects the performance and is not too easy looking the result in a correct way. The main solution is using benchmarks to compare different system running the same operation and production comparable times of execution. The benchmark sets are normally composed of:

- Kernels
- Program Fragments
- Applications

There are multiples solution to analyze the performance one of the most common is the **Amdahl's law** and it based to the comparisons with the older version of the same product. The speedup is calculated by this formula:

$$Speedup = \frac{Performance_{enhancement}}{Performance_{NOenhancement}} \quad (1)$$

the value depends on two factors:

- $Fraction_{Enhanced}$: the fraction of the computation time that take advantages of the enhancement.

- $Speedup_{Enhanced}$: the suze of the enhancement on the part it affects.

A more complete formula is:

$$Speedup_{overall} = \frac{1}{(1 - Fraction_{Enhanced}) + \frac{Fraction_{Enhanced}}{Speedup_{Enhanced}}} \quad (2)$$

an example could be useful:

An enhancement makes one machine 10 times faster for 40% of the programs the machine runs. Which is the overall speedup?

Fraction = 0.4, Speedup = 10.

$$Speedup_{overall} = \frac{1}{(1 - 0.4) + \frac{0.4}{10}} = 1.56 \quad (3)$$

Another important designing evaluation is measuring the time required to execute a program, the are severals approaches:

- Observing the real system: Not easy to be evaluated.
- Simulation: It could be really expensive.
- CPU Equation.

the latest solution use a provided formula to evaluate the CPU time:

$$CPU_{Time} = CLK_{Time} * \sum_{i=1}^n CPI_i * IC_i \quad (4)$$

where:

- CPI_i : Number of clock cycles required by instruction i (depends on hardware and instr set).
- IC_i : Number of times instruction i is executed in the program (depends instruction set and compiler).
- CLK_{Time} : Inverse of clock frequency (depends technology).

in the pipelined processor, CPI_i may vary for a given instruction, therefore the evaluation becomes much harder.

2 Instruction Set Principles

2.1 Introduction

The Instruction Set Architecture (ISA) is hoe the computer is seen by the programmer or the compiler. There are different kind of design and they can be selected by different characteristics. The CPUs are often classified according to the type of their internal storage:

- **Stack:** Is the simplest one it can't access memory during operations.
- **Accumulator:** Similar to 8051 can access the memory during operation.
- **Registers**
 - **Register-memory:** Similar to 8086, have a 16-bit register.
 - **Register-Register (load-store)**
 - **Memory-memory (no real cases)**

Nowadays all processors are General-Purpose Register (GPR) this because registers are faster than memory and are easier for a compiler to use. The CPUs are also classifiable by the number of operand per ALU instruction (2 or 3) and number of memory operand per ALU.

2.2 Memory

The memory has a fundamental work in the CPU world, there are several different type of it and of course different type to how accessing it.

There are many different **memory addresses** that can be implemented:

- **Little Endian:** Byte with lower address at the least significant position. The addresses of the data is that of the LSB.
- **Big Endian:** Byte with lower address at the most significant position. The address of the data is that of the MSB.
- **Aligned:** Allowing only aligned accesses to memory could be a limitation in terms of performance.
- **Misaligned:** This solution requires hardware and performance overhead.

A memory position can be accessed in different ways:

- **Register:** (ADD R4, R3) when a value is in a register.
- **Immediate:** (ADD R4, #3) for constants.
- **Displacement:** (ADD R4, 100(R1)) accessing local variables.
- **Deferred or Indirect:** (ADD R4, (R1)) accessing using a pointer or a computed address.
- **Indexed:** (ADD R3, (R1 + R2)) useful in array addressing: R1=array base, R2=index.
- **Direct or Absolute:** (ADD R1, (1001)) useful for accessing static data.

- **Indirect:** (ADD R1, @(R3)) if R3 is the addr. of a pointer p, then the mode is like p.
- **Autoincrement or decrement:** (ADD R1, (R2) \pm) useful for stepping through array within a loop.

Is important to choosing the correct addressing mode, one can obtain some important consequences, from reducing the number of instruction to increasing the architecture.

There are a lot of instruction that can be done, looking at the 80x86 frequency we can see that most most used are *Load*, *Conditional Branch*, *Compare and Store* that are around the 70% of the total operation performed.

The instruction set encoding depends on which instruction compose the instruction set and which addressing modes are supported. When a high number of addressing modes is supported, and address specifier is used to specified the addressing mode. When the number of addressing modes is low, they can be encoded together with the opcode. The different encoding are:

- **Variable:** Any number of operands, operation with variable length, lower performance and minimum code size.
- **Fixed:** Fixed number of operands, need address specifier, fixed instruction length, maximum performance but larger code size.
- **Hybrid:** Multiple format specified by the opcode and it allows a trading-off between code size and performance.

Assembly-level programs are now produced by compilers only. The CPU designer and the compiler writer must interact and cooperate. A crucial part taken by the compiler is the optimization of the register using, it easier to solve it when the number of register is higher (>16).

3 Pipelining

3.1 Introduction

Pipelining is an a solution to execute multiple instructions at the same time overlapping it. The different units (also called *stage* or *segment*) are completing different part of different instruction in parallel. The figure 3 is an example of this techniques. The **throughput** of a pipelined processor is the number of instructions which exit the pipeline in the time unit. All the pipeline stages are synchronized (they proceed to executing the new task all together); the necessary time for executing one step is called machine cycle and, normally, corresponds to one clock cycle. Of course the duration of a machine cycle is determined by the slowest stage. CPI means clock cycles per instruction. The ideal pipeline have all stages perfectly balanced, in this way the $TRP_{pipelined} = TRP_{unpipelined} * n$ where n is the number of pipeline stages.

Instruction Cycle: Fetch, Decode, Operations, Write

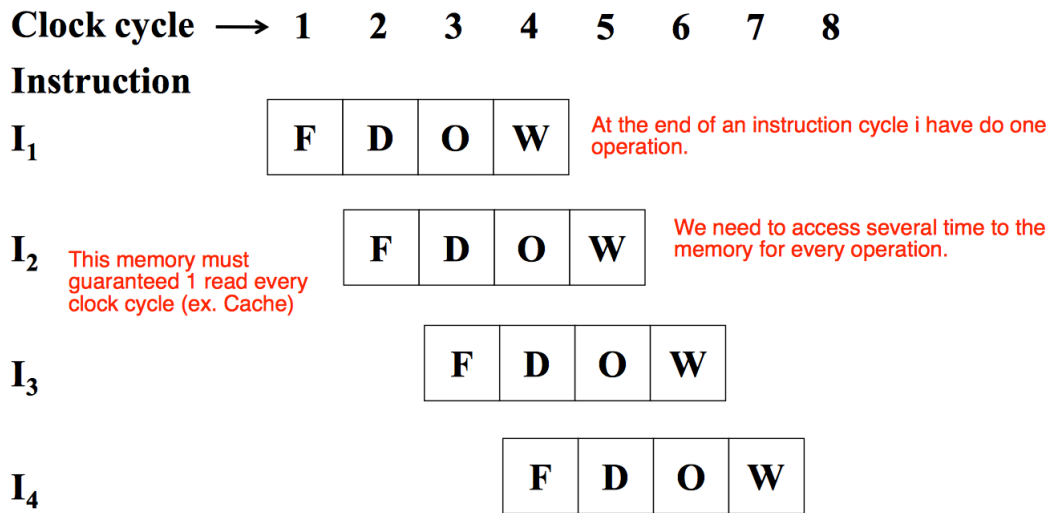


Figura 3: Pipeling example

3.2 Behaviour

The execution of each instruction may be composed of at most five clock cycles:

- **IF**: Fetch the instruction from the memory address of the PC and save it in the IR.
- **ID**: Decode the instruction just fetched.
- **EX**: Execute the instruction.
- **MEM**: Memory access and brach completion.
- **WB**: Write-back.

with the unpipelined processor all instruction require 5 CC, the only optimization to reduce the average CPI is completing the ALU instruction during the MEM cycle, hardware resources could be optimized avoiding duplications. The pipelined version instead can be more powerful beacause a new instrcution is started at each clock cycle and different resources work on different instruction at the same time. There are several consideration to keep in mind during the development of this units; at every clock cycle, each resource can be used for one purpose only, this means that:

- Separate instruction and data memories must be used.
- The register file is used in two stages: for reading (second half of CC) in ID stage and for writing (first half od CC) in WB stage.

- The PC (*Program Counter*) must be changed in the IF stage, What about branches?

A little view of the system behaviour in figure 4.

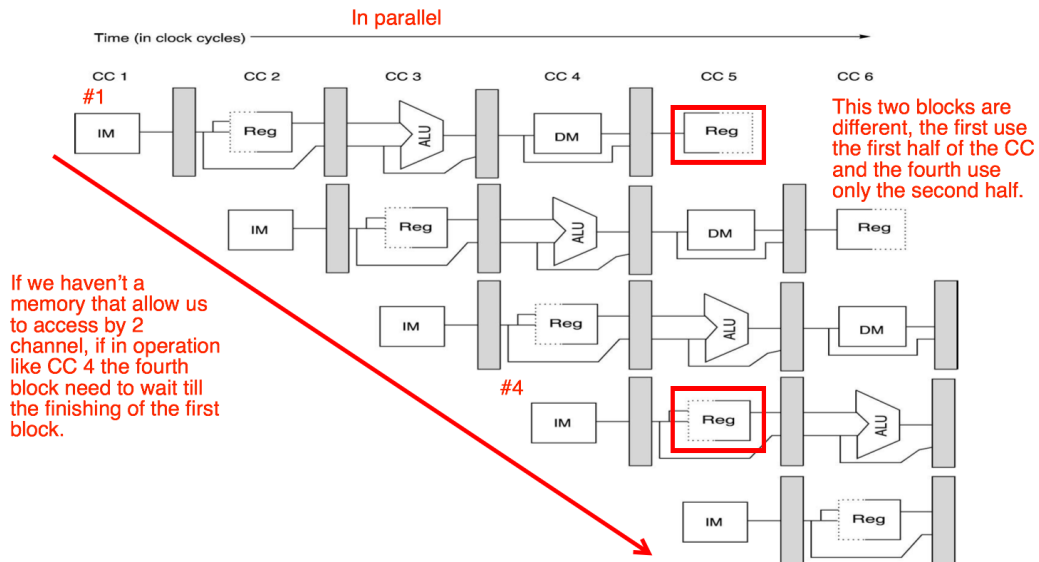


Figure 4: Evolution in time of Pipelining architecture

Performance The first purpose of this architectural type is the performance increase without making single instruction faster. The single instruction are made slower for the overhead introduced by the pipeline control. The limitations of a pipeline come from the necessity of balanced stages and pipelining overhead (pipeline register delay and clock skew).

3.3 Hazards

The hazards are situations that prevent the execution of an instruction during its designated clock cycle. There are 3 classes of hazards:

- **Structural:** Coming from resource conflict.
- **Data:** And instruction depends on the result of a previous instruction.
- **Control:** Depend on pipelining branches and other instructions that change the PC.

One way to dealing with hazards is to force the pipeline to stall, i.e., to block instructions for one or more clock cycles. When an instruction is stalled:

- The instructions following the stalled instruction are also stalled.
- The instructions preceding the stalled instruction continue.

A stall causes the introduction of a bubble in the pipeline.

Structural hazards they may happen when some pipeline unit is not able to execute all the operations scheduled for a given cycle. Example:

- A given unit is not able to complete its task in one clock cycle (division).
- The pipeline owns only one register-file write port, but there are cycles in which two register writes are required.
- The pipeline refers to a single-port memory, and there are cycles in which different instructions would like to access to the memory target.

Solving this problem implies adding or improving the hardware.

Data hazards overlapping the execution of instructions, as it is done by pipelining, changes the order of read/write accesses to operands. This can lead to wrong results and undeterministic behaviour.

An example could be:

```
ADD R1, R2, R3
SUB R4, R1, R5
AND R6, R1, R7
```

The second instruction will read the value of R1 before the first instruction can write the new value. This will cause a wrong result. And is the same of course for the third instruction.

If an **Interrupt effects** occurs during the execution of a critical piece of code (from the pov of data hazards) correctness may be restored. This may cause an undeterministic behaviour.

There are two possible solutions for data hazards problems:

- The first is using the stall till the data will be available.
- Implementing the **forwarding** technique.

Forwarding A special hardware in the datapath detects when a previous ALU operation should write the register corresponding to the source of the current ALU. In this case, the hardware, selects the ALU result as the ALU input rather than the value from the register file. This part must be able to:

- Forward a data from any of the previously started instructions (the data can't be already written in its final location).
- Not to forward anything, if the following instruction is stalled, or an interrupt occurred.

an example can be viewed in the figure 5.

The forwarding can be implemented:

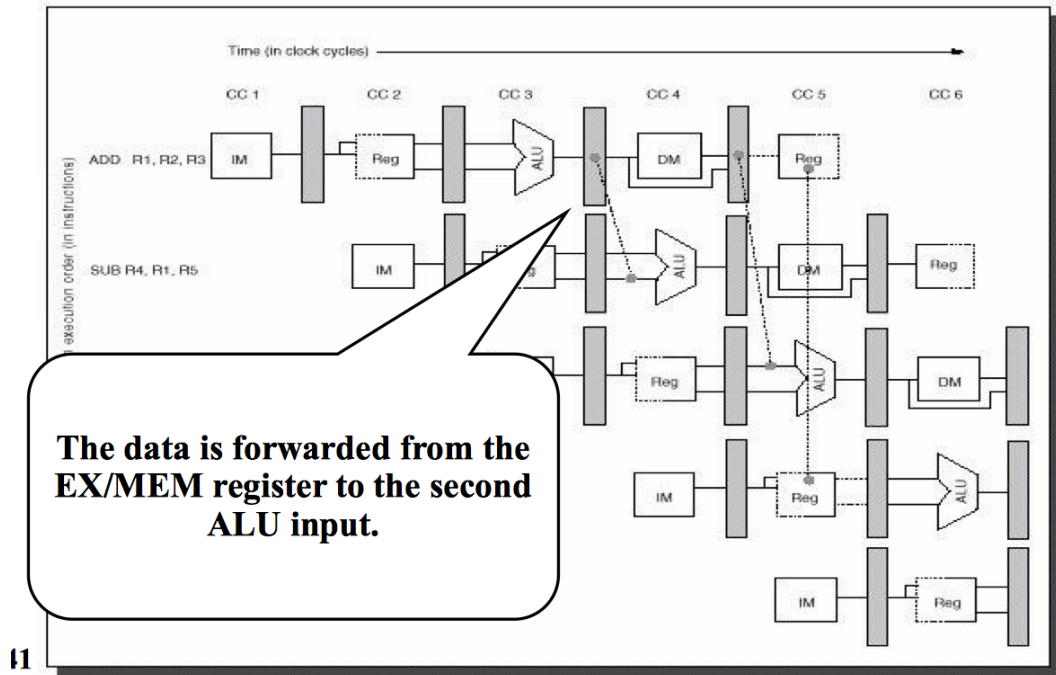


Figura 5: Forwarding Example

- From the ALU or data memory output...
- ...to ALU inputs, data memory inputs, or the zero detection unit (branch instruction).

and the logic must compare:

- The DEST fields of the IR in the EX-MEM and MEM-WB register with...
- ... the SRC fields of the IR in the IF-ID, ID-EX and EX-MEM registers.

Causes of Data Hazards A hazards is created whenever there is dependence between instructions, and they are close enough that the overlap caused by pipelining would change the order of access to an operand. In general, this can happen for:

- Register Operands
- Memory Operands:
 - Accesces to memory by load and store are not made in the same stage.
 - execution can proceed while an instruction waits for a cache miss to be solved.

The forwarding not solve all potential data hazards, in this case to only solution is to use stalls.

Stalls At each clock cycle all test for detecting data hazards are performed (ID stage), if an hazard is detected two actions can be taken:

- The appropriate forwarding is enabled.
- The instruction is stalled before entering in the wrong stage.

This situation can be performed in two ways:

- Forcing all 0s in the ID-EX pipeline register (*nop* instruction).
- Forcing the IF-ID pipeline register to maintain the current value.

Control Hazards are due to branches, which may change the PC after the following instruction has been fetched already. In the case of conditional branches, the decision on whether the PC should be modified or not can be taken even later. In the MIPS implementation, the PC is written with the target address (if the jump is taken) at the end of the ID stage.

A possible solution is based on stalling the pipeline as soon as a branch instruction is detected (ID stage) by:

- Decide earlier whether the branch has to be taken or not.
- Compute earlier the new PC.

Performance there are several techniques for reducing the performance degradation due to branches:

- **Freezing pipeline:** the pipe is stalled until the decision about the branch is known.
- **Predict untaken:** Assume the branch is not taken, avoid any change in the pipe, undo all performed if the branch is taken. *
- **Predict taken:** If the target address is known before the branch outcome, it may be possible to assume the branch as taken. *
- **Delayes branch:** Filling the slot after the branch instruction (branch-delay slot) with instructions which have to be executed no matter the branch outcome. Is a compiler task. Less used nowadays.

* This techniques can be speeded by the compiler which maximizes the chance for the processor to make the right prediction.

3.4 Exception

The exception are fundamental for system. There are a lot of possible causes:

- I/O device request
- System call by user program
- Tracing instruction execution (like debug mode)
- Breakpoint
- Overflow and underflow
- FP anomaly
- Page fault
- Misaligned memory accesses
- Undefined instruction
- Hardware malfunction
- Power failure

they can be classified based on the type:

- **Synchronous - Asynchronous:** Same position in code - External devices.
- **User requested - Coerced:** Like procedures - Out of user control.
- **User maskable - Non maskable:** Force hardware not to answer to exception requests.
- **Within - Between instructions:** Generated by instruction itself.
- **Resume - Terminate:** Terminate or execute something then resume.

There are some machines, called restorable, that are able to handle an exception, save the state, and restart without affecting the execution of the program. Nowadays all processors are restorable.

Stopping execution when an exception occurs, the pipeline must execute the following steps:

- Force a trap instruction into the pipeline on the next IF stage.
- Until the trap is taken, turn off all writes for the instruction that raised the exception and for all the following instructions in the pipeline.
- When the exception-handling procedure receives control, it immediately saves the PC of the faulting instruction.

after the exception has been handled, special instructions return the machine from the exception by reloading the PC and restarting the instruction stream. A processor has precise exceptions if the pipeline can be stopped so that:

- The instructions just before the faulting instruction are completed.
- The instructions following the faulting instruction can be restarted from scratch.

It could be really hard restarting after exception, but precise exception is a must for most architectures, at least for integer.

Contemporary exception for example we can have 2 different instruction LD that thrown an exeception during the MEM phase and a DADD that can generated an arthmetic exception during the EX phase. The handling could manage the first problem, and if its cause is removed, the second occurs. The situation can be different supposing that the DADD generate an error in the IF stage and the LD still in the MEM phase, this case generate before the DADD exception and then the other.

The are multiple solution, a good one could be:

- A status flag is associated to each instruction in the pipeline.
- If an instruction causes an exception, the status flag is set.
- If the status flag is set, the instruction can not perform any write operation.
- When an instruction reaches the last stage, and its status flag is set, an exception is triggered.

In some cases machines have instructions that change the state before they are committed (those using autoincrement addressing modes). If one of these instructions is aborted because of an exception, it leaves the machine state altered. Implementig precise exceptions could be really tough.

Instructions implicitly updating condition codes create complications:

- Data Hazards
- Save and restored in case of exception.
- Hardening the compiler work to filling delay slots between writing conditions and branch.

3.5 Multicycle Operations

Floating points units perform more complex operations than the integer ones. Forcing them to be executed in a single CLK means using a very slow clock or using really complex unit. The main solution is repeating the EX stage as many times as the instruction requires. An example for undestan the number can be viewed in this table below:

Functional Unit	Latency	Init Interval
Integer ALU	0	1
Data Memory	1	1
FP Add	3	1
FP/Int Multiply	6	1
FP/Int Divide	24	25

Latency: It is the number of cycles that should last between an instruction that produces a result and an instruction that uses the same result.

Initiation Interval: It is the number of cycles that must elapse between issuing two operations of the same type to the same unit.

Hazards can be more frequent due to the different structure of the EX stage. This because:

- Unpipelined divide unit, where several instructions could need it at the same time.
- The instructions have varying running times, the number of the register writes required in a cycle can be larger than 1.

solving this problem with more write ports is normally too expensive, the solution is forcing a structural hazard stalling the instructions in the ID stage, or stalling it before the MEM or WB stage. This kind of operations can introduce long period of stall. This introduces new type of hazards, like **RAW** (Read After Write), **WAW** (Write After Write) and structural hazards involving the divide unit and the write port.

A solution could be that, before issuing an instruction to the EX phase, check whether it is going to write on the same register of an instruction still in the EX stage. In this case, stall the new instruction. Of course guaranteeing the precise exception becomes really hard, the possible solutions are:

- Accepting imprecise exceptions.
- Fast, but imprecise operating mode, and slow, but precise one.
- Buffering the results of each instruction until all the previously issued instructions have been completed.
- Forcing the FP units to early determine whether an instruction can cause an exception, and issuing further instructions only when the previous ones are guaranteed not to cause an exception.

An alternative solution is like the MIPS R4000 that implements a 8 cycle pipelined processors. This means more forwarding, increased load delay slot and increased branch delay slot.

4 Instruction Level Parallelism