Laboratory exercises for EITF20 Computer Architecture

Anders Ardö
Electrical and Information Technology
Lund University

December 3, 2016

Contents

1 **Laboratory 1: Pipelined processors** 3

2 **Laboratory 2: Advanced pipelining** 15

3 **Laboratory 3: Cache memory** 21

4 **Laboratory 4: Advanced cache; Processor design trade-offs** 29

A **Appendix: Software** 37

A.1 **Pipeline simulator (mipspipe2000.exe)** 37

A.2 **Cache simulator mips.exe** 37

A.3 **SimpleScalar simulator tool set** 37

A.4 **MatLab routines for plotting results** 43

LUND INSTITUTE OF TECHNOLOGY
Lund University

In the course of these labs you will run a large number of simulations, and it may be difficult to keep track of your results unless you maintain a lab book (hard copy or on-line). This book should contain the description of all the simulation runs you performed, your simulation plans, comparison of results, graphs if any etc. In addition as you will start using more detailed simulators, the simulation time will increase. A lab-book, which documents all the simulation runs you performed already, will help you avoid repeat runs and will save you considerable time. The system allows you to attach an arbitrary ID to each run, as well as it saves all results within a session (typically one lab). Use this intelligently in order to keep better track of what you have done.

The software tools used in this laboratory are of an educational nature. In plain English this means that one may expect all kind of problems with the tools themselves, the installation and the assistants, because 'things can have changed since last time'. We will give you no other guarantee than all the help we can.

Written solutions to home assignments for a lab should handed in to the lab-assistant before the lab starts.

1 Laboratory 1: Pipelined processors

User-name: Your Lucat-ID
Password: Your Password
All software tools are described (and possibly available) at http://dark.eit.lth.se/

1.1 Goals

In this laboratory exercise we practice the pipelining tricks from the lectures. First we will work with the classical pipeline to see the stalls coming up and partly disappear again when we introduce forwarding. Then as the pipelining becomes more complicated we move to the score-board. Finally we take a short look at the Tomasulo technique. On this travel we will encounter a number of different education tools, software specially developed to illustrate the lectured concepts.

After this laboratory exercise, you should understand the basic principles of how pipelining works, including the problems of data and branch hazards and possible remedies like forwarding and delayed branching. Finally, you should have an understanding for how the instructions are used to control different parts of a data path through a control unit.

1.2 Literature

Hennessy and Patterson: Appendix A, Chapter 2
MIPS Lab Environment Reference Manual (section A.1 in this lab-manual)

1.3 Preparations

Read the literature and all laboratory exercises below in detail, and solve the home assignments. Note that you must solve the home assignments, or you will not be allowed to start the laboratory exercise. You need to check whether this laboratory description is understandable for you without having to read the material cited at the end, which reflects information that could and should have been passed onto you at previous courses.

1.3.1 Home assignments

Written solutions to home assignments should handed in to the lab-assistant before the lab! This goes for all 4 labs.

- What is a CPU (or processor)?
- What is an assembly language program?
- How does a computer execute simple machine language instructions?
- What is the relation between assembly language and machine language?
- What does the instruction '`add t0, t1, t2`' do?
- What does the instruction '`beq t0, t1, Dest`' do?
- How does a pipelined CPU differ from a non-pipelined?
- Describe hazards?

1.4 Dry run through the basic simulation environment

The first experiments will be performed using MipsIt (found in the **mips** katalogue (S:)) for the analysis and linkage of assembler programs and MipspipeIt for the subsequent execution. MipsIt Studio is a Windows hosted integrated development environment (IDE) for the IDT MIPS cards (and simulator). Figure I.1 shows the IDE in action.

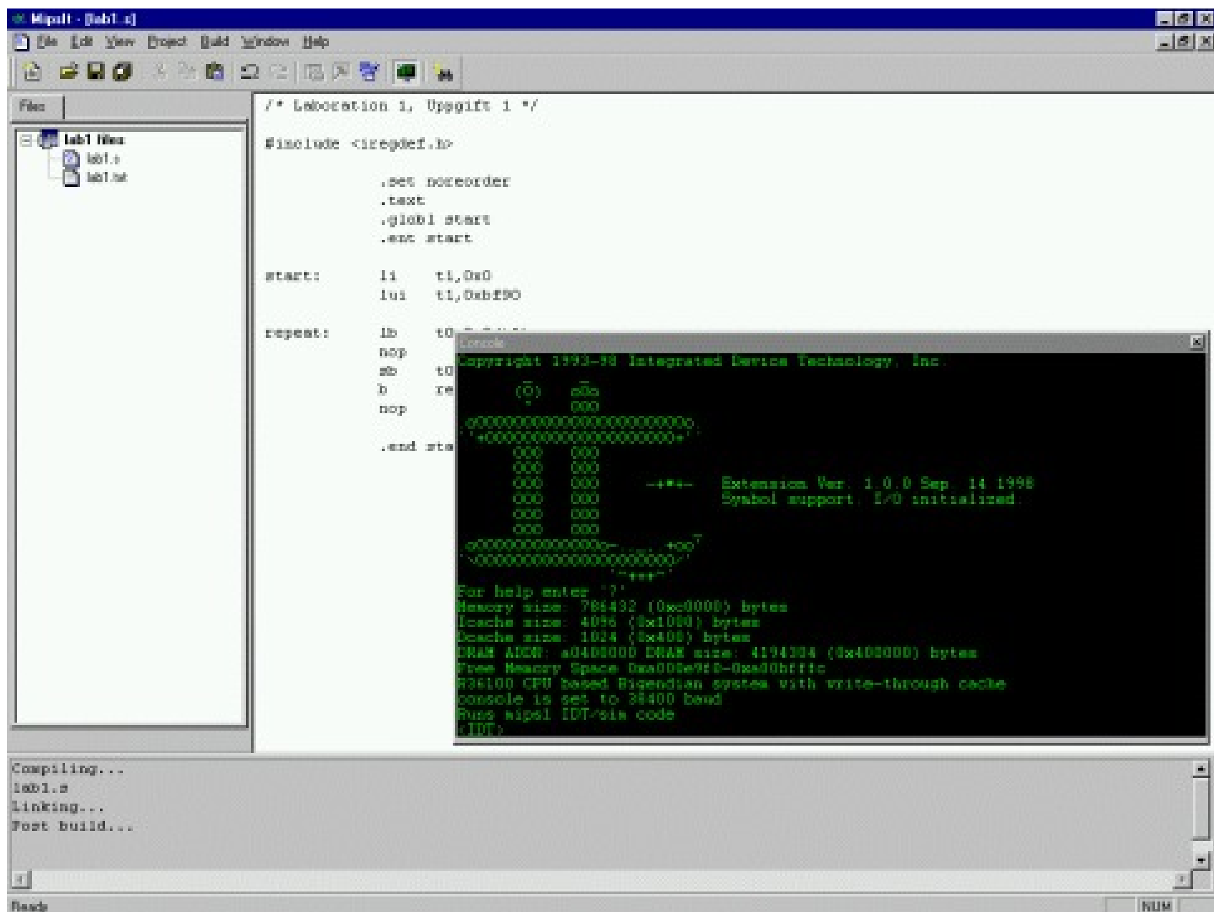

Figure 1: MipsIt Studio 2000.

If you have used Microsoft Developer Studio/Visual C++, you should have a pretty good idea how MipsIt works. But if you are new to IDEs, you need to know what a project is. A project is a collection of interrelated source files that are compiled and linked to make up an executable file that can be uploaded to the simulator. A project may also include text files for informational purposes.

1.4.1 IDE Basics

The IDE consists of the following windows (see Figure 1):

- The project view that contains a list of files included in a project. To open a file for editing double click on it in the list.
- The output window where all output from building etc. is printed.

Many commands also have hot-keys (like most Windows programs) to make work more efficient. There is also a toolbar with some of the commands. Some commands are currently non-implemented and therefore disabled at all times.

To configure the IDE, choose Options from the File menu. You can change COM settings, compiler executable, paths etc. When you start MipsIt the first time it will normally auto-configure correctly except for the COM-port.*This can normally be ignored - just klick OK.*

1.4.2 Creating a Project

To create a new project follow these steps:

- Choose 'New' from the 'File' menu, and then click the Project tab (if it is not already selected) in the resulting New dialog box shown in Figure 2.

- Select the type of project you want to create from the list. The project types are as follows:

  1. Assembler - If your project will only contain assembler files. **This has to be your choice in this lab!**
  2. C/Assembler - If you want a project that will contain only C or C and assembler files.
  3. C(minimal)/Assembler – Same as C/Assembler except with minimal libraries. This is your choice (in lab 3) if you want a project that contains C files.

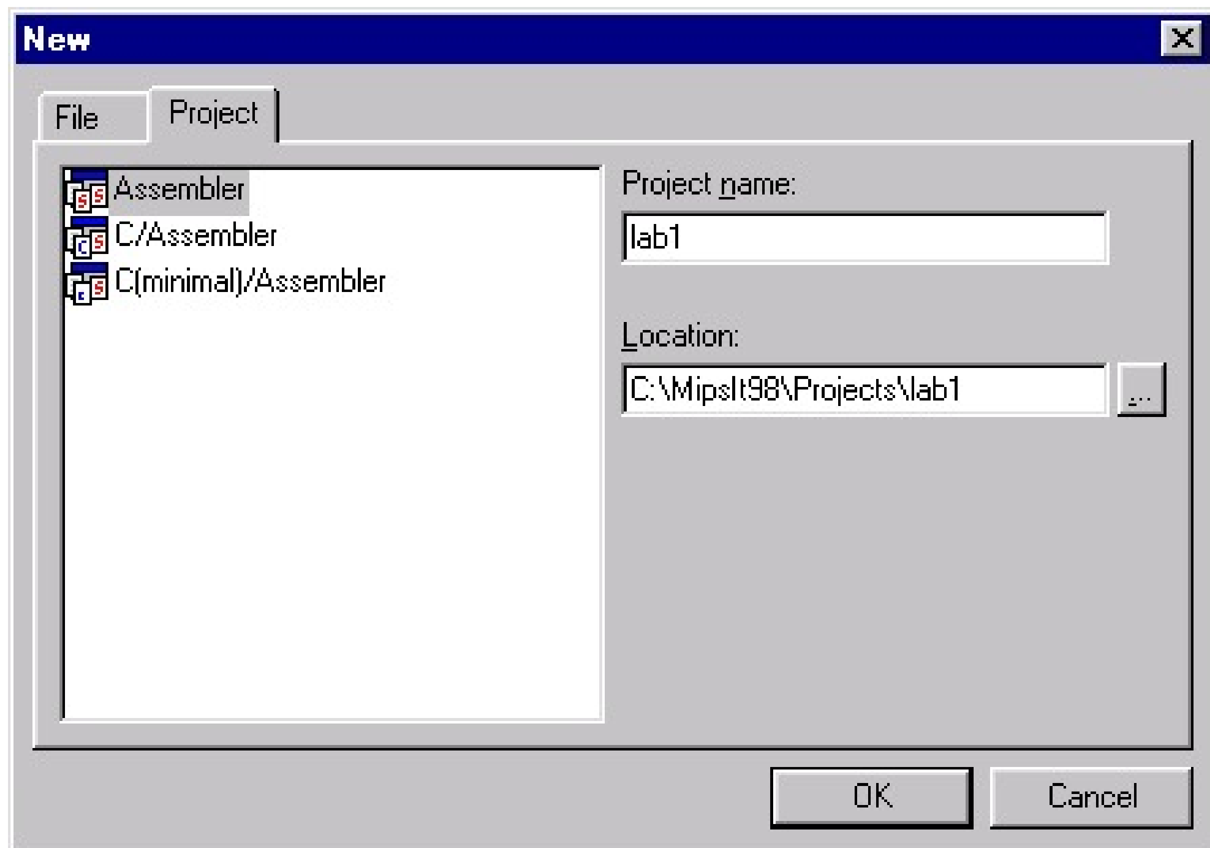

Figure 2: The 'New' dialog, Project tab.

The differences between the project types are the default libraries and modules. A C/Assembler project will link with a couple of libraries and will result in a bigger executable (which won't work with the simulator). A C(minimal)/Assembler project will link with only the absolutely necessary libraries and will result in a smaller executable than with C/Assembler (which will work with the simulator).

- Enter a name for the project and change the location if desired, and then click OK.

1.4.3 Adding Files to a Project

If you followed the steps for creating a new project you should now have an empty project. You can now add files to it by either creating new files or adding existing files. Creating a new file is very similar to creating a new project, except you select the File tab (see Figure 3) instead of the Project tab in the New dialog. If you want to add an existing file choose Add File from the Project menu, and then select the file you want to add. In both cases we are in this laboratory exercise only interested in assembler files, having the extension .s

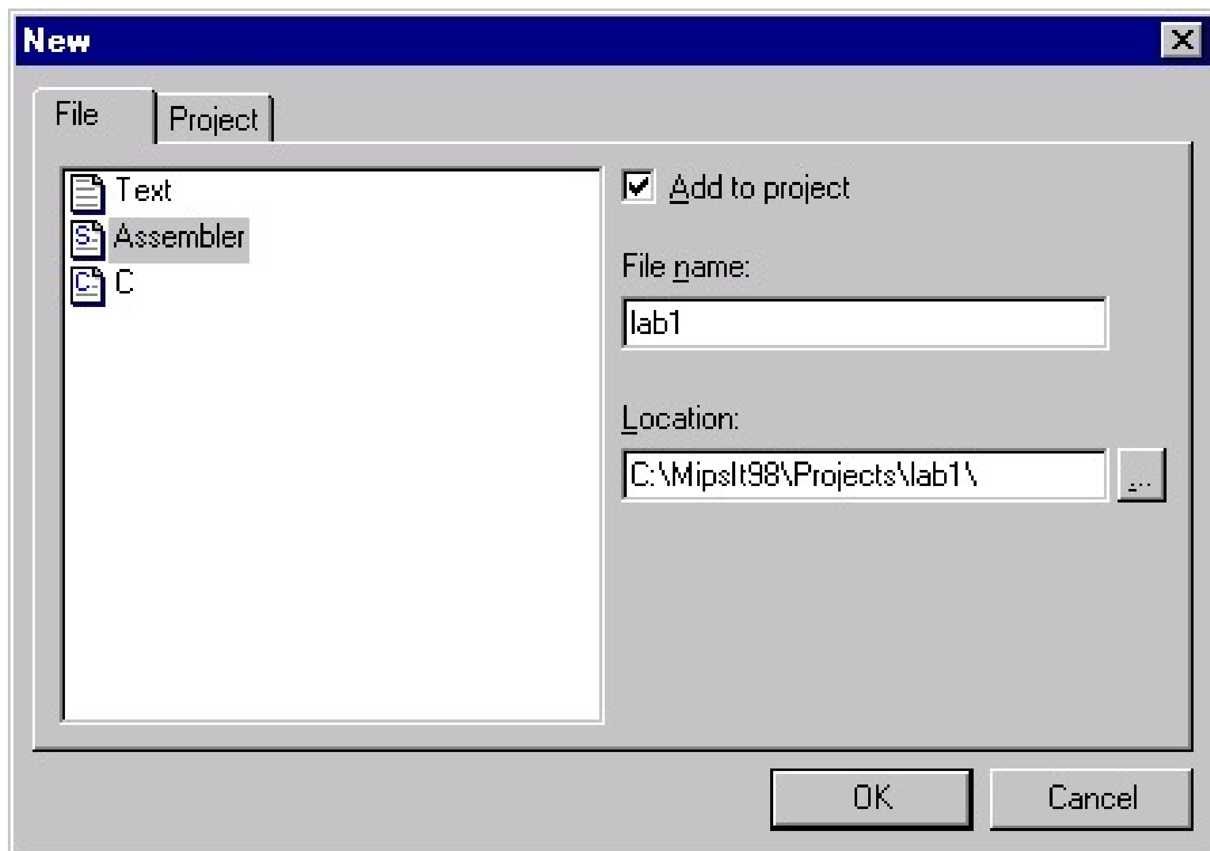

Figure 3: The New dialog, File tab.

As a prototype for the assembler code that we are going to run, enter the following program:

```
1  #include <iregdef.h>
2          .set noreorder
3          .text
4          .globl start
```

```
5       .ent start
6 start: add t0, t1, t2
7       nop
8       nop
9       nop
10      nop
11      .end start
```

1.4.4 Building

In order to prepare your project for simulation, choose Build from the Build menu. Any file that needs compilation will be compiled (or assembled) and finally the executable will be linked. Current status and results of the build process can be seen in the output window (see Figure 1). In case you want to re-compile all files, even those that have not been modified since last build, choose Rebuild All from the Build menu. When the project has been successfully built you can now move to the simulator.

1.4.5 Simulation

Now open the program `mipspipe2000.exe` (in the `mips` katalogue (S:)) to run the pipe simulator. Choose 'Load Pipeline' from the file menu, open the directory called 'S-script' and choose the file called `s.dit` to load the small version of the pipeline. Next choose 'Open' from the file menu or from the toolbar, go to the directory where you saved program, and open the file of '.out' type in the directory called 'Objects'. Figure 4 shows the screen you are going to see. You may play around with the buttons, but we have entered a program without input. So you are going to see very little.

1.5 Arithmetic Instructions

Different classes of instructions use a different selection of the available components in the data path. It is common to group such instructions into four classes: arithmetic, load, store, and branch. Within one such class the instructions are quite similar, and it is often enough to understand one of them in order to understand them all.

The arithmetic instructions are sometimes also called register instructions, because they perform an operation with two source and one destination registers. We will now study how an arithmetic instruction goes through the pipeline.

1.5.1 Experiment 1

Go back to MipsIt, modify your program to insert some distinct values in t1 and t2 (for example by adding new instructions `lui ...` before the `add` instruction), rebuild the project and execute the program MipsPipe2000 by single stepping through each pipeline stage. (Reseting the simulator is done by restarting MipsPipe2000.)

Now answer the following questions, while describing all signals, register changes, and other effects in detail:

- What happens when the instruction goes through the first pipeline stage, the IF stage?
- What happens in the second (ID) stage?

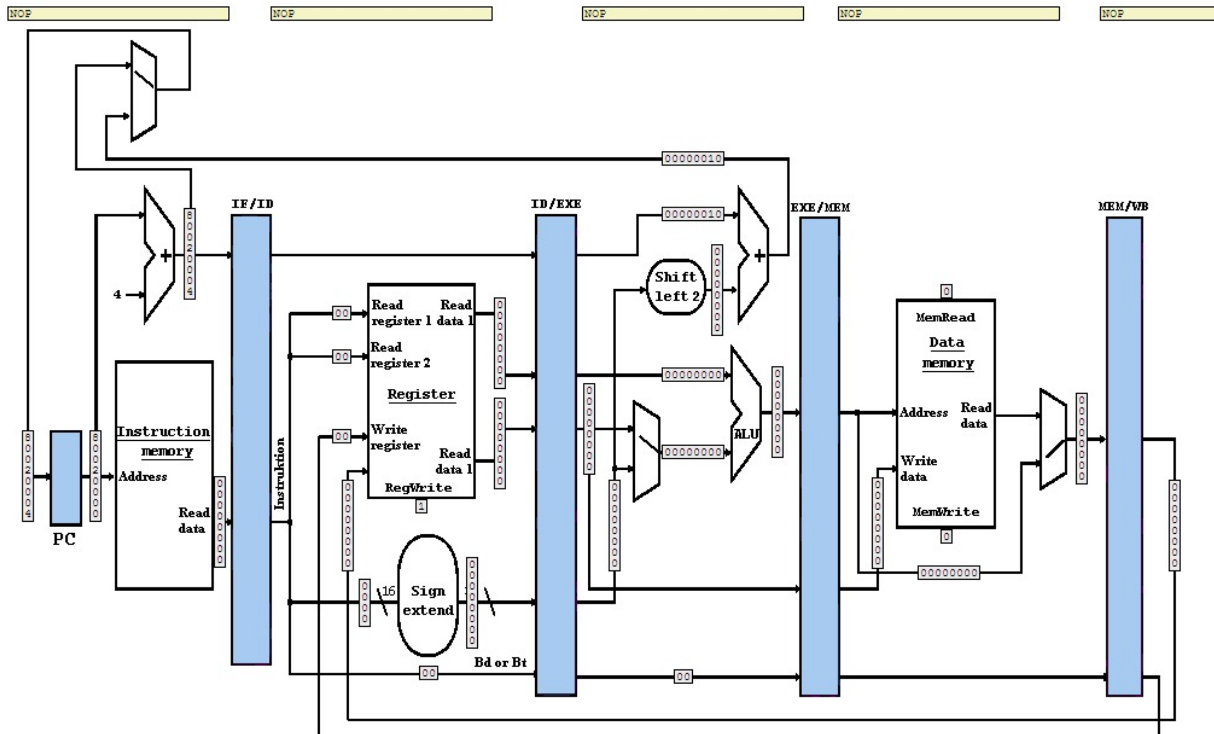

Figure 4: Typical MipsPipe2000 screen.

- What happens in the third (EX) stage?
- What happens in the fourth (MEM) stage?
- What happens in the fifth (WB) stage?
- What does IF, ID, EX, MEM, and WB mean?
- How many clock cycles does it take for the result of the operation to be available in the destination register?
- In which pipeline stages do different arithmetic instructions differ?
- One stage is not used by arithmetic instructions. Which one? Why?

1.5.2 Experiment 2

Replace the instruction **add t0, t1, t2** in the program above with the instruction **lw t0, 0(t1)** , build, upload, and investigate the program. Note that you must add a data variable from which to load a value.

- What happens in the different pipeline stages?
- What arithmetic operation does the ALU perform? Why?
- How many clock cycles does it take for the destination register to receive its value?
- Are all pipeline stages used? Explain!

1.5.3 Experiment 3

Now investigate the instruction **sw t0, 4(t1)** in the same way as with the other instructions above.

- What happens in the different pipeline stages?
- What arithmetic operation does the ALU perform? Why?
- How many clock cycles does it take for the memory to receive its value?
- Are all pipeline stages used? Explain!

1.5.4 Experiment 4

Finally, investigate the instruction **beq t0, t1, Dest** in the same way as with the other instructions above. Note that you must add a label named *Dest* somewhere.

- What happens in the different pipeline stages?
- What arithmetic operation does the ALU perform? Why? (There is a bug in simulator GUI - can you see the problem?)
- How many clock cycles does the instruction use?
- Are all pipeline stages used? Explain!

1.6 A Small Program Example

Pipelining can make processors run up to N times faster than when they are executed one at a time, where N is the number of pipeline stages. However, there are several effects that will cause problems and make it impossible to reach this efficiency in practice. One is that not all instructions will use all pipeline stages.

1.6.1 Experiment 1

So far we have been looking at one instruction at a time, but the real gain with using pipelining is to be able to overlap execution of several instructions. The net effect is that the hardware is busy executing several instructions *at the same time*. We will now investigate a small program with several instructions and determine how it goes through the pipeline so that the different stages of the instructions are executed in parallel. Study the following program.

```
1  #include <iregdef.h>
2
3          .set noreorder      # Avoid reordering instructions
4          .text               # Start generating instructions
5          .globl start        # The label should be globally known
6          .ent start          # The label marks an entry point
7
8  start:  lui     $9, 0xbf90  # Load upper half of port address
9                              # Lower half is filled with zeros
10
```

```
11  repeat: lbu     $8, 0x0($9) # Read from the input port
12          nop                 # Needed after load
13          sb      $8, 0x0($9) # Write to the output port
14          b       repeat      # Repeat the read and write cycle
15          nop                 # Needed after branch
16          li      $8, 0       # Clear the register
17
18          .end start          # Marks the end of the program
```

Upload the program above to the pipeline simulator and execute it step by step. Carefully note when the instructions are launched and when their results are ready. Also note how many instructions are in different stages of their execution at the same time.

1.6.2 Experiment 2

Run the following program on the pipeline simulator. Assign distinct values to t0, t1, and t3, and single step through the instructions.

```
1   #include <iregdef.h>
2
3           .set noreorder
4           .text
5           .globl start
6           .ent start
7
8   start:  add     t2, t0, t1
9           add     t4, t2, t3
10          nop
11          nop
12          nop
13
14          .end start
```

- After how many clock cycles will the destination register of the first add instruction, t2, receive the correct result value?
- After how many clock cycles is the value of t2 needed in the second instruction?
- What is the problem here? What is this kind of hazard called?

1.6.3 Experiment 3

This kind of problem can be solved by code reordering, introduction of **nop** instructions, stalling the pipeline (hazard detection), and by forwarding. Explain when the first three methods can be used and how they work.

Both the hardware MIPS processor and the simulator uses *forwarding* to solve problems with data hazards. So far, you have used a version of the pipeline, which does not have forwarding, S-script. Switch to the pipeline in the directory Xl-script. This version has forwarding too. Then single step through the program above and study how the forwarding works.

- How does the forwarding unit detect that forwarding should be used?
- From where to where is data forwarded in the case above?

1.6.4 Experiment 4

Run the following program on the pipeline simulator. Use the simple version without forwarding, (S-script). Assign the same value to t0 as to t1, and single step through the instructions.

```
1  #include <iregdef.h>
2
3          .set noreorder
4          .text
5          .globl start
6          .ent start
7
8  start:  nop
9          nop
10         beq     t0, t1, start
11         addi    t0, t0, 1
12         nop
13         nop
14         nop
15
16         .end start
```

- How many cycles does it take until the branch instruction is ready to jump? What has happened with the following **addi** instruction while the branch is calculated?
- What is the problem here? What is this kind of hazard called?
- What are the possible solutions to this problem?

  Switch to the forwarding version Xl-script again.
- How does this version handle **beq**?

1.6.5 Experiment 5

Run the following program on the pipeline simulator (simple version). Assign distinct values to t0 and t1, and let t2 contain the address to a memory location where you know the contents. Then single step through the instructions.

```
1  #include <iregdef.h>
2
3          .set noreorder
4          .text
5          .globl start
6          .ent start
7
8  start:  lw      t0, 0(t2)
9          add     t1, t1, t0
```

```
10          nop
11          nop
12          nop
13
14          .end start
```

- After how many clock cycles will the destination register of the load instruction, t0, receive the correct result value?

- After how many clock cycles is the value of t0 needed in the **add** instruction?

- What is the problem here? What is this kind of hazard called?

  This kind of problem can be solved with forwarding or hazard detection and stalling, just as other data hazards, but most MIPS implementations do not have these for load.

- What are the alternative solutions that can be used?

- Does the forwarding version of the simulator, Xl-script, handle the problem with delayed load?

1.7 Instruction Level Parallelism

(Use software tools (http://dark.eit.lth.se/) ScoreBoard.pl for these assignments - works best with Internet Explorer)

In this section we move from compiler-driven to hardware-driven optimizations. We will look at the effect of using the ScoreBoard algorithm. In this line we increase the amount of out-of-order execution to replace the pre-ordering we did so far. The instructions are administrated in three parts. In the first part, one sees the status of the respective instructions while they flow through the pipeline. In the second part, the status of the functional units are shown in nine fields:

- Operation; Operation to perform in the unit (e.g. add or sub)

- Busy; Indicates whether the unit is busy or not

- Fi; Destination register name

- Fj and Fk; Source register names

- Qj and Qk; Name of functional unit producing the data for the source registers

- Rj and Rk; Flags indicating whether the source registers have received their data

In the third part, the status of the registers is shown. Lastly, the history of the program execution is shown.

Record for each instruction in which clock-cycle it is in each of the traversed pipeline stages.

1.7.1 Experiment 1

For the first trial we look at the following program:

```
1 ld F6, 34(R2)
2 ld F2, 45(R3)
3 multd F0, F2, F4
4 subd F8, F6, F2
5 divd F10, F0, F6
6 addd F6, F8, F2
```

- After how many clock cycles can this program branch back to the beginning?
- Does re-ordering influence the execution time of this program and how?
- Is there a Write-after-Read hazard present and how is it solved?

1.7.2 Experiment 2

Another program that regularly has appeared during the lecture is the following (use first simulator not ScoreBoard.pl):

```
1 ld F0, 0(R1)
2 addd F4, F0, F2
3 sd F4, 0(R1)
4 ld F0, -8(R1)
5 addd F4, F0, F2
6 sd F4, -8(R1)
```

- After how many clock cycles can this program branch back to the beginning?
- Does re-ordering influence the execution time of this program and how?
- Is there a Write-after-Read hazard present and how is it solved?

1.7.3 Experiment 3

The last program that we will look at is the sum-of-products that appears in the Fast Fourier transform.

```
1 ld F0, 0(R1)
2 ld F2, 4(R1)
3 multd F8, F0, F2
4 ld F4, 8(R1)
5 ld F6, 10(R1)
6 multd F10, F4, F6
7 addd F10, F8, F10
8 ld F8, 12(R1)
9 addd F10, F10, F8
```

- After how many clock cycles can this program branch back to the beginning?
- Does re-ordering influence the execution time of this program and how?
- Is there a Write-after-Read hazard present and how is it solved?

1.7.4 Experiment 4 - for interested students

Interested students can do the same experiments using the Tomasulo algorithm. Note any difference? Please explain why!

1.8 Conclusions

Before you pass the laboratory exercise, think about the questions below and explain to your supervisor:

- How can a pipelined processor be faster than one without pipelining?
- What are the special problems that appear in pipelining?
- How can these problems be solved?
- Is there a difference in writing compilers for pipelined processors?
- Which is faster, straight code or code with many branches?
- What does RISC mean? What does CISC mean?
- Is the Pentium processor pipelined? AMD Phenom? Intel Core2? ARM Cortex A?

1.9 Literature

Basic computer organization literature such as:

- D. A. Patterson and J. L. Hennessy: Computer Organization and Design - the Hardware/Software Interface, Morgan Kaufmann.

  or

  M. Brorsson: Datorsystem - program och maskinvara, Studentlitteratur, 1999.

and our course literature:

- David Patterson and John Hennessy, “Computer Architecture: A Quantitative Approach”, 4th edition, Kaufmann, 2006, ISBN 978-0-12-370490-0.

2 Laboratory 2: Advanced pipelining

2.1 Goals

After this laboratory exercise, you should understand how program behavior (instruction class profiles) relates to branch prediction efficiency, as well as trade-offs related to their implementation. You should also have an understanding for the relative importance of various advanced pipeline techniques like branch prediction, variable pipeline width and out-of-order execution.

2.2 Literature

Hennessy and Patterson: Appendix A, Chapter 2-3
Section A.3 of this lab-manual

2.3 Preparations

Read section A.3 on simulation and the SimpleScalar tool-set thoroughly. You should be able to answer the home assignment questions. Read through this laboratory assignment and make sure that you have sufficiently familiarized yourselves with the required concepts in pipelining and branch prediction.

Note: In the course of these labs you will run a large number of simulations, and it may be difficult to keep track of your results unless you maintain a lab book (hard copy or on-line). This book should contain the description of all the simulation runs you performed, your simulation plans, comparison of results, graphs if any etc. In addition as you will start using more detailed simulators, the simulation time will increase. A lab book which documents all the simulation runs you performed already will help you avoid repeat runs, and will save you considerable time. The system allows you to attach an arbitrary ID to each run, as well as it saves all results within a session (typically one lab). Use this intelligently in order to keep better track of what you have done.

2.3.1 Home Assignment 1

Answer the following questions:

- What is the role of simulators in processor design?
- Why is it advantageous to have several different simulators?
- For the four branch prediction schemes '`taken|perfect|bimod|comb`', describe the predictor. Your description should include:
  - What information the predictor stores (if any)?
  - How the prediction is made?
- What is out-of-order execution?
- What is the difference between scoreboarding and Tomasulo?

2.4 Program behavior (instruction profiling)

Start a Web-browser (works best with Internet Explorer) and go to the initial lab-page (http://dark.eit.lth.se/) and login with your EFD-id. After the lab is finished and you have recorded all your measurements you should logout of the system.

When you have logged out all your results are unavailable so be sure to record them first!

Choose three of the available benchmarks from the '`Program to run`' menu (they are briefly described in section A.3.3), and run the profiling simulator for each of them, to find out the distribution of instruction classes.

Fill table 1 with all available benchmark programs versus instruction class profiles. (Get values from other groups for those you didn't run yourself!)

benchmark	load	store	uncond branch	cond branch	integer computation	fp computation
anagram						
go						
compress						
applu						
mgrid						
swim						
perl						
gcc	25.8				40.5	

Table 1: Benchmark programs versus instruction class profiles in %

Choose one of the benchmarks for your further assignments based on the following considerations:

- Is your benchmark memory intensive or computation intensive?
- Is your benchmark mainly using integer or floating point?
- What percentage of the instructions executed are conditional branches? Given this percentage, how many instructions on average does the processor execute between each pair of conditional branch instructions (do not include the conditional branch instructions).
- Using your textbook, class notes, other references, and your own opinion, list and explain several reasons why the performance of a processor like the one simulated by sim-outorder (e.g., out-of-order-issue superscalar) will suffer because of conditional branches. For each reason also explain how, if at all, a branch predictor could help the situation.

2.5 Branch Predictors

2.5.1 Experiment 1

We will now use the branch prediction simulator (sim-bpred) to investigate the effects of branch predictors on the execution of your benchmark. This simulator allows you to simulate 5 different types of branch predictors. You can see the list of them by looking at the menu '`branch predictor type`' for the branch prediction simulator sim-bpred. (`bimod` is the 2-bit prediction scheme, figure 2.4, in the course-book. '2lev' is a two level adaptive branch predictor. 'comb'

combines a bimod and a 2-level predictor. The detailed simulator (sim-outorder) also implements the 'perfect' predictor which always make a correct prediction.)

For three of the possible branch prediction schemes, '`nottaken|taken|bimod`', run the simulation for your benchmark as you did above and note the branch prediction statistics for each in table 2.

benchmark	**nottaken**	**taken**	**bimod**
anagram			
go			
compress			
applu			
mgrid			
swim			
perl			
gcc			

Table 2: Branch prediction statistics

Note: the simulator statistics are for all branches - both conditional and unconditional (which are regarded as predicted correctly). For this reason the reported prediction rates for `taken` and `nottaken` do not add to 1. Use the branch-direction measurements and number of updates, both corrected for unconditional branches to calculate accuracy (*hit rate for conditional branches*).

2.5.2 Experiment 2

Use the detailed simulator (sim-outorder) to measure and describe how the prediction rate (bpred_dir_rate) effects the processor CPI (sim_CPI) for your benchmark. Also use this simulator to measure CPI when using the `perfect` branch predictor type. This simulator produces a lot of text (Simulation Statistics) as the result of a simulation, but you can use the browser search function on the result-page in order to find the desired result.

	taken		bimod		comb		perfect	
Benchmark	Branch pred. rate	CPI	Branch pred. rate	CPI	Branch pred. rate	CPI	Branch pred. rate	CPI

Table 3: Branch prediction rate versus CPI

There is a bug in the simulator, such that branch prediction rates for taken/nottaken are wrong. For the taken column only use the CPI from the simulation and use the rates from table 2!

(Again fill in values in table 3 for other benchmarks with the help of other groups.)

For the four branch prediction schemes '`taken|perfect|bimod|comb`', describe the predictor. Your description should include:

- What information the predictor stores (if any)?
- How the prediction is made?
- What the relative accuracy of the predictor is compared to the others.

2.6 Choosing a new branch strategy

Suppose you are choosing a new branch strategy for a processor. Your design choices are:

1. predict branches taken with a branch penalty of 2 cycles and a 1200 MHz clock-rate
2. predict branches taken with a branch penalty of 3 cycles and a 1300 MHz clock-rate
3. predict branches using bimod with a branch penalty of 4 cycles and a 900 MHz clock-rate
4. predict branches using bimod with a branch penalty of 4 cycles and a 1000 MHz clock-rate and half the L1 cache size

Hint - fill in table 4 for your chosen benchmark, with cycle count (sim_cycle) and CPI from detailed simulations. Calculate the execution time using the given clock-rate.

Benchmark		Alternative 1	Alternative 2	Alternative 3	Alternative 4
	Sim_cycle				
	CPI				
	Exe time				
	Sim_cycle				
	CPI				
	Exe time				
	Sim_cycle				
	CPI				
	Exe time				

Table 4: Impact of different branch strategies

Questions:

- What would be your choice for your benchmark? Why?
- How much do you have to be able to increase the clock frequency in order to gain performance when allowing a branch miss-prediction latency of 3 cycles instead of 2 when using the taken predictor?
- Compare your results with other groups using other benchmark programs and discuss your observations.

2.7 In-order vs out-of-order issue

Now you will conduct experiments to find out how the increase in the parallelism in processing instructions affects the CPI of your processor, and how you can improve the performance of memory reference instructions.

In all experiments you will use the default cache and branch predictor configurations.

2.7.1 Experiment 1

Experiment with in-order and out-of-order issue, and the width of the pipeline, by running the simulation with the following combinations of parameters. Measure CPI and total no of cycles. (**Note:** out-of order issue and execution is default. in-order issue is selected with the check-box labeled '`run pipeline with in-order issue`').

- Pipeline width 1, in-order and out-of-order issue
- Pipeline width 4, in-order and out-of-order issue
- Pipeline width 8, in order and out of order issue

Benchmark		Pipeline width=1		Pipeline width=4		Pipeline width=8	
		Sim_cycle	CPI	Sim_cycle	CPI	Sim_cycle	CPI
	Out-of-order						
	In-order						
	Out-of-order						
	In-order						

Table 5: In-order and out-of-order issue versus pipeline width

Questions:

- What is the impact on CPI of the increased pipeline width?
- Explain the impact and their difference for both in-order and out-of-order issue.

2.7.2 Experiment 2

Run the sim-outorder simulator varying the number of memory ports available: 1,2 and 4. Use a pipeline width of 4.

Benchmark		Memory port=1		Memory port=2		Memory port=4	
		Sim_cycle	CPI	Sim_cycle	CPI	Sim_cycle	CPI
anagram	Out-of-order						
anagram	In-order						
compress	Out-of-order						
compress	In-order						

Table 6: In-order and out-of-order issue versus memory ports

Questions

- What is the impact on CPI of the increase in available memory ports?

2.8 Conclusion

Before you pass the laboratory exercise, think about the questions below and explain to your supervisor:

- Why is bimodal branch prediction more expensive to implement than predict nottaken?

- Why is bimodal better than nottaken?
- What, if any, is the impact on CPI by allowing out-of-order issue?
- What, if any, is the impact on CPI by allowing more instructions to be processed in one cycle?
- Is the wider pipeline more effective with in-order or out-of-order issue, and if so - why?

3 Laboratory 3: Cache memory

3.1 Goals

A cache memory is a memory that is smaller but faster than the main memory. Due to the locality of memory references, the use of a cache memory can give the effect on the computer system that the apparent speed of the memory is that of the cache memory, while the size is that of the main memory. The actual efficiency gained by using a cache memory varies depending on cache size, block size, and other cache parameters, but it also depends on the program and data. In short, everything depends on a proper parametrization.

After this laboratory exercise, you should understand the basic principles of cache memories, and how different parameters of a cache memory affects the efficiency of a computer system.

3.2 Literature

Hennessy and Patterson: Appendix C, Chapter 5
MIPS Lab Environment Reference Manual (section A.2 in this lab-manual)
Cache tutorial at http://www.ecs.umass.edu/ece/koren/architecture/Cache/tutorial.html

3.3 Preparations

Use the cache tutorial to familiarize yourself with cache concepts and terminology.

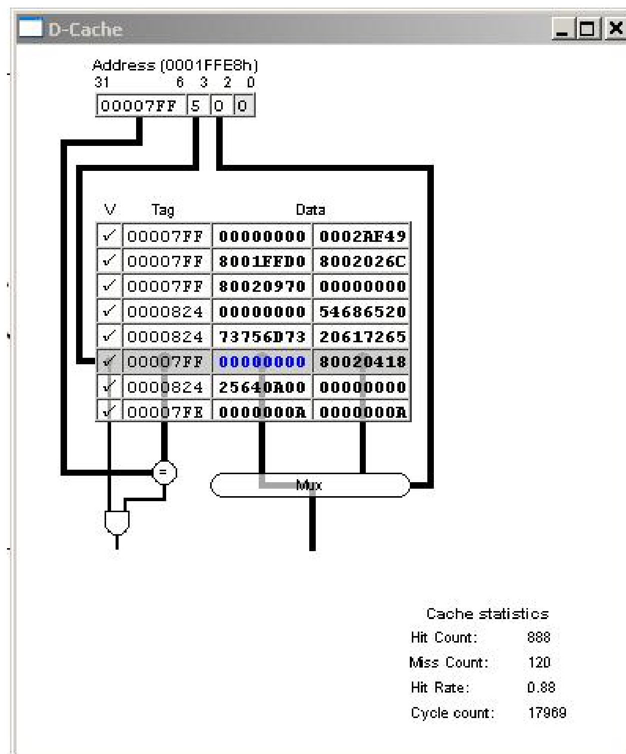

Figure 5: Data cache organization (see also HP fig C.5). (Note there is a bug in the cache simulator that makes the cache statistics counters wrap around at 100 000.)

Read the literature and this laboratory exercise in detail, and solve the home assignments. Note that you must solve the home assignments, or you will not be allowed to start the laboratory

exercise.

It is mandatory to be familiarized with cache concepts and terminology. As a first indication of being initiated on the subject, you should be able to solve conceptual issues like:

3.3.1 Home Assignment 1

Show the address bit partitioning for a memory system with
Memory size = 32MB
Cache size = 64 KB Block size = 16 Bytes
Set associative with 8 blocks per set. (What is a block?)

3.3.2 Home Assignment 2

In which order is the micro-operations done during a successful cache access (cf HP figure C.5).

- choose cache block
- valid address and divide into fields
- data sent to CPU
- compare tag with address field
- choose word within cache block

3.3.3 Home Assignment 3

Explain the following: cache size, block size, number of sets, write policy, and replacement policy.

3.3.4 Home Assignment 4

The following C program contains two subroutines which returns the sum of all the matrix cells. The only difference between the two subroutines is that they visit the matrix elements in a different order. This may seem unimportant, but with a cache memory, it may make a big difference.

Study the C-program carefully so that you understand in which order the matrix elements are used. The C-programs used here are available for downloading from the 'Software tools for Computer Architecture' page at http://dark.eit.lth.se/.

```
#include <stdio.h>
#include <idt_entrypt.h>
#define N 10

int A[N][N];

int SumByColRow (int Matrix[N][N])
{
  int i, j, Sum = 0, Time;

  flush_cache();
  timer_start();
```

```
  for (j = 0; j < N; j ++) {
    for (i = 0; i < N; i ++) {
      Sum += Matrix[i][j];
    }
  }
  Time = timer_stop();
  printf("SumByColRow time: %d\n", Time);
  return Sum;
}

int SumByRowCol (int Matrix[N][N])
{
  int i, j, Sum = 0, Time;

  flush_cache();
  timer_start();
  for (i = 0; i < N; i ++) {
    for (j = 0; j < N; j ++) {
      Sum += Matrix[i][j];
    }
  }
  Time = timer_stop();
  printf("SumByRowCol time: %d\n", Time);
  return Sum;
}

main ()
{
  int a, b;

  printf ("Laboratory Exercise, Home Assignment\n");
  // Run one of the cases below; comment out the other
  // Case 1
  a = SumByColRow (A);
  printf ("The sum is %d\n", a);
  // Case 2
  b = SumByRowCol (A);
  printf ("The sum is %d\n", b);
}
```

3.4 Missing and hitting in a cache

Create your project in the MIPS IDE, type in the above C program, build and upload it to the cache simulator (program Mips in the `mips` catalogue (S:)). Run the program in the cache simulator **with the default settings** and study how the instruction cache works. Fill in the first line of table 8.

Hint: Wall-clock time is smaller if I-cache and D-cache windows are closed.

	I-cache	**D-cache**	**Memory**
Default setting	cache size = 16 block size = 2 blocks in sets = 1	cache size = 16 block size = 2 blocks in sets = 1	read cycles = 50, write cycles = 50 writepolicy=WriteThrough write buffersize = 0 replacementpolicy=Random

Table 7: Default settings of the simulator (All penalty enabled).

Questions:
Give *full* answers!

- How is the full 32 bit address used in the cache memory?
- What happens when there is a cache miss?
- What happens when there is a cache hit?
- How large is the block size?
- What is the function of the tag?

3.5 Parametrization

The parameters of the cache memory can be changed to test the effects of different cases.

Run the program in the cache simulator with the settings in Table 8, and study how the instruction cache works. Record the cycles of cache miss/hit, and total cycles according to the table.

Questions:

- Investigate the effects of different parameter settings. Explain the following: cache size, block size, number of blocks in sets, write policy, and replacement policy.
- If a cache is large enough that all the code within a loop fits in the cache, how many cache misses will there be during the execution of the loop? Is this good or bad? How should the code look like that would benefit the most from a large block size?

Hint: If you don't see any clear difference - change the code to increase the measured difference.

3.6 Optimized parametrization

Compile the C-program with the compiler option 'Optimization high'. Run the program in the in the cache simulator with the settings in Table 9, and study how the data cache works.

Questions:

- Study the subroutines SumByColRow and SumByRowCol . Explain carefully in what order the memory addresses are visited by the two subroutines.
- Execute the program and study how many cache hits the two subroutines have. Is there a difference? Why?
- Also speculate if you can detect in which order the matrix elements are located in physical memory.

Settings		I-cache Hit rate	D-cache Hit rate	Simulation time
All	SumByColRow			
default	SumByRowCol			
I-cache and D-cache cache size =32	SumByColRow			
Others: default	SumByRowCol			
I-cache and D-cache blockSize =8	SumByColRow			
Others: default	SumByRowCol			
I-cache and D-cache Number of blocks in sets =2	SumByColRow			
Others: default	SumByRowCol			
D-cache writepolicy=WriteBack	SumByColRow			
Others: default	SumByRowCol			
D-cache replacementpolicy=FIFO	SumByColRow			
Others: default	SumByRowCol			

Table 8: Settings for parameter experimentation.

3.7 Matrix multiplication

Create a new project with the following C program. Compile your code with the compiler option **Optimization high**.

```
#include <stdio.h>
#include <idt_entrypt.h>
#define N 10
int A[N][N];

int initMatrix (int Matrix[N][N])
{
  int i, j;
```

Settings		I-cache Hit rate	D-cache Hit rate	Simulation time
I-cache: **Disable Penalty(*)**	SumByColRow			
D-cache: **cache size=64** **block size= 16** **Others: Default**	SumByRowCol			

Table 9: Settings for optimization.

```
  for (i = 0; i < N; i ++) {
    for (j = 0; j < N; j ++) {
      Matrix[i][j] = i*N+j;
    }
  }
  return 0;
}

int SumOfProdByRowCol (int Matrix[N][N])
{
  int i, j, k, r, Sum = 0, Time;
  flush_cache();
  timer_start();
  for (i = 0; i < N; i ++) {
    for (j = 0; j < N; j ++) {
      r = 0;
      for (k = 0; k < N; k = k + 1)
        r = r + Matrix[i][k]*Matrix[k][j];
      Sum += r;
    }
  }
  Time = timer_stop();
  printf("SumOfProd time: %d\n", Time);
  return Sum;
}

int main ()
{
  int a;
  initMatrix(A);
  printf ("Laboratory Assignment Matrix multiplication\n");
  a = SumOfProdByRowCol(A);
  printf ("The sum of products is %d\n", a);
}
```

Introduce a blocking factor and change the program accordingly to localize the operations. Run the new version and explain the differences in performance. One of the sample codes like this:

```
#define B 4
#define min(X,Y) (X>Y?Y:X)

int SumOfProdByRowCol_Blockfactor (int Matrix[N][N])
{
  int i, j, k, r, Sum = 0, Time;
  int jj, kk;
  flush_cache();
  timer_start();
  for (jj = 0; jj < N; jj = jj + B)
```

```
    for (kk = 0; kk < N; kk = kk + B)
      for (i = 0; i < N; i ++) {
        for (j = jj; j < min(jj+B, N); j ++) {
          r = 0;
          for (k = kk; k < min(kk+B, N); k = k + 1)
            r = r + Matrix[i][k]*Matrix[k][j];
          Sum += r;
        }
      }
  Time = timer_stop();
  printf("SumOfProd time: %d\n", Time);
  return Sum;
}
```

Simulate both of the programs with the settings shown in table 10.

Settings		I-cache Hit rate	D-cache Hit rate	Simulation time
I-cache: **Disable Penalty(*)**	Without Blocking factor			
D-cache: **block size= 4** **Others: Default**	With Blocking factor			

Table 10: Settings for matrix multiplication program.

Questions:

- What is the blocking factor and how does it work? Draw diagrams to illustrate how the two programs work!
- How many methods do you know to reduce the cache miss rate?

3.8 Reflections/Conclusion

You should now be able to converse on typical cache dimensioning problems, like:

- What is the general idea with cache memory?
- How does block size affect the efficiency of a cache?
- How fast is a cache memory and a DRAM memory in relation to each other?
- Does the optimal cache parameters depend on the program code?
- How can one select good cache parameters?

4 Laboratory 4: Advanced cache; Processor design trade-offs

4.1 Goals

After this laboratory exercise, you should have deeper understanding of how various cache parameters affects performance. Finally you should get an insight into the difficulties and many trade-offs that goes into designing a CPU.

4.2 Literature

Hennessy and Patterson: Chapter 2, 5; Appendix A, C
Section A.3 of this manual

4.3 Preparations

Read section A.3 on simulation and the SimpleScalar tool-set thoroughly.

Make sure that you have sufficiently mastered the concepts in chapters 2, 5 and appendixes A, C of Hennessy and Patterson. You can test some of this by answering the home assignments.

Note: In the course of these lab you will run a large number of simulations, and it may be difficult to keep track of your results unless you maintain a lab book (hard copy or on-line). This book should contain the description of all the simulation runs you performed, your simulation plans, comparison of results, graphs if any etc. In addition as you will start using more detailed simulators, the simulation time will increase. A lab book which documents all the simulation runs you performed already will help you avoid repeat runs, and will save you considerable time. The system allows you to attach an arbitrary ID to each run, as well as it saves all results within a session (typically one lab). Use this intelligently in order to keep better track of what you have done.

4.3.1 Home Assignment 1

- What are the four main categories of cache performance optimizations? Relate these to the formula for average memory access time.
- Which of these categories does associativity affect?
- Which of these categories does block size affect?

4.3.2 Home Assignment 2

- How is associativity, number of blocks, number of sets and cache size related?
- How does these affect the average access time for L1- and l2-caches?

4.3.3 Home Assignment 3

- State as many factors as possible relevant for evaluating and choosing a CPU design for a specific application (like gaming or hearing aids).

4.4 Cache performance

Start a Web-browser and go to the initial lab-page (http://dark.eit.lth.se/) and login with your EFD-id. After the lab is finished and you have recorded all your measurements you should logout of the system.

When you have logged out all your results are unavailable so be sure to record them first!

Use a single run of sim-cheetah to simulate the performance of the following cache configurations for two different benchmarks.

- Unified cache (Reference stream to analyze)
- least-recently-used (LRU) replacement policy
- 16 to 1024 sets
- 1-way to 8-way associativity
- 32-byte cache blocks

Note: sim-cheetah provides results for a continuous range of associativity, in this case 1, 2, 3, 4, 5, 6, 7 and 8. In your analysis of cache behavior ignore the measurements for associativity which is not a power of two, ie. consider only associativity of 1, 2, 4 and 8.

- Using the output from sim-cheetah, for caches of equivalent size, verify if increasing associativity or the number of sets in the cache gives the most benefit. To do so **produce graphs** showing changes in miss rate as associativity/no of sets changes. Matlab routines for producing graphs (see section A.4) can be downloaded from the 'Software tools for Computer Architecture' page at http://dark.eit.lth.se/.
- Repeat this for data only cache, and for instructions only cache ('`Reference stream to analyze`').

4.4.1 Relation block size, miss ratio, and mean access time

Run simulations <sim-outorder> for two different benchmarks with the following configurations:

- unified L1 cache with a size of 32 KB, associativity 2 and block sizes 16, 32, 64, 128, 256 bytes.
- The L2 data cache should be a fixed configuration with a total size of 512 KB and a block-size of at least 256 (choose reasonable parameters).
- Keep other parameters as default.

Note: Remember to set all the cache parameters (program, L2 data cache, unified L1 and L2 caches) *for each simulation*. Record data in table 11 (Hint - hit times for L1 and L2 (cache:dl1lat, cache:dl2lat) are given in the simulation statistics. Since L2 configuration is not changed during your experiment you can estimate a fixed number for L2 miss penalty using L2 block-size, memory latency and memory access bus width.)

- Make plots that show block size vs CPI, and average memory access time vs block size. Matlab routines for producing plots (see section A.4) can be downloaded from the 'Software tools for Computer Architecture' page at http://dark.eit.lth.se/.
- How does average access time vary with block size?

L2 data cache			
fixed total size	No of sets	block size (≥ 256)	associativity
512 KB			

L1 data cache					
L1 size	L1 assoc	No of sets	block size	CPI	average memory access time
32 KB	2		16		
32 KB	2		32		
32 KB	2		64		
32 KB	2		128		
32 KB	2		256		

Table 11: L1 and L2 data cache

4.5 Processor design trade-offs

When designing a processor you have to make trade-offs in order to achieve the best performance for a given cost. The absolute performance measure is the execution time with respect to a program. However, this is not realistic to measure due to an enormous amount of programs. Therefore, CPI (clock cycle per instruction) and clock speed are common parameters used to evaluate the performance of a processor. The design cost may be measured in various respects, such as silicon area, power consumption, fabrication (technology) cost, design effort, and compiler design complexity. Different applications have different constraints on the design cost. For instance, battery-powered embedded processors often have very tight power budget; silicon area and fabrication cost have to be well controlled for processors in consumer devices; whereas high-end gaming processors have much relaxed cost budget, as the performance is their primary goal.

Table 12 gives some processor design choices with corresponding unit costs based on 130 nm silicon technology. Cost measures are gross approximations which vary widely with the silicon technology used to implement the chip. Table 13 lists cost scaling factors for 90 nm and 65 nm silicon technologies with respect to the 130 nm technology. Smaller silicon technology has gains in area, clock speed and switching energy, while it has higher fabrication cost and static power consumption (power consumption without any circuit activity, caused by leakage current of CMOS transistors).

Resource	Area	Clock degradation	Switching power	Value 'N'
Pipeline width	4×N	Free	1×N	1, 2, 4
Integer ALU	1×N	2%	0.2×N	0∼4
Integer multiplier	2×N	5%	0.5×N	0∼4
FP ALU	3×N	8%	0.5×N	0∼4
FP multiplier	4×N	10%	0.8×N	0∼4
Memory port	3×N	2%	0.5×N	1∼4
Taken/Nottaken predictor	Free	Free	Free	Yes/No
Bimod predictor	3	2%	0.5	Yes/No
In-order issue	Free	Free	Free	Yes/No
Out-of-order issue	2×(above area)	10%	1	Yes/No

*'Free' indicates that this feature is part of a base-design, or has no effect on the cost measures.

Table 12: Processor design choices with unit costs based on 130 nm technology.

N is the pipeline width, so a pipeline of width 2 takes an area of $4 \times 2 = 8$ units. Adding one FP-ALU to this CPU increases the area with $3 \times 1 = 3$ units etc.

Technology	Fabrication [1]	Area	Clock [2, 3]	Switching power [2]	Static power [3]
130 nm	1	1	1	1	1
90 nm	2.39	0.69	1.68	0.29	1.96
65 nm	2.43	0.50	2.50	0.09	9.46

Table 13: Cost scaling factors for 130 nm, 90 nm, and 65 nm silicon technologies.

4.5.1 Investigate design optimization

- For your chosen benchmark, optimize your configuration of the pipeline functional units, pipeline width and memory ports using chip area as the cost of interest according to Table 12. Adding up the area costs from each resource, and the total chip area has to be less than or equal to **60**.

  For example, a design with a pipeline width of 2, 4 integer ALUs, 1 integer multiplier, 1 FP ALU, 1 FP multiplier, 2 memory ports, bimod predictor and out-of-order issue would have an area of 60 as seen below

	area/unit	# units (N)	area
Pipeline width	$4 \times N$	2	$4 \times 2 = 8$
integer ALU	$1 \times N$	4	$1 \times 4 = 4$
integer multiplier	$2 \times N$	1	$2 \times 1 = 2$
FP ALU	$3 \times N$	1	$3 \times 1 = 3$
FP multiplier	$4 \times N$	1	$4 \times 1 = 4$
memory port	$3 \times N$	2	$3 \times 2 = 6$
bimod predictor	3		3
sum			30
out-of-order issue	$2 \times (sum)$		$2 \times 30 = 60$

  Fill in Table 14 with your design choices.

- Use simulations both to guide your design and verify your results. Note that simulating all possible reasonable configurations would take several hundreds of simulations, which is not realistic.

  Hint: Consider the instruction distribution of the benchmarks you selected in an earlier lab.

- Describe your design and motivate your choices. Did performance vary much with different configurations?

4.5.2 Example

(In this assignment you can use values from tables 12,13 aswell.)

Now let us consider other cost factors as well when designing processors. As an example, assuming we are going to design an embedded processor for a hearing aid device, which is

Resource	Design 1	Design 2	Design 3
Pipeline width			
Integer ALU			
Integer multiplier			
FP ALU			
FP multiplier			
Memory port			
Taken/Nottaken predictor			
Bimod predictor			
In-order issue			
Out-of-order issue			
Area			
CPI			

Table 14: Your design choices and results.

battery powered and has to be fit into human ears. These requirements set constraints on power consumption and silicon area for a processor design. Based on the fact that hearing aid devices usually have regular data processing operations (linear filtering), we chose to exclude advanced pipeline functional units and all float point arithmetic units to keep the chip area low. Table 15 lists our design choice for the hearing aid processor based on 130 nm technology. Table 16 compares design costs in different silicon technologies. From results shown in Table 16, 65 nm technology seems to be the best choice for smallest silicon footprint and lowest switching power consumption. However, this is not entirely correct, as we should also consider static power consumption of our circuit. Static power consumption is the power cost when circuit is being idle. Because we don't want our batteries to die out before using the hearing aid device, we would like to keep this value as small as possible. Refer to Table 13, 90 nm technology has 4.83 times less static power consumption compared to 65 nm technology, at the expense of 1.38 times more silicon area and 3.22 times more switching power consumption. So using 90 nm technology for our processor in this application might result in a good design compromise between area, switching and static power consumption. In reality, people often use different power control techniques to bring down the static power consumption, such as multiple power domains during chip design, hierarchical power management, etc.

4.5.3 Processor design trade-offs versus application requirements

Design two new processors based on the given requirements below. Enter your design in Table 17. Frequency and switching power are linearly dependent ($P = \alpha_c * V_{dd}^2 * f$).

- Design a processor which will be used in pacemakers. Because patients do not want to have operations often performed just to change batteries inside their pacemaker, it is required that internal circuits must be extremely power efficient. Here we assume that the switching power consumption of the embedded processor inside pacemaker should not draw more than **0.02** units, and assume there are power management units that could reduce static power consumption. Note that pacemakers typically operate on a speed of human heart rate, so clock frequency of the embedded processor is less important in this application.
- Design a processor which will be used for weather forecasting. Because weather forecast needs to process quantitative data collected from current state of the atmosphere that

Resource	Design choice	Area (from Table 12)
Pipeline width	2	8
Integer ALU	2	2
Integer multiplier	2	4
FP ALU	0	0
FP multiplier	0	0
Memory port	2	6
Taken/Nottaken predictor	Yes	0
Bimod predictor	No	0
In-order issue	Yes	0
Out-of-order issue	No	0
Total area		20

Table 15: An example of a hearing aid processor in 130 nm technology.

Technology	Area	Clock speed*	Switching power
130 nm	20	364 MHz	4.4
90 nm	13.85	613 MHz	1.54
65 nm	10	910 MHz	0.5

*Assume a baseline processor is able to run at 400 MHz in 130 nm technology. Considering the total 9% clock degradation of the example processor, clock speed would be $400 \times (1 - 9\%) = 364$ MHz.

Table 16: Cost comparisons of the hearing aid processor design.

involves data analysis, statistic calculations, etc., it is a computationally intensive application. Here we require that the processor should operate at **700 MHz** or faster.

Note: Assume a baseline processor is able to run at 400 MHz in 130 nm technology.

Table 17: Your design choices and results.

Resource	Pacemaker	Weather forecasting
Pipeline width		
Integer ALU		
Integer multiplier		
FP ALU		
FP multiplier		
Memory port		
Taken/Nottaken predictor		
Bimod predictor		
In-order issue		
Out-of-order issue		
Technology		
Area		
Switching power		
Clock speed		

References

[1] Europractice, UMC standard, *http://www.europractice-ic.com/general_prices.php*, 2011.

[2] ITRS, *http://www.itrs.net/Links/2001ITRS/PIDS.pdf*, 2001.

[3] ITRS, *http://www.itrs.net/Links/2003ITRS/PIDS2003.pdf*, 2003.

Conclusions

Before you pass the laboratory exercise, think about the questions below and explain to your supervisor:

- What is the relative gain for the various performance enhancements techniques used in this lab?
- Are the techniques investigated in this laboratory (cache organization, pipeline modifications) independent?
- Is there a "**best**" processor design?

A Appendix: Software

A.1 Pipeline simulator mipspipe2000.exe

See "Mips Lab Environment Reference Manual" section "Pipeline" available from 'Course Material' on the course Web-pages.

A.2 Cache simulator (mips.exe)

See "Mips Lab Environment Reference Manual" available available from 'Course Material' on the course Web-pages.

A.3 SimpleScalar simulator tool set

A.3.1 Getting Started with the SimpleScalar Tool Set

Based on the manual by Ewa Z. Bem, School of Computing and Information Technology, University of Western Sydney Nepean, which was based on the manual by Todd M. Bezenek, University of Wisconsin

Introduction

This document contains background material about the SimpleScalar toolset of simulators used in the Computer Architecture lab. SimpleScalar itself is available for download together with various tools and utilities including detailed documentation from http://www.simplescalar.com/

SimpleScalar and Simulation in Computer Architecture

When computer architecture researchers work to improve the performance of a computer system, they often use an existing system to simulate a proposed system. Although the intent is not always to measure raw performance (estimating power consumption is one alternative), performance estimation is one of the most important results obtained by simulation. The SimpleScalar tool set is designed to measure the performance of several parts of a superscalar processor and its memory hierarchy. This document describes the SimpleScalar simulators. Other simulation systems may be similar or very different.

Overview of SimpleScalar Simulation

The SimpleScalar tool set includes a compiler that creates binaries for a non-existent processor. The binaries can be executed on one of several simulators that are included in the tool set. This section describes the goals of processor simulation.

The execution of a processor can be modelled as a series of known states and the time (or other costs, ie., power) required to make the transition between each pair of states. The state information may include all or a subset of:

- The values stored in all memory locations.
- The values stored in and the status of all cache memories.
- The values stored in and the status of the translation-lookaside buffer (TLB).
- The values stored in and the status of the branch prediction table(s) or branch target buffer (BTB).

- All processor state (ie. the pipeline, execution units (integer ALU, load/store unit, etc.), register file, register update unit (RUU), etc.)

A good way to evaluate the performance of a program on a proposed processor architecture is to simulate the state of the architecture during the execution of the program. By simulating the states through which the processor will pass during the execution of a program and estimating the time (or other measurement) necessary for each state transition, the amount of time that the simulated processor will need to execute the program can be estimated.

The more state that is simulated, the longer a simulation will take. Complex simulations can execute 100s of times slower than a real processor. Therefore, simulating the execution of a program that would take an hour of CPU time on an existing processor can take a week on a complex simulator. For this reason, it is important to evaluate what measurements are desired and limit the simulation to only the state that is necessary to properly estimate those measurements. This is the reason for the inclusion of several different simulators in the SimpleScalar tool set.

Profiling

In addition to estimating the execution time of a program on the simulated processor, profile information may be of use to computer architects. Profile information is a count of the number or frequency of events that occur during the execution of a program. One common example of profile data is a count of how often each type of instruction (ie., branch, load, store, ALU operation, etc.) is executed during the running of a program.

Profile information can be used to gauge the relative importance of each part of a processor's implementation in determining its performance when executing the profiled program.

The SimpleScalar Base Processor

The SimpleScalar tool set is based on the MIPS R2000 processor's instruction set architecture (ISA). The processor is described in MIPS RISC Architecture by Gerry Kane, published by Prentice Hall, 1988. The ISA describes the instructions that the processor is capable of executing - and therefore the instructions that a compiler can generate - but it does not describe how the instructions are implemented. The implementation is what computer architects change in order to improve the performance of a processor.

An existing processor can be chosen as a base processor for several reasons. These may include:

- The architecture of the processor is well known and documented.

- The architecture of the processor is state-of-the-art and therefore it is likely to be useful as a base for the study of future processors.

- The architecture of the processor has been implemented as a real processor, allowing simulations to be compared to executions on a real, physical processor.

An important consideration in the choice of the MIPS architecture for the SimpleScalar tool set was the fact that the GNU GCC compiler was available in source-code form, and could compile to the MIPS architecture. This allowed the use of this public-domain software as part of the SimpleScalar tool set.

Description of the Simulators

The SimpleScalar tool set includes a number of simulators designed for various purposes. They are described below. For those simulators we are using there are also a description of the important profiling options available.

sim-bpred This simulator implements a branch predictor analyser.

sim-cache This simulator implements a functional cache simulator. Cache statistics are generated for a user-selected cache and TLB configuration, which may include up to two levels of instruction and data cache (with any levels unified), and one level of instruction and data TLBs. No timing information is generated.

sim-cheetah This program implements a functional simulator driver for Cheetah. Cheetah is a cache simulation package written by Rabin Sugumar and Santosh Abraham which can efficiently simulate multiple cache configurations in a single run of a program. Specifically, Cheetah can simulate ranges of single level set-associative and fully-associative caches.

```
#-option <args>   # <default> # description
-refs    <string> #      data # reference stream to analyze, {none|inst|data|unified}
-R       <string> #       lru # replacement policy, i.e., lru or opt
-C       <string> #        sa # cache configuration, i.e., fa, sa, or dm
-a       <int>    #         7 # min number of sets (log base 2, line size for DM)
-b       <int>    #        14 # max number of sets (log base 2, line size for DM)
-l       <int>    #         4 # line size of the caches (log base 2)
-n       <int>    #         1 # max degree of associativity to analyze (log base 2)
-in      <int>    #       512 # cache size intervals at which miss ratio is shown
-M       <int>    #    524288 # maximum cache size of interest
-c       <int>    #        16 # size of cache (log base 2) for DM analysis
```

Note that 'line size' above is the same as block size. Most of the parameters above are give as log base 2 of the number, ie a line size of 16 bytes is given as '`-l 4`.

sim-fast This simulator implements a very fast functional simulator. This functional simulator implementation is much more difficult to digest than the simpler, cleaner sim-safe functional simulator. By default, this simulator performs no instruction error checking, as a result, any instruction errors will manifest as simulator execution errors, possibly causing sim-fast to execute incorrectly or dump core. Such is the price we pay for speed!!!!

sim-outorder This simulator implements a very detailed out-of-order issue superscalar processor with a two-level memory system and speculative execution support. This simulator is a performance simulator, tracking the latency of all pipeline operations.

```
# -option        <args>       # <default> # description
-fetch:ifqsize   <int>        #         4 # instruction fetch queue size (in insts)
-fetch:mplat     <int>        #         3 # extra branch mis-prediction latency
-bpred           <string>     #     bimod # branch predictor type
                                          # {nottaken|taken|perfect|bimod|2lev|comb}
-bpred:bimod     <int>        #      2048 # bimodal predictor config (<table size>)
-decode:width    <int>        #         4 # instruction decode B/W (insts/cycle)
```

```
-issue:width     <int>        #       4 # instruction issue B/W (insts/cycle)
-issue:inorder   <true|false> #   false # run pipeline with in-order issue
-issue:wrongpath <true|false> #    true # issue instructions down wrong execution paths
-commit:width    <int>        #       4 # instruction commit B/W (insts/cycle)
-cache:dl1       <string>     # dl1:128:32:4:l # l1 data cache config
-cache:dl1lat    <int>        #       1 # l1 data cache hit latency (in cycles)
-cache:dl2       <string>     # ul2:1024:64:4:l # l2 data cache config
-cache:dl2lat    <int>        #       6 # l2 data cache hit latency (in cycles)
-cache:il1       <string>     # il1:512:32:1:l # l1 inst cache config
-cache:il1lat    <int>        #       1 # l1 instruction cache hit latency (in cycles)
-cache:il2       <string>     #     dl2 # l2 instruction cache config
-cache:il2lat    <int>        #       6 # l2 instruction cache hit latency (in cycles)
-mem:lat         <int list...># 18 2 # memory access latency (<first_chunk> <inter_chunk>)
-mem:width       <int>        #       8 # memory access bus width (in bytes)
-tlb:itlb        <string>     # itlb:16:4096:4:l # instruction TLB config
-tlb:dtlb        <string>     # dtlb:32:4096:4:l # data TLB config
-tlb:lat         <int>        #      30 # inst/data TLB miss latency (in cycles)
-res:ialu        <int>        #       4 # total number of integer ALU's available
-res:imult       <int>        #       1 # total number of integer multiplier/dividers availabl
-res:memport     <int>        #       2 # total number of memory system ports available (to CP
-res:fpalu       <int>        #       4 # total number of floating point ALU's available
-res:fpmult      <int>        #       1 # total number of floating point multiplier/dividers a

The cache config parameter <config> has the following format:

  <name>:<nsets>:<bsize>:<assoc>:<repl>

  <name>   - name of the cache being defined
  <nsets>  - number of sets in the cache
  <bsize>  - block size of the cache
  <assoc>  - associativity of the cache
  <repl>   - block replacement strategy, 'l'-LRU, 'f'-FIFO, 'r'-random

  Examples:   -cache:dl1 dl1:4096:32:1:l
              -dtlb dtlb:128:4096:32:r

Cache levels can be unified by pointing a level of the instruction cache
hierarchy at the data cache hiearchy using the "dl1" and "dl2" cache
configuration arguments.  Most sensible combinations are supported, e.g.,

  A unified l2 cache (il2 is pointed at dl2):
    -cache:il1 il1:128:64:1:l -cache:il2 dl2
    -cache:dl1 dl1:256:32:1:l -cache:dl2 ul2:1024:64:2:l

  Or, a fully unified cache hierarchy (il1 pointed at dl1):
    -cache:il1 dl1
    -cache:dl1 ul1:256:32:1:l -cache:dl2 ul2:1024:64:2:l
```

sim-profile This simulator implements a functional simulator with profiling support.

```
# -option   <args>       # <default> # description
-nice       <int>        #         0 # simulator scheduling priority
-max:inst   <uint>       #         0 # maximum number of inst's to execute
-all        <true|false> #     false # enable all profile options
-iclass     <true|false> #     false # enable instruction class profiling
-iprof      <true|false> #     false # enable instruction profiling
-brprof     <true|false> #     false # enable branch instruction profiling
-amprof     <true|false> #     false # enable address mode profiling
-segprof    <true|false> #     false # enable load/store address segment profiling
-tsymprof   <true|false> #     false # enable text symbol profiling
-taddrprof  <true|false> #     false # enable text address profiling
-dsymprof   <true|false> #     false # enable data symbol profiling
-internal   <true|false> #     false # include compiler-internal symbols during symbol profilin
```

sim-safe This simulator implements a functional simulator. This functional simulator is the simplest, most user-friendly simulator in the simplescalar tool set. Unlike sim-fast, this functional simulator checks for all instruction errors, and the implementation is crafted for clarity rather than speed.

The sim-cache and sim-cheetah simulators simulate only the state of the memory system–they do not keep track of the timings of events. The sim-outorder simulator does. In fact, it simulates everything that happens in a superscalar processor pipeline, including out-of-order instruction issue, the latency of the different execution units, the effects of using a branch predictor, etc. Because of this, sim-outorder runs more slowly, but it also generates much more information about what happens in a processor.

Because sim-outorder keeps track of timing, it can report the number of clock cycles that are needed to execute the given program for the simulated processor with the given configuration.

A.3.2 Running simulation experiments with SimpleScalar

A Web user interface to run simple experiments using SimpleScalar simulators is available at http://dark.eit.lth.se/darklab/

It uses sessions based on a ID (eg your EFD-login) given as login ID at the start to be able to keep track of all the simulations done during the laboratory session.

The main screen (figure 6) allows you to set simulator specific options (defaults are filled in if appropriate), choose which program and which simulator to run. Only a selection of all options are available through this user interface. It also provides access to all results produced earlier in this session. Futhermore it provides functionality for compiling a few programs with a special version of GCC that produces code that the simulator can run.

Note: In the course of these labs you will run a large number of simulations, and it may be difficult to keep track of your results unless you maintain a lab book (hard copy or on-line). This book should contain the description of all the simulation runs you performed, your simulation plans, comparison of results, graphs if any etc. In addition as you will start using more detailed simulators, the simulation time will increase. A lab book which documents all the simulation runs you performed already will help you avoid repeat runs, and will save you considerable time. The system allows you to attach an arbitrary ID to each run, as well as it saves all results within a session (typically one lab). Use this intelligently in order to keep better track of what you have done.

Welcome, tt, to Computer Architecture Simulation Labs

Profile program execution - sim-profile

Program to run: anagram
Short description of benchmarks

Profiling options

all profile options	instruction class	instruction	branch instruction
☐	☐	☐	☐

Simulation ID: [ ] Start simulation

Branch prediction analyses - sim-bpred

Program to run: anagram branch predictor type: bimod
Simulation ID: [ ] Start simulation

Detailed simulation including timing - sim-outorder

Program to run anagram	run pipeline with in-order issue and no speculation: ☐	branch predictor type: bimod	extra branch mis-prediction latency: 3
number of integer ALU's: 4	number of integer multiplier/dividers: 1	number of floating point ALU's: 4	number of floating point multiplier/dividers: 1
Pipeline width: 1	number of memory system ports available to CPU: 2	memory access bus width (in bytes): 8	
L1 data cache: # sets: 128 block size: 32 associativity: 4 algorithm: LRU	L2 data cache: # sets: 1024 block size: 64 associativity: 4 algorithm: LRU	L1 instruction cache: Fully unified: ☐ (*) # sets: 512 block size: 32 associativity: 1 algorithm: LRU	L2 instruction cache: Unified L2: ☐ (**) # sets: 128 block size: 32 associativity: 4 algorithm: LRU

Simulation ID: [ ] Start simulation

Cache levels can be unified by pointing a level of the instruction cache hierarchy at the data cache hiearchy using the "L1 data" and "L2 data" cache configuration arguments
Most sensible combinations are supported, e.g.,

(**) A unified L2 cache - L2 instruction cache is pointed at L2 data cache. L2 instruction cache parameters are disregarded.

(*) Or, a fully unified cache hierarchy - L1 instruction cache is pointed at L1 data cache. L1 and L2 instruction cache parameters are disregarded.

Cache measurements - sim-cheetah

Program to run anagram	Reference stream to analyze Instructions	Replacement policy ○ LRU; ○ Optimal
min number of sets	max number of sets	max degree of associativity

Figure 6: Main screen

A.3.3 Available benchmarks

There is more information available online linked from the main screen.

anagram A program for finding anagrams for a phrase, based on a dictionary.

compress (SPEC) Compresses and decompresses a file in memory.

go (SPEC) Artificial intelligence; plays the game of "Go" agianst itself

applu (SPEC) Parabolic/elliptic partial differential equations

mgrid (SPEC) Multi-grid solver in 3D potential field

swim (SPEC) Shallow water model with 1024x1024 grid

perl Calculates popularity of nodes in a graph based on the PageRank algorithm from Google.

gcc (SPEC) Limited version of GCC

A.3.4 Test programs for compilation

http://dark.eit.lth.se/darklab/anagram.txt
http://dark.eit.lth.se/darklab/stride.txt

A.4 MatLab routines for plotting results

These routines can be downloaded from the 'Software tools for Computer Architecture' page at http://dark.eit.lth.se/.

Use for Laboratory 4 assignment 'Cache performance':

```
set_assoc = [1, 2, 3, 4, 5, 6, 7, 8];
miss_rate_u = [ ]; % miss rates obtained for unified caches
miss_rate_i=[ ]; % miss rates obtained for instruction caches
miss_rate_d=[ ]; % miss rates obtained for data caches
%nbr_sets = [16, 32, 64, 128, 256, 512, 1024];

figure(1);
plot(set_assoc, miss_rate_u(1,:),'bd-',set_assoc, miss_rate_u(2,:), 'cs-', set_assoc,
 miss_rate_u(3,:), 'y^-',set_assoc, miss_rate_u(4,:), 'mx-',set_assoc, miss_rate_u(5,:),
'r+-',set_assoc, miss_rate_u(6,:), 'go-', set_assoc, miss_rate_u(7,:), 'kh-');
legend('16', '32', '64', '128', '256', '512', '1024');
grid on;
title('mgrid (Unified)'); % to be changed depending on the used benchmark
xlabel('Set Associativity (blocks/set)');
ylabel('Miss Rate');

figure(2);
plot(set_assoc, miss_rate_i(1,:),'bd-',set_assoc, miss_rate_i(2,:), 'cs-', set_assoc,
miss_rate_i(3,:), 'y^-',set_assoc, miss_rate_i(4,:), 'mx-',set_assoc, miss_rate_i(5,:),
'r+-',set_assoc, miss_rate_i(6,:), 'go-', set_assoc, miss_rate_i(7,:), 'kh-');
legend('16', '32', '64', '128', '256', '512', '1024');
grid on;
title('mgrid (Instruction)'); % to be changed depending on the used benchmark
xlabel('Set Associativity (blocks/set)');
ylabel('Miss Rate');

figure(3);
plot(set_assoc, miss_rate_d(1,:),'bd-',set_assoc, miss_rate_d(2,:), 'cs-', set_assoc,
miss_rate_d(3,:), 'y^-',set_assoc, miss_rate_d(4,:), 'mx-',set_assoc, miss_rate_d(5,:),
'r+-',set_assoc, miss_rate_d(6,:), 'go-', set_assoc, miss_rate_d(7,:), 'kh-');
legend('16', '32', '64', '128', '256', '512', '1024');
grid on;
title('mgrid (Data)'); % to be changed depending on the used benchmark
xlabel('Set Associativity (blocks/set)');
ylabel('Miss Rate');
```

Use for Laboratory 4 assignment 'Relation block size, miss ratio and mean access time':

```
ht_l1; % Hit Time for L1
```

```
ht_l2; % Hit Time for L2
mp_l2; % Miss Penalty for L2
block_size = [16, 32, 64, 128, 256];
cpi  = [];
mr_l1 = []; % Miss Rate for L1
mr_l2 = []; % Miss Rate for L2
AMAT = ht_l1 + mr_l1.*(ht_l2 + mr_l2 .* mp_l2);%Avarage Memory Access Time

figure(1);
plot(block_size, cpi,'bd-');
legend('L2 cache 128:128:4');% to be changed depending on the settings for L2
set(gca,'xtick',block_size);
title('applu'); % to be changed
grid on;
xlabel('Block size (Bytes)');
ylabel('CPI');

figure(2)
plot(block_size, AMAT,'bd-');
legend('L2 cache 128:128:4');% to be changed depending on the settings for L2
set(gca,'xtick',block_size);
title('applu');% to be changed
grid on;
xlabel('Block size (Bytes)');
ylabel('Average Memory Access Time');
```