Technische Universiteit
Eindhoven
University of Technology

Department of Electrical Engineering
Electronic Systems Group

SIMD Made Explicit in LLVM

Preparation Report

Guus Hendrikus Peter Leijsten

Supervisors:
Prof. dr. ir. Henk Corporaal
Dr. ir. Roel Jordans
Ir. Luc Waeijen

Draft version

Eindhoven, March 2017

Abstract

THIS IS MY ABSTRACT

Keywords: Compilers, SIMD, LLVM, Embedded Systems, Power Efficiency, Explicit Bypassing

Contents

List of Figures

List of Tables

Listings

List of Abbreviations

ALU Arithmetic Logical Unit
AST Abstract Syntax Tree
CP Control Processor
DAG Directed Acyclic Graph
DLP Data Level Parallelism
ES Group Electronic Systems Group
FU Functional Unit
ID Instruction Decode
IF Instruction Fetch
ILP Instruction Level Parallelism
IR Intermediate Representation
ISA Instruction Set Architecture
I-type Immediate-type
J-type Jump-type
LSU Load Store Unit
MIPS Microprocessor without Interlocked Pipeline Stages
N/A Not Applicable
PE Processing Element
pJ peta-Joule
RA Register Allocation
RF Register File
RISC Reduced Instruction Set Computer
SIMD Single Instruction Multiple Data
R-type Register-type
SSA Static Single Assignment
TTA Transport Triggered Architecture
VLIW Very Long Instruction Word
WB Write Back

Chapter 1

Introduction

Nowadays, mobile phones have dedicated processors to support video processing. This embedded streaming processor consumes tens of pJ per operation (pJ/op) and the battery capacity is only sufficient for playing video applications for a few hours [1]. Furthermore, embedded systems like mobile devices have to run high performant applications like video codecs, wireless signal processing and 3D processing [2]. These kind of devices often have a limited power source, and because it is a handheld device, heat produced by power dissipation is a main concern. For these reasons, power efficiency is becoming the bottleneck in the design of such embedded systems.

In general, it is important to improve both performance and energy efficiency. The Very Long Instruction Word (VLIW) architecture is one example architecture designed to improve performance by executing multiple instructions in parallel, exploiting a program's Instruction Level Parallelism (ILP). By exploiting parallelism, the processor requires less cycles to do the same amount of work, thereby improving performance.

The Transport Triggered Architecture (TTA) is similar to the VLIW architecture. However, instead of packing the operations in a single instruction, TTAs pack multiple transports in a single instruction [3]. For VLIWs, each Register File (RF) is connected to every Function Unit (FU). Unlike VLIW, TTAs do not require that each FU has their own private connections to the RFs. Instead, an FU is connected to the RF by means of an interconnect. Another advantage that TTA has over VLIW is that it has explicit datapaths. With explicit datapaths, software bypassing is possible. The compiler can eliminate some RF accesses, as we will show in Chapter 3.

Applications like image or video processing often have a high amount of Data Level Parallelism (DLP). The advantage of the Single Instruction Multiple Data (SIMD) architecture is that it naturally exploits DLP by processing multiple operations in parallel instead of processing them in a sequence. Therefore, the same performance can be achieved at a much lower clock frequency, thereby reducing energy consumption [2]. The SIMD architecture also has large register files that consume around 35% of the total energy consumption [2]. We want to further reduce energy consumption by adding explicit bypassing on top of the SIMD in order to eliminate some RF accesses. Adding explicit bypassing on top of the SIMD architecture resembles the TTA. Namely SIMD with explicit bypassing has explicit datapaths, like TTA also has explicit datapaths.

The ES group has implemented a compiler for the SIMD architecture. This compiler exploits explicit bypassing [2] and can compile a subset of OpenCL and C code [4]. However, this compiler has a custom backend, and in order to improve maintainability, we want to use a compiler framework that supports our needs. To this end we use the LLVM framework to design and build a compiler for the proposed SIMD architecture.

There is another master's project going on about writing a compiler for the proposed SIMD architecture using the LLVM framework [5]. However, this compiler focusses on vector instructions and does not support explicit bypassing. Therefore, we will focus on adding explicit bypassing on top of the new compiler. Furthermore, the proposed SIMD architecture is designed to be configurable. Another master's project will investigate in how to generate hardware code for different combinations of parameters.

1.1 The SIMD Processor Architecture

The advantage of the SIMD architecture is that multiple operations are processed in parallel instead of processing them in a sequence. Therefore, the same performance can be achieved at a much lower clock frequency, thereby reducing energy consumption [2]. Furthermore, because each Processing Element (PE) executes the same instruction, the Instruction Fetch (IF) and Instruction Decode (ID) can be shared amongst the PEs, reducing energy consumption.

We propose a wide SIMD architecture [1] that performs wide vector operations that exploit DLP by executing the same instruction on multiple data simultaneously. Figure 1.1 shows a general overview of the SIMD processor. We have one Control Processor (CP) responsible for scalar operations and control flow e.g. jump/branch instruction. Furthermore there is a wide array of PEs responsible for vector operations. The CP executes in parallel with the PEs, exploiting ILP.

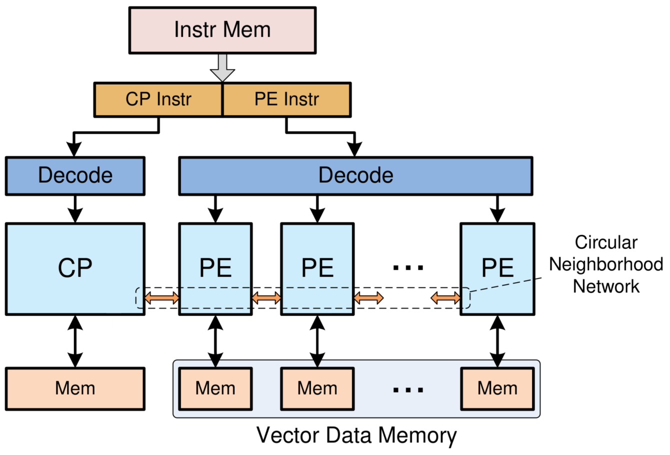


Figure 1.1: General overview of the wide SIMD architecture.

The proposed architecture has a Reduced Instruction Set Computer (RISC) Instruction Set Architecture (ISA) that is divided up in three categories of instructions. In general, instructions have two operands and a destination register. Instructions that take two register files as operands are Register-type (R-type) instructions. Instructions that take a register file and an immediate as operands are called Immediate-type (I-type). The control flow can be controlled by using Jump-type (J-type) instructions, which can only be executed by the CP.

Furthermore, the architecture is designed to be configurable, e.g. width of the PE array, bit width of the wires and registers, the number of stages that the instruction pipeline consists of, and whether it has implicit or explicit bypassing, can be configured. The datawidth of the wires and registers can be configured into 16-bits or 32-bits.

In order to support a configurable number of PE elements, a neighbourhood network topology is chosen for its scalability. With a circular neighbourhood network topology, the connection between the first and last PE does not introduce extra long wires, because the PEs can be places in a circular manner [4].

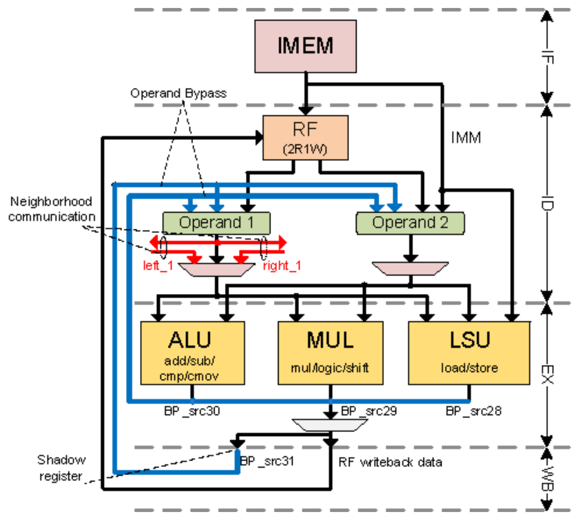


Figure 1.2: 4-stage pipeline processor overview.

1.1.1 Processor Pipeline and Datapath

Generally, each processor (CP or PE) has its own registers and three functional units, i.e. ALU, MUL and LSU.

The instruction pipeline is divided up in four or five stages. Top down, we have an IF-stage, an ID-stage, one or more execution stages and a Write Back (WB) stage. The architecture shown in Figure 1.2 has four stages while the architecture shown in Figure 1.3 has five stages.

The neighbourhood communication network is implemented by overriding the output of *Operand* 1 in ID-stage. Depending on the decoded instruction, data is either selected from another (neighbouring) processor, or from itself. Each FU has private input registers, which keep the result at

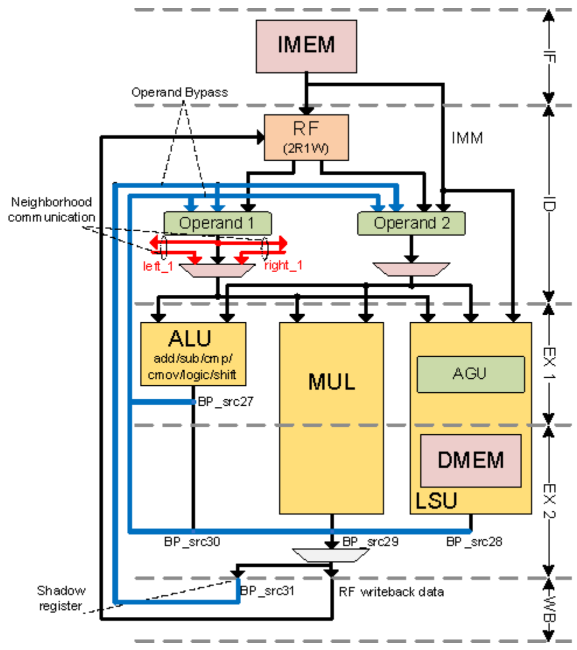
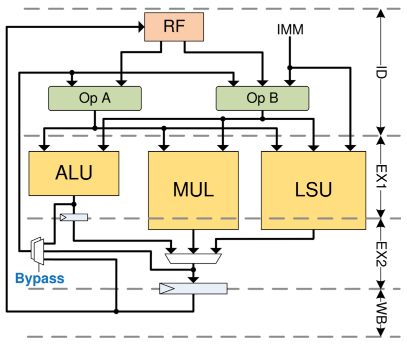


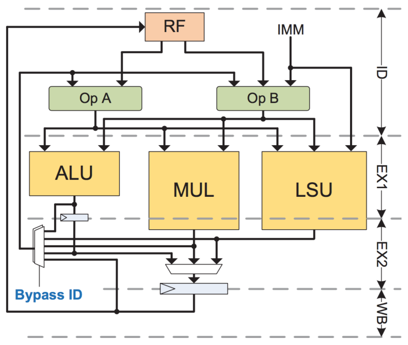
Figure 1.3: 5-stage pipeline with processor overview.

the output of a compute unit valid as long as no new operation or input is assigned to it [2]. The outputs can be used in the bypass network to bypass any of the operands in an instruction.

We can configure the SIMD to have either explicit, or implicit bypassing. With implicit bypassing, also called transparent bypassing it is the hardware's responsibility to handle bypassing. With explicit bypassing on the other hand, it is the compilers responsibility to handle bypassing.



(a) Datapath with implicit bypassing.



(b) Datapath with explicit bypassing.

Figure 1.4: Bypassing network differences between implicit bypassing and explicit bypassing.

One of the advantages of explicit bypassing is that when possible a write to a register can be avoided. Thus reducing the total energy consumption of the register file. Since there are many register files in a wide SIMD, reducing the energy consumption of the register file has a large impact on the overall energy consumption [2]. Because of this, reducing the register file's energy consumption is of great importance. Furthermore, the explicit datapath shown in Figure 1.4b has two extra sources compared to the transparent datapath in Figure 1.4a. These additional bypass sources increase the chance that a result is being bypassed. In the explicit bypassing version, bypassing sources are directly accessible by the instruction. This is done by reserving part of the RF address space for the bypass sources. The disadvantage of this is that the register index space is reduced, however we do not have to change the instruction format in order to specify that an operand of an instruction is bypassed from a previous instruction.

Table 1.1: Special purpose registers, bypassing sources are indicated with BP_src.

Register	**Purpose**
R_0	Constant value zero.
R_1	PE_ID (PE only).
R_3 and R_4	Return value registers.
R_5 through R_8	Argument passing.
R_9	Link register (CP only).
R_{10}	Frame pointer.
R_{11}	Stack pointer.

The total number of registers grows linearly with the number of PEs because each processor has 32 registers. With a wide SIMD, we therefore have many registers that in total consume a considerable amount of energy, namely 35% of the total energy consumption [2]. Some of these registers have a special purpose, e.g. register R_0 is connected to ground is has a static value of 0, these are shown in Table 1.1.

1.2 The LLVM Infrastructure

The LLVM project started in 2000 by Chris Lattner, as a research project at the University of Illinois with the goal of providing a modern, Static Single Assignment (SSA)-based compilation strategy capable of supporting both static and dynamic compilation of arbitrary programming languages. It was first released in 2003, although the project has grown rapidly since then. It has become popular amongst major companies, e.g. Google, Apple and Sony, for its powerful multi-stage compilation strategy and outstanding extendibility. LLVM is a collection of modular and reusable compiler and toolchain technologies. Generally, LLVM follows a 3-phase design, which is divided up in a frontend, a code independent optimizer and a backend, illustrated in Figure 1.5.

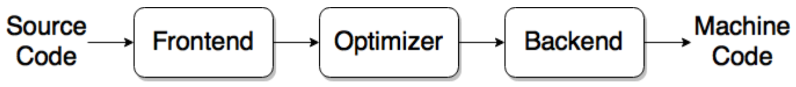


Figure 1.5: 3-phase design: frontend, optimizer and backend.

The frontend is responsible for translating code of an arbitrary language into LLVM's Intermediate Representation (IR) code. The LLVM instruction set represents a virtual architecture that captures the key operations of ordinary processors, but avoids machine specific constraints such as physical registers. Instead, it has an infinite amount of virtual registers in SSA form, which means that each virtual register is assigned only once and each use of a variable is dominated by that variable's definition. This simplifies the dataflow optimizations because only a single definition can reach a particular use of a value, and finding that definition is trivial [6]. As the first phase of a compiler, the main task of the lexical analyzer is to read the input characters of the source program, group them into luxemes, and produce as output a sequence of tokens. These tokens are used by the parser for syntax analysis, where we verify that the sequence of tokens can be reconstructed according to the syntax of the input language. The parser should report any syntax errors during this process and should be able to recover in order to continue processing the rest of the program. The parser constructs a parse tree, and the semantic analyzer uses this parse tree to check for consistency with the language definition. Type checking is also done during this stage, and the information is kept in the syntax tree. The result of these phases is an Abstract Synctax Tree (AST) of the program, which can be translated into three-address IR code.

The optimizer contains a collections of analysis and semantic-preserving transformations that can be used to optimize IR code. One of the advantages of LLVM is that when you build a new backend for any given processor architecture you immediately have access to all of these optimizations. Below we give some of these optimizations that are explained more detailed in literature [7].

- *Constant propagation* computes for each point and each variable in the program, whether that variable has a unique constant value at that point. This can then be used to replace variable references with constant values.
- *Constant folding* recognizes and evaluates constant expressions at compile time rather than runtime. For example, '$add\ 1 + 2$' can be replaced by '3'. Statements like '$add\ 1 + 2$' can be introduced by other optimizations, e.g. constant propagation.
- *Common sub-expression elimination* recognizes that the same expressions appears in more than one place, and that performance can be improved by transforming the code such that the expression appears in one only place.
- *Copy propagation* replaces each target of a copy statement with that of the copied value. For example, if we have a copy statement, $x = y$. Then the uses of x can be replaced by y. Some optimizations require that this optimizations is performed afterwards to cleanup, e.g. common sub-expression elimination requires this pass to run afterwards.

- *Dead code elimination* removes code that do not affect the programs results. This avoids executing irrelevant operations and reduces the code size of a program.
- *Loop invariant code motion* aims at moving code that is independent of the loop iteration out of the loop body. It does this by moving the loop independent statement above the loop, saving it in a temporary variable, and use it in each iteration of the loop. Now the loop independent statement is computed only once instead of every iteration.
- *Function inlining* verifies whether inlining functions in its callees gives a performance benefit. If doing this would give performance benefit, it replaces the call of the function with the function body. This optimization often is useful for small functions because it reduces the overhead that is introduced when a function call is made, e.g. storing frame pointer, storing function parameters and jump in code to where the function is defined.

The backend translates, according to a processor architecture, IR code to a target specific assembly language. It does this by going through a sequence of code generation stages, illustrated in Figure 1.6. The rectangular boxes indicate the data structure that is used by, and produced by a given stage, and the name of each stage is denoted in a rectangular box with rounded corners. During this process, first the IR code is lowered to a Directed Acyclic Graph (DAG) in which each node represents one instruction. However, for some architectures, not all data types and instructions are supported. For this reason, the DAG is legalized to something that is supported by the target architecture. Instruction selection maps each of the nodes into machine nodes, by matching patterns. Then we have a DAG consisting of only target specific machine instructions, in SSA form. Having naive machine instruction, the next step is to schedule them. We schedule the machine instructions according to the resource information of the target processor, and assign each instruction to a specific cycle. We will discuss scheduling in more detail in Chapter 1.2.1. Now the instructions are represented in a list rather than a DAG, but still in SSA form. The Register Allocator (RA) then assigns physical registers to each of the virtual registers, now the list is not in SSA form. We will discuss RA in more detail in Chapter 1.2.2. The post-allocation pass can improve the schedule, taking the physical registers and register pressure, that is known at

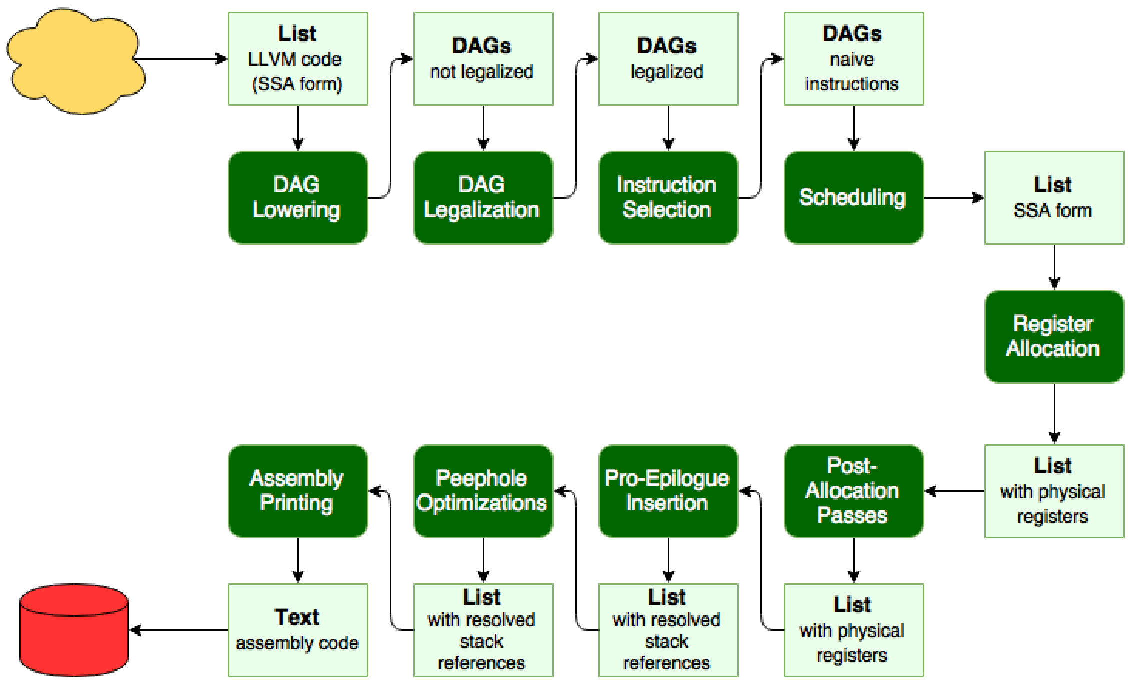


Figure 1.6: Code generation sequence, from LLVM code to assembly code.

this point, into account. After that, some epilogue and prologue code might need to be inserted, e.g. saving/restoring the caller/callee registers and reserving/destroying of the function's stack frame. The peephole optimizations are target specific improvements to the schedule that has been constructed. These optimizations deal with very specific optimizations that can only be done at the end of the process. At last, the assembly printer, prints the assembly code.

1.2.1 Instruction Scheduling

1.2.2 Register Allocation

Register allocation is executed during the code generation phase and consists of finding a mapping of a program with an unlimited number of virtual registers to a program with a limited number of physical registers.

1.3 Problem statement

An SIMD architecture has been designed. The compiler that has been build is too custom and is difficult to maintain. For this reason, a new compiler is being implemented in LLVM. Having the proper tool support we hope to improve maintainability and perhaps also efficiency. Our goal for this thesis is to reduce energy consumption of the SIMD architecture by using explicit datapaths so that we can reduce the pressure on the register files. Since SIMD are known to have a lot of registers, we hope to significantly reduce energy consumption with this method.

We will implement and evaluate different approaches that will be further elaborated in Chapter 3.

To summarise, we want to implement and evaluate different approaches to exploit explicit datapaths for the SIMD architecture, using the LLVM framework.

1.4 Overview

In the remainder of this report we will discuss related works in Chapter 2. We will discuss software bypassing by showing some basic examples in Chapter 3. Proposed solutions for how to achieve our goals will be discussed in Chapter 4. We will provide a planning for the thesis in Chapter 5 and early conclusions are given in Section 6.

Chapter 2

Related Work

Related work goes here.

Chapter 3

Software Bypassing

In Chapter 1.1.1 we have briefly discussed the bypassing sources of the proposed architecture. For convenience we have given them again in Table 3.1.

Table 3.1: Alias for each bypassing source, BP_src in Figure 1.2 and Figure 1.3.

Register:	**Bypass source:**	**Alias:**	
		4 stages	**5 stages**
R_{27}	BP_src27	N/A	ALU1
R_{28}	BP_src28	ALU	ALU2
R_{29}	BP_src29	MUL	MUL
R_{30}	BP_src30	LSU	LSU
R_{31}	BP_src31	WB	WB

With software bypassing the compiler can avoid some accesses to RFs by explicitly specifying the datapath.

Listing 3.1: Example code fragment where no bypassing is specified.

```
add r3, r1, r2 ; r3 = r1 + r2
sub r5, r3, r4 ; r5 = r3 - r4
```

Listing 3.1 shows a code fragment with an addition and a subtraction. There is a flow dependent dependency, namely the result of the addition is used in the subtraction.

Listing 3.2: Example code fragment where the operand is bypassed.

```
add r3, r1, r2  ; r3 = r1 + r2
sub r5, ALU, r4 ; r5 = r3 - r4
```

Listing 3.2 shows the same code fragment, however we have bypassed the result of the addition to the subtraction. Now the subtraction does not require a read access to *r3*, because we can bypass the result of the flow-dependent instructions.

Listing 3.3: Example code fragment avoiding a write access.

```
add --, r1, r2  ; r3 = r1 + r2
sub r5, ALU, r4 ; r5 = r3 - r4
```

When the result of *r3* is not needed anymore, the write access can be removed. By specifying −− as destination, the result will not be written back, as illustrated in Listing 3.3. The freed register can be used for other calculations, which may lead to higher performances.

Introduceer bypassing over loop iterations, example, conclusion below example.

Listing 3.4: Example code fragment bypassing through loop iteration.

```
      lw r6, r10, 5
loop: addrr r7, r6, %src1
        ⋮
      sflts r6, 0
      addri r6, r6, -1
      bf loop
      nop
        ⋮
```

Here we want to apply bypassing to bypass the result of the last add instruction in the loop before we branch to the first instruction in the loop. However, when we go in the loop for the first time, $r6$ comes from the LSU instead of the ALU. Therefore, sometimes it is necessary to insert an instruction before we enter a loop to bypass over loop iterations.

Chapter 4

Proposed Solutions

Proposed solutions: phase ordering problem, discuss each possible implementation here, ordered by difficulty. Indicate that in the available timespan, we can not implement all of these solutions.

Post RA pass to allocate bypass sources by demoting operands to bypass registers where possible. When a bypass is allocated, it replaces the register that it used to take as operand, therefore effectively freeing a register. However in the post-RA stage, spill code has already been inserted, this will not benefit in terms of freeing registers, unless we move spill code explicitly. This will be be the easiest approach, because even without moving the spill code, this method will still work, but with too early inserted spill code as a result.

Bypass-aware register allocator that allocates bypass sources as much as possible, and then allocates any virtual registers not bypass as any normal register allocator. In this approach the freed registers can be utilized which should give overall better results.

If we allocate bypasses after scheduling, but before register allocation, we can change a virtual register into a bypass register when possible. However, during register allocation or any other passes that follow, code might be inserted that break a previously allocated bypass. For this reason, we will need another pass that will identify these broken bypasses, and change them back into virtual registers so that they are dealt with by the register allocator.

Before scheduling, we can identify instruction pairs that we can bypass. Then we change the virtual register that is bypassed into physical (bypass) registers and glue the two instructions together. This way the scheduler will try to keep them together, and if it fails to do so, we might need to undo this bypass allocation in order to ensure that the program still works correctly.

Implement a bypass-aware combined scheduler and register allocator.

Chapter 5

Planning

Course-grain and fine-grain planning for the graduation phase. The graduation phase will take six months, so we distribute that time span over each of the phases in Table 5.1.

Table 5.1: Time table for the graduation phase.

Phase	**Tasks**	**Duration**
Total time		6 months

Chapter 6

Conclusions

Write your conclusions here.

Bibliography

[1] Y. He, Y. Pu, Z. Ye, S. Londono, R. Kleihorst, A. Abbo, and H. Corporaal, "Xetal-Pro: An Ultra Low Energy and High Throughput SIMD Processor," *in Design Automation Conference (DAC), 47th ACM/IEEE*, pp. 543–548, 2010. 1, 2

[2] L. Waeijen, D. She, H. Corporaal, and Y. He, "SIMD Made Explicit," *in Proceedings of the 13th International Conference on Embedded Computer Systems (SAMOS- XIII)*, pp. 330–337, 2013. 1, 2, 4

[3] J. Janssen, "Compiler Strategies for Transport Triggered Architectures," Master's thesis, Delft University of Technology, 2001. 1

[4] D. She, Y. He, L. Waeijen, and H. Corporaal, "OpenCL Code Generation for Low Energy Wide SIMD Architectures With Explicit Datapath," *in Proceedings of the 13th International Conference on Embedded Computer Systems (SAMOS- XIII)*, pp. 322–329, 2013. 1, 2

[5] L. Zhenyuan, "Code Generation of SIMD Architecture With Automatic Bypassing," 2016. 1

[6] C. Lattner and V. Adve, "The LLVM Instruction Set and Compilation Strategy," 2002. 5

[7] A. Aho, M. Lam, R. Sethi, and J. Ullman, *Compilers: Principles, Techniques, & Tools*. Pearson Education, 2 ed., 2006. 5

Appendix A

Supported Operations

This appendix contains descriptions for each of the instructions that are available on the proposed architecture.

Table A.1: List of R-type instructions shared by the CP and PEs.

Opcode	**Operation**	**FU**	**Description**
0	N/A	N/A	Not used.
1	N/A	N/A	Not used.
2	ADD	ALU	Signed addition.
3	SUB	ALU	Signed subtraction.
4	MUL	MUL	Signed multiplication.
5	MULU	MUL	Unsigned multiplication.
6	OR	ALU	Zero extended bitwise or.
7	AND	ALU	Zero extended bitwise and.
8	XOR	ALU	Zero extended bitwise xor.
9	CMOV	ALU	Conditional move.
10	SFEQ	ALU	Set flag if equal.
11	SFNE	ALU	Set flag if not equal.
12	SFLES	ALU	Set flag if less or equal, signed.
13	SFLTS	ALU	Set flag if less, signed.
14	SFGES	ALU	Set flag if greater or equal, signed.
15	SFGTS	ALU	Set flag if greater, signed.
16	SFLEU	ALU	Set flag if less or equal, unsigned.
17	SFLTU	ALU	Set flag if less, unsigned.
18	SFGEU	ALU	Set flag if greater or equal, unsigned.
19	SFGTU	ALU	Set flag if greater, unsigned.
20	SLL	MUL	Shift left logical.
21	SRA	MUL	Shift right arithmetic.
22	SRL	MUL	Shift right logical.
23	ROR	MUL	Rotate right register.

Table A.2: List of I-type instructions, shared by the CP and PEs.

Opcode	**Operation**	**FU**	**Description**
0	SIMM	ALU	Signed upper 18-bit for next immediate instruction.
1	ZIMM	ALU	zero extended upper 18-bit for next immediate instruction.
2	ADDI	ALU	Signed addition.
3	N/A	N/A	Not used.
4	MULI	MUL	Signed multiplication.
5	MULUI	MUL	Unsigned multiplication.
6	ORI	ALU	Zero extended bitwise or.
7	ANDI	ALU	Zero extended bitwise and.
8	XORI	ALU	Zero extended bitwise xor.
9	CMOV	ALU	Conditional move.
10	SFEQ	ALU	Set flag if equal.
11	SFNE	ALU	Set flag if not equal.
12	SFLES	ALU	Set flag if less or equal, signed.
13	SFLTS	ALU	Set flag if less, signed.
14	SFGES	ALU	Set flag if greater or equal, signed.
15	SFGTS	ALU	Set flag if greater, signed.
16	SFLEU	ALU	Set flag if less or equal, unsigned.
17	SFLTU	ALU	Set flag if less, unsigned.
18	SFGEU	ALU	Set flag if greater or equal, unsigned.
19	SFGTU	ALU	Set flag if greater, unsigned.
20	SLLI	MUL	Shift left logical.
21	SRAI	MUL	Shift right arithmetic.
22	SRLI	MUL	Shift right logical.
23	RORI	MUL	Rotate right register.
26	LB	LSU	Load byte.
27	SB	LSU	Store byte.
28	LH	LSU	Load half word.
29	SH	LSU	Store half word.
30	LW	LSU	Load word.
31	SW	LSU	Store word.

Table A.3: List of J-type instruction that can only the CP can execute.

Opcode	**Operation**	**Description**
0	NOP	Do nothing
1	SysCall	System call
2	BF	Branch if flag is set
3	BNF	Branch if flag is not set
4	J	Jump
5	JAL	Jump and link
6	Jr	Jump register
7	JALr	Jump and link register