Efficient MLIR Compiler Design: Vectorization for Presburger Library

Zhou Qi

4th Year Project Report
Computer Science
School of Informatics
University of Edinburgh

2023

Abstract

This report presents a faster implementation for the core pivot function of MLIR's presburger library. Its hot loop is element-wise overflow-checked multiplication and addition on an input matrix of low dimension and mostly small value elements.

The current approach of upstream is element-wise multiplication and addition on transprecision integer matrices, from int64_t to LargeInteger. This can be improved by efficiently utilizing hardware resources, taking advantage of SIMD, and reducing the bit width for every element: the compiler is not capable of automatically generating vectorized instructions for element-wise transprecision computing, and int64_t has a much larger bit width than what is typically used for most of the elements in the matrix. Additionally, extra arithmetics are required to perform overflow checking for int64_t, resulting in significant performance overhead. This report "innovates" the int23_t [1] datatype, a 23-bit integer datatype that utilizes the 23-bit mantissa of a 32-bit floating point, to address these issues. The faster "pivot" performs matrix-wise transprecision computing, targeting 99% TODO: confirm this number of the case where elements fit inside int23_t. Overflow awareness overhead is almost free, as floating point imprecision implies int23_t overflow (See Section **??**) and can be captured by a status register. It takes as low as 1 ns to check the status register in the pipeline, and only takes 9 ns to reset the status register. Additionally the status register is only cleared once before a sequence of pivot calls, making the average cost of clearing the status register per pivot negligible.

On a 30-row by 16-column example matrix, it performs 30 times faster than the upstream scalar implementation. The time cost of a single pivot call is reduced from 550 ns to 18.6 ns. TODO: replace this with actual MLIR benchmark result

TODO:Question: mention int16 here?

TODO:Reminder: we don't use int16 for (1) compatibility (2) slightly faster than float

[1] This is not really a innovation. It it a common technique on GPUs because often they are more capable on floating points than integers. See Section 1 for more history and detail.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Zhou Qi*)

Acknowledgements

Any acknowledgements go here.

TODO: Marisa Kirisame

Table of Contents

Chapter 1

Introduction

```
\begin{preliminary}
     ...
\end{preliminary}
```

Chapter 2

Background

2.1 Modern CPU micro-architecture

Computer architecture textbooks often deceipts

2.2 Citations

Citations (such as [1] or [2]) can be generated using `BibTeX`. For more advanced usage, we recommend using the `natbib` package or the newer `biblatex` system.

These examples use a numerical citation style. You may use any consistent reference style that you prefer, including "(Author, Year)" citations.

Chapter 3

Your next chapter

A dissertation usually contains several chapters.

Chapter 4

Conclusions

4.1 Final Reminder

The body of your dissertation, before the references and any appendices, *must* finish by page 40. The introduction, after preliminary material, should have started on page 1.

You may not change the dissertation format (e.g., reduce the font size, change the margins, or reduce the line spacing from the default single spacing). Be careful if you copy-paste packages into your document preamble from elsewhere. Some LaTeX packages, such as `fullpage` or `savetrees`, change the margins of your document. Do not include them!

Over-length or incorrectly-formatted dissertations will not be accepted and you would have to modify your dissertation and resubmit. You cannot assume we will check your submission before the final deadline and if it requires resubmission after the deadline to conform to the page and style requirements you will be subject to the usual late penalties based on your final submission time.

Bibliography

[1] Hiroki Arimura. Learning acyclic first-order horn sentences from entailment. In *Proc. of the 8th Intl. Conf. on Algorithmic Learning Theory, ALT '97*, pages 432–445, 1997.

[2] Chen-Chung Chang and H. Jerome Keisler. *Model Theory.* North-Holland, third edition, 1990.

Appendix A

First appendix

A.1 First section

Any appendices, including any required ethics information, should be included after the references.

Markers do not have to consider appendices. Make sure that your contributions are made clear in the main body of the dissertation (within the page limit).

Appendix B

Participants' information sheet

If you had human participants, include key information that they were given in an appendix, and point to it from the ethics declaration.

Appendix C

Participants' consent form

If you had human participants, include information about how consent was gathered in an appendix, and point to it from the ethics declaration. This information is often a copy of a consent form.