

Summer Research Reports

Vincent Zhao
vincentzhaorz@gmail.com

August 26, 2015

Abstract

This is the 7th week of my summer research, and this report will contain mainly what I've done in the past 2 weeks, including MaxJ SpMV implementation, Ruby multipumping implementation, and some bug fix in Ruby compiler.

Contents

1	MaxJ SpMV Implementation	3
1.1	CSRP format introduction	3
1.1.1	Redundancy formalization	3
1.1.2	Redundancy Optimization	3
1.2	Performance	4
2	Ruby Multi-Pumping Implementation	4
2.1	Ruby serialization primitives	4
2.2	Different aspects and implementations	4
3	Ruby cyclic loop	4
3.1	Problem: MaxCompiler auto-pipelining	4
3.2	Solution	4
4	Reports Archive	5
4.1	Report of Week 3 (Aug 3rd, 2015)	5
4.1.1	SpMV Implementation	5
4.1.2	Multi-Pumping	6
4.1.3	Ruby Compiler	6
4.2	Report of Week 4 (Aug 4th, 2015)	7
4.2.1	SpMV in MaxJ	7
4.2.2	Multi-pumping in MaxJ(Updated on Aug 25th 2015)	8

1 MaxJ SpMV Implementation

Compressed Sparse Row(CSR) is a famous sparse matrix storage method, it pull all the non zeros to the left of their own row, and store their original column index before they have been aligned. I'm building my SpMV algorithm based on this format.

1.1 CSRP format introduction

In order to reduce complex control logic which could slow down the processing and introduce bugs, there'll be **padding** zeros at the end of each row, in order to make those rows, which will be processed together, have the same number of elements to calculate(although there'll be some useless zeros). That's the idea about CSRP format.

1.1.1 Redundancy formalization

If we align all our rows to the row with maximal number of nonzeros, then we could imagine there'll be lots of redundancy. As we will process only a small number of rows at the same time(number of pipes), we could just align rows to the longest row in each **processing group**.

Let's formalize this description: If we have N rows for one sparse matrix, and for row i it has L_i number of non zeros. And if we group all these rows by R , which is the number of pipes for our design, then all these R rows in one group will be calculated together. So we could devise this formula:

The lower bound of our **redundant** calculation cycles will be:

$$Redundant\ cycles = \sum_i^{N/R} R \times \max_j^R L_{i \times R + j} - \sum_i^N L_i \quad (1)$$

1.1.2 Redundancy Optimization

If we transform formula (1) to this form:

$$Redundant\ cycles = \sum_i^{N/R} (R \times \max_j^R L_{i \times R + j} - \sum_j^R L_{i \times R + j}) \quad (2)$$

So here's a heuristic rule: we need to make sure there's no such case, that a very long row and several short rows coexist in one group. According to this idea, we could simply do a **sorting** on the rows by their length(which is a greedy algorithm). Is this really the best solution? We could easily prove it.

Imagine that we have N/R "buckets", each bucket could contain R rows. And now we need to put rows inside those buckets. We suppose that the best solution is put rows one by one in the sorted sequence, which means that the rows in the first group will be the rows with top R number of non zeros.

If there's a better solution than this one, then it should have swapped some of its elements. If we swap 2 rows between 2 groups, then it will increase redundancy in both groups.

1.2 Performance

I've tested this format with different R value, and 2 extreme sparse matrix format: dense and triangle. The following features have been discovered:

1. The maximal performance I could get is by using design with R equals to 32, and for dense matrix, GFlops is 0.28, for triangle matrix, it's 0.22.
2. With larger R value, the performance increase at first, and then fall down.
3. If R is larger, then redundancy will be larger.

More detailed test and report will be attached once I've finished the final version.

2 Ruby Multi-Pumping Implementation

2.1 Ruby serialization primitives

2.2 Different aspects and implementations

3 Ruby cyclic loop

3.1 Problem: MaxCompiler auto-pipelining

3.2 Solution

4 Reports Archive

4.1 Report of Week 3 (Aug 3rd, 2015)

Basically I've finished running and testing the real hardware implementation of my previous SpMV design, and improved it by enabling large matrix calculation. Also, I've written a simple multi-pumped design by using MaxJ. This one has been checked about its accuracy and time frequency.

4.1.1 SpMV Implementation

CSR format The format that I've devised and implemented could be called as CSR (CSR format with Padding zeros). This format enables multiple rows of elements to be calculated at the same time, and when those simultaneous calculated rows don't have the same number of non-zeros, this format will automatically add padding zeros to ensure the result is correct.

The main advantage of this format is its simplicity in implementation, but it will also face these defects:

1. *Redundant calculation*: For those cases when different rows have quite different number of non-zeros, the total number of calculations will depend on the maximum number of non-zeros among these rows.
2. *Resource duplication*: We need to replicate same kind of resources to enable parallelized calculation.

Implementation The implementation of CSR format is quite simple. We set 2 input streams, one is for value and the other is for column index of each non-zero element. All these input streams are interleaved and zero-padded. Each of the calculation components has one a cyclic-loop which will sum the result for current row, and one ROM which stores the values of vector. We take Matrix Market Format of matrix as input, and generate interleaved data from the original data.

Optimization The detailed statistics will be updated by next week, but here I'll list several ideas of optimization:

1. *Reduce redundant calculation*: Currently the number of redundant calculation depends on the maximal number of non-zeros of all rows in the matrix. We could reorganize our output data and let number of padding zeros depend on the maximal number of non-zeros of all rows in each simultaneously calculated groups. After that, we could also reorder the matrix rows to get the best placement of rows into different groups, and get a globally optimized result.
2. *Summation tree*: Current version of SpMV is based on a cyclic loop, which will sum the results for each row. A summation tree implementation should be better.
3. *Reduce resource usage*: We could do this by using dual-port of ROM, and if possible, using multi-pumping.

TODO

1. Finish the fully implemented baseline by Thursday. Fully implemented means, it could take MMF matrix data as input, and has reasonable space of memory.
2. Finish the first kind of optimization and the dual-port version by the end of this week.
3. If possible, try a multi-pumped version. (Only after the first 2 tasks finished)

4.1.2 Multi-Pumping

Simple implementation This simple implementation will do square on 2 input streams. Those input streams will be configured at 1x clock rate, and the inner multiplier will be configured at 2x clock rate. It could be implemented by using `ManagerClock` in `MaxJ`. The real hardware implementation has also been checked. The result is correct but the frequency is not accurate.

TODO

1. Finish checking this simple implementation. Must get a reasonable clock frequency.
2. Finish a Ruby implementation of multi-pumping by Thursday. Also, generate its `MaxJ` code.

4.1.3 Ruby Compiler

I haven't checked the compiler last week, but I'll start it as soon as possible. Basically, I should first use ruby to do some basic design, compile it into `MaxJ` and run it. After that, try to do some loop in Ruby. And test its `MaxJ` result (Which may not be correct).

TODO

1. *Loops in Ruby*: Test loops in Ruby, including acyclic loops and cyclic loops. Simulate them and check its generated `MaxJ` code.

4.2 Report of Week 4 (Aug 4th, 2015)

This week I've made progress on all the 3 directions: SpMV in MaxJ, multi-pumping in MaxJ and multi-pumping in Ruby. Here're the details.

4.2.1 SpMV in MaxJ

According to the plan, I've finished the baseline version of CSRp SpMV on time. You could find out the definition of CSRp in the archived week 3 report. This version will generate correct result, but it must be improved in many aspects:

1. **Get runtime correct:** Currently, if we add level of parallelization, the run time will increase, not as what we expected. If the code is correct(maybe not), then something wrong might happen during synthesis and implementation. Check the **clock frequency** of the **built kernel**. This will tell whether there's problem during the implementation: if the clock frequency also not follow our expectation.

It might be the case that the current using matrix will never benefit from the level of parallelization, so we need to generate matrix based on our current design specification, and discover their performance.

Todo First check out the frequency of each design with different parallelization level. Also try to organize the code better. Then use randomized benchmark, not only the value, but also the structure to test run time result.

2. **Improved CSRp format:** The CSRp format will perform really bad if the padded zeros occupy most of the computations, which means, we have lots of redundant zeros. Why we have those zeros? As we cannot calculate all the rows simultaneously, only part of the rows will be calculate at the same time. If we call these rows a **block**, then we will calculate the whole matrix block by block. Now, we do not specifically add **splitters** between those blocks. We would know which block we are calculating only by forcing all the blocks has equal width. The way we force this happen is by padding zeros. The number of padding zeros could be enormous if we have a very wide block and most of the blocks are quite narrow. The **redundancy** metrics could be calculated by using:

$$R_{CSRp}(N) = \max_{0 \leq i < N} (block_i.width) \times N \times p - \sum_i^{Np} row_i.width$$

where N is the number of blocks, and p is the level of parallelization. If we have N and p assigned, then the whole number of rows should be Np. The **width** is the number of non-zeros for each row or block. Especially for blocks, their width is the maximal width among all rows in that block.

Obviously, we could improve this. If we assign more information for each clock cycle, like "whether this cycle is in the end of one block", which is the **splitter** that we have mentioned above, then we do not need to do padding on all the blocks to make sure they have same width.

This one could be achieved by adding a new input stream of 0, 1 boolean values. We could verify the improvement in speed and space increase-ment.

Todo: Finish the first part of SpMV work first. And build this design later. Discover their performance differences. And I think this is one should be the standard design of **baseline**. Previous one has really bad performance.

3. **With multi-pumping:** I've already finished a multi-pumping MaxJ version, which has a simple multiplier inside. And we all know that SpMV is all about add and multiply(basically), so we could use the same framework to optimize this SpMV design.

Todo: Last one to do, I think I could do this in the next week.

That's all about this week's SpMV work, I'll attach a full chart about this design's performance in next week's report, as I need to verify some current data this week.

4.2.2 Multi-pumping in MaxJ(Updated on Aug 25th 2015)

In this part I'll just illustrate some result I've gathered in my Multi-pumping MaxJ version's experiment.

First of all, we have these versions to implement the simple multiplication functionality:

1. **UnrollLoop:** This one will use 1x clock rate for multiplier, but will use 2 of them.
2. **ResourceShare:** This one will use one single multiplier to do 2 multiplications sequentially, at 1x clock rate.
3. **Multi-pumping:** This one will use 1 multiplier, running at 2x clock rate. And other parts running at 1x clock rate.

What we expect about these versions are:

1. Multi-pumped version will save number of DSPs compared to the origin version, which will use 2 multiplier to finish 2 multiplication in one 1x clock rate cycle.
2. Multi-pumped version will be 2x faster than the resource sharing version, which will not use more multiplier, just finish the 2 multiplication one by one.

And the result looks the same:

1. **Speed:** UnrollLoop version has the fastest frequency(178MHz), and multi-pumping has similiar result(171MHz). ResourceShare version could just achieve have of the frequency(92.1MHz).
2. **Resource Usage:** For the DSPs, it's obvious that UnrollLoop will use 2 and the others two version will use one. And it should be true that Multi-pumping will have more logics usage(14479) that ResourceShare(13923) as it will have more control logic.