

Binary Analysis for Missed Vectorization Opportunities Detection

AMOS HERZ and ALESSANDRO LEGNANI, ETH Zürich, Switzerland

1 INTRODUCTION

Modern compilers are often able to automatically vectorize code using SIMD instructions. Take, as an example, the following code snippet:

```
1 void copy(long *restrict a, long *restrict b, unsigned long n) {  
2     for (unsigned long i = 0ul; i < n; i++) {  
3         a[i] = b[i];  
4     }  
5 }
```

Listing 1. copy.c

We can compile it with the following set of compiler flags

- -O3: tells the compiler to use the highest level of optimization available.
- -fno-tree-loop-distribute-patterns: prevents replacing the loop with a call to memcpy
- -fno-tree-vectorize: prevents vectorization

Which will produce the following assembly code:

```
1 .L3:  
2     movq    (%rsi,%rax,8), %rcx  
3     movq    %rcx, (%rdi,%rax,8)  
4     addq    $1, %rax  
5     cmpq    %rax, %rdx  
6     jne     .L3
```

Listing 2. copy.c compiled with vectorizations disabled

However, by compiling without the -fno-tree-vectorize flag, the compiler will produce the following vectorized code (note the use of wider instructions and registers):

```
1 .L4:  
2     movdqu   (%rsi,%rax), %xmm0  
3     movups   %xmm0, (%rdi,%rax)  
4     addq     $16, %rax  
5     cmpq     %rcx, %rax  
6     jne     .L4
```

Listing 3. copy.c compiled with vectorizations enabled

The main goal of this project is to develop a method for identifying missed opportunities for vectorization in existing code. That is, given an existing binary, we want to identify pieces of code and procedures that could have been vectorized by the compiler but were not.

2 RELATED WORK

Autovectorization is *difficult*, compilers tend to miss many optimizations (as shown by Feng et al. [3]), and more than often vectorizing a small piece of code requires large changes to the whole code base (as was done for example by Chen et al. [2]).

Compilers have been shown to widely miss out on vectorization opportunities. As an example, Maleki et al. [4] report that in their research only 45-71% of the loops in a benchmark they developed

Authors' address: Amos Herz, amherz@ethz.ch; Alessandro Legnani, alegnani@ethz.ch, ETH Zürich, Zürich, Switzerland, 8092.

and only a few loops from the real applications are vectorized by the compilers they evaluated, which include widely popular compilers such as GCC (version 4.7.0).

Auto-vectorization is still an open field of research: attempts have been made at “fixing” the compiler work by post-processing compiled code (Porpodas and Ratnalikar [6]) or by applying Machine Learning to produce improved vectorization schemes (Mendis et al. [5]).

3 APPROACH

3.1 Dataflow Analysis

Let’s look at Listing 2: there, we can realize that the load and store instructions are completely independent (from other instructions of the same type) and could, theoretically, be computed in parallel, therefore vectorized. This can be concluded by simply looking at the dataflow, an example on how a dataflow graph could look like is given in Figure 1.

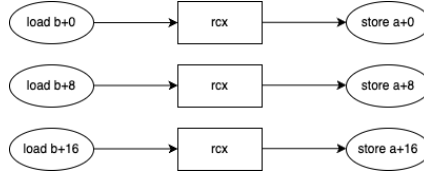


Fig. 1. Example of dataflow graph for copy.c

3.2 Dynamic Analysis

In order to build a dataflow graph we will be using the dynamic analysis framework Intel Pin¹. Intel Pin allows us to instrument single instructions in the assembly to generate the appropriate trace information to construct our dataflow graph.

We opted to use dynamic analysis in our approach as it simplifies the approach of finding missed vectorization when compared to static analysis. Dynamic analysis allows us to generate the trace of a given program, containing all of its memory accesses, which can easily be converted into a dataflow graph. However, when performing dynamic analysis, we will need to be careful not to falsely report code with complex control flow as non-optimized.

```

1 #include <stdint.h>
2 void dyn(float *a, float *b, uint8_t *c, unsigned int n) {
3     for (int i = 0; i < n; ++i) {
4         if (c[i]) {
5             a[i] = 2 * b[i];
6         } else {
7             a[i] = 4 * b[i];
8         }
9     }
10 }

```

Listing 4. dyn function

As an example, assume that in an execution of function dyn from Listing 4 we have that every element of array c is 0. In this case, we would observe that $a = 4 * b$ and vectorization would therefore be trivial. Therefore, if the code were not to be vectorized, we would flag this code as missing a vectorization opportunity, without considering the control flow operation. Note that this example is actually trivially vectorizable if we were to employ masked vector operations. However, this is only an illustrative example and one can consider even more complex control flow operations

¹<https://software.intel.com/sites/landingpage/pintool/docs/98830/Pin/doc/html/index.html>

that make the code impossible to be vectorized. To fix this issue we can either use static analysis or consider the control flow operation in our dataflow graph. We plan on using the second approach as it is less complex to implement and are aware of the loss in precision this could result in.

All in all, we foresee that building the dataflow graph will likely be the greatest challenge this project poses. An alternative, would this challenge turn out to be too big of a one, could be to simply analyze the program trace directly, to track down for example repeating accesses to increasing (or decreasing) memory addresses that would be vectorizable but are not.

3.3 Examples of missed vectorization opportunities

By searching on the different bugs and issues trackers of the clang and gcc compilers we found two examples where the latest version of a compiler does not generate vectorized code. For this we are considering the latests versions of both gcc (13.2) and clang (18.1.0).

3.3.1 Example 1 (gcc). The following example has been taken from the gcc bugs tracker².

```
1 #include <stdint.h>
2
3 void ex1(int8_t *v, int8_t x, const uint64_t *bits, unsigned n) {
4     int num_words = (n + 64 - 1) / 64; // round up to nearest quad-word
5     for (int i = 0; i < n; ++i) {
6         const uint64_t word = bits[i];
7         for (int j = 0; j < 64; ++j) {
8             v[i * 64 + j] += x * (bool)(word & (1UL << j));
9         }
10    }
11 }
```

Listing 5. Piece of code that gcc fails to vectorize

One would expect the compiler to generate a mask from the word variable and perform masked multiplication to vectorize the code. Compiling this code with the -O3 flag on gcc yields no vectorized code, whereas clang unrolls the inner loop and, as expected, makes extensive use of vector instructions. A side-by-side comparison of the two compilers' work can be found at the following link: <https://godbolt.org/z/W8v4Pc463>.

3.3.2 Example 2 (clang). The following example has been taken from the clang issues tracker³.

```
1 void cvt_8_32(uint32_t* op, uint8_t const* ints)
2 {
3     // When this is outlined it generates correct code
4     uint32_t out[16];
5     uint8_t in[16];
6     memcpy(in, ints, sizeof(in));
7     for (int i = 0; i < 16; ++i) {
8         out[i] = in[i];
9     }
10    memcpy(op, out, sizeof(out));
11 }
```

Listing 6. Piece of code that clang fails to vectorize

It appears that when the function in Listing 6 is inlined, clang fails at correctly vectorizing it. However gcc correctly vectorizes the code without any issue. A side-by-side comparison of the two compilers' work can be found at the following link: <https://gcc.godbolt.org/z/PbGfe6da7>.

²https://gcc.gnu.org/bugzilla/show_bug.cgi?id=96888

³<https://github.com/llvm/llvm-project/issues/74380>

3.4 Test Inputs

To test our results, we will use code snippets from an already existing autovectorization test suite: TSVC⁴, first described by Callahan et al. [1]. Other random program generators such as YARPGen⁵ could be used to generate more, less specific, test inputs, even though we predict that such less specific inputs could hardly contain code that can be vectorized but is not.

4 IMPLEMENTATION AND EXPERIMENTAL SETTINGS

As mentioned previously, we will test our implementation (implemented using the latest version of Intel Pin) using the latests versions of both gcc (13.2) and clang (18.1.0). In order to do so, we use pieces of code from the TSVC test suite that are known to be vectorizable, pass them to gcc and clang preventing them to vectorize the resulting binary and then test if our implementation can actually spot the (induced) missing vectorization opportunities. A diagram representing the pipeline of this project is provided in Figure 2.

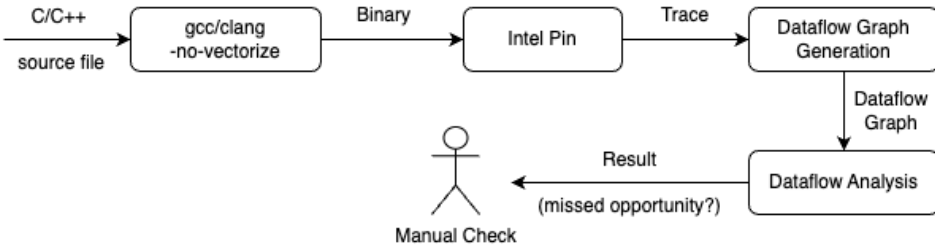


Fig. 2. Sketch of the pipeline of the whole project

As the possibility space of different kinds of missed vectorization opportunities is enormous we decided to focus on some specific patterns. Initially our approach will be to be able to recognize basic missed vectorizations, which don't appear in the wild anymore, like the one shown in Listing 2. To do so more effectively, we will start from the top of the code snippets collection from TSVC⁶ that contains various vectorizable pieces of code of increasing complexity, ranging from trivial sum of vectors to more intricate and convoluted operations.

Afterwards we will be focusing on pattern-matching the access patterns found in the two examples described in 3.3, which is non-trivial work. For this we will be using a combination of Pin's built in Dynamic Control-flow Graph Generation⁷ and our own PinTool to extract the trace.

4.1 Current state of the implementation (as of 28.04.2024)

All the code of this project can be viewed at the following link: <https://github.com/herzamos/ast24-project/>. At the moment, we are able to correctly instrument the binary code and, leveraging Intel Pin tools, generate a code trace that tracks all the memory accesses (both reads and writes).

REFERENCES

- [1] D. Callahan, J. Dongarra, and D. Levine. 1988. Vectorizing compilers: a test suite and results. In *Proceedings of the 1988 ACM/IEEE Conference on Supercomputing* (Orlando, Florida, USA) (*Supercomputing '88*). IEEE Computer Society Press, Washington, DC, USA, 98–105.

⁴https://github.com/UoB-HPC/TSVC_2

⁵<https://github.com/intel/yarpgen>

⁶https://github.com/UoB-HPC/TSVC_2/blob/master/src/tsvc.c

⁷<https://www.intel.com/content/www/us/en/developer/articles/technical/pintool-dcfg.html>

- [2] Yishen Chen, Charith Mendis, and Saman Amarasinghe. 2022. All you need is superword-level parallelism: systematic control-flow vectorization with SLP. In *Proceedings of the 43rd ACM SIGPLAN International Conference on Programming Language Design and Implementation* (San Diego, CA, USA) (*PLDI 2022*). Association for Computing Machinery, New York, NY, USA, 301–315. <https://doi.org/10.1145/3519939.3523701>
- [3] Jing Ge Feng, Ye Ping He, and Qiu Ming Tao. 2021. Evaluation of Compilers’ Capability of Automatic Vectorization Based on Source Code Analysis. *Scientific Programming* 2021 (30 Nov 2021), 3264624. <https://doi.org/10.1155/2021/3264624>
- [4] Saeed Maleki, Yaoqing Gao, María Garzarán, Tommy Wong, and David Padua. 2011. An Evaluation of Vectorizing Compilers. *Parallel Architectures and Compilation Techniques - Conference Proceedings, PACT*, 372–382. <https://doi.org/10.1109/PACT.2011.68>
- [5] Charith Mendis, Cambridge Yang, Yewen Pu, Dr.Saman Amarasinghe, and Michael Carbin. 2019. Compiler Auto-Vectorization with Imitation Learning. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/d1d5923fc822531bbfd9d87d4760914b-Paper.pdf
- [6] Vasileios Porpodas and Pushkar Ratnalikar. 2021. PostSLP: Cross-Region Vectorization of Fully or Partially Vectorized Code. In *Languages and Compilers for Parallel Computing*, Santosh Pande and Vivek Sarkar (Eds.). Springer International Publishing, Cham, 15–31.