Static Profiling of Assembly Code Performance and Optimization Effectiveness using Instructions Performed and Program Latency

Jonathan Paul C. Cempron, Chudrack Shalym Y. Salinas, Roger Luis Uy

Abstract: Software program optimization for improved execution speed can be achieved through modifying the program. Programs are usually written in high level languages then translated into low level assembly language. More coverage of optimization and performance analysis can be performed on low level than high level language. Optimization improvement is measured in the difference in program execution performance. Several methods are available for measuring program performance are classified into static approaches and dynamic approaches. This paper presents an alternative method of more accurately measuring code performance statically than commonly used code analysis metrics. New metrics proposed are designed to expose effectiveness of optimization performed on code, specifically unroll optimizations. An optimization method, loop unroll is used to demonstrate the effectiveness of the increased accuracy of the proposed metric. The results of the study show that measuring Instructions Performed and Instruction Latency is a more accurate static metric than Instruction Count and subsequently those based on it.


I. INTRODUCTION

Software program optimization for improved execution speed can be achieved through modifying the program. Programs are usually compiled from a high level language into machine low level language. More coverage of optimization and performance analysis can be performed on low level than high level language. Discussed are the process of how a program is transformed from the programmer’s code into a language that the processor natively understands. And presents strategic choice on which portion of this process is best for implementing optimizations on.

Optimization improvement is measured in the difference in execution performance. Several methods are available for measuring code performance classified into static and dynamic approaches [1]. Dynamic approaches involves actual program runtime, but less focus on analysis of code. Current disavantages of IC and alternative methods to more accurately measuring code performance statically.

New metrics presented are named Instructions Performed and Program Latency. These metrics are initially designed to expose effectiveness of unroll optimization performed on code. But can be used to more accurately represent code performance.

An optimization method, loop unroll is used to demonstrate the effectiveness of the increased accuracy of the proposed metric. A method for optimizing assembly code [2] used by popular compilers GCC and ICC [3].

II. CODE TRANSLATION AND OPTIMIZATION

Software is most commonly written on High Level Languages then translated into Low Level Language native to the processor that the software will be executed on [4][5][16][18]. Translation can be done either through Interpretation or Compilation. Interpretation is done on a line per line analysis then execution of the written source code. Examples are Python and R. Compilation is when the whole source code is analyzed and the whole software is translated at once. Example of a compiled language is C.

Software optimization is more commonly performed in compiled language than in interpreted language because of the available knowledge of the complete flow of the program presented at compile time. Also there is more time available to perform optimizations during compile time while in interpreting the language, there is significant overhead caused during optimization.

Optimization can be performed on either high level language or in low level language. However we suppose that more coverage, practicality, and effectiveness of optimization can be achieved on low level. There is more coverage because there are more languages being compiled to a single processor architecture [6]. It is also more practical in that there are fewer architectures than languages to perform optimizations on. It is also more effective because of the closer relationship of the optimization to the actual processor hardware.
Static Profiling of Assembly Code Performance and Optimization Effectiveness using Instructions Performed and Program Latency

[15][18][23][24][25][30]. Popular processor architecture families that are subject to optimizations are: x86 architectures, ARM architectures, and MIPS architectures.

Consequently, analysis of the optimization is also best done on low level languages. Measuring the performance of a code and also the effectiveness of an optimization is done using performance metrics. Popular metrics are discussed and also a more accurate proposed metric is also presented.

Optimization could be of different interests [16][17][20][21][26], of which are: code size, code density [7], speed, memory, data, network, and power consumption. The focus of this paper is speed; shorter execution time.

Optimization for speed is best achieved when the software program takes advantage of the processor architecture’s features such as pipelined processors and out of order execution.

### Static Profiling of Assembly Code Performance

Pipeline is a feature wherein several machine instructions can be performed at once. The pipeline is maximized in software by reduction of stalls. A stall processor state is when an execution unit is waiting for a data dependency. Stalls are reduced when dependencies are avoided, this can be done by modifying the assembly code [8][9][10][27]. Figure 1 shows a MIPS pipeline with and without a stall, a stall causes instructions to consume more processor cycles. In an Ideal Pipeline, all the stages: Instruction Fetch (IF), Instruction Decode (ID), Execute (EX), Memory Access (MEM), and Write Back (WB), perform in lockstep; each instruction is completed in 5 cycles. In a Stalled Pipeline however, due to either a dependency on a previous instruction or a lack of resources, an instruction has to be stalled in order for the program to execute correctly.

Software optimization is still closely coupled with knowledge of the computer architecture being used and its actual implementation. One common cause of stalls are jumps in assembly programs. Jumps are typically caused by conditional statements and loops. One method of optimization is by minimizing usage of jumps in loops by performing loop unroll.

Loop unrolling extends source code such that the use of branch instructions is reduced. The method of unrolling a loop will not be discussed but as a summary, loop unrolling increases code length in exchange for reduced latency from branching instructions. As illustrated in Figure 2.

#### CURRENT PERFORMANCE METRICS

The effectiveness of any optimization method must be tested and measured before claiming that the method actually optimizes for a specific interest. This measurement is done by using program performance metrics. Code performance metrics have two categories: static and dynamic.

Dynamic metrics are the measurement of the program performance during actual runtime. This can be done through the use of the clock() function in C. Figure 3 illustrates this method. The clock() function is used to obtain the starting and ending times of the program. The execution time of the program is calculated by subtracting the start and end time.

![Figure 1. MIPS Pipeline example](image_url)

**Figure 1. MIPS Pipeline example**

![Figure 2. Example of an unrolled loop](image_url)

**Figure 2. Example of an unrolled loop**

![Figure 3. Performance measurement using C function clock()](image_url)

**Figure 3. Performance measurement using C function clock()**

The simplicity of this measurement approach is its main advantage. A second advantage is the return of measurements in actual time. One of the disadvantages of this method is its unreliability when a program requires user input. Another disadvantageous situation is when the measurement is performed in a multithreading environment; where the results will always include the effects of the other programs running in the same computer. This causes an inaccuracy where the resulting time is more than what the program actually consumes. There are approaches to overcome the challenges presented by multithreading. One approach is to combine the measurement with static profiling of the code.

The focus of this paper is on Static methods of determining code performance. One of the most popular metrics for comparison is through Instruction Count (IC). IC is the number of lines that are in the Code Segment of an assembly code program [8][9]. White space lines, and assembler directive are not included in the instruction count.
IC is the most simplistic static metric for a program. An advantage of this metric is that it is easy to perform. Another commonly used metric for assembly program code measurement is the CPU Time shown in equation (1). CPU Time includes the actual speeds of the processor thus should yield the actual time [8][9].

\[ \text{Ideal CPI} + \frac{\text{CPU Time}}{\text{Clock Cycle Time}} \]  

(2)

For pipelined processors, the Clock Cycle per Instruction of equation (1) be replaced with Pipeline CPI. Which considers the processor pipeline feature. This yields a more accurate result in case of computing for a pipelined processor [4].

Figure 4. Instructions Performed of a Looping and Non-Looping Program

Figure 4 above will be used to display the limitation of Instruction Count (IC). Program has an IC of 5 while Program B has an IC of 8. Program B has a higher IC. However when the two programs were ran Program A has more instructions that were performed. This limitation of IC is because it does not have consideration of programming loops. Lepak et al. [4] also agree with the unreliability of instruction count in multiprocessor systems and discuss a simulation methodology for improved performance measurement.

Another limitation of Instruction Count is exposed in Figure 5 above. A program has less instructions performed than its IC. This inaccuracy is because IC does not have consideration for branches or jumps.

The inaccuracy of IC is also propagated to the CPU Time equation (1). Since IC is a part of CPU Time. Solving this inaccuracy is important for comparing an optimized and an unoptimized code. That is because an optimized code can have a higher IC but may not necessarily be slower.

IV. OPTIMIZER EFFECTIVENESS

A different metric was developed by this study that would provide a more accurate performance analysis of the program code. The metric takes into consideration programming blocks. The metric uses an analysis of the instructions that will be performed by the code after execution. This solution will be called Instructions Performed for the rest of the discussion.

Instructions Performed is the number of instructions executed by the program during runtime. This takes into consideration blocks of code that are repeated while the program is running. Blocks of code that are ignored on runtime are also accounted. For a more accurate CPU time in equation (1), the Instruction Count can be replaced by Instructions Performed.

Instruction Latency is similar to instructions performed with added consideration to the latency of each instruction. As different instructions take different time in some architectures such as the x86_64 architecture. For the x86_64, latency per instruction information is defined by their document for recommendations for compiler developers in [11] and other processor related documents [19][22][29].

V. METHODOLOGY

The method for computation of proposed metrics Instructions Performed and Program Latency is explained by demonstration. Below are Fibonacci programs written in different processor assembly languages. The computation for the proposed metrics are shown alongside the code.
Figure 6 shows a MIPS64 code for computing 8 Fibonacci numbers and stores those numbers into memory. The steps for computing Instructions Performed is displayed in the table to the right of the code. The steps in computing the Instructions Performed are presented as columns, read from left to right.

Figure 7. Instructions Performed of x86_64 Fibonacci code

Figure 7 shows the same Fibonacci code but implemented for the x86_64 architecture. The following steps are done in order to compute for the Instructions Performed. First step is to separate the code into basic programming blocks. Second is to determine the type of blocks. Third is to count the number of instructions in each block. Fourth is to determine the number of repetitions the block will be performed throughout the program. This can be determined by analyzing the value assigned to the loop counter before entering the block. Typically, only a 2way block has more than 1 repetition. If the number of repetitions cannot be determined, just assume that the block is executed once. A single execution is chosen as default because a block of code would usually be at least used once. Fifth step is to multiply the Instruction Count and block repetition. Sixth and final step is to sum the products and the result is the Instructions Performed.

Figure 8. Program Latency of x86_64 Fibonacci code

Figure 8 shows the computation for Program Latency for the x86_64 Fibonacci program. To compute for Program Latency are the following steps. First is to separate the code into blocks. Second is to determine the type of block. Third is to determine the latency per each instruction (refer to processor documentation for this step). Fourth is to sum the latencies per each block. Fifth is to determine the block repetition. Sixth is to multiply the block repetition and block latency. Seventh and final step is to sum the products and the result is Program Latency.

VI. TEST AND RESULT

Testing the effectiveness of the proposed metrics Instructions Performed and Program Latency is demonstrated via comparison of computed values on an optimized and an unoptimized version of the Fibonacci codes presented. The optimization method used is loops unrolling. The proof of loop unroll effectiveness can only be exposed in increased accuracy offered by the proposed metrics.

Figure 9. Instructions Performed of x86_64 unrolled Fibonacci

Figure 9 shows an unrolled loop from the program shown in Figures 7 and 8 unrolled with a factor of 4 wherein the loop instruction is placed at the bottom and a decrement of the loop counter is placed in between each unroll block. The unrolled block resulted in an instruction count of 16, while the original program has an instruction count of 4 for the same block. The Instruction Count is increased for the unrolled program, but the instructions performed are the same, 33 Instructions Performed.
VII. ANALYZING TIME COMPLEXITIES

Big O notation used for measuring code performance in terms of growth function of an algorithm’s frequency count of its basic operation [12][13][14]. The proposed method for computation can be extended to also express time complexities in terms of big O notation. Not all types of big O notation can be detected and expressed and are limited to the following: O(1), O(n), O(nx) where x is a non-zero positive integer. The extension procedures is placed on the loop detection portion:

1. if no loop is detected, then O(1),
2. if a loop is detected and there is no nested loop, then O(n),
3. if a loop is detected and nested loop, then O(nx) where x is the layer of the deepest nest.

It should also be noted that for O(nx) the iterations of the loops are assumed to be the same for all layers thus nx does not cover all cases. Example 1 assume a 3 layer nested loop where the 1st, 2nd, and 3rd layers have the same n iterations, then our output O(n3) is correct. Example 2, assume a 3 layer nested loop where the layers 1st has n iterations, 2nd has o, and 3rd has p, then our output O(n3) is not correct because n is not a single number, we could average (n+o+p)/3 for a better approximate but it is still not accurate. Example 3 assume a 3 layer nested loop where 1st has n, 2nd has o, 3rd has a constant 7 iterations, then our output O(n3) is not correct because the deepest layer is 3 but the third layer is actually a constant and the correct answer is O(n2).

VIII. CONCLUSION

The presented metrics: Instruction Performed and Program Latency provide a more accurate representation of code performance than Instruction Count based metrics because of its increased accuracy. Approximation of Time Complexities are also presented. The presented methods also provide a tool to prove that a higher Instruction Count does not necessitate a faster running program as shown in the case of an unrolled loop. From the results in Figure 11 it can be seen that the effectiveness of loop unroll can best be statically determined through computing the Program Latency. An unrolled program has longer Instruction Count but has significantly better Program Latency, outperforming the original code.

It is important to note that the relationship between Instructions Performed and Program Latency is dissimilar from that of the Instruction Count and Program Latency. The worst comparative result that Instructions Performed can provide is failing to expose performance gain, but it will not be the reverse from the Program Latency as was in Instruction Count. Thus, Instructions Performed is still more accurate than Instruction Count.

The most accurate metric presented, Program Latency, requires deeper knowledge of the processor as it requires information regarding the latency of each instruction. This information can sometimes be inherent to the processor architecture, as in MIPS64 having uniform latency for each
instruction shown in Figure 12. Latency per instruction can also be provided by the processor architecture manufacturer, as is the case in x86_64 architecture with [11]. But in some cases, this kind of information is not easily available, thus Program Latency cannot be computed for some architectures. Presented in this paper was a more accurate method of statically determining code performance and an approximate of the time complexity of a program designed to display the effectiveness of an optimization by more accurately exposing performance differences. The accuracy is with the cost of tediousness and should be implemented into software that automatically computes for the Investigations Performed and Program Latency. Other future work includes further testing of the static profiling in terms of time complexity.

REFERENCES

22. B. Blaha and D. Wunsch, "Evolutionary programming to optimize an assembly program," Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8660), Honolulu, HI, USA, 2002, pp. 1901-1903 vol.2. doi: 10.1109/CEC.2002.1004533

AUTHORS PROFILE


Chudrack Shalym Y. Salinas. A Graduate of BS Computer Science major in Computer Systems Engineering in De La Salle University. Research focus on Computer Architecture.

Roger Luis Uy. Assistant Professor, Computer Technology Department, College of Computer Studies, De La Salle University, Manila Philippines. Research interest includes applying concepts of Computer Architecture to other interdisciplinary fields such as Bioinformatics.