The first implementation of an optimized HFT algorithm on an ARM chip

Supervisor: Dr. Ruibang Luo

Siddhant Ningoo

2nd October 2022
Table of Contents

Abstract ........................................................................................................................................... 3
1. Introduction & Background ........................................................................................................... 4
   The current situation ................................................................................................................ 4
   The advancement in technology ............................................................................................... 4
   The gap ..................................................................................................................................... 4
2. Related Work ............................................................................................................................. 5
3. Objective ................................................................................................................................... 5
4. Project Methodology ................................................................................................................ 6
   Part 1: Literature review and research on ARM architecture ................................................ 6
   Part 2: Implementation ............................................................................................................. 6
   Part 3: Appraisal of performance ......................................................................................... 6
5. Schedule & Milestones .............................................................................................................. 8
References ..................................................................................................................................... 10
Abstract

With the advancement in ARM based architecture, there has been an increase in the prominence of NEON instruction set based chips in more and more commercial solutions. This project looks toward leveraging ARM based chips’ potential to parallelize computing in the form of vectorization & multithreading, specific to the High Frequency Trading (HFT) field where high-performance computing is pivotal to successful implementation of HFT algorithms.
1. Introduction & Background

Technology, specifically Algorithmic Trading (AT), has revolutionized the way people trade in the markets, with it now accounting for 60-73% of overall US equity trading as per a market research by Mordor Intelligence LLP (2022), with them predicting a further surge of 10.5% over the coming 5 years. High Frequency Trading (HFT) is a form of algorithmic trading that leverages high computational speeds to automate establishment and liquidation of securities in a small-time frame (SEC, 2010).

The current situation

It is estimated that as of 2020, about 50% of the trading volume in US Equity Markets is generated from HFT (Breckenfelder, 2020). HFT is a form of AT that heavily relies on low-latency and the efficiency of the algorithms, in order to achieve high speed and high frequency trading. Apart from the hawkish regulators, firms engaging in HFT need to constantly be aware of the competition from their competitors in the race to produce an algorithm that can be faster by milliseconds ("An Introduction To The HFT Industry And Its Key Players", n.d.). Buchanan (2015) mentions that “as technology advances, trading speed is increasingly limited only by fundamental physics.” However, De Prado (2018) has outlined in his book that while many investment managers think complex machine learning algorithms is key to success for lower latency, it is in fact a mixture of multiple factors, ranging from enhanced data manipulation, vectorization, multiprocessing, etc., which often most firms engaged in HFT fail to leverage.

The advancement in technology

Spanning across decades, x86-64 was the predominant architecture used in PCs while the ARM based architecture was mainly used to power smartphones. However, in recent development, ARM based architecture with the NEON instruction set has been entering the PC space and even the cloud computing space with the likes of Apple M series chips and AWS’s AWS Graviton2 respectively. It has been claimed that customers of AWS (such as Snap, Twitter, Netflix, etc.) have seen improved performance and lowered costs after switching to the ARM based architecture ("AWS & ARM - Partners - ARM", n.d.). Similarly, there are have many studies done to ascertain the positioning of ARM processors in this space and many come to a similar conclusion of ARM performing better energy-to-solution (Gupta & Sharma, 2021; Mantovani et al., 2020; Criado et al., 2020) While Mantovani et al. (2020) even go on to mention ARM based chips could power the next generation of High Performance Computing (HPC) systems.

The gap

When looking at the type of computer architecture on which HFT algorithms run on, the x86-64 based processor segment had the highest revenue share and is estimated to further grow it’s foothold in this space from 2021 to 2028 (Bloomberg, 2021). Coming from the positive signs of using an ARM based architecture in HPCs to the wide praise it has received, it would hence be worth exploring how HFT algorithms would perform on an ARM based chip, especially when firms are looking in every direction to shed a few milliseconds. Despite it’s efficient architecture, there is a current issue of the algorithms still being not as efficient and utilizing the ARM architecture to the maximum.
2. Related Work

HFT and the AT world is such a lucrative industry that most of the industry level code is locked as corporate secret and aren’t available to the general public, hence most of the publicly available bots and the pre-existing work done in this space is very surface level and experimental. Because of that, many of them are not quite optimal, and hence this project aims to produce a better performing algorithm.

While there are tons of research done on ARM based HPC, for instance Project Mont Blanc (n.d.), Fugaku - another ARM based supercomputer, there is no research on optimizing and evaluating HFT algorithms specific to an ARM chip, hence this project aims to fill this gap to optimize & analyze algorithms for ARM based CPUs for industrial purposes (eg. Apple M series chips or AWS Graviton2)

3. Objective

Owing to the productivity of the newer ARM based chips, it is worthwhile to study the performance of HFT algorithms on the newer ARM based chips, with algorithms and the AI models further streamlined to best optimize the total output capabilities of the newer ARM chips. With a focus on Vectorization & Multithreading, the objective of this project is to make such modifications to the HFT models to better fit the ARM processor which in turn makes the algorithm more efficient, with other such modifications that complement the higher performance such as: Information Driven Bars for better accuracy and lower latency to react to market changes, Feature Importance etc.

To truly assess if it has any impact and if these modifications bring about deviant performance changes this project proposes to analyze the performance of the ARM based algorithm in comparison to pre-existing models running on x86-64 processors:

1. it’s latency in order placement
2. the total portfolio performance

From this, the main objectives of the project are to:

- Provide the first fintech application of HFT that leverages the new ARM based architecture to its fullest
- Evaluate and provide recommendations to corporates and funds that engage in HFT to enhance their algorithms’ performance for higher efficiency and lower latency.
4. Project Methodology

The project will be divided in roughly 3 parts, with Part 1 focusing on research of the ARM architecture, Part 2 focusing on the implementation and Part 3 focusing on the analysis.

Part 1: Literature review and research on ARM architecture

In order to better optimize the models it is imperative to understand the ARMv8-A architecture, the NEON instruction set, based on a Reduced Instruction Set Computer (RISC), while the current implementations of x86-64 are all based on a Complex Instruction Set Computer (CISC) and how the NEON instruction set can be utilized to it’s fullest. The literature review will pave a way to the project objective to make the optimization more unique to ARM chips. Additionally, research will also be carried out to understand the scheduling process of the ARM chips’ big.LITTLE technology that can provide answers and hints on how to better implement concurrency at peak efficiency.

Part 2: Implementation

As mentioned, there are several open-source pre-existing HFT models, this project will use these open source projects as the starting point and will build from there. Many of such projects are written in either Python or C++, however, NEON architecture provides vectorizing compiler support for C/C++ to enable higher levels of parallelism for higher performance, therefore C++ would be used for the optimization.

The primary focus of the optimization will be on Vectorizing the existing code, where the operation will be applied at once to the entire data. Vectorization entails the replacement of all the explicit iterators (ie, for loops) with more flexible solutions such as matrix algebra operations, iterators or generators. With vectorization, the process could be parallelized with faster running times, potentially leading to faster HFT algorithms. Similarly, Multithreading is also of great focus in this project to push the ARM chip to its max, especially when targeting vectorization and multithreading simultaneously in order to achieve two levels of parallelization.

Following the implementation of vectorization & concurrency, a certain other few additions will be made to the HFT algorithm to make use of the parallelism that has been deployed to further enhance the throughput. For instance, addition of information-driven bars that could help to make decisions before prices reach an equilibrium following the arrival of a certain news. One such example of information-driven bars is the Tick Imbalance Bars.

As for the general flow of the HFT algorithm, the algorithm would be using APIs such as Yahoo Finance to get access to financial data, the data would then flow into the optimized model and the model would output specific trades to execute.

Part 3: Appraisal of performance

As outlined in the objectives, following the complete optimization of the HFT algorithms, including the vectorization and multithreading, and other HFT model enhancements, a performance review would be carried out.
The first method of reviewing the performance involves comparing the latency of order placement between the ARM chips and the x86-64 chips, where even a difference of milliseconds could make a difference, hence this being a very important test to review.

The second method of reviewing the performance of the optimized model is by evaluating the entire portfolio performance. The HFT algorithm will be connected with a paper trading account, where no real money is at stake, but all other factors emulate the real-world performance, and hence observing real-time changes made to the model. Here the Trader WorkStation (TWS) API from a brokerage firm, Interactive Brokers, that allows paper trading will be used to relay the automated trades to be placed and executed, additionally it provides access to market data, current portfolio, and trades.

From these metrics, there would be a clearer picture of how effective the ARM architecture is to process optimized & parallelized HFT algorithms.
5. Schedule & Milestones

This project entails the following deliverables by the end:
1. A fully working, optimized HFT bot that has been parallelized, with automated order placement to a paper trading account.
2. Research findings on the architecture of ARM & that of X86-64 and the discrepancies amongst them.

Below is the schedule of the project along with the major milestones, along with a Gantt chart illustrating the timeline.

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1: Ideation and research</td>
<td>August – September 2022</td>
</tr>
<tr>
<td>- Background research on existing implementations of HFT</td>
<td></td>
</tr>
<tr>
<td>- Familiarization with NEON instruction set, key differences in x86-64 and in ARM chips for vectorization &amp; multithreading</td>
<td></td>
</tr>
<tr>
<td>Deliverable: Project Plan &amp; Project Website</td>
<td>2nd October, 2022</td>
</tr>
<tr>
<td>Phase 2: Project Development</td>
<td>October – December 2022</td>
</tr>
<tr>
<td>- Literature review on ARM architecture</td>
<td></td>
</tr>
<tr>
<td>- Implementation (50%)</td>
<td></td>
</tr>
<tr>
<td>- Documentation</td>
<td></td>
</tr>
<tr>
<td>Phase 2: (cont’d)</td>
<td>January – February 2022</td>
</tr>
<tr>
<td>- Implementation (100%)</td>
<td></td>
</tr>
<tr>
<td>- Evaluation (30%)</td>
<td></td>
</tr>
<tr>
<td>- Interim Report</td>
<td></td>
</tr>
<tr>
<td>Deliverable: 1st Presentation</td>
<td>9-13 January 2023</td>
</tr>
<tr>
<td>Deliverable: Interim Report; Prelim Implementation</td>
<td>22 January 2023</td>
</tr>
<tr>
<td>Phase 3: Evaluation &amp; Findings</td>
<td>March – April 2023</td>
</tr>
<tr>
<td>- Evaluation (100%)</td>
<td></td>
</tr>
<tr>
<td>- Visualize findings &amp; Recommendations</td>
<td></td>
</tr>
<tr>
<td>- Documentation</td>
<td></td>
</tr>
<tr>
<td>Deliverable: Final Implementation; Final Report</td>
<td>18 April 2023</td>
</tr>
<tr>
<td>Deliverable: Final Presentation</td>
<td>17-21 April 2023</td>
</tr>
<tr>
<td>Deliverable: Project Exhibition</td>
<td>3 May 2023</td>
</tr>
</tbody>
</table>

Table 1: Project Schedule
The first implementation of an optimized HFT algorithm on an ARM chip

---

**Project Plan**

Table 2: Gantt Chart illustrating project schedule

<table>
<thead>
<tr>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ideation &amp; Research</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project Plan &amp; Website</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Literature Review</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Presentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interim Report &amp; Preliminary Implementation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Report, Final Implementation &amp; Final Presentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The first implementation of an optimized HFT algorithm on an ARM chip

References


