BOSTON UNIVERSITY

METROPOLITAN COLLEGE

Thesis

AN EXPERIMENTAL STUDY OF MEMORY MANAGEMENT WITH RUST PROGRAMMING FOR BIG DATA PROCESSING

by

SHINSAKU OKAZAKI

B.S., Seikei University, 2018

Submitted in partial fulfillment of the

requirements for the degree of

Master of Science

2020

Approved by

First Reader

Kia Teymourian, PhD
Assistant Professor of Computer Science

Second Reader

Eugene Pinsky, PhD
Associate Professor of Computer Science

Third Reader

Reza Rawassizadeh, PhD
Assistant Professor of Computer Sicence

Facilis descensus Averni;
Noctes atque dies patet atri janua Ditis;
Sed revocare gradum, superasque evadere ad auras,
Hoc opus, hic labor est. Virgil (from Don's thesis!)

Acknowledgments

Here go all your acknowledgments. You know, your advisor, funding agency, lab mates, etc., and of course your family.

As for me, I would like to thank Jonathan Polimeni for cleaning up old LaTeX style files and templates so that Engineering students would not have to suffer typesetting dissertations in MS Word. Also, I would like to thank IDS/ISS group (ECE) and CV/CNS lab graduates for their contributions and tweaks to this scheme over the years (after many frustrations when preparing their final document for BU library). In particular, I would like to thank Limor Martin who has helped with the transition to PDF-only dissertation format (no more printing hardcopies – hooray !!!)

The stylistic and aesthetic conventions implemented in this LaTeX thesis/dissertation format would not have been possible without the help from Brendan McDermot of Mugar library and Martha Wellman of CAS.

Finally, credit is due to Stephen Gildea for the MIT style file off which this current version is based, and Paolo Gaudiano for porting the MIT style to one compatible with BU requirements.

Janusz Konrad

Professor

ECE Department

AN EXPERIMENTAL STUDY OF MEMORY MANAGEMENT WITH RUST PROGRAMMING FOR BIG DATA PROCESSING

SHINSAKU OKAZAKI

ABSTRACT

Many existing Big Data processing tools are developed with application languages. Dataflow platform such as Hadoop MapReduce, Apache Spark, and Apache Flink developed on Java Virtual Machine (JVM). JVM abstracts hardware and memory management from the developer. This eases development processes and distribution of these tools by eliminating manual memory management and platform dependency.

However, the automated memory strategy, such as Garbage Collection (GC), may lead significant overhead in Big Data processing. Apache Spark and Apache Flink use complex objects to manipulate and transfer tremendous amount of data. Generating many of these complex objects in memory forces GC to rearrange the objects in memory frequently waisting computation. Keeping these objects in memory for long period of time may trigger stop-the-world where the JVM stops application from running to execute a GC.

Considered these problems in memory management in JVM, developing Big Data processing tools with system language can be the solution. By using system language, a developer has control on memory management, so that one can implement systems with better optimized memory management strategies. We selects Rust as good candidate for development of Big Data processing tools, due to its ability to write memory-safe and fearless concurrent codes.

To implement optimal memory management in Rust for such tools, there may be variety of strategies and tools. In this dissertation, we conduct experiments to examine the best memory management strategies in Rust for Big Data processing.

Contents

List of Tables

List of Figures

List of Abbreviations

The list below must be in alphabetical order as per BU library instructions or it will be returned to you for re-ordering.

CAD	.....	Computer-Aided Design
CO	.....	Cytochrome Oxidase
DOG	.....	Difference Of Gaussian (distributions)
FWHM	.....	Full-Width at Half Maximum
LGN	.....	Lateral Geniculate Nucleus
ODC	.....	Ocular Dominance Column
PDF	.....	Probability Distribution Function
$\mathbb{R}^2$	.....	the Real plane

Chapter 1

Introduction

1.1 Motivation

Importance of cluster computing tool for Big Data Analysis has been increasing as amount, value, use of data has increased. Recently, almost all businesses stand on data, from web marketing analysis to factory automations. The leverage of data is ubiquitous, because there are many open source tools to analyze data and cloud computer infrastructure which can support computation for massive amount of data. The improvement of accessibility to these technologies has democratized data driven businesses by eliminating significant amount of initial investment.

However these technologies do not come for free; we need to pay money for use of these resources. Usually, user needs to pay depending on use of computational resources. If your process of data analysis is too long or need to use number of clusters with high speck specification, the cost may end up significantly hight. To address these problems, the quality of analysis tool is critical. If the tool can optimize the runtime performance and usage of computational resources, the cost for running the businesses can become efficient.

Multiple cluster computing analysis tools have been developed, such as Hadoop MapReduce [2], Apache Spark [3], and Apache Flink [1]. These tools have brought reliable and scalable ways to deal massive data. These has become widely popular, in which data-parallel computations are executed on clusters of unreliable machines by systems that automatically provide locality-aware scheduling, fault tolerance, and load balancing.

These tools are constructed on top of Java Virtual Machine (JVM). JVM abstracts hardware and memory management from the developer so that the development is fairly easy. In addition, Java or Scala compiled code is platform-independent, which can run on any ma-

chine with JVM. However, these advantages may be really critical weakness when it comes to processing big data. JVM abstract away most detail regarding memory management from the system designer, including memory deallocation, reuse, and movement, as well as pointers, object serialization and deserialization. Since managing and utilizing memory is one of the most important factors determining Big Data systems' performance, reliance on a managed environment can mean an order-of-magnitude increase in CPU cost for some computations. This cost may be unacceptable for high-performance tool development by an expert.

To overcome these problems, one can use programming languages with more control on hardware, system languages, for development of Big Data tools. For example, C++ is a general-purpose, statically typed, compiled programming language which supports multiple programming paradigm. It is also a system language which gives full control over hardware. There are several researches or projects [12] where developers and researchers implement Big Data tools with this language. These tools shows significantly better performances than those developed with application languages. Although the evidence of the advantage of building high speed computational tools with C++ has been discovered, the steep learning curve and difficulty of writing memory safe codes are barrier to technology diffusion.

Rust is a system language which gives the similar performance and control of hardware to C++ or C and safety of runtime. The memory-safety, and fearless concurrently in Rust programming make the language one of the ideal candidate for development of Big Data tools. Since the design of the language is different from any other programming languages, implementations that can be selected for algorithms can differ from existing ones. In this dissertation, we focus on memory management strategy for Big Data processing algorithms in development with Rust.

1.2 Problem Description

Many of popular open source cluster computing frameworks for large scale data analysis, such as Hadoop and Spark, allow programmers to define objects in a host languages, such

as Java. The objects are then managed in RAM by the language and its runtime, Java Virtual Machine in the case of Java and Scala. Storing objects in memory enables machine to process iterative computation. One of the fundamental tasks for recent big data analysis is analysis using Machine Learning Algorithms, which require iterative process. As the amount of data increases, memory is required to keep many objects. Therefore, memory management plays a critical role in this task.

One of the memory management tool on JVM is garbage collection. The garbage collection brings a significant advantage for programmers by removing responsibility for planning memory management by themselves manually. Instead, JVM monitors the state of memory and performs garbage collection at certain points. However, these monitoring and auto-execution of garbage collection cost additional computation and might consume computation resources which should be used for data processing. This can significantly decrease performance of the computation.

In contrast, memory management in system language, such as C++, relies on programmers' decision for when to allocate and deallocate memory. The functions, malloc/free consume most of the memory management. Proper implementation of system language for big data processing can be overperform the implementation in host language. Nevertheless, implementing C++ performing proper memory management and guaranteeing security can be unproductive and complicated.

Considering the issue of memory management, we introduces solution based on unique memory management methods implemented in Rust, ownership , move and borrowing. This unique concepts in Rust secure codes and perform memory management without monitoring memory. Since developer can select variety of memory management strategies to implement Big Data processing algorithms in Rust. We develop such algorithms with different strategies and compare their runtime performance to study the best algorithm implementation.

Chapter 2

Related Work

2.1 Operating Systems

2.1.1 Memory and Process in Operating Systems

A process is a subsection of computation job. A process can work on a CPU core. We can divide process as well. Basically, each process does not share their memory. However, for multiprocessing, we could avoid this restriction. Processes can be represented as tree structure, because a process may create other child processes. Process has 4 states, new, running, waiting, and ready. Process is represented in process control block (PCB) with state type, process ID, registers, and so on. The scheduling for process assigning to CPU core is implemented in queues containing PCB. There are two main queues in this scheduler: ready queue and wait queue. The head of process in ready queue is selected for execution and once the process requested I/O request or production of child process, the running process will be stored n wait queue. Once the request that the process waiting for end, the waiting process will be pushed tail of ready queue.

Processes executing concurrently in the operating system may be either independent processes or cooperating processes executing in the system. A process is independent if it does not share data with any other processes. A process cooperating if it can affect or be affected by the other processed executing in the system. In cooperating process, there are two kinds, shared memory and message passing. In shared memory, it removes restriction of not interfering memory region. Message passing can be useful for distribution systems as well.

For a pair of processes to communicate through message, a socket is needed to be established. A socket is identified by an IP address concatenated with a port number.

When two process communicate, each process will have socket. If another process of the same machine wants to communicate, we need new socket to be established. The protocol used in the socket connection can be TPC and UDP.

2.1.2 Multi-threading and Parallelism

A thread is a basic unit of CPU utilization, so that a process can have multiple thread. Threads share mainly code and data. Multi-threading is increasingly popular as the multicore programming becomes in common, because we can run multiple thread on different core. Creating thread is much cheaper than creating process and it shares resources so that we do not need additional methods to allow threads to communicate each other, such as sharing memory and message passing.

2.1.3 Memory management in Operating System

In computer storage hierarchy, the closest storage to CPU is register. It is built into each CPU core and accessible within one cycle of the CPU clock. However, the same cannot be said of main memory, which is accessed via a transaction on the memory bus. This takes many cycles of the CPU clock. The remedy is to add fast memory between the CPU and main memory, typically on the CPU chip for faster access. Such a cache plays a role for this.

For the layout of main memory, it must be ensured that each process has a separate memory space, including operation system. The base register and limit register, whose roles are lower bound of memory region and specific size of range respectively, can achieve that goal.

Usually a program resides on a disk as binary executable file. To run, the program must be brought into memory and placed within the context of a process. The process is bound to corresponding parts of the physical memory. Binding program to memory address is staging process. There three stages: compile time, load time execution time. The source program is compiled by compiler producing object file. After the compilation, the object file is linked with other object file by linker creating executable file . Finally, the executable

file will be loaded to run execute. At this run time dynamic library link can be done.

If where the process will reside in memory at compile time, absolute code is generated. If this is unknown at the time, the binding will be done at load time. At this time, the compiler must generate relocatable code. Otherwise, the binding will be done at execution time.

A process does not interact with addresses of physical memory, instead virtual memory. The memory-management unit (MMU) takes roles to map logical address to physical address. OS needs to ensure that any of physical memory spaces of processes do not overlap. Since one process can be created and deleted and the corresponding memory space should be de/allocated, optimization for use of physical memory space is important; we need to allocate memory contiguously avoiding fragmentation.

There are several approaches to deal with this problems. However, we will focus on paging here, which is the most used method OS use to manage memory. A frame and page are a unit of Separated physical and virtual memory space in fixed size (4KB or 8KB) respectively. A process can use as many as pages and corresponding frames obtained by page table matching. This strategy does improve external fragmentation, but not for internal fragmentation. The smaller size of page has smaller fragmentation, but mapping from page to frame has overhead and also disk I/O is more efficient when the amount of data transfered is larger.

2.1.4 Demand Paging

A process can have multiple pages. However, loading entire executable code from secondary storage to memory is not necessarily needed to get jobs done. A strategy used in several operating systems is loading only the portion of programs that are needed, demand page.

In the storage, some pages currently used are in memory and the others are in secondary storage. The page table specifies whether pages are valid or invalid, which means are in memory or not. Access to a page marked invalid causes page fault and some steps to resolve the error will be required.

The first part of process of demand paging would be that we check an internal table to check whether the reference is valid or invalid. If the reference is valid, the process reads the content from the memory. Otherwise, we terminates the process and find the free frame in physical memory. Then, we schedule a secondary storage operation to read the desired page into newly allocated frame. When the storage complete reading the page, we modify the internal table to indicate the page is now in memory. Finally, we restart the instruction that was interrupted.

However, there would be a case where the memory does not have any free frame. In this case, a victim frame that will be replaced with new coming frame should be selected. To perform this selection efficiently, a modify bit is tracked for each frame or page. The modify bit represents whether the page is modified since it is loaded from secondary storage. If the page or frame is modified, when we swap page we need to update the content in the secondary storage. However, it is not modified, we can simply delete the frame and replace with new frame.

2.1.5 Copy on Page

When a parent process creates child process and if these process shares contents on particular page and modify it, only the page which has the context will be copied.

2.2 Linear Algebra Computation

2.2.1 BLAS LAPACK

Basic Linear Algebra Subprograms (BLAS) are standard building blocks for basic vector and matrix operations. There are 3 levels of operation. The level 1 BLAS performs scalar, vector and vector-vector operations, the level 2 BLAS performs matrix-vector operation, and the level 3 BLAS performs matrix-matrix operation.

LAPACK is developed on BLAS and has advanced functionalities such as LU decomposition and Singular Value Decomposition (SVD). Dense and banded matrices are handled, but not general sparse matrices. The initial motivation of development of LAPACK was to

make the widely used EISPACK and LINPACK libraries run efficiently on shared-memory vector and parallel processors. LINPACK and EISPACK are inefficient because their memory access patterns disregard the multi-layered memory hierarchies of the machines, thereby spending too much time moving data. LAPACK addresses this problem by recognizing the algorithms to use block matrix operations, such as matrix multiplication, in the innermost loops. These block operations can be optimized for each architecture to account for the memory hierarchy, and so provide a portable way to achieve high efficiency on diverse modern machines. However, LAPACK requires that highly optimized block matrix operations be already implemented on each machine.

ARPACK is also a collection of linear algebra subroutines which is designed to compute a few eigenvalues and corresponding eigenvectors from large scale matrix. ARPACK is based upon an algorithmic variant of the Arnoldi process called the Implicitly Restarted Arnoldi Method (IRAM). When the matrix is symmetric it reduces to a variant of the Lanczos process called the Implicitly Restarted Lanczos Method(IRLM). The Arnoldi process only interacts with the matrix via matrix-vector multiplies. Therefore, this method can be applied to distributed matrix operations required in big data analysis.

The original BLAS and LAPACK are written in Fortran90. Linear algebra library used for Spark is netlib-java, which is a Java wrapper library for Netlib, C API of BLAS and LAPACK. The reason why the developers addressed to use this package is that the BLAS and LAPACK are already bug free and implementing linear algebra library from scratch can usually buggy.

However, the main advantage of use of BLAS and LAPACK is system optimized implementation. So if we implement original Fortran linear algebra library, it cannot perform as well as BLAS and LAPACK. And the performance would not be such different from one of implementation in Java or Rust. If we want to test only memory management between Rust and Java, it can be enough implementation of linear algebra operation from pure Java and Rust sacrificing the best performance taking advantage of system optimization.

2.2.2 Netlib-Java

Netlib-java is a Java wrapper of BLAS, LAPACK, and ARPACK. Netlib-java choose implementation of linear algebra depending on installation of the libraries. First, if we have installed machine optimised system libraries, such as Intel MKL and OpenBLAS, netlib-java will use these as the implementation to use. Next, it try to load netlib references which netlib-java use CBLAS and LAPCKE interface to perform BLAS and LAPACK native call. The last option is to use f2j which is intended to translate the BLAS and LAPACK libraries from their Fortran77 reference source code to Java class files, instead of calling native libraries by using Java Native Interface (JNI).

We can use JNI to call native libraries from Java. The JNI is a native programming interface which allows Java code that runs inside a Java Virtual Machine (VM) to interoperate with applications and libraries written in other programming languages.

2.2.3 Matrix Computation and Optimization in Apache Spark

Matrix operation is a fundamental part of machine learning. Apache Spark provides implementation for distributed and local matrix operation. To translate single-node algorithms to run on a distributed cluster, Spark addresses separating matrix operations from vector operations and run matrix operations on the cluster, while keeping vector operations local to the driver.

Spark changes its behavior for matrix operations depending on the type of operations and shape of matrices. For example, Singular Value Decomposition (SVD) for a square matrix is performed in distributed cluster, but SVD for a tall and skinny matrix is on a driver node. This is because the matrix derived among the computation of SVD for tall and skinny matrix is usually small so that it can fit to single node.

Spark uses ARPACK to solve square SVD. ARPACK is a collection of Fortran77 designed to solve eigenvalue problems. ARPACK is based upon an algorithmic variant of the Arnoldi process called the Implicitly Restarted Arnoldi Method (IRAM). When the matrix A is symmetric it reduces to a variant of the Lanczos process called the Implicitly

Restarted Lanczos Method (IRLM). ARPACK calculate matrix multiplication by performing matrix-vector multiplication. So we can distribute matrix-vector multiplies, and exploit the computational resources available in the entire cluster. The other method to distribute matrix operations is Spark TFOCS. Spark TFOCS supports several optimization methods.

To allow full use of hardware-specific linear algebraic operations on single node, Spark uses the BLAS (Basic Linear Algebra Systems) interface with relevant libraries for CPU and GPU acceleration. Native libraries can be used in Scala are ones with C BLAS interface or wrapper and called through the Java native interface implemented in Netlib-java library and wrapped by the Scala library called Breeze. Following is some of the implementation of BLAS.

- f2jblas - Java implementation of Fortran BLAS
- OpenBLAS - open source CPU-optimized C implementation of BLAS
- MKL - CPU-optimized C and Fortran implementation of BLAS by Intel

These have different implementation and they perform differently for the type of operation and matrices shape. In Sark, OpenBlas is the default method of choice. BLAS interface is made specifically for dense linear algebra. Then, there are few libraries that efficiently handle sparse matrix operations.

2.2.4 Memory Management of each Linear Algebra Library

The pure Java linear algebra library, such as La4j, EJML, and Apache Common Math, use normal GC performed by JVM to manage memory. This is because the implementation of these libraries are in purely Java.

Netlib-java, Jblas or other simple Java wrapper of BLAS, LAPACK, and ARPACK with Java Native Interface (JNI) use normal GC as well. This is because the native code deals with Java array by obtaining a reference to it. After the operation, the native method releases the reference to the Java array with or without returning new Java array or Java primitive type object.

ND4J has two types of its own memory management methods, GC to pointer of off-heap NDArray, and MemoryWorkspaces. ND4J used off-heap memory to store NDArrats, to provide better performance while working with NDArrays from vative code such as BLAS and CUDA libraries. Off-heap means that the memory is allocated outside of the Java heap so that it is not managed by the JVM's GC. NDArray itself is not tracked by JVM, but its pointer is. The Java heap stores pointer to NDArray on off-heap. When a pointer is dereferenced, this pointer can be a target of JVM's GC and when it is collected, the corresponding NDArray will be deallocated. When using MemoryWorkspaces, NDArray lives only within specific workspace scope. When NDArray leaves the workspace scope, the memory is deallocated unless explicitly calling method to copy the NDArray out of the scope.

2.3 Hadoop MapReduce to Spark

In several Big Data mining tools that have been developed, the most notable ones are MapReduce and Spark. MapReduce is a cluster computing framework which supports locality-aware scheduling, fault tolerance, and load balancing. Spark is designed to improve performance for iterative jobs keeping features of MapReduce.

MapReduce provides a programming model where the user creates acyclic data flow graphs to pass input data through a set of operations. This data flow programming model is useful for many of query applications. However, MapReduce framework struggles from two of main resent data mining jobs: iterative jobs and interactive analysis.

Iterative job is especially common in Machine Learning Algorithm, such as learning a data model using Gradient Descent. In traditional MapReduce framework, each iteration can be expressed as single MapReduce job so that each job must reload the data from disk. This leads I/O overhead and deteriorates performance of iterative algorithms.

Interactive analysis is also an inevitable task in modern data science. A data scientist wants tp perform exploratory analysis in interactive way. Nevertheless, MapReduce is designed in the way more stable for ad-hoc query, so each analysis can be single MapReduce

job. To perform multiple analysis to explore dataset, the data needs to be written to and reload from disk many times.

To overcome these limitations, Spark has been developed as a new cluster computing framework maintaining the innovative characteristics of MapReduce and improving its iterative and interactive jobs with in-memory data structure.

Some of the notable improvements are shown here[9]. For aggregation operation, map output selectivity, which is the ration of the map output size to the job input size, can be significantly reduced by using Spark. Spark uses a map side combiner, hash-based aggregation which is more efficient than sort-based aggregation used in MapReduce. For iterative operation, caching the input as Resilient Distributed Datasets (RDDs) can reduce CRU and disk I/O overheads for sequence iteration. This RDD caching takes a significant role to improve iterative job, because it is more efficient than other low-level caching approaches such as OS buffer caches, and HDFS caching. These caching strategy reduces disk I/O overhead, but CPU overhead, such as parsing text to objects.

2.4 Apache Spark

2.4.1 Resilient Distributed Datasets

Major methods in Spark are Resilient Distributed Datasets (RDDs), a data structure that abstracts distributed memory across different clusters. The immutable coarse-grained transformation, spark-scheduler with lazy-evaluation, and memory management with cacheing achieve computation with fault-tolerance, fast execution, and moderate control on memory efficiency [11].

A RDD is essentially a multi-layer Java data structure. A top RDD object references Java array, which intern, references a set of tuple objects. The coarse-grained transformations and immutability requires a RDD to be deep-copied to produce a new RDD, but efficiently offers fault tolerance. The lost partitions of a RDD can be recomputed in parallel on different nodes rather than rolling back the whole program.

Spark-pipeline consists of sequence of transformations and actions over RDDs. A trans-

formation produces a new RDD from a set of existing RDD. An action is method that computes statistics from an RDD. Due to lazy-evaluation nature, transformations do not materialize the newly created RDD. Instead, RDD Lineages are created. Lineage is a graph among parent and child RDDs which represents logical execution plan. This enhance fault-tolerance and improve ability to optimize execution plan.

RDDs can be cached in memory for faster access by persist method. Developers can specify a storage level for a persisted RDD, in memory with serialized or deserialized, or on disk. Other than persisted RDD, Spark generates a lot of intermediate RDDs during execution. Since RDD is a Java object, they are managed by Garbage Collection (GC) in the JVM. However, persisted RDDs are never collected by GC. This GC might cause significant deterioration of performance of Spark, because GC shows heavy overhead when there are a number of objects.

2.4.2 Memory Management in Spark

Spark framework allocates multiple executors, JVMs, that run sequence of transformations and actions. As we describe the previous section, data in Spark is mainly stored as Java objects in memory, so that they are allocated on JVM heap and managed by JVM Garbage Collection. The data may form three types [10]: Cached data, Shuffled data, and Operator-generated data.

Spark can cache data in memory to reduce disk I/O. This Cached data usually long-lived Java objects and span multiple stages in Spark-pipeline. Spark allocates a logical storage space to store the cached data as shown in Figure. After aggregation, Spark generates Shuffled data. Shuffled data is usually long-lived, because it need to be kept in memory until the task ends. Spark allocates execution space to store Shuffled data. The storage space and execution space spans 60 % of JVM heap space in default. Operator-generated data is data generated by user-defined operations. Since Operator-generated data may or may not be used, after the operation finish, the data object can be both short-lived or long-lived objects. These are stored in user space allocated on default 40 % of JVM heap.

All data of these types on JVM heap is managed by JVM GC. GC check references graph of objects, mark whether the objects are used and deallocate memory space occupied by unused objects. There are three popular GCs: Parallel, CMS and G1. All of these method track generation of object based on the region of memory. The Java logical heap structure is shown in Figure 2·1. As the figure shows, Java logical heap can be separated into three main parts where store objects for each corresponding generation: permanent generation, young generation, and old generation. The region for permanent generation stores metadata required by JVM to describe class and method used in application which will be permanently lived on the region of memory. The overview of Java GC is shown in Figure 2·2. The outer box represents region for particular generation. The inner box and the number represents object and its age in GC.

The region for young generation mainly consists of two parts: Eden and Survivor space. First, Java objects are created in Eden space and promoted to Survivor space when survive from GC. After objects survive several GCs in Survivor space, they are finally promoted to old generation.

In region of old generation, JVM lunches multi-thread to perform GC. GC with multi-threading suffers from Stop-The-World (STW) pauses; GC may suspend application threads while performing object marking and deallocation. Different GC algorithms try to solve this problem with trade-off between GC frequency and memory utilization.

Because of the problem of STW and copying objects to different physical memory pages , JVM GC cause huge overhead when number of objects is large. Therefore, GC become severe issue in Big Data processing where might produce significant number of object.

2.4.3 Garbage Collection Tuning

There are many different ways that one can improve the performance of GC in Java. One of these is for example avoiding pointer-based data structures, such as HashMap and LinkedList. These objects have a "wrapper" object for each entry so that number of object tends to be larger than when when an array is used.

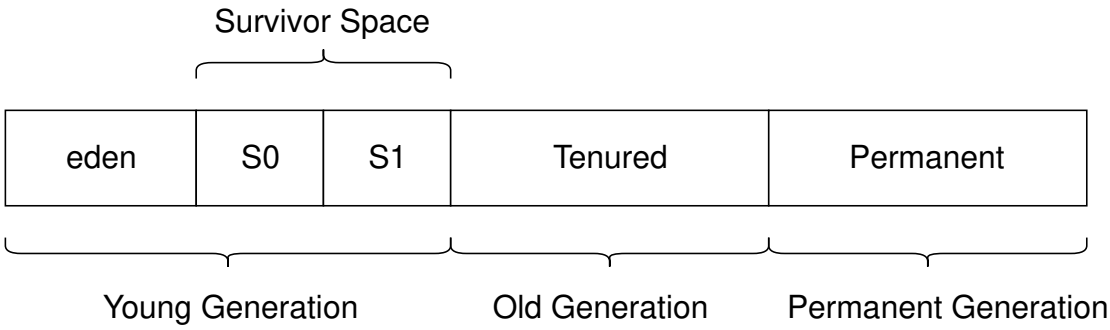

Figure 2·1: Java Heap Structure

Caching serialized object in memory also reduce number of object and memory usage since the set of objects become a byte or binary array. Spark SQL applications use DataFrames[5], whose intermediate data are managed by an optimized memory manager named Tungsten. Tungsten stores the intermediate data in a serialized binary form and performs aggregation functions directly on he serialized objects. Therefore, the number of Java objects in memory is reduced and it reduces the DC frequency and object marking/sweeping.

In addition, developers can allocate data off-heap of JVM to avoid tracking by GC. Facade[8] proposed a compiler and runtime system to bound the number of in-memory data objects, through storing data in an off-heap region and manipulating the data with control interfaces.

Although these memory management solution for GC help developer improve performance of Spark applications, the effort to discover the best GC tuning afflicts developers.

2.5 Application to System Language

Considered overhead produced by Memory Management in JVM, use of system language for development of Big Data tool can be one of solutions rather than application language, such as Java and Scala. System languages, such as C and C++, are languages that give developer total control over the hardware and de/allocation of memory without GC. These features enable program written in system languages to optimized performance taking full advantage of hardware.

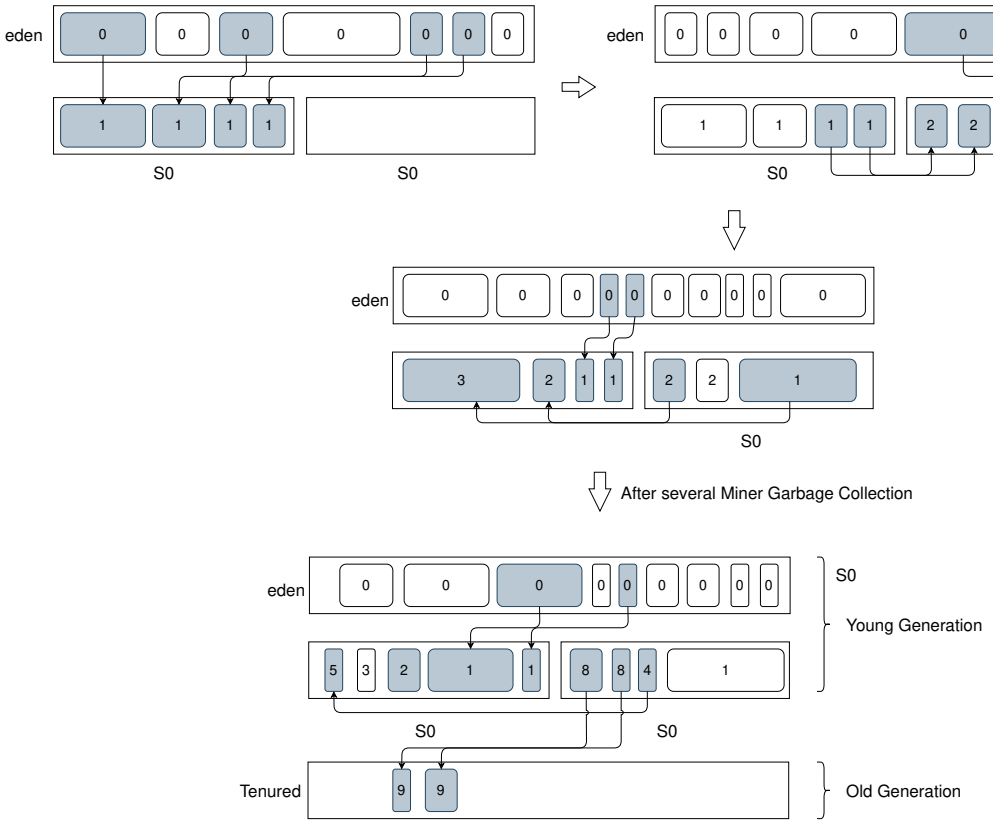

Figure 2·2: Java Garbage Collection

To take example, one can build Big Data tool with C++. C++ is one of the most popular system languages which has Object-oriented features. C++ has functions which provide control over memory to developers. In another word, it is responsibility for one to manage memory properly and safely. The functions, malloc() and free(), take roll for memory allocation and deallocation in respectively. The manual memory de/allocation may cause several problems and require developers attention to the problems with significant effort for debugging and testing. Here, we explain two of the most common problems regarding to memory management in existing system languages.

Dangling pointer or reference is a pointer or reference pointing to object that no longer exists. The situation of dangling pointer happens because of deallocation of memory without modification of value of the pointer. If the memory region is reallocated for other object and the dangling pointer tries to access the original object, the unpredictable behavior may

result.

Memory leak occurs when memory is allocated and no longer referenced so that the object in the memory location cannot be reached and released. This is result of dereferencing object without deallocation. Memory leak consumes more memory than necessary by making unreachable location.

Some solutions are established to address these problems. Actually GC is a high-level solution that guarantees memory safety. C++ has a different solution called Resource Acquisition is Initialization (RAII). In RAII, objects can live within the scope where they are created. The memory is released when the object goes out of scope. This solution is more predictable and deterministic than GC. However, it is problematic When we need the object out of the scope, returning a value from a function. There are several ways to go around this problems, such as smart pointers, copy constructors, and move semantics. Nevertheless, these non-orthogonal concepts makes code in disorganize and leads to error prone implementation.

2.6 Rust Memory Management

Rust is a system programming language which provides memory safety without runtime checking like GC and necessity of explicit memory de/allocation. To ensure memory safety, Rust provides restrictive coding patterns and checks lifetime of value and memory safety at compile-time. The restrictive patterns also enables a developer to write fearless concurrent code that is free of data races. Main concepts of Memory Management in Rust are ownership, move, and borrowing.

2.6.1 Ownership

In ownership feature or Rust, each value has a variable called owner. This owner has information about the value, such as location in memory, length and capacity of the value. For example, the object representation of Vec<i32> is shown in Figure 2·3. The upper boxes represent owner variable in stack frame. The lower boxes represent contiguous memory allocated to store i32. Its capacity is specified 10, but 7 values of i32 are stored. Therefore,

there are still spaces to store 3 values of i32 without reallocation of memory. This owner can live on the scope associated with its lifetime. When the owner is dropped, the value will be dropped too. This feature is similar to how RAII in C++ works. However, acquisition of owner out of the scope where it was constructed is available in Rust with the concept of move.

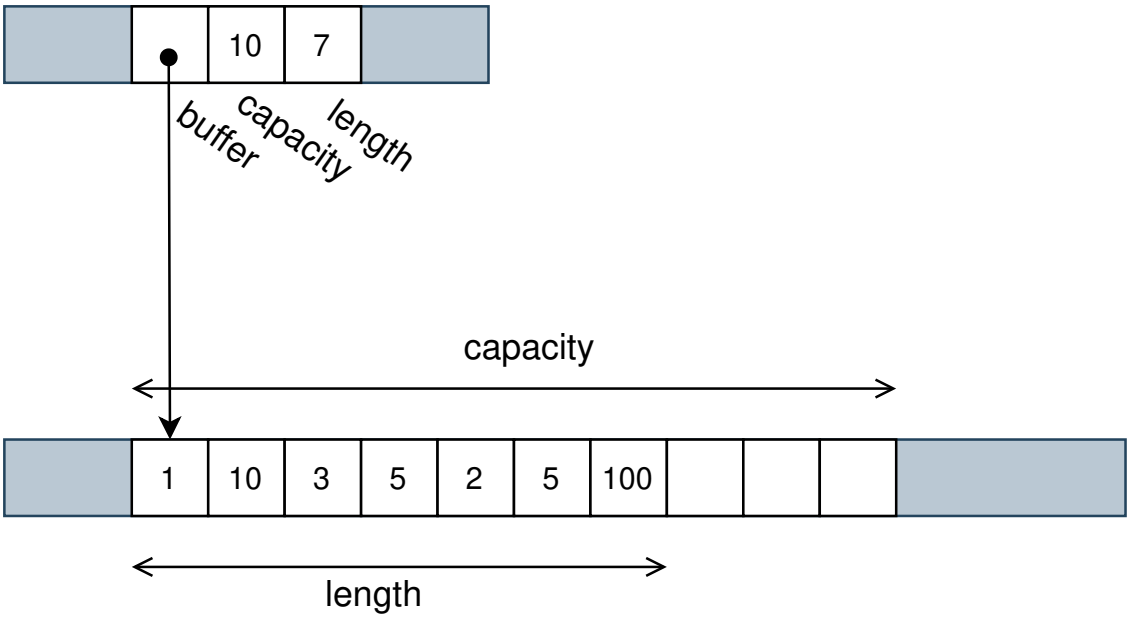

Figure 2·3: Representation of Rust Vec<i32>

2.6.2 Move

In Rust, for most types operations like assigning a value to a variable, passing it to a function, or returning it from a function do not copy the value: they move it. With move, a value can be transferred from one owner to another. The previous variable does not have ownership of the value; it is moved to a newly assigned variable. To understand how this assignment implementation is unique from other programing languages, Java, C++, and Rust code example of assigning list or vector of strings are shown.

Below a few lines of code are initialization of list of string and reassigning it to other variables.

```
List<String >s = new ArrayList<>(
                    Arrays.asList("lemon", "orange", "apple"))
List<String> t = s
List<String> u = s
```

Figure 2·4 represents the memory allocated after a Java ArrayList of strings is initialized and assigned to variable called s. After assigning s to t and to u, the memory representation become Figure 2·5. The assignments are simply setting pointers to the ArrayList object and increment reference count.

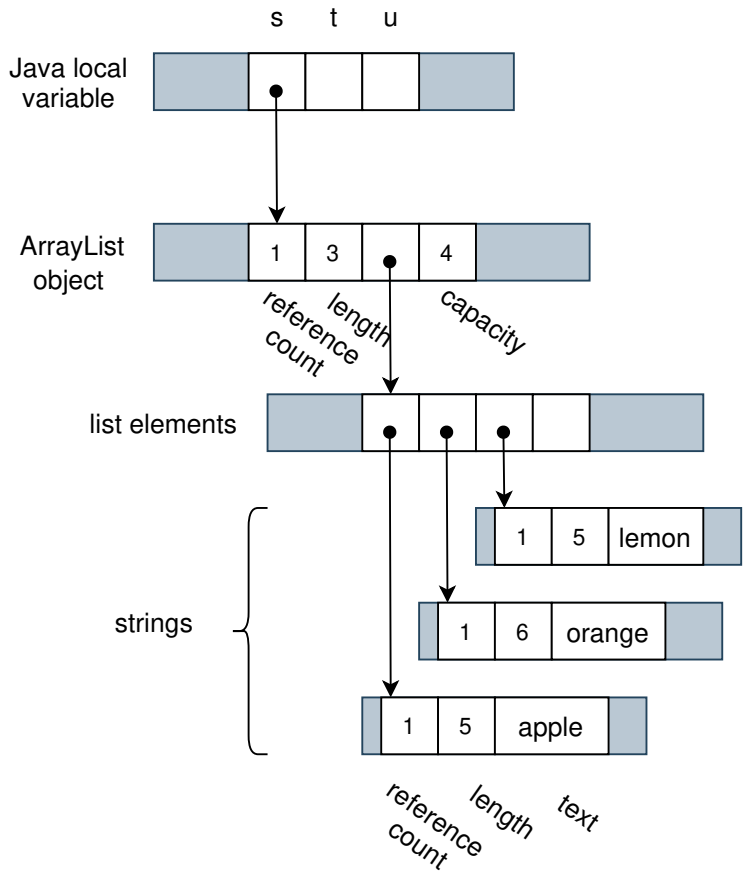

Figure 2·4: Representation of Java ArrayList of String

Now, the analogous C++ code is shown below.

```
vector<string> s = {"lemon", "orange", "apple"};
```

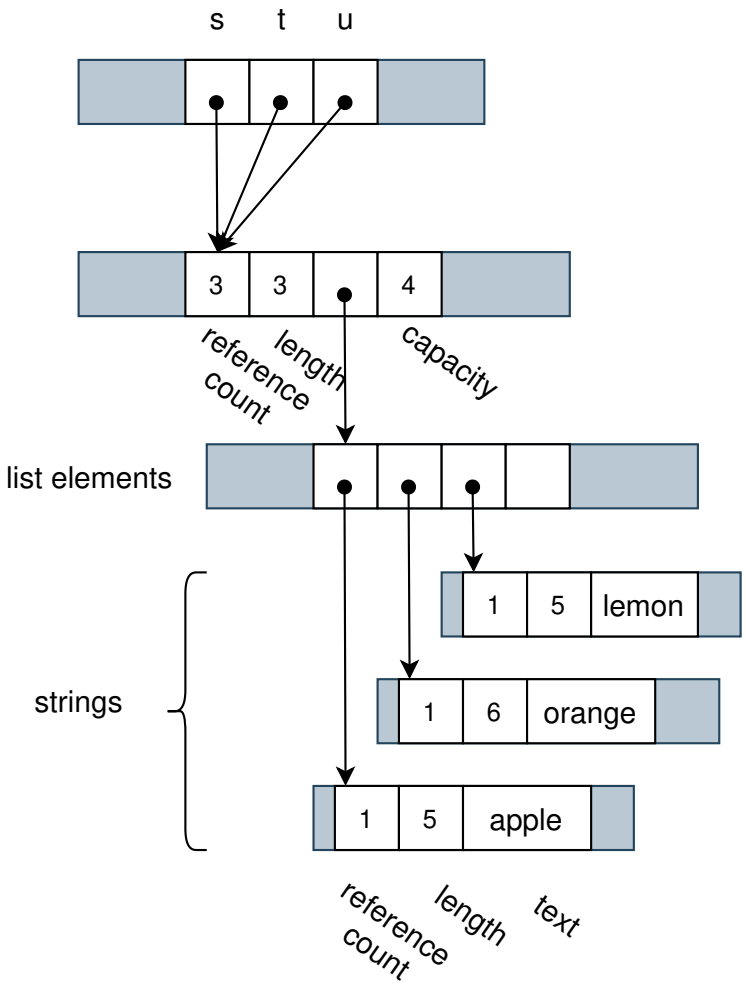

Figure 2·5: Representation of Java ArrayList of String after assignment to another variable

```
vector<string> t = s;
vector<string> u = s;
```

Figure 2·6 shows the memory allocated when the vector is initialized. Figure 2·7 is a representation of after the assignments of s to t and to u. In C++, assigning value of vector to other variables involves allocating memory for new vector and copying the contents of the original vector to newly allocated one.

In Rust code, the code is like below,

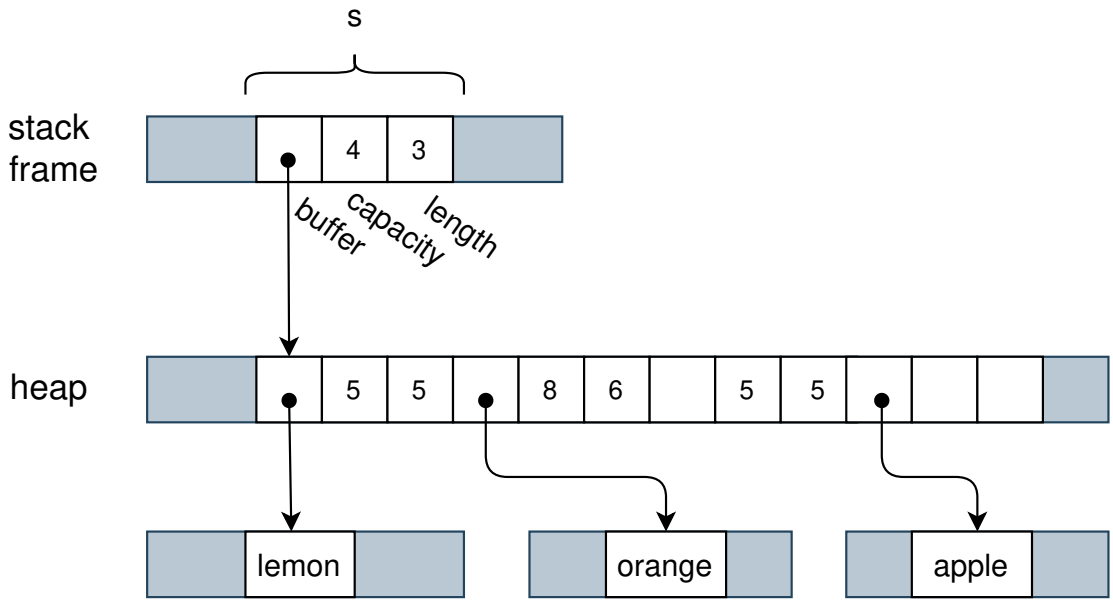

Figure 2·6: Representation of C++ vector of string

```
let s = vec!["lemon".to_string, "orange".to_string, "apple".to_string];
let t = s;
let u = s;
```

The representation of the original Vec of String in Rust is the almost same in C++ vector of string. Figure 2·8 however shows different behavior when Rust assigns s to t. The value is moved to t from s so that the source variable s is uninitialized. The Rust code actually throws compile error, because we are assigning uninitialized variable at the last

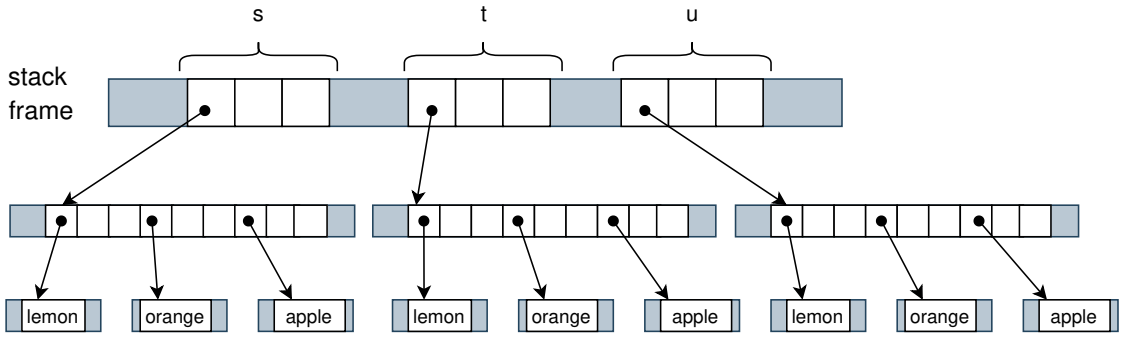

Figure 2·7: Representation of C++ vector of string after assignment to another variable

line.

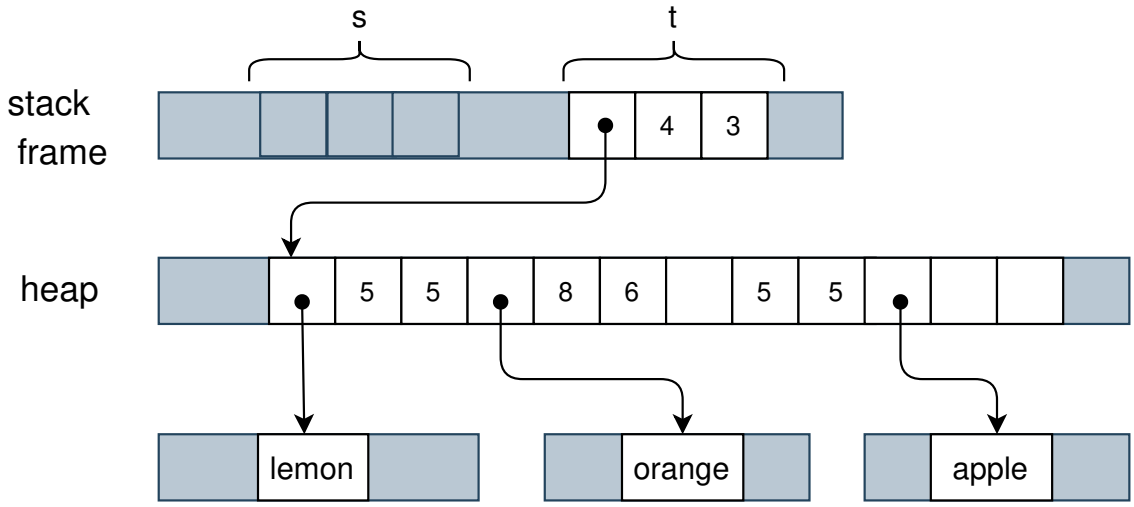

Figure 2·8: Representation of Rust Vec<String> after assignment to another variable

2.6.3 Borrowing

Borrowing lets code use a value temporarily without affecting its ownership so that it reduces unnecessary movement of ownership. One use case is when value is used in function and needed to be passed to the argument. If the argument takes ownership and the function does not return the value, the ownership of value goes out of scope and the memory is deallocated. One can pass reference of the value to the argument instead of owner. The reference goes out of scope, but ownership remains the same.

2.7 LLVM

LLVM (Low Level Virtual Machine) [6] is an umbrella project which contains components of compilation of programming language. Existing compilers have tightly coupled functionalities so that it is not possible to embed them into other applications. However, the abstract framework of LLVM decouples the functionalities into peaces and the peaces of functionalities can be reused.

In structure of a compiler, there are three main components; frontend, optimizer, and

backend. In frontend, a developer designs the interface of source programming language in a way where it can be optimized by optimizer. Then, backend takes optimized code and produce the native machine code.

This separation of functionalities gives reusability to parts of compiler. Establishing new programming language requires new front end, but the existing optimizer and back end can be reused. This speeds up processes of developing new programming languages.

The reusability of compiler leads to support multiple programming languages so that broader set of programmers are involved in the development. For open source project like Rust, This ends up larger community of potential contributors to the development.

Since implementation of front end can be independent on optimizer and back end, development of programming language becomes easy for person without skills required to implement optimizer and back end. This thrives new language developments.

LLVM has a component called LLVM Intermediate Representation (IR), which places itself across frontend to optimizer. IR is designed to host mid-level analyses and transformations that you find in optimizer section of a compiler. High-level language has many common structures and functionalities, so most of all high-level program languages can be represented with IR. Once source code is represented with IR, optimizer can easily find pattern and optimize it in faster time. IR is useful in terms of frontend. This is because developer of language frontend need to know only how the IR works and use the framework to develop a language.

Thanks to emerging of LLVM, many new programming languages and compilers have been developed. Rust compiler is also developed based on LLVM.

Chapter 3

Conceptual Design of Experience

3.1 Type of Variable

In Rust, there are three variable types: owner, reference, and slice (only for sequence of values). A developer is sometimes forced to use specific variable types. For example, some of methods are only implemented to specific variable types. However, one can select any variable types for operation in most case.

These variables have different memory representation shown in Figure 3·1. The owner has a pointer pointing to the memory address of sequence values, length of the values, and capacity allocated to store additional values. Reference and Slice are variables borrowing value owned by other variable. The reference is a pointer that points to the owner. The slice is a pointer that points to memory address of sequence values. It has value such as length of sequence values stored in the memory. Since they have different memory representation, an assumption is that it takes different time to access to the contents of the memory among these pointer types. We examine this by constructing complex objects whose fields are these variable types. The details are explained in section 3.2.

3.2 Reference Count

Reference is useful to avoid movement of ownership. However, one needs to track its lifetime and explicitly includes it in code, because Rust compiler cannot infer it. This can be another encumbrance. We can instead acquire multiple owners to single value by using Reference Counting (Rc). By leveraging Rc, a value can be shared like what borrowing plays the role in Rust programming.

The difference is that Rc checks number of owner pointing to the actual data and makes

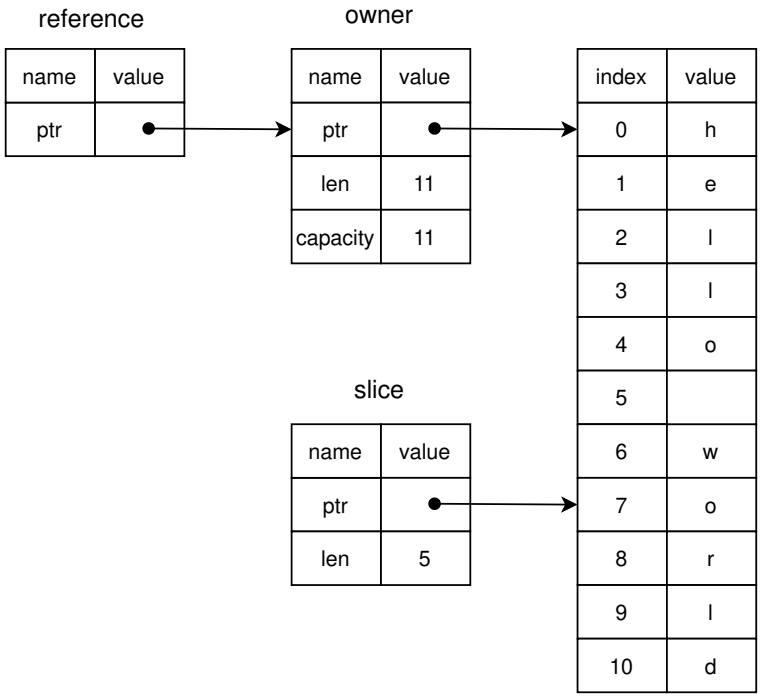

Figure 3·1: Memory Representation of Owner, Reference, and Slice Type

sure the data is not deleted until all the owners are dereferenced. Using Rc is sometimes preferable approach for developers especially when lifetime planning is extremely difficult. However, the possible problems regarding to Rc are the cost for tracking the number of references and allocating memory on heap instead of stack. Having these assumption, an experiment is conducted to examine difference of runtime performance of dropping reference and Rc. This is explained in section 3.2.

3.3 Multithread

In Rust programming, writing concurrent code is relatively easy. The care Rust takes with reference, mutability, and lifetimes is valuable enough in single-threaded programs, but it also is in concurrent programming. Rust has tools to write concurrent code, such as threads, locks, atomic reference. One can implement various concurrent codes for the same purpose with different memory management strategies. The most ubiquitous tool used in Rust concurrent code is Atomic Reference Counting (Arc).

Arc is a simple interface that allows threads to share data. Arc allows multiple variable to have ownerships of a particular value similarly to Rc, but also supports atomic feature enabling the ownerships exist in different threads. In many situation where developer write a multithreading code, the deletion of Arc happens significant amount of times. Similarly to Rc, our assumption is that deletion of Arc has also overhead when we compare to normal reference. To assess runtime performances of algorithm with Arc vs normal reference, we implement merge-sort algorithm in two different ways.

3.4 Tree-aggregate

Finally, we implement some of algorithm common in Big Data processing. One of them is tree-aggregate.

Tree-aggregate is a communication patten heavily used for Machine Learning algorithm in Spark (MLlib [7]). The topology of aggregation patterns in Apache Spark are shown in Figure 3·2. In the traditional aggregation function in Spark, results of aggregation in all executor clusters are sent to the driver. That is why this operation suffers from the CPU cost in merging partial results and the network bandwidth limit. Tree-aggregate is a communication pattern which overcomes these problems by breaking aggregate operation in multi-level represented like tree structure.

Spark generates intermediate objects from RDDs. In Tree-aggregation, aggregated HashMap like data structure is created in each thread or node. When it aggregates objects in RDD, copy of objects should be performed to construct intermediate aggregated data structure. In another possible way that might be implemented, one can use reference to the objects to perform aggregation instead of copying the values themselves. In Rust, we can clone or get Arc (Rc if in single thread) of objects to implement these operations.

Since Tree-aggregation algorithm generates and deletes a lot of intermediate data structures, how the data structures are constructed and how they are deallocated is important concern in memory management in this algorithm. In our experiment, tree-aggregation algorithms are examined in multi-threading. The detail is described in section 3.2.

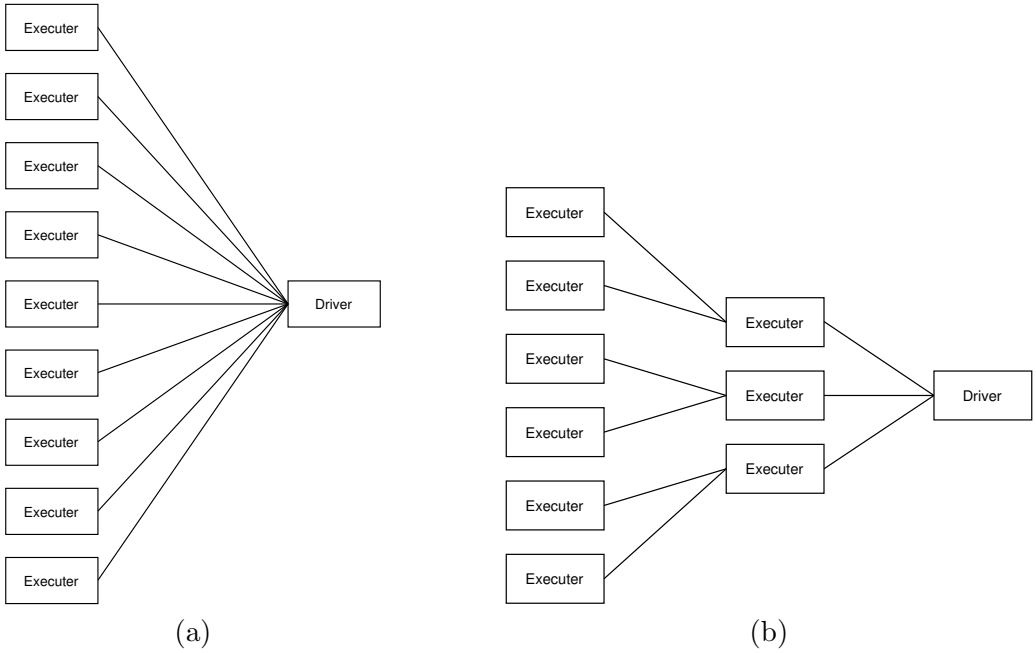

Figure 3·2: Representation of aggregation strategies in Apache Spark: (a) Traditional Aggregation, (b) Tree Aggregation

3.5 K-Nearest-Neighbors

The other algorithm common in Big Data processing is K-nearest-neighbors (KNN). KNN is a traditional Machine Learning algorithm which classify targets into categories. In training KNN, it simply stores all available training data without calculation. At prediction phase, targets are given and similarity measures are calculated. Based on these similarities, the algorithm selects K (user-defined number) training observations similar to the each target. Then, it checks corresponding categories of the K observations to determine predicted category.

In brute force algorithm, KNN calculates similarity measures for all combinations among training and testing observations. If training data has N observations and test data has M observations, KNN needs to calculate $N \times M$ similarity measures. Sort or binary heap can be used to select the K most similar training observations.

In our experiment, KNN algorithms performs document classification and implemented in multithread. There are three phases in our algorithm: preprocessing, query, and combine

phase. In preprocessing phase, the algorithm calculates Term-frequencies (Tfs) to generate numeric feature vectors and matrices. In query phase, similarity measures are calculated and K nearest neighbors are selected. For similarity measure, our choice is cosine similarity(3.1). In combine phase, results of query phase are gathered and combined from each threads.

$$Cos(\vec{x}, \vec{t}) = \frac{\sum_{i=0}^{n}(x_i t_i)}{\sqrt{\sum_{i=0}^{n} x_i^2}\sqrt{\sum_{i=0}^{n} t_i^2}} \tag{3.1}$$

Even though preprocessing phase is not specific process for KNN algorithm, it is common in algorithms used in Natural Language Processing. Therefore, better memory management strategy should be applied in this preprocessing phase. This preprocessing generates many intermediate data structures and copies of String elements are used in these data structure again and again. We can again either clone or get Arc of String.

Our KNN algorithms are implemented with batch processing. In query phase, we control batch size to examine how size of objects allocated simultaneously in memory has impact to algorithm's runtime performance. The detail implementation is explained in section3.2.

3.6 Complex Objects

To conduct experiments for above concepts, we use the 4 types of complex object: CustomerOwned, CustomerBorrowed, CustomerSlice, and CustomerRc. These object contains other type of object: OrderOwned, OrderBorrowed, OrderSlice, and OrderRc. The representation of these objects are shown in Figure 3·3 and Figure 3·4. Three Customer objects have 15 fields: 3 fields for i32, 3 fields for f64, 8 fields for String, and 1 field for Order object. All of fields of CustomerOwned and OrderOwned are owned by the object. On the other hand, fields of CustomerBorrowed, CustomerSlice, OrderBorrowed, and OrderSlice are borrowed. Reference of slice of values are used as the fields and owner of actual values are stored differently in source Vec. CustomerRc acquires Rc of values used for its fields from the source Vec.

```
struct CustomerOwned {
    key: i32,
    age: i32,
    num_purchase: i32,
    total_purchase: f64,
    duration_spent: f64,
    duration_since: f64,
    zip_code: String,
    address: String,
    country: String,
    state: String,
    first_name: String,
    last_name: String,
    province: String,
    comment: String,
    order: OrderOwned
}
```

(a)

```
struct CustomerBorrowed<'a> {
    key: &'a i32,
    age: &'a i32,
    num_purchase: &'a i32,
    total_purchase: &'a f64,
    duration_spent: &'a f64,
    duration_since: &'a f64,
    zip_code: &'a String,
    address: &'a String,
    country: &'a String,
    state: &'a String,
    first_name: &'a String,
    last_name: &'a String,
    province: &'a String,
    comment: &'a String,
    order: &'a OrderBorrowed<'a>
}
```

(b)

```
struct CustomerSlice<'a> {
    key: &'a i32,
    age: &'a i32,
    num_purchase: &'a i32,
    total_purchase: &'a f64,
    duration_spent: &'a f64,
    duration_since: &'a f64,
    zip_code: &'a str,
    address: &'a str,
    country: &'a str,
    state: &'a str,
    first_name: &'a str,
    last_name: &'a str,
    province: &'a str,
    comment: &'a str,
    order: &'a OrderSlice<'a>
}
```

(c)

```
struct CustomerRc {
    key: Rc<i32>,
    age: Rc<i32>,
    num_purchase: Rc<i32>,
    total_purchase: Rc<f64>,
    duration_spent: Rc<f64>,
    duration_since: Rc<f64>,
    zip_code: Rc<String>,
    address: Rc<String>,
    country: Rc<String>,
    state: Rc<String>,
    first_name: Rc<String>,
    last_name: Rc<String>,
    province: Rc<String>,
    comment: Rc<String>,
    order: Rc<OrderRc>
}
```

(c)

Figure 3·3: Representation of Customer objects Whose fields are different variable type: (a) CustomerOwned struct whose fields are all owned (b) CustomerBorrowed struct whose fields are borrowed with reference (c) CustomerSlice struct whose fields are borrowed with slice for sequence value, otherwise reference

```
struct OrderOwned {
    order_id: i32,
    num_items: i32,
    payment: f64,
    order_time: f64,
    title: String,
    comment: String
}
```

(a)

```
struct OrderBorrowed<'a> {
    order_id: &'a i32,
    num_items: &'a i32,
    payment: &'a f64,
    order_time: &'a f64,
    title: &'a String,
    comment: &'a String
}
```

(b)

```
struct OrderSlice<'a> {
    order_id: &'a i32,
    num_items: &'a i32,
    payment: &'a f64,
    order_time: &'a f64,
    title: &'a str,
    comment: &'a str
}
```

(c)

```
struct OrderRc {
    order_id: Rc<i32>,
    num_items: Rc<i32>,
    payment: Rc<f64>,
    order_time: Rc<f64>,
    title: Rc<String>,
    comment: Rc<String>
}
```

(c)

Figure 3·4: Representation of Order objects Whose fields are different variable type: (a) OrderOwned struct whose fields are all owned (b) OrderBorrowed struct whose fields are borrowed with reference (c) OrderSlice struct whose fields are borrowed with slice for sequence value, otherwise reference

Chapter 4

Evaluation Result

4.1 Experimental Set and Detail

4.1.1 Wikipedia Data Sets

Wikipedia page data sets are used to perform document classification with KNN. 10^5 pages are used for training data set, and 18724 pages are used for target.

4.1.2 Experimental Details

Our experiments are run on three types of VM instances on Google Cloud Platform: n1-standard-1 which has 1 vCPU, 3.75 GB RAM, and 10 GB Standard persistent disk, n1-standard-4 which has 4 vCPU, 15 GB RAM, and 10 GB Standard persistent disk, n1-standard-8 which has 8 vCPU, 30 GB RAM, and 10 GB Standard persistent disk, All of the experiments described are performed 5 times. In this thesis, we present the average of 5 separate runs for each experiment.

4.2 Experiment 1: Accessing Object with Different Variable Type

This experiment is conducted to understand two questions. One is how different variable types have impact to runtime performance. The other is how initialization of Vec size has impact to runtime performance. In this experiment, we focus on owner, reference, and slice as a variable of sequence values. Since these variables have different memory representation, there might be differences among time for access to actual values of each variables.

To evaluate this assumption, we use the three types of complex object: CustomerOwned, CustomerBorrowed and CustomerSlice. At first, we generate source Vecs for all fields, Vecs which contain all elements used for corresponding fields of objects. For example, all of i32 elements used for key field in 1 million Customer object are stored in Vec<i32> with 1 million i32 elements. Later, these i32 elements are moved to be owned or borrowed by the object's fields.

Next, 1.5, 1.8, 2.5, 2.8, 3.5, and 3.8 million of Customer objects are created and stored in Vec. When a Customer Vec is created, whether size of Vec is initialized is controlled. Finally, serialization of Customer object is performed as an operation forcing program to access all of fields in the object. This serialization is performed each Customer objects in the Vec. We measure total runtime to serialize all of Customer objects stored in Vec.

4.2.1 Result

The result is shown in Figure 4·1 and Figure 4·2. Figure 4·1 is comparison of the runtime performance among different Customer object types with Vec size initialization. Figure 4·2 is comparison in the same set of experiment except Vec size is not initialized. The blue, yellow, and green chats represent runtime of access to fields of CustomerOwned, CustomerBorrowed, and CustomerSlice objects respectively.

No matter Vec size initialized or not, differences of runtime for accessing objects are not remarkable among different object types. The percent increase in runtime of accessing to fields of from 1.5 to 3.8 million CustomerOwned objects is 121% if the Vec is initialized, 554% if the Vec is not initialized.

Considering the differences of percent increase of runtime whether the Vec is initialized or not, Figure4·3 the comparison of runtime access to fields of CustomerOwned object with and without Vec Initialization. For 3.8 million of objects, the difference of runtime to access to the objects is 38745 secounds; program without Vec initialization is 60% slower than one with the initialization.

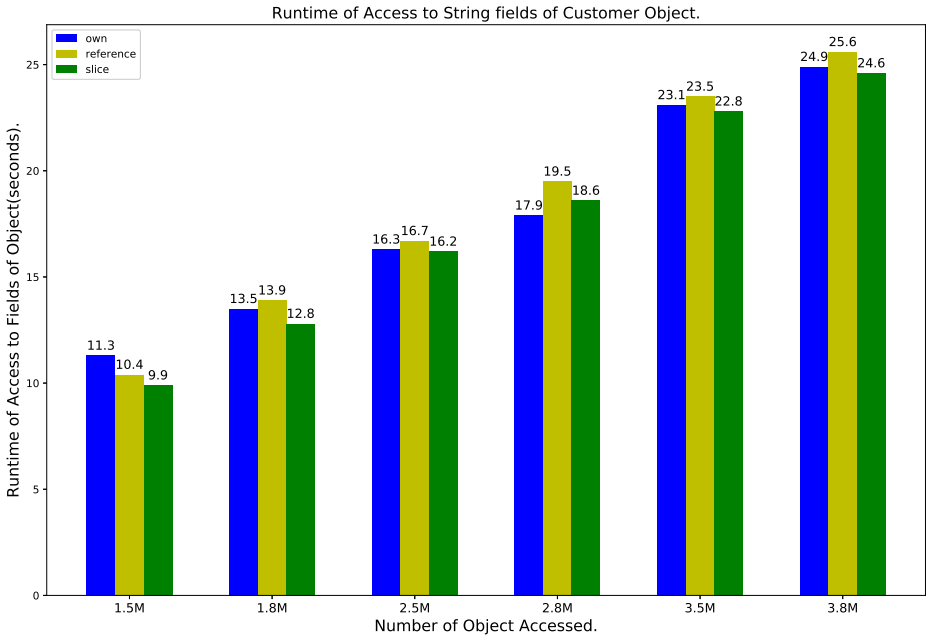

Figure 4·1: Runtime of Access to Different Pointer Types with Vec Size Initialization

4.2.2 Discussion

Difference of variable types does not have huge impact to runtime of accessing to actual value. Even thought owner, reference , and slice have different memory representations, the access time to its value is close to each other. As shown in Figure 3·1, the representations of owner and slice are almost identical except slice does not have capacity for values. Reference is pointer pointing to owner, so it has an additional step to access actual value. However, the result shows this additional step does not have huge impact for runtime to access memory

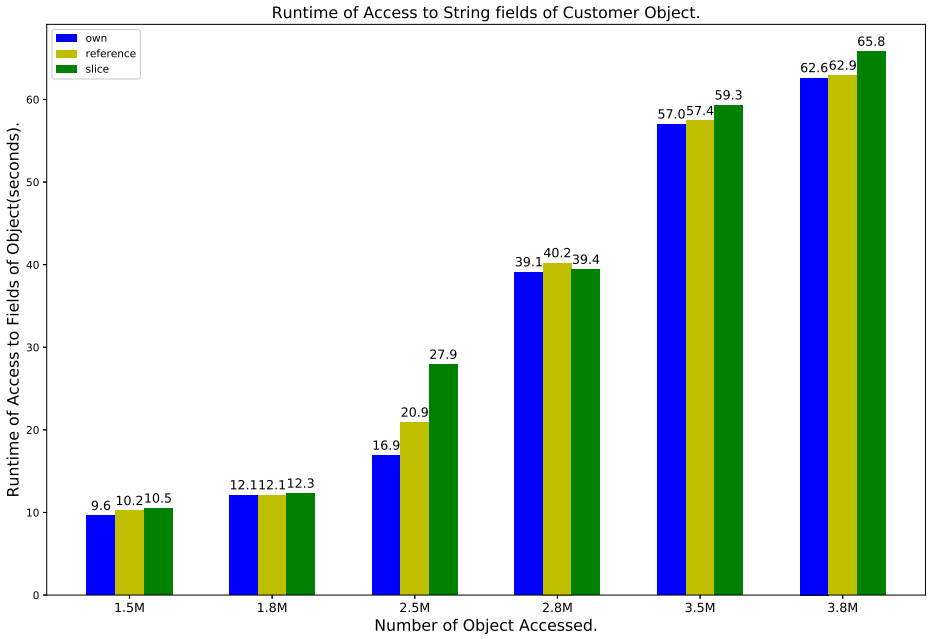

Figure 4·2: Runtime of Access to Different Pointer Types without Vec Size Initialization

region of the value.

Whether initializing Vec size results in disparity of runtime performance to access objects' fields. This is because when Vec uninitialized, the elements of Vec are allocated across different virtual memory pages.

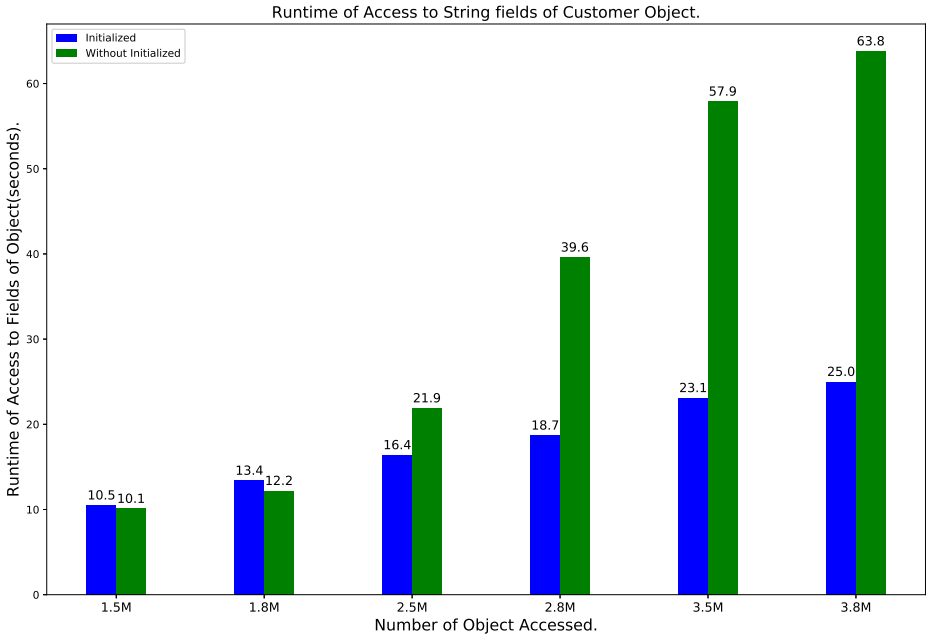

Figure 4·3: Runtime of Access to Fields of Complex Object with Initialization vs without Initialization

4.3 Experiment 2: Assessment of different reference methods in Rust

In this experiment, CustomerBorrowed and CustomerRc are used to see difference of dropping time among reference and Rc. In the CustomerRc and OrderRc struct, all fields take Rc (Rc<T>). Similarly to the experiment in the last section, sets of integer, float, and String vector are created and their elements are borrowed or reference counted to create CustomerBorrowed or CustomerRc objects. The dropping of objects deletes references or Rcs used for fields of the objects. However, it does not deallocate values to which they are pointing. Therefore, the evaluated runtime of dropping objects only consists of dropping time of reference or Rc, but deallocation time. We generated 1, 1.5, 2, 2.5, 3, 3.5 million of CustomerBorrowed and CustomerRc objects and performed drop one by one.

4.3.1 Result

Figure4·4 shows comparison of runtime dropping CustomerBorrowed and CustomerRc objects. The result shows significant difference of dropping time among the two objects; deletion of CustomerRc is much slower than CustomerBorrowed. The runtime of dropping CustomerBorrowed is about 60 times faster than dropping CustomerRc.

4.3.2 Discussion

In this experiment, an assessment is conducted to verify whether there is difference between behavior of reference and Rc. The reason why dropping Rc is much slower than dropping reference is that Rc requires runtime overhead to check some states of the variable, but when to drop reference is already determined at compile time. When dropping Rc, Rc has to check the number of variable pointing to the actual content and decide deallocate the memory or not. However, memory management and lifetime strategy of reference is already determined at compile time. This determination of memory management strategy at compile increases runtime performance of dropping complex object constructed with reference type variable. This may say that we should use reference whenever high performance computation is critical.

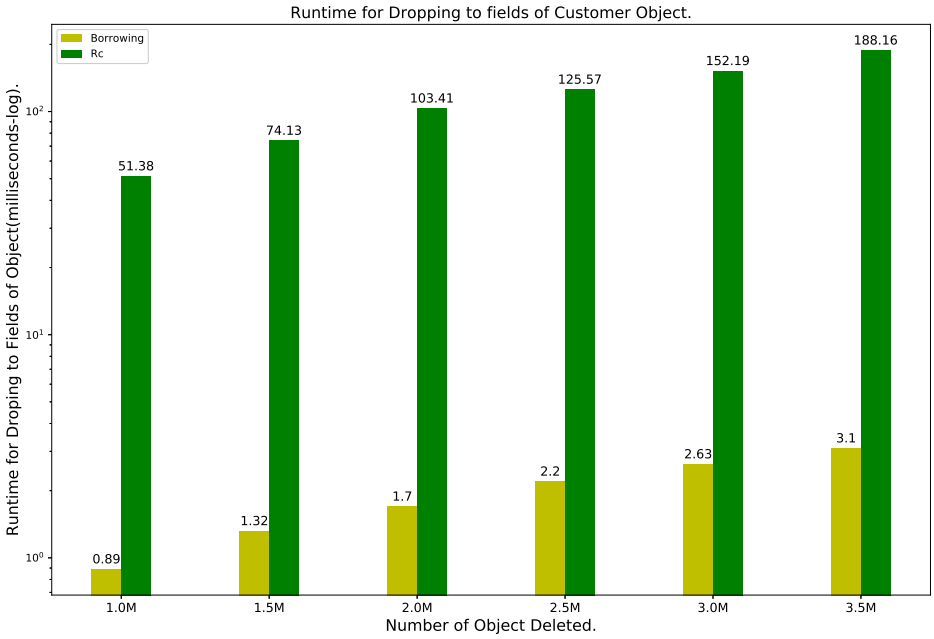

Figure 4·4: Runtime for dropping Customer Object

However, dealing with reference is sometimes cumbersome. Tracking lifetime of reference can be done easily in simple situation. But, if we have complex objects constructed with fields of reference, the lifetime tracking become extremely difficult. For example, constructing nested objects with reference fields require a developer to plan memory management with many lifetime symbols. Using reference counting eliminates developer's responsibility to specify lifetime of variables. This may ease and speed up development process, and increase understandability of codes.

Even though we have stack allocated values, such as i32 and f64, in Rc in our experiment, one should avoid wrapping stack allocated values in Rc. Wrapping value in Rc allocates heap memory so that allocating Rc<i32> or Rc<f64> unnecessarily use space of heap. Additionally, stack allocated values are usually easy to be copied. Therefore, developer does not have to even use reference; one can just copy the value. Copying value in Rust is to copy the original value and to assign the copy to new owner variable.

4.4 Experiment 2: Merge-sort

As explained in last experiment, deletion of Rc has significant overhead than normal reference. By learning this result, our assumption here is that deletion of Arc has also overhead when we compare to normal reference. In many situation where developer writes a multithreading code, Arc is used and the deletion happens multiple times. To assess runtime overhead of algorithm with Arc, We implement merge-sort algorithm in two different ways. One is sharing source vector with Arc. The other is passing reference of source vector to child thread.

Our merge-sort algorithms are implemented with recursion. For each call of recursive function, Arc or slice of the source vector needs to be passed and deleted when the function returns value. These merge-sort algorithms trigger large number of Arc or reference deletion proportional to the number of call recursive function. Merge-sort algorithm can be separated in three phases: splitting phase, copying phase, and merging phase.

The splitting phase is merely acquiring index of range. At this phase, multiple threads are generated and Arc or reference of source vector are passed by calling recursive function. Copying phase occurs in the base case of the recursive call. The element in the source vector in a certain index is deep-copied into newly allocated vector. At merge phase, merge function receives two sorted independent vector and merge them into single new vector.

We use scope method from crossbeam crate to perform multithread programming. Scoped thread can have reference to value from its parent thread by ensuring children threads are joined before their parent thread returns value. By using scoped thread, we can implement two merge-sort algorithms in identical way except whether the function receives Arc or reference of source vector. The representations of source vector for each algorithm are shown in Figure 4·5. The elements of source vector are CustomerOwned objects. We generates source vectors in size of 1, 2, 3 and 4 million. Finally, merge-sort is performed based on value of key field. The figure shows the result for runtime performance of merge-sort algorithms on difference size of vectors. This experiment is run on n1-standard-4 and generates 4 threads to perform multithreading.

```
// Source vector for algorithm with Arc.
arr: Arc<VecDeque<T>>

// Source vector for algorithm with reference (slice).
arr: &[T]
```

Figure 4·5: Representation of Source Vector

4.4.1 Result

Figure 4·5 shows the runtime performance of our merge-sort algorithm with specified Vec sizes. The blue and yellow bar charts represent the runtime performance of merge-sort algorithm using reference and Arc respectively. The result says that algorithms with Arc is about 24% slower than algorithms with reference.

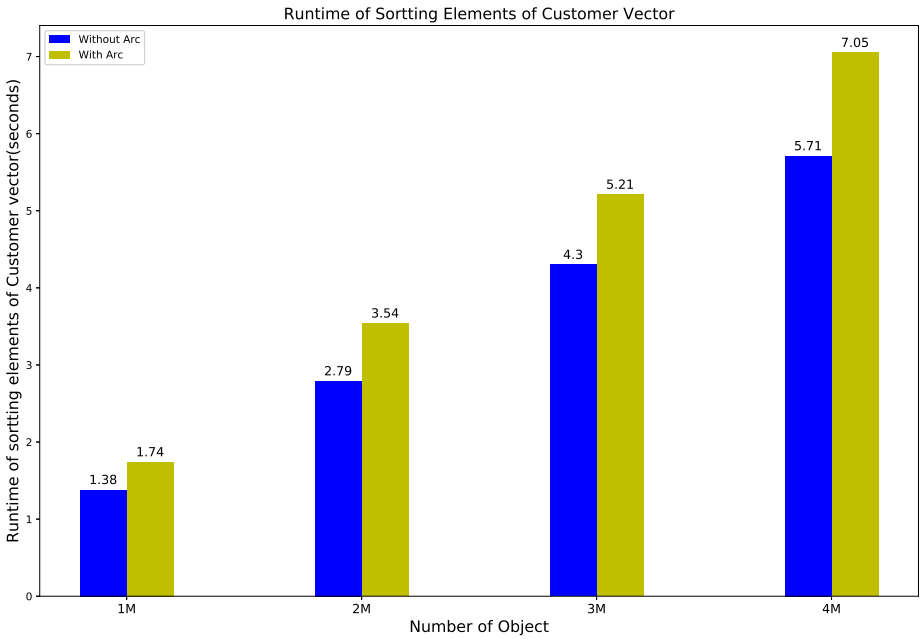

Figure 4·6: Runtime of Sorting Elements of Customer Vector

4.4.2 Discussion

The reason why merge-sort algorithm with Arc is much slower than one with reference is the same for the reason why dropping Rc shows overhead compared to dropping reference. Arc has to check the number of variable pointing to the actual content and decide deallocate the memory or not.

In addition, Arc uses atomic operations for tis reference counting. Atomic operations bring thread-safety, but they are more expensive than ordinary memory accesses. Therefore, when sharing reference counting between threads is not required, using Rc is the recommended way [4].

In the situation like our experiment, normal reference can be used instead of Arc to share data between threads. This solution results in better runtime performance in our experiment. Therefore, one should use reference to share data among different threads whenever it is possible.

4.5 Experiment 4: Tree-aggregation

In our experiment, tree-aggregation algorithms are examined in multi-threading. This experiment is to evaluate the impact of having Arc (Atomic Reference Counting) as elements of vector. In Big Data mining tool, such as Spark, it generates intermediate objects from original source vector. In tree-aggregation, aggregated HashMap like data structure is created in each step or node. Acquisition of elements in source vector is required to perform this aggregation. There are several ways.

One way is deep-copy elements of vector. This solution allocated newly created objects by deep-copy. Aggregation is performed on copied objects, stores them in the data structure and sends it to next node. Deep-copy generates duplicates of objects in vector and aggregated data structure. This can lead to memory intensive moment when we need memory space for the duplicated objects in addition.

The other way is to get reference to the elements. Since an original source vector is deallocated after a local aggregation, Simple reference to elements does not live long enough and allow the aggregation result to be sent to next node. Instead of simple borrowing, we need owner in the aggregation result. Reference Counting (Rc) in Rust is a way to have multiple owners to a value. Since our experiment is implemented in multithreading, Atomic Reference Counting (Arc) is used instead of Rc. With Arc, multiple ownership pointer can be possessed by different variables across multiple threads. Therefore a value is not deallocated until all of owners to it are dropped. This does not require extra memory allocation, because only acquisition of new ownership to value is needed. However, as explained in the last section deletion of Arc type checks whether the value is still owned by other variables. This checking may be an overhead in algorithms where generate a lot of intermediate data structures, because deletion of the data structures occurs in frequent.

Two algorithms are implemented using the above two different methods and evaluated their runtime performance. We perform aggregation to CustomerOwned based on last_name field. Before tree-aggregation algorithms are run, partitions of Vec<CustomerOwned> or of Vec<Arc<CustomerOwned>> are created ,serialized, and stored in disk. A tree-aggregate

algorithm has main three phases: loading, aggregating, and combining phase. At loading phase the algorithm generate threads. In each node, it loads serialized CustomerOwned partition from disk and deserialize them. At aggregating phase, aggregation is performed on each partition by last_name field. Once a node finishes aggregation, it sends result to parent node. After parent nodes receive aggregation results from all of its children nodes, it joins all aggregation results including its and sends next parent. This joining aggregation results is considered as combining phase.

Two kinds of algorithm are implemented. One algorithm performs aggregation by deep-copying elements from source vector loaded from disk. In the other algorithm, each element of source vector is wrapped in Arc, and its reference is acquired while aggregation. The difference of the both algorithm codes are represented in Figure 4·7 and in Figure 4·8. If we glance the codes, the notable difference is only when we acquire an element from CustomerOwned Vec to construct an aggregated data structure. Therefore, there is few difference between two kinds of tree-aggregate algorithm in terms of code appearance.

Numbers of CustomerOwned objects aggregated in our experiment are 1, 2, 3, 4 million. This experiment is run on n1-standard-4.

```
fn aggregate_local(arr :&[Arc<CustomerOwned>])
{
    let mut agg = HashMap::new();
    let n = arr.len();
    for i in 0..n {
        let customer = Arc::clone(&arr[i]);
        let last_name = customer.last_name.clone();
        let vector = agg.entry(last_name).or_insert_with(Vec::new);
        vector.push(customer);
    }
    return agg;
}
```

Figure 4·7: Aggregation function with Arc

```
fn aggregate_local_copy(arr :&[CustomerOwned])
{
    let mut agg = HashMap::new();
    let n = arr.len();
    for i in 0..n {
        let customer = arr[i].clone();
        let last_name = customer.last_name.clone();
        let vector = agg.entry(last_name).or_insert_with(Vec::new);
        vector.push(customer);
    }
    return agg;
}
```

Figure 4·8: Aggregation function with deep-copy

4.5.1 Result

Figure 4·9 shows runtime performance of two tree-aggregate algorithms. The runtime of algorithm with deep-copy is about 33% slower than algorithm with Arc for every vector size.

4.5.2 Discussion

As we explained, Arc has overhead to be deleted because it has to check if the value is still referred. The atomic operations are more expensive than ordinal memory access. Even though the use of Arc slows down runtime performance, deep copy of complex objects has more impact in deterioration of runtime performance.

At the aggregating phase, each object is deep-copied or acquired with Arc once during the runtime in order to construct aggregated data structures. If total number of object is 1 million, the all of 1 million objets are deep-copied or cloned with Arc once during execution. Deep-copy allocates new memory for copied object. On the other hand, clone with Arc is merely acquisition of additional owner. Therefore, deep-copy is more expensive in terms of runtime and also use of memory than clone with Arc. Deep-copy processes all the original objects to generate newly deep-copied objects. This process has overhead and the existence of both original and copied objects doubles its memory usage.

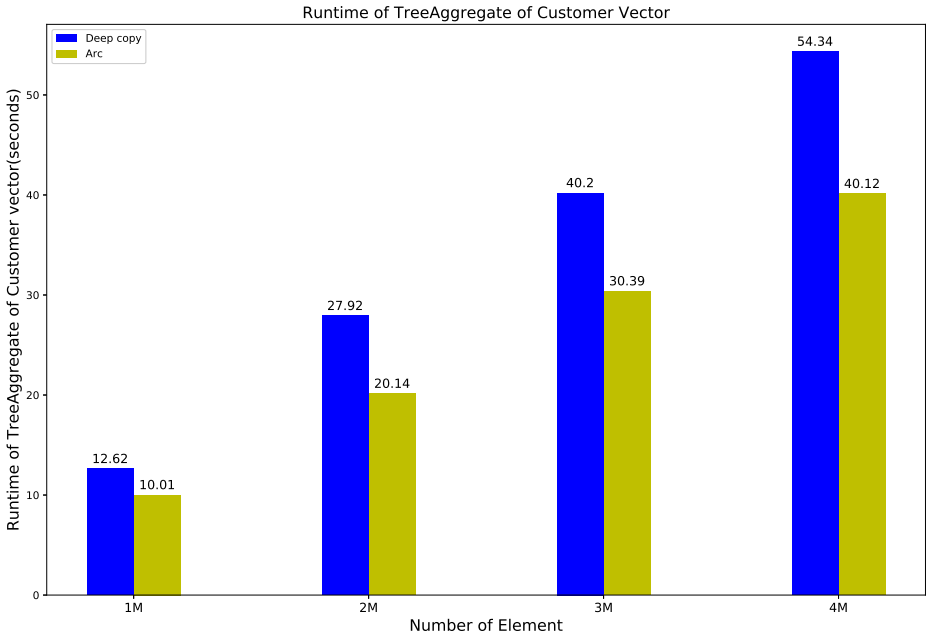

Figure 4·9: Runtime of Tree-aggregate algorithm

After construction of aggregated data structure, the dropping variables of original object occurs at the end of aggregating phase. In the algorithm with deep-copy, these variables are owners so that drop of variables triggers deallocation of actual values. In the algorithm with Arc, the variables are Arc. The aggregated data structure contains Arc pointing to the same values pointed by original Arc. Since the aggregated data structure continue to live after aggregating phase, drop of original Arcs does not triggers deallocation of values. Therefore, drop of original variables can shows overhead of memory deallocation in algorithm with deep-copy, and overhead of checking reference count of Arc in algorithm with Arc.

In addition, memory access from Arc is slower than ordinal variable due to the atomic operations. This may be potential overhead of the algorithm with Arc.

Considered these theoretical analysis and result of our experiment, using Arc improves runtime performance and memory usage compared to algorithm using deep-copy in tree-aggregate algorithm.

Chapter 5

Conclusions

5.1 Summary of the thesis

Time to get philosophical and wordy.

IMPORTANT: In the references at the end of thesis, all journal names must be spelled out in full, except for standard abbreviations like IEEE, ACM, SPIE, INFOCOM, ...

Appendix A

Linear Algebra Computation

A.1 Create Java interface of CBLAS with JNI

1. Download BLAS and build using make file. In the figure2.1, the built file is libblas.a and the header file is blas.h.

2. Download CBLAS and build using make file (I am not sure whether we should build archive file or shared library). In figure, the build file would be libcblas.a or libcblas.dylib and header file is cblas.h.

3. Create java file which will be the Java interface of CBLAS.

4. Compile java file with -h header flag to create class file (CBLASJ.class) and header file (CBLASJ.h).

```
$ javac -h . CBLASJ.java
```

5. Create C file (CBLASJ.c) which will bind Java interface and CBLAS library. And compile it with JNI to create object file (CBLASJ.o).

```
$ gcc -c-fPIC -I${JAVA_HOME}/include -I${JAVA_HOME}/include/darwin CB
```

6. Compile shared library linking library (libcblas.a or libcblas.dylib) object file (CBLASJ.o).

```
$ gcc -o libcblas.dylib (or libcblas.a) CBLASJ.o
```

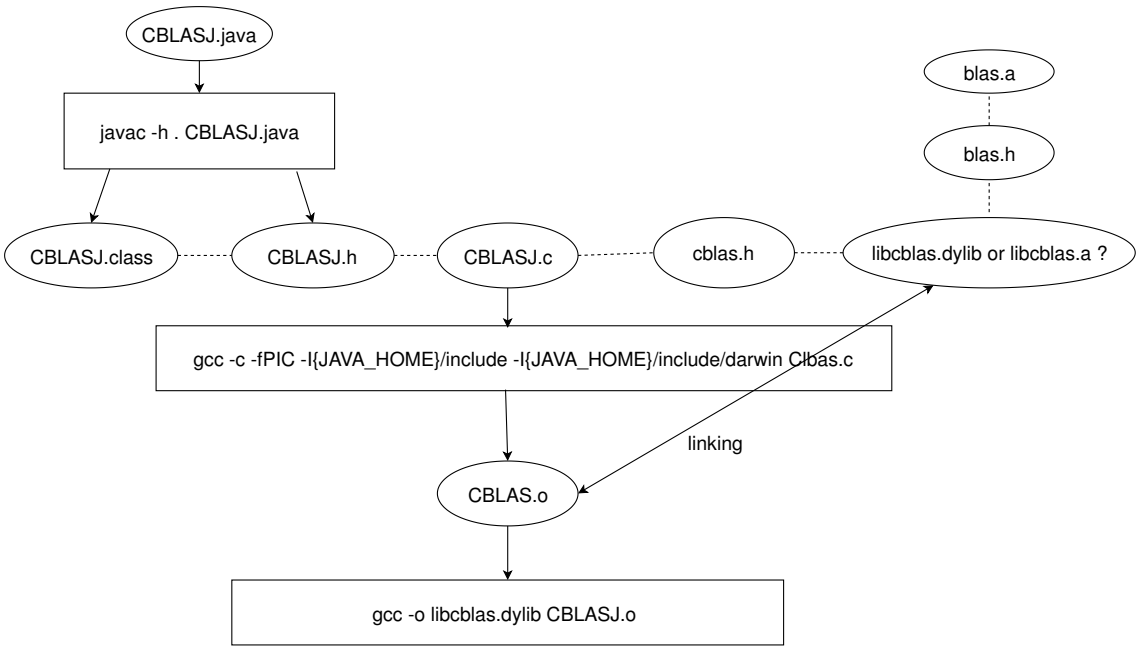

Figure A·1: Integration of Native Methods

References

[1] Apache flink, 2020.

[2] Apache hadoop, 2020.

[3] Apache spark, 2020.

[4] Rust arc documentation, 2020.

[5] Michael Armbrust, Reynold S. Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, and Matei Zaharia. Spark SQL: relational data processing in spark. In Timos K. Sellis, Susan B. Davidson, and Zachary G. Ives, editors, *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 1383–1394. ACM, 2015.

[6] Chris Lattner and Vikram S. Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *2nd IEEE / ACM International Symposium on Code Generation and Optimization (CGO 2004), 20-24 March 2004, San Jose, CA, USA*, pages 75–88. IEEE Computer Society, 2004.

[7] Xiangrui Meng, Joseph K. Bradley, Burak Yavuz, Evan R. Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, D. B. Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. Mllib: Machine learning in apache spark. *J. Mach. Learn. Res.*, 17:34:1–34:7, 2016.

[8] Khanh Nguyen, Kai Wang, Yingyi Bu, Lu Fang, Jianfei Hu, and Guoqing (Harry) Xu. FACADE: A compiler and runtime for (almost) object-bounded big data applications. In Özcan Özturk, Kemal Ebcioglu, and Sandhya Dwarkadas, editors, *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS '15, Istanbul, Turkey, March 14-18, 2015*, pages 675–690. ACM, 2015.

[9] Juwei Shi, Yunjie Qiu, Umar Farooq Minhas, Limei Jiao, Chen Wang, Berthold Reinwald, and Fatma Özcan. Clash of the titans: Mapreduce vs. spark for large scale data analytics. *PVLDB*, 8(13):2110–2121, 2015.

[10] Lijie Xu, Tian Guo, Wensheng Dou, Wei Wang, and Jun Wei. An experimental evaluation of garbage collectors on big data applications. *PVLDB*, 12(5):570–583, 2019.

[11] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Steven D. Gribble and

Dina Katabi, editors, *Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012, San Jose, CA, USA, April 25-27, 2012*, pages 15–28. USENIX Association, 2012.

[12] Jia Zou, R. Matthew Barnett, Tania Lorido-Botran, Shangyu Luo, Carlos Monroy, Sourav Sikdar, Kia Teymourian, Binhang Yuan, and Chris Jermaine. Plinycompute: A platform for high-performance, distributed, data-intensive tool development. In Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein, editors, *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1189–1204. ACM, 2018.

CURRICULUM VITAE

Joe Graduate

Basically, this needs to be worked out by each individual, however the same format, margins, typeface, and type size must be used as in the rest of the dissertation.