MOLLY

by

Richard Drake

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

Masters of Science

in

Faculty of Science

Computer Science

University of Ontario Institute of Technology

Supervisor: Dr. Ken Q. Pu

September 2012

Abstract

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Preface

Background and Motivation

The introduction of keyword search has revolutionized how we find information. Over the years, numerous techniques have been developed which make searching through large amounts of information for one or more keywords extremely fast. With the advent of faster computer hardware, we are able to not only search through small attributes of a document (eg. Title, synopsis, etc.) but rather the entirety of the document itself.

An example of an early system which utilized keyword search would be a library catalogue. Such a system would allow a user to search, by keyword, for the title of a book, manuscript, etc. The results would show item titles matching the keyword(s), as well as other information such as whether or not the item is in circulation, as well as where it is located within the library. This information would come from a relational database.

With the rise of the World Wide Web, much information was placed online. This information would be easy to access if one knew how to locate it. Unfortunately, over time, so much information existed on the World Wide Web that it became difficult to keep track of it all. There was a need to index all of this information and make it accessible. This need was filled by a Web search engine.

Initial Web search engines comprised of simple scripts that gathered listings of files on FTP servers; they were essentially link farms. A few short years later, the first full-text (keyword) search engine, WebCrawler, was released.

While full-text search engines provided an excellent means for locating information, as the

Web grew larger, the volume of noise also grew larger. In addition, every Web page could be structured in a different way; the Web was largely a collection of unstructured documents. That is, there were few obvious links between them.

Search engines such as Google attempted to solve this problem by introducing new algorithms, such as PageRank, to rank Web pages on both the relevance of their content as well as their reputation. The idea was if a page is linked to often, it is considered to be more authoritative on a subject than a page with fewer links. This allowed the relevance of a page to be computed based on not only its contents, but its artificial importance.

Molly attempts to avoid some of the issues plaguing search engines. It deals primarily with structured, filtered data. This allows us to provide results with less noise. In addition, the fact that it deals with structured data means links between documents are explicitly stated. Rather than inferring a link between documents based on hyperlinks, we know when two documents are linked together.

This thesis provides an overview of the Molly system.

Outline

Changed.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Literature Review

Chapter 2

Theoretical

The problem of efficiently searching through a graph is the subject of countless articles and journal papers. Numerous algorithms have been proposed to perform graph search in an efficient manner. Many of these algorithms build upon their predecessors, making assumptions and changing aspects to better suit a particular problem area.

The applications of efficient graph search algorithms are endless. Algorithms such as Dijkstra's are used heavily in path finding, for example in a portable GPS unit. A* Search is used in the area of computer vision to approximate the best path to take. Companies such as Amazon make use of graph search algorithms in order to find related products for consumers to purchase.

This thesis concentrates on building a system which can be utilized in order to discover related information. The system that was built is capable of finding not only related information from neighbouring nodes, but also the best path between two arbitrary nodes.

Section 2.1 provides an overview of how the data is represented in the system. It also defines notation to represent this data. Section 2.2 outlines how the relational database is related to the data in the system. Section 2.3 provides an example data corpus which is utilized throughout this thesis. It details the "mycampus" dataset. Finally, Section 2.4 provides justification for the algorithm chosen to perform the graph search. It discusses multiple different graph search algorithms, as well as why the one used in this thesis was chosen.

2.1 Data Representation & Notation

The data from the database is represented in various data structures. There are separate representations for each type of data: values, entities, and entity groups.

Value

Definition 1. A **Value** represents a single piece of information. To avoid repetition, each value is unique. That is, $\exists!\, v \in V$, where v is a value in the set V of all values.

Entity

Definition 2. An **Entity** is a collection of attributes, a_n, each mapped to a single value, v_n. An entity also includes additional information such as a unique identifier.

$$\begin{array}{ll} \text{id} & T_n | v_{id} \\ a_1 & v_1 \\ a_2 & v_2 \\ \vdots & \vdots \\ a_n & v_n \end{array}$$

Figure 2.1: The structure of an entity

Entities are analogous to rows in a database table. Thus, the unique identifier is generated based on the table name, T_n, as well as unique key in the table, v_{id}. The unique key identifies the row, and the table name identifies the table. Together they uniquely identify the entity within the entire database.

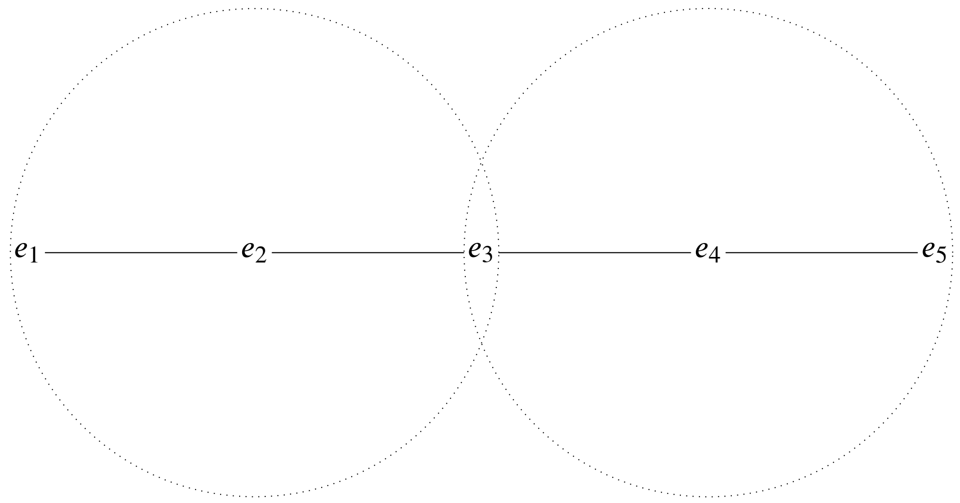
$\exists!\, e_{id} \in E$, where E is the set of all entities.

Entity Group

Definition 3. An **Entity Group** joins together two or more entities. These entity groups can also have attributes, a_n, and values, v_n, associated with them much like entities.

$$\begin{array}{ll} e_L & [e_1, e_2, \dots, e_n] \\ a_1 & v_1 \\ a_2 & v_2 \\ \vdots & \vdots \\ a_n & v_n \end{array}$$

Figure 2.2: The structure of an entity group



2.2 Relational Database Abstraction

2.3 Data Corpus

For illustrative purposes, the mycampus dataset will be used. This data comes from UOIT's course registration system.

There are several different entities which comprise the mycampus dataset:

- Courses
- Instructors
- Schedules
- Sections

- Teaches

The **Courses** entity represents a course. A course has an individual code, along with a title and a description (See Table 2.1). The code uniquely identifies the course.

Column	Type	Description
code	VARCHAR	Unique course code
title	VARCHAR	Title of the course
description	TEXT	A brief description

Table 2.1: Courses entity schema

The **Instructors** entity represents an individual instructor. Each instructor has a unique identifier, as well as a name (See Table 2.2). An instructor can be a Professor, Lecturer, Sessional Instructor, or Teaching Assistant.

Column	Type	Description
id	INT	Unique identifier
name	VARCHAR	The instructor's name

Table 2.2: Instructors entity schema

The **Sections** entity represents a section of a course. Courses may have many sections. For example, one section could be a lecture, while another is a lab. Some courses may have over a dozen sections, depending on the associated term.

Each section contains a unique identifier, capacity information (students enrolled, spots open, etc.), when registration is open for the section, how many credits it is worth, what level (eg. undergraduate or graduate), and what year the section is offered in (See Table 2.3).

The **Schedules** entity represents a scheduled meeting of a section. A section may have many schedules. For example, a lecture section may meet twice a week.

Each schedule has a unique identifier, a date range in which the schedule is active, the time of the schedule, the type, location, and the day which the class takes place (See Table 2.4).

A **Teaches** entity is used to link together an instructor and a schedule. Each link has a unique

Column	Type	Description
id	INT	Unique identifier
actual	INT	Number of people enrolled in the course
campus	VARCHAR	String uniquely identifying the campus
capacity	INT	Maximum number of people that may be enrolled in the section
credits	FLOAT	Number of credits awarded upon successful completion
levels	VARCHAR	The level of the course (eg. undergraduate, graduate, etc.)
registration_start	DATE	Date registration for the section opens
registration_end	DATE	Date registration for the section ends
semester	VARCHAR	String that uniquely identifies the semester
sec_code	INT	Unique section code (called a CRN)
sec_number	INT	Sequential number identifying the number of the section
year	INT	The year the section is offered in

Table 2.3: Sections entity schema

Column	Type	Description
id	INT	Unique identifier
date_start	DATE	First day of class
date_end	DATE	Last day of class
day	VARCHAR	Single character representing the day of the week
schedtype	VARCHAR	Lecture, tutorial, lab, etc.
hour_start	INT	Hour the class starts at
hour_end	INT	Hour the class ends at
min_start	INT	Minute the class starts at
min_end	INT	Minute the class ends at
classtype	VARCHAR	The type of class
location	VARCHAR	Unique location name where class is held

Table 2.4: Schedules entity schema

identifier along with the instructor's position (eg. Teaching Assistant, Lecturer, etc.) (See Table 2.5).

Column	Type	Description
id	INT	Unique identifier
position	VARCHAR	Position of the Instructor with regard to a Schedule

Table 2.5: Teaches entity schema

2.4 Graph Search Algorithms

Careful consideration was given to which graph search algorithm was to be used. Among the choices were:

- Breadth-First
- Bellman-Ford
- Dijkstra
- A*

The first choice, Breadth-First Search, is among the simplest of the algorithms. It has the advantage of being simple to implement. It can only handle fixed costs for travelling between nodes, which can be a disadvantage.

Bellman-Ford is similar to BFS. It has the ability to deal with variable cost. This comes at the cost of increased difficulty in implementation. When the cost between nodes is fixed, Bellman-Ford essentially becomes BFS.

A greedy version of BFS is Dijkstra's Algorithm. It utilizes a priority queue rather than a regular queue, allowing it to be faster. As it is a greedy algorithm, Dijkstra's Algorithm may not return the optimal result.

An extension to Dijkstra's is A* search. Graph search spaces can be rather large. A* attempts to prune the search space based on a heuristic. A* is a natural choice to perform graph search in certain areas such as computer vision where an obvious heuristic exists. It has a disadvantage of consuming large amounts of memory (though IDA* attempts to limit memory consumption).

There are many more graph search algorithms. The above were primarily considered as many of the other algorithms are simple extensions with different data structures.

For this application, there is no obvious heuristic. This eliminates A*. There is also no obvious cost function. This eliminiates Dijkstra's Algorithm. Bellman-Ford is very similar to BFS with the added ability to deal with variable cost. As the cost function is not obvious and thus constant, Bellman-Ford essentially reverts back to BFS.

For these reasons, BFS was chosen as the graph search algorithm. While the other candidates and others provide numerous advantages over BFS in many situations, this is not one of them.

Chapter 3

Implementation

3.1 Functional vs. Procedurial Programming of Algorithms

Key differences between FP and procedurial (for algorithms)

3.1.1 Functional Data Structures

Clojure immutable data structures, versioning (persistent data structures), STM vs. locking

3.2 Tunable Parameters

Tunable parameters

Chapter 4

System Implementation

4.1 Technology Selection

4.1.1 Programming Language

Numerous programming languages were considered for the implementation. Among them were: Ruby, Python, and Clojure. Each of the languages presented both pros and cons. Ruby and Python are rather similar object oriented languages. Clojure is a functional language.

Ruby features a clean syntax and is very object oriented. Its object model was inspired by that of Smalltalk. It has a very active web development community and numerous web frameworks (eg. Rails, Sinatra, Padrino, etc.). In contrast to Python, it features a lean core, instead choosing to depend on third party libraries.

Python also features clean syntax and is object oriented. It embraces a "batteries included" philosophy to its standard library, featuring libraries for everything from serial communications to importing/exporting CSV files. It too has a strong web development community and numerous web frameworks (eg. Django, Flask, web2py, etc.).

Python is used extensively by scientists. As such, it has numerous plotting, computation, and simulation libraries. Examples include Matplotlib, SciPy, NumPy, and NetworkX.

Both of the above languages embrace functional programming elements. Python has filter, reduce, and map. Ruby's collections are also capable of performing filter (select), reduce (inject), and map. They both lack an important aspect of functional programming: immutable objects.

Clojure is a purely functional language built on top of the JVM. Like Python and Ruby, it is dynamically typed. Unlike Python and Ruby, it is a Lisp dialect and as such features macros and code-as-data.

As it runs on the JVM, Clojure allows us to utilize existing Java languages. Using Java libraries through Python or Ruby requires using their respective native code interfaces to Java's JNI. This process is typically automated through a tool such as SWIG.

Clojure was chosen as the implementation language for a number of reasons. First of all, as it is built on top of the JVM, it can utilize JDBC for database access. It can also make use of several other useful libraries. Secondly, its dynamic Lisp nature is a natural way of processing data. Clojure allows us to model our data structures directly after the data. In a language such as Python, one would need to make use of classes.

4.1.2 Relational Database

4.1.3 Full-Text Search Database

4.1.4 Web Stack

4.2 Implementation Issues

Chapter 5

Performance & Evaluation

5.1 Environment

All benchmarks were run on the same machine using the same software versions. The machine has the following specs.

Item	Description
Model	Dell PowerEdge 2950
Processor	2 x Quad-Core Intel® Xeon™ 5300 3.0GHz
Memory	8 GB DDR2 ECC RAM 667 MHz

Along with the following software versions.

Item	Description
Operating System	GNU/Linux Ubuntu 10.04.4 LTS
Kernel	2.6.32-28-generic
Java™ SE Runtime Environment	1.6.0_26-b03
Java HotSpot™ 64-Bit Server VM	20.1b02

5.2 Methodology

Evaluating the performance of code is a difficult problem at best. It is difficult to determine the impact on performance of various uncontrollable factors. Virtual Machines add another layer of abstraction which introduces even more factors.

As this project runs a top of the Java Virtual Machine (JVM), there are many factors to consider with regard to performance. In addition to uncontrollable events such as garbage collection, the JVM's behaviour can differ based on the operating system (OS) it is running on.

5.2.1 Pitfalls of the Java Virtual Machine

Timing Execution

There are two methods provided by the JVM for getting a precise time from the OS; `System.currentTimeMillis()` and `System.nanoTime()`. The former returns the "wall" time. It is possible for time to leap forward or backward using this method. If daylight savings time occurs between calls to `System.currentTimeMillis()`, it can result in a negative time.

The latter uses a variety of methods in order to obtain the precise time. The method it uses depends on both the OS and the hardware itself. Under Windows, `System.nanoTime()` makes use of the `QueryPerformanceCounter(QPC)` call. This call may make use of the "programmable-interval-timer (PIT), or the ACPI power management timer (PMT), or the CPU-level timestamp-counter (TSC)." [3] Accessing the PIT and PMT is a slow operation. Accessing the TSC is a very fast operation, but may result in varying numbers [2].

Linux attempts to use the TSC when possible. If it finds the values to be unreliable (eg. different cores vary too much), it makes use of the High Precision Event Timer (HPET) [2].

Platform	Resolution (ms)
Windows 95/98	55
Windows NT, 2000, XP (single processor)	10
Windows XP (multi processor)	15.625
Linux 2.4 Kernel	10
Linux 2.6 Kernel	1

Table 5.1: Reported resolutions of `currentTimeMillis()` by platform [1].

The JVM's timing functions are one area where the JVM's behaviour differs based on the OS. Different operating systems provide varying degrees of time resolution. The speed at which the OS-level calls to these timing functions return can even differ.

Garbage Collection

Just-in-Time Compilation

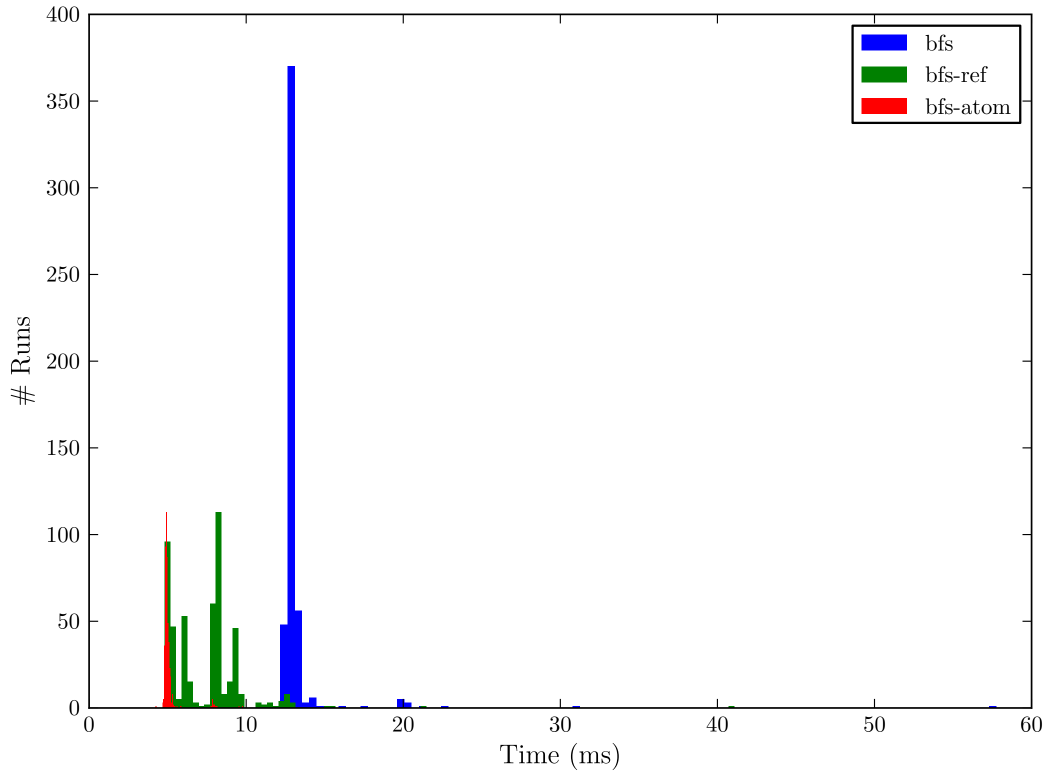


Figure 5.1: 1 Hop

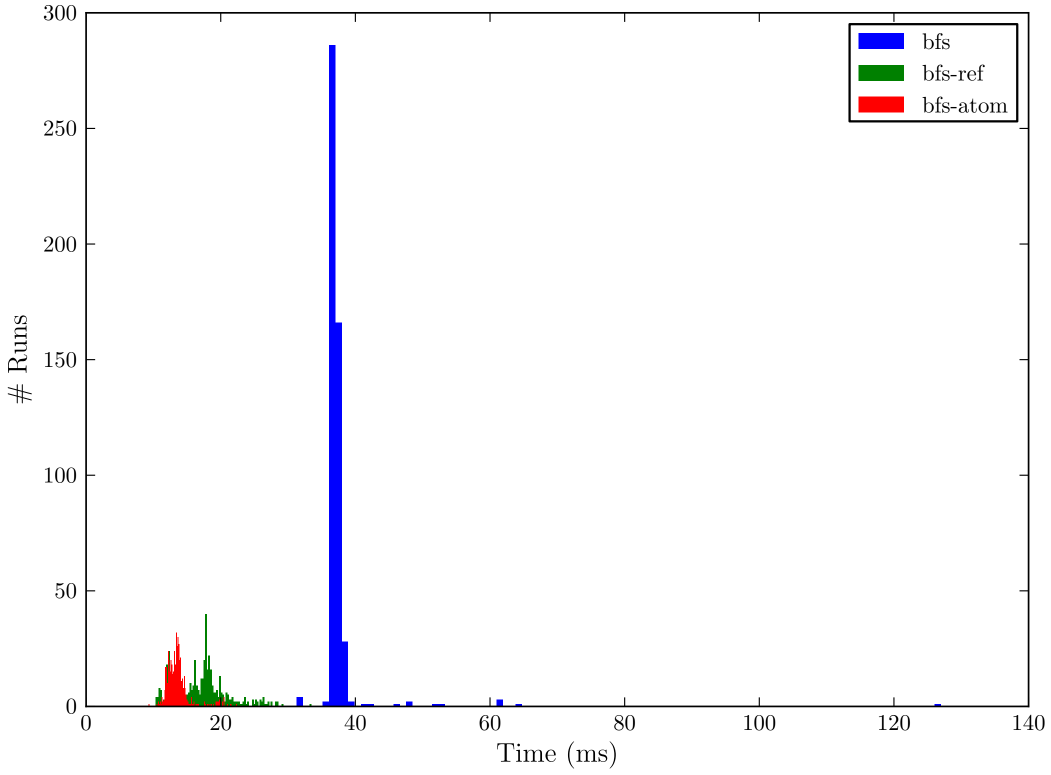


Figure 5.2: 2 Hops

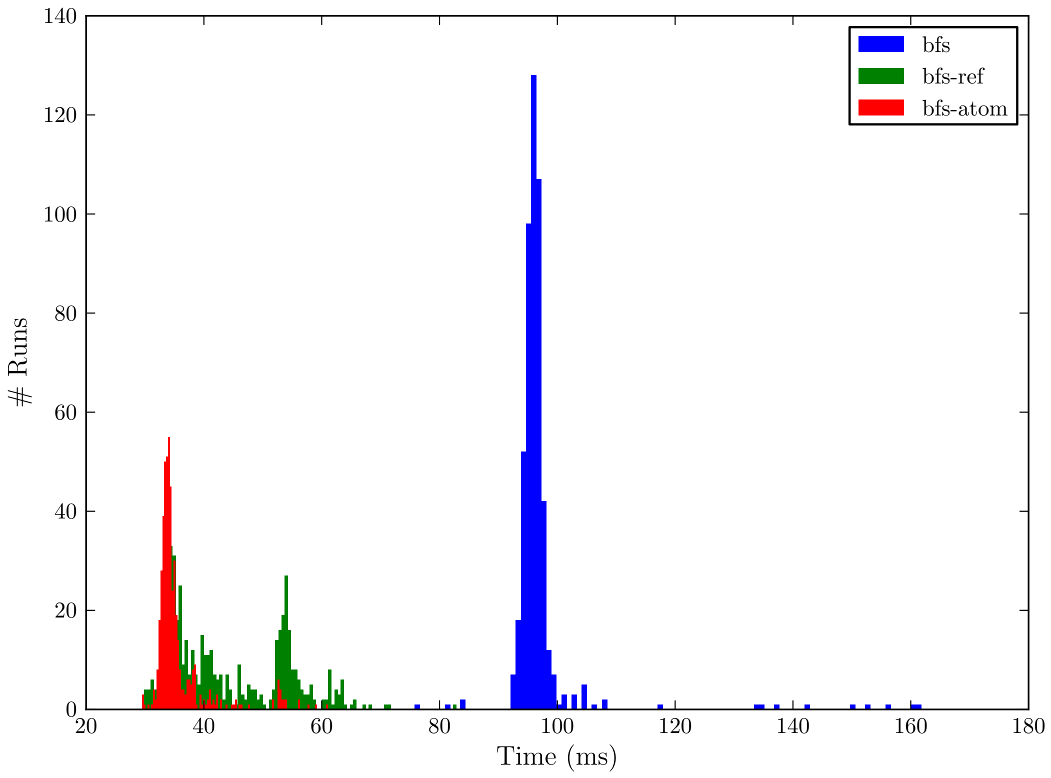


Figure 5.3: 3 Hops

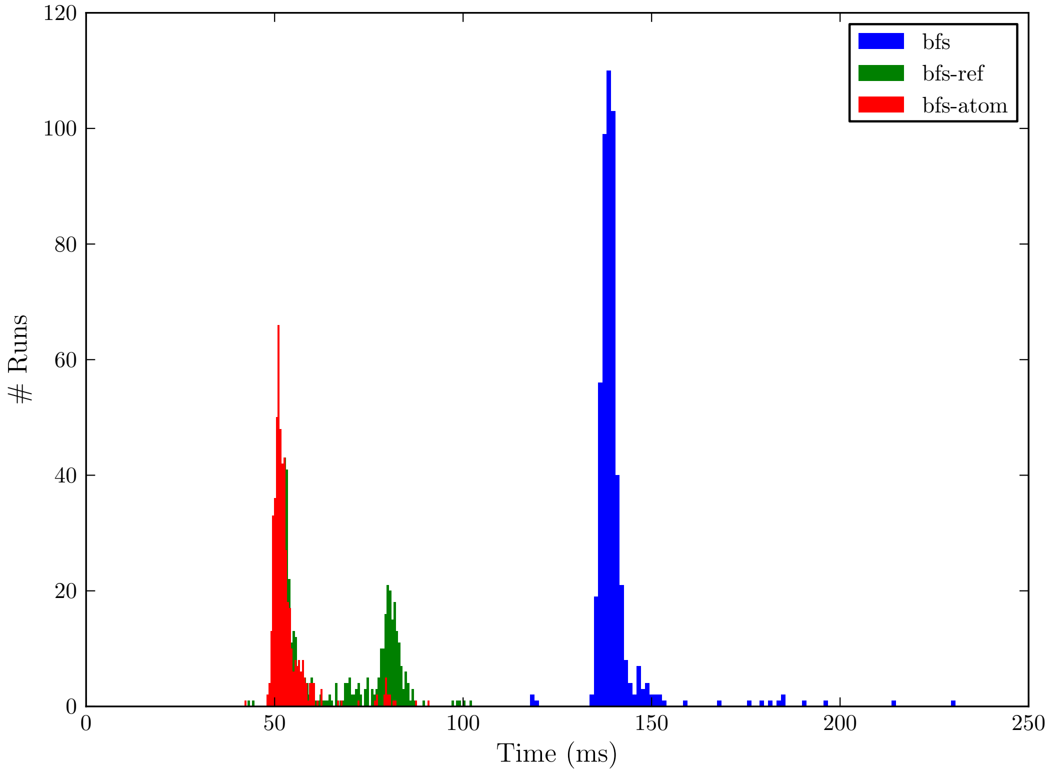


Figure 5.4: 4 Hops

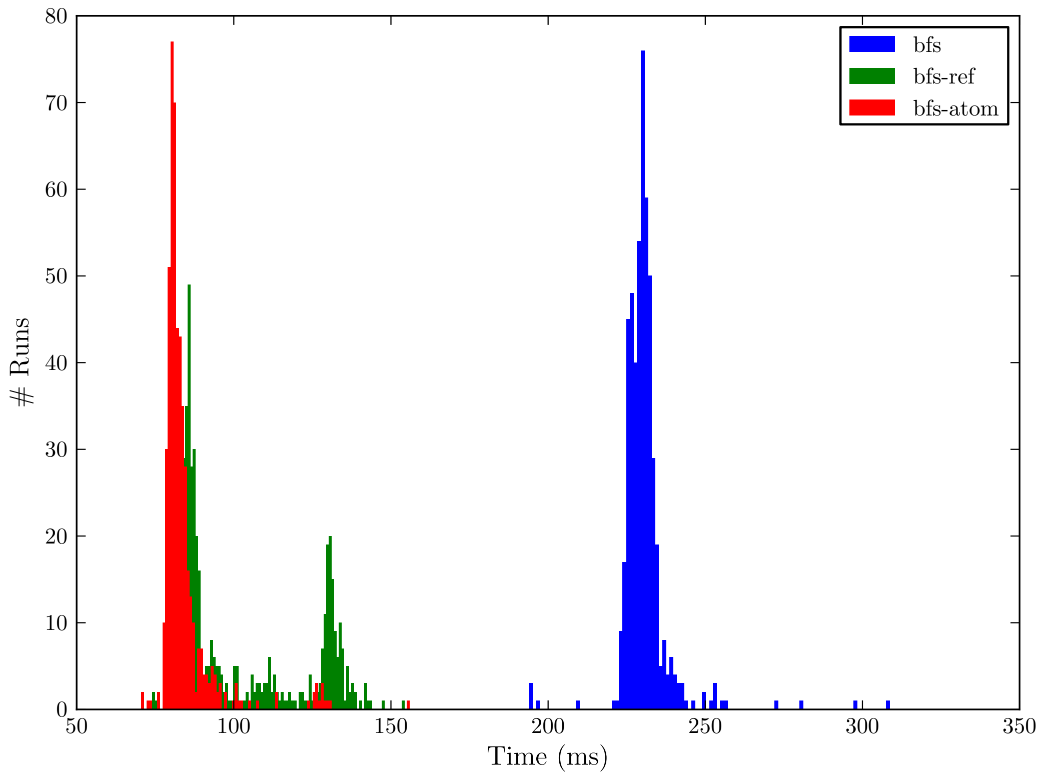


Figure 5.5: 5 Hops

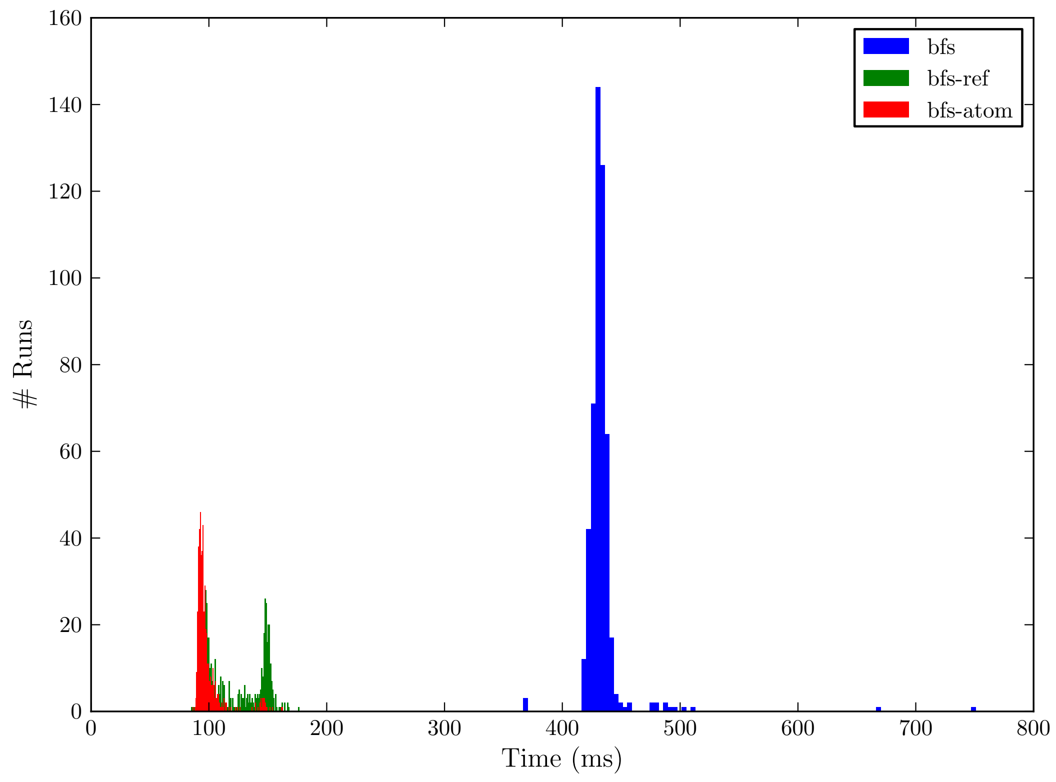


Figure 5.6: 6 Hops

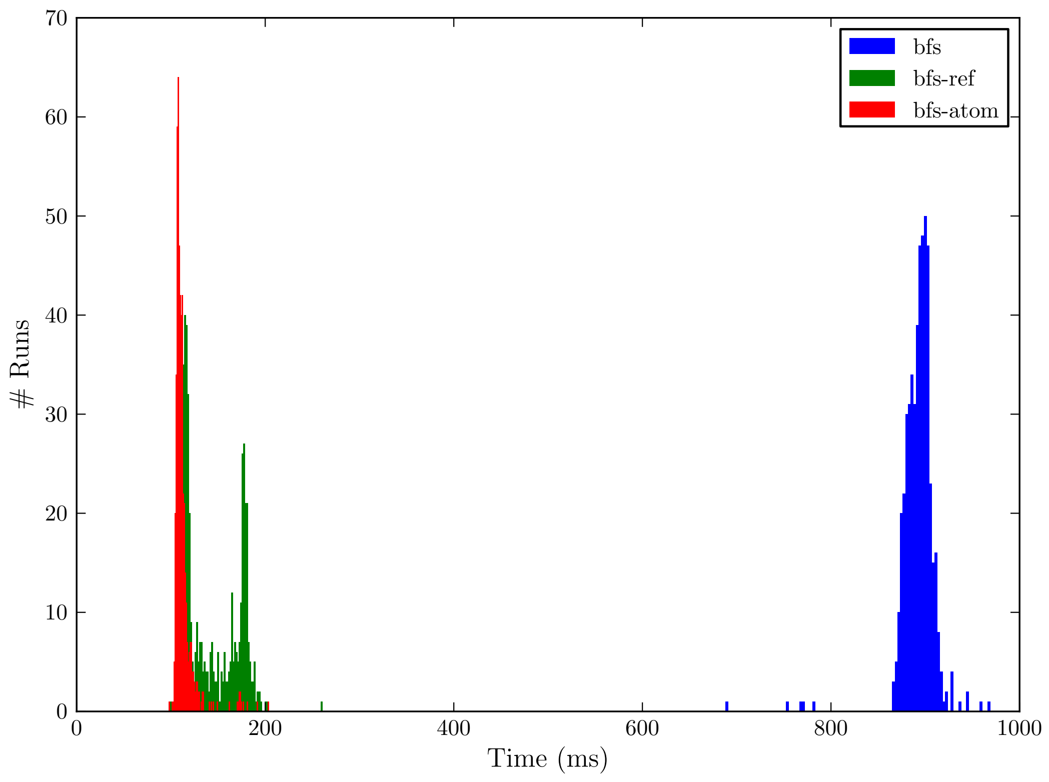


Figure 5.7: 7 Hops

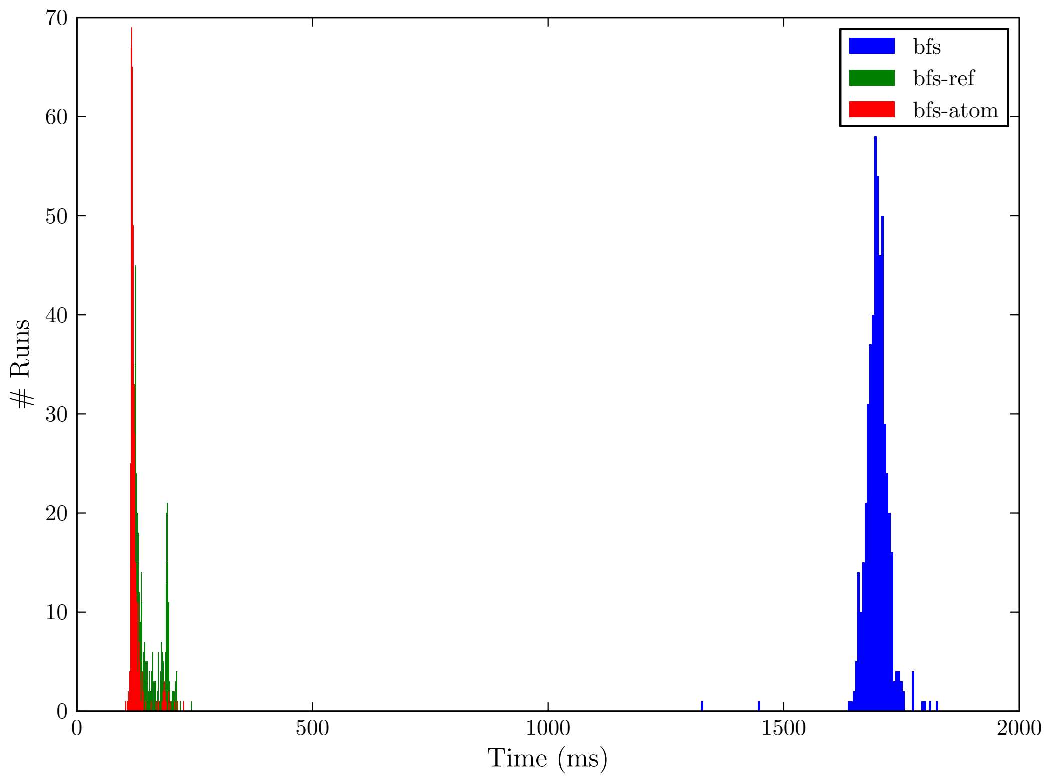


Figure 5.8: 8 Hops

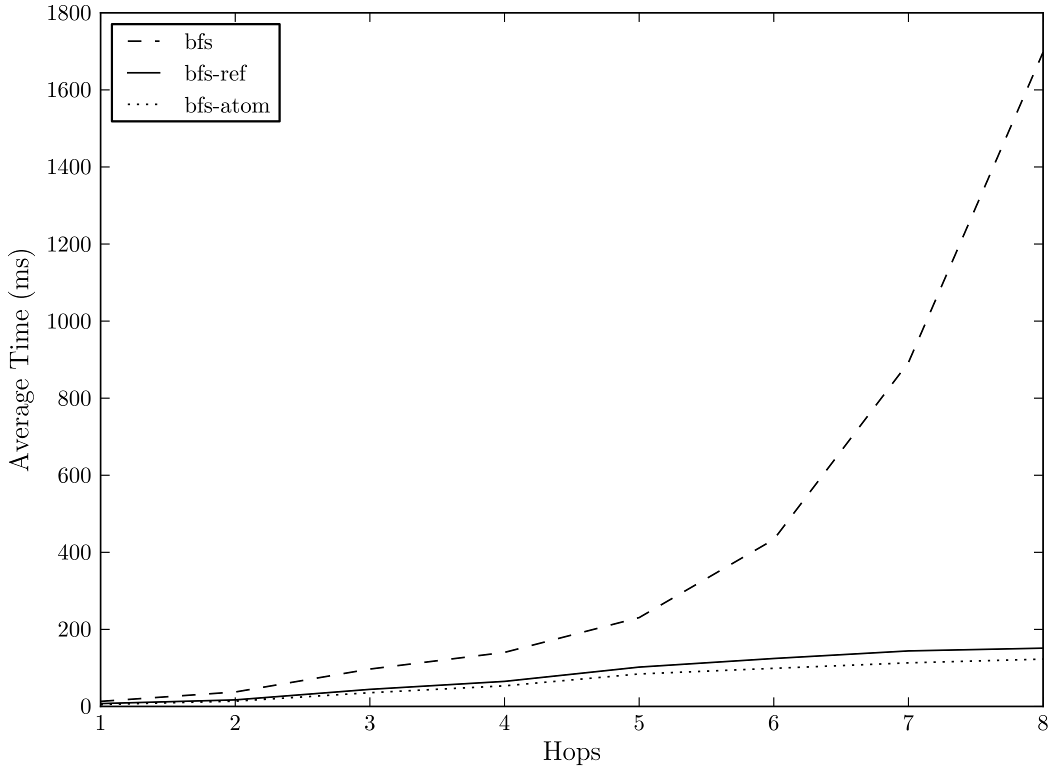


Figure 5.9: Comparison between single threaded and concurrent graph search

Chapter 6

Conclusion

[1]

Bibliography

[1] Brent Boyer. *Robust Java benchmarking, Part 1: Issues*. June 2008. URL: `http://www.ibm.com/developerworks/java/library/j-benchmark1/index.html`.

[2] Jonathan Corbet. *Counting on the time stamp counter*. Nov. 2006. URL: `http://lwn.net/Articles/209101/`.

[3] David Holmes. *Inside the Hotspot VM: Clocks, Timers and Scheduling Events - Part I - Windows*. Oct. 2006. URL: `https://blogs.oracle.com/dholmes/entry/inside_the_hotspot_vm_clocks`.