HASHING IN COMPUTER SCIENCE

Fifty Years of Slicing and Dicing

ALAN G. KONHEIM

WILEY

HASHING IN COMPUTER SCIENCE

FIFTY YEARS OF SLICING AND DICING

Alan G. Konheim

WILEY

JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Konheim, Alan G., 1934–
Hashing in computer science : fifty years of slicing and dicing / Alan G. Konheim.
p. cm.
ISBN 978-0-470-34473-6
1. Hashing (Computer science) 2. Cryptography. 3. Data encryption (Computer science) 4. Computer security. I. Title.
QA76.9.H36K65 2010
005.8′2–dc22

2009052123

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

HASHING IN COMPUTER SCIENCE

To my grandchildren …
Madelyn, David, Joshua, and Karlina

PREFACE

0.1 WHAT IS HASHING

The *Merriam-Webster Dictionary* (1974) provides several definitions for *hash*[1]:

- *noun:* an American food made by combining and compressing *chopped-up* leftovers—meat, eggs, and potatoes.
- *verb:* to chop into small pieces; make into hash; mince.
- *noun:* colloquial for *hashish* (or *hashish*), the resin collected from the flowers of the cannabis plant. The primary active substance is THC (tetrahydrocannabinol), although several other cannabinoids are known to occur. Hash is usually smoked in pipes, water pipes, joints, and hookahs, sometimes mixed with cannabis flowers or tobacco. It can also be eaten.

A search for *hashing* on *Google* yields 1,720,000 hits; of course, it doesn't come close to the 670,000,000 hits for *sex* or even the 129,000,000 hits for *money*, but then ⋯

Although hashing may be exhilarating and even intoxicating, this book deals with its applications in computer science relating to the processing of information that is both *old* (1953+) and *new* (1990+).

0.2 MY HASHING ROOTS

Nat Rochester was the software project manager for the IBM 701 machine; together with Gene M. Amdahl, Elaine M. McGraw (née Boehme), and Arthur L. Samuel, they considered a key-value-to-address machine for the 701 assembler. Several sources agree that Amdahl invented linear open addressing, a form of hashing (*aka* scatter storage), when confronted with the problem of collisions.

I was a Research Staff member in the Department of Mathematical Sciences at the IBM Thomas J. Watson Research Center in Yorktown Heights (New York) from

[1] Denis Khotimsky, a former student, directed me to www.onin.com/hhh/hhhexpl.htm, which gives an entirely different meaning to hashing, as follows:

> Hashing ⋯ it's a mixture of athleticism and sociability, hedonism and hard work, a refreshing escape from the nine-to-five dweebs you're stuck with five days a week. Hashing is an exhilaratingly fun combination of running, orienteering, and partying, where bands of harriers and harriettes chase hares on eight-to-ten kilometer-long trails through town, country, and desert, all in search of exercise, camaraderie, and good times.

This type of hashing began in Kuala Lumpur, Malaysia, in 1938.

1960 to 1982. Sometime in 1964, I was asked by my manager to meet with Wes Peterson (1924–2009) to learn about his work on hashing. Wes was a former employee of and now consultant to IBM; in 1964, he was a faculty member at one of the University of Florida campuses. Wes used simulation to evaluate the performance of *linear probing*, an algorithm developed during the design of the IBM 701 machine. As a result of our conversation, I started to work with Benjamin Weiss, then a post-doc at IBM, now a Professor Emeritus at the Hebrew University in Jerusalem, to develop a combinatorial analysis of hashing with linear probing.

In a footnote on page 529 of *The Art of Computer Programming: Volume 3/ Sorting and Searching* (Addison-Wesley Publishing Company, 1973), Professor Donald Knuth wrote that his running-time analysis of Algorithm L "was the first non-trivial algorithm I had ever analyzed satisfactorily. ···" He also noted "that more than ten years would go by before the derivation got into print!" Algorithm L is central to the performance of *linear probing*, which is essentially the hashing algorithm proposed by Gene Amdahl. Ben Weiss and I found the formula $N_j \equiv (j + 1)^{(j-1)}$ (Theorem 13.2 is Knuth's Algorithm L), which counts the number of hash sequences inserting j keys in a hash table of size $j + 1$ leaving the last hash table address unoccupied. First, we computed by hand the values N_j for $j = 2, 3, 4$ until we recognized the form of the algebraic expression $(j + 1)^{(j-1)}$, a technique that I have suggested to students in other contexts.

I contacted Wes again in 1997, where he was and remains today a Professor of Computer Science at the University of Hawaii. I inquired about a sabbatical leave there in 1999. My wife and I visited Wes and his wife Hiromi at their home in December 1997 while celebrating our 40th anniversary. Wes had just been awarded the Japan Prize, and both Petersons were both contemplating sabbatical leaves in Japan starting in January 1999. Timing is everything! I was hired by Wes and Hiromi to watch over their dog Koko; I was certainly the highest paid dog sitter in the Hawaiian Islands. Koko was a constant source of innocent merriment during my 5-month sabbatical; he curiously developed a significant taste for *Hebrew National* salami, and our acquiescing to his new found tastes was criticized by both Wes and Koko's veterinarian. However, before leaving for Japan, Wes warned my wife that *Koko* might die before their return. The miraculous and curative powers of *Hebrew National* salami—We *Answer* To *a Higher Authority*—added 2 years to Koko's life. We saw Wes again when we returned to Hawaii to celebrate our 50th anniversary in July 2007.

0.3 THE ORIGINAL APPLICATION OF HASHING (DATA STORAGE)

When data are stored and manipulated in a computer system, a mechanism must be provided to access the data easily. We view information in storage as composed of *records* of variable sizes located in the system storage media in a variety of regions, especially in dynamic file systems. A unique *key* is associated with and stored in each record. This key could be a name or a suitable number, but in general, any string can serve as a unique identifier. The key is used by the programs processing the data to locate—to `SEARCH` for the address of—the desired record. When a telephone directory is used to determine a telephone number, `SEARCH` is somewhat easier than the general instance of file management because the records composed of the triple `(Name Address TelephoneNumber)` are stored *sorted* on the `Name` *field*.

The concept of hashing was first mentioned in a January 1953 internal IBM document by Hans Peter Luhn and several years later in an article by Arnold I. Dumey; the word *hashing* was first used in an article by Robert Morris in 1968.

There is considerable literature for the *old* hashing starting at least in 1963 and continuing until 2005. We describe various hashing methods in Chapters 7–15.

0.4 NEW APPLICATIONS FOR HASHING

Because technology evolves quickly, certain problems arise in many diverse new disciplines. Chapter 16 describes the work of Karp and Rabin. In a 1987 paper, they observed that hashing could be applied to the search of data for a particular string, for example, the word *bomb* in monitored e-mail messages.

The tracking of music performed on radio and television is part of agreements between the Music Performance Trust Fund and the recording industry. As part of this process, it is necessary to listen to the performance of a song on radio and to identify a song; is the performance by the Glenn Miller band or U2? Techniques for *audio fingerprinting*—the automated recognition of music—appeared in 2002 and will be described in Chapter 17.

The IBM Corporation introduced the Data Encryption Standard (DES) in 1973. It generated considerable criticism but resulted in important changes in the relationship among the government, the business world and the academic community. The Web offered services to about 300,000 hosts (processing systems) in 1990. Today, millions of users are connected by the Internet. The October 2006 Forrester Research report predicted that "Non-travel online retail revenues ··· **e-commerce** ··· on the Web will top the quarter-trillion-dollar mark by 2011." This has produced, a new criminal activity called identity theft. Cryptography offered a possible solution to these problems.

Transactions on the Web require the following:

- *Secrecy* to hide credit card information transmitted from a buyer to a seller
- *User authentication* to verify to each party in a transaction the identity of the other party
- *Message authentication* to detect unauthorized changes in transmitted data

Cryptography has been ported to the world of e-commerce, and hashing is one of the constituents of the secrecy/authentication protocols for e-commerce used to protect and defend Web transactions. We describe the exciting current work on hashing in e-Commerce in Chapter 18.

Not only are the mom-and-pop stores on main street disappearing, but it seems that only the large-box chain stores may survive. The villain or savior is the Web, which provides a convenient and often cost-effective way of purchasing some items and/or services. Among the products easily distributed in this medium are music, video, software, books, and data.

Because information in its native form can easily be copied, distributed, or altered without the owners or the purchasers of the information being aware, their unrestricted distribution poses serious issues. We show how hashing is used in Chapter 19 to inhibit these forms of misuse.

0.5 ABOUT THIS BOOK

This book will describe the basic applications of hashing, analyze hashing protocols, and illustrate the tools needed by a student to understand basic and fundamental problems in computer systems.

Why should students study hashing, and why this book? The purpose of higher education in computer science and mathematics is to prepare students for careers in industry, government, and the university. Computer science—like music and cooking—is not a spectator sport; it requires participation. We all learn by practice, attempting to solve new problems by adapting the solutions of old problems. This book focuses on mathematical methods in information management, but its structural techniques are applicable in many areas of science and technology.

The intended audience includes upper-division students in a computer science/engineering program. The prerequisites include skill in some programming language and elementary courses in discrete mathematics and probability theory. The book is divided into four parts.

Part I contains brief reviews of the relevant mathematical tools

Part II deals with the original hashing for data storage management

Part III examines the appearance of new applications of hashing ideas

Problems and solutions are included as well as an extensive bibliography which concludes the presentation

0.6 ACKNOWLEDGMENTS

I am in debt to many people who have helped and encouraged me in writing this book.

- My friend of 36 years, Dr. Raymond Pickholtz, Professor Emeritus at George Washington University, who visited UCSB several times, read all the chapters, and provided advice.
- His son, Mr Andrew Pickholtz *Esq.*, who read and added material on his dongle patent in Chapter 19.
- Dr. Michele Covell of Google who was kind enough to read and suggest changes to Chapter 17.
- Mr. Vlado Kitanovski of the Faculty of Electrical Engineering and Information Technologies (Ss Cyril and Methodius University—Skopje) who was kind enough to read and suggest changes to the section on watermarking of images in Chapter 19.
- Dr. Feng Hao of Thales e-Security (Cambridge, United Kingdom) and Professor Ross Anderson of the University of Cambridge Computer Laboratory for their comments on iris identification in Chapter 19.
- Professor Haim Wolfson of Tel-Aviv University who clarified my presentation on geometric hashing.
- My son Keith whose graphic skills were indispensable.

- Finally, to my wife Carol of more than 52 years, who continues to amaze me by her wide-ranging talents. I could not have undertaken this book without her encouragement, assistance, and advice.

It's what you learn after you know it all that counts, attributed to Harry S. Truman.

Anyone who stops learning is old, whether at twenty or eighty, Henry Ford.

Teaching is the highest form of understanding, Aristotle.

CONTENTS

PART I

MATHEMATICAL PRELIMINARIES