# A Collection of Research Processes for Genealogy and Proofs

VOLUME FOUR, SECTION 27

The Paper, "A Comparative Study of Hash Functions for an Effective Hash Coder." was proofread by Ms. Cayo Gamber before Submission to Communications of the ACM (August 21, 1991)

by

Dr. Dong-Keun Shin

June 13, 1996

Submitted to the Chair of
Department of Electrical Engineering and Computer Sciences
College of Engineering
University of California, Berkeley
Berkeley, CA 94720

(2)

# A COMPARATIVE STUDY OF HASH FUNCTIONS FOR AN EFFECTIVE HASH CODER

## ABSTRACT

This study surveys several newly developed hash functions along with well-known hash functions such as algebraic coding, digit analysis, division, folding, midsquare, multiplicative, radix, random, and Pearson's table indexing. The comparative analysis of the hash coders in an open hashing scheme was based on criteria such as speed, distribution, and cost.

The study developed a new, fast, hardware oriented hash function to serve as a hash coder in a processor. The new mapping hash method is designed to take advantage of parallel processing. Moreover, this ~~The~~ new mapping hash method not only has reliable and relatively good key distribution, but it also takes only three clock cycles to calculate a hash address if the mapping hash coder is implemented in hardware. This paper concludes that the new mapping hash method is a reasonable choice for an effective hash coder.

indexed value (H(n)) from the table T becomes the hash address for the buckets ranging 0 through 255.

Figure 6. Pearson's Auxiliary Table T

Figure 7. Pearson's Hash Algorithm

5. An Analysis of Distribution, Speed, and Cost

Table 1 shows each hash function's performance in terms of distribution, in terms of speed when implemented either in software (SW) or in hardware (HW), and in terms of the cost of the hardware implementation of the hash function. For measurement of distribution, mean square deviation (MSD) is provided whenever a hash function is applied to the three different data sets: randomly chosen names (RCN), generally chosen names (GCN), and randomly chosen numeric strings (RNS). The number of clock cycles (clocks) is used in the measurement of the speeds of the hash coders. The cost of building a hardware hash coder is represented according to the number of gates needed.

Distribution performances of the mapping hash method have been developed in cases when each ROM contains both prime numbers and ~~when each ROM contains~~ random numbers. As shown in the Tables 2A and 2B, mean square deviations hover around four, as do those of other relatively good hash methods. Since there is no distinguishable difference between using prime numbers and random numbers for each ROM, there is no clear reason to insist on solely prime numbers. The results do not provide any clue regarding data dependency, since the mapping hash function distributes

numeric string keys as well as other keys. Different groups of eight bits, e.g., 1-8, 2-9, 3-10, 4-11, 5-12, 6-13 bits, are extracted to compose a hash address (The 1-8 means bits 1 through 8 are selected.). In summary, there is no noticeable difference between the distribution performances of the various groups.

I think "needed" would be stronger than here. Would it alter the meaning if you used "needed" rather than "taken"?

Table 2A.

Table 2B.

Table 1. Performances of Hash Functions.

By virtue of byte-by-byte parallel processing, with separate ROM and exclusive-OR module, the mapping hash method can produce a hash address within three clock cycles. Two clock cycles of the MC68030 processor are required in order for the memory read to retrieve a random number from the corresponding ROM, as is specified in the Motorola's users manual <MOTO1>. One clock cycle is [needed] [taken] for the calculation process for hash address bits through the four levels of exclusive-OR gates. The maximum gate delay is n e nanoseconds and the clock frequency is set to 20 MHz (50 nanoseconds per a clock pulse width); thus, the address bit signal can pass through

the four gate levels (4*9 = 36 <= 50 nsec) within a clock cycle.

Based on the stored contents (selected prime numbers) of the ROMs, each mapping hash coder calculates a hash address in its unique way. The hash addresses generated by different mapping hash coders are independent of each other, but the address calculation time for each hash coder is always the same. This characteristic of statistical independence is also valuable in an application environment which uses rehashing scheme. The additive mapping hash method shows competitive distribution performances (MSDs of 4.40, 3.39, 3.58) when it is tested. This result supports the claim that addition and exclusive-ORing produce the same effect in randomizing the bit values.

The distribution performances of the author's fold-shifting hash method, in particular, FS(0,10,20,30) and FS(0,11,22,25), are as good as those of other acceptable hash methods. But other selected fold-shifting methods, such as FS(0,12,17,29), FS(0,13,18,31), and FS(0,15,22,29), show a data dependency problem, such that the distribution performance on the RNS data set is not compatible with the distribution performance on the RCN and GCN data sets, as is demonstrated in Table 3. Therefore, careful selection of the number of partitions and the number of rotated bits is required.

The distribution performance of the division hash method <BUCH1, MAUR1, LUM1> varies, depending on the chosen divisor which is close to the number of buckets, as is shown in Table 4. If an inappropriate divisor is chosen, a data dependency problem may occur. In this experiment, the divisors which are greater than the number of buckets in a table (i.e., 256) are also tested. The divisor 257 is a nonprime number with prime factors less than 20, as recommended by Lum and his colleagues, but it shows very poor distributions (MSDs of 5.67, 11.95, and 122.99). As Maurer

keep this comma in (STET)

and Buchholz suggested <MAUR2, BUCH1>, using the largest prime number (i.e., 241) that also is smaller than the number of buckets, as the divisor yields better results (MSDs of 5.51, 5.35, 4.48).

take this comma out

Table 4. Distribution Performance of the Division Hash Method

T le 3. Distribution Performances of Various Fold-shifting Hash Methods

Several other researchers <BUCH1, LUM1, RAMA1> conducted experiments on typical key sets in order to discover the ideal hash method. Their overall conclusions verify that the simple method of division seems to be the best key to address transformation technique when computational time is not critical. Nevertheless, in this survey of hash methods, the division method is not highly recommended, since either the mapping or the additive mapping method can be used instead, depending on the application environment. In the application, where fast hash address calculation is not required, the additive mapping method is superior to the division method. When using the additive mapping method, one need not worry about selecting a correct divisor; one need only divide the sum or combination by the number of buckets in order to arrive at a remainder for a hash address. On the other hand, when the speed in address calculation is

delete the comma

take this comma out

imperative and the number of buckets can be 2**n, then a hardware hash coder is needed, and the mapping hash coder which is faster and cheaper than the division hash coder is thus recommended.

Pearson's table indexing hash method appears to be erratic owing to its poor distribution performance. The fold-boundary and the midsquare show data dependency problems as shown in tables 5 and 6 respectively. In particular, The multiplicative, the radix, and the random hash fuctions show signs that they may perform poorly for specific data sets. The distribution performance of the digit analysis hash method is measured by using two types of encoded keys: 2 bytes and 4 bytes as shown in table 1. The findings indicate that this hash method may be data dependent. Both Maurer (see table 7 for more information) and Berkovich present new hash methods that have proved to be proficient in distribution performance. Their methods, however, have not been highly recommended for the effective hash coder due to their relatively slow hash address calculation speeds.

The hash functions such as midsquare, multiplicative, radix, random ar algebraic coding use ~~the~~ complex mathematical operations, ~~like~~ e.g., multiplication and division. Their speeds of hash address calculation can be

increased by fast multipliers and/or dividers. These fast multipliers and dividers, however, are quite expensive. Since there are speed versus cost trade-offs, any judgement regarding adaption must be made thoughtfully. For that reason, the gates of these options also are reflected in the costs of a hash coder in order to help a computer designer make the best decision.

## 6. Summary and Conclusions

If huge amounts of data pass through a hash coder, the hash address calculation should be very fast. In order to speed up the hash address computation, efforts should be concentrated on designing a new hash function that will avoid time-consuming serial and/or iterative computations while taking advantage of parallel processing, by means of hardware, for converting a key into a hash address. Moreover, the new hash algorithm should distribute random keys into buckets as uniformly as possible. The ideal hash function design for this application is thus data-independent and calculates a hash address within a few machine cycles with relatively good distribution.

Most of the well known hash functions, and several new ones, including mapping, additive mapping, shift-fold-loading, Hu-Tucker code, and various versions of fold-shifting, are surveyed in this paper. Each hash function has been simulated and applied to two different name data sets (RCN and GCN) and one numeric string data set (RNS) to produce distribution performances measured in terms of mean square deviations. The speed of calculating a hash address is measured in terms of clock cycles for each hash function in both the hardware and software implementation cases. The cost of the hardware implemented hash coder may be calculated and stated in terms of the number of gates used.

As the results illustrated in the above tables indicate, some of the well known hash functions, such as the midsquare and the fold-boundary,

show data dependency problems. Other hash functions, like the multiplicative, the radix, and the random, show signs that they may perform poorly f.. specific data sets. New shift-fold-loading and Hu-Tucker code hash methods have good distribution performances, but they are not fast in hash address calculation. The new Fold-shifting hash methods (FS) are not very reliable in terms of distribution performance; nonetheless, they are fast and inexpensive. This survey also verifies that there is no distinguishable difference in distribution performances of relatively good, data-independent hash functions.

As shown in Table 1, the mapping hash method satisfies all three requirements at the highest rank. The mapping hash does not require an encoding scheme. This hash method involves the combination of the mapping or converting of each character in a key to a corresponding prime-number or random-number technique and the folding technique. The parallel processing of the mapping hash coder transforms each character into a number and calculates each bit value in a hash address by means of hardware in order to produce a hash address within three clock cycles. Other hash methods cannot take advantage of such effective parallel processing due to the algorithmic nature of their hash address calculation. The mapping hash coder in hardware is relatively inexpensive compared to other hardware hash coders which use complex mathematical operations like multiplication and division. Furthermore, compared to other well-known methods, this mapping hash method distributes keys effectively. The mapping hash method is also sensitive to every character in a key producing a hash address; that is, it does not have a data dependency problem in its distribution of similar keys. The new mapping hash method is thus recommended for an effective hash coder in various applications.

CURRICULUM VITA

Tayo Elizabeth Gamber

20 May 1991

You may want to date (and continually update) yours in order to show how recent the information is.

**PERSONAL HISTORY**

Business Address: The Department of English
The George Washington University
Washington, D.C. 20052

Phone: (202) 994-5969

Home Address: 2024 North Scott Street #304
Arlington, VA 22209

Phone: (703) 525-5827

Birthdate: 30 July 1959

**EDUCATIONAL HISTORY**

Ph. D. The George Washington University (February 1991):
Primary area of concentration in British Literature (specializing in Drama and the Twentieth Century),
Secondary area of concentration in Rhetoric and Composition
Dissertation: No Place/Like Home: Feminism, Semiotics, and Staged Space in Osborne, Pinter, Stoppard, and Churchill
Director: Gail Kern Paster

it often is a good idea to include the name of your director.

M. Phil. The George Washington University, (awarded 1986)

B.A. The College of William and Mary (1975-1976, 1977-1979)
The University of Barcelona (1976-1977)
The Normal University of Taiwan (1974)

majors: English and Spanish
minors: Geology and Mandarin

**POSITIONS**

| | |
|---|---|
| The George Washington University: | Lecturer in English 1987-present |
| The George Washington University: | Graduate Teaching Assistant 1983-1986 |
| The George Washington University: | Tutor, the Writing Center 1983 |
| The George Washington University: | Tutor, the Athletic Department 1986-present |
| The George Washington University: | Tutor, Disabled Student Services 1989-present |

**TEACHING EXPERIENCE**

At the George Washington University:

| | |
|---|---|
| English 9: | English Composition: Language as Communication |
| English 10: | English Composition: Language as Communication |
| English 10: | English Composition: Language as Communication: Special Topic: A Community Study of Dupont Circle, Washington, D.C. |
| English 11: | English Composition: Language and the Arts and Sciences: Special Topic: A Community Study of the George Washington University, Washington, D.C. |
| English 11: | English Composition: Language and the Arts and Sciences: Special Topic: Social Constructions of Race, Class, and Sexuality |
| English 12: | English Composition: Language in Literature |
| English 12: | English Composition: Language in Literature: Special Topic: The Theatrical Production: A Collaborative Enterprise |
| English 101: | Advanced Writing |
| English 52: | Introduction to English Literature |

you might want to consider highlighting your teaching experience as I have done.

**APERS AND PRESENTATIONS**

"A 1978 Chevy Parked in the Front Yard: Social Constructions of Class, Race, and Sexuality," First Annual Conference of Teachers of Writing in the Washington Area, the George Washington University, 13 April 1991.

"Lapsing into French: The Unwitting Injuries of Class," Women's Studies Department Colloquia, the George Washington University, 8 April 1991.

"From Vaginas to Bearded Clams: How Synonyms, Euphemisms, Slang, and Slurs Inform Our Language and Our Cultural Perspective," Currents of Change: Feminist Research, 1991, the University of Maryland, 8 March 1991.

"Housewives and Commercials: Who Put the Gravy in the Gravy Train?" Women's Studies Department Colloquia, the George Washington University, 4 December 1990.

"Conjugal Women and Conjugal Beds: The Portrayal of the Bedroom in Osborne, Pinter, Stoppard, and Churchill," the 1990 Conference of the National Women's Studies Association, University of Akron, Ohio, 22 June 1990.

"Bartleby the Scrivener as Woman in Drag: A Deconstructionist Performance in Performance," the 1990 Conference of the National Women's Studies Association, University of Akron, Ohio, 21 June 1990.

"Journals as a Risk-Free, Error-Free Space for Both Developmental Writers and Their Instructors," Annual Conference on Composition and Communication, the George Washington University, Washington, D.C., 24 February 1990.

"Strategizing Doctorate Examinations," the Graduate Student Association Forum, the George Washington University, Washington, D.C., 15 March 1989.

"The Methodology and Philosophy of Teaching Composition Courses with a Thematic Focus," the Annual Conference on Composition and Communication, the George Washington University, 8 February 1986.

"Shopping Lists as Acts of Coherence: Applications of Ann Berthoff's Methodology in the Teaching of Composition," the Writing Center Seminar, the George Washington University, 11 November 1985.

**ARTICLES UNDER CONSIDERATION AND WORKS IN PROGRESS**

Do you have similar scholarship under consideration you need to list that scholarship.

"The Hidden Agenda in Commenting on Student Writing, or, 'Please Do Not Be Discouraged by These Comments,'" under consideration at Journal of Teaching Writing.

"Barbie as the Text of My Life," to be published in a forthcoming issue of On the Issues.

"The Use of Journals for the Developmental Writer," under consideration at Freshman English News.

"Dramatic Representations of the Domestic in Osborne, Pinter, Stoppard, and Churchill," dissertation at the George Washington University, defended 30 November 1990.

change title

note to e. signed this note Donna

**COMMITTEE MEMBERSHIPS**

at the George Washington University:

The Graduate Student Association 1985-1990
Part-time Faculty Representative to Meetings of the Full-time Faculty 1988
Chair of the English Department Colloquia 1985-1987
The Textbook Committee 1985-1989

The Curriculum Committee 1985-1989
Co-chair of the Committee for the Annual Conference on Composition and Communication 1988
Graduate Student Representative to the English Department Renovation Committee 1984-1985

**PROFESSIONAL ORGANIZATIONS**

Modern Language Association
National Council of Teachers of English
National Women's Studies Association
South Atlantic Modern Language Association
Virginia Association of Teachers of English

**LETTERS OF REFERENCE**

Dossier available on request from Career Services Center, the George Washington University, 801 22nd Street, N.W. Suite T509, Washington, D.C. 20052, (202) 994-6495.

you may want to consider rephrasing your reference section accordingly. The reader should have the address and phone number readily available.