НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ УКРАИНЫ
" КИЕВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ "

На правах рукописи

БАРДИС ЕВГЕНИОС
(Греция)

681.322.01

МЕТОДЫ И СРЕДСТВА ПОВЫШЕН
ЭФФЕКТИВНОСТИ ХЕШ-АДРЕСАЦ

Специальность 05.13.08-Вычислительные ины, системы и
сети, элементы и стройства вычисли-
тельной техники и систем управления

А В Т О Р Е Ф Е Р А Т
диссертации на соискание ученой степени
кандидата технических наук

Киев-1995 г.

Диссертацией является рукопи

Работа выполнена в Киевском
вычислительной техники.

Научный руководитель: - доктор технических наук
член-корреспондент НАН Украины
профессор Самофалов Константин Григорьевич

Официальные оппоненты: - доктор технических наук,
профессор Кузьмук Валерий Валентинович
- кандидат технических наук,
Селигей Александр Минович

Ведущая организация - Киевский государственный университет
строительства и архитектуры

Защита состоится 19. 06. 1995 г. в 14-30 часов на заседании специализированного Совета Д 01. 02. 06 в Киевском политехническом институте (г. Киев, пр. Победы, 37, корп. 18, ауд. 306)

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просим направлять по адресу: 252056, г. Киев, пр. Победы, 37, Ученому секретарю КПИ.

С диссертацией можно ознакомиться в библиотеке Киевского политехнического института

Автореферат разослан "18" мая 1995 г.

Ученый секретарь
специализированного Совета
доктор технических наук,
профессор О. В. Бузовский

АВ-32.457

А Н Н О Т А Ц И Я

В диссертационной работе исследуются информационные аспекты организации хеш-памяти и разрабатываются методы и средства повышения ее эффективности с целью уменьшения информационной избыточности хранения данных при хеш-адресации.

Основные задачи исследования определены в следующем:

1. Сравнительный анализ и классификация методов уменьшения (для динамических массивов) или исключения (для статических массивов) коллизий при хеш-адресации.

2. Исследование организации хеш-преобразования и хранения ключей с целью выявления возможностей уменьшения информационной избыточности при хранении ключей в хеш-памяти.

3. Исследование процессов преобразования информации при использовании хеш-алгоритмов различных классов: определение требований к выбору процедур получения хеш-адреса и кода, сохраняемого в памяти по хеш-адресу и названного автором хеш-сверткой.

4. Разработка структурной организации хеш-памяти, основанной на комбинированном использовании хеш-свертки и ограниченного пробинга с дополнительной памятью синонимов, как средств разрешения коллизий при неизменном хеш-алгоритме и динамичном массиве ключей.

5. Разработка структурной организации хеш-памяти без коллизий для хранения статических массивов (perfect hashing или P-хеш-пямяти), основанной на использовании хеш-свертки и многоуровневого P-хеш-преобразования.

6. Разработка математических моделей а также методик анализа и оптимизации предлагаемых структур хеш-памяти.

7. Разработка средств сжатия и защиты информации на основе использования хеш-адресации.

Автор выносит на защиту следующие основные положения и результаты:

1. Метод сокращения информационной избыточности хранения ключей в хеш-памяти на основе использования специальных алгоритмов хеш-преобразования и сохранения в памяти хеш-свертки.

2. Структурную организацию хеш-памяти для динамических массивов на основе использования хеш-сверток и разрешения коллизий

дополнительной памяти ключей-синонимов.

3. Способ сокращения времени нахождения Р-хеш-алгоритмов за счет комбинированного использования хеш-свертки и организации многоуровневого хеш-преобразования.

4. Методику сжатия информации с использованием хеш-преобразований для решения задач защиты информации от несанкционированного доступа.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы: Достигнутое в последние годы повышение производительности средств обработки информации связано с совершенствованием, в первую очередь, процессорных средств и распараллеливанием вычислительных процессов на процессорном уровне. Вместе с тем, для значительной части практически важных в настоящее время и перспективе, применений вычислительной техники повышение производительности связывается с распараллеливанием процессов обработки информации на уровне памяти. К таким применениям можно с полным правом отнести и традиционные, такие как, информационно-поисковые системы, системы баз данных и баз знаний, а также перспективные, такие как, системы распознавания образов, лингвистические процессоры, системы защиты информации.

В упомянутых применениях скорость выполнения операций поиска в памяти является определяющим фактором производительности. Как правило, поиск производится по ключу в информационных структурах, обьемы и сложность которых имеют устойчивую тенденцию к возрастанию, в то время, как требования в времени поиска ужесточаются.

Указанные обстоятельства диктуют необходимость в разработке новых, а также совершенствованию традиционных методов и средств поиска информации в памяти. К числу последних относится хеш-адресация, которая, являясь наиболее быстродействующим средством поиска поиска, широко применяется в информационно-поисковых системах и трансляторах с 60-х годов. В отличие от методов дихотомического поиска и В-деревьев, скорость поиска при использовании хеширования не зависит от обьема поискового массива.

Основным недостатком хеширования является сложность реализации, обусловленная наличием коллизий. Указанная проблема сдерживала развитие и практическое использование хеш-адресации и только в начале 90-х годов, с появлением высокопроизводительных ЭВМ поя-

вились реальные предпосылки ее решения путем использованием perfect хеш-алгоритмов (P - хеш-алгоритмов), которые для фиксированного поискового массива не порождают коллизий.

Достигнутый в последние годы рост производительности вычислительных устройств в сочетании с удешевлением их аппаратной реализации создает предпосылки для использования более сложных алгоритмов хеш-преобразований и структур хеш-памяти, позволяющих резко повысить ее эффективность.

При этом, для реализации этих, потенциально значительных возможностей хеш-памяти , необходимо решить ряд задач, связанных с совершенствованием форм хранения информации, организацией выбора эффективных хеш-алгоритмов, разработкой структур средств аппаратной поддержки хеш-адресации, применением хеширования в составе систем обработки информации. Означенные задачи и определяют предмет исследования данной диссертационной работы.

П Р Е Д М Е Т О М И С С Л Е Д О В А Н И Й в диссертационной работе является структурная организация хеш-памяти, ориентированная на эффективную реализацию хеш-доступа за счет уменьшения информационной избыточности хранения данных.

М Е Т О Д Ы И С С Л Е Д О В А Н И Й. В диссертационной работе использованы теоретические положения и методы теории вычислительных процессов, теории информации, теории вероятностей и математической статистики, теории множеств. Экспериментальная проверка полученных результатов проводилась методами статистического моделирования на ЭВМ.

Н А У Ч Н А Я Н О В И З Н А. Предложены метод уменьшения информационной избыточности хранения данных при хеш-адресации, структурная организация хеш-памяти, алгоритмическое и программное обеспечение реализации метода, а также определены возможности использования полученных результатов для защиты информации.

П Р А К Т И Ч Е С К А Я Ц Е Н Н О С Т Ь работы состоит в разработке структурных и алгоритмических методов построения эффективной хеш-памяти, которая может быть использована для быстрого доступа в информационно-поисковых системах, высокопроизводительной реализации поиска в таблицах трансляторов, лингвистических процессоров и систем распознавания образов, а также для надежной защиты информации от несанкционированного доступа и идентификации цифровой подписи. Практическая полезность и новизна полученных результатов подтверждается авторскими свидетельствами СССР.

П У Б Л И К А Ц И И. По теме диссертации опубликовано 4 работы .

С Т Р У К Т У Р А И О Б Ъ Е М Р А Б О Т Ы.

Диссертационная работа состоит из введения, четырех глав, приложения, заключения и списка используемой литературы из 130 наименований. Работа содержит 128 страниц машинописного текста, 12 рисунков на 7 страницах и 4 таблиц на 3 страницах.

Во введении обоснована актуальность темы диссертационной работы, сформулированы цели и задачи исследований.

В первой главе выполнен обзор литературных источников по теме диссертации ,в котором рассмотрены и критически проанализированы различные подходы к повышению эффективности как хеш-памяти с коллизиями, так и памяти, использующей Р - хеш-адресацию. Проведено исследование зависимости показателей эффективности хеширования от коэффициента использования объема накопителей для хеш-памяти с различной организацией хеш-доступа.

Во второй главе рассмотрены информационные аспекты организации хеш-памяти, в результате чего выявлены резервы повышения ее эффективности за счет сокращения информационной избыточности хранения данных. Исследованы возможные пути реализации этих возможностей с целью уменьшения числа возникающих коллизий и ускорения нахождения совершенных хеш-функций для хеш-памяти без коллизий. Сформулированы требования к хеш-алгоритмам, допускающим возможность сжатия информации в хеш-памяти. Предложены структуры хеш-памяти с коллизиями , в основе которых лежит комбинированное использование хеш-свертки, ограниченного пробинга и памяти переполнения.

В третьей главе исследованы вопросы структурной организации хеш-памяти на основе комбинированного использования хеш-свертки и многоуровневых Р- хеш-функций. Проанализированы возможности использования такой хеш-памяти для организации быстрого доступа к статическим и динамическим информационным структурам. Предложены математические модели, методы расчета и оптимизации структур хеш-памяти рассмотренного класса.

В четвертой главе рассмотрены вопросы практического использования предложенных методов синтеза структур хеш-памяти в информационно-поисковых системах на основе дисковой оптической или магнитной памяти, а также для защиты информации от несанкционированного доступа и идентификации подлинности информации.

СОДЕРЖАНИЕ РАБОТЫ

Методы хеш-адресации, как средства поиска информации по ключу начали развиваться с 60-х годов, одновременно в СССР и США. Основным достоинством хеширования является высокая скорость поиска, практически не зависящая от обьема поискового массива.

Основной проблемой, сдерживавшей долгое время широкое использование хеш-памяти является наличие коллизий, то есть явления, когда в результате хеш-преобразования, разным ключевым словам соответствует один и тот же хеш-адрес. Наличие коллизий многократно усложняет выполнение операций в хеш-памяти, имеет следствием увеличение времени обращения.

Разрешение коллизий в настоящее время выполняется тремя способами: применением различных видов пробинга, введением памяти переполнения, а также использованием хеш-алгоритмов, не порождающих коллизии для данного набора ключей , последний класс алгоритмов хеш-преобразования получил название perfect хеш-алгоритмов (P - хеш-алгоритмов).

Исследования, направленные на разработку методики получения совершенных хеш-алгоритмов начались с конца 80-годов. Основной трудностью нахождения P- хеш-алгоритмов является необходимость в большом обьеме требуемых для этого вычислений. Современные высокопроизводительные процессорные средства позволяют решать эту проблему, при выполнении определенных условий, за приемлимое для практически важных задач, время.

Анализ большинства предложенных к настоящему времени методик формирования P- хеш-алгоритмов позывает, что в их основе лежит использование некоторого ограниченного класса алгоритмов хеш-преобразования, выбор подходящего для заданного массива из которого производится методом перебора. Эффективность методики формирования P- хеш-алгоритма определяется временем, затрачиваемым на выполнение указанного перебора, которое , в конечном итоге оказывается зависимым от выбранного класса функций хеш-преобразования , коэффициента загрузки хеш-памяти и количества ключей в массиве.

Важнейшим фактором, влияющим на эффективность хеш-памяти является коэффициент α загрузки. Применительно к хеш-памяти с коллизиями для анализа влияния этого фактора целесообразно отдельно рассмотреть два случая:

- в хеш-памяти число выполняемых операций записи и исключе-

ния во много раз превышает число хранящихся в ней ключей;

- в хеш-памяти выполняются только операции записи или общее число циклов записи и исключения соизмеримо с количеством хранящихся в хеш-памяти ключей.

Для первого случая среднее число T_{01} циклов обращения к памяти необходимых для поиска по ключу, при использовании хеш-алгоритма, порождающего равномерно распределенных хеш-адреса, определяется формулой:

$$T_{01} = \frac{1}{1-\alpha} \qquad (1)$$

Для второго случая, при тех же условиях, среднее число циклов обращения к памяти определяется выражением:

$$T_{02} \approx -\frac{\ln(1-\alpha)}{\alpha} \qquad (2)$$

Результаты статистического моделирования подтверждают приведенные теоретические оценки.

Применительно к проблеме нахождения P- хеш-алгоритмов, коэффициент α загрузки хеш-памяти оказывает заметное влияние на общее время T_n перебора хеш-алгоритмов. Так как процесс перебора можно описать вероятностной моделью однородных испытаний Бернулли, то среднее время T_n перебора оказывается величиной , обратной по отношению к вероятности $P(\alpha,m)$ того, что данный хеш-алгоритм окажется для заданного массива ключей perfect, которая является функцией от α и m- числа ключей, размещаемых в M ячейках хеш-памяти. Указанная зависимость приближенно выражается следующим образом:

$$P(\alpha,m) = e^{-m\frac{\alpha^2}{2}\cdot e^{-\alpha\frac{3}{3-\alpha}}} \qquad (3)$$

Результаты проведенных экспериментальных исследований подтверждают приведенную теоретическую оценку.

Анализ зависимостей позволяет сделать вывод о том, что при поиске P- хеш-функций методом направленного перебора, численное значение коэффициента загрузки является наиболее весомым фактором, определяющим время поиска P- хеш-функции . Важным представляется также вывод о том, что уже при α=0.5, вероятность нахождения P-хеш-функции достаточно высока, так , что можно осуществить подбор P- хеш-функции приемлимое время.

При использовании хеш-адресации адрес, по которому происходит размещение ключа в памяти находится в определенной функциональной зависимости от кода ключа, так, что можно говорить о том, что с точки зрения теории информации, код хеш-адреса аккумулирует в себе часть информации, содержащейся в коде ключевого слова.

В общем случае, код ключевого слова содержит I_X бит информации. Если признать справедливым предположение о независимости значений (0 или 1) отдельных разрядов ключевого слова и о равновероятном появлении в них нулей и единиц , то количество информации, содержащейся в ключевом слове численно равно его разрядности- n . При хеш-преобразовании имеет место переобразование ключа X в хеш-адрес $A = H(X)$,где $H(X)$- функция хеш-преобразования, причем разрядность кода хеш-адреса равна $[\log_2 M]$.

Пусть количество информации о коде ключевого слова, передаваемое в процессе хеш-преобразования в код хеш-адреса равно I_A . Верхняя граница возможных значений I_A определяется разрядностью d хеш-адреса, хотя на практике, в большинстве случаев $I_A < d$. В силу того, что $I_A > 0$, то, представляется очевидным, что сохранение полного ключа в хеш-памяти является, в информационном плане, избыточным, так как часть информации о коде ключа, а именно I_A , уже содержится в коде адреса A по которому хранится ключ.

В свете изложенного выше, предлагается сохранять в памяти ключей не полный n-разрядный код ключа, а некоторый h-разрядный код S , который назван хеш-сверткой, являющийся результатом функционального преобразования $\Psi(X)$ над кодом X ключевого слова, при выполнении следующих двух условий:

- каждому возможному ключевому слову X однозначно соотвествует пара $\langle A_x , S_x \rangle$, где $A_x = H(X)$, $S_x = \Psi(X)$, то есть код хеш-свертки должен содержать количество информации I_S об исходном ключе, достаточное для того, чтобы выполнялось $I_S + I_A = I_X$;
- разрядность h хеш-свертки всегда меньше разрядности n ключевого слова, то есть $h < n$.

Разрядность h хеш-свертки определяется видом функций $H(X)$ и $\Psi(X)$.

Реализация возможностей, предоставляемых предлагаемым сокращением информационной избыточности может осуществляться в двух формах:

- уменьшение общего обьема хеш-памяти за счет уменьшения разрядности ячеек с n до h . Общий обьем памяти ключей при этом сок-

ращается в n/h раз. В частности, при выполнении условия $I_A = d$ объем хеш-памяти сокращается в $n/(n-[\log_2 M])$ раз. Так как число хранящихся ключей не меняется, то в этом варианте не меняется и коэффициент α загрузки, а, следовательно, и число возникающих коллизий;

- увеличение числа ячеек хеш-памяти с M до M' и сокращение их разрядности с n до h, при сохранении общего объема памяти: $M \cdot n = M' \cdot h$ и числа хранящихся ключей.

Следует указать на тот факт, что в этом случае величина разрядности h хеш-свертки, в общем случае, является функцией от нового значения M' числа ячеек хеш-памяти, так, что нахождение численных значений M' и h сопряжено с необходимостью решения системы уравнений, одно из которых задается условием сохранения неизменным общего объема памяти, а второе, задается функциональной зависимостью между h и M'. В частности, при выполнении условия $I_A = d$, разрядность h хеш-свертки определяется через M' в виде $h = n - [\log_2 M']$, и тогда численные значения величин h и M' определяются как решения следующей системы уравнений:

$$\begin{cases} M' \cdot h = n \cdot M \\ h = n - [\log_2 M'] \end{cases} \qquad (4)$$

Подстановкой второго уравнения системы в первое можно получить следующее уравнение:

$$F(M') = \frac{M'}{M}(n - [\log_2 M']) - n = 0 \qquad (5)$$

Это уравнение не может быть решено относительно M' в явном виде. Поэтому на практике следует использовать известные методы получения приближенного решения.

Коэффициент α загрузки памяти при этом уменьшается с исходного значения m/M до величины $m \cdot h / n \cdot M = \alpha \frac{h}{n}$. Соответственно уменьшается и среднее время поиска в хеш-памяти с коллизиями, а также число проб при выборе Р - хеш-алгоритмов.

При уменьшении объема хеш-памяти за счет хранения хеш-свертки имеет место эффект сжатия информации, то есть для сохранения m n-разрядных ключей общим объемом $m \cdot n$ бит, используется M' h- разрядных ячеек памяти, общим объемом $M' \cdot h$ бит, причем, при

определенных условиях может иметь место $M' \cdot k < m \cdot n$.

Таким образом, предлагаемый подход позволяет реализовать сжатие информации. В отличии от известных методов сжатия информации, основанных на оптимизации кодирования сохраняемых слов, предлагаемый метод базируется на сохранении части информации в виде позиций слов в организованной определенным образом их последовательности.

Для получения наибольшего эффекта от предлагаемого сохранения в хеш-памяти кода хеш-свертки вместо полного кода ключа большое значение имеет выбор алгоритма формирования хеш-адреса. Выбор алгоритма формирования хеш-свертки полностью зависит от принятого алгоритма формирования хеш-адреса, поэтому выбор последнего имеет определяющее значение.

Для эффективного использования предлагаемого способа исключения информационной избыточности в хеш-памяти выбор алгоритма формирования хеш-адреса должен выполняться исходя из следующих требований:

1. Используемый хеш-алгоритм должен обеспечивать высокий уровень эффективности и надежности формирования хеш-адреса. При этом под эффективностью хеш-алгоритма понимается мера близости генерируемой хеш-алгоритмом совокупности хеш-адресов к равномерному распределению. Надежность - это свойство хеш-алгоритма генерировать близкую к равномерно распределенной совокупность хеш-адресов при любых распределения входных ключей.

2. Хеш-алгоритм должен в минимальной степени использовать операции, связанные с потерей информации, с тем, чтобы генерируемый хеш-адрес при заданной разрядности содержал максимальное количество информации о исходном коде ключа.

3. Хеш-алгоритм должен в максимальной степени обеспечивать простоту и малое время, требуемое на его реализацию имеющимися процессорными средствами.

4. Хеш-алгоритм должен обеспечивать простоту и малое время реализации функционально от него зависимого алгоритма формирования хеш-свертки.

Следует указать на противоречивый характер приведенных требований, так, что выбор оптимального хеш-алгоритма всегда связан с принятием некоторого компромиссного варианта.

С точки зрения выполнения первых двух из указанных свойств наиболее предпочтительным является использование хеш-алгоритмов,

в основе которых лежит формирование каждого из d разрядов хеш-адреса $Z=\{z_1,z_2,\dots,z_d\}$ в виде суммы по модулю 2 некоторого подмножества Ω_i разрядов исходного ключа: $z_i = H_i(X) = \sum_{t\in\Omega_i} x_t \pmod 2$ $i=\overline{1,n}$, $\bigcup_{i=1}^{n}\Omega_i = X$. В работе доказано следующее положение: если задано d линейных функций $H_1(X), H_2(X), \dots, H_d(X)$ формирования хеш-адресов, то для любого n-разрядного ключевого слова $X=\{x_1, x_2, \dots, x_n\}$ всегда можно указать $h=n-d$ функций $\Psi_1(X)$, $\Psi_2(X), \dots, \Psi_h(X)$ хеш-свертки, обеспечивающих в совокупности со значениями $z_1, z_2, \dots, z_d$ разрядов хеш-адреса однозначность соответствия коду исходного ключа . При этом , всегда можно сформировать функции $\Phi_1(Z,S), \Phi_2(Z,S), \dots, \Phi_n(Z,S)$ восстановления таким образом, что $S_k = x_g$ то есть код хеш-сверки может быть представлен как подмножество разрядов исходного слова.

На основе приведенного утверждения выработана методика получения функций хеш-свертки и функций восстановления разрядных значений исходного ключа по коду хеш-адреса и хеш-свертки:

$$x_i = \Phi_i(z_1,\dots,z_d,S_1,\dots,S_h) = \sum_{\substack{\mathrm{mod}2\\ t\in\Omega_i}} z_t + \sum_{\substack{\mathrm{mod}2\\ q\in\Gamma_i}} S_q \,,\ i=\overline{1,n} \qquad (6)$$

Методика состоит в последовательном выполнении следующих пунктов:

1. Для заданного множества хеш-функций $H_1(X), \dots, H_d(X)$ формирования разрядов хеш-адреса и соответствующим им двоичным векторам определить множество ϑ функций, линейно зависящих от $H_1(X)$, $\dots, H_d(X)$ и множество Θ векторов, являющихся суммой по модулю 2 всевозможных подмножеств векторов $B_1, B_2, \dots, B_d$.

2. На множестве Θ найти вектор $B_r=\{b_{r1}, b_{r2}, \dots, b_{rn}\}$ содержащий минимальное число γ_r единиц, то есть $\forall\, B_\ell \in \Theta$, $\ell \neq r$ $\sum_{i=1}^{n} b_{\ell i} \geqslant \sum_{i=1}^{n} b_{ri} = \gamma_r$. В работе показано, что $\gamma_r \leqslant n-d+1 = h+1$ Вектор B_r можно, в общем случае, представить в виде поразрядной суммы по модулю 2 подмножества Δ_r векторов $B_r = \sum_{q\in\Delta_r} B_r$. Поэтому найденный вектор B_r соответствует уравнению вида:

$$b_{\ell 1}\cdot x_1 \oplus b_{r2}\cdot x_2 \oplus \dots \oplus b_{\ell n}\cdot x_n = x_{\ell_1} \oplus x_{\ell_2} \oplus \dots \oplus x_{\ell_{h+1}} = \sum_{\substack{\mathrm{mod}2\\ K\in\Delta_r}} z_K \qquad (7)$$
$$\ell_1, \ell_2, \dots, \ell_{h+1} \in \{\overline{1,n}\},\ \ell_1 < \ell_2 < \dots < \ell_{h+1}$$

3. Определить произвольным образом γ_r-1 переменных из входящих в уравнение (7) через соотвествующее количество функций формирования разрядов хеш-сверки в виде $S_q = \Psi_q(X) = x_{\ell_q}$, $q=\overline{1,\gamma_r-1}$, а оставшуюся переменную $x_{\ell_{h+1}}$, используя уравнение (7) представить в виде $x_{\ell_{h+1}} = \sum_{\substack{\mathrm{mod}2\\ q\in\Delta_r}} z_q + \sum_{\substack{\mathrm{mod}2\\ g=\overline{1,\gamma_r}}} S_g$. Таким образом определено γ_r+1 функций хеш-

свертки и γ_r+1 уравнений, составляющих систему (6). Последние определяют множество G разрядов ключа, определяемых через код хеш-свертки и код хеш-адреса.

4. Выделить на множестве Θ вектор, содержащий не более $h+2$ единиц, причем γ_r+1 единичные компоненты которого соответствуют переменным, входящим во множество G. Из уравнения соотвествующего выделенному вектору по способу изложенному выше в п.3 определить $h-\gamma_r+1$ переменных и добавить к множеству G.

Пункт 4 повторить до тех пор, пока множество G не будет включать в себя все n переменных $x_1, x_2, \ldots, x_n$, и система (6) функций восстановления ключа не будет определена полностью.

Необходимо отметить, что при использовании предложенной методики реализация функций формирования кода свертки не требует никаких аппаратных или временных затрат.

Использование пробинга, как средства разрешения коллизий эквивалентно последовательному применению ряда хеш-алгоритмов, каждый из которых является функцией начального хеш-адреса и номера шага пробинга. В работе доказано, что невозможно одновременное с пробингом изменять функции формирования хеш-свертки при условии обеспечения однозначности соотвествия хеш-адреса и хеш-свертки коду исходного ключевого слова.

Поэтому, при использовании хеш-свертки в хеш-памяти с коллизиями необходимым является применение дополнительной памяти в которую записываются ключи-синонимы.

Для уменьшения обьема дополнительной памяти синонимов, в которой, в отличии от основной хеш-памяти, следует сохранять полный код ключевого слова, предлагается совместно с применением хеш-свертки использовать ограниченный пробинг.

Ограниченный пробинг при этом рассматривается как обращение к основной хеш-памяти последовательно по хеш-адресам $H(X), H^1(X), \ldots, H^r(X)$, причем величина r ограничена предельным значением $t-1$. Если, при адресации основной хеш-памяти указанной выше последовательностью хеш-адресов не будет достигнут критерий поиска, то это соответствует ситуации сохранения искомого ключа в памяти переполнения. Уместным представляется заметить, что при введении понятия ограниченного пробинга хеш-поиск без пробинга можно рассматривать, как частный случай хеш-памяти с ограниченным пробингом при $t=1$. Очевидно, что при поиске с ограниченным пробингом, мак-

симальное время поиска в основной памяти не превышает времени, необходимого для r обращений к хеш-памяти.

При предлагаемом комбинированном использовании хеш-свертки и ограниченного пробинга для обеспечения однозначности соответствия хеш-адреса и хеш-свертки n-разрядному коду исходного ключа, в основной памяти необходимо сохранять номер $j \in \{\overline{0,t-1}\}$ использованной хеш-функции из множества $\{H_0(X), H^1(X), \ldots, H^{t-1}(X)\}$ функций пробинга. Число q дополнительных разрядов для хранения указанного номера составляет $q = [\log_2 t]$.

Практически важным, при использовании предложенной организации обработки коллизий, представляется определение необходимого обьема памяти переполнения.

В работе показано, что математическое ожидание $K_п$ и среднеквадратичное отклонение $\sigma_п$ числа ключей, попадающих в дополнительную память синонимов определяются выражениями:

$$K_п \approx \frac{\alpha t (m+1)}{t+1} + t \cdot \alpha, \quad \sigma_п \approx \frac{\alpha t}{t+1}\left(1 - \frac{\alpha t}{t+1}\right) \cdot m. \qquad (8)$$

Если задаться вероятностью $P_{пп}$ того, что ключ не может быть помещен в дополнительную память по причине ограниченного обьема последней, то численное значение $V_п$ можно, опираясь на теорему Муавра-Лапласса, определить из следующего выражения:

$$V_п = \frac{1}{2} \cdot \Psi(P_{пп}) \cdot \sqrt{\frac{\alpha t}{t+1}\left(1 - \frac{\alpha t}{t+1}\right) \cdot m} + \frac{\alpha t \cdot m}{t+1} \qquad (9)$$

где $\Psi(\ldots)$ - обратная функция плотности распределения Гаусса.

Разработаны структуры хеш-памяти с использованием хеш-свертки и ограниченного пробинга вместе с дополнительной памятью синонимов для разрешения коллизий. На рис. 1 приведена структура хеш-памяти, когда в качестве дополнительной памяти используется обычная память при том, в основной используется ограниченный пробинг, а на рис. 2 показана хеш-память без пробинга и с АЗУ - в качестве дополнительной памяти синонимов.

Основными узлами предложенных структур являются: формирователь основного хеш-адреса(ФОХА), формирователь дополнительного хеш-адреса(ФДХА), основная память(ОП) и память переполнения (ПП), счетчики пробинга(СП), формирователи адресов пробинга (ФАП), схемы совпадений (СС) и формирователь кода хеш-свертки (ФКХС).

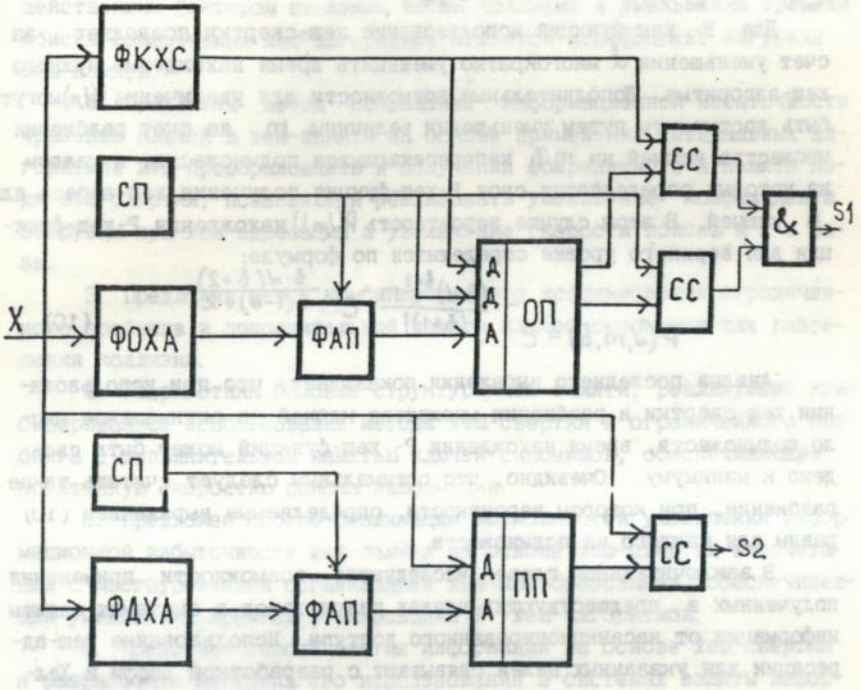
ФКХС
СП
ФОХА
ФАП
ДДА
ОП
СС
СС
&
S1
X
СП
ФДХА
ФАП
ДА
ПП
СС
S2

Рис.1

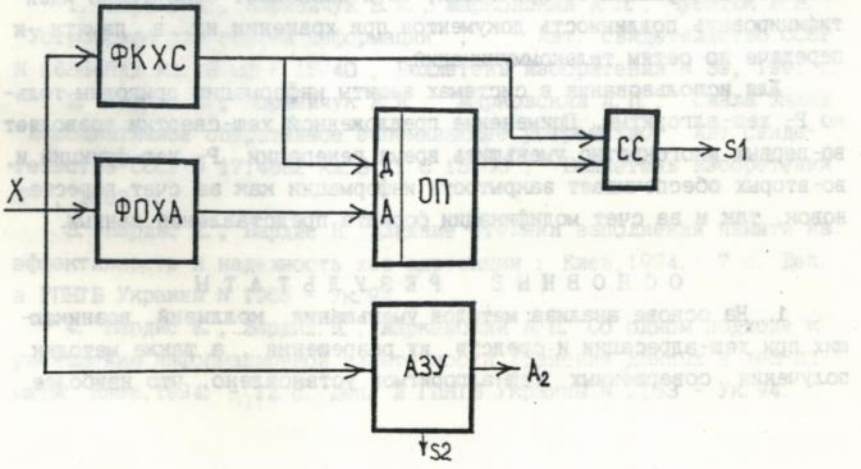
ФКХС
ФОХА
ДА
ОП
СС
S1
X
АЗУ
A2
S2

Рис.2

Для Р- хеш-функций использование хеш-свертки позволяет , за счет уменьшения α многократно уменьшить время нахождения нужного хеш-алгоритма. Дополнительные возможности для увеличения $P(\alpha,m)$ могут быть достигнуты путем уменьшении величины m за счет разбиения множества ключей на m/b непересекающихся подмножеств, в каждом из которых определяется своя Р-хеш-фукция получения хеш-адреса для b ключей. В этом случае, вероятность $P(\alpha,m,b)$ нахождения Р-хеш-функции для верхнего уровня определится по формуле:

$$P(\alpha,m,b) = e^{-\frac{(b\alpha)^{b+1}\cdot m}{(b+1)!}\cdot e^{-\frac{b\cdot\alpha(b+2)}{b(1-\alpha)+2}}} \qquad (10)$$

Анализ последнего выражения показывает, что при использовании хеш-свертки и разбиении множества ключей на оптимальное число подмножеств, время нахождения Р- хеш-функций может быть сведено к минимуму. Очевидно, что оптимальным следует считать такое разбиение, при котором вероятности, определяемые выражением (10) равны для каждого из подмножеств.

В заключительной главе исследуются возможности применения полученных в предшествующих главах результатов в системах защиты информации от несанкционированного доступа. Использование хеш-адресации для указанных целей связывают с разработкой Даффи и Хелманом во второй половине 70-х годов принципов защиты информации с открытым ключом. В системах, построенных на этом принципе хеш-преобразования играют доминирующую роль. Их использование не только защищает информацию от прочтения, но и позволяет эффективно идентифицировать подлинность документов при хранении их в памяти и передаче по сетям телекоммуникаций.

Для использования в системах защиты информации пригодны только Р- хеш-алгоритмы. Применение предложенной хеш-свертки позволяет во-первых многократно уменьшить время генерации Р- хеш-функции и, во-вторых обеспечивает закрытость информации как за счет перестановок, так и за счет модификации формата представления данных.

О С Н О В Н Ы Е Р Е З У Л Ь Т А Т Ы

1. На основе анализа методов уменьшения коллизий возникающих при хеш-адресации и средств их разрешения , а также методик получения совершенных хеш-алгоритмов установлено, что наиболее

действенным фактором снижения числа коллизий и уменьшения времени поиска совершенных хеш-алгоритмов является коэффициент загрузки хеш-памяти.

2. Предложен метод сокращения информационной избыточности хранения ключей в хеш-памяти на основе применения специальных алгоритмов хеш-преобразований и получения сохраняемого в памяти кода хеш-свертки, позволяющий реализовать уменьшение коэффициента загрузки при хеш-адресации и увеличение скорости поиска в 2-4 раза.

3. Предложен метод комбинированного использования ограниченного пробинга и дополнительной памяти ключей-синонимов для разрешения коллизий.

4. Разработаны базовые структуры хеш-памяти, реализующие комбинированное использование метода хеш-свертки и ограниченного пробинга с дополнительной памятью ключей-синонимов, обеспечивающие повышенную скоростью поиска информации.

5. Предложен способ реализации возможностей уменьшения информационной избыточности хеш-памяти на основе хеш-свертки в сочетании с многоуровневой организацией хеш-преобразования, обеспечивающий уменьшение времени формирования Р- хеш-алгоритмов.

6. Предложен способ сжатия информации на основе хеш-свертки и разработана методика его использования в системах защиты информации от несанкционированного доступа.

Основные результаты диссертации опубликованы в следующих работах:

1. Бардис Е., Корнейчук В.И., Марковский А.П., Чубатюк Ю.Н. "Устройство для поиска информации", Авт. свидетельство СССР N 16864641 кл. G 06 F 15/40, Бюллетень изобретений N 39, 1991 г.

2. Бардис Е., Корнейчук В.И., Марковский А.П., Сиала Халед "Ассоциативное оперативное запоминающее устройство" Авт. свидетельство СССР N 1714682 кл. G 11 C 15/00, Бюллетень изобретений N 7, 1992 г.

3. Бардис Е., Бардис Н. Влияние степени заполнения памяти на эффективность и надежность хеш-адресации : Киев,1994. - 7 с. Деп. в ГПНТБ Украины N 1965 - Ук.94.

4. Бардис Е., Бардис Н., Марковский А.П. Об одном подходе к уменьшению информационной избыточности хранения данных в хеш-памяти: Киев,1994. - 12 с. Деп. в ГПНТБ Украины N 2163 - Ук.94.

Бардис Евгениос

Методы и средства повышения эффективности хеш-адресации

Работой является рукопись на соискание ученой степени кандидата технических наук по специальности 05.13.08 - Вычислительные машины, системы и сети, элементы и устройства вычислительной техники и систем управления.

г. Киев, 1995 г.

Целью диссертации является разработка методов и средств повышения эффективности хеш-адресации, базирующихся на уменьшении информационной избыточности хранения данных в хеш-памяти. Для достижения поставленной цели в диссертации разработаны и исследованы способы сжатия информации в хеш-памяти, алгоритмы формирования хеш-адреса и хеш-свертки, структурная организация хеш-памяти с использованием сжатия данных и ограниченного пробинга с дополнительной памятью синонимов, схема многоуровневого P-хеширования для постоянных массивов ключей, использование хеш-функций и хеш-сверток для защиты информации.

Bardis Evgenios

Methods and means to riasing the effectiveness of hashing.

This scientific work is a manuscript to submit one's thesis for candidate's scientific degree in technical scienses in speciality 05.13.08 - Computers, system and network, elements and units of computer technique and control systems.

Kiev, 1995

The airm of the thesis is to develop methods and means for riasing the effectiveness of hashing which are based on decrease of information surplus data storage in hash-memory. For this aim achievement in the thesis a research have been develop of manners to data compression in hash-memory , algorithms for forming of hash adress and hash curdle , structural organization of hash-memory which based on data compression and limited probing with additional memory for key-synonym, composite perfect hashing scheme for static arrays of keys, applications of hash function and hash curdle for information security.

Ключові слова: пошук даних в пам'яті, хеш-адресація, організація файлів, стиснення інформації, архітектура пам'яті , асоціативна пам'ять, інформаційно-пошукові системи.