НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ УКРАИНЫ

" КИЕВСКИЙ ПОЛИТЕХНИЧЕСКИЙ ИНСТИТУТ "

На правах рукописи

АБУ-ХЕНДИ ХАСАН
(Бахрейн)

УДК 681.322.01

АЛГОРИТМИЧЕСКИЕ И СТРУКТУРНЫЕ СПОСОБЫ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ РАНДОМИЗАЦИОННЫХ МЕТОДОВ АССОЦИАТИВНОГО ДОСТУПА К ДАННЫМ

Специальность 05.13.08-Вычислительные машины, системы и сети, элементы и устройства вычислительной 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 рукопись.

Работа выполнена в Национальном техническом университете Украины на кафедре вычислительной техники.

Научный руководитель - кандидат технических наук, профессор Корнейчук Виктор Иванович

Официальные оппоненты - доктор технических наук, профессор Кузьмук Валерий Валентинович,

кандидат технических наук Селигей Александр Минович

Ведущая организация - Институт проблем регистрации информации Национальной Академии наук Украины.

Защита состоится 16.10.1995 г. в 14-30 на заседании специализированного Совета Д 01.02.06 в Национальном техническом университете Украины (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 Национального технического университета Украины.

Автореферат разослан " " сентября 1995 г.

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

А Н Н О Т А Ц И Я

Целью диссертационной работы является исследование факторов, влияющих на эффективность и надежность рандомизационных алгоритмов хеш-адресации и разработка методов получения хеш-алгоритмов, позволяющих уменьшить число коллизий за счет учета статистических характеристик динамических массивов ключей и реализовать perfect хеш-адресацию заданных постоянных массивов ключей, а также разработка структур реализации хеш-алгоритмов на аппаратном уровне.

Основные задачи диссертационной работы определяются поставленной целью и состоят в следующем:

1. Сравнительный анализ и классификация рандомизационных методов ассоциативного доступа с целью выработки общего критерия оценки качественных характеристик эффективности и надежности хеш-алгоритмов на этапе их синтеза.

2. Обзор и классификация методов получения perfect хеш-алгоритмов для заданных постоянных массивов ключей с целью определения возможностей повышения их эффективности.

3. Исследование свойств линейных функций применительно к возможностям их использования для минимальной perfect хеш-адресации.

4. Разработка способа ,методики и программных средств получения хеш-алгоритмов на основе линейных функций для минимальной perfect хеш-адресации заданных постоянных массивов ключей.

5. Разработка методики получения хеш-алгоритмов на основе линейных функций для perfect хеш-адресации заданных постоянных массивов ключей при неполном заполнении хеш-памяти.

6. Разработка способов получения эффективных хеш-алгоритмов для динамических массивов ключей с постоянными характеристиками статистического распределения.

7. Разработка структурно-функциональной организации специализированных процессорных средств реализации хеш-адресации на аппаратном уровне.

Автор выносит на защиту следующие основные положения и результаты :

1. Способ и методику получения perfect хеш-алгоритмов на основе линейных булевых функций для заданных постоянных мас-

сивов ключей при полном и неполном заполнении хеш-памяти.

2. Утверждения и следствия, полученные в результате исследования свойств линейных функций, положенные в основу способа использования функций указанного класса для минимальной perfect хеш-адресации.

3. Математические модели для вероятностного оценивания параметров процедур нахождения perfect хеш-алгоритмов на основе линейных функций.

4. Критерий аналитической оценки эффективности и надежности хеш-алгоритмов и методику его использования для синтеза алгоритма хеш-преобразования.

5. Способ формирования эффективных функций хеш-преобразования для динамических массивов ключей с постоянными характеристиками статистического распределения а также структуру устройств интервального поиска , основанных на использовании указанных хеш-функций.

6. Структурную организацию специализированных процессорных средств, реализующих на аппаратном уровне функции хеш-адресации.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы: Отличительными особенностями современного этапа развития систем обработки информации являются тенденция к увеличению обьемов хранимых данных , усложнение их структурной организации,ужесточение требований к оперативности доступа к ним. Использование традиционных методов поиска данных в больших информационных массивах приводит к существенному увеличению времени, затрачиваемого на эту часто встречающуюся операцию. С другой стороны,характерным является ускоренное развитие и расширение области применения систем обработки информации, в которых операции поиска по ключу занимают значительный удельный вес (до 70 %).

К таким системам, кроме традиционных систем баз данных, информационных и справочных систем , можно отнести системы распознавания образов, большинство которых базируется на использовании структурно-лингвистических методов, языковые процессоры, системы искусственного интеллекта. Характерным для указанных систем является большой обьем поисковых массивов и весьма жесткие требования к времени поиска.

Принципиально, наибольшую скорость доступа к данным по ключу обеспечивают рандомизационные методы (хеш-адресация), широко используемые в большинстве трасляторов, многих инфор - мационно-поисковых системах. Характерным для рандомизационных методов является независимость времени поиска от обьема поискового массива. Их главным недостатком является наличие коллизий, что имеет следствием необходимость в сложных процедурах их разрешения, увеличение максимального времени поиска, неполному использованию обьема накопителя. Следует отметить, что указанные недостатки усугубляются при увеличении обьема поискового массива.

Уменьшение числа коллизий ключей или их полное исключение (для постоянных массивов) может быть достигнуто за счет использования более совершенных и сложных хеш-алгоритмов, разработка которых является важной и актуальной задачей, решение которой определяет предмет исследования настоящей диссертационной работы.

ПРЕДМЕТОМ ИССЛЕДОВАНИЙ диссертационной работы являются способы и методы построения эффективных и надежных хеш-алгоритмов для динамических массивов ключей и perfect хеш-алгоритмов для заданных постоянных массивов ключей, а также структуры их реализации на аппаратном уровне.

МЕТОДЫ ИССЛЕДОВАНИЯ. В диссертационной работе использованы теоретические положения и методы теории множеств, комбинаторики, теории вероятностей и математической статистики , имитационного моделирования , теории цифровых вычислительных машин и систем, а также результаты статистического моделирования на ЭВМ.

НАУЧНАЯ НОВИЗНА. Предложены, теоретически и экспериментально обоснованы способ и методика получения perfect хеш-алгоритмов на основе линейных булевых функций для постоянных массивов ключей. Теоретически доказано существование алгоритмов указанного класса, обеспечивающих минимальную хеш-адресацию любого массива ключей, что отличает предложенный способ от известных. Предложен критерий оценки эффективности хеш-алгоритма на этапе его формирования. Разработаны новые структуры аппаратных средств реализации хеш-адресации. Предложен способ получения эффективного хеш-алгоритма для динамических массивов ключей с постоянной функцией распределения.

ПРАКТИЧЕСКАЯ ЦЕННОСТЬ работы состоит в разработке способов, методик и программных средств получения хеш-алгоритмов, исключающих коллизии для постоянных массивов ключей, а также оптимальных хеш-алгоритмов для массивов с постоянным законом распределения , которые могут быть использованы в быстродействующих лингвистических процессорах, системах распознавания образов, информационно-поисковых и справочных системах, системах обработки изображений при реализации баз данных и знаний.

ПУБЛИКАЦИИ. По теме диссертации опубликовано 3 печатные работы .

АПРОБАЦИЯ РАБОТЫ. Основные научные доклады диссертационной работы докладывались и обсуждались на Всесоюзной научно-технической школе "Устройства и системы хранения информации" (г.Алушта, 1991 г.), конференции AZIA BROWN BOBARY (г.Баден, Швейцария, 1992 г).

СТРУКТУРА И ОБЬЕМ РАБОТЫ.

Диссертационная работа состоит из введения , четырех глав и заключения, изложенных на 120 страницах машинописного текста, содержит список литературы из 93 наименований, 5 рисунков, 7-ми таблиц и приложения.

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

В первой главе приведена классификация алгоритмов хеш-преобразований, проведен критический обзор и сравнительный анализ методов получения perfect (совершенных) хеш-алгоритмов для заданных постоянных массивов ключей, позволившие определить направление исследований. Определены критерии эффективности и надежности рандимизационных методов ассоциативного доступа, предложен подход к аналитической оценке показателей качества хеш-алгоритма на этапе его разработки.

Во второй главе рассмотрены вопросы выбора эффективных хеш-алгоритмов для динамических массивов. Предложен способ формирования алгоритмов монотонных хеш-преобразований для динамических массивов с постоянным статистическим законом распределения. Разработана структура устройств для интервального ассоциативного доступа с использованием монотонных хеш-преобразований.

В третьей главе выполнено обоснование выбора в качестве функциональной основы хеш-преобразования линейных функций исследованы свойства указанного класса функций применительно к задачам perfect хеш-адресации при полном и неполном заполнении хеш-памяти. Разработаны математические модели для определения вероятностных характеристик времени нахождения perfect хеш-алгоритмов. Основным результатом указанных исследования является получение теоретических основ и математических моделей для разработки методики практического получения perfect хеш-алгоритмов на основе линейных функций.

В четвертой главе разрабатывается методика формирования хеш-алгоритмов для perfect хеш-адресации заданных постоянных массивов ключей при полном и неполном заполнении хеш-памяти, исследуются пути уменьшения времени формирования указанных хеш-алгоритмов, предлагаются структуры реализации их на аппаратном уровне, приводится описание программных средств их автоматизированного синтеза.

В заключении сформулированы основные результаты работы.

В приложении приведены листинги программ генерации perfect хеш-алгоритмов для постоянных таблиц ключей.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Важнейшей проблемой эффективного использования ассоциативного доступа в системах обработки данных является построение(или выбор) алгоритма хеш-адресации, т.е. оператора функционального отображения входного (заданного) множества ключей размещаемых записей на множество физических адресов памяти. Рандомизационные методы можно рассматривать как генераторы псевдослучайных чисел в заданном диапазоне адресов. Последние, без потери общности, можно интерпретировать как числа натурального ряда (1,M),где M-количество ячеек памяти. Главным недостатком рандомизационных методов доступа к данны является коллизия, когда для двух или более разных ключей генерируются одинаковые адреса(синонимы). Обработка коллизий требует дополнительных затрат времени.

В качестве критерия эффективности хеш-алгоритма принято использовать степень близости генерируемой им последовательности хеш-адресов к равномерному распределению, которое обеспечивает минимум возникающих коллизий.

Надежность-это свойство хеш-алгоритма сохранять заданный уровень эффективности при любом распределении ключей.

С 50-х годов предложено большое число различных методов генерации хеш-адресов. Однако практический выбор метода затруднен ввиду большого их разнообразия, а также наличием существенных расхождений в имеющихся рекомендациях и оценках Последнее обусловлено тем, что в ряде исследований за основу берутся статистические данные по обработке реальных файлов, выборка которых недостаточно представительна, кроме того, на точность оценок существенно влияла и различная глубина анализа алгоритмов. Для обоснованного выбора и анализа хеш-алгоритмов предложен достаточно общий подход, основанный на определении количества выполняемых при реализации алгоритма операций суммирования. Показано , что при суммировании k бит с вероятностью равенства единице $0.5+\xi$ каждого , вероятность равенства единицы суммы по модулю 2 будет равна $0.5 + 2^{k-1}\cdot\xi^k$ то есть будет ближе к 0.5 чем у слагаемых бит. Для получения бита с вероятностью $0.5+\Delta$ равного единице необходимо сформировать его как сумму по модулю 2 $]Ln(2\Delta)/Ln(2\xi)[$ бит. Получается, что эффективность алгоритма определяется числом суммирований. Такой подход позволяет при заданном значении Δ теоретически оценить пригодность того или иного хеш-алгоритма или на его основе построить алгоритм, удовлетворяющий заданному уровню эффективности.

Проведенный с изложенных позиций анализ наиболее распространенных хеш-алгоритмов показал, что теоретически наибольшей эффективностью обладают методы , использующие операцию деления. Этот вывод подтверждается экспериментальными исследованиями многих авторов.

В общем случае, число возникающих коллизий определяется хеш-алгоритмом и коэффициентом α заполнения хеш-памяти. Поэтому на практике величина α не превышает 0.7 - 0.8, что приводит к недоиспользованию значительного обьема хеш-памяти.

С конца 80-х годов для программной реализации ассоциативного доступа к постоянным или относительно редко изменяемым данным активно используется perfect (совершенная) хеш-адресация, при которой, за счет специальным образом формируемого для заданного массива ключей хеш-алгоритма достигается полное отсутствие коллизий.

При минимальной perfect хеш-адресации хеш-алгоритм должен обеспечивать отсутствие коллизий при полном использовании обьема памяти , то есть при $\alpha=1$. К настоящему времени предложено большое количество методов получения perfect хеш-алгоритмов. Проведенный их критический анализ показал , что все они в той или иной степени используют перебор на некотором подмножестве θ множества Ω всевозможных хеш-преобразований. Пусть, Θ -множество perfect хеш-алгоритмов для заданного массива ключей ($\Theta \subset \Omega$). Метод поиска perfect хеш-алгоритма должен, с одной стороны, обеспечивать выполнение условия $\Theta \cap \theta \neq \varnothing$, что требует расширения множества θ, а с другой стороны -уменьшения числа алгоритмов указанного множества с тем чтобы сократить время,затрачиваемого на их перебор. Для большинства методов формирования perfect хеш-алгоритмов требование сокращения времени формирования является доминирующим и условие $\Theta \cap \theta \neq \varnothing$ выполняется с некоторой вероятностью. Поэтому,важной и актуальной является задача разработка метода получения perfect хеш-алгоритмов, обеспечивающих уменьшение по сравнению с известными методами,времени формирования алгоритма при выполнение условия $\Theta \cap \theta \neq \varnothing$ для любых массивов ключей.

Следует указать и на тот факт, что при рассмотрении различных алгоритмов хеш-адресации существенную роль придавась времени их реализации на универсальном процессоре в ЭВМ в ущерб качественным показателям, что имело следствием ограничение класса исследуемых хеш-алгоритмов относительно простыми. Рост быстродействия современных процессоров, расширение их функциональных возможностей в сочетании со стабильной тенденцией к использованию высокопроизводительных специализированных процессоров является достаточным основанием для расширения класса пригодных для практической реализации хеш-алгоритмов ,прежде всего,за счет их разновидностей обеспечивающих высокую эффективность, но требующих сложных вычислений или плохо приспособленных к системам команд традиционных универсальных процессоров.

Для многих практически важных задач характерным является стабильность статистических характеристик массива ключей Для таких массивов можно указать функцию f(x) плотности вероятности. Если $\mathcal{Y}$ -множество ключей , то для каждого $a \in \mathcal{Y}$ можно сформировать R , принадлежащее множеству равномерно

распределенных чисел в интервале от [0,M] согласно формуле:

$$R = [M \int_{-\infty}^{a} f(x)\, dx] \qquad (1)$$

В частности, если множество ключей распределено по экэкспотенциальному закону, преобразование ключа a в хеш-адрес может выполняться по формуле:

$$\chi = [M \int_{0}^{a} \lambda e^{-\lambda x}\, dx] = [M(1 - e^{-\lambda a})]$$

Если задано значение ключа a, принадлежащего генеральной совокупности, которая распределена по нормальному закону распределения с матожиданием m_x и среднеквадратичным отклонением σ, то можно определить $a_н$, принадлежащее центрированному нормальному распределению с единичным значением σ:

$$a_н = (a - m_x)/\sigma$$

Тогда равномерно распределенный хеш-адрес может быть получен в соответствии со следующей формулой:

$$\xi = 1/\sqrt{(2\pi)} \int_{0}^{a_н} \exp(x^2/2)\, dx = 1/\sqrt{(2\pi)} \sum_{i=1}^{\infty} \frac{(-1)^{i-1}\ a_н^{2i-1}}{(2i-1)\ 2^{i-1}(i-1)!} \qquad (2)$$

$\chi = [M(0.5 + \xi)]$ при $a_н > 0$

$\chi = [M(0.5 - \xi)]$ при $a_н < 0$

Для определения числа элементов суммы было проведено статистическое моделирование. Для оценки соотвествия генерируемых хеш-адресов равномерному закону распределения использовался критерий χ^2. Проведенные экспериментальные исследования показывают, что в сумме число учитываемых членов может быть ограничено 3-4 -мя. Исследовались и другие приближения, которые могут быть использованы для формирования равномерно распределенных хеш-адресов при нормально распределенном массиве ключей. Показано, в частности, что для процессоров, в которых эффективно реализовано вычисление элементарных функций ξ приближенно может быть вычислено в виде: $\xi = 0.5 + \sin(a_н)/2$ при $a_н > 0$ и $\xi = 0.5 - \sin(a_н)$ при $a_н < 0$.

Характерной особенностью хеш-функций, формируемых с использованием формулы (1) является ее монотонность - то

есть: $a_1<a_2$, $a_1,a_2 \in \vartheta \Rightarrow \chi(a_1)<\chi(a_2)$. Эта особенность может быть использована для построения устройств интервального ассоциативного поиска, которые широко используются в системах обработки изображений для кластерного анализа и кодирования видеоизображений. Основой таких устройств является память таблиц пределов, в которой при заданном множестве $\Theta=\{p_1,\ldots,p_k\}$ для каждой пары $p_j,p_{j+1}\in\Theta$, $j<k$ по адресам $A_j : \chi(p_j)<=A_j<\chi(p_{j+1})$ записывается номер предела j. Тогда для отыскания предела в которым заключен поступающий на вход устройства код $x_{вх}$ достаточно вычислить $\chi(x_{вх})$ и произвести считывание кода искомого предела из памяти. При этом время, затрачиваемое на эту операцию существенно меньше, чем при традиционно применяемом дихотомическом поиске. Свойство монотонности эффективно может быть использовано и для реализации других сложных видов ассоциативного поиска. Для того, чтобы коллизии не нарушали упорядоченность ключей в хеш-памяти предложена процедура смещенного пробинга и средства для ее практической реализации. В качестве памяти переполнения может использоваться аппаратно реализуемая ассоциативная память. Это позволяет ограничить максимальное время поиска.

В третьей главе исследуются свойства. линейных булевых функций применительно к использовании их в качестве основы для получения perfect хеш-функций. Выбор линейных функций в качестве функционального базиса perfect хеш-преобразований обосновывается тем , что , как было показано исследованиями первой главы, суммирование по модулю 2 при равном числе операций обеспечивают наибольшее приближение генерируемой совокупности адресов к равномерному распределению. При использовании линейных булевых функций в качестве функционального базиса хеш-преобразований, хеш-адрес формируется как совокуп $k=[\log_2 M\ ]$ линейных функций от разрядных битов ключевых слов В принципе, $k=[\log_2 M]$ хеш-функций может быть получено реализацией k булевых функций от n переменных , определенных на m наборах. Однако , ключевым вопросом при этом будет сложность получаемых таким образом булевых функций. Многовариантность наборов указанных булевых функций ($2^k!$ вариантов) в сочетании со значительными затратами времени на минимизацию по каждому варианту , делают задачу отыскания достаточно простых совершенных хеш-функций перебором вариантов практически

неразрешимой.

Поэтому , следует ограничить класс исследуемых булевых функций, используемых для формирования совершенных хеш-функций, линейными булевыми функциями, которые отличаются простотой реализации и, вместе с тем, как было показано выше, минимизируют число коллизий для динамических массивов.

Для n переменных может быть сформировано 2^n-1 линейных булевых функций, причем, каждая из них L_i $i=1...2^n-1$ может быть представлена вектором $\{l_1,l_2,...,l_n\}$, $l \in \{0,1\}$. Каждая из линейных функций делит заданный набор ключевых слов на $\{X_k\}, k=1,...,m$, $X_k=\{x_1,x_2,...,x_n\}$ два подмножества: на ключевых словах, принадлежащих первому из них линейная булева функция принимает единичное значение, а на ключевых словах принадлежащих второму - функция принимает нулевое значение. При этом d из линейных функций делят заданный набор ключевых слов на две равные части и могут быть использованы в в качестве хеш-функций. Указанные линейные функции образуют подмножество Ω разделяющих функций.

Для анализа возможностей использования линейных функций в качестве основы построения совершенных хеш-алгоритмов приведем, без доказательства, следующие свойства линейных функций:

1. Для двух произвольно выбранных n-разрядных слов значения ровно $2^{n-1}-1$ линейных функций совпадают.
2. Существует не более 2^{n-2} n-разрядных слов для которых значение любой пары линейных функций отлично.
3. Для произвольного набора m n-разрядных ключевых слов всегда существует линейная функция, которая делит набор на две части, причем, число элементов в меньшей части не превышает $2^{n-2}-2(n-4)$.
4. Если линейная функция Φ принимает единичное значение на всех элементах множества A из 2^{n-1} n-разрядных слов, то для всех остальных 2^n-2 линейных функций справедливо:

$$\forall \Psi \neq \Phi: \sum_{\xi \in A} \Psi(\xi) = 2^{n-2}$$

Исходя из приведенных свойств линейных булевых функций можно доказать утверждения , непосредственно используемые при синтезе совершенных хеш-алгоритмов. К таким утверждениям отно-

сится, в частности , утверждение о том, что для любого множества $\mathfrak{I}$ состоящего из 2^{n-1} n-разрядных слов всегда существует линейная функция Φ, принимающая единичное значение ровно на половине его элементов, т.е.:

$$\exists\,\Phi:\sum_{\xi\in\mathfrak{I}}\Phi(\xi)=2^{n-2}$$

Доказательство: из свойства 3 следует,что для множества $\mathfrak{I}$ всегда существует функция Ψ , которая принимает единичное значение на не менее, чем половине элементов множества $\mathfrak{I}$, которые образуют множество Θ. Рассмотрим множество $A: A\subset\mathfrak{I}, \xi\in A\Rightarrow\Psi(\xi)=0$ и множество $B: B\cap\mathfrak{I}=\varnothing,\ \xi\in B\Rightarrow\Psi(\xi)=1$. Очевидно, что количество элементов рассматриваемых множеств A и B одинаково и не превышает $2^{n-1}-2(n-4)$.

Покажем, что всегда существует линейная функция Φ, делящая множества A и B в одинаковых отношениях т.е. так что

$$\sum_{\xi\in A}\Phi(\xi)=\sum_{\zeta\in B}\Phi(\zeta)$$

Для этого рассмотрим четыре n-разрядных слова $\zeta,\xi\in A$ и $\varepsilon,\rho\in B$. Согласно свойству 4 , число линейных функций, принимающих одинаковые значения на парах $\{\zeta,\varepsilon\},\{\zeta,\rho\},\{\xi,\varepsilon\},\{\xi,\rho\}$ равно $2^{n-1}-1$, а число линейных функций, кроме функции Ψ, примающих одинаковые значения на парах $\{\zeta,\xi\},\{\varepsilon,\rho\}$ равно $2^{n-1}-2$. Покажем , что обязательно существует линейная функция Φ, которая на парах $\{\zeta,\varepsilon\},\{\xi,\rho\}$ принимает одинаковые значения, т.е. $\Phi(\zeta)=\Phi(\varepsilon),\ \Phi(\xi)=\Phi(\rho)$. Доказать последнее несложно методом от противного: пусть из общего числа 2^n-2 линейных функций (не учитывая функции Ψ) некоторое подмножество Δ_1 состоящее из $2^{n-1}-1$ функций принимает одинаковое значение на паре $\{\zeta,\varepsilon\}$ и а некоторое подмножество Δ_2 из $2^{n-1}-1$ функций принимает одинаковое значение на паре $\{\xi,\rho\}$. Так, как высказано предположение о том, что не существует функции принимающей одинаковое значение на парах $\{\zeta,\varepsilon\},\{\xi,\rho\}$, то $\Delta_1\cap\Delta_2=\varnothing$, но пара слов $\{\zeta,\xi\}$ должна принимать одинаковое значение на $2^{n-1}-2$ функциях, то есть эти функции должны либо одновременно принадлежать множествам Δ_1 и Δ_2 (что невозможно в силу того , что $\Delta_1\cap\Delta_2=\varnothing$) либо одновременно не принадлежать ни к одному из указанных множеств, что также невозможно в силу того, что эти множества накрывают все множество из 2^{n-1} функций и , следовательно, множество линейных функций , одновременно не принадлежа-

щих Δ_1 и Δ_2 пусто . Таким образом , функция Φ, такая ,что на парах слов $\{\zeta,\varepsilon\}$ и $\{\xi,\rho\}$ принимает одинаковые значения с необходимостью существует. Так как при выборе слов $\zeta, \xi, \varepsilon, \rho$ не накладывалось никаких ограничений , то существование функции Φ, обладающей рассматриваемым свойством, является доказанным фактом для всех других точек множеств A, B, что возможно только в случае существования по крайней мере одной линейной булевой функции Φ, которая обладает свойством : $\Phi(\zeta)=\Phi(\varepsilon)$, для всех пар $\zeta\in A, \varepsilon\in B$, то есть делит множества A и B в одинаковых отношениях.

Согласно свойству 4 , любая линейная функция , отличная от Ψ , в том числе Φ , делит множество $\Theta = B \cup (\vartheta - A)$ на два подмножества с равным количеством элементов, а так как Φ делит множества A, B в одинаковых отношениях, то множество ϑ также делится линейной функцией Φ на два подмножества , содержащих одинаковое число элементов, что и требовалось доказать.

Из доказанного следует, что для набора из m n-разрядных ключевых слов ($m \angle = 2^{n-1}$) всегда можно сформировать дерево хеш-функций, обеспечивающих в совокупности perfect хеширование. По дереву хеш-функций можно построить комбинационную схему заказной БИС perfect хеш-адресации , состоящей из сумматоров по модулю 2 и мультиплексоров.

Достоинствами такого подхода является простота нахождения дерева хеш-функций, возможность обеспечения минимальной perfect хеш-адресации при $2^{n-2} < m <= 2^{n-1}$, что невозможно в общем случае при использовании других методов.

Большее быстродействие и простота реализации достигаются при использовании на каждом каскаде не одной, а нескольких разделяющих линейных функций, хотя это и сопряжено с существенно большей трудоемкостью нахождения подходящих линейных функций. В основе этого подхода лежат следующие утверждения:

1. Множество Θ из k независимых разделяющих линейных функций делит множество $m=2^{n-1}$ n-разрядных ключевых слов ровно на 2^k равных частей.

2. Всякая линейная функция, определяемая как сумма по модулю 2 произвольного подмножества множества Θ независимых разделяющих линейных функций является также разделяющей для множества ϑ m n-разрядных ключевых слов, т.е.

$$\forall \Psi = \underset{\text{mod } 2}{\Sigma} \Phi_i(x), \Phi_i(x)\in\Theta, i\in\{1,\ldots,k\} \Rightarrow \sum_{\xi\in\vartheta} \Psi(\xi) = m/2$$

Из приведенных утверждений следует, что для того, чтобы чтобы для множества $\mathfrak{I}$ m n-разрядных ключевых слов существовал однокаскадный алгоритм perfect кодирования необходимо и достаточно существование не менее k=] $\log_2 m$ [взаимно независимых линейных функций, образующих множество Θ. Поскольку в состав разделяющих для множества $\mathfrak{I}$ с необходимостью должны входить все функции, образованные суммированием по модулю 2 всевозможных подмножеств функций множества Θ, то общее число разделяющих линейных функций, необходимых для существования perfect хеш-алгоритма должно быть не менее R, определяемого формулой:

$$R = k + \sum_{i=2}^{k} C_k^i = \sum_{i=1}^{k} C_k^i = 2^k - 1, \text{ где } k =] \log_2 m [\qquad (3)$$

Определение набора из k НРЛФ для множества $m = 2^{n-1}$ n-разряд-ключей выполняется в следующей последовательности:

1. Генерируется ЛФ(путем перебора)
2. Проверяется, является ли сгенерированная ЛФ зависимой от ранее отобранных НРЛФ. Если ЛФ является зависимой, то выполнить переход на п.1
3. Проверяется, является ли сгенерированная ЛФ разделяющей для множества ключей. Если не является, то выполняется переход на п.1, иначе выбранная ЛФ присоединяется к формируемому набору НРЛФ. Если число найденных НРЛФ меньше k, то выполняется переход на п.1, иначе конец.

Среднее число R_c разделяющих функций определяется посредством формулы:

$$R_c = \frac{(2^n - 1)\,(C_{2^{n-1}}^{m/2})^2}{C_{2^n}^{m}} \qquad (3)$$

Вероятность P_K того, что выбранный набор из k разделяющих функций является эффективным определяется формулой:

$$P_K = \frac{2^{(k-1)2^{k-1}-k}}{(\sqrt{2\pi})^{2^k-k-1}} \qquad (4)$$

Число R_m НРЛФ составляющих эффективный набор определяется как минимальное k при котором выполняется условие:

$$C_R^k P_K - C_{2^k}^k < 3 \cdot \sqrt{C_{R_c}^k \cdot P_K} \qquad (5)$$

Заключительная глава посвящена разработке методики нахождения perfect хеш-алгоритмов для заданных постоянных массивов ключей при полном и неполном использовании обьема хеш-памяти. Указанный процесс может быть оптимизирован по различным критериям, такими, в частности, как минимизация времени нахождения perfect хеш-алгоритма или минимизация числа используемых линейных функций. При использовании второго крите - рия достигается максимальная скорость хеш-преобразования, что обеспечивается следующей методикой построения хеш-алгоритма:

1. Исходя из количества m ключей и их разрядности n по фор - муле 5 определяется максимальное число НРЛФ - $R_m(m,n)$. Если $R_m(m,n) >=]\log_2 m[$ то используется однокаскадная схема. В противном случае используется многокаскадная схема.
2. Для однокаскадной схемы находится $]\log_2 m[$ НРЛФ, образующих эффективный набор. Переход на конец.
3. Для многокаскадной схемы с использованием формулы 5 определяется минимальное число d каскадов при котором выполняется условие:

$$\Sigma R_m(w_i,n) =]\log_2 m[\ , \ w_1 = m, \ w_j = w_{j-1}/2^{R(w,n)}$$

Определяется число k_i НРЛФ на каждом каскаде: для $i = \overline{2,d}$:

$k_i = R_m(w_i,n)$, $k_1 =]\log_2 m[- \Sigma k_i$

Находятся НРЛФ, образующие эффективные наборы для каждого каскада.

На основании предложенных методик разработаны программы автоматизированного получения perfect хеш-алгоритмов.

Проведенный сравнительный анализ предложенного способа получения perfect хеш-алгоритмов показывает, что в отличии от известных способов он обладает следующими преимуществами: обеспечивает нахождение алгоритма минимальной perfect хеш-адресации при любом массиве ключей; требует меньшего времени получения указанного алгоритма; не накладывает жестких ограничений на количество ключей и их разрядность; за счет использования простого операционного базиса достигается большая скорость хеш-преобразования при аппаратной реализации последней; обеспечивает простоту реализации многоуровневой perfect хеш-адресации внешних ассоциативно адресуемых файлов.

О С Н О В Н Ы Е Р Е З У Л Ь Т А Т Ы

1. Выполнены классификация и анализ рандомизационных методов доступа к данным, предложен теоретически обоснованный критерий аналитической оценки эффективности и надежности хеш-алгоритмов на этапе их синтеза.

2. Проведен анализ работ, посвященных методам получения perfect хеш-адресации постоянных массивов ключей, обоснованы критерии эффективности этих методов, выявлены недостатки известных подходов к решению задачи получения perfect хеш-алгоритмов и определены пути их преодоления.

3.Теоретически исследованы свойства линейных булевых функций применительно к задачам минимальной perfect хеш-адресации. Сформулированы и доказаны утверждения и следствия,положенные в основу способа получения perfect хеш-алгоритма на базе линейных функций.

4. Предложены способ и методика получения perfect хеш-алгоритмов на основе линейных булевых функций для заданных постоянных массивов ключей при полном и неполном заполнении хеш-памяти.

5. Предложены структуры специализированных средств для аппаратной реализации функций perfect хеш-адресации на основе линейных булевых функций.

6. Предложены математические модели для вероятностного оценивания параметров процедур нахождения perfect хеш-алгоритмов на основе линейных функций.

7. Предложен способ формирования эффективных функций хеш-преобразований для динамических массивов ключей с заданным статистическим распределением, а также структура устройств интервального поиска, основанная на использовании этих функций

8. Разработан пакет программ для автоматизированного формирования perfect хеш-алгоритмов для постоянных массивов ключей.

РАБОТЫ, ОПУБЛИКОВАННЫЕ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Абу-Хенди Х., Зеебауэр М.,Сон Ки Ен. Ассоциативная память на основе комбинированного использования хеш-адресации и АЗУ небольшого обьема. Рукопись деп. в УкрНИИНТИ № 1009-Ук88. Деп. 25.04.88-1988-19 с.

2. Абу-Хенди Х., Марковский А.П., Хмельницкая Т.П. Ассоциативная память для логического упорядочения динамических информационных массивов. Вест.КПИ Автоматика и приборостроение 1990 - Вып.27 с.54-57.

3. Абу-Хенди Х., Корнейчук В.И. Начинающему пользователю IBM PC.- Киев.Диалектика 1995.- 64 с.

Абу-Хенди Хасан

Алгоритмические и структурные способы повышения эффективности рандомизационных методов ассоциативного доступа к данным.

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 исследование факторов, влияющих на эффективность и надежность хеш-адресации и разработка методов получения хеш-алгоритмов, позволяющих уменьшить число коллизий для динамических массивов и исключить появление коллизий для заданных постоянных массивов ключей, а также структур их реализации на аппаратном уровне. Для достижения поставленной цели, в диссертации исследованы свойства линейных булевых функций, доказаны утверждения и следствия, позволившие разработать способ получения минимальных perfect хеш-алгоритмов. Предложен способ получения оптимальной хеш-функции для динамических массивов с заданным распределением.

Abdulla Bu-Hind Hassan

Algoritmic and structure mode of increasing effectiveness randomize methods associative adressing of date.

This scientific work is a manuscript to submit one's thesis for candidate's scientific degree in tecnical sciense ses in speciality 05.13.08-Computers, system and network, elements and units for compute tecnique and control systems Kiev, 1995.

The aim of the thesis is investigation of factors which influence for effectiveness and reliability of haching and to develop methods for obtain hash-algoritm which can decrease collign number foe dynamic massive or wipe out collign for static massive and structure for realize them. For this aim achivement in the thesis a research have been investigated properties of linear boolian function, demonstration theorems and obtain conclusions permissable to develop method for obtain minimal perfect hash-algoritm. The techique for obtain optimal hash-function for dynamic vassive with prior distribution have been developed.

Ключові слова: пошук даних в пам'яті, хеш-адресація, рандомізаційні методи, асоціативна пам'ять.

КПИ. Проспект Победы,37.
Объем 1 уч. изд.л.
Зак. 456-100.