На правах рукописи

АРАПБАЕВ Русланбек Нурмаматович

АНАЛИЗ ЗАВИСИМОСТЕЙ ПО ДАННЫМ: ТЕСТЫ НА ЗАВИСИМОСТЬ И СТРАТЕГИИ ТЕСТИРОВАНИЯ

05.13.11 – математическое и программное обеспечение
вычислительных машин, комплексов и компьютерных сетей

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Новосибирск – 2008

Работа выполнена в Институте систем информатики
имени А. П. Ершова СО РАН

Научные руководители: Евстигнеев Владимир Анатольевич,
доктор физико-математических наук,
профессор.

Касьянов Виктор Николаевич,
доктор физико-математических наук,
профессор.

Официальные оппоненты: Малышкин Виктор Эммануилович
доктор технических наук, профессор.

Акжолов Маматжан Жолдошевич
кандидат физико-математических
наук, доцент.

Ведущая организация: Южный Федеральный Университет
(г. Ростов-на-Дону).

Защита состоится 12 декабря 2008 г. в15 ч 30 мин на заседании диссертационного совета К003.032.01 в Институте систем информатики им А.П. Ершова Сибирского отделения РАН по адресу:
630090, г. Новосибирск, пр. Лаврентьева, 6.

С диссертацией можно ознакомится в читальном зале ИСИ СО РАН (пр. Лаврентьева, 6)

Автореферат разослан 8 ноября 2008 г.

Ученый секретарь
диссертационного совета
К003.032.01

к.ф.-м.н. Мурзин Ф.А.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Развитие ЭВМ с параллельными архитектурами и высокопроизводительных вычислительных систем ставит перед программистами задачи по созданию новых технологических подходов и их эффективному использованию. В настоящее время успешно развиваются следующие основные направления для решения этой задачи: использование параллельных языков, использование библиотек и автоматическое распараллеливание программ. Первые два пути, несмотря на все их достоинства, оставляют в стороне возможность использования накопленного запаса пакетов прикладных программ, написанных на последовательных языках типа Фортран, а также не облегчают процесс написания параллельных программ. Остается третий путь – создание автоматических распараллеливающих компиляторов, обладающих способностью автоматически преобразовывать последовательную программу в параллельную, функционально эквивалентную, соответствующую заданному типу архитектуры программу.

Однако, разработка эффективных автоматических распараллеливающих компиляторов – это трудоемкий и достаточно длительный процесс. Основная их задача – извлечь как можно больше скрытого параллелизма из последовательной программы. Главным источником такого потенциального параллелизма, как правило, служит гнездо циклов. Извлечение скрытого параллелизма в первую очередь связано с анализом циклов и заключается в нахождении зависимости по данным между итерациями цикла. Таким образом, мощность автоматических распараллеливающих компиляторов весьма зависит от эффективности блока анализа зависимостей по данным.

Тем не менее, прямой подход к решению задачи выявления зависимостей в общем случае невозможен, так как даже для линейных индексных выражений массивов это приводит к NP-полной проблеме отыскания целочисленного решения системы диофантовых уравнений (уравнение зависимости). Один из способов строгого решения этой проблемы был предложен в 1976 г. Тоулем (Towel). Однако метод был слишком трудоемким, чтобы его можно было использовать на практике в распараллеливающих компиляторах. Позднее были разработаны быстрые приближенные методы, которые «ошибочно» предполагают существование решения уравнения зависимости. Конечно, использование таких некорректных предположений никогда не приводит к ошибочному объектному коду, но может мешать некоторым оптимизациям.

В последние годы интерес к этой тематике снова возрос, и были предложены более эффективные методы, которые получили название *тестов на зависимость* (data dependence test). Среди них на практике наибольшее распространение получили НОД-тест и тест на основе неравенства Банержи, специально разработанные Утополом Банержи (Banerjee).

Тесты на зависимость используют различные математические инструменты, и каждый из них имеет различную сложность и разрешающую способность. Мощные алгоритмы могут выявлять зависимости по данным с

большей точностью, но обычно требуют для этого много времени. Поэтому на практике используется алгоритм зависимости по данным, который состоит из серии тестов, исполняемых в определенном иерархическом порядке. Например, в проекте SUIF[1] алгоритм состоит из серии точных тестов, где последним тестом служит метод исключения Фурье-Моцкина. В распараллеливающем компиляторе Parafrase-2[2] используется стратегия применения НОД-теста и теста Банержи, а в системе ОРС[3] применяется тест Банержи-Вольфа, а также поддерживается идея полуавтоматического распараллеливания. Однако до сих пор остается открытым вопрос, **какая последовательность или стратегия лучшая**.

К настоящему времени разработано множество тестов на зависимость, дающих приближенные и точные решения задачи анализа зависимости по данным, что открывает новые возможности. В связи с этим особую актуальность приобретает выработка новых стратегий тестирования для выявления зависимостей по данным, в которых алгоритм стратегии должен быть эффективным при применении на практике, т.е. выбрать “золотую середину” между точностью и использованием ресурсов.

Поэтому в рамках диссертационной работы была предпринята попытка расширить, обобщить и развить существующие подходы с целью преодоления упомянутых выше ограничений.

Все вышесказанное говорит об актуальности проводимых исследований.

Цель работы. Целью диссертационной работы является разработка новых и улучшение имеющихся алгоритмов для анализа зависимостей по данным при распараллеливании и оптимизации последовательных программ.

Достижение цели связано с решением следующих задач:

- Исследование существующих тестов на зависимость и сопоставление их сильных и слабых сторон;
- Разработка новых эффективных тестов для анализа зависимостей по данным, в том числе для анализа ссылок многомерных массивов;
- Реализация библиотеки тестов на зависимость по данным;
- Выработка новой стратегии тестирования для анализа зависимостей по данным;
- Проведение экспериментов, подтверждающих корректность и эффективность предложенных методов.

Методы исследования. В диссертационной работе использовались различные методы и математические инструменты такие, как: теория графов, теория алгоритмов, элементы теории множеств, теории чисел, методы интервального анализа, методы линейного и целочисленного программирования, теория преобразования и оптимизация программ и др.

[1] Система разработана в Стэндфордском университете под руководством M. Lam

[2] Проект разработан в Иллинойском университете под руководством C. Polychronopoulos

[3] Открытая распараллеливающая система разрабатывается в Ростовском государственном университете под руководством Б. Я. Штейнберга.

Научная новизна. Проведены исследования, направленные на изучение применимости различных тестов для выявления зависимостей по данным. Даны сопоставления сильных и слабых сторон тестов, как на примерах, так и по оцениваемым характеристикам отдельных критериев.

Предложен модифицированный эффективный тест для решения проблемы зависимости по данным при анализе ссылок многомерных массивов. Новый модифицированный метод, в отличие от известных способов, позволяет получить ответ о существовании целочисленных решений уравнений зависимости при выявлении зависимости по данным в многомерных массивах, содержащих сцепленные индексы.

Реализована библиотека из новых и модифицированных тестов на зависимость по данным, в состав которой вошли приближенные и точные тесты, рассматривающие одномерные и многомерные случаи.

Выработана новая стратегия тестирования, основанная на новых тестах анализа зависимостей по данным. При построении стратегии использованы факты и результаты некоторых эмпирических и теоретических исследований анализа зависимостей по данным, позволившие оптимизировать общее время выполнения алгоритма новой стратегии. На основе новой стратегии и библиотеки тестов на зависимость создан программный комплекс анализа зависимостей по данным, а также построен алгоритм для индексного анализа зависимости по данным в Sisal-программах в рамках системы функционального программирования (SFP).

Проведены экспериментальные работы для сравнения эффективности предложенных подходов с аналогичными методами анализа зависимостей по данным.

Практическая ценность. Полученные результаты являются неотъемлемой частью системы быстрого прототипирования распараллеливающего компилятора и системы функционального программирования (SFP), разрабатываемых в рамках проекта ПРОГРЕСС. Результаты могут быть использованы при решении практических задач, а именно при разработке распараллеливающих компиляторов. В частности, разработанные автором диссертации методы могут стать основой для построения алгоритмов выявления зависимости по данным между итерациями DO-циклов в блоке зависимостей в системе быстрого прототипирования распараллеливающего компилятора.

Программно реализованные разработки могут использоваться в качестве инструмента для изучения свойств последовательных программ в процессе написания параллельных программ, а также при проведении обучения студентов методам программирования и оптимизации для параллельных архитектур.

Апробация работы. Основные положения диссертации обсуждались на следующих конференциях и семинарах.

1. Международная научная конференция "Параллельные вычислительные технологии" (ПаВТ'2007), Челябинск, Россия, 2007 г.
2. VI Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям (с участием иностранных ученых), Кемерово, 2005.
3. IV Российско-Германская школа по параллельным вычислениям на высокопроизводительных вычислительных системах, Новосибирск, ИВТ СО РАН, 2007.
4. Конференция-конкурс «Технологии Microsoft в теории и практике программирования», Новосибирск, 2006 г.
5. VII Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям (с участием иностранных ученых), Красноярск, 2006.
6. Конференция-конкурс «Технологии Microsoft в теории и практике программирования», Новосибирск, 2008 г.
7. XLIV Международная научная студенческая конференция «Студент и научно-технический прогресс», Новосибирск, 2006 г.
8. Семинары «Конструирование и оптимизация программ», Новосибирск, ИСИ СО РАН, 2003-2008 гг.

Публикации. Основные результаты диссертационной работы опубликованы в 12 работах, среди которых 4 статьи, 1 препринт и 7 тезисов докладов.

Исследования выполнялись в соответствии с планами научно-исследовательских работ ИСИ СО РАН по проекту 3.15 «Методы и средства трансляции и конструирования программ» программы 3.1 СО РАН «Информационное и математическое моделирование в различных областях знаний, задачи поддержки принятия решений, экспертные системы, системное и теоретическое программирование» и частично поддерживались грантом РФФИ (№ 07-07-12050).

Структура диссертации

Диссертационная работа состоит из введения, трех глав и списка литературы. Объем диссертации – 116 стр. Список литературы содержит 109 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность проводимых исследований, сформулирована цель диссертационной работы, показана новизна и практическая значимость результатов, указаны положения, выносимые на защиту, и кратко аннотировано содержание глав.

В главе 1 содержится сравнительный обзор существующих тестов на зависимость по данным, применяемых в распараллеливающих компиляторах. Даны сопоставления сильных и слабых сторон тестов, как на примерах, так и по оцениваемым характеристикам отдельных критериев. Также рассматривается

некоторые проблемы анализа зависимостей по данным для многомерных массивов.

В разделе 1.1. описан ряд определений анализа зависимостей по данным, которые требуются в дальнейшем.

Традиционно под зависимостью по данным понимается зависимость, связанная с совпадением ссылок на элементы массивов. Пусть в гнезде r вложенных DO-циклов, операторы $\mathbf{S_1}$ и $\mathbf{S_2}$ обращаются к d - мерному массиву $\boldsymbol{A}$ индексные выражения которого представлены линейными функциями $h_q(i_1, i_2, \ldots, i_r)$и $g_q(i_1, i_2, \ldots, i_r)$, где $q = 1, \ldots, d$. Зависимость по данным между операторами $\mathbf{S_1}$ и $\mathbf{S_2}$ имеется тогда и только тогда, когда существуют целые $i_1, i_2, \ldots, i_r$ и $j_1, j_2, \ldots, j_r$, такие что

$$\begin{aligned} &h_1(i_1, i_2, \ldots, i_r) = g_1(j_1, j_2, \ldots, j_r),\\ &h_2(i_1, i_2, \ldots, i_r) = g_2(j_1, j_2, \ldots, j_r),\\ &\ldots\\ &h_d(i_1, i_2, \ldots, i_r) = g_d(j_1, j_2, \ldots, j_r) \end{aligned} \qquad (1)$$

и

$$i_p, j_p \in [L_p, U_p], \qquad (2)$$

где p = 1, . . . , r.

Следовательно, проблема зависимости данным представляет собой задачу целочисленного программирования.

Если имеется m-мерный массив $\boldsymbol{A}$ и индексные выражения массива линейны, то тогда система (1) может быть записана в следующем виде:

$$\begin{aligned} &a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n + c_1 = 0\\ &a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n + c_2 = 0\\ &\qquad\qquad \ldots\ldots\ldots\\ &a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n + c_m = 0 \end{aligned} \qquad (3)$$

и

$$L_i \le x_i \le U_i \quad \text{где } i=1,\ldots,n. \qquad (4)$$

В разделах 1.2. и 1.3. приводится сравнительный анализ тестов на зависимость. Также приведена обширная классификация тестов на зависимость по использованию различных математических инструментов и по сложности применения их на практике. В частности сначала описываются три основных метода нахождения зависимости по данным: НОД-тест, тест Банержи и метод исключения переменных Фурье-Моцкина. Далее рассматриваются расширенные варианты предыдущих методов, которые делятся на два вида: приближенные тесты и точные тесты. К приближенным тестам относятся обобщенный НОД-тест, λ-тест, I-тест и их различные модификации. Точные тесты используют более сложные методы: Power-тест, Омега-тест и IR-тест.

В разделе 1.4. отмечены другие тесты на зависимость, которые в рамках

данной диссертационной работе не исследованы. К ним относится СМ-тест, использующий симплекс-метод, модифицированный для решения задач целочисленного программирования, алгоритм Шостака, основанный на использовании теоретико-графового метода для решения систем линейных неравенств, и др.

Раздел 1.5. представляет новый модифицированный λ-тест для анализа ссылок в многомерных массивах. Сравнительный анализ показал, что при анализе зависимостей в цикле ключевой проблемой является работа с многомерными массивами. Общий подход состоит в индивидуальном тестировании уравнений (тестирование "индекс-за-индексом") из (3) вместо проверки существования решения системы в целом. Однако система уравнений зависимости может не иметь решения даже в том случае, когда имеются решения в каждом из отдельных уравнений.

При анализе многомерных массивов основную трудность вызывают часто встречающиеся в реальных программах *сцепленные* индексы. Если потенциальная зависимость включает сцепленные индексы, то для её разрушения необходимо одновременное рассмотрение индексов многомерного массива. Один из таких эффективных тестов анализа зависимости в многомерных массивах, включающих сцепленные индексы, является λ-тест. Однако λ-тест определяет, имеет ли система действительные решения, но не может дать точного ответа о существовании целочисленных решений системы.

В предложенном новом модифицированном варианте λ-теста λ-тест интегрирован с точным IR-тестом, благодаря чему при анализе зависимостей многомерных массивов были получены более точные результаты. IR-тест находит целочисленные решения уравнения зависимости путем сокращения интервала решений переменных с многократным проецированием. Как только эффективный интервал решений какой-нибудь переменной сжимается к пустому, то делается вывод, что линейное уравнение не имеет целочисленного решения.

Геометрически каждое линейное уравнение системы (3) представляет собой гиперплоскость π. в пространстве $\mathbf{R^n}$. Пересечение гиперплоскостей $\boldsymbol{S}$ соответствует общим решениям системы уравнений. Границы циклов (4) соответствуют ограниченному выпуклому множеству **V** в $\mathbf{R^n}$. Если можно найти новую гиперплоскость, которая содержит $\boldsymbol{S}$, и не имеет целочисленных точек пересечения с **V**, то это доказывает, что $\boldsymbol{S}$ не имеет целочисленных значений в **V**, следовательно, зависимости по данным не существует. Данная гиперплоскость является линейной комбинацией уравнений системы.

Теорема 1. $\mathbf{S} \cap \mathbf{V}$ не имеет целочисленных точек тогда и только тогда, когда существует гиперплоскость π, соответствующая линейной комбинации $\left\langle \sum_{i=1}^{m} \lambda_i \vec{a}_i, \vec{x} \right\rangle + \sum_{i=1}^{m} \lambda_i c_i = 0$ системы уравнений (3), такая, что $\pi \cap \mathbf{V}$ не имеет целочисленных точек, где $< \vec{a}_i, \vec{x} >$ – скалярное произведение $\vec{a}_i \equiv (a_{i1}\ a_{i2}\ ,\ldots,\ a_{in})$

и $\vec{x} \equiv (x_1, x_2, \ldots x_n)$.

Предложенный алгоритм с помощью λ-теста генерирует множество линейных комбинаций гиперплоскостей. Затем применяется IR–тест для нахождения гиперплоскости из данного множества, которая не имеет целочисленных точек пересечения с V. Экспериментальные сравнения результатов показали, что модифицированный алгоритм более точен по сравнению с НОД-, Банержи- и λ-тестами на зависимость. Таким образом, модифицированный λ-тест является точным тестом для выявления зависимости по данным в многомерных массивах, содержащих сцепленные индексы.

Вторая глава диссертации посвящена подробному описанию новой выработанной стратегии применения тестов для выявления зависимости по данным, в которой алгоритм состоит из серии эффективных и недорогостоящих тестов на зависимость, имеющих линейную и полиномиальную сложность. При выработке стратегии учтены результаты некоторых эмпирических исследований, а также некоторые ограничения аналогичных работ.

В разделе 2.1. рассматриваются существующие стратегии тестирования анализа зависимостей по данным. Приведется теоретическое и практическое описание Дельта-теста, Эпсилон-теста, алгоритма Майдана и К-теста. Изучаются основные ограничения перечисленных алгоритмов, а также статистические данные существующих эмпирических исследований.

Исследование алгоритмов на зависимость выявило ряд работ в некотором отношении аналогичных представленной диссертационной работе. Во всех исследованиях были представлены наборы тестов на зависимость (библиотека тестов на зависимость), с целью их использования для точного и эффективного решения проблемы в практических ситуациях. Однако наша работа по существу отличается отних.

Дельта–тест разработан для определенных классов ссылок массива, которые часто встречаются в научных программных кодах. Тест сначала классифицирует индексные выражения массивов на следующие категории: ZIV (нулевая индексная переменная), SIV (единственная индексная переменная) и MIV (составная индексная переменная) формы. Соответственно к каждой форме применяются одноименные тесты.

В. Пью (W. Pugh) и Т. Шпейзман (T. Shpeisman) предложили более упрощенный и быстрый вариант Дельта-теста, называемый *Эпсилон – тестом.* В этом тесте рассмотрены только самые простые случаи индексных выражений SIV, не используются сцепленные MIV формы и НОД–тест, а также не рассматриваются треугольные границы цикла при использовании теста Банержи. Хотя эти алгоритмы являются самыми быстрыми, но они уступают по точности предложенному в данной диссертационной работе алгоритму.

Алгоритм Майдана который использован в системе SUIF Стенфордского университета, состоит из серии точных тестов, каждый из которых применим в ограниченной области. Последний тест в алгоритме – метод исключения Фурье–Моцкина, расширенный для решения целочисленных задач. Авторы

показали, что практически зависимость по данным может быть вычислена точно и эффективно. Главное различие между алгоритмом Майдана и предложенным подходом – в том, что в первом случае добивались требуемого результата с использованием дорогих методов. В противоположность этому, наш подход пытается получить те же результаты с использованием более дешевых тестов на зависимость.

K-тест также состоит из библиотеки тестов на зависимость, но, в отличие от других подходов, вместо конкретной стратеги применения тестов, используются методы искусственного интеллекта. Хотя в самой работе также упоминается о NP-полноте методов искусственного интеллекта.

Раздел 2.2. начинается кратким обзором тестов на зависимость, используемых в предложенной новой стратегии тестирования. Тесты можно классифицировать на одномерные и многомерные. Одномерные тесты: SIV-тест, НОД-тест, тест Банержи, I-тест и IR-тест. Многомерные тесты: λ-тест, многомерный I-тест,и модифицированный λ-тест. Идея нашей стратегии опирается на следующие научные факты и результаты.

Основные результаты существующих эмпирических исследований. Объектом автоматического распараллеливания служат большие пакеты научных прикладных программ, написанных на последовательных языках типа Фортран. Согласно эмпирическому изучению Шена (Shen) и др., в реальных программах индексные выражения не очень сложны. Из всех исследованных массивов примерно 56% составляют ссылки одномерных массивов и 36% – ссылки двухмерных массивов. Доля ссылок трехмерных (и выше) массивов составляет около 8 %. Что касается индексных выражений массивов, то 53 % являются линейными, 13 % – частично линейными и 34 % – нелинейными. Поэтому обычно для анализа зависимостей по данным на практике применяются только одномерные тесты, использующие подход тестирования “индекс-за-индексом”.

При анализе многомерных массивов основную трудность вызывают часто встречающиеся в реальных программах сцепленные индексы. Как показано в эмпирических исследованиях, более чем в девяти тысячах пар двухмерных ссылок массивов приблизительно 46% являются сцепленными индексными выражениями. Что касается ссылок массивов большей размерности, то только 2% являются сцепленными индексными выражениями. Поэтому на практике важно иметь эффективный тест для обработки сцепленных индексов, особенно для анализа ссылок двухмерных массивов.

Случаи повышающие точность тестов на зависимость. Точно определяющие методы: Омега-тест, Power-тест, алгоритм Майдана и др. используют линейные и целочисленные методы для решения диофантовых уравнений, например, метод Фурье-Моцкина, Симплекс метод и др., которые не эффективны на практике. В экспериментальных результатах Р. Триоле (Triolet R.) показано, что по сравнению с более простыми методами, метод исключения переменных Фурье-Моцкина выполняется в 22-28 раз дольше.

Одним из стандартных и распространенных тестов на зависимость является тест Банержи. Он является приближенным тестом и принимает во

внимание границы циклов. Эффективность и полноценность теста Банержи при опровержении зависимостей, делают его одним из самых используемых тестов в распараллеливающих компиляторах. В исследованиях Банержи, З. Ли (Z. Li) и Д. Клапхольц (D. Klappholz) показано, что если коэффициенты линейного уравнения удовлетворяют некоторым условиям, то тест Банержи становится точным. Банержи показал, что его неравенства точны, если все коэффициенты индексных переменных равны 1, 0, или -1. Ли и др. показали, что неравенства Банержи точны, если коэффициент одной индексной переменной $|a_k|=1$ и $|a_i| \le (U_i - L_i)$, где $i=1,\ldots,k-1,k+1,\ldots,n$.

Клапхольц и др. доказали, что неравенства Банержи точны, если после упорядочения коэффициентов индексных переменных $|a_1| \le |a_2| \le \ldots \le |a_n|$, коэффициент индексной переменной $|a_1|=1$ и для каждого j выполняется условие $|a_j| \le 1 + \sum_{k=1}^{j-1} |a_k| (U_k - L_k)$, $2 \le j \le n$.

Алгоритм стратегии. Учитывая все вышеприведенные факты и результаты, в настоящей работе предлагается новая стратегия применения тестов для выявления зависимости по данным, в которой алгоритм состоит из серии эффективных и недорогостоящих тестов на зависимость.

В данной стратегии, в зависимости от значений основных параметров задачи (размерность массивов, количество вложенных циклов, значения коэффициентов индексных переменных и значения границ циклов), сначала были выделены часто встречающие и легко разрешимые случаи. Соответственно каждому случаю применяется один быстрый и точный тест или серия эффективных тестов.

На вход алгоритма подается гнездо цикла, в котором **r** – количество вложенных циклов, и операторы цикла обращаются к элементам **d**-мерного массива. Кроме того, считаются постоянными и известными значения коэффициентов индексных переменных a_{11}, a_{12} ,…, a_{mn} и значения границ циклов $L_1, L_2, \ldots, L_n, U_1, U_2, \ldots, U_n$, где $n=2*r$ и $m=d$. Задача нашего алгоритма – выявить зависимости по данным между операторами в итерациях гнезда циклов, т.е. алгоритм должен возвращать ответ «да/нет» о существовании целочисленных решений $i_1, i_2 \ldots, i_n$ системы линейных диофантовых уравнений (3), удовлетворяющих ограничениям (4).

Перечислим часто встречающиеся и легко разрешимые случаи задачи зависимости по данным:

1. ***r***=1, ***d***=1, т.е. внутри единственного цикла операторы обращаются к элементам одномерного массива. В этом случае уравнение зависимости (3) выглядит так: $a_1 x_1 + a_2 x_2 = a_0$ и $L \le x_1, x_2 \le U$. Для уравнения целесообразно применить самый быстрый и точный ***SIV-тест***.

2. ***r***>1, ***d***=1, уравнение зависимости имеет вид $a_1 x_1 + a_2 x_2 + \ldots + a_n x_n = a_0$, где $L_i \le x_i \le U_i$, $i=1,\ldots,n$. Этот случай несколько усложняет решение, поэтому применяется серия одномерных тестов на зависимость: тест Банержи, I-тест и IR-тест. Каждый следующий тест выполняется только

в том случае, если предыдущим тестом был получен неточный ответ (maybe), кроме того, после применения теста Банержи выполняется проверка коэффициентов индексных переменных для уточнения ответов теста.

3. ***d***=2 и имеются *сцепленные индексы*. Система уравнений зависимости имеет вид

$a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n = a_{1,0}$

$a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n = a_{2,0}$

и $L_i \le x_i \le U_i$ где $i=1,\ldots,n$.

Этот случай доминирует в реальных последовательных программах, но применение обычных одномерных тестов на зависимость в этом случае бесполезно, так как имеются сцепленные индексные переменные. Поэтому применяется серия многомерных тестов: λ-тест, многомерный I-тест и модифицированный λ-тест. Метод запоминания результатов предыдущих тестов и использования их для последующих тестов оптимизирует данный случай.

4. В оставшихся случаях уравнение зависимости имеет вид (3) с ограничениями (4). Каждое уравнение рассматривается в отдельности, и для него последовательно применяется серия одномерных тестов: тест Банержи, I-тест и IR-тест. Этот подход дает более точный ответ, если индексные переменные *не сцеплены.* На практике доля сцепленных индексных переменных в ссылках трехмерных массивов и выше незначительна.

Учитывая все случаи, была собрана и реализована библиотека тестов на зависимость. Библиотека состоит из следующих тестов: ZIV-тест, SIV-тест, НОД-тест, Банержи-тест, I-тест, IR-тест, λ-тест, многомерный I-тест и модифицированный λ-тест. Кроме тестов на зависимость в библиотеке имеются алгоритмы для уточнения ответов теста Банержи. Все алгоритмы имеют линейную временную сложность. Из-за высокой стоимости в библиотеку не вошли точные тесты. На рис. 1 приведена общая схема новой стратегии применения тестов на зависимость.

Временная сложность. Новый алгоритм имеет наихудшую временную сложность в двух случаях:

1) при применении серии одномерных тестов на зависимость;

2) при применении серии двухмерных тестов на зависимость.

Рассмотрим отдельно каждый случай. В первом случае тест Банержи имеет временную сложность O(n), I-тест – O(n^2*c+n*c) и IR-тест – O(kn), где $k=\min\{u_i-l_i+1: 1\le i\le n\}$, n – количество переменных в уравнении зависимости, c – константа. В общем случае, где серия одномерных тестов используется для многомерных массивов с подходом тестирования «индекс-за-индексом»,

наихудшая временная сложность – O(d*(n^2*c+n*c + n + k*n + c)), где ***d*** – размерность массива.

Во втором случае используются более сложные механизмы, но в нашем алгоритме они применяются только для двухмерных массивов, где индексы являются сцепленными. Следовательно, трудоемкость методов несколько уменьшается. Например, λ-тест имеет временную сложность O(n*(z+c)), многомерный I-тест – O(n*(z^2*c+z*c+c)) и модифицированный λ-тест – O(*n*(k*z+*c*)), где z – число переменных в сгенерированной λ-плоскости. В общем случае временная сложность – O(n*(z^2*c+z*c+ z+k*z+c)).

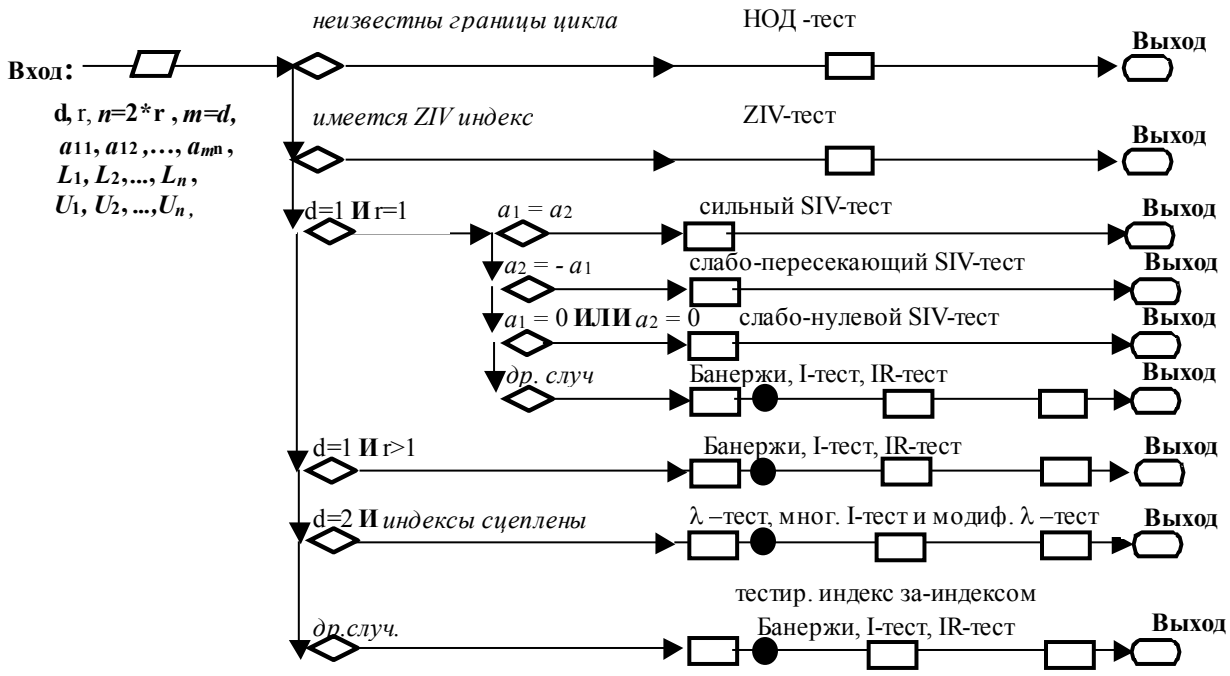

Рис. 1 Общая схема новой стратегии тестирования для выявления зависимости по данным. ● - алгоритм проверки коэффициентов, □ - тесты на зависимость

Третья глава посвящена анализу экспериментальных результатов, подтверждающих эффективность и корректность методов, предлагаемых в рамках диссертационной работы.

В теории распараллеливания и оптимизации постоянно возникают новые подходы и алгоритмы, анализ которых необходимо проводить в контексте уже существующих, что позволяет определить их оптимальность и эффективность. В настоящее время к таким интенсивно развивающимся проектам относятся Vray, POLARIS, SUIF, OPC и др. В качестве базы для проведения исследований в рамках диссертационной работы выбрана система SUIF. На основе предлагаемых нами новых подходов и библиотек системы SUIF был реализован программный комплекс для анализа зависимостей по данным.

Глава начинается (раздел 3.1.) с описания некоторых деталей алгоритма новой стратегии и программной реализации библиотеки тестов на зависимость.

Далее рассматривается характеристика системной среды и структура инструментов для проведения эксперимента. Также описывается конструкция прототипа распараллеливающего компилятора на основе новой стратегии тестирования и библиотек системы SUIF.

В разделе 3.2. проводится экспериментальное сравнение результатов предложенного метода с наиболее известными стратегиями тестирования анализа зависимостей по данным, такими как Эпсилон-тест и алгоритм Майдана. Эксперимент проведен с использованием инструмента Petit V1.2, разработанного в Мэрилендском университете как расширенный вариант инструмента tiny, и с использованием системы SUIF, разработанной в Стенфордском университете. Обе программы были установлены на персональном компьютере с процессором AMD Athlon XP 1700+ с операционной системой Debian GNU/Linux. Для эксперимента использованы два вида данных. Первый – набор тестовых научных программ NASA и PERFECT Club benchmarks (**PERF**ormance **E**valuation for **C**ost-effective **T**ransformations), где каждая программа состоит от 500 до 18000 строк (см. табл. 1). Второй вид – набор из 16 циклов собранный из работ аналогичных нашей. Все циклы являются специальными примерами и созданы для демонстрации мощности некоторых тестов на зависимость по данным.

Таблица 1

Статистические характеристики эталонных тестовых программ

№	Эталонные тестовые программы	Количество строк программ	Количество подпрограмм	Количество DO-циклов	Количество ссылок на массивы
1	**LGSI**	2815	36	161	6389
2	**LWSI**	1430	17	56	906
3	**SDSI**	8446	80	259	658
4	**TISI**	579	8	78	84
5	**btrix**	159	1	14	488
6	**cholsky**	53	1	18	70
7	**gmtry**	117	1	14	369
8	**gosser**	19	2	5	7
	Всего	**13618**	**146**	**605**	**8971**

Сравнение результатов. С помощью пакетов basesuif 1.3.0.1, suifbuilder 1.3.0.1, baseparsuif 1.3.0.1 и suifcookbook 1.3.0.1 системы SUIF собраны два анализатора зависимостей по данным. Первый основан на алгоритме Майдана, а во втором внедрена новая стратегия. Каждый анализатор принимает на входе преобразованный в SUIF формат (с помощью scc драйвера) *.spd файл последовательной программы, а на выходе дает информацию о всех зависимостях по данным в данной программе. При экспериментальном сравнении результатов рассматривались только *истинные (потоковые) циклически порожденные зависимости по данным*.

Результаты каждого анализатора зависимостей по данным для эталонных тестовых программ показаны в табл. 2. Третья колонка табл. 2 представляет общее количество обращений на предмет наличия потоковой зависимости по

данным, здесь тесты должны разрушать зависимости, если в самом деле не существует потоковой зависимости по данным. В пятой и седьмой колонках таблицы показано, на сколько процентов удалось разрушить зависимости с помощью алгоритма Майдана и с помощью алгоритма новой стратегии соответственно.

Таблица 2

Сравнение результатов

№	Эталонные тестовые программы	Всего обрщ. на истин. завис.	Кол-во разрушенных зависимостей			
			Алгоритм Майдана		Новая стратегия	
1	**LGSI**	6168	5546	89,92%	5546	89,92%
2	**LWSI**	795	102	12,83%	102	12,83%
3	**SDSI**	417	154	36,93%	149	35,73%
4	**TISI**	52	0	0,00%	0	0,00%
5	**btrix**	450	208	46,22%	208	46,22%
6	**cholsky**	61	3	4,92%	0	0,00%
7	**gmtry**	258	125	48,45%	119	46,12%
8	**gosser**	5	0	0,00%	0	0,00%
	Всего	**8206**	**6138**	**74,80%**	**6124**	**74,63%**

Сравнение результатов на экспериментальных примерах. Статистические характеристики экспериментальных примеров: количество строк – 66, количество DO-циклов – 26, количество ссылок на массивы – 40. Эти примеры взяты из аналогичных работ описывающих тесты на зависимость такие как Power-тест (авторы М.Вольф и др.), Омега-тест (автор В.Пью) и др. В данных примерах индексные выражения более сложны, чем реальные программы. В этом случае мы сравнивали результаты следующих алгоритмов: Эпсилон-тест, алгоритм Майдана и новая стратегия тестирования. С помощью инструмента Petit к экспериментальным примерам применялся Эпсилон-тест. Для 40 пар ссылок массивов Эпсилон-тест 44% случаев показал независимость по данным, алгоритм Майдана – 63%, а новая стратегия тестирования для выявления зависимости по данным – 56%.

Время выполнения. Для определения того, какой метод требует больше времени исполнения, использован *GNU* профилировщик ***`gprof'*** операционной системы Linux. Чтобы снизить статистические неточности, каждый алгоритм зависимости по данным выполнен 100 раз для различных эталонных тестов и взято их усредненное значение. В общем случае алгоритм Майдана требовал времени на 25% больше, чем алгоритм новой стратегии (см. Рис. 2).

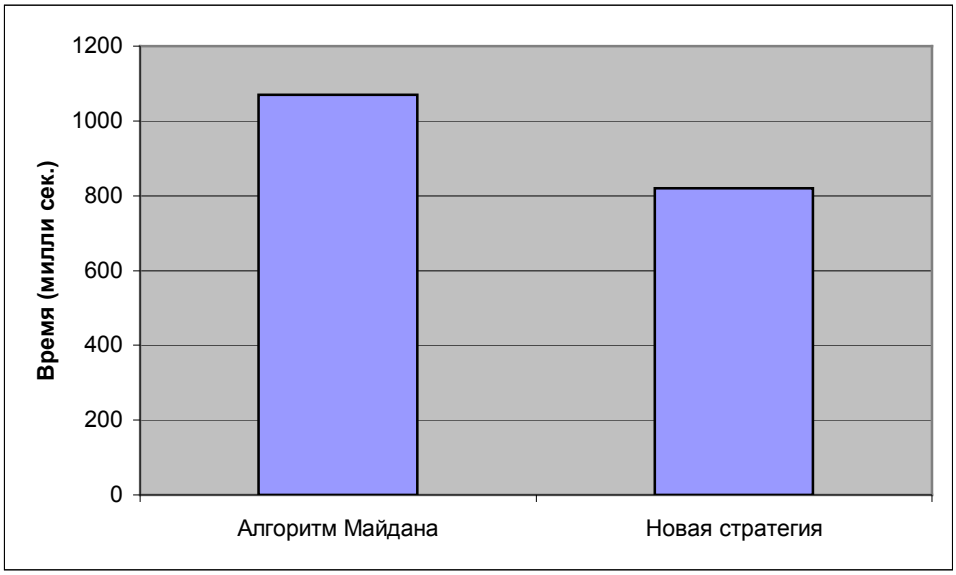

Рис. 2. Время выполнения

Статистические данные *Новой стратегии*. По итогам эксперимента были анализированы статистические данные новой стратегии. Результаты еще раз показывают целесообразность и эффективностью методов, предложенных в предыдущей главе. В большинстве случаев зависимости по данным выявляются с помощью простых тестов (ZIV-тест, SIV-тест, Банержи тест и λ-тест). Помощь более сложных тестов (I-тест, IR-тест, Многомерный I-тест и Модифицированный λ-тест) требовалась в незначительных случаях. Это достигается с помощью алгоритмов, уточняющих ответы теста Банержи. Так в 361 случае с помощью теста Банержи в 54 случаях опровергнуты зависимости по данным, а 301 случае алгоритм проверки коэффициентов доказывает существование зависимости. Это позволяет сократить общее время выполнения новой стратегии, так как после положительного ответа алгоритма проверки исключается выполнение последовательности более сложных тестов. В 71 случае с помощью λ-теста опровергнуты зависимости по данным в 2 случаях, а 67 случае алгоритм проверки коэффициентов доказывает существование зависимости.

Раздел 3.3 посвящен к построению алгоритма для индексного анализа зависимостей по данным в Sisal-программах для SFP системы. Отметим, что она является системой функционального программирования и разрабатывается в Лаборатории конструирования и оптимизации программ ИСИ СО РАН в рамках проекта ПРОГРЕСС, где в качестве начальной версии входного языка выбран Sisal.

В заключении сформулированы основные результаты, полученные в ходе диссертационной работы.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

1. Проведены комплексные исследования существующих тестов на зависимость, позволившие разработать и реализовать новые тесты и усовершенствовать имеющиеся тесты анализа зависимостей по данным.
2. Разработан новый эффективный тест (модифицированный λ-тест) для решения проблемы зависимости по данным при анализе сцепленных ссылок многомерных массивов.
3. Реализована библиотека из новых и модифицированных тестов на зависимость по данным, в состав которой вошли приближенные и точные тесты, включающие одномерные и многомерные случаи.
4. Выработана новая стратегия тестирования, основанная на новых тестах анализа зависимостей по данным. На базе новой стратегии и библиотеки тестов на зависимость создан программный комплекс анализа зависимостей по данным, а также построен алгоритм для индексного анализа зависимости по данным в Sisal-программах в рамках системы функционального программирования (SFP).
5. Проведены экспериментальные исследования для сравнения эффективности предложенных подходов с аналогичными методами анализа зависимостей по данным, выделены наиболее существенные ограничения этих методов.

Личный вклад соискателя заключается в обсуждении постановки задач, разработке адекватных алгоритмов и методов решения, создании и тестировании алгоритмов и программ, проведении расчетов и интерпретации экспериментальных результатов. Все выносимые на защиту результаты принадлежат лично автору. Представление изложенных в диссертации и выносимых на защиту результатов, полученных в совместных исследованиях, согласованно с соавторами.

Благодарности. Автор выражает благодарность научному руководителю д.ф.-м.н. **Владимиру Анатольевичу Евстигнееву** за постоянное внимание и поддержку на всех этапах работы над диссертацией. Автор также благодарит д.ф.-м.н. **Виктора Николаевича Касьянова** за постоянную помощь в работе.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. Евстигнеев В. А., Арапбаев Р. Н., Осмонов Р. А. Анализ зависимостей: основные тесты на зависимость по данным // Сиб. журн. вычисл. математики / РАН, Сиб. отд-е. — Новосибирск, 2007. — Т. 10, № 3. — С. 247–265.
2. Арапбаев Р. Н. Анализ зависимостей по данным: стратегии тестирования и экспериментальное сравнение результатов // Аннотации докл. научной сессии IV Российско-Германской школы по параллельным вычислениям на высокопроизводительных вычислительных системах. — Новосибирск, Академгородок, 9-20 июля 2007г. / Вычислительные технологии — 2007. — Т. 12, №6. — С. 138-142.

3. Арапбаев Р. Н., Осмонов Р. А. Анализ зависимостей по данным для многомерных массивов на базе модифицированного λ-теста // Проблемы интеллектуализации качества систем информатики / РАН, Сиб. отд-е, Ин–т систем информатики. — Новосибирск, 2006. — С. 7–23.
4. Арапбаев Р. Н. Экспериментальное исследование новой стратегии // Методы и инструменты конструирования программ. / РАН, Сиб. отд-е, Ин–т систем информатики. — Новосибирск, 2007. — С. 7–23.
5. Арапбаев Р. Н., Осмонов Р. А. Анализ зависимостей: новая стратегия тестирования // Труды Международной конференции «Параллельные вычислительные технологии (ПаВТ'2007)». — Челябинск: изд-во ЮУрГУ, 2007. — Т.2. — С. 16–27.
6. Арапбаев Р. Н., Евстигнеев В. А., Осмонов Р. А. Сравнительный анализ тестов на зависимость по данным. — Новосибирск, 2006. — 36 с. — (Препр. / РАН. Сиб. отд-е. ИСИ; № 141).
7. Арапбаев Р. Н. Анализ зависимостей по данным: стратегии тестирования и экспериментальное сравнение результатов // Научная сессия IV Российско-Германской школы по параллельным вычислениям на высокопроизводительных вычислительных системах / РАН, Сиб. отд-е, Ин–т выч. техн. — Новосибирск, 2007. — С. 11–14.
8. Арапбаев Р. Н. Новая стратегия применений тестов для выявления зависимостей по данным // VII Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям. — Красноярск, 2006. — С. 78.
9. Арапбаев Р. Н., Осмонов Р. А. Новый алгоритм анализа зависимостей по данным в многомерных массивах // VI Всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям. — Кемерово: изд-во КемГУ, 2005. — С. 57.
10. Арапбаев Р. Н., Осмонов Р. А., Фомин А. С. Программный комплекс для анализа зависимостей по данным // Конференция-конкурс «Технологии Microsoft в теории и практике программирования». — Новосибирск, 2008. — С. 99–101.
11. Арапбаев Р.Н., Осмонов Р.А. Анализ зависимостей по данным для многомерных массивах с учетом векторов направлений // Конференция-конкурс «Технологии Microsoft в теории и практике программирования». — Новосибирск, 2006. — С. 153–155.
12. Арапбаев Р.Н., Осмонов Р. А. Сравнительный обзор тестов на зависимость по данным // XLIV Междунар. науч. студенческая конф. «Студент и научно-технический прогресс»: Информационные технологии / НГУ — Новосибирск, 2006. — С. 4–5.

АРАПБАЕВ Русланбек Нурмаматович

АНАЛИЗ ЗАВИСИМОСТЕЙ ПО ДАННЫМ:
ТЕСТЫ НА ЗАВИСИМОСТЬ И СТРАТЕГИИ ТЕСТИРОВАНИЯ

Автореф. дисс. на соискание учёной степени кандидата физико-математических наук.
Подписано в печать 06.11.2008. Заказ № 99. Формат 60x90/16. Усл.печ.л.1. Тираж 100 экз.
Отпечатано в издательском отделе Института катализа им. Г.К. Борескова СО РАН
630090 Новосибирск, пр. Академика Лаврентьева, 5