П р е д и с л о в и е

В 23-м томе Трудов ИСП РАН публикуются статьи сотрудников Института, их коллег из других организаций, а также аспирантов ИСП РАН, МГУ, МФТИ, СПбГУ. В статьях описываются результаты исследований, выполненных в течение 2012 года. Статьи, публикуемые на английском языке, содержат расширенные варианты текстов докладов, представленных на Восьмом весеннем коллоквиуме исследователей в области баз данных и информационных систем (SYRCoDIS 2012).

Первая группа статей посвящена аспектам технологий компиляции и анализа программ.

В статье М.С. Акопяна «Расширение модели ParJava для случая кластеров с многоядерными узлами» описывается расширение модели параллельной SPMD программы возможностью использования потоков Java. Использование потоков в программе позволяет лучше утилизировать ресурсы многоядерного процессора. Разработанная модель позволяет оценивать время выполнения параллельной программы с явными обращениями к библиотеке MPI, где в каждом процессе можно использовать параллельные потоки Java.

А. Меркулов и А. Белеванцев в статье «Реализация конвейеризации циклов и встраивания присваиваний в трансляторе C-to-HDL» описывают инструмент для трансляции функций языка Си в модули на языке Verilog, процесс трансляции и две реализованных на уровне описания аппаратуры оптимизации: встраивание присваиваний и конвейеризация циклов. Результаты тестирования показывают, что эти оптимизации существенно увеличивают производительность генерируемого кода.

Статья М.Г. Бакулина, С.С. Гайсаряна и др. посвящена описанию результатов экспериментального исследования по восстановлению графа потока управления, запутанного специализированным компилятором на основе LLVM. Средства запутывания и распутывания бинарного кода независимо разрабатывались двумя коллективами ИСП РАН. Помимо того, для количественной оценки стойкости запутывания были получены метрики сложности кода модельных примеров.

В статье Ш. Ф. Курмангалеева, В. П. Корчагина и Р.А. Матевосяна «Описание подхода к разработке обфусцирующего компилятора» приводится обзор запутывающих преобразований программ, формулируют критерии эффективности методов обфускации. Предлагается подход к реализации обфусцирующего компилятора на основе инфраструктуры LLVM. Особенность подхода заключается в одновременном применении преобразований, маскирующих различные аспекты работы запутываемого приложения, что обеспечивает стойкую защиту от статического анализа.

В статье Ш. Ф. Курмангалеева, В. П. Корчагина, В. В. Савченко и С.С. Саргсяна «Построение обфусцирующего компилятора на основе

инфраструктуры LLVM» описываются маскирующие преобразования, реализованные в ходе разработки обфусцирующего компилятора в ИСП РАН, приводится оценка понижения быстродействия и увеличения объема потребляемой приложением памяти, а также оценка возможности восстановления информации об исходном коде.

А. Г. Назаров, М. А. Климушенкова, П. М. Довгалюк и В. А. Макаров статье «Повышение уровня представления трасс выполнения программ» отмечают, что трассы выполнения программ, состоящие из последовательности выполненных процессором инструкций, слишком велики для непосредственного восприятия человеком. Поэтому актуальной является задача повышения уровня представления таких трасс. Предлагается метод повышения уровня представления, предназначенный для построения модели алгоритма по трассе выполнения программы.

Статья Н.Л. Луговского «Подход для проведения рефакторинга «Выделение функции» в инструменте Klocwork Insight» содержит обсуждение подхода для проведения рефакторинга исходного кода на языках Си/Си++, реализованного в инструменте Klocwork Insight. Приводится подробное описание подхода на примере рефакторинга «Выделение функции». Разбираются способы обработки различных языковых конструкций при проведении рефакторинга, показывается, как структурные изменения в синтаксическом дереве отображаются в изменения исходного кода программы.

В статье В.О. Савицкого и Д.В. Сидорова «“Ленивый” анализ исходного кода на языках C и C++» описывается метод построения синтаксического анализатора, позволяющий существенно сократить требуемые для анализа ресурсы. Метод основан на том факте, что каждый исходный файл подключает множество заголовков, из которых используется лишь небольшое количество определений. Отличительной особенностью метода является необходимость внесения лишь небольшого количества изменений в существующий парсер. Метод реализован в статическом анализаторе Klocwork Insight.

В следующей группе статей обсуждаются различные проблемы области управления данными.

Статья П.А. Клеменкова и С.Д. Кузнецова «Большие данные: современные подходы к хранению и обработке» посвящена анализу возможных способов решения проблемы больших данных, ограничений, которые не позволяют сделать это эффективно. Приводится обзор трех современных подходов к работе с большими данными: NoSQL, MapReduce и обработка потоков событий в реальном времени.

В статье А. Посконина «Web-приложения и данные: проблемы абстракции и масштабируемости» отмечается, что несмотря на то, что SQL-ориентированные системы управления базами данных (СУБД) широко применялись ранее и применяются сегодня, масштабирование приложений при использовании этих систем сильно затруднено. В связи с этим появилось

много различных хранилищ данных, стремящихся удовлетворить требованиям современных высоконагруженных Web-приложений. В настоящее время наблюдается тенденция к использованию нескольких технологий и систем в рамках одного приложения для решения различных задач. Рассматриваются основные подходы к реализации доступа к данным и проблемы абстракции от используемых хранилищ, а также предлагаются некоторые принципы организации слоя работы с данными при использовании нескольких хранилищ данных разного типа.

С.Д. Кузнецов и А.А. Прохоров в статье «Алгоритмы управления буферным пулом СУБД при работе с флэш-накопителями» указывают, что флэш-накопители имеют серьезные преимущества по сравнению с традиционными жесткими дисками, главные из которых – более высокая скорость чтения и записи, а также значительно меньшее время доступа к данным. Однако самые распространенные виды флэш-памяти читают данные с большей скоростью, чем записывают. Из-за этой особенности использование классических алгоритмов замещения страниц при кэшировании дисковых данных неэффективно. Приводится обзор современных алгоритмов управления буферным пулом СУБД, которые предназначены для работы с накопителями на флэш-памяти.

В статье С. Д. Кузнецова и Н. А. Мендковича «Обзор развития методов лексической оптимизации запросов» обсуждаются проблемы лексической оптимизации запросов и описываются работы, опубликованные в течение последних четырех десятилетий. Особое внимание уделяется таким подходам к оптимизации, как модификация, украшение и сокращение запросов. Рассматриваются алгоритмы оптимизации для реляционных и нереляционных СУБД.

Статья А. Коршунова и А. Гомзина «Тематическое моделирование текстов на естественном языке» относится к области тематического моделирования — построению модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Переход из пространства терминов в пространство найденных тематик помогает разрешать синонимию и полисемию терминов, а также эффективнее решать такие задачи, как тематический поиск, классификация, реферирование и аннотирование коллекций документов и новостных потоков. Приведён сравнительный обзор различных моделей, описаны способы оценивания их параметров и качества результатов, а также даны примеры открытых программных реализаций.

А.В. Кошкарев, А. А. Медведев и др. в статье «Виртуальная ГИС-лаборатория как инструмент анализа пространственных данных» отмечают, что на основе технологической платформы UniHUB, разработанной в ИСП РАН, в составе Дата-центра РАН создана веб-лаборатория, нацеленная на интеграцию данных дистанционного зондирования в интересах наук о Земле. Информационная система ориентирована на научное и образовательное сообщество и предназначена для совместной научной работы ее участников в едином

рабочем пространстве, обеспечивая поиск источников пространственных данных, формирование хранилищ данных, доступ к данным, в том числе к ресурсам внешних открытых веб-сервисов, к приложениям и обучающим материалам.

В статье А.А. Алексеева и Н.В. Лукашевич «Комбинирование признаков для извлечения тематических цепочек в новостном кластере» предлагается метод для извлечения цепочек семантически близких слов и выражений, описывающих различных участников сюжета – тематических узлов. Метод основан на структурной организации новостных кластеров и анализе контекстов вхождения языковых выражений. Контексты слов используются в качестве базиса для извлечения многословных выражений и построения тематических узлов. Оценка предложенного алгоритма производится в задаче построения обзорных рефератов новостных кластеров.

А. Ю. Пигуль в статье «Сравнительный анализ параллельных алгоритмов соединения для среды MapReduce» отмечает, что для анализа больших объемов данных используются такие методы как параллельные СУБД, парадигма MapReduce, колоночное хранение и различные комбинации этих подходов. В данной работе будут рассмотрены алгоритмы соединения в среде MapReduce. К сожалению, в MapReduce алгоритмы соединения напрямую не поддерживаются. Цель данной работы заключается в том, чтобы обобщить и сравнить существующие алгоритмы соединения по равенству с некоторыми методами оптимизации.

Статья К. Кузнецова « Система интеграции данных на основе наборов RDF-связей пространства Linked Open Data » посвящена разработке системы интеграции данных в пространстве Linked Open Data. Предложено архитектурное решение, осуществляющее весь комплекс задач по публикации данных из множества локальных гетерогенных источников в пространстве Linked Open Data. В системе предлагается применять новый подход исполнения федеративных SPARQL запросов с использованием наборов RDF связей.

В статье Е.А. Иванниковой «Обнаружение периодических наборов событий во временных базах данных» указывается, что повторяемость во временных данных – это важное свойство, которое может быть использовано во многих приложениях. Такая регулярность исследуется в области интеллектуального анализа периодических шаблонов. Ставится проблема обнаружения периодических наборов и предлагается способ ее решения. Подробно рассматриваются существующие алгоритмы обнаружения периодических событий и новый подход. Сравнительный анализ алгоритмов и их производительность демонстрируются через серию экспериментов.

Статьи следующего блока посвящены вопросам тестирования программного обеспечения.

И.Б. Бурдонов и А.С. Косачев в статье «Зависимости между ошибками на коассах тестируемых реализаций» исследуют проблему зависимости между

ошибками, определяемыми спецификацией, и связанную с ней проблему оптимизации тестов. Предлагается формальная модель тестового взаимодействия самого общего вида и конформность типа редукции, для которых зависимость между ошибками практически отсутствует. Показывается, что многие известные конформности в различных семантиках взаимодействия являются частными случаями этой общей модели.

Статья В.В. Кулямина «Комбинаторная генерация программных конфигураций ОС» содержит описание метода генерации тестов для конфигурационного тестирования на основе покрывающих наборов. Новым элементом в предлагаемом методе является учет условий использования отдельных параметров, который вносит коррективы как в учет покрываемых комбинаций, так и в построение отдельных тестов. Описываемый метод использован на практике для генерации тестовых программных конфигураций операционной системы реального времени, приведены результаты этого применения.

В статье Е.М. Новикова и А.В. Хорошилова «Использование аспектно-ориентированного программирования для выполнения запросов по исходному коду программ» отмечается, что запросы по исходному коду программ помогают разработчикам обнаруживать искомые фрагменты кода и определять их взаимоотношения друг с другом. Предлагается подход к выполнению запросов по исходному коду программ на основе аспектно-ориентированного программирования, рассматриваются достоинства и недостатки такого подхода.

А.В. Никешин, Н.В. Пакулин и В.З. Шнитман в статье «Разработка тестового набора для верификации реализаций протокола безопасности TLS» кратко описывается метод формализации требований протокола безопасности TLS, тестовый набор, а также результаты тестирования существующих реализаций сервера TLS. Эти результаты показывают, что предложенный в данной работе метод верификации позволяет эффективно автоматизировать тестирование таких сложных протоколов, как протоколы безопасности.

Д. Бейер и А. К. Петренко в статье «Верификация драйверов операционной системы Linux» приводят аргументы в пользу рассмотрения драйверов операционной системы Linux в качестве области для практического применения различных методов и инструментов верификации программ. Представлен обзор современного состояния дел в области верификации и указаны направления исследований, больше всего нуждающиеся в их дальнейшем развитии.

Статья А.В. Цыварева и В. Мартиросяна «Система Spruce: динамическая верификация качества драйверов файловых систем в ОС Linux» посвящена исследованию подходов к динамической верификации драйверов файловых систем ОС Linux. Помимо рассмотрения существующих систем, которые можно применить для этой цели, представлена система Spruce, которая предназначена для верификации драйверов конкретных файловых систем.

Последние четыре статьи тома посвящены различным теоретическим вопросам, относящимся к тематике системного программирования.

В статье А.В. Шокурова и К.В. Сергеева «Об одном методе построения схемы полного гомоморфного шифрования» предложен вариант метода Джентри для построения полного гомоморфного шифрования.

Статья Д.А. Грушина и Н.Н. Кузюрина «Энергоэффективные вычисления для группы кластеров» посвящена проблеме балансировки нагрузки для множества параллельных задач на группе географически распределенных кластеров уменьшающая количество энергии при вычислениях. Предложены несколько алгоритмов распределения задач и проведена экспериментальная проверка их эффективности.

С.Н. Жук представил статью «О построении расписаний выполнения параллельных задач на группах кластеров с различной производительностью», в которой предложен онлайновый алгоритм распределения параллельных задач на группе кластеров с различными производительностями процессоров и показано, что он гарантирует для любого потока задач построение расписания, отличающегося от оптимального не более, чем в 2e раз.

Наконец, в статье Т.А. Новиковой и В.А.Захарова «Унификация программ» в качестве эквивалентности программ рассматривается отношение логико-термальная эквивалентности, одно из наиболее слабых разрешимых отношений эквивалентности программ, аппроксимирующих отношение функциональной эквивалентности. Опираясь на алгоритм проверки логико-термальной эквивалентности программ, авторы предлагают полиномиальную по времени процедуру вычисления наиболее общего унификатора для произвольной пары последовательных императивных программ относительно логико-термальной эквивалентности.

Академик РАН В.П. Иванников