На правах рукописи

Викторов Юрий Олегович

АНАЛИЗ И ВЕРИФИКАЦИЯ ЗАДЕРЖЕК В МИКРОАРХИТЕКТУРЕ КОММУНИКАЦИОННЫХ ФАБРИК

05.13.05 – «Элементы и устройства вычислительной техники и систем управления»

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Санкт-Петербург – 2013

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Санкт-Петербургский государственный политехнический университет» (ФГБОУ ВПО «СПбГПУ»)

Научный руководитель: МАРАХОВСКИЙ Вячеслав Борисович
доктор технических наук, профессор

Официальные оппоненты: КУПРИЯНОВ Михаил Степанович
доктор технических наук, профессор,
декан ф-та компьютерных технологий
и информатики ФГБОУ ВПО СПбГЭТУ
«ЛЭТИ»

СУВОРОВА Елена Александровна
кандидат технических наук,
зав. лаб. проектирования систем на кристалле
института высокопроизводительных
вычислительных систем ФГАОУ ВПО «ГУАП»

Ведущая организация: Институт электронных управляющих машин
им. И.С. Брука

Защита состоится “03” апреля 2014 г. в 16 часов на заседании диссертационного совета Д 212.229.18 ФГБОУ ВПО «Санкт-Петербургский государственный политехнический университет» по адресу: 195251, Санкт-Петербург, ул. Политехническая, д. 21, 9-й учебный корпус, ауд. 325.

С диссертацией можно ознакомиться в Фундаментальной библиотеке СПбГПУ по адресу 195251, Санкт-Петербург, ул. Политехническая, д. 29.

Автореферат разослан «___» _____ 2014 г.

Ученый секретарь
диссертационного совета
Д 212.229.18 к.т.н., доцент  Васильев Алексей Евгеньевич

Общая характеристика работы

Актуальность темы исследования и степень её разработанности. Рост сложности микропроцессорной архитектуры и требований к её производительности и масштабируемости привел к отказу от общей шины как средства коммуникации между узлами микропроцессора и переходу на использование так называемых коммуникационных фабрик (Communication Fabrics).

Коммуникационные фабрики – это современные системы межсоединений узлов системы на кристалле, для которых характерна одновременная обработка множества запросов, за счет глубокого параллелизма и конвейеризации.

Архитектура коммуникационных фабрик может быть самой разнообразной, в зависимости от области применения. Для северного комплекса микросхем потребительской электроники (исторически «северный мост» – контроллер-концентратор памяти) характерна централизованная структура. Южный комплекс (исторически «южный мост» – контроллер-концентратор ввода-вывода) традиционно имеет древовидную иерархическую структуру. Для серверов применяют распределенные архитектуры типа «сеть на кристалле» (Network on Chip, NoC), такие как кольца и решетки.

Ресурсы в системах на кристалле (размеры очередей, разрядности шин, и т.д.), как правило, ограничены, а производительность и корректность работы системы сильно зависят от эффективности их разделения конкурирующими процессами. Поэтому при проектировании коммуникационных фабрик требуется учитывать минимальный уровень качества обслуживания. Несоблюдение требуемого уровня качества обслуживания приводит к разнообразным последствиям. Например, выпадение кадров при проигрывании видео, или прерывистое воспроизведение звука. Попытка же решить данную проблему увеличением количества ресурсов приводит к удорожанию системы.

В контексте анализа качества обслуживания в коммуникационных фабриках наибольший интерес представляют следующие характеристики:

- форма потока – объем данных, переданных потоком, начиная с некоторого момента времени t_0; для описания формы потока часто используют показатели средней плотности (rate, bandwidth) и величины всплеска (burst);
- задержка передачи (latency) – интервал времени между отправлением данных источником и их получением в точке назначения;
- вариация задержки (jitter) – разница между максимальным и минимальным значениями задержки в потоке данных.

Данная работа посвящена анализу задержек передачи.

Существуют две категории подходов к проектированию коммуникационных фабрик, призванных обеспечить гарантии качества обслуживания “по построению”: подходы на основе дифференциации потоков (traffic differentiation) и на основе передачи данных без конкуренции (contention-free transmisson). Задача выполнения требований качества обслуживания не может быть в полной мере решена использованием этих подходов. Указанные подходы не предлагают ни эффективного способа борьбы с проблемой истощения ресурсов, ни способов оценки задержек на этапе резервирования ресурсов. Требуются надежные и эффективные методы оценки гарантий качества обслуживания, предоставляемых коммуникационной фабрикой, применимые на ранних этапах разработки микроархитектуры, т.к. обнаруженные на поздних стадиях ошибки не всегда могут быть эффективно исправлены. Это приводит к удорожанию и срыву сроков разработки, что неприемлемо в условиях жестких временных ограничений. Проектирование системы на кристалле для потребительской электроники (мобильные устройства и т.д.) не должно занимать более полугода и малейшие срывы сроков являются фатальными. Практически каждое такое решение ориентировано на конкретный конечный продукт, который безнадежно устареет к моменту появления на рынке при больших сроках проектирования.

В настоящее время для анализа производительности систем на кристалле используют имитационное моделирование и набор формальных методов. Однако все существующие на данный момент методы имеют ряд серьезных недостатков: либо не дают никаких гарантий, кроме статистических (имитационное моделирование), либо требуют нереалистичных упрощений исходной задачи (аналитические методы), либо не масштабируются на объекты размеров коммуникационной фабрики (формальная верификация), либо требуют чрезмерных усилий и затрат для их применения (методы доказательства теорем).

Жесткие требования к срокам разработки на практике означают, что этап разработки микроархитектуры системы на кристалле не может занимать более месяца. Соответственно, на первый план выходит получение как можно более оперативного ответа, как изменения микроархитектуры отражаются на её характеристиках. Традиционный подход, когда одна итерация верификации выполняется неделями, здесь неприемлем.

Анализ качества обслуживания тесно связан с подходом к моделированию микроархитектуры коммуникационных фабрик, в качестве которого предлагается использовать среду моделирования xMAS. Среда моделирования xMAS (eXecutable Microarchitecture Specification) была специально разработана для высокоуровневого моделирования микроархитектуры в простой и наглядной форме.

Модели xMAS конструируются из небольшого числа параметризованных стандартных блоков (примитивов), соединенных каналами в синхронную сеть передачи пакетов. Примитивы выполняют простые операции над пакетами данных. Например, queue (очередь) выполняет буферизацию пакетов, сохраняя порядок их следования, а примитивы join и fork играют роль барьеров, синхронизируя передачи на входных и выходных каналах.

Для верификации моделей xMAS используются как специализированные алгоритмы, так и средства верификации моделей общего назначения. Модель xMAS всегда может быть автоматически представлена в виде эквивалентного описания на языке Verilog, к которому применимы методы символьной верификации, такие как ограниченная верификация (bounded model checking, BMC), интерполяция, k-индукция и т.д. Область применения средств верификации общего назначения можно существенно расширить, используя структурный анализ модели для автоматического порождения вспомогательных лемм.

Модели xMAS успешно применялись для обнаружения тупиковых ситуаций (deadlocks) и формального доказательства их отсутствия. Однако моделирование качества обслуживания требует учета большего числа деталей и требует расширения набора примитивов xMAS.

Цель диссертационного исследования - разработка метода автоматического вывода верхней границы на задержку передачи данных в микроархитектурных моделях коммуникационных фабрик, формальной верификации этой границы за время, не превышающее несколько часов, а также исследование возможностей, ограничений и области возможного применения разработанного метода.

Задачи диссертационного исследования:

1. расширение языка моделирования xMAS для представления в моделях временных характеристик микроархитектуры коммуникационных фабрик, в частности, введение в язык новых примитивов;
2. построение математического аппарата для описания задержки передачи данных и формализация суждения о задержке в моделях xMAS;
3. разработка алгоритмов для автоматического анализа структуры модели, получения оценки сверху на задержку передачи данных и верификации этой оценки за малое время (<24 часов);
4. проведение серии экспериментов над моделями коммуникационных фабрик и их частей, имеющих разные размеры, с целью изучения возможностей разработанного подхода.

Объектом исследования является коммуникационная фабрика в составе системы на кристалле. **Предметом исследования** являются временные характеристики коммуникационных фабрик, в зависимости от их микроархитектуры и окружения.

Научная новизна диссертационной работы заключается в следующем:

1. Разработано расширение языка xMAS для моделирования временных характеристик архитектуры коммуникационных фабрик.
2. Разработан алгоритм, позволяющий автоматически получить верхнюю границу задержки передачи данных, анализируя структуру микроархитектурной модели, представленной на языке xMAS.
3. Предложен новый подход, позволяющий в процессе вывода верхней границы задержки, параллельно с выводом также сконструировать доказательство задержки, не выполняя повторно трудоёмкие задачи анализа модели.

Теоретическая значимость работы

1. Предложена математическая модель для описания задержки между событиями и формализовано суждение о задержке для последовательности событий произвольной природы, обладающих определенными свойствами.
2. Предложен новый подход к доказательству задержек с помощью метода k-индукции, при использовании которого глубина индукции не зависит от величины доказываемой задержки.

Практическая значимость работы

1. Предложен новый подход к анализу микроархитектурных моделей, позволяющий получить оценки сверху на задержку передачи данных и эффективно верифицировать такие оценки, используя стандартные средства формальной верификации.
2. На базе предложенных алгоритмов создан программный прототип для автоматического анализа микроархитектурных моделей, записанных на языке xMAS, получения оценки сверху на задержку передачи данных и её верификации. Верификация осуществляется путем конструирования вывода для задержки, каждый шаг которого проверяется с помощью методов k-индукции для малых значений k, не зависящих от величины доказываемой задержки.
3. Экспериментально подтверждена применимость нового подхода к анализу моделей коммуникационных фабрик, используемых в современных микропроцессорах, в то время, как традиционные подходы к верификации оценок не позволяют получить ответ в течение недели на моделях подобных размеров (так для модели, состоящей из ~400 примитивов и ~50 очередей, с оценкой пространства состояний ~10^500,

предлагаемый подход позволяет провести доказательство за ~30 сек., а стандартными средствами доказательство не удаётся выполнить за неделю).

Методология и методы исследования. В работе использовались:

1. язык моделирования xMAS – для моделирования микроархитектуры коммуникационных фабрик и их окружения;
2. язык линейной темпоральной логики (LTL) – как удобное средство спецификации свойств модели xMAS, относящихся к задержке, и построения суждений о задержке;
3. сторонние инструментальные средства: ABC Berkeley – для ограниченной верификации моделей и доказательства утверждений методом k-индукции, Synopsis VCS – для имитационного моделирования описаний на языке Verilog;
4. математический аппарат ранжирующих функций для построения эффективно проверяемых доказательств границы на задержку передачи.

Основные положения и результаты, выносимые на защиту. На защиту выносятся следующие результаты, полученные автором в процессе проведения исследований:

1. создано расширение языка xMAS для представления в микроархитектурных моделях свойств, связанных с временными характеристиками;
2. предложено математическое описание задержки между событиями и формализован процесс суждения о задержке;
3. предложена алгоритмическая схема анализа моделей xMAS для получения оценок сверху на задержку передачи данных из одной точки в другую;
4. разработан подход к построению доказательств полученных временных оценок для широкого класса моделей, на основе метода k-индукции;
5. при доказательстве задержек достигнуто ограничение глубины индукции малым значением k за счет использования ранжирующих функций;
6. продемонстрировано сокращение времени верификации задержек на несколько порядков, при допустимом уровне консервативности получаемых оценок.

Степень достоверности и апробация результатов

Достоверность результатов обеспечивается всесторонним анализом проблемной области, поставленной цели и известных подходов к ее достижению; корректностью используемого математического аппарата; использованием сторонних общедоступных и зарекомендовавших себя средств формальной верификации. Используется подход к верификации полученных оценок задержки, при котором конечный результат проверяется не предполагая правильности работы алгоритма анализа или истинности сделанных в ходе

анализа предположений. Полученные результаты сравнивались как с результатами имитационного моделирования, так и с экспертными оценками.

Внедрение и реализация результатов работы:

Основные результаты работы используются в ЗАО «Интел А/О» (Intel corp.) при проектировании систем на кристалле для технологии 14нм. Реализация основных положений и результатов работы подтверждена соответствующими документами о внедрении. Также результаты работы используются в учебном процессе на кафедре Микропроцессорных Технологий МФТИ в дисциплине «Математические основы САПР»

Апробация работы. Основные теоретические и практические результаты работы были представлены на конференциях:

1. V Всероссийская научно-техническая конференция "Проблемы разработки перспективных микро- и наноэлектронных систем" МЭС-2012, ИППМ РАН, 1 доклад (2012);
2. 37-ая международная научная конференция «Гагаринские чтения», МАТИ, 1 доклад (2011).

Публикация результатов исследования. Результаты диссертации отражены в 5 публикациях, в том числе 3 входят в перечень научных журналов и изданий, рецензируемых ВАК.

Структура и объем работы. Диссертационная работа состоит из введения, 5 глав, заключения и списка литературы. Работа содержит 140 страниц машинного текста, 33 графика и рисунка, 17 таблиц, список литературы из 111 наименований и 2 приложения.

Основное содержание работы

Во **введении** рассматриваются актуальность, научная новизна работы, определяются цели и задачи исследования.

В **первой главе** описывается задача обеспечения качества обслуживания применительно к микроархитектуре коммуникационных фабрик. Рассматриваются существующие подходы к проектированию коммуникационных фабрик, учитывающие требования качества обслуживания, и методы анализа свойств качества обслуживания. Обосновывается необходимость разработки новых методов для получения надежных оценок на задержку передачи данных, применимых к широкому классу моделей коммуникационных фабрик.

Существуют различные подходы к проектированию коммуникационных фабрик, призванные обеспечить гарантии качества обслуживания “по построению”. Их можно

условно разделить на две категории: подходы на основе дифференциации потоков (traffic differentiation) и на основе передачи данных без конкуренции (contention-free transmisson).

Подходы на основе дифференциации потоков решают проблему конкурирующих запросов введением уровней приоритета, соответствующих разным классам трафика. Однако для сложной системы затруднительно определить необходимое количество уровней приоритетов и избежать проблем «инверсии приоритета» (ситуаций, когда низкоприоритетный запрос блокирует выполнение высокоприоритетного, что переводит первый в категорию высокоприоритетных запросов) и «истощения ресурсов» (starvation).

Передача без конкуренции основывается на той или иной схеме резервирования ресурсов перед началом передачи. Статическое резервирование позволяет добиться определенных гарантий производительности, что позволяет оптимизировать архитектуру под конкретную задачу для встраиваемых систем, но приводит к низкой загрузке ресурсов в системах широкого спектра применения. При динамическом резервировании ресурсов трудно оценить задержки на стадии резервирования из-за конкуренции за право зарезервировать ресурс.

Таким образом, задача выполнения требований качества обслуживания не может быть в полной мере решена «по построению». Требуются надежные и эффективные методы для оценки архитектуры с позиции предоставляемых ею гарантий качества обслуживания.

В настоящее время для анализа производительности в системах на кристалле используют имитационное моделирование и формальные методы.

Имитационное моделирование применяется к моделям различной сложности и детализации. Обычно для анализа производительности используют модели высокого уровня, реализованные на C++ или SystemC. Моделирование подробного RTL-описания коммуникационной фабрики и окружения дает наиболее точные результаты, но возможно на поздних стадиях разработки и чрезмерно трудоемко. При этом сложность проектируемых систем растет быстрее, чем производительность средств имитационного моделирования.

Принципиальным ограничением имитационного моделирования является возможность выявить только статистически значимые ошибки и не дает достаточной уверенности в корректной работе системы. Тупиковые состояния и проблемы истощения ресурсов часто возникают в нетипичных для данной системы условиях, поэтому их можно гарантированно обнаружить только для систем небольшого размера путем полного перебора поведений.

Для исчерпывающего анализа качества обслуживания используют формальные подходы, такие как аналитические методы, формальная верификация (верификация моделей) и доказательство теорем. Аналитические методы работают с сильно обобщенной моделью системы и позволяют получать приближенные оценки метрик качества обслуживания в виде

формул. Однако этот подход применим только для сетей определенного вида, с рядом ограничений, которые в коммуникационных фабриках, как правило, не выполняются.

Формальная верификация работает с абстрактной математической моделью системы. Описание модели дополняется спецификацией ее свойств на языке темпоральной логики. Для каждого типа моделей существуют алгоритмы проверки спецификаций, основанные на переборе состояний и поведений системы в той или иной форме. Быструю сходимость алгоритмов можно гарантировать далеко не всегда, так как она зависит от характеристик модели, сложности спецификации и способа организации перебора. Но современные достижения в области задач SAT и SMT (SATisfiability и Satisfiability Modulo Theories), т.е. задач разрешимости для логических формул, существенно расширили применимость методов формальной верификации.

Основной проблемой применения стандартных средств верификации моделей к задачам качества обслуживания является низкая масштабируемость (нелинейное увеличение трудоемкости алгоритмов с ростом сложности задачи). Для получения консервативной оценки задержки L в наихудшем случае можно воспользоваться k-индукцией. Но сходимость метода можно гарантировать только при количестве шагов k, близком к величине оценки L. Значение L может достигать сотен и тысяч тактов, что находится далеко за пределами возможностей существующих средств формальной верификации логических сетей.

Метод доказательства теорем требует описания поведения системы в некоторой формальной логике или в виде особого вида программы. Гарантии качества обслуживания могут быть сформулированы в виде основных теорем, доказательство которых проходит в полуавтоматическом режиме и обычно требует большого числа вспомогательных утверждений (лемм), вводимых вручную. Высокая трудоемкость часто делает доказательство теорем нерентабельным, особенно на ранних этапах проектирования.

Для некоторых видов сетей могут применяться специализированные языки моделирования и средства анализа. Например, для оценки производительности эластичных систем используют маркированные графы; сети на кристалле Æthereal моделируют с помощью диаграмм потоков данных. Однако, такие методы неприменимы для коммуникационных фабрик общего вида.

Таким образом, все существующие на данный момент методы имеют ряд серьезных недостатков: либо не дают никаких гарантий, кроме статистических, либо требуют нереалистичных упрощений исходной задачи, либо не могут эффективно использоваться для анализа объектов размера коммуникационной фабрики, либо требуют чрезмерных усилий и затрат для их применения.

Во **второй главе** рассматривается математический аппарат, разработанный автором для построения суждений о задержке между событиями. Вводятся понятия отношения отклика, связанной с ним границы задержки, интервала и модуля для исчисления задержки. Приводится ряд правил, позволяющих манипулировать отношениями отклика для вывода новых отношений и связанных с ними границ на задержку. Излагается методика использования ранжирующих функций для индуктивного доказательства границ задержки между событиями, связанными отношением отклика. Для каждого из правил вывода задержки вводится двойственное правило, позволяющее параллельно с выводом задержки, шаг за шагом конструировать ранжирующую функцию для верификации этой задержки.

Задержка определяется как количество событий наступления очередного такта между предусловием A и постусловием B, где A и B – произвольные события, связанные отношением отклика:

$$A \rightsquigarrow B \equiv \mathbf{G}(A \rightarrow \mathbf{F}B).$$

Используются стандартные обозначения линейной темпоральной логики (LTL): $\mathbf{G}A$ – "всегда A", $\mathbf{F}A$ – "однажды A", $\mathbf{X}A$ – "в следующий момент времени A". Иначе говоря, за событием A всегда следует событие B. Определим локальную задержку $\mathbf{lat}[B]$ равную числу тактов между текущим моментом времени и следующим появлением события B:

$$(\mathbf{lat}[B] \leq k) \equiv \mathbf{F}^{\leq k}B,$$

Где $\mathbf{F}^{\leq k}B$ определяется как $B + \mathbf{X}B + \cdots + \mathbf{X}^k B$, а $\mathbf{X}^k$ соответсвует k последовательным применениям темпорального оператора $\mathbf{X}$. Если событие B никогда не наступает, т.е. выполнено $\mathbf{G}\neg B$, задержку $\mathbf{lat}[B]$ будем считать равной ∞. С каждым отношением отклика $A \rightsquigarrow B$ можно связать максимальную задержку за все время выполнения

$$(\mathbf{Lat}[A \rightsquigarrow B] \leq k) \equiv \mathbf{G}\big(A \rightarrow (\mathbf{lat}[B] \leq k)\big).$$

Если событие A не наступает ни разу, удобно считать, что $\mathbf{Lat}[A \rightsquigarrow B] = 0$.

Определение отношения отклика и связанной с ним задержки обобщается следующим образом:

$$A \rightsquigarrow_E B \equiv \mathbf{G}(A \cdot \mathbf{GF}E \rightarrow \mathbf{F}B).$$

Событие B следует за A, только если в ходе ожидания периодически наступает событие E, называемое модулем. Локальная задержка $\mathbf{lat}_E[B]$ определяется как число наступлений события E до ближайшего наступления события B (событие E, наступающее одновременно с B, не учитывается). $\mathbf{Lat}[A \rightsquigarrow_E B]$ определяется аналогично $\mathbf{Lat}[A \rightsquigarrow B]$, как максимум $\mathbf{lat}_E[B]$ на всем исполнении. Два типа задержки связаны простым соотношением:

$$\mathbf{Lat}[A \rightsquigarrow B] = \mathbf{Lat}[A \rightsquigarrow_{\mathrm{True}} B].$$

Задержка для сложного отношения отклика может быть сведена к вычислению более простых задержек. Например, пусть известно, что для некоторых событий A, B, C

$$\mathbf{Lat}[A \rightsquigarrow B] \leq l, \qquad \mathbf{Lat}[B \rightsquigarrow C] \leq k.$$

Тогда верно и то, что между появлением события A и последующим появлением события C проходит не более $l + k$ тактов, т.е.

$$\mathbf{Lat}[A \rightsquigarrow C] \leq l + k.$$

Найденную закономерность можно сформулировать в виде синтаксического правила вывода для задержки

$$\frac{\mathbf{Lat}[A \rightsquigarrow B] \leq l, \quad \mathbf{Lat}[B \rightsquigarrow C] \leq k}{\mathbf{Lat}[A \rightsquigarrow C] \leq k + l} \text{ (Seq), где Seq} - \text{имя правила}$$

Над чертой приводятся предпосылки, под чертой – следствие из них. В работе используется ряд таких правил, однако формат автореферата не позволяет приводить их полный перечень.

При построении оценки задержки с использованием синтаксических правил вывода, базовые отношения отклика и их задержки принимаются как предположения. Кроме того, применение тех или иных правил вывода требует дополнительных предположений о свойствах событий. На практике, принятые гипотезы могут оказаться ложными. Для проверки правильности оценки используются средства верификации моделей и метод ранжирующих функций.

Для двух произвольных событий A и B определим интервал от A до B как

$$A{:}\,B = A + \mathbf{pre}(A{:}\,B) \cdot \neg B.$$

Оператор **pre** возвращает значение $(A{:}\,B)$ на предыдущем такте (False в начальный момент времени). Выражение $(A{:}\,B)$ становится истинным при каждом наступлении A и остается таковым до первого наступления события B. Целочисленная неотрицательная функция ϕ называется ранжирующей функцией для отношения отклика $A \rightsquigarrow_E B$, если

$$\mathbf{G}\big((A{:}\,B) \cdot \neg B \rightarrow \phi^+ \leq \phi - [E]\big) \qquad (1)$$

Где $[E]$ – индикатор состояния E. Выражение ϕ^+ дает значение ϕ на следующем такте. Условие $(A{:}\,B) \cdot \neg B$ означает, что событие A уже наступило, а событие B – еще нет. Выражение (1) требует невозрастания ϕ при $E = \text{False}$ и строгого убывания ϕ при $E = \text{True}$. Значение ϕ можно считать “мерой” расстояния от текущего момента времени до наступления события B (по модулю события E).

Для обоснования оценки M на задержку достаточно построить ранжирующую функцию ϕ и доказать, что выполнено условие (1) и $\mathbf{G}(A \rightarrow \phi \leq \text{M})$. На практике часто удается

построить простые ранжирующие функции и проверить выполнение условий средствами одно- или двухшаговой индукции.

Аналогично задержке, вычисление ранжирующей функции для сложного отношения отклика может быть сведено к более простым случаям. Однако при комбинировании ранжирующих функций для разных отношений отклика возникает техническая трудность, связанная с неопределенностью поведения функции ϕ когда на ранжирующую функцию $\mathbf{rk}[A \rightsquigarrow_E B] \doteq \phi$ не накладывается ограничений при $\neg(A{:}\,B) + B$. Пусть:

$$\mathbf{rk}[A \rightsquigarrow B] \doteq \phi, \quad \mathbf{rk}[B \rightsquigarrow C] \doteq \psi.$$

По аналогии с правилом (Seq) для задержки, естественно было бы определить ранжирующую функцию для $A \rightsquigarrow C$ как $\phi + \psi$. Однако поведение функции ψ может быть произвольным на интервале от наступления события A до наступления события B (аналогично, для функции ϕ на интервале от B до C). Это не позволяет гарантировать убывания суммы $\phi + \psi$ на всем интервале от A до C. В таких случаях значение функций за пределами их интервала доопределяется с использованием двух специальных операторов: $\downarrow_{\mathrm{X}}$ и $\uparrow_{\mathrm{X}}^{\mathrm{M}}$.

Для целочисленной функции ϕ, условия X и постоянной M (значение, которым доопределяется функция) определим эти операторы как

$$\phi \downarrow_{\mathrm{X}} \equiv \phi \cdot [\neg \mathrm{X}], \quad \phi \uparrow_{\mathrm{X}}^{\mathrm{M}} \equiv \phi \cdot [\neg \mathrm{X}] + \mathrm{M} \cdot [\mathrm{X}]\,.$$

Заметим, что $\downarrow_{\mathrm{X}} = \uparrow_{\mathrm{X}}^{0}$.

Используя доопределяющие операторы, можно показать, что:

$$\frac{\mathbf{rk}[A \rightsquigarrow_E B] \doteq \phi, \quad \mathbf{rk}[B \rightsquigarrow_E C] \doteq \psi, \quad \mathbf{G}(B \Rightarrow \psi \leq \mathrm{M})}{\mathbf{rk}[A \rightsquigarrow_E C] \doteq \phi \downarrow_{B:C} + \ \psi \uparrow_{\neg(B:C)}^{\mathrm{M}}} \text{ (Rk-Seq)}$$

Аналогично (Rk-Seq), каждому правилу вывода для задержек можно поставить в соответствие правило вывода для ранжирующих функций (полный список правил не позволяет привести формат автореферата). В качестве гипотез при выводе принимаются базовые ранжирующие функции.

В **третьей главе** рассматривается среда моделирования xMAS. приводится описание семантики её примитивов и инфраструктурных расширений. Подробное описание приводится в связи с тем, что описания среды xMAS на русском языке не существует, а в дальнейшей части работы активно используются различные детали её реализации.

xMAS – это графический язык формального моделирования микроархитектуры. Модели xMAS строятся из небольшого набора стандартных блоков (примитивов), соединенных каналами для передачи пакетов с данными. Каждый канал соединяет ровно два примитива,

один из которых является инициатором передачи, а другой – ее получателем. Все примитивы модели работают синхронно по одному тактовому сигналу. В диссертации используется базовый набор примитивов, дополненный специализированными примитивами для моделирования качества обслуживания (рисунок 1).

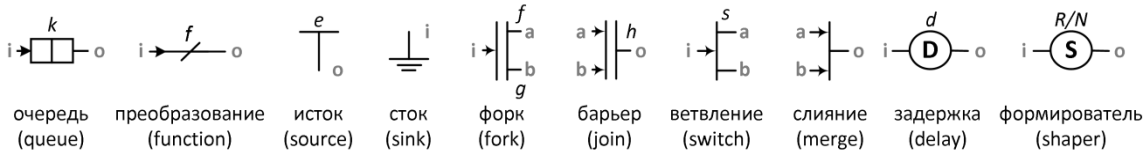

Рисунок 1. Набор примитивов xMAS, расширенный для моделирования задержек передачи данных

Для анализа задержек базовый набор примитивов xMAS необходимо расширить. Например, при моделировании качества обслуживания целесообразно использовать несколько разновидностей примитива слияния, отличающихся алгоритмом арбитража.

Для моделирования задержек обработки в набор примитивов языка xMAS был введен новый примитив ограниченной задержки (delay), задерживающий передачу каждого входного пакета ровно на d тактов. Аналогичным образом может быть смоделирована случайная задержка с верхней границей d.

Другой тип задержек реализуется формирователем потока (shaper). Формирователь гарантирует, что пакеты на выходе появляются не чаще, чем с заданной частотой ρ, задерживая при необходимости входные пакеты. Формирователь также может быть дополнительно параметризован величиной всплеска σ.

На рисунке 2 показан пример модели xMAS использующий как базовые примитивы, так и дополнительные примитивы качества обслуживания.

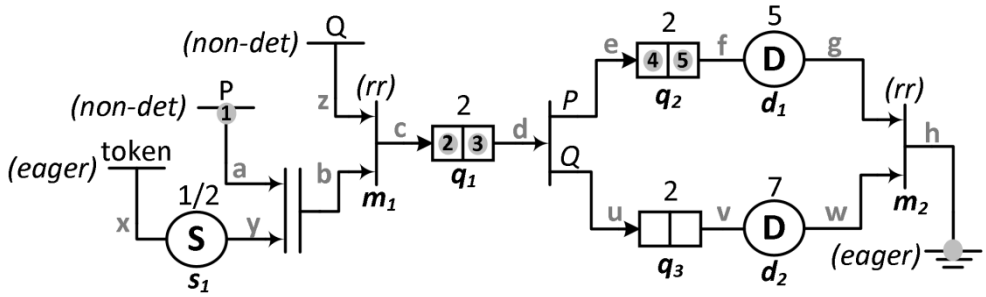

Рисунок 2. Модель коммуникационной фабрики.

Модель на рисунке 2 представляет простейшую коммуникационную фабрику, принимающую запросы от двух агентов P и Q. Пакеты принимаются фабрикой по одному за такт, что моделируется слиянием на входе. Обработка запросов конвейеризирована. На первой ступени запросы обрабатываются последовательно и хранятся в общей очереди. На второй ступени запросы P и Q обрабатываются параллельно с различными задержками.

Перед удалением из фабрики, запросы проходят завершающую фазу по одному за такт, что соответствует слиянию на выходе. Общая задержка обработки запроса P коммуникационной фабрикой определяется как количество тактов с момента появления запроса на канале a до момента его выхода из модели (передачи на канале h).

В **четвертой главе** описан способ применения формализма для вывода суждения о задержке (глава 2) к анализу микроархитектурных моделей на языке xMAS.

Абстрактные понятия предусловия и постусловия в отношениях отклика отображаются на сигналы каналов модели xMAS («готовность инициатора передачи», «готовность получателя передачи», «передаваемые даные»). Приводится список отношений отклика, соответствующих локальным (канальным) задержкам в определенной точке модели и достаточный для анализа задержек в моделях xMAS.

Опираясь на семантику примитивов xMAS и свойства канального протокола (описанного в тексте диссертации), для каждого из примитивов xMAS сформулированы правила “распространения” (propagation) задержек через примитив. Каждое правило распространения задержек соответствует последовательности применения базовых правил вывода и обладает рядом зависимостей – опирается на ранее выведенные задержки для отношений отклика на каналах примитива.

Вывод оценки на задержку заключается в распространении задержек через примитивы. Порядок распространения находится развертыванием зависимостей правил распространения.

При отсутствии управляющих циклов локальные задержки находятся последовательным обходом с использованием соответствующих правил распространения. Оценка для многошаговой транзакции получается суммированием локальных оценок по ее траектории.

Для моделей с управляющими циклами учитывается множество достижимых состояний системы, путём добавления предикатов к ветвям доказательства. Предикат растет по мере развертывания зависимостей между задержками и позволяет избавиться от циклических зависимостей, отсекая недостижимые ветви. Недостижимость ветви определяется на основе решения SAT-задачи для предиката ветви и автоматически генерируемых инвариантов, связывающих значения числа пакетов в различных очередях системы.

В моделях, с маршрутизацией пакетов (примитив switch), задержки, связанные с потоками пакетов различных значений, часто могут быть аппроксимированы по худшему случаю при отсутствии циклов. Анализ для циклических моделей, использующих маршрутизацию пакетов, требует раздельного анализа потоков пакетов различных значений.

В **пятой главе** приводятся и анализируются результаты тестирования разработанного метода для анализа задержек.

Разработанный метод анализа тестировался на двух группах примеров:

1) небольшие примеры, для сравнения с существующими средствами формальной верификации и оценки консервативности получаемых задержек;
2) большие модели, соответствующие микроархитектуре промышленных коммуникационных фабрик и лежащие далеко за пределами возможности существующих средств формальной верификации.

Эти результаты позволяют оценить рост временных затрат предлагаемого решения по мере усложнения модели.

В таблице 1 сравниваются оценки для канальных задержек в модели на рисунке 2.

Таблица 1. Сравнение оценок канальных задержек для модели на рисунке 2

Задержка	BMC	k-индукция	Правила $\mathcal{L}$	
f	6	6	6	Ожидание передачи на канале f
d	7	8	8	Ожидание передачи на канале d
a	13	17	21	Ожидание передачи на канале a
$a \rightsquigarrow$ h	32	34	56	Задержка по траектории *a-b-c-d-e-f-g-h*

На простом примере метод ограниченной верификации моделей (BMC) исследует все исполнения модели из начального состояния за ограниченное число тактов и позволяет получить оценку задержки снизу. k-индукция рассматривает исполнения модели из произвольного состояния (ограниченного дополнительными инвариантами) и дает консервативную оценку сверху.

Оценки для BMC и k-индукции оказываются приблизительно равными. Расхождение обусловлено случайной синхронизацией состояний примитивов (внутренних счетчиков задержек и формирователей, текущих значений приоритетов в алгоритмах арбитража и т.д.). Такие эффекты трудно учесть без полного перебора состояний модели.

Более интересны причины консервативности разработанного метода по сравнению с традиционной k-индукцией. Полная задержка, полученная по правилам вывода, почти вдвое больше оценки методом k-индукции. Расхождение можно сократить, повышая точность правил распространения, но полного совпадения не будет, т.к. оценка для предложенного метода получена по формуле, оценивающей время пребывания в каждой точке траектории по наихудшему случаю. В реальных исполнениях модели, наихудший случай возможен только на подмножестве точек траектории.

В таблице 2 сравниваются оценки для задержки поступления запросов в коммуникационную фабрику большего размера (sb_3e1r), получаемые с помощью метода k-индукции и предлагаемого подхода. В данном эксперименте варьируется время, требуемое

фабрикой на обработку запроса, и помимо границы на задержку приводится время, требуемое на её доказательство.

Таблица 2. Эффективность оценок задержки поступления запросов в фабрику для модели большего размера (sb_3e1r)

Задержка обработки (такты)	Время доказательства (сек.)		Доказываемая граница (такты)		Сравнительная точность
	k-индукция	Правила $\mathcal{L}$	*k*-индукция	Правила $\mathcal{L}$	
0	791	27	33	129	3.909
5	8439	28	68	169	2.485
10	37598	28	115	209	1.817
15	79854	29	144	249	1.729

Как видно из таблицы 2, использование ранжирующих функций даёт значительный выигрыш по производительности и существенно расширяет возможности верификации для сложных примеров.

При росте фактической границы на задержку, использование стандартных подходов на основе k-индукции показывает нелинейный рост времени доказательства, использование же предлагаемого подхода позволяет вывести и доказать задержку за неизменно малое время.

Дальнейшее увеличение задержки обработки в экспериментах таблицы 2 приводит к неприемлемому увеличению времени доказательства. Для задержки обработки равной 20 тактам, по истечении 290 тыс. секунд доказательство занимает всю доступную на вычислительном кластере оперативную память, после чего процесс доказательства замедляется из-за подкачки данных с дисков, и за 4 недели не удается получить результат.

При увеличении задержки обработки, погрешность предлагаемого метода уменьшается. В то время, как при малых задержках погрешность составляет 300-400%. Такое поведение приемлемо, т.к. увеличение задержки обработки запросов фабрикой фактически повышает нагрузку (congestion), а сценарии работы при малой нагрузке не представляют большого практического интереса. Из-за роста времени доказательства, получить с помощью метода k-индукции оценки для сценариев с реалистичными значениями задержек обработки не представляется возможным.

Большое расхождение оценок на задержку в режиме малой нагрузки объясняется тем, что предложенный алгоритм анализирует только структуру модели и не способен обнаружить состояния, которые недостижимы, например, из-за того, что в фабрику инжектируется недостаточное количество запросов, или запросы обслуживаются слишком быстро.

Для больших значений задержек обработки и более сложных моделей в качестве сравнительной оценки используются оценки на задержку, получаемые с помощью

имитационного моделирования на детерминированной тестовой последовательности, соответствующей одному из сценариев работы. Данный подход дает оценку снизу на границу задержки – минимальная величина задержки, которую не удается опровергнуть в ходе имитационного моделирования.

В таблице 3 приводятся оценки для задержки поступления запросов в модели коммуникационных фабрик различного размера, полученные с помощью предлагаемого метода и метода имитационного моделирования. Все эксперименты проводятся при значении задержки обработки запросов равной 40 тактам.

Таблица 3. Результаты оценок задержки поступления запросов в фабрику для различных моделей большого размера

модель				время док-ва (сек)	оценка на задержку (такты)		соотн-е оценок (сравн-я точность)
название	прими-тивов	оче-редей	простр-во состояний		Правила $\mathcal{L}$	имитац-е моделир-е	
sb_small	43	18	10^85	7	101	81	1.246
sb_order	66	22	10^103	23	129	81	1.593
sb_3e1r	435	63	10^499	29	449	237	1.895
sb_4e2r	722	102	10^907	117	1406	454	3.097
sb_3e3r	575	99	10^749	98	529	240	2.204
sb_5e3r	1045	141	10^1541	244	1875	601	3.119

Сравнительная точность для экспериментов, приведенных в таблице 3, для многих задач является достаточной. Кроме того, использование трасс выполнения, воспроизводящих различные сценарии работы, в дополнение к использованной тестовой последовательности должно радикально повысить точность оценки, получаемой в данной серии экспериментов.

По итогам экспериментов можно сказать, что для моделей с числом очередей порядка 10-100, оценка BMC часто далека от точной, т.к. исполнение модели с высокой задержкой может требовать длинной инициализирующей последовательности, а *k*-шаговая индукция не завершается за разумное время. В то же время предлагаемый подход позволяет за малое время получить оценки с приемлемой точностью и верифицировать их. Для простых же моделей большинство алгоритмов верификации позволяют получить результат за короткое время (порядка нескольких секунд) и сравнение их быстродействия не представляет интереса.

Заключение

Основные результаты работы

В ходе выполнения диссертационной работы получены следующие результаты:

1. Разработано расширение языка моделирования микроархитектуры xMAS, предназначенное для отражения в моделях свойств, связанных с качеством обслуживания. Набор стандартных примитивов языка был расширен примитивами для моделирования различных типов задержек и примитивами, реализующими специализированные алгоритмы арбитража;
2. Предложено математическое представление понятия задержки между событиями и формализация процесса суждения о задержке;
3. Разработана алгоритмическая схема анализа моделей xMAS для получения оценок сверху на задержку передачи данных из одной точки в другую;
4. Предложен подход к построению масштабируемых доказательств для полученных временных оценок с помощью метода k-индукции, в котором глубина индукции не зависит от величины оценки.
5. Продемонстрированы возможности использования предложенного подхода к выводу и доказательству задержки в микроархитектурных моделях xMAS, размеры которых делают невозможными применение существующих средств формальной верификации. Проанализированы недостатки и ограничения применения предложенного подхода.

На основании решенных в работе задач можно определить **рекомендации и перспективы дальнейшей разработки темы**:

1. Проработка известных и поиск новых путей снижения консервативности оценок задержки, получаемых с помощью предложенного подхода.
2. Адаптация разработанных алгоритмов для анализа синхронных сетей более общего вида, не являющихся моделями xMAS.
3. Обобщение предложенного подхода для анализа других свойств качества обслуживания, например, пропускной способности, или вариации задержки

Публикации по теме диссертации

1. **Kishinevsky M., Gotmanov A., Viktorov Y.** Challenges in Verifying Communication Fabrics // LNCS. – 2011. – V. 6898. – P. 18-21.
2. **Викторов Ю.О.** Анализ и оптимизация качества обслуживания в системах на кристалле // труды 37-ой молодежной научной конференции «Гагаринские чтения». – 2011. – Т. 4. – С. 42-44.
3. **Викторов Ю.О., Готманов А.Н.** Анализ задержек в микроархитектурных моделях коммуникационных фабрик // Проблемы разработки перспективных микро- и наноэлектронных систем - 2012. Сборник трудов / под общ. ред. академика РАН А.Л. Стемпковского. – М.: ИППМ РАН, 2012. – С. 67-72.
4. **Викторов Ю.О., Готманов А.Н.** Проблемы качества обслуживания при проектировании коммуникационных фабрик в системах на кристалле // Информационные технологии. – 2012. – № 11(195). – С. 15-20.
5. **Викторов Ю.О., Готманов А.Н.** Верификация задержки в микроархитектурных моделях коммуникационных фабрик // Информационно-управляющие системы. – 2012. – №6(61). – С. 43-52.