МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

На правах рукописи

Вдовин Павел Максимович

ЖАДНЫЕ АЛГОРИТМЫ И СТРАТЕГИИ ОГРАНИЧЕННОГО ПЕРЕБОРА ДЛЯ ПЛАНИРОВАНИЯ ВЫЧИСЛЕНИЙ В СИСТЕМАХ С ЖЕСТКИМИ ТРЕБОВАНИЯМИ К КАЧЕСТВУ ОБСЛУЖИВАНИЯ

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

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2016

Работа выполнена на кафедре автоматизации систем вычислительных комплексов факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова».

Научный руководитель: **Костенко Валерий Алексеевич**,
кандидат технических наук, ведущий научный сотрудник Лаборатории Вычислительных Комплексов факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова»

Официальные оппоненты: **Цурков Владимир Иванович**,
доктор физико-математических наук, профессор, заведующий отделом сложных систем Вычислительного центра им. А.А. Дородницына ФИЦ ИУ РАН,

Мокряков Алексей Викторович,
кандидат физико-математических наук, доцент кафедры «Прикладная математика, информационные технологии и электротехника» Института информационных систем и технологий Московского Авиационного Института

Ведущая организация: Акционерное общество «Научно-исследовательский институт вычислительных комплексов им. М.А. Карцева».

Защита диссертации состоится 23 декабря 2016 г. в 11 час. 00 мин. на заседании диссертационного совета Д 501.001.44 при Московском государственном университете имени М. В. Ломоносова по адресу: 119991, г. Москва, ГСП-1, Ленинские горы, д. 1, стр. 52, факультет вычислительной математики и кибернетики, аудитория 685.

С диссертацией можно ознакомиться в Научной библиотеке Московского государственного университета имени М.В. Ломоносва по адресу: 119192, г. Москва, Ломоносовский проспект, д. 27, а также на официальном сайте факультета ВМК МГУ https://cs.msu.ru в разделе «Диссертации».

Автореферат разослан “_____”_____2016 г.

Ученый секретарь
диссертационного совета Д 501.001.44
доктор физико-математических наук, доцент О.В. Шестаков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

В данной работе рассматривается задача планирования вычислений в распределенных системах с жесткими требованиями к качеству обслуживания. Рассматриваемые системы состоят из физических элементов (вычислительных ресурсов и ресурсов хранения данных) и сети обмена данными, состоящей, в свою очередь, из каналов передачи данных и сетевых элементов (коммутаторов и маршрутизаторов).

Задача планирования вычислений заключается в назначении как можно большего числа запросов в распределенной системе. Запросы состоят из виртуальных элементов (виртуальных машин и виртуальных систем хранения данных), которые должны быть размещены на физических элементах распределенной системы, и виртуальных каналов, для которых должны быть проложены маршруты передачи данных между размещенными виртуальными элементами в сети обмена данными с учетом заданных требований к качеству обслуживания.

Ограничения на размещение виртуальных ресурсов (виртуальных элементов и виртуальных каналов) зависят от конкретной решаемой задачи и используемых стандартов/протоколов. Основными ограничениями являются непревышение максимального суммарного значения по каждой характеристике ресурсов для размещаемых на физических элементах виртуальных элементов, ограничения на максимальную пропускную способность каналов передачи данных и сетевых элементов, ограничения на длительность передачи сообщений, передаваемых между виртуальными элементами по виртуальным каналам. Кроме того, могут быть выделены технологические ограничения, обусловленные используемыми стандартами, а также особенностями аппаратных и программных средств системы.

В данной работе задача планирования вычислений решается для двух классов систем: центров обработки данных (ЦОД) с моделью обслуживания Infrastructure-as-a-Service (IaaS) и бортовых сетей на основе стандарта Avionics Full Duplex Ethernet (AFDX).

Модель обслуживания IaaS подразумевает, что планируемыми ресурсами являются виртуальные элементы: виртуальные машины и виртуальные

системы хранения данных (storage-элементы), и виртуальные каналы, объединяющие эти элементы. Основными требованиями, задаваемыми при решении задачи планирования вычислений в ЦОД, являются требования к качеству обслуживания (SLA), которые являются параметрами ограничений на возможные размещения виртуальных ресурсов. В процессе снятия/назначения виртуальных запросов может возникать фрагментация ресурсов, в результате чего новые запросы не могут быть назначены без нарушения SLA. Многие из используемых в настоящее время алгоритмов, такие как жадный алгоритм случайного размещения, используемый в системе Openstack, из-за возникающей фрагментации не позволяют эффективно использовать свободные физические ресурсы системы.

Бортовые сети на основе стандарта AFDX используются в информационно-управляющих системах реального времени (ИУС РВ) для передачи данных между абонентами. При построении сетей AFDX в качестве виртуальных запросов выступают периодически передаваемые сообщения, которые должны передаваться не менее одного раза в период и основными требованиями к которым являются ограничения не только максимальной длительности передачи, но и максимального времени отклонения между соседними экземплярами сообщений. Для обеспечения этих требований сообщения передаются через выделенные соединения –виртуальные каналы, для которых фиксирован маршрут передачи и зарезервирована пропускная способность. Зная параметры виртуальных каналов, можно с помощью специальных техник оценить длительность передачи сообщений и обеспечить выполнение требований.

Актуальность разработки алгоритмов планирования вычислений в распределенных системах с жесткими требованиями к качеству обслуживания подтверждается особенностями задачи, которые отличают ее от других задач планирования вычислений, а именно:

- Необходимость согласованного планирования всех типов ресурсов с учетом ряда жестких требований к качеству обслуживания;
- Необходимость решения нескольких подзадач в процессе решения задачи планирования вычислений, а именно задачи назначения виртуальных элементов на физические элементы распределенной

системы, задачи формирования виртуальных каналов и задачи построения маршрутов виртуальных каналов с учетом заданных ограничений.

Актуальность также подтверждается отсутствием в свободном доступе работ, которые в полной мере позволяют решать поставленные задачи планирования вычислений в ЦОД с моделью обслуживания IaaS и задачи построения сетей AFDX.

Цель работы

Целью данной работы является разработка и исследование алгоритмов, построенных на сочетании жадных стратегий и ограниченного перебора, для решения задач планирования вычислений в распределенных системах с жесткими требованиями к качеству обслуживания.

Для достижения поставленной цели необходимо решить следующие задачи:

- Разработать и исследовать алгоритм планирования вычислений в центрах обработки данных с моделью обслуживания Infrastructure-as-a-Service, позволяющий производить назначение виртуальных запросов с заданными требованиями к качеству обслуживания.
- Разработать и исследовать алгоритм построения соединений для сетей со статически конфигурируемыми коммутаторами, позволяющий для заданного множества периодически передаваемых сообщений строить виртуальные каналы и маршруты для них с учетом заданных ограничений.
- На основе предложенного алгоритма решения задачи построения соединений для сетей, основанных на стандарте Avionics Full Duplex Ethernet, разработать программное инструментальное средство, позволяющее решать поставленную задачу и допускающее настройку на особенности конкретной задачи путем замены отдельных модулей.

Основные результаты работы

Основные результаты диссертационной работы заключаются в следующем:

1. Построен алгоритм согласованного отображения множества запросов на физические ресурсы центров обработки данных, позволяющий

выбирать соотношение между вычислительной сложностью алгоритма и подмножеством размещенных запросов. Получено аналитическое выражение для вычислительной сложности алгоритма в зависимости от глубины ограниченного перебора, обоснована корректность алгоритма, сформулированы условия, при которых алгоритм размещает все запросы.

2. Разработан алгоритм построения соединений для сетей со статически конфигурируемыми коммутаторами с соблюдением ограничений на качество работы сети. Получено аналитическое выражение для вычислительной сложности алгоритма в зависимости от глубины ограниченного перебора, обоснована корректность алгоритма, сформулированы условия, при которых алгоритм строит все соединения.
3. На основе предложенного алгоритма построения соединений для сетей, основанных на стандарте Avionics Full Duplex Ethernet, разработано программное инструментальное средство, предоставляющее возможность запуска алгоритма с различными параметрами используемых процедур и возможность модульной настройки на особенности конкретной задачи.
4. Для построенных алгоритмов проведен экспериментальный анализ влияния жадный критериев и используемых процедур на качество решения. Выявлены наилучшие критерии в зависимости от класса исходных данных, показано улучшение качества получаемого решения при использовании процедуры агрегации соединений и процедуры репликации.

Научная новизна

В диссертационной работе разработаны и исследованы алгоритмы планирования вычисления для двух классов систем, а именно для центров обработки данных (ЦОД) с моделью обслуживания Infrastructure-as-a-Service (IaaS) и для сетей на основе стандарта Avionics Full Duplex Ethernet (AFDX). Алгоритмы основаны на сочетании жадных стратегий и стратегий ограниченного перебора, которые позволяют задавать баланс между сложностью и точностью алгоритма. В алгоритме планирования вычислений

в ЦОД используется возможность создавать реплики назначенных виртуальных систем хранения данных для оптимизации получаемого решения.

Для каждого алгоритма обоснована корректность, получена оценка вычислительной сложности. Доказаны ряд утверждений о точности отдельных процедур алгоритмов и о достаточных условиях назначения всех запросов.

Практическая ценность

В рамках диссертационной работы на основе алгоритма построения сетей AFDX разработана инструментальная система. Инструментальная система представляет возможность запуска алгоритма с различными параметрами, возможность интерактивного редактирования входных данных и отображения результатов работы алгоритма, а также возможность модульной настройки на особенности конкретной задачи.

Полученные в работе результаты могут быть рекомендованы к использованию при проектировании бортовых сетей на основе стандарта AFDX и при построении планировщиков ресурсов облачных платформ.

Методы исследования

При получении основных результатов диссертации использовались методы математического программирования, теории графов, теории расписаний и математической статистики.

Апробация работы

Результаты работы докладывались на научно-исследовательском семинаре кафедры автоматизации систем вычислительных комплектов (АСВК) факультета ВМК МГУ, а также на следующих конференциях:

1. VII международная конференция «Исследование Операций (ORM2013)» (Москва, 2013);
2. Международная научная конференция «Управление и виртуализация в современных сетях (Сети 2014: SDN & NFV)» (Москва, 2014);
3. XIII международная научная конференция «Параллельные компьютерные технологии» (PaCT 2015)» (Петрозаводск, 2015);
4. II международная конференция «Инжиниринг & Телекоммуникации (En&T)» (Долгопрудный, 2015).

5. Всероссийской научно-технической конференции «Авиационные системы в XXI веке» (Москва, 2016).

Личный вклад автора заключается в выполнении основного объема теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку алгоритмов, построение их программных реализаций, проведение экспериментальных исследований, формулировку и доказательство приведенных утверждений, анализ и оформление результатов в виде публикаций и научных докладов.

Публикации

По теме диссертации опубликовано 9 печатных работ, в том числе работы в журналах «Известия РАН. Теория и системы управления», «Вестник МГУ. Вычислительная математика и кибернетика», «Труды МФТИ», входящих в перечень ведущих научных журналов ВАК РФ. Список работ приведен в конце автореферата.

Свидетельства о регистрации программ для ЭВМ

В рамках диссертационной работы получено одно свидетельство о регистрации программы для ЭВМ.

Структура и объем диссертации

Диссертация состоит из введения, шести глав, заключения, списка литературы и четырех приложений. Полный объем диссертации составляет 127 страниц с 16 рисунками и 13 таблицами. Список литературы содержит 70 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность диссертационной работы и дано краткое описание задач, решаемых в работе.

В **первой главе** представлено содержательное описание и математическая постановка задач планирования вычислений в ЦОД с моделью IaaS и построения сетей AFDX.

В *разделе 1.1* приводится описании общей задачи планирования вычислений в распределенных системах, а именно описание входных и выходных данных и ограничений. Распределенная система описывается в виде размеченного графа, в котором вершинами являются физические

элементы (вычислительные узлы и системы хранения данных) и сетевые элементы: коммутаторы, маршрутизаторы и оконечные системы, являющиеся интерфейсами между выдающими и принимающими данные абонентами и непосредственно сетью обмена данными. Ребрами (или дугами) графа являются каналы передачи данных.

Каналы передачи данных и сетевые элементы составляют сеть обмена данными. Для сетевых и физических элементов задаются характеристики в виде вектора доступных ресурсов.

Входными данными задачи планирования вычислений являются виртуальные запросы, для которых заданы ограничения. Виртуальные запросы могут состоять из следующих элементов:

- Виртуальные элементы: виртуальные машины и storage-элементы;
- Коммуникационные каналы между виртуальными элементами, которые могут задаваться следующим образом:
    - в виде виртуальных каналов;
    - в виде периодически передаваемых сообщений, для которых требуется сформировать виртуальные каналы.

Для входных данных задаются требования качества обслуживания, которые должны выполняться гарантированно. Это достигается введением ограничений на возможные варианты размещения виртуальных ресурсов на физические: суммарное значение по каждой характеристике размещенных виртуальных элементов не должно превышать значения физического элемента по данной характеристике. Кроме того, могут задаваться требования на длительность передачи сообщений.

В некоторых случаях (например, в задаче построения сетей AFDX) для виртуальных каналов требуется задать определенные параметры, по которым определяется резервируемая пропускная способность. Кроме того, по виртуальным каналам могут передаваться несколько сообщений. В таких случаях после того, как известно, на какие физические элементы назначены виртуальные элементы, требуется шаг формирования виртуальных каналов с целью определения их параметров.

Задача заключается в назначении как можно большего числа виртуальных запросов и состоит из следующих подзадач:

- Назначение виртуальных элементов на физические с учетом требований к качеству обслуживания.
- Формирование виртуальных каналов и определение значений их параметров для периодически передаваемых сообщений между назначенными виртуальными элементами.
- Построение маршрутов для сформированных виртуальных каналов с учетом требований к пропускной способности и длительность доставки сообщений.

В *разделе 1.2* приводится постановка задачи планирования вычислений в ЦОД с моделью обслуживания IaaS.

Модель физических ресурсов ЦОД задается графом $H = (P \cup M \cup K, L)$, где P – множество вычислительных узлов, M – множество систем хранения данных, K – множество сетевых элементов сети обмена ЦОД, L – множество физических каналов передачи данных. На множествах P, M, K и L определены векторные функции, аргументами которых являются соответственно элементы множеств P, M, K или L. Значениями функций выступают характеристики соответствующего вычислительного узла, системы хранения данных, сетевого элемента или канала передачи данных:

- $(ph_1, ph_2, \ldots, ph_{n1}) = fph(p), p \in P$,
- $(mh_1, mh_2, \ldots, mh_{n2}) = fmh(m), m \in M$,
- $(kh_1, kh_2, \ldots, kh_{n3}) = fkh(k), k \in K$,
- $(lh_1, lh_2, \ldots, lh_{n4}) = flh(l), l \in L$.

Предполагается, что для сетевых элементов и каналов передачи данных используются одинаковые наборы характеристики.

Ресурсный запрос задается графом $G = (W \cup S, E)$, где W – множество виртуальных машин, используемых приложениями, S – множество виртуальных систем хранения данных (storage-элементов), E – множество виртуальных каналов передачи данных между виртуальными машинами и storage-элементами запроса. На множествах W, S и E определены векторные функции, аргументы которых – соответствующие элементы множеств W, S или E. Значениями функций являются характеристики, запрашиваемые для виртуальной машины, storage-элемента или виртуального канала:

- $(wg_1, wg_2, \ldots, wg_{n1}) = fwg(w), w \in W$,
- $(sg_1, sg_2, \ldots, sg_{n2}) = fsg(s), s \in S$,
- $(eg_1, eg_2, \ldots, eg_{n4}) = feg(e), e \in E$.

Набор характеристик SLA элемента запроса совпадают с набором характеристик соответствующего ему физического ресурса.

Назначением ресурсного запроса будем называть отображение

$A: G \to H = \{W \to P, S \to M, E \to \{K, L\}\}$.

Выделим три типа отношений между характеристиками запросов и соответствующими характеристиками физических ресурсов. Обозначим через x_i i-ю характеристику запроса и через y_j – соответствующую ей характеристику физического ресурса j. Тогда эти отношения можно записать следующим образом.

1. Недопустимость перегрузки емкости физического ресурса:

$$\sum_{i \in R_j} x_i \le y_j,$$

здесь R_j – множество запросов, назначенных на выполнение на физическом ресурсе j, x_i – соответствующая характеристика назначенного запроса.

2. Соответствие типа запрашиваемого ресурса типу физического ресурса:

$x_i = y_j$.

3. Наличие требуемых характеристик у физического ресурса:

$x_i \le y_j$.

Для каждой характеристики задан используемый тип отношения t_i, $t_i = \{1, 2, 3\}$.

Отображение $A: G \to H = \{W \to P, S \to M, E \to \{K, L\}\}$ *будем называть корректным*, если для всех физических ресурсов для любой характеристики i выполняются отношения t_i.

Отметим, что на каждый вычислительный узел (на каждую систему хранения данных) может быть назначено такое множество виртуальных машин (виртуальных систем хранения данных), для которого выполняются отношения t_i для всех характеристик.

В качестве *исходных данных* задачи назначения ресурсных запросов на физические ресурсы заданы:

- множество ресурсных запросов $Z = \{G_i\}$, поступивших планировщику;
- граф доступных ресурсов $H_{res} = (P \cup M \cup K, L)$.

Требуется: из множества Z разместить на выполнение в ЦОД максимальное число запросов, таких, что отображение A является корректным.

В случае, когда невозможно построить маршрут к некоторому storage-элементу, возможен вызов процедуры репликации, позволяющей дублировать данные на нескольких физических хранилищах данных. Данная проблема возникает, когда имеются виртуальные хранилища данных с высокой интенсивностью считывания и низкой интенсивностью записи. В этом случае для обеспечения консистентности данных не требуется канал с высокой пропускной способностью, и часть приложений может работать с виртуальной системой хранения данных, а другая – с его копией (или копиями), которая размещается на другой физической системе хранения данных.

Репликацией называется отображение $R: H \to H$, дублирующее данные некоторого $s \in S$, назначенного на систему хранения данных $m \in M$, и создающее виртуальный канал поддержки консистентности данных $(m', l_1, k_1, ..., k_{n-1}, l_n, m): k_i \in K, l_i \in L, m' \in M$, m' – система хранения данных с репликой storage-элемента *s*.

Если storage-элемент является репликацией некоторого виртуальной системы хранения данных *s*, то в граф запрашиваемых ресурсов *G* добавляется вершина *s′*. В граф *G* также добавляется виртуальный канал между вершинами *s* и *s′*, пропускная способность которого определяется исходя из требования обеспечения консистентности репликации и базы данных.

На вход алгоритму назначения запросов на физические ресурсы подаются граф доступных ресурсов H_{res} и множество ресурсных запросов $\{G_i\}$. Множество $\{G_i\}$ формирует диспетчер задач. В него, кроме вновь поступивших запросов, могут входить и запросы, которые выполняются и

для которых допустима миграция. Диспетчер задач также определяет время запуска планировщика.

Выходом алгоритма назначения запросов на физические ресурсы является множество назначений ресурсных запросов на физические ресурсы $\{A_i : G_i \to H, i = \overline{1,n}\}$ и множество репликаций $\{R_i\}, i = 0,1,\ldots$.

В *разделе 1.3* представлены описание стандарта AFDX и постановка задачи построения сетей AFDX.

Согласно стандарту AFDX, сеть состоит из следующих элементов:

- абоненты, передающие сообщения;
- оконечные системы – системы, обеспечивающие интерфейс между абонентами и сетью;
- пакетные коммутаторы, соединяемые каналами передачи данных.

Каждый абонент подключен к оконечной системе, причем к одной оконечной системе может быть подключено несколько абонентов. Соблюдение ограничений на время передачи сообщений в сети AFDX достигается за счет выделения гарантированной пропускной способности соединения между каждой парой оконечных систем. Такое соединение может проходить через несколько пакетных коммутаторов и линий передачи данных. В AFDX соединение между оконечными системами называют виртуальным каналом. Передача данных между абонентами осуществляется путем передачи сообщений по виртуальным каналам, маршруты которых в физической сети определены заранее. Для каждого виртуального канала определена только одна оконечная система-отправитель и одна или более оконечная система-получатель. При этом по одному виртуальному каналу могут передаваться сообщения только от одного абонента.

Важным параметром при построении виртуальных каналов является джиттер – разность между реальным временем начала выдачи кадра и временем начала выдачи, определенным согласно периоду. Джиттер возникает при выдаче кадров, на которые делятся сообщения, в сеть обмена данными. На виртуальных каналах пропускная способность регулируется с помощью разбиения сообщений на кадры и передачи кадров с заданной частотой. Из-за мультиплексирования кадров разных виртуальных каналов кадры не могут быть переданы строго периодически, и отклонения от начала

периода регулируются с помощью задания максимального джиттера виртуального канала.

Сеть AFDX будем описывать ориентированным графом $G = (N \cup K, E)$, где N – множество оконечных систем, K – множество коммутаторов, E – множество дуг, соединяющих элементы из множества $N \cup K$.

Входными данными задачи являются потоки данных между абонентами, которые задаются в виде множества периодически передаваемых сообщений $MSG=\{msg\}$, для каждого из которых задаются размер сообщения $size_{msg}$, период передачи T_{msg}, максимальный джиттер порождения сообщения внутри периода J_{msg}, отправитель $src_{msg} \in A$ и множество получателей $\{dst_{msg}\} \subset A$.

Для каждого сообщения заданы ограничения τ_{msg} на максимальную длительность передачи сообщения с момента выдачи до момента получения сообщения всеми абонентами-получателями и $J*_{msg}$ – максимальный джиттер передачи сообщения (разность между возможными максимальной и минимальной длительностью передачи сообщения).

Задача заключается в построении множества виртуальных каналов $VL=\{vl\}$ и построении для них маршрутов в сети обмена данными. Для каждого виртуального канала должны быть определены следующие параметры: максимальный размер кадра LM_{vl}, минимальный промежуток времени между передачей кадров при нулевом джиттере порождения кадров BAG_{vl} (от 1 до 128 мс и является степенью двойки), максимальный джиттер порождения кадров JM_{vl}, отправитель и получатели кадров и маршрут передачи кадров в сети.

При построении виртуальных каналов и маршрутов для них должны выполняться следующие ограничения:

- Суммарная пропускная способность, зарезервированная под виртуальные каналы, проходящие через физический канал передачи данных e, не превосходит его пропускной способности:

$$\forall e \in E : \sum_{vl \in e} \frac{LM_{vl}}{BAG_{vl}} \le R_e.$$

Здесь R_e – пропускная способность физического канала передачи данных e.

- Частота передачи кадров каждого виртуального канала не превосходит частоты выдачи кадров в канал:

$$\forall vl \in VL: \sum_{msg \in MSG_{vl}} \left(\left\lceil size_{msg} / (LM_{vl} - c) \right\rceil / T_{msg} \right) \le 1 / BAG_{vl} .$$

Данное ограничение возникает из того, что все кадры одного сообщения *msg* должны поступить в канал до выдачи следующего сообщения *msg*, то есть за период T_{msg}.

- Максимальный джиттер на оконечных системах-отправителях, согласно требованиям стандарта, не должен превосходить 500 мкс:

$$\forall vl \in VL: JM_{vl} \le 0.5 .$$

- Максимальная длительность передачи каждого сообщения и максимальный джиттер не превосходят заданных ограничений:

$$\forall msg \in MSG: Dur(msg) \le \tau_{msg}; Jit(msg) \le J^*_{msg} ,$$

здесь *Dur*(*msg*) и *Jit*(*msg*) – процедуры вычисления длительности передачи сообщений и их джиттеров соответственно.

Во **второй главе** приведен обзор возможных подходов к решению не только описанных задач, но и некоторых близких задач и подзадач, которые могут возникать в процессе решения.

В *разделе 2.1* приведены существующие алгоритмы решения задачи планирования вычислений в ЦОД с моделью обслуживания IaaS и близких задач, таких как задача вложения графов и задача построения виртуальных сетей. Основное отличие рассмотренных работ, посвященных указанным близким задачам, от поставленной задачи – отсутствие требований к качеству обслуживания. Любое решения является корректным, и среди них производится поиск оптимального по некоторому параметру. Другое отличие заключается в том, что в указанных задачах допускается назначение не более одного виртуального элемента на любой физических элемент. Ни один из исследованных алгоритмов не позволяет решать все возникающие подзадачи в совокупности.

При решении задачи планирования ресурсов в ЦОД с различными планировщиками ресурсов задача разделяется на подзадачу назначения виртуальных элементов (виртуальных машин и storage-элементов) на физические элементы (вычислительные узлы и системы хранения данных соответственно) и на подзадачу построения маршрутов виртуальных каналов. Первая подзадача соответствует многомерной задаче упаковке в контейнеры, существующие решения которой приведены в *разделе 2.2*. Задача построения маршрутов виртуальных каналов описана в *разделе 2.3*. Некоторые из приведенных подходов использованы при решении поставленной задачи.

Подходы к решению задачи построения сетей AFDX описаны в *разделе 2.4,* согласно исследованию которых сделан вывод, что решение задачи в поставленном виде полностью не представлено, однако существуют подходы к решению отдельных этапов, таких как настройка параметров *BAG, LM* и построение маршрутов виртуальных каналов.

Одним из этапов решения задачи является оценка длительности передачи сообщений, подходы к решению которой описаны в *разделе 2.5*.

В **третьей главе** представлено описание разработанного алгоритма планирования вычислений в ЦОД с моделью обслуживания IaaS. Алгоритм состоит из двух основных шагов: назначение виртуальных машин на вычислительные узлы и назначение storage-элементов на физические системы хранения данных, которые соответствуют многомерной задаче упаковки в контейнеры, и построение маршрутов виртуальных каналов.

Предлагаемый алгоритм основан на сочетании жадных стратегий и стратегий ограниченного перебора. Выбор очередного назначаемого запроса, элемента (виртуального или физического) или виртуального канала для назначения выбирается согласно жадным критериям, при неуспешной попытке назначения вызывается процедура ограниченного перебора, которая заключается в попытке переназначения уже назначенных элементов с целью их переназначения таким образом, чтобы можно было назначить текущий элемент. Для ограниченного перебора задается числом m, которое указывает максимальную мощность множества, среди элементов которых можно производить переназначение.

Построение маршрутов виртуальных каналов производится после того, как получены назначения виртуальных машин и storage-элементов. При построении используется алгоритм построения *k* кратчайших путей в графе. Данный алгоритм позволяет использовать два критерия при построении: каждый из кратчайших маршрутов строится по первому критерию, далее среди не более чем *k* построенных маршрутов по второму критерию выбирается наилучший маршрут.

В случае неуспеха построения маршрута вызывается процедура репликации, которая создает реплику для storage-элемента с целью устранения перегрузки входящих в физическую систему хранения данных каналов обмена. При создании реплики необходимо также проложить канал для поддержания консистентности между storage-элементом и его репликой.

Следует отметить, что процедура репликации в данном алгоритме используется с целью оптимизации назначения виртуальных каналов, т.е. для создания дополнительных возможных путей при назначении запросов.

В главе приведена оценка вычислительной сложности алгоритма, в том числе в зависимости от выбранной глубины процедуры ограниченного перебора. Обоснована корректность алгоритма. Приведены и доказаны следующие утверждения.

Утверждение 3.1. При одномерной упаковки в контейнеры для любого *k* для асимптотической точности *a* жадного алгоритма упаковки с процедурой ограниченного перебора глубины *k* верно: $11/9 \le a \le 1{,}7$.

Под асимптотическая точностью процедуры в данном случае понимается отношение в наихудшем случае количества контейнеров, назначаемое процедурой, к количеству контейнеров, получаемому оптимальным алгоритмом, при стремлении количества элементов к бесконечности.

Пусть производится построение маршрута виртуального канала *vl* с требуемой пропускной способностью *r*(*vl*) и вызывается процедура ограниченного перебора. Пусть *G** граф – доступных ресурсов для назначения *vl*, *G'* – граф *недоступных* ресурсов, *H** – граф, включающий в себя все возможных маршруты без циклов для *vl*, *H* – пересечение графов *G'* и *H**.

Утверждение 3.2. Пусть производится переназначение множества виртуальных каналов $V = \{v\}$ и для всех элементов h из множества H верно: $\sum_{v \in A(h), v \in V} r(h) + r(v) < r(vl)$, где суммирование ведется по всем виртуальным элементам, назначенным на элемент h. Тогда переназначение множества V вместе с элементом vl приведет к неуспеху.

Пусть N – количество физических контейнеров, каждый из которых описывается n характеристиками: $r(e_i) = (r_{i,1}, r_{i,2}, \ldots r_{i,n})$, $i = 1..N$. Пусть производится назначение M виртуальных элементов, каждый из которых также описывается n характеристиками: $v(v_i) = (v_{i.1}, v_{i.2}, \ldots v_{i.n})$, $i = 1..M$, причем для каждой характеристики известны ограничения сверху: $\forall j : v_{i,j} \le vm_j$.

Утверждение 3.3. Пусть значение каждой характеристики ограничено снизу: $\forall j : r_{i,j} \ge rm_j$. Если верно $M \le N \cdot \min_{j \le n} \left\lfloor \frac{rm_j}{vm_j} \right\rfloor$, то алгоритм назначит все виртуальные элементы.

Утверждение 3.4. Пусть входом задачи является ЦОД с древовидной топологией, причем для вычислительных элементов, систем хранения данных и, соответственно, виртуальных машин и storage-элементов выполнены условия утверждения 3.3. Пусть, кроме того, выполнены следующие условия:

- для каждой характеристики физических элементов-контейнеров известны ограничения сверху: $\forall j : r_j \le rn_j$;
- для каждой характеристики виртуальных элементов известны ограничения снизу: $\forall j : v_j \ge vn_j$, причем $\exists j : vn_j > 0$;
- каждый виртуальный элемент присутствует не более чем в k виртуальных каналов;
- каждый виртуальный канал требует пропускную способность, не превышающую RM.
- физическая топология ЦОД представляет собой древовидную структуру глубины l, обладающую следующими характеристиками:

- каждый канал передачи данных уровня $i \le l$ имеет пропускную способность не ниже R_i;
- каждый коммутатор уровня $i \le l$ имеет пропускную способность не ниже K_i;
- каждый уровень i разбивается на кластеры (при этом уровень 0 соответствует вычислительным узлам или системам хранения данных, на остальных уровнях находятся коммутаторы), в кластере находится не более D_{i+1} элементов. При этом коммутаторы уровня i имеют не менее D_i каналов передачи данных[1], идущих к уровню i–1.

Тогда все запросы будут назначены при выполнении следующих условий:

- $R_1 \ge RM \cdot k \cdot \min\limits_{j \le n} \left\lfloor \dfrac{rn_j}{vn_j} \right\rfloor$;
- $K_i \ge D_i \cdot R_i$, $1 < i \le l$;
- $R_{i+1} = K_i$, $1 < i \le l$.

В конце главы приводятся замечания, показывающие, почему утверждения 3.3 и 3.4 могут быть актуальны для реальных ЦОД.

В **четвертой главе** представлено описание алгоритма построения сетей AFDX. Алгоритм состоит из следующих шагов: формирование виртуальных каналов для заданных периодически передаваемых сообщений, построение маршрутов виртуальных каналов, вычисление оценки длительности и джиттера передачи сообщений и итеративное переконфигурирование виртуальных каналов в случае нарушения ограничений.

На первом этапе для каждого сообщения формируется один виртуальный канал. Параметры *BAG* и *LM* формируются согласно предложенному ранее[2] подходу, при котором решается задача минимизации пропускной способности исходя из заданных ограничений. Предложенный подход

[1] Предполагается, что в «строго древовидной» топологии корневые коммутаторы имеют ровно D_{l+1} исходящих каналов передачи данных, остальные коммутаторы имеют ровно $D_i + 1$ (один из каналов идет на уровень выше). В топологии fat tree могут добавляться избыточные каналы передачи данных.

[2] Подход предложен в работе Al Sheikh A. et al. Optimal design of virtual links in AFDX networks // Real-Time Systems. 2013. Т. 49. N. 3. P. 308-336.

расширяется в диссертационной работе с учетом ограничений на джиттер J_{msg} и период T_{msg} формирования сообщения.

На каждой оконечной системе проверяется выполнение ограничения на максимальный длиттер передачи кадров, и в случае его невыполнения вызывается процедура агрегации. Данная процедура объединяет сообщения от одного абонента в один виртуальный канал, при этом выполняется процедура, схожая с формированием параметров *BAG* и *LM* в случае передачи одного сообщения по виртуальному каналу.

На следующее этапе производится построение маршрутов виртуальных каналов, при котором используется алгоритм построения *k* кратчайших путей. Предлагается искать *k* путей «максимальной ширины» (тех путей, у которых остаточная пропускная способность больше), среди которых выбирать путь с минимальной оценкой длительности передачи кадров. Получить надежную оценку на данном этапе невозможно, так как длительность передачи кадра зависит от маршрутов других виртуальных каналов (в результате ожидания передачи кадров других виртуальных каналов в очередях коммутаторов), которые еще могут быть не построены. Поэтому на данном шаге предлагается эвристика, оценивающая стоимость пути согласно оценке длительности передачи кадра исходя из назначенных на текущий момент маршрутов виртуальных каналов.

После построения маршрутов производится вычисление длительности и джиттера передачи сообщений на основе любого из известных методов оценки длительности передачи кадров. При нарушении ограничений на длительность/джиттер передачи сообщений выполняется процедура переконфигурации виртуального канала, которая основана на формировании новых параметров *BAG* и *LM* на основе уже известной оценки длительности передачи кадра. После каждой итерации данная оценка уточняется, и переконфигурация запускается вновь, пока не превышено максимально допустимое количество итераций.

Далее в главе приведена оценка вычислительной сложности алгоритма, обоснована корректность.

Приведены и доказаны следующие утверждения.

Пусть производится назначение виртуального канала *vl* с требуемой пропускной способностью *r*(*vl*) и вызывается процедура ограниченного перебора. Пусть *H** – граф, включающий в себя все возможных маршруты без циклов для *vl*.

Утверждение 4.1. Если для поставленной задачи верно, что из всех оконечных систем выходит не более одного канала передачи данных, то процедуру ограниченного перебора виртуальных каналов достаточно проводить среди тех назначенных виртуальных каналов, которые проходят через хотя бы один элемент из графа *H**.

Утверждение 4.2. Пусть все каналы передачи данных имеют равную пропускную способность *R*, а пропускная способность, требуемая для построения маршрута виртуального канала, не превосходит $R_{\max}<R$. Пусть при построении маршрутов число каналов передачи данных, зарезервированная пропускная способность которых превысила $R - R_{\max}$, равно *l*. Тогда отношение суммарной требуемой пропускной способности маршрутов, построенных жадным алгоритмом, к суммарной пропускной способности при построении маршрутов оптимальным алгоритмом, не меньше $\frac{R - R_{\max}}{R + (l-1) * R_{\max}}$.

Утверждение 4.3 (достаточные условия выполнения ограничений на длительность передачи сообщений). Пусть для некоторой сети AFDX, по которой передается множество сообщений *MSG* по множеству виртуальных каналов *VL*, верны следующие условия:

- по виртуальному каналу *vl* передается множество сообщений MSG_{vl};
- расстояние без циклов между любыми двумя оконечными системами не превышает *k* коммутаторов;
- для любого сообщения[3] *msg* верно $J_{msg}= 0$;
- все коммутаторы сети имеют стратегию буферизацию FIFO на выходных портах;
- Все каналы передачи данных имеют пропускную способность не ниже *R*;

[3] В данном утверждении для упрощения рассматривается случай, когда не задано ограничение на джиттер порождения сообщения. Утверждение может быть расширено с учетом этого ограничения.

- τ_k – наибольшее время обработки кадра в коммутаторе перед постановкой в очередь.

Тогда, если для каждого сообщения *msg*, передающегося по виртуальному каналу *vl*, верно:

$$(\sum_{m\in MSG_{vl}}(\lceil size_m/(LM_{vl}-c)\rceil-1)BAG_{vl})+\frac{(k+1)LM_{vl}}{R}+k\cdot\tau_k+2k\sum_{v\in VL}\left(\frac{LM_v}{R}+gap\right)\le\tau_{msg},$$

где *c* – размер заголовка кадра, *gap* – интервал времени между выдачей двух последовательных кадров, то для всех сообщений будет выполняться ограничение на длительность передачи.

Утверждение 4.4 (достаточные условия построения виртуальных каналов для всех сообщений). Пусть для некоторой сети AFDX производится проектирование виртуальных каналов для множества сообщений *MSG*={*msg*}, причем выполнены следующие условия:

- Все каналы передачи данных имеют пропускную способность не ниже *R*;
- Для каждого сообщения *msg* определены следующие параметры:
  - $n=\min\limits_{i=0..7}(\left\lfloor\frac{\tau_{msg}-\Delta_0}{2^i}\right\rfloor+1,\left\lfloor\frac{T_{msg}}{2^i}\right\rfloor)$ (где Δ_0 – начальное приближение длительности передачи кадра, например, Δ_0 = 1 *мс*), при которых выполнены ограничения:
    - $(n-1)BAG\le\tau_{msg}-\Delta_0$
    - $n\cdot BAG\le T_{msg}$
  - $k(n)=\arg\min\limits_{i=0..7}(\left\lfloor\frac{\tau_{msg}-\Delta_0}{2^i}\right\rfloor+1,\left\lfloor\frac{T_{msg}}{2^i}\right\rfloor)$;
  - $LM(n)=\max(64,\left\lceil\frac{size_{msg}}{n}\right\rceil+c)$, где c = 47 байт – размер заголовка кадра;
  - $BAG(n)=2^{k(n)}$.
- Для указанных параметров выполнены следующие ограничения:
  - $\sum\limits_{msg\in MSG_{es}}(\frac{LM(n)}{R}+gap)\le 500$*мкс*, здесь суммирование ведется по всем сообщениям, исходящим из оконечной системы *es*, *gap* –

интервал времени между выдачей двух последовательных кадров в физический канал.

- $\sum_{msg \in MSG} \frac{LM(n)}{BAG(n)} \leq R$
- Выполнены ограничения утверждения 4.3 (с учетом формирования по одному виртуальному каналу для каждого сообщения).

Тогда алгоритм спроектирует виртуальные каналы для всех сообщений.

В **пятой главе** приведено описание экспериментальных исследований предложенных алгоритмов. Для исследования свойств алгоритмов и эффективности их отдельных процедур проводилась серия запусков для случайно сгенерированных данных. На основе полученных результатов исследовались средние полученные значения количества назначаемых запросов и времени работы алгоритма. Кроме того, для обработки результатов использовался механизм статистических гипотез. Формулировались гипотезы о числовом значении наблюдаемых величин, таких как «разность количества назначаемых запросов, получаемых тем или иным методом». Сформулированные гипотезы подтверждались или опровергались на основе исследования числовых величин полученных наблюдений с помощью статистических методов.

Проведенные исследования показали преимущества приведенного алгоритма планирования вычислений в ЦОД перед стратегиям случайного назначения элементов или стратегии назначения на «первый подходящий» элемент. Кроме того, исследованы преимущества одних жадных критериев над другими, показаны улучшения точности при использовании процедуры ограниченного перебора, которые достигли около 1% на исследуемых данных. Экспериментально показано улучшение точности, получаемые при использовании возможности создания реплик storage-элементов.

Алгоритм построения сетей AFDX исследовался на различных классах данных с различными вариантами запуска: с разными жадными критериями, с включением/выключением процедуры агрегации, с алгоритмом Дейкстры/алгоритмом k кратчайших путей, с различным количеством итераций процедуры переконфигурации виртуальных каналов. Проведенные

исследования показали существенное улучшение количества назначаемых сообщений при использовании процедуры агрегации. Кроме того, для класса данных с критическими ограничениями на длительность передачи сообщений улучшение производит использование алгоритма k кратчайших путей по сравнению с использованием алгоритма Дейкстры (благодаря использованию двух критериев), при этом на используемых входных данных достаточно одной итерации процедуры переконфигурации виртуальных каналов для получения улучшений. Дальнейшее увеличение количества итераций не привело к увеличению назначаемых сообщений.

В **шестой главе** описана разработанная автором инструментальная система построения сетей AFDX, которая обладает следующими свойствами:

- Предоставляет интерактивный графический интерфейс для задания и редактирования входных данных;
- Позволяет запускать алгоритм построения сети AFDX для заданных входных данных и значений параметров алгоритма;
- Имеет графический интерфейс для отображения результатов работы алгоритма;
- Позволяет задавать виртуальные каналы вручную, а также изменять вручную значения параметров спроектированных виртуальных каналов;
- Имеет модульную структуру и позволяет заменять используемые процедуры, а также запускать алгоритм без использования отдельных процедур. Данное свойство позволяет производить настройку инструментальной системы на особенности конкретной задачи путем настройки отдельных модулей.

Также показано отличие от существующих коммерческих и открытых систем для автоматизации проектирования бортовых систем, включающие в себя проектирование сетей AFDX. Существующие системы позволяют проводить моделирование, тестирование и верификацию конфигурации описанной бортовой системы, в то время как реализованное инструментальное средство позволяет по набору входных данных и заданным ограничениям производить построение корректной конфигурации.

В **заключении** сформулированы основные результаты диссертационной работы.

В **Приложении А** представлено доказательство корректности используемой в алгоритме построения сетей AFDX процедуры агрегации сообщений, служащей для организации передачи нескольких сообщений в одном виртуальном канале.

В **Приложении Б** приведено доказательство утверждения 4.2.

В **Приложениях В** представлено доказательство утверждения 4.3.

В **Приложениях Г** приведено доказательство утверждения 4.4.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Вдовин П.М., Костенко В.А.* Алгоритм распределения ресурсов в центрах обработки данных с раздельными планировщиками для различных типов ресурсов // Изв. РАН. ТиСУ. № 6, с. 80-93. 2014.
2. *Вдовин П.М., Костенко В.А.* Исследование эффективности процедуры агрегации виртуальных каналов при построении бортовых коммутируемых сетей // Вестник Московского Университета. Вычислительная Математика и Кибернетика. Т. 39, № 4, с. 32-40. 2015.
3. *Вдовин П.М.* Инструментальная система проектирования сетей AFDX // ТРУДЫ МФТИ. Т. 7, № 2, с. 131-136. 2015.
4. *Вдовин П.М., Зотов И.А., Костенко В.А., Плакунов А.В., Смелянский Р.Л.* Сравнение различных подходов к распределению ресурсов в центрах обработки данных // Изв. РАН. ТиСУ. № 5, с. 71-83. 2014.
5. *P. Vdovin, I. Zotov, V. Kostenko, A. Plakunov.* Congestion elimination on data storages network interfaces in datacenters // Parallel Computing Technologies. Vol. 9251 of Lecture Notes in Computer Science. Springer International Publishing Switzerland. 2015. P. 298-303.
6. *V. Balashov, V. Kostenko, P. Vdovin, R. Smeliansky, A. Shalimov*. An Analysis of Approaches to Onboard Networks Design // Proceedings of the International Science and Technology Conference Modern Networking Technologies (MoNeTec), IEEE. Moscow, Russia, MAKS Press. 2014. P. 20-24.
7. *Вдовин П.М., Костенко В.А.* Свидетельство о государственной регистрации программы для ЭВМ №2015661760 «Программа построения виртуальных каналов в коммутируемых сетях» от 09.11.2015.

8. *Вдовин П.М., Зотов И.А., Костенко В.А., Плакунов А.В., Смелянский Р.Л.* Задача распределения ресурсов центров обработки данных и подходы к ее решению // VII Московская междунар. конф. по исследованию операций (ORM2013). М.: ВЦ РАН, 2013. Т.2. С. 30-32.
9. *Вдовин П.М., Балашов В.В., Костенко В.А.* Управление AFDX-сетями и задачи их конфигурирования // Программные системы и инструменты. Тематический сборник № 13. М.: Изд-во факультета ВМиК МГУ. № 13, с. 30-47. 2013.