На правах рукописи

Рахман Павел Азизурович

РАЗРАБОТКА МЕТОДИКИ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ ПРИ ПРИМЕНЕНИИ ТЕХНОЛОГИИ ВИРТУАЛЬНЫХ МАШИН

Специальность: 05.13.13

“Телекоммуникационные системы и компьютерные сети”

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата технических наук

Москва – 2005

Работа выполнена на кафедре Вычислительных машин, систем и сетей Московского Энергетического Института (Технического университета).

Научный руководитель: кандидат технических наук, профессор
Ладыгин Игорь Иванович

Официальные оппоненты: доктор технических наук, профессор
Топорков Виктор Васильевич

кандидат технических наук, доцент
Бурцев Александр Борисович

Ведущая организация: Московский технический университет связи и информатики

Защита состоится 1 апреля 2005 г. в 16 час. 00 мин. на заседании Диссертационного совета Д 212.157.01 при Московском энергетическом институте (Техническом университете) по адресу: 111250, г. Москва, ул. Красноказарменная, д. 17, ауд. Г-306.

С диссертацией можно ознакомиться в библиотеке Московского энергетического института (Технического университета).

Отзывы в двух экземплярах, заверенные печатью организации, просьба направлять по адресу: 111250, г. Москва, ул. Красноказарменная, д. 14, Ученый совет МЭИ (ТУ).

Автореферат разослан «___» _____ 2005 г.

Ученый секретарь
Диссертационного совета Д 212.157.01
к.т.н. профессор Ладыгин И.И.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. На сегодняшний день существует большое количество организаций, имеющих корпоративную сеть, состоящую из множества конечных рабочих мест пользователей и некоторого, так называемого, серверного парка. Как правило, изначально корпоративная сеть грамотно проектируется специалистами с учетом надежности, безопасности и многофункциональности и руководство организаций крайне отрицательно относится к внесению значительных или даже небольших изменений в инфраструктуру сети, которая уже много лет исправно функционирует и удовлетворяет всем требованиям. Тем не менее, руководство всегда интересует расширение функциональных возможностей, снижение затрат на содержание сети и специалистов, обслуживающих ее, а также получение дополнительной прибыли с используемого технического оборудования. Многолетняя практика эксплуатации серверных систем показала, что на сегодняшний день большинство компьютеров серверного парка достаточно слабо загружены по вычислительным ресурсам.

Соответственно, необходимы новые подходы к решению проблемы повышения эффективности использования вычислительных ресурсов компьютеров серверного парка. Однако, для этого требуется создание формализованного подхода к реорганизации серверного парка с целью повышения эффективности использования ресурсов с применением современных информационных технологий.

Объектом исследований является: современные серверные парки, состоящие из множества компьютеров, связанных сетью передачи данных, и работающих на них логических серверов (серверные ОС вместе со всеми сетевыми приложениями), технология виртуальных машин для повышения эффективности использования вычислительных ресурсов.

Целью диссертационной работы является разработка методики повышения эффективности использования вычислительных ресурсов при применении технологии виртуальных машин.

Основные задачи диссертации.

- Анализ существующих подходов к решению проблемы повышения эффективности использования вычислительных ресурсов.
- Анализ существующих моделей распределения ресурсов.
- Разработка методики реорганизации серверного парка при применении технологии виртуальных машин с целью повышения эффективности использования ресурсов, а также с возможностью предварительной оценки целесообразности проведения реорганизации.
- Разработка математической модели и метода решения задачи поиска оптимального распределения логических серверов по компьютерам при применении технологии виртуальных машин.
- Разработка программной реализации алгоритма поиска оптимального распределения логических серверов по компьютерам.

- Экспериментальное исследование при помощи разработанной методики и программного обеспечения.

Методы исследований базируются на аппарате дискретной оптимизации, динамического программирования, комбинаторного анализа, теории проектирования корпоративной сетевой инфраструктуры и распределенных систем. Экспериментальные исследования, для подтверждения полученных в ходе диссертационной работы результатов, проводились на основе моделирования на компьютере.

Научная новизна работы заключается в следующем:

- разработана методика повышения эффективности использования вычислительных ресурсов при применении технологии виртуальных машин;
- разработана математическая модель и предложен метод решения задачи поиска оптимального распределения логических серверов по компьютерам при применении технологии виртуальных машин.

Практическая ценность работы состоит в следующем:

- предложена подробно детализированная методика реорганизации серверного парка, позволяющая использовать ее для решения проблемы повышения эффективности использования вычислительных ресурсов при применении технологии виртуальных машин;
- разработано программное обеспечение, позволяющее при заданной исходной информации по серверному парку рассчитать оптимальное распределение логических серверов по компьютерам при применении технологии виртуальных машин.

Достоверность и обоснованность научных положений, выводов и рекомендаций, сформулированных в диссертации, подтверждается результатами экспериментальных исследований при помощи разработанной методики и их сопоставлением с результатами программной реализации алгоритма поиска оптимального распределения логических серверов по физическим компьютерам.

Основные научные положения, выносимые на защиту:

- Методика реорганизации серверного парка с целью повышения эффективности использования вычислительных ресурсов при применении технологии виртуальных машин.
- Математическая модель задачи поиска оптимального распределения логических серверов по компьютерам и метод ее решения.

Апробация и внедрение результатов работы. Результаты работы докладывались и обсуждались на международном форуме информатизации МФИ-2004 “Информационные средства и технологии”. Результаты работы были внедрены в производственный процесс отдела системных и сетевых технологий Информационно–Вычислительного Центра МЭИ (ТУ), а также использовались при выполнении учебных научно-исследовательских работ и в дипломных проектах студентов кафедры Вычислительных машин, систем и сетей МЭИ (ТУ).

Публикации. Основные положения диссертационной работы изложены в 4-х печатных работах.

Структура и объем работы.

Основная часть диссертации состоит из введения, четырех глав и заключения и содержит 170 страниц машинописного текста, 33 рисунка и 21 таблицу. Список литературы включает 70 наименований. Дополнительная часть содержит 6 приложений, в том числе исходные тексты программного обеспечения, акт о внедрении. Общий объем приложений составляет 220 страниц машинописного текста, включая 44 рисунка и 55 таблиц.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснованы актуальность, научная новизна и практическая ценность работы, приведены основные результаты, полученные автором, и обозначена область применения разработанной методики.

В первой главе представлен обзор существующих подходов к решению проблемы повышения эффективности использования ресурсов при применении технологии виртуальных машин и моделей распределения ресурсов.

В первой части главы помимо основного критерия эффективности использования вычислительных ресурсов выделены дополнительные критерии для критического анализа существующих подходов к решению проблемы:

- уровень изоляции логических серверов;
- сетевая и антивирусная безопасность серверного парка;
- надежность функционирования серверного парка;
- прозрачность реорганизации серверного парка для конечных пользователей: проведение реорганизации не должно требовать дополнительной настройки компьютеров пользователей.

На сегодняшний день существуют пять основных подходов:

- использование ресурсов для дублирования функций логических серверов или решения дополнительных внутренних задач, приносящих прибыль;
- использование ресурсов для задач сторонних организаций;
- объединение служб и программного обеспечения разных логических серверов с целью снижения количества логических серверов;
- применение адекватных аппаратных решений;
- применение технологии виртуальных машин.

Основным недостатком первого подхода является недостаточная изоляция дополнительных приложений от основных, что порождает проблемы дополнительных уязвимостей ОС с точки зрения сетевой безопасности и конфликтов при работе несовместимых приложений.

Подход, связанный с использованием вычислительных ресурсов для задач сторонних организаций, имеет те же недостатки, что и первый, и при этом обостряется проблема информационной безопасности из-за повышения возможности несанкционированного доступа к данным и нарушения функционирования сервисов корпорации, предоставляющих ресурсы в аренду.

Подход, связанный с объединением сервисов, помимо появления проблем совместимости и информационной безопасности, также сказывается на логической структуре сетевой инфраструктуры, и это, как правило, влечет дополнительную работу по перенастройке рабочих мест пользователей.

Подход, связанный с подбором адекватных аппаратных решений, практически нереализуем в условиях современного рынка компьютерного оборудования, поскольку крайне трудно даже приближенно подбирать компоненты компьютеров с учетом требований логических серверов. Кроме того, при использовании старых компонент для логических серверов с невысокими требованиями может существенно снизиться надежность функционирования компьютеров.

Технология виртуальных машин предоставляет новые возможности для построения нового или реорганизации существующего серверного парка. Технология виртуальных машин позволяет на физическом компьютере под управлением некоторой, так называемой базовой операционной системы, эмулировать работу виртуальных машин, на которых также как и на обычных реальных компьютерах могут функционировать логические серверы. Таким образом, эта технология может обеспечить функционирование нескольких изолированных логических серверов на одном компьютере и тем самым обходить проблемы безопасности, совместимости, изменения привязок сервисов к логическим серверам, а также избавляет от необходимости подбора адекватных аппаратных решений, поскольку размещение нескольких логических серверов позволяет существенно повысить эффективность использования ресурсов компьютера.

Таким образом, при использовании технологии виртуальных машин серверный парк со слабой загрузкой ресурсов может быть реорганизован, что в конечном счете должно привести к уменьшению объема используемого оборудования и затрат на его поддержку. Однако, при применении технологии виртуальных машин возникает ряд проблем:

- неясно, каким образом перераспределять и объединять логические серверы среди компьютеров. При этом необходимо оценивать требования логических серверов, технические характеристики компьютеров, после чего каким-то образом размещать логические серверы на компьютеры, чтобы добиться хорошей загрузки ресурсов используемого оборудования. Однако, эта проблема вполне решаема с помощью создания математической модели задачи поиска оптимального распределения логических серверов на физические компьютеры и поиска метода ее решения.
- реорганизация серверного парка может привести к существенному ухудшению надежности и качества функционирования серверного парка, кроме того, реорганизация может не дать ожидаемой выгоды. Очевидно, что требуется разработка методики, на базе некоторой модели распределения ресурсов, с тщательной проработкой всех технических моментов, возникающих при проведении реорганизации, а также с возможностью предварительной оценки целесообразности проведения реорганизации.

Таким образом, по первой части главы сделан вывод о том, что подход, связанный с применением технологии виртуальных машин, является наиболее приемлемым, однако, требуется разработка новой методики реорганизации серверного парка при использовании этого подхода. Кроме того, требуется поиск подходящей математической модели и метода ее решения для задачи поиска оптимального распределения логических серверов по компьютерам.

Во второй части главы рассмотрены несколько традиционных моделей, в частности, модель задачи о “рюкзаке”, и несколько современных моделей распределения ресурсов, в частности, модель распределения ресурсов в системах жесткого реального времени. В результате обзора сделан вывод о том, что традиционные модели слишком просты и не учитывают ряд особенностей поставленной задачи, в частности, то, что компьютеров, предоставляющих ресурс, множество, а также может возникнуть необходимость исключения возможности размещения определенных логических серверов на один и тот же компьютер. Что касается современных моделей, несмотря на свою универсальность, они также не учитывают ряд особенностей логических серверов и технологии виртуальных машин, в частности, невозможность динамического планирования ресурсов из-за “громоздкости” логических серверов, требующих от одного до нескольких часов для перемещения с одного компьютера на другой. Кроме того, при перемещении работа логического сервера должна быть остановлена, что, как правило, недопустимо. Наконец, программные средства, реализующие технологию виртуальных машин, не имеют возможности встраивания собственных программ динамического управления распределением ресурсов компьютеров серверного парка.

Таким образом, во второй части главы сделан вывод о необходимости разработки новой математической модели.

Во второй главе представлена разработанная методика реорганизации серверного парка при применении технологии виртуальных машин.

Первый подраздел второй главы посвящен разработке общей схемы реорганизации. Поскольку реорганизация серверного парка не всегда может дать ожидаемую выгоду, кроме того, после реорганизации также может существенно снизиться качество и надежность функционирования серверного парка, то возникает необходимость в предварительной оценке целесообразности проведения реорганизации.

Соответственно, был сделан вывод о том, что реорганизация серверного парка должна разбиваться на два этапа.

I) Анализ задачи, сбор первичной информации, разработка первичного проектного решения, оценка качества решения с учетом возможных непредвидимых ситуаций, анализ выгоды, которое дает это решение, и оценка целесообразности проведения реорганизации. Цель первого этапа – максимально обезопасить проектное решение от провала на втором этапе – этапе реализации проектного решения. На первом этапе не допускается внесение каких-либо изменений в серверный парк.

II) Непосредственная реализация первичного проектного решения, выявление негативных последствий реорганизации. В случае неудовлетворительного качества работы серверного парка – корректировка решения и реализация скорректированного решения, далее повторяется анализ качества. Если на каком-либо шаге корректировка невозможна в силу неприемлемого снижения коммерческой выгоды – то поиск компромиссных решений либо отказ от проекта с возвратом серверного парка в исходное состояние.

На рисунках 1 и 2 приведены схемы алгоритмов для первого и второго этапа реорганизации серверного парка.

Остальные подразделы второй главы посвящены подробной детализации каждого из блоков алгоритмов для обоих этапов.

Отметим несколько ключевых моментов для обоих этапов.

Первичная информация по серверному парку включает в себя следующие исходные данные:

- множество типов ресурсов и единицы измерения базовых уровней этих ресурсов. Например, тип ресурса – оперативная память, единица измерения емкости оперативной памяти – мегабайты;
- множество компьютеров парка и базовые уровни их ресурсов;
- множество логических серверов и их требования к ресурсам;
- группы логических серверов, дублирующих функции друг друга, для которых недопустимо размещение на одном и том же компьютере из соображений надежности функционирования;
- требования базовой операционной системы;
- подмножество типов ресурсов, для которых предпочтительно решение проблемы повышения эффективности их использования.

Первичная информация используется для получения первичного распределения логических серверов по компьютерам, которое далее анализируется и принимается решение о целесообразности проведения реорганизации серверного парка.

В случае принятие положительного решения, начинается второй этап, на котором сначала выполняется непосредственная реорганизация серверного парка в соответствии с первичным распределением. При этом логические серверы должны быть подготовлены к переносу на виртуальную платформу. В рамках диссертации была разработана технология переноса операционных систем с физической платформы на виртуальную.

После реорганизации парка проводится оценка качества функционирования и в случае выявления проблем, выполняется поиск их причин. В результате этого информация по серверному парку может быть скорректирована или дополнена: изменение требований логических серверов, добавление новых ограничений по размещению логических серверов на один и тот же компьютер. Далее по скорректированным данным выполняется поиск скорректированного распределения, анализируется целесообразность его внедрения в серверный парк. В случае принятия положительного решения проводится коррекционная реорганизация серверного парка в соответствии со скорректированным распределением.

Рис. 1. Схема алгоритма для первого этапа реорганизации

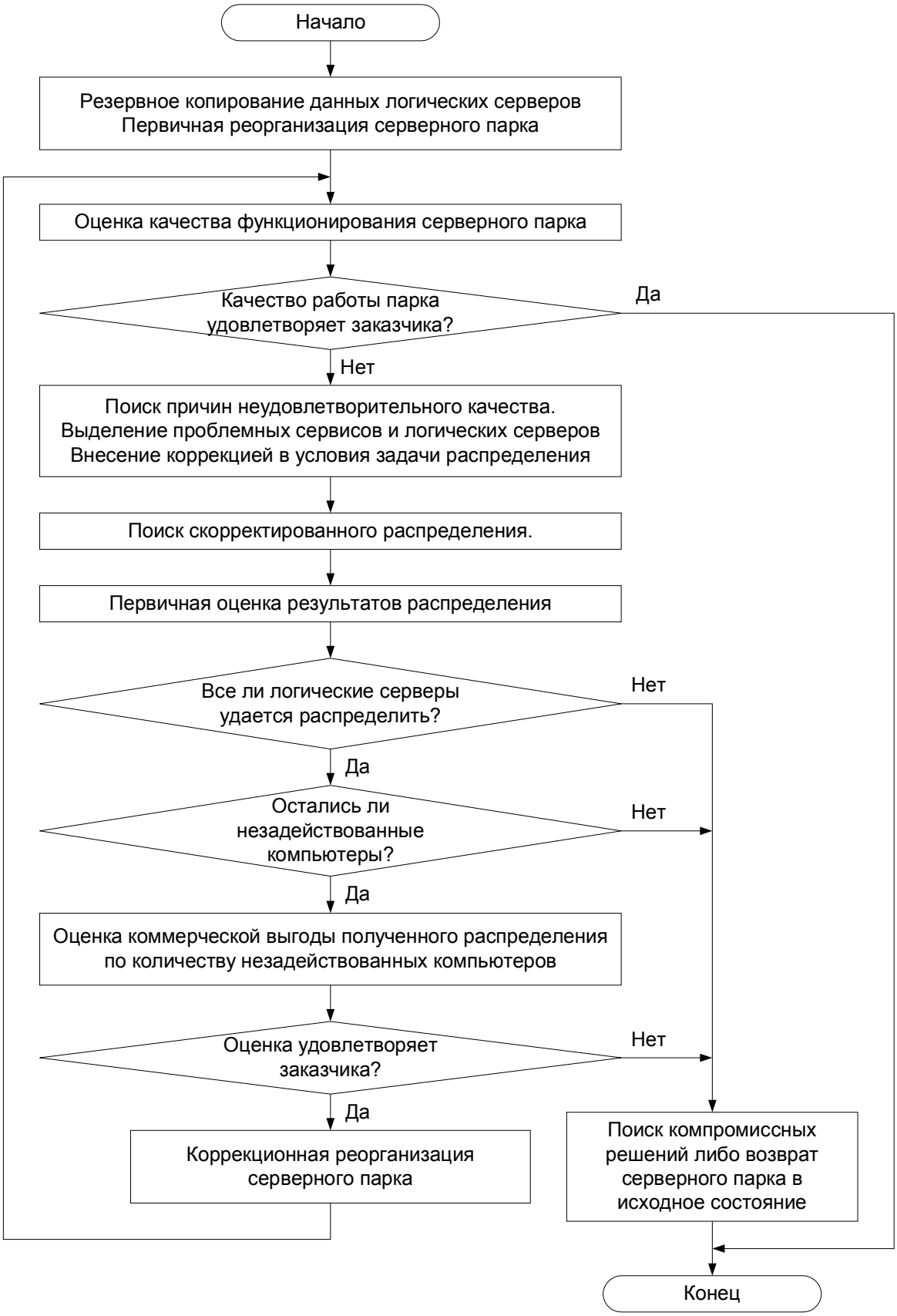

Рис. 2. Схема алгоритма для второго этапа реорганизации

В третьей главе представлены разработанная модель задачи поиска оптимального распределения логических серверов по компьютерам, предложенный метод ее решения, анализ сходимости, оценки объема перебора и качества решений для предложенного метода решения.

Входными данными математической задачи являются:

- множество типов ресурсов $\{\Re_i\}$, i = 1..NC, где NC – число типов ресурсов;
- множество компьютеров $\{H_k\}$, k = 1..NH, где NH – число компьютеров, и матрица базовых уровней ресурсов $\{R_{ik}\}$, i = 1..NC, k = 1..NH, компьютеров;
- множество логических серверов $\{S_j\}$, j = 1..NS, где NS – число логических серверов, и матрица требований $\{Q_{ij}\}$, i = 1..NC, j = 1..NS, серверов;
- вектор требований $\{V_i\}$, i = 1..NC, базовой операционной системы;
- дополнительная булева матрица ограничений $\{E_{dj}\}$, d = 1..NX, j = 1..NS, где NX – число дополнительных ограничений. Строки матрицы соответствуют группам логических серверов, столбцы соответствуют логическим серверам, если элемент E_{dj} равен 1, то это означает, что j-й логический сервер входит в d-ую группу. Для логических серверов, входящих в конкретную группу не допускается их размещение на одном и том же компьютере;
- булев вектор маски $\{O_i\}$. Если элемент O_i равен 1, то это означает, что по i-му типу ресурса необходимо проводить оптимизацию.

Основными выходными данными задачи являются:

- булева матрица распределения $\{X_{kj}\}$, k = 1..NH, j = 1..NS, логических серверов по компьютерам. Если элемент X_{kj} равен 1, то это означает, что j-й логический сервер распределен на k-й компьютер.

Вторичными выходными данными задачи являются:

- матрица загрузки ресурсов компьютеров $\{\eta_{ki}\}$, k = 1..NH, i = 1..NC, которая достаточно просто вычисляется по исходным данным и матрице распределения логических серверов по компьютерам;
- множество $\{K_{REM}\} \in \{H_k\}$ незадействованных компьютеров и множество логических серверов $\{J_{REM}\} \in \{S_j\}$, которые не удалось никуда распределить, и которые достаточно просто определяются по матрице распределения.

Основная цель реорганизации – повышение эффективности использования вычислительных ресурсов компьютеров. Соответственно, с одной стороны в оптимизационной задаче важна загрузка ресурсов каждого компьютера, с другой стороны повышение загрузки ресурсов компьютеров должно приводить к сокращению общего объема задействованного оборудования. Таким образом, в результате распределения на всех задействованных компьютерах загрузка ресурсов должна быть наилучшей, а на всех незадействованных компьютерах загрузка ресурсов должна быть нулевой (изначально все компьютеры серверного парка задействованы). Однако, очевидно, что множество задействованных компьютеров, получаемое в результате распределения, заранее неизвестно и его можно получить, лишь решив задачу. Если пытаться повышать загрузку ресурсов всех компьютеров серверного парка, то это приведет лишь к балансировке загрузки ресурсов и ни один из компьютеров не освободится.

Соответственно, возникает проблема неоднозначности в цели оптимизации. Эту проблему, как будет показано ниже, можно разрешить с помощью использования аппарата динамического программирования. Рассмотрим математическую модель, представленную в первой части главы.

Разработанная модель задачи состоит из следующих компонент:

- для каждого k-го компьютера, k = 1..NH, должны выполняться ограничения по ресурсам и дополнительные ограничения для исключения возможности размещения определенных логических серверов на один компьютер:

$$\begin{cases} \sum\limits_{j=1}^{NS} Q_{ij} x_{kj} \le R_{ik} - V_i, i = 1..NC \\ \sum\limits_{j=1}^{NS} E_{dj} x_{kj} \le 1, d = 1..NX \\ \forall j \in [1, NS] : x_{kj} \in \{0,1\} \\ k = 1..NH \end{cases} \quad (1)$$

- для каждого j-го логического сервера, j = 1..NS, должны соблюдаться ограничения о том, что логический сервер может быть размещен на один и только один компьютер:

$$\begin{cases} \sum\limits_{k=1}^{NH} x_{kj} \le 1 \\ j = 1..NS \end{cases} \quad (2)$$

- целевые функции загрузки вычислительных ресурсов для каждого задействованного компьютера, причем какие именно компьютеры должны быть задействованы – заранее неизвестно:

$$\begin{cases} L_\gamma = \dfrac{1}{\sum\limits_{i=1}^{NC} O_i} \left(\sum\limits_{i=1}^{NC} \dfrac{O_i}{R_{i\gamma} - V_i} \left(\sum\limits_{j=1}^{NS} Q_{ij} x_{\gamma j} \right) \right) \to \max \\ \forall \gamma \in \{k^*\} \end{cases} \quad (3)$$

где, $\{k^*\}$ – неизвестное множество индексов компьютеров, которые должны быть задействованы в соответствии с оптимальным распределением.

Во второй части главы выполнен обзор существующих методов решения исходной задачи целиком и, в силу неоднозначности в основной цели оптимизации, сделан вывод о необходимости в первом приближении разбиения исходной задачи на множество более простых подзадач.

Предложен следующий подход к решению исходной задачи в целом:

- вводится некоторый, так называемый “большой шаг”, обозначаемый буквой T. Вводится множество индексов {K(T)} компьютеров, оставшихся незадействованными на шаге T. Вводится множество индексов {J(T)} логических серверов, оставшихся нераспределенными на шаге T. Очевидно, на нулевом шаге T = 0, {K(0)} = {1,…,NH}, {J(0)} = {1,…,NS}. Считаем, что логические серверы отделены от компьютеров, переведены на виртуальную платформу, и требуется поиск их нового распределения по компьютерам;

- внутри каждого шага T для каждого компьютера с индексом k ∈ K(T), из числа компьютеров, оставшихся на шаге T, формируется математическая модель подзадачи, которая представляет собой задачу условной псевдобулевой оптимизации:

$$\begin{cases} \sum\limits_{j \in \{J(T)\}} Q_{ij} x_{kj} \le R_{ik} - V_i, i = 1..NC \\ \sum\limits_{j \in \{J(T)\}} E_{dj} x_{kj} \le 1, d = 1..NX \\ L(T,k) = \dfrac{1}{\sum\limits_{i=1}^{NC} O_i} \left(\sum\limits_{i=1}^{NC} \dfrac{O_i}{R_{ik} - V_i} \left(\sum\limits_{j \in \{J(T)\}} Q_{ij} x_{kj} \right) \right) \to \max \\ \forall j \in \{J(T)\} : x_{kj} \in \{0,1\} \end{cases} \quad (4)$$

- на каждом шаге T последовательно рассматриваются все компьютеры с индексами k ∈ K(T), оставшиеся к моменту шага T, и выбирается тот, для которого в результате решения соответствующей подзадачи, достигается наивысшее значение целевой функции среди значений, получаемых при решении подзадач для компьютеров с индексами k ∈ K(T). Соответственно, наилучший компьютер с индексом k* выбирается из следующего условия:

$$\begin{cases} L_{k*}(T) = \max\limits_{k \in \{K(T)\}} \{L_{\max}(T,k)\} \\ L_{k*}(T) > 0 \end{cases} \quad (5)$$

где, L_{max} (T, k) – оптимальное значение целевой функции на шаге T при решении подзадачи для k-го компьютера.

L_{k*} (T) – наивысшее оптимальное значение целевой функции на шаге T среди значений L_{max} (T, k) для всех k ∈ {K(T)}.

В условии (5) особенно важно условие того, что L_{k*} (T) не должно быть нулевым, это гарантирует, что на шаге T хотя бы один логический сервер распределится и хотя бы один компьютер будет задействован. Если L_{k*} (T) = 0, то это означает, что дальнейшее распределение логических серверов невозможно и решение исходной задачи должно быть прекращено.

Если же k* успешно найден из условия (5), то выполняются следующие преобразования: из множества оставшихся компьютеров исключается компьютер с индексом k*, а из множества логических серверов – множество серверов, распределенных на этот компьютер:

$$\begin{aligned} \{K(T+1)\} &= \{K(T)\} \backslash k* \\ \{J(T+1)\} &= \{J(T)\} \backslash \{j*\} \end{aligned} \quad (6)$$

где, {j*} – множество индексов логических серверов, которые были распределены на k*-й компьютер. Если в результате преобразования (6), множество K(T + 1) или J(T + 1) окажется пустым, то решения задачи завершается, в противном случае переход к шагу T + 1.

На рисунке 3 представлена схема алгоритма решения задачи в целом.

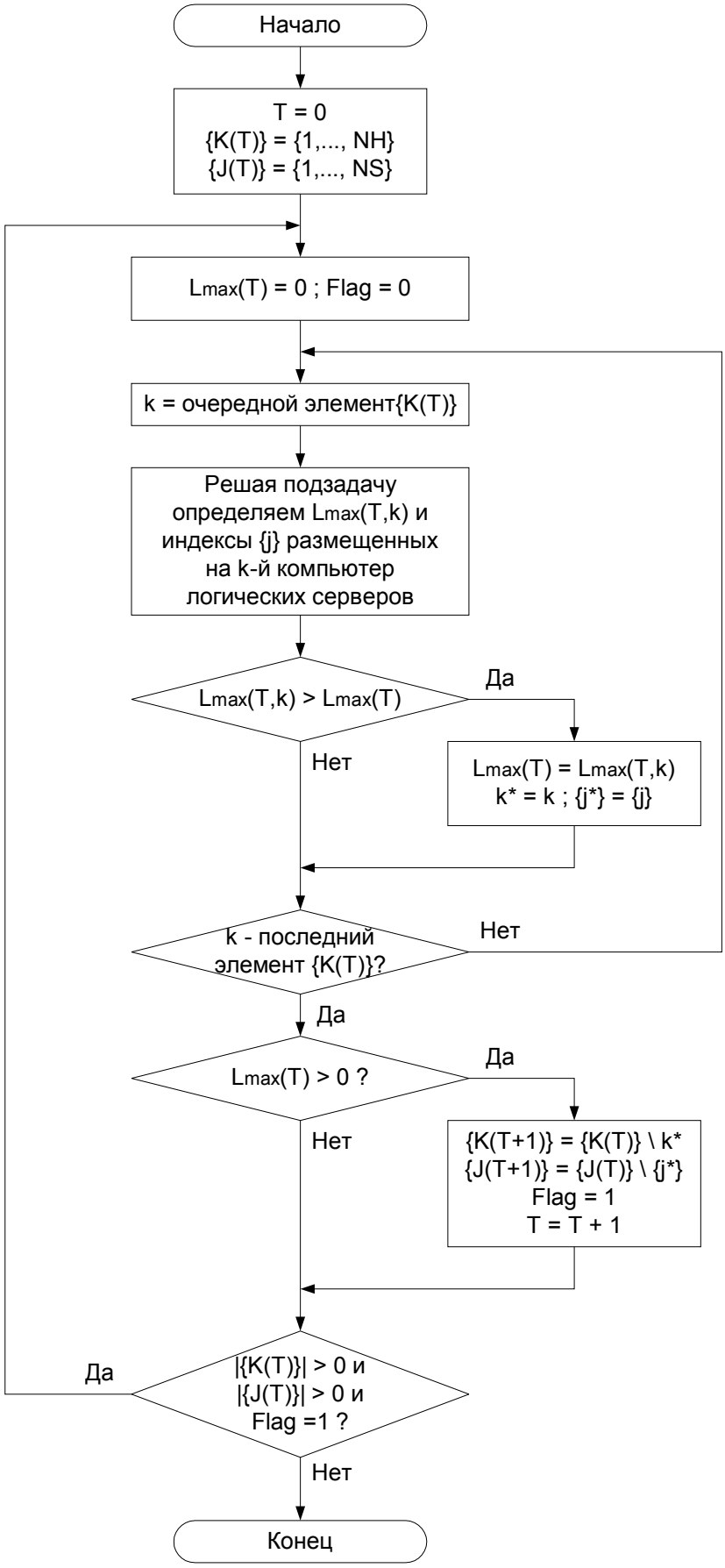

Рис. 3. Схема алгоритма решения задачи в целом

Далее в главе рассмотрены существующие подходы к решению задач условной псевдобулевой оптимизации. Модель задачи выглядит так:

$$\begin{cases} \sum_{j=1}^{N} a_{ij} x_j \le b_i, i = 1..M \\ L = \sum_{j=1}^{N} c_j x_j \to \max \\ \forall j \in [1,N]: x_j \in \{0,1\} \end{cases} \quad (7)$$

где, N – число переменных, M – число ограничений. Матрица $\{a_{ij}\}$, i = 1..M, j = 1..N, – левая часть, а вектор $\{b_i\}$, i = 1..M, – правая часть ограничений. Вектор $\{c_j\}$, j = 1..N, – коэффициенты целевой функции.

Поскольку число логических серверов и число компьютеров в общем случае может быть порядка сотни или даже тысячи, то сделан вывод о целесообразности использования приближенных методов для возможности получения результатов за приемлемое время.

За основу взят приближенный метод локального поиска

Для учета ограничений модель задачи условной псевдобулевой оптимизации преобразуется следующим образом:

$$\begin{cases} F = \sum_{j=1}^{N} c_j x_j - P * \sum_{i=1}^{M} \max\{0, \sum_{j=1}^{N} a_{ij} x_j - b_i\} \to \max \\ \forall j \in [1,N]: x_j \in \{0,1\} \end{cases} \quad (8)$$

где, F – скорректированная с учетом ограничений целевая функция, а P – некоторое достаточно большое положительное число, такое чтобы в любой недопустимой точке (в которой ограничения не выполняются) значение целевой функции было хуже, чем в любой допустимой точке.

Модель (8) задачи условной псевдобулевой оптимизации формируется из модели (4) подзадачи для рассматриваемого k-го компьютера в рамках текущего “большого шага” T, с помощью следующих несложных формул:

M = NC + NX, N = |{J (T)}|.

Для всех j = 1…N и q = j-й элемент из {J(T)}:

- Для всех i = 1…NC: $a_{i,j} = Q_{i,q}$; $b_i = R_{i,k} - V_i$;
- Для всех i = NC + 1…NC + NX: $a_{i,j} = E_{i-NC,q}$; $b_i = 1$;
- $c_j = \frac{1}{\sum_{i=1}^{NC} O_i} \left(\sum_{i=1}^{NC} \frac{O_i Q_{iq}}{R_{ik} - V_i} \right)$.

Здесь особо следует отметить, что модель (8) формируется только в случае, если для любого i = 1..N: $R_{ik} - V_i > 0$. В противном случае, очевидно, что логические серверы не могут быть распределены на заданный компьютер в силу явной нехватки по одному или нескольким типам ресурса компьютера, и подзадачу нет смысла решать.

Поскольку, очевидно, что в случае большой размерности подзадачи и сложных полимодальных ограничений метод локального поиска совсем не гарантирует нахождение глобального оптимума или даже достаточно качественных субоптимальных решений, то для усиления возможностей локального поиска была выполнена следующая его модификация:

- Радиус зон поиска (максимальное число координат, по которым рассматриваемые в зоне точки могут отличаться от точки центра зоны) не равен жестко единице. Радиус задается как дополнительное исходное данное задачи, обозначим его как *MR*.
- Задача решается не один, а множество раз для различных стартовых точек, генерируемых случайным образом. Число стартовых точек задается как дополнительное исходное данное задачи, обозначим его как *MS*. Соответственно, в результате выполнения задачи MS число раз получаем различные решения, сравниваем значения целевой функции в них и выбираем наилучшее решение.

В соответствии с вышесказанным в третьей главе был предложен алгоритм решения подзадач псевдобулевой оптимизации.

Таким образом, используя предложенные методы решения задачи поиска распределения и подзадач условной псевдобулевой оптимизации, мы можем получить матрицу распределения $\{X_{k,j}\}$, k=1…NH, j=1…NS, логических серверов по компьютерам.

Тогда, используя матрицу распределения $\{X_{k,j}\}$ несложно также вычислить матрицу загрузки ресурсов компьютеров $\{\eta_{k,i}\}$, k = 1..NH, i = 1..NC, для оценки загрузки ресурсов:

$$\eta_{ki} = \begin{cases} \sum_{j=1}^{NS} \frac{X_{kj} Q_{ij}}{R_{ik} - V_i}, если: R_{ik} - V_i > 0 \\ 0, в \quad противном \quad случае \end{cases} \quad (9)$$
$$\forall i \in [1, NC], \forall k \in [1..NH]:$$

Наконец, достаточно очевидно, что множество индексов незадействованных компьютеров $\{K_{REM}\}$ – это и есть множество индексов $\{K(T_{FIN})\}$, оставшихся компьютеров на последнем “большом шаге” T_{FIN} алгоритма решения задачи поиска распределения.

Аналогично, $\{J_{REM}\} = \{J(T_{FIN})\}$ – множество индексов логических серверов, оставшихся нераспределенными.

В третьей части главы выполнены анализ сходимости, оценки объема перебора и качества получаемых решений.

Что касается сходимости метода решения задачи поиска распределения, то она достаточно просто обосновывается. На уровне решения задачи в целом возможны только три конечные ситуации: множество логических серверов опустошается, множество компьютеров опустошается, ни один из оставшихся логических серверов не размещается ни на один из оставшихся компьютеров. На уровне решения подзадачи псевдобулевой оптимизации сходимость

определяется тем, что сам метод локального поиска обладает сходимостью. Введенные усложнения метода локального поиска – множество стартовых точек (MS) и управляемый радиус зон поиска (MR) – влияют на объем перебора, но не делают его бесконечным.

Далее, аналитическим путем получена оценка объема перебора для наихудшего случая возможных исходных данных:

$$\sum_{T=0}^{\lambda-1}\left((NH-T)*MS*\left(\left\lceil\frac{(NS-T)}{\psi}\right\rceil*W_1(\psi,NS-T)+W_0(\psi,NS-T)\right)\right) \qquad (10)$$
$$\psi=\min(NS-T,MR),\lambda=\min(NH,NS)$$

где, W_0 – число точек в первой зоне поиска, и оно вычисляется по следующей формуле:

$$W_0(\psi,NS-T)=\sum_{p=0}^{\psi}C_{NS-T}^{p} \qquad (11)$$

Соответственно, W_1 – число точек во всех зонах поиска, кроме первой, и оно вычисляется по следующей формуле:

$$W_1(\psi,NS-T)=\sum_{p=1}^{\psi}\left(\sum_{q=0}^{\lceil p/2\rceil-1}\left(C_{\psi}^{q}*C_{NS-T-\psi}^{p-q}\right)\right) \qquad (12)$$
$$\psi\le NS-T$$

Получить сумму ряда (10) в общем виде не представляется возможным. Тем не менее, были получены некоторые приближенные оценки зависимости объема перебора от тех или иных параметров. Объем перебора полиномиально ~O(NH^2) зависит от числа компьютеров (NH). Объем перебора полиномиально ~O(MS) зависит от числа стартовых точек (MS). Объем перебора полиномиально ~O(NS^{MR+1}) зависит от числа логических серверов (NS) при малых значениях радиуса зон поиска (MR << NS), и экспоненциально ~O(2^{NS}) – при MR = NS.

Также получена оценка нижней границы качества решения в зависимости от радиуса зон поиска (MR) и от числа логических серверов (NS), но только для случая одного компьютера (NH=1), путем анализа решения нескольких представительных примеров с последующим обобщением результатов анализа. Введено понятие точности, как отношение загрузки ресурсов компьютера при субоптимальном решении (распределении логических серверов на компьютер), к загрузке ресурсов компьютера, достигаемых в глобальном оптимуме. Соответственно, получены 2 оценки для наихудшего случая исходных данных:

- При отсутствии дополнительных ограничений (матрицы $\{E_{dj}\}$):

$$\begin{cases}\xi\ge\left(\dfrac{MR}{MR+1}\right)\\ \xi\in[0,1],NS\ge 3,MR\in[1,NS-2]\end{cases} \qquad (13)$$

- При наличии дополнительных ограничений:

$$\begin{cases}\xi\ge\left(\dfrac{MR}{NS-1}\right)\\ \xi\in[0,1],NS\ge 3,MR\in[1,NS-2]\end{cases} \qquad (14)$$

В четвертой главе представлены разработанная программная реализация алгоритма поиска оптимального распределения логических серверов по компьютерам, а также описание проведенного экспериментального исследования с использованием этой программы.

В первой части главы выполнен обзор современных подходов к разработке прикладного программного обеспечения, а также рассмотрены требования, предъявляемые к разрабатываемым программам. В соответствии с ними в рамках диссертации было разработано Windows-приложение в среде Delphi 7.0, которое во второй части главы кратко описано.

В третьей части главы описаны экспериментальное исследование и полученные результаты. Проведено экспериментальное исследование по оценке времени решения задач в зависимости от ее размерности. В частности, на рисунке 4 приведен график зависимости времени решения от числа компьютеров, а на рисунке 5 график зависимости времени решения от числа логических серверов при радиусе зон поиска равном 1 (MR = 1) и при прочих неизменяемых исходных данных задачи. Результаты экспериментального исследования подтвердили оценки объема перебора, полученные аналитическим путем в третьей главе.

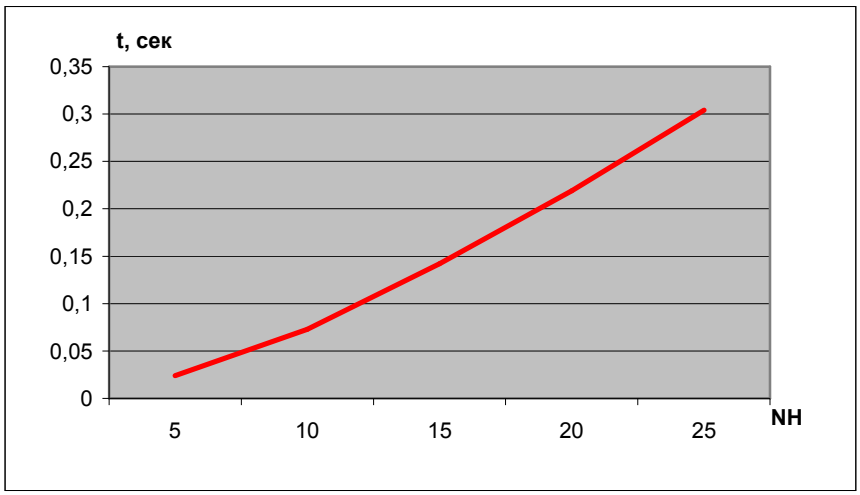

Рис. 4. График зависимости времени решения от числа компьютеров

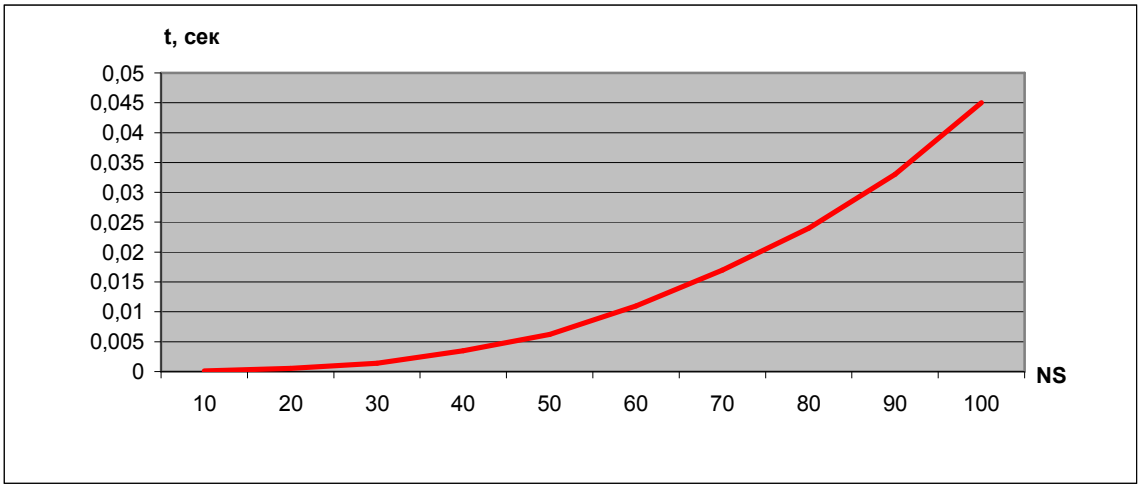

Рис. 5. График зависимости времени решения от числа логических серверов

Проанализирована целесообразность декомпозиции рассматриваемого серверного парка на множество более мелких парков и проведение реорганизации в каждом из них в отдельности. Очевидно, что при использовании декомпозиции размерности математической задачи поиска для каждого более мелкого серверного парка существенно снижаются. Однако, как показали результаты экспериментов, в общем случае, качество получаемых решений для исходного серверного парка при использовании декомпозиции может оказаться существенно худшим, нежели чем при решении задачи для исходного серверного парка без декомпозиции.

Далее рассмотрено несколько производственных задач. Сравнение результатов решения задач при помощи программного инструмента и вручную наглядно показало существенное преимущество качества получаемых решений при использовании программного инструмента, реализующего алгоритм решения задачи поиска распределения, предложенного в главе 3. В частности, в одной из задач было задано 15 логических серверов, работающих на 15 компьютерах. При использовании методики и разработанной программы удалось обойтись всего 6 задействованными компьютерами (сокращение в 2,5 раза). При попытке решения этой же задачи вручную требовалось 9 задействованных компьютеров (сокращение только в 1,66 раза).

Проведена оценка качества решений в зависимости от радиуса зон поиска (MR) и числа логических серверов (NS) на ряде модельных примеров, исходные данные которых выбирались из соображений наихудшего случая. Результаты экспериментов подтвердили оценки нижней границы точности, выведенные в третьей главе. В частности в таблице 1 приведены результаты одной из серии экспериментов по оценке точности субоптимальных решений при различных радиусах зон поиска (MR) при отсутствии дополнительных ограничений.

Таблица 1

λ	**1**	**2**	**3**	…	**10**	…	**17**	**18**	**19**
Загрузка ресурса при MR = λ	50%	66,67%	75%	…	90,91%	…	94,45%	94,74%	95%
Загрузка ресурса при MR = NS	100%	100%	100%	…	100%	…	100%	100%	100%
Точность субоптимального решения	0,5	0,6667	0,75	…	0,9091	…	0,9445	0,9474	0,95

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ.

- Для поставленной проблемы повышения эффективности использования вычислительных ресурсов разработана методика реорганизации серверного парка, позволяющая решить эту проблему, на базе модели задачи поиска оптимального распределения логических серверов по компьютерам при применении технологии виртуальных машин.
- Разработана технология переноса операционных систем с физической аппаратной платформы на виртуальную платформу, которая делает методику реорганизации серверного парка при применении технологии виртуальных машин технически реализуемой.

- Разработана математическая модель задачи поиска оптимального распределения множества логических серверов на множество компьютеров при применении технологии виртуальных машин.
- Предложена схема разбиения, с использованием аппарата динамического программирования, исходной задачи на множество подзадач поиска распределения логических серверов на один компьютер, являющихся задачами условной псевдобулевой оптимизации. Для решения подзадач предложен модифицированный метод локального поиска с использованием функции штрафов, управляемого радиуса зон поиска и множества случайных стартовых точек. Выполнены анализ сходимости, оценки объема перебора и качества решений для предложенного метода решения.
- Разработана программная реализация алгоритма решения задачи поиска оптимального распределения логических серверов по компьютерам.
- Проведено экспериментальное исследование на ряде модельных примеров и производственных задач при внедрении результатов диссертации с использованием разработанного программного обеспечения и предложенной методики.

Основные положения диссертационной работы изложены в 4-х печатных изданиях.

- Рахман П.А. Подходы к повышению эффективности использования вычислительных ресурсов корпоративных сетей // Труды международной конференции “Информационные средства и технологии”. – М.: Янус-К, 2004. – Т. 3. – С. 120-121.
- Рахман П.А. Использование методов дискретной оптимизации для решения задач распределения ресурсов при применении технологии виртуальных машин в корпоративных сетях // Труды международной конференции “Информационные средства и технологии”. – М.: Янус-К, 2004. – Т. 3. – С. 122-123.
- Рахман П.А. Проблемы переноса современных операционных систем с реальной аппаратной платформы на виртуальную // Труды международной конференции “Информационные средства и технологии”. – М.: Янус-К, 2004. – Т. 3. – С. 124-125.
- Рахман П.А. Концептуальный подход к повышению эффективности использования вычислительных ресурсов корпоративных сетей при применении технологии виртуальных машин // Объединенный научный журнал. – М.: Тезарус, 2005. – №2. – С. 59-67.