© 2012 г. И.П. БОЛОДУРИНА, д.т.н., профессор,
Д.И. ПАРФЁНОВ
(Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования "Оренбургский государственный университет", Оренбург)

МОДЕЛИРОВАНИЕ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ И ДИНАМИЧЕСКОЙ БАЛАНСИРОВКИ НАГРУЗКИ В ИНФОРМАЦИОННОЙ СИСТЕМЕ ДИСТАНЦИОННОЙ ПОДДЕРЖКИ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА [1]

Развитие компьютерных технологий в последнее время приводит к все большему использованию облачных вычислений в том числе в системах дистанционного обучения. При этом актуальной задачей является исследование эффективных методов управления производительностью и оптимизации использования программных и аппаратных ресурсов. В рамках представленного исследования построена многоуровневая модель системы дистанционного обучения, проведен анализ характеристик и приведен алгоритм для повышения эффективности использования имеющихся ресурсов с целью улучшения качества предоставления услуг в распределенных информационных системах дистанционного обучения.

Modeling of resource allocation and dynamic load balancing in the information system of remote support educational process / I.P. Bolodurina, D.I. Parfenov (Federal State Educational Government-financed Institution or Higher Professional Education "Orenburg State University", pr. Pobedy, 13, Orenburg 460018, Russia, E-mail: prmat@mail.osu.ru). Development of computer technology in recent years leading to increasing use of cloud computing, including in distance learning. In this case, the actual problem is investigation of effective methods of performance management and optimization of the use of software and hardware resources. As part of the representation of the study built a multilevel model of distance learning, the analysis of the characteristics and the algorithm to improve the utilization of available resources in order to improve the quality of services in a distributed information-dimensional systems of distance learning.

[1] Работа выполнена при финансовой поддержке РФФИ (грант № 11-07-00046).

1. Введение

Получение образования в настоящее время все больше взаимосвязано с применением информационных технологий, что, в свою очередь, порождает потребность в разработке и активном использовании сетевых мультимедийных образовательных услуг. Наиболее активно развивающимся в этом направлении является дистанционное обучение. Лекции, семинары, форумы, трансляции образовательного контента, интерактивные услуги широко внедряются в учебный процесс. При этом мультимедийные образовательные ресурсы должны быть доступны как пользователям локальной вычислительной сети университета, так и через Интернет, внешним пользователям.

С ростом предоставляемых сервисов, а также с увеличением числа конечных пользователей для обеспечения доступа к мультимедийным ресурсам и улучшения качества предоставляемых услуг требуется постоянное наращивание вычислительных мощностей аппаратных платформ и пропускной способности каналов связи. Однако существует определенная статистическая закономерность потребления имеющихся вычислительных мощностей, которая показывает, что 80% ресурсов необходимы лишь в 20% времени (справедливо и обратное утверждение). Как правило, нагрузка на аппаратные и программные ресурсы образовательного учреждения носит плавающий характер, при этом периоды пиковой нагрузки имеют прямую связь с происходящими в ВУЗе событиями и процессами (начало/конец учебного года, проведение промежуточной и итоговой аттестации, вступительных испытаний и научных конференции и т. д.). Большинство событий носят систематический характер, что позволяет рассчитать необходимую нагрузку и подготовить требуемые ресурсы. В то же время расширение ресурсов под конкретную задачу должно быть также оправдано в долгосрочной перспективе, что в свою очередь требует обеспечения масштабируемости внедряемых решений.

В настоящее время для обеспечения хранения мультимедиа контента и доступа к ресурсам наиболее выгодным является применение гибридных облачных систем. Масштабируемость и другие характеристики, присущие облачным вычислениям, являются одним из немаловажных факторов, влияющих на тенденции размещения и предоставления информационных услуг в образовательных учреждениях. Это особенно актуально для дистанционного обучения, при котором основная часть учебного процесса и взаимодействие обучающегося и преподавателя осуществляется по средствам сети Интернет. Применение такого подхода к предоставлению сетевых мультимедийных образовательных услуг, а так же использование возможностей современных веб-технологий позволяет осуществлять доступ и совместное использование функционально насыщенных веб-приложений при помощи веб-браузера, не осуществляя их установку на локальном компьютере. Одним из приоритетных направлений развития мультимедийных услуг является разработка интерактивных веб-сервисов, направленных на трансляцию видеопотоков. К их числу относятся:

- цифровое телевидение (TVoIP);
- видео по запросу (Video on Demand - VoD);
- интернет-трансляции;
- вебинары.

На факультете дистанционных образовательных технологий (ФДОТ) Оренбургского государственного университета (ОГУ) за последнее десятилетие накоплен определенный опыт в автоматизации задач организационно-методического и программно-технического сопровождения дистанционного обучения в ВУЗе. Кроме того, на факультете разработан консолидированный сервис - «Видеопортал дистанционного обучения»,

обеспечивающий доступ к перечисленным ранее услугам с целью реализации образовательных программ высшего профессионального образования, позволяющий организовать взаимодействие преподавателя и студентов на новом уровне путем создания интерактивной обратной связи. Мультимедийный сервис такого класса, помимо обеспечения постоянной доступности, требует высокого качества обслуживания при передаче данных [1].

Ежегодный прирост числа потребителей сетевых мультимедийных услуг, в том числе распределенных внешних пользователей, приводит к значительному росту как внутреннего, так и внешнего трафика и, как следствие, повышению нагрузки на оборудование и каналы связи. Узким местом мультимедийных образовательных сервисов является точка вещания видеопотока ввиду ограниченности пропускной способности выходного канала. Особенно эта проблема актуальна для пользователей, осуществляющих доступ к веб-приложениям из сети Интернет. При этом, клиенты внутренний сети, использующие корпоративные мультимедийные веб-приложения, за счет своей многочисленности, создают дополнительную нагрузку на аппаратное обеспечение, обслуживающее данный сервис. Сама по себе передача видеоконтента требует особого подхода. При доступе к уже существующему контенту создается высокая нагрузка на систему хранения данных. При онлайн вещании (например, видеоконференции) создается высокая нагрузка на службу сжатия и обработки контента. Кроме того, специфика работы Интернет заключается в том, что в глобальных соединениях не поддерживаются сквозные широковещательные трансляции (multicast, broadcast). Отправка пакетов группе пользователей или абсолютно всем пользователям сети возможно только в пределах локальной сети, в глобальных сетях могут отправляться только адресные (unicast) пакеты. Как следствие, для каждого клиента при обращении к сервису трансляции создается персональный поток (точка-точка), что при большом количестве обращений приводит к исчерпанию пропускной способности канала связи.

2. Постаовка задачи

В рамках исследования нами выделено несколько отличительных особенностей обеспечения доступа к мультимедийным образовательным ресурсам в распределенной сети ВУЗа:

1) Нагрузка на сервера периодическая, одновременно происходят обращения к нескольким ресурсам с разными типами. В большинстве случаев существующее оборудование не позволяет без использования распределения нагрузки обслужить всех клиентов, причем загрузка серверов носит неодновременный и наравномерный характер.
2) До 90% нагрузки предопределено, поскольку для доступа к ресурсам используется пре-регистрация (подписка на сервисы), например запись на вещание лекции, а также статистически данные оценки использования ресурсов информационных, полученных на основе ежегодного отчета «об информатизации ВУЗа». При этом использование стандартных средств не позволяет учесть предопределенную нагрузку и распределить ее в условиях ограниченных ресурсов.
3) В пределах локальной сети присутствуют различные категории полезного трафика, но при обращении к корпоративным сервисам не учитывается приоритет обслуживания и выделение полосы пропускания для критически важного трафика.

Для эффективного использования ресурсов необходимо их динамическое выделение в рамках решаемых задач для исключения простоя и перегрузки аппаратного обеспече-

ния. Для повышения надежности и улучшения качества предоставляемых сетевых мультимедийных услуг требуется внедрение эффективных методов обеспечения распределения нагрузки аппаратно-программных ресурсов университетского комплекса.

Традиционно оптимизация использования вычислительных ресурсов осуществляется при помощи процедуры балансировки нагрузки. Как правило, балансировка заключается в распределении запросов определенным компонентам, обработчикам облачной системы на основе оценки загруженности и их состояния. Так как облачная система управляется из единого контроллера, это подразумевает, что поступивший запрос может быть предан на обработку любому из активных устройств, поддерживающих работу выбранного приложения. Однако, работа приложений часто зависит не только от объема оперативной памяти и процессорного времени, требуемых для выполнения запроса пользователя. В настоящее время высоконагруженные приложения, направленные на обработку больших объемов данных, например, видео и мультимедиа контента, невозможно представить без использования масштабируемых систем управления баз данных и распределенных систем хранения данных. Проведенный анализ публикаций по теме исследования показал [2,3,4,5], что на сегодняшний день нет достаточно эффективных универсальных, комплексных методов балансировки и распределения нагрузки, включающих в себя: выделение процессорного времени, оперативной памяти, управление потоком SQL запросов к базе данных, а также динамическое распределение размещения файлов в системе хранения данных (СХД).

Как уже говорилось выше, зачастую «узким» местом в современных приложениях является база данных. В основном, можно выделить два класса проблем: производительность и необходимость хранения большого количества данных. Как правило, для снижения нагрузки на БД предлагается модель распределения на несколько узлов. Однако, при этом остро встаёт проблема синхронизации между ними и обеспечения резервного копирования. Кроме того, такой подход не может гарантировать равномерное распределение нагрузки. На наш взгляд, намного эффективней обеспечивать промежуточное распределение SQL-запросов, используя алгоритм приоретизации обработки и выделения «тяжелых» и «легких» запросов к серверам баз данных на уровне контроллера гибридной облачной системы используя управляющие команды маршрутизации запроса. Кроме того, следует разделять запросы чтения и записи данных.Такая стратегия позволяет значительно легче масштабировать БД, подключая дополнительные мощности по мере необходимости.

Основное отличие облачных СХД заключается в использовании динамически масштабируемого дискового пространства, т.е. в процессе работы сервис по мере необходимости арендует требуемое пространство, задействуя, тем самым, ряд доступных физических устройств. Для эффективного использования арендованных ресурсов облачные СХД вынуждены регулярно производить процедуру масштабирования и переконфигурирования, т.е. изменения количества устройств хранения, используемых системой. При этом процесс оптимизации размещения, в зависимости от изменения объемов размещаемых данных, может занимать достаточно продолжительное время [6,7]. Масштабирование и переконфигурация неразрывно связаны с алгоритмами миграции и распределения элементов данных по устройствам хранения. Выполнение миграции данных не должно приводить к снижению качества обслуживания клиентов СХД, для чего в алгоритмах необходимо учитывать пропускную способность сети и максимальный объем данных, который можно передавать в один момент времени с одного устройства хранения на другое. Существующие алгоритмы миграции данных не учитывают этой особенности, и высвобождение лишних и добавление новых устройств хранения возможно производить лишь после полного завершения процедуры переконфигурации. Во время

выполнения этого длительного этапа лишние устройства остаются задействованными, а новые устройства - не до конца использованными. Существующие подходы не учитывают следующих характеристик работы СХД и размещения данных, а именно объем доступного дискового пространства конечных физических устройств, размеры конечных файлов (блоков данных), отмеченных для миграции между устройствами. Кроме того, сами алгоритмы миграции не учитывают распределенное дублирование данных по устройствам, что значительно снижает отказоустойчивость системы и не позволяет обеспечить динамическую балансировку нагрузки к файловой системе, что является важным показателем при использовании мультимедийный сервисов.

3. Решение задачи

Для детального анализа ресурсов системы дистанционного обучения нами разработана уровневая модель на основе базовых высоконагруженных доступных внешним пользователям подсистем:

- подсистема контроля знаний (уровень 1);
- подсистема предоставления учебно-методических комплексов (электронная библиотека) (уровень 2);
- подсистема трансляции и публикации видео и аудио материалов (видеопортал ДО) (уровень 3).

Выделенные нами базовые компоненты могут быть представлены как комплекс, обеспечивающий работу мультисервисного набора услуг для физически распределенных пользователей [10]. Каждая из подсистем, используемая в системе дистанционного обучения, предъявляет собственные требования к прикладному программному обеспечению оборудования и качеству обслуживания (QOS), что позволяет проводить моделирование с использованием многокритериальных показателей и как следствие создать базу знаний для управления и распределения поступающей нагрузки.

Практика показывает, что большинство информационных систем, работающих с внешними пользователями, при большом количестве обращений испытывают недостаток в потребляемых ими ресурсах. Причем отказ в обслуживании для любой из систем напрямую зависит от объема выделенных для ее работы ресурсов. Большое количество клиентов, осуществляющих одновременное подключение к серверу, приводят к дисбалансу рабочего трафика, что в свою очередь негативно сказывается на буферах маршрутизируешего оборудования и превышению критических объемов ресурсов серверов. Как отмечалось ранее, прогнозирование нагрузки от клиентов позволяет подготовить оборудование и каналы связи для приема трафика. Однако, это не решает проблему непрогнозируемых экстремальных нагрузок, а применение метода, основанного на увеличении времени отклика системы приводит к удлинению очереди заявок что, снижает динамику работы системы. Такой подход невозможно организовать для сервисов реального времени таких как, потоковая передача видео- и аудиоданных. К тому же большинство система работает по принципу First In, First Out (FIFO).

В рамках нашего исследования для системы дистанционного обучения разработан алгоритм приоритетного обслуживания клиентов высоконагруженных приложений с критичным временем отклика. В связи с этим нами решены следующие задачи:

- выделено прикладное программное обеспечение, влияющее на работу каждой из подсистем;
- определена наиболее ресурсоемкая подсистема;

– выставлены индикаторы приоритетов обработки запросов при одновременном функционировании подсистем;
– построена математическая модель для максимизации числа обработанных обращений к СДО.

Работу интерент-приложений часто рассматривают как систему массового обслуживания с ограниченным временем пребывания в очереди и пуассоновским потоком заявок [8,9]. Для формализации работы интернет-приложений механизм обработки запросов будем рассматривать как многоканальное СМО с несколькими очередями (Рисунок 1).

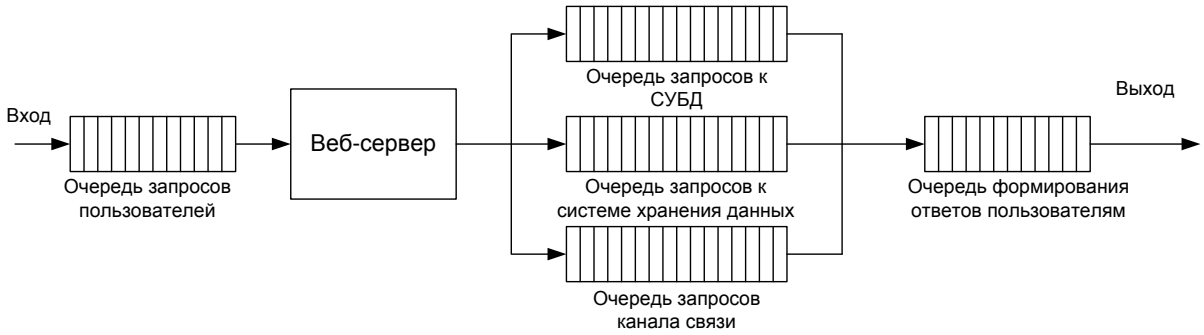

Рисунок 1 – Модель работы интерент-приложения как СМО

В ходе исследования нами установлено, что на всех трех уровнях модели основными факторами, влияющими на скорости обработки запросов пользователей программным обеспечением системы дистанционного обучения, являются:

– обращение к СУБД для получения необходимых данных;
– обращение к дисковому пространству, как самого сервера, так и к системе хранения данных для записи или чтении необходимых данных;
– использование приложением канала связи, заданной пропускной способности в единицу времени для приема и передачи требуемого объема данных.

Для указанных выше факторов нами введены численные показатели классификационных признаков каждого из уровней построенной модели:

– количество запросов в единицу времени, отправленных к СУБД (SQL-запросов/с);
– использование дискового пространства серверного оборудования (Мб/с);
– интенсивность использования входящего/исходящего канала связи (Мбит/с).

Для каждого из уровней численные показатели в процентном соотношении к суммарному показателю использования данного ресурса всеми уровнями модели определяются выражением:

$$R_{i\,исп} = \frac{R_i \cdot 100}{(R_1 + \ldots + R_n)}, \quad (1)$$

где $R_1,\ldots,R_n$ численные показатели использования ресурса по каждому из классификационных признаков, полученные в результате измерений на интервале времени ΔT.

Индикаторы приоритета обслуживания уровней модели определим на основе рейтинга востребованности ресурсов системы в целом. Анализируя интенсивность использования каждого из компонентов ресурсов в СДО построена диаграмма приоритетов

(Рисунок 2) востребованности ключевых сервисов и аппаратного обеспечения, лежащих в основе каждой из подсистем.

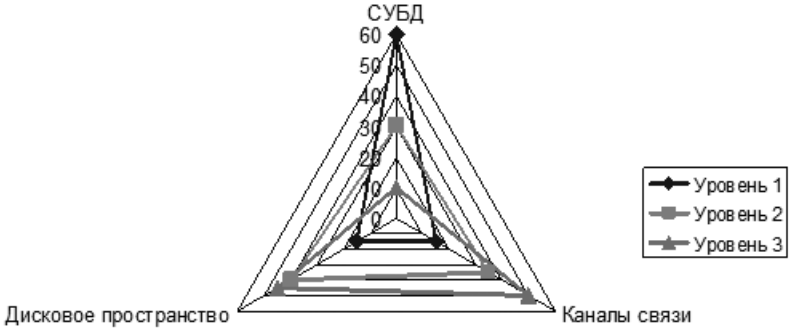

Рисунок 2 – Диаграмма приоритетов востребованности ресурсов СДО

Общую ресурсоемкость системы дистанционного обучения определим как суммарную площадь $U_{сдо}$, занимаемую всеми уровнями модели (U_i). При этом максимально возможные ресурсы сервера обозначим, как площадь, полученную при использовании 100% всех ключевых сервисов[11].

Так как работа подсистем осуществляется непрерывно, поступление заявок к ресурсам системы (СУБД, каналы связи, дисковое пространство), можно описать в дискретном времени:

$I_j(T_j)=\{j:t \in (0,T_i)\}$ – множество номеров заявок, пришедших в интервал времени $(0,T_i)$ на подсистему i (i – уровень подсистемы, i=1,…,M).

Статус обработки j-ой заявки поступившей на i-ый уровень обозначим x_{ij}, причем отказ в обслуживании будем считать $x_{ij}=0$, успех $x_{ij}=1$.

Интенсивность поступления и обработки заявок на каждый из уровней модели обозначим λ_i, при этом она напрямую зависит от ресурсоемкости подсистемы. Кроме того, введем показатель приоритета (P_i) для каждого из уровней, распределение которого зависит от количества одновременно используемых ресурсов. Тогда на нагрузку, создаваемую каждым из уровней, можно наложить ограничение:

$$\sum_{I_j(T_j)} U_i x_{ij} \le H_i \text{ , } i=1,\ldots,M. \quad (2)$$

При задании целевой функции введены следующие ограничения, связанные с предметной областью исследования:

- время обработки (T) любого запроса ограничено;
- мощность сервера (H) фиксирована.

Ввиду неравномерности использования основных ресурсов сервера каждым из уровней системы дистанционного обучения следует определить условия максимальной загрузки сервера, при которой возможна безотказная работа всех приложений:

$$\sum_{i=1}^{M} \sum_{j \in I_j(T_j)} U_i x_{ij} \le H \text{ , } x_{ij}=\{0,1\}. \quad (3)$$

Ввиду неравномерности использования основных ресурсов сервера каждым из уровней системы дистанционного обучения следует определить условия максимальной загрузки сервера, при которой возможна безотказная работа всех приложений:

Таким образом, для обработки максимального количества запросов пользователей в единицу времени получим целевую функцию вида:

$$\sum_{i=1}^{M} \sum_{I_j(T_j)} \lambda_i x_{ij} P_i \rightarrow \max . \quad (4)$$

При выборе приоритетов оцениваются следующие характеристики заявки:

- время нахождения заявки в очереди;
- текущая длина очереди заявок;
- интенсивность обращения к каждому из компонентов ресурса необходимых для выполнения заявки.

Выбор приоритетов и оценка текущей ресурсоемкости задачи производится на основе компонентов ресурса, имеющих индивидуальные пороговые значения связанные с физическими ограничениями оборудования.

В ходе реализации предложенной модели в распределенной информационной системе дистанционного обучения нами получены следующие показатели работы, позволяющие оценить эффективность применения разработанного алгоритма расстановки приоритетов. Анализ производился на промежутке времени ΔT= 60 секунд. Ограничение по времени обусловлено техническими параметрами (максимально допустимым временем отклика) работы приложения. Эффективность работы алгоритма приоритетов будем оценивать путем сравнения очереди (общего количества заявок) одновременно находящихся в системе, и количества отброшенных заявок. На рисунке 3 представлена диаграмма обслуживания заявок в реально работающей системе без использования предложенного алгоритма.

Рисунок 3 - Диаграмма обслуживания заявок без использования алгоритма расстановки приоритетов

Применив алгоритм выбора и расстановки приоритетов для каждого из ресурсов в рамках всей системы дистанционного обучения получим снижение количества отброшенных заявок в каждый момент времени примерно 2,7 раза, при этом общее число необработанных заявок по истечению времени обработки ΔT снизилось с 12 до 5 (Рисунок 4).

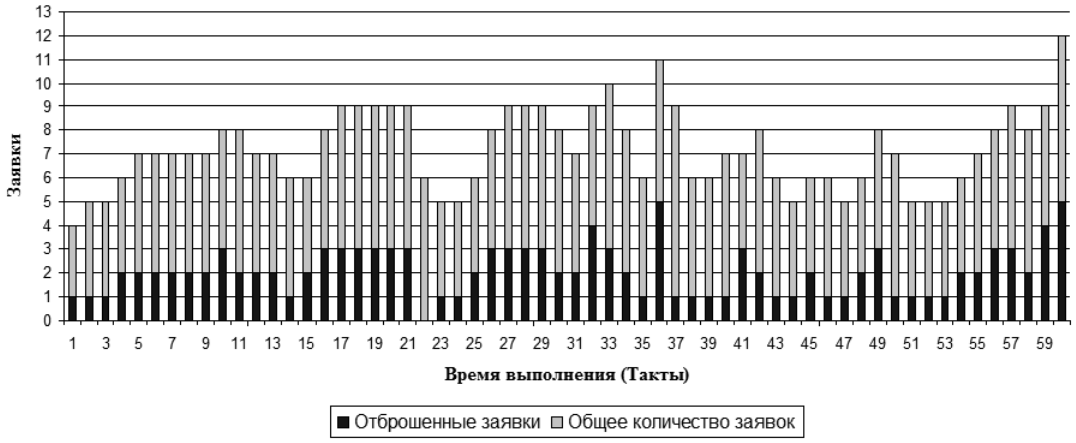

Рисунок 4 - Диаграмма обслуживания заявок с использованием алгоритма расстановки приоритетов

Как можем заметить, наблюдается самоподобие графиков обслуживания заявок в информационной системе. Нами проведено дополнительное исследование по оценке времени отклика системы, показавшее прирост скорости обработки заявок, по сравнению с обычной обработкой, так как средняя дина очереди снизилась с 8,6 до 5,1 (Рисунок 5).

Рисунок 5 - Диаграмма динамики выполнения заявок

Экспериментальная апробация алгоритма проведена на симуляторе, моделирующем распределение нагрузки с использованием имитационной модели процесса взаимодействия пользователей с мультимедийными сервисами. Построенная модель и приведенный алгоритм могут применяться для повышения эффективности использования аппаратных и программных ресурсов с целью улучшения качества предоставления услуг в распределенных информационных системах дистанционного обучения, а также предотвращения перегрузки сервисов в момент пиковой нагрузки.

СПИСОК ЛИТЕРАТУРЫ

1. *Парфёнов Д.И.*, Технологии и инструментальные средства организации и проведения вебинаров в системе дистанционного обучения //Тр. IX всероссийской научно-практической конференции с международным участием «Современные информационные технологии в науке, образовании и практике». – Оренбург.:[Би], 2010. –С. 109–113.
2. *Вашкевич, Н. П.* Активные инфологические модели обработки данных на основе иерархических сетей фреймов / Н.П. Вашкевич, Н.С. Зинкина // Вопросы радиоэлектроники. Серия ЭВТ. Вып. 4, 2009. - С. 54-63.
3. *Зинкина, Н. С.* Агентно-ориентированный подход к проектированию распределенных систем управления базами данных / Н.С. Зинкина // Перспективы науки. № 2. - 2011. - С. 80-86.
4. *Бойченко, И. В.* Управление ресурсами в сервис-ориентированных системах типа «приложение как сервис» / И.В. Бойченко, С.В. Корытников // Доклады Томского государственного университета систем управления и радиоэлектроники Вып. 1-2, 2010. -С. 156-160.
5. *Жевнерчук, Д. В.* "Методика моделирования нагрузки на сервер в открытых системах облачных вычислений" / Д.В. Жевнерчук, А.В. Николаев // Информёи е примен., 2012, 43–50
6. *Петров, Д.Л.* Оптимальный алгоритм миграции данных в масштабируемых облачных хранилищах // Управление большими системами. Вып. 30, 2010. -С.180-197.
7. *Петров, Д.Л.* Динамическая модель масштабируемого облачного хранилища данных // Известия ЛЭТИ, #4, 2010. -С. 17-21
8. *Гусев, О. В.* Проблема адекватной оценки производительности веб-серверов в корпоративных сетях на предприятиях ЦБП / О.В. Гусев, А.В. Жуков, В.В. Поляков, С.В.Поляков // Материалы 6-й научно-технической конференции «Новые информационной технологии в ЦБП и энергетике».- Петрозаводск, 2004. – С. 84-87
9. *Жуков А.В.* Некоторые модели оптимального управления входным потоком заявок в интранет-системах. // Материалы 6-й научно-технической конференции «Новые информационной технологии в ЦБП и энергетике».- Петрозаводск, 2004. – С. 87-90.
10. *Парфёнов Д.И.*, Программно-аппаратный комплекс видеопортала как эффективное средство информационного взаимодействия субъектов образовательного процесса // Тр. V международной научно-практической конференции «Информационная среда ВУЗа XXI Века». – Петрозаводск.:[Би], 2011. – С. 141–144.
11. *Парфёнов Д.И.* , Моделирование востребованности ресурсов в распределенной информационной системе дистанционной поддержки образовательного процесса / И.П. Болодурина // Сборник статей под реакцией А.П. Кудинова «Высокие технологии, экономика, промышленность». – СПб.:[Би], 2012. – С. 30–34.