# МОДЕЛИРОВАНИЕ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ И ДИНАМИЧЕСКОЙ БАЛАНСИРОВКИ НАГРУЗКИ В СИСТЕМЕ ДИСТАНЦИОННОГО ОБУЧЕНИЯ

***И.П. Болодурина, Д.И. Парфёнов***

*Оренбургский госуниверситет*

Повседневное распространение информационных технологий все чаще находит свое применение в образовательных сетевых мультимедийных системах. 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.

## Введение

Повседневное распространение информационных технологий все чаще находит свое применение в образовательных сетевых мультимедийных системах. При этом наиболее активно развивающимся направлением является дистанционное обучение. В последнее время широкое распространение получили такие интерактивные веб-сервисы, построенные на базе современных информационных технологий, как:

- цифровое телевидение (TV over IP – TVoIP);
- видео по запросу (Video on Demand – VoD);
- Интернет-трансляции;
- вебинары;
- веб-конференции.

Однако, несмотря на широкий круг решаемых задач, повсеместное внедрение перечисленных ранее сервисов в образовательных учреждениях остается весьма затруднительным. Для обеспечения необходимого качества предоставляемых услуг требуется выделение специализированных аппаратных ресурсов, а также каналов связи.

Университеты обладают собственными аппаратными (серверными) ресурсами для обеспечения внутренней деятельности и арендуют каналы связи для обмена информацией посредством сети Интернет. 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. Большинство событий носят систематический характер, что позволяет рассчитать необходимую нагрузку и подготовить требуемые ресурсы.

В настоящее время для обеспечения хранения мультимедиаконтента 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 предостав-

ления информационных услуг в образовательных учреждениях. Это особенно актуально для дистанционного обучения, при котором основная часть учебного процесса и взаимодействие обучающегося и преподавателя осуществляется посредством сети Интернет.

На факультете дистанционных образовательных технологий (ФДОТ) Оренбургского государственного университета (ОГУ) 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 приводит к значительному росту трафика и, как следствие, повышению нагрузки на оборудование и каналы связи. Узким местом подобных 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. 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 Интернета заключается в том, что 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.

### 1. Постановка задачи

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

1. Нагрузка на серверы 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 используются пререгистрация (подписка на сервисы), например запись на вещание лекции, а также статистические данные оценки использования информационных ресурсов, полученные 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 обеспечения. 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 передан 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 мультимедиаконтента, 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).

**2. 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 аудиоматериалов (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. 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 систем 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 Интернет-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 многоканальную СМО с несколькими очередями (рис. 1).

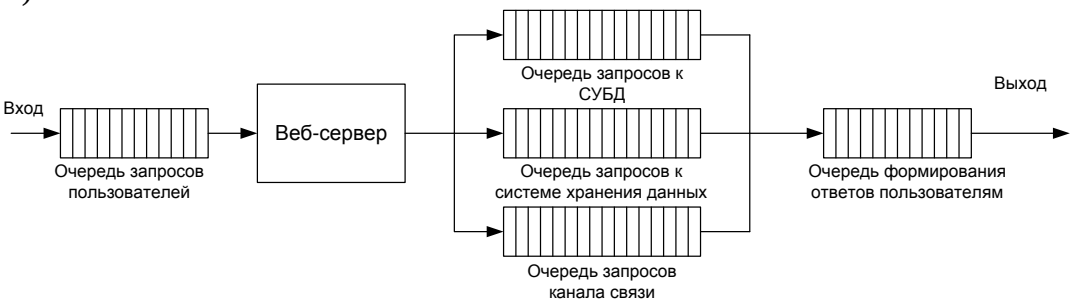


Рис. 1. Модель работы Интернет-приложения как СМО

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 как к дисковому пространству самого сервера, так и к системе хранения данных для записи или чтения 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:

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

где $R_1$, …, $R_n$ – численные показатели использования ресурса по каждому из классификационных признаков, полученные в результате измерений на интервале времени $\Delta T$.

Индикаторы приоритета обслуживания уровней модели определим на основе рейтинга востребованности ресурсов системы в целом. Анализируя интенсивность использования каждого из компонентов ресурсов в СДО, получим рейтинг востребованности ключевых сервисов и аппаратного обеспечения, лежащих в основе каждой из подсистем.

Общую ресурсоемкость системы дистанционного обучения определим как суммарную площадь $U_{\text{сдо}}$, 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, $i = 1, \ldots, M$).

Статус обработки $j$-й заявки, поступившей на $i$-й 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 $\lambda_i$, 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:

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

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:

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

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

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

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 $\Delta T$= 60 секунд. Ограничение по времени обусловлено техническими параметрами (максимально допустимым временем отклика) работы приложения. Эффективность работы алгоритма приоритетов будем оценивать путем сравнения очереди заявок (общего их количества), одновременно находящихся в системе, и количества отброшенных заявок. На рис. 2 представлена диаграмма обслуживания заявок в реально работающей системе без использования предложенного алгоритма.


Рис. 2. 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 в 2,7 раза, при этом общее число необработанных заявок по истечении времени обработки $\Delta T$ снизилось с 12 до 5 (рис. 3).

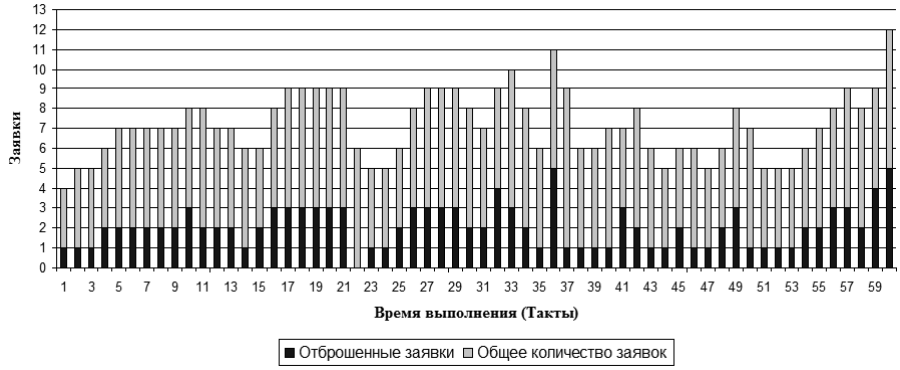


Рис. 3. Диаграмма обслуживания заявок с использованием алгоритма расстановки приоритетов

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 длина очереди снизилась с 8,6 до 5,1.

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.

Работа выполнена при финансовой поддержке программы «Научные и научно-педагогические кадры инновационной России», грант № 14.132.21.1801.

**Литература**

1. Парфёнов Д.И. Технологии и инструментальные средства организации и проведения вебинаров в системе дистанционного обучения // Труды IX Всероссийской научно-практической конференции с международным участием «Современные ин-

формационные технологии в науке, образовании и практике». – Оренбург: [Би], 2010. – С. 109–113.
2. Вашкевич Н.П., Зинкина Н.С. Активные инфологические модели обработки данных на основе иерархических сетей фреймов // Вопросы радиоэлектроники. Серия ЭВТ. 2009. Вып. 4. С. 54–63.
3. Зинкина Н.С. Агентно-ориентированный подход к проектированию распределенных систем управления базами данных // Перспективы науки. 2011. № 2. С. 80–86.
4. Бойченко И.В., Корытников С.В. Управление ресурсами в сервис-ориентированных системах типа «приложение как сервис» // Доклады Томского государственного университета систем управления и радиоэлектроники. 2010. Вып. 1-2. С. 156–160.
5. Жевнерчук Д.В., Николаев А.В. Методика моделирования нагрузки на сервер в открытых системах облачных вычислений // Информ. и её примен. 2012. С. 43–50.
6. Петров Д.Л. Оптимальный алгоритм миграции данных в масштабируемых облачных хранилищах // Управление большими системами. 2010. Вып. 30. С. 180–197.
7. Петров Д.Л. Динамическая модель масштабируемого облачного хранилища данных // Известия ЛЭТИ. №4, 2010. С. 17–21.
8. Гусев О.В., Жуков А.В., Поляков В.В., Поляков С.В. Проблема адекватной оценки производительности веб-серверов в корпоративных сетях на предприятиях ЦБП // Материалы 6-й научно-технической конференции «Новые информационные 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 информационные 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 // Труды V Международной научно-практической конференции «Информационная среда вуза XXI века». – Петрозаводск: [Би], 2011. – С. 141–144.
11. Парфёнов Д.И., Болодурина И.П. Моделирование востребованности ресурсов в распределенной информационной системе дистанционной поддержки образовательного процесса // Сборник статей под реакцией А.П. Кудинова «Высокие технологии, экономика, промышленность». – СПб.: [Би], 2012. – С. 30–34.