УДК 004.94

РАСЧЕТ ХАРАКТЕРИСТИК СЕРВЕРНЫХ КОМПЬЮТЕРНЫХ СИСТЕМ

Аль Абабнех Хасан, Иваница С.В., Аноприенко А.Я.
Донецкий национальный технический университет
anoprien@cs.dgtu.donetsk.ua, isv@cs.dgtu.donetsk.ua

Рассматриваются способы и инструменты расчета характеристик серверных компьютерных систем. Поставлены задачи анализа производительности систем. Выявлена методика экспериментальных оценок и методология математического анализа экспериментальных данных. Рассмотрены подходы к постановке экспериментов, в частности обусловлено проведение факторного эксперимента основанного на методике экспериментальных оценок.

Введение

Вычислительные и коммуникационные системы с каждым днем все глубже входят в повседневную жизнь и становятся незаменимыми средствами для качественной «интеллектуальной активности» человека. Такие системы эффективны только в том случае, когда они основаны на использовании больших распределенных структур: Internet и Web [1]. В последнее время обозначилась следующая тенденция: эффективность работы как коммерческих организаций так и частных лиц все больше зависит от web-служб, причем рост эффективности их работы прямо пропорционален увеличению динамичности и росту функциональных возможностей web-служб.

Работоспособность Web-служб основана на эффективной организации **серверных компьютерных систем** (**СКС**), состоящих из тысяч компьютеров и программных компонентов. Очевидно, что для понимания, анализа, разработки и управления такими системами нужны **количественные методы и модели**, которые помогают оценить различные сценарии функционирования, исследовать структуру и состояние больших систем. Постоянно увеличивающийся спрос на web-службы обеспечит актуальность проблем, связанных с недостаточной производительностью СКС, и, в конце концов, они станут превалирующими при планировании и вводе в эксплуатацию новых web-служб при постоянно увеличивающемся количестве пользователей Internet. Таким образом, чтобы избежать многих проблем, связанных с несоответствием реальной и ожидаемой пользователями производительности СКС, необходимы специальные методики планирования производительности и количественного исследования поведения web-служб.

При количественном подходе к анализу web-служб используют три типа моделей – **модель загрузки** (основана на понимании и описании нагрузки [2]), **модель производительности** (основана на количественном описании поведения системы) и **модель готовности** (основана на описании задержек в сетевых средах). Следует выделить сложность построения модели нагрузки, поскольку процесс описания параметров нагрузки требует существенного объема работ. Некоторые рассматриваемые в модели нагрузки характеристики связаны с распределением размеров файлов, распределением популярности файлов, самоподобии web-трафика и шаблонов пользовательских запросов. Исследования различных web-сайтов показали, что размеры файлов подчиняются закону медленно убывающего распределения, а популярность объектов определяется законом Ципфа [3], т. е. имеет **ципфово распределение** (the Zipf distribution): частота F встречаемости объекта, примерно обратно пропорциональна его рангу r в частотном распределении «ранг–частота» и выражается соотношением

$$F = cr^{-1}, \tag{1}$$

где $c = Fr$ – константа.

Направлением исследований в данной работе являются особенности применения методики экспериментальных оценок производительности к классу трехзвенной (трехуровневой) модели СКС

[4] выражающиеся в выборе средств математического анализа эмпирических данных для составления прогноза производительности.

Задачи анализа производительности

Как правило, при разработке приложений класса клиент-сервер появляются задачи анализа производительности [5], которые указаны на рис. 1. Для решения этих задач применяют техники экспериментального исследования и математического моделирования. При проведении математического моделирования используют **аналитические** или **имитационные модели**, условное различие между которыми заключается лишь в выборе применяемого математического аппарата. При этом техники анализа производительности СКС (*ТАП*) могут быть представлены следующей логической формулой:

$$ТАП = ММ \vee ЭИ = (ИМ \wedge АМ) \vee ЭИ, \quad (2)$$

где $ММ = ИМ \wedge АМ$ – проведение одного из видов математического моделирования *ММ*: имитационного *ИМ* или аналитического *АМ*; *ЭИ* – проведение экспериментального исследования.

Приложение класса клиент-сервер имеет клиент-серверную архитектуру (клиент – web-браузер, сервер – web-сервер, протокол обмена – http-протокол). Поэтому приложение с трехуровневой архитектурой – это приложение, функциональность которого реализована на трех уровнях:

1. Presentation Layer (PL) – уровень представления данных: функциональность приложения заключается в обработке поступающих запросов.
2. Application Layer (AL) – уровень обработки данных: приложение функционирует посредством вызова методов бизнес-логики [6].
3. Storage Layer (SL) – уровень хранения данных: функциональные вызовы реализуются в запросы доступа к данным информационных хранилищ.

Учитывая реализацию функциональности такого приложения на каждом из уровней, можно выделить следующие особенности трехуровневых web-приложений с позиции анализа производительности:

- непредсказуемый характер http-трафика;
- значительный диапазон размера передаваемых по сети данных;
- наличие разноархитектурных серверов, которые требуют отдельного исследования, т. е. в рамках одного приложения звенья архитектуры могут быть разного типа (например, для SL – файловая система; для AL –компоненты бизнес-логики);
- модульное строение web-серверов, позволяющее выбирать оптимальную конструкцию с позиции производительности.

Подходы к постановке экспериментов

При проведении эксперимента, в котором варьируется только один фактор, а другие значения других влияющих параметров просто фиксируются, отсутствует возможность выявить совместное влияние двух или нескольких факторов, и, следовательно, отсутствует возможность получения достаточно точных экспериментальных данных. Такой, называющийся **простым**, эксперимент является недостаточным для проведения экспериментального исследования анализа производительности СКС. Поэтому необходимым и достаточным условием для получения достоверных результатов

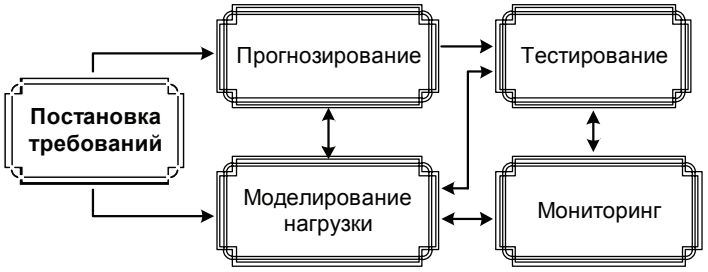

Рисунок 1. Основные задачи анализа производительности модели для СКС

является проведение **факторного эксперимента**, в котором варьируются значения интересуемых и влияющих друг на друга факторов. Для получения экспериментальных данных для СКС можно адаптировать **методику экспериментальных оценок** производительности, описанную в [5], которая выработана на основе практического опыта по исследованию производительности реальных систем. Методика экспериментальных оценок применима для решения задач тестирования и прогнозирования производительности трехуровневой серверной инфраструктуры и предполагает следующий порядок действий:

- исследование особенностей применения методики к реальному web-приложению с трехуровневой архитектурой;
- постановка экспериментов с целью получения оценок производительности;
- выбор средств математического анализа эмпирических данных для прогнозирования производительности;
- составление прогноза производительности исходя из анализа полученных данных;
- исследование применимости регрессии для прогнозирования производительности web-приложений.

В качестве содержательной части подобного эксперимента выступают параметры, метрики, методы измерения и анализа, а в качестве целей – формулируются следующие задачи: проверка требований по производительности; оптимизация; предсказание поведения при различных внешних условиях.

Рассмотрим более подробно составляющие содержательной части эксперимента. Величины, влияющие на производительность, определяют **параметры производительности**, которые принято разделять на системные (аппаратные характеристики, конфигурация ПО, и т. п.) и параметры нагрузки.

Величина, характеризующая производительность целого приложения или его составляющих является **метрикой производительности**, и в своей совокупности представляет следующий ряд параметров:

- Response Time (время обработки запроса) – время между отправкой первого байта запроса и получения последнего байта ответа;
- Reaction Time (время реакции) – время между отправкой первого байта запроса и получения первого байта ответа;
- Latency (латентность) – время обработки запроса на сервере;
- Throughput (пропускная способность) – количество запросов, обрабатываемых в единицу времени;
- Utilization (утилизация ресурса) – время, которое ресурс тратит на обработку запросов из очереди;
- Service demand (сервисное время ресурса) – время, затрачиваемое ресурсом на обработку одного запроса (http-запрос).

Измерение метрик производительности можно осуществлять, используя следующие стратегии [7]: **event-driven** (управляемые событием измерения) – для регистрации редкослучающихся событий; **sampling** (осуществление выборок) – периодическое получение «снимков» состояния СКС.

При проведении прямых измерений утилизации ресурсов (например, средствами ps, sar, iostat, NT Performance Monitor) косвенно вычисляется среднее сервисное время и пропускная способность СКС по графику утилизации от частоты запросов к системе. Значения времени ответа, отклика и латентность можно получить путем выборки необходимых данных из лог-файлов web-сервера и/или программы-генератора нагрузки.

Проведение качественных экспериментов предполагает генерирование реалистичной нагрузки, которая бывает двух видов: **реальная** – нагрузка, которую испытывает система в действительности и **синтетическая** – искусственная, перенимающая существенные свойства реальной нагрузки. Поскольку реальная нагрузка обладает сложногенерируемостью и неповторяемостью, то при получении экспериментальных оценок производительности обычно используют синтетическую

нагрузку – по сути являющуюся **моделью реальной нагрузки**. Таким образом, синтетическая нагрузка разрабатывается либо по результатам измерений реальной нагрузки на систему, либо на основе некоторых общих предположений о будущем поведении пользователей СКС.

В рамках моделирования нагрузки решаются следующие задачи:

– построение модели нагрузки (модель трафика, модель поведения пользователей, и др.);
– разработка методов генерирования и верификации свойств нагрузки.

Схема проведения экспериментов (простая, полнофакторная, дробнофакторная) [7] определяется **планом наблюдений**, в котором также фиксируется количество повторных наблюдений для каждого заданного набора значений факторов.

Стоит отметить, что экспериментальный подход для оценки и прогнозирования производительности можно применять для исследования сложных систем, неподдающихся математическому моделированию.

Методология математического анализа эксперименталь-ных данных

Математический анализ экспериментальных данных предполагает получение и исследование расчетных характеристик СКС и представляет собой математическую модель для прогнозирования производительности системы. В рамках анализа производительности СКС решаются следующие задачи:

1. Capacity Planning (планирование мощностей) – прогнозирование вычислительных и операционных ресурсов системы;
2. Сравнительный анализ различных проектных решений с позиции требований к производительности.

В рамках повышения эффективности СКС математический анализ экспериментальных данных позволяет также решать **прямую** и **обратную задачи** [8], как в статическом, так и в динамическом режимах.

Для качественного построения математической модели для прогнозирования производительности СКС целесообразно использовать следующие техники математического анализа экспериментальных данных:

– анализ фрактальности;
– кластерный анализ;
– пуассоновский процесс;
– анализ грубых погрешностей;
– статистическая оценка параметров распределения (с последующей проверкой гипотезы о типе распределения);
– регрессионный анализ;
– корреляционный анализ;
– дисперсионный анализ.

При исследовании сетевого трафика актуальной задачей является учет его **фрактальных свойств**, т. е. фактически применение фрактальных методов к анализу временных рядов (совокупности наблюдаемых параметров изучаемой системы во времени). Одним из самых перспективных направлений фрактального анализа является изучение динамики во времени такой характеристики, как фрактальная размерность (D). Один из способов для исследования фрактальных временных рядов был предложен Бенуа Мандельбротом и базируется на исследованиях проведенных английским исследователем Херстом и носит название **R/S метода** [10]. Он построен на анализе размаха параметра (наибольшим и наименьшим значением на изучаемом отрезке) и среднеквадратичного отклонения.

Для фрактальных временных рядов на интервале $t_0 < t < T$ размах параметра R $R(\tau) = \max\limits_{1 \le t \le \tau} B(t) - \min\limits_{1 \le t \le \tau} B(t)$ зависит от времени t степенным образом:

$$R(t) = R(t_0) \cdot \left(\frac{t}{t_0} \right)^{2-D}, \qquad (3)$$

где $D = \lim_{\delta \to 0} \frac{\ln B(\delta)}{\ln\left(\frac{1}{\delta}\right)}$ – фрактальная размерность временного ряда B.

Исходя из выражения (3) можно предсказать возможное значение размаха интересующего параметра в будущем. Фрактальная размерность, является показателем сложности кривой. Анализируя чередование участков с различной фрактальной размерностью и тем, как на систему воздействуют внешние и внутренние факторы, можно научиться предсказывать поведение СКС. И что самое главное, диагностировать и предсказывать ее нестабильные состояния.

Кластерный анализ используется для объединения в группы (кластеры) близких по некоторым характеристикам объектов (например, экспериментальные точки в многопараметрическом пространстве). Перед проведением кластерного анализа данные обычно нормируют на минимум-максимум, в пространстве параметров определяют метрику, например, евклидово расстояние. Евклидово расстояние d между двумя точками $\alpha_1 = (D_{i1}, D_{i2}, \ldots, D_{in})$ и $\alpha_2 = (D_{j1}, D_{j2}, \ldots, D_{jn})$ в n-мерном пространстве ($1 \le i, j \le p$) определяется выражением

$$d = \sqrt{\sum_{k=1}^{n} (D_{ik} - D_{jk})^2}. \quad (4)$$

Широко используются следующие два алгоритма [11]: **минимальное стягивающее дерево** (MST – Minimal Spanning Tree) и **метод k-средних** (k-means algorithm).Цели кластеризации – понимание данных путем выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).

Последующие техники математического анализа экспериментальных данных представляют собой классические статистические методы исследования и широко освещены в работах [12–17].

Таким образом, использование вышеперечисленных техник математического анализа экспериментальных данных позволяет построить качественную математическую модель. Причем качество прогнозов зависит от погрешности и количества измерений, уровня статистической достоверности. Средствами статистической обработки эмпирических данных могут быть получены точечные и интервальные оценки производительности СКС в целом и отдельных компонент ее архитектуры. Исследование влияния параметров конфигурации системы и нагрузки на производительность приложения можно провести, используя методы дисперсионного анализа. Средствами линейной и нелинейной регрессии составляется предсказание производительности системы при различной интенсивности нагрузки.

Выводы

В современном мире проблемы, связанные с оценкой производительности СКС усугубляются с каждым годом, поскольку интенсивно возрастает сложность развивающихся сетей, коими является Web и Internet. Поэтому в сфере оценки производительности и анализа компьютерных и сетевых систем можно выделить три существенные критические области:

1. Возрастающая сложность современных и будущих СКС.
2. Необходимость повышения уровня образования специалистов, занимающихся вопросами производительности.
3. Необходимость принятия методик оценки производительности в качестве стандарта при разработке и реализации СКС.

Использованная в докладе методика позволяет с минимальными затратами на проведение экспериментов получить максимальную информацию об эффективности использования серверных компьютерных систем. Фактически в докладе предпринята попытка покрыть все упомянутые критические области.

Литература

[1] National Academy of Sciences, Making IT Better: Expanding Information Technology Research to Meet Society's Needs, National Academy Press, Washington, D.C., 2000.

[2] Аноприенко А.Я., Аль Абабнех Хасан. Модели нагрузки в веб-ориентированных компьютерных сетях // Научные труды Донецкого национального технического университета. Серия «Проблемы моделирования и автоматизации проектирования динамических систем» (МАП–2008). Выпуск 7 (150): Донецк: ДонНТУ, 2008. С. 258–274.

[3] Li Wentian Random Texts Exhibit Zipf's-Law-Like Word Frequency Distribution. IEEE Transactions on Information Theory. – Santa Fe Institute, 1660 Old Pecos Trail, Suite A. – Santa Fe, NM 87501: 1992. В. 38. № 6. С. 1842–1845.

[4] Аль Абабех Хасан, Аноприенко А.Я. Способы и инструменты расчета параметров серверных компьютерных систем // Материалы третьей международной научно-технической конференции «Информационные управляющие системы и компьютерный мониторинг», Донецк, ДонНТУ, 2011. Т.3. С. 276–282.

[5] Raj J. The art of computer systems performance analysis. – NY.: Wiley Computer Publishing, 1992.

[6] Бизнес-логика. Материал из Википедии – свободной энциклопедии. Электронный ресурс. Страница доступа: http://ru.wikipedia.org/wiki/Бизнес-логика.

[7] Lilja D. J. Measuring Computer Performance: a practioner's guide. – UK.: Cambridge University Press, 2000.

[8] Аноприенко А.Я., Аль Абабнех Хасан. Повышение эффективности Интернет-ориентированной сетевой инфраструктуры: Методы, задачи и инструменты // Научные труды Донецкого национального технического университета. Серия «Проблемы моделирования и автоматизации проектирования динамических систем» (МАП-2007). Выпуск 6 (127): Донецк: ДонНТУ, 2007. С. 228-233.

[9] Leland W.E., Taqqu M.S., Willinger W., Wilson D.V. On the selfsimilar nature of ethernet traffic (extended version). IEEE/ACM Transactions of Networking, 2(1): 1–15, February 1994.

[10] Цветков И.В. Фрактальный анализ и его применение к исследованию временных рядов. // Материалы для семинара с преподавателями Кентского университета, 1 апреля 2002 года. Электронный ресурс. Страница доступа: http://russeca.kent.edu/SeminarTsvetkovEng.pdf.

[11] Menasce D.A., Almeida V.A.F. Capacity Planning for Web Performance: Metrics, Models and Methods. – NJ.: Prentice Hall, 1998.

[12] Парфенов В.Г., Статистические методы исследования в оптическом приборостроении. – Учебное пособие, ЛИТМО, 1980.

[13] Парфенов В.Г. Регрессионный и корреляционный анализ. Обработка результатов наблюдений при измерениях. – Учебное пособие, ЛИТМО, 1983.

[14] Гмурман В.Е., Теория вероятностей и математическая статистика: Учебное пособие для вузов. – М.: Высшая школа, 2002.

[15] Бородин А.Н. Элементарный курс теории вероятностей и математической статистики. Серия « Учебники для вузов. Специальная литература» – СПб.: Издательство «Лань», 1999. – 224 с.

[16] Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. – 2-е изд. – М.: ФИЗМАТЛИТ, 2005. – 296 с.

[17] Вуколов Э.Л. Основы статистического анализа. Практикум по статистическим методам и исследованию операции с использованием пакетов STATISTICA и EXCEL: учебное пособие. – 2-е изд., исправ. и доп.. – М.: ФОРУМ. 2008. – 464 с.