Повышение эффективности Интернет-ориентированной Сетевой инфраструктуры: Методы, задачи и инструменты

Аноприенко А.Я., Аль Абабнех Хасан
Кафедра ЭВМ ДонНТУ
anoprien@cs.dgtu.donetsk.ua

Abstract

Anoprienko A., Ababneh H. Providing of efficiency for Internet-focused network infrastructure: methods, problems and tools. *Main variants of Internet-focused network infrastructure efficiency providing on the base of simulation and Excel-tools are described.*

Введение

Усложнение современной инфраструктуры Интернет и чрезвычайно быстрые темпы роста интенсивности ее использования приводят к постоянному росту требований к аппаратным и программным средствам. В этих условиях особую актуальность приобретает эффективность использования имеющихся ресурсов. В работах [1-4] рассмотрены различные подходы к решению данной проблемы. В данной статье анализируются соответствующие методы и инструменты и предлагается подход, основанный на реализации численных моделей инфраструктуры на базе электронных таблиц, что является развитием идей, предложенных в работе [5].

Прямая и обратная задача повышения эффективности

При решении проблем, связанных с повышением эффективности Интеренет-ориентированной сетевой инфраструктуры приходится решать одну из следующих 2-х задач:

Прямая задача: при заданных характеристиках аппаратных средств (и вообще инфраструктуры) определить оптимальную и пиковую нагрузку.

Обратная задача: при заданной нагрузке определить требуемые характеристики инфраструктуры.

Обе задачи могут решаться как в статике, так и в динамике. В последнем случае речь может идти об определенных критических интервалах, для которых необходимо найти достаточно эффективное решение.

Соответственно можно вести речь о **прямом методе** решения задачи обеспечения эффективности, предполагающем расчет возможной нагрузки

при заданных характеристиках аппаратных и программных средств, и **обратном**, позволяющем определить требования к аппаратным и программным средствам исходя из характеристик имеющейся и прогнозируемой нагрузки. На практике наиболее целесообразным является использование **комбинированного метода**, в рамках которого вначале определяется диапазон возможных нагрузок для некоторого набора характеристик инфраструктуры, а затем уточняются конкретные требования к аппаратным и программным средств исходя из планируемой нагрузки.

Модели рабочей нагрузки

При реализации любого из перечисленных выше методов одной из важнейших составляющих является достаточно адекватная модель рабочей нагрузки.

Наиболее просто могут быть реализованы идеализированные модели, ориентированные на какой-либо отдельный вид нагрузки:

вычислительная нагрузка, измеряемая в единицах операций в секунду (например, в MIPS – миллионах операций в секунду) или более сложными способами (например, в миллионах операций с плавающей запятой на каждый сеанс связи или на каждый запрос);

нагрузка, связанная с обращением к базам данных и измеряемая либо в эквивалентной вычислительной нагрузке (например, в виде MIPS) или в более специфических единицах (например, в количестве транзакций в секунду);

нагрузка, связанная с хранением и выдачей определенного контента (при этом в байтах могут учитываться и статическая, и динамическая составляющие).

В последнем случае основными следует считать следующие виды контента:

Текстовый – архивированный (zip, rar и др.), неразмеченный, HTML, pdf, rtf (типичные объемы для текстов 1, 10, 100 Кбайт, 1 Мбайт).

Графический – неоптимизированный (bmp, tiff, pcx) и оптимизированный (jpg, gif, png, jv) объемами от 10 Кбайт до 10 Мбайт.

Аудиоконтент – от 10 Кбайт до 10 Мбайт с ярко выраженным максимумом в диапазоне 2-5 Мбайт, представленный в виде фиксированных записей или потоков (радио и телефония).

Видеоконтент – записи и сеансовые потоки объемом от 1 Мбайта до 10 Гбайт, а также непрерывные потоки (web-камеры, видеотелефония).

На практике наиболее целесообразным является использование различных видов комбинированных моделей контента:

текст (в HTML) и графика;

текст, графика и аудио;

текст, графика и видео (с различной степенью преобладания последнего).

Соответственно и модели нагрузки также наиболее целесообразно использовать комбинированные, например, следующие:

сочетание вычислительной нагрузки с интенсивным обращением к базам данных (в этом случае можно использовать общие единицы измерения в виде, например, MIPS, но специфику баз данных учитывать крайне желательно);

сочетание вычислительной нагрузки с выдачей определенного контента, характеристики которого существенно влияют на эффективность системы в целом.

И, хотя в простейшем случае нагрузку можно считать неизменной, следует в обязательном порядке учитывать определенные особенности ее динамики, в том числе ярко выраженную суточную и недельную периодичность. Наиболее целесообразной при этом является ориентация на пиковую нагрузку.

В современной сетевой среде одной из наиболее крупных единиц вычислительной рабочей нагрузки можно считать виртуальную машину с запущенной одной копией приложения с высокой интенсивностью вычислений.

Особенности современной сетевой инфраструктуры

Необходимость повышения эффективности сетевой инфраструктуры связана в первую очередь со стремлением владельцев серверных ресурсов снизить совокупную стоимость владения. Основная проблема при этом заключается в слишком низкой полезной загрузке серверов. Например, известно, что в настоящее время в большинстве корпоративных серверов Intel вычислительные возможности центральных серверов использовались менее чем на 20%. Причина в том, что в большинстве фирм исторически сложилось так, что на каждом сервере работает всего одно приложение [6].

Одно из решений заключается в консолидации нескольких задач на одном сервере. При этом существенно снижается совокупная стоимость владения парком серверов. Еще одно решение заключается в создании сервис-ориентированных центров данных. Речь идет о гибкой и оперативной вычислительной среды, основанной на виртуализации. При этом предполагается автоматическое распределение задач по виртуальным машинам. Такая модель сервис-ориентированной среды оптимально реализуется на современных многоядерных процессорах [6].

Опыт создания таких сред показал, что пока количество виртуальных машин не превышает количество ядер наблюдается линейный рост общей производительности и стабильное время выполнения задач. При равном числе машин и ядер обеспечивается примерно 90% эффективность использования вычислительных ядер. В дальнейшем происходит

практически линейный рост времени выполнения, существенное влияние на который оказывает и рост затрат времени на переключение виртуальных машин между имеющимися ядрами [6].

Аналогичные проблемы возникают и при реализации различных кластерных систем, а также в процессе формирования GRID-инфраструктуры.

Подход Multi-Core Chip становится в настоящее время универсальным подходом при развитии аппаратных средств. При этом реализуются как универсальные вычислительные ядра, так и спецядра для видео, графики, DSP и прочих целей. В целом следует констатировать, что заканчивается «эра гонки ГГерц» и начинается «эра гонки ядер», в том числе специализированных. А для эффективного использования преимуществ этого этапа необходимы не только новые алгоритмы на 100, 1000 и более ядер, но и достаточно адекватные модели и инструменты моделирования, позволяющие оперативно оценивать уровень соответствия инфраструктуры реальной и прогнозируемой нагрузке.

Вычислительная модель серверной инфраструктуры

Как показали проведенные авторами исследования, для достижения требуемого уровня эффективности в большинстве случаев достаточным является оценочное моделирования инфраструктуры при известных параметрах нагрузки.

Соответствующие инструменты достаточно эффективно могут быть реализованы в среде Excel. При могут быть использованы и усовершенствованы решения предложенные в работе [5].

Пример модели типичной серверной инфраструктуры, разработанной на основе подходов, предложенных в работе [5], представлен на рис. 1. Обозначения на рисунке приняты следующие: ISP – Internet Service Provider, обеспечивающий подключение серверного пула к глобальной инфраструктуре Интернет; LAN1 и LAN2 – локальные сети, обеспечивающие взаимодействие серверов, WS – Web-server, обслуживающий запросы по HTTP-протоколу; AS – Application-server, реализующий бизнес-логику серверного приложения; DB – DataBase-server, обеспечивающий работу с базами данных приложения. При этом CPU означает процессорную часть соответствующих серверов, а IO – подсистему ввода-вывода.

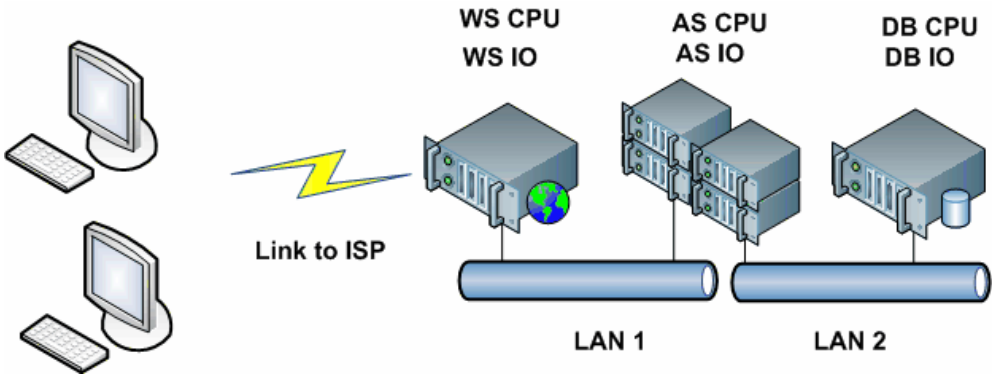

Рисунок 1 - Модель типичной серверной инфраструктуры

На рисунке 2 показана реализация вычислительной модели представленной на рис. 1 серверной инфраструктуры в среде Excel. При этом в качестве приложения используется рассмотренный авторами работы [5] (в своих последующих публикациях) пример сервера электронной коммерции, ориентированного на торговлю автомобилями. Но, в отличие от исходного варианта, предлагаемый вариант предполагает возможность достаточного гибкого задания как параметров загрузки, так и характеристик аппаратных средств.

Производительность			**1000**	**500**	**2500**	**2500**	**2000**	**1800**	**Пропускная способность**			
			MIPS	MIPS	MIPS	MIPS	MIPS	MIPS	**LAN 1**	**LAN 2**	**Link to ISP**	
Базовое количество		**90 000**							**10**	**100**	**1,5**	
запросов		запр/час							Mbps	Mbps	Mbps	
			Время обслуживания одного запроса (мс)									
E-business Function	**F**	**Кол-во запросов**	**WS CPU**	**WS IO**	**AS CPU**	**AS IO**	**DB CPU**	**DB IO**	**LAN 1**	**LAN 2**	**Link to ISP**	**Общее время**
Выбор автомобиля	1	90 000	5,00	10,00	21,60	7,60	14,50	16,11	4,92	0,53	16,38	96,64
Выбор опций	2	85 500	4,80	9,60	19,20	19,20	24,00	26,67	3,28	0,35	12,01	119,11
Выбор цвета	3	63 000	4,70	9,40	19,20	11,20	14,00	15,56	2,87	0,33	12,01	89,27
Доп.сервис	4	36 000	5,10	10,20	21,20	9,20	11,50	12,78	2,95	0,49	11,47	84,89
Ввод перс.данных	5	3 600	30,50	7,00	12,20	12,20	0,00	0,00	6,55	0,00	32,77	101,22
Оплата	6	2 700	32,00	6,40	12,80	12,80	0,00	0,00	7,78	0,00	21,85	93,63
Заказ доставки	7	2 700	30,00	6,00	12,00	12,00	0,00	0,00	4,10	0,00	19,11	83,21
Доставка	8	2 700	31,00	6,20	12,40	12,40	15,50	17,22	8,19	0,90	43,69	147,51
Проверка состояния	9	1 800	5,20	10,40	8,84	8,84	11,05	12,28	2,05	0,23	27,31	86,19
Отмена заказа	10	900	5,30	10,60	9,20	9,20	11,50	12,78	2,46	0,26	30,04	91,33
Процент утилизации мощности			47,6%	77,3%	**159,1%**	98,1%	**131,3%**	**145,8%**	30,4%	3,3%	**113,7%**	

Процент утилизации мощности системы при распараллеливании						
Коэфициент эффективности распараллеливания						**0,75**
Кол-во пар. комп.	**WS CPU**	**WS IO**	**AS CPU**	**AS IO**	**DB CPU**	**DB IO**
1	47,6%	77,3%	**159,1%**	98,1%	**131,3%**	**145,8%**
2	31,8%	51,6%	**106,1%**	65,4%	87,5%	97,2%
4	15,9%	25,8%	53,0%	32,7%	43,8%	48,6%
8	7,9%	12,9%	26,5%	16,3%	21,9%	24,3%

Рисунок 2 - Реализация вычислительной модели серверной инфраструктуры в среде Excel

Желтым цветом в таблице отмечены те значения, которые показывают критическую перегрузку соответствующих ресурсов. При представленной в таблице комбинации исходных значений необходимой и

достаточной является серверная система, состоящая из одного веб-сервера, 4-х серверов приложений и 2-х серверов баз данных.

Заключение и перспективы исследований

Проведенные исследования показали эффективность предложенного подхода на фоне относительной простоты его реализации, что позволяет признать целесообразной реализацию целого комплекса моделей, подобных описанной в данной работе, позволяющих производить достаточно точную оценку адекватности различных элементов сетевой инфраструктуры реальной и прогнозируемой нагрузке.

Литература

1. Аноприенко А.Я., Святный В.А. Высокопроизводительные информационно-моделирующие среды для исследования, разработки и сопровождения сложных динамических систем // Научные труды Донецкого государственного технического университета. Выпуск 29. Серия "Проблемы моделирования и автоматизации проектирования динамических систем" - Севастополь: «Вебер». - 2001. - С. 346-367.
2. Аноприенко А. Я., Джон С. Н. Рычка С. В. Особенности моделирования и оценки эффективности работы сетевой инфраструктуры // Наукові праці Донецького державного технічного університету. Серія: Обчислювальна техніка та автоматизація. Випуск 38. – Донецьк: РВА ДонДТУ, 2002. – С. 205-210.
3. Аноприенко А. Я., Потапенко В.А. WEB-ориентированная среда для интеграции моделирующих, вычислительных и информационных сервисов // Научные труды Донецкого национального технического университета. Выпуск 70. Серия: «Информатика, кибернетика и вычислительная техника» (ИКВТ-2002): - Донецк: ДонНТУ, 2003. - С. 61-70.
4. Аноприенко А. Я., Рычка С. В., Хасан Аль Абабнех. Способы и средства моделирования вычислительных сетей с целью обеспечения эффективности функционирования web-сервисов. – Моделирование и компьютерная графика: Материалы 1-й международной научно-технической конференции, г. Донецк, 04-07 октября 2005 г. – Донецк, ДонНТУ, Министерство образования и науки Украины, 2005. – С. 156-159.
5. Менаске Д., Алмейда В. Производительность Web-служб. Анализ, оценка и планирование: Пер. с англ. – СПб: ООО «ДиаСофтЮП», 2003. – 480 с.
6. Карпентер Р. Сравнение многоядерных процессоров для виртуализации серверов. Intel, август 2007.

Дата надходження до редакції 29.12.2007 р.