УДК 004.72

МОДЕЛЬ ПРЕДОСТАВЛЕНИЯ УСЛУГ ПО РАЗМЕЩЕНИЮ РЕСУРСОВ В КОРПОРАТИВНЫХ ЦЕНТРАХ ОБРАБОТКИ ДАННЫХ

***Аль-Хаками Али Мохаммед Омар*[1],**
аспирант
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

Определяется эффективная архитектура центра обработки данных с помощью анализа соотношений между временем ответа на клиентский запрос, производительностью, показателем использования ресурса, размещаемого в центре обработки данных, и его пропускной способностью.

Ключевые слова *— центр обработки данных, архитектура центра обработки данных, модель центра обработки данных, характеристики центра обработки данных.*

Введение

С увеличением объемов баз данных предприятий, развитием глобальных коммуникаций и электронной коммерции приобретенные пару лет назад компьютеры не справляются с новыми приложениями, а модернизация корпоративных серверов перерастает в дорогостоящую и трудоемкую процедуру. Не менее остро стоит задача обеспечения постоянного доступа к информации. Разрушительность последствий при возможной потере информации и доступа к ней диктует необходимость искать пути снижения такого риска.

Одним из наиболее эффективных механизмов решения этой проблемы является построение корпоративных центров обработки данных (ЦОД), в которых концентрируются важные вычислительные и информационные ресурсы, поддерживающие работу бизнес-приложений. Автор статьи попытался определить наилучшую архитектуру ЦОД, обеспечивающую минимальное время отклика на запрос клиента, не прибегая к масштабированию.

Рассмотрим ЦОД, в котором среднее время ответа на запрос клиента T больше, чем требуется. Пусть имеются некоторые возможности изменения параметров системы, приводящие к снижению T. Определим эффективную архитектуру ЦОД с помощью анализа соотношений между временем ответа, производительностью ЦОД, показателем использования ресурса и пропускной способностью ЦОД [1].

Обозначим через C пропускную способность ресурса в операциях в секунду. Пусть $1/\mu$ — среднее число операций, необходимых для выполнения задания. Обозначим через λ среднее число заданий, поступающих к ресурсу за единицу времени. Будем считать, что время ответа системы T равно времени между моментом поступления задания и моментом полного выполнения этого задания. Среднее время ожидания задания обозначим как W. Тогда $T = W + 1/\mu C$. Дисциплина обслуживания очереди — **FIFO. Показатель использования ресурса** $\rho = \lambda/\mu C$.

Модели предоставления услуг по размещению ресурсов в корпоративных центрах обработки данных

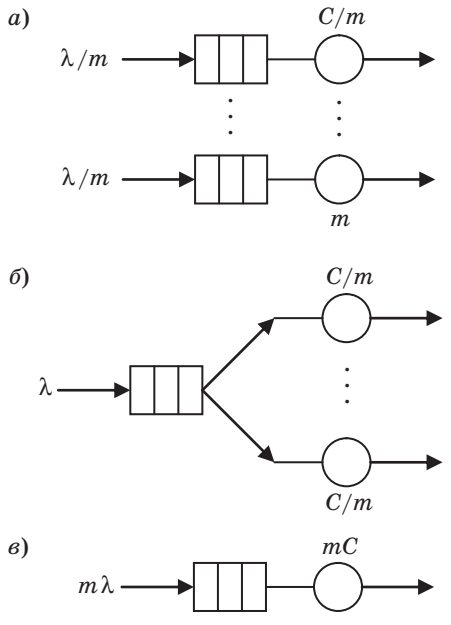
Оценим некоторые возможные структуры распределения ресурса и его коллективного использования [2, 3].

Рассмотрим первый вариант: совокупность m ресурсов, каждый из которых имеет пропускную способность C/m, что характерно для случая «новая задача — новый сервер». На каждый

[1] Научный руководитель – доцент кафедры информационных управляющих систем Санкт-Петербургского государственного университета телекоммуникаций им. проф. М. А. Бонч-Бруевича *Т. М. Татарникова*.

из этих ресурсов поступает поток заданий на выполнение работы с интенсивностью λ/m. Такая структура (рис. 1, *а*) соответствует набору m систем массового обслуживания (СМО) $G|G|1$ с суммарной пропускной способностью C. Эта система не эффективна, так как задания могут выстраиваться в очередь перед одним из ресурсов, в то время как какой-то другой ресурс будет простаивать. В связи с этим рассмотрим второй вариант, характерный для использования кластера компьютеров без виртуальных сред (рис. 1, *б*): одна очередь ко всему набору m ресурсов с суммарной интенсивностью λ, что моделируется СМО $G|G|m$.

Хотя система с единой очередью более эффективна, чем система с m разделенными средствами, все же остается некоторая нерациональность, когда очередь отсутствует, но не все ресурсы заняты. Чтобы преодолеть эту нерациональность, рассмотрим систему (рис. 1, *в*), где объединен как поток заданий, так и ресурсы, что моделируется СМО $G|G|1$ с интенсивностью потока на входе $m\lambda$ и пропускной способностью ресурса mC. Отличие системы с одним ресурсом состоит в том, что в ней интенсивность на входе и пропускная способность умножены на m при неизменном коэффициенте использования ρ. Такая система может быть реализована в виде кластера вычислительных устройств с развернутыми виртуальными средами, позволяющими максимально эффективно использовать имеющиеся ресурсы.



■ ***Рис. 1.*** *Модели СМО предоставления услуг по размещению ресурсов в корпоративных ЦОД: а — m систем G/G/1; б — система G/G/m; в — система G/G/1*

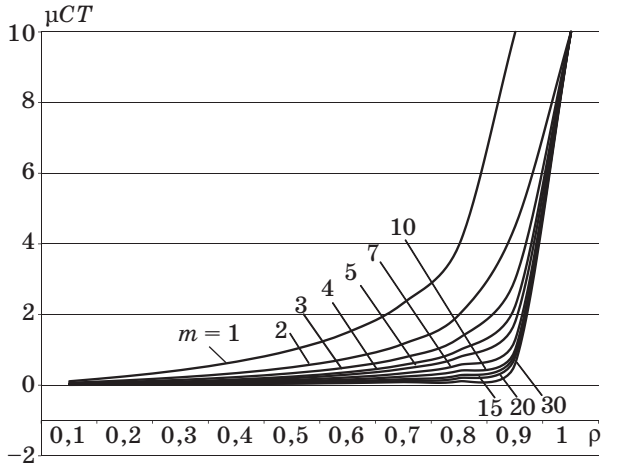
Для доказательства оптимальности данной архитектуры для ЦОД проведем анализ.

Среднее время ответа системы можно определить следующим образом:

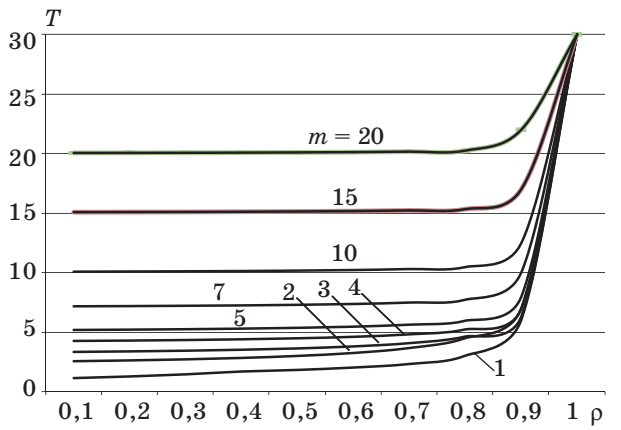
$$T = m\mu C + pm\mu C(1 - \rho). \qquad (*)$$

Графики зависимостей нормированного среднего времени ответа от коэффициента использования ресурса для разных значений m (рис. 2) показывают, что кривые проходят через точку 0 при $\rho = 0$, так как в этой точке p_m, $m = 1, 2, \ldots$ с ростом m при заданном значении ρ нормированная задержка быстро убывает, и при $m \to \infty$ кривые стремятся к прямой, описывающей детерминированную систему $D|D|1$, в которой очереди не образуются.

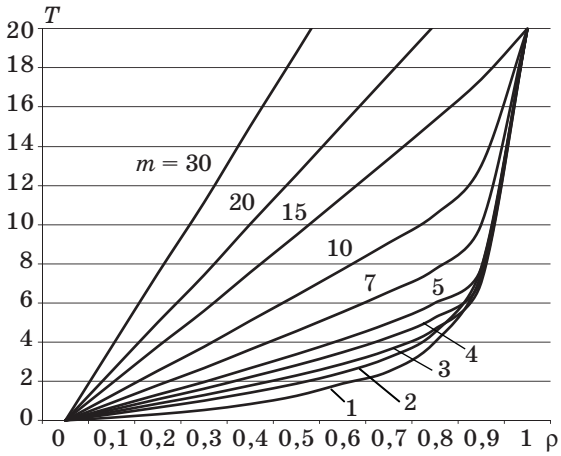
Для системы $M|M|m$ функция (*) представлена на рис. 3. Здесь предполагается, что суммарная пропускная способность удерживается постоянной ($C_0 = 1$) и делится поровну между m ресурса-



■ ***Рис. 2.*** *Зависимость нормированного среднего времени ответа от коэффициента использования ресурса для разных значений m*



■ ***Рис. 3.*** *Зависимость среднего времени ответа от ρ при $\mu = \mu_0 = 1$ и $C = C_0 = 1$*



■ ***Рис. 4.*** *Среднее время ответа при фиксированной интенсивности поступления заявок*

■ *Показатель эффективности системы с одним ресурсом*

ρ	$T = 1$	$T = 2$	$T = 3$	$T = 4$	$T = 5$
0,1	0,10	0,06	0,037	0,03	0,02
0,2	0,25	0,12	0,083	0,06	0,05
0,4	0,70	0,33	0,22	0,17	0,13
0,6	1,50	0,75	0,50	0,37	0,3
0,8	4,00	2,00	1,33	1,00	0,80
0,9	9,00	4,50	3,00	2,25	1,80

ми. Система показывает лучшие результаты по сравнению с первой системой (см. рис. 2), а именно время ответа возрастает с увеличением m при постоянном ρ. Это объясняется тем, что μC постоянно, и, меняя λ, можно менять ρ, т. е. нужно поддерживать постоянным λ, а не μC.

Для того чтобы допустить изменение ρ при изменении μC, положим $\lambda = \lambda_0 = 0{,}8$ (рис. 4). Имеет место увеличение времени ответа с ростом m при постоянном ρ, что также показывает преимущество одного ресурса по сравнению с множеством отдельных ресурсов, обладающих заданной суммарной пропускной способностью.

Эффективность системы можно продемонстрировать другим способом, если сосредоточить внимание на системе с одним ресурсом $M|M|1$ [3]. В таблице приведены значения эффективности системы — увеличение коэффициента использования с ростом масштабного коэффициента при постоянном среднем времени ответа. Представленная здесь функция является решением следующего уравнения относительно ρ:

$$\rho = \lambda T/(1 + \lambda T).$$

Заключение

Полученные результаты исследований говорят о том, что для больших систем $M|M|m$ можно получить выигрыш в среднем времени ответа, который пропорционален масштабному коэффициенту. При заданном масштабном коэффициенте система с единым ресурсом лучше, чем система с разделяемым ресурсом. В целом, улучшение среднего времени ответа системы можно получить при использовании большой системы коллективного использования с единым ресурсом.

Литература

1. **Кормильцев А. И.** Как построить оптимальную систему хранения данных // Сети и системы связи. 2002. № 11. С. 52–58.
2. **Клейнрок Л.** Вычислительные системы с очередями: Пер. с англ. — М.: Мир, 1979. — 600 с.
3. **Chakrvarthy S.** The batch markovian arrival process: a review and future work // Advances in probability theory and stochastic processes. 2001. P. 21–39.