Сведения об авторе

Людмила Александровна Муравьева-Витковская — канд. техн. наук; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники;
E-mail: muravyeva-vitkovskaya@yandex.ru

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 23.12.13 г.

УДК 004.89: 002.53

В. В. Соснин

ВРЕМЯ ОЖИДАНИЯ В НЕОДНОРОДНЫХ СИСТЕМАХ С ОЧЕРЕДЯМИ ПРИ ОБСЛУЖИВАНИИ ЗАЯВОК В ПОРЯДКЕ ПОСТУПЛЕНИЯ

Для системы массового обслуживания с заявками двух классов с помощью имитационного моделирования рассчитано среднее время ожидания в очереди при использовании бесприоритетной дисциплины обслуживания. Показаны условия, при которых различается среднее время ожидания заявок разных классов.

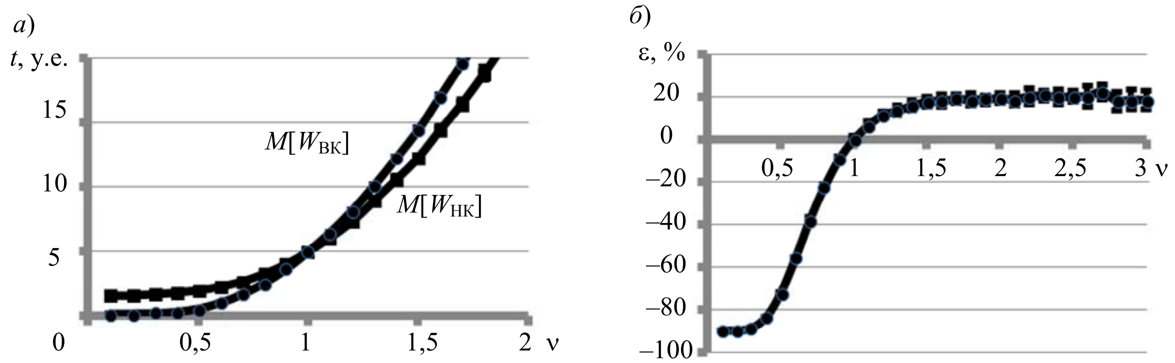
Ключевые слова: *неоднородная система массового обслуживания, среднее время ожидания в очереди, бесприоритетная дисциплина обслуживания.*

Введение. В теории массового обслуживания важное место занимает бесприоритетная дисциплина обслуживания (ДОБП). Традиционно считается [1], что при ДОБП качество обслуживания заявок разных классов одинаково (показателем качества считается среднее время ожидания заявки в очереди). В работе [2] показано, что в системе M/G/1 ДОБП ни один из классов не имеет преимуществ в качестве обслуживания, т.е. если значения времени ожидания в очереди заявок k классов суть случайные $W_1, W_2, ..., W_k$, то их математические ожидания равны: $M[W_1] = M[W_2] = ... = M[W_k]$. Однако цель настоящей работы — проверить, обладает ли этим свойством весь класс систем GI/GI/1 с ДОБП [3]. Для подтверждения корректности полученных автором результатов были проведены дополнительные исследования.

Поставленная задача решалась с помощью имитационного моделирования. Рассмотрим пример исследования системы массового обслуживания (СМО) GI/GI/1 с заявками двух классов (НК — низконагружающий, ВК — высоконагружающий класс), которые создают загрузки $\rho_{\text{ВК}} = 0,3$ и $\rho_{\text{НК}} = 0,03$. Время обслуживания заявок ВК и НК — случайная величина $B_{\text{ВК}}$ и $B_{\text{НК}}$ такая, что $M[B_{\text{ВК}}] = M[B_{\text{НК}}] = 10$ у.е. Время между приходом заявок ВК и НК — случайная величина $A_{\text{ВК}}$ и $A_{\text{НК}}$. Для моделирования $A_{\text{ВК}}$, $A_{\text{НК}}$, $B_{\text{ВК}}$ и $B_{\text{НК}}$ используется гамма-распределение, каждая из этих величин имеет фиксированное значение математического ожидания, а коэффициент вариации ν изменяется от 0 до 3 с шагом 0,1 так, что $\nu = \nu[A_{\text{ВК}}] = \nu[A_{\text{НК}}] = \nu[B_{\text{ВК}}] = \nu[B_{\text{НК}}]$. При проведении имитационных экспериментов измеряются время ожидания в очереди заявок каждого из классов ($W_{\text{ВК}}$ и $W_{\text{НК}}$), а также относительное различие их средних значений:

$$\varepsilon = \frac{M[W_{\text{ВК}}] - M[W_{\text{НК}}]}{M[W_{\text{НК}}]} \cdot 100\,\%.$$

Результаты имитационного моделирования этой системы приведены на рисунке. Высота значков на кривых соответствует величине 99 %-ного доверительного интервала измеренных в соответствующей точке величин (по Стьюденту). На рисунке *а* видно, что средние задержки НК и ВК заметно различаются за пределами доверительного интервала. Относительное значение этого различия представлено на рисунке *б*: видно, что среднее время ожидания заявок ВК может быть на 90 % меньше или на 20 % больше среднего времени ожидания заявок НК.



Аналогичный характер имеет соотношение вариации времени ожидания, отличие состоит в том, что относительная разница их значений несколько меньше. Приведенный пример позволяет однозначно утверждать, что „свойство бесприоритетности“ ДОБП в СМО M/G/1 не может быть распространено на весь класс систем GI/G/1. Анализ экспериментальных данных показал, что заявки НК получают тем лучшее качество обслуживания, чем ближе параметры системы к следующим предельным значениям: $\nu[A_{НК}] \to 0$, $\nu[A_{ВК}] \to \infty$, $\nu[B_{ВК}] \to 0$, $\nu[B_{НК}] \to 0$, $\dfrac{\rho_{ВК}}{\rho_{НК}} \to \infty$. Указанное свойство проверялось для $\nu[A_{ВК}] \le 5$. При $\nu[A_{ВК}] > 5$ для получения результатов с приемлемым доверительным интервалом требуются существенные вычислительные мощности. Примем, что полученное свойство верно для всей области значений $0 < \nu[A_{ВК}] < \infty$, тогда с помощью аналитических преобразований (см. подробный вывод в [3]) можно сформулировать следующее неравенство:

$$M[W_{НК}] > \rho M[W_{ВК}]. \quad (1)$$

Этот результат подтверждается имитационными экспериментами с различными законами распределения (двухфазное Кокса, гамма, равномерное). Приведем пример возможного применения полученного результата. Рассмотрим процессы, протекающие на выходном порте некоторого маршрутизатора в компьютерной сети. В качестве потока заявок ВК рассматривается сетевой трафик, который проходит через порт. Современные маршрутизаторы позволяют в режиме реального времени получить данные о текущих характеристиках передачи пакетов, поэтому будем считать информацию о средних задержках пакетов потока ВК доступной и достоверной. В качестве потока заявок НК рассматривается сетевой трафик низкой интенсивности, который планируется дополнительно пустить через рассматриваемый порт. Тогда неравенство (1) позволяет оценить максимальное преимущество в качестве обслуживания, которое могут получить заявки НК. Например, пусть заявки ВК создают 60 %-ную нагрузку, а их среднее время ожидания в очереди равно 200 мс. Пусть требуется пустить в эту СМО еще один поток заявок, увеличивающий ее нагрузку на 3 %. Тогда формула (1) позволяет утверждать, что время ожидания в очереди заявок нового потока не может быть менее 123 мс.

Таким образом, в настоящей работе получены следующие результаты:

1) с помощью имитационных экспериментов показано, что при бесприоритетной дисциплине обслуживания заявки разных типов могут иметь разные средние значения времени ожидания в очереди, а также разные вариации этого времени;

2) получена численная оценка нижней границы для показателей ожидания низконагружающих потоков заявок.

СПИСОК ЛИТЕРАТУРЫ

1. *Bolch G., Greiner S., Meer H., Triverdi K.* Queueing networks and Markov chains: modeling and performance evaluation with computer science applications. NY: John Wiley & Sons, 1998.

2. *Клейнрок Л.* Вычислительные системы с очередями. М.: Мир, 1979. 600 с.

3. *Соснин В. В.* Свойства бесприоритетной дисциплины обслуживания в системах вида GI/G/1 // Тр. 5-й Всерос. науч.-практ. конф. „Имитационное моделирование. Теория и практика“ (ИММОД 2011). СПб, 2011. Т. 2. С. 355—360.

Сведения об авторе

Владимир Валерьевич Соснин — канд. техн. наук; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; E-mail: vsosnin@mail.ru

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 23.12.13 г.

УДК 681.3

В. А. Богатырев, А. В. Богатырев, С. В. Богатырев

ОПТИМИЗАЦИЯ ПЕРЕРАСПРЕДЕЛЕНИЯ НАГРУЗКИ В КЛАСТЕРАХ ПРИ ИЗМЕНЯЮЩЕЙСЯ АКТИВНОСТИ ИСТОЧНИКОВ ЗАПРОСОВ

Предложено решение задачи динамической оптимизации перераспределения запросов через сеть в общедоступный кластер, выполняемой с целью минимизации среднего времени пребывания запросов и позволяющей учитывать задержки отображения числа активных узлов, формирующих запросы.

Ключевые слова: *отказоустойчивость, распределение запросов, кластер, оптимизация, адаптация.*

Введение. Распределенные вычислительные системы должны характеризоваться минимальными задержками обслуживания при максимальных надежности, отказоустойчивости и производительности системы [1—3]. Эффективность распределенных компьютерных систем достигается при их адаптации к отказам, к изменениям параметров потоков запросов и состояний очередей узлов [3—5], в том числе в результате динамического перераспределения запросов между узлами компьютерной системы [4—9].

Адаптивное перераспределение запросов в реальном времени приводит, с одной стороны, к балансировке загрузки и росту эффективности системы (уменьшению среднего времени пребывания запросов), а с другой — к ее падению из-за дополнительных задержек, связанных с отображением состояний узлов. Это и определяет необходимость оптимизации процесса перераспределения запросов в вычислительных системах, включая объединение через распределенную инфраструктуру множества ресурсов, доступных из любой точки системы.