ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ, СВЯЗАННЫЕ С АНАЛИЗОМ ПРОИЗВОДИТЕЛЬНОСТИ СИММЕТРИЧНЫХ БОЛЬШИХ ЗАМКНУТЫХ КОМПЬЮТЕРНЫХ СЕТЕЙ ОБСЛУЖИВАНИЯ

Вячеслав Абрамов

●

School of Mathematical Sciences, Monash University,
Clayton Campus, Building 28, Level 4, Wellington road, Victoria 3800, Australia.
e-mail: vyacheslav.abramov@sci.monash.edu.au

Аннотация

В работе изучаются замкнутые сети обслуживания, состоящие из станции-сервера и k идентичных станций-клиентов. Станции-клиенты являются ненадежными. Время жизни станции-клиента не зависит от времен жизни других станций-клиентов и имеет одно и то же распределение $G(x)$ для всех таких станций. Станция-сервер – это бесконечно-канальная система массового обслуживания. Станции-клиенты являются одноканальными системами массового обслуживания с автономным обслуживанием, т. е. каждая станция-клиент принимает на обслуживание требования в некоторые фиксированные моменты времени, порожденные стационарной (в узком смысле) и эргодичной последовательностью случайных величин. Общее число требований в системе равно N. Длительности обслуживания на станции-сервере являются независимыми экспоненциально распределенными случайными величинами с параметром λ. Средние времена обслуживание на каждой станции-клиенте $(N\mu)^{-1}$. После завершения обслуживания на станции-сервере требование направляется равновероятно на одну из доступных (т. е. работоспособных) станций-клиентов, и будучи там обслужена, оно возвращается на станцию-сервер. Параметр N предполагается большим. Цель статьи – изучение перегруженных станций-клиентов и нахождение доверительных интервалов, связанных с достижением величиной очереди некоторого высокого уровня, пропорционального величине N.

1. Введение

Рассматривается большая замкнутая сеть массового обслуживания состоящая из станции-сервера (бесконечно-канальной системы массового обслуживания) и k идентичных станций-клиентов (одноканальных систем обслуживания). Общее число требований в сети равно N. Величина N предполагается большой. Процесс обслуживания требований на станциях-клиентах предполагается автономным. Системы с автономным обслуживанием были введены Боровковым [6, 7]. Формальное определение этих систем в простейшем случае одиночных поступлений и обслуживаний требований является следующим. Пусть $A(t)$ - точечный процесс поступления требований, $S(t)$ - точечный процесс обслуживания. Обозначим $Q(t)$ число требований в системе в момент t и предположим, что все эти процессы начинаются в нуле $(A(0) = S(0) = Q(0) = 0)$. Тогда, автономное обслуживание определяется уравнением:

$$Q(t) = A(t) - \int_0^{\infty} \mathrm{I}\{Q(s-) > 0\} dS(t).$$

Системы с автономным обслуживанием изучались во многих работах (см. [1, 3, 4, 8, 9, 10]) . В настоящей работе изучается замкнутая сеть типа client/server (см. Рисунок 1)

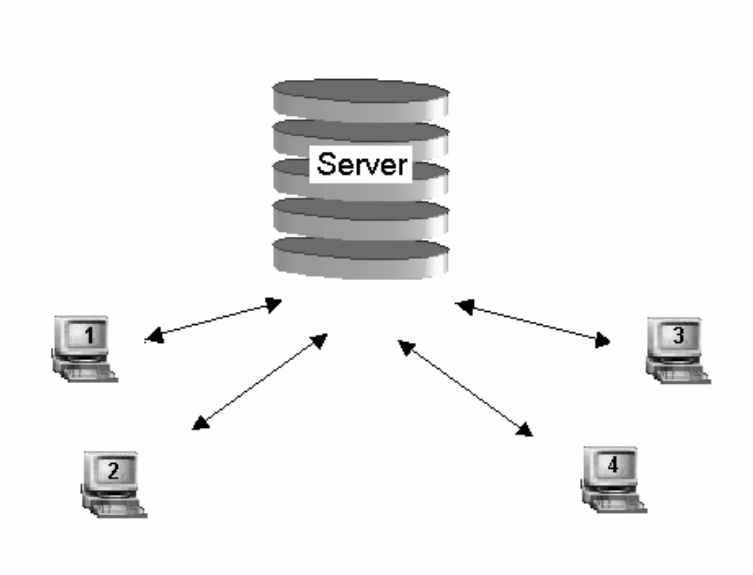

Рисунок 1. Пример сети типа client/server

Процесс поступления требований со станции-сервера на i-ю станцию-клиент обозначается $A_{i,N}(t)$. Время обслуживания каждого требования на станции-сервере распределено экспоненциально с параметром λ. Поэтому, скорость поступления на станции-клиенты зависит от числа требований, имеющихся на станции-сервере. Если в момент времени t на станции-сервере имеется N_t требований, то скорость выхода требований со станции-сервера равна λN_t. Всего имеется k станций-клиентов, и каждая из них является ненадежной. Продолжительность работы до выхода из строя для каждой станции-клиента не зависит от продолжительностей работы других таких станции и имеет функцию распределения $G(x)$.

Все станции-клиенты предполагаются идентичными, и требования, отправляемые со станции-сервера, “выбирают" каждую станцию-клиент с равной вероятностью. (По этой причине сеть называется симметричной.) Поэтому, если имеется l работоспособных станций-клиентов в момент t, то скорость поступления к каждой из этих станций равна $\lambda N_t / l$.

Возможные моменты выхода требований из j-й станции-клиента $(j = 1,2,...,k)$ $\xi_{j,N,1}, \xi_{j,N,1} + \xi_{j,N,2}, \xi_{j,N,1} + \xi_{j,N,2} + \xi_{j,N,3}, \ldots$, где каждая из последовательностей $\xi_{j,N,1}, \xi_{j,N,2}, \ldots$ является стационарной (в узком смысле) и эргодичной последовательностью случайных величин (N является параметром серии). Соответствующий точечный процесс, связанный с выходом требований с j-й станции-клиента , обозначается

$$S_{j,N}(t) = \sum_{i=1}^{\infty} \mathrm{I}\left(\sum_{l=1}^{i} \xi_{j,N,l} \le t \right),$$

и удовлетворяет условию

$$\mathrm{P}\left(\lim_{t\to\infty} \frac{S_{j,N}(t)}{t} = \mu N \right) = 1.$$

Тогда соотношения между параметрами λ, μ и k предполагаются следующими:

(1.1) $$\frac{\lambda}{k\mu} < 1,$$

и

(1.2) $$\frac{\lambda}{\mu} > 1.$$

Условие (1.1) означает, что в начальный момент времени, когда все станции- клиенты работоспособны, скорость обслуживания на каждой стации-клиенте выше скорости поступления требований на эти станции. В свою очередь условие (1.2) означает, что после выхода из строя нескольких станций клиентов, скорость обслуживания на стациях-клиентах становятся ниже скоростей поступления требований на эти станции. Обозначим:

$$l_0 = \max\left(l : \frac{\lambda}{l\mu} > 1 \right).$$

наибольшее число работоспособных станций, когда скорость обслуживания на стациях-клиентах ниже скоростей поступления требований. Тогда для всех $l \le l_0$ это также будет иметь место.

Длина очереди в момент времени t на j-й станции-клиенте определяется из уравнения

$$Q_{j,N}(t) = A_{j,N}(t) - \int_0^t \mathrm{I}\{Q_{j,N}(s-) > 0\} dS_{j,N}(s),$$

где $A_{j,N}(t)$ - это процесс поступления на j-ю станцию-клиент.

Пусть $\alpha < 1$ некоторое положительное число. Будем говорить, что сеть обслуживания под угрозой риска, если общее число требований в станциях-клиентах превышает величину αN. Предположим, что в начальный момент времени все требования находятся на станции-сурвере. Цель статьи – найти доверительный интервал $[0,\theta)$ такой, что с данной вероятностью P (например, $P = 0.95$) сеть не будет в состоянии риска в течении этого интервала. Для сетей с произвольным числом станций-клиентов такая задача является сложной, потому, что поведение очередей в перегруженной сети является многовариантным. (См. следующий раздел, где приведены явные результаты.) Поэтому, в настоящей статье мы изучаем эту задачу только для двух станций- клиентов.

Большая замкнутая перегруженная сеть массового обслуживания изучалась во многих работах. Марковская перегруженная сеть анализировалась в [11]. В [1, 2, 3] результаты [11]

были обобщены на немарковские сети. В частности, в [1] изучался вариант сети с автономным механизмом обслуживания на станциях-клиентах. Результаты [1] затем были распространены на сети с двумя типами станций и многими классами требований в [4]. Однако, вклад всех упомянутых статей является чисто теоретическим. Во всех этих статьях анализ перегруженных сетей проводится в частном случае, когда имеется только одна перегруженная станция-клиент и когда в начальный момент времени все требования находятся на станции-сервере.

Детальный анализ перегруженных сетей, включающий все возможные случаи перегруженных станций-клиентов, а также произвольные начальные условия, касающиеся числа требований на станциях-клиентах в начальный момент времени, сделан в [5]. Результаты [5] являются многообещающими для решения многих прикладных задач. В частности, анализ [5] посвящен замкнутым сетям обслуживания в полумарковской среде, требующий изучения этих сегей при наиболее общих предположениях. Асимптотическое решение проблемы настоящей статьи, когда N неограниченно возрастает, основано на результатах статьи [5].

Следует отметить, что надежность самих компьютерных сетей изучалась во многих статьях. Мы цитируем книгу [12], где читатель может найти детальную информацию, относящуюся к данному вопросу. Доверительные интервалы, изучаемые в настоящей статье имеют отношение к надежности информации, которая зависит от надежности элементов сети.

Статья мотивирована важными практическими задачами в телекоммуникационных сетях. Поддержка и обмен информации являются весьма дорогими и часто превышают стоимость самого оборудования. С другой стороны, надежная поддержка информации является производной от высоконадежного оборудования и непосредственно зависит от этой надежности. Особый цикл практических проблем относится к поддержке больших баз данных. Тогда “требование" ассоциируется с единицей информации (записью) в базе данных. “Станция-клиент" ассоциируется с пользователем базы данных. “Неисправная станция-клиент" ассоциируется с отсутствием связи или очень низкой скоростью обмена. Низкая скорость обмена на некоторых станциях становится причиной перегруженности сети и может привести к разрушению базы данных.

Статья организована следующим образом. В разделе 2 напоминаются некоторые результаты [5], необходимые для цели настоящей работы, которые затем адаптированы для случая симметричной сети, рассмотренной здесь. В разделе 3 мы выводим соотношения для предельных распределений нормализованных длин очередей на исправных станциях-клиентах. В разделе 4 мы устанавливаем результаты для доверительных интервалов в частном случае двух станций клиентов. В разделе 5 приводится простой численный пример. Раздел 6 является заключением.

2. Перегруженные станции-клиенты

В этом разделе мы напоминаем некоторые результаты анализа перегруженных станций-клиентов [5], соответствующие случаям изучаемым в настоящей статье. Мы начинаем с элементарного случая l эквивалентных перегруженных станций-клиентов, т. е. со случая, когда в начальный момент времени $t = 0$ имеется точно l перегруженных станций-клиентов. Для простоты предположим, что все эти l станций являются абсолютно надежными, и в начальный момент времени $t = 0$ имеется $(1-\beta)N$ требований на станции- клиенте, $0 < \beta \le 1$, и остальные

βN требований распределены между l станциями-клиентами. Так как сеть является симметричной, то предположение, что имеется примерно $\beta N / l$ требований на каждой станции-клиенте в момент $t = 0$, в соответствии с законом больших чисел является резонным. Предположение, что станции-клиенты являются перегруженными, означает, что $\lambda(1-\beta) > l\mu$.

Результат об асимптотическом поведении нормированных длин очередей следует из Предложения 5.3 [5], относящихся к общему случаю несимметрической перегруженной сети с произвольными начальными длинами очередей. Напомним этот результат.

Lemma 2.1. *Предположим, что все станции-клиенты являются перегруженными в начальный момент времени, и начальные длины очередей асимптотически равны* $N\beta_1, N\beta_2, ..., N\beta_k$ *соответственно* $(\beta_1 + \beta_2 + ... + \beta_k \le 1)$, *когда* $N \to \infty$. *Тогда нормированная длина очереди на* j*-й станции-клиенте в пределе, когда* $N \to \infty$, *определяется как*

$$(2.1) \qquad q_j(t) = \beta_j + \left(1 - \sum_{j=1}^{k} \beta_j\right)\left([\lambda_j(0) - \mu_j]t - \lambda_j(0)\int_0^t r(s)ds\right),$$

$$(2.2) \qquad r(t) = \left(\frac{\sum_{j=1}^{k}(\lambda_j(0) - \mu_j)}{\sum_{j=1}^{k}\lambda_j(0)}\right)\left(1 - \exp\left[-t\sum_{j=1}^{k}\lambda_j(0)\right]\right),$$

где $q_j(t)$ *обозначает процесс нормированной длины очереди на* j*-й станции-клиенте в пределе, т. е.* $q_j(t)$ *- это предел по вероятности* $Q_{j,N}(t)/N$ *при* N *стремящемся к бесконечности.*

В обозначениях этой леммы $\lambda_j(0)N$ - это мгновенная скорость поступления требований на j-ю станцию-клиент в момент времени $t = 0$, и $\mu_j N$ - это скорость обслуживания на j-й станции-клиенте. В нашем случае число станций-клиентов равно l, мгновенная скорость поступления требований на каждую станцию-клиент равно $(1-\beta)\lambda N / l$ и скорость обслуживания равна μN и все $q_j(t)$ равны, т. е. $q_j(t) \equiv g(t)$ для всех $j = 1,2,...,l$. Поэтому, в нашем случае из этой Леммы 2.1 мы имеем следующее утверждение.

Предложение 3.2. *Имеем:*

$$(2.3) \qquad g(t) = \frac{\beta}{l} + (1-\beta)\left(\left[\frac{(1-\beta)\lambda}{l} - \mu\right]t - \frac{(1-\beta)\lambda}{l}\int_0^t r(s)ds\right),$$

где

$$(2.4) \qquad r(t) = \left(1 - \frac{\mu l}{(1-\beta)\lambda}\right)\left(1 - e^{-(1-\beta)\lambda t}\right).$$

3. Предельная нормированная накопленная длина очереди

В этом разделе мы изучаем предельную (при $N \to \infty$) нормированную накопленную длину очереди на станциях-клиентах. Она обозначена $q(t)$. В начальный момент времени $t = 0$ имеется k работоспособных станций-клиентов. Пусть $\tau_1, \tau_2, ..., \tau_k$ -это моменты отказов, $0 \le \tau_1 \le \tau_2 \le ... \le \tau_k$. Эти моменты отказов связаны с поведением сети обслуживания с ингенсивностями, зависящими от времени, которая может рассматриваться как сеть в полумарковской среде. Поэтому, мы можем применить Теорему 5.4 [5].

Случайный интервал времени [0, τ_k] является временем жизни всей сети. Поэтому процесс $q(t)$ должен рассматриваться в течение этого интервала времени [0, τ_k]. Напомним, что $l_0 = \max(l : \lambda/(l\mu) > 1)$. Поэтому, согласно Теореме 5.4 [5] на интервале [0, τ_{k-l_0}) мы имеем $q(t) = 0$. Далее, на случайном интервале $[\tau_{k-l_0}, \tau_{k-l_0+1})$ уравнение для $q(t)$ является следующим

$$(3.1) \qquad q(t) = (\lambda - l_0\mu)(t - \tau_{k-l_0}) - \lambda \int_{\tau_{k-l_0}}^{t} r(s - \tau_{k-l_0})ds,$$

где $r(t)$ задано соотношением (2.4). Уравнение (3.1) следует из (2.1) и (2.2) следующим образом. Полагая $\beta = 0$, $l = l_0$ и заменяя t разностью $(t - \tau_{k-l_0})$ из (2.1) получаем:

$$g(t) = \left(\frac{\lambda}{l_0} - \mu\right)(t - \tau_{k-l_0}) - \frac{\lambda}{l_0}\int_{\tau_{k-l_0}}^{t} r(s - \tau_{k-l_0})ds.$$

Следовательно, принимая во внимание, что $q(t) = l_0 g(t)$ приходим к (3.1).

Для следующего интервала $[\tau_{k-l_0+1}, \tau_{k-l_0+2})$, $l_0 > 1$, мы имеем уравнение:

$$(3.2) \qquad g(t) = \frac{q(\tau_{k-l_0+1})}{l_0 - 1} + [1 - q(\tau_{k-l_0+1})]\left\{\left[\frac{[1 - q(\tau_{k-l_0+1})]\lambda}{l_0 - 1} - \mu\right](t - \tau_{k-l_0+1}) - \frac{[1 - q(\tau_{k-l_0+1})]\lambda}{l_0 - 1}\int_{\tau_{k-l_0+1}}^{t} r(s - \tau_{k-l_0+1})ds\right\}.$$

Поэтому, на этом интервале $[\tau_{k-l_0+1}, \tau_{k-l_0+2})$

$$(3.3) \qquad q(t) = q(\tau_{k-l_0+1}) + [1 - q(\tau_{k-l_0+1})]\{([1 - q(\tau_{k-l_0+1})]\lambda - \mu(l_0 - 1))(t - \tau_{k-l_0+1}) - [1 - q(\tau_{k-l_0+1})]\lambda \int_{\tau_{k-l_0+1}}^{t} r(s - \tau_{k-l_0+1})ds\}.$$

Для произвольного интервала $[\tau_i, \tau_{i+1}),\ i = k - l_0, k - l_0 + 1, ..., k - 1,$ имеем:

$$q(t) = q(\tau_i) + [1 - q(\tau_i)]\{([1 - q(\tau_i)]\lambda - \mu(k - i))(t - \tau_i) - [1 - q(\tau_i)]\lambda \int_{\tau_i}^{t} r(s - \tau_i)ds\}.$$

В крайней точке τ_k мы полагаем $q(t) = 1.$

4. Доверительные интервалы

Соотношение для предельной нормированной накопленной величины очереди является очень сложным. Поэтому, в этом разделе мы получаем доверительные интервалы в частном случае, когда имеются только две станции-клиенты. В этом случае только используется соотношение (3.1), которое в случае двух станций-клиентов выглядит следующим образом:

$$q(t) = (\lambda - \mu)(t - \tau_1) - \lambda \int_{\tau_1}^{t} r(s - \tau_1)ds, \tag{4.1}$$

где

$$r(s) = \left(1 - \frac{\mu}{\lambda}\right)\left(1 - e^{-\lambda t}\right). \tag{4.2}$$

Доверительный интервал состоит из двух подынтервалов. Первый подынтервал - это $[0,\ \tau_1)$, где предельная нормированная накопленная очередь равна нулю. Второй подынтервал - это $[\tau_1, \theta]$, где θ - это некоторая точка, где $q(\theta) \le \alpha.$ Уравнения (4.1) и\(4.2) определены для $t < \tau_2,$ где τ_2 - это случайная точка отказа второй станции-клиента.

Пусть θ^* - это точка, где $q(\theta^*) = \alpha.$ Точка θ^* - это случайная точка, зависящая от $\tau_1.$ Однако, при предположении, что та или иная станция-клиент является активной, длина интервала $[\tau_1, \theta^*]$ фиксированна и однозначно определена из (4.1) и (4.2).

Найдем распределение процесса $q(t).$ Ясно, что вероятность того, что $q(t) = 0$ совпадает с вероятностью того, что длина интервала $[0, \tau_1)$ более $t.$ Поэтому,

$$\mathrm{P}\big(q(t) = 0\big) = [1 - G(t)]^2. \tag{4.3}$$

Далее,

$$\mathrm{P}\big(q(t) \le \gamma < 1\big) = [1 - G(t)][1 - G(t - t_\gamma)], \tag{4.4}$$

где t_γ - это такое значение t при котором

(4.5) $$(\lambda-\mu)t-\lambda\int_0^t r(s)ds=\gamma.$$

Уравнения (4.3) и (4.4) легко следуют из элементарных фактов теории вероятностей.

Тогда, вероятность того, что предельная нормированная накопленная длина очереди достигнет уровня γ до того как попадет в 1 равна

$$\frac{\int_0^\infty [1-G(t)][1-G(t-t_\gamma)]dt}{\int_0^\infty [1-G(t-t_\gamma)]^2 dt}.$$

Задача состоит в том, чтобы найти такое $\gamma \le \alpha$, что

(4.6) $$\frac{\int_0^\infty [1-G(t)][1-G(t-t_\gamma)]dt}{\int_0^\infty [1-G(t-t_\gamma)]^2 dt} \ge P.$$

(4.6) – это неравенство потому, что точное равенство может иметь место при $\gamma > \alpha$, в то время как для $\gamma \le \alpha$ это должно быть неравенство (4.6).

5. Численный пример

Рассмотрим следующий пример. Пусть $\lambda=4,\ \mu=3,\ \alpha=0.2,\ P=0.95,\ G(x)=1-e^{-2x}$.

Из соотношения (4.6) имеем:

$$\frac{\int_0^\infty e^{-2(t-t_\gamma)}e^{-2t}dt}{\int_0^\infty e^{-4(t-t_\gamma)}dt}=e^{-2t_\gamma}.$$

Решение уравнения $e^{-2t_\gamma}=0.95$ приводит нас к $t_\gamma=0.025647$. Из (4.5) получаем:

$$\gamma=\int_0^{t_\gamma} e^{-4t}dt=\int_0^{0.025647} e^{-4t}dt=0.25-0.25e^{-0.102588}=0.024375.$$

Полученная величина γ меньше, чем $\alpha=0.2$, и следовательно значение $\gamma=0.024375$ - это требуемое значение для доверительного интервала.

6. Заключительные замечания

Мы нашли доверительные интервалы, связанные с достижением некоторого высокого уровня. Доверительные интервалы, найденные в статье, являются случайными. Они получены в терминах параметра γ, который является предельной величиной нормированной накопленной величины очереди, при которой вероятность того, что сеть массового обслуживания будет функционировать, является не меньшей, чем заданное значение P. Таким образом, стратегия состоит в наблюдении за процессом очереди до тех пор пока общее число требований в очередях не достигнет величины γN. Как только эта величина будет превышена, нет достаточных гарантий, что система и/или информация будут доступны.

Благодарность

Автор благодарит Australian Research Council (грант № DP0771338) за поддержку этого исследования.

Литература

1. **V. M. Abramov,** 2000. A large closed queueing network with autonomous service and bottleneck. *Queueing Systems*, 35, 23-54.
2. **V. M. Abramov,** 2001. Some results for large closed queueing networks with and without bottleneck: Up- and down-crossings approach. *Queueing Systems*, 38, 149-184.
3. **V. M. Abramov,** 2004. A large closed queueing network containing two types of node and multiple customers classes: One bottleneck station. *Queueing Systems*, 48, 45-73.
4. **V. M. Abramov,** 2006. The effective bandwidth problem revisited. arXiv: math. 0604182.
5. **V. M. Abramov,** 2007. Large closed queueing networks in semi-Markov environment and their application. arXiv: math. 0612118.
6. **A. A. Borovkov,** 1976. *Stochastic Processes in Queueing Theory.* Springer-Verlag, Berlin.
7. **A. A. Borovkov,** 1984. *Asymptotic Methods in Queueing Theory.* John Wiley, New York.
8. **C. Fricker,** 1986. Etude d'une file GI/G/1 á service autonomé (avec vacances du serveur). *Advances in Applied Probability,* 18, 283-286.
9. **C. Fricker,** 1987. Note sur un modele de file GI/G/1 á service autonomé (avec vacances du serveur). *Advances in Applied Probability,* 19, 289-291.
10. **E. Gelenbe and R. Iasnogorodski,** 1980. A queue with server of walking type (autonomous service). *Ann. Inst. H. Poincare*, 16, 63-73.
11. **Ya. Kogan and R. Sh. Liptser,** 1993. Limit non-stationary behaviour of large closed queueing networks with bottlenecks. *Queueing Systems*, 14, 33-55.
12. **M. Xie, Y. S. Dai and K. L. Poh,** 2004. *Computer Systems Reliability: Models and Analysis.* Kluwer, Dordrecht.