Информационные процессы, Том 5, № 5, 2005, стр. 432–442

Система массового обслуживания с динамической маршрутизацией и распределением Вейбулла времени обслуживания

А.В. Аленичев

Московский Физико-Технический Институт, Долгопрудный, Россия

Поступило в редколлегию 17.11.2005

Аннотация—В статье рассматривается процесс динамической маршрутизации заявок, поступающих в систему из N серверов в соответствии с распределением Пуассона с интенсивностью $N\lambda$ и имеющих Вейбулловское распределение времен обслуживания τ. Маршрутизация производится следующим образом. При поступлении в систему заявки в момент времени t, случайно выбираются K серверов из N, и заявка становится в очередь сервера с минимальной длинной. Изучается поведение вероятности переполнения системы в асимптотике большого входного буфера, в смысле вероятности попадания виртуальной заявки в очередь с временем обслуживания находящихся в ней заявок большим уровня z_0, ($z_0 \to \infty$). Рассматривается модель дискретного времени.

1. ВВЕДЕНИЕ

Появление новых высокоскоростных каналов передачи данных и все большее распространение компьютерных сетей, таких как Интернет, повлекло за собой интенсивные исследования стохастической природы сетевого трафика, которые показали, что традиционные предположения о нем уже не верны. Наряду с пакетным характером трафика, исследования современных сетей(см. [5],[6]) показали существование сеансовых зависимостей, которые приводят к значительной корреляции параметров трафика в течении длительного промежутка времени(времени сеанса). Производительность сетей и способность обеспечивать требуемое качество обслуживания может быть оценена посредством изучения вероятности потери и времени распространения пакета. Исследования сетей обеспечивающих высокое качество обслуживания приводят к тому, что требуется оценивать вероятность редких событий, на которую сильно влияют сеансовые зависимости трафика.

Есть множество объяснений существования сеансовых зависимостей трафика. Например, причиной их появления могут быть как длительные Web соединения в Интернете, так и копирование файлов с удаленного сервера. Наиболее подходящей моделью для описания сеансовых зависимостей трафика является ВКЛ-ВЫКЛ источники с длиной периода активности, имеющей степенное распределение. Многочисленные исследования таких систем показали, что для того чтобы обеспечить такое же качество обслуживания, как при классической модели трафика, необходимо значительно увеличить размер входного буфера. Это вызвано тем, что при классической модели трафика хвост стационарного распределения нагрузки буфера убывает экспоненциально, а сеансовые зависимости приводят к более медленному степенному убыванию.

Как с практической, так и с теоретической точки зрения интересно рассмотреть класс источников, на которых происходит переход от классического поведения нагрузки буфера к

степенному. Для однородного входного потока, как показано в [18], переход происходит на источниках с Вейбулловским распределением длины сеанса τ при $\alpha \in (0,1)$:

$$\mathbb{P}(\tau > x) \sim e^{-\gamma x^{\alpha}}.$$

Системы с субэкспоненциальным и степенным законом распределения длины сеанса хорошо изучены множеством авторов при различных гипотезах о структуре входного потока, распределениях скоростей сеансов и других параметрах системы(см. [7], [16], [14],[10],[17]).
Системы же с Вейбулловским распределением длин сеансов менее изучены. Это вызвано тем, что для оценки вероятности большого уклонения приходится использовать более тонкие методы. Для однородных источников результаты, в виде нижней и верхней границы были получены рядом авторов (см. [7], [16], [14]). Случай неоднородных источников рассмотрен в [19].

В данной статье рассматривается система с Вейбулловским распределением времени обслуживания заявок, что отражает существование в сети пакетов с существенно неоднородным временем обслуживания. Модель входного трафика, а также методы подсчета вероятности переполнения системы, используемые в данной статье, выбраны исходя из вышеизложенных соображений и базируются на модели предложенной в [17].
Расширение сетей передачи данных делает актуальной проблему выбора маршрута прохождения данных, т.е. маршрутизации. Протоколы маршрутизации могут быть построены на основе разных алгоритмов, отличающихся способами построения таблиц маршрутизации, способами выбора наилучшего маршрута и другими особенностями своей работы. Одним из видов маршрутизации является динамическая маршрутизация, т.е. маршрутизация, которая зависит от текущего состояния системы и относится к семейству протоколов, балансирующих нагрузку в сети (load-balance protocol). В статье рассматривается следующий вид динамической маршрутизации. При поступлении в систему заявки в момент времени t, случайно выбираются K серверов из N, и заявка становится в очередь сервера с минимальной длинной. Исследования систем с похожим типом маршрутизации могут быть найдены в [1],[2],[3].

Работа организована следующим образом: в части 2 мы опишем модель и сформулируем предварительные результаты, в части 3 получим формулу для вероятности переполнения системы, а в заключении обсудим основные результаты.

2. МОДЕЛЬ И ПРЕДВАРИТЕЛЬНЫЕ РЕЗУЛЬТАТЫ

В статье рассматривается модель системы обслуживания дискретного времени. Моменты возникновения запросов являются пуассоновским процессом с интенсивностью $N\lambda$. Обозначим через θ_t, количество запросов, поступивших в систему в момент времени t. Имеем:

$$\mathbb{P}(\theta_t = n) = \frac{(\lambda N)^n}{n!} e^{-\lambda N}, \lambda > 0.$$

Через $\tau_{t,j}$ обозначим длину j-го запроса, поступившего в систему в момент времени t. Полагаем, что $\tau_{t,j}$ случайная величина, имеющая распределение Вейбулла. То есть верно, что

$$\mathbb{P}(\tau_{t,j} > x) \sim e^{-\gamma x^{\alpha}} \text{ при } x \to \infty,$$

где $\alpha \in (0,1), \gamma > 0$. Случайные величины $\tau_{t,j}$ являются независимыми и не зависят от входного процесса поступления заявок.
Здесь и в дальнейшем под выражениями $A(x) \sim B(x)$ при $x \to \infty$ понимается, что

$$\lim_{x \to \infty} A(x)/B(x) = 1,$$

а $A(x) \preceq (\succeq) B(x)$ обозначает неравенства в асимптотическом смысле, т.е. $A(x) \leq (\geq) \sim B(x)$.

Входной процесс Y_t, поступающий в систему в каждый момент времени, определим как

$$Y_t = \sum_{j=1}^{\theta_t} \tau_{t,j}.$$

Средняя нагрузка ρ в системе будет:

$$\rho = \mathbb{E}[Y_t] = N\lambda\mathbb{E}[\tau_{0,1}].$$

В статье рассматривается система обслуживания, состоящая из N обслуживающих приборов, каждый из которых работает в соответствии с дисциплиной обслуживания FIFO, имеет скорость обслуживания $C_i = 1, i = 1..N$ и бесконечный входной буфер.

Заявки, поступающие в систему, распределяются между обслуживающими приборами по следующему правилу маршрутизации: при поступлении j-й заявки в момент времени t, из N обслуживающих приборов случайно выбираются K, множество номеров которых обозначим как $\Omega_{t,j}$, и заявка направляется на выбранный прибор с минимальной длиной очереди, номер которого обозначим через $\chi_{t,j}$. Длина очереди прибора будет определена позже.

Входной процесс, поступающий в очередь i-го обслуживающего прибора в каждый момент времени t,- $Y_{t,i}$ может быть задан как

$$Y_{t,i} = \sum_{j=1}^{\theta_t} \tau_{t,j} I(\chi_{t,j} = i),$$

где $I(A)$- индикатор события A.

Длину очереди i-го обслуживающего прибора определим, как время, необходимое для обслуживания всех заявок, находящихся в очереди и на обслуживании в момент времени t, т.е.

$$W_{t,i} = (W_{t-1,i} + Y_{t,i} - C_i)^+,$$

где $x^+ = max(0, x)$.

В соответствии со способом маршрутизации, с.в. $\chi_{t,j}$, может быть задана следующей формулой:

$$\chi_{t,j} = arg \min_{i \in \Omega_{t,j}} W_{t,i}.$$

Остаточное время обслуживания заявок, находящихся в очереди выбранной при маршрутизации виртуальной заявки в момент времени t, обозначим через D_t и определим как

$$D_t = W_{t-1,\chi_{t,1}}.$$

В дальнейшем, под переполнением системы будет подразумеваться попадание виртуальной заявки при маршрутизации в очередь сервера с остаточным временем обслуживания находящихся в нем заявок, превышающим уровень z_0, т.е. $D_t > z_0$, для больших z_0 без учета времени обслуживания поступившей заявки.

Нас будет интересовать поведение стационарной вероятности переполнения системы:

$$\mathbb{P}(D > z_0) = \mathbb{P}(W_{t-1,\chi_{t,1}} > z_0).$$

Основной результат сформулируем в виде теоремы.

Теорема 1. *Если существует n_1 такое, что*

$$n_1 = arg \min_{N-\rho \le n < K} (n \cdot g^{\alpha}(n) : n \cdot g^{\alpha}(n) < K) \tag{1}$$

$$g(n) = \left(1 + \frac{N-n}{\rho - (N-n)}\right),$$

то при $z_0 \to \infty$ верно:

$$\mathbb{P}(D > z_0) \sim \frac{K!(N-K)!}{n_1!N!} \left(\frac{N\lambda}{\alpha\gamma}\right)^{n_1} (g(n_1))^{n_1(1-\alpha)} e^{-n_1\left(g^{\alpha}(n_1)\gamma z_0^{\alpha} + (\alpha-1)\ln z_0\right)}. \tag{2}$$

Иначе, при $z_0 \to \infty$ верно, что

$$\mathbb{P}(D > z_0) \sim \frac{(N-K)!}{N!} \left(\frac{N\lambda}{\alpha\gamma}\right)^{K} e^{-K\left(\gamma z_0^{\alpha} + (\alpha-1)\ln z_0\right)}. \tag{3}$$

Полученные результаты позволяют проанализировать поведение системы с динамической маршрутизацией, балансирующей нагрузку в сети, и заявками с Вейбулловским распределением времени обслуживания.

Как видно из формул (2) и (3), вероятность переполнения системы имеет существенно другой вид зависимости от параметров маршрутизации по сравнению со случаем, рассмотренным в статье [20] и описывающим поведение системы со степенным распределением времени обслуживания заявок. В отличае от медленного, степенного убывания вероятности переполнения с увеличением параметра K, полученного в [20], вероятность переполнения в системе с Вейбулловским распределением времени обслуживания заявки убывает значительно быстрее,– экспоненциально.

Доказательство полученных результатов основывается на выборе оптимальной конфигурации системы, приводящей к переполнению.

В отличае от случая заявок со степенным распределением времени обслуживания, для которых при $\rho < N - K$ оптимальной была конфигурация, когда в системе в момент t находится K заявок с остаточным временем обслуживания, превышающим пороговое значение z_0, при Вейбулловском распределении времени обслуживания, оптимальной будет конфигурация, когда эти заявки пришли в систему практически одновременно (на отрезке порядка $o(z_0)$). Это является следствием более резкого убывания вероятности возникновения заявки с большим временем обслуживания в распределении Вейбулла, по сравнению со степенным.

Построение оптимальной конфигурации для случая $\rho > N - K$ гораздо сложнее и приводит к необходимости решать нелинейную оптимизационную задачу, выраженную формулой (1). По сравнению с оптимальной конфигурацией для степенного распределения времени обслуживания, когда переполнение происходило если "большие" заявки стояли в

$$n_1 = arg \min_{N-\rho \le n < K} (n)$$

очередях, а остальные очереди переполнялись "маленькими", в случае распределения Вейбулла, даже при $\rho > N - K$, оптимальной может быть конфигурация из K заявок с остаточным временем обслуживания, превышающим пороговое значение z_0.

Доказательство производится методом граничных оценок.

Наиболее технически сложным, является получение верхней границы вероятности переполнения. При доказательстве производится декомпрессия входного потока на "короткие"($\tau \le \varepsilon z$) и "длинные"($\tau > \varepsilon z$) заявки, и рассмотрение вклада этих процессов в переполнение системы.

ДОКАЗАТЕЛЬСТВО

В последующих доказательствах активно используются декомпрессия входного процесса на процессы с "длинными" $Y_t^{T,h}$ и "короткими" $Y_t^{T,l}$ заявками.
Процесс $Y_{t,i}^{T,h}$ соответствует входному потоку заявок с длиной $\tau > \varepsilon T$ в i-ую очередь, т.е.

$$Y_{t,i}^{T,h} = \sum_{j=1}^{\theta_t} \tau_{t,j} I(\chi_{t,j} = i, \tau_{t,j} > \varepsilon T).$$

Процесс $Y_{t,i}^{T,l}$ соответствует входному потоку заявок с длиной $\tau \leq \varepsilon T$ в i-ую очередь, т.е.

$$Y_{t,i}^{T,l} = \sum_{j=1}^{\theta_t} \tau_{t,j} I(\chi_{t,j} = i, \tau_{t,j} \leq \varepsilon T).$$

Процессы $Y_t^{T,l}$ и $Y_t^{T,h}$ независимы по построению и верно, что

$$Y_{t,i} = Y_{t,i}^{T,l} + Y_{t,i}^{T,h},$$

$$Y_t = \sum_{i=1}^{N} Y_{t,i}.$$

В дальнейшем влияние процессов Y_t^l и Y_t^h на вероятность переполнения буфера будет рассматриваться отдельно. Как будет показано, несмотря на то, что $\mathbb{E}Y_{t,i}^{T,l} \to \mathbb{E}Y_{t,i}$ а $\mathbb{E}Y_{t,i}^{T,h} \to 0$ при $T \to \infty$, процесс Y_t^h будет давать существенный вклад в вероятность переполнения буфера.

Кроме того, определим суммарную входную работу поступившую на отрезке $[t-k, t]$, как

$$\overline{S}_{k,i}^{T,l} = \sum_{n=t-k}^{t} Y_{n,i}^{T,l} \text{ и } \overline{S}_{k,i}^{T,h} = \sum_{n=t-k}^{t} Y_{n,i}^{T,h},$$

$$\overline{S}_{k,i} = \overline{S}_{k,i}^{T,l} + \overline{S}_{k,i}^{T,h},$$

$$\overline{S}_k = \sum_{i=1}^{N} \overline{S}_{k,i}.$$

В доказательствах также используется следующие величины:

$$S_{k,i} = \sum_{n=t-k-\varepsilon z}^{t-k} (\tau_{n,j} - ((t-k)-n))^+ I(\chi_{n,j} = i) + \sum_{n=t-k}^{t-\varepsilon z} Y_{n,i} +$$

$$\sum_{n=t-\varepsilon z}^{t} \sum_{j=1}^{\theta_n} min(\tau_{n,j}, t-n) I(\chi_{n,j} = i),$$

$$S_k = \sum_{i=1}^{N} S_{k,i}.$$

Сформулируем и докажем Лемму 1, определяющую нижнюю границу переполнения системы.

Лемма 1. *Если существует n_1 такое, что*

$$n_1 = arg \min_{N-\rho \le n < K} (n \cdot g^\alpha(n) : n \cdot g^\alpha(n) < K)$$

$$g(n) = \left(1 + \frac{N-n}{\rho - (N-n)}\right),$$

то верно, что

$$\mathbb{P}(D > z_0) \succeq \frac{K!(N-K)!}{n_1!N!} \left(\frac{N\lambda}{\alpha\gamma}\right)^{n_1} (g(n_1))^{n_1(1-\alpha)} e^{-n_1\left(g^\alpha(n_1)\gamma z_0^\alpha + (\alpha-1)\ln z_0\right)}. \quad (4)$$

Иначе верно, что

$$\mathbb{P}(D > z_0) \succeq \frac{(N-K)!}{N!} \left(\frac{N\lambda}{\alpha\gamma}\right)^{K} e^{-K\left(\gamma z_0^\alpha + (\alpha-1)\ln z_0\right)}. \quad (5)$$

Доказательство. Для нахождения нижней границы вероятности переполнения системы достаточно построить типичную конфигурацию заявок, вызывающую переполнение, и оценить ее вероятность. Точность нижней границы будет зависеть от обоснованности выбора типичной конфигурации и будет подтверждена совпадением нижней границы переполнения с далее получаемой верхней.

Сначала рассмотрим случай $\rho < N - K$.

Типичной будем считать следующую конфигурацию системы: в момент времени t в K разных очередях, множество номеров которых обозначим через B_1, находятся заявки с длинами больше чем z_0. Понятно, что в следствии условий маршрутизации данная конфигурация может вызвать переполнение и верно, что

$$\mathbb{P}(D > z_0) \succeq \mathbb{P}(\exists B_1 : \forall i \in B_1 : W_{t,i} > z_0) \cdot \mathbb{P}(\Omega_{t,1} = B_1).$$

Оценим вероятность $\mathbb{P}(\exists B_1 : \forall i \in B_1 : W_{t,i} > z_0)$. Для этого рассмотрим событие $A_t(n, z_0)$, состоящее в том, что в системе в момент времени t находятся хотя бы n заявок таких, что

$$(\tau_{t_j} - (t - t_j)) > z_0 \text{ для } \forall j = \{1, ..., n\}, \quad (1.1)$$

где t_j время поступления j-ой заявки в систему.

Докажем, что если в системе имеет место событие $A_t(n, z_0)$, то в момент времени t, верно что

$$\exists B_1 : \forall i \in B_1 : W_{t,i} > z_0. \quad (1.2)$$

Понятно, что если все заявки, формирующие событие $A_t(n, z_0)$ при маршрутизации поступили в разные очереди, то из условия (1.1) следует существование (1.2).

Допустим, что какая то заявка j_l поступила в очередь, в которой уже стоит заявка из события $A_t(n, z_0)$, тогда в момент времени t_{j_l} поступления этой заявки, из условий маршрутизации следует, что есть хотя бы n очередей таких, что

$$W_{t,k_i} > z_0 + (t - t_{j_l}) \text{ для } \forall i \in \{1, ..., n\}, k_1 < k_2 < ... < k_n,$$

откуда и следует (1.2).

Поэтому верно, что

$$\mathbb{P}(\exists B_1 : \forall i \in B_1 : W_{t,i} > z_0) \ge \mathbb{P}(A_t(K, z_0)).$$

Далее, в условиях сформулированной модели, исходя из структуры входного потока и свойств распределения Пуассона, следует, что количество заявок $\theta_t^{z_0}$, находящихся в системе в момент времени t и удовлетворяющих условию (1.1),– также пуассоновская случайная величина с интенсивностью Λ:

$$\mathbb{P}(\theta_t^{z_0} = n) = \frac{\Lambda^n}{n!} e^{-\Lambda},$$

$$\Lambda = \int_{z_0}^{\infty} N\lambda e^{-\gamma x^\alpha} dx \sim \left(\frac{N\lambda}{\alpha\gamma}\right) e^{-\gamma z_0^\alpha + (1-\alpha)\ln z_0},$$

откуда следует, что

$$\mathbb{P}(A_t(K, z_0)) \sim \frac{\Lambda^K}{K!}.$$

Из условий маршрутизации следует, что

$$\mathbb{P}(\Omega_{t,1} = B_1) = \frac{K!(N-K)!}{N!}.$$

Окончательно получаем, что

$$\mathbb{P}(D > z_0) \geq \frac{(N-K)!}{N!} \cdot \Lambda^K,$$

т.е верна формула (4). Перейдем к доказательству второго случая.

Пусть $\exists n < K : \rho > N - n$.

В качестве типичной конфигурации системы будем считать следующее состояние: в n очередях в момент времени t находятся одиночные заявки с остаточным временем обслуживания превышающим уровень z_0. Все они пришли не позднее чем $t - z_1$. В систему, находящуюся в этом состоянии, поступают только заявки с длиной $\tau < \varepsilon T$. Понятно, что находящаяся в данном состоянии система будет переполняться, и найдем вероятность переполнения.

Определим через $A_t(z_1, n, z_0)$ событие, состоящее в том, что в системе в момент времени t находятся хотя бы n заявок, удовлетворяющих условию (1.1) и таких, что

$$\min_{j=\overline{1..n}} (t - t_j) = z_1, \tag{1.3}$$

$$z_1 = \left(\frac{N-n}{\rho - (N-n)} + \delta\right) z_0, \ \delta > 0.$$

По предположению, в системе имеет место событие $A_t(z_1, n, z_0)$, и в систему поступают только заявки из входного процесса Y_t^l.

Понятно, что после маршрутизации заявок, формирующих событие $A_t(z_1, n, z_0)$, найдется хотя бы n разных очередей, в которых будут стоять заявки, удовлетворяющие(1.1) и (1.3)(см. выше). Множество номеров этих очередей обозначим через B_1.

Из $N - n$ неперегруженных очередей, множество номеров которых обозначим $B_2 = B_0/B_1$, выберем $K - n$ с максимальной длинной в момент времени t– множество B_3. Через B_0 здесь обозначено множество номеров всех очередей.

Из определений $W_{t,i}$ и $\overline{S}_{k,i}$ следует, что

$$\sum_{i \in B_2} W_{t,i} \geq \sum_{i \in B_2} \left(\overline{S}_{z_1,i}^{T,l} - z_1\right). \tag{1.4}$$

Замечая, что средняя длина всех неперегруженных очередей (множество B_2) меньше чем средняя длина очередей из множества B_3, и используя соотношение (1.4), имеем:

$$\sum_{i \in B_3} W_{t,i} \geq \frac{K-n}{N-n} \sum_{i \in B_2} \left(\overline{S}_{z_1,i}^{T,l} - z_1\right). \tag{1.5}$$

Заметим также, что из условий маршрутизации следует соотношение:

$$\min_{i\in B_3} W_{t,i} + \varepsilon T \geq \max_{i\in B_3} W_{t,i}. \tag{1.6}$$

Объединяя (1.5) и (1.6), получаем:

$$\min_{i\in B_3} W_{t,i} \geq \frac{1}{N-n} \sum_{i\in B_2} \left(\overline{S}^{T,l}_{z_1,i} - z_1\right) - \varepsilon T,$$

и учитывая, что

$$\sum_{i\in B_2} \overline{S}^{T,l}_{z_1,i} \geq S^{T,l}_{z_1-\varepsilon T},$$

получаем

$$Pr\left(\min_{i\in B_3} W_{t,i} > z_0\right) \geq 1 - Pr\left(\left(S^{T,l}_{z_1-\varepsilon T} - (z_1-\varepsilon T)(N-n)\right) \leq (1+2\varepsilon)(N-n)z_0\right). \tag{1.7}$$

Замечая, что при условии

$$\delta \geq \frac{(2\varepsilon+\delta_2)(N-n)\rho}{(\rho-\delta_1)-(N-n)},$$

верно неравенство:

$$\left(S^{T,l}_{z_1-\varepsilon T} - (z_1-\varepsilon T)(N-n)\right) - (1+2\varepsilon)(N-n)z_0 \leq \left(S^{T,l}_{z_1-\varepsilon T} - (\rho-\delta_1)(z_1-\varepsilon T)\right) + \delta_2 z_0,$$

можно применить лемму 2.1 из [19]:

Лемма. *Для любой функции $g(T) \geq T$, любой константы $c_1 > 0$ и достаточно малых $\delta_1 > 0, \delta_2 > 0, \varepsilon > 0, \varepsilon < (\gamma/(4c_1))^{1/(1-\alpha)}, \varepsilon < \delta_2/(2\rho)$ верно, что*

$$\mathbb{P}\left(\left|S^{T,l}_k - k\rho\right| > k\delta_1 + \delta_2 g(T)\right) \leq e^{-c_1\tilde{\delta}_2 T^{\alpha-1} g(T)},$$

для достаточно большого T и $\tilde{\delta}_2 = \delta_2 - (\rho+\delta_1)\varepsilon$.

Из леммы и из (1.7) следует, что при $T = z$

$$\mathbb{P}(\min_{i\in B_3} W_{t,i} > z) \geq \left(1 - e^{-O(z)}\right).$$

Окончательно получаем, что

$$\mathbb{P}(D > z_0) \succeq \mathbb{P}(A_t(n, z_0+z_1)) \cdot \mathbb{P}(\min_{i\in B_3} W_{t,i} > z_0) \cdot \mathbb{P}(\Omega_{t,1} = B_1 \cup B_3) \blacktriangleleft$$

Далее сформулированы 4 леммы, в совокупности являющиеся доказательством верхней границы вероятности переполнения, и указаны ключевые моменты доказательства.

Лемма 2. *Пусть в момент времени $t = 0$ заданы все очереди системы $W_{i,0}$, и в систему поступают заявки только из входного процесса $Y_{t,l}$, тогда при достаточно большом z_0 и достаточно малом $\varepsilon > 0$ на отрезке времени $[0,T], \varepsilon z_0 \ll T \ll z_0$, с вероятностью $1 - e^{-c_1 z_0^{\alpha}}$ (c_1– сколь угодно большое число, независящее от z_0;$\delta_1 > 0, \delta_2 > 0$–сколь угодно малые константы) выполняется:*

$$W_{t,i} = \max\left((W_{i,0} - t + \rho\mu_{t,i}t + \delta_1 t), 0\right) + \delta_2 z_0, \tag{6}$$

где

$$\mu_{t,i} = \frac{(N-K)!n_{t,i}!}{N!(n_{t,i}-K+1)!},$$

$n_{t,i}$– *число таких очередей* j, *что* $W_{t,j} \geq W_{t,i}$, *в противном случае имеет место следующее неравенство:*

$$\mathbb{P}(|W_{T,i} - W_{0,i}| > x| \text{ формула (6) не верна }) \leq e^{-c_1(x^\alpha - z_0^\alpha)} \tag{7}$$

при достаточно большом z_0, $x > z_0$, c_1– *сколь угодно большое число, независящее от* z_0.

Доказательство леммы основано на применении леммы 2.1 из [19].

Лемма 3. *Пусть* $\{A_k\}$ *событие, состоящее в том, что в момент времени* t_0 *в системе имеется* K *или больше очередей с длиной большей* $z_0(W_{t,i_j} > z_0, i_1, i_2, ...i_n, n \geq K)$, *тогда существует такая константа* c_0, *независящая от* z_0, *что для момента времени* $t_2 : t_2 = t_0 - c_0 z_0$ *верно, что*

$$\mathbb{P}(A_k) \sim \mathbb{P}(A_k | W_{t_2,0} = 0, ..., W_{t_2,N} = 0) \text{ при } z_0 \to \infty. \tag{8}$$

Доказательство леммы производится следующим образом. Во-первых, сначала делается довольно грубая оценка сверху длины очереди в точке t_2, в качестве ограничивающего процесса используется суммирующий нагрузку процесс $S_{t_2,k}$. Затем оценивается вклад в вероятность переполнения системы в точке t_0 от заявок, поступивших на отрезке $[t_2, t_0]$ и от накопленных к моменту времени t_2. Оказывается, что вероятность переполнения в точке t_0 слабо зависит от предистории в точке t_2.

Лемма 4. *Для события* $\{A_k\}$ *и момента времени* t_2, *определяемых Леммой 3, справедливо:*

$$\mathbb{P}(A_k) \sim \mathbb{P}(A_k| \text{ на интервале } (t_2, t_0) \text{ всюду выполняется условие (6) из Леммы 2}). \tag{9}$$

Для доказательства этой леммы достаточно посчитать вероятность выброса при поступлении в систему входного процесса Y_t^l. Легко показать, что эта вероятность порядка $z_0 e^{-c_1 x^\alpha}$, где x величина выброса, а c_1 достаточно большая константа.

Лемма 5. *Рассмотрим процесс поступления длинных заявок*($\tau > \varepsilon z_0$) *в систему на интервале* (t_2, t_0). *Пусть событие* $\{B_k\}$ *заключается в том, что на интервале* (t_2, t_0) *поступило* K *длинных заявок и моменты их возникновения* $x_1, ..., x_K$ *и длин* $\tau_1, ..., \tau_K$ *такие, что* $|x_i - t_0| = o(z_0), x_i < t_0, \tau_i > z_0 + x_i, |\tau_i - x_i - z_0| = o(z_0)$, *а событие* $\{D_n\}$ *заключается в том, что на интервале* (t_2, t_0) *поступило* n *длинных заявок и для моментов их возникновения* $x_1, ..., x_n$ *и длин* $\tau_1, ..., \tau_n$, *имеем:* $x_i < t_0 - d, |x_i - t_0 + d| = o(z_0), \tau_i > d + z_0, |\tau_i - d - z_0| = o(z_0)$, *где* $d = \frac{z_0}{(\rho/(N-n))-1}$, *тогда имеем либо:*

$$\mathbb{P}(A_k) \sim \mathbb{P}(A_k|B_k) \text{ при } z_0 \to \infty \ ,$$

либо $\exists n > 0, K > n$:

$$\mathbb{P}(A_k) \sim \mathbb{P}(A_k|D_k) \text{ при } z_0 \to \infty \ .$$

Для доказательства леммы используется метод детальной оптимизации, т.е. рассматривается некоторая конфигурация, вызывающая переполнение, и варьируются длины и моменты прихода заявок.

В результате приходим к тому, что при $\rho < N - K$ оптимальной будет конфигурация, когда в системе в момент t находится K заявок с остаточным временем обслуживания, превышающим

пороговое значение z_0, и все эти заявки пришли в систему практически одновременно (на отрезке порядка $o(z_0)$).

В случае $\rho > N - K$ оптимальной будет либо предыдущая конфигурация, либо конфигурация, когда в n очередях системы стоят длинные заявки, а остальные очереди переполнялись "маленькими"($\tau \leq \varepsilon z_0$). Все зависит от решения нелинейной оптимизационной задачи (1).

3. ЗАКЛЮЧЕНИЕ

Итак, мы получили асимптотически точные формулы для оценки вероятности переполнения в системе с динамической маршрутизацией, которые позволяют проанализировать зависимость вероятности от параметров маршрутизации.

Во-первых, стоит заметить, что зависимость вероятности переполнения от параметра маршрутизации K более резкая, чем в случае системы со степенным времени обслуживания. Вероятность уменьшается экспоненциально, а не по степенному закону.

В случае, когда нагрузка в системе невелика, а именно $\rho < N-K$, вероятность переполнения сильно зависит от количества выбираемых при маршрутизации очередей, т.к. их количество K стоит в показателе экспоненты. Поэтому, увеличивая количество выбираемых при маршрутизации очередей, можно существенно уменьшить вероятность переполнения системы.

В случае же, когда в систему поступает существенная нагрузка, а именно $\rho > N - K$, вероятность переполнения имеет другой вид зависимости от количества выбираемых при маршрутизации очередей. Если решение нелинейной оптимизационной задачи (1) существует, то увеличивать K больше n_1 нет необходимости,– к существенному уменьшению вероятности переполнения это не приведет. Если же решения нет, то увеличивая K, можно значительно снизить вероятность переполнения.

СПИСОК ЛИТЕРАТУРЫ

1. Н.Д. Введенская, Р.Л. Добрушин, Ф.И. Карпелевич, *Система обслуживания с выбором наименьшей из двух очередей - асимптотический подход*, Пробл. передачи информ., **32**, No 1, стр. 15-27, 1996.

2. P. Jacquet, N.D. Vvedenskaya, *ON/OFF Sources in an Interconnection Network: Performance Analysis when Packets are Routed to the Shortest Queue of two Randomly Selected Nodes*, Rapport de Recherche No 3570, INRIA, pp 1-35, December 1998.

3. J.B. Martin, Yu. M. Suhov, *Fast Jackson networks*, J. Appl. Prob., **9**, pp 854-870, 1999.

4. S. N. Ethier, T. G. Kurtz, *Markov Processes Characterization and Convergence*, N.Y. etc: John Willey and Stons ,1986.

5. W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, On the self-similar nature of Ethernet traffic(extended version),*IEEE/ACM Trans. on Networking*, 1994, no.2, pp.1–15.

6. V. Paxon, S. Floyd, Wide area traffic: the failure of Poisson modeling, *IEEE/ACM Trans. on Networking*, 1993, no.3, pp.226–244.

7. N. G. Duffied and N. O'Connel, Large deviation and overflow probabilities for the general single-server queue with applications,*Math. Proc. Camb. Phil. Soc*, 1995, no.118(1), pp.363–374.

8. A. Botvich and N. G. Duffied, Large deviation, economies of scale and the shape of the loss curve in large multiplexers,*Queueing Systems*, 1995, no.20, pp.293–320.

9. C. Courcoubeties and R. Weber,Buffer overflow asymptotics for a switch handling many traffic soutces,*J. Appl. Prob.*, 1996, vol.33, no.3, pp.886–903.

10. N. Likhanov and R. Mazumdar,Cell loss asymptotics in buffers fed with a large number of independent stationary sources,*J. Appl. Prob.*, 1999, no.36, pp.86–96.

11. A. Simonian and J. Guibert, Large deviations approximation for fluid queues fed by a large number of ON/OFF sources,*IEEE J. Sel. Areas Commun.*, 1995, vol. 13, no.6, pp.1017–1027.

12. I. Norros, A storage model with self-similar input,*Queueing Systems*, 1994, no.16, pp.387–396.

13. J. Choe and N. B. Shroff, On the supremum distribution of integrated stationary Gaussian processes with negative linear drift,*Advances in Applied Probability*, 1999, no.31, pp.135–157.

14. M. Parulekar and A. M. Makowski, Tail probabilities for $M/G/\infty$ input processes,*Queueing Systems*, 1994, no.27, pp.271–296.

15. N. Likhanov, Bounds on the buffer occupancy with self-similar input traffic,in *Self-similar network traffic and performance evaluation*, K. Park and W. Willinger eds., Wiley, 2000, pp.193–214.

16. Z. Liu,P.Nain, D.Towsley and Z-L. Zhang, Asymptotics behavior of a multiplexer fed by long-range dependent process, *Journal of Applied Probability*, , 1999, no.36, pp.105–118.

17. N. Likhanov and R. Mazumdar,Loss asymptotics in large buffers fed by heterogeneous long-tailed sources, *Advances in Applied Probability*, 2000, no.32, pp.1168–1189.

18. S. Asmussen and C. Kluppelberg,Stationary M/G/1 excursions in the presence of heavy tails,*J. Appl. Prob.*, 1997, no.34, pp.208–212.

19. N. Likhanov, R. Mazumdar, O. Ozturk, Large buffer asymptotics for fluid queues whith heterogeneous $M/G/\infty$ Weibullian inputs,*Queueing Systems*, 2003, no.45, pp.333–356.

20. А. В. Аленичев , Н. Б. Лиханов, Динамическая маршрутизация в системе с заявками, имеющими степенной закон распределения времени обслуживания, *Информационные процессы*, 2005, т. 5, №3, стр. 213-226.