УДК 621. 395

Ложковский А.Г.
Loshkovski A.G.

МЕТОД РАСЧЕТА СТАЦИОНАРНОГО РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СОСТОЯНИЙ В МОДЕЛИ ПАЛЬМА

CALCULATION METHOD OF STATIONARY PROBABILITIES DISTRIBUTION OF STATE IN MODEL A PALM

Аннотация. Исследована зависимость стационарного распределения состояний системы от дисперсии интенсивности нагрузки, образованной рекуррентным потоком вызовов. Предложен метод расчета стационарных вероятностей состояний системы типа GI/D/m.

Summary. The dependence of stationary distribution of system state from a traffic dispersion obtained by reneval calls flow is probed. The calculation method of stationary probabilities of state of system such as GI/D/m is offered.

Решение проблемы разработки методов анализа и расчета систем распределения информации, адекватных текущему уровню развития телекоммуникаций и реальным моделям потоков нагрузки, циркулирующих в них, имеет важное народно-хозяйственное значение для отрасли связи. Поэтому, в расчётах оборудования сетей связи кроме методов, основанных на модели простейшего и примитивного потока вызовов, необходимо использовать и методы, базирующиеся на модели рекуррентного потока типа Пальма. Такой поток, являясь обобщением «идеализированного» простейшего, по своим свойствам ближе к реальным потокам вызовов современных телекоммуникационных сетей или точнее их описывает [1–3]. В обобщённом рекуррентном потоке функция распределения длительности промежутка времени между вызовами $P(z \le t) = F(t)$ может быть произвольной, а не только экспоненциальной, как у простейшего потока (пуассоновского потока первого рода). При этом в случайном процессе формирования потока вызовов дисперсия количества вызовов в условную единицу времени σ^2, как обобщённая характеристика неравномерности процесса поступления вызовов, может быть существенно больше его математического ожидания Λ. Именно эта особенность и является весьма типичной для реальных потоков вызовов нынешних сетей связи. Для частного случая рекуррентного потока, известного под названием простейшего, все задачи данной проблемы в основном решены Эрлангом. Для случая обобщенного рекуррентного потока при экспоненциальной длительности обслуживания решение найдено Такачем [4], однако для регулярного закона распределения длительности обслуживания применение данного метода невозможно. Целью данной работы является разработка метода расчета полнодоступной системы с постоянной длительностью занятия её приборов, обслуживающей рекуррентный поток вызовов.

При проектировании систем обслуживания в телекоммуникациях часто необходимо знать отдельные значения функции распределения вероятностей состояний системы P_i. Например, в случае простейшего потока вызовов вероятность занятия всех приборов полнодоступной системы соответствует вероятности потерь по времени или вероятности потери вызова, которые являются основными характеристиками качества обслуживания. В случае рекуррентного потока вызовов эти же характеристики также находятся исходя из распределения вероятностей состояний системы, однако метод их расчёта значительно сложнее. В книге Такача [4] даны формулы определения стационарных вероятностей состояний системы в модели Пальма при экспоненциальной длительности обслуживания, полученные на основе метода вложенных цепей Маркова. В общем случае для полнодоступного пучка из m приборов вероятности P_i рассчитываются так:

$$P_i = \sum_{j=i}^{m} (-1)^{j-i} \cdot C_j^i \cdot B_j \,, \tag{1}$$

где B_j – биноминальный момент обслуженной нагрузки. Для его нахождения используется формула:

$$B_j = \frac{m \cdot X_{j-1}}{j} \cdot \frac{\sum_{k=j-1}^{m-1} C_{m-1}^k \cdot X_k^{-1}}{(1 + m \cdot \frac{t}{T}) \cdot \sum_{k=0}^{m-1} C_{m-1}^k \cdot X_k^{-1}}, \tag{2}$$

где T – средняя длительность обслуживания вызова; t – средняя длительность промежутка времени между вызовами (естественно, t / T = 1 / Λ). Следует учесть, что T имеет экспоненциальную функцию распределения, а t – произвольную, и при этом они стохастически и взаимно независимы. Если и t экспоненциально, то P_i подчиняется распределению Эрланга.

Для вычисления формального параметра X_j используется формула:

$$X_j = \prod_{k=1}^{j} \frac{F(k \cdot t)}{1 - F(k \cdot t)} \,. \tag{3}$$

По данному способу расчета следует сделать два замечания:

1. Как видно из (3), в довольно сложном способе расчета вероятностей P_i требуется задание функции распределения длительности промежутка времени между вызовами $F(t)$. Эта функция, являясь достаточной для описания рекуррентного потока, в реальных условиях, как правило, неизвестна. В телекоммуникационных сетях при практическом исследовании параметров нагрузки, создаваемой потоками вызовов, измеряется не длительность промежутка времени между вызовами, а интенсивность нагрузки Λ или, иначе, среднее количество вызовов в условную единицу времени. По всем значениям периодических отсчётов легко рассчитать степень разброса отдельных значений от среднего или дисперсию выборки. Следовательно, рекуррентные потоки вызовов лучше описывать с помощью распределения вероятностей количества вызовов в условную единицу времени [5, с. 23].

2. В телекоммуникациях кроме экспоненциального закона распределения длительности обслуживания (например, для описания случайной длительности телефонного разговора) используют ещё и регулярный закон (например, для описания длительности работы управляющих устройств узлов коммутации или для описания длительности обработки пакетов в сетях с коммутацией пакетов). В этом случае применение упомянутого способа расчета невозможно, поскольку в методе вложенных цепей Маркова количество неэкспоненциально распределённых величин не может быть более одной.

Поэтому в данной работе разработан новый метод расчета стационарных вероятностей состояний полнодоступной системы P_i, обслуживающей рекуррентный поток вызовов с интенсивностью нагрузки Λ и дисперсией σ^2 при постоянной длительности занятия приборов T.

Рассматривается модель полнодоступной системы с отказами, состояние которой определяется количеством занятых приборов. Случайный процесс поступления вызовов модифицирует состояния системы со скоростью, определяемой интенсивностью нагрузки Λ. Интенсивность нагрузки Λ – это суммарное количество поступивших вызовов j за время, равное средней длительности занятий приборов T. Таким образом, интенсивность перехода системы из одного состояния в другое зависит от свойств потока вызовов, который описывается соответствующим распределением вероятностей поступающих вызовов Q_j, где j – количество вызовов за время T, принимающее значения от 0 до ∞. Все события поступления вызовов принадлежат пространству состояний Q. События занятий приборов образуют новое дискретное подпространство P и описываются распределением вероятностей P_i, где i – число занятых приборов, принимающее значения от 0 до m (m – ёмкость системы). Пространство P, несомненно, меньше пространства Q, поскольку состояния $i > m$ для системы невозможны, а j может быть сколь угодно велико.

Так как система обслуживания с отказами (потерями), то независимо от её текущего состояния для любого из случаев поступления $j > m$ вызовов происходит событие «потеря вызовов». Это очевидно, поскольку за время T, равное постоянной длительности занятия приборов, поступит $j > m$ вызовов и ни один из вновь занявшихся приборов за это же время не освободится. Если система находится в начальном состоянии $i = 0$ (все приборы свободны), то в этом случае для любого из вариантов поступления за тот же отрезок времени $j \le m$ вызовов событие «потеря вызовов» не происходит. События, состоящие в поступлении за время T точно j вызовов, образуют полную группу несовместных гипотез $H_0, H_1, \ldots, H_j$ с априорными вероятностями $Q(H_0), Q(H_1), \ldots, Q(H_j)$ и поэтому $\sum_{j=0}^{\infty} Q(H_j) = 1$. Событие A, состоящее в «отсутствии потерь вызовов» может происходить только вместе с одной из гипотез группы $H_0, H_1, \ldots, H_m$. Условные вероятности гипотез поступления $j \le m$ вызовов (апостериорные), при условии осуществления события A (отсутствие потерь вызовов) вычисляются по формуле Бейеса:

$$Q(H_j \mid A) = \frac{Q(H_j) \cdot Q(A \mid H_j)}{\sum_{k=0}^{m} Q(H_k) \cdot Q(A \mid H_k)}. \qquad (4)$$

Поскольку событие A (отсутствие потерь), появляющееся с любой из гипотез группы $H_0, H_1, \ldots, H_m$ является достоверным, то условные вероятности $Q(A|H_j) = 1$. Поэтому, если для осуществления события возможна только часть гипотез, а остальные невозможны, то для получения апостериорных вероятностей нужно каждую из априорных вероятностей этой части возможных гипотез разделить на их сумму. Следовательно:

$$Q(H_j \mid A) = \frac{Q(H_j)}{\sum_{k=0}^{m} Q(H_k)}. \qquad (5)$$

Для системы, находящейся в начальном состоянии (все приборы свободны), условные вероятности $Q(H_j|A)$ соответствуют вероятностям занятия $i = j$ приборов системы, т.е. вероятностям P_i. Они также образуют полную группу несовместных событий подпространства P и поэтому

$\sum_{i=0}^{m} P_i = 1$. Таким образом, вероятности занятия приборов P_i выражены через вероятности поступления вызовов $Q(H_j)$, т.е.

$$P_i = \frac{Q(H_i)}{\sum_{k=0}^{m} Q(H_k)}. \qquad (6)$$

Условием стационарности полученного распределение является *эргодичность* процесса обслуживания вызовов. Это значит, что полученные вероятности состояний системы (числа занятых приборов) не должны зависеть от того, в каком состоянии система была в начальный момент (было принято, что в начальный момент все приборы свободны). Известно, что свойством эргодичности обладают *марковские* процессы и для любого такого процесса после достаточно длительного времени функционирования системы обязательно наступит стационарный режим, где вероятность того, что система будет в *i*-м состоянии, не зависит от того, в каком состоянии она находилась в начальный момент времени. Согласно теореме Маркова любой транзитивный (из любого состояния можно перейти в любое другое) однородный (вероятности перехода из состояния в состояние не зависят от того, в какой момент времени начало перехода) процесс с конечным числом состояний обладает эргодическим свойством [6, с. 149]. Согласно указанному в эргодической теореме [7, с. 33] достаточному критерию эргодичности *марковского* процесса, процесс будет *эргодичен*, если выполняется условие:

$$\Lambda = \lambda/\mu < m, \qquad (7)$$

где Λ – поступающая нагрузка; λ – интенсивность поступления вызовов; μ – интенсивность обслуживания вызовов; m – число приборов. Это значит, что в среднем должно поступать меньше вызовов, чем их может обслужиться или среднее число вызовов за время $T = 1 / \mu$ не должно превышать числа приборов, тогда при этом система работает в состоянии статистического равновесия.

Примером *марковского* процесса, обладающего *эргодическим* свойством, является обслуживание полнодоступной системой простейшего потока вызова, где *марковский* процесс однородный по причине отсутствия последействия, поскольку промежутки времени между вызовами распределены экспоненциально. Если интервалы времени между событиями распределены экспоненциально, то количество событий в единичном интервале имеет распределение Пуассона. Подставив в (6) вместо вероятностей $Q(H_i)$ плотность распределения Пуассона $Q_i = \frac{\Lambda^i}{i!} \cdot e^{-\Lambda}$, получим известное первое распределение Эрланга:

$$P_i = \frac{\frac{\Lambda^i}{i!}}{\sum_{k=0}^{m} \frac{\Lambda^k}{k!}}, \quad i = 0, 1, 2, \ldots m. \qquad (8)$$

Однако в этом случае, непременно, математическое ожидание случайного процесса поступления вызовов Λ равно его дисперсии σ^2.

Как показывают практические исследования параметров потоков вызовов, дисперсия количества вызовов в условную единицу времени может во много раз превосходить его математическое ожидание [1–3]. При этом реальное распределение интенсивности нагрузки лучше согласуется с нормальным (Гаусса) законом распределения, а не с пуассоновским. (Уже при $\Lambda > 10$ имеет место хорошее совпадение между огибающими закона Пуассона и нормальным законом распределения для случая $\Lambda = \sigma^2$). После подстановки в (6) вместо вероятностей $Q(H_i)$ плотности вероятности нормального закона получим:

$$P_i = \frac{\frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(i-\Lambda)^2/2\sigma^2}}{\sum_{k=0}^{m} \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-(k-\Lambda)^2/2\sigma^2}} = \frac{1}{\sum_{k=0}^{m} \exp\left[\frac{-(k-2\cdot\Lambda+i)\cdot(k-i)}{2\sigma^2}\right]}. \qquad (9)$$

Поскольку при $\Lambda = \sigma^2$ нормальный случайный процесс также является *марковским* [6, с. 238], то в этом случае в соответствии с (7) при $\Lambda < m$ расчетные значения, получаемые по формулам (8) и (9) весьма близки – расхождение не более 5%.

Формула (8) справедлива для простейшего потока, где распределение промежутков времени между вызовами экспоненциально и $\sigma^2 = \Lambda$, однако в реальных потоках $\sigma^2 > \Lambda$ для большинства видов телекоммуникационных сетей. По результатам исследований [1–3] установлено, что неплохая степень согласия реальных потоков вызовов и теоретических законов распределения наблюдается

при аппроксимации их рекуррентным потоком Пальма с гиперэкспоненциальным распределением промежутков времени между вызовами (достаточно двух экспонент). Поскольку при $\sigma^2 > \Lambda$ распределение промежутков между вызовами уже не экспоненциально, то поток вызовов теряет свойство отсутствия последействия. При этом процесс обслуживания становится сложнее и возникает зависимость вероятностей состояний системы от её начального состояния. Для апробации адекватности распределения (9) реальным потоком вызовов использовано статистическое моделирование, схема которого дана в [8]. Цель моделирования – определить степень зависимости стационарного распределения состояний системы P_i от её начального состояния.

На рис. 1 показаны гистограммы, полученные в результате моделирования обслуживания рекуррентного потока в системе *GI/D/m* при $\sigma^2 / \Lambda = 4$ и огибающие (сплошные линии) распределений состояний системы P_i, рассчитанные по (9). Интенсивность нагрузки составляет $\Lambda = 100$ вызовов за постоянную длительность обслуживания T, а ёмкость системы $m = 120$, 130, 140 и 150 приборов.

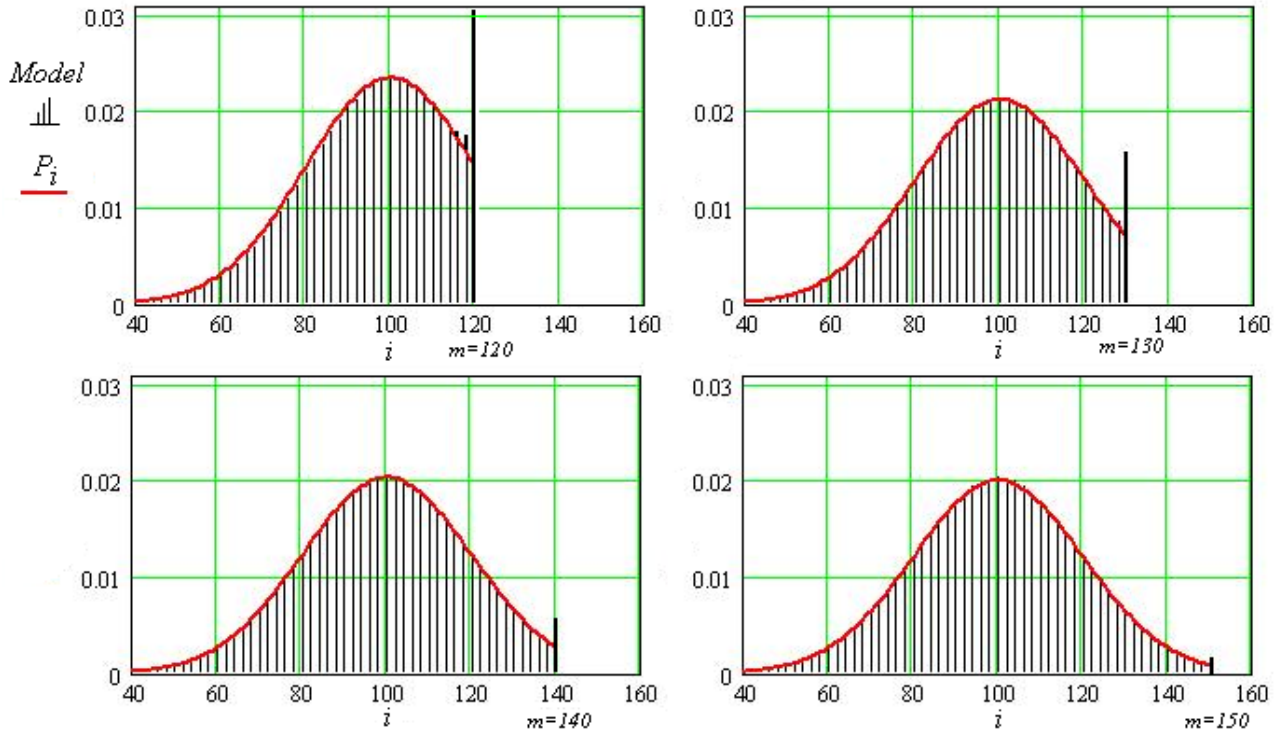

Рисунок 1 – Гистограммы результатов моделирования
и огибающие распределения состояний системы P_i

Данные гистограммы свидетельствуют о достаточно точной аппроксимации статистики реального поведения системы функцией распределения (9). В сериях опытов имитационного моделирования в широком диапазоне задаваемых параметров потока Λ, σ^2 и ёмкостей системы m относительная ошибка аппроксимации не превышала 1% для всех значений P_i, кроме $P_{i=m}$. Вероятность $P_{i=m}$, т.е. состояния занятия всех m приборов системы (выделено жирной линией), совершенно не вписывается в предлагаемое распределение при некоторых соотношениях Λ, σ^2 и m. При $\Lambda = \sigma^2$ (экспоненциальное распределение промежутков) и любом m результаты моделирования и расчета (9) в этой точке достаточно близки. С ростом σ^2 вызовы потока имеют тенденцию «скучиваться» или группироваться, поскольку короткие промежутки времени между вызовами в гиперэкспоненциальном распределении еще более вероятны, чем длинные. В этих же условиях при неограниченном количестве линий ($m = \infty$) вызовы обслуживаются без потерь. Из-за постоянного времени обслуживания всех вызовов без потерь свойства потока освобождений приборов совпадают со свойствами потока поступления вызовов, так как происходит только сдвиг по времени на величину T между моментом поступления вызова и моментом окончания его обслуживания. При этом состояния системы обслуживания полностью определяются свойствами потока вызовов, и её начальное состояние никак не отражается на функции распределения количества вызовов в системе (состояние системы P_i). По мере уменьшения ёмкости системы её начальное состояние уже начинает оказывать влияние на вероятности P_i, но в наибольшей степени на вероятность $P_{i=m}$, и очень незначительно на все остальные вероятности распределения состояний системы (см. рис. 1 для

m = 150, 140, 130 и 120). Здесь очевидно, что в каждом из опытов по мере приближения m к Λ влияние начального состояния системы возрастает. Объясняется это тем, что бóльшая из-за повышенной дисперсии σ^2 мгновенная «скученность» вызовов на интервале T приводит к увеличению относительно среднего значения Λ общего количества вызовов, приходящегося на данный интервал времени. Поскольку от начального состояния системы (сколько приборов уже было занято) зависит количество свободных мест для вновь поступающих вызовов, то тем больше зависимость от этого состояния именно вероятности занятия всех m приборов системы. Вызовы, получающие отказ в обслуживании ($> m$), или покидающие систему (уже обслуженные) тут же «восполняются» новыми из большой группы «скучившихся» на данном интервале времени. Они поддерживают систему более продолжительное время в состоянии «насыщения» (все приборы заняты) и тем самым увеличивают вероятность $P_{i=m}$. При этом возникает кратковременная перегрузка системы и не выполняется условие эргодичности процесса (7).

По результатам статистического моделирования [8] установлено, что по найденному в (9) значению $P_{i=m}$ можно рассчитать реальную вероятность P_m следующим образом:

$$P_m = P_{i=m} \cdot \left[\left(\frac{\sigma^2}{\Lambda} \right)^{0.8} \cdot f(x) \right], \qquad (10)$$

где $f(m)$ – линейная функция, принимающая значения от 1 до 0,1 при изменении значений m от Λ до 2Λ (при $\sigma^2 = \Lambda$ функция $f(m) = 1$ при любом m).

Итак, предложенное распределение (9) позволяет непосредственно вести расчеты стационарных вероятностей состояний полнодоступной системы P_i, обслуживающей рекуррентный поток вызовов с параметрами нагрузки Λ и σ^2 при постоянной длительности занятия приборов T (во всех точках, за исключением точки P_m).

Особо актуально распределение (9) для современных мультимедийных сетей, которые, как правило, являются сетями с коммутацией пакетов и постоянной длительностью их обслуживания. Здесь при широком диапазоне скоростей передачи (от сотен бит/с до сотен Мбит/с) источники каждой службы характеризуются максимальной (пиковой) и средней скоростями передачи, их соотношением. Значит, потоки вызовов здесь нестационарные и не простейшие. Интегральной оценкой всех этих факторов в этом случае может быть дисперсия интенсивности нагрузки. Распределение (9), позволяющее учитывать эту дисперсию, может быть применено, например, для расчета условного количества выходных портов пакетного коммутатора, среднего числа задержанных источников (пакеты которых находятся в буфере дольше одного цикла), среднего заполнения входного буфера и пр.

Литература

1. *Ложковский А.Г., Захарченко Н.В., Горохов С.М.* Экспериментальная оценка модели потока вызовов на современных телефонных сетях // Наукові праці ОНАЗ ім. О.С. Попова. – 2001. – №2. – С. 40-43.
2. *Ложковский А.Г.* Исследование параметров телефонной нагрузки на сотовой сети мобильной радиосвязи // Труды УНИИРТ. – 2001. – №3. – С. 10–14.
3. *Ложковский А.Г., Гайворонская Г.С*. Исследование параметров потоков вызовов на национальной телефонной сети Украины // Труды II международной НПК «Современные информационные и электронные технологии». – 2001. – С. 162–163.
4. *Takach L*. Introduction to the theory of queues, Oxford University Press, Oxford, 1962.
5. *Клейнрок Л.* Теория массового обслуживания: Пер. с англ. – М.: Машиностроение, 1979. – 432 с.: ил.
6. *Вентцель Е.С., Овчаров Л.А*. Теория случайных процессов и её инженерные приложения. – М.: Наука, 1991.– 384 с.
7. *Гнеденко Б.В., Беляев Ю.К.. Соловьев А.Д.,* Математические методы в теории надежности: Основные характеристики надежности и их статистический анализ. – М.: Наука, 1965.
8. *Ложковский А.Г.* Статистическое моделирование полнодоступного пучка с потерями // Наукові праці ОНАЗ ім. О.С. Попова. – 2003. – С. 75-82.