УДК 681.324

Г.А. Кучук[1], А.А. Можаев[1], А.А. Коваленко[2]

[1]*Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков*
[2]*Харьковский национальный университет радиоэлектроники, Харьков*

ОБЗОР ПОДХОДОВ К ВЫБОРУ ВРЕМЕННЫХ ШКАЛ ПРИ ПРОВЕДЕНИИ АНАЛИЗА ОЧЕРЕДЕЙ

Представлены результаты, относящиеся к возможности выбора временных шкал для построения адекватных моделей современного трафика. Использование таких моделей, в частности, позволяет изучать динамику очередей активных сетевых устройств, что чрезвычайно важно для планирования и распределения загрузки сети. Использование статистических характеристик трафика на небольшом количестве временных масштабов позволяет расширить теоретические концепции для критических временных масштабов, что делает такой подход применимым к любому трафиковому процессу, включая трафик с долговременной зависимостью.

Ключевые слова: *протокол TCP, трафик, фрактальность, масштабная инвариантность, сетевые протоколы, модель, телекоммуникационная сеть.*

Введение

Постановка задачи и анализ литературы. Многочисленные исследования трафика современных высокоскоростных сетей передачи данных подтверждают наличие фрактальной природы или присутствия долговременной зависимости (ДВЗ) трафика [1 – 3] в этих сетях. Особенностями современного трафика также являются наличие последействия и масштабная инвариантность его статистических характеристик [1, 4 - 7]. Такой характер трафика в больших масштабах времени может привести к неравномерности загрузки существующих каналов сетей передачи данных, которая характеризуется или перегрузкой или недостаточной используемостью выделенных каналов.

Будем рассматривать очередь маршрутизатора как очередь бесконечной длины с постоянным коэффициентом обслуживания [8] для исследования вероятности того, что длина очереди Q превысит пороговое значение b, $P\{Q > b\}$ (вероятность возникновения хвоста очереди (ВХО)). ВХО является метрикой для различных приложений, включая методику поддержания маленьких задержек постановки пакетов в очередь и флуктуаций размеров очередей при управлении доступом и инициализации сети [9, 10]. Маленькие задержки в сети являются критичными для живучести приложений, использующих потоки данных реального времени, таких как телефония, телемедицина, видеоконференции, экономические транзакции и т.д.

До сих пор существующие подходы и модели не позволяли проводить всегда адекватное прогнозирование ситуации, когда $P\{Q > b\}$. Таким образом, разработка нового подхода к анализу процесса постановки данных в очередь, реализующего возможность прогнозирования трафикового процесса, непосредственно исходя из измеренных статистических характеристик трафика, является **актуальной**.

Целью данной статьи является исследование различных подходов к выбору временных шкал, использующихся при изучении организации очередей современных высокоскоростных сетей передачи данных.

Результаты теоретических исследований

Классические пуассоновские и марковские подходы к организации очередей неприменимы для ДВЗ трафика, для которого необходимо создание новых аналитических средств. До сих пор точных выражений для задержки постановки в очередь ДВЗ процессов, отличных от асимптотически больших задержек [11 – 13], не существовало, и, следовательно, существовала необходимость использования аппроксимаций.

Большинство аппроксимаций для вероятности хвоста очереди ДВЗ процессов основаны на понятии критической временной шкалы [11 – 18]. При заданном пороговом значении длины очереди b критической временной шкалой наиболее вероятно является количество времени, необходимое для заполнения очереди до значения, большего чем b. Поскольку критическая временная шкала является мощным теоретическим средством, ее вычисление непосредственно из эмпирических результатов является неосуществимым, вследствие необходимости наличия статистики трафика во всех временных шкалах.

Используя статистические характеристики трафика на конечном наборе временных масштабов θ, видится возможной разработка подхода, предоставляющего три практических аппроксимации для

$P\{Q > b\}$: аппроксимацию максимума, аппроксимацию произведения и аппроксимацию суммы. У всех них есть следующие важные свойства:

• применимость к любому конечному порогу очереди b, то есть, неасимптотичность;

• применимость к любой модели трафика, включая нестационарные;

• простота реализации, что обусловлено необходимостью знать статистические характеристики трафика только на нескольких масштабах времени θ.

А. Размер очереди как многошкальная функция

Рассмотрим непрерывную во времени текучую очередь с постоянной интенсивностью обслуживания c и трафиковым процессом $X_t, t \in R$ на входе. Обозначим через

$$K_t[\tau] := \int_{t-\tau}^{t} X_\omega d\omega \qquad (1)$$

трафиковый процесс с временной шкалой τ. Для идентификации трафикового процесса в обозначениях будем добавлять надстрочный индекс $K_t^{\{X\}}[\tau]$. Также для удобства обозначений опустим нижний индекс для всех инвариантных во времени величин. Положим, что очередь была пуста в некоторый момент времени ранее t. Тогда размер очереди Q_t равен разности между общим количеством трафика, прибывшего в очередь, и общим количеством трафика, обслуженным с того момента времени, когда очередь была пуста последний раз. Это кратко излагается формулой Рейча

$$Q_t := \sup_{\tau>0}\left(K_t[\tau] - c\tau\right), \qquad (2)$$

т.е. Q_t равно функции $K_t[\tau]$ трафикового процесса на всех временных шкалах τ.

Б. Аппроксимация критической шкалы времени

Большинство предложенных аппроксимаций $P\{Q > b\}$ для очередей ДВЗ трафика основаны на одной временной шкале, называемой критической временной шкалой [11 – 18]:

$$\lambda_t(b) := \arg \sup_{\tau>0} P\{K_t[\tau] - c\tau > b\}. \qquad (3)$$

Определим аппроксимацию критической временной шкалы как

$$\begin{aligned} C_t(b) &:= \sup_{\tau>0} P\{K_t[\tau] - c\tau > b\} = \\ &= P\{K_t[\lambda_t(b)] - c\lambda_t(b) > b\}. \end{aligned} \qquad (4)$$

Очевидно, что $C_t(b)$ является нижней границей $P\{Q_t > b\}$, поскольку, согласно выражению (2), $K_t[\lambda_t(b)] - c\lambda_t(b) \le Q_t$, следовательно

$$C_t(b) \le P\{Q_t > b\}. \qquad (5)$$

В работах, основанных на теории больших отклонений, было показано, что $C_t(b)$ имеет такое же лог-асимптотическое затухание, как и $P\{Q > b\}$ при $b \to \infty$ для обширного класса входных трафиковых процессов включая фракционное Броуновское движение [11, 12]. Таким образом, если $\{Q_t > b\}$ является редким событием и размер очереди превышает значение b, то это происходит в момент времени $\lambda_t(b)$, когда это наиболее правдоподобно. То есть, основываясь на $\{Q_t > b\}$, получим, что Q_t приблизительно равно $K[\lambda_t(b)] - c\lambda_t(b)$.

Хотя критическая временная шкала является мощным инструментом, имеющим внедрение в теории очередей, ее использование на практике не является непосредственным.

Во первых, рассмотрим задачу вычисления $C_t(b)$ для очереди случайного процесса, исключительно из эмпирических замеров трафика. Из выражения (4) видно, что требуется знать распределение $K_t[\tau]$ для всех возможных τ, что невозможно получить эмпирически. Даже если полностью заменить эмпирические схемы техникой, использующей и эмпирические статистики и аналитические модели, подобные вычислительные проблемы по-прежнему могут присутствовать. Например, если использовать модели трафика, для которых аналитические выражения для $C_t(b)$ неизвестны, то, возможно, для вычисления $C_t(b)$ необходимо будет применять сложные вычислительные алгоритмы.

Во-вторых, нужно вычислить аппроксимацию критической временной шкалы $C_t^{\{X+Y\}}(b)$, когда два независимых процесса X и Y мультиплексированы и попадают на вход очереди. Такой сценарий часто возникает при управлении доступом и сетевой инициализации [9, 10], а при вычислении непосредственно используются статистики процессов X и Y.

В. Многошкальные аппроксимации

Здесь приводятся три варианта аппроксимации организации очередей, не имеющие вычислительных сложностей, связанных с использованием аппроксимаций критических временных шкал. Ключевым фактором, упрощающим их вычисление, является то, что они используют статистические характеристики трафика только на фиксированном ограниченном наборе временных шкал $\theta \subset R_+$. Необходимо заметить, поскольку некоторые теоретические результаты касаются счетно бесконечных наборов θ, на практике всегда принимается укороченный конечный набор θ при вычислении аппроксимаций. Обычно набор θ выбирается для охвата диапазона временных шкал, на которых ожидается расположение критической временной шкалы $\lambda(b)$ для значений b, подходящих для конкретной аппроксимации.

Аппроксимация максимума. По аналогии с выражением для длины очереди и критической временной шкалы (согласно (2) и (3)), определим

$$Q_t^{[\theta]} := \sup_{\tau\in\theta}\left(K_t[\tau] - c\tau\right) \qquad (6)$$

и $$\lambda_t^{[\theta]} := \arg\sup_{\tau\in\theta} P\{K_t[\tau] - c\tau > b\} \quad (7)$$

для $\theta \subset R_+$, что приводит к следующей аппроксимации максимума

$$M_t^{[\theta]}(b) := \sup_{\tau\in\theta} P\{K_t[\tau] - c\tau > b\} = \\ = P\left\{K_t\left[\lambda_t^{\theta}(b)\right] - c\lambda_t^{\theta}(b) > b\right\}. \quad (8)$$

При сравнении выражений (4) и (8) видно, что аппроксимация максимума сходна с аппроксимацией критической временной шкалы. Разница заключается в том, что верхняя граница выбрана большей конечного набора в (8), вместо большей всех временных шкал в (4). Согласно выражениям (4), (5) и (8), получим границы

$$M_t^{[\theta]}(b) \le C_t(b) \le P\{Q_t > b\}. \quad (9)$$

Необходимо заметить, что, согласно выражениям (2) и (6),

$$Q_t = Q_t^{[R_+]} \ge Q_t^{[\theta]}. \quad (10)$$

и согласно (6), (8) и (10), получим

$$M_t^{[\theta]}(b) \le P\left\{Q_t^{[\theta]} > b\right\} \le P\{Q_t > b\}. \quad (11)$$

Аппроксимация максимума является практической заменой $C_t(b)$. Поскольку такая аппроксимация требует оценок $P\{K_t[\tau] - c\tau > b\}$ только для $\tau \in \theta$, сложностей, описанных ранее и связанных с вычислением $C_t(b)$, не возникает.

Рассмотрим задачу получения аппроксимации максимума из эмпирических замеров трафика. Достаточно просто вычислить гистограммы трафика на временных шкалах $\tau \in \theta$ и затем оценить $P\{K_t[\tau] - c\tau > b\}$.

Рассмотрим задачу вычисления аппроксимации максимума, когда два независимых процесса создают очередь. Простой операцией свертки распределений $K_t^{\{X\}}[t]$ и $K_t^{\{Y\}}[\tau]$ для $\tau \in \theta$ можно получить соответствующие распределения $K_t^{\{X+Y\}}[\tau]$, которые непосредственно дают аппроксимацию максимума.

Аппроксимации произведения и суммы. Две традиционных аппроксимации $P\{Q_t > b\}$, основанные на наборе временных шкал θ, являются аппроксимацией произведения

$$P_t^{[\theta]}(b) := 1 - \prod_{\tau\in\theta} P\{K_t[\tau] - c\tau < b\} \quad (12)$$

и аппроксимацией суммы

$$S_t^{[\theta]}(b) := \sum_{\tau\in\theta} P\{K_t[\tau] - c\tau > b\}. \quad (13)$$

Аппроксимация произведения равна $P\left\{Q_t^{[\theta]} > b\right\}$ когда события $\{K_t[\tau] - c\tau > b\}$, $\tau \in \theta$ независимы, а аппроксимация суммы равна $P\left\{Q_t^{[\theta]} > b\right\}$ когда они взаимоисключающие.

Точность аппроксимаций. Три вышеприведенные аппроксимации на практике наследуют точность аппроксимаций критической временной шкалы. Если существует элемент θ достаточно близкий к критической временной шкале, то $M_t^{[\theta]}(b)$ будет близко к $C_t(b)$ (согласно выражениям (4) и (8)). Кроме того, если единственно вероятный терм преобладает при суммировании в (13), то аппроксимации произведения и суммы будут близко аппроксимировать $M_t^{[\theta]}(b)$ и, следовательно, $C_t(b)$.

Выводы

В данной статье проведено исследование различных подходов к выбору временных шкал, использующихся при изучении организации очередей современных высокоскоростных сетей передачи данных. В дальнейшем планируется разработка моделей трафика наряду с разработкой аппроксимаций для вычисления вероятности возникновения хвоста очереди ДВЗ процессов.

Список литературы

1. Crovella M. Self-similarity in World Wide Web traffic: evidence and possible causes / M. Crovella, A. Bestavros // IEEE/ACM Transactions on Networking. – 1997. – Vol. 5. – P. 835-846.

2. Erramilli A. Experimental Queuing Analysis with Long-Range Dependent Traffic / A. Erramilli, O. Narayan, W. Willinger. // IEEE/ACM Transactions on Networking. – 1996. – № 7. – P. 135-141.

3. Willinger W. Self-Similarity Through High-Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level / W. Willinger, M.S. Taqqu, R. Sherman, D.V. Wilson. // ACM SIGCOMM'91. – 1991. – P. 149-157.

4. Leland W. On the self-similar nature of IP-trafic / W. Leland, M. Taqqu, W. Willinger // IEEE/ACM Transactions on Networking. – 1997. – № 3. – P. 423-431.

5. Фрактальный анализ процессов, структур и сигналов: Коллективная монография / Г.А. Кучук, А.А. Можаев, Р.Э. Пащенко и др. – Х.: ЭкоПерспектива, 2006. – 360 с.

6. Воробьев О.В. Моделювання самоподібного трафіку синтезом ансамблю стохастичних квазіперіодичних джерел та ON/OFF моделі / О.В. Воробьев // Системи озброєння і військова техніка .– 2006. – 3 (6). – С.97-105.

7. Кучук Г.А., Можаєв О.О., Воробйов О.В. Аналіз та моделі самоподібного трафіка // Авиационно-космическая техника и технология. – 2006. – № 9 (35). – С. 173-180.

8. Papagiannaki K. Analysis of measured single-hop delay from an operational backbone network / K. Papagiannaki, S. Moon, C. Fraleigh, P. Thiran, F. TobagiC. Diot // Proc. IEEE INFOCOM. – Jun. 2002. – P. 535-544.

9. Breslau L. Comments on the performance of measurement-based admission control / L. Breslau, S. Jamin, S. Shenker // Proc. IEEE INFOCOM. – Mar. 2000. – P. 1233-1242.

10. Fraleigh C. Provisioning IP backbone networks to support latency sensitive traffic / C. Fraleigh, F. Tobagi, C. Diot // Proc. IEEE INFOCOM. – Apr. 2003. – P. 375-385.

11. Кучук Г.А. Побудова черги при самоподібному трафіку / Г.А. Кучук // Системи обробки інформації. – Х.: ХВУ. – 2004. – Вип. 6. – С. 134 – 137

12. Кучук Г.А. Метод дослідження фрак-тального мережного трафіка / Г.А. Кучук // Системи обробки інформації. – Х.: ХУ ПС, 2005. – Вип. 5(45). – С. 74 – 84.

13. Кучук Г.А. Метод визначення моментів зміни характеристик трафікового процесу / Г.А. Кучук // Системи озброєння і військова техніка. – 2008. – № 4(16). – С. 151-155.

14. Кучук Г.А. Модель процесу буферізації пакетів в апаратних засобах мультиплексування / Г.А. Кучук // Системи обробки інформації. – Х.: ХУ ПС, 2006. – Вип. 2(51). – С. 65-70.

15. Grossglauser M. On the relevance of long-range dependence in network traffic / M. Grossglauser, J.-C. Bolot // Comput. Commun. Rev. – Oct. 1996. – Vol. 26, no. 4. – P. 15-24.

16. Choe J. Queueing analysis of high-speed multiplexers including long-range dependent arrival processes / J. Choe, N.B. Shroff // Proc. IEEE INFOCOM. – Mar. 1999. – P. 617-624.

17. Кучук Г.А. Конвергенція різнорідного трафіка в інформаційно-телекомунікаційних мережах / Г.А. Кучук, Ю.В. Паржин, І.І.Сидоренко // Системи обробки інформації. – Х.: ХУ ПС, 2008. – Вип. 6 (73). – С. 75 – 77.

18. Debicki K. A note on transient Gaussian fluid models / K. Debicki, T. Rolski // Queueing Syst. – 2002. – Vol. 41. – P. 321-342.

Поступила в редколлегию 12.12.2008

Рецензент: д-р техн. наук, проф. Ю.В. Стасев, Харьковский университет Воздушных Сил им. И. Кожедуба, Харьков.

ОГЛЯД ПІДХОДІВ ДО ВИБОРУ ЧАСОВИХ ШКАЛ ПРИ ПРОВЕДЕННІ АНАЛІЗУ ЧЕРГ

Г.А. Кучук, О.О. Можаєв, А.А. Коваленко

Представлені результати, що відносяться до можливості вибору часових шкал для побудови адекватних моделей сучасного трафіку. Використання таких моделей, зокрема, дозволяє вивчати динаміку черг активних мережевих пристроїв, що надзвичайно важливе для планування і розподілу завантаження мережі. Використання статистичних характеристик трафіку на невеликій кількості часових масштабів дозволяє розширити теоретичні концепції для критичних часових масштабів, що робить такий підхід застосовним до будь-якого трафіковому процесу, включаючи трафік з довготривалою залежністю.

Ключові слова: *протокол TCP, трафік, фрактальність, масштабна інваріантність, мережеві протоколи, модель, телекомунікаційна мережа.*

REVIEW OF APPROACHES TO CHOICE OF TEMPORAL SCALES DURING CONDUCTING OF QUEUING ANALYSIS

G.A. Kuchuk, A.A. Mogaev, A.A. Kovalenko

Results, related to possibility of choice of temporal scales for the construction of adequate models of modern traffic, are presented. The use of such models, in particular, allows to study the dynamics of turns of active networkings issues, that is extraordinarily important for planning and distributing of load of network. The use of statistical descriptions of traffic on the two-bit of temporal scales allows to extend theoretical conceptions for critical temporal scales, that does such approach applicable to any traffic process, including a traffic with of long duration dependence.

Keywords*: protocol of TCP, traffic, fractal, scale invariance, protocols of networks, model, telecommunication network.*