УДК 621.398 В555

Халимон Н. Ф., к.т.н. *(Национальный авиационный университет)*

СТАТИСТИЧЕСКАЯ ДИНАМИКА ПРОЦЕССА ЗАПРОСОВ С ПЕРЕКЛЮЧЕНИЯМИ В СЕТЯХ ХРАНЕНИЯ ДАННЫХ

Халімон Н. Ф. Статистична динаміка процесу запитів з перемиканнями в мережахзберігання даних. У даній роботі проведений аналіз ефективності системи поллінгу, яку доцільно використовувати для обслуговування мереж зберігання даних. За наслідками аналізу процесів опиту в системі поллінгу встановлено, що вибір найприйнятнішого порядку опиту елементів мережі зберігання залежить від об'єму запрошуваного пакету даних, тобто від довжини черги запитів. Іншими словами, цей вибір визначається видом розподілу даних по окремих елементах сховища даних.

Як інформативні параметри для вибору порядку опитування і інтенсивності обслуговування конкретної черги запропоновано використовувати комплексний показник – функцію довжини черги і швидкості її зростання або убування.

Ключові слова: мережа зберігання даних, система поллінгу, черга, сховище даних, вкладений ланцюг Маркова, динаміка зміни довжини черги

Халимон Н. Ф. Статистическая динамика процесса запросов с переключениями в сетях хранения данных. В данной работе проведен анализ эффективности системы поллинга, которую целесообразно использовать для обслуживания сетей хранения данных. По результатам анализа процессов опроса в системе поллинга установлено, что выбор наиболее приемлемого порядка опроса элементов сети хранения зависит от объема запрашиваемого пакета данных, т.е. от длины очереди запросов. Другими словами, этот выбор определяется видом распределения данных по отдельным элементам хранилища данных.

В качестве информативных параметров для выбора порядка опроса и интенсивности обслуживания конкретной очереди предложено использовать комплексный показатель – функцию длины очереди и скорости ее роста или убывания.

Ключевые слова: сеть хранения данных, система поллинга, очередь, хранилище данных, вложенная цепь Маркова, динамика изменения длины очереди

Khalimoh N. F. Statistical dynamics of queries process with switchingin in the data storage networks. The analysis of efficiency of the polling system, which it is expedient to use for maintenance of networks of data storage, is conducted in this work. On results the analysis of processes of questioning it is set in the polling system that the choice of the most acceptable order of questioning of elements of network of storage depends on the volume of the inquired package of information, I.e. from length of turn of queries. In other words, this choice is determined by the type of distributing of information on the separate elements of information depository.

As informing parameters for the choice of order of questioning and intensity of maintenance of concrete turn it is suggested to use a complex index – function of length of turn and speed of its growth or decrease.

Keywords: network of data storage, polling system, queue, information depository, nested Markov chain, dynamics of the queue length change

I. Введение. Сети хранения данных предназначены для управления большими объемами данных для большого количества пользователей. Однако с ростом объемов данных возникает проблема оптимизации скорости обработки запросов и выдачи запрашиваемых данных. Наиболее перспективный путь решения этой задачи – применение статистических методов, в частности, методов теории массового обслуживания.

Для оценки эффективности функционирования сетей хранения данных как систем массового обслуживания предлагается применять стохастические модели поллинга или упорядоченного опроса [1,2]. Порядок опроса очередей определяется правилом выбора сервером следующей очереди. Наиболее распространенные виды порядка опроса [2, 3]:

- циклический – установлена последовательность прохода очередей;
- периодический – опрос осуществляется на основе таблицы поллинга;
- случайный;
- приоритетный.

В работе [4] отмечается, что с практической точки зрения значительный интерес представляют системы с адаптивным механизмом поллинга, в которых при опросе очередей учитывается состояние очереди, то есть если очередь была пуста в данном цикле опроса, то на следующем цикле опроса эта очередь будет пропущена. Однако такой подход можно применять только в системах поллинга, в которых поток заявок представляет собой стационарный эргодический процесс [3, 5]. Следовательно, необходимо учитывать нестационарности текущей величины конкретной очереди. Поэтому на практике целесообразно применять механизм поллинга с обратной связью и адаптацией не только к текущему размеру очереди, но и к скорости роста (убывания) очереди. Такие задачи в доступных нам источниках не встречаются ни в общем плане, ни, в частности, применительно к системам поллинга. В данной работе сделана попытка восполнить этот пробел.

II. Математическое описание системы поллинга. Модель системы поллинга для сети хранения описывается следующим образом. Система имеет один обслуживающий прибор и N $(N \geq 2)$ очередей. Каждый из N буферов имеет ограниченный объем памяти в L ячеек. Заявки поступают в общем нестационарном входном потоке. В k-ю очередь поступает нестационарный поток заявок с функцией распределения $f_k(t)$ и мгновенной интенсивностью $\lambda_k(t)$. Максимальное число заявок на интервале наблюдения T_s равно M_k, причем $L > M_k$, $k = \overline{1, N}$. При опросе k-го элемента сети хранения обслуживается $f_k(n) \leq M_k$ заявок.

Считаем, что времена обслуживания τ_{sk} заявок в очереди независимы и одинаково распределены с функцией распределения $w_k(t)$, которая является непрерывной и дифференцируемой, с математическим ожиданием

$$m_k = \int_0^\infty t dw_k(t)$$

и вторым начальным моментом

$$\sigma_k^2 = \int_0^\infty t^2 dw_k(t).$$

Интеграл понимается в смысле Стилтьеса. Предполагается, что потоки заявок и длительности обслуживания заявок представляют собой взаимно независимые процессы. Сервер посещает очереди, следуя выбранному порядку опроса и обслуживая их в соответствии с выбранной дисциплиной. Время подключения сервера к k-й очереди τ_{qk} –

случайная величина с плотностью распределения $\nu_{qk}(t)$, математическим ожиданием m_{qk} и вторым начальным моментом σ_{qk}^2. Текущая длина k-й очереди составляет m_k запросов, объем буфера (число ячеек) равен L. Время обслуживания t_s – случайная величина с математическим ожиданием m_{ts} и дисперсией σ_{ts}^2. За время t_s в k-й очереди обслуживается часть запросов длиной δ_{kl}.

После обслуживания k-й очереди сервер переключается на обслуживание i-й очереди из общего числа N очередей. Время переключения с k-й на i-ю очередь обозначим t_{ki}.

Введем дополнительные обозначения:

$T_{est} << T_s$ – интервал оценивания параметров, длительность которого значительно меньше длительности интервала наблюдения T_s;

$\lambda(t)$ – мгновенная интенсивность на интервале оценивания;

p_k – вероятность того, что вызов из входящего потока будет направлен в k-ю очередь;

$\Delta_{kl} = \lim\limits_{l_k \to L} \delta_{kl}$ – часть обслуженных запросов при максимально возможной загрузке очереди; N_k – среднее число посещений k-й очереди за один цикл опроса;

$T_{k\Sigma} = \sum\limits_{\substack{k=1 \\ i=1 \\ k \neq i}}^{N} t_{ki}$ – среднее суммарное время переключений с k-й очереди на i-ю очередь за цикл опроса.

Маршрут обслуживания в системе поллинга с нестационарным потоком заявок с функцией распределения $f_i(t)$ и мгновенной интенсивностью $\lambda_i(t)$ представляет собой вложенную цепь Маркова [6], состояния которой s_{ki} характеризуются переходами от k-й очереди к i-й очереди с вероятностями π_{ki}, $i,k = \overline{1,N}$, $i \neq k$. Если на интервале наблюдения T_s выполняется условие

$$\lambda_k(t)\left(m_{ts} + \sum_{k=1}^{N} \frac{p_k}{\Delta_{kl}\pi_{ki}}|t_{ki}|\right) < 1, \qquad \textbf{(1)}$$

(где $|t_{ki}|$ – математическое ожидание времени переключения k-й очереди к i-й очереди), то средняя длина очереди на интервале наблюдения T_s будет принимать некоторое конечное, хотя и нестационарное значение (напомним, что справедливо условие $T_{est} << T_s$).

Если же условие (1) не выполняется:

$$\lambda_k(t)\left(m_{ts} + \sum_{k=1}^{N} \frac{p_k}{\Delta_{kl}\pi_{ki}}|t_{ki}|\right) \geq 1, \qquad \textbf{(2)}$$

то вероятность неограниченного роста общей длины очереди $M_\Sigma = \sum\limits_{k=1}^{N} M_k$ будет стремиться к единице (при этом некоторые из очередей могут иметь постоянную длину или даже уменьшаться).

III. Система поллинга с динамическим регулированием длины очереди. Системы поллинга как системы упорядоченного опроса, по существу, представляют собой специальные системы массового обслуживания с приоритетами. Однако для назначения приоритета запросу (или группе запросов) необходимо учитывать не только класс запроса, но и среднее время пребывания запросов в системе. Для нахождения среднего времени ожидания чаще всего используется взвешенная сумма средних времен ожидания $T_{\Sigma w}$, под которой будем понимать величину

$$T_{\Sigma w} = \sum_{k=1}^{N} \rho_k \hat{t}_k \ , \tag{3}$$

где $\rho_k = \lambda_k(t) \cdot t_{sk}$;

t_{sk} – среднее время обслуживания запросов в k-й очереди;

$\hat{t}_k = T_{k\Sigma} + (N-1)t_{sk}$ – среднее время ожидания в k-й очереди, равное сумме средних времен переключений $T_{k\Sigma}$ и суммарному среднему времени обслуживания запросов в других очередях, равному $(N-1)t_{sk}$.

Здесь предполагается, что по умолчанию запросы обслуживаются в соответствии с дисциплиной «первый пришел – первый обслужен» (*FCFS – First Come*, *First Served*).

Без потери общности можно считать, что процессы формирования очередей в системе и процессы обслуживания являются стохастически эквивалентными, т.е. имеющими одинаковые характеристики, стационарные или нестационарные. В этом смысле система поллинга является статистически однородной.

Пусть мгновенная длительность k-й очереди равна $m_k(t)$ для непрерывной системы или $m_k(n)$, $n = 0,1,2,\ldots$ для дискретной системы поллинга. Соответственно, мгновенная скорость изменения длины очереди есть производная $\dot{m}_k(t)$ длины очереди для непрерывной системы. Для дискретной системы получаем конечную разность

$$\Delta m_k = m_k(n) - m_k(n-1).$$

В дальнейшем будем рассматривать дискретные системы поллинга.

Рассмотрим задачу оптимизации среднего времени обслуживания запросов в сети хранения. Возьмем в качестве весовых коэффициентов эффективности c_k произведения средней длины k-й очереди на скорость ее роста:

$$c_k = \mathrm{E}\left[m_k(n)\right]\mathrm{E}\left(\Delta m_k\right), \tag{4}$$

где E – символ математического ожидания.

Таким образом, мы приходим к задаче минимизации функционала для аддитивной меры множества параметров системы поллинга:

$$\Psi\left(\mathrm{T}_{\Sigma w}\right) = \Psi\left[N, L, \lambda_k, \varphi(\rho_k)\right] \to \min_{\mathrm{V}_\Psi} \sum_{k=1}^{N} c_k t_{sk} \ , \tag{5}$$

где $\mathbf{V}_{\Psi}^{\mathrm{T}} = \left[N, L, \lambda_k, \varphi(\rho_k) \right]$ – вектор параметров, по которым минимизируется функционал Ψ;

Т – символ транспонирования.

Таким образом, минимизация функционала (5) сводится к нелинейной свертке критериев, в качестве которых используются весовые коэффициенты c_k. Частота обращений к той или иной очереди зависит от величин произведений (4).

IV. Выводы. По результатам анализа процессов опроса в системе поллинга установлено, что выбор наиболее приемлемого порядка опроса элементов сети хранения зависит от объема запрашиваемого пакета данных. В свою очередь, объемом запрашиваемого пакета определяется длина очереди запросов. Другими словами, этот выбор определяется видом распределения данных по отдельным элементам хранилища данных.

В качестве информативных параметров для выбора порядка опроса и интенсивности обслуживания конкретной очереди предложено использовать комплексный показатель длины очереди и скорости ее роста или убывания.

В дальнейшем планируется, исходя из вида зависимости переходных вероятностей от времени, будем правило опроса вида “*m* из *n*” – за *n* циклов опроса производится m_i опросов *i*-й очереди – в зависимости от скорости роста *i*-й очереди.

Литература

1. Вишневский В. М. Системы поллинга: теория и применение в широкополосных беспроводных сетях / В. М. Вишневский, О. В. Семенова. – М.: Техносфера, 2007. – 312 с.

2. Халимон Н. Ф. Асимптотические характеристики процессов обработки запросов к сетям хранения данных на основе методов поллинга / Н. Ф. Халимон // Вісник Державного університету інформаційно-комунікаційних технологій. – 2012. – Т.10, №4. – С. 87-91.

3. Borovkov A., Schassberger R. Ergodicity of a polling network // Stoch. Proc. Appl., 1994. – V.50. – PP. 253-262.

4. Синюгина Ю. В. О времени ожидания в системе с ограниченным шлюзовым обслуживанием и адаптивными отдыхами / Ю. В. Синюгина // Проблемы физики, математики и техники. – 2012. – № 4 (13). – С. 61-65.

5. Фосс С. Г. Теоремы сравнения и эргодические свойства систем поллинга / С. Г. Фосс, Н. И. Чернова // Проблемы передачи информации. – 1996. – Т.32, Вып. 4. – С. 46-71.

6. Гнеденко Б.В. Введение в теорию массового обслуживания / Б. В. Гнеденко, И. Н. Коваленко. – М.: Наука, 1987. – 336 с.