УДК 621.395

Е.Г. Игнатенко, В.И. Бессараб, И.В. Дегтяренко
Донецкий национальный технический университет, г. Донецк
кафедра автоматики и телекоммуникаций
E-mail: kovalenko_evg@mail.ru

АДАПТИВНЫЙ АЛГОРИТМ МОНИТОРИНГА ЗАГРУЖЕННОСТИ СЕРВЕРОВ WEB-КЛАСТЕРА В СИСТЕМЕ БАЛАНСИРОВКИ НАГРУЗКИ

Аннотация
Игнатенко Е.Г., Бессараб В.И., Дегтяренко И.В. Адаптивный алгоритм мониторинга загруженности серверов web-кластера в системе балансировки нагрузки. *Проведен анализ проблемы мониторинга загруженности серверов. Усовершенствован метод определения пачечности во входящем потоке. Предложен новый адаптивный алгоритм мониторинга состояния серверов web-кластера, позволяющий сократить количество служебной информации.*
Ключевые слова: *система балансировки нагрузки, мониторинг, самоподобие, пачечность.*

Общая постановка проблемы. В связи с интенсивным развитием кластерных web-серверов, усложнением их структуры и увеличением количества используемых приложений, а также необходимостью обеспечения высокого качества обслуживания и доступности, требуется производить постоянный мониторинг состояния кластера. Одна из задач мониторинга — сбор статистической информации о загруженности web серверов. Система мониторинга предназначена для измерения и регистрации основных состояний системы, в том числе перегрузок [1]. Мониторинг позволяет своевременно принимать необходимые меры в условиях высокой интенсивности входящей нагрузки, анализировать и прогнозировать состояние системы. Основной из существующих проблем мониторинга является его точность. При попытке достижения высокой точности, служебная информация приобретает избыточный характер[2].

Оценка загрузки сервера может быть произведена несколькими способами. Одним из способов, который обычно используют при статической балансировке загрузки, состоит в приблизительной оценке загрузки каждого объекта на основе знаний о приложении в целом [1]. Вместе с тем он может быть неточным в случае, если априорная аналитическая модель для оценки скорости выполнения приложений неточна [2].

Другой известный способ сбора данных о загрузке состоит в текущем измерении загрузки процессоров и задач. Достоинство метода состоит в том, что он является более точным. К недостаткам можно отнести следующее: алгоритмы балансировки, основанные на этом методе, учитывают прошлое распределение нагрузок, т.е. если загрузка меняется случайным образом, то метод является не достаточно точным.

Постановка цели исследования. Целью данной статьи является разработка адаптивного алгоритма мониторинга загруженности серверов информационного кластера, учитывающего самоподобную структуру трафика web приложений и снижающего количество служебной информации.

Решение задач и результаты исследований. Информационный кластер представлен группой серверов и системой балансировки, осуществляющей распределение запросов на основе информации мониторинга состояния серверов.

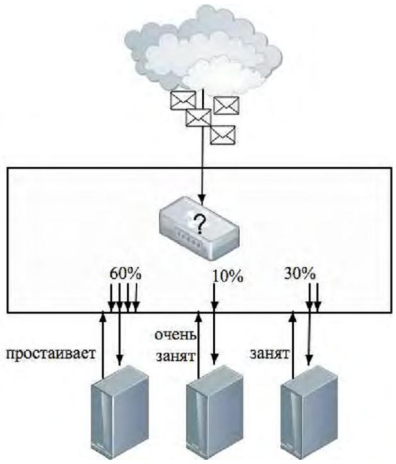

Рисунок 1 – Структурная схема кластера

Полученная информация о загрузке используется в качестве базы данных для процесса балансировки, во-первых, для определения возникновения дисбаланса, во-вторых, для определения нового распределения запросов, путем вычисления объема работ, необходимого для обработки запросов. Математически задачу БН можно представить следующей зависимостью:

$$Q = f(N, G, \Lambda), \tag{1}$$

где N – множество серверов кластера $N = \{n_i\}$;

G – множество характеристик серверов $G = \{P_i, S_i\}$;

P_i - производительность сервера;

S_i - текущая загруженность сервера;

Λ - множество характеристик входящего потока запросов $\Lambda = \{\mu, \lambda\}$, μ - вычислительная сложность запроса, λ - интенсивность входящего потока.

Мониторинг входящего потока запросов в систему позволит отслеживать резкие изменения в нагрузке на кластер. Интенсивность поступления запросов – является главным параметром мониторинга, т.к. позволяет избегать перегрузок на web кластере. Кроме интенсивности входящего потока проводится мониторинг показателей загруженности серверов таких как, загруженность процессора, время простоя процессора, фоновая загрузка процессора, размер свободной памяти, число операций ввода-вывода и т.д. Таким образом, загрузку сервера (U) можно представить в следующем виде:

$$U = f(I_1, I_2, ...I_m), \tag{2}$$

где $I_1, I_2, ...I_m$ - показатели загруженности сервера.

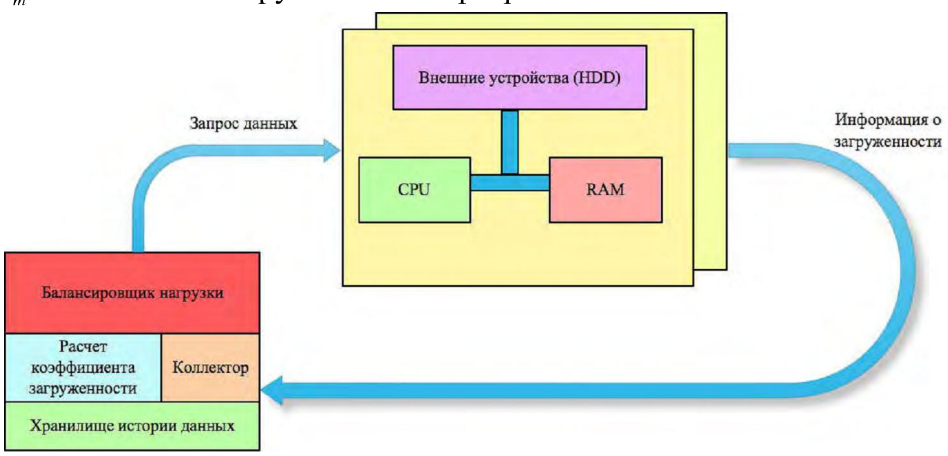

Рисунок 2 – Мониторинг загруженности серверов

Предлагаемый мониторинг можно осуществить следующими способами:

- после каждого поступившего запроса;
- в фиксированные промежутки времени, определяемые статическим алгоритмом;
- в нефиксированные промежутки времени, определяемые динамическим алгоритмом.

Служебная информация, полученная 1 способом, является наибольшей по объёму, т.к. измерения проводятся после каждого поступившего запроса. При 2 способе мониторинга служебная информация постоянна, в то время как при 3 способе - она зависит от частоты интервалов контроля. Недостатком 2 метода является выбор интервала. Достаточно сложно определить временной интервал, который смог бы приспосабливаться к потоку поступающих запросов, учитывая их самоподобную структуру.

На основании вышесказанного, использование динамически меняющегося интервала, является наиболее приемлемым с точки зрения уменьшения избыточности данных. При таком способе частота мониторинга будет зависеть от количества всплесков (пачечности) входящего потока. Наличие пачек в web-трафике является одной из его особенностей [3,4,5]. Интервал мониторинга должен сокращаться, если во входящем потоке обнаружен всплеск, и увеличиваться при наблюдении уменьшения интенсивности.

Для определения величины значимого всплеска возможно применение двух методов: первый метод предложен в [7], второй метод – предлагаемая модификация первого. Сравним их на примере. Для этого используем экспериментальную статистику с web-сервера, полученную на некотором интервале наблюдений [6]. Данная статистика представляет собой последовательность поступления http-запросов. Статистика обработана, путем агрегирования по 1 секунде (см.рис.3). Представленные данные свидетельствуют о наличии неравномерности интенсивности поступления запросов. Запросы не плавно рассредоточены по различным интервалам времени и группируются в «пачки». Из-за этого в пачечном трафике, при сравнительно небольшом среднем значении интенсивности поступления запросов, присутствуют относительно большие выбросы.

Для определения пачечности первым методом необходимо знать среднюю интенсивность поступления http-запросов. Этот способ предполагает нахождение 2 параметров, описывающих пачечность – a и b [7]:

- a – отношение пиковой интенсивности процесса поступления заявок на обслуживание к его среднему значению в наблюдаемом интервале;
- b – доля времени, в течении которого мгновенная интенсивность поступления запросов превышает среднюю интенсивность (значение параметра пачечности b заключено в интервале (0;1]).

Алгоритм определения параметров a и b состоит в следующем:

- подсчитывается количество HTTP запросов – L, попавших в интервал длиною τ;
- скорость поступления запросов λ в интервал τ определяется как $\lambda = \frac{L}{\tau}$;
- интервал τ разбивается на n равных подинтервалов, длиною $k = \frac{\tau}{n}$;
- определяется arr(k) - количество HTTP запросов, попавших в интервал k;
- находится λ_k - скорость поступления запросов за интервал длиною k, $\lambda_k = \frac{arr(k)}{k}$;
- вычисляется arr* - общее количество HTTP запросов в интервалах, удовлетворяющих условию $\lambda_k > \lambda$;

- определяется g – количество интервалов, удовлетворяющих условию $\lambda_k > \lambda$;
- рассчитываются параметры пачечности:

$$a = \frac{\lambda^*}{\lambda} = \frac{arr^*}{b \cdot \tau \cdot \lambda}, \tag{3}$$

$$b = \frac{g}{n}. \tag{4}$$

Полученный параметр b предупреждает систему балансировки о пачках входящего потока. Между величиной пачечности и длительностью интервала мониторинга существует некоторая функциональная связь.

Рассмотрим весь интервал наблюдений T. Разделим его на несколько слотов различной длины. Длительность каждого слота d изменяется в зависимости от значений пачечности, т.е. длительность d(k+1) слота зависит от размера пачечности двух предыдущих слотов b(k) и b(k-1):

$$d(k+1) = \frac{b(k-1)}{b(k)} \cdot d(k). \tag{5}$$

Таким образом, длительность следующего слота уменьшается, если был обнаружен всплеск, т.е. b(k)>b(k-1). И наоборот - увеличивается, d в случае b(k)<b(k-1).

На графиках отмечены моменты времени, в которые наблюдались пики интенсивности входящего потока. Как видно из рисунка 3, первый способ предоставляет эту информацию с некоторой задержкой.

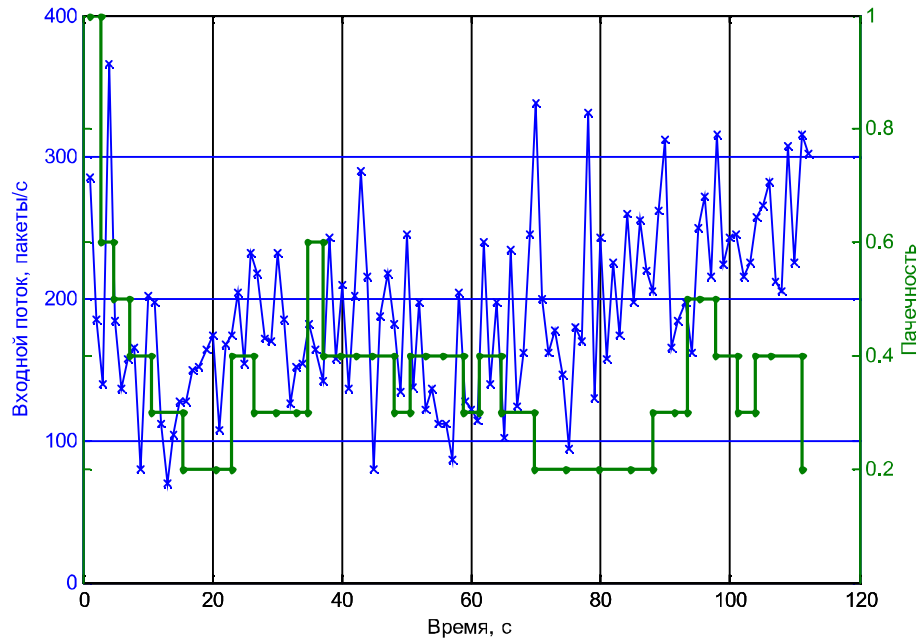

Рисунок 3 – Интенсивность входящего потока и пачечность, определенная 1-м способом

Кривая всплесков сглаживает кривую поступления запросов. На рисунке 3 показано, что кривая всплесков соответствует кривой поступления, но при этом она не описывает резкие изменения характера кривой входящего потока.

Второй способ определения пачечности является его модификацией и включает относительную разность входящих потоков за два предыдущих слота. Таким образом, размер пачечности зависит от того, на сколько увеличился или уменьшился входной поток:

$$b_m = \frac{g}{n} \cdot \left(1 + \frac{\lambda(k) - \lambda(k-1)}{\lambda(k-1)}\right). \quad (6)$$

Следует отметить, что при вычислении параметра пачечности bm по формуле (6), значение параметра может превышать 1. Для алгоритма управления необходимо ограничить значение пачечности в переделах (0;1].

На рисунке 4 показано как кривая пачечности, с использованием модифицированного способа, повторяет изменения во входящем потоке:

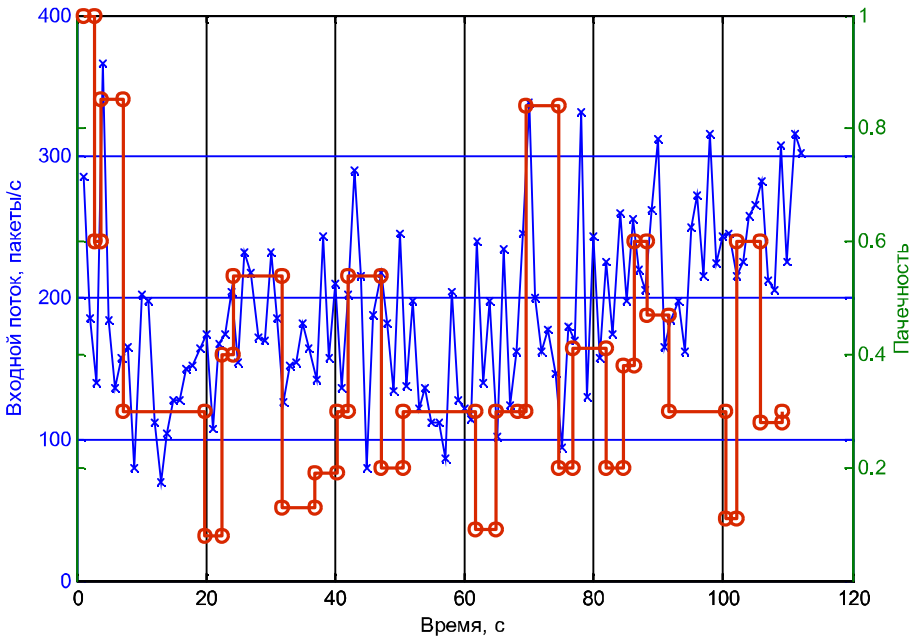

Рисунок 4 – Интенсивность входящего потока и пачечность, определенная модифицированным способом

На рисунке 5 приведена сравнительная характеристика двух способов определения пачечности:

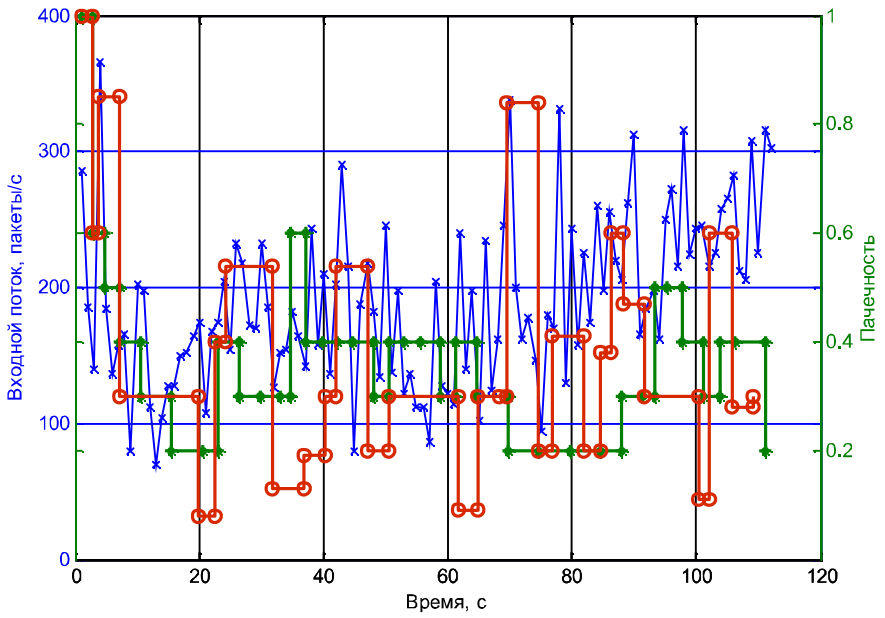

Рисунок 5 – Сравнительная характеристика способов определения пачечности

На рисунке 6, на некотором интервале, показано изменение шага мониторинга в зависимости от изменения интенсивности входящего потока. Например, изменение пачечности с 0,3 до 0,5 свидетельствует об увеличении интенсивности входящего потока, тем самым интервал наблюдения уменьшается со значения d1=3,33 до d2=1,85 секунд.

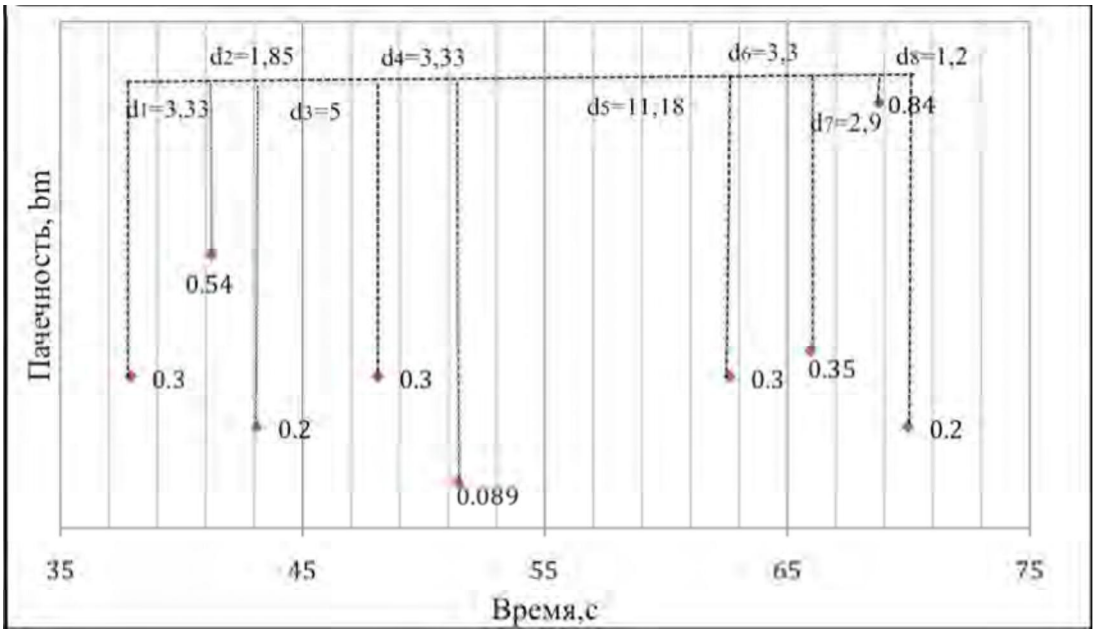

Рисунок 6 – Изменение шага мониторинга в зависимости от пачечности

Сравнив два способа определения пачечности, можно сделать вывод об эффективности применения модифицированного способа, т.к. кривая bm точнее описывает изменения во входящем потоке. Кроме того, количество интервалов наблюдения при 1-м способе составляет - 36, при модифицированном - 28, что сокращает объем служебной информации.

При сравнении разработанного адаптивного алгоритма мониторинга с плавающим шагом и алгоритма с постоянным дискретным шагом t=2 с получены следующие результаты:

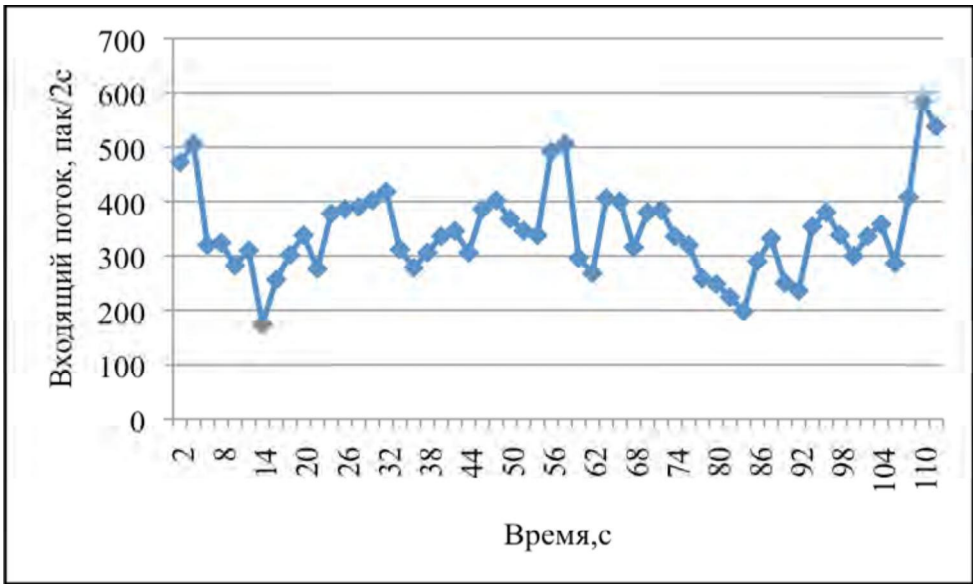

Рисунок 7 – Мониторинг с дискретным постоянным шагом мониторинга t=2с

Мониторинг с постоянным дискретным шагом сглаживает кривую входящего потока (см.рис.3), не отображая при этом всех пиков нагрузки. Для мониторинга 112 секунд входящего потока с использованием такого метода понадобится 56 отчетов, что в 2 раза превышает количество отчетов при адаптивном алгоритме мониторинга.

На рисунке 8 приведена блок-схема адаптивного алгоритма мониторинга загруженности серверов web-кластера:

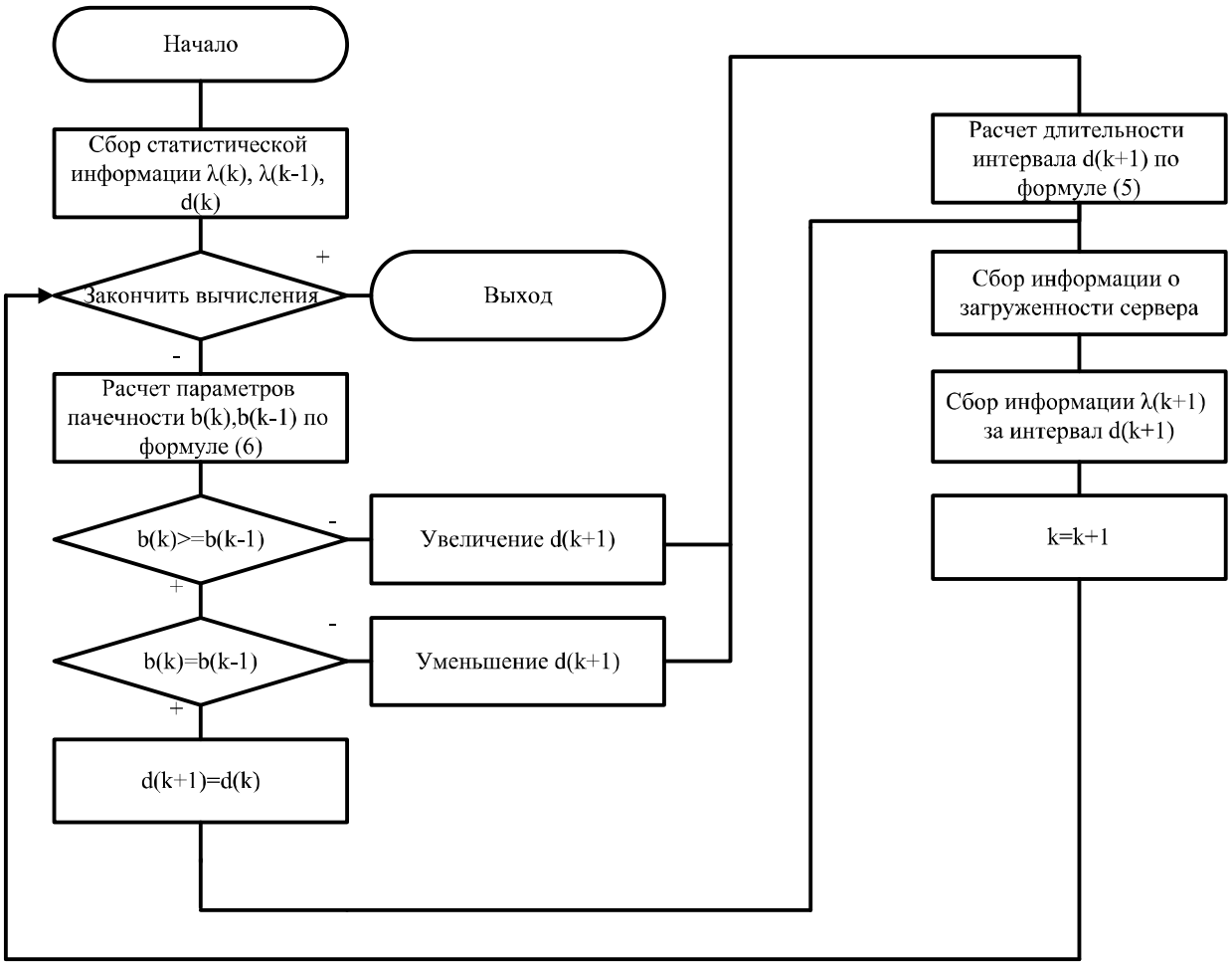

Рисунок 8 – Блок-схема адаптивного алгоритма мониторинга

Выводы. В статье проведен анализ проблемы мониторинга загруженности серверов, рассмотрены основные параметры загруженности и способы их мониторинга. Усовершенствован метод определения пачечности во входящем потоке. На основании модифицированного метода предложен новый адаптивный алгоритм мониторинга состояния серверов web-кластера. Алгоритм позволяет сократить количество служебной информации. При этом алгоритм описывает изменения входящего потока с достаточной точностью, что позволяет системе балансировки избегать перегрузок в web-кластере.

Проведено моделирования разработанного алгоритма. Полученные результаты показывают преимущества использования в алгоритме модифицированного метода определения пачечности, т.к. при данном методе кривая пачечности наиболее точно показывает изменения во входящем потоке при меньшем количестве точек мониторинга.

Литература

1. Электронный ресурс. Способ доступа: http://www.intuit.ru/department/algorithms/distrsa/13/.
2. Борисенко Н.П. Модель контроля и управления производительностью web-сервера /Н.П. Борисенко, Д.А. Васинев, Д.Л. Жусов // Высокие технологии, фундаментальные и прикладные исследования, образование: сб. Второй международной науч. - практ. конф. ["Исследование, разработка и применение высоких технологий в промышленности"]. – С-Пб.: Политехн. ун-та, 2006. Т.7. – С. 69-71.
3. Banga G. and P. Druschel, "Measuring the Capacity of a Web server," USENIX Symposium on Internet Technology and Systems, Dec. 1997.
4. Ложковский А.Г. Оценка параметров качества обслуживания самоподобного трафика энтропийным методом / А.Г. Ложковский, Р.А. Ганифаев // Наукові праці ОНАЗ ім. О.С. Попова. – 2008. – No 1. – С.57–62.
5. Kun-Chan and John Heidemann. A measurement study of correlations of Internet flow characteristics. Computer Networks, 50(1):46-62, January 2006.
6. Бессараб В.И. Генератор самоподобного трафика для моделей информационных сетей / В.И. Бессараб, Е.Г. Игнатенко, В.В. Червинский // Вісник Східноукраїнського Національного Университету ім. Володимира Даля. – 2010. – № 2(144).
7. Menascé D. A. and V. A. F. Almeida, Capacity Planning for Web Performance: metrics, models, and methods, Prentice Hall, Upper Saddle River, 1998.

Надійшла до редакції:
18.02.2011

Рекомендовано до друку:
д-р техн. наук, проф. Скобцов Ю.О.

Abstract

Ignatenko E.G., Bessarab V.I., Degtyarenko I.V. Adaptive algorithm for servers utilization monitoring within load-balanced web-cluster. *Servers utilization monitoring problem was analyzed. The incoming stream burstisness identification method was improved. New adaptive algorithm for web-cluster servers utilization monitoring was proposed. It allows to decrease the amount of overhead data.*

Keywords: *load balancing system, monitoring, self-similarity, burstiness.*

Анотація

Ігнатенко Є.Г., Бессараб В.І., Дегтяренко І.В. Адаптивний алгоритм моніторингу завантаженості серверів web-кластеру в системі балансування навантаження. *Проведено аналіз проблеми моніторингу завантаженості серверів. Удосконалено метод визначення пачечності у вхідному потоці. Запропоновано новий адаптивний алгоритм моніторингу стану серверів web-кластеру, який дозволяє скоротити кількість службової інформації.*

Ключові слова: *система балансування навантаження, моніторинг, самоподібність, пачечність.*