УДК 004.7:621.395:338.46
DOI 10.17513/snt.39947

АЛГОРИТМ БАЛАНСИРОВКИ НАГРУЗКИ ЦЕНТРА ОБРАБОТКИ ДАННЫХ НА ОСНОВЕ НЕЛИНЕЙНОЙ ПРОГНОЗНОЙ МОДЕЛИ

Мочалов В.П., Братченко Н.Ю., Гостева Д.В.

ФГАОУ ВО «Северо-Кавказский федеральный университет», Ставрополь, e-mail: mochalov.valery2015@yandex.ru

Аннотация. Целью статьи является повышение эффективности функционирования системы распределения и балансировки нагрузки центров обработки данных облачных сред за счет разработки и применения механизма прогнозирования состояний сетевого трафика, характеризуемого фрактальным самоподобием. При разработке прогнозной модели использованы методы нелинейной динамики, учитывающие статистическое самоподобие нагрузки и обеспечивающие решение задачи прогнозирования моментов ее предполагаемых всплесков. Проверка на хаотичность сетевого трафика выполнена путем расчета спектра показателей Ляпунова. Для восстановления фазового портрета процесса применена теорема Такенса – Мане. Для сглаживания сетевого трафика, устранения его шумовых компонент, выделения наиболее информативных гармоник и исключения случайных возмущений использован метод сингулярного спектрального анализа. Математическая модель и динамический алгоритм прогнозной модели состояния нелинейной системы представлены в виде системы дискретных отображений предыдущих и последующих значений временного ряда и связующих их регрессионного аппроксимирующего полинома. Представленный алгоритм отличается от существующих учетом особенностей фрактального самоподобия входной нагрузки, негативно влияющего на показатели качества, использованием прогнозной модели, разработанной на основе методов нелинейной динамики, а также возможностью выбора рациональных параметров балансировки по критерию равномерной загрузки ресурсов серверов. В статье показано, что динамический алгоритм балансировки нагрузки, построенный на нелинейных подходах и прогнозных моделях, позволяет более качественно, по сравнению с традиционными методами, решать задачи распределения нагрузки между серверами кластеров центров обработки данных. Обоснован вывод об эффективности разработанного алгоритма.

Ключевые слова: балансировка нагрузки, фрактальный сетевой трафик, самоподобие, прогноз, сингулярный спектральный анализ

ALGORITHM FOR LOAD BALANCING OF A DATA PROCESSING CENTER BASED ON A NONLINEAR FORECAST MODEL

Mochalov V.P., Bratchenko N.Yu., Gosteva D.V.

North Caucasus Federal University, Stavropol, e-mail: mochalov.valery2015@yandex.ru

Annotation. The purpose of the article is to increase the efficiency of the cloud data center load distribution and balancing system by developing and applying a mechanism for predicting network traffic conditions characterized by fractal self-similarity. In developing the predictive model, methods of nonlinear dynamics were used, taking into account the statistical self-similarity of the load and providing a solution to the problem of predicting the moments of its expected bursts. Checking for the randomness of network traffic was performed by calculating the spectrum of Lyapunov exponents. The Takens-Manet theorem is applied to reconstruct the phase portrait of the process. To smooth network traffic, eliminate its noise components, highlight the most informative harmonics and eliminate random disturbances, the method of singular spectral analysis was used. The mathematical model and the dynamic algorithm of the predictive model of the state of a nonlinear system are presented as a system of discrete mappings of previous and subsequent values of a time series and their connecting regression approximating polynomial. The presented algorithm differs from the existing ones by taking into account the features of fractal self-similarity of the input load, which negatively affects quality indicators, using a predictive model developed on the basis of nonlinear dynamics methods, as well as the possibility of choosing rational balancing parameters according to the criterion of uniform loading of server resources. The article shows that a dynamic load balancing algorithm based on nonlinear approaches and predictive models allows solving load distribution problems between servers of data center clusters more efficiently than traditional methods. The conclusion about the effectiveness of the developed algorithm is substantiated.

Keywords: load balancing, fractal network traffic, self-similarity, prediction, singular spectral analysis

Задачи эффективного использования информационных ресурсов, их оптимальная загрузка, сокращение времени вычисления и возможность гарантированного обеспечения требуемого качество сервиса (SLA) являются ключевыми для распределенной вычислительной среды центра обработки данных (ЦОД) облачных сред. Как показали многочисленные исследования, между подсистемами ЦОД передаются информационные потоки телекоммуникационного трафика, характеризуемого хаотической структурой, самоподобием, долговременной зависимостью. Широко известные алгоритмы балансировки нагрузки используют приближенные линейные или эвристические под-

ходы, которые не учитывают особенности фрактальной структуры нагрузки современных мультисервисных сетей и не обеспечивают статистически равномерную загрузку серверов кластеров ЦОД. Возможным подходом к решению данных задач является использование методов нелинейной динамики, учитывающих статистическое самоподобие нагрузки и обеспечивающих решение задачи прогнозирования моментов предполагаемых всплесков нагрузки, используемых для своевременного выделения необходимых ресурсов для ее обработки. В нелинейной динамике подобные системы исследуются методом реконструкции фазового пространства по набору значений одномерного временного ряда, являющегося отражением ее состояния. Данный подход дает возможность определять свойства нелинейного процесса по отдельным элементам описывающего его временного ряда, является основой построения прогнозной модели и динамического распределения нагрузки ЦОД в условиях фрактальной сетевой нагрузки.

Материалы и методы исследования

Одним из направлений решения задачи рационального использования аппаратно-программных ресурсов ЦОД в условиях неоднородной нагрузки является ее статистически равномерное распределение в среде серверов ЦОД. Фрагмент рассматриваемой системы представлен на рис. 1 и используется как основа построения ЦОД облачных сред. Обеспечить своевременное выделение необходимых ресурсов ЦОД можно путем прогнозирования предполагаемых всплесков нагрузки и рационального ее распределения по серверам кластеров. Задача прогнозирования решается методом реконструкции фазового пространства нелинейной системы по порожденному временному ряду динамики ее развития, основанному на поиске и выделении k ближайших фазовых траекторий и последующем восстановлении фазового пространства [1–3]. При этом распределение нагрузки осуществляется с использованием известных алгоритмов балансировки RR, WRR, CAP, LARD, AMLB, TLoB с добавлением элементов прогнозных решений.

Распределитель нагрузки, построенный на базе программируемого контроллера [4], реализует динамический алгоритм балансировки нагрузки, характеризуемой специальными свойствами второго порядка, самоподобием, последействием, осуществляя при этом прогнозные оценки интенсивности входящего потока запросов, а также фильтрацию, исключение аномальных отсчетов, сглаживание данных, например, с помощью методов сингулярного спектрального анализа. Критерий оценки эффективности алгоритма основан на показателях загрузки серверов кластеров ЦОД, при условии минимального отклонения их загруженности от заданного значения.

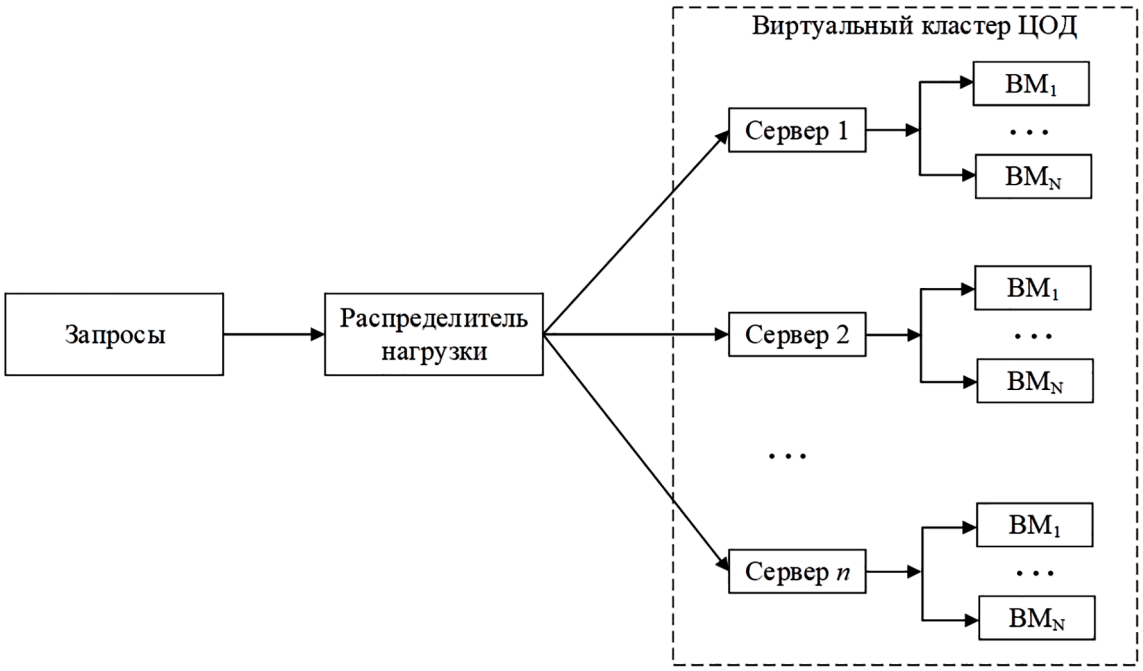

Рис. 1. Структурная схема фрагмента ЦОД

Модель балансировки нагрузки в кластере ЦОД имеет вид

$$P_{ij}\left(k_1,k_2,...,k_N\right)=f\left(k,N,\lambda\right),$$

где $P_{ij}\left(k_1,k_2,...,k_N\right)$ – матрица распределения i-х запросов по $j=\overline{1,N}$ серверам кластера;

N – число серверов;

λ – интенсивность входящего потока запросов.

При реализации алгоритма прогнозирования данное выражение будет иметь вид

$$P_{ij}=\left[x_{ij}\right],\quad \left(i=1,M;\ j=1,N\right)$$

$$R_j\left(k\right)=R_j\left(k-1\right)+\sum_{i=1}^{N}\lambda_{ij}\cdot x_{ij}\left(k\right)$$

где x_{ij} – матрица распределения ресурсов;

$\lambda_{ij}(k)$– прогноз интенсивности нагрузки на k-м шаге.

При этом отклонение по нагрузке каждого из серверов должно быть минимальным

$$H=\frac{\sum_{j=1}^{N}\left(\overline{R}-R_j(k)\right)^2}{N}\to\min$$

при условии

$$\sum_{j=1}^{N}x_{ij}\left(k\right)=1,\quad i=\overline{\left(1,M\right)};$$

$$\sum_{j=1}^{N}\lambda_{ij}\left(k\right)=\lambda,\quad i=\overline{\left(1,M\right)}.$$

Прогнозная модель представляется в виде

$$x_{i+1}=f_1\left(x_i,x_{i-1},...,x_{i-m+1}\right);$$

$$x_{i+2}=f_2\left(x_i,x_{i-1},...,x_{i-m+1}\right);$$

...

$$x_{i+n}=f_n\left(x_i,x_{i-1},...,x_{i-m+1}\right),$$

где x_i – элементы числового ряда;

m – размерность фазового пространства;

$i=m-1,m,...,M-m-1$;

f_n – нелинейный полином

$$f\left(x\right)=\sum_{l_1,l_2,...,l_n=0}^{V}C_{l_1,l_2,...,l_n}\prod_{i=1}^{m}x_i^{l_i},\quad \sum_{i=1}^{n}l_i\le V,$$

где $C_{l_1,l_2,...,l_n}$ – коэффициенты V-го полинома;

m – размерность фазового пространства;

l_n – числовые значения полинома.

Время упреждения прогноза $T=m\cdot\tau$, где τ – временная задержка.

В [5] представлен алгоритм построения реконструированного аттрактора динамической системы.

Временная задержка τ определяется из условия равенства нулю автокорреляционной функции $B(\tau)$

$$B\left(\tau\right)=\frac{1}{m}\sum_{i=0}^{m-1}\left(x_{j,i}-\overline{x}\right)\left(x_{j,i+\tau}-\overline{x}\right),$$

$$m=M-\tau,$$

где x_j – фазовая координата $j=\overline{1,n}$;

m – размерность фазового пространства;

x_{ji} – одномерная реализация временного ряда $i=\overline{1,M}$.

Другой подход определения τ предполагает использование метода средней взаимной информации и функции

$$I\left(\tau\right)=-\sum_{i,j}P_{ij}\left(\tau\right)\ln\frac{P_{ij}\left(\tau\right)}{P_iP_j},$$

где $P_{ji}(\tau)$ – совместная вероятность нахождения точек P_i и $P_j(\tau)$ аттрактора. Пример зависимости $I(\tau)$ от τ приведен на рис. 2. Значение τ определяется по первому минимуму функции $I(\tau)$.

Размерность вложения m вычисляется на основании свойств корреляционного интеграла $C(e)$

$$C\left(e\right)=\lim_{M\to\infty}\frac{1}{M\left(M-1\right)}\sum_{i,j=1}^{N}Q\left(e-p\left(x_i,x_j\right)\right),$$

где $Q\left(x\right)=\begin{cases}1, & x<0\\ 0, & x\ge 0\end{cases}$;

$p\left(x_i,x_j\right)=\left\|x_i-x_j\right\|$;

$N=M-\left(m-1\right)\tau$ – количество точек аттрактора;

e – радиус окружности вокруг точек аттрактора.

При этом необходимо исключить точки числового ряда, расположенные на расстоянии меньше ω (окно Тейлора)

$$\omega>\tau\left(2/N\right)^{2/m}.$$

Числовое значение m определяется по тангенсу угла наклона графика зависимости $\log C(e)$ от $\log(e)$.

В основе другого подхода вычисления размерности m, допускающего его простую программную реализацию, лежит теорема о вложении, определяющая, что в восстанавливаемой системе должны быть исключены самопересекающиеся фазовые траектории и все точки фазового пространства должны пересекаться только одной траекторией.

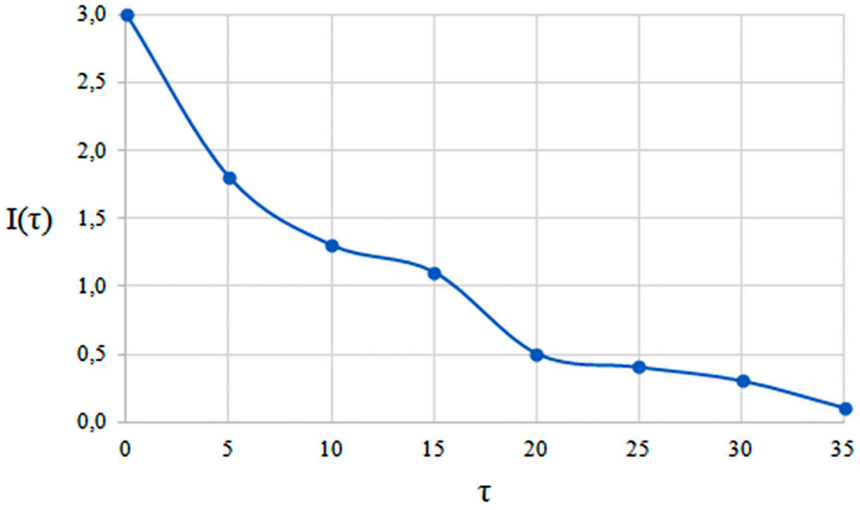

Рис. 2. Зависимость по времени функции I(τ)

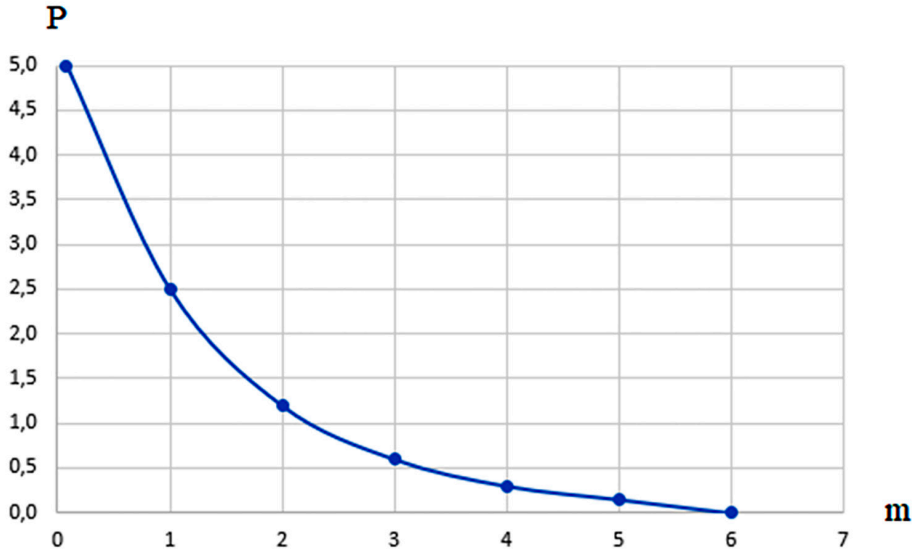

Рис. 3. Зависимость P от m

При таком подходе необходимо многократно вычислять расстояние между точками временного ряда в реконструируемом фазовом пространстве

$$R_i = \frac{\left\|\bar{x}(i+1) - \bar{x}(j+1)\right\|}{\left\|\bar{x}(i) - \bar{x}(j)\right\|},$$

увеличивая на каждом шаге значение m и определяя количество ближайших соседей P. При $P/N = 0$ получаем оптимальное значение m. Пример зависимости количества ближайших соседей P от размерности вложения m приведен на рис. 3.

Проверка на хаотичность временного ряда осуществляется с помощью показателя Ляпунова λ_1, положительное значение которого определяет хаотическое поведение процесса. Максимальный показатель Ляпунова характеризует скорость расхождения фазовых траекторий и определяется выражением [6]

$$\lambda_1(i,k) = \frac{1}{k\Delta t}\left(\ln\frac{d_j(i+k)}{d_j(i)}\right) = \frac{1}{k\Delta t}\left[\ln d_j(i+k) - \ln d_j(i)\right],$$

где Δt – период ряда; $d_j(i)$ – расстояние между близкими соседями, $d(t) = Ce^{\lambda_1 t}$, $d_j(i) = C_j e^{\lambda_1(i\Delta t)}$ – скорость расхождения близких траекторий.

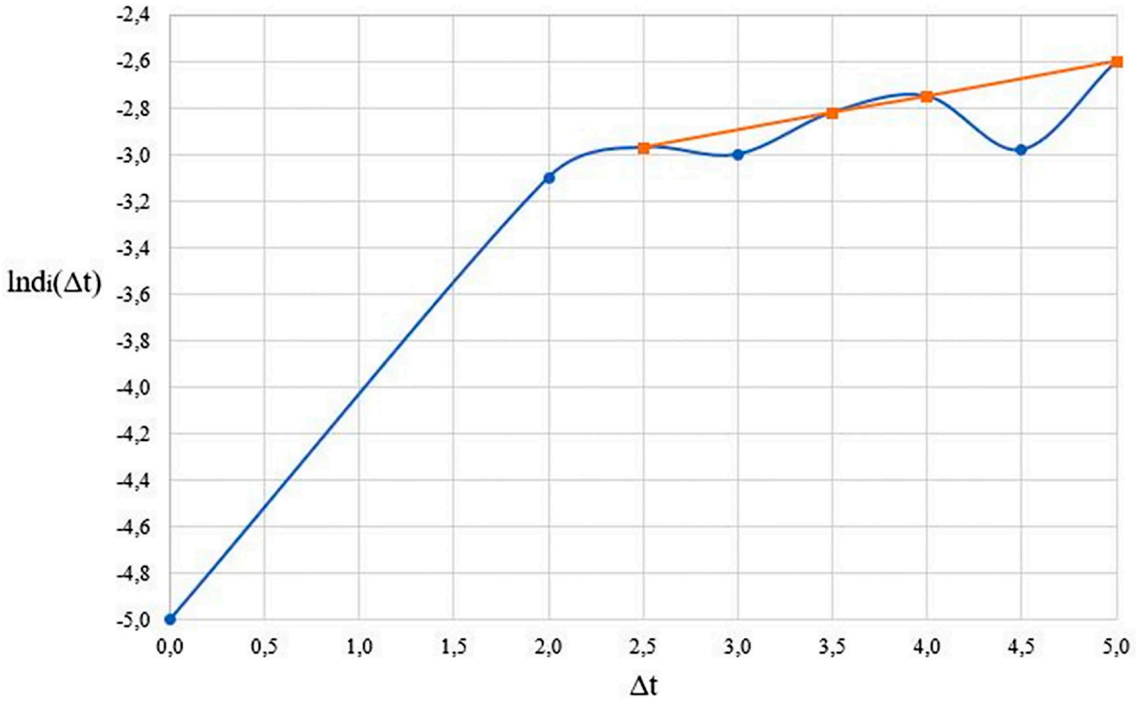

Рис. 4. График определения λ_1

Отсюда следует $\ln d_j(i) \approx \ln C_j + \lambda_1 (i\Delta t)$, то есть λ_1 определяется как тангенс угла наклона прямой, аппроксимирующей данную зависимость. С использованием программ пакета TISEAN 3.0.1 определены значения спектра показателей Ляпунова.

На рис. 4 приведен пример графика расхождения траектории аттрактора и его аппроксимирующая прямая.

Для повышения точности прогноза необходимо реализовать алгоритм устранения случайных шумовых компонент входного телекоммуникационного трафика. Выделить наиболее информативные компоненты временного ряда, порождаемого сетевым трафиком, устранить шумы и случайные возмущения предлагается с помощью метода сингулярного спектрального анализа (SSA) [7, 8]. Метод реализует процедуру декомпозиции скалярного временного ряда $\{x1, x2, \ldots, xN\}$ и получение $K = N - L + 1$ векторов вложения, $1 < L < N$, где N – длина временного ряда. Из полученных векторов вложения составляется траекторная матрица X. Для матрицы $S = X \cdot X^T$ получаем N собственных чисел $\lambda_1, \lambda_2, ..., \lambda_N$ и N ортонормированных собственных векторов

$$U_1, U_2, ..., U_N, \quad V_i = \frac{X^T U_i}{\sqrt{\lambda_i}}.$$

Выражение $\left(\sqrt{\lambda_i}, U_i, V_i\right)$ является i-й собственной тройкой сингулярного разложения, а выражение $(\sqrt{\lambda_i} \cdot V_i = X^T \cdot U_i)$ – вектором *i*-й главной компоненты. Уровень наиболее значимых составляющих временного ряда зависит от параметров собственных троек сингулярного разложения. При исследовании системы, построения и управления матрицами и векторами алгоритма SSA используются программные пакеты линейной алгебры Maple, linalg, LinalgAlgebra, MTT. Пример численных значений главных компонент временного ряда приведен на рис. 5.

Очевидно, что первые три компоненты практически полностью определяют поведение ряда. Дальнейшая группировка собственных троек и диагональное усреднение результирующей матрицы обеспечивают разложение исходного ряда на сумму восстановленного ряда и шума.

При этом доля самого незначительного вклада в общую дисперсию данных приходится на шумовые компоненты [9]. SVD-разложение представляется в виде

$$X = U \sum V^T = \sum_{i=1}^{L} x_i, \quad x_i = \sigma_i u_i v_i^T,$$

где $\sum$ – диагональная матрица, x_i – элементарная матрица.

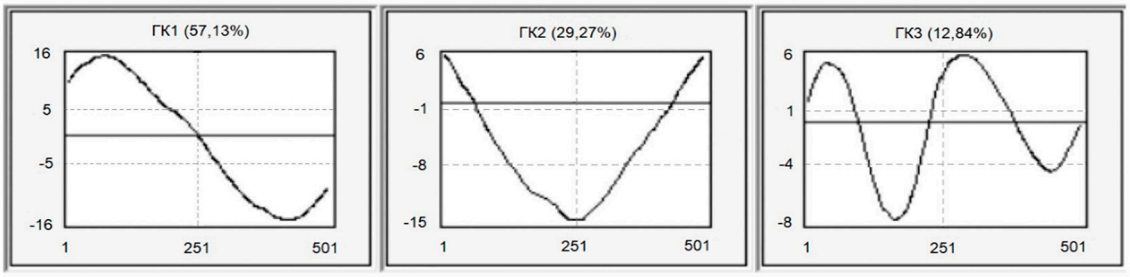

Рис. 5. Главные компоненты временного ряда

Сингулярный спектр определяется выражением $c_i = \sigma_i^2 \Big/ \sum_{j=1}^{L} \sigma_j^2$.

На этапе восстановления L матриц $x_i \left(i = 1, ..., L\right)$ делится на τ непересекающихся групп, а матрица X переходит в матрицу $\widetilde{X} = \sum_{k=1}^{\tau} X_{I_k}$, где $X_{I_k} = \sum_{i \in I_k} x_i$.

Диагональное усреднение матриц X_{I_k} обеспечивает преобразование ряда S_N в ряд $\widetilde{S_N}$ с элементами x_n^k

$$x_n^k = \begin{cases} \dfrac{1}{n}\sum_{m}^{n} X_m, n-m+1 \quad при \quad 1 \le n < L \\ \dfrac{1}{L}\sum_{m=1}^{L} X_m, n-m+1 \quad при \quad L \le n < K \\ \dfrac{1}{N-n+1}\sum_{m=n-k+1}^{n} X_m, n-m+1 \quad при \quad K+1 \le n < \end{cases}$$

и, следовательно, исходный временной ряд раскладывается на сумму восстановленного ряда и выделенного шума

$$f_n = \sum_{k=1}^{m} f_n^k .$$

Показателем качества, определяющим долю не устраненных шумовых компонент, является выражение [10]

$$W = \frac{\sum_i \left(A_i - x_i\right)^2}{\sum_i \left(R_i\right)^2} \cdot 100\% ,$$

где A_i – элементы восстановленного ряда,
R_i – элементы шума.

Динамический алгоритм балансировки нагрузки, построенный на основе локального метода ее прогнозирования, представлен ниже и включает в себя следующие шаги:

1. Оценка требуемого количества элементов ряда $N_{\min}$, проверка его на хаотичность по старшему показателю Ляпунова

$$\lambda_1\left(i,k\right) = \frac{1}{k\Delta t}\left(\ln\frac{d_j\left(i+k\right)}{d_j\left(i\right)}\right) =$$

$$= \frac{1}{k\Delta t}\left[\ln d_j\left(i+k\right) - \ln d_j\left(i\right)\right],$$

$\lambda_1 > 0$ – хаотическое движение, $\lambda_1 \le 0$ – регулярное движение. Реализация фильтрующего алгоритма с использованием метода SSA.

2. Реконструкция фазового пространства с определением лага τ с помощью метода средней взаимной информации и функции

$$I\left(\tau\right) = -\sum_{i,j} P_{ij}\left(\tau\right)\ln\frac{P_{ij}\left(\tau\right)}{P_i P_j},$$

а также вычисление размерности вложения m с использованием корреляционного интеграла $C(e)$

$$C\left(e\right) = \lim_{M\to\infty}\frac{1}{M\left(M-1\right)}\sum_{i,j=1}^{N} Q\left(e - p\left(x_i, x_j\right)\right).$$

3. Разработка прогнозной модели на базе алгебраического полинома степени V одинакового для всех его переменных

$$f\left(x\right) = \sum_{l_1,l_2,...,l_n=0}^{V} C_{l_1,l_2,...,l_n}\prod_{i=1}^{m} x_i^{l_i}, \quad \sum_{i=1}^{n} l_i \le V .$$

4. Прогнозная оценка интенсивности нагрузки на k-м шаге $\lambda_{ij}(k)$.

5. Распределение запросов по серверам

$$R_j\left(k\right) = R_j\left(k-1\right) + \sum_{i=1}^{N}\lambda_{ij}\left(k\right)\cdot x_{ij}\left(k\right).$$

Результаты исследования и их обсуждение

В настоящее время для решения задач управления нагрузкой ЦОД широко применяются методы математической статистики. Данные методы не учитывают хаотическую структуру нагрузки, характеризующуюся нелинейным характером и фрактальным самоподобием. Большое количество публикаций посвящено также исследованию поведения детерминированных динамических систем методами нелинейного анализа и теории хаоса. Однако подобные работы практически не содержат реализуемые методики решения нелинейных прикладных задач. Решить некоторые из них, получить численные результаты исследований и стало одной из целей данной работы. Возможным подходом к решению данных задач является использование алгоритмов прогнозирования, разработанных на основе методов нелинейной экстраполяции состояний сетевого трафика, размерность фазового пространства которого неизвестна. Восстановление фазового пространства процесса методами нелинейной динамики дает возможность определять его свойства по отдельным параметрам описывающего его временного ряда, является основой построения прогнозной модели и динамического распределения нагрузки ЦОД.

Заключение

В настоящей работе сформулирована и решена задача распределения и балансировки нагрузки ЦОД, отличающаяся применением механизма прогнозирования состояний сетевого трафика, характеризуемого фрактальным самоподобием. В основу решения задачи положены особенности фрактального сетевого трафика, отражающие характер изменения информационной нагрузки ЦОД. Реконструкция фазового пространства подобного динамического процесса осуществлена в соответствии с теоремой Такенса – Мане, с использованием метода задержек. Разработаны математическая модель и динамический алгоритм прогнозной модели состояния нелинейной системы. Прогнозная модель представлена в виде системы дискретных отображений предыдущих и последующих значений временного ряда и связующих их аппроксимирующего полинома степени v. При этом горизонт прогноза меняется в соответствии с изменением интенсивности входящей нагрузки. Проверка на хаотичность временного ряда осуществляется путем определения значений спектра показателей Ляпунова с использованием программ пакета TISEAN 3.0.1. Алгоритм фильтрации шумов входного телекоммуникационного трафика реализован методом сингулярного спектрального анализа (SSA). Разработанный динамический алгоритм распределения и балансировки нагрузки обеспечивает выбор рациональных параметров балансировки по критерию равномерной загрузки ресурсов серверов.

Список литературы

1. Анищенко В.С. Знакомство с нелинейной динамикой. М.: Издательство ЛКИ, 2008. 224 с.

2. Курников П.А., Крапухина Н.В. Реконструкция фазового пространства динамической системы высоконагруженного кэширующего механизма информационных систем // Информационные технологии и вычислительные системы. 2019. Вып. 1. С. 49–65.

3. Fowler H.J., Leland W.E. Local area network traffic characteristic, with implications for broadband network congestion management // IEEE Journal on Selected Areas in Communications. 2021. Vol. 9. P. 1139–1149.

4. Mochalov V.P., Linets G.I., Bratchenko N.Y., Govorova S.V. An analytical model of a corporate software-controlled network switch // Scalable Computing. 2020. Vol. 21 (2). Is. 2. P. 337-346.

5. Мальцев Г.Н., Назаров А.В., Якимов В.Л. Алгоритм реконструкции фазового пространства динамической системы и его применение // Информационно-управляющие системы. 2014. № 2. С. 33–39.

6. Никульчев Е.В., Паяин С.В., Питиков Д.А., Плужник Е.В. Вычисление характеристик динамического хаоса по трафику компьютерных сетей // Фундаментальные исследования. 2014. № 8. С. 812–816.

7. Поршнев С.В., Рабайа Ф. Исследование особенностей применения метода сингулярного спектрального анализа в задаче анализа и прогнозирования временных рядов: монография. Ульяновск: Зебра, 2016. 167 с.

8. Зиненко А.В. Прогнозирование финансовых временных рядов с использованием сингулярного спектрального анализа // Бизнес-информатика. 2023. Т. 17, № 3. С. 87–100.

9. Simar P.S. Analysis of Load Balancing Algorithms using Cloud Analyst // International Journal of Grid and Distributed-Computing. 2016. № 9. P. 11–24.

10. Кучерявый А.Е., Маколкина М.А., Киричек Р.В. Тактильный Интернет. Сети связи со сверхмалыми задержками // Электросвязь. 2016. № 1. С. 44–46.