УДК 519.876.5:621.395

І.В. Дегтяренко (канд. техн. наук, доц.), О.О. Абраменко (студент), О.С. Чекунков (студент)
Донецький національний технічний університет, м. Донецьк
кафедра автоматики та телекомунікацій
E-mail: ilya.degtyarenko@gmail.com

ПРОГНОСТИЧНЕ КЕРУВАННЯ НАВАНТАЖЕННЯМ НА СЕРВЕРИ З ВИКОРИСТАННЯМ НЕЙРОМЕРЕЖІ

Проведено аналіз впливу завантаженості серверів телекомунікаційної мережі на параметри QoS. Удосконалено методи балансування навантаження на сервери кластера, завдяки впровадженню прогностичного керування з використанням нейромережі. Розроблена структура та алгоритм роботи системи прогностичного керування. Проведена оцінка ефективності запропонованих рішень шляхом імітаційного моделювання.
Ключові слова: *трафік, параметри QoS, кластер серверів, прогностичне керування, нейромережа, імітаційне моделювання.*

Загальна постановка проблеми

Одна з основних тенденцій розвитку телекомунікаційного ринку — зростання попиту користувачів на різноманітні інтелектуальні послуги. У зв'язку з цим набуває популярності концепція Triple Play, згідно з якою на базі трьох сервісів (передача аудіо реального часу, відео реального часу, високошвидкісних даних) комбінуються різноманітні мультимедійні послуги: телефонія, високошвидкісний доступ до Інтернету, телевізійне мовлення, відеоконференції, онлайн-ігри, відео за запитом тощо, які надаються абонентові через єдину інфраструктуру мультисервісної конвергентної мережі.

Необхідність розгортання послуг Triple Play спонукає операторів зв'язку модернізувати власні мережі згідно з концепцією мереж наступного покоління (NGN) [1]. Однією з важливих проблем, яка виникає при побудові NGN мережі, є необхідність забезпечення великої кількості абонентів високо якісним доступом до ресурсів, що надають мультимедійні послуги. Це вимагає не тільки високої пропускної здатності мережі, а і виконання вимог до якості обслуговування (QoS) трафіку даних послуг [2]. Ключову роль при цьому відіграє якість роботи серверного устаткування рівня послуг мережі. В години пікового навантаження „лавина" запитів може призвести до непередбачуваних наслідків — від надмірно довгого обслуговування запитів абонентів до перевантаження і відмови в обслуговуванні. Одним із способів усунення цієї проблеми є придбання більш потужного обладнання, яке зможе витримувати більше навантаження. Але цей спосіб вимагає від операторів значних капіталовкладень та не в повної мірі вирішує дану проблему. Іншим способом є підвищення ефективності роботи наявних серверних ресурсів, для чого використовують засоби статичного та динамічного балансування навантаження. Максимальна якість обслуговування при умові суттєвої утилізації ресурсів досягається, коли навантаження розподіляється між існуючими серверами рівномірно, тому метою застосування системи балансування є досягнення якомога меншого розбалансу утилізації серверів кластера.

Проблема балансування навантаження актуальна не тільки на рівні послуг, але і на рівні керування мережі NGN. Як правило, Softswitch який здійснює керування послугами має також кластерну архітектуру, отже виникає необхідність рівномірного розподілення навантаження між елементами кластеру для забезпечення високої швидкості обслуговування запитів. Загалом балансування навантаження може бути впроваджене на будь-яких елементах, що реалізують функції, критичні до швидкості виконання і доступності,

наприклад на серверах аутентифікації, авторизації і тарифікації. Керування навантаженням на серверні ресурси слід проводити з урахуванням прогностичних моделей трафіка мережі. Це дозволить більш ефективно використовувати наявні технічні ресурси мережі.

Метою даної роботи є поліпшення параметрів якості обслуговування трафіку мультимедійних послуг за рахунок розробки системи прогностичного керування навантаженням на серверні ресурси телекомунікаційних мереж.

Для досягнення поставленої мети необхідно вирішити наступні **задачі**:

1) розробити математичну модель, що дозволяє оцінити вплив навантаження серверів телекомунікаційної мережі на параметри QoS трафіку мультимедійних послуг;
2) розробити засоби прогнозу характеристик запитів на сервери;
3) розробити структуру та алгоритми роботи системи прогностичного керування навантаженням на сервери;
4) провести аналіз ефективності роботи даної системи.

Вирішення задач та результати дослідження

Параметри QoS на рівні послуг та керування передусім пов'язані з роботою кластера серверів. Час обробки запиту на сервері $\tau_{обробки}$ загалом є випадковою величиною, що залежить від параметрів кластерних ресурсів та потоку запитів. Аналіз літературних джерел [3] показав, що ключовий вплив на цю величину має значення поточної утилізації сервера W і час, необхідний на виконання однієї операції без урахування черг $\tau_{запиту}$ (тривалість обробки самого запиту). Аналітична залежність для $\tau_{обробки}$ може бути описана формулою [3]:

$$\tau_{обробки} = \tau_{запиту} + \frac{\mathrm{W} \cdot \tau_{\text{запиту}}}{1 - \mathrm{W}}. \quad (1)$$

У свою чергу, $\tau_{обробки}$ впливає на такі параметри QoS як затримка, джиттер та ймовірність відмови в обслуговуванні. Дослідження доводять, що імовірність втрат (P) в системі масового обслуговування (СМО) з обмеженим часом очікування в черзі зростає зі збільшенням величини відношення інтенсивності вхідного потоку до інтенсивності обслуговування A, яка пропорційна утилізації сервера [4]:

$$P = \frac{e^{-(V-A)\mu \cdot \tau_{крит}}}{\frac{1}{E_v(A)} + \frac{A}{V-A}(1 - e^{-(V-A)\mu \cdot \tau_{крит}})}, \quad (2)$$

де V — кількість каналів СМО; μ — інтенсивність обслуговування; $E_v(A)$ — перша формула Ерланга; $\tau_{крит}$ — критичний час очікування у черзі, після якого виклик втрачається.

Отже, для покращення параметрів QoS на рівні послуг необхідним є зменшення ймовірності перевантаження серверів.

Кожен запит на обслуговування характеризується навантаженням, яке він створює на сервер (w(i) — ємність серверних ресурсів, необхідних для обслуговування цього запиту) та тривалістю обслуговування (z(i) — час, що витрачає сервер на обслуговування запиту). Якщо час обробки перевищує максимально допустимий ($Z_{доп}$), то відбувається втрата запиту ($N_{пот}$ — кількість втрачених запитів). Величина навантаження від потоку запитів характеризується імовірнісним законом розподілу. Сервери кластера виконують обробку запитів та характеризуються ємністю та продуктивністю ресурсів.

Математична модель роботи сервера представлена графом станів (див. рис. 1), де 1 — початковий стан; 2 — надходження нового запиту; 3 — запит не отримав обслуговування через перевантаження процесора; 4 — запит надіслано на обробку; 5 — запит відкинуто через перевищення припустимого часу обробки; 6 — завершення роботи з запитом.

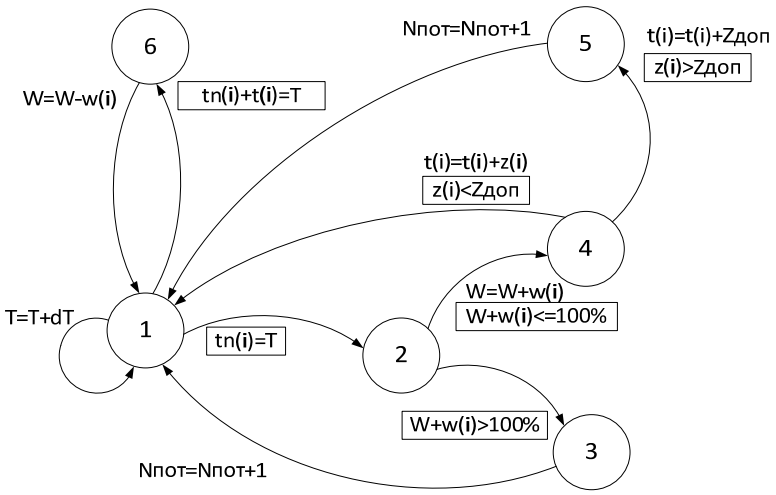

Рисунок 1 — Граф станів сервера

Система балансування виконує динамічний розподіл потоку запитів, що надходить, на сервери. Робота системи балансування повністю визначається закладеним алгоритмом. Найбільш відомі алгоритми балансування здійснюють вибір сервера для спрямування чергового запита на основі:

- порядкового номера запита та сервера (алгоритм Round Robin);
- інформації про завантаженість серверів (алгоритми Least Connection, Least Load).

Можливість приймати рішення на основі інформації про навантаження від потоку запитів може значно підвищити ефективність балансування. Але у зв'язку із наявністю багатьох факторів обчислити навантаження від кожного з запитів має можливість лише сервер, тому система балансування не має змоги безпосередньо врахувати властивості навантаження потоку запитів.

Проте у мережі NGN з наданням мультимедійних послуг характерна наявність прихованих залежностей між навантаженням, що створюють серії запитів. Це зумовлене властивостями циклічності коливання та самоподібності трафіку. Завдяки цьому можливим є прогнозування навантаження для наступних a запитів на основі реальних даних про попередні p запитів. Для задачі прогнозування навантаження пропонується використовувати фокусовану рекурентну нейронну мережу [5], побудовану на основі багатошарового персептрону (див. рис. 2).

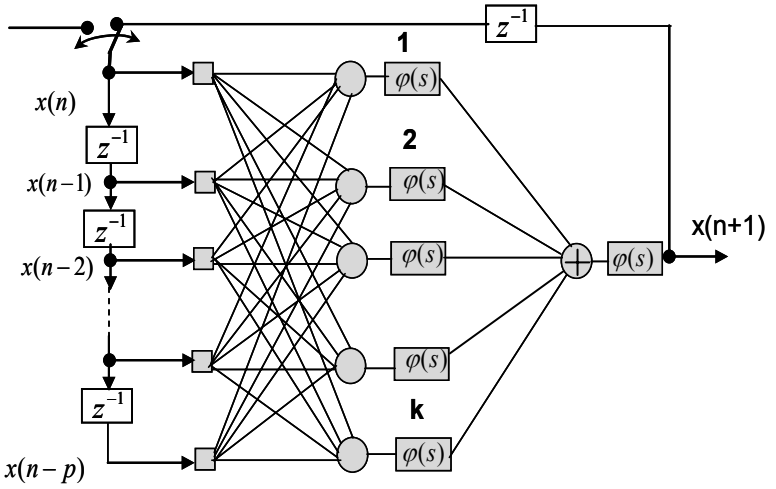

Рисунок 2 — Стуктура нейронної мережі

В процесі синтезу архітектури нейронної мережі, важливим етапом є вибір функції активації, виходячи зі специфіки задачі, що вирішується. Для задачі прогнозування навантаження було розглянуто декілька функцій активації, серед яких перевага була віддана саме логістичній функції , зважаючи на наступні чинники:

- логістична функція є обмеженою, що не дозволяє вихідному сигналу необмежено зростати при будь-якому значенні аргументу (лежить інтервалі від 0 до 1);
- логістична функція є невід'ємною на всій області допустимих значень аргументу (у реальній системі параметри трафіку не можуть приймати від'ємні значення, тому це є необхідною умовою фізичної реалізації);
- логістична функція має похідні, що можуть бути виражені через саму функцію.

Аналітично логістичну функцію можна формалізувати наступним чином:

$$\varphi(s) = \frac{1}{1+e^{-\alpha S}} \ . \tag{3}$$

Таким чином, якщо в нейронній мережі використовується логістична функція активації, процес прогнозування, з точки зору апроксимації функцій, задовольняє умові теореми про повноту нейронної мережі [5], яка говорить, що будь яка безперервна функція на замкнутій обмеженій множині може бути рівномірно приближена функціями, що обчислюються нейронними мережами, якщо функція активації безперервна та має другу похідну .Окрім того, існування похідних обумовлює можливість навчання нейронної мережі за допомогою методу зворотного розповсюдження помилки.

Багатокрокове прогнозування відбувається за ітераційною концепцією (рис. 3). Якщо розглядати модель нейромережі в динаміці, з точки зору цієї концепції, то можливість такого прогнозування обґрунтовано теоремою Такенса про вкладні затримки [5]. Після того, як вхідний шар було ініціалізовано значеннями навантаження попередніх сесій, що зберігаються у пам'яті лінійної затримки порядку *p*, обчислюється значення вихідного нейрона, що відображає прогнозне навантаження наступної *(n+1)* сесії.З точки зору алгоритму прямого розповсюдження, сигнал на виході нейронної мережі можна формалізувати таким чином:

$$x^*(n+1) = \varphi\Big(\sum_{j=1}^{k}\Big[\omega_j \varphi\Big(\sum_{l=0}^{p}\omega_j(l)x(n-l)+b_j\Big)+b_0\Big]\Big). \tag{4}$$

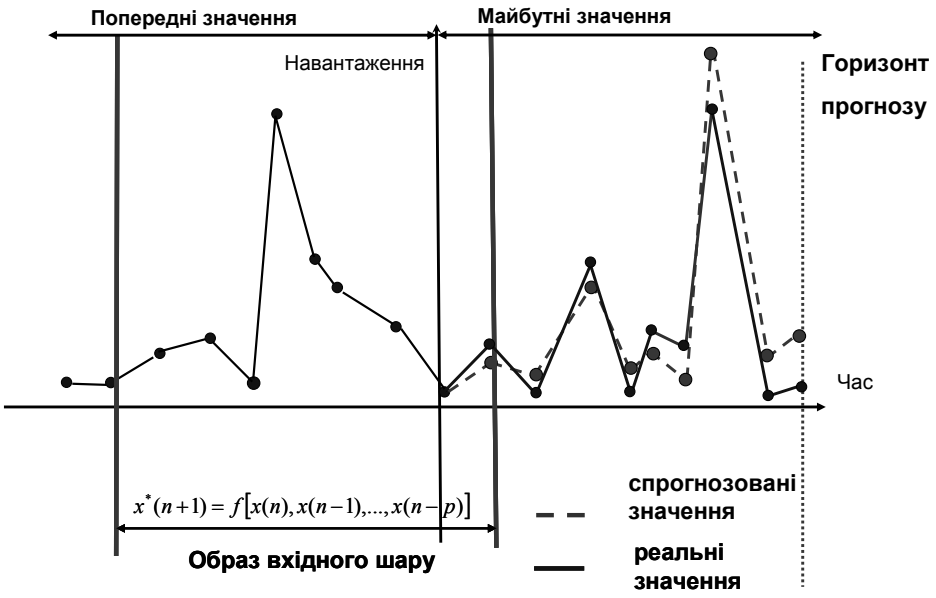

Рисунок 3 — Концепція ітераційного прогнозування

Це значення приймається за дійсне та додається до вхідного вектора на наступному кроці для того щоб обчислити прогнозне значення *x(n+2)*. Процедура прогнозу повторюється ітераційно до досягнення горизонту прогнозу. Похибка прогнозування має властивість зростати з кожною ітерацією, тому горизонт прогнозу обирається за критерієм перевищення порогу помилки на кроці *m* або середньої помилки за всі кроки.

Важливою задачею є визначення кількості нейронів в прихованих шарах. Слід зауважити, що чим більше їх кількість, тим в середньому краще буде апроксимована вибірка даних у процесі навчання. Але це не завжди добре, через те що при надто великій кількості нейронів, нейронна мережа «запам'ятає» тільки ту інформацію на якій вона навчилася. Задача ж навчання, навпаки, ставить на меті запам'ятання не однієї конкретної вибірки, а узагальнення прихованої інформації, що міститься у вхідних даних. Невелика кількість нейронів у прихованих шарах створить не достатньо високий ступінь нелінійності нейронної мережі, що також зробить прогнозування неефективним. Тому для оцінки оптимальної кількості нейронів в прихованих шарах можна скористатися наступною формулою [6]:

$$\frac{mN}{1+\log_2 N} \le L_w \le m\left(\frac{N}{m}+1\right)(p+m+1)+m, \qquad (6)$$

де: m — розмір вихідного сигналу, N — розмір вибірки для навчання p — розмір вхідного сигналу; L_w — кількість синаптичних зв'язків.

Для нашого випадку :

$$\frac{1*5000}{1+\log_2 5000} \le L_w \le 1\left(\frac{5000}{1}+1\right)(20+1+1)+1,$$

$$390 \le L_w \le 110023.$$

На основі кількості синаптичних зв'язків,можна встановити кількість нейронів для мережі з одним прихованим шаром. Для розрахунку вибираємо нижню границю, запобігаючи ефекту перенавчання:

$$k = \frac{L_w}{p+m} = \frac{390}{20+1} = 18{,}57 \qquad (7)$$

Округлюючи кількість нейронів до найближчою десятки отримаємо k=20.

Слід зауважити, що вибір порядку мережі p, є компромісним між обчислювальною складністю (складність та тривалість процесів функціонування та тренування) нейронної мережі і якості прогнозу. Для систем багатокрокового прогнозування бажано, щоб виконувалася умова $p>2a$, де a — кількість кроків прогнозування.

Система прогностичного керування навантаженням (див. рис. 4) складається з двох підсистем та елементу керування (рис. 4). Підсистема 1 є основною та реалізує прогностичне балансування з використанням нейронної мережі, друга є допоміжною та реалізує балансувальник з динамічним зважуванням (модифікований Round Robin).

Інформація $\bar{I} = \{\overline{W}, \overline{X}\}$ про поточний стан кластера надходить до системи з інтервалом T. Стан серверів описується векторами утилізації $\overline{W} = \{w_1, ..., w_S\}$ та характеристик запитів $\overline{X} = \{x_n, x_{n-1}, ..., x_{n-p}\}$. Результатами роботи даних підсистем є адреса сервера (A) ,на який буде надіслано наступний запит.

На початковому етапі для балансування навантаження використовується модифікований Round Robin з динамічним зважуванням (підсистема 2). Паралельно проводиться навчання нейромережі. Результатом навчання є вектор вагових коефіцієнтів синаптичніх зв'язків $\overline{\omega}$, що передається нейронній мережі. Після закінчення навчання проводиться оцінка достовірності прогнозу. Якщо помилка e не перевищує граничного

значення $e_{гр}$ керування передається підсистемі з прогностичним балансуванням навантаження.

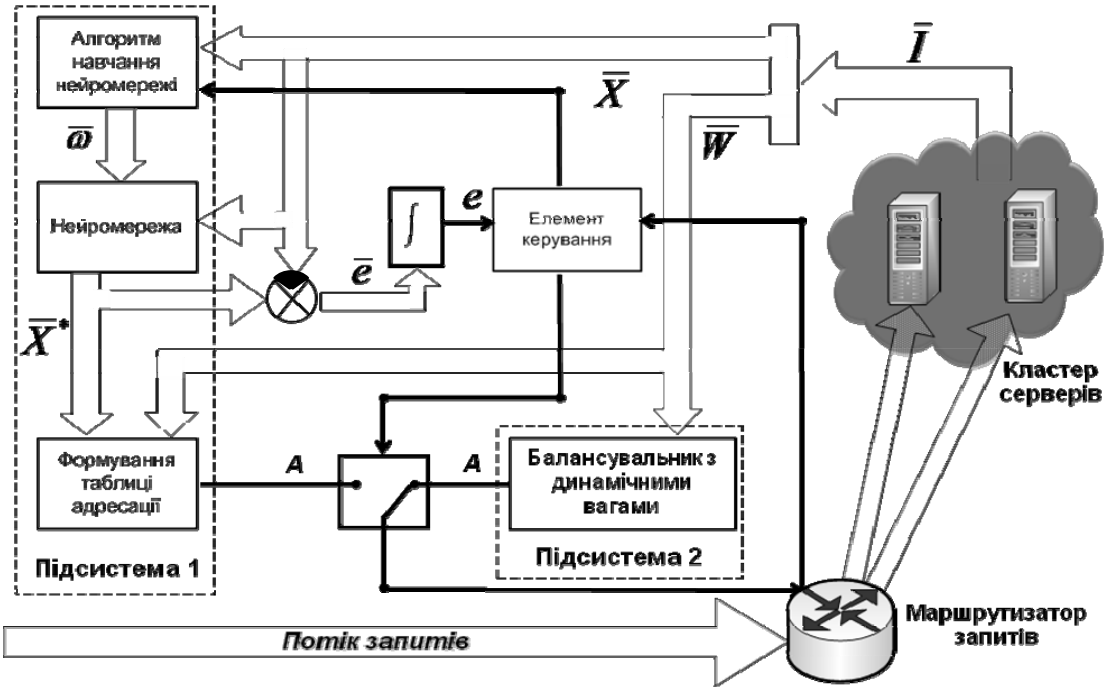

Рисунок 4 — Структура системи прогностичного керування навантаження на сервери

На відміну від стандартного зваженого Round Robin, вагові коефіцієнти якого є статичними та задаються ззовні на основі даних про різну обчислювальну потужність серверів, модифікований зважений Round Robin реалізований в підсистемі 2 здійснює періодичну зміну ваг. Ваги розраховуються в залежності від поточного завантаження серверів:

$$K_i = \frac{k_i}{\min\limits_{j=0..S}(k_j)}, \tag{8}$$

де k_i — частка вільних ресурсів i-го сервера, K_i — значення вагового коефіцієнта i-го сервера. перераховує вагові коефіцієнти з періодом T в залежності від поточного навантаження на сервери.

Алгоритм прогностичного балансування, що використовується в підсистемі 1, базується на m-кроковому прогнозі навантаження сесій $\overline{X}^* = \{x^*_{n+1}, x^*_{n+2}, ..., x^*_{n+m}\}$, який формується нейронною мережею. Ця інформація використовується для складення таблиці адресації за допомогою методів алгоритму Least Load, а саме ведеться пошук на кожному кроці мінімально завантаженого сервера з урахуванням прогнозованого навантаження сесій, що надходять. У випадку, коли горизонт прогнозу вичерпано, тобто таблиця розподілення закінчилася, керування передається підсистемі 2.

Система балансування постійно веде моніторинг середньої помилки прогнозування e, та при досягненні нею граничного значення приймає рішення про передачу керування підсистемі 2 та перенавчання нейронної мережі. На підставі описаного підходу було складено комплексний алгоритм балансування навантаження на сервери. Було проведено імітаційне моделювання роботи системи прогностичного керування з використанням нейромережі за допомогою спеціально розробленого на мові C++ програмного забезпечення.

Для проведення порівняльної оцінки ефективності розроблених засобів прогностичного керування було також проведене імітаційне моделювання роботи системи балансування з використанням стандартних найпоширеніших алгоритмів:

- алгоритму Round Robin — розподілення навантаження відбувається методом перебору серверів по круговому циклу;
- алгоритму пошуку мінімально завантаженого сервера з захистом від стадного ефекту (Exclude) — полягає у виборі мінімально завантаженого сервера серед доступних, після чого сервер стає недоступним до початку наступного циклу;
- модифікованого алгоритму Round Robin з динамічним зважуванням (Weights);
- алгоритму розподілу навантаження з циклами, що розширюються (Cycles) — кожний цикл розподілення навантаження поділений на N вкладених циклів, де N — кількість серверів. Після отримання інформації про завантаженість серверів, вони впорядковуються за зростанням завантаженості. Перший вкладений цикл містить лише один найменш завантажений сервер і запит відправляється до нього. Другий — 2 сервери, які є найменш завантаженими зі всього набору. Після закінчення вкладеного циклу, його розмір збільшується на 1, добавляється наступний найменш завантажений сервер, запити відправляються на сервери за зростанням завантаженості. Останній вкладений цикл включає всі сервери, після його закінчення сервери знову впорядковуються за зростанням і починається перший вкладений цикл.

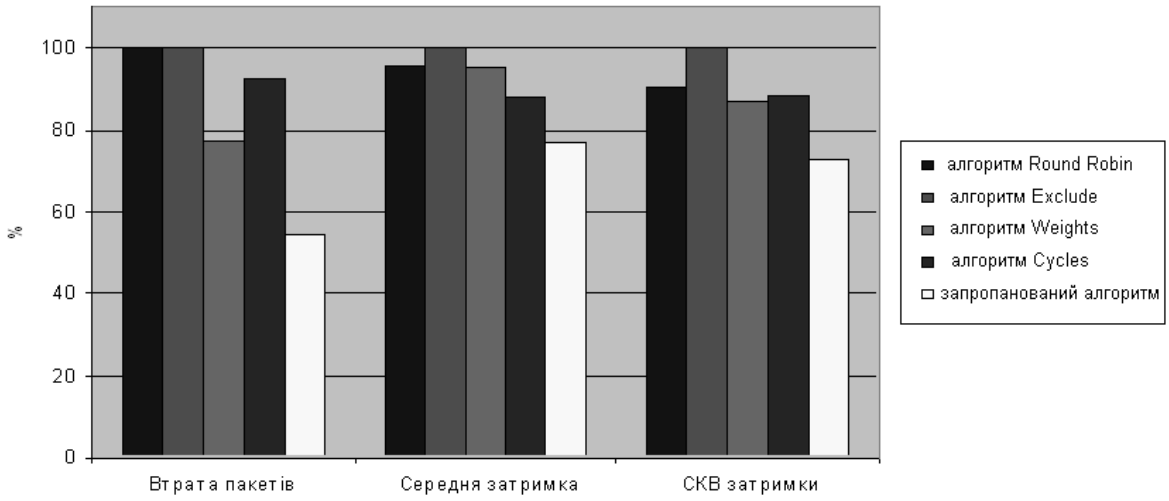

Рисунок 5 — Результати імітаційного моделювання

Для кожного з алгоритмів оцінювалися параметри: середня затримка обробки пакету, СКВ затримки, а також кількість втрачених пакетів. Ці величини безпосередньо впливають на такі параметри QoS як затримка, джіттер та вірогідність втрат. Порівняння результатів роботи різних алгоритмів балансування показано на рис.5. Тут за 100% взяті найгірші з точки зору якості обслуговування значення параметрів, що отримано в ході моделювання.

Запропонований в даній роботі алгоритм балансування з використанням нейронної мережі для прогнозування навантаження показує найкращі результати серед усіх алгоритмів, що моделювались. Порівняно зі стандартним алгоритмом балансування Round Robin затримка і джиттер знижуються на 22%, а ймовірність втрати пакетів майже вдвічі.

Дані дослідження було проведено в рамках всеукраїнської програми „Професіонали майбутнього-2011” організатором якої є ПрАТ „МТС –Україна”.

Висновки

Розроблено математичну модель кластера серверів, яка дозволила провести аналіз впливу завантаженості серверів рівня послуг та керування телекомунікаційної мережі на параметри якості обслуговування трафіку мультимедійних послуг.

Отримали подальший розвиток методи динамічного балансування навантаження на сервери телекомунікаційної мережі, завдяки їх адаптації до характеристик потоку запитів та впровадженню прогностичного керування з використанням нейромережі.

Розроблена структура та алгоритм роботи системи прогностичного керування навантаженням на сервери. Проведено імітаційне моделювання роботи даної системи за допомогою програмної моделі. Аналіз результатів моделювання показав, що використання даної системи дозволяє покращити показники якості надання послуг в телекомунікаційних мережах.

Список використаної літератури

1. ITU-T Rec. Y.2001 General overview of NGN / ITU [Електронний ресурс]. — Вільний доступ з мережі Internet. — http://www.itu.int/itudoc/itu-t/aap/sg13aap/history/y2001/y2001.html. — (30.05.2011).
2. ITU-T Rec. Y.1541 Network performance objectives for IP-based services / ITU [Електронний ресурс]. — Вільний доступ з мережі Internet. — http://www.itu.int/rec/T-REC-Y.1541-200205-S/en. — (30.05.2011).
3. Лихтциндер Б.Я. Интеллектуальные сети связи / Б.Я. Лихтциндер, М.А. Кузякин, А.В. Росляков и др. — М : Эко-Трендз, 2000. — 206 с.
4. Ершов В.А. Мультисервисные телекоммуникационные сети / В.А. Ершов, Н.А. Кузнецов. — М : Издательство МГТУ им. Н.Э. Баумана, 2003. — 424 с.
5. Хайкин С. Нейронные сети: полный курс / С.Хайкин. — М : ВИЛЬЯМС, 2006. —1104 с.
6. Круглов В.В. Искусственные нейронные сети. Теория и практика / В.В.Круглов. — М : Горячая линия — Телеком, 2002. — 382 с.

Надійшла до редакції:
06.02.2012р.

Рецензент:
д-р техн.наук, проф. Скобцов Ю.О.

I. Degtyarenko, O. Abramenko, O. Chekunkov. Predictive Control of Servers Loads Using a Neural Network. *The analysis of servers loads influence on the telecommunications network QoS parameters were considered. Methods of servers load balancing, were improved by the introduction of predictive control using neural networks. Structure and algorithm of the predictive control system was proposed. Simulations results that demonstrated the effectiveness of the proposed solutions were presented.*
Keywords*: traffic, QoS parameters, cluster of servers, predictive control, neural network, simulation.*

И.В. Дегтяренко, А.А. Абраменко, А.С. Чекунков. Прогностическое управление нагрузкой на серверы с использованием нейросети. *Проведен анализ влияния загрузки серверов телекоммуникационной сети на параметры QoS. Усовершенствованы методы балансировки нагрузки на серверы кластера, благодаря внедрению прогностического управления с использованием нейросети. Предложены структура и алгоритм работы системы прогностического управления. Проведена оценка эффективности предложенных решений путем имитационного моделирования.*
Ключевые слова: *трафик, параметры QoS, кластер серверов, прогностическое управление, нейросеть, имитационное моделирование.*