Харківський національний університет радіоелектроніки

Кузьміних Євгенія Дмитрівна

УДК 621.391

МЕТОДИ ОБСЛУГОВУВАННЯ ВИКЛИКІВ НА SIP-СЕРВЕРАХ В УМОВАХ ВЕЛИКОГО НАВАНТАЖЕННЯ

05.12.02 – Телекомунікаційні системи та мережі

Автореферат дисертації на здобуття наукового ступеня
кандидата технічних наук

Харків – 2013

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Науковий керівник: доктор технічних наук, доцент

Агеєв Дмитро Володимирович,

Харківський національний університет радіоелектроніки, професор кафедри телекомунікаційних систем.

Офіційні опоненти: доктор технічних наук, професор

Бараннік Володимир Вікторович,

Науковий центр Повітряних сил Харківського університету Повітряних сил Міністерства оборони України, провідний науковий співробітник;

кандидат технічних наук, доцент

Акулинічев Артем Аркадійович,

Національний аерокосмічний університет
ім. М.Є. Жуковського "ХАІ" МОН України,
зам. декана факультета РТСЛА.

Захист відбудеться «__» _____ 2013 р. о _____ годині на засіданні спеціалізованої вченої ради Д 64.052.09 в Харківському національному університеті радіоелектроніки за адресою:
Україна, 61166, м. Харків, пр. Леніна, 14.

З дисертацією можна ознайомитися у бібліотеці Харківського національного університету радіоелектроніки за адресою:
Україна, 61166, м. Харків, пр. Леніна, 14.

Автореферат розісланий «____»_____2013 р.

Учений секретар
спеціалізованої вченої ради Є.В. Дуравкін

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Вимоги щодо зменшення вартості телефонії та інтеграція послуг класичної мережі загального користування PSTN з послугами мережі IP призвели до створення IP-телефонії. На даний час протокол SIP є базовим при побудові VoIP-мереж, використовується для миттєвої передачі повідомлень, відео та голосових конференцій. Провайдери безпроводового зв'язку теж використовуюсь SIP для побудови IP мультимедіа систем (IMS) стандарту 3GPP. SIP як найважливіший протокол для передачі сигнального трафіку по IP-мережі має декілька елементів, які відіграють велике значення для забезпечення надійності та масштабованості мережі. Один з таких елементів у мережі SIP – це сервер обслуговування викликів SIP проксі-сервер. Незважаючи на те, що SIP проксі-сервер відповідальний тільки за сигнальний трафік, і голосовий потік не проходить через цей сервер, у масштабних SIP-мережах навіть сигнальне навантаження може викликати перевантаження на сервері. Виникає проблема, пов'язана з контролем перевантажень на SIP-серверах, які виникають через відсутність достатніх ресурсів для встановлення і завершення сесій користувачів. Перевантаження SIP-сервера може стати причиною неможливості встановлення нових з’єднань, білінга та обслуговування поточних викликів.

У протоколі SIP (Session Initiation Protocol) існує метод боротьби з перевантаженням, відповідно до якого у разі перевантаження проксі-сервера передбачена генерація повідомлення 503 (Service Unavailable) та передача його на клієнтську сторону. Але у цьому методі є істотні недоробки, прикладами яких є проблема посилення перевантаження (load amplification) і проблема неповного використання кластера серверів (underutilization).

Перевантаження в мережі SIP може призвести до небажаних наслідків, таких як тривалі затримки і втрата викликів, що створює серйозну проблему для контролю якості функціонування мереж NGN (Next Generation Network). Проблема настільки істотна, що для її вирішення IETF вже підготував і прийняв ряд стандартів (RFC 5390, RFC 3665, RFC 6357), але ще більша частина документів знаходиться на стадії розгляду проектів. З аналізу методів управління та боротьби з перевантаженням на маршрутизаторах можна зробити висновок, що їх не завжди можна застосувати на серверах, які обробляють сигнальний трафік та виклики. Серед методів, які можуть бути застосовані на SIP-серверах, можна виділити методи боротьби з перевантаженням на основі моніторингу ресурсів SIP-сервера по завантаженню буфера повідомлень (Bang Bang Control), по завантаженню процесора (Occupancy Control) або по середній довжині черги повідомлень (Signalling RED). Але ці методи не враховують особливості функціонування протоколу SIP та обміну повідомленнями, зокрема, поточну фазу встановлення з'єднання і пріоритет користувача.

У зв'язку з цим у роботі сформульовано та розв’язано актуальну науково-прикладну задачу, яка полягає у підвищенні продуктивності мережі SIP в умовах великого навантаження за рахунок розробки і вдосконалення методів обслуговування черг на серверах обробки викликів та методів балансування навантаження у кластері SIP-серверів.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота безпосередньо пов'язана з реалізацією основних положень «Концепції національної інформаційної політики», «Концепції Національної програми інформатизації», «Концепції конвергенції телефонних мереж і мереж з пакетною комутацією в Україні» та «Основних засад розвитку інформаційного суспільства в Україні на 2007-2015 роки». Результати дисертаційної роботи були використані на кафедрі телекомунікаційних систем Харківського національного університету радіоелектроніки (ХНУРЕ) у науково-дослідній роботі «Методи проектування телекомунікаційних мереж NGN та управління їх ресурсами» (тема №235-1).

Мета роботи полягає у підвищенні продуктивності SIP-мереж шляхом збільшення пропускної здатності SIP-сервера та зменшення часу встановлення з’єднання в умовах великого навантаження.

Задачами дослідження є:

- аналіз існуючих методів боротьби з перевантаженням на серверах SIP;
- синтез математичної моделі процесу обслуговування викликів на SIP-сервері та оцінка адекватності розроблених моделей;
- розробка та удосконалення локальних методів боротьби з перевантаженням та методів балансування навантаження на SIP-сервері;
- оцінка ефективності запропонованих методів обслуговування викликів на SIP-серверах в умовах великого навантаження;
- розробка рекомендацій щодо практичного застосування методів боротьби з перевантаженням на SIP-серверах.

Об’єкт дослідження – процес обслуговування викликів на серверах обробки сигнальних повідомлень SIP.

Предмет дослідження – математичні моделі процесу обслуговування викликів на серверах обробки сигнальних повідомлень SIP, методи боротьби з перевантаженням на SIP-серверах, методи балансування навантаження у кластері SIP-серверів.

Методи дослідження. Проведені дослідження базуються на основних положеннях теорії масового обслуговування, теорії систем, теорії телетрафіку, імітаційному моделюванні розфарбованими мережами Петрі, статистичному аналізі.

Наукова новизна отриманих результатів. Під час розв’язання поставленої наукової задачі автором були отримані такі нові наукові результати:

1. Отримано подальший розвиток математичної моделі системи клієнт-сервер та системи обміну сигнальними повідомленнями між користувачем та SIP-сервером, новизна якої полягає у використанні апарату мереж Петрі та врахуванні полів заголовків повідомлень SIP-протоколу, що дозволило оцінити ймовірнісні характеристики SIP-сервера у різних режимах роботи, а також описати процеси сигнального обміну, що протікають в системі.

2. Вдосконалено метод обслуговування викликів на SIP проксі-серверах за рахунок врахування поточної фази встановлення з'єднання. Новизна полягає у тому, що в режимі перевантаження сигнальні повідомлення, які знаходяться у фазі діалогу, обслуговуються за принципом «останній прийшов - перший вийшов», що дозволило підвищити пропускну здатність сервера в умовах великого навантаження.

3. Вдосконалено локальні методи боротьби з перевантаженням на SIP-сервері за рахунок обслуговування повідомлень за схемою з динамічними пріоритетами, що дозволило зменшити час встановлення з'єднання.

4. Вдосконалено метод балансування навантаження в кластері SIP-серверів, новизна якого полягає у тому, що він враховує поділ сесій на транзакції і відносну вагу цих транзакцій, що дозволило поліпшити продуктивність кластера серверів і зменшити час встановлення з'єднання.

Практичне значення результатів роботи. Запропоновані методи обслуговування викликів на SIP-серверах в умовах великого навантаження можуть бути використанні при проектуванні нового та модернізації існуючого програмного забезпечення серверів обробки сигнальних повідомлень SIP, що дозволить підвищити якість SIP-систем, та не потребують зміни устаткування, бо зміни виконуються на програмному рівні.

Запропонований метод був застосований на серверному обладнанні провайдера IP-телефонії України, що підтверджується відповідним актом. Отримані в дисертаційній роботі результати були використані у навчальному процесі кафедри телекомунікаційних систем Харківського національного університету радіоелектроніки, зокрема, в дисципліні «Системи управління, сигналізації та синхронізації в ТКС», що підтверджується актом впровадження.

Особистий внесок здобувача. Дисертаційна робота виконана на кафедрі телекомунікаційних систем ХНУРЕ. Основні результати роботи належать особисто автору і повністю опубліковані у фаховій літературі [1–5]. У роботі [1], виконаній у співавторстві, особисто Кузьміних Є.Д. належить апробування застосування апарату розфарбованих мереж Петрі в телекомунікаційних мережах і обґрунтування використання цього апарату для систем реального часу.

Апробація результатів дисертації проводилася на 7 наукових конференціях і форумах: 11-му, 12-му і 16-му Міжнародному молодіжному форумі «Радіоелектроніка і молодь у XXI столітті »(м. Харків, 2007, 2008, 2012); I Всеросійській науково-технічній конференції (м. Туапсе, Росія, 2007); 9th and 11th In-

ternational Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science. TCSET (м. Львів-Славське, 2008, 2012); І Міжнародній науково-практичній конференції молодих вчених «Інфокомунікації - сучасність і майбутнє» (м. Одеса, 2011).

Публікації. Основні результати дисертаційної роботи опубліковані у 12 наукових працях, зокрема, у 5 наукових статтях [1-5] у наукових збірниках, що входять до переліку фахових видань, затверджених МОН України.

Структура дисертації. Дисертація складається зі вступу, чотирьох розділів та висновків. Загальний обсяг дисертаційної роботи становить 148 сторінок, у тому числі 46 рисунків, 18 таблиць, 115 бібліографічних джерел, викладених на 12 сторінках.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** розкрито загальний стан проблем, які виникають на сервері обробки викликів в умовах великого навантаження у мережі SIP, обґрунтовано актуальність теми досліджень, сформульовано мету та наукові задачі дослідження, визначено наукову новизну та практичне значення отриманих у роботі результатів.

У **першому розділі** на підставі аналізу мережі SIP показано, що сервери обробки викликів SIP є вразливими елементами в умовах великих навантажень і схильні до перевантажень з різних причин. Серед основних причин їх перевантаження можна виділити наступні: випадкові сплески трафіку, відмова устаткування, навмисні дії зловмисників при DoS (Denial of Service) атаках та інші. Результати аналізу існуючих методів боротьби з перевантаженням на серверах дозволили констатувати, що застосування локальних методів боротьби с перевантаженням на основі моніторингу ресурсів сервера дозволяє підвищити продуктивність сервера в порівнянні з закладеним в протоколі SIP механізмом 503 (Service Unavailable). Водночас, існуючи методи не враховуюсь особливості сигнального трафіку SIP.

Наведено огляд локальних методів боротьби з перевантаженням. Зроблено аналіз методів на основі моніторингу ресурсів SIP-сервера на прикладі методів по завантаженню буфера повідомлень (Bang Bang Control), по завантаженню процесора (Occupancy Control), по середній довжині черги повідомлень (Signalling RED). Зроблено висновок про те, що ці методи не враховують особливості функціонування протоколу SIP та обміну повідомленнями, зокрема, поточну фазу встановлення з'єднання і пріоритет користувача.

Наведено аналіз алгоритмів балансування навантаження на серверах SIP та зроблено висновок про те, що існуючі алгоритми при розподілі навантаження не беруть до уваги ієрархічну структуру сесій та поділ сесій на транзакції.

Показано, що основним напрямком щодо покращення ефективності функціонування проксі-сервера у режимі великого навантаження є перегляд і удо-

сконалення існуючих математичних моделей процесу обробки викликів на SIP-серверах, а також оптимальне балансування навантаження у кластері SIP-серверів. Сформульовано наукову задачу та здійснено її декомпозиції на окремі задачі дослідження.

У **другому розділі** розроблено метод, який враховує особливості обміну сигнальними повідомленнями для встановлення, підтримки і роз'єднання сесій за протоколом SIP. Розроблений метод спрямований на підвищення продуктивності SIP-сервера в умовах великого навантаження. Він враховує фазу встановлення з'єднання та обслуговує повідомлення у чергах з динамічними пріоритетами.

Структурна схема запропонованого методу обслуговування викликів на сервері SIP наведена на рис.1. Він реалізується двома паралельними процесами: процесом обробки повідомлень та процесом збору даних щодо завантаження ресурсів сервера у блоці контролю перевантаження.

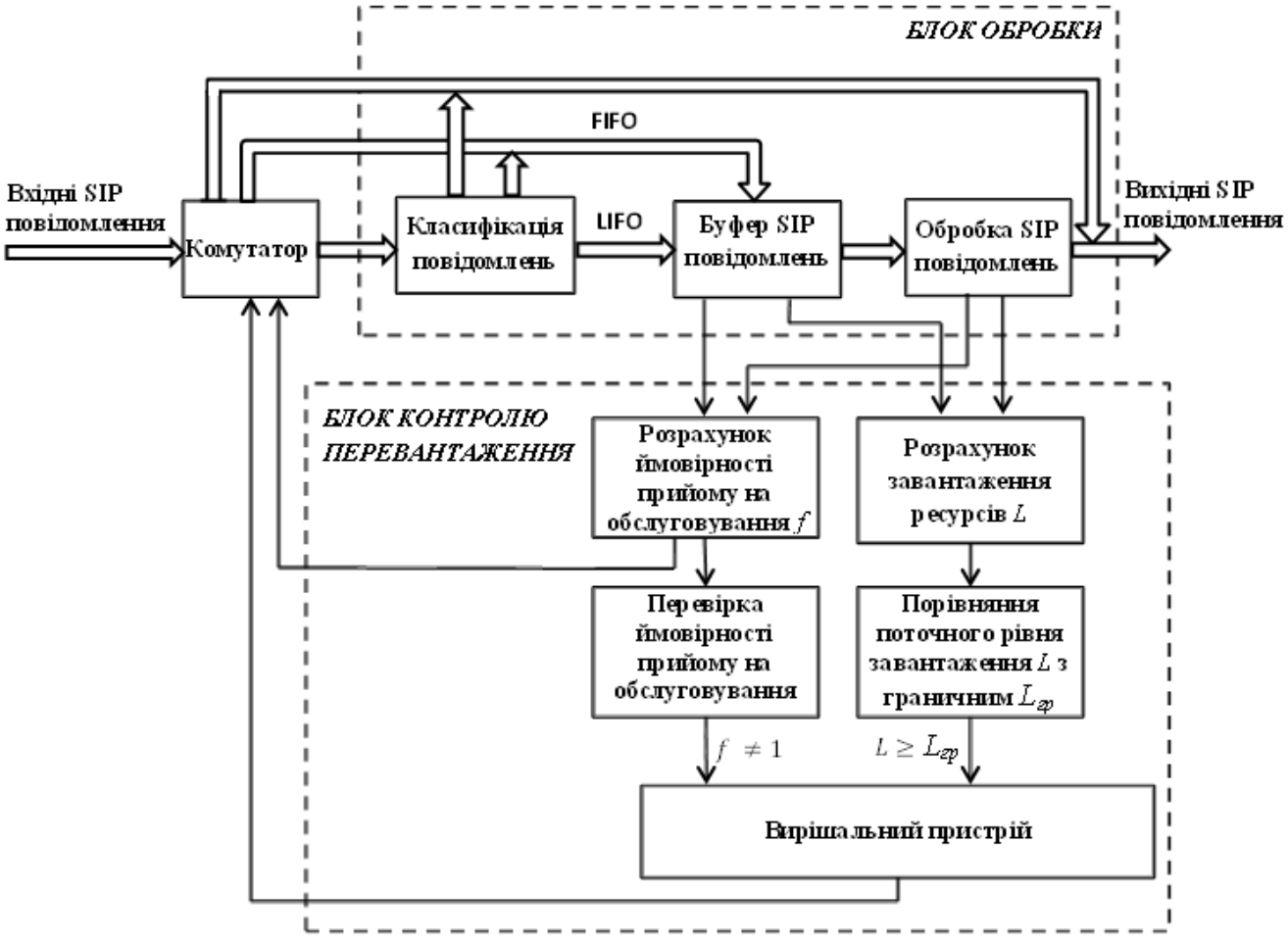


Рис.1. Структурна схема методу обслуговування викликів на сервері SIP

Запропонований метод враховує такі особливості, як:

- розподіл повідомлень за транзакціями (INVITE та не-INVITE транзак-

ції);

- трифазний обмін повідомленнями: запит-відповідь-підтвердження (INVITE-OK-ACK);

- систему таймерів ретрансляції для надійної доставки повідомлень.

Згідно зі схемою на рис. 1, після того як навантаження перевищило граничне значення $L_{гр}$ або імовірність прийому повідомлень на обслуговування $f_t \neq 1$, SIP-сервер переходить у стан перевантаження. У режимі перевантаження повідомлення спочатку потрапляють на класифікатор (рис. 2), в якому вони розподіляються за типом транзакції на INVITE чи не-INVITE і розподіляються по чергах у залежності від пріоритету. Для вибору повідомлення з черги було запропоновано використовувати динамічний пріоритет, тому що статичний може привести до недостатнього обслуговування черг з низьким пріоритетом.

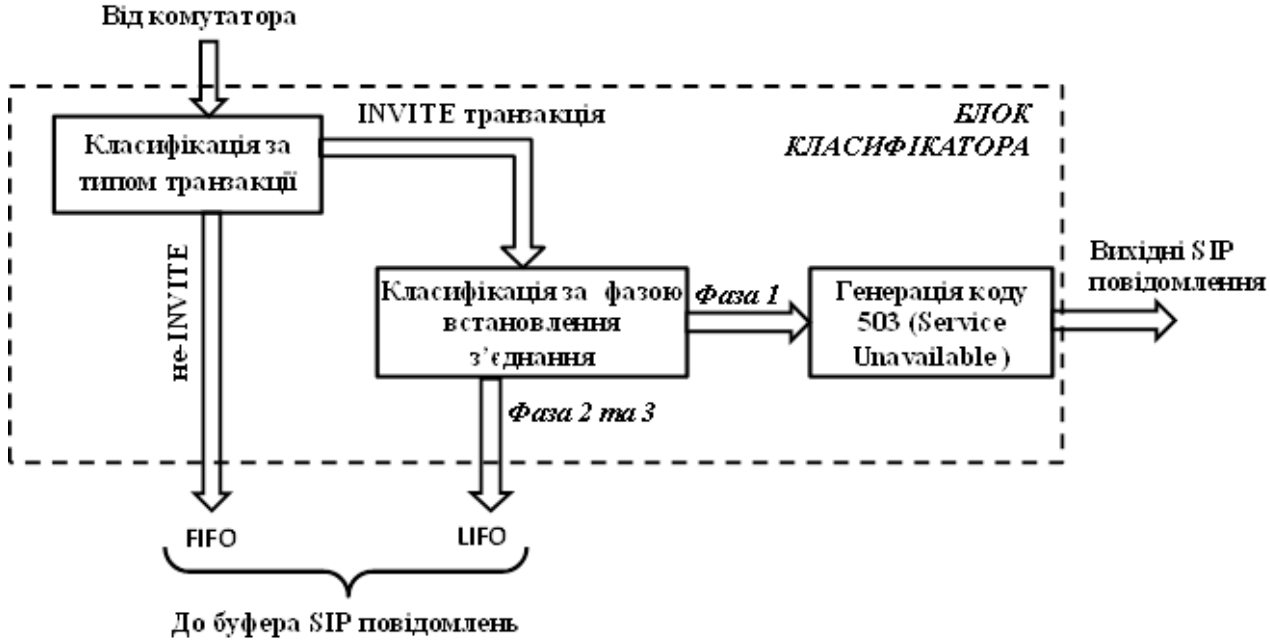


Рис.2. Блок класифікатора повідомлень

Для вибору черги використовуємо два параметри, які розраховуються для кожної черги *i =1, 2..k*: кількість запитів, які очікують у черзі (N_i) та коефіцієнт значущості черги (U_i). Тоді пріоритет черги можна визначити як N_i xU_i. Коефіцієнт значущості черги вибирається у залежності від важливості повідомлень, що знаходяться у цій черзі. Значення цього коефіцієнта може бути отримано зі статистичних даних, що описують запити користувачів, може призначатися адміністратором або залежати від сервера призначення.

Повідомлення, що належать INVITE-транзакції фази «у діалозі», обслуговуються за принципом LIFO (Last In First Out), всі інші – за принципом FIFO (First In First Out). Обслуговування за цим принципом отримало назву LIFO-PRIO.

Процес відхилення повідомлень реалізується детермінованою схемою прийняття рішення про відмову в обслуговуванні. Ця схема вперше застосовується у процесі обслуговування сигнальних повідомлень та у методах боротьби з перевантаженням на SIP-сервері. Рішення прийняти або відхилити повідомлення приймається за алгоритмом, наведеним на рис. 3, де f - ймовірність прийняття повідомлення на обслуговування, отримана згідно з методом боротьби з перевантаженням.

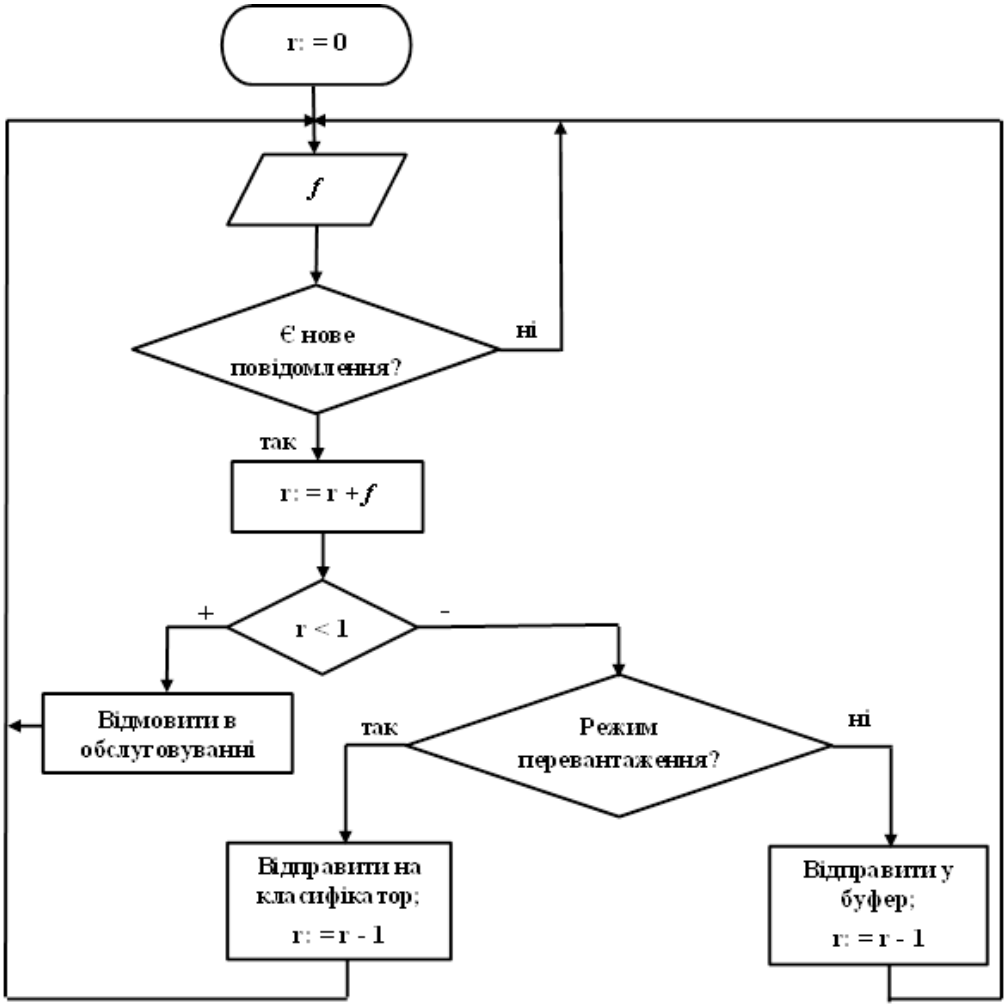


Рис.3. Детермінована схема прийняття повідомлень на обслуговування

Запропонований у дисертаційній роботі метод обслуговування викликів на SIP-сервері з метою аналізу його ефективності був застосований до методів боротьби з перевантаженням, які були локальними і базувались на моніторингу ресурсів сервера.

Першим методом для аналізу був метод BBC (Bang Bang Control), заснований на моніторингу завантаження буфера як індикатора перевантаження.

Другим методом для аналізу був метод OCC (Occupancy Control), заснований на моніторингу завантаження процесора сервера як індикатора перевантаження. Алгоритм роботи методу OCC представлений нижче:

$$f_{t+1} = \begin{cases} f_{min}, & \text{якщо } \phi_t f_t \leq f_{min} \\ 1, & \text{якщо } \phi_t f_t \geq 1 \\ \phi_t f_t, & \text{якщо } f_{min} \leq \phi_t f_t \leq 1 \end{cases}, \quad (1)$$

де f_t , f_{t+1} – імовірність прийому повідомлень на обслуговування в момент часу *t* та *t+1* відповідно, f_{min} – поріг мінімальної частки трафіку, який можна прийняти на обслуговування. Мультиплікативний коефіцієнт зміни (збільшення / зменшення) ϕ_t знаходиться з формули

$$\phi_t = min\{U_{\text{пор}}/U_t, \phi_{max}\}, \quad (2)$$

де $U_{\text{пор}}$ - граничне значення завантаження процесора сервера, U_t – завантаження процесора у момент часу *t*, ϕ_{max} визначає максимально можливе збільшення ймовірності *f* в період часу між моментами *t* та *t +1*.

Третім методом для аналізу був метод SiRED (Signalling RED), заснований на моніторингу середньозваженої довжини черги у буфері сервера як індикатора перевантаження, яка визначається методом ковзного середнього за формулою

$$Q_{t+1} = (1 - w)Q_t + wq_t , \quad (3)$$

де Q_t , Q_{t+1} - середньозважена довжина черги у момент часу *t* та *t+1* відповідно, q_t - поточне значення довжини черги, *w* - ваговий коефіцієнт.

Імовірність скидання повідомлення визначається формулою

$$f_{t+1} = \begin{cases} f_{min}, & \text{якщо } Q_t \geq Q_{max} \\ 1, & \text{якщо } Q_t \leq Q_{min} \\ \max\left(f_{min}, \dfrac{Q_{max} - Q_t}{Q_{max} - Q_{min}}\right), & \text{якщо } f_{min} \leq Q_t \leq 1 \end{cases}, \quad (4)$$

де Q_{max} та Q_{min} – максимальний та мінімальний пороги відкидання повідомлень, f_{min} - поріг мінімальної частки трафіку, який можна скинути.

Якщо $f_{min} \leq Q_t \leq 1$, то запити на встановлення нового з'єднання INVITE (фаза 1) відкидаються відповіддю 503 з імовірністю (1 - f_t); запити INVITE, які вже обробляються сервером (фаза 2 та 3), обслуговуються у порядку LIFO-PRIO, а повідомлення не-INVITE транзакцій – у порядку FIFO-PRIO.

В дисертаційній роботі запропоновано метод балансування навантаження між серверами SIP, який обирає сервер, куди слід направити виклик, за меншим числом транзакцій, що обробляє сервер на даний момент часу. Зазвичай балансування навантаження для SIP-серверів використовувало такі показники завантаження сервера, як завантаження процесора, доступність ресурсів пам’яті, се-

редній час відгуку сервера тощо. Для нового методу показником завантаження виступає кількість активних транзакцій на сервері. Лічильники транзакцій на сервері балансування навантаження зберігають інформацію про те, скільки на кожному SIP-сервері активних транзакцій. Коли запит на встановлення нового з'єднання INVITE надходить до сервера балансування навантаження, він направляється на сервер з найменшим числом транзакцій, і лічильник для цього сервера збільшується. Також при балансуванні навантаження враховується вага кожної транзакції

Рівень завантаження E_i для *i*-го SIP-сервера залежить від завантаження ресурсів цього сервера:

$$E_i = 1 + \alpha - \sum_{j=1}^{k} r_j \{r_1, r_2 .. r_k\}^{-1} \begin{bmatrix} L\,(P_{1i}) \\ L\,(P_{2i}) \\ \dots \\ L\,(P_{ki}) \end{bmatrix}^{-1}, \quad (5)$$

де α – дуже маленьке число, наприклад 0,0001; r_j – зважені параметри ресурсів, за якими проводиться моніторинг завантаження сервера ($j = 1, 2 .. k$, де k – число параметрів). Зважені параметри ресурсів конфігуруються адміністратором в залежності від параметрів сервера, його функцій та продуктивності. $L\,(P_{ki})$ - завантаження *k*-го ресурса *i*-го SIP-сервера. Для запропонованого методу для визначення завантаження *i*-го сервера за кількістю транзакцій на ньому з урахуванням відносної ваги цих транзакцій маємо:

$$E_i = 1 + \alpha - \; r_i \,^{-1} \left[\sum_{j=1}^{n} L_{ji}(c_j, T_j) \right]^{-1}, \quad (6)$$

де c_j – відносна вага *j*-ої транзакції T_j ($j = 1, 2 .. n$, де n – число типів транзакцій, наприклад, INVITE, BYE, OPTION, REGISTER транзакції для мережі SIP).

Цей метод дозволяє уникнути зворотного зв’язку та постійного обміну інформацією між серверами кластера та сервером балансування навантаження щодо фізичних показників завантаження ресурсів, а базується лише на показаннях лічильників транзакцій на сервері балансування.

У **третьому розділі** запропонована модель процесу обслуговування викликів на сервері у формі розфарбованих мереж Петрі та проведено дослідження розробленої моделі, зроблена перевірка її адекватності.

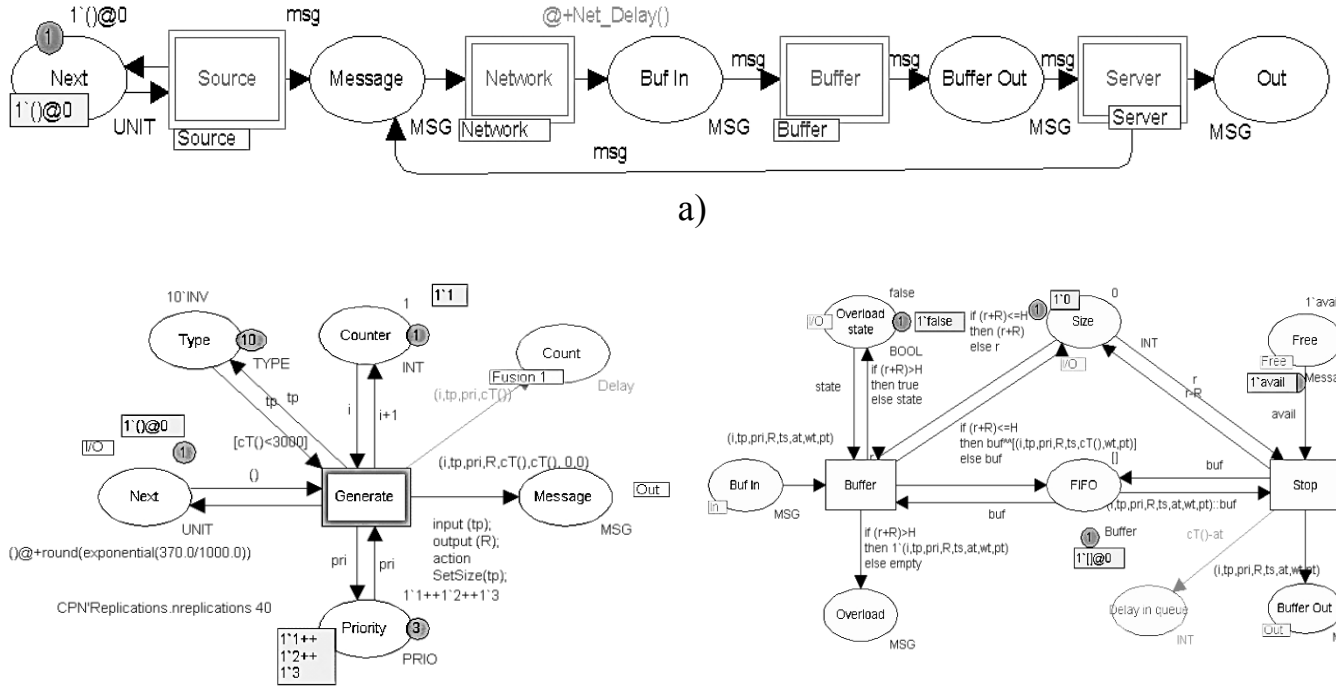


Рис.4. Модель обслуговування викликів на сервері SIP

Модель являє собою ієрархічну мережу Петрі, яка складається з двох рівнів та п’яти окремих блоків. На рис. 4,а зображено верхній рівень ієрархії моделі, до якого входять блок генератора трафіку *Source*, представлений у розгорнутому вигляді на рис. 4,б; ділянка мережі *Network*, що вносить затримку; блок буфера сервера *Buffer*, в якому відбувається розподіл повідомлень по чергах (рис. 4,в); блок *Server*, в якому відбувається обслуговування повідомлень і генерація відповіді, якщо це необхідно.

Позиція моделі Петрі *Next* задає закон розподілу надходження повідомлень в мережу. Трафіком можна управляти, задаючи йому різні закони розподілу, варіюючи інтенсивність і процентне співвідношення запитів-відповідей.

У ході перевірки адекватності моделі для порівняння результатів було зібрано сигнальний трафік з діючої мережі одного з найбільших українських VoIP операторів. Зібраний трафік був різного типу, серед його повідомлень можна виділити повідомлення на процедуру встановлення з'єднання, на реєстрацію/перереєстрацію, на завершення з'єднання. За результатами аналізу було встановлено, що 70% всього сигнального трафіку займають повідомлення на встановлення з'єднання, тому основну увагу при проведенні досліджень буде приділено саме цій процедурі. Була отримана послідовність сигнальних повідомлень на реальній мережі, зафіксовано час приходу запитів та час відправлення відповідей. З даних моніторингу проксі-сервера під час обробки повідомлень були отримані дані по завантаженню процесора і середньому часу обробки

сигнальних повідомлень на сервері. У результаті час обробки повідомлень INVITE у 4-5 разів перевищував час обробки інших повідомлень SIP. Це пов'язано з тим, що для аналізу полів заголовків повідомлення INVITE потрібно більше часу, зокрема, для аналізу поля URI SIP, яке вказує, куди слід надіслати запит, і для звернення до баз даних сервера.

Побудована модель була налагоджена й протестована в покроковому режимі імітації динаміки мережі Петрі. Під час перевірки адекватності моделі вхідними даними моделі була реалізація вхідного трафіку, отриманого на реальній мережі, а вихідними – дані по параметрам завантаження процесора сервера і затримці у встановленні з'єднання, які моніторилися у блоці *Server*. З метою забезпечення точності результатів, число експериментів було обрано таким чином, щоб забезпечити точність у 0,5 відсотків при довірчому інтервалі 0,9. Розбіжність між результатами імітаційного моделювання і даними, отриманими на реальній мережі оператора інтернет-телефонії, склала в середньому: за параметром завантаження процесора - 0,5%, за параметром середнього часу встановлення з'єднання – 1,3%.

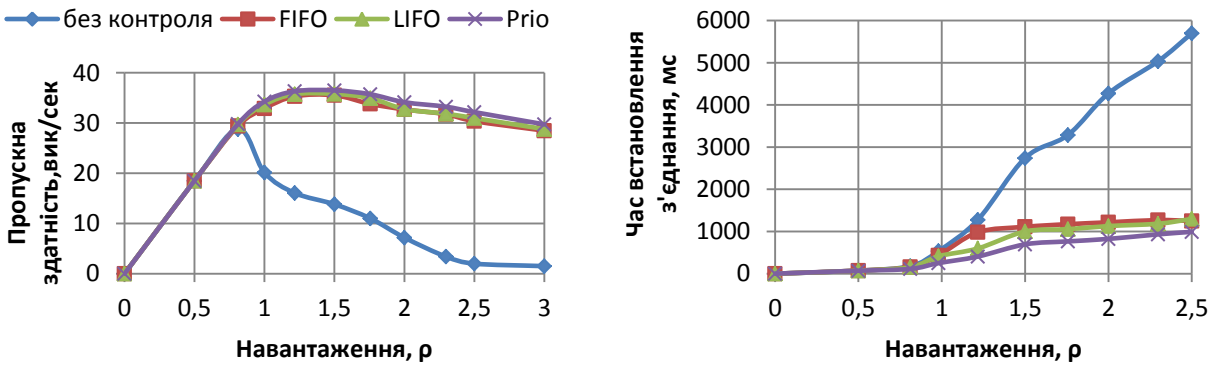
У **четвертому розділі** була проведена оцінка запропонованого методу обслуговування викликів на сервері SIP та дослідження продуктивності SIP-сервера в умовах великого навантаження, розроблено рекомендації щодо практичного використання результатів роботи.

У цьому розділі було обґрунтовано вибір показників ефективності SIP-сервера, таких як час встановлення з'єднання в мережі та доля встановлених з'єднань. Час встановлення з'єднання складається із затримки в черзі буфера сервера, затримки на обробку повідомлення процесором і з затримки передачі по мережі. Відношення кількості встановлених з'єднань до загальної кількості запитів на встановлення з'єднання є складовою частиною ефективної пропускної здатності сервера.

За результатами порівняльного аналізу запропонованого та існуючих методів обслуговування викликів на сервері SIP було зроблено висновок, що новий метод обслуговування викликів на 10-16% зменшує затримку в черзі буфера, в 1,5 рази зменшує затримку встановлення з'єднання в режимі перевантаження SIP-сервера та надає виграш по пропускній спроможності сервера SIP в режимі перевантаження у 2-4% в порівнянні з роботою інших методів боротьби з перевантаженням. У порівнянні з обслуговуванням повідомлень без застосування методів боротьби з перевантаженням час встановлення з'єднання із застосуванням таких методів зменшується у 5 - 6 разів, а пропускна здатність зростає на 40%.

Залежності пропускної здатності сервера та часу встановлення з'єднання в мережі SIP від інтенсивності сигнального трафіку для методу боротьби з перевантаженням на основі двохпозиційного регулювання BBC (Bang Bang Control) показано на рис. 5. Для методу з контролем завантаженості процесора OCC

(Occupancy Control) та для методу на основі алгоритму раннього виявлення перевантаження SiRED (Signalling RED) були отримані аналогічні результати.



а) б)

Рис. 5. Залежність пропускної здатності сервера (а) та часу встановлення з’єднання (б) від вхідного навантаження

Крім того, було досліджено вплив на продуктивність SIP-сервера таких параметрів, як інтервал моніторингу завантаження процесора та граничні значення порогів для скидання повідомлень згідно з методами боротьби з перевантаженням. Найкращі результати по пропускній здатності і часу встановлення з'єднання дає інтервал моніторингу в 200-250 мс або половина значення таймера першої повторної передачі повідомлень ½ T_A. Розкид порогів у межах 20-30% незначно впливає на ефективність роботи сервера, але при низьких значеннях мінімального порогу скидання повідомлень спостерігається недовикористання ресурсів процесора і менше число успішних з'єднань.

Метою дослідження ефективності роботи серверів обслуговування викликів також було вивчення того, як швидко методи боротьби з перевантаженням на SIP-сервері можуть реагувати на раптову появу перевантаження і раптове припинення перевантаження. З результатів проведеного порівняльного аналізу було зроблено висновок, що застосування нового методу, який враховує фазу встановлення з’єднання та пріоритет повідомлення, дозволяє збільшити швидкість реакції на припинення перевантаження у 2-3 рази у порівнянні з методами, які не враховують ці особливості.

На рис. 6 представлена динаміка показників часу встановлення з'єднання при використанні методу з контролем завантаженості процесора OCC при нестаціонарному навантаженні, де у період часу з 10000мс до 30000мс спостерігається різке збільшення навантаження.

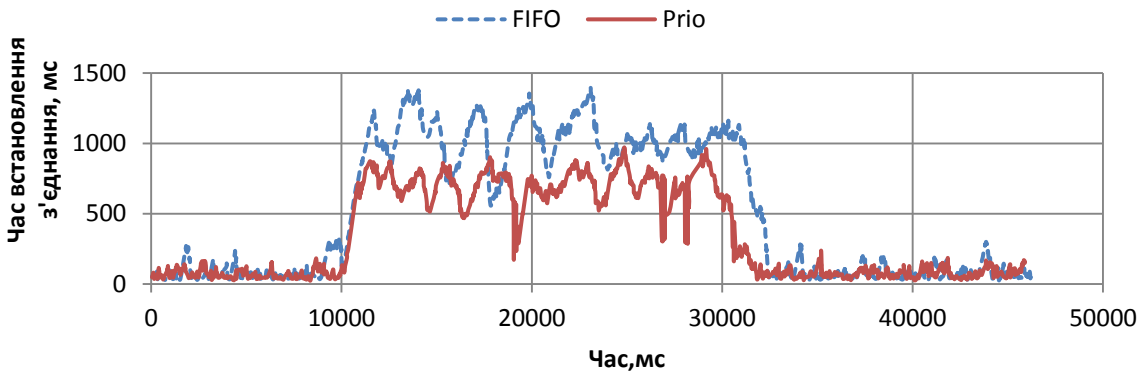
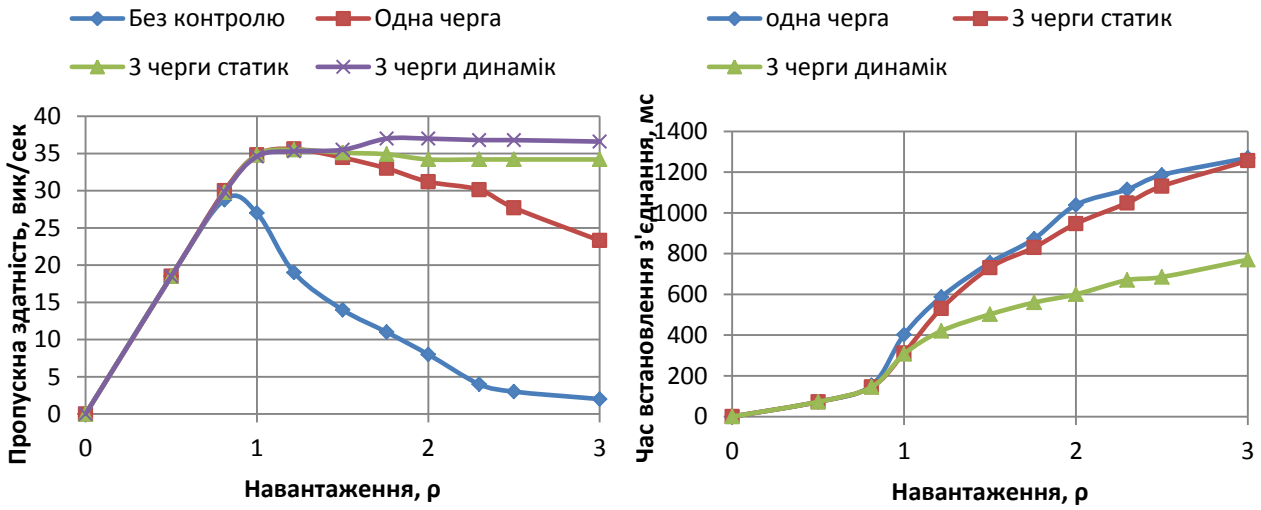


Рис. 6. Динаміка часу встановлення з'єднання при нестаціонарному навантаженні

В роботі досліджено ефективність обслуговування повідомлень з динамічним пріоритетом, в якому враховується коефіцієнт значущості черги. Результати проведеного дослідження у вигляді графів залежності показано на рис. 7.



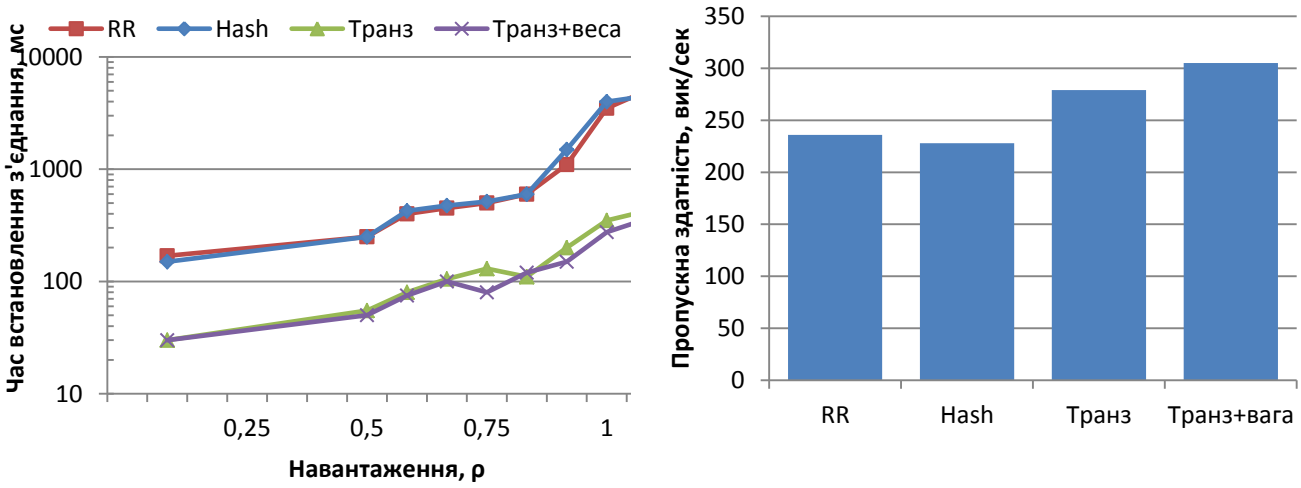
а) б)

Рис. 7. Залежність пропускної здатності сервера (а) та часу встановлення з’єднання (б) від вхідної навантаження

З результатів можна зробити висновок, що застосування обслуговування повідомлень у чергах з динамічним пріоритетом при обробці викликів на серверах SIP є більш ефективним, ніж обслуговування у чергах з фіксованою довжиною. У порівнянні з обслуговуванням в одній черзі або в декількох, але з фіксованою довжиною, обслуговування з динамічним пріоритетом в 1,6 разів зменшує затримку встановлення з'єднання в умовах великого навантаження. Виграш по пропускній спроможності сервера SIP в умовах великого навантаження з динамічним обслуговуванням у чергах становить 3 - 6% порівняно з обслуговуванням в статичних чергах. Цей метод можна застосовувати на серве-

рах обробки повідомлень SIP, які одночасно виконують функції декількох елементів SIP-мережі і потребують формування черги до кожного з цих логічних елементів. Наприклад, окремі черги до сервера реєстрації, до сервера визначення місцезнаходження та до проксі-сервера SIP.

У роботі було проведено дослідження методів балансування навантаження на серверах SIP. У результаті метод балансування, що розподіляє навантаження між серверами за показником активних транзакцій і враховує відносну вартість цих транзакцій, дав найкращі показники за часом встановлення з'єднання і пропускною здатністю кластеру, що відображено на рис. 8.



а) б)

Рис. 8. Залежність часу встановлення з'єднання (а) від навантаження та пропускна здатність кластеру серверів (б) при піковому навантаженні

У результаті дослідження доведено, що запропонований метод має час встановлення з'єднання в умовах великого навантаження до 10 разів менший, ніж інші підходи, а в умовах середнього навантаження - у 5-6 разів менший. Використання цього методу для балансування навантаження на 25% забезпечує кращу продуктивність серверів кластеру, ніж стандартний метод, заснований на hash-алгоритмі; і на 14% – кращу продуктивність, ніж циклічний розподіл за алгоритмом Round Robin (RR).

ВИСНОВКИ

У дисертаційній роботі вирішена актуальна науково-прикладна задача, яка полягає у підвищенні продуктивності мережі SIP в умовах великого навантаження за рахунок розробки і вдосконалення методів обслуговування черг у буфері серверів обробки викликів та методів балансування навантаження у кластері SIP-серверів. За підсумками вирішення поставлених наукової та прикладної задач зроблено наступні висновки:

1. Показано, що на даному етапі розвитку мереж IP-телефонії сервери SIP є вразливим місцем у мережі і можуть бути перевантажені за рядом причин, таких як випадкові сплески трафіку, відмова устаткування, навмисні дії зловмисників при DoS (Denial of Service) атаках та інші. Існуючий метод боротьби з перевантаженням у протоколі SIP не справляється зі своїм завданням і може навіть погіршити стан перевантаження. У зв'язку з цим було проведено аналіз методів боротьби з перевантаженням на серверах, і зроблено висновок про те, що застосування цих методів дозволяє підвищити продуктивність сервера в 2-5 разів у порівнянні з закладеним в протоколі SIP механізмом.

2. У результаті проведення аналізу локальних методів боротьби з перевантаженням було виявлено, що підвищити продуктивність SIP-серверів в умовах великого навантаження можна за рахунок наступних методів та схем: 1) змінити порядок обслуговування повідомлень у черзі при перевантаженні з FIFO (First In First Out) на LIFO (Last In First Out), який враховує фазу встановлення з'єднання по протоколу SIP; 2) використовувати динамічну схему для вибору повідомлення з черги, яка заснована на довжині черги і коефіцієнті значущості повідомлення; 3) застосувати детерміновану схему прийняття рішення про відмову в обслуговуванні, яка дозволяє зменшити витрати ресурсів на генерацію випадкових чисел.

3. У результаті проведення аналізу методів балансування навантаження у мережі SIP було виявлено, що жоден з існуючих методів не враховує таких особливостей створення сесій по протоколу SIP, як поділ сесій на транзакції та відносна вартість цих транзакцій. Тому був розроблений новий метод балансування навантаження, який обирає сервер, куди слід направити виклик, за меншим числом активних транзакцій, що обробляє сервер на даний момент часу.

4. Побудовано математичну модель процесу обслуговування викликів на сервері SIP, що дало можливість оцінювати методи боротьби з перевантаженням на серверах SIP. Модель була побудована відповідно до рекомендації RFC 3261 за допомогою математичного апарату розфарбованих мереж Петрі. Розроблені блоки моделі, такі як генератор SIP-повідомлень, блок буфера сервера, блок обробки повідомлень, можна використовувати у якості компонентів для моделювання мереж передачі даних зі складною топологією, а також при дослідженні та проектуванні різного телекомунікаційного устаткування.

5. Зроблено збір статистичних даних трафіку сигнального протоколу SIP на діючій мережі крупного оператора IP-телефонії, аналіз яких показав, що 70% всього сигнального трафіку займають повідомлення на встановлення з'єднання. Були проаналізовані технічні характеристики SIP-прокси сервера і отримані дані по завантаженню процесора і середньому часі обробки сигнальних повідомлень на сервері. У результаті час обробки повідомлення INVITE в 4-5 разів перевищувало час обробки інших повідомлень SIP. Це пов'язано з тим, що для аналізу полів заголовків повідомлення INVITE потрібно більше часу, зокрема,

для аналізу поля URI SIP, яке вказує, куди слід надіслати запит, і для звернення до баз даних сервера.

6. Для оцінки адекватності побудованої моделі мережі SIP було проведено імітаційне моделювання з такими вхідними даними, які відповідали отриманим на реальній мережі. В результаті моделювання було отримано сигнальний трафік на вході сервера, що має властивості самоподібності, що характерно для реального сигнального трафіку. Розбіжність між результатами імітаційного моделювання та практичними дослідженнями склала у середньому: за параметром завантаження процесора - 0,5%, за параметром середнього часу встановлення з'єднання – 1,3%, що говорить про те, що модель обміну повідомленнями в мережі SIP у формі розфарбованих мереж Петрі є адекватною і може бути застосована для моделювання поведінки сервера в різних режимах і для оцінки методів боротьби з перевантаженнями та методів балансування на SIP-серверах.

7. У дисертації проведено порівняльний аналіз запропонованого і відомих методів обслуговування викликів на SIP-сервері з кількісною оцінкою основних показників якості обслуговування. У ході порівняльного аналізу було встановлено, що застосування принципу LIFO-PRIO в методі боротьби з перевантаженням з обслуговуванням повідомлень з урахуванням фази встановлення з'єднання в 1,4 рази зменшує затримку встановлення з'єднання в режимі перевантаження SIP-сервера та покращує пропускну здатність сервера на 2,6% у порівнянні з LIFO і 3,8% - в порівнянні з FIFO.

8. У ході досліджень встановлено, що час відновлення після перевантаження при застосуванні принципу LIFO-PRIO майже в 2 рази менше, ніж при застосуванні FIFO, при навантаженні в 2 рази більшим, ніж пропускна здатність сервера.

9. Результати дослідження методу балансування, що розподіляє навантаження між серверами за показником активних транзакцій і враховує відносну вартість цих транзакцій, показали, що застосування цього методу має кращі показники ефективності в умовах великого навантаження. Так, час встановлення з'єднання у 5-10 разів менший, ніж при застосуванні існуючих методів балансування за hash-алгоритмом чи за круговим принципом Round Robin. Показник ефективної пропускної здатності сервера при піковому навантаженні у кластері на 20% вищий, ніж при застосуванні інших методів балансування навантаження.

10. Розроблено рекомендації щодо практичної реалізації запропонованого методу обслуговування викликів на сервері SIP. Виконання запропонованих рекомендацій передбачають внесення додаткових параметрів конфігурування SIP-сервера в модулі контролю перевантаження. Запропоновано використання цього методу спільно з системою моніторингу мереж сигналізації та системами моніторингу ресурсів вузлів мережі SIP для забезпечення контролю стану мережі та фіксації перевантажень і поломок.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Лысяк Т.Н. Исследование методов профилирования трафика в условиях нестационарной загрузки / Т.Н. Лысяк, Е.Д. Кузьминых // Радиотехника: Всеукр. межвед. науч.-техн. сб. – 2007. – Вып. 148. – С.128-134.

2. Кузьминых Е.Д. Борьба с отказами и перераспределение нагрузки на основе протокола SIP в IP-телефонии / Е.Д. Кузьминых // Радиотехника: Всеукр. науч.-техн. сб. — 2007. — Вып.151. – С. 57-64.

3. Кузьминых Е.Д. Модель процедуры управления соединением в протоколе SIP в форме раскрашенных сетей Петри / Е.Д. Кузьминых // Радиотехника: Всеукр. науч.-технич. сб. — 2010. — Вып.163. – С. 92-98.

4. Кузьминых Е.Д. Оценка параметров SIP-сети с использованием раскрашенных сетей Петри / Е.Д. Кузьминых // Проблемы телекоммуникаций. – 2012. – №1(6). – С. 20–40. – Режим доступа: http://pt.journal.kh.ua/2012/1/1/121_kuzminykh_sip.pdf

5. Кузьминых Е.Д. Метод борьбы с перегрузкой на SIP-сервере с учетом текущей фазы установления соединения и динамических приоритетов / Е.Д. Кузьминых // Проблемы телекоммуникаций. – 2012. – № 2 (7). – С. 68 – 77. – Режим доступа: http://pt.journal.kh.ua/2012/2/1/122_kuzmenyh_sip.pdf

6. Кузьминых Е.Д. Процесс установления соединения в сетях следующего поколения / Е.Д. Кузьминых // 11-й Междунар. молод. форум «Радиоэлектроника и молодежь в XXI веке»: Сб. мат-лов. – Х.: ХНУРЭ, 2007. – С. 78.

7. Кузьминых Е.Д. Процесс перераспределения ролей основного и дублирующего SIP-серверов при отказе основного в IP-телефонии / Е.Д. Кузьминых, Д.В. Агеев // I Всеросс. науч.-техн. конф.: Сб.тезисов. - Туапсе, 2007. - С.113.

8. Kuzminykh I. Failover and Load Sharing in SIP-based IP telephony / I.Kuzminykh // 9th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET’2008): Сб. материалов. – Львов-Славское, 2008. – С. 420-422.

9. Кузьминых Е.Д. Модель работы прокси-сервера в SIP-сетях / Е.Д.Кузьминых // 12-й Междунар. молодежный форум «Радиоэлектроника и молодежь в XXI веке»: Сб. материалов – Х.: ХНУРЭ, 2008. – С.117.

10. Кузьминых Е.Д. Модель обслуживания вызовов на SIP-сервере в условиях перегрузки / Е.Д. Кузьминых // I Международная научно-практич. конференция молодых ученых «Инфокоммуникации-современность и будущее»: Сб. материалов. – Одесса: ОНАС им. А.С. Попова, 2011. – С. 80-83.

11. Kuzminykh I. A Combined LIFO-Priority Algorithm for Overload Control of SIP Server / I. Kuzminykh // 11th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science: Сб. материалов. – Львов-Славское: НУ «Львовская политехника», 2012. – С. 330.

12. Кузьминых Е.Д. Локальные методы борьбы с перегрузкой на SIP-сервере / Е.Д. Кузьминых // 16-й Междунар. молодежный форум «Радиоэлектроника и молодежь в XXI веке»: Сб. материалов. – Х.: ХНУРЭ, 2012. – С. 64-65.

АНОТАЦІЯ

Кузьміних Є.Д. Методи обслуговування викликів на SIP-серверах в умовах великого навантаження. – Рукопис. Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.12.02 – Телекомунікаційні системи і мережі. – Харківський національний університет радіоелектроніки, Харків, 2013.

Дисертацію присвячено підвищенню ефективності функціонування мереж IP-телефонії в режимі перевантаження за показниками продуктивності сервера обслуговування викликів і часу встановлення з'єднання за рахунок розробки та вдосконалення методів обслуговування викликів на SIP-сервері та методів балансування навантаження.

Було проведено аналіз методів боротьби з перевантаженням і зроблено висновок про те, що застосування цих методів дозволяє підвищити продуктивність сервера у 2-5 разів в порівнянні з закладеним в протоколі SIP механізмом.

Для підвищення ефективності роботи методів обслуговування викликів на SIP-серверах в умовах великого навантаження розроблено сукупність методів, алгоритмів, схем і програмних засобів аналізу з використанням апарата мереж Петрі, які є науково обґрунтованою базою для боротьби з перевантаженням на сервері обслуговування викликів, а також аналізу таких характеристик, як час встановлення з'єднання; відсоток відмов у обслуговуванні; інтенсивність потоку викликів з необхідним пріоритетом.

Для підвищення продуктивності мережі SIP було розроблено метод балансування навантаження, який розподіляє виклики між серверами за числом активних транзакцій та враховує вартість цих транзакцій.

Ключові слова: SIP-сервер, методи боротьби з перевантаженням, балансування навантаження, транзакції, розфарбована мережа Петрі, час встановлення з'єднання, продуктивність.

АННОТАЦИЯ

Кузьминых Е.Д. Методы обслуживания вызовов на SIP-серверах в условиях высокой нагрузки. – Рукопись. Диссертация на соискание ученой степени кандидата технических наук по специальности 05.12.02 – Телекоммуникационные системы и сети. –Харьковский национальный университет радиоэлектроники, Харьков, 2013.

Диссертация посвящена повышению эффективности функционирования сетей IP-телефонии в режиме перегрузки по показателям производитель-

ности сервера обслуживания вызовов и времени установления соединения посредством разработки и усовершенствования методов борьбы с перегрузкой на SIP-сервере с учетом фазы установления соединения и приоритета сообщения.

В связи с этим был проведен анализ методов борьбы с перегрузкой и сделан вывод о том, что применение этих методов позволяет повысить производительность сервера в 2-5 раз по сравнению с заложенным в протоколе SIP механизмом.

Предложенный метод обслуживания вызовов на сервере отличается от существующих тем, что: 1) учитывает фазу установления соединения по протоколу SIP, использует алгоритм обслуживания сообщений в очереди FIFO при нормальной нагрузке и LIFO при перегрузке; 2) использует динамическую схему для выбора сообщения из очереди, которая основана на длине очереди и коэффициенте значимости сообщения; 3) использует детерминированную схему принятия решения об отказе в обслуживании, которая позволяет уменьшить расходы ресурсов на генерацию случайных чисел.

В результате проведения анализа методов балансировки нагрузки в сети SIP было обнаружено, что ни один из существующих методов не учитывает таких особенностей создания сессий по протоколу SIP, как разделение сессий на транзакции и относительную стоимость этих транзакций. Поэтому был разработан новый метод балансировки нагрузки, который выбирает сервер, куда следует направить вызов, по наименьшему числу активных транзакций, которые обрабатывает сервер в данный момент времени.

Для оценки методов борьбы с перегрузкой на серверах SIP была построена математическая модель обмена сообщениями в сети SIP в форме раскрашенных сетей Петри в имитационной среде CPN Tools. Новизна разработанных моделей состоит в том, что благодаря преимуществам выбранного математического аппарата в них одновременно отражены не только функциональные, но и структурные свойства элементов сети SIP. Более точно и детально рассмотрены алгоритмы обслуживания вызовов на серверах обработки вызовов. Разработанные блоки модели можно использовать в качестве компонентов для моделирования сетей передачи данных со сложной топологией, а также при исследовании и проектировании различного телекоммуникационного оборудования. Достоверность результатов подтверждается совпадением значений, полученных методами имитационного моделирования, со значениями, полученными на реальной сети, с точностью до 0,5..1,3%.

Произведен сбор статистических данных трафика сигнального протокола SIP на действующей сети крупного оператора IP-телефонии, анализ которых показал, что 70% всего сигнального трафика занимают сообщения на установление соединения.

Сравнительный анализ существующих и модифицированных методов борьбы с перегрузкой и методов балансировки нагрузки показал, что применение усовершенствованных методов позволяет уменьшить время установления соединения, повысить пропускную способность сервера в условиях высокой нагрузки, а также уменьшить время восстановления системы после перегрузки.

Полученные результаты использованы в учебном процессе ХНУРЭ, в частности, в дисциплинах «Системы управления, сигнализации и синхронизации в ТКС», реализованы на сервере обработки вызовов провайдера IP-телефонии. К практическим результатам также относятся имитационные модели SIP-сетей в графической системе CPN Tools, что позволяет проводить эксперименты не на специализированном сетевом оборудовании, а на обычных компьютерах.

Ключевые слова: SIP-сервер, методы борьбы с перегрузкой, методы балансировки нагрузки, транзакции, раскрашенные сети Петри, время установления соединения, производительность сервера.

ABSTRACT

Kuzminykh I.D. Methods of call processing by SIP servers under high load. - Manuscript. Thesis for candidate's degree of technical science by specialty 05.12.02 – Telecommunication systems and networks. – Kharkov National University of Radio-electronics, Kharkov, 2013.

The thesis is devoted to improvement the efficiency of IP-telephony performance under overload by SIP-proxy server throughput and setup delay values. This improvement can be reached by the development and improvement of SIP calls processing by the proxy-server and new load balancing method.

The analysis of local overload control methods showed that the application of these techniques can improve server performance by 2-5 times in comparison with SIP protocol overload control mechanism.

For improvement the efficiency of methods of call processing by SIP servers under high load a system of methods, algorithms, schemes and software is developed using coloured Petri Net tool. This system represents a science-based groundwork for SIP server congestion avoidance and for evaluation of such characteristics as setup delay; service deny ratio; flow intensity of calls served with required priority.

Comparative analysis of the proposed and existing overload control and load balancing methods under high load showed that proposed methods allow increase SIP server throughput, decrease setup delay and reduce recovery time after a system overload.

Keywords: SIP server, overload control methods, load balancing, transaction, colored Petri nets, setup delay, server throughput.