



УКРАЇНА

(19) **UA** (11) **77142** (13) **U**
(51) МПК
H04L 12/56 (2006.01)
H04N 21/647 (2011.01)

ДЕРЖАВНА СЛУЖБА
ІНТЕЛЕКТУАЛЬНОЇ
ВЛАСНОСТІ
УКРАЇНИ

(12) ОПИС ДО ПАТЕНТУ НА КОРИСНУ МОДЕЛЬ

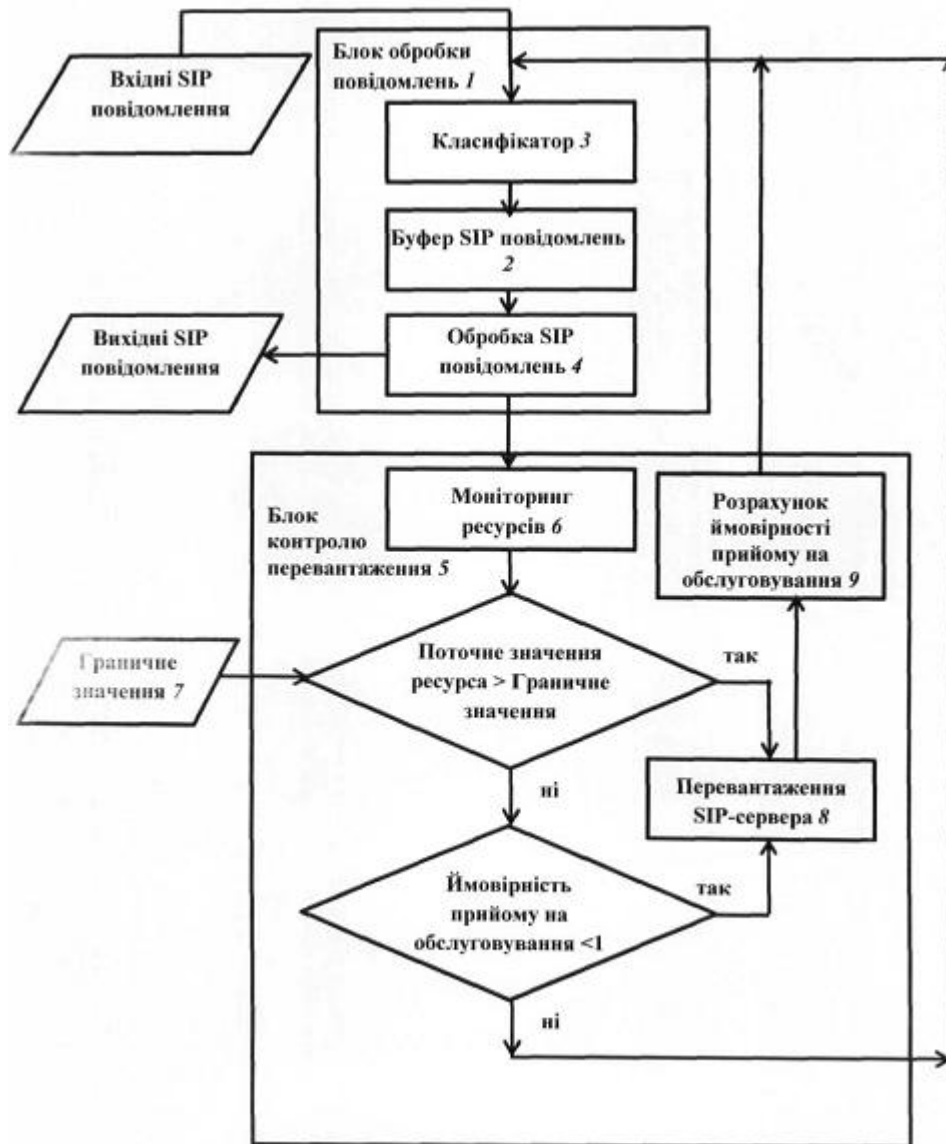
(21) Номер заявки:	u 2012 10038	(72) Винахідник(и):	Кузьмініх Євгенія Дмитрівна (UA), Агеєв Дмитро Володимирович (UA)
(22) Дата подання заявки:	21.08.2012	(73) Власник(и):	ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ, пр. Леніна, 14, м. Харків, 61166 (UA)
(24) Дата, з якої є чинними права на корисну модель:	25.01.2013		
(46) Публікація відомостей про видачу патенту:	25.01.2013, Бюл.№ 2		

(54) СПОСІБ БОРОТЬБИ З ПЕРЕВАНТАЖЕННЯМ НА СЕРВЕРІ SIP

(57) Реферат:

Спосіб боротьби з перевантаженням на сервері SIP включає виявлення перевантаження за завантаженням ресурсів сервера та реакцію на перевантаження з пробалістичною схемою регулювання відмовами. При цьому у режимі перевантаження сервер SIP обробляє вхідні повідомлення, враховуючи фазу встановлення з'єднання, використовує динамічний пріоритет при обслуговуванні повідомлень у черзі буфера та детерміновану схему прийняття рішення про відмову в обслуговуванні.

UA 77142 U



Фіг. 1

Корисна модель належить до області програмно-алгоритмічних рішень для боротьби з перевантаженням у мережах IP-телефонії та може бути використана як локальний механізм боротьби з перевантаженням на SIP-серверах обробки сигнальних повідомлень спільно з іншими методами та механізмами виявлення та боротьби з перевантаженням.

Відомий спосіб боротьби з перевантаженням на сервері SIP, зазначений у рекомендації RFC 3261 [1], передбачає базовий механізм, який використовує код 503 (Service Unavailable). У випадку, якщо SIP-сервер не може обробити запит через тимчасове перевантаження, він відхиляє запит з кодом помилки 503. Це повідомлення означає, що відправник повинен повторити відправку повідомлення через деякий час або направити запит на інший сервер. Перевантажений сервер також може додати Retry-After заголовок до SIP-повідомлення, вказавши час у секундах, протягом якого на сервер не слід звертатися. Відправник припиняє відправку повідомлень на цей період часу, повідомлення направляються на альтернативний сервер, а відправка повідомлень на перевантажений сервер поновлюється після закінчення інтервалу часу, зазначеного в заголовку. Якщо альтернативний сервер відсутній і заголовок Retry-After не зазначений, то сервер SIP повинен відправити відповідь 500 (Server Internal Error) користувачеві, який припиняє повторну передачу запиту.

Під час впровадження та експлуатації даного способу було виявлено ряд недоліків: цей механізм може призвести до коливань трафіку, до зниження середньої пропускної здатності сервера і погіршити стан перевантаження через перенаправлення трафіку на інший проксі-сервер.

Відомий також вибраний як найближчий аналог спосіб боротьби з перевантаженням на сервері SIP, що включає виявлення перевантаження за завантаженням ресурсів сервера та реакцію на перевантаження з пробалістичною схемою регулювання відмовами [2].

Цей спосіб передбачає, що для виявлення перевантаження на сервері SIP вводяться граничні значення завантаження ресурсів сервера (центрального процесора, пам'яті сервера, довжини черги у буфері), за якими визначається тривалість режиму перевантаження. Як реакція на перевантаження вводиться ймовірність відмови в обслуговуванні кодом 503 для всіх повідомлень на встановлення з'єднання. Ймовірність відмови визначається за ступенем завантаження ресурсів і реалізується на сервері за допомогою пробалістичної схеми регулювання відмовами. Ця схема використовує генератор псевдовипадкових чисел для прийняття рішення про відмову в обслуговуванні, що призводить до додаткових відхилень у розрахованій ймовірності відмови в обслуговуванні і до витрати ресурсів сервера на генерацію випадкових чисел.

Цей спосіб або зовсім не враховує пріоритет повідомлень SIP, або використовує розподіл повідомлень за пріоритетом по чергах зі статичною довжиною, що призводить до недовикористання черг з низьким пріоритетом.

В основу корисної моделі поставлена задача: створити спосіб боротьби з перевантаженням на сервері SIP, який дозволив би враховувати фазу встановлення з'єднання, використовувати динамічний пріоритет обслуговування у черзі буфера та детерміновану схему прийняття рішення про відмову в обслуговуванні, що призвело б до зменшення часу встановлення з'єднання у мережі SIP, збільшення пропускної здатності сервера, зменшення витрати ресурсів на генерацію випадкових чисел у режимі перевантаження.

Поставлена задача вирішується тим, що в способі боротьби з перевантаженням на сервері SIP, що включає виявлення перевантаження за завантаженням ресурсів сервера та реакцію на перевантаження з пробалістичною схемою регулювання відмовами, згідно з корисною моделлю, у режимі перевантаження сервер SIP обробляє вхідні повідомлення, враховуючи фазу встановлення з'єднання, використовує динамічний пріоритет при обслуговуванні повідомлень у черзі буфера та детерміновану схему прийняття рішення про відмову в обслуговуванні.

Заявлений спосіб боротьби з перевантаженням на сервері SIP описується наступними кресленнями:

Фіг. 1. Загальний алгоритм управління перевантаженням на SIP-сервері.

Фіг. 2. Блок класифікатора повідомлень.

Фіг. 3. Блок моніторингу ресурсів.

Фіг. 4. Алгоритм детермінованої схеми регулювання відмовами.

Заявлений спосіб реалізується в алгоритмі управління перевантаженням на SIP-сервері наступним шляхом (Фіг. 1):

Крок 1. Кожне нове повідомлення надходить на SIP-сервер у блок обробки повідомлень 1. Цей блок містить буфер SIP повідомлень 2, класифікатор повідомлень 3 і модуль обробки SIP повідомлень 4.

Буфер SIP повідомлень 2 умовно має дві частини: для вхідних повідомлень і для повідомлень-таймерів. У нормальному режимі роботи SIP-сервер приймає повідомлення на обслуговування в порядку надходження їх у чергу за принципом FIFO (First In First Out), а повідомлення в буфері пам'яті для таймерів стоять в порядку закінчення таймерів. Вхідні повідомлення проходять через блок контролю перевантаження 5. Блок на основі заздалегідь заданих критеріїв приймає рішення про просунення повідомлення до черги або про відхилення повідомлення. По завершенню обслуговування проводиться відправка повідомлення далі по мережі і одночасно з цим відбувається установка таймера у буферну пам'ять. Після спливу таймера повідомлення ставляться в чергу на повторну відправку.

У режимі перевантаження повідомлення спочатку потрапляють на класифікатор 3 (Фіг. 2), в якому вони розподіляються за типом транзакції на INVITE і не-INVITE і розподіляються по чергах в залежності від пріоритету. Повідомлення з більшим пріоритетом мають більшу ймовірність бути обслугованими. Для вибору повідомлення з черги будемо використовувати динамічний пріоритет, тому що статичний може призвести до недостатнього обслуговування черг з низьким пріоритетом. Для вибору черги використовуємо два параметри, які розраховуються для кожної черги:

- кількість запитів, які очікують у черзі (N_i);
- коефіцієнт значущості черги (U_i).

Тоді пріоритет черги можна визначити як $N_i \times U_i$. Коефіцієнт значущості черги вибирається в залежності від важливості повідомлень, що знаходяться в цій черзі. Значення цього коефіцієнта може бути отримано зі статистичних даних, що описують запити користувачів.

Крім того, в режимі перевантаження повідомлення, що належать не-INVITE транзакції, обслуговуються за принципом FIFO. Якщо повідомлення належить INVITE-транзакції, то в цьому випадку враховується фаза встановлення з'єднання. Якщо це запит на нове з'єднання (фаза 1), то повідомлення відхиляється кодом 503. Якщо запит INVITE вже обробляється сервером (фаза 2 та 3), повідомлення надходить до голови черги і обслуговуються за принципом LIFO (Last In First Out). Дисципліна обслуговування LIFO забезпечує кращу пропускну здатність і менші затримки в черзі при перевантаженні в порівнянні з FIFO.

Крок 2. Блок моніторингу ресурсів 6 визначає стан сервера по завантаженню ресурсів і є індикатором перевантаження. Ресурси, за якими може проводитися моніторинг стану сервера, можуть бути різні, наприклад навантаження центрального процесора, затримка в черзі буфера, поточна довжина черги в буфері і т.п. Приклад розгорнутого блока моніторингу ресурсів 6 представлений на Фіг. 3. Процес визначення перевантаження на SIP-сервері визначається методами управління перевантаженням.

Крок 3. Після того, як навантаження перевищило граничне значення 7, SIP-сервер переходить у стан перевантаження SIP-сервера 8 і змінює порядок обслуговування повідомлень, враховує пріоритет обслуговування. В класифікаторі 3 відбувається поділ повідомлень та запити на встановлення нової сесії відхиляються відповіддю 503 або 500 для кінцевих вузлів згідно з [1]. Процес відхилення повідомлень реалізується детермінованою схемою прийняття рішення про відмову в обслуговуванні (Фіг. 4). Він має кращу продуктивність, ніж пробалістичні схеми, маючи менший коефіцієнт варіації значення ймовірності відмови в обслуговуванні. У цьому алгоритмі змінна g має початкове значення, рівне 0. Рішення прийняти або відхилити повідомлення приймається у блоці 9 за наступним алгоритмом з Фіг. 4, де f - ймовірність прийняття повідомлення на обслуговування, отримана згідно з методом управління перевантаженням.

Врахування фази встановлення з'єднання і використання динамічного пріоритету при обслуговуванні черг дозволить підвищити пропускну здатність сервера SIP, зміна порядку обслуговування повідомлень у режимі перевантаження з FIFO на LIFO дозволить скоротити час встановлення з'єднання, використання детермінованої схеми управління відмовами дозволить зменшити витрати ресурсів сервера на відмови у режимі перевантаження. Одночасне використання цих способів у будь-яких методах управління перевантаженням на сервері SIP дозволить підвищити ефективність цих методів та ефективність роботи сервера.

Заявлений спосіб може бути здійсненим таким чином.

Кожен сервер SIP має блок контролю перевантаженням 5, який відповідає безпосередньо за вибір методів управління перевантаженням і керування параметрами вибраного методу. Цей блок має параметри, які відповідають за вибір типу вимірювання, що використовуються для контролю перевантаження, за значення, які визначають умови переходу сервера в режим перевантаження і виходу з цього режиму. Враховуючи запропонований спосіб боротьби з перевантаженням на сервері SIP для поліпшення ефективності функціонування методів управління перевантаженням, доцільно додати в цей блок контролю перевантаження такі

параметри: 1) queue-policy, який визначає політику обслуговування повідомлень у черзі буфера сервера; 2) measuring-interval, який визначає інтервал вимірювань завантаження ресурсів сервера для методів управління перевантаженням, які цього вимагають; 3) throttling-method, який визначає схему управління відмовами повідомлень в режимі перевантаження.

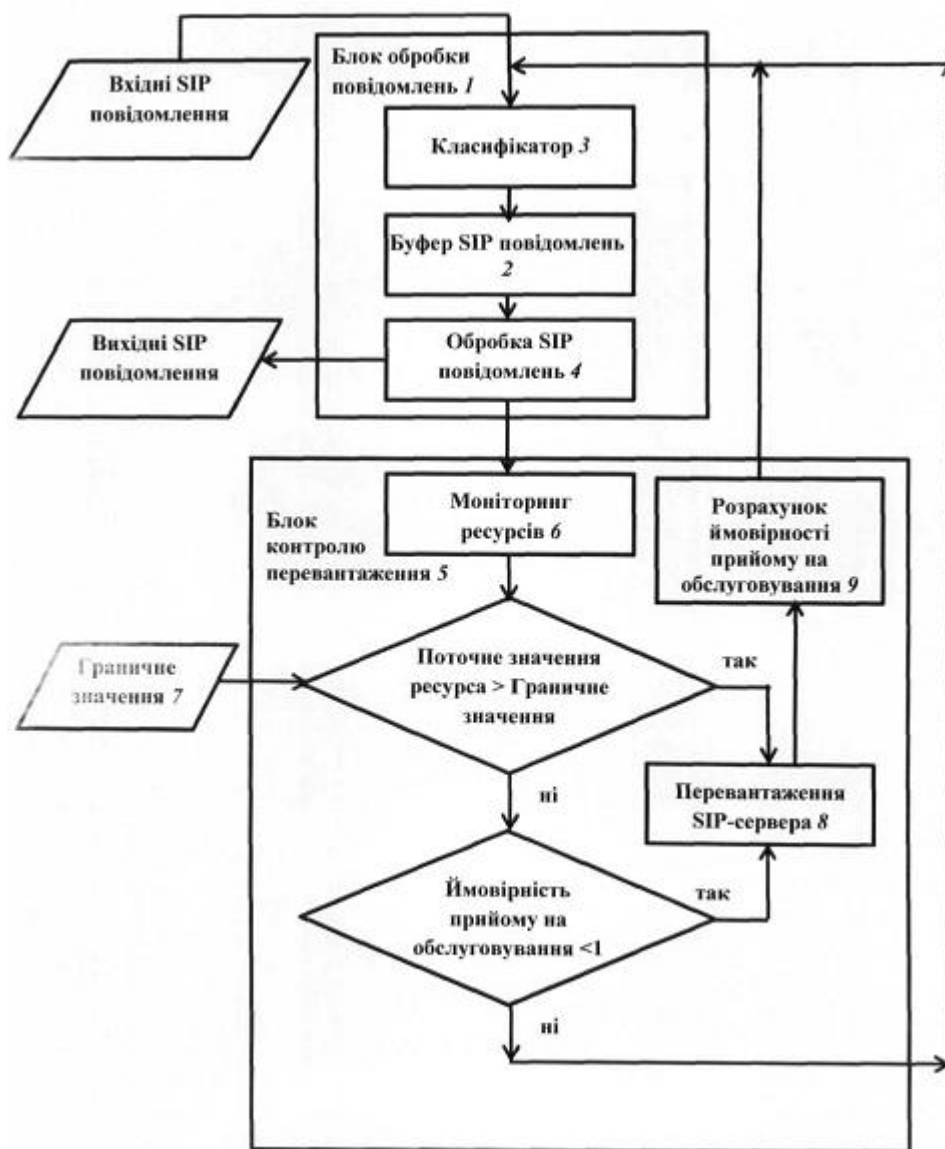
Джерела інформації:

1. Rosenberg J., Schulzrinne H. SIP: Session Initiation Protocol // IETF, RFC 3261. - June 2002. - 202 p.

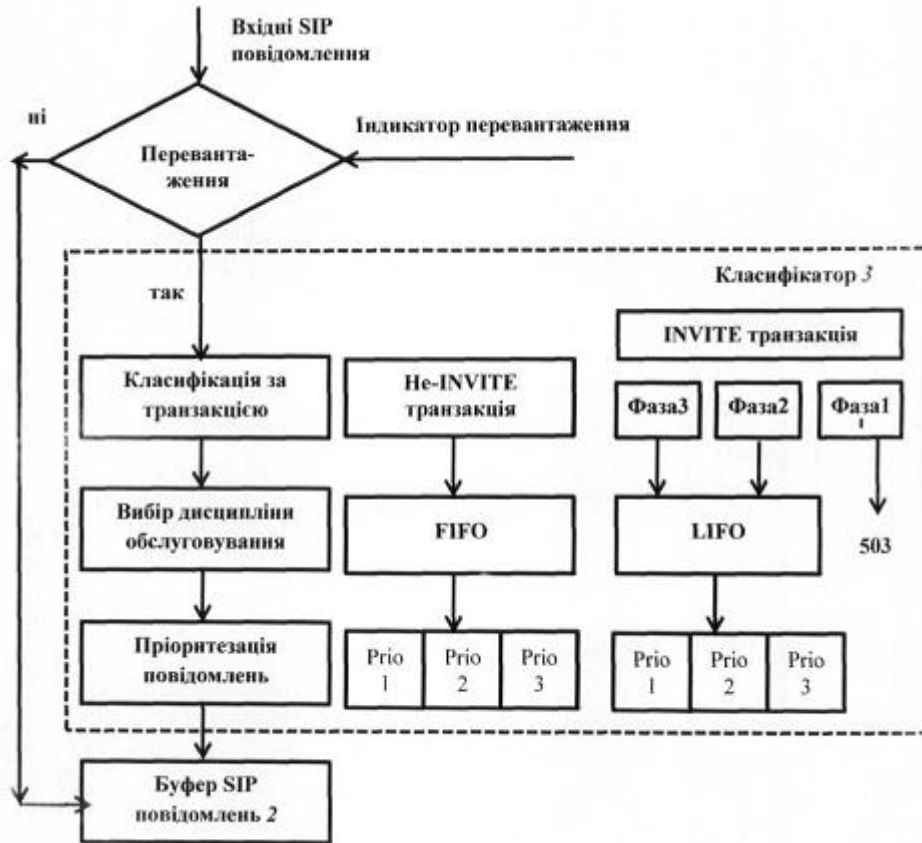
2. Masataka Ohta. Overload Control in a SIP Signaling Network // International Journal of Electrical and Electronics Engineering.-2009. - Vol. 3:2. - P. 87-92.

ФОРМУЛА КОРИСНОЇ МОДЕЛІ

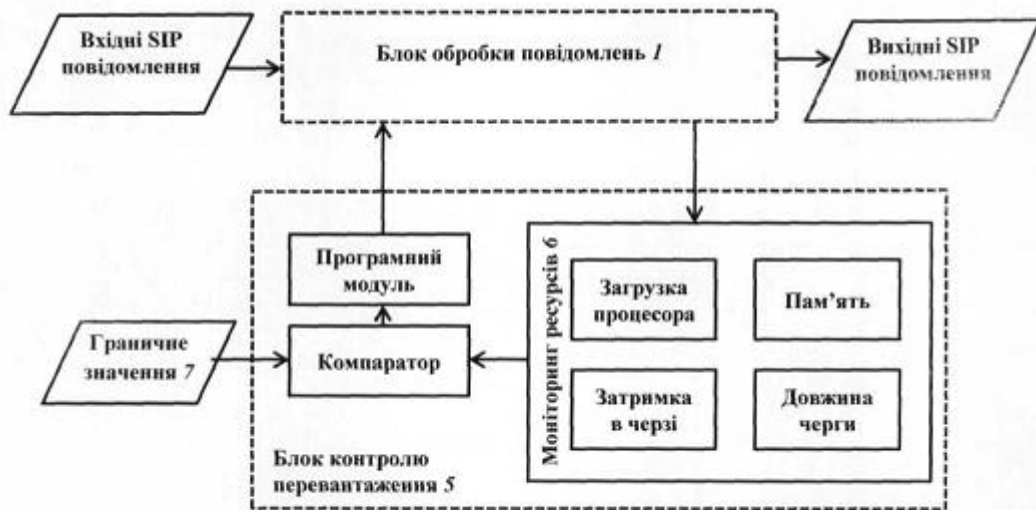
Спосіб боротьби з перевантаженням на сервері SIP, що включає виявлення перевантаження за завантаженням ресурсів сервера та реакцію на перевантаження з пробалістичною схемою регулювання відмовами, який **відрізняється** тим, що у режимі перевантаження сервер SIP обробляє вхідні повідомлення, враховуючи фазу встановлення з'єднання, використовує динамічний пріоритет при обслуговуванні повідомлень у черзі буфера та детерміновану схему прийняття рішення про відмову в обслуговуванні.



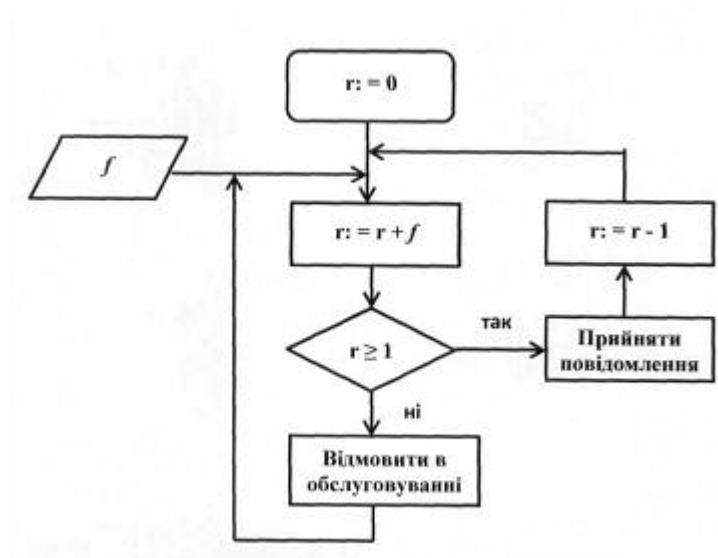
Фіг. 1



Фіг. 2



Фіг. 3



Фіг. 4