УДК 004.2

Скулиш М. А., канд. техн. наук (Тел.: +380 (50) 607 42 29. E-mail: mb_s@ukr.net)
(Національний технічний університет України «КПІ», Інститут телекомунікаційних систем, м. Київ)

МЕТОД КОНТРОЛЮ ЗА ПЕРЕВАНТАЖЕННЯМИ В ПРОЦЕСІ БАГАТОЕТАПНОЇ ОБРОБКИ ЗАЯВОК

Скулиш М. А. Метод контролю за перевантаженнями в процесі багатоетапної обробки заявок. Запропоновано метод керування вхідним потоком заявок на тарифікацію, особливістю якого є контроль кількості заявок, що знаходяться на ресурсозатратному етапі обслуговування. У разі перевищення їх допустимої кількості заявки затримуються на вході в систему обслуговування, що дозволяє уникнути перевантаження обчислювальних ресурсів та не допустити неефективного завантаження ресурсу на обслуговування заявок, які будуть втрачені внаслідок перевищення допустимого часу обслуговування.

Ключові слова: телекомунікаційна система, розподілена система, тарифікація послуг, динаміка навантаження, якість послуг

Скулиш М. А. Метод контроля перегрузок в процессе многоэтапной обработки заявок. Предложен метод управления входным потоком заявок на тарификацию. Особенностью которого, является контроль количества заявок, которые находятся на ресурсозатратном этапе обслуживания. В случае превышения их допустимого количества заявки задерживаются на входе в систему обслуживания, что позволит избежать перегрузки вычислительных ресурсов и не допустить неэффективного использования ресурсов на обслуживание заявок, которые потеряются по причине превышения допустимого времени обслуживания.

Ключевые слова: телекоммуникационная система, распределенная система, тарификация услуг, динамика нагрузки, качество услуг

Вступ. Надання телекомунікаційних послуг операторами зв'язку невід'ємно пов'язано з процесом їх тарифікації. Зростання кількості послуг, які надаються, викликає зростання навантаження на систему тарифікації. Виділяють два способи тарифікації: в режимі реального часу ("online") та в режимі "offline". Основним принципом боротьби з перевантаженнями в процесі тарифікації є переведення частини заявок з режиму online тарифікації в режим offline, тобто в періоди часу коли очікується перевантаження системи частина заявок не тарифікується, послуги надаються у повному об'ємі, а списання грошей відбувається у менш навантажені періоди, за принципами offline тарифікації [1, 2]. Однак через структуру вхідного потоку та особливості багатоетапного процесу обробки заявок на тарифікацію можуть виникати короткочасні тимчасові перевантаження.

Для операторів зв'язку достатньо гостро постає проблема перевищення часу обслуговування заявки саме на сервері тарифікації, оскільки не враховується принцип розподілу ресурсів технічних засобів, що у моменти пікових навантажень є критичними для якості обслуговуваня. Вирішення цієї проблеми може полягати у постійному збільшенні потужності обчислювальних ресурсів серверів, що є досить затратним, а через це неприйнятним на практиці, або у запобіганні появі пікових навантажень за рахунок керування вхідним потоком заявок на тарифікацію. Перевищення допустимої тривалості обслуговування призводить до відхилення виклику, відповідно до екотомічних втрат та зниження репутації компанії.

Внаслідок цього є актуальною наукова-технічна задача удосконалення процесу керування вхідним потоком викликів, яке б враховувало потреби у технічних ресурсах системи тарифікації та навантаження, яке створюється різними типами послуг, а також включало відповідні механізми, методи, моделі та алгоритми, протоколи, інтерфейси та засоби керування процесами обробки викликів та тарифікації для їх практичного впровадження, що дозволило б подолати описані вище недоліки.

Дана стаття є продовженням раніше опублікованих робіт [3...5], тому структура процесу тарифікації та особливості розподілу обчислювальних ресурсів між різними сервісами та етапами обслуговування не розглядається.

Постановка задачі методу керування вхідним потоком на сервер мобільного зв'язку. В основу методу керування вхідним потоком заявок на тарифікацію покладений контроль кількості заявок, що знаходяться на етапі обслуговування. У разі перевищення їх допустимої кількості заявки затримуються на вході, що дозволяє уникнути перевантаження ресурсів та не допустити неефективного завантаження ресурсу на обслуговування заявок, які оброблятимуться більш ніж за відведений в системі час.

Вхідними даними в задачі керування потоком заявок, які надходять на обслуговування на сервер мобільного оператора є:

- інформація про об'єм ресурсу, який є необхідним для здійснення операцій, передбачених функціональним блоком для обслуговування заявки заданого типу сервісу;
- інформація про тривалість використання ресурсів при обслуговуванні заявки заданого типу сервісу у кожному функціональному блоці;
- статистична інформація про тривалість обслуговування заявки заданого типу сервісу у кожному функціональному блоці;
- об'єм ресурсів виділений для обслуговування заданого типу сервісу.

Параметри серверу, які характеризуються як ресурси системи, що обслуговує заявки, як правило розраховані для середніх значень параметрів вхідного потоку, однак в системі наявні пікові значення кількості заявок, що надійшли одночасно.

Під сплеском навантаження вхідного потоку розуміємо одночасне надходження такої кількості заявок, яка більшою розрахованого вище допустимого значення.

Для керування процесом обробки заявок з метою запобігання дефіциту ресурсу в системі керування пропонується використання наступної стратегії:

– два і більше сплески навантаження вхідного потоку не обслуговувалися одночасно у функціональних блоках, обробка в яких потребує значної кількості ресурсів;

– для цього вводиться затримка частини заявок, надходження яких співпало зі сплеском навантаження. Час затримки визначати так, щоб затримані заявки не поступали в систему доти, доки попередній сплеск навантаження не буде успішно обслуженим в ресурсозатратному функціональному блоці.

Заявки на обслуговування надходять до системи *за заданим законом*. Процес обслуговування однієї заявки включає в себе перебування (обслуговування) заявки в одному з *n* функціональних блоків, для обслуговування використовується *G* типів ресурсів. Нехай на обслуговування поступають заявки від *m* типів сервісів. Відома статистика часу перебування заявки *i*-го типу сервісу $(i = \overline{1,m})$ в *j*-му функціональному блоці $(j = \overline{1,n})$.

Відоме математичне очікування (t_{ij}) часу перебування заявки *i*-го типу сервісу $(i = \overline{1,m})$ в *j*-му функціональному блоці $(j = \overline{1,n})$, ці дані зведені в матрицю **T**=$\{t_{ij}\}$. Відомо, що протягом обслуговування заявки *i*-го типу сервісу $(i = \overline{1,m})$ у *j*-му функціональному блоці $(j = \overline{1,n})$ ресурс g-го типу займається на час τ_{ij}^{Sg} $(\tau_{ij}^{Sg} \le t_{ij})$. Інформація про тривалість обслуговування зведена в матриці $T^{Sg} = \{\tau_{ij}^{Sg}\}_{i=\overline{1,m},j=\overline{1,n}}$. Всього таких матриць *G* штук, кожна матриця відповідає одному з ресурсів, що розглядається.

Відома матриця $V^{Sg} = \{v_{ij}^{Sg}\}$, кожний елемент v_{ij}^{Sg} якої відповідає об'єму ресурсу g-го типу який використовується при обслуговуванні заявки *i*-го типу $(i = \overline{1,m})$ в *j*-му функціональному блоці $(j = \overline{1,n})$.

В рамках цього дослідження не розглядається деталізація бізнес процесів, які відбуваються у функціональному блоці, тобто не уточнюється на якому саме етапі обслуговування заявки в середині функціонального блока який ресурс використовується. Тому зроблено припущення: всі заявки які у поточний час обслуговуються у функціональному блоці використовують ресурси рівномірно, тобто об'єм g-го ресурсу,

що використовується *i*-м типом сервісу у *j*-му ФБ зменшується пропорційно відношенню часу використання ресурсу до часу перебування заявки у функціональному блоці, тоді справедлива формула: $v_{ij\ new}^{Sg} = v_{ij}^{Sg}\frac{\tau_{ij}^{Sg}}{t_{ij}}$, де $v_{ij\ new}^{Sg}$ – індексований об'єм ресурсу g-го типу, який використовується протягом часу обслуговування заявки і-го типу сервісу у j-му ФБ. Формуються нові матриці $V_{new}^{Sg} = \{v_{ij\ new}^{Sg}\}$.

Необхідно визначити метод керування вхідним потоком заявок, що дозволяє уникнути дефіциту ресурсів системи.

Алгоритм методу. Розшифрування позначень застосованих у алгоритмі наводиться після нього.

1. Для кожного *i*-го типу сервісу задати допустиму кількість заявок які можуть одночасно поступати на обслуговування в систему ($k_{i\ доп}$). Кількість допустимих заявок залежить від інтервалу дискретизації часу, система відліку дискретного часу має бути єдиною для всієї системи.

Зауваження: У подальших роботах буде розглянутий метод визначення допустимої кількості заявок, що розв'язується як задача динамічного програмування (задача про загрузку машини).

2. Задається множина **F**={∅}. Для кожного типу ресурсу $g = \overline{1,G}$ в матриці V_{new}^{Sg} знаходиться максимальний елемент $v_{i_g j_g}^{Sg} = \max\{v_{11\ new}^{Sg},\ v_{12\ new}^{Sg}, \dots, v_{mn\ new}^{Sg}\}$, пари індексів $(i_{g1}j_{g1})$ відповідних елементів додаються до множини **F**. Якщо у матриці присутні два або більше ($gmax \geq 1$) максимальних елементи $v_{i_{g\,1}j_{g\,1}}^{Sg} = \dots = v_{i_{g\ gmax}j_{g\ gmax}}^{Sg}$, тоді в множину **F** додаються всі пари індексів, та позначаються $(i_{g1}j_{g1},\ \dots,\ i_{g\ gmax}j_{g\ gmax})$. Індекси максимальних значень об'єму для різних типів ресурсів можуть співпадати; значення, що повторюються, до множини **F** не додаються.

Наприклад, $i_{11}j_{11} = i_{21}j_{21} = 2\ 3$, це означає, що для ресурсу 1 і для ресурсу 2 перший максимальний елемент відповідає процесу обслуговування заявки 2-го типу сервісу у третьому функціональному блоці, тобто це обслуговування є найбільш витратним для ресурсів першого та другого типу, в такому випадку пара (2, 3) увійде до множини **F** один раз. Таким чином, множина **F** заповнюється парами, де на першій позиції стоїть номер сервісу, обслуговування якого є ресурсозатратним у функціональному блоці, номер якого стоїть на другій позиції.

Зауваження: Пари номерів не зберігають тип ресурсу, оскільки це не має значення для даного методу керування.

3. Елементи множини **F** впорядковуються за першим елементом. Множина **F** розділяється на *m* підмножин, таким чином, щоб в $\mathbf{F_1}$ увійшли пари де перший елемент дорівнює 1, в $\mathbf{F_2}$ увійшли пари де перший елемент дорівнює 2, тощо. Якщо деяка *r*-та підмножина $(r\epsilon\overline{1,m})$ буде порожньою ($\mathbf{F}_r$={∅}), тоді для заявок *r*-го типу сервісу не будуть застосовуватися затримки заявок, що надійшли у сплесках навантаження вхідного потоку. Для всіх підмножин $\mathbf{F}_d$ $(d\epsilon\overline{1,m})$, де міститься один елемент, виконуються дії п.4. Для всіх підмножина $\mathbf{F}_p$ $(p\epsilon\overline{1,m})$, де міститься два і більше елементи, виконуються дії з п. 5.

4. Завдання цього пункту полягає в тому, щоб визначити максимальну затримку надлишкової кількості заявок *d*-го типу сервісу, які надійшли у моменти пікових навантажень вхідного потоку. Якщо в множині $\mathbf{F}_d$ міститься 1 елемент (d, f_d), це означає, що для заявки *d*-го типу сервісу не можна допускати двох піків навантаження протягом часу обслуговування у функціональному блоці f_d, тривалість якого визначається з матриці **T** та дорівнює $t_{d\ f_d}$. Перехід до п. 6.

5. Завдання цього пункту не тільки не допустити припадання двох піків навантаження на один критичний (ресурсозатратний) функціональний блок, але й уникнути суперпозиції, коли два піки навантаження обслуговуються у двох ресурсозатратних функціональних блоках. Для цього елементи підмножини $\mathbf{F}_p$ впорядковуються за другим елементом. Припустимо, що множина $\mathbf{F}_p$ складається з двох елементів: $(p,\ f1_{\text{р}})$ і $(p,\ f2_{\text{р}})$, задачі з більшою кількістю елементів мало ймовірні та вирішуються у аналогічний спосіб. Це означає, що при обслуговуванні заявок p-го типу сервісу, ресурсозатратними є функціональні блоки з номерами $f1_{\text{р}}$ і $f2_{\text{р}}$. З матриці **T** обираються елементи з відповідними індексами: $t_{p\,f1_p}$, $t_{p\,f2_p}$. Умови при яких два піки навантаження не припадуть на один функціональний блок наступні:

A) відстань між піками навантаження не може бути меншою ніж значення $t_{p\,f1_p}$;

B) відстань між піками навантаження не може бути меншою ніж значення $t_{p\,f2_p}$;

C) якщо $f2_{\text{р}} - f1_{\text{р}} = x > 1$, тоді не допускається відстань між піками навантаження більша ніж $\sum_{q=1}^{x-1} t_{p\,(f1_p+q)}$.

6. В процесі роботи системи моніторингу фіксуються моменти пікових навантажень, коли у систему надійшла кількість заявок, що є більшою за допустиме значення (відповідно до п. 1). Моменти часу, коли виявлено сплеск навантаження, додаються до множин $\mathbf{T}_{i\ \max}$, де i – тип сервісу, для якого зафіксовано сплеск навантаження. Для сервісів r-того типу (див. п.3) множина $\mathbf{T}_{r\ \max}$ не створюється.

Для сервісів типу d, для елементів множини $\mathbf{T}_{d\ \max}$ перевіряється умова п.4, тобто для кожного нового елементу множини $t_{d\,\max w+1}$ перевіряється значення $t_{d\,f_d} - (t_{d\,\max w+1} - t_{d\,\max w}) = y1$, якщо y1>0, тоді частина заявок $(k_d(t_{d\,\max w+1}) - k_{d\,\text{доп}})$ затримується на час $y1$. Якщо у момент часу $(t_{d\,\max w+1} + y1)$, кількість заявок що надійшла $k(t_{d\,\max w+1} + y1)$ плюс залишок $(k_d(t_{d\,\max w+1}) - k_{d\,\text{доп}})$ в сумі дають значення більше допустимого $k_{d\,\text{доп}}$. Тоді надлишок передається у наступний момент часу $(t_{d\,\max w+1} + y1 + 1)$, *процедура згладжування навантаження.*

Для сервісів типу p, для елементів множини $\mathbf{T}_{p\ \max}$ перевіряються умови п.5, тобто для кожного нового елементу множини $t_{p\,\max w+1}$ перевіряються умови:

– значення $t_{p\,f1_p} - \left(t_{p\,\max w+1} - t_{p\,\max w}\right) = y2$, якщо y2>0, тоді частина заявок $\left(k_p\left(t_{p\,\max w+1}\right) - k_{p\,\text{доп}}\right)$ затримується на час $y2$, у разі потреби застосовується процедура згладжування навантаження;

– значення $t_{p\,f2_p} - \left(t_{p\,\max w+1} - t_{p\,\max w}\right) = y3$, якщо y3>0, тоді частина заявок $\left(k_p\left(t_{p\,\max w+1}\right) - k_{p\,\text{доп}}\right)$ затримується на час y2, у разі потреби застосовується процедура згладжування навантаження.

Якщо $f2_{\text{р}} - f1_{\text{р}} = x > 1$, тоді досліджується значення $\sum_{q=1}^{x-1} t_{p\,(f1_p+q)} - \left(t_{p\,\max w+1} - t_{p\,\max w}\right) = y4$, якщо $y4<0$, тоді частина заявок $\left(k_p\left(t_{p\,\max w+1}\right) - k_{p\,\text{доп}}\right)$ затримується на час $(t_{p\,f2_p} + \text{y4})$, у разі потреби застосовується процедура згладжування навантаження.

Таким чином, у разі реєстрації події надходження другого піку навантаження протягом часу, який визначено умовами, здійснюється затримка надлишкової кількості заявок на час визначений алгоритмом методу, після чого затримані заявки надсилаються в систему так, щоб не допустити створення піку навантаження.

Алгоритм методу керування наведено на Рис. 1.

Кількість заявок, що є допустимою, для заданого типу сервісу розраховується методом перерозподілу технічних засобів між заявками різних типів послуг описаним вище, при цьому враховується ефективність обслуговування всіх типів сервісів при наявному об'ємі ресурсів системи.

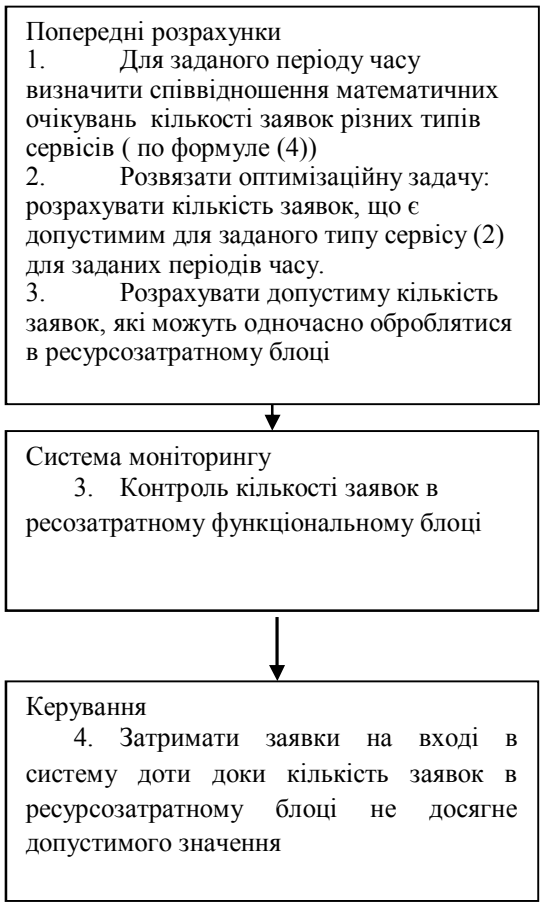

Рис. 1. Алгоритм методу керування вхідним потоком заявок для запобігання дефіциту серверних ресурсів

Моделювання. Проведено імітаційне моделювання методу керування потоком заявок на тарифікацію. Для моделювання було використано пакет GPSS.

В процесі імітаційного моделювання досліджувалася модель для двох ресурсів і потоку сервісів. Ресурси, що враховувалися під час моделювання – RAM and Permanent storage.

Процес обробки заявки включає в себе чотири функціональні блоки. Робота функціональних блоків імітувала такі операції як: вилучення інформації абонента з бази даних, розрахунок вартості послуги, формування нотифікації для абонента, фінальний підрахунок та списання коштів.

Для забезпечення обслуговування був виділений заданий об'єм ресурсів, розрахований на одночасне обслуговування 50 тисяч заявок на тарифікацію, за умови рівномірного розподілу кількості заявок між функціональними блоками. Під час обслуговування заявки в функціональному блоці відповідна кількість ресурсу блокувалася, та звільнялася при переході до наступного функціонального блоку. Якщо заявка надходить на обслуговування у функціональний блок, але ресурсу не достатньо для здійснення обслуговування, то заявка затримується до звільнення необхідної кількості ресурсу. На кожному етапі перевіряється час перебування заявки в системі та порівнюється з допустимим часом обслуговування. Значення були обрані максимально наближеними до реальних систем.

Вхідний потік був змодельований за законом Пуасона. Виходячи з аналізу роботи реальних системи найбільше ресурсів витрачається під час формування повідомлення абоненту. Тому в моделі здійснювався контроль за кількістю заявок, які обслуговувалися у поточний момент часу в третьому функціональному блоці та здійснювалася затримка

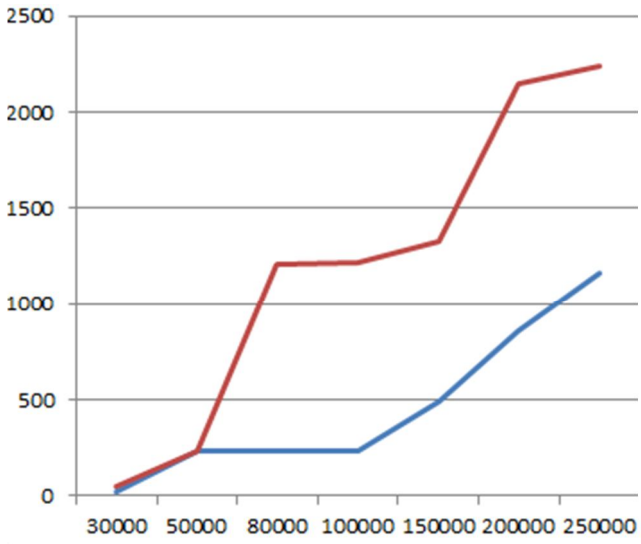

Рис. 2. Кількість втрачених заявок

повідомлень доти, доки кількість заявок не стане меншою за максимально допустиму кількість.

Динаміку залежності кількості втрачених заявок від інтенсивності вхідного потоку показано на Рис. 2. На рисунку видно зменшення втрат заявок через перевищення допустимого часу обслуговування.

Верхня лінія позначає втрати пакетів без застосування запропонованого методу керування вхідним потоком.

Нижня лінія позначає результати моделювання за допомогою запропонованого методу керування.

Висновки. В статті розглянуті проблеми організації роботи системи онлайн тарифікації. Удосконаленно спосіб керування чергами вхідних заявок на сервер тарифікації, який враховує вимоги до технічних ресурсів на кожному етапі обслуговування, дозволяє зменшити втрати заявок за рахунок введення затримок, які суттєво не впливають на загальний час обслуговування заявки на тарифікацію, однак унеможливлюють одночасне перебування значної кількості заявок, що потребують великої кількості ресурсів, на етапі обслуговування.

Проведено імітаційне моделювання засобами пакету GPSS, що дозволило отримати висновок: при використанні методів керування вхідним потоком заявок на тарифікацію кількість не обслужених заявок на тарифікацію, через перевищення часу обслуговування зменшилася до 3-х разів.

Література

1. Pilipenko A. Y. Method of reducing the billing system load in critical mode / A. Y. Pilipenko, V. F. Cherdyntseva // 22nd Int. Crimean Conf. "Microwave & Telecommunication Technology" (CriMiCo'2012). Sevastopol. – 2012. – C. 403-404.

2. Mussel K. M.: Provision and billing of communication services. System Integration / K. M. Mussel. – Moscow : Eco-Trendz, 2003.

3. Скулиш М. А. Організація роботи групи серверів для забезпечення потреб розподіленої системи тарифікації послуг / М. А. Скулиш // Наукові записки Українського науково-дослідного інституту зв'язку. – 2014. – №5(33). – С. 56-64.

4. Скулиш М. А. Метод складання розкладу залучення ресурсів длявисоконавантажених інформаційних систем / М. А. Скулиш // Наукові записки Українського науково-дослідного інституту зв'язку. – 2014. – №6(34). – С. 65-70.

5. Скулиш М. А. Метод розподілу ресурсів сервера оператора мобільного зв'язку / М. А. Скулиш, А. А. Заставенко // Вісник НТУУ «КПІ». Серія Радіотехніка, Радіоапаратобудування. – 2015. – № 60. – С. 35-45.

Дата надходження в редакцію: 17.01.2015 р. Рецензент: д.т.н., проф. К. С. Сундучков