УДК 004.9

Скулиш М. А.,
Інститут телекомунікаційних систем НТУУ «КПІ»

МЕТОД ЗГЛАДЖУВАННЯ ВХІДНОГО НАВАНТАЖЕННЯ НА СЕРВЕР ON-LINE ТАРИФІКАЦІЇ

Skulysh Mariia,
National Technical University of Ukraine “Kiev Polytechnic Institute”

METHOD OF SMOOTHING INPUT LOAD ON THE ONLINE CHARGING SYSTEM

Charging users for a server operator billing involves performing a number of operations . Each service provided by the operator or service provider customers , charged at a separate scheme , as the bottleneck in the process of providing services are bilingual system.

Operation performed is limited. In case of exceeding its application is considered lost and the user is notified of the inability to obtain services. In this situation, the operator has monetary losses. In systematic failures in service deteriorates reputation . Therefore, it is important to make a control input streams so as to reduce the possibility of exceeding the service time of the user's requests.

The process of managing incoming applications to the server should consider charging the features of charging large companies:

• charging process involves a sequence of standard operations , speed and resources for the implementation of which depends on the type of service that serviced

• a variety of services that are billed differently ,

• extremely high number of requests for billing ,

• irregularity of the input stream , applications received according to Poisson.

Scheme smoothing the input load is a set of values of the maximum number of requests (the sequence {k_i}), arriving at the input of the system in a short time interval Δt_i in a given sequence. Number of elements in a sequence of n is chosen so that the equation was carried out $t = \sum_{i=1}^{n} \Delta t_i$, where t - the average time of the system.

It’s necessary to choose a sequence {k_i} to satisfy conditions:

1. Applications that are both served by the system must use the volume V of resources close to the total maximum number of resource Vmax. Dispersion sequences such amounts shall be minimized.

2. Dispersion elements of the sequence {ki} should be minimal.

The method of solution of choice sequence {ki} by using genetic algorithm.

When using the proposed scheme smoothing the input load is provided by the maximum allowable flow uniformity requests. Since overlooked the sequence of operations performed on the application server to the mobile operator, the amount of resource that is required for these operations. Selected sequence of maximum permissible values of number of applications coming into the system in a short period of time.

Keywords – flow management system, server’s resources, service processing, charging system

У статті запропоновано метод згладжування вхідного потоку заявок на сервер тарифікації, що враховує вимоги до ресурсів кожного з етапів обслуговування заявок,що дозволить рівномірно навантажити ресурси системи протягом часу обслуговування заявки.

Ключові слова – система керування потоками, ресурсі серверу, процес обслуговування, система тарифікації

Обслуговування абонентів на сервері тарифікації є необхідною складовою частиною процесу обслуговування абонентів та надання їм телекомунікаційних послуг. Оскільки кожна послуга є платною, оператор зв'язку залежно від типу тарифного плану здійснює тарифікацію заявки або в перед наданням послуги в режимі реального часу або в режимі offline має перевірити наявність коштів на рахунку абонента, зробити перерахунок, тобто здійснити ряд стандартних операцій.

Тарифікація абонентів на стороні оператора зв'язку передбачає виконання ряду операцій. Кожна послуга, яку надає оператор або провайдер послуг абонентам, тарифікується за окремою схемою, тому вузьким місцем у процесі надання послуг є білінгва система.

Час виконання операцій є обмеженим, у разі його перевищення заявка вважається втраченою і абонент отримує повідомлення про неможливість отримання послуги. У такій ситуації оператор несе збитки, а у разі систематичних відмов у наданні послуг псується репутація компанії. Тому важливо здійснити керування вхідним потоком таким чином, щоб зменшити ймовірність перевищення часу обслуговування абонентських заявок.

Процес керування вхідним потоком заявок на сервер тарифікації має враховувати особливості систем тарифікації великих компаній:

- процес тарифікації передбачає послідовність стандартних операцій, швидкість та ресурси для виконання яких, залежать від типу сервісу, який обслуговується,
- різноманітність сервісів, що по-різному тарифікуються,
- надзвичайно велика кількість запитів на тарифікацію,
- нерівномірність вхідного потоку, заявки надходять за законом Пуассона.

Розглянемо більш детально кожну особливість.

Схему обслуговування заявок на сервері тарифікації показано на рис. 1. Заявки надходять на сервер за різними протоколами (Diameter, CAP2, MAP),у відповідних модулях EDP та RES (Enhanced Diameter Proxy та) розподіляються між пулом адаптерів відповідних протоколів. Після чого декодуються у відповідних адаптерах CPA, DPA (CAP Protocol Adapter та Diameter Protocol Adapter) та приводяться до єдиного виду.

Розкодовану справу модуль передає на сервер бізнес логіки SEE (Service Execution Environment) через модуль маршрутизації BUS. Сервер бізнес логіки SEE являється ядром системи тарифікації та забезпечує середовище для виконання послідовності операцій, які передбачає процес обслуговування заявок.

Послідовність операцій які повинні бути виконані для успішного обслуговування заявки включає в себе наступні дії:

1. Вилучення інформації про абонента. Для цього модуль SEE звертається до бази даних керування абонентами (CM DB Customer Management data base).
2. Вилучення інформації про місце розташування абонента. Для цього модуль SEE звертається до бази даних, що зберігає структуру мережі (RE DB Resource Inventory data base).
3. Вилучення інформації про стан рахунку абонента. Для цього модуль SEE звертається до бази даних рахунків абонентів (Balance Storage).

4. Здійснюється розрахунок вартості послуги на основі тарифної книги та поточної тарифної моделі абонента. Для цього модуль SEE звертається до модулю розрахунків (RE – Rating Engine).
5. З рахунку абонента знімається плата за послугу, залежно від типу сервісу паралельно може здійснюватися резервування заданої суми, тільки для сервісів з резервацією (SCUR - Session Charging with Unit Reservation та ECUR - Event Charging with Unit Reservation), коли тривалість надання послуги є не відомою та не можна підвести остаточний розрахунок. Для здійснення операцій дебету та резервації коштів модуль SEE звертається до бази даних рахунків абонентів (Balance Storage).
6. Формується звіт про надання послуги CDR (call data record). Для цього модуль SEE звертається до програмного модулю формування CDR (TTS – Toll Ticket Server).
7. Якщо це передбачено послугою, відправляється повідомлення абоненту про результати наданих послуг. Для цього модуль SEE звертається до модулю для посилки повідомлень абоненту (MG- message GateWay).

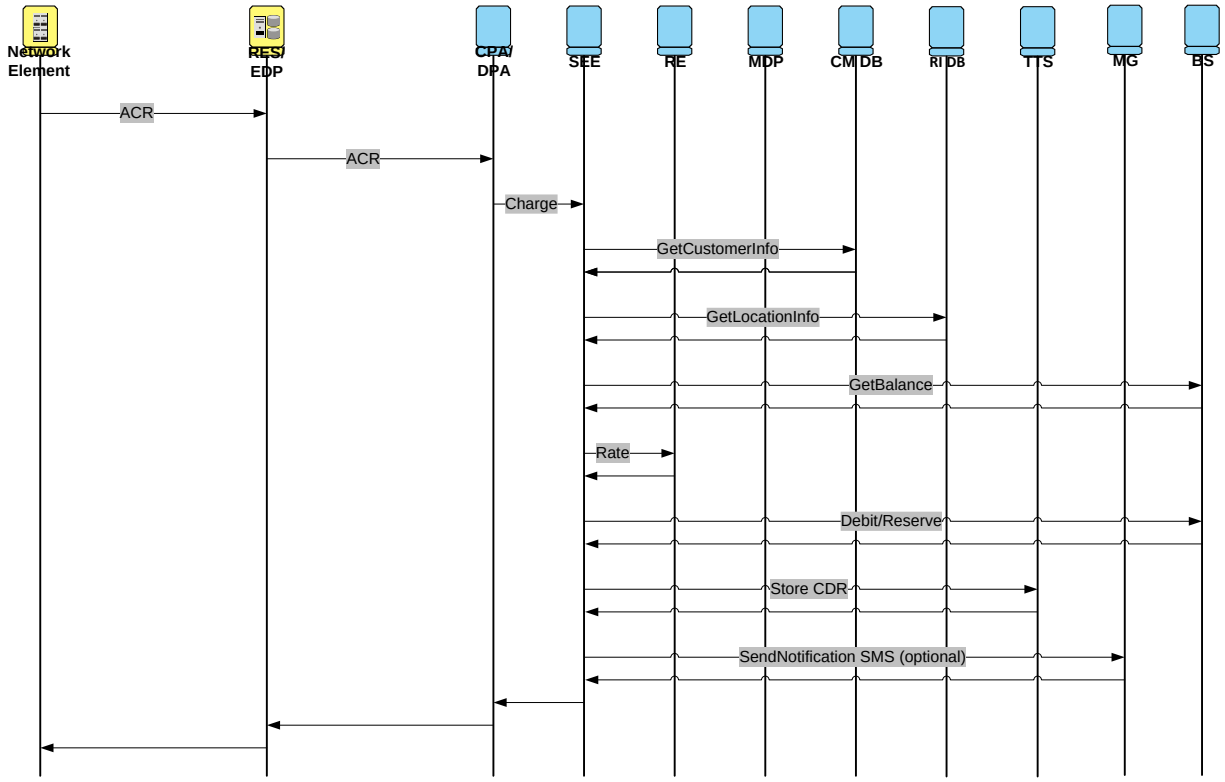

Рис. 1. Загальна схема online тарифікації послуг на сервері тарифікації

Про цес тарифікації завершено.

Як видно процес тарифікації є багатоетапним, при цьому операції, які послідовно виконуються у ядрі бізнес логіки SEE із залученням різних підсистем, різноманітні, відповідно потребують різної кількості оперативної пам'яті, процесорного часу та дискового простору. При вирішенні задачі керування вхідним потоком необхідно звернути увагу на час використання ресурсів. Необхідно врахувати як загальну кількість ресурсів, що обслуговує сервер в цілому, так і розділення ресурсів (методами віртуалізації), які з одного боку забезпечують ефективне обслуговування кожного етапу обробки, а з іншого є обмеженням, оскільки підсистема використовує лише ресурси їй відведені та не має доступу до інших ресурсів. Є також необхідність врахувати розподіл середнього часу виконання операцій.

Друга особливість полягає у тому, що кожний тип послуг, незважаючи на стандартність операцій, які виконуються в процесі тарифікації, потребує для здійснення розрахунків різної кількості ресурсів.

З точки зору необхідної процедури обслуговування всі сервіси можна розділити на три групи:

- Сесія тарифікації з резервацією (SCUR - Session Charging with Unit Reservation) оперативна пам'ять зайнята на всю тривалість сесії (може тривати до доби – наприклад, GPRS)
- Моментальна тарифікація події (IEC Immediate Event Charging) не зберігає стан свого виконання в пам'яті – виконується оцінка та списання грошових коштів в один момент (SMS).
- Тарифікація події з резервацією (ECUR - Event Charging with Unit Reservation) – оперативна пам'ять зайнята на період резервації (наприклад, час доставки контенту абоненту: відео, музика, MMS).

Таким чином, сервіси SCUR та ECUR виконуються у декілька етапів. Стан заявки або стан виклику зберігається на у підсистемі MDP (Memory DataBase Provider). MDP – це модуль для збереження поточного стану, що представляє собою програмно-апаратний комплекс, який забезпечує швидкий доступ до оперативної пам'яті (запис, зчитування, пошук).

На рис. 1 було наведено загальну схему online тарифікації всі сім етапів одноразово проходять заявки групи сервісів IEC. При обслуговуванні сервісів SCUR та ECUR перший та другий етапи, що включають в себе вилучення інформації про абонента та про місце його розташування виконуються один раз після чого вся інформація про абонента та стан заявки-виклику зберігається у підсистемі MDP.

Ресурси, які потребує система при обслуговуванні 7 етапів залежать не тільки від групи сервісів але й від його типу. Відмінність у швидкості та ресурсозатратності операцій виникає при розрахунку вартості послуги та при вилученні інформації про стан заявки-виклику із системи MDP. Ці умови мають бути враховані при розрахунку плану керування вхідним потоком на сервер тарифікації.

Третя особливість полягає в тому, що одночасно надходить велика кількість заявок на тарифікацію різних типів ресурсів.

На сьогоднішній день оператори мобільного зв'язку надають послуги мільйонам абонентів, наприклад компанія «Київстар» обслуговує до 26 мільйонів абонентів. При цьому за рахунок привабливих пакетних умов все більше абонентів користуються послугами мобільного Інтернету, та дзвінків. За умови централізованого обслуговування система тарифікації одночасно обслуговує до одного мільйона абонентів, які замовляють або продовжують користуватися послугами. Для кожної заявки абонента ініціюється ланцюг операцій описаних вище. У часи найбільшого навантаження кількість абонентських заявок збільшується у декілька разів.

Четвертою особливістю є неоднорідність вхідного потоку заявок. Системи обслуговування абонентів прийнято розглядати як системи із пуассонівським вхідним потоком заявок. Основними особливостями якого є значна дисперсія кількості заявок, що надходять на тарифікацію. Для розподілу Пуассона дисперсія дорівнює математичному очікуванню. Тобто можливі сплески навантаження на короткий період часу, що є меншим за час обслуговування заявки на сервері тарифікації. Такі сплески призводять до тимчасового перевантаження серверу навіть в умовах не часу пік.

З точки зору реалізації логіки процесу обслуговування, роботу підсистем серверу оператора мобільного зв'язку можна представити як багаторівневу систему масового обслуговування, де керування потоком заявок здійснюється на двох рівнях.

Перший логічний рівень прикладних програм. Тут заявки, які надійшли у систему, розділяються за типом сервісу, який вони представляють, обслуговування черг ведеться відповідно до схеми обслуговування розробленої для відповідного типу сервісу. Процес керування включає в себе формування черг за типом сервісу, застосування методів групи WRAD, а також інших схем керування, які враховують специфіку обслуговування сервісів [вставить ссылку про патент]. Таким чином, система масового обслуговування першого рівня представляє собою заявки різних типів сервісів, які надходять на обслуговування до системи, будемо називати їх заявками першого рівня. Обслуговуючими пристроями у такій системі виступають ланцюги функціональних блоків, де

здійснюється послідовне обслуговування заявок, кожний тип сервісу обслуговується у окремому ланцюзі.

Другий рівень – рівень технічної обробки. Схема обслуговування заявок за типом сервісу передбачає послідовне виконання операцій, для виконання яких потрібна задана кількість апаратних ресурсів, кожну операцію можна представити як заявку на обслуговування, будемо говорити про заявки другого рівня, де обслуговуючими пристроями виступають апаратні ресурси. Тут заявки другого рівня організовуються у черги до відповідних ресурсів. Політики використання ресурсів визначаються методами керування ресурсами обчислювальної системи. Архітектури розподілу ресурсів, організація обслуговування заявок другого рівня, суттєво впливають на швидкість обслуговування. Однак, така архітектура системи обробки заявок другого рівня є постійною, про її роботу можна судити по статистичним даним затримок в обслуговуванні заявок першого рівня.

Оскільки, вхідний потік заявок другого рівня однозначно визначається кількістю заявок першого рівня, що обслуговуються у системі. Тому система керування вхідним потоком заявок першого рівня, яка побудована з урахуванням статистики завантаженості ресурсів системи другого рівня, дозволить зменшити втрати заявок через затримки пов'язані з нестачею ресурсів.

Архітектура дворівневої системи керування представлена на рис. 2.

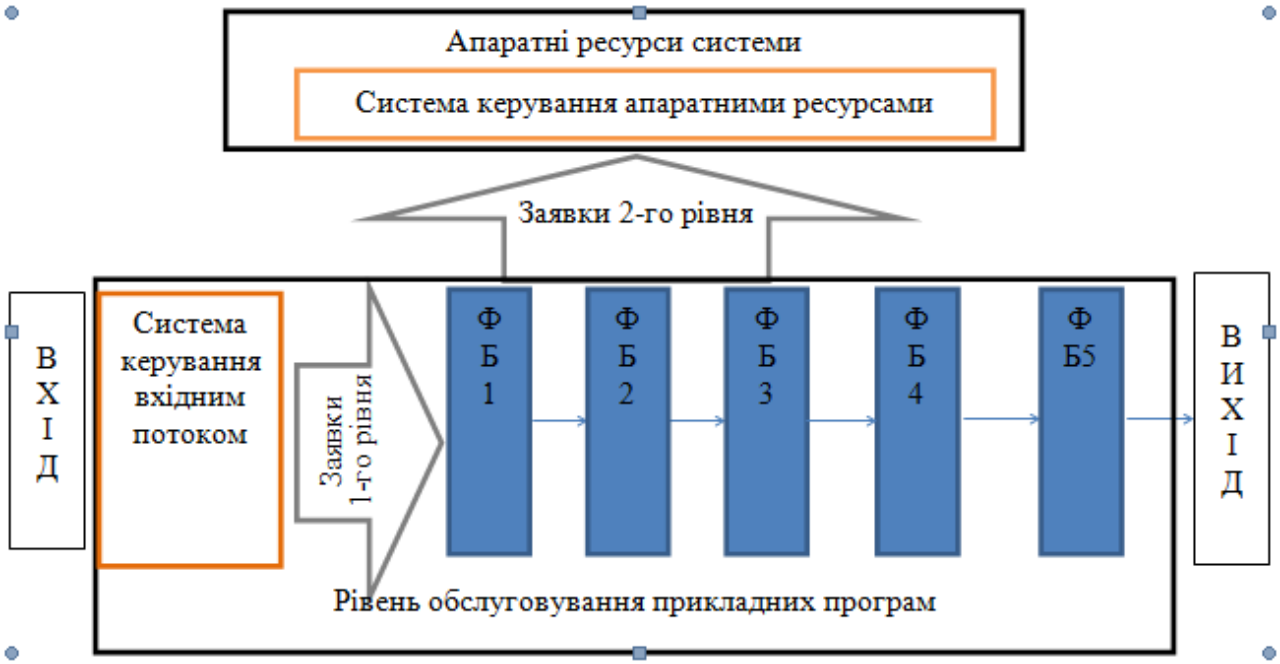

Рис. 2. Схема обслуговування заявок на сервері оператора мобільного зв'язку

Постає питання, яким чином організувати роботу системи керування вхідним потоком заявок, щоб потік заявок другого рівня був як найбільше рівномірним.

Обслуговування заявок першого рівня у функціональних блоках породжує потік заявок другого роду, для виконання яких використовуються задана кількість ресурсів серверу. Отже, якщо у деякому функціональному блоці обробляється одночасно велика кількість заявок першого роду, при цьому заявки другого роду породжені відповідним функціональним блоком потребують для свого виконання значної кількості ресурсів, тоді може постати проблема браку ресурсів серверу, що призведе до затримки в обслуговуванні заявок першого роду, и як наслідок перевищення допустимого часу обслуговування, втрати заявок, зниження якості обслуговування абонентів.

В [1] була запропонована стратегія керування вхідним потоком заявок, яка полягає в тому, щоб не допускати двох сплесків навантаження протягом часу обслуговуванні заявок у ресурсозатратних функціональних блоках. Запропонований метод передбачає відслідковування піків навантаження, та введення затримки для деякої частини заявок другого піку, що дозволяє уникнути перевантаження ресурсів серверу. Однак реалізація даної стратегії та запропонованого

методу потребує постійного моніторингу роботи серверу, а на практиці з метою заощадження ресурсів система моніторингу не завжди активна через брак ресурсів, тому відслідковування моментів надходження надлишкової кількості заявок не завжди можлива. Саме тому доцільно розробити *схему згладжування вхідного навантаження.*

Схема згладжування вхідного навантаження представляє собою набір значень максимально допустимої кількості заявок (послідовність $\{k_i\}$), що поступають на вхід системи за малий інтервал часу Δt_i у заданій послідовності. Кількість елементів послідовності *n* підбирається таким чином, щоб виконувалось рівняння

$t = \sum_{i=1}^{n} \Delta t_i$,

де t – середній час перебування заявки першого рівня у системі.

Необхідно підібрати таку послідовність $\{k_i\}$, щоб виконувалось дві умови:

1. Заявки, які одночасно обслуговуються у системі повинні використовувати об'єм ресурсів *V* близький до загальної максимально можливої кількості ресурсу V_{max}. Дисперсія послідовностей таких об'ємів має бути мінімальною.
2. Дисперсія елементів послідовності $\{k_i\}$ повинна бути мінімальною.

Тривалість перебування заявки першого рівня у функціональних блоках (ФБ) є випадковою величиною, що залежить від швидкості обробки породжуваних у даному ФБ заявок другого рівня. Спираючись на середньо статистичні значення отримані системою моніторингу, будемо говорити, що час перебування заявки у функціональному блоці є відомим (t_j, де j- номер ФБ). $t = \sum_{j=1}^{m} t_j$, де m – кількість функціональних блоків у системі.

Відома кількість ресурсу (v_j, де j – номер ФБ), яка необхідна для обслуговування заявок другого рівня породжуваних заданим ФБ.

Яким чином розраховувати об'єм ресурсу *V,* який використовується у поточний момент часу було показано в [2], основний принцип полягає в тому, що використовується зворотня система відліку часу. Приймемо за нульовий час закінчення обслуговування заявки, тобто t^0=*t*, далі позначимо періоди часу коли заявки переходять між функціональними блоками: $t^1 = t^0 - t_1$, …, $t^j = t^{j-1} - t_j$, …, $t^m = t^{m-1} - t_m = 0$. Всі заявки, які надійшли протягом інтервалу $[t^1, t^0]$ в момент часу t^0 обслуговуються у першому функціональному блоці. Заявки, які надійшли до системи протягом інтервалу $[t^j, t^{j-1}]$ в момент часу t^0, обслуговуються у *j*-му функціональному блоці.

Таким чином, об'єм ресурсу V^0, що використовується у момент часу t^0 – це сума об'ємів ресурсу v_j^0 зайнятого заявками, які перебувають у j-му ФБ ($j = \overline{1,m}$) в момент часу t^0.

$$V^0 = \sum_{j=1}^{m} v_j^0$$

Значення v_j^0 залежить від кількості заявок, які надійшли до системи протягом часу $[t^j, t^{j-1}]$, та визначається як добуток кількості заявок, на об'єм ресурсу, що необхідних для обслуговування однієї заявки у відповідному функціональному блоці. Оскільки застосовується *схема згладжування вхідного навантаження*, то максимально дупустима кількість заявок які припадають на цей інтервал часу відома. v_j^0 – це добуток кількості заявок, що у момент часу t^0 перебувають у j-му ФБ, на об'єм ресурсу, який потрібен для обслуговування заявок другого рівня породжених j-му ФБ.

Задля забезпечення ефективного згладжування необхідно, щоб умова 1 виконувалася не тільки для об'єму V^0, але й для всіх V^i, $(i = \overline{1,n})$.

Цільова функція має три складові:

- Дисперсія елементів послідовності $\{k_i\}$ повинна бути мінімальною.
- Дисперсія елементів послідовностей $\{V^i\}$ мінімальна
- Середнє значення елементів послідовності {Vj} прямує до максимально можливої кількості ресурсу V_{max} заданого типу , що виділяється для обслуговування заявок вибраного типу сервісу

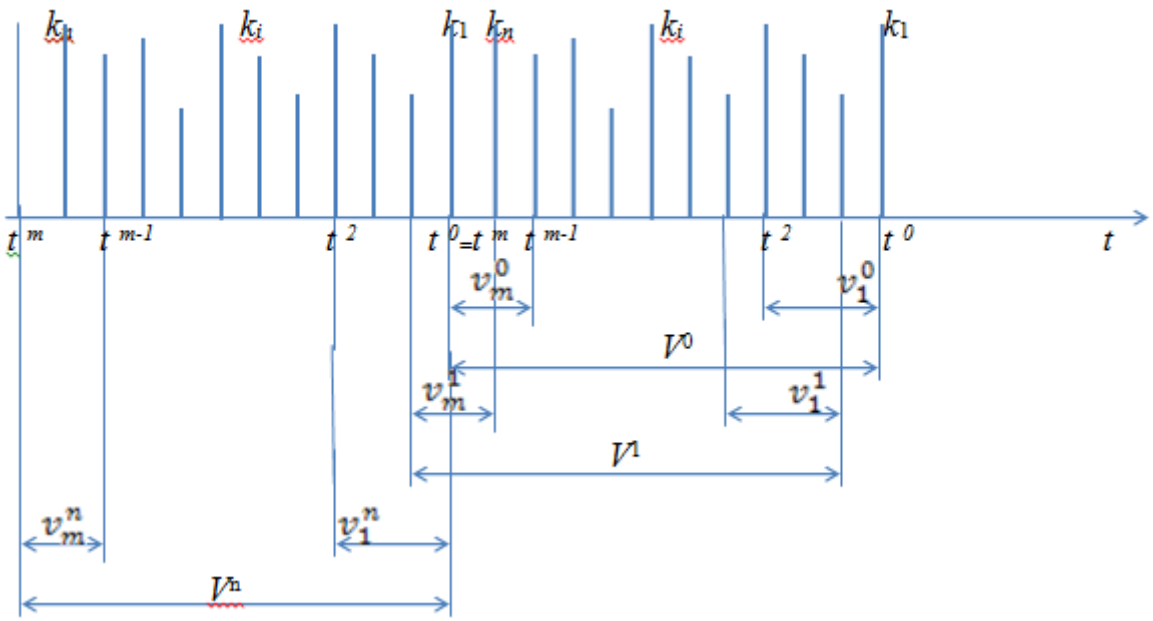

Рис. 3 Схема згладжування вхідного навантаження, з урахуванням об'єму ресурсу, що використовується

Методом розв'язку вибору послідовності {ki} здійснюється за допомогою генетичного алгоритму, таким чином щоб задовольнялися умови 1 і 2:

Геномом виступають елементи послідовності *{k_i}.*

Кросовер: зміна значень елементів послідовності {k_i}.

Завершення алгоритму:

- за часом,
- кількістю розглянутих поколінь
- виродження популяції

В результаті отримуємо послідовність

При використанні запропонованої схеми згладжування вхідного навантаження забезпечується максимально допустима однорідність потоку заявок другого рівня. Оскільки береться до уваги послідовність операцій, які виконуються з заявкою на сервері мобільного оператора, об'єму ресурсу, що потрібен для забезпечення цих операцій. Підбирається послідовність максимально допустимих значень кількості заявок, що надходять у систему за малий інтервал часу. Критеріями оцінки є наближеність загального об'єму, що використовується до максимально допустимого при цьому забезпечується мінімальне середньоквадратичне відхилення від середнього значення кількості заявок, що пропускаються у систему. Фактично приведено спосіб керування чергою заявок, які надходять на серверу оператора.

1 Larysa Globa, Maria Skulish, Andrei Reverchuk. Server resources load monitoring serving different types of services// International Conference TCSET'2014 - February 2014.

2 Larysa Globa, Maria Skulish, Andrei Reverchuk. Manage of incoming application flow to prevent shortage of server resources // International Conference TCSET'2014 - February 2014.

3 Totok, V. Karamcheti. Exploiting Service Usage Information for Optimizing Server Resource Management //ACM Transactions on Internet Technology, Vol. 11, No. 1, Article 1, Publication date: July 2011