МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ «КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
НАЦІОНАЛЬНИЙ ЦЕНТР «МАЛА АКАДЕМІЯ НАУК УКРАЇНИ»

Л.С. Глоба, О.М. Дяденко, А.Ю. Пилипенко, М.А. Скулиш

МАТЕМАТИЧНІ МЕТОДИ АНАЛІЗУ ТА КЕРУВАННЯ ТЕЛЕКОМУНІКАЦІЙНИМИ МЕРЕЖАМИ

монографія

Київ
2017

УДК 621.391+004.7
ББК 32.81
Г 54

Рекомендовано до друку
Вченою радою Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» (протокол № 10 від 3 жовтня 2016),
Вченою радою Національного центру «Мала академія наук України»
(протокол № 8 від 31 жовтня 2016)

Відповідальний редактор: Стрижак О.Є., доктор технічних наук, старший науковий співробітник.

Рецензенти: Лобур М. В., доктор технічних наук, професор; Безрук В. М., доктор технічних наук, професор; Трофимчук О. М., доктор технічних наук, професор.

Автори: Глоба Л. С., доктор технічних наук; Дяденко О. М., кандидат технічних наук; Пилипенко А. Ю., доктор фізико-математичних наук; Скулиш М. А., кандидат технічних наук.

Г 54 Математичні методи аналізу та керування телекомунікаційними мережами : монографія / Л.С. Глоба, О.М. Дяденко, А.Ю. Пилипенко, М.А. Скулиш. – К.: Інститут обдарованої дитини НАПН України, 2017. – 236 с.
ISBN 978-966-2633-84-9

Охоплено весь спектр проблем, які виникають в процесі функціонування телекомунікаційних систем, а також забезпечення процесу надання телекомунікаційних послуг, запропоновано підходи до аналізу та керування телекомунікаційними системами, а також ряд методів та моделей для обслуговування інформаційних потоків в рамках телекомунікаційної системи. Розглядаються проблеми та задачі транспортування інформаційних потоків, задачі керування мультисервісним потоком в комутаційному центрі транспортної мережі, розглядаються питання обслуговування викликів в центрах керування мобільною мережею зв'язку, розглядаються задачі організації роботи та зменшення навантаження на систему тарифікації оператора телекомунікаційних послуг.

Для студентів технічних спеціальностей вищих закладів освіти, аспірантів, які навчаються за напрямком телекомунікації, а також фахівців, які працюють у сфері інформаційно-комунікаційних систем та мереж.

УДК 621.391+004.7
ББК 32.81

ISBN 978-966-2633-84-9

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

ABR	Available Bit Rate
AC	Attached Circuit
ACR	Actual cell rate
ACT-L3	Відео кодек
AF	Блок-функція прикладної програми (Application Function)
AGW	Шлюз доступу (Access Gateway)
API	Інтерфейс програмування прикладних програм (Application Program Interface)
ATM	Asynchronous Transfer Mode
AVP	Формат запису параметрів протоколу Diameter (Attribute Value Pair)
CBQ	Class-Based Queuing
CBR	Constant Bit Rate
CBWFQ	Class-Based Distributed WFQ
CDR	Формат запису і зберігання тарифікаційної інформації (Charging Data Record)
CDV	Cell delay variation
CE	Customer Edge
CGF	Функція збору тарифікаційної інформації (Charging Gateway Function)
CLR	Cell loss ratio
CSCF	Блок управління сесією (Call Session Control Function)
CTD	Cell transfer delay
CTF	Функція тригера тарифікації (Charging Trigger Function)
DDN	Digital Delivery Network
DiffServ	Differentiated Services
DMShort	Назва протоколу між блоками PCRF та OSC
DRM	Цифрове управління авторськими правами (Digital Rights Management)
DRR	Deficit Round Robin
DSCP	DiffServ Code Point
DSL	Digital subscriber line
DWFQ	Flow-Based Distributed Weighted Fair Queuing
ECN	Explicit Congestion Notification
ETSI	Європейський інститут телекомунікаційних стандартів (European Telecommunications Standards Institute)
E-UTRAN	Удосконалена універсальна мережа наземного радіо зв’язку
FR	Frame Relay
FTP	File Transfer Protocol
FWRD	Комутації, якій відповідає механізм пересилання
GCRA	Generic Rate Algorithm
GGSN	Шлюзовий вузол, що обслуговує GPRS трафік (GPRS Gateway Service Node)
GPRS	Пакетний протокол передачі в GSM

GSM	Глобальний цифровий стандарт для мобільного стільникового зв'язку
GTP	Протокол передачі даних, що працює в мережах GSM , UMTS, LTE, оснований на IP протоколі (GPRS Tunnelling Protocol)
HDTV	High-definition television
HSSI	High-Speed Serial Interface
IETF	Відкрите міжнародне співтовариство проектувальників (Internet Engineering Task Force)
IGP	Interior gateway protocol
IMS	Мультимедійна підсистема управління послугами на базі IP протоколу
IP	Internet Protocol
IPER	IP packet error ratio
IPLR	IP packet loss ratio
ISP	Провайдер Інтернету (Interner Service Provider)
IWF	Interworking Function
LSP	Label Switched Path
LTE	Мобільний протоколу передачі даних, довгостроковий розвиток (Long Term Evolution)
MCR	Minimum Cell Rate
MDRR	Modified DRR
MMS	Мультимедійне повідомлення (Multimedia Message)
MPLS	Multi Protocol Label Switching
NGN	Next Generation Network
NSP	Natural Switching Point
OFDMA	Доступ на базі ортогонального частотного розділення каналів с мультиплексуванням ОЧД (Orthogonal Frequency-Division Multiple Access)
OSC	Підсистема онлайн тарифікації (On-/off-line Charging System)
OSI/ISO	Open System Interconection/ International Organization of standartization
PAN	Protected Area Network
PCC	Система управління обробкою і тарифікацією викликів (Policy Control and Charging)
PCEF	Блок виконання правил обслуговування та тарифікації викликів (Policy and Charging Enforcement Function Policy)
PCN	Personal Communication Network
PCR	Peak Cell ate
PCRF	Блок формування правил обробки та тарифікації викликів (Policy Charging and Rules Function)
PDU	Protocol Data Uunit
PE	Provider Edge
PPP	Point –Point Protocol
PQ	Priority Queuing
PSN	Packet Switched Networks
PW	Pseudo Wire
PWE3	Pseudo Wire Emulation Edge-to-Edge
QoS	Quality of Service
QoS	Якість обслуговування (Quality of Service)

RAN	Radio access network
RED	Random Early Detection
RM	Resource management
RNC	Radio Network Controller
RSVPR	Resource Reservation Protocol
SDH	Synchronous Digital Hierarchy
SDP	Мережевий протокол, що застосовується для опису сесії передачі даних (Session Description Protocol)
SGSN	Вузол обробки GPRS інформації (Serving GPRS Support Node)
SIP	Протокол встановлення сесії (Session Initiation Protocol)
SMS	Short Massage System
SONET	Synchronous Optical Networking
SPD	Selective Packet Discard
STM-1	Synchronous Transport Module
TCP	Transmission Control Pprotocol
TDM	Time division management
TISPAN	Підрозділ ETSI по стандартизації, що спеціалізується на конвергентних послугах і протоколах телекомунікаційного та Інтернет середовища (Telecommunications and Internet converged Services and Protocols for Advanced Networking)
Ty	Інтерфейс між блоками PCRF та OSC
UBR	Unspecified Bit Rate
UMTS	Universal Mobile Telecommunication System
VBR	Variable bit rate
VoIP	Voice over IP
VPI	Virtual Circuit Identifier
VPN	Virtual Private Network
WFQ	Weighted Fair Queuing
WiMAX	Всесвітній доступ для взаємодії мікрохвильових мереж (Worldwide Interoperability for Microwave Access)
WRED	Weighted Random Early Detection
WRR	Modified Weighted Round Robin
АСШС	Асимптотично самоподібний в широкому сенсі
БС	Базова станція
ВК	Вузол комутації
ВМ	Взаємне мовчання
КБС	Контролер базових станцій
КК	Комутація каналів
КП	Комутація повідомлень
КП-В	Комутація пакетів – віртуальний режим
КП-Д	Комутація пакетів – датаграмний режим
КЦ	Комутаційний центр
МІ	Мовний імпульс
НМ	Неперервна мова
НОР	Незалежні та однаково розподілені
ПМ	Пауза мовчання
ПНМ	Пауза неперервної мови

ССШС	Строго самоподібний в широкому сенсі
ТЕ	Traffic Engineering
ТКМ	Телекомунікаційна мережа

ПЕРЕДМОВА

Комунікаційні мережі так швидко розвиваються, що скоро стане можливим отримувати довільні сервіси у кожному місці так просто, як отримуються сервіси електрики та телефонії. Всепроникаючий комп'ютинг це середовище, в якому люди взаємодіють з вбудованими пристроями (комп'ютерами), які об'єднані у мережу пристроїв, що знають про своє оточення та мають надавати сервіси або отримувати сервіси від інших вузлів мережі. Така можливість забезпечується за рахунок насичення обчислювальними ресурсами, що з'єднані з користувачами бездротовим зв'язком.

Системи зв'язку за технологією 5G, яка активно розробляється провідними вченими світу та України, використовують мікро-, піко-, фемто- сот малих розмірів, для процесів обробки інформації поєднання малих сот за допомогою розподілених центрів обробки даних. В таких системах усі технології як комунікаційні, так і інформаційні мають працювати разом.

Все більшого поширення набувають хмарні технології, які передбачають обслуговування різноманітних мультисервісних інформаційно-обчислювальних процесів на потужностях крупних Дата центрів. Поява концепції програмно керованих мереж та концепції віртуалізації мережевих функцій відкриває нові можливості для світу телекомунікаційних систем, в той же час виникає необхідність у нових підходах, методах та концепціях організації процесу обслуговування сервісів. На конгресі в Барселоні в 2013 році (Mobile World Congress) компанія Huawei представила версію SDN для сукупності технологій GSM/UMTS/LTE/Wi-Fi в рамках SDN-ініціативи, яку названо SoftCOM, а у 2014 році комерційну версію хмарно-орієнтованої платформи SoftMobile. Так функції керування розподіленням радіоресурсів в мережі радіодоступу LTE, а також розподілення ресурсів мереж радіодоступу різних стандартів (GSM, UMTS, LTE) реалізуються у віртуальній програмній мережі RAN на базі Центру Обробки Даних (ЦОД).

Усі елементи базової мережі LTE (EPC) реалізуються у вигляді програм на базі високонадійного Центру Обробки Даних. Керування мережею здійснює єдиний контролер, де зберігають інформацію щодо стану усієї мережі. Мобільні та мережеві сервіси реалізуються на Центрах Обробки Даних операторів, IT або сервіс-провайдерів, та взаємодіють з мобільною мережею через стандартні програмні інтерфейси.

Отже, інфраструктурний оператор втрачає свої позиції як обов'язкового власника усієї технологічної інфраструктури та стає власником як програмного забезпечення віртуальної інфраструктури, так і частини інфраструктури, що залишилася поза хмарною інфраструктурою.

Розвиток комунікаційних систем сьогодення є стрімким й через впровадження на ряду з сучасними апаратно-технічними рішеннями програмних комплексів, які здійснюють не лише моніторинг, але й частково заміщують апаратні засоби в процесі надання комунікаційних послуг. В той же час нарощування обчислювальних ресурсів у хмарних Центрах Обробки Даних збільшують можливості процесів аналізу мереж та їх керування. Тому розробка відповідних математичних методів є актуальною задачею. У монографії розглянуто нові

математичні методи розв'язку актуальних задач, які виникають в процесі функціонування комунікаційних систем, направлених на підвищення якості обслуговування інформаційних потоків на різних етапах їх передачі.

Як правило, передача інформаційних потоків в комунікаційній мережі включає декілька ключових етапів обслуговування: доступ до мережі, організація транспорту інформації та організація процесів керування, встановлення з'єднання, тарифікація послуг, тощо. Постійне збільшення кількості послуг, які надаються в телекомунікаційній системі, різні вимоги до їх якості, які швидко змінюються, потребують аналізу процесу функціонування складного телекомунікаційного комплексу на кожному з етапів та своєчасного керування.

В монографії охоплено весь спектр проблем, які виникають в процесі функціонування телекомунікаційних систем, а також забезпечення процесу надання телекомунікаційних послуг, запропоновано підходи до аналізу та керування телекомунікаційними системами, а також ряд методів та моделей для обслуговування інформаційних потоків в рамках телекомунікаційної системи.

В розділах розглядаються проблеми та задачі транспортування інформаційних потоків, задачі керування мультисервісним потоком в комутаційному центрі транспортної мережі;

розглядаються питання обслуговування викликів в центрах керування мобільною мережею зв'язку;

розглядаються задачі організації роботи та зменшення навантаження на систему тарифікації оператора телекомунікаційних послуг.

Монографія підготовлена колективом авторів кафедри інформаційно-телекомунікаційних мереж Інституту телекомунікаційних систем КПІ ім. Ігоря Сікорського у складі проф. Глоба Л.С., Дяденко О.М., Пилипенко А.Ю., Скулиш М.А. Монографія значно поліпшена завдяки науковим роботам аспірантів кафедри інформаційно-телекомунікаційних мереж Інституту телекомунікаційних систем КПІ ім. Ігоря Сікорського, а саме аспірантки В. Ф. Чердинцевої (підрозділ 4.1) та аспірантки С. В. Суліми (розділ 5).

Автори сподіваються, що матеріал монографії буде корисним студентам старших курсів, аспірантам, які навчаються за напрямком телекомунікації, а також фахівцям, які працюють у сфері інформаційно-комунікаційних систем та мереж.

1. МОДЕЛЬ ФУНКЦІОНУВАННЯ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ СИСТЕМ

1.1. Архітектура рішень

На сучасному етапі розвитку процесів проектування інформаційно-комунікаційних систем останні все частіше розглядаються як розподілені обчислювально-інформаційні системи, керовані розподіленими потоками даних в інтегрованому, територіально розподіленому інформаційно-комунікаційному середовищі. Такий підхід вимагає нових концепцій, технологій, архітектурних рішень щодо побудови інформаційно-комунікаційної системи, яка базується на територіально-розподіленому інформаційно-комунікаційному середовищі, що використовує динамічні, гнучкі процеси одержання та обробки корпоративної інформації. Ці процеси спираються на розподілені обчислювальні та інформаційні ресурси, які транспортуються в інформаційно-комунікаційному середовищі.

На сьогоднішній день обробка значних об’ємів різнорідних потоків інформації для мереж зв’язку є звичайною справою. Для абонентів мережа представляється, як чорна скринька, що забезпечує надання необмеженого набору послуг із гнучкими можливостями щодо їх керування, персоналізації та створення нових послуг. З іншого боку інформаційно-комунікаційна мережа – це система, яка об’єднує різнорідне обладнання з розподіленою системою керування, це середовище, де одночасно функціонують різні технології націлені на виконання своєї підзадачі у розрізі широкого спектру надання послуг зв’язку. Розроблено ряд концепцій та технології, в рамках яких передбачено реалізацію універсальної транспортної мережі з розподіленою комутацією, винос функцій надання послуг у вузли мереж та інтеграцію з традиційними мережами зв’язку.

Під розподіленими системами будемо розуміти територіально розподілені інформаційно-комунікаційні середовища (DCE), що забезпечують надійне та ефективне функціонування територіально-розподіленої, взаємозалежної через інформаційно-комунікаційну мережу, інтегрованої програмно-технічної системи. Такі системи засновані на незалежно працюючих технічних засобах, що використовують спільні розподілені дані і обчислювальні процеси, які інтегруються за допомогою інформаційно-комунікаційної мережі, а також інформаційних потоків у розподілені динамічні маршрути обробки інформації, виконують спільне завдання або спільну групу завдань, об'єднаних метою ефективного функціонування системи передачі інформації.

Загальними рисами сучасних технологій розподілених інформаційно-комунікаційних систем є:

- наявність територіально-розподіленого інформаційно-комунікаційного середовища, яким є фактично інформаційно-комунікаційна мережа та програмно-апаратні пристрої та компоненти, що керують нею;
- просторовий розподіл інформації (даних), а також просторовий розподіл її обробки – фізично і логічно об'єднаної у рамках бізнес-процесів, що відбуваються у системі;

- наявність єдиної сфери надання послуг в цілому та можливість регламентованого надання інформаційних послуг конкретному користувачеві зокрема;
- велика кількість взаємодіючих частин і елементів, що складають систему;
- можливість поділу на групи найбільш тісно взаємодіючих елементів (підсистем) і виділення належних цим підсистемам бізнес-процесів;
- стійкість до зовнішніх та внутрішніх завад і наявність самоорганізації і адаптації до всіляких збуджувань;
- саморегуляція процесів у системі;
- наявність зовнішніх систем-учасників процесу надання комунікаційних послуг;
- уніфікація інтерфейсів взаємообміну системи щодо зовнішніх систем і у середині самої системи;
- можливість контролю і керування безпосередньо самими процесами, а не тільки послідовністю виконуваних робіт (просуванням документів);
- контроль якості і термінів виконання підпроцесів у рамках процесу, який виконується, як і самого процесу в цілому;
- наявність систем розрахунку та біллінгу інформаційних потоків, які передаються системою;
- збереження і доступність інформації щодо всього процесу, термінів його виконання і т.д.

Під програмним забезпеченням (ПЗ) інформаційно-комунікаційних систем розуміють відкрите програмне забезпечення, архітектура якого містить усі складові, необхідні для організації «прозорої» взаємодії прикладних програм з метою надання інформаційних послуг, що виконуються на різних платформонезалежних вузлах інформаційно-комунікаційної мережі, і яке характеризується:

- можливістю інтеграції територіально-розподілених інформаційних процесів і ресурсів на базі інформаційно-комунікаційної мережі;
- можливістю повноцінно і комплексно функціонувати на об'єднаних різнорідних комп'ютерних і інформаційно-комунікаційних платформах;
- розробляється на основі затверджених стандартів, що забезпечує сумісність прикладного програмного забезпечення для різних платформ;
- відсутністю впливу методів і засобів інтеграції територіально-розподілених процесів і ресурсів на функціонування самої системи і засоби отримання інформації в ній;
- забезпеченням для користувача ілюзії роботи з терміналом єдиного комплексу, у той час як ресурси такого комплексу територіально розподілені.

Отже, відкрите програмне забезпечення інформаційно-комунікаційних систем має такі характеристики:

- сумісність різних інформаційно-комунікаційних та програмних платформ;
- керованість всіх обчислювальних та інформаційних ресурсів системи з будь-якого місця і з будь-якої точки;
- інтегрованість, тобто структурна розподіленість компонентів системи поряд з їхньою інтегрованістю за допомогою інтерфейсів взаємодії;

- можливість зміни масштабу програмного забезпечення в умовах різних апаратних та інформаційно-комунікаційних платформ;
- доступність програмного забезпечення для модифікації, розвитку і реструктуризації;
- стандартизація програмних компонентів системи;
- можливість представлення системи, фактично розподіленої як інформаційно, функціонально, так і за ресурсами, як логічно єдиного функціонального комплексу.

При побудові розподілених інформаційно-комунікаційних систем використовуються, насамперед, базові моделі взаємодії відкритих систем OSI/Internet/DoD, що дозволяє розглядати сервіси незалежно від програмної платформи вузлових комп'ютерів. Деталі реалізації мережі залишаються схованими від користувачів. Користувач взаємодіє з інтерфейсом DCE, що працює поза стандартних мережних операційних систем (ОС) і забезпечує такі інструментальні засоби і служби:

- віддалений виклик процедур, що дозволяє ініціалізувати виконання прикладних програм на кожному з комп'ютерів мережі. При цьому, з погляду користувача, віддалені ресурси мають такий самий вигляд, як і локальні;
- паралельну обробку незалежних програмних сегментів в різних вузлах інформаційно-комунікаційної мережі;
- засоби захисту даних, що працюють в масштабах усієї мережі. Її користувачі автоматично виявляються захищеними від нападу комп'ютерних вірусів і неавторизованого доступу з інших комп'ютерів;
- усі вузли мережі з архітектурою DCE, синхронізовані спільним тактовим генератором;
- прозоре подання інформаційних та обчислювальних ресурсів мережі, яке гарантує використання їх з максимальною ефективністю і збільшує ймовірність того, що конкретний ресурс виявиться доступним саме тоді, коли в ньому виникне потреба.

Комунікаційна обчислювальна мережа – це мережа обміну та розподіленої обробки інформації, що утворюється множиною взаємопов’язаних абонентських систем та засобами зв’язку, які є постачальниками або споживачами інформації. Засоби зв’язку та обробки інформації орієнтовані на колективне використання загальних ресурсів мережі – апаратних, інформаційних, програмних.

Основними компонентами апаратного забезпечення є комп’ютери різних типів та класів або інші термінали мережі, комутаційні вузли та засоби зв’язку.

Поняття «архітектура» в інформаційних технологіях поєднує в собі склад і структуру керованих користувачем інтегрованих апаратних і програмних засобів, що визначають ефективність і функціональні можливості комплексу програмно-технічних засобів переробки інформації і маніпулювання нею.

Елементи архітектури дозволяють користувачеві змусити комплекс програмно-технічних засобів робити те, що потрібно для кожного конкретного алгоритму в процесі переробки інформації і маніпулювання нею.

Аналіз можливих типів архітектур дозволяє говорити про використання архітектури «client-server» як базового принципу побудови розподілених систем, у тому числі і систем паралельних обчислень. При цьому технології Intranet, побудовані на групі протоколів TCP/IP і HTTP, повинні розглядатись не стільки як

середовище для пошуку інформації, скільки як сучасна реалізація зазначеної архітектури і технологічної основи для створення розподілених прикладних програм у мережах Internet/Intranet.

В процесі проектування необхідної структури апаратних засобів, об'єднаних у єдину інформаційно-комунікаційну мережу, для розподілених інформаційно-комунікаційних систем виникає проблема кількісної оцінки якості функціонування цих засобів в умовах повсякденного вирішення фіксованого набору функціональних задач за визначеною технологією, що відповідає повсякденним потребам корпорації, де така розподілена система використовується.

Оцінити якість функціонування технічних засобів і сформувати на базі цих засобів оптимальну структуру інформаційно-комунікаційної мережі під конкретне застосування дозволяє теорія систем.

Під системою у даній теорії розуміють оптимальну архітектуру мережного інформаційно-комунікаційного комплексу, що забезпечує необхідну ефективність вирішення фіксованої сукупності прикладних задач за обумовленою технологією та надання інформаційних послуг віддаленому користувачеві.

Розглядувана система функціонує на базі інформаційно-комунікаційної мережі, яка є її апаратною платформою. Ця мережа є системою розподіленої обробки інформації, яка складається як мінімум з двох комп'ютерів, які виконують роль термінального обладнання мережі та взаємодіють між собою за допомогою засобів зв'язку.

Термінальне обладнання, що входить до складу мережі, виконує досить широке коло функцій, основними з яких є:

- організація доступу до мережі;
- управління передачею інформації;
- надання обчислювальних та інформаційних ресурсів і послуг абонентам мережі.

Для організації доступу до мереж NGN використовують:

- інтегровані мережі доступу, що під'єднані до термінальних пристроїв мультисервісних інформаційно-комунікаційних мереж та забезпечують підключення користувачів як до мультисервісних, так і до традиційних мереж (наприклад, телефонних мереж загального користування – ТфЗК);
- традиційні мережі (ТфЗК), абоненти яких отримують доступ до мультисервісних мереж через вузли, під'єднані до шлюзів (Media Gateway).

На ТфЗК для доступу використовують абонентський вузол, для збільшення пропускної здатності якого застосовують технологію xDSL, на мережах рухомого зв'язку (3G) – технологію широкосмугового доступу.

Особливостями NGN щодо керування є те, що ці мережі складаються з великої кількості компонентів різного типу, а не з відносно невеликої кількості менш різноманітних великих комутаційних пристроїв, як нині. Крім того, в NGN підтримуватиметься більша кількість інтерфейсів, ніж в існуючих мережах, та більша пропускна здатність.

Система керування NGN – набір рішень, що забезпечують керування мережами, реалізованими на базі різних технологій (фіксовані та мобільні телефонні мережі, мережі передачі даних, сигналізації тощо), які надають різні послуги та побудовані на обладнанні різних виробників. Системи керування будуватимуть з використанням об'єктно-орієнтованої розподіленої структури.

Однією з головних особливостей систем керування NGN є відкрита модульна архітектура, що дозволяє розробляти та впроваджувати нові модулі, працювати з прикладними програмами та модернізувати модулі, які вже існують. Для реалізації інтегрованого керування системами та мережами незалежно від їхнього виробника та технології можна використовувати різні стандарти та протоколи, а саме SNMP, OSI, ASCII, CORBA, SOAP. Наприклад, стандартом керування де-факто в мережах передачі даних є протокол SNMP. У моделі TNM передбачалось використання моделі OSI як основи побудови інформаційно-комунікаційної мережі. Однак реалізація систем керування на базі TNM виявилась складною, повільною та дорогою. В системі на базі TNM недостатньо відпрацьовані питання керування послугами. Останнім часом активно розвиваються та реалізуються рішення з організації керування на базі web-технологій, зокрема web-сервісів та протоколів SOAP, що як раз і спроектовані для організації як мультисервісного доступу, так і для використання широкої гами термінальних пристроїв.

У мережах NGN системи керування насамперед мають вирішити конкретні задачі операторів, рівнева структура TNM вже не буде мати першочергового значення та відійде на другий план. Особливого значення набувають питання керування послугами.

Як засоби зв'язку можна використовувати різні фізичні середовища: коаксіальний кабель, виту пару, оптоволоконний кабель, телефонну лінію. Нині широко застосовують бездротові технології, які передають інформацію за допомогою радіохвиль або інфрачервоного випромінювання. Засоби зв'язку мають забезпечувати надійну передачу інформації між абонентами мережі.

При розгляді інформаційно-комунікаційних мереж широко використовують поняття «клієнт» і «сервер».

На структурному рівні під сервером розуміють комп'ютер або інший тип термінального обладнання, який надає свої ресурси іншим терміналам, які називають клієнтами.

На програмному рівні під сервером і клієнтом розуміють програми**,** що виконують відповідно функції надання та використання мережних ресурсів. Крім того, комп'ютери, за допомогою яких користувачі отримують доступ до ресурсів інформаційно-комунікаційної мережі, називаються *робочими станціями*.

Відповідно до функціонального призначення комп'ютерів мережі прийнято розподіляти на одноранговi мережі та мережі на основі серверів (серверні). В одноранговій мережі всі комп'ютери рівноправні, кожний з них може виступати як в ролі клієнта, так і в ролі сервера. При цьому ресурси кожного комп'ютера умовно розподіляються на локальні та мережні.

Локальними називаються особисті ресурси кожного з комп'ютерів, незалежно від того, підключений він до мережі чи ні.

Мережні ресурси – це та частина локальних ресурсів, які кожний комп'ютер надає в загальне користування іншим комп'ютерам. Якщо один з комп'ютерів мережі використовує ресурси іншого комп'ютера, то він виступає як клієнт. Відповідно, комп'ютер, який надає ресурси, розглядається в такий момент як сервер. Одноранговa організація, як правило, використовується в невеликих мережах, які включають не більше 10 комп'ютерів.

В мережах на основі серверів виділяються окремі комп'ютери для серверів і для клієнтів. Для кожного виду мережних ресурсів може бути створено свій сервер,

наприклад файловий сервер (файл-сервер), сервер друку, сервер бази даних, сервер прикладних програм тощо.

В свою чергу інформаційно-комунікаційні мережі поділяються на глобальні та локальні мережі. Глобальні мережі охоплюють значні території, найвідомішою серед яких є глобальна мережа Internet. Локальна мережа функціонує, як правило, в межах окремих будівель, підрозділів тощо. Одним з сучасних напрямів розвитку інформаційно-комунікаційних мереж є поєднання локальних і глобальних мереж в межах корпорації (створення так званих корпоративних мереж) на основі Internet-технологій.

Кожна інформаційно-комунікаційна мережа має свою визначену архітектуру, що характеризується топологією, протоколами, інтерфейсами, мережними програмними та технічними засобами. Схематично основні компоненти такої мережі наведено на рис. 1.1.

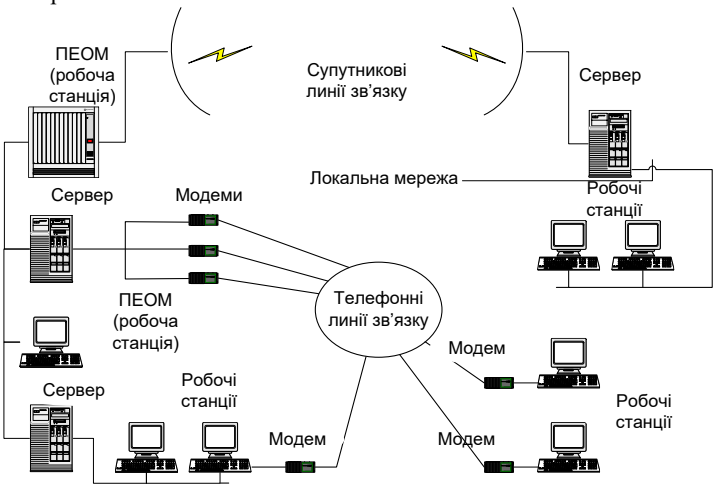


Рис. 1.1 Основні компоненти інформаційно-комунікаційної мережі

Топологія інформаційно-комунікаційної мережі відображає структуру зв'язків між її основними елементами.

Протоколами є правила взаємодії функціональних елементів мережі.

Інтерфейси – це засоби поєднання функціональних елементів мережі. Слід звернути увагу на те, що як функціональні елементи можуть виступати і окремі пристрої, і програмні модулі. Відповідно до цього розглядають апаратні і програмні інтерфейси.

Під мережними технічними засобами розуміють пристрої, що забезпечують поєднання комп'ютерів в єдину інформаційно-комунікаційну мережу. До них належать мережні контролери, вузли комутації, хаби, свічі тощо.

Мережні програмні засоби здійснюють керування роботою інформаційно-комунікаційної мережі та забезпечують відповідний інтерфейс з користувачами. До мережних програмних засобів належать мережні операційні системи та допоміжні

сервісні програми. Кожна зі складових архітектури мережі характеризує її окремі властивості, а тільки їхня сукупність характеризує всю мережу в цілому.

Перша і головна задача розглядуваної теорії полягає у виробленні методології аналізу інформаційно-комунікаційних систем, застосування якої дозволить оцінювати ефективність використання їхніх ресурсів у конкретних умовах експлуатації.

Друга цільова задача – створення методології синтезу оптимальної архітектури системи, зокрема архітектури інформаційно-комунікаційної мережі, для фіксованих умов експлуатації – заснована цілком на методології аналізу. Інструментарій теорії передбачає оперування тільки з абстрактними моделями реальних об'єктів.

Концепція моделювання процесів функціонування системи підпорядкована головній цільовій задачі – оцінюванню ефективності використання ресурсів системи.

Склад досліджуваних ресурсів, а також змінних і параметрів, необхідних для опису їх, обумовлюється призначенням інформаційно-комунікаційної мережі та особливостями її архітектури.

Архітектуру мережі зв’язку, побудованої відповідно до концепції NGN, показано на рис. 1.2.

В основу мережі NGN покладено універсальну транспортну мережу, що реалізує функції транспортного рівня та рівня керування комутацією і передачею.

До складу транспортної мережі NGN можуть входити:

- транзитні вузли, які виконують функції переносу та комутації;
- кінцеві вузли, що забезпечують доступ абонентів до мультисервісної мережі;
- контролери сигналізації, які виконують функції обробки інформації сигналізації, керування викликами та з’єднаннями;
- шлюзи, що дозволять підключити традиційні мережі зв’язку.

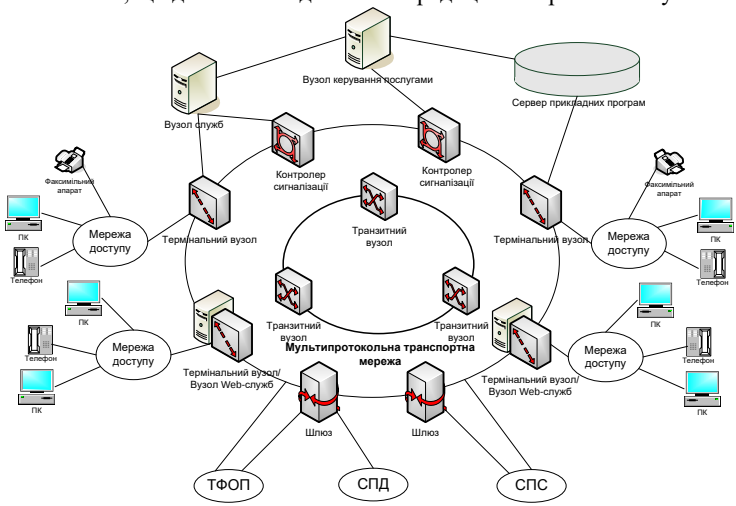


Рис. 1.2 Архітектура мереж зв’язку NGN

Контролери сигналізації можна винести в окремі пристрої, які обслуговують декілька вузлів комутації. Використання загальних контролерів дає змогу розглядати їх як єдину систему комутації, розподілену в мережі.

Розподілена система комутації не тільки спрощує алгоритми встановлення з'єднань, але й є найекономічнішою для операторів і постачальників послуг, оскільки дозволяє замінити дорогі системи комутації великої ємності невеликими, гнучкими і доступними за вартістю навіть дрібним постачальникам послуг.

Призначення транспортної мережі – надання послуг переносу інформації. *Інформаційно-комунікаційні послуги реалізуються на базі вузлів служб (SN) і/або вузлів керування послугами (SCP).*

SN – устаткування постачальників послуг, яке можна розглядати як сервер прикладних програм для інформаційно-комунікаційних послуг, клієнтська частина яких реалізується термінальним устаткуванням користувача.

SCP – елемент розподіленої інформаційно-комунікаційної платформи, який виконує функції керуваня логікою і атрибутами послуг.

Сукупність декількох вузлів служб або вузлів керування послугами, задіяних для надання тієї самої послуги, утворюють платформу керування послугами. До складу платформи також можуть входити вузли адміністративного керування послугами і сервери різних прикладних програм.

Термінальні/термінально-транзитні вузли транспортної мережі можуть виконувати функції вузлів служб, тобто склад функцій кінцевих вузлів можна розширити завдяки додаванню функцій надання послуг. Для побудови таких вузлів застосовують технологію гнучкої комутації (Softswitch).

1.2. Характеристика системи управління комунікаційними мережами

Продуктивність телекомунікаційних систем, побудованих на принципах NGN, багато в чому залежить від ефективності реалізації функцій системи мережевого управління. В мережах наступного покоління система мережевого управління здійснює управління на кожному з трьох рівнів (транспортному, комутації та передачі інформації) в підсистемах адміністративного управління, технічної експлуатації та динамічного управління (рис. 1.3).

Система динамічного управління забезпечує розподіл потоків інформації з метою найкращого використання ресурсів мережі при задоволенні вимог абонентів. Вона заснована на ієрархічній структурі протоколів. Ця система здійснює управління потоками пакетів як таких, що знаходяться в мережі на обслуговуванні, шляхом вибору допустимих маршрутів передачі повідомлень, так і на етапі введення в мережу за рахунок заборони доступу. Технічною базою системи динамічного управління являються вузли комутації.

Система адміністративного управління призначена для вибору варіантів реорганізації структури мережі у випадку масових руйнувань об’єктів мережі, відображення стану мережі на робочих місцях операторів, ведення статистики передачі службової інформації. Крім того, вона забезпечує відображення стану елементів і всієї мережі, а також прийняття рішення на усунення виникаючих несправностей.

В процесі набору статистики повинні бути отримані такі інтегральні ймовірнісно-часові характеристики, як ймовірність та час доставки повідомлень, а також дані про використання каналів і центрів комутації в різний час. Дана інформація необхідна для прийняття рішення про реконструкцію мережі.

У випадку виходу з ладу елементів мережі потрібно визначити ступінь руйнування та можливості відновлення елементів, можливість отримання додаткових ресурсів з каналів і трактів первинної передачі, використання резервних вузлів комутації та ліній зв'язку. У випадку недостачі ресурсів мережі повинна бути передбачена можливість примусового відключення деяких користувачів. Технічну базу системи адміністративного управління складає сукупність пунктів управління мережею та елементами мережі, суміщених з вузлами зв'язку.

Ефективність комунікаційних мереж, побудованих на принципах NGN, багато в чому залежить від ефективності рішення задач управління мережевими ресурсами, котрі в першу чергу напрямлені на забезпечення QoS.

Засоби управління мережевими ресурсами повинні приймати саму активну участь в процесі функціонування рівня транспорту та рівня доступу.

Рис. 1.3 Структура системи мережевого управління

1.3. Характеристика служби якості обслуговування

Необхідність у множинному доступі до розділюваних ресурсів обумовила створення систем, які повинні не тільки надавати користувачу необідні послуги, але і забезпечувати виконання вимог по ймовірнісно-часовим характеристикам доставки повідомлень – «якість обслуговування» (Quality of Service, QoS).

У контексті взаємодії відкритих систем можна визначити групи характеристик, прив'язавши їх до специфіки багаторівневої взаємодії. В результаті можна прийти до наступної класифікації характеристик:

- часові характеристики – затримки, час появи та тривалість виконання процесів передачі та обробки інформації;
- характеристики продуктивності – завантаженість обладнання, швидкість передачі, алгоритми обробки інформації і т.д.;
- характеристики зв'язності – часова, просторова та протокольна узгодженість передачі, прийому та обробки інформації;
- характеристики цілісності – стійкість, достовірність, а також точність передачі та обробки інформації;
- характеристики збереженості – об'єм, тривалість та спосіб зберігання інформації;
- характеристики безпеки – різні аспекти захисту інформації, управління доступом, аутентифікації;
- характеристики надійності – стійкість до несправностей, час відновлення і т.д.

Шляхом управління мережевими ресурсами необхідно забезпечити виконання вимог за вказаними характеристиками для кожного каналу.

До основних показників QoS відносять (рис. 1.4) показники продуктивності (пропускної здатності), показники часової прозорості (часові показники) та показники семантичної прозорості (показники надійності доставки пакетів).

Основний вплив на наведені показники якості обслуговування здійснюють засоби та технології управління мережевим трафіком.

Рис. 1.4 Структура системи мережевого управління

1.4. Особливості мереж нового покоління (NGN)

Мережа нового покоління, або Next Generation Network (NGN), – поняття не нове і вже кілька років обговорюється в середовищі фахівців в галузі зв'язку. Про мережах NGN відомими авторами написані книги [48, 49]. Однак для багатьох фахівців NGN все ще асоціюються просто з новими мережами: адаптивними, інтелектуальними, мультисервісними або IP-мережами. Автор підготував короткий огляд особливостей NGN.

Передумови появи NGN

Почнемо з того, що розвиток ринку послуг зв'язку призвів до наступних передумов появи мереж NGN:

- масовому впровадженню сучасних систем і засобів зв'язку, характерні риси яких – мультисервісність і мультипротокольність;
- істотної зміни мережевих архітектур: відмови від жорсткої ієрархії, характерної для класичних телефонних мереж загального користування (ТМЗК), під впливом впровадження нових засобів зв'язку, принципів передачі та обробки інформації;
- функціональному поділу рівнів транспортної комутованої мережі і рівня формування послуг, що виник в результаті впровадження інтелектуальних мереж (IN) і був закріплений в NGN (завдяки Інтернету, оператору необов'язково мати власну транспортну мережу, а спектр послуг вийшов за рамки традиційних послуг зв'язку; розмитим виявилося і поняття-концепція «телематичні служби»);

- загострення конкуренції в динамічних секторах ринку, таких як мобільний зв'язок, Інтернет, послуги для корпоративних користувачів;
- розділенню бізнес-моделі оператора нових послуг на дві частини: інфраструктурну (створення та обслуговування мережі) та сервісну (пов'язану з маркетингом);
- наявності проміжних ланок – віртуальних операторів, які формують і реалізують пакети послуг з доданою вартістю, як це роблять системні інтегратори в IT;
- зміни статусу інфокомунікаційних послуг: власне мережа втрачає свою цінність, її набувають послуги;
- зменшення ролі / частки голосових послуг в сучасних пакетах Triple Play (TP) і Quadruple Play (QP);
- використанню умовно безкоштовних послуг, заснованих на експлуатації мережі Інтернет (наприклад, послуга, що надається по Skype);
- зниження інвестиційної привабливості, конкурентоспроможності та рентабельності традиційних систем зв'язку.

Основні поняття

Синонімами NGN, на думку ряду авторів, можуть бути поняття: адаптивні мережі (AN – Adaptive Networks), інтелектуальні мережі (IN – Intelligent Networks) і мультисервісні мережі (MN – Multiservice Networks). Однак поняття AN може однаково добре ставитися до традиційних і NGN-мереж. Поняття IN впроваджено в 1986 г. (Ameritech, США) у зв'язку з появою системи сигналізації SS7 (ОКС-7), а поняття MN відображає суть конвергенції мереж з комутацією каналів і пакетів. З позицій систем передачі даних (СПД) мережа NGN це мережа Інтернет наступного покоління на базі протоколу IPv6 з його новою (і без обмежень) структурою IP-адреси. З позицій мобільних мереж це мережі покоління 3G і вище, що використовують для свого управління підсистеми моделі ОКС-7. З позицій класичної телефонії це мережа IP-телефонії (IPT), керована програмним комутатором (softswitch).

Фактично ж NGN базується на уявленні про новий тип мережі, введеному в рек. ITU-T Y.100 (6.98) у зв'язку із завданням створення глобальної інформаційної інфраструктури (GII – Global Information Infrastructure), де зазначено, що в цій мережі «всі види інформації, включаючи голос, дані або відео / мультимедіа просто зводяться до цифрових потоків біт для передачі їх по шляху поширення (або по цифровій мережі)». Більш того, було підкреслено, що «вказане не виключає можливості розриву зв'язків між мережами і їх корисними навантаженнями». Сама ж мережа NGN «розглядалася як реалізація GII або принаймні деяких її компонентів» (Y.2011).

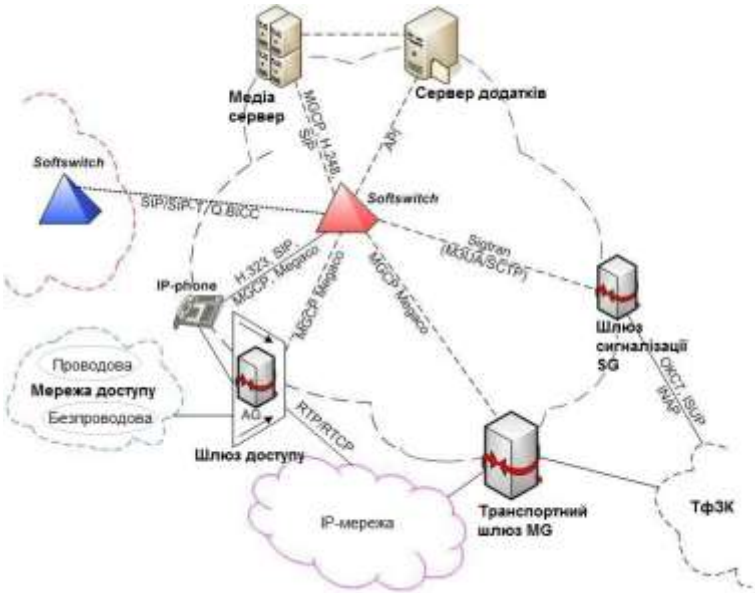


Рис. 1.5 Приклад мережі NGN

Модель OSI і NGN

Хоча NGN, як зрозуміло з визначення, є мережею пакетної комутації, структура рівнів і ієрархія протоколів моделі OSI не може бути застосована безпосередньо в мережі NGN і може інтерпретуватися лише певним чином.

Основна відмінність мережі NGN

Функціональна модель мережі NGN може бути грубо представлена у вигляді двох взаємодіючих, але незалежних шарів: шару сервісів NGN і транспортного шару NGN, див. рис. 1.6.

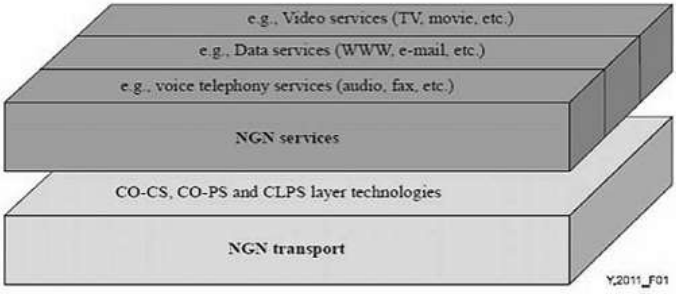


Рис. 1.6 Двошарова модель мережі NGN

Вказані типи сервісу включають: телефонію (класи: audio, fax та ін.); дані (www, e-mail і ін.) і відео (TV, movie та ін.). Кожен шар двошарової моделі

представлений трьома площинами (див. рис. 1.7): площиною користувача; площиною управління і площиною менеджменту.

Рис. 1.7 Базова еталонна модель мережі NGN (рек. МСЕ-Т Y.2011)

Визначення NGN дано в рек. Y.2001 (12.04). NGN – мережа пакетної передачі, здатна забезпечити телекомунікаційні сервіси і використовувати багато широкосмугових, що підтримують QoS транспортні технології, в яких сервісні функції незалежні від цих базових транспортних технологій, каналів. Вона дає вільний доступ до мереж і конкурентним сервіс-провайдерам та / або сервісів на свій вибір. Вона підтримує узагальнену мобільність, яка дозволить послідовно і повсюдно забезпечити сервіс користувачам.

Другим наріжним каменем є взаємозв'язок мережі NGN з принципами мережевої архітектури, викладеними в рек. G.805 (базова функціональна архітектура транспортних мереж), рек. G.809 (функціональна архітектура багаторівневих мереж, без з'єднань між рівнями) і рек. Y.110 (принципи глобальної інформаційної інфраструктури).

Відомо, що рек. G.805 описує функціональну архітектуру транспортних мереж незалежно від технологій. Ця загальна функціональна архітектура може бути використана в якості базової для гармонізації набору архітектур таких транспортних мереж, як ATM, SDH і PDH, а також сполучної ланки для відповідних рекомендацій з менеджменту, аналізу робочих характеристик і специфікації обладнання. Рек. G.809 описує функціональну архітектуру транспортних мереж без попереднього встановлення з'єднань з точки зору їх здатності передавати інформацію. Функціональна і структурна архітектура цих мереж описується незалежно від мережевих технологій. А значить, ці рекомендації повинні бути взяті за основу для опису транспортних мереж, що не використовують попереднього встановлення з'єднань, але реалізують певну технологію.

Транспортні функції

Існує набір транспортних функцій, які відповідають за перенесення цифрової інформації між будь-якими двома географічно розділеними точками. У транспортному шарі може перебувати складний набір багаторівневих мереж, складених з рівнів 1-3 моделі OSI. Транспортні функції в першу чергу забезпечують можливість підключення. Зокрема, транспортний шар полегшує можливість наступних типів підключень: користувача до користувача; користувача до сервісної платформі; сервісної платформи до сервісної платформі.

У загальному випадку всі типи мережевих технологій можуть бути розгорнуті в транспортному шарі, включаючи технології з комутацією ланцюгів (CO-CS) і пакетів (CO-PS), розраховані на попереднє встановлення з'єднань, а також багаторівневі технології пакетної комутації без попереднього встановлення з'єднання (CLPS) відповідно до рек. G.805 і G.809. Для мереж NGN передбачається, що IP може бути розглянутий як протокол, переважний для забезпечення не тільки сервісів NGN, але і легальних супутніх сервісів. Сервісні платформи забезпечують такі сервіси, як телефонія, Web-сервіси та ін. Сервісний шар може (в загальному випадку) охопити цілий ряд складних географічно рознесених платформ, а в найпростішому випадку обмежитися сервісними функціями двох сайтів кінцевих користувачів.

Прикладні функції

Існує набір прикладних функцій, що відносяться до сервісу, який викликається / активізується. Серед сервісів можуть бути, наприклад, голосові сервіси (включаючи телефон), сервіс передачі даних (включаючи сервіси на основі Web, але не обмежуючись ними), відеосервіси (включаючи фільми і TV-програми, але не обмежуючись ними) або комбінації перерахованого вище (наприклад, мультимедійні сервіси, такі як відеотелефонія та ігри). Так як існує багато інших схем класифікації типів сервісу (наприклад, сервіси пакетні / реального часу або сервіси унікастінга / мультікастінга / бродкастінг), на рис. 1.5(верхній шар) наведено приклади сервісів, які можуть працювати в мережі NGN.

Кожен шар охоплює один або кілька рівнів, причому кожен рівень концептуально складається з площини даних (або площини користувача), площини управління і площини менеджменту. У загальному випадку кожен шар буде мати свій набір рольових функцій, гравців та адміністративних доменів (рек. Y.110). Ролі, залучені в сервісне забезпечення, не залежать від тих, що залучені до забезпечення можливості з'єднуватися за допомогою транспорту. Кожен шар повинен оброблятися окремо (з технічної точки зору). Це досягається обов'язковим розбиттям площин користувача (або даних) на дві, розміщені в двох шарах (рис. 1.6). Спираючись на викладене, в NGN і виділені два шари.

Шари NGN-мережі

Сервісний шар NGN

Це частина NGN, що забезпечує функції користувача, які передають сервісні дані, а також функції, які управляють і адмініструють сервісні ресурси та мережеві сервіси, обслуговуючи тим самим різноманітні користувальницькі сервіси та прикладні програми.

Сервіси користувача можуть бути реалізовані рекурсивно за допомогою багатьох сервісних рівнів, наявних у даному шарі. Сервісний шар NGN займається програмами та їх сервісами, які функціонують між однорангових об'єктами. Наприклад, сервіси можуть бути пов'язані з голосовими сервісами, дані або відео, організованими окремо або в комбінації у разі мультимедійних потокыв. З точки зору архітектурних перспектив будь-який рівень шару сервісів розглядається як рівень, що має власні площині користувача, управління та менеджменту.

Транспортний шар NGN

Ця частина NGN забезпечує функції користувача, які передають дані, а також функції, які управляють і адмініструють транспортні ресурси так, щоб переносити ці дані між термінальними закінченнями / вузлами / об'єктами.

Передані таким чином дані можуть бути інформацією користувача або даними управління та адміністрування. Може бути встановлено динамічне або статичне відповідність з інформацією управління або менеджменту, переданої між такими закінченнями / вузлами / об'єктами. Транспорт шару NGN реалізується багаторівневими мережами ітеративно, як описано в рек. G.805 і G.809. З точки зору архітектури кожен рівень транспортного шару розглядається як рівень, що має власні площині користувача, управління та менеджменту.

Площині користувача (або даних), управління та менеджменту існують завжди і для кожного рівня.

На практиці площині управління та менеджменту можуть бути нульовими для даного конкретного рівня.

У мережі NGN, що використовує технології з уніфікованою площиною управління, згідно рек. G.807 / Y.1302, такі як ASON (автоматично перемикається оптична мережа) і GMPLS (узагальнена багатопротокольна комутація по мітках), еквівалентні функції площин управління, реалізовані у всіх рівнях, можуть піддаватися обробці в рамках одного протоколу.

У мережі NGN, що використовує технології з уніфікованою площиною менеджменту, згідно рек. M.3010, еквівалентні функції площин менеджменту, реалізовані на всіх рівнях, можуть піддаватися обробці в рамках одного протоколу (всередині і між шарами NGN).

Для обох шарів NGN (сервісного і транспортного) загальні архітектурні концепції площин даних (або користувача), управління та менеджменту можуть логічно збігатися (як показано на рис. 1.7). На цьому ж малюнку видно, що відбулося розділення не тільки площин користувача на кілька шарів сервісу і транспорту, але також і площин управління та менеджменту. Відносно цих двох площин важливо визначити:

- площину менеджменту NGN як загальну частину, що складається з площин менеджменту сервісного і транспортного шарів;
- площину управління NGN як загальну частину, що складається з площин управління сервісного і транспортного шарів.

Ці визначення, можливо, перекривають загальні для менеджменту та / або управління функції. Важливо те, що концепція площин NGN не має на увазі якусь їх вертикальну інтеграцію, хоча вимагає визначення еталонних точок між площинами різних верств. Все це вводиться, щоб полегшити перехід від функціональних аспектів архітектури NGN до її реалізації шляхом обліку менеджменту та управління. Реалізація, менеджмент і управління мережами NGN не розглядаються в рек. Y.2011.

Зв'язок між основною моделлю NGN і рек. ITU-T

Функціональні принципи архітектури рек. G.805 і G.809 можуть бути застосовані до вертикальної зв'язку між багаторівневими мережами в рамках однієї мережі NGN, а підходи, викладені в рек. Y.110, – до оцінки ролі, гравців та організацій в корпоративної моделі (Enterprise Model), до сервісів та прикладних програм у структурній моделі (Structural Model), до функцій і інтерфейсів в функціональній моделі (Functional Model) і до компонентів в моделі реалізації (Implementational Model).

Узагальнена мобільність – можливість для користувача (або іншого мобільного об'єкта) спілкуватися і мати доступ до сервісів незалежно від зміни їх

становища або технічного оточення. Ступінь можливості бути обслужених може залежати від ряду факторів, включаючи можливості конкретної мережі доступу, угод про рівень обслуговування (якщо такі є) між базовою / домашньою мережею користувача і візитною мережею і т.д. Мобільність включає можливість зв'язку з безперервним обслуговуванням або без обслуговування.

Загальна функціональна модель

Рек. Y.110 формалізує структурну модель, де сервіси та їх компоненти описуються окремо, забезпечуючи:

- корпоративну модель, яка встановлює гравців та їх структурні та інфраструктурні ролі, тобто бізнес-активності в рамках послідовності нарахування вартості;
- модель реалізації, яка сконцентрована на тому, як функції моделі розподіляються і реалізуються обладнанням; вона визначає протоколи, які обслуговують інтерфейси між елементами обладнання. У даному контексті це розглядається як фізична реалізація мережі NGN.

Як і GII, мережа NGN повинна розділяти аналіз сервісів і функцій. Рек. Y.110 може бути використана як керівництво для декомпозиції на інфраструктурні та прикладні сервіси, сервіси Middleware і Baseware. Рек. G.805, G.809, G.807 / Y.1302, M.3010, M.3400, M.3050.x, X.700 і X.701 були розроблені для освітлення функціональних аспектів (транспортної) мережевий операції. При вивченні NGN їх слід брати до уваги, а їх зв'язки між функціями, сервісами і ресурсами повинні бути встановлені для обох шарів NGN.

Ці сервіси і функції пов'язані між собою, оскільки функції зазвичай вбудовані в сервіси. Більше того, існує певна схожість між підтипами цих сервісів і функцій. Однак не існує однозначної відповідності між функціями та сервісами, і це одна з причин, чому вони повинні розглядатися окремо. Одна і та ж функція (наприклад, аутентифікація користувача) може бути використана для доставки двох різних сервісів (див. рек. Y.110, де представлені інфраструктурні та прикладні сервіси; сервіси middleware і baseware, включаючи зв'язкові сервіси та ресурси – компоненти сервісів обробки та збереження).

Зручно об'єднати ці функції в дві групи або площині: одна охоплює всі функції управління, інша – всі функції менеджменту. Групування функцій одного і того ж типу (тобто управління і менеджменту) дає можливість визначити функціональні взаємозв'язки всередині заданої групи, а також інформаційні потоки між функціями в цій групі. Узагальнено це показано на рис. 1.8, який дає уявлення про загальну функціональної моделі.

Цей малюнок показує (в тривимірному вигляді) зв'язок між сервісними ресурсами та функціями сервісного шару NGN, з одного боку, і транспортними ресурсами та функціями транспортного шару NGN – з іншого. Зауважимо, що малюнок показує розділення (рис. 1.7) площині управління та менеджменту, але не показує можливі загальні функції для сервісного і транспортного шарів.

Ресурси

Ресурси забезпечують фізичні та логічні елементи (наприклад, лінії зв'язку, пристрої обробки, ЗУ і т.д.), які, в свою чергу, забезпечують сервіс та функціонування мережі. Як і в GII, забезпечення ресурсами має йти окремо від реалізації функцій і сервісів.

Ресурси можуть бути транспортними, які забезпечують, наприклад, складання списків (комутаторів, маршрутизаторів, ліній зв'язку тощо), або ресурсами обробки і пам'яті, такими, як обробні платформи, на яких можуть бути запущені сервісні програми та прикладні програми (сервісні платформи) або бази даних для зберігання контенту прикладних програм.

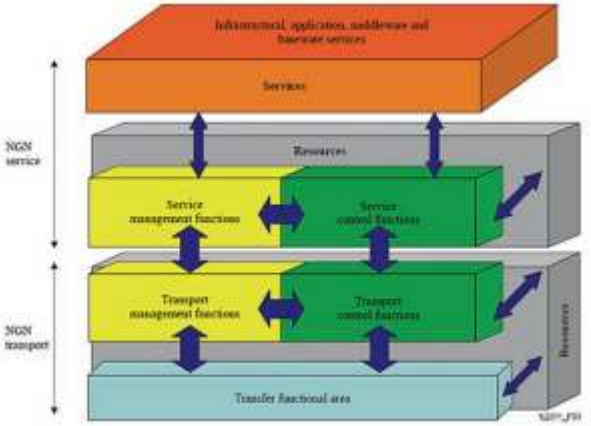


Рис. 1.8. Загальна модель функціонування мережі NGN

1.5. Основні принципи мереж наступного покоління NGN

Термін «мережі наступного покоління» NGN з'явився в телекомунікаційній літературі на початку нового тисячоліття. Ідею розробки NGN, запропоновану в 2001 р Європейським інститутом стандартів електрозв'язку ETSI (European Telecommunications Standards Institute), підтримав Сектор стандартизації телекомунікацій Міжнародного союзу електрозв'язку (МСЕ-Т). У липні 2003 р на спеціальному семінарі з NGN в рамках дослідної комісії ІК 13 МСЕ-Т була утворена Змішана група доповідачів (Joint Rapporteur Group, JRG) по NGN, яка підготувала проекти перших рекомендацій по NGN. Перші дві рекомендації МСЕ-Т – Y.2001 і Y.2011 – були затверджені в кінці 2004 р в новій серії Y. 2000, спеціально виділеної для рекомендацій про NGN. На початок 2011 року в цій серії вже було 70 рекомендацій, які відносяться до так званої першої версії NGN (NGN release 1). Останнім часом в МСЕ-Т розпочато роботи по другій версії (NGN release 2).

Основними об'єктивними передумовами виникнення ідеї мереж наступного покоління NGN є:

- успіхи пакетних технологій передачі інформації, що зумовили бурхливе зростання цифрового трафіку, насамперед за рахунок розширення використання Інтернет;
- збільшення попиту на рухомий зв'язок і на нові мультимедійні служби Triple Play (спільної передачі голосу, відео, даних);
- конвергенція (взаємопроникнення) мереж електрозв'язку та інформаційно-обчислювальних мереж, розвиток інфокомунікаційних мереж.

Слід особливо відзначити одну з основних причин появи ідеї NGN – завершення життєвого циклу експлуатованих цифрових комутаційних станцій телефонної мережі і бажання не замінювати їх такими ж станціями, а радикально модернізувати мережу з метою надання всього комплексу послуг Triple Play. Таким чином, технологія NGN є новим способом розвитку і модернізації існуючих мереж зв'язку і, в першу чергу, телефонних мереж зв'язку загального користування.

Згідно з визначенням, наведеним в Рекомендації MCE-T Y.2001, мережа наступного покоління (NGN) – це мережа з пакетною комутацією, здатна забезпечити користувачів різноманітними вузькосмуговими і широкосмуговими послугами, включаючи послуги телефонного зв'язку, заснована на широкосмугової мережі з пакетною технологією транспортування, що забезпечує необхідну якість послуг QoS (Quality of Service), в якій функції, пов'язані з наданням послуг, не залежать від технологій транспортування інформації. Мережа NGN дає користувачам необмежений доступ до різноманітних послуг провайдерів і підтримує узагальнену мобільність, яка дозволяє користувачам отримати доступ до послуг у будь-якому місці і в будь-який час.

У рекомендації MCE-T Y.2012 перераховані основні принципи функціональної архітектури NGN:

1. Підтримка безлічі технологій доступу – функціональна архітектура NGN повинна володіти гнучкою конфігурацією, необхідної для підтримки безлічі технологій доступу.
2. Розподілене управління – повинен використовуватися принцип розподіленої обробки в пакетних мережах і підтримуватися прозорість розташування для розподілених обчислень.
3. Відкрите управління – мережеві інтерфейси управління повинні бути відкриті для підтримки процесів створення нових і зміни існуючих послуг та підтримки засобів забезпечення логіки послуг сторонніх постачальників.
4. Незалежність надання послуг – процес надання послуг повинен бути розділений між функціями транспортної мережі, що працює з використанням зазначеного вище механізму розподіленого відкритого управління. Це призведе до підтримки конкурентного оточення при розвитку NGN, яке сприятиме прискоренню процесів впровадження нових послуг.
5. Підтримка послуг конвергентних мереж – це необхідно для створення гнучких, простих у використанні мультимедійних послуг для заміщення технічних можливостей конвергентних фіксовано-мобільних мереж за допомогою функціональної архітектури NGN.
6. Розширені можливості безпеки і захисту – це базовий принцип відкритої архітектури, він вимагає обов'язкового захисту мережевої інфраструктури за допомогою механізмів забезпечення відповідних рівнів безпеки і живучості мережі.

Функціональність рівнів базової еталонної моделі NGN (рис 1.6) розкривається в загальній функціональній архітектурі NGN першої версії (NGN release 1), наведеною в рекомендації MCE-T Y.2012 (рис 1.9). На кожному з рівнів використовується декілька функцій. Так для надання послуг кінцевим користувачам використовуються функції підтримки прикладних програм і функції підтримки послуг і відповідні керуючі функції. NGN підтримує точку сполучення з функціональною групою прикладних програм, яка називається інтерфейсом

прикладних програм мережі ANI (Application Network Interface), який реалізує канал взаємодії та обміну інформацією між прикладними програмами і елементами мережі NGN. ANI забезпечує можливості і ресурси, необхідні для реалізації прикладних програм. Транспортний рівень забезпечує послуги IP-з'єднань для користувачів мережі NGN за допомогою функцій управління транспортом, включаючи функції управління мережевими підключеннями NACFs (Network Attachment Control Functions) і функції управління ресурсами та доступом RACFs (Resource and Admission Control Functions).

Відповідно до Рекомендації MCE-T Y.2011 функції транспортного рівня включають безпосередньо транспортні функції і функції управління транспортом.

Транспортні функції забезпечують з'єднання всіх компонент і фізично розділених функцій всередині NGN. Ці функції підтримують передачу медіаінформації, а також інформацією управління (сигналізації) та технічного обслуговування. Транспортні функції включають функції мережі доступу, прикордонні функції, функції транспортного ядра і функції шлюзів.

Функції мережі доступу забезпечують підключення кінцевих користувачів до мережі, а також збір і агрегацію трафіку, що надходить з мережі доступу в транспортну магістраль (ядро). Ці функції також реалізують механізми управління якістю обслуговування QoS, пов'язані безпосередньо з користувальницькою трафіком, включаючи управління буферами, чергами і розкладами, пакетну фільтрацію, класифікацію трафіку, маркування трафіку, визначення політик обслуговування і формування профілю передачі трафіку.

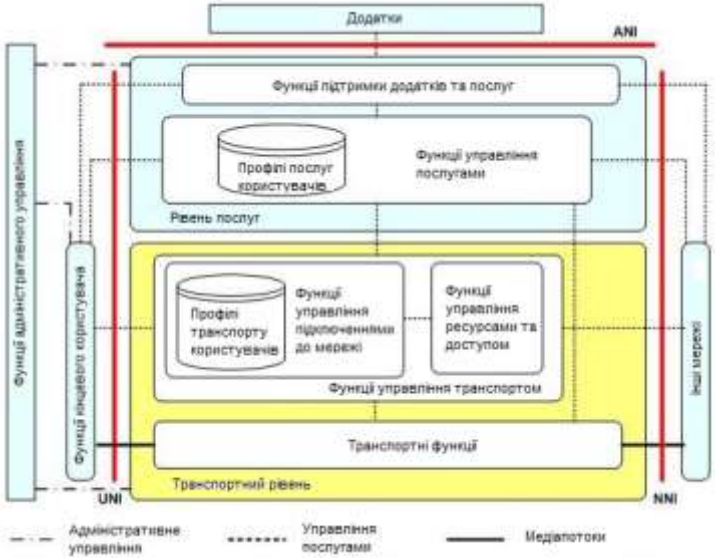


Рис. 1.9 Загальна функціональна архітектура NGN (із рекомендації MCE-T Y.2012)

Функції мережі доступу залежать від використовуваної технології доступу, наприклад, вони розрізняються для бездротової технології CDMA та провідної

технології доступу xDSL. Залежно від технології, яка використовується для доступу до послуг NGN, мережа доступу включає функції, пов'язані з:

- кабельним доступом;
- доступом за технологіями xDSL;
- бездротовим доступом (наприклад, технології IEEE 802.11 (WiFi) і 802.16 (WiMAX) та доступ 3G RAN);
- оптичним доступом.

Прикордонні функції використовуються для обробки трафіку, який виходить шляхом агрегування трафіку, що надходить з різних мереж доступу і передається в магістральну транспортну мережу, вони включають функції, пов'язані з підтримкою якості обслуговування QoS і управління трафіком. Прикордонні функції використовуються також між магістральними транспортними мережами.

Магістральні транспортні функції відповідають за гарантовану передачу інформації через транспортну мережу з різним рівнем якості. Вони забезпечують механізми реалізації заданого рівня якості передачі QoS для користувача трафіку включаючи управління буферами, чергами і розкладом, фільтрацію пакетів, класифікацію, маркування і формування трафіку, контроль дотримання правил обслуговування, управління шлюзами і функції міжмережевих екранів.

Функції шлюзів забезпечують можливості взаємодіяти з функціями кінцевих користувачів і / або іншими мережами, включаючи інші типи мереж NGN та безліч існуючих мереж, таких як ТфЗК / ISDN, публічний Інтернет та інші. Функції шлюзів можуть управлятися або безпосередньо функціями рівня управління або через функції управління транспортною мережею.

Функції обробки медіаінформації забезпечують обробку медіаінформації при наданні послуг, таких як генерація тональних сигналів і перекодування. Ці функції реалізуються спеціальними ресурсами обробки медіаінформації на транспортному рівні.

Функції управління транспортною мережею включають функції управління ресурсами та доступом та функції управління приєднанням до мережі.

Функції управління ресурсами та доступом RACFs (Resource and Admission Control Functions) діють як арбітр між функціями управління послугами і транспортними функціями для підтримки QoS і пов'язані з керуванням транспортними ресурсами в мережі доступу і в магістральної транспортної мережі. Рішення з управління ґрунтується на інформації про необхідному транспорті, угодах про заданий рівень обслуговування SLA, правилах мережної політики, пріоритети усулг та інформації про стан і використання транспортних ресурсів. Функції RACF забезпечують абстрактний підхід до інфраструктури транспортної мережі для функцій управління послугами SCFs (Service Control Functions) і забезпечують сервіс-провайдерам незалежність від мережевої топології, зв'язності, завантаження ресурсів, механізмів/технологій QoS та ін. Функції RACF взаємодіють з функціями SCF і транспортними функціями для різних прикладних програм (наприклад, SIP-виклики, потокове відео й ін.), що вимагає керування транспортними ресурсами NGN, включаючи управління QoS, управління NAPT/firewall і проходження трансляції мережевих адрес на рівні портів NAPT.

Функції управління підключенням до мережі NACFs (Network Attachment Control Functions) забезпечують реєстрацію на рівні доступу та ініціалізацію функцій кінцевого користувача для послуг доступу NGN. Ці функції забезпечують

транспортний рівень ідентифікацією / авторизацією, керуючи простором IP-адрес в мережі доступу і аутентифікації сесій доступу. Вони також повідомляють кінцевим користувачам про контактної точок до функцій NGN на рівні послуг. Функції NACF включають транспортний профіль користувача, який зберігатися у вигляді функціональної бази даних, що включає інформацію користувача, а також інші дані управління.

Рівень послуг включає:

1. Функції управління послугами, включаючи функції профілів послуг користувачів.
2. Функції підтримки прикладних програм і функції підтримки послуг.

Функції управління послугами включають управління ресурсами, функції реєстрації, аутентифікації та авторизації для різних послуг на рівні послуг. Вони також можуть включати функції управління медіаресурсами, такими як спеціалізовані пристрої та шлюзи на сигнальному рівні. Функції управління послугами підтримують профілі послуг користувачів, які являють собою комбінацію користувальницької інформації та інших даних управління, творчу індивідуальний профіль кожного користувача та об'єднані у функціональні бази даних.

Функції підтримки прикладних програм і функції підтримки послуг включають функції шлюзів, реєстрації, аутентифікації та авторизації на прикладному рівні. Ці функції доступні функціональних групах «*прикладні програми*» і «кінцеві користувачі». Вони працюють спільно з функціями управління слугами для забезпечення кінцевих користувачів і прикладних програм необхідними послугами NGN. Через інтерфейс «користувач-мережа» UNI функції підтримки прикладних програм і функції підтримки послуг забезпечують точку доступу до функцій кінцевих користувачів. Взаємодія прикладних програм з даними функціями здійснюється через точку доступу, реалізовану інтерфейсом «прикладна програма-мережа» ANI.

Функції кінцевих користувачів не визначають ніяких обмежень на користувача інтерфейси і мережі кінцевих користувачів, які можуть бути з'єднані з мережею доступу NGN. Термінальні пристрої користувачів послуг NGN можуть бути будь-якими мобільними або стаціонарними пристроями.

Функції адміністративного управління (management functions) забезпечують можливість управляти мережею NGN для надання послуг із заданим рівнем якості, безпеки та надійності. Ці функції розподіляються децентралізовано по всім функціональним блокам (FE) і вони взаємодіють з функціональними блоками управління мережевими елементами, управління мережею і управління послугами. Функції адміністративного управління використовуються на транспортному рівні і рівні послуг і для кожного цього рівня вони реалізують такі завдання:

- управління процесом усунення відмов;
- управління конфігурацією мережі;
- управління розрахунками з користувачами і постачальниками послуг;
- контроль продуктивності мережі;
- забезпечення безпеки роботи мережі.

З метою більш простого розуміння принципів побудови мереж наступного покоління в більшості публікацій з NGN наводиться узагальнена 4-х рівнева архітектура NGN, в якій виділяються такі рівні (рис. 1.10):

- рівень доступу, який містить мережу абонентського доступу до транспортної пакетної мережі;
- транспортний рівень, який включає магістральну пакетну мережу (мережу, побудовану на базі протоколів пакетної комутації IP або ATM, у теперішній час найчастіше на базі технології MPLS та протоколу IP);
- рівень керування комутацією, включає сукупність функцій з керування усіма процесами обслуговування викликами в телекомунікаційній мережі;
- рівень послуг та експлуатаційного керування, який містить логіку виконання послуг та/або прикладних програм та керує цими послугами, має відкриті інтерфейси для використання сторонніми організаціями (для розробки программ і нових послуг).

Термінальне обладнання не входить у склад мережі NGN та в принципі може бути любим з набору абонентського обладнання рівня доступу. Безпосереднє підключення до мережі можливо тільки для пакетних абонентських терміналів, які працюють з використанням протоколів SIP та H.323.

Слід відмітити, зо в деяких публікаціях зустрічаються ще більш проста 3х рівнева архітектура NGN, в якій функції рівня доступу та транспортної мережу об'єднані в один транспортний рівень.

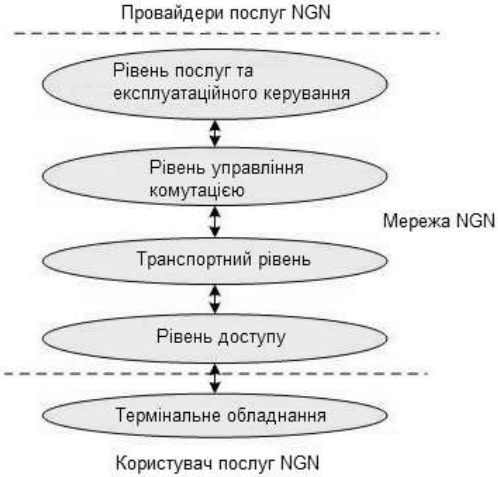


Рис. 1.10 Чотирьохшарова модель NGN

1.6. Принципи взаємодоповнюючого розвитку (конвергенції) мереж рухомого та фіксованого зв'язку майбутнього

На сьогоднішній день спостерігаються таки напрямки розвитку рухомого зв'язку:

- розгортання систем рухомого зв'язку 3G в поєднанні з подальшим розвитком існуючих мереж 2G +;

- розгортання обладнання підсистем платформи IMS, здатних забезпечити надання користувачам широкого спектру інфокомунікаційних послуг на базі технології мультимедіа в загальних послугах NGN мереж наступного покоління;
- розширення ресурсів міжміських мереж зв'язку з метою забезпечення взаємодії опорних мереж 2G+/3G в процесі надання користувачам телекомунікаційних послуг на базі технології комутації пакетів NGN;
- розширення можливостей систем OSS / BSS у відповідності з основними положеннями концепції NGOSS з метою підвищення ефективності експлуатаційного управління мережевими ресурсами та створення бази для управління бізнесом операторів зв'язку у сфері надання інфокомунікаційних послуг;
- розвиток нормативно-правової бази як в частині регламентації вимог до побудови систем рухомого зв'язку, орієнтованих на надання інфокомунікаційних послуг, так і правил надання таких послуг.

У відповідності зі специфікаціями, виробленими в рамках європейського інституту стандартів (ETSI) і в 3GPP на мережеву підсистему GPRS системи GSM ph2 +, використання технології пакетної передачі дозволяє надати користувачам доступ в Інтернет, забезпечуючи передачу на швидкості від 14 до 115 кбіт / с і вище. Система рухомого зв'язку 3G/UMTS в порівнянні з системою 2G/GSM ph2 + дозволяє додатково розширити спектр послуг за рахунок збільшення швидкості передачі даних (до 384 кбіт/с при пересуванні зі швидкістю до 120 км/год і до 2048 кбіт/с при пересуванні зі швидкістю до 10 км/год). Так, наприклад, тільки в системах 3G можлива підтримка відеоконференції в реальному часі. Крім того, в системах 3G використовуються механізми для оптимізації навантаження – управління якістю наданих послуг QoS. У мережах рухомого зв'язку припинилося прирощення зростання доходів (у ряді випадків в операторів фіксованої мережі спостерігається тенденція припинення росту доходів), як наслідок, оператори зв'язку зацікавлені в нових джерелах доходу. На сьогодні ці джерела, значною мірою, пов'язані з широкосмуговими послугами. Інтенсивний розвиток IP телебачення, а також пропозиції користувачам послуг демонстрації один одному відеопотоків в реальному часі, зумовлюють попит на засоби широкосмугового доступу до мережевих ресурсів, а також до потенційних постачальників послуг NGN. Спостерігається зсув доходів операторів рухомого зв'язку у бік надання послуг передачі даних і послуг з доданою вартістю (VAS), зокрема, мобільного контенту. Тенденція падіння прибутку GPRS-WAP і плавного зростання прибутку від послуг GPRS-Internet збережеться в наступних періодах 2016-2018 р. Однак зростання прибутку від GPRS-Internet не здатний компенсувати втрату доходів від GPRS-WAP. На даному етапі розвитку засобів телекомунікації VAS-напрям є частиною загальної маркетингової стратегії операторів, одним із способів підвищення ARPU. На сучасному рівні розвитку інфокомунікаційних технологій генератором послуг зв'язку все більше стає користувач, у той час як постачальник обладнання втрачає цю функцію. При цьому зусиллями постачальників відповідного обладнання та прикладних програм буде створена інфокомунікаційних інфраструктура, в якій оператори зв'язку перестануть відігравати провідну роль. Спостерігається поступова зміна бізнес-моделі надання послуг рухомого зв'язку, а саме, перехід від традиційної бізнес-моделі «оператор-абонент-споживач» до бізнес-моделі надання інфокомунікаційних послуг в мережах NGN «розробник (власник) контенту-постачальник-агрегатор-Дистриб'юторських оператор сервісів-оператор доступу

користувач». Послуги мереж рухомого зв'язку 3G (розгорнуті в багатьох країнах) виявляються десяткам мільйонів абонентів, проте ефективність цього бізнесу регулярно відстає від бажаного рівня. Найбільший попит на послуги високошвидкісної рухомого зв'язку очікується у великих містах з населенням більше 1 млн. Чоловік. Розгортання мереж 3G відкриє нові можливості для Location Based Service - LBS, «важкого» контенту, мобільного банкінгу та маркетингу. У міру розгортання мереж 3G поняття «Додаткові Послуги Зв'язки VAS» знайде новий сенс (причому, діаметрально протилежний існуючого). У мережах 3G послуга передачі мови (радіотелефонного зв'язку) стане додатковою, тоді як послуга передачі даних придбає характер основною. Зокрема, телефонний зв'язок може стати безкоштовним додатком до інших сервісів для абонентів мереж рухомого зв'язку 3G / 4G. Операторам мереж 3G варто врахувати світовий досвід експлуатації подібних мереж разом з мережами GSM. Оптимальною є стратегія використання єдиних тарифних планів при обслуговуванні в мережі будь-якого стандарту (GSM або WCDMA). При цьому абонент автоматично почне використовувати мережу WCDMA саме для передачі даних (за рахунок високих швидкостей), а мережа GSM для телефонного зв'язку (за рахунок більшої території покриття).

Відповідно до основних положень концепції NGN мережі рухомого та фіксованого зв'язку вбудовуються в загальну інфраструктуру, що задовольняє функціональної архітектурі NGN. При цьому існуючі мережі рухомого та фіксованого зв'язку розглядаються як мережі доступу до послуг NGN, а відмінність цих мереж один від одного полягатиме тільки в технологіях абонентського доступу.

На рівні опорних мереж фіксованого та рухомого зв'язку вже активно впроваджуються конвергентні технології, зокрема, планомірно здійснюється перехід на технологію комутації пакетів на базі протоколу IP при наданні будь-яких телекомунікаційних та інфокомунікаційних послуг. На міжміському рівні поставлена і вирішується завдання створення таких мультисервісних магістральних мереж також на базі технології комутації пакетів і протоколу IP, які будуть здатні підтримати всі процедури надання послуг NGN широкому колу користувачів.

Однією з основних проблем на шляху переходу до мереж наступних поколінь є розгортання підсистем IMS, які здатні будуть управляти процедурами надання телекомунікаційних та інфокомунікаційних послуг на базі технології мультимедіа. Видається, що успішне вирішення цієї проблеми з'явиться вирішальним внеском на шляху взаємодоповнююче розвитку мереж рухомого та фіксованого зв'язку.

Разом з тим слід зазначити, що одне з основних положень концепції NGN, а саме, відділення функцій управління викликами і сесіями від функцій транспорту не була своєчасно підтримана відповідним набором стандартів і специфікацій.

Це призвело до того, що основні мережеві елементи, здатні реалізувати функціональні елементи архітектури NGN, що поставляються різними виробниками, виявилися несумісними між собою як по інтерфейсах, так і за своїми функціями. Саме це визначило активність міжнародних організацій (в першу чергу ETSI і 3GPP), які почали розробку нових принципів побудови і стандартів мереж рухомого зв'язку 3G, грунтуючись на рівневій архітектурі NGN.

Спочатку IMS була розроблена в 3GPP як функціональна архітектура управління послугами в мережах зв'язку наступних поколінь. Ця архітектура була розроблена з метою забезпечити можливість для операторів зв'язку надавати

користувачам широкий діапазон мультимедійних послуг на базі технології комутації пакетів і надавати їх у реальному часі, відстежувати процедури надання послуг як шляхом традиційного збору інформації про тривалість часу надання послуги, так і збору інформації про профіль наданої послуги і кількості переданих пакетів.

Підсистема IMS має структуру, орієнтовану на введення і надання як основних, так і розширених послуг зв'язку, зокрема: Voice Call Continuity (VCC); передача повідомлень на базі технології мультимедіа; web інтеграція (chat text, shared online whiteboards та ін.); Push to talk over Cellular (PoC). Оператори зв'язку очікують, що запровадження IMS скоротить CapEx і OpEx допомогою використання конвергентної магістральної IP мережі та відкритої архітектури IMS:

- архітектура IMS визначає безліч спільних компонент (наприклад, для управління викликом і зберігання конфігурації профілю послуг), як наслідок, буде потрібно менше зусиль для створення нових послуг, оскільки платформа IMS може багаторазово використовуватися для цих цілей;
- використання стандартних інтерфейсів має збільшити конкуренцію між постачальниками обладнання, що зніме небезпеку для операторів зв'язку потрапити в залежність від обладнання конкретного постачальника, яке підтримує лише власні закриті інтерфейси.

Як наслідок, розгортання обладнання платформи IMS повинно забезпечити більш швидке і дешеве введення нових послуг зв'язку в порівнянні з традиційними монолітними конструкціями телефонних послуг.

Технології NGN відкривають перед операторами зв'язку нові способи ефективної передачі трафіку, підключення та обслуговування клієнтів (користувачів, постачальників послуг, постачальників контенту та ін.), Взаємодії постачальників сервісу. Формування інфокомунікаційної кооперації, що володіє величезним потенціалом, обумовлює завершення етапу розвитку телекомунікацій, при якому оператор зв'язку відіграє ключову роль (будує мережу, наповнює послугами та обслуговує користувачів тільки оператор зв'язку). Процес конвергенції (зближення) мережевих технологій зачіпає як фіксовані, так і рухливі мережі зв'язку, зокрема:

- при наданні послуг фіксованого та рухомого зв'язку в перспективі будуть використовуватися опорні й транзитні мережі, побудовані на базі єдиної технології, що базується на протоколі IP;
- у мережах фіксованого зв'язку все ширше застосовуються засоби і технології бездротового широкосмугового доступу (зокрема, WiFi, WiMAX), що поступово перетворює їх в мережі рухомого зв'язку;
- мережі рухомого зв'язку модернізуються (згідно концепції NGN) шляхом переходу на IP технологію в процесі всередині мережевої конвергенції за підтримки послуг і прикладних програм, що залучаються при обслуговуванні клієнтської бази;
- введення багатомодових (наприклад, GSM / UMTS / WiFi) терміналів користувача, забезпечення безшовного хендовера, а також конвергенція послуг, підтримуваних у фіксованих та рухомих мережах, стирає межі між відповідними системами зв'язку в процесі міжмережевої конвергенції (FMC).

Основні положення концепції NGN знаходять своє практичне відображення в міжнародних стандартах і специфікаціях, які розробляються на обладнання підсистеми IMS, що орієнтується на підтримку безлічі різного роду інфокомунікаційних мультимедійних послуг на базі протоколу IPv6. Процес

міжмережевої конвергенції в поєднанні з розгортанням платформ IMS обумовлює тісну взаємодію (роботу в кооперації) різних сторін, представлених в бізнес-моделі надання інфокомунікаційних мультимедійних послуг. Процес конвергенції фіксованого та рухомого зв'язку обумовлює розробку і впровадження нових принципів білінгу, що надають користувачеві можливості вибору в реальному часі оптимального тарифу в будь-якому місці для будь-якої послуги, що в свою чергу вплине на конкуренцію між постачальниками сервісу. В процесі впровадження технологій NGN неминуче збільшиться кількість сервісів на базі єдиної транспортної інфраструктури, що надасть можливість користувачам не думати про мережеві особливостях, а зосередитися на виборі між сервісами з різними характеристиками якості та відповідними тарифами.

Основні інновації в мережевих підсистемах рухомого зв'язку на етапі просуванні їх до складу мереж NGN пов'язані з переходом на IP технологію в CS-сегментах опорних мереж, а також зі створенням мультисервісних міжміських (транзитних) мереж, здатних підтримати взаємодія опорних мереж (CS- і PS-сегментів) при наданні користувачам послуг NGN. При перекладі на IP технологію CS -сегменті опорних мереж виникає ряд проблем, пов'язаних з розміщенням конвергентного обладнання по території країни і відповідності різних варіантів розміщення існуючій нормативній базі. При вирішенні проблем, пов'язаних із взаємодією конвергентних опорних мереж окремих операторів зв'язку, виникає ряд завдань, що відносяться до стратегії і тактиці модернізації існуючих міжміських мереж.

Вважається, що в міру розширення і ускладнення ринку рухомого зв'язку ростуть запити користувачів на послуги рухомого зв'язку. Очікується, що розгалужені і комплексні послуги зажадають зовсім інших характеристик трафіку і рівнів якості послуг порівняно з трафіком мови або тексту систем 3G. Майбутні технології рухомого зв'язку повинні надавати розгалужені послуги з різноманітними параметрами трафіку, забезпечуваними незалежно від технології радіоінтерфейсу. В даний час встановлено, що найефективнішим рішенням для радіоінтерфейсу, що забезпечує мобільному користувачеві доступ до послуг, порівнянним з послугами, наданими провідними мережами, є повний перехід на технологію IP як універсального методу передачі всіх видів трафіку і прикладних програм. Тільки перехід на технологію IP всіх елементів систем рухомого зв'язку дозволяє організувати в майбутньому їх глобальне взаємодія.

Тим часом саме мережі радіодоступу в силу фізичних обмежень радіоканалів є основним чинником, що визначає можливості забезпечення мобільного радіодоступу до послуг NGN, так як опорні мережі не мають принципових обмежень ні на швидкості передачі даних, ні на можливості доступу користувачів до будь-яких видів послуг. Більш того, вже існуючі IP мережі в принципі готові до перекладу на них всіх видів трафіку – пакетного мовного і даних, і вирішальною умовою успішної реалізації NGN є вироблення рішень, що визначають організацію і використання мереж радіодоступу.

На основі досвіду переходу від 2G (GSM) до 3G (UMTS) можна очікувати, що майбутні системи радіодоступу дозволять реалізувати істотно більш високі швидкості користувача трафіку, ніж є в даний час. Однак при цьому слід мати на увазі, що для швидко рухається користувача ці швидкості завжди будуть менше швидкостей, доступних для нерухомого користувача. У зв'язку з цим майбутні

мережі радіодоступу повинні бути оптимізовані стосовно до різних конкретних категорій користувачів (за ступенем рухливості, швидкості передачі, якості послуг QoS та ін.).

Питання побудови мереж мобільного доступу в мережах наступного за 3G покоління систем рухомого зв'язку (NGN) розглядаються у двох міжнародних організаціях – МСЕ-Р і 3GPP поки на рівні постановки задачі та визначення основних вимог як до опорних мереж, так і до мереж радіодоступу. Зростаюча конкуренція на ринку послуг рухомого зв'язку ще більш загостриться при розгортанні систем 3G і впровадженні у фіксованих мережах ресурсів радіодоступу 4G/5G. Неминуче розширення бізнес-моделей, пов'язаних з наданням послуг рухомого зв'язку, зумовить появу відносно вільних ніш для альтернативних учасників.

Однією з таких ніш стає ніша віртуальних операторів рухомого зв'язку (MVNO). Оператори рухомих мереж 2G+/3G, маючи надлишок мережевих ресурсів, можуть надавати його частина операторам MVNO за схемами аутсорсингу та франчайзингу. Це зніме протиріччя між наявністю в операторів рухомих мереж надлишку мережевих ресурсів і недостатню зацікавленість у веденні власного бізнесу і, зокрема, у сфері надання послуг IMS.

Оператор транзитної мультисервісної мережі крім виконання функцій з пропуску трафіку, може стати також і оператором MVNO. У цьому випадку такий оператор зв'язку здатний надавати мультимедійні широкосмугові послуги абонентам різнорідних мереж на великій території, використовуючи ресурси радіодоступу існуючих і знову з'являються операторів фіксованих і рухомих мереж. Залежно від оснащеності віртуального оператора мережевими ресурсами доцільно розрізняти три рівня проникнення в сферу надання послуг рухомого зв'язку, а саме:

- управління бізнес-процесом (орендується частина ресурсів всієї мережевої інфраструктури базового оператора);
- управління бізнес-процесом і управління послугами (орендується частина ресурсів опорної мережі та мережі радіодоступу базового оператора);
- управління бізнес-процесом, послугами та опорною мережею (орендується частина ресурсів мережі радіодоступу базового оператора).

Застосування технології MVNO дозволяє забезпечити:

- зниження ризиків для традиційних операторів зв'язку на шляху просування в напрямку надання послуг IMS за рахунок розширення відповідних бізнес-моделей;
- залучення на ринок надання послуг IMS додаткових учасників, що підвищить конкуренцію і, як наслідок, призведе до зниження тарифів на послуги рухомого зв'язку.

1.7. 5G бездротові системи зв'язку

Поява концепції NGN та її підсистем відобразилась на принципах розвитку структури мереж стільникового рухомого зв’язку 5G.

Основна відмінність, з точки зору користувача між нинішнім поколінням і очікуваною системою 5G має бути не тільки в збільшенні максимальної пропускної спроможності; інші вимоги включають в себе:

- більш низьке споживання енергії;

- кілька одночасних шляхів передачі даних;
- нижче ймовірність відмови; краще покриття і високі швидкості передачі даних, доступних на границі стільника;
- 1Гбіт/с (і вище) швидкість передачі даних при мобільності;
- більш висока безпека; краща ступінь когнітивності/радіо, що визначається програмно (Software-Defined Radio – SDR);
- більш висока спектральна ефективність системного рівня;
- не завдає шкоди здоров'ю людини;
- переносні пристрої з можливостями штучного інтелекту (AI);
- дешевший трафік в зв'язку з низькими витратами на розгортання інфраструктури;
- системи розумних променевих антен;
- світова бездротова мережа (World Wide Web Wireless – WWWW), бездротові веб-додатки, які включають повні мультимедійні можливості за межами швидкостей 4G;
- покращене та інноваційне кодування даних і методи модуляції, що включають в себе набір фільтрів з декількома несучими в схемах.

У табл. 1.1 наведено порівняння між системами бездротового зв'язку 3G, 4G, і 5G [111].

Табл. 1.1

Порівняння між 3G, 4G, 5G системами бездротового зв'язку

Технологія/Характеристики	**3G**	**4G**	**5G**
Визначення	Цифрова широкосмугова мережа, пакетні дані	Цифрова широкосмугова мережа, пакетні дані, все IP	Цифрова широкосмугова мережа, пакетні дані, все IP, дуже висока пропускна здатність
Швидкість передачі даних	2 Мбіт/с	2 Мбіт/с – 1 Гбіт/с	1 Гбіт/с і вище (за потребою)
Стандарти	WCDMA, CDMA2000, TD-SCDMA	Конвергенція всіх видів доступу, включаючи: OFDMA, MC-CDMA Network-LMPS	CDMA та BDMA
Технологія	Широка смуга пропускання, CDMA, IP технологія	Уніфіковане IP та неперервна комбінація широкосмугового LAN/WAN/PAN та WLAN	Уніфіковане IP та неперервна комбінація широкосмугового, LAN/WAN/PAN/WLAN та wwww
Сервіси	Інтегроване високоякісне аудіо, відео та дані	Динамічний доступ до інформації, носимі пристрої, HD стримінг, глобальний роумінг	Динамічний доступ до інформації, носимі пристрої, HD стримінг; будь-які вимоги користувачів; усі майбутні

			технології, м'який глобальний роумінг
Множинний доступ	CDMA	CDMA	CDMA та BDMA
Опорна мережа	Пакетна мережа	Все IP мережа	Більш плоска IP мережа та інтерфейси з 5G мережею (5G-NI)
Хендовер	Горизонтальний	Горизонтальний та вертикальний	Горизонтальний та вертикальний

На рис. 1.11 представлений огляд задач, систем забезпечення, а також відповідних принципів побудови 5G [112].

Добре відомо, що користувачі, які використовують бездротові системи зв'язку, перебувають у приміщенні близько 80% свого часу, в той час як перебування за межами приміщень (на свіжому повітрі) складає близько 20% часу. В даний час традиційна стільникова архітектура зазвичай використовує базову станцію поза приміщенням в середині стільника, що взаємодіє з мобільними користувачами, незалежно від того, залишаються вони в приміщенні чи ні. Для користувачів всередині приміщення, які взаємодіють з базовою станцією поза приміщенням, сигнали повинні проходити через стіни будівлі, і це призводить до дуже високих втрат на проходження, що істотно знижує швидкість передачі даних, спектральну ефективність і ефективність використання енергії бездротової передачі даних.

Основною ідеєю проектування стільникової архітектури 5G є поділ зовнішніх і внутрішніх сценаріїв, так що втрат на проникнення сигналу через стіни будівель якимось чином можна уникнути. Це забезпечать розподілена антенна система (Distributed Antenna System – DAS) і масивна технологія MIMO, де розгорнуті географічно розподілені антенні решітки з десятками або сотнями антенними елементами. У той час як більшість сьогоденних систем MIMO використовують від двох до чотирьох антен, мета масивних систем MIMO полягає в використанні потенційно великого виграшу ємності, що буде мати місце в більшій кількості антен. Зовнішні базові станції будуть оснащені великою кількістю антенних решіток з деякими елементами антен (також великих антенних решіток), розподілених по стільнику і підключених до базової станції через оптичні волокна, отримуючи вигоду з обох DAS і масивних технологій MIMO. Зовнішні мобільні користувачі зазвичай обладнані обмеженою кількістю антенних елементів, але можуть об'єднуватись один з одним формуючи віртуальну велику антенну решітку, що разом з антенними решітками базової станції будуватиме віртуальні масивні MIMO канали. Великі масиви антен будуть також встановлюватися зовні кожної будівлі, щоб взаємодіяти з зовнішніми базовими станціями для розподілених антенних елементів базових станцій, можливо, з компонентами прямої видимості (line of sight – LoS). Великі масиви антен мають кабель, підключений до бездротових точок доступу всередині будівлі, що взаємодіють з внутрішніми користувачами. Це, безумовно, збільшить вартість інфраструктури в короткостроковій перспективі в той час як, при цьому значно покращить середню пропускну здатність стільника, спектральну ефективність, ефективність використання енергії та швидкість передачі даних стільникової системи в довгостроковій перспективі.

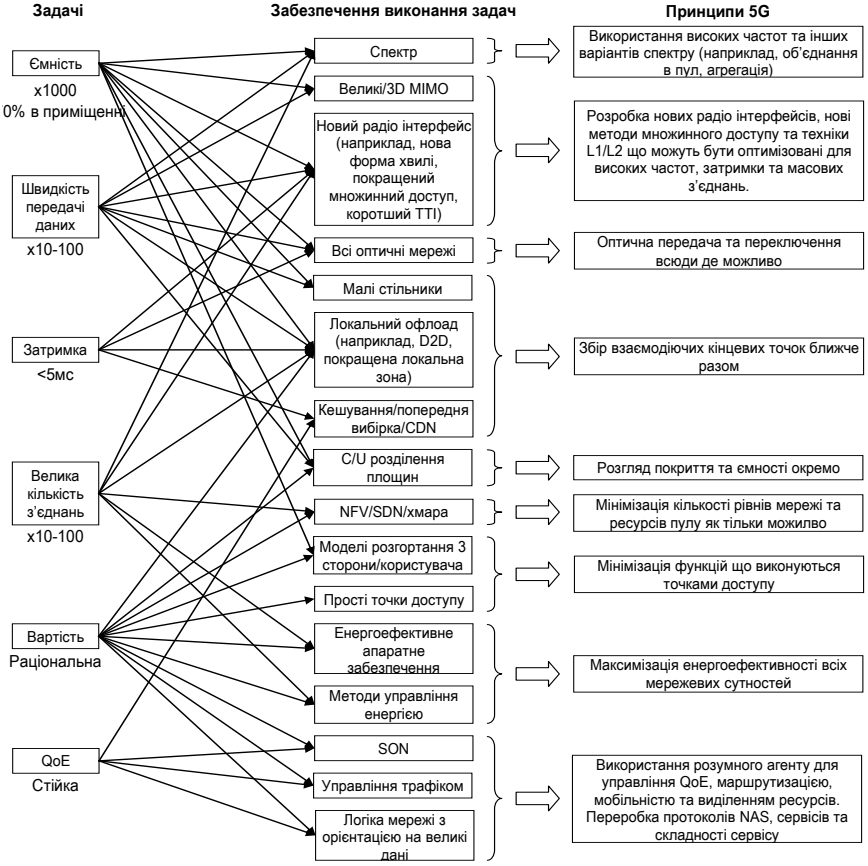


Рис. 1.11 Задачі, потенційне забезпечення та принципи побудови 5G [112]

За допомогою такої стільникової архітектури, оскільки в приміщенні користувачам необхідно тільки зв'язуватися з бездротовими точками доступу в приміщенні(не зовнішні базові станції) з великими антенними решітками, встановленими поза будівлями, багато технологій можуть бути використані, які підходять для зв'язку малого радіусу з високими швидкостями передачі даних. Деякі приклади включають в себе Wi-Fi, femtocell, надширокі смуги (UWB), міліметровий діапазон частот (3–300 ГГц), і зв'язок на основі видимого світла (VLC) (400-490 ТГц). Варто відзначити, що міліметрові хвилі і технології VLC використовують більш високі частоти, які зазвичай не використовуються для стільникового зв'язку. Ці високочастотні хвилі не проникають через тверді матеріали дуже гарно і можуть бути легко поглинені або розсіяні газами, дощем і листям. Тому важко використовувати ці хвилі для зовнішніх і далеких відстаней.

Стільникова архітектура 5G також повинна бути гетерогенною, з макростільниками, малими стільниками і релейною передачею. Для користувачів з

високою мобільністю, таких як користувачів у транспортних засобах і високошвидкісних поїздах, були запропоновані мобільні фемтостільники (MFemtocell), що поєднує концепції мобільної передачі і фемтостільник. MFemtocells розташовуються всередині транспортних засобів, щоб забезпечити зв'язок з користувачами всередині транспортного засобу, в той час як великі антенні решітки розташовані поза транспортним засобом для зв'язку з зовнішніми базовими станціями. MFemtocell і асоційовані з ним користувачі розглядаються як єдине ціле для базової станції. З точки зору користувача, MFemtocell розглядається як звичайна базова станція. Це дуже схоже на ідею поділу сценаріїв в приміщенні (всередині транспортного засобу), а також поза ним. Користувачі, які використовують MFemtocells можуть користуватися послугами з високою швидкістю передачі даних зі зменшиними сигнальним трафіком. Рис. 1.12 показує гетерогенную бездротову стільникову архітектуру 5G [111].

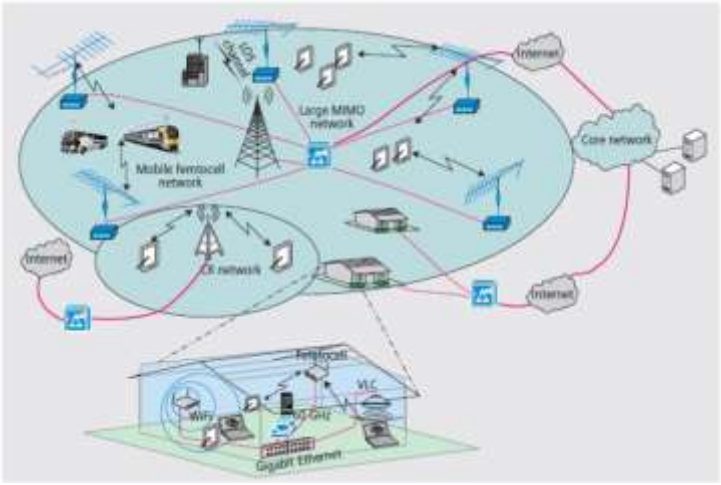


Рис. 1.12 Безпроводова стільникова гетерогенна архітктура 5G [111]

Рівень прикладних програм	Прикладна програма (Сервіси)
Рівень представлення	
Сеансовий рівень	Open Transport Protocol (OTP)
Транспортний рівень	
Мережевий рівень	Верхній мережевий рівень
	Нижній мережевий рівень
Рівень каналу передачі даних (MAC)	Open Wireless Architecture (OWA)
Фізичний рівень	

Рис. 1.13 Стек протоколів 5G

Фізичний рівень і рівень управління доступом до середовища (MAC) тобто OSI рівень 1 і OSI рівень 2 визначають технологію бездротового зв'язку і показано на рис. 1.13. Для цих двох рівнів мобільні мережі 5G, ймовірно, будуть спроектовані на основі відкритої бездротової архітектури.

Мережевий рівень буде IP (Internet Protocol), тому що сьогодні на цьому рівні немає нічого іншого. IPv4 (версія 4) поширений по всьому світу, і у нього є кілька проблем, таких як обмежений адресний простір і відсутність реальної можливості для підтримки QoS для кожного потоку.

Ці проблеми вирішуються в IPv6, але в обмін на значно більший заголовок пакету. Також мобільність як і раніше залишається проблемою. Існує стандарт Mobile IP, з одного боку, а також безліч рішень мікромобільності (наприклад, Cellular IP, HAWAII і т.д.). Всі мобільні мережі будуть використовувати Mobile IP в 5G, і кожен мобільний термінал буде FA (Foreign Agent), зберігаючий CoA (Care of Address) відображаючи між його фіксованим адресою IPv6 і CoA-адресою для поточної бездротової мережі. Проте, мобільний пристрій може бути приєднано до декількох мобільних або бездротових мереж одночасно. У такому випадку, він буде підтримувати різні IP-адреси для кожного з радіоінтерфейсів, в той час як кожен з цих IP-адрес буде CoA-адресою для FA, розміщеного в мобільному телефоні. Фіксований IPv6 буде реалізований в мобільному телефоні виробниками 5G пристроїв. Мобільний телефон 5G має підтримувати віртуальне мультибезпроводове мережеве середовище. Для цього має бути розподіл мережевого рівня на два підрівні в 5G пристроях, тобто: Нижній мережевий рівень (для кожного інтерфейсу) і Верхній мережевий рівень (для мобільного терміналу). Це пов'язано з початковою структурою Інтернету, де вся маршрутизація базується на IP-адресах, які повинні бути різними в кожній IP-мережі по всьому світу. Проміжне програмне забезпечення між верхнім і нижнім мережевими рівнями (рис. 1.13) повинне підтримувати перетворення адрес з верхнього мережевого адреси (IPv6) до різних IP-адрес нижньої мережі (IPv4 або IPv6) і навпаки. На рис. 1.14 показано мережевий рівень 5G [113].

Мобільні та бездротові мережі відрізняються від провідних мереж що стосується транспортного рівня. У всіх версіях TCP передбачається, що сегменти втрачаються через перевантаження мережі, в той час як в бездротових мережах втрати можуть виникати через більш високий коефіцієнта бітових помилок в радіоінтерфейсі. Таким чином, модифікації TCP пропонуються для мобільних і бездротових мереж, які заново передають втрачені або пошкоджені сегменти TCP тільки по бездротовій лінії зв'язку. Для мобільних терміналів 5G підійде наявність транспортного рівня, який можна завантажити і встановити. Такі мобільні пристрої повинні мати можливість завантажувати (наприклад, TCP, RTP і т.д. або новий транспортний протокол) версію, яка орієнтована на конкретну технологію бездротового зв'язку, встановлену на базових станціях. Це називається OTP (Open Transport Protocol).

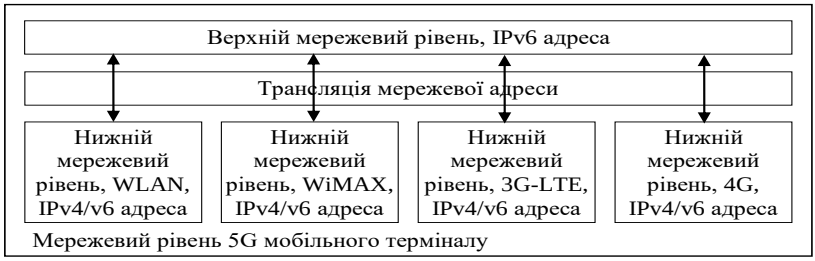


Рис. 1.14 Мережевий рівень 5G мобільного терміналу

Що стосується прикладних програм, вимогою мобільних терміналів 5G є забезпечення інтелектуального управління QoS по безлічі мереж. Сьогодні в мобільних телефонах користувачі вручну вибирають бездротовий інтерфейс для конкретної Інтернет-послуги, не маючи можливості використовувати історію QoS для вибору найкращого бездротового з'єднання для даної послуги. Пристрій 5G має забезпечити можливість тестування якості обслуговування і зберігання вимірювальної інформації в інформаційних базах даних в мобільному терміналі. Параметри QoS, такі як затримка, джитер, втрати, пропускна здатність, надійність, будуть зберігатися в базі даних в пристрої 5G з метою використання інтелектуальних алгоритмів, що працюють в мобільному терміналі як системні процеси, що в кінцевому рахунку має забезпечити найкраще бездротове з'єднання при необхідному QoS і персональних обмеженнях вартості. З 5G буде доступний цілий ряд нових послуг і моделей. Ці послуги та моделі повинні бути додатково перевірені в їх взаємозв'язку з будовою систем 5G. Процес вичерпування адрес IPv4, як очікується, буде на завершальній стадії на той час, як розгорнеться 5G. Таким чином, підтримка IPv6 для 5G має важливе значення для того, щоб підтримувати велику кількість бездротових пристроїв. IPv6 усуває необхідність в NAT (Network Address Translation) за рахунок збільшення кількості IP-адрес. За допомогою доступного адресного простору і числа біт адресації в IPv6, багато інноваційних схем кодування можуть бути розроблені для 5G пристроїв і прикладних програм, які можуть бути корисні в розгортанні мережі та послуг 5G. Нове покоління обіцяє виконати мета персонального комп'ютингу та комунікацій – бачення, яке за доступною ціною забезпечує високу швидкість передачі даних в будь-якому місці по бездротовій мережі. У майбутніх бездротових мережах повинна бути низька складність реалізації і ефективні засоби переговорів між кінцевими користувачами і бездротовою інфраструктурою. Інтернет є рушійною силою для більш високих швидкостей передачі даних і високошвидкісного доступу для мобільних користувачів бездротової мережі. Це буде мотивацією для еволюції опорної мережі на основі мобільного все IP [113].

Висновки

Сьогодні комунікаційні системи являються галуззю, що швидко розвивається. У розділі досліджено особливості архітектурних рішень сучасних телекомунікаційних систем, систем керування мережама зв'язку як розподілених інформаційно обчислювальних систем. Досліджено особливості архітектури мереж наступного покоління NGN, яка передбачає розділення ситеми організації зв'язку на рівень транспорту та рівень послуг. Ефективність сучасних та перспективних комунікаційних систем визначається конкретною реалізацією їх системи управління. Описана структура системи управління комунікаційної системи. Однією з найважливіших основних підсистем системи управління комунікаційної системи являється система динамічного управління комунікаційної системи. Основна задача цієї системи – забезпечення якості обслуговування абонентів. Дано характеристику основних параметрів якості обслуговування. Описані характеристики служби якості обслуговування. Досягнення заданих параметрів якості обслуговування абонентів комунікаційної системи досягається за рахунок ефективного використання засобів управління трафіком та мережевими ресурсами. Проаналізовані основні функції

NGN, та окреслено поняття ресурсів у мережах нового покоління. Концепція мереж нового покоління лягла в основу сучасних технологій забезпечення рухомого зв'язку, що широко впроваджуються на сьогоднішній день, таких як IMS, або організація управління яких знаходиться на етапі розробки, зокрема 5G, SDN, NFV.

2. ОБСЛУГОВУВАННЯ ІНФОРМАЦІЙНИХ ПОТОКІВ У ТРАНСПОРТНИХ МЕРЖАХ ЗВ'ЯЗКУ

Обслуговування інформаційного потоку – це багатоскладовий процес. В цьому розділі описана математична модель процесу обслуговування інформаційних потоків у транспортних мережах зв'язку. Розглянуто метод підвищення ефективності обслуговування інформаційних потоків в комутаційному центрі транспортної мережі, удосконалення механізму зваженого кругового обслуговування черг та принцип керування інформаційними потоками в комутаційному центрі PWE3.

2.1. Математична модель процесу обслуговування інформаційних потоків у транспортних мережах зв'язку

Система керування пакетами інформації у вузлах комутаційних центрів має враховувати особливості трафіку. В цьому підрозділі висвітлюються характеристики інформаційного трафіку, які впливають на показники якості передачі інформації.

2.1.1. Передача інтерактивного відео

Відеоінформація складається з послідовності нерухомих зображень (кадрів), які відтворюються з частотою 25-30 кадр/с. При передачі кожний кадр інтерпретується як об'єднання деякої кількості фіксованих точок (пік- селів) з певною яскравістю та кольором. Конкретна швидкість передачі залежить від кількості пікселів у кадрі, частоти кадрів (кількості кадрів в секунду). Кількість інформації для передавання кожного пікселю (біт на піксель), в залежності від потрібної якості зображення, може змінюватися у широких межах: від одиниць до сотень мегабіт на секунду.

Наприклад, для виробничих умов, де зображення повинні передаватися в реальному часі з постійною швидкістю, для досягнення високої якості відеосигналу, може знадобитися швидкість передачі до 1,8Гбіт/с. Вказана пропускна здатність необхідна для цифрового кіно.

Необхідно відмітити, що відео інформація володіє досить великою надлишковістю і при її передачі можуть застосовуватися різні методи ущільнення. В цьому випадку передача сигналу здійснюється зі змінною швидкістю, тобто об'єм даних для передавання кожного кадру різний [13]. Вибір стандарту ущільнення (наприклад, MPEG-2, M-JPEG, H.261, ACT-L3, RMS) визначає ступінь стиснення відеоінформації і відповідно якість сигналу, що передається, а також необхідну смугу пропускання, наприклад, 4-5Мбіт/с (MPEG-2) або 64-8000кбіт/с у випадку застосування ACT-L3.

Відео інформація висуває досить високі вимоги до мережі зв'язку: якість каналу та його пропускна здатність повинні бути достатньо високими [13]. Наприклад, мережа ATM, яка транспортує потік MPEG-2, повинна гарантувати долю втрачених комірок (Cell Loss Ratio – CLR) меншу ніж $1{,}7*10^{-9}$, значення постійної складової затримки комірок (Cell Transfer Delay – CTD) рівне 4мс (не

більше 150 мкс на комутатор) і флуктуацію змінної складової мережевої затримки (Cell Delay Variation – CDV) не більше 500мкс на з‘єднання типу «точка-точка».

Величина CTD може змінюватися в широкому діапазоні [13]. У той час коли низько швидкісна 64 кбіт/с відео конференція може допускати величину транзитної затримки CTD=300 мкс, високошвидкісна відеоконференція 1,5 Мбит/с потребує гарантії CTD=5 мкс, а для відео HDTV повинна гарантуватися величина CTD=1 мкс.

Для побудови математичної моделі трафіку необхідно мати чітке уявлення про характер тарафіку. Отже, можна виділити наступні ознаки відео трафіку:

1. Характер надходження заявок на передачу відеоповідомлень аналогічна до характеру надходження телефонних заявок та описується найпростішим потоком.

2. Тривалість потоку пакетів одного повідомлення є значною, наприклад, 1,5 години один кінофільм або 20-30 хвилин тривалість відеоконференції.

3. Потік пакетів одного повідомлення, з метою забезпечення найвищої якості передачі, повинен бути неперервним [16].

4. Довжина пакетів одного потоку на виході з кодеку є змінною, залежно від картинки, що передається з кадром. Диперсія довжини пакетів може бути значною, тому часто у сучасних кодеках застосовують процедуру шейпінгу, яка дозволяє зменшити дисперсію відео пакетів, але при цьому може створювати деяку затримку в передачі кадрів на вихід кодеку.

У переважній більшості робіт по теорії масового обслуговування розглядається найпростіший випадок потоків, коли ймовірність надходження за проміжок часу t рівно k вимог задається формулою

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \qquad (2.1)$$

де $\lambda>0$ – постійне число, інтенсивність надходження заявок. Потік, що надходить при цьому вважається таким, що для будь-якої скінченої групи відрізків часу, які не перетинаються, заявки, що з‘явилися протягом них являють представляють собою взаємно незалежні випадкові величини.

Класичним прикладом пуасонівського потоку є потік заявок на телефонну станцію. Потік Пуасона не лише підтверджується статистично при спостереженні багатьох реальних потоків, але й з‘являється в якості граничного об‘єкту у багатьох ймовірнісних схемах [1, 2].

2.1.2. Передача Інтернет трафіку в режимі off-line

Передача Інтернет трафіку характеризується значно м‘якшими вимогами до передачі порівняно з телефонним трафіком або трафіком інтерактивного відео. Основні особливості передачі даних по мережі полягають у наступному:

1.Потрібна висока достовірність передачі, не допускаються вставки або випадання окремих порцій інформації. Необхідно застосування надійних способів виявлення помилок та повторної передачі відповідних блоків даних.

2.Відсутні жорсткі вимоги до величини постійної затримки інформації в мережі та до її дисперсії, хоча для деяких інтерактивних програм можуть існувати обмеження на транзитну затримку, які визначаються вимогами часу відгуку.

3.Припускається довільний та незалежний темп передачі та прийому даних в мережі.

4.Потрібна організація багаторежимного обміну даними (діалогова передача, передача файлів, тощо) та розгорнута система пріоритетів.

5.Канали зв‘язку використовуються, як правило, високої якості з ймовірністю помилки нижче ніж 10^{-4}.

6.Вимоги до ширини смуги пропускання лежать у широких діапазонах: від десятків кбіт/с для низькошвидкісних прикладних програм до тисяч Мбіт/с для прикладних програм, орієнтованих на роботу з графічними даними.

Цікавим фактом є те, що як показали досліди, потік передачі даних не можна описувати моделлю найпростішого потоку. Виміри на реальних локальних мережах передачі даних, вимірювання потоку інформації при передачі зображень, вимірювання WWW-мережі Інтернет та в інших мережах передачі даних призвели до відкриття того, що трафік в них є самоподібним випадковим процесом. На якісному рівні самоподібність проявляється в тому, що присутня повільно спадаюча залежність між величинами трафіку в різні моменти часу, а також в тому, що трафік гуртується у пачки даних і ці пачки виглядають статистично подібними в широкому діапазоні зміни масштабу по шкалі часу.

Запис вимірювань, виконаних компанією Beilcore в локальній мережі Ethernet, відтворена в [50]. Результати статистичної обробки записаних даних дають право стверджувати, що виміряний трафік є самоподібним.

Самоподібний характер трафіку, що спостерігається у сучасних високошвидкісних мережах зв‘язку обумовлений тим, що мережі є інтегральними та використовуються для передачі мови, даних, зображень (в тому числі по факсу), файлів та інших видів інформації, що передається у формі стандартизованих пакетів. Такий трафік спричиняє значний вплив на характеристики систем зв‘язку, а саме, як встановлено в [51], при збільшенні розміру буферу на вході каналу ймовірність втрат падає значно повільніше, ніж по експоненційному закону, притаманному класичним моделям телетрафіку, що широко використовується.

2.1.3. ATM-трафік

В ATM передбачено декілька категорій послуг, які наведені в Табл. 2.1.

Табл. 2.1

Типи категорій ATM-услуг

Клас	Опис	Приклад
Cbr	Постійна швидкість передачі	Канал T1(E1)
Rt-vbr	Змінна швидкість передачі (реальний час)	Відеоконференції
nrt-vbr	Змінна швидкість передачі (нереальний час)	Мультимедіа по электронній пошті
Abr	Доступна швидкість передачі	Прегляд web-інформації
Ubr	Не специфікована швидкість передачі	Пересилка файлів у фоновому режимі

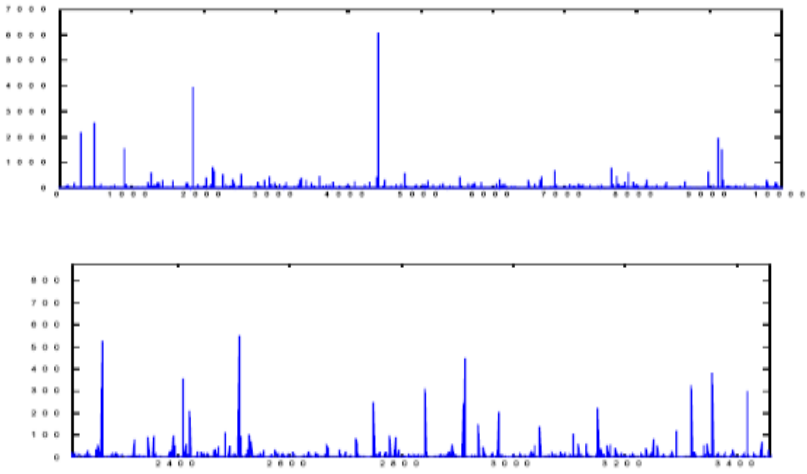


Рис. 2.1 Модель самоподібного трафіку для інтервалів часу різної довжини

CBR не передбачає контролю за помилками, керування трафіком або будь-якої обробки. Клас CBR придатний для мультимедійних програм реального часу.

Клас VBR містить в собі два підкласи – звичайний та для реального часу (див. табл. 2.1). ATM в процесі доставки не вносить зміни послідовності комірок в часі. Випадкові втрати ігноруються.

Клас ABR передбачено для роботи в умовах миттєвих варіацій трафіку. Система гарантує деяку пропускну здатність, але протягом короткого часу допускає й більше навантаження. Цей клас передбачає наявність зворотнього зв‘язку між приймачем та відправником, що дозволяє знизити завантаженість каналу, якщо це необхідно.

Клас UBR добре підходить для пересилки IP-пакетів (немає гарантії доставки і у випадку перевантаження, невиключені втрати).

ATM використовує виключно модель з встановленням з‘єднання. Це означає труднощі для керування трафіком з метою забезпечення потрібного рівня якості обслуговування. Для розв‘язку цієї задачі використовується алгоритм GCRA (Generic Rate Algorithm).

Багато джерел комірок ATM надходять до системи з фіксованою швидкістю (наприклад, відеоконференція). Вимога знизити швидкість передачі тут досить беззмістовна. З цієї причини в ATM розумніше передбачати перевантаження. Але для трафіку типів CBR, VBR и UBR не існує ніякого динамічного керування перевантаженням, тож адміністративне керування є єдиною можливістю. Коли система бажає встановити віртуальний канал, вона повинна охарактеризувати очікуваний трафік. Мереа аналізує можливість обробки додаткового трафіку з урахуванням різних маршрутів. Якщо реалізувати додатковий трафік не можна, запит анулюється. За відсутності адміністративного контролю декілька широкосмугових користувачів можуть блокувати роботу маси вузькосмугових клієнтів мережі, наприклад тих, що читають пошту.

Резервування ресурсів за своєю сутністю близьке адміністративному контролю та виконується на фазі формування віртуального каналу. Резервування

здійснюється уздовж всього маршруту (по всіх комутаторах) в ході реалізації процедури setup. Параметрами резервування може бути значення пікового значення смуги пропускання та/або середнє завантаження.

Для типів сервіса CBR та VBR відправником навіть у випадку перевантаження не може бути знижено рівень трафіку. У випадку UBR втрати не мають ніякого значення.

Керування перевантаженням для послуг типу ABR базується на тому, що кожний відправник має поточну швидкість передачі (ACR – Actual Cell Rate), що лежить між MCR (Minimum Cell Rate) та PCR (Peak Cell Rate) коли відбувається перевантаження, ACR зменшується, але не нижче MCR. При зникненні перевантаження ACR збільшується, але не вище PCR. Кожна RM- комірка містить значення завантаження, що пропонує реалізувати відправник. Це значення називається ER (Explicit Rate). На шляху до місця призначення ця величина може бути зменшена комутаторами, що знаходяться на шляху слідування інформації. Жоден з комутаторів не може збільшити ER. Модифікація ER може відбуватися як на прямому шляху, так і на зворотньому. При отриманні RM-комірки відправник може скорегувати значення ACR, якщо це необхідно.

Для моделювання вибрано два типи категорій ATM-послуг: Rt-vbr та nrt-VBR.

Фактично категорія Rt-vbr передбачається для передачі інтерактивного відео, тому, при моделюванні інформаційних потоків, що мають відповідну категорію, необхідно виходити з наступних позицій: моменти надходження нових пакетів до системи (***ti***, рис. 2.2) розподілені по пуасонівському закону, час тривалість обробки заявок даного трафіку (рис. 2.2) розподілений по закону Парето, заявка передається протягом тривалго часу інформаційними блоками рівної довжини, оскільки мова йде про ATM трафік.

Категорія nrt-VBR, що характеризується змінною швидкістю передачі у нереальному часі, передбачається для передачі мультимедійних інформаційних потоків. Як показали дослідження передача мультимедійних потоків найліпше характеризується поняттям самоподібного трафіку, нижче буде наведена його більш детальна матетматична модель.

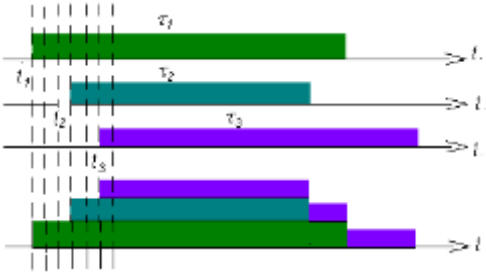


Рис. 2.2. Модель самоподібного трафіку

Для моделювання в роботі вибрані тільки ці дві категорії послуг, оскільки, вони є найменш дослідженими на сьогоднішній день. При цьому вимоги до передачі таких трафіків є високими і немає оптимального способу боротьби з перевантаженнями. Оптимізація розподілу інформаційних потоків між чергами вихідних за критерієм переповнення черг може зменшити випадки перевантаження.

2.1.4. Формалізація самоподібних інформаційних потоків. Модель мультимедійного трафіку

Як вже було показано в першій главі, на сьогоднішній день в мережах мобільних операторів зв‘язку циркулюють трафіки різної природи з різними вимогами до передачі. Крім цього, відмінним є і характер вхідних потоків. Вже доведено, що Інтернет трафік є самоподібнім процесом, тобто розрахунки, які проводилися для пуасонівського вхідного потоку не завжди відповідають дійсності.

В [52], [53], [54] наводиться опис дослідів, які показують, що мультимедійні – це самоподібні процеси.

Основна маса математичних моделей будується виходячи з припущення, що вхідний інформаційний потік описується пуаснівським розподілом з параметром λ (інтенсивність надходження заявок). Як показано в [55] пуасонівський процес є частим випадком самоподібного процесу, для якого параметр Херста H=0,5.

На якісному рівні самоподібність проявляється в тому, що наявна повільно спадаюча залежність між величинами трафіку в різні моменти часу, а також в тому, що трафік гуртується у пачки даних і ці пачки виглядають статистично подібними в широкому діапазоні зміни масштабу по шкалі часу.

Загальна модель трафіку [56, 57] Трафік Y, що розглядається, – це потік бітів інформації. Пакети складаються з деякої кількості бітів, вони генеруються джерелами таким чином, що трафік Y являє собою суперпозицію бітів інформаційних пакетів, які генеруються джерелами.

Розглянемо випадкові джерела визначеного процесу Y=(…,$Y_{-1,}Y_0$ ...)на дискретній осі часу t. Позначимо через $\xi_t \in I_0$кількість джерел, що надіслали пакети у момент часу t. Відповідно до цього приходу, джерела інформації були занумеровані індексами s ($s \in Z = \{\ldots, -1, 0, 1 \ldots\}$). Позначимо через ω_s час утворення джерела s, $\omega_s \le \omega_{s+1}$. Це передбачає, що з кожною точкою s асоціюється випадкова функція (дискретного) часу t, це послідовність$\{\theta_s(t-\omega_s+1), \omega_s \le t \le \omega_s+\tau_s-1\} = (\theta_s(1), \ldots, \theta_s(\tau_s))$, де $\theta_s(\tau)$– це активний період джерела s з довжиною $\tau_s \in I_1$. До моменту ω_s і після моменту ω_{s+i-1} джерело пакети не генерує, $\theta_s(i) = 0$ при i≤0 та i≥$\tau_s + 1$. Таким чином, $\theta_s(t-\omega_s+1), t \in Z$, – послідовність числа пакетів, які генерує джерело s у послідовні моменти часу. Точки незалежні однаково розподілені для різних s, також вони незалежні для послідовностей ξ_t таω_s.

Отже процес Y може бути перевизначений як

$$\boldsymbol{Y_t = \sum_{s \in Z} \theta_s (t - \omega_s + 1), t \in Z} \qquad (2.2)$$

Для постановки задачі зручно зафіксувати такі поняття пов‘язані з Y:

s – джерело інформації;

ω_s- час находження джерела s;

$\theta_s(i)$- представляє собою кількість бітів інформації згенерованих джерелом s за час

τ_s– довжина активного періоду джерела s.

Позначимо

$\gamma_s = \theta_s(1) + \cdots + \theta_s(\tau_s)$– загальна кількість бітів, що була згенерована джерелом s за активний період γ_s, називається розміром джерела s.

Процес Y визначений як (2.2) називається тут як процес-джерело. Y – це суперпозиція активних періодів джерел інформації. Y_t – загальна кількість бітів інформації згенерованих всіма активними джерелами в момент t.

Тепер, ξ_t визначає кількість нових джерел, що надійшли в момент t. Покладається, що ξ_t – незалежні одинаково розподілені за законом Пуасона з параметром $\lambda = M\xi_t$,0<λ<∞. Це дає те, що пари (довжина активного періоду джерела s, активні періоди джерела s) $(\tau_s, \theta_s(1), \dots, \theta_s(\tau_s))$ є незалежними одинаковорозподіленими для різних s, а також вони не залежать від послідовностей ξ_t таω_s.

В [17] розглянуті наступні часткові випадки процесу Y з різними активними періодами:

1) $Y^{(1)}$ константа $\theta_s(i) = R \in N, 1 \le i \le \tau_s$, де R – швидкість джерела;

2) $Y^{(2)}$, випадкова константа $\theta_s(i) = R$, де R=R(τ_s);

3) $Y^{(3)}$, незалежні та однаково розподіленні (НОР) $\theta_s(i)$ приймають значення 0 та 1 з ймовірностями p_0 і p_1 відповідно;

4) $Y^{(4)}$, НОР $\theta_s(i)$ приймають значення множини {0,1,…,k} з біноміальним розподілом або з Z геометричним, пуасонівським або з якимось іншим заданим розподілом;

Умова самоподібності моделі трафіка. Необхідна і достатня умова того, що Y є строго самоподібним в широкому сенсі (ССШС) процесом представлена в наступній теоремі, що використовує позначення $\mu^{(1)} = M\theta(t)$ (інтенсивність надходження заявок) і $B^{(l)}(k) = M\theta(t)\theta(t+k), k \in Z$, де (…,θ(-1),θ(0),θ(1),…) – залежний від l випадковий стаціонарний процес, розподіл якого на довжині l співпадає з умовним розподілом активного періоду джерела при τ = l (тут і на далі τ – випадкова величина, яка має такий саме розподіл, як τ_s).

Теорема 1. Процес Y буде ССШС з 0<β<1 (H=1(β/2)) тоді і тільки тоді, коли $\Pr\{\tau = l\}$, $\mu^{(t)}$ і $B^{(l)}(k)$ задовольняє умовам

$$\frac{\sum_{l=k+1}^{\infty}(l-k)\Pr\{\tau=l\}B^{(l)}(k)}{\sum_{l=1}^{\infty} l\Pr\{\tau=l\}B^{(l)}(0)} = \frac{1}{2}\delta^2\left(k^{2-\beta}\right), k \in N \quad (2.3)$$

$$\sum_{l=1}^{\infty}\Pr\{\tau = l\}\mu^{(1)} < \infty. \quad (2.4)$$

А саме, якщо $\mu^{(1)}$ не залежить від l, а $B^{(l)}(k)$ постійна величина, яка не залежить від l і k, то Y буде процесом ССШС тоді і тільки тоді, коли

$$\Pr\{\tau = l\} \frac{3^{2-\beta} - 2^{4-\beta} + 7}{4 - 2^{4-\beta}}; \quad (2.5)$$

$$\Pr\{\tau = l\} \frac{\delta^4(k^{2-\beta})}{4 - 2^{4-\beta}}, k \in \{2,3,..\}, \quad (2.6)$$

де

$$\delta(f(x)) = f\left(x + \frac{1}{2}\right) - f(x - \frac{1}{2})$$

$$\delta^4(k^{2-\beta}) = (k+2)^{2-\beta} - 4(k+1)^{2-\beta} + 6k^{2-\beta} - 4(k-1)^{2-\beta} + (k+2)^{2-\beta}, k \in \{2,3,\dots\}, \quad .$$

і розподілPr$\{\tau = k\}$, який визначається (2.5) та (2.6), має скінчене середнє значення $M\tau = (2 - 2^{1-\beta})^{-1}$, нескінченну дисперсію і наступну експоненційну асимптотику

$$\mathbf{Pr}\{\boldsymbol{\tau = l}\}\frac{(\boldsymbol{\beta+1})\boldsymbol{\beta}(\boldsymbol{\beta-1})(\boldsymbol{\beta-2})}{\boldsymbol{4-2^{2-\beta}}}, \boldsymbol{k^{-(2+\beta)}}, \boldsymbol{k\to\infty},$$

Достатня умова того, що Y – процес асимптотично самоподібний в широкому сенсі (АСШС), виражена у наступній теоремі.

Теорема 2 [88]. Процес Y являється АСШС з $0 < \beta < 1$ тоді, коли $\mathbf{Pr}\{\boldsymbol{\tau = l}\}$, $\mu^{(1)}$ і $B^{(l)}(k)$ такі, що

$$\mathbf{Pr}\{\boldsymbol{\tau = l}\}\, B^{(l)}(l) \sim L(l) l^{-(\beta+2)}, l\to\infty$$

$$\sum_{l=1}^{\infty} \mathbf{Pr}\{\boldsymbol{\tau = l}\}\, B^{(l)}(0) < \infty$$

$$\sum_{l=1}^{\infty} \mathbf{l Pr}\{\boldsymbol{\tau = l}\}\, \mu^{(1)} < \infty$$

де L(x) – будь яка функція, що повільно змінюється на нескінченності.

А саме, відповідно до теореми 2, процес Y стає АСШС з $H = (3 - \alpha)/2$, якщо $\mu^{(1)}$ не залежить від l, $B^{(l)}(k)$ не залежить від l і k, а

$$\mathbf{Pr}\{\boldsymbol{\tau = l}\} = \boldsymbol{c_0 l^{-\alpha-1}}, \mathbf{1} < \boldsymbol{\alpha} < 2, l \in \boldsymbol{N}$$

$$c_0 = \ln l,$$

c_o – повільно спадаюча функція.

Тобто τ розподілено за законом Парето.

2.1.5. Модель системи зв‘язку та визначення верхньої межі для імовірності переповнення буферу

Як показали дослідження, визначення показників якості обслуговування для самоподібного трафіку є задачею складною, яка не має простих розв‘язків. В [17] Б. Цибаковим та Н. Георганасом наводиться ряд моделей ATM трафіку, для яких наводиться розрахунок верхньої границі переповнення буферу для черги ATM трафіку. Основні ідеї задач, поставлених в статті, вказані у попередньому пункті. Ці моделі були взяті за основу для оцінки якості роботи комутаційного центру.

Для оцінки роботи комутаційного центру PE (Provider Edge), що може забезпечувати сервіси технології ATM довжина пакету еквівалентна одному ATM-пакету.

Отже, розглядається задача дискретного часу для системи Y1/DC/1/h, де вхідний трафік – це трафік Y1, а час обслуговування C^{-1} , обслуговування ведеться одним сервером, скінчений буфер складається з h біт інформації. Тут приймається, що система Y1/DC/1/h ~ Y1/D/C/h з C обслуговуючими пристроями та одиничним часом обслуговування.

Послідовність подій за один момент часу наступна:

{Закінчення обслуговування в слоті t-1}

{Закінчення слота t-1}

{Прихід нових бітів інформації якщо інформація надходить у t}

{Вибір нових бітів інформації для передачі}
{Відкидання бітів інформації, якщо це потрібно}
{Постановка комірок, що не були відкинуті в чергу}
{Початок слота t}
{Початок обслуговування слота t}.

Розглядаються дві дисципліни обслуговування:

(i) Якщо $y_t + z_t > 0$, де y_t – кількість нових бітів інформації, що надійшли в момент t , z_t – кількість бітів у буфері на момент t, тоді $\min(y_t+z_t,C)$ переходить на обслуговування в t.

(ii) Якщо$y_t + z_t \leq C + h$, тоді в момент t біти не втрачаються. Якщо $y_t + z_t > C + h$, тоді $y_t + z_t - C - h$ комірок втрачаються в момент t.

Позначимо $D_C(h)$ клас наведених дисциплін обслуговування.

Подія $\{y_t + z_t > C + h\}$ називається переповненням буферу в момент t, а ймовірність

$P_{over}^t = \Pr\{y_t + z_t > C + h\}$– визначає ймовірність переповнення.

Нас цікавить оцінка верхньої межі для стаціонарного розподілу ймовірності переповнення $P_{over} = limsup\, P_{over}^t$

.

Верхня межа P_{over} для скінченого буферу складається з верхньої межі для ймовірності знаходження h або більше бітів інформації у нескінченому буфері трафіку Y1.

Нехай $N_t(h)$ – кількість бітів в буфері обміну G/D/C/h/d_1 в системі зразу після позиції {Прихід нових бітів інформації якщо інформація надходить у t}, $d_1 \in D_c(h), N_t(\infty) = N_t$ – кількість бітів в буфері обміну G/D/C/h/d_2 в системі, $d_2 \in D_c(\infty)$, границю N_t при $t\to\infty$ позначимо через N_∞, якщо границя існує.

Твердження. Якщо вхідний трафік для системи G/D/C/h/d_1, }, $d_1 \in D_c(h)$ та G/D/C/h/d_2, $d_2 \in D_c(\infty)$, однаковий, тоді

$$P_{over} \leq \Pr\{N_\infty > h\}, \tag{2.7}$$

$\Pr\{N_\infty > h\} = limsup Pr\{ N_t > h\}$.

Зараз треба звернути увагу на систему Y1/D/C/$_\infty$ де Y1 – самоподібний трафік з розподілом активних періодів τ джерел інформації по закону Парето. Закон розподілу ймовірності τ асимптотично такий:

$$\Pr\{\tau = l\} \approx c_0 l^{-\alpha-1}, \tag{2.8}$$

де c_0 – повільно спадаюча функція.

Задача знайти асимптотичну ($h\to\infty$) верхню межу для $Pr\{ N_\infty > h\}$. Стаціонарний розподіл ймовірності для більше ніж h біт у нескінченому буфері. Коли межа буде знайдена, можна буде застосувати Твердження для знаходження ймовірності переповнення скінченого буферу Y1/D/C/h.

Пропонується ввести інтервал для n успішних моментів часу (t-n, …, t-1)= i_n для декількох t і n. Позначимо

$$T_n = \sum_{j=i_n} Y_j,$$

T_n – загальна кількість бітів інформації що надійшли з трафіком Y1 на інтервалі i_n.

Як відомо [18]:

$$\boldsymbol{Pr\{ N_t > h\} = Pr\{sup(T_n - nC) > h\},} \tag{2.9}$$

де права сторона ймовірності є стаціонарним розподілом та не залежить від t.

Для будь-якого цілого n_0 , $1 \le n_0 < \infty$ справедливо

$$Pr\{sup(T_n - nC) > h\} \le Q_1 + Q_2, \quad (2.10)$$

де $Q_1 = \Pr\{\max_{1 \le n \le n_0} (T_n - nC) > h\}$, (2.11)

$$Q_2 = \Pr\left\{\sup_{1 \le n \le n_0} (T_n - nC\right) > h, \quad (2.12)$$

Отже треба знайти дві верхні межі: одну для Q_1, а другу для Q_2. Варто відмітити, що асимптотично $Q_1 \le const h^{-\alpha}$ і $Q_2 \le const h^{-\alpha+1}$, виходячи з (2.10)-(2.12), тому отримуємо асимптотичну границю

$Pr\{ N_t > h\} \le const h^{-\alpha+1}, h \to \infty.$

Спочатку знайдемо ймовірність Q_1.

Позначимо $T_n = T_n(1) + T_n(2)$, (2.13)

де $T_n(1)$ – загальна кількість бітів, отриманих в T_n від джерел інформації, які щойно надійшли, тобто тих, що стартували свої активні періоди протягом i_n, $T_n(2)$ – стартували передачу до моменту часу t-n.

Нехай Aj позначає множину джерел, що надійшли у момент часу j. Потужність множини |Aj| еквівалентна ξ_j .

Тоді

$$T_n(1) \le \sum_{s \in A_{t-n}} \tau_s + \cdots + \sum_{s \in A_{t-1}} \tau_s = \overline{T_n(1)} \quad (2.14)$$

Нехай $A^{(t-n)}$ позначає множину джерел, які надійшли раніше за час t-n, але не закінчили своїх активних періодів до моменту t-n. Потужність $|A^{(t-n)}|$ – це пуасонівська випадкова величина із скінченим параметром λ_0 (може бути показано, що) $\lambda_0 = \lambda(M\tau) - \lambda)\backslash$

$$T_n(2) \le \sum_{s \in A_{t-n}} \tau_s = T_n(2) \quad (2.15)$$

Доданки в (2.14) і (2.15) є незалежними однаково розподіленими, а потужності $|A^{(t-n)}|$, $|A_j|$, $1 \le j \le n$ незалежні та пуасонівські.

Таким чином

$$T_n(1) + T_n(2) = \sum_{s=1}^{\chi} \tau_s \quad (2.16)$$

де τ_s – незалежні однаково розподілені з розподілом ймовірностей $\mathbf{Pr}\{\boldsymbol{\tau} = \boldsymbol{l}\}$ та кількістю доданків в (2.16) χ , де χ – пуасонівська випадкова величина з параметром $\lambda^* = n\lambda + \lambda_0$.

Використовуючи позначення (2.12), (2.13)-(2.16) отримуємо

$$Q_1 = \Pr\{max\, T_n > h\} \le \Pr\left\{\max\left(\overline{T_n}(1) + \overline{T_n}(2)\right) > h\right\} \le \sum_{n=1}^{n_0} Pr\{\sum_{s=1}^{x} \tau_s > h\} \quad (2.17)$$

Нехай $T=X_1+\ldots+X_{\chi}$ незалежна однаково розподілена сума змінних X_s з імовірністю $\Pr\{X>x\}$ визначаються як

$$Pr\{X > x\} \le x^{-\alpha} L(x),\ x \to \infty,\ \alpha > 0. \quad (2.18)$$

χ- пуасонівська випадкова величина з параметром$\lambda*$, і χ не залежить від X_1, X_2,… В (2.18) L(x) –повільно спадаюча функція на нескінченності.

Тоді

$$Pr\{T > x\} \approx \lambda^* Pr\{X > x\} \quad (2.19)$$

Аналізуючи (2.7) та застосовуючи (2.18) і (2.19), для ймовірності в правій частині (2.16) отримуємо

$$Q_1 \le c_2 h^{-\alpha},\ c_2 = \frac{h_0(n_0\lambda + \lambda_0)c_0}{\alpha},\ h \to \infty \qquad (2.20)$$

Тепер повернемося до Q2. У відповідності (2.12)-(2.16)

$$Q_2 = Pr\left\{\max_{n>n_0}(\bar{T}_n(1) + \bar{T}_n(2)) > h + nC\right\} \le Pr\{\sup \textstyle\sum_{s=1}^{x} \tau_s > h + N_c\} \qquad (2.21)$$

З (2.21) отримуємо

$$Q_2 \le Pr\{\sup(\textstyle\sum_{s=1}^{x'} \tau_s - nC) > h\} \qquad (2.22)$$

де χ' - пуасонівська випадкова величина з параметром. $n\left(\lambda + \frac{\lambda_0}{n_0}\right)$

Тепер розглянемо систему $M_C/G/1/\infty$/FIFO на неперервній вісі часу , де M_C – пуасонівський прихід джерел інформації з інтенсивністю $\left(\lambda + \frac{\lambda_0}{n_0}\right)/C$. На неперервному часі, а G позначає порядок обслуговування абонентів розподілений за законом $\Pr\{\tau = l\}$. Для систем з неперервним часом позначимо $T_{t'}^{const}$ як загальний об'єм обслуговування, що запитується джерелами інформації, які щойно надійшли на інтервалі $(t\text{-}t', t)$.

У відповідності до специфікації $M_C/G/1/\infty$/FIFO, випадкова змінна $\sum_{s=1}^{\chi} \tau_s$ та $T_{t'}^{const}$ однаково розподілені. Враховуючи це, отримуємо

$$\Pr\left\{\sup_{n>n_0}\left(\textstyle\sum_{s=1}^{\chi} \tau_s - nC\right) > h\right\} \le \Pr\left\{\sup_{t\in[0,\infty]}\left(T_{t'}^{const} - t'\right) > h\right. \qquad (2.23)$$

З правої сторони (2.23) стаціонарний розподіл ймовірностей (при$\left(\lambda + \frac{\lambda_0}{n_0}\right)/C$)<1, тому віртуальний запит часу очікування для $M_C/G/1/\infty$/FIFO більший за h. Якщо τ розподілено за законом Парето, тобто $M\tau = c_0 \sum_{l=1}^{\infty} l^{-\alpha}$, тоді остання ймовірність може бути асимптотично знайдена за допомогою формул для системи M/G/1/∞ /FIFO, де ймовірність перевищення часу обслуговування більша заданої величини: $P_{over}\{\tau > x\} \sim \frac{\lambda}{\alpha(\alpha-1)(1-\rho)} x^{-\alpha} L(x), \alpha > 1, x \to \infty$,ρ<1 якщо $P_{over}\{\tau > x\} \sim x^{-\alpha} L(x), \alpha > 1, x \to \infty$, $\mu^{-1} = \int_0^{\infty} Pr\{\tau > y\}\, dy$.

Потім, використовуючи (2.22) та (2.23), ми отримуємо асимптотичну границю для Q_2. Нарешті, використовуючи границю для Q_2 разом з (2.7), (2.9), (2.10) та (2.20) та уточнюючи, що λ_0скінчена і n_0 довільне, отримуємо відповідну границю для ймовірності переповнення:

$$P_{over} \le \frac{c_{0\lambda}}{\alpha(\alpha-1)(C - \lambda M_\tau)} h^{-\alpha+1}, h \to \infty, \lambda M_\tau < C$$

Це відповідає для Y1/D/C/h/d, $d \in D_c(h)$, тобто для асимптотично самоподібного вхідного трафіку з параметром Херста $H = \frac{3-\alpha}{2}$, τ розподілено по закону Парето, ємність вихідного каналу C та розмір буферу h, $c_0 = ln\ l$ – повільно спадаюча функція.

Якщо, швидкість R джерела інформації не одинична, а τ задана деяким чином, самоподібній трафік $Y^{(4)}$, тобто крім розподілу для використовується розподіл для γ (швидкість джерела)

$$\gamma = \textstyle\sum_{i=1}^{r} \theta(i),$$

де $\theta(i)$ – узагальнене позначення для $\theta_s(i)$, та якщо $Pr\{\gamma = l\}$ має розподіл Парето або експоненційний розподіл. Тоді для трафіку Y(4)/D/C/h/d,$d \in D_c(h)$, отримуємо

$$P_{over} \leq \frac{\lambda}{C - \lambda E_\gamma} \sum_{n=h}^{\infty} Pr\{\gamma > n\}, h \to \infty, \lambda M_\gamma < C.$$

В роботі буде розглядатися наступний частий випадок трафіку Y(4), коли $\theta(i) = R,\ R \in I_1$. В цьому випадку Y(4) є Y(1). Якщо $Pr\{\tau = l\}$ визначається за формулою (2.8), тоді використовуючи $Pr\{\gamma > x\} = Pr\left\{\tau > \frac{x}{R}\right\}$, отримуємо

$$P_{over} \leq \frac{c_0 \lambda R^\alpha}{\alpha(\alpha-1)(C-\lambda R M_\tau)} h^{-\alpha+1}, h \to \infty, \lambda R M_\tau < C, \quad (2.24)$$

де P_{over} – верхня границя переповнення черги у вихідному каналі; c_0 – мінімальна довжина повідомлення; λ – інтенсивність вхідного потоку, що направляється у вихідний канал; R – швидкість передачі джерела інформації; α – параметр самоподібного потоку, який є рішенням рівняння:

$$M_\tau = c_0 \sum_{l=1}^{\infty} l^{-\alpha},$$

M_τ – середня довжина повідомлення; h – довжина черги в вихідному каналі; C – швидкість передачі інформації в каналі зв'язку.

2.2. Метод підвищення ефективності обслуговування інформаційних потоків в комутаційному центрі транспортної мережі

Відповідно до проведеного аналізу мереж мобільних операторів зв'язку та технологій, які використовуються для передачі інформаційних потоків, з'ясувалося, що найбільш перспективною технологією є протокол емуляції постійного з'єднання PWE3. Ця технологія дозволяє підтримувати якість передачі інформаційних потоків на рівні таких технологій як ATM, Fast Ethernet, Frame Relay, які основані на способі організації з'єднання каналів (технологія комутацій каналів), при цьому передача інформаційних потоків ведеться через мережу з комутацією пакетів (PNS – Packet Switched Networks). Аналіз технології показав, що може бути ефективною оптимізація передачі інформаційних потоків через комутаційний центр PE, що є воротами між базовими станціями, які до нього під'єднані, та мережею PNS, де організовані тунелі для передачі інформаційних потоків між двома комутаційними центрами PE1 та PE2.

2.2.1. Вибір способу керування інформаційними потоками в комутаційному центрі

Аналіз роботи мереж за протоколом PWE3 показав, що розподіл інформаційних потоків від базових станцій по тунелях відведених для передачі інформації є неефективним. Передача здійснюється таким чином, що для кожної базової станції прокладається окремий тунель – емульоване постійне з'єднання, по якому організується передача всіх інформаційних потоків від однієї базової станції.

На рис. 2.3 показана схема передачі інформації через мережу з комутацією пакетів PSN між двома комутаційними центрами PE.

Така схема може бути не ефективною через те, що навантаження на базові станції, залежно від часу може змінюватися. Наприклад, до одного комутаційного центру під‘єднані базова станція, що охоплює територію, де знаходяться бізнес-центри, та базова станція, що стоїть у спальному районі. Очевидно, що протягом дня навантаження на першу БС буде значно більшим, ніж навантаження на другу, а ввечері та вночі, навпаки, перевантаженою буде друга БС. Тобто за схемою, яка пропонується протоколом PWE3 тунель організований для першої базової станції у робочі години буде перевантаженим, в той час коли тунель для другої базової станції навпаки буде напівпорожнім.

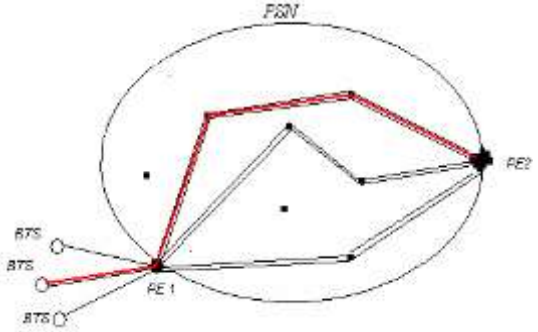


Рис. 2.3 Організація PWE3

Такі спостереження обумовили виникнення науково-технічної задачі ефективного розподілу канального ресурсу мережі, яка полягає у ефективній передачі сумарного інформаційного потоку від всіх базових станцій по всіх тунелях, що виходять з комутаційного центру та прямують до головного контролеру базових станцій.

Аналіз існуючих методів обслуговування інформаційних потоків в КЦ показав, що фактично йдеться про задачу вузлової маршрутизації в мережах з комутацією пакетів [20].

2.2.2. Загальна характеристика алгоритмів керування потоками в мережах

Процедури керування інформаційними потоками в PSN мережах визначаються протоколами різних рівнів архітектури мережі у відповідності до еталонної моделі OSI/ISO.

До функцій міжвузлового керування потоками мережевого рівня (PE1–PE2) відносяться протоколи, що дозволяють уникати блокування в мережі та виходити з нього.

Керування на рівні комутаційних центрів PE1–PE2 здійснюється децентралізовано в кожному вузлі шляхом обмеження навантаження на КЦ, коли перевищується деякий поріг, що обмежує максимальну довжину черги. Функції контролю довжини черги, зриву та повторної передачі пакетів часто виконуються протоколом керування у вузлі зв‘язку. Існує декілька моделей цього рівня керування

потоками. Перша група моделей названа схемами обмеження черги до каналу. В цьому випадку на довжину черги для кожного вихідного каналу накладається обмеження. Якщо черга досягла границі, то пакети відкидаються. Схема обмежень черги до каналу має наступні модифікації [59]:

1.Модель повного розділення, для якої справедливе обмеження $0 \le n_i \le B/n,\ \forall_i$, де n – число вихідних каналів , n_i – число пакетів i-ї черги, B – розмір КЦ.

2.Модель розподілу по максимальних чергах $0 \le n_i \le b_{max}, \forall_i,\ \sum_i n_i \le B$, де $b_{max} > B/n$ - максимально допустимий розмір черги.

3.Модель розподілу за мінімальним розміщенням $\sum_i \max(0, n_i - b_{min}) \le B - nb_{min}$, b_{min} – мінімальний розмір буферу, який гарантується для кожної черги (як правило, $b_{min} \le B/n$).

4.Модель розподілу по мінімальному розміщенню та максимальній черзі (з‘єднання моделей 2 і 3).

Методи керування потоками на рівні доступу до мережі пов‘язані з обмеженням вхідного навантаження з метою виключення перевантаження мережі. Перевантаження може бути локальним (заняті буфери на вихідних каналах), глобальним (заняті всі буфери, доступні в мережі), селективним (заблоковано один з шляхів, що веде до PE2).

Таким чином, для організації ефективного керування потоками в КЦ, що з‘єднує базові станції з контролером базових станцій, необхідно розібратися з наступними положеннями:

- характер вхідних інформаційних потоків, можливості їх обмеження;
- оцінка вихідних трактів (емульованих постійних з‘єднань);
- вибір ефективного способу маршрутизації.

2.2.3. Задача вузлової маршрутизації PWE3

Задача вузлової маршрутизації для мереж мобільних операторів зв‘язку, де використовується протокол емуляції постійного з‘єднання PWE3, повинна враховувати мультисервісний характер потоку. Задача вузлової маршрутизації для однотипного пуасонівського потоку вирішується при розв‘язанні наступної оптимізаційної задачі.

На вхід вузла комутації поступає АТМ трафік інтенсивністю λ. Трафік повинен бути розподілений по n трактам, що виходять з PE1. Тракти характеризуються кількістю каналів k_j. В кожному з каналів тракта $(j = \overline{1, n})$ покладається, що вхідний потік пуасонівський. Тоді пуасонівським будуть і потоки в трактах.

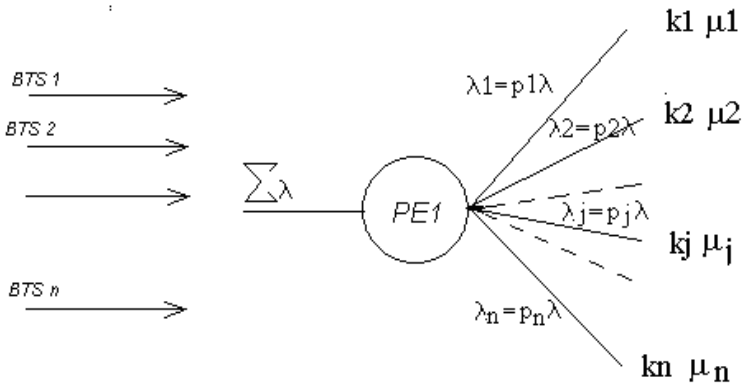


Рис. 2.4 Задача вузлової маршрутизації

Параметрами керування є ймовірності x_j, які пов‘язують інтенсивності потоків в трактах λ_j з величиною загального трафіку

$$(\lambda_j = x_j\lambda, j = \overline{1,n}) \tag{2.25}$$

Для оцінки якості обслуговування абонентських заявок, ефективності розподілу повідомлень між вихідними каналами за критерій вибрано верхню межу ймовірності переповнення буферу($P_{over\ j}$), тобто ймовірність того, що пакети, які надійшли до PE1 та були направлені в j-й канал загубляться через те, що буфер цього каналу був переповнений. $P_{over\ j}$ – розраховується за формулою (2.24) на основі даних про інтенсивність надходження заявок в j-й каналі зв‘язку.

Оцінити ефективність розподілу інформаційних потоків в PE1 можна за допомогою наступного адитивного критерію:

$$W_{over} = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda} P_{over\ j}$$

Або

$$W_{over} = \sum_{j=1}^{n} x_j P_{over\ j}$$

Розподіл потоку буде оптимальним, якщо знайти такі$\{x_j\}$, що

$$W_{over} = \sum_{j=1}^{n} x_j P_{over\ j} \Rightarrow \min_{\{p_j\}}, \tag{2.26}$$

та задовольняють умові

$$\sum_{j=1}^{n} x_j = 1,\ x_j \geq 0,\ j = \overline{1,n.}$$

2.2.4. Задача вузлової маршрутизації вхідного потоку для комутаційного центру PE

Обробка заявок в комутаційному центрі PE мережі PWE3 є аналогічною до роботи будь якого комутаційного центру.

На рис. 2.5 показана схема роботи комутаційного центру, де видно, що всі вхідні потоки, які поступають на вхід до комутаційного центру спершу комутуються, тобто розподіляються по каналах зв‘язку, а потім класифікуються та утворюють черги відповідно до типу трафіку. Таким чином, якщо на вході спостерігається сплеск навантаження заявок заданого типу, тоді втрати пакетів з перевантаженої черги зростають.

Нещодавні дослідження показують, що видалення повідомлень має серйозні надлишкові ефекти. Наприклад, коли повідомлення втрачено, прикладна програма відправник може розглядати це як сигнал про те, що воно передає пакети занадто швидко. TCP реагує на такий сигнал уповільненням відправки повідомлень. Але, коли черга переповнена, тоді декілька повідомлень відкидаються одне за одним – в результаті цілий ряд прикладних програм не допускаються до обслуговування одне за одним, і як наслідок більшість прикладних програм вирішує уповільнити передачу. Після цього прикладні програми зондують мережу для визначення її завантаженості та буквально через декілька секунд поновлюють передачу з попереднім темпом, що знову приводить до перевантаження. Як показано, такі механізми керування трафіком, глобально не вирішують задачу боротьби з перевантаженням.

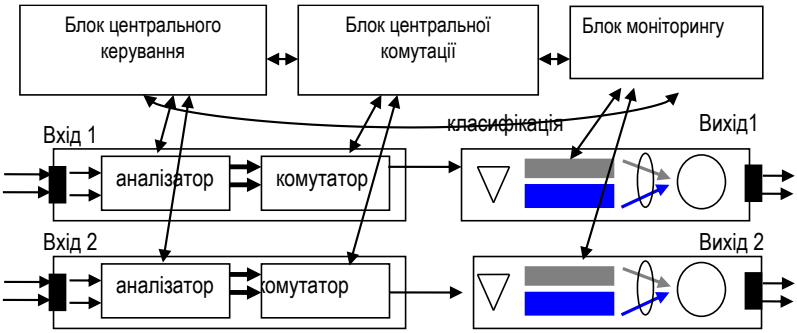


Рис. 2.5 Класична схема обробки заявок в комутаційному центрі PE

Фактично на сьогоднішній день маршрутизація інформаційних трафіків від PE1 до PE2 в комутаційному центрі мережі PWE3 здійснюється за принципом «точка-точка» тому, що організується емуляція постійного з‘єднання. Канали «точка-точка», наприклад PPP або HSSI є повною протилежністтю локальним мережам, оскільки тут ми маємо справу тільки з двома ділянками. Деякі архітектури маршрутизації розглядають їх як внутрішні інтерфейси між двома половинками маршрутизатора, в той час коли інші – як вироджений випадок локальної мережі. В конфігурації «точка-точка» система має змогу повністю контролювати

характеристики трафіку. А також допускає зміну схеми роботи комутаційного центру РЕ (рис. 2.6)

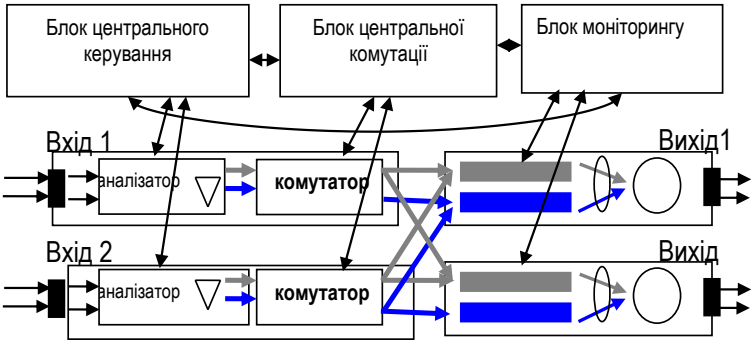


Рис. 2.6 Вдосконалена схема роботи комутаційного центру

Класифікація трафіку до процесу комутації дозволить контролювати наповнення черг заданих класів трафіків у вихідних каналах. Тобто у разі переповнення черг розв‘язок наступної задачі динамічно маршрутизації буде ефективним.

2.2.5. Задача вузлової маршрутизації для мультисервісного трафіку

Як показали дослідження, трафік, що надходить до РЕ1 є мультисервісним трафіком, тобто складається з інформаційних пакетів різної природи, наприклад VoIP, video, ftp, email, тощо. Крім різного змісту інформації, трафіки відрізняються вимогами до передачі, зокрема трафік інтерактивного відео дуже чутливий до затримок інформаційних пакетів, у разі перевищення часу допустимої межі затримки суттєво втрачається якість передачі інформації. До того ж на сьогоднішній день спостерігається різке зростанні об‘ємів трафіку інтерактивного відео, все частіше абоненти з мобільних станцій користуються відеоконференцзв‘язком, або фільтрами в режиму реального часу. Відео трафік може бути описаний за допомогою самоподібного трафіку.

Трафік, що передається по цій смузі, є надзвичайно різнорідним, від коротких SMS повідомлень до цілих фільмів, тобто це Інтернет трафік. Як було сказано в першій главі даної роботи, Інтернет трафік не може бути описаний за допомогою класичних моделей найпростішого потоку, для дослідження потоку Інтернет також найбільше підходить модель самоподібного трафіку.

Різниця моделей, що описує характер інформаційних потоків при розв‘язку оптимізаційної задачі, що буда описана далы, задає характер розрахунку ймовірності переповнення черги. Оскільки весь вхідний потік, що складається з трафіку інтерактивного відео та Інтернет трафіків розділяється на дві черги, то задача оптимізації повинна вирішуватися в комплексі для системи, з урахуванням вимог до якості передачі кожного класу трафіку окремо.

Необхідно знайти такий розподіл коефіцієнтів x_{ij} , що мінімізує цільову адитивну функцію від показника втрат пакетів (ймовірності переповнення черги) заданого класу трафіку у вихідних каналах комутаційного центру PE1.

У загальному випадку на вхід комутаційного центру надходять трафіки типу s (для передачі інтерактивного відео та Інтернет трафіків s=2), від n БС з інтенсивностямиλ_{ij} $(i = \overline{1,s}, j = \overline{1,n})$.

Таким чином, на вхід PE1 надходить трафік заданого типу з відповідною інтенсивністю $\lambda_i = \sum_{i=1}^{n} \lambda_{ij}$. Трафік повинен бути розподілений по n трактах (кількість вихідних каналів дорівнює кількості PW), що з'єднують PE1 і PE2. Тракти характеризуються швидкістю передачі C_j (j=$\overline{1,n}$), тобто кількісттю ATM пакетів, що можуть бути передані в канал зв'язку PW за один такт.

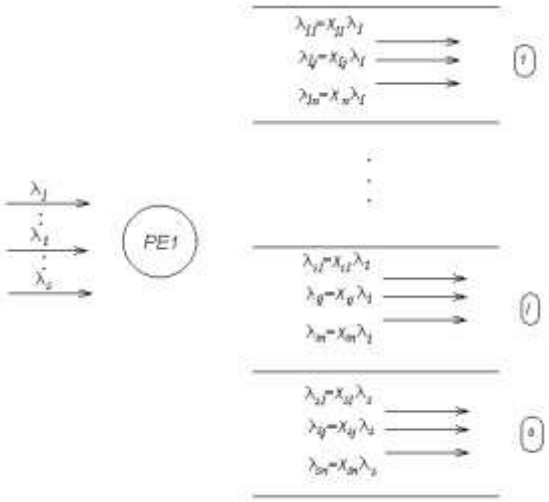


Рис. 2.7 Задача вузлової маршрутизації

Параметри керування – коефіцієнти x_{ij}, які зв'язують інтенсивності потоків в трактах λ_{ij} з величиною загального трафіку λ_i:

$$\lambda_{ij} = x_{ij}\lambda_i, \qquad j = \overline{1,n}, \qquad i = \overline{1,s}$$

Доля потоку *i*-го трафіку λ_{ij} , яка направляється в *j*-й канал фіксується, до моменту, коли система сигналізує про рівень втрат пакетів вище заданого значення. При надходженні відповідної інформації про втрати пакетів система заново перераховує значення x_{ij} .

Тоді оптимізаційна задача (2.26) приймає наступний вигляд:

$$W_{over}(\lambda) = \sum_{i=1}^{s} a_i \left(\sum_{j=1}^{n} x_{ij} P_{over\ ij} \right) \Rightarrow \min_{\{p_{ij}\}}, \ \sum_{j=1}^{n} x_{ij} = 1 \ \ x_{ij} \geq 0$$

де $P_{over\ ij} = \dfrac{c_{0i} x_{ij} \lambda_i R_i^{\alpha_i}}{\alpha_i(\alpha_i - 1)(C - x_{ij}\lambda_i RM\tau_i)} h^{-\alpha_i + 1}$

$P_{over\ ij}$ – розрахункова верхня границя переповненя черги і-го класу трафіку у j-му вихідному каналі (значеня параметрів c_{0i}, R_i, α_i, C_{ij}, $M\tau_i$ були описані раніше); a_j – параметри трафіку задаються адміністратором мережі.

Задача пошуку x_{ij}, що мінімізують цільову функцію (ймовірність переповнення черг), розв‘язується класичними методами теорії оптимізації.

Розв‘язок даної оптимізаційної задачі вирішує проблему вибору каналу PW в якому буде організовано віртуальне з‘єднання на час передачі інформаційної заявки.

Аналіз досліджень проведений в області самоподібних трафіків, показав, що є потреба розробити механізм комплексного керування інформаційними потоками, який буде передбачати можливість сплесків вхідного навантаження, для забезпечення адекватної реакції системи. Наведена оптимізаційна задача ляже в основу методу комплексного керування.

Розв‘язок даної оптимізаційної задачі вирішує проблему вибору каналу PW, в якому буде організовано віртуальне з‘єднання на час передачі інформаційної заявки.

Результат розв‘язку задачі вузлової маршрутизації схематично зображений на рис. 2.8 а) та б), де O – момент виникнення пікового навантаження, P_q, P_i – максимально допустимі втрати пакетів, k – номер вихідного каналу. На рис. 2.7 а) показаний характер залежності втрат трафіку q (трафік для якого зафіксовано перевантаження) в каналі j від часу. На рис. 2.8 б) – характер залежності втрат пакетів трафіків i, таких що ($i \neq q, k = \overline{1,N}$) and ($i = q, k = \overline{1,\ j-1,\ j+1,N}$)

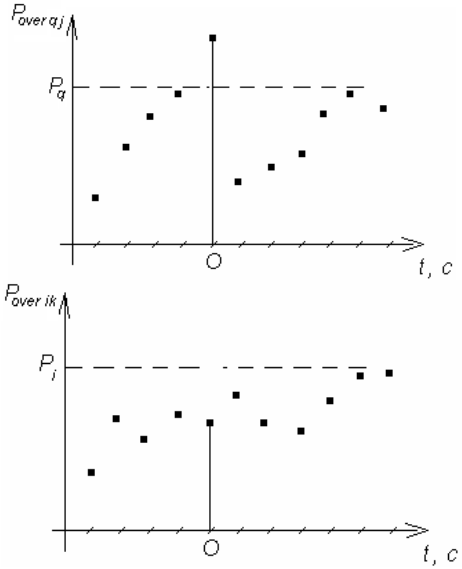


Рис. 2.8 Характер залежності втрат пакетів від часу

Після сигналізації про перевантаження здійснюється перерозподіл інформаційних потоків між вихідними каналами зв‘язку відповідно до задачі вузлової маршрутизації. Це спричиняє зменшення втрат пакетів трафіку, для якого

було виявлено перевантаження, за рахунок незначного збільшення втрат пакетів решти трафіків в вихідних каналах зв‘язку.

2.3. Удосконалення механізму зваженого кругового обслуговування черг

Мета забезпечення якості обслуговування абонентів у мережах мобільних операторів зв‘язку – надання гарантованих та диференційованих послуг у масштабах мережі одного оператора. Одним з перспективних протоколів транспорту інформаційних потоків є протокол PWE3, що імітує роботу технологій ATM, Fast Ethernet, Frame Relay, тощо, для передачі інформаційних потоків через мережу з комутацією пакетів (для даної роботи вибрана імітація технології ATM). Відповідно, для забезпечення якісного обслуговування абонентів застосовуються такі ж принципи, як у відповідних технологіях. Диференціація послуг обумовлює наявність декількох рівнів забезпечення якості обслуговування, кожному з яких відповідає власна архітектурна модель забезпечення функцій QoS.

Отже, далі буде проведений математичний аналіз систем, що надають такі послуги абонентам, і відслідковано взаємовплив між трафіками, яким присвоюються різні категорії якості обслуговування.

2.3.1. Методи обробки черг

Оптимізація керування інформаційними потоками проводиться в пограничному комутаційному центрі (КЦ), який організує зв‘язок базових станцій (БС) з контролером базових станцій (КБС). Канали які зв‘язують БС з КЦ це виділені лінії (радіо або стаціонарні канали). Канали від КЦ до КБС організовані на основі мережі з комутацією пакетів за допомогою протоколу PWE3 засобами емуляції постійного з‘єднання.

В першій главі були розглянуті етапи роботи КЦ та запропонована модель комутації повідомлень, яка основана на класифікації трафіків та оптимізації розподілу інформаційних потоків між вихідними каналами так, щоб вихідні черги рівномірно заповнювалися. Наступним етапом роботи КЦ є обробка черг повідомлень у вихідних каналах. Основна мета керування чергами повідомлень – забезпечення потрібного рівня якості обслуговування абонентів.

Забезпечення якісного обслуговування абонентів – одна з найважливіших задач, що ставляться перед операторами зв‘язку. Перш за все це пов‘язано з конкуренцією на ринку телекомунікацій. Оператор, який не в стані забезпечити заявлений рівень обслуговування одразу програє своїм конкурентам. Примхливий абонент без вагань перейде до іншого оператора, послуги якого більш якісні, і кількість таких абонентів буде зростати, якщо якість послуг буде низькою. Відповідно оператор втрачає прибутки.

На сьогоднішній день розроблено ряд стратегій, які дозволяють організувати контроль за рівнем QoS (Quality of Service). Серед них забезпечення диференційованих послуг (diffserv), інтегрованих послуг (intserv), резервування каналів зв‘язку (RSVP) та алгоритми раннього виявлення перевантажень системи (RED, WRED). Ці механізми пов‘язані з організацією обробки черг повідомлень.

Що передбачає організація обробки черг? Перш за все те, яким чином повідомлення формуються у черги. Розрізняють різні порядки:

- Нормальний метод обробки (базовий) – це FIFO (First In = First Out) – першим надійшов – першим обслужений.
- Додаткові методи включають:

- черги з пріоритетами;
- черги абонентів;
- зважене рівномірне обслуговування.

- Черги з приортетами надають можливість здійснювати приорітизацію основану на:

- мережевому протоколі;
- вхідному інтерфейсі;
- розмірі пакета;
- джерела;
- відстані.

- Обслуговування черг дозволяє розподіляти смугу пропускання вихідного каналу між прикладними програмами.

Абсолютно оптимального розподілу вихідного каналу з забезпеченням всім класам трафіку передачі без втрат та затримок не існує. Але мета апарату керування інформаційними потоками оптимізувати роботу системи таким чином, щоб втрати та затримки були мінімальними. В даній роботі за критерій оптимізації вибрано ймовірність переповнення черг, мінімізація якої зменшить втрати заявок, що пов‘язані з переповненням буферів вихідних каналів комутаційних центрів, для забезпечення заданого рівня QoS для кожної групи абонентів.

2.3.2. Забезпечення якісного обслуговування абонентів

З огляду на обмеженість мережевих ресурсів (буферного простору на вузлах мережі, пропускної здатності трактів передачі, розрахункових потужностей та часу на прийняття керівних рішень), мультисервісна телекомунікаційна мережа повинна мати ефективні засоби забезпечення QoS. Термін «якість послуг» в рекомендації Е.800 МСЕ розуміється як деяка інтегральна оцінка, що визначає ступінь задоволення користувача послугою зв‘язку, яка йому надається [43]. У мережах NGN рішення задач QoS здійснюється на кожному з трьох рівнів у підсистемах адміністративного керування, керування технічною експлуатацією та динамічного керування [41]. Ефективність телекомунікаційних мереж, побудованих на принципах NGN, багато в чому залежить від коректності розв‘язку задач QoS [44].

Якісне обслуговування забезпечується різними способами. Мета забезпечення якості обслуговування абонентів у ТКМ – надання гарантованих та диференційованих послуг у масштабах мережі одного оператора. В рамках даної роботи розглядається IP-мережа. Диференціація послуг обумовлює наявність декількох рівнів якості обслуговування, кожному з яких відповідає власна архітектурна модель забезпечення функцій QoS.

Типове рішення задач QoS охоплює наступні області [44]:

- класифікація прикладних програм з призначенням пріоритетів та диференціюванням трафіку;
- профілювання інформаційного трафіку;
- обмеження інформаційних потоків, що надходять від користувачів;
- керування чергами із встановленням послідовності обробки пакетів на мережевих вузлах;
- маршрутизація мережевого трафіку.

Класифікувати основні послуги, що надаються в мультисервісних мережах в рамках забезпечення QoS, можна по трьох основних характеристиках [45]:

- чутливість до величини пропускної здатності, що надається;
- чутливість до затримок;
- чутливість до втрат.

Кількісно ступінь чутливості програмних програм оцінюється по відповідних показниках якості обслуговування інформаційних трафіків. До основних показників QoS відносяться показники продуктивності (швидкісні варіанти), показники часової прозорості (часові показники), показники семантичної прозорості (показники надійності та достовірності).

До основних показників продуктивності мережі відносять ефективну, пікову, стійку та мінімальну швидкості передачі, що вимірюється, як правило, в біт/с. Мінімальне значення продуктивності, зазвичай, гарантується поставщиком послуг, який, в свою чергу, повинен мати гарантії від мережевого провайдера. Параметри, пов‘язані з ефективною швидкістю передачі можуть бути визначені через дескриптор трафіку IP-мережі, що описаний в рекомендаціях МСЕ-Т.1221.

Другий важливий критерій класифікації прикладних програм за типом трафіку – їх чутливість до затримок пакетів. В рекомендації МСЕ-Т.1540 чутливість до втрат є основним параметром, що характеризує доставку пакетів IP-мереж.

2.3.3. Задача диференційованого обслуговування абонентів

При розв‘язку задачі оцінки загальної затримки повідомлень у мережі, слід зазначити, що введення груп обслуговування абонентів, що передбачає послуга diffserv, вносить деякі корективи у політику обробки потоків заявок. А саме: відповідно до політики надання послуг абонентам мережі можуть надаватися ті самі послуги з різною якістю. Тобто, за умовами договору на обслуговування, оператор для певної послуги гарантує деяку швидкість передачі, деякі показники надійності. Нехай оператор пропонує g пакетів послуг. Відповідно абоненти через базові станції та КЦ надсилають до контролера базових станцій s типів інформаційних потоків.

Отже, за такої постановки задачі, можна говорити про утворення більшої кількості груп навантаження. Нехай R – множина груп навантаження, тоді потужність множини R складає $\| R \| = gS$, тобто передача кожного типу трафіку для однієї з груп обслуговування абонентів утворює групу навантаження, обслуговування якої ведеться за заданими правилами.

Ймовірності переповнення черг, відповідно, можуть бути розраховані для кожної з груп навантаження.

Інтенсивності потоків заявок на передачу інформації для розрахунку показників якості визначаються для кожної групи навантаження окремо та залежать від кількості активних абонентів, закріплених за базовою станцію у заданий час.

$$\lambda_{ij} = f(A_{ij}),$$

де λ_i– інтенсивність надходження заявок на передачу інформації *i*-го типу трафіку на *j*-ту базову станцію, A_{ij} – величина, що знаходиться статистично та позначає середню кількість активних абонентів в *j*-й БС, які роблять заявки на передачу потоків *i*-ї типу.

Середній об‘єм повідомлень береться із статистичних даних окремо для кожної групи навантаження.

Швидкість передачі інформації в одному каналі є постійною величиною, оскільки визначається вихідною смугою пропускання. Однак розділ смуги пропускання між різними групами навантаження (чергами) може бути різним та визначається способом обробки черг. Алгоритми резервування черг були розроблені для надання гарантій в обслуговуванні. Вся ємність каналу, яку може забезпечити емульоване постійне з‘єднання, розділяється відповідно до вимог трафіків. Резервування ємності здійснюється на замовленням. Потік, що зарезервував ємність ставиться в окрему чергу, якщо черга порожня, тоді зарезервований частотний ресурс розділяється між рештою черг. Передача відео трафіку вимагає індивідуального резервування смуги частот, для передачі голосового трафіку достатньо групового резервування. Ємність каналу, яку необхідно зарезервувати, обумовлюється швидкістю обробки інформації.

2.3.4. Алгоритми обробки черг

Якщо в системі ведеться обслуговування заявок не лише за типами інформаційних потоків, а з урахуванням групи абонентського договору, тоді інформаційні потоки розподіляються фактично для R типів черг. Необхідно також підкреслити, що для забезпечення належного рівня якості обслуговування, комутація здійснюється не по-пакетно, а на рівні повідомлень. Це має право на існування, оскільки протокол PWE3 імітує сервіс ATM.

Таким чином, після того як на вихідних каналах були утворені черги повідомлень інформаційних потоків, постає проблема: який з існуючих способів обробки черг повідомлень використовувати для передачі інформації.

У випадках, коли мережа близька до стану перевантаження, для рішення задач QoS для окремого потоку інформаційного трафіку, використовуються засоби керування чергами. Умовно їх можна поділити на алгоритми обслуговування черг та алгоритми превентивного обмеження черг. На практиці найбільше поширення отримали наступні алгоритми:

- PQ (Priority Queuing) – алгоритм обслуговування черг з абсолютним пріоритетом;
- CBQ (Class-Based Queuing) – алгоритм обслуговування черг на основі класів трафіків, що передбачає виділення всім класам номінальної пропускної здатності;

- WFQ (Weighted Fair Queuing) – зважений алгоритм рівномірного обслуговування черг на основі розрахунку порядкового номеру пакету, який передбачає збільшення або зменшення розміру черги в залежності від рівня пріоритету;

- DWFQ (Flow-Based Distributed WFQ) – розподілений зважений алгоритм рівномірного обслуговування черг на основі потоку;

- CBWFQ (Class-Based Distributed WFQ) – розподілений зважений алгоритм рівномірного обслуговування черг на основі класу, що дозволяє вказати потрібну мінімальну пропускну здатність для кожного класу трафіку;

- WRR (Weighted Round Robin) – зважений алгоритм кругового обслуговування;

- DRR – алгоритм кругового обслуговування з дефіцитом.

Алгоритми обслуговування черг в IP-мережах та ATM-мережах є одним з рішень задач QoS шляхом завчасного встановлення порядку використання канальних та буферних ресурсів мережі, тобто сам процес розподілу має чітко виражений статистичних характер.

Пріоритетне обслуговування черг дозволяє гарантувати надання всієї смуги пропускання трафіку, необхідному для рішення критично важливих задач, при ігноруванні решти трафіків. Для використання замовленого обслуговування черг гарантується надання визначеної смуги пропускання трафіку, необхідному для виконання критично важливих задач і в той самий час враховуючи решту трафіків.

2.3.5. Зважений механізм кругового обслуговування (WRR)

Принцип роботи *механізму зваженого кругового обслуговування* черг полягає у способі вибора кількості бітів інформації різних типів трафіків, що забираються в у тракт каналу зв‘язку за один такт передачі.

Зважений механізм кругового обслуговування (Weighted Round Robin – WRR) являє собою розширення планувальника кругового обслуговування, відповідно до якого кожному потоку трафіку призначається своя вага. Алгоритм WRR обробляє потік трафіку пропорційно до його ваги. Найбільш органічно WRR-планувальник працює з механізмом комутації ATM (Asynchronous Transfer Mode – режим асинхронної передачі), відповідно до якого пакет представляється у вигляді комірок, а алгоритм WRR використовується для обробки черг, що складаються з комірок, що несуть інформацію. Кожній черзі виділяється частина смуги пропускання інтерфейсу у відповідності до ваги потоку трафіку, не залежно від розміру пакету. Таким чином, моделювання алгоритму WRR в рамках даної роботи є доцільним.

Зважений алгоритм кругового обслуговування черг передбачає призначення певному класу трафіку своєї ваги. Вага є параметром, на основі якого розраховується кількість пакетів, що буде забрана з черги *i*-го типу трафіку, що буде передана у вихідний канал з один такт передачі.

Ефективна ширина смуги пропускання черги прямо пропорційна її вазі та розраховується за наступною формулою:

ефективна ширина смуги = (вага черги $\times$ *ширина смуги пропускання інтерфейсу)* $\div$ *сума ваг всіх активних черг.*

Для оцінки верхньої границі ймовірності переповнення черг трафіку і-го типу в j-му вихіному каналі при порядку обробки черг за принципом WRR треба виконати наступні кроки:

1. Задана загальна швидкість обслуговування інформації в j-му каналі C_j.

2. Для кожного типу трафіку задана вага w_i (натуральні числа), що визначає смугу пропускання:

$$\frac{w_{D\,j} v_j}{w_{D\,j} + w_{v\,j}} = C_{D\,j} \tag{2.27}$$

$$\frac{w_{V\,j} v_j}{w_{D\,j} + w_{v\,j}} = C_{V\,j}. \tag{2.28}$$

3. З огляду на те, що трафік інтерактивного відео Інтернет характеризуються самоподібними моделями, розрахунок верхньої границі переповнення черг розраховується за формулою

$$P_{over\ ij} \leq \frac{c_{0i} R_i \lambda_{ij}}{\alpha_i(\alpha_i - 1)(C_j - R_i \lambda_{ij} M\tau_i)} h_{ij}^{-\alpha_i + 1}, \tag{2.29}$$

де λ_{ij} – інтенсивність і-го типу трафіку, що направляється в j-й вихідний канал, h_{ij} – довжина черги і-го типу трафіку в j-му вихідному каналі КЦ РЕ.

Параметр c_{0i} – це кількість трансакцій, протягом яких може бути передано мінімальне повідомлення і-го типу трафіку. Параметр потоку α_i розраховується з відомостей про середнє значення кількості трансакцій протягом яких передаються повідомленні і-го типу трафіку ($M\tau_i$). Тобто α_i знаходиться з рівняння: $M\tau_i = c_{0i} \sum_{l=1}^{\infty} l^{-\alpha_i}$.

Таким чином, можна розраховувати верхню межу ймовірності переповнення черг вихідних каналів, де обслуговування інформаційних потоків здійснюється за принципом WRR.

2.3.6. Вдосконалений алгоритм WRR

Задача оптимізації передачі інформаційних потоків є особливо актуальною. Аналіз статистики комутаційних центрів мобільних операторів та характеру зміни інтенсивності вхідних трафіків наштовхнув на думку реалізації такого порядку обробки черг заявок, коли в моменти різкого зростання кількості пакетів, пакети трафіку інтерактивного відео перенаправляються в чергу низькопріоритетного трафіку.

Кількість черг, що підтримує обладнання комутаційних центрів є обмеженим, тому при передачі великих об'ємів трафіків, та обробці великої кількості заявок не завжди є можливість організувати окрему чергу для заявки інтерактивного відео, тому слід розглядати обробку інтерактивного відео в одній черзі.

Вимоги до якості обслуговування абонентів диктують збільшення ймовірності обслуговування заявок кожного виду трафіку та зменшення часу затримок трафіків при проходженні через КЦ, який є пограничним між базовими станціями та мережею з комутацією пакетів, де організовані РЕ, що ведуть до РЕ2.

В розділі 2 були наведені математичні моделі для оцінки верхньої границі ймовірності переповнення черги мультимедійного трафіку та трафіку Інтернет. Задача цього розділу показати яким чином міграція заявок між чергами впливає на якість обслуговування абонентів.

Оскільки моделюється робота АТМ-мережі, то довжина черги визначається кількістю АТМ-пакетів рівної довжини, які поміщаються у пам‘яті буферу вихідного каналу. Як показали дослідження, характер самоподібного трафіку є бурхливим, тобто присутні короткочасні сплески інтенсивності (см. [47]).

Тому може бути ефективним перенаправлення комірок відео трафіку на початок черги Інтернет трафіку. Фактично це може бути представлено, як перерозподіл кількості пакетів, що забираються з черг відео трафіку та Інтернет трафіку, за одну транзакцію, в момент виникнення переповнення черги відео трафіку. Тобто, зменшення кількості пакетів низькопріоритетного трафіку, що передаються за одну трансакцію, на кількість пакетів, що надійшли з черги високопріоритетного потоку.

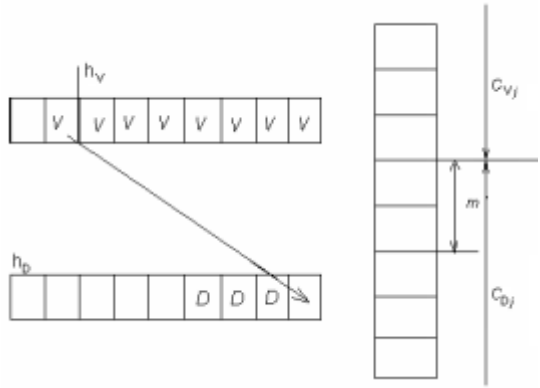


Рис. 2.9 Вдосконалений алгоритм WRR

На рис. 2.9 показано основний принцип розподілу канального ресурсу за вдосконаленим алгоритмом WRR. При наповненні черги відео трафіку заданим числом заявок, наприклад h_V, всі пакети даних, що надходять, поступають на перші позиції в чергу Інтеренет трафіку, до того ж встановлюється ліміт на кількість пакетів, що можуть бути перенесені між чергами, деяке $m \le C_{Dj}$, де C_{Dj} –кількість пакетів, що забираються – черги Інтернет трафіку за одну трансакцію, відповідно до алгоритму WRR, або кількість АТМ-пакетів низькопріоритетного трафіку, що можуть бути передані за одиницю часу. Така міграція пакетів між чергами рівносильна тому, що черга буде збільшена на (h_V +k) позицій і смуга пропускання відео трафіку буде збільшена до (C_{Vj}+k) пакетів за одиницю часу, де $k = \overline{1,m}$ залежить від кількості пакетів, які надійшли зверх черги h_V.

Розрахунок ймовірності переповнення черги, яка фактично оцінює ймовірність втрат заявок заданого типу, є нетривіальною задачею. Оскільки необхідно розглядати відео и Інтернет трафік одночасно, кількість пакетів, що забираються за одну трансакцію з черги Інтернет трафіку є функцією від кількості пакетів відео трафіку, що надійшли в систему, аналогічно швидкість обробки заявок відео трафіку залежить від кількості заявок в його черзі. Розрахунок ймовірностей двох сильно залежних випадкових подій (подія переповнення буферу трафіку, подія переповнення буферу мультимедійного трафіку) є складною математичною задачею. Тому для оцінки роботи системи було прийнято рішення розв'язати задачу пошуку ймовірності переповнення черги відео трафіку та верхньої межі ймовірності переповнення черги трафіку даних кожного *k-го* випадку $(k = \overline{1,m})$ за формулами (2.29), та розрахунку смуг пропускання C_{Vj} і C_{Dj} за формулами (2.27) і (2.28). В результаті отримуємо послідовність виду

$$\{P_{overV}(C_{Vj} - k, h_V), P_{overD}(C_{Dj} + k, h_D + k)\}, k = \overline{0,m}.$$

Розрахунок величини навантаження m, що задовільняють вимогам до якості передачі інформаційних трафіків наводиться у наступному пункті даного розділу.

2.3.7. Пошук оптимальної величини навантаження, що перерозподіляється між чергами високопріоритетного та низькопріоритетного трафіків

Для роботи запропонованого вдосконаленого алгоритму кругового обслуговування черг необхідно розрахувати m – кількість АТМ пакетів високопріоритетного трафіку, що переміщаються на початок черги низькопріоритетного трафіку, що дозволить суттєво покращити якість передачі високопріоритетного трафіку при збереженні допустимої якості передачі низькопріоритетного трафіку. Для цього необхідно виконати наступні кроки:

- для смуги пропускання (кількості АТМ пакетів, що забирається за одну транзакцію), яка виділена для передачі високопріоритетного відео трафіку та низькопріоритетного трафіку даних, розрахувати довжину черги, що задовольняє вимогам до часу затримки пакетів у комутаційному центрі. Тобто знайти таке максимальне h, щоб виконувалася умова $\frac{h}{C} \leq T$, де C – швидкість передачі каналу зв'язку (кількість АТМ-пакетів, що забираються з відповідної черги за одну трансакцію), T – допустимий час затримки розділений на довжину інтервалу часу, через який здійснюється відправлення інформації в канал зв'язку;
- розрахувати матрицю значень ймовірностей (верхньої межі) переповнення черг, де кількість пакетів, що забирається з черги відповідного трафіку змінюється на один АТМ-пакет, в результаті при незмінній сумарній кількості пакетів всіх типів трафіків, що відправляються в вихідний PW, кількість пакетів, що забирається з черги високо пріоритетного відео трафіку збільшується з C_V до (C_V+C_D-1), а кількість пакетів, що забираються з черги Інтернет трафіку зменшується відповідно від C_D АТМ-пакетів, до 1 АТМ-пакета:

$$M = \{P_{kj}\} = \begin{array}{|c|c|}\hline P_{overV}(C_V + 0\ ATM\ пакетів) & P_{overD}(C_D - 0\ ATM\ пакетів) \\ \hline P_{overV}(C_V + 1\ ATM\ пакет) & P_{overD}(C_D - 1\ ATM\ пакет) \\ \hline \dots & \dots \\ \hline P_{overV}(C_V + k\ ATM\ пакетів) & P_{overD}(C_D - k\ ATM\ пакетів) \\ \hline \dots & \dots \\ \hline P_{overV}(C_V + C_D - ATM\ пакет) & P_{overD}(ATM\ пакет) \\ \hline \end{array} =$$

$$= \left\{P^k_{overV}(C_V + k\ ATM \text{ пакетів}), P^k_{overD}(C_D - k\ ATM \text{ пакетів})\right\}$$

- обирається таке максимальне i, для якого виконуються умови $\sum_{j=1}^{2} P_{ij} < \sum_{j=1}^{2} P_{0\,j}$, $P_{k1} \le P_1$, $P_{k2} \le P_2$, де P_1 та P_2 - порогові значення для ймовірності переповненні високопріоритетного трафіку та низькопріоритетного трафіку;
- присвоюється m = i АТМ пакетів.

Відповідь: навантаження (m), яке перерозподіляється, береться у розмірі i АТМ пакетів.

2.4. Принцип керування інформаційними потоками в комутаційному центрі PWE3

2.4.1. Принцип комплексного керування інформаційними потоками в комутаційному центрі PE

Для здійснення керування інформаційними потоками в комутаційному центрі PE необхідно мати інформацію про параметри системи, такі як:

- кількість базових станцій N, що обслуговуються комутаційним центром, відповідно до протоколу PWE3, кількість вихідних тунелів PW відповідно також дорівнює N;
- інтенсивність надходження заявок і-го типу трафіку від j-ї БС (кількість заявок одиницю часу);
- середня тривалість обслуговування одного повідомлення і-го типу трафіку;
- швидкість передачі джерела інформації;
- швидкість передачі інформації в j-му каналі PW;
- довжина черги і-го типу трафіку в j-му каналі PW;
- повинні бути задані порогові значення ймовірностей втрат всіх типів трафіків, обслуговування яких здійснюється.

Також необхідно задати інтервал дискретизації часу, тобто через який час буде передаватися інформація, розрахувати значення всіх параметрів системи у одиницях виміру часу.

Змінні у системі – це множина долей x_{ij} інформаційного потоку і-го типу, що направляється в j-й вихідний канал.

Механізм керування інформаційними потоками передбачає розрахунок доль інформаційного потоку, що направляється в вихідні канали щоразу, коли ймовірність втрат принаймні у одному з каналів перевищує заданий поріг.

I. При роботі комутаційного центру РЕ ведеться моніторинг за якістю обслуговування інформаційних потоків. Механізм запускається в роботу при сигналізації про перевантаження однієї з черг вихідних PW каналів (рис. 2.10).

Рис. 2.10 Принцип комплексного керування інформаційними потоками. Крок I.

II. Якщо реальні втрати пакетів хоча б в одному з вихідних каналів перевищують заданий поріг, це означає, що покращеного алгоритму кругового обслуговування не достатньо для забезпечення якості обслуговування, тоді включається в роботу програма оптимізації розподілу інформаційних потоків (Рис. 2.10). Розв'язується задача мінімізації цільової функції , $W_{over} = \sum_{i=1}^{s} a_i\left(\sum_{j=1}^{n} x_{ij} P_{over\ ij}\right) \Rightarrow \min_{\{p_{ij}\}}$,

де $P_{over\ ij}$ – верхня межа ймовірності переповнення черги і-го типу трафіку у j-му каналі розраховується за формулою

$$P_{over\ ij} \le \frac{c_{0i}\lambda_{ij}R_i}{\alpha_i(\alpha_i-1)(C_j-\lambda_{ij}R_iM\tau_i)} h_{ij}^{-\alpha_i+1} \qquad (2.29)$$

x_{ij} – долі сумарного інформаційного потоку і-го типу, що направляються в j-й канал. Оптимізаційна задача має обмеження

$$\sum_{i=1}^{n} x_{ij} = 1,\ \ x_{ij} \ge 0,\ \ j = \overline{1,3}.$$

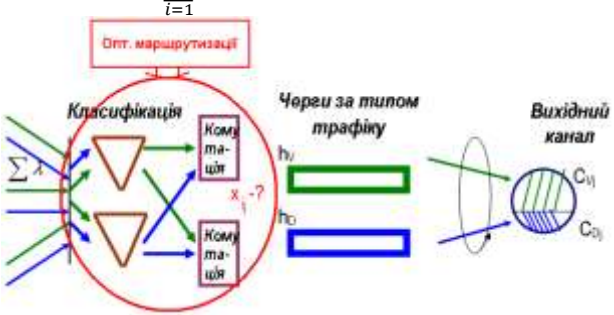


Рис. 2.11 Принцип комплексного керування інформаційними потоками. Крок II.

III. Після того, як була проведена оптимізація, необхідно перерахувати значення m за принципом наведеним в п. 2.4.9 для зміненого значення інтенсивності надходження інформаційного потоку: $\lambda_{ij\,new} = x_{ij}\lambda_i$, де $\lambda_i = \sum_{k=1}^{N} \lambda_{ik}$ – кількість заявок на передачу і-го типу трафіку, що надходять від усіх базових станцій. Після того, як m було розраховано, може бути запущено роботу покращеного алгоритму кругового обслуговування заявок (Рис. 2.12).

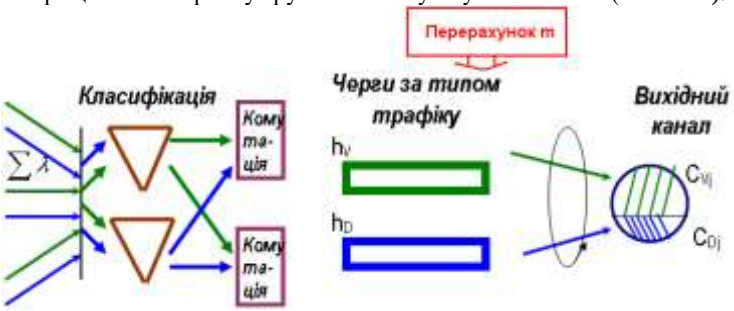


Рис. 2.12 Принцип комплексного керування інформаційними потоками. Крок III.

IV. Мережа працює відповідно до прийнятих керівних рішень. Здійснюється моніторинг мережі (рис. 2.13).

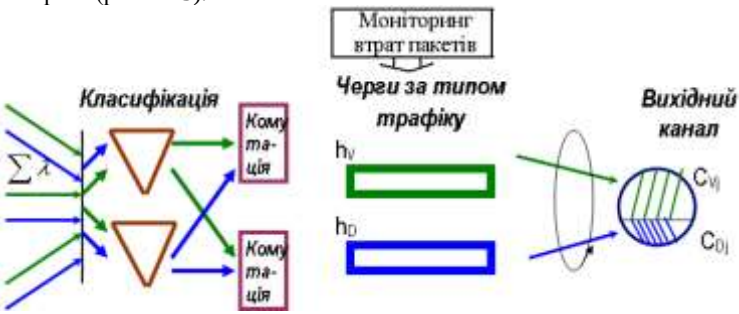


Рис. 2.13 Принцип комплексного керування інформаційними потоками. Крок IV.

2.4.2. Структура багаторівневої системи керування маршрутизацією трафіку та ємністтю каналів його передачі

На сьогоднішній день робота мережі за протоколом PWE3 може бути представлена у вигляді блок-схеми, наведено на рис. 2.14.

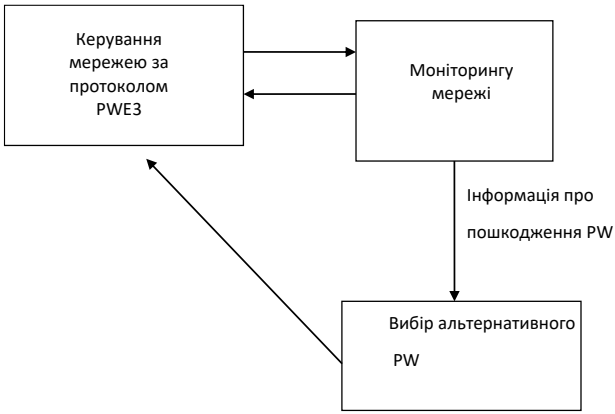


Рис. 2.14 Система керування інформаційними потоками PWE3

Система керування інформаційними потоками за протоколом PWE3 включає в себе:

Блок керування мережею за протоколом PWE3, що забезпечує імітацію з‘єднань між комутаційними центрами PE1 та PE2, тобто організацію PW для кожної БС, що обслуговується PE1.

Блок моніторингу мережі, що збирає відслідковує інформацію про якість обслуговування абонентів, про зв‘язність мережі, тощо. Якщо блок моніторингу мережі виявляє пошкодження ліній PW, то інформація про це передається в блок «Вибір альтернативного з‘єднання PW». Інформація про вибраний шлях PW передається в блок керування мережею за протоколом PWE3.

Отже спрощено роботу системи забезпечення якості комутаційного центру PE можна охарактеризувати як таку, що забезпечує організацію зимітованих з‘єднань PW, розподіл канального ресурсу в рамках одного PW здійснюється за заданим принципом обслуговування інформаційних потоків різних типів.

При введенні модифікації багаторівневої системи керування інформаційними потоками та розподілом канальних ресурсів з’являється можливість розподіляти інформаційний потік від всіх БС між вихідними PW відповідно до інформації про втрати пакетів в каналах зв’язку, а також динамічно розподіляти ємність каналу PW між інформаційними потоками різних типів, за в виділення *m-позицій* у блоці з C_j ATM-пакетів, що відправляються для передачі у вихідний канал, які, в разі необхідності, надаються високопріоритетному трафіку або низькопріоритетному трафіку, залежно від зкорегованої інтенсивності вхідних потоків.

При введенні в дію запропонованого механізму комплексного керування інформаційними потоками блок-схема системи керування зміниться (рис. 2.15)

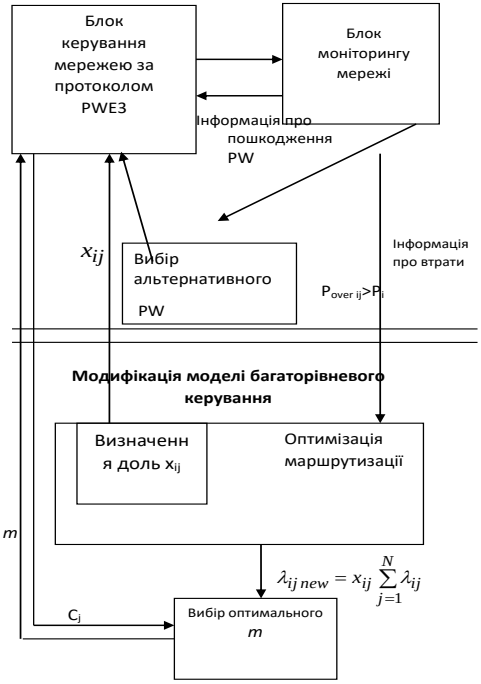


Рис. 2.15 Блок-схема системи керування інформаційними потоками PWE3 з модифікацією моделі керування

З модифікацією моделі багаторівневого керування на блок-схемі з'явилися блоки Оптимізація маршрутизації, що включає в себе блок Визначення долей x_{ij}, що можуть бути знайдені при розв'язанні оптимізаційної задачі

$$W_{over} = \sum_{i=1}^{s} a_i \left(\sum_{j=1}^{n} x_{ij} P_{over\ ij} \right) \Rightarrow \min_{\{p_{ij}\}} \text{ за умови } \sum_{j=1}^{n} x_{ij} = 1.$$

Коли оптимізаційна задача розв'язана, інформація про значення доль трафіків різних типів, які можуть бути направлені в вихідні PW тунелі, направляється в блок керування мережею за протоколом PWE3.

Ще один блок, який додається при модифікації системи керування, – Вибір оптимального значення m, який на основі отриманої інформації про змінені за формулою $\lambda_{ij\ new} = x_{ij} \sum_{j=1}^{N} \lambda_{ij}$ інтенсивності надходження заявок та інформації про ємності вихідних каналів PW, розраховує оптимальне значення величини позицій m.

Таким чином, додаткові блоки системи керування інформаційними потоками та канальними ресурсами впливають на розподіл трафіків між каналами PW і на розподіл смуги пропускання в середині PW.

2.4.3. Багаторівнева система. Алгоритм керування

На рис. 2.16 наведено алгоритм керування багаторівневою системою керування з запропонованою в роботі модифікацією.

Основними блоками алгоритму є «Керування за протоколом PWE3», «Оптимізація маршрутизації», «Розрахунок оптимального m». В рамках «Керування за протоколом PWE3» здійснюється моніторинг за показниками втрат інформаційних пакетів. Якщо виконується умова $P_{over\ ij}>P_i$, де $P_{over\ ij}$ – статистично визначений показник втрат пакетів і-го типу в j-му каналі зв'язку, P_i – задані максимальні значення втрат пакетів і-го типу трафіку, тоді за алгоритмом здійснюється оптимізація маршрутизації, що включає в себе розрахунок доль x_{ij}, що показує, яка частина сумарного інформаційного потоку і-го типу трафіку була направлена в j-й канал зв'язку.

Значення x_{ij} передаються в блок керування за протоколом PWE3 для здійснення оптимального розподілу інформаційних трафіків між вихідними PW каналами, а також передаються в блок для оптимізації розподілу кількості ATM пакетів різних типів трафіків, що будуть передаватися за одну трансакцію в канал PW. Для динамічного розподілу ємності каналу за вдосконаленим механізмом кругового обслуговування розраховується також значення величини m.

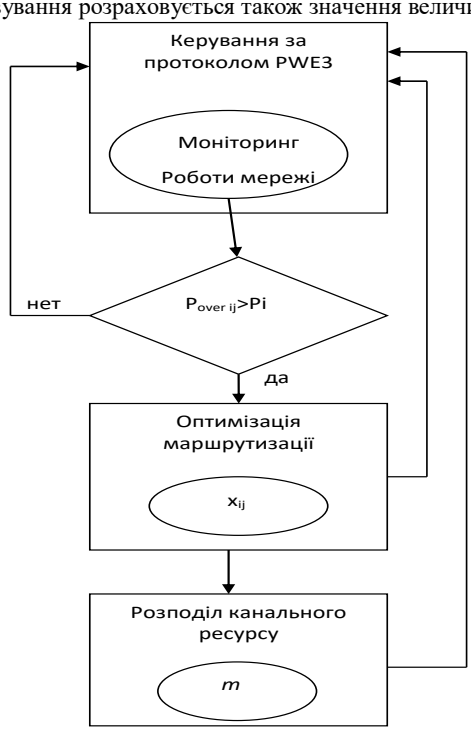


Рис. 2.16 Алгоритм багаторівневої вдосконаленої системи керування за протоколом PWE3

Висновки

1. Досліджені основні категорії АТМ трафіку, що передається в мережах мобільного оператора та потребує моделювання при розв‘язку задачі керування інформаційними потоками в комутаційному центрі РЕ, що дозволило вибрати дві найпоширеніші категорії (Rt-vbr та nrt-VBR) для моделювання. Rt-vbr описує трафік інтерактивного відео, має особливі вимоги до передачі, які необхідно враховувати при організації системи керування потоками в комутаційному центрі.

2. Підібрано математичну модель для інтерактивного відео та Інтернет трафіку. Вона характеризується як самоподібний трафік, що може бути описаний довготривалими залежностями між вхідними пакетами, тобто одне джерело передає низку пакетів, довжина таких сеансів передачі розподілена за законом Парето, а моменти стартів сеансів передачі від абонентів мають експоненційний розподіл. Підкреслено, що різниця між трафіком інтерактивного відео та Інтернет трафіком полягає у вимогах до якості передачі, однак порядок надходження та обслуговування потоків однаковий.

3. Запропоновано спосіб оцінки якості обслуговування інформаційних трафіків, що прямують між двома комутаційними центрами РЕ1 та РЕ2 через мережу з комутацією пакетів, та дисципліну їх обслуговування, яка за рахунок передачі інформаційних потоків не по одному емульованому з’єднанню PW, а по всіх, що організовані між РЕ1 та РЕ2, дозволяє рівномірно завантажувати черги вихідних каналів комутаційного центру РЕ1.

4. Розроблена постановка оптимізаційної задачі розподілу інформаційних потоків між вихідними каналами комутаційного центру РЕ1 для одного класу трафіку, що лягла в основу мультисервісної задачі оптимізації. За критерій оптимізації вибрано ймовірність переповнення буферу.

5. Доведена можливість використання запропонованого способу оцінки якості обслуговування інформаційних потоків, це дозволило вдосконалити процес керування роботою комутаційного центру РЕ.

6. Запропонована дисципліна обслуговування вхідного мультисервісного інформаційного потоку вихідними каналами комутаційного центру РЕ, що забезпечує перерозподіл інформаційних потоків у разі переповнення черг вихідних каналів, за рахунок розв‘язку задачі мінімізації втрат пакетів для одного з типів трафіків, які обслуговуються.

7. Запропоновано рішення задачі пошуку максимальної ємності каналу передачі низькопріоритетного трафіку, яка тимчасово виділяється для передачі високопріоритетного трафіку, що дозволило підвищити ефективність передачі за рахунок зниження рівня втрат пакетів, які пов’язані з перевищенням допустимого часу очікування обслуговування, оскільки довжина черги розрахована таким чином, щоб всі заявки з черги могли бути вчасно обслужені виділеним каналом.

8. Запропонована блок-схема та алгоритм багаторівневої системи керування маршрутизацією трафіку і пропускною здатністю трактів передачі інформації, що дозволило формалізувати отримані результати.

9. На основі досліджених принципів диференційованого обслуговування абонентів для мережі мобільних операторів зв’язку, які працюють за протоколом PWE3, зроблено висновок про необхідність модернізації існуючих способів обробки черг з метою забезпечення необхідного рівня якості обслуговування мультисервісного трафіку.

10. Розроблено модель оцінки втрат пакетів мультисервісного потоку в пограничному комутаційному центрі, який організує зв‘язок базових станцій з контролером базових станцій, при застосуванні принципу зваженого кругового обслуговування черг (WRR) з урахуванням самоподібної природи вхідних трафіків, що дозволило оцінити якість роботи даного принципу обробки черг та виявити його недоліки.

11. На основі зваженого принципу кругового обслуговування черг (WRR) був запропонований вдосконалений алгоритм WRR обслуговування абонентських заявок, основна ідея якого полягає у перенесенні деякої кількості пакетів з черги високопріоритетного трафіку на початок черги низькопріоритетного трафіку, які могли б загубитися внаслідок переповнення відповідної черги. Це дозволило підвищити ефективність передачі високопріоритетного трафіку із збереженням рівня якості передачі низькопріоритетного трафіку.

3. ЗАБЕЗПЕЧЕННЯ ПРОЦЕСІВ КЕРУВАННЯ МОБІЛЬНИМИ МЕРЕЖАМИ

3.1. Аналіз надання послуг в сучасних мобільних мережах

3.1.1. Аналіз послуг та їх надання в сучасних мобільних мережах в Україні

Традиційно в процесі надання послуг в мобільних мережах працює ланцюжок: клієнт – оператор зв'язку – провайдер, що надає доступ до конкретного контента. З можливістю повсюдного мобільного доступу в Інтернет, де провайдер (контент-провайдер) є невизначеним суб'єктом, який надає послуги в Інтернеті. У випадку, якщо абоненти вважатимуть їх послуги кращими, ніж послуги операторів, останні перетворяться на «комунікаційну трубу», яка приноситиме дохід лише від кількості переданої інформації, надаючи доступ до сервісів Інтернет-компаній [1-5].

На сьогоднішній день основними провайдерами GSM/UMTS мобільних послуг в Україні є оператори: «Укртелеком», «Український мобільний зв'язок», «КиївСтар», «Астеліт».

Аналіз послуг показує наступне:

- виділяються 7 основних груп послуг;
- групи послуг сформовані згідно потреб, з метою зручного усвідомлення потенційним споживачем, і не враховують в системі класифікації технічні аспекти надання послуг;
- окремо виділена послуга доступу до мережі Інтернет.

На рис. 3.1 показано шість основних видів послуг, які виділяють провідні глобальні європейські оператори в мобільних мережах 3-го покоління UMTS форуму.

Телекомунікаційна послуга – продукт діяльності оператора або провайдера телекомунікацій, направлений на задоволення потреб споживачів у сфері телекомунікації [9].

Послуги зв’язку – продукт діяльності по прийому, обробці та доставки повідомлень електрозв’язку [10].

Контент – продукт, що представляє собою вміст наповнення послуг зв’язку [11]. З терміну контента витікає поняття «веб контент» [9], текстовий, візуальний, аудіо, відео тощо – контент, котрий користувач може отримати в Веб середовищі.

Порівнюючи перелік видів послуг в українських мережах з послугами в Інтернеті, відмітимо, що станом на сьогоднішній день кількість послуг, які користувач може отримати в Інтернеті набагато більша, ніж кількість послуг, що надаються безпосередньо оператором зв’язку.

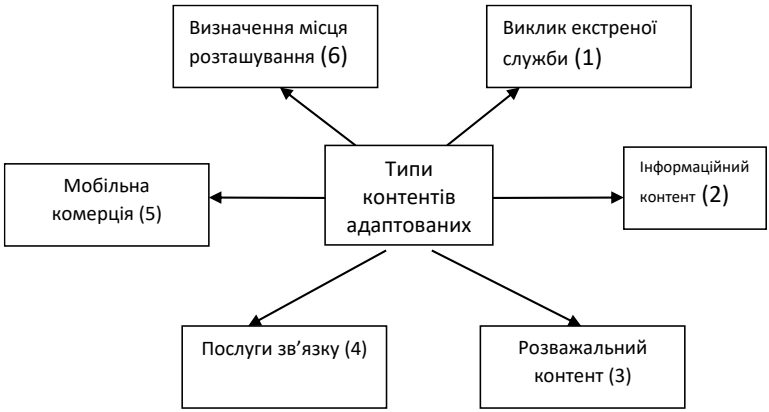


Рис. 3.1 Види послуг в мобільних мережах 3/4-го покоління

Порівнюючи послуги оператора зв'язку та конкурентних послуг Інтернету необхідно відмітити наступні переваги отримання послуг в Інтернет мережі:

- зручність у передачі мультимедійної інформації (ICQ, Skype тощо) завдяки широкому різновиду програмного забезпечення в Інтернеті;
- контроль якості за допомогою використання відповідних кодеків та форматів інформації;
- «безкоштовність» інформації (у деяких випадках).

Існують наступні недоліки використання послуг в Інтернеті:

- Інтернет оператор не гарантує якість отримання послуг в Інтернеті: затримка пакетів, коефіцієнт втрат пакетів тощо. Це суттєво впливає на якість послуг, які надаються в реальному масштабі часу;
- обмежена кількість і, відповідно, порівняно висока вартість термінальних пристроїв, що підтримують програмні засоби Інтернету;
- не всі послуги можуть буди надані в мережі Інтернет, наприклад, – послуга місцезнаходження.

На рис. 3.2 показана тенденція росту пропускної здатності, ціни за Мбіт/с в залежності від технології передачі [6, 7].

Рис. 3.2 Тенденція росту пропускної здатності в залежності від радіо технології

Як видно з рис. 3.2, при переході до нових технологій, швидкість передачі зростає, в той час як ціна за 1 Мбіт/с знижується. Це призводить до неефективної тарифікації послуг за об'ємом або швидкістю передачі, тому необхідно визначити безпосередньо послуги, а також їх характеристики. Класифікація сервісів за категоріями [9], яка включає 6 наступних характеристик контента:

- тип контента: відео, голос, відформатований та невідформатований текст, їх комбінації. В Інтернет середовищі існує необмежена кількість форматів контента, значно більша, ніж в мережі мобільного оператору;
- схема кодування. Вибір схеми кодування залежить від можливостей терміналу, наявності відповідного програмного забезпечення, смуги пропускання щодо доступу до контента, характеристик каналу. У даному випадку в Інтернеті користувач може впливати на кодек або формат, який використовується для надання контента, або перетворювати формат;
- якість обслуговування, яка залежить не лише від сервісів і типів контента, а також від кодека, який використовується. Наприклад, високоякісні аудіо матеріали можуть бути представлені за допомогою кодека Mpeg-1 Layer 3 (Mp3), потокова швидкість 128 kbit/s або Mpeg-2 Advanced Audio Coding, швидкість потоку 96 kbit/s або ще менша. Аналогічно для відео інформації, де характеристиками якості можуть бути глибина кольору, швидкість/частота кадрів, дозвіл, схема стискування. Все це визначатиме швидкість потоку;
- прийнятна затримка очікування послуги. Наприклад, виклик екстреної служби повинен відбуватися фактично негайно, тоді як e-mail аркуш, електронна пошта, «терпить» затримку та час її передачі може бути значно більшим;
- наявність управління правами на контент – DRM (Digital Right Management);
- масштабованість інформації контента.

Порівняльний аналіз категорій контента підтверджує, що:

- по-перше, в Інтернеті, мобільний користувач має доступ до контентів, що апріорі більш різноманітні та мають ширші параметри;
- по-друге, користувач може впливати на адаптивність контента;
- по-третє, поширеність різного програмного забезпечення породжує кількість контента та послуг, що користувач може отримати в мережі.

Слід відмітити, що в Інтернеті та мережах мобільних операторів має місце тенденція появи персоналізованих послуг (інша назва розширених послуг) [12, 13, 15]. Ці послуги враховують характеристики та особливості користувача, такі як параметри термінального обладнання, належність до відповідної привілейованої групи користувачів тощо.

Шляхом аналізу мобільних послуг на ринку України сформульовані основні вимоги, що висувають персоналізовані широкосмугові послуги до мережі мобільного телекомунікаційного оператору:

1.Пропускна здатність у межах: 64k – 20 Мбіт/с;

2.Гарантування якості передачі (постійної швидкості передачі, затримки, джитеру, коефіцієнту втрат пакету, коефіцієнту доступності послуги) ;

3.Адаптація контента в залежності від технічних характеристик термінального обладнання користувача;

4.Гнучкість тарифікації послуг, включаючи контент-послуги, що враховує якість передачі, підписку користувача;

5.Секретність зв'язку, що передбачає шифрування каналу зв'язку тощо

Персоналізовані послуги спрямовані на повне задоволення потреби користувача з урахуванням його технічних можливостей та стану бюджету. Одночасно з цим, оператор зв'язку прагне підвищити ефективність системи обробки викликів.

Відтак, в процесі надання послуги необхідно враховувати технічні параметри послуги й економічні характеристики з метою одночасного задоволення потреб користувача послуг і підвищення ефективності системи обробки викликів оператора.

Доцільним є порівняння можливостей сучасних мереж щодо максимальної відповідності надання персоналізованих послуг реалізації параметрів. В табл. 3.1 надано порівняльну характеристику реалізації персоналізованих послуг у сучасних мобільних мережах [12].

Табл. 3.1

Порівняння технічних особливостей реалізацій персоналізованих послуг у мобільних мережах

Характеристика надання послуг	**2-ге покоління мережі GSM**	**3-тє покоління мережі UMTS**	**4-те покоління мережі LTE/IMS**
Смуга пропускання	До 384 кбіт/с	384 кбіт/с – 2 Мбіт	100Мбіт/с – при мобільному доступі 1 Гбіт при стаціонарному
Технологія комутації	Каналів	Каналів і пакетів	Пакетів
Якість обслуговування пакетних послуг в режимі "end-to-end"	Не підтримується	Підтримується частково	Підтримується
Секретність передачі пакетів зв'язку	Не підтримується	Не підтримується	Підтримується
Багатоадресна передача	Не підтримується	Не підтримується	Підтримується
Тарифікація пакетних сервсів	Підтримується без урахування якості	Підтримується без урахування якості	Підтримується з урахуванням якості
MIMO технології	Не підтримуються	Підтримуються частково	Підтримуються
Технологія доступу	TDMA	CDMA	OFDMA різновиди

Як видно з табл. 3.1, технології четвертого покоління 4G найбільше задовольняють вимогам персоналізованих послуг. Абревіатура 4G використовується для узагальнення переліку стандартів нових безпровідних мереж [12]: LTE/LTE-Adavnced, UMB, WiMAX/802.16e.

У джерелі [13, 14] автори характеризують технологію LTE як найбільш потенціальну для впровадження серед систем 4-го покоління, враховуючи технічні характеристики та зацікавленість операторів мереж і виробників обладнання.

Таким чином, у даному підрозділі наведений аналіз послуг на українському ринку телекомунікаційних мобільних послуг. Окрім того, приведена класифікація послуг в мобільних мережах і наданий порівняльний аналіз телекомунікаційних послуг мережі Інтернет і операторів мобільних мереж. Досліджені основні характеристики найбільш перспективних, персоналізованих послуг: смуга пропускання, якість надання, тарифікація, адаптація контента.

Аналіз надання персоналізованих послуг у мережах GSM, UMTS, довів, що технічні можливості мережі LTE найбільше відповідають вимогам персоналізованих послуг наступного покоління. З'ясовано, що в процесі обслуговування послуги необхідно одночасно враховувати характеристики послуги і системні параметри мережі з метою одночасного задоволення потреб користувача послуг і підвищення ефективності системи обслуговування (обробки) викликів в мережі оператора.

3.1.2. Обробка викликів в системах масового обслуговування

Обслуговування викликів в мобільних мережах є складною задачею, що описується в термінах систем масового обслуговування. Основною проблемою формалізації процесів управління викликами є неможливість записати в явному вигляді цільову функцію оптимізації, що враховує структурні та економічні параметри системи, і як наслідок, неможливість застосувати аналітичні розрахунки [16]. В деяких випадках застосовують математичне модулювання для оптимізації [17].

В СМО з декількома вхідними потоками при заданих структурних параметрах більш вузьким класом задач управління є застосування дисциплін обслуговування. Широке застосування отримали пріоритетні системи обслуговування [18, 19]. Тому розглянемо цей клас задач більш детально.

3.1.2.1. Дисципліни обслуговування в СМО

Дисципліна обслуговування визначається правилами, за якими заявки вибираються для подальшого обслуговування. Існує багато різновидів дисциплін обслуговування, з них основними є [20]:

- прямий порядок обслуговування («перший прийшов – перший вийшов»);
- інверсний («перший прийшов – останній вийшов»);
- випадковий порядок обслуговування ;
- груповий порядок обслуговування;
- циклічний порядок обслуговування;
- порядок обслуговування з пріоритетом;

- інші.

Порядок обслуговування з пріоритетом найбільш поширений в інфо-телекомунікаційних системах, тому в подальшому будемо розглядати даний вид пріоритету. В теорії СМО розглядається два основних пріоритети – абсолютний і відносний. Відмінність між ними полягає в тому, що при відносному пріоритету не можливе переривання обслуговування заявки або виклику, що обслуговується, при надходженні більш пріоритетної заявки і в момент відсутності вільних місць обслуговування [16].

Пріоритет має аналітичний вигляд – індекс або ваговий коефіцієнт. Розрахунок даного параметру може виконуватись на базі несистемних або внутрішньо системних параметрах СМО.

Згідно джерел [11, 21] найбільш поширеними несистемними параметрами в телекомунікаційних системах є:

- категорія абонента;
- рівень якості надання послуги;
- номер заявки в черзі, інші.

При аналізі пріоритетів, що формуються на базі системних параметрів, а саме – пріоритет, що чергується, динамічний пріоритет, ситуаційний пріоритет, виявив науковий потенціал і можливу практичну цінність даного виду пріоритетів [16, 17, 18].

Суть дисципліни з пріоритетом, що чергується в тому, що рішення вибору обслуговування заявки одночасно залежить від пріоритетного класу, до якого відноситься заявка, і від класу останньої обслугованої заявки.

Дисципліна з динамічним пріоритетом визначає заявки, що береться на обслуговування, визначається не тільки пріоритетною шкалою, але і часом очікування даної заявки в черзі.

В системах оборки викликів перспективним є підхід к формуванню пріоритету на основі параметрів, що залежить від поточного стану системи. Такий спосіб впливу дає можливість підвищити ефективність на СМО в конкретних умовах, ситуації, а пріоритет отримав назву – ситуаційний пріоритет.

3.1.2.2. Ситуаційні пріоритети в інформаційних системах

Ситуаційні пріоритети були вперше введені в роботі [22] і набули подальшого розвитку в роботі [16]. Стан системи з очікуванням характеризується вектором $x=(x_1,x_2,\ldots,x_N)$, *де* x_i – кількість заявок *i-го* типу в системі. Множини таких векторів з цілими невід'ємними компонентами утворюють простір станів системи $x=\{x=(x_1,x_2,\ldots,x_N)\}$.

Цей простір розбитий на множини, що не перетинаються: x_k $(k=0,1..N)$,

$x = \bigcup_{k=0}^{N} x_k$, $x_k \cap x_s = \notin$, $k \neq s$. Для кожного такого розбиття $x=\{x_k, k=0,1..N\}$ визначається стратегія управління. Якщо в момент прийняття управляючого рішення система заходиться в одному із станів підмножин x_k, приймається управління $V = V(x_k) = k$. Обмеження на моменти управління вводяться окремо в залежності від типу системи. Система, в якій стратегія управління вибирається описаним вище способом за допомогою розбиття на простори станів, називається системою з ситуаційними пріоритетами.

Система з очікуванням і відносними ситуаційними пріоритетами буде оптимальною, якщо вона оптимальна при відносному пріоритеті.

В інформаційних системах з великим обсягом заявок, що надходять на обробку, дуже складно або зовсім неможливо зробити попереднє сортування за формальним ознакою α, що характеризує кожну заявку. В залежності від α і кількості зайнятих обслуговуючих пристроїв (каналів) приймається рішення брати або не брати заявку на обслуговування для мінімізації втрат в СМО.

Завдяки високій ефективності, ситуаційні пріоритети знаходять все більш широке застосування на практиці у Центрах обробки даних (ЦОД), в системах АСУ, промислових і транспортних проектах.

З практичної точки зору перспективним є управління СМО по ситуації (стану), при якому процес функціонування об'єкта визначається таблицею рішення, де вхідним рядкомє ситуація (стан), вихідним – рішення.

3.1.2.3. Критерії ефективності організації пріоритетного обслуговування

На ефективність функціонування будь-якої системи СМО впливають наступні фактори [16]:

- характеристики і параметри вхідних потоків заявок;
- характеристики і параметри обслуговування заявок різних типів;
- структура і об'єм буферних накопичувачів;
- дисципліна вибору із черги.

Дисципліну обслуговування можна оцінити за допомогою лінійного функціонала, що характеризує величину сумарного штрафу за одиницю часу функціонування системи [16]. (3.1)

$$C_w^{(s)} = \sum_{i=1}^{(s)} \alpha_i \cdot \lambda_i \cdot W_i^{(s)}$$

де α_i, $W_i^{(s)}$ – штраф за одиницю часу очікування і середній час очікування заявки i-го типу; λ_i – інтенсивність потоку i-го заявок; N – кількість типів заявок.

В системах з обмеженою чергою можливі втрати заявок до обслуговування. При переповненні черги заявки i- го типу втрачаються з ймовірністю Rf_i і втрати ефективності при даній дисципліні обслуговування можна визначити за допомогою функціоналу:

$$C^{(s)} = \sum_{i=1}^{N} \alpha_i \cdot \lambda_i \cdot W_i^{(s)} + \sum_{i=1}^{N} \alpha_i^1 \cdot \lambda_i \cdot Rf_i^{(s)} \quad (3.2)$$

де α_i^1 – штраф за втрату однієї заявки переповнення черги.

Для загального випадку різні втрати ефективності в СМО можно визначити як втрати знаходження системи у відповідних станах. Тоді критерій ефективності:

$$C^{(s)} = \sum_{i=1}^{N} \lambda_i \cdot (\sum_k \alpha_k \cdot \pi_k) \quad (3.3)$$

де N – множина можливих станів системи станів; α_k- штраф за прибування системи в стані k, π_k – стаціонарна ймовірність знаходження системи в стані k.

Таким чином, за допомогою вибору дисципліни обслуговування можливо досягти знаходження системи більшу частину часу в станах з мінімальним штрафом, і підвищити її ефективність.

3.1.3. Сучасні системи обробки та тарифікації викликів в мобільних мережах зв'язку

Задачі управління якістю надання послуг в сучасних мобільних мережах зв'язку вирішуються на різних рівнях.

Перший тип задач – транспорт послуги, передбачає передачу послуги з гарантованою швидкістю передачі, затримкою, джитером, коефіцієнтом втрат пакетів. Такі задачі вирішуються на 1–4 рівнях OSI, за допомогою протоколів маршрутизації [12, 45], методів пріоритетної обробки пакетів на проміжних вузлах і шлюзах – WFQ, CBQ, FIFO, LIFO, описані [46], алгоритмів розподілу ресурсу [21, 45, 47], гарантування якості при підвищення утилізації ресурсів [48, 49], методів оцінки моніторингу та технічних показників якості мережі [49, 50] в рамках систем мобільного зв'язку 2-го, 3-го, 4-го покоління [60 – 62].

Другий тип задач відноситься до управління якістю послуги на 5-му і вищих рівнях OSI. На цих рівнях вирішуються наступні задачі:

- забезпечення встановлення з'єднання з заданим коефіцієнтом відмов і затримки;
- узгодження рівня якості послуги з профілем абонента, станом його балансового рахунка та завантаженням мережі;
- управління якістю відповідно до запитів користувача;
- управління форматом представлення інформації;
- управління транспортним рівнем мережі для резервування необхідного ресурсу з метою гарантування якості подальшого обслуговування послуги; моніторинг параметрів якості, що впливають на тарифікацію послуг.

Задачі встановлення з'єднання розглядаються в роботах [63–66]. У джерелі [64] досліджується затримка роботи кожної команди SIP протоколу [67, 68]. Автори статті виявили, що максимальне навантаження щодо обслуговування викликів в IMS системі припадає на функціональний блок обслуговування SIP викликів. У роботі [73] запропоновано модель продуктивності SIP протоколу, що базується на середній затримці очікування встановлення виклику і витраченій пропускній здатності для реалізації протоколу. Наведено порівняння теоретичної моделі з вимірами і підтверджено адекватність моделі.

Зроблені висновки спонукають до більш детального вивчення варіантів реалізації з'єднання на базі SIP протоколу. Питання дослідження перевірки заданих параметрів якості при надходженні SIP запиту на з'єднання досліджені в роботі [68]. У роботі показано важливість протоколу перевірки ресурсу мережі доступу для впровадження послуги – Resource Availability System (перевірки ресурсів системи). На основі результатів перевірки автори пропонують систему адаптації мультимедійного потоку, що пов'язана зі створенням адаптивного кодеку для передачі такого виду інформації.

Питання узгодження якості послуги з профілем абонента, станом його балансового рахунка і перевантаженням мережі розглядаються в роботах [64, 65, 69-72]. У [69] описується нова архітектура управління якістю та тарифікацією, що дозволяє гнучко і незалежно надавати доступ до послуг згідно до сформованих правил.

У роботі [63] пропонується удосконалення системи за рахунок можливості введення динамічного управління послугами та тарифікацією шляхом переговорів.

Користувач може задавати параметри тарифікації як діапазон вхідних параметрів для переговорного процесу. Наприклад, мінімальну та максимальну прийнятну ціну послуги. В залежності від заданих параметрів може бути впроваджена конфігурація, найближча до мінімальної або до максимальної ціни. Після встановлення результуючої конфігурації, параметри відправляють до блоку прийняття рішень щодо активації послуги. Конфігурація може бути змінена в переговорному режимі на будь-якому етапі надання послуги через зміну ресурсів мережі або зміну початкових вхідних параметрів. Реалізація такого підходу базується:

- по-перше, на модифікації обміну інформацією між користувачем, системою управління якістю та тарифікацією, мережею доступу;
- по-друге, на модифікації алгоритмів роботи блоків впровадження послуг і прийняття рішень (PCEF, PCRF,) що описані авторами на базі діаграм станів системи;
- по-третє, на створенні нового інтерфейсу взаємодії між блоком PCRF і підсистемою тарифікації OSC (система тарифікації), що пропонується для реалізації переговорного процесу між функціональними блоками в PCC. Також пропонується модифікація процедури встановлення сесії. Переваги такого підходу – в гнучкості присвоєння рівня якості обслуговування і адаптації відповідного тарифу.

Автори даної статі [63] не розкривають реалізації своєї пропозиції щодо управління якістю та тарифікацією. Недоліки такого підходу:

- відсутні правила формування залежності тарифу від якості та стану мережі, обмеження переговорного процесу;
- непрозорість отримання параметрів мережі, на основі яких будуються можливі конфігурації;
- не визначені механізми впливу користувача на параметри тарифікації, а саме: принцип, протоколи і програмні засоби.

Основними елементами управління пакетними послугами в сучасних системах є вузол GGSN (Gateway Service Node) – Обслуговуючий шлюзовий вузол, та SGSN (Serving GPRS Support Node) – Вузол обслуговування абонентів GPRS. Функції вузлів GGSN і SGSN детально описані [3]. Функціональна архітектура тарифікації пакетних послуг в мережах GSM зображена на рис. 3.3.

Основною функцією шлюзу тарифікації є збір, зберігання, форматування та передача тарифікаційної інформації від вузлів GGSN і SGSN до білінгової системи. CGF підтримує два режими роботи. Перший режим – централізований, що підтримує передачу інформація від GGSN і SGSN в один елемент. Другий режим – децентралізований, в даному випадку функція CGF вбудована в вузли від GGSN і SGSN.

Детальний опис інтерфейсів U_m, U_u ,G_b, G_n, G_a, IuPS наведений в джерелі [50]. Інтерфейс G_a є інструментом передачі інформації CDR (Charging Data Record) – тарифікаційний запис.

Наступна тарифікаційна інформація, визначена стандартом [28], що входить до CDR:

- кількість даних переданих і прийнятих користувачем, вимірюється в мегабайтах;
- протокол і якість обслуговування QoS в кожному з напрямків передачі інформації;
- тривалість передачі інформації;

– інформація щодо адреси відправника і адресату інформації;
– місце знаходження (номер БС) під час користування послугою;
– інша інформація.

При тарифікації CGF повинна враховувати зміну тарифу при наступних [27], подіях:

– зміна якості обслуговування,
– зміна часу дії тарифу;
– досягнення границі кількості переданої інформацій, при котрій наступає зміна тарифу.

Для передачі необхідної інформацій між вузлами CGF і GGSN/SGSN використовують протокол GTP' – протокол, що спеціально розроблений для передачі тарифікаційної інформації, при передачі GPRS трафіку за протоколом тунелювання GPRS Tunnelling Protocol [73].

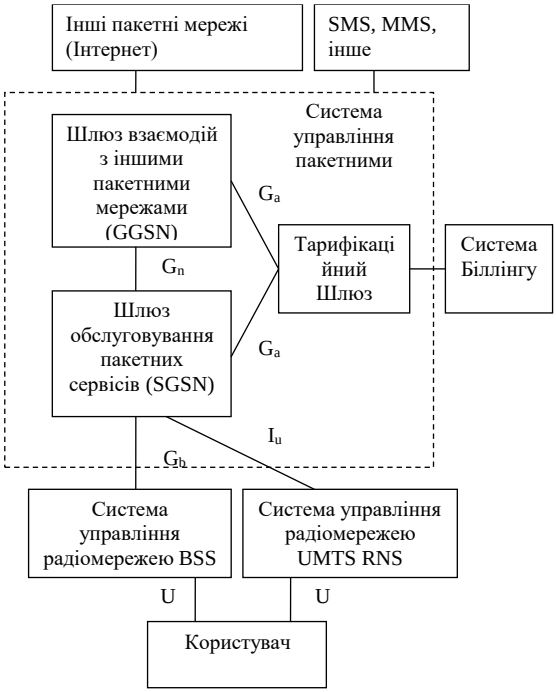


Рис. 3.3 Функціональна архітектура тарифікації пакетних сервісів GPRS/UMTS

Основним недоліками системи передачі пакетних сервісів GPRS є:

– по-перше, невелика швидкість передачі, максимально теоретична на абонента 172 кбіт/с;

– по-друге, існуючі механізми гарантування якості [21] не можуть гарантувати належних значень параметрів якості обслуговування мультимедійних послуг. Однією з найголовніших причин є те, що клас тарфіку GPRS є менш пріоритетним порівняно з голосовими каналами. Тому неможливо гарантувати

якість мультимедійних послуг, що надаються, наприклад, в Інтернеті. Додатково необхідно відмітити, що постійний ріст GPRS каналів в діапазоні 900 МГц або 1800 Мгц не можливий в зв'язку з обмеженістю смуги.

Основні переваги реалізації сервісів в архітектурі системи тарифікації пакетних послуг GPRS [73]:

– точна настройка кожного сервісу окремо (SMS, MMS, доступ до Інтернет тощо), необхідних параметрів для тарифікації та інтерфейсів;

– порівняно просте управління тарифікацією. Тарифікація кожного сервісу – це окремий рівень (рис. 3.4);

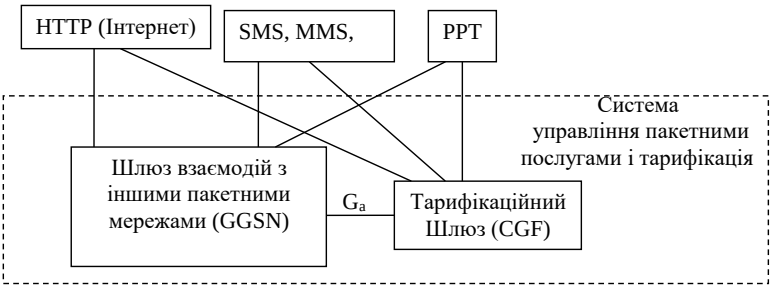


Рис. 3.4 Існуюча архітектура тарифікації пакетних сервісів в системі управління послугами GPRS/UMTS

Основними недоліками реалізації існуючої архітектури тарифікації пакетних сервісів в системі управління пакетними послугами GPRS/UMTS сервісів в архітектурі системи тарифікації пакетних послуг GPRS/UMTS [66] є:

– обмежена масштабованість системи. Для впровадження кожної послуги необхідна складна інтеграція в тарифікаційну систему;

– складна тарифікації при використанні одночасно багатьох послуг;

– неврахування якості при тарифікації пакетних послуг, що є загальним недоліком системи;

– тривалий час для введення послуги в експлуатацію;

– не враховує частотний ресурс при тарифікації послуг в мережах 4- го покоління з ОЧД.

Існуючі недоліки тарифікації пакетних послуг в мережах операторів GPRS/UMTS а саме: немасштабованість, неврахування якості при обслуговуванні та тарифікації [67–69], складність надання персоналізованих сервісів та їх тарифікації, – є одним з ключових факторів, що мотивували до розробки нових систем управління обробкою викликів в мобільних мережах зв'язку на базі протоколу IP.

Така система, запропонована в рамках проекту співробітництва 3GPP [74], отримала назву IMS система управління послугами на базі IP протоколу. Основними ініціаторами створення системи стали компанії Ericsson, Alcatel-Lucent, Huawei, Nokia-Simens, інші.

3.1.4. Процес обробки викликів в мобільних мережах зв'язку

Алгоритм методу обробки викликів послуг в мобільних мережах 4-го покоління. Передумови:

– термінал користувача зареєстрований в мережі оператора;

– запит передається після процедури ініціації з'єднання RRC і NAS [69] для бездротової мережі доступу;

– сервер і користувач знаходяться в «домашній» мережі.

Опис методу по крокам.

1.Блок P-CSCF отримує SIP запит на встановлення з'єднання, що включає опис програми прикладного рівня в тілі протоколу SDP[87]. P-CSCF присвоює сесії унікальний ICID, для визначення сесії, в тому числі тарифікації сесії, пересилає запит на блок-функцію S-CSCF.

2.S-CSCF приймає SIP запит, аналізує адресу прикладного інтерфейсу AF, наплавляє запит на API (інтерфейс програмування прикладних програм)

3.AF приймає та аналізує SDP тіло, формує вимоги обслуговування послуги в мережі – швидкість, рівень якості, кодек. AF пересилає данні до PCRF

4.PCRF отримує від AF через *Rx* ідентифікатор прикладної програми з детальним описом вимог щодо якості і швидкості передачі, в тому числі тарифікаційний ключ, що визначає модель тарифікації: за часом, за Мбіт/с, інше.

5.PCRF створює правила обслуговування, включаючи тарифікаційні параметри, що мають відстежуватись для даного виду послуги. PCRF пересилає правила для впровадження PCEF.

6.PCEF отримує правила та впроваджує їх. PCEF ініціює процедуру перевірки або резервування каналу між абонентом і сервером послуг відповідно заданих PCRF правилами.

7.Якщо PCEF процедура, описана в п. 6, відбулася успішно, то автоматично відкривається тарифікаційний запис CDR, що зберігається у базі даних блоку PCEF.

8.Початок передачі послуги Користувач – Сервер на основі IP протоколу. Послуга повинна мати відповідний API інтерфейс і підтримувати протокол Diameter.

9.Під час передачі генеруються проміжні тарифікаційні записи CDR на основі відповіді на запит команди протоколу Diameter «UPDATE» до вузла PCEF.

10. Кінець передачі. SIP повідомлення про закінчення сеансу отримали P-CSCF, S-CSCF, AF, користувач.

11. Закривається тарифікаційний запис.

12. Відбувається кореляція записів, що отримана від елементів учасників з'єднання, в основному від PCEF. Здійснюється передача в CDR, в систему біллінга.

Аналіз методу показує, що однією з основних процедур є виконання правил передачі (5,6), встановлення перевірки каналу за заданими параметрами і тарифікація, що відбувається в функціональних блоках PCEF, PCRF.

Існує два варіанти управління правилами, в разі встановлення з'єднання, для вхідних і вихідних викликів[1, 2]:

- примусове (push) управління політиками запитів;
- управління правилами «по запиту» користувача.

На рис. 3.5 показана діаграма послідовності примусового управління політиками запитів.

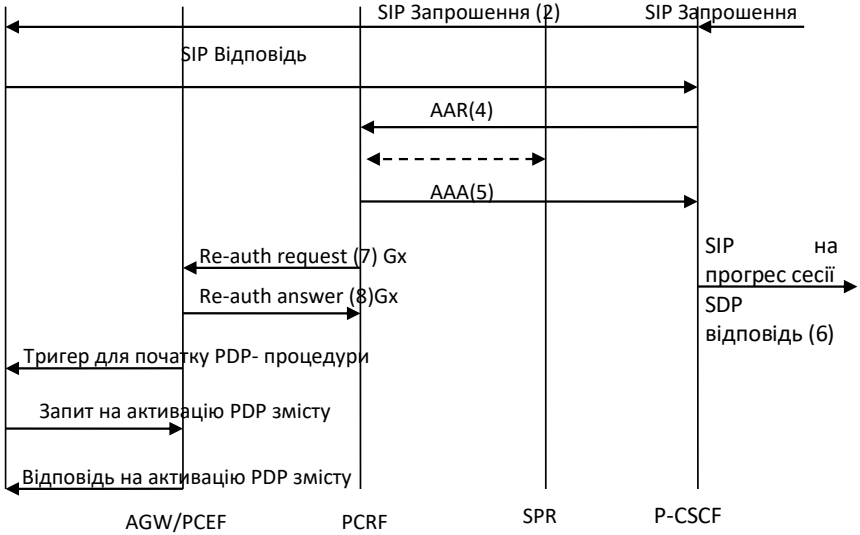


Рис. 3.5 Метод примусового управління правилами обробки вхідних запитів архітектури PCC

Метод управління обробкою викликів описаний нижче по крокам.

Блок P-CSCF отримує «SIP запрошення» на встановлення сесії з SDP описом сесії.

1.P-CSCF направляє запит відповідному терміналу на встановлення сесії.

2.Термінал відповідає на запит за допомогою повідомлення «SIP Відповідь». Повідомлення містить згоду або відмову на встановлення з'єднання.

3.P-CSCF направляє за допомогою Diameter протоколу запит AAR (Authorization-Authentication(AA)-Request) – запит аутентифікації користувача і авторизації послуги та ресурсу для неї в PCRF. Запит включає опис медіа компоненту та суб-компоненту. Процедура перевіряє можливість надання послуги (через взаємодію з SPR) з необхідними характеристиками в мережі доступу.

4.PCRF відповідає на запит повідомленням «AA-Answer» – «успішно» або «не успішно». Під час цієї процедури будуються правила тарифікації і обслуговування послуги користувача.

5.P-CSCF посилає повідомлення – «SDP відповідь» щодо прогресу сесії на сервер ініціації виклику.

6.PCRF посилає запит в PCEF на виконання політики обслуговування та тарифікації викликів.

7. PCEF, через взаємодію з AGW, перевіряє можливість виконання політики обслуговування викликів, посилає відповідь до PCRF – «успішно» або «не успішно».

8.PCEF посилає активаційний тригерний запит на термінал через PDP протокол (Packet Data Protocol – пакетний протокол передачі даних та команд до

користувача).

9.Термінал користувача посилає запит PDP з підтвердженням готовності початку передачі.

10. Шлюз відповідає – підтверджує активацію PDP запита.

Розглянуті механізми тарифікації та управління політиками вхідних запитів показують зв'язок між цими процесами.

Виходячи з аналізу сучасних систем тарифікації, слід відмітити особливості нових систем тарифікації, в тому числі на базі платформи IMS [1,2]. Ці системи є конвергентними, можуть проводити тарифікацію послуг в будь-яких мережах доступу. Також нові системи тарифікації побудовані на основі функціональних архітектур, що включають функції управління тарифікацією і якістю передачі. Такі системи тарифікації враховують параметри, що детально описують послуги та субпослуги, використані кодеки при наданні послуги, якість обслуговування, період передачі, швидкість передачі, кількість переданої інформації.

Аналіз параметрів даних систем [1–3, 14, 28] показав, що не враховані величини частотного ресурсу при тарифікації і наданні послуг в системах LTE, що базуються на ОЧД доступі.

3.1.5. Обґрунтування необхідності врахування смуги частот в системі обробки викликів та тарифікації сучасних мобільних мереж

Важливо показати необхідність врахування радіоресурсів – ширини смуги частот в мережах LTE та мережах на базі технології доступу SC-OFDMA/OFDMA при тарифікації. Тому розглянемо методи організації каналів і доступу до каналів у LTE мережах.

В E-UTRAN [89] існує два основних: дуплексний режим з розділенням по частоті (FDD) та по часу (TDD). Особливістю E-UTRAN є використання OFDMA технології доступу із окремою модуляцією піднесучих. Спектр частот розподілений на множину піднесучих, що є ортогональними. Кожна піднесуча окремо модулюється потоком із заданою швидкістю.

На рис. 3.7 показано сигнал OFDM [88] зі смугою пропускання 5 МГц. Для інших смуг принцип не змінюється. Символи даних незалежно промодулювані та передані по великій кількості поруч розташованих ортогональних піднесучих. В E-UTRAN існують наступні схеми модуляції при передачі від БС до рухомого об’єкту: QPSK, 16QAM, і 64QAM [88].

Інформація передається в елементах ресурсних блоків (рис. 3.6). Фізично, ресурсний блок складається із 12(24) послідовних піднесучих в частотній області. Відстань між ним складає 15 КГц або 7.5 КГц, в залежності від обраного режиму. У часовій області ресурсний блок складається з N_{sym}^{Dl} послідовних OFDM символів. Розмір ресурсного блоку однаковий для всіх смуг пропускання, тому кількість доступних ресурсних блоків залежить від ширини смуги пропускання.

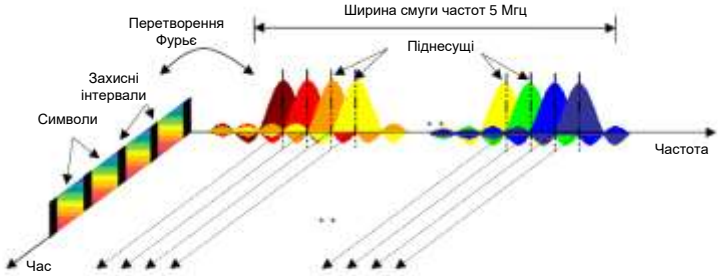


Рис. 3.6 Частотно-часове представлення OFDM сигналу

БС назначає один або більше ресурсних блоків абонентської станції в кожному часовому інтервалі передачі довжиною 1 мс. Кількість назначених ресурсних блоків залежить від необхідної швидкості передачі. Планування виділення ресурсних блоків виконується БС.

Дані користувача передаються по каналу *PDSCH*(*Physical Downlink Shared Channel)* – фізичний загальний канал передачі даних зверху вниз. (рис. 3.7)

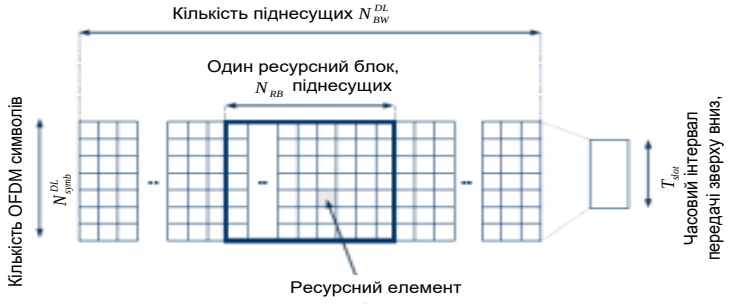


Рис. 3.7 Структура фізичного каналу передачі зверху вниз радіомережі E-UTRAN

Контроль за передачею інформації виконується за допомогою фізичного каналу управління – PDCCH (Physical Downlink Control Channel), що розташований в першому OFDM символі слота.

Один слот передачі зверху вниз складається з N_{symb}^{DL} OFDM символів. До кожного символу прикріплюється циклічна приставка (ЦП), як цикл синхронізації

Довжина N_{symb}^{DL} залежить від ЦП (табл. 3.2). Структура стандартного кадру з нормальною довжиною ЦП включає 7 символів. Відповідно, ЦП для першого символу дорівнює 7.2 мкс, і 4.7 мкс для інших 6-ти символів. Для великих стільниць з високим рівнем затримки розповсюдження сигналу може використовуватись додаткова розширена приставка довжиною 16.7 мкс, що включає 6 символів з кроком піднесущих 15 кГц. Кожна піднесуща модулюється 4-,16- або 64-позиційною квадратурною модуляцією. Відповідно, один символ на одній піднесущій може нести 2, 4 або 6 біт.

Табл. 3.2

Параметри для downlink структури загального кадру

Конфігурація	**Кількість символів** N_{symb}^{DL}	**Довжина ЦП, символів**	**Довжина у часі**
Стандартний префікс ЦП, відстань між під несущими – 15 КГц	7	160 для 1-го символу 144 для інших	7.2 мкс для 1-го символу 4.7 для інших
Розширений префікс ЦП, відстань між під несущими – 15 КГц	6	512	16.7 мкс
Розширений префікс ЦП, відстань між під несущими – 7.5 КГц	3	1024	33.3 мкс

При стандартному префіксу, символьна швидкість дорівнює 14000 символів/с, що відповідає агрегативній швидкості від 28 до 84 кбіт/с. Детальний розрахунок швидкості передачі наведений в табл. 3.3. Сигнал зі смугою 20 МГц включає 100 ресурних блоків або 1200 піднесучих, що дає загальну агрегативну швидкість у радіоканалі від 32 до 96 Мбіт/с.

Це обумовлює використання різної кількості піднесучих при наданні однієї послуги, а відповідно й кількість ресурсних блоків R_0. Приклад розрахунку для організації передачі он-лайн послуги з симетричною швидкістю передачі 512 кбіт/с і максимальною затримкою пакетів не більше 100 мс, наведений в табл. 3.3.

Табл. 3.3.

Розрахунок швидкості передачі при різних видах модуляції для LTE технології

Тип модуляції	**Символів на піднесучу**	**Швидкість модуляції, біт/с**	**Швидкість на одну піднесучу, кбіт/с**	**Агрегативна швидкість передачі в каналі 20Мгц, Мбіт/с**
QPSK	2	14000	27.34	32.07
QAM16	4	14000	54.69	64.09
QAM64	6	14000	82.03	96.13

Таблиця показує, що при наданні он-лайн послуг з швидкостями більше ніж 512 кбіт/с і використанням 12 піднесучих можна надати тільки одну послугу з даними параметрами, та при модуляції вищій за QPSK. Для передачі послуги з зазначеними параметрами та використанні 24 піднесучих можливо передати 1-3 таких послуг.

Табл. 3.4
Розрахунок ємності каналів

Модуляція	Швидкість на 12 піднесучу, 20xRo, кбіт/с	Кількість послуги на 12 піднесучих	Швидкість на 24 піднесучу 40xRo, кбит/с	Кількість послуги на 24 піднесучих
QPSK	328.13	1	656.25	1.00
QAM16	656.25	1.00	1,312.50	2.00
QAM64	984.38	1.00	1,968.75	3.00

Тому для адекватного відображання спожитого ресурсу мережі, що споживається при наданні таких послуг, необхідно враховувати частотний ресурс.

Таким чином, аналізуючи структуру каналу в мережі LTE, маємо відмінності порівняно з доступом в режимі передачі даних в UMTS і GSM. В LTE не використовується доступ на множинній основі з кодовим розподіленням абонентів. Доступ до мережі надається за допомогою часового розподілення та розподілення частотного ресурсу – кількості частотних каналів, що залежить від режиму роботи. Також фіксується певний частотний ресурс, що надається абоненту на визначений час. Виходячи з цього, з'являється можливість визначати обсяг частотного ресурсу (кількість піднесучих), що задіяний при наданні послуг. Обсяг частотного ресурсу є одним з основних параметрів, що має бути врахований при тарифікації мудьтимедійних послуг.

3.1.6. Інтегрована система обробки та тарифікації викликів

Загальна схема мобільної мережах 4-го покоління IMS/LTE [1] зображена на рис. 3.8. До складу мережі входить:

- користувачі;
- базова станції (БС);
- зонова мережа, що включає маршрутизатори (R);
- зонові шлюзи доступу (AGW)
- національна мережа IP/MPLS, що об'єднує вузли AGW;
- сервери обслуговування викликів, маршрутизації, послуг, і сервери база даних інформації користувачів.

Особливістю мереж 4-го покоління є використання пакетних технологій передачі, зокрема IP протоколу на транспортному рівні [15]. IMS детально описана в стандарті 3GPP TS 23.228 [74] і джерелах [1, 3]. Архітектурні особливості IMS-мережі обумовлюють зміну існуючого процесу розробки та впровадження послуг. IMS забезпечує високу гнучкість налаштування послуг за рахунок використання єдиного профілю абонента, що зберігається в базі властивостей абонента.

Профіль містить індивідуальні налаштування для кожної послуги, інформацію по регулюванню рівня якості обслуговування для кожного абонента, що вирішує або забороняє використання мережевих ресурсів.

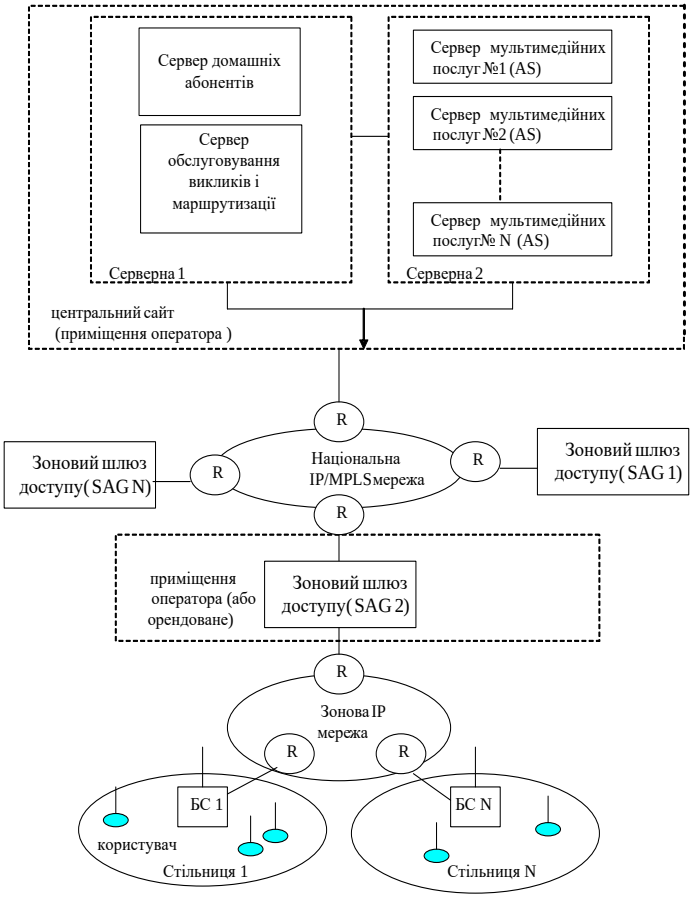


Рис. 3.8 Загальна схема мобільної мережі 4-го покоління

У IMS-мережі також стають доступні нові способи тарифікації абонентів. Окрім традиційної тарифікації на основі тривалості користування послугою, кількості переданих повідомлень або об'єму переданої/отриманої інформації, можливе здійснення тарифікації залежно від сервісів, що дозволяє створювати нові моделі тарифікації, а відтак і нові бізнес-моделі, що дозволяють операторові підвищити ефективність системи обробки викликів оператора .

Деталізована схема мережі 4-го покоління LTE/IMS зображена на рис. 3.9. Архітектура мобільної мережі 4-го покоління (рис. 3.8) включає три основні складові. Перша – мережа доступу – складається з: терміналу користувача (Terminal); базової станції, що є основною складовою мережі радіодоступу 4-го покоління (eNodeB);

обслуговуючого шлюзу доступу (AGW – Access Gateway) – точка концентрації та термінації сесій радіомережі, управління мобільністю користувача. Друга складова – ядро мережі – включає: блок управління сесією (CSCF – Call Sesion Contriol Function), що відповідає за з'єднання користувачів, відкриття сесії користування послугою й т. і.; сервер-реєстратор домашніх абонентів (Home Subscriber Server), що зберігає та оновлює інформацію про підписку користувача, місце знаходження, стан (активний/ неактивний), інше; шлюз взаємодії з іншими мережами (IGW- Inteaction Gateway): телефонними мережами загального користування, мережами Інтернет тощо. Третя складова – це платформи: сервера послуг (AS- Application Server) – програмно-апаратні платформи, що надають послуги абоненту; тарифікаційна платформа (OSC – Online/offline charging system), платформа білінга (Billing) – виконують функції генерації тарифікаційних записів, розрахунок тарифу, формування рахунку абоненту за використання послуг.

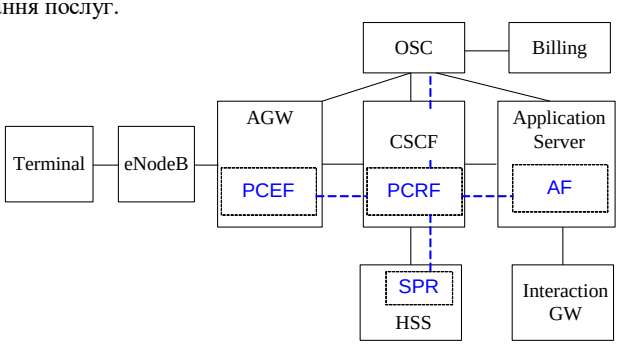


Рис. 3.9 Система обробки викликів та тарифікації в мережі 4-го покоління IMS/LTE

Розглянемо модель надання послуги: *клієнт – сервер*. Користувач має клієнтську програму, що зберігається в терміналі, з переліком послуг в мережі мобільного оператора. Користувач звертається до серверу послуг (AS) з метою отримати мультимедійну послугу *Y* в реальному часі зі швидкістю більше 256kbit/s.

Умови:

– користувач знаходиться в домашній мережі і зареєстрований в HSS;

– користувач не обмежений в доступі до послуги *Y;*

– у користувача достатньо коштів на рахунку для ініціалізації послуги;

– в мережі достатньо ресурсів для надання послуги.

Опис по крокам процесу надання доступу до послуги *Y*:

1. Встановлення з'єднання терміналу з AGW для запиту послуги:

1.1. Термінал користувача ініціює процедуру виділення ресурсу для встановлення з'єднання з БС – процедура RRC –Radio Resource connection request. Успішно.

1.2. Водночас БС запитує дозвіл на вставлення з'єднання з AGW/MME за процедурою NAS (Non Access Stratum). Відбувається аутентифікація термінала користувача, і назначається IP адрес і порт APN (TCP) *для передачі сигнальної інформації*.

1.3. Якщо дві попередні процедури проведені успішно, то перехід до п.2.

2. Клієнтська програма користувача (з підтримкою SIP протокола) формує запит на отримання послуги до CSCF

2.1. Термінал посилає SIP (Session Initiation Protocol) запит - команда «Invite-Запит» інкапсулюваний в IP пакет до AGW. AGW отримує запит і пересилає до CSCF.

2.2. CSCF приймає SIP запит, аналізує адресу серверу і прикладну функцію програми AF з прикладним інтерфейсом API (інтерфейс програмування прикладних програм) – middleware (програмне забезпечення для адаптації сесії під користувача з урахуванням можливості терміналу - розширення екрану, рівень відтворення аудіо інформації тощо, і рівнем якості).

3. Формування правил обробки викликів і тарифікації (підсистема PCC).

4. Початок передачі послуги:

4.1. PCRF посилає підтвердження про початок передачі до CSCF і AF.

4.2. CSCF формує SIP запит «відповідь на встановлення з'єднання» і передає до терміналу користувача.

4.3. CSCF отримує SIP повідомлення, – ACK, що користувач готовий до прийому/передачі.

4.4. CSCF направляє SIP повідомлення з кодом 180 – початок передачі.

4.5. Відбувається передача інформації між користувачем (клієнтською програмою) і AS на базі IP протоколу.

4.6. Під час передачі генеруються проміжні тарифікаційні записи CDR на основі відповіді на запит блоку OSC, команди протоколу Diameter "UPDATE" до вузла AGW, CSCF, AS.

4.7. При он-лайн тарифікації, OSC з заданою періодичністю розраховує вартість використаного ресурсу і виконує контроль рахунку абонента з метою не перевищення ліміту коштів.

5. Кінець передачі. Користувач (клієнтська програма) посилає SIP повідомлення про закінчення сеансу на CSCF, AS. Закривається тарифікаційний запис. Відбувається кореляція записів, що отримані від елементів учасників з'єднання. Здійснюється формування єдиного CDR в OSC відповідно сесії та передача його в систему біллінга.

Розрізняють дві моделі тарифікації [1, 27]: офф-лайн і он-лайн тарифікацію. У разі офф-лайн схеми тарифікації, звіт, що генерується мережею посилається в біллінг домейн (BD) по закінченню моменту користування ресурсом. У цьому режимі немає прямої взаємодії з процедурами відшкодування при використанні ресурсів мережі, що є основною характеристикою механізму офф-лайн.

Слід зазначити різницю між поняттям офф-лайн тарифікації та пост-пейд біллінгу. Останнє визначає тільки метод оплати рахунків за послуги, отримані абонентом. Під час он-лайн тарифікації відбувається безперервна взаємодія в он-лайн режимі між процесом тарифікації та використовуваним сервісом з метою реалізації контролю стану рахунку абонента, та розпізнавання тарифікаційних подій. Під час процесу тарифікації відбувається авторизація користувача перед наданням доступу до ресурсу.

Система для вирішення недоліків в архітектурі обробки викликів та тарифікації послуг – PCC [27], в тому числі в мережі мобільних операторів з ОЧД [66] представлена на рис. 3.10. PCC визначає уніфіковані інтерфейси, протоколи, правила перетворення сигнальної інформації на рівні сесії в сигнальну інформацію

на мережевому рівні, встановлення з'єднання. PCC визначає правило забезпечення QoS і правила тарифікації для кожного потоку (послуги і субпослуги), що визначається специфікою даних прикладної програми користувача і передається на рівні з'єднання. Базовим сигнальним протоколом, який зв'язує функціональні блоки IMS, є протокол Diameter [5], що представляє собою протокол нового покоління аутентифікації, авторизації і обліку. Протокол базується на принципі «запит – відповідь». Повідомлення включають AVP параметри [62]. Параметри AVP можуть бути розширені. Функція проксі, Рис. 3.12 – управління виклику впродовж сесії – Proxy - Call Session Control Function (P-CSCF) – вузол що знаходиться між ядром системи IMS і обладнанням користувача [89], і приймає запит за протоколом Session Initiation Protocol (SIP) [27, 91] – протокол ініціації сесії. P-CSCF пересилає інформацію щодо сесії в PCC вузол.

Рис. 3.10 Архітектура системи обробки викликів та тарифікації

Блок PCRF формує набір правил тарифікації обслуговування сесії користувача. Блок PCRF приймає рішення щодо вибору тарифікаційної моделі щодо обслуговування визначеного потоку даних в пакетній мережі.

Рішення залежить від профайла абонента, що зберігається в базі даних абонента, наявності коштів на рахунку абонента, адресата виклику. Кожне правило PCC включає тарифікаційний ключ, який визначає модель тарифікації: на базі часу, події, кількості переданої інформації, тощо, і механізму тарифікації (он-лайн чи офф-лайн).

Функціональний блок впровадження тарифікації та політики обробки викликів послуг (Policy and Charging Enforcement Function) – PCEF знаходиться в шлюзі управління, на рівні вузлів GGSN в GPRS, AGS в WiMax/LTE. PCEF відповідає за отримання та впровадження PCC правил від блоку PCRF через Gx за допомогою Diameter протокола [14]. PCEF резервує мережевий ресурс і виконує тарифікацію взаємодіючи з офф-лайн/он-лайн тарифікаційними підсистемами.

Таким чином, на основі проведених досліджень сформульована актуальна наукова задача удосконалення системи обробки викликів за рахунок введення

ситуаційних пріоритетів, що базуються на системних параметрах, зокрема ширині смуг частот, та несистемних параметрах (інформаційній важливості контенту) мереж 4-го покоління і складається з:

- синтезу нової дисципліни обслуговування, яка б враховувала тариф і ресурс послуги;
- розробки нового методу обробки викликів;
- удосконалення моделі архітектури системи обслуговування викликів і тарифікації.

3.2. Дисципліни обслуговування викликів в центрі керування мобільною мережею

Управління обслуговуванням викликів в мобільних мережах це складний процес, який має ряд обмежень порівняно з традиційними СМО. До ряду таких обмежень можна віднести наступні: неможливість контролювати находження викликів (джерело), час обслуговування викликів, обмежений ресурс системи (кількість обслуговуючих каналів). Розділ присвячений дослідженню ефективності систем обслуговування викликів при використанні різних дисциплін обслуговування при умові перевантаження мережі (обмеженому канальному ресурсу).

3.2.1. Управління процесом доступу до послуг

Послідовність процеса обробки та тарифікації викликів в мобільних мережах 4-го покоління показана на рис. 3.11.

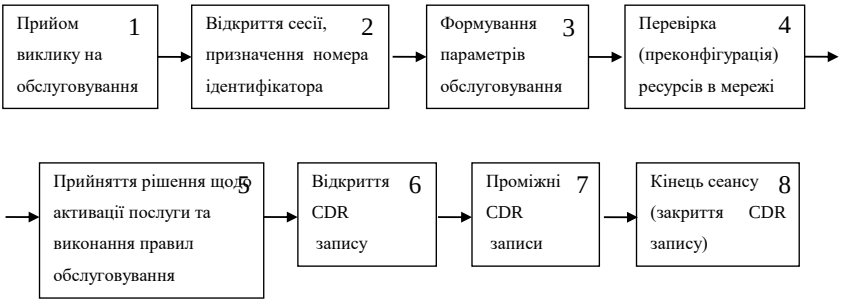


Рис. 3.11 Процес обробки та тарифікації викликів

Особливості процесу обробки викликів в мобільних мережах 4-го покоління:

- по-перше, набір параметрів (3), що характеризують послугу і необхідні для організації обслуговування;
- по-друге, процедура перевірки ресурсів в мережі доступу(4), відрізняється від існуючої особливістю організації каналу зв'язку в БС;
- по-третє, прийняття рішення щодо активації послуги (6), базується на можливості забезпечення необхідної якості надання послуги, наявності коштів на рахунку та підписки на послуги;

- по-четверте, спеціальний механізм активації та впровадження правил обслуговування послуги.

Процес зображений на рис. 3.12 можна розглянути як процес надходження заявок на обслуговування в ТКС представлена на рис. 11.2 [40].

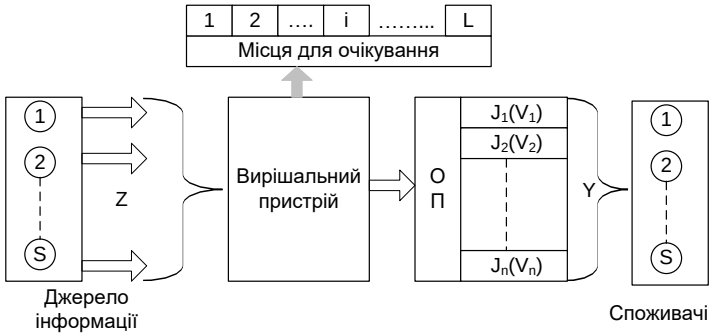


Рис. 3.12 Модель процесу надходження та обслуговування заявок в ТКС

Опис моделі. Нехай існують джерела інформації, що формують сумарну загрузку на ТКС. Інформація від джерел надходить у систему розподілення інформації, що забезпечує її прийом, зберігання, розподілення відповідно до адреси доставки. Система розподілення інформації включає розподільчий пристрій та місця для обслуговування. Обслуговуючі пристрої (ОП) розподіляються за групами – V_1, $V_2 \ldots V_i$ або по напрямкам – J_1, $J_2..J_i$ зв'язку.

Параметри моделі: кількість джерел (абонентів) - S, кількість напрямів зв'язку – I, кількість місць для очікування – L, максимальний час очікування обслуговування заявки – t; кількість ОП в *i-м* напрямі (групі), що розраховується:

$$\text{ОП} - V = \sum_{i=1}^{I} V_i \qquad (3.4)$$

Існують наступні способи обслуговування заявок: без втрат і очікування; з втратами; з очікуванням і необмеженою кількістю місць у черзі; з очікуванням та обмеженою кількістю місць в черзі; з формалізованим очікуванням. Способи обслуговування детально описані в [86]. У центрах управління мобільними мережами найчастіше використовуються способи з формалізованим очікуванням і з втратами.

Дисципліна обслуговування викликів або заявок визначає, за яких умов припиняється обслуговування заявок, як обирається для обслуговування наступна заявка, а також, що сталося з частково обслугованою заявкою. Розрізняють безпріоритетні і пріоритетні дисципліни обслуговування.

У випадку безпріоритетного обслуговування порядок обслуговування визначається за дисципліною вибору вимог з черги, наприклад FIFO, LIFO [86]. Під час пріоритетного обслуговування для кожної вимоги задається відповідний числовий параметр, значення якого визначає його пріоритет (рис. 3.13)

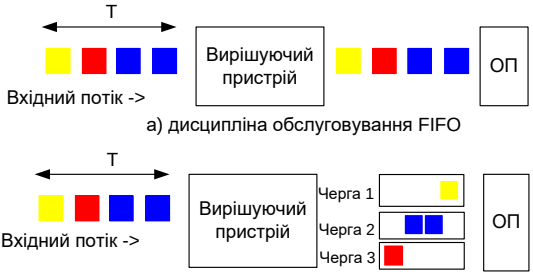


Рис. 3.13 Принцип роботи дисциплін обслуговування

Значення пріоритету може бути незмінними, статичним, або являти собою функцію, яка залежить від часу перебування пріоритету в системі – динамічний пріоритет.

Пріоритет також може бути абсолютним або відносним. Відносний пріоритет передбачає, що надходження вимог з вищим пріоритетом не перериває обслуговування менш пріоритетних. Вимоги з однаковими пріоритетами утворюють чергу. Якщо в системі задається абсолютний пріоритет, то поява заявки з більш високим пріоритетом перериває обслуговування менш пріоритетної заявки. У таких системах можуть утворюватися вкладені переривання. Перервані в процесі обслуговування заявки можуть або залишати систему, або знову ставати в чергу для додаткового обслуговування [86].

Сучасні системи обробки викликами використовують різні дисципліни обслуговування: FIFO, LIFO, з абсолютним пріоритетом, з відносним пріоритетом, з пріоритетом що є функцією параметрів – часу очікування.

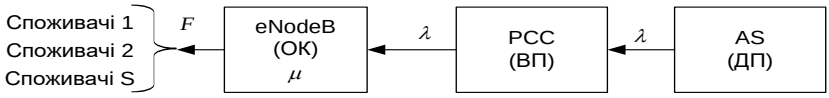
Реалізація дисципліни обслуговування є дуже важливим інструментом, за допомогою якого оператори зв'язку підвищують ефективність системи обробки викликів, класифікують абонентів у пріоритетні групи, наприклад:

- по номеру: відповідним номерам назначається фіксований пріоритет, наприклад номер VIP абонентів – 1, корпоративних абонентів – 2, загальних абонентів – 3;
- пріоритети на базі параметрів встановлення сесії: тип послуги (виклик екстрених служб – 1, дзвінок–2 тощо).

Такі пріоритети стали можливими в нових мобільних комутаційних центрах на базі технологій NGN і IMS. За допомогою програмного комутатора компанії Huawei Technologies Co.,Ltd., софтсвітч SoftX3000 [90].

3.2.2. Удосконалення дисципліни обслуговування в системі обробки викликів в мобільних мережах з ОЧД

Використаємо модель процесів надходження та обслуговування заявок в ТКС (рис. 3.14), для опису обслуговування викликів у мережі доступу LTE.



3.14 Модель обслуговування послуг в LTE з однією БС

Опис моделі: сервер надання послуг (AS) генерує з інтенсивністю λ виклик щодо надання послуг, які надходять на систему управлінням політиками обробки та тарифікації викликів (РСС), яка виступає в ролі вирішального і розподільчого пристрою. РСС перевіряє можливість організації послуг, формує черги щодо порядку обслуговування викликів згідно з дисципліною, і направляє виклик на обслуговування в БС (базова станція – eNodeB). БС формує обслуговуючі канали (ОК) на базі технології передачі в мережі доступу.

Припустимо, що умови роботи системи та надходження заявок наступні:

1. AS генерує виклики з інтенсивністю λ, які відповідають пуасонівському потоку.
2. Нехай на РСС від базової станції оператору надходить стаціонарний пуасоновський потік викликів.
3. РСС розраховує коефіцієнт *w*, що характеризує пріоритет виклику. РСС формує послідовність обслуговування викликів на БС згідно з пріоритетом *w*.
4. РСС має буферну пам'ять (з повнодоступним принципом організації) достатню для зберігання викликів, що надійшли за час *t*, буфер обнуляється з періодичністю *t*.
5. БС має n рівнозначних вільних каналів, кожен має фіксовану ємність *c,* і інтенсивність обслуговування μ. Процес звільнення каналів у БС в період t є пуасовніським процесом.
6. Для обслуговування виклику необхідно задіяти від 1-го до *n* каналів.

За рахунок введення додаткових параметрів – ширини смуги частот та інформаційної важливості послуги, пропонується нова удосконалена дисципліна обробки заявок, яка підвищує ефективність системи обробки викликів оператора у порівнянні з моделлю FIFO і моделлю з пріоритетами, за умови, що ймовірність доступності послуги не погіршується.

Удосконалення дисципліни полягає у введенні ситуаційного пріоритету, на основі якого проводиться вибір заявок на обслуговування – першою обирається заявка з найбільшим коефіцієнтом $_{\mathrm{W}}$. Новий запропонований коефіцієнт розраховується за формулою:

$$W_k = \frac{Tr_k}{\Delta f_k}$$

деTr_k– тариф послуги, Δf_k– необхідна ширина смуги частот для надання послуги в радіомережі з ОЧД.

Принцип обслуговування приведений на рис. 3.15.

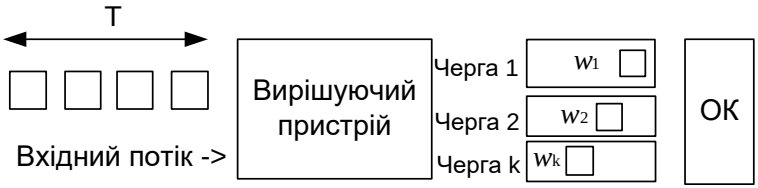


Рис. 3.15 Принцип обслуговування

Формулювання дисципліни обслуговування для моделі «RF» з одним обслуговуючим каналом :

1.Якщо за період від *0* до *t*, надійшло $N(t) > 0$– *заявок,* та ОК вільний, в першу чергу обслуговується заявка *j*, для якої справедливо $w_j > w_i$. Інші заявки отримують відмову.

2.Якщо за період від *0* до *t*, надійшло $N(t) > 0$ – *заявок,* і ОК зайнятий, заявка отримує відмову.

3.Якщо за період від *0* до *t*, надійшло $N(t) = 0$ – заявок, то система чекає час t , далі перехід до п.1, 2 або 3.

Формулювання дисципліни обслуговування для моделі «RF» з n обслуговуючими каналами :

1.Якщо за період [0;t] надійшло $k > 0$ заявок, в першу чергу обслуговується заявка *j*, де *j =1..k*, і для якої справедливо:

$$j^1(w_j) = \max\{1(w_1),..i(w_i),...,k(w_k)\} \quad (3.6)$$

при умові: $\Delta f_j \le F$, де *F* – загальна ширина смуги частот системи, кратна *c*; в другу чергу l^2 , для якої справедливо $l^2 = \max\{1(w_1),..i(w_i),...,k(w_k)\}$, де $l \in (1,k)$ та $i \notin (1,k)$, і т.д. поки в каналі є місце; інші заявки відкидаються.

2. Якщо за період *[0;t]* надійшло *k>0* заявок, і вільна ємність у каналі відсутня, тоді заявки отримують відмову.

3. Якщо за період *[0;t]* не надійшло заявок, то система чекає час t, після чого перехід до п.1,2 або 3.

Для доведення ефективності запропонованої дисципліни обробки використаємо критерій ефективності в СМО, введений автороми в [16, 22]. Функціонал, запропонований автором [16] і описаний далі, дозволяє мінімізувати або зменшити втрати в системі.

Використаємо критерій сумарної (інтегральної) інформаційної ваги обслужених викликів, що характеризує сумарну вагу всіх викликів, що пішли на обслуговування за такт t;

Задачу розрахунку інтегральної ваги обслужених викликів за дисциплінами «RF» та FIFO вирішимо в декілька етапів:

1) Розрахуємо математичне сподівання інтегральної ваги обслужених заявок за час t при дисциплінах обслуговування FIFO і RF, якщо вільний один канал і всі заявки, займають канал на весь час обслуговування;

2) Розрахуємо математичне сподівання інтегральної ваги обслужених заявок

при дисциплінах обслуговування FIFO і RF, якщо вільно *n* каналів, при умові, *що заявки з різними тарифами займають однакову кількість каналів*;

3) Розрахуємо математичне сподівання інтегральної ваги обслужених заявок при дисципліні RF, *якщо заявки з різними тарифами займають різну визначену постійну кількість смуг при обслуговуванні*;

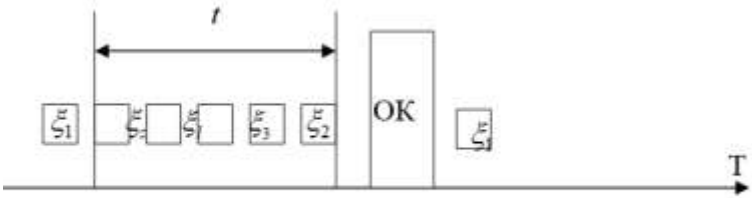
4) Розрахуємо математичне очікування інтегральної ваги обслуговування заявок при дисципліні RF, *якщо заявки з однаковим тарифом можуть займати різну кількість смуг при обслуговуванні.*

Задача зводиться до знаходження аналітичного виразу ймовірності обробки заявок в усіх чотирьох випадках.

3.2.2.1. Розрахунок сумарної ваги обслужених викликів в системі з одним каналом

Розрахуємо математичне сподівання інтегральної ваги обслужених заявок при дисциплінах обслуговування FIFO і RF, якщо вільний один канал і всі заявки, незалежно від тарифу, займають канал на весь час обслуговування.

Розрахуємо математичне очікування інтегральної ваги обслужених заявок при FIFO. Принцип роботи дисципліни FIFO зображений на рис. 3.16.



$$\xi_i \to (X_i, \lambda_i), i = 1..m, X_1 > X_2 > \ldots > X_m$$

Рис. 3.16 РСС з дисципліною обслуговуванням FIFO з одним вільним каналом

З умов задачі, процес надходження заявок – пуасонівський процес. Нехай інтенсивність надходження всіх заявок дорівнює $\lambda = \sum_{i=1}^{n} \lambda_i$, де λ_i– інтенсивність надходження *i-го* типу заявок, *i=0…m. Знайдемо математичне сподівання інтегральної ваги обслуженої заявки за один такт для моделі FIFO з одним каналом.* Нехай ξ – заявка, що надійшла першою за період часу *0* до *t*, тоді *математичне сподівання інтегральної ваги обслуженої заявки* ξ дорівнює добутку ймовірності, що заявки будуть, і першою буде заявка *i-го* типу, помножена на вагу X_i заявки:

$$M^1_{f_1 f_0}[\xi_1] = \sum_{i=1}^{m} X_i * P(t, N \geq 1, i) \quad (3.7)$$

де

$M^1_{f_1 f_0}[\xi_1]$– математичне сподівання інтегральної ваги обслуженої заявки для моделі FIFO з одним каналом;

m– кількість типів заявок;

X_i– вага *i*-ї заявки;

$P(t, N \geq 1, i$ – імовірності, що за період від *0* до *t* заявки надійдуть, і першою буде заявка *i-го* типу.

Ймовірність того що за період *0* до *t* заявки будуть і першою надійде заявка *i-го* типу – $P(t, N \geq 1, i)$, розраховується за формулою:

$$P(t, N \geq 1, i) = P(t, N \geq 1, i * P(t, i) \quad (3.8)$$

де

$P(t, N \geq 1, i)$ – ймовірність, що за період від *0* до *t* заявки надійдуть;

$P(t, i)$ – ймовірність, що першою надійде заявка *i-го* типу.

Ймовірність того, що за період за період від *0* до *t* заявки будуть дорівнює, для пуасонівського процесу, різниці між усіма можливими варіантами, тобто одиницею, та ймовірністю, що не буде жодної заявки:

$$P(t, N \geq 1) = 1 - e^{-\sum_{k=1}^{m} \lambda_k t} \quad (3.9)$$

де

λ_k– інтенсивність надходження заявок типу k з ціною X_k;

m – кількість заявок що надійшло за час від *0* до *t*.

Ймовірність того, перша заявка буде *i*-го типу дорівнює вибірковій ймовірності:

$$P(t, i) = \frac{\lambda_i}{\lambda} \quad (3.10)$$

де

λ_i – інтенсивність надходження заявок типу *i*;

λ – сумарна інтенсивність надходження усіх заявок.

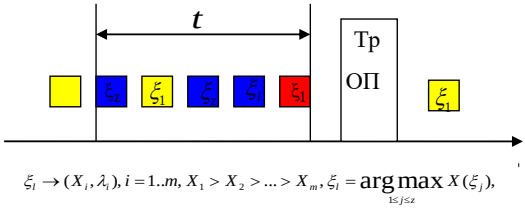
Таким чином, враховуючи вирази (3.9), (3.10) математичне сподівання інтегральної ваги обслужених заявок для моделі FIFO за 1 такт приймає вираз:

$$M_{f_1 f_0}^1 \left[\xi_1\right] = \sum_{i=1}^{m} X_i * 1 - e^{-\sum_{k=1}^{m} \lambda_k t}) \frac{\lambda_i}{\lambda} \quad (3.11)$$

Після ряду перетворень вираз (3.10), математичне очікування інтегральної ваги обслуговування заявок для моделі FIFO з одним каналом запишемо

$$M_{f_1 f_0}^1 \left[\xi_1\right] = \frac{1 - e^{-\lambda t}}{\lambda} \sum_{i=1}^{m} X_i * \lambda_i \quad (3.12)$$

На рис. 3.17 наведена графічна модель обслуговування заявок за *дисципліною «RF»*.



$$\xi_i \to (X_i, \lambda_i), i = 1..m, X_1 > X_2 > \ldots > X_m, \xi_l = \arg\max_{1 \leq j \leq z} X(\xi_j),$$

Рис. 3.17 РСС з дисципліною обслуговуванням «RF» з одним вільним каналом

Розрахуємо математичне сподівання інтегральної ваги обслуговування заявок $M_{RF}^1[X]$ для моделі «RF». $M_{RF}^1[X]$ дорівнює сумі добутків значень тарифу заявки X на ймовірність надходження заявок і на ймовірність, що не надійдуть заявки вищого пріоритету:

$$M_{RF}^1\left[\xi_1\right] = \sum_{i=1}^{m} X_i * P(t, N \geq 1, i) * P(t, N_1(t) = \cdots = N_l(t) = \cdots = N_{i-1}(t) = 0) \quad (3.13)$$

де

$M_{RF}^{1}[\xi_1]$– математичне сподівання інтегральної ваги обслужених заявок для моделі «RF»;

X_i– ціна *i*-ї заявки;

$P(t, N \geq 1, i)$– ймовірність надходження заявок, розраховується за формулою (3.9);

$P(t, N_1(t) = \cdots = N_l(t) = \cdots == N_{i-1}(t) = 0)$- ймовірність, що не надійдуть заявки вищого пріоритету.

Ймовірність, що не надійдуть заявки вищого пріоритету за проміжок часу від 0 до *t* дорівнює:

$$P(t, N_1(t) = N_2(t) = \cdots = N(i-1) = P(t, N_1(t) == 0) * \ldots * P(t, N_2(t) = 0) * \ldots * P(t, N_{i-1}(t) = 0) \quad (3.14)$$

Вираз (3.14) еквівалентний:

$$P\big(t, N_1(t) = N_2(t) = N_{i-1}(t)\big) = e^{-\lambda t} \quad (3.15)$$

Після математичних перетворень (3.15) приймає вид:

$$P(t, N_1(t) = N_2(t) = \cdots = N_{i-1}(t)) = e^{-t(\lambda_1+\lambda_2\ldots\lambda_{i-1})} \quad (3.16)$$

Тоді математичне очікування інтегральної ваги обслуговування заявок – $M_{RF}^{1}[X]$записується:

$$M_{RF}^{1}[X] = \sum_{i=1}^{m} X_i * (1 - e^{-\lambda t})\, e^{-t(\lambda_1+\lambda_2+\cdots\lambda_{i-1})} \quad (3.17)$$

Таким чином, знайдені вирази математичного очікування інтегральної ваги обслуговування заявок за один такт за дисциплінами FIFO і RF при умові, що вільний рівно один канал. У даному випадку модель RF еквівалентна пріоритетній дисципліні обслуговування моделі при умові: якщо $i \geq k$, тоді $w_i \geq w_k$.

3.2.2.2. Розрахунок інтегральної ваги обслужених викликів в системі з n каналами

Загальний потік надходження заявок *N(t)* – пуасонівський процес з інтенсивністю $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_m$, де *m* – кількість типів викликів. Інформаційна вага обслуженого виклику типу *i* дорівнює X_i і є постійною для всіх викликів *i*-го типу, будь-який виклику займає один канал для обслуговування .

Дисципліна обслуговування FIFO. Розрахуємо математичне сподівання інтегральної інформаційної ваги обслужених викликів за один такт при дисципліні обслуговування FIFO.

Принцип роботи системи з дисципліною обслуговування FIFO показаний рис. 3.18.

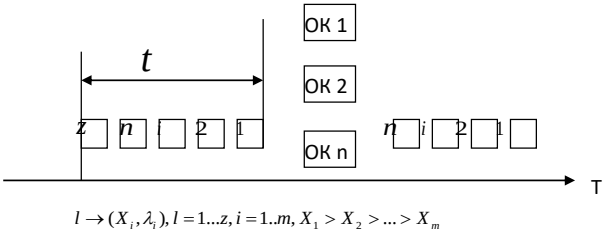


Рис. 3.18 РСС з дисципліною обслуговуванням FIFO, M/M/n

Нехай ξ_1 – інформаційна вага 1-го виклику, ξ_2 – інформаційна вага 2-го виклику, ξ_i – інформаційна вага *i*- го виклику, ξ – випадкова величина. Максимальна кількість викликів, що може буте обслужена за час *t* дорівнює кількості каналів в системі при FIFO.

Математичне сподівання інтегральної ваги обслужених заявок, що надійшли за час *t* для СМО з FIFO , розраховується як сума доданків множення математичного очікування інтегральної ваги від заявок типу *i, i=1..k* на ймовірність що за цей час звільниться *k* каналів, *k=1..n*:

$$M_{FIFO}^{n}[\xi_1 + \xi_2 + \xi_3 + \cdots + \xi_n] = \sum_{k=1}^{n} P_k(t) * \sum_{i=1}^{k} M[\xi_i] \quad (3.18)$$

де

$M_{FIFO}^{n}[\xi_1 + \xi_2 + \xi_3 + \cdots + \xi_n]$– сумарне математичне сподівання ваги від обслужених заявок за такт t при FIFO;

$M[\xi_i]$– математичне сподівання інтегральної ваги від заявок типу i, де $i = 1..n$,що надійшли за час$_t$;

$P_k(t)$– ймовірність того що за час $_t$ звільниться *k* каналів, де *k=1..n.*

Ймовірність $P_k(t)$ знаходиться як розв'язок системи рівнянь Колмогорова [91]:

$$P_k(t) = \frac{(n*\mu*t)^k}{k!} e^{-n\mu t} \quad (3.19)$$

де

μ– інтенсивність обслуговування одного каналу;

n – загальна кількість каналів;

t – час за який звільняться *k* каналів

Математичне сподівання ваги від заявки ξ_k дорівнює сумі всіх добутків тарифу виклику X_i на ймовірність того що *k*-й виклик має тип *i* і надійшов в період часу t. Розраховуються за формулою:

$$M[\xi_k] = \sum_{i=1}^{m} X_i * P(k_i; t) \quad (3.20)$$

де

X_i– тариф виклику *i*-го типу*;*

$P(k_i; t)$– ймовірність, що *k* –й виклик має тип *i* і надійшов в період часу t;

m – кількість типів викликів, що мають однаковий тариф.

Ймовірність $P(k_i; t)$ еквівалентна добутку ймовірності, що *k* –й виклик має тип *i,* і ймовірності, що *k* -й виклик надійшов в період часу *t*.

$$P(k_i; t) = P(i) * P(N(t) \geq k) \quad (3.21)$$

де

$P(i)$– ймовірність, що *k -й* виклик має тип *i*;

$P(N(t) \geq k)$- ймовірність, що *k* -й виклик надійшов в період часу *t*.

Ймовірність того, що *k*-й виклик має тип *i,* дорівнює відношенню інтенсивності надходження викликів типу *i* до інтенсивності усіх викликів, і розраховується:

$$\boldsymbol{P(i) = \frac{\lambda_i}{\lambda}} \quad (3.22)$$

Знайдемо ймовірність того, що *k*-й виклик надійшов в період часу *t*. Вираз $P(N(t) \geq k)$ еквівалентний різниці між одиницею (сумою ймовірностей всіх подій) та сумою ймовірностей, що не надійшли викликни 1,2,..,*i* ,*k-1* в період часу *t*.

$$P(N(t) \geq k) = 1 - \sum_{i=1}^{k-1} P(N(t) = i) \quad (3.23)$$

де

$P(N(t) = i)$– ймовірність того, що надійде *i* –й виклик в період часу t.

Після математичних перетворень вираз (3.23) приймає вид:

$$P(N(t) \geq k) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda_1 t)^i}{i!} e^{-\lambda_i t} \quad (3.24)$$

Враховуючи вирази (3.22) – (3.24) і провівши математичні перетворення, сподівання ваги від виклику *k*, де *k=1..n,* з ціною X_i для випадку *n* вільних каналів, при дисципліні обслуговування FIFO, розраховується за виразом:

$$M_{FIFO}[\xi_k] = \sum_{i=1}^{m} [X_i * \frac{\lambda_i}{\lambda}] * [1 - \sum_{i=0}^{k-1} \frac{(\lambda_i t)^i}{i!} e^{-\lambda_i t}] \quad (3.25)$$

де

$M_{FIFO}[\xi_k]$– математичне сподівання ваги від *k*-го,

X_i– вага від обслуговування виклику типу *i*,

λ_i– інтенсивність надходження виклику типу *i*,

λ – сумарна інтенсивність надходження викликів усіх типів,

t – час за який надійшов виклик типу *i.*

Сумарне математичне очікування ваги обслужених викликів при FIFO і n каналів, приймає вигляд:

$$M_{FIFO}^{n}[\xi_1 + \xi_2 + \xi_3 + \cdots + \xi_n] = \sum_{i=0}^{m} [X_i * \frac{\lambda_i}{\lambda}] * \sum_{k=1}^{n} P_k [1 - \sum_{i=0}^{k-1} \frac{(\lambda_i t)^i}{i!} e^{-\lambda_i t}] \quad (3.26)$$

Дисципліна обслуговування «RF». Різниця в обслуговувані порівняно з FIFO полягає в тому, що виклики йдуть на обслуговування в порядку пріоритетності, що визначається абсолютною величною тарифу виклику -X_i, *i* – кількість типів викликів з однаковим тарифом.

Знайдемо математичне сподівання інформаційної ваги від обслуговування викликуξ_k, який характеризується тарифомX_i. Нехай на обслуговування пішло *j* викликів (*j=1..n)*, що більш пріоритетні ніж викликξ_k. Виклик ξ_kбуде обслужений тільки у випадку, якщо кількість викликів типу менше *i* (*i* (N(t)<i)) прийшло менше *k*, і кількість пакетів типа *i* більше або дорівнює різниці між кількістю каналів (*k*) та кількістю пакетів типа менше *c*

Тоді, математичне сподівання інформаційної ваги від обслуговування викликуξ_k для «*RF*» дорівнює :

$$M^{RF}[\xi_k] = \sum_{i=0}^{m} X_i * \sum_{j=0}^{k-1} P(N(t))_{<1} = j, N(t)_{=1} \geq k - j \quad (3.27)$$

де$M^{RF}[\xi_k]$– математичне сподівання інтегральної від обслуженого *k*-го виклику;

$\sum_{j=0}^{k-1} P(N(t))_{<1} = j, N(t)_{=1} \geq k - j$– ймовірність, того що викликів типа *i* прийшло менше *k* , і кількість викликів типа *i* більше або дорівнює різниці між *k* та *j;*

j – кількість викликів типа менше *i*.

Зробивши відповідні математичні перетворення, отримуємо, що в даному випадку ймовірність події дорівнює добутку ймовірностей двох подій:

$$M^{RF}[\xi_k] = \sum_{i=0}^{m} X_i * \sum_{j=0}^{k-1} \frac{((\lambda_1 + \cdots + \lambda_{i-1}))*t)^j}{j!} (1 - P(N(t)_{-i} \geq k - j)) \quad (3.28)$$

де $N(t)_{<i}$ – кількість заявок типа менше *i*, що прийшли за час *t*, $N(t)_{=i}$– кількість заявок типа i , що прийшли за час *t*.

Вираз $P(N(t))_{<1} = j, N(t)_{=1} \geq k - j)$ можна записати як добуток ймовірностей:

$$P(N(t))_{<1} = j, N(t)_{=1} \geq k - j) = P(N(t))_{<1} = j * P(N(t)_{=1} \geq k - j) \quad (3.29)$$

де

$P(N(t))_{<1} = j$ – ймовірність, того що викликів типа *i* прийшло менше *k;*

$P(N(t))_{=1} = j$ – ймовірність, того що кількість викликів типа i більше або дорівнює різниці між k та *j;*

j - кількість викликів типа менше *i.*

Враховуючи вираз (3.24), залишимо математичне сподівання інтегральної ваги від обслуговування виклику:

$$M[\xi_k] = \sum_{i=1}^{m} x_i \sum_{j=0}^{k} \frac{\left((\lambda_1 + \cdots + \lambda_{i-1})\right) * t)^j}{j!} e^{-(\lambda_1 + \cdots + \lambda_{t-1})t} * (1 - \sum_{l=0}^{k-j-1} \frac{(\lambda_i t)^l}{l!} e^{-\lambda_i t}) \quad (3.30)$$

Інтегральне математичне сподівання інформаційної ваги для n каналів за принципом обслуговування RF приймає вираз:

$$M[\xi_1 + \xi_2 \ldots + \xi_n] = \sum_{k=l}^{n} P_k \sum_{i=1}^{m} x_i \sum_{j=0}^{k} \frac{((\lambda_1 + \cdots + \lambda_{i-1})) * t)^j}{j!} e^{-(\lambda_1 + \cdots + \lambda_{t-1})t} * (1 -$$

$$- \sum_{l=0}^{k-j-1} \frac{(\lambda_i t)^l}{l!} e^{-\lambda_i t} \quad (3.31)$$

Таким чином буле отримані аналітичні вирази для розрахунку математичного очікування при дисциплінах обслуговування FIFO і «RF» за СМО із моделлю M/M/n. Необхідно відмітити, що модель RF і модель з пріоритетами можуть мати однакові розрахункові вирази з припущенням, що кількість пріоритетів співпадає з кількості типів тарифів викликів.

3.2.3. Розрахунок інтегральної ваги обслужених викликів дискретної нефіксованої ємності в системі з n каналами

Постановка задачі. Нехай на систему надходить пуасонівський потік викликів $\{\xi_k\}$, де *k=1…N, N* – загальна кількість викликів, що надійшли за період *[0;t]*. Виклик ξ_k характеризується наступними параметрами: , де $\omega_i = \frac{X_i}{c_i}$, X_i – інформаційна вага виклику, c_i – наперед відома велична і визначає кількість каналів, яка необхідна для обслуговування *i*–го типу виклику, *i=1….m,* де *m* – кількість типів викликів; λ_i– інтенсивність надходження викликів *i*–го типу.

Нехай виклик ξ_k більш пріоритетний ніж виклик ξ_l. Дисципліна обслуговування “RF” полягає в тому, що в першу чергу відправляються на обслуговування виклики з коефіцієнтом ω_k потім ω_{k+1} і так далі, поки є канали.

Задача: розрахувати математичне сподівання середньої ваги обслужених викликів за час $_t$.

Задача вирішується в два етапи:

- допоміжна задача з припущенням: надходить один тип викликів, що займає фіксовану кількість каналів;
- основна задача: надходить *m* типів заявок, кожний тип заявок займає

фіксовану кількість каналів при обслуговуванні.

Допоміжна задача. Розглянемо випадок коли надходить тільки один вид заявок, $\xi^{(1)} = \xi^{(2)} = \cdots = \xi^{(k)} = \xi$ які при обслуговуванні займають тільки одну стільницю ресурсного блоку, тобто справедливо $\xi^{(k)} = const$. Загальна кількість місць для обслуговування (ОК) дорівнює g , рис. 3.19.

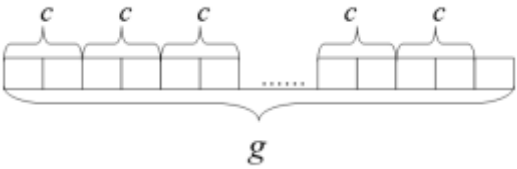


Рис. 3.19 Загальна кількість місць для обслуговування (ОК)

Ймовірність надходження k заявок за час t, що характеризуються пуасонівським потоком, розраховується за формулою:

$$P(N(t) = k) = e^{-\lambda_1 t} * \frac{(\lambda_1 t)^k}{k!} \quad (3.32)$$

Знайдемо ймовірність того, що буде оброблено всі k заявок, при $k < \left[\frac{g}{c}\right]$, де g – кількість ОК.

$$P_k^{(1)} = P(N(t) = k) = e^{-\lambda_1 t} * \frac{(\lambda_1 t)^k}{k!} \quad (3.33)$$

де λ_1 – інтенсивність обслуговування заявок в каналі.

Ймовірність обслуговування заявок при $k = \left[\frac{g}{c}\right]$ за час t:

$$P_k^{(1)} = P(\text{буде обслужено к заявок}) = 1 - \sum_{k=0}^{\left[\frac{g}{c}\right]-1} e^{-\lambda_1 t} * \frac{(\lambda_1 t)^k}{k!} \quad (3.34)$$

де

g – загальна кількість каналів для обслуговування;

c – необхідна кількість каналів для обслуговування однієї заявки.

Тоді математичне очікування інтегральної ваги $M_g^1[\xi]$ для одного типу заявок з фіксованим параметром $\omega_1 = \frac{X}{c}$ потоку, розраховується:

$$M_g^1[\xi] = p_g \sum_{k=0}^{\left[\frac{g}{c}\right]} k * c * \omega_1 * P_k^{(1)} = p_g \sum_{k=0}^{\left[\frac{g}{c}\right]} k * X * P_k^{(1)} \quad (3.35)$$

де p_g- ймовірність, що буде вільно g обслуговуючих каналів

Аналогічно розрахунки можна провести для будь-яких ω_k. Розглянемо наступний випадок $\xi^{(k)} \to c_k$ і будемо вирішувати задачу з кінця.

Основна задача. $M\left[X^{(k)}\right]$ - математичне сподіваняя ваги від заявок типа *k, k+1,..,m,* якщо вільно j каналів, де *j=0..n, n* – максимальна кількість каналів. Розраховуючи з кінця, спершу знайдемо $M\left[X^{(m)}\right]$, що розраховуються аналогічно формулі (3.35):

$$M\left[X^{(m)}\right] = \sum_{i=0}^{\left[\frac{j}{c_m}\right]} P_i^{(m)} * i * c_i * \omega_i \quad (3.36)$$

де

$$P_i^{(m)} = \begin{cases} e^{-\lambda_m t} * \frac{(\lambda_m t)^i}{i!}, i < \left[\frac{j}{c_m}\right] \\ 1 - \sum_{i=0}^{\left[\frac{j}{c_m}\right]-1} e^{-\lambda_m t} * \frac{(\lambda_m t)^i}{i!}, i = \left[\frac{j}{c_m}\right] \end{cases} \quad (3.37)$$

В даному випадку математичні сподівання інтегральної ваги $M\left[X^{(k)}\right]$ і $M\left[X^{(k+1)}\right]$ пов'язані формулою:

$$M\left[X^{(k)}\right] = \sum_{i=0}^{\left[\frac{j}{c_k}\right]} P_i^{(k)} * (i * c_k * \omega_k + M\left[X_i^{(k+1)}\right] * (j - c_k * i) \quad (3.38)$$

Сумарне математичне очікування обслуговування N заявок розраховується за виразом:

$$M\left[X^{(k)}\right] = \sum_{j=1}^{g} P_j \sum_{i=0}^{\left[\frac{j}{c_k}\right]} P_i^{(k)} * (i * c_i * \omega_i + M\left[X_i^{(k+1)}\right] * (j - c_k * i) \quad (3.39)$$

Вираз використовується при умові: $\omega_i > \omega_j, i < j$, тому існують обмеження при формуванні інформаційної ваги виклику, а саме

$\frac{X_i}{c_k} > \frac{X_j}{c_j}, i < j.$

Вирішення комплексної задачі в загальному виді. Нехай відомо ймовірність $P_{i,j}^{(k)}$ того, що за час t надійде k викликів і буде обслуговано i викликів, водночас звільниться j стільниць. Тоді математичне сподівання інтегральної ваги:

$$M_c\left[X_j^{(k)}\right] = \sum_{i=0}^{j} P_{i,j}^{(k)} * i * X^{(t)} \quad (3.40)$$

Ймовірність $P_{i,j}^{(k)}$ виражається як множення імовірності надходження k заявок на ймовірність що буде оброблено i заявок, і на ймовірність що звільниться j стільниць.

$$P_{i,j}^{(n)} = \sum_{k=0}^{N} P(N(t) = k) * P(\text{занято\_}i\text{\_із\_}k, \text{вільно\_}j) \quad (3.41)$$

Нехай

$$P_{i,j}^{(n)} = P(\text{занято\_}i\text{\_із\_}k, \text{вільно\_}j) \quad (3.42)$$

При умові, що

$$P_{i,j}^{(0)} = \begin{cases} 1, i = 0 \\ 0, \ i \neq 0 \end{cases} \quad (3.43)$$

Для знаходження $P_{i,j}^{(k)}$ розглянемо обслуговування заявок:

а) якщо перша заявка $l \leq i$, то вона потрапляє на обробку. *K* заявок що залишилися мають зайняти $i - l$, вільно $j - l$ місць.

б) якщо перша заявка $l > i$, то вона не потрапляє на обробку. Заявки що залишилися повинні займати *i* місць, тоді вільно *j* місць.

Ймовірність $P_{i,j}^{(k)}$ розраховується:

$$P_{i,j}^{(k)} = \sum_{l=1}^{i} P(\xi = l) * P_{i-l,j-l}^{(k)} + P(\xi > i) * P_{i,j}^{(k)} \quad (3.44)$$

Таким чином, знайдено аналітичні вирази розрахунку математичного сподівання інтегральної ваги при дисциплінах обслуговування FIFO, Priority, RF.

3.2.4. Ефективність використання дисциплін з відносними ситуаційними пріоритетами

При порівнянні пріоритетних дисциплін використовуються два основних критерії: економічної ефективності та еквівалентної продуктивності [16].

Критерій *економічної ефективності* порівнює втрати до і після застосування дисципліни. Такий виграш від використання дисципліни, також повинен перекривати витрати на введення більш складної дисципліни.

Нехай сумарні втрати в системі при першій дисципліні Z_1, при другій – Z_2. Тоді ефективність першої дисципліни порівняно з другою дорівнює відношенню:

$$L_{1,2} = \frac{Z_1 - Z_2}{Z_1} \quad (3.45)$$

Критерій *еквівалентної продуктивності.* Суть критерію полягає в наступному: на скільки більше необхідно використовувати технічних засобів для отримання однакових втрат при різних дисциплінах, припускаючи що ціни втрат для різних заявок однакові. Наприклад, якщо для першої пріоритетної дисципліни необхідно r_1 місць в черзі , а для другої r_2, то ефективність першої дисципліни порівняно з другою

$$L_{1,2} = \frac{r_1 - r_2}{r_1} \quad (3.46)$$

Даний критерій є ефективним при виборі технічних засобів, що управляють однорідними об'єктами. При управлінні різнорідними об'єктами його ефективності низька.

3.3. Процеси обробки викликів в мобільних мережах 4-го покоління

В даному розділі розроблено новий метод обробки викликів, який враховує ширину смуги частот при формуванні тарифу в мережах 4-го покоління з технологією ортогонального частотного доступу. Запропонований метод обробки заявок підвищує ефективність системи обробки викликів оператора при перевантаженні мережі.

3.3.1. Метод обробки викликів з урахуванням ширини частотної смуги

Тенденці розвитку систем обробки викликів та тарифікації є об'єднання їх в одну конвергентну архітектуру обробки викликів і тарифікації. Удосконалення методів обробки викликів впливає на гнучкість тарифікації та навпаки. Тому, архітектури систем управління правилами обробки викликів (PCC) [27] та методи обробки викликів потребують удосконалення.

Запропонований метод обробки викликів є унікальним. Підхід до розробки такого методу базується на проведеному детальному дослідженні параметрів та характеристик послуг в мобільних мережах 4-ого покоління/LTE.

Алгоритм роботи запропонованого методу обробки викликів приведений на рис. 3.21. Опис методу:

1) старт – абонент запитує доступ до послуги через термінал;
2) сервер вхідних запитів отримує SIP повідомлення користувача;

3) на основі SDP, що включає URI відбувається пошук адреси AS в SIP/IP таблиці маршрутизації, передача SIP запиту на AS;

4) AS(AF) перевіряє можливість адаптації контента (кодек) для сприймання користувачем, назначає ключ тарифікації та направляє запит до PCRF.

5) перевірка авторизації користувача на використання послуг; у випадку успішної перевірки – перехід до кроку 5, якщо абонент не авторизований, то кінець обслуговування виклику з повідомленням абонента про відмову;

6) формування політик обслуговування: пріоритету виклику, інформаціної важливості, методу тарифікації;

7) преконфігурація радіоканалу з метою перевірки можливості організації каналу передачі з заданою якістю, визначення ширини смуги частоти необхідної для передачі послуги; у випадку успішної процедури – перехід до кроку 6, якщо ресурси відсутні – відмова за недостачею каналів;

8) формування ситуаційного пріоритету, запит на активацію(впровадження правил осбулговування);

9) формування порядку обслуговування згідно ситуаційного пріоритету: в першу чергу обслуговуються виклики, що мають найбільший ситуаційний коефіцієнт, що дорівнює відношенню інформаційної ваги послуги до ширини смуги частоти, необхідної для надання послуги;

10) процес обслуговування послуг: передача даних від сервера до користувача;

11) кінець обслуговування ініціюється користувачем або мережею, у випадку неможливості продовжити обслуговування.

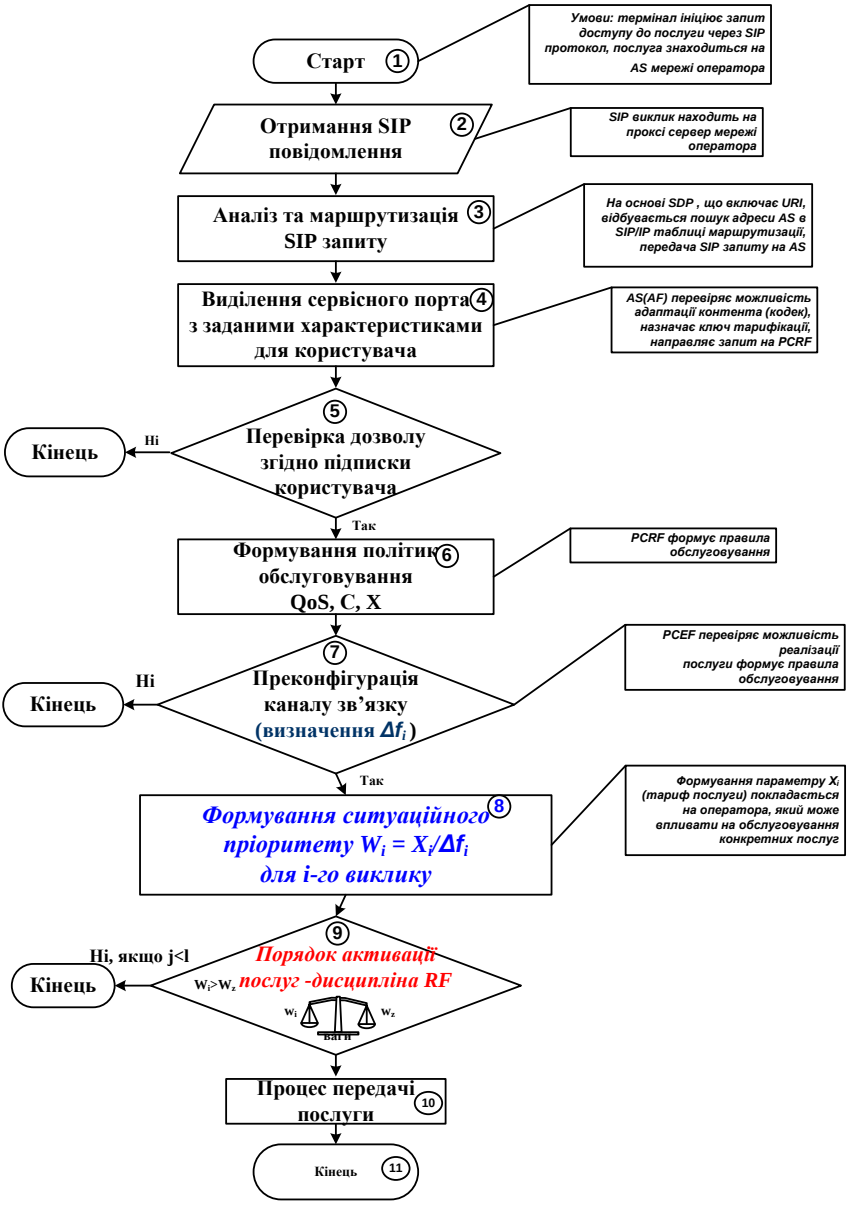


Рис. 3.21 Алгоритм роботи запропонованого методу обробки викликів послідовності процедур реалізації РСС

Запропонований метод обробки викликів враховує інформаційну вагу послуги та ширину смуги частот за домогою модифікації системи РСС. Для

реалізації методу необхідно провести модифікацію методу примусового управління якістю і тарифікацією вхідних запитів архітектури PCC. Для цього необхідно виконати наступні кроки:

1.Модифікувати процедуру перевірки ресурсу мережі для організації послуги з метою отримання параметрів попередньої конфігурації смуги ресурсу в каналі вниз і вверх для організації послуги. Це передбачає модифікацію сигнально протоколу GTPv2-C [69] ініціалізації виділення ресурсу радіосмуги.

2.Адаптувати протокол Diameter [60] для передачі додаткових параметрів в блок прийняття рішень щодо надання послуги та блок тарифікації. Для цього необхідно створити новий композиційний формат опису атрибутів послуги і передати відповідні параметри атрибутів в блок прийняття рішень PCRF.

3.Створити новий інтерфейс і протокол обміну інформацією між блоком прийняття рішень PCRF і блоком он-лайн тарифікації OSC. Створення інтерфейсу дозволить розраховувати тариф послуги на етапі прийняття рішення щодо надання послуги даному абоненту.

На рис. 3.22 представлені процедури методу управління правилами обробки і тарифікації викликів.

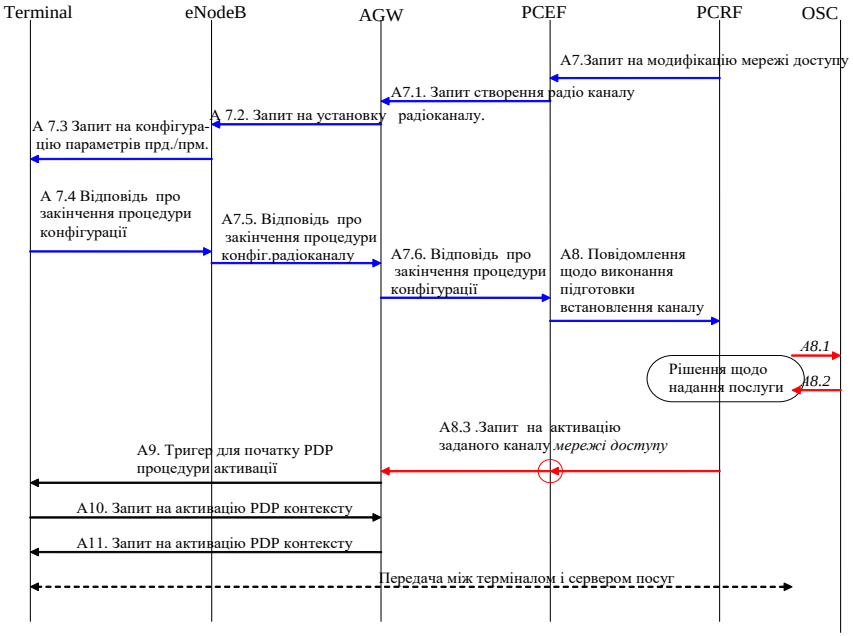


Рис. 3.22 Модифікована схема послідовності процедур реалізації PCC

Модифіковано дві процедури: процедуру перевірки наявності ресурсу в мережі доступу, RRC (Radio Resource Connection) – A7.1-A7.6, і процедуру активації послуги, що включає новий блок прийняття рішення щодо надання послуги і взаємодію з OSC (A8.1, A8.2). Введено нову дисципліну “RF” активації послуг в блоці впровадження політик обслуговування (A8.3).

Опис запропонованого методу обробки викликів і тарифікації.

Крок А7. PCRF відправляє запит Re-auth request по Diameter протоколу на перевірку можливості виконання політики обслуговування і тарифікації виклику, що включає перелік параметрів AVP, в тому числі новий розроблений композиційний AVP [10415: 450], який включає 14 параметрів: ширину смуги частот, кількість піднесучих, які необхідні для передачі послуги тощо.

Крок А7.1. PCEF приймає повідомлення *RAR,* Diameter протоколу, і на базі композиційного AVP [10415: 450] формує запит щодо створення радіоканалу до AGW.

Крок А7.2. AGW формує повідомлення-запит на встановлення радіоканалу - *Create Session Request з* параметрами *[Request accepted, F_d – ширина смуги частот вниз, Ncr_d – кількість піднесучих вниз, F_u – ширина смуги частот вверх, Ncr_u – кількість піднесучих вверх]* на базі модифікованого формату кадру протоколу GTPv2 до e-NodeB.

Крок А7.3 і А7.4. Не змінні – процедура вибору піднесучих і тестування каналу RRC.

Крок А7.5. E-NodeB формує повідомлення у відповідь – *Create Session Response, з* параметрами преконфігурації піднесучих зверху вниз і знизу вверх - *[Request accepted, F_d, Ncr_d, F_u, Ncr_u].*

Крок А7.6. AGW транслює повідомлення *Create Session Response* прозоро в PCEF.

Крок А8. PCEF формує відповідь AAR поверх *Gx* інтерфейсу, що містить данні нововведеного композиційного AVP, передає відповідь до PCRF.

Крок А8.1. PCRF отримує інформацію про підтвердження можливості організації радіоканалу з установленими параметрами ширини смуги частот вверх і вниз. PCRF формує запит RTI (Request tarif information), і за допомогою запропонованого протоколу DMShort та інтерфейсу *Ty* передає до OSC.

Крок А8.2. OSC отримує інформацію про сесію із нового композиційного AVP і, за допомогою вбудованої rating-функції, розраховує тариф послуги. OSC формує повідомлення ARP (Answer for tarif request), що включає тариф надання послуги, і пересилає до PCRF.

Крок А8.3 *Рішення щодо активації послуги.* PCRF приймає рішення щодо активації послуги і посилає повідомлення – активацію до PCEF. PCEF створює послідовність обслуговування викликів з використанням запропонованої дисципліни «RF».

3.3.2. Модифікація процедури преконфігурації ресурсу в мережі доступу

Дана процедура може бути розділена на дві: перша, процедура управління встановленням сесії, що реалізується в системі обробки викликів та тарифікації [27], друга, преконфігурація раідоканалу реалізується в системі радіомережі E-UTRAN. Модифікована процедура показана на рис. 3.23.

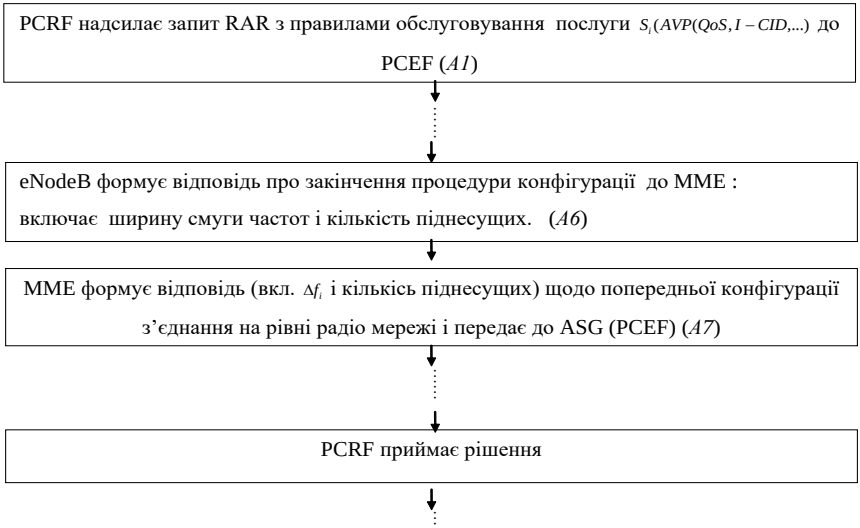


Рис. 3.23 Модифікований алгоритм попередньої конфігурації радіо мережі E-UTRAN

Модифікація потрібна в блоці *A6*: необхідно інкапсулювати параметри попередньої конфігурації, що були додані при процедурі RRR описаній (рис. 3.25) також параметри можуть бути отримані при процедурі тестування каналу зв'язку [51].

Ці параметри можна представити в бінарному вигляді (табл. 3.5).

Табл. 3.5

Представлення параметрів передачі радіо мережі в бінарному виді

Параметри	**Значення в десятковій системі**	**Значення в бінарній системі**
F_d, КГц	360	101101000
Ncr_d, одиниць	24	11000
F_u, КГц	180	10110100
Ncr_u,, одиниць	12	1100

Дані параметри пропонується передати через інтерфейс S1-U [54], що з'єднує MME і eNodeB. Для передачі інформації використати протокол GTPv2 (GPRS Tunnelling Protocol) [73], з додатковими технічними параметрами преконфігурації терміналу користувача і базової станції, визначені в модифікованому повідомленні в табл. 3.6.

Табл. 3.6

Модифіковане повідомлення – відповідь на встановлення сесії протоколу GPTv2

Відповідь Create Session Response, GPTv2 протоколу на запит Create Session Request	Відповідь Create Session Response, GPTv2 протоколу на запит Create Session Request	Приклад представлення в бінарному виді
Request accepted	Request accepted, F_d, Ncr_d, F_u, Ncr_u	{001}, {101101000, 11000, 10110100, 1100}

Передача сигнальної інформації між MME і ASG теж базується на протоколі GTPv2, тому ASG передасть нове повідомлення Create Session Response з параметрами [Request accepted, F_d, Ncr_d, F_u, Ncr_u]без змін. Діаграма роботи модифікованого протоколу представлена на рис. 3.24.

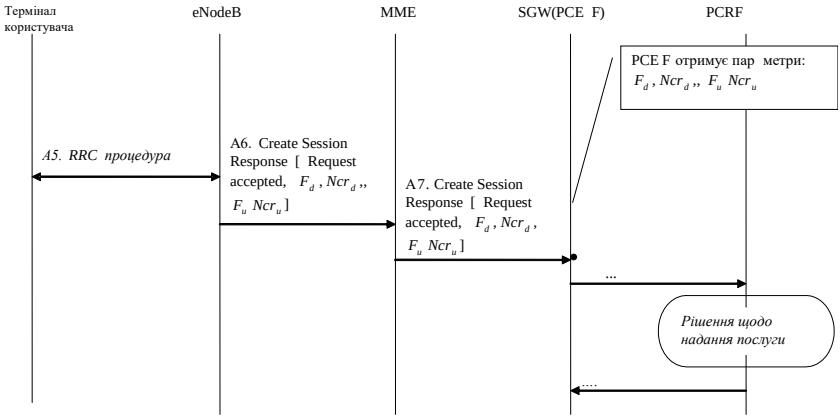


Рис. 3.24 Діаграма роботи модифікованого протоколу GTPv2

Для модифікації протоколу необхідні додаткові поля для передачі нових бітів інформації. Розрахуємо максимальну кількість додаткових бітів : максимальна кількість піднесучих в системі, що може бути виділена терміналу – максимальна кількість піднесучих, визначена системою радіо-інтерфейсів мережі E-UTRAN і дорівнює 1200, що еквівалентно в бінарній системі числення 10010110000. Таким чином додатково необхідно 52 біт або додаткових 7 байт. На рис. 3.25 представлений формат кадру протоколу GTPv2.

Октет	Біти							
	8	7	6	5	4	3	2	1
1	Версія			P	T	З	З	З
2	Тип повідомлення							
3	Тіло повідомлення октет 1й							
4	Тіло повідомлення октет 2й							
m до m+3	Якщо флаг T=1, то ідентифікатор кінця тунелю (TEID) повинен бути в 5-8 октетах. В іншому випадку TEID відсутній							
n до n+2	Номер послідовності							
n+3	З							

Рис. 3.25 Формат кадру протокол GTPv2

Повідомлення CSR [RA, F_d, Scr_d, F_u, Scr_u] в форматі кадру GTPv2 представлене на рис. 3.26.

Октет	Біти							
	8	7	6	5	4	3	2	1
1	GTPv2			P	0	З	З	З
2	Create session response							
3	F_d							
4	F_d							
5	F_d			Ncr_d				
6	Ncr_d					F_u		
7	F_u							
8	F_u				Ncr_u			
9	Ncr_u							

Рис. 3.26 Модифікований формат кадру протоколу GTPv2 для повідомлення CSR

Для введення додаткових байтів для передачі інформації в кадрі протоколу GTPv2 пропонується розширення заголовку протоколу. Перевагою GTPv2 порівняно з протоколом GTPv1 є можливість розширення полів даних, в тому числі заголовку. Враховуючи, що формат заголовку GTPv2 [73] змінної довжини, тому нове повідомлення – Create Session Response [Request accepted, F_d, Ncr_d, F_u, Ncr_u], далі CSR, може бути передано за допомогою розширення заголовка відповідно на 7 октет, що дорівнює 7 байт.

Транспортним протоколом для GTPv2 є IP [73] протокол. Розмір поля даних більше за 65000 біт, тому розширення поля даних протоколу GTPv2 для повідомлення CSR $[RA, F_d, Ncr_d, F_u, Ncr_u]$ не ускладнить передачу інформації наступному рівню. Якщо кількість піднесучих в системі фіксована і відома ширина

смуги частот однієї піднесучої, то формат повідомлення в кадрі можна скоротити до 22 біт, що еквівалентно 3 байтам рис. 3.27.

Октет	Біти							
	8	7	6	5	4	3	2	1
1	GTPv2			P	0	З	З	З
2	Create session response							
3	Ncr_d							
4	Ncr_u						Ncr_d	
5		Ncr_u						

Рис. 3.27 Модифікований формат кадру протоколу GTPv2 для повідомлення CSR(2)

Таким чином, реалізовано п.1 модифікації методу тарифікації, введенням нових параметрів в протокол передачі відповіді на встановлення з'єднання GTPv2, а саме: ширини смуги частот і кількості піднесущих при передачі «зверху-вниз» і «знизу-вверх». Для цього розраховано необхідну додаткову кількість біт: 52 для випадку чотирьох параметрів, 22 у випадку двох параметрів. Для розміщення додаткової інформації було модифіковано заголовок кадру протоколу GTPv2, що передбачає можливість розміщення додаткових октетів. Також має місце можливість роботи модифікованого протоколу GTPv2 поверх IP-протоколу. Модифікація протоколу дала можливість передати необхідні параметри, F_d, Ncr_d, F_u, Ncr_uдо функції впровадження тарифікації і якості послуги, для подальшої обробки.

3.3.3. Модифікація протоколу Diameter для передачі параметрів в блок прийняття рішень PCRF і блок он-лайн тарифікації OSC

В розділі 2 запропоновано новий метод обробки заявок, що підвищує ефективність системи обробки викликів оператора в мережах з ортогональним частотним доступом. Для реалізації методу необхідно змінити існуючий метод примусового (push) управління правилами і тарифікацією вхідних запитів архітектури PCC [27], що описані в розділі 1, підрозділ 9.3. По-перше, необхідно модифікувати запити RAR (4) і RAA (8) (рис. 3.28) таким чином, щоб запит RAR при впроваджені правил тарифікації та обробки викликів включав додаткові параметри F_d, Ncr_d, F_u, Ncr_u, а відповідь RAA включала значеннях цих параметрів, і передати додаткову тарифікаційну інформацію від PCRF до підсистем он-лайн і офф-лайн тарифікації. По-друге, створити протокол DMShort для онлайн розрахунку тарифу послуги з подальшим визначенням у коефіцієнта w_i. Вирішимо ці задачі послідовно.

3.3.3.1. Процедура протоколу Diameter для передачі додаткових параметрів в блок прийняття рішень PCRF

На рис. 3.28 проілюстрований модифікований метод примусового (push) управління правилами і тарифікацією вхідних викликів. Запити RAR (4) і RAA (8),

що працюють поверх інтерфейсу *Gx* [71]. Модифікація запитів з метою розширення запитуваних параметрів передбачає пошук додаткових AVP (Atribute Vaule Paired) [71] – параметри величин або характеристик послуги існування якостей, які отримують в результаті роботи протоколу Diameter, що побудований на принципі запит-відповідь

В результаті аналізу використання AVP, встановлено можливість використання спеціальних AVP, що зарезервовані для виробника(ів) обладнання[60].

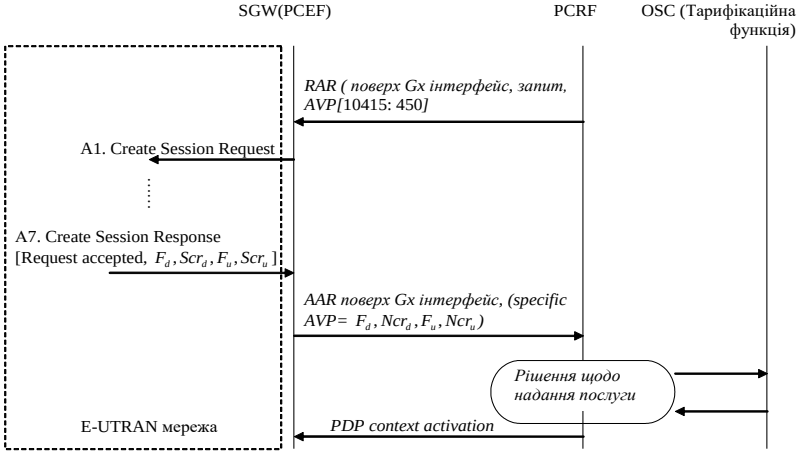


Рис. 3.28 Модифікований метод примусового управління якістю і тарифікацією вхідних запитів інтегрованої архітектури PCC

Код AVP Code і код виробника однозначно ідентифікують параметри AVP. Перші 256 номерів AVP мають код виробника – 0, і зарезервовані для сумісності з протоколом RADIUS [59], що використовується в інших системах, крім 3GPP. Згідно [60] коди спеціальних AVP можуть бути 10415: 400-499. Назначимо новому AVP код - 10415: 450, рис. 3.29.

32 Біти (x)	
AVP Код	
V M P x x x x x	Довжина AVP
Ідентифікатор виробника- Vendor-ID	
AVP дані: F_d, Ncr_d, F_u, Ncr_u	

Рис. 3.29 Формат нового AVP [10415: 450]

Параметри AVP, що використовуються Diameter протоколом, описані в [59, 60]. Використання нового композиційного AVP є кращим підходом порівняно з можливістю розширення попереднього AVP. Недоліком останнього є необхідність відслідковувати AVP, що передаються під час окремої сесії. Diameter протокол визначає наступні типи даних: Integer32, Unsigned32, Integer64, Unsigned64, Float32, Float64, Float128, OctetString і Grouped. Останній тип передбачає, що в полі даних

розташована група AVP.

Розроблений композиційний код AVP (табл. 3.7, що складається з 14 AVP параметрів, три з яких нововведені:

– тариф послуги, AVP: 10415: 452;

– конфігурація радіоканалу, визначається параметрами F_d, Ncr_d, F_u, Ncr_u, AVP: 10415: 453, 454;

– індекс лояльності, AVP: 10415: 455.

Табл. 3.7

Новий розроблений композиційний AVP

Номер	**Опис**	**Поле даних AVP**
1	Глобальний ідентифікатор сесії	ICID
2	Ідентифікатор рахунку користувача	Users account identifier
3	Ідентифікатор сесії	Session identifier
4	Ключ тарифікації сесії	Charging key
5	Індекс лояльності	Loyalty index
6	Опис субпослуг	Submedia description
7	Тариф послуги	Tariff
8	Клас якості послуги	QoS-Class-Identifier
9	Пріоритет надання послуги	Allocation-Retention-Priority
10	Тип радіо мережі	Radio Access Type
11	Конфігурації радіоканалу «вниз»	Radio Bear Session Downlink Setup Information
12	Конфігурації радіоканалу «вверх»	Radio Bear Session Uplink Setup Information
13	Швидкість передачі «вниз»	Guaranteed-Bitrate-DL AVP
14	Швидкість передачі «вверх»	Guaranteed-Bitrate-UL AVP

В результаті використання нового AVP, можна адаптувати Diameter протокол:

– запит *RAR*, що включає новий композиційний AVP[10415: 451] передає параметри 1, 2, 3, 4, 8, 13, 14, 15, із них параметри 11 і 12 нововведені відповідно до табл. 12.1;

– відповідь на запит, *AAR*, включає наступні параметри якості і тарифікації 1, 2, 3, 4, 7, 10, 11, 12, 13, 14, AVP$[F_d, Ncr_d, F_u, Ncr_u]$, і відповідно відповідь *AAR*.

Загальний формат адаптованого Diameter протоколу з командами *RAR* і *AAR*, що включає новий композиційний AVP[10415:451], наведено в табл. 3.10.

Табл. 3.10
Адаптація команд протоколу Diameter для передачі композиційного AVP
[AVP: 10415: 451]

Формат запиту тарифу - RAR:	Формат відповіді тарифу – RAA:
<RA-Request> ::= Diameter Header: 258, REQ, PXY >/ *команда протоколу діаметр* < Session-Id >/ *ідентифікатор сесії* { Auth-Application-Id }/ *ідентифікатор прикладної програми* { Origin-Host } { Origin-Realm } { Destination-Realm } { Destination-Host } { Re-Auth-Request-Type } **[Supported-Features: Vendor AVP=1]** [Session-Release-Cause] [Origin-State-Id] [Event-Trigger] [Event-Report-Indication] [Charging-Rule-Remove] [Charging-Rule-Install] [Default-EPS-Bearer-QoS] [QoS-Information] [Revalidation-Time] [Proxy-Info] [Route-Record] **[AVP:** 10415: 451**]**	<RA-Answer> ::= < Diameter Header: 258, PXY > < Session-Id > { Origin-Host } { Origin-Realm } **[Supported-Features Vendor AVP =1]** [Result-Code] [Experimental-Result] [Origin-State-Id] [IP-CAN-Type] [RAT-Type] [AN-GW-Address] [3GPP-SGSN-MCC-MNC] [3GPP-SGSN-Address] [3GPP-SGSN-IPv6-Address [RAT] [3GPP-User-Location-Info] [Charging-Rule-Report] [Access-Network-Charging-Address] [Access-Network-Charging-Indent-Gx] [Error-MesAGWe] [Error-Reporting-Host] [Failed-AVP] [Proxy-Info] **[AVP 10415: 451]**

Створений композиційний AVP може бути використаний для передачі інформації через інтерфейси *Gx, Gy, Sp, Rx, Ty* системи. Транспортним протоколом для Diameter є IP протокол, розмір поля даннях якого більше за 65000 біт, тому збільшення поля даних в повідомленнях протоколу Diameter не ускладнить передачу інформації між блоками. Адаптація протоколу дала можливість передати необхідні параметри F_d, Ncr_d, F_u, Ncr_u в блок прийняття рішень щодо надання послуги.

3.3.3.2. Розробка нового інтерфейсу Ty в системі обробки та тарифікації викликів

Для реалізації запропонованого методу обробки заявок необхідно розрахувати нововведений коефіцієнт w_i, що дорівнює відношенню тарифу послуги, вимірюється в грн., до ширини смуги частот, кГц, для впровадження послуги. За допомогою адаптованого протоколу Diameter і розробленого композиційного параметру, PCRF отримує параметри: ширину смуги частот і кількість піднесущих.

Для отримання параметра тарифу послуги, між PCRF і підсистемою он-лайн тарифікації, створюємо новий інтерфейс *Ty*, рис. 3.30.

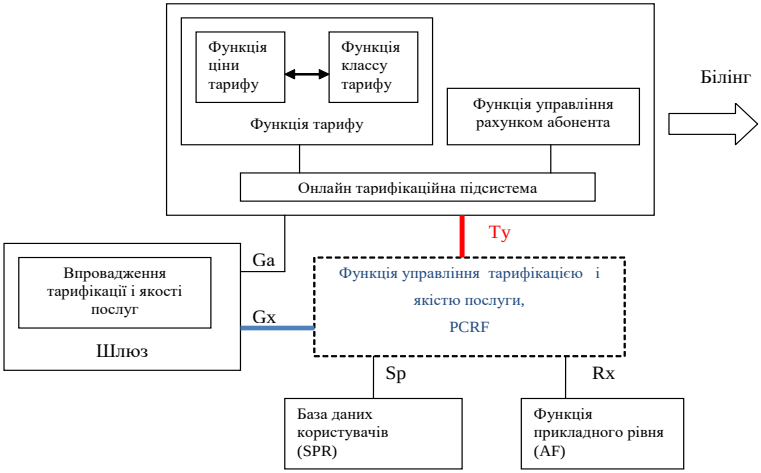


Рис. 3.30 Модифікована система обробки та тарифікації викликів

Функція інтерфейсу *Ty* – передача необхідних параметрів, визначених в *AVP [10415: 450]*, і повернення параметрів – розрахованого тарифу послуги, індексу лояльності . Інтерфейс пропонується реалізувати на новому протоколі DMSHORT, що базується на принципі запит – відповідь, і включає необхідні параметри. Вводимо дві команди запиту тарифу і відповіді – RTI (Request tarif information) і ARP (Answer for tarif request) відповідно. Робота нового протоколу показана на рис. 3.31.

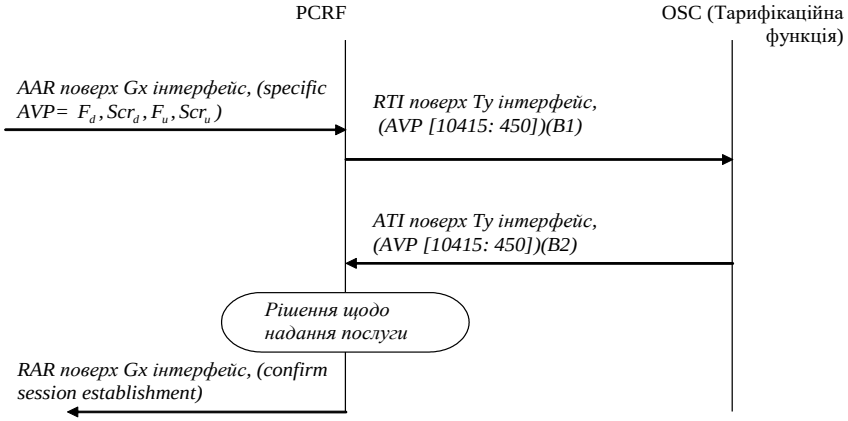


Рис. 3.31 Робота нового протоколу DMSHORT поверх Gy інтерфейсу

3.3.3.3. Модифікація інтерфейсу Gy для передачі параметрів радіо ресуру

Інтерфейс *Gy* визначає параметри що передаються до підсистеми он-лайн тарифікації з метою контролю рахунку абонента. Методи кредитного контролю абонента описані в [62, 63]. Для передачі параметрів *Gy* використовується Diameter протокол для контролю кредиту абонента [62]. Модифікована робота протоколу представлена на рис. 3.34. Команди CCR і CCA описані в специфікації [62], модифіковані записи, що працюють поверх інтерфейсу *Gx*. Модифікація запитів з метою розширення запрошуваних параметрів передбачає пошук додаткових AVP (AtributeVaulePaired) [60] – параметри величин або існування якостей, які можна отримати в результаті роботи протоколу Diameter, що побудований на принципі запит-відповідь.

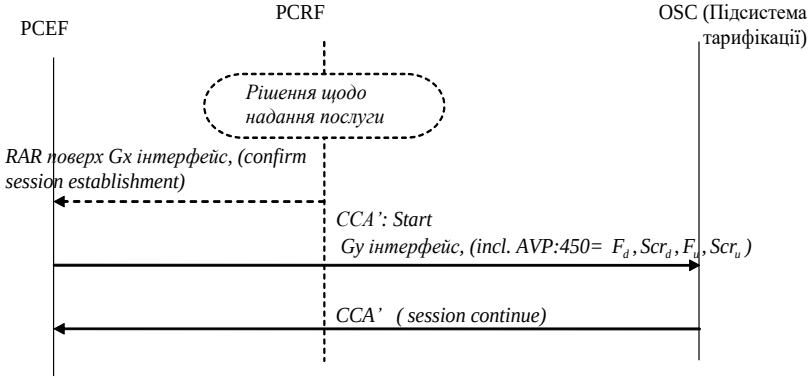


Рис. 3.32 Модифікований протокол передачі тарифікаційної інформації в підсистему OSC

За допомогою нового розробленого композиційного AVP450 передаємо інформацію про тарифікаційні події із блока PCEF до підсистеми тарифікації OSC. На основі даної інформації, оператор розраховує плату за використання послуги. Розрахунок плати базується на різних параметрах: часі використання послуги, величині частотного ресурсу, типу послуги, якості надання послуги, інше. Після, оператор перевіряє рахунок абонента і надає відповідь: продовження сесії або закінчення сесії.

3.3.3.4. Врахування обсягу частотного ресурсу при тарифікації послуг

В роботі [25] запропоновано формалізувати тарифікацію мультимедійної послуги за допомогою представлення її як набору простих послуг на рівні прикладної програми. Модифікований опис послуг з урахуванням ширини смуги частот субпослуги представлений в табл. 3.9.

Табл. 3.9

Модифікований опис послуг

Мультимедійна послуга A(t)	**Якість передачі**	**Вага субпослуги**	**Смуга частот**	**Тариф субпослуги**
СубпослугаS1(t)	**a**	**1**	F_1	X_1

СубпослугаS2(t)	b	3	F_2	X_2
…	…	…	…	
Субпослуга Si(t)	b	4	F_i	X_i
Субпослуга Sk(t)	d	2	F_k	X_k

Модифікований аналітичний опис тарифікації на рівні прикладної програми A(t)' яка надає мультимедійну послугу, включає набір субпослуг, якими користується абонент в різні моменти часу:

$$A(t) = w_1 \times S_1(t_{2,}F_2) + .. + w_i \times S_i(t_{i,}F_i) \ldots + w_k \times S_k(t_{k,}F_k) \quad (3.47)$$

де

$S_i(t_{i,}F_i)$- субпослуга типу, що входить до складу прикладної програми $A(t)$,

t_i– період часу користування субпослугою,

F_i– необхідна смуга частот для надання i -ї послуги

w_i– вага субпослуги i, k– кількість субпослуг в мультимедійній послузі;

Таким чином, з'являється можливість враховувати обсяг і вартість частотного ресурсу під час контролю і резервування коштів на рахунку абонента. Це дає подальший поштовх до удосконалення методів приведених в роботах [26-29].

3.4. Удосконалена система обробки викликів та тарифікації в мобільних мережах з IMS

З метою розширення спектру послуги і удосконалення якості розроблено систему IP Multimedia Subsystem. IMS-архітектура базується на багаторівневій схемі: мережа доступу і контролю передачі, рівень управління сесією, рівень управління послугами. IMS-рішення сучасних виробників підтримують різні технології доступу, такі як GPRS, UMTS/LTE, WiFi, xDSL/LAN, забезпечуючи уніфікацію засобів реєстрації і аутентифікації, реалізацію функцій управління сесіями, і виконує «безшовне» продовження сесії при зміні зон роботи різних технологій.

Більшість виробників рішення і обладнання є членами організацій 3GPP, 3GPP2, ETSI и ITU-T, в тому числі Huawei Technologies Ltd, Erricsson, Alcatel-Lucent. Наприклад, рішення IMS від Huawei Technologies Ltd відповідає вимогам і стандартам вище згаданих інститутів стандартизації.

3.4.1. Удосконалення процесів тарифікації мультимедійних послуг

Задачі тарифікації мультимедійних послуг розглядаються в роботах [15, 23–26]. У роботі [15] запропоновано формалізувати тарифікацію мультимедійної послуги за допомогою представлення прикладних програм набором простих послуг.

У [23] показано, що в процесі тарифікації треба враховувати, що послуга складається із субпослуг. У випадку, якщо в процесі надання послуг використовується он-лайн тарифікація, що потребує резервування коштів на

рахунку абонента, максимальна сума резервування коштів дорівнює максимальному платежу за використання послуги в IP мережі [15] та може бути розрахована за формулою (3.48):

$$Rs(A(t)) = \sum_{i=1}^{k} Xi(Q, S_i(t)) \quad (3.48)$$

Де Rs(A(t))– максимальна сума резервування коштів на рахунку абонента.

Даний підхід не враховує обсяг частотного ресурсу, що використовується в мережі LTE при наданні послуг.

Задача, поставлена в роботах [26–29], зводиться до контролю коштів на рахунках абонентів при он-лайн тарифікації. У роботах [27, 28] описується алгоритм здійснення контролю рахунку абонентів для мінімізації втрат оператора. Алгоритми TICA 1.0, TICA2.0 базуються на моделях статистичного прогнозування. У реальній ситуації складно спрогнозувати поведінку користувача при роботі зі складними прикладними програмами. Інше рішення задачі полягає в повідомленні тарифікаційної системи щодо змін в мультимедійній сесії користувача, що призводить до підвищення навантаження на систему тарифікації. В роботі [28] визначені оптимальні умови роботи тарифікаційної системи (3.49):

$$C_{tot} = E[C_c] + E[R_l] + \varepsilon \quad (3.49)$$

деC_{tot}– втрати на обслуговування послуг, $E[C_c]$ – втрати при контролі балансу рахунку користувача, $E[R_l]$– втрати через додаткове навантаження на систему при контролі балансу рахунку користувача, ε– випадкові загальні втрати системи.

Он-лайн тарифікаційна система працює оптимально при мінімальному значенніC_{tot}. Дилема полягає в тому, що при мінімізації $E[C_c]$ підвищується $E[R_l]$, що показано в роботах [72, 73].

Відмітимо, що вираз (3.49) справедливий для стаціонарного режиму системи тарифікації, тобто коли ресурсу мережі достатньо для обслуговування заявок, що надходять.

У новій системі обробки викликів і тарифікації, на блок PCRF покладені задачі формування пріоритетів і розв'язання конфліктів.

Рішення, щодо сповіщення системи тарифікації про всі тарифікаційні події базується на методах тарифікації послуг, що включають інтерфейси, протоколи та алгоритми, щодо збору, зберігання, форматування, тарифікаційної інформації. Активація тарифікаційної інформації будується на тригерних тарифікаційних подіях [73].

Методи тарифікації послуг є актуальною науково-технічною задачею, що висвітлюється в технічних специфікаціях тарифікаційних систем і телекомунікаційних систем управління [27, 71, 74].

Проведений аналіз міжнародних (європейських і американських) патентів [89], що відповідають міжнародній класифікації МПК, вияв патенти, в яких пропонуються нові або удосконалюються сучасні методи тарифікації в бездротових широкосмугових мережах [75–84].

У патенті [75], автор запропонував метод офф-лайн тарифікації, що враховує локальний час користувача(ів) при офф-лайн тарифікації. Необхідність такої інновації полягає в тому, що програмно-пакетний комутатор (в системах NGN або IMS), який встановлює з'єднання, обслуговує територіальну зону в більше, ніж один

часовий пояс. Метод включає отримання сигналу виклику від мережі доступу, що обслуговує термінал користувача. Сигнал виклику містить ідентифікатор мережі доступу. Далі, метод включає визначення інформації щодо розташування мережі доступу на основі ідентифікатору. На основі інформації розташування мережі доступу має місце встановлення локального часу користувача. Після цього, відбувається генерація тарифікаційного повідомлення для виклику та вставка часової мітки в повідомлення на основі визначеного локального часу користувача.

Тарифікаційне повідомлення, яке включає часову мітку, передається до біллінгової системи, що забезпечує гнучку тарифікацію і біллінг на основі реального часу абонента.

Винахід, описаний в [76], дозволяє гнучко, в залежності від інформації в базі даних користувача, створити конфігурацію для системи тарифікації відповідно до послуги, що надається користувачу. За винаходом, тип тарифікації, що використовується для відповідного типу послуги та абонента, зберігається в базі даних користувача на сервері. При виклику послуги сервер посилає інформацію – адресу тарифікаційної функції та тип тарифікації, до функції управління викликами. Остання визначає адресу тарифікаційної функції для відповідного серверу прикладної програми та перенаправляє адресну інформацію даному серверу. Метод реалізується для систем управління на базі IMS [75].

У роботі [76] запропонований розподілений метод тарифікації для систем управління мультимедійними послугами, що надаються в пакетних мережах. Метод дозволяє користувачам провести переговори щодо розподілення оплати компонентів мультимедійного сесії під час встановлення сесії або модифікації параметрів сесії. Метод включає передачу повідомлення до терміналу абонента на початку процедури виклику, що містить першу пропозицію співвідношення оплати послуги між користувачами, та отримання наступного повідомлення, яке включає змінену пропозицію співвідношення оплати послуги між користувачами. Також метод передбачає обробку другого повідомлення – прийняття або відмови від правил розподіленої тарифікації користувачами та системою управління послугами, дозвіл на подальше узгодження правил тарифікації між користувачами сесії.

3.4.2. Модифікована архітектура PCC в системі IMS

Узагальнена архітектура мережі операторського класу на базі IMS, розробленої компанією Huawei Technologies Co.,Ltd. представлена на Рис. 3.35. Всі компоненти IMS-системи можуть фізично розташовуватись в одній уніфікованій шафі. До складу шафи входять плати, що працюють на базі протоколу SIP з підтримкою прикладних послуг SIP AS, контролю управління мультимедійними ресурсами MRFC, і сервером баз даних Home Subscriber Server. Одна шафа обслуговує до 200 тисяч абонентів одночасно. Для повноцінного використання всіх можливостей IMS необхідна модернізація всієї інфраструктури з метою підтримки IP на всіх рівнях. До складу архітектури входять, зліва на право: мережі доступу кабельні, бездротові, оптичні, гібридні, задача яких полягає в організації останньої милі для кінцевих споживачів; рівень управління потоками інформації та доступом до послуг – NACF, RACK, CLF, SPDF, PCRF; ядро IMS системи – проксі сервер P-CSCF; модулі управління викликами – I/C–CSCF, HSS, MRF; платформи надання

послуг – IP Centrix, ATC, ENP (ігри, обмін інформацією, PushX); бізнес рівень включає платформу тарифікації (OSC), портал управління послугами абонента (Portal), інструменти управління рівнями рішення IMS (рис. 3.33).

До складу рішення IMS входить архітектура PCC, що включає блоки SPDF і PCRF. Функції цих блоків реалізуються на базі обладнання RM9000 Huawei Technologies Co.,Ltd.. RM9000 виступає ланцюгом, що зв'язує ядро IMS і мережу доступу, наприклад LTE, і реалізує гнучке управління послугами та їх тарифікацією. Функція RM9000 – це контроль виконання політик передачі сервісних потоків: відео, голосу, мультимедійних сесій, передачі даних. Політики (правила), що застосовуються для певного або всього набору типів сесій, створюються і активуються вузлом PCRF. Цей процес відбувається до тих пір, поки модуль контролю передачі здатен забезпечити відповідний рівень їх виконання. Вузол PCRF(RM9000) взаємодіє с блоком P-CSCF впродовж сесії, і відповідно реагує на дії користувача в середині сесії. P-CSCF включає Прикладну функцію(AF). AF працює як проксі сервер, запитує необхідний рівень QoS і тарифікаційної політики від терміналу. AF може розташовуватися в мережевому елементі, що ініціює з'єднання, наприклад платформі додаткових послугах. AF напряму звертається до PCRF, за допомогою Diameter протоколу і формує запит на встановлення сесії, що обробляється CSCF. В залежності від змісту сесії виклику, CSCF звертається до PCRF у випадку, якщо само CSCF визначило подію прийняття нового рішення щодо правил обслуговування послуги.

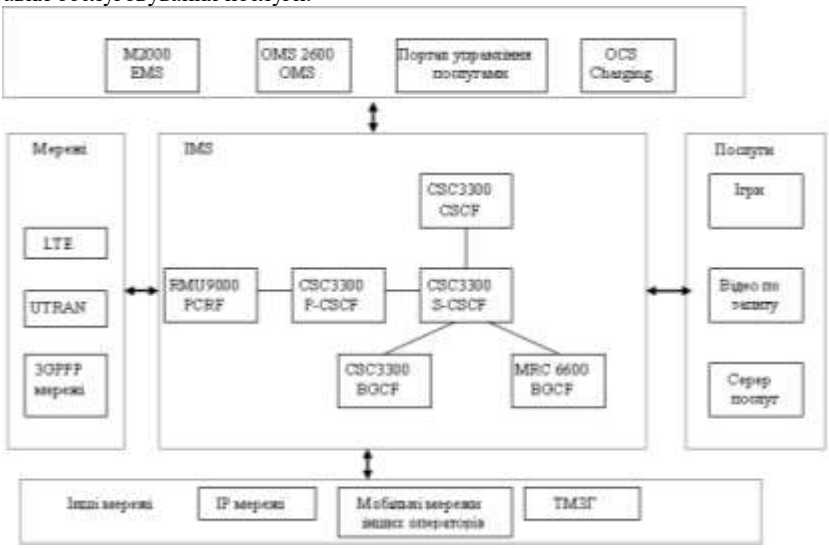


Рис. 3.33 Архітектура рішення IMS операторського класу

Запропонована модифікована архітектура наведена на рис. 3.36. В системі модифіковано зв'язок між рівнями контролю мережею управління введенням і системою тарифікації на бізнес рівні. Модифікація зв'язку *a* з метою отримання додаткового параметру споживання ресурсів від мережі радіо доступу при наданні послуг – ширини смуги частот.

Модифікована функціональна схема реалізації PCC для LTE наведена на рис.

3.34. На функціональному рівні показано:

– новий інтерфейс *Ty*, який введений для он-лайн розрахунку вартості на базі введених характеристик, для перевірки стану рахунку і передачі в блок PCRF для прийняття рішення щодо подальшого обслуговування послуги;

– удосконалення інтерфейсу *Gy* за рахунок введення додаткового комплексного універсального AVP в Diameter протокол для передачі в систему OSC інформації про споживання ресурсів;

– модифікація інтерфейсу *Gx* за рахунок введення додаткового комплексного AVP в Diameter протокол для передачі в систему OSC інформації про споживання ресурсів;

Рис. 3.34 Модифікована архітектура PPC для мережі LTE

– GTP-Cv2 з метою отримання певних параметрів, кількості піднесучих, у відповідь на запит створення преконфігурації і перевірки каналу для передачі інформації.

Обслуговуючий шлюз S-GW виконує роль управління радіомережею, включаючи резервування/модифікацію радіо ресурсу IP мережі.

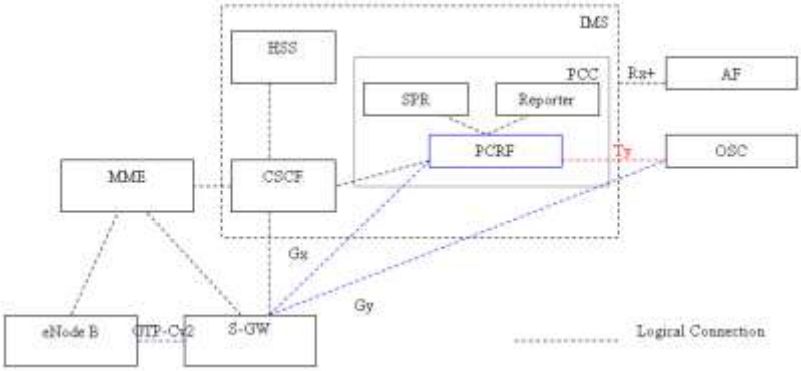


Рис. 3.35 Модифікована функціональна модель архітектури PCC для LTE

S-GW включає функцію PCEFs (Policy Control Enforcement Functions), що формує політики обслуговування послуг в радіомережі LTE. Функція PCEF отримує політики виконання з вузла PCRF за допомогою Diameter протоколу.

Функція «Reporter», реалізована в RM9000, збирає звітності і аналізує стан мережевих вузлів, формує аналіз даних для оператора . За допомогою такого аналізу оператор визначає політику обслуговування послуг і тарифікацію послуг. База даних SPR (Subscription Profile Repository) містить політики обслуговування послуг і тарифікаційну базу даних. SPR може бути реалізована разом з HSS в одній захищеній базі даних, за рахунок введення додаткової інформації – політики обслуговування послуг і методів тарифікації користувача розширенням бази даних.

В рамках пілотного проекту була модифікована програмна структура PRCF блоку, рис. 3.38, що дало змогу запровадити нову дисципліну обробки викликів. Нова дисципліна передбачає пріоритетну політику обслуговування на базі коефіцієнту w. Даний параметр розраховується як відношення тарифу послуги до ширини смугуи частот, необхідної для надання послуги.

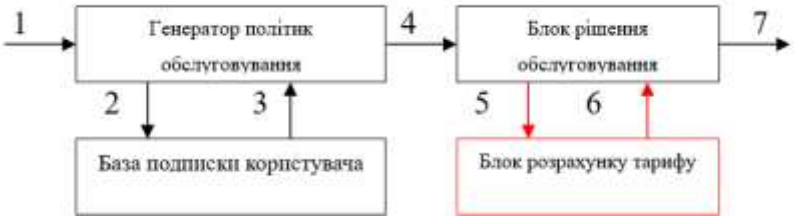


Рис. 3.34 Модифікована структура програмних засобів

Нова архітектура відрізняється від існуючої тим, що включає модуль розрахунку тарифу . Даний програмний модуль враховує ширину смуги частот при формуванні для тарифу для он-лайн послуг. Тарифікаційна функція T в даному випадку може бути записана:

$$T=F(t,\Delta f,S,C) \qquad (3.50)$$

Де t – час користування послугою, включає початок і кінець виклику, Δf – ширина смуги частот, S – тип послуги, C – ідентифікатор користувача, що впливає на тарифікацію послуги.

На рис. 3.35 представлена модифікована схема послідовності процедур реалізації PCC для LTE. Модифікований алгоритм управління викликами і тарифікацією:

Крок А1. RMU 9000 (блок PCRF) відправляє запит на перевірку можливості виконання політик якості RAR, що включає перелік параметрів AVP , в тому числі новий розроблений композиційний AVP [10415: 450].

Крок А2. S-GW (PCEF) приймає повідомлення *RAR*, і на базі композиційного AVP [10415: 450] формує запит створення радіоканалу.

Крок А3. S-GW формує повідомлення-запит на встановлення радіоканалу – *Create Session* .

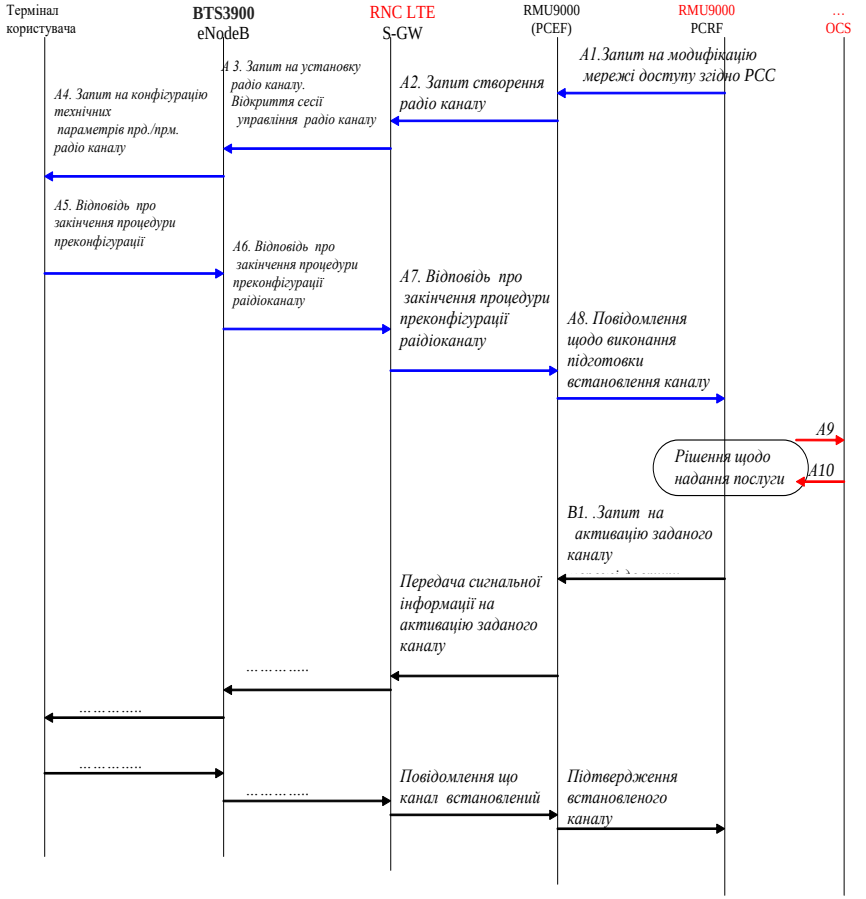


Рис. 3.35 Модифікована схема послідовності процедур реалізації PCC для LTE

Крок А4 і А5. Не змінні – вибір піднесучих і тестування каналу.

Крок А6. E-NodeB формую повідомлення у відповідь – *Create Session Response.*

Крок А7. Вузол MME є опціальним в даній схемі, присутній у разі необхідності підтримки високого рівня мобільності мережі. MME транслює повідомлення *Create Session Response* прозоро.

Крок А8. PCEF формує відповідь AAR поверх *Gx* інтерфейсу, що містить данні нововведеного композиційного AVP, передає відповідь до PCRF.

Крок А9. PCRF отримує інформацію про підтвердження можливості організація радіоканалу з параметрами ширини смуги частот вверх і вниз. PCRF формує запит RTI (Request tariff information), що належить до нового протоколу DMSHORT, і передає до OSC поверх нового *Ty* інтерфейсу.

Крок А10. OSC отримує інформацію про сесію із нового композиційного AVP і за допомогою вбудованої rating функції розраховує вартість послуги. OSC формує повідомлення ATR (Answer for tariff request), що включає тариф надання послуги, і пересилає до PCRF.

Таким чином, запропонований новий метод тарифікації, що враховує додатковий параметр – ширину смуги частот «зверху вниз» і «знизу вверх» при зборі та передачі інформації в систему тарифікації. Новий метод реалізований в рамках архітектури управління викликами і тарифікацією PCC для LTE, на прикладі обладнання RMU-9000 Huawei Technologies Co.,Ltd., і система тарифікації OSC. З метою реалізації даного методу в рамках рішення модифіковано стандартні протоколи: протокол резервування радіо ресурсів GTP-Cv2, модифіковано інтерфейси *Gx* і *Gy*, з метою передачі композиційного AVP. Створено новий інтерфейс *Ty*, що утворює взаємодію вузлів.

3.4.3. Розрахунок ефективності методу тарифікації

Для доведення ефективності запропонованого методу зроблений відповідний експеримент в рамках пілотного проекту тестування надання мультимедійних послуг в радіо мережах LTE на базі платформ IMS.

3.4.3.1. Опис методу збору даних про використаний ресурс

Експеримент включає стенд з апаратно-програмним забезпеченням. Користувачі ініціюють одночасно доступ до різних мільтимедійних видів послуг: голосовий виклик, відео конференція, послуги VOD з якістю HDTV і SDTV.

Дані про надання послуг отримані за допомогою функції Reporter, що виконує функцію реєстратора статусу послуг. Reporter інтегрована в блок PRCF (рис. 3.36) і має зручний Веб інтерфейс.

Початкові умови стану системи: система не обслуговує послуги. В короткий період 3–5 секунд на систему надходять запити щодо надання послуг. Необхідний ресурс системи: ширина смуги частот 5 МГц.

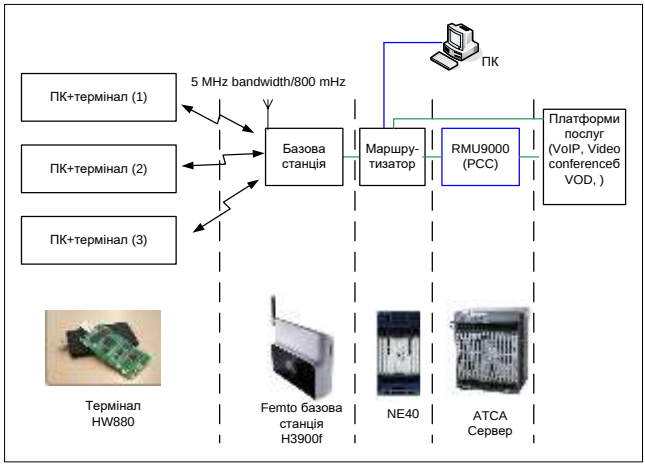


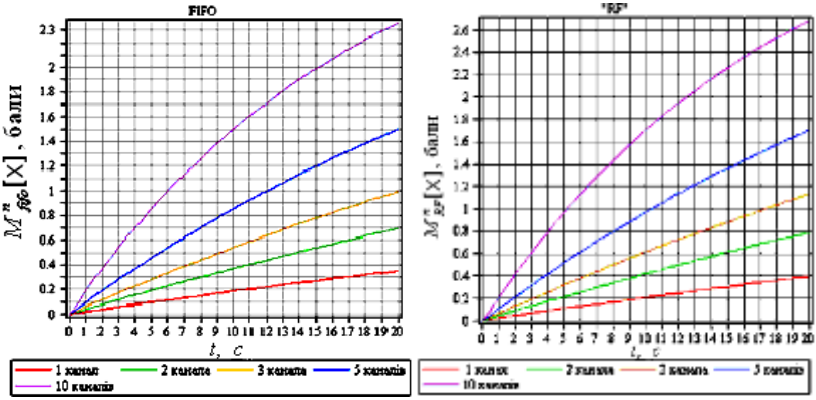
Рис. 3.36 Експериментальний стенд збору даних про послуги

3.4.3.2. Порівняння ефективності методу обробки викликів з дисциплінами RF, Priority та FIFO.

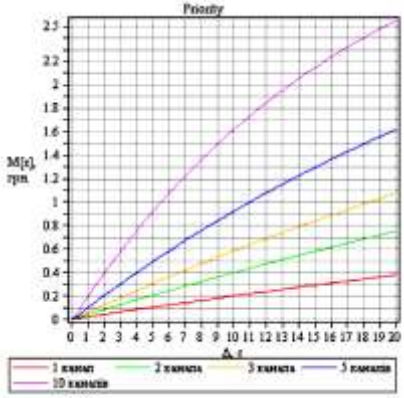
Графіки залежності математичного сподівання ефективності системи обробки викликів оператора при обслуговуванні викликів в залежності від часу очікування та від кількості каналів для дисциплін обслуговування FIFO, Priority та RF приведені на рис. 3.41.

З рис. 3.41 видно залежність інтегральної ваги обслуговування викликів від часу очікування та кількості каналів. Зрозуміло, що чім більше кількість обслуговуючих каналів (ширина смуги), тим більше викликів можна обслужити та отримати більшу сумарну інформаційну вагу.

На рис. 3.42 наведено відносне порівняння математичного сподівання дисциплін «RF» і FIFO, Priority.

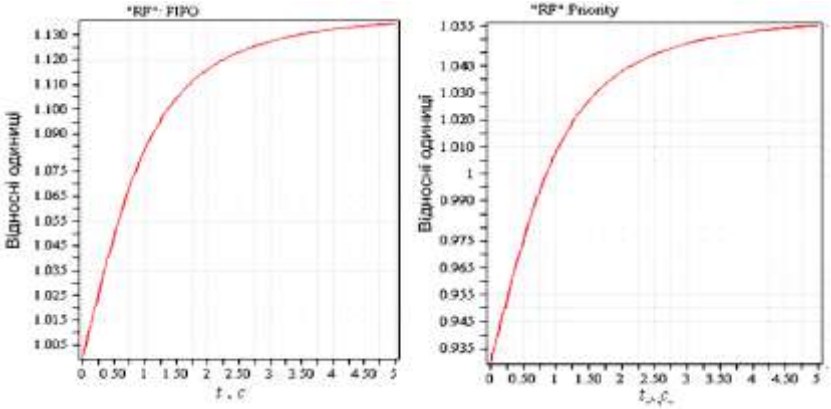


а) б)



в)

Рис. 3.37 Математичне очікування інтегральної ваги оброблених виликів оператора залежності від часу та від кількості каналів



а) б)

Рис. 3.38 Відносне порівняння математичного сподівання інформаційного ваги обслужених викликів при дисциплінах в залежності від часу очікування: а) - «RF» до FIFO , б) - «RF» з Priority

З рис. 3.38 видно:

– запропонована дисципліна «RF» дає виграш порівняно з дисциплінами FIFO та Priority відповідно на 13.% та 5.4% відсотки при обслуговуванні з часом очікування 5 с;

– швидкість росту виграшу зменшується із збільшення часу очікування. Виграш суттєво залежить від інтенсивності надходження викликів і середнього часу їх обслуговування.

На рис. 3.39 показана діаграма практично отриманих результатів сумарної кількості балів обслужених викликів.

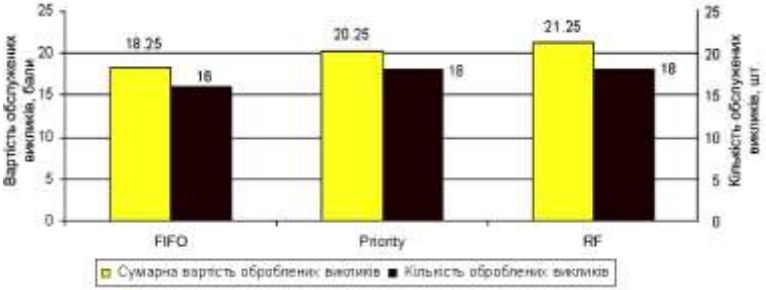


Рис. 3.39 Діаграма значення сумарної вартості обслужених викликів

З рис. 3.39 видно, що запропонована дисципліна «RF» дозволяє за визначений період обслужити заявки з більшим значенням сумарної вартості.

Таким чином, ефективність запропонованого методу та дисципліни підтверджується результатами практичних вимірів, що дозволяє зробити висновок про адекватність отриманих виразів та ефективність роботи системи в цілому.

3.5. Метод організації функціональних вузлів мережі LTE з EPC віртуалізацією

Трансформація сьогоднішньої мережевої інфраструктури в бік більш надійних, економічно ефективних рішень, що підтримує швидкі зміни, включає в себе два ключових елементи: по-перше, перехід від складних апаратних платформ до більш дешевих рішень, які підтримують функції зв'язку і, по-друге, перехід до програмно-керованої мережевої архітектури.

Віртуалізація EPC має потенціал, щоб реконструювати класичний функціональну архітектуру мобільних мереж [4].

В архітектурі Evolved Packet Core (EPC), яка є останньою архітектурою базової мережі для стільникової системи, приклади мережевих функцій, що можуть бути віртуалізовані, включають MME (Mobility Management Entity), S/P-GW (Serving/Packet Gateway) та інші.

В мультимедійній IP підсистемі (IMS), яка є архітектурою контролю сесій для підтримки постачання мультимедійних сервісів через EPC і інші мережі на основі IP, приклади мережевих функцій включають P-CSCF, S-CSCF та інші. HSS і PCRF — інші мережеві функції, які необхідні для повноцінної архітектури, щоб працювати у зв'язці з EPC і IMS для надання повноцінного сервісу, також можуть бути віртуалізовані. Аналогічним чином, об'єктами технології NFV можуть бути онлайн та офлайн системи тарифікації (OCS і OFCS) — системи, які охоплюють записи про тарифікації як частину управління сесією зв'язку.

Ядро мобільної мережі EPC може бути реалізована із віртуалізацією функцій Mobility Management Entity (MME), Serving Gateway (SGW), Packet Data Network Gateway (PGW). Віртуалізація IP Multimedia Subsystem в якості платформи, яка надає послуги в EPC та інших пакетних доменів можна також розглядати як об'єкт віртуалізації [2].

3.5.1. Network Functions Virtualization у сфері Evolved Packet Core

В [3] Network Functions Virtualization визначається як концепція мережевої архітектура, спрямованої на перетворення таким чином, що оператори зв'язку планують розвиток мережі із застосуванням стандартів віртуалізації програмного забезпечення, що вирішують питання об'єднання різних типів мережевого обладнання та стандартних потужних серверів, незалежно від точки їх розташування по мережі.

Програмне забезпечення для ядра мобільного зв'язку допоможе створити більш гнучкі послуги і радикально змінити інфраструктуру і функціонування мереж. Мережа віртуалізації, як очікується, надасть ключові переваги, такі як підвищення надійності зв'язку для високих навантажень та під час стихійних лих, а також у разі апаратного збою. Він також прискорить доставку нових послуг та гнучкість залучення інвестицій в інфраструктуру [5].

При віртуалізації архітектури EPC, функції IP Mobile Core можуть бути декомпоновані на віртуалізовані підфункції, які розміщуються на окремих віртуальних машинах. Різні типи віртуальних машин використовуються для виконання різних завдань віртуальних підфункцій.

Як показано в [6] Підхід до віртуалізації EPC архітектури повинен забезпечувати різноманітні аспекти, серед яких можна виділити наступні:

1.Мережеві функції EPC повинні мати хмарно-оптимізовану архітектуру, що дасть можливість найбільш повно реалізувати переваги архітектур NFV/SDN.

2.Масштабованість для забезпеченя максимальної операційної гнучкості при розгортанні віртуалізованих функції EPC у необхідному масштабі.

3.Підтримка потенціалу для послуг мобільного широкосмугового доступу для забезпечення ствлого рівня продуктивності для задоволення кінцевих потреб користувачів.

4.Підтримка кращої доступність і надійність у порівнянні з існуючим пакетного ядра мережі завдяки новим схемам захисту;

5.Підтримка загальних операцй управління мережевими функціями у всій виртуалізованій і фізичній площині EPC.

При цьому серед очікувані переваг від віртуалізації для мобільного ядра можна виділити наступні:

1.Покращена операційна ефективність: NFV інфраструктура буде збільшувати ефективність операційних процедур за рахунок зниження мережевих витрат і спрощених операцій.

2.Оптимізована конфігурація та/або топологія мережі за допомогою моніторингу продуктивності: автоматизоване підключення віртуальної машини і оптимізація можуть бути забезпечена використанням політики маршрутизації.

3.Підтримка так званої мульти-оренди: кілька мережевих функції можуть бути сконфігуровані на базі однієї й тієї ж NFV інфраструктури.

4.Скорочення часу виходу на ринок нових послуг: «хмарна» автоматизація дозволяє прискорити впровадження нових сервісів.

5.Введення цільових сервісів на основі географії або місцезнаходження клієнта.

В першу чергу, віртуалізоване ядро EPC, як правило, повинно бути розгорнуто паралельно основному EPC і бути зосередженим на сервісах типу

машина-машина та налаштованих сервісах рівня підприємства. Ці види послуг можуть мати певні профілі трафіку, які можуть надати бажані переваги з використанням конкретної конфігурації EPC (і, зокрема, конфігурації віртуального вузля PGW), і звісно ж таки надати можливість для операторів експериментувати з віртуальним ядром перед впровадженням технології до масового ринку послуг [7].

Як вже зазначалося вище, розгортання віртуалізованого EPC і пов'язаного з ним надання послуг потребує забезпечення складного оркестрування віртуальних ресурсів як усередині ЦОД, так і у фізичній мобільній транзитній мережі.

З точку зору організації елементів мережі, а зокрема елементів ядра, мережевий сервіс, що забезпечується роботою певного функціонального вузла (або групою вузлів) можна розглядати як граф передачі мережевих функцій (Network Functions, NFs), з'єднаних між собою за допомогою мережевої інфраструктури. Ці мережеві функції можуть бути реалізовані в одній мережі оператора або у взаємодії між різними мережами оператора. Логіка функціонування базової функції мережі обумовлює логіку функціонування сервісу вищого рівня. Отже, функціонування мережевого сервісу в цілому може бути представлене як поєднання логік функціонування блоків, що входять до її складу, які в свою чергу можуть включати окремі мережеві функції, набори мережевих функцій та/або мережеву інфраструктуру. Рис. 3.40 відображає основну схему цього підходу [10].

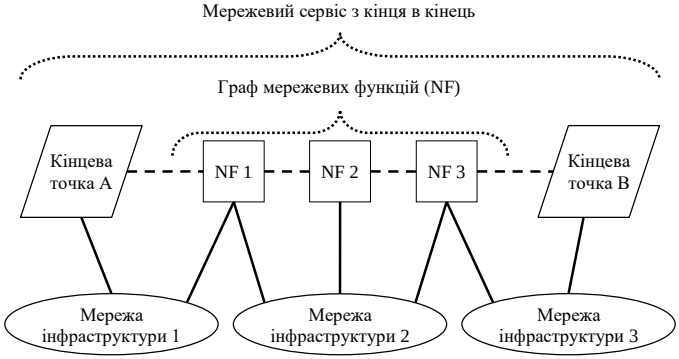


Рис. 3.40 Представлення у вигляді графу мережевого сервісу з-кінця-в-кінець

Сьогодні, більшість сервісів потребує використання декількох мережевих функцій. Оркестрування допомагає інтегрувати і використовувати віртуалізовані мережеві функції найбіль зручним і ефективним чином. Система використовує шаблони оркестрування сервісів, які визначають технологічний процес автоматизації для різних функціональних процесів. Окрім цього, архітектура NFV також може забезпечувати формування нового набору вимог до управління операціями в області оркестрування.

Головне питання в полягає у віртуалізації ядра мобільних мереж полягає в тому, якого рівня ефективності вдасться досягти, яка в свою чергу залежить від одного фактора – де будуть розміщені віртуалізовані функції. Впровадження віртуалізації в EPC в цілому повинно бути спрямовано на завдання підтримки необхідного рівня продуктивності та якості обслуговування (QoS).

Віртуалізація EPC передбачає [9], що архітектура NFV буде реалізувати одну або декілька віртуалізованих мережевих функцій (VNFs), а це означає віртуалізацію мережевої функції в області базової мережі без віртуалізації.

При цьому VNF може складатися з декількох внутрішніх компонентів і сама по собі автоматично не забезпечувати необхідний функціональний сервіс. Рис. 3.41 відображає загальну схему віртуалізації ресурсів із NFV, що запропонована у [9].

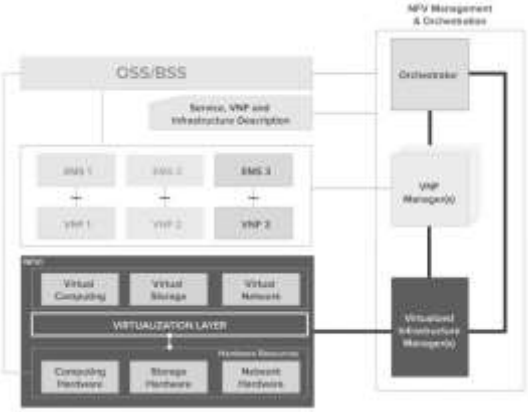


Рис. 3.41. Схема Network Functions Virtualization

Хмарна інфраструктура передбачає використання віртуальних машин (VM) для організації ресурсів віртуалізованих функцій, що дозволить застосувати методи для підвищення доступності ресурсів за допомогою механізмів управління, що автоматично застосовуються до екземплярів віртуальних сутностей в хмарній інфраструктурі шляхом використання найбільш ефективного ядра процесора, пам'яті та інтерфейсів, повторної ініціалізації та міграції віртуальних машин. Оскільки кожна віртуальна машина в контексті фізичних ресурсів працює відокремлено, а отже не залежить від інших віртуальних машин і не впливає на їх продуктивність. Ця особливість дає змогу за допомогою обраних інтерфейсів динамічно сконфігурувати ресурси ядра мережі (посилання, топологію мережі і т.д.), виходячи з факторів необхідної потужності і необхідних транспортних моделей.

3.5.2. Підхід до віртуалізації Evolved Packet Core

Основна ідея, яка пропонується у підході до віртуалізації EPC – це використання декількох менеджерів VNF для забезпечення оперативного контролю в точці розподілу віртуальних функцій в рамках загальної схеми. Зрозуміло, що, в такому випадку, для автоматизації оркестрування, всі операції налаштування, які було передбачено виконувати вручну тепер повинні бути розроблені зу рахуванням підтримки машинних форматів опису. Хмарна інфраструктура використовує описи для організації віртуальної мережі, для цього менеджерів VNF здійснюють моніторинг ресурсів в усіх областях і шарах функціональності.

Вище було описано формальний підхід, який пропонується використовувати при віртуалізації ядра мобільної мережі. Для ефективного

впровадження віртуалізації необхідно визначити, які функції доцільно віртуалізовувати при запропонованому підході. Для цього розглянемо схему організації функціональних вузлів ядра мережі LTE (рис. 3.42).

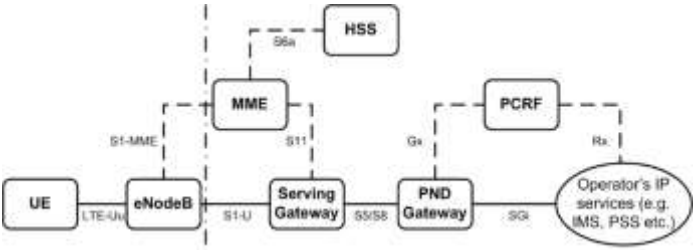


Рис. 3.42. Базова схема EPC

MME (блок управління мобільністю, Mobility Management Entity) – це основний елемент керування в мережі LTE. Він здійснює лише функції управління і не працює з даними користувачів. Має безпосередній зв'язок з терміналом користувача (UE, User Entity) через протокол сигналізації поза рівнем доступу (NAS, Non Access Stratum).

Функції, що виконуються MME, можна розділити на наступні два набори:

- управління потоком (на пре'явника управління);
- управління підключеннями (управління з'єднання).

Загалом же до функціоналу MME відноситься сигналізація між мережею EPC і UE, сигналізація хендоверу між різними мережами, вибір PGW і SGW, аутентифікація при реєстрації UE в мережі, управління каналами на інтерфейсах до інших елементів мережі.

SGW (Serving Gateway, обслуговуючий шлюз) призначений для обробки і маршрутизації пакетних даних надходили з/в підсистему базових станцій.

SGW маршрутизує і направляє пакети з даними користувачів, в той же час виконуючи роль вузла управління мобільністю (mobility anchor) для користувацьких даних при хендовері між базовими станціями, а також вузла управління мобільністю між мережею LTE і мережами з іншими технологіями 3GPP. Коли UE вільний і не зайнятий викликом, SGW підключає спадний канал даних (DownLink, DL) і виконує пейджинг, якщо потрібно передати дані по DL в напрямку UE. Він керує і зберігає стан UE (наприклад вимоги по пропускній здатності для IP-сервісів, внутрішню інформацію щодо мережевої маршрутизації).

PGW (Пакетний шлюз, Packet Data Network Gateway) забезпечує з'єднання від UE до зовнішніх пакетних мереж даних, будучи точкою входу і виходу трафіку для UE. UE може мати одночасно з'єднання з більш ніж одним PGW для підключення до декількох мереж. PGW виконує функції захисту, фільтрації пакетів для кожного користувача, підтримку білінгу, узаконеного перехоплення і сортування пакетів.

Виходя з типу функціонал вузлів сервісного і пакетного шлюзів може бути розділений на дві частини: площину управління і площину даних. Таке розділення дозволяє перемістити функції площини управління (SGW-Ctrl та PGW-Ctrl) до централізованої хмарної платформи. При цьому VNF менеджери будуть обслуговувати віртуалізовані функції SGW та PGW, що працюють на відокремлених

віртуальних машинах, і відповідати за оркестрацію життєвого циклу VNF. Рис. 3.43 схематично ілюструє підхід до організації EPC, що пропонується.

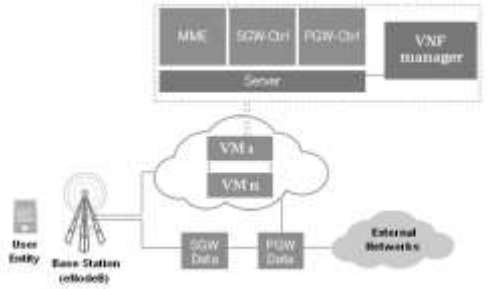


Рис. 3.43. Віртуалізація EPC з декількома менеджерами VNF

Підхід до організації EPC, який пропонується, передбачає, що робота мережі в цілому буде грунтуватися на декількох серверах або центрах обробки даних. При цьому, пропонується організувати SGW/PGW і MME функції на спільно на відособлених комплексах центрів обробки даних.

Взаємодія вузлів мережевих функцій забезпечується інкапсульовано стандартними інтерфейсами (s6s для взаємодії MME-HSS, s11 для MME-S/P-GW і s5 для окремих S/P-GW вузлів). Мережеві операції представлені на рис. 3.44.

Таким чином, підхід дозволяє масштабувати S/P-GW і мережеві функції MME відповідно до їх власних потреб в ресурсах. Це ефективно, коли, наприклад, може виникнути необхідність збільшити обсяг ресурсів у площині користувача, не впливаючи на рівень управління, і навпаки.

Також зазначимо, що при динамічній реконфігурації віртуальної функції мережі унаслідок збою або перевантаження з автоматичним або ручним режимом управління, сеанси та/або сесії управління повинні бути оброблені відповідним чином для досягнення необхідної надійності послуг.

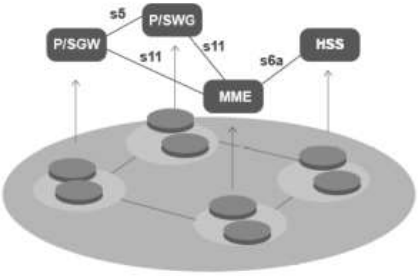


Рис. 3.44. Взаємодія мережевих блоків

3.5.3. Організація функціональніх вузлів мережі LTE з EPC віртуалізацією

PCRF (Policy and Charging Rules Function) – функціональний елемент в мережах зв'язку рекомендований групою 3GPP, основною функцією якого є впровадження політик обслуговування абонентів, наприклад, дозвіл/заборона сервісів або встановлення параметрів якості обслуговування (QoS), тобто застосування певних правил до корисного інформаційного потоку. Крім того, функціональний елемент PCRF встановлює правила тарифікації залежно від різних умов, таких як: параметри абонентського профілю, час доби, місць розташування абонента, обсяг спожитого трафіку та інших.

На сьогоднішній день функції PCRF реалізовані на програмно-апаратному забезпеченні, що обмежує можливості його масштабування. При цьому всі інформаційні потоки виявляються прив'язаними до топології мережі оператора, так як для забезпечення підрахунку кількості трафіку і для контролю якості обслуговування всі інформаційні потоки повинні взаємодіяти з блоком PCRF. Саме тому після появи концепції програмно-керованих мереж і віртуалізації мережевих функція, стало можливим застосування даних концепцій для удосконалення роботи функціонального блоку PCRF в рамках архітектури EPC.

Запропонований метод має на меті віртуалізацію деяких функцій PCRF, яка полягає в тому, що ряд функцій реалізується на базі віртуальних машин розміщених в хмарі. При цьому стає можливим масштабування самого PCRF при розгортанні мережі без необхідності організації окремих реальних фізичних серверних потужностей.

Організація взаємодії віртуальних машин PCRF з низкою вузлів мережі зв'язку є завданням, яке заслуговує особливої уваги. При цьому важливим питанням є розподіл функцій між віртуальною PCRF та іншими функціональними елементами системи 3GPP які пропускають через себе інформаційний потік.

Функціональні вузли мережі, з якими взаємодіє PCRF, обслуговують певний сегмент мережі, відповідно пропонується організовувати віртуалізацію вузлів, що відповідають за один сегмент на одному хмарному «сервері» для оптимізації взаємодії. Для оптимізації адміністрування функціональні вузли-контролери, що обслуговують різні сегменти мережі будуть організовані на окремих віртуальних машинах. Саме вузли-контролери функціональних елементів PCRF будуть здійснювати взаємодію з іншими елементами архітектури сервера оператора.

Розглянемо детально функціональні елементи і інтерфейси, з якими взаємодіє PCRF.

Перший елемент, PCEF (Policy and Charging Enforcement Function) — функціональний елемент в 3GPP мережах зв'язку, який здійснює застосування PCC-правил, отриманих від PCRF, до трафіку, що проходить через нього. Здійснює тарифікацію цього трафіку в системі тарифікації оператора зв'язку OCS/OFCS.

Даний компонент в якості функціонального блоку PGW виносится в хмарну інфраструктуру. Взаємодія хмарного компонента PCEF у складі PGW-Control з PCRF здійснюватиметься за Gx-інтерфейсом, який використовує протокол Diameter і призначений для надання службових даних по реалізації Flow Based Charging — FBC правил білінгових розрахунків з абонентами. По цьому інтерфейсу PCEF передає на PCRF інформацію, необхідну для прийняття PCC-рішень: ідентифікатор

абонента, інформацію щодо місцезнаходження і часового пояса, в якому знаходиться абонент, IP-адресу пристрою, з якої здійснюється доступ, параметри каналу, та інші.

На реальному фізичному обладнанні буде працювати варіант «тонкого клієнта» PGW – PGW-Data, завдання якого полягає у фільтрації трафіку і застосуванні правил в залежності від інструкцій, одержуваних від PGW-Control. Взаємодія клієнта і хмарної реалізації PGW здійснюється за допомогою інтерфейсу s5.

Функції фільтрації пакетів по користувачам і законного перехоплення трафіку (позначимо даний набір функцій як F1) здійснюються на боці «тонкого клієнта» PGW, в хмарну інфраструктуру виноситися функціонал розподілу пулу IP-адрес для пристроїв користувачів UE (F2).

Призначення PCC-правил, якими обмінюються між собою елементи мережі PCRF і PCEF – поділ фізичного потоку даних (IP-CAN) на логічні сесії SDF (Service Data Flow), визначення того до яких додатків і послуг відноситься трафік, надання параметрів QoS та інформації для тарифікації. Використовується два типи PCC-правил: динамічні PCC-правила, які передаються з PCRF на PCEF через Gx-інтерфейс і визначені на PCEF. Визначені правила можуть бути активізовані або PCRF, або самим PCEF.

Наступний елемент, який буде віртуалізовано – BBERF (Bearer Binding and Event Reporting Function). Це функціональний елемент в 3GPP мережах зв'язку, який здійснює нотифікацію PCRF про встановлення сесії з посилкою ідентифікатора абонента і додаткових параметрів для коректного визначення QoS-правил обслуговування. Функціонал даного компонента буде винесено в хмарну інфраструктуру в якості функціоналу вузла SGW-Control.

«Тонкий клієнт» SGW-Data буде виконувати наступні функції: базову маршрутизацію пакетного трафіку і перехоплення пакетного трафіку (F3), а також функціонал «якірної» точки (точки об'єднання трафіку) для хендовера між базовими станціями NodeB всередині однієї мережі доступу в зоні обслуговування базових станцій згідно набору правил та інструкцій (F4), що надходитимуть від хмарної «серверної» частини SGW-Control.

Елемент SGW-контроль в свою чергу забезпечує виконання таких функцій: «якірна» точка для хендовера між різними мережами доступу стандартів LTE/LTE і LTE/UMTS (F4), і обробка функціоналу BBERF (F5).

Взаємодія з PCRF здійснюється вузлом SGW-Control (також як і у випадку з PGW-Control), але по інтерфейсу Gxx. Взаємодія між компонентами SGW-Control і SGW-Data здійснюється по інтерфейсу s5.

Існує, також, ряд вузлів, функціонал яких пропонується розміщувати в хмарній інфраструктурі (з поділом сервером за сегментами мережі). Серед них можна виділити:

TDF (Traffic Detection Function) (F6) – функціональний елемент в 3GPP мережах зв'язку, який здійснює визначення трафіку певних додатків і нотифікацію про нього PCRF. Залежно від отриманих правил здійснює пропуск даного трафіку абоненту, перенаправлення і обмеження швидкості. Взаємодіє з PCRF по інтерфейсу Sd, який використовується для встановлення ADC (Application Detection and Control) правил керування параметрами трафіку конкретних програм.

UDR (User Data Repository) (F7) – функціональний елемент, який здійснює зберігання даних користувачів. Взаємодіє з PCRF по інтерфейсу Ud, який використовується для отримання/зміни профілів, в яких зберігається інформація про сервіси, що доступні абоненту, параметрах QoS та інших, необхідних для прийняття PCC-рішень. Також інтерфейс Ud використовується для організації передплати та отримання нотифікацій про зміни у профілях абонентів.

AF (Application Function) (F8) – функціональний елемент, який надає опис потоку даних сервіса і здійснює інформування про необхідні сервісу ресурси. Взаємодіє з PCRF по інтерфейсу Rx.

OCS (Online Charging System) (F9) – сервер кредитного контролю в режимі реального часу, який здійснює тарифікацію послуг, контролює баланс абонента, обробляє інформацію про нарахування та списання коштів на балансі абонента, застосовує знижки, здійснює підрахунок обсягу спожитих послуг. Взаємодіє з PCRF по інтерфейсу Sy, який використовується для обліку обсягів спожитих послуг і нотифікації про подолання порогів лічильників з OCS на PCRF. Крім PCRF, OCS взаємодіє з PCEF по інтерфейсу Gy за допомогою якого здійснюється тарифікація послуг.

На рис. 3.45 представлена запропонована схема організації функціональних вузлів мережі (функції, що відмічені червоним кольором віртуалізовані в «хмарну» інфраструктуру).

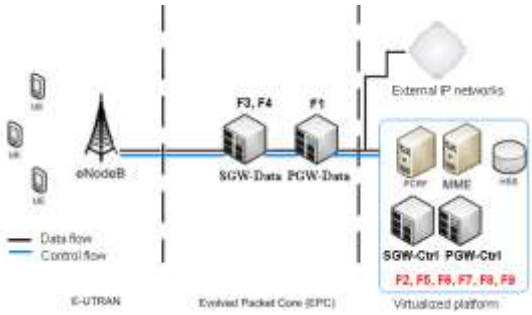


Рис. 3.45. Організація мережевих функціональних вузлів

3.5.4. Передача повідомлень управління при віртуалізації EPC

У ході взаємодії компонентів мережі LTE мова йде про логічні канали, які надають послуги середнього рівня управління доступом MAC (Medium Access Control) в межах структури протоколу LTE. Логічні канали по типу інформації, що передається, діляться на логічні канали управління та логічні канали трафіку. Логічні канали управління використовуються для передачі різних сигнальних та інформаційних повідомлень. По логічним каналам трафіку передаються дані користувачів.

Підхід до віртуалізації EPC, що пропонується, дає можливість: по-перше, спростити організацію логічних каналів, так як при даному варіанті організації, по

суті, існують тільки логічний канал трафіку UE-Base station-SGW-Data-PGW-Data з виходом на пакетні мережі для обміну даними між різними ділянками мережі і аналогічний логічний канал управління, який використовує вихід на пакетні мережі для організації взаємодії з хмарною віртуалізованою архітектурою; по-друге, максимально уніфікувати обладнання, що здійснює фільтрацію трафіку (SGW і PGW) за допомогою обмеження його функціоналу; і, по-третє, надає можливість резервування серверів віртуалізованих функцій.

Рис. 3.46 ілюструє схему встановлення підключення пристрою користувача до мережі LTE при реалізації методу організації ресурсів EPC.

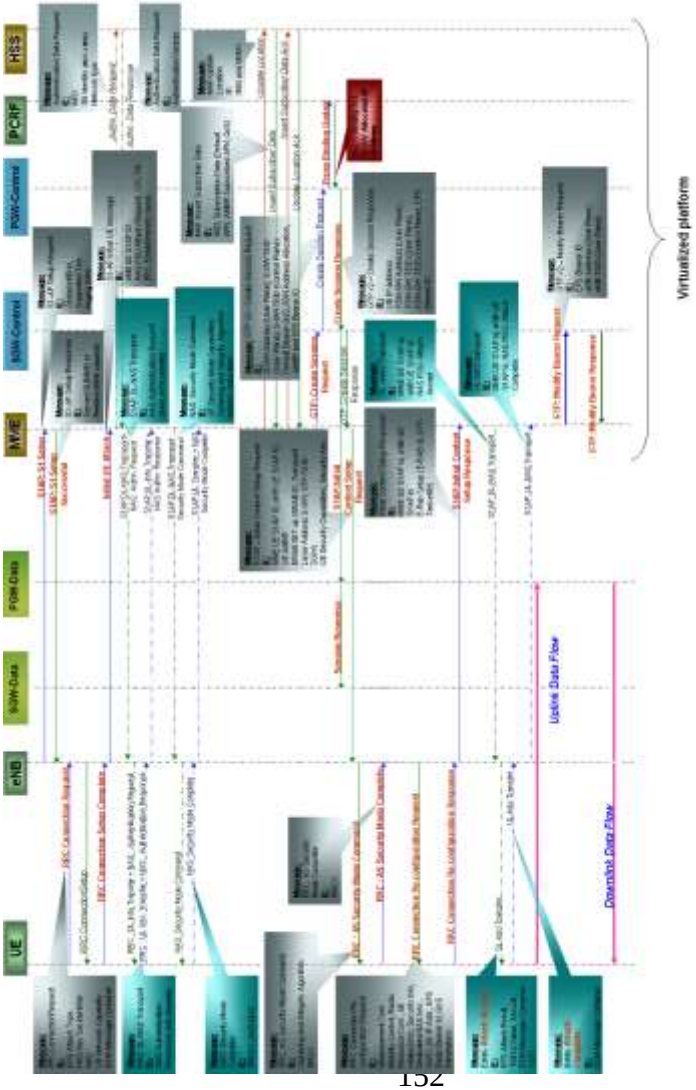


Рис. 3.46. Діаграма повідомлень під час підключення абонента до мережі

В цілому послідовність взаємодії в мобільній мережі передачі даних при запропонованому варіанті віртуалізації можна описати наступним чином:

1.Абонент починає сесію передачі даних (Запит на підключення RRC). Дані від базової станції надходять до віртуалізованого блоку MME (S1 Setup) і базової станція розпочинає приєднання UE (UE Attach).

2.MME надсилає до HSS запит на авторизацію даних (Authorization Data Request).

3.У хмарній віртуалізованій інфраструктурі функціонал BBERF надсилає до PCEF запит на створення сесії для пропуску трафіку (IP-CAN) (блоки розміщені на одній віртуалізованій платформі).

4.PCEF формує запит по інтерфейсу Gx і посилає його на PCRF. Запит полягає у формуванні Diameter CCR (Credit-Control-Request) запиту з інформацією про абонента і запитуваних послугах.

5.PCRF здійснює запит профілю абонента по інтерфейсу Ud.

6.Отримує профіль з параметрами послуг абонента.

7.Здійснює підписку на нотифікацію про зміни профілю.

8.PCRF приймає PCC-рішення про можливість надання послуг абоненту і з якими параметрами якості. Формує PCC-правила, які відправляє на PCEF по інтерфейсу Gx. Це полягає у формуванні Diameter CCA (Credit-Control-Answer) відповіді з включеним набором PCC-правил.

9.При отриманні відповіді PCEF встановлює сесію кредитного контролю з OCS по інтерфейсу Gy за допомогою обміну повідомленнями Diameter CCR/CCA.

10. PCEF дозволяє встановлення IP-CAN сесії і передає відповідні повідомлення на SGW-Control і PGW-Control.

11. SGW-контроль надсилає повідомлення Initial Context Setup на PGW-Data і Session Response (відповідь сесії) на SGW-Data.

12. Потік трафіку (Service Data Flow) починає проходити між пристроєм абонента і зовнішніми мережами зв'язку.

13. Через деякий час абонент завершує сесію передачі даних і BBERF посилає на PCEF запит на розрив IP-CAN сесії.

14. PCEF здійснює завершення Diameter сесій на PCRF по інтерфейсу Gx. Завершення сесій полягає також в обміні повідомленнями Diameter CCR/CCA.

PCEF здійснює завершення Diameter сесій на OCS по інтерфейсу Gy.

Висновки

1. Наведений аналіз послуг в мобільних мережах на українському ринку дозволив визначити, що найбільш перспективними є персоналізовані послуги, які характеризуються вимогами до обсягу смуги пропускання та якості обслуговування. Особливістю процесу надання персоналізованих послуг є складність процесів обробки викликів та тарифікації, а також необхідність адаптації контента до технічних вимог користувача. Визначено, що послуга тарифікації телекомунікаційної мережі є базовою для персоналізованих послуг.

2. Проведений аналіз надання персоналізованих мультимедійних послуг в мережах GSM, UMTS, LTE/IMS вияв, що наявні технічні можливості мережі LTE найповніше відповідають вимогам надання персоналізованих послуг наступного покоління.

3. Система обробки викликів та тарифікації не є досконалою через те, що не враховує технічні параметри функціонування мережі доступу, обсяг частотного ресурсу, необхідний для її надання.

4. Тарифікаційна система стає інструментом впливу при наданні доступу до мультимедійних послуг на базі IP протоколу в залежності від наявності ресурсів і характеристик самої послуги. При наданні послуг на базі IP протоколу тарифікаційна система оператора має враховувати: по-перше, набір параметрів, що описують саму послугу та ресурс мережі для надання послуги (час, пропускна спроможність, кількість переданих пакетів, сумарний трафік). По-друге, має місце тенденція одночасної реалізації вимог 3GPP та EITF у рамках однієї системи.

5. Існуючі недоліки тарифікації пакетних послуг в мережах операторів GPRS/UMTS а саме: немасштабованість, неврахування якості при обслуговуванні та тарифікації, складність надання персоналізованих сервісів і їх тарифікації, – є ключовими факторами, що призвели до розробки нових інтегрованих систем обробки викликів в мобільних мережах зв'язку.

6. При дослідженні основних параметрів, що використовуються при тарифікації (час, пропускна здатність, кількість переданих пакетів, сумарний трафік), було виявлено, що відсутній найважливіший параметр мереж 4-го покоління з ОЧД – ширина смуги частот, кількість піднесущих при наданні послуг.

7. Сформульована задача удосконалення системи обробки викликів за рахунок введення ситуаційних пріоритетів, що базуються на системних параметрах, зокрема ширині смуги частот, та несистемних параметрах (інформаційній важливості контенту) мереж 4-го покоління.

8. Наведений аналіз послуг в мобільних мережах на українському ринку дозволив визначити, що найбільш перспективними є персоналізовані послуги, які характеризуються вимогливістю до обсягу смуги пропускання та якості обслуговування.

9. Виявлено можливість введення пріоритетів при наданні персоналізованих послуг за рахунок удосконалення системи обробки викликів та тарифікації під час перевантаження мережі, що дозволяє зробити систему тарифікації інструментом впливу на обробку викликів при наданні мультимедійних послуг на базі IP протоколу.

10. З метою удосконалення системи обробки викликів та тарифікації і при наданні послуг запропонована нова дисципліна обслуговування заявок – «RF». Дисципліна дозволяє підвищити ефективність системи обробки викликів оператора під час перевантаження мережі за рахунок застосування нового коефіцієнту, який характеризує послугу, і відрізняється від існуючих принципом вибору заявок для подальшого обслуговування.

11. Дисципліна «RF» базується на системі з ситуаційними пріоритетами, та обслуговує в першу чергу заявки з максимальним питомим коефіцієнтом w_k, при умові що $w_k > w_r$, тобто необхідна якість обслуговування для заявок типу k вища ніж заявок для типу r.

12. Запропонований коефіцієнт w_k являє собою відношення тарифу послуги до ширини смуги (частот/пропускання)необхідної для реалізації послуги, для розрахунку її питомої вартості.

13. Проведено математичний розрахунок з використанням теорії масового обслуговування і теорії ймовірності, який підтверджує ефективність дисципліни

обслуговування «RF» порівняно з моделями FIFO, з пріоритетами для одно-канальних і багатоканальних систем. Використання дисципліни «RF» дозволяє підвищити ефективність системи обробки викликів оператора в період перевантаження мережі за умови, що середня кількість відмов не збільшується.

14. Розроблено новий метод тарифікації, що враховує ширину смуги часот при тарифікації мультимедійних широкосмугових послуг та базується на запропонованій дисципліні обслуговування «RF» .

15. Для реалізації запропонованого методу тарифікації модифіковано сигнальний протокол управління радіомережею GTPv2-C для організації послуги з метою отримання параметрів попередньої конфігурації смуги ресурсу в каналі вниз і вверх для передачі послуги. Також обґрунтовано використання модифікації протоколу GTPv2-C, що дозволяє вводити додаткові поля для передачі інформації

16. Адаптовано протокол Diameter для передачі додаткових параметрів в блок прийняття рішень щодо надання послуги і блок тарифікації. Для цього розроблено новий універсальний композиційний блок опису атрибутів сесії – AVP 10415:450, що включає 14 AVP, з яких три нових запропонованих, які описують параметри послуги і параметри користувача: ширина смуги частот, індекс лояльності, тариф послуги. Адаптовано команди RAR і AAR Diameter протоколу для передачі композиційного AVP.

17. Розроблений новий інтерфейс Ty і протокол DMSHORT, що передбачає обмін інформацією на базі нового розробленого AVP 10415:450 між блоком прийняття рішень (PCRF) і он-лайн підсистемою тарифікації OSC.

18. Створений інтерфейс, який описується параметрами AVP 10415:450 і протоколом DMSHORT, та дозволяє розраховувати тариф послуги на етапі прийняття рішення щодо надання послуги даному абоненту.

19. Запропоновані модифікації протоколів GTPv2-C, Diameter, інтерфейс Ty і створений протокол DMSHORT надають змогу в подальшому удосконалювати інтегровані систем управління викликами і тарифікацією, дозволяють узгоджувати тариф послуги з якістю її надання.

20. Удосконалено архітектуру систему обробки викликів та тарифікації (PCC) в системі операторського класу для управління сервісами в мережах радіодоступу LTE/IMS, на прикладі обладнання компанії Huawei Technologies Co.,Ltd..

21. Модифікована програмна архітектура включає новий функціональний блок розрахунку тарифу, безпосередня функція якого – це розрахунок коефіцієнту інформаційної ваги послуги на базі отриманих технічних параметрів функціонування радіомережі. Запропонована архітектура підвищує інтегральну інформаційну вагу оброблених викликів від 10 до 15% відсотків в момент перевантаження мережі без погіршення якості обслуговування, за рахунок реалізації обробки в першу чергу заявок, які мають найбільший коефіцієнт відношення тарифу до ширини смуги частот.

22. Результати роботи впроваджено в компанії МТС на базі обладнання компанії Huawei, які показали ефективність удосконаленої системи обробки викликів та тарифікації в мобільних мережах з ОЧД.

23. Основна перевага віртуалізації з використанням EPC NFV очікується, що деякі віртуальні машини можуть мати різні програмні компоненти, які працюють на віртуальної інфраструктурі. Головною перевагою NFV є здатність складати різні

VNFs та забезпечувати різноманітні та гнучкі послуги кінцевому споживачу за допомогою спеціальних служб.

24. Використання NFV допоможе консолідувати мережеві функції на економічних високопродуктивних серверах, вимикачем і зберігання, скорочення часу виходу на ринок мережевих операторів та витрати з цим.

25. Віртуалізація дасть готову платформу для міграції елементів мережі у хмару. Масштабованість і мультиорендні можливості по віртуалізації платформ дозволить легко розгортати, модернізації та адмініструвати систему.

26. Запропонована послідовність взаємодії елементів в мобільній мережі передачі даних із застосуванням запропонованих розподілених функцій.

27. Запропонований підхід може бути використаний та для віртуалізації різних мережевих елементів, які реалізують протоколи та процедури керування трафіком. Концепція, яка дається в пропонованому підході може бути додатково продовжений на іншу категорію мережевих елементів, що реалізують протокол площині управління і процедури обробки або руху.

4. ОПТИМІЗАЦІЯ РОБОТИ БІЛІНГОВИХ СИСТЕМ

4.1. Метод зменшення навантаження на білінгову систему в режимі критичного навантаження

4.1.1. Тарифікація абонентів з післяплатою в автономному режимі в момент критичного навантаження на білінгову систему

Проаналізуємо ситуацію, коли в момент критичного навантаження на білінгову систему абоненти тарифікуються в автономному режимі за допомогою Fall back CDR. Тобто, транспортна мережа оператора працює в нормальному режимі, а білінгова система знаходиться в непрацездатному або критичному стані. Це означає, що в даний момент абоненти мають можливість скористатися послугами оператора незалежно від показників платоспроможності свого рахунку.

Щоб уникнути ситуації, коли абоненти з нульовим балансом або з невеликим залишком на рахунку використовують надлишкові послуги, через що ефективність функціонування системи знижується, зупинимось більш докладно на абонентах з післяплатою (post-paid) та абонентах з передплаченими послугами (pre-paid).

Зауважимо, що post-paid абоненти згідно з договором про надання телекомунікаційних послуг розраховуються за надані послуги на кінець певного періоду, тобто і рахунок цим абонентам можна виставляти по закінченню деякого часу після завершення надання послуги. Таким чином, оператор має певні гарантії отримати плату за надані послуги, і в жодному разі не несе ніяких збитків. І навпаки, якщо у pre-paid абонентів виникає негативний баланс на рахунку – оператор завжди несе збитки.

Враховуючи все вище згадане, запропонуємо новий метод тарифікації абонентів: при збільшенні навантаження на білінгову систему **всіх post-paid абонентів**, що ініціювали у цей момент запит на надання різних послуг, необхідно тарифікувати в автономному режимі за допомогою створення CDR – тобто фіксувати сам факт надання послуг та їх тривалість. Це знизить навантаження, збільшить ефективність і дозволить білінговій системі обробити більшу кількість вхідних заявок від абонентів передплаченої послуги, і, тим самим, гарантувати тарифікацію в режимі реального часу. В свою чергу тарифікація абонентів передплаченої послуги в режимі реального часу допоможе зменшити кількість оброблених заявок з недостатнім залишком на рахунку і зменшити втрати.

Коли робота білінгової системи налагодиться, тобто вийде з критичного стану – абоненти з післяплатою будуть протарифіковані. Час, який може знадобитися для тарифікації – від декількох мілісекунд до декількох годин. Ще раз підкреслимо, що збитків від post-paid абонентів не буде.

Необхідно визначити середні збитки оператора в моменти, коли білінгова система знаходиться в критичному стані, при використанні звичайної схеми тарифікації всіх абонентів в автономному режимі, і для запропонованої схеми з тарифікацією post-paid абонентів у режимі off-line під час критичного навантаження.

Під час зниження ефективності роботи білінгової системи операторами телекомунікаційного зв'язку зазвичай використовують два варіанти обслуговування абонентів:

– оператори з тарифікацією всіх заявок від абонентів в автономному режимі;

– оператори з пріоритетним обслуговування заявок від абонентів (обслуговуються заявки з вищим пріоритетом, інші – відкидаються).

Тому проаналізуємо окремо операторів з тарифікацією абонентів в автономному режимі та операторів з пріоритетним обслуговування заявок від абонентів під час критичного навантаження на білінгву систему.

4.1.2. Оператори з тарифікацією абонентів в автономному режимі

Припустимо спочатку, що вхідні потоки post-paid і pre-paid абонентів є пуасонівськими з інтенсивностями $\lambda_{\text{post-paid}}$ і $\lambda_{\text{pre-paid}}$, відповідно, та інтенсивність надходження абонентів не залежить від часу. Тоді сумарна інтенсивність надходження пакетів:

$$\lambda = \lambda_{\text{post-paid}} + \lambda_{\text{pre-paid}}. \quad (4.1)$$

Нехай всі абоненти обслуговуються одним пристроєм, обробка абонентів являється марківським процесом з інтенсивністю μ. Припустимо, що ємність буфера дорівнює $(n_0 - 1)$.

Якщо заявка від абонента надходить в момент, коли буфер зайнятий, то вона буде протарифікована в режимі off-line. Позначимо через p ймовірність утворення негативного балансу на рахунку у pre-paid абонента, через m – середні збитки.

Звичайна схема тарифікації

Нехай $\rho_1 = \frac{\lambda}{\mu}$. Стаціонарний розподіл кількості заявок в системі дорівнює:

$$\pi_k = \frac{\lambda^k}{\mu^k}\pi_0;$$

$$\sum_{k=0}^{n_0+1} \pi_k = 1;\ \frac{\lambda^k}{\mu^k} = \rho_1^k;$$

$$\pi_0 \sum_{k=0}^{n_0} \rho_1^k = 1;$$

$$\pi_0 = \frac{1-\rho_1}{1-\rho_1^{n_0+1}};\ \pi_k = \frac{\rho_1^k(1-\rho_1)}{1-\rho_1^{n_0+1}}, k = \overline{0, n_0}. \quad (4.2)$$

де n_0 – максимальна кількість заявок в системі,

π_{n_0} – ймовірність того, що буфер повністю зайнятий.

Нехай пройшов час T. Тоді в середньому $T\pi_{n_0}$ одиниць часу буфер був зайнятий.

За цей час в середньому надійшло $T\pi_{n_0}\lambda_{\text{pre-paid}}$ pre-paid абонентів. Отже, математичне сподівання збитків від абонентів передплаченої послуги, на рахунку яких утворився негативний баланс дорівнює:

$$M(Збитки1) = T\pi_{n_0} \cdot \lambda_{\text{pre-paid}} \cdot p \cdot m = T\frac{\rho_1^{n_0}(1-\rho_1)}{1-\rho_1^{n_0+1}} \cdot \lambda_{\text{pre-paid}} \cdot p \cdot m \quad (4.3)$$

Запропонована схема тарифікації

У даній схемі post-paid абоненти не беруть участь при тарифікації в режимі реального часу.

Нехай $\rho_2 = \frac{\lambda_{\text{pre-paid}}}{\mu}$. Тоді аналогічно звичайній схемі тарифікації середня кількість збитків складає:

$$M(Збитки2) = T\frac{\rho_2^{n_0}(1-\rho_2)}{1-\rho_2^{n_0+1}} \cdot \lambda_{\text{pre-paid}} \cdot p \cdot m \tag{4.4}$$

Порівняємо дві схеми тарифікації абонентів в момент критичного навантаження на білінгову систему та підрахуємо як вплине використання другої системи на економічну ефективність функціонування системи оператора.

Середній виграш за одиницю часу при використанні запропонованої схеми тарифікації в порівнянні зі звичайною схемою складатиме:

$$V(\lambda) = \lambda_{\text{pre-paid}} \cdot p \cdot m\left(\frac{\rho_1^{n_0}(1-\rho_1)}{1-\rho_1^{n_0+1}} - \frac{\rho_2^{n_0}(1-\rho_2)}{1-\rho_2^{n_0+1}}\right) \tag{4.5}$$

$$\frac{Збитки1}{Збитки2} = \frac{\pi_{n_0}^{(1)}}{\pi_{n_0}^{(2)}} \tag{4.6}$$

Зокрема, якщо ρ_1 та $\rho_2 > 1$ і n_0 має достатньо велике значення, то

$$\frac{\rho_1^{n_0}(1-\rho_1)}{1-\rho_1^{n_0+1}} = \frac{1-\rho_1}{\frac{1}{\rho_1^{n_0}}-\rho_1} \approx \frac{1-\rho_1}{-\rho_1} = 1 - \frac{1}{\rho_1};$$

$$\frac{\rho_2^{n_0}(1-\rho_2)}{1-\rho_2^{n_0+1}} = \frac{1-\rho_2}{\frac{1}{\rho_2^{n_0}}-\rho_2} \approx \frac{1-\rho_2}{-\rho_2} = 1 - \frac{1}{\rho_2}. \tag{4.7}$$

$$V(\lambda) = \lambda_{\text{pre-paid}} \cdot p \cdot m\left(1 - \frac{1}{\rho_1} - \left(1 - \frac{1}{\rho_2}\right)\right) = \lambda_{\text{pre-paid}} \cdot p \cdot m\left(\frac{\rho_1-\rho_2}{\rho_1\rho_2}\right) =$$

$$= \lambda_{\text{pre-paid}} \cdot p \cdot m \cdot \frac{\lambda_{\text{post-paid}} \cdot \mu}{(\lambda_{\text{pre-paid}}+\lambda_{\text{post-paid}})\lambda_{\text{pre-paid}}} = \frac{p \cdot m \cdot \lambda_{\text{post-paid}} \cdot \mu}{\lambda} \tag{4.8}$$

Припустимо, що доля абонентів з післяплатою складає всього 5 відсотків від усіх абонентів оператора N:

$$\lambda_{\text{pre-paid}} = 0{,}95\lambda, \quad \lambda_{\text{post-paid}} = 0{,}05\lambda.$$

Тоді:

$$V(\lambda) \approx \frac{p \cdot m \cdot 0{,}05\lambda \cdot \mu}{\lambda} = 0{,}05 \cdot p \cdot m \cdot \mu. \tag{4.9}$$

Отже, на стільки відсотків збільшиться кількість протарифікованих абонентів передплаченої послуги в режимі реального часу, і на стільки ж відсотків збільшиться технічна і економічна ефективність білінгової системи.

Необхідно також відмітити, що зазвичай інтенсивність надходження заявок від абонентів (пакетів λ) не є постійною і змінюється протягом дня: $\lambda = \lambda(t)$, де $t \in [0; 24]$. Тому позначимо щільність розподілу інтенсивності надходження пакетів через $p(\lambda)$, тобто для будь-яких a, b частка часу за добу, коли інтенсивність надходження належить відрізку $[a, b]$, дорівнює $\int_a^b p(\lambda)d\lambda$.

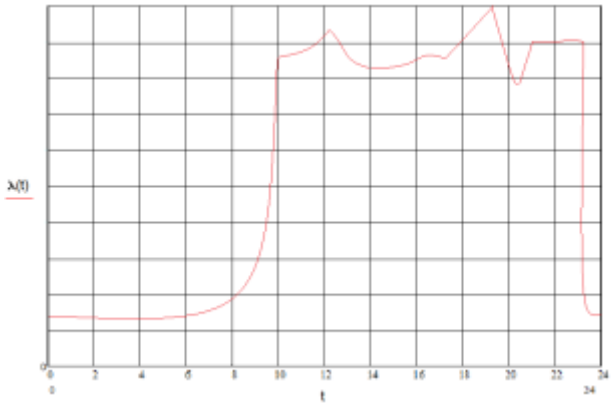


Рис. 4.1 Інтенсивності надходження пакетів впродовж доби $\boldsymbol{\lambda(t)}$

Припустимо, що на невеликих проміжках часу λ змінюється «плавно». Тоді через відсутність довгих переходів стаціонарний режим досягається досить швидко.

Зазначимо, що відношення інтенсивностей надходження post-paid і pre-paid абонентів $\lambda_{\text{post-paid}}/\lambda_{\text{pre-paid}}$ не повинно залежати від часу, так як відношення кількості абонентів з передплаченою послугою і абонентів з післяплатою, які подзвонили в будь-який момент часу, – однакове, і не залежить від типу абонента і часу. Величина ж даного відношення, у свою чергу, коливається залежно від часу доби (починаючи з 18:00 до 21:00 години – воно досягає свого максимуму).

Тому середній виграш за час T приблизно дорівнює:

$$V_T = \int_0^\infty V(\lambda) \cdot p(\lambda) d\lambda \cdot T, \qquad (4.10)$$

де T= 365 днів,

$V(\lambda)$ – обчислюється за формулою (4.5).

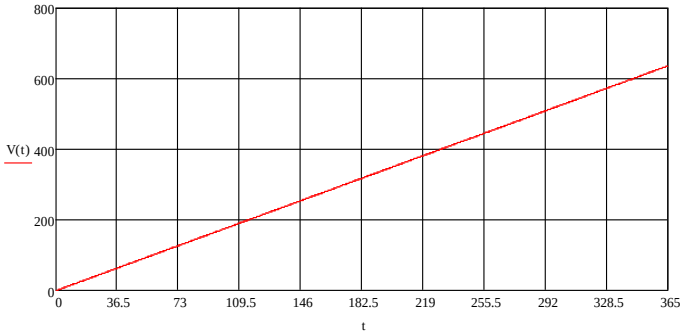


Рис. 4.2 Середній виграш від впровадження запропонованої схеми тарифікації за час T (365 днів)

В деяких випадках $p(\lambda)$ можна вирахувати наступним чином. Нехай $\lambda(t)$ в момент t являється гаусівською випадковою величиною з математичним сподіванням $a(t)$ і дисперсією $\sigma^2(t)$. Тоді:

$$\mathrm{P}(\lambda(t) > \mu) = \mathrm{P}\big(N\big(a(t), \sigma^2(t)\big) > \mu\big) = \mathrm{P}(a(t) + \sigma(t)N(0,1) > \mu)$$
$$= \mathrm{P}\left(N(0,1) > \frac{\mu - a(t)}{\sigma(t)}\right) = \bar{\Phi}\left(\frac{\mu - \bar{a}(t)}{\bar{\sigma}(t)}\right);$$
$$\bar{\Phi}(x) = \mathrm{P}(N(0,1) > x).$$

Отже, середня кількість моментів, коли інтенсивність надходження заявок перевищує інтенсивність обробки $\lambda > \mu$ дорівнює $\int_0^{24} \bar{\Phi}\left(\frac{\mu - \bar{a}(t)}{\bar{\sigma}(t)}\right) dt$.

$$\mathrm{P}(\lambda(t) \in [x, x + \Delta]) = \cdots = \mathrm{P}\left(N(0,1) \in \left[\frac{x - a(t)}{\sigma(t)}, \frac{x + \Delta - a(t)}{\sigma(t)}\right]\right) \approx$$
$$\approx \frac{1}{\sigma(t)} p_{N(0,1)}\left(\frac{x - a(t)}{\sigma(t)}\right) \cdot \Delta$$

Тому

$$\boldsymbol{p(\lambda) = \frac{1}{24}\int_0^{24} \frac{1}{\sigma(t)} p_{N(0,1)}\left(\frac{\lambda - a(t)}{\sigma(t)}\right) dt}. \qquad \textbf{(4.11)}$$

Зауваження. Якщо $\lambda > \mu$, то середній виграш запропонованої схеми тарифікації у порівнянні зі звичайною схемою приблизно дорівнює $C \approx const$ (див. формулу (19.9)) і не залежить від λ.

Якщо ж $\lambda < \mu$, то можна помітити, що середній виграш досить малий. Тому середній виграш від впровадження нової системи приблизно дорівнює:

$$V \approx C \cdot \int_\mu^\infty p(\lambda) d\lambda \cdot 24 = C \cdot \mathrm{P}(\mu) \cdot 24 = C \cdot \tau, \qquad (4.12)$$

де $C \approx const$ – див. формулу (4.9);

$p(\lambda)$ – щільність розподілу інтенсивності надходження пакетів,

$\mathrm{P}(\mu)$ – імовірність того, що інтенсивність надходження пакетів більше інтенсивності обробки,

τ – середній час за день, коли інтенсивність надходження пакетів більше інтенсивності обробки.

Розглянемо більш детально затрати на впровадження запропонованої схеми тарифікації, щоб зрозуміти що є більш доцільним для оператора: модернізувати білінгову систему і оптимізувати (зменшити) кількість відкинутих абонентів у момент критичного навантаження та збільшити прибуток, або заощадити кошти на модернізації білінгової системи та терпіти деякі збитки від безоплатного користування послугами в момент критичного навантаження. Вигода від впровадження запропонованої схеми тарифікації буде завжди, проте можливо незначна. З часом сумарна вигода перевищить затрати на впровадження, але цей проміжок часу може бути занадто великим, що робить модернізацію нерентабельною. Порівняти невеликі, проте регулярні збитки з великою разовою затратою на модернізацію дозволяє підхід, який оснований на «дисконтуванні», або часовій цінності грошей.

Часова цінність грошей – це концепція, на якій основане припущення про те, що гроші повинні приносити відсотки – цінність сьогоднішніх грошей вище, ніж цінність тієї ж суми, яка буде одержана в майбутньому. Кожен віддасть перевагу отримати певну суму грошей сьогодні, ніж ту ж саму кількість у майбутньому. Це пояснюється тим, що якщо покласти якусь суму грошей у банк, то в наступному

році можна отримати не тільки цю суму грошей, але й відсотки від неї. Іншими словами, можна вважати, що сума x у.о. в момент часу t рівносильна $x\alpha^t$у.о. в момент 0, де константа $\alpha \in (0; 1)$ називається коефіцієнтом дисконтування [48].

Це означає, що для перерахунку майбутніх потоків доходів в єдину величину поточної вартості необхідно використовувати інтегральний ефект, дисконтований до теперішнього моменту часу, або ж інтегральний дисконтований ефект [48].

Нехай A – вартість впровадження нової системи в даний момент часу, n – кількість років, V_T – середній виграш за час T (див. формулу 19.10). Тоді загальний дохід, приведений до даного моменту часу з урахуванням коефіцієнта дисконтування, дорівнює:

$$W = -A + \sum_{n=0}^{\infty} V_T \cdot \alpha^n = -A + \frac{V}{1-\alpha}. \quad (4.13)$$

Таким чином, якщо загальний дохід W позитивний, тобто $\frac{V}{1-\alpha} > A$, тоді модернізація системи в перспективі принесе прибуток, якщо ж дохід негативний, то модернізація для системи не рентабельна.

4.1.3. Оператори з пріоритетним обслуговуванням заявок

При використанні дисципліни обслуговування абонентів з пріоритетом під час перевантаження білінгової системи, оператор не несе прямих збитків. Зупинимось на цьому більш детально: якщо навантаження на систему значне, проте не критичне, то всі нові заявки, які поступають і мають менший пріоритет – будуть відкидатися. Якщо ж навантаження на білінгову систему досягає критичного рівня, то всі нові заявки відкидаються. Таким чином, оператор не несе ніяких збитків, оскільки повністю унеможливлюється доступ до послуг незалежно від стану рахунку і пріоритету заявки (виключення складають тільки дзвінки на екстрені номери).

Проте існують так звані непрямі збитки, які проявляються в таких важливих показниках, як якість обслуговування, доступність послуги, зниження яких веде до зростання незадоволеності абонентів і можливої їх втрати у майбутньому.

Пропозиція застосувати метод тарифікації **всіх абонентів з <u>післяплатою</u> в автономному режимі** підвищить технічну і економічну ефективність функціонування білінгової системи і дозволить зменшити кількість втрачених заявок на обслуговування серед абонентів з передплаченими послугами, а абоненти з післяплатою зможуть скористатися послугою не залежно від технічного стану системи тарифікації в режимі реального часу. Це зменшить непрямі збитки оператора і покращить показники якості надання послуг.

Аналогічно до попереднього пункту, якщо доля абонентів з післяплатою складає 5% від усіх абонентів, то кількість протарифікованих абонентів передплаченої послуги в режимі реального часу збільшиться в середньому за час T на $V = \frac{T \cdot p \cdot m \cdot 0{,}05\lambda \cdot \mu}{\lambda} = 0{,}05 \cdot T \cdot p \cdot m \cdot \mu$ (%) (де p – ймовірність утворення негативного балансу на рахунку у pre-paid абонента, m – середні збитки pre-paid абонента, μ – інтенсивність обслуговування), і на стільки ж відсотків збільшиться технічна і економічна ефективність білінгової системи.

4.1.4. Класифікації абонентів з передплаченого послугою за рівнем ризику

Запропонуємо ще один можливий підхід для оптимізації кількості некоректно тарифікованих заявок і підвищення технічної та економічної ефективності функціонування білінгової системи при навантаженій роботі, якщо використання методу, описаного в попередньому параграфі, не дає бажаного результату.

Нагадаємо, що при збоях у роботі білінгової системи тарифікація абонентів відбувається в автономному режимі і грошовий залишок на їх рахунку не контролюється, тому виникає вірогідність надання послуг у кредит (поява негативного балансу на рахунку у абонента). В залежності від ряду умов (наприклад, сума боргу, тип підключення і т.д.) даний кредит може бути погашений користувачем, а може бути проігнорований, оскільки оператор не має додаткових важелів впливу на абонентів з передплаченою послугою, окрім часткового або повного призупинення надання послуг (згідно попереднього параграфу, всі абоненти з післяплатою вже обслуговуються в автономному режимі).

Звернемо увагу на те, що ймовірність того, що у абонентів з незначним залишком коштів на рахунку утвориться негативний баланс вище, ніж таж сама ймовірність у абонентів з грошима на рахунку. Тому розділимо абонентів з передплаченою послугою на категорії:

– абоненти, у яких на поточному рахунку «мало» грошей – «ризиковані» абоненти;

– абоненти, у яких на рахунку «багато» грошей – «не ризиковані», або «надійні» абоненти.

Будемо тарифікувати «надійних» абонентів в автономному режимі, а «ризикованих» – в режимі реального часу. Це зменшить можливі втрати від абонентів, які з нульовим або незначним балансом намагаються отримати послуги в моменти, коли система знаходиться в непрацездатному або критичному стані. Можливий ризик, який виникає при даному рішенні, полягає в тому, що «надійні» абоненти вичерпають усі засоби на рахунку, і утвориться негативний баланс в той момент, коли робота білінгової системи вже буде налагоджена і протікати в нормальному режимі.

Для поділу абонентів на «надійних» і «ризикованих», необхідно визначити деякий рівень залишку коштів на рахунку абонентів з передплаченою послугою, так званий «поріг ризику». У моменти критичного навантаження на білінгову систему спочатку обробляти абонентів з високим ступенем ризику, а абонентів, у яких грошей більше зазначеного порогу – тарифікувати в режимі off-line. Тоді ймовірність втрати грошей у передплачених абонентів знижується. Зі зменшенням навантаження на білінгову систему будемо тарифікувати «не ризикованих» абонентів в режимі реального часу, а якщо потужності дозволяють, то і абонентів з післяплатою.

Для спрощення розрахунків припустимо, що телекомунікаційний оператор надає один вид послуг, наприклад, телефонні дзвінки.

Припустимо, що в даний момент білінгова система знаходиться в критичному стані, і абоненти тарифікуються в автономному режимі. Обчислимо середні збитки, які понесе оператор від обслуговування одного абонента.

Нехай Y – кількість грошей на рахунку у абонента, C – тариф, або вартість одиниці наданої послуги (грн./хв.), час розмови абонента – випадкова величина ξ з щільністю розподілу $p(x)$.

Таким чином, функція розподілу телефонних дзвінків дорівнює:

$$F(x) = \mathrm{P}(\xi \le x),\ \bar{F}(x) = \mathrm{P}(\xi > x). \tag{4.14}$$

Система неефективна і оператор несе збитки, якщо $\xi \cdot C > Y$. Тобто ймовірність збитків дорівнює:

$$\mathrm{P}(C\xi > Y) = \mathrm{P}\left(\xi > \frac{Y}{C}\right) = \bar{F}\left(\frac{Y}{C}\right) \tag{4.15}$$

Обчислимо математичне сподівання збитків:

$$\text{Збитки } = \begin{cases} 0, \text{якщо } \xi \le \frac{Y}{C}, \\ \left(\xi - \frac{Y}{C}\right) \cdot C, \text{якщо } \xi > \frac{Y}{C}; \end{cases} = \begin{cases} 0, \text{якщо } \xi \le \frac{Y}{C}, \\ C\xi - Y, \text{якщо } \xi > \frac{Y}{C}. \end{cases} \tag{4.16}$$

$$\text{Позначим, } \mathrm{g(x)} = \begin{cases} 0, \text{якщо x} \le \frac{\mathrm{Y}}{\mathrm{C}}, \\ \mathrm{Cx - Y}, \text{якщо x} > \frac{\mathrm{Y}}{\mathrm{C}}. \end{cases} \tag{4.17}$$

Тоді:

$$M[\textit{Збитки}] = \int_0^\infty g(x)p(x)dx = \int_0^{\frac{Y}{C}} 0 \cdot p(x)dx + \int_{\frac{Y}{C}}^\infty (Cx - Y)p(x)dx =$$

$$= \int_{\frac{Y}{C}}^\infty (Cx - Y)p(x)dx \tag{4.18}$$

Знайдемо рішення даного інтеграла методом інтегрування по частинах. Нехай $u = Cx - Y,\ dv = p(x)dx$, тоді $du = Cdx$, $v = -\int_x^\infty p(y)dy = -\bar{F}(x)$.

Отже,

$$M[\textit{Збитки}] = \int_{\frac{Y}{C}}^\infty (Cx - Y)p(x)dx = (Cy - Y)\left(-\bar{F}(y)\right)\Big|_{\frac{Y}{C}}^\infty - \int_{\frac{Y}{C}}^\infty -\bar{F}(x)Cdx$$

$$= 0 - 0 + \int_{\frac{Y}{C}}^\infty \bar{F}(y)dy = \int_{\frac{Y}{C}}^\infty \bar{F}(y)dy.$$

Позначимо даний вираз через $f_{\text{аб}} = \int_{\frac{Y}{C}}^\infty \bar{F}(y)dy$.

Зауважимо, що для кожного абонента $\bar{F}(x)$ і Y мають свої власні значення, які необхідно знайти.

Для функції розподілу телефонних дзвінків потрібно скласти статистичну таблицю активності абонента, наприклад, за останні три місяці (якщо абонент новий, тобто обслуговується в даній мережі менше трьох місяців, то його завжди тарифікуємо в режимі реального часу). Для складання такої таблиці активності

абонента необхідно встановити графік періодичного перерахунку, наприклад, раз на місяць для кожного абонента, під час найменшого навантаження на білінгову систему (нічний час з 2:00 до 6:00).

Для визначення залишку коштів на рахунку абонентів необхідно за деякий час до початку ймовірного критичного навантаження створити таблицю з інформацією про кількість грошей на рахунку кожного абонента. Це робиться заздалегідь (наприклад, за 20 хвилин) для того, щоб не створювати додаткове навантаження на білінгову систему в критичні моменти, але і не надто завчасно, щоб цей залишок не сильно змінився за відповідний час.

Поділ абонентів на «ризикованих» і «не ризикованих» будемо робити з деякою періодичністю, наприклад, раз або два рази на добу (в надії, що «не ризикований» абонент не встигне витратити всі кошти на рахунку за час, поки білінгова система буде перебувати в стані критичного навантаження). У деяких випадках такий поділ доцільно робити перед настанням критичного моменту, коли навантаження незначне, щоб запобігти утворення додаткового навантаження на білінгову систему.

Припустимо, що вхідний потік заявок – марківський з інтенсивністю надходження λ та інтенсивністю обробки μ. Припустимо, що в даний момент інтенсивність надходження є критичною, тобто $\lambda > \mu$. Необхідно знизити кількість заявок в системі, які тарифікуються в on-line режимі таким чином, щоб:

$$\lambda_0 < \mu. \tag{4.19}$$

Введемо деякий «поріг ризику» абонента L, і залежно від того $f_{\text{аб}} > L$, чи $f_{\text{аб}} < L$ приймемо рішення про тарифікацію користувача в on-line або off-line режимі.

Константу L виберемо так, щоб інтенсивність надходження абонентів, для яких $f_{\text{аб}} > L$, дорівнювала λ_0. Тоді «ризиковані» абоненти будуть тарифікуватися в режимі реального часу, оскільки ймовірність появи заборгованості досить велика у порівнянні з ймовірність погашення цієї заборгованості.

Також це дозволить тарифікувати абонентів, у яких $f_{\text{аб}} < L$ – в автономному режимі, і за рахунок цього зменшити інтенсивність надходження абонентів від критичної до нормальної.

При цьому для того щоб не тарифікувати всіх абонентів, у яких виконується вище наведена нерівність, в автономному режимі і обробляти CDR, коли робота білінгової системи налагодиться – поріг ризику абонента L повинен бути таким, щоб навантаження було трохи менше одиниці.

4.1.5. Метод оптимізації ємності буфера очікування білінгової системи

Як вже відмічалося вище, білінгові системи є вендорним продуктом і часто, якщо технічні і економічні показники функціонування системи не задовольняють всіх вимог оператора, – повна заміна системи і придбання нової тягне за собою значні фінансові вливання та різноманітні технічні ускладнення при впровадженні системи в мережу оператора. Тому, якщо система лиш частково не задовольняє всіх вимог оператора – цілковита її заміна просто нерентабельна. В цьому випадку необхідно знайти методи, що дозволять підвищити техніко-економічну ефективність

функціонування системи та мінімізувати втрати при її модернізації, виключаючи можливість заміни системи.

Одним із методів вирішення даної проблеми може бути запропонований метод нарощення буфера очікування білінгової системи. Необхідно розрахувати оптимальне значення приросщення і визначити, коли витрачені кошти почнуть приносити прибуток, в порівнянні з ціною витрат на модернізацію.

Щоб розрахувати майбутню економічну ефективність, потрібно враховувати коефіцієнт дисконтування – це процентна ставка, яка використовується для перерахунку майбутніх потоків доходів в єдину величину поточної вартості.

Оскільки потік службової інформації від і до білінгової системи є досить великим – це і різноманітні тарифні плани, і послуги, і платіжний баланс абонентів, і т.д., тому при збої в роботі білінгової системи абоненти з передплаченою послугою, тарифікація яких здійснюється в режимі реального часу, можуть отримувати послугу безкоштовно. Тобто, коли буфер очікування системи переповнений, і білінгова система не встигає обробляти нові поступаючі пакети, тоді ці пакети отримують запитувані сервіси незалежно від стану платоспроможності їх рахунку і тарифікуються в автономному режимі.

Припустимо, що робота білінгової системи описується марковским процесом. Нехай у систему випадковим чином незалежно надходять потоки пакетів типу $1, \dots, m$ з інтенсивностями, $\lambda_1, \lambda_2 \dots \lambda_m$, відповідно. Позначимо через C_k – ціну пакету з k-го потоку. Тоді сумарна інтенсивність надходження пакетів:

$$\lambda = \lambda_1 + \cdots + \lambda_m \tag{4.20}$$

Ймовірність того, що надійшов пакет типу k, дорівнює:

$$p_k = \frac{\lambda_k}{\lambda} \tag{4.21}$$

Зауважимо, що середня ціна пакету, що надійшов, дорівнює:

$$C = \sum_k p_k C_k \tag{4.22}$$

Оскільки, обробка будь-якого пакета, що надходить до системи, здійснюється одним пристроєм з інтенсивністю μ і $x(t)$ – кількість пакетів в системі у момент t. Тоді $x(t)$ описується системою M/M/1/n_0, де n_0 – початкова ємність буфера.

Добре відомо, що при тривалій роботі системи, її розподіл сходиться до стаціонарного. Тому далі припустимо, що $x(t)$ має стаціонарний розподіл.

Нехай $\rho = \frac{\lambda}{\mu}$. Тоді стаціонарний розподіл дорівнює [47]:

$$\begin{gathered} \pi_k = \frac{\lambda^k}{\mu^k}\pi_0; \\ \sum_{k=0}^{n_0+1} \pi_k = 1;\ \frac{\lambda^k}{\mu^k} = \rho^k; \\ \pi_0 \sum_{k=0}^{n_0+1} \rho^k = 1; \\ \pi_0 = \frac{1-\rho}{1-\rho^{n_0+2}};\ \pi_k = \frac{\rho^k(1-\rho)}{1-\rho^{n_0+2}}, k = \overline{0, n_0+1}. \end{gathered} \tag{4.23}$$

Зауваження. Якщо $\rho = 1$, то стаціонарний розподіл $\pi_k = \frac{1}{n_0+2}$. Цей випадок розглядати не будемо.

Пакет буде відкинутий, якщо застане систему в стані $n_0 + 1$, тобто ймовірність відмови пакету, який надійшов, дорівнює:

$$P_{\text{відмови}} = \pi_{n_0+1} = \frac{\rho^{n_0+1}(1-\rho)}{1-\rho^{n_0+2}} \tag{4.24}$$

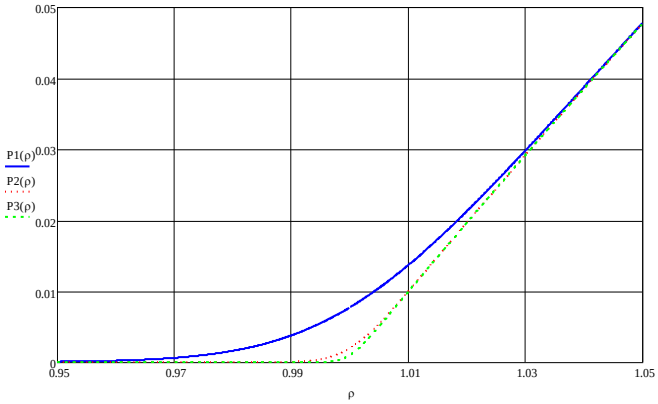


Рис. 4.3 Залежність ймовірності відмови пакету від ρ при початковій ємності буфера 128, 512, 1024 відповідно

Згідно графіку, зображеному на Рис. 4.3 нарощення буфера очікування білінгової системи доцільне при $\rho \in (0;1)$. Якщо $\rho > 1$, то збільшення буферу марне.

Позначимо середню кількість втрачених пакетів за час T через l, а середні збитки за час T через W. Тоді

$$l = \pi_{n_0+1} \cdot \lambda T,$$
$$W = l \cdot \sum_k p_k C_k = \pi_{n_0+1} \lambda T \cdot C. \tag{4.25}$$

Розглянемо нескінченно малий відрізок часу від t до $t + \Delta t$. Тоді:

Середні збитки оператора від втрачених пакетів за інтервал часу $[t, t + \Delta t]$ дорівнює $\pi_{n_0+1} \cdot \lambda \Delta t$.

Таким чином, $F_t = \pi_{n_0+1} \cdot \lambda C$ можна інтерпретувати, як швидкість збільшення збитків від втрати пакетів.

Щоб зрозуміти, що більш доцільно для оператора: наростити ємність буфера білінгової системи, і збільшити тим самим кількість обслуговуваних абонентів та техніко-економічну ефективність системи, або заощадити кошти на модернізації білінгової системи і терпіти збитки від втрати пакетів – розглянемо тимчасову цінність грошей.

За нескінченний інтервал часу середня вартість втрачених пакетів, з урахуванням коефіцієнта дисконтування дорівнюватиме:

$$\int_0^\infty F_t\, \alpha^t dt = \int_0^\infty \pi_{n_0+1} \lambda C \alpha^t\, dt = \frac{\pi_{n_0+1} \lambda C}{\ln \frac{1}{\alpha}},$$

де $\alpha \in (0;1)$ – коефіцієнт дисконтування.

Нехай A – це вартість одиниці буфера в поточний момент часу,

$C = \sum_k p_k C_k$ – середня ціна пакета, n_0 – початкова ємність буфера, n – ємність буфера після нарощування.

Тоді збільшення загального доходу від збільшення буфера, приведене до даного моменту часу, дорівнює:

$$f(n) = -\frac{\pi_{n+1}\lambda C}{\ln\frac{1}{\alpha}} + \frac{\pi_{n_0+1}\lambda C}{\ln\frac{1}{\alpha}} - A(n - n_0).$$

Щоб знайти найбільший прибуток від збільшення буфера необхідно дослідити дану функцію на максимум.

Для спрощення розрахунків введемо позначення $B = \frac{\rho(1-\rho)\lambda C}{\ln\frac{1}{\alpha}}$, тоді:

$$f(n) = -\frac{\rho^n}{1-\rho^{n+2}} \cdot B + \frac{\rho^{n_0}}{1-\rho^{n_0+2}} \cdot B - A(n - n_0) \to \max.$$

Трансформуємо дану функцію до функції, що залежить від змінної x:

$$f(x) = -\frac{\rho^x}{1-\rho^{x+2}} \cdot B + \frac{\rho^{n_0}}{1-\rho^{n_0+2}} \cdot B - A(x - n_0).$$

Для того, щоб знайти максимум функції $f(x)$ дослідимо її похідну:

$$f'(x) = \left(-\frac{\rho^x}{1-\rho^{x+2}} \cdot B + \frac{\rho^{n_0}}{1-\rho^{n_0+2}} \cdot B - A(x - n_0)\right)';$$

$$f'(x) = -\frac{B[\rho^x \cdot \ln\rho \cdot (1-\rho^{x+2}) + \rho^{x+2} \cdot \ln\rho \cdot \rho^x]}{(1-\rho^{x+2})^2} - A = -\frac{\rho^x \cdot \ln\rho}{(1-\rho^{x+2})^2} \cdot B - A =$$

$$= -\frac{\rho^x \cdot \ln\rho \cdot B + A(1-\rho^{x+2})^2}{(1-\rho^{x+2})^2}.$$

Обчислюючи другу похідну і враховуючи, що $B > 0$ при $\rho \in (0;1)$, $B < 0$ при $\rho > 1$, нескладно перевірити, що $f''(x) < 0$ при $x > -2$. Тому функція f строго опукла вгору.

Оскільки $\lim\limits_{x\to -2+0} f(x) = \lim\limits_{x\to+\infty} f(x) = -\infty$, то f має єдиний максимум на інтервалі $(-2; +\infty)$. Для його пошуку вирішимо рівняння $f'(x) = 0$. Таким чином, похідна дорівнює 0, якщо:

$$-\rho^x \cdot \ln\rho \cdot B - A(1-\rho^{x+2})^2 = 0.$$

Зробимо заміну $z = \rho^x$, тоді

$$-z \cdot \ln\rho \cdot B - A(1 - z\rho^2)^2 = 0,$$

$$A\rho^4 z^2 + z(\ln\rho \cdot B - 2A\rho^2) + A = 0.$$

Дане рівняння є квадратним відносно z. Дискримінант цього рівняння дорівнює:

$$D = \ln\rho \cdot B(\ln\rho \cdot B - 4A\rho^2) > 0,$$

$$\text{так як } \ln\rho \cdot B = \ln\rho \cdot \frac{\rho(1-\rho)\lambda C}{\ln\frac{1}{\alpha}} < 0.$$

Корені квадратного рівняння:

$$z_{\pm} = \begin{bmatrix} \dfrac{-(\ln\rho \cdot B - 2A\rho^2) + \sqrt{\ln\rho \cdot B(\ln \mathrm{c} \cdot B - 4A\rho^2)}}{2A\rho^4}, \\ \dfrac{-(\ln\rho \cdot B - 2A\rho^2) - \sqrt{\ln\rho \cdot B(\ln\rho \cdot B - 4A\rho^2)}}{2A\rho^4}. \end{bmatrix}$$

Тому оптимальне значення x дорівнює:

$$x_+^* = \log_\rho z_+ = \frac{\ln z_+}{\ln \rho} \text{ или } x_-^* = \log_\rho z_- = \frac{\ln z_-}{\ln \rho}.$$

Як згадувалося вище, f досягає єдиного максимуму на інтервалі $x \in (-2; +\infty)$. Тому рівняння $f'(x) = 0, x \in (-2; +\infty)$ має рівно один корінь, а значить

$$x^* = \max\{x_+^*, x_-^*\} = \begin{cases} \dfrac{\ln z_+}{\ln \rho}, \text{якщо } \rho \in (1; +\infty), \\ \dfrac{\ln z_-}{\ln \rho}, \text{ якщо } \rho \in (0; 1). \end{cases}$$

Таким чином, якщо $x^* \le n_0$, то нарощування буфера очікування нерентабельне. Якщо ж $x^* > n_0$ – збільшення буфера мінімізує втрати оператора телекомунікаційного зв'язку та дозволить йому отримати найбільший прибуток.

Зауваження. Якщо витрати на роботу з заміни обладнання рівні $C > 0$, то функція прибутку буде мати вигляд:

$$g(x) = -\frac{\rho^x}{1-\rho^{x+2}} \cdot B + \frac{\rho^{n_0}}{1-\rho^{n_0+2}} \cdot B - A(x - n_0) - C.$$

Відмітимо, що $g'(x) = f'(x)$. Тому дані функції мають один і той же максимум в точці x^*.

Тоді, якщо $g(x^*) > 0$, то буфер очікування білінгової системи необхідно збільшити для отримання більшого прибутку оператором зв'язку. В іншому випадку, коли $g(x^*) < 0$ – нарощування буфера недоцільно.

4.2. Розподіл ресурсів в системі online тарифікації сервісів

4.2.1 Обслуговування заявок на тарифікацію

Розглянемо більш детально роботу системи онлайн тарифікації (OCS – Online Charging System). Обслуговування абонентів на сервері тарифікації пов'язано з виконанням ряду стандартних операцій, які потребують використання фізичних ресурсів системи, таких як оперативна пам'ять, постійна пам'ять, мережевий ресурс та процесорний час системи, саме тому коли йдеться про перевантаження системи тарифікації, спостерігається нестача одного або декількох типів фізичних ресурсів, що тягне за собою відмову у обслуговування заявок на тарифікацію. 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.

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 підібрати ресурси серверу таким чином, щоб зменшити ймовірність перевищення часу обслуговування абонентських заявок.

В даному розділі досліджується спосіб розподілу наявних ресурсів між різними типами заявок на тарифікацію та різними функціональними блоками системи тарифікації, що забезпечить ефективну роботу системи тарифікації.

Складність процесу розрахунку кількості ресурсів необхідної для тарифікації різних типів послуг пояснюється чотирма основними передумовами:

1. Процес тарифікації передбачає послідовне виконання операцій, які потребують різної кількості ресурсів.

2. Кожний тип послуг, незважаючи на стандартність операцій, які виконуються в процесі тарифікації, потребує для здійснення розрахунків різної кількості ресурсів.

3. Одночасно надходить велика кількість заявок на тарифікацію різних типів ресурсів.

4. Інтенсивність надходження заявок залежить від дня тижня та часу доби.

Зробимо дослідження кожної передумови.

Схему обслуговування заявок на сервері тарифікації показано на рис. 4.4.

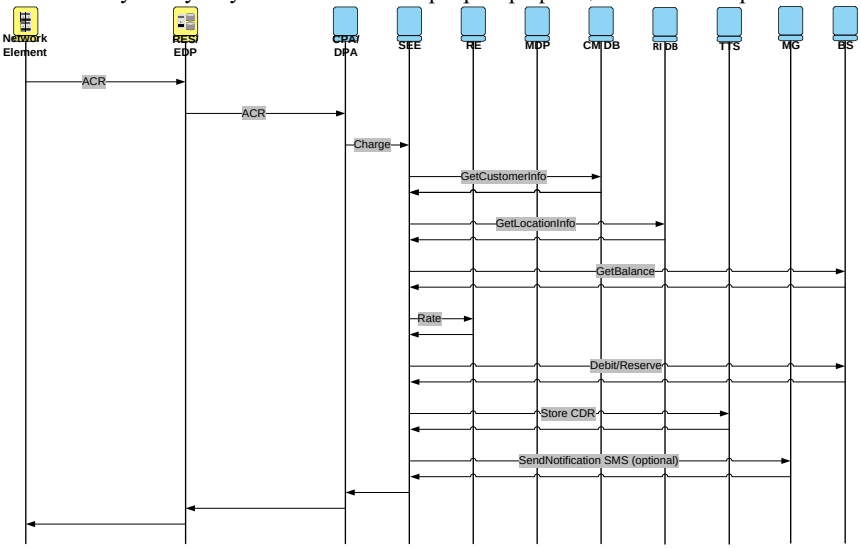


Рис. 4.4 Загальна схема online тарифікації послуг на сервері тарифікації

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).

8.Процес 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 вибору оптимальної кількості ресурсів необхідно звернути увагу на спосіб 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.

Друга передумова 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. 4.5-4.7 показана послідовність запитів, які надходять до системи тарифікації (OCS – Online Charging System) при обслуговування групи сервісів SCUR, ECUR, IEC відповідно.

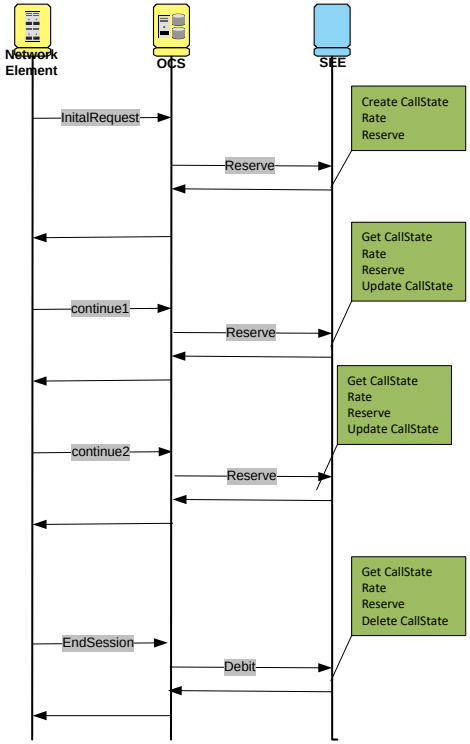


Рис. 4.5 Процес виконання заявки для сервісу SCUR

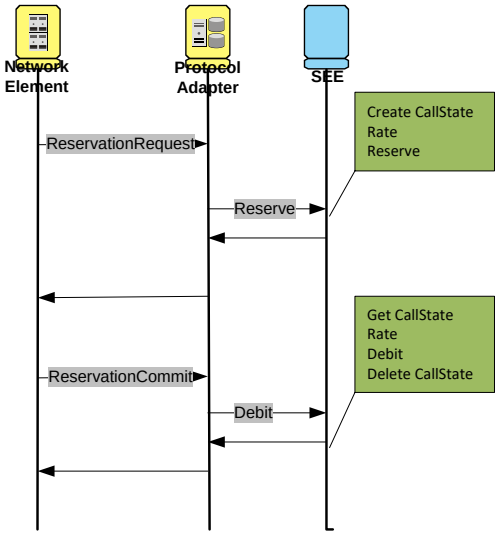


Рис. 4.6 Процес виконання заявки для сервісу ECUR

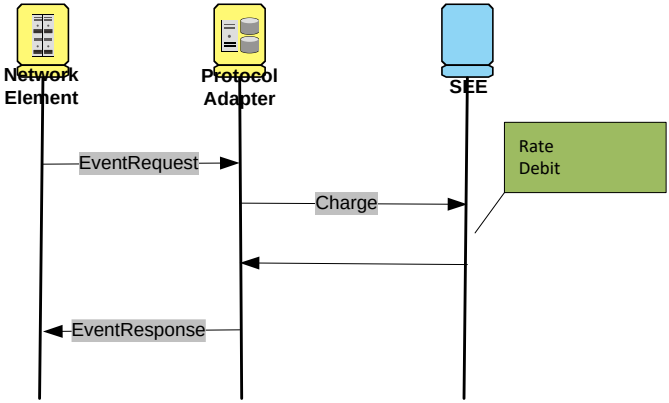


Рис. 4.7 Процес виконання заявки для сервісу IEC

На рис. 4.4 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.

Третя передумова 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 систем тарифікації 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 пуасонівським 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.

4.2.2. Модель контролю за навантаженням на підсистеми в процесі обслуговування заявок на тарифікацію

Враховуючи складність процесу обслуговування заявок на тарифікацію в системах онлайн тарифікації, необхідно розробити модель контролю за навантаженням, що створює той чи інший тип сервісу, з метою забезпечення ефективного керування системою.

Сервер мобільного зв'язку обслуговує m типів сервісів. Заявкою будемо називати запит, який надсилає клієнт до серверу на обслуговування одного з m типів сервісів. Заявки одного типу сервісу обслуговується за однаковою схемою. Для кожного типу сервісу розроблено схему обслуговування заявок на сервері. Узагальнено схему обслуговування заявки можна представити як набір функціональних блоків (n штук). Будемо говорити, що заявка знаходиться в системі, якщо вона поступила до серверу та знаходиться на обслуговуванні в одному з функціональних блоків. Успішне проходження всіх функціональних блоків у заданій послідовності забезпечує успішне обслуговування заявки на сервері. Час обслуговування заявки на сервері є обмеженим, тому якщо заявка перебуває в системі довше заданого часу, тоді вона знімається з обслуговування, і клієнту повідомляється, що мережа зайнята.

Для обслуговування абонентських заявок у функціональних блоках застосовуються ресурси, такі як оперативна пам'ять, процесорний час, мережевий ресурс (зайнятість каналу сигнальним трафіком), об'єм постійної пам'яті на дисках, (всього G ресурсів позначаються індексом Sg). Обслуговування кожної заявки у заданому функціональному блоці потребує заданого об'єму ресурсу. Відомо, протягом якого часу використовується кожний з ресурсів при обслуговуванні у заданому функціональному блоці (статистичні дані). Якщо ресурс зайнятий, тоді

заявка очікує звільнення ресурсу. Таким чином, виникають затримки в обслуговуванні, які призводять до втрати успішно обслужених заявок.

Негативна дія заявок, які обслуговувалися в системі, але процедуру їх обслуговування було завершено не успішно, полягає в тому, що протягом часу обслуговування, такі заявки займали ресурс, створюючи додаткове безрезультатне навантаження на систему.

Саме тому для запобігання перевантаженню система керування вхідним трафіком будується з використанням даних системи моніторингу щодо використання ресурсів різними типами сервісу, таким чином актуальним є зменшення кількості необслужених заявок через брак ресурсів серверу.

На рис. 4.8 показана схема обслуговування однієї заявки (позначена трикутником), яка надійшла на обслуговування в систему. Обслуговування кожної заявки складається з виконання набору операцій, які виконуються у заданій послідовності, послідовність операцій розділяється на етапи. Логічно завершений етап називатимемо функціональним блоком. В даній моделі передбачається, що сервер має певні ресурси, які використовуються функціональними блоками у необхідному обсязі залежно від кількості операцій, що виконуються у межах відповідного функціонального блоку, та кількості заявок, які знаходяться на відповідному етапі обслуговування (у розглядуваному функціональному блоці). Під сервером розуміємо всі операції, що виконуються над заявками всіх типів сервісів у всіх функціональних блоках, та всі наявні для цього ресурси. На рис. 4.8 заявка обслуговується у першому функціональному блоці (Б1), відповідно обслуговування потребує використання групи ресурсів, які позначено як S1, S2, … . Стрілками показано необхідність використання певного ресурсу. Обслуговування заявки буде успішним, якщо вона пройде від «Входу» до «Виходу» в межах заданого часу.

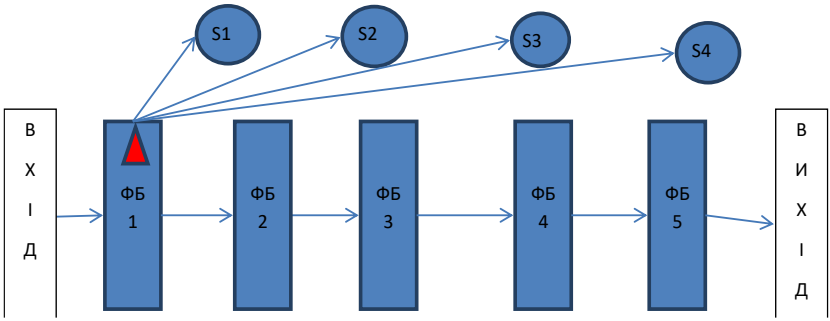


Рис. 4.8 Схема обслуговування заявки

4.2.3. Розподіл технічних засобів для забезпечення обслуговування заявок на тарифікацію різних типів сервісів

Задача розподілу ресурсів між різними типами сервісів з урахуванням економічної ефективності передбачає визначення оптимальної долі ресурсів, які необхідні для обслуговування заявок кожного типу сервісу. Вирішення такої задачі дозволить максимізувати економічну ефективність обслуговування викликів та врахувати статистичні дані розподілу навантажень різних типів сервісів.

Вхідні дані.

1. Об'єм кожного виду ресурсу, необхідний для обслуговування певного типу заявок у функціональному блоці.

2. Прибуток від кожного типу заявок.

3. Максимально допустимий об'єм ресурсів серверу тарифікації, що розподіляється.

Необхідно визначити розподіл кількості ресурсів, що виділяється для обробки певного типу заявок.

Нехай k_i – шукана кількість заявок і-го типу, які перебувають у системі тарифікації;

v_{ij}^{Rg} – кількість g-го ресурсу, необхідних для обслуговування однієї заявки і-го типу у j-му функціональному блоці;

S_i – кількість прибутку, який можна отримати від однієї заявки і-го типу.

Тоді g-го ресурсу, необхідних для проходження однієї заявки і-го типу через всі функціональні блоки буде рівна $\sum_j v_{ij}^{Rg}$. Загальна кількість прибутку, отриманого від обслуговування усіх заявок і-го типу дорівнює

$$S = \sum_i k_i S_i$$

Загальна кількість ресурсу, яку займають всі заявки *i*-го типу, які одночасно знаходяться у системі по всіх функціональних блоках дорівнює

$$v^{Rg} = \sum_i k_i (\sum_j v_{ij}^{Rg}) \quad (4.26)$$

Оскільки необхідно максимізувати прибуток, тому цільова функція буде наступною:

$$\sum_i k_i S_i \to max \quad (4.27)$$

При умові, що не буде використано об'єм ресурсу більший ніж доступний. Додаткові обмеження задачі, отримані в наслідок аналізу добової статистики, представляють відношення середньостатистичної кількості заявок різних типів сервісів

$$\sum_i k_i \Big(\sum_j v_{ij}^{Rg} \Big) \le V_{Rg,}\ g = \overline{1,G} \quad (4.28)$$

де V_{Rg} – об'єм g-го ресурсу, який може надати сервер тарифікації;

ki– шукана кількість заявок на тарифікацію, що надійшли від і-го типу сервісів;

При вирішенні задачі розподілу ресурсів необхідно також враховувати статистичні дані про надходження заявок на сервер у різні періоди часу.

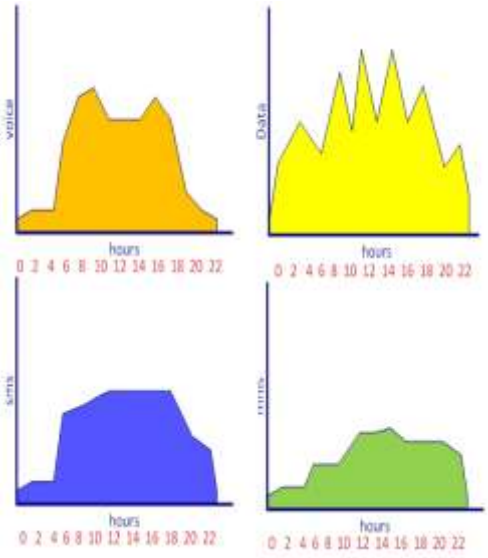


Рис. 4.9 Статистика надходження заявок протягом доби

На Рис. 4.9 зображена статистика надходження заявок на сервер протягом доби. Оскільки співвідношення між кількістю різних заявок є відомою величиною, нескладно накласти додаткові обмеження на кількості заявок, які будуть отримані на виході задачі.

Додаткові обмеження задачі, отримані в наслідок аналізу добової статистики – це відношення середньостатистичної кількості заявок різних типів сервісів:

$$a_{ij} \leq \frac{k_i}{k_j} \leq b_{ij}$$

де k_i та k_j – шукані кількості заявок на тарифікацію, що надійшли від i-го та j-го типів сервісів відповідно; a_{ij} та b_{ij} – числові границі розраховані на основі статистичних даних про середньогодинну кількість заявок на тарифікацію, що надходять до системи; m – кількість сервісів, для яких здійснюється тарифікація.

Результатом розв'язку поставленої оптимізаційної задачі на умовний екстремум є послідовність $\{k_i\}$ $(i = \overline{1,m})$, кількість заявок кожного типу сервісу, що може одночасно обслуговуватися у системі тарифікації. Тоді за формулою (4.26) можна розрахувати кількість ресурсу, щоповинна виділятися для обслуговування.

В результаті вирішення задачі визначено розподіл загального об'єму технічних ресурсів між заявками на тарифікацію різних типів сервісів під час обробки викликів в OCS, що забезпечує отримання максимального прибутку.

Оцінка ефективності

Імітаційна модель, створена для реалізації запропонованого методу показала, що при фіксованому розподілі доступних технічних ресурсів між різними типами заявок, економічна ефективність роботи системи покращена на 5%, крім того, кількість втрачених заявок через перевищення допустимого часу обслуговування, а також внаслідок роботи алгоритмів раннього попередження перевантажень, які

традиційно застосовуються за для уникнення перевантажень скоротилася в середньому на 1%.

Було проведено два експерименти роботи імітаційної моделі для трьох типів заявок на тарифікацію: текстові повідомлення (sms), мультимедійні повідомлення (mms) та дзвінки. В першому на обслуговування виділявся обмежений технічний ресурс, який спільно використовувався для обслуговування заявок на тарифікацію, в процесі обслуговування застосовувався алгоритм RAD (Random early detection), який передбачав відкидання заявок в разі перевантаження серверу, крім того заявки втрачалися якщо час перебування їх у системі перевищував максимально допустиме значення.

В другому експерименті технічні ресурси розділялися відповідно до запропонованого методу. На основі статистичних даних оператора мобільного зв'язку про кількість заявок, що надходить на сервер протягом дня та кількість прибутку від обслуговування заявки певного типу, був розрахований об'єм ресурсу, що виділяється для обслуговування кожного з трьох типів заявок. Наприклад, кількість постійної пам'яті складає 1,25Мбайт для голосу, 9,15Мбайт для sms і 17,5Мбайт для mms- повідомлень. В ході експериментів була імітована робота системи тарифікації з вхідним мультисервісним потоком максимально наближеним до реального.

Не зважаючи не те, що в сумі кількість втрачених заявок суттєво не змінилася, однак спостерігався перерозподіл кількості заявок між різними типами, які були втрачені, за рахунок чого спостерігалось зменшення розміру сумарного втраченого прибутку в другому експерименті.

Розрахунок проводився виходячи з того, що одночасно в середньому надходить по 10 тисяч заявок на тарифікацію, кожного з типів сервісу. В результаті в ході першого експерименту середні втрати заявок склали 2%, в тому числі втрати заявок на тарифікацію sms сервісів склали – 2,0%, mms сервісів – 2,55%, голосу – 1,45%. В ході другого експерименту середні втрати заявок склали 1,8% в тому числі втрати заявок на тарифікацію sms сервісів склали – 1,65%, mms сервісів – 1,25%, голосу – 2,5%.

В результаті експерименту було визначено, що для першого випадку кількість втрат складає: 870 у.о. для заявок 1 типу (дзвінки), 1800 у.о. для заявок 2 типу (sms) та 3825 у.о. для заявок 3 типу (mms). Теоретично можлива кількість отриманого прибутку при умові, що всі заявки, які надійшли на сервер будуть оброблені становить 100000 у.о. Загальна кількість втрат складає 6495 у.о., що становить 6,5% від можливого прибутку. Для випадку з застосуванням запропонованого методу були отримані наступні дані про кількість втраченого прибутку: 1500 у.о. для заявок 1 типу (дзвінки), 1485 у.о. для заявок 2 типу (sms) та 1875 у.о. для заявок 3 типу (mms). Загальна кількість втрат складає 4860 у.о., що становить 4,86% від можливого прибутку.

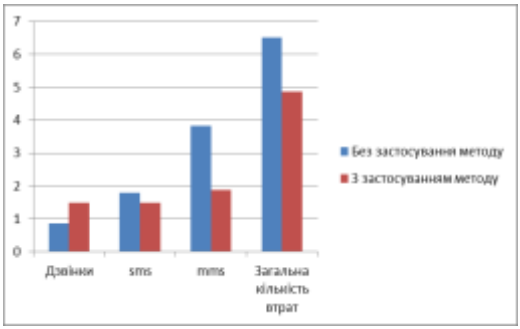


Рис. 4.10 Порівняння розміру втраченого прибутку у відсотках

На рис. 4.10 представлено порівняння розміру втраченого прибутку у відсотках для кожного типу заявок та загальної кількості втрат до та після застосування методу розподілу ресурсів між різними типами заявок. Результат експерименту показав, що загальна кількість втраченого прибутку зменшилася на 1,64%.

4.3. Керування вхідним потоком заявок на тарифікацію в OCS

4.3.1. Рівні керування в системі онлайн тарифікації

Роботу підсистем серверу тарифікації (Online Charging System – OCS) 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, відрізняються типом сервісу, який вони представляють, обслуговування черг передбачає послідовне виконання операцій з модулем SEE, які наведені вище. Процес керування передбачає керування чергами вхідних заявок: формування черг за типом сервісу, застосування методів групи WRAD, тощо. Таким чином, перший рівень - система масового обслуговування, яка обслуговує 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. А розрахунок схеми розділення ресурсів, що дозволить у періоди високого навантаження розподіляти ресурси залежно від економічної ефективності та співвідношень кількості заявок різних типів, що надходять.

Архітектура дворівневої системи керування представлена на рис. 4.11.

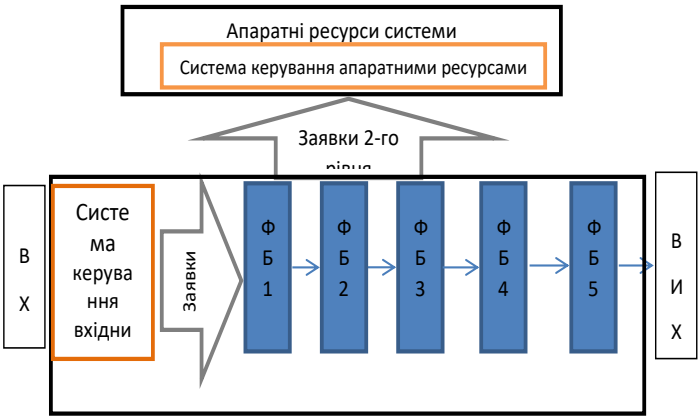


Рис. 4.11 Схема обслуговування заявок на сервері оператора мобільного зв'язку

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 рівня, 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.

Розроблено ряд методів керування вхідним потоком заявок на сервер тарифікації з відомими технічними характеристиками ресурсів. Застосування того чи іншого методу залежить від способу функціонування системи, перший метод передбачає контроль кількості заявок, які у поточній момент обслуговуються у ресурсозатратному функціональному блоці. Другий метод передбачає розрахунок схеми подавання заявок до системи тарифікації, що дозволяє без завантаження системи моніторингу забезпечити оптимальний план обслуговування заявок, який забезпечує ефективне використання наявних ресурсів серверу.

4.3.2. Передумови створення методів згладжування вхідного навантаження на сервер on-line тарифікації

Випадковість процесу надходження запитів на обслуговування до серверу мобільного зв'язку обумовлює нерівномірність вхідного потоку. Вважатимемо, що вхідний потік запитів на обслуговування описується за допомогою закону Пуассона.

Параметри серверу, що обслуговує заявки розраховані для середніх значень параметрів вхідного потоку, однак наявні пікові значення кількості заявок, що надійшли одночасно. Затримки у обслуговуванні заявок під час проходження через відповідні функціональні блоки виникають внаслідок суперпозиції великої кількості заявок, що потребують значної кількості ресурсу, потрапивши у певний функціональний блок.

Назвемо ресурсозатратним функціональним блоком, такий блок якому для виконання операцій із заявками необхідна значна кількість ресурсу. Більш детально можна розглянути на прикладі.

Приклад 1. Нехай m=1, n=4. Тобто система обслуговує 1 тип заявок у 4-х функціональних блоках, для обслуговування потрібна тільки оперативна пам'ять (G=1, позначимо s1).

Нехай відомі значення математичного очікування часу перебування заявки у кожному функціональному блоці: t_{11}=1, t_{12} =3, $t_{13} = 4$, $t_{14} = 2$, які зведені до матриці T.

$$T = \begin{pmatrix} 1 \\ 3 \\ 4 \\ 2 \end{pmatrix}$$

Тобто в середньому заявка проводить у системі 10 одиниць часу, позначимо його за T_Σ (T_Σ=10).

Нехай відомо, що для виконання операцій з інформаційним потоком 1-го типу у 1-му функціональному блоці потрібен об'єм оперативної пам'яті 1 одиниця ($v_{11}^{s1} = 1$), для виконання операцій з 1-м інформаційним потоком у функціональному блоці 2 потрібен об'єм оперативної пам'яті 0 одиниць ($v_{12}^{s1} = 0$), відповідно: $v_{13}^{s1} = 4$, $v_{14}^{s1} = 2$. Позначимо дані , що характеризують потребу у ресурсі S1 у відповідному функціональному блоці через матрицю V^{S1} , яка має вигляд:

$$V^{S1} = \begin{pmatrix} 1 \\ 0 \\ 4 \\ 2 \end{pmatrix}$$

На Рис. 4.12 наведено використання ресурсу s1 протягом часу обслуговування однієї заявки 1-го типу в усіх 4-х функціональних блоках.

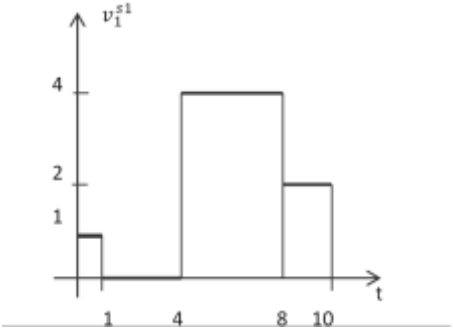


Рис. 4.12 Приклад використання ресурсу в часі

Тепер змоделюємо випадковий процес надходження заявок. Застосуємо зворотною системою відліку часу. Приймемо за нульовий час закінчення обслуговування заявки, тобто t_0=10, далі позначимо періоди часу коли заявки переходять між функціональними блоками: $t_1 = t_0 - t_{11} = 9$, $t_2 = t_1 - t_{12} = 6$, $t_3 = t_2 - t_{13} = 2$, $t_4 = t_3 - t_{14} = 0$. 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 $[t_2, t_1]$ в момент часу t_0, обслуговуються у другому функціональному блоці, $[t_3, t_2]$ $(t_4, t_3]$ – у четвертому.

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

- щоб уникнути дефіциту ресурсів необхідно, щоб два і більше сплески навантаження вхідного потоку не обслуговувалися одночасно у функціональних блоках, що потребують значної кількості ресурсів:

- для запобігання цьому застосовується затримка частини заявок, які припали на сплеск навантаження. Час затримки визначається так, щоб затримані заявки не поступали в систему доти, доки попередній сплеск навантаження не буде успішно обслуженим в ресурсозатратному функціональному блоці.

Повертаючись до прикладу, із матриці V^{S1} видно, що найбільш ресурсозатратним функціональним блоком є третій функціональний блок, оскільки потребує найбільшого об'єму ресурсу $v_{13}^{s1} = 4$. Відповідно до запропонованого методу не мають допускатися в систему сплески навантаження, коли різниця між часом надходження двох сплесків менша тривалості обслуговування заявки у ресурсозатратному функціональному блоці ($t_{13} = 4$).

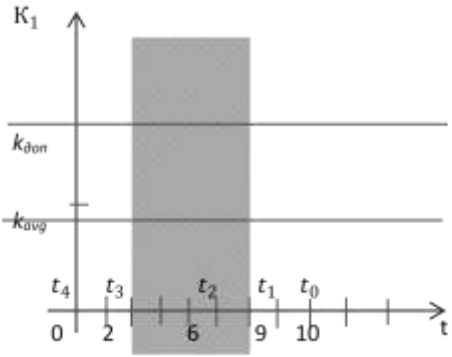


Рис. 4.13 Інтервали часу, в які не допускається більше одного сплеску навантаження вхідного потоку

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

Припустимо, що у прикладі інтервал дискретизації часу дорівнює 1. В табл. 4.1 наведені розрахунки потреби ресурсу у момент часу t_0=10, якщо заявки надходили у кількості К1 (другий стовпчик таблиці). Як видно сумарний об'єм ресурсу, у цей момент часу дорівнює 216 одиниці. Загальний об'єм ресурсу складає 200 одиниць.

Табл. 4.1
Розрахунки

Время t	К-ть заявок, що надійшли (К1)	об'єм ресурсу для однієї заявки $(v_{ij}^{s1}, j = \overline{1,n})$	загальний об'єм ресурсу
1	5	2	10
2	1	2	2
3	21	4	84
4	4	4	16
5	20	4	80
6	5	4	20
7	3	0	0
8	2	0	0
9	6	0	0
10	4	1	4
		сумарний об'єм потрібного ресурсу	**216**

Це означає, що виникла ситуація з дефіцитом ресурсу, яка негативно впливає на обслуговування не тільки заявок, які прийшли у моменти $t = t_{max1} = 3$ і $t = t_{max2} = 5$, коли спостерігалось пікове навантаження, але й на всі інші заявки. Наслідком цього може бути збільшення часу користування ресурсом, або очікування

деякої частини заявок звільнення потрібної кількості ресурсу. Що свідчить про неефективне використання ресурсу.

Припустимо, що в прикладі допустиме значення кількості заявок, що можуть одночасно надходити на обслуговування в систему дорівнює 13 ($k_{\text{доп}} = 13$). Це означає, що за наведеним методом допустима частина заявок (13 шт), що надійшли у момент t=5 поступають на обслуговування, а 7 заявок затримуються на дві одиниці часу, тобто на стільки, щоб час затримки не перевищував t_{13}:

$$t_{del} = t_{13} - (t_{max2} - t_{max1}) = 4 - (5 - 3) = 2.$$

Табл. 4.2
Розрахунки

ремя t	К-ть заявок, що надійшли (К1)	Об'єм ресурсу для однієї заявки ($v_{1j}^{s1}, j = \overline{1,4}$)	Загальний об'єм ресурсу
	5	2	10
	1	2	2
	21	4	84
	4	4	16
	13	4	52
	5	4	20
	10	0	0
	2	0	0
	6	0	0
0	4	1	4
		сумарний об'єм потрібного ресурсу	188

При введенні затримки зміниться сумарний об’єм ресурсу, який потрібен в момент часу t_0=10, щоб обслуговувати всі заявки наявні в системі, і буде дорівнювати 188 . Результати представлені в табл.4.2.

4.3.3. Метод контролю перевантажень в системі онлайн тарифікації (забезпечується неперервною роботою системи моніторингу)

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 є більшою допустимого значення.

Для керування процесом обробки заявок з метою запобігання дефіциту ресурсу в системі керування пропонується використання стратегії наведеної у попередньому пункті.

Функціональні блоки, в яких постійно виникають затримки через нестачу технічних ресурсів серверу тарифікації під час обробки певних типів заявок, що потребують значних обсягів обчислень за алгоритмами, покладеними в основу програмного забезпечення сервісів, вважатимемо ресурсозатратними ФБ.

Визначення ресурсозатратних ФБ для вибраного типу сервісу здійснюється на основі інформації про об'єм ресурсу, який є необхідним для здійснення операцій, передбачених функціональним блоком для обслуговування заявки заданого типу сервісу.

Для всіх ресурсів системи, для кожного типу сервісу знаходяться функціональні блоки, де відповідне обслуговування є ресурсозатратним.

Проводиться аналіз статистичної інформація про тривалість обслуговування заявки заданого типу сервісу у кожному функціональному блоці. Розраховується середнє значення тривалості обслуговування заявки вибраного типу у ресурсозатратному функціональному блоці.

Постановка задачі методу керування вхідним потоком на сервер мобільного зв'язку.

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} ($\tau_{ij}^{Sg} \leq t_{ij}$). Інформація про тривалість обслуговування зведена в матриці $\mathcal{T}^{Sg} = \left\{\tau_{ij}^{Sg}\right\}_{i=\overline{1,m},j=\overline{1,n}}$. 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 і-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 і-го типу сервісу задати допустиму кількість заявок які можуть одночасно поступати на обслуговування в систему ($k_{i\,\text{доп}}$). 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 ($i_{g1}j_{g1}$, …, $i_{g\,gmax}j_{g\,gmax}$). 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 F_1 увійшли пари де перший елемент дорівнює 1, в F_2 увійшли пари де перший елемент дорівнює 2, тощо. Якщо деяка r-та підмножина ($r\epsilon\overline{1,m}$) буде порожньою (F_r={∅}), тоді для заявок r-го типу сервісу не будуть застосовуватися затримки заявок, що надійшли у сплесках навантаження вхідного потоку. Для всіх підмножин F_d ($d\epsilon\overline{1,m}$), де міститься один

елемент, виконуються дії п.4. Для всіх підмножина 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 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 F_p впорядковуються за другим елементом. Припустимо, що множина F_p складається з двох елементів: (p, $f1_p$) і (p, $f2_p$), 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 $f1_p$ і $f2_p$. 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 $f2_p - f1_p = x > 1$, 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 $T_{i\ max}$, де i – тип сервісу, для якого зафіксовано сплеск навантаження. Для сервісів r-того типу (см. п.3) множина $T_{r\,max}$ не створюється.

Для сервісів типу d, для елементів множини $T_{d\ max}$ 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 $(k_d(t_{d\,\max w+1}) - k_{d\,доп})$ 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 $(k_d(t_{d\,\max w+1}) - k_{d\,доп})$ в сумі дають значення більше допустимого $k_{d\,доп}$. Тоді надлишок передається у наступний момент часу $(t_{d\,\max w+1} + y1 + 1)$, *процедура згладжування навантаження.*

Для сервісів типу p, для елементів множини $T_{p\ max}$ 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 $\left(k_p\left(t_{p\,\max w+1}\right) - k_{p\,доп}\right)$ 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 $(k_p(t_{p\max w+1}) - k_{p\,\text{доп}})$ затримується на час y2, у разі потреби застосовується процедура згладжування навантаження.

якщо $f2_{\text{р}} - f1_{\text{р}} = x > 1$, тоді досліджується значення $\sum_{q=1}^{x-1} t_{p\,(f1_p+q)} - (t_{p\max w+1} - t_{p\max w}) = y4$, якщо y4<0, тоді частина заявок $(k_p(t_{p\max w+1}) - k_{p\,\text{доп}})$ 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.

Схема алгоритму методу керування наведена на рис. 4.14.

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.

4.3.4. Імітаційна модель методу контролю перевантажень в системі онлайн тарифікації

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.

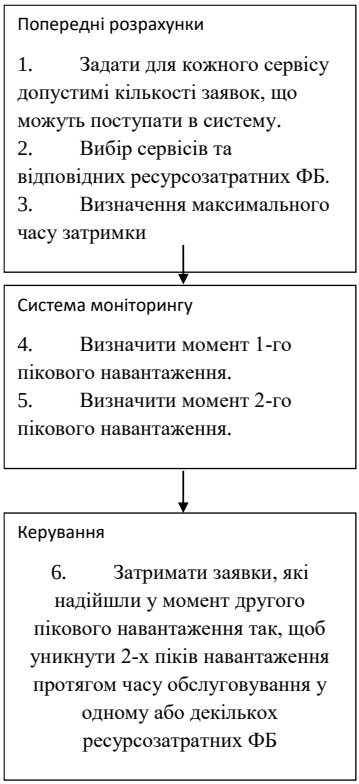


Рис. 4.14 Алгоритм методу контролю перевантажень в системі онлайн тарифікації

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

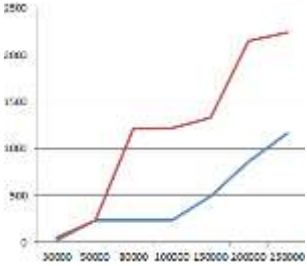


Рис. 4.15 Кількість втрачених заявок

Динаміку залежності кількості втрачених заявок від інтенсивності вхідного потоку показано на Рис. 4.15. На Рис. 4.15 видно зменшення втрат заявок через перевищення допустимого часу обслуговування. Червоним позначена лінія втрат пакетів без застосування запропонованого методу керування вхідним потоком, синім позначені результаті моделювання за допомогою запропонованого методу керування.

4.3.5. Метод керування вхідним потоком заявок на тарифікацію (працює незалежно від системи моніторингу)

Запропонований вище метод контролю перевантажень в системі онлайн тарифікації передбачає відслідковування піків навантаження, та введення затримки для деякої частини заявок другого піку, що дозволяє уникнути перевантаження ресурсів серверу. Однак реалізація стратегії 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:

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.

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

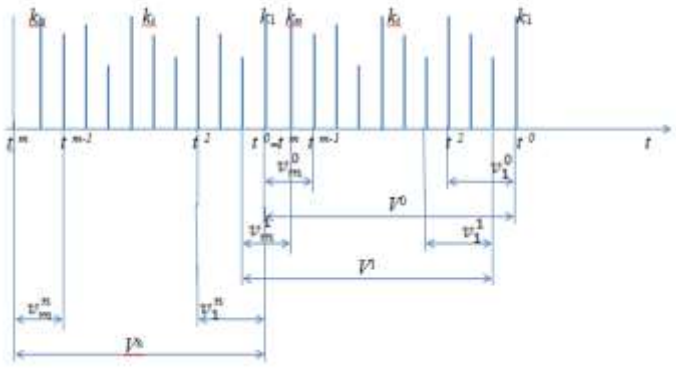


Рис. 4.16 Схема згладжування вхідного навантаження, з урахуванням об'єму ресурсу, що використовується

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 $\{k_i\}$.

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.

4.4. Метод організації розкладу включення серверів і можливості використання необмеженої кількості ресурсів для забезпечення потреб біллінгової системи

4.4.1. Обслуговування заявок на сервері з необмеженим ресурсом

Сьогодні все більшої популярності набуває використання так званих віртуальних серверів. Віртуальний сервер (або Cloud Server) – це повноцінна інфраструктура, побудована за моделлю хмарних обчислень (cloud computing).

На відміну від моделі зберігання даних на власних виділених серверах, у випадку використання віртуальних серверів, їх структура і кількість в загальному випадку не видна користувачеві. Уся інформація зберігається і обробляється у хмарі, яка являє собою, з точки зору клієнта, один великий віртуальний сервер. Основні переваги віртуальних серверів:

Гнучкість ресурсів. Оренда сервера у хмарі дозволяє забезпечити його високу масштабованість. Такий сервер легко налаштовується під збільшення навантаження, наприклад, можна легко додати оперативної пам'яті або дискового простору. Так само легко можна зменшити дані параметри віртуальної системи. Сервер стає «резиновим» у відношенні своїх ресурсів.

Швидкодія. Віртуальний сервер працює значно швидше ніж звичайний, а також знижуються часові затрати на впровадження і оперативний перерозподіл ресурсів. Висока швидкість розгортання системи.

Безпека. На віртуальному сервері клієнт має можливість власноручно контролювати користувачів, процеси, а також використовувати власну політику безпеки. Підвищення рівня безпеки Cloud Server відбувається також за рахунок зведення до мінімуму «людського фактору».

Мобільність. Доступ до серверу можна отримати з будь-якої точки земної кулі.

4.4.2. Архітектура віртуального сервера

Хмарне зберігання розвивається в трьох напрямах, один з яких допускає злиття двох інших для досягнення економічної ефективності і безпеки (Рис. 4.17).

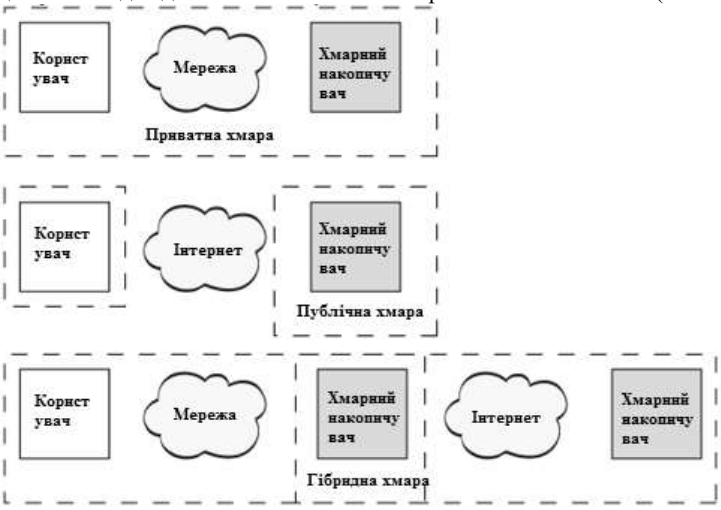


Рис. 4.17 Моделі зберігання даних у хмарах

Приватна хмара – інфраструктура, призначена для використання однією організацією, що включає декілька споживачів(наприклад, підрозділів однієї організації), можливо також клієнтами і підрядниками цієї організації. Приватна хмара може знаходитися у власності, управлінні і експлуатації як самої організації, так і третьої сторони(чи якій-небудь їх комбінації), і воно може фізично існувати як усередині, так і поза юрисдикцією власника.

Публічна хмара – інфраструктура, призначена для вільного використання широкою публікою. Публічна хмара може знаходитися у власності, управлінні і експлуатації комерційних, наукових і урядових організацій. Публічна хмара фізично існує в юрисдикції власника - постачальника послуг.

Гібридна хмара – це комбінація з різних хмарних інфраструктур, що залишаються унікальними об'єктами, але пов'язаних між собою стандартизованими або приватними технологіями передачі даних і додатків (наприклад, короткочасне використання ресурсів публічних хмар для балансування навантаження між хмарами).

Постачальники традиційних IT-послуг, хостингових систем і хмарних служб використовують для створення своїх пропозицій передусім мережеві системи і технології обробки і зберігання даних. Для виконання широкого спектру вимог до хмарних систем обробки і зберігання даних потрібні різні типи серверів, мереж і облаштувань зберігання даних (наприклад, щільні стійкові і блейд-сервери з різною кількістю роз'ємів і потоків, ядрами різної тактової частоти, різним об'ємом пам'яті і можливостями розширення входів/виходів).

Різні рівні зберігання даних забезпечуються надпродуктивними твердотілими накопичувачами і продуктивними жорсткими дисками середньої і великої місткості.

Функції управління зберіганням включають захист даних (забезпечення високої доступності, резервне копіювання і засоби відновлення після аварій), а також скорочення займаних площ з оптимізацією використання простору (стискування, дедуплікація і динамічний розподіл), що дозволяє зберігати більше інформації впродовж тривалішого часу при менших витратах.

Загальнодоступні хмари і служби бувають платними і безкоштовними і надають різні набори функцій. В якості прикладів можна привести Amazon Web Services (AWS), Google Docs і ПЗ для резервного копіювання Seagate® EVault®. Управляються загальнодоступні хмари їх власниками, а клієнти тільки використовують служби, що надаються. Приватними хмарами володіють або ж користуються і управляють самі організації; це нагадує механізм надання традиційних IT-служб. Приватні хмари, для створення яких використовуються загальнодоступні компоненти або служби, а також видалені ресурси від різних постачальників, називаються гібридними.

4.4.3. Тарифні плани в Cloud

На сьогоднішній день існує велика кількість компаній, які пропонують послугу оренди хмари для створення віртуальних серверів. Вартість послуги залежить від об'єму орендованого сервера, потужності процесора та тривалості оренди. В табл. 4.3 наведені тарифні плани при оренді віртуального сервера компанії «LikeHost».

Табл. 4.3

Тарифні плани при оренді віртуального сервера компанії «LikeHost»

Тарифний план	Base-VDS-10	Base-VDS-20	Base-VDS-40	Base-VDS-60	Base-VDS-80
Об'єм диску (GB)	10	20	40	60	80
Відказостійкість	Потрійна реплікація даних	Потрійна реплікація даних	Потрійна реплікація даних	Потрійна реплікація даних	Потрійна реплікація даних
Ліміт трафіка	Необмеж.	Необмеж.	Необмеж.	Необмеж.	Необмеж.
CPU (MHz)	2000	4000	4000	8000	8000
RAM (MB)	256	512	768	1024	2048
Об'єм бекапа диску	10	20	40	60	80
Вартість за 1 рік ($)	150	210	265	395	565

- Об'єм диску – простір на жорсткому диску сервера, виділений для віртуального сервера;
- Ліміт трафіка – об'єм даних, щи проходять через сервер;
- CPU – потужність процесора;
- RAM – об'єм оперативної пам'яті;

Об'єм бекапа диску – прості на жорсткому диску сервера, розташованого в дата-центрі іншої країни для бекапа (резервного копіювання) даних віртуального сервера.

4.4.4. Балансування навантаження у хмарах

При проектуванні і створенні хмарних рішень важливе місце займає забезпечення надійності і оптимальності використання ресурсів - моніторинг, або отримання інформації про доступність і завантаження апаратних ресурсів платформ з копіями розподіленого застосування, і балансування, тобто розподіл призначених для користувача запитів, що надходять, між наявними апаратними платформами.

Можна виділити наступні класи рішень, використовувані при побудові багатовузлових систем: балансування з пропуском трафіку через один пристрій балансування; балансування засобами кластера; балансування без пропуску трафіку через один пристрій балансування.

Рис. 4.18 Принципи балансування навантаження : а) балансування з пропуском трафіку через один пристрій балансування; б) балансування засобами кластера; в) балансування без пропуску трафіку через один пристрій балансування

1) Трафік через один пристрій

В цьому випадку взаємодія усіх користувачів проходить через виділений пристрій, який на підставі заздалегідь заданих статичних правил розподілу навантаження або, орієнтуючись на час обробки запитів серверами, робить комутацію встановлюваних сесій зі зміною окремих полів заголовків пакетів, що пересилаються.

2) Балансування засобами кластеру

Кластеризація дозволяє управляти групою незалежних серверів як єдиною системою, що підвищує відмовостійкість, спрощує управління і дозволяє добитися більшої масштабованості. Сервіс балансування в кластері забезпечує розподіл потоку IP- даних між вузлами. У кластері декілька апаратних платформ найчастіше розділяють єдину IP- адресу.

Для кожного вузла, що бере участь в розподілі навантаження, адміністратор може вказати його конкретну долю навантаження, або за замовчуванням навантаження рівномірно розподілятиметься між вузлами. Запити клієнтів статистично розподіляються між вузлами, щоб кожен сервер обробляв рівно свою частину, а розподіл навантаження змінюється при включенні або видаленні вузла кластера. Розподіл навантаження не змінюється при зміні параметрів навантаження на апаратні платформи серверів. Для програм з великою кількістю клієнтів і породжуваних ними коротких запитів, таких як веб-сервіси, подібний механізм

розподілу навантаження забезпечує ефективне балансування навантаження і швидку реакцію на зміну складу вузлів кластера.

3) Балансування без жодного пристрою балансування

Цей варіант балансування організовується шляхом відправки первинного запиту від користувача (наприклад, запиту IP- адреси для встановлення з'єднання або першого запиту на встановлення HTTP- сесії) на виділений сервер-балансувальник. Після отримання запиту балансувальник вказує, з яким сервером додатків з інфраструктури хмари продовжувати роботу(повертаючи відповідну IP-адресу або перенаправляючи її засобами HTTP Redirect). Подальша взаємодія відбувається без участі балансувальника.

4.4.5. Використання віртуальних серверів для обслуговування викликів у системах мобільного зв'язку

На сьогоднішній день оператори зв'язку лише починають впроваджувати хмарні технології. Так, багато мобільних операторів пропонують своїм клієнтам таку послугу як віртуальна АТС. Віртуальна АТС- це послуга для компаній, яка замінює фізичну офісну МІНІ-АТС і навіть call-центр. Суть послуги полягає в тому, що клієнт (компанія) отримує в повне користування IP-АТС, фізично розміщену у провайдера. Віртуальна АТС надає усі стандартні можливості IP-АТС: багатоканальний номер, запис розмов, голосові вітання, переклад виклику - усе це і багато що інше доступне через Інтернет без придбання устаткування.

Так само віртуальні сервери можна використовувати для обслуговування і тарифікації викликів. Це може значно спростити процес розподілу ресурсів на сервері, забезпечить його масштабованість, підвищить безпеку і відказостійкість. Віртуальний сервер може бути використаний двома способами. По-перше, він може бути додатком до основного фізичного сервера і використовуватись коли виникає потреба у збільшенні кількості ресурсів при збільшенні навантаження. В іншому варіанті все обладнання для обслуговування і тарифікації викликів може розміщуватись у хмарі. В такому випадку фізичний сервер взагалі не потрібен і процес обслуговування запитів буде повністю проходити на віртуальному сервері.

Обидва способи значно спрощують процес розподілу ресурсів, необхідних для обробки викликів, так як ресурси хмарних технологій являються практично необмежними.

4.4.6. Проблеми організації роботи біллінгової системи на множені технічних засобів

Однією з перешкод на шляху до массового використання в системах онлайн тарифікації залученних зовнішніх технічних ресурсів є відсутність відповідних методів та алгоритмів організації злагодженої роботи групи серверів. В даній монографії ми розглянемо два алгоритма оцінки кількості серверів, які мають одночасно обслуговувати заявки на тарифікацію. Технічна система, що забезпечує потреби системи он-лайн тарифікації є складною. Залучення додаткових серверів

може здійснюватися як для забезпечення всієї системи в цілому так і для покращення роботи окремих функціональних блоків.

4.4.7. Розробка розкладів включення обладнання. Динамічна система моніторингу

В умовах необхідності заощаджувати витрати на утримання серверів, будь то орендна плата за використання ресурсів, які розташовані у хмарах, або енергетичні ресуси для забезпечення роботи власних серверів. Компанії оператору необхідно розробити план включення серверів, який би задовольняв всі потреби користувачів у поточний період часу.

Отже, якщо відома поточна статистика навантаження, що створюється абонентськими заявками на тарифікацію, також відома верхня межа допустимої кількості заявок, які одночасно можуть оброблятися на потужностях серверів, які включені у поточний момент часу, тоді можна визначити ймовірність того що протягом заданого часу кількість заявок, що надійде не перевищить допустимого значення.

Алгоритм методу контролю достатності ресурсів системи для обробки заявок на тарифікацію

Вхідні дані:

1) інтервал T1 – інтервал часу протягом якого буде проводитися аналіз статистики;
2) інтервал dt – інтервал дискретизації часу (малий інтервал);
3) дані системи моніторингу про кількість заявок на тарифікацію;
4) час протягом якого може бути завантажено додатковий сервер – T;
5) M – максимально допустима кількість заявок, які можуть обслуговуватися при поточній кількості ввімкнених серверів (ресурсів).

Табл. 4.4
Вхідні дані

t	t_0- dt	t_0-2dt	t_0-3dt	…	T_1
X	x(t)	x(t)	x(t)		

На рис. 4.19 x_0 – кількість заявок що перебуває у системі в момент часу t_0, x=at + b – пряма пубудована на основі даних статистики за допомогою метода найменших квадратов; точки тренду – це можливі значення кількості заявок що надійдуть на обслуговування в систему протягом настояного часу T. Задача оцінити ймовірність того, що випадковий процес вийде за межі допустимого значення M.

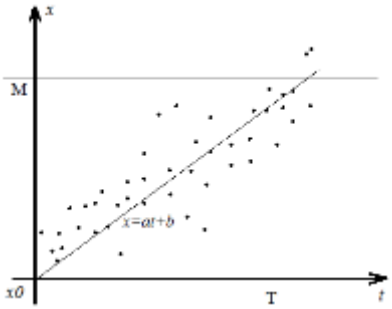


Рис. 4.19 Аналіз динаміки навантаження на сервер онлайн тарифікації

Алгоритм методу включає в себе три основні кроки:

1.Аналіз статистичних даних за час, де t_0 – поточний час, для якого виконується розрахунок. На основі статистичних даних системи моніторингу для пари значень (t,x) за методом найменших квадратів розраховується оцінка коефіцієнту *a* для прямої, яка апроксимує значення вхідного навантаження з табл. 4.4 x=at+b

2.Оцінити ймовірність P_T того, що протягом заданого часу T кількість заявок, яка надійде в систему не перевищить допустиме значення M. Оцінка ймовірності розраховується за формулою

$$P_{T=}P(\bar{x} + 3\sigma > M)$$

де σ – середньоквадратичне відхилення:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}=$$

$\mathbf{x_i}$– поточне навантаження на систему,

$\bar{x}$– середнє значення кількість заявок

3.Якщо ймовірність помилки P_T протягом допустимого часу T перевищує допустимий поріг, тоді відбувається включення додаткового серверу.

Застосування даного алгоритму дозволить контролювати динаміку зростання навантаження на сервер онлайн тарифікації та визначити момент коли необхідно залучити для обслуговування додаткові ресурси. Час T обирається виходячи з міркувань скільки потрібно часу від запуску до повної працездатності додаткового серверу.

Як видно з рис. 4.19 лінія $x=at+b$ проходить нижче від точки з координатами (T, M), оскільки даний метод враховує структуру випадкового процесу та середньоквадратичне відхилення яким характеризується випадковий працес надходження заявок на тарифікацію.

Дана функція перевірки є частиною системи моніторингу, запускається із заданою періодичністю для забезпечення надійності та безперебійної роботи системи, забезпечує контроль за достатністю ресурсів системи.

4.4.8. Оцінка ефективності методу контролю достатності ресурсів системи для обробки заявок на тарифікацію

Спектр проблем, які можуть вирішуватися за допомогою запропонованого методу та алгоритмів досить широкий. Основною ознакою систем, для яких може бути застосований запропонований метод, є виконання великої кількості процедур, ініціатором яких є люди або інші програми. Виконання процедур здійснюється за допомогою програмного забезпеченні серверу, при цьому використовуються технічні ресурси системи.

На сьогоднішній день проблема перевантаження вирішується за рахунок не допуску в систему надлишкової кількості запитів на тарифікацію, тобто якщо сервер є перевантаженим, то поступає сигнал на керуючий пристрій і тимчасово заявки на тарифікацію не приймаються. При цьому одночасно задіяні (знаходяться у режимі очікування) всі ресурси серверу, завантаження обладнання на 20-30% є нормальним в процесі обслуговування.

Був проведений експеримент роботи імітаційної моделі в двох режимах:

1. На обслуговування виділявся обмежений технічний ресурс, у разі перевантаження заявки відкидалися (Режим без застосування запропонованих методів).

2. Обслуговування заявок може проводитися від одного до трьох аналогічних серверів з обмеженими технічними ресурсами, на основі статистичних вибірки від оператора зв'язку був сформований розклад включення серверів відповідно до алгоритму пошуку моментів переключення, далі було згенеровано вхідний потік максимально наближений до реального. Включення серверів відбувалося за розкладом, а за методом включення додаткового технічного обладнання кожну три хвилини відбувалась перевірка достатності ресурсів для обслуговування вхідного потоку.

За результатами роботи імітаційної моделі було проведено дослідження використання електричної енергії, яке показало скорочення споживання енергії на 60% (рис. 4.20 а), В той самий час кількість втрачених заявок скоротилася в 5 разів від п'яти відсотків до одного (рис. 4.20 б).

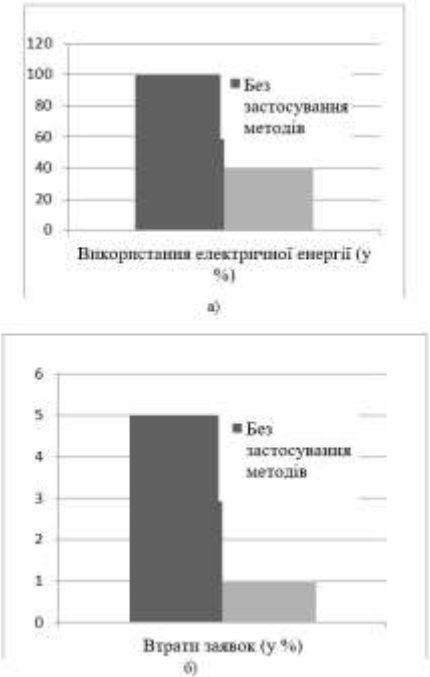


Рис. 4.20 Порівняння використання електричної енергії та кількості втрачених заявок а)Порівняння використання електричної енергії; б) Порівняння кількості втрачених заявок на тарифікацію

4.4.9. Метод складання розкладу включення серверів

Для забезпечення безперебійної роботи серверу тарифікації необхідно розробити комплекс засобів, які враховують не тільки поточну ситуацію, але й статистику набрану за тривалий період часу. Це дозволить планувати ресурси, розробляти методи балансування навантаження, тощо.

Побудова таблиці розкладу включення серверів.

Для рішення даної задачі необхідно провести аналіз статистики протягом тривалого періоду. Наприклад, аналіз статистики за понеділок протягом 15-20 тижнів. Результатом таких статистичних даних стане матриця де, для кожного малого інтервалу часу, наприклад, для часу 8:00:00-8:00:01, такий інтервал позначатимемо значенням закінчення інтервалу: 8:00:01. В матриці будуть зібрані дані про кількість заявок (навантаженні), яке потребувало обслуговування на всіх доступних серверах для яких складається даний розклад. Розглядатимемо систему з єдиним входом, задача розподілу вхідного навантаження між доступними серверами не розглядається.

Табл. 4.5

Таблиця розкладу включення серверів

n\t	00:00:00	00:00:01	00:00:02	00:00:03	00:00:04	…	…	23:59:58	23:59:59
1	x(n,t)								
2									
N									

де t – час вимірювання, n=1,..N – номер тижня для якого збиралася статистика, проводилося вимірювання навантаження, x(n,t) – кількість заявок на тарифікацію, або навантаження на сервер, що було заміряно на n-му тижні в момент часу t.

Необхідно провести розбиття на відрізки, які характеризуються однаковою динамікою зміни навантаження. Оцінка динаміки зміни навантаження буде проводитися на основі аналізу прямої, яка апроксимує статистичні дані.

Алгоритм розбиття на відрізки.

Вхідні дані: Статистика навантаження x(n,t), яка була описана в табл. 4.4.

Крок 1. Для кожного значення t знаходимо середнє значення $\bar{x}(t)=\sum_n x(n,t)/N$ дані зводяться в таблицю, як показано в табл. 4.6.

Табл.4.6

Дані

T	00:00:00	00:00:01	00:00:02	00:00:03	…
$\bar{x}$	$\bar{x}$(t)	$\bar{x}$(t)	$\bar{x}$(t)		

Крок 2. Задати час завантаження серверу T, задати мале число ε_1

Крок 3. Необхідно розділити множину допустимих значень t на підмножини t_i, таким чином, щоб $t_{i+1}-t_i=T$. Окремо розглядається окремі частини матриці з табл. 4.3, (j+1)-я матриця буде мати вигляд такий як показано в табл. 4.7:

Табл. 4.7

Дані

T	t_i +00:00:01	t_i +00:00:02	…	t_{i+1}- 00:00:01	t_{i+1}
$\bar{x}$	$\bar{x}$(t)	$\bar{x}$(t)	$\bar{x}$(t)		

Кількість отриманих частин матриць I=24години/T.

Крок 4. Методом найменших квадратів знаходимо для кожної i-й матриці, для пар значень ($\bar{x}$,t) знаходимо оцінку коефіцієнту апроксимуючої прямої $\widehat{a_i}$.

Крок 5. Для всіх i=1,..I проводиться аналіз $\widehat{a_i}$. Если $|\widehat{a_i}-\widehat{a_{i+1}}|<\varepsilon_1$, тоді множини *i* и *(i+1)* об'єднуються. Отримані множини отримують нову нумерацію, кількість нових множин позначається, як I_{new}. Перехід на Крок 4. Інакше, якщо для всіх i=1,..I_{new}:$|\widehat{a_i}-\widehat{a_{i+1}}|>\varepsilon_1$, тоді розбиття знайдене.

Результатом роботи даного алгоритму є розбиття на ділянки з однаковою динамікою зростання навантаження. Для отриманих значень t_i (i=1,..I_{new}, I_{new} – кількість інтервалів, якщо більше немає можливості поєднувати підмножини) відбувається порівняння середньостатистичного навантаження $\bar{x}(t_i)$ з урахуванням можливого середньоквадратичного відхилення.

Розрахунок середньоквадратичного відхилення.

Вхідні дані:

1.Час t_i, для якого виконується оцінка дисперсії.

2.Значення із стовпця матриці $x(n,t_i)$ (див. табл. 4.4).

Розрахунок:

$Mx(t_i)= \bar{x}(t_i)=\frac{x_1(ti)+x_2(ti)+\cdots x_N(ti)}{N}$

$Mx^2(t_i)= \frac{x_1{}^2(ti)+x_2{}^2(ti)+\cdots x_N{}^2(ti)}{N}$

$D(t_i)= Mx^2(t_i)- Mx(t_i)$.

Другим етапом побудови розкладу включення серверів є визначення кількості серверів, яка необхідна для обслуговування заявок. Результатом роботи даного етапу є інформація про час, коли необхідно змінити кількість обслуговуючих пристроїв та відповідна кількість серверів, яка має бути включена в той чи інший момент часу. Ілюстрація динаміки зміни кількості заявок протягом доби наведена на рис. 4.21

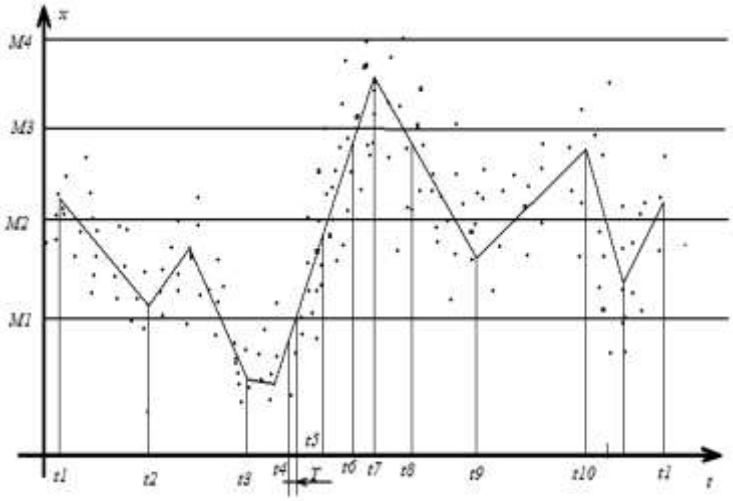


Рис.4.21 Динаміки зміни кількості заявок протягом доби

Алгоритм пошуку моментів переключення:

Вхідні дані

1. $\bar{x}(t_i)$ – середньостатичне навантаження в момент часу t_i (масив $\{\bar{x}(t_i)\}$ кількість елементів: i=1,.. I_{new}, де I_{new} –кількість моментів, в які змінюється динаміка вхідного навантаження (переломні моменти).

2. $D(t_i)$ – середньоквадратичне відхилення (масив $\{D(t_i)\}$ кількість елементів – I_{new}).

3. Множина M={M_1, M_2…M_K} – значення допустимого навантаження, яку може обслужити 1, 2 .. K серверів відповідно.

4. Масив значень $\widehat{a_i}$, які відповідають куту нахилу прямої апроксимації статистичних даних до моменту можливого переключення t_i.

5. T – час завантаження серверу.

Крок 1. Для кожного i=1,.. I_{new} знайти x'(t_i) = $\bar{x}(t_i)$+ D(t_i).

Крок 2. Створити матрицю R={r_{qp}}$_{q=1,..3,p=1,.(2Inew)…}$ Змінній p присвоїти значення 1 (p=1).

Крок 3. Для кожного i=1,.. I_{new} знаходимо пару (M_k, M_{k+1}), так щоб виконувалось x'(t_i) є [M_k, M_{k+1}].

Крок 4. Для всіх i=1,.. I_{new} перевірити умову: якщо x'(t_i) є [M_k, M_{k+1}] ти x'(t_{i+1}) є [M_k, M_{k+1}], тоді i=i+1 повторити крок 3, інакше в момент часу t_i потрібно буде переключення (перехід на крок 4).

Крок 5. Якщо x'(t_i) є [M_k, M_{k+1}] і x'(t_{i+1})> M_{k+1}, і x'(t_{i+1})< M_{k+s}, тоді в момент часу t_i необхідно включення серверу. Якщо s>2 перехід на крок 6, інакше додати в матрицю R елементи: r_{1p}= t_i , r_{2p}= k+2, r_{3p} =1, p=p+1, Якщо i= I_{new} перейти на крок 8, інакше i= i+1 перейти на крок 5.

Якщо x'(t_i) є [M_k, M_{k+1}] і x'(t_{i+1})< M_k, і x'(t_{i+1})> $M_{k-s'}$, тоді в момент часу t_i необхідно включення додаткового серверу. Якщо s>0, тоді перехід на крок 7, інакше добавити в матрицю R елементи r_{1p}= t_i , r_{2p}= k+1, r_{3p} =0, p=p+1, Якщо i= I_{new} перейти на Крок 8, інакше i= i+1 перейти на крок 5.

Крок 6. Для кожного M_{k+1}… M_{k+s-1} знайти моменти включення відповідних серверів. Доки g=1, g<s

t_g=(M_{k+g}– x'(t_i))/$\widehat{a_{i-1}}$

додати в матрицю R елементи r_{1p}= t_i+t_g-T , r_{2p}= k+g, r_{3p} =1, p=p+1, g=g+1. Якщо i= I_{new} перейти на крок 8, інакше i= i+1 перейти на крок 5.

Крок 7. Для кожного M_{k-1}… $M_{k-s'+1}$ знайти моменти відключення відповідних серверів. Поки g=1, g<s'+1

t_g=(x'(t_i) – M_{k-g})/$\widehat{a_{i-1}}$

додати в матрицю R елементи r_{1p}= t_i+t_g-T , r_{2p}= k-g, r_{3p} =0, p=p+1, g=g+1. Якщо i= I_{new} перейти на крок 8, інакше i= i+1 перейти на крок 5.

Крок 8. Завершення роботи програми, виведення матриці R, відображує розклад включення серверів.

Результатом роботи даного алгоритму є матриця яка відображає час переключення, кількість серверів які мають перший рядок відображає моменти часу в які необхідно проводити операцію включення або виключення серверу, другий рядок означає кількість активних серверів які повинні бути будуть обслуговувати вхідне навантаження у наступний період часу, третій рядок означає процес який має бути ініційований у відповідний момент часу: 1 – включення одного з серверів, 0 – виключення одного з серверів.

Слід зауважити, що робота даного алгоритму є лише складовою частиною процесу технічного забезпечення системи онлайн тарифікації. Оскільки крім статичного розкладу включення серверів потребує розгляду питання балансування навантаження між серверами.

4.4.10. Оцінка ефективності статистичного методу розподілу кількості хмарних ресурсів.

Спектр проблем, які можуть вирішуватися за допомогою запропонованого методу та алгоритмів досить широкий. Основною ознакою систем, для яких може бути застосований запропонований метод, є виконання великої кількості процедур, ініціатором яких є люди або інші програми. Виконання процедур здійснюється за допомогою програмного забезпеченні серверу, при цьому використовуються технічні ресурси системи.

На сьогоднішній день проблема перевантаження вирішується за рахунок не допуску в систему надлишкової кількості запитів на тарифікацію, тобто якщо сервер є перевантаженим, то поступає сигнал на керуючий пристрій і тимчасово заявки на тарифікацію не приймаються. При цьому одночасно задіяні (знаходяться у режимі очікування) всі ресурси серверу, завантаження обладнання на 20-30% є нормальним в процесі обслуговування.

Був проведений експеримент роботи імітаційної моделі в двох режимах:

1. На обслуговування виділявся обмежений технічний ресурс, у разі перевантаження заявки відкидалися (Режим без застосування запропонованих методів).

2. Обслуговування заявок може проводитися від одного до трьох аналогічних серверів з обмеженими технічними ресурсами, на основі статистичних вибірки від оператора зв'язку був сформований розклад включення серверів відповідно до алгоритму пошуку моментів переключення, далі було згенеровано вхідний потік максимально наближений до реального. Включення серверів відбувалося за розкладом, а за методом включення додаткового технічного обладнання кожну три хвилини відбувалась перевірка достатності ресурсів для обслуговування вхідного потоку.

За результатами роботи імітаційної моделі було проведено дослідження використання електричної енергії, яке показало скорочення споживання енергії на 60% (рис. 4.21).

Рис. 4.21 Порівняння використання електричної енергії

Висновки

1. Запропоновано модель обслуговування заявок на тарифікацію абонентів, яка полягає в тому щоб в моменти перевантаження не обслуговувати post-paid абонентів в режимі реального часу, доведено ефективність запропонованої моделі, а саме розраховано оцінку прибутку від впровадження моделі, досліджено витрати на впровадження моделі. Запропоновано підхід до класифікації post-paid абонентів за рівнем ризику, розраховані середні збитки від обслуговування відповідно до запропонованої моделі обслуговування.

2. Запропоновано метод оптимізації ємності буфера очікування заявок на тарифікацію, в якому наведені умови економічної доцільності нарощування буферу білінгової системи, залежно від навантаження на систему та прибутку від збільшення потужностей обладнання оператора зв'язку.

3. В даному розділі досліджено роботу системи онлайн тарифікації OCS, та особливості обслуговування різних типів сервісів (SCUR, ECUR, IEC). Запропоновано метод розподілу ресурсів системи онлайн тарифікації для забезпечення ефективної обробки заявок, який дозволяє здійснювати контроль за якістю надання послуг, враховує необхідну кількість ресурсів для обслуговування однієї заявки, що дозволяє виділяти ресурси пропорційно вимогам сервісу, враховує статистичні дані про кількості заявок різних типів сервісів, які надходять у заданих інтервалах часу, що дозволяє налаштовувати розподіл технічних ресурсів, що обслуговують заявки, відповідно до складу вхідного навантаження, а також забезпечити максимізацію економічної ефективності процесу надання послуг. за критерій доцільності вибрана економічна ефективність надання послуг.

4. У розділі досліджено тенденції залучення хмарних ресурсів для забезпечення роботи серверу тарифікації. Запропоновано метод визначення моменту включення додаткового технічного засобу, який дозволяє проводити оцінку динаміки вхідного навантаження та поточного стану технічних засобів, та завчасно увімкнути додатковий ресурс, та запобігли перевантаженню існуючих ресурсів. Запропоновано метод складання розкладу включення серверів, який на основі довгострокових статистичних даних, дозволяє скласти розклад залучення додаткових технічних ресурсів (включення серверів) отже спланувати роботу технічних засобів у періоди часу з різним очікуваним навантаженням.

5. В даному розділі описано двохрівневу модель системи керування на сервері онлайн тарифікації. Запропоновано метод контролю перевантажень в системі онлайн 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, що при використанні запропонованого методу кількість втрачених заявок на тарифікацію, через перевищення часу обслуговування зменшується до 3-х разів. З метою формування дисципліни подачі заявок на тарифікацію із буфера на обслуговування, запропоновано метод керування вхідним потоком, який полягає формуванні послідовності кількостей заявок на тарифікацію, які подаються на обслуговування, що забезпечує згладжування використання фізичних ресурсів системи тарифікації.

5. ВИДІЛЕННЯ РЕСУРСІВ ДЛЯ ВІРТУАЛІЗОВАНИХ МЕРЕЖЕВИХ ФУНКЦІЙ В ГІБРИДНОМУ СЕРЕДОВИЩІ

5.1. Вступ

Сьогодні мобільні абоненти бажають залишатися на зв'язку в будь-якому місці, в будь-який час, і використовуючи будь-який пристрій. Це явище спонукає операторів мобільного зв'язку до побудови складних мережевих архітектур з включенням нових можливостей і розширень, якими важче управляти і працювати [92]. Два поняття знаходяться в центрі досліджень і розробок в даний момент, а саме Віртуалізація Мережевих Функцій (Network Functions Virtualization – NFV) і Програмно Конфігуровані Мережі (Software Defined Networking – SDN) [93].

Існуючі мобільні мережеві інфраструктури складаються з виділених мережевих вузлів, які зазвичай розміщуються в різних точках мережі, і кожному вузлу призначається надання певного набору функцій і сервісів. Така архітектура є негнучкою з точки зору впровадження нових сервісів, здійснює неоптимальну маршрутизацію трафіку і неефективне використання ресурсів мережі. Крім того, масштабованість і економічність такої архітектури стає проблемою у світлі останніх прогнозів трафіку і очікувань користувачів [94].

З поширенням потужних мобільних пристроїв (наприклад, смартфонів, планшетних ПК і ноутбуків), а також зі зростаючою популярністю мобільних мультимедійних прикладних програм, очікуються підвищені вимоги до пропускної здатності. Згідно з [95] очікується, що загальний мобільний трафік даних зросте до 24.3 екзабайт на місяць до 2019 року, майже в десять разів більше в порівнянні з 2014. Мобільний трафік даних буде рости з середнім темпом річного зростання (CAGR) у 57 відсотків з 2014 до 2019 (рис. 5.1).

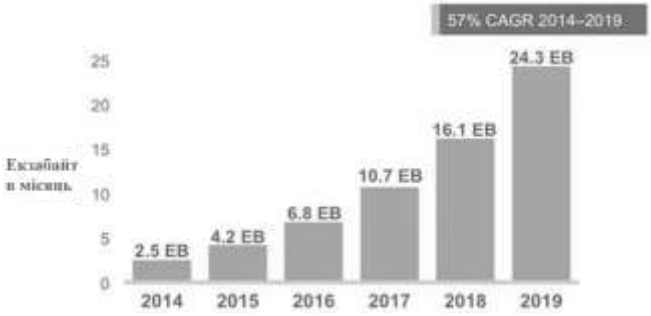


Рис. 5.1 Прогноз Cisco у 24.3 екзабайти на місяць трафіку мобільних даних до 2019 [95]

Надання послуг в телекомунікаційній галузі традиційно ґрунтується на тому, що мережеві оператори впроваджують фізичні пропрієтарні пристрої та обладнання для кожної функції, яка є частиною певного сервісу. Крім того, сервісні компоненти мають чіткі ланцюги і/або порядок, які повинні бути відображені в топології мережі і в локалізації сервісних елементів. Це, в поєднанні з вимогами до високої якості, стабільності і строгим дотриманням протоколу, привело до тривалих циклів

продукту, дуже низької гнучкості обслуговування і сильної залежності від спеціалізованих апаратних засобів.

Проте, вимоги користувачів до більш різноманітних і нових (короткоживучих) послуг з високими швидкостями передачі даних продовжують збільшуватися. Таким чином, телекомунікаційні сервіс-провайдери (Telecommunication Service Provider – TSP), повинні відповідним чином і постійно набувати, зберігати і експлуатувати нове фізичне обладнання. Це не тільки вимагає високих і швидко змінюваних навичок для техніків що експлуатують та управляють цим обладнанням, а й вимагає щільного розміщення мережевого обладнання, такого як базові станції. Все це призводить до високих капітальних і експлуатаційних витрат для TSP.

Більш того, навіть з цими високими вимогами абонентів, зростання капітальних і експлуатаційних витрат не може бути переведене в більш високу абонентську плату, так як TSP відомо, що через високу конкуренцію, як між собою, так і від over-the-top послуг на їх каналах передачі даних, підвищення цін призводить лише до відтоку клієнтів. Тому TSP були змушені шукати шляхи побудови більш динамічних і сервіс-обізнаних мереж з метою скорочення життєвих циклів продукції, операційних і капітальних витрат і підвищення оперативності обслуговування [114].

Значна залежність мереж від апаратного забезпечення та існування різних спеціалізованих апаратних пристроїв, таких як брандмауери, обладнання глибокої інспекції пакетів (DPI), і маршрутизаторів в мережевій інфраструктурі, посилили проблеми, що стоять перед провайдерами послуг мережі [96].

Як правило, оператори мобільного зв'язку справляються з підвищеними навантаженнями трафіку шляхом розширення/покращення загальної пропускної здатності мережі відповідним чином. Проте, це все більш і більш важко реалізувати за рахунок збільшення капітальних/операційних витрат (CAPEX/OPEX) в світлі низької рентабельності інвестицій (ROI). Окрім низького ROI, надмірне резервування ресурсів більше не вважається життєздатною стратегією, щоб задовольнити збільшення трафіку, так як відповідно до [97] до 80% обчислювальної потужності базових станцій і до половини потужності ядра мережі є невикористаними. Це призводить до низького використання мережевих ресурсів, а також до високого рівня споживання енергії, що знижують економічну ефективність мережі для операторів мобільного зв'язку [94].

Принцип NFV спрямований на перетворення мережевих архітектур шляхом впровадження мережевих функцій в програмному забезпеченні, що може працювати на стандартній апаратній платформі. Крім того, він спрямований на перетворення традиційних мережевих операцій, оскільки програмне забезпечення може бути легко переміщене, або створено сутність в різних місцях без необхідності використовувати нове обладнання. NFV має багато переваг, від поліпшення операційної ефективності і зниження енергоспоживання до коротших інтервалів розгортання/оновлення і майже оптимального використання мережевих ресурсів, оскільки будівельні блоки можуть виділятись і перерозподілятись під час виконання в залежності від вимог [98, 99].

NFV прокладає шлях до ряду відмінностей в способах реалізації надання мережевого сервісу в порівнянні з існуючою практикою. Таким чином, ці відмінності можна охарактеризувати наступним чином [114]:

• Відв'язка програмного забезпечення від апаратного. Оскільки елемент мережі більше не є об'єднанням інтегрованих апаратних і програмних сутностей, еволюція обох є незалежною один від одного. Це дозволяє мати окремі терміни розробки і технічного обслуговування програмного і апаратного забезпечення.

• Гнучке розгортання мережевих функцій. Відрив програмного забезпечення від апаратного допомагає перерозподілити і спільно використовувати ресурси інфраструктури, таким чином, разом, апаратне і програмне забезпечення, може виконувати різні функції в різний час. Це допомагає мережевим операторам розгортати нові мережеві сервіси швидше по тій же фізичній платформі. Таким чином, компоненти можуть бути створені в будь-якому NFV-сумісному пристрої в мережі і їх з'єднання можуть бути встановлені на гнучкій основі.

• Динамічне масштабування. Розділення функціональності мережевої функції на створювані програмні компоненти забезпечує більшу гнучкість масштабування реальної продуктивності VNF більш динамічно і з більшою деталізацією, наприклад, відповідно до фактичного трафіку, для якого оператор мережі повинен надавати ємність.

На практиці, більш дорогі спеціалізовані апаратні засоби часто працюють швидше і ефективніше ніж віртуалізовані сутності, навіть хоча останні є більш гнучкими. Оскільки спеціалізовані апаратні засоби на даний час широко використовуються, цілком імовірно, що гібридні сценарії розгортання стануть поширеними, коли частина сервісів надається фізичним обладнанням (рис. 5.2). У NFV мережах, набір ланцюгів сервісів повинен розташовуватись на фізичних вузлах мережі. Ланцюг сервісів – це набір з одного або декількох сервісів або віртуальних машин, які з'єднані разом для забезпечення певної функціональності, і можуть бути представлені у вигляді графа, що містить сервіси і мережеві вимоги між цими сервісами. У гібридному мережевому середовищі, ланцюги сервісів можуть розміщуватись або з використанням фізичного обладнання, або з використанням віртуалізованих сутностей. Успіх цього підходу залежить від наявності та продуктивності алгоритмів, що визначають де і як ці структурні блоки створюються [99].

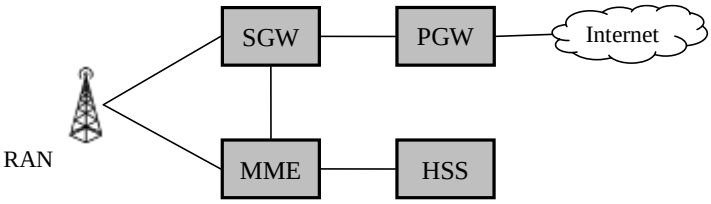
5.1.1. Високорівнева платформа NFV

NFV передбачає реалізацію мережевих функцій як сутностей програмного забезпечення, які працюють на Інфраструктурі NFV (NFV Infrastructure – NFVI). На рис. 5.3 показано платформу NFV високого рівня. Таким чином, три основні робочі області визначені в NFV [115]:

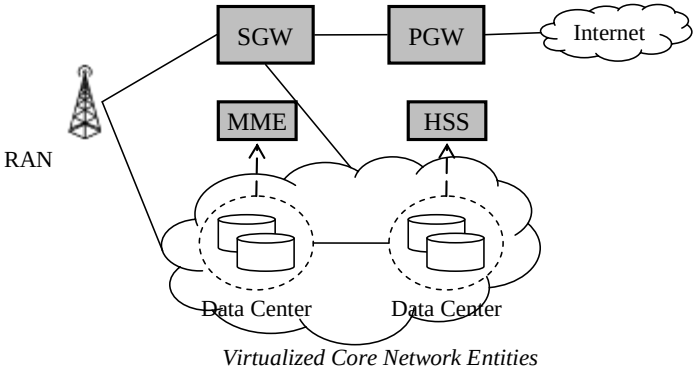
• Віртуалізована Мережева Функція (Virtual Network Function – VNF), як програмна реалізація функції мережі, яка здатна працювати на NFVI;

• Інфраструктура NFV (NFVI), включаючи різноманітність фізичних ресурсів і як вони можуть бути віртуалізовані. NFVI підтримує виконання VNF;

• Управління і Оркестровка NFV, що охоплює оркестровку і управління життєвим циклом фізичних і/або програмних ресурсів, що підтримують віртуалізацію інфраструктури, та управління життєвим циклом VNF. Управління і Оркестровка NFV фокусується на всіх специфічних для віртуалізації завданнях управління необхідних в платформі NFV.



(а) Типова мережа LTE/EPC



(б) Часткова віртуалізація мобільного ядра

Рис. 5.2 Мобільне ядро мережі

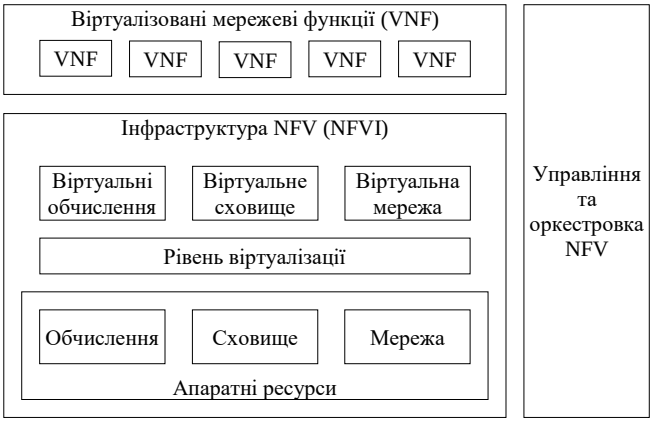


Рис. 5.3 Високорівнева платформа NFV [115]

5.1.2. NFV платформа

Основними компонентами віртуалізованих платформ, де розгорнуто NFV є:

• фізичний сервер: фізичний сервер це машина, яка має всі фізичні ресурси, такі як CPU, сховище і оперативну пам'ять;

• гіпервізор: гіпервізор або система контролю віртуальної машини, це програмне забезпечення, яке управляє фізичними ресурсами. Він забезпечує віртуальне середовище, в якому працюють гостьові віртуальні машини;

• гостьова віртуальна машина: частина програмного забезпечення, яка емулює архітектуру і функціональні можливості фізичної платформи, на якій виконується бажана прикладна програма.

Віртуальні машини розгортаються на серверах з великими об'ємами, які можуть бути розташовані в датацентрах, в вузлах мережі і в приміщеннях кінцевих користувачів. Крім того, більшість віртуальних машин забезпечують обчислювальні ресурси за вимогою, використовуючи хмару. Послуги хмарних обчислень пропонуються в різних форматах: інфраструктура як сервіс (IaaS), що також називають апаратне забезпечення як сервіс (HaaS), платформа як сервіс (PaaS), програмне забезпечення як сервіс (SaaS) і мережа як сервіс (NaaS). Немає домовленості по стандартному визначенню NaaS. Однак часто вважається що його надають під IaaS. Технологія NFV використовує переваги інфраструктури та мережевих сервісів (IaaS і NaaS) для формування інфраструктури віртуалізації мережевих функцій (NFVI).

Для досягнення поставлених NFV цілей, таких як гнучкість в призначенні віртуальних мережевих функцій (VNF) апаратному забезпеченню, швидкі інновації в послугах, підвищення ефективності роботи, знижене споживання енергії та відкриті стандартні інтерфейси між VNF, кожна VNF повинна працювати на платформі, що включає динамічне ініціювання та оркестровку екземплярів VNF. Крім того, також необхідно управляти середовищем хостингу NFVI на технологіях віртуалізації для задоволення всіх вимог VNF до даних, розподілу ресурсів, залежностей, доступності та інших атрибутів [96].

5.1.3. Архітектура NFV

A. Інфраструктура NFV (NFVI)

NFVI є поєднанням апаратних і програмних ресурсів, які складають середовище, в якому розгортаються VNF. Фізичні ресурси включають в себе готову комерційну (COTS) обчислювальну техніку, сховище і мережу (що складаються з вузлів і каналів), які забезпечують обробку, зберігання і можливість підключення для VNF. Віртуальні ресурси є абстракціями обчислювальних ресурсів, ресурсів зберігання і мережевих ресурсів. Абстракція досягається з використанням рівня віртуалізації (на основі гіпервізора), який відділяє віртуальні ресурси від нижчерозташованих фізичних ресурсів. У середовищі датацентру обчислювальні ресурси і ресурси зберігання можуть бути представлені у вигляді однієї або декількох віртуальних машин, в той час як віртуальні мережі складаються з віртуальних каналів і вузлів. Віртуальний вузол являє собою програмний компонент з функціональністю хостингу або маршрутизації, наприклад, операційну систему інкапсульовану в віртуальній машині. Віртуальний канал являє собою логічне з'єднання двох віртуальних вузлів, що виглядають для них як пряма фізична лінія з динамічно змінюваними властивостями.

В. Функції та сервіси віртуальної мережі

Мережева функція є функціональним блоком в межах мережевої інфраструктури, яка має чітко визначені зовнішні інтерфейси і чітко визначену функціональну поведінку. Прикладами мережевих функцій є елементи в домашній мережі, наприклад абонентський шлюз (Residential Gateway – RGW); і традиційні мережеві функції, наприклад, DHCP-сервери, брандмауери і т.д. Таким чином, VNF є реалізацією мережевої функції, яка розгорнута на віртуальних ресурсах, таких як віртуальна машина. Один VNF може складатися з декількох внутрішніх компонентів і, отже, він може бути розгорнутий на декількох віртуальних машинах, і в цьому випадку кожна віртуальна машина містить один з компонентів VNF. Сервіс є пропозицією від TSP, що складається з однієї або декількох мережевих функцій. У випадку NFV, мережеві функції, які утворюють сервіс, віртуалізуються і розгортаються на віртуальних ресурсах, таких як віртуальна машина. Проте, з точки зору користувачів, сервіси, що працюють на функціях спеціалізованого обладнання або на віртуальних машин, повинні мати однакову продуктивність. Кількість, тип і порядок VNF, що створюють його, визначаються функціональною і поведінковою специфікацією сервісу. Таким чином, поведінка сервісу залежить від складових VNF [114].

Мережева служба архітектурно може розглядатися як граф мережевих функцій, з'єднаних між собою за допомогою відповідної мережевої інфраструктури. Ці мережеві функції можуть бути реалізовані в мережі одного оператора або у взаємодії між різними операторами мереж. Поведінка нижчерозташованої мережевої функції вносить вклад у поведінку сервісу вищого рівня. Таким чином, поведінка мережевого сервісу являє собою поєднання поведінки її складових функціональних блоків, які можуть включати в себе окремі мережеві функції, набори мережевих функцій, графи мережевих функцій і/або мережу інфраструктури.

Кінцеві точки і мережеві функції мережевого сервісу представлені у вигляді вузлів і відповідають пристроям, прикладним програмам і/або прикладним програмам фізичних серверів. Граф мережевих функцій може містити вузли мережевих функцій, з'єднаних логічними каналами, які можуть бути однонаправленими, двонаправленими, багатонаправленими та/або широкомовними. Простим прикладом графа є ланцюг мережевих функцій. Приклад такого мережевого сервісу з кінця в кінець може включати в себе смартфон, бездротову мережу, брандмауер, балансувальник навантаження і набір серверів CDN. Область діяльності NFV знаходиться всередині ресурсів, якими володіє оператор. Таким чином, пристрій користувача, наприклад, мобільний телефон знаходиться поза зоною дії, так як оператор не може управляти ним. Проте, віртуалізація і мережевий хостинг функцій абонента можливі і знаходяться в рамках NFV (наприклад, див. випадки використання Віртуальної Платформи як Сервіс (Virtual Network Platform as a Service – VNPaaS) і віртуалізацію домашнього середовища в GS NFV 001 [117]).

Рис. 5.4 ілюструє представлення мережевого сервісу з кінця в кінець, що включає в себе другий вкладений граф мережевих функцій, як показано блоками вузлів мережевих функцій в середині рисунку, з'єднаних логічними каналами. Кінцеві точки підключаються до мережевих функцій через мережеву інфраструктуру (дротову або бездротову), в результаті чого ми бачимо логічний

інтерфейс між кінцевою точкою і функцією мережі. Ці логічні інтерфейси представлені на рисунку пунктиром. На рис. 5.4, зовнішній мережевий сервіс з кінця в кінець складається з кінцевої точки А, внутрішнього графа мережевих функцій і кінцевої точки В, в той час як внутрішній граф мережевих функцій складається з мережевих функцій NF1, NF2 і NF3. Вони з'єднані між собою за допомогою логічних каналів, що надаються мережею інфраструктури 2 [115].

C. NFV Управління та Оркестровка (NFV Management and Orchestration – NFV MANO)

Згідно з платформою MANO від ETSI, NFV MANO забезпечує функціональні можливості, необхідні для управління VNF, і пов'язаних з цим операцій, такі як конфігурація VNF та інфраструктури, на яких ці функції працюють. Вона включає в себе оркестровку і управління життєвим циклом фізичних і/або програмних ресурсів, що підтримують віртуалізацію інфраструктури, а також управління життєвим циклом VNF. Вона також включає в себе бази даних, які використовуються для зберігання моделей інформації і даних, які визначають як властивості розгортання, так і властивості життєвого циклу функцій, сервісів і ресурсів. NFV MANO фокусується на всіх завданнях управління віртуалізації необхідних платформі NFV. Крім того, платформа визначає інтерфейси, які можуть використовуватися для зв'язку між різними компонентами NFV MANO, а також координації з традиційними системами управління мережею, такими як системи підтримки операцій (Operations Support System – OSS) і системи підтримки бізнесу (Business Support System – BSS), для того, щоб забезпечити управління як VNF, так і функціями, які працюють на традиційному обладнанні [114].

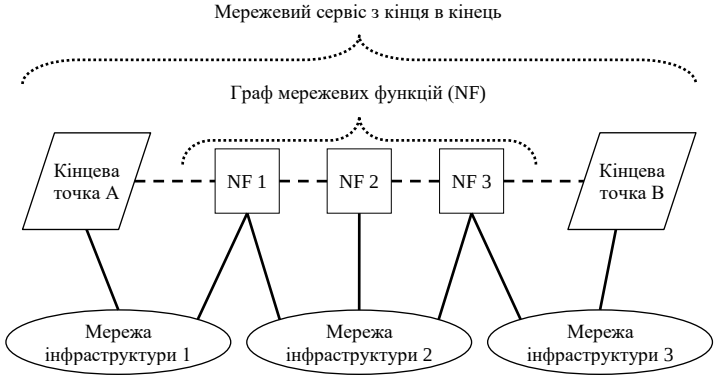


Рис. 5.4 Представлення мережевого сервісу з кінця в кінець у вигляді графа [115]

5.1.4. Задачі та проблеми віртуалізації мережевих функцій

NFV є важливим нововведенням і перспективним підходом для операторів і провайдерів послуг. Тим не менш, він також стикається з низкою проблем. Далі

розглянуто відповідні завдання, відкриті проблеми і пов'язані з ними рішення, які зведено в табл. 5.1 [116].

Табл. 5.1
Задачі NFV

Задачі	**Опис**	**Рішення**
Віртуалізація функцій	Віртуалізовані функції повинні задовільняти певні вимоги для підтримки обробки пакетів з лінійною швидкістю: (1) Висока продуктивність (висока швидкість вводу/виводу, швидка обробка пакетів, малі затримки передачі і т.д.) (2) Підтримка мульти-тенантності (3) Незалежність від операційної системи	Важливі пов’язані з цим роботи: (1) DPDK, набір бібліотек та драйверів для швидкої обробки пакетів. (2) NetVM, система для роботи функціональності мережі та проміжних пристроїв (middlebox) з лінійною швидкістю на загальному апартному забезпеченні. (3) ClickOS, мала, швидко завантажувана, з низькою затримкою, віртуалізована платформа програмних проміжних пристроїв.
Переносимість	Очікується, що платформа NFV буде завантажувати, виконувати та переміщати VNF по різним але стандартних серверам в багатовендорних середовищах. Ця властивість відома як переносимість.	Розгортання мережевих функцій за допомогою віртуального програмного середовища покращує переносимість. Цей підхід гарантує, що VNF незалежні від операційної системи, а також гарантується ізоляція ресурсів.
Стандартні інтерфейси	Стандартизовані API повинні бути розроблені для забезпечення можливості NFV з‘єднуватися з абонентами за допомогою	І VNF, і обчислювальні ресурси описуються за допомогою стандартних шаблонів.

	нижчерозташованої інфраструктури та центрально контролюватися та управлятися.	Нормалізовані північний і південний інтерфейси повинні бути розроблені муж цими рівнями.
Розгортання функцій	Деталізоване розгортання, управління та контроль мережевих функцій необхідні в контексті мережевих вузлів з NFV для різних оптимізаційних цілей.	Повинна бути розгорнута система моніторингу, яка збирає і повідомляє дані про поведінку ресурсів, а також уніфікована система контролю та оптимізації з різними двигунами оптимізації.
Направлення трафіку	У такій, що визначається програмно, архітектурі NFV, направлення трафіку повинно бути оптимізоване разом з розгортанням функцій, роблячи при цьому задачу оптимізації складною для вирішення.	Для досягнення онлайн обчислень направлення трафіку, евристичні алгоритми повинні бути розроблені для зменшення обчислювальної складності.

5.1.5. Вбудовування віртуальної мережі

У віртуалізації мережі основним об'єктом є віртуальна мережа. Віртуальна мережа є поєднанням активних і пасивних елементів мережі (вузлів мережі і мережевих каналів) на фізичній мережі. Віртуальні вузли з'єднані між собою через віртуальні з'єднання, формуючи віртуальну топологію. Віртуалізуючи ресурси і вузли, і канали фізичної мережі, багато віртуальних мережевих топологій з широко змінюваними характеристиками можуть бути створені і розташовані на одному фізичному обладнанні. Крім того, абстракція, введена механізмами віртуалізації ресурсів, дозволяє мережевим операторам керувати і змінювати мережі дуже гнучко і динамічно.

Майбутні Інтернет-архітектури будуть базуватися на бізнес-моделі Інфраструктури як Сервіс (IaaS), яка роз'єднує роль поточних провайдерів Інтернет-послуг (ISP) на дві нові ролі: Провайдера Інфраструктури (Infrastructure Provider – InP), який розгортає і підтримує мережеве обладнання та Провайдера Послуг (Service Provider – SP), що відповідає за розгортання мережевих протоколів і пропонує сервіси з кінця в кінець. Впровадження віртуалізації мережі розділяє

управління і бізнес ролі SP шляхом визначення трьох основних гравців: Провайдера віртуальної мережі (Virtual Network Provider – VNP), який збирає віртуальні ресурси одного або декількох InP, Оператора віртуальної мережі (Virtual Network Operator – VNO), який встановлює і управляє віртуальною мережею відповідно до потреб SP, і СП, який є вільним від управління і концентрується на бізнесі, використовуючи віртуальні мережі, щоб запропонувати індивідуальні послуги.

Проблема вбудовування віртуальних мереж у фізичну мережу є основною проблемою розподілу ресурсів в області віртуалізації мережі і зазвичай згадується як проблема вбудовування віртуальної мережі (Virtual Network Embedding – VNE). За допомогою динамічного відображення віртуальних ресурсів на фізичне обладнанні вигода, отримана від існуючого обладнання може бути максимізована. Оптимальний динамічний розподіл ресурсів, що призводить до самостійної конфігурації і організації майбутніх мереж, буде необхідний для надання індивідуальних гарантованих послуг з кінця в кінець кінцевим користувачам. Це оптимальність може бути обчислена стосовно різних цілей, починаючи від якості обслуговування, економічної вигоди або живучості до енергоефективності та безпеки мереж. Рис. 5.5 показує, як віртуалізація мережі використовує алгоритми вбудовування з метою виділення віртуальних ресурсів на фізичну інфраструктуру оптимальним чином. VNO використовує алгоритми вбудовування, щоб вирішити, які віртуальні ресурси можна вимагати від VNP, який, в свою чергу, створює їх за допомогою фізичних ресурсів InP [103].

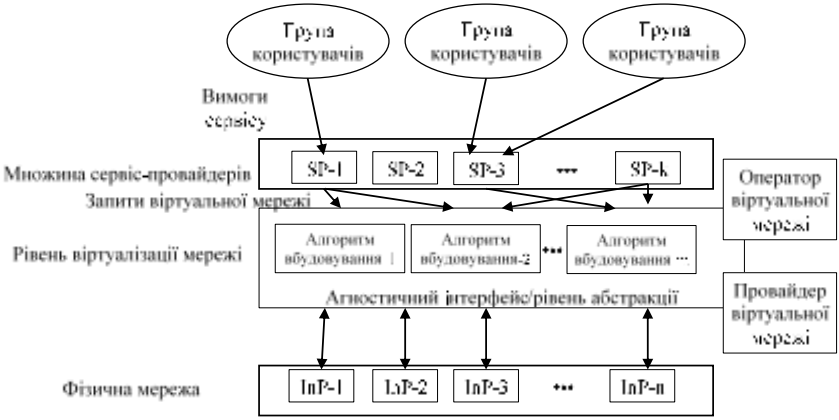


Рис. 5.5 Виділення ресурсів в майбутніх комунікаційних системах [103]

5.1.6. Задача розміщення та формування ланцюга мережевих функцій

Як коротко пояснювалося раніше, розміщення та формування ланцюга мережевих функцій складається з зв'язування набору мережевих функцій (наприклад, брандмауеру, балансувальника навантаження і т.д.) через мережу, щоб забезпечити правильне обслуговування мережевих потоків. Ці потоки повинні

пройти з шляхи з кінця в кінець, що проходять через певний набір функцій. По суті, ця проблема може бути розкладена на три етапи: (I) розміщення, (II) призначення, і (III) формування ланцюга.

Етап розміщення полягає у визначенні кількості екземплярів мережевої функції, необхідних для задоволення поточної/очікуваної вимоги, та їх місцерозташування в інфраструктурі. Віртуальні мережеві функції, як очікується, будуть розміщені на мережевих точках присутності (Network Point of Presence – N-PoP), які представляють собою групи (загальних) серверів в певних місцях інфраструктури (з ресурсами обробки). N-PoP, в свою чергу, потенційно будуть встановлені або в місцях з раніше встановленими пристроями комутації і/або маршрутизації або в приміщеннях таких як датацентри.

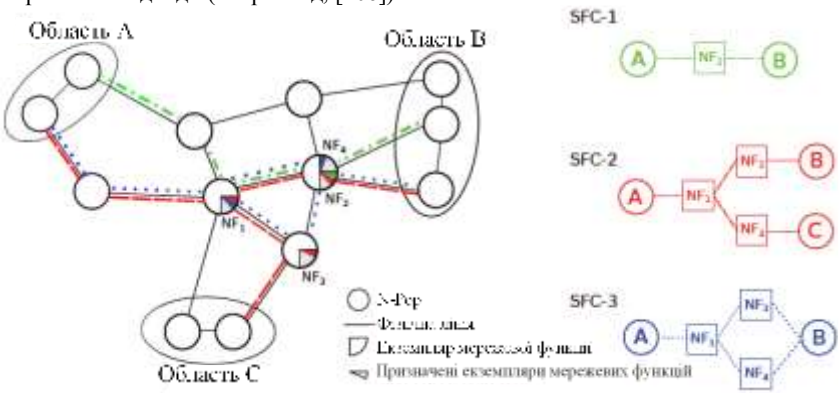
Етап призначення визначає, які розміщені екземпляри віртуальних мережевих функцій (в N-PoP) будуть відповідати за кожен потік. На основі джерела і призначення потоку, їм призначаються екземпляри таким чином, щоб запобігти недопустимим затримкам через час обробки. Наприклад, може бути більш ефективним призначати запити мережевих функцій найближчому екземпляру віртуальної мережевої функції або просто розділити запитувану вимогу між двома або більше віртуальними мережевими функціями (коли можливо).

На третьому і завершальному етапі формують ланцюги запитуваних функцій. Цей процес складається з створення шляхів, які зв'язують мережеві функції, розміщені і призначені на попередніх етапах. Цей етап враховує два важливі чинники, а саме затримки на шляху з кінця в кінець і різні затримки обробки, що додаються різними віртуальними мережевими функціями. На рис. 5.6 показані основні елементи, які беруть участь у віртуальному розміщенні та побудові ланцюга віртуальних мережевих функцій. Фізична мережа складається з N-PoP, з'єднаних за допомогою фізичних каналів. Існує множина запитів SFC (Service Function Chain), які містять логічні послідовності мережевих функцій, а також кінцеві точки, які неявно визначають шляхи. Крім того, провайдер має набір образів віртуальних мережевих функцій, екземпляри яких він може створ.вати. На рисунку, великі півкола представляють екземпляри мережевих функцій, що працюють на N-PoP, в той час як вписані півкола представляють запити мережевих функцій, призначені розміщеним екземплярам. Сіра область у великих півколах представляє собою обчислювальну потужність, виділену для мережевих функцій, які в даний час не використовуються. Штрихові лінії позначають шляхи, що формують ланцюг кінцевих точок і мережевих функцій [110].

Хоча NFV обіцяє істотну економію коштів, гнучкість і простоту розгортання, потенційні проблеми в реалізації віртуалізованих мережевих елементів, які можуть підтримувати вимоги до продуктивності реального світу, і досі залишаються відкритим питанням [100], й в даний час NFV все ще перебуває на початкових етапах реалізації [101]. Предмет віртуалізованого EPC, що може динамічно реконфігуруватись, є досить новим сам по собі. Саме тому в цій сфері не так багато напрацювань [94].

Розподіл ресурсів в NFV мережах подібний до розміщення прикладних програм в датацентрах і хмарах [102]. Задача розміщення функцій також тісно пов'язана з вкладенням віртуальної мережі (Virtual Network Embedding – VNE) [103]. В останні роки багато наукових досліджень [104] розглядали проблему VNE або

застосовуючи математичні моделі оптимізації (наприклад, [93], [99], [105]) або алгоритмічні підходи (наприклад, [106]).



а) Фізична інфраструктура **б) SFC запити**

Рис. 5.6 Приклад розгортання SFC на фізичній інфраструктурі для задоволення ряду запитів [110]

Наприклад, проблема розміщення контролерів розглядалась в [107] і [108], але з основним акцентом на досягнення мінімальної затримки SDN управління, а також стійкість надання послуг. У дослідженні [93] розглядається оптимальне розміщення датацентрів, на яких розміщуються віртуалізовані шлюзи, а також вирішується задача застосування віртуалізації та SDN декомпозиції на шлюзах мобільного ядра. Однак задача розміщення виникає і для інших функцій мережі, зокрема мобільного ядра EPC. Подібно до проблеми розміщення контролерів в області SDN, такі технології як NFV вимагають відповідних алгоритмів, які можуть вирішувати проблеми, що виходять за рамки однокритеріальних проблем розміщення. Ці проблеми можуть вносити додаткові складності через можливі взаємозалежності між мережевими функціями, як у випадку об'єднання функцій у ланцюжки, і нові потенційні обмеження відносно додаткових аспектів, таких як безпека [109].

Сучасні дослідження в основному розглядають тільки статичні запити віртуальної мережі, де запит і вимоги на ресурси є фіксованими і не змінюються з часом. Тим не менше, більшість запитів до датацентрів мають динамічні характеристики.

Далі ми розширюємо підхід VNE, визначаючи модель розташування мережевих функцій, що включає в себе поняття гібридних мереж, які містять як фізичні засоби обслуговування, так і віртуалізовані сервісні сутності. Схожим чином [99] також фокусується на розгортанні віртуальних мережевих функцій в гібридному середовищі, проте не враховує той факт, що продуктивність сервісної сутності залежить від виділених їй ресурсів. Також розглядається питання про те, як оптимально реконфігурувати розгорнуту віртуальну мережу в умовах зміни навантаження.

Також важливо відзначити, що наша робота відрізняється від більш загальних стратегій управління ресурсами в хмарних середовищах, оскільки EPC

являє собою граф взаємозалежних вузлів, які не можуть розглядатися ізольовано [94].

Таким чином, метою дослідження є підвищення ефективності мобільної мережі за рахунок оптимального розміщення ресурсів в середовищі гібридних датацентрів. У цьому напрямку пропонується підхід до моделювання і дослідження динамічного виділення ресурсів для мережевих функцій у мережі телекомунікаційного оператора.

Задача виділення фізичних ресурсів в NFV може бути розділена на дві частини: (1) вкладення/відображення віртуальних машин на фізичні машини, яка пов'язана з VNE;(2) відображення і планування VNF на створених віртуальних вузлах.

5.2. Опис методу відображення віртуальних вузлів на фізичні вузли

5.2.1. Модель

Запропонований підхід базується на спільному розташуванні індивідуальних ланцюгів сервісів ядра мережі на фізичній мережі, де ланцюг сервісів ядра мережі позначає послідовність мережевих функцій мобільного ядра, яку потік трафіку повинен пройти [104]. Припускаємо, що віртуальні мережеві функції мобільного ядра мають таку ж функціональність і інтерфейси як і мережеві елементи ядра архітектури 3GPP LTE Evolved Packet Core (EPC).

Фізична мережа задана у вигляді графа $SN = (N,L)$, де N є множиною фізичних вузлів і L – множиною каналів. Кожен канал $l = (n_1,n_2) \in L$, $n_1,n_2 \in N$ має максимальну пропускну здатність $c(n_1,n_2)$ і кожен вузол $n \in N$ пов'язаний з певними ресурсами c_n^i, $i \in R$, де R – множина типів ресурсів. Множина усіх точок агрегації трафіку (Traffic Aggregation Point – TAP), тобто кластерів eNodeB, в мережі позначається $T \subseteq N$. Для кожного вузла $n \in N$, $virt_n$ є бінарним параметром, який вказує, чи є вузол n віртуальним, $phys_n^j$ – бінарний параметр, який вказує, чи є вузол n виділеним апаратним блоком функції типу $j \in F$, де F є множиною типів мережевих функцій.

Віртуальне ядро мобільної мережі представлено множиною ланцюгів сервісів (один ланцюг на TAP), які вбудовуються в фізичну мережу.

Вимоги смуги пропускання каналу між двома функціями, $j1$ і $j2$, що відносяться до ланцюга сервісів TAP $t \in T$ позначається як $d_t^{(j1,j2)}$. $d_t^{j,i}$ – кількість ресурсу типу i, що виділяється для мережевої функції j TAP t. $s_t^{j,i}$ позначає час обробки запиту на ресурсі типу i мережевої функції j TAP t однією одиницею ресурсу для випадку віртуальної мережевої функції. Задано значення оброблюваних запитів в секунду виділеного апаратного функціонального блоку n типу j, і позначено як μc_n^j. Вимоги до запитів в секунду мережевої функції j, що відноситься до ланцюга сервісів TAP t, позначаються як M_t^j.

Метою оптимізації є знаходження розташування ланцюгів сервісів ядра мережі (тобто розміщення мережевих функцій ядра мережі та розподіл ресурсів, а також визначення шляхів передачі трафіку між ними), так щоб мінімізувати витрати

на зайняті ресурси каналів і вузлів у фізичній мережі, при цьому задовільняючи вимоги трафіку. Сформулюємо цільову функцію (формула (5.1)) у вигляді лінійної комбінації (з ваговими коефіцієнтами *a, b, c*) трьох вартісних виразів: базова вартість *cost(n)*, яка має місце якщо якась мережева функція розміщується на фізичному вузлі $n \in N$, вартість зайнятої одиниці ресурсів *cost(i,n)* на фізичному вузлі *n* і вартість зайнятої одиниці пропускної здатності $cost(n_1,n_2)$ на фізичному каналі $(n_1,n_2) \in L$.

Наступні формули (5.1-5.10) представляють собою постановку оптимізаційної задачі нелінійного програмування. Змінні ${x_n}^{t,j}$ вказують, чи мережева функція *j* пов'язана з ТАР $t \in T$ розташовується на фізичному вузлі *n*. Для *j*=ТАР, ${x_n}^{t,TAP}$ – не змінні, а вхідні параметри, які вказують де ТАР $t \in T$ знаходиться, тобто

$$ {x_n}^{t,TAP} = \begin{cases} 1 \; if \; t = n, \\ 0 \; else. \end{cases} $$

Аналогічно, змінні ${f_{(n_1,n_2)}}^{t,(j_1,j_2)}$ вказують, чи фізичний канал $(n_1,n_2) \in L$ використовується для шляху між j_1 і j_2 для ТАР $t \in T$.

$$ \min_{{x_n}^{t,j},{f_{(n_1,n_2)}}^{t,(j_1,j_2)},{d_t}^{j,i}} \left(a \cdot \sum_{n\in N} x_n \cdot cost(n) + b \cdot \sum_{n\in N} \sum_{t\in T} \sum_{j\in V} \sum_{i\in R} {x_n}^{t,j} \cdot {d_t}^{j,i} \cdot cost(i,n) + c \cdot \sum_{(n_1,n_2)\in L} cost(n_1,n_2) \cdot \sum_{t\in T} \sum_{(j_1,j_2)\in E} {f_{(n_1,n_2)}}^{t,(j_1,j_2)} \cdot {d_t}^{(j_1,j_2)}\right) \tag{5.1} $$

$$ \text{Subject to } \sum_{n\in N} {x_n}^{t,j} = 1 \;\; \forall t \in T, j \in V \tag{5.2} $$

$$ {x_n}^{t,j} \le {d_t}^{j,i} \;\; \forall t \in T, j \in V, n \in N, i \in \{R \backslash bdw\} \tag{5.3} $$

$$ \sum_{(w,n)\in L} \sum_{t\in T} \sum_{(j_1,j_2)\in E} {f_{(w,n)}}^{t,(j_1,j_2)} \cdot {d_t}^{(j_1,j_2)} \le {c_n}^{bdw} \;\; \forall n \in N \tag{5.4} $$

$$ \sum_{t\in T} \sum_{j\in V} {x_n}^{t,j} \cdot {d_t}^{j,i} \le {c_n}^{i} \;\; \forall n \in N, i \in \{R \backslash bdw\} \tag{5.5} $$

$$ \sum_{t\in T} \sum_{(j_1,j_2)\in E} {f_{(n_1,n_2)}}^{t,(j_1,j_2)} \cdot {d_t}^{(j_1,j_2)} \le c(n_1,n_2) \;\; \forall (n_1,n_2) \in L \tag{5.6} $$

$$ \sum_{(n,w)\in L} {f_{(w,n)}}^{t,(j_1,j_2)} - {f_{(n,w)}}^{t,(j_1,j_2)} = {x_n}^{t,j_1} - {x_n}^{t,j_2} \;\; \forall t \in T, n \in N, (j_1,j_2) \in E \tag{5.7} $$

$$ {x_n}^{t,j}, {f_{(n_1,n_2)}}^{t,(j_1,j_2)} \in \{0,1\} \;\; \forall t \in T, j \in V, n \in N, (j_1,j_2) \in E, (n_1,n_2) \in L \tag{5.8} $$

$$ {\mu_t}^{j} \ge {M_t}^{j} \;\; \forall t \in T, j \in V \tag{5.9} $$

$$ {\mu_t}^{j} = \sum_{n\in N} \left({x_n}^{t,j} \cdot \sum_{i\in R} \frac{{d_t}^{j,i}}{{s_t}^{j,i}} \cdot virt_n + {x_n}^{t,j} \cdot {\mu c_n}^{j} \cdot {phys_n}^{j} \right) \;\; \forall t \in T, j \in V $$

$$ \sum_{(j_1,j_2)\in E} \sum_{(n_1,n_2)\in L} {f_{(n_1,n_2)}}^{t,(j_1,j_2)} \cdot L(n_1,n_2) \le L_t \; \forall t \in T \tag{5.10} $$

Змінні x_n є булевими змінними, що позначають чи яка-небудь мережева функція розміщується на вузлі $n \in N$, наприклад x_n=0 означає, що немає жодної мережевої функції з будь-якого ланцюга сервісів, яка би розміщувалася в цьому вузлі.

Вираз (5.2) гарантує, що для кожної ТАР/ланцюга сервісів розміщується тільки одна мережева функція кожного типу. Вираз (5.3) гарантує, що розміщення ресурсів здійснюється на фізичних вузлах, які було обрано для розташування

відповідних мережевих функцій. Вирази (5.4), (5.5) і (5.6) являють собою обмеження на ресурси фізичних вузлів і каналів. Слід зауважити, що канал між двома мережевими функціями відображається на шлях у фізичній мережі. Таким чином, його вимоги до полоси пропускання впливають не тільки на мережеві ресурси фізичних вузлів, де мережеві функції розміщуються, але й на мережеві ресурси проміжних вузлів, які лежать на шляху (вираз (5.4)). Вираз (5.7) представляє собою обмеження щодо збереження потоку для всіх шляхів у фізичній мережі. Вираз (5.8) гарантує, що змінні у задачі розміщення функції мережі та відображення шляху є булевими.

Для того, щоб урахувати у моделі продуктивність, обмеження на значення запитів в секунду, виражені у виразі (5.9), необхідні, щоб гарантувати, що загальна швидкість обслуговування мережевої функції j ТАР t, яка позначається як μ_t^j, перевищує необхідну величину.

Щоб обмежити затримки на каналах, обмеження на затримку, показане в виразі (5.10), також додається, щоб обмежити загальну затримку на всьому шляху.

На рис. 5.7 представлено приклад системи розподілу мережевих запитів.

Передбачається вирішення задачі (5.1)-(5.10) в офлайн режимі на початковому етапі. Згідно з рішенням, кожному блоку мережевої функції виділяється певна кількість ресурсів, на основі грубої оцінки його потреби в ресурсах; миттєві потреби різних мережевих функцій динамічно задовольняються під час виконання таким чином, щоб задовольнити гарантії передбачені для кожної мережевої функції.

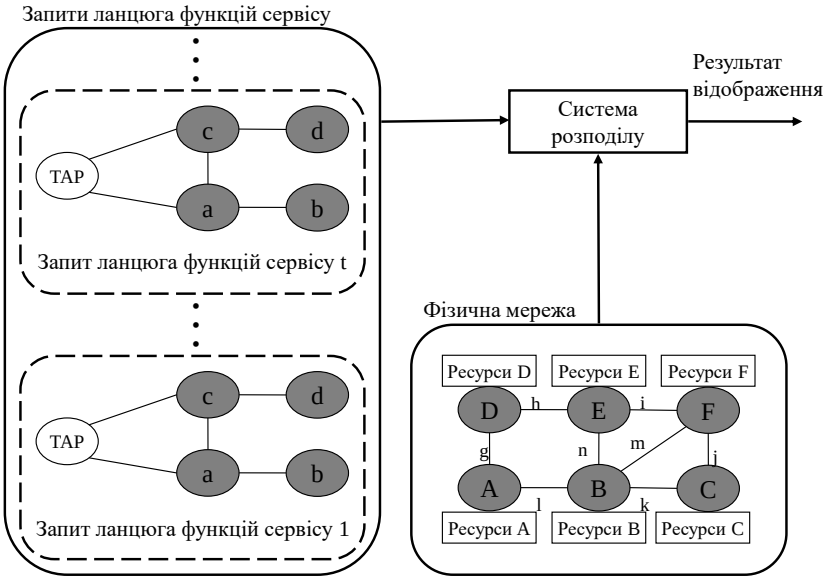


Рис. 5.7 Система виділення мережевих ресурсів – приклад топології

5.2.2. Оцінка

Вирішення задачі (5.1)-(5.10) здійснювалось з використанням генетичного алгоритму у системі MATLAB. Фрагмент коду програми представлено на рис. 5.8.

```
>>ObjectiveFunction=@fitness;
>>nvars=108;% Number of variables
>>LB=zeros(1,108);% Lower bound
>>UB=ones(1,90);% Upper bound
>>IntCon=ones(1,90);% Integer variables
>>ConstraintFunction=@constraint;
>>[x,fval]=ga(ObjectiveFunction,nvars,[],[],[],[],LB,UB,ConstraintFunction,IntCo
n,options)
```

Рис. 5.8 Фрагмент коду MATLAB

Простий приклад моделювання системи з десятьма вузлами, трьома функціональними блоками та двома типами ресурсів показав, що у порівнянні з фіксованим виділенням ресурсів можна отримати виграш у 3 рази. Зміну загальних витрат при зміні вимог до обслуговування вхідного навантаження проілюстровано на рис. 5.9.

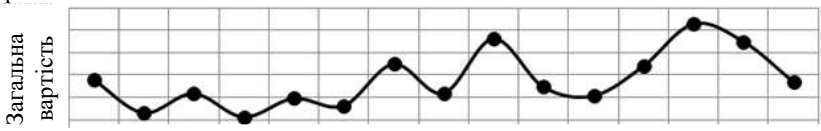


Рис. 5.9 Загальна вартість розміщення при зміні вимог до інтенсивності обслуговування

Графік показує, що вартість оптимального виділення ресурсів значно змінюється зі зміною потрібної інтенсивності обслуговування запитів, тобто, відповідно, від інтенсивності надходження запитів на обслуговування, що має місце протягом дня. Різниця у вартості може сягати трьох разів.

5.2.3. Підхід до динамічного розподілу ресурсів

Розглянемо випадок, коли необхідно збільшити кількість ресурсів, виділених блоку мережевої функції.

Метою виконання алгоритму перерозподілу ресурсів (рис. 5.10) буде знаходження кількості ресурсів, які слід додати до мережевої функції, а також їх місцерозташування. Передбачається три кроки алгоритму. На першому кроці перевіряється можливість виділення необхідної кількості ресурсів на вузлі, на якому функціональний блок *j* для ТАР *t* розгорнуто в поточний момент. Якщо ресурсів недостатньо, то на другому кроці здійснюється перевірка наявності необхідної кількості ресурсів на вузлах, де розгорнуто функціональні блоки *j* для інших ТАР. Якщо ресурсів на таких вузлах недостатньо, то на третьому кроці здійснюється перевірка на решті вузлів. Якщо на вузлі, відмінному від поточного (кроки 2 і 3), кількість необхідних для блоку мережевої функції ресурсів виявляється вільною, то здійснюється міграція мережевої функції на такий вузол.

Для випадку, коли необхідно відняти ресурси, то виконується просте звільнення ресурсів.

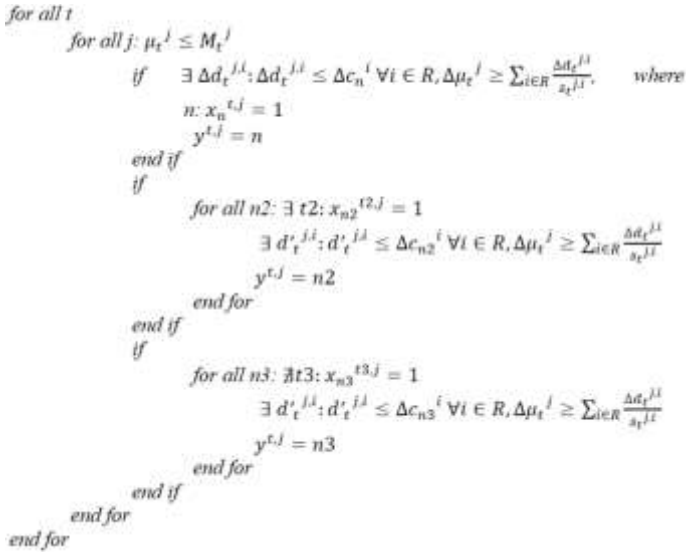


Рис. 5.10 Алгоритм динамічного виділення ресурсів

Де на рис. 5.10 як $\Delta d_t^{j,i}$ позначену зміну у вимогах до ресурсів, $d'^{j,i}_t = d_t^{j,i} + \Delta d_t^{j,i}$, Δc_n^i – залишкові ресурси типу i на вузлі n, $\Delta\mu_t^j$ – різниця у вимогах до обслуговування між попереднім і наступним моментом часу функціонального блоку j для ТАР t, $y^{t,j}$ – змінна, яка вказує розташування функціонального блоку j для ТАР t.

Увесь процес виділення ресурсів зображено на рис. 5.11.

Процес виділення ресурсів починається з приходом запиту ланцюга функцій сервісу, як показано на рис. 5.11. Метод виділення ресурсів початкового етапу (тобто (5.1) – (5.10)) використовується для вбудовування запитів ланцюга функцій мережі, він приймає на вхід поточний стан фізичної мережі (наприклад, доступні ресурси CPU, пам'яті та пропускної здатності) і самі запити. З приходом нового запиту, тобто якщо досягається межа рівня обслуговування, для вбудовування запитів використовується метод виділення ресурсів з рис. 5.10.

Метод виділення ресурсів направлений на мінімізацію використання ресурсів, при цьому забезпечуючи заданий рівень якості обслуговування. Запропонований метод дозволяє телекомунікаційному оператору мінімізувати капітальні та операційні витрати використовуючи парадигму хмарних обчислень та значно покращити якість сприйняття.

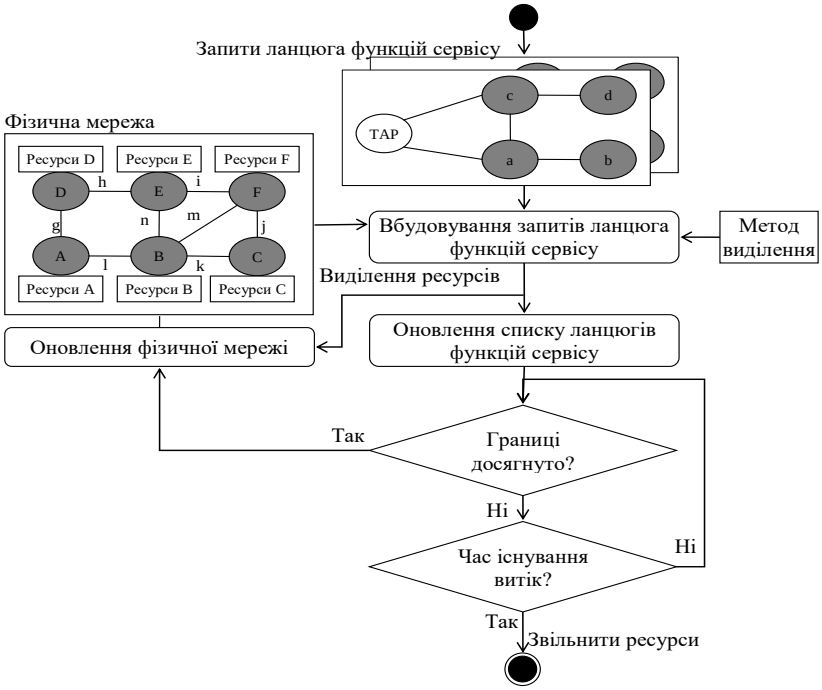


Рис. 5.11 Життєвий цикл мережевого запиту – діаграма активності

5.3. Опис методу відображення і планування мережевих функцій

Для задачі відображення і планування мережевих функцій (Network Function Mapping And Scheduling – NFMS) є багато можливостей для спільного використання ресурсів. Однією з них є те, що для кожного VNF використовується виділена віртуальна машина. Це явно не буде доцільним, оскільки фізичні ресурси будуть легко виснажуватися, і буде марнотратством ресурсів, так як більшість функцій є "легкими", і, отже, можуть бути оброблені однією віртуальною машиною, наприклад, за допомогою контейнерів у віртуальній машині. Далі розглядається підхід до спільного використання ресурсів, який дозволяє для даної віртуальної машини обробляти декілька VNF, один за одним (можливо) з черги.

Тому NFMS складається з необхідності обробляти мережеві сервіси онлайн (кожен сервіс створюється і вбудовується в міру його необхідності) з використанням набору $F = \{1, ..., f\}$ з f віртуальних мережевих вузлів. Будь-який заданий сервіс мережі S складається з послідовності $SF = \{1, ..., m\}$ з m VNF, де функція $1 \le i \le m$ повинна бути оброблена на множині $F(i) \subseteq F$ вузлів. Функції $\{1, ..., m\}$ повинні бути оброблені одна за одною в певній послідовності, і кожен віртуальний вузол може обробляти не більше однієї функції одночасно. Час обробки для функції i на вузлі $j \in F(i)$ складає $\rho_{ij}>0$, де $1 \le j \le f$, і під час обробки або в черзі для обробки, функція i

використовує буфер δ_i на вузлі, на який вона відображається. У будь-який момент, заданий вузол j має доступний розмір буфера B_j. Для кожної комбінації вузлів та функцій, визначаємо бінарну змінну $\beta_{i,j}$, яка приймає значення 1, якщо вузол j може обробляти функцію i, і 0 в іншому випадку. У цьому випадку задача полягає у виборі для кожної VNF i віртуального вузла $j \in F(i)$ і часу завершення t_i, коли її обробка буде завершена (час t_s, при якому обробка функції i на вузлі j починається, може бути отриманий з $t_s = t_i - \rho_{ij}$). Також визначаємо граничний термін t_l для обробки даного сервісу. Нарешті, для кожного віртуального вузла j, визначаємо очікуваний час завершення π_j останньої функції у черзі на обробку на вузлі, і для кожного сервісу визначаємо час прибуття t_a, який є часом, коли запит на відображення і планування сервісу отримується фізичною мережею.

Задачу NFMS можна розглядати як таку, що складається з двох частин: вирішення на які віртуальні вузли кожна VNF повинна бути відображена (проблема відображення), і для кожного вузла вирішення порядку, в якому відображені VNF повинні бути оброблені (проблема планування).

Алгоритм відображення та планування віртуальних мережевих функцій представлено на рис. 5.12 та рис. 5.13.

Алгоритм виконується наступним чином: після прибуття запиту на обслуговування, функції сервісу відображаються і диспетчеризуються послідовно. Для кожної функції i визначаються всі вузли $F(i) \subseteq F$, які мають можливість обробити її. Ці вузли потім ранжуються на основі жадібного критерію. Потім вузол з найкращим рангом обирається для відображення, а функція планується на обробку в кінці черги вузла. Фактичний час початку обробки основується і на доступності вузла (завершення обробки раніше поставлених в чергу функцій), і на завершенні обробки попередньої функції (якщо це застосовно). Крім того, на кожному етапі планування, визначається завершення обробки, щоб гарантувати, що воно знаходиться в межах встановленого граничного терміну для сервісу. Також цілком можливо, що під час відображення останньої функції сервісу час завершення перевищує граничний термін, або що функція не має жодного вузла-кандидата (через те, що всі вузли-кандидати повністю завантажені). У такому випадку пропонується здійснювати додавання ресурсів згідно з алгоритмом "Додавання VNF". На рис. 5.12 показується псевдокод алгоритму, де C позначає критерій, за яким ранжуються вузли.

Пропонується, що алгоритм базується на тому, що функції відображаються на вузли таким чином, щоб їх завантаженість була близькою до оптимального за критерієм енергоефективності значення.

Алгоритм 1. Жадібне відображення функцій (S, F, C)

Старт
Здійснити резервне копіювання стану фізичної мережі
for $i \in S$ *do*
Ініціалізація: Множина можливих вузлів $F'=0$
if $(i=1)$ *then*
  $t_{i-1}=t_a$
end if
for $j \in F$ *do*
  $t_e=\rho_{ij}+max(\pi_j,t_{i-1})$
  if $((\beta_{ij}==1)\wedge(B_j \geq \delta_i)\wedge(t_e \leq t_l))$ *then*
    $F'=F' \cup f$
  end if
end for
if $F' \equiv 0$ *then*
  Неуспіх відображення та планування
  Додавання VNF
  return
end if
  Відсортувати F' *відповідно до* C
  Обрати найвищий вузол j^* *з* F'
  Відобразити функцію i *на* j^*
  Встановити $t_i=max(\pi_j,t_{i-1})$
  Оновити B_j, π_j, t_{i-1}
end for
Відображення та планування завершене
Кінець

Рис. 5.12 Алгоритм відображення мережевих функцій

Алгоритм 2. Додавання VNF

$F=F \cup (f+1)$
$B_{f+1}=\delta_i$
$\pi_{f+1}=0$

Рис. 5.13 Алгоритм додавання мережевих функцій

Висновки

1. Запропоновано метод динамічного управління виділенням ресурсів мережевих функцій для визначення оптимальної кількості ресурсів, виділених мережевій функції у мережі телекомунікаційного оператора.

2. Представлено алгоритм управління виділенням ресурсів для задачі відображення та планування віртуалізованих мережевих функцій.

3. Метод може застосовуватись при управлінні розгортанням мережевих функцій у гетерогенному апаратному середовищі для мінімізації витрат оператора зв'язку та покращення якості обслуговування абонентів.

4. В подальших дослідженнях, запропонований метод може бути розширений до комплексного методу динамічного виділення ресурсів в умовах змінного навантаження та архітектури управління мережею мобільного зв'язку.

ЗАКЛЮЧЕННЯ

В монографії розглянуто математичні методи аналізу та керування телекомунікаційними мережами, зокрема ряд задач пов'язаних з процесом надання послуг абонентам телекомунікаційних мереж, досліджено особливості архітектурних рішень сучасних телекомунікаційних систем, систем керування мережами зв'язку як розподілених інформаційно обчислювальних систем, а також особливості архітектури мереж наступного покоління NGN, базові функції та принципи.

Досліджено **проблеми та задачі транспортування інформаційних потоків**, задачі керування мультисервісним потоком в комутаційному центрі транспортної мережі:

- Вперше запропоновано метод підвищення ефективності обслуговування інформаційних потоків та його застосування в процесі прийняття керівних рішень по розподілу інформаційних потоків між вихідними каналами пограничного комутаційного центру транспортної мережі, які за рахунок класифікації та динамічного перерозподілу трафіків в моменти перевантаження, дозволяють підвисити ефективність передачі мультисервісного інформаційного потоку та зменшити втрати пакетів на 10-15% без зниження якості.

- Запропоновано удосконалений спосіб кругового обслуговування черг заявок, що дозволяє мінімізувати втрати пакетів високопріоритетних трафіків з одночасним збереженням якості передачі низькопріоритетного трафіку.

- Розроблено принцип комплексного керування інформаційними потоками в комутаційному центрі із застосуванням запропонованого методу підвищення ефективності обслуговування трафіків, який дозволяє ефективно керувати маршрутизацією та розподілом навантаження в вихідних тунелях транспортної мережі, враховувати вимоги до часу затримки інтерактивних трафіків шляхом обмеження довжини черги, контролювати та вчасно покращувати показники втрат пакетів.

Досліджено **проблеми та задачі обслуговування викликів в центрах керування мобільною мережею** зв'язку:

- Проведено аналіз стану та проблем при наданні послуг в мережах мобільного зв'язку, який показав, що визначальними характеристиками послуг для систем обробки викликів та тарифікації є смуга пропускання, якість надання, тарифікація, адаптація та інформаційна важливість контенту, можливість враховування яких виявлено при обробці послуг та їх тарифікації в мережах LTE/PCC, завдяки чому можна підвищити технічну й економічну ефективність механізмів системи обробки викликів та гнучкість тарифікації мобільних операторів зв'язку.

- Вперше запропоновано метод обробки викликів та тарифікації, який відрізняється від існуючих тим, що враховує технічні параметри функціонування мережі (ширину смуги частот, кількість піднесучих) та, за рахунок модифікації протоколів і створення нових інтерфейсів, дозволяє отримати більш гнучку систему обробки викликів та тарифікації, і збільшити ефективність функціонування системи: зменшити сумарну кількість штрафів на 10% і підвищити інтегральну вагу обслугованих заявок на 15% без зниження середньої якості обслуговування.

- Запропоновано дисципліну обслуговування викликів в системі обробки викликів та тарифікації, особливістю якої є введення ситуаційного пріоритету –

відношення тарифу послуги (значення інформаційної ваги) до ширини смуги радіочастотного ресурсу при наданні персоналізованих послуг, що дозволяє обслуговувати в першу чергу заявки, які мають найбільше значення коефіцієнту. Запропоновано удосконалену архітектуру системи обробки викликів і тарифікації в системі IMS, яка включає модифікований протокол Diameter, новий інтерфейс *Ty* і протокол DMShort, що дозволило передати параметри ширини смуги частот з метою їх врахування при тарифікації послуги.

Досліджено **задачі організації роботи системи онлайн тарифікації**, запропоновано підходи щодо зменшення навантаження на систему тарифікації оператора телекомунікаційних послуг.

- Запропоновано модель обслуговування заявок на тарифікацію абонентів, яка полягає в тому щоб в моменти перевантаження не обслуговувати post-paid абонентів в режимі реального часу, доведено ефективність запропонованої моделі, а саме розраховано оцінку прибутку від впровадження моделі, досліджено витрати на впровадження моделі. Запропоновано підхід до класифікації post-paid абонентів за рівнем ризику, розраховані середні збитки від обслуговування відповідно до запропонованої моделі обслуговування.

- Запропоновано метод оптимізації ємності буфера очікування заявок на тарифікацію, в якому наведені умови економічної доцільності нарощування буферу білінгової системи, залежно від навантаження на систему та прибутку від збільшення потужностей обладнання оператора зв'язку.

- Запропоновано метод розподілу ресурсів системи онлайн тарифікації для забезпечення ефективної обробки заявок, який дозволяє здійснювати контроль за якістю надання послуг, враховує необхідну кількість ресурсів для обслуговування однієї заявки, що дозволяє виділяти ресурси пропорційно вимогам сервісу, враховує статистичні дані про кількості заявок різних типів сервісів, які надходять у заданих інтервалах часу, що дозволяє налаштовувати розподіл технічних ресурсів, що обслуговують заявки, відповідно до складу вхідного навантаження, а також забезпечити максимізацію економічної ефективності процесу надання послуг. за критерій доцільності вибрана економічна ефективність надання послуг.

- Запропоновано метод контролю перевантажень в системі онлайн 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.

- З метою формування дисципліни подачі заявок на тарифікацію із буфера на обслуговування, запропоновано метод керування вхідним потоком, який полягає формуванні послідовності кількостей заявок на тарифікацію, які подаються на обслуговування, що забезпечує згладжування використання фізичних ресурсів системи тарифікації

Досліджено тенденції залучення хмарних ресурсів для забезпечення роботи серверу тарифікації. Запропоновано метод визначення моменту включення додаткового технічного засобу, який дозволяє проводити оцінку динаміки вхідного навантаження та поточного стану технічних засобів, та завчасно увімкнути додатковий ресурс, та запобігли перевантаженню існуючих ресурсів. Запропоновано метод складання розкладу включення серверів, який на основі довгострокових

статистичних даних, дозволяє скласти розклад залучення додаткових технічних ресурсів (включення серверів) отже спланувати роботу технічних засобів у періоди часу з різним очікуваним навантаженням.

Досліджено **задачі організації роботи системи виділення ресурсів для віртуалізованих мережевих функцій в гібридному середовищі**, запропоновано підходи щодо розміщення мережевих функцій та їх диспетчеризації.

ЛІТЕРАТУРА

1. Antonio Cuevas, Jose Ignacio Moreno, Hans Einsiedler. IMS Service Platform: A Solution for Next-Generation Network Operators to Be More than Bit Pipes.// IEEE Communication Magazine .2006. Aug. P.75-81.
2. Syed A. Ahson, Mohammed Ilyas. IP Multimedia subsystem (IMS) handbook/ CRC Press, 2009. P. 250
3. Ільченко М.Ю. Сучасні телекомунікаційні системи./М.Ю. Ільченко, С.О. Кравчук– К.: НВП «Видавництво «Наукова Думка» НАН України. – 328 с.: іл.
4. Gonzalo Camarillo, Miguel A. Garcia Martin. The 3G IP Multimedia subsystem (IMS): merging the Internet and cellular world.- 3rd ed., 2009. ISBN 978-0470-51662-1.
5. Khalid Al-Begain. IMS:development an deployment perspective, 1st edition. John Wiley& Sons, 2009 . P. 320.
6. Послуги Київстар [Електронний ресурс]: http://www.kyivstar.net/ personal/prepaid/services/.
7. Послуги МТС [Електронний ресурс]: http://www.mts.com.ua/ ukr/services.php.
8. Послуги УТЕЛ [Електронний ресурс]: http://utel.ua/private/services.php.
9. Support of Third Generation Services using UMTS in a Converging Network Environment./ UMTS Forum, 2004. Report 14. [Електронний ресурс]: http://umts-forum.org /reports/ report14.pdf.
10. Закон України «Про телекомунікації» від 18.11.2003 № 1280-IV.
11. Величко В.В. Передача данных в сетях мобильной связи./ В.В. Величко – М: Радио и Связь Горячая линия-Телеком, 2005.- 322 с.
12. Hao Wang. 4G wireless video communication, 1st edition./ John Wiley and Sons Ltd , 2009. P. 320.
13. Груздев А. В., Мухин П. В. Аспекты развития систем сотовой связи 4–го поколения. III Науково-практична конференція «Актуальні питання регулювання у сфері телекомнікацій та користування радіочастотним ресурсом України»/ /ГП УНИИРТ, Одесса 2009. – с. 25-40.
14. Глоба Л.С. Послуги в сучасних мобільних мережах України./ Л.С Глоба, О.О. Зінченко, О.М. Дяденко, І.М. Попова.// Электроника и связь. Тематический выпуск «Электроника и нанотехнологии». – ч. 2. – 2009. – С. 291-296.
15. Oumina D. Ranc. Towards a Real Time Charging Framework for Complex Applications in 3GPP IP Multimedia Subsystem.// Proceedings of the Conference on Next Generation Mobile Applications, Services and Technologies. – September 2007. – P . 145–150.
16. Мова В.В. Организация приоритетного обслуживания в АСУ./ Мова В.В. Л.А. Пономаренко, А.М Калиновски – Киев: «Техника»,1977. – 160 с.
17. Бусленко Н.П.Лекции по теории сложных систем./ Н.П. Бусленко, В.В. Калашников, И.Н Коваленко – М.: «Сов. радио», 1973.- 440с.
18. Джейсуол Н. Очереди с приоритетами./ Н. Джейсуол– М.: «Мир», 1973.- 279 с.

19. Конвей Р.В. Теория расписаний./ Р.В. Конвей, В.Л. Максвелл, Л.В.Миллер – М.: «Энергия», 1975.- 360 с.

20. Вишневський В.М. Системы полинга теория и применении в широкополосных беспроводных сетях./ В.М. Вишневський, О.В. Семенова – М.: Техносфера, 2007.-312с. ISBN 978-5-94836-166-6.

21. G.Priggouris, S.Hadjiefthymiades, LMarakos. Supporting IP QoS in the GPRS.// IEEE Network. – September 2009. –P.22-31.

22. Рыков В.В.Об оптимальных динамических приоритетах в однолинейных системах массового обслуживания./ В.В. Рыков, Э.Е. Лемберг// Изв. АН СССР. Техническая кибернетика.– 1967. – №1. – С. 25-34.

23. Maria Kuotsoploul, Alexandros Kaloxylos, Athanassia Alonistioti, Lazaros Merakos. Communication Networks Laboratory.//IEEE Communications. Surveys & Tutorials, Q12004. – P. 50-58.

24. P. Kurtansky. State of the Art Prepaid Charging for IP Services. // TIK-ReportNr. 236. – WWIC 2006, LNCS 3970. – P. 143–154

25. P. Kurtansky, P. Reichl, J. Fabini, T. Lovric, B. Stiller. Efficinet prepadai charging for 3GPP IMS.// Society for Design and Process Science. – 2006, June. – P. 43.

26. Kurtansky P., Stiller B. Prepaid Charging for QoS-enabled IP Services based on Time Intervals.// TIK-Report Nr. 222. – 2005, June. – P 16.

27. ETSI 123.203 Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); LTE; Policy and charging control architecture, p. 118.

28. 3GPP TS 32.299 R9. Telecommunication management.Charging management. Diameter Charging applicatio, P.143.

29. M. Falkner, M. Devetsikiotis, I. Lambadaris. An Overview of Pricing Concepts for Broadband IP Networks. // IEEE Commun.Surveys. – 2nd Quarter, 2000.

30. Aboba, J. Arkko, and D. Harrington. Introduction to Accounting Management. // RFC 2975, Oct. 2000.

31. C. Rigney. RADIUS Accounting.// RFC 2866, June 2000.

32. B. Aboba, J. Arkko, and D. Harrington.Introduction to Accounting Management.// RFC 2975, Oct. 2000.

33. G. Zorn D. Mitton B. Aboba. RADIUS Accounting Modifications for Tunnel Protocol Support.// RFC 2139, August 1999

34. H. Jonkers and S. Hille. Accounting Context: Application and Issues. Oct. 2000. [Електронний ресурс]: www.aaaarch.org/doc06/file-11249.pdf

35. Васильев М. Математическое моделирование систем связи : учебное пособие / М. Васильев, Н. Служивый – К.: УлГТУ, 2008. – 170 с. ISBN 978-5-9795-0100-0.

36. Вентцель Е. С. Теория вероятностей: учебник для втузов. Издание 8-е, перераб. и доп./ Е. С. Вентцель – М. : Физматлит, 1999. – 576 с.

37. Pascal Kurtansky, Burkhard Stiller.State of the Art Prepaid Charging for IP Services.// TIK-Report Nr. 236, November 2005.

38. Maria Koutsopouloiu. Charging, accounting and billing scheme in mobile telecommunication networks and the Internet./ Maria Koutsopouloiu, Alexandros Kaloxylos, Athanassia Alonostioti, Lazaros Merakos. // Communications Surveys & Tutorials . – First Quarter, 2004. – P.50, 58.

39. H. Oumina and D. Ranc.Towards a Real Time Charging Framework for Complex Applications in 3GPP IP Multimedia Subsystem.//Proceedings of the Conference on Next Generation Mobile Applications, Services and Technologies, September 2007. P. 145–150.

40. Романов А.И. Основы теории телелкоммуникационных сетей: учебное пособие для вузов./ А.И. Романов– К.,2002 -152, ил. 84.

41. Основи управління мережами та послугами телекомунікацій: Підруч. для студ. вищ. навч. закл. за напрямком «Телекомунікації»/З ред. проф. Стеклова В.К.- К:Техніка, 2002.- 438 с.

42. Беркман Л.Н. Методи розрахунку показників якості конвергентних мереж на базы теорії ігор./ Л.Н. Беркман, О.М. Ткаченко, Григорович, П.Ю. Дещинський// Электроника и Связь. Тематический выпуска «Проблемы электроники» – ч.1. – 2008. – С. 220-221.

43. Вінницький В. П. Термінальне устаткування та передавання інформації в телекомунікаційних системах: Підручник./ В. П. Вінницький, В. Г. Поліщук – 2004. – 436 с.:аб.

44. Зайченко О.Ю. Аналіз иа оптимізація показників якості та структур копютерних мереж з технологією ATM: дисертація на здобуття наук.ступеня д-ра техн.наук: спец. 05.13.06/О.Ю. Зайченко. – К., 2005 р.

45. Elwalid A. MATE: MPLS Adaptive Traffic Engineering / A. Elwalid, C. Jin, S. Low, I. Widjaja // IEEE INFOCOM – 2003.

46. Урывский Л.А. Методика управления характеристиками обслуживания при изменении требований к качеству связи./ Л.А. Урывский, Е.А Прокопенко// Наукові записки УНДІЗ.– №3(11).– 2009.

47. Perkins C. IP Mobility Support for IPv4 / C. Perkins // RFC 3344. - August 2002. Checland P.B. Soft systems methodology: an overview // J. Appl. Syst. Anal. – 1988. –15. – P.27–36

48. Fernando A. Hernández Solana. QoS and Radio Resource Management in Multimedia Packet Transmission for 3G Wireless IP Networks./ Fernando A. Hernández Solana, A. Valdovinos Bardaji, Casadevall Palacio.// IEEE VTC 2004 Spring. – Milán, Italia.

49. Бакланов И.Г. NGN: принципы построения и организации. – М.: Эко-Трендз, 2007. – 400 с.

50. Сети следующего поколения NGN / Под ред. А.В. Рослякова. – М: Эко-Трендз, 2008. – 424 с.

51. Beran J. Statistics for Long-Memory Processes / J. Beran. - New York, Chapman and Hall. – 1994.

52. Boxma O. J. Regular Variation in a Multi-Source Fluid Queue / O. J. Boxma. - CWI Report BS-R9614. – 1996.

53. Estimators for long range dependence an empirical study by Murad S. Taqqu, Vadim Teverovsky and Walter Willinger in "Fractals". // Murad S. Taqqu, V. Teverovsky, W. Willinger. – Vol 3, No. 4. – 1995. – C. 785-788.

54. Maximum Likelihood estimation of stationary univariate fractionally integrated time series models by Fallaw Sowell. // Maximum Likelihood. – Journal of Econometrics 53. – 1992. – C. 165–188.

55. http://buggy.itm.hk-r.se/ttn410/models/

56. Олифер В.Г. Компьютерные сети. Принципы, технологии, протоколы: Учебник для вузов. 2-е изд. / Олифер В.Г., Олифер Н.А. – СПб: Питер. - 2005. – 864 с.

57. Estimators for long range dependence an empirical study by Murad S. Taqqu, Vadim Teverovsky and Walter Willinger in "Fractals". // Murad S. Taqqu, V. Teverovsky, W. Willinger. – Vol 3, No. 4. – 1995. – С. 785-788.

58. Maximum Likelihood estimation of stationary univariate fractionally integrated time series models by Fallaw Sowell. // Maximum Likelihood. – Journal of Econometrics 53. – 1992. – С. 165–188.

59. А. Лазарев В. Г. Динамическое управление потокими информации в сетях связи. / Лазарев В. Г., Лазарев Ю.В. – М.: Радио и связь. - 1983.

60. ETSI 123.401. General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access.

61. Friedhelm Hillebrand. GSM and UMTS: The Creation of Global Mobile Communication //.Wiley ,2002 -P .371.

62. Generation Services using UMTS in a Converging Network Environment.//UMTS Forum, 2002 . – Report 14.

63. Maja Matijasevic, Lea Skorin-Kapov, Miran Mosmondor. Application-level QoS Negotiation and Signaling for Advanced Multimedia Services in the IMS.// IEEE Communications Magazine - July 2007, p. 25-33.

64. Plane Siddhartha Pandey, Vikram Jain, Debabrata Das Vincent Planat, Ragupathi Periannan. Performance Study of IMS Signaling.// IEEE Xplore Communications Magazine - July 2006, p. 123-135.

65. Основы передачи голосовых данных по сетям IP, 2-е изд.: Пер. с. англ. – М .: ООО «И.Д. Вильямс», 2007.-400с.:ил. – Парал. тит. анагл. УДК. 681.3.07/ ISBN 978-5-8459-1281-7.

66. Руководство по технологиям объеденeнных сетей, 4-е издание.: Пер. с анг.: - М.: Издательский дом «Вильямс», 2005. – 1040 с.: ил.-Парал. тит. англ. ISBN 5-8459-0787-X.

67. M. Handley, H. Schulzrinne, E. Schooler, J. Rosenberg. SIP: Session Initiation Protocol.//RFC 2543, March 1999.

68. Tanir Ozcelebi, Igor Radovanovich. Multimedia Adaptation with SIP Resource Availability Signalling in IMS network.// IEEE 2007 International Conference on Convergence Information Technology. P. 1714-1719.

69. 3GPP Charging principles. [Електронний ресурс]: http://www.scribd.com/doc/18148552/3GPP-Charging-Principles.

70. H. Hakala, L. Mattila, J-P. Koskinen. Diameter Credit-Control Application.//RFC 4006, August 2005.

71. Robert Lloyd-Evans. QoS in intergrated 3G networks.//Artech House mobile communication series,2002. ISBN 1-58053-351-5.

72. Priggouris, G. Hadjiefthymiades, S. Merakos L. Supporting IP QoS in the GPRS.// IEEE Network September/October.Vol.14

73. 3GPP Evolved Packet System (EPS); Evolved General Packet Radio Service (GPRS); Tunnelling Protocol for Control plane (GTPv2-C); Stage 3 (Release 8).// 3GPP TS 29.274, 3rd Genral Partneship Project.P.143.

74. UMTS, LTE, Policy and control over Gx refernce point.// 3GPP TS 29.212 version 8.5.0 Release 8. [Електронний ресурс]:

http://www.etsi.org/deliver/etsi_ts/129200_129299/129212/08.05.00_60/ts_129212v080500p.pdf.

75. M. Chang, Y. Lin, Wei-Zu Yang. Performance of hot billing mobile prepaid service.// Computer Networks, Vol. 36, No. 2-3, July 2001, P. 269—290.

76. M. Chang, W. Yang. Y. Lin. Performance of Service-Node-Based Mobile Prepaid Service.// IEEE Transactions on vehicular technology, Vol. 51, No. 3, May 2002. P. 597—612.

77. IP Miltimedia Subsystem (IMS); Stage 2.// 3GPP TS 23.228., Release 10 SP 2010-11. P. 252.

78. User equipment time stamp for offline charging in IMS Networks.// United States Patant Application Publication. – Pub. No.: US 2008/0159499 A1. –Jul,2008.

79. Methoud for flexible configuration charging modes in IMS systems. // United States Patant Application Publication. – Pub. No.: US 2008/0195535 A1. – Aug. ,2008.

80. Flexible charging mechanism for IP Multimedia service.// United States Patant Application Publication. – Pub. No.: US 2008/0126230 A1.– May 2008.

81. Charging control in IP Multimedia susbystmes.//United States Patant Application Puplication. – Pub. No.: US 2008/0056304 A1. –Mar, 2008.

82. Methoud for flexible charging of IP multimedia communication sessions, telecommunication systems and network elements for applying such a method.// United States Patant Application Puplication. – Pub. No.: US 7,145,994 B2. – Dec. 2006.

83. Methoud for flexible charging in LTE/EPC communication network.//United States Patant Application Puplication. – Pub. No.: US 2009/0264097 A1. –Oct. 2009 .

84. Methoud and system for charging correlation.// United States Patant Application Puplication. – Pub. No.: US 2009/0089208 A1. – Apr. 2009.

85. Charging for roaming users IMS networksn. // United States Patant Application Puplication. – Pub. No.: US 2009/0088129 A1. – Apr. 2009.

86. Charging split negotiation in IMS networks.// United States Patant Application Puplication. – Pub. No.: US 2009/00116627 A1. – May 2009.

87. M. Chang, Y. Lin, Wei-Zu Yang. Performance of hot billing mobile prepaid service.// Computer Networks, Vol. 36, No. 2-3, July 2001. P. 269—290.

88. Tsybkov B. Self-similar Processes in Communications Networks. / Tsybkov B., Georganas N. D. – IEEE Trns. On Information Theory. – 1998. - v. 44, №5. - P. 1713-1725.

89. Universal Mobile Telecommunications System (UMTS).// 3GPP TS 23.401. V. 9.2.0, Release 9.

90. C.Gessner. UMTS Long Term Evolution (LTE) Technology Introduction.// Rohde & Schwarz . – 2008. –P.32.

91. M. Handley. SIP: Session Initiation Protocol.//RFC: 2543. – ACIRI Category: Standards Track. – March 1999. – p.40.

92. Bradai A. Cellular software defined networking: a framework / A. Bradai, K. Singh, T. Ahmed, T. Rasheed // IEEE Communications Magazine. – 2015. – Vol. 53, No. 6. – P. 36-43.

93. Basta A. Applying NFV and SDN to LTE mobile core gateways, the functions placement problem / A. Basta, W. Kellerer, M. Hoffmann, H. Morper et al. //

4th workshop on All things cellular: operations, applications, & challenges. – Chicago, USA, 2014.– P. 33-38.

94. Yousaf F. Z. SoftEPC – Dynamic instantiation of mo-bile core network entities for efficient resource utilization / F. Z. Yousaf, J. Lessmann, P. Loureiro, S. Schmid // 2013 IEEE International Conference on Communications (ICC). – Bu-dapest, Hungary, 2013.– P. 3602-3606.

95. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2014–2019 White Paper [Online]. – Available at: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

96. Hawilo H. NFV: state of the art, challenges, and imple-mentation in next generation mobile networks (vEPC) / H. Hawilo, A. Shami, M. Mirahmadi, R. Asal // IEEE Network. – 2014. – Vol. 28, No. 6. – P. 18-26.

97. Liquid Core [Online]. – Available at: http://networks.nokia.com/portfolio/liquidnet/liquidcore.

98. Soares J. Cloud4nfv: A platform for virtual network functions / J. Soares, M. Dias, J. Carapinha, B. Parreira, S. Sargento // 2014 IEEE 3rd International Conference on Cloud Networking (CloudNet). – Luxembourg, 2014. – P. 288-293.

99. Moens H. VNF-P: A model for efficient placement of virtualized network functions / H. Moens, F. De Turck // 10th International Conference on Network and Service Manage-ment. – Rio de Janeiro, 2014. – P. 418-423.

100. Rajan A.S. Understanding the bottlenecks in virtualiz-ing cellular core network functions / A.S. Rajan, S. Gobriel, C. Maciocco, K.B. Ramia et al. // 2015 IEEE International Workshop on Local and Metropolitan Area Networks (LANMAN). – Beijing, 2015.– P. 1-6.

101. Ferrer Riera J. Virtual network function scheduling: Concept and challenges / J. Ferrer Riera, E. Escalona, J. Ba-talle, E. Grasa, J. A. Garcia-Espin // 2014 International Con-ference on Smart Communications in Network Technologies (SaCoNeT). – Vilanova i la Geltru, 2014. – pp. 1-5.

102. Jennings B. Resource management in clouds: Survey and research challenges / B Jennings, R Stadler // Journal of Network and Systems Management. – 2014. – P. 1-53.

103. Fischer A. Virtual Network Embedding: A Survey / A. Fischer, J. Botero, M. Till Beck, H. de Meer and X. Hesselbach // IEEE Communications Surveys & Tutorials. – 2013. – Vol. 15, No. 4. – P. 1888-1906.

104. Baumgartner A. Mobile core network virtualization: A model for combined virtual core network function placement and topology optimization / A. Baumgartner, V.S. Reddy, T. Bauschert // 2015 1st IEEE Conference on Network Softwari-zation (NetSoft). – London, 2015. – P. 1-9.

105. Mehraghdam S. Specifying and placing chains of vir-tual network functions / S. Mehraghdam, M. Keller, H. Karl // 2014 IEEE 3rd International Conference on Cloud Network-ing (CloudNet). – Luxembourg, 2014. – P. 7-13.

106. Lischka J. A virtual network mapping algo-rithm based on subgraph isomorphism detection / J. Lischka, H. Karl // Proceedings of the 1st ACM workshop on Virtualized infra-structure systems and architectures. – 2009. – P. 81-88.

107. Heller B. The controller placement problem / B. Heller, R. Sherwood, N. McKeown // ACM HotSDN. – Helsinki, Fin-land, 2012. – P. 1-6.

108. Hock D. Pareto-optimal resilient controller placement in SDN-based core networks / D. Hock, M. Hartmann, S. Gebert, M. Jarschel et al. // 25th International Teletraffic Congress (ITC). – Shanghai, 2013. – P. 1-9.

109. Lange S. Heuristic Approaches to the Con-troller Placement Problem in Large Scale SDN Net-works / S. Lange, S. Gebert, T. Zinner, P. Tran-Gia et al. // IEEE Transactions on Network and Service Management. – 2015. – Vol. 12, No. 1. – P. 4-17.

110. Luizelli M. C. Piecing together the NFV provisioning puzzle: Efficient placement and chaining of virtual network functions / M. C. Luizelli, L. R. Bays, L. S. Buriol, M. P. Barcellos et al. // 2015 IFIP/IEEE International Symposium on Integrated Network Management. – Ottawa, 2015. – P. 98-106.

111. Albreem M. A. M. 5G wireless communication systems: Vision and challenges / M. A. M. Albreem // 2015 International Conference on Computer, Communications, and Control Technology (I4CT). – Kuching, 2015. – P. 493-497.

112. Agyapong P. K. Design considerations for a 5G network architecture / P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, A. Benjebbour // IEEE Communications Magazine. – 2014. – Vol. 52, No. 11. – P. 65-75.

113. Gohil A. 5G technology of mobile communication: A survey / A. Gohil, H. Modi, S. K. Patel // 2013 International Conference on Intelligent Systems and Signal Processing (ISSP). – Gujarat, 2013. – P. 288-292.

114. Mijumbi R. Network Function Virtualization: State-of-the-art and Research Challenges / R. Mijumbi, J. Serrat, J. Gorricho, N. Bouten et al. // IEEE Communications Surveys & Tutorials. – 2015. – Vol. 18, No. 1. – P. 236-262.

115. Network Functions Virtualisation (NFV); Architectural Framework [Online]. – Available at: http://www.etsi.org/deliver/etsi_gs/nfv/001_099/002/01.01.01_60/gs_nfv002v010101p.pdf .

116. Li Y. Software-Defined Network Function Virtualization: A Survey / Y. Li, M. Chen // IEEE Access. – 2015. – Vol. 3. – P. 2542-25.

Математичні методи аналізу та керування телекомунікаційними мережами

Монографія

Глоба Л.С., Дяденко О.М., Пилипенко А.Ю., Скулиш М.А.

Відповідальний редактор: д.т.н. Стрижак О.Є.

Підписано до друку 12.10.2017. Формат 60x84$^{1/16}$
Папір офс. 80 г/м^2. Друк цифровий. Ум. друк. арк. 14
Наклад 300 прим.

Видавництво Інститут обдарованої дитини НАПН України
вул. Артема, 52-Д, м. Київ, 04053
тел./факс.: (044) 481-27-27
Свідоцтво про внесення до Державного реєстру
Серія ДК №3366 від 13.01.2009 р.