УДК 621.396 DOI: 10.31673/2412-4338.2020.031526

Максимюк Т.А., Шубин Б.П., Андрущак В.С., Думич С.С., Климаш М.М.
Національний університет "Львівська політехніка"

МЕТОД ІНТЕЛЕКТУАЛЬНОГО УПРАВЛІННЯ КОНТЕНТОМ У МЕРЕЖАХ МОБІЛЬНОГО ЗВ'ЯЗКУ

Анотація: *З появою мереж мобільного зв'язку 5G, користувачі очікують абсолютно нових послуг з багатим мультимедійним контентом. Забезпечення таких вимог в рамках фізичних обмежень доступності інфраструктури та радіочастотного ресурсу є складним завданням, яке заважає операторам правильно масштабувати свої послуги. На сьогодні, оператори мобільного зв'язку змушені інвестувати значні суми у свою інфраструктуру, щоб максимізувати пропускну здатність за рахунок ущільнення мережної інфраструктури та більшого коефіцієнту повторного використання частот. Проте, це призводить до зростання вартості управління мережею для оператора, і відповідно, вищій вартості послуг для кінцевих користувачів. На сьогоднішній час, із розвитком штучного інтелекту, хмарних та крайових обчислень, мережа стає більш гнучкою, що відкриває багато можливостей операторам для підвищення продуктивності та якості функціонування мережі. В даній статті запропоновано новий підхід до управління контентом у мережі мобільного зв'язку за допомогою прогнозованого кешування громіздкого мультимедійного контенту на граничних серверах. Запропонований метод базується на прогнозуванні популярності контенту за допомогою рекурентних нейронних мереж. Це дає змогу гарантувати доставку необхідного контенту у безпосередній близькості до цільового користувача до того часу, коли він буде необхідний. Результати моделювання показують, що запропонована модель забезпечує точність понад 90% як в межах доби, так і протягом кількох діб. Крім того, розроблено метод персоналізованого кешування контенту з використанням апаратних ресурсів пристроїв користувачів, який працює на основі аналізу вподобань користувача та забезпечує проактивну доставку контенту. Таким чином, підхід до управління контентом, запропонований у статті дає змогу підвищити загальну продуктивність мережі за допомогою попереднього кешування контенту в періоди низького завантаження мережі. Додатковою перевагою є те, що проактивне кешування дає змогу завантажувати контент у найкращій якості, незалежно від перевантажень та вузьких місць у мережі.*

Ключові слова: *5G, кешування контенту, граничні обчислення, хмарні обчислення рекурентні нейронні мережі.*

Maksymyuk T., Shubyn B., Andrushchak V., Dumych S., Klymash M.
Lviv Polytechnic National University

METHOD OF INTELLIGENT CONTENT MANAGEMENT IN MOBILE NETWORKS

Abstract: With the advent of 5G, the market has been expecting the immersive user experience with rich multimedia content. Meeting such requirements within the physical constraints of limited spectrum and infrastructure availability is a challenging task, which prevents operators to scale their services properly. Currently, mobile operators are forced to invest large amount of money in their infrastructure, in order to maximize the capacity by network densification and higher frequency reuse factors. The dark side of such trend is that infrastructure becomes more expensive, spectrum price is getting higher and total cost of ownership for operator increases drastically. Nowadays, with the rise of artificial intelligence, cloud and edge computing the network becomes more flexible that opens many opportunities to enhance the performance and user experience. In this paper, we propose a new approach for content management in mobile network by using predictive caching of rich multimedia content in edge servers. Proposed approach is based on the content popularity prediction by using recurrent neural networks, that allows to deliver corresponding content in the close proximity to the target end users by the time it will be needed. Simulation results show that the proposed model is more than 90% accurate for both daily and weekly timeframes. Furthermore, we develop a method of personalized content caching in user devices based on their subscriptions and preferences, to make sure that user will have the best experience. Proposed approach for

content management allows to improve the overall network performance by proactive content caching during the time of low network load. Moreover, the proactive caching allows to download the content in the best quality, regardless of the network congestions and bottlenecks.

Keywords: 5G, content caching, edge computing, cloud computing, recurrent neural networks.

Максимюк Т.А., Шубин Б.П., Андрущак В.С., Думич С.С., Климаш М.М.
Национальный университет «Львовская политехника»

МЕТОД ИНТЕЛЛЕКТУАЛЬНОГО УПРАВЛЕНИЯ КОНТЕНТОМ В СЕТЯХ МОБИЛЬНОЙ СВЯЗИ

Аннотация: *С появлением сетей мобильной связи 5G пользователи ожидают абсолютно новых услуг с богатым мультимедийным контентом. Обеспечение таких требований в рамках физических ограничений доступности инфраструктуры и радиочастотного ресурса является сложной задачей, которая мешает операторам правильно масштабировать свои услуги. На сегодня, операторы мобильной связи вынуждены инвестировать значительные суммы в свою инфраструктуру, чтобы максимизировать пропускную способность за счет уплотнения сетевой инфраструктуры и большого коэффициента повторного использования частот. Однако, это приводит к росту стоимости управления сетью оператора и, соответственно, высокой стоимости услуг для конечных пользователей. На сегодняшнее время, с развитием искусственного интеллекта, облачных и граничных вычислений, сеть становится более гибкой, что открывает много возможностей операторам для повышения производительности и качества функционирования сети. В данной статье предложен новый подход к управлению контентом в сети мобильной связи с помощью прогнозируемого кэширования громоздкого мультимедийного контента на граничных серверах. Предложенный метод основан на прогнозировании популярности контента с помощью рекуррентных нейронных сетей. Это позволяет гарантировать доставку необходимого контента в непосредственной близости от целевого пользователя к тому времени, когда он будет необходим. Результаты моделирования показывают, что предложенная модель обеспечивает точность более 90% как в пределах суток, так и в течение нескольких суток. Кроме того, разработан метод персонализированного кэширования контента с использованием аппаратных ресурсов устройств пользователей, который работает на основе анализа предпочтений пользователя и обеспечивает проактивную доставку контента. Таким образом, подход к управлению контентом, предложенный в статье позволяет повысить общую производительность сети с помощью предварительного кэширования контента в периоды низкой загрузки сети. Дополнительным преимуществом является то, что проактивное кэширование позволяет загружать контент в наилучшем качестве, независимо от перегрузок и узких мест в сети.*

Ключевые слова: *5G, кэширование контента, граничных вычисления, облачные вычисления рекуррентные нейронные сети.*

1. Вступ

Мобільний зв’язок став важливою частиною людського життя. Зростання кількості мобільних пристроїв було експоненціальним протягом останніх двох десятиліть, де новітні сервіси та програмні продукти відіграють роль каталізатора. Ця зміна не тільки призвела до необхідності збільшення пропускної здатності в мережі, але й вимагає тісної інтеграції різноманітних мережних технологій. Сучасні мережі 5G істотно розширюють сфери застосування безпровідного зв’язку, створюючи нові можливості для бізнесу, нових послуг та інновацій. В останні роки спостерігається стрімке зростання кількості розумних пристроїв у мережах мобільного зв’язку, що у поєднанні із розвитком технологій 5G та Інтернету речей (IoT) суттєво розширює спектр мобільних додатків та мультимедійних послуг [1, 2].

Більшість таких послуг значною мірою потребують великої швидкості передавання даних та низької затримки, що спричиняє ряд проблем для мереж мобільного зв’язку. Централізована архітектура мереж мобільного зв’язку та обмежені радіочастотні ресурси не дають змогу операторам у повній мірі задовольнити зростаючі обсяги трафіку [3]. Тому,

потрібна нова парадигма функціонування мережі мобільного зв'язку, яка дасть змогу вирішити завдання масової доставки контенту до кінцевих користувачів.

Перспективним напрямком вдосконалення мережної архітектури є використання контекстно-орієнтованої архітектури мережі на основі технології мобільних граничних обчислень (MEC – Mobile edge Computing). Такий підхід забезпечує можливість кешування контенту на граничних серверах, які розташовані у безпосередній близькості до кінцевих користувачів, що, в свою чергу, дає змогу знизити затримку надання послуг, а також оптимізувати процес передавання трафіку в мережі вцілому. Для ефективного впровадження кешування контенту використовується гібридна архітектура, яка використовує переваги граничних та хмарних обчислень [4]. При цьому, важливо враховувати фізичний рівень, який є визначальним з точки зору підвищення спектральної ефективності в мережах 5G за рахунок кешування контенту [5]. Враховуючи динаміку мобільних користувачів, соціальні та економічні аспекти теж істотно впливають на обсяги трафіку та доставку контенту в мережах 5G [6]. Зокрема, в ряді робіт було запропоновано використання соціальних характеристик кінцевих користувачів для підтримки обміну контенту між кеш-пам'яттю мобільних пристроїв [7], а також економічних підходів для стимулювання постачальників послуг до кешування з метою ефективного розповсюдження популярного контенту [8].

Розвиток технологій програмно-конфігурованих мереж (SDN – Software Defined Networks) та віртуалізації мережевих функцій (NFV – Network Functions Virtualization) відкриває нові можливості для інтеграції мереж мобільного зв'язку та хмарних технологій [9]. Така інтеграція дає змогу оптимізувати процес кешування контенту в мережах 5G та підвищити ефективність послуг, які потребують великих обчислювальних ресурсів та своєчасної доставки контенту.

Однак покращенням можливостей кешування, за рахунок обчислювальних ресурсів, часто нехтують. Наприклад, у додатках доповненої реальності загальноприйнятою практикою є вилучення деяких ключових функцій із спочатку захоплених відеозаписів, щоб заощадити ресурси на кешування та передавання [10]. Тому важливо використовувати обчислювальні ресурси, для того щоб зменшити навантаження на ресурси кешування для вузлів обмеженою ємністю накопичувачів. Незважаючи на те, що кешування на граничних серверах пропонує контент для користувачів у безпосередній близькості, постійно зростаюча кількість мобільних пристроїв може сильно вплинути на стратегії кешування та ускладнити процес доставки контенту. Тому, прогнозування мобільності абонентів у мережах 5G із високою густиною базових станцій, під час запитів на контент великого розміру може суттєво підвищити спектральну ефективність мережі 5G за рахунок оптимального кешування контенту в пам'яті усіх базових станцій вздовж шляху користувача. Таким чином, користувачі зможуть отримувати контент з мінімальною затримкою у потрібний момент часу [11].

2. Проблематика та переваги кешування контенту в мережах мобільного зв'язку

Кешування контенту на граничних вузлах дає змогу знизити обсяги трафіку трафіку на віддаленому вузлі до 35%. Проте, обсяги пам'яті, яких потребує сучасний контент (4K відео, VR360 та ін.) є значно більшими від можливостей кешування граничних серверів. Тому, визначальним аспектом постає пошук оптимального контенту для кешування, з метою забезпечення максимальної вигоди від кешування. Таким чином, популярність контенту є основним критерієм для прийняття рішення про його кешування. Проте, доволі складно визначити фактори, які безпосередньо впливають на популярність контенту.

Одним із найпростіших рішень є використання інформації користувачів для виявлення контексту. Проте, такий метод обмежується правилами конфіденційності персональних даних користувачів, що може спричинити ряд законодавчих обмежень в залежності від країни в якій працює мережа. Тому, більш доцільним технічним рішенням є прогнозування на рівні базових станцій та граничних серверів. В такому випадку, має місце врахування територіального та функціонального розділення на житлові зони, ділові райони, туристичні

та рекреаційні зони, у яких користувачі мають різні інтереси. Реальні дослідження показують, що розподіл популярності контенту навіть для сусідніх точок доступу Wi-Fi або базових станцій 5G відрізняється. Проте, існуючі схеми кешування контенту не враховують таких незначних відмінностей.

Ефективність кешування на основі граничних серверів у значній мірі залежить від кількості користувачів, які запитують однаковий контент. Це характерно для деяких сценаріїв з високою густиною абонентів та схожими запитами користувачів, таких як аеропорти, залізничні вокзали, стадіони, тощо. Ймовірність кешування контенту в гетерогенній мережі залежить від коефіцієнта його подібності та кількості користувачів у цільовій зоні обслуговування, як показано на рис. 1. При цьому, вплив кількості абонентів є менш відчутним, ніж вплив коефіцієнта подібності контенту.

При достатньо високому коефіцієнті подібності контенту та великій кількості користувачів ймовірність кешування різко зростає. Крім того, експериментально визначено, що при подібності контенту вище 80%, ймовірність кешування різко зростає незалежно від кількості пристроїв. При коефіцієнті подібності контенту меншому, ніж 50% ймовірність кешування істотно не змінюється, навіть при зростанні кількості пристроїв [12].

Окрім кешування, в мережах мобільного зв'язку також використовується технологія багатоадресного передавання (мультикастинг), яка була включена в специфікації 3GPP. Мультикастинг є ефективним під час подій з великою концентрацією абонентів, які мають спільні інтереси, наприклад, під час спортивних ігор, концертів та публічних демонстрацій із десятками тисяч відвідувачів. В такому випадку контент може передаватись до великої кількості користувачів використовуючи спільні радіочастотні ресурси. Це також характерно для різноманітних сервісів, таких як платформи соціальних мереж (Tweeter, Facebook, тощо).

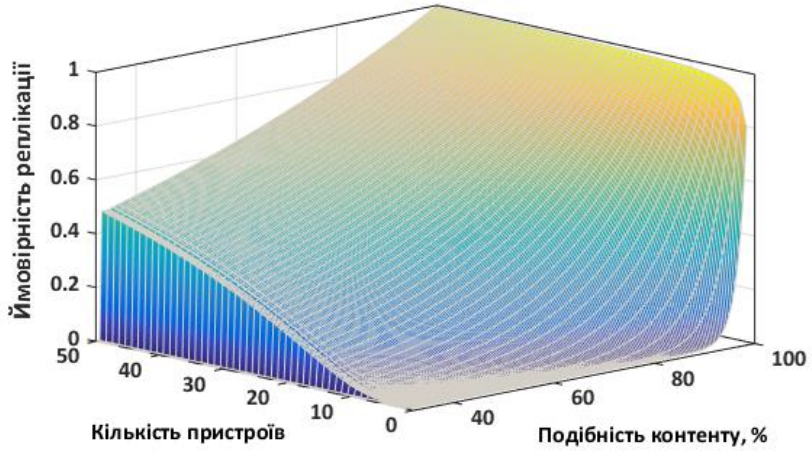


Рис. 1. Залежність ймовірності кешування контенту від кількості користувачів та коефіцієнта подібності контенту

Таким чином, можна зробити висновок, що тип контенту, має важливе значення з точки зору його кешування. Зокрема, хмарні сервіси мають найвищу ймовірність кешування, за рахунок контекстної орієнтованості хмарних та граничних систем. Ймовірність того, що користувачі в одні зоні покриття будуть використовувати один і той же сервіс (сервіс прогнозу погоди, інформаційні системи громадського транспорту, тощо.) є достатньо високою. Веб-сервіси також характерні високою ймовірністю кешування контенту, оскільки дуже ймовірно, що користувачі будуть звертатись до одних і тих самих Інтернет ресурсів. В свою чергу, мультимедійні сервіси характерні найнижчою ймовірністю кешування, оскільки є багато копій одного контенту, які несуттєво відрізняються для кінцевих користувачів (назви файлів, розмір, формат, тощо), але вважаються повністю різними для обчислювальної системи. Проте, незважаючи на складність, кешування мультимедійного контенту в мережі є найбільш перспективним з точки зору ефективного використання пропускної здатності [12].

Тому, для подальшого вдосконалення процесу передавання контенту в мережах мобільного зв'язку, важливо розробити модель прогнозування популярності контенту з використанням інформації про його просторово-часову актуальність.

3. Управління контентом у мережах мобільного зв'язку на основі методів штучного інтелекту

3.1. Загальна архітектура мережі

Не зважаючи на те, що поведінка абонентів є дуже передбачуваною та корельованою, прогнозувати точний час їх запитів на певний контент є надзвичайно складним завданням. Враховуючи низку випадкових соціальних факторів, які можуть вплинути на прийняття рішень користувачів стосовно інформаційних сервісів та просторових переміщень. Тому, в даній статті пропонується макромодель прогнозування популярності контенту на рівні базових станцій мережі мобільного зв'язку. Дана модель аналізує часовий ряд кількості запитів на певний тип контенту з використанням методів штучного інтелекту.

Використання методів машинного навчання та штучного інтелекту дає змогу використовувати перевагу зворотного зв'язку та аналітики даних для оптимізації доставки контенту в мережах 5G [13]. Із збільшенням обсягів інформації критично важливою є здатність швидко та точно прогнозувати параметри мережі та проактивно реагувати на зміни шляхом оновлення конфігурації мережної інфраструктури. Зокрема, важливими аспектами є визначення найбільш популярного контенту, вибір оптимальної базової станції для кешування та багато інших параметрів. Це в свою чергу потребує наявності системи моніторингу трафіку, аналізу контенту та визначення трендів.

Відповідно, наявність інфраструктури хмарних та граничних обчислень, для передавання, збереження, систематизації та аналізу великих обсягів даних, є важливим завданням для автоматизації процесу кешування контенту у мережах мобільного зв'язку 5G.

Технологія хмарних обчислень забезпечує зручний доступ до спільного пулу гнучко конфігурованих обчислювальних ресурсів, таких як мережі, сервери, сховища даних, програмні додатки, які можна швидко розгорнути та налаштувати з мінімальними зусиллями. За рахунок перенесення основних обчислювальних функцій мережі 5G у центри обробки даних, оператори мають змогу підвищити гнучкість розгортання та конфігурації базових станцій, що дає змогу їм забезпечити більш ефективну модель надання інформаційних послуг та ведення бізнесу. Таким чином, оператори можуть легко задовольнити потреби ринку, що швидко змінюються, і забезпечувати якісний зв'язок для своїх абонентів [14] .

Для запропонованої архітектури мережі 5G використовується модель «інфраструктура як сервіс» (IaaS – Infrastructure as a Service) постачальника хмарних послуг Google Cloud Platform (GCP). Дана модель включає в собі проміжне програмне забезпечення, мережу, процесори, графічні процесори, оперативну пам'ять, зовнішнє сховище даних, сервери та образи базових операційних систем. Операторам мобільного зв'язку в такому випадку достатньо забезпечити об'єднання усіх вищезгаданих елементів в одну цілісну інфраструктуру, що задовольняє вимоги мережі 5G.

Варто зазначити, що в масштабах держави використання єдиного централізованого хмарного сервісу є доволі складним, оскільки сучасні сервіси в мережі 5G потребують низької затримки та високої пропускної здатності. Тому, для розвантаження мережі пропонується модель з використанням розподілених серверів мобільних граничних обчислень (MEC – Mobile Edge Computing), яка передбачає розміщення обчислювальних ресурсів у радіусі географічної близькості до кінцевих користувачів. Такий підхід дає змогу підвищити ефективність функціонування мережі 5G, особливо з точки зору надання контекстно-орієнтованих сервісів. Це, в свою чергу, дає змогу знизити навантаження на опорну мережну інфраструктуру.

Використання запропонованої архітектури дає змогу підвищити гнучкість управління інфраструктурою мережі 5G з точки зору розподілу радіочастотних ресурсів, контролю політик якості обслуговування, координованого кешування контенту, а також багатьох інших параметрів мережі.

3.2. Побудова хмарної інфраструктури для системи управління контентом у мережах мобільного зв'язку

Для побудови хмарної інфраструктури використано сервіс GCP Virtual Private Cloud (VPC), який забезпечує функціонал для управління хмарними ресурсами. VPC – це аналог фізичної мережі у віртуалізованому представленні в середовищі Google Cloud. Зв'язок між всіма хмарними ресурсами реалізований через VPC мережу.

Середовищем для розробки та роботи системи обрано сервіс Google Kubernetes Engine (GKE). GKE забезпечує кероване середовище для розгортання, керування та масштабування контейнеризованих програм за допомогою інфраструктури Google. GKE середовище складається з множини віртуальних машин сервісу Google Compute Engine, згрупованих в кластер. Кожен кластер керується системою менеджменту кластерів Kubernetes. За допомогою ресурсів та команд Kubernetes можна розгортати та керувати програмами, виконувати завдання адміністрування, встановлювати політики та контролювати стан розгорнутих застосувань. Kubernetes базується на аналогічній структурі, яка використовується популярними службами Google, що надає йому ряд переваг, таких як автоматичне керування, моніторинг та контроль активності для контейнерів з додатками, автоматичне масштабування, впровадження оновлень тощо. Загальна архітектура кластера GKE складається з як мінімум одного головного вузла (master node) та з кількох робочих вузлів (worker nodes).

Головний вузол керує площиною управління Kubernetes, включаючи сервер Kubernetes API, планувальник та контролери основних ресурсів. Життєвим циклом головного вузла керує GKE під час створення або видалення кластера. Цей цикл включає в себе оновлення версії Kubernetes, що працює на головному вузлі, яке GKE виконує автоматично, або вручну за запитом. Всі взаємодії з кластером проводяться через запити Kubernetes API напряму через HTTP/gRPC, або за посередництвом клієнта командної стрічки kubectl чи за допопогою інтерфейсу користувача в хмарній консолі GCP. Головний вузол також відповідає за розподіл завдань та функцій між робочими вузлами кластера. Як правило, кластер має один або декілька робочих вузлів, які виконують контейнеризовані програми та функції. Робочими вузлами керує головний вузол, який отримує оновлення про стан кожного вузла. Це включає як безпосередній контроль над життєвим циклом вузла, так і автоматичні оновлення на усіх вузлах кластера. Робочий вузол запускає служби, необхідні для підтримки контейнерів Docker, які складають навантаження кластера. До них відносяться Docker runtime та агент вузла Kubernetes (kubelet), який спілкується з головним вузлом і відповідає за запуск та підтримку контейнерів Docker, запланованих на відповідному робочому вузлі. При флуктуації робочого навантаження, яке пов'язано із кількістю активних абонентів кластер автоматично змінює кількість вузлів в пулі, базуючись на поточних потребах.

Оскільки реалізація інтелектуального управління контентом у мережі 5G передбачає використання методів машинного навчання та штучного інтелекту, необхідно забезпечити обчислювальну інфраструктуру з графічними та тензорними процесорами для підтримки відповідних обчислень. Архітектура GKE дає можливість створювати пули вузлів, які оснащені відповідним апаратними ресурсами (наприклад NVIDIA Tesla).

Важливим аспектом системи є системи збереження великих обсягів даних мереж 5G. Вибір варіантів зберігання даних програм, що працюють в Google Kubernetes Engine (GKE) залежить від їх гнучкості та простоти використання. Крім цього, Kubernetes надає абстракцію сховища даних, яка використовується для підримки функціонування кластера. Загальна архітектура побудованої системи управління мережами 5G представлена на рис. 2.

Для реалізації алгоритмів на основі нейронних мереж в запропонованій системі розгортається фреймворк TensorFlow, який складається із набору програмних бібліотек для машинного навчання та вирішення завдань побудови і тренування нейронної мережі. Основний інтерфейс для роботи з даною бібліотекою реалізований на мові Python. Процес функціонування мережі відбувається наступним чином. Кінцеві користувачі підключаються до базових станцій 5G, створюючи тим самим певний паттерн популярності контенту.

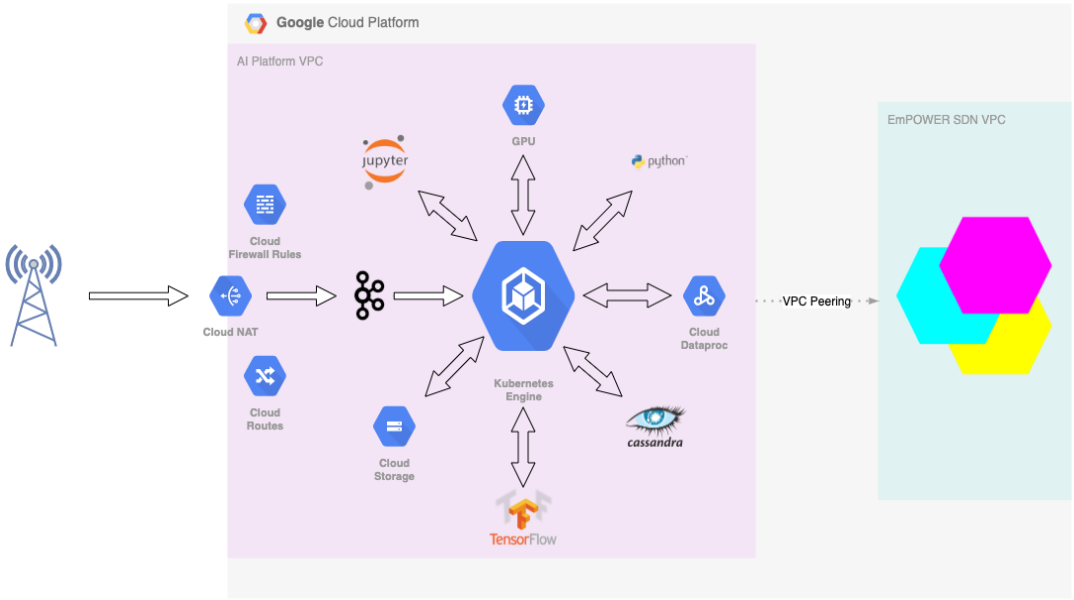


Рис.2. Загальна архітектура хмарної інфраструктури інтелектуальної системи кешування контенту

В залежності від розташування базової станції та типу контенту, здійснюється процедура кешування з метою зниження обсягів трафіку, які завантажуються базовою станцією з мережі Інтернет при обслуговуванні користувачів. Вся статистика при цьому записується у відповідні лог-файли, які перенаправляються на IP адресу інтелектуальної хмарної системи управління. Весь трафік, що поступає до Cloud NAT фільтрується за правилами мережевого екрану та правилом пересилання, і направляється до сервісу Apache Kafka в GKE кластері. Дані з базових станції перенаправляються в Dataproc кластер на обробку і після цього зберігаються в базі даних Cassandra. Відповідно, навчальні вибірки для нейронних мереж формуються безпосередньо із даних, які доступні в Cassandra. Навчені нейронні мережі використовуються для прогнозування популярності контенту на кожній базовій станції в мережі 5G, і безпосередньо передаються до контролера SDN (Software Defined Network) через інтерфейс VPC Peering для прийняття рішення стосовно кешування відповідного контенту [9], як показано на рис. 3.

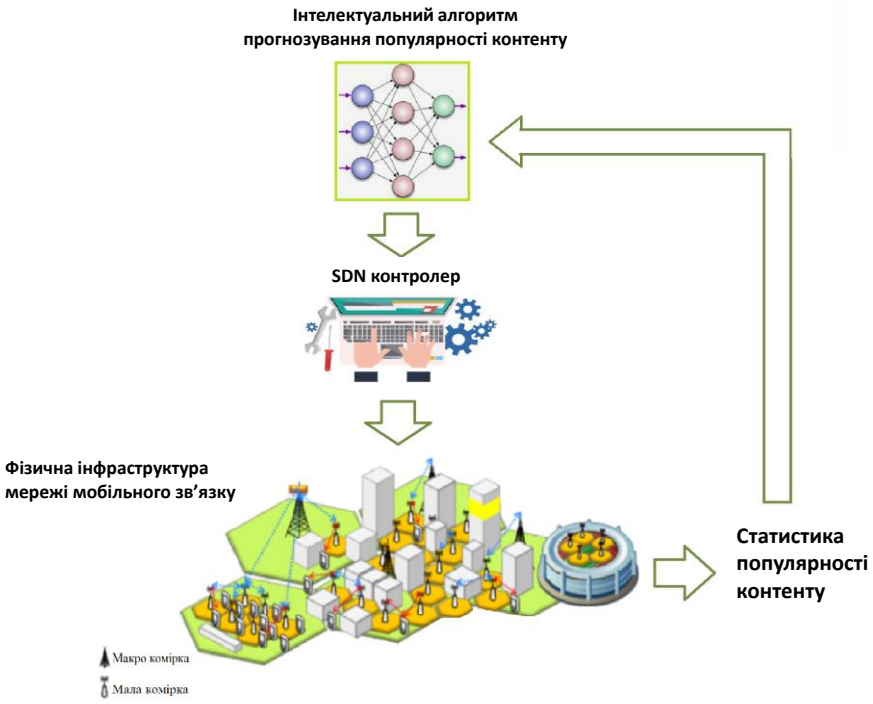


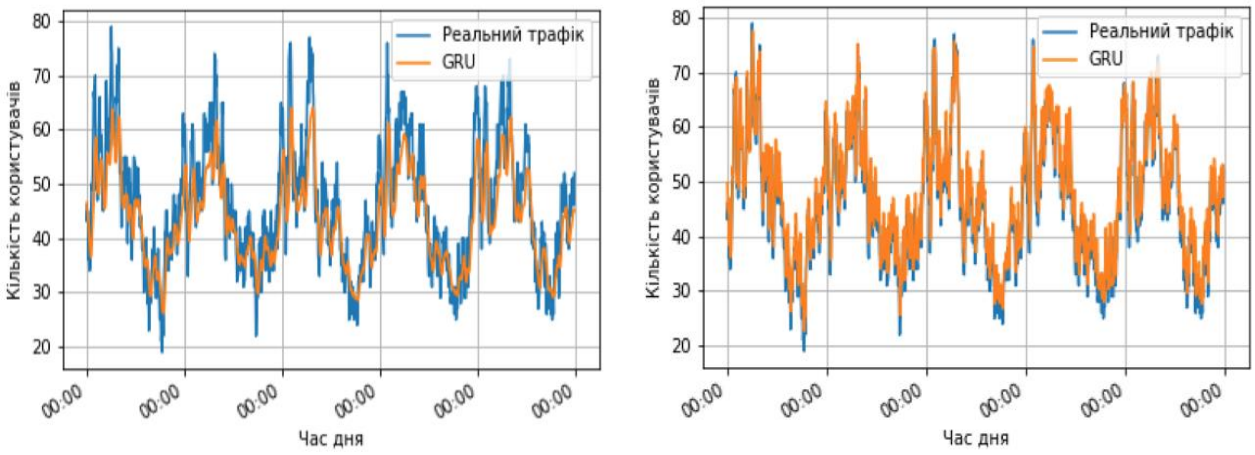
Рис. 3. Архітектура мережі 5G на основі технологій хмарних та граничних обчислень

Таким чином, запропонована архітектура мережі 5G на основі технологій хмарних та граничних обчислень дає змогу забезпечити ефективне кешування контенту у мережах мобільного зв'язку з гнучким управлінням на основі технологій SDN та методів штучного інтелекту.

3.3. Прогнозування популярності контенту на основі рекурентних нейронних мереж

Для проведення експерименту необхідно зібрати статистичну інформацію про кількість запитів контенту в залежності від часу. Для проведення експерименту згенеровано навчальну вибірку, яка імітує кількість запитів до інформаційного ресурсу для кожної комірки в залежності від часу доби. Для обробки такого типу даних створено модель рекурентної нейронної мережі на основі архітектури GRU, яка має доведену ефективність для розв'язку задач регресії [15].

При здійсненні експерименту було здійснено декілька моделювань, для того щоб визначити, яким чином розроблена модель може передбачити кількість запитів до певної категорії контенту в конкретній комірці протягом різних часових інтервалів та різної кількості епох. На рис. 4. представлено результати прогнозування кількості запитів до однієї категорії контенту протягом 5 робочих днів для 50 та 100 епох навчання рекурентної нейронної мережі GRU. При навчанні протягом 50 епох, точність прогнозування становила 76%, у той час як при навчанні протягом 100 епох – 91%. Варто зазначити, що подальше збільшення кількості епох не призводить до суттєвого покращення точності. Наприклад для 200 епох точність становить 92%, що є несуттєвим виграшем для вдвічі довшого періоду навчання моделі.



а) б)

Рис. 4. Результати прогнозування кількості запитів до однієї категорії контенту протягом 5 робочих днів для 50 та 100 епох

На рис. 5. представлено результати моделювання для однієї доби при 100 епохах навчання рекурентної нейронної мережі GRU. Як можна побачити з отриманих результатів, висока точність прогнозування зберігається в масштабі доби, що дає змогу кешувати контент на стороні базової станції враховуючи флуктуації протягом доби.

На основі запропонованої моделі реалізується стратегія кешування контенту в мережі 5G на кожному рівні агрегації трафіку. Наприклад, контент, який є поширеним у значній територіальній зоні, спочатку кешується на серверах вищого рівня агрегації, а потім за необхідності кешується у нижчих рівнях агрегації в залежності від локальних прогнозів моделі.

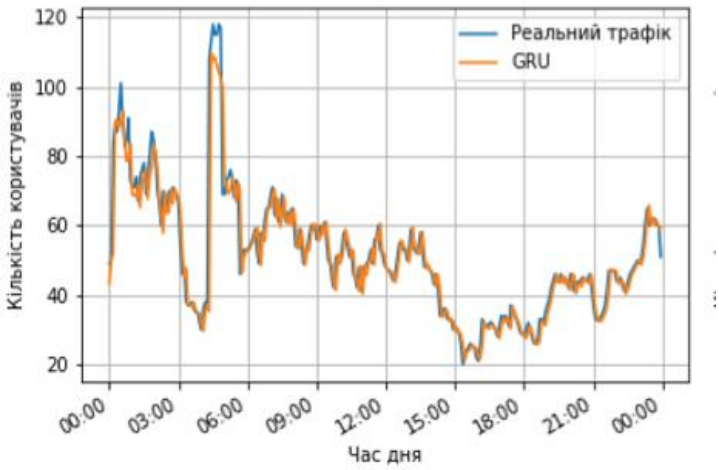


Рис. 5. Результати прогнозування кількості запитів до однієї категорії контенту протягом доби для 100 епох.

Така стратегія дає змогу забезпечити більш ефективне використання пропускної здатності як безпровідних каналів мережі 5G, так і каналів опорної транспортної інфраструктури. Крім того, запропонована система кешування може адаптивно змінювати частку кешованого контенту в залежності від поточних потреб трафіку у мережі 5G.

4. Метод персоналізованого кешування контенту на основі кінцевих пристроїв

Сучасні мобільні пристрої зазвичай мають вражаючі характеристики з точки зору обчислювальних ресурсів, постійної та оперативної пам’яті. Це дає їм можливість заздалегідь кешувати контент, який буде потрібен абоненту безпосередньо у пам’яті пристрою. Таким чином, у момент коли користувач почне завантажувати даний контент з мережі (наприклад, перегляд відео із сервісу YouTube), він одразу буде відтворюватись із локальної кеш пам'яті без затримок та з максимальною якістю. Схема функціонування персоналізованого кешування контенту представлена на рис. 6.

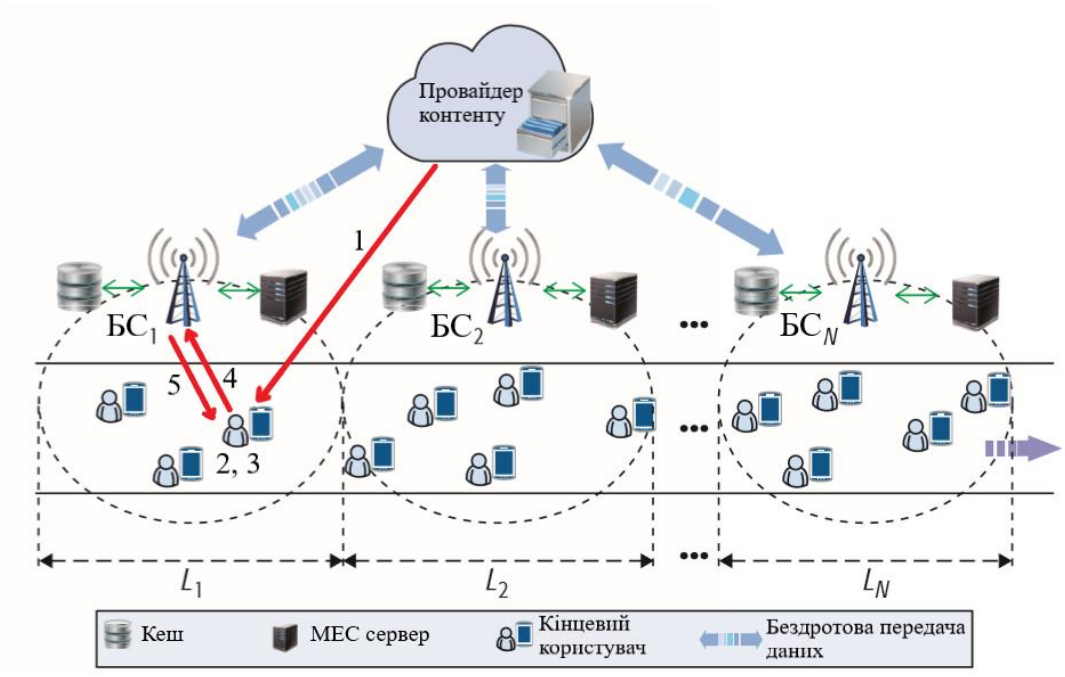


Рис. 6. Схема функціонування методу персоналізованого кешування контенту у мережах мобільного зв’язку

Крок 1. Клієнтський додаток Інтернет-сервісу (Youtube, Instagram, тощо) у пристрої користувача отримує сповіщення про новий актуальний контент для користувача на основі його підписок та вподобаних сторінок.

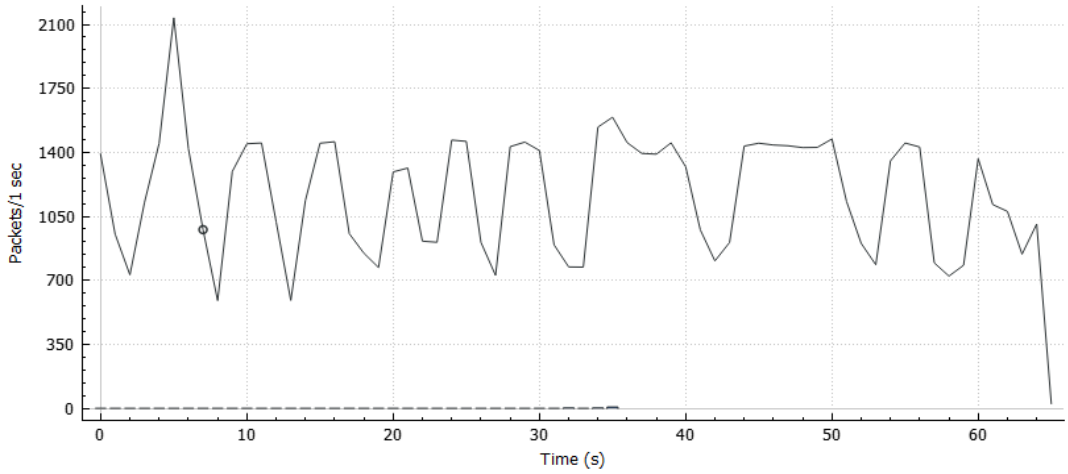
Крок 2. AI-агент, який знаходиться у пристрої користувача, прогнозує час, коли даний контент буде потрібен користувачеві на основі його попередньої статистики, часу доби, актуальності та приорітетності даного контенту.

Крок 3. Пристрій користувача аналізує свої поточні параметри заряду акумулятора, активність користувача у мережі, завантаження CPU/RAM, обсяг доступної пам'яті та доступні ресурси пропускної здатності поточного з'єднання з мережею Інтернет.

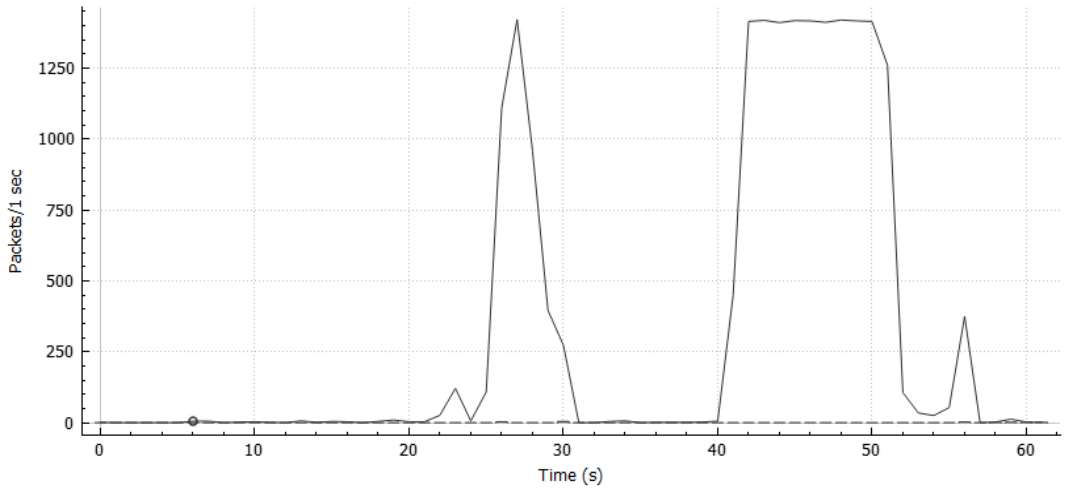
Крок 4. Пристрій користувача надсилає запит до оператора про необхідність попереднього кешування даного контенту в пам'яті телефону, edge-серверах чи базових станціях. При цьому надсилається інформацію про потенційне переміщення користувача протягом доби та очікуваний час, коли контент повинен бути доступний.

Крок 5. Оператор приймає рішення про найбільш оптимальну доставку даного контенту до користувача, враховуючи поточне навантаження на мережу, пріоритет користувача, який прописується на сервері HSS (Home Subscriber Server), а також важливість даного контенту, яка визначається з таблиці QCI (Quality Channel Identifier).

На рис. 7 показано експериментальні результати перехоплення мережевого трафіку при передаванні контенту в мережі мобільного зв'язку з використанням кешування (рис. 7.а) та звичайного завантаження контенту в режимі онлайн (рис.7.б).



а)



б)

Рис. 7. Результати перехоплення мережевого трафіку при передаванні контенту в мережі мобільного зв'язку звичайним методом – а) та методом з використанням персоналізованого кешування – б)

Дослідження проводилися шляхом завантаження відеофайлу із сервісу YouTube. Для імітації процесу кешування використано попередню буферизацію 50% відеофайлу з подальшим довантаженням безпосередньо після початку перегляду користувачем. Для аналізу трафіку використано програмне забезпечення Wireshark.

Як можна побачити із результатів експериментального дослідження, при використанні звичайного методу передавання (рис.7,а), спостерігається пилоподібний характер трафіку, який пов'язаний із періодичним завантаженням контенту. Крім того, інтенсивність трафіку не падає нижче ніж на 50% від пікових значень.

Натомість, при використанні персоналізованого кешування, є лише 2 окремих сплески довантаження контенту, які за інтенсивністю навіть не досягають пікових значень, що характерні для звичайного методу передавання контенту без кешування (рис. 7,б). Враховуючи потенційну кількість контенту, який може кешуватись безпосередньо у пристроях, можна припустити, що запропонований метод в перспективі дасть змогу суттєво розвантажити ресурси у мережах мобільного зв'язку.

Крім того, персоналізоване кешування контенту може ефективно доповнювати кешування в граничних вузлах вищих рівнів, даючи змогу балансувати обсяг кешованого контенту між пам'яттю кінцевих пристроїв та граничних серверів.

Висновки. В статті запропоновано метод кешування контенту в мережах мобільного зв'язку на основі технології граничних обчислень, яка передбачає розміщення обчислювальних ресурсів у радіусі географічної близькості до кінцевих користувачів. Запропонований підхід дає змогу підвищити ефективність функціонування мережі 5G з точки зору надання контекстно-орієнтованих сервісів та підвищити гнучкість управління ресурсами та процесом передавання контенту. Запропоновано макромодель прогнозування популярності контенту на рівні базових станцій мережі мобільного зв'язку на основі рекурентних нейронних мереж GRU. Результати моделювання показали, що запропонована модель дає змогу прогнозувати кількість запитів до контенту з точністю понад 90%, як в межах доби, так і протягом кількох діб. Крім того, запропоновано метод персоналізованого кешування контенту з використанням апаратних ресурсів пристроїв користувачів, який працює на основі аналізу вподобань користувача та забезпечує проактивну доставку контенту. Результати експериментальних досліджень показують, що запропонований метод суттєво знижує навантаження користувачів на мережу мобільного зв'язку при використанні сервісів з громіздким мультимедійним контентом.

References

1. Zhang Z. 6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies/ Zhang Z., Xiao Y., Ma Z., Xiao M., Ding Z., Lei X., Karagiannidis G., Fa P.// IEEE Vehicular Technology Magazine, Sep. 2019. – vol. 14 – №3 – P. 28-41.

2. Zhang L. 6G Visions: Mobile ultra-broadband, super internet-of-things, and artificial intelligence/ Zhang L., Liang Y., Niyato D. // China Communications, 2019 – vol. 16 – № 8 – P. 1-14.

3. Wang X., Chen M., Taleb T., Ksentini A., Leung V. Cache in the air: exploiting content caching and delivery techniques for 5G systems/Wang X., Chen M., Taleb T., Ksentini A., Leung V.// IEEE Communications Magazine, 2014 – vol. 52, № 2 – P. 131-139.

4. Tandon R. Harnessing cloud and edge synergies: toward an information theory of fog radio access networks / Tandon R. Simeone O.// IEEE Communications Magazine, 2016. – vol. 54 – № 8 – P.44-50.

5. Han W. PHY-caching in 5G wireless networks: design and analysis/ Han W., Liu A., Lau V. //IEEE Communications Magazine, 2016. – vol. 54 – № 8 – P. 30-36.

6. Wang X. Tag-assisted social-aware opportunistic device-to-device sharing for traffic offloading in mobile social networks / Wang X., Sheng Z., Yang S. Leung V //IEEE Wireless Communications, 2016, – vol. 23 – № 4 – P. 60-67.

7. Zhang X. Information Caching Strategy for Cyber Social Computing Based Wireless Networks/ Zhang X., Li Y., Zhang Y., Zhang J., Li H., Wang S., Wang D.// IEEE Transactions on Emerging Topics in Computing, 2017 – vol. 5 – № 3 – P. 391-402.

8. Hajimirsadeghi M. Joint Caching and Pricing Strategies for Popular Content in Information Centric Networks/ Hajimirsadeghi M., Mandayam N., Reznik A. // IEEE Journal on Selected Areas in Communications, 2017 – vol. 35 – № 3 – P. 654-667.

9. Ma L. An SDN/NFV based framework for management and deployment of service based 5G core network/ Ma L., Wen X., Wang L., Lu Z., Knopp R. // China Communications, Oct. 2018 – vol. 15 – №10 – P. 86-98.

10. Wang X. Cloud-assisted adaptive video streaming and social-aware video prefetching for mobile users / Wang X., Kwon T., Choi Y., Wang H., Liu J. // IEEE Wireless Communications, 2013. – vol. 20 – № 3 – P. 72-79.

11. Mahmood A. Mobility-aware edge caching for connected cars / Mahmood A., Casetti C., Chiasserini C., Giaccone P., Harri J.//12th Annual Conference on Wireless On-demand Network Systems and Services (WONS), (Cortina d'Ampezzo, 2016) – P. 1-8.

12. Jo M. Device-to-device-based heterogeneous radio access network architecture for mobile cloud computing / Jo M., Maksymyuk T., Strykhalyuk B., Cho C. //IEEE Wireless Communications, 2015. – vol. 22 – №. 3 – P. 50-58.

13. Letaief K. The Roadmap to 6G: AI Empowered Wireless Networks/ Letaief K., Chen W., Shi Y., Zhang J., Zhang Y.// IEEE Communications Magazine, 2019 – vol. 57 – № 8 – P. 84-90.

14. Song S. Clustered Virtualized Network Functions Resource Allocation based on Context-Aware Grouping in 5G Edge Networks / Song S., Lee C., Cho H., Lim G., Chung J.// IEEE Transactions on Mobile Computing, 2020. – vol. 19 – № 5 – P. 1072-1083.

15. Shubyn B. Intelligent Handover Management in 5G Mobile Networks based on Recurrent Neural Networks/ Shubyn B., Maksymyuk T.// 3rd IEEE International Conference on Advanced Information and Communications Technologies (AICT), (Lviv, Ukraine, July 2019), P. 348-351.