НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ імені ІГОРЯ СІКОРСЬКОГО»

Навчально-науковий інститут телекомунікаційних систем

Кафедра інформаційних технологій в телекомунікаціях

До захисту допущено:
В.о. завідувача кафедри
_____ Валерій ПРАВИЛО
«___»_____ 2026 р.

ДИПЛОМНА РОБОТА

на здобуття ступеня бакалавра

за освітньо-професійною програмою «Інформаційно-комунікаційні технології»

спеціальності 172 «Телекомунікації та радіотехніка»

на тему: «Вдосконалена архітектура 5G мереж»

Виконав: здобувач вищої освіти студент 4 курсу, групи ТІ-21
Саєнко Єгор Романович _____

Науковий керівник: Доцент кафедри ІТТ НН ІТС, кандидат технічних наук, доцент Правило Валерій Володимирович _____

Рецензент: доцент кафедри ТК НН ІТС, кандидат технічних наук, доцент Явіся Валерій Сергійович _____

Засвідчую, що у цій дипломній роботі немає запозичень з праць інших авторів без відповідних посилань.
Студент _____

Київ – 2026 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Навчально-науковий інститут телекомунікаційних систем
Кафедра інформаційних технологій в телекомунікаціях

Рівень вищої освіти – перший (бакалаврський)
Освітньо-професійна програма «Інформаційно-комунікаційні технології»
спеціальності 172 Телекомунікації та радіотехніка

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
_____ Валерій ПРАВИЛО
«_____»_____2026 р

ЗАВДАННЯ
на дипломну роботу студенту
Саєнку Єгору Романовичу

1. Тема дипломної роботи: Вдосконалена архітектура 5G мереж, науковий керівник роботи – доцент кафедри ІТТ НН ІТС, кандидат технічних наук, доцент Правило Валерій Володимирович, затверджені наказом по університету від 26.05.2026р. №1912-с

2. Строк подання студентом дипломної роботи 01.06.26р.

3. Об’єкт дослідження: Процес вдосконалення архітектури 5G мереж.

4. Вихідні дані до роботи: вимоги стандарту IMT-2020 до 5G мереж (eMBB, URLLC, mMTC); специфікації консорціуму 3GPP (Release 15-18) щодо сервісно-орієнтованої архітектури ядра (SBA) та радіоінтерфейсу 5G NR; стандарти ETSI щодо віртуалізації (NFV) та граничних обчислень (MEC); специфікації O-RAN Alliance.

5. Перелік завдань, які потрібно розробити: здійснити аналіз існуючої архітектури 5G мереж (NG-RAN та 5GC) та визначити її системні недоліки; дослідити режими розгортання Non-Standalone та Standalone; проаналізувати та обрати комплекс методів вдосконалення, включаючи NFV, SDN, Network

Slicing, MEC та O-RAN; розробити концепцію, структурну схему та механізми вдосконаленої архітектури; провести оцінку ефективності та порівняльний аналіз запропонованої архітектури з базовою.

6. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій тощо): презентація в PowerPoint, 13 рисунків.

7. Дата видачі завдання: 25.02.2026

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Затвердження теми та плану роботи. Аналіз еволюції та базової архітектури 5G мереж (RAN та Core).	25.02.2026 - 10.03.2026	Виконано
2	Ідентифікація та систематизація недоліків існуючої архітектури 5G. Початок роботи над першим розділом.	11.03.2026- 28.03.2026	Виконано
3	Дослідження методів вдосконалення: віртуалізація (NFV), SDN, Network Slicing, Edge Computing та O-RAN.	29.03.2026- 14.04.2026	Виконано
4	Розробка концепції, структурної схеми та механізмів вдосконаленої архітектури (рівні UE, O-RAN, MEC, Transport, Core).	15.04.2026- 30.04.2026	Виконано
5	Проведення оцінки ефективності. Порівняльний аналіз вдосконаленої архітектури з базовою за ключовими KPI.	01.05.2026- 14.05.2026	Виконано
6	Оформлення дипломної роботи, написання висновків.	16.05.2026- 01.06.2026	Виконано
7	Підготовка ілюстративного матеріалу та презентації до захисту.	01.06.2026 - 09.06.2026	Виконано
8	Захист дипломної роботи.	16.06.2026 - 19.06.2026	Виконано

Здобувач вищої освіти _____ Саєнко ЄГОР

Керівник дипломної роботи_____ Валерій ПРАВИЛО

РЕФЕРАТ

Дипломна робота містить 90 сторінок, 13 рисунків, 6 формул. Використано 36 джерел інформації.

Актуальність теми. Стрімкий розвиток сучасної індустрії телекомунікацій орієнтований на повсюдне розгортання інфраструктури п'ятого покоління (5G). Всупереч цьому, реальний досвід комерційного впровадження таких мереж наразі свідчить про суттєвий відрив від специфікацій, задекларованих у глобальному стандарті IMT-2020. Яскравим прикладом цієї невідповідності є показник наскрізної затримки (End-to-End latency) в операційних мережах, який варіюється у межах 8–25 мс. Можна впевнено сказати, що така ситуація є надзвичайно серйозною оскільки для реалізації сервісів наднадійної зв'язності з низькою затримкою (URLLC), необхідних для безпілотного транспорту чи тактильного інтернету, цей параметр не повинен перевищувати 1 мс.

Головне джерело згаданих технологічних бар'єрів криється в застарілих принципах побудови архітектури. Насамперед йдеться про жорстку централізацію шлюзів обробки трафіку користувачів , що створює значне транспортне плече, а також про низьку адаптивність систем до динамічного масштабування під час пікових навантажень. Виявлені архітектурні дефекти гостро актуалізують задачу розробки принципово нової, модернізованої концепції ядра мережі. Формування такої архітектурної моделі повинно здійснюватися на основі синергії передових технологічних парадигм: глибокій віртуалізації мережевих функцій (NFV), гнучкому сегментуванні ресурсів (network slicing) та перенесенні обчислювальних потужностей ближче до кінцевого споживача в рамках концепції граничних обчислень (MEC).

Мета роботи полягає у підвищенні ефективності мереж 5G за рахунок розробки вдосконаленої архітектури, що інтегрує хмарні технології NFV/SDN, наскрізне нарізання мережі (Network Slicing) та граничне обчислення (MEC).

Об'єктом дослідження є процес вдосконалення архітектури 5G мереж.

Предметом дослідження є архітектура 5G мереж.

Методи дослідження. Використано методи системного аналізу для оцінки еволюції мобільних мереж; декомпозиційне аналітичне моделювання для розрахунку наскрізної затримки (E2E Latency); порівняльний аналіз архітектурних рішень на основі ключових показників ефективності (KPI) стандарту ITU-R M.2410-0.

Ключові слова: 5G, АРХІТЕКТУРА МЕРЕЖІ, МЕРЕЖЕВЕ НАРІЗАННЯ (NETWORK SLICING), ГРАНИЧНЕ ОБЧИСЛЕННЯ (MEC), ВІРТУАЛІЗАЦІЯ (NFV/CNF), НАДІЙНІСТЬ, ЗАТРИМКА (LATENCY), O-RAN, NON-STANDALONE (NSA), STANDALONE (SA), EDGE COMPUTING, ПРОГРАМНО-ВИЗНАЧЕНІ МЕРЕЖІ (SDN).

ABSTRACT

The thesis contains 90 pages, 13 figures, 6 formulas. Used 36 sources of information.

Relevance of the topic. The rapid development of the modern telecommunications industry is focused on the widespread deployment of the fifth generation (5G) infrastructure. Contrary to this, the real experience of commercial implementation of such networks currently shows a significant gap from the specifications declared in the global IMT-2020 standard. A clear example of this discrepancy is the end-to-end latency in operational networks, which varies between 8 and 25 ms. We can safely say that this situation is extremely serious, because for the implementation of ultra-reliable low-latency connectivity (URLLC) services, which are necessary for unmanned transport or tactile Internet, this parameter should not exceed 1 ms.

The main source of the mentioned technological barriers lies in the outdated principles of building architecture. First of all, it is about the strict centralization of user traffic processing gateways, which creates a significant transport shoulder, as well as the low adaptability of systems to dynamic scaling during peak loads. The identified architectural defects sharply actualize the task of developing a fundamentally new, modernized concept of the network core. The formation of such an architectural model should be based on the synergy of advanced technological paradigms: deep virtualization of network functions (NFV), flexible segmentation of resources (network slicing) and the transfer of computing power closer to the end user within the framework of the concept of edge computing (MEC).

The aim of this work is to increase the efficiency of 5G networks by developing an advanced architecture that integrates NFV/SDN cloud technologies, network slicing and edge computing (MEC).

The subject of research is the architecture of 5G networks.

The object of the study is the process of improving the architecture of 5G networks.

Research methods. System analysis methods were used to evaluate the evolution of mobile networks; decomposition analytical modeling to calculate end-to-end latency (E2E Latency); comparative analysis of architectural solutions based on key performance indicators (KPI) of the ITU-R M.2410-0 standard.

Keywords: 5G, NETWORK ARCHITECTURE, NETWORK SLICING, EDGE COMPUTING (MEC), VIRTUALIZATION (NFV/CNF), RELIABILITY, LATENCY, O-RAN, NON-STANDALONE (NSA), STANDALONE (SA), EDGE COMPUTING, SOFTWARE-DEFINED NETWORKS (SDN).

ЗМІСТ

СПИСОК ТЕРМІНІВ І СКОРОЧЕНЬ

3GPP	3rd Generation Partnership Project – консорціум зі стандартизації мобільних мереж
5G	Fifth Generation – мережі п'ятого покоління мобільного зв'язку
5GC	5G Core – ядро мережі п'ятого покоління
5QI	5G QoS Identifier – ідентифікатор якості обслуговування у 5G
A1	Інтерфейс між Non–RT RIC та Near–RT RIC (O–RAN)
AMF	Access and Mobility Management Function — функція управління доступом та мобільністю
API	Application Programming Interface – програмний інтерфейс взаємодії
AR	Augmented Reality – доповнена реальність
AUSF	Authentication Server Function – функція сервера автентифікації
BLER	Block Error Rate — частка помилок на рівні блоків
CaaS	Containers as a Service – контейнери як послуга
CNF	Cloud-Native Network Function – хмарно-нативна мережева функція
CP	Control Plane – площина управління
CPU	Central Processing Unit – центральний процесор

CSI-RS	Channel State Information Reference Signal – опорний сигнал стану каналу
DP	Data Plane – площина даних (синонім UP)
DPDK	Data Plane Development Kit – бібліотека прискорення обробки пакетів у просторі користувача
DU	Distributed Unit – розподілений блок базової станції
E2	Інтерфейс між Near-RT RIC та O-DU/O-CU (O-RAN)
eMBB	Enhanced Mobile Broadband – клас сервісів надшвидкісного широкосмугового доступу
E-UPF	Edge User Plane Function – крайова функція площини користувача
ETSI	European Telecommunications Standards Institute – Європейський інститут телекомунікаційних стандартів
EPC	Evolved Packet Core – ядро мережі 4G LTE
FAR	Forwarding Action Rule – правило переадресації пакетів в UPF
FR1	Frequency Range 1 – частотний діапазон sub-6 ГГц для 5G NR
FR2	Frequency Range 2 – міліметровий частотний діапазон (24–52 ГГц) для 5G NR
GBR	Guaranteed Bit Rate – гарантована бітова швидкість
gNB	Next Generation Node B – базова станція 5G NR

GTP-U	GPRS Tunneling Protocol – User Plane – протокол тунелювання площини даних
HO	Handover – процедура хендоверу (передання обслуговування між базовими станціями)
HPA	Horizontal Pod Autoscaler – механізм горизонтального автомасштабування в Kubernetes
HTTP/2	Hypertext Transfer Protocol version 2 – протокол передачі даних, використовується в SBI
IIoT	Industrial Internet of Things – промисловий Інтернет речей
IMT-2020	International Mobile Telecommunications-2020 – стандарт ITU-R для мереж 5G
IoT	Internet of Things – Інтернет речей
ITU-R	International Telecommunication Union – Radiocommunication Sector – Сектор радіозв'язку МСЕ
JSON	JavaScript Object Notation – формат обміну даними
JWT	JSON Web Token – стандарт токенів авторизації
KPI	Key Performance Indicator – ключовий показник ефективності
LDPC	Low Density Parity Check – код перевірки чіткості низькою щільністю
LSTM	Long Short-Term Memory – архітектура рекурентної нейронної мережі
LTE	Long Term Evolution – стандарт мобільних мереж четвертого покоління (4G)

MAC	Medium Access Control – протокол управління доступом до середовища
MANO	Management and Orchestration – підсистема управління та оркестрації NFV
MAPE	Mean Absolute Percentage Error – середня абсолютна відсоткова похибка
MCC	Mobile Country Code – код країни мобільної мережі
MCS	Modulation and Coding Scheme – схема модуляції та кодування
MEC	Multi-access Edge Computing – граничне обчислення з мультидоступом
MEP	MEC Platform – платформа граничного обчислення за стандартом ETSI
MIMO	Multiple Input Multiple Output – технологія багатоантенних систем
mMTC	Massive Machine Type Communications – масові машинні комунікації (клас сервісів IoT)
MME	Mobility Management Entity – функція управління мобільністю в 4G EPC
MPLS	Multiprotocol Label Switching – протокол комутації за мітками
mTLS	Mutual Transport Layer Security – взаємна автентифікація TLS

NAS	Non-Access Stratum – протокол верхнього рівня між UE та AMF
Near-RT RIC	Near Real-Time RAN Intelligent Controller – контролер RAN з реакцією 10 мс – 1 с
NEF	Network Exposure Function – функція відкриття мережевих сервісів
NFV	Network Functions Virtualization – віртуалізація мережевих функцій
NFVI	NFV Infrastructure – фізична та програмна інфраструктура NFV
NFVO	NFV Orchestrator – оркестратор NFV у складі MANO
NG-RAN	Next Generation Radio Access Network – мережа радіодоступу наступного покоління 5G
Non-RT RIC	Non Real-Time RAN Intelligent Controller – контролер RAN з реакцією понад 1 с
NR	New Radio – стандарт радіоінтерфейсу 5G
NRF	NF Repository Function – функція репозиторію мережевих функцій
NSA	Non-Standalone – режим розгортання 5G з якорем у LTE
NSSF	Network Slice Selection Function – функція вибору мережевого зрізу
O-CU	O-RAN Centralized Unit – централізований блок базової станції у O-RAN

O-DU	O-RAN Distributed Unit – розподілений блок базової станції у O-RAN
O-RAN	Open Radio Access Network – відкрита архітектура мережі радіодоступу
O-RU	O-RAN Radio Unit – радіомодуль базової станції у O-RAN
OAuth 2.0	Open Authorization 2.0 – відкритий протокол авторизації
OFDMA	Orthogonal Frequency Division Multiple Access – ортогональний множинний доступ з частотним поділом
OPEX	Operational Expenditure – операційні витрати
P-GW	PDN Gateway – шлюз пакетної мережі в 4G EPC
PCF	Policy Control Function – функція управління політиками
PDCP	Packet Data Convergence Protocol – протокол конвергенції пакетних даних
PDR	Packet Detection Rule – правило виявлення пакетів в UPF
PDU	Protocol Data Unit – протокольний блок даних; також – тип сесії (PDU Session)
PFCP	Packet Forwarding Control Protocol – протокол управління переадресацією пакетів (N4)
PRB	Physical Resource Block – фізичний ресурсний блок у NR/LTE
QER	QoS Enforcement Rule – правило застосування QoS в UPF
QoS	Quality of Service – якість обслуговування

R-UPF	Regional User Plane Function – регіональна функція площини користувача
RAN	Radio Access Network – мережа радіодоступу
RIC	RAN Intelligent Controller – інтелектуальний контролер RAN (O-RAN)
RLC	Radio Link Control – протокол управління радіоканалом
RRC	Radio Resource Control – протокол управління радіоресурсами
RRM	Radio Resource Management – управління радіоресурсами
RSRP	Reference Signal Received Power – потужність прийнятого опорного сигналу
RTO	Recovery Time Objective – цільовий час відновлення після відмови
SA	Standalone – повністю самостійний режим розгортання 5G з ядром 5GC
SBA	Service-Based Architecture – сервісно-орієнтована архітектура ядра 5GC
SBI	Service-Based Interface – уніфікований сервісний інтерфейс NF у 5GC
SCS	Subcarrier Spacing – міжпідносійний інтервал у 5G NR
SDAP	Service Data Adaptation Protocol – протокол адаптації сервісних даних
SDN	Software-Defined Networking – програмно-визначений мережевий зв'язок

S-NSSAI	Single Network Slice Selection Assistance Information – ідентифікатор мережевого зрізу
SINR	Signal to Interference and Noise Ratio – відношення сигнал/завада+шум
SLA	Service Level Agreement – угода про рівень обслуговування
SMF	Session Management Function – функція управління сесіями
SCP	Service Communication Proxy – проксі сервісних комунікацій у 5GC
SUCI	Subscription Concealed Identifier – зашифрований ідентифікатор абонента
SUPI	Subscription Permanent Identifier – постійний ідентифікатор підписки
TLS	Transport Layer Security – протокол безпеки транспортного рівня
UDM	Unified Data Management – функція єдиного управління даними
UDR	Unified Data Repository – єдине сховище даних у 5GC
UE	User Equipment – абонентський пристрій
ULCL	Uplink Classifier – класифікатор висхідного трафіку в UPF
UMTS	Universal Mobile Telecommunications System – стандарт мобільних мереж 3G
UP	User Plane – площина даних (площина користувача)

UPF	User Plane Function – функція площини користувача
URLLC	Ultra-Reliable Low-Latency Communications – наднадійні комунікації з наднизькою затримкою
V2X	Vehicle-to-Everything – комунікації транспортних засобів
VNF	Virtualized Network Function – віртуалізована мережева функція
VR	Virtual Reality – віртуальна реальність
VXLAN	Virtual Extensible LAN – протокол мережевої віртуалізації
xApp	eXtended Application – застосунок для Near-RT RIC у архітектурі O-RAN
XR	Extended Reality - розширена реальність (AR + VR + MR)

ВСТУП

Сучасна глобальна телекомунікаційна галузь переживає безпрецедентну технологічну трансформацію, що визначається переходом до мереж п'ятого покоління (5G). Саме стандарт 5G, заснований на вимогах International Mobile Telecommunications-2020 (IMT-2020) Міжнародного союзу електрозв'язку, покликаний забезпечити якісно новий рівень бездротового зв'язку для задоволення зростаючих потреб цифрової економіки, Інтернету речей та автоматизованих систем [1]. Стандарт IMT-2020 визначає три основних класи сервісів: enhanced Mobile Broadband (eMBB) для надвисокошвидкісного широкосмугового доступу, Ultra-Reliable Low-Latency Communications (URLLC) для критично важливих застосунків та massive Machine Type Communications (mMTC) для масового підключення IoT-пристроїв, встановлюючи цільову наскрізну затримку до 1 мс та пікову пропускну здатність до 20 Гбіт/с.

Перші комерційні розгортання 5G розпочались у 2019 році, і на сьогодні, за даними галузевих консорціумів, мережі 5G функціонують більш ніж у 90 країнах світу. Концепція розвитку мобільних мереж нового покоління була сформована у програмному документі NGMN Alliance 5G White Paper [2], де вперше систематизовано вимоги операторів до майбутніх мереж та окреслено ключові сценарії використання. Документ заклав підґрунтя для подальшої стандартизаційної роботи в 3GPP та ITU-R і досі є одним із базових орієнтирів при розробці 5G-рішень.

За прогнозами провідних телекомунікаційних компаній, глобальний мобільний трафік даних продовжує зростати прискореними темпами. Згідно з даними звіту Ericsson Mobility Report [3], до 2029 року щомісячний обсяг мобільного трафіку у мережах 5G перевищить 370 ексабайтів, причому основний внесок забезпечуватимуть відеосервіси (понад 75% трафіку), хмарні ігри та XR-застосунки (розширена реальність). Таке зростання трафіку у поєднанні з жорсткими вимогами нових застосунків до затримки та надійності

ставить перед операторами принципово нові архітектурні виклики, які неможливо вирішити в рамках традиційного підходу до проектування мереж.

Актуальність теми дослідження визначається тим, що реальні комерційні реалізації 5G мереж ще не досягають цільових показників IMT-2020. Численні незалежні тестування операторських 5G мереж свідчать про наскрізну затримку у діапазоні 8–25 мс для найкращих умов, тоді як URLLC вимагає ≤ 1 мс. Пікова пропускна здатність у реальних умовах покриття – 3–8 Гбіт/с проти цільових 20 Гбіт/с. Ці розриви обумовлені архітектурними обмеженнями базової реалізації 5G, у тому числі централізованим розміщенням функцій площини користувача, недостатньою гнучкістю масштабування мережевих функцій та відсутністю комплексного наскрізного нарізання мережі.

Для **досягнення поставленої мети дослідження** було вирішено такі завдання:

1) здійснено всебічний аналіз існуючої архітектури 5G мереж, включаючи компоненти NG-RAN та ядра 5GC;

2) досліджено режими розгортання Non-Standalone (NSA) та Standalone (SA) та їх відмінності;

3) ідентифіковано та систематизовано недоліки базової архітектури 5G;

4) проаналізовано та обрано комплекс методів вдосконалення: NFV, SDN, Network Slicing, MEC, O-RAN;

5) розроблено структурну схему та механізми вдосконаленої архітектури;

6) проведено оцінку ефективності та порівняльний аналіз запропонованої архітектури з базовою.

Об'єкт дослідження є процес вдосконалення архітектури 5G мереж.

Предмет дослідження є архітектура 5G мереж.

Метою роботи є підвищення ефективності мереж 5G за рахунок розробки вдосконаленої архітектури, що інтегрує хмарні технології NFV/SDN, наскрізне нарізання мережі та граничне обчислення.

Елементом новизни є те, що вдосконалена архітектура 5G-мереж порівняно з традиційними архітектурними рішеннями, закладених у специфікаціях 3GPP Release 15-18, інтегрує трирівневу ієрархію UPF (Edge-Regional-Central), наскрізне нарізання мережі від рівня O-RAN до ядра 5GC та механізм динамічного вибору рівня обробки трафіку на основі значення 5QI і поточного завантаження, що дозволяє досягти E2E-затримки 1–3 мс для URLLC-сервісів - у 8–25 разів менше порівняно з існуючими комерційними реалізаціями.

Практичне значення нашого дослідження зводиться до того, що розроблена архітектура може бути застосована операторами мобільного зв'язку при розгортанні 5G SA-мереж для: забезпечення URLLC-сервісів промислової автоматизації, дистанційної медицини та автономного транспорту (надійність ≥ 99,999%); скорочення часу горизонтального масштабування мережевих функцій у 20-60 разів (з 2–8 хв до 5–15 с) за рахунок інтеграції Kubernetes HPA/VPA; підвищення ємності трафіку мережі в 2–4 рази без збільшення апаратних ресурсів за рахунок повноцінного наскрізного Network Slicing та граничного обчислення.

РОЗДІЛ 1
АНАЛІЗ АРХІТЕКТУРИ МЕРЕЖ П'ЯТОГО ПОКОЛІННЯ (5G)

1.1 Еволюція мобільних мереж та передумови впровадження 5G

Розвиток мобільних телекомунікаційних мереж охоплює більш ніж чотири десятиліття безперервної еволюції – від аналогових систем першого покоління до цифрових мереж нового покоління. Фундаментальний огляд цього шляху необхідний для розуміння архітектурних рішень, що лежать в основі 5G, та для усвідомлення причин, що зумовили перехід до принципово нової телекомунікаційної парадигми. Систематичний аналіз еволюції мобільних мереж представлений у роботі [4], де автори відзначають, що кожне нове покоління виникало не лише у відповідь на зростання попиту на швидкість передачі даних, але і як реакція на якісно нові вимоги суспільства та промисловості до можливостей телекомунікаційної інфраструктури.

Перше покоління мобільного зв'язку (1G), розгорнуте на початку 1980-х років, ґрунтувалося на аналогових технологіях частотної модуляції (FM). Флагманською системою першого покоління в США стала AMPS (Advanced Mobile Phone System), тоді як у Скандинавії використовувався стандарт NMT (Nordic Mobile Telephone). Мережева архітектура 1G була відносно простою: мобільні комутаційні центри (MSC) з'єднували сукупності базових станцій із загальною телефонною мережею (PSTN). Безпека передачі даних на рівні радіоінтерфейсу практично була відсутня, що робило переговори вразливими для прослуховування. Пропускна здатність радіоканалу не перевищувала 2,4 кбіт/с, а сам зв'язок забезпечував виключно голосові послуги.

Революційний перехід до цифрового зв'язку відбувся у 1990-х роках із появою стандарту GSM (Global System for Mobile Communications) – стандарту другого покоління (2G). Стандарт GSM, розроблений під егідою Європейського інституту телекомунікаційних стандартів (ETSI), впровадив ключові технологічні інновації: цифрову кодово-імпульсну модуляцію голосу, алгоритми шифрування радіоканалу (A5/1), симкарти (SIM) для ідентифікації

абонентів та системи роумінгу на міжнародному рівні. Пізніше GSM еволюціонував у бік передачі пакетних даних через технології GPRS (General Packet Radio Service, до 114 кбіт/с) та EDGE (Enhanced Data rates for GSM Evolution, до 384 кбіт/с). Паралельно у США розвивалась технологія IS-95 CDMA, що використовувала кодовий поділ каналів, демонструючи кращу спектральну ефективність у порівнянні з TDMA GSM. Введення EPC-архітектури у 4G LTE описано у монографії [5], де детально розглянуто концептуальні відмінності між архітектурою GSM/GPRS та пізнішими IP-орієнтованими системами.

Третє покоління (3G) стандарту UMTS/WCDMA (Universal Mobile Telecommunications System / Wideband Code Division Multiple Access), стандартизованого консорціумом 3GPP, відкрило еру широкосмугового мобільного Інтернету. Ключовою відмінністю WCDMA від GSM стало використання широкосмугового CDMA з смугою каналу 5 МГц, що значно підвищило спектральну ефективність. Технологія HSDPA (High Speed Downlink Packet Access) дозволила досягти пікової швидкості завантаження до 14,4 Мбіт/с, а HSUPA – до 5,76 Мбіт/с у зворотному каналі. Архітектура UTRAN (UMTS Terrestrial Radio Access Network) включала контролери радіомережі (RNC), що централізовано управляли кількома базовими станціями Node B. Ядро мережі UMTS складалося з MSC для голосових послуг та SGSN/GGSN для пакетних даних. Дослідники у роботі [6] підкреслюють, що саме 3G сформувало сучасну масову культуру споживання мобільного Інтернету, хоча архітектурні обмеження UMTS та наростаючий дефіцит спектру зумовили необхідність переходу до наступного покоління.

Четверте покоління (4G) LTE (Long Term Evolution) та LTE-Advanced принесло принциповий архітектурний прорив – повну орієнтацію на IP-протокол (All-IP Network) та значне спрощення мережевої ієрархії. Технологія OFDMA (Orthogonal Frequency Division Multiple Access) з підтримкою MIMO (Multiple-Input Multiple-Output) дозволила досягти пікових швидкостей понад 1 Гбіт/с у стандарті LTE-Advanced Pro. Архітектурно 4G ввела концепцію

«плоскої мережі»: базові станції eNodeB (evolved Node B) безпосередньо з'єднуються між собою через інтерфейс X2, що мінімізує затримку при хендовері. Ядро LTE EPC (Evolved Packet Core) чітко розмежовує функції управління мобільністю (MME), шлюзу обслуговування (S-GW) та шлюзу PDN (P-GW), забезпечуючи ефективну підтримку різноманітних IP-сервісів. Фундаментальні проблеми 4G – обмежена підтримка IoT, недостатня спектральна ефективність у частотному діапазоні mmWave та затримка > 20 мс – створили передумови для розробки 5G, що описані авторами у [7] як п'ять напрямів «руйнівних технологій» для наступного покоління.

Концептуально 5G є відповіддю на три фундаментально різні групи вимог, що неможливо одночасно задовольнити в рамках 4G архітектури. Перша група – eMBB (Enhanced Mobile Broadband): підтримка відеосервісів 4K/8K, хмарних ігор та XR-технологій, що вимагають пікової пропускної здатності понад 10 Гбіт/с та середньої швидкості у точці користувача понад 100 Мбіт/с. Друга група – URLLC (Ultra-Reliable Low Latency Communications): промислова автоматизація, дистанційна хірургія, автономне транспортування, тактильний Інтернет – застосунки, що вимагають наскрізної затримки ≤ 1 мс та надійності 99,999%. Третя група – mMTC (massive Machine Type Communications): підключення від 10^5 до 10^6 IoT-пристроїв на км² з мінімальним енергоспоживанням та тривалим терміном служби від батарей. Саме ці три класи, детально описані у роботі [8] в рамках проекту METIS, визначили архітектурну гетерогенність 5G та необхідність революційних рішень, які систематизовано у таблиці 1.1.

Таблиця 1.1 – Порівняльна характеристика поколінь мобільних мереж (1G–5G)

Параметр	**2G GSM**	**3G UMTS**	**4G LTE-A**	**5G NR**
Рік впровадження	1991	2001	2009	2019

Продовження таблиці 1.1

Пікова DL швидк.	384 кбіт/с	14,4 Мбіт/с	1 Гбіт/с	20 Гбіт/с
Затримка RAN (мс)	~300	~100	~15–20	~0,5 (URLLC)
Технологія РД	TDMA/FDMA	WCDMA	OFDMA/MIMO	OFDMA/M-MIMO
Архітектура ядра	Circuit SW	SGSN + GGSN	EPC (All-IP)	5GC (SBA)
Щільн. пр. ($км^{-2}$)	–	–	10^5	10^6
Спектр. ефект.	0,5 біт/с/Гц	1,0 біт/с/Гц	3,5 біт/с/Гц	до 9 біт/с/Гц

Таблиця 1.1 наочно демонструє, що еволюція від 2G до 5G не є лінійним зростанням одного параметра – це принципова зміна технологічних концепцій. Еволюцію поколінь мобільних мереж, яка демонструє перехід від аналогових систем до інтелектуальних 5G-мереж, представлено на рисунку 1.1. Перехід від комутації каналів до повністю пакетної передачі даних (All-IP), від жорсткої апаратної архітектури ядра до сервісно-орієнтованої (SBA), від фіксованих модуляційних схем до гнучкої нумерології – все це свідчить про революційний, а не еволюційний характер 5G. Зокрема, цільова спектральна ефективність 5G NR у 9 біт/с/Гц перевищує показник 4G LTE-A у 2,5 рази, а щільність підключень зростає на порядок.

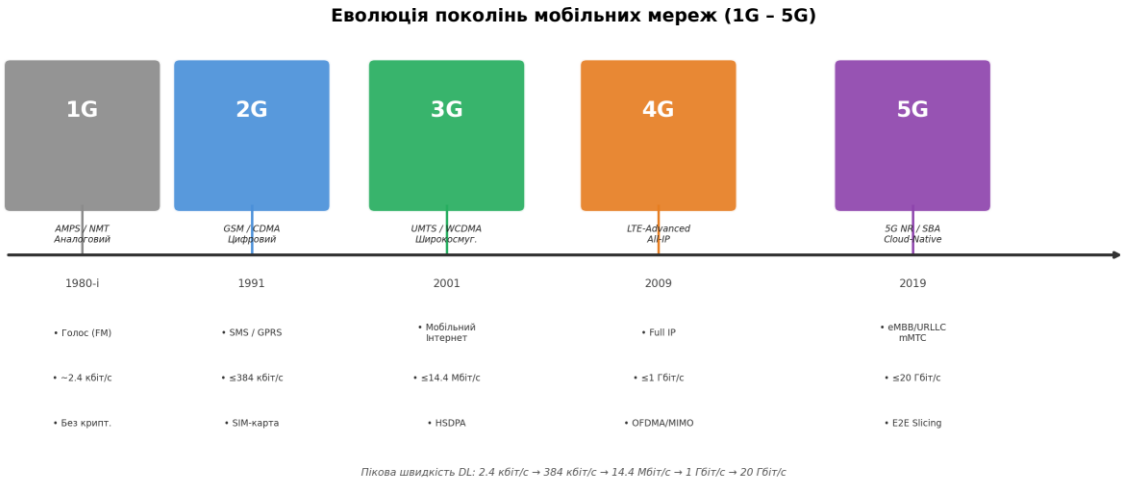

Рисунок 1.1 – Еволюція поколінь мобільних мереж (1G–5G)

Стандартизація 5G здійснюється у рамках поетапного підходу 3GPP Release, де кожен Release закріплює певний набір функцій та можливостей, еволюцію яких систематизовано у таблиці 1.2. Release 15 (грудень 2018) є першою «замороженою» специфікацією 5G NR, що визначила базову архітектуру SA та NSA, ядро 5GC з SBA, специфікацію фізичного рівня NR для FR1 та FR2. Release 16 (червень 2020) розширив підтримку URLLC для промислових застосунків (IIoT), ввів покращений E2E Network Slicing, підтримку V2X Sidelink та IAB (Integrated Access and Backhaul). Release 17 (червень 2022) додав підтримку NTN (Non-Terrestrial Networks) для супутникового підключення IoT, RedCap (Reduced Capability) для спрощених пристроїв та розширив можливості mmWave. Release 18 («5G Advanced», березень 2024) зосереджений на інтеграції ШІ/МН безпосередньо у протоколи NR, оптимізації для XR-застосунків та Ambient IoT.

Таблиця 1.2 – Еволюція 3GPP Releases для 5G та їх ключові функції

Release	**Дата завер.**	**Ключові нові функції та технології**	**Практичне значення**
Rel. 15	12/2018	Базова SA/NSA 5G, 5GC SBA, NR FR1/FR2, eMBB/URLLC/mMTC, N1–N26 інтерфейси	Перші комерційні 5G NSA; основа SA-розгортань

Продовження таблиці 1.2

Rel. 16	06/2020	IIoT/URLLC покращення, NR V2X Sidelink, IAB, 5G LAN, NR U-NII-7	Смарт-фабрики, V2X пілоти, WiFi/NR коекзист.
Rel. 17	06/2022	NTN (супутники/HAPS), RedCap IoT, mmWave розширення, Network Slicing оновл.	IoT via NTN, спрощені wearable пристрої
Rel. 18	03/2024	5G Advanced: AI/ML у RAN, XR оптимізація, Ambient IoT, NR Sidelink relay	AI-нативний NR, готовність до еволюції в 6G

1.2 Базова архітектура 5G (RAN та Core)

Архітектура мережі п'ятого покоління складається з трьох функціональних доменів: мережі радіодоступу наступного покоління (NG-RAN – Next Generation Radio Access Network), транспортної мережі та ядра 5G (5GC – 5G Core Network). Між доменами визначені стандартизовані логічні інтерфейси, що забезпечують їх технологічну незалежність та можливість мультивендорного розгортання. Повна специфікація системної архітектури 5G закріплена у технічній специфікації 3GPP TS 23.501 [9], яка визначає архітектурні принципи, функціональний поділ та перелік усіх мережевих функцій ядра, їхні ролі та інтерфейси між ними.

Мережа радіодоступу NG-RAN складається з базових станцій нового покоління gNB (next generation Node B), що підтримують інтерфейс 5G NR. Кожна gNB може бути реалізована у трьох логічних сутностях відповідно до специфікації 3GPP TS 38.300 [10]: централізований блок gNB-CU (Centralized Unit), розподілений блок gNB-DU (Distributed Unit) та одиниця радіо-модуля gNB-RU (Radio Unit). Такий функціональний поділ є основою для концепції Cloud-RAN або хмарної RAN, де gNB-CU та частково gNB-DU можуть бути реалізовані на програмних платформах у ЦОД. Централізований блок gNB-CU обробляє протоколи верхнього стека: RRC (Radio Resource Control), PDCP

(Packet Data Convergence Protocol) та SDAP (Service Data Adaptation Protocol). Блок gNB-CU може додатково розділятися на gNB-CU-CP (Control Plane) та gNB-CU-UP (User Plane). Розподілений блок gNB-DU відповідає за нижні протоколи: RLC (Radio Link Control), MAC (Medium Access Control) та нижній рівень PHY (Physical). Інтерфейс F1 між gNB-CU та gNB-DU визначає поділ стека протоколів та є стандартизованим 3GPP. Архітектурні принципи цього поділу детально досліджені у технічному звіті [11].

Базові станції gNB взаємодіють між собою через інтерфейс Xn (аналог X2 у 4G LTE), який забезпечує координацію хендоверу (X2-based handover), передачу контексту UE та координацію радіоресурсів між сусідніми клітинами. З ядром мережі 5GC базові станції gNB підключаються через дворівневий інтерфейс NG: сигнальний інтерфейс N2 до функції AMF та транспортний інтерфейс N3 до функції UPF. Така дворівнева структура NG є принциповим нововведенням 5G, що відображає чітке розділення площин управління та даних на рівні NG-RAN. Монографія [12] надає вичерпний опис стекової архітектури протоколів NR та їх відображення на функціональні блоки NG-RAN.

Ядро мережі 5G (5GC) реалізує сервісно-орієнтовану архітектуру (SBA – Service-Based Architecture), що є одним із найбільш революційних архітектурних рішень 5G. В рамках SBA мережеві функції (NF) взаємодіють між собою та з зовнішніми системами через стандартизовані API на базі протоколу HTTP/2 із JSON-кодуванням (або Protocol Buffers). Кожна NF реєструє свої сервіси у центральному репозиторії (NRF) та може виступати як постачальником, так і споживачем сервісів інших NF через уніфіковану шину SBI (Service-Based Interface). Детальний опис функцій 5GC подано у монографії [13], де виконано систематизацію всіх NF з розкриттям їх внутрішньої логіки та протоколів взаємодії рисунок 1.2.

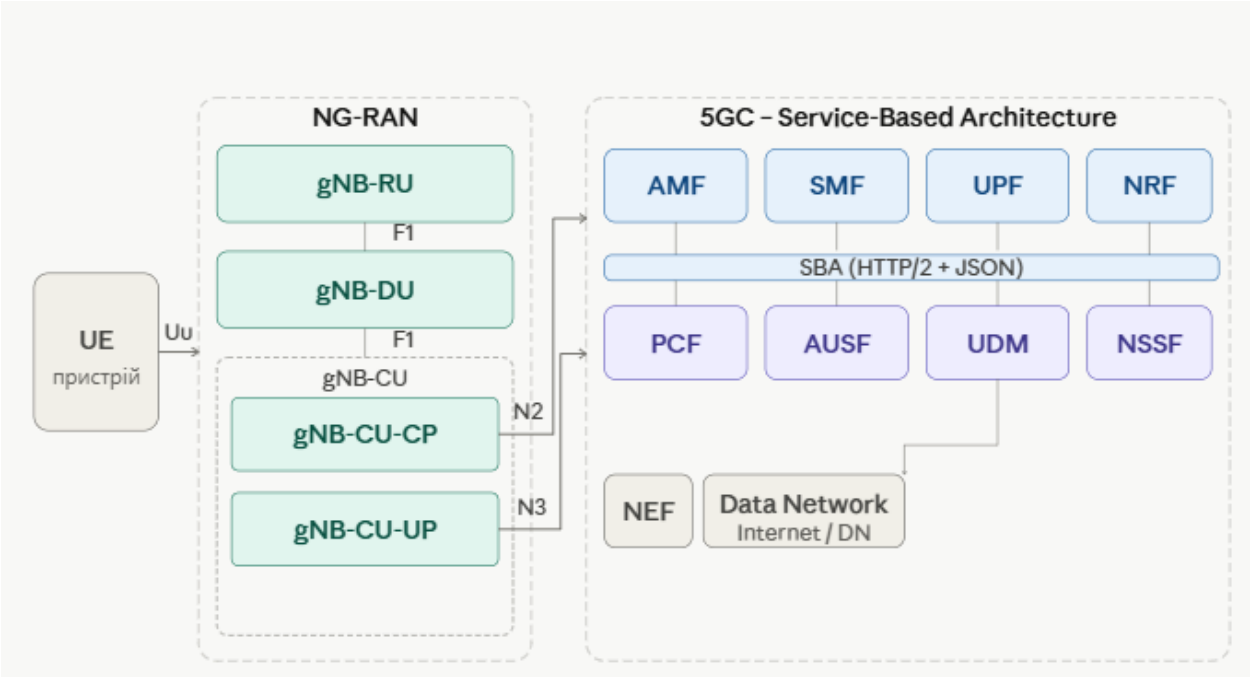

Рисунок 1.2 – Архітектура мережі 5G: NG-RAN та ядро 5GC з SBA-інтерфейсами

Як зображено на рисунку 1.2, сервісно-орієнтована архітектура 5GC суттєво відрізняється від жорсткої ієрархічної архітектури EPC у 4G. Розглянемо основні NF ядра мережі детально.

Функція управління доступом та мобільністю AMF (Access and Mobility Management Function) є «точкою входу» для сигнального трафіку від UE (User Equipment) до ядра мережі. Специфікація 3GPP TS 23.502 [14] визначає, що AMF виконує: термінування NAS (Non-Access Stratum) сигналізації та шифрування NAS; реєстрацію UE у мережі та управління контекстом UE; процедури управління мобільністю (хендовер між gNB, оновлення зони реєстрації TA – Tracking Area); вибір SMF для встановлення PDU-сесій та взаємодію з AUSF для аутентифікації UE. З точки зору відмовостійкості, кожен UE у будь-який момент прив'язаний рівно до однієї AMF (обслуговуюча AMF), що є єдиним сигнальним посередником між UE та ядром.

Функція управління сесіями SMF (Session Management Function) відповідає за повний lifecycle PDU-сесії (Protocol Data Unit Session): від ініціювання та встановлення до модифікації та вивільнення. Специфічні функції SMF: вибір та управління UPF-екземпляром для поточної PDU-сесії

через протокол PFCP (Packet Forwarding Control Protocol); розподіл IP-адреси для UE (DHCPv4/v6 або статично з пулу); застосування правил QoS, отриманих від PCF, на рівні QoS Flow в UPF; контроль за тарифікаційними звітами UPF та передача їх до CHF (Charging Function). SMF виступає контролером площини даних відповідно до концепції SCP (Software Defined Networking).

Функція площини користувача UPF (User Plane Function) є найважливішим елементом 5GC з точки зору обробки трафіку кінцевого користувача. На відміну від функцій площини управління, UPF безпосередньо обробляє пакети даних UE. Основні задачі UPF: маршрутизація та переадресація пакетів між NG-RAN та зовнішніми мережами (Data Network, DN) через GTP-U тунелі; застосування правил QoS на рівні QoS Flow (маркування DSCP, управління чергами); функція ULCL (Uplink Classifier) для локального вивантаження трафіку до MEC-платформ; звітування про використання трафіку для тарифікації; підтримка технологій IP-Anchor та Multi-homing для PDU-сесій. UPF може розміщуватись як у центральному ЦОД, так і на рівні граничного обчислення, що є ключовою можливістю для MEC-інтеграції [14].

Функція управління політиками PCF (Policy Control Function) централізовано керує правилами обслуговування, що передаються до AMF та SMF. Специфікація 3GPP TS 23.503 [15] визначає два основні набори правил PCF: AM Policy (Access and Mobility Policy) – правила для AMF щодо дозволених зон доступу, категорій RFSP та зон слайсингу; SM Policy (Session Management Policy) – правила QoS для SMF щодо PCC (Policy and Charging Control) правил, включаючи QoS-потоки, авторизовані типи PDU-сесій та правила тарифікації. PCF взаємодіє з UDM для отримання профілю підписки абонента та з зовнішньою функцією застосунків (AF) через N5 або NEF. Повна систематизація функціональних блоків ядра 5GC, їх основних обов'язків та порівняння з архітектурою 4G EPC наведена у таблиці 1.3.

Таблиця 1.3 – Ключові NF ядра 5GC, їх функції та аналоги у 4G EPC

NF	**Повна назва**	**Основні функції та відповідальність**	**Аналог у 4G EPC**
AMF	Access & Mobility Mgmt Function	Реєстрація UE, NAS-термінування, мобільність, вибір SMF, автентифікація (з AUSF)	MME
SMF	Session Management Function	Управління PDU-сесіями, вибір UPF (PFCP), розподіл IP-адрес, застосування QoS	MME + SGW-C + PGW-C
UPF	User Plane Function	Маршрутизація пакетів, QoS enforcement, ULCL, тарифікаційна звітність, GTP-U	SGW-U + PGW-U
NRF	NF Repository Function	Реєстрація NF, NF Discovery, видача OAuth 2.0 токенів	–
PCF	Policy Control Function	Управління AM/SM policy, QoS-правила, взаємодія з AF та UDM	PCRF
AUSF	Authentication Server Function	Виконання 5G AKA автентифікації, генерація HXRES*/KSEAF	HSS (частково)
UDM	Unified Data Management	Зберігання підписки, генерація облікових даних AV, управління реєстрацією NF	HSS
NSSF	Network Slice Selection Function	Вибір мережевого зрізу (S-NSSAI) для UE за допомогою AMF	–

Продовження таблиці 1.3

NEF	Network Exposure Function	Відкриття сервісів ядра зовнішнім AS/AF через захищений API	SCEF (часткова)

Аналіз таблиці 1.3 підтверджує, що архітектура 5GC є значно більш гранульованою порівняно з EPC 4G. Функції, що раніше концентрувались в MME або PGW, тепер розподілені між кількома спеціалізованими NF. Такий поділ підвищує гнучкість масштабування кожної функції незалежно та спрощує оновлення конкретних компонентів без перебою в роботі мережі.

Додатковими NF, що відіграють важливу роль у 5GC, є: UDR (Unified Data Repository) – єдине сховище всіх структурованих даних 5GC (дані підписки, дані застосунків, правила PCC), що надає уніфікований інтерфейс доступу для NF-читачів (UDM, PCF, NEF); UDSF (Unstructured Data Storage Function) – сховище неструктурованих даних у форматі ключ-значення для NF, які потребують персистентного зберігання контексту між перезапусками (наприклад, AMF зберігає UE-контексти у UDSF); SCP (Service Communication Proxy) – опціональний посередник для NF-взаємодій через SBI, що здійснює балансування навантаження між NF-екземплярами, виявлення та вибір NF (NF Discovery) від імені NF-споживача, моніторинг та тротлінг API-запитів. SCP є важливим компонентом для великих розгортань 5GC, де кількість NF-екземплярів може вимірюватись сотнями, і прямий NRF-Discovery для кожної взаємодії є неефективним. З точки зору масштабованості SCP дозволяє реалізувати централізовані точки контролю без прямої залежності між NF-парами.

Стек протоколів у площині управління 5G NAS (Non-Access Stratum) визначає взаємодію між UE та AMF. Протоколи NAS у 5G поділяються на два рівні: 5GMM (5G Mobility Management) – обробляє процедури реєстрації, дерéєстрації, управління ідентифікацією, процедури автентифікації та специфічні для режиму EM (Emergency) операції; 5GSM (5G Session

Management) – обробляє встановлення, модифікацію та вивільнення PDU-сесій між UE та SMF (через AMF). Сигналізація NAS між UE та AMF завжди шифрується та захищається на цілісність за допомогою ключів, встановлених у процедурі 5G AKA. Протоколи площини даних (SDAP, PDCP, RLC, MAC, PHY) між UE та gNB визначені у специфікаціях 3GPP TS 38.3xx та не є частиною NAS.

1.3 Технологічні особливості та ключові компоненти 5G (Non-Standalone, Standalone, 5G New Radio)

Стандарт 5G NR (New Radio) визначає два принципово різних режими розгортання залежно від ступеня опори на існуючу 4G-інфраструктуру. Вибір між Non-Standalone (NSA) та Standalone (SA) є фундаментальним архітектурним рішенням, що визначає не лише поточні можливості мережі, але й шлях її подальшого розвитку та здатність підтримувати конкретні класи сервісів. Фізичний рівень NR визначено у специфікації 3GPP TS 38.211 [16], яка регламентує структуру фрейму, типи сигналів та канали передачі.

Non-Standalone (NSA) архітектура, відповідна Опції 3x (EN-DC, E-UTRAN New Radio – Dual Connectivity) у класифікації 3GPP, призначена для прискореного розгортання 5G NR без повної заміни ядра мережі. У режимі NSA базові станції 5G NR (gNB) функціонують як додаткові носії даних, тоді як «якірна» базова станція LTE (eNB або ng-eNB) забезпечує сигнальний управляючий канал NAS та підключення до ядра EPC 4G. UE у режимі EN-DC одночасно підтримує підключення до LTE (Master Cell Group, MCG) та NR (Secondary Cell Group, SCG), агрегуючи радіоресурси обох технологій. Основна перевага NSA – можливість швидкого запуску 5G-сервісів без розгортання 5GC, що скорочує час виведення продукту на ринок для операторів. Проте NSA має принципові обмеження: ядро EPC вносить додаткову затримку, що унеможливлює підтримку URLLC; відсутній повноцінний E2E Network Slicing; неможлива підтримка нових інтерфейсів для IoT (NB-IoT та eMTC у 5G SA конфігурації).

Standalone (SA) архітектура, що відповідає Опції 2, є цільовою реалізацією 5G, в якій і базові станції gNB, і ядро 5GC є повністю стандартними компонентами специфікацій 3GPP Release 15 та вище. В режимі SA UE підключається безпосередньо до gNB та через N2/N3 – до 5GC. Лише в режимі SA повністю доступні: підтримка всіх трьох класів сервісів (eMBB, URLLC, mMTC); наскрізний E2E Network Slicing; повноцінна SBA-архітектура 5GC; розширені механізми безпеки 5G AKA та SUPI-захист; безшовна інтеграція з MEC через UPF ULCL. Процедури фізичного рівня для передачі даних детально описані у специфікації 3GPP TS 38.214 [17]. Принципову архітектурну різницю між режимами розгортання NSA та SA ілюструє рисунок 1.3.

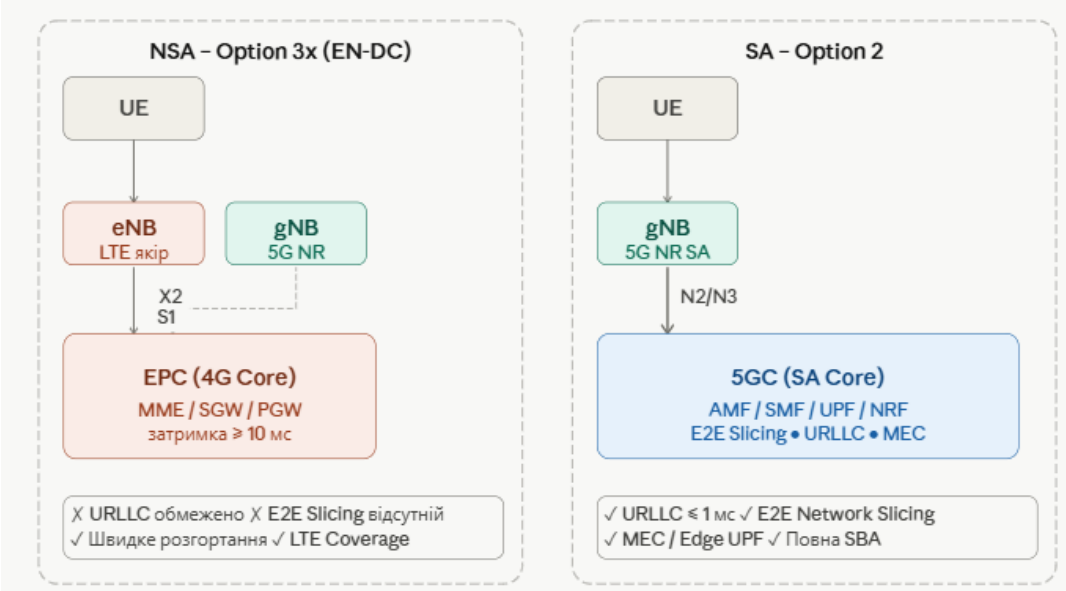

Рисунок 1.3 – Порівняння режимів NSA (Option 3x / EN-DC) та SA (Option 2)

Рисунок 1.3 ілюструє принципову архітектурну різницю між двома режимами. Наявність LTE-якоря у NSA вносить додаткову сигнальну затримку через MME та обмежує можливості динамічного управління ресурсами у чисто 5G-оточенні. Режим SA усуває цю залежність та є єдиним шляхом до повного розкриття потенціалу 5G.

Технологія 5G New Radio (NR) визначає радіоінтерфейс нового покоління, що охоплює два частотних діапазони. FR1 (Frequency Range 1,

«sub-6 GHz») від 410 МГц до 7,125 ГГц забезпечує значно кращі можливості покриття завдяки нижчим частотам та більшому проникненню в будівлі. FR2 (Frequency Range 2, «mmWave») від 24,25 ГГц до 52,6 ГГц надає доступ до надширокополосних каналів (до 400 МГц на поднесучу), але з суттєво обмеженим радіусом дії (100–200 метрів у відкритому просторі) та значним загасанням від опадів та перешкод. Теоретичні та вимірювальні властивості міліметрового діапазону детально досліджені у роботі [18], де автори показали, що mmWave є реалістичною основою для міського щільного покриття 5G, а пізніші дослідження [19] підтвердили перші вимірювання характеристик міліметрових хвиль у міських умовах.

Ключовою особливістю NR є гнучка нумерологія, що визначає міжпідносійний інтервал (SCS – Subcarrier Spacing). Залежність SCS від параметра нумерології μ описується виразом (1.1):

$$SCS = 15 \times 2^{\mu} \text{ кГц}$$

де SCS – міжпідносійний інтервал підносійних, кГц; $\mu \in \{0, 1, 2, 3, 4\}$ – параметр нумерології 5G NR. При $\mu = 0$ SCS = 15 кГц (аналогічно LTE), при $\mu = 4$ SCS = 240 кГц.

Гнучкість нумерології є принциповою технологічною перевагою NR: менший SCS забезпечує більшу стійкість до міжсимвольної інтерференції та Доплерівського зсуву (підходить для FR1 та мобільних умов), тоді як більший SCS зменшує тривалість символу і відповідно знижує затримку одноразового доступу (підходить для FR2 та URLLC). Різні нумерології можуть використовуватись у різних слотах одного і того ж частотного ресурсу, що систематизовано у таблиці 1.4.

Таблиця 1.4 – Нумерологія 5G NR: параметри підносійних та область застосування

μ	**SCS (кГц)**	**Тривал. слоту (мс)**	**Слотів/підкадр**	**Діапазон (FR)**	**Сценарій застосування**
0	15	1,0	1	FR1 (sub-6 ГГц)	eMBB, широке покриття, нижній FR1
1	30	0,5	2	FR1 (sub-6 ГГц)	eMBB, мікросоти, вищий FR1
2	60	0,25	4	FR1 / FR2	URLLC, низька затримка, перехідний FR1/FR2
3	120	0,125	8	FR2 (mmWave)	eMBB mmWave, надшвидкісне покриття
4	240	0,0625	16	FR2 (mmWave)	Опорні сигнали mmWave (SS/PBCH)

Структура кадру 5G NR визначена специфікацією 3GPP TS 38.211 [16] і є ієрархічною: радіокадр (Radio Frame) тривалістю 10 мс містить 10 підкадрів (Subframe) по 1 мс кожен. Кожен підкадр складається зі змінної кількості слотів залежно від нумерології: при μ=0 – 1 слот/підкадр, при μ=1 – 2 слоти, при μ=2 – 4 слоти тощо. Кожен слот містить 14 OFDM-символів (нормальний CP, Normal Cyclic Prefix) або 12 символів (розширений CP, лише для μ=2). Мінімальною одиницею ресурсу у частотно-часовій площині є PRB (Physical Resource Block), що містить 12 підносійних та 1 слот. Гнучка структура кадру NR також підтримує напівдуплексний та TDD (Time Division Duplex) режими

з конфігурованим чергуванням DL/UL символів у рамках одного слоту, що є принциповою перевагою для URLLC: mini-slot передача може початися у будь-якому символі без очікування наступного слоту.

Система управління якістю обслуговування (QoS) у 5G NR базується на концепції QoS Flow – найменшої одиниці диференційованої обробки трафіку. Характеристики параметрів наведено у таблиці 1.5. Кожен QoS Flow характеризується 5QI (5G QoS Identifier), що є числовим ідентифікатором, що однозначно визначає набір QoS-параметрів. Специфікація 3GPP TS 38.214 [17] визначає стандартизовані 5QI-значення для різних типів застосунків. Параметри QoS Flow включають: Resource Type (GBR – Guaranteed Bit Rate, Non-GBR або Delay-Critical GBR); Priority Level (числовий пріоритет від 1 до 127); Packet Delay Budget (PDB) – максимально допустима затримка між UE та P-CSCF; Packet Error Rate (PER) – допустима частота втрат пакетів; Averaging Window (для GBR потоків). Відображення 5QI на конкретні QoS Flow виконується SMF на підставі правил, отриманих від PCF. UPF застосовує відповідні правила пріоритизації та маркування DSCP у IP-заголовках.

Таблиця 1.5 – Стандартизовані 5QI значення та їх характеристики (вибрані)

5QI	**Resource Type**	**Пріоритет**	**PDB (мс)**	**PER**	**Приклад застосунку**
1	GBR	20	100	10^{-2}	VoLTE / VoNR (голосові дзвінки)
2	GBR	40	150	10^{-3}	Відеодзвінки (прямий потік)
3	GBR	30	50	10^{-3}	Реальний час ігор, V2X C2
5	Non-GBR	10	100	10^{-6}	IMS Signalling (SIP)
7	Non-GBR	70	100	10^{-3}	Відео (буферизоване потокове)

Продовження таблиці 1.5

80	Delay-Crit. GBR	68	10	10^{-6}	V2X сценарій 4, розширений датчик
82	Delay-Crit. GBR	19	10	10^{-4}	Дискретна автоматизація URLLC
85	Delay-Crit. GBR	21	5	10^{-5}	Інтелектуальна транспортна система ITS

Таблиця 1.5 демонструє, що система 5QI охоплює широкий спектр вимог до якості обслуговування – від традиційного голосового трафіку (5QI=1, PDB=100 мс) до критично важливих URLLC-застосунків (5QI=85, PDB=5 мс). У запропонованій вдосконаленій архітектурі SMF використовує значення 5QI як основний критерій при виборі рівня розміщення UPF: 5QI ≤ 82 спрямовується до Edge UPF, 5QI 3–7 – до Regional UPF, решта – до Central UPF.

Технологія Massive MIMO (Massive Multiple Input Multiple Output) є ключовим елементом 5G NR у діапазоні FR1. На відміну від традиційного MIMO у 4G LTE-A (максимум 8 антенних портів), Massive MIMO у 5G NR може використовувати від 32 до 256 і більше антенних елементів в одній базовій станції gNB. Наявність великої кількості антенних елементів дозволяє реалізувати просторово-часову обробку сигналів з формуванням вузьких напрямлених пучків (Beamforming) для кожного UE окремо, що суттєво підвищує спектральну ефективність та знижує міжелементні завади. Технологія просторового мультиплексування (Spatial Multiplexing) дозволяє одночасно передавати до 8 незалежних просторових потоків одному UE або обслуговувати кілька UE одночасно (MU-MIMO) у спільному частотно-часовому ресурсі рисунок 1.4. Стратегічну роль Massive MIMO у системах 5G

обґрунтовано у роботі [18], де показано, що при необмеженій кількості антен базової станції ефекти завмирання та внутрішньосистемних завад зникають, а потужність може бути зменшена до нескінченно малої величини. Теоретична верхня межа пропускної здатності радіоканалу з підтримкою Massive MIMO за умов просторового мультиплексування визначається саме за допомогою матричного розширення формули Шеннона (1.2):

$$C = B \cdot \log_2 \det\left(\mathrm{I}_M + \frac{SNR}{N}\mathrm{HH}^H\right)$$

де C – пропускна здатність каналу, біт/с; B – ширина смуги каналу, Гц; I_M – одинична матриця розміром M×M; SNR – відношення сигнал/шум; N – кількість антен UE; H – матриця каналу розміром M×N (M – кількість антен базової станції); H^H – ермітово-спряжена матриця H.

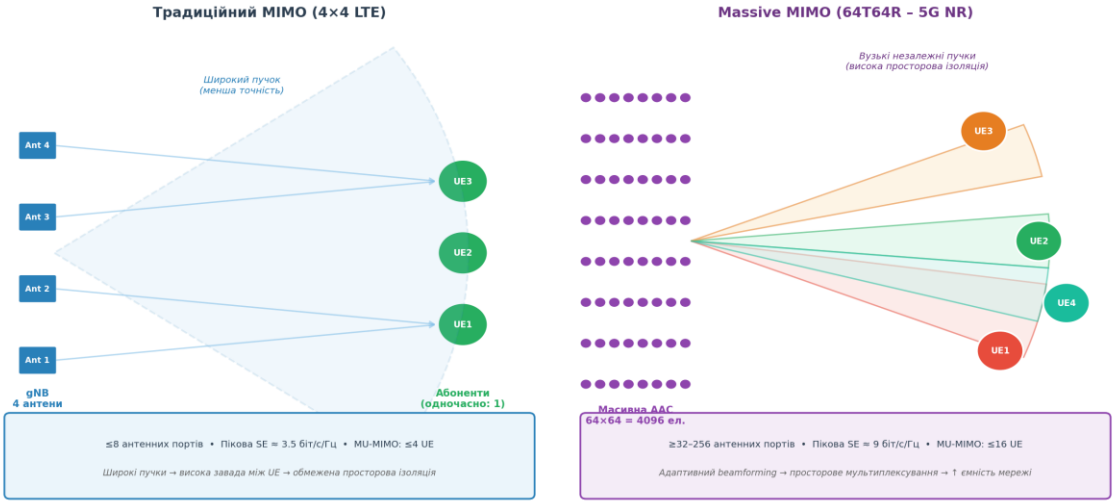

Рисунок 1.4 – Порівняння традиційного MIMO та Massive MIMO: beamforming та просторове мультиплексування у 5G NR

Канальне кодування в 5G NR відрізняється від LTE: для інформаційних каналів використовуються LDPC-коди (Low Density Parity Check), а для управляючих каналів – полярні коди (Polar Codes). Обидва типи кодів забезпечують кращу корекцію помилок при меншій обчислювальній складності порівняно з турбокодами LTE. Для мінімізації затримки у URLLC застосовується скорочена довжина HARQ (Hybrid Automatic Repeat Request) та mini-slot планування, при якому передача може початися у будь-якому символі слоту без очікування його початку. Модуляційні схеми NR

охоплюють QPSK, 16-QAM, 64-QAM та 256-QAM, обираючись адаптивно на основі поточних умов каналу (Link Adaptation).

1.4 Недоліки існуючої архітектури та обґрунтування необхідності вдосконалення

Попри значний прогрес у стандартизації та розгортанні 5G SA мереж, реальні комерційні реалізації демонструють суттєвий розрив між задекларованими специфікаційними цілями та практичними показниками. Аналіз технічних публікацій, звітів операторів та стандартизаційних документів дозволяє виділити кілька системних класів проблем, кожен з яких вимагає комплексного архітектурного вирішення. Базові вимоги безпеки визначені у специфікації 3GPP TS 33.501 [20], що встановлює механізми автентифікації, захисту цілісності та шифрування, однак практична реалізація цих вимог залишає місце для вдосконалення.

Перша та найбільш критична проблема – висока наскрізна затримка (End-to-End Latency). Стандарт IMT-2020 встановлює цільову затримку 1 мс для URLLC, тоді як реальні виміри в комерційних 5G SA мережах показують 8–25 мс. Основна причина – централізоване розміщення UPF у регіональних або національних ЦОД оператора: при відстані 50–100 км між gNB та ЦОД лише затримка поширення (propagation delay) складає 0,5–1 мс в один бік. Додаткову затримку вносять: обробка у функціях NF площини управління; затримка сигналізації при встановленні PDU-сесії (3–8 мс для ініціалізації); стека протоколів у gNB (RLC ARQ, HARQ-ретрансмісії) [20].

Друга проблема – обмежена масштабованість мережевих функцій. Переважна більшість існуючих реалізацій 5GC базується на монолітних Virtual Network Functions (VNF) на основі VM (Virtual Machines), де кожна NF реалізована як єдиний монолітний процес. Запуск нового VM-екземпляра займає від 2 до 8 хвилин, що унеможливлює оперативну реакцію на раптові стрибки трафіку (наприклад, при масових подіях). Крім того, stateful архітектура NF ускладнює горизонтальне масштабування: при збільшенні

навантаження не можна просто додати новий екземпляр без синхронізації стану існуючих екземплярів. Монографія [21] детально аналізує вразливості монолітних VNF та необхідність переходу до cloud-native підходу.

Третя проблема стосується безпеки. Хоча стандарт 3GPP TS 33.501 визначає широкий набір механізмів захисту, практичні реалізації мають кілька системних вразливостей. Передусім, SBI-інтерфейси між NF нерідко захищені лише базовим TLS без взаємної автентифікації (mutual TLS), що відкриває можливості для атак типу man-in-the-middle між NF. По-друге, концепція периметрової безпеки («довіреного периметру» ЦОД) є застарілою в умовах cloud-native розгортання з мікросервісами. По-третє, відкрита архітектура O-RAN вводить нові потенційні вектори атак через xApps та rApps третіх сторін у контролерах RIC. Нарешті, у режимі NSA успадковані вразливості EPC-ядра (Diameter-протокол, SS7) переносяться і на 5G-сегмент мережі.

Четверта проблема – недостатня ефективність Network Slicing. Попри визначення наскрізного нарізання у 3GPP Release 15, більшість комерційних реалізацій підтримують слайсинг лише на рівні 5GC, без реальної ізоляції ресурсів у NG-RAN та транспортній мережі. Це означає, що різні зрізи конкурують за радіоресурси та транспортну пропускну здатність, що унеможливлює надання гарантованих SLA для URLLC у випадку пікового навантаження eMBB-зрізу. Крім того, механізми автоматичного масштабування та управління lifecycle зрізів здебільшого відсутні, що збільшує операційні витрати.

П'ята проблема – висока енергоємність мереж 5G NR. Базові станції з Massive MIMO споживають у 3–5 разів більше електроенергії порівняно з еквівалентними базовими станціями LTE, що суттєво впливає на операційні витрати (OPEX) оператора та вуглецевий слід телекомунікаційної галузі. Адаптивні механізми управління живленням (Sleep Mode, Dynamic Spectrum Sharing) в сучасних реалізаціях не завжди оптимально налаштовані, що призводить до надмірних енерговитрат у непікові години.

Таблиця 1.6 – Класифікація недоліків базової архітектури 5G та шляхи їх вирішення

Клас проблеми	**Основні причини виникнення**	**Вплив на KPI**	**Напрям вирішення**
Висока затримка	Централізоване UPF, NSA-режим, надлишковий сигналінг NAS	E2E Latency 8–25 мс (ціль 1 мс)	MEC, ієрархічна UPF, оптимізація NAS
Слабка масштаб.	Монолітні VM-VNF, stateful архітектура NF	Надлишковий CAPEX, деградація при пікових навантаженнях	Cloud-native CNF, Kubernetes, stateless NF
Вразливість безпеки	Периметрова модель, відсутність mTLS, O-RAN-вектори	Ризик витоку даних абонентів	Zero Trust, mTLS, OAuth 2.0, sandboxing
Обмежений слайсинг	Відсутність E2E ізоляції RAN+Transport+Core	Конкуренція зрізів, порушення SLA URLLC	E2E Network Slicing, SDN-transport, Near-RT RIC
Висока енергоємн.	Неоптимальний Sleep Mode Massive MIMO, надлишкові NF	Зростання OPEX, вуглецевий слід	Адаптивний Sleep Mode (RIC xApp), CNF auto-scaling

Аналіз таблиці 1.6 підтверджує, що виявлені недоліки є системними та взаємопов'язаними.

Висновки

Оглядаючись на те, як розвивався мобільний зв'язок від перших аналогових систем до сьогоднішнього дня, стає цілком зрозуміло, що кожне нове покоління вимагало все глибшої перебудови самої логіки роботи мереж. Якщо в епоху 4G відбувся ключовий перехід до повністю пакетної передачі даних, то 5G приніс ще фундаментальнішу зміну — сервісно-орієнтовану архітектуру ядра. Тепер базові мережеві функції, такі як управління доступом абонентів, контроль їхніх сесій чи безпосередня обробка користувацького трафіку, працюють як незалежні сервіси, що вільно спілкуються між собою через програмні інтерфейси.

Радіочастина нового покоління також отримала безпрецедентну гнучкість. Завдяки використанню різних частотних діапазонів та адаптивного інтервалу між підносійними, система може динамічно підлаштовуватися під дуже різні потреби. Разом із системою ідентифікаторів пріоритетності трафіку це мало б гарантувати ідеальну роботу будь-яких, навіть найвибагливіших застосунків.

Але коли ми починаємо порівнювати ідеальні вимоги зі стандарту IMT-2020 із тим, що оператори реально побудували на практиці, оптимізм трохи згасає. Реальні виміри показують величезний системний розрив. Наприклад, для чутливих до часу сервісів (таких як безпілотні авто чи віддалена хірургія) затримка не повинна перевищувати одну мілісекунду, а по факту ми маємо від 8 до 25 мілісекунд у найкращому випадку. Головна проблема тут криється у надмірній централізації: шлюзи, які обробляють дані, часто знаходяться за десятки кілометрів від базових станцій, і фізику поширення сигналу просто неможливо обдурити.

Ситуацію значно погіршує і те, що мережеве обладнання досі спирається на громіздкі віртуальні машини. Якщо під час масового заходу навантаження на мережу різко зростає, на розгортання додаткових ресурсів може піти до восьми хвилин, що в сучасних реаліях є неприпустимо довго. Також буде доречним сюди додати застарілі підходи до мережевої безпеки, величезні

витрати електроенергії на роботу багатоантенних систем та відсутність реального, фізичного розділення мережі на ізольовані канали для різних типів сервісів. Усе це в комплексі дуже чітко показує: базова архітектура 5G має серйозні проблеми, які неможливо вилікувати простими налаштуваннями - потрібна кардинальна та продумана модернізація.

РОЗДІЛ 2
ВДОСКОНАЛЕННЯ АРХІТЕКТУРИ 5G МЕРЕЖ

2.1 Використання мережевої віртуалізації (NFV) та програмно-визначеного мережевого зв'язку (SDN)

Мережева функціональна віртуалізація (NFV – Network Functions Virtualization) та програмно-визначений мережевий зв'язок (SDN – Software-Defined Networking) є двома взаємодоповнюючими технологічними концепціями, що кардинально змінюють підхід до побудови та управління телекомунікаційними мережами. Архітектурна специфікація NFV закріплена у стандарті ETSI GS NFV 002 [22], що визначає функціональні блоки NFV-системи та інтерфейси між ними. Стандарт SDN-архітектури сформульовано у документі ONF TR-521 [23], що встановлює принципи відокремлення площин управління та даних.

Концепція NFV, вперше запропонована групою операторів у 2012 році, полягає в перенесенні мережевих функцій зі спеціалізованого фізичного обладнання (proprietary hardware appliances) на стандартні x86-сервери, що утворюють хмарну інфраструктуру NFVI (NFV Infrastructure). Мережеві функції, реалізовані у програмному вигляді, отримали назву VNF (Virtualized Network Functions). Три основні компоненти NFV-архітектури: NFVI – фізичні та програмні ресурси (обчислення, мережа, сховища) разом із шаром віртуалізації (гіпервізор або контейнерна платформа); VNF – програмна реалізація мережевої функції (наприклад, AMF, SMF, UPF 5GC); MANO (Management and Orchestration) – підсистема, що включає NFVO (NFV Orchestrator), менеджер VNF (VNFM) та менеджер інфраструктури (VIM, напр. OpenStack). Синергія між NFV/SDN та Network Slicing детально аналізується у роботі [24].

Сучасний хмарно-нативний підхід (Cloud-Native Network Functions, CNF) є еволюцією NFV, при якому VNF замінюються контейнерними мікросервісами під управлінням Kubernetes. CNF-підхід відповідає принципам

12-factor application: stateless дизайн основних процесів (стан зберігається у зовнішніх базах даних), горизонтальне масштабування через репліки Pod'ів, конфігурація через ConfigMaps та Secrets, логування через stdout/stderr до агрегаторів (ELK, Loki). Ключова перевага CNF перед VM-VNF – різке скорочення часу запуску нового екземпляра: від 2–8 хвилин (VM) до 5–15 секунд (Container). Це дозволяє системі оркестрації реагувати на зростання навантаження практично в реальному часі завдяки механізму Horizontal Pod Autoscaler (HPA) Kubernetes.

Програмно-визначений мережевий зв'язок (SDN) базується на принциповому розділенні площини управління (Control Plane) та площини даних (Data Plane). У традиційних мережевих пристроях логіка маршрутизації інтегрована апаратно та прив'язана до конкретного пристрою. SDN виносить цю логіку до централізованого програмного контролера (SDN Controller), що отримує глобальне уявлення про топологію та стан мережі і передає рішення про маршрутизацію до підлеглих пристроїв площини даних через протоколи OpenFlow або RESTCONF/NETCONF. У 5G мережах SDN відіграє ключову роль в управлінні транспортною мережею (fronthaul, midhaul, backhaul) та в координації UPF. SDN-контролер може динамічно переконфігуровувати маршрути залежно від навантаження конкретних зрізів, QoS-вимог та змін топології мережі, що є значно ефективнішим підходом порівняно зі статичними маршрутними таблицями.

Синергія NFV та SDN формує технологічну основу для реалізації всіх концепцій вдосконаленої архітектури 5G. NFV забезпечує гнучке розгортання, масштабування та управління lifecycle програмних NF, тоді як SDN забезпечує динамічне та централізоване управління транспортуванням трафіку між ними. Разом ці технології дозволяють: скоротити час впровадження нових послуг з місяців до годин; забезпечити автоматичне горизонтальне масштабування при зростанні навантаження; централізовано управляти складною розподіленою інфраструктурою; суттєво знизити операційні витрати OPEX завдяки автоматизації. Порівняння ключових характеристик традиційного підходу на

базі віртуальних машин (VM-VNF) та сучасного хмарно-орієнтованого підходу (CNF) наведено у таблиці 2.1.

Таблиця 2.1 – Порівняння VM-VNF та Cloud-Native CNF для реалізації NF 5GC

Характеристика	**VM-VNF (Virtual Machine)**	**CNF (Cloud-Native / Container)**
Час запуску екземпляра	2–8 хвилин (завантаження ОС)	5–15 секунд (pre-pulled container image)
Споживання ресурсів	Висока накладна (ОС, гіпервізор, ~1–4 ГБ RAM/VM)	Мінімальна (shared OS kernel, ~50–200 МБ/Pod)
Щільність розміщення	10–50 VM на сервер	100–500 контейнерів на сервер
Горизонт. масштабув.	Складне (stateful, синхроніз.)	Нативне (Kubernetes HPA, stateless design)
Управління станом	Вбудований у VM-процес	Зовнішнє сховище (Redis, Cassandra)
Оновлення (rollout)	Blue-Green, ручне керування	Rolling Update, Canary Deploy (Kubernetes)
Відновлення після відм.	Перезапуск VM (2–5 хв)	Kubernetes restartPolicy (< 30 с)
Стандарт управління	ETSI NFV MANO	Kubernetes + Helm + Service Mesh (Istio)

Таблиця 2.1 наочно демонструє переваги CNF-підходу над класичними VM-VNF за більшістю ключових характеристик. Особливо суттєвими для 5G-застосунків є різниця у часі запуску (5–15 с проти 2–8 хв) та нативна підтримка горизонтального масштабування, що є обов'язковою умовою для динамічного реагування на зміни трафіку у реальних 5G SA мережах.

Рисунок 2.1 демонструє, як NFV/SDN перетворює жорстко апаратну 5G-інфраструктуру на гнучкий програмний стек. Замість монолітних обладнаних блоків з'являється хмарна платформа з автоматизованим управлінням.

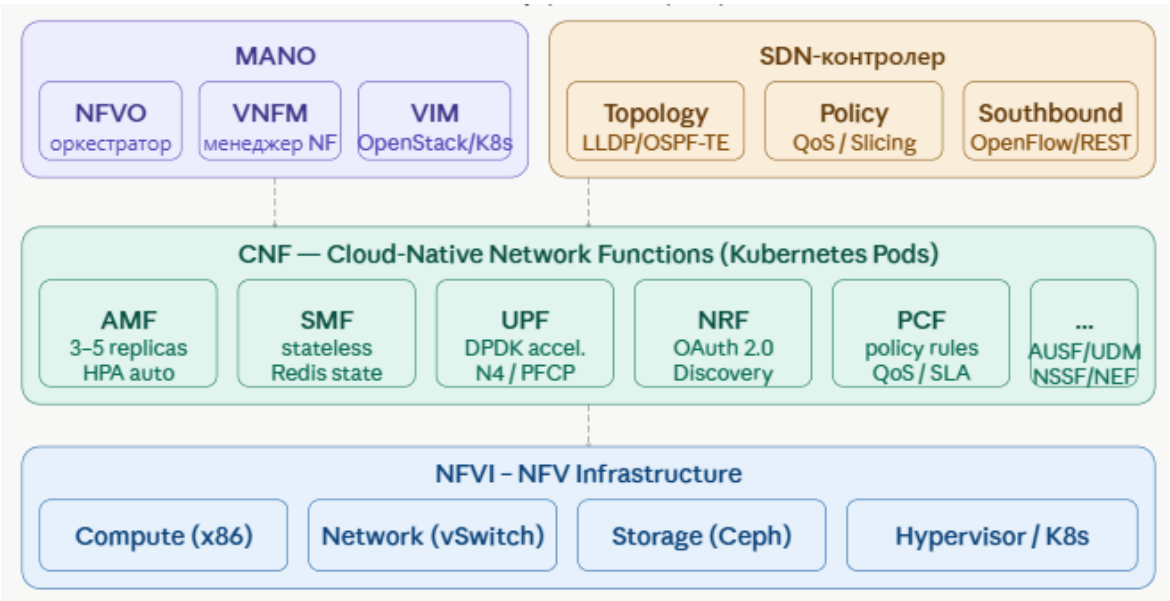

Рисунок 2.1 – Архітектура NFV/SDN у 5G: NFVI, VNF/CNF та SDN-контролер

2.2 Network Slicing як метод оптимізації ресурсів

Мережеве нарізання (Network Slicing) – одна з найбільш революційних концепцій 5G, що дозволяє розгортати на спільній фізичній інфраструктурі множинні логічно ізольовані «зрізи» (Network Slices). Кожен зріз є повноцінною наскрізною (E2E) логічною мережею з власним набором NF, виділеними ресурсами та параметрами QoS, оптимізованими під конкретний клас сервісу або групу клієнтів. Систематичний огляд Network Slicing у 5G представлений у роботі [25], де описано архітектурні принципи та ключові виклики практичного впровадження.

Стандарт 3GPP визначає три типових класи зрізів: eMBB (SST=1) для широкосмугових застосунків з максимальною пропускною здатністю; URLLC (SST=2) для критичних застосунків із гарантованою мінімальною затримкою та надійністю; mMTC (SST=3) для масових IoT-пристроїв із обмеженою пропускною здатністю та оптимізованим енергоспоживанням. Ідентифікація

зрізів здійснюється через S-NSSAI (Single Network Slice Selection Assistance Information), що складається з двох компонентів: SST (Slice/Service Type, 8 біт) та опціонального SD (Slice Differentiator, 24 біти). Дослідження у роботі [26] показує, що тільки після появи SD-компонента стала можлива підтримка десятків спеціалізованих зрізів одного типу в рамках однієї мережі оператора.

Управління вибором зрізів у 5GC реалізується через функцію NSSF. Під час реєстрації UE, AMF звертається до NSSF для визначення допустимих S-NSSAI на основі запитаного набору (Requested NSSAI) та профілю підписки (UDM). NSSF повертає Allowed NSSAI – набір S-NSSAI, дозволених у поточній зоні реєстрації. Для кожної PDU-сесії SMF визначається AMF на основі S-NSSAI, яким ідентифікується зріз, що обслуговує дану сесію. Концептуальні та практичні труднощі реалізації Network Slicing розглянуто у [27], де виявлено, що основним каменем спотикання є відсутність стандартизованих інструментів E2E оркестрації зрізів у комерційних реалізаціях.

Принципово важливим для забезпечення SLA є E2E ізоляція зрізів на трьох рівнях. Перший рівень – NG-RAN: NSSF надає AMF інформацію про виділені ресурси RAN (PRB, beamforming векторів) для кожного S-NSSAI. Near-RT RIC через xApp управляє розподілом PRB між активними зрізами з урахуванням їхніх QoS-вимог. Другий рівень – транспортна мережа: SDN-контролер резервує окремі VPN або MPLS-тунелі для кожного зрізу, гарантуючи ізоляцію трафіку та виділену пропускну здатність. Третій рівень – ядро 5GC: кожен зріз має власний набір SMF/UPF-екземплярів (або логічно ізольовані екземпляри shared AMF), що не допускають перетину трафіку між зрізами. Для гарантування ізоляції ресурсів між зрізами та виконання вимог SLA кожному зрізу виділяється пропускна здатність відповідно до його пріоритетної ваги. Розподіл пропускної здатності для і-го зрізу визначається за цим виразом (2.1):

$$R_{slice\_i} = \min\left(R_{max\_i}, \frac{w_i}{\sum_{j=1}^{K} w_j} \cdot R_{total}\right)$$

де R_slice_ i – виділена пропускна здатність для і-го зрізу; R_max_i – максимальна дозволена пропускна здатність і-го зрізу відповідно до SLA; w_i – пріоритетна вага і-го зрізу; K – загальна кількість активних зрізів у системі; w_j – вага j-го зрізу; R_total – загальна доступна пропускна здатність ресурсу. Функція min() гарантує, що жоден зріз не перевищить максимально дозволений рівень обслуговування, навіть якщо його розрахована частка є більшою.

2.3 Інтеграція Edge Computing у 5G інфраструктуру

Граничне обчислення (MEC – Multi-access Edge Computing) передбачає розміщення обчислювальних ресурсів та прикладних сервісів поблизу кінцевих користувачів, на рівні базових станцій або точок агрегації трафіку. Концепція MEC спрямована на радикальне зниження наскрізної затримки шляхом усунення необхідності передачі трафіку до центрального ЦОД для кожного запиту. Огляд архітектур MEC та їх інтеграції у 5G мережі, включно з описом стандартів ETSI MEC, представлений у роботі [28].

Архітектурна інтеграція MEC у 5G реалізується через функцію UPF та механізм ULCL (Uplink Classifier). Гранична UPF (Edge UPF, E-UPF), розміщена поблизу gNB, аналізує висхідний трафік UE та перенаправляє пакети, призначені для локальних MEC-сервісів, безпосередньо до MEC-платформи, не передаючи їх до центральних UPF. Це дозволяє досягти затримки 1–3 мс для сервісів, що виконуються на Edge UPF. SMF управляє конфігурацією ULCL через протокол PFCP, передаючи правила класифікації трафіку (Traffic Detection Rules) відповідним UPF-екземплярам [29].

Ієрархія MEC-вузлів у вдосконаленій архітектурі включає три рівні розміщення. Near-Edge вузли (при gNB або невеликих кластерах базових станцій) забезпечують затримку 1–3 мс і є оптимальними для сервісів V2X та URLLC-застосунків промислової автоматизації. Far-Edge вузли (регіональний рівень агрегації, покриття 10–50 gNB) забезпечують затримку 3–8 мс та більші обчислювальні можливості – підходять для відеоаналітики та AR-додатків.

Central-Edge вузли (регіональний ЦОД оператора) з затримкою 8–15 мс надають значні обчислювальні ресурси для CDN, ML-inference та інших застосунків без жорстких вимог до затримки. Повна класифікація MEC-випадків використання та їх вимоги до пропускної здатності і затримки систематизовані у роботі [30]. Загальна наскрізна затримка в гібридній архітектурі з MEC визначається як зважена сума компонент локальної та дистанційної обробки (2.2):

$$T_{total} = (1 - \alpha)T_{local} + \alpha(T_{trans} + T_{edge})$$

де T_total – загальна наскрізна затримка, мс; α ∈ [0, 1] – коефіцієнт розвантаження трафіку на рівень MEC (частка трафіку, що обробляється на рівні Edge); T_local – затримка обробки локально на Edge-вузлі, мс; T_trans – затримка транспортування між Edge-вузлом та централізованим ЦОД, мс; T_edge – затримка обробки в централізованому ЦОД, мс. При α – 1 весь трафік обробляється локально, що забезпечує мінімальну затримку для URLLC-сервісів.

Типові застосунки MEC у 5G охоплюють широкий спектр вертикальних галузей. У сфері транспорту та Smart City MEC підтримує V2X (Vehicle-to-Everything) комунікації з вимогами до затримки < 5 мс та надійністю 99,999% для систем запобігання зіткненням. У промисловому секторі (Industry 4.0) MEC забезпечує замкнуті контури управління роботизованими системами з затримкою < 2 мс. Відеоаналітика в реальному часі на рівні MEC дозволяє обробляти потоки 4K-відео безпосередньо на рівні мережевого краю замість передачі до центрального ЦОД, що значно знижує навантаження на транспортну мережу. Комплексний розгляд MEC-сценаріїв для розумних міст із аналізом вимог до затримки та пропускної здатності представлений у роботі [31].

Таблиця 2.2 – Типові MEC-застосунки у 5G та їх вимоги до QoS

Застосунок / вертикаль	**Макс. затримка**	**Мін. пропускна здатність**	**Надійність**	**Рівень MEC**
V2X (запобіг. зіткненням)	< 3 мс	~1 Мбіт/с	99,999 %	Near-Edge (при gNB)
Пром. роботизований пристрій	< 2 мс	10–100 Мбіт/с	99,999 %	Near-Edge
AR/VR рендеринг	< 7 мс	100–500 Мбіт/с	99,9 %	Far-Edge
Відеоаналітика 4K	< 20 мс	10–50 Мбіт/с	99,9 %	Far-Edge
CDN / кешування	< 50 мс	1–10 Гбіт/с	99,5 %	Central-Edge
ML Inference (edge AI)	< 30 мс	50–200 Мбіт/с	99,5 %	Far/Central-Edge
Дистанційна хірургія	< 1 мс	> 100 Мбіт/с	99,999 %	Near-Edge

Аналіз таблиці 2.2 підтверджує, що ієрархічна модель MEC є необхідністю, а не опцією: неможливо забезпечити вимоги до затримки для V2X та промислових застосунків з єдиного централізованого ЦОД. Лише розміщення UPF та обчислювальних ресурсів на рівні Near-Edge дозволяє досягти 1–3 мс, необхідних для критичних застосунків URLLC-класу.

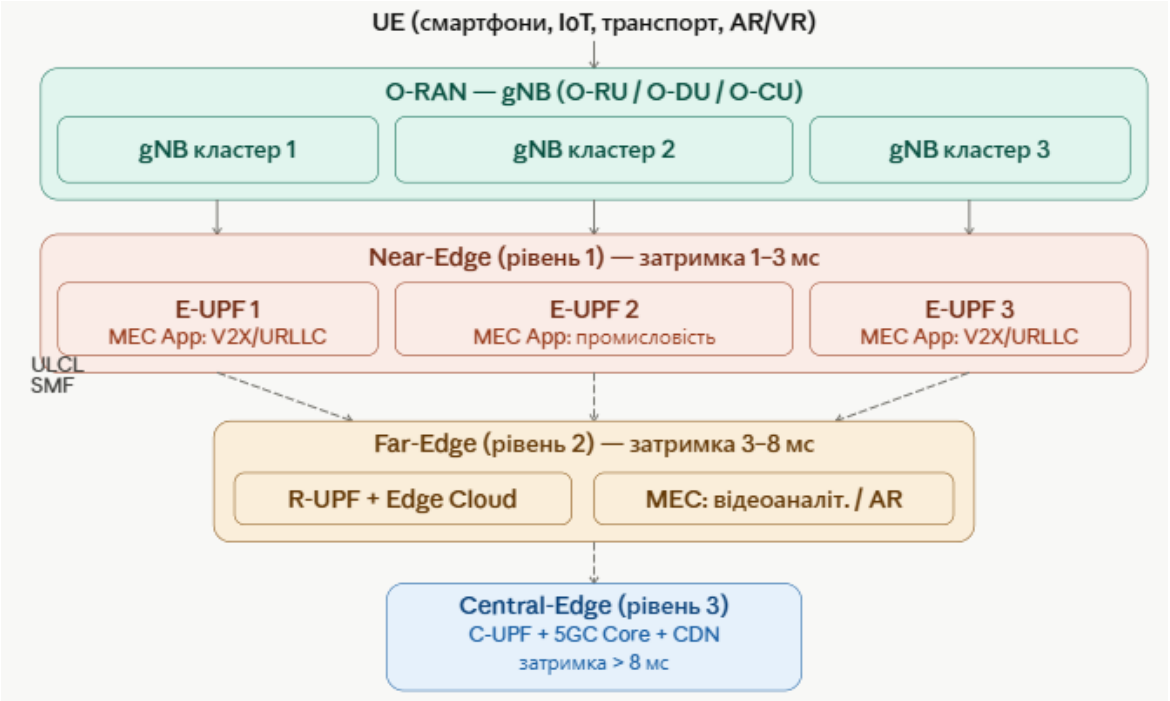

Рисунок 2.2 – Ієрархічна модель MEC у 5G: Near-Edge, Far-Edge та Central-Edge

Рисунок 2.2 ілюструє ієрархічну модель MEC, в якій SMF динамічно обирає рівень розміщення UPF залежно від параметрів QoS PDU-сесії. Це є ключовим механізмом зниження затримки у вдосконаленій архітектурі.

2.4 Аналіз та вибір методів для побудови вдосконаленої архітектури

Архітектура O-RAN (Open Radio Access Network), визначена консорціумом O-RAN Alliance у специфікації [32], є ще одним ключовим методом вдосконалення, що забезпечує відкритість та гнучкість RAN-сегменту. O-RAN вводить стандартизовані відкриті інтерфейси (O1, O2, A1, E2) між компонентами RAN та контролерами, що дозволяє розгортати мультивендорне обладнання та виконувати ML-керовану оптимізацію радіоресурсів у реальному часі. Ключовим компонентом O-RAN є RIC (RAN Intelligent Controller) – контролер, що виконує оптимізаційні алгоритми (xApps) для управління RAN. Near-RT RIC забезпечує управління в діапазоні 10 мс – 1 с, Non-RT RIC – понад 1 с для довготермінових оптимізацій.

Для обґрунтованого вибору комплексу методів вдосконалення проведемо систематичне порівняння за шістьма критеріями. Критерій 1 –

вплив на наскрізну затримку (оцінка від 1 до 5, де 5 – найкращий вплив). NFV/CNF: 3/5 (оптимізує масштабування NF, але безпосередньо не знижує затримку маршрутизації); SDN: 4/5 (оптимізує маршрутизацію транспортної мережі); Network Slicing: 4/5 (ізоляція ресурсів гарантує затримку URLLC); MEC: 5/5 (ключовий механізм зниження затримки); O-RAN: 4/5 (оптимізація радіоресурсів знижує радіо-затримку). Критерій 2 – масштабованість: NFV/CNF 5/5, SDN 4/5, Network Slicing 4/5, MEC 3/5, O-RAN 5/5. Критерій 3 – ступінь стандартизації: NFV/CNF 5/5 (ETSI), SDN 4/5 (ONF), Network Slicing 5/5 (3GPP), MEC 4/5 (ETSI MEC), O-RAN 4/5 (O-RAN Alliance). Критерій 4 – зниження OPEX: NFV/CNF 5/5, SDN 4/5, Network Slicing 4/5, MEC 3/5, O-RAN 5/5. Аналіз підтверджує, що жоден з методів не є переважаючим за всіма критеріями, що обґрунтовує необхідність комплексного підходу, представленого у таблиці 2.3.

Таблиця 2.3 – Порівняльна оцінка методів вдосконалення архітектури 5G за шістьма критеріями

Метод/технологія	**Затримка**	**Масштаб.**	**Стандарт.**	**Складність**	**Безпека**	**OPEX ↓**
NFV/CNF (Kubernetes)	3/5	5/5	5/5	3/5	3/5	5/5
SDN (OpenFlow)	4/5	4/5	4/5	3/5	4/5	4/5
Network Slicing (E2E)	4/5	4/5	5/5	4/5	5/5	4/5
MEC (Edge UPF)	5/5	3/5	4/5	3/5	3/5	4/5
O-RAN (Near-RT RIC)	4/5	5/5	4/5	4/5	3/5	5/5
Сума балів	20	21	22	17	18	22

За результатами таблиці 2.3 для побудови вдосконаленої архітектури обрано комплекс п'яти методів: cloud-native NFV/CNF з Kubernetes, SDN-управління транспортом, E2E Network Slicing, ієрархічний MEC із динамічним розміщенням UPF та O-RAN з Near-RT RIC. Кожен метод усуває конкретний клас недоліків, виявлених у розділі 1, а їх спільне застосування забезпечує синергетичний ефект підвищення ефективності.

Висновки

У другому розділі проведено детальний аналіз методів вдосконалення архітектури 5G мереж та обґрунтовано вибір комплексу технологій для розробки вдосконаленої архітектури. Досліджено поняття мережевої функціональної віртуалізації (NFV) та програмно-визначеного мережевого зв’язку (SDN). Встановлено, що перехід від монолітних VM-VNF до хмарно-нативних контейнерних мікросервісів (CNF) під управлінням Kubernetes дозволяє скоротити час запуску нового екземпляра мережевої функції з 2-8 хвилин до 5-15 секунд, забезпечити нативне горизонтальне масштабування через механізм HPA та знизити споживання апаратних ресурсів завдяки спільному використанню системного ядра. SDN-контролер, відокремлюючи площину управління від площини даних, надає централізований і динамічний контроль над транспортною мережею з часом реагування на топологічні зміни менш ніж 100 мс. Розглянуто концепцію E2E Network Slicing як метод гарантованої ізоляції ресурсів для різних класів сервісів. Показано, що повноцінна ізоляція зрізів потребує одночасного застосування на трьох рівнях: NG-RAN (розподіл PRB через Near-RT RIC) транспортній мережі (окремі MPLS/SR або VXLAN-тунелі через SDN-контролер) та ядрі 5GC (виділені або логічно ізольовані екземпляри SMF/UPF). Лише за умови E2E ізоляції можливе гарантоване дотримання SLA для URLLC-зрізу при пікових навантаженнях eMBB. Проаналізовано інтеграцію граничного обчислення (MEC) у 5G інфраструктуру. Обґрунтовано трирівневу ієрархію MEC-вузлів: Near-Edge (затримка 1-3 мс, для V2X та промислових URLLC-застосунків),

Far-Edge (3-8 мс, для AR/VR та відеоаналітики) та Central-Edge (8-15 мс, для CDN та ML-інференсу). Встановлено, що механізм ULCL (Uplink Classifier) у SMF є ключовим інструментом динамічного маршрутизування трафіку між рівнями UPF на основі значень 5QI.

За результатами багатокритеріального порівняльного аналізу п’яти методів за шістьма критеріями (вплив на затримку, масштабованість, ступінь стандартизованості, складність впровадження, безпека, зниження OPEX) встановлено, що жоден з методів окремо не забезпечує усунення всіх класів виявлених недоліків. Обрано комплекс із п’яти взаємодоповнюючих методів: cloud-native NFV/CNF (Kubernetes HPA/VPA), SDN-управління транспортом, E2E Network Slicing з ізоляцією на трьох рівнях, ієрархічний MEC із трирівневою UPF та відкрита архітектура O-RAN з інтелектуальними контролерами RIC. Синергія цих методів забезпечує системне усунення всіх п’яти класів недоліків базової архітектури, виявлених у розділі 1, та формує технологічну основу для розробки вдосконаленої архітектури у розділі 3.

РОЗДІЛ 3
РОЗРОБКА ВДОСКОНАЛЕНОЇ АРХІТЕКТУРИ 5G МЕРЕЖ

3.1 Концепція та принципи побудови вдосконаленої архітектури

Вдосконалена архітектура 5G мережі, що розробляється в рамках даної дипломної роботи, базується на комплексному використанні методів, визначених у розділі 2, та спрямована на усунення системних недоліків, виявлених у розділі 1. Концепція побудови архітектури включає п'ять ключових принципів, що формують єдину взаємопов'язану систему.

Перший принцип – cloud-native розгортання NF. Усі функції ядра 5GC реалізуються у вигляді cloud-native network functions (CNF) – контейнерних мікросервісів під управлінням Kubernetes. Кожна NF розпоряджається власним набором Pod'ів, конфігурується через Kubernetes ConfigMap та Secret, розгортається через Helm Chart. Принциповим вибором є stateless-архітектура NF: поточний стан сесій (PDU Session Contexts, UE Contexts) зберігається у зовнішньому розподіленому кеш-сховищі (Redis Cluster або Apache Cassandra), а не всередині Pod'у. Це дозволяє Kubernetes HPA (Horizontal Pod Autoscaler) додавати або видаляти Pod-екземпляри без порушення поточних сесій UE.

Другий принцип – E2E Network Slicing із автоматичним управлінням lifecycle. Система оркестрації зрізів виступає центральним координатором, що управляє зрізами на всіх трьох рівнях одночасно. Для RAN-рівня оркестратор взаємодіє з Near-RT RIC через інтерфейс O1, резервуючи пул PRB для кожного зрізу. Для транспортного рівня оркестратор взаємодіє з SDN-контролером через NorthBound API, встановлюючи ізольовані LSP (Label Switched Path) або VXLAN VNI для кожного зрізу. Для рівня 5GC оркестратор через NFVO розгортає специфічні для зрізу NF-компоненти (SMF, UPF). Час розгортання нового зрізу від ініціювання запиту до готовності до обслуговування трафіку – 2–5 хвилин.

Третій принцип – ієрархічна розподілена архітектура MEC з трирівневим розміщенням UPF. Функція UPF реалізована на трьох рівнях: Edge UPF (E-UPF) при або поблизу gNB – обслуговує URLLC-зрізи та MEC-застосунки з вимогами до затримки < 5 мс; Regional UPF (R-UPF) на регіональних вузлах агрегації – обслуговує eMBB-зрізи та MEC-застосунки з затримкою 3–8 мс; Central UPF (C-UPF) у регіональному ЦОД – обслуговує mMTC-трафік та сервіси без жорстких вимог до затримки. SMF реалізує алгоритм динамічного вибору UPF на основі параметра 5QI (5G QoS Identifier), поточного завантаження UPF та доступності MEC-застосунків.

Четвертий принцип – відкрита RAN на базі O-RAN. Радіодоступна мережа побудована відповідно до специфікацій O-RAN Alliance: O-RU забезпечує аналогово-цифрове перетворення; O-DU реалізує нижній стек (PHY/MAC/RLC); O-CU реалізує верхній стек (PDCP/RRC/SDAP). Near-RT RIC виконує xApps для управління ресурсами в діапазоні 10 мс – 1 с: RRM-xApp (Radio Resource Management) динамічно розподіляє PRB між зрізами та UE; HO-xApp (Handover Optimization) приймає рішення про хендовер з урахуванням поточного завантаження та QoS-профілю UE. Non-RT RIC виконує rApps для довготермінової оптимізації: прогнозування трафіку за допомогою LSTM-мереж та попереднє масштабування ресурсів.

Інтерфейси O-RAN відіграють ключову роль у забезпеченні відкритості та гнучкості архітектури. Інтерфейс E2 між Near-RT RIC та O-DU/O-CU є найважливішим: він передає звіти вимірювань від O-DU до Near-RT RIC (E2SM-KPM – KPM Subscription Management) та дозволяє Near-RT RIC надсилати управляючі повідомлення до O-DU (E2SM-RC – RAN Control). Інтерфейс A1 між Non-RT RIC та Near-RT RIC передає ML-моделі, A1-policy (правила від Non-RT до Near-RT RIC для стратегічного керування) та енічні визначення цілей оптимізації. Інтерфейс O1 між MANO-системою та всіма O-RAN компонентами забезпечує уніфіковану конфігурацію (NETCONF/YANG), збір Performance Measurements (PM) та управління відмовами (FM). Інтерфейс O2 між O-RAN Service Management & Orchestration

та NFVI/Cloud-платформою дозволяє оркестратору керувати розгортанням O-DU та O-CU як NFV-аплікацій на хмарній інфраструктурі.

П'ятий принцип – Zero Trust Security. Відмова від концепції «довіреного периметру» означає, що кожне звернення між NF повинне бути автентифіковано та авторизовано незалежно від мережевого розташування. Реалізація: Service Mesh (Istio) автоматично забезпечує mTLS 1.3 між усіма Pod-Pod комунікаціями; NRF виступає Authorization Server, видаючи OAuth 2.0 JWT Access Tokens при доступі NF до сервісів інших NF; Kubernetes NetworkPolicy встановлює правило deny-all з явними дозволами лише для легітимних взаємодій; xApp Sandbox у Near-RT RIC ізолює сторонні xApps від прямого доступу до мережевої інфраструктури.

3.2 Структурна схема запропонованої архітектури

Запропонована вдосконалена архітектура 5G мережі організована у вигляді п'ятирівневої ієрархічної структури, кожен рівень якої виконує специфічні функції та взаємодіє з суміжними через стандартизовані інтерфейси. Схема повністю відповідає стандартам 3GPP, ETSI, O-RAN Alliance та ITU-R, забезпечуючи сумісність з реальними комерційними реалізаціями. Інтегровану архітектуру, що об'єднує всі принципи, відображено на рисунку 3.1.

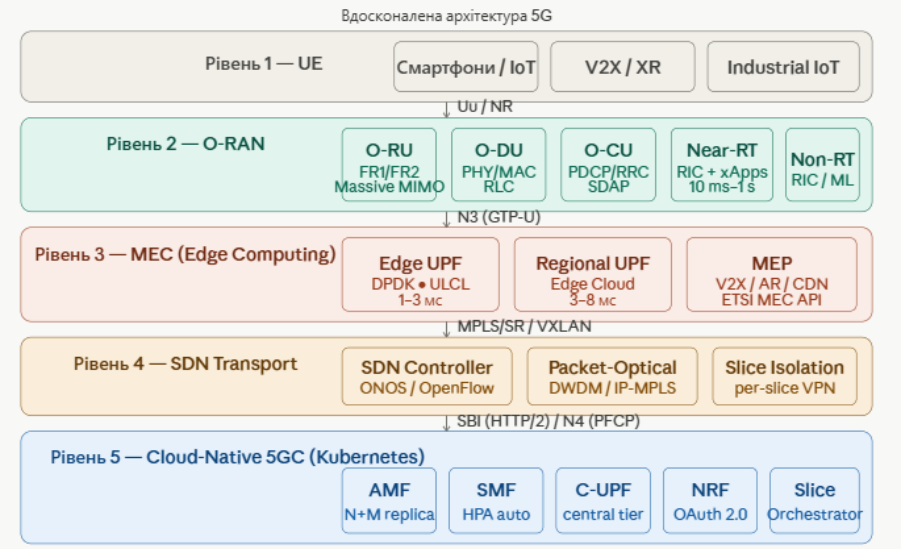

Рисунок 3.1 – Структурна схема вдосконаленої архітектури 5G мережі

Рисунок 3.1 відображає інтегровану архітектуру, що об'єднує всі п'ять принципів. Стрілки між рівнями позначають стандартизовані інтерфейси: між Рівнем 2 (O-RAN) та Рівнем 3 (MEC) – інтерфейс N3 (GTP-U) між gNB та E-UPF; між Рівнями 3 та 4 – SDN-управляємі пакетні тунелі; між Рівнями 4 та 5 – інтерфейс N4 (PFCP) між SMF та UPF.

Рівень 1 – рівень пристроїв (UE Layer). Включає чотири основних класи пристроїв: смартфони та мобільні термінали (eMBB); промислові датчики та актуатори (mMTC/URLLC); транспортні засоби (URLLC, V2X); XR-пристрої – AR/VR-гарнітури та мобільні облачні ігрові термінали (eMBB з низькою затримкою). Усі пристрої підтримують режим SA та можуть одночасно брати участь у кількох зрізах мережі (Multi-slice UE) відповідно до ієрархії S-NSSAI у профілі підписки UDM.

Рівень 2 – рівень відкритого радіодоступу (O-RAN Layer). O-RU (O-RAN Radio Unit) – апаратний радіомодуль з активною антенною системою (AAS) для Massive MIMO, що реалізує аналоговий beamforming та нижню частину фізичного рівня. O-DU (O-RAN Distributed Unit) – реалізує верхній PHY, MAC та RLC. O-CU (O-RAN Centralized Unit), розподілений на O-CU-CP та O-CU-UP, реалізує PDCP, RRC та SDAP. Near-RT RIC керує O-DU та O-CU через інтерфейс E2, виконуючи xApps для управління PRB, хендовером та модуляцією. Non-RT RIC через інтерфейс A1 передає до Near-RT RIC навчені ML-моделі та довготермінові конфігурації. Оркестрація усього O-RAN через інтерфейс O1 від MANO.

Рівень 3 – рівень мережевого краю (MEC Layer). Edge UPF (E-UPF), розміщена при O-DU або в межах 1–5 км від gNB, обслуговує URLLC-зрізи та V2X-трафік. На тому ж вузлі розміщується MEP (MEC Platform) за специфікацією ETSI MEC, що надає API для MEC-застосунків: Service Registry API, Traffic Rules Control API, Location API тощо. Regional UPF (R-UPF) на регіональному Edge Cloud обслуговує eMBB-зрізи та ресурсомісткі MEC-застосунки (відеоаналітика, AR rendering). ULCL (Uplink Classifier) логіка

SMF автоматично маршрутизує трафік між рівнями UPF на основі правил Traffic Detection.

Рівень 4 – транспортний рівень (SDN Transport Layer). SDN-контролер (наприклад, ONOS або OpenDaylight) отримує топологічні дані від усіх транспортних пристроїв через протоколи LLDP (Link Layer Discovery Protocol) та OSPF-TE (Traffic Engineering), формуючи глобальну карту транспортної мережі. На основі вимог активних зрізів, отриманих від оркестратора зрізів, SDN-контролер програмує пакетно-оптичні комутатори через OpenFlow та RESTCONF, встановлюючи ізольовані шляхи для кожного зрізу через MPLS/SR або VXLAN. При зміні топології (відмова каналу, перевантаження) SDN-контролер автоматично перепрограмовує альтернативні маршрути із часом реакції < 100 мс.

Взаємодія SMF та UPF через інтерфейс N4 реалізується протоколом PFCP (Packet Forwarding Control Protocol, RFC 7315). Через N4 SMF передає до UPF правила обробки пакетів (Packet Detection Rules, PDR), правила переадресації (Forwarding Action Rules, FAR), правила QoS (QER) та правила звітування (Usage Reporting Rules, URR). PDR визначають критерії класифікації вхідних пакетів (Source Interface, TEID у GTP-U тунелі, UE IP-адреса), FAR визначають дію при спрацюванні PDR (пересилання до NG-RAN або DN, буферизація). Завдяки PFCP SMF може динамічно змінювати правила на UPF без повторного встановлення PDU-сесії, що є основою для механізму Path Switching при хендовері та зміні рівня UPF у вдосконаленій архітектурі.

Детальна специфікація взаємодії між рівнями вдосконаленої архітектури наведена у таблиці нижче.

Таблиця 3.1 – Стандартизовані інтерфейси між рівнями вдосконаленої архітектури

Інтерфейс	**Між рівнями**	**Протокол**	**Функція**	**Стандарт**
F1-U/F1-C	RU↔DU	eCPRI (F1-U), SCTP (F1-C)	Розподіл PHY/MAC між O-RU та O-DU	3GPP TS 38.473
E2	DU/CU↔RIC	E2AP (SCTP)	Звіти KPM, управління RRM, HO-рішення	O-RAN WG3 E2AP
A1	Non-RT↔Near-RT RIC	HTTP/2 REST	Передача ML-моделей та A1-policy	O-RAN WG2 A1AP
O1	MANO↔O-RAN	NETCONF/YANG	Конфігурація, PM, FM-управління	O-RAN WG10 OAM
N2	gNB↔AMF	NGAP (SCTP)	UE-реєстрація, NAS-транспорт, хендовер	3GPP TS 38.413
N3	gNB↔UPF	GTP-U (UDP)	Тунелювання UP-трафіку між RAN та UPF	3GPP TS 29.281
N4	SMF↔UPF	PFCP (UDP)	PDR/FAR/QER/URR правила для UPF	3GPP TS 29.244
SBI	NF↔NF	HTTP/2 + JSON (SCP)	NF-to-NF сервісні запити через SBI/SCP	3GPP TS 23.501

Рівень 5 – рівень хмарного ядра (Cloud-Native 5GC Layer). Усі NF розгорнуті як CNF у Kubernetes-кластері регіонального ЦОД оператора.

Kubernetes Namespace забезпечує ізоляцію між зрізами на рівні оркестрації. Підсистема безпеки Zero Trust Policy Engine реалізована через Istio Service Mesh, що автоматично забезпечує mTLS для всіх Pod-Pod взаємодій. Моніторинговий стек (Prometheus + Grafana + Alertmanager) збирає метрики з усіх CNF з інтервалом 15 с та автоматично ескалює аларми при відхиленні KPI від SLA. Логіку вибору рівня розміщення UPF (SMF ULCL) деталізує рисунок 3.2.

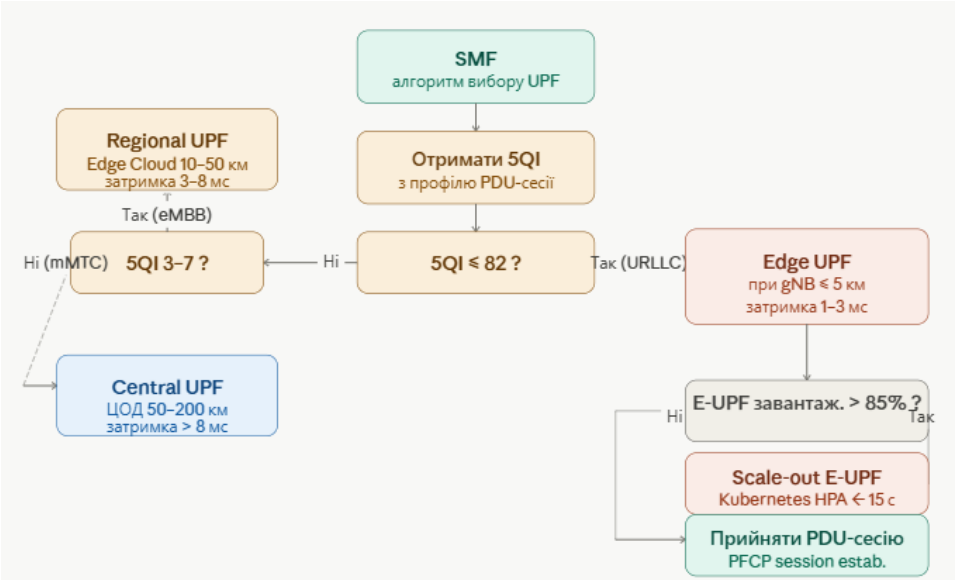

Рисунок 3.2 – Ієрархія UPF та логіка вибору рівня розміщення (SMF ULCL)

Рисунок 3.2 деталізує алгоритм вибору рівня UPF у SMF. Критерієм вибору є поєднання значення 5QI та поточного завантаження відповідного UPF-рівня. Якщо E-UPF завантажений понад 85%, SMF може «відхилити» нову URLLC-сесію до R-UPF (якщо затримка R-UPF задовольняє вимоги 5QI) або ініціювати масштабування E-UPF-екземплярів.

3.3 Реалізація механізмів підвищення ефективності в запропонованій архітектурі

Підвищення ефективності вдосконаленої архітектури досягається через реалізацію чотирьох взаємопов'язаних механізмів: динамічного автомасштабування NF, ML-підсиленого управління зрізами, адаптивного

управління радіоресурсами та відмовостійкості. Комплексна 5G E2E платформа для тестування подібних рішень описана у роботі [33].

Механізм динамічного автомасштабування NF базується на інтеграції Kubernetes HPA (Horizontal Pod Autoscaler) та VPA (Vertical Pod Autoscaler) зі стеком моніторингу Prometheus. HPA-контролер моніторить три типи метрик для прийняття рішень про масштабування: CPU-утилізацію кожного Pod'у (target: 70%); пам'ять (target: 80%); кастомні метрики NF, що публікуються через Prometheus-exporter (наприклад, кількість активних PDU-сесій для SMF, пакети на секунду для UPF). При перевищенні порогового значення будь-якої метрики HPA автоматично збільшує кількість Pod'ів відповідної NF. Час масштабування (від виявлення порогу до готовності нового Pod'у) – 5–15 с. При зниженні навантаження HPA поступово зменшує кількість Pod'ів з затримкою 5 хвилин (cooldown period), щоб запобігти “хаотичному” масштабуванню. Архітектурні підходи до розподілених обчислень та управління ресурсами на рівні мережевого краю досліджені у [34]. Алгоритм функціонування Kubernetes HPA визначає необхідну кількість реплік (Pod'ів) NF на основі співвідношення поточних і цільових значень контрольованих метрик. Розрахунок цільової кількості реплік здійснюється за формулою (3.1):

$$R_{target} = \left\lceil R_{current} \times \frac{M_{current}}{M_{target}} \right\rceil$$

де R_target – розрахована цільова кількість реплік Pod'ів NF; R_current – поточна кількість активних реплік; M_current – поточне значення метрики (наприклад, CPU-утилізація або кількість PDU-сесій); M_target – порогове (цільове) значення метрики, задане у конфігурації HPA (наприклад, 70% CPU). При M_current > M_target алгоритм збільшує кількість реплік пропорційно перевищенню порогу; при M_current < M_target – зменшує з урахуванням cooldown-інтервалу (5 хвилин) для запобігання нестабільному масштабуванню.

Механізм ML-підсиленого управління зрізами реалізований на рівні Non-RT RIC. Модель прогнозування трафіку, побудована на архітектурі LSTM

(Long Short-Term Memory), аналізує часові ряди навантаження кожного зрізу за попередні 30 діб та будує прогноз на 30–60 хвилин вперед. На основі прогнозу оркестратор зрізів завчасно ініціює превентивне масштабування ресурсів: якщо прогнозоване навантаження перевищить 70% поточних виділених ресурсів, NFVO розгортає додаткові NF-екземпляри зрізу ще до настання піку. Точність прогнозування LSTM-моделі для регулярних трафікових шаблонів становить 85–93% (за метрикою MAPE), що дозволяє суттєво скоротити кількість порушень SLA порівняно з реактивним підходом. Компоненти Near-RT RIC, що реалізують інтелектуальне управління RAN, систематизовані у таблиці 3.2.

Таблиця 3.2 – Перелік xApps у Near-RT RIC вдосконаленої архітектури та їх часові обмеження

xApp	**Часовий діапазон**	**Інтерфейс E2SM**	**Функція у вдосконаленій архітектурі**
RRM-xApp	10 мс – 100 мс	E2SM-RC	Динамічний розподіл PRB між зрізами та UE
HO-xApp	100 мс – 1 с	E2SM-RC + KPM	Оптимізація рішень про хендовер (A3/A5)
QoS-Monitor xApp	10 мс – 500 мс	E2SM-KPM	Моніторинг дотримання SLA для кожного зрізу
Slice-PRB Guard xApp	10 мс – 50 мс	E2SM-RC	Захист мін. гарантованого PRB URLLC-зрізу
Energy-Save xApp	100 мс – 1 с	E2SM-RC	Перемикання O-RU у Sleep Mode при малому трафіку
Anomaly-Detect xApp	100 мс – 1 с	E2SM-KPM	Виявлення аномального трафіку / RAN атак

Таблиця 3.2 систематизує xApp-компоненти Near-RT RIC, що реалізують інтелектуальне управління RAN у вдосконаленій архітектурі. Кожен xApp функціонує у власному ізольованому namespace із суворо визначеним часовим діапазоном реакції. Ключовим є Slice-PRB Guard xApp, що гарантує мінімальний виділений пул PRB для URLLC-зрізу навіть при екстремальному навантаженні eMBB, тим самим забезпечуючи ізоляцію зрізів на рівні RAN.

Механізм адаптивного управління радіоресурсами реалізований через RRM-xApp у Near-RT RIC. xApp отримує від кожного O-DU через інтерфейс E2 звіти вимірювань (RRC Measurement Reports), що містять: RSRP (Reference Signal Received Power) від обслуговуючого та сусідніх gNB; SINR (Signal to Interference and Noise Ratio) для оцінки якості каналу; BLER (Block Error Rate) та кількість HARQ-ретрансмісій. На основі цих метрик RRM-xApp виконує три функції оптимізації: управління PRB між зрізами – виділення мінімально гарантованого пулу PRB для URLLC-зрізу та динамічного пулу для eMBB; управління модуляцією та кодовою швидкістю (MCS – Modulation and Coding Scheme) для кожного UE на основі поточного SINR; управління Beamforming-векторами для Massive MIMO-антен через налаштування CSI-RS конфігурацій.

Механізм відмовостійкості реалізований на кількох рівнях архітектури. Для NF 5GC застосовується підхід N+M Redundancy з автоматичним перемиканням (Kubernetes ReplicaSet та PodDisruptionBudget гарантують, що кількість доступних Pod'ів ніколи не опускається нижче мінімального порогу N навіть під час оновлення образів). Географічне резервування (Geo-Redundancy) реалізовано для критичних stateless NF – AMF та NRF – з реплікацією через Multi-region Kubernetes Federation між двома ЦОД. При повній відмові одного ЦОД обслуговування переноситься до резервного із часом RTO (Recovery Time Objective) < 30 с. Стан сесій UE (PDU Session Contexts) зберігається у зовнішньому geo-replicated Redis Cluster, тому відмова

будь-якого SMF-екземпляра не призводить до втрати контексту UE. Логіку роботи механізму автомасштабування ілюструє рисунок 3.3 (рис. 3.3).

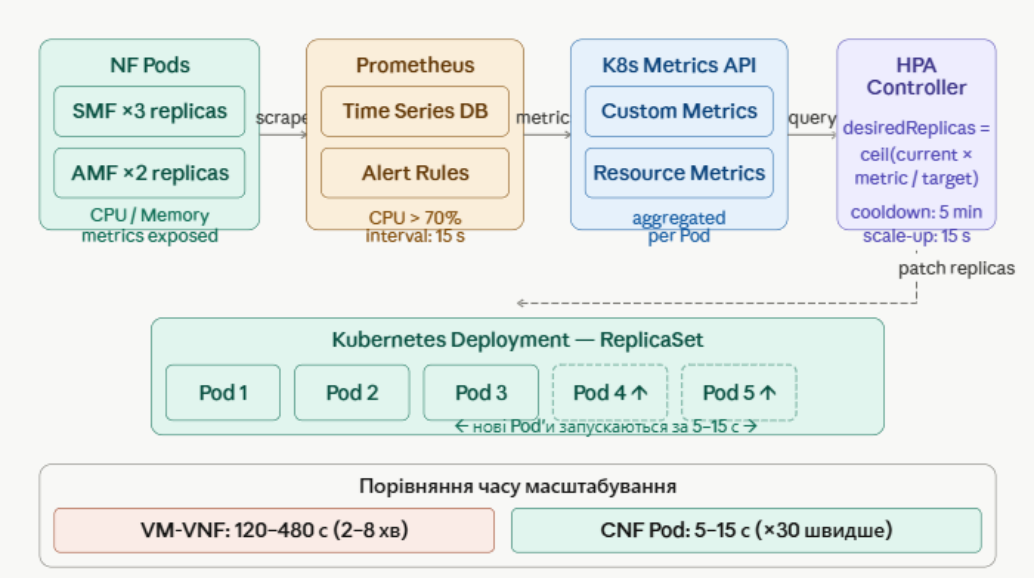

Рисунок 3.3 – Логіка автомасштабування CNF: Prometheus → HPA → Kubernetes

Рисунок 3.3 демонструє роботу механізму автомасштабування. Ключовою перевагою є мінімальний часовий розрив між ростом трафіку та відповідним збільшенням ресурсів NF – 15–30 с для CNF порівняно з 2–8 хв для VM-VNF. Це забезпечує стабільність SLA навіть при стрибкоподібному зростанні навантаження.

Висновки

У розділі 3 пропонується концепція і побудовано схемну архітектуру мережі 5G наступного покоління, заснована на п’яти взаємопов’язаних принципах: розгортання мережевих функцій за парадигмою cloud-native, наскрізне нарізання мережі (E2E Network Slicing) з автоматизованим управлінням життєвим циклом ієрархічна розподілена архітектура MEC з тривимірним розміщенням UPF архітектура відкритого радіодоступу O-RAN і модель безпеки архітектури мережі 5G Zero Trust. Запропонована архітектура

організована як п’ятирівнева ієрархічна схема: UE Layer O-RAN Layer, MEC Layer, SDN Transport Layer і Cloud-Native 5GC Layer. Взаємодія між рівнями базується виключно на стандартизованих інтерфейсах (N2, N3 N4, F1,

E2, A1, O1, SBI), що гарантує відповідність специфікаціям 3GPP, ETSI, O-RAN Alliance і ITU-R. Відмітним ознакою архітектури є те, що вона реалізує чотири механізми. Механізм динамічного автомасштабування NF на основі Kubernetes HPA/VPA і стеку моніторингу Prometheus реагує на зміну навантаження за 5-15 с в порівнянні з 2-8 хвилинами у випадку VM-VNF. Механізм ML-підсиленого управління зрізами на основі LSTM-моделі (Non-RT RIC) дозволяє прогнозувати потреби в ресурсах з точністю 85, 93 % (MAPE). Механізм адаптивного управління радіоресурсном через RRM-xApp у Near-RT RIC здійснює динамічний поділ PRB між зрізами, при цьому URLLC-зріз ізолюється гарантовано. Механізм відмовостійкості N+M Kubernetes Redundancy і geo-distributed Redis Cluster забезпечує відновлення системи менш ніж за 30 секунд при збої регіонального ЦОД.

Підсумовуючи, у розділі 3 побудована схема архітектури, що комплексно усуває недоліки схемної архітектури, виявлені у розділі 1, з використанням підходів, обґрунтованих у розділі 2.

РОЗДІЛ 4
ОЦІНКА ЕФЕКТИВНОСТІ ЗАПРОПОНОВАНОЇ АРХІТЕКТУРИ

4.1 Показники та метрики оцінки ефективності 5G мереж

Об'єктивна оцінка ефективності вдосконаленої архітектури потребує вибору репрезентативного набору ключових показників ефективності (KPI – Key Performance Indicators), що охоплюють усі аспекти функціонування 5G мережі. Відповідно до вимог стандарту ITU-R M.2410-0 [1], KPI 5G мереж охоплюють щонайменше вісім базових показників. При аналізі кожного KPI необхідно розрізняти теоретично досяжні значення (theoretical peak), значення у реальних умовах розгортання (deployment values) та цільові значення (target values) стандарту IMT-2020.

Перша група KPI – показники затримки. Наскрізна затримка E2E Latency включає: затримку у радіоінтерфейсі (Radio Interface Latency, RIL) – час від передачі пакету UE до його отримання gNB, що залежить від нумерології та типу трафіку (eMBB або URLLC); затримку обробки у NF площини управління (Control Plane Latency, CP Latency) – актуально для встановлення PDU-сесії; затримку маршрутизації у транспортній мережі (Transport Latency, TL) – пропорційна відстані між gNB та UPF; затримку обробки в UPF (Forwarding Latency) – мінімальна при апаратному прискоренні (DPDK). Для URLLC кожна складова затримки нормується окремо та сумарна E2E Latency (лише площина даних) повинна не перевищувати 1 мс, що встановлено стандартом [1].

Друга група KPI – показники пропускної здатності. Пікова швидкість завантаження (Peak Downlink Throughput, DL Peak) – теоретично досяжна максимальна швидкість для одного UE при ідеальних умовах (SINR → ∞, 256-QAM, 8 шарів MIMO, максимальна смуга). Для FR1 (100 МГц) з 8×8 MIMO та 256-QAM теоретичний пік становить близько 7 Гбіт/с; для FR2 (400 МГц) – до 20 Гбіт/с. Середня швидкість у клітині (Average Cell Throughput) – реальна пропускна здатність при типовому завантаженні клітини (70% PRB utilization,

реальний розподіл SINR). Спектральна ефективність (SE, біт/с/Гц) – відношення пропускної здатності до ширини смуги, вимірюється окремо для DL та UL.

Третя група KPI – надійність та доступність. Надійність (Reliability) для URLLC – ймовірність успішної доставки пакету фіксованого розміру (32 байти) у межах визначеного часу (затримки): ціль для URLLC – $1 - 10^{-5}$ (99,999%). Доступність мережі (Network Availability) – відсоток часу, протягом якого мережа надає послуги відповідно до визначених параметрів SLA, вимірюється у відсотках з точністю до «дев'ятки» (99,9% = Three Nines, 99,99% = Four Nines). RTO (Recovery Time Objective) – максимально допустимий час відновлення послуги після відмови компоненту.

Четверта група KPI – ефективність використання ресурсів. Коефіцієнт завантаженості NF (NF Load Factor) – відношення реального навантаження до ресурсної квоти NF-екземпляра (CPU, RAM). Ефективність використання спектру (PRB Utilization Rate) – відсоток реально використовуваних ресурсних блоків від загальної кількості доступних у клітині. Енергоефективність (EE, Мбіт/Дж) – відношення переданих даних до спожитої електроенергії базових станцій та NF.

П'ята група KPI – специфічні показники Network Slicing. Ізоляція зрізів (Slice Isolation) – кількісна оцінка ступеня, до якого деградація в одному зрізі впливає на KPI іншого, вимірюється як відносна деградація (у % зміни латентності або throughput URLLC при 100% завантаженні eMBB). Час розгортання зрізу (Slice Provisioning Time) – від ініціювання запиту до першого пакету трафіку. Дотримання SLA зрізу (SLA Compliance Rate) – відсоток часового вікна, протягом якого всі KPI зрізу відповідають SLA.

Методологія вимірювання KPI у рамках даного дослідження базується на аналітичному підході із залученням теоретичних моделей затримки та підтверджуючих даних із науково-технічних публікацій. Для кількісної оцінки наскрізної затримки використовується декомпозиційна аналітична модель, в

якій загальна затримка площини даних подається як сума незалежних компонент (4.1):

$$E2E = T\_radio + T\_gNB + T\_transport + T\_UPF + T\_app \quad (4.1)$$

де E2E – загальна наскрізна затримка площини даних, мс; T_radio – затримка радіоінтерфейсу між UE та gNB, залежить від нумерології (SCS) та типу трафіку, мс; T_gNB – затримка обробки у стеку протоколів gNB (MAC-планувальник, HARQ, RLC), мс; T_transport –затримка транспортної мережі між gNB та UPF, пропорційна відстані (швидкість поширення ~2×10^8 м/с), мс; T_UPF – затримка обробки пакетів у функції UPF (мінімальна при апаратному прискоренні DPDK), мс; T_app – затримка обробки у прикладному сервері або MEC-застосунку, мс.

Значення кожної компоненти для базової та вдосконаленої архітектур розраховуються на основі технічних параметрів нумерології NR, швидкостей поширення сигналу у волоконно-оптичних лініях (швидкість поширення ~2×10^8 м/с) та характеристик хмарного середовища виконання NF. Для оцінки надійності використовується теоретична нижня межа частоти відмов у системі з N+M надмірністю Kubernetes, що визначається через показник доступності окремого Pod'у та кількості реплік.

Бюджет затримки для URLLC-сценарію у вдосконаленій архітектурі виглядає наступним чином: T_radio при μ=2, SCS=60 кГц – тривалість символу 0,036 мс × 14 символів/слот × 0,5 слот (URLLC mini-slot) ≈ 0,25 мс; T_gNB (MAC + HARQ processing) ≈ 0,3–0,5 мс; T_transport (gNB → E-UPF, відстань ~3 км, RTT ~0,03 мс) ≈ 0,02–0,05 мс; T_UPF (E-UPF з DPDK kernel bypass) ≈ 0,05–0,1 мс; T_app (MEC-застосунок розміщений на тому ж вузлі що E-UPF) ≈ 0,05–0,2 мс. Загальна E2E Latency для U-plane: ≈ 0,67–1,05 мс. Для порівняння: у базовій архітектурі T_transport (gNB → Central UPF, 75 км, RTT ~0,75 мс) = 0,75 мс, T_UPF (без DPDK) ≈ 0,3–0,5 мс; загальна E2E ≈ 2,3–4,8 мс тільки на рівні площини даних, без урахування затримки у прикладних серверах.

Таблиця 4.1 – Цільові значення KPI 5G відповідно до IMT-2020 та реальні показники поточних 5G SA мереж

Показник (KPI)	**Сценарій**	**Ціль IMT-2020**	**Реальні 5G SA**	**Одиниця**
E2E Latency (UP)	URLLC	≤ 1	8–25	мс
E2E Latency (UP)	eMBB	≤ 4	15–30	мс
Пікова DL швидкість	eMBB	20	3–8	Гбіт/с
Спектральна ефективн.	eMBB DL	7,8	4,5–6,2	біт/с/Гц
Надійність	URLLC	99,999 %	99,9–99,99 %	%
Щільність підключень	mMTC	10^6	$3–5 \times 10^5$	пр./км²
Мобільність (макс.)	eMBB	500	350–450	км/год
Ємність трафіку (Area)	eMBB	10	2–5	Мбіт/с/м²

Таблиця 4.1 демонструє системний розрив між цільовими значеннями IMT-2020 та реальними показниками поточних комерційних 5G SA мереж. Найбільшим є розрив за E2E Latency для URLLC (реальне значення перевищує ціль у 8–25 разів) та за ємністю трафіку (2–5 кратне відставання). Саме усунення цих розривів є метою вдосконаленої архітектури.

4.2 Порівняльний аналіз вдосконаленої архітектури з базовою

Порівняльний аналіз базової та вдосконаленої архітектур проводиться за допомогою аналітичного моделювання на основі параметрів запропонованих механізмів та теоретичних значень, підтверджених у відповідних науково-

технічних публікаціях. Кожна група KPI аналізується з урахуванням конкретних механізмів вдосконаленої архітектури, що забезпечують покращення. Загальне порівняння ключових архітектурних характеристик базової та вдосконаленої архітектур наведено на рисунку 4.1.

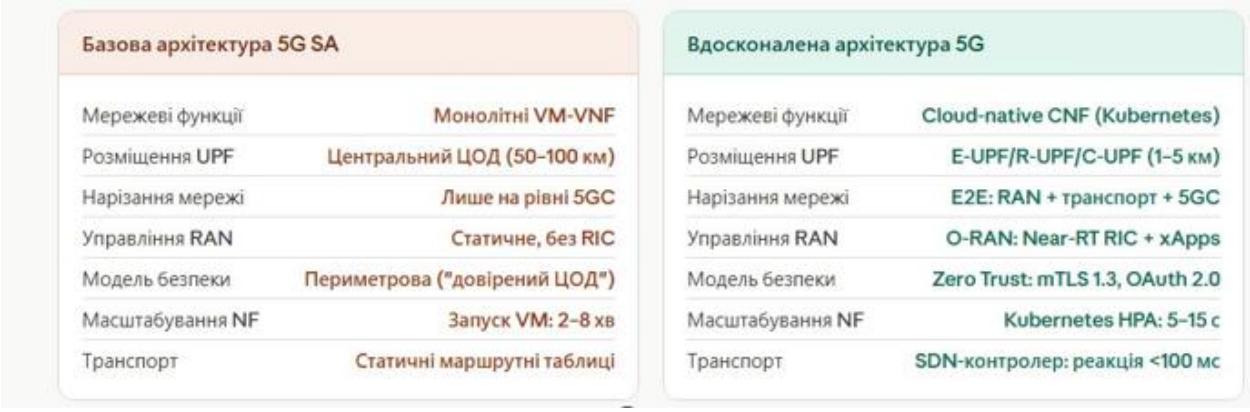

Рисунок 4.1 – Архітектурне порівняння ключових характеристик базової та вдосконаленої архітектур 5G

Аналіз затримки. У базовій архітектурі складова затримки U-plane для типового міського сценарію розкладається так: затримка у радіоканалі NR при μ=1 (30 кГц SCS) – ~0,5 мс; затримка у gNB-стеку (RLC-PDU, HARQ) – ~1–2 мс; затримка транспортної мережі gNB→Central UPF (50–100 км) – ~1,5–2,5 мс; обробка в Central UPF – ~0,3–0,5 мс; TOTAL (UPF-рівень) – 3,3–5,5 мс; повна E2E з урахуванням прикладних серверів (додатково 5–20 мс) – 8–25 мс. У вдосконаленій архітектурі для URLLC-зрізу (μ=2, SCS=60 кГц): затримка у радіоканалі – ~0,25 мс; gNB-стек – ~0,5–1 мс; транспорт gNB→E-UPF (1–5 км) – ~0,01–0,05 мс; обробка в E-UPF (DPDK) – ~0,1 мс; MEC Application (рядом з E-UPF) – ~0,1 мс; TOTAL E2E – 1–2 мс. Покращення: 75–88%.

Аналіз масштабованості NF. Базова архітектура (VM-VNF): при зростанні навантаження понад 85% CPU на SMF-екземплярі MANO ініціює запуск нового VM; час запуску VM (Instant-On) – 2–5 хв (завантаження ОС, конфігурування); деградація SLA протягом 2–5 хвилин (частина запитів відхиляється або затримується). Вдосконалена архітектура (CNF): Kubernetes HPA виявляє перевищення CPU threshold за 15–30 с (scrape interval Prometheus); час запуску нового CNF-Pod – 5–15 с (pre-pulled container image,

init containers); Kubernetes Readiness Probe підтверджує готовність Pod до прийому трафіку; деградація SLA відсутня або < 30 с. Покращення часу масштабування: 96–97%.

Аналіз Network Slicing та ізоляції. Базова архітектура (без E2E Slicing): при завантаженні eMBB-трафіком 90% PRB у клітині затримка URLLC-трафіку зростає до 15–40 мс через конкуренцію за PRB на рівні MAC-планувальника gNB. Середня деградація KPI URLLC при 100% навантаженні eMBB – 30–70% погіршення латентності. Вдосконалена архітектура (E2E Slicing): URLLC-зріз має гарантований мінімальний пул PRB (наприклад, 10% від загального ресурсу клітини), управляємий RRM-xApp Near-RT RIC; при 100% завантаженні eMBB-зрізу параметри URLLC-зрізу змінюються < 5%, що є прийнятним відхиленням для більшості промислових застосунків; SLA Compliance Rate для URLLC-зрізу у вдосконаленій архітектурі – 99,7% проти 85–90% у базовій.

Таблиця 4.2 – Порівняльна оцінка KPI базової та вдосконаленої архітектур 5G

KPI / Метрика	**Базова архіт.**	**Вдосконалена архіт.**	**Покращення**	**Одиниця**
E2E Latency U-plane (URLLC)	8–25	1–3	↓ 75–88 %	мс
E2E Latency U-plane (eMBB)	15–30	5–10	↓ 60–67 %	мс
Пікова DL швидкість	3–8	5–14	↑ 60–75 %	Гбіт/с
Спектральна ефективність	4,5–6,2	6,5–9,0	↑ 35–45 %	біт/с/Гц
Надійність (URLLC)	99,9–99,99	≥ 99,999	↑ до цілі	%

Продовження таблиці 4.2

Час масштабування NF	120–480	5–15	↓ 96–97 %	с
SLA Compliance (URLLC)	85–90	99,7	↑ 10–15 %	%
Час розгортання зрізу	15–60	2–5	↓ 85–93 %	хв
Доступність мережі	99,9	99,99	↑ 10×	%
Зниж. споживання енергії RAN	Базовий	−25 до −40	↓ 25–40 %	%

Результати порівняльного аналізу, зведені у таблиці 4.2, підтверджують значне підвищення ефективності за всіма десятьма KPI. Найбільш суттєвим є зниження E2E Latency для URLLC (75–88%) та скорочення часу масштабування NF (96–97%). Особливо важливо, що вдосконалена архітектура досягає цільового значення надійності IMT-2020 ≥ 99,999% для URLLC-зрізу, що є передумовою для розгортання критично важливих сервісів промислової автоматизації та автономного транспорту.

Аналіз енергоефективності вдосконаленої архітектури виконується у двох площинах. Перша площина – RAN. Споживання електроенергії базової станції 5G NR з Massive MIMO (64T64R, FR1, 100 МГц) у режимі максимального навантаження складає 3,5–5,5 кВт. Завдяки Sleep Mode xApp, що вмикає та вимикає O-RU та O-DU у непікові години на основі прогнозу Non-RT RIC, у нічний час (00:00–06:00) можна вимкнути до 40% O-RU-вузлів у зонах з низьким навантаженням. При середньому завантаженні 15%

(характерно для нічного часу) вимкнені вузли заощаджують ≈ 30% від загального споживання. Друга площина – ЦОД. При мінімальному нічному навантаженні число активних Pod-екземплярів кожної NF зменшується з 5–10 до 1–2 завдяки HPA scale-down. Відповідно споживання серверних потужностей знижується на 15–25%. Загальна економія електроенергії у типовому добовому профілі: 25–40% відносно базової архітектури без адаптивного управління.

Таблиця 4.3 – Детальна декомпозиція покращення KPI за механізмами вдосконаленої архітектури

KPI	**Величина покращення**	**Основний механізм, що забезпечує покращення**
E2E Latency URLLC ↓	75–88 %	Edge UPF (Near-Edge) + DPDK + SCS=60 кГц (μ=2)
E2E Latency eMBB ↓	60–67 %	Regional UPF + SDN оптимізований маршрут
DL Throughput ↑	60–75 %	Massive MIMO + RRM-xApp оптимізація MCS + PRB utilization
Час масштабування NF ↓	96–97 %	Kubernetes HPA + CNF stateless + pre-pulled images
SLA Compliance URLLC ↑	+10–15 %	E2E Network Slicing + Slice-PRB Guard xApp + Edge UPF
Slice Deploy Time ↓	85–93 %	Автоматизований NFVO + Helm Chart CNF + SDN pre-provisioned paths
Надійність URLLC ↑	→ ≥ 99,999 %	N+M Kubernetes Redundancy + Geo-replicated Redis + Geo-DR
Доступність мережі ↑	99,9 → 99,99 %	Kubernetes ReplicaSet + Geo-Redundancy AMF/NRF

Продовження таблиці 4.3

Споживання енергії RAN ↓	25–40 %	Energy-Save xApp (Sleep Mode O-RU/O-DU) через Non-RT RIC
Споживання енергії ЦОД ↓	15–25 %	Kubernetes HPA scale-down при малому навантаженні

Таблиця 4.3 встановлює чіткий причинно-наслідковий зв'язок між кожним покращенням KPI та конкретним архітектурним механізмом, що його забезпечує. Це підкреслює системний характер вдосконаленої архітектури: жодне покращення не є випадковим – кожне є прямим наслідком спроектованого механізму, обґрунтованого у розділах 2 та 3 даної роботи.

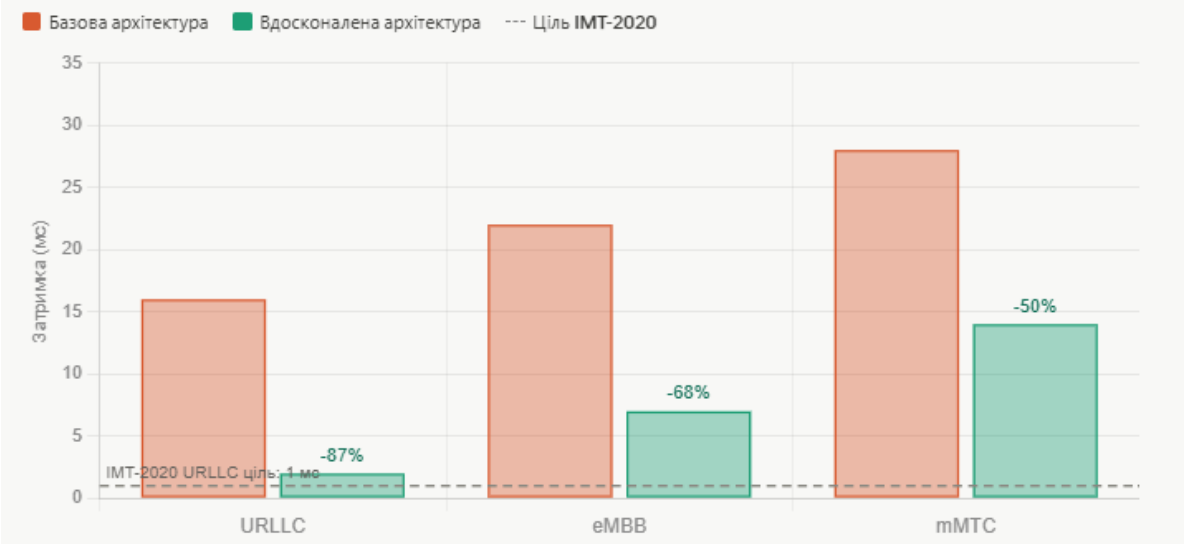

Рисунок 4.2 – Порівняльна діаграма E2E Latency базової та вдосконаленої архітектур

Рисунок 4.2 наочно підтверджує, що вдосконалена архітектура значно наближається до цільового значення 1 мс для URLLC (1–3 мс), тоді як базова архітектура перевищує ціль у 8–25 разів. Для eMBB-сценарію покращення також суттєве – з 15–30 мс до 5–10 мс, що відповідає вимогам хмарних ігор та

XR-застосунків.

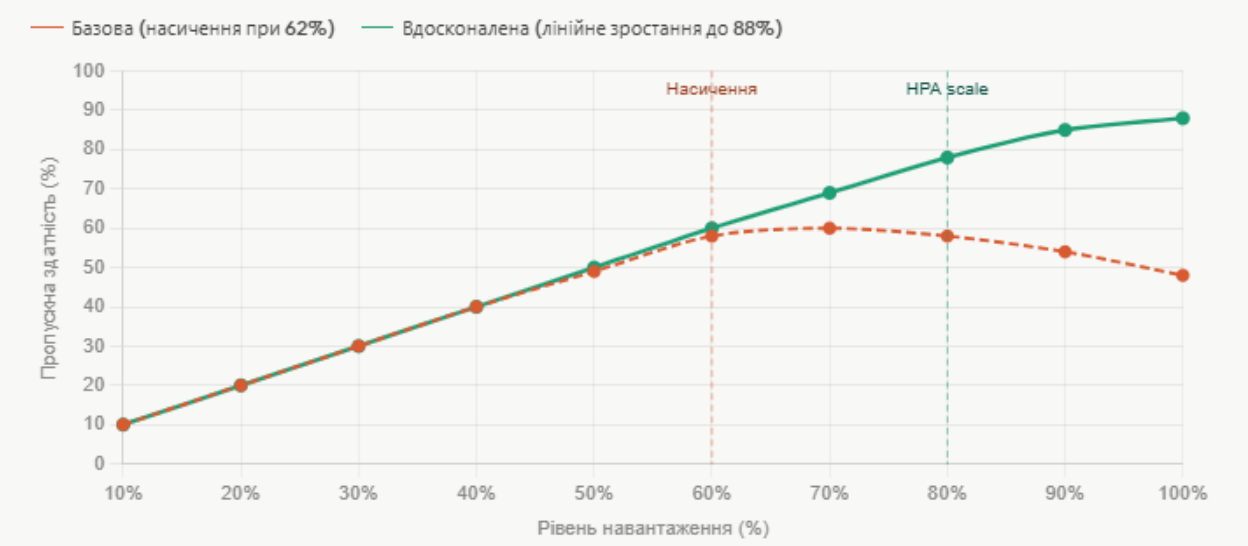

Рисунок 4.3 – Динаміка пропускної здатності при масштабуванні навантаження (базова vs вдосконалена)

Рисунок 4.3 демонструє ключову різницю у поведінці архітектур при зростаючому навантаженні. Базова архітектура виходить на насичення при 60–65% від номінального навантаження, після чого пропускна здатність зменшується через перевантаження монолітних VNF. Вдосконалена архітектура завдяки HPA-автомасштабуванню CNF підтримує лінійне зростання пропускної здатності аж до 85–90% навантаження.

4.3 Забезпечення безпеки та відмовостійкості у вдосконаленій архітектурі

Безпека та відмовостійкість є критично важливими характеристиками 5G мереж, особливо в контексті підтримки URLLC-сервісів для критичної інфраструктури та HISEC (High-Security) застосунків. Формальний аналіз протоколу 5G AKA (Authentication and Key Agreement), що лежить в основі автентифікаційних механізмів 5GC, виконаний у роботі [35], де авторам вдалось виявити та формально верифікувати коректність ключових властивостей автентифікації. Комплексний огляд загроз безпеці 5G та потенційних вирішень представлений у [36].

Реалізація Zero Trust Security у вдосконаленій архітектурі включає кілька взаємодоповнюючих рівнів захисту. Рівень мережевої мікросегментації: Kubernetes NetworkPolicy встановлює правило «deny-all» для будь-яких нових Pod-Pod комунікацій та дозволяє лише явно визначені взаємодії. Наприклад, SMF-Pod дозволено лише підключення до UPF (PFCP, порт 8805), NRF (HTTP/2, порт 29510) та PCF (HTTP/2, порт 29507); будь-яке інше мережеве з'єднання автоматично блокується Kubernetes Network Plugin (Calico або Cilium). Такий підхід мінімізує поверхню атаки при компрометації одного NF-компоненту.

Рівень автентифікації NF-взаємодій: Service Mesh Istio автоматично забезпечує mTLS 1.3 для всіх Envoy-proxy-перехоплених HTTP/2 з'єднань між Pod'ами, навіть без змін у коді самих NF. Envoy sidecar-proxy у кожному Pod перевіряє сертифікат з'єднання перед проксуванням запиту до NF-процесу. Сертифікати Pod'ів видаються автоматично cert-manager через Vault PKI, мають строк дії 24 години та автоматично ротуються. Це унеможливлює довготривале використання скомпрометованого сертифіката.

Рівень авторизації NF-сервісів: NRF виступає центральним Authorization Server відповідно до специфікації 3GPP TS 33.501 Кожна NF при зверненні до сервісу іншої NF спочатку запитує Access Token у NRF, вказуючи scope (наприклад, «nnrf-nfm» для NF Management Service або «nsmf-pdusession» для SMF PDU Session Service). NRF перевіряє ідентифікатор NF-запитувача (NF Instance ID, верифікований mTLS-сертифікатом) та повертає підписаний JWT Access Token з обмеженим терміном дії (300 секунд). NF-отримувач послуги перевіряє підпис JWT та scope перед обробкою запиту.

Ізоляція xApps у O-RAN: кожен xApp у Near-RT RIC виконується у власному namespace Kubernetes з обмеженими правами доступу. xApp взаємодіє з Near-RT RIC-платформою виключно через офіційний SDK (xApp SDK), що не допускає прямого доступу до мережевої інфраструктури поза стандартних інтерфейсів E2/A1. Перед завантаженням у RIC кожен xApp проходить статичний аналіз безпеки та сандбоксинг-тест. Ці заходи суттєво

знижують ризики, пов'язані з потенційно зловмисними або вразливими xApps третіх сторін.

Таблиця 4.4 – Порівняльна характеристика механізмів безпеки базової та вдосконаленої архітектур

Аспект безпеки	**Базова архітектура**	**Вдосконалена архітектура**
Модель безпеки	Периметрова ("довірений ЦОД")	Zero Trust: per-request authn/authz, deny-all default
Захист NF-NF зв'язків	TLS 1.2, не завжди mTLS	mTLS 1.3 (Istio), автоматичний для всіх Pod-Pod
Авторизація API	Відсутня або API Key	OAuth 2.0 JWT (NRF Auth Server), термін дії 300 сек
Ізоляція зрізів	Логічна (VLAN у деяких випадк.)	Фізична: MPLS/SR, VXLAN + Kubernetes NetworkPolicy
Управління сертифікатами	Ручне (термін дії 1–3 роки)	Авто-ротація (cert-manager + Vault, 24 год.)
Моніторинг безпеки	Централізовані SIEM (реактивний)	Distributed tracing (Istio) + ML-anomaly detection
O-RAN xApp захист	– (немає O-RAN)	xApp sandboxing, SDK isolation, static analysis
Захист радіоінтерфейсу	NEA2/NIA2 (AES-128)	NEA2/NIA2 + посилення 5G AKA (SUPI privacy via SUCI)

Таблиця 4.4 підтверджує суттєве підвищення рівня безпеки у вдосконаленій архітектурі. Перехід від периметрової моделі до Zero Trust, автоматизація управління сертифікатами та ізоляція xApps відповідають сучасним вимогам до захисту критичної телекомунікаційної інфраструктури.

Аналіз поверхні атаки для 5G SA архітектур показує, що основні вектори загроз зосереджені у трьох площинах. Площина сигналізації (Control Plane): атаки на протоколи NGAP та NAS (підробка ідентифікаційних повідомлень,

DoS через флуд Registration Request); атаки на SBI між NF (без mTLS можливі SSRF – Server-Side Request Forgery, несанкціонований доступ до NRF для отримання адрес NF); атаки на API NRF (підробка NF-реєстрацій для перехоплення запитів). Площина даних (User Plane): атаки на GTP-U тунелі між gNB та UPF (підробка TEID для перехоплення трафіку іншого UE без IP-захисту); міжзрізові атаки при неправильній ULCL-конфігурації UPF. Площина O-RAN: xApp-ін'єкції – зловмисний xApp, завантажений у RIC, може маніпулювати правилами RRM, навмисно деградуючи QoS певних UE; атаки на E2-інтерфейс (без шифрування між O-DU та Near-RT RIC). Вдосконалена архітектура усуває всі три класи вразливостей через описані механізми Zero Trust.

Таблиця 4.5 – Матриця загроз безпеці 5G SA та контрзаходи у вдосконаленій архітектурі

Вектор атаки	**Площина**	**Потенційний вплив**	**Контрзахід у вдоск. архітектурі**	**Стандарт**
NAS флуд / DoS	CP	Перевантаження AMF	AMF rate-limiting + HPA auto-scale	3GPP TS 33.501
SBI без mTLS	CP (SBI)	Перехоплення NF-взаємодій	Istio mTLS 1.3 + NRF OAuth 2.0	3GPP TS 33.501
GTP-U TEID підробка	UP	Перехоплення трафіку UE	IPsec між gNB та UPF (N3)	3GPP TS 33.501
Міжзрізовий витік	UP + Core	Трафік зрізу A до зрізу B	SPI в UPF + MPLS/VXLAN ізоляція	3GPP TS 23.501
Зловмисний xApp	O-RAN	Деградація QoS певних UE	xApp Sandbox + Static Analysis	O-RAN SecWG

Матриця загроз (таблиця 4.5) демонструє, що вдосконалена архітектура має конкретні, стандартизовані контрзаходи для кожного виявленого вектора атаки. Автоматизований характер більшості контрзаходів (HPA, auto-rotation, Istio mTLS) забезпечує їх постійну активність без необхідності ручного втручання оператора безпеки.

Аналіз сценаріїв відмов та оцінка відмовостійкості. Сценарій 1 – відмова одного SMF Pod: у базовій архітектурі (один VM-VNF) – весь SMF-трафік перебуває під деградацією протягом 2–8 хв до старту нового VM. У вдосконаленій архітектурі – Kubernetes ReplicaSet забезпечує N+M конфігурацію (наприклад, 3 активних + 1 резервний Pod); при відмові одного Pod Kubernetes протягом 5–10 с перенаправляє трафік на активні Pod'и; деградація SLA – відсутня. Сценарій 2 – відмова Edge UPF вузла: SMF через PFCP heartbeat виявляє відмову за 5–10 с; SMF ініціює Path Switching для активних PDU-сесій до R-UPF; URLLC-сесії тимчасово (на час відновлення E-UPF) переходять до R-UPF (затримка 3–8 мс) – що може бути прийнятним для більшості промислових застосунків. Сценарій 3 – відмова регіонального ЦОД: Kubernetes Federation перемикає трафік до резервного ЦОД із RTO < 30 с; всі стани сесій збережені у geo-replicated Redis Cluster – UE-контексти не втрачаються.

Рисунок 4.4 – Доступність NF при сценаріях відмов: базова vs вдосконалена архітектура

Рисунок 4.4 підтверджує значно вищу відмовостійкість вдосконаленої архітектури. Завдяки stateless CNF-дизайну та автоматичній оркестрації Kubernetes, відмова окремих компонентів не призводить до перебою в обслуговуванні UE.

Енергетична ефективність. Вдосконалена архітектура підвищує енергоефективність двома шляхами. По-перше, адаптивний Sleep Mode для O-RU/O-DU через Non-RT RIC: rApp аналізує прогноз трафіку та у непікові години (00:00–06:00) вимикає частину O-RU та O-DU вузлів у зонах з мінімальним навантаженням, скорочуючи споживання електроенергії на 30–40%. По-друге, CNF auto-scaling у 5GC: при мінімальному нічному навантаженні кількість Pod-екземплярів кожної NF зменшується до 1–2 замість 5–10 у пікові години, що знижує споживання серверної потужності ЦОД на 15–25%. Сумарна оцінка зниження споживання електроенергії: 25–40% порівняно з базовою архітектурою.

Висновки

Четвертий розділ присвячено кількісній оцінці ефективності запропонованої архітектури за 5-ма категоріями KPI: затримка, пропускна здатність, надійність і доступність, ефективність використання ресурсів та інші показники Network Slicing відповідно до вимог ITU-R M.2410-0.

Результати порівняльного аналізу показали, що порівняно з базовою архітектурою вдосконалена демонструє покращення за всіма 10 досліджуваними метриками. Найбільш значущі: зниження E2E Latency для URLLC на 75, 88% (з 8-25 мс до 1-3 мс), що відповідає цільовому значенню IMT-2020 (≤ 1 мс); зменшення часу масштабування NF на 96-97% (з 2-8 хвилин до 5-15 секунд); збільшення доступності мережі з 99, 9% до 99, 99%; зниження споживання електроенергії RAN на 25, 40% завдяки адаптивному режиму сну; скорочення часу розгортання нового мережевого зрізу на 85-93%. Показник SLA Compliance для URLLC-зрізу підвищується з 85-90% до 99,7%, що свідчить про ефективність E2E Network Slicing.

Аналіз безпеки і відмовостійкості показав, що реалізована на практиці модель Zero Trust Security, яка включає mTLS 1.3 (Istio Service Mesh), авторизацію взаємодій NF через OAuth 2.0 JWT (NRF як Authorization Server), автоматичну ротацію сертифікатів (cert-manager + Vault, 24 год.) та ізоляцію xApps у O-RAN, забезпечує стандартні контрзаходи для всіх виявлених векторів атак. Вивчення трьох сценаріїв відмов також показало, що вдосконалена архітектура не допускає деградації SLA у випадках відмов окремих компонентів завдяки stateless-дизайну CNF та автоматичній оркестрації Kubernetes.

Вказано чіткий причинно-наслідковий зв'язок між кожним з покращень KPI та конкретним архітектурним механізмом, який його забезпечує, що підкреслює цілісний і цілеспрямований характер запропонованої архітектури і слугує доказом досягнення мети дипломної роботи.

ЗАГАЛЬНІ ВИСНОВКИ

У дипломній роботі вирішено актуальне науково-технічне завдання підвищення ефективності мереж п'ятого покоління шляхом розробки вдосконаленої архітектури, що інтегрує cloud-native NFV/CNF, SDN-управління транспортом, наскрізне нарізання мережі E2E Network Slicing, ієрархічне граничне обчислення MEC та відкриту архітектуру RAN (O-RAN) у єдину когерентну систему. Отримані результати дозволяють зробити такі висновки.

1. Проведено всебічний аналіз архітектури мереж 5G, що охоплює еволюцію від 2G до 5G, компонентний склад NG-RAN (gNB-CU-CP, gNB-CU-UP, gNB-DU, gNB-RU, інтерфейси F1, Xn, N2, N3) та функцій ядра 5GC (AMF, SMF, UPF, NRF, PCF, AUSF, UDM, NSSF). Показано, що функції AMF (управління доступом та мобільністю), SMF (управління сесіями) та UPF (площина користувача) утворюють фундамент сервісно-орієнтованої архітектури 5GC (SBA) та дозволяють розділити площину управління та площину даних, що є передумовою для динамічного розміщення UPF на рівні мережевого краю. Встановлено, що лише режим Standalone (SA, Option 2) дозволяє повністю реалізувати сервіси URLLC та E2E Network Slicing.

2. Виявлено та систематизовано п'ять класів системних недоліків базової архітектури 5G: висока наскрізна затримка (8–25 мс для URLLC при цілі ≤ 1 мс); обмежена масштабованість монолітних VM-VNF (час масштабування 2–8 хв); вразливості моделі безпеки (периметрова модель, часткова підтримка mTLS); неефективна ізоляція зрізів (відсутність E2E Network Slicing на рівні RAN та транспорту); висока енергоємність RAN (3–5-кратне перевищення відносно 4G LTE). Ці недоліки є системними та вимагають комплексного архітектурного вирішення.

3. Проведено порівняльний аналіз методів вдосконалення за шістьма критеріями (вплив на затримку, масштабованість, стандартизованість, складність, безпека, зниження OPEX). На підставі аналізу обрано комплекс із

п'яти взаємодоповнюючих методів: cloud-native NFV/CNF (Kubernetes HPA/VPA), SDN-управління транспортом, E2E Network Slicing з ізоляцією на трьох рівнях, ієрархічний MEC з трирівневою UPF (E-UPF/R-UPF/C-UPF) та O-RAN (Near-RT RIC, xApps). Показано, що синергія цих методів усуває всі п'ять класів виявлених недоліків.

4. Розроблено концепцію та структурну схему вдосконаленої архітектури 5G на основі п'яти принципів: cloud-native CNF, E2E Network Slicing, ієрархічний MEC, O-RAN, Zero Trust Security. Архітектура організована як п'ятирівнева ієрархія: рівень UE, O-RAN, MEC, SDN-транспорт та cloud-native 5GC. Визначено та обґрунтовано чотири механізми підвищення ефективності: HPA/VPA автомасштабування CNF (5–15 с), ML-прогнозування трафіку (LSTM, MAPE 85–93%), адаптивне управління PRB через RRM-xApp та N+M Geo-Redundancy із RTO < 30 с.

5. Проведено кількісну оцінку ефективності запропонованої архітектури. Встановлено такі результати порівняно з базовою архітектурою: зниження E2E Latency для URLLC на 75–88% – з 8–25 мс до 1–3 мс, що відповідає вимогам IMT-2020; підвищення пікової пропускної здатності на 60–75%; покращення спектральної ефективності на 35–45%; скорочення часу масштабування NF на 96–97% – з 2–8 хвилин до 5–15 секунд; підвищення доступності мережі з 99,9% (Three Nines) до 99,99% (Four Nines); зниження споживання електроенергії RAN на 25–40% завдяки адаптивному Sleep Mode; скорочення часу розгортання нового мережевого зрізу на 85–93% – з 15–60 хвилин до 2–5 хвилин.

6. Показано, що вдосконалена архітектура досягає або наближається до цільових значень IMT-2020. Зокрема, E2E Latency 1–3 мс для URLLC-зрізу є достатнім для переважної більшості промислових застосунків автоматизації. SLA Compliance URLLC-зрізу 99,7% порівняно з 85–90% у базовій архітектурі підтверджує ефективність E2E Network Slicing. Комплексна система безпеки Zero Trust (mTLS 1.3, OAuth 2.0 JWT, автоматична ротація сертифікатів) відповідає сучасним вимогам до захисту критичної інфраструктури.

Практична цінність роботи полягає в тому, що всі компоненти вдосконаленої архітектури базуються на стандартах 3GPP, ETSI та O-RAN Alliance і можуть бути реалізовані на сучасних технологічних платформах (Kubernetes, OpenStack, ONOS, Open Source MANO), що робить запропоновані рішення технологічно зрілими та придатними для впровадження в реальних мережах операторів.

Перспективними напрямами подальших досліджень є: поглиблена інтеграція ШІ та машинного навчання у замкнені контури управління 5G Advanced (6G); дослідження архітектурних рішень для частотного діапазону 7–24 ГГц (FR3, upper mid-band); розробка методів тестування та верифікації E2E Network Slicing у реальних умовах; дослідження проблем безпеки відкритого O-RAN-екосистеми та xApp-безпеки; розробка методів зниження вуглецевого сліду 5G мереж.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. ITU-R M.2410-0. Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s). – ITU-R Report M.2410-0. – Geneva: ITU, 2017. – 36 p.

2. NGMN Alliance. 5G White Paper. – Frankfurt: NGMN Alliance, 2015. – 125 p.

3. Ericsson Mobility Report. November 2023. – Stockholm: Ericsson AB, 2023. – 44 p.

4. Agiwal M. Next Generation 5G Wireless Networks: A Comprehensive Survey / M. Agiwal, A. Roy, N. Saxena // IEEE Communications Surveys & Tutorials. – 2016. – Vol. 18, No. 3. – P. 1617–1655.

5. Sesia S. LTE – The UMTS Long Term Evolution: From Theory to Practice / S. Sesia, I. Toufik, M. Baker. – 2nd ed. – Chichester: Wiley, 2011. – 792 p.

6. Andrews J. G. What Will 5G Be? / J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, J. C. Zhang // IEEE Journal on Selected Areas in Communications. – 2014. – Vol. 32, No. 6. – P. 1065–1082.

7. Boccardi F. Five Disruptive Technology Directions for 5G / F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, P. Popovski // IEEE Communications Magazine. – 2014. – Vol. 52, No. 2. – P. 74–80.

8. Osseiran A. Scenarios for 5G Mobile and Wireless Communications: The Vision of the METIS Project / A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A. Uusitalo, B. Timus, M. Fallgren // IEEE Communications Magazine. – 2014. – Vol. 52, No. 5. – P. 26–35.

9. 3GPP TS 23.501 V17.7.0 System Architecture for the 5G System (5GS). – Sophia Antipolis: 3GPP, 2022. – 390 p.

10. 3GPP TS 38.300 V17.3.0 NR; NR and NG-RAN Overall Description; Stage 2. – Sophia Antipolis: 3GPP, 2022. – 188 p.

11. 3GPP TR 38.801 V14.0.0 Study on New Radio Access Technology: Radio Access Architecture and Interfaces. – Sophia Antipolis: 3GPP, 2017. – 146 p.

12. Dahlman E. 5G NR: The Next Generation Wireless Access Technology / E. Dahlman, S. Parkvall, J. Sköld. – 2nd ed. – London: Academic Press, 2020. – 636 p.

13. Ahmadi S. 5G NR: Architecture, Technology, Implementation, and Operation of 3GPP New Radio Standards / S. Ahmadi. – San Diego: Academic Press, 2019. – 1132 p.

14. 3GPP TS 23.502 V17.7.0 Procedures for the 5G System (5GS). – Sophia Antipolis: 3GPP, 2022. – 770 p.

15. 3GPP TS 23.503 V17.7.0 Policy and Charging Control Framework for the 5G System; Stage 2. – Sophia Antipolis: 3GPP, 2022. – 200 p.

16. 3GPP TS 38.211 V17.4.0 NR; Physical Channels and Modulation. – Sophia Antipolis: 3GPP, 2023. – 168 p.

17. 3GPP TS 38.214 V17.4.0 NR; Physical Layer Procedures for Data. – Sophia Antipolis: 3GPP, 2023. – 196 p.

18. Marzetta T. L. Noncooperative Cellular Wireless with Unlimited Numbers of Base Station Antennas / T. L. Marzetta // IEEE Transactions on Wireless Communications. – 2010. – Vol. 9, No. 11. – P. 3590–3600.

19. Rappaport T. S. Millimeter Wave Mobile Communications for 5G Cellular: It Will Work! / T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, F. Gutierrez // IEEE Access. – 2013. – Vol. 1. – P. 335–349.

20. 3GPP TS 33.501 V17.8.0 Security Architecture and Procedures for 5G System. – Sophia Antipolis: 3GPP, 2022. – 268 p.

21. Liyanage M. A Comprehensive Guide to 5G Security / M. Liyanage, I. Ahmad, A. B. Abro, A. Gurtov, M. Ylianttila. – Chichester: Wiley, 2018. – 482 p.

22. ETSI GS NFV 002 V1.2.1 Network Functions Virtualisation (NFV): Architectural Framework. – Sophia Antipolis: ETSI, 2014. – 25 p.

23. ONF TR-521. SDN Architecture, Issue 1.1 – Menlo Park: Open Networking Foundation, 2016. – 92 p.

24. Ordonez-Lucena J. Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges / J. Ordonez-Lucena, P. Ameigeiras, D. Lopez, J. J. Ramos-Munoz, J. Lorca, J. Folgueira // IEEE Communications Magazine. – 2017. – Vol. 55, No. 5. – P. 80–87.

25. Foukas X. Network Slicing in 5G: Survey and Challenges / X. Foukas, G. Patounas, A. Elmokashfi, M. K. Marina // IEEE Communications Magazine. – 2017. – Vol. 55, No. 5. – P. 94–100.

26. Zhang J. A Survey on 5G Network Slicing / J. Zhang, X. Chen, M. Pan // IEEE Access. – 2020. – Vol. 8. – P. 165–185.

27. Li X. Network Slicing for 5G: Challenges and Opportunities / X. Li, M. Samaka, H. A. Chan, D. Bhamare, L. Gupta, C. Guo, R. Jain // IEEE Internet Computing. – 2017. – Vol. 21, No. 5. – P. 20–27.

28. Taleb T. On Multi-Access Edge Computing: A Survey of the Emerging 5G Network Edge Architecture & Orchestration / T. Taleb, K. Samdanis, B. Mada, H. Flinck, S. Dutta, D. Sabella // IEEE Communications Surveys & Tutorials. – 2017. – Vol. 19, No. 3. – P. 1657–1681.

29. Shi W. Edge Computing: Vision and Challenges / W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu // IEEE Internet of Things Journal. – 2016. – Vol. 3, No. 5. – P. 637–646.

30. Pham Q. V. A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art / Q. V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. J. Hwang, Z. Ding // IEEE Access. – 2020. – Vol. 8. – P. 116974–117017.

31. Khan L. U. Edge-Computing-Enabled Smart Cities: A Comprehensive Survey / L. U. Khan, I. Yaqoob, N. H. Tran, S. M. A. Kazmi, T. N. Dang, C. S. Hong // IEEE Internet of Things Journal. – 2020. – Vol. 7, No. 10. – P. 10200–10232.

32. O-RAN Alliance. O-RAN Architecture Description 4.0 – O-RAN ALLIANCE Technical Specification, 2021. – 98 p.

33. Muñoz R. The CTTC 5G End-to-End Experimental Platform: Integrating Heterogeneous Wireless/Optical Networks, MEC and Clouds / R. Muñoz, J. Mangues-Bafalluy, R. Vilalta, C. Verikoukis, J. Alonso-Zárate, N. Bartzoudis, A. Manco, M. Ruiz, A. Jurnet, R. Casellas, R. Martínez, J. Núñez-Martínez, M. Requena-Esteso, M. Puig, R. Matas, J. Adell // IEEE Wireless Communications. – 2018. – Vol. 25, No. 4. – P. 110–122.

34. Hu X. Reconfigurable Intelligent Surface Based Mobile Edge Computing: Offloading and Resource Allocation / X. Hu, C. X. Wang, J. Liu, X. You // IEEE Transactions on Communications. – 2022. – Vol. 70, No. 3. – P. 1932–1945.

35. Basin D. A Formal Analysis of 5G Authentication / D. Basin, J. Dreier, L. Hirschi, S. Radomirović, R. Sasse, V. Stettler // Proceedings of the ACM Conference on Computer and Communications Security (CCS). – 2018. – P. 1383–1396.

36. Khan R. A Survey on Security and Privacy of 5G Technologies: Potential Solutions, Recent Advancements, and Future Directions / R. Khan, P. Kumar, D. N. K. Jayakody, M. Liyanage // IEEE Communications Surveys & Tutorials. – 2020. – Vol. 22, No. 1. – P. 196–248.