МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

ЖАРІКОВ ЕДУАРД В'ЯЧЕСЛАВОВИЧ

УДК 004.75

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ УПРАВЛІННЯ ІТ-ІНФРАСТРУКТУРОЮ ХМАРНОГО ЦЕНТРУ ОБРОБЛЕННЯ ДАНИХ

05.13.06 – інформаційні технології

АВТОРЕФЕРАТ

дисертації на здобуття наукового ступеня

доктора технічних наук

Київ – 2020

Дисертацією є рукопис.

Робота виконана на кафедрі автоматики та управління в технічних системах Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського” Міністерства освіти і науки України.

Науковий консультант: доктор технічних наук, професор
Теленик Сергій Федорович,
Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”, декан факультету інформатики та обчислювальної техніки.

Офіційні опоненти: доктор технічних наук, старший науковий співробітник
Гуляницький Леонід Федорович,
Інститут кібернетики ім. В.М. Глушкова НАН України,
завідувач відділу методів комбінаторної оптимізації та інтелектуальних інформаційних технологій № 180;

доктор технічних наук, професор
Снитюк Віталій Євгенович,
Київський національний університет імені Тараса Шевченка,
декан факультету інформаційних технологій;

доктор технічних наук, професор
Купін Андрій Іванович,
Криворізький національний університет,
завідувач кафедри комп’ютерних систем та мереж.

Захист відбудеться “ 09 ” липня 2020 р. о 12 годині на засіданні спеціалізованої вченої ради Д 26.002.29 в Національному технічному університеті України “Київський політехнічний інститут імені Ігоря Сікорського” за адресою: 03056, Київ, просп. Перемоги, 37, корп. 11, ауд. 215.

З дисертацією можна ознайомитися у бібліотеці Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського” за адресою: 03056, м. Київ, пр. Перемоги, 37.

Автореферат розісланий “ 09 ” червня 2020 р.

В.о. вченого секретаря
спеціалізованої вченої ради  О. І. Ролік

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. Цифрова трансформація суттєво впливає на всі сфери життєдіяльності людини і вимагає вирішення низки проблем організації обчислень, оброблення, передачі і зберігання даних, що потребує розроблення і впровадження нових підходів, технологій і методів управління ІТ-інфраструктурою. Від ефективного функціонування ІТ-інфраструктури напряму залежить реалізація місії компанії і її конкурентоспроможність.

Основне навантаження, що виникає при наданні сучасних ІТ-послуг, обслуговується ресурсами ІТ-інфраструктури центрів оброблення даних (ЦОД). При цьому в процесі еволюції ІТ-інфраструктури компанії виникають проблеми технічного і організаційного характеру, рішення яких залежить від рівня розвитку архітектури, сучасних тенденцій в області віртуалізації і програмно-визначених систем, хмарних обчислень і аутсорсингу.

Дослідження і розробки в галузі управління ІТ-інфраструктурою ЦОД проводяться міжнародними організаціями (Distributed Management Task Force, The Open Group, Cloud Native Computing Foundation, Storage Networking Industry Association, Uptime Institute, IEEE та ін.), аналітичними агенціями (Gartner, IDC, Forrester Research, RightScale), великими корпораціями (Microsoft, Google, Amazon, IBM, Hewlett-Packard, Cisco та ін.). Управлінню ІТ-інфраструктурою присвячено ряд праць (Глушков В.М., Палагін О.В. та ін.). Управління ІТ-інфраструктурою спирається на такі напрями наукових досліджень як: системний аналіз (Берталанфі Л. фон, Вінер М., Вернадський В.І., Клір Дж., Згуровський М.З. та ін.), математичні методи вирішення задач оптимізації (Нейман Д. фон, Данціг Д.Б., Канторович Л.В. та ін.), інформаційні технології та системи (Гроувер Д., Сатер Р., Баррозо Л. та ін.), управління якістю послуг (Демінг У., Джуран Д., Ісікава К., Кросбі Ф. та ін.), теорія ієрархій (Петті Г., Уайт Л., Аллен Т. та ін.), теорія управління багато-об’єктними багаторівневими системами (Кунцевич В.М., Лебедев Д.В., Месарович М., Беллман Р., Воронов Є.М. та ін.), теорія штучного інтелекту (Хопфілд Д.Д., Голдберг Д. та ін.).

Але наведені дослідження не вирішують всі проблеми, пов’язані з управлінням ІТ-інфраструктурою ЦОД провайдерів хмарних послуг. Недостатньо опрацьовані аспекти управління, пов’язані з управлінням ресурсами з дотриманням заданих вимог угоди про рівень обслуговування (SLA) в умовах змін режимів роботи і навантажень, раціонального використання ресурсів ІТ-інфраструктури хмарного ЦОД з урахуванням різних стратегій і критеріїв, зменшення операційних витрат на підтримку ефективної роботи ІТ-інфраструктури та ін. Висока динаміка розвитку інформаційних процесів в цифровому суспільстві, поява новітніх парадигм та підходів до їх реалізації, співіснування програмно-визначених і гібридних ІТ-інфраструктур різних поколінь обумовлюють невідповідність існуючих рішень і підходів до управління ІТ-інфраструктурою хмарного ЦОД сучасним потребам і викликам бізнесу.

Набула поширення концепція хмарних обчислень, яка реалізує обчислення на замовлення у спосіб використання хмарних ЦОД з метою забезпечення споживачів високоякісними і високопродуктивними ІТ-послугами. Хмарні ЦОД є найбільш динамічними, складними системами з централізованим управлінням, що забезпечують надання тарифікованого доступу до інформаційних платформ, процесів, застосунків та обчислювальних ресурсів на основі сервісних моделей надання хмарних послуг. Забезпечення заданої якості надання хмарних послуг в умовах змінних навантажень одночасно зі зменшенням операційних витрат є одним з основних завдань при

управлінні ІТ-інфраструктурою хмарного ЦОД. Отже, виникає науково-практична проблема забезпечення ефективного функціонування ІТ-інфраструктури хмарних ЦОД через створення методології та на її основі математичних моделей, методів та інформаційної технології управління з метою надання хмарних послуг із заданими показниками якості кінцевому користувачеві.

Таким чином, виключна актуальність питань теоретико-методологічного обґрунтування впровадження підходів і методів управління ІТ-інфраструктурою хмарних ЦОД і розроблення інформаційної технології, що реалізує зазначені методи і підходи, зумовили і визначили мету, завдання та зміст дисертації.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана у відповідності до плану наукових досліджень кафедри автоматики та управління в технічних системах Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського" в рамках науково-дослідних робіт: "Хмарна платформа розроблення і управління функціонуванням критичних ІТ-інфраструктур, що опрацьовують великі обсяги даних", 2017-2019 рр., номер державної реєстрації 0117U000537; "Розробка та впровадження системи управління ІТ-інфраструктурою з консолідованими інформаційно-обчислювальними ресурсами", 2014-2016 рр., номер державної реєстрації 0115U000322; "Розроблення і дослідження моделей, методів та технологій проектування, програмування і управління хмарними ІТ-інфраструктурами", 2013-2015 рр., номер державної реєстрації 0113U002285; "Платформа розроблення, експлуатації і розвитку критичних ІТ-інфраструктур для роботи з великими даними", 2016-2018 рр., номер державної реєстрації 0116U003801.

Мета і задачі дослідження. Метою роботи є створення методології та на її основі математичних моделей, методів та інформаційної технології як системного підходу до вирішення проблеми ефективного управління ІТ-інфраструктурою хмарного ЦОД в умовах невизначеності та змінних навантажень, що дозволить забезпечити заданий рівень обслуговування користувачів та зниження операційних і капітальних витрат.

Для досягнення зазначеної мети необхідно обґрунтувати і розробити теоретичні положення, методологічні підходи і науково-практичні рекомендації щодо розвитку теорії автоматизованого управління ІТ-інфраструктурою хмарного ЦОД на основі системного підходу до оброблення інформації, прогнозування і оптимізації.

Досягнення поставленої мети передбачає розв'язання таких теоретичних, методологічних і практичних задач:

1) аналіз існуючих підходів, технологій, моделей і методів управління ІТ-інфраструктурою ЦОД провайдерів хмарних послуг, уточнення проблеми і задач дослідження;
2) розроблення методології управління ІТ-інфраструктурою хмарного ЦОД як системи новітніх підходів, моделей і методів з використанням стратегій управління і схем їх реалізації для відповідних умов функціонування;
3) розроблення моделей і методів прогнозування змішаних навантажень на ресурси ІТ-інфраструктури хмарного ЦОД в умовах невизначеності з використанням принципів адаптації та комбінування;
4) розроблення комплексу моделей і методів інтегрованого управління ресурсами, потужністю і сховищем ІТ-інфраструктури хмарного ЦОД на базі алгоритмів і методів стохастичного локального пошуку з використанням прогнозування змінних

навантажень, нових метрик оцінювання стану та дотриманням вимог угоди про рівень обслуговування;

5) подальший розвиток алгоритмів і методів стохастичного локального пошуку з урахуванням особливостей управління ресурсами і навантаженням ІТ-інфраструктури хмарного ЦОД в умовах невизначеності;
6) розроблення архітектури багаторівневої програмно-визначеної системи управління ІТ-інфраструктурою хмарного ЦОД і концепції інформаційної технології на базі підходів і методів інтегрованого та ієрархічного управління;
7) подальший розвиток декомпозиційно-компенсаційного підходу до управління хмарним ЦОД з урахуванням адаптивності, багаторівневості та програмно-визначених підходів його функціонування;
8) розроблення і реалізація інформаційної технології управління ІТ-інфраструктурою хмарного ЦОД та експериментальне дослідження її ефективності, а також розроблених моделей, алгоритмів і методів.

Об'єктом дослідження є процеси управління ІТ-інфраструктурою центру оброблення даних провайдера хмарних послуг.

Предметом дослідження є методологія, моделі і методи процесів управління ІТ-інфраструктурою центру оброблення даних провайдера хмарних послуг.

Методи дослідження. Для вирішення проблеми ефективного управління ІТ-інфраструктурою хмарного ЦОД використовувались теорія систем, методи теорії ієрархій, методи математичного програмування, методи дослідження операцій і теорії прийняття рішень, методи математичного та імітаційного моделювання, методи теорії штучного інтелекту, стохастичні і евристичні методи пошуку, методи прогнозування, методи математичної статистики, сервісні моделі хмарних обчислень. Достовірність та обґрунтованість отриманих результатів обумовлені коректним використанням математичного апарату, а також підтверджуються результатами обчислювальних експериментів.

Наукова новизна одержаних результатів Основний науковий результат дисертаційної роботи полягає у вирішенні наукової проблеми ефективного управління ІТ-інфраструктурою хмарного ЦОД на основі розроблення інформаційної технології, що базується на теоретико-методологічних положеннях, моделях і методах процесів управління ІТ-інфраструктурою як складовою інформаційної системи провайдера.

Основні положення дисертаційної роботи, що визначають її наукову новизну, полягають у наступному:

– *вперше:*

1) розроблено методологію управління ІТ-інфраструктурою хмарного ЦОД на основі операторної форми постановки, аналізу і розв'язання задач управління в умовах невизначеності і змінних навантажень, яка відрізняється вибором стратегій управління в різних режимах роботи хмарного ЦОД за рахунок визначення відповідних моделей через комбінування критеріїв і обмежень, а також за рахунок визначення відповідних схем реалізації, що дозволяє при управлінні хмарним ЦОД перейти від традиційного підходу з одним визначеним методом до програмно-визначеного підходу з використанням множини відповідних схем реалізації, моделей і методів;
2) розроблено структурно-функціональну модель багаторівневої ієрархічної програмно-визначеної системи управління ІТ-інфраструктурою хмарного ЦОД через поєднання стратегічного і оперативного управління з автоматичним

керуванням, а також надання програмно-визначених властивостей, яка дозволяє реалізувати задане управління ресурсами і навантаженням ІТ-інфраструктури з вибором стратегій, плануванням і управлінням їх реалізацією, автоматичним керуванням з урахуванням структури системи, ретроспективних даних її функціонування та дотриманням заданих показників якості угоди про рівень обслуговування;

3) розроблено оригінальну інформаційну технологію управління ІТ-інфраструктурою хмарного ЦОД на основі запропонованої методології, моделей і методів яка відрізняється адаптивністю та використанням стратегічного управління з прогнозуванням за критеріями енергозбереження, забезпечення заданого рівня обслуговування, зниження операційних і капітальних витрат, що дозволяє підвищити ефективність використання ресурсів в умовах змінних навантажень;
4) розроблено адаптивний метод комбінованого прогнозування навантаження на обчислювальні ресурси хмарного ЦОД з використанням усередненого, зваженого та оптимізаційного комбінування оцінок прогнозів, обчислених за альтернативними методами прогнозування та з адаптацією розміру навчальної вибірки, який відрізняється обчисленням та використанням певних типів комбінованих прогнозів (усередненого та зваженого) в реальному часі, отриманих за множиною альтернативних методів, що дає можливість оцінювати високоякісне прогнозоване значення для відомих видів змішаних навантажень в ІТ-інфраструктурі;
5) розроблено метод інтегрованого управління ресурсами ІТ-інфраструктури хмарного ЦОД на основі динамічної моделі його станів із застосуванням стохастичного пошуку для виконання міграцій віртуальних машин, вивільнення, увімкнення і вимкнення фізичних серверів, а також прогнозування для адаптації до змін навантаження з метою досягнення сталого режиму роботи згідно визначених критеріїв, що дозволяє ефективніше управляти ресурсами ІТ-інфраструктури хмарного ЦОД із дотриманням заданих показників якості угоди про рівень обслуговування;
6) розроблено метод управління розподіленим сховищем хмарного ЦОД на основі моделі дворівневого сховища з реплікацією та урахуванням кешування за розміром файлу і за кількістю транзакцій доступу до файлу для управління міграцією і з урахуванням кількості транзакцій доступу до блоків даних, об’єму вільного місця та затримки передачі даних між вузлами зберігання даних для управління реплікацією, який відрізняється застосуванням принципу гіперконвергентності, що забезпечило підвищення рівня надійності збереження даних, відмовостійкості та продуктивності їх оброблення в сучасних апаратно-програмних комплексах ЦОД;
7) розроблено метод управління потужністю хмарного ЦОД на основі динамічної моделі його станів, який використовує запропоновані метрики оцінки стану хмарного ЦОД (коефіцієнт життєздатності віртуальної машини, індикатор дисбалансу фізичного сервера, коефіцієнт відношення необхідних ресурсів до середнього об’єму наявних ресурсів, поріг вільних ресурсів та метрика ємності ЦОД), враховує гетерогенність фізичних серверів, їх тип і кількість, який на відміну від існуючих підвищує якість обслуговування користувачів і зменшує споживання електроенергії, що дає можливість автоматично забезпечити ресурс потрібного типу, а також мінімальну затримку розгортання сервісів у хмарі;

– *отримали подальший розвиток:*

8) декомпозиційно-компенсаційний підхід через надання адаптивності, багаторівневості та програмно-визначених властивостей за рахунок реалізації ефективного управління ресурсами IT-інфраструктури хмарного ЦОД з переходом від вибору стратегій до планування і управління їх втіленням з урахуванням структури системи та її параметрів;
9) алгоритми і методи стохастичного пошуку через розроблення нових методів управління ресурсами IT-інфраструктури, а саме методу рівномірної консолідації віртуальних машин з використанням ідеї імітації відпалу, двостадійного методу управління ресурсами хмарного ЦОД на основі алгоритму променевого пошуку, методу динамічної консолідації і розміщення віртуальних машин на основі алгоритму навчання з підкріпленням, що відрізняються врахуванням процедур визначення станів системи на основі нових метрик, врахуванням достатньої кількості видів ресурсів і дискретності вимірів, що дозволяє ефективно визначати схему реалізації стратегії управління для різних режимів роботи хмарного ЦОД;
10) модель управління IT-інфраструктурою з координатором через застосування програмно-визначеного керування підсистемами, де координатор генерує керуючі впливи для програмно-визначених контролерів мережі, сховища і гіпервізорів, погоджуючи їх роботу з визначеною стратегією управління хмарним ЦОД, що дозволило застосовувати її в середовищах з програмно-визначеним функціонуванням.

Практичне значення одержаних результатів. Розв’язані у дисертаційному дослідження завдання, розроблені методи і підходи становлять методичну базу розроблення і реалізації систем управління IT-інфраструктурою хмарних ЦОД і підвищення ефективності їх функціонування. Створена інформаційна технологія може бути застосована для розроблення підсистем, компонентів та інших складових систем управління IT-інфраструктурою провайдерів хмарних послуг.

До числа результатів, які мають найбільше практичне значення, належать: методологія управління на основі операторної форми вибору стратегії управління і схеми її реалізації; підхід до управління багаторівневою ієрархічною системою ресурсів IT-інфраструктури хмарного ЦОД; методи прогнозування навантаження хмарного ЦОД; методи управління ресурсами, навантаженням і потужністю хмарного ЦОД з урахуванням прогнозів; методи управління реплікацією та міжрівневою міграцією даних у сховищі хмарного ЦОД.

Розроблені теоретичні положення, методи і алгоритми використані при: модернізації систем управління функціонуванням інформаційно-телекомунікаційної інфраструктури в компаніях-членах Асоціації «ТЕЛАС»; розробленні системи управління функціонуванням інформаційно-телекомунікаційної інфраструктури ТОВ «АМ ІНТЕГРАТОР ГРУП»; модернізації системи управління IT-інфраструктурою ТОВ «СІТІУС ПРО». Впровадження результатів досліджень дозволило на 28% скоротити операційні витрати на управління IT-інфраструктурою без порушень SLA; зменшити споживання електроенергії на 17% в середовищі з гомогенними конфігураціями фізичних серверів; скоротити в середньому на 18% кількість фізичних серверів, що обслуговують навантаження клієнтів; зменшити кількість порушень SLA на 27% при наданні IT-послуг; скоротити витрати на експлуатацію серверного парку на 19% при забезпеченні виконання заданих вимог угоди про рівень обслуговування.

Теоретичні і практичні результати дисертаційної роботи склали основу нових спецкурсів, що викладаються автором на кафедрі автоматизованих систем обробки

інформації та управління Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”: “Технології віртуалізації та хмарних обчислень”, “Сучасні технології розроблення програмного забезпечення”, “Інтелектуальні системи управління технічними пристроями”, “Програмування інтернету речей”, “Методи та системи штучного інтелекту”, “Обробка надвеликих масивів інформації”.

Особистий внесок здобувача. Усі наукові результати дисертаційної роботи отримані автором самостійно у друкованих працях: [6, 17, 7] – теоретичне обґрунтування еволюційних процесів розвитку і становлення сучасної ІТ-інфраструктури та підходів до її оптимізації, [21] – двостадійний метод управління консолідацією віртуальних машин хмарного ЦОД з використанням променевого пошуку, [22] – метод консолідації віртуальних машин хмарного ЦОД з використанням модифікованого методу навчання з підкріпленням, [23] – модель динаміки хмарного ЦОД, яка враховує різні типи фізичних серверів, віртуальних машин і ресурсів в гетерогенній ІТ-інфраструктурі хмарного ЦОД, [24] – метод інтегрованого управління ресурсами ІТ-інфраструктури ЦОД, що забезпечує необхідну еластичність на рівні фізичного сервера з урахуванням споживання енергії та з урахуванням порушень угоди про рівень обслуговування і заснований на моделі хмарного ЦОД в просторі станів, моделі енергоспоживання, моделі порушень SLA і методі управління ємністю хмарного ЦОД, [25] – адаптивний метод комбінованого прогнозування навантаження, який забезпечує меншу помилку прогнозу у порівнянні з методом прогнозу за моделлю, отриманою на основі навчальної вибірки фіксованого розміру, [26, 5] – метод управління і модель дворівневого сховища з реплікацією, [2, 21, 23, 24] – метрики оцінювання стану і динаміки змін ресурсів хмарного ЦОД, [15, 16] – прикладні аспекти використання ІТ-інфраструктури для розміщення і оброблення знань в СППР.

У друкованих працях, опублікованих у співавторстві, автору належать: [12, 8] – архітектурні рішення щодо побудови апаратно-програмних комплексів для реалізації прикладних інформаційних процесів, [11, 13] – підхід до оптимізації передачі даних в мережі провайдера хмарних послуг та його прикладні аспекти, [3] – концептуальні основи інформаційної технології управління ресурсами ІТ-інфраструктури ЦОД, структурно-функціональна модель типової програмно-визначеної системи управління і адаптивний програмно-визначений підхід до розподілу віртуальних машин хмарного ЦОД, який полягає в управлінні фізичними і віртуальними ресурсами через вибір стратегій управління з метою адаптації до зовнішніх впливів, [9] – підхід до управління ІТ-інфраструктурою інтернету речей із забезпеченням заданої якості надання сервісу при раціональному використанні ресурсів хмарного ЦОД, методи управління ресурсами екосистеми інтернету речей з архітектурою мікрохмари, [18] – стратегії управління і алгоритми адаптивного програмно-визначеного методу до розподілу ресурсів ІТ-інфраструктури між віртуальними машинами хмарного ЦОД, [1] – підхід до організації крайових (Edge) обчислень на основі архітектури системи інтернету речей у вигляді мікрохмари, розвиток декомпозиційно-компенсаційного підходу при управлінні ІТ-інфраструктурою екосистеми інтернету речей, [2] – метод динамічного розподілу ресурсів ІТ-інфраструктури між віртуальними машинами хмарного ЦОД на основі стохастичного локального пошуку з використанням різних евристик, урахуванням багатовимірності, стохастичності, збалансованого навантаження на фізичний сервер, обмеження на кількість одночасних міграцій і обмеження апаратного забезпечення, [19] – метод консолідації віртуальних машин хмарного ЦОД на базі променевого пошуку,

[20] – розвиток принципів координації керуючих впливів при управлінні рівнем послуг в екосистемі інтернету речей, [4] – дворівнева система управління гіперконвергентною інфраструктурою ЦОД на основі принципів координації, [10] – двостадійний метод управління ресурсами хмарного ЦОД на базі променевого пошуку, [14] – прикладні аспекти використання ІТ-інфраструктури для розміщення і оброблення наукових публікацій, [27] – підхід до декомпозиції управління технологіями, застосунками і процесами в складі ІТ, [28] – операторна форма постановки, аналізу і розв'язання задач управління ІТ-інфраструктурою хмарного ЦОД.

Апробація результатів дисертації. Основні результати наукових досліджень доповідалися та обговорювалися на закордонних та міжнародних наукових конференціях і форумах, зокрема: «IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications» (**Metz, France, 2019**), «Університет і регіон: проблеми сучасної освіти» (м. Луганськ, 2009), «Комп'ютерні науки для інформаційного суспільства» (м. Луганськ, 2010, 2011, 2012, 2013), «Информационно-компьютерные технологии в экономике, образовании и социальной сфере» (м. Сімферополь, 2013), «Комп'ютерні інтелектуальні системи та мережі» (м. Кривий Ріг, 2015, 2016, 2017), «ABIA-2015» (м. Київ, 2015), «Computer Science and Information Technologies (CSIT)» (м. Львів, 2016, 2017, 2018), «International Conference Radio Electronics & Info Communications (UkrMiCo)» (м. Київ, 2016, м. Одеса, 2017, 2018), «Problems of Infocommunications Science and Technology (PIC S&T)» (м. Харків, 2016, 2017, 2018), «World Forum on Internet of Things (WF-IoT)» (**Reston, VA, USA, 2016**), «Cloud Computing, GRIDs, and Virtualization» (**Athens, Greece, 2017**), «Інформатика та обчислювальна техніка (ІОТ)» (м. Київ, 2017, 2018), «Electrical and Computer Engineering (UKRCON)» (м. Київ, 2017), «Automatic Control and Information Technology (ICACIT'17)» (**Cracow, Poland, 2017, 2019**), «Computer Science, Engineering and Education Applications (ICCSEEA)» (м. Київ, 2018, 2019), «Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)» (Lviv-Slavske, Ukraine, 2018). Результати дисертаційних досліджень доповідались на наукових семінарах: кафедри автоматики та управління в технічних системах, кафедри математичних методів системного аналізу Національного технічного університету України "Київський політехнічний інститут імені Ігоря Сікорського", кафедри інтелектуальних та інформаційних систем Київського національного університету імені Тараса Шевченка.

Публікації. За результатами досліджень опубліковано 71 наукова праця, у тому числі 1 монографія, 27 статей у наукових фахових виданнях (з них 9 статей у виданнях іноземних держав, 7 у виданнях України, які включені до міжнародних наукометричних баз), 43 тези доповідей в збірниках матеріалів конференцій.

Структура та обсяг дисертації. Дисертаційна робота складається зі вступу, восьми розділів, висновків, списку використаних джерел із 353 найменувань та 4 додатків. Загальний обсяг дисертації становить 402 сторінки, з яких 383 сторінки основного тексту, 75 рисунків, 20 таблиць.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність теми дисертації – розроблення інформаційної технології управління (ІТУ) ІТ-інфраструктурою хмарного центру оброблення даних, сформульовано мету, задачі, об'єкт, предмет та методи дослідження, показано зв'язок з науковими програмами, визначено наукову новизну та практичне значення одержаних

результатів, наведено дані про впровадження результатів, публікації та особистий внесок здобувача.

У **першому розділі** виконано аналіз проблеми забезпечення ефективного функціонування ІТ-інфраструктури хмарного ЦОД в цілому, а також проблем управління ресурсами і підсистемами. Проаналізовано особливості надання хмарних послуг в Україні і світі, сучасні технології, методи і моделі ієрархічного, інтегрованого та програмно-визначеного управління ресурсами ІТ-інфраструктури ЦОД та інфраструктури інтернету речей на основі методів штучного інтелекту, стохастичного пошуку та прогнозування. На основі системного підходу проаналізовано інфраструктуру хмарних ЦОД як об'єктів керування, сформульовано науково-практичну проблему та задачі дослідження.

Аналіз наукових праць вітчизняних і зарубіжних авторів показав, що ІТ-інфраструктура хмарних ЦОД знаходиться в стадії інтенсивного розвитку і модернізації, що потребує покращення роботи існуючих систем управління ІТ-інфраструктурою (СУІ) і впровадження нових підходів і методів для підвищення ефективності систем управління ІТ-інфраструктурою з метою забезпечення заданої якості надання послуг при раціональному використанні ресурсів. Складність організації процесів взаємодії елементів ІТ-інфраструктури не дозволяє створити їх адекватні моделі без додаткових обмежень, інструментальних засобів і галузевих стандартів. Системні властивості елементів ІТ-інфраструктури не досліджені належним чином.

Виконаний в дисертації аналіз основних концепцій і підходів до управління ІТ-інфраструктурою хмарного ЦОД дозволив зробити такі висновки:

- сучасний ЦОД, що реалізує концепцію хмарних обчислень, характеризується складністю, ієрархічністю, багатовимірністю і багатоаспектністю;
- існуючі підходи і методи управління ІТ-інфраструктурою хмарного ЦОД часто не передбачають врахування взаємозв'язків і взаємного впливу на всіх рівнях ієрархії побудови інформаційних процесів, не адаптуються до інтенсивності використання та вивільнення ресурсів, а орієнтуються лише на вирішення певної задачі на певному рівні (інфраструктури, платформи або програмного забезпечення (ПЗ);
- недостатньо уваги приділяється розробленню методів і моделей управління ІТ-інфраструктурою з використанням прогнозів споживання обчислювальних ресурсів і прогнозів навантаження на елементи інфраструктури, а використані методи і моделі прогнозу орієнтовані на певну комбінацію навантажень і не адаптуються до поточних умов роботи хмарного ЦОД, які характеризуються невизначеністю;
- недостатньо уваги приділяється розробленню інтегрованих методів управління ІТ-інфраструктурою, які б враховували одночасно аспекти віртуалізації, програмно-визначеного функціонування ІТ-інфраструктури, особливості взаємодії обчислювальних ресурсів, сховищ даних і мереж, а також впливу гіперконвергентних архітектур.

Результати проведеного аналізу обумовлюють необхідність вирішення проблеми забезпечення ефективного функціонування ІТ-інфраструктури хмарного ЦОД через створення методології управління та на її основі відповідної ІТУ із застосуванням новітніх підходів, моделей і методів, що базуються на методах інтегрованого і ієрархічного управління, штучного інтелекту, моделях і методах прогнозування навантажень і споживання ресурсів ІТ-інфраструктури.

Таким чином, розроблення методології управління і на її основі ІТУ ІТ-інфраструктурою хмарного ЦОД передбачає розв'язання складного комплексу задач управління і керування на всіх рівнях архітектури, в їх взаємозв'язку, з урахуванням стратегічного і оперативного планування, виконання схем реалізації у поєднанні з управлінням в реальному часі.

Результати за розділом наведено у [12, 17-30, 33, 35, 40, 68].

У **другому розділі** обґрунтовано необхідність врахування суттєвих характеристик хмарних обчислень при розробленні ІТУ ІТ-інфраструктурою хмарного ЦОД; розроблено методологію управління ІТ-інфраструктурою хмарного ЦОД на основі операторної форми постановки, аналізу і розв'язання задач управління в умовах невизначеності і змінних навантажень; операторний підхід до управління ІТ-інфраструктурою хмарного ЦОД; узагальнену схему реалізації функцій управління ІТ-інфраструктурою хмарного ЦОД; запропоновано розглядати ІТ-інфраструктуру хмарного ЦОД як нелінійний нестаціонарний дискретний об'єкт зі змінною структурою (ННДОЗС); розроблено схему реалізації операторної форми управління ІТ-інфраструктурою; розроблено структурно-функціональну модель СУІ хмарного ЦОД; розроблено задачі і стадії управління ІТ-інфраструктурою хмарного ЦОД з урахуванням ієрархії підсистем і стратегій управління; розроблені стратегії управління і моделі вибору стратегій при управлінні ресурсами хмарного ЦОД; обґрунтовано необхідність використання інтегрованих рішень і гіперконвергентних систем при розгортанні хмарних ЦОД.

Поставлені задачі обумовлюють розроблення методології управління ІТ-інфраструктурою в умовах невизначеності та в її основі відповідних моделей і методів які, з одного боку забезпечать задану якість надання хмарних послуг в ієрархічній структурі інформаційних процесів, а з іншого – адаптацію до змін навантаження через прогнозування споживання обчислювальних ресурсів ІТ-інфраструктури.

Узагальнена схема управління ІТ-інфраструктурою хмарного ЦОД, для якого розроблено ІТУ, показана на рис. 1. Хмарний ЦОД являє собою ієрархічну програмно-апаратну систему, що складається з трьох рівнів, які відповідають трьом сервісним моделям надання хмарних послуг. Запропонована СУІ включає в себе: підсистему диспетчеризації, моніторингу і прогнозування; підсистему визначення стану ЦОД; підсистему визначення стратегій управління і вибору схем їх реалізації; підсистему генерації керуючих впливів, що впливають на програмно-апаратні засоби ЦОД.

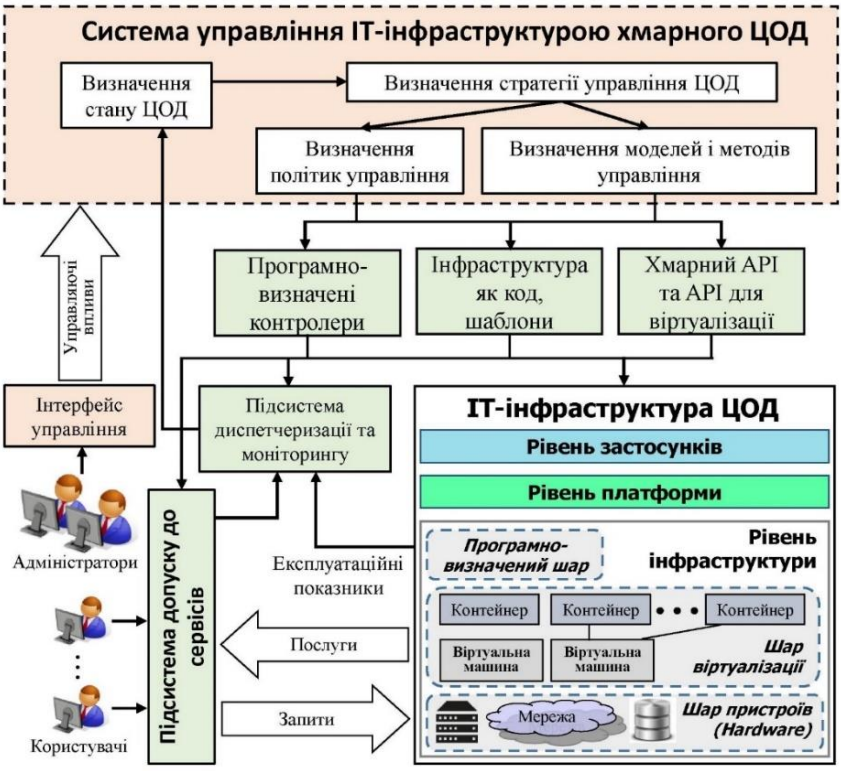

Рис. 1. Узагальнена схема управління ІТ-інфраструктурою хмарного ЦОД.

Стан ІТ-інфраструктури хмарного ЦОД, як складної багаторівневої системи управління, пропонується визначати в трьох вимірах: вимір ресурсів (надлишок або нестача); вимір навантажень (змінне або стале); вимір динаміки навантажень (з трендом на збільшення або на зменшення). Належність поточного стану до певних значень цих вимірів визначається сукупністю метрик, обчислених з використанням даних моніторингу, і впливає на визначення стратегії управління ресурсами ІТ-інфраструктури. При цьому, ІТ-інфраструктура хмарного ЦОД включає в себе: мережу ЦОД, сховище ЦОД і множину ФС, які безпосередньо обробляють навантаження згідно сервісних моделей IaaS (Інфраструктура як сервіс), PaaS (Платформа як сервіс), SaaS (ПЗ як сервіс). Для забезпечення ефективного управління ресурсами хмарних ЦОД, необхідного рівня сервісу, масштабування і адаптації до розгортання нових послуг, пропонується *структурно-функціональна модель типової програмно-визначеної СУІ хмарного ЦОД*, яка показана на рис. 2.

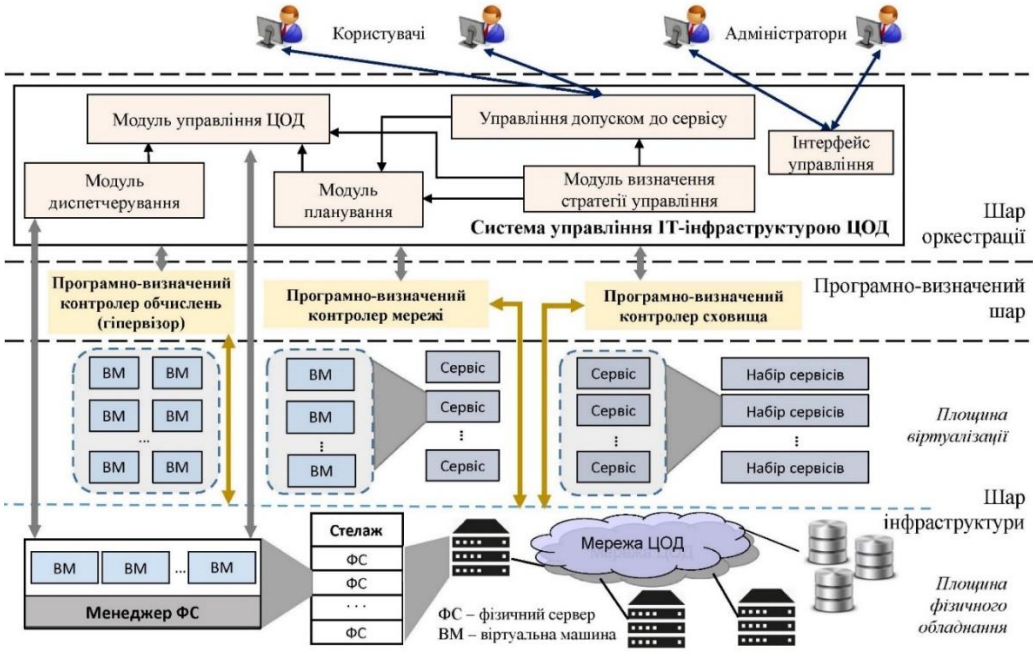

Рис. 2. Структурно-функціональна модель типової програмно-визначеної СУІ хмарного ЦОД.

В дисертації визначені функції кожного шару запропонованої моделі (оркестрації, програмно-визначеного шару і шару інфраструктури), вхідні і вихідні дані, інтерфейси, а також функціональність компонентів і підсистем у зв'язку із розробленими в розділах 3-7 ІТУ ІТ-інфраструктурою хмарного ЦОД, підходами, стратегіями, методами, моделями і алгоритмами.

Сформульовано загальну постановку задачі управління ресурсами ІТ-інфраструктури хмарного ЦОД яка відноситься до класу задач управління ННДОЗС. Для кожного кроку управління *t* визначена матриця $\mathbf{A}: m \times n$ (1), яка позначає загальний стан системи (ЦОД). Її елементи $a_{ij}(t) \in \{0,1\}$ вказують, чи працює *j*-а ВМ з множини *V* на *i*-му ФС з множини *P* ($a_{ij}(t) = 1$), або не використовується ($a_{ij}(t) = 0$), $n=N(t)$, $m=M(t)$. Кількість змін значень в стовпчиках матриці **A** з 0 на 1 дорівнює кількості завершених міграцій ВМ між ФС, $\sum_{i=1}^{m} a_{ij} = 1, \forall n$.

$$\mathbf{A}(t) = \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22(t)} & \cdots & a_{2n}(t) \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1}(t) & a_{m2}(t) & \cdots & a_{mn(t)} \end{bmatrix} \quad (1)$$

Наступний стан системи визначається рівнянням $\mathbf{A}(t+1) = F\left[\mathbf{A}(t), \hat{N}(t+1), U(t), Z(t)\right]$, де $F[\bullet]$ є функціоналом управління, який реалізується обраною стратегією управління; $U(t) = \{k(t), p(t)\}$ – керуючі впливи; $\hat{N}(t+1) = N(t) + \hat{q}(t) - \hat{r}(t) - \varepsilon(t+1)$ – кількість ВМ на наступному кроці управління, $Z(t)$ – план міграцій ВМ між ФС, $q(t)$ – кількість нових ВМ для розміщення; $r(t)$ – кількість ВМ, які зупинені; $p(t)$ – кількість ВМ, які мігрують; $k(t)$ – кількість увімкнених/вимкнених ФС; $\delta(t+1), \varepsilon(t+1)$ – кількість ФС і ВМ відповідно, які вийшли з ладу, $\delta, \varepsilon \geq 0$. Кількість ФС на наступному кроці визначається рівнянням $M(t+1) = M(t) + k(t) - \delta(t+1)$. Критерії, які визначаються обраною стратегією управління: мінімізація кількості активних ФС, $M(t)$; мінімізація міграцій ВМ, $p(t)$; мінімізація кількості випадків повного завантаження ресурсу; мінімізація затримки розгортання ВМ та ін. Обмеження, які враховано при реалізації різних стратегій: кількість вхідних/вихідних міграцій на ФС, обмеження використання кожного ресурсу в межах ФС, обмеження швидкості мережевої взаємодії, обмеження швидкості роботи з даними в системі збереження даних (СЗД) та ін.

При цьому, в залежності від поточного стану хмарного ЦОД і визначеної стратегії управління, склад критеріїв і обмежень змінюються під впливом технологічних особливостей хмарного ЦОД та необхідності надання сервісів із зазначеними показниками якості. Таким чином, кожного разу обирається стратегія управління через побудову нової моделі хмарного ЦОД і застосовування різних методів і алгоритмів управління у вигляді схем реалізації з метою адаптації до поточних умов функціонування і зменшення невизначеності.

Зазначені моделі і положення лягли в основу процесу розроблення методології управління, яка забезпечує розв'язання класу задач управління ІТ-інфраструктурою хмарного ЦОД за допомогою операторної форми. Цільову функцію управління ІТ-інфраструктурою хмарного ЦОД в операторній формі представлено у такий спосіб:

$$\min\left[\mathbf{C}_1\mathbf{F}_i\left[P,V\right] + \mathbf{C}_2\mathbf{M}_i\left[P,V\right] + \mathbf{C}_3\mathbf{L}_i\left[P,V\right]\right], \quad (2)$$

де $\mathbf{F}_i\left[P,V\right] = (b_1^F, b_2^F, ..., b_m^F)^T$ – оператор i-ї стратегії, що визначає ФС з множини P з когерентними ВМ з множини V, $b^F \in \{0,1\}$; $\mathbf{M}_i\left[P,V\right] = (b_1^M, b_2^M, ..., b_m^M)^T$ – оператор i-ї стратегії, що визначає розподіл некогерентних ВМ з множини V між ФС з множини P, $b^M \in \{0,1\}$; $\mathbf{L}_i\left[P,V\right] = (b_1^L, b_2^L, ..., b_m^L)^T$ – оператор i-ї стратегії, що визначає ФС з множини P які переведені в режим сну, $b^C \in \{0,1\}$; $\mathbf{C}_1 = (c_1^1, c_2^1, ..., c_m^1)$ – вектор коефіцієнтів витрат на підтримку роботи ФС; $\mathbf{C}_2 = (c_1^2, c_2^2, ..., c_m^2)$ – вектор коефіцієнтів витрат на підтримку роботи ФС, виділених для обслуговування некогерентних ВМ; $\mathbf{C}_3 = (c_1^3, c_2^3, ..., c_m^3)$ – вектор коефіцієнтів витрат на підтримку ФС у вимкнутому стані. Результатом цільової функції управління є значення операційних витрат на підтримку роботи хмарного ЦОД. Результатом дії операторів **F** і **M**, які отримують на вході вектори характеристик ФС і ВМ, є матриця розподілу ВМ між ФС $\mathbf{A}(t+1)$. Результатом дії оператора **L**, який отримує на вході метрики стану хмарного ЦОД, є план міграцій ВМ $Z(t)$ і керуючі впливи $U(t)$.

Схеми реалізації оператора залежать від історичних даних, від значень метрик та від результатів прогнозування і являють собою певну визначену стратегію управління. Параметри моделі управління, склад обмежень і критеріїв визначаються в процесі управління. Стратегії S управління визначаються множиною моделей $\Omega(D,O,G)$ і

методів управління з множиною змінних D, множиною обмежень O і множиною критеріїв G які визначаються в процесі функціонування системи. Кожна i-та стратегія управління $S_i = \Omega(D_i, O_i, G_i), D_i \in D, O_i \in O, G_i \in G$ реалізує певний оператор цільової функції (2). При цьому кожна підмножина складається з довільної кількості елементів, $D_i = \{d_{0,i}, d_{1,i}, \ldots\}$, $O_i = \{o_{0,i}, o_{1,i}, \ldots\}, G_i = \{g_{0,i}, g_{1,i}, \ldots\}$, $D_i \neq \varnothing, O_i \neq \varnothing, G_i \neq \varnothing \ \forall i$.

Таким чином, схема реалізації операторів функції (2) визначається системою.

$$\begin{cases} \mathbf{F}_i[P,V]: \langle \ \sum_{i=1}^{N(t)} res_i^{VM}, \sum_{j=1}^{M(t)} Res_j^{PM}, \Omega(D_i, O_i, G_i) \ \rangle \to \mathbf{A}(t+1) \\ \mathbf{M}_i[P,V]: \langle \ K, \Omega(D_i, O_i, G_i) \ \rangle \to M(t+1) \\ \mathbf{L}_i[P,V]: \langle \ K, \Omega(D_i, O_i, G_i) \ \rangle \to Z(t) \end{cases}, \quad (3)$$

де $\mathbf{F}_i$, $\mathbf{M}_i$, $\mathbf{L}_i$ – оператори цільової функції (2), res_i^{VM} – ресурси, які споживає i-та ВМ, Res_i^{PM} – ресурси, які споживає j-й ФС, K – множина метрик оцінювання стану хмарного ЦОД, $\Omega(D_i, O_i, G_i), D_i \in D, O_i \in O, G_i \in G$ – вибрана модель в процесі управління ресурсами хмарного ЦОД. Конкретні значення D_i, O_i, G_i вибираються відповідно до конфігурації хмарного ЦОД, значень табл. 1 і вимог реалізації кожного методу управління.

Схема реалізації завдань управління для визначеної стратегії реалізується методами, розробленими в розділах 3, 4, 5. Застосування тієї чи іншої схеми реалізації стратегії, моделі або їх комбінацій у залежності від стану ресурсів і навантажень хмарного ЦОД показано у табл. 1.

З метою автоматичного визначення стратегії управління в дисертації запропонована модель нечіткого виводу, яка дозволяє визначити стратегію управління (значення вихідної змінної X) в залежності від метрик оцінки стану хмарного ЦОД (n вхідних змінних множини K). Вхідні змінні моделі: K_1 – середній коефіцієнт життєздатності ВМ, K_2 – індикатор дисбалансу ФС, K_3 – коефіцієнт відношення необхідних ресурсів до середнього об'єму наявних ресурсів, K_4 – поріг вільних ресурсів та K_5 – метрика ємності хмарного ЦОД. При цьому, метрики K_1, K_2 K_3 вперше запропоновані в дисертації. Вихідна змінна приймає значення, відповідні назвам стратегій: S_1 – з нестачею ресурсів при сталому навантаженні, S_2 – з надлишком ресурсів при сталому навантаженні, S_3 – з нестачею ресурсів і трендом на зменшення навантаження, S_4 – з надлишком ресурсів і трендом на зменшення навантаження, S_5 – з нестачею ресурсів і трендом на збільшення навантаження, S_6 – з надлишком ресурсів і трендом на збільшення навантаження. За допомогою правил виду "Якщо $K_1 \in A_1 \wedge K_2 \in A_2 \wedge \ldots \wedge K_n \in A_n$, то $X \in S_i$", де A_n – множина значень вхідної змінної K_n, реалізований нечіткий вивід Мамдані для визначення стратегії управління хмарним ЦОД на поточному кроці управління.

Для побудови функцій належності використовується підхід, представлений в роботі [35], який модифікований через розроблення нової методики вибору піків, яка полягає у поєднанні декількох сусідніх інтервалів для визначення піку. Крім того, розроблено модель "гри з природою" з використанням вищезазначених змінних K і S, де перший гравець – модуль вибору стратегії управління хмарним ЦОД, а другий гравець – "природа" у вигляді навантаження на ресурси хмарного ЦОД. При дослідженні обох моделей перевагу віддано моделі "гри з природою".

Табл. 1. Вибір стратегії і застосування схем реалізації управління ресурсами у залежності від стану ЦОД.

Передумови	Стратегії	Схеми реалізації стратегій	Базові методи
Середній коефіцієнт життєздатності ВМ; індикатор дисбалансу ФС; коефіцієнт відношення необхідних ресурсів до середнього об’єму наявних ресурсів; поріг вільних ресурсів; метрика ємності ЦОД.	Управління з нестачею ресурсів при сталому навантаженні, S_1	Рівномірна консолідація ВМ.	Модифікований метод відпалу
	Управління з надлишком ресурсів при сталому навантаженні, S_2	Управління міграцією та розміщенням ВМ в сталому режимі. Двостадійний метод управління ресурсами.	Модифікований метод променевого пошуку.
	Управління з нестачею ресурсів і трендом на зменшення навантаження, S_3	Інтегроване управління ресурсами (ІУР).	Метод інтегрованого управління ресурсами
	Управління з надлишком ресурсів і трендом на зменшення навантаження, S_4	Інтегроване управління ресурсами.	Метод інтегрованого управління ресурсами. Метод управління потужністю ЦОД.
	Управління з нестачею ресурсів і трендом на збільшення навантаження, S_5	Динамічна консолідація і розміщення ВМ.	Модифікований метод навчання з підкріпленням (НП).
	Управління з надлишком ресурсів і трендом на збільшення навантаження, S_6	Динамічна консолідація і розміщення ВМ з управлінням потужністю	Модифікований метод НП. Метод управління потужністю ЦОД.

Таким чином, методологія управління ІТ-інфраструктурою хмарного ЦОД об’єднує запропоновані в дисертації стратегії, методи, моделі і алгоритми, а також існуючі принципи і технології.

Результати досліджень, подані в цьому розділі, опубліковано у працях [6, 7, 3, 4, 24, 27, 28, 61, 37, 40, 42].

У **третьому розділі** розроблено схему реалізації стратегій управління ресурсами ІТ-інфраструктури; структуру ієрархічної системи управління хмарним ЦОД на основі декомпозиції хмарної ІТ-інфраструктури на рівні інфраструктури, платформи та застосунків; розроблено архітектуру і метод управління сервісами в ієрархічній системі хмарного ЦОД; обґрунтовано необхідність інтегрованого управління ресурсами ІТ-інфраструктури хмарного ЦОД; розроблено модель динаміки хмарного ЦОД на основі простору станів та інтегровану модель, що враховує споживання електроенергії, порушення умов SLA і планування ресурсної потужності; сформульовано і розв’язано оптимізаційну задачу перерозподілу ресурсів в ІТ-інфраструктурі в складі методу інтегрованого управління ресурсами (ІУР); розроблено схему дворівневої системи управління гіперконвергентною ІТ-інфраструктурою і підходу до адаптивного розміщення ВМ з використанням прогнозу навантаження.

Складність хмарного ЦОД визначається тим, що з плином часу: зростає кількість інформаційних процесів на кожному рівні в ієрархії; управляючі впливи на програмне та апаратне забезпечення генеруються різними користувачами та різними підсистемами управління; з’являються нові інформаційні процеси, нове програмне забезпечення та обладнання на кожному рівні в ієрархії. Для зменшення впливу складності в дисертаційній роботі виконано декомпозицію хмарного ЦОД на рівні, що складаються з процесів, якими можна керувати окремо, беручи до уваги різні цілі.

Застосування операторного підходу для розроблення ІТУ ІТ-інфраструктурою хмарного ЦОД обумовлено високою складністю програмно-апаратної платформи,

гетерогенністю програмного і апаратного забезпечення, наявністю декількох видів послуг на одному рівні ієрархії, необхідністю забезпечення високої надійності та безпеки, різною швидкістю плину інформаційних процесів, необхідністю використання різних моделей, критеріїв і обмежень на кожному рівні ієрархії.

За результатами декомпозиції хмарного ЦОД на основі моделей хмарних обчислень, запропонованих NIST, SaaS, PaaS і IaaS розроблено структурно-функціональну модель багаторівневої ієрархічної СУІ для керування ІТ-послугами за критерієм мінімізації витрат і мінімізації відхилень параметрів якості послуг від заданих.

Розроблена модель складається з трьох рівнів (рівень інфраструктури, рівень платформи та рівень ПЗ), як показано на рис. 3, і дозволяє виділити оптимальну кількість ресурсів для реалізації ІТ-послуг на кожному рівні системи таким чином, щоб ефективно керувати послугами та застосунками на вищих рівнях, впливаючи на параметри продуктивності рівня інфраструктури та впливаючи на конфігурації та структури рівня платформи та прикладного рівня. Модель враховує дві моделі споживання послуг: моделі хмарних сервісів і моделі послуг хостингу. Провайдер хмарних послуг надає користувачам змогу створювати окремі інфраструктури I на кожному рівні.

Рівень інфраструктури представлений набором інфраструктур I^I і відповідає за управління фізичними ресурсами, включаючи ФС, стійкові комутатори, обладнання мережі та систем зберігання даних. Рівень інфраструктури надає послуги за запитом з платформного рівня, впливає на конфігурації та визначає управління, обумовлене процесами на рівні платформи. Функції управління, такі як оновлення, відображення віртуальних ресурсів на фізичні, підключення до мережі, забезпечення роботи СЗД виконуються провайдером, використовуючи керуючі впливи U^I. Функції управління, такі як конфігурація віртуального ресурсу, вибір шаблону об'єкта, визначення розмірів ВМ, використання мережі та СЗД, виконуються користувачем за допомогою керуючих впливів U^{I*}. Об'єкт управління *на рівні платформи* представлений елементом з набору платформ I^P. Основна функція рівня платформи – створення екземпляру середовища для запуску сервісів і застосунків. Рівень платформи повністю управляється і регулярно оновлюється провайдером послу, використовуючи впливи управління U^P. Користувач керує розгорнутими застосунками та конфігураціями контейнерів, використовуючи

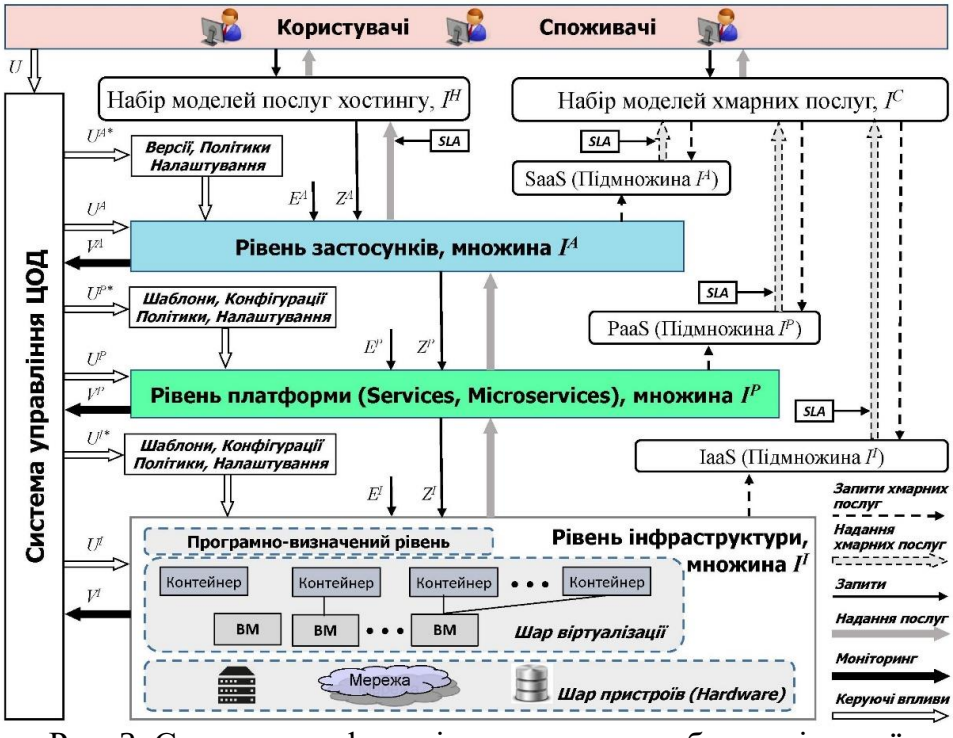

Рис. 3. Структурно-функціональна модель багаторівневої ієрархічної СУІ.

впливи керування U^{P*}, такі як керування сеансами та контентом, інтеграцією пристрою та різними незалежними від платформи реєстрами для забезпечення функціонування застосунків. Рівень платформи надає послуги за запитом з прикладного рівня і вимагає налаштування та управління, викликаного процесами прикладного рівня.

Об'єкт управління на *прикладному рівні* представлений елементом з набору застосунків і служб I^A. Основна функція прикладного рівня – надання користувачеві застосунку. Користувачеві доступні лише інструменти конфігурування на рівні застосунку в якості впливів U^{A*} для керування набором програм. Інші функції управління, такі як оновлення та конфігурування безпеки, виконуються провайдером послуг за допомогою впливу управління U^A.

Таким чином, з метою застосування схем реалізації стратегій управління необхідна генерація керуючих впливів на кожному рівні за критерієм мінімізації витрат на підтримання ІТ-інфраструктури і мінімізації відхилень параметрів якості послуг від заданих. Задача керування якістю послуг, що надаються ІТ-інфраструктурою хмарного ЦОД, представлена як проблема прийняття рішень.

Змінні моделі багаторівневої ієрархічної СУІ: A_i^h – i-й сервіс, наданий на рівні h, $h \in H$; H – набір типів рівнів, таких як інфраструктура (I), платформа (P) і застосунок (A); $R_i^{h,k}$ $i = \overline{1, M^h}$, $k \in K$ – споживання ресурсів i-м сервісом на h-му рівні; M^h – це кількість сервісів на h-му рівні ЦОД; K – набір типів ресурсів; U_i^h – набір управлінських впливів, які керують i-ю послугою на h-му рівні ЦОД; Q_i^h – якість надання сервісу i на рівні h; Q_i^{h*} – задана якість надання сервісу i на рівні h; Z_i^h – запити i-го сервісу на h-му рівні моделі; $q_{i,1}^h \in Q_i^h$, $l = \overline{1, L_i^h}$ – l-й індикатор якості i-го сервісу на h-му рівні ЦОД; L_i^h – кількість якісних показників i-го сервісу на h-му рівні ЦОД; V^h – сигнали зворотного зв'язку, отримані з h-го рівня в якості даних моніторингу; E^h – збурюючі впливи, які впливають на h-й рівень ЦОД.

Якість надання i-го сервісу на рівні інфраструктури визначається як $Q_i^I = f(U_i^I, U_i^{I*}, R_i^{I,k})$. Якість надання i-го сервісу на рівні платформи визначається як $Q_i^P = f(U_i^I, U_i^{I*}, U_i^P, U_i^{P*}, R_i^{I,k}, R_i^{P,k})$. Якість надання i-го сервісу на рівні застосунків визначається як $Q_i^A = f(U_i^I, U_i^{I*}, U_i^P, U_i^{P*}, U_i^A, U_i^{A*}, R_i^{I,k}, R_i^{P,k}, R_i^{A,k})$. Ресурси $R_i^{P,k}$ і $R_i^{A,k}$ також виділяються на рівні інфраструктури і використовуються для підтримки сервісів і застосунків на рівні платформи і застосунків відповідно.

Метою впливів управління U_i^h є виділення оптимальної кількості ресурсів $R_i^{h,k}$ для роботи сервісу A_i^h на рівні h. Управляючі впливи, які контролюють сервіс i на рівні інфраструктури, визначаються як система $< U_i^I, U_i^{I*} >$. Управляючі впливи, які контролюють сервіс i на рівні платформи, визначаються як система $< U_i^I, U_i^{I*}, U_i^P, U_i^{P*} >$. Управляючі впливи, які контролюють послугу i на рівні застосунків, визначаються як система $< U_i^I, U_i^{I*}, U_i^P, U_i^{P*}, U_i^A, U_i^{A*} >$. Набір U сигналів управління, що впливають на всі три рівні на кожному кроці управління, представлений у вигляді декартового добутку шести множин $U = U_i^I \times U_i^{I*} \times U_i^P \times U_i^{P*} \times U_i^A \times U_i^{A*}$.

СУІ ЦОД отримує сигнали зворотного зв'язку V^I, V^P, V^A у вигляді даних моніторингу, які функціонально залежать від управляючих впливів U, входів Z і

збурюючих впливів *E*. СУІ ЦОД управляє сервісами на кожному рівні *h* таким чином, що різниця між фактичними значеннями $q_{i,k}^{h} \in Q_i^h$ *i*-го показника якості надання послуг і заданими значеннями $q_{i,k}^{h*} \in Q_i^{h*}$ є мінімальною: $\sum_{i=1}^{M^h} \sum_{k=1}^{L_i^h} \left(q_{i,k}^{h} - q_{i,k}^{h*} \right)^2 \to \min,\ \forall h \in H.$

Ефективність управління сервісами оцінюється за якістю наданих послуг, за витратами на використання ресурсів ЦОД та за витратами (штрафами за порушення SLA). Вартість штрафу за порушення SLA P^{SLA} визначається як $P^{SLA} = p_1 P^Q + p_2 P^d$, де P^Q – штраф за відхилення параметрів якості обслуговування від визначених, P^d – штраф за затримку розгортання ресурсу, $p_1 > 0$ і $p_2 > 0$ вагові константи.

В дисертаційній роботі пропонується визначити експлуатаційні витрати на використання ресурсів ЦОД як сумарні експлуатаційні витрати у термінах енергоспоживання ЦОД за формулою $P^E = e(E^I + E^U)$, де E^I – загальне споживання електроенергії всім працюючим обладнанням на рівні ІТ-інфраструктури, а E^U – загальне споживання електроенергії, викликане управляючими впливами, *e* – ціна електроенергії. E^I визначено як $E^I = E_{PM} + E_N + E_S + E_C$, де E_{PM} – споживання електроенергії ФС, E_N – споживання електроенергії всім мережевим обладнанням, E_S – споживання електроенергії всім обладнанням СЗД, а E_C – споживання електроенергії всім обладнанням для охолодження. E^U визначено як $E^U = E_{mVM} + E_M$, де E_{mVM} – споживання електроенергії на міграцію ВМ, E_M – споживання електроенергії для застосування всіх інших впливів управління. Таким чином, наступна функція витрат повинна бути мінімізована:

$$J = \alpha \left| P^{SLA} \right|^2 + \beta \left| P^E \right|^2, \tag{4}$$

де α і β – вагові коефіцієнти, які визначаються провайдером ЦОД.

Тому в дисертації вирішується задача розробки схем реалізації стратегій управління і відповідних методів, що базуються на моделі динаміки хмарного ЦОД. Сформульована задача є нелінійною задачею цілочисельного програмування (4) і розв'язується в дисертації на рівні інфраструктури методом Монте-Карло оскільки велика кількість вхідних/вихідних змінних і обмежень не дозволяє знайти оптимальне рішення за допомогою класичних методів оптимізації.

Модель динаміки хмарного ЦОД. Змінні моделі: $z_i(t) \in \{0,1\}$ – стан *i*-го ФС в момент *t*; $r_{ij}^k(t) \in [0,1]$ – використання *k*-го ресурсу *j*-ю ВМ на *i*-му ФС, $i = \overline{1,M}$, $j = \overline{1,N}$, $k \in K$, де *K* – це набір типів ресурсів, *M* – кількість ФС, *N* – кількість ВМ. ВМ перестає існувати, якщо для кожного *k* $r_{ij}^k(t) = 0$. Значення $r_{ij}^k(t)$ нормалізується до найбільшої *k*-ї ресурсної ємності ФС. Використання *k*-го ресурсу на *i*-му ФС – $R_i^k(t) \in [0,1]$ $i = \overline{1,M}$, $k \in K$, $a_{ij}^h(t) \in \{0,1\}$ – цілочисельна змінна, яка вказує, чи працює *j*-а ВМ типу *h* на *i*-му ФС ($a_{ij}^h(t) = 1$) або не використовується ($a_{ij}^h(t) = 0$), $h = \overline{1,H}$, *H* – число типів ВМ; $g = \overline{1,G}$ – тип ФС, *G* – кількість типів ФС; $h_{ij}(t)$ – тип *j*-ї ВМ, що працює на *i*-му ФС; $c_j^k(t) \in [0,1]$ – необхідний об'єм ресурсу *k* для *j*-ї ВМ, (нормалізований до найбільшої *k*-ї ресурсної спроможності ФС). Використання ресурсу *k* *j*-ю ВМ не повинно перевищувати замовлений об'єм, $r_j^k(t) \le c_j^k(t)$, крім випадків, коли змінено тип ВМ. Об'єм наявного ресурсу *k* *i*-го ФС (нормалізований до найбільшого об'єму ресурсу *k* на найбільш

потужному ФС) – $C_i^k \in [0,1]$; $u_{ij}(t) \in \{0,1\}$ – змінна, яка вказує на міграцію *j*-ї ВМ з *i*-го ФС. Міграція відбувається тоді, коли $u_{ij}(t) = 1$. Індекс ФС, до якого *j*-та ВМ мігрує з *i*-го ФС – $x_{ij}(t) \in \{0,...,M\}$, якщо *j*-та ВМ не мігрує, тоді $x_{ij}(t) = i$. Індекс ВМ, яка буде мігрувати з *i*-го ФС до ФС з індексом $x_{ij}(t)$ – $y_i(t) \in \{0,...,N\}$, якщо *j*-та ВМ не мігрує, тоді $y_i(t) = j$. Сигнал управління зміною типу *j*-ї ВМ – $v_j(t) \in \{0,1\}$, якщо $v_j(t) = 1$, то потрібно змінити тип *j*-ї ВМ. Зміна стану *i*-го ФС – $S_i(t) \in \{-1,0,1\}$. При $S_i(t) = -1$ *i*-й ФС повинен бути вимкнений, при $S_i(t) = 1$ *i*-й ФС повинен бути включений, при $S_i(t) = 0$ стан *i*-го ФС не змінюється. Зміна стану *j*-ї ВМ – $s_{ij}(t) \in \{-1,0,1\}$, При $s_{ij}(t) = -1$ *j*-та ВМ повинна бути вимкнена, при $s_{ij}(t) = 1$ *j*-та ВМ повинна бути розгорнута на *i*-му ФС, при $s_{ij}(t) = 0$ стан *j*-ї ВМ не змінюється.

Наступний стан *i*-го ФС виражений у такий спосіб:

$$z_i(t+1) = z_i(t) + S_i(t). \tag{5}$$

Динаміка роботи *j*-ї ВМ на *i*-му ФС виражена у такий спосіб:

$$a_{ij}(t+1) = (a_{ij}(t) + s_{ij}(t))(1-u_{ij}(t)) + a_{x_{ij}(t)y_{ij}(t)}(t)u_{ij}(t), u_{ij}(t) = 0 \mid a_{ij}(t) = 0. \tag{6}$$

Динаміка використання *k*-го ресурсу *j*-ю ВМ на *i*-му ФС:

$$r_{ij}^k(t+1) = (a_{ij}(t) + s_{ij}(t))(r_{new}^k(t)s_{ij}(t) + r_{ij}^k(t))(1-u_{ij}(t)) + r_{x_{ij}(t)y_{ij}(t)}^k(t)u_{ij}(t). \tag{7}$$

Зміна типу *j*-ї ВМ виражена у такий спосіб:

$$h_j(t+1) = \Big((a_{ij}(t) + s_{ij}(t))\Big(h_{ij}(t)\big(1-v_j(t)\big) + h_{new}v_j(t)\Big)\Big)\big(1-u_{ij}(t)\big) + h_{x_{ij}(t)y_{ij}(t)}(t)u_{ij}(t). \tag{8}$$

Умова обмеження використання ресурсів *i*-го ФС для кожного ресурсу:

$$\sum_{j=1}^{N} r_{ij}^k(t) \le C_i^k. \tag{9}$$

Нехай $W_{PM}(t)$ – невідомий процес, що повільно змінюється і впливає на кількість ФС, що обслуговують робоче навантаження. Число ФС, що працюють в момент *t*, – $M_{PM}(t) = \sum_{i=1}^{M} z_i(t)$, $M_{PM}(t) \in \mathbb{N}$, $M_{PM}(t) < M$. Число ВМ, що працюють в момент *t*, – $N_{VM}(t) = \sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}(t)z_i(t)$, $N_{VM}(t)$, $N_{VM}(t) \in \mathbb{N}$, $N_{VM}(t) < N$. Число ВМ, які мігрують в момент *t*, – $N_{mig}(t) = \sum_{i=1}^{M}\sum_{j=1}^{N} u_{ij}(t)z_i(t)$, $N_{mig}(t)$, $N_{mig}(t) \in \mathbb{N}$, $N_{mig}(t) < N$. Число ВМ, що працюють в момент *t* на *i*-му ФС, – $N_i(t) = \sum_{j=1}^{N} a_{ij}(t)$.

Щоб оцінити кількість ВМ N' кожного типу $h \in H$, які теоретично можуть бути створені в ЦОД, пропонується використовувати функцію $N' = \max_{h \in H}\left\{\max_{k \in K}\left\{\sum_{i=1}^{M} C_i^k \Big/ c^k\right\}\right\}$.

Для управління кількістю активних ФС, для вибору нового місця розташування ВМ та для керування міграцією ВМ враховуються такі вимоги: 1) мінімізувати кількість порушень SLA, що виникають у зв'язку з перевантаженням як мінімум одного з ресурсів ФС; 2) мінімізувати кількість порушень SLA, що виникають у зв'язку із затримкою розгортання нової ВМ; 3) мінімізувати споживання енергії в ЦОД з урахуванням гетерогенного та нестаціонарного стохастичного середовища IaaS.

Інтегрована модель включає в себе модель електроспоживання і модель порушень SLA. В цій моделі, а також в моделях розділу 5, порушення SLA враховуються і

обчислюються тільки для сервісної моделі IaaS. Нехай $e_{idle}^{g} \in \mathbb{R}^{+}$ – споживання електроенергії ФС типу g коли він працює в режимі очікування; ρ^{gk} – коефіцієнт енергоефективності ФС типу g для ресурсу k. $e_i^g(t) \in \mathbb{R}^{+}$ – споживання електроенергії i-го ФС типу g; таким чином: $e_i^g(t) = e_{idle}^g + \sum_{k \in R} R_i^k \rho^{gk}$. Сумарне споживання електроенергії усіх працюючих ФС – $E_{PM}(t) = \sum_{i=1}^{M} z_i(t) e_i^g(t)$, де $z_i(t) \in \{0,1\}$ – стан i-го ФС.

Нехай $e_j^h(t)$ – споживання електроенергії при міграції ВМ типу h, тоді загальне електроспоживання при міграціях ВМ в момент t визначається як $E_{mVM}(t) = \sum_{i=1}^{M} \sum_{j=1}^{N} u_{ij}(t) e_j^h(t)$. Таким чином, споживання електроенергії ЦОД, $E(t)$, в момент t, визначається як:

$$E(t) = E_{PM}(t) + E_{mVM}(t). \tag{10}$$

Штраф за порушення SLA, $P^{SLA}(t)$, є сумою штрафів за перевантаження ФС, $P^o(t)$, та штрафів за затримки розгортання нової ВМ, $P^d(t)$:

$$P^{SLA}(t) = p_1 P^o(t) + p_2 P^d(t), \tag{11}$$

де p_1 – узгоджена з користувачем вага одиниці штрафу за порушення SLA на одній ВМ, $p_1 > 0$; p_1 – узгоджена з користувачем вага одиниці штрафу за затримку розгортання ВМ, $p_2 > 0$.

ФС i перевантажений, якщо $\exists k \in K : \sum_{j=1}^{N} r_{ij}^k(t) + R_i^k(t) = C_i^k$. Значення штрафу $P^o(t)$ за перевантажений стан ФС для ресурсу $k \in K$ – $P^o(t) = N_{VM}^v(t)$, $N_{VM}^v(t)$ – кількість ВМ, на яких впливає перевантаження відповідного ФС, яка визначається таким чином: $N_{VM}^v(t) = \sum_{i=1}^{M} z_i(t) \psi(i)$, якщо ресурс $k \in K$ існує такий, що $\psi(i)$ визначається як

$$\psi(i) = \begin{cases} 0, & C_i^k \Big/ \sum_{j=1}^{N} r_{ij}^k(t) + r_i^k(t) < 1, \\ \sum_{j=1}^{N} a_{ij}(t), & \textit{інакше} \end{cases}.$$

Нехай $P^d(t)$ – штраф за затримку розгортання нової ВМ. Кількість ВМ на наступному кроці $t+1$ управління визначається як: $N_{VM}(t+1) = N_{VM}(t) + N_{VM}^{on}(t) - N_{VM}^{off}(t)$, де $N_{VM}^{on}(t)$ – кількість ВМ, визначених на розгортання, $N_{VM}^{off}(t)$ – кількість ВМ, визначених на вимикання. Значення $N_{VM}^{off}(t)$ коректується кожного разу на наступному кроці управління ресурсами хмарного ЦОД, щоб врахувати аномальне вимкнення ВМ.

Кількість ВМ, запланованих до розгортання, визначається таким чином: $N_{VM}^{on}(t) = \sum_{j=1}^{N} \gamma_1(s) s_{ij}(t)$, де $\gamma_1(s)$ – це функція, визначена так, що $\gamma_1(s) = 1$, якщо $s = 1$, та $\gamma_1(s) = 0$ в іншому випадку. Кількість ВМ, запланованих для завершення роботи, визначається таким чином: $N_{VM}^{off}(t) = \sum_{j=1}^{N} \left| \gamma_2(s) s_{ij}(t) \right|$, де $\gamma_2(s)$ – це функція, визначена так, що $\gamma_2(s) = -1$, якщо $s = -1$, та $\gamma_2(s) = 0$ в іншому випадку.

За кожний додатковий час τ, витрачений на розгортання ВМ, постачальник хмарних послуг сплачує штраф p_2. Нехай $w(t)$ – число ВМ, які не були розгорнуті на попередньому кроці від $t-1$ до t. Таким чином, за визначений період $\Upsilon = \{1,2,...,T\}$ $P^d(t) = \sum_{t=1}^{T} w(t)$.

З метою підтримки ресурсної потужності на необхідному рівні, достатньому для розгортання нових ВМ, пропонується використовувати *дві нові метрики*. Перша з них – миттєвий коефіцієнт життєздатності ВМ V_{VM}^{inst} використовується для прийняття рішень на наступний крок управління ресурсами хмарного ЦОД в момент t (12) при управлінні міграціями ВМ. Перша метрика визначається так:

$$V_{VM}^{inst} = N_{VM}^{on}(t) / N_{VM}^{off}(t). \quad (12)$$

Оскільки миттєве значення не дає інформацію про минулі зміни в кількості ВМ, та одного інтервалу управління не достатньо для включення ФС, то пропонується використовувати другу метрику – середній коефіцієнт життєздатності ВМ V_{VM}^{mid}. Вона характеризує динаміку змін кількості ВМ в хмарному ЦОД і обчислюється одним з варіантів зваженого рухомого середнього. На початку кожного інтервалу управління t, кількість ФС, які будуть потрібні на наступний момент t+1, визначається з використанням моделі динаміки хмарного ЦОД, моделі планування потужності, та методу управління потужністю. Таким чином, запропонована *інтегрована модель* включає в себе всі показники, що використовуються провайдерами хмарних послуг в реальних умовах. Запропоновані моделі враховують динаміку ЦОД, гетерогенність ресурсів, споживання електроенергії і міграції ВМ, що дозволяє застосувати метод ІУР.

Досягнення цілі ІУР ЦОД пов'язане з управлінням кількістю ФС згідно критерія енергозбереження з урахуванням обмежень та дотриманням SLA через розв'язання відповідної оптимізаційної задачі. В дисертації запропонована модель управління кількістю ФС за критерієм мінімізації суми штрафів, пов'язаних з міграцією ВМ та затримкою планування виконання ВМ (11) і вартості спожитої електроенергії (10). Задача ІУР зводиться до мінімізації на кожному кроці управління t функції

$$J = \alpha \left| P^{SLA}(t) \right|^2 + \beta \left| E(t) \right|^2 \quad (13)$$

за умов (5) – (9), де α і β – це ваги, визначені користувачем через експертну оцінку, яка позначає відносну важливість складових критерію.

Задача ІУР ЦОД (13) є задачею нелінійного цілочисельного програмування. Використання класичних методів оптимізації для її розв'язання в режимі онлайн обмежено великою кількістю змінних та обмежень, тому для великомасштабних ЦОД, пропонується вирішувати задачу оптимізації (13) приблизно, за рахунок використання оптимізації Монте-Карло.

Таким чином, в дисертації розроблений метод ІУР, призначений для розв'язання задач консолідації ВМ, розміщення нової ВМ та планування ресурсної місткості в рамках стратегій S_3 та S_4. Метод забезпечує визначення необхідної кількості ФС для оброблення навантаження на основі описаної вище моделі. В реальних умовах високої динаміки ЦОД використання методу дозволяє зменшити витрати на електроспоживання і штрафи, обумовлені порушеннями угоди SLA.

Результати досліджень за розділом опубліковано у працях [7, 3, 4, 23, 24, 37, 27, 44, 46 - 57, 61, 29, 70, 40].

У **четвертому розділі** обґрунтовано необхідність прогнозування навантаження на ресурси IT-інфраструктури хмарного ЦОД, сформульовано задачу прогнозування навантаження на обчислювальні ресурси, проаналізовано існуючі моделі прогнозування, розроблено моделі і методи адаптивного прогнозування навантаження на ресурси хмарного ЦОД із застосуванням набору альтернативних методів і моделей прогнозування, змінного розміру навчальної вибірки та евристик комбінування прогнозних значень.

Обґрунтування необхідності *прогнозування навантаження* на ресурси IT-інфраструктури пов'язане з пошуком компромісу між бажанням вивільнити максимально можливу кількість ФС (для зменшення витрат на енергоспоживання) і залишити їх у гарячому резерві через острах перед зростанням штрафів за порушення SLA. Точний прогноз дозволяє зменшити енергоспоживання і штрафи за рахунок прийняття випереджаючих управлінських рішень щодо розміщення нових ВМ та міграції існуючих в рамках кожної стратегії управління.

Оскільки визначити розподіл вибірки та вибрати параметри і коефіцієнти моделі прогнозування при будь-яких комбінаціях навантажень не уявляється можливим, залишається через певний проміжок часу створювати нову модель, застосовуючи нові дані підсистеми моніторингу. Розмір навчальної вибірки для отримання моделі прогнозу треба теж адаптувати до поточних умов функціонування хмарного ЦОД.

Відсутність строгого критерію застосування того чи іншого методу прогнозування навантаження обумовлює використання методу з адаптацією параметрів моделі і комбінуванням прогнозів. Адаптивний метод комбінованого прогнозування полягає у застосуванні декількох (альтернативних) методів або моделей прогнозування з адаптацією розміру навчальної вибірки і подальшим вибором прогнозу на основі визначеного критерію. Склад альтернативних методів прогнозування визначається емпіричним шляхом за результатами попередніх досліджень їх застосування до навчальних вибірок типового навантаження. Для вибору управління на поточному кроці вибирається прогноз, отриманий методом або моделлю і розміром навчальної вибірки, що дали мінімальну середню абсолютну похибку прогнозу в процентах на попередньому кроці.

Задача *прогнозування навантаження представлена таким чином:* альтернативні конфігурації адаптивного методу прогнозування позначено як $\langle A,W \rangle$, де A – множина альтернативних методів/моделей прогнозування, $|A| = n$, n – кількість елементів множини A, W – множина розмірів навчальних вибірок, $|W| = m$, m – кількість елементів множини W. Навчальна вибірка – це набір значень визначеного показника (прогнозованого параметру), який отриманий з системи моніторингу починаючи на попередньому кроці управління. Нехай $Q_{i,j} = f(a_i, w_j)$, $i = \overline{1,n}$, $j = \overline{1,m}$ – середня абсолютна помилка в процентах (САПП), що отримана при використанні альтернативного методу a_i, $a_i \in A$, з розміром навчальної вибірки w_j, $w_j \in W$, $w_{\min} \le w_j \le w_{\max}$, $K_{t|t-1} = \{a_i, w_j, Q_{i,j}\}$ – список конфігурацій адаптивного методу прогнозування, а $K^p_{t|t-1} = \{a^p_i, w^p_j, Q^p_{i,j}\}$ – альтернатива p для вибору конфігурації адаптивного методу прогнозування, що буде обчислений на поточному кроці за допомогою конфігурації, що отримана на попередньому кроці t-1. Тоді в систему

управління для прийняття рішень на поточному кроці *t* передається прогноз, отриманий за допомогою двійки $\{a_i, w_j\}_{t+1|t}$, яка визначається у такий спосіб:

$$\{a_i, w_j\}_{t+1|t} = \{a_i^p, w_j^p\}_{t|t-1} \Big| p = \arg\min_Q \{K_{t|t-1}\}. \quad (14)$$

Список конфігурацій $K_{t|t-1}$, відсортований в порядку збільшення $Q_{i,j}$, позначимо як $C_{t|t-1}$. Таким чином, найкраща конфігурація для отримання прогнозу на поточному кроці *t* визначається як $C_{t|t-1}^1 = K_{t|t-1}^p \Big| p = \arg\min_Q \{K_{t|t-1}\}$.

Метод усередненого комбінованого прогнозування. Нехай $\hat{y}_{t+1|t}^p$ – прогноз, отриманий за допомогою конфігурації $K_{t|t-1}^p$ (14). Відповідно, прогнози, отримані за допомогою конфігурацій $C_{t|t-1}^1, C_{t|t-1}^2, C_{t|t-1}^3, \ldots, C_{t|t-1}^k$, позначені як $\hat{y}_{t+1|t}^1, \hat{y}_{t+1|t}^2, \hat{y}_{t+1|t}^3, \ldots, \hat{y}_{t+1|t}^k$. Комбінований прогноз, що обчислений з урахуванням *k* найкращих конфігурацій $C_{t|t-1}^i, i = \overline{1,k}$, обчислюється у такий спосіб: $\hat{y}_{t+1|t}^c = 1/k \sum_{i=1}^k \hat{y}_{t+1|t}^i$.

Якщо окремі прогнози незміщені (це повинен забезпечувати метод прогнозування), то комбінований прогноз також буде незміщеним. Помилка комбінованого прогнозу: $e_{t|t}^c = y_{t|t} - \hat{y}_{t|t-1}^c = 1/k \sum_{i=1}^k e_{t|t}^i$, де $y_{t|t}$ – фактичне значення прогнозованої змінної.

Метод зваженого комбінованого прогнозування. Нехай S_i - сума квадратів різниць між дійсним значенням, отриманим з системи моніторингу, і прогнозованим значенням, отриманим *i*-м методом (або моделлю) прогнозування, $S_i = \sum_{r=1}^l \left(y_r - \hat{y}_r^i\right)^2$, де l – кількість попередніх кроків управління, y_r – дійсне значення змінної, $r = \overline{1,l}$, $\hat{y}_r^i$ – прогнозоване значення змінної, отримане *i*-м методом/моделлю прогнозування з найменшою помилкою САПП, із застосуванням розміру навчальної вибірки w_j, $i = \overline{1,n}$, $j = \overline{1,m}$, на *r*-му кроці управління.

Найкращим серед усіх отриманих прогнозів вважається такий, помилка САПП якого найменша. Відповідно, найкраща модель прогнозу і найкращий розмір навчальної вибірки на даному кроці будуть ті, за допомогою яких отриманий найкращий прогноз. Нехай q_i – вага прогнозу, отриманого *i*-м методом/моделлю на кроці управління *t*. Вага q_i визначається у такий спосіб: $q_i = (1/S_i) \Big/ \sum_{k=1}^n (1/S_k)$.

Зважений комбінований прогноз $\hat{y}_{t+1|t}^{cw}$ визначається за допомогою зважених прогнозів, отриманих за допомогою *n* конфігурацій *K*, $\{a_i, w_j\}$, $\hat{y}_{t+1|t}^{cw} = \sum_{i=1}^n q_i \hat{y}_{t+1|t}^i$.

Прогноз $\hat{y}_{t+1|t}^i$ для системи управління отримується при розмірі навчальної вибірки $w_{t+1|t}^i = \sum_{k=1}^l w_{t-k|t}^i \Big/ l$, де *i* – метод/модель прогнозу, $w_{t-k|t}^i$ – розмір навчальної вибірки на попередніх кроках, з якою був отриманий найкращий прогноз. Інші прогнози з використанням інших розмірів навчальної вибірки і альтернативних методів/моделей обчислюються з метою отримання вагових коефіцієнтів на наступних кроках управлыння.

Таким чином, в дисертації розроблений адаптивний метод комбінованого прогнозування навантаження на ресурси хмарного ЦОД із застосуванням набору альтернативних моделей і методів прогнозування, змінного розміру навчальної вибірки та евристик комбінування прогнозних значень для визначення керуючих впливів на основі результатів попередніх кроків управління. Запропонований метод використовується в схемах реалізації усіх стратегій управління ресурсами.

Результати досліджень, наведені в цьому розділі, опубліковано у працях [7, 3, 25, 24, 37, 71].

У **п'ятому розділі** розроблено схеми реалізації стратегій управління з використанням запропонованих в дисертації методів консолідації ВМ і управління ресурсами на основі моделі динаміки хмарного ЦОД. Розроблено *метод рівномірної консолідації ВМ* з використанням ідеї імітації відпалу. З використанням алгоритму променевого пошуку розроблено *двостадійний метод управління ресурсами хмарного ЦОД*. З використанням алгоритму навчання з підкріпленням розроблено модель і *метод динамічної консолідації і розміщення ВМ*. Для управління потужністю хмарного ЦОД розроблено метод, який відрізняється урахуванням динаміки станів ЦОД та їх прогнозів. Для управління розподіленим дворівневим сховищем з реплікацією в IT-інфраструктурі хмарного ЦОД розроблено оригінальні моделі і методи управління, що враховують комбінацію метрик стану сховища і мережі передачі між вузлами.

Метод рівномірної консолідації ВМ з використанням ідеї імітації відпалу реалізує стратегію S_1 і використовує карту розподілу ВМ MAP_{best}, набір ВМ G^{VM}, а також набір ФС G^{PM}, які визначають конфігурацію системи. Модифікований в дисертації алгоритм імітації відпалу дозволяє отримати близьку до оптимальної карту розподілу ВМ. Для цього на кожному кроці на основі індикатора дисбалансу $IB = \sum_{i=1}^{K} IB_i, IB_i = 1/\left(CPU_i \cdot RAM_i \cdot NET_i\right)$, де IB_i – індикатор дисбалансу i-го ФС, CPU_i, RAM_i, NET_i – відповідно завантаження процесора, пам'яті та мережевого інтерфейсу i-го ФС, K – кількість ФС у початковій конфігурації, вибираються перевантажені і недовантажені ФС, для яких із застосуванням спеціальної випадкової схеми генеруються і оцінюються на основі індикатора прийнятності до 100 конфігурацій. Остаточна конфігурація визначається на основі відомих метрик:

– відсоток часу, протягом якого активні ФС використовували 100% ЦП (*SLATAH*)

$$SLATAH = 1/M \sum_{i=1}^{M} \left(T_{S_i} / T_{a_i}\right), \quad (15)$$

– загальне падіння продуктивності ВМ через міграції (*PDM*)

$$PDM = 1/N \sum_{j=1}^{N} \left(C_{d_j} / C_{r_j}\right), \quad (16)$$

– комбіноване порушення SLA (*SLAV*)

$$SLAV = SLATAH \cdot PDM, \quad (17)$$

– інтегрована метрика врахування споживання енергії і рівня порушень SLA (*ESV*)

$$ESV = E \cdot SLAV. \quad (18)$$

У формулах (11) – (14) використовуються такі позначення: M – кількість ФС; T_{S_i} – загальний час, протягом якого i-й ФС завантажений на 100%; T_{a_i} – загальний час роботи i-го ФС; N –кількість ВМ; C_{d_j} – викликана міграціями оцінка зниження продуктивності роботи j-ї ВМ; C_{r_j} – загальна потужність процесора, яку вимагає j-та ВМ; C_{d_j} оцінюється як 10% від використання ЦП в MIPS під час всіх міграцій j-ї ВМ; E – споживання електроенергії.

На основі побудованої карти модуль управління хмарним ЦОД керує процесом міграції. Нова карта будується по завершенню міграцій ВМ, передбачених попередньою. Модифікація методу імітації відпалу полягає у розробці нового алгоритму пошуку станів і їх оцінки.

Розроблений в дисертації *двостадійний метод управління ресурсами хмарного ЦОД* реалізує стратегію S_2 і забезпечує зменшення кількості міграцій ВМ під час циклу управління та збільшення кількості ФС, що переключаються в режим сну. Його перевагою є можливість застосування для кластерів з ФС різної конфігурації, які описуються в такий спосіб: *M* ФС; *N* ВМ, де $N, M \in \mathbb{N}^+$.

Метод базується на модифікованому алгоритмі променевого пошуку. Вхідні дані: *n* – ширина променя; *A* – список ФС, які є претендентами для перемикання в режим сну; *B* – список ФС з вільними ресурсами. Перша стадія методу (формування списку *A*) здійснюється за допомогою підалгоритмів *нижньої границі* та *порогу вільних ресурсів*. Перший з них визначає кількість ФС, які неможливо вимкнути в результаті міграції з них усіх ВМ. Другий підалгоритм формує список *A* із ФС, загальна кількість невикористаних ресурсів яких перевищує об'єм ресурсів одного з ФС кластеру.

Робота другої стадії методу полягає у визначенні ФС зі списку *A*, з яких доцільна міграція ВМ, та пошуку ФС зі списку *B*, куди можлива міграція. Результатом є план міграцій у вигляді матриці *U*, елемент якої $U_{ij} \in \{0,1\}$, визначає міграцію ВМ *j* на ФС *i*. В дисертації розроблений алгоритм реалізації другої стадії методу з використанням променевого пошуку. Конфігурації порівнюються за формулою $J = \sum_{i=1}^{m} u_i + \sum_{i=1}^{n} f_i$, де u_i – кількість використовуваних ресурсів на *i*-му ФС зі списку *B*, f_i – кількість вільних ресурсів на *i*-му ФС списку *A*, *m* – кількість ФС в списку *B*, *n* – кількість ФС в списку *A*. Модифікація методу променевого пошуку полягає у розробці нового алгоритму підготовки вхідних даних на першій стадії методу і нового алгоритму оцінювання міграцій ВМ.

Розроблений в дисертації *метод динамічної консолідації і розміщення ВМ хмарного ЦОД* з використанням навчання з підкріпленням (НП) реалізує стратегії S_5 та S_6 і забезпечує зменшення кількості порушень SLA, кількості увімкнутих ФС та кількості міграцій ВМ через навчання на результатах попередніх дій в певних станах системи з урахуванням прогнозованого навантаження. Метод може бути застосований для гетерогенних конфігурацій ФС.

На кожній ітерації алгоритму НП, агент отримує дані про поточний стан середовища $s_t \in S$ (поточне навантаження на ресурси всіх ФС як суму навантажень всіх ВМ по кожному ресурсу на кожному ФС) та вибирає дію $a_t \in A$ – перемикання в/з режиму сну ФС, що впливає на середовище. Після виконання дії середовище переходить до наступного стану $s_{t+1} \in S$, і агент отримує штраф p_t. На початку наступної ітерації *t*+1 агент спостерігає поточний стан системи $s_{t+1} \in S$, оновлює *Q*-значення кроку *t* на основі наступного рівняння: $Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha[p_t + \gamma \min_{a \in A} Q(s_{t+1}, a_{t+1})]$, де $Q(s_t, a_t)$ – очікуваний довгостроковий штраф за виконання дії $a_t \in A$ в стані $s_t \in S$; $\alpha \in [0,1]$ – швидкість навчання (learning rate); $\gamma \in [0,1]$ – коефіцієнт знецінювання (discount factor); $\min_{a \in A} Q(s_{t+1}, a_{t+1})$ – оцінка майбутнього *Q*-значення після виконання керуючого впливу $a_{t+1} \in A$ на *t*+1 кроці. Коли агент знову отримає стан $s_t \in S$, він

вибере дію з мінімальним *Q*-значенням. Політика π вибору найкращої дії у стані $s_t \in S$ визначається у такий спосіб: $\pi(s_t) = \arg\min_a(Q(s_t,a))$. Таким чином, метою агента НП є знаходження оптимальної політики відображення $S \to A$.

Простір станів *S* представлений поточним навантаженням на процесор, оперативну пам'ять, мережевий адаптер та підсистему зберігання даних усіх ВМ на кожному *i*-му ФС, $s_{ti} = \{CPU_{PMi}, RAM_{PMi}, NET_{PMi}, STIO_{PMi}\}$, $i = \overline{1,M}$. Показник використання кожного ресурсу розбитий на чотири інтервали $\{[0,0.25),[0.25,0.5),[0.5,0.75),[0.75,1]\}$. Простір дій визначається як набір $A_t = \{a_{t1}, a_{t2}, ..., a_{ti}, a_{tm}\}$, де $a_{ti} \in \{-1,0,1\}$, $i = \overline{1,M}$. Кожна дія з *A* переводить ФС у сплячий або активний режим роботи до наступного інтервалу часу $t+1$. Значення штрафу p_t в момент *t* складається зі штрафу за порушення SLA p_t^{SLA} та штрафу за надмірне споживання електроенергії p_t^{power}: $p_t = \beta p_t^{SLA} + \delta p_t^{power}$, де β і δ – ваги, що визначають відносну важливість складових рівняння. Кожен показник використання ресурсу нормалізується відносно максимального об'єму відповідного ресурсу ФС.

Штраф за порушення SLA призначається, якщо $res_{ija}^k(t) > res_{ije}^k(t)$, де $res_{ija}^k(t)$ – фактичне використання, а $res_{ije}^k(t)$ – очікуване використання *k*-го ресурсу *j*-ю ВМ на *i*-му ФС. Він розраховується у такий спосіб: $p_t^{SLA} = \begin{cases} \sum_{i=1}^m T_t^{SLA} / T_{t-1}^{SLA}, & T_{t-1}^{SLA} > 0 \\ 0, & T_{t-1}^{SLA} = 0 \end{cases}$, де T_t^{SLA} – час порушень SLA після виконання дії $a_t \in A$; T_{t-1}^{SLA} – час порушень SLA після завершення виконання дії $a_{t-1} \in A$. Якщо $P_t^{SLA} < 1$ то це означає, що агент вибрав правильну дію, щоб мінімізувати кількість порушень SLA. *Штраф за збільшення споживання електроенергії* розраховується у такий спосіб $p_t^{power} = \sum_{i=1}^m P_{i,t} / P_{i,t-1}$, де $P_{i,t}$ – значення споживання електроенергії на поточному кроці; $P_{i,t-1}$ – значення споживання електроенергії на попередньому кроці управління.

В дисертації розроблений алгоритм реалізації методу динамічної консолідації і розміщення ВМ з використанням НП в реальному режимі часу результатом роботи якого є список ВМ L^{VM}, які потрібно розмістити в ЦОД. Модифікація методу НП полягає у розробці нової моделі обчислення штрафу і нової схеми оцінювання стану, що дозволило зменшити розмір задачі.

Розроблений в дисертації *метод управління потужністю хмарного ЦОД* реалізує стратегії S_4 та S_6 і дозволяє обчислити кількість ФС для обслуговування прогнозованого навантаження у вигляді ВМ. Запропонований метод проілюстровано на прикладі застосування тільки до одного ресурсу *k*. Нехай D^k – сума вимог до *k*-го ресурсу ЦОД; F^k – вільна ємність *k*-го ресурсу ЦОД; $\Delta^k = D^k - F^k$ – дефіцит ресурсу; $C^{g,k}$ – ємність ресурсу *k* на ФС типу *g*; $M_{PM}^{+}(t)$ – кількість ФС типу *g*, які повинні бути увімкнені; $M_{PM}^{g,off}(t)$ – кількість ФС типу *g*, які знаходяться в режимі сну.

Якщо $\Delta^k \le 0$ і є наявний ресурс для розміщення максимального запитуваного ресурсу $c_{\max,d}^k$ $(c_{\max,d}^k < C_{\text{free}}^k)$, то немає необхідності вмикати додатковий ФС. Якщо $\Delta^k > 0$, то деяка кількість додаткових ФС повинна бути увімкнена. Враховано два

випадки. Перший випадок: якщо $0 \le \Delta^k \le 1$, то тільки один ФС з ємністю ресурсу k, що є найближчою до Δ^k (для таких g, коли $C^{g,k}/\Delta^k \ge 1$), повинен бути увімкнений. Другий випадок: якщо $\Delta^k > 1$, то кількість додаткових ФС обчислюється так: $M^+_{PM}(t) = \sum_{i=1}^{M}\left(\left\lceil \frac{M^{g,off}_{PM}(t)}{2}\right\rceil \prod_{j=1}^{i} f^g(t)\right)$, $f^g(t) = \begin{cases} 1, & \left\lfloor \Delta^k_j / C^{g,k} \right\rfloor - \left\lceil M^{g,off}_{PM}(t)/2 \right\rceil \ge 0 \\ \left\lfloor \Delta^k_j / C^{g,k} \right\rfloor + 1, & \text{інакше} \end{cases}$,

$\Delta^k_{j+1} = \Delta^k_j - C^{g,k} M^{g,off}_{PM}(t)$.

Таким чином, отримана кількість ФС для увімкнення $M^+_{PM}(t)$ складається з $\left\lceil M^{g,off}_{PM}(t)/2 \right\rceil$, $g = \overline{1,G}$, ФС кожного типу g. Запропонований метод пропонується використовувати для оцінки нижньої границі кількості ФС, які повинні знаходитися в робочому режимі для розміщення нових ВМ з урахуванням кожного ресурсу k, який враховується при розміщенні ВМ і є критичним для її роботи.

Розроблений в дисертації *метод управління реплікацією та міжрівневою міграцією даних у сховищі хмарного ЦОД* використовується при всіх стратегіях управління і дозволяє підвищити рівень надійності збереження даних, відмовостійкості та продуктивності їх оброблення в сучасних апаратно-програмних комплексах ЦОД за рахунок розбиття сховища на швидкий і повільний рівні, а також за рахунок врахування кількості транзакцій доступу до блоків даних, об'єму вільного місця та затримки передачі даних між вузлами зберігання даних для управління реплікацією.

Нехай M – кількість ФС в ЦОД; n_m – кількість файлів, які зберігаються ВМ в локальному сховищі поточного ФС; t_m – кількість блоків, в яких зберігаються файли m-го ФС; u_m – кількість пристроїв швидкого рівня на m-му ФС; v_m – кількість пристроїв повільного рівня; f_{im} – розмір i-го файлу на ФС m, $i = \overline{1,n_m}$, $m = \overline{1,M}$; c – розмір блоку даних; q_{jm} – кількість разів доступу до j-го блоку ФС m на швидкому рівні протягом заданого часу, $j = \overline{1,t_m}$; p_{jm} – кількість разів доступу до j-го блоку на повільному рівні протягом заданого часу; **B** – матриця розбиття файлів на блоки (b_{ij}=1 якщо j-ий блок є частиною i-го файлу, b_{ij}=0 – в іншому випадку); **R** – матриця розміщення блоків (реплікації) на ФС (r_{jm}=1 якщо j-й блок розміщений на ФС m, r_{jm}=0 – в іншому випадку); s_{km} – розмір k-го швидкого пристрою на ФС m, $k = \overline{1,u_m}$; **X** – матриця розміщення блоків на швидких пристроях, x_{kj} – логічна змінна розміщення j-го блоку на k-ому швидкому пристрої; h_{lm} – розмір l-го повільного пристрою на ФС m, $l = \overline{1,v_m}$; **Y** – матриця розміщення блоків на повільних пристроях, y_{lj} – логічна змінна розміщення j-го блоку на l-му повільному пристрої, $l = \overline{1,v_m}$; **Z** – вектор значень затримки передачі даних з поточного ФС m на інші ФС ЦОД, що зберігають копії блоків (репліки) ФС m, z_m=0; **E** – вектор показників роботи кожного вузла зберігання даних, що зберігає значення вільного місця на повільному рівні для кожного ФС ЦОД; **W** – вектор показників роботи кожного вузла зберігання даних, що зберігає значення кількості транзакцій доступу до блоків даних на обох рівнях для кожного ФС ЦОД; C_k – вартість використання k-го швидкого пристрою; G_k – вартість зберігання одного блоку на k-му швидкому пристрої; D_l – вартість використання l-го повільного пристрою; g_l – вартість зберігання блоку на l-му повільному пристрої; d_{vk} – вартість міграції блоку з l-ого повільного пристрою на k-ий швидкий пристрій.

Задача управління міжрівневою міграцією. Оскільки причиною міграції між рівнями є наявність/відсутність доступу до блоків даних протягом деякого часу, в дисертації вирішується задача максимізації середньої кількості транзакцій роботи з блоками, які розміщуються на пристроях швидкого рівня $\max\sum_{k=1}^{u_m}\left[\sum_{j=1}^{t_m} q_{jm}x_{kj}\Big/\sum_{j=1}^{t_m} x_{kj}\right]$. Кількість блоків на поточному ФС визначено як $t_m=\sum_{i=1}^{n_m} f_i\big/c$, а на всіх вузлах зберігання даних, – як $t=\sum_{i=1}^{m} t_i$ (без урахування реплік). При цьому в моделі враховано такі обмеження: 1) наявність вільного місця на пристроях швидкого рівня $\sum_{k=1}^{u_m}\sum_{j=1}^{t_m} x_{kj}c<\sum_{k=1}^{u_m} s_k$; 2) наявність вільного місця на пристроях повільного рівня $\sum_{l=1}^{v_m}\sum_{j=1}^{t_m} y_{lj}c<\sum_{l=1}^{v_m} h_l$; 3) на пристроях швидкого та повільного рівнів *j*-й блок повинен зберігатись лише в один раз $\sum_{k=1}^{u_m} x_{kj}=1,\ \sum_{l=1}^{v_m} y_{lj}=1$; 4) кожен з блоків *j* належить тільки одному файлові $\sum_{i=1}^{n_m} b_{ij}=1,\ j=\overline{1,t_m}$.

Враховуючи вищезазначені обмеження, критерій мінімізації вартості збереження даних у локальному сховищі ФС представлений у вигляді:

$$\min\left[\sum_{k=1}^{u_m}(C_k+\sum_{j=1}^{t_m}G_k x_{jk})+\sum_{l=1}^{v_m}(D_l+\sum_{j=1}^{t_m}g_l y_{lj})+\sum_{l=1}^{v_m}\sum_{k=1}^{u_m}d_{lk}\right]. \qquad (19)$$

У критерії (15) введені такі позначення: вартість збереження даних на всіх швидких пристроях – $\sum_{k=1}^{u_m} C_k$; вартість збереження даних на всіх повільних пристроях – $\sum_{l=1}^{v_m} D_l$; вартість зберігання блоків даних на *k*-му швидкому пристрої – $\sum_{j=1}^{t_m} G_k x_{kj}$; вартість зберігання блоків даних на *l*-му повільному пристрої – $\sum_{j=1}^{t_m} g_l y_{lj}$; вартість міграції між двома пристроями різних рівнів – $\sum_{l=1}^{v_m}\sum_{k=1}^{u_m} d_{lk}$.

Для вирішення сформульованої задачі нелінійного булевого програмування великого розміру в дисертації запропонований варіант алгоритму на основі методу Монте-Карло. При цьому акцент робиться на балансуванні між міграцією відповідних блоків даних з повільного рівня на швидкий (якщо виникають запити до певних файлів протягом деякого часу) і міграцію файлів на повільний рівень (якщо запити до них зникають). Для реалізації метода в дисертації розроблено також алгоритм сортування файлів за двома критеріями.

Задача управління міжвузловою реплікацією. Елементи векторів **E** і **W** визначено у такий спосіб: $e_m=\sum_{l=1}^{v_m} h_{lm}+\sum_{k=1}^{u_m} s_{km}-\sum_{j=1}^{t_m} cr_{jm}$, $w_m=\sum_{j=1}^{t_m}(p_{jm}+q_{jm})r_{jm}$. При цьому враховуються природні обмеження на: 1) кількість даних, які зберігаються на вузлі, включаючи репліки даних з інших вузлів, не повинні перевищувати загальний розмір накопичувачів даного вузла ($\exists m=\overline{1,M}:\sum_{j=1}^{t} cr_{jm}\le\sum_{l=1}^{v_m} h_{lm}$); 2) кількість копій блоків на вузлах збереження даних повинна задовольняти рівнянню $\exists j=\overline{1,t}:\sum_{m=1}^{M} r_{jm}=RF$, де *RF* – кількість копій даних, що зберігається в розподіленому сховищі.

Тоді, з урахуванням обмежень 1) – 4), бажаний рівномірний розподіл реплік блоків даних по вузлах досягається мінімізацією стандартного відхилення значень вільного

місця на повільному рівні і стандартного відхилення значень кількості транзакцій доступу до блоків даних на кожному ФС хмарного ЦОД:

$$\min\left[\sigma_E + \sigma_W\right]. \tag{20}$$

Запропонований в дисертації метод реплікації даних виконується в три етапи: визначення якісних показників кожного ФС; обрання ФС з найкращими показниками за критерієм (18); реплікація даних на обрані ФС.

На першому етапі для кожного *i*-го вузла, крім поточного, обчислюється коефіцієнт реплікації p^r за формулою:

$$\begin{aligned} p_i^r &= p^e(1/e_i) + p^w w_i + p^z z_i, \\ p^e &+ p^w + p^z = 1, \end{aligned} \tag{21}$$

де p^w – коефіцієнт, що враховує важливість покажчика кількості транзакцій доступу до блоків даних, p^e – коефіцієнт, що враховує важливість покажчика об'єму вільного місця, p^z – коефіцієнт, що враховує важливість покажчика затримки передачі даних.

На другому етапі виконуються стандартні обчислення, а на третьому коригуються значення коефіцієнтів у рівнянні (19) через оцінювання стандартного відхилення на поточному кроці методом найменших квадратів і збільшення коефіцієнту для того параметру, значення стандартного відхилення якого зростає. Таким чином, мінімізація критерію (20) досягається підбором коефіцієнтів рівняння (21).

Результати досліджень, подані у розділі, опубліковано в роботах [18 - 19, 21, 22, 10, 26, 5, 44, 53, 30, 67, 68 - 36, 38 - 70, 41, 42].

У **шостому розділі** розроблено узагальнену модель системи інтернету речей (IoT) і архітектуру системи IoT, що базується на використанні мікрохмари з метою розподілу ресурсів, що забезпечують роботу сервісів в IT-інфраструктурі хмарних ЦОД. Розвинуто декомпозиційно-компенсаційний підхід для управління якістю послуг в системі IoT на основі мікрохмари, розподілено функції управління системою IoT між мікрохмарою і застосунками в хмарі, визначені керуючі і координуючі впливи.

Враховуючи необхідність розподілу ресурсів між ядром системи та периферійними зонами, архітектуру IoT системи доцільно будувати за схемою, що поєднує хмару та мікрохмари. В дисертаційній роботі декомпозиційно-компенсаційний підхід розвинутий з метою вирішення завдань розподілу і перерозподілу ресурсів в системі IoT з архітектурою мікрохмара-хмара (МХ-Х) в такий спосіб. Компанії, що працюють в сфері IoT, прагнуть отримати множину $\mathcal{G} = \{s_i\}$, $i = \overline{1,K}$ необхідних послуг IoT з максимальною якістю $\mathcal{Q}$ і мінімальними витратами $\mathcal{C}$. Управління рівнем послуг в системі IoT з архітектурою МХ-Х пропонується здійснювати інтегрованою взаємодією трьох процесів: узгодження рівня послуг, планування ресурсів і управління рівнем послуг.

На рис. 4 в кожній *n*-й мікрохмарі дані надходять з давачів, обробляються в сервері V_n і запам'ятовуються в сховище X_n. Узагальнена інформація результатів моніторингу в окремій мікрохмарі, отримана після оброблення в сервері V_n, надходить на сервер хмари V_c і в сховище хмари X_c. Сервери хмари, обробляючи інформацію моніторингу від усіх мікрохмар, надають можливість отримати глобальну картину в масштабах міста або регіону про стан контрольованих параметрів. Процес узгодження рівня послуг запускається з ініціативи замовників сервісу IoT і закінчується формуванням або оновленням елементів множини $\mathcal{G}$ і матриці $Q = \|q_{ki}\|$, елемент q_{ki}, $k = \overline{1,L_i}$, $i = \overline{1,K}$, якої відповідає значенню *k*-го показника якості *i*-ї послуги. Для функціонування послуг $\mathcal{G}$

їм виділяється сумарна кількість ресурсів r в системі IoT як у мікрохмарі, так і в самій хмарі. Отримане в результаті ресурсне забезпечення IoT у вигляді системи $<Q,r>\in$ основою для вирішення завдань на рівні нижче.

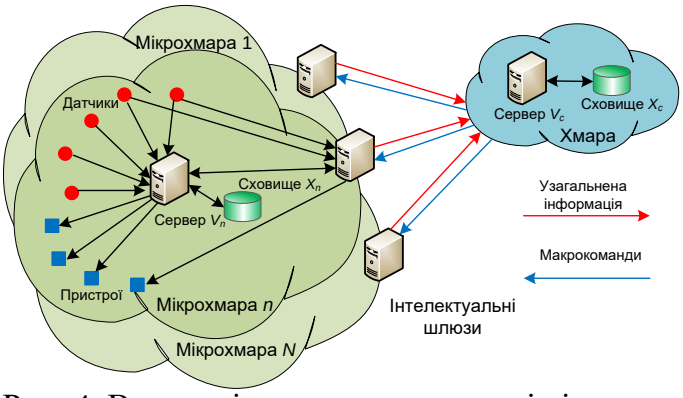

Рис. 4. Взаємодія локальних серверів і систем зберігання даних в системі IoT.

Процес планування полягає у виділенні і закріпленні за кожною послугою s_i, $i=\overline{1,K}$ частини ресурсів мікрохмари та хмари з ресурсів $R_1,...,R_m$, виділених для підтримки послуг. При цьому обсяги або кількість $r_1,...,r_m$, ресурсів $R_1,...,R_m$ відповідно, визначаються у такий спосіб: $r_j=\sum_1^N r_j^{(M_n)}+r_j^{M_c}$, де $r_1^{(M_n)},...,r_m^{(M_n)}$ – кількість ресурсу R_j, $j=\overline{1,m}$, що виділені в мікрохмарі M_n, $r_j^{M_c}$ – кількість ресурсу R_j, $j=\overline{1,m}$, виділених у хмарі. Одиниця j-го ресурсу має вартість c_j. При цьому $c_j^{M_n}$ – вартість одиниці j-го ресурсу в мікрохмарі M_n, $c_j^{M_c}$ – вартість одиниці j-го ресурсу в хмарі. Тоді кількість всіх ресурсів обчислюється як $r=\sum_{j=1}^m r_j$, а вартість c ресурсів визначається у такий спосіб: $c=\sum_{j=1}^m r_j\cdot c_j$. Використання певними послугами призначених ресурсів задається матрицею $\mathrm{P}=\left\|\rho_{ij}\right\|$, де ρ_{ij} дорівнює кількості виділеного послузі s_i ресурсу R_j, $j=\overline{1,m}$, або 0, якщо ресурс не потрібен.

Процес управління рівнем послуг здійснює управління системою IoT так, щоб фактичні значення q_{ki}^*, $k=\overline{1,L_i}$, $i=\overline{1,K}$ показників якості послуг відповідали узгодженим значенням з матриці Q, тобто, щоб виконувалася рівність

$$q_{ki}-q_{ki}^*=0,\quad k=\overline{1,M_i},i=\overline{1,K}. \tag{22}$$

При невиконанні умови (22) визначаються елементи матриці фактичних значень показників якості $Q^*=\left\|q_{ki}^*\right\|$, для яких $q_{ki}^*<q_{ki}$. Система управління намагається вирішити задачу на нижньому рівні, змінюючи значення параметрів функціонування системи IoT або перерозподіляючи ресурси між мікрохмарою та хмарою так, щоб збільшити значення q_{ki}^*. Якщо в результаті відновлювальних заходів вдалося забезпечити виконання рівності (22), то функціонування системи IoT триває з новими налаштуваннями, в іншому випадку здійснюється ескалація проблеми на рівень планування ресурсів.

Процес узгодження рівня послуг з ініціативи процесу планування здійснює перегляд спочатку значення q_{ki}, для якого $q_{ki}^*<q_{ki}$, а потім, можливо, і значень всіх елементів q_{ki}, матриці якості послуг Q у бік зменшення. Якщо вдається сформувати матрицю $Q'=\left\|q'_{ki}\right\|$ з новими значеннями показників якості послуг, то вона передається на рівень нижче, де проводиться вивільнення ресурсів і виділення їх для послуг, для яких виконується умова $q_{ki}^*<q_{ki}$. Якщо процес узгодження рівня послуг не має повноважень для формування матриці $Q'=\left\|q'_{ki}\right\|$, то проводиться ескалація проблеми на

рівень бізнесу, який або згенерує матрицю $Q' = \|q'_{ki}\|$ з новими значеннями, або збільшує загальний обсяг ресурсів $r_1,...,r_m$.

Таким чином, запропонована архітектура IoT на основі MX-X дозволяє ефективно використовувати ресурси для забезпечення ІТ-послуг в екосистемі IoT. В дисертації запропонована ІТУ, що за рахунок розроблення комплексної, інтегрованої СУІ в екосистемі IoT дає можливість розробляти нові послуги IoT, що використовуються в певній галузі.

Результати досліджень за розділом опубліковано у працях [1, 9, 20, 31, 32, 34, 36, 41].

У **сьомому розділі** реалізовано ІТУ ІТ-інфраструктурою хмарного ЦОД, проаналізовано технологічні аспекти та апаратні рішення гіперконвергентної архітектури, обґрунтовано необхідність моделювання ІТ-інфраструктури за допомогою ArchiMate, розроблено підхід і модель ІТ-інфраструктури провайдера хмарних послуг, розроблено архітектуру системи управління і моніторингу продуктивності програмно-визначеної ІТ-інфраструктури хмарного ЦОД.

Архітектура ІТУ базується на методології управління, структурно-функціональній моделі багаторівневої ієрархічної СУІ хмарного ЦОД та існуючих поширених практиках розроблення і надання хмарних послуг, використовуваних на практиці. Архітектура ІТУ (рис. 5) включає в себе основні блоки ПЗ СУІ, які опрацьовують управляючі впливи адміністраторів, працюючих в локальній мережі ЦОД, і запити на управління і обслуговування з боку споживачів (користувачів) хмарних послуг, працюючих в мережі Інтернет. Функції управління хмарними послугами на всіх рівнях сервісної моделі виконуються через портал провайдера після відповідної аутентифікації і авторизації. Провайдер хмарних послуг використовує систему надання хмарних послуг, яка створюється за модульним принципом з метою залучення тільки тих модулів управління, які забезпечують виконання певної множини хмарних послуг користувачеві і сервісних (забезпечуючих) послуг. При необхідності розгортання нових хмарних послуг або нових службових сервісів система надання хмарних послуг дозволяє підключити відповідний модуль і за рахунок використання відповідних API інтегрувати його в СУІ.

Рис. 5. Архітектура ІТУ ІТ-інфраструктурою хмарного ЦОД.

Структурна схема ІТУ ІТ-інфраструктурою хмарного ЦОД (рис. 6) включає в себе модулі управління, які вже впроваджені в систему надання хмарних послуг, і модулі управління, які реалізують інформаційну технологію, запропоновану в дисертаційній роботі. Розроблені в розділах 2-7 методологія, інформаційна технологія, операторний підхід, стратегії управління та схеми їх реалізації, методи і алгоритми реалізуються у вигляді окремих модулів: глобальний менеджер ЦОД і МФС.

Глобальний менеджер ЦОД розгортається і інтегрується на рівні центральної частини системи надання хмарних послуг, а МФС розгортається на кожному ФС, який знаходиться в пулі хмарних ресурсів. Розгортання окремих частин і компонентів ІТУ показано на рис. 7. Для виконання функцій глобальний менеджер ЦОД використовує базу даних, в якій зберігає дані моніторингу, поточних параметрів методів і алгоритмів, значення метрик і інші службові дані.

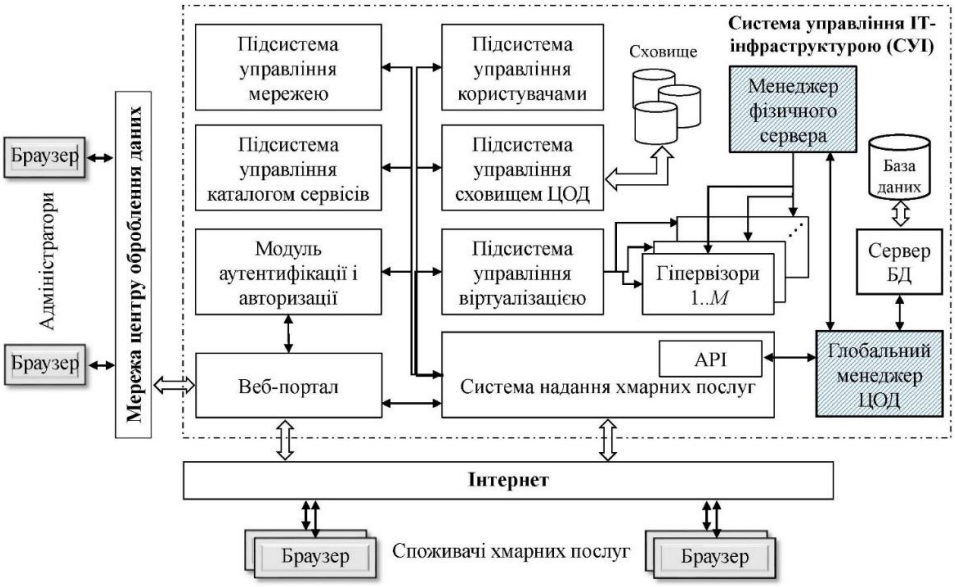

Рис. 6. Структурна схема ІТУ ІТ-інфраструктурою хмарного ЦОД.

SQL Server Express Edition запропоновано використати як СУБД, оскільки ця система має кращий показник ціна/якість, є достатньо потужною для виконання поставлених функцій, нескладною в керуванні і розширюваною. Модулі управління, які реалізують розроблені в дисертації стратегії управління, схеми реалізації і відповідні методи (табл. 1) реалізовані у вигляді докер-контейнерів, що дозволяє додавати інші оптимізаційні методи управління, ізолювати вплив виконання методів один на одного, використати поширену технологію управління віртуалізацією застосунків, заощадити ресурси сервера управління, на якому розгорнутий глобальний менеджер СУІ. Система надання хмарних послуг реалізована на базі ПЗ OpenStack, яке є дуже поширеним серед провайдерів хмарних послуг в світі, зокрема в Україні, Німеччині та Литві. Модулі управління глобального менеджера ЦОД взаємодіють з модулями системи OpenStack через API як самої системи, так і її модулів, що розробляються ІТ-спільнотою OpenStack.

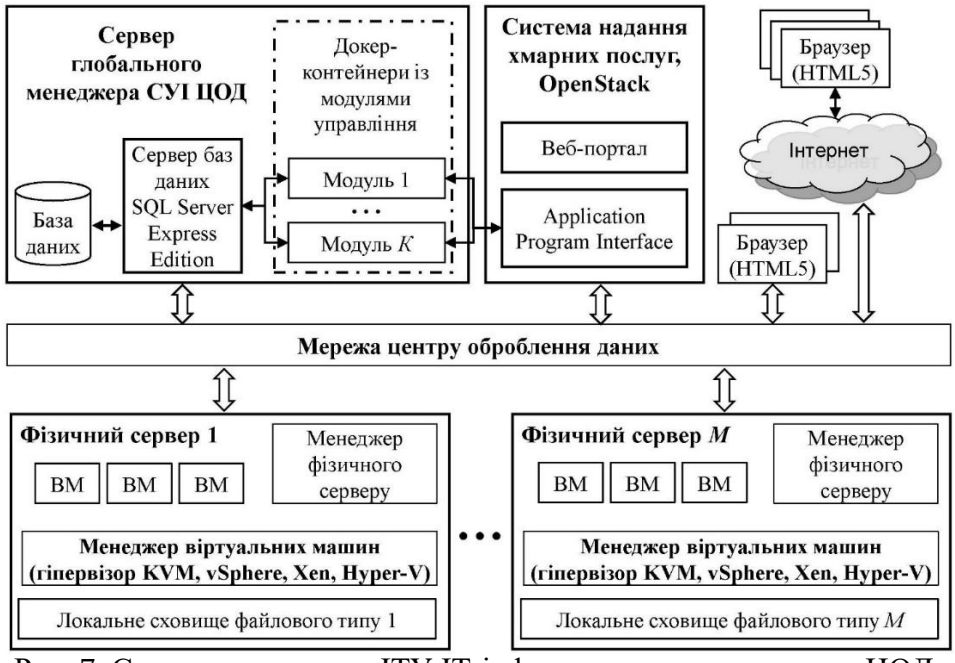

Рис. 7. Схема розгортання ІТУ ІТ-інфраструктурою хмарного ЦОД.

Робота користувачів і адміністраторів з СУІ і хмарними послугами виконується через браузер з підтримкою HTML5. МФС реалізований у вигляді спеціальної ВМ (dom0) і взаємодіє з гіпервізором Xen. Після

певної модифікації МФС може бути пристосований для використання з іншими гіпервізорами. Оскільки в функції МФС входить прогнозування навантаження треба зберігати дані моніторингу, отримані з гіпервізора, в локальному файловому сховищі.

Результати досліджень за розділом опубліковано у працях [1, 4, 14-16, 27, 29-58, 67, 70].

У **восьмому розділі** представлено результати експериментального дослідження запропонованих у дисертації схем реалізації, моделей і методів та розробленої на їх основі ІТУ. Виконаний аналіз використання моделей динаміки та інтегрованого управління ІТ-інфраструктурою хмарного ЦОД, а також аналіз результатів роботи схем реалізації стратегій управління ресурсами ІТ-інфраструктури, а саме: методу рівномірної консолідації ВМ з використанням ідеї імітації відпалу, двостадійного методу управління ресурсами хмарного ЦОД на основі алгоритму променевого пошуку, методу динамічної консолідації і розміщення ВМ на основі алгоритму навчання з підкріпленням. Представлені результати експериментальних досліджень адаптивного методу комбінованого прогнозування з різними евристиками комбінування прогнозів, виконана оцінка продуктивності їх роботи. Представлені результати моделювання управління реплікацією та міжрівневою міграцією даних у сховищі хмарного ЦОД.

Модель динаміки хмарного ЦОД та двостадійний метод на базі алгоритму променевого пошуку реалізовані за допомогою мови C#. Для дослідження методу рівномірної консолідації ВМ з використанням ідеї імітації відпалу та методу динамічної консолідації і розміщення ВМ на основі алгоритму навчання з підкріпленням використовується модульний і розширюваний інструментарій з відкритим кодом CloudSim. Для дослідження запропонованої моделі та методу управління сховищем розроблено програмний застосунок на мові Java. Адаптивний метод прогнозування споживання обчислювальних ресурсів і комбіновані методи прогнозування (усереднений і зважений) з адаптацією параметрів моделі реалізовані мовою R. При аналізі та моделюванні використані дані з журналів GCT та з набору даних Bitbrains. Симуляції і дослідження проводились на комп'ютері з процесором Intel i7-3632QM та 8 Гб оперативної пам'яті під керуванням Windows 10 Pro 64bit.

Метод інтегрованого управління ресурсами ІТ-інфраструктури хмарного ЦОД дозволяє розподіляти ВМ на ФС в середньому на 17% ефективніше, ніж аналогічний Power Aware Best Fit Decrease. Після застосування запропонованої моделі управління потужністю, затримка планування зменшується до 37% у деяких сценаріях. Але даний метод потребує для своєї роботи в середньому 12 хвилин (в залежності від кількості станів системи), що потребує збільшення часу між керуючими впливами.

Метод рівномірної консолідації віртуальних машин з використанням ідеї імітації відпалу з обмеженням міграцій дозволяє зменшити кількість міграцій і, як наслідок, зменшити метрику *PDM* (16) на 16,7% у порівнянні з існуючим в літературі. Також, він дозволяє знизити загальне падіння продуктивності роботи ВМ через міграції, внаслідок чого порушення SLA (метрика (17)) зменшено на 17,2%. Рівномірне завантаження ФС призводить до зменшення порушень SLA через резервування деяких ресурсів ФС для реагування на зростаючі випадкові вимоги до ресурсів у найближчому майбутньому. Об'єднана метрика, яка фіксує як споживання енергії, так і рівень порушень SLA (метрика (18)) покращилась на 13,6%.

Двостадійний метод управління ресурсами хмарного ЦОД на основі алгоритму променевого пошуку дозволяє переключити в режим сну в середньому 56% ФС, що потенційно визначені для переключення в режим сну за допомогою верхньої оцінки необхідної ємності ресурсів при врахуванні допустимої кількості міграцій ВМ. Також

встановлено, що рекомендована ширина променя в алгоритмі променевого пошуку складає від 5 до 8, в залежності від умов роботи методу, обмежень на кількість міграцій ВМ та обмежень на час виконання алгоритму. Також встановлено, що методика з порогом вільних ресурсів показала більш ефективні результати при вирішенні задачі консолідації ВМ, що дає можливість адаптуватися до стану кластера через зміну порогу перед циклом управління, враховуючи інші ресурси, час міграції ВМ та прогноз надходження заявок на створення нових ВМ, що визначений метрикою (12).

Метод динамічної консолідації і розміщення віртуальних машин на основі алгоритму НП, усуває два недоліки політик, запропонованих в літературі: високу затримку при плануванні створення нової ВМ через відсутність доступних ФС; збільшення споживання енергії при перемикання ФС з активного стану в сплячий режим і навпаки та коли кількість ВМ часто змінюється протягом відносно короткого періоду часу. Встановлено (рис. 8), що запропонований метод значно перевершує відомі конкурентні методи за показниками часу порушення SLA і кількості міграцій ВМ, оскільки запропонований метод виконує менше міграцій ВМ і переключає в режим сну менше ФС, ніж у конкурентних методах. Таким чином, цей метод рекомендується використовувати для стратегії управління з високою інтенсивністю змін кількості ВМ у сторону збільшення.

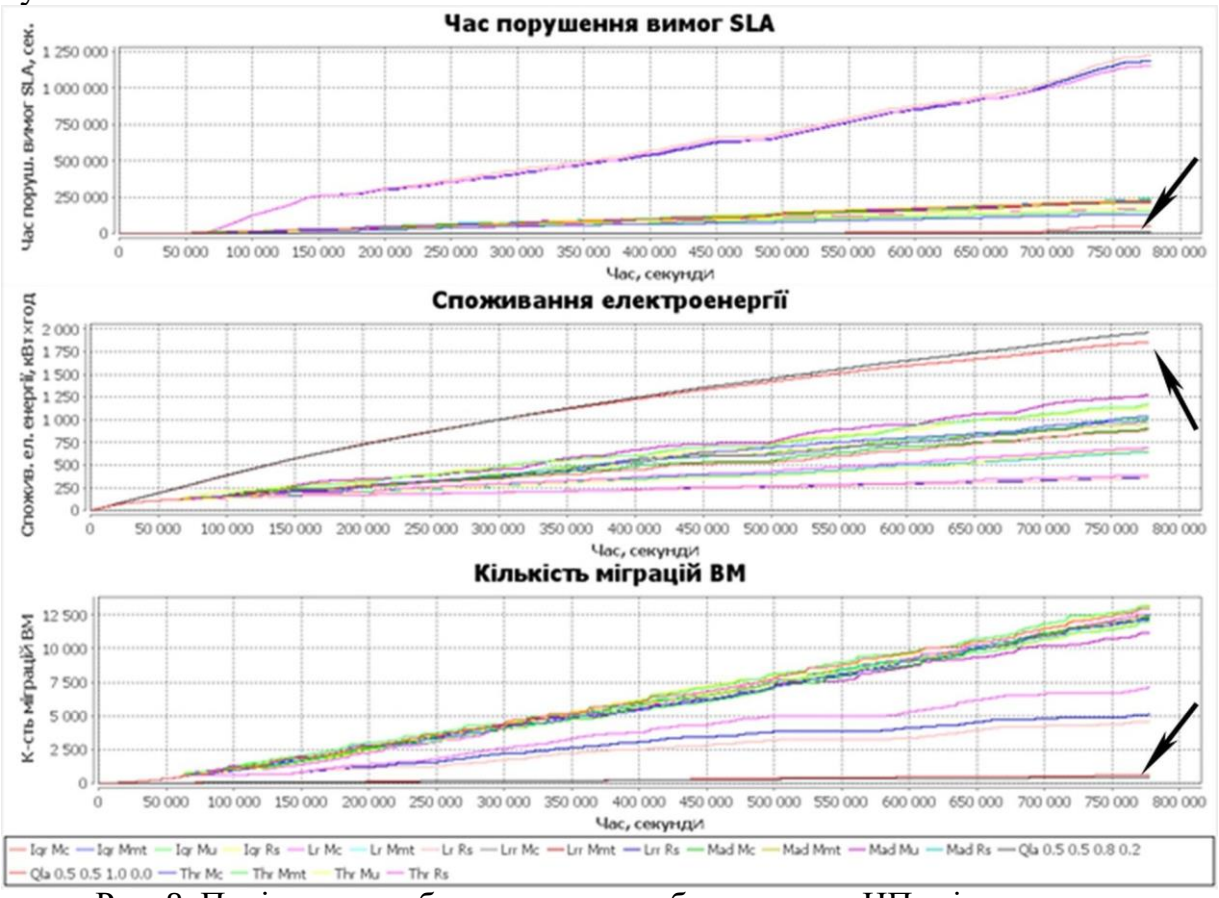

Рис. 8. Порівняння роботи методу, що базується на НП, з існуючими.

Аналіз та моделювання *адаптивного методу комбінованого прогнозування навантаження на обчислювальні ресурси хмарного ЦОД* виконані з використанням статистичних даних Bitbrains на прикладі прогнозування споживання процесорного ресурсу. Якісний аналіз результатів дослідження показав, що точність прогнозу значною мірою залежить від статистичних характеристик часового ряду. Аналіз також показав, що прогнозування сплесків практично неможливе, оскільки вони можуть бути викликані збоями ФС і нестаціонарними моделями використання ресурсів.

Результати кількісного аналізу показали, що точність прогнозу, отриманого за допомогою спрощеного адаптивного методу, збільшується в середньому на 2.4 – 23.6% в залежності від часового ряду. Точність прогнозу, отриманого адаптивним методом зваженого комбінованого прогнозування, збільшується в середньому на 6.6% – 21.2% (в залежності від часового ряду) у порівнянні з точністю спрощеного адаптивного методу прогнозування. Найкращий прогноз споживання процесорного ресурсу отриманий за допомогою адаптивного методу зваженого комбінованого прогнозування з трансформацією, показники роботи якого виділені на рис. 9.

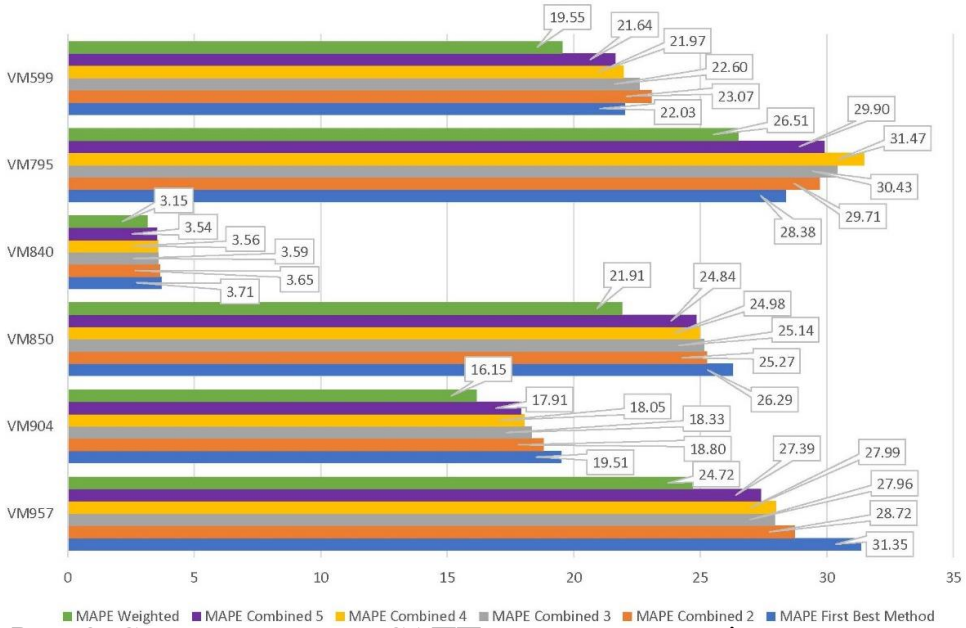

Рис. 9. Середня помилка САПП для часових рядів навантажень.

Результати дослідження *методу управління реплікацією та міжрівневою міграцією даних у сховищі хмарного ЦОД* показують, що використання дворівневих сховищ із запропонованим методом управління призводить до зменшення необхідного об’єму пристроїв швидкого рівня (зниження вартості збереження даних) та зменшення часу очікування завершення транзакцій доступу до файлів при одночасній роботі ВМ зі сховищем ФС. В середньому, при використанні дворівневого сховища запис виконується на 51%, а читання – на 16% швидше. Запропонований метод реплікації дозволяє рівномірно розмістити репліки даних по вузлах ЦОД не створюючи вузьких місць при створенні нових і модифікації існуючих блоків даних. Стандартне відхилення об’ємів даних, що зберігаються на вузлах, зменшилося на 35%.

Результати досліджень, подані у розділі, опубліковано в працях [18 - 19, 21, 22, 10, 26, 5, 30, 68, 35, 38 - 70, 42].

У **додатках** наведено діаграми класів застосунків для моделювання та довідки про впровадження результатів дисертаційної роботи.

ВИСНОВКИ

У дисертаційній роботі вирішено науково-практичну проблему забезпечення ефективного функціонування ІТ-інфраструктури хмарного ЦОД через створення методології управління та на її основі відповідної ІТУ із застосуванням розроблених підходів, моделей, алгоритмів і методів, які базуються на методах інтегрованого і ієрархічного управління, штучного інтелекту, моделях і методах прогнозування навантажень і споживання ресурсів ІТ-інфраструктури, що дозволило забезпечити виконання заданих вимог угоди про рівень обслуговування та зниження операційних і капітальних витрат в умовах невизначеності та змінних навантажень. У процесі вирішення поставлених задач в дисертаційній роботі отримані такі наукові результати:

1. Аналіз існуючих підходів, технологій, моделей і методів управління ІТ-інфраструктурою ЦОД провайдера хмарних послуг виявив необхідність розроблення методології управління ІТ-інфраструктурою та на її основі інформаційної технології, моделей і методів з метою досягнення таких показників ефективності, як: зменшення споживання електроенергії, підвищення якості обслуговування споживачів, зменшення капітальних та операційних витрат, зменшення кількості збоїв та простоїв.

2. На основі операторної форми постановки, аналізу і розв'язання задач управління ІТ-інфраструктурою хмарного ЦОД запропоновано методологію управління для реалізації програмно-визначеного підходу в умовах невизначеності і змінних навантажень через використання множини напрацьованих схем реалізацій з можливістю їх розширення, а також комбінування моделей і методів запропонованого комплексу в залежності від початкових і поточних умов функціонування з урахуванням результатів прогнозування.

3. На основі запропонованої методології з використанням комплексу моделей і методів управління розроблено оригінальну інформаційну технологію, яка забезпечує збір, накопичення, оброблення і використання інформації для ефективного управління ІТ-інфраструктурою хмарного ЦОД в умовах змінних навантажень, яка на відміну від відомих враховує суттєві характеристики хмарних обчислень, гетерогенність ІТ-інфраструктури, нові архітектури організації обчислень, забезпечує адаптацію до змінних навантажень за рахунок прогнозування і може бути застосована для реалізації функцій, підсистем, компонентів та інших складових інформаційної системи управління.

4. Розроблено адаптивний метод комбінованого прогнозування навантаження на обчислювальні ресурси хмарного ЦОД з використанням усередненого, зваженого та оптимізаційного комбінування оцінок прогнозів, обчислених за альтернативними методами прогнозування та з адаптацією розміру навчальної вибірки, який відрізняється обчисленням та використанням певних типів комбінованих прогнозів (усередненого та зваженого) в реальному часі, отриманих альтернативними моделями і методами, що дозволяє збільшити точність прогнозу в середньому на 6.6% – 21.2% в залежності від часового ряду та застосувати його в умовах провайдера хмарних послуг для прогнозування відомих видів змішаних навантажень в ІТ-інфраструктурі.

5. Розроблено метод інтегрованого управління фізичними і віртуальними машинами в ІТ-інфраструктурі хмарного ЦОД на основі динамічної моделі станів із застосуванням стохастичного пошуку для виконання міграцій ВМ, вивільнення, увімкнення і вимкнення ФС та прогнозування для адаптації до змін навантаження з метою досягнення сталого режиму роботи згідно визначених критеріїв. Запропонований метод інтегрованого управління дозволяє розподіляти ВМ на ФС в середньому на 17% ефективніше, ніж за допомогою відомого аналогу, що дозволяє застосувати його при розробленні модулів управління в складі ПЗ з відкритим кодом, що застосовується для побудови хмарних ЦОД.

6. Удосконалено комплекс алгоритмів і методів стохастичного пошуку через розроблення нових методів управління ресурсами ІТ-інфраструктури хмарного ЦОД в складі запропонованої методології, а саме методу рівномірної консолідації ВМ з використанням ідеї імітації відпалу, двостадійного методу управління ресурсами хмарного ЦОД на основі алгоритму променевого пошуку, методу динамічної консолідації і розміщення ВМ на основі алгоритму навчання з підкріпленням, що відрізняються врахуванням методик вибору станів системи на основі нових метрик,

врахуванням достатньої кількості видів ресурсів і дискретності вимірів, а також особливостей функціонування ІТ-інфраструктури хмарного ЦОД в умовах віртуалізації із застосуванням програмно-визначеного управління, що дозволяє застосувати їх при розробленні і модернізації існуючих комплексів управління ресурсами ІТ-інфраструктури хмарного ЦОД.

7. Розроблено архітектуру та структурно-функціональну модель багаторівневої ієрархічної програмно-визначеної СУІ хмарного ЦОД, яка за рахунок поєднання стратегічного і оперативного управління з автоматичним керуванням, а також надання програмно-визначених властивостей дозволяє застосувати її при управлінні ресурсами і навантаженням ІТ-інфраструктури хмарного ЦОД з вибором стратегій, плануванням і управлінням їх реалізацією, автоматичним керуванням з урахуванням структури системи, історичних даних її функціонування та дотриманням заданих показників якості угоди про рівень обслуговування.

8. Декомпозиційно-компенсаційний підхід до управління ІТ-інфраструктурою отримав розвиток через надання адаптивності, багаторівневості та програмно-визначених властивостей, що дозволило реалізувати ефективне управління ресурсами ІТ-інфраструктури хмарного ЦОД з переходом від вибору стратегій до планування і управління їх реалізацією з урахуванням структури системи. Запропонований декомпозиційно-компенсаційний підхід реалізовано для управління інфраструктурою IoT на основі використання мікрохмари через виділення рівнів координації послуг, планування ресурсів та управління рівнем обслуговування в інтегрованій системі управління, що дозволило забезпечити задану якість ІТ-послуг при раціональному використанні ресурсів.

9. Розроблено метод управління реплікацією та міжрівневою міграцією даних сховища хмарного ЦОД у складі інформаційної технології управління, який використовує кешування за розміром файлу і за кількістю транзакцій доступу до файлу для управління міграцією і використовує кількість транзакцій доступу до блоків даних, об’єм вільного місця та затримку передачі даних між вузлами зберігання даних для управління реплікацією, що дозволяє у середньому виконувати запис на 51%, а читання – на 16% швидше і застосувати його в умовах хмарного ЦОД для підвищення рівня надійності збереження даних, відмовостійкості та продуктивності їх оброблення.

10. Для оцінювання стану хмарного ЦОД засобами інформаційної технології управління запропоновано і експериментально обґрунтовано ефективність використання нових метрик, зокрема миттєвого і середнього коефіцієнтів життєздатності віртуальної машини, індикатору дисбалансу фізичного сервера, коефіцієнту відношення необхідних ресурсів до середнього об’єму наявних ресурсів, порогу вільних ресурсів та метрики ємності ЦОД, що відрізняються більш чітким оцінюванням стану і динаміки змін ресурсів хмарного ЦОД і дозволяє ефективніше визначати змінні, обмеження і критерії на множині моделей і методів управління нижнього рівня.

11. Розроблено метод управління потужністю хмарного ЦОД у складі інформаційної технології управління на основі динамічної моделі його станів, який використовує запропоновані метрики, враховує гетерогенність ФС і визначає тип і кількість ФС, що треба увімкнути для обслуговування прогнозованого навантаження, що дає можливість провайдеру завчасно автоматично вмикати необхідну кількість ФС потрібної конфігурації і зменшити затримку розгортання сервісів користувача у хмарі на 18%.

12. Отримані в дисертаційній роботі результати досліджень використані при модернізації систем управління функціонуванням інформаційно-телекомунікаційної інфраструктури в компаніях-членах Асоціації «ТЕЛАС», при розробленні системи управління функціонуванням інформаційно-телекомунікаційної інфраструктури ТОВ «АМ ІНТЕГРАТОР ГРУП» та використані при модернізації системи управління IT-інфраструктурою ТОВ «СІТІУС ПРО». Впровадження результатів досліджень дозволило на 28% скоротити операційні витрати на управління IT-інфраструктурою без порушень SLA; зменшити споживання електроенергії на 17% в середовищі з гомогенними конфігураціями ФС; скоротити в середньому на 18% кількість ФС, що обслуговують навантаження клієнтів; зменшити кількість порушень SLA на 27% при наданні IT-послуг; скоротити витрати на експлуатацію серверного парку на 19% при забезпеченні виконання заданих вимог угоди про рівень обслуговування.

13. Теоретичні і практичні результати дисертаційної роботи склали основу нових спецкурсів, що викладаються автором на кафедрі автоматизованих систем обробки інформації і управління Національного технічного університету України “Київський політехнічний інститут імені Ігоря Сікорського”: “Технології віртуалізації та хмарних обчислень”, “Сучасні технології розроблення програмного забезпечення”, “Інтелектуальні системи управління технічними пристроями”, “Програмування інтернету речей”, “Методи та системи штучного інтелекту”, “Обробка надвеликих масивів інформації”.

СПИСОК ОПУБЛІКОВАНИХ ПРАЦЬ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Монографія, яка індексується Scopus

1. Rolik O., Telenyk S., Zharikov E. IoT and Cloud Computing: The Architecture of Microcloud-Based IoT Infrastructure Management System / In P. Kocovic, R. Behringer, M. Ramachandran, & R. Mihajlovic (Eds.), Emerging Trends and Applications of the Internet of Things. – Hershey, PA: IGI Global. 2017. – С. 198-234. ISBN10: 1522524371

Статті у закордонних виданнях, які індексуються Scopus

2. Telenyk S., Zharikov E., Rolik O. Consolidation of Virtual Machines Using Stochastic Local Search // Advances in Intelligent Systems and Computing. – Springer, Cham, 2017. – Vol. 689. – С. 523-537. (**Scopus**).

3. Telenyk S., Zharikov E., Rolik O. Architecture and conceptual bases of cloud IT infrastructure management // Advances in Intelligent Systems and Computing. – Springer, Cham, 2017. – Vol. 512. – С. 41-62. (**Scopus**).

4. Rolik O., Telenyk S., Zharikov E. Management of services of a hyperconverged infrastructure using the coordinator // Advances in Intelligent Systems and Computing. – Springer, Cham, 2018. – Vol. 754. – С. 456-467. (**Scopus**).

5. Zharikov E., Telenyk S., Rolik O. Method of Distributed Two-Level Storage System Management in a Data Center // Advances in Intelligent Systems and Computing. – Springer, Cham, 2019. – Vol. 938. – С. 301-315. (**Scopus**).

Статті у закордонних виданнях

6. Zharikov E. Topical questions of implementation of information services in a network of university // TEKA Kom. Mot. i Energ. Roln. – 2010, Vol 10B. – С. 331-337. **(Bibliografia Geografii Polskiej, EBSCO, Index Copernicus).**

7. Zharikov E. Analysis of the theoretical and applied aspects of modern IT infrastructure // Teka. Commission of motorization and energetics in agriculture. – 2013, Vol. 13, №4. – С. 297-306. **(Bibliografia Geografii Polskiej, EBSCO, Index Copernicus).**

8. Samozdra M., Zharikov E., Samozdra O. Implementation of automated informational interactions as a part of integrated information-processing system // Teka. Commission of

motorization and energetics in agriculture. – 2014, Vol. 14, №1. – С. 229-237. **(Bibliografia Geografii Polskiej, EBSCO, Index Copernicus).**

9. Rolik O., Telenyk S., Zharikov E. Service quality management in microcloud-based IoT infrastructure //Czasopismo Techniczne. – 2016. – Vol. 2-E(12). – С. 231-244.

10. Telenyk S., Rolik O., Zharikov O., Serdiuk Y. Energy efficient data center resources management using beam search algorithm // Czasopismo Techniczne. – 2018. – Т. 2018. – №. 4. – С. 127-138.

Статті у фахових виданнях

11. Жариков Э. В., Овчинников В. Л. Система передачи мультимедийных данных в локальных сетях // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2008. – №9. Частина 1. – С. 142-146.

12. Зубов Д. А., Ульшин В. А., Жариков Э. В., Григоренко М. С. Концепция программно-аппаратного комплекса долгосрочного прогнозирования средней температуры воздуха на базе статистико-гидродинамических моделей // Збірник наукових праць Укр. держ. геологорозвідувального інституту. 2007. – №4. – С. 223-226.

13. Жаріков Е.В., Солдатенко В.Ю. Методика порівняльного аналізу бездротових технологій // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2011. – № 11. Частина 2. – С. 250-253.

14. Жариков Э.В., Шмитько А.А. Автоматизация подготовки научных изданий к печати // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2006. – №1(95). – С. 237-240.

15. Жариков Э.В. Публикация и использование знаний в глобальной сети // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2003. – №4(62). – С. 36-40.

16. Жариков Э.В. Разработка методов представления и организации знаний в распределенной экспертной системе. // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2005. – №3(85). – С. 89-95.

17. Жариков Э. В. Основные направления оптимизации ИТ-инфраструктуры учебных заведений // Вісник Східноукраїнського національного університету імені Володимира Даля. – 2011. – № 3. – С. 66–71.

18. Теленик С.Ф., Ролик А.И., Жариков Э.В. Управление распределением виртуальных машин в ЦОД // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. – К.: Век+. – 2016. – № 64. – С. 90–99.

19. Жаріков Е.В., Сердюк Е.А. Метод консолидации виртуальных машин на основе лучевого поиска / Е.В.Жаріков // Автомобіль і електроніка. Сучасні технології. – 2017. – No12. – С. 180–186.

20. Ролік О.І., Теленик С.Ф., Жаріков Е.В. Управління рівнем послуг в системі інтернету речей з мікрохмарною архітектурою // Вісник НТУУ «КПІ». Інформатика, управління та обчислювальна техніка: Зб. наук. пр. – К.: Век+. – 2017. – № 65. – С. 110–117.

21. Жаріков Е.В. Керування ресурсами хмарних центрів обробки даних на основі евристичного пошуку // Проблеми програмування. – 2017. – № 4. – С. 16-27.

22. Жаріков Е. В. Динамічне розміщення віртуальних машин на основі навчання з підкріпленням в хмарних центрах обробки даних / Е. В. Жаріков, А. А. Коваль, Р. А. Терентьєв. // Наукові вісті Далівського університету. – 2017. – № 13. – Режим доступу: http://nbuv.gov.ua/UJRN/Nvdu_2017_13_4

23. Жаріков Е. В., Моделювання динаміки хмарного центру обробки даних у просторі станів // Моделювання та інформаційні технології. – 2018. – № 84(6). – С. 125-134

24. Жаріков Е.В. Інтегроване управління ресурсами хмарного центру обробки даних на основі віртуальних машин // Матем. машини і системи. – 2018. – № 2. – С. 21-32.

25. Жаріков Е. В., Структурна оптимізація моделей прогнозу споживання обчислювальних ресурсів в умовах віртуалізації // Електронне моделювання. – 2018. - №5. - С. 49-66.

26. Жаріков Е. В. Метод управління дворівневим сховищем віртуалізованого центру обробки даних // Проблеми програмування. – 2018. – № 4. – С. 3-14.

27. Telenyk S., Nowakowski G., Zharikov E., Vovk Y. Information technology for web-applications design and implementation // Адаптивні системи автоматичного управління, К: Політехніка. – 2019. – Т.1, №34. – С. 138-151. (Google Scholar, WorldCat).

28. Telenyk S., Zharikov E. Operator form to formulate, analyze and solve the cloud data center IT infrastructure management tasks // Адаптивні системи автоматичного управління, К: Політехніка. – 2019. – Т.2, №35. – С. 25-39. (Google Scholar, WorldCat).

Статті в збірниках матеріалів конференцій, які індексуються Scopus та IEEE

29. Telenyk S. An approach to Software Defined Cloud Infrastructure management / S. Telenyk, E. Zharikov, O. Rolik // Proc. of the XI International Scientific and Technical Conference “Computer Science and Information Technologies Congress on Information Technology” (CSIT 2016) 6–10 September, Lviv, Ukraine. – 2016. – С. 21–26.

30. Telenyk S., Zharikov E., Rolik O. An approach to virtual machine placement in cloud data centers // In proc. of the 2016 International Conference Radio Electronics & Info Communications (UkrMiCo) 11–16 September, Kyiv, Ukraine. – 2016. – С. 1–6.

31. Rolik O., Zharikov E., Telenyk S. Microcloud-based architecture of management system for IoT infrastructures // Problems of Infocommunications Science and Technology (PIC S&T), Third International Scientific-Practical Conference. – IEEE, 2016. – С. 149-151.

32. Rolik, O., Telenyk, S., Zharikov, E., & Yasochka, M. Decomposition-compensation approach to microcloud-based IoT infrastructure management // In Proc. of 3rd World Forum on Internet of Things (WF-IoT). – IEEE, 2016. – С. 603-608.

33. Rolik O., Telenyk S., Zharikov E., Samotyy V., Dynamic Virtual Machine Allocation Based on Adaptive Genetic Algorithm // The Eighth International Conference on Cloud Computing, GRIDs, and Virtualization. – 2017. – С. 108-114.

34. Rolik O., Zharikov E., Kolesnik V., Yasochka M. Rule-based Algorithmic Approach for Solving Problems of Impact Analysis in Access Networks // In proc. of the 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) May 29 – June 2, Kyiv, Ukraine. – 2017. – С. 1161–1166.

35. Telenyk S., Zharikov E., and Rolik O. Consolidation of Virtual Machines Using Simulated Annealing Algorithm // in Proc. of the XIIth International Scientific and Technical Conference “Computer Science and Information Technologies” (CSIT 2017) 5–8 September, Lviv, Ukraine. – 2017. – С. 117–121.

36. Telenyk S., Bidyuk P., Zharikov E. and Yasochka M.,Assessment of cloud service provider quality metrics // 2017 International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), Odesa, 2017. – С. 1-5.

37. Telenyk S., Zharikov E., Rolik O. An Integrated Approach to Cloud Data Center Resource Management // Problems of Infocommunications Science and Technology (PIC S&T), 4th International Scientific-Practical Conference. – IEEE, 2017. – С. 211-218.

38. Telenyk S., Rolik O., Zharikov O., Serdiuk Y. Consolidating Virtual Machines with the Use of Beam Search Algorithm // The Fourth International Conference on Automatic Control and Information Technology (ICACIT’17), 2017. – С. 34-46.

39. Rolik O., Zharikov E., Koval A. and Telenyk S. Dynamic management of data center resources using reinforcement learning // 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, 2018. – С. 237-244.

40. Zharikov E., Rolik O., Telenyk S. A Decomposition Approach to Hierarchical Management of Cloud Data Center Services // 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). – IEEE, 2018. – T. 1. – C. 42-47.

41. Rolik O., Zharikov E., Yasochka M., Butenko M. The Method of Impact Analysis for Access Networks with RIP and OSPF Protocols // 2018 International Conference on Information and Telecommunication Technologies and Radio Electronics (UkrMiCo), Odesa, Ukraine, 2018. – C. 1-7.

42. Telenyk S., Zharikov E. and Rolik O. Modeling of the Data Center Resource Management Using Reinforcement Learning // 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T), Kharkiv, Ukraine, 2018. – C. 289-296. doi: 11109/INFOCOMMST.2018.8632064

Статті і тези доповідей в збірниках матеріалів конференцій

43. Жаріков Е. В. Технології віртуалізації на базі VIRTUAL SERVER в навчальному процесі // Збірник наукових праць Східноукраїнського національного університету імені Володимира Даля, Міжнародні Далівські читання XV Нуково-практична конференція "Університет і регіон:проблеми сучасної освіти"/ За заг.ред.проф. Голубенка О.Л.– Луганськ: вид-во Східноукраїнського національного університету імені Володимира Даля, 2009. – С. 247-249.

44. Жариков Э. В. Актуальные вопросы оптимизации IT-инфраструктуры учебных заведений // Комп'ютерні науки для інформаційного суспільства: Матеріали міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 22-23 грудня 2010 р.). – Луганськ: Вид-во «Ноулідж», 2011. – С. 98-102.

45. Солдатенко В. Ю., Жариков Э. В. Методика сравнительного анализа беспроводных технологий для проектирования сетей передачи данных // Комп'ютерні науки для інформаційного суспільства: Матеріали міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 22-23 грудня 2010 р.). – Луганськ: Вид-во «Ноулідж», 2011. – С. 126-128.

46. Кордубайло С. А., Жариков Э. В. Планирование внедрения технологий виртуализации на предприятии // Комп'ютерні науки для інформаційного суспільства: Матеріали II Міжнар. науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 23-24 листопада 2011 р.). – Луганськ: Вид-во «Ноулідж», 2011. – С. 109-112.

47. Альсаясні М., Жаріков Е. В. Підвищення якості роботи інформаційних сервісів корпоративної мережі // Комп'ютерні науки для інформаційного суспільства: Матеріали II Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 23-24 листопада 2011 р.). – Луганськ: Вид-во «Ноулідж», 2011. – С. 82-85.

48. Грищенко Е. В., Жариков Э. В. Уровневое разбиение элементов ИТ-инфраструктуры // Комп'ютерні науки для інформаційного суспільства: Матеріали III Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 12-13 грудня 2012 р.). – Луганськ: Вид-во «Ноулідж», 2012. – С. 107-108.

49. Жариков Э. В., Альсаясни М., Лукутин О. В. Анализ систем мониторинга ИТ-инфраструктуры и разработка подхода к их интеграции // Комп'ютерні науки для інформаційного суспільства: Матеріали III Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 12-13 грудня 2012 р.). – Луганськ: Вид-во «Ноулідж», 2012. – С. 128-132.

50. Удовенко Д. В., Жариков Э. В. Облачные вычисления как инновационное направление информационно-коммуникационных технологий // Комп'ютерні науки для інформаційного суспільства: Матеріали III Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 12-13 грудня 2012 р.). – Луганськ: Вид-во «Ноулідж», 2012. – С. 208-210.

51. Жариков Э. В., А. Салем Повышение качества информации в современной ИТ-инфраструктуре // Информатика и вычислительная техника: сборник научных трудов 5-й Всероссийской научно-технической конференции аспирантов, студентов и молодых ученых ИВТ-2013 / под ред. Н. Н. Войта. –Ульяновск: УлГТУ, 2013. – С. 69-74.

52. Жариков Э. В., Альсаясни М. Об одном подходе к интеграции систем мониторинга в современной ИТ-инфраструктуре // Информатика и вычислительная техника: сборник научных трудов 5-й Всерос. научно-техн. конф. аспирантов, студентов и молодых ученых ИВТ-2013. –Ульяновск: УлГТУ, 2013. – С. 65-69.

53. Жариков Э. В. Критерии оптимизации ИТ-инфраструктуры предприятия // Информационно-компьютерные технологии в экономике, образовании и социальной сфере. Выпуск 8. – Симферополь : ФЛП Бондаренко О.А., 2013. – С. 6-7.

54. Obinna J., Zharikov E. Architectural Development of Pivate Cloud for Mid Business Enterprises // Комп'ютерні науки для інформаційного суспільства: Матеріали IV Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 11-12 грудня 2013 р.). – Луганськ: Вид-во «Ноулідж», 2013. – С. 42-46.

55. Adetola R., Zharikov E. Challenges in cloud computing // Комп'ютерні науки для інформаційного суспільства: Матеріали IV Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 11-12 грудня 2013 р.). – Луганськ: Вид-во «Ноулідж», 2013. – С. 33-35.

56. Жариков Э. В., Алдееб С. Неструктурированная информация предприятия, качественный и количественный аспект // Комп'ютерні науки для інформаційного суспільства: Матеріали IV Міжн. наук.-практ. конф. аспірантів та молодих вчених (м. Луганськ, 11-12 грудня 2013 р.). – Луганськ: Вид-во «Ноулідж», 2013. – С. 52-56.

57. Жариков Э. В., Альсаясни М., Лукутин О. В. Идентификация инцидента методом распознавания образов в системах мониторинга // Комп'ютерні науки для інформаційного суспільства: Матеріали IV Міжнародної науково-практичної конференції аспірантів та молодих вчених (м. Луганськ, 11-12 грудня 2013 р.). – Луганськ: Вид-во «Ноулідж», 2013. – С. 56-58.

58. Жариков Э. В., Алдееб С. Моделирование ИТ-инфраструктуры с использованием Archimate // Комп'ютерні науки для інформаційного суспільства: Матеріали V Міжнар. науково-практичної конфер. аспір. та молодих вчених (м. Сєвєродонецьк, 23 грудня 2014 р.). – Сєвєродонецьк: Вид-во СНУ ім.В.Даля, 2014. – С. 21-26.

59. Жариков Э. В., Алдееб С. Анализ компонентов ИТ-инфраструктуры в контексте обработки информации предприятия // Комп'ютерні інтелектуальні системи та мережі: Матеріали VIII Всеукраїнської науково практичної WEB конференції аспірантів, студентів та молодих вчених (24-26 березня 2015 р.). – Кривий Ріг: ДВНЗ «Криворізький національний університет», 2015. – С. 23-26.

60. Жариков Э. В. Архитектура системы управления и мониторинга производительности программно-определяемой ИТ-инфраструктуры // «АВІА-2015»: Матеріали XII Міжнародної науково-технічної конференції (28-29 квітня 2015 р.). – Київ: Національний авіаційний університет, 2015. – С. 638-641.

61. Жариков Э. В. Гиперконвергентная архитектура, анализ возможностей применения в ЦОД // Комп'ютерні інтелектуальні системи та мережі: Матеріали IX Всеукр. науково практичної WEB конференції аспірантів, студентів та молодих вчених (22-24 березня 2016 р.). – Кривий Ріг: ДВНЗ «Кривор. нац. універ.», 2016. – С. 14-19.

62. Коваль А., Жаріков Е. Порівняльний аналіз існуючих методів управління ресурсами в умовах хмарних обчислень // Комп'ютерні інтелектуальні системи та мережі: Матеріали IX Всеукраїнської науково практичної WEB конференції КІСМ-2017 (22-24 березня 2017 р.). – Кривий Ріг: ДВНЗ «Кривор. нац. універ.», 2017. – С. 8-10.

63. Сягайло Т., Жаріков Е. Порівняльний аналіз алгоритмів для управління розміщенням віртуальних машин // Комп'ютерні інтелектуальні системи та мережі: Матеріали IX Всеукраїнської науково практичної WEB конференції КІСМ-2017 (22-24 березня 2017 р.). – Кривий Ріг: ДВНЗ «Кривор. нац. універ.», 2017. – С. 10-13.

64. Терентьєв Р., Жаріков Е. Порівняльний аналіз засобів моделювання інфраструктури хмарних обчислень // Комп'ютерні інтелектуальні системи та мережі: Матеріали IX Всеукраїнської науково практичної WEB конференції КІСМ-2017 (22-24 березня 2017 р.). – Кривий Ріг: ДВНЗ «Кривор. нац. універ.», 2017. – С. 13-16.

65. Andrii Koval, Roman Terentiev, Eduard Zharikov, Comparative Analysis of Modeling Methods of Infrastructure of Cloud Computing, Science and Technology of the XXI Century: the XVIII All-Ukrainian Students R&D Conference, (Kyiv, December 07, 2017) / NTUU „Igor Sikorsky Kyiv Polytechnic Institute". – Part IV. – Kyiv, 2017. – С. 124.

66. Terentiev R., Koval A., Zharikov E. Comparative Analysis of Resource Management Methods in Cloud Computing, Science and Technology of the XXI Century: the XVIII All-Ukrainian Students R&D Conference, (Kyiv, December 07, 2017) / NTUU „Igor Sikorsky Kyiv Polytechnic Institute". – Part V. – Kyiv, 2017. – С. 125.

67. Сягайло Т.А. Способи налаштування сховищ в гіперконвергентних системах./ Жаріков Е.В., Сягайло Т.А., Коваль А.А. // Актуальні наукові дослідження в сучасному світі. - 2018. Вып. 4(36), ч. 3 - С. 41 - 51.

68. Жаріков Е.В., Сердюк Є.О. Консолідація віртуальних машин з використанням променевого пошуку // Інформатика та обчислювальна техніка ІОТ-2017. – Київ: НТУУ "КПІ ім. І. Сікорського", 2017. – С. 79-84

69. Коваль А.А., Жаріков Е.В. Метод розміщення віртуальних машин на основі навчання з підкріпленням // Інформатика та обчислювальна техніка: Матеріали наукової конференції студентів, магістрантів та аспірантів ІОТ-2018 (23 – 24 квітня 2018). – Київ: НТУУ "КПІ ім. І. Сікорського", 2018. – С. 29-35.

70. Сягайло Т.А., Жаріков Е.В. Управління процесом збереження даних в гіперконвергентних системах // Інформатика та обчислювальна техніка: Матеріали наукової конференції студентів, магістрантів та аспірантів ІОТ-2018 (23 – 24 квітня 2018). – Київ: НТУУ "КПІ ім. І. Сікорського", 2018. – С. 119-122.

71. Терентьєв Р.А., Жаріков Е.В. Прогнозування потреби ресурсів серверної системи в умовах хмарних обчислень // Інформатика та обчислювальна техніка: Матеріали наукової конференції студентів, магістрантів та аспірантів ІОТ-2018 (23 – 24 квітня 2018). – Київ: НТУУ "КПІ ім. І. Сікорського", 2018. – С. 123-126.

АНОТАЦІЯ

Жаріков Е. В. Інформаційна технологія управління ІТ-інфраструктурою хмарного центру оброблення даних. На правах рукопису.

Дисертація на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.06 – інформаційні технології. – Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", Міністерство освіти і науки України, Київ, 2020.

У дисертаційній роботі вирішено науково-практичну проблему забезпечення ефективного функціонування ІТ-інфраструктури хмарного ЦОД через створення методології управління та на її основі розроблення і застосування інформаційної технології управління з метою надання хмарних послуг із заданими показниками якості кінцевому користувачеві. Розроблено інформаційну технологію, що забезпечує ефективне функціонування ІТ-інфраструктури ЦОД провайдера хмарних послуг, яка на відміну від відомих базується на методології управління ІТ-інфраструктурою, а також

враховує суттєві характеристики хмарних обчислень, гетерогенність ІТ-інфраструктури, нові архітектури організації обчислень, її багаторівневість та ієрархічність і використовує адаптацію до непередбачуваних навантажень за рахунок застосування розроблених адаптивного методу комбінованого прогнозування та методів управління для різних стратегій управління, що дозволило забезпечити виконання заданих вимог угоди про рівень обслуговування та зниження операційних і капітальних витрат.

Практична цінність інформаційної технології, а також розроблених алгоритмів, методів і підходів полягає у створенні методологічної бази розроблення і реалізації систем управління ІТ-інфраструктурою хмарних ЦОД і підвищення ефективності їх функціонування з подальшим їх застосуванням для розроблення підсистем, компонентів та інших складових систем управління ІТ-інфраструктурою провайдерів хмарних послуг.

Ключові слова: центр оброблення даних, хмарні обчислення, система управління, прогнозування, стохастичний пошук, оптимізація, дворівневе сховище з реплікацією, інтернет речей.

АННОТАЦИЯ

Жариков Э. В. Информационная технология управления ИТ-инфраструктурой облачного центра обработки данных. На правах рукописи.

Диссертация на соискание ученой степени доктора технических наук по специальности 05.13.06 – информационные технологии. – Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского», Министерство образования и науки Украины, Киев, 2020.

В диссертационной работе решена научно-практическая проблема обеспечения эффективного функционирования ИТ-инфраструктуры облачного ЦОД путем создания методологии управления и на ее основе разработки и применения информационной технологии управления с целью предоставления облачных услуг с заданными показателями качества конечному пользователю. Разработана информационная технология, обеспечивающая эффективное функционирование ИТ-инфраструктуры ЦОД провайдера облачных услуг, которая в отличие от известных базируется на методологии управления ИТ-инфраструктурой, а также учитывает существенные характеристики облачных вычислений, гетерогенность ИТ-инфраструктуры, новые архитектуры организации вычислений, ее многоуровневость и иерархичность и использует адаптацию к непредсказуемым нагрузкам за счет применения разработанных адаптивного метода комбинированного прогнозирования и методов управления для различных стратегий управления, что позволило обеспечить выполнение заданных требований соглашения об уровне обслуживания и снижение операционных и капитальных затрат.

Практическая ценность информационной технологии, а также разработанных алгоритмов, методов и подходов заключается в создании методологической базы разработки и реализации систем управления ИТ-инфраструктурой облачных ЦОД и повышения эффективности их функционирования с последующим их применением для разработки подсистем, компонентов и других составляющих систем управления ИТ-инфраструктурой провайдеров облачных услуг.

Ключевые слова: центр обработки данных, облачные вычисления, система управления, прогнозирование, стохастический поиск, оптимизация, двухуровневое хранилище с репликацией, интернет вещей.

ABSTRACT

Zharikov E. V. Information technology for the cloud data center IT infrastructure management. – A manuscript.

The thesis for scientific degree of Doctor of Technical Sciences on the specialty 05.13.06 – Information technologies. – National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ministry of Education and Science of Ukraine, Kyiv, 2020.

The dissertation solves the scientific and practical problem of effective management of the IT infrastructure of cloud data centers under conditions of uncertainty and variable workloads by creating a management methodology and on its basis the development and application of management information technology to provide cloud services with specified quality indicators to the end user. The proposed information technology is developed ensuring the efficient functioning of the IT infrastructure of the cloud service provider's data center by increasing the resource usage efficiency under conditions of variable workload, which unlike the known methods takes into account the developed methodology, essential characteristics of cloud computing, heterogeneity of IT infrastructure, its multilevel and hierarchy and uses adaptation to unpredictable and mixed workloads at the expense of forecasting, which allowed to ensure SLA requirements while reducing operating and capital costs.

The developed information technology is based on the concept of management of IT infrastructure of a cloud data center, which combines the following developed findings, approaches, and methods: the operator form of setting, analyzing and solving problems of IT infrastructure of the cloud data center; identifying and implementing dedicated resource and workload management strategies, as well as their implementation schemes; decomposition of cloud IT infrastructure at three levels (infrastructure, platforms, and applications); taking into account traditional, convergent and hyperconverged architectures of modern cloud data centers; application of the adaptive method of combined workload forecasting to determine the control influences on the IT infrastructure of cloud data centers; developing the Method of Integrated Resource Management for heterogeneous data centers based on SLA violations, power consumption and required power at the next management step; application of stochastic methods (beam search, simulated annealing, reinforcement learning) for implementation of particular strategies for IT infrastructure management of cloud data centers; application of a distributed two-level storage management method for hyperconverged systems; taking into account new metrics of the IT infrastructure state namely instantaneous and average viability coefficients of virtual machine, physical server imbalance indicator, the ratio of necessary resources to the average available resources, threshold of available resources) to determine the current management strategy; accounting of software-defined controllers for management of three primary data center resources namely computing, storage, network; the combination of centralized management using global data center manager and the decentralized management on the physical server level, depending on the chosen management strategy at the current management step.

To solve the problem of effective management of the IT infrastructure of the cloud data center and solving the correspondent set of tasks, the following were used: systems theory, methods of hierarchical systems theory, methods of mathematical programming, methods of operations research and decision theory, methods of mathematical and simulation modeling, methods of artificial intelligence theory, stochastic and heuristic search methods, forecasting

methods, mathematical statistics methods, and cloud service models. The reliability and validity of the obtained results are conditioned by the correct use of the mathematical apparatus and are confirmed by the results of computational experiments.

The scientific novelty of the obtained results is determined by the following theoretical and practical results obtained by the author. For the first time the following has been developed: a methodology for managing the IT infrastructure of a cloud data center; a structural and functional model of a multilevel hierarchical software-defined system for managing the IT infrastructure of a cloud data center; an original information technology for managing the IT infrastructure of a cloud data center; an adaptive method of combined workload forecasting for cloud data center computing resources; a method for integrated management of IT infrastructure of cloud data center; a distributed datacenter cloud management method; a cloud datacenter power management method. The following has been further improved: the decomposition-compensation approach; the algorithms and methods of stochastic local search; an IT infrastructure management model with a coordinator.

Practical value of the information technology, as well as developed methods and approaches, is to create a methodical basis for the development and implementation of management systems for IT infrastructure of cloud data centers and increase their efficiency with their subsequent application for the development of subsystems, components and other parts of the IT infrastructure management systems for cloud service providers.

The most practical results include: management methodology based on the operator form of the management strategy choice and scheme of its implementation; an approach to managing a multi-level hierarchical cloud infrastructure data center IT infrastructure resources; methods of forecasting the workload in cloud data center; methods of managing the resources, workload and power of the cloud data center using forecasts; methods for managing replication and cross-level data migration in a cloud datacenter storage systems.

Keywords: data center, cloud computing, management system, forecasting, stochastic search, optimization, two-level storage with replication, internet of things.