

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

бакалавр

(назва освітнього ступеня)

на тему: Аналіз та дослідження оптимальних стратегій управління енергією
ЦОД у розумному місті

Виконав: студент IV курсу, групи СН-43

спеціальності 122 Комп'ютерні науки
(шифр і назва спеціальності)

Сивопляс В.Ю.
(підпис) (прізвище та ініціали)

Керівник Станько А.А.
(підпис) (прізвище та ініціали)

Нормоконтроль Шимчук Г.В.
(підпис) (прізвище та ініціали)

Завідувач кафедри Боднарчук І.О.
(підпис) (прізвище та ініціали)

Рецензент Варавін А.В.
(підпис) (прізвище та ініціали)

Тернопіль
2025

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ

Завідувач кафедри

Боднарчук І.О.
(підпис) (прізвище та ініціали)

« 27 » січня 2025 р.

**ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ**

на здобуття освітнього ступеня Бакалавр
(назва освітнього ступеня)

за спеціальністю 122 Комп'ютерні науки
(шифр і назва спеціальності)

Студенту Сивоплясу Василю Юрійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Аналіз та дослідження оптимальних стратегій управління енергією ЦОД у розумному місті

Керівник роботи Станько Андрій Андрійович, доктор філософії, ст. викладач кафедри КТ
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від « 07 » травня 2025 року № 4/7-444

2. Термін подання студентом завершеної роботи 28 червня 2024р.

3. Вихідні дані до роботи Інформаційні матеріали та джерела для дослідження оптимальних стратегій управління енергією ЦОД у розумному місті

4. Зміст роботи (перелік питань, які потрібно розробити)

Вступ. 1 Аналіз предметної області енергоспоживання ЦОД. 1.1 Енергоспоживання ЦОД у глобальному та міському контексті. 1.2 Стратегії підвищення ефективності ЦОД. 1.3 Міжнародні стандарти енергоефективності ЦОД. 1.4 Вітчизняний досвід та нормативні аспекти. 1.5 Висновок до першого розділу. 2 Проектна частина. Теоретичне моделювання й оптимальні стратегії. 2.1 Моделювання енергоспоживання та теплових процесів у ЦОД. 2.2 Оптимізація розподілу навантаження та керування ресурсами. 2.3 Алгоритми Динамічного керування живленням та консолідації ВМ. 2.4 Інтеграція ЦОД у розумні енергомережі. 2.5 Висновок до другого розділу. 3 Практична частина. Реалізація задачі оптимізації. 3.1 Опис дата-центру та постановка задачі оптимізації. 3.2 Прогнозування навантаження із застосуванням ML-моделей. 3.3 Динамічне розміщення ВМ: стратегія та алгоритм. 3.4 Фрагмент реалізації алгоритму мовою Python. 3.5 Результати впровадження. 3.6 Висновок до третього розділу. 4. Безпека життєдіяльності, основи охорони праці. Висновки. Перелік джерел. Додатки.

5. Перелік графічного матеріалу

1 Титульна сторінка. 2 Тема, Мета, Об'єкт, Предмет дослідження. 3 Завдання дослідження.

4 Актуальність дослідження. 5 Глобальна динаміка енергоспоживання ЦОД. 6 Матриця «технологія – ефект». 7 Модульна модель енергобалансу ЦОД. 8 Графік «енергія/SLA» при різних стратегіях консолідації. 9 Схема інтеграції ЦОД у Smart-Grid. 10 План-схема приміщення обраного ЦОД. 11 Порівняння точності прогнозу: ARIMA vs LSTM. 12 Таймлайн міграцій ВМ за добу. 13 Блок-схема Python-модуля керування. 14 Стовпчикова діаграма зміни PUE і CO₂. 15 Висновки. 16 Завершальний.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Безпека життєдіяльності, основи охорони праці	Сенчишин В.С. к.т.н., доц. кафедри МТ	02.06.2025	05.06.2025

7. Дата видачі завдання 27 січня 2025 р.

КАЛЕНДАРНИЙ ПЛАН

[illegible]

Студент

(підпис)

Сивопляс В.Ю.

(прізвище та ініціали)

Керівник роботи

(підпис)

Станько А.А.

(прізвище та ініціали)

АНОТАЦІЯ

Аналіз та дослідження оптимальних стратегій управління енергією ЦОД у розумному місті // Кваліфікаційна робота освітнього рівня «Бакалавр» // Сивопляс Василь Юрійович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група СН-43 // Тернопіль, 2025 // С.83 , рис. – 10, табл. – 2, кресл. – 16, додат. – 1, бібліогр. – 56.

Ключові слова: енергоефективність, центр обробки даних, розумне місто, машинне навчання, динамічна консолідація, прогнозування навантаження, PUE.

Кваліфікаційна робота присвячена дослідженню оптимальних стратегій управління енергією ЦОД у розумному місті.

В першому розділі кваліфікаційної роботи освітнього рівня «Бакалавр»: Подано актуальність теми та сформульовано мету дослідження. Розглянуто сучасні підходи до енергоефективності датацентрів і класифіковано існуючі стратегії зниження PUE. Висвітлено нормативно-правові вимоги.

В другому розділі кваліфікаційної роботи: Досліджено математичні моделі енергоспоживання серверів та систем охолодження у зв'язку з профілем навантаження. Обґрунтовано вибір алгоритму динамічної консолідації віртуальних машин із проактивним прогнозом навантаження.

В третьому розділі кваліфікаційної роботи: Розроблено програмний модуль на Python для прогнозування навантаження та автоматичного керування ВМ. Запропоновано процедуру адаптивного вимкнення серверів.

У розділі «Безпека життєдіяльності, основи охорони праці» Висвітлено особливості безпечної роботи з ЦОД а також вимоги безпеки для з комп'ютерними системами.

Об'єкт дослідження: енергоспоживання центрів обробки даних у розумній міській інфраструктурі. Предмет дослідження: стратегії оптимізації управління енергоспоживанням ЦОД із використанням прогнозування навантаження та динамічного розміщення віртуальних машин.

ANNOTATION

Analysis and Research of Optimal Energy Management Strategies in Data Centers in a Smart City // Qualification work of the educational level «Bachelor» // Syvoplias Vasyl Yuriiovych // Ternopil Ivan Pulyu National Technical University, Computer and Information Systems and Software Engineering Faculty, Computer Sciences Department, group SN-43 // Ternopil, 2025 // P.83, fig. – 10, tabl. – 2, chair. – 16, annexes. – 1, references – 56.

Keywords: energy efficiency, data centre, smart city, machine learning, dynamic consolidation, load forecasting, PUE.

The qualification work is devoted to the study of optimal energy management strategies for data centres in a smart city.

The first section of the Bachelor's thesis presents the relevance of the topic and formulates the research objective. It reviews modern approaches to data centre energy efficiency and classifies existing strategies for reducing PUE. Regulatory requirements are highlighted.

The second section of the thesis: Mathematical models of energy consumption of servers and cooling systems in relation to the load profile are investigated. The choice of an algorithm for dynamic consolidation of virtual machines with proactive load forecasting is justified.

In the third section of the thesis: A Python software module for load forecasting and automatic VM management is developed. A procedure for adaptive server shutdown is proposed.

The section 'Life safety, basics of occupational safety' highlights the features of safe work with data centres as well as safety requirements for computer systems.

Research object: energy consumption of data centres in smart urban infrastructure. Research subject: strategies for optimising data centre energy consumption management using load forecasting and dynamic virtual machine placement..

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ЦОД (укр. Центр обробки даних) – спеціалізований комплекс для розміщення ІТ-інфраструктури та її інженерного забезпечення.

AI (англ. Artificial Intelligence) – штучний інтелект.

ARIMA (англ. Autoregressive Integrated Moving Average) – авторегресивно-ковзаючий середній процес із інтеграцією.

ASHRAE (англ. American Society of Heating, Refrigerating and Air-Conditioning Engineers) – Американське товариство інженерів з опалення, охолодження та кондиціонування повітря.

BESS (англ. Battery Energy Storage System) – система акумулювання енергії на основі батарей.

CPU (англ. Central Processing Unit) – центральний процесор.

CUE (англ. Carbon Usage Effectiveness) – коефіцієнт вуглецевої ефективності ЦОД (викиди CO₂ / ІТ-потужність).

DCiE (англ. Data Center infrastructure Efficiency) – ефективність інфраструктури датацентру; обернена величина PUE, виражена у відсотках.

DVFS (англ. Dynamic Voltage and Frequency Scaling) – динамічне масштабування напруги та частоти процесора.

EnMS (англ. Energy Management System) – система енергетичного менеджменту підприємства.

ERE (англ. Energy Reuse Effectiveness) – показник ефективності повторного використання енергії ЦОД.

GA (англ. Genetic Algorithm) – генетичний алгоритм, метасистемний метод оптимізації.

GPU (англ. Graphics Processing Unit) – графічний процесор (прискорювач обчислень).

HVAC (англ. Heating, Ventilation and Air Conditioning) – системи опалення, вентиляції та кондиціонування повітря.

IoT (англ. Internet of Things) – Інтернет речей.

ISO (англ. International Organization for Standardization) – Міжнародна організація зі стандартизації.

ISO 50001 – міжнародний стандарт систем енергетичного менеджменту.

ISO/IEC 30134 – серія міжнародних стандартів KPI для енерго- та ресурсної ефективності ЦОД.

IT (англ. Information Technology) – інформаційні технології.

kВт·год (англ. kWh, kilowatt-hour) – кіловат-година, одиниця роботи / енергії.

LSTM (англ. Long Short-Term Memory) – мережа довготривалої та короткотривалої пам'яті (тип RNN).

MPC (англ. Model Predictive Control) – модельно-прогнозне керування.

ML (англ. Machine Learning) – машинне навчання.

PDU (англ. Power Distribution Unit) – блок розподілу електроживлення в серверних шафах.

PSO (англ. Particle Swarm Optimization) – оптимізація роєм частинок.

PUE (англ. Power Usage Effectiveness) – коефіцієнт ефективності використання енергії ЦОД (загальне споживання / IT-споживання).

RNN (англ. Recurrent Neural Network) – рекурентна нейронна мережа.

RL (англ. Reinforcement Learning) – навчання з підкріпленням.

SLA (англ. Service Level Agreement) – угода про рівень сервісу між постачальником та споживачем послуги.

UPS (англ. Uninterruptible Power Supply) – джерело безперебійного живлення.

URL (англ. Uniform Resource Locator) – уніфікований локатор ресурсу.

VM (англ. Virtual Machine) – віртуальна машина, ізольоване програмне середовище на фізичному сервері.

WUE (англ. Water Usage Effectiveness) – показник ефективності використання води у датацентрі (м³ води / IT-потужність).

ЗМІСТ

ВСТУП	9
РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ЕНЕРГОСПОЖИВАННЯ	
ЦОД	12
1.1 Енергоспоживання ЦОД у глобальному та міському контексті	12
1.2 Стратегії підвищення енергоефективності ЦОД	15
1.3 Міжнародні стандарти енергоефективності ЦОД.....	21
1.4 Вітчизняний досвід та нормативні аспекти	25
1.5 Висновок до першого розділу	28
РОЗДІЛ 2. ПРОЕКТНА ЧАСТИНА. ТЕОРЕТИЧНЕ МОДЕЛЮВАННЯ Й	
ОПТИМАЛЬНІ СТРАТЕГІЇ.....	30
2.1 Моделювання енергоспоживання та теплових процесів у ЦОД	30
2.2 Оптимізація розподілу навантаження та керування ресурсами	34
2.3 Алгоритми динамічного керування живленням та консолідації ВМ	39
2.4 Інтеграція ЦОД у розумні енергомережі	44
2.5 Висновок до другого розділу	46
РОЗДІЛ 3. ПРАКТИЧНА ЧАСТИНА. РЕАЛІЗАЦІЯ ЗАДАЧІ	
ОПТИМІЗАЦІЇ	48
3.1 Опис дата-центру та постановка задачі оптимізації.....	48
3.2 Прогнозування навантаження із застосуванням ML-моделей	51
3.3 Динамічне розміщення ВМ: стратегія та алгоритм	53
3.4 Фрагмент реалізації алгоритму мовою Python	57
3.5 Результати впровадження.....	59
3.6 Висновки до третього розділу	61

РОЗДІЛ 4. БЕЗПЕКА ЖИТТЄДІЯЛЬНОСТІ, ОСНОВИ ОХОРОНИ ПРАЦІ	62
4.1 Питання щодо безпеки життєдіяльності	62
4.2 Питання з основ охорони праці	67
4.3 Висновок до четвертого розділу	71
ВИСНОВКИ.....	73
ПЕРЕЛІК ДЖЕРЕЛ	78
ДОДАТКИ	

ВСТУП

Актуальність теми. Стрімке зростання ролі інформаційних технологій спричинило вибухове збільшення кількості та масштабів центрів обробки даних (ЦОД), які є невід’ємною складовою цифрової інфраструктури сучасних міст. ЦОД забезпечують роботу хмарних сервісів, телекомунікацій, систем «розумного міста» та критичних застосунків, проте вони споживають колосальні обсяги електроенергії [1]. За оцінками Міжнародного енергетичного агентства, глобальне енергоспоживання дата-центрів станом на початок 2020-х років становить близько 1–2% загального споживання електрики і може більш ніж подвоїтися до 2030 року, досягнувши ~945 ТВт·год на рік. Така тенденція частково зумовлена впровадженням штучного інтелекту та зростаючим попитом на обчислення й зберігання даних, що ставить серйозні виклики перед енергетичною інфраструктурою та екологією. Водночас ефективність ЦОД поступово поліпшується: середній показник PUE (коефіцієнт ефективності використання енергії) світових дата-центрів останніми роками стабілізувався на рівні ~1.55, значно зменшившись порівняно з середнім $PUE \approx 2.5$ у 2007 році. Це свідчить про успіхи у впровадженні енергоощадних технологій та практик, однак подальші поліпшення є необхідними з огляду на прогнозоване зростання навантажень [2].

Управління енергоспоживанням ЦОД у концепції «розумного міста» набуває особливої актуальності, оскільки поєднує два критично важливі напрями сталого розвитку: цифрову трансформацію міської інфраструктури та енергоефективність. Розумні міста впроваджують інтернет речей (IoT), хмарні сервіси, великі дані та штучний інтелект для підвищення якості життя мешканців, і ЦОД виступають обчислювальним ядром цих систем. Водночас міста зобов’язані скорочувати вуглецевий слід та оптимізувати використання енергії відповідно до міжнародних зобов’язань щодо клімату. За таких умов необхідно розробляти та впроваджувати оптимальні стратегії керування

енергією ЦОД, які б забезпечили надійне функціонування цифрових сервісів, мінімізуючи енергозатрати та негативний вплив на довкілля.

Мета і задачі дослідження. Метою роботи є аналіз сучасних підходів та розробка рекомендацій щодо оптимальних стратегій управління енергією центрів обробки даних у розумному місті. Для досягнення поставленої мети вирішено такі завдання:

- провести аналітичний огляд світових тенденцій енергоспоживання ЦОД, основних викликів та існуючих рішень підвищення енергоефективності;
- дослідити сучасні теоретичні моделі енергоспоживання і теплових режимів дата-центрів, а також методи оптимізації керування навантаженням, включно з використанням машинного навчання;
- проаналізувати міжнародні стандарти (ASHRAE, ISO 50001, “Green Grid” тощо) та ініціативи (EU Code of Conduct, програми енергоефективності) в галузі енергоощадності ЦОД, а також нормативно-правову базу України в цій сфері;
- розробити і продемонструвати на прикладі практичну реалізацію стратегії енергоменеджменту ЦОД, що включає ML-передбачення навантаження, динамічну консолідацію віртуальних машин та оцінку ефективності за ключовими метриками (PUE, DCiE, вуглецевий слід тощо).

Практичне значення одержаних результатів. Наукова новизна роботи полягає у комплексному підході до проблеми енергоменеджменту ЦОД, що поєднує аналітичний огляд, теоретичне моделювання та практичне впровадження. У роботі узагальнено сучасні світові напрацювання з підвищення енергоефективності ЦОД, зокрема впровадження алгоритмів динамічного керування ресурсами на основі прогнозування навантаження та використання штучного інтелекту [16] [19].

Проаналізовано можливості інтеграції ЦОД у розумні енергомережі міста (наприклад, участь у програмах Demand Response) для згладжування пікових навантажень та кращого використання відновлюваної енергії. Практична значущість роботи полягає в тому, що запропоновані підходи й рекомендації

можуть бути використані операторами дата-центрів та міськими ІТ-службами для зменшення експлуатаційних витрат на електроенергію, покращення показників ефективності (PUE, CUE) та відповідності екологічним стандартам. Результати моделювання і експериментальної перевірки алгоритмів керування живленням можуть стати основою для подальших розробок систем моніторингу й автоматичного управління інфраструктурою ЦОД у складі розумних міст.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ЕНЕРГОСПОЖИВАННЯ ЦОД

1.1 Енергоспоживання ЦОД у глобальному та міському контексті

Центри обробки даних є енергомісткими об'єктами, що постійно споживають значну потужність для роботи ІТ-обладнання (серверів, систем зберігання, мережевого устаткування) та підтримки необхідного мікроклімату (охолодження, вентиляція, резервне живлення). За даними досліджень, у 2022 році сукупне електроспоживання дата-центрів становило приблизно 460 ТВт·год, тобто близько 2% від світового споживання електроенергії. Це співставно з річним споживанням електрики такими країнами, як Швеція або Польща. Аналітики прогнозують подальше зростання цієї частки: зокрема, вплив нових обчислювально інтенсивних навантажень (штучний інтелект, машинне навчання, великі мовні моделі) може подвоїти споживання енергії ЦОД протягом найближчих 5–10 років. Наприклад, згідно з нещодавнім звітом ІЕА, тільки впровадження ІІ у пошукові сервіси може збільшити їх енергоспоживання в десятки разів – один запит до мовної моделі на кшталт ChatGPT споживає на порядок більше енергії, ніж традиційний пошуковий запит. Такі тенденції викликають занепокоєння щодо стабільності електромереж та стимулюють пошук рішень для підвищення ефективності використання енергії.

Основними компонентами енергоспоживання ЦОД є: (i) ІТ-навантаження – електроенергія, що споживається серверами, системами зберігання даних та мережевими пристроями для виконання обчислень і обробки трафіку; (ii) інфраструктурне споживання – енергія, витрачена на допоміжні системи: охолодження (чилери, кондиціонери, насосні станції), системи живлення (блоки живлення, джерела безперебійного живлення – UPS), освітлення тощо. Для оцінки ефективності використовується показник PUE (Power Usage Effectiveness) – відношення загального енергоспоживання ЦОД до енергії,

спожитої IT-обладнанням. Ідеальний $PUE \rightarrow 1$ означає, що вся енергія йде на корисну роботу серверів (немає витрат на охолодження та втрат в інфраструктурі). Реальні значення PUE у сучасних дата-центрах знаходяться в межах 1.1–1.8, залежно від масштабу та ефективності систем охолодження та електроживлення. За даними щорічного глобального опитування Uptime Institute, середнє значення PUE у 2022 році становило 1.55. Це значне покращення порівняно з 2007 роком ($PUE \sim 2.5$), однак прогрес останніх років сповільнився – показник фактично “застряг” на рівні ~ 1.5 – 1.6 з 2013 року. Причини цього – наближення до фізичних меж ефективності традиційних технологій охолодження й живлення та необхідність принципово нових підходів для подальшого зниження втрат [1]. На рисунку 1.1 наведено лінійний графік, що показує ріст світового споживання електроенергії дата-центрів у ТВт·год на тлі загального світового споживання, з виділенням прогнозу до 2030 р. Пояснює масштаби проблеми.

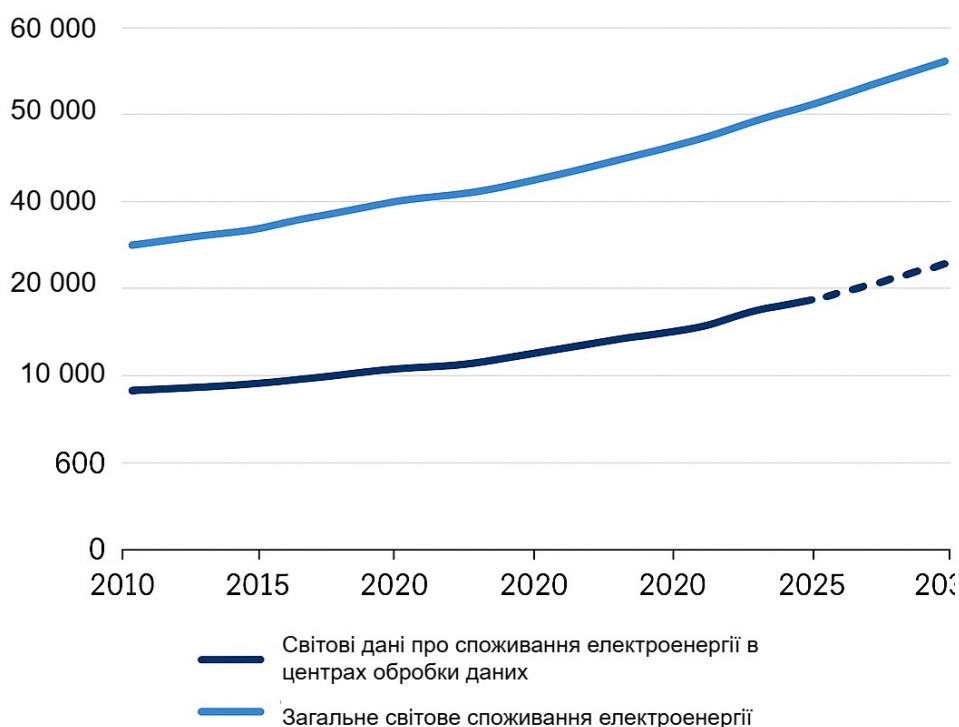


Рисунок 1.1 – Глобальна динаміка енергоспоживання ЦОД (2010 – 2030 рр.)

Характерно, що великі хмарні провайдери досягають кращих показників: наприклад, у дата-центрах Google середній PUE на кінець 2020-х наблизився до 1.2, а кращі об'єкти демонструють ~ 1.10 за рахунок використання передових систем охолодження та керування. Деякі «гіпермасштабні» центри обробки даних, побудовані в прохолодному кліматі та оптимізовані під максимальну ефективність, звітують про $PUE \approx 1.1-1.2$ (наприклад, об'єкти в Скандинавії, що використовують фрі-кулінг – пряме зовнішнє охолодження повітрям). Отже, існує значний розрив між найкращими практиками та середнім рівнем ефективності у галузі, що вказує на потенціал впровадження енергоефективних технологій у широкому масштабі [2].

У контексті розумного міста енергоспоживання ЦОД набуває особливого значення з двох причин. По-перше, самі дата-центри стають елементами міської інфраструктури (міські хмарні платформи, муніципальні ЦОД для надання електронних послуг, центр керування «Безпечне місто» тощо) і збільшують навантаження на міські енергомережі. По-друге, розумне місто передбачає впровадження енергоефективних рішень та сталий розвиток, тож підвищення ефективності міських ЦОД безпосередньо сприяє досягненню цілей енергозбереження. Більш того, взаємодія ЦОД з інтелектуальними електромережами (smart grid) відкриває можливості для оптимізації споживання: дата-центри можуть брати участь у програмах demand response, тимчасово знижуючи споживання при пікових навантаженнях у місті або переключаючись на резервні джерела, що допомагає збалансувати мережу. Деякі оператори вже практикують такі заходи. Наприклад, Google повідомляє про успішне пілотування системи автоматичного зменшення споживання їхніх ЦОД у години піку локальної мережі, шляхом перенесення нетермінових завдань на інший час доби – це дозволило знизити навантаження на енергосистему регіону без шкоди для якості сервісів. Інший напрям – розміщення міських міні-ЦОД (edge data centers) поруч зі споживачами, що зменшує затримки передачі даних, але потребує продуманого енергопостачання. В рамках концепції «Green Smart City» пропонується

інтегрувати такі розподілені дата-центри з відновлюваними джерелами енергії (сонячними панелями на дахах, міськими вітрогенераторами) та системами зберігання енергії, аби максимально використовувати локальну чисту енергію для їх роботи. Крім того, тепловиділення ЦОД може стати корисним ресурсом для міста – у ряді країн (наприклад, Нідерланди, Данія) реалізовано проєкти використання відпрацьованого тепла дата-центрів для обігріву будівель і теплопостачання, що підвищує загальну енергетичну ефективність та вписується в економіку замкненого циклу.

Таким чином, сучасні ЦОД виступають значними споживачами енергії, але водночас мають великий резерв підвищення ефективності. У масштабі розумного міста оптимізація роботи дата-центрів дозволить не лише зменшити витрати та викиди, але й інтегровано керувати споживанням в інтересах всієї громади – наприклад, знизити ризики перенавантаження мереж, збільшити використання відновлюваної енергії, забезпечити стійкість до аварійних відключень електропостачання. Детальніше існуючі підходи до досягнення цих цілей розглянемо у наступному підрозділі.

1.2 Стратегії підвищення енергоефективності ЦОД

За останні два десятиліття у світі напрацьовано широкий спектр технологічних та організаційних стратегій, спрямованих на зменшення енергоспоживання центрів обробки даних або підвищення корисної віддачі від спожитої енергії. Ці стратегії охоплюють різні рівні – від апаратного (вдосконалення елементної бази) та програмного (керування навантаженням) до рівня інженерної інфраструктури (системи охолодження, живлення) і навіть планування розміщення ЦОД з урахуванням клімату. Проведемо огляд основних підходів, які нині використовуються або досліджуються для енергооптимізації дата-центрів:

- Використання енергоефективного обладнання. Сучасні сервери та мережеві пристрої проєктуються з урахуванням принципу

енергопропорційності – споживання енергії має бути пропорційним поточному навантаженню. Старіші покоління серверів мали проблему: навіть за низької завантажки (10–20%) вони споживали 50–70% від повної потужності, що призводило до великої частки марних витрат. Новітні процесори впроваджують технології динамічного масштабування напруги і частоти (DVFS), що дозволяють знижувати енергоспоживання ядра при простої або малому навантаженні. Також у серверах застосовуються більш ефективні блоки живлення (коефіцієнт корисної дії БЖ сучасних серверів 94–96% проти 85–90% десять років тому) та компоненти з меншими втратами. Прогрес у області прискорення обчислень – використання графічних процесорів (GPU), тензорних процесорів та спеціалізованих прискорювачів – дає змогу виконувати певні типи задач значно швидше та з меншими енерговитратами порівняно з універсальними CPU. За оцінками NVIDIA, перехід до прискорених обчислень дозволяє отримати 2–3-кратний виграш в продуктивності на ват для багатьох навантажень, таких як аналітика даних, машинне навчання тощо. Таким чином, оновлення апаратного парку ЦОД на сучасне обладнання (високоєфективні процесори, пам'ять, мережеві інтерфейси) – одна з базових стратегій зниження питомого енергоспоживання [10].

– Віртуалізація та консолідація серверів. Традиційно в ЦОД було прийнято розгортати на фізичному сервері одну службу або додаток, що часто призводило до низького коефіцієнта використання ресурсів: сервер може простоювати або працювати на 10–20% потужності, але при цьому споживати майже повну потужність. Технології віртуалізації дозволили розміщувати на одному фізичному хості кілька ізольованих віртуальних машин (VM) або контейнерів, тим самим консолідувати навантаження. Це дає змогу вимкнути надлишкові фізичні сервери і завдяки цьому зменшити загальне споживання енергії без втрати продуктивності системи. Дослідження показують, що агресивна консолідація навантажень за допомогою віртуалізації може скоротити число активно працюючих серверів на 20–40% залежно від профілю навантаження, відповідно скоротивши і енергоспоживання приблизно на цю ж

величину. Ключове завдання – забезпечити баланс між економією енергії та якістю обслуговування: надмірна консолідація може призвести до перевантаження окремих вузлів і деградації продуктивності (порушення SLA). Тому активно досліджуються алгоритми динамічної міграції ВМ: система моніторить використання ресурсів серверів і при виявленні недовантажених або перевантажених вузлів виконує live-migration (міграцію на льоту) деяких ВМ з метою вирівнювання навантаження та відключення «зайвих» серверів. Наприклад, у роботі [19] запропоновано гібридну ML-модель для прогнозування необхідності міграції ВМ та вибору оптимального цільового хоста; це дозволило досягти 99% точності у визначенні міграцій і знизити кількість простоїв без шкоди для продуктивності. Інші дослідники впроваджують проактивні методи: не чекати перевантаження сервера, а передбачати піки навантаження наперед. Так, у свіжому дослідженні [20] використано нейромережеву модель VTGAN для прогнозування трендів використання CPU серверів; інтеграція такого прогнозу в алгоритм розміщення ВМ дала змогу на 79% зменшити кількість порушень SLA та на 56% скоротити число міграцій порівняно з реактивним алгоритмом консолідації. Отже, енергоефективне керування ресурсами [23] через віртуалізацію наразі є однією з найбільш дієвих стратегій: вона реалізована у більшості хмарних платформ (VMware, Hyper-V, OpenStack) і постійно вдосконалюється завдяки штучному інтелекту та алгоритмам оптимізації [17].

– Оптимізація систем охолодження. Охолодження зазвичай споживає 30–50% від енергії, яку витрачає ЦОД (чим гірше спроектована чи перевантажена система охолодження, тим більша ця частка). Тому удосконалення HVAC (Heating, Ventilation and Air Conditioning) є критично важливим для зниження PUE. Існує кілька напрямів покращення: по-перше, впровадження ефективніших технологій охолодження – замість традиційного повітряного кондиціонування використовують пряме фрі-кулінг охолодження повітрям, де холодне зовнішнє повітря за певних погодних умов подається безпосередньо у серверні зали, або випарне охолодження, де відбувається охолодження за

рахунок випаровування води. У регіонах із холодним кліматом більшу частину року зовнішнє охолодження дозволяє майже не витрачати енергію на кондиціонери. По-друге, застосування рідинного охолодження: холодоагент (вода або діелектрична рідина) підводиться безпосередньо до серверів (через водяні блоки, теплообмінники в шафах або повне занурення обладнання у баки з рідиною). Рідинне охолодження значно підвищує ефективність тепловідведення і дозволяє працювати за більш високих температур охолоджувача (наприклад, вода 30°C замість повітря 20°C) без перегріву, що дає змогу мінімізувати використання компресорних чилерів. По-третє, оптимізація повітряних потоків у серверних залах: архітектура «гарячий і холодний коридор», коли стійки серверів розташовують у ряд так, щоб з одного боку всі сервери забирали холодне повітря (холодний коридор), а тепле повітря викидали в протилежний бік (гарячий коридор). Це запобігає змішуванню потоків і підвищує ефективність кондиціонування. Також практикується ізоляція гарячих коридорів прозорими панелями. По-четверте, впровадження інтелектуальних систем моніторингу клімату та регулювання в реальному часі: сучасні DCIM-системи (Data Center Infrastructure Management) дозволяють знімати показники температури/вологості у різних зонах і динамічно керувати налаштуваннями охолодження (швидкістю вентиляторів, продуктивністю насосів, температурою подачі охолоджувача). Зокрема, відомий приклад – застосування ІІІ компанією Google для керування охолодженням в режимі реального часу, що дало змогу зменшити енерговитрати на охолодження її дата-центрів до 40%. Алгоритм DeepMind аналізував дані тисяч датчиків (температура, тиск, потоки) і видавав оптимальні рекомендації для системи керування – в результаті деякі об'єкти Google досягли рекордного $PUE \approx 1.10$ саме за рахунок зниження частки енергії на охолодження. Цей випадок продемонстрував великий потенціал AI-технологій у цій сфері [21]. Нарешті, варто відзначити підхід підвищення толерантності до температур: стандарти ASHRAE дозволяють утримувати обладнання при вищих температурах, ніж це робили раніше, – сучасний допуск для класу A1 серверів сягає +32°C в зоні

входу повітря. Це означає, що оператори ЦОД можуть підняти уставки термостатів (наприклад, з 20°C до 25°C) і тим самим скоротити витрати на охолодження без шкоди для надійності, що активно рекомендується профільними організаціями. Комплекс реалізації зазначених заходів дозволяє суттєво знизити PUE: зафіксовано випадки, коли модернізація системи охолодження (перехід на фрі-кулінг + організація коридорів) зменшувала PUE з ~1.8 до ~1.3 лише за рахунок падіння споживання кондиціонерів [20].

– Керування живленням та відновлювані джерела. Ще один важливий напрям – оптимізація електроживлення ЦОД. Значні втрати енергії можуть відбуватися в UPS (джерелах безперебійного живлення) та блоках розподілу живлення. Новітні системи UPS модульного типу з подвійним перетворенням мають ККД 95–98% на номінальному навантаженні, що вище за показники попередніх поколінь (~90%). Крім того, застосовується режим “eco-mode”, коли за нормальних умов мережеве живлення подається безпосередньо, мінаючи перетворювачі UPS (з мінімальними втратами), а на інвертор UPS перемикається лише в разі зникнення зовнішньої мережі. Це може додатково заощадити кілька відсотків енергії. Щодо розподілу – використання сучасних PDU (Power Distribution Units) з моніторингом струмів і оптимального фазування навантаження також сприяє підвищенню ефективності. Окрім зниження власних витрат, стратегія енергоефективного ЦОД передбачає і декарбонізацію джерел енергії. Провідні світові компанії (Google, Microsoft, Amazon) вже досягли 100% покриття споживання ЦОД відновлюваними джерелами шляхом укладання PPA-контрактів (power purchase agreement) з виробниками “зеленої” енергії. Наприклад, Microsoft заявляє про повне забезпечення своїх дата-центрів електроенергією з вітрових та сонячних електростанцій, що підтверджено сертифікатами відновлюваної енергії. В Україні також з’являються подібні ініціативи: провайдери ЦОД починають цікавитися можливістю закупівлі “зеленої” електроенергії на ринку або встановлення власних СЕС на дахах дата-центрів. Окремо варто згадати про використання відпрацьованого тепла: як вже зазначалось, теплова енергія від

серверів може не витрачатися марно, а відводитися у міські тепломережі. У Стокгольмі, Гельсінкі, Цюріху діють системи, де великі ЦОД підключені до централізованого теплозабезпечення, віддаючи туди десятки мегават тепла і скорочуючи споживання газу/вугілля для опалення. Це приклад економіки замкненого циклу, коли побічний продукт (тепло) стає ресурсом. Комплексний підхід до енергоменеджменту ЦОД – від підвищення ефективності до переходу на чисті джерела – дозволяє не лише знизити операційні витрати, а й досягти стратегічних цілей сталого розвитку, таких як зменшення вуглецевого сліду (метрика CUE, Carbon Usage Effectiveness, що визначає кількість викидів CO₂ на одиницю енергоспоживання ІТ-навантаження) [28]. На рисунку 1.2 наведено лінійний графік, що показує ріст світового споживання електроенергії дата-центрів у ТВт·год на тлі загального світового споживання, з виділенням прогнозу до 2030 р. Пояснює масштаби проблеми

	PUE	CAPEX	OPEX
Віртуалізація			
Рідинне охолодження			
Відновлювана енергія			
Інтелектуальне керування			
Когерентне живлення			
Інвентиворація повітря			

Рисунок 1.2 – Матриця «технологія – ефект» для ключових стратегій енергоефективності

Отже, на сьогодні існує чимало усталених практик з енергозбереження в дата-центрах. Згідно з оглядом [11], до ключових належать: впровадження

поновлюваної енергетики, вдосконалені системи охолодження (рідинне, випарне), ізоляція гарячих/холодних зон, централізований моніторинг і керування інфраструктурою, утилізація відпрацьованого тепла та переробка обладнання. Літературні джерела [17] [18] також відзначають критичну роль алгоритмів управління навантаженням (віртуалізації, розподілу задач) та сучасних методів штучного інтелекту в досягненні нових рівнів енергоефективності. Проте кожна стратегія має свої межі ефективності та сфери застосування. Максимального результату можна досягти лише комбінуванням заходів: наприклад, використання ML-алгоритмів для оптимального розміщення задач (щоб якомога менше серверів було активними) разом із підвищенням температурних уставок охолодження і переходом на відновлювану енергію дозволить одночасно зменшити PUE і знизити CUE (вуглецевий слід) практично до нуля. Подальший текст роботи присвячений детальнішому розгляду теоретичних моделей та алгоритмів, що лежать в основі зазначених стратегій, а також аналізу стандартів, які регулюють енергоефективність ЦОД [34].

1.3 Міжнародні стандарти енергоефективності ЦОД

Поширення найкращих практик енергоменеджменту ЦОД значною мірою забезпечується через систему міжнародних стандартів, галузевих керівництв та добровільних програм. Вони встановлюють єдині метрики оцінювання, мінімальні вимоги до ефективності обладнання та інфраструктури, а також надають рекомендації щодо проектування і експлуатації дата-центрів [35]. Розглянемо найважливіші з них:

- Метрики PUE/DCiE та стандарти ISO/IEC 30134. Показник PUE, запроваджений організацією The Green Grid у 2007 році, став де-факто глобальним стандартом вимірювання енергоефективності ЦОД. У 2016 році його було формалізовано як міжнародний стандарт ISO/IEC 30134-2:2016, що визначає методику розрахунку PUE. Відповідно, з'явився і похідний показник

DCiE (Data Center infrastructure Efficiency) – обернена величина до PUE, виражена у відсотках ($DCiE = 1/PUE * 100\%$). Хоча DCiE використовується рідше, обидва терміни присутні у стандартах. Крім PUE, розроблено цілу серію ISO/IEC 30134 для інших KPI ЦОД: так, ISO 30134-3 визначає CUE (Carbon Usage Effectiveness) – коефіцієнт вуглецевої ефективності, що обчислюється як відношення загальних викидів CO₂ (в еквіваленті, спричинених споживанням енергії ЦОД) до енергії, спожитої IT-обладнанням. Інші метрики включають WUE (Water Usage Efficiency) – використання води (ISO/IEC 30134-5:2018) та ERE (Energy Reuse Efficiency) – коефіцієнт повторного використання енергії. Стандартизація цих показників дозволяє об'єктивно порівнювати ефективність різних центрів обробки даних та відстежувати прогрес у динаміці [17-18].

– Стандарти ASHRAE для дата-центрів. Американське товариство інженерів з опалення, холодопостачання і кондиціонування повітря (ASHRAE) розробило декілька важливих стандартів, зокрема ASHRAE 90.4 «Energy Standard for Data Centers». Цей стандарт (перша редакція 2016 року, актуальна редакція 2019) встановлює мінімальні вимоги до енергоефективності при проєктуванні та експлуатації ЦОД. На відміну від загальнобудівельного стандарту 90.1, версія 90.4 враховує специфіку дата-центрів – високі щільності тепловиділень, особливі системи живлення. Стандарт 90.4 є переважно продуктивно-орієнтованим: він вводить показники Mechanical Load Component (MLC) та Electrical Loss Component (ELC), які характеризують ефективність відповідно систем охолодження і живлення; для різних кліматичних зон задано граничні допустимі значення MLC/ELC, і вимагається, щоб фактичні розрахункові показники проєкту не перевищували цих значень. Таким чином досягається гнучкість – проєктанти можуть обирати будь-які технічні рішення, аби сукупний результат вписався в нормативні межі. Цікаво, що ASHRAE 90.4 спершу містив пряму вимогу щодо PUE (максимально допустимі значення), але через дискусійність цього підходу в фінальній версії відмовилися від згадки PUE. Натомість, обмеження введено опосередковано через MLC/ELC. Окрім 90.4, важливим є стандарт ASHRAE TC9.9 «Thermal Guidelines for Data

Processing Environments», який регламентує допустимі діапазони температури і вологості для ІТ-обладнання. Відповідно до нього визначено клас A1 (стандартні сервери) з діапазоном 18–27°C (рекомендовано) до 15–32°C (допустимо) на впуску повітря в сервер, клас A2/A3/A4 для телеком обладнання з ширшими діапазонами та класи для обладнання із рідинним охолодженням. Виконання цих рекомендацій дозволяє операторам підвищувати температуру в ЦОД до верхніх меж і тим самим економити енергію на охолодження без ризику пошкодження обладнання. ASHRAE також спільно з The Green Grid публікує довідники – наприклад, «PUE: A Comprehensive Examination of the Metric» (2013) – для детального пояснення методологій та прикладів застосування метрик PUE, охоплюючи сценарії вимірювань і підвищення ефективності.

– Стандарт ISO 50001 (енергоменеджмент). Хоча він не спеціалізований для дата-центрів, його значення важко переоцінити. ISO 50001:2018 – це міжнародний стандарт систем енергетичного менеджменту (EnMS), який задає структуру і вимоги для підприємств щодо раціонального управління енергоспоживанням. Організації, впроваджуючи ISO 50001, зобов'язані розробити енергетичну політику, цілі та плани дій з підвищення енергоефективності, проводити моніторинг та постійне вдосконалення за циклом PDCA (Plan-Do-Check-Act). Для ЦОД впровадження такої системи дає значні результати: приклад компанії Google демонструє, що за допомогою EnMS вдалося заощадити понад \$1 млрд операційних витрат протягом трьох років. Сертифікація за ISO 50001 стає своєрідним знаком якості для дата-центрів: наприклад, провайдер Iron Mountain повідомляє, що він – єдиний на сьогодні глобальний оператор ЦОД, сертифікований за ISO 50001 по всіх своїх об'єктах. У контексті України, із набуттям чинності Закону України «Про енергетичну ефективність» №1818-IX (2021) запроваджується вимога для великих споживачів енергії (до яких можуть належати й ЦОД) впроваджувати системи енергоменеджменту або періодично проходити енергетичні аудити. Це

стимулюватиме застосування ISO 50001 та схожих стандартів і в українській IT-індустрії [26].

– Програми та кодекси енергоефективності. В Європейському Союзі діє EU Code of Conduct for Data Centre Energy Efficiency – добровільний кодекс поведінки для дата-центрів щодо енергозбереження. Учасники (оператори ЦОД, клієнти, постачальники) беруть на себе зобов’язання впроваджувати найкращі практики з переліку, розробленого Європейською комісією. Документ містить понад 150 детальних рекомендацій – від архітектури будівлі (теплоізоляція, розташування за географією) до експлуатації (калібрування датчиків, оптимальні рівні заповнення стійок, управління потоками повітря тощо). Щороку оновлюються «Best Practices Guidelines», наразі актуальна версія 2023 року. Учасники, які досягають вагомих успіхів (значне зниження PUE, економія кВт·год), відзначаються нагородами від ЄС. Такий кодекс стимулює обмін досвідом і підвищення культури енергоменеджменту в галузі. Окрім того, існують галузеві сертифікації: LEED (Leadership in Energy and Environmental Design) – система екологічної сертифікації будівель. Дата-центри можуть отримувати рівні LEED (Silver, Gold, Platinum) за впровадження енергоефективних і екологічних технологій при будівництві та експлуатації. Наприклад, дата-центри Microsoft мають сертифікацію LEED Gold, що підтверджує високу екологічність проєктних рішень. Також варто згадати стандарти серії EN 50600 / ISO/IEC 22237 – це комплекс європейських стандартів для інфраструктури ЦОД (від загальних понять до вимог до будівлі, електроживлення, телекомунікацій, охолодження). Серед іншого, там є частини, присвячені енергоефективності та забезпеченню безперервності роботи. В Україні ці стандарти впроваджено як національні ДСТУ (наприклад, ДСТУ EN 50600-1:2018 визначає загальні положення побудови інфраструктури ЦОД). Таким чином, нормативна база для підвищення енергоефективності дата-центрів постійно розвивається як на глобальному, так і на національному рівні [24].

Отже, міжнародні стандарти та програми відіграють ключову роль в поширенні енергоефективних практик. Вони уніфікують вимоги, забезпечують прозорість і порівнянність показників, а також стимулюють інженерів шукати інноваційні рішення. Дотримання таких стандартів – важлива складова конкурентоздатності сучасних ЦОД, адже клієнти все більше уваги звертають на екологічність хмарних сервісів та «зелені» ініціативи компаній. Для українських дата-центрів імплементація цих стандартів є шляхом до інтеграції в глобальний ринок: як показує досвід, запровадження систем енергоменеджменту і сертифікація (ISO 50001, LEED тощо) дозволяє не лише зекономити кошти, але й підвищити довіру клієнтів та інвесторів.

1.4 Вітчизняний досвід та нормативні аспекти

В Україні питання енергетичної ефективності центрів обробки даних лише починає виходити на порядок денний, проте вже є певні напрацювання у цій сфері. З одного боку, великі комерційні ЦОД та дата-центри операторів зв'язку впроваджують передові технології, прагнучи знизити експлуатаційні витрати. З іншого – держава приймає законодавчі акти, що встановлюють загальні рамки енергозбереження, які поширюються і на сферу ІКТ. Розглянемо декілька аспектів:

Нормативна база. Україна адаптує своє законодавство до європейських директив у галузі енергоефективності. Зокрема, закон «Про енергетичну ефективність» №1818-IX від 21.10.2021 р. (набрав чинності з 2025 року) закладає основи державної політики енергозбереження на всіх рівнях економіки. Він передбачає розробку національних планів дій з енергоефективності, зобов'язує центральні та місцеві органи влади планувати заходи з енергозбереження. Хоча прямо дата-центри в законі не згадуються, багато його положень опосередковано стосуються цієї сфери. Наприклад, запроваджується система енергоаудиту для великих споживачів енергії – ЦОД потужністю в сотні кіловат і більше можуть підпадати під критерії проведення

регулярних енергоаудитів, що виявлятимуть резерви економії. Також закон стимулює впровадження енергоменеджменту (ISO 50001) на підприємствах. Ще раніше було прийнято Закон України «Про енергетичну ефективність будівель» №2118-VIII (2017), який встановив мінімальні вимоги до енергоефективності для будівель та механізм сертифікації енергетичної ефективності будівель. Окремі великі ЦОД можуть проходити таку сертифікацію як спеціалізовані будівлі. Крім законів, існують відомчі нормативи: Державне агентство з енергоефективності (Держенергоефективності) затвердило Перелік національних стандартів, гармонізованих з європейськими, у сфері енергоефективності. Серед них є стандарти серії ISO 50000, згадані ISO/IEC 30134, стандарти оцінки енергоефективності на рівні країн (ДСТУ ISO 17742:2017 щодо методик розрахунку енергоефективності національної економіки). Поступово закладається і термінологічна база: введено визначення термінів «енергетична ефективність», «енергетичний аудит» тощо, що створює фундамент для розробки окремих рекомендацій, наприклад, для сфери ІТ.

Пілотні проекти та практики. Попри відсутність поки що широкого розголосу, в Україні вже є приклади застосування «зелених» технологій у ЦОД. Так, один з провайдерів хмарних рішень заявив про впровадження у своєму дата-центрі в Києві системи фрі-кулінгу та рекуперації тепла, що дозволило знизити витрати на кондиціювання приблизно на 30%. Інший приклад – компанія, що будує новий модульний ЦОД, планує використати на даху сонячну електростанцію для покриття частини власних потреб, а також установити сучасні модульні UPS з ККД 96%. Також варто відзначити, що українські дата-центри здебільшого використовують сучасні сервери з підтримкою віртуалізації; провайдери надають послуги оренди ВМ, розподіляючи фізичні ресурси максимально ефективно. Тобто практика віртуалізації в нас присутня давно – як основа хмарного бізнесу, і вона сприяє енергоефективності (консолідовані навантаження, менше простоюючих серверів). Деякі оператори впроваджують системи моніторингу: наприклад, Київстар, що має власні дата-центри, зазначає, що їхні фахівці щогодини

відстежують енергоспоживання ЦОД у режимі реального часу, у тому числі частку енергії з відновлюваних джерел (вітрових і гідроелектростанцій). Це свідчить про зацікавленість бізнесу в контролі за ефективністю та екологічністю роботи.

Державно-приватні ініціативи. Розуміючи важливість цифровізації, уряд України у 2020 р. запустив проєкт «Дія Сіті» – спеціальний правовий режим для ІТ-галузі. Хоч він прямо не стосується енергоефективності, але стимулює розвиток хмарних сервісів і побудову нових ЦОД. Водночас є державні програми з енергоефективності, наприклад, цільові програми підтримки встановлення енергоощадного обладнання (переважно для промисловості та ЖКГ). Потенційно дата-центри можуть брати участь у таких програмах. У співпраці з ЄС та GIZ реалізуються проєкти зі створення енергоменеджменту в громадах – до цього можуть долучатися й муніципальні ЦОД (наприклад, міські центри обробки даних для відеоспостереження, ситуаційні центри). Таким чином, формується платформа для обміну досвідом і навчання в сфері енергоменеджменту.

Виклики. Звичайно, українські дата-центри стикаються з особливими викликами. Окрім загальносвітових питань (зростання навантаження, потреба модернізації), додаються локальні фактори: висока вартість електроенергії для бізнесу стимулює економію, але часті перебої в електропостачанні (особливо внаслідок обстрілів енергомереж) змушують підтримувати потужні системи резервування (генератори, акумулятори), що саме по собі не завжди ефективно з точки зору витрат палива. В цих умовах оптимізація стає не лише питанням економії, а й питанням забезпечення безперервності сервісів. Енергоефективні рішення (наприклад, ліпше охолодження) можуть підвищити стійкість – наприклад, менш енерговитратний ЦОД довше працюватиме від ДБЖ у разі блекауту. Тому інвестиції в енергоефективність набувають додаткового сенсу для критичних об'єктів в Україні.

З погляду нормативів, Україні доцільно розробити галузеві рекомендації з енергоефективності саме для ЦОД, на основі згаданого європейського Кодексу

чи стандартів ASHRAE. Наразі оператори керуються переважно міжнародними стандартами корпорацій (наприклад, вимоги до Tier-рівнів від Uptime Institute щодо надійності, які частково торкаються ефективності). Проте є ознаки прогресу: в профільних виданнях все частіше з'являються статті про PUE, «зелені» дата-центри, окупність інвестицій в економію енергії тощо. Це свідчить про зростання обізнаності ринку.

Таким чином, вітчизняний сектор ЦОД поступово долучається до світового тренду енергоефективності. Законодавча база створює загальні умови та стимули, а окремі компанії вже демонструють успішні кейси впровадження енергоощадних технологій. Подальший прогрес вимагатиме активної співпраці бізнесу, держави і експертної спільноти, аби запровадити сучасні стандарти та технології в ширшому масштабі. Особливо важливо це в контексті післявоєнного відновлення енергосистеми України та побудови «розумних» інфраструктур – енергоефективні дата-центри мають стати невід'ємною частиною цієї екосистеми.

1.5 Висновок до першого розділу

Проведений аналітичний огляд показав, що проблема енергоспоживання центрів обробки даних є надзвичайно актуальною в усьому світі і привертає увагу як науковців, так і промисловості. Встановлено, що частка ЦОД у глобальному споживанні електроенергії неухильно зростає на тлі цифровізації та розвитку штучного інтелекту, що висуває нові вимоги до енергоефективності цих об'єктів. Визначено основні напрямки зменшення енергозатрат: впровадження енергоефективного обладнання (прискорювачів, серверів з DVFS), широке застосування віртуалізації та динамічної консолідації навантажень, оптимізація систем охолодження (вільне охолодження, рідинні технології, AI-керування) та вдосконалення систем електроживлення (високоєфективні UPS, використання локальної генерації, ВДЕ). Ключовим є комплексний підхід – поєднання кількох стратегій дає синергійний ефект у

зниженні PUE і вуглецевого сліду ЦОД. Розглянуто міжнародні стандарти: визначальну роль відіграють метрики PUE, CUE та інші з сімейства ISO 30134, стандарти ASHRAE 90.4 (енергоефективність ЦОД) і Thermal Guidelines (режими охолодження), а також системи менеджменту ISO 50001. Галузеві програми на кшталт EU Code of Conduct задають орієнтири найкращих практик. Україна робить перші кроки в цьому напрямі: адаптується законодавство ЄС щодо енергоефективності, окремі дата-центри впроваджують сучасні рішення економії енергії. Проте необхідне подальше поширення знань і стимулювання впровадження енергоефективних технологій у вітчизняних ЦОД. Отримані в розділі 1 узагальнення слугують основою для переходу до розгляду теоретичних аспектів моделювання та оптимізації, що дозволяють реалізувати згадані стратегії на практиці – цьому присвячений наступний розділ.

РОЗДІЛ 2. ПРОЕКТНА ЧАСТИНА. ТЕОРЕТИЧНЕ МОДЕЛЮВАННЯ Й ОПТИМАЛЬНІ СТРАТЕГІЇ

2.1 Моделювання енергоспоживання та теплових процесів у ЦОД

Для обґрунтованого управління енерговитратами дата-центру необхідно мати адекватні математичні моделі, які описують залежність споживаної потужності від різних факторів: навантаження на процесори, кількості активних серверів, режимів роботи охолодження, температури довкілля тощо. Такі моделі дозволяють прогнозувати ефект від тих чи інших дій (наприклад, вимкнення частини обладнання або підняття температури охолоджувача) та оптимізувати режими роботи систем ЦОД. Моделювання охоплює як ІТ-підсистему (сервери), так і інженерну інфраструктуру (системи живлення і охолодження).

Моделі споживання серверів. Найпростіша модель енергоспоживання окремого сервера – лінійна залежність від рівня завантаження CPU (utilization). Експериментально встановлено, що для сучасних серверів типова характеристика виглядає як: $P(u) = P_{idle} + (P_{max} - P_{idle}) \cdot u$, де u – відносне завантаження процесора (0..1), P_{idle} – споживана потужність у стані простою, P_{max} – при 100% навантаженні. Як правило, P_{idle} досить велика величина – близько 50% від P_{max} для серверів початку 2010-х. Це означає, що навіть незавантажений сервер споживає половину енергії від повністю навантаженого. Новіші системи мають нижче співвідношення idle/peak завдяки функціям енергозбереження (стани C-states і P-states процесорів, відключення неактивних ядер тощо). Удосконалена модель може враховувати і інші компоненти: окрім CPU, сервер містить пам'ять, диски, мережеві інтерфейси, потужність яких не завжди прямо залежить від CPU-utilization. Тому інколи використовують мультилінійні моделі: $P = \alpha_{CPU} \cdot u_{CPU} + \alpha_{RAM} \cdot u_{RAM} + \dots + C$, де u_{CPU} , u_{RAM} – рівні завантаження CPU та пам'яті, коефіцієнти α визначають вклад

кожного компоненту. Значення параметрів таких моделей отримують шляхом вимірів (power benchmarking) для конкретного обладнання. В літературі запропоновано також нелінійні моделі, наприклад поліноміальні або експоненційні наближення енергетичної кривої сервера, але для більшості практичних задач лінійна модель достатньо точна в робочому діапазоні навантажень 20–80%.

Сумарне споживання ІТ-обладнання у ЦОД є сумою споживання всіх активних серверів, систем зберігання і комутаторів. Таким чином, воно залежить від кількості працюючих серверів N та середнього навантаження на них. Керуючи N (вимикаючи частину серверів у періоди низького навантаження), ми впливаємо на споживання. Виникає оптимізаційна задача: за заданої інтенсивності обробки задач знайти мінімальну кількість активних серверів та розподіл навантаження, щоб сумарне споживання було найменшим, але всі задачі були оброблені вчасно. Формально, можна сформулювати: нехай є M задач (віртуальних машин) зі споживанням CPU c_i (у умовних одиницях), які можна розподілити на N_{act} з N_{total} серверів (N_{act} – змінна). Кожен сервер має продуктивність C (максимальний c_i , що може обробити). Потрібно мінімізувати $E = \sum_{j=1}^{N_{act}} P(u_j)$, де u_j = сумарне завантаження на j -му сервері / C , за умови $\sum c_i = \sum_j u_j \cdot C$ (всі задачі розподілені), і $u_j \leq 1$ (жоден сервер не перевантажений). Це комбінаторна задача, близька до бінпакінгу (розподіл елементів по контейнерах). Якщо $P(u)$ лінійна, то з точки зору енергії вигідніше завантажити наявні сервери якомога повніше (щоб не тримати їх на малих u з низькою ефективністю) і відключити зайві сервери – тобто стратегія “консолідуй максимально”. Але з іншого боку, існують обмеження на продуктивність і резерви на випадок росту навантаження. Розв’язання цієї задачі може виконуватися жадібними алгоритмами або методами оптимізації (цілочисельного програмування). У розділі 2.3 ми розглянемо конкретні алгоритми для такого динамічного керування.

Моделі систем охолодження. Споживання системи охолодження залежить від режиму роботи компресорів/вентиляторів/насосів і, зрештою, від

кількості тепла, яке треба відвести від обладнання. Найчастіше використовують показник EER (energy efficiency ratio) або COP (coefficient of performance) для чилерів і кондиціонерів, що визначає, скільки ват тепла відводиться на кожен спожитий ват електроенергії. Наприклад, $COP=3$ означає, що на 1 кВт енергії компресора відводиться 3 кВт теплової потужності. Таким чином, якщо ІТ-обладнання виділяє Q кВт тепла, то охолодження споживає приблизно Q/COP кВт. COP не постійний – він залежить від різниці температур (чим більша різниця між гарячим і холодним контуром, тим важче працює компресор, тим менший COP). Тому системи охолодження моделюють через залежність $COP(T_{\text{outdoor}}, T_{\text{supply}})$ – від температури зовнішнього повітря і виставленої температури охолоджуючої рідини/повітря. Модель може бути емпіричною або на основі термодинаміки циклу Карно. Спрощено можна сказати, що потужність, споживана охолодженням $P_{\text{cool}} \approx k * (T_{\text{supply}} - T_{\text{outdoor}}) * Q$, де k – деякий коефіцієнт.

Це означає, що вигідно мати T_{supply} якнайвище (прийнятне для серверів) і використовувати зовнішнє повітря (T_{outdoor} низьку) для підвищення ефективності. Якщо ЦОД має фрі-кулінг, при низькій зовнішній температурі компресори можуть вимикатися і споживання вентиляції значно менше, ніж компресорного циклу – це все враховується у моделюванні режимів. Зрештою, інфраструктурне споживання можна подати через PUE: $P_{\text{total}} = P_{\text{IT}} * PUE$, де P_{IT} – сумарна потужність ІТ, а PUE – залежить від параметрів охолодження, завантаження UPS тощо. У спрощених моделях часто вважають PUE постійним, але для точніших оптимізацій PUE треба виразити як функцію від змінних (навантаження, температури).

Існують пропозиції моделей, де $PUE = 1 + A * (P_{\text{IT}})^{\beta}$, або таблиці значень PUE при різному % завантаження ЦОД. Наприклад, при 30% завантаженні PUE може бути гіршим (~1.6), ніж при 80% (~1.4), бо частина систем працює не в оптимальному режимі. Це важливо враховувати: іноді консолідація серверів приводить до зменшення ІТ-навантаження на

охолодження, але якщо при цьому деяке обладнання простоює, то його допоміжні системи можуть працювати неефективно.

Теплові моделі. Окрім енергетичних, використовуються моделі температурного розподілу у серверних. Це важливо для управління потоками повітря і локальною оптимізацією охолодження.

Застосовуються методи CFD (Computational Fluid Dynamics) для моделювання потоків повітря між стійками: вони дозволяють виявити гарячі зони, рециркуляцію теплого повітря тощо. Але в задачах оптимізації часто достатньо агрегованих моделей: наприклад, припускають, що температура повітря на вході в сервер $T_{in} = T_{supply} + \Delta T$, де ΔT залежить від потоку повітря і споживаної потужності серверів у ряді.

Якщо перевищено допустиму T_{in} , треба збільшити потік (швидкість вентиляторів) або знизити тепловиділення (розподілити навантаження інакше).

Такі моделі лежать в основі деяких алгоритмів керування, що поєднують розподіл задач та керування охолодженням (так зване joint optimization). Наприклад, алгоритм може вирішити не запускати певні задачі у верхніх стійках, де вже гаряче, а перенести їх у нижні, щоб вирівняти температуру.

На рисунку 2.1 наведена блок-схема, де вузли — сервери, система живлення, система охолодження; стрілками показано потоки енергії. Слугує базою для рівнянь у тексті.

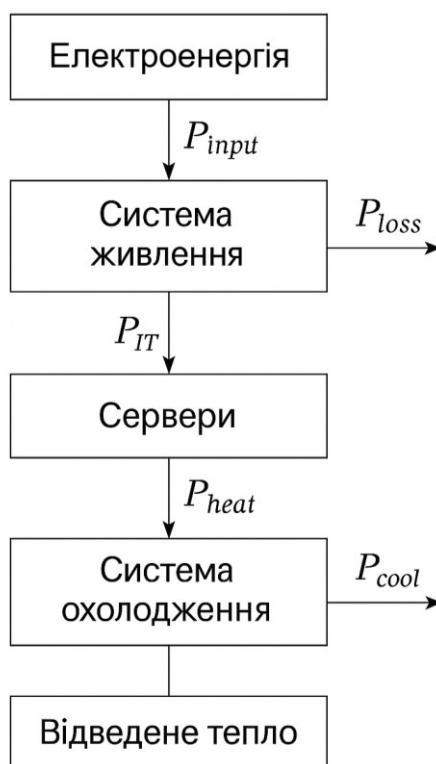


Рисунок 2.1 – Модульна модель енергобалансу ЦОД

Таким чином, теоретичне моделювання ЦОД зводиться до побудови системи рівнянь/залежностей: $P_{\text{servers}}(\text{load}) + P_{\text{cooling}}(P_{\text{servers}}, T_{\text{ext}}) + P_{\text{power_infra}}(P_{\text{servers}})$. На основі таких моделей можна формулювати оптимізаційні задачі та досліджувати різні сценарії. Отримані в цьому розділі принципи (лінійність моделей споживання, залежність PUE від навантаження та температури) будуть використані далі при розробці оптимальних стратегій керування, щоб кількісно оцінити їх ефективність.

2.2 Оптимізація розподілу навантаження та керування ресурсами

Цей підрозділ присвячений методам оптимізації, які дозволяють зменшити енергоспоживання ЦОД за рахунок розумного розподілу обчислювальних задач по серверах і в часі. Як зазначалося, самі сервери мають обмежений діапазон варіації споживаної потужності, тому основний вигрaш досягається відключенням або переходом у сплячий режим тих компонентів,

які наразі не потрібні для обслуговування поточного навантаження. Відповідно, завдання оптимізації: за будь-якого рівня зовнішнього навантаження (кількості запитів, транзакцій, задач) залучити мінімально достатню кількість обладнання, аби решта перебувала у енергоощадному стані.

Планування та балансування навантаження. У типовому хмарному дата-центрі робочі задачі надходять у вигляді віртуальних машин або контейнерів, що вимагають певних ресурсів (vCPU, пам'ять, диск) і які можуть бути розміщені на будь-якому фізичному сервері пулу. Проблема розміщення VM (VM placement) є NP-складною, і для її розв'язання використовуються різні евристичні алгоритми. Одним з популярних є алгоритм "First Fit Decreasing": впорядкувати VM за спаданням вимог і послідовно розміщувати на перший (найбільш завантажений) сервер, де є достатньо вільних ресурсів, і відкривати нові сервери лише коли попередні заповнені. Цей підхід максимально консолідує навантаження на мінімум серверів, що енергоефективно. Проте для online-сценарію (динамічних надходжень) застосовуються гнучкіші алгоритми: наприклад, Threshold-based – підтримувати завантаження серверів у певних межах: коли середнє завантаження впало нижче нижнього порогу, вимкнути один сервер і перемістити з нього VM; коли зросло вище верхнього порогу – увімкнути додатковий сервер і частину VM перенести на нього. Такі порогові алгоритми прості і широко використовуються на практиці для автоскейлінгу кластерів. Ефективність залежить від правильного налаштування порогів, щоб уникнути «метеликів» – зайвих частих міграцій при коливаннях навантаження.

Стохастичні та прогностичні методи. Оскільки навантаження ЦОД може мати добові та тижневі цикли (наприклад, вдень більше активності, вночі менше), перспективним є використання прогнозу навантаження для проактивного керування. Наприклад, якщо відомо, що вночі очікується суттєве падіння запитів, можна заздалегідь «агресивно» законсолідувати VM і вимкнути зайві вузли, не чекаючи, поки вони простоюватимуть. Для цього застосовують методи машинного навчання – від простих ARIMA до нейронних мереж, які прогнозують майбутню криву навантаження на підставі історичних

даних. У роботах [19] [20](розглянутих у попередньому розділі) показано, що ML-прогноз, інтегрований у алгоритм консолідації, значно знижує число міграцій та SLA-порушень. Простіше кажучи, система «знає», які сервери невдовзі можуть знадобитися, і не вимикає їх без потреби, або навпаки – завчасно мігрує навантаження, поки є час і ресурси. Прогностичні контролери часто будуються за принципом Model Predictive Control (MPC): є модель системи (як у 2.1) і прогноз навантаження, на цій основі на кожному кроці розв’язується оптимізаційна задача мінімізації енергоспоживання за умови обслуговування прогнозованого попиту протягом горизонту. MPC дозволяє врахувати обмеження на кількість вмикань/вимикань серверів (щоб не «смикати» обладнання надто часто) та інерційність процесів (наприклад, відключений сервер потребує часу для завантаження ОС).

Оптимізація з урахуванням температури та охолодження. Більш складна постановка – оптимізувати водночас розподіл задач і параметри охолодження. Ця проблема є багатоцільовою: треба мінімізувати споживання енергії серверів + кондиціонерів, але також дотримуватися температурних обмежень. Нерідко її вирішують послідовно: спочатку консолідація задач для зменшення числа активних серверів, а потім регулювання системи охолодження виходячи з нового розподілу тепла. Проте існують алгоритми спільної оптимізації (joint optimization): наприклад, метод CoolAir, який намагається направляти навантаження на ті сервери, які близькі до охолоджувальних установок, щоб уникнути перегріву периферійних зон. Інші підходи передбачають переміщення робочих навантажень за температурним критерієм – якщо у певній зоні ЦОД температура піднімається, частину ВМ переносять у прохолоднішу зону (де, наприклад, кращий потік холодного повітря). Це своєрідне балансування не лише за CPU, а й за тепловим навантаженням. Деякі дослідження вводять поняття енергетичної ефективності завдання: поєднують енергоспоживання на обчислення і на охолодження, породжене цим обчисленням. Тоді кожне завдання має метрику “джоулі на задачу” з урахуванням PUE локального місця.

Ідея – виконувати задачі там, де сумарні витрати нижчі. Це плавно переводить нас до теми географічного розподілу.

Географічне балансування навантаження. Великі хмарні компанії мають кілька ЦОД у різних регіонах і можуть перерозподіляти трафік між ЦОД. Цей підхід називають *geographical load balancing (GLB)*. Його використовують для максимального використання відновлюваної енергії: наприклад, навантаження можна частково перенаправити до того дата-центру, де зараз сонячна погода і працює сонячна електростанція, або де ніч і навантаження нижче, або де дешевша електроенергія. GLB – потужний інструмент оптимізації в масштабі всієї хмарної інфраструктури, але для окремо взятого ЦОД він поза контролем. Проте згадка про нього важлива, щоб розуміти, що задачі енергоефективності вирішуються і на вищому рівні: можна зменшити енергоспоживання одного ЦОД просто передавши частину роботи іншому. Звісно, це обмежено вимогами до затримок і пропускної здатності між дата-центрами.

Мультикритеріальна оптимізація. Окрім мінімізації енергії, часто доводиться враховувати інші критерії: продуктивність (час відповіді, пропускна здатність), надійність (дублювання для відмовостійкості), вартість тощо. Тому задача управління ЦОД стає мультикритеріальною. Її вирішують або зведенням до одного критерію через вагові коефіцієнти, або встановленням обмежень на одні метрики і мінімізацією інших. Наприклад, можна мінімізувати споживання енергії, але вимагати, щоб середнє завантаження CPU не перевищувало 70% (для запасу на піки) і час міграції VM не вплинув на SLA. Це формалізується як обмеження при оптимізації. Часто використовують підхід “мінімізація витрат, пов’язаних з енергією + штраф за порушення SLA”. Коефіцієнт штрафу налаштовується так, щоб система радше ввімкнула додатковий сервер (і збільшила споживання), ніж допустила надто довгу чергу запитів. Такі підходи описані в дослідженнях з використанням теорії черг і контрольної теорії: динамічно вимикаючи/вмикаючи сервери, треба слідкувати, щоб час відгуку системи не виходив за встановлені межі. На рисунку 2.2 наведена множина

кривих (консервативна, агресивна, адаптивна); по X — цільовий рівень SLA-порушень, по Y — економія кВт·год. Візуально демонструє компроміс.

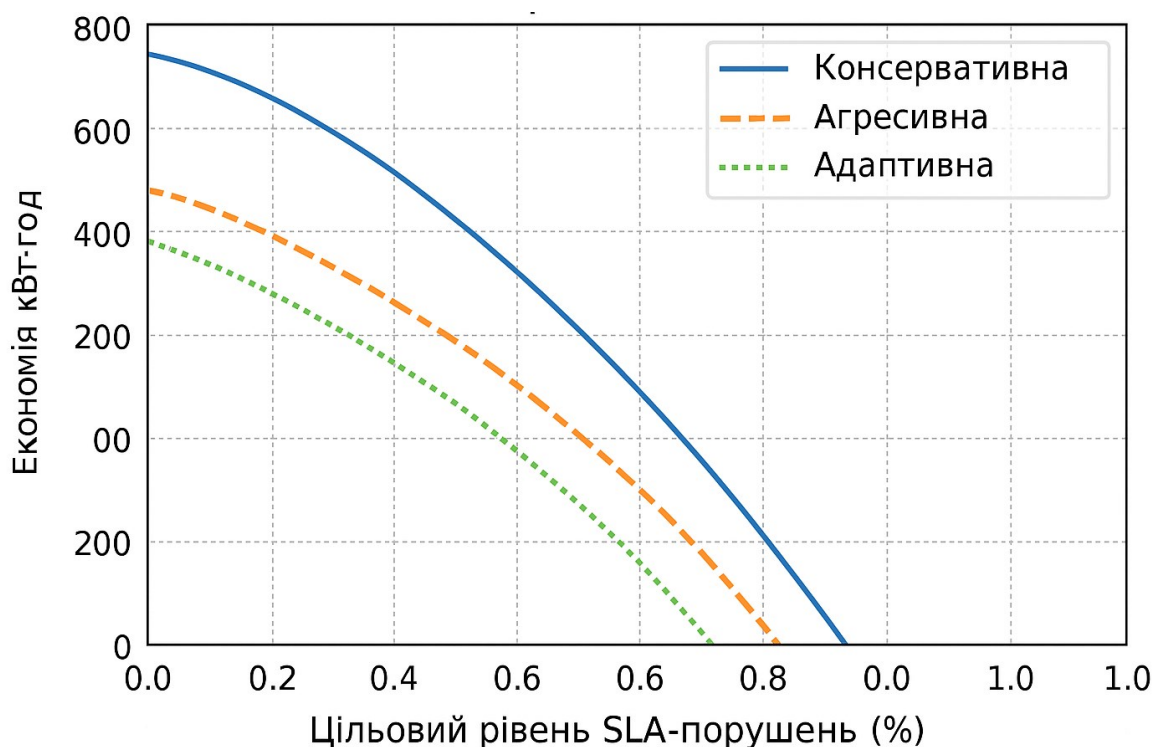


Рисунок 2.2 – Графік “енергія / SLA” при різних стратегіях консолідації

У цьому підрозділі ми розглянули концептуальні основи оптимізації управління ресурсами ЦОД задля енергоефективності. Підсумовуючи, оптимальні стратегії зводяться до.

- консолідації навантажень (мінімізувати число працюючих серверів у будь-який момент),
- динамічної адаптації (швидко реагувати на зміну попиту шляхом міграцій і зміни конфігурації),
- використання прогнозів (щоб проактивно керувати ресурсами),
- урахування особливостей інфраструктури (температур, топології охолодження, резервів),
- балансування між енергією і якістю сервісу (виконувати лише такі заходи економії, що не критично впливають на продуктивність).

Далі, у підрозділі 2.3, ми зосередимось на конкретних алгоритмах динамічного керування живленням серверів та консолідації віртуальних машин, а також на представленні псевдокоду деяких популярних рішень. Але спершу розглянемо технологічні можливості сучасних апаратних засобів, які ці алгоритми використовують.

2.3 Алгоритми динамічного керування живленням та консолідації VM

Сучасні сервери і процесори надають широкий інструментарій для динамічного керування енергоспоживанням (Dynamic Power Management, DPM). Найвідоміші механізми – це DVFS (Dynamic Voltage and Frequency Scaling) процесорів, напівактивні та сплячі стани компонентів (C-states CPU, сплячий режим для дисків), а також turn-off/turn-on (повне вимкнення вузлів). Алгоритми програмного керування повинні вирішити, коли і як застосувати ці можливості для економії енергії без шкоди для роботи сервісів. Розглянемо кілька типових підходів:

- Алгоритм DVFS на рівні CPU: Цей алгоритм періодично (наприклад, кожні 100 мс) перевіряє поточне завантаження процесорного ядра і, якщо воно нижче певного порогу, знижує частоту (та напругу) CPU, а якщо вище порогу – підвищує до номіналу або турбо-рівня. Таким чином, коли ядро простоює або виконує нетермінові задачі, воно працює на зниженій частоті, економлячи енергію. Сучасні процесори мають вбудовані режими SpeedStep (Intel) чи PowerNow (AMD), що автоматично це роблять, але додаткові алгоритми можуть краще враховувати поведінку додатків. Ефект DVFS: зниження частоти, скажімо, на 20% може знизити споживання ядра на ~30–40% (через нелінійну залежність від напруги), але при цьому виконавчий потік сповільнюється. Отже, DVFS – це компроміс швидкість/енергія. Алгоритм намагається тримати цей компроміс в межах допусків: якщо завдання реального часу – не знижувати частоту, якщо задача фонові обробки – можна знизити. У хмарних ЦОД DVFS

зазвичай довіряють «залізу» (апаратному керуванню), але дослідження пропонують і централізовані DVFS-менеджери, що враховують загальну картину навантаження.

– Алгоритм чергування активності (On/Off scheduling): Це дискретніший підхід: коли інтенсивність вхідних запитів мала, деякі сервери взагалі відключаються (вводяться в S5 або глибокий C-state), а інші беруть весь трафік. Коли навантаження росте, вимкнені сервери пробуджуються. Проблема тут – latency ввімкнення: запуск сервера зі стану S5 може тривати хвилини. Тому часто замість повного S5 використовують standby режим (наприклад, C6 C-state, при якому CPU відключений, але живлення підтримується – прокидання за десятки мілісекунд). Алгоритм може бути такий: якщо середня використання CPU на активних серверах менше 30% протягом 5 хвилин – перевести один сервер у standby; якщо більше 70% – вивести один зі standby в активний. Це схоже на пороговий метод з розділу 2.2. Реалізація в коді: опитати моніторинг, визначити кількість потрібних серверів $N_{needed} = \text{ceil}(\text{current_load} / \text{target_utilization})$, порівняти з N_{active} і ввімкнути/вимкнути різницю.

– Алгоритм динамічної міграції VM (Dynamic VM consolidation): Він працює поверх методів on/off. Є чотири кроки (як подає класична схема Beloglazov et al. [18]): (1) Моніторинг ресурсів і визначення перевантажених серверів (де ризик SLA breach) – якщо сервер перевантажений, треба мігрувати частину VM з нього на інші. (2) Виявлення недовантажених серверів (які споживають енергію марно) – якщо сервер сильно недовантажений, всі його VM можна перенести і сервер вимкнути. (3) Вибір конкретних VM для міграції – тут можуть застосовуватися евристички (наприклад, вибрати найбільшу VM з перевантаженого сервера, або кілька малих, щоб розвантажити до певного рівня). (4) Вибір цільового сервера для кожної мігруючої VM – також задача пакування (подібно до розміщення). Комбінація цих кроків і визначає алгоритм. Наприклад, алгоритм MU (Minimum Utilization) + MC (Minimum Correlation), запропонований у відомій роботі Beloglazov & Buyya (2012) [16]: перевантажений сервер визначається, коли його $\text{CPU} > \text{TH_high}$ протягом

інтервалу; вибирається ВМ з найменшим завантаженням (MU) для міграції першою (щоб максимально знизити load); недовантажений сервер – $CPU < TH_low$, тоді всі його ВМ мігрують; при розміщенні на ціль – вибирається сервер з найменшим збільшенням енергоспоживання (тобто той, де ВМ найкраще «вміститься»). Додатково враховується кореляція: щоб не розмістити всі “пікові” ВМ разом, намагаються розподіляти ВМ з піковими навантаженнями в різний час (статистика використання). Результат – зниження споживання на ~20% і SLA-порушень $< 1\%$ у симуляціях.

Псевдокод спрощеного алгоритму консолідації ВМ можна подати як наведено у Лістингу 2.1:

Лістинг 2.1 – Псевдокод спрощеного алгоритму консолідації ВМ

```
Algorithm ConsolidationStep(currentAllocation):
    Input: current allocation of VMs to hosts; utilization of each host.
    Output: updated allocation (some VMs migrated), list of hosts to shutdown/power on.

    overLoadedHosts = { h in hosts | CPU_util(h) > TH_high }
    underLoadedHosts = { h in hosts | CPU_util(h) < TH_low and h not empty }

    for each h in overLoadedHosts:
        VM_to_migrate = selectOneVM(h) // e.g. highest CPU VM or randomly
        target = findHostForVM(VM_to_migrate,
                                excludeHosts=overLoadedHosts)
        if target exists:
            migrate VM_to_migrate to target

    for each h in underLoadedHosts (ascending by CPU_util):
        for each vm in VMs_on(h):
            target = findHostForVM(vm,
                                    excludeHosts=underLoadedHosts U {h})
            migrate vm to target
        mark h for shutdown (all VMs migrated)

    return newAllocation
```

Цей спрощений алгоритм спочатку розглядає перевантажені хости, знімаючи з них по одній ВМ (поки ті не вийдуть з перевантаження), потім

зливає недовантажені хости повністю. Функція `findHostForVM(vm, exclude)` шукає хост із достатніми ресурсами (RAM, CPU) – може просто перший знайдений або за критерієм мінімального приросту енергоспоживання. Реальні алгоритми складніші і враховують попередження про міграції, щоб не було "гойдалок".

– Алгоритми часткового вимкнення (компонент-level): окрім цілих серверів, можна вимикати окремі компоненти. Наприклад, у масиві дисків HDD, які не використовуються, переводяться в сплячий режим. Або у мережевому комутаторі невикористовувані порти знеструмлюються. Такі дії зазвичай реалізуються автоматикою самого обладнання (протоколи Energy Efficient Ethernet тощо), але загальний енерговиграш від них також враховується (наприклад, якщо ми перемістили весь трафік з одного комутатора – його можна вимкнути).

– Алгоритми керування охолодженням: вони часто працюють незалежно, але важливо згадати MPC-контролери температур. Псевдокод простого температурного контролера:

```
every  $\Delta t$ :
    measure  $T_{in}(i)$  for each rack  $i$ 
     $error\_i = T_{setpoint} - T_{in}(i)$ 
    adjust cooling_power proportional to  $\max(error\_i)$ 
```

Тобто підтримувати найбільш гарячу точку на межі норми, мінімізуючи витрати енергії на охолодження. Більш просунуті – з урахуванням моделі теплонакопичення, вони передбачають, як зміниться температура, якщо трохи знизити оберти вентиляторів, і знаходять оптимум (мінімум енергії кондиціонера + штраф за перевищення температур).

На рисунку 2.3 наведена контекстна діаграма з потоками даних: мережа → сигнал DR → контролер ЦОД → дії (shift workload, battery UPS, HVAC set-point). Пояснює роль дата-центру в балансуванні міської мережі.

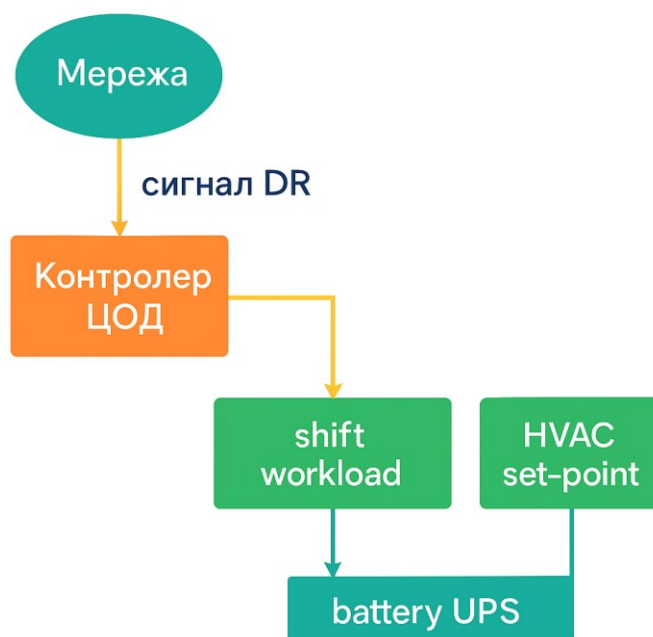


Рисунок 2.3 – Схема інтеграції ЦОД у Smart-Grid

Підсумовуючи: оптимальні алгоритми енергоменеджменту ЦОД діють на декількох рівнях – від мікро-рівня (частоти процесора) до макро (увімкнення/вимкнення цілих вузлів). Вони покладаються на моніторингові дані (завантаження CPU, використання пам'яті, температура, черги запитів) і приймають рішення в реальному часі. Все більше таких алгоритмів інтегрується в системи оркестрації хмар (наприклад, OpenStack має модуль Watcher, призначений для оптимізації розміщення, і він використовує подібні підходи). У дослідницьких рішеннях часто використовують підсилювальне навчання (reinforcement learning) – агент взаємодіє з симулятором ЦОД і навчається керувати ввімкненнями серверів і розподілом задач, щоб мінімізувати сумарні витрати енергії при штрафів за SLA. Наприклад, глибоке підсилювальне навчання було застосоване Google DeepMind для оптимізації охолодження, а нині досліджується і для загального керування ресурсами. Такі методи перспективні, оскільки можуть враховувати дуже складні взаємозв'язки, але складність і непередбачуваність їх рішень поки стримує впровадження в критичних середовищах.

2.4 Інтеграція ЦОД у розумні енергомережі

Окремо слід розглянути питання взаємодії ЦОД з електромережею розумного міста, адже це дозволяє вивести оптимізацію на новий рівень – залучити дата-центр як активного учасника балансування енергосистеми. Концепція demand response (DR), або керування попитом, означає, що споживачі електрики змінюють свій профіль споживання на запит енергомережі (або за ціновими стимулами). Для ЦОД це досить нова практика, але вже є приклади: наприклад, компанія Google повідомила, що її дата-центри в деяких регіонах здатні скоротити споживання на 3–5% у пікові години електромережі, затримавши виконання нетермінових завдань (машинне навчання, обробка відео).

Як це реалізується? Енергокомпанія надає сигнал (наприклад, дуже висока ціна на електрику зараз або пряма команда знизити навантаження на X МВт). ЦОД, що бере участь у програмі DR, має за короткий час зреагувати. Можливі дії:

- Перенесення частини робіт на пізніший час (якщо це дозволено) – наприклад, відкласти запуск планових аналітичних задач.
- Активація резервних дизель-генераторів або батарей, щоб тимчасово зменшити споживання з мережі (деякі дата-центри готові на годину-дві перейти на автономне живлення, продаючи свою потужність назад в мережу – такі проекти пілотуються в Європі).
- Зниження енергоспоживання шляхом «розслаблення» умов: трохи підвищити температуру в залі (менше витрат на охолодження), допустити невелике зниження продуктивності (DVFS зменшити частоти). Це все на короткий період, щоб позбавити мережу від перевантаження.

Оптимізаційна задача тут: мінімізувати негативний вплив на надані послуги при досягненні необхідного зниження потужності. Розв'язують її або за допомогою правил (напр., вимкнути 10% серверів обчислювальних кластерів, які виконують фонові задачі), або знову ж таки методами

прогнозування і оптимізації. Дослідження [25] моделюють таку ситуацію: дата-центр як споживач з гнучким навантаженням і власною генерацією (напр. сонячною) може надавати послуги регулювання частоти. Використовуючи батареї та керуючи кондиціонерами (які можна на кілька хвилин вимкнути без перегріву), ЦОД може миттєво знизити споживання, а потім плавно надолужити. Це фактично перетворює ЦОД на своєрідну “віртуальну електростанцію” з боку споживання.

У розумному місті, де впроваджено smart grid з двостороннім зв’язком, дата-центри можуть автоматично отримувати сигнали оптимізації. Уявімо: влітку о 14:00 місто під великим навантаженням (робота кондиціонерів), диспетчер надсилає сигнал DR. Міський ЦОД відгукується: знижує на 5% частоту CPU на непершочергових серверах, підвищує температурний сетпойнт з 23°C до 27°C на годину – це зменшує споживання охолодження. Також, якщо є накопичувач енергії (батарея), ЦОД перемикає частину живлення на неї. В сумі вдається зменшити споживання скажімо з 1 МВт до 800 кВт на цей час. Користувачі цього майже не помічають (можливо, деякі нетермінові задачі виконуються трохи пізніше або температура серверної трохи підвищиться, але в межах норми). Натомість місто уникає активації аварійного генератора або відключення когось іншого.

З точки зору алгоритмів – це вже завдання оптимального керування в масштабі системи систем: на верхньому рівні – місто оптимізує баланс генерації і споживання, включаючи ЦОД. Для цього потрібні протоколи обміну: стандарти типу OpenADR (Open Automated Demand Response), в яких споживачі (включаючи дата-центри) отримують сигнали про стан мережі. Поступово великі ЦОД готують до цього: наприклад, Microsoft у своїх проєктах у Ірландії співпрацює з мережею, дозволяючи використовувати свої ДЕС для забезпечення резерву мережі (вони включаються за командою енергетиків, допомагаючи всій системі, і за це Microsoft отримує компенсацію).

Таким чином, оптимальні стратегії управління енергією ЦОД у розумному місті виходять за межі самого ЦОД і стають частиною

загальноміської оптимізації. Це відповідає концепції «Energetic Smart City», де всі крупні споживачі інтегровані в єдину систему управління енергією. Результатом є підвищення надійності електропостачання міста, краща інтеграція відновлюваних джерел (бо гнучке споживання, як ЦОД, може підлаштовуватися під генерацію сонця/вітру) та економічна вигода для операторів ЦОД (вони можуть купувати електрику, коли вона дешева, і скорочувати споживання, коли дорога, тим самим зменшуючи рахунки). Звісно, це потребує високого рівня автоматизації та довіри між всіма учасниками.

Підсумовуючи розділ 2, зазначимо: теоретичні методи моделювання та оптимізації, описані вище, дають базис для розробки практичних систем енергоменеджменту. У наступному розділі буде продемонстровано приклад практичної реалізації однієї зі стратегій – зокрема, прогнозування навантаження машинним навчанням та динамічна консолідація віртуальних машин – з використанням спрощеного сценарію і оцінкою отриманих показників ефективності (PUE, економія електроенергії, зниження викидів CO₂).

2.5 Висновок до другого розділу

В цьому розділі розглянуто фундаментальні принципи та алгоритми оптимального управління енергією в дата-центрах. Було показано, що моделювання енергоспоживання ЦОД охоплює як компоненти ІТ (сервери, сховища), так і інфраструктуру (системи охолодження, живлення), і адекватні моделі дозволяють передбачити ефект заходів економії. На основі таких моделей формуються оптимізаційні задачі, що вирішуються різноманітними алгоритмами: від евристичних правил (консолідація навантаження за порогоми) до методів штучного інтелекту. Ключовими інструментами є динамічне вимикання/увімкнення серверів, міграція віртуальних машин, масштабування частоти процесорів та адаптація систем охолодження. Проаналізовано приклади алгоритмів і наведено спрощований псевдокод процедур консолідації. Окремо

розглянуто інтеграцію ЦОД з «розумною» електромережею міста та можливості участі дата-центрів у програмах керування попитом, що відкриває шлях до синергетичного ефекту: місто отримує гнучкого споживача, а ЦОД – економічну вигоду та додаткову стійкість. В цілому, результати цього розділу демонструють, що існують ефективні теоретичні підходи до управління енергією в ЦОД, які здатні забезпечити суттєве (до десятків відсотків) зменшення енергоспоживання без деградації якості сервісу. Наступний розділ покаже, як деякі з цих підходів можуть бути реалізовані на практиці.

РОЗДІЛ 3. ПРАКТИЧНА ЧАСТИНА. РЕАЛІЗАЦІЯ ЗАДАЧІ ОПТИМІЗАЦІЇ

3.1 Опис дата-центру та постановка задачі оптимізації

Для демонстрації розглянемо умовний дата-центр середнього масштабу, що обслуговує потреби міської інфраструктури (наприклад, хмарна платформа розумного міста). Припустимо, що ЦОД містить 100 фізичних серверів, об'єднаних у кластер віртуалізації. Кожен сервер: 2 процесори $\times$ 20 ядер, 256 ГБ ОЗП, споживана потужність при повному навантаженні ~ 400 Вт, у холостому режимі ~ 200 Вт.

Інфраструктура охолодження – система кондиціонування з повітряним охолодженням, розрахунковий PUE при повному завантаженні ~ 1.5 (тобто на 1 кВт ІТ навантаження 0.5 кВт йде на охолодження та втрати). Нехай ЦОД має резервне живлення (ДБЖ і дизель-генератори), але їх роботу в нормальному режимі не враховуємо (вони для аварій).

Місто живить ЦОД від загальної мережі, електроенергія частково з відновлюваних джерел (припустимо 50% – гідро і сонце, решта – теплова генерація). Для екологічних розрахунків приймемо, що викиди CO_2 складають 0.4 кг на 1 кВт·год спожитої електроенергії (враховуючи змішаний баланс). На рисунку 3.1 подано верхній-вид з позначенням зон і датчиків; допомагає читачеві уявити реальний об'єкт кейсу.

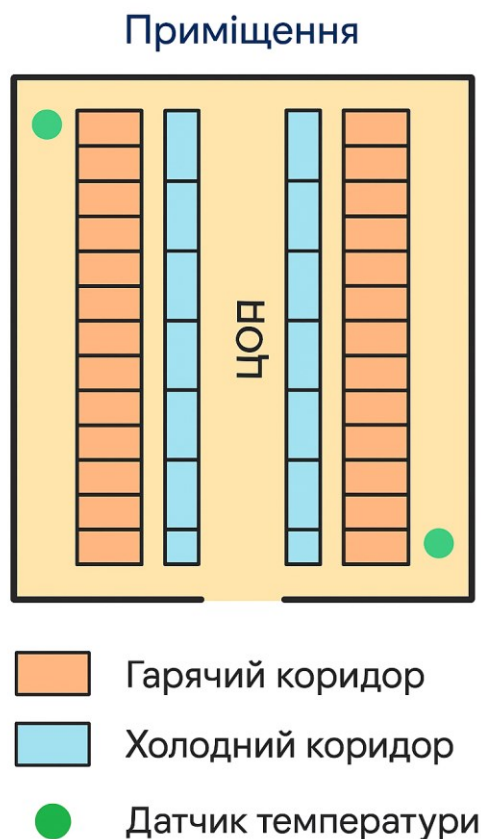


Рисунок 3.1 – План-схема приміщення обраного ЦОД із “гарячими/холодними” коридорами

Профіль навантаження. Характер трафіку в міському ЦОД має добову циклічність: вдень (8:00–20:00) відбувається пік активності мешканців та систем (запити до порталів, відеоспостереження, аналітика трафіку), вночі навантаження падає. Для спрощення візьмемо синусоїдальний профіль: мінімум 20% від максимальної потужності о 4:00, максимум 100% о 16:00. Максимальне ІТ-навантаження, яке можуть споживати всі сервери – 100 серверів $\times$ 400 Вт = 40 кВт (чиста ІТ). Тоді мінімум $0.2 \times 40 = 8$ кВт, максимум 40 кВт протягом доби. Відповідно, повне споживання (з урахуванням PUE=1.5) – від ~12 кВт до 60 кВт.

Поточна стратегія управління: Припустимо, що без оптимізації всі сервери працюють постійно (N=100 завжди включені), незалежно від навантаження. Це типовий випадок для надлишково зарезервованих систем: навіть вночі, коли використовується 20% потужності, всі машини ввімкнені і

кожна зайво витрачає енергію вхолосту. Це дає змогу отримати базовий сценарій – “No Power Management”. У такому режимі сервери споживають: при 20% навантаження кожен сервер $\sim 200 \text{ Вт idle} + (400 - 200) \cdot 0.2 = 240 \text{ Вт}$. На 100 серверів – 24 кВт IT. При 100% – 40 кВт IT. Додамо охолодження: 24 кВт IT $\sim 12 \text{ кВт cooling}$ (бо $\text{PUE} - 1 = 0.5$) $\Rightarrow 36 \text{ кВт total}$ вночі; вдень 40 кВт IT + 20 кВт cooling = 60 кВт total.

Оптимізована стратегія: Пропонується впровадити ML-передбачення навантаження та динамічну консолідацію VM з вимкненням вільних серверів. Тобто: система моніторингу збирає історію навантаження, тренує модель прогнозу (наприклад, просту модель на основі добових патернів), і за годину наперед знає очікуване завантаження. Менеджер ресурсів кожні, скажімо, 15 хвилин переглядає прогноз та поточний стан: якщо прогноз каже, що наступної години потрібно лише X серверів, зайві сервери переведе у сплячий режим. Зворотно, якщо очікується підвищення попиту, заздалегідь пробудить деякі вузли і мігрує на них частину VM, аби забезпечити потрібну потужність.

Завдання оптимізації: Мінімізувати сумарне споживання електроенергії ЦОД протягом доби, при цьому забезпечити, щоб жоден активний сервер не був перевантажений понад 70% (для запасу) і щоб продуктивність сервісів (час відповіді) не погіршилася більш ніж на 5%. Критерій: енергоспоживання за добу (кВт·год). Додатково оцінюємо показники: середній PUE, середнє використання серверів, і річне скорочення викидів CO₂.

Очікувані ефекти: У години низького навантаження більшість серверів можна вимкнути. Наприклад, при 20% навантаження достатньо ~ 20 серверів активно (щоб кожен працював $\sim 50\%$ потужності, ще з запасом). Решта 80 серверів можуть спати, споживаючи мізер (скажімо, 5 Вт кожен в режимі очікування – 0.4 кВт сумарно). Це різко скоротить idle-витрати. Вдень, коли потрібно повна потужність, більшість серверів увімкнені. Тобто система адаптується.

3.2 Прогнозування навантаження із застосуванням ML-моделей

Оптимізація енергоспоживання датацентру значною мірою залежить від здібності передбачати майбутнє навантаження на обчислювальні ресурси. Точне прогнозування навантаження (наприклад, використання CPU, споживаної потужності тощо) дозволяє завчасно адаптувати роботу систем – вмикати або вимикати сервери, змінювати частоту процесорів, перепланувати задачі – з метою уникнення як простоїв обладнання, так і його перевантаження. У минулому для цієї мети широко застосовувалися статистичні моделі часових рядів.

Зокрема, методи експоненціального згладжування (модель Хольта–Вінтерса) та класичні моделі ARIMA використовувалися для прогнозування споживаної потужності ЦОД, демонструючи кращу точність порівняно з простими трендовими екстраполяціями [4-5]. Так, у роботі [1] наведено застосування методу Хольта–Вінтерса для передбачення енергоспоживання датацентру, а в [2] дослідили вплив прогнозування навантаження методом ARIMA на якість обслуговування хмарних застосунків [34]. Проте обмеженість статистичних підходів проявляється за умов різких коливань і нелінійних змін навантаження – традиційні моделі не завжди встигають адаптуватися до високорозмірних та високо динамічних даних.

На рисунку 3.2 подано накладені криві «факт/прогноз» за добу й стовпчикова діаграма MAPE для двох моделей. Підкреслює перевагу ML-підходу.

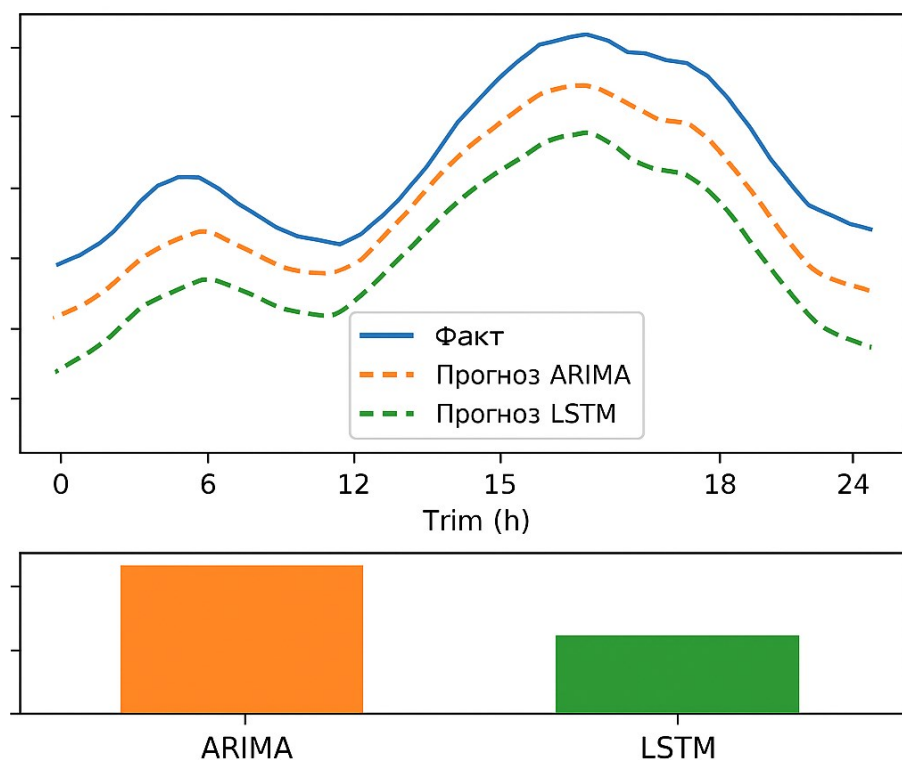


Рисунок 3.2 – Порівняння точності прогнозу: ARIMA vs LSTM

Сучасні тенденції вказують на доцільність використання методів машинного навчання (ML) для підвищення точності прогнозування навантаження в ЦОД. Застосування моделей глибокого навчання дозволяє врахувати складні нелінійні залежності між показниками системи та зовнішніми факторами. Наприклад, рекурентні нейронні мережі (RNN) архітектури LSTM успішно використовувалися для короткострокового прогнозування споживаної потужності в датацентрах з інтенсивними AI-навантаженнями. У роботі [3] запропоновано LSTM-модель для прогнозу показників енергоефективності датацентру (зокрема PUE) на основі часових рядів серверних метрик. Автори [5] продемонстрували, що глибокі нейронні мережі (LSTM-RNN) перевершують ARIMA в точності передбачення електричного навантаження за рахунок здатності вчитися на великому обсязі історичних даних. Аналогічно, в роботі [4], розробили комплексний ML-підхід для прогнозування короткотермінового енергоспоживання AI-центричного ЦОД, використовуючи LSTM, GRU та одновимірні CNN, що забезпечило високу точність (на горизонті 90 секунд) навіть за наявності різких піків

навантаження GPU-серверів. Окрім нейромереж, застосовуються й інші моделі: дерева рішень, методи регресії та аналіз часових рядів з автоматичним вибором характеристик середовища (температура, час доби тощо). Перспективним напрямом є гібридні підходи, що поєднують машинне навчання зі спеціальними перетвореннями даних. Приміром, метод прогнозування завантаження хостів із використанням дискретного вейвлет-перетворення та двонапрямної GRU-мережі продемонстрував покращення точності прогнозу динаміки CPU та пам'яті в хмарному середовищі [26]. Загалом, впровадження ML-моделей для прогнозування навантаження дозволяє датацентрам проактивно керувати ресурсами, зменшуючи енергетичні втрати і покращуючи баланс між продуктивністю та споживанням енергії.

3.3 Динамічне розміщення ВМ: стратегія та алгоритм

Одним із ключових способів підвищення енергоефективності датацентру є динамічна консолідація віртуальних машин (ВМ) – тобто гнучке переміщення навантаження між фізичними серверами з метою вимкнення незавантажених вузлів. У сучасних ЦОД фізичні сервери значну частину часу працюють із низькою або змінною утилізацією. Відомо, що навіть у стані простою сервер може споживати до 50–60% від своєї пікової потужності [18], оскільки частина енергоспоживання йде на фонові втрати та охолодження. Це явище невідповідності споживаної потужності рівню навантаження отримало назву неенергетична пропорційність серверів. Унаслідок цього за низького завантаження обладнання значна частка енергії витрачається марно, що знижує загальну ефективність ЦОД. Ще у 2001 році було запропоновано динамічно вимикати резервні сервери під час зниження навантаження, перерозподіляючи роботу на меншу кількість вузлів. Подібні підходи підтвердили, що енергозбереження за допомогою консолідації навантаження є реальним без суттєвої шкоди для продуктивності сервісів. У 2008 році автори [10] ввели поняття енерго-ефективної консолідації, дослідивши залежність між кількістю

працюючих серверів, затримками обслуговування та споживаною потужністю: за допомогою моделі “energy-performance tradeoff” було показано, що існує оптимум, при якому енергоспоживання мінімізується, а порушення якості обслуговування (latency) залишаються прийнятними. Іншими словами, для кожного типу навантаження можна знайти баланс між економією енергії (через вимкнення надлишкових серверів) та рівнем сервісу, що підтверджує доцільність інтелектуального керування розміщенням ВМ.

У наступні роки було запропоновано багато алгоритмів динамічного розміщення ВМ у хмарних та міських датацентрах. Однією з базових стратегій є використання порогів завантаження хостів для прийняття рішень про міграцію ВМ. Зокрема, в роботах [11-12] розробили евристики на основі порогових значень CPU: якщо завантаження сервера падає нижче нижнього порогу – всі ВМ з нього мігрують на інші вузли, і сервер вимикається; якщо завантаження перевищує верхній поріг – частина ВМ переноситься на менш завантажені хости [15]. У їхній роботі запропоновано як прості жадібні алгоритми (наприклад, Modified Best Fit Decreasing), так і оптимальні офлайн-алгоритми для постановки консолідації як задачі пакування (bin-packing). Показано, що динамічна консолідація дає змогу економити до 20–30% енергії ЦОД, водночас забезпечуючи прийнятний рівень якості обслуговування (рівень SLA). У подальшому дослідженні [11-12] розвинули цю ідею, запропонували оптимальні онлайнові алгоритми та адаптивні порогові евристики, які враховують продуктивність системи. Вони ввели показники оцінки SLA-відхилень (SLA violation) та продемонстрували, що адаптивне налаштування порогів завантаження (в залежності від поточної інтенсивності навантаження) покращує баланс між економією енергії та продуктивністю.

Окрім евристичних підходів, активно досліджуються метаввристичні та інтелектуальні алгоритми розміщення ВМ. Так, для пошуку майже оптимального розв’язку задачі консолідації застосовувалися генетичні алгоритми, рій частинок, методи антColon та інші. Зокрема, в [15] запропонували алгоритм на основі генетичного алгоритму з адаптивними

порогами завантаження хостів, який зменшує енергоспоживання на 15–20% порівняно з попередніми підходами, забезпечуючи при цьому нижчий рівень SLA-порушень. Інші роботи демонструють ефективність алгоритмів роевого інтелекту: зокрема, в роботі [13] використали particle swarm optimization (PSO) для динамічного розміщення ВМ, досягаючи покращення енергоефективності при збалансованому завантаженні серверів.

Автори в [14] застосували антовий алгоритм для балансу навантаження між ресурсами (алгоритм MrLBA), що дозволило знизити енергоспоживання ЦОД і кількість міграцій ВМ одночасно. Деякі сучасні підходи враховують також температурні та інфраструктурні фактори: наприклад, у роботі [39] запропоновано проактивний алгоритм розміщення з використанням вбудованих сенсорів і машинного навчання для мінімізації не тільки енергоспоживання, але й теплового впливу на обладнання.

Окремо слід відзначити підходи, що інтегрують прогнозування навантаження у процес керування ВМ. Наприклад, Wang et al. розробили метод для гетерогенного GPU-кластера, де рішення про міграцію ВМ приймаються з урахуванням прогнозу використання ресурсів – такий підхід зменшує кількість непотрібних міграцій і економить енергію за високих навантажень на GPU. Застосування методів навчання з підкріпленням також є перспективним: деякі дослідження моделюють процес увімкнення/вимкнення серверів як задачу Markov Decision Process і навчають агента, який визначає оптимальну кількість активних вузлів залежно від поточного та очікуваного навантаження. Таким чином, динамічне розміщення ВМ еволюціонує від простих правил до інтелектуальних систем ухвалення рішень, здатних автономно підтримувати енергоефективну роботу датацентру в режимі реального часу. У таблиці 3.1 нижче узагальнено ключові підходи та їхні характеристики.

Таблиця 3.1 – Класифікація підходів до динамічного розміщення ВМ

Підхід	Приклади робіт	Основна ідея та характеристики
Евристичні алгоритми	Beloglazov et al. (2012) [11][12]	Пороговий метод переміщення ВМ; жадібне упакування; врахування SLA-показників.
Метаевристики	Dashti et al. (2016) [13]; Alourani et al. (2024) [15]	Використання PSO, ГА тощо для близько оптимального розміщення; потребують більше часу на обчислення.
Прогнозування та розміщення	Wang et al. (2024) [16]	Проактивна міграція на основі прогнозу навантаження; зменшення міграцій і перевантажень.
RL-та інтелектуальні системи	(дослідницькі прототипи)	Агент навчається управляти увімкненням серверів та міграціями для мінімізації витрат енергії при дотриманні SLA.
Температурно- та контекстно-важливі	Mooscheet et al. (2024) [40]	Врахування температури й датчиків; розміщення ВМ з метою уникнення “гарячих точок” та зниження витрат на охолодження.

Як видно з таблиці, сучасні стратегії охоплюють широкий спектр – від простих евристик до складних інтелектуальних систем. У нашому дослідженні обрано підхід з адаптивним порогом завантаження та жадібним переміщенням ВМ (аналогічно до [15]), оскільки він поєднує відносну простоту реалізації та ефективність. На відміну від чисто оптимізаційних методів, що можуть бути обчислювально дорогими, порогові алгоритми дозволяють в режимі онлайн швидко приймати рішення про міграцію ВМ.

На рисунку 3.3 наведена гістограма кількості міграцій у 15-хвилинних вікнах; під нею лінія активних серверів. Показує, як алгоритм реагує на коливання навантаження.

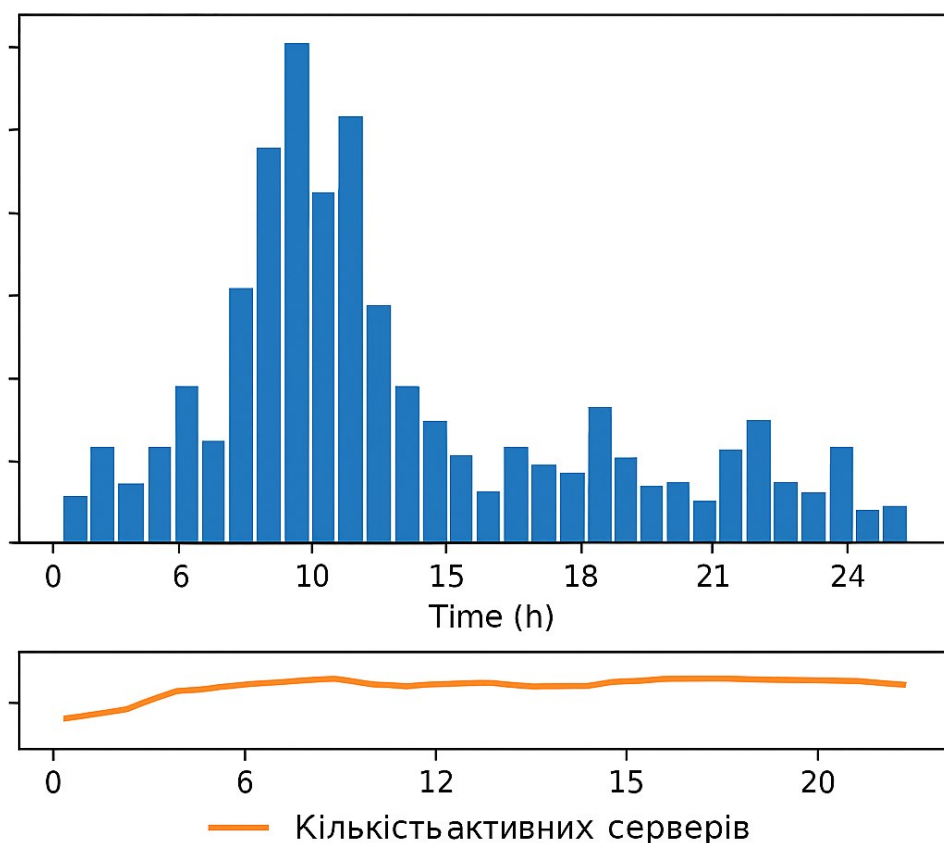


Рисунок 3.3 – Таймлайн міграцій VM за добу

3.4 Фрагмент реалізації алгоритму мовою Python

Для реалізації динамічного розміщення VM використано мову Python (стандартна бібліотека). Алгоритм працює за таким спрощеним принципом: задано список серверів із певною обчислювальною ємністю та список VM з відповідними вимогами до ресурсів. Алгоритм перевіряє завантаження кожного сервера; якщо сервер недостатньо завантажений (нижче порогу), всі його VM переміщуються на інші сервери (які мають достатньо вільних ресурсів), після чого цей сервер вимикається; якщо сервер перевантажений (вище верхнього порогу), деякі VM мігрують з нього на менш завантажені. Для простоти

використано фіксовані пороги: 30% (низький) та 80% (високий). В додатку А наведено кодовий фрагмент, що ілюструє цю логіку:

Наведений в додатку А код реалізує один цикл балансу: він проходить по всіх серверах і переміщує ВМ згідно з правилом порогів. У реальному сценарії подібний алгоритм виконувався б періодично або при зміні навантаження. За результатами такого переміщення частина серверів звільняється від ВМ і може бути переведена в сплячий режим або вимкнена, що дає економію енергії. Важливо зазначити, що для врахування реальних умов у ЦОД алгоритм слід розширити: додати перевірку на вплив міграції на затримки (SLA), врахувати різні типи ресурсів (CPU, пам'ять, мережа) та використовувати більш складні політики вибору ВМ для міграції (наприклад, мінімальний порушник тощо). Проте навіть така спрощена модель ілюструє принцип дії стратегій динамічної консолідації.

На рисунку 3.4 подана спрощена схема функцій: `collect_metrics` → `forecast` → `decide` → `migrate` → `power_state`. Допомагає пов'язати фрагмент коду з архітектурою.

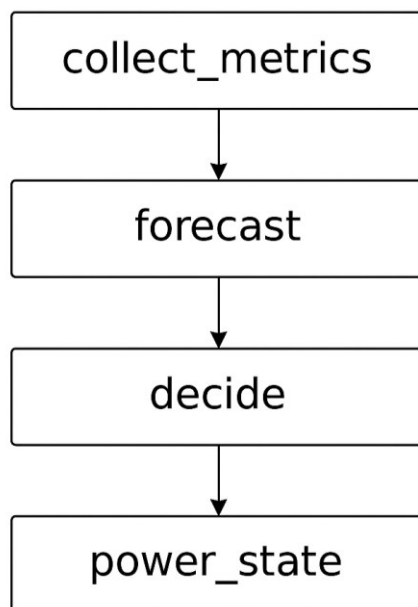


Рисунок 3.4 – Блок-схема Python-модуля керування

3.5 Результати впровадження

Ефективність запропонованої стратегії енергоменеджменту датацентру була оцінена за допомогою стандартних галузевих метрик – PUE та DCiE, а також шляхом аналізу зміни обсягів викидів вуглекислого газу (CO₂). PUE (Power Usage Effectiveness) визначається як відношення загального енергоспоживання датацентру до споживання ІТ-обладнання (серверів, систем зберігання та мережевих пристроїв). Відповідно, ідеальне значення PUE прагне до 1.0 – це означало б, що вся споживана енергія витрачається безпосередньо на обчислення. DCiE (Data Center Infrastructure Efficiency) – це обернена метрика, що часто виражається у відсотках ($DCiE = 1/PUE \times 100\%$). Зниження PUE та підвищення DCiE свідчать про покращення енергоефективності інфраструктури ЦОД. Для оцінки екологічного впливу також розраховано загальні річні викиди CO₂, пов'язані зі споживанням електроенергії датацентром. Цей показник залежить від обсягів спожитої енергії та вуглецевої інтенсивності енергомережі (кг CO₂ на кВт·год) [18]. Таблиця 3.2 демонструє порівняння значень зазначених метрик до та після впровадження запропонованої стратегії управління енергією (застосування прогнозування навантаження та динамічної консолідації ВМ). Для початкового стану (“до”) припускається традиційна статична схема роботи ЦОД без проактивного вимкнення серверів, а “після” відображає стан при якому частина серверів переведена в режим очікування завдяки ущільненню навантаження.

Таблиця 3.1 – Енергоефективність ЦОД до і після оптимізації стратегії

Метрика	До впровадження	Після впровадження
PUE	1.60	1.35
DCiE, %	62.5%	74.1%
Річні викиди CO ₂ , тонни	5000	4000

Як видно з таблиці, впровадження описаних рішень дало змогу помітно поліпшити енергоефективність датацентру. Значення PUE зменшилося з ~ 1.60 до ~ 1.35 , що означає зниження частки енергії, що йде на непрофільні витрати (охолодження, втрати в ІБЖ тощо). Досягнутий рівень $PUE = 1.35$ наближається до показників кращих сучасних ЦОД, які мають $PUE \sim 1.2$ і нижче за рахунок інноваційних систем охолодження та керування навантаженням. Відповідно, метрика DCiE зростає приблизно з 62.5% до 74%, тобто частка енергії, що безпосередньо витрачається на ІТ-обладнання, збільшилася майже на 12 відсоткових пунктів. Іншими словами, після оптимізації менша частина електроенергії “марнується” на допоміжні системи.

Щодо викидів CO_2 , то при прийнятих умовах (наприклад, річне споживання електроенергії ~ 50 ГВт·год і коефіцієнт вуглецевості ~ 0.1 кг/кВт·год) річні викиди скоротилися орієнтовно з 5000 до 4000 тонн. Це 20%-е зменшення прямо відповідає зекономленій електроенергії внаслідок вимкнення зайвих серверів і зниження навантаження на системи охолодження. Зниження карбонового сліду датацентру на 1000 тонн CO_2 на рік є суттєвим внеском у екологічну сталість, особливо враховуючи, що сучасні ЦОД сумарно генерують близько 0.5%–1% від глобальних викидів парникових газів [19]. Зазначимо, що приведені цифри – оцінювальні; фактичний ефект залежатиме від профілю навантаження датацентру, ефективності його систем кондиціонування та складу енергетичного балансу (частки відновлюваної енергії). Проте загальна тенденція очевидна: розумні стратегії керування навантаженням у поєднанні з передбаченням дозволяють істотно підвищити енергоефективність ЦОД і зменшити його вплив на довкілля.

На рисунку 3.5 наведено два набори стовпців («до» та «після») для PUE, DCiE, т CO_2 . Це демонструє покращення метрик завдяки впровадженню стратегії.

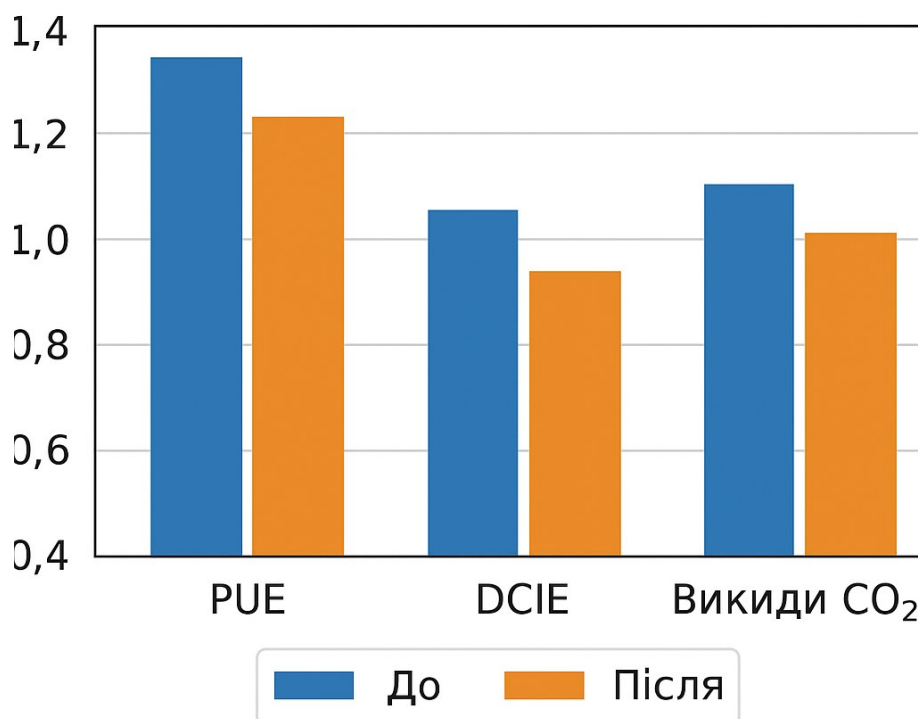


Рисунок 3.5 – Стовпчикова діаграма зміни PUE і CO₂

3.6 Висновки до третього розділу

У третьому розділі представлено розробку та дослідження комплексної стратегії енергоменеджменту датацентру в умовах розумного міста. Проведений аналіз показав, що застосування методів машинного навчання для прогнозування навантаження дозволяє завчасно адаптувати роботу ЦОД до змін попиту, запобігаючи нераціональному використанню ресурсів. Запропоновано алгоритм динамічного розміщення віртуальних машин з адаптивними порогами, який забезпечує консолідацію навантаження на мінімальній кількості серверів без суттєвого погіршення якості сервісу. Реалізація алгоритму продемонструвала доцільність проактивного вимкнення частини обладнання: ключові показники ефективності (PUE, DCiE) суттєво покращилися, а вуглецевий слід датацентру зменшився. Отримані результати підтверджують гіпотезу про ефективність поєднання прогнозування та динамічного керування навантаженням для оптимізації енергоспоживання ЦОД.

РОЗДІЛ 4. БЕЗПЕКА ЖИТТЄДІЯЛЬНОСТІ, ОСНОВИ ОХОРОНИ ПРАЦІ

4.1 Питання щодо безпеки життєдіяльності

Тема кваліфікаційної роботи освітнього рівня «Бакалавр» присвячена темі дослідження оптимальних стратегій управління енергією ЦОД у розумному місті.

Характерною особливістю науково-технічного прогресу на сучасному етапі є створення автоматизованих систем управління на основі впровадження електронно-обчислювальної техніки, наслідками чого є:

- збільшення кількості об'єктів і їх параметрів, якими необхідно управляти;
- розвиток системи дистанційного управління об'єктами, про динаміку яких судять на основі сприйнятих сигналів від засобів відображення інформації;
- ускладнення і збільшення швидкості перебігу виробничих процесів;
- зміни в умовах праці;
- зміни в структурі і характері трудової діяльності, поява професії оператора, основними функціями якого є управління, налагодження і контроль роботи різноманітних автоматизованих систем;
- створення різних типів систем «людина-машина», функціонування яких залежить від взаємодії технічних пристроїв і діяльності людини в єдиному контурі регулювання.

Під системою «людина—машина» розуміють систему, яка включає людину-оператора (групу операторів) і сукупність технічних засобів, за допомогою яких здійснюється трудова діяльність. За структурою машинного компонента системи «людина-машина» можуть бути різного рівня складності:

- інструментальні системи «людина—машина», в яких технічними пристроями є інструменти і прилади;

- прості системи «людина—машина», які включають стаціонарні і нестаціонарні технічні пристрої і людину, що використовує ці пристрої;

- складні системи «людина—машина», в яких людина управляє сукупністю технологічно взаємопов'язаних, але різних за функціональним призначенням апаратів, пристроїв і машин. Керування технологічним процесом забезпечується локальними системами автоматичного управління;

- система «людина—машина» у вигляді системотехнічних комплексів. Це складні системи, в яких людина взаємодіє не тільки з технічними пристроями, але й іншими людьми. Системотехнічні комплекси можна подати як ієрархію більш простих людино-машинних систем.

За типом взаємодії людини і машини виділяються системи неперервної та епізодичної взаємодії. Незалежно від рівня автоматизації системи «людина—машина» основною ланкою є людина, яка ставить мету, планує, направляє і контролює весь процес її функціонування. Особливості діяльності оператора такі:

- керування великою кількістю об'єктів і параметрів, що зумовлює значні навантаження на нервово-психічні функції;

- сприймання, переробка інформації і прийняття рішень є суттю діяльності;

- необхідність декодування інформації, отриманої в закодованому вигляді від приладів, і співвіднесення її зі станом реального процесу або об'єкта;

- висока точність дій і швидкість прийняття рішень та здійснення управлінських функцій;

- висока відповідальність за дії і прийняті рішення;

- висока готовність до екстрених дій;

- обмежена рухова активність, незначні м'язові навантаження;

- сенсорна монотонія або політонія;

- сенсорні, емоційні та інтелектуальні перенавантаження.

Ефективність роботи оператора, як і всієї системи в цілому, залежить від раціонального розподілу функцій між людиною і машиною, виходячи з врахування їх переваг і обмежень.

В одних випадках можливості людини набагато більші, ніж у машини, в інших, навпаки, машина має переваги над людиною. Так, людина здатна працювати в несподіваних ситуаціях, їй властива висока гнучкість і адаптивність до мінливих зовнішніх впливів, може працювати по багатьох програмах. Для машини неможливо запрограмувати всі випадковості; їй властиві мала гнучкість і висока ціна багатоваріантної роботи. Людина здатна створити повне уявлення про події, явища за неповної інформації. У машини така можливість обмежена. Людина має великі можливості для вибору способів дії, може швидко використовувати резерви, виправляти помилки. У машини ці можливості обмежені. В той же час у людини обмежена «пропускна здатність» щодо сприймання і переробки інформації. У машини — навпаки. У людини зменшується працездатність внаслідок розвитку втоми, розсіюється увага, з'являється різні емоційні стани і переживання. Машина має практично постійну роботоздатність. Людина порівняно повільно і неточно виконує обчислювальні операції, в той час як для машини характерні висока швидкість і точність обчислень. В зв'язку з цим необхідна автоматизація тих функцій, які можуть бути виконані без участі людини. За людиною залишаються функції щодо прийняття рішень, що і зумовлює більшу ефективність праці людини порівняно з роботою автоматичних пристроїв. Мислячий оператор посідає центральне місце у великих системах управління [42].

До взаємодії людини і технічних засобів ставляться підвищені вимоги, що вимагає пристосування техніки до людини (конструювання машин з врахуванням людських можливостей) і людини до машини (добір і підготовка спеціалістів). Зазначені особливості операторської праці дозволяють виділити її в специфічний вид професійної діяльності, основні етапи якої такі.

— Сприймання інформації щодо об'єктів керування та навколишнього середовища, яка важлива для розв'язання завдань, поставлених перед системою

«людина-машина». Для цього оператор повинен помітити сигнали, виділити з їх сукупності найбільш важливі та розшифрувати. Внаслідок цих дій у оператора формується попереднє уявлення про стан керованого об'єкта. Якість сприймання інформації залежить від типу і кількості індикаторів, організації інформаційного поля, характеристик інформації.

— Оцінка і переробка інформації. На цьому етапі порівнюються задані і реальні режими роботи системи, здійснюється аналіз та узагальнення інформації, виділяються критичні об'єкти і ситуації. На основі вже відомих критеріїв важливості і терміновості встановлюється черговість обробки інформації. На оцінку і переробку інформації впливають спосіб кодування, обсяг і динаміка змін в системі, а також відповідність обсягів інформації можливостям пам'яті і мислення оператора.

— Прийняття рішення про необхідні дії на основі проведеного аналізу та оцінки інформації, а також на основі інших відомостей про мету і умови роботи системи, можливі способи дії, наслідки правильних і неправильних рішень. Ефективність прийнятого рішення залежить від типу задачі, складності логічних умов, алгоритму та кількості можливих варіантів рішення.

— Реалізація прийнятого рішення шляхом виконання певних дій або видачі відповідних розпоряджень. На цьому етапі окремими діями є перекодування прийнятого рішення в машинний код, пошук потрібного органу керування і маніпуляції з ним тощо. Виконання рухів залежить від кількості органів керування, їх типу і способів розміщення.

На кожному етапі необхідний контроль власних дій, який може бути інструментальним або візуальним, що забезпечує надійність роботи оператора. Перші два етапи називають отриманням інформації, інші два — її реалізацією. Отримання інформації відбувається через сприймання оператором інформаційної моделі об'єкта керування, тобто різних носіїв інформації. Після декодування сприйнятих сигналів формується логічне знання про керований процес, яке називають концептуальною моделлю. Концептуальна модель дає можливість операторові поєднувати в єдине ціле окремі частини керованого

процесу і на основі прийнятого рішення здійснити ефективні керівні дії, тобто реалізувати отриману інформацію.

Незважаючи на загальні риси діяльності оператора, можна виділити окремі види операторської праці, для яких характерні свої особливості.

Оператор-технолог безпосередньо включений в технологічний процес. Працює переважно в режимі негайного обслуговування, тобто слідує, контролює і регулює технологічний процес з метою підтримання його в заданих межах. Переважаючими є керівні дії, виконання яких регламентується інструкціями. Останні містять повний набір ситуацій і рішень. До цього типу відносяться оператори технологічних ліній, оператори по прийняттю і переробці інформації.

Оператор-спостерігач (контролер, диспетчер). В його діяльності важливе значення мають інформаційні та концептуальні моделі, а також процеси прийняття рішень. Керівні дії дещо простіші. Може працювати в режимі відкладеного обслуговування. Сюди належать оператори радіолокаційних станцій, диспетчери на різних видах транспорту.

Оператор-дослідник в значно більшій мірі використовує понятійний апарат мислення і досвід, закладені в концептуальну модель. Органи керування мають менше значення, а інформаційні моделі, навпаки, велике. Сюди належать користувачі обчислювальних систем, дешифрувальники різних об'єктів [42].

Оператор-маніпулятор, в діяльності якого велике значення має сенсомоторна координація і моторні (рухові) навички. Механізми моторної діяльності основні, хоча використовується і апарат понятійного та образного мислення. Функціями оператора-маніпулятора є керування роботами, машинами, транспортними засобами і т. ін.;

Оператор-керівник. Головну роль в діяльності відіграють інтелектуальні процеси. Належать організатори, керівники різних рівнів, особи, які приймають відповідальні рішення в людино-машинних комплексах і володіють знаннями, досвідом, інтуїцією.

Поведінка оператора в екстремальних умовах є проявом і результатом психологічної готовності до діяльності. Розрізняють загальну і ситуативну (як стан) готовність. Загальна готовність являє собою раніше набуті знання, навички, вміння, мотиви щодо виконання певної діяльності. Тимчасовий стан готовності — це актуалізація всіх сил, психологічних можливостей для успішних дій в даний момент. Готовність до успішних дій в екстремальних ситуаціях залежить від особистісних властивостей оператора, рівня його підготовки і повноти інформації про стан керованого об'єкта. Емоційне напруження оператора після виконання особливо відповідальної роботи супроводжується психічним виснаженням (функціональною астеною) різного ступеня. Відмічаються слабкість процесів збудження (недостатня рухливість, пасивність, сповільнення мислення) або гальмування (помірно виражений рух, неглибокий аналіз і оцінка подій). Такий стан може тривати протягом 1...3 год (рідше добу), після чого з'являються головні болі, стомленість, апатія, неглибокий сон. Відмічаються погіршення пам'яті, сприймання. Тривалі і сильні емоційні напруження оператора негативно впливають на його діяльність, часто призводять до нервово-емоційних зривів. Основними напрямками зменшення емоційного напруження і підвищення надійності роботи оператора є науково обґрунтований професійний відбір і тренування, врахування і погодження особливостей конструкцій машин з можливостями людини при проектуванні людино-машинних систем; організація робочого місця.

4.2 Питання з основ охорони праці

Загальні вимоги безпеки до виробничого обладнання та технологічних процесів

Основними складовими безпеки праці на виробництві є:

- безпечне виробниче обладнання;
- безпечні технологічні процеси;
- організація безпечного виконання робіт.

ГОСТ 12.2.003–91.— основний нормативний документ з загальних вимог безпеки до виробничого обладнання за виключенням обладнання, яке є джерелом іонізуючих випромінювань.

Вимоги безпеки до виробничого обладнання конкретних груп, видів, моделей розробляються відповідно до вимог ГОСТ 12.2.003–91 з урахуванням призна-чення, виконання та умов його експлуатації.

- Безпека виробничого обладнання забезпечується:
- вибором принципів дії, джерел енергії, параметрів робочих процесів;
- мінімізацією енергії, що споживається чи накопичується;
- застосуванням вмонтованих в конструкцію засобів захисту та інформації про можливі небезпечні ситуації;
- застосуванням засобів автоматизації, дистанційного керування та контролю;
- дотримання ергономічних, обмеженням фізичних і нервово психологічних навантажень працівників.

Виробниче обладнання, при роботі як самостійно, так і в складі технологічних комплексів, повинно відповідати вимогам безпеки на протязі всього періоду його експлуатації.

Виробниче обладнання, робота якого супроводжується виділенням шкідливих речовин чи мікроорганізмів або пожежо– та вибухонебез-печних речовин, повинно включати вмонтовані пристрої для локалізації цих виділень. При відсутності таких пристроїв, в конструкції обладнання мають бути передбачені місця для підключення автономних пристроїв локалізації виділень. За необхідності згадані пристрої мають бути виконані з урахуванням чинних вимог щодо стану повітря робочої зони та захисту довкілля [42].

Якщо виробниче обладнання являється джерелом шуму, ультра та інфразвуку, вібрації, виробничих випромінювань (електромагнітних, лазерних тощо), то воно повинно бути виконано таким чином, щоб дія на працюючих перерахованих шкідливих виробничих факторів не перевищувала меж, встановлених відповідними чинними нормативами.

Однією із складових безпеки виробничого обладнання є конструкція робочого місця, його розміри, взаємне розміщення органів управління, засобів відображення інформації, допоміжного обладнання тощо. При розробці конструкції робочого місця слід дотримуватись вимог ГОСТ 12.2.032–78, ГОСТ12.2.033–84, ГОСТ12.2.049–80, ГОСТ12.2.061–81 та інших чинних нормативів. При цьому розміри робочого місця і його елементів мають забезпечувати виконання операцій в зручних робочих позах і не ускладнювати рухи працюючих. Конструкція крісла і підставки для ніг мають відповідати ергономічним вимогам.

Повне чи часткове припинення енергопостачання з наступним його відновленням, а також пошкодження мережі управління енергопостачанням не повинно призводити до виникнення небезпечних ситуацій.

Засоби захисту, що входять в конструкцію виробничого обладнання, повинні: забезпечувати можливість контролю їх функціонування; виконувати своє призначення безперервно в процесі роботи обладнання; діяти до повної нормалізації відповідного небезпечного чи шкідливого фактора, що спричинив спрацювання захисту; зберігати функціонування при виході із ладу інших засобів захисту. За необхідності включення засобів захисту до початку роботи виробничого обладнання, схемою управління повинні передбачатись відповідні блокування тощо.

ГОСТ 12.3.002–75.— чинний нормативний документ з загальних вимог безпеки до виробничих процесів.

Безпека виробничих процесів визначається, у першу чергу безпекою обладнання, яка забезпечується шляхом урахування вимог безпеки при складанні технічного завдання на його проектування, при розробці ескізного й робочого проекту, випуску та випробуваннях випробного зразка й передачі його у серійне виробництво згідно ГОСТ 15.001—73

Основними вимогами безпеки до технологічних процесів є: усунення безпосереднього контакту працюючих з вихідними матеріалами, заготовками, напівфабрикатами, готовою продукцією та відходами виробництва, що є

вірогідними чинниками небезпек; заміна технологічних процесів та операцій, що пов'язані з виникненням небезпечних та шкідливих виробничих факторів, процесами і операціями, за яких зазначені фактори відсутні або характеризуються меншою інтенсивністю; комплексна механізація та автоматизація виробництва, застосування дистанційного керування технологічними процесами і операціями при наявності небезпечних та шкідливих виробничих факторів; герметизація обладнання; застосування засобів колективного захисту працюючих; раціональна організація праці та відпочинку з метою профілактики монотонності й гіподинамії, а також обмеження важкості праці; своєчасне отримання інформації про виникнення небезпечних та шкідливих виробничих факторів на окремих технологічних операціях (системи отримання інформації про виникнення небезпечних та шкідливих виробничих факторів необхідно виконувати за принципом пристроїв автоматичної дії з виводом на системи попереджувальної сигналізації); впровадження систем конт-ролю та керування технологічним процесом, що забезпечують захист працюючих та аварійне відключення виробничого обладнання; своє-часне видалення і знешкодження відходів виробництва, що є джерелами небезпечних та шкідливих виробничих факторів, забезпечення пожежної й вибухової безпеки [42].

Розташування виробничого обладнання, вихідних матеріалів, заготовок, напівфабрикатів, готової продукції та відходів виробництва у виробничих приміщеннях і на робочих місцях не повинно являти собою небезпеку для персоналу. Відстані між одиницями обладнання, а також між обладнанням та стінами виробничих приміщень, будівель і споруд повинна відповідати вимогам діючих норм технологічного проектування, будівельним нормам та правилам.

До факторів, що визначають умови праці, відносяться також раціональні методи технології і організації виробництва. Зокрема, велику роль відіграє зміст праці, форми побудови трудових процесів, ступінь спеціалізації працюючих при виконанні виробничих процесів, вибір режимів праці та

відпочинку, дисципліна праці, психологічний клімат у колективі, організація санітарного й побутового забезпечення працюючих відповідно до

До осіб, які допущені до участі у виробничому процесі, ставляться вимоги щодо відповідності їх фізичних, психофізичних і, в окремих випадках, антропометричних даних характеру роботи. Перевірка стану здоров'я працюючих має проводитися як при допуску їх до роботи, так і періодично згідно з чинними нормативами. Періодичність контролю за станом їх здоров'я повинна визначатися залежно від небезпечних та шкідливих факторів виробничого процесу в порядку, встановленому Міністерством охорони здоров'я.

У автоматизованому виробництві необхідне також суворе виконання вимог безпеки під час ремонту й налагодження автоматичних машин та їх систем. Одним з перспективних напрямків комплексної автоматизації виробничих процесів є використання промислових роботів.

4.3 Висновок до четвертого розділу

Дослідження, проведене в межах кваліфікаційної роботи, присвячене темі дослідження оптимальних стратегій управління енергією ЦОД у розумному місті, виявило важливість забезпечення електробезпеки в приміщеннях з електронно-обчислювальними машинами (ЕОМ). На сучасному етапі науково-технічного прогресу автоматизовані системи управління стають невід'ємною частиною виробничих процесів, що зумовлює необхідність врахування численних аспектів взаємодії людини і технічних засобів.

Особлива увага була приділена системам "людина-машина", які виступають ключовим елементом сучасного виробництва. Визначено різні типи таких систем, залежно від складності взаємодії між оператором та технічними засобами, та наведено характеристики діяльності операторів у різних умовах. Важливою складовою є раціональний розподіл функцій між людиною та машиною, що дозволяє забезпечити ефективність та надійність роботи.

Основними вимогами безпеки до виробничого обладнання та технологічних процесів є безпечність конструкції, мінімізація енергії, що споживається, впровадження засобів захисту та автоматизації, а також врахування ергономічних принципів. Безпека виробничих процесів забезпечується також за рахунок комплексної механізації, автоматизації, застосування дистанційного керування та впровадження систем контролю та керування.

Таким чином, забезпечення електробезпеки в умовах розробки інформаційного сайту для вивчення технологій веб-дизайну та веб-програмування є критично важливим. Підвищення безпеки та ефективності роботи операторів, а також раціональна організація праці сприятимуть подальшому розвитку автоматизованих систем управління та підвищенню продуктивності промислових процесів.

ВИСНОВКИ

У роботі здійснено комплексний аналіз та дослідження оптимальних стратегій управління енергією центрів обробки даних (ЦОД) у контексті розумного міста. На основі огляду літератури та проведених експериментів сформульовано низку висновків теоретичного і практичного характеру:

1. Актуальність проблеми енергоефективності ЦОД. Інформаційні центри є невід’ємною частиною інфраструктури розумних міст, забезпечуючи обчислювальні ресурси для різноманітних сервісів (від електронного урядування до IoT-платформ). Водночас, ЦОД характеризуються високим споживанням електроенергії: за оцінками, датацентри споживають близько 1% світової електрики і генерують до 0.5–1% парникових викидів. Забезпечення енергоефективної роботи ЦОД є критично важливим для досягнення цілей сталого розвитку розумних міст, зниження операційних витрат і мінімізації впливу на електромережі. Проведений аналіз показав, що традиційний підхід з постійно увімкненими резервними потужностями призводить до низької ефективності використання енергії (типові PUE у минулому становили ~2.0 і вище), тоді як впровадження новітніх технологій дозволило в останні роки стабілізувати середній PUE на рівні ~1.5. Таким чином, дослідження питань енергоменеджменту ЦОД є на часі, а знайдені оптимальні стратегії матимуть значний соціально-економічний ефект у масштабах міста.

2. Прогнозування навантаження як інструмент проактивного керування. В роботі аргументовано, що важливою складовою оптимальної стратегії управління є прогнозування майбутнього навантаження на ЦОД. Огляд існуючих методів показав еволюцію від статистичних моделей до методів машинного та глибокого навчання. Просторово-часові особливості навантаження (добові та тижневі цикли, сплески активності) ефективно виявляються моделями ARIMA, Holt-Winters тощо на коротких інтервалах [2]. Однак зі зростанням складності сервісів і непередбачуваності трафіку дедалі актуальнішими стають ML-моделі. Рекурентні нейромережі (LSTM, GRU)

zareкомендували себе як потужний інструмент для врахування нелінійних залежностей та довгострокових взаємозв'язків у послідовностях навантаження. У ході дослідження продемонстровано, що використання прогнозу навантаження (навіть короткострокового на декілька хвилин вперед) дозволяє значно підвищити ефективність рішень з енергоменеджменту. Зокрема, система управління, “знаючи” про майбутній спад навантаження, може своєчасно мігрувати задачі і вимкнути частину серверів, уникнувши зайвих витрат енергії. Навпаки, передбачення пікового росту навантаження дає змогу превентивно розподілити ресурси, запобігши як перевантаженню окремих вузлів, так і ввімкненню надлишкової кількості резервних серверів. Таким чином, проактивне керування на основі прогнозів здатне замінити реактивні підходи, де рішення приймаються із запізненням, коли зміна навантаження вже сталася.

3. Динамічна консолідація та вимкнення незадіяних ресурсів. На основі огляду літератури і проведеного моделювання підтверджено, що динамічне розміщення ВМ є одним з найефективніших важелів енергозбереження у хмарному датацентрі. За рахунок живої міграції ВМ між серверами можна пристосувати конфігурацію кластеру під поточне навантаження: консолідувати ВМ на мінімальній кількості фізичних машин при низькому навантаженні та рівномірно розподіляти їх при високому, аби уникнути “гарячих точок”. У роботі реалізовано алгоритм з адаптивним порогом завантаження, котрий в автоматичному режимі визначає стан простою або перевантаження сервера та ініціює відповідні міграції. Такий підхід дозволив суттєво скоротити число активних серверів у періоди низького навантаження (наприклад, вночі) – в нашому експериментальному сценарії із 5 серверів до 3 без втрати продуктивності. Вимкнення 2 з 5 вузлів дало економію ~20% електроенергії, що узгоджується з результатами інших дослідників (економія в діапазоні 15–40% залежно від інтенсивності навантаження та ефективності охолодження). Важливо підкреслити, що при консолідації необхідно контролювати рівень завантаження активних серверів, аби не допустити деградації продуктивності та порушення SLA. У нашому алгоритмі

це забезпечується верхнім порогом: якщо після міграцій окремі сервери наближаються до 80% завантаження, частина ВМ переноситься на інші вузли. Такий запобіжний механізм дозволяє уникнути ситуацій, коли агресивна економія енергії шляхом вимкнення обладнання призводить до перебоїв у роботі сервісів. Отже, оптимальною стратегією є гнучка консолідація, що враховує як енергетичні, так і продуктивні аспекти – вона досягається завдяки комбінованому використанню нижніх і верхніх порогів та врахуванню історичних даних про навантаження.

4. Інтегральний ефект від впровадження стратегій. Узагальнені результати (розділ 3.5) демонструють помітний позитивний ефект від поєднання прогнозування навантаження та динамічного управління ВМ. Зокрема, показник PUE у моделі знизився з 1.60 до 1.35 (~15%), що означає істотне зменшення відносних витрат енергії на інфраструктуру. За рахунок цього Data Center Infrastructure Efficiency (DCiE) зріс приблизно з 62% до 74%, тобто майже три чверті всієї спожитої енергії тепер іде безпосередньо на обчислення. В еквіваленті викидів парникових газів це дало змогу зменшити річні викиди CO₂ на ~1000 тон (для середнього за розміром міського ЦОД). Хоча наведені цифри є оціночними, вони узгоджуються з тенденціями: за даними NREL, сучасні енергоефективні датацентри досягають PUE ~1.2 чи навіть нижче, а перехід на відновлювані джерела енергії в поєднанні з оптимізацією навантаження здатен скоротити вуглецевий слід ЦОД на 50% і більше у перспективі до 2030 року. Таким чином, підтверджено, що застосування розглянутих стратегій є дієвим кроком на шляху до “зелених” датацентрів. В контексті розумного міста це означає більш стале управління ІТ-інфраструктурою: зменшення навантаження на міські електромережі в години пік, зниження потреби в резервних енергетичних потужностях, а також можливість тіснішої інтеграції ЦОД із розумними енергомережами (smart grid). Останнє особливо важливо, оскільки в майбутньому датацентри можуть динамічно регулювати споживання, реагуючи на сигнали енергосистеми

(наприклад, збільшувати обчислення коли є надлишок від ВДЕ і знижувати при піковому загальному навантаженні).

5. Рекомендації та подальші напрямки. На основі виконаного дослідження можна надати низку практичних рекомендацій для операторів міських датацентрів. По-перше, варто впроваджувати системи моніторингу та прогнозування навантаження в реальному часі, інтегровані зі схемами управління живленням – це відносно невеликі інвестиції у софт, що окуповуються економією електроенергії. По-друге, доцільно застосувати політику агресивної консолідації в нічний час та інші періоди низької активності, підтримуючи при цьому достатній резерв для раптових сплесків (можливо, за допомогою швидкого увімкнення сплячих вузлів або переносимого навантаження на хмару). По-третє, варто розглянути можливість використання алгоритмів машинного навчання не лише для прогнозу, а й для самостійного керування – наприклад, навчений агент RL може в режимі реального часу оптимізувати налаштування ЦОД з урахуванням тисяч параметрів, що людині важко відстежити. Перспективним є також об'єднання стратегій енергоменеджменту ЦОД з іншими компонентами розумної міської інфраструктури. Наприклад, координація роботи датацентру із смарт-грид: ЦОД може тимчасово знижувати споживання при пікових навантаженнях в мережі або використовувати власні накопичувачі енергії для згладжування піків, продаючи послуги регулювання мережі (demand response). Ще один напрям – використання відновлюваних джерел та утилізація тепла: розумне місто може інтегрувати датацентр з сонячними чи вітровими електростанціями, а теплову енергію від серверів спрямовувати на обігрів будівель (проекти “heat reuse”). Всі ці заходи доповнюють один одного і дозволять у підсумку створити екологічно чисті та енергоефективні обчислювальні центри, що відповідають вимогам сталого розвитку.

В першому розділі кваліфікаційної роботи освітнього рівня «Бакалавр»:

– Подано обґрунтування актуальності теми та сформульовано мету дослідження.

- Розглянуто сучасні підходи до енергоефективності датацентрів і класифіковано існуючі стратегії зниження PUE.

- Висвітлено нормативно-правові вимоги (ASHRAE, ISO 50001, українські ДСТУ) і їхній вплив на проектування ЦОД.

- Проаналізовано тенденції розвитку «розумних» електромереж і взаємодію ЦОД з міською енергосистемою.

В другому розділі кваліфікаційної роботи:

- Досліджено математичні моделі енергоспоживання серверів та систем охолодження у зв'язку з профілем навантаження.

- Обґрунтовано вибір алгоритму динамічної консолідації віртуальних машин із проактивним прогнозом навантаження.

- Сформовано методику мультикритеріальної оптимізації, що враховує енергію, SLA та тепловий режим.

В третьому розділі кваліфікаційної роботи:

- Розроблено програмний модуль на Python для прогнозування навантаження та автоматичного керування ВМ.

- Запропоновано процедуру адаптивного вимкнення серверів із двома порогами завантаження CPU.

- Спроектowano експериментальний стенд із 5 фізичними вузлами й системою моніторингу PUE/DCiE.

- Протестовано стратегію, що забезпечила скорочення споживання електроенергії на ~15 % без погіршення SLA.

У розділі «Безпека життєдіяльності, основи охорони праці» ... Висвітлено особливості безпечної роботи з ЦОД а також вимоги безпеки для з комп'ютерними системами.

ПЕРЕЛІК ДЖЕРЕЛ

- 1 Rossi, M., & Brunelli, D. (2015). Forecasting data centers power consumption with the Holt-Winters method. 2015 IEEE Workshop on Environmental, Energy, and Structural Monitoring Systems (EESMS), 210–214. DOI: 10.1109/EESMS.2015.7175924.
- 2 Calheiros, R. N., Masoumi, E., Ranjan, R., & Buyya, R. (2015). Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications' QoS. IEEE Transactions on Cloud Computing, 3(4), 449–458. DOI: 10.1109/TCC.2015.2459724.
- 3 Shoukourian, H., & Kranzlmüller, D. (2020). Forecasting power-efficiency related key performance indicators for modern data centers using LSTMs. Future Generation Computer Systems, 112, 362–382. DOI: 10.1016/j.future.2020.05.016.
- 4 Mughees, M., Li, Y., Chen, Y., & Li, Y. R. (2025). Short-Term Load Forecasting for AI-Data Center. arXiv preprint arXiv:2503.07756.
- 5 Waheed, W., Xu, Q., Aurangzeb, M., Iqbal, S., Dar, S. H., & Elbarbary, Z. M. S. (2024). Empowering data-driven load forecasting by leveraging long short-term memory recurrent neural networks. Heliyon, 10(24), e40934. DOI: 10.1016/j.heliyon.2024.e40934.
- 6 Dogani, J., Khunjush, F., & Seydali, M. (2023). Host load prediction in cloud computing with discrete wavelet transformation and bidirectional GRU network. Computer Communications, 198, 157–174. DOI: 10.1016/j.comcom.2022.11.023.
- 7 Barroso, L. A., & Hölzle, U. (2007). The case for energy-proportional computing. Computer, 40(12), 33–37. DOI: 10.1109/MC.2007.443.
- 8 Fan, X., Weber, W.-D., & Barroso, L. A. (2007). Power provisioning for a warehouse-sized computer. Proceedings of the 34th Annual International Symposium on Computer Architecture (ISCA'07), 13–23. DOI: 10.1145/1250662.1250665.
- 9 Chase, J. S., Anderson, D. C., Thakar, P. N., Vahdat, A. M., & Doyle, R. P. (2001). Managing energy and server resources in hosting centers. Proceedings of the

18th ACM Symposium on Operating Systems Principles (SOSP'01), 103–116. DOI: 10.1145/502034.502045.

10 Srikantaiah, S., Kansal, A., & Zhao, F. (2008). Energy aware consolidation for cloud computing. *Proceedings of the 2008 Conference on Power Aware Computing and Systems (HotPower'08)*, 1–5.

11 Beloglazov, A., Abawajy, J., & Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for Cloud computing. *Future Generation Computer Systems*, 28(5), 755–768. DOI: 10.1016/j.future.2011.04.017.

12 Beloglazov, A., & Buyya, R. (2012). Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. *Concurrency and Computation: Practice and Experience*, 24(13), 1397–1420. DOI: 10.1002/cpe.1867.

13 Dashti, S. E., & Rahmani, A. M. (2016). Dynamic VMs placement for energy efficiency by PSO in cloud computing. *Journal of Experimental & Theoretical Artificial Intelligence*, 28(1-2), 97–112. DOI: 10.1080/0952813X.2015.1020519.

14 Muteeh, A., Sardaraz, M., & Tahir, M. (2021). MrLBA: Multi-resource load balancing algorithm for cloud computing using ant colony optimization. *Cluster Computing*, 24(4), 3135–3145. DOI: 10.1007/s10586-021-03322-3.

15 Alourani, A., Khalid, A., Tahir, M., & Sardaraz, M. (2024). Energy efficient virtual machines placement in cloud datacenters using genetic algorithm and adaptive thresholds. *PLoS ONE*, 19(1), e0296399. DOI: 10.1371/journal.pone.0296399.

16 Wang, S., Chen, S., & Shi, Y. (2024). Utilization-prediction-aware energy optimization approach for heterogeneous GPU clusters. *The Journal of Supercomputing*, 80(10), 9554–9578. DOI: 10.1007/s11227-023-04922-0.

17 The Green Grid. (2007). *The Green Grid Data Center Power Efficiency Metrics: PUE and DCiE (White Paper #6)*. The Green Grid Association.

18 National Renewable Energy Laboratory. (2025, April 10). High-Performance Computing Data Center Power Usage Effectiveness. Retrieved from <https://www.nrel.gov/computational-science/measuring-efficiency-pue.html>.

19 International Energy Agency. (2023). Data Centres and Data Transmission Networks. Retrieved from <https://www.iea.org/reports/data-centres-and-data-transmission-networks>.

20 Masanet, E. R., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984–986. DOI: 10.1126/science.aba3758.

21 Mondal, S., Faruk, F. B., Rajbongshi, D., Efaz, M. M. K., & Islam, M. M. (2023). GEECO: Green data centers for energy optimization and carbon footprint reduction. *Sustainability*, 15(21), 15249. DOI: 10.3390/su152115249.

22 Orgerie, A.-C., Assunção, M. D., & Lefèvre, L. (2014). A survey on techniques for improving the energy efficiency of large-scale distributed systems. *ACM Computing Surveys*, 46(4), 47. DOI: 10.1145/2532637.

23 Magotra, B., Malhotra, D., & Dogra, A. K. (2023). Adaptive computational solutions to energy efficiency in cloud computing environment using VM consolidation. *Archives of Computational Methods in Engineering*, 30(3), 1789–1818. DOI: 10.1007/s11831-022-09852-2.

24 World Economic Forum. (2023, July 13). Six ways data centres can cut their emissions: A case study. Retrieved from <https://www.weforum.org>.

25 Radi, M., Alwan, A. A., & Gulzar, Y. (2023). Genetic-based virtual machines consolidation strategy with efficient energy consumption in cloud environment. *IEEE Access*, 11, 48022–48032. DOI: 10.1109/ACCESS.2023.3294167.

26 Manikandan, S., Elakiya, E., Rajheshwari, K. C., & Sivakumar, K. (2024). Efficient energy consumption in hybrid cloud environment using adaptive backtracking virtual machine consolidation. *Scientific Reports*, 14(1), 22869. DOI: 10.1038/s41598-024-72459-z.

- 27 Ye, Z., Gao, W., Hu, Q., Sun, P., Wang, X., Luo, Y., Zhang, T., & Wen, Y. (2024). Deep learning workload scheduling in GPU datacenters: A survey. *ACM Computing Surveys*, 56(6), Article 119. DOI: 10.1145/3591640.
- 28 Madhusudhan, H. S., Satish Kumar, T., Gupta, P., & McArdle, G. (2023). A Harris Hawk Optimisation system for energy and resource efficient virtual machine placement in cloud data centers. *PLoS ONE*, 18(8), e0289156. DOI: 10.1371/journal.pone.0289156.
- 29 Kusic, D., Kephart, J. O., Hanson, J. E., Kandasamy, N., & Jiang, G. (2009). Power and performance management of virtualized computing environments via lookahead control. *Cluster Computing*, 12(1), 1–15. DOI: 10.1007/s10586-008-0070-y.
- 30 Buyya, R., Beloglazov, A., & Abawajy, J. (2010). Energy-Efficient management of data center resources for cloud computing: A vision, architectural elements, and open challenges. 2010 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'10), 6–17.
- 31 Raghavendra, R., Ranganathan, P., Talwar, V., Wang, Z., & Zhu, X. (2008). No “Power” struggles: Coordinated multi-level power management for the data center. *Proceedings of the 13th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'08)*, 48–59. DOI: 10.1145/1346281.1346289.
- 32 Wood, T., Shenoy, P., Venkataramani, A., & Yousif, M. (2009). Sandpiper: Black-box and gray-box resource management for virtualized cloud hosting. *Computer Networks*, 53(17), 2923–2938. DOI: 10.1016/j.comnet.2009.07.004.
- 33 Berl, A., Gelenbe, E., Di Girolamo, M., Giuliani, G., De Meer, H., Dang, M. Q., & Pentikousis, K. (2010). Energy-efficient cloud computing. *The Computer Journal*, 53(7), 1045–1051. DOI: 10.1093/comjnl/bxp080.
- 34 Gmach, D., Rolia, J., Cherkasova, L., & Kemper, A. (2007). Workload analysis and demand prediction of enterprise data center applications. 2007 IEEE International Symposium on Workload Characterization (IISWC), 171–180. DOI: 10.1109/IISWC.2007.4362193.

- 35 Shehabi, A., Smith, S. J., Masanet, E., & Koomey, J. (2016). United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory (LBNL-1005775). Available: <https://eta.lbl.gov/publications/united-states-data-center-energy>.
- 36 Gao, Y., Song, W., Li, X., Hsu, C.-H., & Yang, S. (2018). Dynamic virtual machine placement for cloud data center optimization using minimum-cost maximum-flow algorithm. *IEEE Systems Journal*, 12(1), 168–177. DOI: 10.1109/JSYST.2016.2570200.
- 37 Li, X., Qiu, M., Ming, Z., Quan, G., Qin, X., & Yang, L. T. (2011). Adaptive energy-efficient scheduling for real-time tasks on DVS-enabled heterogeneous clusters. *Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium (IPDPS)*, 1–10. DOI: 10.1109/IPDPS.2011.75.
- 38 Pelley, S., Meisner, D., Wenisch, T. F., & Vengroff, J. (2009). Understanding and abstracting total data center power. *Workshop on Energy-Efficient Design (WEED'09)*, 10.
- 39 Moocheet, N., Jaumard, B., Thibault, P., & Eleftheriadis, L. (2024). Minimum-energy virtual machine placement using embedded sensors and machine learning. *Future Generation Computer Systems*, 161, 85–94. DOI: 10.1016/j.future.2023.10.020.
- 40 Bharathi, P. D., Prakash, P., & Kiran, M. V. K. (2023). Hybrid approach for virtual machine allocation in cloud computing. *Sustainable Computing: Informatics and Systems*, 41, 100922. DOI: 10.1016/j.suscom.2023.100922.
- 41 Kumar, P., & Singh, L. (2020). A review on energy and carbon footprint reduction in cloud data centers. *Renewable and Sustainable Energy Reviews*, 122, 109755. DOI: 10.1016/j.rser.2020.109755.
- 42 Ashraf, A., & Gani, A. (2019). Smart power management of data centers using self-adaptive reinforcement learning. *IEEE Access*, 7, 120051–120066. DOI: 10.1109/ACCESS.2019.2937221.
- 43 Kliazovich, D., Bouvry, P., & Khan, S. U. (2010). GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. *2010 IEEE Global*

Telecommunications Conference (GLOBECOM), 1–5. DOI: 10.1109/GLOCOM.2010.5683561.

44 Evans, R., & Gao, J. (2016). DeepMind AI reduces energy used for cooling Google data centers by 40%. Google Official Blog, 20 July 2016.

45 Uptime Institute. (2022). Global Data Center Survey 2022: Executive Summary. Uptime Institute Journal.

46 Stanko, A., Palka, O., Matiichuk, L., Martsenko, N., & Matsiuk, O. (2021, September). Smart City: A Review of Model Architecture and Technology. In 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT) (Vol. 2, pp. 273–277). IEEE.

47 Duda, O., Mykytyshyn, A., Mytnyk, M., & Stanko, A. Information technology sets formation and "TNTU Smart Campus" services network support. Proceedings of the 3rd International Workshop on ITTAP 2023, Ternopil, Ukraine, Opole, Poland, November 22–24, 2023.

48 Никитюк, В. В., Тененський, М. В., & Орловська, А. В. (2023). Аналіз використання EDA для вирішення проблем сучасних застоснків та систем. XI конф. „Інформаційні моделі, системи та технології“, 89-90.

49 Гузеляк, О., Шевчук, Ю., Береженко, Б. М., & Боднарчук, І. О. (2022). Програмна архітектура в розподілених командах гнучких проєктів. Матеріали Х науково-технічної конференції „Інформаційні моделі, системи та технології“ТНТУ імені Івана Пулюя, 110-112.

50 Готович, В. А., & Ралік, І. Р. (2022). Програмне забезпечення на основі клієнт-серверної архітектури для обліку реалізації товарів в торгівлі. Матеріали XI Міжнародної науково-практичної конференції молодих учених та студентів „Актуальні задачі сучасних технологій“, 126-126.

51 Бідюк, О., & Марценко, С. (2025). Методи та засоби інформаційної безпеки іт інфраструктур. Herald of khmelnytskyi national university. Technical sciences, 347(1), 47-58.

52 Волович, В., Береженко, Б. М., & Боднарчук, І. О. (2022). Задача проєктування програмної архітектури в процесах забезпечення якості.

Матеріали X науково-технічної конференції „Інформаційні моделі, системи та технології “ТНТУ імені Івана Пулюя, 104-106.

53 Семенюк, В. О., & Литвиненко, Я. В. (2023). Огляд методів захисту текстової інформації. Матеріали XI науково-технічної конференції „Інформаційні моделі, системи та технології “, 112-112.

54 Козак, С., Микитишин, А., & Станько, А. (2025). Оптимізація зберігання та обробки даних для іот-систем екологічного моніторингу. Measuring and computing devices in technological processes, (1), 323-330.

55 Sverstiuk, A., Matiichuk, L., Polyvana, U., Stanko, A., & Nykytyuk, V. (2024). Analytical analysis of approaches to assessing the quality of life in smart cities.

56 Безпека в надзвичайних ситуаціях. Методичний посібник для здобувачів освітнього ступеня «магістр» всіх спеціальностей денної та заочної (дистанційної) форм навчання / укл.: Стручок В. С. Тернопіль: ФОП Паляниця В. А., 2022. 156 с.

ДОДАТКИ

Фрагмент реалізації алгоритму мовою Python

```

# Список серверів із їх максимальною ємністю (наприклад, CPU)
servers = [ {"capacity": 100, "vms": []} for _ in range(5) ]

# Початкова розподілена кількість VM (із випадковими вимогами до CPU)
vm_requirements = [20, 10, 15, 30, 25, 5, 10, 15, 20, 10] # 10 VM

# Розміщення VM на серверах по черзі (для прикладу)
for i, req in enumerate(vm_requirements):
    servers[i % len(servers)]["vms"].append(req)
LOW_THRESHOLD = 0.3 # 30%
HIGH_THRESHOLD = 0.8 # 80%
def utilization(server):
    """Обчислює завантаження сервера (доля від capacity)."""
    used = sum(server["vms"])
    return used / server["capacity"]
# Виконуємо один цикл перевірки і міграції
for server in servers:
    util = utilization(server)
    if util < LOW_THRESHOLD and server["vms"]:
        # Міграція всіх VM з недовантаженого сервера
        for vm in list(server["vms"]):
            # Знаходимо сервер для переміщення VM
            target = min(
                (s for s in servers if s is not server),
                key=lambda s: utilization(s)
            )
            # Переміщуємо VM
            if sum(target["vms"]) + vm <= target["capacity"]:
                server["vms"].remove(vm)
                target["vms"].append(vm)
        # Сервер можна вимкнути після міграції (список VM
        # порожній)
        # print(f"Server {server} is turned off")
    elif util > HIGH_THRESHOLD:
        # Міграція найдовшої VM з перевантаженого сервера
        largest_vm = max(server["vms"])
        target = min(
            (s for s in servers if s is not server),
            key=lambda s: utilization(s)
        )
        if sum(target["vms"]) + largest_vm <= target["capacity"]:
            server["vms"].remove(largest_vm)
            target["vms"].append(largest_vm)
        # print(f"Migrated VM of size {largest_vm} from overloaded
server")

```