УДК.681.12

В.О.Віталійович[1]

Л.В.Олександрович[2]

Оптимізація мережевої взаємодії в розподілених системах реального часу на прикладі чат-ботів

[1] Вінницький національний технічний університет;

Анотація

У роботі досліджено методи оптимізації мережевої взаємодії у високонавантажених розподілених чат-платформах реального часу. Проведено порівняльний аналіз традиційних HTTP-підходів із неблокуючими механізмами передачі даних. Розглянуто застосування протоколу WebSocket для двостороннього зв'язку, Unix Domain Sockets для локальної міжпроцесної взаємодії та бінарної серіалізації Protocol Buffers у межах асинхронної event-driven архітектури. Результати моделювання доводять, що інтеграція цих технологій істотно зменшує затримки, скорочує мережеві накладні витрати та оптимізує використання ресурсів сервера. Також приділено увагу питанням масштабованості, відмовостійкості та інформаційної безпеки систем.

Ключові слова: розподілені системи, чат-боти, WebSocket, Unix Domain Sockets, Protocol Buffers, event-driven architecture, low latency, IPC.

Abstract

This paper investigates methods for optimising network interaction in high-load distributed real-time chat platforms. A comparative analysis of traditional HTTP approaches with non-blocking data transfer mechanisms is presented. The use of the WebSocket protocol for bidirectional communication, Unix Domain Sockets for local inter-process communication, and binary serialization with Protocol Buffers within an asynchronous event-driven architecture is explored. Simulation results demonstrate that integrating these technologies significantly reduces latency, lowers network overhead, and optimises server resource usage. Attention is also given to issues of scalability, fault tolerance, and information security of the systems.

Keywords: distributed systems, chatbots, WebSocket, Unix Domain Sockets, Protocol Buffers, event-driven architecture, low latency, IPC.

Вступ

Інтеграція чат-ботів з NLP та LLM вимагає створення масштабованих клієнт-серверних архітектур із мінімальною затримкою. Це є особливо критичним для реалізації механізмів потокової передачі даних (streaming) під час генерації відповідей сучасними великими мовними моделями, де мережева затримка безпосередньо впливає на користувацький досвід (UX). Стандартна інфраструктура таких платформ містить клієнтську частину, балансувальник, gateway та мікросервіси бізнес-логіки. Традиційна HTTP-модель неефективна під високим навантаженням через постійне перевстановлення TCP-з’єднань, надсилання службових заголовків та блокування потоків введення-виведення, що обмежує масштабованість системи. Ефективною альтернативою є перехід до асинхронної event-driven архітектури, яка дозволяє обслуговувати велику кількість одночасних з’єднань із мінімальними витратами ресурсів.

Актуальність

Під час масштабування розподіленої системи та зростання кількості активних користувачів класичні HTTP-механізми формують надлишкове мережеве навантаження. HTTP Polling вимагає

регулярного повторного надсилання запитів для перевірки готовності даних, що створює службовий трафік та збільшує навантаження на CPU і RAM серверного вузла. Додатковою проблемою є накопичення TCP-з’єднань та збільшення навантаження на пул файлових дескрипторів операційної системи. Це знижує ефективність використання обчислювальних ресурсів і обмежує масштабованість сервісу.

Таблиця 1. Порівняння архітектурних підходів

Технологія / Протокол	Напрямок передачі	Накладні витрати (Overhead)	Механізм утримання з'єднання	Ефективність використання CPU/RAM
HTTP Polling	Односторонній (Клієнт -Сервер)	Критичні (Повторні HTTP-заголовки, TCP-handshake)	Відсутній (Постійні нові запити)	Низька (Забиває пул дескрипторів та CPU)
AJAX Long Polling	Напівдвосторонній (Емуляція)	Високі (Утримання HTTP-сесії, реконнекти)	Очікування відповіді сервера (Timeout)	Середня (Високе навантаження на RAM сервера)
WebSocket (JSON)	Повнодвосторонній (Duplex)	Низькі (Мінімальний фрейм, один handshake)	Постійне TCP-з'єднання (Keep-Alive)	Висока (Асинхронний неблокуючий I/O)
WebSocket + Protobuf + UDS	Повнодвосторонній (Оптимізований)	Мінімальні (Бінарний формат, без мережевого стеку в IPC)	Локальний Unix-сокет + WSS-сесія	Максимальна (Ефективна серіалізація, low latency)

Основні задачі

дослідження зосереджені на комплексному аналізі та оптимізації асинхронної мережевої взаємодії для забезпечення високоефективної двосторонньої комунікації в реальному часі, мінімізації мережевих накладних витрат шляхом раціоналізації протоколів передачі даних, модернізації механізмів міжпроцесної взаємодії (IPC) з метою зниження латентності обміну інформацією, гарантуванні горизонтальної та вертикальної масштабованості архітектури в умовах інтенсивного зростання навантажень, а також на підвищенні інформаційної безпеки системи через впровадження концепції нульової довіри (zero-trust), сучасних методів криптографічного захисту та механізмів автентифікації на всіх рівнях функціонування платформи.

Шляхи вирішення

Запропонована архітектура розподіленої системи організована на трьох рівнях: локальному (міжпроцесна взаємодія IPC через високопродуктивні Unix Domain Sockets), транспортному (повнодуплексна асинхронна WebSocket-комунікація для обміну даними в реальному часі) та кластерному (динамічне балансування навантаження та координація мікросервісів для забезпечення горизонтального масштабування).

Ключовими технологіями, що забезпечують ефективне вирішення поставлених задач у межах асинхронної event-driven моделі, є використання неблокуючого введення-виведення (Non-blocking I/O), яке мінімізує навантаження на обчислювальні ресурси сервера, та впровадження компактних протоколів передачі даних для зниження латентності обміну інформацією.

Додатково застосовується предиктивне управління чергами повідомлень для оптимізації потоків даних, а також сучасні засоби зворотного проксіювання, що гарантують високу відмовостійкість і доступність сервісів. Надійний захист інформаційних каналів на всіх рівнях функціонування платформи забезпечується інтеграцією архітектури безпеки нульової довіри (zero-trust) із механізмами взаємної автентифікації (mTLS) та наскрізного шифрування.

Рисунок 1. Схема архітектури оптимізованої розподіленої чат-платформи реального часу.

Результати моделювання. Для оцінки ефективності запропонованої архітектури було проведено навантажувальне тестування з використанням інструментарію Apache JMeter.

Таблиця 2. Результати benchmark

Архітектура	Avg latency	CPU	Connections
HTTP Polling	430 ms	78%	~1200
Long Polling	210 ms	61%	~3000
WebSocket+ JSON	95 ms	44%	~8500
WebSocket + Protobuf + UDS	58 ms	37%	>12000

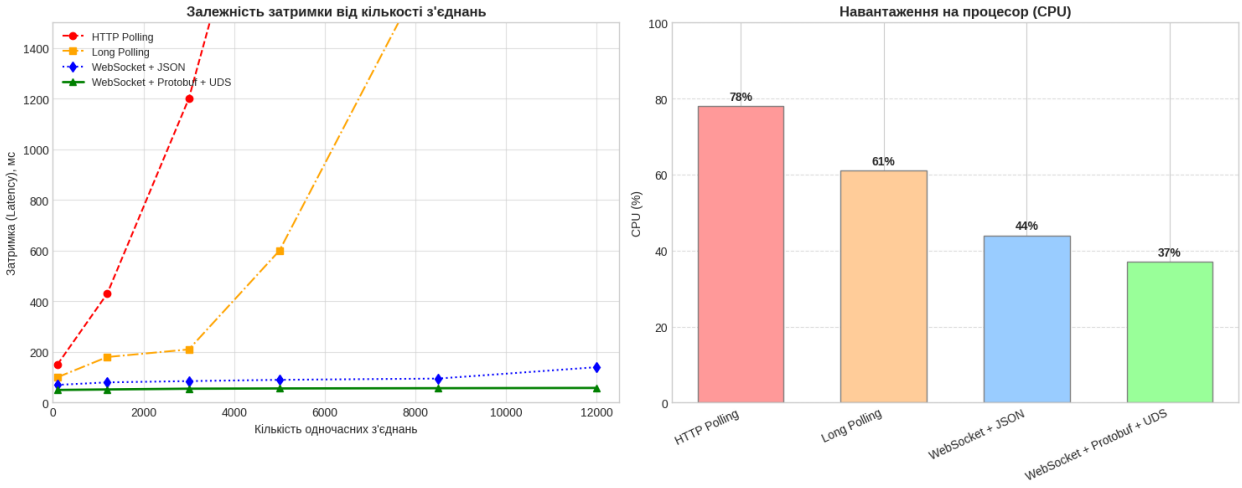

Рисунок 2. Графік результатів benchmark, навантаження на CPU

Шляхи забезпечення безпеки системи**.** Комплексний захист транспортного рівня базується на багаторівневій архітектурі, що включає наскрізне TLS-шифрування, динамічний Rate Limiting для протидії DDoS-атакам, сувору валідацію вхідних даних проти ін'єкцій та моніторинг аномалій у реальному часі. Проєктування системи відповідно до стандарту OWASP ASVS гарантує її високу стійкість до поширених уразливостей і надійний захист інформаційних потоків.

Висновки

Проведене дослідження доводить високу доцільність та перспективність інтеграції WebSocket, Unix Domain Sockets, Protocol Buffers та event-driven архітектури при розробці сучасних високонавантажених систем реального часу. Отримані результати підтверджують, що використання неблокуючих механізмів обміну даними дозволяє вирішити проблему надлишкового мережевого трафіку, притаманного класичним HTTP-моделям, забезпечуючи гарантоване зниження затримок передачі повідомлень, безперешкодне горизонтальне масштабування та оптимізацію споживання ресурсів серверних вузлів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Leshchenko Yu., Yukhimchuk M., Lesko V., Ivanov Yu. Integrating Clustering and Artificial Intelligence for Improved Efficiency in Last-Mile Logistics. Measuring and Computing Devices in Technological Processes. 2025. Vol. 84 (4). pp. 346-350. https://doi.org/10.31891/2219-9365-2025-84-41.

2. Юхимчук М.С., Лесько В.О., Дубовой В.М., Іванов Ю.Ю. Інтелектуальна система автоматичного керування процесом сушіння зернових культур на основі IoT-технологій. Наукові праці ВНТУ. Вінниця: ВНТУ, 2025. №4. С. 1-8. https://doi.org/10.31649/2307-5376-2025-4-46-53.

3. Development and Research of the Hardware and Software Architecture of an IoT-Node for Monitoring Technological Parameters Based on Nodemcu V3 and Prometheus / M.S. Yukhymchuk, V.O. Lesko, Yu.Yu. Ivanov, P.P. Strembitskiy. Measuring Technology and Metrology. Lviv: Lviv Polytechnic National University, 2026. Issue 87, № 1. pp. 59-62. https://doi.org/10.23939/istcmtm2026.01.059.

4. Проєктування системи автоматичного управління технологічним процесом сушіння зерна / М.С. Юхимчук, В.О. Лесько, Ю.Ю. Іванов, Ю.А. Горчук, О.В. Климчук. Наукові праці ВНТУ. Вінниця: ВНТУ, 2026. № 1. С. 1-17.

Валько́в Олександр Віталійович — студент групи 2ПКТ-24б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: valkovalexsander4@gmail.com;

Лесько Владислав Олександрович — канд. техн. наук, доцент кафедри електричних станцій і систем, Вінницький національний технічний університет, e-mail: leskovlad@ukr.net;

Науковий керівник: ***Юхимчук Марія Сергіївна*** — д-р техн. наук, професор кафедри комп'ютерних систем управління, Вінницький національний технічний університет, e-mail: umcmasha@gmail.com;

Oleksandr V. Vitaliyovych — student of group 2PKT-24b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia. *E-mail:* valkovalexsander4@gmail.com;

Vladyslav O. Lesko – Ph.D. in Technical Sciences, Associate Professor of the Department of Electric Stations and Systems, Vinnytsia National Technical University, e-mail: leskovlad@ukr.net;

Supervisor: ***Mariia S. Yukhymchuk*** – Doctor of Technical Sciences, Professor of the Department of Computer Control Systems, Vinnytsia National Technical University, e-mail: umcmasha@gmail.com;