

МОНІТОРИНГ ТА ПРОГНОЗУВАННЯ НАВАНТАЖЕННЯ СЕРВЕРУ ІР-ТЕЛЕФОНІЇ У ЦИФРОВИХ НАВЧАЛЬНИХ ЕКОСИСТЕМАХ

А.В. Хижняк¹, А.Ю. Милиця², П.Г. Бивойно^{* 3}, О.В. Заворотний

ORCIDs: [0009-0008-8655-3736](https://orcid.org/0009-0008-8655-3736)¹; [0009-0003-5026-0601](https://orcid.org/0009-0003-5026-0601)²; [0000-0001-8145-8459](https://orcid.org/0000-0001-8145-8459)³

Національний університет "Чернігівська Політехніка", Чернігів

* Листування: p.g.byvoino@gmail.com

Open Access under [CC BY-NC 4.0](https://creativecommons.org/licenses/by-nc/4.0/) License

UDC 004.7:004.032.26

Monitoring and Load Forecasting of IP Telephony Server in Digital Educational Ecosystems

Andrii Khyzhniak, Artem Mylytsia, Pavlo Byvoino^{*}, Oleksandr Zavorotnyi

Chernihiv Polytechnic National University, Chernihiv, Ukraine

* Correspondence: p.g.byvoino@gmail.com

Introduction. The article addresses the relevant scientific and practical problem of developing and substantiating a subsystem for intelligent server monitoring and load forecasting in IP telephony networks. It highlights the critical importance of transitioning from traditional reactive observation of telecommunication nodes to proactive infrastructure management. This is particularly vital for digital educational ecosystems, where the time series of telephone traffic exhibit a complex structure with pronounced daily and weekly seasonality, directly influenced by academic schedules, consultations, and examination periods.

The purpose of the paper is to develop a software module utilizing the Holt–Winters method (triple exponential smoothing) for short-term traffic forecasting.

Results. A detailed comparative analysis of existing monitoring tools (Zabbix, Prometheus, Homer) was conducted, identifying a significant gap between hardware resource tracking (CPU, RAM) and specific SIP signaling parameters.

It was established that most systems focus on historical data analysis, lacking tools for predicting future peak loads.

The scientific novelty lies in the experimental determination of the optimal model parameters ($\alpha = 0.2$; $\beta = 0.05$; $\gamma = 0.1$), ensuring high stability against anomalous spikes and noise typical of VoIP networks. It is demonstrated that using this method achieves a MAPE accuracy of $< 15\%$ without the need for resource-intensive neural network architectures.

The practical value of the results involves integrating the developed algorithm into the modern VictoriaMetrics and Grafana technology stack. This solution automates the detection of overload risks and ensures the stability of communication services during peak periods of educational activity.

Conclusions. The findings establish a methodological foundation for the further implementation of decision support systems within the IT infrastructure of educational institutions.

Keywords: IP telephony, server monitoring, load forecasting, Holt–Winters method, time series, VictoriaMetrics, Grafana, Asterisk.

Анотація

Вступ. У статті розв'язано актуальну науково-практичну задачу розроблення та обґрунтування підсистеми інтелектуального моніторингу серверів і прогнозування навантаження в мережах ІР-телефонії. Обґрунтовано доцільність переходу від традиційного реактивного підходу до спостереження за станом телекомунікаційних вузлів до проактивного управління інфраструктурою. Це є особливо важливим для цифрових навчальних екосистем, де часові ряди телефонного трафіку мають складну структуру з вираженою добовою та тижневою сезонністю, зумовленою розкладом занять, консультацій та екзаменаційних періодів.

Мета роботи. Метою дослідження є розробка програмного модуля, що використовує метод Холта–Вінтерса (потрійне експоненційне згладжування) для короткострокового передбачення трафіку.

Результати. Проведено детальний порівняльний аналіз існуючих інструментів (Zabbix, Prometheus, Homer) та виявлено проблему розриву між моніторингом апаратних ресурсів (CPU, RAM) та специфічних параметрів SIP-сигналізації. Встановлено, що більшість систем фокусуються на аналізі історичних подій, не надаючи інструментів для виявлення майбутніх пікових навантажень.

Наукова новизна полягає в експериментальному визначенні оптимальної конфігурації параметрів моделі ($\alpha = 0,2$; $\beta = 0,05$; $\gamma = 0,1$), що забезпечує високу стійкість до аномальних сплесків та шумів, характерних для VoIP-мереж. Доведено, що використання даного методу дозволяє досягти точності $MARE < 15\%$ без залучення ресурсоемних нейромережевих архітектур.

Практичне значення результатів полягає в інтеграції розробленого алгоритму в сучасний стек VictoriaMetrics та Grafana. Таке рішення дозволяє автоматизувати процес виявлення ризиків перевантаження та забезпечити стабільність комунікаційних сервісів у пікові періоди освітньої активності.

Висновки. Отримані результати формують методичну основу для подальшого впровадження систем підтримки прийняття рішень в IT-інфраструктурах закладів освіти.

Ключові слова: IP-телефонія, моніторинг серверів, прогнозування навантаження, метод Хольта–Вінтерса, часові ряди, VictoriaMetrics, Grafana, Asterisk.

Вступ. В умовах стрімкої цифровізації освітнього процесу та переходу до змішаних форм навчання суттєво зростає роль IP-телефонії як засобу комунікації в корпоративних мережах та освітніх екосистемах. Забезпечення стабільності та безперервності роботи серверів телефонії (зокрема на базі платформи Asterisk) [1] стає пріоритетним завданням, оскільки навіть незначні затримки передавання голосу або збої в маршрутизації викликів призводять до порушення навчальних процесів. Існуючі підходи до моніторингу переважно обмежуються реактивним спостереженням за поточним станом, що не дозволяє адміністраторам завчасно реагувати на пікові навантаження та запобігати відмовам в обслуговуванні. Характер телефонного трафіку в цифрових освітніх екосистемах відрізняється вираженою добовою та тижневою сезонністю, що зумовлена розкладом занять, дистанційними консультаціями та екзаменаційними сесіями.

Сучасний ринок програмного забезпечення пропонує широкий спектр інструментів моніторингу, які можна розділити на дві категорії: системи загального призначення (Nagios, Zabbix, Prometheus [2–4]) та спеціалізовані VoIP-інструменти (Homer) [5]. Системи загального призначення ефективно відстежують ресурси (CPU, RAM), але не враховують специфіку SIP-сигналізації. Спеціалізовані засоби надають детальну аналітику якості дзвінків (RTP, jitter), але ігнорують стан апаратних ресурсів сервера. Головною проблемою є відсутність єдиного інтегрованого рішення, яке б поєднувало збір різномірних метрик із можливостями прогнозування навантаження. Більшість існуючих систем фокусуються на аналізі історичних даних ("що сталося"), а не на предиктивну аналітику ("що станеться"), що ускладнює реалізацію проактивного керування інфраструктурою.

Питання прогнозування часових рядів і моніторингу IT-інфраструктури розглядаються у працях багатьох закордонних та вітчизняних вчених. Теоретичні основи методів прогнозування, зокрема експоненційного згладжування, висвітлено у роботах Р. Хіндмана та Дж. Бокса [6, 7]. Особливості використання Prometheus та Grafana для моніторингу детально описано Дж. Тернбуллом [8]. Архітектуру та налаштування Asterisk розглянуто у працях Р. Брайанта [9]. Водночас, аналіз наукових джерел свідчить, що питання інтеграції методу Хольта – Вінтерса безпосередньо у pipeline моніторингу контейнеризованих VoIP-систем для прогнозування resource-intensive процесів (таких як транскодинг) залишається недостатньо дослідженим.

Запропоновані в [10] підходи до побудови модульних аналітичних компонентів використані як концептуальну основу для розроблення підсистеми моніторингу та прогнозування навантаження серверів IP-телефонії.

Виділення недосліджених частин загальної проблеми. Попри наявність потужних математичних моделей прогнозування (ARIMA, LSTM), їх застосування в реальному часі для моніторингу телефонії часто є надлишковим з точки зору обчислювальних ресурсів або складним у впровадженні.

Невирішеною залишається задача створення легкої та адаптивної системи, здатної поєднувати метрики операційної системи та бізнес-метрики телефонії (кількість дзвінків) для побудови короткострокових прогнозів із урахуванням добової та тижневої сезонності, характерної для навчальних закладів.

Мета дослідження. Мета роботи – розробка та експериментальна перевірка підсистеми моніторингу та прогнозування навантаження на сервери IP-телефонії. Досягнення мети передбачає реалізацію програмного модуля, який використовує метод потрійного експоненційного згладжування (Хольта – Вінтерса) для передбачення пікових навантажень і візуалізації результатів у середовищі Grafana.

Наукова новизна та методичний внесок роботи полягають у тому, що вперше експериментально визначено збалансовану конфігурацію параметрів методу потрійного експоненційного згладжування Хольта – Вінтерса ($\alpha = 0,2$; $\beta = 0,05$; $\gamma = 0,1$), яка забезпечує компроміс між чутливістю моделі до змін навантаження та стійкістю до шуму й аномальних сплесків, характерних для телефонного трафіку в цифрових навчальних екосистемах. Отримана параметризація розглядається не як емпіричний тюнінг, а як методично обґрунтований результат, підтверджений порівняльним аналізом точності прогнозування. Також методично обґрунтована інтеграція класичного методу потрійного експоненційного згладжування (Хольта – Вінтерса) безпосередньо в pipeline системи моніторингу IP-телефонії, що дозволяє поєднати збір метрик у реальному часі з короткостроковим прогнозуванням навантаження без застосування ресурсоемних моделей машинного навчання.

Виклад основного матеріалу. Методологічною основою дослідження є порівняльний аналіз методів прогнозування часових рядів, що застосовуються для оцінювання та передбачення навантаження в інформаційних і телекомунікаційних системах. У працях вітчизняних авторів розглянуто класичні статистичні та адаптивні підходи до прогнозування, зокрема методи ковзних середніх, авторегресійні моделі та методи експоненційного згладжування, а також проаналізовано їхні переваги й обмеження з точки зору точності, стійкості до шуму та обчислювальної складності [11, 12]. Показано, що для практичних задач з вираженою сезонністю та обмеженими обчислювальними ресурсами доцільним є використання адаптивних методів експоненційного згладжування, які забезпечують баланс між точністю прогнозу та простотою реалізації [13].

З огляду на специфіку оброблення великих обсягів часових даних у розподілених і моніторингових системах, у дослідженні також враховано підходи до збору, агрегування та аналітичної обробки даних у сучасних інформаційних інфраструктурах [14, 15]. Методи інтелектуального аналізу даних розглядаються як потенційне розширення базової методології, однак у межах даної роботи вони не застосовуються через підвищену складність та ресурсоемність у сценаріях реального часу [16].

Архітектура запропонованої системи моніторингу. Для розв'язання задачі прогнозування навантаження було розроблено дворівневу архітектуру, що забезпечує логічне розмежування операційного трафіку телефонії та трафіку системи моніторингу (рис. 1).

Ключовим елементом системи є сервер IP-телефонії на базі Asterisk, який генерує детальні записи про дзвінки (CDR) та зберігає їх у реляційній базі даних MariaDB. Паралельно функціонує розроблений на базі фреймворку FastAPI (Python) API модуль, що здійснює збір системних метрик (CPU, RAM) і зберігає їх у документо-орієнтованій базі MongoDB. Така гетерогенна структура баз даних забезпечує ефективну роботу як зі структурованими даними дзвінків, так і з часовими рядами системного навантаження.

Для зберігання та візуалізації часових рядів обрано стек VictoriaMetrics [17] та Grafana [18]. VictoriaMetrics забезпечує високу продуктивність запису та сумісність із протоколом Prometheus, що дозволяє зберігати як фактичні, так і прогнозовані значення метрик у єдиному сховищі [4].

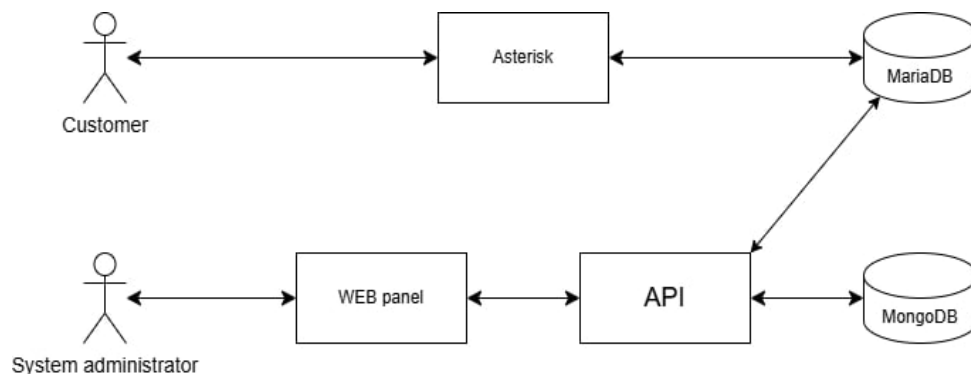


РИС. 1. Архітектура запропонованої системи моніторингу

Математична модель прогнозування. Експоненційне згладжування розглядається як один із базових і водночас ефективних підходів до прогнозування часових рядів, що підтверджується класичними дослідженнями в цій галузі [19]. Адаптивні модифікації цього методу широко застосовуються для прогнозування навантаження в системах із вираженою циклічністю, зокрема в інфраструктурних і сервісних середовищах [20].

З урахуванням циклічної природи телефонного трафіку (добова та тижнева сезонність), для прогнозування обрано метод потрібного експоненційного згладжування Хольта – Вінтерса (Holt–Winters), який дозволяє врахувати рівень $L_t = \alpha(y_t - S_{t-m}) + (1 - \alpha)(L_{t-1} + T_{t-1})$, тренд $T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}$ і сезонну компоненту часового ряду $S_t = \gamma(y_t - L_t) + (1 - \gamma)S_{t-m}$.

Рівняння прогнозу має вигляд:

$$\tilde{y}_{t+h} = L_t + h * T_t + S_{t+h-m},$$

де h – горизонт прогнозування; m – довжина сезонного циклу; α , β , γ – параметри згладжування.

Вибір добової сезонності ($m = 24$) зумовлений специфікою функціонування IP-телефонії в цифрових освітніх екосистемах, де характер навантаження визначається розкладом занять, дистанційними консультаціями та іншими синхронними активностями. Така структура трафіку формує повторювані добові цикли з вираженими піками в робочі години та зниженим навантаженням у нічний період, що робить використання фіксованої добової сезонності методично обґрунтованим.

Експериментальний підбір параметрів моделі. Одним із ключових етапів дослідження став підбір коефіцієнтів згладжування: α (для рівня), β (для тренду) та γ (для сезонності). З цією метою проведено серію експериментів на реальних даних CDR, спрямованих на досягнення балансу між чутливістю моделі до змін та стійкістю до шуму. Результати одного із таких експериментів показано на рис. 2. В таблиці наведено результати оцінювання впливу різних комбінацій коефіцієнтів згладжування на статистичні характеристики відхилень актуальних даних від прогнозованих.

Аналіз результатів проведених експериментів дозволив зробити такі висновки:

- консервативна модель ($\alpha = 0,1$; $\beta = 0,01$; $\gamma = 0,1$) показала найменшу чутливість до раптових сплесків, але мала значну затримку у визначенні тренду (MAPE $\approx 13,9$ %);
- агресивна модель ($\alpha = 0,7$; $\beta = 0,2$; $\gamma = 0,05$) швидко реагувала на зміни, однак давала хибні прогнози через перенавчання на випадковому шумі (MAPE $\approx 17,5$ %);
- збалансована модель ($\alpha = 0,2$, $\beta = 0,05$, $\gamma = 0,1$) забезпечила оптимальне співвідношення точності та стабільності (MAPE $\approx 14,6$ %). Саме цей набір параметрів було імплементовано у фінальну версію системи.

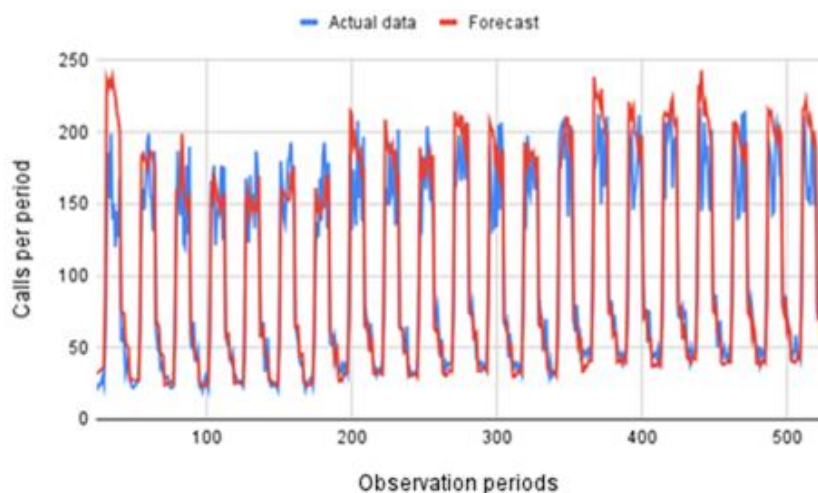


РИС. 2. Порівняння графіків прогнозу з актуальними даними для коефіцієнтів згладжування $\alpha = 0,2$; $\beta = 0,05$; $\gamma = 0,1$

ТАБЛИЦЯ. Результати обчислення статистичних метрик прогнозування для різних комбінацій коефіцієнтів згладжування

Variant(α, β, γ)	MAPE %	Median APE %	Share % (APE > 25)	RMSE
$\alpha = 0,1$; $\beta = 0,01$; $\gamma = 0,1$	13,88	11,80	13,07	17,15
$\alpha = 0,1$; $\beta = 0,01$; $\gamma = 0,05$	14,13	12,24	14,39	17,39
$\alpha = 0,2$; $\beta = 0,05$; $\gamma = 0,1$	14,59	11,89	16,48	17,83
$\alpha = 0,2$; $\beta = 0,2$; $\gamma = 0,2$	17,03	13,43	22,54	19,23
$\alpha = 0,7$; $\beta = 0,2$; $\gamma = 0,05$	17,53	13,77	23,67	21,69

Дослідження поведінки системи при аномаліях. Окрему увагу приділено аналізу реакції моделі на аномальні події, такі як DDoS-атаки або технічні збої (пропуски даних). Експериментально встановлено, що агресивні налаштування ($\alpha > 0,5$) призводять до ефекту "positive bias", коли модель сприймає атаку як зміну тренду і необґрунтовано завищує прогнози на майбутні періоди (рис. 3).

Для мінімізації цього ефекту в системі реалізовано механізм попередньої обробки даних, який передбачає фільтрацію екстремальних викидів перед подачею даних на вхід алгоритму прогнозування.

Візуалізація результатів. Завершальним етапом роботи стала інтеграція розробленого API з дашбордами Grafana. Це дозволило адміністраторам у режимі реального часу спостерігати кореляцію між кількістю активних дзвінків і завантаженням CPU/RAM (рис. 4).

Аналіз візуалізації моніторингу в Grafana підтвердив, що прогнозовані значення використання RAM мають високу точність завдяки стабільному тренду, тоді як прогноз CPU ефективно згладжує миттєві піки, надаючи оцінку середнього очікуваного навантаження.

Обмеження дослідження. Проведене дослідження має низку обмежень, які необхідно врахувати при інтерпретації отриманих результатів та їх практичному застосуванні. Зокрема:

- обмеження математичної моделі. Метод Хольта – Вінтерса орієнтований на часові ряди з відносно стабільною сезонною структурою та поступовими змінами тренду. У випадках різких структурних зламів, тривалих аномалій або нерегулярної сезонності точність прогнозування може знижуватися;

– обмеження даних. Модель навчалася та тестувалася на даних, отриманих з конкретної інфраструктури IP-телефонії (Asterisk), що накладає обмеження на безпосереднє перенесення результатів на системи з іншими характеристиками трафіку або архітектурою.

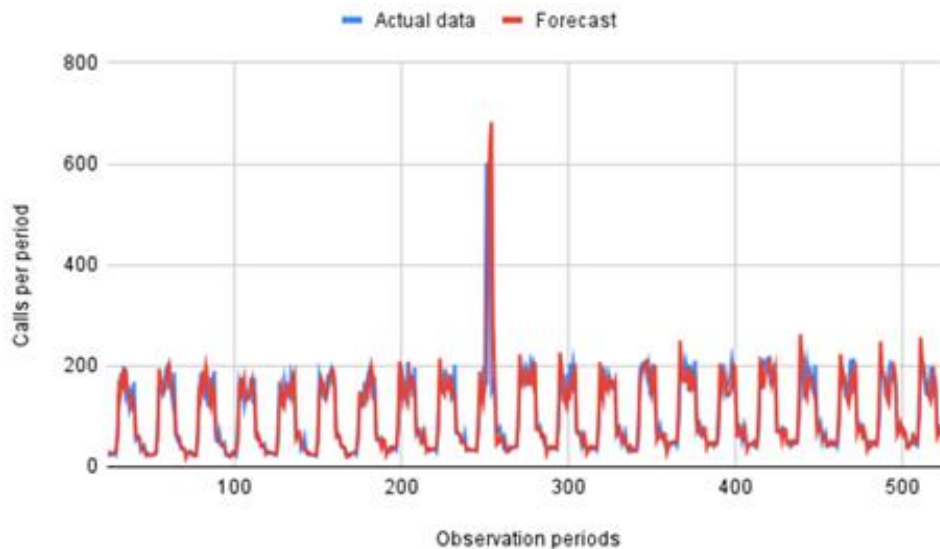


РИС. 3. Демонстрація зниження точності прогнозування при імітації мережевої атаки (різкий пік навантаження)

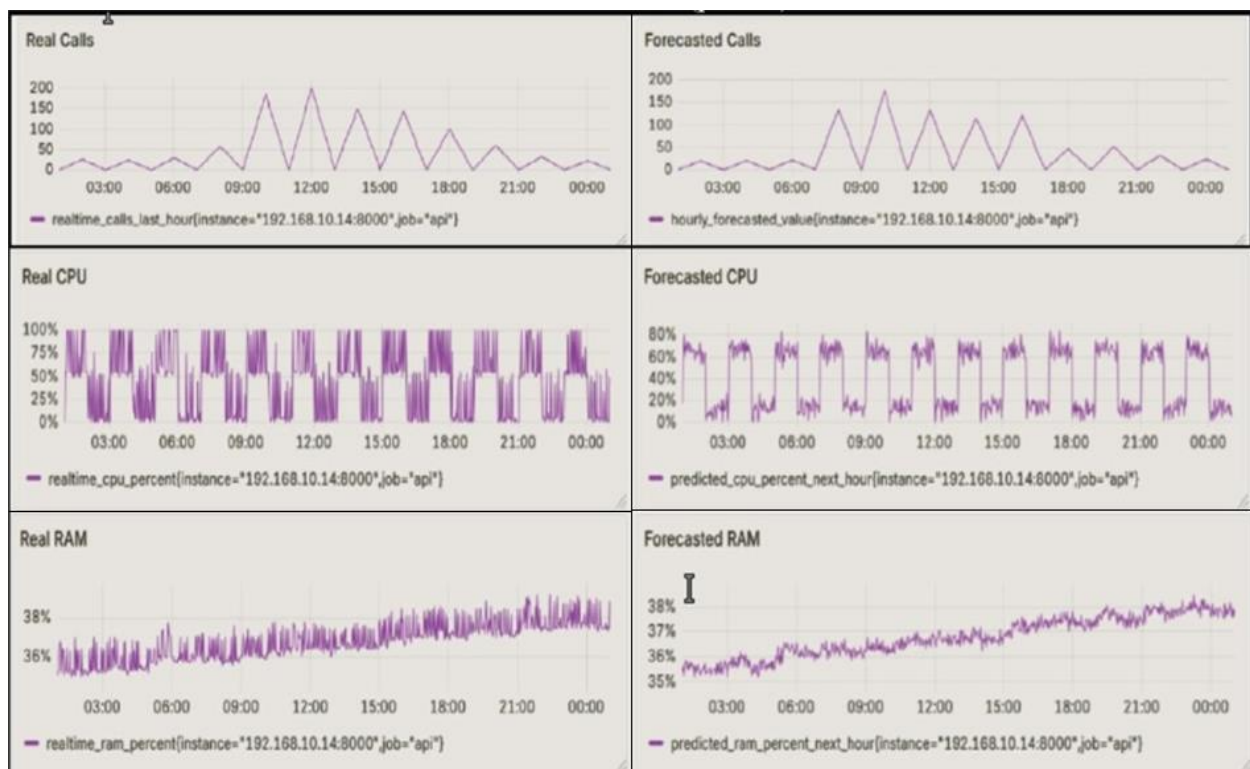


РИС. 4. Панель візуалізації моніторингу в Grafana: порівняння реальних (ліворуч) та прогнозованих (праворуч) метрик

Висновки. У статті розв’язано науково-практичну задачу розроблення та обґрунтування підсистеми інтелектуального моніторингу та прогнозування навантаження на сервери ІР-телефонії, орієнтованої на умови функціонування цифрових навчальних екосистем.

Досягнення поставленої мети забезпечено шляхом поєднання збору системних і бізнес-метрик у реальному часі з методами аналізу часових рядів, що дозволяє перейти від реактивного моніторингу до проактивного управління інфраструктурою.

Отримано нові результати щодо поведінки експоненційних моделей прогнозування в умовах аномальних подій (пропуски даних, пікові навантаження, імітація атак), які демонструють залежність стабільності прогнозу від швидкості оновлення компонентів рівня, тренду та сезонності. Використання методу Хольта – Вінтерса з експериментально підібраними параметрами ($\alpha = 0,2$; $\beta = 0,05$; $\gamma = 0,1$) дозволило досягти прийнятної точності прогнозування ($MAPE < 15\%$) для добових циклів навантаження. Інтеграція прогнозування в стек технологій Docker/VictoriaMetrics/Grafana дозволяє трансформувати процес адміністрування з реактивного в проактивний, завчасно виявляючи ризики вичерпання ресурсів. Запропонований підхід є особливо релевантним для цифрових навчальних екосистем, де прогнозування навантаження забезпечує стабільність комунікаційної інфраструктури під час пікових освітніх активностей.

Практичне значення отриманих результатів полягає у можливості застосування розробленого підходу для підтримки стабільної роботи серверів ІР-телефонії в цифрових навчальних екосистемах, де навантаження визначається розкладом занять, дистанційними консультаціями та сесійними періодами. Інтеграція прогнозованих метрик у середовище Grafana створює основу для завчасного виявлення ризиків перевантаження ресурсів і підвищує ефективність управління освітньою ІТ-інфраструктурою.

Отримані результати формують методичну основу для подальших досліджень у напрямі контекстно-адаптивного прогнозування навантаження та узагальнення запропонованого підходу на інші сервіси цифрових навчальних екосистем.

Перспективи подальших досліджень пов’язані з розширенням запропонованого підходу в кількох напрямках:

- по-перше, доцільним є включення до моделі прогнозування додаткових контекстних факторів, характерних для цифрових навчальних екосистем, зокрема академічного календаря, розкладу занять і сесійних періодів, що дозволить перейти від фіксованої сезонності до контекстно-адаптивного прогнозування;

- по-друге, перспективним є розвиток механізмів обробки аномалій шляхом поєднання експоненційного згладжування з процедурами попередньої фільтрації викидів або гібридними моделями, що підвищить стійкість прогнозу в умовах атак або нестабільних мережевих подій;

- по-третє, подальші дослідження можуть бути спрямовані на узагальнення запропонованого підходу для інших сервісів цифрових навчальних екосистем (відеоконференції, системи дистанційного навчання), що дозволить сформулювати єдину методичну основу прогнозування навантаження в освітніх ІТ-інфраструктурах.

Заява про використання генеративного штучного інтелекту. У процесі підготовки рукопису автори застосували інструменти на основі штучного інтелекту (зокрема, Gemini 3 Pro) з метою усунення стилістичних та граматичних помилок. Увесь згенерований або модифікований контент був ретельно перевірений, відредагований і схвалений авторами, які несуть повну відповідальність за остаточний зміст публікації.

Авторські внески: Хижняк В.В. – концептуалізація, методологія, формальний аналіз; Милиця А.Ю. – програмне забезпечення, написання – оригінальна чернетка; Бивойно П.Г. – узагальнення, візуалізація, редагування, ресурси; Задорожний А.В. – дослідження.

Наявність даних. Дані, що підтверджують результати цього дослідження, доступні за запитом у автора (Хижняк В.В.) через електронну пошту alf.andrey@stu.cn.ua.

Фінансування. Автори не отримували фінансування для проведення досліджень та написання статті.

Список літератури

1. Asterisk Documentation. <https://www.asterisk.org> (звернення: 25.01.2025)
2. Nagios Documentation. <https://www.nagios.org> (звернення: 25.01.2025)
3. Zabbix Documentation. <https://www.zabbix.com> (звернення: 25.01.2025)
4. Prometheus Documentation. <https://prometheus.io> (звернення: 25.01.2025)
5. Homer Documentation. <https://sipcapture.org> (звернення: 25.01.2025)
6. Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. 2nd ed. Melbourne : OTexts, 2018. <https://otexts.com> (звернення: 20.01.2025).
7. Box G. E. P., Jenkins G. M., Reinsel G. C. Time Series Analysis: Forecasting and Control. 5th ed. Hoboken : Wiley, 2015. 712 p.
8. Turnbull J. Monitoring with Prometheus. Brooklyn: Turnbull Press, 2018. 280 p.
9. Bryant R., Madsen L., Van Meggelen J. Asterisk: The Definitive Guide. 5th ed. Sebastopol: O'Reilly Media, 2019. 850 p.
10. Khyzhniak A.V., Prila O.A. Designing a system for automated generation and automated assessment of parameterized practical assignments. *Technical Sciences and Technologies*. 2023. Vol. 2 (32). P. 124–132.
11. Широкопетлева М.С., Пономаренко А.А., Дудар З.В. Порівняння методів прогнозування часових рядів. *Біоніка інтелекту*. 2018. № 1 (90). С. 91–96.
12. Воробець І. Порівняння методів прогнозування часових рядів. *Матеріали X науково-технічної конференції* (Тернопіль, 2022). Тернопіль: ТНТУ, 2022. С. 23.
13. Ніколенко В.В. Прогнозування часових рядів методами експоненційного згладжування. *Містобудування та територіальне планування*. 2022. Вип. 80. С. 101–106.
14. Олещенко Л.М. Технології оброблення великих даних: навч. посіб. Запоріжжя: ЗНУ, 2021. 120 с.
15. Журавель С., Думич С., Шпур О. Дослідження методів збору та обробки даних у розподілених системах. *Вісник Національного університету «Львівська політехніка»*. Серія: Інформаційні системи та мережі. 2021. Вип. 10. С. 154–162.
16. Гороховський В.О., Творошенко І.С. Методи інтелектуального аналізу та оброблення даних: навч. посіб. Харків: ХНУРЕ, 2023. 184 с.
17. VictoriaMetrics Documentation. <https://victoriametrics.com> (звернення: 25.01.2025)
18. Grafana Documentation. <https://grafana.com> (звернення: 25.01.2025)
19. Gardner E.S. Exponential smoothing: The state of the art. *Journal of Forecasting*. 1985. Vol. 4, No. 1. P. 1–28.
20. Taylor J.W. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*. 2003. Vol. 54, No. 8. P. 799–805.

Received/Одержано 11.02.2026

Accepted/Прийнято 26.05.2026

Published/Надруковано 01.06.2026