

УДК 004.415.2:004.382:004.75

DOI <https://doi.org/10.36994/2788-5518-2025-02-10-07>

МЕТОД ДИНАМІЧНОГО БАЛАНСУВАННЯ ПРОДУКТИВНОСТІ СЕРВЕРІВ У ІНТЕРАКТИВНИХ СИСТЕМАХ

Гнатюк В. О., к.т.н., доцент, Державний університет «Київський авіаційний інститут», Київ, Україна; ДержНДІ технологій кібербезпеки, Київ, Україна. <https://orcid.org/0000-0002-4916-7149>, viktor.hnatiuk@npp.kai.edu.ua

Зандер К. Ю., PhD аспірант, Державний університет «Київський авіаційний інститут», Київ, Україна. <https://orcid.org/0009-0006-4944-9249>, 8390983@stud.kai.edu.ua

Анотація. У статті запропоновано метод динамічного балансування продуктивності серверів в інтерактивних системах, що функціонують в умовах непередбачуваних змін інтенсивності клієнтських звернень. Показано, що традиційні підходи до розподілу навантаження між серверами, орієнтовані на статичні або квазістатичні режими роботи, не забезпечують стабільної якості обслуговування користувачів при високій динамічності трафіка. Запропонований метод базується на принципах теорії автоматичного регулювання та використовує апарат аналітичного синтезу регуляторів для реалізації механізму авторегулювання у реальному часі. У рамках дослідження сформульовано оптимізаційну задачу вирівнювання поточних коефіцієнтів завантаження серверів шляхом мінімізації інтегрального квадратичного функціоналу, який враховує відхилення реальної траєкторії станів системи від оптимальної. Розроблено математичну модель процесу балансування, що представлена у вигляді системи диференціальних рівнянь із відповідними обмеженнями, які гарантують сталість роботи системи під час різких сплесків трафіка. Здійснено синтез регулюючого функціоналу, що поєднує функцію вирівнювання, функцію Ляпунова-Беллмана та механізм керування на основі рівняння Ріккати. Отримано рівняння настроювання у замкненій формі, яке описує процес динамічного вирівнювання навантаження між серверами. Доведено коректність роботи механізму балансування та показано, що запропонований метод забезпечує рівномірний розподіл навантаження, зменшення ризику перенавантаження окремих серверів, скорочення часу відгуку системи та підвищення ефективності використання обчислювальних ресурсів. Отримані результати можуть бути використані при розробленні адаптивних серверних платформ, хмарних систем оброблення запитів та інтелектуальних керуючих модулів для систем реального часу.

Ключові слова: інтерактивна система, динамічний режим, оптимальний розподіл потоків, сервер, балансування.

DYNAMIC SERVER PERFORMANCE BALANCING METHOD FOR INTERACTIVE SYSTEMS

Viktor Gnatyuk, Ph.D., Ass. Prof., State University “Kyiv Aviation Institute”; State Scientific and Research Institute of Cybersecurity Technologies and Information Protection, Kyiv, Ukraine. <https://orcid.org/0000-0002-4916-7149>, viktor.hnatiuk@npp.kai.edu.ua

Kostiantyn Zander, State University “Kyiv Aviation Institute”, Kyiv, Ukraine. <https://orcid.org/0009-0006-4944-9249>, 8390983@stud.kai.edu.ua

Abstract. The paper proposes a dynamic server performance balancing method for interactive systems operating under unpredictable variations in client request intensity. It is shown that traditional load distribution approaches designed for static or quasi-static operating modes fail to ensure stable quality of service under highly dynamic traffic conditions. The proposed method is based on the principles of automatic control theory and applies analytical regulator synthesis to implement self-regulation mechanisms in real time. The study formulates an optimization problem aimed at equalizing the current server load coefficients by minimizing an integral quadratic functional that accounts for deviations of the real system state trajectory from the optimal one. A mathematical model of the balancing process is developed in the form of a system of differential equations with constraints that guarantee system stability during traffic surges. The control functional is synthesized as a combination of a balancing function, the Lyapunov–Bellman function, and a control mechanism based on the Riccati equation. A closed-form tuning equation describing the dynamic equalization of server loads has been derived. The correctness of the balancing mechanism is proved, and it is demonstrated that the proposed method ensures uniform load distribution, reduces the risk of individual server overloads, shortens system response time, and increases the efficiency of computing resource utilization. The obtained results can be applied to the design of adaptive server platforms, cloud request-processing systems, and intelligent control modules for real-time applications.

Key words: interactive system, dynamic mode, optimal flow distribution, server, balancing.

Вступ

Сучасні інтерактивні системи характеризуються високою динамічністю трафіку та значною варіативністю навантаження на серверну інфраструктуру. Традиційні методи розподілу ресурсів між серверами, орієнтовані на статичні або слабкозмінні умови, не забезпечують ефективної роботи системи у реальному часі, що може призводити до перевантажень окремих серверів, збільшення часу обробки запитів та зниження якості обслуговування користувачів.

Сучасні підходи до адаптивного та інтелектуального керування ресурсами демонструють високу ефективність, однак більшість із них застосовуються або в умовах передбачуваних навантажень, або для конкретних типів серверних систем. В умовах інтерактивних систем з непостійною інтенсивністю клієнтських потоків залишається актуальним завдання розробки методів динамічного перерозподілу продуктивності серверів, що враховують поточний стан завантаженості кожного сервера та характер змін трафіку.

Отже, наукова проблема полягає у створенні методу адаптивного балансування продуктивності серверної лінійки, який забезпечував би оптимальне використання обчислювальних ресурсів, підвищував загальну ефективність роботи системи та підтримував стабільний рівень якості обслуговування при змінній інтенсивності клієнтських запитів.

Аналіз останніх джерел

Аналіз останніх наукових джерел у сфері управління динамічними системами та оптимізації продуктивності серверів свідчить про високий рівень розробок у галузі теорії автоматичного керування, адаптивних методів регулювання та сучасних алгоритмів оптимізації. Традиційні методи управління та теорія зворотного зв'язку. Класичні підходи до управління динамічними системами представлені у працях Kulmamirov та Smakova, Dorf та Bishop, Franklin et al. У цих джерелах детально розглядаються принципи побудови регуляторів, системи зворотного зв'язку, стійкість та адаптація параметрів. Вони формують теоретичну базу для розробки методів динамічного балансування продуктивності серверів, де необхідно забезпечити стабільність та ефективність реакції системи на змінні навантаження. Актуальним напрямом є використання PID-регуляторів та їх розширених модифікацій для управління продуктивністю складних систем (Aström & Hägglund). Такі методи дозволяють гнучко налаштовувати параметри регуляторів у реальному часі, що є критично важливим для інтерактивних серверних систем з непередбачуваним навантаженням. Ascher та Petzold пропонують ефективні чисельні методи для розв'язання диференціальних та диференціально-алгебраїчних рівнянь, що дозволяє моделювати динамічні процеси у серверних системах і прогнозувати поведінку системи при зміні навантаження. Це забезпечує основу для валідації та тестування алгоритмів балансування. Сучасні адаптивні та оптимізаційні методи. Ряд статей у IEEE (Wang et al.; Chen et al.; Zhuang et al.) присвячено методам адаптивного керування, підкріпленого навчанням (Reinforcement Learning) та ітераційного навчання для лінійних та нелінійних систем. Ці підходи дозволяють розробляти алгоритми, які самонавчаються та оптимізують продуктивність серверів у реальному часі, враховуючи змінні умови експлуатації та обмеження ресурсів. Застосування подібних методів до інтерактивних серверних систем може суттєво підвищити ефективність балансування навантаження, скоротити час відгуку та знизити ризик перевантажень.

Аналіз останніх джерел показує, що сучасні підходи до управління та оптимізації продуктивності систем поєднують класичну теорію управління, адаптивні регулятори та алгоритми машинного навчання. Для задач динамічного балансування серверів у інтерактивних системах найбільш перспективним є інтегрований підхід, що поєднує теоретичні основи стабільності та управління з сучасними методами адаптації та оптимізації на основі даних у реальному часі.

Метою статті є розробка та обґрунтування методу динамічного балансування продуктивності серверів інтерактивної системи, який: забезпечує адаптивний перерозподіл обчислювальних ресурсів між серверами залежно від поточного рівня їх завантаженості; враховує характер та динаміку змін інтенсивності клієнтських потоків; підвищує загальну ефективність серверної лінійки та стабільність обробки запитів користувачів; може бути реалізований для практичних інтерактивних систем з мінімальними втратами продуктивності та затримками у реагуванні на змінні умови навантаження.

Виконання досліджень

Моделювання процесу розподілу потоку звернень між серверами. Для досягнення поставленої цілі – оптимізувати роботу інтерактивної системи в умовах різних сплесків трафіка клієнтських звернень у напрямку підвищення її реактивності – у даному дослідженні запропоновано представляти інтерактивну систему як динамічну, яку слід розглядати у двох режимах – статичному (устойчивому) і динамічному (перехідному). У статичному режимі роботи системи, визначались оптимальні значення ємності буферної стекової пам'яті, що призначені для організації черг запитів, та кількості прикладних серверів, що здійснюють одночасну обробку клієнтських запитів, тобто оптимальні значення тих параметрів, котрі не змінюються у часі в процесі її функціонування. За результатами такої параметричної оптимізації була задана оптимальна траєкторія зміни станів інтерактивної системи у реальному часі відповідно до розробленої моделі обслуговування запитів як системи масового обслуговування (СМО), яка враховує сталі характеристики трафіка клієнтських запитів. Однак сплески трафіка звернень призводять до непередбачуваних відхилень реальної траєкторії зміни станів системи від оптимальної і за певних умов

можуть вивести систему із області її сталої роботи. Тому виникла потреба у мінімізації таких відхилень з тим, щоб додержувались критерії сталості інтерактивної системи.

Перш ніж здійснити моделювання процесу розподілу потоку звернень між серверами доцільно врахувати наступні міркування. Механізм динамічного розподілу потоку звернень між серверами має відтворювати у реальному часі процес вирівнювання навантаження на ці сервери. При цьому слід мати на увазі, що ні характеристики трафіка звернень, ні тривалість опрацювання звернень кожним із серверів не піддаються прогнозуванню. Саме тому із множини можливих методів розподілу доцільно обрати метод авторегулювання, що відслідковує тенденції щодо поточних рівнів завантаженості серверів і формує відповідний вектор керуючих впливів у напрямку вирівнювання ступеню їхнього навантаження. Отже, задачу керування розподілом потоку запитів між серверами на основі спостереження за рівнями їхнього завантаження у реальному часі доцільно представити як задачу динамічного вирівнювання коефіцієнтів завантаження цих серверів. Для її рішення, як показано далі, зручно використати результати теорії побудови аналітичних регуляторів, що широко висвітлені в багатьох публікаціях.

У теорії побудови аналітичних регуляторів при формалізації алгоритмів динамічного регулювання прийнято користуватися векторними формами запису змінних, що підлягають керуванню. У даному випадку це значення інтенсивності потоків клієнтських звернень, що просуваються на вводи кожного із серверів лінійки. Однак в якості керованих більш зручно узяти безрозмірні змінні – поточні значення коефіцієнтів навантаження на сервери лінійки $k_1, k_2, k_3, \dots, k_n$ або їхні зворотні значення $\eta_1, \eta_2, \eta_3, \dots, \eta_n$:

$$k_1 = X_1 / w_1, \quad k_2 = X_2 / w_2, \quad k_3 = X_3 / w_3, \quad \dots,$$

$$k_n = X_n / w_n; \quad \eta_1 = w_1 / X_1, \quad \eta_2 = w_2 / X_2, \quad \eta_3 = w_3 / X_3, \quad \dots, \quad \eta_n = w_n / X_n, \quad (1)$$

де $X_1, X_2, X_3, \dots, X_n$ – поточні значення інтенсивності пульсуючих потоків звернень від клієнтів відповідно на вході кожного із серверів лінійки; $w_1, w_2, w_3, \dots, w_n$ – продуктивність кожного із серверів лінійки; n – кількість серверів у лінійці.

Отже, вектор-стовпець поточних значень коефіцієнтів завантаження серверів:

$$\dot{\mathbf{k}} = \begin{pmatrix} k_1 \\ k_2 \\ \cdot \\ \cdot \\ \cdot \\ k_n \end{pmatrix} \quad (2)$$

показує рівні завантаження кожного із серверів лінійки в той час, як вектор-стовпець зворотних коефіцієнтів завантаження серверів

$$\dot{\boldsymbol{\eta}} = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \cdot \\ \cdot \\ \cdot \\ \eta_n \end{pmatrix} \quad (3)$$

показує, наскільки пропускні спроможності серверів щодо обробки звернень перевищують поточні інтенсивності реальних потоків клієнтських звернень, що ними обслуговуються.

Так що, процес регулювання в узагальненому вигляді можливо моделювати наступними диференціальними рівняннями:

$$\begin{aligned} \dot{\mathbf{k}} &= \Upsilon_{\mathbf{k}}; \dot{\boldsymbol{\eta}} = \Upsilon_{\boldsymbol{\eta}}. \\ \mathbf{k}' &= \Upsilon_{\mathbf{k}}; \boldsymbol{\eta}' = \Upsilon_{\boldsymbol{\eta}}. \end{aligned} \quad (4)$$

Як бачимо, праві частини цих рівнянь відображають у неявному вигляді функціонал управління процесом регулювання, а ліві частини – вектор керованих змінних. Процес регулювання в даному випадку це є процес покрокового рішення одного із наведених вище рівнянь (4). Як наслідок, значення коефіцієнтів завантаження серверів (або їхніх зворотних значень) будуть поступово у реальному часі вирівнюватися шляхом відповідного перерозподілу часток загального потоку звернень між цими серверами за умови дотримання ряду обмежень, що наведені далі.

Зрозуміло, що для конкретного відображення процесу регулювання слід розкрити в явному вигляді праві частини рівнянь (4), зокрема обрати вид керування, задати вектор керуючих впливів та формалізувати функціональний взаємозв'язок між керованими змінними та керуючими впливами, а також задати обмеження, що пов'язані із забезпеченням сталого функціонування інтерактивної системи з урахуванням якості перехідних процесів, оскільки мова йде про динамічну систему.

Постановка задачі оптимального розподілу потоку звернень на обслуговування між серверами інтерактивної системи. У багатьох застосуваннях інтенсивність трафіка клієнтських звернень швидко змінюється у реальному часі непередбаченим чином. Тривалість обробки цих звернень кожним із серверів лінійки також є непередбачуваною величиною. В процесі роботи одні сервери можуть бути перенавантажени обробкою клієнтських звернень тоді, як інші недонавантажени. Зрозуміло, що бажано уникати таких ситуацій, інакше у результаті перенавантаження будь-якого із паралельно працюючих серверів звернення будуть втрачатися не зважаючи на те, що поточне значення загального коефіцієнту навантаження на лінійку серверів буде на відносно низькому рівні.

Таким чином, доцільною є постановка наступної оптимізаційної задачі: розробити метод динамічного перерозподілу загальної продуктивності лінійки серверів між її серверами в залежності від ступеню поточної їхньої завантаженості трафіком з урахуванням характеру змін інтенсивності потоків клієнтських звернень, що надходять до серверної частини обладнання інтерактивної системи. Тобто, ставиться задача вирівнювання поточних значень коефіцієнтів навантаження серверів лінійки між собою у динамічному режимі в реальних умовах їхнього застосування, коли мають місце різкі сплески трафіка звернень, що мають непередбачуваний характер.

Отже, представимо інтерактивну систему як динамічну систему із непередбачуваною траєкторією змін її стану у часі. Для кожної j -ої точки цієї траєкторії визначимо показник оптимальності у вигляді

$$\min K = (X_1 - s^0_1)^2 + (X_2 - s^0_2)^2 + \dots + (X_i - s^0_i)^2 + \dots + (X_n - s^0_n)^2, \quad (5)$$

де K – критерій оптимальності у задачі оптимізації, яка має враховувати невизначеність як у характері сплесків трафіку звернень, так і у тривалості їхньої обробки серверами; n – кількість серверів, що одночасно обробляють звернення клієнтів у j -ій точці траєкторії; $X_1, X_2, X_3, \dots, X_n$ – поточні значення інтенсивності пульсуючих потоків звернень від клієнтів на вході кожного із серверів у j -ій точці траєкторії; $s^0_1, s^0_2, s^0_3, \dots, s^0_n$ – значення інтенсивності потоків оброблених серверами звернень від клієнтів серверів у j -ій точці траєкторії в устояному (статичному) режимі, коли пульсації трафіка є відсутніми.

З фізичної точки зору вираз (5) означає, що у кожній точці траєкторії змін стану інтерактивної системи ставиться завдання мінімізації суми поточних відхилень реальної траєкторії змін стану від оптимальної траєкторії, що визначена для устояного режиму роботи системи.

Визначимо фізичні обмеження, що накладені на простір пошуку оптимального керування станом інтерактивної системи.

Обмеження 1. Сумарна інтенсивність реального трафіка звернень у будь-який момент часу має не перевищувати загальну продуктивність лінійки серверів. Тобто, має виконуватися наступне обмеження:

$$X_1 + X_2 + X_3 + \dots + X_n \leq \Delta W, \quad (6)$$

де ΔW – загальна продуктивність лінійки серверів.

$$\Delta W = w_1 + w_2 + w_3 + \dots + w_n = \text{const}, \quad (7)$$

де $w_1, w_2, w_3, \dots, w_n$ – продуктивність кожного із серверів лінійки.

Якщо обмеження (6) буде порушено, то виникнуть втрати звернень, які бажано не допускати, зокрема, шляхом підключення до процесу паралельної обробки звернень додаткових серверів, що знаходяться у гарячому резерві. Для цього треба мати інструмент вияву моменту часу, коли таке підключення треба здійснювати. Зокрема, це може бути момент переповнення черги на вході будь-якого із серверів лінійки або момент перевищення припустимого часу очікування у черзі будь-якого із звернень. В залежності від сфери застосування інтерактивної системи визначення цього показника здійснюється розробником системи. Отже, оптимальний розподіл має здійснюватися за умови, коли кількість серверів у складі лінійки та значення їхньої пропускну здатності є заданими величинами.

Обмеження 2. Слід увести наступне обмеження:

$$K^0_1 = K^0_2 = K^0_3 = \dots = K^0_n, \quad (8)$$

де $K^0_1, K^0_2, K^0_3, \dots, K^0_n$ – значення коефіцієнтів завантаження серверів лінійки в устояному режимі, що визначаються як

$$K^0_1 = s^0_1 / w_1, \quad K^0_2 = s^0_2 / w_2, \quad K^0_3 = s^0_3 / w_3, \dots, \quad K^0_n = s^0_n / w_n. \quad (9)$$

Критерій оптимальності K у даному випадку пов'язано з умовою забезпечення рівності поточних значень коефіцієнтів навантаження серверів в устояному режимі, коли інтенсивності запитів, що надходять до серверів, є постійними та не відчують флуктуацій.

Зрозуміло, що діапазон припустимих значень коефіцієнтів завантаження кожного із серверів має бути в межах від 0 до 1.

Обмеження 3 щодо вибору механізму розподілу потоку звернень. Процес відслідковування має представлятися як ітеративний процес покрокового рішення у реальному часі відповідним чином синтезованого диференціального рівняння (4), в якості вихідних даних для якого використовуються виміряні на даному кроці інтегрування рівняння поточні значення інтенсивності оброблених потоків на кожному із серверів лінійки, а в якості результату рішення – значення інтенсивності увідних потоків звернень, що мають опрацьовуватися серверами на наступному кроці інтегрування. Синтезоване диференціальне рівняння забезпечить відстеження тенденцій у змінах інтенсивності оброблюваних потоків. Внаслідок інтегрування поточні зміни потоків будуть впливати на величини коефіцієнтів завантаження серверів.

З урахуванням вище зазначеного показник оптимальності має представлятися як:

$$\min K = (X_1 - s^p_1)^2 + (X_2 - s^p_2)^2 + \dots + (X_n - s^p_n)^2, \quad (10)$$

де $s^p_1, s^p_2, s^p_3, \dots, s^p_n$ – рекомендовані значення інтенсивності потоків оброблюваних серверами клієнтських звернень, що визначаються на кожному кроці інтегрування диференціального рівняння відслідковування відхилень реальної траєкторії змін у стані інтерактивної системи відносно траєкторії, що обчислена для устояного режиму її роботи.

Позитивний ефект від наближення до режиму однаковості значень коефіцієнтів навантаження серверів полягає у тому, що в процесі перерозподілу створюється ситуація, коли робота з обробки пульсацій трафіка рівномірно розподіляється між серверами. Це зменшує ризики перенавантаження будь-якого із них, якщо величини цих пульсацій знаходяться в межах запасу пропускної здатності інтерактивної системи.

Однак реально інтенсивності потоків звернень швидко змінюються непередбачуваним чином. Вони не можуть бути визначені та виміряні миттєво. Існує можливість оцінювання лише поточних тенденцій у змінах швидкостей потоків на коротких відрізках часу. Тому за цих умов механізм розподілу потоку звернень між серверами має спиратися не на вимірювані поточні значення інтенсивності потоків, а відслідковувати тенденції щодо напрямку змін інтенсивності цих потоків – у бік збільшення або зменшення.

Обмеження 4 щодо забезпечення сталої роботи інтерактивної системи в умовах різких сплесків трафіка звернень. Синтез диференціального рівняння (4), інтегрування якого у реальному часі забезпечує відслідковування відхилень реальної траєкторії змін станів інтерактивної системи від оптимальної траєкторії, має здійснюватися з урахуванням трьох факторів. По-перше, покрокове інтегрування цього рівняння має відтворювати процес відслідковування в напрямку мінімізації показника оптимальності (10). По-друге, відтворювана траєкторія змін у стані системи має знаходитись у зоні сталого функціонування цієї системи. По-третє, із множини можливих показників якості перехідних процесів, перш за все, слід враховувати швидкість перехідних процесів при зміні станів системи. Інші показники якості перехідних процесів у даному випадку суттєвої ролі не відіграють. Отже, необхідно синтезувати функціонал, що обмежує область припустимих рішень щодо керування механізмом перерозподілу трафіку звернень між серверами. Цей функціонал має зв'язувати керовані параметри динамічної системи з керуючими впливами механізму перерозподілу трафіка таким чином, щоб забезпечувалось виконання умов обраного критерію сталості та контрольована швидкість перехідних процесів.

Слід сподіватися, що у рамках визначених обмежень результатом рішення даної оптимізаційної задачі буде зменшення ймовірності перенавантаження серверів трафіком і, як наслідок, зменшення величини коефіцієнту втрат необроблених клієнтських звернень.

Теорія автоматичного регулювання (ТАР) пропонує кілька рівноцінних методів рішення для класу задач, до якого відноситься поставлена задача оптимального розподілу потоку звернень на обслуговування між серверами інтерактивної системи. Зокрема, дану задачу можливо вирішити за допомогою класичного варіаційного числення, методом принципу максимуму Л. С. Понтрягіна або методом динамічного програмування. Проте, найбільш прийнятним у даному випадку є метод, що дозволяє отримати рішення у замкнутому вигляді, тобто у вигляді однозначно визначеного рівняння настроювання. Тому у даній роботі узгодження усіх параметрів системи для забезпечення її сталої роботи в оптимальному режимі згідно з критерієм (10) здійснено шляхом використання функціоналу, формалізований вигляд якого представлений далі. Процедура формального синтезу цього функціоналу розглянута у багатьох роботах, що присвячені проблемам керування динамічними об'єктами, і не потребує додаткового опрацювання. Завдання даної роботи – обґрунтувати вибір одного із відомих видів цього функціоналу в якості основи для розроблюваного методу рішення задачі оптимального розподілу потоку звернень.

Синтез методу рішення задачі оптимального розподілу потоку звернень між серверами інтерактивної системи. Даний метод передбачає наступний порядок рішення задачі оптимального розподілу потоку звернень: вибір регулюючого функціоналу; синтез матриці регулюючих зв'язків для функції вирівнювання; синтез функції вирівнювання; обґрунтування виду рівняння настроювання; визначення вектору керування; синтез рівняння настроювання у замкненій формі; формальний доказ коректності роботи механізму вирівнювання.

Вибір виду регулятора для розподілу потоку звернень. У даному випадку синтез методу рішення задачі оптимального розподілу полягає у виборі виду типового регулятора з відомими параметрами та у визначенні оптимальних з точки зору обраних показників якості значень цих параметрів. Тобто, спочатку в рамках даного методу здійснюється обґрунтування вибору оптимальної структури (закону регулювання) типового регулятора для заданого об'єкту керування та обраного критерія оптимальності, а потім вирішується задача параметричного синтезу обраної структури.

Існує велика кількість методів побудови регуляторів для керування об'єктами різноманітного призначення щодо яких висувається широкий спектр вимог, котрі враховують конкретні умови функціонування цих об'єктів. Відповідні класифікатори цих методів широко представлені в наукових публікаціях. Усю множину методів побудови регуляторів доцільно розділити на дві великі групи. До першої групи відносять чисто експериментальні методи визначення параметрів регулятора, коли із евристичних міркувань задають його базову структуру і далі здійснюють ітеративний підбір бажаних значень параметрів цієї структури методом проб та помилок. Зокрема, спостерігають, як поводить себе регулятор з експериментально визначеними коефіцієнтами регулювання, і в залежності від того, що має місце – недорегулювання чи перерегулювання – змінюють коефіцієнти та повторюють експеримент. При цьому користуються відомими методиками підбору коефіцієнтів, що створені за методом Зіглера-Ніколса. Зрозуміло, що ці методи не потребують детальної аналітичної проробки задачі регулювання і не можуть претендувати на оптимальність отриманих рішень, але у багатьох випадках можуть слугувати основою для первісного налагоджування регулятора. До другої групи відносять методи синтезу регуляторів на основі побудованих моделей системи управління об'єктом регулювання. Ці методи дозволяють більш точно розрахувати параметри регулятора і в багатьох випадках отримати оптимальні рішення з точки зору обраних критеріїв якості його роботи. У свою чергу, методи другої групи щодо математичного опису моделей систем регулювання розділяють на частотні та часові.

Лінійні та нелінійні системи. На практиці більшість об'єктів регулювання за своєю природою є багатовимірними та нелінійними. Особливо, якщо розглядати ці об'єкти у широкому діапазоні змін режимів їхньої роботи та в умовах дії значних збурень. Вичерпний аналіз нелінійних систем регулювання надано у працях з теорії нелінійних систем автоматичного регулювання та управління. Зроблено висновок, що формальний синтез регуляторів для таких об'єктів у загальному випадку не уявляється можливим, а використання лінеаризованих моделей при синтезі регуляторів не завжди дозволяє отримати задані показники сталості та точності їхньої роботи. Більш пізні роботи з результатами досліджень нелінійних, стохастичних, робастних та адаптивних систем зроблений вище висновок лише підтверджують.

Особливістю задачі оптимального розподілу потоку звернень є те, що на практиці у більшості випадків спостерігаються відносно невеликі відхилення траєкторії зміни станів об'єкту регулювання від заданого режиму роботи. Окрім того, вплив такого явища як запізнення реакції на клієнтські запити (внаслідок кінцевої швидкості обробки інформаційних потоків) на сталість та якість роботи більшості прикладних систем, в даному випадку не є суттєвим. З цього витікає можливість вважати шукану систему регулювання лінійною. Так що, будемо вважати лінійними як об'єкт регулювання, так і регулятор. А аналітичне конструювання регулятора будемо здійснювати в класі задач оптимального лінійного регулювання.

Вибір регулюючого функціоналу. Даний метод рішення задачі оптимального розподілу потоку звернень передбачає необхідність вибору математичного апарату для формалізації процесу такого розподілу.

Зокрема, якщо користуватися термінами теорії авторегулювання, метод передбачає необхідність вибору виду керування. Тобто, необхідно задати функціонал, що оптимізує у реальному часі сталу роботу замкненої системи авторегулювання в заданих обмеженнях щодо обраних показників якості перехідних процесів. Цей функціонал має задавати на формальному рівні вид функціональних зв'язків між керуючими впливами та керованими змінними з урахуванням визначених вище обмежень, а також структуру та алгоритм роботи регулятора (механізму керування), що задає оптимальну траєкторію зміни стану інтерактивної системи із додержанням критеріїв сталості її функціонування. Умовою дотримання режиму узгодженої роботи усіх елементів системи авторегулювання має бути забезпечення мінімальних значень обраного виду функціоналу. Цей функціонал має задавати також вид рівняння відслідковування тренду змін у стані інтерактивної системи.

У даному випадку вибір виду оптимізуючого функціоналу здійснено виходячи із наступних міркувань. По-перше, необхідно, щоб у підінтегральному виразі функціоналу містилась складова, що має відображати квадратичні відхилення поточних значень інтенсивності потоку звернень від їх устояних значень згідно виразу (10). По-друге, необхідно, щоб функціонал мав вигляд інтегрального критерію, оскільки використання інтегральних критеріїв дозволяє відобразити закон управління у замкнутій формі та задати вимоги до якості перехідних процесів таким чином, щоб керована система не втратила асимптотичну сталість. Зокрема, шляхом майже довільного вибору початкових значень відносно невеликої кількості вагових коефіцієнтів при функції регулювання забезпечується можливість здійснювати цілеспрямований ітераційний перебір значень вагових коефіцієнтів в області сталого режиму роботи системи аж доки не буде досягнута задана якість перехідних процесів. По-третє, інтегральні функціонали зайшли широке відображення в теорії оптимального управління, так що даний спосіб формалізації дозволяє підібрати вид типового регулятора з відомими характеристиками, що є оптимальним для розподілу потоку звернень між серверами інтерактивної системи. Застосування типового регулятора дозволяє позбавитись необхідності здійснення розрахункових робіт з метою визначення елементів його структури.

Якщо вважати, що система регулювання забезпечує можливість спостереження за змінами вектору стану системи, то сутність рішення задачі оптимального лінійного регулювання на основі інтегрального критерію оптимізації полягає в наступному.

У класичній постановці цієї задачі лінійну систему задають двома векторними диференційними рівняннями:

$$\begin{aligned}\dot{x} &= A(t) x(t) + B(t) u(t); \\ y(t) &= C^T(t) x(t),\end{aligned}\quad (11)$$

де $x(t_0) = x^0$; $A(t)$, $B(t)$, $C(t)$ – матричні функції часу розмірністю $(n \times n)$, $(n \times m)$, $(n \times p)$ відповідно; $x(t)$ – вектор стану системи, $u(t)$ – вектор керуючих впливів, $y(t)$ – вектор виміряних значень змінних, що визначає поточний стан системи.

Необхідно знайти вектор $u(t)$, що забезпечує перехід системи протягом проміжку часу $b = [t_0, t_1]$ з початкового стану $x(t_0) \neq 0$ в кінцевий стан $x(t_1) = 0$.

Як показник якості регулювання використовують інтегральний квадратичний критерій наступного виду:

$$I[x(t_0), u(t), t_0] = \int_{t_0}^{t_1} [x^T Q x + u^T R u] dt + x^T(t_1) G x(t_1), \quad (12)$$

де Q , R , G – матриці масштабування, фізичний зміст котрих буде роз'яснено далі.

Як бачимо із (12), значення функціоналу залежить не тільки від вектору керуючих впливів $u(t)$, але і від початкового стану $x(t_0)$, що не є обмежуючою умовою для роботи механізму перерозподілу потоку звернень. У цьому випадку більш реалістичним є інтегральний квадратичний критерій, що не залежить від початкового стану системи та часу її переходу в кінцевий стан, а інтервал оптимізації є нескінченним, а саме:

$$I[u(t)] = \int_0^{\infty} [x^T(t) Q(t) x(t) + u^T(t) R(t) u(t)] dt, \quad (13)$$

де $Q(t)$, $R(t)$ – позитивно визначені матриці.

Перша складова підінтегрального виразу відображає механізм регулювання, а друга – механізм керування. Доведено [6]: якщо система регулювання є повністю керованою у будь-який момент часу t , то оптимальне рішення щодо вектора керування має наступний вигляд:

$$u^*(t) = -R^{-1}(t) B^T(t) \left[\lim_{T \rightarrow \infty} P(t, t_1) \right] x(t). \quad (14)$$

У виразі (14) $P(t, t_1)$ є позитивно визначеним рішенням допоміжного рівняння Ріккати, якщо виконується обмежувальна умова $P(t_1, t_1) = 0$.

Рівняння Ріккати – це матричне нелінійне алгебраїчне рівняння із змінними коефіцієнтами, що має наступний вигляд:

$$- \dot{P} = PA + A^T P - PBR^{-1}B^T P + Q. \quad (15)$$

Вперше рівняння Ріккати було використано А.М. Летовим для побудови лінійного регулятора. Доведено, що рівняння (15) має стале єдино можливе рішення, якщо регульований об'єкт, що визначається матрицями A і B , знаходиться під керуванням, а симетричні матриці Q та R є позитивно визначеними. У нашому випадку для задачі оптимального розподілу потоку звернень названі вище умови виконуються, оскільки поточні значення коефіцієнтів завантаження паралельно працюючих серверів у процесі регулювання прагнуть до їхніх сталих значень.

При векторі керування, що визначається виразом (14), мінімальне значення функціоналу (15) буде представлятися наступним виразом:

$$I^*[x(t_0), t_0] = x^T(t_0) \left[\lim_{T \rightarrow \infty} P(t_0, t_0) \right] x(t_0). \quad (16)$$

Слід звернути увагу на те, що мінімальному значенню функціонала (16) відповідає широка область оптимальних траєкторій змін стану системи (11), оскільки не враховуються обмеження на керування механізмом регулювання. Найбільш суттєвим є те, що вираз (13) не враховує обмеження, що пов'язане із забезпеченням необхідного рівню якості перехідних процесів. Щоб інтегральний критерій відображав сталу роботу системи регулювання із заданими показниками якості перехідних процесів, необхідно, у підінтегральний вираз функціоналу (13) додати третю складову, що синтезована на основі застосування функції Ляпунова-Беллмана. Синтез виду шуканого функціоналу, що враховує показник якості перехідного процесу і дозволяє визначити вектор керування, котрий забезпечує оптимальне регулювання в межах заданого значення цього показника, здійснено наступним чином.

В теорії побудови аналітичних регуляторів для динамічних систем широко використовуються квадратичні матричні форми та операція транспонування матриць. Для формального представлення взаємозв'язків між елементами системи авторегулювання зазвичай утворюють квадратичну форму у вигляді:

$$K^T S \cdot Y \cdot S^T K, \quad (17)$$

де K – вектор-стовпець керованих змінних; S – матриця регулюючих зв'язків між керуючими впливами та керованими змінними; Y – матриця масштабування керуючих впливів; T – операція транспонування матриці.

Для нашого випадку доцільно використати метод побудови системи авторегулювання. Цей метод орієнтовано на виконання умов як сталості, так і якості перехідних процесів. Відповідно до цього методу функціонал, що має зв'язувати усі елементи механізму відслідковування сплесків трафіка звернень і задавати оптимальну траєкторію змін у стані інтерактивної системи з урахуванням визначених вище обмежень, представляється як:

$$\min \Phi = \int_0^{\infty} (K^T S \cdot W_R \cdot S^T K + \zeta K^T S \cdot Q \cdot S^T K + U^T R U) dt, \quad (18)$$

де $W_R - m \times m$ – мірна квадратна позитивно визначена симетрична матриця вагових коефіцієнтів, яка задає рівень «важливості» кожній парі регулюючих зв'язків між керованими змінними;

$Q - m \times m$ – мірна позитивно визначена симетрична матриця квадратичної форми, що є складовим елементом функції Беллмана;

ζ – позитивна константа (множник при функції Беллмана), що визначає ступінь загасання функції Беллмана;

$U - m$ – мірний вектор керуючих впливів;

$R - m \times m$ – мірна позитивно визначена симетрична матриця вагових коефіцієнтів, яка задає рівень «важливості» кожного керуючого впливу;

T – символ операції транспонування матриці.

Як бачимо із (18), підінтегральний вираз функціоналу Φ на відміну від функціоналу (13) складається із трьох членів.

Оптимізація процесу авторегулювання полягає у виборі параметрів механізму вирівнювання навантаження на сервери, за яких значення цього функціоналу будуть мінімізовані.

Перша складова підінтегрального виразу Φ_1 – це формальне відображення функції вирівнювання. Φ_1 представляє у неявному вигляді роботу механізму вирівнювання керованих змінних між собою (тобто, коефіцієнтів навантаження на сервери лінійки), а саме:

$$\Phi_1 = K^T S \cdot W_R \cdot S^T K. \quad (19)$$

Як бачимо, функція вирівнювання згідно (10) являє собою зважену суму квадратів відмінків (через коефіцієнти матриці W_R) керованих змінних. Чим більша ця сума, тим буде більше значення функціоналу Φ . Механізм вирівнювання має діяти у напрямку поступового зменшення цієї суми. При цьому слід мати на увазі, що чим більше значення функціоналу, тим довше за часом здійснюється процес вирівнювання.

Друга складова підінтегрального виразу – це є функція Ляпунова-Беллмана Φ_2 , яка визначає швидкість дії механізму вирівнювання. Вона уведена у функціонал Φ із заданим показником загасання ζ . Функція Ляпунова-Беллмана на оптимальних траєкторіях зміни станів динамічної системи убуває із швидкістю не меншій за здобуток її самої та показника загасання ζ . Тому, шляхом вибору значень показника ζ є можливим задати швидкодію оптимальної системи авторегулювання (точніше, задати час існування перехідного процесу) за якої ця система буде замкнутою та сталою. Формальне

представлення функції Ляпунова-Беллмана Φ_2 у вигляді квадратичної форми в даному випадку співпадає із представленням функції вирівнювання, а саме:

$$\Phi_2 = K^T S \cdot Q \cdot S^T K. \quad (20)$$

Функцію Ляпунова-Беллмана можливо трактувати як поточну відстань між збуреним станом замкненої системи регулювання та незбуреним, що є характерним для статичного режиму її роботи. Динаміка перехідного процесу тим вища, чим швидше убуває ця відстань. Отже, експериментальним шляхом підбираючи значення показника загасання ζ , є можливим задати прийнятну для даних конкретних умов використання інтерактивної системи швидкість убування функції Ляпунова-Беллмана уздовж траєкторії змін стану замкненої системи регулювання.

Третя складова підінтегрального виразу Φ_3 – це формальне представлення роботи аналітичного регулятора, а саме:

$$\Phi_3 = U^T R U. \quad (21)$$

Його уведення у підінтегральний вираз функціоналу (18) дозволяє визначити керування та формально замкнути ланцюг авторегулювання, тобто представити систему регулювання у вигляді одного математичного виразу.

Синтез матриці регулюючих зв'язків для функції вирівнювання. Після вибору регулюючого функціоналу згідно з методом оптимального розподілу потоку звернень між серверами інтерактивної системи необхідно сформувати матрицю регулюючих зв'язків S для функції вирівнювання Φ_1 , яка задає порядок перерозподілу керуючих впливів між змінними під час регулювання.

Матриця повинна відображати можливість перерозподілу навантаження між усіма серверами лінійки, тобто встановлювати функціональні зв'язки між будь-якими двома серверами. Кількість стовпців у матриці дорівнює числу змінних, що підлягають керуванню (кількість паралельно працюючих серверів), а кількість рядків – числу унікальних пар серверів. Для зручності сервери нумеруються послідовно, щоб кожний рядок матриці характеризував взаємодію певної пари серверів; порядок розташування пар у таблиці не має значення.

Матриця описує не абсолютні значення потоків звернень, а напрямки змін навантаження. Знак «+» означає збільшення потоку на сервер, «-» – його зменшення, «0» – відсутність взаємодії між відповідною парою. Таким чином, кожний рядок матриці інтерпретується як передача частини ресурсу від сервера зі знаком «-1» до сервера зі знаком «1», тоді як сервери з позначкою «0» у цьому процесі участі не беруть.

Якщо лінійка складається із чотирьох серверів, то маємо чотири змінні, що потребують керування – k_1, k_2, k_3, k_4 . Якщо лінійка складається із п'яти серверів, то маємо п'ять змінних – k_1, k_2, k_3, k_4, k_5 . Тоді для відображення взаємозв'язків між чотирма серверами матриця S_4 буде складатися із шести рядків, а для відображення взаємозв'язків між п'ятьма серверами матриця S_5 буде складатися із десяти рядків:

$$S_4 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}; S_5 = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}. \quad (22)$$

Фізичний зміст, зокрема, другого рядка матриці S_4 відображає ситуацію, коли частка потоку звернень передається від першого сервера на другий сервер. Третій рядок цієї матриці відображає передачу частки потоку звернень від другого сервера на третій. Аналогічно відображається перерозподіл ресурсу і у матриці S_5 . Зокрема, третій рядок S_5 відповідає передачі частки потоку з другого сервера на третій. Порядок розташування рядків або стовпців у матрицях значення не має. Стовпці і рядки можуть мінятися місцями.

Синтез функції вирівнювання. Після того, як в явному вигляді задана матриця регулюючих зв'язків S згідно запропонованого методу рішення задачі оптимального розподілу потоку звернень, необхідно визначити функцію вирівнювання в явному вигляді. Ця функція Φ_1 в неявному вигляді представляється першою складовою регулюючого функціоналу (18), тобто виразом (19). Як бачимо, для

формального відображення функції вирівнювання використана квадратична форма, у котрій змінні, що потребують регулювання, задані у вигляді вектор-стовпця поточних значень коефіцієнтів завантаження $\mathbf{K}$ (це вираз (2)), а механізм перерозподілу керуючих впливів між змінними в процесі регулювання задано у вигляді матриці регулюючих зв'язків $\mathbf{S}$ (це вираз (22)).

Спочатку в явному вигляді представимо перший елемент Φ_1 , а саме добуток $\mathbf{K}^T \mathbf{S}$, тобто добуток між матрицею регулюючих зв'язків $\mathbf{S}$ та транспонованим вектором керованих змінних $\mathbf{K}$. Аналогічним чином представляється третій елемент Φ_1 , тобто добуток $\mathbf{S}^T \mathbf{K}$. Другим елементом Φ_1 є матриця вагових коефіцієнтів $\mathbf{W}_R$, яка задає «вагу» кожному регулюючому зв'язку між керованими змінними. Ця матриця в явному вигляді представляється як позитивно визначена діагональна матриця:

$$\mathbf{W}_R = \begin{pmatrix} w_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{33} & 0 & 0 & 0 \\ - & - & - & - & - & - \\ - & - & - & - & - & - \\ 0 & 0 & 0 & 0 & 0 & w_{nn} \end{pmatrix}. \quad (23)$$

При побудові системи розподілу потоку звернень «вагу» кожного регулюючого зв'язку доцільно прирівняти до 1, так що

$$w_{11} = w_{22} = w_{33} = \dots = w_{nn} = 1. \quad (24)$$

Якщо маємо лінійку із чотирьох серверів, то перший елемент добутка Φ_1 в явному вигляді представиться як

$$\mathbf{K}^T \mathbf{S}_4 = (k_1, k_2, k_3, k_4) \cdot \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix} = (k_1 - k_2, k_3 - k_1, k_2 - k_3, k_1 - k_4, k_4 - k_2, k_3 - k_4). \quad (25)$$

Як бачимо, добуток (25) є транспонованим вектором чотирьох керованих змінних, тобто вектором-рядком. Кожен компонент вектору-рядка представляється як відмінок між значеннями відповідної пари керованих змінних, що складають вектор $\mathbf{K}$.

Третій елемент Φ_1 , а саме добуток $\mathbf{K}^T \mathbf{S}$, в явному вигляді представиться як

$$\mathbf{S}^T \mathbf{K} = \begin{pmatrix} k_1 - k_2 \\ k_3 - k_1 \\ k_2 - k_3 \\ k_1 - k_4 \\ k_4 - k_2 \\ k_3 - k_4 \end{pmatrix}. \quad (26)$$

Як бачимо, добуток (26) є вектором-стовпцем чотирьох керованих змінних. Кожен компонент вектору-стовпця представляється як відмінок між значеннями відповідної пари керованих змінних, що складають вектор $\mathbf{K}$.

У нашому прикладі, коли маємо чотири керовані змінні, матриця вагових коефіцієнтів $\mathbf{W}_R$ в явному вигляді представляється як

$$\mathbf{W}_R = \begin{pmatrix} w_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & w_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & w_{33} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{44} & 0 & 0 \\ 0 & 0 & 0 & 0 & w_{55} & 0 \\ 0 & 0 & 0 & 0 & 0 & w_{66} \end{pmatrix}. \quad (27)$$

$$w_{11} = w_{22} = w_{33} = \dots = w_{nn} = 1. \quad (28)$$

Підставивши (25), (26) та (27) в (19) отримаємо в явному вигляді кінцевий вираз для функції вирівнювання:

$$\begin{aligned} K^T S \cdot W_R \cdot S^T K = p_{11}(k_1 - k_2)^2 + p_{22}(k_3 - k_1)^2 + p_{33}(k_2 - k_3)^2 + \\ p_{44}(k_1 - k_4)^2 + p_{55}(k_4 - k_2)^2 + p_{66}(k_3 - k_4)^2. \end{aligned} \quad (29)$$

Порівнявши вираз (5) з отриманим виразом (29), можливо зробити висновок, що цей вираз являє зважену суму квадратів відмінків між усіма можливими значеннями керованих змінних і може бути використаний в якості показника оптимальності процесу вирівнювання навантаження на сервери, що оброблюють звернення клієнтів інтерактивної системи. Так що, вираз (29) і визначає функцію вирівнювання.

Обґрунтування виду рівняння настроювання. Вище було показано, що задача динамічного настроювання регулятора зводиться до задачі вирівнювання коефіцієнтів завантаження пошукових серверів. Якщо в якості керованих змінних узяті поточні значення коефіцієнтів навантаження на сервери лінійки $k_1, k_2, k_3, \dots, k_n$, то процес вирівнювання має здійснюватися таким чином, щоб під час його роботи сума керованих змінних була постійною величиною. Так що, поводження даної системи регулювання доцільно підкорити рівнянню настроювання у наступному виді:

$$K = SU, \quad (30)$$

де K – вектор керованих змінних, S – матриця регулюючих зв'язків, а U – вектор керуючих впливів, який у даному випадку має визначатися як лінійна функція компонентів вектору K .

Визначення вектору керування. Аналіз функціоналу показує, що вектор керування має залежати від функції вирівнювання Φ_1 і функції Ляпунова-Беллмана Φ_2 . Функція Φ_1 в явному вигляді представлена виразом (29), а функція Φ_2 у вигляді квадратичної форми (20). Отже, першим кроком визначення вектору керування має бути отримання виразу для функції Ляпунова-Беллмана в явному вигляді. Для цього використовують допоміжне рівняння Беллмана, що має наступний вигляд:

$$\min_u (K^T S W_R S^T K + \zeta K^T S Q S^T K + U^T R U + \dot{K}^T (\partial V / \partial K) + (\partial V / \partial K^T) \dot{K}) = 0. \quad (31)$$

Якщо підставити у рівняння Беллмана (31) рівняння настроювання (30), то одержимо

$$\min_u (K^T S W_R S^T K + \zeta K^T S Q S^T K - U^T R U - U^T S^T S Q S^T K - K^T S Q S^T S U) = 0. \quad (32)$$

Для того, щоб отримати кінцевий вираз для вектору керування, у наступному кроці визначення матриці Q необхідно про диференціювати по вектору керування вираз (32) і результат прирівняти нулеві. У результаті вектор керування відобразиться як

$$\begin{aligned} U &= -R^{-1} S^T S Q S^T K; \\ U^T &= -K^T S Q S^T S R^{-1}. \end{aligned} \quad (33)$$

Усі складові елементи виразів (33), окрім матриці Q , що входить до квадратичної форми функції Беллмана, мають математичне відображення у явному вигляді. Для визначення цієї матриці реалізують наступну відому процедуру. Спочатку у рівняння Беллмана (31) підставляють вираз (33) і отримують наступне рівняння:

$$\begin{aligned} K^T S \cdot W_R \cdot S^T K + \zeta K^T S Q S^T K + K^T S Q S^T S R^{-1} R R^{-1} S^T S Q S^T K - \\ - K^T S Q S^T S R^{-1} R R^{-1} S^T S Q S^T K - K^T S Q S^T S R^{-1} R R^{-1} S^T S Q S^T K = 0. \end{aligned} \quad (34)$$

Потім після спрощень і винесення за дужки однакових лівих та правих множників у виразі (34) отримують наступне рівняння:

$$K^T S \cdot (W_R + \zeta Q - Q S^T S R^{-1} S^T S Q) \cdot S^T K = 0. \quad (35)$$

Оскільки рівність (35) є справедливою для будь-яких значень винесених за дужки множників, то вираз у дужках має дорівнювати нулеві. Так що,

$$W_R + \zeta Q - Q S^T S R^{-1} S^T S Q = 0. \quad (36)$$

Якщо порівняти отриманий вираз із (15), то можливо зробити висновок, що вираз (36) являє собою рівняння Ріккати, тобто матричне квадратне рівняння, що використовують для пошуку матриці функції Беллмана. Існує декілька ефективних обчислювальних алгоритмів для його рішення. У результаті рішення рівняння (36) отримують матрицю Q в явному вигляді. Якщо підставити знайдену матрицю у вираз (33), то одержимо в явному вигляді усі елементи вектору керування.

Синтез рівняння настроювання у замкненій формі. Для того, щоб отримати вираз для системи вирівнювання у замкненій формі, достатньо підставити знайдений вектор керування (33) у рівняння настроювання (30). Отже, замкнута система вирівнювання поточних значень коефіцієнтів завантаженості пошукових серверів визначається наступним виразом:

$$K = -SR^{-1}S^T S Q S^T K. \quad (37)$$

Таким чином, отримані вище вирази являють формальний опис механізму вирівнювання навантаження між паралельно працюючими серверами лінійки пошукових серверів. Для завершення формального опису слід навести доказ, що при будь-яких початкових значеннях вектору керування змінних в системі (37) досягається ситуація, коли стає значення керованого вектору є таким, що його компоненти є однаковими, а їхня сума протягом усього процесу вирівнювання є незмінною.

Формальний доказ коректності роботи механізму вирівнювання. Для того, щоб синтезоване рівняння настроювання (37) при будь-якому довільному векторному керуванні забезпечило рішення задачі вирівнювання керованих перемінних, необхідно, щоб вектор керування $\mathbf{U}$ виглядав як лінійна функція компонентів вектору $\mathbf{K}$, а матриця регулюючих зв'язків $\mathbf{S}$ відслідковувала тенденції щодо поточних рівнів завантаженості пошукових серверів.

Доказ того, що за наведених вище умов система вирівнювання забезпечує під час настроювання збереження суми компонент вектору $\mathbf{K}$, міститься, зокрема, в [13–15] на прикладі системи регулювання другого порядку. Доказ для систем настроювання більш високого порядку здійснюється аналогічно.

Для обраного прикладу система настроювання у розгорнутому вигляді представлялась як

$$\begin{pmatrix} \dot{k}_1 \\ \dot{k}_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \mathbf{U}. \quad (38)$$

А вектор керування для (38) як лінійна функція керованих змінних у вигляді

$$\mathbf{U} = \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \cdot \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}. \quad (39)$$

де w_{11} , w_{12} , w_{21} , w_{22} – вагові коефіцієнти масштабування матриці $\mathbf{W}_R$, які задають «вагу» кожному регулюючому зв'язку між керованими змінними.

Завдання полягало в тому, щоб визначити конкретні значення цих коефіцієнтів і упевнитись, що похідні від керованих перемінних будуть однакові за величиною і протилежні за знаком. У цьому разі їхня сума буде зберігатися.

Отже, з урахуванням (39) рівняння настроювання у замкненій формі буде мати наступний вигляд:

$$\begin{pmatrix} \dot{k}_1 \\ \dot{k}_2 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{W}_{11}\mathbf{W}_{12} \\ \mathbf{W}_{21}\mathbf{W}_{22} \end{pmatrix} \cdot \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}. \quad (40)$$

А у скалярному записі система (40) матиме наступний вигляд:

$$\dot{k}_1 = -(w_{11} k_1 + w_{12} k_2) + (w_{21} k_1 + w_{22} k_2); \quad \dot{k}_2 = (w_{11} k_1 + w_{12} k_2) - (w_{21} k_1 + w_{22} k_2). \quad (41)$$

Далі для даної системи настроювання призначається стає допоміжне рівняння у вигляді

$$\dot{V} + cV \leq 0, \quad (42)$$

де V – допоміжна функція, що уведена у вигляді квадратичної форми

$$V = q (k_1 - k_2)^2. \quad (43)$$

а q – позитивний скалярний коефіцієнт.

У виразі (42) c – показник загасання допоміжної функції.

Похідна за часом від допоміжної функції V має наступний вигляд:

$$\dot{V} = 2q (k_1 - k_2) (\dot{k}_1 - \dot{k}_2). \quad (44)$$

Далі визначають похідні від керованих змінних з рівняння (41) і отриманий результат підставляють в допоміжне рівняння (42). І після спрощень одержують наступний вираз:

$$4 (w_{21} - w_{11}) k_1 + 4 (w_{22} - w_{12}) k_2 + ck_1 - ck_2 = 0. \quad (45)$$

Дорівнюють коефіцієнти при керованих змінних і з виразу (45) отримують систему рівнянь, яка дозволяє визначити «вагу» кожного регулюючого зв'язку між керованими змінними:

$$w_{21} - w_{11} = -(c/4); \quad w_{22} - w_{12} = c/4. \quad (46)$$

Отже, отримано два рівняння (46) для визначення чотирьох «вагових» коефіцієнтів, значення двом із котрих призначають довільно. Зокрема, у такий спосіб:

$$w_{11} = w_{22} = 1. \quad (47)$$

Тоді із системи рівнянь (46) з урахуванням (47) одержують значення двох інших «вагових» коефіцієнтів:

$$w_{21} = 1 - (c/4); \quad w_{12} = 1 - (c/4). \quad (48)$$

Підставляють визначені коефіцієнти у рівняння настроювання (40) і шляхом перемноження матриць одержують:

$$\begin{pmatrix} \dot{k}_1 \\ \dot{k}_2 \end{pmatrix} = \frac{c}{4} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} k_1 \\ k_2 \end{pmatrix}. \quad (49)$$

Запис (49) вказує на те, що похідні від керованих змінних однакові за величиною і протилежні за знаком. Отже, їхня сума зберігається, що доводить сформульоване вище твердження.

Висновок

У результаті проведеного дослідження розроблено метод динамічного балансування продуктивності серверів інтерактивної системи, який базується на використанні математичного апарату теорії автоматичного регулювання, зокрема на принципах синтезу аналітичних регуляторів і застосуванні інтегрального квадратичного функціоналу як критерію оптимізації.

Розроблена модель динамічного балансування відображає процес перерозподілу потоків клієнтських звернень між серверами на основі аналізу їх поточного стану, враховуючи інтенсивність трафіка та продуктивність серверної лінійки. Уведення у функціонал регулювання функції Ляпунова-Беллмана дало змогу врахувати показники якості перехідних процесів та забезпечити сталу роботу системи навіть за умов різких змін навантаження. Отримане рівняння настроювання у замкненій формі описує поведінку інтерактивної системи як динамічного об'єкта з властивістю саморегулювання.

Доведено, що застосування запропонованого методу сприяє зменшенню ризику перевантаження окремих серверів, зниженню часу обробки запитів, стабілізації показників якості обслуговування користувачів і підвищенню ефективності використання обчислювальних ресурсів. Результати роботи можуть бути використані для побудови інтелектуальних систем керування в розподілених серверних інфраструктурах, хмарних середовищах та інтерактивних додатках реального часу.

Подальші дослідження будуть спрямовані на технічну реалізацію запропонованого методу, проведення експериментальної валідації його ефективності в умовах змінної інтенсивності трафіка та розроблення програмного прототипу системи динамічного балансування для інтерактивних серверних платформ.

Література

1. Kulmamirov S. A., Smakova N. S. Theory of Automatic Control. Linear Systems and Their Links. Astana, 2024. 150 p.
2. Dorf R. C., Bishop R. H. Modern control systems. 14th ed. Pearson Education, 2022. 1024 p. ISBN 978-0137307258.
3. Franklin G. F., Powell J. D., Emami-Naeini A. Feedback control of dynamic systems. 8th ed. Pearson, 2018. 928 p. ISBN 978-0-13-468571-7.
4. Astrom K., Hägglund T. Advanced PID Control. Research Triangle Park, NC : ISA – Instrumentation, Systems, and Automation Society, 2006. ISBN 978-1-55617-942-6.
5. Ascher U. M., Petzold L. R. Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations. Philadelphia : SIAM, 1998. 331 p. ISBN 161197139X, 9781611971392.
6. Wang Y. S., Matni N., Doyle J. C. A System-Level Approach to Controller Synthesis. *IEEE Transactions on Automatic Control*. 2019. № 10. Vol. 64. P. 4079–4093. DOI: <https://doi.org/10.48550/arXiv.1610.04815>.
7. Chen C., Modares H., Xie K., Lewis F. L., Wan Y., Xie S. Reinforcement Learning-Based Adaptive Optimal Exponential Tracking Control of Linear Systems With Unknown Dynamics. *IEEE Transactions on Automatic Control*. 2019. № 11. Vol. 64. P. 4423–4438. DOI: <https://doi.org/10.1109/TAC.2019.2905215>.
8. Wang Y., Sun J., He H., Sun C. Deterministic Policy Gradient With Integral Compensator for Robust Quadrotor Control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 2020. № 10. Vol. 50. P. 3713–3725. DOI: <https://doi.org/10.1109/TSMC.2018.2884725>.
9. Zhuang Z., Tao H., Chen Y., Stojanovic V., Paszke W. An Optimal Iterative Learning Control Approach for Linear Systems With Nonuniform Trial Lengths Under Input Constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 2023. № 6. Vol. 53. P. 3461–3473. DOI: <https://doi.org/10.1109/TSMC.2022.3225381>.

References

1. Kulmamirov, S. A., & Smakova, N. S. (2024). Theory of automatic control: Linear systems and their links.
2. Dorf, R. C., & Bishop, R. H. (2022). Modern control systems (14th ed.). Pearson Education.
3. Franklin, G. F., Powell, J. D., & Emami-Naeini, A. (2018). Feedback control of dynamic systems (8th ed.). Pearson.
4. Astrom, K., & Hägglund, T. (2006). Advanced PID control. ISA – Instrumentation, Systems, and Automation Society.
5. Ascher, U. M., & Petzold, L. R. (1998). Computer methods for ordinary differential equations and differential-algebraic equations. SIAM.
6. Wang, Y. S., Matni, N., & Doyle, J. C. (2019). A system-level approach to controller synthesis. *IEEE Transactions on Automatic Control*. 64 (10). 4079–4093. DOI: <https://doi.org/10.48550/arXiv.1610.04815>.

7. Chen C., Modares H., Xie K., Lewis F. L., Wan Y., Xie S. (2019). Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics. *IEEE Transactions on Automatic Control*. 64 (11). 4423–4438. DOI: <https://doi.org/10.1109/TAC.2019.2905215>.

8. Wang Y., Sun J., He H., Sun C. (2020). Deterministic policy gradient with integral compensator for robust quadrotor control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 50 (10). 3713–3725. DOI: <https://doi.org/10.1109/TSMC.2018.2884725>.

9. Zhuang Z., Tao H., Chen Y., Stojanovic V., Paszke W. (2023). An optimal iterative learning control approach for linear systems with nonuniform trial lengths under input constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. 53 (6). 3461–3473. DOI: <https://doi.org/10.1109/TSMC.2022.3225381>.

Дата першого надходження статті до видання: 28.10.2025

Дата прийняття статті до друку після рецензування: 27.11.2025

Дата публікації (оприлюднення) статті: 26.12.2025