АНОТАЦІЯ

Бугров А.А. **Моделі і методи вдосконалення високонавантажених розподілених систем.** – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 126 «Інформаційні системи та технології». – Київський національний університет будівництва і архітектури. – Київ, 2025.

Наукова новизна одержаних результатів. Наукова новизна дисертаційної роботи полягає у тому, що:

вперше розроблено:

– математичну модель динамічного масштабування ресурсів, що поєднує експоненційне згладжування для прогнозу із нечітким агрегуванням лінгвістичних оцінок та дозволяє уникнути фіксованих порогів, скорочуючи час реакції високонавантажених розподілених систем;

– адаптивний нечіткий регресивний метод (АНРМ) застосування зазначеної моделі, що забезпечує безперервне формування керуючих впливів шляхом нечіткої регресивної оцінки параметрів масштабування ресурсів;

– метод коригування моменту масштабування, який, аналізуючи вихідні величини АНРМ, визначає найкращий час зміни конфігурації ресурсів і забезпечує баланс між продуктивністю та експлуатаційними витратами;

удосконалено:

– метод оцінювання ефективності масштабування ресурсів, доповнений багатокритеріальним аналізом технічних і економічних показників, що забезпечує комплексну оцінку результативності керування ресурсами;

набули подальшого розвитку:

– концепція комбінованих стратегій управління ресурсами, у межах якої прогнозні процедури експоненційного згладжування інтегровано з нечіткими реактивними механізмами, що підвищує гнучкість і своєчасність адаптації системи;

– методи аналізу впливу динаміки навантаження на результативність масштабування, що уточнюють критерії ініціювання процесу залежно від швидкості та амплітуди зміни навантаження.

Основний зміст дисертаційної роботи

Дисертаційна робота присвячена проблемам адаптивного управління ресурсами високонавантажених розподілених систем, що виникають в умовах динамічних змін навантаження, необхідності балансування між продуктивністю, масштабованістю та відмовостійкістю, а також економічної ефективності.

За результатами дослідження розроблено:

– математичну модель динамічного масштабування ресурсів, що поєднує експоненційне згладжування для прогнозу із нечітким агрегуванням лінгвістичних оцінок та дозволяє уникнути фіксованих порогів, скорочуючи час реакції високонавантажених розподілених систем;

– адаптивний нечіткий регресивний метод (АНРМ) застосування зазначеної моделі, що забезпечує безперервне формування керуючих впливів шляхом нечіткої регресивної оцінки параметрів масштабування ресурсів;

– метод коригування моменту масштабування, який, аналізуючи вихідні величини АНРМ, визначає найкращий час зміни конфігурації ресурсів і забезпечує баланс між продуктивністю та експлуатаційними витратами.

У першому розділі на основі аналізу сучасних високонавантажених розподілених систем встановлено, що їх ефективність і стійкість залежать від архітектурних рішень, механізмів керування ресурсами і вибору технологічного стеку; охарактеризовано широке коло застосувань, зокрема хмарні обчислення, обробку потокових даних і мікросервісні архітектури; з’ясовано, що головними проблемами є баланс між продуктивністю, масштабованістю та відмовостійкістю, що вимагає оптимізації як на рівні інфраструктури, так і програмного забезпечення; досліджено методи масштабування й управління ресурсами, серед яких виділено вертикальне та горизонтальне масштабування: перше дає змогу підвищити продуктивність окремих вузлів, однак має фізичні обмеження, тоді як друге потребує складніших механізмів балансування навантаження, зате забезпечує вищу

відмовостійкість; розглянуто критерії надійності та метрики ефективності, що охоплюють пропускну здатність, час відгуку, хвостові затримки, коефіцієнт відмов і середній час відновлення після збою, а також економічну ефективність, яка вимірюється рівнем використання ресурсів та вартістю обробки запитів; окреслено, що ефективне управління ресурсами й методи адаптивного масштабування дають можливість знизити витрати без втрати продуктивності, а масштабованість системи безпосередньо впливає на її спроможність адаптуватися до змін навантаження.

У другому розділі проаналізовано існуючі методи автоматичного масштабування високонавантажених систем і визначено необхідність адаптивного управління ресурсами в умовах аномальних і непередбачуваних навантажень; розглянуто вплив динамічних змін у запитах на ефективність масштабування та виявлено, що жорсткі порогові значення можуть призводити до неефективного розподілу ресурсів, що підтверджує доцільність використання більш гнучких стратегій масштабування; досліджено реактивні методи автоматичного масштабування, серед яких метод порогових правил, автоматичне масштабування на основі теорії черг та адаптивні реактивні методи; показано, що ці методи є ефективними у випадках різких змін навантаження, однак вони мають певні обмеження, пов’язані з запізненням у реакції на зміну системних параметрів; обґрунтовано необхідність комбінування реактивних методів із проактивними підходами для підвищення ефективності управління ресурсами; розглянуто можливості застосування проактивних методів автоматичного масштабування, зокрема машинного навчання, розширеного навчання з підкріпленням, нечіткої логіки та аналізу часових рядів; показано, що використання проактивних методів дозволяє передбачати майбутні навантаження та ініціювати масштабування завчасно, що зменшує ризик перевантаження системи; обґрунтовано необхідність розробки комбінованих стратегій, які б дозволяли поєднувати прогнозні підходи з механізмами негайного реагування на зміну навантаження.

У третьому розділі запропоновано адаптивну нечітку регресивну модель динамічного масштабування ресурсів; на основі цієї моделі розроблено адаптивний нечіткий регресивний метод (АНРМ), який забезпечує гнучке керування

обчислювальними ресурсами високонавантажених розподілених систем; обґрунтовано доцільність застосування нечіткої логіки для оцінювання поточного стану навантаження, що дає змогу уникнути жорстких порогових значень і здійснювати плавне масштабування; у межах АНРМ розроблено систему нечіткого виведення й сформовано матрицю ваг термів; проведено дослідження впливу характеристик хвилі навантаження на ефективність масштабування і показано, що невдалий вибір моменту зміни конфігурації може спричинити перевитрати або зниження продуктивності; розроблено метод коригування моменту масштабування, який аналізує вихідні величини АНРМ і визначає найкращий час зміни конфігурації, запобігаючи частому перемиканню станів системи; описано принцип використання розрахованого оптимального часу згортання ресурсів; показано, що застосування запропонованого методу, який інтегрує експоненційне згладжування з подвійною корекцією й нечітку регресивну оцінку, усуває надмірну реактивність класичних методів та забезпечує стабільну динаміку адаптації ресурсів до змін навантаження.

У четвертому розділі проаналізовано результати експериментальних досліджень, які підтвердили ефективність запропонованого адаптивного нечіткого регресивного методу динамічного масштабування ресурсів, що поєднує експоненційне згладжування для прогнозу навантаження з нечітким агрегуванням лінгвістичних оцінок; метод дає змогу своєчасно виділяти ресурси, уникати фіксованих порогів і забезпечувати стабільну роботу високонавантажених розподілених систем за оптимального використання обчислювальних потужностей; на прикладі геоінформаційного модуля системи відновлення пошкоджених будівель продемонстровано переваги проактивного масштабування, завдяки якому короткочасні пікові періоди нівелювалися завчасним збільшенням ресурсів і подальшим поверненням до економнішого режиму; для інтелектуальної системи керування транспортним трафіком великого міста показано, що повторюваність годин «пік» полегшує прогнозування, а метод адаптивно коригує обсяг ресурсів під поточний рівень навантаження; на високонавантаженій платформі аналізу даних із соціальних мереж підтверджено здатність методу плавно масштабувати ресурси при поступовому довготривалому зростанні користувацької активності; запропонований

метод може бути інтегрований у практичні системи з високою варіативністю навантаження, забезпечуючи гнучкий і своєчасний розподіл обчислювальних потужностей.

Ключові слова: балансування навантаження, високе навантаження, інформаційні системи, керування ресурсами, нечітка логіка, проактивне масштабування, прогнозування, регресивне моделювання, розподілені системи, часові ряди.