УДК 519.876.2:004.822

ВОВК Є. А.,
КСЕНЗОВЕЦЬ Д. С.

МОДЕЛІ ПРОГНОЗУВАННЯ ДЛЯ РОЗПОДІЛЕННЯ РЕСУРСІВ І НАВАНТАЖЕННЯ В IT-ІНФРАСТРУКТУРАХ

Центри оброблення даних сьогодні стали широко розповсюдженими. Вони є складовою частиною інфраструктури інформаційних технологій (IT-інфраструктури).

Однією з головних задач управління IT-інфраструктурою є ефективний розподілення віртуальних машин (ВМ) між фізичними машинами (ФМ) з урахуванням вимог і ресурсів.

Домінуючою тенденцією у розподіленні ресурсів є управління з прогнозуючими моделями. У статті розроблено новий метод розподілення ресурсів і навантаження IT-інфраструктури, що базується на прогнозуванні. Розподілення ресурсів і навантаження в IT-інфраструктурах визначає жорсткі вимоги до часу та точності прогнозування. Існує багато моделей прогнозування, які можуть бути використані для управління IT-інфраструктурою, наприклад, ARIMA, GARCH і NARX. Проте для їх ефективного використання для розподілення ресурсів і навантаження необхідно знати, наскільки добре вони вирішують ці завдання відповідно до вимог. У статті зазначені моделі прогнозування досліджуються на реальних даних ЦОД. На основі узагальнення результатів цих досліджень висунуто пропозиції щодо інтеграції моделей прогнозування у механізм розподілення ресурсів і навантаження IT-інфраструктури. Для аналізу моделей прогнозування використано дані Google Cluster Data.

Data centers (DC) have become widespread in our life today. It is integral part of Information technologies infrastructure (IT-infrastructure).

One of the main IT-infrastructure management problems is the one of correct scheduling of virtual machines (VMs) between physical machines (PMs).

The dominant trend in resource allocation is the use of Model Predictive Control. In the article the new based on forecast mechanism of IT-infrastructure resources and load allocation have been developed. The IT-infrastructure resources and load allocation determines the hard requirements to time and accuracy of forecast. There are a lot of forecasting models which can be used for IT-infrastructure management, for example, ARIMA, GARCH and NARX. But for their effective using for resources and load allocation one needs to know how good are they working to solve this tasks according to requirements. In the article above mentioned models of forecast have been explored on the real data of DC. Results of exploration have been summarized and propositions for integration of forecast models in mechanism of IT-infrastructure resources and load allocation have been worked out. The Google cluster data have been used to analyze the forecast models.

Ключові слова: Центр оброблення даних, віртуальна машина, прогнозування навантаження, міграція, розподілення ресурсів.

Вступ

Хмарні обчислення на сьогодні є основним постачальником інфраструктур (IaaS), програмного забезпечення (SaaS) та платформ (PaaS) як сервісів. Потреби користувачів у широкому спектрі обчислювальних середовищ задовольняються за мінімальною ціною і необхідним рівнем обслуговування. Крім того, користувач може контролювати виділення ресурсів, збільшуючи чи зменшуючи їх обсяг відповідно до поточних потреб. Ці ресурси та сервіси доступні за вимогою, і користувач може отримати їх з будь-якого місця за допомогою мережі Інтернет.

Постачальник послуг і користувач домовляються про якість обслуговування (Quality of Service — QoS) за допомогою угоди про рівень послуг (Service-Level Agreement — SLA). Для того, щоб гарантувати погоджений рівень QoS, постачальник має правильно розподілити свої ресурси між користувачами. Це є комплексною проблемою, адже SLA має гарантовано виконуватися незалежно від рівня навантаження, можливих відмов обладнання, перебоїв з електропостачанням і будь-яких інших непередбачуваних несприятливих подій, що можуть статися в IT-інфраструктурі [1].

Існують різні алгоритми для адаптування керування гетерогенними робочими

навантаженнями. Наприклад, пропонується обрати різні алгоритми, або «політики», управління фізичними ресурсами та віртуальними машинами за допомогою адаптивного програмно-визначеного перемикача політик [2].

Задача додатково ускладняється потребою підтримувати багато додатків з гетерогенними робочими навантаженнями, починаючи з невеликих пакетних завдань і закінчуючи веб-сервісами, що працюють у режимі реального часу.

Ефективне вирішення цієї проблеми вимагає точного відображення вимог користувача на ресурси постачальника послуг. Це, у свою чергу, потребує не тільки підтримки моніторингу ресурсів і навантаження, а й передбачення навантаження. Іншими словами, постачальнику потрібен вичерпний підхід до прогнозування навантаження на всіх рівнях IT-інфраструктури: фізичний рівень обладнання, рівень ЦОД, рівень IT-інфраструктури.

Підхід, що пропонується, базується на інтеграції модулів прогнозування, планування та оцінки. Перший надає прогноз для другого. Планувальник розподіляє ресурси та навантаження, використовуючи прогноз навантаження. Третій модуль оцінює параметри QoS після виконання планування на реальних даних і покращує стратегію й алгоритми планування та умови використання моделі прогнозування. Реальні дані використовуються для оновлення бази даних прогнозування в режимі реального часу.

Планувальник потребує прогнозування різних параметрів IT-інфраструктури та її компонентів. Різні статистичні характеристики цих параметрів потребують набору різних моделей та методів прогнозування. Ці моделі та методи будуть використовуватися відповідно до статистичних характеристик часового ряду. Точність прогнозування навантаження залежить від методів прогнозування та характеристик робочого навантаження.

Час виділення ресурсів фізичних машин обмежений секундами. Навантаження ФМ є короткостроковим і несталим процесом. Тому в цьому випадку найкращі результати прогнозування надає інтегрована авто регресійна модель — ковзного середнього (Autoregressive Integrated Moving Average — ARIMA). Методологія ARIMA не надає чіткої моделі передбачення часового ряду. Тоді визначення параметрів моделі ARIMA для ряду реального часу стає проблемою.

Іншими словами, для різних значень періоду та типів параметрів IT-інфраструктури є потреба визначити кращі параметри моделей ARIMA, які буде використано для побудови короткострокового прогнозу.

Але моделі ARIMA не можуть бути використані, коли навантаженню на ЦОД властиві різкі зміни. У реальній IT-інфраструктурі не тільки навантаження, а й ресурси можуть різко змінюватися. У таких умовах може бути використана волатильність — статистичний індекс, що характеризує частоту раптових відхилень від середніх значень параметрів. Така можливість може бути реалізована за допомогою авторегресивних умовно гетероскедастичних (Autoregressive Conditional Heteroskedasticity — ARCH) моделей і узагальнених ARCH моделей (Generalized ARCH — GARCH). Для підрахунку волатильності використовується стандартне відхилення.

Добре відомо, що лінійні моделі мають тенденцію до зниження при підвищенні значень горизонту прогнозування [3]. Тому середньо- та довгострокові моделі прогнозування ефективно використовують нелінійне прогнозування. Головною перевагою нелінійних моделей є те, що вони можуть оброблювати розподіли з важкими хвостами та випадковими відхиленнями фактів, що є характерним для часових рядів.

Найкращою з відомих моделей, що може бути використаною для середньо- та довгострокових нелінійних прогнозувань, є нелінійна авторегресійна екзогенна модель (Nonlinear Autoregressive Exogenous Model — NARX) [4].

Наступним кроком є аналіз пов’язаних праць з метою визначення ефективних умов використання цього та інших методів прогнозування.

Огляд існуючих праць

Через стрімкий розвиток хмарних архітектур і появу додатків, що абстраговані від фізичних ресурсів, задача розподілення ресурсів і навантаження стала доволі складною. Іншими причинами ускладнення даної задачі є змінність потреб і запитів клієнтів, піки навантаження і швидкий ріст та значний розмір IT-інфраструктури.

У цій ситуації ефективне розподілення ресурсів і навантаження має базуватися на моніторингу ресурсів і робочого навантаження, прогнозуванні навантаження і врахуванні динаміки зміни навантаження.

Розв'язання задачі розподілення ресурсів і навантаження має приймати до уваги мету постачальника послуг (зменшення вартості IT-інфраструктури) і користувача (необхідні послуги з низькою вартістю). Тому пов'язані праці цікавлять не тільки як джерело для пошуку та обрання ефективного методу прогнозування. Специфіка полягає, в тому числі, і в пошуку та аналізі, у першу чергу, праць, в яких моделі прогнозування є частиною інтегрованих рішень, націлених на управління IT-інфраструктурою.

Також важливо порівняти методи прогнозування на базі важливих характеристик, наприклад, точності прогнозування. Також необхідно ідентифікувати точність методів прогнозування в залежності від типів об'єктів — процесор, пам'ять, пристрої введення-виведення і запам'ятовуючі пристрої ВМ і ФМ. Важливо і відшукати залежність точності прогнозування від таких параметрів, як довжина вікна прогнозування, кількість точок прогнозування, характеристик часового ряду в цілому.

У праці [5] пропонується модель прогнозування використання процесорного часу та оперативної пам'яті на основі подвійного експоненційного згладжування. У праці [6] також використовується подвійне експоненційне згладжування, але для прогнозування робочого навантаження. Результати, отримані цими авторами, можуть бути використані для визначення параметрів моделі ARIMA.

У праці [7] автори пропонують модель самоадаптивного процесу прогнозування для забезпечення негайного розподілення хмарних ресурсів за потребою. В одну адаптаційну модель були об'єднані переваги кількох технік прогнозування, а саме: авторегресійна модель (autoregressive model — AR), модель ковзного середнього (moving average — MA), моделі штучних нейронних мереж (artificial neural networks — ANN), метод опорних векторів (support vector machine — SVM) і метод програмування генних виразів (gene expression programming — GEP). Недоліками даного методу є велика кількість обчислень і відсутність можливості обрахунку волатильності.

Метод неавтономної самореконфігурації (online self-reconfiguration approach) розподілення віртуальних машин, що запропоновано в [8], використовує генетичні алгоритми для обрання оптимальної політики реконфігурації і подвійне експоненційне згладжування Брауна для прогнозування навантаження. У цьому підході великі обсяги обчислень також супроводжуються відсутністю технік визначення волатильності.

У працях [8, 9] з точки зору точності прогнозування порівнюються різні алгоритми, а саме: адаптивна мережа на основі системи нечіткого виведення (Adaptive Neuro-Fuzzy Inference System — ANFIS), нелінійна авторегресійна екзогенна модель (NARX), інтегрована модель авторегресії — ковзного середнього (ARIMA), метод регресії за допомогою опорних векторів (SVR). Продуктивність цих методів порівнюється за допомогою наступних статистичних інструментів: середньоквадратична похибка RMSE (root-mean-square error), середня абсолютна похибка MAE (mean absolute error), середня абсолютна відсоткова похибка MAPE (mean absolute percentage error), метод найменших квадратів SSE (sum of squared error), нормалізована середня квадратична помилка NMSE (normalized mean squared error). Результати експерименту показали перевагу моделі NARX у порівнянні з моделями ANFIS, ARIMA і SVR. На жаль, автори не порівнювали техніки прогнозування на основі відхилення параметрів від середнього значення.

У праці [10] було запропоновано метод прогнозування навантаження на декілька кроків наперед під назвою KSwSVR. Базуючись на теорії статистичного навчання, він поєднує покращений метод SVR і згладжуючий фільтр Калмана. Метод KSwSVR було порівняно з точки зору точності прогнозування з методами AR, BPNN (штучна нейронна мережа, навчена методом зворотного поширення помилки) і звичайним SVR. Результати експерименту показали, що KSwSVR не може розглядатися як повне узагальнення методів AR, BPNN і SVR.

У праці [11] було розроблено алгоритм прогнозування навантаження на основі фільтру Калмана і методу ANFIS — K-ANFIS. Алгоритм поєднує попередню обробку даних робочого навантаження у хмарному середовищі з використанням фільтру Калмана і модель ANFIS для прогнозування навантаження.

Метод K-ANFIS було порівняно з точки зору точності прогнозування з оригінальним алгоритмом ANFIS і методом ARIMA. Використання алгоритму K-ANFIS значно покращило точність прогнозування у порівнянні з цими двома алгоритмами. Проте, на жаль, у праці не наведено жодних порівнянь з точки зору відхилень параметрів від середніх значень.

Як можна побачити, автори розглянутих робіт провели порівняння відносно невеликої попарно кількості методів прогнозування, і тільки деякі методі прогнозування були порівняні одне з одним на реальних даних Google Cluster Data і Intel Netbatch [12]. Оцінка точності моделей прогнозування була проведена за допомогою прогнозування поза вибірки. Також була проведена перехресна перевірка. Автори зазначили, що одна й та сама модель дає кращу точність на одних наборах даних і гіршу — на інших. Точність прогнозу також залежить від періоду прогнозування. Наприклад, дані Google Cluster Data менш передбачувані на проміжку 10 хвилин, ніж 1 години, при тому на даних Intel Netbatch на обох проміжках спостерігалася приблизно однакова похибка.

Залежність точності методів прогнозування від інших параметрів також була продемонстрована в інших працях. У праці [13] автори показали, що точність прогнозування навантаження залежить від методів прогнозування та характеристик робочого навантаження. У цій статті були досліджені такі методи, як лінійна регресія (linear regression — LR), ANFIS і NARX. Для порівняння цих методів було обрано такі параметри, як використання пам'яті, процесорного часу та дискових ресурсів для машин Google на різних часових проміжках.

Результати свідчать, що метод NARX дає більшу точність, аніж ANFIS і LR, для передбачення на один крок наперед, і метод ANFIS дає найкращі результати для багатокрокового прогнозування у порівнянні з іншими алгоритмами.

Автори [13] також обчислили час роботи методів LR, ANFIS і NARX в залежності від кількості вхідних даних у діапазоні від 500 до 30000 точок. Час роботи методу LR не залежить від кількості вхідних даних (він варіювався від 0,016 с для набору з 500 точок до 0,024 с для набору з 30000 точок). У той же час час роботи методів NARX і ANFIS збільшується пропорційно до кількості вхідних даних (від 0,22 с для набору з 500 точок до 13 с для набору з 30000 точок).

Автор [14] запропонував модуль прогнозування робочого навантаження з використанням моделей ARIMA і зворотного зв'язку для оновлення моделей у режимі реального часу. Використання такого підходу для прогнозування динамічного навантаження на ВМ в еластичних хмарних середовищах має покращити такі параметри QoS, як час відклику та частота відмов. Автор підкреслив, що оцінка точності методів прогнозування може допомогти обрати найкращий серед них і, відповідно, підвищити ефективність розподілення ресурсів.

У праці [15] було запропоновано стандарт короткострокового прогнозування навантаження. Він включає не тільки вибір алгоритму прогнозування, а й створення моделі відбору та попередньої обробки даних. На думку авторів публікації, аналіз моделей прогнозування навантаження за цією схемою сформує базис для порівняння методів прогнозування, оцінки ефекту методу прогнозування на загальну продуктивність і можливу гібридизацію моделей.

На основі аналізу даних публікацій можна зробити наступні висновки:

- Регресійні й авторегресійні моделі прогнозування і їх комбінації, що представлені у моделі ARIMA, є найкращими для короткострокового прогнозування.
- Нейронні мережі показують кращі результати для середньо- та довгострокового прогнозування.
- На жаль, було розглянуто лише моделі прогнозування типового робочого навантаження. Моделі прогнозування навантаження в умовах різких змін умов, наприклад, моделі оцінки волатильності, не розглядалися.
- У розглянутих працях було запропоновано лише класичні методи прогнозування, їх поєднання та модифікації.
- У розглянутих працях не було представлено інтегрований підхід до прогнозування на різних рівнях IT-інфраструктури.

Отже, ефективна реалізація механізму розподілення ресурсів і навантаження в IT-інфраструктурі потребує:

- стандартизації процесу обрання алгоритму прогнозування;
- модель обрання даних для прогнозування;
- модель попередньої обробки даних;
- обрання моделі прогнозування на основі особливостей IT-інфраструктури.

Прогнозування в загальній системі планування

Аналіз даних Google Cluster Data показує, що обсяг ресурсів, які замовляють користувачі, значно відрізняються від тих, що дійсно використовуються. Планувальник управляє навантаженням на ФМ шляхом переміщення ВМ між ФМ у кластері. Схема на рис. 1 показує основні блоки та об'єкти для збирання,

оброблення і передавання потрібної для планування інформації.

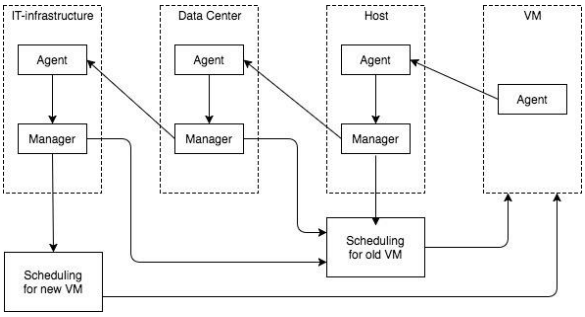

Рис. 1. Схема блоків і об'єктів планування

Основним завданням агентів є збирання інформації та передавання її до відповідних об'єктів. На різних рівнях можуть бути використані різні види інформації. На рівні ВМ агент отримує інформацію про потреби ВМ у ресурсах (процесор, постійна та оперативна пам'ять). На рівні хост-машини (ФМ) агент отримує інформацію від усіх агентів попереднього рівня і формує з неї масив даних, що передається до агенту даного рівня.

Перед тим, як розглянути агента на рівні ЦОД, розглянемо менеджера на рівні ФМ. Основним його завданням є отримання даних від ФМ і ВМ, їх аналіз і проведення прогнозування для ЦОД.

Таким чином, менеджер на хост-рівні отримує дані з усіх ВМ, що виконуються на даній ФМ і робить висновки щодо навантаження даної ФМ у майбутньому. Ця інформація використовується блоком планування і також передається до агенту наступного рівня.

Агент на рівні ЦОД збирає всю інформацію про ресурси, що були використані та що були необхідні для всіх ВМ у цьому ЦОД. Ця інформація збирається у простий для оброблення масив даних і передається до менеджера даного рівня.

Менеджер на рівні ЦОД збирає інформацію про всі ресурси, що були використані та що були необхідні, і робить висновки щодо навантаження цього ЦОД у майбутньому. Ці висновки використовуються блоком планування і передаються до агента наступного рівня.

Агент на рівні IT-інфраструктури отримує всю інформацію про ресурси, що були використані та що були необхідні всіма ВМ в інфраструктурі. Ця інформація збирається в простий для обробки масив даних і передається до менеджера даного рівня.

Менеджер на рівні IT-інфраструктури отримує всю інформацію про ресурси, що були використані та що були необхідні, і робить висновки щодо навантаження на інфраструктуру в майбутньому. Ці висновки використовуються блоком планування.

Крім того, є два блоки планування: «Планування для нової ВМ» і «Планування для існуючої ВМ». Блок «Планування для нової ВМ» активується, коли є потреба додати нову ВМ до інфраструктури, наприклад, з'явилися нові користувачі. Блок «Планування для існуючої ВМ» активується, коли менеджер робить прогноз щодо високого навантаження на IT-інфраструктуру і є потреба перерозподілити вже створені ВМ.

Ця інформація є повною та має невеликий обсяг. Вона дозволяє отримати високоякісний прогноз за невеликий час.

Реалізація алгоритму планування для використання інформації, отриманої модулем визначення стратегії та параметрів планування, може значно скоротити різницю між необхідними і дійсно використаними ресурсами. Це може значно покращити ефективність фізичних машин у кластері.

Прогнозування в системі в основному відбувається і використовується в блоках планування.

Решта цієї статті організована таким чином. У частині 3 розглядається коректність міграції завдань між машинами кластеру з точки зору стратегії та параметрів планування. У частині 4 запропоновано алгоритм прогнозування для визначення необхідності міграції завдань між ФМ у кластері. Важлива для розуміння отриманих результатів схема взаємодії алгоритму прогнозування і механізму розподілення ресурсів і навантаження, зображена на рис. 2.

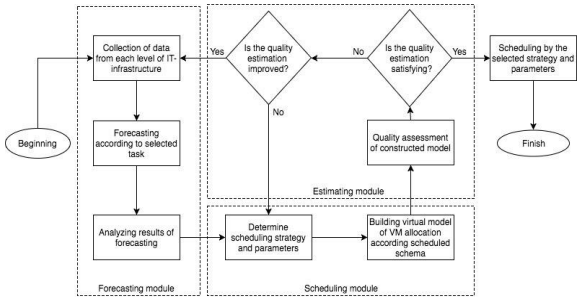

Рис. 2. Схема взаємодії між алгоритмом прогнозування і механізмом розподілення ресурсів і навантаження

Спочатку алгоритм прогнозування збирає та обробляє інформацію про ресурси, що були використані та що були необхідні в рамках системи. На основі результатів аналізу

активується прогнозування навантаження. Якщо прогноз відповідає вимогам, визначається, що система ефективно працює і перерозподілення ВМ між ФМ не потрібне. Інакше, активується метод розподілення ресурсів і навантаження.

Ресурси, що використовуються віртуальними машинами, аналізуються для кожної ФМ окремо. На основі цього і базується прогнозування навантаження на кожну ФМ. Спочатку визначаються проблематичні ФМ. Вони поділяються на дві основні групи: недостатньо навантажені ФМ і занадто навантажені ФМ.

Визначення недостатньо навантажених ФМ. Нехай на ФМ виконуються n ВМ. Кожна ВМ *i*, *i=1,..,n,* вимагає ресурси *j*, *j=1,..,m.* Нехай ці ресурси позначаються як Rd_{ij}. Нехай ВМ *i*, *i*=1,..,*n,* реально використовують Rr_{ij} ресурсів *j*, *j*=1,..,*m*. Для коректної роботи системи, сума необхідних ресурсів j, j=1,..,m на всіх ВМ не може перевищувати ресурс j даної ФМ $R_{host\,j}$:

$$\sum_{i=1}^{n} Rd_{ij} \le R_{host\,j},\ j=1,\dots,m \qquad (1)$$

Тим не менш, можливо, що сума реально використаних усіма ВМ даної ФМ ресурсів j, j=1,..,m значно менша за ресурси j, j=1,..,m даної ФМ:

$$\sum_{i=1}^{n} Rr_{ij} \ll R_{host\,j} \qquad (2)$$

Таким чином, значна частина ресурсів *j* даної ФМ, що може бути обчислена за формулою

$$R_{host\,j} - \sum_{i=1}^{n} Rr_{ij}, \qquad (3)$$

не використовується віртуальними машинами.

ФМ називається недостатньо навантаженою, якщо умова (2) виконується для кожного ресурсу *j, j*=1,..,*m.*

Визначення занадто навантажених ФМ. Поняття занадто навантаженої ФМ в деякому сенсі протилежне поняттю недостатньо навантаженої ФМ. Якщо для недостатньо навантаженої ФМ значна частина ресурсів не використовується, то ресурсів занадто навантаженої ФМ не достатньо для ефективної роботи всіх ВМ на цій ФМ. Також слід зазначити, що ця умова має виконуватися хоча б для одного з ресурсів. Іншими словами, нерівність (2) має стати рівністю для хоча б одного ресурсу j, або хоча б одна з умов (4) має бути виконана.

$$\sum_{i=1}^{n} Rd_{ij} \approx R_{host\,j}, \qquad (4)$$

Якщо в IT-інфраструктурі є недостатньо або занадто навантажені ФМ, частина ВМ має бути перенесена (мігрована) на інші ФМ IT-інфраструктури. Значна частина даних Google Cluster Data пов'язана з міграцією ВМ. Статистична інформація щодо міграцій ВМ дуже важлива для розподілення ресурсів і навантаження. Для отримання цієї інформації необхідно зібрати відповідні дані з Google Cluster Data і отримати корисні статистичні оцінки. На рис. 3 показано кількість міграцій усіх ВМ у ЦОД в рамках однієї доби.

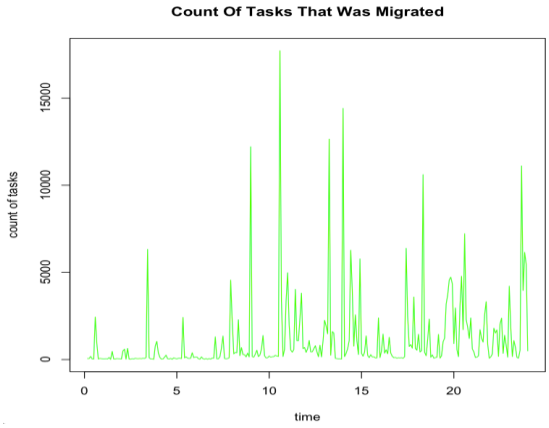

Рис. 3. Статистика міграцій ВМ в рамках однієї доби

Видно, що кількість ВМ, що підпадають під процес міграції, значно залежить від часу доби. Методи прогнозування мають визначати великі сплески в міграціях ВМ і розглядати їх, таким чином покращуючи якість управління IT-інфраструктурою.

Схожим чином можна отримати і корисну інформацію шляхом статистичної обробки реально використаних обсягів оперативної пам'яті, процесорного часу, дискових операцій віртуальної машини. Ця інформація має великий вплив на якість роботи планувальника.

Таким чином, дані Google Cluster Data є вхідними даними для методів прогнозування. У свою чергу, вихідні дані модуля прогнозування є вхідними даними для модуля планування. Відповідно до висновків, зроблених у частині II, є необхідність визначити, які дані мають бути зібрані з метою прискорення роботи цього модуля, і які властивості повинна мати модель прогнозування для обробки цих даних.

Таким чином, вище зазначений новий механізм розподілення ресурсів і навантаження в IT-інфраструктурі може бути реалізований.

Виходячи з визначених у частині II особливостей розподілення ресурсів і навантаження в IT-інфраструктурі, є необхідність в побудові прогнозів на основі наступних даних:

- використані ресурси IT-інфраструктури в цілому (для реалізації модулів прогнозування, планування та оцінювання в системі управління IT-інфраструктурою);
- використані ресурси ЦОД (для реалізації модулів прогнозування, планування та оцінювання у менеджері ЦОД);
- використані ресурси ФМ (для реалізації модулів прогнозування, планування та оцінювання у менеджері ФМ);
- використані ресурси ВМ (для реалізації модуля прогнозування на рівні ВМ).

Залишається обрати методи прогнозування, що будуть брати до уваги особливості цих даних, специфіку рівнів IT-інфраструктури та жорсткі вимоги до часу і точності прогнозування. Це буде виконано у частині IV.

Модель прогнозування

Є три різних завдання коротко-, середньо- та довгострокового прогнозування, а також прогнозування волатильності.

Крім того, існує необхідність планувати ресурси IT-інфраструктури між різними ВМ у двох різних сценаріях: недостатньо і занадто навантажені ФМ. Таким чином, планувальник отримує в режимі реального часу інформацію щодо обсягів ресурсів і навантаження на ФМ, ЦОД та IT-інфраструктуру в цілому.

Беручи до уваги різкі сплески міграцій ВМ та інші параметрі IT-інфраструктури, модуль прогнозування має надавати можливість оцінювати волатильність. У загальному сенсі, волатильність означає флуктуації, що спостерігаються з певним явище протягом часу. Економісти використовують це поняття з метою формально описати без певної непрямої метрики мінливість випадкової (непередбачуваної) компоненти в часовому ряді [16].

Для короткострокового неавтономного прогнозування оцінка волатильності є дуже важливою. Отже, система отримує певні сповіщення щодо необхідності планування ресурсів ВМ на ФМ або ЦОД, таким чином покращуючи використання ресурсів ВМ. Таким чином система управління запобігає перенавантаженню IT-інфраструктури.

Тепер система управління може визначати час, коли слід активувати механізм планування ресурсів ФМ. Проте після визначення часу активації цього механізму з'являється нова проблема, що є не менш складною, — визначення того, як має працювати цей механізм. У цьому допоможе прогнозування. Короткострокове прогнозування дозволяє отримати інформацію щодо ресурсів, що будуть використовуватися віртуальними машинами у майбутньому. Наприклад, обсяг ресурсів, що потребують усі ВМ на одній ФМ, може перевищувати реальні ресурси ФМ. Механізм короткострокового прогнозування визначить занадто навантажені ФМ і активує перерозподілення ВМ між ФМ або ЦОД. Ці механізми також мають працювати в режимі реального часу паралельно з прогнозуванням волатильності.

Короткострокове прогнозування та прогнозування волатильності мають супроводжуватися прогнозуванням навантаження на довші періоди часу. Це допоможе вчасно активувати механізм планування і запобігти невірному розподіленню ВМ в IT-інфраструктурі. Точність довгострокового прогнозування гірша за точність короткострокового, але якщо довгострокові прогнози будуть підкріплені короткостроковими та прогнозами волатильності, IT-інфраструктура працюватиме коректно та швидко.

Для розроблення моделей прогнозування можна використовувати дані робочих навантажень Google Cluster Data. Набір даних Google Cluster Data включає дані обчислювальних ресурсів та робочих навантажень кластера з приблизно 12000 ФМ протягом 29 днів [2]. Ці дані розбиті на шість таблиць: job_events; machine_attributes; machine_events; task_constraints; task_events і task_usage. Дане експериментальне дослідження базується на таблицях machine_events, task_events і task_usage.

Набір даних Google також містить деяку інформацію про самі ФМ, наприклад їх нормалізовані (приведені до числових значень від 0 до 1) ресурси центрального процесора та оперативної пам'яті.

Розклад ряду даних навантаження на центральний процесор наведено на рис. 4.

Табл.1. Порівняння методів прогнозування на різних часових проміжках

Використання ресурсу процесора				
Модель	Похибка	Робоче навантаження 1 (5 хвилин)	Робоче навантаження 2 (1 година)	Робоче навантаження 3 (1 доба)
ARIMA	MAPE (%)	0.9	0.1	0.1
GARCH	MAPE (%)	0.5	0.9	0.2
NARX	MAPE (%)	0.1	0.4	0.5

Використання ресурсу оперативної пам'яті				
Модель	Похибка	Робоче навантаження 1 (5 хвилин)	Робоче навантаження 2 (1 година)	Робоче навантаження 3 (1 доба)
ARIMA	MAPE (%)	0.1	0.7	0.7
GARCH	MAPE (%)		0.6	0.3
NARX	MAPE (%)	0.0517	0.3	0.05

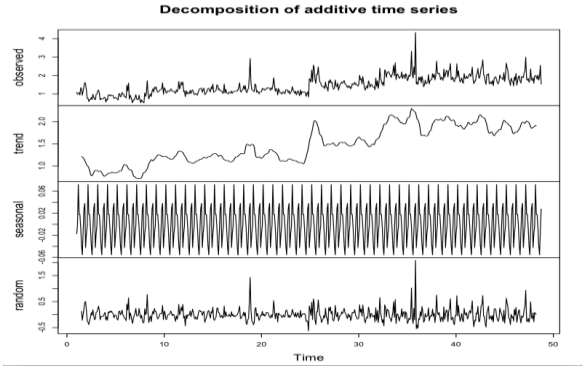

Рис. 4. Розклад ряду даних навантаження на центральний процесор

Як видно зі схеми, часовий ряд можна розкласти на три основні компоненти: тренд (або тенденція), сезонна та випадкова частина.

У нашому випадку, основна увага наділяється тренду та випадковій компоненті. Обробка цих компонент може надати прогноз з високою якістю. Виведення оцінок обраних моделей базується на прогнозуванні, наступному зборі і порівнянні з реальними значеннями, та експерименті. Таким чином було виведено оцінювання якості прогнозування моделей ARIMA, GARCH і NARX. Результати дослідження підсумовано в таблиці 1.

Якщо часовий ряд є стаціонарним, для довгострокового прогнозування можна використати модель ARIMA. Але оскільки у даних Google Cluster Data числовий ряд не є стаціонарним, для довгострокового прогнозування краще використати модель

Табл.2. Час роботи моделей прогнозування на різній кількості точок

Модель	**Кількість прогнозованих точок**		
	12	300	2000
ARIMA (s)	0,0001	0,02	0,28
GARCH (s)	0,001	0,21	0,31
NARX (s)	0,002	0,23	0,5

NARX, що є штучною нейронною мережею і дає добру якість прогнозування на даних Google Cluster Data.

Відповідно до таблиці 2, використання моделі NARX для короткострокового прогнозування не є ефективним через великий час виконання обчислень.

Висновки

У статті представлено загальний метод порівняння моделей прогнозування. Він дає

можливість обрати найкращу модель на базі потреб ІТ-інфраструктури, завдань прогнозування та вимог користувачів.

Крім того, було розширено набір параметрів оцінювання прогнозів з метою покращити механізми прогнозування ті планування в системах управління ІТ-інфраструктурою. Так, було запропоновано використовувати волатильність для того, щоб прогнозувати не тільки обсяги використання ресурсів, а й їх динаміку.

На базі порівняння основних моделей прогнозування було визначено моделі, що краще підходять для вирішення різних задач прогнозування у різних умовах.

Наступні дослідження пов'язані з покращенням і модифікацією алгоритмів прогнозування, щоб забезпечити їх роботу для прогнозування параметрів ІТ-інфраструктури в умовах високого навантаження та обробки великих об'ємів даних (big data).

Перелік посилань

1. Sedaghat M. Cluster Scheduling and Management for Large-Scale Compute Clouds – Umeå, 2015.
2. Telenyk S., Zharikov E., Rolik O. An approach to virtual machine placement in cloud data centers, Radio Electronics & Info Communications // UkrMiCo, 2016 International Conference. – К.: 2016.
3. Hansen P., Lunde A. A comparison of volatility models: Does anything beat a GARCH(1,1)? // Brown Univ. Economics Working Paper. – 2004. – №1-4.
4. Diaconescu E. The use of NARX Neural Networks to predict Chaotic Time Series // Electronics, Communications and Computer Science Faculty University of Pitesti Targu din Vale. – 2008. – №1.
5. Huang J., Li C., and Yu J. Resource prediction based on double exponential smoothing in cloud computing // 2nd International Conference on Consumer Electronics, Communications and Networks, (CECNet'12). – IEEE: 2012. – С. 2056-2060.
6. Mi H., Wang H., Yin G., Zhou Y., Shi D, Yuan L. Online self-reconfiguration with performance guarantee for energy-efficient large-scale cloud computing data centers // 2010 IEEE International Conference on Services Computing (SCC). – IEEE: 2010. – С. 514-521.
7. Jiang Y., Perng C.-S., Li T., and Chang R. ASAP: a selfadaptive prediction system for instant cloud resource demand provisioning // 11th IEEE International Conference on Data Mining (ICDM '11). – IEEE: 2011. – С. 1104-1109.
8. Rasheduzzaman M., Islam M.A., Islam T., Hossain T., Rahman R.M. Study of Different Forecasting Models On Google Cluster Trace // 16th International Conference on Computer and Information Technology (ICCIT). – 2014. – С. 414-419.
9. Rasheduzzaman M., Islam M.A., Islam T., Hossain T., Rahman R.M. Workload Prediction on Google Cluster Trace // Journal International Journal of Grid and High Performance Computing Volume 6. – 2014. – № 3. – С. 34-52.
10. Rongdong H., Jingfei J., Guangming L., Lixin W. Efficient Resources Provisioning Based on Load Forecasting in Cloud // Hindawi Publishing Corporation and Scientific World Journal Volume 2014. – 2014.
11. Sun J., Zhuang Y. The Cloud Computing Load Forecasting Algorithm Based on Kalman Filter and ANFIS, 2016 4th International Conference on Machinery, Materials and Computing Technology. – 2016.
12. Vazquez C, Krishnan R., John E. Time Series Forecasting of Cloud Data Center Workloads for Dynamic Resource Provisioning // Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, Volume 6. – San Antonio: 2015. – № 3. – С. 87-110.
13. Le T.A. Workload prediction for resource management in data centers // Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications. – San Antonio: 2016.
14. Calheiros R.N., Masoumi E, Ranjan R., Buyya R. Workload Prediction Using ARIMA Model and Its Impact on Cloud Applications' QoS // IEEE Transactions on Cloud Computing (Volume 3). – 2015. – № 4.
15. Lopez M., Valero, S., Senabre C., Gabaldon A. Standardization of short-term load forecasting models, European Energy Market (EEM) 9th International Conference. – 2012.
16. Andersen T.G., Bollerslev T., Christoffersen P.F., Diebold F.X. Volatility and Correlation Forecasting // Handbook of Economic Forecasting. – Amsterdam: North-Holland, 2006. – С. 778-878.