OPTIMIZATION OF TELECOMMUNICATION NETWORK DATA CENTER RESOURCES ALLOCATION

Yaremko I.

The problems of resource allocation data center telecommunications network, resulting in the provision of services to customers. An approach to the dynamic allocation of the required number of servers at variable load on the service.

Стрімкий розвиток інформаційно-комунікаційних технологій спричинив глобальну інформатизацію в усіх сферах бізнесу, промисловості, наукових досліджень. Інформаційно-обчислювальні центри в нових умовах поступово трансформуються в центри обробки даних - дата-центри. Сучасні дата-центри дозволяють не лише консолідовано обробляти всю корпоративну інформацію, але й забезпечують роботу соціальних мереж, web-конференцій, інтернет-магазинів та ін., надають хостингові послуги, фізичну і віртуальну інфраструктуру і т.п. технології, що застосовуються в дата-центрах, доволі ефективні для вирішення аналітичних задач з метою підтримки бізнес-проектів, ERP- та білінгових систем, систем мобільних платежів [1]. Насиченість IT-інфраструкгури дата-центрами породжує проблеми ефективного її використання, що можливе лише за умови сучасного інструментарію та методик збору, аналізу інформації, прийняття рішень та контролю їх реалізації. Створення такого інструментарію вимагає глибокого розуміння процесів функціонування IT-інфраструктури, чіткої постановки конкретних задач дослідження, розробки математичних моделей і відповідних методів розв'язання задач і реалізації їх у структурі системи управління.

Замовника не цікавлять, на яких типах серверів вирішуються його задачі, які об'єми систем зберігання даних, яка пропускна здатність каналів зв'язку та ін. - важливіше, щоб користувачі отримували відповідь на запит за час, що не перевищує обумовлених значень.

Забезпечення якості послуг дата-центру на належному рівні вимагає підтримки обчислювальної системи, системи збереження даних та ін. у робочому стані із заданим рівнем надійності. Для цього необхідно забезпечити постійну наявність вільних ресурсів, які оперативно задіюються для нарощування продуктивності сервісу при збільшенні кількості одночасних запитів. З одного боку, якість, своєчасність та обсяг послуг, що надаються, з іншого – необхідність мінімізації затрат обумовлюють актуальність задачі раціонального управління ресурсами дата-центру, коли кількість зарезервованих ресурсів динамічно змінюється в залежності від поточного навантаження на сервіс.

Вимоги до якості послуг замовники фіксують у SLA-угодах. Для надання цих послуг у дата-центрі виділяються необхідні ресурси, якщо їх недостатньо для вирішення задач користувачів, і обчислювальна система не здатна виконувати запити користувачів із заданими обмеженнями у часі, відбувається функціональна відмова, яка призводить до суттєвого зниження якості послуг або ж недоступності сервісу. Стійкість ресурсів до таких відмов будемо описується розрахунковим показником надійності – коефіцієнтом ресурсної готовності, який визначає ймовірність того, що у довільний момент часу окремому замовнику виділено достатні об'єми ресурсів для виконання запитів користувачів, які виключають появу функціональних відмов конкретного сервісу. Система повинна забезпечувати стійкість сервісів до функціональних відмов, виділяючи та резервуючи ресурси для сервісів з подальшим перерозподілом незадіяних ресурсів при зміні кількості запитів користувачів таким чином, щоб забезпечити підтримку коефіцієнту ресурсної готовності на заданому рівні.

Потік запитів користувачів носить статистичний характер, тому для виконання вимог необхідно управляти динамічним виділенням і перерозподілом ресурсів. У такому разі необхідно, маючи в наявності інформацію щодо замовлених послуг і граничних параметрів якості їх надання, наприклад, у вигляді максимального часу обслуговування окремого запиту, відслідковувати динаміку надходження запитів диференційовано за кожною послугою. Для підсилення обчислювальних потужностей, що підтримують послуги, коефіцієнт готовності

яких наближається до граничних значень, необхідно виділяти додаткові ресурси. Для цього частина спільних для усіх послуг ресурсів, що є резервом дата-центру та знаходиться у вимкненому стані або у стані очікування, вмикається, проходить підготовку до роботи та доповнює задіяні на даний час ресурси. Виділення додаткових ресурсів може відбуватися також шляхом перерозподілу задіяних ресурсів з урахуванням пріоритету послуг.

На відміну від традиційного підходу, де спочатку визначається пікове навантаження, а потім резервуються ресурси в об'ємі, достатньому для подолання пікового навантаження, пропонується, що обсяг зарезервованих ресурсів динамічно змінюється у залежності від інтенсивності надходження запитів протягом певного часового інтервалу, і збільшується відповідно до зростання інтенсивності. При зниженні інтенсивності об'єм зарезервованих ресурсів скорочується, а звільнені ресурси можуть бути передані для вирішення інших задач або ж тимчасово відключені з метою заощадження електроенергії та зменшення витрат на обслуговування.

У ролі ресурсів дата-центру на першому етапі можна розглядати кількість серверів, що складають серверні кластери дата-центру, а другому процесорну ємність серверу, об'єм оперативної пам'яті та дискового простору серверу, пропускну здатність каналів зв'язку, ємність системи зберігання даних, канали запитів до СУБД та ін. [2]. Часто беруть до уваги лише частину процесорної ємності та об'єму пам'яті, які виділяються окремій послузі чи окремому замовнику. У якості ресурсу виступають і віртуальні машини (ВМ), у разі, якщо замовникам виділяється деяка кількість ВМ, необхідна для підтримки надання послуг користувачам.

Нехай кожен кластер серверного комплексу дата-центру має у наявності n фізичних серверів, що утворюють множину $N=\{N_1, \ldots, N_n\}$. Оскільки однією з найсуттєвіших витратних статей обслуговування дата-центру вважаються затрати на енергозабезпечення, то організація резерву, де частина зарезервованих обчислювальних ресурсів знаходиться у вимкненому стані або ж у стані зниженого енергоспоживання, має переваги з економічної точки зору. Тоді систему критеріїв для оптимізації можна записати у вигляді [3]:

$$\left.\begin{aligned} \max_{\{N\}} C(N_i) &= \sum_{i=1}^{n} \frac{N_i}{t_i} \\ \min_{\{N\}} OpEx(N_i) &= \sum_{i=1}^{n} OpEx_i \cdot N_i \end{aligned}\right\}$$

з обмеженнями:

$$\sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \frac{t_i}{N_i - q_i} \le t^{опт}$$

$$q_i < N_i \le N_i^{max}$$

де λ-інтенсивність запитів; λ_i-інтенсивність запитів у кластер i; n - кількість кластерів у дата-центрі; $OpEx$ -оперативні витрати дата-центру; $OpEx_i$ – оперативні витрати на сервер; N_i - кількість серверів у кластері i; C – продуктивність дата-центру; t_i-середній час обробки запиту слабозавантаженим сервером кластера i; q_i - номінальне завантаження кластера i із одним сервером при навантаженні λ_i; N_i^{max} - максимальна кількість серверів у кластері i; $t^{опт}$ - оптимальний середній час відповіді.

Таким чином в результаті розв'язання даної задачі може бути отримана кількість серверів по кластерах серверного комплексу дата-центру і вирішена задача першого етапу розподілу ресурсів у системі дата-центру телекомунікаційної мережі.

Література

1. Центр обработки данных основа ИТ-инфраструктуры предприятия [електроний ресурс]/ – Центр Информационных Технологий. – режим доступа: http://www.ci.ru/inform08_04/p_04.htm

2. *Теленик С.Ф., Ролік О.І., Букасов М.М., Андросов С.А., Римар Р.В.* Управління навантаженням і ресурсами центрів оброблення даних при віртуальному хостингу // Вісник Тернопільського держ. техніч. ун-ту. – 2009. – № 4. – С. 198 – 210.

5. *Яремко І.М.* Управління розподілом ресурсів центрів обробки даних телекомунікаційної мережі/І.М. Яремко, В.В Турупалов.// «Искусственный интеллект», №4, 2011, с.380-385.