УДК 621.391

І.М. Яремко (канд. техн. наук, доц.)
ДВНЗ «Донецький національний технічний університет», м. Донецьк,
кафедра автоматики і телекомунікацій

ОПТИМІЗАЦІЯ ФУНКЦІОНУВАННЯ СЕРВЕРНОГО КОМПЛЕКСУ ДАТА-ЦЕНТРУ

Розроблені моделі серверного комплексу дата-центру телекомунікаційної мережі як системи масового обслуговування відбивають особливості функціонування дата-центру. Удосконалено методи оптимізації серверного комплексу дата-центру телекомунікаційної мережі, що дозволяє мінімізувати експлуатаційні витрати і знайти оптимальну кількість серверів при максимізації продуктивності обслуговування, що при змінюваному навантаженні дата-центру телекомунікаційної мережі дозволяє змінювати кількість задіяних серверів.

Ключові слова: *телекомунікаційна мережа, дата-центр, моделі серверного комплексу, оптимізація.*

Вступ

Розвиток інформаційних технологій пов'язаний із реалізацією концепції дата-центрів – комплексних організаційно-технічних рішень для створення високопродуктивної, відмовостійкої IT-інфраструктури. До їх головних завдань належать консолідоване зберігання і опрацювання даних користувачів, надання їм прикладних сервісів, підтримка функціонування застосувань.

Концепція дата-центрів втілена багатьма великими корпораціями переважно для забезпечення доступу великої кількості користувачів до певних ресурсів. Зростаючі вимоги користувачів до рівня сервісу, керованості, надійності, доступності і масштабованості IT-інфраструктури ускладнює управління такою нею, а створення дата-центрів на основі міжнародного стандарту TIA-942 [1], вимагає значних коштів, ефективної підтримки IT-інфраструктури і наявності у штаті висококваліфікованих фахівців.

Послугами дата-центрів користуються компанії, які займаються електронною комерцією, надають інформаційні послуги, фінансові організації, оператори зв'язку та телекомунікацій та ін., а також і фізичні особи. Загальна ємність дата-центрів в Україні оцінюється в 40–50 тис. юнітів, і ця цифра постійно збільшується.

Постановка проблеми

Для забезпечення підтримання параметрів якості системи на заданому рівні і обслуговування запитів користувачів з параметрами часу, що не перевищують обумовлених значень, за суттєвої динаміки запитів необхідно мати надлишок необхідних ресурсів. Водночас, сучасні дата-центри телекомунікаційної мережі характеризуються високою вартістю не тільки створення, але й обслуговування, а отже, значна кількість незадіяних або надміру зарезервованих ресурсів є невиправданою з точки зору економіки.

Тобто постає проблема забезпечення ефективного функціонування дата-центрів - очікують зменшення витрат на експлуатацію, зниження вартості обслуговування користувачів. Зазвичай цю проблему розбивають на ряд складових. Однією з них є проблема управління ресурсами і навантаженням дата-центрів.

Сучасний дата-центр включає серверний комплекс, систему зберігання даних, систему експлуатації й систему інформаційної безпеки, які інтегровані між собою й об'єднані високопродуктивною локальною обчислювальною мережею. Найбільш поширеною моделлю серверного комплексу є модель із багаторівневою архітектурою, у якій виділяється кілька груп серверів (рис.1). Таким чином одним зі шляхів розв'язку поставленого завдання є оптимізація серверного комплексу – кількості серверів в кластерах дата-центру.

Аналіз шляхів вирішення задачі

Комплекс задач дослідження та їх постановки залежать від багатьох чинників, які тією чи іншою мірою впливають на згадані вище процеси. Першим чинником є модель хостингу. При виділеному хостингу застосуванню відведена необхідна кількість серверів, за які клієнти не змагаються, а мають можливість використовувати резервовані для них ресурси. При сумісному хостингу кластери серверів обслуговують велику кількість застосувань.

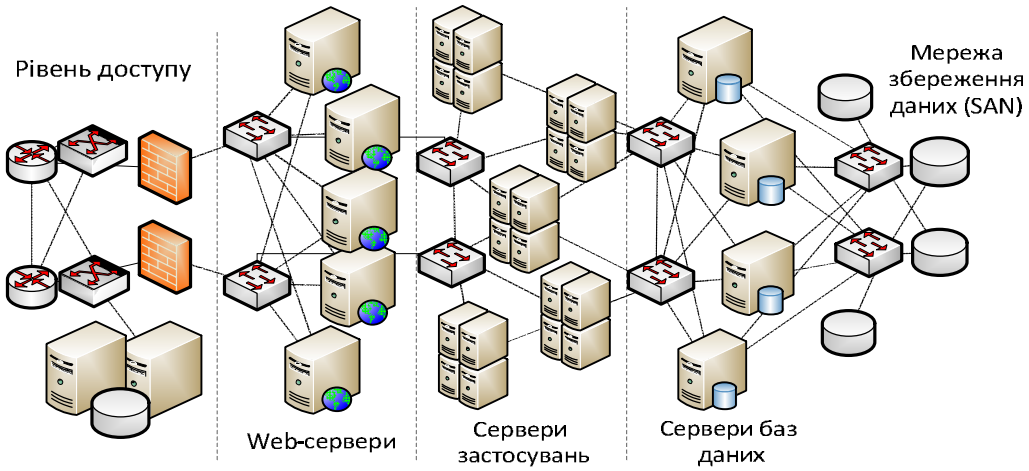

Рисунок 1 - Структура серверного комплексу дата-центру телекомунікаційної мережі

Сучасні кластерні технології дозволяють сервіс-провайдеру масштабувати систему й підвищувати її готовність. Однак не завжди є можливим спрогнозувати навантаження й заздалегідь підготувати достатню кількість обчислювальних ресурсів. Як наслідок, надмірність виділення ресурсів призводить до їх неефективного використання і зайвих економічних витрат. В структурі цих витрат, за даними аналітиків, значну частку займають витрати на електроживлення, витрати на персонал і технічне обслуговування, ліцензії на серверне обладнання і програмне забезпечення. В свою чергу в структурі витрат на електроенергію більш ніж 50% складає електроенергія власне серверного комплексу і комутаційного обладнання. А за даними аналітиків, резерви економії електроенергії можуть бути знайдені в наступних напрямах: до 40 % – при використанні методів оптимізації серверних потужностей; до 15 % – при виборі ефективної архітектури кондиціонування; до 12% – при правильному плануванні фальшполу; до 10 % – при виборі ефективного устаткування електроживлення.

Ці цифри слушні для дата-центрів з високим рівнем резервування (2N), які зазвичай функціонують при навантаженні 30%. Для дата-центрів з низьким рівнем резервування показники економії можуть становити половину від наведених вище.

Тобто один зі шляхів зменшення витрат на утримання дата-центру – оптимізація серверного комплексу, що є основною складовою дата-центру телекомунікаційної мережі.

Незважаючи на численні публікації, існуючі роботи розглядають або оптимізацію за одним основним критерієм – з точки зору отримання прибутку, або розглядають питання розподілу ресурсів з урахуванням надійності апаратно-програмних засобів, але в них не

розкриваються питання якості обслуговування користувачів [2]. Все вищевикладене підкреслює важливість і актуальність розглянутої проблеми.

Побудова моделі серверного комплексу дата-центру телекомунікаційної мережі

Сучасна телекомунікаційна система – об'єкт високої складності, теорія побудови якої знаходиться ще на стадії становлення. Системи обслуговування з розділенням процесора стали основними моделями функціонування web-вузлів комп'ютерних мереж, фактично замінивши собою класичну модель M/G/1 з обслуговуванням в порядку надходження (FCFS) [3]. Системи EPS виступають моделями багатьох інформаційно-обчислювальних систем, загальні ресурси яких використовуються користувачами, запити яких виконуються одночасно. На сьогодні інтерес до них виріс завдяки їх чисельним застосуванням щодо аналізу вузлів комп'ютерних мереж. Розглядаючи серверний комплекс дата-центру телекомунікаційної мережі, що має багатоланкову архітектуру, де кожною ланкою є кластер серверів можливе застосування цієї моделі [4].

В моделі серверного комплексу (рис.2) обслуговуючим приладом є кластер серверів, а заявкою – запити користувачів.

Запити надходять на обслуговування на першу ланку web-серверів з інтенсивністю λ.

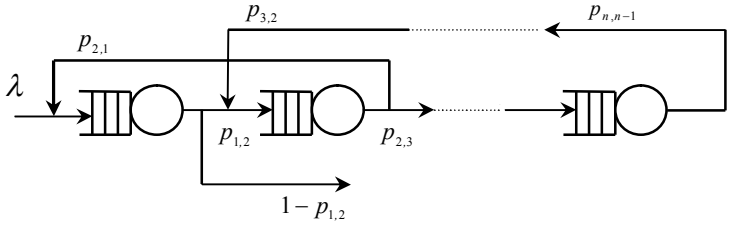

Рисунок 2 - Модель серверного комплексу дата-центру

Після обробки запиту ланкою web-сервера заявка переходить з імовірністю p_{12} до другої ланки кластера серверів застосувань або з імовірністю $(1-p_{12})$ запит виходить із системи - користувачеві надсилається відповідь. Згідно моделі, запит користувача може пройти через усі ланки системи, навіть деякі може відвідати кілька разів. Знаючи ймовірності переходів заявки між вузлами, можна визначити інтенсивності надходжень запитів на кожну ланку λ_i і середній час відповіді на запит - час, що пройшов з моменту надходження запиту в систему до моменту його виходу із системи, тобто сумарний час проходження запиту через усі ланки системи:

$$t_{cp} = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \left(\frac{t_i}{N_i - \rho_i} \right), \tag{1}$$

де t_i – час обробки запиту одним сервером кластеру i;ρ_i – номінальне завантаження кластеру; N_i – кількість серверів у кластері i.

В моделі серверного комплексу з урахуванням класів запитів (рис.3) для моделювання процесу обмірковування вводиться віртуальний сервер з нескінченою кількістю паралельних незалежних каналів обслуговування, що характеризує час обмірковування користувача. Це дозволяє відбити в моделі незалежність часу обмірковування від часу обробки запиту.

Нехай також є $k = 1,\ldots K$ класів запитів, G різних типів сесій, і ресурсів $g = 1,...,G$.

Кожний тип сесії відповідає одному ресурсу. Так само, як і в моделі серверного комплексу без урахування класів запитів, припустимо, що користувацькі сесії типу g надходять у систему з інтенсивністю l_g й починаються із запиту класу k, де $g = 1,...,G$.

Після виконання запиту класу k користувачі з типом сесії g витрачають на обмірковування випадковий час t_g. Після цього вони або вертаються в систему із запитом

класу $\overline{k'}$ з імовірністю $p'_{kk'}$, або виходять із системи, завершуючи сесію, з імовірністю $1-\sum_{k'=1}^{K} p^{g}{}_{kk'}$.

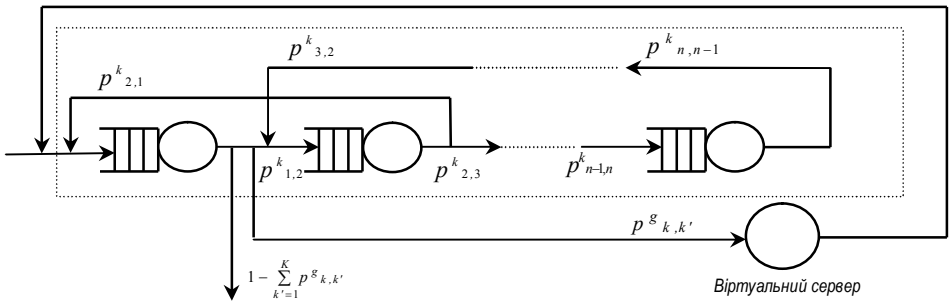

Рисунок 3 - Модель серверного комплексу дата-центру з урахуванням класів запитів

Припустимо, що матриця розмірністю $K \times K$ $P^{g}=\left[p^{g}_{kk'}\right]$ є матрицею ймовірностей переходів користувачів по ресурсах дата-центра. Ця матриця визначає послідовність надходжень запитів у СеМО (мережу масового обслуговування) в межах користувацької сесії g і відбиває зв'язок між надходженнями запитів класу k і k' від того самого користувача.

Нехай Λ^{g}_{k} означає інтенсивність надходжень запитів класу k із сесії g, яка буде дорівнювати:

$$\Lambda^{g}_{k}=\sum_{k'=1}^{K}\Lambda^{g}_{k'}p^{g}_{k'k}+l_{g},\ g=\overline{1,G}.$$

Сумарна інтенсивність надходжень запитів класу k із усіх сесій буде дорівнювати:

$$\lambda^{k}=\sum_{g=1}^{G}\Lambda^{g}_{k}.$$

Запити класу k в багатоланковій системі за одне відвідування можуть мати різний маршрут і кілька раз відвідувати різні кластери. Знаючи ймовірності переходів запиту класу k між кластерами, можна визначити інтенсивності надходжень запитів класу k в кожний кластер.

Час відповіді на запит класу k - t^{k}_{cp} - час, що пройшов з моменту надходження запиту в систему до моменту його виходу із системи, є сумарним часом проходження запиту через усі ланки:

$$t^{k}_{cp}=\sum_{i=1}^{n}\frac{\lambda^{k}_{i}}{\lambda^{k}}\frac{t^{k}_{i}}{N_{i}-\rho_{i}}\ , \qquad (2)$$

де n-кількість кластерів серверного комплексу; N_i – кількість серверів у кластері i; ρ_i – номінальне завантаження кластеру i; t^{k}_{i} – середній час відповіді на запит класу k сервером кластеру.

Отримані моделі є основою для постановки задач оптимізації серверного комплексу дата-центру телекомунікаційної мережі.

Критерії оптимізації

При оптимізації серверного комплексу дата-центру телекомунікаційної мережі задаються обмеження на значення показників якості. Показники надійності і вірогідності сучасних дата-центрів забезпечуються на високому рівні, завдяки стандартам на проектування. На сьогодні, один з найбільш відомих методів врахування витрат в IT-галузі є метод, заснований на сукупній вартості володіння. Сукупна вартість володіння – це

методика, яка призначена для визначення витрат на інформаційні системи та обчислювальні комплекси, які розраховуються на всіх етапах їх життєвого циклу [5]. Один з підходів до обліку витрат базується на їхньому поділі на капітальні й експлуатаційні витрати:

$$S_{TCO} = CapEX + OpEx \cdot T, \tag{3}$$

де *CapEx* (Capital Expenditures) – капітальні витрати організації, які створюють її майбутню вигоду. Вони виникають, коли організація витрачає кошти на придбання нових активів або відновлення існуючих. *CapEx* можна розрахувати на основі обліку вартості устаткування (включаючи монтаж, конфігурацію й необхідне ПЗ) і споруджень по балансовій вартості; *OpEx* (Operational Expenditures) – це вартість бізнес-операцій, які відносять до експлуатаційних витрат організації на утримання активів. Розраховується на основі обліку поточних витрат дата-центру, які можуть бути віднесені до експлуатаційних витрат поточного періоду. Експлуатаційні витрати можуть бути розраховані як витрати прямі (у тих випадках, коли можливий їхній безпосередній облік) і непрямі (у випадку неможливості прямого віднесення витрат на утримання комплексу, у цьому випадку використовують відсоток від деякої бази – наприклад, фонд заробітної плати); T – період, на який відносяться витрати.

З урахуванням вищевикладеного основними критеріями оптимізації обрано максимум продуктивності обробки запитів і мінімум оперативних (експлуатаційних) витрат на серверний комплекс дата-центру.

Таким чином, для вимог якості без поділу запитів на класи, де задається обмеження на середній час відповіді, постановка задачі має вигляд:

$$\left.\begin{aligned} &\max_{\{N\}} C(N_i) = \sum_{i=1}^{n} \frac{N_i}{t_i}; \\ &\min_{\{N\}} OpEx\,(N_i) = \sum_{i=1}^{n} OpEx\ \cdot N_i. \end{aligned}\right\} \tag{4}$$

при обмеженнях

$$\sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \frac{t_i}{N_i - \rho_i} \le t^{onn}\ ; \tag{5}$$

$$\rho_i < N_i \le N_i^{max}, \tag{6}$$

де n – кількість кластерів у дата-центрі; N_i – кількість серверів у кластері i; C – критерій продуктивності дата-центру; $OpEx$ – оперативні витрати серверного комплексу дата-центру; $OpEx_i$ – оперативні витрати на один сервер кластеру i; λ – інтенсивність надходження запитів у систему; λ_i – інтенсивність запитів у кластер i; ρ_i – номінальне завантаження кластеру i з одним сервером при навантаженні λ_i; N_i^{max} – максимальна кількість серверів у кластері i; t_i – середній час обробки запиту слабонавантаженим сервером кластеру i; t^{onm} – оптимальний середній час відповіді.

Для вимог якості з поділом запитів на класи при обмеженні на середній час відповіді:

$$\left.\begin{aligned} &\min_{\{N\}} OpEx\,(N) = \sum_{i=1}^{n} OpEx_{\,i} \cdot N_i; \\ &\max_{\{N\}} C(N) = \sum_{i=1}^{n} \frac{N_i}{\frac{1}{K}\sum_{j=1}^{K} t_i^j}. \end{aligned}\right\} \tag{8}$$

при обмеженнях

$$\sum_{i=1}^{n} \frac{\lambda^j{}_i}{\lambda^j} \frac{t^j{}_i}{N_i - \rho_i} \le T^{*j},\ j = \overline{1,K} \tag{9}$$

$$\rho_i < N_i \le N_i^{max}, \quad (10)$$

де n – кількість кластерів серверного комплексу дата центру; K – кількість класів запитів; N_i – кількість серверів у кластері i ; C – критерій продуктивності дата-центру; $OpEx$ – оперативні витрати дата-центру; $OpEx_i$ – оперативні витрати на один сервер кластеру i; λ^j – інтенсивність надходження запитів класу j у систему; $\lambda^j{}_i$ – інтенсивність надходження запитів класу j у кластер i; N_i^{max} – максимальне можливе число серверів у кластері i; t_i^j – середній час обробки запиту j слабонавантаженим сервером кластеру; T^{*j} – обмеження на середній час відповіді на запит класу j; N_i^{max} – максимальна можлива кількість серверів у кластері i.

Оптимізація серверного комплексу

На основі встановлених вимог було обрано методи і методику, що дозволили б розв'язати поставлену задачу [6].

Процедура оптимізації серверного комплексу дата-центру телекомунікаційної мережі без урахування класів запитів складається з ряду етапів.

1. Визначається середній час обробки запиту сервером у кожному кластері при малих величинах навантаження.

2. Визначається інтенсивність надходжень запитів до кожного кластеру серверів λ_i.

3. Визначається обмеження на максимальний припустимий середній час відповіді на запит.

4. Задаються вагові коефіцієнти критеріїв продуктивності й оперативних витрат $\mu = \{\mu_1, \mu_2\}$, після чого із застосуванням пропонованого методу проводиться оптимізація з використанням моделі за формулами (3-5). Якщо вимогами якості задаються обмеження на максимальний час відповіді для заданої частки запитів, то розрахунки проводяться за формулами (3,5,6). Процедура оптимізації серверного комплексу дата-центру телекомунікаційної мережі з урахуванням класів запитів аналогічна.

Висновки

Розроблено математичні моделі серверного комплексу дата-центру телекомунікаційної мережі як системи масового обслуговування, що відбивають особливості багаторівневої кластерної структури і функціонування дата-центру. Оптимізація серверного комплексу дата-центру телекомунікаційної мережі, дозволяє зменшити кількість задіяних серверів серверного комплексу при змінюваному навантаженні дата-центра телекомунікаційної мережі. Застосування розроблених процедур оптимізації серверного комплексу дозволяє скоротити оперативні витрати на 6-15%, що підвищує ефективність функціонування дата-центру телекомунікаційної мережі.

Список використаної літератури

1. Телекоммуникационная инфраструктура Центров Обработки Данных. Документ SP–3-0092: Стандарт TIA-942, редакция 7.0, (февраль 2005) [Электронный ресурс]. – Режим доступа: http://www.ups-info.ru/etc/tia_russkii.pdf.
2. Кученко Ю. ЦОД как объект системной и структурной оптимизации [Электронный ресурс]. - Режим доступа: http://www.rvip.ru/1065/document1546.shtml.
3. Яшков С.Ф. О методе анализа системы M/G/1-EPS и моментах времени пребывания [Электронный ресурс] / Ю. Кученко // Информационные процессы. - Т. 9, №4. - 2009. - С. 368-375. - Режим доступа: http://www.jip.ru/2009/368-375-2009.pdf.
4. Яремко І.М. Моделі масового обслуговування в ЦОД / І.М. Яремко, В.В. Турупалов // Інформаційно-керуючі системи на залізничному транспорті. – 2011. – №6. – С. 23-26.

5. Равшанов Я. Сколько стоит корпоративный ЦОД: методики расчета ТСО [Электронный ресурс] / Я. Равшанов // Технологии и средства связи. – 2010. – № 4. – Режим доступа: http://www.tssonline.ru/articles2/fix-corp/skolko-stoit-korporativnii-cod-metodiki-rascheta-tso.
6. Яремко І.М. Управління розподілом ресурсів центрів обробки даних телекомунікаційної мережі / І.М. Яремко, В.В. Турупалов // Искусственный интеллект. – 2011. – №4. – С. 380-385.

Надійшла до редакції
31.03.2013

Рецензент:
д-р техн. наук, проф. Скобцов Ю.О.

И. Н. Яремко
ГВУЗ «Донецкий национальный технический университет»

Оптимизация функционирования серверного комплекса дата-центра. *Разработаны модели серверного комплекса дата-центра телекоммуникационной сети как системы массового обслуживания, отражающие особенности функционирования дата-центра. Усовершенствованы методы оптимизации серверного комплекса дата-центра телекоммуникационной сети, что позволяет найти оптимальное количество серверов при минимизации эксплуатационных расходов и максимизации производительности обслуживания, что позволяет изменять количество задействованных серверов серверного комплекса при изменяющейся нагрузке дата-центра телекоммуникационной сети.*
Ключевые слова: *телекоммуникационная сеть, дата-центр, модели серверного комплекса, оптимизация.*

I.M. Yaremkot
Donetsk National Technical University

Functioning Optimization of a Data Center Server Farm. *To maintain the quality parameters of a system at a given level and the service of users' requests with the parameters of time, which should not exceed certain values, when queries dynamic is significant, you must have an excess of resources. At the same time, today's data centers of telecommunications networks are characterized by high cost of creation and service. Hence, significant amount of unused or unnecessary reserved resources is useless from economic point of view. EPS systems are models for many information systems, their shared resources are used by users and queries are done simultaneously. In a server farm of a telecommunications network data center which has multilink architecture, where each link is a cluster of servers, the use of this model is possible. In a server farm model the server is a cluster of servers, and applications are requests. In a server farm model a virtual server with infinite number of parallel independent channels of service is introduced taking into account the class of query. This allows the model to reflect the independence of the time of thinking from the time of query. Thus, on the basis of a model with multi-level architecture we developed models of server complex of data center of telecommunications network as a queuing system, reflecting the peculiarities of the data center. Various models of server farm are developed, which take and do not take into account classes and queries. The advanced algorithms to optimize telecommunications network data center server complexes allows you to quickly find the desired solution. And it allows you to find the optimal number of servers while minimizing operating costs and maximizing the performance of service that allows you to change the number of servers of the server farm with varying load of the telecommunications network data center . The procedures developed to optimize the server farm can reduce operational costs by 6-15%, which increases the efficiency of the telecommunications network data center.*
Key words: *telecommunications network, data center, server farm models, optimization.*