УДК 621.39
© І. Яремко[1], 2011

МЕТОДИКА ОПТИМІЗАЦІЇ РЕСУРСІВ ЦЕНТРІВ ОБРОБКИ ДАНИХ

Розглянуто проблеми розподілу і оптимізації ресурсів центрів обробки даних. Запропоновано метод, що забезпечує гарантовану якість обслуговування запитів користувачів і раціональне використання обчислювальних ресурсів.

The problems of resource allocation data centres are considered. Proposed a method that provides guaranteed quality of service for user requests and managing computing resources.

1. ВСТУП.

Розвиток мережі Інтернет і доступність високошвидкісних каналів зв'язку дозволили об'єднати розосереджені інформаційні та обчислювальні системи підприємств, організацій і державних установ у інтегровані центри обробки даних (ЦОД), послугами яких користуються як телекомунікаційні компанії, представники банківського та фінансового секторів, так і корпорації, великі організації зі значною територіально розподіленою інфраструктурою, а також загальнодержавні й регіональні структури влади. В умовах глобальної інформатизації суспільства і бізнесу зростають вимоги користувачів до рівня сервісу, керованості, надійності, доступності і масштабованості IT-інфраструктури [1].

Існують корпоративні та комерційні спеціалізовані ЦОД. Незважаючи на те, що в першу чергу особливий інтерес до використання комерційних ЦОД проявляє малий та середній бізнес, для великих корпорацій з власними ЦОД комерційні центри можуть застосовуватись в якості резерву потужностей. Послугами комерційних централізованих ЦОД користуються компанії, які займаються електронною комерцією, надають інформаційні послуги, фінансові організації, оператори зв'язку та телекомунікацій та ін., а також і фізичні особи. Якість послуг, що надаються користувачам з боку ЦОД, зазвичай обумовлена договором про гарантоване обслуговування – SLA (Service Level Agreement), який включає в себе опис послуг, тарифи на послуги, рівень сервісу, специфікацію показників якості сервісу, параметри рівня обслуговування і т. д. Готовність, доступність, неперервність та цілісність обчислювальних ресурсів ЦОД або сервісів, що надаються, – од-

[1] Донецький національний технічний університет

ні з найважливіших показників надійності. Щоб забезпечити підтримання параметрів надійності системи на заданому в SLA рівні, а також для обслуговування запитів користувачів з параметрами часу, що не перевищують обумовлених у SLA значень, за суттєвої динаміки надходження запитів, необхідно мати надлишок обчислювальних ресурсів. Водночас, сучасні ЦОД характеризуються високою вартістю не тільки створення, але й обслуговування, а отже, значна кількість незадіяних або надмірно зарезервованих ресурсів є невиправданою з економічної точки зору. У зв'язку з цим задача управління зарезервованими ресурсами, необхідними для надання послуг з показниками якості та надійності згідно SLA, є актуальною.

2. ПОСТАНОВКА ЗАДАЧІ.

Однією з основних задач у галузі сучасних ЦОД можна вважати підвищення ефективності їх роботи. Це пояснюється значними витратами на проектування, будівництво, оснащення ЦОД, а також поточними витратами на підтримку їх працездатності. Одним із напрямків підвищення ефективності роботи ЦОД вважається застосування технології віртуальних серверів [2], яка водночас зі зменшенням кількості фізичних серверів та підвищенням ефективності використання обчислювальних ресурсів дозволяє скоротити витрати на адміністрування, а також оптимізувати енергоспоживання завдяки тимчасовому відключенню незадіяних фізичних серверів та відповідному зниженню витрат на системи енергопостачання та кондиціювання. Однак застосування віртуалізації породжує низку проблем, пов'язаних як із необхідністю управління розподілом ресурсів ЦОД [3], так і з вирішенням задачі оптимізації кількості зарезервованих ресурсів, необхідних для дотримання вимог до рівня обслуговування запитів користувачів, заданих у SLA. Остання задача тісно пов'язана із задачею планування розширення ЦОД, оскільки останнім часом переважає тенденція їх поетапного розвитку, що передбачає поступове нарощування сумарної потужності обчислювальних систем та об'єму систем зберігання даних відповідно до вимог замовників [2, 3, 4]. Однак вказані роботи не приділяють уваги питанням визначення необхідного об'єму ресурсів гарячого резерву для забезпечення необхідної якості обслуговування запитів користувачів.

Управління ресурсами ЦОД спрямоване на забезпечення визначенної в SLA-угодах якості обслуговування. Це може досягатися різними засобами, у тому числі за рахунок оптимального розподілу ресурсів між різними прикладними програмами в ЦОД.

Даній проблемі залежно від виду хостинга присвячене ряд робіт [2-6], у яких поставлені однокритериальні завдання оптимізації з різ-

ною глибиною проробки. Критеріями оптимізації в цих завданнях є витрати, доходи від SLA-угод, пропускна здатність та ін. При цьому в кожнім завданні вид SLA-угод точно визначений, тоді як на практиці можливі різні варіанти завдання SLA-угод. Тому пропонується методика оптимального розподілу ресурсів ЦОД з урахуванням кількох критеріїв, яка може використовуватися для різних видів SLA-угод з різними видами хостинга.

3. МЕТОДИКА ОПТИМІЗАЦІЇ РЕСУРСІВ ЦОД.

Методика оптимального розподілу ресурсів ЦОД складається з декількох етапів. Спочатку проводиться аналіз SLA-угод. Виділяються основні способи завдання вимог на якість обслуговування в SLA-угод: без виділення рівнів обслуговування й класів запитів; з виділенням рівнів обслуговування, без виділення класів запитів; без виділення рівнів обслуговування, але з виділенням класів запитів; з виділенням рівнів обслуговування й виділенням класів запитів.

У першому випадку в SLA-угоді компанії із сервіс-провайдером розглядається загальний клас запитів, для якого задається граничне значення середнього часу відповіді на запит. У випадку якщо значення середнього часу відповіді на запит не перевищує граничного значення, то компанія платить сервісу-провайдеру певну величину винагороди, якщо середній час відповіді перевищує граничне значення, то сервіс-провайдер виплачує штраф компанії. За винагороду приймається плата сервісу-провайдеру за послуги із заданою якістю обслуговування.

У випадку виділення рівнів обслуговування, задається кілька граничних значень середнього часу відповіді на запит, при досягненні яких змінюється величина винагород і штрафів. Такий вид SLA-угод представляється у вигляді східчастої функції корисності [4].

Способи завдання обмежень на якість обслуговування для двох видів SLA-угод, що залишилися, аналогічні розглянутим, тільки значення якості обслуговування задається для декількох класів запитів. Після цього будується агрегована східчаста функція корисності для всіх SLA-угод. Оскільки робота з переривчастою східчастою функцією є складною, то пропонується апроксимувати східчасту функцію корисності гладкою диференційованою функцією. Для зручності добору функціональних залежностей замість єдиної функції корисності пропонується розглядати функції винагород (f_B) і штрафів ($f_Ш$) і апроксимувати їх.

В SLA-угодах не вказуються витрати сервісу-провайдера на забезпечення необхідного рівня середнього часу відповіді на запит. Тому для знаходження оптимального середнього часу відповіді по SLA-

угодам доцільно також розглядати функцію витрат від середнього часу відповіді на запит:

$$f_{ВИТР}(t_{SLA}) = \left(\frac{\sum_{c=1}^{C} \sqrt{TCO_c \cdot m_{Tc} \cdot q_c}}{t_{SLA} - \sum_{c=1}^{C} m_{Tc}} \right)^2, \quad (1)$$

де t_{SLA} - середній час відповіді на запит в системі;

TCO_c - сукупна вартість володіння сервером кластера c, $c = \overline{1,C}$;

m_{Tc} - середній час відповіді на запит сервером кластера c ;

$q_c \approx m_{T_c} \lambda_c$ - номінальне завантаження кластера c за заданим навантаженням λ_c .

Математична постановка задачі находження оптимального середнього часу відповіді на запит за SLA-угодам:

$$f_{SLA}(t^*_{SLA}) = \max_{t_{sla}} f_{SLA}(t_{SLA}), \quad (2)$$

при обмеженнях $t^{min}_{SLA} \le t_{SLA} \le t^{max}_{SLA}$,

де

$$f_{SLA}(t_{SLA}) = f_B(t_{SLA}) - \left(f_{ШІ}(t_{SLA}) + \left(\frac{\sum_{c=1}^{C} \sqrt{TCO_c \cdot m_{Tc} \cdot q_c}}{t_{SLA} - \sum_{c=1}^{C} m_{Tc}} \right)^2 \right) -$$

функція прибутку від SLA-угод;

t^{min}_{SLA} - минимальний можливий середній час відповіді на запит;

t^{max}_{SLA} - максимально можливий середній час відповіді на запит.

Отриманий середній час відповіді на запит є обмеженням на другому етапі оптимізації, де визначається оптимальна кількість серверів без урахування класів запитів.

На другому етапі визначається оптимальна кількість серверів без урахування класів запитів за критеріями сукупної вартості володіння серверами ЦОД і пропускної здатності. Постановка двокритеріальної задачі оптимізації:

$$\max_{\{S_1,\ldots,S_C\}} F_{ПЗ}(S_1,\ldots,S_C) = \sum_{c=1}^{C} \frac{S_c}{m_{Tc}}, \quad (3)$$

$$\min_{\{S_1,...,S_C\}} F_{TCO}(S_1,...,S_c) = \sum_{c-1}^{C} TCO_c \cdot S_c \ , \qquad (4)$$

при обмеженнях $\sum_{c=1}^{C} \frac{\lambda_c}{\lambda} \frac{m_{tC}}{S_c - q_c} \le t^*_{SLA}$,

$q_c < S_c \le S_c^{max}$,

$c = \overline{1,C}$

де C - кількість кластерів в ЦОД;

S_c - число серверів в кластері c ;

$F_{ПС}(S_1,\ldots, S_C)$ - критерий пропускної здатності;

$F_{TCO}(S_1,\ldots, S_C)$ - критерій сукупної вартості володіння серверами ЦОД;

TCO_c - сукупна вартість володіння одним сервером кластера c,

$\frac{\lambda_c}{\lambda}$ - середня кількість відвідувань кластера c запитом за час його перебування в системі;

q_c - номінальне завантаження кластера c з одним сервером при заданому навантаженні λ_c ;

S_c^{max} - максимально можлива кількість серверів в кластері c ;

m_{Tc} - середній час відповіді на запит сервером кластера c;

t^*_{SLA} - оптимальний середній час відповіді, визначений з анализу SLA-угод.

Для обмеження по максимальному часу відповіді на запит для заданої долі всіх запитів, постановка задачі буде такою:

$$\max_{\{S_1,...,S_C\}} F_{ПЗ}(S_1,...,S_C) = \sum_{c=1}^{C} \frac{S_c}{m_{Tc}} \qquad (5)$$

$$\min_{\{S_1,...,S_C\}} F_{TCO}(S_1,...,S_c) = \sum_{c-1}^{C} TCO_c \cdot S_c \qquad (6)$$

при обмеженнях $P[T \ge V] \le \varepsilon$; $q_c < S_c \le S_c^{max}$.

На третьому етапі методики оптимізації знаходимо оптимальну кількість серверів для випадку, коли в SLA-угодах прописується середній час відповіді на запит і/або максимальний час відповіді на запит для заданої долі всіх запитів з поділенням цих запитів на класи.

Для обмеження на середній час відповіді на запит, постановка задачі буде такою:

$$\max_{\{S_1,\ldots,S_C\}} F_{TCO}(S_1,\ldots,S_c) = \sum_{c=1}^{C} TCO_c \cdot S_c \tag{7}$$

$$\max_{\{S_1,\ldots,S_C\}} F_{ПЗ}(S_1,\ldots,S_C) = \sum_{c=1}^{C} \frac{S_c}{\frac{1}{K}\sum_{k=1}^{K} m_c^k} \tag{8}$$

при обмеженнях

$$\sum_{c=1}^{C} \frac{\lambda^k{}_c}{\lambda^k} \frac{m^k{}_c}{S_c - q_c} \le T^k{}_{SLA},\ k = \overline{1,K} \tag{9}$$

$$q_c < S_c \le S_c^{max},\ c = \overline{1,C} \tag{10}$$

де C - число кластерів ЦОД;

S_c - число серверів в кластері c;

$F_{ПС}(S_1,\ldots, S_C)$ - критерій пропускної здатності;

$F_{TCO}(S_1,\ldots, S_C)$ - критерій сукупної вартості володіння;

TCO_c - сукупна вартість володіння одним сервером кластера c; q_c - номінальне завантаження кластера c з одним сервером при заданому навантаженні;

m_c^k - середній час відповіді на запит класа k сервером кластера с;

$\frac{\lambda^k{}_c}{\lambda^k}$ - середня кількість відвідувань кластера c запитом класу k за час його перебування в системі;

T_{SLA}^k - середній час часу відповіді на запит класса k, яке задається в SLA-угоді;

S_c^{max} - максимально можлива кількість серверів в кластері c.

Для обмеження за максимальним часом відповіді на запит для заданої частки всіх запитів, постановка задачі буде такою:

$$\max_{\{S_1,\ldots,S_C\}} F_{TCO}(S_1,\ldots,S_c) = \sum_{c=1}^{C} TCO_c \cdot S_c \tag{11}$$

$$\max_{\{S_1,\ldots,S_C\}} F_{ПЗ}(S_1,\ldots,S_C) = \sum_{c=1}^{C} \frac{S_c}{\frac{1}{K}\sum_{k=1}^{K} m_c^k} \tag{12}$$

при обмеженнях $P\left[T^k \ge V^k\right] \le \varepsilon^k,\ k = \overline{1,K}$,

$$q_c < S_c \le S_c^{max},\ c = \overline{1,C}.$$

4. ВИСНОВКИ.

Запропонована методика може бути включена сервіс-провайдером у контур управління ресурсами ЦОД. При цьому план розподілу ресурсів ЦОД повинен складатися при змінах навантаження, при внесенні змін у систему угод, а також склад прикладних програм, наприклад, додаванні нової прикладної програми.

Пропонована методика оптимального розподілу ресурсів ЦОД дозволить сервіс-провайдеру, що надає послуги хостинга прикладних програм з різними видами SLA-угод, знаходити оптимальну кількість серверів у кластерах, і тим самим виконувати свої зобов'язання з забезпечення якості обслуговування.

1. Павлов О.А., Теленик С.Ф. Інформаційні технології та алгоритмізація в управлінні. – К., 2002. 2. Теленик С.Ф., Ролік О.І., Букасов М.М., Косован О.А., Кобец О.І. Моделі і методи розподілу ресурсів в системах з серверною віртуалізацією // Збірник наукових праць Військового інституту телекомунікацій та інформатизації Національного технічного університету України „Київський політехнічний інститут". – Випуск № 3. – Київ: ВІТІ НТУУ „КПІ", 2009. – С. 100 – 111. 3. Теленик С.Ф., Ролік О.І., Букасов М.М., Римар Р.В., Ролік К.О. Управління навантаженням і ресурсами центрів оброблення даних при виділених серверах // Автоматика. Автоматизація. Електротехнічні комплекси та системі. – 2009. – №2 (24). –С. 122 – 136. 4. D. Ardagna, M. Trubian, L. Zhang. SLA based profit optimization in multi-tier web application systems / Proc. of Int'l Conference On Service Oriented Computing, New York, NY. 2004. P. 173–182. 5. Zhang L., Ardagna D., SLA Based Profit Optimization in Web Systems,- 13th International Conference on World Wide Web (WWW'04), New York, USA, 2004. -pp. 462-463. 6. Urgaonkar B., Dynamic Resource Management in Internet hosting Platforms, Ph.D., University of Massachusetts Amherst, 2005. 7. Liu Z., Squillante M.S., Wolf J.L., Optimal Control of Resource Allocation in e-Business Environments with Strict Quality-of-Service Performance Guarantees, Proceedings of the 41st IEEE Conference on Decision and Control, 2002, pp. 4431- 4439 vol.4.