УДК 004.942(045)

В.М. Боровик, к.т.н., с.н.с.

МОДЕЛІ СИТУАЦІЙНОГО КЕРУВАННЯ ЗАПИТАМИ В СИСТЕМАХ «КЛІЄНТ–СЕРВЕР»

Національний авіаційний університет
E-mail: vborovik@ukr.net

Досліджено модель «клієнт–сервер» як систему масового обслуговування з ситуаційними пріоритетами. Визначено метод оптимізації пріоритетів та принципи практичного застосування результатів дослідження в системахі з розподіленим обробленням даних.

A models of client-server is explored with as a system of mass service with situational priorities. Determine method to optimize the priorities and principles for practical application of research results in systems with distributed data processing.

Исследована модель «клиент–сервер» как система массового обслуживания с ситуационными приоритетами. Определены метод оптимизации приоритетов и принципы практического применения результатов исследования в системах с распределенной обработкой данных.

Вступ

Моделі «клієнт–сервер» дозволили будувати гнучкі системи розподіленого оброблення даних. Розміщення серверів та клієнтів як на окремих комп’ютерах, так і на одному, де система одночасно є сервером і клієнтом [1], породжує проблеми їх взаємодії як у технічному аспекті побудови комп’ютерних мереж, так і у прикладному для побудови інформаційних систем та її важливої частини – системи баз даних.

Розглядаючи сучасні методи розрахунку та оптимізації параметрів традиційних чи мультисервісних систем і систем зв’язку, можна визначити аналогії у підході до моделі як системи масового обслуговування (СМО) з пріоритетами.

Загальним підходом дослідження таких систем є марковські моделі прийняття рішення [2; 3].

Предметом дослідження є моделі «клієнт–сервер» та їх математична модель СМО, де клієнти підключаються до реального серверу через проміжний керуючий орган – диспетчер [4]. Перевагою такої системи є гнучкість керування запитами за можливості оптимізації їх доступу до серверу.

Метод оптимізації керування такою системою як СМО ґрунтується на теорії марковських процесів із ситуаційними процесами.

Клас моделей враховує практичні ситуації роботи систем «клієнт–сервер».

Метою роботи є розроблення алгоритму роботи диспетчера з керування запитами.

Побудовану стохастичну модель розглянуто у стаціонарному режимі як детермінований процес, що дозволяє спростити процес керування та розрахунок пріоритетів.

Постановка проблеми

Під час розглядання системи «клієнт–сервер» як СМО введено проміжний словник переходу математичних термінів теорії СМО та означених комп’ютерних систем:

– система: мережа заданої структури;

– прилад: комп’ютер чи процесор;

– заява: запит;

– вхідні потоки заяв: запити на якусь дію;

– процес обслуговування: час виконання запитів;

– черга: набір запитів, які конкурують на виконання приладом (комп’ютером);

– типи вхідних заяв: типи запитів;

– дисципліна обслуговування (пріоритет): правило, яке дозволяє визначати послідовність виконання запитів;

– орієнтація (розігрів, підготовка): підготовча дія до виконання запиту чи типу запитів.

Черга може бути обмежена на загальну довжину за кількістю запитів або на довжину за кожним типом.

Час виконання запиту на приладі (комп'ютері) залежить від типу приладу (його потужності).

Якщо черга до приладу (комп'ютера) повністю заповнена, можливі дві ситуації:

1) запит не потрапляє у чергу і більше не розглядається, знову надійшовши до системи;

2) оптимізаційна задача виштовхування інших запитів із черги на основі першої групи ситуаційних пріоритетів.

Усі заяви з черги надходять на прилади (комп'ютери), якщо ті звільняються від попередньої заяви (запиту). Якщо є черга, розглядається друга оптимізаційна задача використання з попереднім розрахунком ситуаційних пріоритетів, які дозволяють оптимізувати цей процес.

Якщо на приладі зміняється тип запитів, можливе витрачання часу орієнтації на підготовчу дію з обслуговування наступного запиту. Можлива ситуація, коли орієнтація виконується завжди у разі зміни запитів на приладах.

Теоретично можливе враховування відмови приладів, якщо їх розглядати як потік запитів на прилади з найвищим пріоритетом.

Набір указаних параметрів, який впливає на структуру моделі, її складність та алгоритм існування, показано на рис. 1. На схемі виділяються базові параметри як обов'язкові для формування будь-якої моделі.

Розширені параметри розглядаються як додаткові, що вказують на більш складні структури моделі.

Таким чином будується структура моделі – її якісна складова.

На складність моделі впливають і кількісні параметри: кількість типів запитів, кількість приладів, тип орієнтації, врахування відмов приладів.

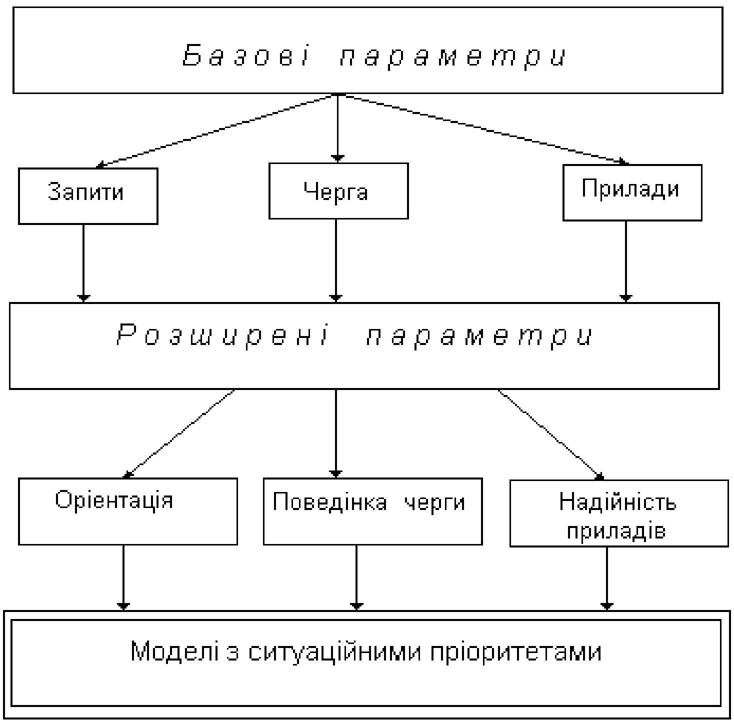

Рис.1. Види моделей

Математична модель

Усі заяви (запити) надходять на прилад (сервер чи серверний процес) у деякі випадкові моменти часу $t_0 < t_1 < t_2 \ldots < t_k \ldots < \ldots$.

Інтервал між їх появою

$$\tau_k = t_k - t_{k-1}$$

є незалежним випадковим значенням з експоненціальним законом розподілу:

$$F(t) = P\{\tau_k \le t\} = 1 - e^{-\lambda t},$$

де $\frac{1}{\lambda}$ – середній час між послідовними моментами появи двох однотипних запитів (рис. 2).

Далі будемо розрізняти типи заяв від відповідного клієнта, а саме, $\lambda\ (\overline{1,n})$ – інтенсивність появи заяв i-го типу (від i-го клієнта).

Таким чином, на сервер надходять n пуассонівських вхідних потоків різнотипних запитів.

Усі заяви, які надходять на прилад, можуть бути виконані на будь-якому з них, причому час їх виконання буде випадковим значенням з густиною ймовірності

$$\varphi(t) = \mu e^{-\mu t},$$

де $\frac{1}{\mu}$ – середній час виконання, який залежить від типу заяви.

У загальному випадку у разі багатоплатформеної архітектури час виконання залежить від типу запиту та номера приладу (комп'ютера, процесора), тому $\mu_{ij}\,(i = \overline{1,n};\ j = \overline{1,m})$.

У нашому випадку за однотипних серверів час виконання запиту не залежить від номеру приладу, тому $\mu_i\,(i = \overline{1,n})$.

За n>1 треба вже вирішувати альтернативні ситуації призначення заяв на обслуговування (виконання) на відповідному приладі.

Для формального визначення механізму вибору введемо поняття простору станів системи.

Простір станів системи описується векторами $(\vec{k},\vec{i})$, де $\vec{\boldsymbol{k}} = (k_1, k_2, ..., k_m)$ – номери типів заяв, які обслуговуються на кожному з m приладів, $\vec{\boldsymbol{i}} = (i_1, i_2, ..., i_n)$ – кількість заяв по кожному s типу в черзі.

У подальшому індексований параметр «i» буде визначати кількість заяв даного типу в черзі, а без індексу –номер заяви.

За $\vec{\boldsymbol{k}}\Big|_{k_j = 0,\,(j=\overline{1,m})}$ прилад вільний,

за $\vec{\boldsymbol{i}}\Big|_{i_s = 0,\,(s=\overline{1,n})}$ – у черзі немає заяви s-го типу.

Обмеження на довжину черги задаються двома параметрами:

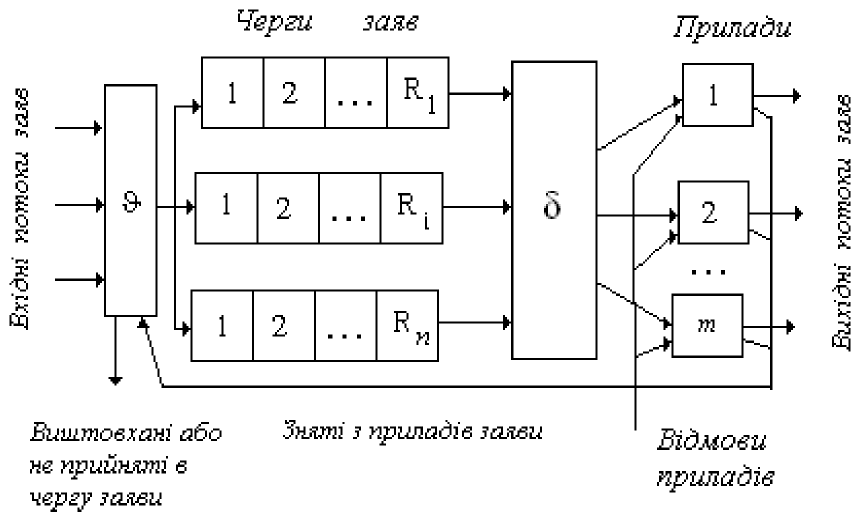

Рис. 2. Структура моделі СМО з ситуаційними пріоритетами

1) для роздільної черги – за кожним типом заяв

$$i_s \le r_s, \overline{(s=1,n)},$$

де r_s – максимально можливий розмір черги для заяв s-го типу;

2) для загальної черги всіх типів заяв

$$\sum_{s=1}^{n} i_s \le R,$$

де R – обмеження за розміром загальної черги.

У разі появи сигналу «прилад вільний» може статися ситуація, коли в черзі перебувають заяви більше одного типу. Для вирішення таких конфліктних ситуацій використовується пріоритетний параметр

$$\delta_s^j(\vec{\boldsymbol{k}}\Big|_{k_j=0,}\vec{\boldsymbol{i}}), (s=\overline{1,n};\ \ j=\overline{1,m}),$$

що визначає можливість вибору заяви s-го типу для обслуговування на j-му приладі.

Цей параметр є першою групою ситуаційних пріоритетів, які використовуються залежно від стану простору (ситуації) для постанови заяви для виконання на приладі.

Потрапляння у чергу є обов'язковою умовою обслуговування заяви, тому існує можливість її не обслужити на певному відрізку часу.

Таким чином, включення найбільш цінних заяв у чергу є гарантією їх участі у разі прийняття рішення про призначення чергової заяви на обслуговування. Для цього вводиться пріоритетний параметр

$$\vartheta_s^t(\vec{\boldsymbol{k}},\vec{\boldsymbol{i}}), (s,t=\overline{1,n}; s\ne t),$$

що визначає можливість виштовхування заяви t-го типу заявою s-го типу. Цей параметр є другою групою ситуаційних пріоритетів для встановлення заяв (запитів) у чергу.

На етапах «загибелі» (сидіння чи перескоків) одержуємо ланцюг, що має марковську властивість.

Для визначення стаціонарних імовірностей станів використовується граф-схема можливих переходів марковського процесу. Після цього для збалансованого випадку складаються рівняння для усіх стаціонарних імовірностей станів.

У разі неальтернативних ситуацій на етапах «розмноження» надходять пуассонівські вхідні потоки заяв, при тому виконуються обмеження на кількість місць у черзі.

У момент закінчення обслуговування неальтернативна ситуація відповідає існуванню у черзі заяв тільки одного типу. Для цієї ситуації, яка відноситься до тривіального випадку моделей «розмноження та загибелі», стан системи являє собою лінійний ланцюг. Закон вибору заяв на обслуговування у цьому випадку може бути вільним, що не змінює розподіл стаціонарних імовірностей станів системи.

Для визначеності будемо задавати, що для черги з заявами одного типу буде виконуватися дисципліна обслуговування FCFS – «перший прийшов – перший обслугований».

У разі надання інформації «прилад вільний» може з'явитися ситуація, коли у черзі перебувають заяви більше одного типу. Для вирішення таких конфліктних ситуацій використовується пріоритетний параметр $\delta^s(\vec{i}),(s=\overline{1,n})$, який визначає можливість вибору заяви s-го типу для обслуговування на приладі. Звичайно, що якась заява обов'язково буде вибрана, тому

$$\sum_{s=1}^{n} \delta^s(i_1, i_2, \dots, i_n) u(i_s) = 1,$$

$$\text{де } u(x) = \begin{cases} 1, \text{якщо } x > 0, \\ 0, \text{якщо } x \le 0. \end{cases}.$$

Тепер можна у векторній формі записати систему рівнянь для моделі

$M_n \left| M_1 \right| r_i \, (i=\overline{1,n})$ чи R:

$$\pi(0,\vec{0})\sum_{s=1}^{n}\lambda_s = \sum_{s=1}^{n}\mu_s\pi(s,\vec{0}); \qquad (1)$$

$$\pi(k,\vec{\mathbf{0}})[\sum_{s=1}^{n}\lambda_s + \mu_k] = \pi(0,\vec{\mathbf{0}})\lambda_k +$$

$$+\sum_{s=1}^{n}\mu_s\pi(s,\vec{\mathbf{0}}\Big|_{0_k=1}); \qquad (2)$$

$$\pi(k,\vec{\boldsymbol{i}})[\sum_{s=1}^{n}\lambda_{s}u(\Omega^{s})+\mu_{k}]=$$

$$=\sum_{s=1}^{n}u(i_{s})\lambda_{s}\pi(k,\vec{\boldsymbol{i}}\Big|_{i_{s}=i_{s}+1})+$$

$$+u(\Omega^{k})[u(\sum_{\substack{t=1\\t\neq k}}^{n}i_{t})\delta^{k}(\vec{\boldsymbol{i}}\Big|_{i_{k}=i_{k}+1})+$$

$$+[1-u(\sum_{\substack{t=1\\t\neq 1}}^{n}i_{t})]]\sum_{s=1}^{n}\mu_{s}\pi(s,\vec{\boldsymbol{i}}\Big|_{i_{k}=i_{k}+1}); \quad (3)$$

$$k=\overline{1,n};\quad \sum_{s=1}^{n}i_{s}>0; i_{s}=\overline{0,\Omega^{s}};\quad s=\overline{1,n};$$

$$\Omega^{s}=\begin{cases}\dfrac{(r_{s}-i_{s}),(s=\overline{1,n})-\text{роздільна}}{R-\sum\limits_{s=1}^{n}i_{s}-\text{загальна}}\end{cases}. \quad (4)$$

Нормувальну умову можна надати у вигляді:

$$\pi(0,\vec{\mathbf{0}})+\sum_{k=1}^{n}\sum_{\vec{i}}\pi(k,\vec{\boldsymbol{i}})=1. \quad (5)$$

В одержаній системі рівнянь $u(\Omega^{s})$ задає обмеження на кількість місць у черзі.

Запис $\pi(k,\vec{\boldsymbol{i}}\big|_{i_s=i_s-1})$ визначає перехід із стану $(k,i_1,i_2,\ldots,i_s,\ldots,i_n)$ у $(k,i_1,i_2,\ldots,i_s-1,\ldots,i_n)$, а $\pi(k,\vec{\boldsymbol{i}}\big|_{i_s=i_s+1})$ визначає перехід із стану $(k,i_1,i_2,\ldots,i_s,\ldots,i_n)$ у $(k,i_1,i_2,\ldots,i_s+1,\ldots,i_n)$, що описує відповідно вихід заяви з черги чи її появу в ній.

Перед початком обслуговування кожній заяві може бути необхідна «орієнтація», тобто підготовка приладу до обслуговування наступної заяви. Цей час у моделі також розподілено за експоненціальним законом із середнім $\frac{}{\mu^{i}}$ $(i=\overline{1,n};\ j=\overline{1,m})$.

Час обслуговування та орієнтації є незалежними випадковими значеннями. Ці процеси не можуть перериватися іншими заявами, тобто керування виконується тільки в класі відносних пріоритетів. Будемо розглядати два типи орієнтації:

1) орієнтація потрібна для обслуговування кожної заяви, що задає функціювання системи (обнуління орієнтації);

2) орієнтація потрібна для обслуговування заяви нового типу, що визначається теорією СМО як «дивись вперед».

Таким чином, з урахуванням часу орієнтації середній час обслуговування в таких моделях зміниться на

$$\frac{}{\mu^{i}}=\frac{}{\mu^{i}}+\frac{}{\mu^{i}}(i=\overline{1,n};\ j=\overline{1,m}).$$

У розглянутих моделях ураховується можливість відмови приладів у межах концепції побудови всіх моделей. Будемо враховувати, що з кожного приладу надходять відмови, які теж задаються пуассонівським законом із параметром

$$\lambda_{i}(i=\overline{n+1,N};N=n+m),$$

залежним від номера приладу.

У загальному випадку з кожного приладу надходять відмови різного типу, які потребують деякий час поновлення – випадкового значення з середнім часом обслуговування

$$\frac{}{\mu^{i}}(i=\overline{n+1,N};N=n+m).$$

Для того, щоб не ускладнювати дослідження моделей зайвою індексацією, розглянемо узагальнену відмову, яка зумовлює закриття приладу, але потребує негайного його поновлення.

Поява відмови по-різному впливає на заяву, яка в поточний момент обслуговується.

Вона буде безповоротно загублена чи потрібно її повторне оброблення.

У будь-якому випадку вводиться новий керуючий параметр

$$\sigma_{s}^{t}(\vec{\boldsymbol{k}},\vec{\boldsymbol{i}})(s\ t=\overline{1,n};\ s\neq t),$$

що дозволяє визначати можливість помістити в чергу s-й заяви, знятої з обслуговування через відмову приладу, замість t-ї, яка перебуває у черзі, але в ній немає вільних місць.

Далі будується модель за розглянутою методикою.

Визначення ситуаційних пріоритетів

Головним математичним об'єктом досліджень роботи є СМО зі скінченним числом можливих станів у стаціонарному режимі. Процес керування такими системами розглядається як пошук керуючих параметрів безупинного марковського ланцюга.

Для стаціонарних імовірностей моделі застосовують рівняння (1) – (5). Із теорії скінченних ланцюгів Маркова відомо, що всі стаціонарні ймовірності станів $\pi(\cdot)>0$, тобто більше нуля. Оптимальне значення керуючих елементів – $\delta(\cdot)$ можуть бути тільки 0 чи 1. Таким чином, оптимізаційна задача зводиться до пошуку екстремуму функції:

$$L^* = f(\pi^*(\cdot)) = \max\{L = f(\pi(\cdot)):$$

$$Q(\pi(\cdot), \delta(\cdot)) = 0;$$

$$0 < \pi(\cdot) < 1;$$

$$\delta(\cdot) \in \{0,1\};\ \sum_{P} \pi(\cdot) = 1\},$$

де $\pi^*(\cdot)$ – шукані оптимальні стаціонарні ймовірності стану системи;

L^* – оптимальне значення критерію ефективності;

Q – функціональний зв'язок стаціонарних імовірностей стану та керуючих пріоритетів $\delta(\cdot)$;

P – множина всіх можливих стаціонарних імовірностей стану, що визначаються типом та розмірністю моделі.

У найбільш складному варіанті додається пошук $\sigma(\cdot)$ та $\upsilon(\cdot)$ для підключення інших моделей.

Головним привабливим результатом теорії марковських процесів є оптимальне керування ним як детермінованим процесом, що дозволяє вирішувати задачу як відому задачу лінійного програмування. Це дозволяє складну математичну задачу звести до відносно простої практичної реалізації [6].

Висновки

Розглянута методологія може бути використана для побудови пріоритетного керування запитами у системах «клієнт–сервер».

Результати розрахунків подаються у вигляді налаштованої таблиці переходів чи бази даних ситуаційних пріоритетів, які заздалегідь розраховують за параметрами системи. За більш потужних систем із сервером додатків ця задача вирішується у реальному часі.

Такий підхід дозволяє розглядати систему як багаторівневу [5]. Подібну концепцію оброблення даних пропонують фірми Oracle, Sun, Borland та ін.

Література

1. *Дейт К. Дж.* Введение в системы баз данных / К. Дж. Дейт. – К.: Вильямс, 2006. – 1328 с.

2. *Меликов А.З.* Телетрафик: Модели, методы, оптимизация / А.З. Меликов, Л.А. Пономаренко, В. В. Паладюк. – К.: Політехніка, 2007. – 256 с.

3. *Меликов А.З.* Математические модели многопотоковых систем обслуживания / А.З. Меликов, Л.А. Пономаренко, П.А. Рюмшин. – К.: Техніка, 1991. – 265 с.

4. *Карпова Т.* Базы данных: модели, разработка, реализация / Т. Карпова. – С.Пб.: Питер, 2003. – 304 с.

5. *Петров В.Н.* Информационные системы / В. Н. Петров. – С.Пб.: Питер, 2002. – 688 с.

6. *Wolf P.* Linear programming in Markov chain / P. Wolf, G.B. Danzig // Operation research, 1963. – Vol.10, No.5. – P.702–710.

Стаття надійшла до редакції 30.08.10.