ЛЕКЦІЯ 8

ОБСЛУГОВУВАННЯ ЗАЯВОК ЗА ПРІОРИТЕТНИМИ ДИСЦИПЛІНАМИ

Література

1. Омельченко А.В. Основи аналізу систем розподілу інформації. Навч. посібник. – Харків: ХНУРЕ, 2008. – С 65-72

Основні поняття

- Дисципліна обслуговування з очікуванням, згідно з якою вибір заявок для обслуговування проводиться у відповідності зі ступенем їх важливості, називається пріоритетною дисципліною.

- У будь-якій пріоритетній дисципліні мають бути визначені правила для прийняття таких рішень.
- 1. Яку заявку брати на обслуговування в момент готовності приладу для прийняття наступної заявки.
- 2. Продовжити або перервати обслуговування заявки, що перебуває в приладі.

- Вважатимемо, що ступінь важливості заявки встановлюється за допомогою приписування кожному класу пріоритетного індексу i:

$$1 \leq i \leq r,$$

де 1 позначає найвищий ступінь важливості,

а *r* – найнижчий.

Дисципліни пріоритетного обслуговування

Можливі такі дисципліни пріоритетного обслуговування:

- 1. Відносний пріоритет (пріоритет без переривання обслуговування): обслуговування заявки будь-якого класу триває до повного завершення.
- 2. Абсолютний пріоритет (пріоритет, що перериває обслуговування): обслуговування заявки нижчого класу негайно переривається, і прилад починає обслуговувати заявку більш важливого класу.
- 3. Динамічні пріоритети: кожній вхідній заявці призначається певний індекс пріоритету залежно від стану CPI.

Закон збереження роботи

- Незавершеною роботою R(t) у момент часу t у теорії черг називається час, який має пройти до повного звільнення системи від усіх заявок, якщо після моменту часу t на її вхід не поступають нові заявки.

- Консервативною називається система, в якій заявки не зникають усередині системи і прилади, що обслуговують, не простоюють при непустій черзі.

- Закон збереження роботи для пріоритетних дисциплін стверджує, що для консервативної системи незакінчена робота R(t) в СРІ в будь-який момент часу t не залежить від порядку обслуговування.

- Розподіл часу очікування в загальному випадку істотно залежить від порядку обслуговування. Однак, якщо дисципліна обслуговування вибирає заявки незалежно від їхнього часу обслуговування, то середній час очікування у черзі є інваріантним щодо порядку обслуговування.

Закон збереження роботи

- Закон збереження для системи типу M/G/1 формулюється так.
- Для будь-якої системи M/G/1 і будь-якої відносної дисципліни обслуговування, що зберігає роботу, має виконуватися рівність

$$\sum_{i=1}^{r} \frac{y_i}{y} \cdot \overline{T}_{оч.i} = \overline{T}_{оч}, \qquad (1)$$

де y_i – інтенсивність навантаження, що створюється заявками i-ї категорії;

$\overline{T}_{оч.i}$ – середній час очікування для заявок i-ї категорії;

y – загальне навантаження;

$\overline{T}_{оч}$ – загальний середній час очікування заявок.

Середній час очікування в черзі

Для однолінійної системи з формули Полячека-Хінчина випливає компактний вираз для середнього часу очікування в черзі

$$\overline{T}_{оч} = \frac{\lambda \cdot \overline{T_{об}^{\,2}}}{2(1-y)}, \qquad (2)$$

де $T_{об}$ – час обслуговування однієї заявки;

$\overline{T}_{об}$ – середній час обслуговування заявки;

$\overline{T_{об}^{\,2}}$ – другий початковий момент часу обслуговування заявки.

Відносний пріоритет. Одноканальна СРІ

Нехай заявки з r пріоритетами надходять в одноканальну СРІ. Заявки кожного з класів утворять пуассонівські потоки з параметрами λ_i, $i = 1,...,r$.

Отже, загальний потік є пуассонівським з параметром $\lambda = \sum_{i=1}^{r} \lambda_i$.

Припустимо, що в межах даного пріоритету заявки обслуговуються у порядку надходження (дисципліна черги FIFO).

Функція розподілу часу обслуговування заявок i-го класу має довільний вигляд $F_i(t)$ і характеризується середнім часом обслуговування $\overline{T}_{об} = 1/\mu_i$. При цьому функція розподілу загального часу обслуговування

$$F(t) = \frac{1}{\lambda} \sum_{i=1}^{r} \lambda_i \cdot F_i(t). \qquad (3)$$

Необхідно визначити середній час очікування $\overline{T}_{оч.i}$ у черзі для заявок i-го пріоритету.

Вирішення сформульованої задачі

Згідно з формулою Літтла, середній час очікування заявки і-го пріоритету задовольняє рівнянню

$$\overline{T}_{оч.i} = \overline{T}_o + \sum_{k=1}^{i} \frac{1}{\mu_k} \cdot \lambda_k \cdot \overline{T}_{оч.k} + \sum_{k=1}^{i-1} \frac{1}{\mu_k} \cdot \lambda_k \cdot \overline{T}_{оч.i}, \qquad (4)$$

де $\overline{T}_o$ – середній час, необхідний для закінчення обслуговування заявки, що вже перебуває на обслуговуванні на момент надходження заданої (міченої) заявки;

$\sum_{k=1}^{i} \frac{1}{\mu_k} \cdot \lambda_k \cdot \overline{T}_{оч.k}$ – середній час обслуговування заявок із пріоритетом не нижче i, що вже перебувають у черзі;

$\sum_{k=1}^{i-1} \frac{1}{\mu_k} \cdot \lambda_k \cdot \overline{T}_{оч.i}$ – середній час обслуговування заявок із пріоритетом вище i, що надходять у чергу за час очікування міченої заявки.

Відносний пріоритет. Одноканальна система Середній час очікування у черзі.

З рівняння (4) одержимо

$$\overline{T}_{оч.i} = \frac{\overline{T}_o + \sum_{k=1}^{i} y_k \cdot \overline{T}_{оч.k}}{1 - \sigma_{i-1}}, \qquad (5)$$

де

$$\sigma_i = \sum_{k=1}^{i} y_k, \quad i > 0; \quad \sigma_0 = 0; \qquad (6)$$

$$y_k = \frac{\lambda_k}{\mu_k}, \quad k = \overline{1,r}.$$

Звідси методом математичної індукції одержимо необхідне співвідношення

$$\overline{T}_{оч\ i} = \frac{\overline{T}_o}{(1 - \sigma_{i-1}) \cdot (1 - \sigma_i)}. \qquad (7)$$

Суть методу математичної індукції

Застосування методу математичної індукції полягає в тому, що якщо справедлива рівність

$$T_{оч\ i-1} = \frac{\overline{T}_o}{(1-\sigma_{i-2})\cdot(1-\sigma_{i-1})}, \qquad (8)$$

то згідно з (5), маємо

$$\overline{T}_{оч.i} = \frac{\overline{T}_o + \sum_{k=1}^{i-1} y_k \cdot \overline{T}_{оч.k} + y_i \cdot \overline{T}_{оч.i}}{1-\sigma_{i-1}} = \frac{(1-\sigma_{i-2})\cdot \overline{T}_{оч.i-1} + y_i \cdot \overline{T}_{оч.i}}{1-\sigma_{i-1}}. \qquad (9)$$

З урахуванням (8) одержимо

$$\overline{T}_{оч\ i}(1-\sigma_{i-1}-y_i) = \frac{\overline{T}_o}{(1-\sigma_{i-1})}. \qquad (10)$$

Звідки одержуємо шукане співвідношення (7).

Відносний пріоритет. Одноканальна система Середній час очікування у черзі.

Таким чином, середній час очікування початку обслуговування заявок i-го пріоритету визначається формулою

$$\overline{T}_{оч\ i} = \frac{\overline{T}_o}{(1-\sigma_{i-1})\cdot(1-\sigma_i)}, \qquad (11)$$

де $\overline{T}_o$ – час, необхідний для закінчення обслуговування заявки, що перебуває на обслуговуванні в довільний момент часу.

При $r=1$ маємо класичний випадок одноканальної системи без пріоритету й $\overline{T}_{оч} = \frac{\overline{T}_o}{(1-y)}$, звідки з врахуванням (2) знаходимо величину

$$\overline{T}_o = \frac{\lambda}{2}\cdot\int_0^\infty t^2 dF(t). \qquad (12)$$

З формули (11) видно, що зі зменшенням пріоритетності, тобто зі зростанням індексу i, збільшується й середній час очікування для відповідного класу.

Багатоканальна система

У даному випадку аналіз системи проводиться аналогічно. Розглянемо вхідний потік того самого типу, і нехай час обслуговування в кожному з v каналів має однаковий розподіл з параметром μ . Дослідження проводиться, як і раніше, з тією лише відмінністю, що коли всі канали зайняті, заявки, що обслужені, залишають систему випадково з інтенсивністю $v\mu$.

Аналогічно попередньому маємо

$$\overline{T}_{оч\_i} = \frac{\overline{T}_o}{(1-\frac{1}{v}\cdot\sum_{k=1}^{i-1} y_k)\cdot(1-\frac{1}{v}\cdot\sum_{k=1}^{i} y_k)}. \qquad (13)$$

Багатоканальна система

У цьому виразі $\overline{T}_o$ можна знайти за допомогою формули для середнього часу очікування у багатоканальній системі типу M/M/v/W, що перебуває в стаціонарному стані, при обслуговуванні заявок у порядку надходження. При $r = 1$ з виразу (13) маємо

$$\overline{T}_{оч} = \frac{\overline{T}_o \cdot v}{(v - y)},$$

а з іншого боку,

$$\overline{T}_{оч} = \frac{1}{\mu \cdot (v - y)} \cdot \frac{E_v(y)}{1 - \frac{y}{v} \cdot (1 - E_v(y))}.$$

Звідси маємо

$$\overline{T}_o = \frac{1}{\mu} \cdot \frac{E_v(y)}{v - y \cdot (1 - E_v(y))}. \qquad (14)$$

Висновок

- На основі аналізу отриманих співвідношень можна зробити такий висновок.

- Якщо система з пріоритетом завантажена слабко, то різниця в середньому часі очікування для заявок з різними пріоритетами мала, але вона стає помітною, коли навантаження на систему зростає. Наприклад, збільшення інтенсивності потоку заявок з найбільшим пріоритетом приводить до зростання часу очікування як для заявок із цим пріоритетом, так і для інших заявок. Тому важливо контролювати призначення високих пріоритетів і за можливістю зменшувати час обслуговування заявок.

Абсолютний пріоритет. Одноканальна СРІ

Розглянемо випадок абсолютних пріоритетів з дообслуговуванням заявок.

Позначимо через $\overline{T}_i$ середній загальний час перебування в системі заявок із класу i. Він складається з трьох складових.

1. Середній час обслуговування $\overline{T}_{об.i}$.

2. Затримка через обслуговування тих заявок рівного або більш високого пріоритету, що дана (мічена) заявка застала в системі. Відповідно до закону збереження роботи для обслуговування цих заявок мічена заявка в середньому затримується в черзі на час $\frac{\sum_{k=1}^{i} \lambda_k \cdot \overline{T^2_{об.k}}}{2(1-\sigma_i)}$.

3. Середня затримка міченої заявки за рахунок заявок, що належать більш високим пріоритетним класам і надходять у систему за час перебування у черзі міченої заявки. Середнє число таких заявок з k-го класу дорівнює $\lambda_k \cdot \overline{T}_i$ й кожна з них затримується на час $\overline{T}_{об.k}$.

Абсолютний пріоритет. Одноканальна СРІ

Таким чином,

$$\overline{T}_i = \overline{T}_{об.i} + \frac{\sum_{k=1}^{i} \lambda_k \cdot \overline{T^2_{об.k}}}{2(1-\sigma_i)} + \sum_{k=1}^{i-1} y_k \cdot \overline{T}_i . \qquad (15)$$

Отже, рішення для $\overline{T}_i$ має вигляд

$$\overline{T}_i = \frac{\overline{T}_{об.i}.}{(1-\sigma_{i-1})} + \frac{\sum_{k=1}^{i} \lambda_k \cdot \overline{T^2_{об.k}}}{2 \cdot (1-\sigma_{i-1}) \cdot (1-\sigma_i)} . \qquad (16)$$

Слід зазначити, що для системи з відносними пріоритетами

$$\overline{T}_i = \overline{T}_{об.i}. + \frac{\sum_{k=1}^{r} \lambda_k \cdot \overline{T^2_{об.k}}}{2 \cdot (1-\sigma_{i-1}) \cdot (1-\sigma_i)} . \qquad (17)$$

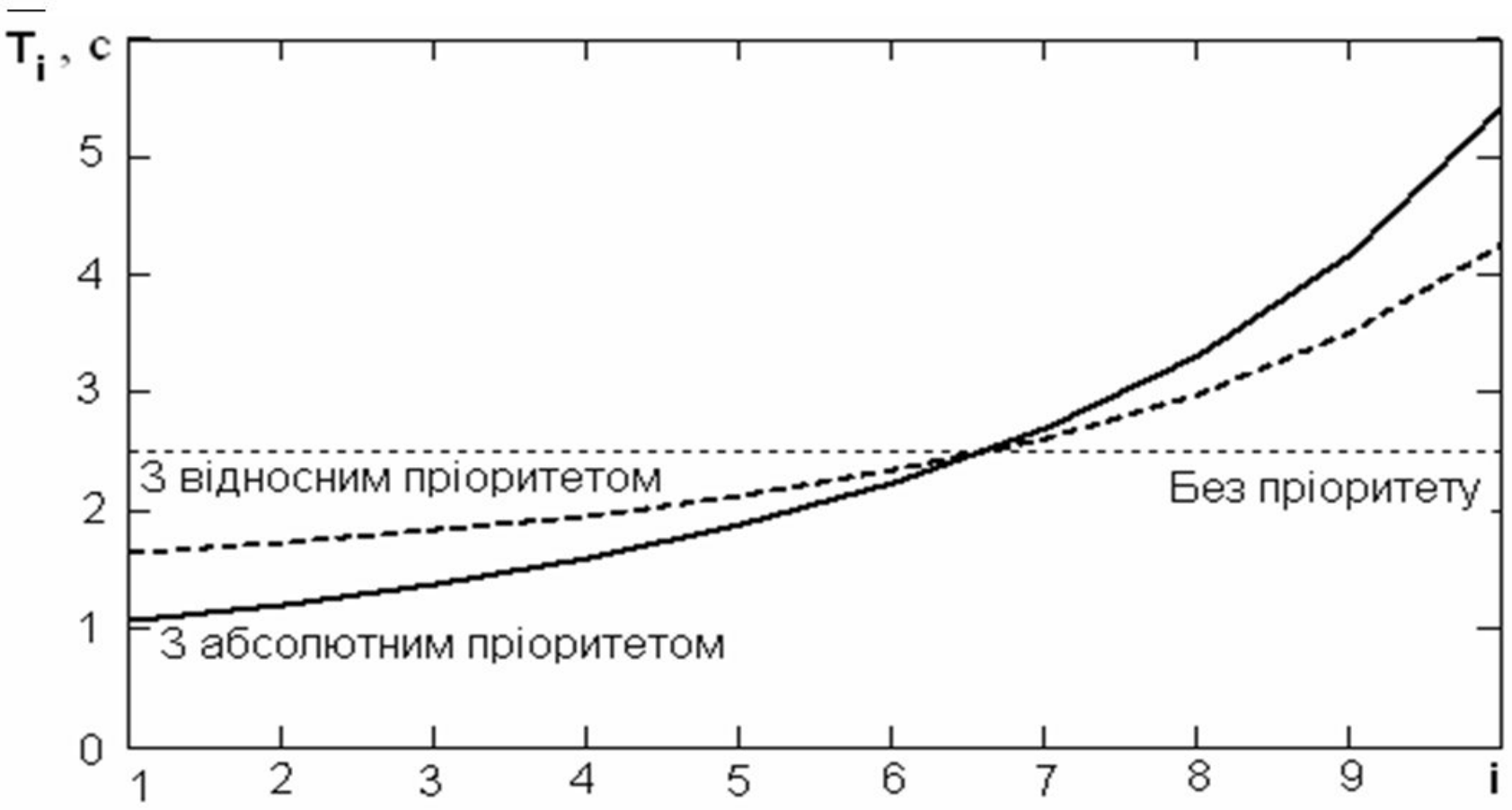
Порівняльної оцінки дисциплін обслуговування

- Для порівняльної оцінки дисциплін обслуговування на наступному слайді наведені залежності часу затримки заявок у системі від їх пріоритету. При цьому використані такі дані:

$$r = 10;\ \lambda_i = 0{,}06\ \ c^{-1};\ T_{об.i} = 1\ c\ \ ,\ i = \overline{1,r}\,.$$

- Вважається, що час обслуговування заявок є випадковим і має експоненціальний закон розподілу.

Залежність часу очікування в черзі від пріоритету



Висновок: Дисципліна обслуговування з абсолютним пріоритетом у більшій мірі зменшує затримки високопріоритетних заявок за рахунок збільшення затримок заявок з низьким пріоритетом.

Задача оптимізації призначення пріоритетів

Розглянемо випадок відносних пріоритетів. Припустимо, що вартість використання системи дорівнює C_i за кожну секунду затримки заявок i-го класу. Зрозуміло, що середня вартість використання системи на протязі секунди, яку позначимо через C, дорівнює

$$C = \sum_{i=1}^{r} C_i \cdot \lambda_i \cdot \overline{T}_i, \qquad (18)$$

де λ_i – інтенсивність потоку викликів i-го класу;

$\overline{T}_i = \overline{T}_{оч.i} + \overline{T}_{об.i}$ – середня затримка заявок і-го класу в системі.

Отже, справедливе співвідношення

$$C = \sum_{i=1}^{r} y_i \cdot C_i + \sum_{i=1}^{r} \lambda_i \cdot \overline{T}_{оч.i} \cdot C_i . \qquad (19)$$

Необхідно відшукати таку дисципліну обслуговування з відносним пріоритетом, що мінімізує С.

Призначення пріоритетів

Вирішимо цю задачу для системи M/G/1 з r пріоритетними класами.

Сформульована задача зводиться до мінімізації другої складової в (19), яка може бути представлена у вигляді добутку функцій

$$\sum_{i=1}^{r} \lambda_i \cdot \overline{T}_{оч.i} \cdot C_i = \sum_{i=1}^{r} f_i \cdot g_i , \qquad (20)$$

де функції $g_i = y_i \cdot \overline{T}_{оч.i}$; $f_i = \dfrac{C_i}{\overline{T}_{об.i}} \cdot$

Тут відповідно до закону збереження роботи $\sum_{i=1}^{r} g_i = const$ стосовно дисципліни обслуговування.

Оптимальне призначення пріоритетів

Необхідно призначенням пріоритетів так перерозподілити значення позитивні величини g_i, щоб найменші значення g_i відповідали найбільшим f_i. Ця умова виконується, якщо найвищі пріоритети призначаються найбільшим значенням $f_i = \frac{C_i}{\overline{T}_{об.i}}$, тобто тим класам, для яких відношення вартості затримки до середнього часу обслуговування заявок максимальне. Після призначення пріоритетності одержимо $f_1 \geq f_2 \geq \ldots \geq f_r$.

Таким чином, вищий рівень пріоритету раціонально призначати тим класам, які мають вищу вартість затримок заявок і вимагають меншого середнього часу обслуговування.

Приклад системи з пріоритетами в системах зв'язку

Розглянемо систему передачі інформації з двома типами пакетів, що відповідають передачі мови й даних. Припустимо, що прибуття цих пакетів утворює два пуассонівські потоки з інтенсивностями λ_1 для пакетів мови та λ_2 для пакетів з даними. Припустимо, що час передачі для пакетів мови розподілений як випадкова величина із середнім $\overline{T}_{об.1}$, а час передачі пакетів з даними розподілений як випадкова величина з середнім $\overline{T}_{об.2}$.

Пакети мови мають відносний пріоритет стосовно пакетів даних.

Приклад системи з пріоритетами в системах зв'язку

Тоді середній час очікування початку обслуговування становить:

– для пакеті мови

$$\overline{T}_{оч.1} = \frac{\lambda_1 \cdot M[T^2_{об.1}] + \lambda_2 \cdot M[T^2_{об.2}]}{2(1 - y_1)};$$

– для пакетів даних

$$\overline{T}_{оч.2} = \frac{\lambda_1 \cdot M[T^2_{об.1}] + \lambda_2 \cdot M[T^2_{об.2}]}{2(1 - y_1)(1 - y_1 - y_2)},$$

де $y_i = \lambda_i \cdot T_{об.i}$, $i = 1,2$.

Приклад системи з пріоритетами в системах зв'язку

Припустимо, наприклад, що $\lambda_1 = 300\ \frac{пакетів}{с}$, $\lambda_2 = 10\ \frac{пакетів}{с}$, час обслуговування пакетів мови є постійнім $T_{об.1} = 2\ мс$, а час обслуговування пакетів даних має експоненціальний розподіл з середнім значенням $\overline{T}_{об.2} = 20\ мс$. Тоді $M[T^2_{об.2}] = 8 \cdot 10^{-4}$ с^2. Підставляючи дані в наведені вище вирази, одержимо $\overline{T}_{оч.1}$=11,5 мс і $\overline{T}_{оч.2}$=57,5 мс.

Порівняння цих величин демонструє вплив призначення пріоритетів на середній час затримки заявок у черзі.