



МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Дніпропетровський національний університет
залізничного транспорту імені академіка
В. Лазаряна

Кафедра «Електронні обчислювальні машини»

АНАЛІТИЧНІ МОДЕЛІ МАСОВОГО ОБСЛУГОВУВАННЯ В ЗАДАЧАХ ПРОЕКТУВАННЯ ІНФОРМАЦІЙНИХ СИСТЕМ

Навчально-довідковий посібник
для курсового проектування, контрольних і практичних робіт

Укладач: А.А. Косолапов

*Для студентів IV курсу денної форми навчання
спеціальностей «Комп'ютерні системи»
(напрямок підготовки 6.050102 «Комп'ютерна
інженерія») і «Безпека інформаційних і комуні-
каційних систем» (напрямок підготовки 6.170101
«Безпека інформаційних і комунікаційних сис-
тем»)*

Дніпропетровськ 2015

ББК 629.08
К 33
УДК 681.5

Укладач:
Косолапов Анатолій Аркадійович

Рецензенти:
д-р техн. наук, проф. *О. І. Міхальов*
(Національна металургійна академія України)
д-р техн. наук, проф. *В.С. Хандецький* (ДНУ)

К 33 Аналітичні моделі масового обслуговування в задачах проектування інформаційних систем [Текст] : навчально-довідковий посібник / уклад.: А. А. Косолапов; Дніпропетр. нац. ун-т залізн. трансп. ім. акад. В. Лазаряна. – Д.: «LikePrint», ФОП Гечка Т.О., 2015. – 186 с.

Розглядаються теоретичні основи побудови і базовий набір аналітичних моделей систем і мереж масового обслуговування (СМО і ММО), а також приклади їх використання в процесі аналізу і проектування комп'ютерних систем і мереж. Простота і прийнятна точність моделей дозволяють реалізувати їх на мобільних пристроях і планшетах і використовувати результати для оперативного прийняття рішень. Посібник можна використовувати як довідковий посібник з аналітичних моделей масового обслуговування. Посібник забезпечено матеріалами для самостійної роботи: узагальненням матеріалу для модульного контролю, тестами, задачами, питаннями для самоконтролю, які допоможуть студентам сформулювати необхідні практичні навички.

Структура посібника відповідає навчальній програмі дисципліни «Комп'ютерні системи» для студентів за напрямком підготовки «Комп'ютерна інженерія».

Розраховано на студентів, магістрів та аспірантів, викладачів, проектувальників інформаційних систем та мереж на транспорті.

ББК 629.08

© Косолапов А. А., укладання, 2015

ЗМІСТ

ВСТУП.....	6
РОЗДІЛ I	11
МОДЕЛІ СМО. ЗАГАЛЬНІ ПОЛОЖЕННЯ.....	11
1.1. Система обслуговування сервера-одинака	11
1.1.1. Параметри СМО.....	12
1.1.2. Основні позначення	13
1.2. Система обслуговування мульти-сервера.....	17
1.3. Загальні положення СМО.....	19
1.3.1. Базові співвідношення	19
1.3.2. Задачі аналізу СМО.....	20
1.3.3. Опис СМО в нотації Кендала.....	21
1.3.4. СМО сервера-одинака	21
РОЗДІЛ II	26
СМО З ОДНОРІДНИМ ПОТОКОМ ЗАЯВОК.....	26
2.1. ОСМО - одноканальні СМО	26
2.1.1. СМО сервера-одинака з однорідним потоком заявок.....	26
2.1.2. Характеристики ОСМО М/М/1	27
2.1.3. Характеристики ОСМО М/Г/1	29
2.1.4. Характеристики ОСМО Г/М/1	30
2.1.5. Характеристики ОСМО Г/Г/1	32
2.1.6. Аналіз властивостей ОСМО з ОПЗ	34
2.2. Багатоканальні СМО (БСМО).....	35
2.2.1. Характеристики БСМО М/М/К	36
2.2.2. Основні формули	37
2.2.3. Аналіз властивостей БСМО	38
2.3. Самоконтроль. Перелік питань і завдань	41
РОЗДІЛ III.....	43
ОРГАНІЗАЦІЯ РОБОТИ СМО З НЕОДНОРІДНИМ ПОТОКОМ ЗАЯВОК.....	43
3.1. Класифікація стратегій управління в СМО з НПЗ	43
3.2. Методи управління в аналітичних моделях СМО.....	49

3.3. Характеристики та властивості ДО БП	50
3.4. Характеристики та властивості ДО ВП.....	53
3.5. Характеристики та властивості ДО АП	57
3.6. Закони збереження.....	60
3.7. Самоконтроль. Перелік питань і завдань	62
РОЗДІЛ IV	66
МЕРЕЖІ МАСОВОГО ОБСЛУГОВУВАННЯ. ТЕОРЕТИЧНІ ЗАСАДИ	66
4.1. Поділ та об'єднання потоків трафіка	67
4.2. СМО тандемів. Теорема Бьорке.....	68
4.3. Теорема і відкриті мережі Джексона (Jackson)	71
4.4. Розкладання часу мережевої затримки по каналах	75
4.5. Двопараметрична модель часу затримки повідомлень.....	79
4.6. Задача вибору пропускних здатностей каналів	82
РОЗДІЛ V	88
ЕКСПОНЕНЦІЙНІ ММО З ОПЗ	88
5.1. Розімкнені експоненційні ММО з ОПЗ.....	88
5.1.1. Опис розімкнених ММО	88
5.1.2. Розрахунок коефіцієнтів передач і інтенсивностей потоків заявок у вузлах РММО	90
5.1.3. Перевірка умови відсутності перевантажень в ММО.....	91
5.1.4. Розрахунок вузлових характеристик РММО.....	91
5.1.5. Розрахунок мережевих характеристик РММО.....	93
5.1.6. Аналіз властивостей розімкнених ММО	97
5.2. Замкнуті експоненційні ММО з ОПЗ.....	102
5.2.1. Опис замкнутих ММО	102
5.2.2. Розрахунок коефіцієнтів передач у вузлах ЗММО	104
5.2.3. Розрахунок характеристик ЗММО	105
5.2.4. Аналіз властивостей ЗММО.....	110
5.3. Самоконтроль. Перелік питань і завдань	115
РОЗДІЛ VI	118
УЗАГАЛЬНЕННЯ МАТЕРІАЛУ ДЛЯ МОДУЛЬНОГО КОНТРОЛЮ	118
РОЗДІЛ VII	128
ПРАКТИКУМ. РІШЕННЯ ЗАДАЧ	128
СПИСОК ЛІТЕРАТУРИ.....	145
Додаткова література.....	146
ДОДАТОК I.....	147

ЕЛЕМЕНТИ ТЕОРІЇ ЙМОВІРНОСТІ.....	147
Д.1. Основні поняття і визначення	147
Д.1.1. Подія, ймовірність	147
Д.1.2. Випадкова величина.....	148
Д.2. Закони розподілення випадкових величин	149
Д.2.1. Закон розподілу дискретної випадкової величини	150
Д.2.2. Закон розподілу безперервної випадкової величини	151
Д.3. Числові характеристики випадкових величин	155
Д.3.1. Початкові моменти.....	155
Д.3.2. Центральні моменти.....	156
Д.4. Твірна функція і перетворення Лапласа	159
Д.4.1. Твірна функція	159
Д.4.2. Перетворення Лапласа	160
Д.5. Типові розподіли випадкових величин	160
Д.5.1. Розподіл Пуасона	161
Д.5.2. Геометричний розподіл	162
Д.5.3. Рівномірний закон розподілу	163
Д.5.4. Експоненційний закон розподілу	164
Д.5.5. Розподіл Ерланга	166
Д.5.6. Нормований розподіл Ерланга	168
Д.5.7. Гіперекспоненційний розподіл	169
Д.5.8. Гіперерлангівський розподіл.....	173
Д.6. Практикум. Рішення задач.....	176
Д.7. Самоконтроль. Перелік питань і задач	180
ПРЕДМЕТНИЙ ПОКАЖЧИК	183

ВСТУП

Аналіз систем масового обслуговування (СМО) - один з найбільш важливих процесів для тих, хто пов'язаний з аналізом комп'ютерів і мереж. Він може використовуватися для отримання наближених відповідей на безліч запитань, таких, як:

- що трапиться під час отримання файлу, якщо диск буде переважаний?
- чи зміниться час відповіді інформаційної системи, якщо подвоїти тактову частоту її процесора з одночасним подвоєнням кількості клієнтів, що обслуговуються?
- скільки ліній може обслужити система, що має одну комутовану лінію?
- скільки терміналів можна підключити до он-лайн центру обслуговування і який час очікування будуть мати оператори терміналів центру?

Ряд питань, які можна вирішити за допомогою аналізу СМО, нескінченний і зачіпає практично всі галузі комп'ютерних наук і інформаційних технологій. Можливість зробити такий аналіз за допомогою цього інструментарію є важливим для тих, хто бере участь в цих галузях.

Незважаючи на те, що теорія черг є математично складною, застосування її для аналізу продуктивності у багатьох випадках надзвичайно просте. Для цього потрібні знання елементарних статистичних понять (середні значення і стандартні відхилення), а також розуміння основних положень теорії СМО.

Озброївшись цим, аналітик може робити аналіз СМО «на зворотному боці конверта» або на калькуляторі мобільного телефону з допомогою простих формул аналізу СМО, або з використанням простих комп'ютерних програм, які займають лише кілька рядків коду.

Але чому саме аналіз систем масового обслуговування?

Існує багато випадків, коли важливо вміти прогнозувати вплив де-

яких змін під час проектування або модернізації інформаційних систем: або очікуване збільшення навантаження на систему, або зменшення часу обслуговування заявок, або інші зміни в проекті, що передбачаються. Наприклад, організація підтримує низку терміналів, персональних комп'ютерів і робочих місць на 100 Мбітній локальній мережі (LAN). В будівлі створюється новий офіс з додатковими робочими місцями. Чи зможе існуюча мережа впоратися із підвищенням обсягу роботи, чи краще організувати нову LAN і встановити між ними мост?

Є й інші випадки, коли створюється нова система и необхідно оцінити її характеристики, враховуючи існуючий досвід в цій галузі. Наприклад, департамент має намір оснастити всіх своїх співробітників персональними комп'ютерами (ПК) і об'єднати їх у локальну мережу з файловим сервером. Спираючись на досвід в інших системах, можна вирахувати навантаження файлового серверу кожним ПК.

Важливою характеристикою є продуктивність систем. В інтерактивному режимі або в реальному часі роботи додатків суттєвою характеристикою є час відгуку. В інших випадках, важливою є пропускна здатність системи. У будь-якому випадку, прогнози продуктивності повинні бути зроблені на основі наявної інформації навантаження або на основі розрахункового навантаження для нового обладнання.

Для вирішення подібних завдань існує велика кількість різноманітних підходів:

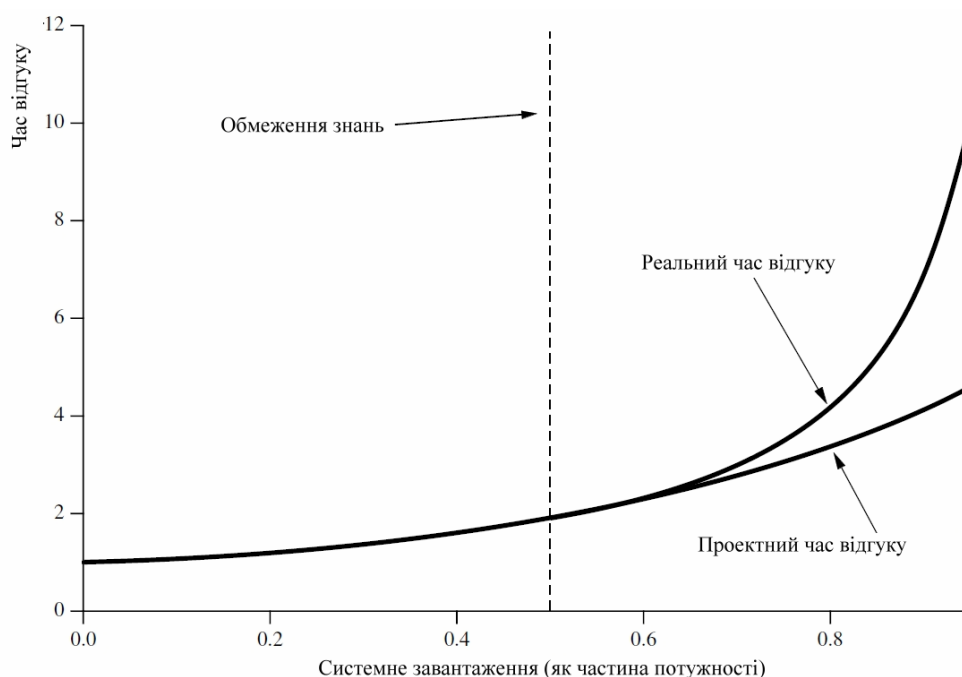
1. Впровадити систему і на основі реальних даних розрахувати її характеристики.
2. Зробити простий експеримент і його результати відобразити на характеристики реального обладнання.
3. Розробити аналітичну модель на основі теорії СМО.
4. Використати програму імітаційного моделювання системи (AnyLogic, Extendsim, тощо).

Варіант 1 не варіант взагалі: ми почекаємо впровадження системи (!) і тоді дивимось, що відбувається. Це призводить до нещасних випадків або до нерозумних витрат.

Варіант 2 здається більш перспективним. Аналітик може зайняти позицію, що неможливо спроектувати майбутню систему з яким-небудь ступенем визначеності. У зв'язку з цим безглуздо намагатися робити деякі точні процедури моделювання. Швидше за все, груба проекція моделі буде готова забезпечити тільки приблизні оцінки.

Проблема такого підходу полягає в тому, що поведінка більшості систем при зміні навантаження не може бути передбачена інтуїтивно.

Якщо є система, в якій спільно використовуються різні об'єкти (наприклад, мережі, лінії електропередачі, системи з поділом часу), то продуктивність цієї системи зазвичай експоненційно залежить від збільшення попиту. Наведений рисунок являє собою типовий приклад [1].



Проектний та реальний час відгуку

Верхня лінія показує, що відбувається з часом відгуку користувачів, коли збільшується навантаження на об'єкт. Навантаження виражається як частка потужності. Таким чином, якщо ми розглядаємо диск, який здатний передавати 1000 блоків за секунду, то навантаження 0,5 являє собою передачу 500 блоків за секунду, а час відгуку – це час, необхідний для ретрансляції цих блоків. Нижня лінія проста проекція на основі знань про поведінку системи до навантаження 0,5. Але в реальних умовах спостерігається колапс системи, коли навантаження близько до 0,8 до 0,9. Таким чином, потрібен інструмент для більш точного прогнозу.

Варіант 3 полягає у використанні аналітичної моделі, яка може мати вигляд набору рівнянь. Вони можуть бути вирішені з отриманням бажаних параметрів (час відгуку, пропускну здатність і т.д.). Для

комп'ютерів, операційних систем і мережних проблем, та для багатьох практичних реальних проблем, аналітичні моделі, засновані на теорії масового обслуговування (ТМО), забезпечують результати, які достатньо добре узгоджуються з реальністю. Недоліком ТМО є те, що необхідно зробити ряд спрощуючих припущень для отримання рівнянь для обчислення необхідних характеристик.

Останній підхід є використання імітаційної моделі. Тут, з огляду на досить потужну і гнучку мову моделювання, аналітик може моделювати реальність в найдрібніших подробицях і уникнути багатьох припущень, що вимагає ТМО. Тим не менш, в більшості випадків, імітаційна модель не потрібна, або принаймні, не рекомендується в якості першого кроку в процесі аналізу. З одного боку, існуючі вимірювання, і прогнози майбутнього навантаження несуть в собі певні похибки. Таким чином, незалежно від того, наскільки добрі імітаційні моделі, якість значень отриманих результатів обмежується якістю вхідних параметрів. З іншого боку, незважаючи на численні припущення, які вимагає ТМО, результати, що вираховуються, часто підходять зовсім близько до тих, які будуть отримані за рахунок більш ретельного імітаційного моделювання. Крім того, СМО в буквальному сенсі дозволяють в кілька хвилин отримати чіткі рішення певних проблем, у той час як імітаційні експерименти можуть зайняти кілька днів, тижнів або довше на програмування та роботу з моделлю.

У зв'язку з цим, аналітику, інженеру-системотехніку необхідно оволодіти основами аналізу СМО.

Метою даного навчального посібника є надання студентам практичного керівництва з використання аналітичних моделей одноканальних та багатоканальних СМО з однорідними та неоднорідними потоками заявок з різними дисциплінами обслуговування та замкнених і розімкнутих ММО під час проведення практичних робіт з дисциплін «Обчислювальні системи», «Проектування інформаційно-керуючих комплексів», а також в курсовому проекті з проектування інтегрованої автоматизованої системи управління підприємством.

У посібнику введена авторська система позначень основних параметрів і характеристик систем і мереж масового обслуговування, яка, на думку автора, що базується на багаторічному досвіді роботи зі студентами, є інтуїтивно зрозумілою.

Для кращого засвоєння студентами матеріалу, що викладається, в посібнику дані зведені таблиці аналітичних моделей, готових до за-

стосування, наведено практикум рішення задач. Крім того, представлені узагальнюючі матеріали для модульного контролю знань, переліки питань та завдань для самоконтролю. У Додатку розглядаються деякі необхідні елементарні поняття з теорії ймовірності і статистики.

МОДЕЛІ СМО. ЗАГАЛЬНІ ПОЛОЖЕННЯ

1.1. Система обслуговування сервера-одинака

Найпростіша система масового обслуговування зображена на рисунку 1.1.

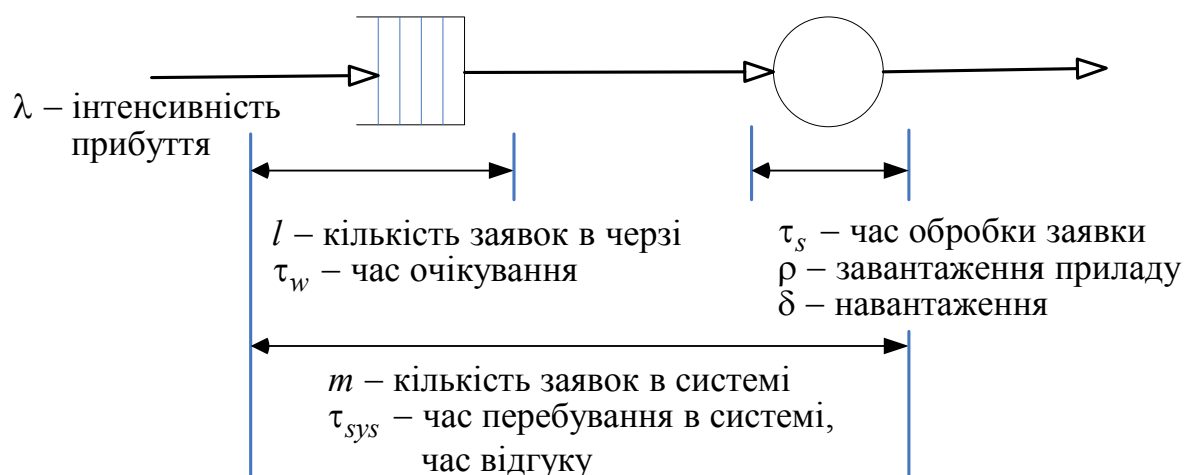


Рис. 1.1. Структура СМО і параметри для сервера-одинака

Центральним елементом системи є сервер, який надає деякі послуги за заявками, що надходять в систему. Якщо сервер знаходиться в режимі очікування і черга – пуста, заявка негайно подається на обробку в сервері. В іншому випадку, заявки, що прибувають, розташовуються в накопичувачі і чекають звільнення сервера. Коли сервер завершує обробку заявки, вона відбуває з системи. Якщо в цей час є

Прибуття
заявок
Накопичувач
(черга)
Дисципліна
буферизації

заявки в черзі, то одна з них поступає на обробку до сервера. Сервер в моделі може представляти будь-який пристрій, що виконує деякі функції або надає послуги за відповідними заявками, що надходять в систему. Наприклад: процесор обслуговує процеси; лінії забезпечують послуги з передачі пакетів або кадрів даних; пристрої вводу/виводу забезпечують зчитування або запис інформації тощо.

1.1.1. Параметри СМО

На рис. 1.1 ілюструються деякі важливі параметри, пов'язані з моделями СМО. Заявки прибувають в систему з деякою середньою швидкістю λ (заявок за секунду).

В будь-який момент часу, певна кількість заявок буде чекати в черзі (нуль або більше); середня кількість заявок, що очікують, l , а середній час очікування для кожної заявки τ_w . τ_w є усереднення для всіх вхідних заявок, в тому числі і для тих, які не чекали в черзі взагалі.

Сервер обробляє вхідні заявки з середнім часом обслуговування (обробки, Serve) τ_s . Це проміжок часу між надходженням заявки на обробку до сервера і часом, коли ця заявка залишає сервер (систему).

Завантаження ρ - це частка часу, коли сервер зайнятий, яка вимірюється по деякому інтервалу часу.

Нарешті, два параметра відносяться до системи в цілому. Середня кількість заявок, що перебувають в системі на обслуговуванні в сервері і в черзі в накопичувачі є m (резидентні заявки), а середній час, що заявка проводить в системі (SYStem) в черзі і на обробці - τ_{sys} .

Якщо припустити, що довжина черги нескінченна, то в системі не можна втратити заявок. Вони всі просто відкладаються в накопичувачі, поки не зможуть бути обслужені. В цих умовах інтенсивність вибуття заявок з системи λ_1 дорівнює швидкості прибуття λ . Оскільки швидкість прибуття λ - є швидкість трафіку, що проходить через цю систему, то його зростання призводить до заторів: черга стає більшою, час очікування зростає. При $\rho = 1$ сервер стає насиченим, працює 100% часу. Таким чином, теоретичний максимум вхідного потоку, який може бути оброблений системою:

$$\lambda_{\text{макс}} = 1/\tau_s$$

Проте, вже в цьому випадку система максимально насичена і її черга зростає необмежено при $\rho = 1$. Практичні міркування, такі як час відгуку, вимоги до розміру буфера, як правило, обмежують вхідний рівень заявок для одного сервера до 70-90% від теоретичного максимуму (для процесорів в інформаційних системах цей рівень складає 60 %, а для каналів передачі даних – 30-40 %).

Зараз, для продовження, ми повинні зробити деякі припущення для цієї моделі.

Популяція заявок на обслуговування. Як правило, ми припускаємо нескінченний потік заявок. Це означає, що інтенсивність потоку постійна і немає втрат заявок. Якщо популяція кінцева, то обробка заявок зазвичай пропорційно зменшує швидкість їх прибуття.

Розмір черги. Зазвичай, ми припускаємо, нескінченний розмір черги. Таким чином, черга може зростати необмежено. В кінцевій черзі можлива втрата заявок в системі. На практиці будь-яка черга кінцева, але у багатьох випадках це не суттєво для аналізу.

Дисципліни диспетчеризації. Коли сервер стає вільним, і, якщо є більш ніж одна заявка в черзі, має бути прийняте рішення про те, яку заявку спрямувати далі на обробку. Найпростіший підхід (дисципліна) «першим увійшов - першим обслуговується», або «First Input – First Output» (FIFO). Інша можливість «останній увійшов - перший вийшов» або «Last Input – First Output» (LIFO). На практиці використовуються і більш складні дисципліни обслуговування, що враховують характеристики заявок. Наприклад, в мережах з комутацією пакетів першими на обробку можуть вибиратися найкоротші пакети (для генерації найбільшої кількості вихідних пакетів) або першими вибираються самі довгі пакети (щоб звести до мінімуму час обробки по відношенню до часу передачі). На жаль, дисципліни, засновані на врахуванні часу обслуговування, дуже важко аналітично моделювати.

1.1.2. Основні позначення

У таблиці 1.1 наведені позначення, які використовуються в нав-

чальному посібнику.

Таблиця 1.1

Позначення для систем масового обслуговування

t_i	i -й момент часу
τ_f	інтервал часу f ($f \in \{in, s, out, ws, wi, w, pr, sys, net, neti, netw\}$)
τ_{in}	час між заявками у вхідному (<i>input</i>) потоці
τ_s	тривалість обробки (<i>service</i>) заявки
τ_{out}	час між заявками у вихідному (<i>output</i>) потоці
τ_{ws}	час очікування (<i>waiting</i>) початку обслуговування
τ_{wi}	час очікування в перерваному стані (<i>interrupt</i>)
τ_w	повний час очікування: $\tau_w = \tau_{ws} + \tau_{wi}$
τ_{pr}	час перебування заявки на обробці (<i>processing</i>): $\tau_{pr} = \tau_{wi} + \tau_s$
τ_{sys}	час перебування заявки в системі: $\tau_{sys} = \tau_w + \tau_s$
τ_d	середній час затримки (<i>delay</i>) заявок в мультисервері
$F(\tau)$	функція розподілу випадкової величини τ_f
$f(\tau)$	щільність розподілу випадкової величини τ_f : $f(\tau) = F'(\tau)$
$F^*(s)$	перетворення Лапласа щільності розподілу $f(\tau)$
$X^*(z)$	твірна функція розподілу $p_k = P(X = k)$
$f^{(n)}$	n -й початковий момент розподілу $F(\tau)$
$\hat{f}^{(n)}$	n -й центральний момент розподілу $F(\tau)$
$f = M[\tau_f]$	математичне очікування випадкової величини τ_f : $f = M[\tau_f] = f^{(1)}$
D_f	дисперсія випадкової величини τ_f : $D_f = \hat{f}^{(2)}$
σ_f	середнє квадратичне відхилення випадкової величини τ_f : $\sigma_f = \sqrt{D_f}$
v_f	коефіцієнт варіації випадкової величини τ_f : $v_f = \sigma_f / f$

v_{in}	коефіцієнт варіації інтервалів τ_{in} у вхідному потоці
v_s	коефіцієнт варіації тривалості обслуговування τ_s
v_{out}	коефіцієнт варіації інтервалів τ_{out} в вихідному потоці
λ	інтенсивність потоку заявок
$\lambda_{\hat{1}}$	інтенсивність потоку обслужених заявок
$\lambda_{\hat{a}}$	інтенсивність потоку втрачених заявок
Λ	сумарна інтенсивність об'єднаного потоку заявок
$\hat{\lambda}$	продуктивність замкнутої ММО (мережі масового обслуговування), пропускна здатність замкнутої ММО
$\pi_{\hat{a}}$	ймовірність втрати заявки
$\pi_{\hat{1}}$	ймовірність обслуговування заявки
l	середня довжина черги заявок в СМО
L	сумарна довжина черги заявок в ММО
m	середня кількість заявок в СМО
M	сумарне число заявок в ММО
H	кількість класів заявок
N	число вузлів в ММО
$\hat{N}$	кількість каналів в ММО з N вузлами
K	число обслуговуючих приладів в СМО
k	задана кількість заявок в СМО М/М/1 (таблиця 1.3)
Q	матриця пріоритетів: $Q = [q_{ij}(i, j = \overline{1, H})]$
q_{ij}	елемент матриці пріоритетів, що приймає значення: 0 – немає пріоритету, 1 - пріоритет відносний і 2 - пріоритет абсолютний
$r_g(i, h)$	коефіцієнти, що дозволяють виділити класи заявок з однаковим співвідношенням пріоритетів
μ	інтенсивність обслуговування: $\mu = 1/\tau_s$
θ	середня ресурсомісткість обслуговування заявки
δ	навантаження: $\delta = \lambda\tau_s$
Δ	сумарне навантаження
ρ	коефіцієнт завантаження (завантаження): $\rho = \min(\delta / K; 1)$
β	коефіцієнт простою: $\beta = 1 - \rho$
R	сумарне завантаження
α_j	коефіцієнт передачі j -го вузла ММО
p_{ij}	ймовірність передачі заявки з вузла i у вузол j ММО

q_{ij}	ймовірність переходу марківського процесу з дискретним часом зі стану i в стан j
g_{ij}	інтенсивність переходу марківського процесу з безперервним часом зі стану i в стан j
$\mathbf{D}=[d_{ij}(i, j = \overline{1, N})]$	матриця ймовірностей переходу марківського випадкового процесу з дискретним часом
$\mathbf{G}=[g_{ij}(i, j = \overline{1, N})]$	матриця інтенсивностей переходу марківського випадкового процесу з неперервним часом
$p_i(t)$	ймовірність того, що марківський випадковий процес в момент часу t перебуває в стані i
$p_i(0)$	початкова ймовірність - ймовірність того, що марківський випадковий процес в момент часу $t = 0$ знаходиться в стані i
$\mathbf{P}=(p_1, \dots, p_n)$	вектор стаціонарних ймовірностей станів
S	вартість
τ_{net}	середній час перебування заявки в мережі
τ_{neti}	середній час перебування заявки в вузлі i за час перебування в мережі
τ_{netw}	середній час очікування заявок в мережі
$m_x(y)$	у процентіль; значення y , нижче якого відбувається x відсотків значень
ПРИМІТКИ:	
Обробка	Service, s
Обслуговування, перебування	SYStem, sys Це ОЧІКУВАННЯ + ОБРОБКА
Процентиль розподілу	Процентиль розподілу - термін був вперше використаний Галтоном в 1885 р. Це таке число xp , що значення p -ї частини сукупності менше або рівні xp . Наприклад, 25-а процентиль змінної (також звана квантиль 0,25 або нижня квартиль) - це таке значення (xp), що 25% (p) значень змінної потрапляють нижче цього значення. Аналогічно обчислюється 75-а процентиль (також звана квантиль 0,75 або верхня квартиль) - таке значення, нижче якого потрапляють 75% значень змінної.

1.2. Система обслуговування мульти-сервера

На рис.1.2 показана узагальнення простої моделі для кількох серверів, що мають загальну чергу, а на рис. 1.3 – мультисерверна система з розподіленими чергами. З точки зору функціонування систем, перша модель відповідає багатопроцесорним обчислювальним системам, а друга – багатомашинним [2].

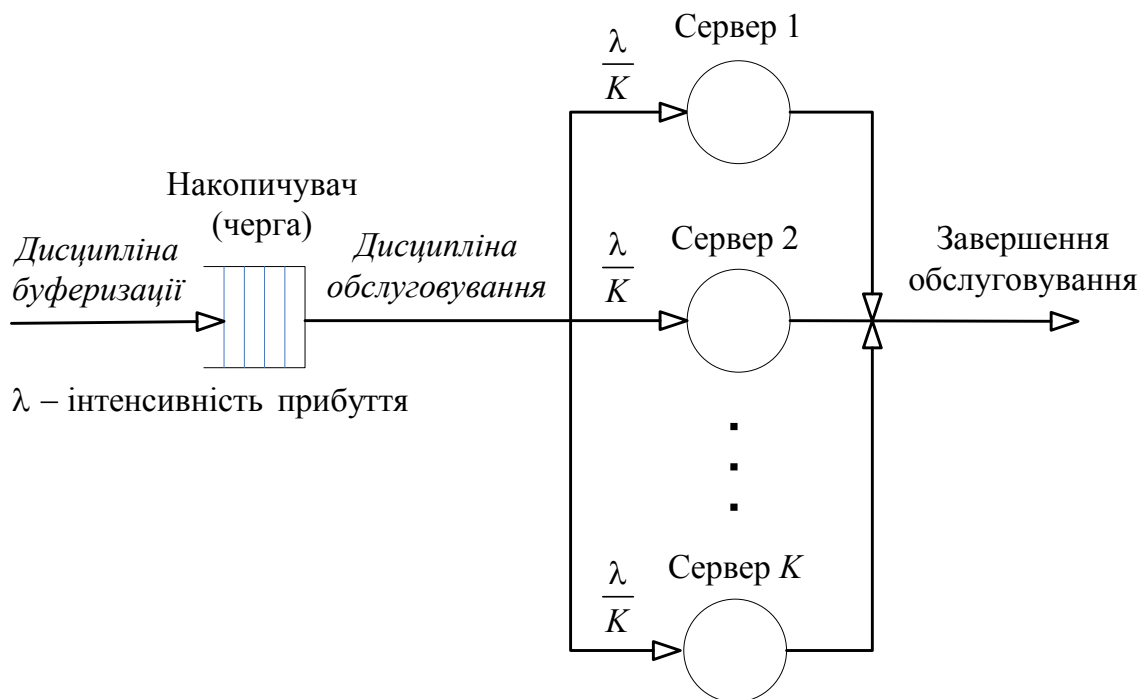


Рис. 1.2. Мультисерверна СМО з однією чергою

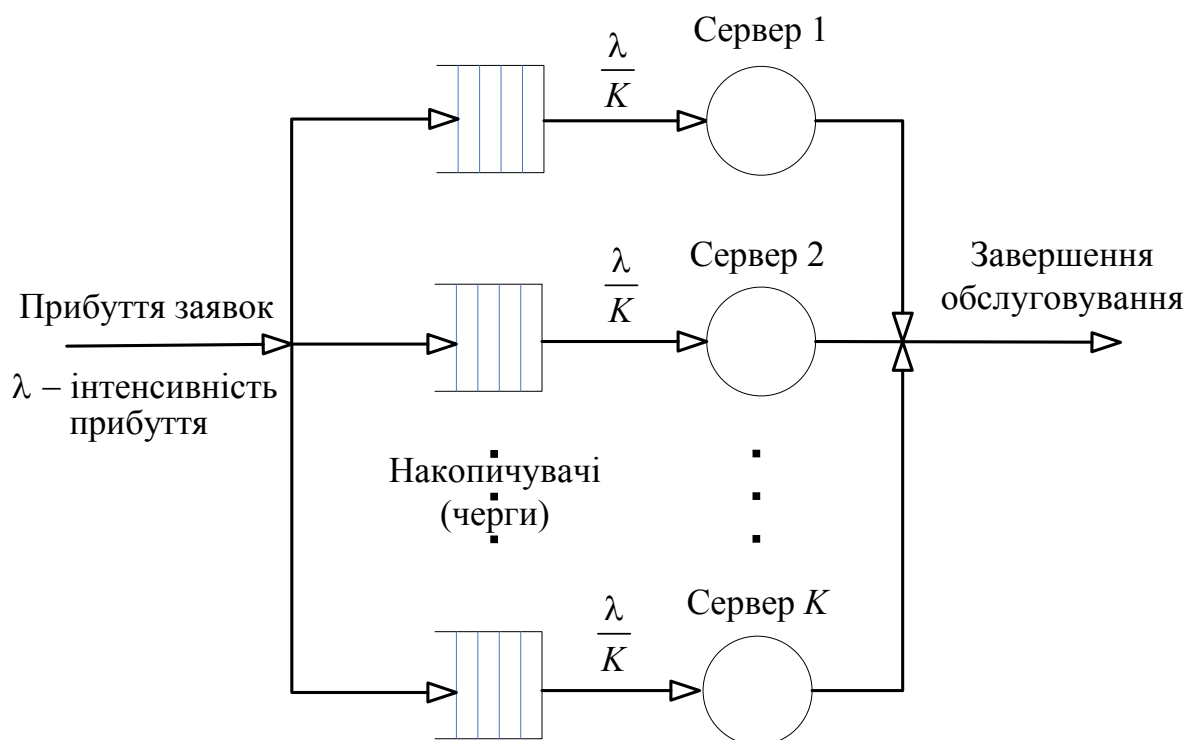


Рис. 1.3. Мультисерверна СМО з K чергами

Якщо в систему надходить заявка і принаймні один сервер вільний, то вона негайно відправляється на сервер. Передбачається, що всі сервери ідентичні, таким чином, якщо більш ніж один сервер доступний, то не має значення, який сервер обраний для дані заявки. Якщо всі сервери зайняті, то починає формуватися черга. Як тільки сервер стає вільним, заявка з черги, згідно з дисципліною обслуговування, поступає на обробку.

За винятком завантаження, всі параметри, показані на рисунку 1.1 переносяться на багатосерверний випадок з тією ж інтерпретацією. Якщо ми маємо K ідентичних серверів і ρ - завантаження кожного сервера, то ми можемо розглядати завантаження системи в цілому як $K\rho$. Цю величину часто називають інтенсивністю трафіка λ_{tr} . Останній термін часто порівнюється з інтенсивністю дорожнього руху. Таким чином, теоретичний максимум завантаження $(100\%)*K$, а теоретична максимальна вхідна інтенсивність становить:

$$\lambda_{\text{макс}} = \frac{K}{\tau_s}$$

Основні характеристики для багатосерверної СМО зазвичай вибирають так, як для СМО сервера-одинака. Тобто, ми припускаємо, нескінченний потік заявок і нескінченний розмір черг. З цієї однієї нескінченної черги заявки розподіляються між усіма серверами за дисципліною диспетчеризації FIFO, якщо не вказано інше. Для мультисерверного випадку, якщо всі сервери вважаються ідентичними, вибір конкретного сервера для заявки, що очікує, не впливає на час її обслуговування. Для порівняння, на рисунку 1.3 показана структура декількох серверами-одинаками. Як ми побачимо, ці очевидні незначні зміни в структурі, суттєво впливають на продуктивність системи.

1.3. Загальні положення СМО

1.3.1. Базові співвідношення

Для продовження розгляду аналітичних моделей, необхідно зробити деякі спрощення і припущення. Хоча вони дещо обмежують актуальність моделей, але на щастя, в більшості випадків, результат буде достатньо точним для планування і проектування. Є, однак, деякі співвідношення, які є істинними в загальному випадку (див. табл. 1.2).

Таблиця 1.2

Базові співвідношення СМО

Загальні формули	Один сервер	Багато серверів
$m = \lambda \tau_{sys}$ Формула Літла	$\rho = \lambda \tau_s$	$\rho = \frac{\lambda \tau_s}{K}$
$l = \lambda \tau_w$ Формула Літла	$m = l + \rho$	$\lambda_{tr} = \lambda \tau_s = \rho K$
$\tau_{sys} = \tau_w + \tau_s$		$m = l + K\rho$

1.3.2. Задачі аналізу СМО

Основна задача аналізу СМО наступна.

Задано вхідні дані:

- середня інтенсивність вхідних заявок (λ);
- середній час обробки (τ_s).

Забезпечити розрахунок вихідних характеристик:

- середня кількість заявок, що чекають в черзі (l);
- середній час очікування в черзі (τ_w);
- середня кількість резидентних заявок, що знаходяться в системі (m);
- середній час перебування заявки в системі (τ_{sys}).

Що конкретно ми хотіли б знати про ці результати? Звичайно, треба знати їх середні значення (l , τ_w , m , τ_{sys}). Крім того, було б корисно знати дещо про стандартне відхилення кожної з характеристик (σ_l , σ_{τ_w} , σ_m , $\sigma_{\tau_{sys}}$), а також інші дані. Наприклад, при проектуванні буфера для моста або мультиплексора було б корисно знати, який розмір буфера буде забезпечувати ймовірність його переповнення менше 0,001. Тобто, яке значення повинна мати M , щоб $\Pr[l < M] = 0,999$?

Для відповіді на ці питання необхідно знати закони розподілу ймовірностей випадкових значень інтенсивностей прибуття заявок і часу їх обробки. Крім того, навіть при тому, що ми будемо мати ці параметри, отримані формули є надзвичайно складними. Таким чином, щоб спростити цю проблему, ми повинні зробити деякі спрощуючі припущення.

Найбільш важливими з цих припущень є те, що швидкість прибуття підпорядковується розподіленню Пуасона, яке еквівалентно тому, що інтервали часу між двома заявками підпорядковуються експоненційному закону, заявки прибувають випадково і незалежно одна від одної. Ці припущення будемо робити завжди, без цього аналітичні моделі непрактичні. Вони дозволяють отримати багато корисних результатів, якщо ми маємо тільки середні значення та стандартні відхилення швидкості прибуття і часу обробки. Задачі можна зробити ще простіше і отримати більш докладні результати, якщо припустити, що і час обробки є експоненційним або постійним.

1.3.3. Опис СМО в нотації Кендала

Зручні позначення для ММО були запропоновані Д. Кендалом [3], які узагальнюють основні припущення в моделях. Вони мають наступний вигляд:

$$IN / S / K / L,$$

де **IN** відноситься до розподілення інтервалів часу між заявками τ_{in} , **S** відноситься до розподілення часу обробки τ_s , **K** вказує на кількість серверів, **L** – кількість місць в накопичувачі (відсутність **L** свідчить, що ємність його - необмежена).

Найбільш поширені закони розподілу позначаються так:

- **G** = загальне незалежне прибуття або час обробки;
- **M** = експоненційне (показове) розподілення;
- **D** = детерміноване прибуття або фіксований час обробки.

Таким чином, модель **M/M/1** вказує на модель з одним сервером, з Пуасонівським законом прибуття заявок і експоненційним часом їх обробки.

1.3.4. СМО сервера-одинака

В таблиці 1.3 а) наведені деякі рівняння для сервера-одинака, які відповідають моделі **M/G/1**, тобто, швидкість прибуття за Пуасоном, час обробки – загальне розподілення. Введемо коефіцієнт масштабування **A**, що спростить рівняння для отримання характеристик системи. Зверніть увагу, що ключовим фактором у масштабуванні є відношення стандартного відхилення часу обробки до середнього значення, і ніякої іншої інформації про час обробки не потрібно.

Цікавими є два наступних випадки. Коли стандартне відхилення дорівнює середньому, розподіл часу обробки є експоненційним, то отримуємо модель **M/M/1**. Це найпростіший випадок для розрахунку результатів. В таблиці 1.3 б) показана спрощена версія рівняння для стандартного відхилення m , τ_{sys} , а також деякі інші корисні параметри. Інший цікавий випадок, коли стандартне відхилення часу обробки

дорівнює нулю, тобто час обробки – постійний. Отримуємо модель M/D/1. Відповідні рівняння для цього випадку наведені в таблиці 1.3 с).

На рисунках 1.4 і 1.5 відображені залежність значень середнього розміру черги і часу перебування заявок в системі від завантаження при трьох значеннях σ_{τ_s} / τ_s [1].

Ця остання величина називається коефіцієнтом варіації і дає нормовану міру мінливості. Зверніть увагу, що найгірша продуктивність проявляється в експоненційний час обробки ($\sigma_{\tau_s} = \tau_s$) і краща за все при постійному часі обробки ($\sigma_{\tau_s} = 0$). Можна вважати, що, як правило, експоненційний час обробки буде відповідати найгіршому випадку. Аналіз, заснований на цьому припущенні, дає консервативний (песимістичний) результат. І це добре, тому що аналітичні моделі для M/M/1 достатньо прості.

Які значення може приймати σ_{τ_s} / τ_s ? Розглянемо чотири області [1].

Таблиця 1.3

Моделі масового обслуговування для сервера-одинака

a). Загальний час обробки (M/G/1)	b). Експоненційний час обробки (M/M/1)		c). Постійний час обробки (M/D/1)
$A = \frac{1}{2} * \left[1 + \left(\frac{\sigma_{\tau_s}}{\tau_s} \right)^2 \right]$	$m = \frac{\rho}{1-\rho}$	$l = \frac{\rho^2}{1-\rho}$	$m = \frac{\rho^2}{2 * (1-\rho)} + \rho$
$m = \rho + \frac{\rho^2 + A}{1-\rho}$	$\tau_{sys} = \frac{\tau_s}{1-\rho}$	$\tau_w = \frac{\rho * \tau_s}{1-\rho}$	$l = \frac{\rho^2}{2 * (1-\rho)}$
$l = \frac{\rho^2 + A}{1-\rho}$	$\sigma_m = \frac{\sqrt{\rho}}{1-\rho}$	$\sigma_{\tau_{sys}} = \frac{\tau_s}{1-\rho}$	$\tau_{sys} = \frac{\tau_s * (2-\rho)}{2 * (1-\rho)}$
$\tau_{sys} = \tau_s + \frac{\rho * \tau_s * A}{1-\rho}$	$\Pr[m_k = k] = (1-\rho) * \rho^k$	$\tau_w = \frac{\rho * \tau_s}{2 * (1-\rho)}$	
$\tau_w = \frac{\rho * \tau_s * A}{1-\rho}$	$\Pr[m_k \leq k] = \sum_{i=0}^k (1-\rho) * \rho^i$	$\sigma_m = \frac{1}{1-\rho} * \sqrt{\rho - \frac{3 * \rho^2}{2} + \frac{5 * \rho^3}{6} - \frac{\rho^4}{12}}$	
Припущення в моделях	$\Pr[\tau_k \leq t] = 1 - e^{-(1-\rho) * t / \tau_s}$	$\sigma_{\tau_{sys}} = \frac{\tau_s}{1-\rho} * \sqrt{\rho - \frac{\rho^2}{3} - \frac{\rho^2}{12}}$	
	$m_{\tau_{sys}}(y) = \tau_{sys} * \ln\left(\frac{100}{100-y}\right)$		
	$m_{\tau_w}(y) = \frac{\tau_w}{\rho} * \ln\left(\frac{100 * \rho}{100-y}\right)$		
	1. Пуасонівська швидкість прибуття заявок. 2. Дисципліна диспетчеризації не віддає переваги заявкам за часом їх обробки. 3. Формули для стандартного відхилення відповідають дисципліні обслуговування FIFO (першим увійшов, першим оброблений). 4. Заявки з черги не видаляються (накопичувач необмежений).		

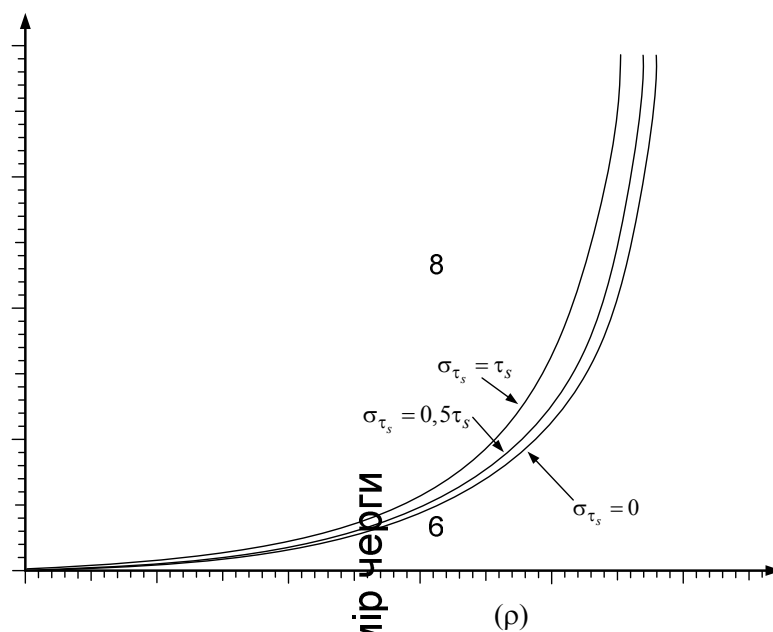


Рис. 1.4. Середній розмір черги для сервера-одинака

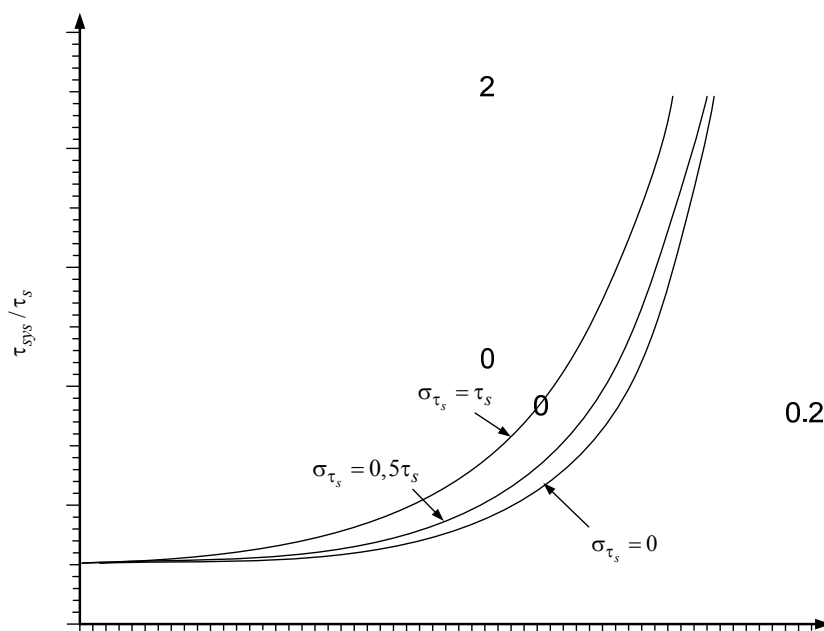


Рис. 1.5. Середній час обслуговування для сервера-одинака

- Область **Zero**. Це рідкісний випадок постійного часу обробки заявок. *Наприклад, якщо всі передані повідомлення мають однакову довжину, вони підходять під цю категорію.*
- **Співвідношення менше 1**. Це співвідношення краще, ніж експоненційний випадок, тому використовуючи модель M/M/1 ми будемо мати розміри черг і часу обробки трохи більші, ніж вони повинні бути. *Прикладом такої категорії може бути введення даних додатків для заповнення конкретних форм.*
- **Співвідношення близько до 1**. Це є звичайне явище, що відповідає експоненційному часу обробки, тобто, по суті випадковому часу обробки. Розглянемо повідомлення, що надходить в систему і має об'єм, що дорівнює місткості екрану терміналу комп'ютера: повний екран може бути 1920 символів і всі повідомлення можуть бути різного розміру в цьому діапазоні. *Під цю категорію задач підходять бронювання авіаквитків, пошук файлів по запиту, процеси в локальних мережах і в мережах з комутацією пакетів.*
- **Співвідношення більше 1**. Якщо ви помітили це, то повинні використовувати модель M/G/1 і не покладатися на моделі M/M/1. В загальному випадку це бімодальне розподілення з широким розмахом між піками. *Прикладом може служити система, яка отримує багато коротких повідомлень, багато довгих повідомлення і лише деякі інші між ними.*

Те ж саме відноситься і до розгляду швидкості прибуття заявок. Для швидкості прибуття за законом Пуассона, час між двома заявками - експоненційний і співвідношення стандартної девіації до середнього значення становить 1. Якщо спостерігається співвідношення набагато менше, ніж 1, то прибулі заявки, як правило, рівномірно розподілені (не дуже мінливі) і припущення Пуассона буде переоцінювати розміри черг і затримок. З іншого боку, якщо коефіцієнт більше ніж 1, то прибулі заявки, як правило, утворюють кластери і все більш гострою стає проблема заторів.

РОЗДІЛ II

СМО З ОДНОРІДНИМ ПОТОКОМ ЗАЯВОК

2.1. ОСМО - одноканальні СМО

2.1.1. СМО сервера-одинака з однорідним потоком заявок

Розглянемо одноканальну СМО з однорідним потоком заявок (ОПЗ) при наступних припущеннях (рис.2.1):

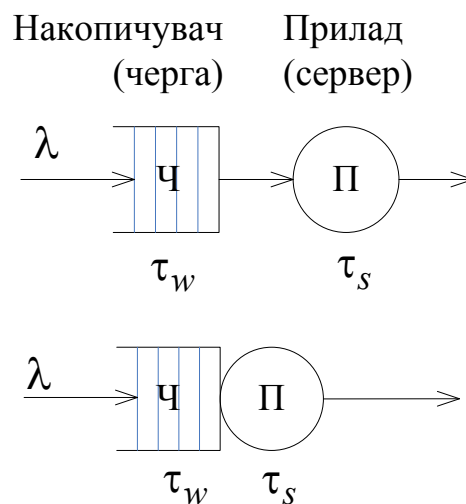


Рис. 2.1. Варіанти зображення одноканальних СМО з ОПЗ

- 1) СМО містить один обслуговуючий прилад (П), в якому в кожний момент часу може обслуговуватися тільки одна заявка;
- 2) перед приладом є накопичувач Н необмеженої ємкості, що означає відсутність відмов заявкам, що надходять, при їх поста-

новці в чергу Ч, тобто будь-яка заявка, що надходить, завжди знайде в накопичувачі місце для очікування не залежно від того, скільки заявок вже знаходиться в черзі;

- 3) заявки надходять в СМО з інтенсивністю λ ;
- 4) середня тривалість обслуговування однієї заявки в приладі дорівнює τ_s , причому тривалості обслуговування різних заявок не залежать один від другого;
- 5) обслуговуючий прилад не простоє, якщо в системі (накопичувачі) є хоча б одна заявка, причому після завершення обслуговування чергової заявки миттєво з накопичувача вибирається наступна заявка;
- 6) заявки з накопичувача вибираються відповідно до беспріоритетної дисципліни обслуговування в порядку надходження за правилом «першим прийшов - першим обслужений» (FIFO - First In First Out).
- 7) у системі існує стаціонарний режим, що припускає відсутність перевантажень, тобто навантаження і завантаження системи менше 1:

$$\delta = \rho = \lambda \tau_s < 1.$$

В якості розрахункової характеристики обслуговування заявок в СМО будемо використовувати середній час очікування заявок τ_w . Значення інших характеристик функціонування СМО легко можуть бути розраховані з використанням наведених в розділі 1 фундаментальних співвідношень.

Розглянемо чотири моделі СМО з однорідним потоком заявок: експоненційну СМО М/М/1 і три неекспоненційних СМО типу М/Г/1, Г/М/1, Г/Г/1.

2.1.2. Характеристики ОСМО М/М/1

Нехай заявки, що надходять в одноканальну СМО, утворюють найпростіший потік з інтенсивністю λ , а тривалість обслуговування

заявок розподілена за експоненційним законом із середнім значенням τ_s , причому $\rho = \lambda \cdot \tau_s$, тобто система працює в сталому режимі. Така СМО з однорідним потоком заявок називається експоненційною. З використанням методу середніх значень або Марківської моделі [4] можна отримати наступні вирази для розрахунку середніх значень:

- час очікування заявок

$$\tau_w = \frac{\rho \tau_s}{1 - \rho} \quad (2.1)$$

- час перебування заявок в системі

$$\tau_{sys} = \tau_w + \tau_s = \frac{\tau_s}{1 - \rho} \quad (2.2)$$

- довжина черги заявок

$$l = \lambda \cdot \tau_w = \frac{\rho^2}{1 - \rho} \quad (2.3)$$

- число заявок в системі (в черзі і на обслуговуванні)

$$m = \lambda \cdot \tau_{sys} = \frac{\rho}{1 - \rho} \quad (2.4)$$

З останніх виразів випливає, що середнє число заявок в системі

$$m = l + \rho,$$

де другий доданок ρ визначає середнє число заявок, що перебувають на обслуговуванні в приладі. Крім того, порівнюючи вирази (2.1) і (2.2) отримаємо, що

$$\tau_{sys} = \rho \cdot \tau_w.$$

2.1.3. Характеристики ОСМО M/G/1

Нехай заявки, що надходять в одноканальну СМО, утворюють найпростіший потік з інтенсивністю λ , а тривалість обслуговування заявок розподілена за довільним законом $B(\tau_s)$ з середнім значенням τ_s і коефіцієнтом варіації v_{τ_s} .

З використанням методу середніх значень можна показати, що середній час очікування заявок визначається за формулою Поллачека-Хінчина[4]:

$$\tau_w = \frac{\lambda \cdot \tau_s^2 (1 + v_{\tau_s}^2)}{2(1 - \rho)}, \quad (2.5)$$

де $\rho = \lambda \cdot \tau_s < 1$ - завантаження системи.

Середній час перебування заявок в системі:

$$\tau_{sys} = \tau_w + \tau_s = \frac{\lambda \cdot \tau_s^2 (1 + v_{\tau_s}^2)}{2 \cdot (1 - \rho)} + \tau_s, \quad (2.6)$$

Слід відзначити цікаву особливість представлених виразів, а саме: середні значення характеристик обслуговування заявок залежать тільки від двох перших моментів тривалості обслуговування заявок і не залежать від моментів більш високого порядку. Іншими словами, для того щоб розрахувати середні характеристики обслуговування, не обов'язково знати закон розподілу тривалості обслуговування заявок - достатньо знати тільки два перші моменти розподілу. Можна показати, що для розрахунку других моментів характеристик обслуговування заявок досить задати три перші моменти тривалості обслуговування і т.і. Резюмуючи, можна стверджувати, що для **СМО з найпростішим потоком заявок** для розрахунку перших k моментів характеристик обслуговування необхідно задати $(k + 1)$ моментів тривалості обслуговування заявок.

2.1.4. Характеристики ОСМО G/M/1

Нехай в одноканальну СМО з інтенсивністю λ надходить випадковий потік заявок довільного виду, що задається функцією розподілу інтервалів між заявками $A(\tau_{in})$, а тривалість обслуговування заявок розподілена за експоненційним законом $B(\tau)$ з середнім значенням τ_s (інтенсивністю $\mu = 1/\tau_s$).

СМО G/M/1 є симетричною по відношенню до СМО M/G/1, розглянутої в попередньому пункті. Однак отримання кінцевого результату у вигляді аналітичного виразу для розрахунку середнього часу очікування, за аналогією з попередньою моделлю, в загальному випадку, виявляється неможливим. Це обумовлено тим, що середній час очікування, втім, як і інші числові моменти, залежить не тільки від двох перших моментів інтервалів між заявками, що надходять, але і від моментів більш високого порядку, тобто від закону розподілу інтервалів між заявками. Середній час очікування заявок в черзі може бути розрахованим наступним чином [3]:

$$\tau_w = \frac{\zeta \cdot \tau_s}{1 - \zeta}, \quad (2.6)$$

де ζ - єдиний в області $0 < \zeta < 1$ корінь рівняння

$$\zeta = A^*(\mu - \mu \cdot \zeta). \quad (2.7)$$

Тут $A^*(s)$ - перетворення Лапласа щільності розподілу $a(\tau)$ інтервалів між заявками, що надходять в систему:

$$A^*(s) = \int_0^{\infty} e^{-s\tau} a(\tau) d\tau \quad (s \geq 0).$$

Основна складність при дослідженні СМО G/M/1 полягає в тому, що рівняння (2.7) для знаходження ζ , в загальному випадку, є

трансцендентним, і неможливо отримати вираз для ζ в явному вигляді. Проте в кожному конкретному випадку корінь рівняння (2.7) може бути знайдений з використанням чисельних методів.

Як сказано вище, середні значення характеристик обслуговування заявок залежать не тільки від двох перших моментів інтервалів між заявками, що надходять, але і від моментів більш високого порядку, причому ступінь впливу відповідних моментів зменшується з збільшенням порядку моментів. Іншими словами, вплив моментів 4-го порядку менш істотно, ніж моментів 3-го порядку і т.і.

Приклад 2.1. Проілюструємо застосування описаного методу розрахунку до розглянутої вище СМО М/М/1 з найпростішим потоком заявок.

У найпростішому потоці інтервали часу між послідовними заявками розподілені за експоненційним законом, перетворення Лапласа якого має вигляд:

$$A(s) = \frac{\lambda}{\lambda + s}$$

Тоді рівняння (2.7) прийме вигляд:

$$\zeta = \frac{\lambda}{\lambda + \mu - \mu\zeta}$$

Після деяких перетворень отримаємо квадратне рівняння:

$$\mu\zeta^2 - (\lambda + \mu)\zeta + \lambda = 0$$

Розділивши ліву і праву частину цього рівняння на μ , отримаємо:

$$\zeta^2 + (1 + \rho)\zeta + \rho = 0$$

З двох коренів $\zeta_1 = 1$ і $\zeta_2 = \rho$ останнього рівняння умові $0 < \zeta < 1$ задовольняє тільки другий корінь. Підставляючи його у вираз (2.6) остаточно отримаємо вираз для середнього часу очікування, що збігається з відомим для СМО М/М/1 виразом (2.1).

2.1.5. Характеристики ОСМО G/G/1

Найбільш загальним випадком одноканальних систем масового обслуговування є СМО типу G/G/1, в яку надходить довільний потік заявок загального виду з функцією розподілу інтервалів між заявками $A(\tau_{in})$. Тривалість обслуговування заявок в приладі розподілена за довільним законом $B(\tau)$.

Розрахунок таких систем вимагає завдання конкретних законів розподілів, що не дозволяє отримати аналітичне рішення в загальному вигляді. Аналітичне рішення можливе тільки для деяких приватних розподілів, пов'язаних, наприклад, з експоненційним розподілом. Для більшості законів розподілів інтервалів між заявками, що надходять в систему, і тривалостей їх обробки в приладі неможливо отримати точне рішення в аналітичній формі. У той же час, на практиці при дослідженні реальних систем рідко бувають відомі закони розподілів зазначених величин. Зазвичай при описі процесів надходження заявок в систему і їх обслуговування в приладі обмежуються кількома моментами відповідних розподілів, найчастіше - двома першими моментами, які задаються в вигляді математичного очікування і середньоквадратичного відхилення або коефіцієнта варіації шуканої випадкової величини. Однак при цьому виявляється неможливим отримання точного результату. Це обумовлено тим, що в разі довільного (відмінного від найпростішого) потоку заявок, що надходять в систему, характеристики функціонування СМО, зокрема середній час очікування, залежать не тільки від двох перших моментів, але і від моментів вищого порядку - третього, четвертого і т.і. Причому ця залежність тим менше, чим вище порядок числового моменту. Таким чином, всі результати, отримані в аналітичній формі при завданні інтервалів між заявками, що поступають в систему, і тривалостей їх обслуговування в приладі двома першими моментами - середніми значеннями $\tau_{in} = 1/\lambda$ і $\tau_s = 1/\mu$ і коефіцієнтами варіації $v_{\tau_{in}}$ і v_{τ_s} , є наближені залежності.

Як показав аналіз численних опублікованих результатів, одним з найбільш вдалих наближень для розрахунку середнього часу очікування в СМО G/G/1 є наступна формула [5]:

$$\tau_w = \frac{\rho \tau_s (v_{\tau_{in}}^2 + v_{\tau_s}^2)}{2(1-\rho)} \cdot f(v_{\tau_{in}}), \quad (2.8)$$

де $\rho = \lambda \cdot \tau_s < 1$ - завантаження системи; $\lambda, v_{\tau_{in}}$ - інтенсивність потоку заявок і коефіцієнт варіації інтервалів між заявками, що потсупають в систему; τ_s, v_{τ_s} - середнє значення і коефіцієнт варіації тривалості обслуговування заявок; $f(v_{\tau_{in}})$ - коригувальна функція, що розраховується в залежності від значення коефіцієнта варіації $v_{\tau_{in}}$:

$$f(v_{\tau_{in}}) = \begin{cases} \exp \left[-\frac{2(1-\rho)(1-v_{\tau_{in}}^2)^2}{3\rho(v_{\tau_{in}}^2 + v_{\tau_s}^2)} \right], & v_{\tau_{in}} < 1 \\ \exp \left[-(1-\rho) \frac{v_{\tau_{in}}^2 - 1}{(v_{\tau_{in}}^2 + 4v_{\tau_s}^2)} \right], & v_{\tau_{in}} \geq 1 \end{cases}.$$

При вирішенні багатьох практичних задач (в мережах СМО) вихідний потік заявок з однієї СМО є вхідним потоком в іншу СМО. В цьому випадку для розрахунку характеристик функціонування другої СМО необхідно знати характер вхідного потоку, що найбільш повно описується законом розподілу інтервалів між послідовними заявками. У той же час, для проведення оціночних розрахунків в багатьох випадках достатньо знання перших двох моментів цих інтервалів: математичного очікування і коефіцієнта варіації.

Очевидно, що в СМО з *накопичувачем необмеженої ємності, яка працює без перевантажень*, інтенсивність потоку заявок, що виходить, дорівнює інтенсивності вхідного потоку, тобто математичні очікування інтервалів між послідовними заявками на виході і вході СМО збігаються.

Можна показати, що для експоненційної СМО М/М/1 коефіцієнт варіації вихідного потоку дорівнює одиниці.

У загальному випадку для СМО G/G/1 коефіцієнт варіації вихідного потоку заявок може бути розрахований за такою наближеною формулою [5]:

$$v_{\lambda_{out}}^2 = v_{\tau_{in}}^2 + 2\rho v_{\tau_s}^2 - 2\rho(1-\rho)\frac{\tau_w}{\tau_s}. \quad (2.9)$$

2.1.6. Аналіз властивостей ОСМО з ОПЗ

Аналіз властивостей одноканальної СМО з однорідним потоком заявок будемо проводити з використанням представлених вище математичних моделей у вигляді формул (2.1 - 2.4), що визначають залежності характеристик обслуговування заявок від параметрів надходження та обробки заявок для сталого (стаціонарного) режиму роботи системи.

1. Середній час очікування заявок в черзі є мінімальним при постійній (детермінованій) тривалості обслуговування заявок, коли коефіцієнт варіації тривалості обслуговування $v_{\tau_s} = 0$, і збільшується із зростанням коефіцієнта варіації (дисперсії) тривалості обробки. Зауважимо, що залежність середнього часу очікування від коефіцієнта варіації v_{τ_s} носить нелінійний характер. Так, наприклад, при експоненційно розподіленій тривалості обслуговування, коли $v_{\tau_s} = 1$, середній час очікування заявок збільшується в 2 рази, а при $v_{\tau_s} = 2$ - в 5 разів, порівняно з детермінованим обслуговуванням.

2. Середній час очікування заявок істотно залежить від навантаження δ (завантаження ρ) системи (рис.2.2).

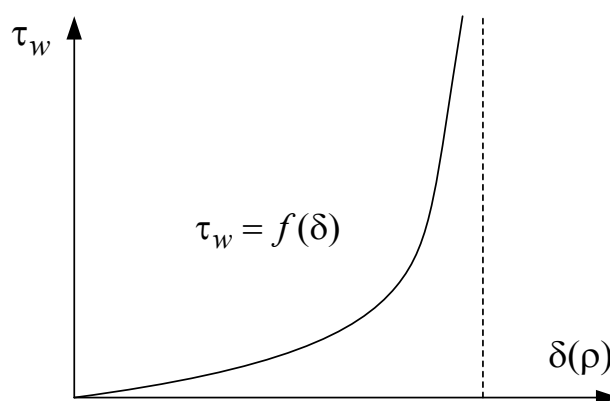


Рис. 2.2. Залежність середнього часу очікування від навантаження

При $\delta \geq 1$ ($\rho \rightarrow 1$) час очікування заявок зростає необмежено: $\tau_w \rightarrow \infty$, тобто заявки можуть очікувати обслуговування як завгодно довго. Відзначимо, що збільшення навантаження може бути обумовлено двома факторами: збільшенням інтенсивності надходження заявок в систему або збільшенням тривалості обслуговування заявок (наприклад, в результаті зменшення швидкості роботи обслуговуючого приладу).

3. Можна показати, що для безпріоритетних дисциплін обслуговування в зворотному порядку та обслуговування у випадковому порядку середні часи очікування заявок будуть такими ж, як і при обслуговуванні в порядку надходження, але дисперсії часу очікування будуть більше. Це обумовлено, зокрема для дисципліни обслуговування в зворотному порядку, тим, що частина заявок, що надійшли останніми, будуть очікувати незначний час, в той же час як заявки, що потрапили в початок черги, можуть очікувати обслуговування достатньо довго, тобто збільшується розкид значень часу очікування щодо середнього значення [6].

2.2. Багатоканальні СМО (БСМО)

Розглянемо багатоканальну СМО з K ідентичними обслуговуючими приладами і накопичувачем необмеженої ємності, в яку надходять заявки, що утворюють найпростіший потік з інтенсивністю λ (рис.2.3).

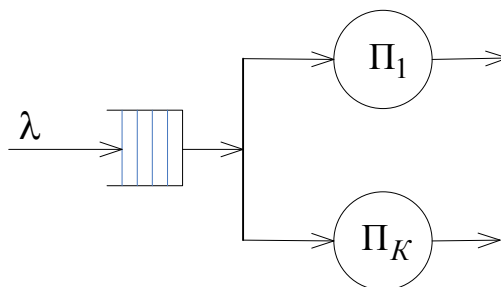


Рис. 2.3. Багатоканальна СМО

Тривалість обробки заявок розподілена за експоненційним зако-

ном із середнім значенням τ_s . Вибір заявок з черги на обслуговування здійснюється відповідно до безприоритетних дисциплін обслуговування в порядку надходження за правилом «першим прийшов - першим обслужений» (FIFO - First In First Out).

2.2.1. Характеристики БСМО М/М/К

В якості основної характеристики функціонування СМО, будемо використовувати середній час очікування заявок τ_w .

Точний метод розрахунку характеристик обслуговування заявок в багатоканальній СМО розроблений при наступних припущеннях:

- потік заявок - найпростіший;
- тривалість обробки заявок розподілена за експоненційним законом із середнім значенням τ_s ;
- всі K приладів - ідентичні, і будь-яка заявка може бути оброблена будь-яким приладом;
- ємність накопичувача - не обмежена;
- в системі відсутні перевантаження, тобто завантаження системи менше 1:

$$\rho = \frac{\lambda \tau_s}{K} < 1$$

При цих припущеннях середній час очікування заявок визначається наступним чином:

$$\tau_w = \frac{P_K \tau_s}{K(1 - \rho)}, \quad (2.10)$$

де P_K - ймовірність того, що всі K приладів зайняті обробкою заявок.

Ймовірність P_K визначається як:

$$P_K = \frac{(K\rho)^K}{K!(1 - \rho)} \cdot P_0,$$

де P_0 - ймовірність простою багатоканальної СМО, тобто є ймовірність того, що в системі немає заявок:

$$P_0 = \left[\frac{(K\rho)^K}{K!(1-\rho)} + \sum_{i=0}^{K-1} \frac{(K\rho)^i}{i!} \right]^{-1}.$$

2.2.2. Основні формули

У таблиці 2.1 наведені формули для деяких ключових параметрів мультисерверних конфігурацій. Зверніть увагу на обмеженість припущень. Корисні статистичні дані для цієї моделі були отримані тільки для випадку М/М/К, де експоненційний час обробки однаковий для K -серверів.

Таблиця 2.1

Формули для мультисервера

Припущення	1. Пуасонівська швидкість прибуття. 2. Експоненційний час обробки 3. Всі сервери в рівній мірі завантажені 4. Всі сервери мають однаковий середній час обробки 5. Дисципліна обслуговування заявок - FIFO (першим увійшов, першим оброблений). 6. Заявки з черги не видаляються (накопичувач необмежений).
Функція співвідношення Пуасона	$Z = \frac{\sum_{i=0}^{K-1} \frac{(K * \rho)^i}{i!}}{\sum_{i=0}^K \frac{(K * \rho)^i}{i!}}$
Ймовірність того, що всі сервери зайняті (С-функція Ерланга)	$C = \frac{1 - Z}{1 - \rho * Z}$

Середнє число заявок, що перебувають в системі (які очікують і обслуговуються)	$m = C * \frac{\rho}{1 - \rho} + K * \rho$
Середня кількість заявок в черзі, які очікують обробки	$l = C * \frac{\rho}{1 - \rho}$
Середній час перебування заявки в системі (обслуговування)	$\tau_{sys} = \frac{C}{K} * \frac{\tau_s}{1 - \rho} + \tau_s$
Середній час очікування (в тому числі заявки, які повинні чекати, і заявки, з часом очікування = 0)	$\tau_w = \frac{Z}{K} * \frac{\tau_s}{1 - \rho}$
Стандартне відхилення τ_{sys}	$\sigma_{\tau_{sys}} = \frac{\tau_s}{K * (1 - \rho)} * \sqrt{C * (2 - C) + K^2 * (1 - \rho)^2}$
Стандартне відхилення l	$\sigma_l = \frac{1}{1 - \rho} * \sqrt{C * \rho * (1 + \rho - C * \rho)}$
Ймовірність того, що середній час очікування заявок в черзі не буде перевищувати заданого значення	$\Pr[\tau_w \geq t] = C * e^{-K(1-\rho)*t/\tau_s}$
Вирахування процентілей для τ_w (див. розділ VII)	$m_{\tau_w}(y) = \frac{\tau_s}{K * (1 - \rho)} * \ln\left(\frac{100 * C}{100 - y}\right)$
Середній час затримки (delay) заявок в мультисервері	$\tau_d = \frac{\tau_s}{K * (1 - \rho)}$

2.2.3. Аналіз властивостей БСМО

Аналіз властивостей багатоканальної СМО з однорідним потоком заявок і накопичувачем необмеженої ємності може бути виконаний з використанням представлених вище математичних моделей, які ви-

значають залежність характеристик обслуговування заявок від параметрів надходження і обробки заявок для сталого (стаціонарного) режиму роботи системи.

1. На рис.2.4 показаний характер залежності середнього часу очікування τ_w і середнього часу перебування заявок в системі τ_{sys} від числа обслуговуючих приладів K .

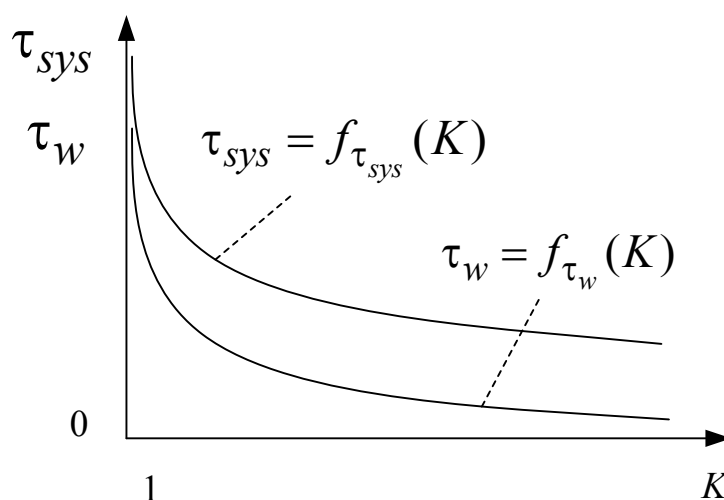


Рис. 2.4 Залежності часу очікування і перебування заявок від числа приладів

Очевидно, що зі збільшенням числа обслуговуючих приборів часи очікування і перебування заявок зменшуються, при цьому в межі при $K \rightarrow \infty$ час очікування прагне до нуля, а час перебування досягає свого найменшого значення, рівного тривалості обробки заявок.

2. На рис.2.5 показані аналогічні залежності, але за умови, що при

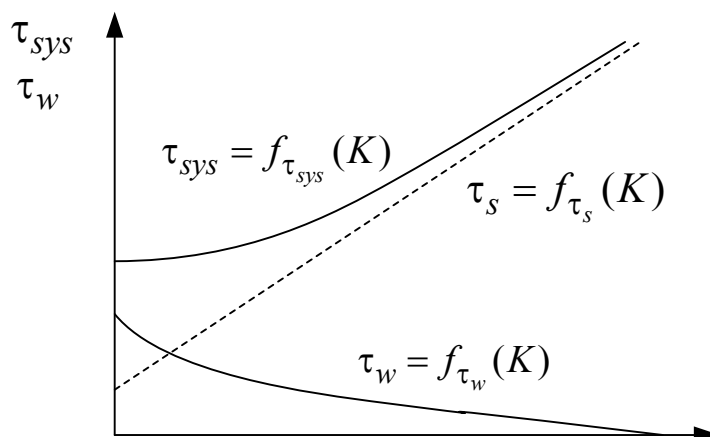


Рис. 2.5. Залежність часу перебування заявок від числа приладів при $V_{\Sigma} = const$

збільшенні числа обслуговуючих приладів K їх сумарна продуктивність (швидкість роботи) залишається постійною, тобто

$$V_{\Sigma} = KV_K = const ,$$

де V_K - продуктивність одного приладу при наявності в системі K обслуговуючих приладів.

З представлених графіків видно, що середній час очікування заявок τ_w , як і в попередньому випадку, зменшується зі збільшенням числа приладів, однак час перебування заявок в системі τ_{sys} збільшується. Останнє пояснюється тим, що зі збільшенням числа приладів K продуктивність кожного з них для збереження сумарної продуктивності системи зменшується пропорційно K і, отже, лінійно збільшується тривалість обслуговування заявки в приладі. При цьому швидкість збільшення тривалості обслуговування більше швидкості зменшення часу очікування, що в сумі призводить до збільшення часу перебування заявок в системі. У межі при $K \rightarrow \infty$ час перебування заявок асимптотично прагне до тривалості обслуговування заявок.

Таким чином, при проектуванні систем обслуговування слід мати на увазі, що з точки зору затримок (часу перебування заявок) більш ефективною є одноканальна система, ніж багатоканальна, при рівності сумарної продуктивності. Основною перевагою багатоканальної системи є більш висока надійність, що проявляється

в тому, що при виході з ладу одного або навіть декількох обслуговуючих приладів система продовжує функціонувати, хоча і з меншою ефективністю, що полягає в збільшенні часу перебування заявок в системі.

3. Можна показати, що середній час очікування заявок, як і для одноканальних систем, істотно залежить від навантаження ρ (завантаження ρ) системи. При $\delta \geq K$ ($\rho \rightarrow 1$) час очікування заявок зростає необмежено: $\tau_w \rightarrow \infty$, тобто заявки можуть очікувати обслуговування настільки завгодно довго.

2.3. Самоконтроль. Перелік питань і завдань

1. Як залежить середній час очікування заявок в СМО від коефіцієнта варіації тривалості обслуговування?

2. У скільки разів зміниться середній час очікування заявок при переході від постійної тривалості обслуговування до експоненційно розподіленої?

3. У скільки разів зміниться середній час очікування при переході від експоненційного розподілу тривалості обслуговування до гіперекспоненційного розподілу з коефіцієнтом варіації, рівним 2?

4. Який розподіл тривалості обслуговування заявок в СМО є кращим для зменшення середнього часу очікування заявок?

5. Чи зміниться різниця між середнім часом перебування і середнім часом очікування заявок в СМО при зміні: а) швидкості роботи (швидкодії) приладу; б) інтенсивності потоку заявок; в) кількості приладів?

6. Чи зміниться різниця між середнім числом заявок в системі і середньою довжиною черги при зміні: а) швидкості роботи (швидкодії) приладу; б) інтенсивності потоку заявок; в) кількості приладів?

7. Заявки надходять в одноканальну СМО з інтервалом 2,5 хвилин, тривалість обслуговування заявок в приладі 45 секунд. Визначити завантаження і коефіцієнт простою системи.

8. Інтенсивність надходження заявок в трьохканальну СМО - 21 заявка в секунду, інтенсивність обслуговування - 10 заявок в секунду. Визначити: а) ймовірність того, що обслуговуючий прилад працює;

б) ймовірність того, що обслуговуючий прилад простоє; в) середнє число заявок, що перебувають на обслуговуванні; г) середнє число працюючих приладів?

9. Інтенсивність надходження заявок в чотирьохканальний СМО дорівнює інтенсивності обслуговування заявок одним приладом. Визначити: а) ймовірність того, що система простоє, б) середнє число простоючих приладів; в) на яку величину середнє число заявок в системі відрізняється від середньої довжини черги.

10. Тривалість обслуговування заявок в одному приладі чотирьохканальної СМО дорівнює 4 хвилини. Визначити граничну інтенсивність надходження заявок в систему, при якій в системі існує стаціонарний режим?

11. Інтенсивність надходження заявок в СМО - 15 заявок на секунду, тривалість обслуговування однієї заявки - 5 секунд. Визначити число обслуговуючих приладів, при якому в системі існує стаціонарний режим?

12. Заявки надходять в одноканальну СМО з інтервалом 0,5 секунд, інтенсивність обслуговування - 2,5 заявки в секунду, середній час перебування заявок в системі - 2 секунди. Визначити середню довжину черги заявок.

13. Інтенсивності надходження і обслуговування заявок в СМО відповідно дорівнюють 4 і 5 заявок в секунду. Визначити середній час перебування заявок в системі, якщо відомо, що середня довжина черги дорівнює 6.

14. Інтенсивність обслуговування заявок в СМО дорівнює 5 заявок в секунду. Визначити граничну інтенсивність надходження заявок в систему, при якій середній час перебування заявок в системі не перевищить 1 секунду і середня довжина черги не перевищить 2.

РОЗДІЛ III

ОРГАНІЗАЦІЯ РОБОТИ СМО З НЕОДНОРІДНИМ ПОТОКОМ ЗАЯВОК

3.1. Класифікація стратегій управління в СМО з НПЗ

Розглянемо одноканальну СМО з неоднорідним потоком заявок, в яку надходять N класів заявок, що утворюють найпростіші потоки з інтенсивностями $\lambda_1, \dots, \lambda_H$. Тривалість обслуговування заявок класу h розподілена за довільним законом із середнім значенням τ_{sh} і коефіцієнтом варіації $\nu_{\tau_{sh}}$.

Для систем масового обслуговування з неоднорідним потоком заявок ускладнюється процес їх функціонування і в даному випадку говорять про стратегію управління такими СМО.

Стратегія управління потоками заявок в моделях масового обслуговування задається у вигляді:

- дисципліни буферизації (ДБ);
- дисципліни обслуговування (ДО).

ДБ та ДО можуть бути класифіковані за такими ознаками:

- наявність пріоритетів між заявками різних класів;
- спосіб (режим) витіснення заявок з черги (для ДБ);
- спосіб призначення заявок на обслуговування (для ДО);
- правило витіснення або вибору заявок на обслуговування;
- можливість зміни пріоритетів.

Одна з можливих класифікацій дисциплін буферизації в відповідності до перерахованих ознак представлена на рис.3.1.

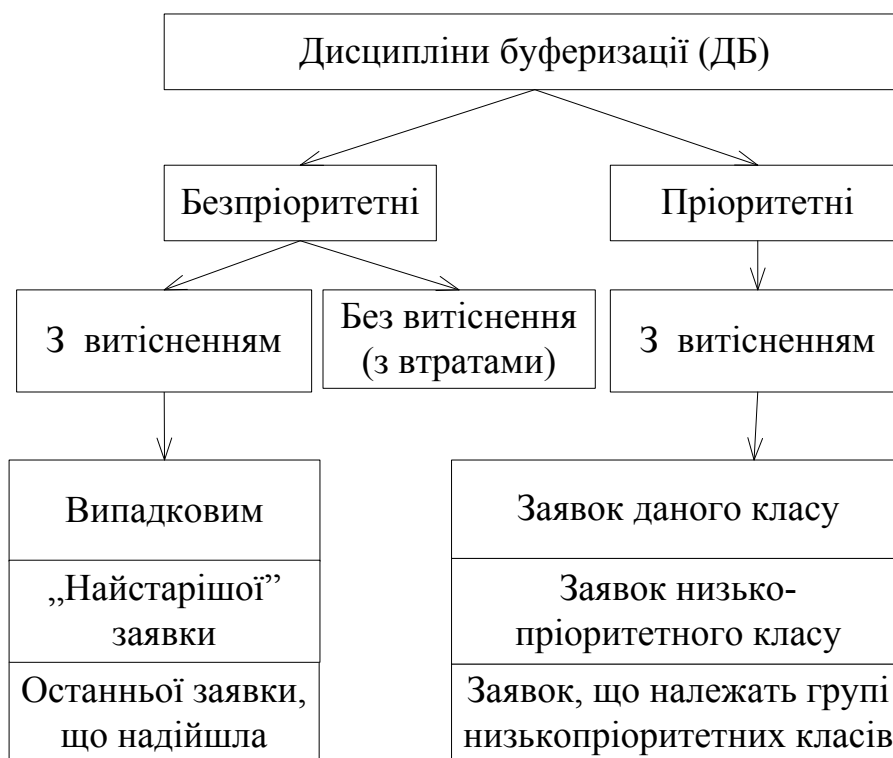


Рис. 3.1. Класифікація дисциплін буферизації

В залежності від наявності або відсутності пріоритетів між заявками різних класів усі ДБ можуть бути розбиті на дві групи:

- безпріоритетні;
- пріоритетні.

За способом витіснення заявок з накопичувача можна виділити наступні класи ДБ:

- без витіснення заявок: заявки, що надійшли в систему і застали накопичувач заповненим до кінця, втрачаються;
- з витісненням заявки даного класу, тобто такого ж класу, що і надійшла;
- з витісненням заявки самого низькопріоритетного класу;
- з витісненням заявки, що належить групі низькопріоритетних класів.

Два перших класи відносяться до безпріоритетних ДБ, а решта - до пріоритетних.

ДБ можуть використовувати такі правила витіснення заявок з накопичувача:

- витіснення випадкове;
- витіснення останньої заявки, тобто яка надійшла в систему пізніше за всіх;
- витіснення «довгої» заявки, «найстарішої», тобто що знаходиться в накопичувачі довше всіх.

Часто ємність накопичувача в моделях передбачається необмеженою, незважаючи на те, що в реальній системі відповідна ємність обмежена. Таке припущення виправдане в тих випадках, коли ймовірність втрати заявки в реальній системі через переповнення накопичувача обмеженої ємності менше 10^{-3} , оскільки в цьому випадку ДБ практично не впливає на характеристики обслуговування заявок.

На рис.3.2 представлена класифікація дисциплін обслуговування заявок за тими ж ознаками, що і для ДБ.

В залежності від наявності або відсутності пріоритетів між заявками різних класів усі ДО, як і ДБ, можуть бути розбиті на дві групи:

- безпріоритетні;
- пріоритетні.

За способом призначення заявок на обслуговування ДО можуть бути розділені на дисципліни:

- одиночного режиму;
- групового режиму;
- комбінованого режиму.

У ДО одиночного режиму всякий раз на обслуговування призначається тільки одна заявка (перегляд черг з метою призначення на обслуговування в приладі чергової заявки виконується після обслуговування кожної заявки).

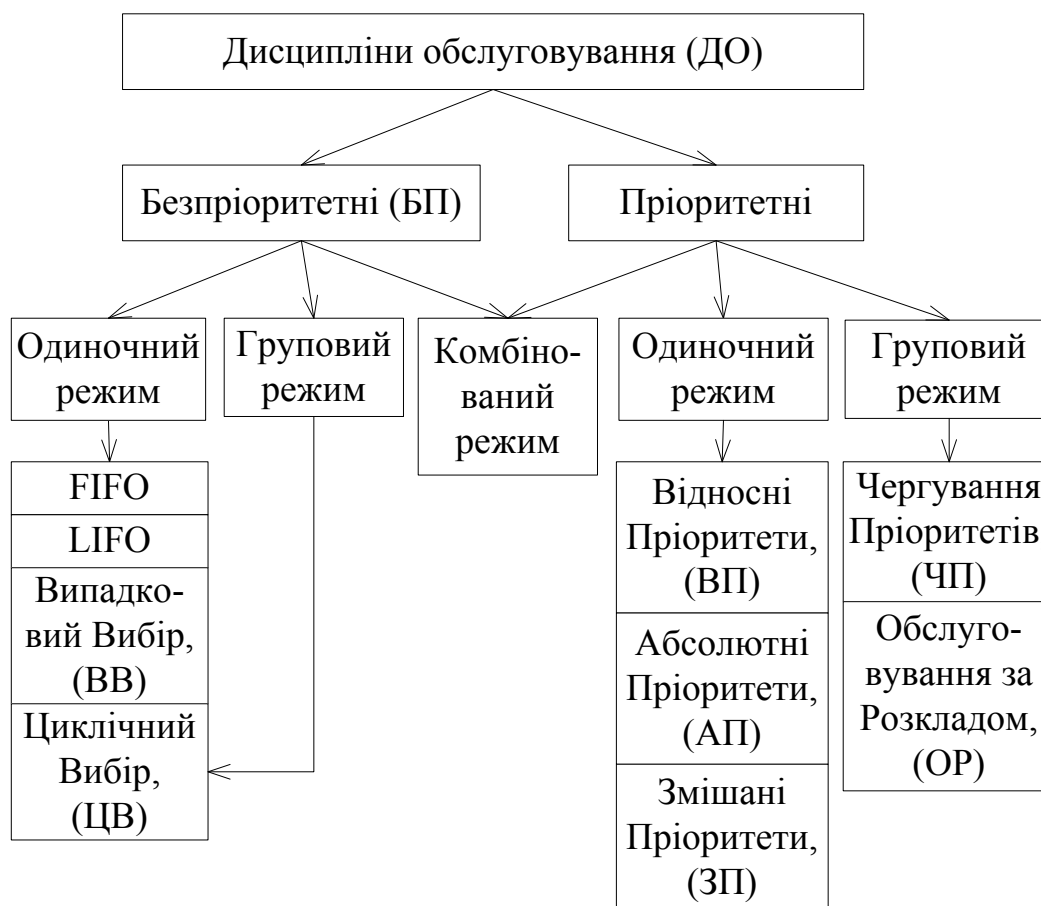


Рис. 3.2. Класифікація дисциплін обслуговування

У ДО групового режиму всякий раз на обслуговування призначається група заявок однієї черги (перегляд черг з метою чергового призначення на обслуговування виконується тільки після обслуговування всіх заявок раніше призначеної групи). У граничному випадку група заявок, яка призначається на обслуговування, може включати в себе всі заявки даної черги. Заявки призначеної на обслуговування групи послідовно вибираються з черги і обслуговуються приладом, після чого на обслуговування призначається наступна група заявок іншої черги відповідно до заданої ДО.

Комбінований режим - комбінація одиночного і групового режимів, коли частина черг заявок обробляється в одиночному режимі, а інша частина - в груповому.

ДО можуть використовувати такі правила вибору заявок на обслуговування:

1) безпріоритетні:

а) обслуговування в порядку надходження (FIFO – First In First

Out), коли на обслуговування вибирається заявка, що надійшла в систему раніше за інших;

б) обслуговування в зворотному порядку (LIFO - Last In First Out), коли на обслуговування вибирається заявка, що надійшла в систему пізніше інших;

в) обслуговування у випадковому порядку (ВВ), коли на обслуговування заявка вибирається випадковим чином;

г) обслуговування в циклічному порядку (ЦВ), коли на обслуговування заявки вибираються в процесі циклічного опитування накопичувачів в послідовності 1, 2, ..., Н (Н - кількість накопичувачів), після чого зазначена послідовність повторюється;

2) пріоритетні:

а) з відносними пріоритетами (ВП), які враховуються тільки в моменти завершення обслуговування заявок при виборі нової заявки на обслуговування і не впливають на процес обслуговування низькопріоритетних заявок в приладі; іншими словами, надходження в систему заявки з більш високим пріоритетом порівняно з тим, що обслуговується в приладі, не призводить до переривання заявки, яка обслуговується;

б) з абсолютними пріоритетами (АП), які, на відміну від ВП, при надходженні високопріоритетної заявки обслуговування заявка з низьким пріоритетом переривається і на обслуговування приймається високо пріоритетна заявка; при цьому перервана заявка може бути повернута в накопичувач або видалена з системи; якщо заявка повернута в накопичувач, то її подальше обслуговування може бути продовжено з перерваного місця або розпочато заново, тобто з самого початку;

в) зі змішаними пріоритетами (ЗП), що представляють собою будь-яку комбінацію безпріоритетних дисциплін обслуговування, ВП і АП;

г) з пріоритетами, що чергуються (ЧП), які є аналогом ВП і проявляються тільки в моменти завершення обслуговування групи заявок однієї черги і призначення нової групи;

д) обслуговування за розкладом (ОР), коли заявки різних класів (що знаходяться в різних накопичувачах) вибираються на обслуговування відповідно з деяким розкладом (планом), що задає послідовність опитування черг заявок, наприклад, у випадку трьох класів заявок (накопичувачів) розклад може мати вигляд: {1, 2, 1, 3, 1, 2}.

Дисципліни FIFO, LIFO, ВВ, ЦВ, ВП, АП і ЗП відносяться до дис-

циплін одиночного режиму. Очевидно, що дисципліни групового режиму ЦВ, ЧП і ОР, в окремому випадку можуть бути реалізовані як ДО одиночного режиму, якщо розмір групи, що призначається на обслуговування, дорівнює 1, при цьому ДО ЧП вироджується в ДО ВП.

Серед представлених ДО особливе місце займають дисципліни зі змішаними пріоритетами (ЗП), що мають багато спільного з перерахованими ДО одиночного режиму [7].

Для математичного опису ДО ЗП використовується матриця пріоритетів (МП), що представляє собою квадратну матрицю:

$Q = [q_{ij} | i, j = 1, \dots, H]$, де H - число класів заявок, що надходять в систему.

Елемент q_{ij} матриці задає пріоритет заявок класу i по відношенню до заявок класу j і може приймати такі значення:

- 0 - немає пріоритету;
- 1 - пріоритет відносний (ВП);
- 2 - пріоритет абсолютний (АП).

Елементи МП повинні відповідати таким вимогам:

- $q_{ii} = 0$, так як між заявками одного і того ж класу не можуть бути встановлені пріоритети;
- якщо $q_{ij} = 1$ або 2 , то $q_{ji} = 0$, тому що якщо заявки класу i мають пріоритет до заявок класу j , то останні не можуть мати пріоритет до заявок класу i ($i, j = \overline{1, H}$).

В залежності від можливості зміни пріоритетів в процесі функціонування системи пріоритетні дисципліни буферизації і обслуговування діляться на два класи:

- зі статичними пріоритетами, які не змінюються з часом;
- з динамічними пріоритетами, які можуть змінюватися в процесі функціонування системи залежно від різних чинників, наприклад, при досягненні деякого критичного значення довжини черги заявок якогось класу, що володіє низьким пріоритетом; йому може бути наданий більш високий пріоритет.

3.2. Методи управління в аналітичних моделях СМО

В цьому розділі і надалі ми будемо використовувати ті дисципліни обслуговування заявок, які дозволяють отримати аналітичні вирази для оцінки характеристик СМО.

Вибір заявок з черги на обслуговування здійснюється відповідно до заданої дисципліни обслуговування, в якості якої будемо розглядати:

- дисципліну обслуговування безпріоритетну (ДО БП), при якій заявки вибираються на обслуговування в порядку надходження;
- дисципліну обслуговування заявок з відносними пріоритетами (ДО ВП);
- дисципліну обслуговування заявок з абсолютними пріоритетами (ДО АП).

В якості основної характеристики, що описує ефективність функціонування системи, будемо розглядати середні часи очікування заявок різних класів, на основі яких легко можуть бути розраховані всі інші характеристики з використанням фундаментальних залежностей, представлених в даному розділі.

При цьому слід мати на увазі, що представлені нижче формули були отримані при наступних припущеннях:

1) СМО містить *один обслуговуючий прилад*, який в кожен момент часу може обслуговувати тільки одну заявку;

2) СМО має *накопичувач заявок необмеженої ємності*, що означає відсутність відмов заявкам, що надходять, при їх постановці в чергу, тобто будь-яка вхідна заявка завжди знайде у накопичувачі місце для очікування незалежно від того, скільки заявок вже знаходиться в черзі;

3) заявки різних класів, що надходять в СМО незалежно один від одного, утворюють *найпростіші потоки*;

4) тривалість обробки заявок кожного класу в приладі розподілені *за довільним законом* і не залежать один від одного;

5) обслуговуючий *прилад не простоює*, якщо в системі (накопичувачі) є хоча б одна заявка будь якого класу, причому після завершення обслуговування поточної заявки миттєво з накопичувача вибирається наступна заявка відповідно до заданої дисципліни обслуговування;

6) при використанні ДО БП заявки різних класів вибираються на обслуговування тільки в залежності від часу надходження в систему за правилом *«раніше прийшов - раніше обслужений»*, незалежно від номера класу, до якого належить заявка;

7) при використанні пріоритетних дисциплін (ДО ВП та ДО АП) пріоритети класам заявок призначені за принципом *«клас з меншим номером має більш високий пріоритет»*, тобто найвищий пріоритет мають заявки класу 1;

8) у разі ДО АП заявка, обслуговування якої перервано більш високопріоритетною заявкою, *повертається в накопичувач*, де очікує подальшого обслуговування, причому її *дообслуговування триває з перерваного місця*.

3.3. Характеристики та властивості ДО БП

При безпріоритетних ДО середні часи очікування однакові для всіх класів заявок і визначаються за такою формулою:

$$\tau_{wh}^{БП} = \tau_w^{БП} = \frac{\sum_{i=1}^H \lambda_i \tau_{si}^2 (1 + v_{\tau_{si}}^2)}{2(1 - R)} \quad (h = 1, \dots, H), \quad (3.1)$$

де $R = \sum_{i=1}^H \rho_i = \sum_{i=1}^H \lambda_i \tau_{si}$ - сумарне завантаження системи.

Вираз (3.1) отримано в припущенні, що в системі існує стаціонарний режим і відсутнє перевантаження: $R < 1$.

Аналіз представленої аналітичної залежності (3.1) дозволяє виявити властивості ДО БП і сформулювати наступні *висновки*.

1. Середній час очікування заявок різних класів при використанні ДО БП однаковий за будь-яких інтенсивностях надходження

$\lambda_1, \dots, \lambda_H$ і законах розподілів $B_1(\tau_s), \dots, B_H(\tau_s)$ тривалостей обробки заявок:

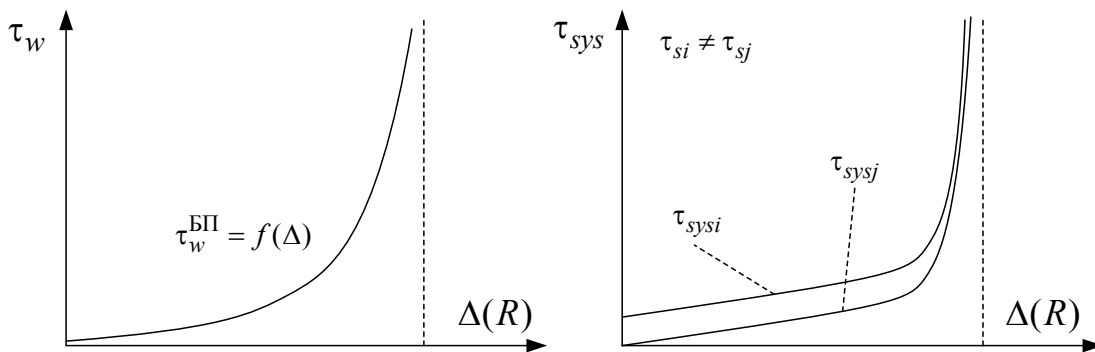
$$\tau_{wh}^{\text{БП}} = \tau_w^{\text{БП}} \text{ для всіх } h = 1, \dots, H.$$

Зазначимо, що середні часи перебування в системі заявок різних класів, в загальному випадку, різні, так як різні тривалості обробки:

$$\tau_{sysh}^{\text{БП}} = \tau_w^{\text{БП}} + \tau_{sh} \quad (h = 1, \dots, H).$$

2. Середній час очікування заявок в черзі мінімальний при постійній (детермінованій) тривалості обслуговування заявок кожного класу, коли коефіцієнт варіації тривалості обслуговування $v_{\tau_{sh}} = 0$ і збільшується з ростом коефіцієнта варіації (дисперсії) тривалості обслуговування. Зауважимо, що залежність середнього часу очікування від коефіцієнта варіації $v_{\tau_{sh}}$ носить нелінійний характер. Так, наприклад, при експоненційно розподіленій тривалості обслуговування, коли $v_{\tau_{sh}} = 1$, середній час очікування заявок збільшується в 2 рази, а при $v_{\tau_{sh}} = 2$ - в 5 разів, порівняно з детермінованим обслуговуванням.

3. Середній час очікування заявок істотно залежить від сумарного навантаження Δ (завантаження R) системи (рис.3.3, а).



3.3. Залежність середнього часу очікування (а) і середнього часу перебування (б) від навантаження при ДО БП

При $\Delta \geq 1$ ($R \rightarrow 1$) час очікування заявок всіх класів зростає необмежено: $\tau_w^{\text{БП}} \rightarrow \infty$, тобто заявки можуть очікувати обслуговування як завгодно довго. Відзначимо, що збільшення сумарного навантаження може бути обумовлено двома факторами: збільшенням інтенсивностей надходження в систему заявок різних класів або збільшенням тривалості обслуговування заявок (наприклад, за рахунок зменшення швидкості роботи обслуговуючого приладу).

Залежність середнього часу перебування в системі заявок різних класів від сумарного навантаження аналогічна залежності часу очікування (рис.3.3, б). Єдина відмінність полягає в тому, що середні часи перебування в системі заявок різних класів, в загальному випадку, різні, тобто $\tau_{\text{sys}i} \neq \tau_{\text{sys}j}$ ($i \neq j$), тому на графіку, на відміну від часу очікування, можуть відображатися кілька залежностей. Ця відмінність обумовлена розходженням тривалостей обслуговування заявок різних класів.

Аналогічно, на графіках, що відображають залежності середніх довжин черг і числа заявок в системі від сумарного навантаження, в загальному випадку, будуть зображуватися кількома кривими, що відповідають різним класам заявок. Відзначимо, що середні довжини черг заявок різних класів, незважаючи на однаковий час очікування, в загальному випадку, різні і, відповідно до формули Літтла ($l_h = \lambda_h \tau_w^{\text{БП}}$), збігаються тільки в разі рівності інтенсивностей надходження заявок різних класів в систему.

4. Можна показати, що для безприоритетних дисципліни обслуговування в зворотному порядку (LIFO), коли заявки на обслуговування вибираються за правилом «останній прийшов - перший обслужений», середні часи очікування заявок будуть такими ж, як і при обслуговуванні в порядку надходження (FIFO), але дисперсія часу очікування буде більше. Це обумовлено тим, що заявки, що надійшли останніми, будуть очікувати незначний час, в той час як заявки, що потрапили в початок черги, можуть очікувати обслуговування достатньо довго, що обумовлює великий розкид значень часу очікування.

5. Аналітичне дослідження дисципліни обслуговування в циклічному порядку (ДО ЦП) досить складно і пов'язано з громіздкими математичними викладками. Тому, не виписуючи громіздких формул, відзначимо лише найбільш характерні особливості цієї ДО.

Для дисципліни обслуговування в циклічному порядку середній час очікування заявок різних класів в загальному випадку не однаковий (рис.3.4).

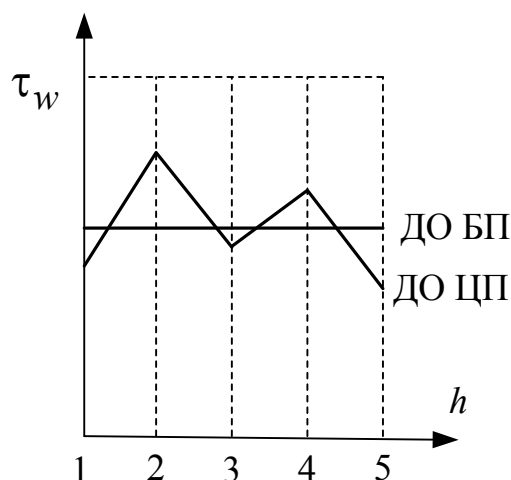


Рис. 3.4. Залежність часу очікування від номеру класу заявок

Ця відмінність залежить від співвідношення параметрів потоків ($\lambda_1, \dots, \lambda_H$) та обслуговування ($\tau_{s1}, \dots, \tau_{sH}$) заявок різних класів. В деяких випадках ДО ЦП дозволяє забезпечити меншу сумарну довжину черги заявок, ніж ДО БП. Залежність середнього часу очікування заявок кожного класу від сумарного навантаження Δ має такий же вигляд, як і для ДО БП (рис.3.3).

3.4. Характеристики та властивості ДО ВП

Пріоритети називаються відносними, якщо вони враховуються тільки в момент вибору заявки на обслуговування і не позначаються на роботі системи в період обробки заявки будь-якого класу (пріоритету). Відносність пріоритету пов'язана з наступним. Після завершення обслуговування будь якої заявки, з черги на обслуговування вибирається заявка класу з найвищим пріоритетом, що надійшла раніше інших заявок цього класу (такого ж пріоритету). Якщо в процесі її обробки в систему надійдуть заявки з більш високим пріоритетом, то

обробку даної заявки не буде припинено, тобто ця заявка, захопивши прилад, виявляється як би більш пріоритетною. Таким чином, пріоритет відносний в тому сенсі, що він має місце лише в момент вибору заявок на обслуговування і відсутній, якщо прилад зайнятий обробкою іншої заявки.

Введення відносних пріоритетів (ВП) дозволяє зменшити, по порівнянню з ДО БП, час очікування високопріоритетних заявок. При описі властивостей, для визначеності, будемо вважати, що відносні пріоритети призначені за правилом: «більш високий пріоритет класу заявок - з меншим номером». Для ДО ВП середній час очікування заявок класу h визначається за такою формулою:

$$\tau_{wh}^{ВП} = \frac{\sum_{i=1}^H \lambda_i \tau_{si}^2 (1 + v_{\tau_{si}}^2)}{2(1 - R_{h-1})(1 - R_h)} \quad (h = 1, \dots, H), \quad (3.2)$$

де R_{h-1} і R_h - сумарні завантаження, створювані заявками, які мають пріоритет не нижче $(h - 1)$ і h відповідно:

$$R_{h-1} = \sum_{i=1}^{h-1} \rho_i; \quad R_h = \sum_{i=1}^h \rho_i \quad (3.3)$$

Аналіз представленої залежності (3.2) дозволяє виявити властивості ДО ВП і сформулювати наступні висновки.

1. Введення відносних пріоритетів у порівнянні з ДО БП призводить до зменшення часу очікування високопріоритетних заявок першого класу і до збільшення часу очікування низькопріоритетних заявок класу H :

$$\tau_{w1}^{ВП} < \tau_1^{БП} \quad \text{і} \quad \tau_{wH}^{ВП} > \tau_{wH}^{БП}.$$

2. При використанні ДО ВП середні часи очікування заявок монотонно збільшуються зі зменшенням пріоритету при будь-яких інтенсивностях надходження $\lambda_1, \dots, \lambda_H$ і законах розподілу тривалостей обслуговування:

$$\tau_{w1}^{ВП} < \tau_{w2}^{ВП} < \dots < \tau_{wH}^{ВП}.$$

Зазначимо, що для середніх часів перебування заявок різних класів останнє співвідношення, в загальному випадку, може і не виконуватися.

Властивості, сформульовані вище, ілюструються рис.3.5, а). Графік показує характер залежності середнього часу очікування заявок τ_{wh} від номера класу h при використанні ДО БП та ДО ВП.

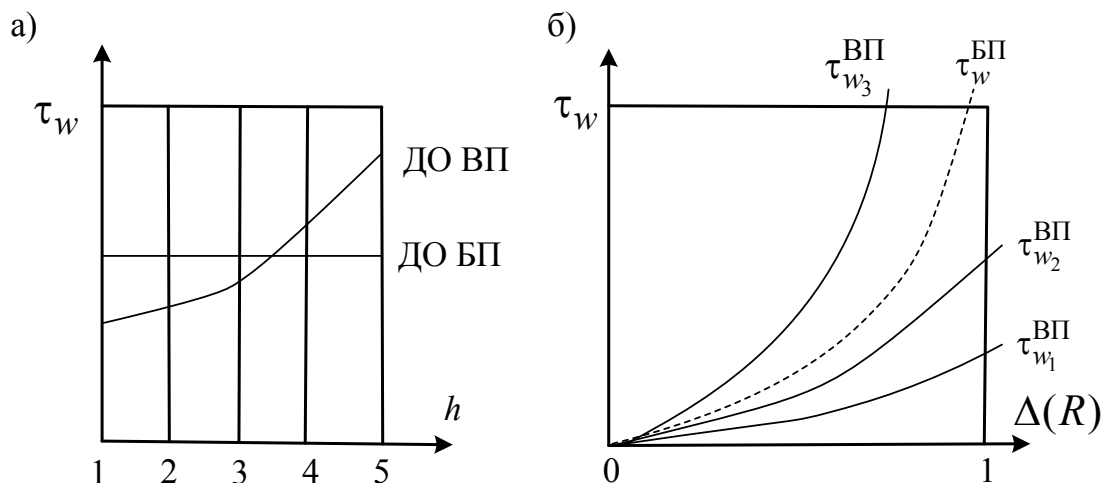


Рис. 3.5. Залежність середнього часу очікування від номера класу (а) і від навантаження (б) для ДО ВП та ДО БП

3. На рис.3.5 б) показані залежності середнього часу очікування заявок різних класів від сумарного навантаження Δ системи при використанні ДО ВП. Тут же для порівняння наведена аналогічна залежність для ДО БП (штрихова лінія). Характер залежностей свідчить про те, що для ДО ВП при $\Delta \rightarrow 1$ різко збільшується час очікування заявок низькопріоритетних класів, у той час як для високопріоритетних заявок це збільшення незначне. Більш того, для високопріоритетних заявок забезпечується досить гарна якість обслуговування, тобто невеликий час очікування навіть при виникненні перевантажень, коли сумарне навантаження стає більше одиниці: $\Delta \geq 1$. Це властивість зветься захистом від перевантажень і забезпечується за рахунок відмови в обслуговуванні низькопріоритетним заявками, час очікування яких при цьому різко зростає. При ДО БП захист від перевантажень відсутній для всіх класів заявок.

4. Розглянемо більш детально властивість захисту від перевантажень при ДО ВП, для чого побудуємо залежності середнього часу

очікування заявок трьох класів при значному зростанні навантаження Δ (рис.3.6).

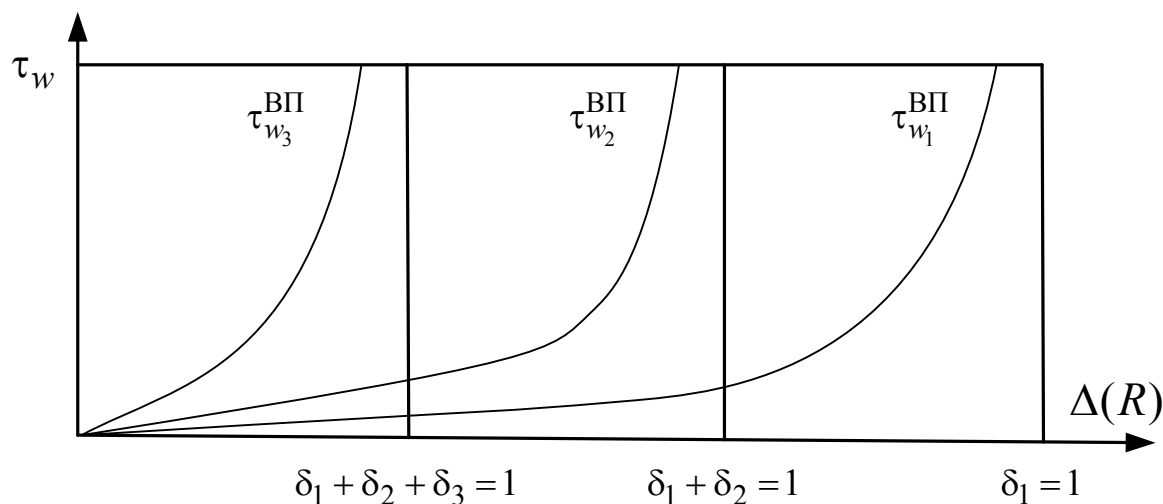


Рис. 3.6. Залежності середнього часу очікування заявок різних класів від навантаження

При досягненні сумарного навантаження, що створюється заявками всіх трьох класів, значення 1 ($\delta_1 + \delta_2 + \delta_3 = 1$) час очікування заявок третього класу спрямовується в нескінченність, що означає відмову в обслуговуванні, при цьому заявки класів 1 і 2 продовжують обслуговуватися і мають кінцевий час очікування. Подальше збільшення навантаження призводить до відмови в обслуговуванні заявок другого класу при $\delta_1 + \delta_2 = 1$, тобто коли навантаження, яке створюється заявками 1-го і 2-го класів, досягне значення 1. Заявки першого класу отримають відмову в обслуговуванні при $\delta_1 = 1$. Таким чином, на відміну від ДО БП при ДО ВП система повністю перестане обслуговувати заявки, тобто функціонувати, тільки в тому випадку, якщо навантаження, що створюється заявками самого високопріоритетного (першого) класу, досягне значення 1.

3.5. Характеристики та властивості ДО АП

Іноді час очікування заявок деяких класів необхідно зменшити в такій мірі, яка недосяжна при використанні ДО ВП. Можна припустити, що час очікування зменшиться, якщо при надходженні високопріоритетної заявки обслуговування заявки, яка прийшла раніше з низьким пріоритетом, переривається, і прилад переходить до обслуговування високопріоритетної заявки. Пріоритет, перериваючий обслуговування низькопріоритетної заявки, називається абсолютним, а відповідна дисципліна – дисципліною обслуговування з абсолютними пріоритетами (ДО АП). Перервана заявка може бути втрачена або повернута в накопичувач, де вона буде чекати подальшого обслуговування. В останньому випадку можливі два варіанти продовження обслуговування перерваної заявки:

- обслуговування з початку, тобто перервана заявка буде оброблятися заново з самого початку;
- дообслуговування, коли обробка перерваної заявки в приладі буде виконуватися з перерваного місця.

Надалі, якщо не обумовлено інше, будемо припускати дообробку перерваної заявки.

Для ДО АП середній час очікування заявок класу h визначається за такою формулою:

$$\tau_{wh}^{АП} = \frac{\sum_{i=1}^h \lambda_i \tau_{si} (1 + v_{\tau_{si}}^2)}{2(1 - R_{h-1})(1 - R_h)} + \frac{R_{h-1} \tau_{sh}}{1 - R_{h-1}} \quad (h=1, \dots, H) \quad (3.4)$$

де R_{h-1} і R_h - сумарні завантаження, що створюються заявками, які мають пріоритет не нижче $(h-1)$ і h відповідно, і визначаються за формулами (3.3).

Аналіз виразу (3.4) для розрахунку середнього часу очікування заявок при використанні ДО АП і його зіставлення з аналогічним виразом (3.2) для ДО ВП дозволяє виявити властивості ДО АП та сформулювати наступні висновки.

1. Вираз (3.4) містить два доданки:

$$\tau_{wh}^{АП} = \tau_{wh}^{поч} + \tau_{wh}^{пер},$$

відображають середній час очікування початку обслуговування $\tau_{wh}^{поч}$ середній час очікування в перерваному стані $\tau_{wh}^{пер}$ відповідно:

$$\tau_{wh}^{поч} = \frac{\sum_{i=1}^h \lambda_i \tau_{si} (1 + v_{\tau_{si}}^2)}{2(1 - R_{h-1})(1 - R_h)}, \quad \tau_{wh}^{пер} = \frac{R_{h-1} \tau_{sh}}{1 - R_{h-1}} \quad (h = 1, \dots, H)$$

2. Час очікування заявок класу h залежить тільки від значень параметрів класів $1, \dots, h$ заявок, що мають більш високий або такий же пріоритет, і не залежить від параметрів класів заявок $h + 1, \dots, H$, які мають більш низький пріоритет.

3. Для заявок класу 1, які мають найвищий абсолютний пріоритет, забезпечується мінімально можливий час очікування порівняно з усіма іншими ДО, тобто при будь-якій іншій ДО середній час очікування заявок першого класу не може бути менше, ніж при ДО АП. Це пояснюється тим, що в разі ДО АП заявки першого класу обслуговуються як би в ізоляції, незалежно від заявок інших класів.

4. Часи очікування початку обслуговування $\tau_{wh}^{поч}$ монотонно збільшуються зі зменшенням пріоритету: $\tau_{w1}^{поч} < \tau_{w2}^{поч} < \dots < \tau_{wH}^{поч}$, проте час очікування високопріоритетної заявки в перерваному стані $\tau_{wh}^{пер}$ може виявитися більше часу очікування $\tau_{w(h+1)}^{пер}$ заявки з більш низьким пріоритетом, якщо тривалості обслуговування пов'язані співвідношенням $\tau_{sh} \gg \tau_{s(h+1)}$, так як кількість переривань заявками вищого пріоритету і, отже, час очікування в перерваному стані, прямо пропорційно залежить від тривалості обслуговування заявок даного класу. Внаслідок цього, повний час очікування заявок високопріоритетного класу, що складається з часу очікування початку обслуговування і часу очікування в перерваному стані, може виявитися більше, ніж у заявок класу з низьким пріоритетом:

$$\tau_{wh}^{АП} \gg \tau_{w(h+1)}^{АП}$$

Очевидно, що

$$\tau_{w1}^{АП} < \tau_{w2}^{АП} < \dots < \tau_{wH}^{АП},$$

якщо тривалості обслуговування заявок різних класів пов'язані співвідношенням

$$\tau_{s1} \leq \tau_{s2} \leq \dots \leq \tau_{sH}$$

5. Введення АП порівняно з ВП призводить до зменшення середнього часу очікування самих високопріоритетних заявок першого класу і до його збільшення для заявок класу Н:

$$\tau_{w1}^{АП} < \tau_{w1}^{ВП} \text{ і } \tau_{wH}^{АП} > \tau_{wH}^{ВП}$$

Два останніх результати ілюструються рис.3.7 а). Для ДО АП пунктиром показано випадок, коли

$$\tau_{w3}^{АП} \gg \tau_{w4}^{ВП},$$

з чого випливає, що

$$\tau_{s3} \gg \tau_{s4}.$$

Залежність повного часу очікування від сумарного навантаження Y системи при використанні ДО АП аналогічна залежності для ДО ВП (див. рис.3.7, б) з тією лише відмінністю, що при ДО АП високопріоритетні заявки краще захищені від перевантажень.

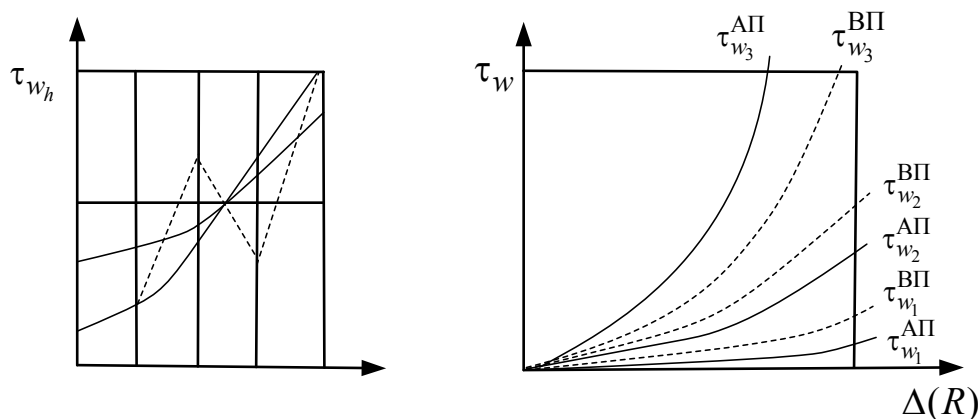


Рис. 3.7. Залежності середнього часу очікування від номера класу (а) і від навантаження (б) для ДО АП і ДО ВП

3.6. Закони збереження

Зміна ДО дозволяє зменшити час очікування високопріоритетних заявок за рахунок збільшення часу очікування нізкопріоритетних заявок. Очевидно, що за рахунок зміни ДО не можна домогтися того, щоб зменшився або збільшився час очікування заявок всіх класів. Цей факт сформульований у вигляді **закону збереження часу очікування**.

Формулювання закону збереження часу очікування. Для будь-якої дисципліни обслуговування (ДО)

$$\sum_{i=1}^H \rho_i \tau_{wi} = \text{Const} / \text{ДО}, \quad (3.5)$$

тобто сума добутків завантажень ρ_i на середній час очікування τ_{wi} ($i = \overline{1, H}$) заявок всіх класів інваріантна щодо ДО.

Закон збереження часу очікування виконується при наступних умовах:

- система без втрат - всі заявки на обслуговування задовольняються;
- система простоє лише в тому випадку, коли в ній немає заявок;
- при наявності переривань тривалість обслуговування перерваних заявок розподілена за експоненційним законом;
- всі потоки заявок, що надходять, найпростіші, і тривалості обслуговування не залежать від інтенсивностей потоків заявок.

Значення константи в законі збереження можна визначити наступним чином. Оскільки закон збереження справедливий для будь-яких ДО, що задовольняють перерахованим умовам, то він справедливий і для ДО БП, для якої

$$\tau_{wh}^{\text{БП}} = \tau_w^{\text{БП}} \text{ для всіх } (h = 1, \dots, H).$$

Звідси знаходимо значення константи:

$$Const = \tau_w^{БП} \sum_{i=1}^H \rho_i = \tau_w^{БП} R$$

Якщо підставити отримане значення константи у формулу (3.1) для розрахунку $\tau_w^{БП}$ в закон збереження, остаточно отримаємо:

$$\sum_{i=1}^H \rho_i \tau_{wi} = \frac{R \sum_{i=1}^H \lambda_i \tau_{si}^2 (1 + v_{\tau_{si}}^2)}{2(1 - R)} . \quad (3.6)$$

Закон збереження часу очікування універсальний і справедливий для всіх ДО, що задовольняють зазначеним умовам. Його можна використовувати для оцінки достовірності наближених результатів, отриманих при дослідженні складних ДО та проведення імітаційного моделювання, а також при вирішенні задач синтезу.

Модифікація закону збереження. Закон збереження може бути модифікований стосовно до часу перебування заявок в системі з урахуванням того, що $\tau_{wi} = \tau_{sysi} - \tau_{si}$. Підставивши цей вираз в закон збереження часу очікування (3.6) після деяких перетворень, отримаємо закон збереження часу перебування:

$$\sum_{i=1}^H \rho_i \tau_{sysi} = \frac{R \sum_{i=1}^H \lambda_i \tau_{si}^2 (1 + v_{\tau_{si}}^2)}{2(1 - R)} + \sum_{i=1}^H \rho_i \tau_{si} \quad (3.7)$$

Зауважимо, що зміна ДО призводить лише до зміни часу очікування і часу перебування, а решта величин, що входять до виразу (3.6) і (3.7), не змінюються.

Розглянемо випадок, коли середні тривалості обслуговування заявок різних класів однакові: $\tau_{si} = \tau_s = const$ для всіх $i = \overline{1, H}$. Тоді вираз (3.6) може бути перетворений в такий спосіб:

$$\sum_{i=1}^H \lambda_i \tau_s \tau_{wi} = \tau_s \sum_{i=1}^H \lambda_i \tau_{wi} = \tau_s \sum_{i=1}^H l_i = \tau_s L = Const / ДО ,$$

звідки отримаємо нове формулювання закону збереження у вигляді **закону збереження сумарної довжини черги заявок**:

$$\sum_{i=1}^H \lambda_i \tau_{wi} = L = \text{Const} / \text{ДО} . \quad (3.8)$$

Таким чином, якщо середні тривалості обслуговування заявок різних класів однакові, то зміна ДО не призводить до зміни сумарної довжини L черг заявок всіх класів, яка залишається постійною. У той же час довжини черг l_i ($i = \overline{1, H}$) заявок кожного класу змінюються зі зміною ДО.

3.7. Самоконтроль. Перелік питань і завдань

1. При якій дисципліні обслуговування заявок середні часи очікування заявок різних класів однакові?
2. За яких умов середній час очікування заявок для ДО ВП є зростаючою (спадною) функцією від номера класу заявок?
3. За яких умов середній час перебування заявок для ДО ВП є зростаючою функцією від номера класу заявок?
4. Чи може середній час перебування заявок для ДО ВП бути спадною функцією від номера класу заявок, якщо середній час очікування заявок є зростаючою функцією від номера класу заявок? Відповідь пояснити.
5. Чи може заявка з більш високим відносним пріоритетом мати більший час перебування, ніж низькопріоритетним? Відповідь обґрунтувати.
6. За яких умов середній час очікування заявок для дисципліни обслуговування з абсолютними пріоритетами є зростаючою функцією від номера класу заявок?
7. У яких випадках середній час очікування заявок вищого пріоритету при ДО АП може мати більше значення, ніж заявок низького пріоритету?
8. Чи може середній час очікування заявок бути спадною функцією від номера класу (пріоритету) заявок?

9. Чому середній час очікування заявок найвищого пріоритету при використанні ДОАП відрізняється від нуля?

10. При яких умовах характер залежності середнього часу перебування заявок від номера класу збігається з характером залежності середнього часу очікування? Чи можуть залежності середнього часу перебування і середнього часу очікування заявок від номера класу мати протилежний характер?

11. У яких випадках характер залежності середньої довжини черги заявок від номера класу відрізняється від характеру залежності середнього часу очікування заявок? При яких умовах характер залежності середньої довжини черги заявок від номера класу збігається з характером залежності середнього часу очікування заявок?

12. Чи може середнє число заявок в СМО відрізнитися від середньої довжини черги заявок більше, ніж на одиницю? Чи зміниться різниця між ними при зміні дисципліни обслуговування?

13. Чи зміниться різниця між середнім часом перебування і середнім часом очікування для заявок сумарного потоку при зміні дисципліни обслуговування?

14. Чи можна стверджувати, що якщо при якоїсь дисципліни обслуговування значення однієї з характеристик сумарного потоку (наприклад, час очікування) менше значень аналогічної характеристики при іншій дисципліні обслуговування, то і значення решти характеристик при першій дисципліні також будуть менше? Проілюструвати на прикладі.

15. Чи зміняться характеристики обслуговування всіх класів заявок або тільки деяких з них при ДО БП (ДО ВП, ДО АП), якщо змінити: а) інтенсивність надходження заявок будь-якого класу; б) середню тривалість обслуговування будь-якого класу заявок?

16. Намалювати залежності середнього часу очікування та середнього часу перебування заявок в СМО від номера класу заявок для безпріоритетної дисципліни обслуговування. Пояснити характер цих залежностей.

17. Намалювати залежності середнього часу очікування заявок в СМО від сумарного завантаження в разі трьох класів заявок і дисципліни обслуговування з відносними й абсолютними пріоритетами. Провести аналіз цих залежностей.

18. Намалювати залежності середнього часу очікування та середнього часу перебування заявок від номера класу заявок для дисциплін

обслуговування з відносними і з абсолютними пріоритетами. Пояснити характер цих залежностей.

19. Фізичний сенс і математичний запис закону збереження часу очікування. Сформулювати умови, що супроводжують закон збереження. За якої умови у випадку ДО АП не виконується закон збереження часу очікування?

20. Як отримується значення константи в законі збереження часу очікування?

21. За якої умови закон збереження часу очікування вироджується в закон збереження: а) сумарної довжини черги заявок: $L = \text{Const}$; б) сумарного числа заявок в системі: $M = \text{Const}$; в) середнього часу очікування; г) середнього часу перебування?

22. Чи можна, змінивши дисципліну обслуговування, змінити: а) час очікування заявок всіх класів; б) час перебування заявок всіх класів, в) тривалість обслуговування заявок; г) час очікування заявок тільки одного класу; д) час очікування заявок тільки двох класів? Показати на прикладах.

23. У яких випадках слід застосовувати пріоритетні дисципліни обслуговування заявок з відносними й абсолютними пріоритетами?

24. Що таке "захист від перевантажень" і для яких дисциплін обслуговування він існує? Проілюструвати на графіку властивість "захисту від перевантажень".

25. В одноканальну СМО надходять 2 найпростіших потоки заявок з інтенсивностями 0,1 і 0,2 заявок за секунду; тривалості їх обробки відповідно 2 і 4 секунди. Чому дорівнюватиме середній час очікування заявок 1-го класу при використанні безпріоритетної дисципліни?

26. В одноканальну СМО надходять 2 класи заявок з інтенсивностями 0,1 і 0,2 заявок за секунду. Тривалості їх обслуговування відповідно 2 і 3 секунди. Середній час очікування заявок при використанні безпріоритетних дисципліни обслуговування - 5 секунд. Після введення пріоритетів середній час очікування заявок 1-го класу стало дорівнювати 2 секундам. Чому дорівнює середній час очікування заявок 2-го класу?

27. У систему надходять заявки трьох класів з інтенсивностями 2, 1 і 0,5 заявок за секунду відповідно. Сформулювати умови, за яких середній час перебування заявок всіх класів буде однаковим.

28. В одноканальну систему обслуговування надходять заявки двох класів з інтенсивностями 0,5 і 2 заявки за секунду. Інтенсивності

їх обробки відповідно дорівнюють 5 і 1,25 заявок за секунду. а) При яких умовах час перебування заявок 2-го класу буде дорівнює 0,8 секунди? б) Чому дорівнюватиме час перебування заявок 2-го класу, якщо при тих же умовах інтенсивність надходження заявок 1-го класу збільшиться в два рази? в) Чому дорівнюватиме час перебування заявок 2-го класу, якщо при тих же умовах інтенсивність їх надходження збільшиться в два рази? г) Чому дорівнюватиме час перебування заявок 2-го класу, якщо при тих же умовах інтенсивність їх обслуговування збільшиться в два рази?

МЕРЕЖІ МАСОВОГО ОБСЛУГОВУВАННЯ. ТЕОРЕТИЧНІ ЗАСАДИ

У розподіленій інформаційній системі ізольовані черги, на жаль, не є єдиною проблемою для аналітиків. Часто проблеми, які мають бути проаналізовані, пов'язані з декількома взаємопов'язаними чергами. Рисунок 4.1 ілюструє цю ситуацію, де вузли є СМО з чергами, а сполучні лінії являють собою потоки трафіку. З двома елементами такої мережі пов'язані методи аналізу:

- розділення і злиття трафіку, як показано на вузлах 1 і 5, відповідно;
- наявність черг в тандемі, або в серії черг, як показано на вузлах 3 і 4.

У цьому розділі ми розглянемо теоретичні засади, теореми та методи аналізу мереж масового обслуговування (ММО) з відповідними практичними задачами оцінки характеристик мереж СМО.

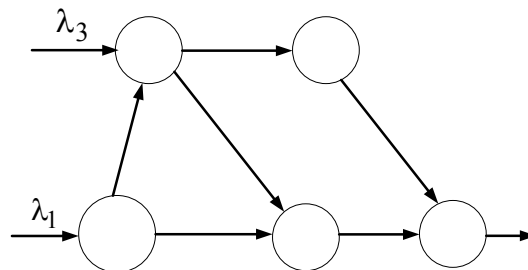
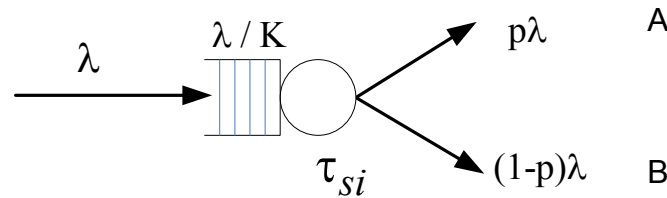


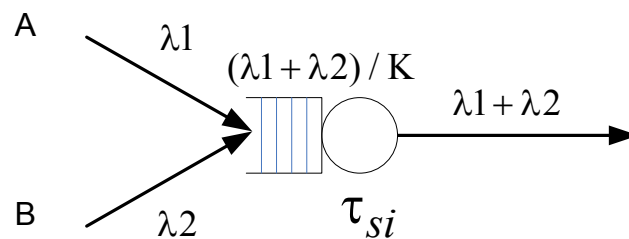
Рис. 4.1. Приклад мережі СМО

4.1. Поділ та об'єднання потоків трафіка

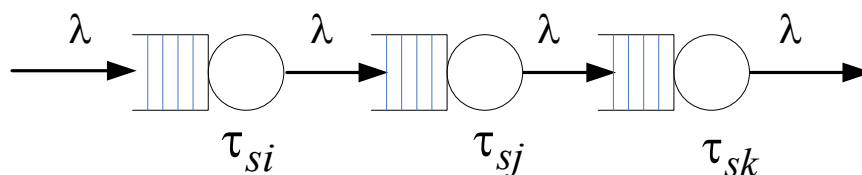
Припустимо, що трафік надходить до черги вузла з середньою швидкістю прибуття λ , і що є два шляхи, А і Б, в яких заявки можуть йти далі (рис. 4.2 а).



(а) Розподілення трафіка



(б) Об'єднання трафіка



(в) Послідовно розташовані черги

Рис. 4.2. Елементи мереж масового обслуговування

Коли заявка оброблена, то вона йде далі за напрямком А з імовірністю p , або у напрямку В з ймовірністю $(1-p)$. Загалом, розподілення потоків трафіку А і В буде відрізнятися від вхідного розподілення. Однак, якщо вхідне розподілення Пуасона, то і два вихідні потоки трафіку також мають розподілення Пуасона із середніми інтенсивностями $p\lambda$ і $(1-p)\lambda$.

Аналогічна ситуація складається при злитті потоків трафіку (рис.

4.2 б). Якщо два потоки Пуасона з середніми інтенсивностями λ_1 і λ_2 будуть об'єднані, то в результаті ми отримаємо потік Пуасона з середньою швидкістю $\lambda_1 + \lambda_2$. Ці результати можна використовувати при розділі трафіку на більше, ніж два потоки, або при об'єднанні більше ніж двох потоків в єдиний трафік.

4.2. СМО тандемів. Теорема Бьорке

СМО, що розглядалися раніше, можна назвати однофазними, тому що кожна заявка проходила одну операцію обслуговування. В мережах СМО ми маємо справу з багатофазним (багато етапним) обслуговуванням, коли заявка обслуговується більш ніж в одному приладі (вузлі).

На рисунку 4.2 с) наведено приклад набору односерверних СМО в ланцюзі, який зветься **тандем**: входом для кожної СМО, крім першої, є вихід попередньої СМО. Припустимо, що на вході першої черги – Пуасонівський потік. Тоді, якщо час обробки в кожній СМО - експоненційний і черги - нескінченні, то вихід в кожній СМО в тандемі - Пуасонівський потік, що статистично збігається з вихідним.

Коли цей потік попадає на поточну СМО в тандемі, час обробки заявок в потоці такий, ніби до цього на існувало ніяких СМО, тобто ця черга перша в ланцюзі. Таким чином, усі СМО незалежні і можуть бути проаналізовані окремо. Середня сумарна затримка для тандему СМО дорівнює сумі середніх затримок в СМО на кожному етапі. Цей результат може бути поширений на випадок, коли деякі або всі вузли в тандемі – мультисерверні СМО.

Ці положення спираються на дві базові теореми для аналізу мереж СМО.

Теорема Бьорке (Burke). Розглянемо найпростішу послідовну систему з двома вузлами, кожен з яких представляє СМО типу М/М/1 (рис. 4.3).

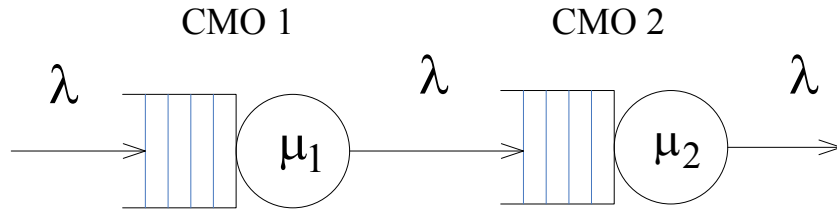


Рис. 4.3. Проста тандемна мережа

Фрагмент часової діаграми для цієї мережі виглядає так, як це показано на рис. 4.4

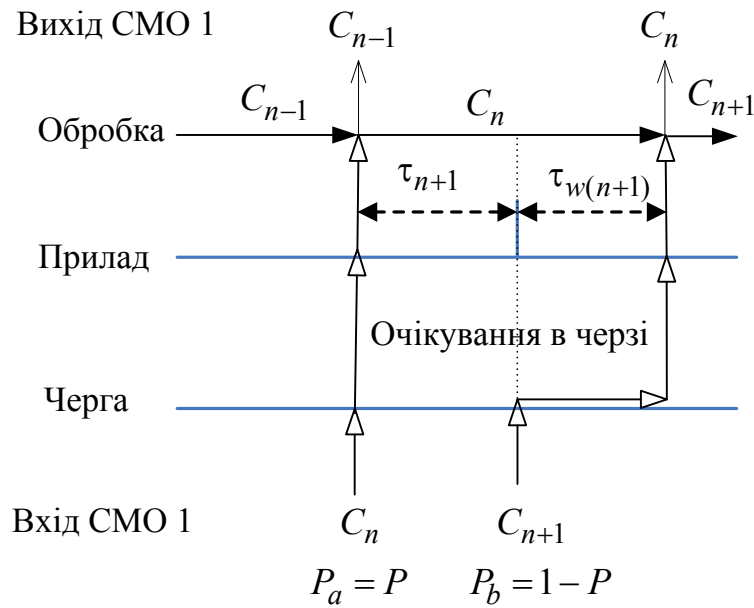


Рис. 4.4. Фрагмент часової діаграми найпростішої тандемної мережі

Нехай $d(\tau_{out})$ – щільність розподілу ймовірностей проміжків часу між послідовними заявками на виході, а $D^*(S)$ – перетворення Лапласа $d(\tau_{out})$. Коли заявка залишає вузол 1, можлива одна з двох подій:

- (а) в черзі є хоча б одна вимога (вузол 1 не порожній) $P_a = P$;
- (б) в черзі немає вимог (вузол 1 порожній) $P_b = 1 - P$.

У першому випадку – проміжок часу, через який наступна заявка покине систему 1, розподілений так само, як і час обслуговування:

$$D^*(S)|_a = B^*(S).$$

У другому випадку – проміжок часу, після закінчення якого, наступна заявка покине систему, є сумою $\tau_{w(n+1)}$ – часу очікування прибором надходження наступної заявки і τ_{n+1} – часу обробки цієї заявки.

Так як ці два проміжки часу розподілені незалежно, то щільність розподілу ймовірностей їх суми дорівнює згортку щільностей розподілу величин, що додаються. Відповідне перетворення Лапласа щільності розподілу:

$$D^*(S)|_b = A^*(S) \cdot B^*(S).$$

Так як проміжок часу τ_{in} між послідовними заявками, що надходять, є показово розподіленою випадковою величиною, то

$$A^*(S) = \frac{\lambda}{S + \lambda}$$

Крім того, враховуючи, що ймовірність того, що заявка покине порожню систему, дорівнює ймовірності того, що заявка, що надходить, застане систему порожньою і дорівнює $1 - \rho$. Можна записати наступне безумовне перетворення Лапласа для щільності розподілу проміжків часу між заявками, що йдуть з системи:

$$\begin{aligned} D^*(S) &= (1 - \rho) \cdot D^*(S)|_{\text{вузол 1 порожній}} + \rho \cdot D^*(S)|_{\text{вузол 1 не порожній}} = \\ &= B^*(S) \cdot \left[(1 - \rho) \cdot \frac{\lambda}{\lambda + S} + \rho \right] = B^*(S) \cdot \left[\frac{(\mu - \lambda) \cdot \lambda}{\mu \cdot (\lambda + S)} + \frac{\lambda}{\mu} \right] = \\ &= B^*(S) \cdot \left[\frac{\lambda}{\mu} \cdot \left(\frac{\mu - \lambda + \lambda + S}{\lambda + S} \right) \right] = B^*(S) \cdot \frac{\lambda \cdot (\mu + S)}{\mu \cdot (\lambda + S)} \end{aligned}$$

Враховуючи, що час обслуговування - показово розподілена випадкова величина, то можна записати

$$B^*(S) = \frac{\mu}{\mu + S}$$

і, остаточно, маємо:

$$D^*(S) = \frac{\lambda}{\lambda + S} = A^*(S)$$

і, отже, розподіл проміжків часу між заявками, що йдуть, так само як і проміжки часу між заявками, що надходять, розподілені за показовим законом з одним і тим же параметром.

Цей результат, в 1956 році встановлений Бьорке, справедливий і для СМО типу М/М/К.

4.3. Теорема і відкриті мережі Джексона (Jackson)

З теореми Бьорке випливає, що будь-яка ациклічна мережа масового обслуговування (в такій мережі виключається можливість повторного відвідування заявкою вузла) з показовим розподілом часу обслуговування, обслуговуючи незалежні пуасоновські джерела заявок, генерує для кожного вузла незалежний Пуасонівський потік.

Для мереж загального виду (у такій мережі можливі петлі) теорема Джексона стверджує, що кожний вузол поводить себе так, ніби на його вхід надходив Пуасонівський потік. Зокрема, якщо позначити через $p(m_1, m_2, \dots, m_N)$ - стаціонарну ймовірність того, що у i -ому вузлі знаходяться m_i заявок, то при $\rho_i/K_i < 1$, згідно теореми Джексона:

$$p(m_1, m_2, \dots, m_N) = \prod_{i=1}^N p_i(m_i),$$

де:

N - число вузлів мережі;

$p_i(m_i)$ - стаціонарна ймовірність того, що m_i заявок знаходяться в системі М/М/К:

$$p_i(m_i) = \begin{cases} p_i(0) \cdot \rho_i^{m_i} / m_i!, & \text{якщо } m_i < K_i; \\ p_i(0) \cdot \rho_i^{m_i} / (K_i! \cdot K_i^{m_i - K_i}), & \text{якщо } m_i = K_i + N \end{cases}$$

Відкриті мережі Джексона. Розглянемо мережу з N вузлами, у i -

На рисунку використовуються наступні позначення:

p_{ij} – ймовірність того, що після обслуговування в i -тому вузлі заявка надійде в вузол j ;

γ_i – інтенсивність зовнішнього потоку заявок;

λ_i – повна інтенсивність потоку заявок в i -му вузлі.

У 1967 році Гордон і Ньюелл (Gordon W.J. Newell G. F.) опублікували опис моделі довільної замкненої Марківської мережі. Фактично це було розвитком окремого випадку, розглянутого в 1963 році Джексоном. У цій моделі міститься m заявок і не допускається надходження заявок ззовні і їх вихід з мережі (тобто $\sum_{j=1}^N p_{ij} = 1$ для всіх i).

Так як $\gamma_i = 0$, то справедливе векторне рівняння: $\Lambda = \Lambda \cdot P$.

Це твердження визначає λ_i не єдиним чином, визначаючи відносні значення компонент λ_i (тобто дозволяє вираховувати з точністю до постійного множника).

У 1971 Мур (Moore F.R.) запропонував використовувати замкнуті Марківські мережі для моделювання роботи обчислювальної системи з багатьма ресурсами, в якому кожен ресурс моделюється вузлом мережі.

У цій моделі заявка проходить в обчислювальній системі від одного ресурсу до іншого, послідовно запитуючи обслуговування. Якщо кілька заявок вимагають обслуговування одного ресурсу, то цей конфлікт вирішується шляхом утворення черги. Модель дозволила передбачати появу «вузького місця» – це такий вузол мережі, в якому відношення ρ_i/K_i , що називають відносним використанням i -ого вузла, має найбільше значення. При деякому $K_i = K_i^{\text{кр}}$ у вузькому місці утворюється нескінченна черга. Крім того, Мур показав, що балансування системи (тобто приведення всіх ρ_i/K_i до одного знаменника) максимізує $K_i^{\text{кр}}$.

Приклад 4.1. Приклад відкритої мережі. Розглянемо приклад розрахунку мережі масового обслуговування (рис. 4.6) - інтенсивностей заявок, що обробляються в кожному вузлі Λ при заданих зовнішніх потоках Γ і матриці маршрутизації потоків $P = \|p_{ij}\|$.

Початкові дані.

$$\Gamma = (16, 0, 16); \quad P = \|p_{ij}\| = \begin{vmatrix} 0 & 1/2 & 0 \\ 0 & 0 & 1/4 \\ 4/5 & 0 & 0 \end{vmatrix}$$

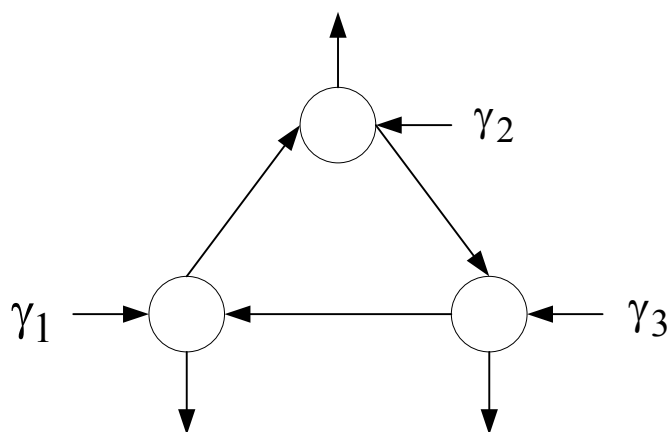


Рис. 4.6. Структура мережі СМО

Рішення задачі.

Система описується рівнянням $\Lambda = \Gamma + \Lambda P$. Для кожного вузла мережі використовуємо рівняння $\lambda_i = \gamma_i + \sum_{j=1}^3 \lambda_j \cdot p_{ji}$ і отримуємо систему рівнянь

$$\begin{cases} \lambda_1 = \gamma_1 + \lambda_3 \cdot \frac{4}{5} \\ \lambda_2 = \gamma_2 + \lambda_1 \cdot \frac{1}{2} \\ \lambda_3 = \gamma_3 + \lambda_2 \cdot \frac{1}{4} \end{cases} \quad \begin{matrix} 0,5 \\ 1 \end{matrix}$$

Вирішуючи її отримуємо три потоки $\Lambda = (32, 16, 20)$.

Зауважимо, що вектор Λ залежить від Γ і P , і не залежить від припущення про показове розподілення.

0,5

4.4. Розкладання часу мережевої затримки по каналах

Введемо позначення γ_{jq} - трафік (traffic), що надходить в мережу із зовнішніх джерел для тих повідомлень, які виникають у вузлі j і призначені для вузла q . Повний зовнішній трафік, що надходить в мережу, визначимо як:

$$\gamma = \sum_{j=1}^N \sum_{q=1}^N \gamma_{jq},$$

де N - число вузлів.

Позначимо T^z - середня затримка повідомлення, T_{jq}^z - затримка повідомлення, яке виникло в j і призначене для вузла q . Ясно, що ці дві середні величини пов'язані рівністю:

$$T^z = \sum_{j=1}^N \sum_{q=1}^N \frac{\gamma_{jq}}{\gamma} \cdot T_{jq}^z$$

так як частка $(\frac{\gamma_{jq}}{\gamma})$ повного вхідного трафіку повідомлень має в середньому затримку рівну T_{jq}^z . Відзначимо, що остання рівність представляє розкладання мережі по парам джерело-адресат. Позначимо через π_{jq} шлях, по якому йдуть повідомлення, що виникають у вузлі j і мають в якості вузла призначення вузол q . Кажуть, що i -ий канал (з пропускнуою здатністю C_i) включений в шлях π_{jq} , якщо повідомлення, що йдуть цим шляхом, проходять вказаний канал ($C_i \in \pi_{jq}$). Тоді середня кількість повідомлень в одиницю часу, що проходять по i -му каналу зв'язку, дорівнює:

$$\lambda_i = \sum_j \sum_q \gamma_{jq},$$

де для усіх $j, q: C_i \in \pi_{jq}$.

Крім того, зауважимо що T_{jq}^z - сума (T_i^z) середніх затримок повідомлення при передачі його по різних каналах шляху π_{jq} .

$$T_{jq}^z = \sum_{i: C_i \in \pi_{jq}} T_i^z.$$

Отже:

$$T^z = \sum_{j=1}^N \sum_{q=1}^N \frac{\gamma_{jq}}{\gamma} \sum_{i: C_i \in \pi_{jq}} T_i^z.$$

Змінимо тепер порядок підсумовування (умову на i змінимо на пару jq)

$$T^z = \sum_{i=1}^{\hat{N}} \frac{T_i^z}{\gamma} \sum_j \sum_q \gamma_{jq} = \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\gamma} T_i^z,$$

де $j, q: C_i \in \pi_{jq}$; $\hat{N}$ - кількість каналів в мережі, в якій N вузлів.

Тепер середня затримка являє розкладання по окремих каналах.

Скористаємося відомими результатами досліджень Джексона. Канал мережі можна розглядати як такий же канал, діючий незалежно від мережі, але з пуасонівським вхідним потоком, інтенсивність якого дорівнює інтенсивності, що задається мережею.

$$T_i^z = \frac{1}{\mu \cdot C_i - \lambda_i}$$

$$T^z = \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\gamma} \cdot \left[\frac{1}{\mu C_i - \lambda_i} \right]$$

і $1/\mu$ - середня довжина повідомлення, $1/\mu C_i$ - середній час передачі повідомлення по каналу C_i .

Можна провести розрахунок для випадку непоказових довжин повідомлень. Але при цьому буде проявлена дуже велика сміливість, тому що для систем M/G/1 результат Джексона непридатний. Однак це наближення призводить до моделі, яка дуже добре узгоджується з моделюванням на ЕОМ.

Приклад 4.2. Дана структура мережі з трьох вузлів (рис. 4.7) с сумарним зовнішнім трафіком

$$\gamma = \gamma_1 + \gamma_3 = 16 + 16 = 32 \text{ пак/с}$$

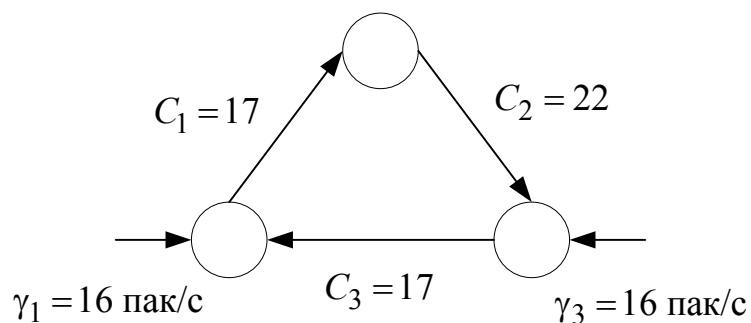


Рис. 4.7. Структура мережі

Матриця розподілення трафіку по вузлах мережі

$$\gamma_{ij} = \begin{vmatrix} 0 & 12 & 0 \\ 0 & 0 & 4 \\ 16 & 0 & 0 \end{vmatrix}$$

Пропускна здатність каналів

$$C_1 = C_{1-2} = 17 \text{ кбіт/с}$$

$$C_2 = C_{2-3} = 22 \text{ кбіт/с}$$

$$C_3 = C_{3-1} = 17 \text{ кбіт/с}$$

Сумарна попускна здатність каналів $C = 56 \text{ кбіт/с}$

В розрахунках, для спрощення, приймаємо $\mu = 1 \text{ пак/кбіт}$

Топологію мережі можна описати матрицею

$$st = \begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix}$$

Відмінний від нуля елемент вказує, що існує шлях довжиною в 1 канал з вузла i у вузол j .

$$st^2 = \begin{vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix}$$

Відмінний від 0 елемент квадрата цієї матриці вказує число шляхів топологічною довжиною в 2 одиниці з вузла i у вузол j

Аналізуючи обидві матриці (st і st^2) можна сформуувати таблицю характеристик каналів:

Марш-рути	Канали		
	I	II	III
1-2	12		
1-3	4	4	
3-1			16
Загалом λ_i	16	4	16

$$\begin{aligned}\lambda_1 &= 16 \text{ пак/с,} \\ \lambda_2 &= 4 \text{ пак/с,} \\ \lambda_3 &= 16 \text{ пак/с.}\end{aligned}$$

Тоді середня затримка повідомлення в мережі

$$\begin{aligned}T^z &= \frac{1}{\gamma} \cdot \sum_{i=1}^3 \frac{\lambda_i}{\mu \cdot C_i - \lambda_i} = \frac{1}{32} \left(\frac{16}{1 \cdot 17 - 16} + \frac{4}{1 \cdot 22 - 4} + \frac{16}{1 \cdot 17 - 16} \right) = \\ &= 1,007 \frac{\text{с}}{\text{пак}}\end{aligned}$$

Який вплив мають зовнішні потоки на середню затримку повідомлень в мережі представлено на рис. 4.8.

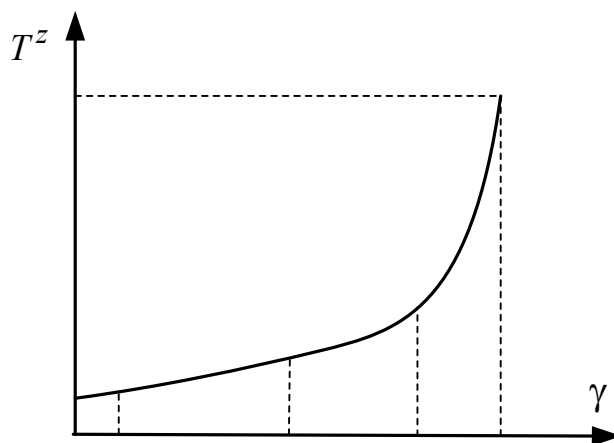


Рис. 4.8. Вплив зовнішніх потоків на середню затримку повідомлень

4.5. Двопараметрична модель часу затримки повідомлень

Визначимо n_{jq} - довжину шляху π_{jq} , як число каналів в цьому шляху, тоді середня довжина всіх шляхів в мережі

$$\bar{n} = \sum_{j=1}^N \sum_{q=1}^N \frac{\gamma_{jq}}{\gamma} \cdot n_{jq}.$$

Розглянемо тепер повний трафік у мережі. Вклад трафіку (jq) в повному трафіку дорівнює $\gamma_{jq} \cdot n_{jq}$, бо γ_{jq} повідомлень в секунду пройдуть n_{jq} ділянок при русі по мережі. Отже

$$\lambda = \sum_{j=1}^N \sum_{q=1}^N \gamma_{jq} \cdot n_{jq}.$$

Середня довжина шляху виражається у вигляді:

$$\bar{n} = \frac{\lambda}{\gamma}.$$

Цей результат в 1964 р. отримав Клейнрок (Kleinrock) [3,8]. Виражаючи середню затримку повідомлення T^z через середню довжину шляху, одержимо

$$T^z = \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\gamma} \cdot \left(\frac{1}{\mu \cdot C_i - \lambda_i} \right) = \bar{n} \cdot \sum_{i=1}^{\hat{N}} \frac{\lambda_i / \lambda}{\mu \cdot C_i - \lambda_i}.$$

Відношення $p_i = \lambda_i / \lambda$ - ймовірність використання i -го каналу не залежить від навантаження мережі і є лише функцією процедури вибору маршрутів. Відтак, в слабо навантаженій мережі ($\gamma \rightarrow 0$) затримка повідомлення обчислюється за формулою

$$T_0^z = \bar{n} \cdot \sum_{i=1}^{\hat{N}} p_i \cdot \frac{1}{\mu \cdot C_i}, \quad \text{бо } \lambda_i \rightarrow 0.$$

Навантаження насиченої мережі $\gamma \rightarrow \gamma_{\text{кр}}$ можна знайти наступним чином [9].

1. Для деякого значення навантаження γ визначити найбільше відношення

$$\max_{1 \leq i \leq \hat{N}} \frac{\lambda_i}{\mu \cdot C_i}.$$

2. Найменше значення навантаження, при якому насичується цей канал $\gamma_{\text{кр}}$

$$\gamma_{\text{кр}} = \frac{\gamma}{\max_{1 \leq i \leq \hat{N}} \frac{\lambda_i}{\mu \cdot C_i}}.$$

Ця двопараметрична спрощена порогова модель описує поведінку мережі в першому наближенні. Продовжимо розгляд попереднього **прикладу 4.2** (рис. 4.7).

Мережа має наступний розподіл потоків повідомлень

$$\gamma_{ij} = \begin{vmatrix} 0 & 12 & 0 \\ 0 & 0 & 4 \\ 16 & 0 & 0 \end{vmatrix} \Rightarrow \gamma = 32$$

Потоки по каналах, як було вираховано,

$$\lambda_1 = 16; \lambda_2 = 4; \lambda_3 = 16 \Rightarrow \lambda = 36$$

Тоді $\bar{n} = \frac{\lambda}{\gamma} = \frac{36}{32} = \frac{9}{8}$. Далі знаходимо p_i та $\frac{1}{\mu C_i}$

$$p_1 = \frac{\lambda_1}{\lambda} = \frac{16}{36} = \frac{4}{9}; \quad \frac{1}{\mu \cdot C_1} = \frac{1}{17};$$

$$p_2 = \frac{\lambda_2}{\lambda} = \frac{4}{36} = \frac{1}{9}; \quad \frac{1}{\mu \cdot C_2} = \frac{1}{22};$$

$$p_3 = \frac{\lambda_3}{\lambda} = \frac{16}{36} = \frac{4}{9}; \quad \frac{1}{\mu \cdot C_3} = \frac{1}{17};$$

$$T_0^z = \frac{9}{8} \left[\frac{4}{9} \cdot \frac{1}{17} + \frac{1}{9} \cdot \frac{1}{22} + \frac{4}{9} \cdot \frac{1}{17} \right] = 0,065 \text{ с},$$

$$\gamma_{\text{кр}} = \frac{32}{\max \left[\frac{16}{17}, \frac{21}{22}, \frac{16}{17} \right]} = \frac{32 \cdot 17}{16} = 34 \frac{\text{пак}}{\text{с}}.$$

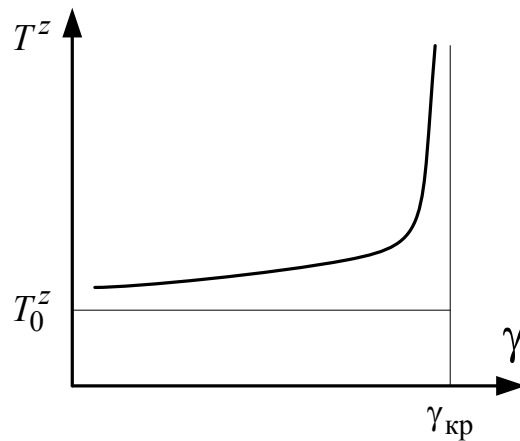


Рис. 4.9. Критичне значення зовнішніх потоків

Однак у будь-якій реальній мережі ЕОМ процедура вибору маршрутів залежить від поточного трафіку (тобто є адаптивною). В результаті, якщо який-небудь канал досягає насичення, повідомлення надсилається по обхідним каналам. Тому, при аналізі мереж, дуже важливою виявляється теорема Форда-Фолкерсона про мінімальний перетин і максимальний потік, який мережа може переносити між даною парою вузлів [10].

4.6. Задача вибору пропускних здатностей каналів

Припустимо, що задані потоки $\{\lambda_i\}$ і топологія мережі. Необхідно мінімізувати час перебування заявки в мережі, що задовольняє обмеженню

$$D = \sum_{i=1}^{\hat{N}} d_i \cdot C_i.$$

Для вирішення завдання будемо варіювати пропускними здатностями каналів $\{C_i\}$, d_i – вартість пропускної здатності для i -ого каналу, λ_i / μ – середня кількість біт в секунду, які проходять по i -ому каналу, і, отже, будь-яке вирішення завдання повинно бути таким, щоб i -тий канал мав пропускну здатність, не менше зазначеної величини, тобто $C_i > \lambda_i / \mu$.

Для мінімізації T – часу перебування вимоги в мережі складемо функцію Лагранжа:

$$\begin{aligned} G &= T + \beta \cdot \left[\sum_{i=1}^{\hat{N}} d_i \cdot C_i - D \right] = \\ &= \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\gamma} \cdot \left[\frac{1}{\mu C_i - \lambda_i} \right] + \beta \cdot \left[\sum_{i=1}^{\hat{N}} d_i \cdot C_i - D \right] \end{aligned}$$

Як завжди, використовуючи метод Лагранжа, отримуємо наступну систему з $\hat{N}$ рівнянь ($i \in \{1 \dots \hat{N}\}$)

$$\frac{\partial G}{\partial C_i} = \left(-\frac{\lambda_i}{\gamma} \right) \cdot \frac{\mu}{(\mu \cdot C_i - \lambda_i)^2} + \beta \cdot d_i = 0;$$

Вирішуючи відносно C_i , отримаємо

$$C_i = \frac{\lambda_i}{\mu} + \frac{\sqrt{\lambda_i/d_i}}{\sqrt{\beta \cdot \gamma \cdot \mu}} \quad i \in \{1 \dots \hat{N}\}.$$

Визначимо β , склавши обмеження шляхом перемноження на d_i і підсумувавши по i :

$$\sum_{i=1}^{\hat{N}} d_i \cdot C_i = D = \sum_{i=1}^{\hat{N}} \frac{\lambda_i \cdot d_i}{\mu} + \frac{\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i}}{\sqrt{\beta \cdot \gamma \cdot \mu}}.$$

Отже

$$1/\sqrt{\beta \cdot \gamma \cdot \mu} = \frac{D - \sum_{i=1}^{\hat{N}} \frac{\lambda_i \cdot d_i}{\mu}}{\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i}} = \frac{D_e}{\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i}}.$$

де D_e – додаткова вартість.

$$C_i = \frac{\lambda_i}{\mu} + \frac{D_e}{d_i} \frac{\sqrt{\lambda_i \cdot d_i}}{\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i}}.$$

При такому наборі пропускних здатностей кожен канал буде мати, принаймні, пропускну здатність λ_i/μ і, крім того, деяку додаткову пропускну здатність.

Мінімальна середня затримка мережі, пропускні здатності якої обрані оптимально, може бути записана у вигляді:

$$\begin{aligned}
T^z &= \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\gamma} \cdot \left(\frac{1}{\mu \cdot C_i - \lambda_i} \right) = \\
&= \frac{1}{\gamma} \cdot \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\mu \cdot \left(\frac{\lambda_i}{\mu} + \frac{D_e}{d_i} \frac{\sqrt{\lambda_i \cdot d_i}}{\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i}} \right) - \lambda_i} = \\
&= \frac{1}{\gamma} \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\left(\lambda_i + \frac{\mu \cdot D_e}{d_i} \frac{\sqrt{\lambda_i \cdot d_i}}{\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i}} \right) - \lambda_i} = \\
&= \frac{1}{\gamma \cdot \mu \cdot D_e} \sum_{i=1}^{\hat{N}} \frac{\lambda_i \cdot d_i}{\sqrt{\lambda_i \cdot d_i}} \sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i} = \\
&= \frac{\left(\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i \cdot d_i} \right)^2}{\gamma \cdot \mu \cdot D_e}, \quad \text{бо} \quad \sum_{i=1}^{\hat{N}} a_i \cdot \sum_{i=1}^{\hat{N}} a_i = \left(\sum_{i=1}^{\hat{N}} a_i \right)^2
\end{aligned}$$

На рис. 4.10 наведено якісний графік відповідної залежності.

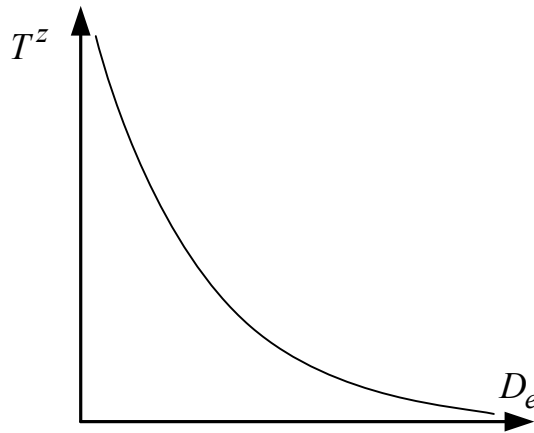


Рис. 4.10. Мінімальна середня затримка в мережі

Повернемося до розглянутого раніше **прикладу 4.2** (рис. 4.7).

$$\gamma_{ij} = \begin{vmatrix} 0 & 12 & 0 \\ 0 & 0 & 4 \\ 16 & 0 & 0 \end{vmatrix};$$

$$\mu = 1 \frac{\text{пак}}{\text{кбіт}};$$

$$\gamma = 32 \frac{\text{пак}}{\text{с}};$$

$$\lambda_1 = 16 \frac{\text{пак}}{\text{с}}; \quad \lambda_2 = 4 \frac{\text{пак}}{\text{с}}; \quad \lambda_3 = 16 \frac{\text{пак}}{\text{с}}; \quad \lambda = 36 \frac{\text{пак}}{\text{с}};$$

$$C_1 > 16; \quad C_2 > 4; \quad C_3 > 16; \quad C > 36.$$

Якщо $d_i = 1$, що має місце при розгляданні каналів зв'язку геостаціонарних супутників, то в цьому разі:

$$D = \sum_{i=1}^{\hat{N}} C_i = C \frac{\text{кбіт}}{\text{с}};$$

$$D_e = C - \sum_{i=1}^{\hat{N}} \frac{\lambda_i}{\mu} = C_e,$$

тоді $T^z = \frac{\left(\sum_{i=1}^{\hat{N}} \sqrt{\lambda_i} \right)^2}{\gamma \cdot \mu \cdot C_e}$ (рис. 4.11).

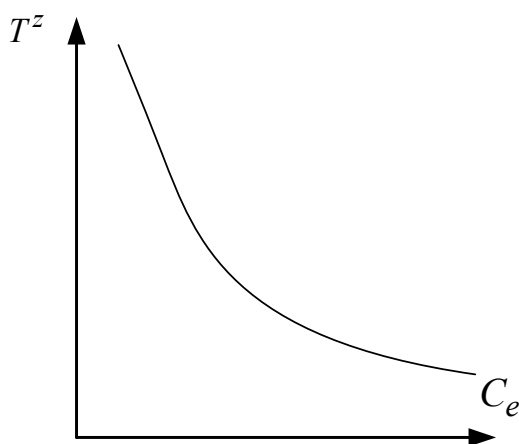


Рис. 4.11. Залежність $T^z = f(C_e)$

Якщо вимагати, що $T^z \leq 0,157$ с, то необхідно вибрати:

$$\tilde{N}_a = 20 \text{ і } C_e = 56.$$

Тоді:

$$C_1 = \frac{\lambda_1}{\mu} + C_e \frac{\sqrt{\lambda_1}}{\sum_{i=1}^{\tilde{N}} \sqrt{\lambda_i}} = 16 + 20 \frac{\sqrt{16}}{\sqrt{16} + \sqrt{4} + \sqrt{16}} = 24;$$

$$C_2 = 4 + 20 \frac{\sqrt{4}}{\sqrt{16} + \sqrt{4} + \sqrt{16}} = 8;$$

$$C_3 = 16 + 20 \frac{\sqrt{16}}{\sqrt{16} + \sqrt{4} + \sqrt{16}} = 24;$$

$$\begin{aligned} T_0 &= \bar{n} \cdot \sum_{i=1}^{\tilde{N}} p_i \frac{1}{\mu \cdot C_i} = \\ &= \frac{36}{32} \left(\frac{16}{36} \cdot \frac{1}{24} + \frac{4}{36} \cdot \frac{1}{8} + \frac{16}{36} \cdot \frac{1}{24} \right) = 0,0573; \\ \gamma_{кр} &= \gamma / \max \left(\frac{\lambda_i}{\mu \cdot C_i} \right) = \frac{32 \cdot 24}{16} = 48 \frac{\text{пак}}{\text{с}}; \end{aligned}$$

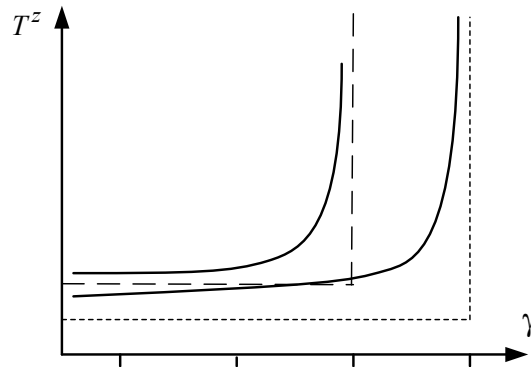


Рис. 4.11. Залежність часу затримки від $\gamma_{кр}$

Для розглянутого раніше прикладу T_0^z і $\gamma_{кр}$ мали значення:

$$T_0^z = 0,064 \text{ с}; \quad \gamma_{кр} = 34 \frac{\text{пак}}{\text{с}}.$$

ЕКСПОНЕНЦІЙНІ ММО З ОПЗ

5.1. Розімкнені експоненційні ММО з ОПЗ

5.1.1. Опис розімкнених ММО

Розглянемо розімкнену експоненційну мережу масового обслуговування (ММО) з однорідним потоком заявок при наступних припущеннях:

- 1) розімкнена ММО (РММО) довільної топології містить N вузлів;
- 2) після завершення обслуговування в будь-якому вузлі передача заявки в інший вузол відбувається миттєво;
- 3) окремі вузли можуть бути як одноканальні, так і багатоканальні СМО;
- 4) всі прилади багатоканального вузла є ідентичними, і будь-яка заявка може обслуговуватися будь-яким приладом;
- 5) заявка, що надійшла в багатоканальний вузол, коли всі або кілька приладів вільні, направляється випадковим чином в будь який вільний прилад;
- 6) у кожному вузлі РММО є накопичувач заявок необмеженої ємності, що означає відсутність відмов заявкам, що поступають в СМО, від постановки в чергу, тобто будь-яка заявка, яка надходить у вузол, завжди знайде в накопичувачі місце для очікування незалежно від того, скільки заявок вже знаходиться в черзі;
- 7) заявки надходять в РММО із зовнішнього незалежного джерела і утворюють найпростіший потік заявок;
- 8) тривалості обслуговування заявок у всіх вузлах мережі представляють собою випадкові величини, розподілені за експоненційним законом;

9) обслуговуючий прилад будь-якого вузла не простоює, якщо в його накопичувачі є хоча б одна заявка, причому після завершення обслуговування чергової заявки, наступна заявка миттєво вибирається з накопичувача;

10) в кожному вузлі мережі заявки з накопичувача вибираються відповідно до безпріоритетних дисциплін обслуговування в порядку надходження за правилом «першим прийшов - першим обслужений» (FIFO - First In First Out).

Для опису лінійних розімкнених однорідних експоненційних ММО необхідно задати таку сукупність параметрів:

- число вузлів в мережі: N ;
- число обслуговуючих приладів у вузлах мережі: $K_1, \dots, K_N$;
- матрицю ймовірностей передач: $P = [p_{ij} | i, j = 0, 1, \dots, N]$, де ймовірності передач p_{ij} повинні задовольняти умові :
сума елементів кожного рядка повинна дорівнювати 1

$$\sum_{j=0}^N p_{ij} = 1 \quad (i = \overline{0, N});$$

- інтенсивність λ_0 джерела заявок, що надходять в РММО;
- середні тривалості обслуговування заявок у вузлах мережі: $\tau_{1s}, \dots, \tau_{Ns}$.

На основі перерахованих параметрів можуть бути розраховані вузлові і мережеві характеристики, що описують ефективність функціонування відповідно вузлів і РММО в цілому.

Розрахунок характеристик функціонування лінійних розімкнених однорідних експоненційних ММО базується на еквівалентному перетворенні мережі і **проводиться в чотири етапи**:

- розрахунок коефіцієнтів передач α_j і інтенсивностей потоків заявок λ_j у вузлах $j = \overline{1, N}$ ММО;
- перевірка умови відсутності перевантажень в ММО;
- розрахунок вузлових характеристик;
- розрахунок мережевих характеристик.

5.1.2. Розрахунок коефіцієнтів передач і інтенсивностей потоків заявок у вузлах РММО

Покажемо, що інтенсивності $\lambda_0, \dots, \lambda_N$ потоків заявок, що надходять у вузли $0, \dots, N$ мережі, однозначно визначаються ймовірностями передач p_{ij} ($i, j = 1, \dots, N$), що задають маршрути заявок в ММО.

Будемо розглядати тільки сталий режим.

Так як в лінійній ММО заявки не розмножуються і не губляться, то інтенсивності вхідного і вихідного потоків для будь-якого вузла будуть рівні між собою.

Інтенсивність потоку заявок, що входять в будь-який вузол j мережі, дорівнює сумі інтенсивностей потоків заявок, що надходять до нього з інших вузлів $i = \overline{0, N}$ (Рис.5.1). Оскільки заявки з вузла i надходять у вузол j з ймовірністю p_{ij} , то інтенсивність потоку заявок, що надходять з i в j , дорівнює $p_{ij}\lambda_i$, де λ_i - інтенсивність потоку заявок, що входить, і, отже, виходить з вузла i . З урахуванням цього, на вході вузла j є потік з інтенсивністю

$$\lambda_j = \sum_{i=0}^N p_{ij}\lambda_i \quad (i = 0, 1, \dots, N) \quad (5.1)$$

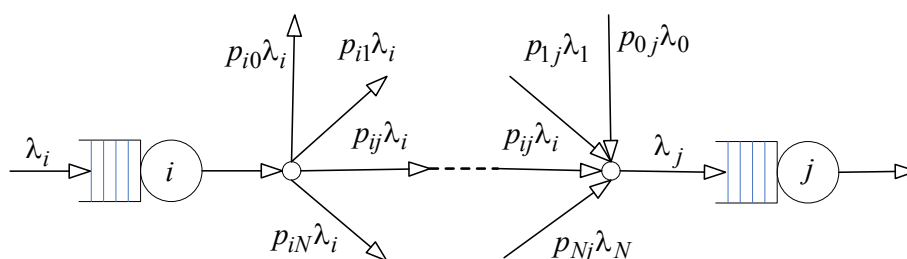


Рис. 5.1. До розрахунків інтенсивностей потоків заявок у вузлах РММО

Вираз (5.1) являє собою систему лінійних алгебраїчних рівнянь $(N + 1)$ -го порядку, з якої можуть бути знайдені інтенсивності потоків заявок у вигляді співвідношень

$$\lambda_j = \alpha_j \lambda_0 \quad (j = \overline{1, N}).$$

Коефіцієнт α_j називається коефіцієнтом передачі і визначає середнє число влучень заявки в вузол j за час її перебування в мережі, причому $\alpha_0 = 1$.

Для розімкненої ММО відома інтенсивність джерела заявок λ_0 . Можна показати, що система рівнянь для розрахунку інтенсивностей має єдине рішення виду $\lambda_j = \alpha_j \lambda_0$, де λ_0 - задана величина.

5.1.3. Перевірка умови відсутності перевантажень в ММО

В [6] показано, що в розімкненій ММО відсутні перевантаження, якщо виконується умова:

$$\lambda_0 < \min \left(\frac{K_1}{\alpha_1 \tau_{s1}}, \frac{K_2}{\alpha_2 \tau_{s2}}, \dots, \frac{K_N}{\alpha_N \tau_{sN}} \right) \quad (5.2)$$

Якщо зазначена умова не виконується, то, як впливає з нього, стаціонарний режим в розімкнутій ММО може бути реалізований одним із таких способів:

- зменшенням інтенсивності λ_0 зовнішнього джерела заявок до значення, при якому ця умова буде виконуватися;
- збільшенням кількості обслуговуючих приладів K_j в перевантажених вузлах;
- зменшенням тривалості τ_{sj} обслуговування заявок в перевантажених вузлах;
- зменшенням коефіцієнтів передачі α_j в перевантажених вузлах.

5.1.4. Розрахунок вузлових характеристик РММО

Один і той самий об'єкт, що розглядається на різних рівнях деталізації, можна уявити різними моделями масового обслуговування, характеристики яких однакові або відрізняються на величину, що не

перевищує заданої похибки. При виконанні певних умов такі моделі легко перетворюються одна в одну.

Для мережевих моделей у вигляді розімкнених і замкнутих ММО можуть використовуватися два види перетворень:

- еквівалентне перетворення;
- толерантне перетворення.

Дві мережеві моделі еквівалентні, якщо порівнювані характеристики цих моделей не відрізняються одна від одної.

Дві мережеві моделі толерантні (подібні), якщо значення певних характеристик відрізняються один від одного на величину, не переважаючу задану.

Використання властивостей еквівалентних і толерантних моделей дозволяє спростити розрахунок характеристик моделей шляхом заміни складних мережевих моделей більш простими.

Еквівалентними можуть бути мережеві моделі одного типу (наприклад, дві замкнуті мережі), толерантними - моделі як одного, так і різних типів [2].

Розрахунок характеристик функціонування лінійних розімкнених однорідних експоненційних ММО базується на еквівалентному перетворенні мережі, що полягає в поданні розімкненої ММО з N вузлами у вигляді N незалежних експоненційних СМО типу М/М/К (найпростіший потік заявок, тривалість обслуговування розподілена за експоненційним законом, K обслуговуючих приладів). При цьому інтенсивність вхідного потоку заявок в СМО, що відображає вузол j ($j = \overline{1, N}$) мережі, визначається із системи алгебраїчних рівнянь (5.1) через інтенсивність вхідного в мережу потоку і коефіцієнт передачі вузла: $\lambda_j = \alpha_j \lambda_0$, а середня тривалість обслуговування заявок в СМО дорівнює тривалості обслуговування τ_{sj} заявок у відповідному узлі ММО.

Характеристики всіх N СМО (час очікування заявок в черзі і перебування в системі, довжина черги і число заявок в системі, середнє число зайнятих приладів і т.д.) є вузлові характеристики ММО.

Середній час очікування заявок в черзі може бути розрахован з використанням виразу (2.10) для багатоканальних СМО типу М/М/К або виразе (2.1) для одноканальних СМО типу М/М/1, інші характеристики вузла j ($j = \overline{1, N}$) - з використанням фундаментальних співвідношень, а саме:

- навантаження у вузлі j , що показує середнє число зайнятих приладів: $\delta_j = \lambda_j \tau_{sj}$;
- завантаження вузла j : $\rho_j = \min(\delta_j / K_j; 1)$, де K_j - число обслуговуючих приладів у вузлі j ;
- коефіцієнт простою вузла: $\beta_j = 1 - \rho_j$;
- час перебування заявок у вузлі: $\tau_{sysj} = \tau_{wj} + \tau_{sj}$;
- довжина черги заявок: $l_j = \lambda_j \tau_{wj}$;
- число заявок у вузлі (в черзі і на обслуговуванні в приладі): $m_j = \lambda_j \tau_{sysj}$.

Розраховані таким чином характеристики окремих СМО в точності відповідають вузловим характеристикам вихідної ММО. Таким чином, щодо своїх характеристик, модель масового обслуговування мережі представляє собою сукупність незалежних СМО (кожна СМО розглядається незалежно від інших) і строго еквівалентна вихідній розімкненій ММО в цілому.

5.1.5. Розрахунок мережевих характеристик РММО

Мережеві характеристики, що описують ефективність функціонування ММО в цілому, розраховуються на основі отриманих значень вузлових характеристик.

До складу мережевих характеристик входять:

- середня кількість заявок, які очікують обслуговування в мережі, і середнє число заявок, що знаходяться в мережі:

$$L = \sum_{j=1}^N l_j; \quad M = \sum_{j=1}^N m_j,$$

де l_j - середня довжина черги і m_j - середнє число заявок у вузлі j ;

- середній час очікування і середній час перебування заявок в мережі:

$$\tau_{netw} = \sum_{j=1}^N \alpha_j \tau_{wj}; \quad \tau_{net} = \sum_{j=1}^N \alpha_j \tau_{sysj}$$

де τ_{wj} і τ_{sysj} - відповідно середній час очікування і середній час перебування заявок у вузлі j ; α_j - коефіцієнт передачі для вузла j , показує середнє число влучень заявки в вузол j за час її знаходження в мережі.

Приклад 5.1. Проілюструємо викладений метод розрахунку характеристик функціонування лінійних розімкнених однорідних експоненційних ММО на прикладі ММО з чотирма вузлами ($N = 4$), граф якої представлений на рис. 5.2.

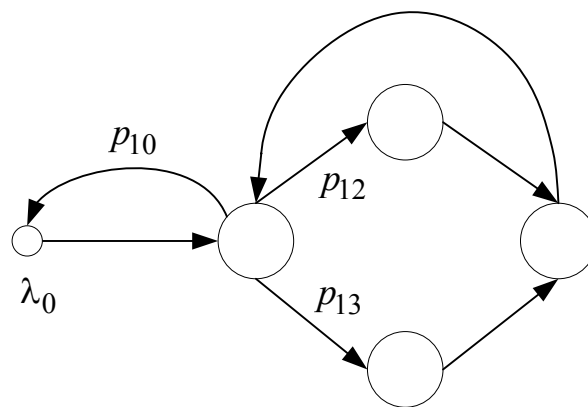


Рис. 5.2. Граф розімкненої ММО

Зв'язки між вузлами ММО описуються наступної матрицею ймовірностей передач:

$$P = \begin{array}{c|ccccc} & 0 & 1 & 2 & 3 & 4 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0,1 & 0 & 0,2 & 0,7 & 0 \\ 2 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 0 & 0 & 1 \\ 4 & 0 & 1 & 0 & 0 & 0 \end{array}$$

В РММО надходить найпростіший потік заявок з інтенсивністю $\lambda_0 = 0,1 \text{ с}^{-1}$. Покладемо, що всі вузли ММО - одноканальні, а середні тривалості обслуговування заявок у вузлах відповідно рівні:

$$\tau_{s1} = 0,8 \text{ с}; \tau_{s2} = 2 \text{ с}; \tau_{s3} = 0,4 \text{ с}; \tau_{s4} = 0,3 \text{ с}$$

Система лінійних алгебраїчних рівнянь для розрахунку інтенсивностей потоків заявок у вузлах ММО, згідно (5.1), має вигляд:

$$\begin{cases} \lambda_0 = p_{10}\lambda_1 = 0,1\lambda_1 \\ \lambda_1 = p_{01}\lambda_0 + p_{41}\lambda_4 = \lambda_0 + \lambda_4 \\ \lambda_2 = p_{12}\lambda_1 = 0,2\lambda_1 \\ \lambda_3 = p_{13}\lambda_1 = 0,7\lambda_1 \\ \lambda_4 = p_{24}\lambda_2 + p_{34}\lambda_3 = \lambda_2 + \lambda_3 \end{cases}$$

Вирішуючи цю систему рівнянь, одержимо наступні значення інтенсивностей:

$$\lambda_1 = 1 \text{ с}^{-1}, \lambda_2 = 0,2 \text{ с}^{-1}, \lambda_3 = 0,7 \text{ с}^{-1}, \lambda_4 = 0,9 \text{ с}^{-1}$$

Тоді коефіцієнти передач будуть рівні:

$$\alpha_1 = \lambda_1 / \lambda_0 = 10; \alpha_2 = \lambda_2 / \lambda_0 = 2; \alpha_3 = \lambda_3 / \lambda_0 = 7; \alpha_4 = \lambda_4 / \lambda_0 = 9$$

Визначимо граничну інтенсивність надходження заявок в разомкнуту ММО, при якій в мережі відсутні перевантаження. Для

цього скористаємося виразом (5.2), що визначає умову відсутності перевантажень в РММО:

$$\lambda_0 < \min\left(\frac{K_1}{\alpha_1 \tau_{s1}}, \frac{K_2}{\alpha_2 \tau_{s2}}, \frac{K_3}{\alpha_3 \tau_{s3}}, \frac{K_4}{\alpha_4 \tau_{s4}}\right) = 0,125 \text{ c}^{-1}$$

РММО працює без перевантажень, оскільки дана умова виконується. Відповідно еквівалентними перетвореннями уявімо розглянуту експоненційну розімкнену ММО у вигляді 4-х незалежних СМО типу М/М/1, в які надходять найпростіші потоки заявок відповідно з інтенсивностями:

$$\lambda_1 = 1 \text{ c}^{-1}, \lambda_2 = 0,2 \text{ c}^{-1}, \lambda_3 = 0,7 \text{ c}^{-1}, \lambda_4 = 0,9 \text{ c}^{-1},$$

а середні тривалості обслуговування заявок в СМО збігаються з тривалостями обслуговування у відповідних вузлах ММО:

$$\tau_{s1} = 0,8 \text{ c}; \tau_{s2} = 2 \text{ c}; \tau_{s3} = 0,4 \text{ c}; \tau_{s4} = 0,3 \text{ c}.$$

Значення вузлових характеристик ММО, розраховані з використанням виразу (2.1) для середнього часу очікування заявок в черзі СМО типу М/М/1 і фундаментальних співвідношень, представлених в п.2.1.2, наведені в табл.5.1.

Таблиця 5.1

Вузлові характеристики ММО

Вузлові характеристики	Розрахункові формули	Вузол 1	Вузол 2	Вузол 3	Вузол 4
Навантаження	$\delta_j = \lambda_j \tau_{sj}$	0,8	0,4	0,28	0,27
Завантаження	$\rho_j = \min(\delta_j / K_j; 1)$	0,8	0,4	0,28	0,27
Коеф-т простоя	$\beta_j = 1 - \rho_j$	0,2	0,6	0,72	0,73
Час очікування	$\tau_{wj} = \rho_j \tau_{sj} / (1 - \rho_j)$	3,2	1,33	0,16	0,11
Час перебування	$\tau_{sysj} = \tau_{wj} + \tau_{sj}$	4	3,33	0,56	0,41

Длина черги	$l_j = \lambda_j \tau_{wj}$	3,2	0,27	0,11	0,1
Число заявок у вузлі	$m_j = \lambda_j \tau_{sysj}$	4	0,67	0,39	0,37

У табл.5.2 представлено математичні залежності та отримані на їх основі значення мережових характеристик, розраховані з урахуванням знайдених значень вузлових характеристик.

Таблиця 5.2

Математичні моделі та результати

Мережеві характеристики	Розрахункові формули	Значення
Час очікування в мережі	$\tau_{netw} = \sum_{j=1}^N \alpha_j \tau_{wj}$	36,75
Час перебування в мережі	$\tau_{net} = \sum_{j=1}^N \alpha_j \tau_{sysj}$	54,25
Число заявок в стані очікування	$L = \sum_{j=1}^N l_j$	3,68
Число заявок в мережі	$M = \sum_{j=1}^N m_j$	5,43

5.1.6. Аналіз властивостей розімкнених ММО

Властивості розімкнених ММО визначаються значеннями вузлових і мережових характеристик, пов'язаних між собою залежностями, представленими в цьому розділі. Найбільший інтерес представляють властивості мережі в цілому, оскільки властивості окремих вузлів ММО аналогічні властивостям відповідних одноканальних та багатоканальних СМО.

На рис. 5.3 показана залежність основної мережової характеристики РММО - середнього часу перебування τ_{net} заявок в мережі від інтенсивності λ_0 надходження заявок в мережу.

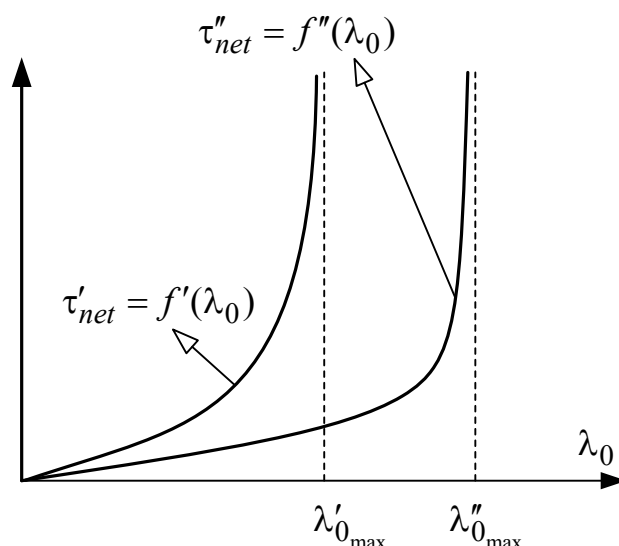


Рис. 5.3. Час перебування заявок в РММО

Залежність $\tau'_{net} = f'(\lambda_0)$ аналогічна залежності середнього часу перебування заявок в СМО від завантаження системи, зміна якого може бути обумовлена, зокрема, зміною інтенсивності надходження заявок в СМО. Як і в СМО, є деяке граничне значення інтенсивності λ'_{0max} , при якому середній час перебування заявок в мережі стає нескінченно великим, що свідчить про перевантаження в ММО. Вище показано (5.2), що в РММО відсутні перевантаження, якщо вони відсутні у всіх вузлах мережі, тобто перевантаження в розімкненій ММО наступає в тому випадку, коли завантаження одного з вузлів мережі стає рівним одиниці. Такий вузол називається «вузьким місцем» і характеризується тим, що черга заявок перед ним з часом зростає до нескінченності і, як наслідок, стає нескінченним число заявок в розімкненій ММО.

Для того, щоб позбутися в РММО від перевантаження, необхідно розвантажити «вузьке місце». Це може бути досягнуто наступними способами:

- збільшенням швидкості роботи (швидкодії) обслуговуючого приладу;
- збільшенням числа обслуговуючих приладів у вузлі.

Будь-який з цих способів дозволяє збільшити продуктивність ММО в цілому і, як наслідок, поліпшити характеристики мережі. За-

лежність середнього часу перебування τ_{net} заявок в мережі від інтенсивності λ_0 надходження заявок в мережу приймає вигляд $\tau''_{net} = f''(\lambda_0)$, тобто час перебування заявок при одній і тій же інтенсивності λ_0 стає менше (оскільки мережа має велику продуктивність), а граничне значення інтенсивності $\lambda''_{0_{max}}$, при якому настає перевантаження ММО, стає більше: $\lambda''_{0_{max}} > \lambda'_{0_{max}}$. При цьому з'являється нове вузьке місце в ММО, і подальше поліпшення мережі може бути досягнуто шляхом розвантаження нового вузького місця. Очевидно, що якщо ММО є моделлю реальної технічної системи, розвантаження вузького місця за рахунок збільшення швидкості роботи обслуговуючого приладу або числа приладів у вузлі означає збільшення вартості реальної системи.

Існує ще один спосіб розвантаження вузького місця ММО, який полягає в зменшенні ймовірності передачі заявок до вузла, що є вузьким місцем. Цей спосіб часто використовується в реальних системах і зазвичай не пов'язаний зі збільшенням вартості системи. Наприклад, в обчислювальній системі зміна ймовірностей передачі до накопичувачів зовнішньої пам'яті може бути досягнуто за рахунок перерозподілу файлів між накопичувачами: найбільш часто використовувані файли, розташовані у найбільш завантаженому накопичувачі, переносяться в найменш завантажений накопичувач. При цьому зменшується кількість звернень до завантаженого накопичувача (коефіцієнт передачі відповідного вузла ММО). Характер залежностей інших мережових характеристик (часу очікування, числа заявок в мережі і в стані очікування) розімкненої ММО від інтенсивності надходження заявок аналогічний показаному на рис. 5.3.

Приклад 5.2. Проілюструємо способи розвантаження вузького місця і одержуваний від цього ефект для чотирьох вузлової розімкненої ММО, що розглядалась в **прикладі 5.1**. Там же було показано, що інтенсивність надходження заявок до розімкненої ММО, при якій в мережі відсутні перевантаження, повинна задовольняти умові:

$$\lambda_0 < 0,125 \text{ с}^{-1}.$$

1. Розрахуємо спочатку характеристики РММО, що працює в області завантажень, близьких до 1, для чого покладемо, що інтенсив-

ність потоку заявок, що надходять у мережу, дорівнює $\lambda_0 < 0,12 \text{ с}^{-1}$. Тоді інтенсивності потоків заявок у вузлах РММО відповідно будуть рівні: $\lambda_1 = \alpha_1 \lambda_0 = 1,2 \text{ с}^{-1}$, $\lambda_2 = \alpha_2 \lambda_0 = 0,24 \text{ с}^{-1}$, $\lambda_3 = \alpha_3 \lambda_0 = 0,84 \text{ с}^{-1}$, $\lambda_4 = \alpha_4 \lambda_0 = 1,08 \text{ с}^{-1}$, а середні тривалості обслуговування заявок, як і раніше, будуть рівні: $\tau_{s1} = 0,8 \text{ с}$; $\tau_{s2} = 2 \text{ с}$; $\tau_{s3} = 0,4 \text{ с}$; $\tau_{s4} = 0,3 \text{ с}$.

Розраховані значення вузлових та мережевих характеристик ММО приведені в табл. 5.3.

Таблиця 5.3

Вузлові та мережеві характеристики ММО

Характеристики	Вузол 1	Вузол 2	Вузол 3	Вузол 4	ММО
Навантаження	0,96	0,48	0,336	0,324	2,10
Завантаження	0,96	0,48	0,336	0,324	2,10
Час очікування	19,2	1,85	0,202	0,144	198,4
Час перебування	20	3,85	0,602	0,144	215,9
Довжина черги	23,04	0,44	0,170	0,155	23,8
Число заявок	24	0,92	0,506	0,479	25,9

Аналіз представлених результатів показує, що збільшення інтенсивності надходження заявок в РММО всього лише на 20% до значення $\lambda_0 < 0,12 \text{ с}^{-1}$, призвело до різкого зростання значень мережевих характеристик. Зокрема, середній час перебування заявок в мережі зріс в 4 рази, а число заявок, що знаходяться в чергах - майже в 6,5 разів. Це говорить про те, що ММО працює в області великих завантажень, де незначне збільшення навантаження призводить до істотної зміни характеристик обслуговування заявок. Найбільш завантаженим вузлом ММО, тобто вузьким місцем, є вузол 1, завантаження якого багато більше завантажень інших вузлів і становить $\rho_1 = 0,96$. Саме в цьому вузлі характеристики обслуговування заявок виросли найбільш істотно: середній час перебування заявок в 5 разів (з 4 до 20 секунд), а середня довжина черги - більш ніж в 7 разів (з 3,2 до 23 заявок).

2. Для поліпшення характеристик обслуговування заявок в РММО необхідно розвантажити вузьке місце мережі, яким є вузол 1. Для цього збільшимо швидкість роботи обслуговуючого приладу в 2 рази,

що, в кінцевому рахунку, призведе до зменшення тривалості обслуговування заявок в 2 рази, яка стане рівною $\tau_{s1} = 0,4$ с.

Розраховані значення вузлових та мережних характеристик ММО після розвантаження вузького місця приведені в табл.5.4.

Таблиця 5.4

Вузлові та мережеві характеристики ММО після розвантаження

Характеристики	Вузол 1	Вузол 2	Вузол 3	Вузол 4	ММО
Навантаження	0,48	0,48	0,336	0,324	1,62
Завантаження	0,48	0,48	0,336	0,324	1,62
Час очікування	0,369	1,846	0,202	0,144	10,1
Час перебування	0,769	3,846	0,602	0,144	23,6
Довжина черги	0,443	0,443	0,170	0,155	1,21
Число заявок	0,923	0,923	0,506	0,479	2,83

Аналіз представлених результатів показує, що розвантаження вузького місця дозволило суттєво зменшити значення мережних характеристик: середній час перебування заявок в мережі зменшився більше ніж у 9 разів, а число заявок, що знаходяться в чергах - майже в 20 разів. Зазначимо, що зміна тривалості обслуговування заявок у вузлі 1 привело до зміни вузлових характеристик тільки цього вузла; вузлові характеристики інших вузлів не змінилися. Це є наслідком незалежного функціонування вузлів експоненційної розімкненої ММО, що фактично і дозволяє використовувати метод розрахунку характеристик мережі, заснований на декомпозиції, тобто такого типу мережі у вигляді сукупності незалежних СМО.

3. Для порівняння виконаємо розвантаження вузького місця іншим способом, а саме: збільшимо число обслуговуючих приладів у вузлі 1 з одного до двох: $K_1 = 2$, зберігши колишнє значення тривалості обслуговування одним приладом: $\tau_{s1} = 0,8$ с.

Розраховані значення вузлових та мережних характеристик ММО після розвантаження вузького місця приведені в табл. 5.5.

Таблиця 5.5

Вузлові та мережеві характеристики ММО після розвантаження вузького місця

<i>Характеристики</i>	<i>Вузол 1</i>	<i>Вузол 2</i>	<i>Вузол 3</i>	<i>Вузол 4</i>	<i>ММО</i>
Навантаження	0,96	0,48	0,336	0,324	2,10
Завантаження	0,48	0,48	0,336	0,324	1,62
Час очікування	0,288	1,846	0,202	0,144	9,28
Час перебування	1,088	3,846	0,602	0,144	26,78
Довжина черги	0,346	0,443	0,170	0,155	1,11
Число заявок	1,306	0,923	0,506	0,479	3,21

Порівняємо отримані значення мережевих характеристик зі значеннями, представленими в табл. 5.4 для першого способу розвантаження вузького місця за рахунок зменшення тривалості обслуговування заявок. При другому способі розвантаження вузького місця за рахунок збільшення числа обслуговуючих приладів ($K_1 = 2$; $\tau_{s1} = 0,8$ с) середній час очікування заявок в мережі дещо зменшився в порівнянні з першим способом ($K_1 = 1$; $\tau_{s1} = 0,4$ с). У той же час, середній час перебування заявок в РММО збільшився більш ніж на 10%, що зумовлено більшою тривалістю обслуговування заявок ($\tau_{s1} = 0,8$ с) у кожному з приладів двоканального вузла 1 порівняно з одноканальним вузлом при першому способі ($\tau_{s1} = 0,4$ с). Як і в попередньому випадку, зміна числа обслуговуючих приладів у вузлі 1 привела до зміни вузлових характеристик тільки цього вузла.

5.2. Замкнуті експоненційні ММО з ОПЗ

5.2.1. Опис замкнутих ММО

Розглянемо замкнуту експоненційну мережу масового обслуговування з однорідним потоком заявок при наступних припущеннях:

- 1) замкнута ММО (ЗММО) довільної топології містить N вузлів;

2) після завершення обслуговування в будь-якому вузлі передача заявки в інший вузол відбувається миттєво;

3) всі вузли замкнутої ММО одноканальні;

4) у ММО циркулює постійне число заявок;

5) тривалості обслуговування заявок у всіх вузлах мережі являють собою випадкові величини, розподілені за експоненційним законом;

6) ємність накопичувача в кожному вузлі ММО достатня для зберігання всіх заявок, що циркулюють в мережі, що означає відсутність відмов заявкам, що надходять, при їх постановці у чергу будь-якого вузла (зокрема, можна вважати, що ємність накопичувача в кожному вузлі дорівнює числу заявок, що циркулюють в мережі);

7) обслуговуючий прилад будь-якого вузла не простоює, якщо в його накопичувачі є хоча б одна заявка, причому після завершення обслуговування чергової заявки миттєво з накопичувача вибирається наступна заявка;

8) у кожному вузлі мережі заявки з накопичувача вибираються відповідно до безпріоритетної дисципліни обслуговування в порядку надходження за правилом «першим прийшов - першим обслужений» (FIFO - First In First Out).

Для опису лінійних замкнутих однорідних експоненційних ММО необхідно задати таку ж сукупність параметрів, як і для розімкнутих ММО, з єдиною відмінністю, яка полягає в тому, що замість інтенсивності джерела заявок слід задати число заявок, циркулюючих в ЗММО. Таким чином, сукупність параметрів для замкнутих ММО буде мати наступний вигляд:

- число вузлів в мережі: N ;
- число обслуговуючих приладів у вузлах мережі: $K_1, \dots, K_N$;
- матриця ймовірностей передач: $P = [p_{ij} \mid i, j = 0, 1, \dots, N]$, де p_{ij} - ймовірність передачі заявки з вузла i у вузол j ;

- число заявок M , циркулюючих в ЗММО;
- середні тривалості обслуговування заявок у вузлах мережі:
 $\tau_{s1}, \dots, \tau_{sN}$.

На основі перерахованих параметрів можуть бути розраховані вузлові і мережеві характеристики, що описують ефективність функціонування відповідно вузлів і ЗММО в цілому.

Розрахунок характеристик функціонування лінійних замкнутих однорідних експоненційних ММО з одноканальними вузлами базується на так званій «теоремі про прибуття» і проводиться з використанням *методу середніх значень* в два етапи:

- розрахунок коефіцієнтів передач у вузлах замкнутої ММО;
- розрахунок характеристик ЗММО.

5.2.2. Розрахунок коефіцієнтів передач у вузлах ЗММО

Для замкнутої ММО на першому етапі розраховуються тільки коефіцієнти передач. Інтенсивності потоків заявок у вузлах ЗММО НЕ можуть бути розраховані, як в РММО, оскільки для ЗММО спочатку не відома інтенсивність λ_0 , яка є не параметром, що задається у складі вихідних даних, а характеристикою, що представляє собою продуктивність ЗММО і визначається в процесі аналізу ефективності функціонування ЗММО.

Для розрахунку коефіцієнтів передач $\alpha_1, \dots, \alpha_N$ після деяких перетворень можна скористатися тією ж системою лінійних алгебраїчних рівнянь (5.1). Для цього в лівій і правій частині виразу (5.1) представимо інтенсивності у вигляді $\lambda_j = \alpha_j \lambda_0$. Якщо розділити ліву і праву частин виразу (5.1) на λ_0 , остаточно отримаємо систему лінійних алгебраїчних рівнянь щодо $\alpha_1, \dots, \alpha_N$:

$$\alpha_j = \sum_{i=0}^N p_{ij} \alpha_i \quad (i = 0, 1, \dots, N). \quad (5.3)$$

Вважаючи $\alpha_0 = 1$, можна знайти коріння системи рівнянь, чисельно визначають значення $\alpha_1, \dots, \alpha_N$.

5.2.3. Розрахунок характеристик ЗММО

Характеристики ЗММО можуть бути розраховані з використанням марковських процесів, оскільки кількість станів марківського процесу, на відміну від РММО, не нескінченно, а дорівнює числу сполучень C_{M+N-1}^M , де N - число вузлів в ЗММО і M - число заявок, що циркулюють в ЗММО. При цьому основна складність полягає у визначенні ймовірностей станів мережі $P(M_1, \dots, M_N)$ в разі великої її розмірності ($N > 5$; $M > 5$), коли число станів виявляється значним. При виконанні розрахунків на ЕОМ це, у багатьох випадках, призводить до втрати значущості в процесі проміжних обчислень і, отже, до неможливості отримання кінцевих результатів.

Від зазначеного недоліку вільний метод середніх значень, дозволяє обчислювати середні характеристики функціонування експоненційних ММО на основі порівняно простих рекурентних співвідношень.

Покладемо, що замкнута однорідна ММО містить N одноканальних вузлів, тривалості обслуговування заявок в яких розподілені за експоненційним законом із середніми значеннями $\tau_{s1}, \dots, \tau_{sN}$ відповідно. Нехай для кожного вузла i мережі відомо середнє число влучень заявки в даний вузол за час її перебування в мережі, тобто коефіцієнт передачі α_i , який, якщо конфігурація мережі задана матрицею ймовірностей передач $P = [p_{ij} \mid i, j = 0, 1, \dots, N]$, визначається в результаті рішення системи лінійних алгебраїчних рівнянь (5.3).

Позначимо: τ_{sysi} - середній час перебування заявки у вузлі i за час перебування в мережі; m_i - середнє число заявок у вузлі i ($i = 1, \dots, N$); λ_0 - продуктивність замкнутої мережі. Очевидно, що ці величини залежать від числа заявок M , циркулюючих в замкнутій мережі, тобто $\tau_{sysi} = \tau_{sysi}(M)$; $\lambda_0 = \lambda_0(M)$.

Можна показати, що мають місце наступні співвідношення:

$$\tau_{sysi} = \tau_{si}[1 + m_i(M - 1)]; \quad (5.4)$$

$$\tau_{net}(M) = \sum_{i=1}^N \alpha_i \tau_{sysi}(M); \quad (5.5)$$

$$\lambda_0(M) = \frac{M}{\tau_{net}(M)}; \quad (5.6)$$

$$m_i(M) = \alpha_i \lambda_0(M) \tau_{sysi}(M), \quad (5.7)$$

де $\tau_{net}(M)$ - середній час перебування заявок в мережі за умови знаходження в ній M заявок; $m_i(0) = 0$.

Вираз (5.4) отримано на основі так званої теореми про прибуття [11], яка стверджує, що в замкнутій експоненційній мережі з одноканальними вузлами, в якій циркулюють M заявок, стаціонарна ймовірність стану будь-якого вузла в момент надходження в нього нової заявки збігається зі стаціонарною ймовірністю того ж стану розглянутого вузла в мережі, в якій циркулює на одну заявку менше, тобто $(M - 1)$ заявок. Це означає, що в мережі з M заявками середнє число заявок $m_i(M)$, що знаходяться у вузлі i в момент надходження в цей вузол нової заявки, дорівнює $m_i(M - 1)$. Тоді середній час перебування у вузлі i заявки, що надійшла, буде складатися із середнього часу обслуговування всіх $m_i(M - 1)$ заявок, які раніше надійшли і знаходяться у вузлі i , та середньої тривалості обслуговування даної заявки:

$$\tau_{sysi}(M) = \tau_{si} m_i(M - 1) + \tau_{si} = \tau_{si}[1 + m_i(M - 1)].$$

У цьому виразі враховано, що середній час дообслуговування заявки, що знаходиться в приладі на момент надходження даної заявки, дорівнює середній тривалості обслуговування τ_{si} в силу властивості відсутності післядії, притаманної експоненційному закону. Середній час перебування заявки у вузлі i за час її перебування в мережі, враховує число влучень α_i заявки в даний вузол, і дорівнює

$$\tau_{neti}(M) = \alpha_i \tau_{sysi}(M).$$

Вирази (5.5) і (5.6) являють собою формули Літтла для мережі, а вираз (5.7) - для вузла i , де $\lambda_i(M) = \alpha_i \lambda_0(M)$ - інтенсивність потоку заявок у вузол i ($i = 1, \dots, N$).

На основі рекурентних співвідношень (5.4) - (5.7) послідовно для $M = 1, 2, \dots, M^*$, де M^* - задане число заявок в замкнутій мережі, можуть бути розраховані середні значення характеристик замкнутої експоненційної ММО.

Зауважимо, що наведений метод розрахунку є точним для замкнутих експоненційних ММО з одноканальними вузлами.

Приклад 5.3. Розрахуємо характеристики замкнутої однорідної експоненційної ММО, отриманої шляхом перетворення розімкнутої ММО (рис. 5.2), розглянутої в Прикладі 5.1, в замкнуту. Покладемо, що «нульова точка», що відображає завершення обслуговування заявок в мережі та миттєве формування нової заявки. Заявка на дузі виходить з вузла 1 і входить знову в цей же вузол (рис.5.4).

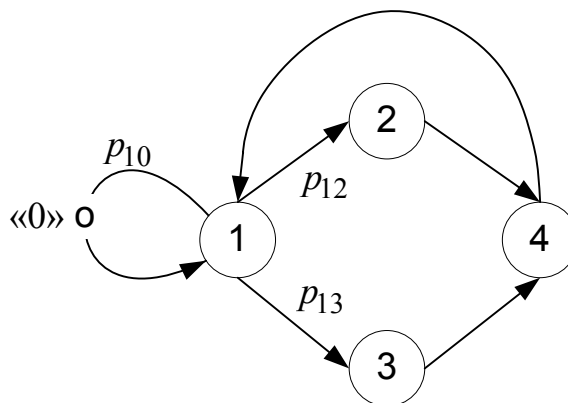


Рис. 5.4. Граф замкнутої ММО

Нагадаємо, що в ЗММО щодо «нульової точки» розраховуються тимчасові мережеві характеристики: час перебування в стані очікування і час перебування заявок в мережі, а також продуктивність ЗММО.

ЗММО містить $N = 4$ одноканальних вузла, зв'язки між якими описуються тією ж матрицею ймовірностей передач:

$$P = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{vmatrix} 0 & 1 & 0 & 0 & 0 \\ 0,1 & 0 & 0,2 & 0,7 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix} \end{matrix}$$

Отже, коефіцієнти передач для всіх вузлів, що розраховуються шляхом вирішення системи лінійних алгебраїчних рівнянь (5.3), будуть мати ті ж самі значення: $\alpha_1 = 10$; $\alpha_2 = 2$; $\alpha_3 = 7$; $\alpha_4 = 9$.

В ЗММО циркулює M заявок, середні тривалості обслуговування яких у вузлах є:

$$\tau_{s1} = 0,8 \text{ с}; \tau_{s2} = 2 \text{ с}; \tau_{s3} = 0,4 \text{ с}; \tau_{s4} = 0,3 \text{ с}.$$

Нижче в табл. 5.6 представлені значення часу перебування $\tau_{sysi}(M)$ і числа заявок $m_i(M)$ у вузлах мережі, а також середнього часу перебування $\tau_{net}(M)$ заявок в мережі і продуктивності $\lambda_0(M)$, розрахованих на основі виразів (5.4) - (5.7), для числа циркулюючих в мережі заявок $M = 1, 2, \dots, 6$. Коректність виконаних розрахунків підтверджується тим, що для всіх $M = 1, 2, \dots, 6$ виконується перевірка:

$$\sum_{i=1}^4 m_i(M) = M.$$

Таблиця 5.6

Характеристики ЗММО

M	i	$\tau_{sysi}(M)$	$\tau_{net}(M)$	$\lambda_0(M)$	$m_i(M)$
1	1	0,8	17,5	0,057	0,46
	2	2,0			0,23
	3	0,4			0,16
	4	0,3			0,15

2	1	1,17	22,94	0,087	1,02
	2	2,46			0,43
	3	0,46			0,28
	4	0,35			0,27
3	1	1,61	28,87	0,104	1,68
	2	2,87			0,59
	3	0,51			0,37
	4	0,38			0,36
4	1	2,14	35,29	0,113	2,43
	2	3,19			0,72
	3	0,55			0,44
	4	0,41			0,42
5	1	2,74	42,14	0,119	3,25
	2	3,45			0,82
	3	0,57			0,48
	4	0,42			0,45
6	1	3,40	49,35	0,122	4,14
	2	3,63			0,88
	3	0,59			0,50
	4	0,44			0,48

На рис. 5.5 представлені залежності продуктивності даної замкнутої ММО і середнього часу перебування заявок в мережі від кількості $M = \overline{1, 10}$ циркулюючих заявок. Аналіз отриманих результатів показує, що всі характеристики, включаючи продуктивність λ_0 , ростуть із збільшенням M .

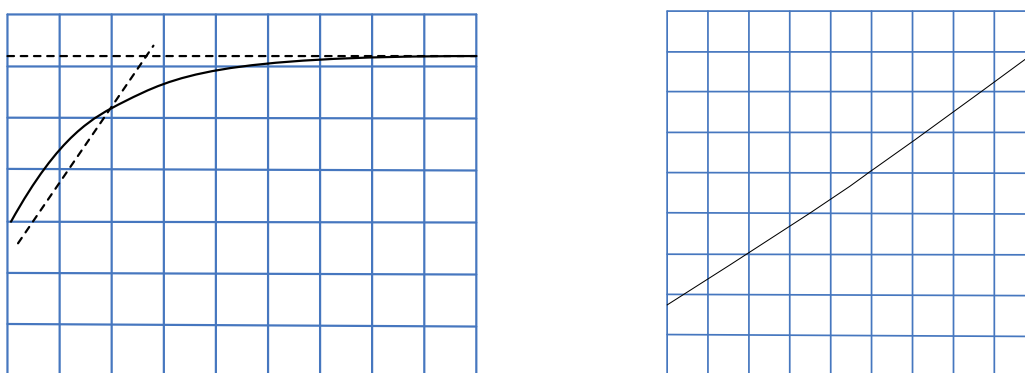


Рис. 5.5. Продуктивність і час перебування заявок у ЗММО

Продуктивність мережі асимптотично наближається до максимально можливої продуктивності (пропускної спроможності ЗММО), що збігається з гранично допустимою інтенсивністю надходження заявок в аналогічній розімкненій ММО (див. Приклад 5.1), при якій в мережі відсутні перевантаження, і дорівнює $\lambda_0 = 0,125 \text{ с}^{-1}$.

Середній час перебування заявок в ЗММО зростає необмежено зі збільшенням кількості заявок з мережі.

Інші характеристики замкнутої ММО (завантаження і коефіцієнти простою вузлів, час очікування, довжини черг і число заявок в вузлах мережі, повний час очікування в мережі) можуть бути розраховані з використанням фундаментальних співвідношень, представлених в розділі 2.

5.2.4. Аналіз властивостей ЗММО

Для замкнутих ММО, як і для розімкнених, найбільший інтерес представляють властивості мережі в цілому, зокрема, вплив циркулюючого в ЗММО числа заявок на такі мережеві характеристики як продуктивність λ_0 замкнутої ММО і середній час перебування τ_{net} заявок в мережі.

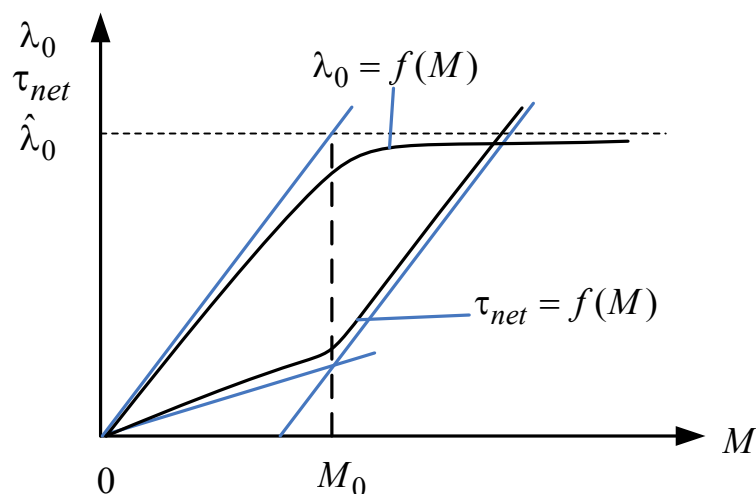


Рис. 5.6. Характеристики ЗММО

Аналіз представлених на рис.5.6 залежностей дозволяє сформулювати наступні висновки.

1. Залежність $\lambda_0 = f(M)$ продуктивності ЗММО λ_0 від числа M циркулюючих заявок спочатку зростає зі збільшенням M до деякого значення M_0 , після якого зростання продуктивності сповільнюється, а з подальшим збільшенням M продуктивність мережі асимптотично прагне до деякого граничного значення $\hat{\lambda}_0$, яке представляє собою *пропускну здатність* ЗММО. Для пояснення цієї залежності згадаємо, що продуктивність замкнутої мережі вимірюється як інтенсивність потоку заявок, що проходять через деяку умовну точку, що позначається як «0», і розташовану на одній з дуг ММО, що відображає завершення обслуговування заявок в мережі і миттєве формування нової заявки, що надходить в мережу. Вище (див. приклад 5.3) було показано, що збільшення числа заявок в замкнутій ММО призводить до збільшення значень всіх мережевих характеристик, включаючи продуктивність λ_0 . В свою чергу, збільшення продуктивності призводить до збільшення завантажень вузлів ММО, пов'язаних з інтенсивністю λ_0 залежністю:

$$\rho_j = \frac{\alpha_j \lambda_0 \tau_{sj}}{K_j},$$

де α_j, τ_{sj}, K_j - відповідно коефіцієнт передачі, середня тривалість обслуговування і кількість приладів у вузлі $j = \overline{1, N}$.

Коли число заявок в ЗММО досягає деякого значення M_0 , загрузка одного з вузлів стає близькою до 1, при цьому практично припиняється зростання продуктивності, яка при $M \rightarrow \infty$ досягає свого граничного значення - пропускну здатності $\hat{\lambda}_0$. Такий вузол являє собою «вузьке місце» мережі, і значення пропускну здатності $\hat{\lambda}_0$ визначається пропускну здатністю вузького місця за умови, що завантаження ρ_y цього вузла дорівнює 1:

$$\rho_y = \frac{\alpha_y \lambda_0 \tau_{sy}}{K_y} = 1.$$

Звідси пропускну здатність замкнутої ММО:

$$\hat{\lambda}_0 = \frac{K_y}{\alpha_y \tau_{sy}},$$

де α_y , τ_{sy} , K_y - відповідно коефіцієнт передачі, середня тривалість обслуговування і кількість обслуговуючих приладів в вузлі y , що є вузьким місцем.

Права частина останнього висловлювання є пропускнуою здатністю вузла, що є вузьким місцем мережі:

$$\mu_y = \frac{K_y}{\alpha_y \tau_{sy}}.$$

Дійсно, $\alpha_y \tau_{sy}$ являє собою повний час обслуговування однієї заявки в даному вузлі з урахуванням того, що заявка за час перебування в мережі в середньому α_y раз побуває в даному вузлі. Тоді величина, зворотня $\alpha_y \tau_{sy}$, являє собою інтенсивність обслуговування заявок

одним приладом в даному вузлі: $\mu_1 = \frac{1}{\alpha_y \tau_{sy}}$, а $\mu_y = K_y \mu_1$ - інтенсивність обслуговування заявок вузлом, тобто всіма приладами.

Цей же результат можна отримати наступними міркуваннями [6]. Якщо завантаження деякого вузла, що є вузьким місцем ММО, стає рівним 1, то це означає, що всі прилади даного вузла постійно обслуговують заявки, тобто не простоюють. Тоді інтенсивність потоку заявок, що виходить з цього вузла, буде дорівнювати інтенсивності обслуговування:

$$\lambda_y = \mu_y = K_y \mu_1.$$

Нагадаємо, що інтенсивність потоку заявок у вузлі λ_y пов'язана з продуктивністю ЗММО λ_0 залежністю $\lambda_y = \alpha_y \lambda_0$. Звідси випливає, що продуктивність ЗММО дорівнює

$$\lambda_0 = \frac{\lambda_y}{\alpha_y} = \frac{K_y \mu_1}{\alpha_y} = \frac{K_y}{\alpha_y \tau_{sy}}.$$

2. Середній час перебування заявок (рис. 5.6) в замкнутій ММО, як і продуктивність, зростає зі збільшенням числа M циркулюючих в мережі заявок, причому спочатку спостерігається незначне зростання, а потім, після значення $M = M_0$, спостерігається лінійне зростання часу перебування.

Дійсно, якщо в мережі циркулює тільки одна заявка, то в такій мережі не може бути черг, і час перебування заявок в ММО складається тільки з часів обслуговування заявок у вузлах з урахуванням коефіцієнтів передач:

$$\tau_{net} = \sum_{i=1}^N \alpha_i \tau_{si}.$$

Зі збільшенням числа заявок M у вузлах ЗММО з'являються черги, причому очевидно, що чим більше заявок в мережі, тим довші черги утворюються у вузлах і тим більше час очікування, а, отже, і час перебування заявок в ЗММО.

Зіставляючи залежності продуктивності та середнього часу перебування заявок від їх числа в ЗММО, можна зробити наступний висновок: збільшення числа заявок в мережі, з одного боку, призводить до збільшення продуктивності, що може розглядатися як позитивний фактор, а, з іншого боку, - до збільшення часу перебування заявок в мережі, що є небажаним чинником.

Точка $M = M_0$ характеризує деяке граничне значення числа заявок в ЗММО. Подальше збільшення числа заявок в мережі виявляється недоцільним, оскільки призводить до різкого збільшення часу перебування заявок в ЗММО при незначному збільшенні продуктивності мережі.

3. Коли завантаження вузького місця стає рівною одиниці, подальше зростання продуктивності за рахунок збільшення числа заявок в ЗММО неможливе. Для збільшення продуктивності ЗММО, як і в РММО, необхідно розвантажити вузьке місце, тобто зменшити завантаження:

$$\rho_y = \frac{\alpha_y \lambda_0 \tau_{sy}}{K_y} = 1, \text{ що при одній і тій же продуктивності може}$$

бути досягнуто:

- зменшенням тривалості обслуговування заявок τ_{sy} , наприклад за рахунок збільшення швидкості роботи (швидкодії) обслуговуючого приладу;

- збільшенням числа обслуговуючих приладів K_y у вузлі;

- зменшенням коефіцієнта передачі α_y або, що те ж саме, ймовірності передачі заявок до вузла, що є вузьким місцем.

Якщо до розвантаження вузького місця залежність продуктивності ЗММО від числа заявок в мережі мала вигляд $\lambda'_0 = f(M)$ (рис. 5.7), а пропускна здатність дорівнювала $\hat{\lambda}'_0$, то після розвантаження – залежність продуктивності від числа заявок буде мати вигляд $\lambda''_0 = f(M)$, а пропускна здатність стане рівною $\hat{\lambda}''_0 > \hat{\lambda}'_0$. При цьому граничні значення числа заявок в ЗММО збільшаться: $M''_0 > M'_0$.

Слід зазначити, що до розглянутої залежності продуктивності ЗММО λ_0 від числа M циркулюючих в мережі заявок може бути застосована лінійна апроксимація $\tilde{\lambda}''_0 = \tilde{f}(M)$, показана на рис. 5.7 у вигляді пунктирних ліній і представляє собою верхню межу продуктивності ЗММО. Останнє означає, що продуктивність ЗММО буде не більше, ніж розраховане верхнє значення.

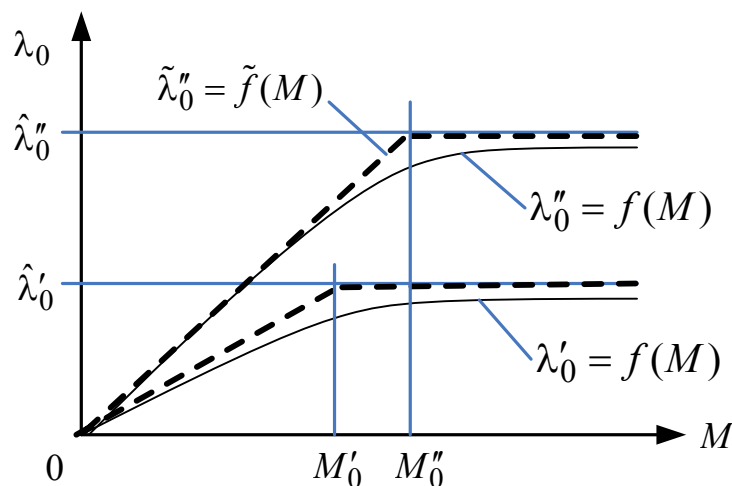


Рис. 5.7. Завантаження «вузького місця»

Неважко уявити собі й зобразити на графіку, як зміниться залежність середнього часу перебування заявок в замкнутій ММО від числа циркулюючих в мережі заявок після розвантаження вузького місця.

Відзначимо, що в деяких випадках розвантаження вузького місця не призводить до поліпшення характеристик ММО, зокрема, до збільшення продуктивності. Зазвичай це пов'язано з тим, що в ММО може існувати декілька вузлів, які є «вузькими місцями». Умовою

цього є рівність завантажень вузлів: $\rho_i = \rho_j$ або $\frac{\alpha_i \lambda_0 \tau_{si}}{K_i} = \frac{\alpha_j \lambda_0 \tau_{sj}}{K_j}$,

звідки остаточно отримаємо: $\frac{\alpha_i \tau_{si}}{K_i} = \frac{\alpha_j \tau_{sj}}{K_j}$ ($i \neq j$). В цьому випадку

для поліпшення характеристик ЗММО необхідно одночасно розвантажити всі «вузькі місця».

Послідовно розвантажуючи вузькі місця ММО, ми можемо прийти до деякої «ідеальної» мережі, в якій завантаження всіх вузлів однакові. **ММО, в якій завантаження всіх вузлів рівні, називається збалансованою.** Збалансована ММО володіє найкращими характеристиками в порівнянні з незбалансованою. При побудові реальних систем, моделями яких служать ММО, необхідно, по-можливості, будувати збалансовані системи, хоча на практиці з багатьох причин досягти цього не вдається.

5.3. Самоконтроль. Перелік питань і завдань

1. Що показує коефіцієнт передачі в ММО?
2. Дати фізичне тлумачення значення коефіцієнта передачі вузла ММО, рівне: а) 4; б) 0,4.
3. У чому відмінність між еквівалентним і толерантним перетвореннями ММО? Привести приклад еквівалентного перетворення ММО.
4. Що таке «вузьке місце» і які способи використовуються для розвантаження «вузького місця»?
5. Показати на графіку, як зміниться залежність продуктивності та середнього часу перебування заявок в замкнутій ММО від числа циркулюючих в мережі заявок після розвантаження вузького місця. В яких випадках розвантаження вузького місця може не дати позитивного ефекту?

6. В двовузловій замкнутій ММО циркулює 1 заявка. Визначити завантаження вузлів 1 і 2, якщо відомо, що завантаження вузла 1 в 3 рази більше, ніж завантаження вузла 2.

7. В замкнутій двовузловій ММО циркулює одна заявка, яка послідовно переходить з одного вузла в інший. Тривалість обслуговування в вузлах розподілена за експоненційним законом з одним і тим же середнім значенням, рівним 5 хвилин. За яким законом розподілено час перебування заявки в мережі? Визначити продуктивність замкнутої ММО.

8. У замкнутій двовузловій ММО циркулює одна заявка, яка послідовно переходить з одного вузла в інший. Тривалості обслуговування в вузлах 1 і 2 мережі відповідно рівні 2 і 3 с. Визначити: а) коефіцієнти простою вузлів замкнутої ММО; б) середнє число заявок, що знаходяться в кожному з вузлів ММО.

9. У замкнутій двовузловій ММО циркулює 4 заявки, які послідовно переходять з одного вузла в інший. Тривалості обслуговування заявок у вузлах мережі однакові і рівні 2 с. Середній час очікування заявок у вузлі 1 дорівнює 3 с. Визначити: а) продуктивність замкнутої ММО; б) завантаження вузлів мережі; в) середнє число заявок, що знаходяться в стані очікування.

10. В розімкнену ММО надходять заявки з інтервалом 5 секунд. Час перебування заявок в мережі дорівнює 15 секунд. Визначити середнє число заявок в мережі та інтенсивність виходу потоку заявок з мережі.

11. Середні часи очікування заявок у вузлах трьохвузлової ММО відповідно рівні: 1, 2 і 4 секунди, а коефіцієнти простою вузлів дорівнюють 0,8; 0,4; 0,7. Визначити середній час очікування заявок в мережі, якщо відомо, що тривалості обслуговування заявок у всіх вузлах однакові і коефіцієнт передачі вузла 1 дорівнює 2.

12. Відомі ймовірності станів двовузлової замкнутої ММО: $P(0,4) = 0,1$; $P(1,3) = 0,4$; $P(2,2) = 0,2$; $P(3,1) = 0,1$; $P(4,0) = 0,2$, де стан (i_1, i_2) задає число заявок в одноканальному вузлі 1 і трьохканальному вузлі 2 відповідно. Визначити середнє число заявок в ММО, що знаходяться в стані очікування.

13. Відомі ймовірності станів трьохвузлової замкнутої ММО: $P(0,0,2) = 0,1$; $P(0,1,1) = 0,2$; $P(0,2,0) = 0,15$; $P(1,0,1) = 0,35$; $P(1,1,0) = 0,05$; $P(2,0,0) = 0,15$, де стан (i_1, i_2, i_3) задає число заявок у вузлі 1, 2, 3 відповідно. Визначити середнє число паралельно працюючих ву-

злів мережі.

14. Відомі ймовірності станів трьохвузлової замкнутої ММО: $P(0,0,2) = 0,1$; $P(0,1,1) = 0,3$; $P(0,2,0) = 0,4$; $P(1,0,1) = 0,05$; $P(1,1,0) = 0,05$; $P(2,0,0) = 0,1$. Тривалості обслуговування заявок у всіх одноканальних вузлах однакові. Визначити значення коефіцієнтів передачі другого і третього вузлів мережі, якщо відомо, що коефіцієнт передачі першого вузла дорівнює 2.

15. Відомі ймовірності станів трьохвузлової замкнутої ММО: $P(0,0,2) = 0,1$; $P(0,1,1) = 0,3$; $P(0,2,0) = 0,4$; $P(1,0,1) = 0,05$; $P(1,1,0) = 0,05$; $P(2,0,0) = 0,1$. Визначити продуктивність ММО, якщо відомо, що коефіцієнт передачі першого вузла (чотирьохканального) дорівнює 2, а середня тривалість обслуговування заявок в цьому вузлі дорівнює 0,1 с.

УЗАГАЛЬНЕННЯ МАТЕРІАЛУ ДЛЯ МОДУЛЬНОГО КОНТРОЛЮ

1. **Одноканальна експоненційна СМО М/М/1** є найбільш простою з точки зору аналітичного розрахунку. Середні часи очікування та перебування заявок в **СМО М/М/1** розраховуються за порівняно простим формулами:

$$\tau_w = \frac{\rho \tau_s}{1 - \rho} \quad \text{і} \quad \tau_{sys} = \frac{\tau_s}{1 - \rho},$$

де $\rho = \lambda \tau_s < 1$ - завантаження системи; λ - інтенсивність надходження заявок в систему; τ_s - середня тривалість обробки заявок в приладі. Для **СМО М/Г/1** середній час очікування заявок визначається за формулою Поллачека-Хінчина:

$$\tau_w = \frac{\lambda \tau_s^2 (1 + v_s^2)}{2(1 - \rho)},$$

де v_s - коефіцієнт варіації тривалості обслуговування.

Для загального випадку одноканальних СМО типу **Г/Г/1** з однорідним потоком застосовуються наближені аналітичні методи розрахунку.

Властивості одноканальної СМО з однорідним потоком заявок:

- середній час очікування заявок в черзі мінімальний при детермінованій тривалості обслуговування заявок з коефіцієнтом варіації $v_s = 0$ і збільшується нелінійно зі зростанням коефіцієнта варіації (дисперсії) тривалості обслуговування;
- середній час очікування заявок істотно залежить від навантаження δ (завантаження ρ) системи і при $\delta \geq 1$ ($\rho \rightarrow 1$) зростає необмеже-

но: $\tau_w \rightarrow \infty$, тобто заявки можуть очікувати обслуговування як за-
вгодно довго;

- для дисциплін обслуговування в зворотному порядку (LIFO) і об-
слуговування у випадковому порядку (ВВ) середні часи очікування
заявок будуть такими ж, як і при обслуговуванні в порядку надхо-
дження (FIFO), але дисперсії часу очікування будуть більше.

2. У разі **багатоканальних СМО** з однорідним потоком заявок
точний метод розрахунку середнього часу очікування заявок розроб-
лений тільки для **СМО типу М/М/К**:

$$\tau_w = \frac{P\tau_s}{K(1-\rho)},$$

де $\rho = \frac{\lambda\tau_s}{K}$ - завантаження системи; P - імовірність того, що всі K
приладів зайняті обслуговуванням заявок:

$$P = \frac{(K\rho)^K}{K!(1-\rho)} P_0,$$

де P_0 - імовірність простою багатоканальної СМО, тобто ймовірність
того, що в системі немає заявок:

$$P_0 = \left[\frac{(K\rho)^K}{K!(1-\rho)} + \sum_{i=0}^{K-1} \frac{(K\rho)^i}{i!} \right]^{-1}$$

**Властивості багатоканальної СМО з однорідним потоком за-
явок:**

- зі збільшенням числа обслуговуючих приладів часи очікування
та перебування заявок зменшуються, при цьому в межі при $K \rightarrow \infty$
час очікування прагне до нуля, а час перебування стає рівним трива-
лості обробки заявок;

- при збільшенні числа обслуговуючих приладів K і збереженні
їх сумарної продуктивності (швидкості роботи) час очікування заявок
зменшується, однак час перебування заявок в системі збільшується й
при $K \rightarrow \infty$ асимптотично прагне до тривалості обслуговування за-
явок, тобто з точки зору затримок (часу перебування заявок) більш

ефективною є одноканальна система, ніж багатоканальна, при рівності сумарної продуктивності; перевагою багатоканальної системи є більш висока надійність;

- середній час очікування заявок, як і для одноканальних систем, залежить від навантаження δ (завантаження ρ) системи і при $\delta \geq K$ ($\rho \rightarrow 1$) час очікування заявок зростає *необмежено*: $\tau_w \rightarrow \infty$, тобто заявки можуть очікувати обслуговування як завгодно довго.

3. Для **одноканальних СМО з неоднорідним потоком заявок і безпріоритетній ДО** середні часи очікування однакові для всіх класів заявок і визначаються за такою формулою:

$$\tau_{wh}^{БП} = \tau_w^{БП} = \frac{\sum_{i=1}^H \lambda_i \tau_{si}^2 (1 + v_{si}^2)}{2(1 - \rho)} \quad (h = 1, \dots, H)$$

де $R = \sum_{i=1}^H \rho_i = \sum_{i=1}^H \lambda_i \tau_{si}$ - сумарне завантаження системи ($R < 1$).

Властивості одноканальної СМО з безпріоритетною ДО:

- середній час очікування заявок різних класів при використанні ДО БП однаково;
- середній час очікування заявок в черзі мінімальний при детермінованій тривалості обслуговування заявок кожного класу і збільшується нелінійно зі зростанням коефіцієнта варіації (дисперсії) тривалості обробки;
- середній час очікування заявок залежить від сумарного навантаження Δ (завантаження R) системи і при $\Delta \geq 1$ ($R \rightarrow 1$) час очікування заявок всіх класів зростає *необмежено*: $\tau_w^{БП} \rightarrow \infty$, проте середні часи перебування в системі і середні довжини черг заявок різних класів, в загальному випадку, різні, оскільки різні тривалості обробки та інтенсивності надходження заявок різних класів;
- для безпріоритетних дисципліни обслуговування в зворотному порядку (LIFO) середні часи очікування заявок будуть такими ж, як і при обслуговуванні в порядку надходження (FIFO), але дисперсія часу очікування буде більше;

- для дисципліни обслуговування в циклічному порядку середні часи очікування заявок різних класів в загальному випадку *не рівні*.

4. Для ДО ВП середні часи очікування заявок h -го класу визначаються за такою формулою:

$$\tau_{wh}^{\text{ВП}} = \frac{\sum_{i=1}^H \lambda_i \tau_{si}^2 (1 + v_{si}^2)}{2(1 - R_{h-1})(1 - R_h)} \quad (h = 1, \dots, H),$$

де R_{h-1} і R_h - сумарні завантаження системи з боку заявок, які мають пріоритет не нижче $(h - 1)$ і h відповідно.

Властивості ДО ВП:

- введення відносних пріоритетів у порівнянні з ДО БП призводить до зменшення часу очікування високопріоритетних заявок і до збільшення часу очікування низькопріоритетних заявок;

- середні часи очікування заявок при використанні ДО ВП монотонно збільшуються зі зменшенням пріоритету:

$$\tau_{w1}^{\text{ВП}} < \tau_{w2}^{\text{ВП}} < \dots < \tau_{wH}^{\text{ВП}};$$

- ДО ВП має властивість захисту від перевантажень, яка полягає в тому, що високопріоритетні заявки навіть при виникненні перевантаження ($\Delta \geq 1$) мають кінцевий час очікування за рахунок відмови в обслуговуванні низькопріоритетним заявками, час очікування яких при цьому різко зростає і прагне до нескінченності.

5. Для ДО АП середні часи очікування заявок h -го класу визначаються за наступною формулою:

$$\tau_{wh}^{\text{АП}} = \frac{\sum_{i=1}^h \lambda_i \tau_{si}^2 (1 + v_{si}^2)}{2(1 - R_{h-1})(1 - R_h)} + \frac{R_{h-1} \tau_{sh}}{1 - R_{h-1}} \quad (h = 1, \dots, H),$$

де R_{h-1} і R_h - сумарні завантаження системи з боку заявок, які мають пріоритет не нижче $(h - 1)$ і h відповідно.

Властивості ДО АП:

- середній час очікування заявок складається з двох складових: середнього часу очікування початку обслуговування і середнього часу очікування в перерваному стані;
- час очікування заявок класу h залежить тільки від значень параметрів класів заявок, що мають більш високий або такий же пріоритет, і не залежить від параметрів класів заявок, що мають нижчий пріоритет;
- для заявок з найвищим абсолютним пріоритетом забезпечується мінімально можливий час очікування порівняно з усіма іншими ДО;
- повний час очікування у заявок високопріоритетного класу h може виявитися більше, ніж у заявок класу $(h + 1)$ з більш низьким пріоритетом, якщо тривалості обслуговування заявок цих класів пов'язані співвідношенням $\tau_{sh} \gg \tau_{s(h+1)}$;
- введення АП порівняно з ВП призводить до зменшення середнього часу очікування самих високопріоритетних заявок і до його збільшення для заявок класу найнижчого пріоритету;
- при ДО АП високопріоритетні заявки краще захищені від перевантажень, ніж при ДО ВП.

6. Відповідно до **закону збереження** часу очікування зміна ДО дозволяє зменшити час очікування високо пріоритетних заявок за рахунок збільшення часу очікування низькопріоритетних заявок:

$$\sum_{i=1}^N \rho_i \tau_{wi} = \text{Const}(\text{ДО}).$$

Закон збереження виконується при наступних умовах:

- система без втрат;
- система простоює лише при відсутності в системі заявок;
- при наявності переривань тривалість обслуговування перерваних заявок розподілена за експоненційним законом;
- всі потоки заявок, що надходять - найпростіші, і тривалість обслуговування не залежить від параметрів потоків заявок.

7. **Лінійних розімкнених однорідних експоненційних ММО (РММО)** описуються наступними параметрами:

- число вузлів в мережі: N ;
- число обслуговуючих приладів у вузлах мережі: $K_1, \dots, K_N$;

- матриця ймовірностей передач: $P = [p_{ij} | i, j = 0, 1, \dots, N]$, де p_{ij} - ймовірність передачі заявки з вузла i у вузол j ;
- інтенсивність λ_0 джерела заявок, що надходять в РММО;
- середні тривалості обслуговування заявок у вузлах мережі: $\tau_{s1}, \dots, \tau_{sN}$.

Умова відсутності перевантажень в розімкненій ММО:

$$\lambda_0 < \min \left(\frac{K_1}{\alpha_1 \tau_{s1}}, \frac{K_2}{\alpha_2 \tau_{s2}}, \dots, \frac{K_N}{\alpha_N \tau_{sN}} \right).$$

Розрахунок характеристик функціонування лінійних розімкнених однорідних експоненційних ММО базується на еквівалентному перетворенні мережі і проводиться в три етапи.

I - розрахунок інтенсивностей потоків заявок λ_j у вузлах $j = \overline{1, N}$ ММО шляхом вирішення системи лінійних алгебраїчних рівнянь:

$$\lambda_j = \sum_{i=0}^N p_{ij} \lambda_i \quad (j = 0, 1, \dots, N);$$

II - розрахунок вузлових характеристик:

- навантаження у вузлі j , що показує середнє число зайнятих приладів: $\delta_j = \lambda_j \tau_{sj}$;
- загрузка вузла j : $\rho_j = \min(\delta_j / K_j; 1)$, де K_j - число обслуговуючих приладів у вузлі j ;
- коефіцієнт простою вузла: $\delta_j = 1 - \rho_j$;
- час перебування заявок у вузлі: $\tau_{netj} = \tau_{wj} + \tau_{sj}$;
- довжина черги заявок: $l_j = \lambda_j \tau_{wj}$;
- число заявок у вузлі (в черзі і на обробці в приладі): $m_j = \lambda_j \tau_{sysj}$.

III - розрахунок мережових характеристик:

- середнє число заявок, що очікують обслуговування в мережі, і середнє число заявок, що знаходяться в мережі:

$$L = \sum_{j=1}^N l_j ; \quad M = \sum_{j=1}^N m_j ,$$

де l_j - середня довжина черги і m_j - середнє число заявок у вузлі j ;

- середній час очікування і середній час перебування заявок в мережі:

$$\tau_{netw} = \sum_{j=1}^N \alpha_j \tau_{wj} ; \quad \tau_{net} = \sum_{j=1}^N \alpha_j \tau_{sysj} ,$$

де τ_{wj} і τ_{sysj} - відповідно середній час очікування і середній час перебування заявок у вузлі j ;

$\alpha_j = \lambda_j / \lambda_0$ ($j = \overline{1, N}$) - коефіцієнт передачі для вузла j , що показує середнє число влучень заявки в вузол j за час її перебування в мережі.

Властивості розімкнених ММО:

- властивості окремих вузлів ММО аналогічні властивостям відповідних одноканальних та багатоканальних СМО;

- зі збільшенням інтенсивності λ_0 надходження заявок в мережу мережеві характеристики збільшуються, причому є граничне значення інтенсивності $\lambda'_{0_{\max}}$, при якому, зокрема, середній час перебування заявок в мережі стає нескінченно великим, що свідчить про перевантаження в ММО;

- вузол мережі, завантаження якого зі збільшенням λ_0 прагне до одиниці, називається «вузьким місцем» і характеризується нескінченним зростанням черги заявок перед ним і, як наслідок, нескінченним зростанням числа заявок в ММО;

- способи розвантаження «вузького місця»:

а) збільшення швидкості роботи (швидкодії) обслуговуючого приладу;

б) збільшення числа обслуговуючих приладів у вузлі;

в) зменшення ймовірності передачі заявок до вузла, що є вузьким місцем.

8. Лінійні замкнуті однорідні експоненційні ММО (ЗММО) описуються наступними параметрами:

- число вузлів в мережі: N ;

- число обслуговуючих приладів у вузлах мережі: $K_1, \dots, K_N$;
- матриця ймовірностей передач: $P = [p_{ij} | i, j = 0, 1, \dots, N]$, де p_{ij} - ймовірність передачі заявки з вузла i у вузол j ;
- число заявок M^* , циркулюючих в ЗММО;
- середні тривалості обслуговування заявок у вузлах мережі: $\tau_{s1}, \dots, \tau_{sN}$.

У замкнутих ММО завжди існує сталий режим.

Розрахунок характеристик функціонування лінійних замкнутих однорідних експоненційних ММО з одноканальними вузлами проводиться з використанням методу середніх значень в два етапи.

I - Розрахунок коефіцієнтів передач у вузлах замкнутої ММО шляхом рішення системи лінійних алгебраїчних рівнянь щодо $\alpha_1, \dots, \alpha_N$ з урахуванням того, що ($\alpha_0 = 1$):

$$\alpha_j = \sum_{i=0}^N p_{ij} \alpha_i \quad (i = 0, 1, \dots, N);$$

II - Розрахунок характеристик ЗММО з використанням наступних рекурентних співвідношень для значень $M = 1, 2, \dots, M^*$:

$$\tau_{neti}(M) = \tau_{si}[1 + m_i(M - 1)];$$

$$\tau_{net}(M) = \sum_{i=1}^N \alpha_i \tau_{neti}(M);$$

$$\lambda_0(M) = \frac{M}{\tau_{net}(M)};$$

$$m_i(M) = \alpha_i \lambda_0(M) \tau_{neti}(M),$$

де M^* - задане число заявок в замкнутій мережі; $m_i(0) = 0$.

Властивості замкнутих ММО:

- залежність продуктивності ЗММО λ_0 від числа M циркулюючих заявок зростає зі збільшенням M і прагне до деякого граничного значення $\hat{\lambda}_0$, яке представляє собою пропускну здатність ЗММО;

- середній час перебування заявок в замкнутій ММО, як і продуктивність, зростає зі збільшенням числа циркулюючих в мережі заявок, причому зростання часу перебування спочатку незначне, а потім приймає лінійний характер;

- збільшення числа заявок в мережі, з одного боку, призводить до збільшення продуктивності (позитивний фактор), а, з іншого боку - до збільшення часу перебування заявок в мережі (небажаний фактор);

- для кожної замкнутої ММО існує деяке граничне значення числа заявок в мережі, після якого різко збільшується час перебування заявок в ЗММО при незначному збільшенні продуктивності мережі;

- коли завантаження вузького місця стає рівною одиниці, подальше зростання продуктивності за рахунок збільшення числа заявок в ЗММО неможливе; для збільшення продуктивності ЗММО необхідно розвантажити вузьке місце одним з наступних способів:

- а) зменшенням тривалості обслуговування заявок (збільшенням швидкості роботи обслуговуючого приладу);

- б) збільшенням числа обслуговуючих приладів у вузлі;

- в) зменшенням коефіцієнта передачі.

- якщо в ММО існує декілька вузлів, які є одночасно «вузьким місцем», для поліпшення характеристик функціонування ЗММО необхідно одночасно розвантажити всі вузькі місця;

- при побудові реальних систем, моделями яких служать ММО, слід, по-можливості, будувати збалансовані системи, в яких завантаження всіх вузлів однакові.

9. Теорема Джексона, що може бути використана для аналізу ММО, базується на трьох припущеннях:

- ММО складається з N вузлів, кожен з яких забезпечує незалежне експоненційне обслуговування;

- заявки, що приходять в ММО, в будь який її вузол, являють Пуасонівський потік;

- заявка, що обслужена в вузлі, миттєво поступає в чергу на один з інших вузлів з фіксованою ймовірністю або покидає ММО.

Теорема Джексона стверджує, що в такій мережі масового обслуговування, кожен вузол є самостійною СМО з Пуасонівським входом, що обумовлює принципи розподілення, злиття потоків і аналізу тандемів СМО. Таким чином, кожен вузол може бути проаналізовано окремо від інших, використовуючи М/М/1 або М/М/Н моделі, і результати можуть бути об'єднані звичайними статистичними метода-

ми. Середня затримка в кожному вузлі може бути додана до сумарної середньої затримки в системі, але нічого не можна сказати про вищі моменти системної затримки (наприклад, стандартне відхилення).

Теорема Джексона є привабливою для застосування до мереж з пакетною комутацією. Модель мережі з комутацією пакетів можна представити як ММО. Кожен пакет являє собою індивідуальну заявку на обслуговування. Ми припускаємо, що кожен пакет передається окремо, і в кожному вузлі комутації пакет становиться в чергу для передачі його у напрямку до вузла-адресата. Час обробки пакета в СМО пропорційний його довжині.

Недолік цього підходу в тому, що умови теореми можуть порушуватись, а саме: не факт, що розподілені обробки є незалежними, оскільки довжини пакетів в кожному з каналів відрізняються і процес прибуття заявок в кожному СМО корелює з процесом її обробки. Тим не менш, Клейнрок показав [8], що через усереднення під час злиття і розподілення потоків, припущення щодо незалежності часу обробки забезпечує хорошу апроксимацію.

ПРАКТИКУМ. РІШЕННЯ ЗАДАЧ

Задача 1. В одноканальну систему обслуговування надходять заявки двох класів з інтенсивностями 0,3 і 1 заявок в секунду. Інтенсивність їх обслуговування відповідно рівні 0,5 і 5 заявок в секунду.

а) Сформулювати умови, при яких час перебування заявок 1-го класу буде дорівнює 2 секунди?

б) Чому дорівнюватиме час перебування заявок 1-го класу, якщо при тих же умовах інтенсивність їх надходження збільшиться в два рази?

в) Чому дорівнюватиме час перебування заявок 1-го класу, якщо при тих же умовах інтенсивність їх обслуговування збільшиться в два рази?

Дано: Одноканальна СМО:

- кількість класів заявок: $H = 2$;

- інтенсивності потоків: $\lambda_1 = 0,3 \text{ c}^{-1}$; $\lambda_2 = 1 \text{ c}^{-1}$

- інтенсивності обслуговування: $\mu_1 = 0,5 \text{ c}^{-1}$; $\mu_2 = 5 \text{ c}^{-1}$.

Потрібно:

а) сформулювати умови, при яких $\tau_{\text{sys}1} = 2 \text{ c}$;

б) визначити $\tau'_{\text{sys}1} = ?$ при $\lambda'_1 = 2\lambda_1$;

в) визначити $\tau''_{\text{sys}1} = ?$ при $\mu''_1 = 2\mu_1$.

Рішення.

а) Час перебування заявок класу 1: $\tau_{\text{sys}1} = \tau_{\text{w}1} + \tau_{\text{s}1}$, де $\tau_{\text{w}1}$ - час очікування; $\tau_{\text{s}1} = \frac{1}{\mu_1} = 2 \text{ c}$ - тривалість обслуговування. Очевидно, що $\tau_{\text{sys}1} = 2 \text{ c}$, якщо $\tau_{\text{w}1} = 0$, тобто заявки 1-го класу не повинні утворювати чергу. Для цього необхідно, щоб:

- заявки 1-го класу мали абсолютний пріоритет по відношенню до заявок 2-го класу, це означає, що заявки 2-го класу не зможуть впливати на характеристики обслуговування заявок 1-го класу, однак це не виключає утворення черги заявок 1-го класу;

- для того щоб заявки 1-го класу не утворювали чергу, процеси надходження і обслуговування заявок 1-го класу повинні бути детермінованими, тобто інтервали між заявками 1-го класу, що надходять в систему, та тривалості їх обслуговування повинні бути детермінованими (не випадковими) величинами;

- навантаження, що створюється заявками 1-го класу, не повинно перевищувати 1, в іншому випадку система буде перевантажена і не зможе впоратися з обслуговуванням заявок 1-го класу, час очікування яких буде рости до нескінченності.

Перевіримо виконання останньої умови: $\delta_1 = \lambda_1 / \mu_1 = 0,6$ - система працює без перевантажень. Таким чином, для того щоб $\tau_{sys1} = 2$ с, необхідно виконання двох перших умов.

б) Визначимо $\tau'_{sys1} = ?$ при $\lambda'_1 = 2\lambda_1$. Якщо інтенсивність надходження заявок 1-го класу збільшиться в 2 рази, то створюване заявками навантаження теж збільшиться в 2 рази і стане рівним $\delta'_1 = 2\lambda_1 / \mu_1 = 1,2$, що означає перевантаження системи, отже, час очікування і час перебування заявок 1-го класу зростуть до нескінченності: $\tau'_{sys1} = \infty$.

Зауважимо, що якщо б навантаження не перевищило значення 1, то час перебування заявок 1-го класу залишився б тим самим: $\tau_{sys1} = 2$ с.

в) Визначимо $\tau''_{sys1} = ?$ при $\mu''_1 = 2\mu_1$.

Збільшення інтенсивності обслуговування заявок 1-го класу приведе до зменшення навантаження в 2 рази: $\delta''_1 = \lambda_1 / (2\mu_1) = 0,3$, тобто система працюватиме без перевантаження. З іншого боку, тривалість обробки заявок теж зменшиться в 2 рази: $\tau_{s1} = 1 / (2\mu_1) = 1$ с, отже, час перебування стане рівним $\tau''_{sys1} = \tau''_{s1} = 1$ с.

Задача 2. В двоканальну СМО надходить найпростіший потік заявок із середнім інтервалом між сусідніми заявками $\tau_{in} = 10$ с, причому кожна друга заявка направляється до другого приладу (див. рис. 7.1). Чому дорівнює інтенсивність потоку заявок до другого приладу?

Чому дорівнює коефіцієнт варіації інтервалів між заявками потоку до другого приладу?

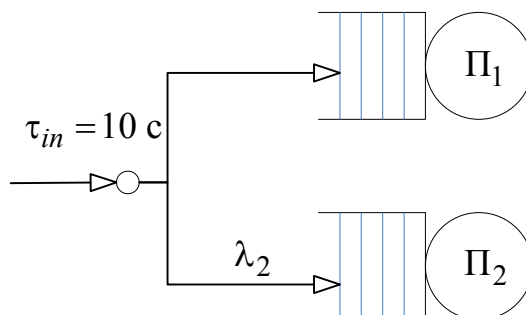


Рис. 7.1. Двоканальна СМО з розрідженням потоків

Дано: СМО; $K = 2$; потік - найпростіший; $\tau_{in} = 10 \text{ с}$.

Потрібно:

Визначити λ_2 ;

Визначити v_2 .

Рішення.

1) Інтенсивність потоку заявок в СМО: $\lambda_0 = \frac{1}{\tau_{in}} = 0,1 \text{ с}^{-1}$. Оскільки

кожна друга заявка направляється до другого приладу, то очевидно, що інтенсивність надходження заявок до другого приладу буде в два рази менше, ніж вхідна інтенсивність λ_0 , тобто $\lambda_2 = 0,5\lambda_0 = 0,05 \text{ с}^{-1}$.

2) Для визначення коефіцієнта варіації v_2 знайдемо вид закону розподілу інтервалів між заявками до другого приладу, для чого побудуємо часову діаграму (рис. 7.2), що відображає процес надходження заявок в систему (а) і до другого приладу (б).

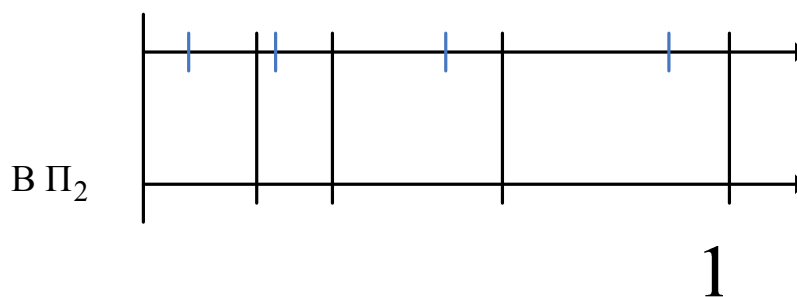


Рис. 7.2. Розрідження потоку заявок

Як видно з діаграми, інтервали між заявками до другого приладу являють собою суму двох часових інтервалів початкового найпростішого потоку заявок, що надходять в систему. Кожен такий часовий інтервал у разі найпростішого потоку являє собою випадкову величину, розподілену за експоненційним законом. Таким чином, інтервали між заявками до другого приладу представляють собою випадкову величину, яка дорівнює сумі двох експоненційно розподілених величин, що відповідає розподілу Ерланга 2-го порядку ($k=2$).

Коефіцієнт варіації випадкової величини, розподіленої за законом Ерланга (Додаток 1), залежить від порядку k і визначається за формулою:

$$v_2 = v_{E_2} = \frac{1}{\sqrt{k}} = \frac{1}{\sqrt{2}} = 0,71.$$

Слід розрізняти розглянуте вище детерміноване розрідження потоку від імовірнісного розрідження. У разі імовірнісного розрідження, коли заявки направляються до другого приладу з імовірністю $p_2 = 0,5$, інтенсивність надходження заявок до другого приладу буде такою ж, як і при детермінованому розрідженні, тобто $\lambda_2 = p_2 \lambda_0 = 0,5 \lambda_0 = 0,05 \text{ с}^{-1}$. Проте коефіцієнт варіації в цьому випадку дорівнює одиниці: $v_2 = 1$, оскільки, відповідно до однієї з сформульованих чудових особливостей найпростішого потоку при імовірнісному розрідженні утворюються найпростіші потоки, в яких інтервали між послідовними заявками розподілені за експоненційним, а не за законом Ерланга.

Задача 3. Інтенсивність надходження заявок в розімкнену трьох-вузлову ММО дорівнює 2 заявки в секунду. Середнє число заявок у вузлах ММО відповідно дорівнює: 2, 4 і 6. Визначити середній час перебування заявок в мережі.

Дано: РММО; $N = 3$; $\lambda_0 = 2 \text{ с}^{-1}$; $m_1 = 2$; $m_2 = 4$; $m_3 = 6$.

Потрібно: визначити τ_{net} .

Рішення.

Середній час перебування заявок в ММО визначається за формулою:

$$\tau_{net} = \sum_{j=1}^N \alpha_j \tau_{sysj} = \sum_{j=1}^N \alpha_j \frac{m_j}{\lambda_j} = \sum_{j=1}^N \alpha_j \frac{m_j}{\alpha_j \lambda_0} = \sum_{j=1}^3 \frac{m_j}{\lambda_0} = \frac{2+4+6}{2} = 6 \text{ с.}$$

Цей же результат можна отримати, виходячи з формули Літтла, що зв'язує середній час перебування і число заявок в мережі:

$$\tau_{net} = \frac{M}{\lambda_0} = \frac{1}{\lambda_0} \sum_{j=1}^3 m_j = 6 \text{ с.}$$

Задача 4. Сервер баз даних. В підрозділі Управління залізниці розгорнута локальна мережа зі 100 персональними комп'ютерами і одним сервером баз даних, що забезпечує даними комп'ютери за їх запитамі. Середній час відповіді сервера на запит становить 0,6 секунди, а стандартне відхилення, за оцінками, дорівнює середньому. У години пік, швидкість надходження запитів від локальної мережі сягає 20 запитів за хвилину.

Ми хотіли б відповісти на наступні питання:

- 1) який в системі середній час відгуку сервера?
- 2) якщо 1,5 секунди вважається максимально допустимим часом відгуку, то на скільки процентів треба знизити навантаження сервера, щоб відбулося досягнення максимуму?
- 3) якщо на 20% збільшиться завантаження сервера, то чи буде час реакції збільшуватись на 20%?

Рішення.

1) Припустимо, що моделлю сервера баз даних буде модель М/М/1. Ми знехтуємо впливом локальної мережі, вважаючи що її внесок у затримку обслуговування незначний.

Рахуємо завантаження серверу:

$$\rho = \lambda \tau_s = \frac{20 \frac{\text{запит}}{\text{хв}} * 0.6 \frac{\text{с}}{\text{запит}}}{60 \frac{\text{с}}{\text{хв}}} = 0,2$$

Тепер можна обчислити першу характеристику, середній час відгуку, за формулою:

$$\tau_{sys} = \frac{\tau_s}{1 - \rho} = \frac{0.6}{1 - 0.2} = 0,75 \text{ с.}$$

2) Другу характеристику отримати більш важко. Замість неї, вирахуємо, коли 90% всіх відповідей будуть менше 1,5 секунд з допомогою рівняння з таблиці 1.3:

$$m_{\tau_{sys}}(y) = \tau_{sys} \times \ln\left(\frac{100}{100 - y}\right)$$

$$m_{\tau_{sys}}(90) = \tau_{sys} \times \ln(10) = \frac{\tau_s}{1 - \rho} \times 2,3 = 1,5 \text{ с}$$

Ми маємо $\tau_s = 0,6$. Тоді, вирішуємо рівняння відносно ρ і отримуємо $\rho = 0,08$. Тобто, щоб 90 % запитів до сервера баз даних мали час відгуку менше 1,5 секунди, необхідно знизити завантаження сервера з 20 % до 8 %.

3) Третя задача: знайти співвідношення між завантаженням сервера і часом відгуку.

Оскільки завантаження сервера 0,2 знаходиться в рівнинній частині кривої, то час відгуку буде зрости повільніше, ніж завантаження. В цьому випадку, якщо завантаження зростає з 20% до 40%, що є 100%-ним збільшенням, то значення τ_{sys} зростає від 0,75 секунд до 1,0 секунди, збільшившись лише на 33%.

$$\tau_{sys} = \frac{\tau_s}{1 - 2\rho} = \frac{0,6}{1 - 2 * 0,2} = \frac{0,6}{0,6} = 1 \text{ с.}$$

Задача 5. Сильно-зв'язаний мультипроцесор. Розглянемо викорис-

тання декількох тісно пов'язаних процесорів в одній комп'ютерній системі. Одне з проектних рішень може бути пов'язане з розподілом заявок між процесорами. Якщо заявку постійно віднести до одного процесору з моменту активації до його завершення, то для кожного процесора ми будемо мати окремі короткі черги. У цьому випадку у одного процесора може бути порожньою черга і він простоює, в той час як інший процесор має велике τ_s . Щоб запобігти цієї ситуації, можна запропонувати загальну чергу. Всі заявки йдуть в одну чергу і обслуговуються в будь-якому доступному процесорі. Таким чином, протягом терміну дії заявки, вона може бути оброблена в будь-якому процесорі в будь який час.

Рішення.

Давайте спробуємо з'ясувати, на скільки збільшиться продуктивність при використанні загальної черги. Розглянемо систему з п'ятьма процесорами ($K=5$), в якій середня кількість процесорного часу на обробку заявки становить $\tau_s = 0,1$ с. Припустимо, що стандартне відхилення часу обробки процесу дорівнює $v_{\tau_s} = 0,094$ с. Оскільки стандартне відхилення близько до середнього, ми припускаємо експоненційний час обробки. Також припустимо, що заявки прибувають в систему зі швидкістю $\lambda = 40$ заявок в секунду.

Підхід сервера-одинака. Якщо заявки рівномірно розподілені між процесорами, то навантаження на кожен процесор $40/5 = 8$ заявок за секунду. Таким чином,

$$\rho = \lambda \tau_s = 8 \times 0,1 = 0,8;$$

$$\tau_w = \frac{\rho \tau_s}{1 - \rho} = \frac{0,8 * 0,1}{1 - 0,8} = 0,4 \text{ с.}$$

Тоді легко обчислити резидентний час:

$$\tau_{sys} = \frac{\tau_s}{1 - \rho} = \frac{0,1}{1 - 0,8} = \frac{0,1}{0,2} = 0,5 \text{ с}$$

Підхід мульти-сервера. Тепер припустимо, що існує одна черга для всіх процесорів. Тепер у нас є сумарна швидкість прибуття заявок - 40 заявок за секунду. Тим не менше, завантаження процесорів залишається 0,8 ($\lambda \tau_s / K$). Для розрахунку резидентного часу τ_{sys} за

формулою з таблиці 2.1, ми повинні спочатку обчислити С-функцію Ерлангу. Цю функцію можна подивитися в таблиці для завантаження 0,8 для 5 серверів, отримуємо $C = 0,554$. Підставляємо у формулу

$$\tau_{sys} = 0,1 + \frac{0,554 \times 0,1}{5 \times (1 - 0,8)} = 0,1544 \text{ с};$$

$$\tau_w = \frac{C}{K} \frac{\tau_s}{1 - \rho} = \frac{0,554}{5} * \frac{0,1}{1 - 0,8} = 0,0554 \text{ с}.$$

Таким чином, використання багатосерверної черги скорочує середній час обслуговування з 0,5 с до 0,1544 с, що більше, ніж в 3 рази. Якщо дивитися тільки на середній час очікування, у багатосерверному випадку це 0,0544 секунди в порівнянні з 0,4 секундами для односерверного випадку, що в 7 разів менше. Можна не бути експертом в теорії масового обслуговування, щоб знати, що одна мультисерверна черга краще, ніж багато односерверних черг.

Задача 6. Розрахунок процентилей. Розглянемо конфігурацію корпоративної мережі в Управлінні залізниці, в якій пакети передаються з комп'ютерів в одній локальній мережі в комп'ютер в іншій локальній мережі. Всі ці пакети повинні проходити через маршрутизатор, який з'єднує ці локальні мережі у корпоративну мережу і з зовнішнім світом. Давайте подивимося на трафік, що проходить через маршрутизатор між локальними мережами. Пакети приходять із середньою швидкістю 5 пакетів за секунду. Середня довжина пакета становить 144 октета і передбачається, що довжина пакета розподілена за експоненційним законом. Лінійна швидкість від маршрутизатора до глобальної мережі - 9600 біт за секунду. Нас цікавлять наступні питання.

- 1) Який середній час перебування пакетів в маршрутизаторі?
- 2) Скільки в середньому пакетів знаходяться в маршрутизаторі, у тому числі тих, що чекають передачі і в даний час передаються?
- 3) Те ж питання, як (2), на 90 відсотків.
- 4) Те ж питання, як (2), на 95 відсотків.

Рішення.

Розрахунки:

$$\begin{aligned}
\lambda &= 5 \text{ пак/с} \\
\tau_s &= (144 \text{ октета} \times 8 \frac{\text{біт}}{\text{октет}}) / 9600 \frac{\text{біт}}{\text{с}} = 0,12 \text{ с} \\
\rho &= \lambda \tau_s = 5 \times 0,12 = 0,6 \\
\tau_{sys} &= \frac{\tau_s}{(1-\rho)} = \frac{0,12}{(1-0,6)} = 0,3 \text{ с} \\
m &= \frac{\rho}{(1-\rho)} = \frac{0,6}{(1-0,6)} = 1,5 \text{ пак}
\end{aligned}$$

Для отримання процентилі, будемо використовувати рівняння з таблиці 1.3:

$$\Pr[m_k = k] = (1-\rho)\rho^k$$

Для обчислення y -ї процентилі розміру черги, запишемо попереднє рівняння в кумулятивному вигляді:

$$\frac{y}{100} = \sum_{k=0}^{m_m(y)} (1-\rho)\rho^k = 1 - \rho^{1+m_m(y)}.$$

Тут $m_m(y)$ представляє максимальне число пакетів в черзі, що очікують, у відсотках часу. Тобто, $m_m(y)$ є те значення, нижче якого k відбувається у процентів часу. У даній формі, ми можемо визначити процентиль для будь-якого розміру черги.

Ми хочемо зробити навпаки: нехай дано y , знайти - $m_m(y)$. Таким чином, логарифмування обох сторін дає (перетворення див. Додаток 2):

$$m_m(y) = \frac{\ln(1 - \frac{y}{100})}{\ln \rho} - 1.$$

Якщо $m_m(y)$ є дробовим, то приймаємо найближче більше ціле, якщо воно негативне, встановлюємо його нулем. В нашому прикладі, $\rho = 0,6$, і ми хочемо, щоб знайти $m_m(90)$ і $m_m(95)$:

$$m_m(90) = \frac{\ln(1 - 0,90)}{\ln(0,6)} - 1 = 3,5$$

$$m_m(95) = \frac{\ln(1 - 0,95)}{\ln(0,6)} - 1 = 4,8$$

Таким чином, 90% часу, в черзі буде менше 4 пакетів і 95% часу - менше 5 пакетів. Якщо б ми розробляли систему за критерієм 95-ї процентиля, то буфер довелося би обирати розміром не менше 5 пакетів.

Задача 7. АСУ СС. На сортувальній станції працює інформаційна система АСУ СС, яка обслуговує запити своїх абонентів. Відомо, що кількість абонентів в системі дорівнює $n=10$, інтенсивність надходження запитів від кожного абонента становить $\lambda_i = 0,02 \text{ с}^{-1}$ ($i = \overline{1,10}$), а середній час обробки заявок процесором, розподілений за показовим законом, складає $\tau_s = 5 \text{ с}$ [13].

Потрібно:

1) Визначити основні показники роботи системи: навантаження процесора ρ , середній час очікування заявки τ_w і середній час відгуку системи τ_{sys} .

2) Вирахувати максимальну кількість абонентів, яку може обслуговувати система при завантаженні процесора $\rho \leq 0,6$.

3) Яким повинно бути завантаження системи, щоб середній час відгуку системи був $\tau_{sys} \leq 1,26 \text{ с}$.

Рішення.

1) Для вирішення завдання будемо використовувати модель М/М/1. Середня інтенсивність загального потоку запитів буде складати $\lambda_0 = n * \lambda_i = 10 * 0,02 = 0,2 \text{ с}^{-1}$. Середнє завантаження системи $\rho = \lambda_0 * \tau_s = 0,2 * 5 = 1$. Я ми бачимо, це навантаження є критичним і потребує коригування конфігурації інформаційної системи.

2) Перший шлях – зменшити кількість абонентів, що обслуговується, з урахуванням обмеження $\rho \leq 0,6$.

$$\begin{aligned} \rho &= n * \lambda_i * \tau_s \leq 0,6 \\ n &\leq \frac{0,6}{\lambda_i * \tau_s} = \frac{0,6}{0,02 * 5} = \frac{0,6}{0,1} = 6 \text{ абон} \end{aligned}$$

Тобто, система може обслуговувати не більше 6 абонентів при завантаженні процесора $\rho \leq 0,6$.

Другий шлях – збільшити продуктивність процесора при $n=10$ абонентів:

$$\rho = \lambda_0 * \tau_s \leq 0,6$$

$$\tau_s \leq \frac{0,6}{\lambda_0} = \frac{0,6}{n * \lambda_i} = \frac{0,6}{10 * 0,02} = \frac{0,6}{0,2} = 3 \text{ с}$$

При існуючому $\tau_s = 5 \text{ с}$ продуктивність процесора необхідно збільшити у 1,7 разі щоб отримати $\tau_s = 3 \text{ с}$

Тоді, середній час очікування заявки $\tau_w = \frac{\rho \tau_s}{1 - \rho} = \frac{0,6 * 3}{1 - 0,6} = \frac{1,8}{0,4} = 4,5 \text{ с}$

Час відгуку буде $\tau_{sys} = \tau_s + \tau_w = 3 + 4,5 = 7,5 \text{ с}$

3) Визначимо завантаження процесора, при якому буде

$$\tau_{sys} \leq 1,26 \text{ с.}$$

$$\tau_s + \tau_w \leq 1,26$$

$$\frac{\rho}{\lambda_0} + \frac{\rho \tau_s}{1 - \rho} \leq 1,26$$

$$\frac{\rho}{\lambda_0} + \frac{\rho * \rho}{(1 - \rho) \lambda_0} \leq 1,26$$

$$\rho(1 - \rho) + \rho^2 \leq 1,26 \lambda_0 (1 - \rho)$$

$$\rho - \rho^2 + \rho^2 \leq 1,26 \lambda_0 - 1,26 \lambda_0 \rho$$

$$\rho \leq 1,26 * 0,2 - 1,26 * 0,2 * \rho$$

$$\rho \leq 0,252 - 0,252 \rho$$

$$1,252 \rho \leq 0,252$$

$$\rho \leq \frac{0,252}{1,252} = 0,2$$

Таким чином, щоб час відгуку був сприйнятним абонентам і не дратував їх затримками, завантаження процесору в інформаційній системі АСУ СС повинно бути менше 0,2!

Задача 8 [14]. Аналіз інформаційно-довідкової системи залізничного вокзалу. Інформаційно-довідкова система залізничного вокзалу дозволяє пасажиром отримувати інформацію, що стосується розкладу руху поїздів, наявності вільних місць, вартості квитків і т. і. Для доступу до системи на вокзалі встановлюються термінали користувачів. Пасажир, який бажає отримати довідку, в режимі діалогу з ЕОМ формулює свій запит на екрані монітора. Відповідь системи та-

кож виводиться на екран монітора. Нехай на вокзалі встановлено досить велике число моніторів, так що черги пасажирів у терміналів не виникає. В цьому випадку для аналізу роботи системи може бути прийнята модель СМО ММ1. Припустимо, що час пошуку на один запит - випадкова величина, розподілена за експоненційним законом з інтенсивністю 10 запитів за хвилину ($\mu = 10 \frac{\text{запит}}{\text{хв}}$). Знайти залежність середнього часу реакції системи (середнього часу отримання відповіді на запит) і середньої довжини черги від інтенсивності потоку заявок, що надходить в систему.

Рішення.

Запишено відповідні залежності.

$$l = \frac{\rho^2}{1-\rho} = \frac{(\lambda/\mu)^2}{1-\lambda/\mu} = \frac{\lambda^2}{\mu^2 - \lambda\mu} = \frac{\lambda^2}{100 - 10\lambda};$$

$$\tau_{sys} = \frac{\tau_s}{1-\rho} = \frac{\frac{1}{\mu}}{1-\frac{\lambda}{\mu}} = \frac{1}{\mu-\lambda} = \frac{1}{10-\lambda}.$$

Отримані вирази справедливі тільки в стаціонарному режимі, коли $\lambda < 10 \frac{\text{запит}}{\text{хв}}$. Графічні залежності $l = f(\lambda)$ і $\tau_{sys} = f(\lambda)$ наведені на рисунку 7.3.

Аналіз отриманих залежностей показує, що існують критичні точки інтенсивності вхідного потоку заявок, в яких починається різке погіршення експлуатаційних характеристик системи. Середня довжина черги починає різко зростати при $\lambda = 8,6$. При цьому середній час відгуку зберігає стабільність до $\lambda = 9,8$, після чого час зростає експоненціально.

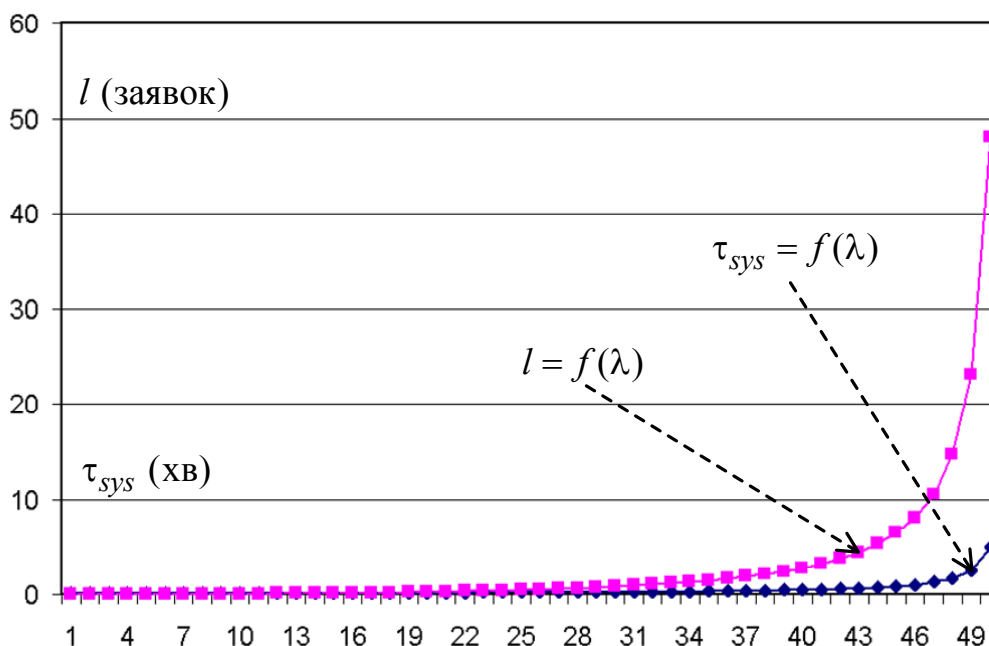


Рис. 7.3. Залежність часу відгуку і середньої довжини черги запитів в системі від потоку заявок

Задача 9 [14]. Аналіз інформаційно-довідкової системи залізничного вокзалу при обмеженому числі терміналів доступу. Розглянемо інформаційно-довідкову систему вокзалу типу описаній в задачі 8, але в дещо інших умовах функціонування. Нехай число терміналів в системі відносно невелике і всі вони постійно зайняті. Нехай також пасажери, отримавши і осмисливши відповідь на перший запит, можуть послати в систему наступний запит (наприклад, з'ясувавши наявність квитків на інші поїзди, їх вартість і т. і.). В цих умовах середній час реакції системи залежить (крім інтенсивності обслуговування μ) ще й від числа терміналів доступу n та інтенсивності надходження запитів від кожного терміналу, або, по-іншому, середнього часу обдумування відповіді на попередній запит і формування нового запиту.

Нехай, як і в задачі 8, інформаційно-довідкова система реалізована на базі одно процесорної обчислювальної системи, інтенсивність обслуговування $\mu = 10 \frac{\text{запит}}{\text{хв}}$, а тривалість обмірковування – випадкова величина, розподілена за експоненціальним законом з параметром

$\lambda = 2 \frac{\text{запит}}{\text{хв}}$. Тобто середній час обмірковування складає 0,5 хвилин (див. рис. 7.4).

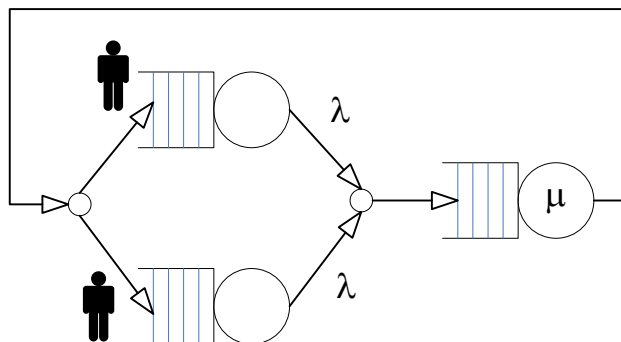


Рис. 7.4. Зімкнута мережа масового обслуговування

Знайти залежність середнього часу реакції системи від числа терміналів n .

Рішення.

Для виконання завдання скористайтесь формулами, наведеними в п. 5.1. Результати розрахунків показані на рис. 7.5.

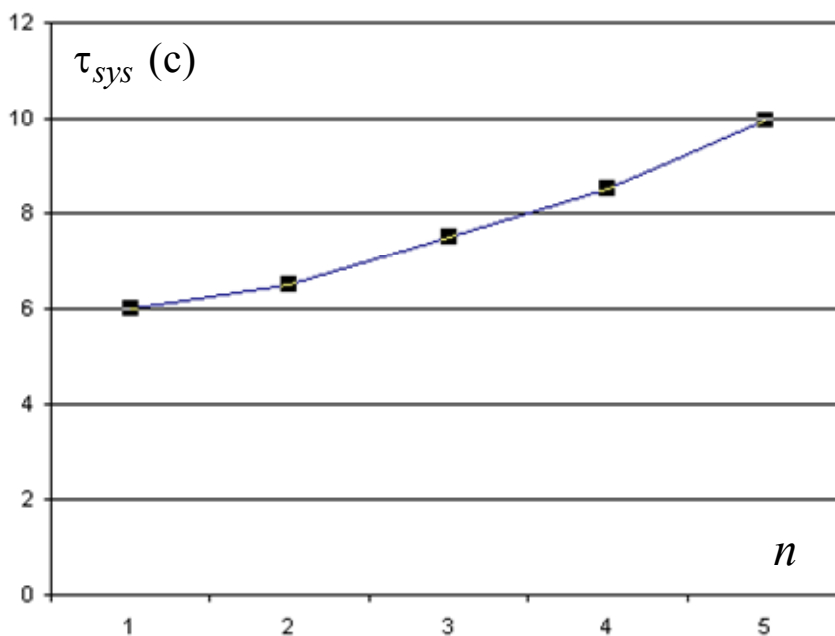


Рис. 7.5. Залежність середнього часу реакції системи від числа терміналів n .

Задача 10. Оцінка пропускної здатності касового залу системи "Експрес". Припустимо, що в залі для продажу і резервування залізничних квитків системи "Експрес" встановлено десять терміналів касирів. Зал пов'язаний каналами передачі даних з обчислювальним комплексом "Експрес". Покладемо, що є один виділений канал в прямому (від касового залу до обчислювального комплексу) і в зворотному (від обчислювального комплексу до касового залу) напрямках. Покладемо, також, що тривалості передачі повідомлень в прямому і зворотному напрямках - випадкові величини, розподілені експонентійно з параметром $\mu_{1,2} = \frac{1}{20} \text{ с}^{-1}$. На обчислювальний комплекс, крім заявок від розглянутого касового залу, надходять повідомлення від інших кас. На обслуговування цього залу виділені ресурси продуктивності обчислювального комплексу, що забезпечують середній час обробки запитів, рівний 60 с. Вважаємо час обробки запитів випадковим, що підкоряється експоненційному закону розподілу (при цьому параметр закону розподілу $\mu_3 = \frac{1}{60} \text{ с}^{-1}$).

І останнє припущення: час, що витрачається касиром на переговори з пасажиром (з'ясування потреб, прийом оплати) - випадковий, розподілений експоненційно з параметром $\lambda = 0,5 \text{ хв}^{-1}$, тобто середній час переговорів - 2 хвилини.

Необхідно: знайти пропускну здатність касового залу, тобто середнє число обслужених пасажирів протягом зміни (8 годин) при повному завантаженні касирів.

Рішення.

Для рішення задачі необхідно скористатись співвідношеннями (5.4) – (5.7) для замкнутих ММО, наведеними в п. 5.2, а також прикладами рішення подібних задач (приклад 5.3, стор. 100). Для рішення задачі є і інші підходи, що описані в [2,14].

При цьому маємо тримати наступні результати.

На рис. 7.6 приведена модель описаної системи у вигляді замкнутої мережі масового обслуговування.

Повне завантаження касирів означає, що в системі постійно перебуває $M = 10$ заявок. Пропускна здатність касового залу визначається при цьому інтенсивністю потоку обслужених заявок λ_0 . В результаті

розрахунків отримуємо $\lambda_0 = 0,0166 \text{ с}^{-1}$. Це означає, що пропускна здатність касового залу складе 60 пас/година, або 480 пас/зміна. На рис 7.7 приведена залежність λ_0 від числа n терміналів касирів в залі продажу квитків. Пропускна здатність при $n \geq 8$ не збільшується. Це означає, що в залі з зазначеними параметрами не доцільно встановлювати більше 8 терміналів.

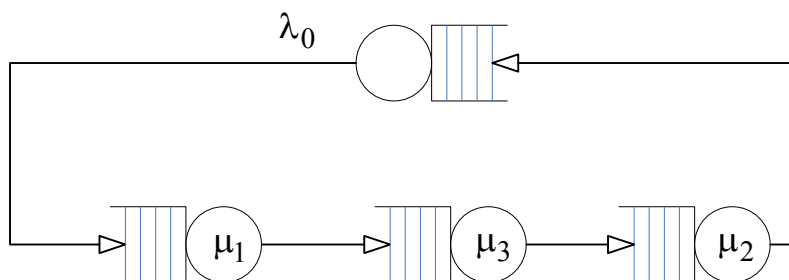


Рис. 7.6. Структурна модель замкненої ММО

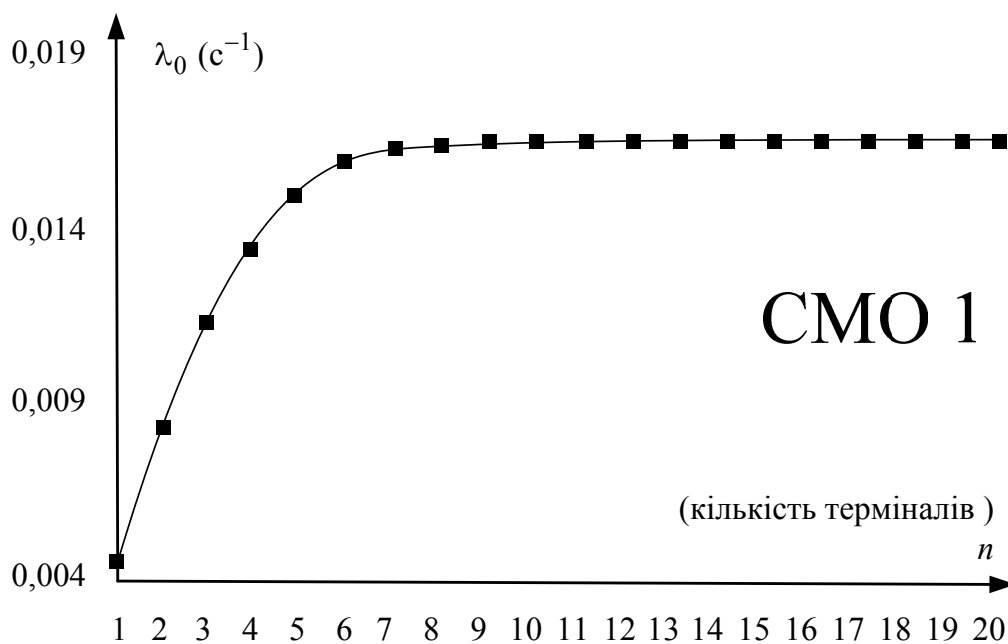


Рис. 7.7. Залежність λ_0 від n - числа терміналів касирів в залі продажу квитків

Збільшення числа терміналів при повному завантаженні касирів призводить до збільшення часу τ_{s0} отримання квитка (без урахування часу очікування в черзі біля каси). На відміну від τ_{s0} , τ_{sys0} враховує середній час переговорів пасажирів з касиром, тобто середній час перебування заявки в групі користувачів. На рис. 7.8 показана залежність τ_{sys0} від числа n терміналів касирів при повному завантаженні системи.

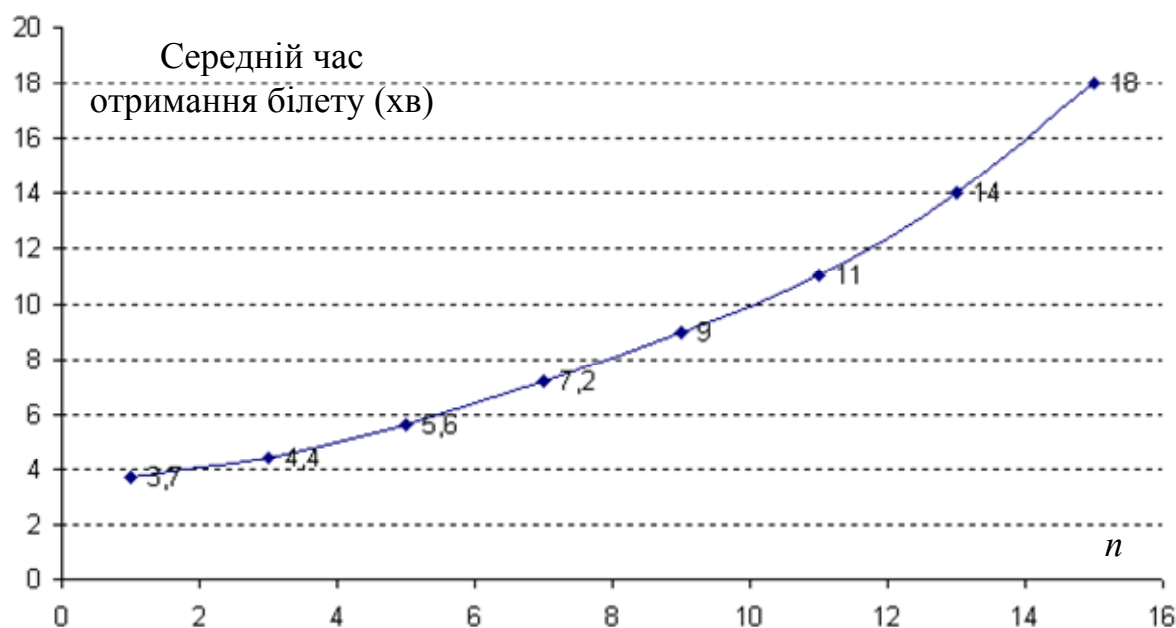


Рис. 7.8. Залежність середнього часу отримання білету від кількості касирів в касовому залі

Виконайте детальне дослідження системи з використанням виразів (5.4) – (5.7).

СПИСОК ЛИТЕРАТУРЫ

1. Stallings, W. Queuing Analysis. [Электрон. ресурс] / W. Stalling. – 2000. – 30 с. - Режим доступа: <http://WilliamStallings.com/>.
2. Основы теории вычислительных систем. Под ред. С. А. Майорова. [Текст] / Учеб.пособие для вузов. – М.: Высш. школа, 1978. – 408 с.
3. Клейнрок, Л. Теория массового обслуживания. Пер.с англ. [Текст] / Л. Клейнрок. – М.: Машиностроение, 1979. – 432 с.
4. Алиев Т.И. Математические методы теории вычислительных систем.– Л.: ЛИТМО, 1979. – 92 с.
5. Шнепс М.А. Системы распределения информации. Методы расчета. - М.: Связь, 1979.
6. Алиев, Т. И. Основы моделирования дискретных систем. Учебное пособие. [Текст] / Т. И. Алиев. – Спб.: СПбГУ ИТМО, 2009. – 363 с.
7. Алиев Т.И. Исследование методов диспетчеризации в цифровых управляющих системах. Уч. пособие. – Л.: ЛИТМО, 1986. – 82 с.
8. Клейнрок, Л. Вычислительные системы с очередями. Пер.с англ. [Текст] / Л. Клейнрок. – М.: Мир, 1979. – 600 с.
9. Шварц М. Сети ЭВМ. Анализ и проектирование: Пер. с англ./ Под ред. В.А.Жожикашвили. – М.: Радио и связь, 1981. – 336 с.: ил.
10. Форд Л.Р., Фалкерсон Д.Р. Потоки в сетях: Пер. с англ./ – М.: Мир, 1966. – 276 с.
11. Авен О.И., Гурин Н.Н., Коган Я.А. Оценка качества и оптимизация вычислительных систем. – М.: Наука, 1982. – 464 с.
12. Расчёты автоматизированных систем управления (на примерах АСУ железнодорожным транспортом) / Г.В. Дружинин, Э.К. Лецкий, В.И. Панкратов и др. Под ред. Г.В. Дружинина. – М. Транспорт, 1985, 223 с.
13. Информационные технологии на железнодорожном транспорте: Учеб. для вузов ж.-д. трансп. / Э.К. Лецкий, В.И. Панкратов, В.В.

Яковлев и др. Под ред Э.К. Лецкого, Э.С. Поддавашкина, В.В. Яковлева и др. – М, УМК МПС России, 2000. – 680 с.

14. Вентцель Е. С. Теория вероятностей. - 10-е изд., стер.. - М.: «Академия», 2005. - 576 с.

Додаткова література

1. Вентцель, Е. С. Исследование операций: Задачи, принципы, методология. Учебное пособие [Текст] / Е. С. Вентцель. - М.: Дрофа, 2004.

2. Гнеденко, Б. В. Введение в теорию массового обслуживания. [Текст] / Б. В. Гнеденко, И. Н. Коваленко. – М.: КомКнига, 2005.

3. Жожикашвили В.А., Вишневский В.М. Сети массового обслуживания. Теория и применение к сетям ЭВМ. – М.: Радио и связь, 1988. – 192 с.: ил.

4. Феррари Д. Оценка производительности вычислительных систем. – М.: Мир, 1981.

5. Рыжиков Ю.И. Теория очередей и управление запасами: Учебник для вузов. – СПб.: Питер, 2001 год. – 384 с.

6. Советов Б.Я., Яковлев С.А. Моделирование систем: Учебник для вузов. – 4-е изд., стер. – М.: Высшая школа, 2005. – 343 с.: ил.

7. Томашевський, В. М. Моделювання систем. [Текст] / В. М. Томашевський. – К.: Видавнича група ВНУ, 2005. – 352 с.

ЕЛЕМЕНТИ ТЕОРІЇ ЙМОВІРНОСТІ

Математичне моделювання дискретних систем зі стохастичним характером функціонування передбачає використання моделей масового обслуговування, описаних у термінах апарату теорії ймовірностей. В даному Додатку [6, 14], не претендуючи на повноту, розглядаються деякі елементи теорії ймовірностей, знання яких необхідно для розуміння та засвоєння матеріалу розділів навчального посібника, пов'язаного з грамотним описом і розрахунком ймовірнісних моделей, а також вдумливим аналізом отриманих результатів.

Д.1. Основні поняття і визначення

Базовими поняттями в теорії ймовірностей є «подія», «ймовірність», «випадкова величина».

Д.1.1. Подія, ймовірність

Подія - всякий факт, який в результаті досліду може відбутися або не відбутися.

Ймовірність події є чисельна міра ступеня об'єктивної можливості цієї події.

Припустимо, що розглядається деякий дослід або явище, в якому в залежності від випадку відбувається або не відбувається деяка подія А.

Якщо умови досліду можуть бути відтворені багато разів, так що в принципі здійснюється ціла серія однакових і незалежних один від

одного випробувань, то ймовірність події A може бути обчислена за такою формулою:

$$P(A) = m / n,$$

де n - загальне число взаємно виключаючих один одного випадків;
 m - число результатів, які призводять до настання події A .

Ймовірність може приймати значення від 0 до 1.

Подія, ймовірність якої дорівнює 0, називається **неможливою**, а подія, ймовірність якої дорівнює 1, називається **достовірною**.

Кілька подій утворюють **повну групу подій**, якщо в результаті досліду має неодмінно з'явитися хоча б одна з них.

Кілька подій називаються **несумісними** в даному досліді, якщо ніякі два з них не можуть з'явитися разом.

Кілька подій називаються **рівноможливими** в даному досліді, якщо жодна з цих подій не є об'єктивно більш можливою, ніж інша.

Події називаються **незалежними**, якщо поява однієї з них не залежить від того, чи відбулися інші події.

Д.1.2. Випадкова величина

Випадковою величиною називається величина, яка може приймати те чи інше значення, невідоме заздалегідь.

Випадкові величини можуть бути двох типів:

- дискретні (перервні), що приймають тільки відокремлені один від одного значення, які можна пронумерувати;
- безперервні (аналогові), які можуть приймати будь яке значення з деякого проміжку.

Прикладами дискретних випадкових величин можуть служити:

- кількість завдань, які виконуються обчислювальною системою (ОС) за день;
- кількість звернень до зовнішньої пам'яті в процесі вирішення задачі;
- кількість повідомлень, переданих в комп'ютерній мережі за одиницю часу, і т.і.

Прикладами безперервних випадкових величин є:

- інтервали часу між моментами надходження в ОС запитів на вирішення задач або між моментами формування повідомлень, переданих в телекомунікаційну мережу;
- час виконання завдань в ОС і т.і.

Іноді випадкові величини, які мають дискретну природу, розглядаються як безперервні. Така заміна виправдана в тих ситуаціях, коли випадкова величина приймає безліч значень, які незначно відрізняються одне від одного, так що заміна дискретної випадкової величини безперервною практично не впливає на результати розрахунків. Наприклад, час передачі пакета по каналу зв'язку в обчислювальній мережі, який визначається як відношення довжини переданого пакета (в бітах) до пропускної здатності каналу зв'язку (біт/с), яке є дискретною випадковою величиною, зазвичай розглядається як безперервна випадкова величина, що змінюється, в загальному випадку, в інтервалі від нуля до деякого граничного значення, визначається максимально можливою довжиною пакета.

Випадкові величини часто позначають великими буквами, а їх можливі значення - відповідними малими літерами. Наприклад, випадкова величина X - число звернень до накопичувача на магнітному диску в процесі виконання завдання в обчислювальній системі - може приймати значення $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3, \dots$.

Д.2. Закони розподілення випадкових величин

Математичний опис випадкових величин припускає завдання закону розподілення, що встановлює відповідність між значеннями випадкової величини та ймовірністю їх появи.

Розглянемо дискретну випадкову величину X , що приймає значення $x_1, x_2, \dots, x_n$. Величина X може прийняти кожне з цих значень з деякою вірогідністю. Позначимо через p_i ($i = \overline{1, n}$) ймовірність того, що випадкова величина X прийме значення x_i : $p_i = P(X = x_i)$. Якщо в результаті досліду величина X приймає тільки одне з цих значень, то маємо повну групу несумісних подій і сума ймовірностей всіх можливих значень випадкової величини дорівнює одиниці:

$$\sum_{i=1}^n p_i = 1.$$

Ця сумарна ймовірність якимось чином розподілена між окремими значеннями. Випадкова величина буде повністю описана з ймовірнісної точки зору, якщо ми задамо це розподілення, тобто встановимо так званий закон розподілу.

Законом розподілу випадкової величини називається всяке співвідношення, що встановлює зв'язок між можливими значеннями випадкової величини і відповідними їм ймовірностями. Про випадкову величину говорять, що вона підпорядкована цьому закону розподілення.

Д.2.1. Закон розподілу дискретної випадкової величини

Закон розподілу дискретної випадкової величини X (дискретний закон розподілу), що приймає значення $x_1, x_2, \dots, x_n$, може бути заданий одним із таких способів:

- аналітично у вигляді математичного виразу, що відображає залежність ймовірності від значення випадкової величини:

$$p_i = f(x_i) \quad (i = 1, n);$$

- таблично у вигляді ряду розподілення випадкової величини, в якому перераховані можливі значення випадкової величини і відповідні їм ймовірності:

Значення випадкової величини X	x_1	x_2	$\dots$	x_n
Вірогідність P	p_1	p_2	$\dots$	p_n

- графічно у вигляді багатокутника розподілу, де по осі абсцис відкладаються можливі значення випадкової величини, а по осі ординат ймовірності цих значень (рис. Д.2.1).

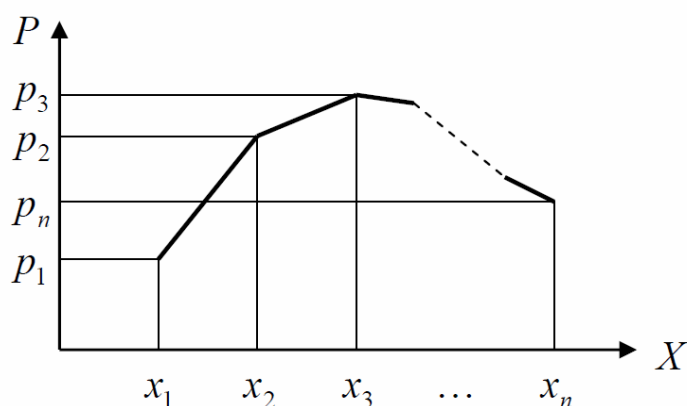


Рис. Д.2.1 Багатокутник розподілення

Графічне представлення закону розподілення дискретної випадкової величини має наочність і дозволяє судити про близькість до того чи іншого типового закону.

Як приклади дискретних законів розподілу нижче розглядаються закони розподілення Пуасона і геометричний, які широко використовуються в теорії масового обслуговування. Але спочатку – про закони розподілу безперервної випадкової величини.

Д.2.2. Закон розподілу безперервної випадкової величини

Для безперервної випадкової величини неможливо поставити закон розподілу в тому вигляді, в якому він задається для дискретної величини, оскільки безперервна випадкова величина має нескінченну кількість можливих значень, що суцільно заповнюють деякий проміжок, і ймовірність появи будь-якого конкретного значення дорівнює нулю.

У зв'язку з цим, для опису безперервних випадкових величин використовується інший спосіб встановлення відповідності між значеннями випадкової величини і ймовірностями їх появи у вигляді функції розподілу ймовірностей.

Функція розподілу ймовірностей (або просто функція розподілу) $F(x)$ випадкової величини X являє собою ймовірність того, що випадкова величина X прийме значення менше, ніж деяке задане значення x :

$$F(x) = P(X < x). \quad (\text{Д.2.1})$$

Функція розподілу випадкової величини X , що приймає будь-які значення з деякого інтервалу, може бути представлена:

- аналітично у вигляді математичного виразу (Д.2.1), відбиває залежність ймовірності від значення випадкової величини;
- графічно у вигляді безперервної функції (рис.Д.2.2, а), відображаючої залежність (Д.2.1), або у вигляді гістограми функції розподілу (рис.Д.2.2, б), отриманої експериментально, наприклад в процесі імітаційного моделювання, і представляє собою дискретний графік, в якому по осі абсцис відкладаються частотні інтервали, що охоплюють всі можливі значення випадкової ве-

личини, а по осі ординат - накопичена частота потрапляння випадкової величини в ці частотні інтервали.

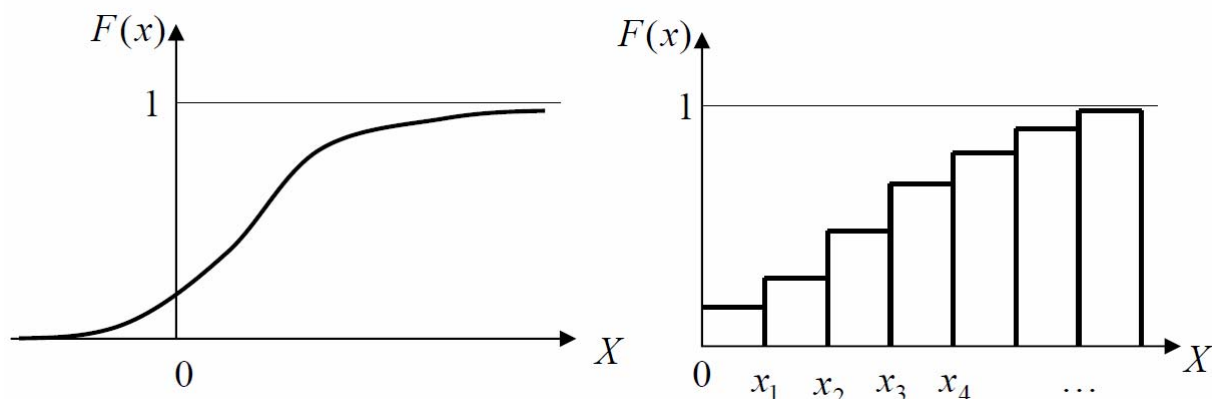


Рис. Д.2.2. Графік (а) та гістограма функції розподілу

Накопичена частота потрапляння в i -й частотний інтервал визначається відношенням кількості випадкових величин, значення яких знаходяться в інтервалі $(-\infty; x_i)$, до загальної кількості випадкових величин, отриманих в процесі експериментів.

Властивості функції розподілу:

- функція розподілу $F(x)$ є неспадною функцією свого аргументу, тобто якщо $x_j > x_i$, то $F(x_j) \geq F(x_i)$;
- $F(-\infty) = 0$;
- $F(+\infty) = 1$.

Якщо випадкова величина визначена тільки в області позитивних значень, її функція розподілу дорівнює нулю на всьому проміжку від мінус нескінченності до нуля.

Ймовірність того, що випадкова величина прийме значення з деякого інтервалу (a, b) , визначається через функцію розподілу як

$$P(a < x < b) = F(b) - F(a).$$

Функція розподілу $F(x)$ є **універсальною** характеристикою випадкової величини і існує як для **безперервних**, так і для **дискретних** величин. Функція розподілу дискретної випадкової величини X , що приймає значення $x_1, x_2, \dots, x_i, \dots$ визначається як

$$F(x_m) = P(X < x_m) = \sum_{i=1}^{m-1} p_i,$$

де p_i - імовірність того, що випадкова величина X прийме значення x_i .

На практиці замість функції розподілу частіше використовують інший спосіб представлення закону розподілу неперервної випадкової величини у вигляді **щільності розподілу ймовірностей**, яка на відміну від функції розподілу має більшу наочність і дозволяє отримати уявлення про близькість того чи іншого розподілу до одного з відомих теоретичних розподілів, який має аналітичний вираз.

Щільність розподілу ймовірностей $f(x)$ визначається як похідна від функції розподілу $F(x)$ по x :

$$f(x) = F'(x) = \frac{dF(x)}{dx}.$$

Розмірність щільності розподілу $f(x)$ зворотня розмірності випадкової величини, в той час як функція розподілу $F(x)$, як всяка імовірність, є величина *безрозмірна*.

Щільність розподілу неперервної випадкової величини X , як і функція розподілу, може бути представлена:

- аналітично у вигляді математичного виразу $y = f(x)$;
- графічно у вигляді безперервної функції (графіка), відображає залежність $y = f(x)$ (рис.Д.2.3, а), або у вигляді гістограми щільності розподілу, в якій на відміну від гістограми функції розподілу по осі ординат відкладається частота (або число) влучень випадкової величини в кожний з частотних інтервалів (рис.Д.2.3, б).

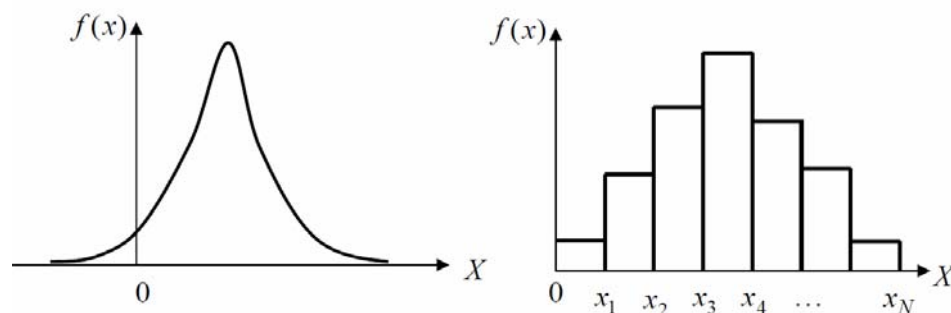


Рис. Д.2.3. Графік (а) і гістограма (б) щільності розподілу

Властивості щільності розподілу:

- щільність розподілу є функція невід'ємна: $f(x) \geq 0$;
- інтеграл в нескінченних межах від щільності розподілу дорівнює одиниці:

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Функція і щільність розподілу випадкової величини однозначно пов'язані між собою. Зокрема, функція розподілу визначається через щільність розподілу таким чином:

$$F(x) = \int_{-\infty}^x f(x)dx. \quad (\text{Д.2.2})$$

Тоді ймовірність того, що випадкова величина прийме значення з деякого інтервалу (a, b) , може бути визначена через щільність розподілу як

$$P(a < x < b) = \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx.$$

Таким чином, закон розподілу випадкової величини (безперервний закон розподілу) може бути заданий у вигляді:

- **функції розподілу** $F(x)$ випадкової величини X , званої також **інтегральним законом розподілу**;
- **щільності розподілу** $f(x)$ випадкової величини X , званої також **диференціальним законом розподілу**.

Д.3. Числові характеристики випадкових величин

Числові характеристики дозволяють висловити у стислій формі найбільш суттєві особливості розподілу випадкової величини, наприклад:

- середнє значення, біля якого групуються можливі значення випадкової величини;
- ступінь розкиданості цих значень щодо середнього;
- асиметрію (або «скошеність») щільності розподілу;
- «крутість», тобто гостро вершинні або плоско вершинні щільності розподілу і так далі.

У теорії ймовірностей використовуються різні числові характеристики, які мають різне призначення і різні області застосування. З них на практиці найбільш часто застосовуються *початкові* і *центральні* моменти різних порядків, кожен з яких описує ту чи іншу властивість розподілу. *Початкові моменти* розглядаються відносно *початку координат*, а *центральні моменти* - відносно *середнього значення* (математичного очікування), тобто центру розподілу.

У загальному випадку для опису випадкової величини використовується нескінченна безліч початкових і центральних моментів. Між числовими моментами і законом розподілу випадкової величини існує взаємна відповідність, яка означає, що, знаючи закон розподілу, можна обчислити будь-які моменти, число яких нескінченно. У той же час, знаючи кінцеве число початкових або центральних моментів, можна шляхом апроксимації підібрати закон розподілу випадкової величини у вигляді функції або щільності розподілу, причому, чим більше відомо моментів, тим точніше апроксимація закону розподілу.

На практиці зазвичай обмежуються застосуванням декількох перших початкових або центральних моментів, що виявляється цілком достатнім для отримання коректних кінцевих результатів.

Д.3.1. Початкові моменти

Покладемо, що випадкова величина X описується імовірностями $p_1, p_2, \dots, p_n$ появи значень $x_1, x_2, \dots, x_n$, якщо X - дискретна величина, і

щільністю розподілу $f(x)$ $x \in (-\infty; +\infty)$, якщо X – безперервна величина.

Початковий момент s -го порядку $\alpha_s[X]$ випадкової величини X визначається наступним чином ($s = 1, 2, \dots$):

$$\alpha_s[X] = \begin{cases} \sum_{i=1}^n x_i^s p_i & \text{– для дискретної випадкової величини;} \\ \int_{-\infty}^{+\infty} x^s f(x) dx & \text{– для неперервної випадкової величини.} \end{cases}$$

Перший початковий момент $\alpha_1[X]$ випадкової величини X :

$$\alpha_1[X] = \begin{cases} \sum_{i=1}^n x_i p_i & \text{– для дискретної випадкової величини;} \\ \int_{-\infty}^{+\infty} x f(x) dx & \text{– для неперервної випадкової величини.} \end{cases}$$

називається математичним очікуванням або середнім значенням випадкової величини і позначається $M[X]$: $M[X] = \alpha_1[X]$. Математичне очікування характеризує становище випадкової величини на числовій осі, тобто показує деяке середнє ймовірнісне (не плутати зі середнім арифметичним) значення, біля якого групуються всі можливі значення випадкової величини.

Другий початковий момент $\alpha_2[X]$ випадкової величини X характеризує *розсіювання*, тобто *розкид (віддаленість)* значень випадкової величини *відносно початку координат*, і має розмірність квадрата випадкової величини.

Д.3.2. Центральні моменти

Центральний момент s -го порядку $\beta_s[X]$ випадкової величини X визначається наступним чином ($s = 1, 2, \dots$):

$$\beta_s[X] = \begin{cases} \sum_{i=1}^n (x_i - M[X])^s p_i & \text{— для дискретної випадкової величини;} \\ \int_{-\infty}^{+\infty} (x - M[X])^s f(x) dx & \text{— для неперервної випадкової величини.} \end{cases}$$

Різниця між значеннями випадкової величини та її математичним очікуванням ($X - M[X]$) являє собою відхилення випадкової величини X від її математичного очікування і називається центрованою випадковою величиною. Тоді центральний момент s -го порядку випадкової величини X можна визначити як математичне очікування s -ого ступеня відповідної центрованої випадкової величини:

$$\beta_s[X] = M[(X - M[X])^s].$$

Для будь якої випадкової величини *центральний момент першого порядку дорівнює нулю*, так як математичне очікування центрованої випадкової величини завжди дорівнює нулю.

Другий центральний момент називається **дисперсією** випадкової величини і позначається $D[X]$: $D[X] = \beta_2[X]$.

Дисперсія обчислюється за формулами:

$$D[X] = \begin{cases} \sum_{i=1}^n (x_i - M[X])^2 p_i & \text{— для дискретної випадкової величини;} \\ \int_{-\infty}^{+\infty} (x - M[X])^2 f(x) dx & \text{— для неперервної випадкової величини.} \end{cases}$$

Можна показати, що дисперсія і другий початковий момент пов'язані наступною залежністю:

$$D[X] = \alpha_2[X] - (M[X])^2. \quad (\text{Д.2.3})$$

Дисперсія випадкової величини, як і другий початковий момент, характеризує розкид значень випадкової величини, але, на відміну від другого початкового моменту, відносно математичного очікування, і має розмірність квадрата випадкової величини.

При вирішенні різних завдань зручно користуватися характеристикою розкиду, розмірність якого збігається з розмірністю випадкової величини. Такою характеристикою є середньоквадратичне відхилення $\sigma[X]$, яке визначається як корінь квадратний з дисперсії:

$$\sigma[X] = \sqrt{D[X]}.$$

В якості *безрозмірної* характеристики розкиду випадкових величин, визначених в області позитивних значень, часто використовують **коефіцієнт варіації** $v[X]$, який визначається як відношення середньоквадратичного відхилення до математичного очікування:

$$v[X] = \frac{\sigma[X]}{M[X]} \text{ за умови, що } M[X] > 0.$$

Застосування числових характеристик істотно полегшує вирішення багатьох імовірнісних задач, зокрема, при вирішенні складних задач, коли використання законів розподілів призводить до громіздких викладок і не дозволяє отримати результати в явному вигляді. Дуже часто вдається вирішити задачу до кінця, залишаючи в стороні закони розподілу і оперуючи одними числовими характеристиками. Якщо в задачі фігурує велика кількість випадкових величин, то для вичерпного судження про результуючі закони розподілу не потрібно знати закони розподілу окремих випадкових величин, що фігурують в задачі, а достатньо знати лише деякі числові характеристики цих величин.

Крім того, на практиці (і в повсякденному житті) рідко оперують законом розподілу для опису конкретних фізичних величин, вважаючи за краще використовувати такі поняття як середнє значення і, в деяких випадках, розкид значень випадкової величини або мінімальне і максимальне значення. Дійсно, наприклад для пасажирів, очікуючих на зупинці автобус, становить інтерес закон розподілу інтервалів між автобусами. Більш важливим і зрозумілим є знання про середній або максимальний інтервал. У той же час при моделюванні транспортних потоків для отримання коректних і достовірних результатів може знадобитися знання закону розподілу або, принаймні, кількох моментів розподілу шуканих інтервалів.

Альтернативою випадковій величині є не випадкова величина, звана детермінованою. У деяких задачах детерміновану величину $X = x$ розглядають як випадкову, яка з імовірністю $p = 1$ приймає одне і те ж значення x .

Д.4. Твірна функція і перетворення Лапласа

Аналітичне дослідження складних систем з випадковим характером функціонування в багатьох випадках можна істотно спростити, якщо дії над функціями розподілів замінити діями над відповідними твірними функціями і перетвореннями Лапласа.

Твірні функції використовуються для дискретних, а перетворення Лапласа - для безперервних випадкових величин.

Д.4.1. Твірна функція

Твірною функцією розподілу $p_k = P(X = k)$ дискретної випадкової величини X , що приймає невід'ємні цілочислові значення $k = 0, 1, 2, \dots$, називається ряд

$$X^*(z) = \sum_{k=0}^{\infty} z^k p_k \quad |z| \leq 1 \quad (\text{Д.2.4})$$

Розподіл ймовірностей однозначно визначається своєю твірною функцією:

$$p_k = \frac{1}{k!} X^{*(k)}(0), \quad X^{*(k)}(0) = \frac{d^k}{dz^k} X^*(z) \Big|_{z=0} \quad (k = 0, 1, 2, \dots).$$

На основі твірної функції (Д.2.4) можуть бути обчислені початкові і центральні моменти випадкової величини, зокрема математичне очікування і дисперсія визначаються як

$$M[X] = X^{*(1)}(1); \quad D[X] = X^{*(2)}(1) + X^{*(1)}(1) - [X^{*(1)}(1)]^2. \quad (\text{Д.2.5})$$

Твірна функція $X^*(z)$ суми $X = X_1 + X_2 + \dots + X_n$ незалежних випадкових величин дорівнює добутку твірних функцій доданків:

$$X^*(z) = X_1^*(z)X_2^*(z)\dots X_n^*(z).$$

Д.4.2. Перетворення Лапласа

Перетворенням Лапласа щільності розподілу $f(x)$ невід'ємної безперервної випадкової величини X називається функція

$$F^*(s) = \int_0^{\infty} e^{-sx} f(x) dx \quad (s \geq 0). \quad (\text{Д.2.6})$$

Щільність розподілу однозначно визначається своїм перетворенням Лапласа.

Диференціюючи перетворення Лапласа по s в точці $s = 0$, можна визначити *початкові моменти* випадкової величини:

$$\alpha_k[X] = \frac{(-1)^k}{k!} \left. \frac{d^k F^*(s)}{ds^k} \right|_{s=0} \quad (k = 1, 2, \dots) \quad (\text{Д.2.7})$$

Перетворення Лапласа $F^*(s)$ суми $X = X_1 + X_2 + \dots + X_n$ незалежних випадкових величин дорівнює добутку перетворень Лапласа доданків:

$$F^*(s) = F_1^*(s)F_2^*(s)\dots F_n^*(s).$$

Д.5. Типові розподіли випадкових величин

Моделювання технічних систем з дискретним характером функціонування припускає застосування різних законів розподілу, як дискретних, так і безперервних випадкових величин. Нижче розглядаються типові закони розподілів випадкових величин, які широко використовуються в моделях масового обслуговування.

В якості законів розподілів дискретних випадкових величин найбільш широко використовуються:

- розподіл Пуассона;
- геометричний розподіл.

Оскільки в математичних моделях масового обслуговування безперервною випадковою величиною зазвичай є час, найбільший інтерес представляють закони розподілів безперервних випадкових величин, визначених в області позитивних значень:

- рівномірний;
- експоненційний;
- Ерланга;
- Ерланга нормований;
- гіперекспоненційний;
- гіперерланговській.

Д.5.1. Розподіл Пуасона

Дискретна випадкова величина X розподілена за законом Пуасона, якщо ймовірність $P(X = k)$ того, що вона прийме певне значення $x_k = k$ виражається формулою:

$$p_k = P(X = k) = \frac{a^k}{k!} e^{-a} \quad (k = 0, 1, 2, \dots), \quad (\text{Д.2.8})$$

де a - деяка позитивна величина, яка називається параметром розподілу Пуассона.

На рис.Д.2.4 показані багатокутники розподілу Пуассона для трьох значень параметра розподілу: $a = 0,5$; $a = 1$; $a = 2$.

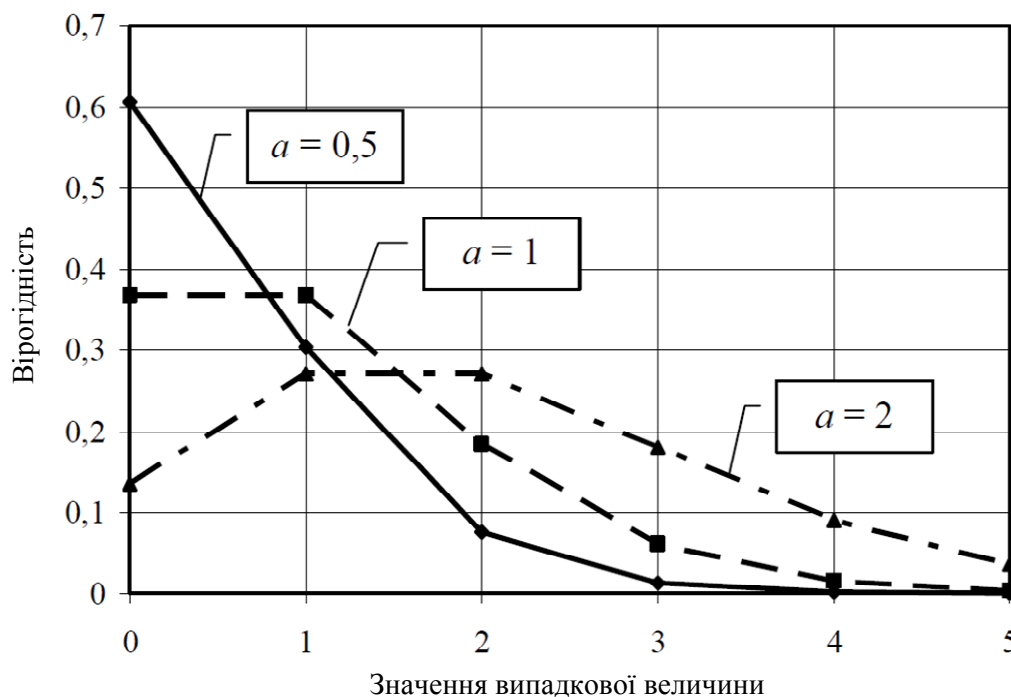


Рис. Д.2.4. Багатокутники розподілу Пуассона

Твірна функція розподілу Пуассона:

$$X^*(z) = e^{-a(1-z)} \quad (0 \leq z \leq 1).$$

Д.5.2. Геометричний розподіл

Розподіл дискретної випадкової величини $X = k$ виду

$$p_k = P(X = k) = \rho^k (1 - \rho) \quad (k = 0, 1, 2, \dots), \quad (\text{Д.2.9})$$

де ρ - параметр розподілу ($0 < \rho < 1$), називається геометричним.

Розподіл (Д.2.9) може бути записано в дещо іншому вигляді, якщо параметр ρ замінити параметром $\gamma = 1 - \rho$:

$$p_k = \gamma(1 - \gamma)^k \quad (0 < \gamma < 1; \quad k = 0, 1, 2, \dots).$$

На рис.Д.2.5 показані багатокутники геометричного розподілу для трьох значень параметра розподілу: $\gamma = 0,2$; $\gamma = 0,5$; $\gamma = 0,8$.

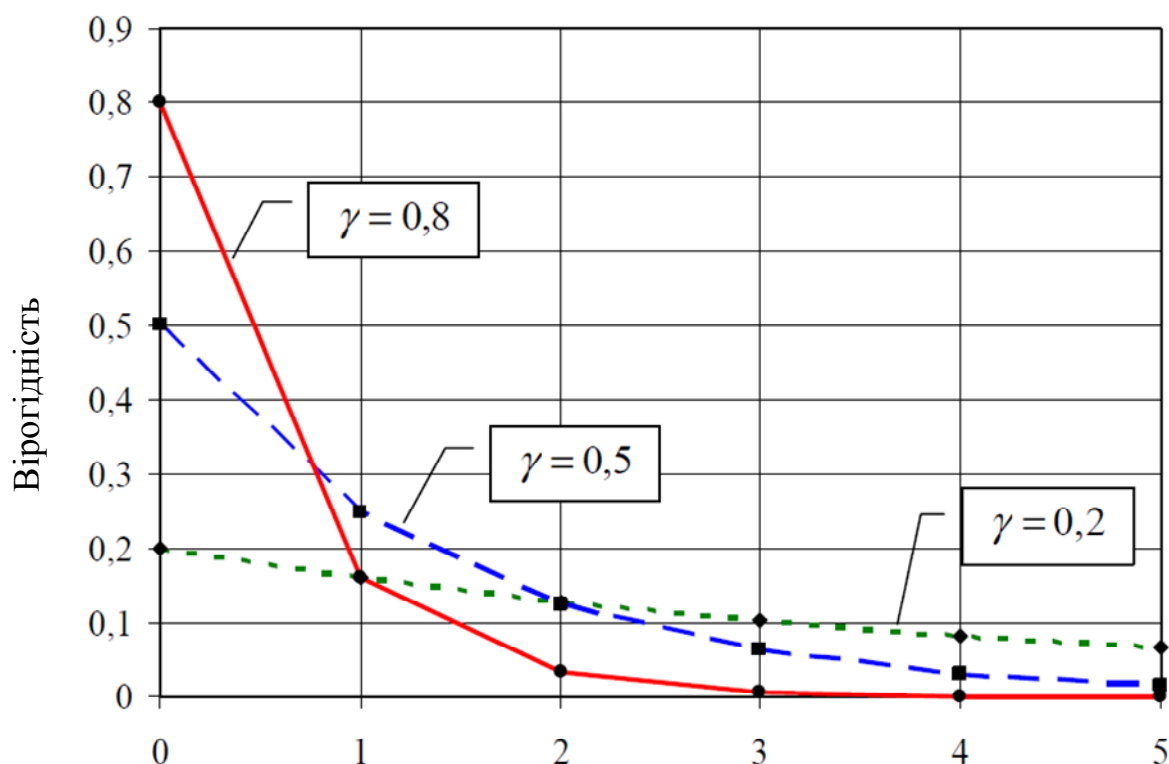


Рис. Д.2.5. Багатокутники геометричного розподілу

Твірня функція геометричного розподілу:

$$X^*(z) = \frac{1-\rho}{1-\rho z} \quad \text{або} \quad X^*(z) = \frac{\gamma}{1-(1-\gamma)z} \quad (0 \leq z \leq 1).$$

Д.5.3. Рівномірний закон розподілу

Безперервна випадкова величина X розподілена рівномірно в інтервалі $(a;b)$, де $a < b$, якщо функція $F(x)$ і щільність $f(x)$ розподілу відповідно мають вигляд:

$$F(x) = \begin{cases} 0 & \text{при } x < a; \\ \frac{x-a}{b-a} & \text{при } a < x < b; \\ 1 & \text{при } x > b. \end{cases}$$

Значення ви

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{при } a < x < b; \\ 0 & \text{при } x > b. \end{cases}$$

На рис. Д.2.6 показані функція і щільність рівномірного розподілу.

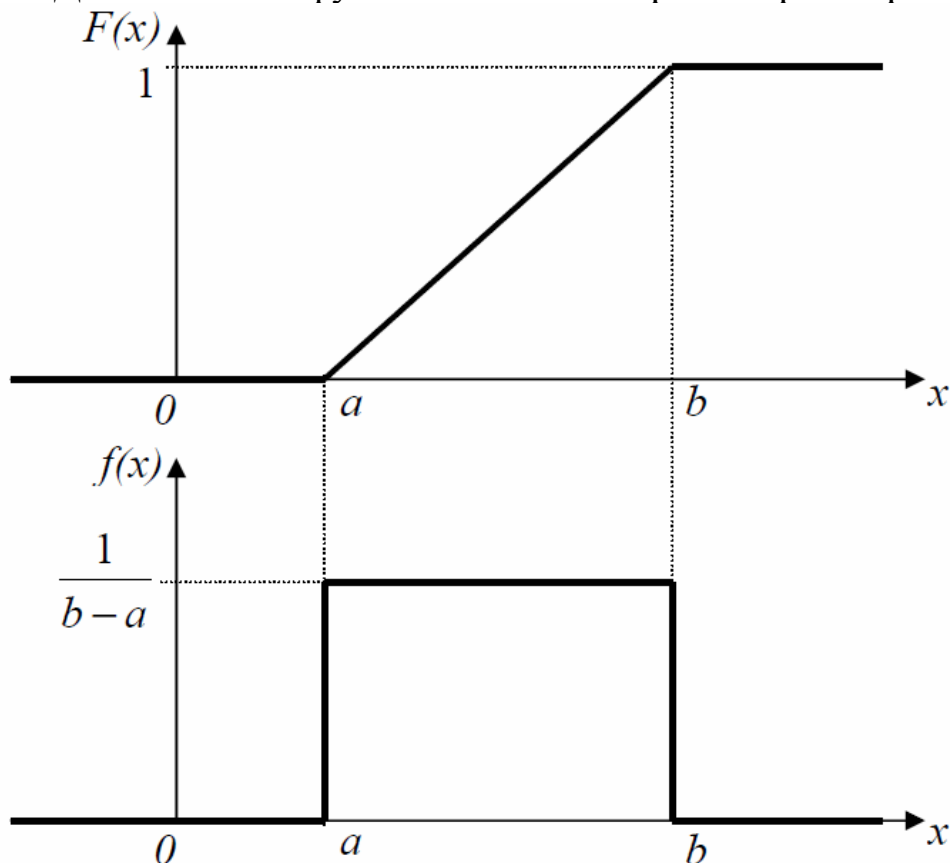


Рис. Д.2.6. Функція і щільність рівномірного розподілу

Д.5.4. Експоненційний закон розподілу

Безперервна випадкова величина X , яка приймає позитивні значення в нескінченному інтервалі $(0; +\infty)$, розподілена по експоненційному (показовому) закону, якщо функція $F(x)$ і щільність $f(x)$ розподілу відповідно мають вигляд:

$$F(x) = 1 - e^{-\alpha x}, \quad f(x) = \alpha e^{-\alpha x}, \quad (\text{Д.2.12})$$

де $\alpha > 0$ - параметр розподілу; $x \geq 0$ - безперервна випадкова величина.

Чудовою особливістю експоненційного розподілу є те, що його коефіцієнт варіації не залежить від параметра α і завжди дорівнює одиниці: $v_{\text{експ}}[X] = 1$.

На рис. Д.2.7 показані функція і щільність експоненційного розподілу для трьох значень параметра: $\alpha = 0,5$; $\alpha = 1$; $\alpha = 2$.

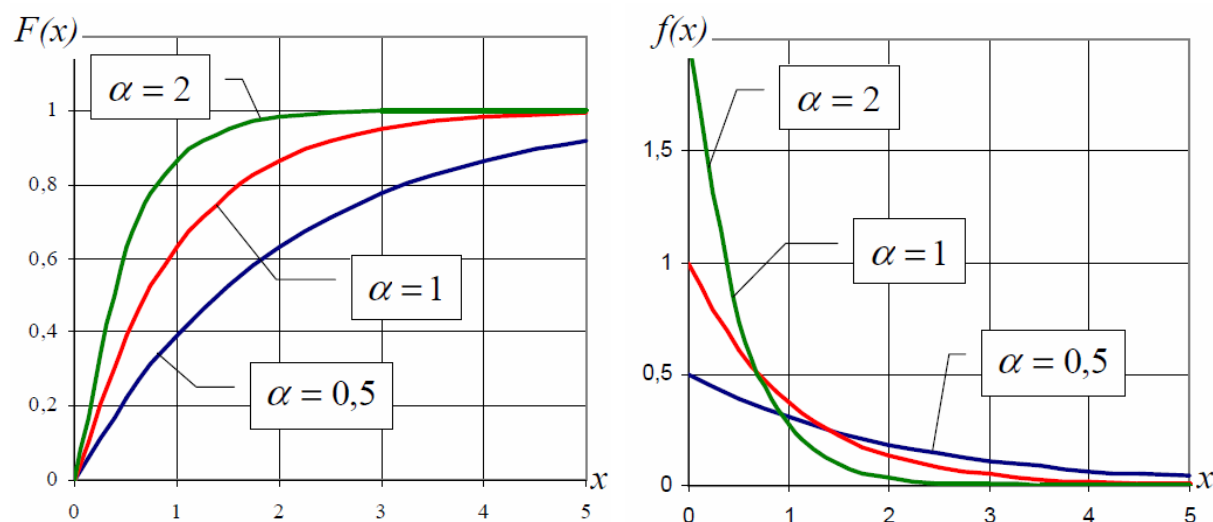


Рис. Д.2.7. Функція та щільність експоненційного розподілу

Перетворення Лапласа експоненційного розподілу

$$F^*(s) = \frac{\alpha}{\alpha + s}. \quad (2.13)$$

Експоненційний розподіл широко застосовується в теорії масового обслуговування при описі випадкових процесів, що відбуваються в моделях масового обслуговування. Це пояснюється тим, що експоненційний розподіл має чудову властивість, яка притаманна тільки цьому розподілу, завдяки якому для багатьох моделей масового обслуговування вдається отримати достатньо прості аналітичні результати в явному вигляді. З цим же розподілом тісно пов'язаний особливий клас дискретних випадкових процесів, званих марковським процесами, в яких переходи між станами не залежать від передісторії процесу і визначаються тільки станом процесу в даний конкретний час. Цю властивість іноді називають властивістю відсутності пам'яті у експоненційному розподілі (точніше, у експоненційно розподілених випадкових величин), а в теорії масового обслуговування використовується

термін «відсутність післядії» (див. п.Д.2.6).

Можливість отримання порівняно простих аналітичних результатів при використанні припущення про експоненційний характер випадкових процесів зумовила появу розглянутих нижче специфічних законів розподілів, що представляють собою композиції експоненційних розподілів і дозволяють спростити вирішення багатьох завдань, пов'язаних з дослідженням моделей масового обслуговування. До них, зокрема, належать такі розподіли: Ерланга, гіперекспоненційне, гіперерлангівське.

Д.5.5. Розподіл Ерланга

Розподілом Ерланга k -го порядку називається розподіл, що описує безперервну випадкову величину X , яка приймає позитивні значення в інтервалі $(0; +\infty)$ і представляє собою суму k незалежних випадкових величин, розподілених по одному і тому ж експоненційному закону з параметром α . Функція і щільність розподілу Ерланга k -го порядку мають вигляд:

$$F_k(x) = 1 - e^{-\alpha x} \sum_{i=0}^{k-1} \frac{(\alpha x)^i}{i!}; \quad f_k(x) = \frac{\alpha(\alpha x)^{k-1}}{(k-1)!} e^{-\alpha x}, \quad (\text{Д.2.14})$$

де α і k - позитивні параметри розподілу ($\alpha \geq 0; k = 1, 2, \dots$); $x \geq 0$ - безперервна випадкова величина.

На рис. Д.2.8 показані щільності розподілу Ерланга при $\alpha = 1$ для трьох значень параметра: $k = 1; k = 2; k = 4$.

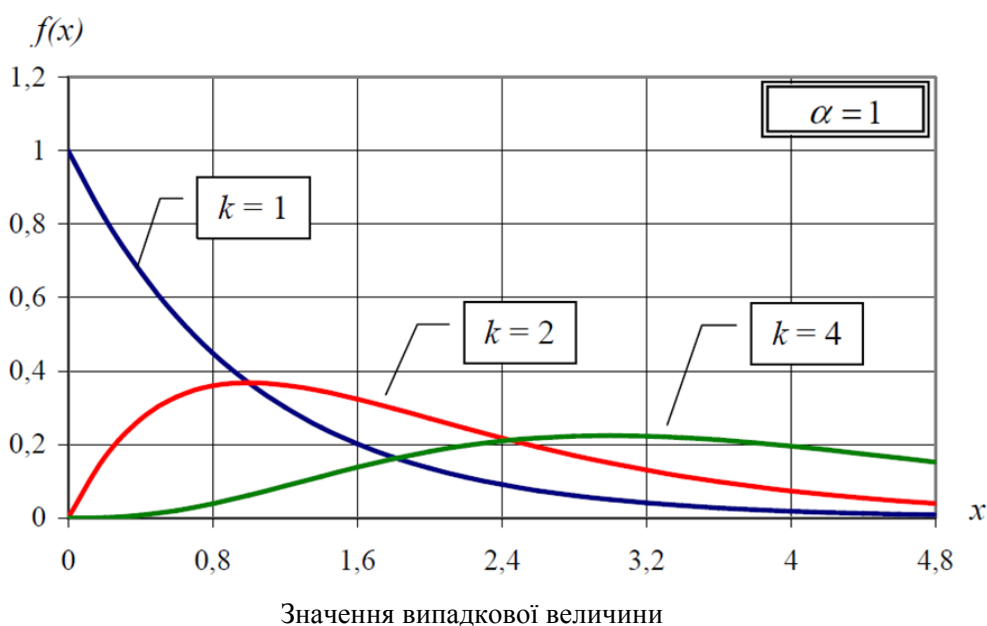


Рис. Д.2.8. Щільність розподілу Ерланга

При $k = 1$ розподіл Ерланга вироджується в експоненційний, а при $k \rightarrow \infty$ - наближається до нормального розподілу.

Перетворення Лапласа розподілу Ерланга k -го порядку

$$F^*(s) = \left(\frac{\alpha}{\alpha + s} \right)^k. \quad (\text{Д.2.15})$$

Оскільки розподіл Ерланга є двопараметричним, то він може використовуватися для апроксимації реальних розподілів по двох перших моментах.

Коефіцієнт варіації розподілу Ерланга залежить від параметра k і приймає значення **менші або рівні одиниці**:

$$v_{Ek}[X] = \frac{1}{\sqrt{k}} \leq 1 \quad (k = 1, 2, \dots).$$

Зазначимо, що математичне очікування розподілу Ерланга залежить від значення параметра k , що створює певні проблеми при апроксимації реальних розподілів законом Ерланга. Ці проблеми відсутні при апроксимації нормованим розподілом Ерланга.

Д.5.6. Нормований розподіл Ерланга

Нормований розподіл Ерланга являє собою розподіл суми k незалежних випадкових величин, кожна з яких розподілена за експоненційним законом з параметром $k\alpha$, що залежить від k . Іншими словами, підсумовуються k експоненційно розподілених випадкових величин, кожна з яких має математичне очікування в k раз менше, ніж вихідне математичне очікування реального розподілу, що призводить до незалежності математичного очікування нормованого розподілу Ерланга від параметра k .

Математичні вирази для функції і щільності нормованого розподілу Ерланга можна отримати з (Д.2.14), замінивши параметр α на $k\alpha$:

$$F_k(x) = 1 - e^{-k\alpha x} \sum_{i=0}^{k-1} \frac{(k\alpha x)^i}{i!}; \quad f(x) = \frac{k\alpha (k\alpha x)^{k-1}}{(k-1)!} e^{-k\alpha x},$$

Коефіцієнт варіації нормованого розподілу Ерланга так само, як і ненормованого, залежить від параметра k і приймає значення **менші або рівні одиниці**:

$$v_{нЕk} = \frac{1}{\sqrt{k}} \leq 1 \quad (k = 1, 2, \dots).$$

На рис. Д.2.9 показані щільності розподілу Ерланга при $\alpha=1$ для трьох значень параметра: $k = 1$; $k = 2$; $k = 16$.

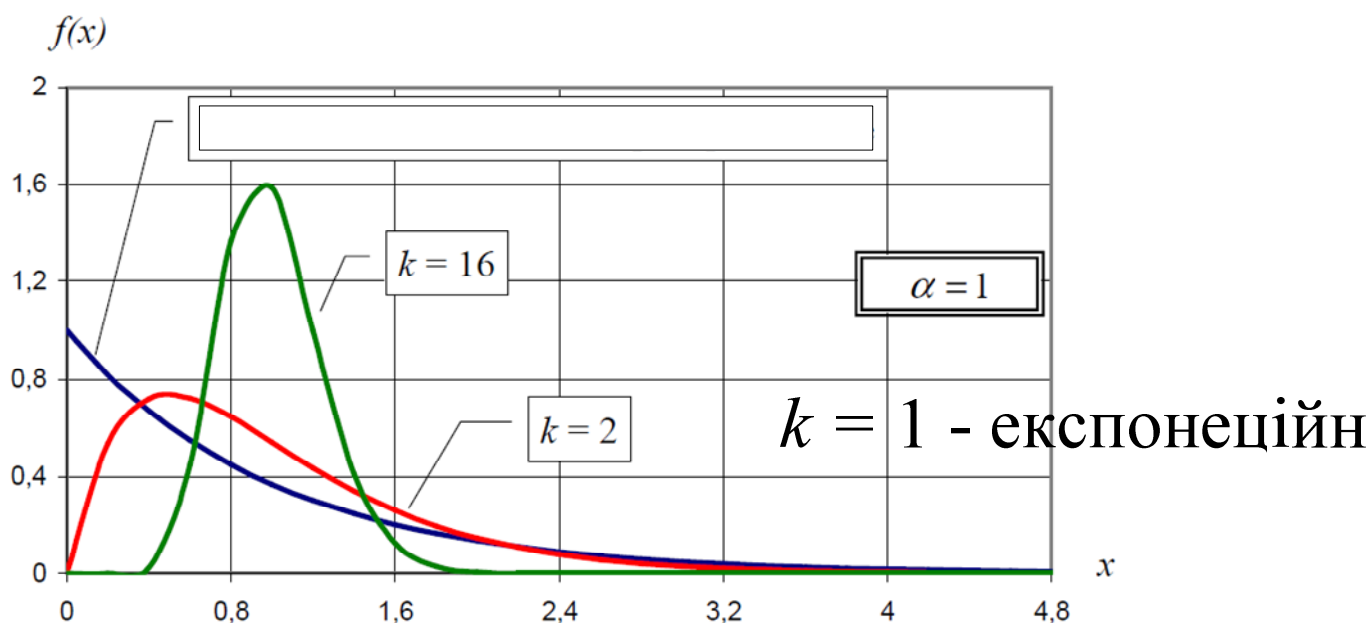


Рис. Д.2.9. Щільність нормованого розподілу Ерланга

Нормований розподіл Ерланга при $k \rightarrow \infty$, на відміну від простого розподілу Ерланга, призводить до *детермінованої величини* $1/\alpha$.

Перетворення Лапласа нормованого розподілу Ерланга

$$F^*(s) = \left(\frac{k\alpha}{k\alpha + s} \right)^k.$$

Д.5.7. Гіперекспоненційний розподіл

У тих випадках, коли якийсь реальний розподіл безперервної випадкової величини, що приймає невід'ємні значення, має **коефіцієнт варіації більше одиниці**, для його апроксимації може використовуватися гіперекспоненційний розподіл.

Як впливає з назви, гіперекспоненційний розподіл деяким чином пов'язан з експоненційним і являє собою адитивну суміш різних експоненційних розподілів. Процес формування випадкових величин з гіперекспоненційним розподілом з експоненційно розподілених випадкових величин може бути представлений наступним чином.

Покладемо, що є n різних генераторів експоненційно розподілених випадкових величин з параметрами $\alpha_1, \dots, \alpha_n$ відповідно з математич-

ними очікуваннями $M_1 = 1/\alpha_1, \dots, M_n = 1/\alpha_n$, причому $\alpha_i \neq \alpha_j$ для всіх $i \neq j$ ($i, j = \overline{1, n}$). Нехай в результаті одного досліду з ймовірністю q_i виробляється тільки одна випадкова величина i -м генератором з параметром α_i ($i = \overline{1, n}$), причому $q_1 + \dots + q_n = 1$. Сукупність випадкових величин, отриманих в результаті проведення безлічі таких дослідів, буде розподілена по гіперекспоненційному закону:

$$\left. \begin{aligned} F(x) &= \sum_{i=1}^n q_i (1 - e^{-\alpha_i x}) = 1 - \sum_{i=1}^n q_i e^{-\alpha_i x}; \\ f(x) &= \sum_{i=1}^n q_i \alpha_i e^{-\alpha_i x}; \end{aligned} \right\} \quad (\text{Д.2.17})$$

Гіперекспоненційний розподіл (Д.2.17) містить $(2n - 1)$ параметрів: $\alpha_1, \dots, \alpha_n, q_1, \dots, q_{n-1}$, оскільки $q_n = 1 - \sum_{i=1}^{n-1} q_i$.

Перетворення Лапласа гіперекспоненційного розподілу

$$F^*(s) = \sum_{i=1}^n q_i \frac{\alpha_i}{\alpha_i + s}.$$

У найпростішому варіанті випадкові величини з гіперекспоненційним розподілом можуть бути отримані з використанням тільки двох експоненційних розподілів: $n = 2$. Тоді функція і щільність гіперекспоненційного розподілу будуть мати вигляд:

$$\left. \begin{aligned} F(x) &= qp(1 - e^{-\alpha_1 x}) + (1 - q)(1 - e^{-\alpha_2 x}); \\ f(x) &= q\alpha_1 e^{-\alpha_1 x} + (1 - q)\alpha_2 e^{-\alpha_2 x} \end{aligned} \right\} \quad (\text{Д.2.18})$$

Зауважимо, що гіперекспоненційний розподіл (Д.2.18) є трипараметричним, тобто містить три незалежних параметри: q, α_1, α_2 ($0 < q < 1; \alpha_1 \geq 0; \alpha_2 \geq 0$). Отже, апроксимація реальних розподілів гіперекспоненційним може здійснюватися за трьома моментами розподілу, а не по двох, як в розподілі Ерланга.

На рис. Д.2.10 показані щільності гіперекспоненційного розподілу випадкової величини X з математичним очікуванням, рівним 1, для

двох значень коефіцієнта варіації: $v[X]=2$ і $v[X]=4$. Параметри розподілу (Д.2.18) мають такі значення:

- $\alpha_1 = 0,183$; $\alpha_2 = 1,506$ для розподілу з $v[X]=2$;
- $\alpha_1 = 0,091$; $\alpha_2 = 4,022$ для розподілу з $v[X]=4$,

причому параметр q однаковий для обох розподілів і дорівнює 0,07.

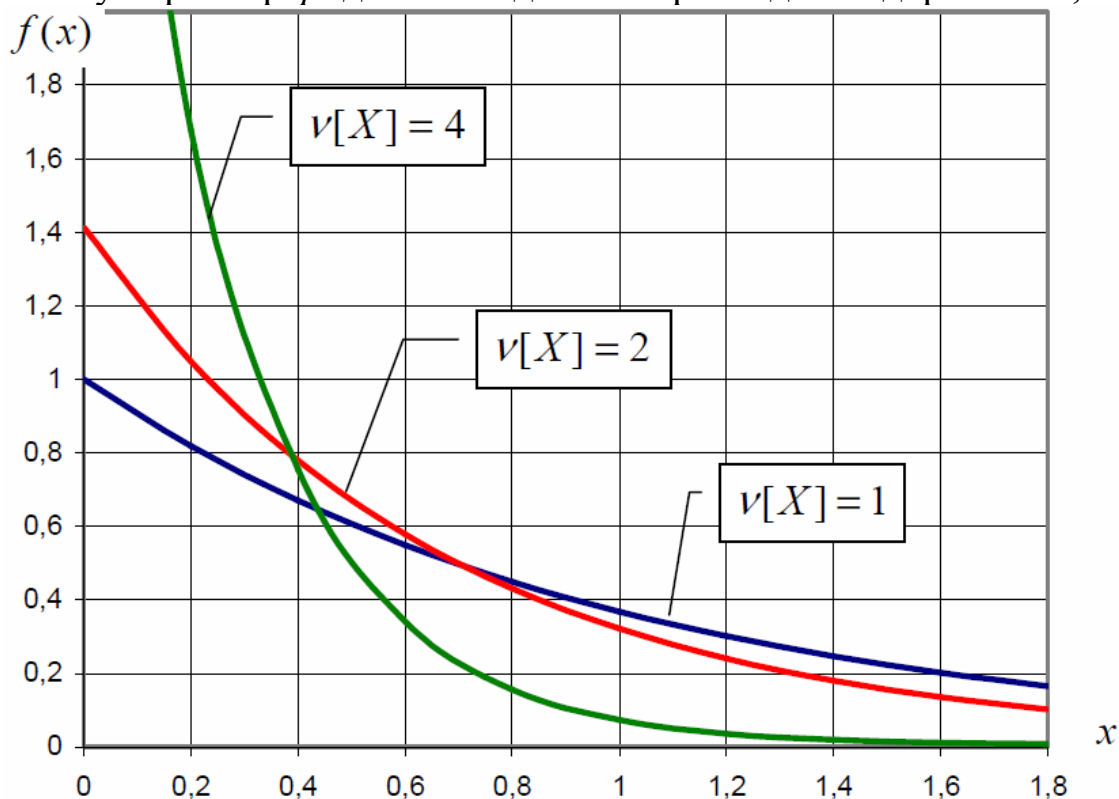


Рис. Д.2.10. Щільність гіперекспоненційного розподілу

Тут же для порівняння показана щільність експоненційного розподілу з тим же математичним очікуванням $M=1$.

Як видно з представлених графічних залежностей, щільність гіперекспоненційного розподілу в порівнянні з експоненційним розподілом характеризується більш різким спадом в області малих значень випадкової величини, причому чим більше коефіцієнт варіації випадкової величини, тим крутіше ця залежність.

Можна показати, що ймовірність появи маленьких значень випадкової величини для гіперекспоненційного розподілу набагато більше ймовірності появи великих значень. Визначимо ймовірність того, що випадкова величина прийме значення менше математичного очікування M . Для цього розрахуємо значення функції розподілу в

точці $x = M$: $\Pr(X < M) = F(x = M) = F(M)$. Тоді для розглянутих вище гіперекспоненційних розподілів отримаємо:

$$\Pr(X < M) = F(M) = \begin{cases} 0,739 & \text{для розподілу з } \nu[X] = 2; \\ 0,919 & \text{для розподілу з } \nu[X] = 4. \end{cases}$$

Таким чином, більше 73% значень випадкової величини, розподіленої за гіперекспоненційним законом з коефіцієнтом варіації, рівним 2, потрапляє в інтервал $(0; M)$ і тільки 27% значень виявляться більше математичного очікування. Для випадкової величини, з коефіцієнтом варіації, рівним 4, ймовірність попадання в інтервал $(0; M)$ ще вище і складає майже 92%. Очевидно, що чим більше коефіцієнт варіації, тим більше вірогідність появи маленьких значень випадкової величини.

Для порівняння обчислимо цю ж ймовірність для експоненційного розподілу: $\Pr(X < M) = F(M) = 0,632$. Таким чином, ймовірність появи маленьких значень експоненційно розподіленої випадкової величини більше ймовірності появи великих значень і становить 63%, але при цьому вона значно менше, ніж при гіперекспоненційному розподілі.

Уявлення про гіперекспоненційний розподіл буде не повним, якщо не звернути уваги на «хвіст» цього розподілу. На рис.Д.2.11 представлений графік щільності гіперекспоненційного розподілу для значень випадкової величини більше 4. Нагадаємо, що математичне очікування випадкової величини дорівнює 1.

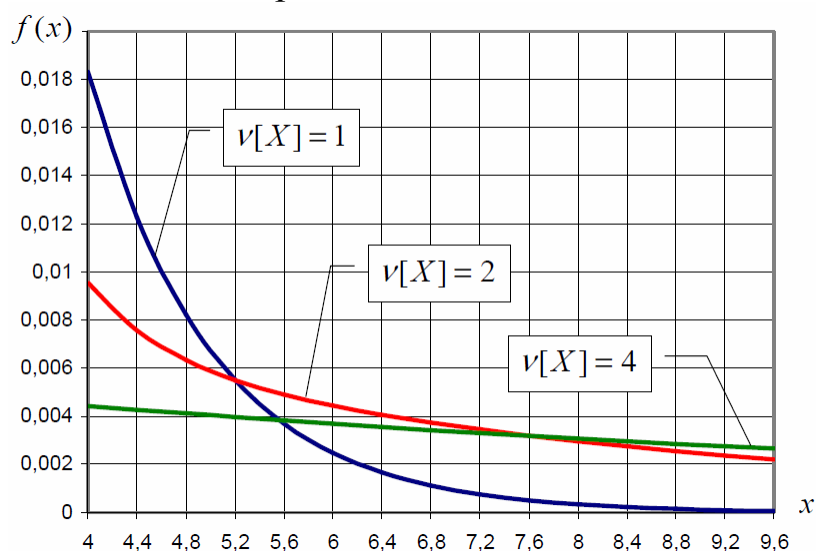


Рис. Д.2.11. «Хвіст» гіперекспоненційного розподілу

З графіка видно, що крива щільності гіперекспоненційного розподілу з коефіцієнтом варіації, рівним 4, має довгий так званий «важкий хвіст», що характеризується малою зміною. Це означає, що при гіперекспоненційному розподілі ймовірність появи великих значень випадкової величини значно вище, ніж, наприклад, для експоненційного розподілу. Таким чином, основна відміна гіперекспоненційного розподілу від експоненційного полягає в тому, що гіперекспоненційний розподіл характеризується більшою ймовірністю появи маленьких значень випадкової величини і, в той же час, більшою ймовірністю появи великих значень випадкової величини.

Д.5.8. Гіперерлангівський розподіл

Гіперерлангівський розподіл являє собою адитивну суміш нормованих розподілів Ерланга і є найбільш загальним розподілом невід'ємних безперервних випадкових величин, оскільки має коефіцієнт варіації в інтервалі від 0 до ∞ .

Складовими гіперерлангівського розподілу, на відміну від гіперекспоненційного, є нормовані розподіли Ерланга.

Щільність гіперекспоненційного розподілу

$$f(x) = \sum_{i=1}^n q_i \frac{k_i \alpha_i (k_i \alpha_i x)^{k_i-1}}{(k_i-1)!} e^{-k_i \alpha_i x} \quad (x \geq 0). \quad (\text{Д.2.19})$$

Перетворення Лапласа гіперерлангівського розподілу

$$F^*(s) = \sum_{i=1}^n q_i \left(\frac{k_i \alpha_i}{k_i \alpha_i + s} \right)^{k_i}.$$

Нижче в таблиці представлені основні числові характеристики розглянутих розподілів дискретних і безперервних випадкових величин:

- математичне очікування $M[X]$;
- другий початковий момент $\alpha_2[X]$;
- дисперсія $D[X]$;
- середньоквадратичне відхилення $\sigma[X]$;
- коефіцієнт варіації $v[X]$.

У графі «Примітки» вказані значення або діапазон зміни параметрів відповідних розподілів.

Нагадаємо, що представлені числові характеристики пов'язані між собою досить простими співвідношеннями:

$$D[X] = \alpha_2[X] - (M[X])^2;$$

$$\sigma[X] = \sqrt{D[X]};$$

$$v[X] = \frac{\sigma[X]}{M[X]}.$$

Числові характеристики розподілів

Розподіл	$M[X]$	$\alpha_2[X]$	$D[X]$	$\sigma[X]$	$v[X]$	Примітки
Пуссона	a	$a(a+1)$	a	$\sqrt{a}$	$1/\sqrt{a}$	$a > 0$
Геометричний	$\frac{1-\gamma}{\gamma}$	$\frac{2(1-\gamma)^2}{\gamma^2}$	$\frac{(1-\gamma)^2}{\gamma^2}$	$\frac{1-\gamma}{\gamma}$	1	$0 < \gamma < 1$
Рівномірний	$\frac{a+b}{2}$	$\frac{a^2 + ab + b^2}{3}$	$\frac{(b-a)^2}{12}$	$\frac{b-a}{2\sqrt{3}}$	$\frac{b-a}{\sqrt{3}(a+b)}$	$b > a$
Експоненційний	$\frac{1}{\alpha}$	$\frac{2}{\alpha^2}$	$\frac{1}{\alpha^2}$	$\frac{1}{\alpha}$	1	$\alpha > 0$
Ерланга	$\frac{k}{\alpha}$	$\frac{k(k+1)}{\alpha^2}$	$\frac{k}{\alpha^2}$	$\frac{\sqrt{k}}{\alpha}$	$\frac{1}{\sqrt{k}}$	$k = 1, 2, \dots$
Ерланга нормований	$\frac{1}{\alpha}$	$\frac{k+1}{k\alpha^2}$	$\frac{1}{k\alpha^2}$	$\frac{1}{\alpha\sqrt{k}}$	$\frac{1}{\sqrt{k}}$	$k = 1, 2, \dots$
Гіперекспоненційний	$\sum_{i=1}^n \frac{q_i}{\alpha_i}$	$2 \sum_{i=1}^n \frac{q_i}{\alpha_i^2}$	$\alpha_2[X] - (M[X])^2$	$\sqrt{D[X]}$	$v[X] \geq 1$	$\sum_{i=1}^n q_i = 1 \quad \alpha_i > 0$
Гіперерлангівський	$\sum_{i=1}^n \frac{q_i}{\alpha_i}$	$\sum_{i=1}^n q_i \frac{k_i + 1}{k_i \alpha_i^2}$	$\alpha_2[X] - (M[X])^2$	$\sqrt{D[X]}$	$v[X] \geq 0$	$\sum_{i=1}^n q_i = 1 \quad \alpha_i > 0$ $k_i = 1, 2, \dots$

Д.6. Практикум. Рішення задач

Задача 1. Дискретна випадкова величина X приймає значення: 1; 2, 3 з імовірностями 0,2; 0,3; 0,5 відповідно.

1) Намалювати графік функції розподілу дискретної випадкової величини X .

2) Обчислити математичне очікування, дисперсію, другий початковий момент, середньоквадратичне відхилення і коефіцієнт варіації випадкової величини X .

Дано: $x_1 = 1$; $x_2 = 2$; $x_3 = 3$;

$p_1 = 0,2$; $p_2 = 0,3$; $p_3 = 0,5$

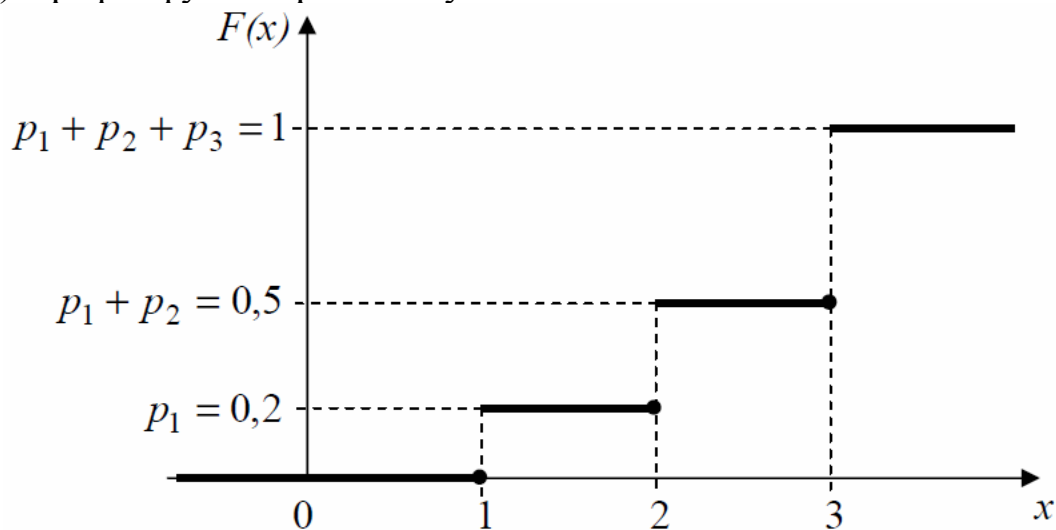
Потрібно:

1) намалювати $F(x)$;

2) обчислити $M[X]$, $D[X]$, $\alpha_2[X]$, $\sigma[X]$, $\nu[X]$.

Рішення.

1) Графік функції розподілу випадкової величини X :



Слід зазначити, що значення функції розподілу $F(x)$ для кожного значення випадкової величини x_i збільшуються на величину, рівну відповідній ймовірності p_i появи цього значення, причому саме верхнє значення завжди дорівнює 1.

Відзначимо також, що, як показано на графіку (у вигляді чорних кружечків), значення функції розподілу в точках $x = 1$, $x = 2$ та $x = 3$ відповідно рівні: $F(1) = 0$, $F(2) = 0,2$ і $F(3) = 0,5$, оскільки функція розподілу $F(x)$ визначається як імовірність появи випадкової величини, значення якої строго менше (а не менше або дорівнює) x : $F(x) = P(X < x)$.

2) Математичне очікування:

$$M[X] = p_1x_1 + p_2x_2 + p_3x_3 = 0,2 \times 1 + 0,3 \times 2 + 0,5 \times 3 = 2,3$$

Другий початковий момент:

$$\alpha_2[X] = p_1x_1^2 + p_2x_2^2 + p_3x_3^2 = 0,2 \times 1 + 0,3 \times 4 + 0,5 \times 9 = 5,9$$

$$\text{Дисперсія: } D[X] = \alpha_2[X] - (M[X])^2 = 5,9 - 5,29 = 0,61$$

$$\text{Середньоквадратичне відхилення: } \sigma[X] = \sqrt{D[X]} = 0,78$$

$$\text{Коефіцієнт варіації: } v[X] = \frac{\sigma[X]}{M[X]} \approx 0,34.$$

Задача 2. Чому дорівнює математичне очікування, дисперсія, другий початковий момент і коефіцієнт варіації детермінованої величини $x = 10$? Намалювати графік функції і щільності розподілу випадкової величини.

Дано: детермінована величина: $x = 10$.

Потрібно:

- 1) обчислити $M[X]$, $D[X]$, $\alpha_2[X]$, $v[X]$;
- 2) намалювати $F(x)$ і $f(x)$.

Рішення.

1) Детерміновану величину можна розглядати як випадкову величину, приймаючи одне і те ж значення $x = 10$ з ймовірністю $p = 1$. Тоді:

- математичне сподівання: $M[X] = px = 10$;

- другий початковий момент: $\alpha_2[X] = px^2 = 100$;

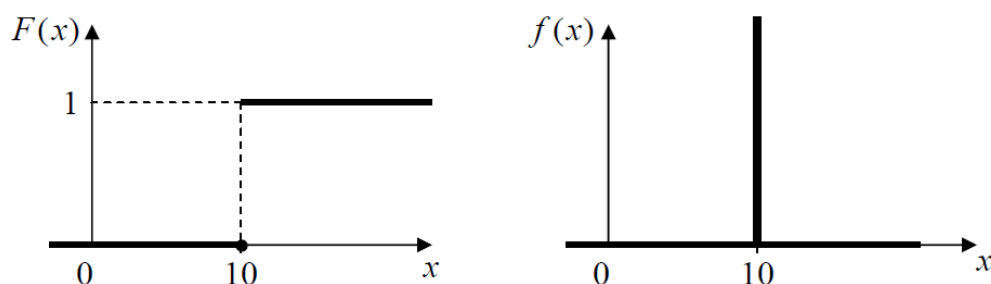
- дисперсія: $D[X] = \alpha_2[X] - (M[X])^2 = 0$;

- коефіцієнт варіації: $v[X] = \frac{\sqrt{D[X]}}{M[X]} = 0$.

Отримані досить тривіальні результати стають очевидними, якщо згадати фізичне тлумачення представлених величин. Математичне очікування, що представляє собою середнє значення випадкової величини, природно, збігається з єдиною можливим значенням $x = 10$.

Дисперсія, середньоквадратичне відхилення і коефіцієнт варіації, що визначають розкид значень щодо математичного очікування, очевидно завжди дорівнюють нулю для детермінованої величини, оскільки розкиду значень просто немає. Однак слід звернути увагу, що другий початковий момент не дорівнює нулю, хоча теж визначає розкид значень, але, на відміну від попередніх характеристик, щодо початку координат. Дійсно, єдине значення $x = 10$ знаходиться від початку координат на «відстані», не рівному нулю і, отже, другий початковий момент відмінний від нуля.

2) Графіки функції та щільності розподілу детермінованої величини:



Як випливає з представлених графіків, функція розподілу детермінованої величини являє собою функцію Хевісайда, а щільність розподілу - дельта-функцію: $F(x) = H(x-M)$; $f(x) = \delta(x-M)$, де M - математичне очікування, рівне значенню детермінованої величини (у нашому випадку $M = 10$).

Задача 3. Безперервна випадкова величина рівномірно розподілена в інтервалі $(-30; +20)$. Намалювати графік щільності та функції розподілу випадкової величини. Визначити: а) математичне очікування випадкової величини; б) ймовірність того, що випадкова величина приймає позитивні значення.

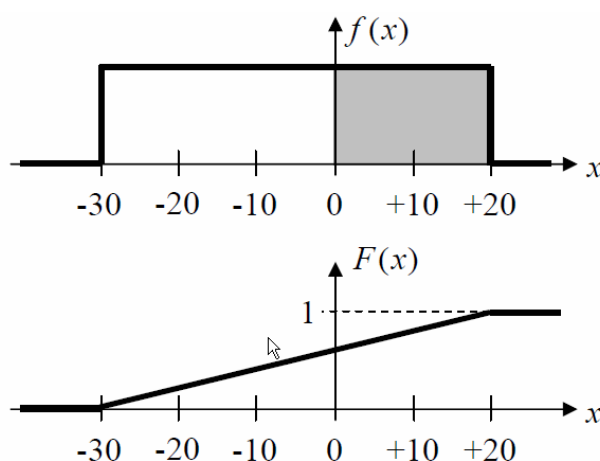
Дано: рівномірно розподілена випадкова величина в інтервалі $(-30; +20)$.

Потрібно:

- 1) намалювати $f(x)$ і $F(x)$;
- 2) обчислити $M[X]$;
- 3) визначити $\Pr(X \geq 0) = 1 - F(0)$.

Рішення.

1) Графіки щільності та функції рівномірно розподіленої випадкової величини:



2) Очевидно, що математичне очікування рівномірно розподіленої випадкової величини знаходиться в середині заданого інтервалу $(-30; +20)$ і дорівнює: $M = -5$. Цей же результат може бути отриманий з використанням формули для розрахунку математичного очікування рівномірно розподіленої випадкової величини:

$$M = \frac{a+b}{2} = \frac{-30+20}{2} = -5,$$

де a і b - відповідно ліва і права межі інтервалу.

3) Ймовірність того, що випадкова величина приймає позитивні значення, також може бути визначена кількома способами.

По-перше, через значення функції розподілу:

$$\Pr(X \geq 0) = 1 - F(0) = 1 - 0,6 = 0,4.$$

По-друге, з графіка щільності розподілу як площа під щільністю розподілу, що обмежена ліворуч значенням $x = 0$ і справа значенням $x = +20$ (на графіку виділена сірим кольором). Пам'ятаючи, що площа під щільністю розподілу на всьому інтервалі значень випадкової величини дорівнює 1, можна зробити висновок, що площа на інтервалі значень $(0, +20)$ становить $2/5$, тобто дорівнює $0,4$.

Д.7. Самоконтроль. Перелік питань і задач

1. Що розуміється під випадковою величиною?
2. Наведіть приклади випадкових величин.
3. Чи є випадковою величиною ...:
 - ... Число днів у році?
 - ... Зріст людини?
 - ... Кількість пасажирів в автобусі?
 - ... Інтервал між поїздами в метро?
 - ... Температура повітря на вулиці?
 - ... Напруга в електричній мережі?
 - ... Кількість студентів у групі?
 - ... Оцінка на іспиті?
 Відповіді супроводити необхідними поясненнями. Які з перерахованих величин є безперервними?
4. Наведіть приклади дискретних і безперервних випадкових величин.
5. Що характеризує ймовірність?
6. Як розрахувати ймовірність якої-небудь події?
7. Розрахуйте ймовірність того, що в ретельно перемішаній колоді з 36 карт нижньої картою виявиться ...:
 - ... Туз піковий?
 - ... Туз будь-якої масті?
 - ... Будь-якої масті туз або король?
 - ... Піковий туз або піковий король?
 - ... Піковий туз або хрестовий король?
 - ... Будь-яка карта червоної масті?
8. Розрахуйте ймовірність того, що в ретельно перемішаній колоді з 36 карт дві верхні карти виявляться ...:
 - ... Тузами?
 - ... Туз і король однієї масті?
 - ... Піковий туз і піковий король?
 - ... Піковий туз або король будь-якої масті?
 - ... Червоної масті?
9. З ретельно перемішаної колоди, яка містить 36 карт, відкинута половина карт. Яка ймовірність того, що в половині колоди, яка залишилася знаходиться ...:
 - ... Туз піковий?

- ... Туз будь-якої масті?
 - ... Будь-якої масті туз або король?
 - ... Піковий туз або піковий король?
 - ... Піковий туз або хрестовий король?
 - ... Будь-яка карта червоної масті?
10. Поняття та способи завдання закону розподілу випадкової величини.
 11. Поняття і властивості функції розподілу випадкової величини.
 12. Поняття і властивості щільності розподілу ймовірностей.
 13. Що характеризує і яку розмірність має математичне очікування (дисперсія, другий початковий момент, середньоквадратичне відхилення, коефіцієнт варіації, функція розподілу, щільність розподілу) випадкової величини?
 14. Для чого використовуються твірна функція і перетворення Лапласа? Для яких випадкових величин використовується перетворення Лапласа?
 15. Назвіть відомі Вам дискретні і безперервні закони розподілів.
 16. Чому дорівнює коефіцієнт варіації: а) експоненційного розподілу; б) розподілу Ерланга 9-го порядку?
 17. В якому інтервалі знаходиться коефіцієнт варіації розподілу: а) Ерланга; б) гіперекспоненційного; в) гіперерлангівського?
 18. Намалювати графік щільності та функції розподілу: а) експоненційного; б) Ерланга; в) гіперекспоненційного.
 19. Показати на графіку і пояснити, в чому відмінність між щільностями розподілів експоненційного і гіперекспоненційного законів.
 20. Показати на графіку і пояснити, в чому відмінність між щільностями розподілів Ерланга 2-го і 4-го порядку.
 21. Дискретна випадкова величина приймає значення 1; 2; 3 з імовірностями 0,5; 0,4; 0,1 відповідно. Обчислити математичне очікування, дисперсію, другий початковий момент, середньоквадратичне відхилення, коефіцієнт варіації.
 22. Чому дорівнює математичне очікування, дисперсія, другий початковий момент і коефіцієнт варіації детермінованої величини $X = 25$?
 23. Визначити значення детермінованої величини X , якщо відомо, що її другий початковий момент дорівнює 25?
 24. Визначити коефіцієнт варіації детермінованої величини X , якщо відомо, що її другий початковий момент дорівнює 121?

25. Математичне очікування експоненційно розподіленої випадкової величини дорівнює 0,1. Визначити середньоквадратичне відхилення, другий початковий момент і коефіцієнт варіації.
26. Дисперсія експоненційно розподіленої випадкової величини дорівнює 16. Визначити математичне очікування і другий початковий момент.
27. Чому дорівнює коефіцієнт варіації розподілу Ерланга 4-го порядку?
28. Чому дорівнює дисперсія випадкової величини, розподіленої за законом Ерланга 9-го порядку з математичним очікуванням, рівним 15?
29. Дискретна випадкова величина приймає значення 1; 2; 3 відповідно з імовірностями 0,3; 0,1; 0,6. Намалювати графік функції розподілу випадкової величини. Визначити математичне очікування і другий початковий момент.
30. Безперервна випадкова величина приймає значення в інтервалах (1, 2) і (3, 4), причому ймовірність появи значення з інтервалу (3, 4) в три рази більше ймовірності появи значення з інтервалу (1, 2). Вважаючи, що в межах кожного з інтервалів випадкова величина має рівноймовірний розподіл, побудувати графіки функції і щільності розподілу. Обчислити математичне очікування, дисперсію, другий початковий момент, середньоквадратичне відхилення, коефіцієнт варіації.
31. Випадкова величина може приймати тільки два значення 10 і 90. Яка ймовірність появи цих значень, якщо відомо, що математичне очікування випадкової величини дорівнює 80?
32. У чому полягає властивість відсутності післядії, притаманна експоненційному закону розподілу випадкових величин?

ПРЕДМЕТНИЙ ПОКАЖЧИК

Burke.....	68
Jackson.....	4, 71
M/M/1	3, 21, 22, 25, 26, 27, 31, 93, 96, 118, 126, 132
АСУ СС.....	137, 138
багатоканальна СМО	35
базові співвідношення СМО	19
БСМО	3, 35, 36, 38
вузлові характеристики	93, 96, 101
геометричний розподіл.....	5, 161, 162
гіперекспоненційний розподіл.....	169, 170, 172, 173
гіперерлангівський розподіл	173
двопараметрична модель	79
дисципліни обслуговування.....	13, 27, 35, 43, 49, 52, 53, 60, 62, 63, 64, 103
дисципліни диспетчеризації.....	13
ДО АП	49, 50, 57, 58, 59, 62, 63, 64, 122
ДО БП.....	49, 50, 51, 53, 54, 55, 56, 60, 63, 120, 121
ДО ВП	48, 49, 50, 53, 54, 55, 56, 57, 59, 62, 63, 121, 122
задачі аналізу СМО	20
закони збереження	60
закони розподілення	149, 151
закон збереження часу очікування	60, 61, 64, 120
ЗММО	102, 103, 104, 105, 107, 108, 109, 110, 111, 112, 113, 114, 115, 124, 125, 126
ймовірність	15, 16, 20, 36, 37, 38, 41, 42, 45, 70, 71, 73, 79, 103, 106, 119, 123, 125, 147, 148, 149, 150, 151, 154, 161, 171, 172, 173, 178, 18
інформаційно-довідкова система.....	138, 140
Кендал	21

класифікація	43, 44, 45, 46
мережеві характеристики	89, 93, 97, 100, 101, 102, 104, 107, 110, 124
мережі Джексона	71
ММО 9, 15, 21, 66, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 122, 123, 124, 125, 126, 127, 132, 143	
мультисерверна СМО	17, 18
нормований розподіл Ерланга	168, 169
ОСМО	26, 27, 29, 30, 32, 34
основні позначення	13
параметри СМО	12
подія	14, 147, 148
популяція заявок на обслуговування	13
ПРАКТИКУМ	10, 176
припущення в моделях	21, 23
РММО 88, 89, 90, 91, 93, 95, 96, 97, 98, 99, 100, 102, 104, 105, 113, 122, 123, 132	
розмір черги	13, 24
розподіл Ерланга	166, 167, 168, 169
розподіл Пуассона	161
самоконтроль	41, 62, 115, 180
сервер баз даних	132
система обслуговування мульти-сервера	17
СМО тандемів	68
структура СМО	11
твірна функція	14, 159, 160, 162, 163, 181
теорема Бьорке	68
типові розподіли	160
узагальнення матеріалу	118
формула Літла	19
формули для мультисервера	37
числові характеристики розподілів	175

Навчально-довідковий посібник

Косолапов Анатолій Аркадійович

Аналітичні моделі масового обслуговування
в задачах проектування інформаційних систем

Редактор О.Б. Ременюк

Формат 60x84x1/16 Ум. Друк. Арк.. 1,16. Обл.-вид. арк. 8,5
Тираж 50 пр. Зам. № 17

Віддруковано ФОП Гечка Т.О.
Свідоцтво платника єдиного податку №2224000000089371
ЄДРПОУ № 3212715263

Адреса видавництва:
49080, м. Дніпропетровськ, вул.. Коцюбинського 2/4