МІНІСТЕРСТВО ТРАНСПОРТУ ТА ЗВ'ЯЗКУ УКРАЇНИ

ОДЕСЬКА НАЦІОНАЛЬНА АКАДЕМІЯ ЗВ'ЯЗКУ ім. О.С. ПОПОВА

А.Г. Ложковський

ТЕОРІЯ МАСОВОГО ОБСЛУГОВУВАННЯ В ТЕЛЕКОМУНІКАЦІЯХ

Затверджено Міністерством транспорту та зв'язку України як підручник для студентів вищих навчальних закладів, які навчаються за напрямом „Телекомунікації"

Одеса 2010

План НМВ 2010 р.

УДК 621.39

ББК 32.81

Л71

Затверджено Міністерством транспорту та зв'язку України

(*Лист № 6778/23/14-08 від 22.09.2008 р.*)

Рецензент – Захарченко М. В.

Ложковський А.Г.

Л71 Теорія масового обслуговування в телекомунікаціях /

А.Г. Ложковський. – Одеса: ОНАЗ ім. О.С. Попова, 2010. – 112 с.: іл.

Викладено основні положення та методи аналізу теорії масового обслуговування в телекомунікаціях (теорії телетрафіка), на яких базуються процедури проектування телекомунікаційних систем і мереж. Розглянуто математичні моделі систем розподілу інформації з втратами, з чергою та з пріоритетами. Наведено методи дослідження цих систем в умовах ідеалізованої моделі пуассонівського потоку та реальних потоків вимог мультисервісних мереж зв'язку.

Підручник призначено для студентів та аспірантів, які навчаються за напрямом „Телекомунікації", а також для фахівців, що займаються практичним застосуванням теорії телетрафіка.

Рекомендовано
до видання вченою радою
ОНАЗ ім. О.С. Попова
Протокол № 10
від "30" травня 2008 р.

ББК 32.81

ISBN 978-966-7595-43-3

ЗМІСТ

ПЕРЕДМОВА

Виникнення телефонного зв'язку спонукало розробку спеціальних математичних методів оцінки якості функціонування телефонних систем, внаслідок чого науковими роботами датського вченого Агнера Крарупа Ерланга (1878-1929) на початку XX сторіччя закладено основи теорії телетрафіка. Шведський вчений К. Пальм узагальнив дослідження А. К. Ерланга і в своїй докторській дисертації навів важливі результати з вивчення змінюваності телефонного навантаження. Подальший розвиток цієї роботи дозволив в середині XX сторіччя російському математику А.Я. Хінчину започаткувати новий науковий напрям прикладної математики, що окрім телекомунікацій охоплює ще й процеси у системах виробництва, обслуговування, керування тощо. Даний напрям названо теорією масового обслуговування (ТМО). Ця теорія, що «виросла» з теорії телетрафіка, розглядає більш широке коло питань кількісної оцінки процесів масового обслуговування. Таким чином теорія телетрафіка є окремим розділом ТМО, а теорія масового обслуговування в телекомунікаціях – є теорією телетрафіка.

Головним змістом теорії телетрафіка є дослідження пропускної здатності телекомунікаційних систем. Крім того методами цієї теорії розробляються нові науково обґрунтовані методи оцінки характеристик якості обслуговування. Теорія телетрафіка забезпечує оцінку всіх параметрів телекомунікаційних систем, причому насамперед враховується стохастичний (випадковий) характер потоків вимог, що надходять до системи на обслуговування.

Оцінка прогнозованої пропускної здатності та якості обслуговування є найважливішим етапом проектування телекомунікаційних систем і мереж. Тут аналітичні розрахунки базуються на математичному описі реакції системи на зовнішні впливи. Під реакцією системи розуміється її стан (кількість зайнятих серверів або місць очікування, час затримки й ін.), а під зовнішніми впливами – потоки вимог, збої, відмови із-за ненадійності тощо. Зовнішнім фактором впливу в мультисервісних мережах зв'язку є різнорідність інформації, що передається у рамках єдиної мережі – дані, мова, відео. Оскільки потоки цієї інформації істотно відрізняються між собою за пріоритетами, механізмами обслуговування, особливостями протоколів і т.д., то адекватною неминуче є багатомірна модель. Тому для складних систем аналітичні розрахунки, виконуються з обмеженням зовнішніх факторів, роздільно для кожного типу (групи) впливів або із застосовуванням багатопотокових моделей.

Опис реакції системи на сукупність усіх зовнішніх впливів є надзвичайно важкою задачею, яка у загальному виді навряд чи розв'язувана. По-перше, число зовнішніх впливів може бути дуже великим, по-друге, кожен вплив не завжди однозначно описується простими формулами, що дозволяють отримати кінцевий результат із зрозумілим фізичним змістом, по-третє, опис зовнішніх впливів не завжди адекватний реальним процесам, які відбуваються у системі.

В математичних моделях теорії телетрафіка враховано вид вхідного потоку, схему системи та дисципліну обслуговування.

Науково обґрунтоване планування й оптимізація телекомунікаційних систем та мереж, які забезпечують надання запитуваних послуг із заданими показниками якості обслуговування, є дуже складною науково-технічною й економічною проблемою, без вирішення якої неможливе створення інформаційної інфраструктури, що відповідає потребам розвинутого суспільства. В розвитку бізнесу окремих телекомунікаційних компаній цей фактор є найважливішим при обґрунтуванні дій адміністрації, спрямованих на підвищення ефективності роботи мережі і якості обслуговування користувачів.

Вирішення даної проблеми ґрунтується на розв'язанні задач аналізу і синтезу телекомунікаційних систем. Комплексне розв'язання цих задач дозволяє оптимізувати структуру мережі на тривалу перспективу. В умовах розвитку телекомунікацій у відповідності до основних положень концепції мереж наступного покоління *NGN*, які забезпечують надання необмеженого набору послуг із заданими характеристиками якості обслуговування *QoS*, зазначені питання стають ще більш актуальними. Обрана технологія розподілу інформації в *NGN* визначає ступінь складності вузлів комутації, що, безумовно, впливає на якість обслуговування обміну інформацією між терміналами користувачів. Крім того, якість обслуговування потоків інформації впливає й на самі характеристики передачі інформації (наприклад, затримки пакетів IP-телефонії призводять до зниження якості телефонного зв'язку). Таким чином, розширення спектру надаваних послуг та зростаюча складність телекомунікаційних систем і мереж вимагає вирішення проблеми розробки адекватних методів аналізу і синтезу цих систем з метою отримання достовірних оцінок їх характеристик, реалізації задач їх оптимізації щодо обраного критерію якості обслуговування та розробки відповідних алгоритмів керування ними.

Процеси функціонування мереж та систем зв'язку можна представити тією чи іншою сукупністю систем масового обслуговування (СМО), для яких визначаються характеристики *QoS*. Одним із класів СМО в телекомунікаціях є системи розподілу інформації (СРІ), до яких належать мережі зв'язку в цілому або окремі комутаційні вузли чи, наприклад, пакетні комутатори, що обслуговують за певним алгоритмом повідомлення телекомунікаційних служб. Кількісна сторона процесів обслуговування потоків повідомлень (трафіка) у СРІ є предметом теорії телетрафіка. Ця теорія, як самостійна наукова дисципліна, являє собою набір імовірнісних методів вирішення проблем проектування нових та експлуатації діючих систем телекомунікацій.

Заснування теорії телетрафіка розпочато з наукових праць Ерланга А.К. (визначні *B*- і *C*-формула Ерланга). Її розвивали такі вчені, як Burce P.J, Crommelin C. D., Kleinrock L., O'Dell G.F., Palm C., Pollaczek F., Wilkinson R.I. Суттєвий вклад в розвиток теорії телетрафіка внесли представники російської наукової школи: Хінчин А. Я., Башарин Г.П., Лівщиц Б.С., Харкевич А.Д., Севаст'янов Б.А., Нейман В.І, Степанов С.М. та академіки АН України Гніденко Б.В., Коваленко І.М., Королюк В.В. Здобутком теорії є підручники і наукові роботи Шнепса М.А., Корнишева Ю.М. та інших.

Пропускна здатність СМО тісно пов'язана з оцінкою показників якості обслуговування трафіка, що вимагає обліку багатьох факторів для побудови адекватних, науково обґрунтованих методів їх розрахунку. Методи оцінки характеристик якості обслуговування базуються на математичних моделях СРІ. Різноманіття видів та топологій мереж, структур систем та способів виділення мережного ресурсу для обслуговування трафіка вимагає розробки моделей, які враховують ще й реальний характер потоків повідомлень і деталі обслуговування мультисервісного трафіка різних комунікаційних додатків (мова, відео, дані). Через те неможливо побудувати єдину модель, яка б давала відповіді на всі питання стосовно функціонування нових мереж зв'язку. Саме на основі застосовуваних моделей СРІ розробляються методи оцінки характеристик *QoS*, достовірність яких залежить від адекватності моделі реальній ситуації, що може виникнути при проектуванні та експлуатації.

Оцінка якості обслуговування трафіка є одним із найважливіших наукових напрямом в дослідженнях телекомунікаційних мереж. На цьому базується продумана та цілеспрямована стратегія модернізації сучасних мереж на етапі їх конвергенції та заміни технології комутації каналів на комутацію пакетів. Принципи функціонування мережі обумовлені режимами переносу інформації, а якість обслуговування – реальним характером трафіка. За цих обставин необхідна розробка нових методів аналізу і синтезу СРІ, що адекватно віддзеркалюють реальні процеси обміну інформацією в мережі. Це надасть подальший розвиток теорії телетрафіка та збагатить практичний інструментарій середовища проектування інфокомунікаційних мереж, що в свою чергу дозволить забезпечити відчутну економію витрат на будівництво та експлуатацію мереж зв'язку. Завдяки більш точним розрахункам підвищиться якість обслуговування й пропускна здатність СРІ. Нові методи оцінки характеристик QoS необхідні в системах динамічного керування мережами для перерозподілу їх ресурсів та оптимізації трафіка і мережі в цілому на основі заданої (нормованої) якості обслуговування.

Таким чином, при системному підході до проблеми планування й оптимізації телекомунікаційних систем та мереж неможливо обійтися без математичних методів аналізу, синтезу та оцінки якості надання інформаційних послуг в умовах реальних потоків повідомлень. Відсутність таких методів призводить до прийняття неоптимальних рішень у процесі розробки, проектування й експлуатації телекомунікаційних систем, та мереж оскільки виникає різка невідповідність між очікуваними (проектними) показниками та реальною якістю обслуговування.

Телетрафік – це не тільки класичні телефонні повідомлення, але й потоки повідомлень у нових інфокомунікаційних мережах. Специфічні особливості різних СРІ збільшують проблеми розробки універсальних методів їх аналізу і синтезу. Особливо складна ця проблема для моделей трафіка, які є адекватними реальним процесам формування його потоків в мережі. Природа надходження потоків і їх обслуговування залежить від конкретного виду системи та мережі, структурного складу абонентів, спектру надаваних послуг та інших факторів.

В теорії телетрафіка розроблено низку математичних моделей і методів вирішення задач аналізу і синтезу СРІ для умов ідеалізованої пуассонівської моделі трафіка. Однак набір цих методів поки є недостатньо повним з погляду структурних особливостей реальних СРІ, дисциплін обслуговування і особливо характеру трафіка. Реальному трафіку сучасних мультисервісних мереж зв'язку властивий значно більший рівень нерівномірності інтенсивності навантажень, ніж це передбачено класичною моделлю пуассонівського потоку. Для таких моделей трафіка теорія телетрафіка не має відповідних методів розрахунку і на практиці оцінка характеристик якості обслуговування мультисервісних мереж зв'язку ведеться наближеними методами та засобами імітаційного моделювання. „Нерезультативність" існуючих методів полягає в тому, що вони орієнтовані на використання лише перших моментів розподілів випадкових величин, що визначають інтенсивність трафіка та функціонування СРІ. При обслуговуванні реального трафіка суттєвий вплив на характеристики *QoS* мають і вищі моменти розподілів названих величин, які визначають характер та ступінь нерівномірності трафіка.

Аналіз наукових публікацій показує, що багато теоретичних розробок не можливо використовувати на практиці. Це пов'язано із такими недоліками теоретичних досліджень, як: застосування математичного апарату, що не адекватно відображає процеси в телекомунікаційних мережах; невдалий вибір показників і критеріїв оцінки пропонованих рішень; спроба отримання аналітичних залежностей характеристик телекомунікаційних мереж в межах, де аналітичний апарат не працює; розробка методів, які, покращуючи один з параметрів телекомунікаційної системи, зрештою, знижують ефективність функціонування всієї системи в цілому; визначення закономірностей, у складі яких є початкові дані, одержати які не вбачається можливим.

З урахуванням значимості даної наукової проблеми можна констатувати, що проблема розвитку теорії телетрафіка шляхом її збагачення новими методами аналізу і синтезу систем розподілу інформації, які функціонують в умовах обслуговування реальних потоків трафіка, є особливо актуальною в сучасних умовах конвергенції мереж та побудови на цій основі мереж нового покоління *NGN*. Така постановка задачі відповідає загальному підходові, прийнятому в міжнародній практиці і сформульованому в рекомендаціях *ITU*.

1 ЕЛЕМЕНТИ ТЕОРІЇ ЙМОВІРНОСТЕЙ

1.1 Випадкові величини та ймовірнісні розподіли

Теорія ймовірностей – математична наука, яка вивчає закономірності, властиві масовим випадковим явищам. При цьому досліджувані явища розглядаються в абстрактній формі, незалежно від їхньої конкретної природи. Тобто теорія ймовірностей розглядає не самі реальні явища, а їхні спрощені схеми – математичні моделі. Предметом теорії ймовірностей є математичні моделі випадкових явищ. При цьому під випадковим явищем розуміють явище, передбачати результат якого неможливо (при неодноразовому відтворенні того самого досліду воно протікає щораз трохи по-іншому). Приклади випадкових явищ: випадання герба при підкиданні монети, виграш по придбаному лотерейному квитку, результат виміру якої-небудь величини, тривалість роботи телевізора й т.п.

Мета теорії ймовірностей – здійснення прогнозу в області випадкових явищ, вплив на хід цих явищ, контроль їх, обмеження сфери дії випадковості. У цей час немає практично жодної області науки, у якій у тому або іншому ступені не застосовувалися б імовірнісні методи.

Задачі, результат яких не можна передбачити з упевненістю, вимагають вивчення не тільки основних, головних закономірностей, що визначають явище в цілому, але й випадкових, другорядних факторів. Виявлені в таких задачах (дослідах) закономірності називаються *статистичними* (або *ймовірнісними*). Статистичні закономірності досліджуються методами спеціальних математичних дисциплін – теорії ймовірностей і математичної статистики.

Імовірність P є мірою можливості здійснення результату або події *A*. Формально міра ймовірності є функцією $P(A)$, що ставить у відповідність результатам деякі числа і задовольняє наступним вимогам:

- $0 \le P(A) \le 1$ для будь-якого результату A;
- $P(A) = 1$ у разі достовірного результату;
- $\sum_{i=1}^{n} P(A_i) = 1$, де n – увесь простір вибірки можливих результатів.

Функція, що ставить у відповідність кожному результатові з простору вибірки деяке дійсне число, називається *випадковою величиною*. Дискретними називаються ті випадкові величини, що належать кінцевій або ліченій множині значень. Безперервні випадкові величини можуть належати континуумові значень. Наприклад, інтервал часу між надходженням вимог на обслуговування є безперервною випадковою величиною, а кількість вимог, обслужених за інтервал часу, – дискретною.

Імовірнісний закон розподілу являє собою деяке правило завдання імовірності для кожного з усіх можливих значень випадкової змінної. Правило завдання імовірності має дві різні форми в залежності від того, є випадкова величина *дискретною* або *безперервною*.

Для дискретної випадкової величини *K* функція ймовірності (закон розподілу) задається ймовірностями кожного її значення k_i:

$$P(K = k_i) = p(k_i), \qquad \text{де } \sum_i p(k_i) = 1.$$

Для кожного можливого значення k_i закон розподілу встановлює конкретну імовірність того, що дискретна випадкова величина K приймає значення k_i.

Кумулятивна функція розподілу $F(k)$ позначається у такий спосіб:

$$F(k) = P(K \le k)$$

Функція $F(k)$ визначає імовірність того, що випадкова величина K прийме значення не більше, ніж k. Кумулятивна функція розподілу пов'язана з функцією ймовірності так:

$$F(k) = \sum_{k_i \le k} p(k_i)$$

Для безперервних випадкових величин необхідна інша форма представлення ймовірнісного розподілу. Оскільки випадкова величина може приймати будь-яке з нескінченної множини значень, імовірність конкретного значення дорівнює нулеві. Це говорить не про те, що дане значення неможливе, а про те, що воно вкрай неймовірне внаслідок нескінченної кількості альтернативних значень. При цьому, звичайно, імовірність того, що випадкова величина прийме значення в інтервалі між точками a і b, у більшості випадків не буде дорівнювати нулеві Отже, функція імовірності для дискретного випадку заміняється на безперервну функцію густини ймовірності $f(x)$, обумовлену наступним виразом:

$$P(a \le X \le b) = \int_a^b f(x)dx, \qquad \text{де } \int_{-\infty}^{\infty} f(x)dx = 1$$

Таким чином, функція густини ймовірності при інтегруванні на інтервалі від a до b дає ймовірність того, що безперервна випадкова величина X прийме значення з цього інтервалу.

Функція розподілу $F(x)$ для безперервних випадкових величин визначається у такий спосіб:

$$P(X \le x) = F(x) = \int_{-\infty}^{x} f(x)dx$$

Функція $F(x)$ визначає ймовірність того, що безперервна випадкова величина X прийме значення, не більше ніж x.

Густина ймовірності $f(x)$ безперервної випадкової величини X визначається як похідна від її функції $F(x)$:

$$f(x) = F'(x).$$

Найбільш поширеними ймовірнісними розподілами дискретної випадкової величини є розподіли Пуассона, Бернуллі, біноміальний, геометричний та логарифмічний розподіли.

Найбільш поширеними розподілами безперервної випадкової величини є розподіли Гаусса (нормальний закон), Парето, Пірсона, Вуйбулла, Релея, гамма-розподіл, логарифмічно нормальний, експонентний розподіли.

1.2 Математичне сподівання та моменти розподілу

Часто необхідно охарактеризувати випадкову величину одним або декількома значеннями, що підсумовують інформацію, яка є у функції розподілу імовірності. *Математичним сподіванням* випадкової величини X, що позначається $M(X)$, є значення, обумовлене в такий спосіб:

$$M(X) = \sum_{i=0}^{n} x_i p_i \text{ , якщо величина } X \text{ дискретна;} \qquad (1.1)$$

$$M(X) = \int_{0}^{\infty} xf(x)dx \text{, якщо величина } X \text{ безперервна.} \qquad (1.2)$$

Математичним сподіванням є зважена по імовірності середня величина всіх можливих значень X, що визначає міру центральності розподілу. Тому ця величина часто називається *середнім значенням* $\bar{x}$.

Повним комплектом числових характеристик випадкової величини X є *моменти* розподілу або математичне сподівання функцій цих випадкових величин. Зокрема, математичне сподівання функції X^k називається k-м *початковим* моментом або початковим моментом k-го порядку випадкової величини X і визначається в такий спосіб:

$$m_k = M(X^k) = \sum_{i=0}^{n} x_i^k p_i \text{ , де } X \text{ дискретна;}$$

$$m_k = M(X^k) = \int_{0}^{\infty} x^k f(x)dx \text{, де } X \text{ безперервна.}$$

З цього видно, що при $k = 1$ перший початковий момент m_1 випадкової величини X є її математичним сподіванням або середнім значенням $\bar{x}$.

Варіацією k-го початкового моменту є k-й *центральний* момент μ_k, який визначається виразом

$$M[(X - m_1)^k]$$

Отже, для обчислення k-го центрального моменту випадкової величини X від її значення віднімається перший початковий момент m_1 або середнє значення $\bar{x}$. Центральні моменти k-го порядку для дискретної та безперервної випадкової величини X відповідно визначаються так:

$$\mu_k = \sum_{i=0}^{n} (x_i - \bar{x})^k p_i = \frac{1}{n} \sum_{i=0}^{n} (x_i - \bar{x})^k ;$$

$$\mu_k = \int_{0}^{\infty} (x - \bar{x})^k f(x)dx .$$

Особливе значення має другий центральний момент μ_2, називаний *дисперсією* X і який позначається як $D(X)$ або σ^2. Дисперсія випадкової величини X є мірою розкиду її ймовірнісного розподілу. Якщо дисперсія випадкової величини мала, то уся вибірка лежить поблизу математичного сподівання. Квадратний корінь із дисперсії σ^2 називається *стандартним відхиленням* випадкової величини або середньоквадратичним відхиленням σ.

Центральний момент третього порядку μ_3, який нормовано такою ж ступеню середньоквадратичного відхилення, тобто σ^3, називається *асиметрією* випадкової величини *Sk* і для дискретної величини

$$\mu_3 = Sk = \frac{1}{n\sigma^3}\sum_{i=0}^{n}(x_i - \bar{x})^3 .$$

Центральний момент четвертого порядку μ_4, який нормовано такою ж ступеню середньоквадратичного відхилення, тобто σ^4, називається *ексцесом* випадкової величини *Ex* і для дискретної величини

$$\mu_4 = Ex = \frac{1}{n\sigma^4}\sum_{i=0}^{n}(x_i - \bar{x})^3 - 3$$

Моменти більш високих порядків використовуються рідко. При цьому, якщо по заданих законах розподілу випадкової величини моменти розподілу можна визначити однозначно, те зворотна задача розв'язувана не завжди.

Коефіцієнти асиметрії (показник зсуву вліво-вправо вершини функції розподілу або скошеності) та ексцесу (показник гостроти піка цієї функції) використовуються для порівняння закону розподілу будь-якої випадкової величини з нормальним законом, для якого $Sk = Ex = 0$. Їх можна визначити через центральні моменти відповідного порядку:

$$Sk = \frac{\mu_3}{\mu_2^{3/2}}; \qquad Ex = \frac{\mu_4}{\mu_2^2} - 3 .$$

На діапазон розкиду окремих значень випадкової величини від її середнього значення вказує дисперсія, але у випадку порівняння цих діапазонів для двох випадкових величин різної розмірності краще використовувати нормоване значення дисперсії випадкової величини її середнім значенням, яке називається *коефіцієнтом варіації* випадкової величини:

$$v_x = \frac{\sigma}{\bar{x}} .$$

Статистична обробка масиву значень випадкової величини дозволяє компактно описати її, зрозуміти структуру даних, провести класифікацію, побачити закономірності в потоці випадкових подій. Замість розгляду всіх значень досліджуваної величини складаються описові статистики, що дають загальне уявлення про значення, які приймає ця величина. Серед них основними є наступні: максимум, мінімум і середнє значення, дисперсія і стандартне відхилення, медіана, квартілі і квантілі, мода, асиметрія й ексцес.

Для дослідження зв'язку між двома випадковими величинами обчислюється коваріація та коефіцієнт кореляції між ними. При цьому можуть використовуватися рангові кореляції, статистика Спірмена *R*, статистика Кендалла, Гамма-статистика, кореляція Пірсона. Для визначення, чи є результат дослідження дійсно значимим, оцінюється міра впевненості в його правильності за рівнем значимості, для чого використовується метод найменших квадратів.

2 ЗАГАЛЬНІ ПОЛОЖЕННЯ ТЕОРІЇ ТЕЛЕТРАФІКА

Зростаюча складність телекомунікаційних систем та мереж вимагає вирішення проблеми розробки адекватних методів розрахунку цих систем з метою отримання достовірних оцінок їх характеристик, реалізації задач їхньої оптимізації щодо вибраного критерію якості обслуговування та розробки відповідних алгоритмів керування ними.

Математичні моделі телекомунікаційних систем та мереж, як правило, будуються на основі теорії *систем масового обслуговування* (СМО). В загальному випадку СМО обслуговують вимоги, що надходять до системи через випадкові інтервали часу, причому тривалість обслуговування також може бути випадковою. Методами теорії СМО досліджується вплив випадкових факторів на процеси функціонування системи.

Одним із класів СМО є *системи розподілу інформації* (СРІ), які характеризуються наявністю розподільчої мережі, подібно транспортним системам або системам енергопостачання. При передаванні інформації аналогом розподільчої мережі є телекомунікаційна мережа, яка складається з каналів передачі інформації та вузлів комутації. Сукупність цих засобів зв'язує джерела інформації з їх споживачами. Каналами зв’язку передається інформація, яка безпосередньо є предметом передачі й розподілу, і допоміжна, яка необхідна в процесі керування роботою всієї системи. Вузли комутації забезпечують з'єднання каналів передавання інформації і в них за певними алгоритмами обслуговуються повідомлення телекомунікаційних служб мережі. При цьому обслуговування повідомлення ототожнюється з вимогою на його передачу або обробку і прикладами таких вимог можуть бути виклики телефонної станції або пакети пакетного комутатора. В якості СРІ може розглядатися не тільки мережа зв'язку в цілому, а й пучок каналів або ліній, окремий комутатор або весь комутаційний вузол.

Кількісна сторона процесів обслуговування потоків вимог (трафіка) в СРІ досліджується теорією телетрафіка (інша назва – теорія розподілу інформації).

Предметом теорії телетрафіка є встановлення залежностей між характером потоку вимог, кількістю каналів обслуговування, продуктивністю окремого каналу та ефективним обслуговуванням з метою визначення найкращих шляхів керування цими процесами.

Задача теорії телетрафіка полягає у встановленні залежності результуючих показників роботи СРІ (наприклад, середньої кількості вимог, що обслуговуються; середньої кількості вимог, що очікують обслуговування в черзі і т.д.) від вхідних показників (кількості каналів у системі, параметрів вхідного потоку вимог і т.д.). Результуючими показниками або досліджуваними характеристиками СРІ є показники ефективності, що описують, чи здатна дана система впоратися з потоком вимог.

Методами теорії телетрафіка можна вирішувати задачі оптимізації, які спрямовані на визначення такого варіанта системи, за якого буде забезпечений мінімум сумарних витрат від очікування обслуговування, втрат часу і ресурсів на обслуговування, та простоїв каналів обслуговування.

Теорія телетрафіка являє собою набір імовірнісних методів аналізу, синтезу та оптимізації СРІ і, таким чином, вирішення проблем проектування нових та експлуатації діючих мереж зв’язку. Без вирішення задач аналізу, синтезу і на цій основі оптимізації телекомунікаційних мереж та систем неможливий їх подальший розвиток.

Задача аналізу – це встановлення залежностей і значень величин, які характеризують якість обслуговування, від характеристик і параметрів вхідного потоку вимог, схеми і дисципліни обслуговування. Задача аналізу виникає в тих випадках, коли телекомунікаційна мережа або система вже побудована і функціонує. Цілями аналізу є отримання реальних характеристик СРІ, порівняння їх із проектними характеристиками, надання об'єктивних оцінок якості роботи системи. Аналіз дозволяє визначити причини зниження якості обслуговування і видати рекомендації щодо усунення цих причин. Іноді аналіз робиться після внесення змін у систему або після підключення нових джерел навантаження (реконструкції). Розробка методів оцінки якості функціонування телекомунікаційних мереж та систем є основною метою теорії телетрафіка.

Задача синтезу – це визначення структурних параметрів мережі або, наприклад, схеми комутаційного вузла цієї мережі при заданих потоках, дисципліні і якості обслуговування. Задача синтезу певною мірою є зворотною до задачі аналізу. Синтез (проектування) телекомунікаційних мереж може складатися з кількох етапів. З позицій системної методології, основними етапами вирішення задачі синтезу мереж та систем зв'язку є: аналіз проблеми; визначення системи; визначення цілей, критеріїв, ресурсів; визначення альтернативних варіантів; оцінка, порівняння і вибір варіантів; реалізація рішення. Задачі проектування й планування телекомунікаційних мереж виникають з необхідності завчасного вибору технічних засобів, що забезпечують задоволення потреб у передаванні інформаційних повідомлень. Метою проектування є оптимальна структура мережі на тривалу перспективу з урахуванням поточного стану розвитку телекомунікаційної техніки і технологій.

Задачі оптимізації є близькими до задач аналізу і синтезу. Як правило, при проектуванні телекомунікаційних мереж та систем вони формулюються в такий спосіб: визначити структурні параметри або алгоритми функціонування мережі (системи), для яких:

- при заданих потоках, якості і дисципліні обслуговування вартість або обсяг мережі (системи) мінімальні;
- при заданих потоках, дисципліні обслуговування і вартості якісні показники функціонування мережі (системи) оптимальні.

При експлуатації телекомунікаційних мереж та систем задача оптимізації формулюється як задача керування потоками вимог або структурою мережі для досягнення найкращих показників якості функціонування. Через великі обчислювальні труднощі задачі оптимізації телекомунікаційних мереж та систем вирішуються за допомогою ЕОМ.

Аналіз, синтез і оптимізація СРІ виконуються із застосуванням теорії ймовірностей, математичної статистики, комбінаторних й алгебраїчних методів, теорії множин, теорії графів, принципів системного підходу (системотехніки) та ін.

Основними методами вирішення задач у теорії телетрафіка є аналітичний, числовий та метод статистичного моделювання.

Аналітичні методи дозволяють розв’язувати задачі теорії телетрафіка в тих випадках, коли структура системи, характеристики потоку і дисципліна обслуговування відносно прості. При цьому розглядаються всі можливі стани системи, обумовлені, наприклад, положенням кожної точки комутації або кількістю зайнятих каналів. Такі стани називаються мікростанами системи. Щораз, коли надходить нова вимога, закінчується яка-небудь фаза роботи керуючого пристрою по встановленню з'єднання або закінчується з'єднання, система змінює свій мікростан. Для кожного мікростану записується рівняння статистичної рівноваги. Розв'язуючи систему таких рівнянь, знаходять точне рішення задачі в межах прийнятої моделі.

Числові методи використовують спеціальні алгоритми, що дозволяють знаходити наближені рішення ітераційними або іншими методами. Вони застосовуються для складних систем, де кількість мікростанів настільки велика, що розв'язати систему рівнянь статистичної рівноваги неможливо навіть за допомогою швидкодіючих ЕОМ. Тому застосовується так званий макропідхід. У складній системі з дуже великою кількістю мікростанів є та або інша ознака, за якою мікростани поєднуються в класи-макростани. Шляхом усереднення визначаються інтенсивності переходів з одних макростанів в інші. Для кожного макростану записується рівняння статистичної рівноваги. В результаті розв'язання системи таких рівнянь виводяться наближені формули для ймовірностей макростанів.

Методи статистичного моделювання є найбільш універсальними методами, які придатні для розв’язання задач практично будь-якої складності. Метод полягає в побудові математичної моделі системи, реалізація якої здійснюється у виді програми для ЕОМ. Моделювання дозволяє одержати числові результати, що характеризують якість обслуговування при заданих параметрах потоку, схеми і дисципліни обслуговування. Однак через специфіку методу він менш зручний порівняно з аналітичним і числовим методами при визначенні неявних закономірностей функціонування або залежностей між окремими характеристиками системи.

Для детального аналізу досліджуваних СРІ можливе поєднання аналітичних і числових методів з методом статистичного моделювання. Наприклад, якщо за малих значень параметрів системи вдається отримати рішення точними аналітичними методами і проаналізувати граничні випадки при асимптотичній поведінці характеристик досліджуваної системи, то потім отримані відомості доповнюються результатами статистичного моделювання в області реальних значень параметрів системи.

3 МОДЕЛІ СИСТЕМ РОЗПОДІЛУ ІНФОРМАЦІЇ

Теорія телетрафіка оперує не із самими системами розподілу інформації, а з їхніми математичними моделями. Для повного опису СРІ необхідно вказати імовірнісні процеси, що описують вхідний потік вимог, структуру системи та дисципліну обслуговування. Таким чином математична модель СРІ містить такі основні елементи:

1. *Вхідний потік вимог на обслуговування* (*трафік*) – класифікується за ознаками стаціонарності, ординарності та післядії. Основними характеристиками потоку вимог є його параметр та інтенсивність.

2. *Структура системи розподілу інформації* – це інформація про кількість обслуговуючих пристроїв або серверів (пристроїв, що надають послуги), їх взаємне з'єднання (схему) та доступність для вхідних вимог.

3. *Дисципліна обслуговування потоку вимог* – характеризує взаємодію потоку вимог із системою розподілу інформації. В теорії телетрафіка дисципліна обслуговування описується:

- способом обслуговування вимог;
- порядком обслуговування вимог;
- режимами пошуку виходів схеми (наприклад, довільний або груповий);
- законами розподілу тривалості обслуговування;
- наявністю переваг (пріоритетів) в обслуговуванні деяких вимог;
- наявністю обмежень при обслуговуванні (наприклад, за тривалістю очікування або обслуговування, кількості вимог, що очікують);
- законами розподілу імовірностей виходу з ладу елементів схеми.

У загальному випадку вхідний потік вимог на обслуговування описується функцією розподілу ймовірностей інтервалів часу між сусідніми вимогами $A(z)$:

$$A(z) = P(\leq z), \tag{3.1}$$

де $P(\leq z)$ – імовірність того, що час між послідовними вимогами $\leq z$.

Якщо інтервали часу між послідовними вимогами є незалежними та однаково розподіленими випадковими величинами, то вхідний потік вимог утворює стаціонарний *процес відновлення*. Це є ознакою незмінності в часі імовірнісних характеристик випадкових процесів, що у більшості випадків добре віддзеркалює реальні процеси в СМО за короткі проміжки часу. Таким чином, функція розподілу інтервалів $A(z)$ є достатньою для опису потоку вимог.

Час, протягом якого вимога перебуває у сервері, описується функцією розподілу ймовірностей тривалості обслуговування $B(x)$:

$$B(x) = P(\leq x), \tag{3.2}$$

де $P(\leq x)$ – імовірність того, що час обслуговування $\leq x$.

Для опису інтервалу часу між послідовними вимогами або тривалості обслуговування застосовуються різні закони. Найбільше з них використовуються розподіли, що подано далі та позначено певними літерами:

- *M* – експонентний (*M – марковська* модель);
- *H* – гіперекспонентний (*Hyper-exponential*);
- *D* – детермінований (*Determined*);

– *U* – рівномірний (*Uniform*);
– *E* – розподіл Ерланга;
– *G* – довільний або узагальнений (*General*).

Дисципліна обслуговування потоку вимог визначає правила обслуговування та долю вимог при їх надходженні до системи на обслуговування. Розрізнюють такі типи СМО, які визначаються способом обслуговування вимог:

1. *Системи з втратами* – вимоги, які при надходженні до системи не знаходять в ній жодного вільного сервера, отримують відмову в обслуговуванні та втрачаються.

2. *Системи з чергами* – вимоги, які не можуть бути обслужені відразу через зайнятість всіх серверів системи, стають в чергу, і за допомогою деякої дисципліни обслуговування черги визначається, у якому порядку вимоги, що очікують, вибираються із черги для обслуговування. Найбільш поширеними дисциплінами обслуговування черги є:
– FF (*FIFO – first in first out*) – вимоги з черги обслуговуються в порядку їхнього надходження (упорядкована черга);
– LF (*LIFO – last in first out*) – щоразу перевагу для обслуговування має вимога, що надійшла до черги останньою;
– SR (*SIRO – service in random order*) – наступна вимога для обслуговування із черги вибирається випадково (випадкова черга).

3. *Комбіновані системи* з чергами та втратами (системи з чергою при обмеженнях). Наприклад, очікувати може тільки кінцева кількість вимог, обумовлена кількістю місць очікування, меншою за нескінченність. Можливо й так – вимога втрачається тоді, коли час очікування в черзі або перебування в системі перевищує задані межі.

4. *Пріоритетні системи* – для вимог передбачено різні пріоритети в обслуговуванні. Якщо вимога, що надійшла, має високий пріоритет, а всі сервери зайняті, то вона або займає одне з перших місць у черзі, або тимчасово припиняє обслуговування вимоги низького пріоритету і займає її місце в сервері. При цьому можуть бути застосовані такі пріоритетні правила:
– абсолютний пріоритет з перериванням (*pre-emptive discipline*) – вимога високого пріоритету переривається обслуговування вимоги низького пріоритету. Може бути: абсолютний пріоритет із втратами (*pre-emptive loss discipline*), абсолютний пріоритет з дообслуговуванням (*pre-emptive resume discipline*) і абсолютний пріоритет з обслуговуванням заново (*pre-emptive repeat different discipline*);
– відносний пріоритет (*head of the line priority discipline*) – вимога високого пріоритету посідає перше місце в черзі і переривань немає.

Змішані пріоритети зумовлюють вибір абсолютного або відносного пріоритетного правила залежно від уже реалізованої частини тривалості обслуговування, а динамічні – залежно від типу поточних вимог і співвідношення кількості вимог різних пріоритетів, що є у серверах та в черзі.

Основні характеристики, що представляють структуру СРІ такі:

- кількість обслуговуючих пристроїв (серверів, ліній, каналів, портів);
- кількість місць очікування або максимальна довжина черги (ємність пам'яті, у якій накопичуються вимоги, що очікують);
- доступність – спосіб включення серверів, за якого кожній вимозі доступні всі або не всі (хоча всім вимогам у сукупності доступні всі) сервери. Схема може бути повнодоступною або неповнодоступною;
- взаємне з'єднання (схема) – спосіб включення серверів, за якого кожна вимога обслуговується одним сервером або декількома, але поетапно. Схема може бути однокаскадною або багатокаскадною (ланцюговою).

Структурні характеристики системи частково зумовлюють дисципліну обслуговування потоку вимог. Наприклад, за кількості місць очікування $r = 0$ буде система з втратами, при $0 < r < \infty$ – комбінована система з чергою та з втратами, а при $r = \infty$ – чиста система з чергою.

Для стислого запису досліджуваної системи Д. Кендаллом запропоновано спеціальне *умовне позначення базової моделі*, в якому зі всіх наведених параметрів математичної моделі СРІ представлено чотири елементи: *A*/*B*/*m*/*r*.

Елемент *A* характеризує потік вимог і певною літерою з наведених вище видів розподілів позначується функція розподілу ймовірностей інтервалів часу між сусідніми вимогами.

Елемент *B* характеризує випадкові послідовності тривалості обслуговування на окремих серверах системи і аналогічно до попереднього може використовувати такі ж розподіли.

Елементи *m* та *r* характеризують відповідно кількість обслуговуючих пристроїв і місць очікування у системі.

Умовне позначення базової моделі крім цих основних позначень може містити ще й додаткові символи, які вказуються після знаку „:” і можуть уточнювати особливості системи. Наприклад. Запис *M*/*D*/120/$r = \infty$ означає, що СМО обслуговує найпростіший потік вимог (*M*) за допомогою $m = 120$ обслуговуючих пристроїв, де кожна вимога має постійну тривалість обслуговування (*D*). У системі є нескінченна кількість місць очікування ($r = \infty$), що й визначає дисципліну обслуговування з чергами. Запис *G*/*M*/120:*Loss* означає, що СРІ обслуговує довільний потік вимог з експонентним розподілом їх тривалості. Ємність накопичувальної пам'яті, у якій вимоги очікують у разі зайняття всіх 120 серверів системи дорівнює нулю (не записується), через те дана система має дисципліну обслуговування із втратами (*Loss* – втрати).

Отже, базова математична модель СМО позначається послідовністю символів: перший – вказує функцію розподілу інтервалів часу між вимогами, другий – функцію розподілу тривалості обслуговування, третій і наступний (необов’язковий) символи – схему і дисципліну обслуговування.

Побудова математичної моделі (рис. 3.1), що адекватно відображає реальну систему розподілу інформації, у багатьох випадках є непростою задачею. Від правильного вибору моделі залежить точність вирішення задач аналізу, синтезу та оптимізації.

Рисунок 3.1 – Класифікація систем масового обслуговування

4 МАТЕМАТИЧНА МОДЕЛЬ ПОТОКУ ВИМОГ

У теорії масового обслуговування одним із основних понять є випадкова послідовність вимог, що надходять до системи і які необхідно обслужити. Сукупність (послідовність) подій надходження до системи в моменти $t_1 \ldots t_n$ вимог на обслуговування утворюють *потік вимог* (рис. 4.1).

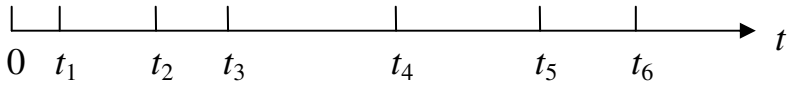


Рисунок 4.1 – Потік вимог на обслуговування

Потоки вимог визначаються моментами надходжень t_x і кількістю вимог k_n, що надійшли у момент t_n. При цьому k_n і t_n в загальному випадку випадкові. У *рекурентного потоку вимог* $k_n = 1$ для всіх $n = 1, 2, \ldots$, а інтервали часу між подіями надходження вимог $z_n = t_n - t_{n-1}$ є стохастично незалежними позитивними й однаково розподіленими випадковими величинами.

За незмінного інтервалу часу між вимогами z_n потік є *детермінованим*, однак в телекомунікаціях в основному потоки є *випадковими*.

Випадковий потік вимог може бути описаним у два способи.

1. Опис випадкового потоку вимог функцією розподілу ймовірностей інтервалів часу між сусідніми вимогами $F(t)$:

$$F(t) = P(z_n \le t), \tag{4.1}$$

де $P(z_n \le t)$ – імовірність того, що час між послідовними вимогами $z_n \le t$.

Основна характеристика такого потоку – це середнє значення інтервалів z, що для випадкової величини є математичним сподіванням $\bar{z}$. Параметр, зворотний до математичного сподівання $\bar{z}$, визначається як *інтенсивність потоку* надходження вимог λ за одну одиницю часу, якими вимірюється $\bar{z}$:

$$\lambda = \frac{1}{\bar{z}}. \tag{4.2}$$

Наприклад, при $\bar{z}$ = 0,1 с інтенсивність потоку λ = 10 вимог на секунду, а при $\bar{z}$ = 100 мс – інтенсивність потоку λ = 0,01 вимоги на мілісекунду.

Найпоширенішою математичною моделлю потоку вимог у телефонних мережах зв'язку є модель експонентного розподілу інтервалів часу між вимогами (викликами АТС) з параметром λ:

$$P(z_n \le t) = 1 - e^{-\lambda t}. \tag{4.3}$$

Густина (щільність) цього розподілу дозволяє розрахувати імовірність будь-якої тривалості $z_x = t$ випадкової величини z (інтервалів між вимогами) за заданої інтенсивності надходження вимог λ:

$$p(t) = \lambda e^{-\lambda t}. \tag{4.4}$$

Середнє значення випадкової величини t, розподіленої за експонентним законом (4.4), дорівнює λ^{-1} і тому з (4.2) випливає, що параметр даного розподілу λ – це теж середня кількість вимог за одиницю часу, в яких

вимірюється $\bar{z}$. Потік, де всі інтервали z_n мають однаковий експонентний розподіл з параметром λ, є дуже важливим прикладом рекурентного потоку.

На рис. 4.2 представлено два графіки експонентного розподілу (4.4), які позначено $p(z)$ та $p1(z)$ з параметрами λ = 0,5 та λ1 = 0,25 відповідно.

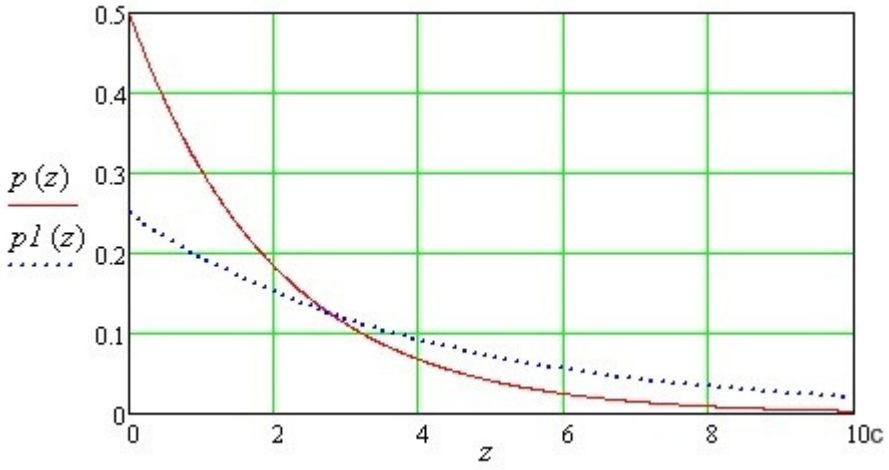


Рисунок 4.2 – Експонентний розподіл інтервалу z

З графіків видно, що для потоку, заданого функцією $p(z)$ є значно більшою імовірність (частка) коротких інтервалів між вимогами, ніж для потоку, заданого функцією $p1(z)$. Це свідчить про більшу інтенсивність потоку вимог, яка у першому випадку складає λ = 0,5, а у другому – λ1 = 0,25 вимог на одиницю часу. Відповідно до цього середнє значення інтервалу часу між вимогами складає $\bar{z} = 2$ та $\bar{z}1 = 4$ одиниць часу.

2. Опис випадкового потоку вимог функцією $P_i(t)$ – розподілом імовірностей кількості вимог i за умовну одиницю часу t. Наприклад, якщо діаграму процесу надходження вимог, що представлено на рис. 4.1, умовно поділити на однакові проміжки часу тривалістю t, що у рази або десятки разів перевищує середнє значення інтервалів $\bar{z}$, то на кожний з таких умовних інтервалів припаде випадкова кількість вимог i. Функція розподілу випадкової величини i й буде описувати потік вимог, що надходить до системи на обслуговування.

Відомо, якщо інтервал часу між подіями (вимогами) z розподілений за експонентним законом, то кількість таких подій i за умовну одиницю часу t буде розподілена за законом Пуассона:

$$P_i(t) = \frac{(\lambda t)^i}{i!} e^{-\lambda t}. \qquad (4.5)$$

Величина λt є параметром розподілу Пуассона. За цим розподілом можна розрахувати імовірність надходження до системи точно i вимог за умовну одиницю часу тривалістю t за заданої інтенсивності надходження вимог λ.

Для наведеного вище прикладу експонентного потоку з інтенсивністю λ = 0,5 побудовано розподіл Пуассона випадкової кількості вимог, що припадає на умовні інтервали часу, наприклад, тривалістю t = 20 с (рис. 4.3).

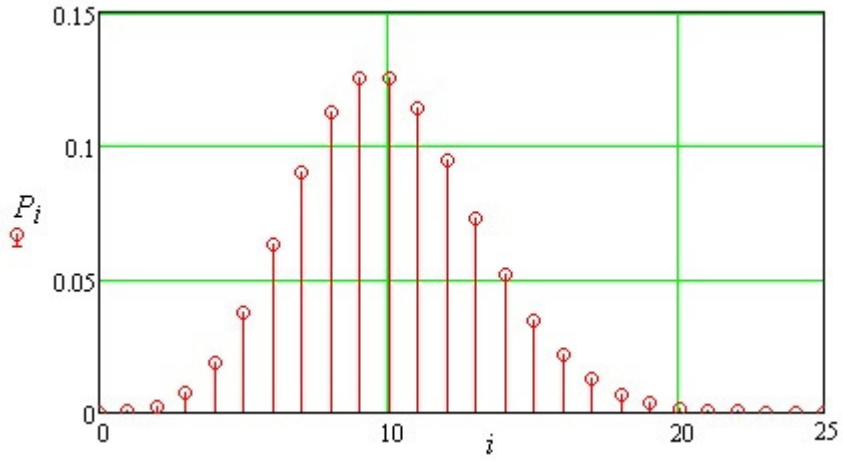


Рисунок 4.3 – Розподіл Пуассона при $\lambda = 0{,}5$ та $t = 20$ с

Середнє значення випадкової величини *i*, розподіленої за законом (4.5), визначається як $\bar{i} = \lambda t$, і у цьому випадку $\lambda t = 10$. З рис. 4.3 видно, що імовірність саме такого значення кількості вимог *i* за умовну одиницю часу *t* є найбільшою, і це є імовірність середнього значення $\bar{i}$. При зростанні *i* від нуля до $\bar{i}$ імовірність $P_i(t)$ поступово зростає, а далі – зменшується. Даний графік достатньо симетричний, а форма апроксимуючої кривої наближується до форми кривої нормального (Гаусса) закону розподілу випадкової величини.

Отже математичну модель потоку вимог, що надходить до СРІ на обслуговування, можна відобразити у два способи за допомогою імовірнісних функцій розподілу:

- інтервалів часу між сусідніми вимогами *z*, наприклад, (4.4);
- кількості вимог *i* за умовну одиницю часу *t*, наприклад, (4.5).

У першому випадку, як правило, застосовуються неперервні закони, а другому – дискретні.

Свою назву кожний із видів потоку запозичує від назви імовірнісного закону розподілу інтервалів часу між вимогами або їх кількості за умовну одиницю часу. Тому модель потоку, яка визначається розподілом (4.5) і використовується в телефонних мережах, називається *пуассонівським потоком*.

Пуассонівські потоки вимог поділяються на потоки *першого* та *другого роду*. Для потоків першого роду імовірність надходження вимог у систему не залежить від того, скільки або які вимоги вже є в ній. СМО, в які надходять такі потоки, називаються системами з нескінченним числом джерел або відкритими системами. Потоки першого роду виникають із накладення багатьох потоків вимог окремих джерел при тому, що поведінка кожного джерела невідома.

Потоки вимог другого роду утворюються в СМО з кінцевим числом джерел або в так званих замкнутих системах. Оскільки в момент, коли вимога обслуговується або очікує, її джерело вже не може породжувати нових вимог, то імовірність того, що надійде вимога із сукупності вимог всіх джерел, залежить від того, скільки вимог у системі й від яких вони джерел. Такі потоки називаються примітивними.

Очевидно, що оцінка якості обслуговування або пропускної здатності СРІ потребує врахування всіх елементів її моделі. Найбільш складним при цьому є врахування математичної моделі вхідного потоку вимог. Саме з цієї причини весь пакет задач аналізу й синтезу СРІ для будь-яких із її схем та дисциплін обслуговування вирішено тільки для випадку найпростішої моделі трафіка – моделі *пуассонівського* потоку. Для цієї моделі відомі всі аналітичні формули розрахунку основних характеристик якості обслуговування в системах розподілу інформації [1–3].

Стрімкий розвиток телекомунікаційних технологій, нові принципи побудови мереж зв'язку, зміна структурного складу абонентів і спектру надаваних послуг – все це дуже позначається на характері трафіка в мережах. Ці фактори, насамперед, збільшують нерівномірність інтенсивності потоків вимог, яка вимірюється дисперсією інтенсивності. Результати статистичних вимірів, виконуваних на різних етапах еволюції розвитку мереж і послуг, дають змогу виділити 3 типи трафіка, до яких слід вживати певні математичні моделі:

І тип – в моносервісних мережах з однорідним трафіком. Такими є суто телефонні мережі з єдиною послугою телефонного зв'язку, що й зумовило однорідність трафіка. Найпростіша модель пуассонівського потоку, в основному, відповідає таким умовам, а значення інтенсивності трафіка та її дисперсії співпадають або достатньо близькі.

ІІ тип – в мультисервісних мережах з різнорідним трафіком. Інтегральний характер мультисервісної мережі з розширеним спектром надаваних послуг зумовлює різнорідність трафіка, яка сильно змінює його параметри та математичну модель. Реальним потокам властива підвищена нерівномірність трафіка, за якої дисперсія інтенсивності трафіка перевищує її математичне сподівання від 2 до 15 раз. Іноді дане перевищення буває й більшим, але це відбувається або за межами ГНН, або на невеликих пучках каналів [4].

ІІІ тип – в пакетних мережах з мультисервісним трафіком. Трафік має довгострокові залежності в інтенсивності та ще більш суттєво відрізняється від пуассонівського потоку. Адекватною моделлю потоків в таких мережах є самоподібні процеси. В мультисервісних пакетних мережах трафік є різнорідним і з певними вимогами до *QoS*. Тут передачу потоків різних служб забезпечує одна і та ж сама мережа з єдиними протоколами та законами управління. Оскільки джерела кожної служби можуть мати різні швидкості передавання інформації або змінювати її в процесі сеансу зв'язку, то об'єднаному потоку пакетів властива так звана «пачковість» трафіка (*burstness*), вимірювана коефіцієнтом пачковості [1]. Ця пачковість обумовлює ще більшу нерівномірність трафіка, за якої дисперсія інтенсивності трафіка перевищує її математичне сподівання від 20 до 60 раз і більше.

Незалежно від способу надання математичної моделі потоку вимог вибрана модель обов'язково має бути адекватною реальним потокам трафіка телекомунікаційних мереж, оскільки від цього суттєво залежить точність розрахунку характеристик якості обслуговування та пропускної здатності СРІ при їх аналізі, синтезі та оптимізації.

5 НАВАНТАЖЕННЯ ТА ЙОГО ВИДИ

Сумарний час обслуговування всіх вхідних вимог є *навантаженням* для серверів (приладів, ліній, каналів) СМО. Чим більший цей час – тим більше навантаження „обслуговують" сервери системи. В теорії телетрафіка розрізняють *вхідне* (*traffic offered*), *обслужене* (*traffic carried*) та *надлишкове* (*overflow traffic*) навантаження. Надлишкове навантаження – це різниця між вхідним і обслуженим навантаженнями, яка для систем з втратами буде *втраченим* навантаженням.

5.1 Визначення та інтенсивність навантаження

Навантаження вимірюється в *годино-заняттях*. Наприклад, навантаження в одне годино-заняття (1 *г.-зан.*) утворюється безперервним обслуговуванням вимог одним сервером протягом однієї години або двома серверами терміном по півгодини кожний і т.д. Тому параметр „навантаження" не дає чітких відомостей щодо напруженості роботи системи, оскільки невідомо, за який час воно виконане. Сумарний час обслуговування всіх вимог, який дорівнює, наприклад, 20 г.-зан. може свідчити про роботу в системі 1 сервера протягом 20 годин або 20 серверів протягом 1 години кожний. Тому введено поняття *інтенсивності обслуженого навантаження Y*, яка визначається як приведений час навантаження. Він розраховується як сумарний час обслуговування всіх вимог x_i на деякому інтервалі часу $t_2 - t_1$, поділений на величину цього інтервалу ($t_2 - t_1$ може дорівнювати, наприклад, одній годині):

$$Y = \frac{\sum x_i}{t_2 - t_1}. \qquad (5.1)$$

Кількість серверів системи, що має бути залученою до обслуговування вимог вхідного потоку, теж є *навантаженням* для системи. При надходженні вимоги у систему один з серверів займається, а в кінці її обслуговування – цей сервер звільняється і вимога „виходить із системи". Через випадковість даних подій у моменти надходження вимог або моменти їх виходу в системі буде зайнятою різна кількість серверів. Таким чином, кожної миті кількість зайнятих серверів системи або кількість одночасно обслуговуваних вимог визначає *миттєве значення* обслуговуваного навантаження. Через те що ця кількість є випадковою, то основною характеристикою обслуговуваного навантаження є її середнє значення, що так само, як і (5.1), називається *інтенсивністю обслуженого навантаження Y*. Якщо в багатосерверній системі за час її роботі від t_1 до t_2 з будь-якою періодичністю, наприклад $\tau = 1$ с, $n = (t_2 - t_1) / \tau$ разів обчислити (сканувати) кількість зайнятих серверів C_i і у підсумку розрахувати їх середнє значення

$$Y = \frac{1}{n}\sum_{i=1}^{n} C_i, \qquad (5.2)$$

то воно однозначно співпаде зі значенням, розрахованим для цієї ж системи за формулою (5.1).

На діаграмі рис. 5.1. наочно показано функціонування 4-серверної системи при обслуговуванні вхідного навантаження На осі ординат відображено точки, що символізують номери певних серверів системи, а на осі абсцис – час, на який кожний із цих серверів займається вимогою для її обслуговування. Тут відображено тільки обслужене навантаження і не видно того, надходили до системи ще вимоги або ні.

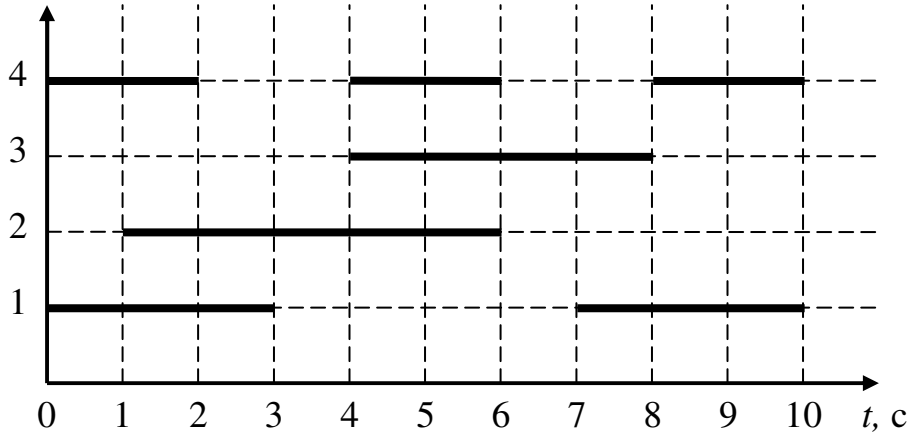


Рисунок 5.1 – Діаграма функціонування 4-серверної системи

Сумарний час зайняття всіх серверів системи визначиться так (індексами позначені номери серверів, що відображені на осі ординат):

$$\sum x_i = (3\text{с} + 3\text{с})_1 + (5\text{с})_2 + (4\text{с})_3 + (2\text{с} + 2\text{с} + 2\text{с})_4 = 21\,\text{с}\,.$$

Інтенсивність обслуженого навантаження за 10 с відповідно до формули (5.1) складе $Y = 21 / 10 = 2{,}1$ у.о. (умовних одиниць).

Середня кількість зайнятих серверів за інтервалу сканування 1 с (i – це номер відліку сканування на осі абсцис) складе:

$$Y = \frac{1}{10}\sum_{i=0}^{9} C_i = \frac{2+3+2+1+3+3+1+2+2+2}{10} = 2{,}1\ \text{ó.î.}$$

Отже інтенсивність обслуженого навантаження – це приведений час сумарного обслуговування всіх вимог (5.1) або середнє значення кількості зайнятих серверів (5.2). Обидва способи визначення інтенсивності обслуженого навантаження можуть використовуватись в засобах вимірювання параметрів навантаження на діючих мережах зв’язку. Середню кількість зайнятих серверів ще називають *завантаженістю системи*.

За одиницю виміру інтенсивності навантаження прийнято 1 *Ерланг* (1 Ерл), яку названо на честь засновника теорії телетрафіка А.К. Ерланга.

Навантаження на сервери системи є результатом спільного процесу надходження та обслуговування вимог. Оскільки вимоги надходять до системи через випадкові інтервали часу (або кількість вимог за умовну одиницю часу є випадковою) і тривалість обслуговування також може бути випадковою, то й *навантаження є випадковою величиною*. У зв’язку з цим *інтенсивність вхідного навантаження* Λ нормується середнім часом обслуговування вимог $\bar{x}$

(не плутати з інтенсивністю потоку надходження вимог λ). Ця інтенсивність визначається у відповідності до другого способу подання математичної моделі вхідного потоку вимог, наприклад (4.5), але при цьому за умовну одиницю часу t береться середня тривалість обслуговування вимог $\bar{x}$. Таким чином, інтенсивність вхідного навантаження

$$\Lambda = \lambda \bar{x}. \tag{5.3}$$

Для випадку, наведеному на рис. 4.3, величина $\lambda t = 10$ є інтенсивністю вхідного навантаження в 10 Ерл тільки за умови, якщо середня тривалість обслуговування $\bar{x} = t = 20$ с. Отже *інтенсивність вхідного навантаження* Λ, вимірювана в *Ерлангах*, це середня кількість вимог, що надходить до системи за середній час обслуговування однієї вимоги. Інакше, це є інтенсивність потоку надходження вимог λ, яка віднесена до середньої тривалості обслуговування $\bar{x}$.

Інтенсивність вхідного навантаження можна визначати як інтенсивність обслуженого навантаження у припущенні того, що втрати вимог вхідного потоку відсутні, тобто кожній вимозі, що надійшла, надається вільний сервер за будь-яких умов (наприклад, у системі нескінченна кількість серверів).

В основному при дослідженнях розглядаються стаціонарні потоки вимог. У цьому випадку як вхідне, так і обслуговуване навантаження описуються стаціонарними випадковими процесами і їхні статистичні параметри у ймовірнісному змісті не залежать від часу.

Інтенсивність вхідного навантаження Λ – це середня кількість вимог, що надходить до системи за середній час обслуговування однієї вимоги. Інтенсивність обслуговуваного навантаження Y – це середня кількість зайнятих серверів. В обох випадках дані інтенсивності є оцінками навантажень, які є випадковими величинами.

У відповідності до (1.1) математичне сподівання, а отже й середню кількість вимог, що надходить до системи за середній час обслуговування однієї вимоги, можна розрахувати так:

$$\Lambda = \sum_{i=0}^{\infty} iP_i, \tag{5.4}$$

де P_i – імовірність того, що за інтервал часу $t = \bar{x}$ до системи надійде точно i вимог, де $i = 0, \ldots, \infty$.

У відповідності до (1.1) математичне сподівання, а отже й середню кількість зайнятих серверів системи, можна розрахувати так

$$Y = \sum_{j=0}^{m} jp_j, \tag{5.5}$$

де P_j – імовірність того, що в довільний момент часу у системі з m серверів зайнято точно j серверів, де $j = 0, \ldots, m$.

Математичне сподівання інтенсивності навантаження іноді називають просто навантаженням, що є неточним визначенням.

5.2 Дисперсія та скупченість навантаження

Крім інтенсивності навантаження (математичне сподівання) важливою характеристикою випадкової величини навантаження є дисперсія, яка для вхідного та обслуговуваного навантаження відповідно визначиться так:

$$D_\Lambda = \sum_{i=0}^{\infty} (i-\Lambda)^2 P_i ; \qquad D_Y = \sum_{i=0}^{m} (j-Y)^2 P_j . \tag{5.6}$$

Якщо потік вхідних вимог є експонентним (4.4), то створюване їм навантаження, як випадкова величина, має розподіл Пуассона (4.5). Для випадкової величини, що описується цим розподілом, характерна однаковість перших двох моментів, тобто дисперсія навантаження D_Λ збігається з її математичним сподіванням Λ. Таке навантаження називається пуассонівським навантаженням першого роду і воно вважається умовно рівномірним.

Якщо дисперсія навантаження менше її математичного сподівання, то навантаження називають згладженим, оскільки його відхилення від середнього значення будуть менше, ніж для пуассонівського навантаження.

Навантаження, у якого дисперсія більше математичного сподівання отримало назву скупченого. У цьому випадку вимоги надходять не рівномірно: для деяких інтервалів часу кількість вхідних вимог мала, а на інших інтервалах їх кількість сягає значної величини, тобто вимоги групуються на коротких інтервалах часу. Наприклад, скупчене навантаження створюється так званим надлишковим потоком вимог, що втрачені (не обслужені) у системі *A* і надходять для обслуговування на іншу систему *B*. Цей потік є переривчастим, тому що на систему *B* вимоги можуть надходити тільки за умови, що в системі *A* відсутні вільні сервери.

Скупченість навантаження *S* визначається як відношення дисперсії навантаження D_Λ до її математичного сподівання Λ:

$$S = \frac{D_\Lambda}{\Lambda}. \tag{5.7}$$

Величина *S*, що називається коефіцієнтом скупченості навантаження, дорівнює одиниці для пуассонівського навантаження, менше одиниці для вирівняного (згладженого) навантаження і більше одиниці для скупченого (надлишкового) навантаження.

Якщо на сервери системи надходять відразу *n* потоків вимог, то інтенсивність об'єднаного потоку буде дорівнювати підсумку математичних сподівань Λ_i. Для статистично незалежних потоків дисперсія об'єднаного потоку буде дорівнювати підсумку дисперсій D_i відповідних навантажень. Таким чином, математичне сподівання Λ і дисперсія D_Λ сумарного навантаження розраховуються за наступними формулами:

$$\Lambda_\Sigma = \sum_{i=0}^{n} \Lambda_i ; \qquad D_\Sigma = \sum_{i=0}^{n} D_i . \tag{5.8}$$

Слід розрізняти скупченість вхідного та обслуженого навантажень. Безумовно, $S_\Lambda > S_Y$, оскільки в системі обслуговується тільки частка вхідного навантаження та й обслуговуване навантаження згладжується системою.

6 ХАРАКТЕРИСТИКИ ЯКОСТІ ОБСЛУГОВУВАННЯ

Для будь-якої телекомунікаційної системи важливою є оцінка ступеня задоволення потреби в обслуговуванні, або якість обслуговування (*QoS – Quality of Service*). В теорії телетрафіка якість обслуговування потоку вимог характеризується можливістю негайного обслуговування вимоги або тривалістю очікування початку обслуговування. З математичної моделі СРІ випливає, що ці можливості визначаються обраною дисципліною обслуговування вимог. Через те для кожної дисципліни обслуговування вимог властивий певний набір основних і допоміжних характеристик якості обслуговування.

6.1 Системи з втратами

З економічних причин СРІ можуть проектуватися з дисципліною обслуговування із втратами, за якої вимозі, що надходить до системи в момент відсутності вільних обслуговуючих пристроїв, відмовляється в обслуговуванні і вона відразу ж покидає її та втрачається. Основною кількісною оцінкою якості обслуговування в цьому випадку є *імовірність втрати вимоги* P_B (*B – blocking*, втрати). Імовірність P_B на відрізку часу (t_1, t_2) визначається як відношення кількості втрачених за цей відрізок часу вимог $C_B(t_1, t_2)$ до загальної кількості вимог, що надійшли за той самий час до системи $C(t_1, t_2)$:

$$P_{\hat{A}} = \frac{C_{\hat{A}}(t_1, t_2)}{C(t_1, t_2)}. \qquad (6.1)$$

Допоміжними характеристиками *QoS* є імовірність втрати навантаження (кількість вимог на умовну одиницю часу) та імовірність втрат за часом, які використовуються рідко. Імовірність втрати навантаження визначається як відношення інтенсивності втраченого навантаження до вхідного, а імовірність втрат за часом – це сумарна частка часу з проміжку часу (t_1, t_2), за якої були зайняті всі сервери системи.

Середня кількість вимог у системі N характеризує ступінь завантаженості системи і співпадає з середньою кількістю зайнятих серверів, що є інтенсивністю обслуговуваного навантаження *Y*. Всі інші вимоги, що надходять і не знайшли вільного сервера, втрачаються і до системи не потрапляють.

Середній час перебування вимоги в системі T співпадає з середнім часом обслуговування вимоги $\bar{x}$.

Для систем із втратами інтенсивність обслуженого навантаження менша вхідного навантаження на величину ΛP_B, тобто

$$Y = \Lambda(1 - P_B).$$

Величина ΛP_B є інтенсивністю надлишкового навантаження і його потік за своєю структурою суттєво відрізняється від потоку навантаження, що надходить до системи, більш нерівномірним характером.

6.2 Системи з чергами

Для кількісної оцінки якості обслуговування систем з чергою розраховують такі основні характеристики:

- імовірність очікування $P_{w>0}$ або середню частку затриманих вимог;
- середню довжину черги Q;
- середню тривалість очікування для затриманих вимог t_q;
- середню тривалість очікування для будь-якої вимоги W.

Імовірність $P_{w>0}$ на відрізку часу $(t_1,\ t_2)$ визначається як відношення кількості вимог, що потрапили за цей відрізок часу в чергу $C_Q(t_1,\ t_2)$ до загальної кількості вимог, що надійшли за той же час до системи $C(t_1, t_2)$:

$$P_{w>0} = \frac{C_Q(t_1, t_2)}{C(t_1, t_2)}. \tag{6.2}$$

Довжина черги є ключовим параметром якості обслуговування (та показником ефективності функціонування CPI) і визначається кількістю вимог, що очікують обслуговування. Довжина черги залежить від того, коли і скільки вимог надійшло в систему, скільки часу витрачено на обслуговування вимог, що надійшли, і т.д. Оскільки довжина черги є випадковою величиною, то як показник довжини черги використовується її математичне сподівання Q.

Середній час очікування в черзі t_q утворюється за рахунок затримки вимог у черзі. Він залежить від кількості вимог, що знаходяться в даний момент у черзі, часу закінчення обслуговування всіх попередніх вимог і т.д.

Середній час очікування в системі W являє собою середнє значення часу очікування, віднесене до всіх вимог – затриманих і незатриманих. Цей параметр вводиться через те, що не всі вимоги потрапляють до черги, а частина з них за наявності вільних серверів системи обслуговується негайно.

Допоміжними характеристиками *QoS* є середня кількість вимог у системі N та середній час перебування вимоги в системі T. Вони є допоміжними тому, що їх можна розрахувати із основних характеристик.

Середня кількість вимог у системі N визначає ступінь завантаженості системи і за необмеженої черги складається з середньої кількості вимог, що надходять до системи Λ, та тих, що очікують в черзі Q:

$$N = \Lambda + Q. \tag{6.3}$$

Середній час перебування вимоги в системі T – це час, проведений однією вимогою в системі й усереднений за всіма вимогами (затриманим і незатриманим). Він складається із середнього часу обслуговування $\bar{x}$ і середнього часу очікування вимог у системі W:

$$T = \bar{x} + W. \tag{6.4}$$

Для кожної моделі потоку всі характеристики якості обслуговування перебувають у певній функціональній залежності.

6.3 Комбіновані системи (з чергами та втратами)

Система з чергою, в якій введено обмеження на максимальну кількість вимог, що можуть перебувати в черзі, або на максимальний час очікування початку обслуговування є системою з комбінованою дисципліною обслуговування. За обмеженої кількості місць очікування (максимальна довжина черги) у випадку надходження вимоги в момент, коли всі сервери і місця очікування зайняті попередніми вимогами, дана вимога втрачається. За обмеженого часу очікування якщо вимога перебуває в черзі понад допустимий час, то їй відмовляється в обслуговуванні і вона теж втрачається. Тому окрім характеристик *QoS* чистої системи з чергами розраховуються й такі:

– імовірність втрати вимоги P_B (із-за обмеженої довжини черги);
– імовірність очікування понад припустимий час $P_{W>t}$.

Імовірність $P_{W>t}$ залежить від дисципліни обслуговування черги і найбільш простим для розрахунку є випадок упорядкованої черги *FIFO*. Цю імовірність ще називають *умовними втратами*, оскільки вимога, що очікує понад припустимий час t може втратити свою актуальність для користувача.

6.4 Пріоритетні системи

Сучасні телекомунікаційні системи і мережі характеризуються наявністю пріоритетного обслуговування переданих і оброблюваних даних. Аналітичні методи дослідження пріоритетних дисциплін обслуговування вимог розроблені, в основному, для дисциплін з одним класом пріоритетів, а більшість результатів отримані при різних допущеннях і припущеннях, що обмежують їхнє застосування на практиці. Детальне дослідження пріоритетних систем обслуговування припускає застосування комбінованого підходу до моделювання, що дозволяє, як показує практика, отримати результати, що значно змінюють уявлення про вплив пріоритетів на якість функціонування пріоритетних систем обслуговування

Комбіновані системи з обмеженнями на довжину черги та час очікування є найбільш поширеними в телекомунікаціях. В умовах різнорідного трафіка (мова, відео, дані) цю систему доповнюють механізмом пріоритетів, в якому всі вимоги поділяють на категорії і вимоги більш високої категорії при обслуговуванні мають певні переваги (пріоритети) перед вимогами більш низької категорії. Для кількісної оцінки якості обслуговування систем із пріоритетами розраховуються такі ж характеристики, як і для системи з чергами, але для кожного з введених пріоритетів окремо. Наприклад, середній час очікування в черзі вимоги k-го пріоритету, середня кількість вимог у системі k-го пріоритету тощо.

Прикладом системи з втратами може бути телефонна станція – абонент, що є ініціатором телефонного виклику, отримує відмову в обслуговуванні, якщо необхідний канал зв'язку вже зайнятий (технологія комутації каналів). Модель системи з чергою, комбінованої системи та з пріоритетами застосовувана для мереж, що базуються на технології комутації пакетів.

6.5 Пропускна здатність і продуктивність

У будь-якій з наведених СРІ якість обслуговування істотно впливає на такі характеристики системи, як пропускна здатність і продуктивність. При цьому критеріями якості обслуговування для систем із втратами є імовірність втрати вимоги, а для систем з чергами – імовірність очікування. Чим більше припустима норма втрат, тим менша якість обслуговування.

Пропускна здатність – це максимальна інтенсивність навантаження, що може надходити до системи при забезпеченні заданої якості обслуговування.

За малої імовірності втрат інтенсивність обслуженого навантаження близька до інтенсивності вхідного навантаження (пропускною здатністю часто вважають інтенсивність обслуженого навантаження). Однак при великих втратах ця відмінність суттєва і тому, чим менша якість обслуговування, тим більшою буде пропускна здатність (і навпаки). Крім того, чим вище норма якості обслуговування, тим більше серверів необхідно для забезпечення певної пропускної здатності.

Пропускна здатність системи не дорівнює кількості серверів в ній, оскільки інтенсивність обслуженого навантаження – це середня кількість зайнятих серверів. Через випадковість потоків навантаження ця кількість не досягає кількості каналів в системі. Задача лише полягає в наближенні середньої кількості зайнятих серверів до кількості серверів системи.

Продуктивність – це гранична, статистично усереднена кількість вимог, які будуть обслужені системою за одиницю часу при заданій якості обслуговування. Ця характеристика використовується, як правило, для оцінки систем управління та керуючих пристроїв.

Пропускна здатність та продуктивність СРІ залежать не тільки від імовірності втрат, але й від структури системи (кількості серверів і схеми їх включення) та дисципліни обслуговування і закону розподілу тривалості обслуговування. Крім того на пропускну здатність та якість обслуговування суттєво впливає вид потоку вимог або його математична модель.

При вимірі пропускної здатності системи розподілу інформації важливим є поняття блокування вимоги, що позначає подію, яка складається у відсутності вільних і доступних шляхів з'єднання до необхідного напрямку в момент надходження вимоги (або, точніше, у момент спроби встановити з'єднання). У системі з втратами блоковані вимоги втрачаються і основним показником пропускної здатності є частка втрачених вимог.

Для системи з чергою розрізняють випадок необмеженої та обмеженої черги (кількість місць очікування). Якщо нагромаджувач черги має необмежену ємність, то основними характеристиками є середній час очікування та імовірність очікування (імовірність блокування). У випадку нагромаджувача з обмеженою ємністю додається ще одна характеристика – імовірність втрати вимоги, тоді імовірність блокування дорівнює сумі двох ймовірностей – очікування та втрати вимоги.

У системі з повторними спробами підраховується кількість повторних спроб на одну вимогу та інші характеристики.

7 АНАЛІЗ СМО З ПУАССОНІВСЬКИМ ПОТОКОМ ВИМОГ

Дослідження функціонування СМО під впливом випадкових факторів можливо тільки за допомогою випадкових процесів. Випадковий процес – це функція $X(t)$, значення якої випадкові величини. Вибір випадкових процесів, використовуваних для опису й аналізу систем, залежить від структури й типу системи, від припущення про незалежність або залежність випадкових величин у системі, від виду їхніх функцій розподілів.

Якщо всі функції розподілу, що характеризують поведінку елементів системи, експонентні, то систему можна описати за допомогою однорідних безперервних *марковських ланцюгів* або його підвиду однорідних процесів *розмноження та загибелі*, як моделі СМО. При цьому аналітичне визначення величин, що характеризують систему, є відносно простим.

В умовах випадкових потоків вимог розрахунки ступеня завантаженості або пропускної здатності та характеристик якості обслуговування виконуються на основі імовірнісних функцій розподілу станів системи, які й визначають ці характеристики. За пуассонівського потоку вимог для визначення *стаціонарних* імовірностей станів системи використовується математичний апарат марковського процесу, який потребує складання та розв'язання системи лінійних алгебраїчних рівнянь рівноваги. *Стан системи* – це поточна кількість вимог системи, що обслуговуються в серверах або очікують в черзі. Зміна станів системи – це випадковий процес, який є результатом спільного процесу надходження та обслуговування вимог. Закон розподілу станів системи, на відміну від середньої кількості вимог в ній (завантаженості), більш повно характеризує функціонування СМО під впливом випадкових факторів.

Найбільш ефективний математичний апарат аналізу розроблений для систем, функціонування яких описується однорідними ланцюгами або процесами Маркова. Основні характеристики марковських моделей визначаються рішеннями систем лінійних рівнянь (алгебраїчних, диференціальних або інтегральних). Однак припущення про марковість функціонування системи є досить обмежуючим, тому в якості математичних моделей досить складних реальних систем використовуються більш загальні класи випадкових процесів. Разом з тим марковості моделі іноді вдається досягти ускладненням фазового простору моделюючого процесу.

Процес є марковським (без післядії), якщо для будь-якого моменту часу поведінка системи в майбутньому залежить тільки від поточного стану системи і не залежить від того, коли і як система цього стану дісталася. Марковський процес з *дискретним простором* станів називається ланцюгом Маркова. Множина $\{S\}$ утворює ланцюг Маркова, якщо імовірність того, що наступний стан буде S_{k+1} залежить тільки від поточного стану S_k, або вплив всієї передісторії процесу цілком зосереджено на поточному стані.

Таким чином в ланцюзі Маркова потік випадкових величин визначається тільки імовірністю переходу від попереднього значення випадкової величини (стану системи) до наступного. Для *однорідного* ланцюга Маркова імовірності цих переходів не залежать від номера кроку, на якому система дістається

певного стану. Крім того, згідно з *теоремою Маркова*, якщо кількість станів системи *S* кінцева, і з кожного стану можна перейти за певну кількість кроків в будь-який інший стан, то граничні імовірності станів існують і не залежать від початкового стану системи.

Умовою застосування ланцюгів Маркова для визначення ймовірностей станів системи є експонентний розподіл інтервалів часу між вимогами потоку. Експонентний розподіл єдиний зі всіх неперервних законів, якому властива *відсутність післядії*, яка полягає в наступному: якщо тривалість інтервалу між вимогами експонентна, то з будь-якого моменту від початку інтервалу залишок часу до кінця інтервалу не залежить від подій до цього моменту і буде мати експонентний розподіл з тим же параметром, що і весь інтервал.

Оскільки експонентний розподіл інтервалів часу між вимогами призводить до пуассонівського розподілу кількості вимог за умовну одиницю часу, то властивість відсутності післядії переноситься й на пуассонівський потік вимог, який разом з тим є:

- *стаціонарним*, тому що імовірність кількості вимог на відрізку часу залежить тільки від тривалості цього відрізку і не залежить від того, де саме на осі часу він розміщений;
- *ординарним*, тому що імовірність надходження більше однієї вимоги за нескінченно малий відрізок часу є нескінченно малою порівняно з імовірністю надходження точно однієї вимоги (це означає, що вимоги надходять тільки по одній);
- *без післядії*, тому що для будь-яких відрізків часу, що не перетинаються, кількість вимог одного відрізку не залежить від того, скільки вимог надійшло на інший.

Для такого потоку інтенсивність λ, тобто середня кількість вимог на одиницю часу, є величина незмінна.

Якщо в моделі не всі розподіли експонентні, шукають такі аналітичні прийоми, які призводять досліджувані процеси до *марковських*, або шукають такі моменти часу, в яких процес стає *марковським*. Далі простими методами дискретних *марковських* ланцюгів обчислюються шукані величини або імовірності, що характеризують стан системи, після чого вони перетворюються у відповідні величини вихідного процесу. Для розв'язання задач такого типу застосовуються методи напівмарковських процесів або вкладених ланцюгів Маркова. Однак реальні результати з застосуванням цих методів отримані для невеликого класу систем (наприклад, формула Поллачека-Хінчина для системи типу $M/G/1/\infty$), або отримано тільки часткові та наближені результати (для систем типу $M/G/m/\infty$, $GI/G/1/\infty$ та ін.).

Прагнення до максимальної точності опису функціонування реальних складних систем приводить до того, що відповідні математичні моделі стають усе більше й більше складними. При цьому ускладнюється їх математичний аналіз, апарат аналізу стає громіздким і часто недоступним в інженерній практиці.

7.1 Система з втратами *M*/*M*/*m*

Повнодоступна схема СРІ з m серверами обслуговує пуассонівський потік вимог за дисципліною із втратами. Інтервал часу між вимогами та тривалість їх обслуговування мають експонентні закони розподілу. Параметр потоку вимог λ, а середнє значенням тривалості обслуговування – $\bar{x}$. Необхідно визначити стаціонарний розподіл імовірностей станів системи P_k ($k = 0 \ldots m$) та характеристики якості обслуговування (*QoS*).

Для ординарного потоку *інтенсивність надходження* вимог співпадає з параметром потоку λ. Сукупність моментів завершення обслуговування вимог утворюють потік звільнення серверів системи з параметром μ. Цей параметр називається *інтенсивністю обслуговування* вимог і розраховується як зворотна величина до тривалості обслуговування:

$$\mu = \frac{1}{\bar{x}}. \tag{7.1}$$

Дискретні стани системи S_k змінюються при кожній події надходження вимоги або закінченні її обслуговування. Ланцюг Маркова з кінцевою кількістю станів відображається у виді діаграми переходів (рис. 7.1).

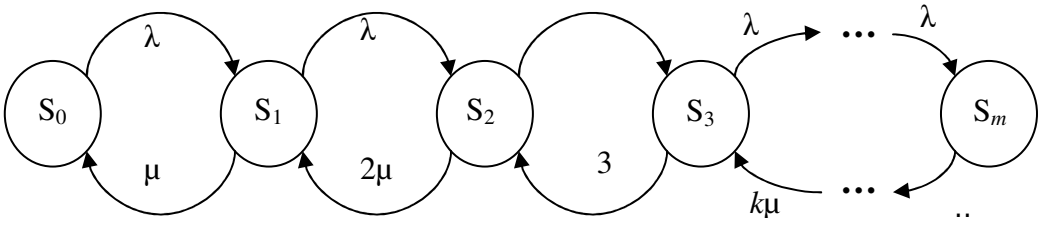


Рисунок 7.1 – Діаграма переходів m-серверної Системи з втратами

Для пуассонівського потоку інтенсивність надходження вимог λ є величина незмінна. Якщо система знаходиться в стані S_k і надходить вимога, то система переходить з однаковою інтенсивністю λ в стан S_{k+1} за будь-якого k. В разі закінчення обслуговування вимоги інтенсивність, з якою система переходить із стану S_k в стан S_{k-1} залежить від кількості зайнятих серверів або поточного стану системи. Якщо, наприклад, система знаходиться в S_1, то по закінченні обслуговування вимоги, яка займає один сервер, система переходить в стан S_0 з інтенсивністю μ. Якщо система в стані S_2, то вона може перейти в стан S_1 з інтенсивністю вже 2μ, тому що може закінчитися обслуговування будь-якої з двох вимог, які займають два сервери системи, і т.д. Таким чином, якщо обслуговуванням зайняті k серверів, то потік звільнення серверів, що переводить систему із станів S_k в S_{k-1}, буде в k раз інтенсивніше.

Стаціонарний розподіл імовірностей станів системи P_k для однорідного дискретного ланцюга Маркова визначається системою лінійних алгебраїчних рівнянь:

$$P_k = \sum_i P_i p_{i,k}, \tag{7.2}$$

де $p_{i,k}$ – імовірність переходу системи зі стану i в стан k [5]. При цьому розподіл (7.2) має задовольняти умові нормування, за якої $\sum_{k=0}^{m} P_k = 1$.

Сума в (7.2) обчислюється для всіх можливих варіантів станів i, з яких можна перейти до стану k. За ординарного потоку в цей стан можна перейти тільки зі станів $k-1$, k та $k+1$. Тому за рахунок зменшення кількості можливих імовірностей $p_{i,k}$ система рівнянь (7.2) скоротиться:

$$P_k = P_{k-1} p_{k-1,k} + P_k p_{k,k} + P_{k+1} p_{k+1,k}. \qquad (7.3)$$

Перехід із стану $k-1$ в стан k можливий в разі надходження вимоги, імовірність чого пропорційна параметру λ, а перехід із $k+1$ в стан k можливий в разі закінчення обслуговування вимоги, імовірність чого за експонентного розподілу тривалості обслуговування x пропорційна параметру μ. Таким чином імовірність $p_{k-1,k} = \lambda$, а імовірність $p_{k+1,k} = \mu(k+1)$ (див. рис. 7.1).

Імовірність переходу системи зі стану k в стан k очевидна – система не має піти зі стану k ні в стан $k-1$, ні в стан $k+1$, бо її стан незмінний:

$$p_{k,k} = 1 - p_{k,k-1} - p_{k,k+1} = 1 - \mu k - \lambda. \qquad (7.4)$$

Отже, із (7.3) з урахуванням (7.4) у підсумку маємо

$$P_k = \lambda P_{k-1} + P_k[1 - \mu k - \lambda] + \mu(k+1) P_{k+1}. \qquad (7.5)$$

Після розкриття дужок і об'єднання послідовно отримуємо:

$$P_k = \lambda P_{k-1} + P_k - \mu k P_k - \lambda P_k + \mu(k+1) P_{k+1},$$

$$0 = \lambda P_{k-1} - (\lambda + \mu k) P_k + \mu(k+1) P_{k+1},$$

а після переносу:

$$(\lambda + \mu k) P_k = \lambda P_{k-1} + \mu(k+1) P_{k+1}. \qquad (7.6)$$

Система рівнянь (7.6) описує стаціонарний режим в ланцюзі Маркова і є математичною моделлю процесу обслуговування вимог потоку. В цьому режимі виконується закон збереження інтенсивності імовірностей – інтенсивність потоку імовірностей в стан k дорівнює інтенсивності потоку із цього стану в стани $k-1$ або $k+1$. Тому (7.6) є рівнянням рівноваги або балансу.

Розв'язання рівняння балансу розпочнемо зі значення $k = 0$, за якого із (7.6) отримуємо

$$\lambda P_0 = \lambda P_{-1} + \mu P_1.$$

Імовірність $P_{-1} \equiv 0$ як імовірність неіснуючого стану і тому $\lambda P_0 = \mu P_1$, звідки випливає, що:

$$P_1 = \frac{\lambda}{\mu} P_0.$$

При $k = 1$ з (7.6) отримуємо $(\lambda + \mu) P_1 = \lambda P_0 + 2\mu P_2$ і тому

$$P_2 = \frac{\lambda}{2\mu} P_1 = \frac{\lambda}{2\mu}\frac{\lambda}{\mu} P_0 = \left(\frac{\lambda}{\mu}\right)^2 \frac{1}{2\cdot 1} P_0 .$$

При $k = 2$ з (7.6) отримуємо $(\lambda + 2\mu)P_2 = \lambda P_1 + 3\mu P_3$ і тому

$$P_3 = \frac{\lambda}{3\mu} P_2 = \frac{\lambda}{3\mu}\frac{\lambda}{2\mu}\frac{\lambda}{\mu} P_0 = \left(\frac{\lambda}{\mu}\right)^3 \frac{1}{3\cdot 2\cdot 1} P_0 .$$

Без продовження з наведених розрахунків (індукції) видно, що

$$P_k = \frac{\lambda}{\mu k} P_{k-1}, \tag{7.7}$$

та

$$P_k = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} P_0 \qquad (k = 1, 2, \ldots m). \tag{7.8}$$

З урахуванням умови нормування, внесеної до стаціонарного розподілу імовірностей станів системи (7.2), імовірність P_0 визначиться так:

$$P_0 = \left[\sum_{i=0}^{m}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!}\right]^{-1} . \tag{7.9}$$

Розподіл (7.8) є шуканим стаціонарним розподілом імовірностей станів повнодоступної системи з втратами, яка обслуговує пуассонівський потік вимог. Дана задача вперше розв’язана А.К. Ерлангом в припущенні про експонентний розподіл тривалості обслуговування, але згодом доведено, що (7.8) вірно за будь-якого довільного закону, тобто й для моделі *M*/*G*/*m*.

Відносна простота розв’язання даної задачі пояснюється наступним. Тут досліджується сталий режим системи, який досягається системою на нескінченному відрізку часу, коли система обслуговування працює в стані статистичної рівноваги. Для того щоб отримані імовірності станів системи (кількості зайнятих серверів) не залежали від того, у якому стані система була у початковий момент, даний процес має бути *ергодичним*. Відповідно до теореми Маркова будь-який *транзитивний* (з будь-якого стану можна перейти в будь-який інший) *однорідний* (імовірності переходу зі стану в стан не залежать від того, у який момент часу початок переходу) процес з кінцевою кількістю станів має ергодичну властивість [6, с. 149]. Для будь-якого марковського процесу, що має ергодичну властивість, після досить тривалого часу функціонування системи обов'язково наступить стаціонарний режим, де імовірність того, що система буде в *k*-му стані, не залежить від того, у якому стані вона знаходилася в початковий момент часу. Відповідно до зазначеного в ергодичній теоремі [7] достатньому критерію ергодичності марковського процесу, процес буде ергодичним, якщо виконується умова:

$$\frac{\lambda}{\mu} < m. \qquad (7.10)$$

Це означає, що в середньому повинно надходити менше вимог, ніж їх можна обслужити, або середня кількість вимог за одиницю часу обслуговування не повинна перевищувати кількості серверів.

З урахуванням (5.3) і (7.1) з'ясовується, що параметром стаціонарного розподілу імовірностей станів системи (7.8) і (7.9) є інтенсивність вхідного навантаження Λ, і таким чином можна записати:

$$P_k = \frac{\frac{\Lambda^k}{k!}}{\sum_{i=0}^{m} \frac{\Lambda^i}{i!}} \qquad (k = 1, 2, \ldots m). \qquad (7.11)$$

Розподіл (7.11) називається *першим розподілом Ерланга*, за яким можна розрахувати імовірність зайнятості k серверів повнодоступної m-серверної системи з втратами при обслуговуванні пуассонівського потоку вимог.

Для стану $k = m$ з (7.11) отримуємо *B-формулу Ерланга*, що позначається $E_m(\Lambda)$, і за якою можна розрахувати основну характеристику якості обслуговування потоку вимог – імовірність втрати вимоги P_B (6.1). Для спрощення обчислень значення $E_m(\Lambda)$ наводяться в таблицях Ерланга.

Фактично при $k = m$ з (7.11) отримуємо значення імовірності зайняття всіх серверів системи або імовірності втрат за часом (див. розд. 6). Чергова вимога, що надходить до системи, буде втрачена (заблокована) в тому випадку, коли всі сервери зайняті. Але тільки за умов пуассонівського потоку значення імовірності зайняття всіх серверів системи співпадає з імовірністю втрати вимоги незалежно від того, надходить в цей момент вимога або ні. Через властивість відсутності післядії інтенсивність надходження вимог не залежить від стану системи.

Для односерверної системи $M/G/1$ із (7.11) за умови $m = 1$ можна розрахувати:

– імовірність того, що сервер вільний P_0:

$$P_0 = \frac{1}{1+\Lambda};$$

– імовірність того, що сервер зайнятий P_1:

$$P_1 = 1 - P_0 = P_B = E_m(\Lambda) = Y = N = \frac{\Lambda}{1+\Lambda};$$

Як бачимо, імовірність P_1 співпадає з імовірністю втрати вимоги P_B і з інтенсивністю обслуженого навантаження Y, а також із середньою кількістю вимог у системі N.

7.2 Система з необмеженою чергою *M/M/m/*∞

Повнодоступна схема СРІ з m серверами та необмеженою кількістю місць очікування в черзі обслуговує пуассонівський потік вимог з дисципліною обслуговування черги *FIFO*. Інтервал часу між вимогами та тривалість їх обслуговування мають експонентні закони розподілу. Параметр потоку вимог λ, а середнє значенням тривалості обслуговування – $\bar{x}$. Визначити стаціонарний розподіл імовірностей станів системи P_k ($k = 0 \ldots \infty$) та характеристики *QoS*.

Дискретні стани системи S_k змінюються при кожній події надходження вимоги або закінченні її обслуговування. Ланцюг Маркова з кінцевою кількістю станів відображається у виді діаграми переходів (рис. 7.2).

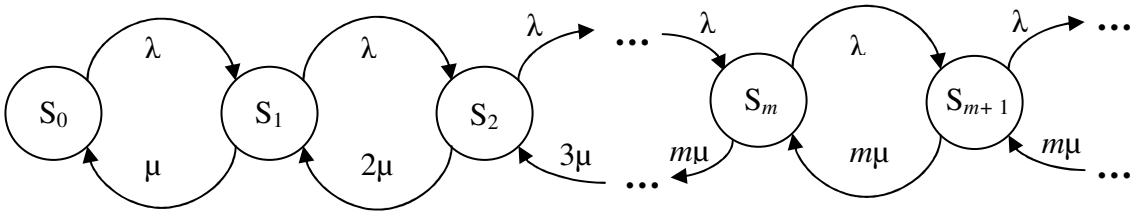


Рисунок 7.2 – Діаграма переходів m-серверної системи з чергою

Параметр потоку звільнення серверів системи μ розраховується за формулою (7.1). Інтенсивність потоку звільнення для станів системи $k = 0 \ldots m$ залежить від k і визначається аналогічно до її визначення на діаграмі переходів рис. 4.3 – вона є змінною та дорівнює $k\mu$. Для станів системи $k = m \ldots \infty$ ця інтенсивність незмінна і дорівнює $m\mu$, оскільки при цьому вже є черга і зміна її стану залежить від звільнення будь-якого зі всіх m серверів.

Тому визначення стаціонарних імовірностей станів системи P_k складається з двох частин:

- для станів системи $k = 0 \ldots m$ – кількість зайнятих серверів або вимог, що обслуговуються;
- для станів системи $k = m \ldots \infty$ – кількість зайнятих місць очікування черги, або вимог, що очікують.

Розподіл кількості зайнятих серверів визначається за алгоритмом формул (7.2) – (7.8) і для станів $k = 0 \ldots m$ маємо:

$$P_k = \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} P_0 \qquad (k = 1, 2, \ldots m). \tag{7.12}$$

Для станів $k = m \ldots \infty$ рівняння балансу системи подібне до (7.6), але за незмінної інтенсивності потоку звільнення серверів, що дорівнює $m\mu$, маємо:

$$(\lambda + m\mu) P_k = \lambda P_{k-1} + m\mu P_{k+1}. \tag{7.13}$$

Звідси, якщо для станів $k = 0 \ldots m$ аналогічно до (7.7)

$$P_k = \frac{1}{k} \frac{\lambda}{\mu} P_{k-1}, \tag{7.14}$$

то для станів $k = m \ldots \infty$

$$P_k = \frac{1}{m}\frac{\lambda}{\mu} P_{k-1}. \qquad (7.15)$$

Результат розв'язання рівняння (7.15) очевидний. Шляхом послідовної підстановки в (7.15) значень $k = m + 1$; $k = m + 2$; $k = m + 3$ і т.д. перевіряємо:

$$P_{k=m+1} = \frac{\lambda}{m\mu} P_m = \left(\frac{\lambda}{m\mu}\right)^{k-m} P_m;$$

$$P_{k=m+2} = \frac{\lambda}{m\mu} P_{m+1} = \frac{\lambda}{m\mu}\frac{\lambda}{m\mu} P_m = \left(\frac{\lambda}{m\mu}\right)^{k-m} P_m;$$

$$P_{k=m+3} = \frac{\lambda}{m\mu} P_{m+2} = \frac{\lambda}{m\mu}\frac{\lambda}{m\mu} P_{m+1} = \frac{\lambda}{m\mu}\frac{\lambda}{m\mu}\frac{\lambda}{m\mu} P_m = \left(\frac{\lambda}{m\mu}\right)^{k-m} P_m,$$

а це (індукція) підтверджує, що для будь-якого $k = m \ldots \infty$ правильне

$$P_k = \left(\frac{\lambda}{m\mu}\right)^{k-m} P_m. \qquad (7.16)$$

Оскільки з (7.12) для P_m маємо:

$$P_m = \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} P_0, \qquad (7.17)$$

то у підсумку з (7.16) і (7.17) для станів $k = m \ldots \infty$ виходить, що

$$P_k = \left(\frac{\lambda}{m\mu}\right)^{k-m}\left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} P_0, \quad (k = m \ldots \infty). \qquad (7.18)$$

Рівняння (7.12) і (7.18) остаточно визначають стаціонарний розподіл імовірностей станів системи P_k для всіх можливих значень k або станів серверів та місць очікування черги. Імовірність P_0 для цього випадку визначається з умови нормування $\sum_{k=0}^{\infty} P_k = 1$, звідки підстановкою всіх значень P_k формул (7.12) та (7.18) дістаємо (доданки, що не залежать від k винесено за знак другої суми):

$$P_0\left[\sum_{k=0}^{m-1}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}\sum_{k=m}^{\infty}\left(\frac{\lambda}{m\mu}\right)^{k-m}\right] = 1. \qquad (7.19)$$

У цій формулі друга сума від $k = m$ до ∞ існує тільки за умови $(\lambda / m\mu) < 1$ (збіжність ряду за ознакою Даламбера) і тому вона може бути розрахованою за відомою формулою суми членів нескінченно спадної геометричної прогресії:

$$\sum_{k=m}^{\infty}\left(\frac{\lambda}{m\mu}\right)^{k-m} = \sum_{k=0}^{\infty}\left(\frac{\lambda}{m\mu}\right)^{k} = \frac{1}{1-\frac{\lambda}{m\mu}} = \frac{m\mu}{m\mu-\lambda}. \qquad (7.20)$$

Підставивши цей вираз в (7.19) отримуємо:

$$P_0 = \left[\sum_{k=0}^{m-1}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} + \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!}\frac{m\mu}{m\mu-\lambda}\right]^{-1}. \qquad (7.21)$$

Використана в (7.19) умова збіжності ряду $(\lambda / m\mu) < 1$ співпадає з (7.10), яка є умовою ергодичності процесу. Це в свою чергу не дозволяє нескінченно зростати черзі, що є запорукою нормального функціонування всієї системи.

Оскільки $\Lambda = \lambda / \mu$ (тому з (7.10) обов'язково $\Lambda < m$), то рівняння (7.12), (7.18) і (7.21), що остаточно визначають стаціонарний розподіл імовірностей станів P_k, можна записати в одну систему рівнянь так:

$$\begin{cases} P_k = \dfrac{\Lambda^k}{k!} P_0 & \text{їðè} \quad k = 1 \ldots m \qquad (7.22) \\ P_k = \left(\dfrac{\Lambda}{m}\right)^{k-m} \dfrac{\Lambda^m}{m!} P_0 & \text{їðè} \quad k = m \ldots \infty \qquad (7.23) \\ P_0 = \left[\sum\limits_{k=0}^{m-1} \dfrac{\Lambda^k}{k!} + \dfrac{\Lambda^m}{m!} \dfrac{m}{m-\Lambda}\right]^{-1} & \qquad (7.24) \end{cases}$$

Це є *другий розподіл Ерланга*, за яким можна розрахувати імовірність станів (зайнятості серверів та місць очікування в черзі) повнодоступної системи з необмеженою чергою при обслуговуванні пуассонівського потоку вимог.

Основною характеристикою якості обслуговування цієї системи є імовірність очікування $P_{w>0}$ ($W > 0$ означає, що час очікування є), яка може бути визначена з функції розподілу станів системи P_k. В умовах пуассонівського потоку $P_{w>0}$ дорівнює імовірності зайнятості всіх m серверів системи з урахуванням усіх можливих станів черги (m серверів зайнято і 0 вимог у черзі, m серверів зайнято і 1 вимога в черзі, 2 вимоги в черзі і т.д., а тому не плутати зі станом P_m):

$$P_{w>0} = \sum_{k=m}^{\infty} P_k \,, \qquad (7.25)$$

де k – стан системи ($0 < k \le m$ – сервери, $m < k \le \infty$ – черга).

Підставлянням (7.23) в (7.25) з урахуванням властивості (7.20) отримаємо:

$$P_{w>0} = \sum_{k=m}^{\infty} P_k = \frac{\Lambda^m}{m!} P_0 \sum_{k=m}^{\infty} \left(\frac{\Lambda}{m}\right)^{k-m} = \frac{\Lambda^m}{m!} \frac{m}{m-\Lambda} P_0 \,. \qquad (7.26)$$

Підставляння (7.24) в (7.26) дає

$$P_{w>0} = \frac{\dfrac{\Lambda^m}{m!} \dfrac{m}{m-\Lambda}}{\sum\limits_{k=0}^{m-1} \dfrac{\Lambda^k}{k!} + \dfrac{\Lambda^m}{m!} \dfrac{m}{m-\Lambda}} \,. \qquad (7.27)$$

Формула (7.27) називається *С-формулою Ерланга* і позначається $C_m(\Lambda)$. Для системи $M/M/m/\infty$ за нею можна розрахувати імовірність очікування $P_{w>0}$ (6.2). Дану формулу можна переписати так:

$$P_{w>0} = C_m(\Lambda) = \frac{\frac{\Lambda^m}{m!}\frac{m}{m-\Lambda}}{\sum_{k=0}^{m}\frac{\Lambda^k}{k!} - \frac{\Lambda^m}{m!} + \frac{\Lambda^m}{m!}\frac{m}{m-\Lambda}} = \frac{\frac{\Lambda^m}{m!}\frac{m}{m-\Lambda}}{\sum_{k=0}^{m}\frac{\Lambda^k}{k!} + \frac{\Lambda^m}{m!}\left(\frac{m}{m-\Lambda}-1\right)}.$$

Поділивши чисельник і знаменник на $\sum_{k=0}^{m}\frac{\Lambda^k}{k!}$ (див. 7.11) отримуємо:

$$C_m(\Lambda) = \frac{E_m(\Lambda)\frac{m}{m-\Lambda}}{1+E_m(\Lambda)\left(\frac{m}{m-\Lambda}-1\right)} = \frac{E_m(\Lambda)m}{m-\Lambda[1-E_m(\Lambda)]}.$$

Дана формула встановлює взаємозв'язок між *B*– та *C*-формулами Ерланга, а це дозволяє розраховувати значення імовірності $C_m(\Lambda)$ за значеннями $E_m(\Lambda)$ формули (7.11), які наведено у відповідних таблицях. Очевидно, що за однакових умов завжди імовірність $C_m(\Lambda) > E_m(\Lambda)$.

Якщо позначимо

$$\rho = \frac{\Lambda}{m}, \tag{7.28}$$

де ρ – *інтенсивність питомого навантаження* (навантаження на один сервер), то цю формулу можна записати так:

$$C_m(\Lambda) = \frac{1}{\frac{1-\rho}{E_m(\Lambda)}+\rho}. \tag{7.29}$$

Середня довжина черги Q (середня кількість вимог у черзі) може бути знайдена з розподілу ймовірностей станів системи P_k при $k = m+1, m+2, \ldots, \infty$ (в черзі 1, 2 і т.д. вимог) за правилом обчислення математичного сподівання:

$$Q = \sum_{k=m+1}^{\infty}(k-m)P_k .$$

Підставивши з (7.23) імовірності станів черги P_k маємо (доданки, що не залежать від k винесено за знак суми):

$$Q = \frac{\Lambda^m}{m!}P_0\sum_{k=m+1}^{\infty}(k-m)\left(\frac{\Lambda}{m}\right)^{k-m} .$$

Відомо, що $\sum_{i=0}^{\infty} ix^i = \frac{x}{(1-x)^2}$, а тому

$$Q = \frac{\Lambda^m}{m!} P_0 \frac{\frac{\Lambda}{m}}{\left(1-\frac{\Lambda}{m}\right)^2}. \tag{7.30}$$

З (7.17) і (7.26) випливає, що імовірність стану P_m визначиться як

$$P_m = \left(1-\frac{\Lambda}{m}\right) P_{w>0}. \tag{7.31}$$

Підставивши до (7.30) замість імовірності $P_0 = P_m \left(\frac{\Lambda^m}{m!}\right)^{-1}$, яку отримано з формули (7.17), а замість P_m її значення з (7.31), після скорочень маємо:

$$Q = \frac{\Lambda}{m-\Lambda} P_{w>0}. \tag{7.32}$$

Середню тривалість очікування для будь-якої вимоги W отримаємо із формули Літтла – за W одиниць часу очікування до черги надійде в середньому $Q = \Lambda W$ вимог, а отже з урахуванням (7.32)

$$W = \frac{Q}{\Lambda} = \frac{1}{m-\Lambda} P_{w>0}. \tag{7.33}$$

Середню тривалість очікування для затриманих вимог t_q отримаємо із тлумачення величини Q, як інтенсивності навантаження, яке створюється вимогами, що очікують в черзі. Величина $\Lambda P_{w>0}$ – це інтенсивність потоку вимог з черги, де кожна вимога в середньому очікує час t_q. Отже, $Q = \Lambda P_{w>0}\, t_q$, звідки з урахуванням (7.32)

$$t_q = \frac{Q}{\Lambda P_{w>0}} = \frac{1}{m-\Lambda}. \tag{7.34}$$

Значення W та t_q подано в одиницях середньої тривалості обслуговування. Такий нормований вид показує, в скільки разів ці значення зростають або убувають порівняно із середнім часом обслуговування $\bar{x}$. Із (7.33) та (7.34) виплаває очевидне співвідношення – $W \le t_q$.

Таким чином всі основні характеристики *QoS* знайдено. Середня кількість вимог у системі N (сумарно тих, що обслуговуються та тих, що очікують обслуговування) визначається за формулою (6.3), а середній час перебування вимоги в системі T (складається з середнього часу обслуговування та середнього часу очікування) – за формулою (6.4).

При проектуванні систем розподілу інформації часто необхідно знати окремі значення функції розподілу часу очікування вимог у системі. Функція розподілу часу очікування в системі $P_{w>t}$ – це імовірність того, що вимога, яка

надходить до системи в будь-якому з її станів буде очікувати час понад t. Для розрахунку цієї імовірності використаємо умовну імовірність $P_{k(w>t)}$, яка тлумачиться як і попередня, але за умови надходження вимоги в стані системи $k = m \ldots \infty$. Оскільки очікування можливе тільки у випадку, коли зайняті всі m серверів системи, то за формулою повної імовірності отримаємо:

$$P_{w>t} = \sum_{k=m}^{\infty} P_k P_{k(w>t)} . \qquad (7.35)$$

Вимога, яка очікує в черзі останньою, почне обслуговуватись тільки за умови, якщо за час її очікування t закінчиться обслуговування $k - m + 1$ вимог ($k - m$ є довжиною черги). Таким чином, час очікування буде більше t в тому випадку, коли за цей час з черги піде не більше $k - m$ попередніх вимог. Отже, імовірність того, що вимога, яка надходить до системи в стані k буде очікувати час понад t дорівнює імовірності того, що за цей же час звільниться не більше $i = k - m$ серверів.

За експонентного розподілу тривалості обслуговування потік звільнень серверів системи є пуассонівським з параметром $m\mu$ (рис. 7.2). Імовірність того, що за час t звільниться точно i серверів розраховується за формулою Пуассона (4.5), а імовірність того, що звільниться не більше $k - m$ серверів дорівнює сумі:

$$P_{k(w>t)} = \sum_{j=0}^{k-m} \frac{(m\mu t)^j}{j!} e^{-m\mu t} .$$

Підставивши цей вираз та (7.23) в формулу (7.35) і послідовно скорочуючи [2, с. 110] отримаємо:

$$P_{w>t} = \sum_{k=m}^{\infty} P_k \sum_{j=0}^{k-m} \frac{(m\mu t)^j}{j!} e^{-m\mu t} = P_{w>0} e^{-(m\mu-\Lambda)t} .$$

З цього за правилом визначання математичного сподівання розрахуємо середню тривалість очікування в системі W:

$$W = \int_0^{\infty} P_{w>t} dt = P_{w>0} \int_0^{\infty} e^{-(m\mu-\Lambda)t} = \frac{P_{w>0}}{m\mu - \Lambda},$$

що за умови $\mu = 1$ (означає, що W вимірюється в одиницях середньої тривалості обслуговування $\bar{x} = 1/\mu$) дає аналогічний до (7.33) результат.

У системах розподілу інформації не завжди вимоги з черги надходять на обслуговування в порядку надходження. Є чимало випадків, коли вибір з черги випадковий. Хоча дисципліна вибору з черги не впливає на середній час очікування W, але випадковий вибір з черги збільшує дисперсію часу очікування порівняно з вибором у порядку надходження.

7.3 Система з обмеженою чергою *M*/*M*/*m*/*r*

Повнодоступна схема СРІ з m серверами та обмеженою кількістю місць очікування в черзі r обслуговує пуассонівський потік вимог з дисципліною обслуговування черги *FIFO*. Інтервал часу між вимогами та тривалість їх обслуговування мають експонентні закони розподілу. Інтенсивність вхідного навантаження Λ. Необхідно визначити стаціонарний розподіл імовірностей станів системи P_k ($k = 0 \ldots m + r$) та характеристики *QoS*.

У разі обмеженої черги, де кількість місць очікування $r < \infty$, отримуємо комбіновану систему *M*/*M*/*m*/*r* – з чергами та втратами. Якщо в момент надходження вимоги в системі є хоча б один вільний сервер, то вона негайно починає обслуговуватись. Якщо ж всі сервери системи зайняті, то вимога стає в чергу за умови, що в ній очікує менше, як r вимог (є місця для очікування). Якщо ж в черзі є вже r вимог, що надійшли раніше, то поточна вимога втрачається. Таким чином, вимозі відмовляється в обслуговуванні у тому разі, коли в системі знаходяться $l = m + r$ вимог. З цих l вимог m обслуговуються, а r очікують своєї черги.

Очевидно, якщо $r = 0$, то система описується моделлю *M*/*M*/*m* і задача вирішується за допомогою першого розподілу Ерланга (7.11). Якщо $r = \infty$, то система описується моделлю *M*/*M*/*m*/∞ і задача зводиться до попередньої.

Для комбінованої системи *M*/*M*/*m*/*r*, яка частково є аналогічною до системи *M*/*M*/*m*/∞, рівняння (7.22) і (7.23) визначають стаціонарний розподіл імовірностей станів системи P_k для $k = 1 \ldots m$ та $k = m \ldots l$ відповідно. Тому імовірність P_0 визначається з умови нормування $\sum_{k=0}^{l} P_k = 1$, звідки підстановкою всіх значень P_k формул (7.22) та (7.23) дістаємо:

$$P_0 \left[\sum_{k=0}^{m-1} \frac{\Lambda^k}{k!} + \frac{\Lambda^m}{m!} \sum_{k=m}^{l} \left(\frac{\Lambda}{m} \right)^{k-m} \right] = 1. \qquad (7.36)$$

Оскільки сума від $k = m$ до l може бути розрахованою за формулою суми $r = l - m$ членів геометричної прогресії зі знаменником Λ / m, то

$$P_0 = \left[\sum_{k=0}^{m-1} \frac{\Lambda^k}{k!} + \left(\frac{\Lambda^m}{m!} \right) \frac{1 - \left(\frac{\Lambda}{m} \right)^{r+1}}{1 - \frac{\Lambda}{m}} \right]^{-1}. \qquad (7.37)$$

Імовірність повної зайнятості системи $P_{\text{ЗН}}$ (очікування або втрати вимоги) у цьому випадку визначається аналогічно (7.26), але сума має бути не до ∞, а тільки до l. Замінивши цю суму подібно, як у формулах (7.36) і (7.37), маємо:

$$P_{\text{çí}} = \sum_{k=m}^{l} P_k = \frac{\Lambda^m}{m!} P_0 \sum_{k=m}^{l} \left(\frac{\Lambda}{m} \right)^{k-m} = \frac{\Lambda^m}{m!} P_0 \frac{1 - \left(\frac{\Lambda}{m} \right)^{r+1}}{1 - \frac{\Lambda}{m}}. \qquad (7.38)$$

Підставляння P_0 з (7.37) в (7.38) дає

$$P_{\text{çí}} = \frac{\left(\frac{\Lambda^m}{m!}\right)\frac{1-\left(\frac{\Lambda}{m}\right)^{r+1}}{1-\frac{\Lambda}{m}}}{\sum_{k=0}^{m-1}\frac{\Lambda^k}{k!}+\left(\frac{\Lambda^m}{m!}\right)\frac{1-\left(\frac{\Lambda}{m}\right)^{r+1}}{1-\frac{\Lambda}{m}}}.$$

Враховуючи, що $\frac{1}{1-\frac{\Lambda}{m}} = \frac{m}{m-\Lambda}$, дану формулу можна переписати так:

$$P_{\text{çí}} = \frac{\frac{\Lambda^m}{m!}\frac{m}{m-\Lambda}\left[1-\left(\frac{\Lambda}{m}\right)^{r+1}\right]}{\sum_{k=0}^{m-1}\frac{\Lambda^k}{k!}+\frac{\Lambda^m}{m!}\frac{m}{m-\Lambda}\left[1-\left(\frac{\Lambda}{m}\right)^{r+1}\right]}. \qquad (7.39)$$

Очевидно, якщо $r = \infty$, то формула (7.37) співпадає з (7.24), а формула (7.39) – з *C*-формулою Ерланга (7.27). Якщо $r = 0$, то (7.37) співпадає з (7.9), а формула (7.39) – з *B*-формулою Ерланга (7.11). З цього випливає, що при зміні довжини черги від ∞ до 0 величина імовірності $P_{\text{ЗН}}$ моделі *M*/*M*/*m*/*r* знаходиться в діапазоні значень, отриманих за формулами (7.27) та (7.11) для моделей *M*/*M*/*m*/∞ та *M*/*M*/*m*/0 відповідно. Таким чином, (7.39) є узагальнюючим розв'язком для повнодоступних систем з втратами, з необмеженою та обмеженою чергами за умови обслуговування пуассонівського потоку вимог та експонентної тривалості їх обслуговування.

Формулу (7.39) можна спростити у такий же спосіб, як це отримано в (7.29) з урахуванням (7.28)

$$P_{\text{çí}} = \frac{1-\rho^{r+1}}{\frac{1-\rho}{E_m(\Lambda)}+\rho-\rho^{r+1}}. \qquad (7.40)$$

У системі з обмеженою до *r* чергою подія втрати вимоги відбудеться у тому випадку, коли при надходженні вимоги у черзі вже є *r* вимог, що надійшли раніше. Отже імовірність втрати вимоги P_B дорівнює імовірності того, що в системі вже є $l = m + r$ вимог (отже $r = l - m$) або система в стані P_l. Для пуассонівського потоку імовірність $P_B = P_l$ за будь-якого закону тривалості обслуговування.

Тому відповідно до (7.23) та (7.37) отримаємо:

$$P_B = P_l = \left(\frac{\Lambda}{m}\right)^{l-m} \frac{\Lambda^m}{m!} P_0 = \frac{\left(\frac{\Lambda}{m}\right)^r \frac{\Lambda^m}{m!}}{\sum_{k=0}^{m-1} \frac{\Lambda^k}{k!} + \left(\frac{\Lambda^m}{m!}\right) \frac{1-\left(\frac{\Lambda}{m}\right)^{r+1}}{1-\frac{\Lambda}{m}}}. \qquad (7.41)$$

Імовірність очікування за обмеженої до r черги $P_{w>0}(r)$ або частку вимог, які очікують, розрахуємо так: $P_{w>0}(r) = P_{\text{ЗН}} - P_B$. Підставивши сюди (7.39) і (7.41) відповідно маємо

$$P_{w>0}(r) = \frac{\frac{\Lambda^m}{m!} \frac{m}{m-\Lambda}\left[1-\left(\frac{\Lambda}{m}\right)^r\right]}{\sum_{k=0}^{m-1} \frac{\Lambda^k}{k!} + \frac{\Lambda^m}{m!} \frac{m}{m-\Lambda}\left[1-\left(\frac{\Lambda}{m}\right)^{r+1}\right]}. \qquad (7.42)$$

Цю імовірність можна розрахувати аналогічно до (7.38), але з верхньою межею підсумку $l - 1$, що дає аналогічний результат:

$$P_{w>0}(r) = \sum_{k=m}^{l-1} P_k = \frac{\Lambda^m}{m!} P_0 \sum_{k=m}^{l-1} \left(\frac{\Lambda}{m}\right)^{k-m} = \frac{\Lambda^m}{m!} \frac{m}{m-\Lambda}\left[1-\left(\frac{\Lambda}{m}\right)^r\right] P_0.$$

Формулу (7.41) з урахуванням (7.38) можна записати так:

$$P_B = \left(\frac{\Lambda}{m}\right)^{l-m} \frac{\Lambda^m}{m!} P_0 = P_{\text{çí}} \frac{\left(\frac{\Lambda}{m}\right)^r \left(1-\frac{\Lambda}{m}\right)}{1-\left(\frac{\Lambda}{m}\right)^{r+1}}.$$

Враховуючи (7.28), дану формулу можна спростити:

$$P_B = P_{\text{çí}} \frac{\rho^r (1-\rho)}{1-\rho^{r+1}}. \qquad (7.43)$$

Підставивши до (7.43) формулу (7.40) маємо

$$P_B = \frac{\rho^r}{\frac{1}{E_m(\Lambda)} + \frac{\rho - \rho^{r+1}}{1-\rho}}.$$

Формулу (7.42) можна переписати не як різницю (7.39) і (7.41), а як різницю (7.40) та (7.43), що дає

$$P_{w>0}(r) = \frac{1-\rho^r}{\dfrac{1-\rho}{E_m(\Lambda)} + \rho - \rho^{r+1}}.$$

Із (7.43) і урахуванням (7.16), де $k = l$ та $r = l - m$, випливає, що

$$P_{\text{çí}} = P_l \frac{1-\rho^{r+1}}{\rho^r(1-\rho)} = P_m \frac{1-\rho^{r+1}}{1-\rho}.$$

Для односерверної системи $M/M/1/r$ із (7.37), (7.39), (7.42) та (7.43) можна розрахувати:

– імовірність P_0 (сервер і місця в черзі вільні) із (7.37) –

$$P_0 = \left[1 + \rho\frac{1-\rho^{r+1}}{1-\rho}\right]^{-1} = \frac{1-\rho}{1-\rho^{r+2}};$$

– імовірність $P_{\text{зн}}$ із (7.39) або $P_{\text{зн}} = 1 - P_0$, що дає

$$P_{\text{çí}} = \frac{\rho(1-\rho^{r+1})}{1-\rho^{r+2}};$$

– імовірність $P_{w>0}(r)$ з (7.42) та підстановкою (7.28) дає

$$P_{w>0}(r) = \frac{\rho(1-\rho^r)}{1-\rho^{r+2}};$$

– імовірність P_B із (7.43) та підстановкою вище знайденого $P_{\text{зн}}$ дає

$$P_B = \frac{\rho^{r+1}(1-\rho)}{1-\rho^{r+2}}.$$

Із (7.18) випливає, що в системі $M/M/1/r$ стани системи мають розподіл

$$P_k = \rho^k P_0 = \frac{\rho^k(1-\rho)}{1-\rho^{r+2}}.$$

В системі $M/M/1/\infty$ з цієї формули маємо геометричний розподіл станів системи

$$P_k = \rho^k(1-\rho).$$

Отже характеристики якості обслуговування системи $M/M/m/r$ знайдено.

7.4 Система з необмеженою чергою *M*/*D*/*m*/∞

У телекомунікаційних системах, що базуються на технології комутації пакетів, застосовується дисципліна обслуговування з чергами. Є системи, в яких час обслуговування постійний, наприклад, процес обслуговування вимог керуючими пристроями вузлів комутації або обслуговування пакетів однакової довжини. В моделі *M*/*M*/*m*/∞ з експонентним часом обслуговування вимог (пакетів) імовірність очікування розраховується за допомогою *C*-формули Ерланга (7.27). Для моделі *M*/*D*/*m*/∞ за дисципліни черги *FIFO* розв'язання отримано К.Д. Кроммеліном [3], в якому використано рівняння станів Фрая. Система таких рівнянь розв'язується методом похідних функцій. Ці обчислення настільки складні, що на практиці для інженерних розрахунків замість точних аналітичних формул застосовуються відповідні діаграми (криві Кроммеліна).

У [8] запропонований більш простий метод розрахунку моделі *M*/*D*/*m*/∞, в якому використовується *C*-формула Ерланга.

Необхідно визначити основні характеристики якості обслуговування *QoS*:

– середню тривалість очікування вимог у системі *W*;
– середню довжину черги *Q*;
– середню кількість вимог у системі *N*;
– імовірність очікування $P_{w>0}$;
– середню тривалість очікування вимог у черзі t_q.

Основні характеристики якості обслуговування моделі *M*/*M*/*m*/∞ перебувають між собою в певній функціональній залежності. Середня тривалість очікування для будь-якої вимоги *W* (що очікує і що не очікує) є середнім значенням часу очікування, віднесеним до всіх вимог. Якщо ж відома середня тривалість очікування тільки затриманих вимог t_q, то для визначення *W* необхідно помножити цю тривалість на імовірність $P_{w>0}$, яка показує середню частку затриманих вимог. Отже з (7.34) та (7.33) отримуємо

$$W = t_q P_{w>0}. \tag{7.44}$$

За формулою Літтла за *W* одиниць часу очікування до черги надійде ΛW вимог, отже

$$Q = \Lambda W. \tag{7.45}$$

З наведених співвідношень характеристик *QoS* випливає, що для аналізу моделі *M*/*M*/*m*/∞ з експонентним розподілом тривалості обслуговування необхідно лише розрахувати середню тривалість очікування вимог у черзі t_q (7.34) та імовірність очікування $P_{w>0}$ (7.27). При цьому середня тривалість очікування вимог $W_{(M)}$ (*M* – експонента) визначається з (7.33).

Для розрахунку $P_{w>0}$ в моделі *M*/*D*/*m*/∞ з постійною тривалістю обслуговування викладений в п. 7.2 алгоритм визначення станів статистичної рівноваги системи застосовувати неможливо. Тепер імовірність закінчення обслуговування вимоги, яка в разі експонентного розподілу $\bar{x}$ пропорційна $m\mu$, визначається не кількістю зайнятих серверів *m* (рис. 7.2), а моментами початку обслуговування, і тому буде залежати від розміщення досліджуваного інтервалу на осі часу – процес не буде ергодичним (див. п. 7.1).

Коли точне теоретичне розв'язання виявляється складним, то нерідко досить отримати наближений вираз, на якому може ґрунтуватися весь метод розрахунку. Визнаним методом аналітичних наближень є імітаційне моделювання, яке дозволяє ефективно перевіряти якість використовуваних припущень. За результатами імітаційного моделювання з використанням алгоритму [9] встановлено, що час очікування W за постійної (D) та експонентної (M) тривалості обслуговування перебуває в такому співвідношенні:

$$W_{(D)} = W_{(M)} 2\left(\frac{m}{m+\Lambda}\right)^2 = C_m(\Lambda)\frac{1}{m-\Lambda}2\left(\frac{m}{m+\Lambda}\right)^2. \qquad (7.46)$$

Формула (7.45) визначає, що для багатоканальної системи при $m = \Lambda$ середня тривалість очікування за постійної та експонентної тривалості обслуговування відрізняються в 2 рази, що відповідає такому самому співвідношенню, встановленому формулою Поллачека-Хінчина для односерверної системи. Зі зростанням ємності системи m це співвідношення спадає і імітаційне моделювання свідчить, що в широкому діапазоні зміни m та Λ точність оцінки середньої тривалості очікування (7.46) не гірше ±10 %.

Імовірність очікування $P_{w>0}$ визначається з функції розподілу станів системи P_k і тут $P_{w>0}$ дорівнює імовірності того, що при надходження вимоги вона застає всі m серверів зайнятими (7.25). Це можна записати так:

$$P_{w>0} = \sum_{k=m}^{\infty} P_k = 1 - \sum_{k=0}^{m-1} P_k . \qquad (7.47)$$

В умовах необмеженої кількості серверів ($m = \infty$) вимоги обслуговуються без втрат. За постійної тривалості обслуговування, коли немає втрат, властивості потоку звільнень збігаються з властивостями потоку надходження вимог, тому що відбувається тільки зсув у часі на величину $\bar{x}$ між моментом надходження вимоги й моментом закінчення її обслуговування. При цьому стан системи повністю визначається властивостями потоку вимог, а функції розподілу кількості вимог у системі P_k і кількості вимог P_i, що надійшли за час $\bar{x}$, цілком збігаються з розподілом Пуассона

$$P_k = P_i = \frac{\Lambda^i}{i!} \cdot e^{-\Lambda}. \qquad (7.48)$$

При кінцевому значенні m і необмеженій кількості місць очікування вимоги також обслуговуються без явних втрат. Однак у цьому випадку вимоги, що надходять після зайняття всіх серверів системи, попадають у чергу на очікування, і у випадку звільнення хоча б одного з m зайнятих серверів відразу подаються із черги на обслуговування. Тепер на канали системи надходять вимоги з первинного потоку з інтенсивністю Λ та із черги з інтенсивністю ΛW.

Таким чином, з урахуванням (7.45) сумарна інтенсивність навантаження на канали системи збільшується до величини

$$\Lambda_\Sigma = \Lambda + Q. \qquad (7.49)$$

Інтенсивність навантаження Λ_Σ дорівнює середній кількості вимог у системі N (6.3).

З наведених аргументів випливає простий метод розрахунку основних характеристик якості обслуговування системи $M/D/m/\infty$, який складається з шести кроків:

1. За (7.29) для заданих Λ й m розраховується $C_m(\Lambda)$ для моделі $M/M/m/\infty$.
2. За (7.46) для заданих Λ й m визначається $W_{(D)}$ (апроксимація).
3. За (7.45) для Λ і розрахованого значення $W_{(D)}$ визначається Q.
4. За (7.49) розраховується сумарна інтенсивність навантаження Λ_Σ .
5. За (7.47) і (7.48) для Λ_Σ розраховується імовірність очікування $P_{w>0}$.
6. За розрахованими $W_{(D)}$ й $P_{w>0}$ з формули (7.44) визначається t_q.

Слід зазначити, що при m близьких до Λ, тобто в діапазоні ємностей системи $\Lambda < m \le (\Lambda + \sqrt{\Lambda}/2)$, додатковий потік вимог із черги суттєво збільшує не тільки сумарну інтенсивність навантаження Λ_Σ , але і її сумарну дисперсію σ_Σ^2, що для пуассонівського потоку дорівнює $\sqrt{\Lambda}$. Тому в даному діапазоні ємностей системи для підвищення точності розрахунку $P_{w>0}$ та t_q рекомендується на кроці 5 замість розподілу Пуассона (7.48) використати розподіл нормального закону (Гаусса) з підставленням (7.49) Λ_Σ та $\sigma_\Sigma^2 = \sqrt{\Lambda} + Q/2$. Для виключення нескінченної черги обов'язково $m > \Lambda$.

Оцінку ступеня точності запропонованого методу розрахунку характеристик якості обслуговування моделі $M/D/m/\infty$ перевірено імітаційним моделюванням, результати якого [9] засвідчили, що в широкому діапазоні зміни m і Λ відносна помилка розрахунку всіх характеристик QoS не перевищує ±10 %.

Результатами імітаційного моделювання з такою ж точністю оцінки визначене й співвідношення імовірностей очікування при експонентній і постійній тривалості обслуговування в моделях $M/M/m/\infty$ і $M/D/m/\infty$:

$$P_{w>0} \approx C_m(\Lambda) \cdot \frac{1}{2 \cdot 2^0 \left(\frac{m}{m+\Lambda} \right)^1}, \qquad (7.50)$$

де $P_{w>0}$ – імовірність очікування при постійній, а $C_m(\Lambda)$ – при експонентній тривалості обслуговування.

З (7.44) випливає, що імовірність очікування за будь-яких умов – це співвідношення середніх тривалостей очікування в системі W та черзі t_q. Тому з (7.46) і (7.50) маємо середню тривалість очікування вимог у черзі $t_{q(D)}$ при постійній тривалості обслуговування:

$$t_{q(D)} = \frac{W_{(D)}}{P_{w>0}} = \frac{C_m(\Lambda) \cdot \frac{1}{m-\Lambda} \cdot 2^1 \left(\frac{m}{m+\Lambda} \right)^2}{C_m(\Lambda) \cdot 2^{-1} \left(\frac{m}{m+\Lambda} \right)^{-1}} = \frac{1}{m-\Lambda} 2^2 \left(\frac{m}{m+\Lambda} \right)^3. \qquad (7.51)$$

Аналізуючи формули (7.46), (7.50) і (7.51) помічаємо, що однойменні характеристики *QoS* в моделях *M*/*M*/*m*/∞ і *M*/*D*/*m*/∞ при експонентній і постійній тривалості обслуговування пов'язані між собою наступною апроксимуючою функцією:

$$F(k) = 2^{k-1}\left(\frac{m}{m+\Lambda}\right)^k, \tag{7.52}$$

де k = 1, 2 і 3 для характеристик $P_{w>0}$, $W_{(D)}$ та $t_{q(D)}$ відповідно.

Отже, шляхом застосування апроксимуючої функції (7.52) можуть бути розраховані основні характеристики якості обслуговування в моделі *M/D/m/*∞ через *C*-формулу Ерланга, справедливу для моделі *M/M/m/*∞:

$$P_{w>0} = \frac{C_m(\Lambda)}{2 \cdot F(k)},$$

$$W_{(D)} = \frac{C_m(\Lambda)}{m-\Lambda} \cdot F(k+1), \tag{7.53}$$

$$t_{q(D)} = \frac{1}{m-\Lambda} F\left(k+2\right)$$

де k = 1 для всіх наведених характеристик.

Імітаційним моделюванням встановлено, що при заміні в (7.53) коефіцієнта $k = 1$ на $k = 0{,}01\Lambda$ (може бути не цілим) точність оцінки основних характеристик якості обслуговування підвищується до ±5 %.

Кроммеліном розглянуто систему з чергою, в якій обслуговування вимог, що очікують, виконується у порядку надходження (впорядкована черга). Але є приклади систем з чергою, у яких обслуговування вимог, що очікують, виконується при випадковому виборі їх з черги.

Однолінійна система з постійною тривалістю обслуговування і випадковим вибором з черги вимог, що очікують, досліджена Берком [3]. Аналіз розподілу часу очікування *W* в однолінійній системі з постійною тривалістю обслуговування при випадковому виборі вимог з черги і виборі вимог у порядку надходження показує, що для невеликих значень тривалості очікування (від 0,1 до $3\,\bar{x}$) якість обслуговування вимог вище при випадковому виборі з черги. Для великих значень тривалості очікування ($> 3\,\bar{x}$) якість обслуговування вимог при випадковому виборі з черги істотно гірша, ніж при обслуговуванні викликів у порядку черги.

Слід нагадати висновок, зроблений в кінці підрозділу 7.2, що дисципліна вибору з черги (у порядку надходження, випадковому порядку або будь-якій іншій дисципліні) не впливає на середній час очікування вимоги початку обслуговування *W*, але впливає на дисперсію часу очікування.

7.5 Система з необмеженою чергою *M*/*G*/1/∞

Односерверна система з необмеженою кількістю місць очікування (рис. 7.3) обслуговує пуассонівський потік вимог. Тривалість обслуговування вимог має довільний (будь-який) закон розподілу $B(x)$ з математичним сподіванням $\bar{x}$. Інтенсивність потоку вимог λ. Визначити характеристики *QoS*.

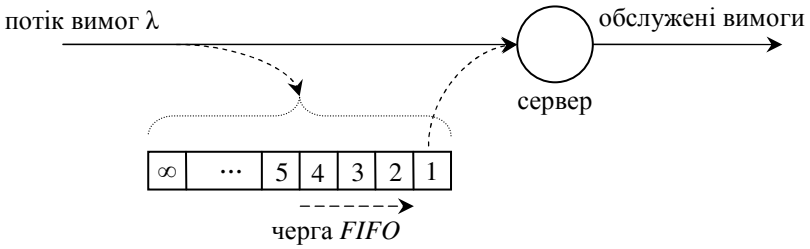


Рисунок 7.3 – Модель односерверної системи з чергою

У разі не експонентного розподілу тривалості обслуговування є тільки один експонентний розподіл моделі – інтервал часу між вимогами в потоці вимог. Тому дана задача розв'язується методом вкладених ланцюгів Маркова.

Поняття вкладеного ланцюга Маркова зводиться до наступного. Нехай (g_t) – випадковий процес з неперервним часом, що набуває тільки цілих числових значень з деякої безлічі G, та існує послідовність t_1, t_2, ... (теж випадкова) моментів часу така, що процес (ξ_n)

$$\xi_n = g_{t_n}, \qquad n = 1, 2, \ldots \tag{7.54}$$

є однорідним марковським ланцюгом з дискретним часом, тобто для $n = 2, 3, \ldots$ має місце:

$$\Psi(\xi_n = j \mid \xi_{n-1} = i, \xi_{n-2} = i_{n-2}, \ldots \xi_1 = i_1) =$$
$$= \Psi(\xi_n = j \mid \xi_{n-1} = i) = \Psi(\xi_2 = j \mid \xi_1 = i), \tag{7.55}$$

де $j, i, i_1 \ldots \in G$.

Процес (ξ_n) є *марковським ланцюгом, вкладеним в* (g_t). Основний зміст методу вкладених ланцюгів Маркова полягає в тому, що для даного випадкового процесу, що описує поведінку розглянутої моделі, конструюється зручний марковський ланцюг, досліджуваний аналітичними методами, звичайними для ланцюгів Маркова, і отримані результати (наприклад, стаціонарні імовірності станів) перераховуються для величин вихідної системи.

Нехай досліджувана модель приймає стани 0, 1, 2, ..., а g_t – кількість вимог, що перебувають в системі в момент часу t або g_t є стан системи в момент часу t; (g_t) – не марковський ланцюг. Визначимо характеристики *QoS* – середні значення кількості вимог і часу очікування в системі.

Метод вкладених ланцюгів складається з наступних чотирьох кроків:

К р о к 1. Визначимо послідовності моментів часу (t_n), $0 < t_1 < t_2 < \ldots < \infty$, що дозволяє конструювання вкладеного ланцюга Маркова. Зрозуміло, якщо всі розподіли експонентні, то (g_t) – однорідний марковський ланцюг і в якості (t_n)

можна вибрати будь-яку зростаючу послідовність моментів часу $0 \le t_n < \infty$. Якщо не всі розподіли, що характеризують поведінку елементів системи, експонентні, то (g_t) – не марковський ланцюг, і при цьому коли тільки один розподіл не експонентний, то стан системи в момент часу t можна описати за допомогою величини ξ^*_t:

$$\xi^*_t = (g_t, r_t), \tag{7.56}$$

де ξ^*_t – перетворення Лапласа-Стілтьєса неперервної функції розподілу ξ_t, за допомогою якого шляхом диференціювання можна визначити відповідні *моменти* цього розподілу.

Для додаткової змінної r_t покладається таке: r_t дорівнює залишковому часові неекспонентно розподіленої величини в момент часу t, якщо такий існує (наприклад залишковий час обслуговування, залишковий час паузи); $r_t = 0$, якщо відсутній такий залишковий час (наприклад, не обслуговується жодна вимога; немає паузи з неекспонентним розподілом).

Значення t_n знаходиться в такий спосіб. Залишковим часом r_t є залишкова тривалість обслуговування. Вона буде дорівнювати нулеві, якщо деяка вимога залишає систему. Тому покладається, щоб t_n дорівнювало моментові часу, коли n-а вимога полишає систему (момент виходу).

Сконструйований за допомогою додаткової змінної r_t двовимірний стохастичний процес (ξ^*_t) розглядається тільки в певних точках t_n, тобто в моменти виходу вимог із системи, а тому він є однорідним марковським і розподіл цього процесу можна визначити.

К р о к 2 . Обчислимо перехідні імовірності для вкладеного ланцюга Маркова. Нехай ξ_n – кількість вимог, що знаходяться в системі безпосередньо після моменту t_n, тобто $\xi_n = g_{t_n+0}$. При цьому (ξ_n) – однорідний марковський ланцюг. Нехай $p_{i,j}$ – незалежна від n імовірність переходу системи в момент виходу із неї вимоги:

$$p_{i,j} = \Psi(\xi_{n+1} = j \mid \xi_n = i), \qquad i, j = 0, 1, 2, \ldots$$

Позначимо через k_j імовірність того, що за час обслуговування деякої вимоги в систему надходить j нових вимог; тоді

$$p_{i,j} = \begin{cases} k_j & \text{ïðè} \quad j = 0,1,2,\ldots, \quad i = 0; \\ k_{j-i+1} & \text{ïðè} \quad j > i-1, \qquad i = 1,2,\ldots: \\ 0 & \text{â ³íøèõ âèïàäêàõ.} \end{cases} \tag{7.57}$$

Оскільки вимоги надходять за пуассонівським законом з інтенсивністю λ, то для них імовірність того, що на інтервалі часу тривалістю t надійде точно j вимог, дорівнює (4.5), і тому

$$k_j = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} dB(t), \quad j = 0,1, \ldots \tag{7.58}$$

Твірна (*рос.* – производящая) функція $K^*(z)$ розподілу (k_j) за визначенням:

$$K^*(z) = \sum_{j=0}^{\infty} k_j z^j . \qquad (7.59)$$

де z – комплексна змінна. Із (7.59) з урахуванням (7.58) і згідно з [5, с.205]

$$K^*(z) = B^*(\lambda - \lambda z) . \qquad (7.60)$$

Оскільки $B(x)$ є функцією розподілу тривалості обслуговування, то отримане з (7.58) рівняння встановлює взаємозв'язок між твірною функцією K^* дискретного розподілу випадкової величини k_j (кількості вимог за час обслуговування) і перетворенням Лапласа-Стілтьєса B^* густини неперервного розподілу випадкової величини x (тривалості обслуговування) в точці $(\lambda - \lambda z)$. Перша похідна твірної функції в точці $z = 1$ та перетворення Лапласа-Стілтьєса в точці $z = 0$ дозволяє обчислити перші моменти випадкової величини, що розглядається. Тому з цього для даного рівняння випливає, що

$$K^{*'}(1) = -\lambda B^{*'}(0) = \lambda \bar{x} = \rho \qquad (7.61)$$

В (7.61) останнє рівняння випливає з (5.3) та (7.28).

К р о к 3 . Обчислимо стаціонарний розподіл для вкладеного ланцюга Маркова. Вкладений марковський ланцюг непривідний, аперіодичний та у випадку $\rho < 1$ – ергодичний (доведено за допомогою теореми Фостера). Тому існує точно один стаціонарний початковий розподіл (P_j) ланцюга, де P_j – імовірність того, що в стаціонарному стані вимога, що покидає систему, після себе залишає j вимог. Тому імовірності P_j задовольняють системі рівнянь, аналогічній (7.2):

$$P_j = \sum_i P_i p_{i,j} , \qquad (7.62)$$

тобто

$$P_0 = k_0(P_0 + P_1); \qquad P_j = P_0 k_j + \sum_{i=0}^{j+1} P_i k_{j-i+1}, \qquad j = 1, 2, \ldots$$

Аналогічно до (7.59) твірна функція $P^*(z)$ розподілу $P = (P_j)$

$$P^*(z) = \sum_{j=0}^{\infty} z^j P_j, \qquad |z| < 1,$$

маємо:

$$P^*(z) = P_0 \sum_{j=0}^{\infty} k_j z^j + \sum_{j=0}^{\infty} z^j \sum_{i=1}^{j+1} P_i k_{j-i+1} = P_0 K^*(z) + \sum_{i=1}^{\infty} P_i z^{i-1} K^*(z) .$$

Після скорочень виходить, що

$$P^*(z) = P_0 K^*(z) + \frac{K^*(z)}{z}(P^*(z) - P_0) = \frac{P_0(1-z)K^*(z)}{K^*(z) - z}. \qquad (7.63)$$

У (7.63) входить невідома поки що константа P_0. Її можна визначити за допомогою наступних обчислень, типових при роботі з твірними функціями.

Оскільки $\lim\limits_{z\to 1} P^*(z) = \lim\limits_{z\to 1} K^*(z) = 1$, а

$$\lim_{z\to 1} \frac{K^*(z) - z}{1 - z} = \lim_{z\to 1}\left(\frac{K^*(z) - 1}{1 - z} + \frac{1 - z}{1 - z}\right) = K^{*'}(1) + 1 = 1 - \rho,$$

то з (7.63) при $z \to 1$ випливає рівняння: $1 = P_0 / (1 - \rho)$, з чого

$$P_0 = 1 - \rho, \qquad (7.64)$$

$$P^*(z) = \frac{(1-\rho)(1-z)K^*(z)}{K^*(z) - z} = \frac{(1-\rho)(1-z)}{1 - \dfrac{z}{B^*(\lambda - \lambda z)}}. \qquad (7.65)$$

Формулу (7.65) називають співвідношенням Поллачека-Хінчина, яке задає твірну функцію стаціонарного розподілу вкладеного марковського ланцюга в точках, коли вимоги залишають систему *M*/*G*/1/∞, як функцію від перетворення Лапласа-Стілтьєса функції розподілу тривалості обслуговування.

К р о к 4. Перерахуємо отримані на кроці 3 результати у шукані величини системи: 1. Середня кількість вимог у системі 2. Середній час очікування у системі (в тому числі і для точок, відмінних від t_n).

1. Середня кількість вимог у системі *N*.

У розглянутій моделі *M*/*G*/1/∞ імовірності того, що в стаціонарному стані надходячи до системи вимога застає в ній *j* інших вимог, дорівнюють стаціонарним ймовірностям P_j наявності в системі *j* вимог безпосередньо після закінчення обслуговування деякої вимоги, і тому вони задаються формулою (7.65). Через властивість пуассонівського потоку вимог ці імовірності дорівнюють стаціонарним ймовірностям P_j у будь-який момент часу.

Математичне сподівання (середнє значення) кількості вимог *N*, що знаходяться в системі, дорівнює похідній від (7.65):

$$N = P^{*'}(1) = \rho + \frac{\lambda^2 \mu_x^2}{2(1-\rho)}, \qquad (7.66)$$

де

$$\mu_x^2 = \int_0^\infty x^2 dB(x) = \bar{x}^2 + \sigma_x^2. \qquad (7.67)$$

Параметри $\bar{x}^2$ та σ_x^2 характеризують математичне сподівання (середнє значення) та дисперсію або другий центральний момент функції розподілу випадкової тривалості обслуговування вимог *x*.

Формулу (7.66) називають формулою Поллачека-Хінчина.

2. Середній час очікування у системі *W*.

Нехай *W*(*t*) – стаціонарна функція розподілу часу очікування і *W**(s) – її перетворення Лапласа-Стілтьєса:

$$W^*(s) = \int_0^\infty e^{-st} dW(t).$$

Нехай *V* – функція розподілу часу перебування в стаціонарному стані; час перебування дорівнює часу очікування плюс час обслуговування. Оскільки час очікування і наступна тривалість обслуговування стохастично незалежні, то для перетворення Лапласа-Стілтьєса *V**(*s*) часу перебування в стаціонарному стані на підставі теореми згортки (множення) маємо:

$$V^*(s) = W^*(s)B^*(s). \tag{7.68}$$

Кількість вимог, що перебувають у системі, коли її залишає деяка вимога, дорівнює кількості вимог, що надходять до системи за час перебування в системі вимоги, що йде з неї. Тому між функцією розподілу часу перебування *V*(*t*) *i* розподілом P_j існує такий зв'язок:

$$P_j = \int_0^\infty \frac{(\lambda t)^j}{j!} e^{-\lambda t} dV(t). \tag{7.69}$$

Для твірної функції *P**(z) розподілу (P_i) із (7.69) з урахуванням (7.68) та (7.60) отримуємо:

$$P^*(z) = \sum_{j=0}^\infty z^j P_j = \int_0^\infty e^{-\lambda t(1-z)} dV(t) = V^*(\lambda - \lambda z) = W^*(\lambda - \lambda z)B^*(\lambda - \lambda z).$$

З цього та із (7.65) випливає, що

$$W^*(s) = \frac{P^*(1 - s/\lambda)}{B^*(s)} = \frac{(1-\rho)s}{s - \lambda + \lambda B^*(s)}.$$

Оскільки математичне сподівання $W = -W^{*\prime}(0)$, то взявши похідну для стаціонарного стану отримуємо:

$$W = \frac{\lambda \mu_x^2}{2(1-\rho)} = \frac{\lambda(\bar{x}^2 + \sigma_x^2)}{2(1-\rho)}. \tag{7.70}$$

Для характеристики другого центрального моменту розподілу тривалості обслуговування σ_x^2 часто використовується коефіцієнт варіації ν_x:

$$\nu_x = \frac{\sigma_x}{\bar{x}}. \qquad (7.71)$$

Суму $\bar{x}^2 + \sigma_x^2$ з урахуванням (7.71), можна представити як $\bar{x}^2(1+\nu_x^2)$. Підставивши це до (7.66) і (7.70) отримаємо остаточні формули розрахунку *N* та *W*. З урахуванням (6.3), (6.4), (7.44) та (7.45) отримуємо всі інші характеристики *QoS*, формули яких занесені до табл. 7.1. Значення *W*, t_q та *T* нормуються за $\bar{x}$.

Таблиця 7.1 – Характеристики якості обслуговування моделі *M*/*G*/l/∞

Хар-ка *QoS*	Модель		
	M/*G*/1/∞	*M*/*D*/1/∞	*M*/*M*/1/∞
$P_{w>0}$	ρ	ρ	ρ
Q	$\frac{\rho^2(1+\nu_x^2)}{2(1-\rho)}$	$\frac{\rho^2}{2(1-\rho)}$	$\frac{\rho^2}{1-\rho}$
W	$\frac{\rho(1+\nu_x^2)}{2(1-\rho)}$	$\frac{\rho}{2(1-\rho)}$	$\frac{\rho}{1-\rho}$
t_q	$\frac{1+\nu_x^2}{2(1-\rho)}$	$\frac{1}{2(1-\rho)}$	$\frac{1}{1-\rho}$
N	$\rho + \frac{\rho^2(1+\nu_x^2)}{2(1-\rho)}$	$\rho + \frac{\rho^2}{2(1-\rho)}$	$\rho + \frac{\rho^2}{1-\rho} = \frac{\rho}{1-\rho}$
T	$1 + \frac{\rho(1+\nu_x^2)}{2(1-\rho)}$	$1 + \frac{\rho}{2(1-\rho)}$	$1 + \frac{\rho}{1-\rho} = \frac{1}{1-\rho}$

До табл. 7.1 записано й формули всіх характеристик *QoS* моделей *M*/*D*/l/∞ та *M*/*M*/l/∞, для яких коефіцієнт варіації ν тривалості обслуговування регулярного та експонентного розподілів дорівнює 0 та 1 відповідно. З таблиці видно, що за постійної тривалості значення *W*, t_q та *Q* в два рази менші, ніж за експонентної. Наведені в табл. 7.1 значення характеристик *QoS* для моделі *M*/*M*/l/∞ відповідно збігаються з (7.27), (7.32), (7.33) і (7.34).

Для *m*-серверної системи *M*/*G*/*m*/∞ простого розв'язання не існує, однак є наближені оцінки. Зокрема, середній стаціонарний час очікування W_m задовольняє нерівності

$$W_m \le \frac{\bar{x}^2 + \sigma_x^2}{2m - 2\lambda\bar{x}}.$$

Доказ даного співвідношення наведено в [7].

7.6 Система з пріоритетами *M*/*G*/1/∞

Встановлення пріоритетів для вимог, що очікують, – один із ефективних способів керування розмірами черги і часом перебування в ній. При надходженні до системи вимоги з високим пріоритетом обслуговування вимоги з більш низьким пріоритетом або переривається (абсолютний пріоритет), або вимога з високим пріоритетом стає в початок черги вимог, що очікують, (відносний пріоритет). Розглянемо ці два випадки пріоритетів докладніше, при цьому у випадку абсолютного пріоритету візьмемо переривання з продовженням обслуговування (дообслуговування). Більш докладно з пріоритетними системами можна ознайомитися в [1, 2].

7.6.1 Відносний пріоритет

Нехай вимоги з r пріоритетами (чим менше номер, тим вище пріоритет) надходять до однолінійної системи і утворюють r пуассонівських потоків з інтенсивністю λ_k, де $k = 1, \ldots, r$. Отже, сукупний потік є пуассонівським з інтенсивністю $\lambda_{\Sigma} = \sum_{k=1}^{r} \lambda_k$. Нехай у межах даного пріоритету вимоги обслуговуються в порядку надходження. Функція розподілу загального часу обслуговування вимог всіх пріоритетів має вид $B(x) = \frac{1}{\lambda} \sum_{k=1}^{r} \lambda_k B_k(x)$, де $B_k(x)$ – функція розподілу часу обслуговування вимоги з k-м пріоритетом, що має середній час обслуговування $1/\mu_k$. Переривання обслуговування не допускається. Після закінчення обслуговування будь-якої вимоги з черги вибирається вимога з вищим пріоритетом.

Позначимо інтенсивність питомого навантаження k-го пріоритету:

$$\rho_k = \frac{\lambda_k}{\mu_k}, \qquad k = 1, \ldots, r; \tag{7.72}$$

та сумарне навантаження всіх потоків з 1 по k-й пріоритет:

$$R_k = \sum_{i=1}^{k} \rho_i, \text{ де } R_0 = 0, \quad k = 1, \ldots, r. \tag{7.73}$$

Припустимо, що сумарне навантаження потоків всіх пріоритетів $R_r < 1$ (умова ергодичності процесу). З використанням позначення (7.67), де $\int_0^{\infty} x^2 dB(x) = \bar{x}^2 + \sigma_x^2$, визначимо середній час очікування вимоги з k-м пріоритетом:

$$W_k = \frac{\lambda_{\Sigma}(\bar{x}^2 + \sigma_x^2)}{2(1 - R_{k-1})(1 - R_k)}. \tag{7.74}$$

Для виявлення рівня впливу відносних пріоритетів на середній час очікування W_k кожного з пріоритетів розглянемо випадок організації п’яти пріоритетів, де потоки кожного з пріоритетів однакової інтенсивності $\lambda_k = \lambda$, і

тривалість обслуговування вимог кожного з пріоритетів однакова та постійна. Звідси відповідно інтенсивність обслуговування $\mu_k = \mu$, а $\sigma_x^2 = 0$. Оскільки W_k будемо вимірювати в одиницях середньої тривалості обслуговування $\bar{x}$, то в чисельнику (7.74) при $\bar{x}^2 = 1/\mu = 1$ можна записати $\lambda_{\tilde{N}}\bar{x}^2 = \rho$.

На рис. 7.4 показано графіки залежності середнього часу очікування вимог *k*-го пріоритету від загальної інтенсивності навантаження ρ потоків усіх пріоритетів. Оскільки всі інтенсивності однакові, то в кожній точці графіка інтенсивність навантаження потоку *k*-го пріоритету буде 1 / 5 від загальної інтенсивності ρ.

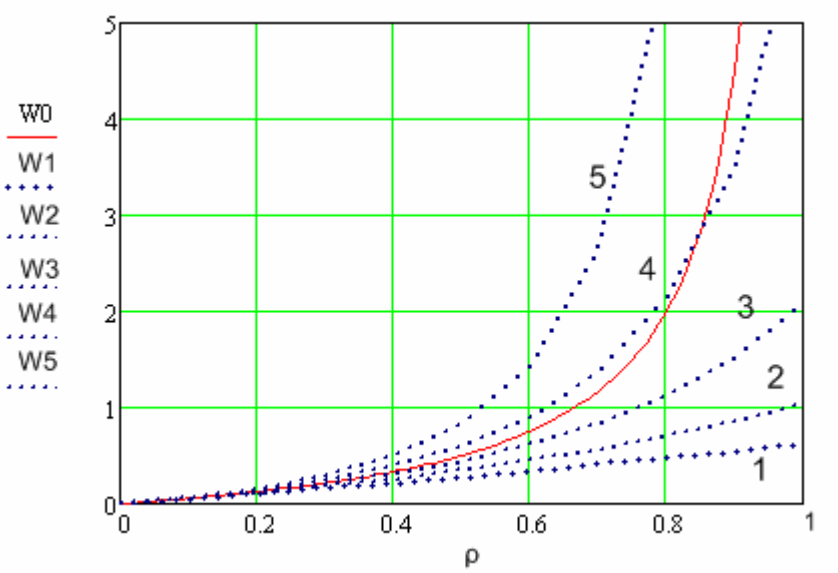


Рисунок 7.4 – Вплив відносних пріоритетів на середній час очікування

Тут пунктирними лініями зображені графіки середнього часу очікування W_k, вираженого в одиницях середньої тривалості обслуговування $\bar{x}$ при відносних пріоритетах (1, …, 5) і постійному часі обслуговування, однаковому для всіх п'яти однакових за інтенсивністю пуассонівських потоків вимог.

З рисунка видно, що при введенні пріоритетів середній час очікування для окремих потоків можна зробити менше, ніж це було б у випадку без пріоритетного обслуговування (показано безперервною лінією).

Якщо, наприклад, *r* = 2, то вимоги першого пріоритету є пріоритетними, а другого – непріоритетними. Середній час очікування пріоритетних і непріоритетних вимог з (7.74) визначиться:

$$W_1 = \frac{\lambda_{\tilde{N}}(\bar{x}^2 + \sigma_x^2)}{2(1-\rho_1)}. \qquad (7.75)$$

$$W_2 = \frac{\lambda_{\tilde{N}}(\bar{x}^2 + \sigma_x^2)}{2(1-\rho_1)(1-\rho_1-\rho_2)}. \qquad (7.76)$$

Середній час перебування вимоги *k*-го пріоритету в системі T_k визначається як сума W_k та середнього часу обслуговування 1 / μ_k.

7.6.2 Абсолютний пріоритет з дообслуговуванням

Можливі різні режими переривання обслуговування вимоги низького пріоритету у разі, коли до системи надходить вимога більш високого пріоритету. Розглянемо випадок з дообслуговуванням (див. розд. 3).

З (7.74) з урахуванням (7.73) та (7.72) випливає середній час очікування вимог першого пріоритету

$$W_1 = \frac{\lambda_1(\bar{x}_1^2 + \sigma_1^2)}{2(1-\rho_1)}.$$

і згідно з (7.66) середня кількість вимог першого пріоритету у системі

$$N_1 = \rho_1 + \frac{\lambda_1^2(\bar{x}_1^2 + \sigma_1^2)}{2(1-\rho_1)}.$$

Загальний вираз для середнього часу очікування вимог інших пріоритетів $1 \le k \le r$ виходить аналогічно (7.74):

$$W_k = \frac{\sum_{i=1}^{r} \lambda_i(\bar{x}_i^2 + \sigma_i^2)}{2(1-R_{k-1})(1-R_k)}.$$

Якщо, наприклад, $r = 2$, то є пріоритетні (пріоритет 1) і непріоритетні (пріоритет 2) вимоги. Середній час очікування непріоритетних вимог (до початку обслуговування)

$$W_2 = \frac{\lambda_1(\bar{x}_1^2 + \sigma_1^2) + \lambda_2(\bar{x}_2^2 + \sigma_2^2)}{2(1-\rho_1)(1-\rho_1-\rho_2)}. \qquad (7.77)$$

Звідси середній час перебування в системі непріоритетних вимог T_2 (з урахуванням переривань і дообслуговування) виходить додаванням середньої тривалості обслуговування, так що

$$T_2 = W_2 + \frac{1}{\mu_2(1-\rho_1)}. \qquad (7.78)$$

Порівняння (7.76) і (7.77) показує, що при $\bar{x}_1^2 + \sigma_1^2 = \bar{x}_2^2 + \sigma_2^2$ середній час очікування при абсолютному пріоритеті з дообслуговуванням такий самий, як і при відносному пріоритеті, але відмінні часи перебування у системі T. При відносному пріоритеті час перебування у системі дорівнює $(1 / \mu_2) + W_2$, тобто менше, ніж T_2 у формулі (7.78).

Складніше знайти характеристики черги непріоритетних вимог. Середня кількість непріоритетних вимог у системі

$$N_2 = \frac{1}{1-\rho_1}\left[\rho_2 + \frac{\lambda_1\lambda_2(\bar{x}_1^2 + \sigma_1^2) + \lambda_2(\bar{x}_2^2 + \sigma_2^2)}{2(1-\rho_1-\rho_2)}\right].$$

де λ_2 – інтенсивність потоку непріоритетних вимог. Тривалість обслуговування визначається розподілом $B_2(x)$ із середнім значенням $\bar{x}_2 = 1/\mu_2$ та $\rho_2 = \lambda_2/\mu_2$.

7.7 Система з втратами $M_B/M/m$ – мультисервісний вузол доступу

При створенні пакетних мереж нового покоління виникає проблема розрахунку пропускної здатності мереж широкосмугового мультисервісного доступу. На практиці при цьому використовують математичне моделювання, або, без належного обґрунтування, традиційні формули теорії телетрафіка. Загального аналітичного або інженерного методу вирішення проблеми немає.

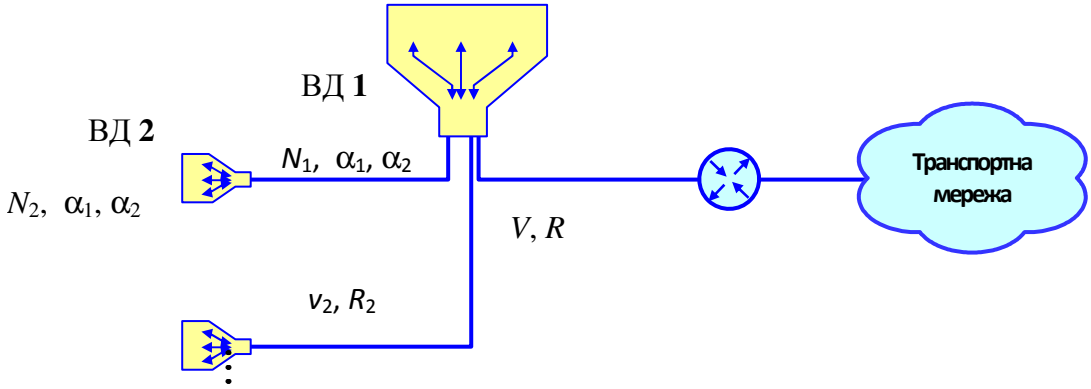


Рисунок 7.5 – Автономний сегмент мережі доступу

Типова структура автономного кластера мережі доступу (рис. 7.5) передбачає каскадне підключення вузлів доступу (ВД) і, у ряді випадків, можливість взаємного зв'язку абонентів кластера. Вона містить транзитний вузол доступу ВД1 і каскадно підключені через нього $ВД_2 \ldots ВД_m$. Вузлами доступу, залежно від конкретної технології, можуть бути мультиплексори ліній *xDSL* (*DSLAM*), базові станції *WiMAX* і/або *WiFi* та інше обладнання.

Потоки вимог внаслідок кінцевої кількості джерел описуються моделлю примітивного потоку (пуассонівського потоку другого роду). За примітивного потоку розподіл інтервалів між вимогами теж експонентний, але його параметр λ не постійний, а пропорційний до кількості вільних джерел навантаження.

Стани послідовно з'єднаних ВД залежні, тому що кожна вимога займає в них певну пропускну здатність, і до того ж через втрати змінюється характер навантаження, що надходить на транзитний ВД. Для отримання точних результатів слід розраховувати кожний автономний сегмент мережі доступу з каскадним включенням ВД (кластер) у цілому, не розчленовуючи його на окремі ВД, що пов'язано з певними обчислювальними труднощами. Позначимо:

$N_1, N_2 \ldots N_m$ – ємності відповідних вузлів доступу;

$R_1 \ldots R_m$ – швидкості передачі, які необхідно забезпечити для $ВД_1 \ldots ВД_m$;

R – загальна швидкість передачі, яка необхідна в напрямку до транспортної мережі від всіх вузлів доступу розглянутого кластера;

$v_1 \ldots v_m$ – розрахункова кількість умовних каналів для $ВД_1 \ldots ВД_m$ відповідно;

V – розрахункова загальна кількість умовних каналів для напрямку до транспортної мережі;

α_1, α_2 – інтенсивність одного джерела навантаження (абонента) у вільному стані відповідно усередині кластера та до транспортної мережі.

Треба визначити кількість умовних каналів так, щоб не перевищувалися нормативні втрати вимог, а потім визначити для кожного такого каналу швидкість передачі так, щоб не перевищувалися норми втрат пакетів, тоді визначимо необхідну смугу пропускання в Мбіт/с кожного напрямку зв'язку.

Вимоги обслуговуються з явними втратами, час обслуговування кожної вимоги є експонентним, а потік вимог від кожної групи абонентів є примітивним, тобто має властивості ординарності, стаціонарності та простої післядії. При цьому стани системи розподілені за законом Енгсета, який ще називається розподілом Ерланга-Бернуллі або усіченим розподілом Бернуллі. Тому модель потоку позначено символом M_B – експонентний розподіл інтервалу між вимогами з усіченим розподілом Бернуллі станів системи. Така модель характерна для невеликих груп абонентів (до 300).

Позначимо $P(k_1, l_1; \ldots ; k_m, l_m)$ – імовірність наявності на ВД$_1$…ВД$_m$ відповідно $k_1 \ldots k_m$ з'єднань усередині кластера й $l_1 \ldots l_m$ з'єднань до транспортної мережі. Виходячи з формули Енгсета [2] для конкретного i-го ВД імовірність наявності $k_i + l_i$ з'єднань дорівнює:

$$\tilde{N}_{N_i}^{l_i} \alpha_2^{l_i} C_{N_i - l_i}^{k_i} \alpha_1^{k_i} p_0, \tag{7.79}$$

де p_0 – імовірність відсутності з'єднань. Тоді для сталого режиму маємо:

$$P(k_1, l_1; \ldots; k_m, l_m) = \frac{1}{C} \prod_{i=1}^{m} \tilde{N}_{N_i}^{l_i} \alpha_2^{l_i} C_{N_i - l_i}^{k_i} \alpha_1^{k_i}, \tag{7.80}$$

де константа C визначається з умови, що нормує $\sum P(k_1, l_1; \ldots k_m, l_m) = 1$, а сума включає всі стани, для яких виконується: $0 \le k_i + l_i \le v_i$, $\sum_{i=1}^{m} l_i \le V$.

Розрахунки за формулою (7.80) громіздкі. Їх можна спростити, якщо припустити симетричність кластера мережі доступу, тобто задати $N_i = N$, $v_i = v$ для всіх значень i ($0 \le i \le m$). Спрощення досягається за рахунок появи множин станів кластера, у кожній з яких всі стани однаково ймовірні [15].

Розглянемо кластер доступу з урахуванням допущень щодо його симетричності. Позначимо i_{kl} – кількість ВД, що мають по k з'єднань усередині кластера й по l з'єднань до транспортної мережі. Для окремо взятого ВД імовірність наявності $k + l$ з'єднань визначається формулою (7.79) з врахуванням на те, що $N_i = N$. Тоді ймовірність того, що із всіх ВД i_{00} мають по $k = 0$ і $l = 0$ з'єднань, i_{01} – по $k = 0$ і $l = 1$ з'єднань і так далі:

$$P(i_{00}, i_{01}, \ldots, i_{v0}) = \tilde{N}_m^{i_{00}} (C_N^0 \alpha_2^0 C_{N-0}^0 \alpha_1^0)^{i_{00}} \times C_{m-i_{00}}^{i_{01}} (C_N^1 \alpha_2^1 C_{N-1}^0 \alpha_1^0)^{i_{01}} \times \ldots$$
$$\times C_{m - \sum_{k=0}^{v-1} \sum_{l=0}^{v-k} i_{kl}}^{i_{v0}} (C_N^0 \alpha_2^0 C_{N-0}^v \alpha_1^v)^{i_{v0}} \times P(m, 0, \ldots, 0). \tag{7.81}$$

Тут $P(m, 0 \ldots 0 \ldots 0) = P_0$ – імовірність того, що в кластері немає жодного з'єднання, тобто $i_{00} = m$. Можна показати, що

$$\tilde{N}_m^{i_{00}} C_{m-i_{00}}^{i_{01}} ... C_{m-\sum_{k=0}^{v-1}\sum_{l=0}^{v-k} i_{kl}}^{i_{v0}} = \frac{m!}{i_{00}! i_{01}! .. i_{v0}!}.$$

Тоді рівняння (7.81) запишеться у такий спосіб:

$$P(i_{00}, i_{01}, ..., i_{v0}) = \frac{m!}{\prod_{k=0}^{v-l}\prod_{l=0}^{v} i_{kl}!} \prod_{l=0}^{v}\prod_{k=0}^{v-l} (C_N^l \alpha_2^l C_{N-l}^k \alpha_1^k)^{i_{kl}} \times P_0. \quad (7.82)$$

Таким чином, імовірності макростанів кластера доступу описуються мультиноміальним розподілом. Очевидні наступні обмеження:

$$0 \le i_{kl} \le m, \quad 0 \le k + l \le v, \ \sum_{l=0}^{v}\sum_{k=0}^{v-l} i_{kl} = m, \quad 0 \le \sum_{l=0}^{v}\sum_{k=0}^{v-l} l i_{kl} \le V.$$

Імовірність наявності j з'єднань ($0 \le j \le V$) від всіх m ВД у загальному напрямку

$$P_j = \sum_j P(i_{00}, i_{01}, ..., i_{v0}) = B_j(m) P_0, \quad (7.83)$$

де підсумовування виконується по k і l, що задовольняє рівнянню $\sum_{l=0}^{v}\sum_{k=0}^{v-l} l i_{kl} = j$.

Перетворимо вираз для $B_j(m)$ у вид, зручний для розрахунків. Оскільки ймовірність наявності j з'єднань у загальному напрямку можна представити добутком імовірності наявності l з'єднань від певного ВД на ймовірність наявності $j - l$ з'єднань від інших $m - 1$ ВД, то має місце рекурентне співвідношення:

$$B_j(m) = \sum_{l=0}^{v}\sum_{k=0}^{v-l} C_N^l \alpha_2^l C_{N-l}^k \alpha_1^k B_{j-l}(m-1). \quad (7.84)$$

З урахуванням рівняння $j \frac{m!}{\prod_{k=0}^{v-l}\prod_{l=0}^{v} i_{kl}!} = m \sum_{s=0}^{v}\sum_{r=0}^{v-s} s \frac{i_{rs}(m-1)!}{\prod_{l=0}^{v}\prod_{k=0}^{v-l} i_{kl}!}$ отримаємо друге рекурентне співвідношення:

$$jB_j(m) = \sum_j m \sum_{s=0}^{v}\sum_{r=0}^{v-s} s \frac{i_{rs}(m-1)!}{\prod_{k=0}^{v-l}\prod_{l=0}^{v} i_{kl}!} \prod_{l=0}^{v}\prod_{k=0}^{v-l} (C_N^l \alpha_2^l C_{N-l}^k \alpha_1^k)^{i_{kl}} = \quad (7.85)$$

$$= m \sum_{s=1}^{v}\sum_{r=0}^{v-s} s C_N^s \alpha_2^s C_{N-s}^r \alpha_1^r B_{j-s}(m-1).$$

Після кількох перетворень отримаємо:

$$jB_j(m) = mN\alpha_2 \sum_{s=1}^{v} C_{N-1}^{s-1} \alpha_2^{s-1} B_{j-s}(m-1) \sum_{r=0}^{v-s} C_{N-s}^r \alpha_1^r. \quad (7.86)$$

З урахуванням рівняння $C_N^l = C_{N-1}^l + C_{N-1}^{l-1}$ з (7.86) треба

$$mNB_j(m) = mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-s}(m-1)\sum_{r=0}^{v-s} C_{N-s}^r \alpha_1^r + \\ + mN\alpha_2 \sum_{s=0}^{v} C_{N-1}^{s-1} \alpha_2^{s-1} B_{j-s}(m-1)\sum_{r=0}^{v-s} C_{N-s}^r \alpha_1^r. \tag{7.87}$$

З порівняння (7.86) і (7.87) випливає:

$$(mN - j)B_j(m) = mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-s}(m-1)\sum_{r=0}^{v-s} C_{N-s}^r \alpha_1^r.$$

Тоді

$$\alpha_2[(mN-(j-1)]B_{j-1}(m) = \alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1)\sum_{r=0}^{v-s} C_{N-s-1}^r \alpha_1^r + \\ + \alpha_1\alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1)\sum_{r=0}^{v-s} C_{N-s-1}^{r-1} \alpha_1^{r-1}. \tag{7.88}$$

Розглянемо перший доданок правої частини виразу (7.88):

$$\alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1)\sum_{r=0}^{v-s} C_{N-s-1}^r \alpha_1^r = \\ = \alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1)\sum_{r=0}^{v-s-1} C_{N-s-1}^r \alpha_1^r + \\ + \alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1) C_{N-s-1}^{v-s} \alpha_1^{v-s}. \tag{7.89}$$

Позначимо $s + 1 = s'$. Тоді у формулі (7.89)

$$\alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1)\sum_{r=0}^{v-s-1} C_{N-s-1}^r \alpha_1^r = \\ = \alpha_2 mN\sum_{s'=1}^{v+1} C_{N-1}^{s'-1} \alpha_2^{s'-1} B_{j-s'}(m-1)\sum_{r=0}^{v-s'} C_{N-s'}^r \alpha_1^r = \\ jB_j(m) + \alpha_2 mNC_{N-1}^v \alpha_2^v B_{j-v-1}(m-1) C_{N-v-1}^{-1} \alpha_1^{-1} = jB_j(m). \tag{7.90}$$

Розглянемо другий доданок правої частини виразу (7.88) і при цьому позначимо $r - 1 = r'$. Тоді

$$\alpha_1\alpha_2 mN\sum_{s=0}^{v} C_{N-1}^s \alpha_2^s B_{j-1-s}(m-1)\sum_{r=0}^{v-s} C_{N-s-1}^{r-1} \alpha_1^{r-1} = \\ \alpha_1\alpha_2 mN\sum_{s'=1}^{v+1} C_{N-1}^{s'-1} \alpha_2^{s'-1} B_{j-s'}(m-1)\sum_{r'=0}^{v-s'} C_{N-s'}^{r'} \alpha_1^{r'} = \alpha_1 jB_j(m). \tag{7.91}$$

Підставивши (7.89), (7.90) і (7.91) в (7.88), одержимо:

$$\alpha_2[(mN-(j-1)]B_{j-1}(m)=$$

$$jB_j(m)+\alpha_2 mN\sum_{s=0}^{v}C_{N-1}^{s}\alpha_2^{s}B_{j-1-s}(m-1)C_{N-s-1}^{v-s}\alpha_1^{v-s}+\alpha_1 jB_j(m).$$

З огляду на те, що $\tilde{N}_{N-1}^{s}C_{N-s-1}^{v-s}=C_{N-1}^{v}C_{v}^{s}$ й замінивши індекс підсумовування колишнім позначенням, одержимо третє, найбільш зручне для розрахунків рекурентне співвідношення:

$$jB_j(m)=\frac{\alpha_2}{1+\alpha_1}[(mN+j+1)B_{j-1}(m)-mNC_{N-1}^{v}\sum_{l=0}^{v}C_v^l\alpha_2^l\alpha_1^{v-l}B_{j-1-l}(m-1)], \quad (7.92)$$

$$B_{-j}\equiv 0.$$

Для $j=0$ з (7.87) потрібно, щоб $B_0(m)=B_0(m-1)\sum_{k=0}^{v}\tilde{N}_N^k\alpha_1^k$. Очевидно також, що $B_0(1)=\sum_{k=0}^{v}\tilde{N}_N^k\alpha_1^k$. Цю суму можна виразити через формулу Енгсета, наведену таблицях [2]: $\varepsilon(N,v,\alpha_1)=\dfrac{C_N^v\alpha_1^v}{\sum_{k=0}^{v}C_N^k\alpha_1^k}$. Тоді для $1\le i\le m$ отримаємо:

$$B_0(1)=\frac{C_N^v\alpha_1^v}{\varepsilon(N,v,\alpha_1)}, \qquad B_0(i)=[B_0(1)]^i. \quad (7.93)$$

З урахуванням (7.93) з виразу (7.92) послідовно визначаються всі значення $B_j(m)$, за якими розраховуються характеристики якості обслуговування абонентів у схемі рис. 7.5. Позначимо:

- P_i – імовірність зайнятості i умовних каналів для розглянутого ВД при зв'язку всередині кластера;
- $П_j$ – імовірність зайнятості j умовних каналів у напрямку до транспортної мережі;
- P_{ij} – імовірність одночасної зайнятості i умовних каналів для розглянутого ВД при зв'язку всередині кластера та j умовних каналів у напрямку до транспортної мережі.

Можна показати, що мають місце співвідношення:

$$P_i=\frac{\sum_{l=0}^{i}C_N^l\alpha_2^lC_{N-l}^{i-l}\alpha_1^{i-l}\sum_{j=0}^{V-l}B_j(m-1)}{\sum_{x=0}^{V}B_x(m)}, \qquad Ï_j=\frac{B_j(m)}{\sum_{x=0}^{V}B_x(m)}, \quad (7.94)$$

$$P_{ij} = \frac{\sum_{l=0}^{i} C_N^l \alpha_2^l C_{N-l}^{i-l} \alpha_1^{i-l} B_{j-l}(m-1)}{\sum_{x=0}^{V} B_x(m)}.$$

Звідси визначаються втрати за часом P_t – імовірності зайнятості всіх доступних умовних каналів. Для зв'язку всередині кластера $P_t = P_v$. Для зв'язку в напрямку транспортної мережі величина втрат за часом визначається сумою двох імовірностей – імовірності P_v і імовірності зайнятості V умовних каналів за умови, що для розглянутого ВД є хоча б один вільний умовний канал зв'язку із транзитним ВД, тобто $P_t = P_v + \sum_{i=0}^{v-1} P_{iV} = P_v + \ddot{I}_V - P_{vV}$. Однак реально важливі втрати не за часом, а за вимогами, обумовлені відношенням інтенсивностей потоків втрачених і вхідних вимог. Для конкретного ВД інтенсивність потоків вимог всередині кластера $\lambda_к$ та зовнішнього $\lambda_в$ відповідно дорівнюють:

$$\lambda_{\hat{e}} = \sum_{i=0}^{v} \alpha_1 (N-i) P_i, \qquad \lambda_{\hat{a}} = \sum_{i=0}^{v} \alpha_2 (N-i) P_i. \tag{7.95}$$

Інтенсивності втрачених вимог становлять:

- для потоку всередині кластера – $\alpha_1(N-v)P_v$;
- для зовнішнього потоку на ділянці до транзитного вузла – $\alpha_2(N-v)P_v$;
- на ділянці до транспортної мережі – $\sum_{i=0}^{v-1} \alpha_2 (N-i) P_{iV}$.

Тоді втрати вимог P_B для зв'язку всередині кластера і для зовнішнього зв'язку становлять відповідно:

$$P_{B\hat{e}} = \frac{\alpha_1 (N-v) P_v}{\lambda_{\hat{e}}}, \qquad P_{B\hat{a}} = \frac{\alpha_2 (N-v) P_v + \sum_{i=0}^{v-1} \alpha_2 (N-i) P_{iV}}{\lambda_{\hat{e}}}. \tag{7.96}$$

Для будь-якого ВД інтенсивності обслуженого навантаження при зв'язку всередині кластера та зовнішнього відповідно дорівнюють:

$$Y_{\hat{e}} = \sum_{i=0}^{v} i P_i, \qquad Y_{\hat{a}} = \sum_{j=0}^{V} j P j. \tag{7.97}$$

Таким чином, використовуючи рекурентні співвідношення (7.92) і (7.93), можна розрахувати втрати вимог і відповідну їм пропускну здатність (Ерл) мережі мультисервісного абонентського доступу. Рекурентний метод розрахунку кількості умовних каналів дозволяє оцінити похибку наближених методів розрахунку, оскільки дає точні результати в окремому випадку симетричності мережі доступу, тобто при однакових ємностях вузлів доступу й однакових параметрів абонентського навантаження кожного вузла.

7.8 Модель обслуговування мультисервісного трафіка

У мультисервісних мережах зв'язку, що базуються на пакетних технологіях передавання інформації, одними й тими самими трактами зв'язку передаються мовні потоки, потоки даних, відео й ін. При цьому передавання окремих видів інформації вимагає різних швидкостей передачі. Тому в залежності від категорії вимоги на надання певного сервісу (послуги) для кожного з потоків інформації необхідно зайняти певний ресурс із спільного ресурсу пропускної здатності (Кбіт/с). Наприклад, передавання мови вимагає гарантованої швидкості передавання 64 Кбіт/с, зв'язок відео-конференції з кодеком H.263 – 320 Кбіт/с, обмін файлами, наприклад, 1024 Кбіт/с [17]. Для розподілу спільного ресурсу швидкості передавання між всіма цими послугами весь ресурс можна представити певною кількістю умовних портів, де швидкість передавання кожного складає мінімальну швидкість із послуг, що надаються. У даному разі швидкість передавання одного умовного порту складе 64 Кбіт/с і в залежності від категорії вимоги на надання послуги виконується одночасне зайняття декількох портів, через які передаються пакети із заданою бітовою швидкістю. Таким чином, при надходженні вимоги на надання послуги передавання мови буде одночасно зайнято 1 порт, на відео-конференцзв'язок – 5, а на передавання файлів – 16 умовних портів.

Примітка.

1. Без зміни сутності моделі масштаб швидкості передавання умовного порту може бути й іншим. Наприклад, вона може бути кратною швидкості передавання одного або декількох пакетів переданої інформації. Таким чином, змінюється тільки кількість займаних умовних портів, кожний з яких представляє собою комірку пам’яті накопичування й передавання пакетів.
2. При надходженні вимоги на надання певного сервісу встановлюється з’єднання між відповідними споживачами сервісу, а потім в межах цього з’єднання передаються інформаційні пакети із заданою для сервісу швидкістю. Кожна вимога вимагає встановлення одного з’єднання.

За такої моделі обслуговування мультисервісного трафіка необхідно відрізняти потік вимог на надання сервісу від потоку зайняття портів, оскільки вони суттєво відрізняються за своїми властивостями. Очевидно, якщо потік вимог на надання сервісу є ординарним, то потік зайняття портів буде неординарним, тому що для обслуговування окремих з’єднань порти займаються групами. З урахуванням зазначеної обставини варто розрізняти й два поняття навантаження: навантаження за вимогами і навантаження на порти.

Для СМО (системи серверів) миттєва інтенсивність обслуговуваного навантаження за вимогами у момент часу t має значення $j(t)$, що дорівнює кількості одночасно обслуговуваних вимог (з’єднань). На відміну від цього, інтенсивність обслуговуваного навантаження на порти – це кількість $i(t)$ одночасно зайнятих портів у момент t. У загальному випадку $i(t) \neq j(t)$, тому що одна вимога (з’єднання) може займати відразу кілька портів.

Аналогічно необхідно розрізняти відповідні види вхідного навантаження: за вимогами і на порти. Миттєве значення інтенсивності вхідного навантаження за вимогами у момент часу t є випадкова величина, яка дорівнює кількості вимог, що обслуговуються в момент часу уявлюваною СМО з нескінченною кількістю одночасних з'єднань (див. п. 5.1: один сервер – одно з'єднання).

Розглянемо окремий випадок, коли для обслуговування кожної вхідної вимоги необхідна однакова кількість m вільних портів. Тоді в будь-який момент часу обслуговуване навантаження на порти буде в m раз більше навантаження, що обслуговується за вимогами: $i(t) = mj(t)$. Якщо потік вимог експонентний і характеризується параметром λ, то вхідне навантаження за вимогами буде пуассонівським, а його математичне сподівання $\Lambda_{\text{В}}$ і дисперсію $D_{\text{В}}$ можна обчислити за допомогою наступного співвідношення:

$$\Lambda_{\hat{\text{А}}} = D_{\hat{\text{А}}} = \lambda\bar{x},$$

де $\bar{x}$ – середній час обслуговування однієї вимоги (з'єднання).

В уявлюваній системі з нескінченною кількістю портів, як і в реальній, кількість зайнятих портів буде в m раз більше кількості обслуговуваних вимог (з'єднань). Звідси випливають формули, за якими можна визначити інтенсивність вхідного навантаження на порти і його дисперсію:

$$\Lambda_{\ddot{\text{I}}} = m\Lambda_{\hat{\text{А}}} = m\lambda\bar{x}, \qquad D_{\ddot{\text{I}}} = m^2 D_{\hat{\text{А}}} = m^2\lambda\bar{x}. \tag{7.98}$$

Тут застосовується наступне правило, відоме з теорії ймовірностей: якщо випадкова величина збільшується на постійний коефіцієнт, то на цей же коефіцієнт потрібно помножити її математичне сподівання, а дисперсія повинна збільшуватися на коефіцієнт у квадраті.

З наведених формул видно, що $D_{\text{П}} > \Lambda_{\text{П}}$, і це навантаження є скупченим через неординарність потоку займань портів. Коефіцієнт скупченості $S_{\text{П}}$ (5.4) дорівнює кількості портів, які необхідні для обслуговування однієї вимоги:

$$S_{\ddot{\text{I}}} = \frac{D_{\ddot{\text{I}}}}{\Lambda_{\ddot{\text{I}}}} = \frac{m^2\lambda\bar{x}}{m\lambda\bar{x}} = m. \tag{7.99}$$

При надходженні вимог від джерел різних категорій формули (7.98) справедливі для математичного сподівання і дисперсії навантаження на порти, що створюється вимогами i-ї категорії, $i = 1, \ldots, n$:

$$\Lambda_i = m_i\lambda_i\bar{x}_i, \qquad D_i = m_i^2\lambda_i\bar{x}_i.$$

Підставляючи ці співвідношення у формулу (5.8) та (5.7), можемо знайти коефіцієнт скупченості об'єднаного навантаження на порти:

$$S_{\ddot{\text{I}}} = \frac{\sum_{i=0}^{n} D_i}{\sum_{i=0}^{n} \Lambda_i} = \frac{\sum_{i=0}^{n} m_i^2\lambda_i\bar{x}_i}{\sum_{i=0}^{n} m_i\lambda_i\bar{x}_i}. \tag{7.100}$$

Тут коефіцієнт скупченості мультисервісного навантаження дорівнює середньозваженій кількості портів m, які потрібні для обслуговування вимог окремих категорій, з вагами $m_i\lambda_i\bar{x}_i$, що дорівнюють інтенсивності навантаження на порти, створюваного вимогами цих категорій, $i = 1, \ldots, n$.

7.8.1 Апроксимація Хейворда

На повнодоступну систему із V портів надходить пуассонівський потік вимог з інтенсивністю λ, причому кожна вимога потребує для свого обслуговування одночасно m портів, $m > 1$. Порти займаються на випадковий час обслуговування, середня тривалість якого дорівнює $\bar{x}$. У разі завершення обслуговування всі порти з групи одночасно звільняються. Якщо в момент надходження вимоги в системі відсутня необхідна кількість вільних портів, то вимога втрачається.

При визначенні ймовірності втрат вимог π для аналізованої системи, що назвемо системою U, безпосереднє застосування B-формули Ерланга неможливе у силу неординарного потоку займань портів. Замість цього досліджуємо модифіковану систему U', що складається з $v = V / m$ комплектів, кожний з яких містить m портів. Тепер кожній вхідній вимозі для обслуговування буде необхідний один такий комплект і тому потік зайняття комплектів буде ординарним. Навантаження в системі U' визначаються кількістю зайнятих комплектів (а не портів), тобто збігається з навантаженням за вимогами і є пуассонівським з інтенсивністю $\Lambda_B = \lambda \bar{x}$. Тому для розрахунку ймовірності втрат вимог в системі U' можна вжити B-формулу Ерланга:

$$\pi' = E_v(\Lambda_{\hat{A}}) = \frac{\dfrac{\Lambda_{\hat{A}}}{v!}}{\sum\limits_{i=0}^{v} \dfrac{\Lambda_i}{i!}}.$$

З погляду статистичних характеристик процесу обслуговування вимог, системи U і U' цілком еквівалентні і, зокрема, очевидно, що $\pi = \pi'$. Звідси з урахуванням співвідношень (7.99) і (7.100) випливає:

$$\pi = E_{\left(\frac{V}{S_{\ddot{I}}}\right)}\left(\frac{\Lambda_{\ddot{I}}}{S_{\ddot{I}}}\right) \qquad (7.101)$$

Отже, коли мультисервісне навантаження створюється декількома категоріями джерел з різною кратністю вимог m_i, суперпозицію вхідних потоків можна замінити одним потоком, що має такі самі значення математичного сподівання $\Lambda_П$ і дисперсії $D_П$ навантаження на порти. Хоча в дійсності непуассонівські потоки мають досить складні статистичні властивості, і для повного опису таких потоків потрібне використання більшої кількості характеристик, на практиці звичайно припускають, що імовірність втрати вимоги слабко залежить від моментів навантаження більш високого порядку і їх можна не враховувати.

Таким чином, якщо відома імовірність втрат вимог π у деякій системі з V портів при мультисервісному навантаженні з інтенсивністю $\Lambda_П$ і коефіцієнтом скупченості $S_П$, то в будь-якій системі з такими ж параметрами V, $\Lambda_П$ і $S_П$ втрати вимог будуть приблизно дорівнювати π.

Обчисливши коефіцієнт скупченості об'єднаного потоку вимог за формулою (7.100), можемо потім за допомогою співвідношення (7.101)

визначити імовірність втрати будь-якої вимоги, що дає наближену оцінку середніх (або загальних) втрат. Для різних категорій джерел навантаження вимоги втрачаються при різних станах системи і, як наслідок, імовірнісні характеристики якості обслуговування вимог будуть відрізнятися. Для розрахунку індивідуальних втрат, тобто імовірності втрат вимог *i*-ї категорії, при $i = 1, \ldots, n$, можна вжити наближене співвідношення:

$$\pi_i = \frac{m_i}{S_i}\pi .$$

Таким чином, при обслуговуванні мультисервісного навантаження, що має непуассонівський характер, розрахунок втрат вимог у вихідній системі заміняється аналогічною задачею для еквівалентної системи, де така задача може бути вирішена з використанням *B*-формули Ерланга. Порівняльний аналіз такого підходу і імітаційного моделювання свідчить, що пропоновані наближені формули мають точність, цілком достатню для розв'язання інженерних задач.

Формула (7.101) називається формулою або апроксимацією Хейворда і її обґрунтування досліджено в деяких наукових роботах [18].

7.8.2 Приклад розрахунку імовірності втрат

На систему з $V = 45$ портів надходять вимоги $n = 2$ категорій.

Потік вимог сервісу телефонії (1-ї категорія) з параметрами:

$$\lambda_1 = 0{,}14 \text{ вимоги за 1 с, } \bar{x}_1 = 100 \text{ с, } m_1 = 1.$$

Потік вимог сервісу відео-телефонії (2-ї категорія) з параметрами:

$$\lambda_2 = 0{,}05 \text{ вимоги за 1 с, } \bar{x}_2 = 56 \text{ с, } m_2 = 5.$$

Визначити загальні (середні) і індивідуальні втрати вимог.

Розв'язання задачі.

Загальна інтенсивність навантаження на порти від джерел усіх категорій:

$$\Lambda_\Pi = m_1\lambda_1\bar{x}_1 + m_2\lambda_2\bar{x}_2 = 1 \cdot 0{,}14 \cdot 100 + 5 \cdot 0{,}05 \cdot 56 = 28 \text{ (Ерл)}.$$

Дисперсія навантаження на порти від джерел усіх категорій:

$$D_\Pi = m_1^2\lambda_1\bar{x}_1 + m_2^2\lambda_2\bar{x}_2 = 1^2 \cdot 0{,}14 \cdot 100 + 5^2 \cdot 0{,}05 \cdot 56 = 84.$$

Коефіцієнт скупченості навантаження:

$$S_\Pi = D_\Pi / \Lambda_\Pi = 84 / 28 = 3.$$

За формулою Хейворда і з таблиць зі значеннями *B*-формули Ерланга визначаємо середню імовірність втрат для загального вхідного потоку вимог:

$$\pi = E_{\left(\frac{V}{S_\Pi}\right)}\left(\frac{\Lambda_\Pi}{S_\Pi}\right) = E_{\left(\frac{45}{3}\right)}\left(\frac{28}{3}\right) = E_{15}(9{,}33) = 0{,}027.$$

Індивідуальні ймовірності втрат для вимог 1-ї і 2-ї категорій:

$$\pi_1 = (m_1 / S_\Pi) \cdot \pi = (1 / 3) \cdot 0{,}027 = 0{,}009,$$
$$\pi_2 = (m_2 / S_\Pi) \cdot \pi = (5 / 3) \cdot 0{,}027 = 0{,}045.$$

Із такою якістю система з пропускною здатністю 45 · 64 Кбіт/с = 2,88 Мбіт/с обслужить навантаження сервісів телефонії та відео-телефонії.

8 АНАЛІЗ СМО В УМОВАХ РЕАЛЬНОГО ПОТОКУ ВИМОГ

Інтегральний характер мультисервісних мереж з розширеним спектром надаваних послуг зумовлює різнорідність трафіка, яка сильно змінює його параметри та математичну модель. Параметрами трафіка є інтенсивність навантаження Λ (середня кількість вимог, що надійшла до системи за середню тривалість обслуговування) та дисперсія інтенсивності навантаження σ^2. Математичною моделлю трафіка є імовірнісна функція розподілу випадкової величини кількості вимог i за середню тривалість обслуговування x.

У математичній моделі пуассонівського потоку вимог інтервал часу між вимогами z розподілений за експонентним законом. Ступінь відхилення інших потоків від моделі пуассонівського потоку можна оцінити за коефіцієнтом варіації v_z функції розподілу інтервалу z. Для експонентного розподілу $v_z \equiv 1$. Модель пуассонівського потоку не завжди адекватно описує реальні потоки вимог в телекомунікаціях і тому необхідно вибирати інші розподіли для їхнього опису, що забезпечують кращу згоду з даними вимірів. Заміна експонентного розподілу будь-якими іншими функціями значно ускладнює математичну модель, а складні моделі не завжди піддаються аналітичному розв'язанню.

Реальні потоки вимог у мультисервісних мережах зв'язку формуються множиною джерел з різною питомою інтенсивністю навантаження (різнорідні потоки). В процесі створення потоку вимог беруть участь джерела, що належать до тієї або іншої групи споживачів сервісів з близькими інтенсивності навантаження. Значення інтенсивності результуючого потоку вимог в кожну мить залежить від того, до якої групи за інтенсивністю навантаження належить джерело і яке співвідношення чисельності цих джерел з іншими. Отже, більш адекватно описати такий потік або розподіл інтервалів часу між вимогами можна не експонентним розподілом (*M*), а їхньою сумішшю – гіперекспонентним розподілом (*HM*):

$$P(z) = \sum_{i=1}^{k} p_i \lambda_i e^{-\lambda_i z} \qquad \text{при} \sum_{i=1}^{k} p_i = 1.$$

Даному розподілу відповідає перервний пуассонівський потік k-го порядку. Практичні виміри свідчать, що реальні потоки достатньо апроксимувати з $k = 2$. Цей розподіл описує більший розкид величини інтервалу часу між вимогами z і забезпечує значення коефіцієнтів $v_z \geq 1$, а це, у свою чергу, дозволяє описувати реальні потоки з дисперсією інтенсивності навантаження σ^2, що перевищує її математичне сподівання Λ від одиниць до десятків разів. Співвідношення σ^2 та Λ визначає пікфактор трафіка.

Гіперекспонентний інтервал часу між вимогами призводить до такого розподілу кількості вимог, що надходять до системи за середню тривалість їх обслуговування, який досить вдало апроксимується нормальним (Гаусса) законом. Отже для реальних потоків вимог мультисервісних мереж зв'язку адекватною є математична модель з гіперекспонентним розподілом інтервалу часу між вимогами, що апроксимується нормальним розподілом інтенсивності навантаження Λ.

8.1 Функція розподілу станів системи з втратами *HM/D/m*

Повнодоступна схема СМО з m серверами функціонує за дисципліною обслуговування з втратами. Тривалість обслуговування постійна й дорівнює $\bar{x}$. До системи надходить потік вимог з інтенсивністю Λ, в якому інтервал між вимогами має гіперекспонентний розподіл, а кількість вимог Λ на одиницю часу $\bar{x}$ розподілена за нормальним законом. Визначити стаціонарний розподіл імовірностей станів системи P_k ($k = 0, \ldots, m$).

Стан системи визначається кількістю зайнятих серверів. Випадковий процес надходження вимог модифікує стан системи зі швидкістю, обумовленою інтенсивністю навантаження Λ. Інтенсивність навантаження Λ – це сумарна кількість вимог j, які надійшли за час, що дорівнює середній тривалості обслуговування $\bar{x}$. Таким чином, інтенсивність переходу системи з одного стану в другий залежить від властивостей потоку вимог, що описується відповідним розподілом імовірностей Q_j, де j – кількість вимог за час $\bar{x}$, яка може бути в діапазоні від 0 до ∞. Усі події надходження вимог належать просторові станів Q. Події зайняття серверів утворюють новий дискретний підпростір P, який описано розподілом імовірностей P_i, де i – кількість зайнятих серверів ($i = 0, \ldots, m$). Простір P, безсумнівно, менше простору Q, оскільки стани $i > m$ для системи неможливі, а j може бути як завгодно велике.

Оскільки система обслуговування із втратами, то незалежно від її поточного стану для кожного з випадків надходження $j > m$ вимог відбувається подія «відмова в обслуговуванні». Це очевидно, оскільки за час $\bar{x}$ (постійна тривалість зайняття серверів) надійде $j > m$ вимог і жоден зі знову зайнятих серверів за цей же час не звільниться. Якщо система перебуває в початковому стані $i = 0$ (усі сервери вільні), то в цьому випадку для кожного з варіантів надходження за той самий відрізок часу $j \le m$ вимог подія «відмова в обслуговуванні» не відбувається. Події, які полягають в надходженні за час $\bar{x}$ точно j вимог, утворюють повну групу неспільних гіпотез H_0, H_1, … , H_j з апріорними ймовірностями $Q(H_0)$, $Q(H_1)$, … , $Q(H_j)$ і тому $\sum_{j=0}^{\infty} Q(H_j) = 1$. Подія A, яка полягає у „відсутності відмови в обслуговуванні” може відбуватися тільки разом з однією із гіпотез групи H_0, H_1, …, H_m. Умовні імовірності гіпотез надходження $j \le m$ вимог (апостеріорні), за умови здійснення події A (відсутність втрат вимог) обчислюються за формулою Байеса:

$$Q(H_j \mid A) = \frac{Q(H_j)Q(A \mid H_j)}{\sum_{k=0}^{m} Q(H_k)Q(A \mid H_k)}.$$

Через те що подія A (відсутність відмови в обслуговуванні), яка з'являється з кожною із гіпотез групи H_0, H_1, …, H_m є достовірною, то умовні імовірності $Q(A|H_j) = 1$. Тому, якщо для здійснення події можлива тільки частина гіпотез, а інші неможливі, то для отримання апостеріорних ймовірностей треба кожну з апріорних ймовірностей цієї частини можливих гіпотез розділити на їхню суму. Отже:

$$Q(H_j \mid A) = \frac{Q(H_j)}{\sum_{k=0}^{m} Q(H_k)}.$$

Для системи, що перебуває в початковому стані (усі сервери вільні), умовні імовірності $Q(H_j|A)$ відповідають ймовірностям зайняття $i = j$ серверів системи, тобто ймовірностям P_i. Вони також утворять повну групу неспільних подій підпростору P і тому $\sum_{i=0}^{m} P_i = 1$. Таким чином, імовірності зайняття серверів P_i представлено через імовірності надходження вимог $Q(H_j)$, тобто

$$P_i = \frac{Q(H_i)}{\sum_{k=0}^{m} Q(H_k)}. \qquad (8.2)$$

Умовою стаціонарності отриманого розподілу є ергодичність процесу обслуговування вимог. Це значить, що отримані імовірності станів системи (кількості зайнятих серверів) не повинні залежати від того, в якому стані система була у початковий момент (було прийнято, що в початковий момент усі сервери вільні). Відомо, що властивість ергодичності мають *марковські* процеси і для будь-якого такого процесу після досить тривалого часу функціонування системи обов'язково настане стаціонарний режим, де імовірність того, що система буде в *i*-му стані, не залежить від того, у якому стані вона знаходилася в початковий момент часу. Відповідно до теореми Маркова умовою ергодичності процесу є (7.10) і тоді система працює в стані статистичної рівноваги.

Підставивши в (8.2) замість імовірностей $Q(H_i)$ імовірності з розподілу Пуассона (7.48), отримаємо перший розподіл Ерланга (7.11), справедливий для моделі *M/G/m*, і за яким можна розрахувати імовірності всіх станів системи. Однак у цьому випадку, неодмінно, математичне сподівання випадкового процесу надходження вимог з інтенсивністю Λ дорівнює її дисперсії σ^2. Але якщо при розрахунках (8.2) замість імовірностей $Q(H_i)$ і $Q(H_k)$ підставити імовірності, що розраховано за нормальним законом, то отримаємо формулу розрахунку станів системи саме в моделі *HM/D/m*:

$$P_i = \frac{\frac{1}{\sqrt{2\pi\sigma}} e^{-(i-\Lambda)^2/2\sigma^2}}{\sum_{k=0}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-(k-\Lambda)^2/2\sigma^2}} = \frac{1}{\sum_{k=0}^{m} \exp\left[\frac{-(k-2\Lambda+i)(k-i)}{2\sigma^2}\right]}. \qquad (8.3)$$

При $\sigma^2 = \Lambda$ нормальний випадковий процес є *марковським* і при цьому відповідно до (7.10) при $\Lambda < m$ розрахункові значення, отримувані за формулами розподілу Ерланга (7.11) і (8.3) досить близькі – розбіжність не більше 1 %. При $\sigma^2 > \Lambda$ розподіл інтервалів часу між вимогами вже не експонентний і потік вимог втрачає властивість відсутності післядії. Процес обслуговування стає складнішим і виникає залежність ймовірностей станів

системи від її початкового стану. Крім того, виникає залежність і від виду розподілу тривалості обслуговування вимог, на відміну від розподілу Ерланга, точному за будь-якого закону розподілу тривалості обслуговування. Через це імовірність зайняття всіх серверів системи вже не збігається з імовірністю відмови в обслуговуванні, як це спостерігається за пуассонівського потоку.

Дану задачу для моделі *HM/D/m* можна вирішити і в інший спосіб.

В умовах необмеженої кількості серверів (модель *HM/D/*∞) вимоги обслуговуються без втрат. За постійної тривалості обслуговування *x*, коли нема втрат, властивості потоку звільнень збігаються із властивостями потоку надходження вимог, тому що відбувається тільки зсув у часі на величину *x* між моментом надходження вимоги та моментом її виходу із системи (завершення обслуговування). При цьому стани системи повністю визначаються властивостями потоку вимог, а імовірнісні функції розподілу кількості вимог у системі *i* і вхідної кількості вимог *j* за час *x*, повністю збігаються. За нормального розподілу кількості вимог, що надходять в систему за середню тривалість їх обслуговування, нормальна функція розподілу потоку вимог P_j визначить функцію розподілу станів системи P_i:

$$P_j = P_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(i-\Lambda)^2}{2\sigma^2}}. \qquad (8.4)$$

У разі обмеженої до *m* кількості серверів простір станів системи буде також обмеженим від 0 до *m*. У цьому випадку моделі *HM/D/m* імовірнісна функція розподілу станів системи може бути апроксимована усіченим нормальним законом, що визначає імовірності станів системи P_i в межах $0 \le j \le m$:

$$P_i = \frac{A}{\sqrt{2\pi}\sigma} e^{-\frac{(j-\Lambda)^2}{2\sigma^2}}, \qquad (8.5)$$

де

$$A = \frac{1}{\frac{1}{\sqrt{2\pi}}\left[\int_0^{\frac{m-\Lambda}{\sigma}} e^{-\frac{u^2}{2}} du - \int_0^{\frac{0-\Lambda}{\sigma}} e^{-\frac{u^2}{2}} du\right]}.$$

Якщо $m = \infty$, то $A = 1$ і тому (8.4) та (8.5) збігаються. При всіх інших *m* значення формул (8.3) та (8.5) співпадають, що підтверджує правильність обох способів вирішення задачі. Для (8.3), (8.4) та (8.5) виконується $\sum_{i=0}^{m} P_i = 1$.

На рис. 8.1 наведені графіки апроксимації функції розподілу станів системи, яка обслуговує потік вимог з параметрами: інтенсивність $\Lambda = 100$ Ерл та дисперсія $\sigma^2 = 400$ (пікфактор $\sigma^2 / \Lambda = 4$). Пунктирною лінією показана система з необмеженою кількістю серверів, де $m = \infty$, безперервною – система з $m = 125$, штриховою – з $m = 115$ та штрихпунктирною – з $m = 110$ серверів.

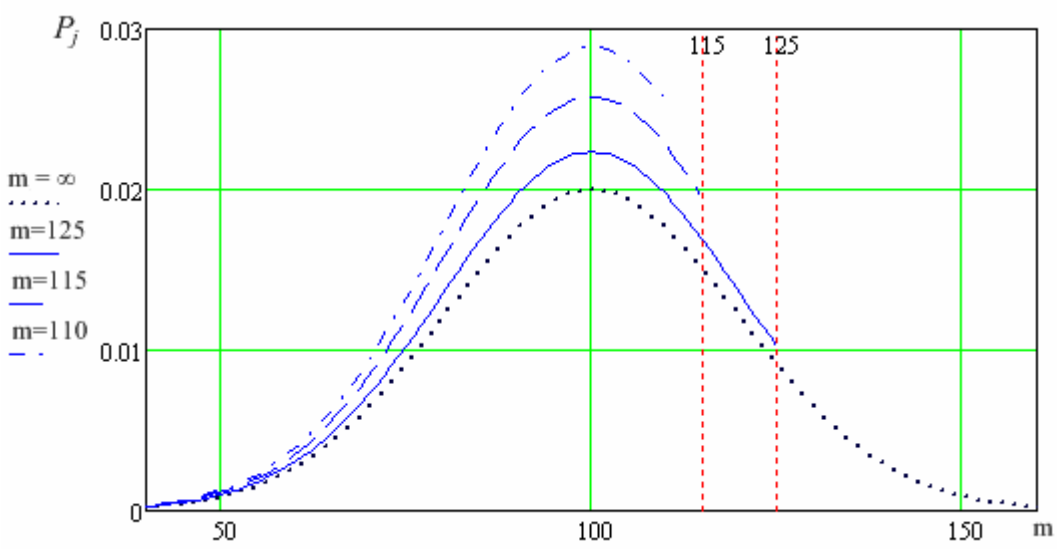


Рисунок 8.1 – Апроксимація станів усіченим нормальним законом

В широкому діапазоні зміни параметрів трафіка Λ і σ^2 та кількості серверів m відносна похибка запропонованої апроксимації усіченим нормальним законом не перевищує 1% для всіх значень P_i, окрім $P_{j=m}$. Процес обслуговування вимог у системі є ергодичним або не залежить від початкового стану системи за умови $m > \Lambda$ [2]. Якщо кількість серверів m суттєво перевищує значення інтенсивності Λ, то початковий стан системи не дуже впливає на функцію розподілу станів системи. При зменшенні m початковий стан системи виявляє свій найбільший вплив саме на імовірність $P_{i=m}$. Це відбувається через „скупченість" потоку вимог реального трафіка, яка випливає з його властивості $\sigma^2 > \Lambda$. Саме у разі „перевантаження" системи, коли зайняті всі сервери, при звільненні будь-якого з серверів він тут же займається черговою вимогою через скупченість вимог на даному інтервалі часу, а це підтримує систему більш тривалий час у стані „насичення", тобто у стані $P_{i=m}$.

За умови $\sigma^2 \equiv \Lambda$ закони Гаусса та Пуассона досить близько збігаються вже при $\Lambda > 20$. Тому тут даний усічений нормальний розподіл дає значення ймовірностей станів системи, які дуже близькі до значень, що отримуються за розподілом Ерланга (7.11). За цим розподілом визначаються стани системи типу $M/G/m$ з експонентним розподілом інтервалу часу між вимогами потоку, для якого $\sigma^2 \equiv \Lambda$. Гіперекспонентий розподіл за $k = 1$ обертається в експонентний і тому він є його узагальненням. Але на відміну від розподілу Ерланга, який дійсний за будь-якого (G) закону розподілу тривалості обслуговування вимог, усічений нормальний розподіл станів системи при $\sigma^2 > \Lambda$ справедливий тільки для моделі $HM/D/m$, тобто за постійної тривалості обслуговування.

Регулярним законом розподілу тривалості обслуговування (постійна тривалість) може бути описана робота керуючих пристроїв вузлів комутації або тривалість оброблення пакетів в пакетних мережах передавання даних, де потоки не є пуассонівськими. Саме для цих випадків застосування запропонованої апроксимації буде адекватним.

8.2 Імовірність втрат системи *HM/D/m*

Співвідношення перших двох моментів випадкової кількості вимог за середню тривалість обслуговування (інтенсивність навантаження) визначає скупченість інтенсивності навантаження або пікфактор трафіка

$$S = \frac{\sigma^2}{\Lambda}.$$

У найбільш часто застосовуваній математичній моделі пуассонівського потоку вимог $\sigma^2 \equiv \Lambda$, оскільки інтервал часу між вимогами z розподілений за експонентним законом, а кількість вимог за середню тривалість обслуговування розподілена за законом Пуассона. Реальним потокам властива підвищена нерівномірність трафіка, за якої дисперсія інтенсивності навантаження σ^2 перевищує її математичне сподівання Λ від 2 до 15 разів. Іноді дане співвідношення буває ще більше, але це відбувається або за межами ГНН, або на невеликих пучках каналів. Отже для реальних потоків, як правило, $S > 1$.

Визначення основної характеристики *QoS* – імовірності відмови в обслуговуванні через зайнятість всіх серверів системи – базується на визначенні імовірнісної функції розподілу станів системи P_i.

У моделі *HM/D/m* імовірність відмови в обслуговуванні залежить від пікфактора інтенсивності навантаження S і за умови $m > \Lambda$ дорівнює імовірності зайняття всіх m серверів системи (8.3), помноженій на S:

$$P_B = \frac{1}{\sum_{k=0}^{m} \exp\left[\frac{-(k-2\Lambda+m)(k-m)}{2\sigma^2}\right]} \frac{\sigma^2}{\Lambda}. \qquad (8.6)$$

У [10] показано, що регулярний закон розподілу тривалості обслуговування з точки зору якості обслуговування є найгіршим і тому формула (8.6) є не що інше, як верхня оцінка можливих значень імовірності відмови в обслуговуванні. Для інших законів розподілу тривалості обслуговування вимог до формули (8.6) застосовується спеціальна апроксимаційна функція, яка отримана за результатами імітаційного моделювання [9], і в якій коефіцієнтом h задається вид закону розподілу тривалості обслуговування:

$$P_B = \frac{S}{\sum_{k=0}^{m} \exp\left[\frac{-(k-2\Lambda+m)(k-m)}{2\Lambda S}\right]} \left[1 - \frac{(S^2-1)(\sigma-\Lambda+m)}{\sigma(S^2 h - h + 5)}\right], \qquad (8.7)$$

де h дорівнює 4.25, 3.55 і 2.85 для рівномірного, експонентного і логарифмічно нормального законів розподілу тривалості обслуговування відповідно. Видно, що при $S = 1$ (тобто $\sigma^2 = \Lambda$) дана формула перетворюється у формулу (8.3) для випадку $i = m$ – стан системи m (зайняті всі сервери).

На рис. 8.2 наведено залежності імовірності втрат P_B від кількості серверів у системі $HM/G^*/m$ та виду закону розподілу тривалості обслуговування при $\Lambda = 100$ Ерл та $S = 8$ (G^* – регулярний, рівномірний, експонентний та логарифмічно нормальний закони розподілу тривалості обслуговування).

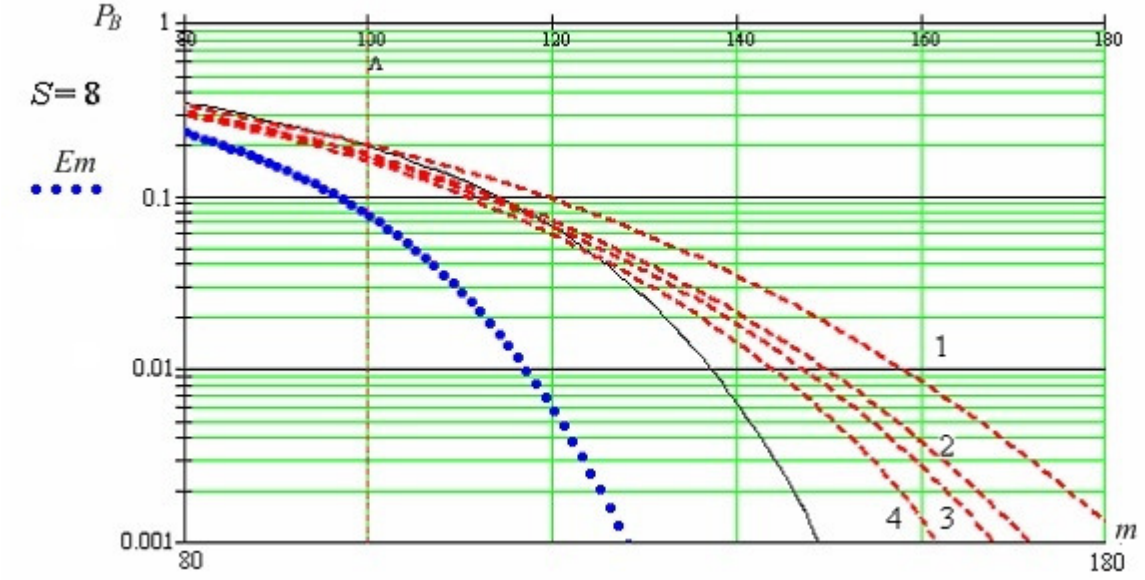


Рисунок 8.2 – Залежність P_B від m та закону розподілу тривалості зайняття

Пунктирна крива $E_m(\Lambda)$ репрезентує залежність втрат від m, розраховану за B-формулою Ерланга. Штриховими лініями 1, 2, 3 та 4 показано залежність імовірності втрат вимог для регулярного, рівномірного, експонентного та логарифмічно нормального законів розподілу тривалості зайняття відповідно.

Дані графіки демонструють, що, наприклад, за ємності системи $m = 125$ серверів імовірність втрати вимоги, що розрахована за B-формулою Ерланга складе $P_B = 0{,}002$. У той самий час, якщо до системи надходить не пуассонівський потік, а більш нерівномірний потік з пікфактором $S = 8$, то за постійної тривалості обслуговування вимог реальні втрати складуть $P_B = 0{,}07$, що у 35 разів більше. Для того, щоб за цих умов втримати якість обслуговування на заданому рівні $P_B = 0{,}002$, то необхідно в системі мати реально $m = 172$ сервери, а не 125. Саме такими великими є помилки в розрахунках СМО, коли використовуються неадекватні методи, які не враховують характеру вхідного потоку вимог.

З рисунка видно, що за умов значної ($S = 8$) дисперсії інтенсивності навантаження краща якість обслуговування буде там, де в законі розподілу більша частка (імовірність) коротших за тривалістю обслуговування вимог. Коротші за тривалістю обслуговування вимоги скоріше йдуть із системи, звільняючи сервери на використання наступними вимогами.

В табл. 1 Додатку наведено дані щодо інтенсивності навантаження Y, обслуженого m-серверним повнодоступним пучком при втратах 1, 3 і 5 ‰ та коефіцієнті скупченості вхідного навантаження $S = 1, \ldots, 5$. При $S = 1$ надані значення співпадають зі значеннями, розрахованими за B-формулою Ерланга.

8.3 Система з необмеженою чергою *HM/D/m/*∞

До повнодоступної *m*-серверної системи з необмеженою чергою надходить гіперекспонентний потік вимог з інтенсивністю Λ, пікфактором $S > 1$ та нормальним розподілом кількості вимог на одиницю часу *x*, де *x* – постійна тривалість обслуговування вимоги. Вибір із черги – в порядку надходження. Необхідно визначити основні характеристики *QoS*:

- імовірність очікування $P_{w>0}$;
- середню довжину черги *Q*;
- середню тривалість очікування вимог у системі *W*;
- середню тривалість очікування вимог у черзі t_q.

Розподіл імовірностей P_i випадкової кількості вимог *i*, що надходять до системи за одиницю часу *x* описується *нормальним* законом розподілу:

$$P_i = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{\frac{-(i-\Lambda)^2}{2\cdot\sigma^2}}. \qquad (8.8)$$

Незалежно від виду потоку вимог *W* та *Q* мають розраховуватись за формулами (7.44) та (7.45) на підставі визначення та формули Літтла відповідно. З цих формул видно, що дані характеристики *QoS* залежать від $P_{w>0}$ та t_q, які можуть бути визначені з функцій розподілу станів системи P_j та розподілу часу очікування вимог у черзі $P(t_q)$. Однак для непуассонівського потоку не існує загального методу отримання таких функцій, і формули для них не є простими.

Для розрахунку t_q вживаємо такі, встановлені раніше, результати:

- з *C*-формули Ерланга випливає, що в системі *M/M/m/*∞ середня тривалість очікування вимог у черзі $t_{q(M)} = 1 / (m - \Lambda)$;
- з формули Поллачека-Хінчина випливає, що в системі *M/D/*1/∞ середня тривалість очікування вимог у черзі $t_{q(D)} = t_{q(M)} / 2$.

Очевидно, що в шуканому виразі для розрахунку t_q системи *HM/D/m/*∞ мають враховуватися дані наслідки. Перший – тому, що пуассонівський потік (*M*) є окремим випадком гіперекспонентного (*H*). Другий – тому що односерверна система ($m = 1$) є окремим випадком багатосерверної.

В [11] для системи *HM/D/m/*∞ показано, що $t_{q(D)} > t_{q(M)}$ у $S / (k = 2)$ разів за кількості серверів $m = \Lambda$. Це добре узгоджується з наведеними вище співвідношеннями – через пікфактор *S* враховано відмінність гіперекспонентного потоку від пуассонівського, і відмінність у два рази середньої тривалості очікування при постійній та експонентній тривалості обслуговування, але віднесено це до характерної точки $m = \Lambda$.

Зі зростанням ємності системи *m* коефіцієнт $k = 2$ спадає приблизно зі швидкістю $k(m) \approx (m + \Lambda) / m$. За результатами імітаційного моделювання системи *HM/D/m/*∞ встановлено, що точність розрахунку t_q підвищується при заміні даної залежності на $k(m) \approx (m + \Lambda + 1 + \Lambda / m) / m$. Остаточна формула для розрахунку t_q системи *HM/D/m/*∞ має вид:

$$t_q \approx \frac{1}{m-\Lambda} \cdot S \cdot \frac{m}{m+\Lambda+1+\Lambda/m} = \frac{S}{(m+1)\cdot[1-(\Lambda/m)^2]}. \qquad (8.9)$$

Для розрахунку $P_{w>0}$ застосуємо такі аргументи. Імовірність $P_{w>0}$ – це імовірність зайнятості всіх m серверів (7.47).

Аналогічно до моделі $M/D/m/\infty$, при кінцевому числі m і необмеженій кількості місць очікування вимоги обслуговуються без втрат. На сервери системи надходять вимоги з первинного потоку з інтенсивністю Λ та із черги з інтенсивністю $\Lambda \cdot P_{>0} \cdot t_w$ (див. 7.34). Тому загальна інтенсивність навантаження на систему збільшується до величини $\Lambda_2 = \Lambda + Q$ (див. п. 7.4).

В умовах гіперекспонентного потоку функція розподілу кількості вимог у системі (обслуговуються та у черзі) або станів системи P_j відрізняється від функції розподілу P_i кількості вимог, що надходять. На рис. 8.3 подано розподіли станів системи за гіперекспонентного потоку вимог з параметрами $\Lambda = 100$ Ерл та $S = 4$ для ємності системи $m = 105$, 110 та 120 серверів ($m > \Lambda$).

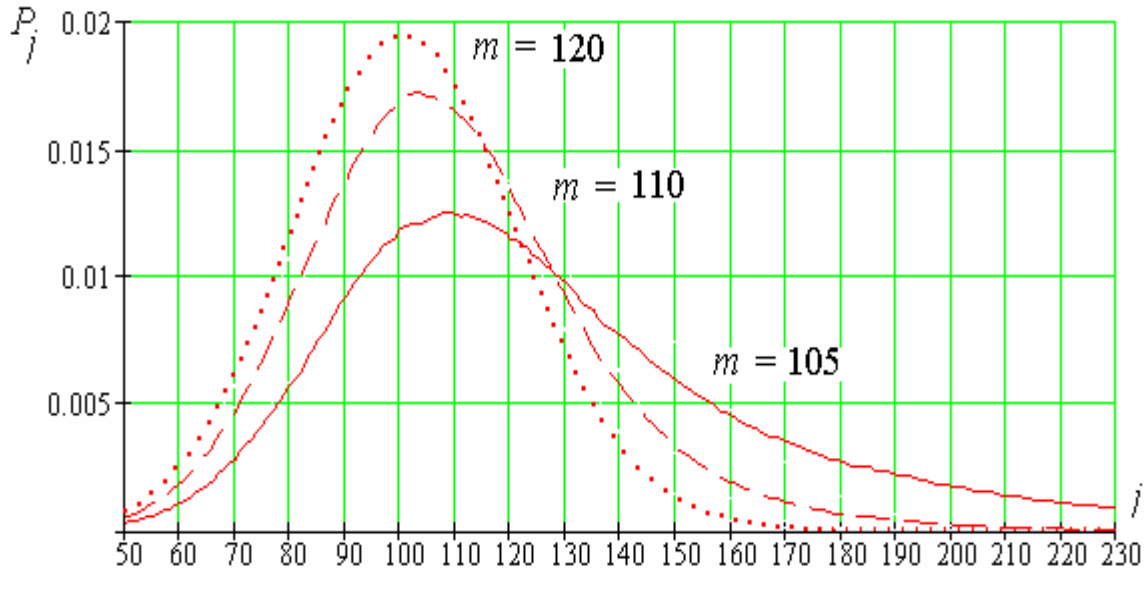


Рисунок 8.3 – Розподіл станів системи *HM/D/m/∞*

З рисунка видно, що при зменшенні m, розкид окремих значень функції розподілу станів системи від середнього значення збільшується. З чого випливає, що додатковий потік вимог із черги збільшує загальну інтенсивність навантаження Λ_2 та її дисперсію σ_2^2. Помітно, що вже при $m = 120$ (пунктирна лінія) функція розподілу станів системи доволі симетрична, що дозволяє її цілком апроксимувати нормальним законом розподілу. Апроксимація цих функцій нормальним законом (8.8) з параметрами

$$\Lambda_2 = \Lambda + Q; \qquad (8.10)$$

$$\sigma_2 \approx \sigma + \frac{Q}{2}. \qquad (8.11)$$

дає достатньо точні результати на лівому відрізку функції розподілу станів системи, обумовленому межами підсумовування у (7.47), тобто від 0 до m – 1:

$$P_{w>0} = 1 - \sum_{i=0}^{m-1} P_i \, .$$

З цього випливає простий ітераційний алгоритм розрахунку основних характеристик якості обслуговування системи *HM/D*/*m*/∞:

– із (8.9) для заданих Λ, S та m розраховується t_q;

– із (7.47) та (8.8) для заданих Λ і σ^2 визначається первинна імовірність $P_{w>0}$ (нібито для випадку, коли вимоги із черги не йдуть в систему і не збільшують навантаження на неї);

– для розрахованих t_q і $P_{w>0}$ відповідно до (7.44) і (7.44) визначаються первинні значення W і Q;

– для розрахованих із (8.10) і (8.11) значень Λ_2 та σ_2 відповідно до (7.47) і (8.8) визначається уточнена імовірність $P_{w>0}$, тобто з урахуванням впливу додаткового навантаження на сервери системи із черги. (При цьому довжина черги більш реальна, оскільки вимоги, що не йдуть із системи негайно, сприяють зростанню черги);

– за уточненим $P_{w>0}$ із (7.44) і (7.45) уточнюються значення W та Q.

Шляхом імітаційного моделювання встановлено, що реалізація цього алгоритму у великому діапазоні варіювання параметрів Λ, S і m дає завжди занижену оцінку імовірності $P_{w>0}$, однак при цьому відносна помилка ніколи не перевищує -10 %. Тому, оскільки на останньому кроці знову уточнюються значення W і Q, то можна ще раз перерахувати $P_{w>0}$ з більш точними значеннями Λ_3 і σ_3. Перевірка цього кроку показала, що результати розрахунків після третьої ітерації завжди дають верхню оцінку імовірності $P_{w>0}$, яка не перевищує +10 % [11].

8.4 Система з необмеженою чергою *fBM*/*D*/1/∞

Трафік мультисервісних мереж з комутацією пакетів характеризується наявністю довгострокових залежностей в інтенсивності навантаження й істотною відмінністю статистичних властивостей потоків пакетів від пуассонівського потоку. Адекватною моделлю потоків в таких мережах вважаються самоподібні (*self-similarity*) процеси, де вхідний потік описується фрактальним броунівським рухом (модель *fBM*). Однак дослідження характеристик якості обслуговування CPI в цих умовах є дуже складною математичною задачею. Причиною цьому є слабка формалізованість моделі самоподібних потоків, внаслідок чого й неможливо отримати аналітично обґрунтовані результати для оцінки параметрів *QoS* у системах розподілу інформації.

Для односерверної системи з нескінченною чергою та постійним часом обслуговування (модель *fBM*/*D*/1/∞) наближене рішення наведено в [1], де показано, що кількість вимог у розглянутій системі в будь-який момент часу *t* може бути представлено випадковою величиною

$$N(t) = \sup_{s \le t} (A(t) - A(t) - \mu(t - s)),$$

де

$$A(t) = \lambda t + \sqrt{a\lambda} Z(t).$$

Випадковий процес $Z(t)$ є нормалізованим фрактальним броунівським рухом з параметром Херста H ($H = 0{,}5 \ \ldots \ 1$), а позитивний коефіцієнт a є деяким множником масштабу.

В стані статистичної рівноваги при $\rho = \dfrac{\lambda}{\mu} < 1$ імовірність того, що кількість вимог у системі N перевищить задану величину x, представлено у виді функції:

$$\Pr(N > x) = \Pr\left(\sup_{t>0} \left(Z(t) - \frac{\mu - \lambda}{\sqrt{a\lambda}} t \right) > \frac{x}{\sqrt{a\lambda}} \right) = f\left(\left(\frac{\mu - \lambda}{\sqrt{a\lambda}} \right)^{(1-H)/H} \frac{\mu - \lambda}{\sqrt{a\lambda}} \right).$$

Для випадку, коли ця імовірність дорівнює наперед заданій величині $\Pr(N > x) = \varepsilon$ із вищенаведеного випливає, що

$$\frac{1-\rho}{\rho^{\frac{0.5}{H}}} \mu^{\frac{H-0.5}{H}} x^{\frac{1-H}{H}} = \frac{a^{\frac{0.5}{H}}}{f(\varepsilon)} = \text{const},$$

а це означає, що

$$N = x = \frac{(1-\rho)^{\frac{H}{H-1}}}{\rho^{\frac{0.5}{H-1}}}. \qquad (8.12)$$

Величину x знайдено з попереднього в припущенні, що $\text{const} = 1$. Імовірність, яка дорівнює одиниці – це достовірна подія і, отже, x – це кількість вимог у системі, яку не може бути перевищено, тобто це може бути верхня оцінка середньої кількості вимог N у системі $fBM/D/1/\infty$.

Оскільки з формули Літтла випливає, що $T = N / \lambda$, то середній час перебування вимоги в системі при $\mu = 1$, тобто в одиницях часу тривалості обслуговування визначається формулою:

$$T = \frac{(1-\rho)^{\frac{H}{H-1}}}{\rho^{\frac{0.5}{H-1}}} \frac{1}{\rho} = \frac{(1-\rho)^{\frac{H}{H-1}}}{\rho^{\frac{H-0.5}{H-1}}}. \qquad (8.13)$$

Виходячи з того, що для будь-якої односерверної системи середня довжина черги $Q = N - \rho$, то з урахуванням (8.12) отримаємо:

$$Q = \frac{(1-\rho)^{\frac{H}{H-1}}}{\rho^{\frac{0.5}{H-1}}} - \rho. \qquad (8.14)$$

Результат (8.12), (8.13) і (8.14) вважається аналітичним розв'язанням для системи $fBM/D/1/\infty$. Однак при аналізі цього результату можна помітити, що при завданні коефіцієнта Херста $H = 0{,}5$ (несамоподібний процес) маємо відомий результат для середньої кількості вимог, середньої тривалості перебування і середньої довжини черги в системі типу $M/M/1/\infty$. Це є нелогічним результатом, оскільки спочатку досліджувалася система з детермінованим часом обслуговування $fBM/D/1/\infty$. При зміні коефіцієнта Херста від значення $H = 1$ (максимальне значення) до $H = 0{,}5$ (мінімальне значення), безсумнівно, видозмінюється потік вимог і відповідна функція розподілу імовірності інтервалу часу між вимогами, але не змінюється функція розподілу тривалості обслуговування. Зрозуміло, при $H = 0{,}5$ потік утрачає властивості самоподібності, але в цьому випадку результати (8.12–8.14) мають корелюватися з результатами для деякої моделі з постійною тривалістю обслуговування, а не експонентною [12].

З наведеного на рис. 8.6 графіка залежності середнього часу перебування вимог в системі T витікає, що при навантаженні $\rho > 0{,}3$ система $fBM/D/1/\infty$ із вхідним потоком вимог, що має характер самоподібного процесу, буде витрачати на обробку більше часу, чим при відсутності самоподібності, тобто чим системи $M/M/1/\infty$ і $M/D/1/\infty$.

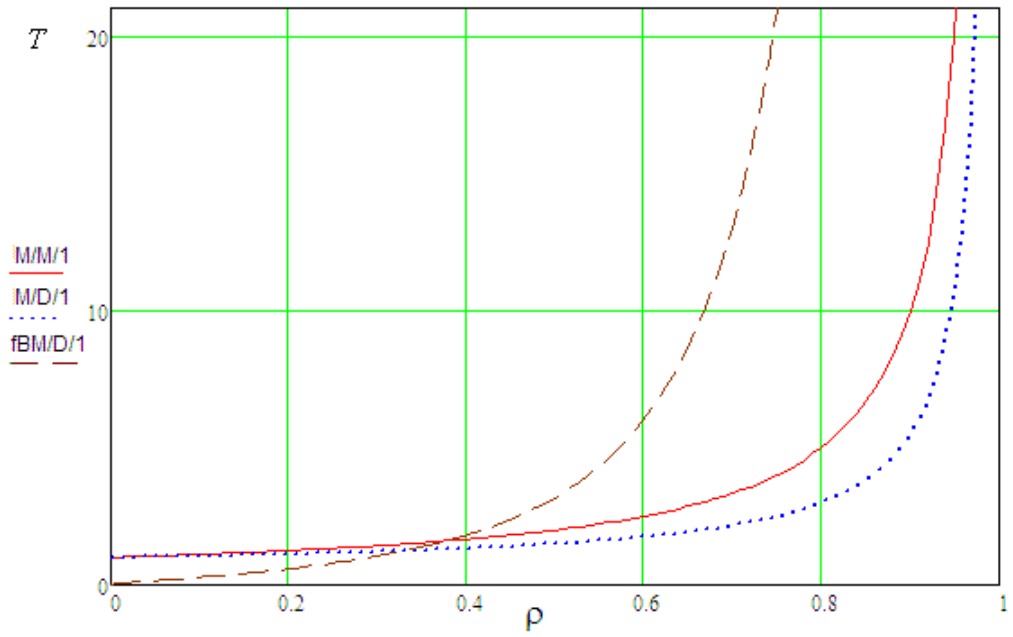


Рисунок 8.6 – Залежність середнього часу перебування вимог в системі T для моделей $M/M/1/\infty$, $M/D/1/\infty$ і $fBM/D/1/\infty$ при $H = 0{,}7$

З приведеного на рис. 8.7 графіка залежності середньої кількості вимог у системі N для систем $M/M/1/\infty$, $M/D/1/\infty$ і $fBM/D/1/\infty$ витікає аналогічний висновок, що при навантаженні $\rho > 0{,}3$ в системі із вхідним потоком вимог, що має характер самоподібного процесу, буде більше вимог, чим при відсутності самоподібності.

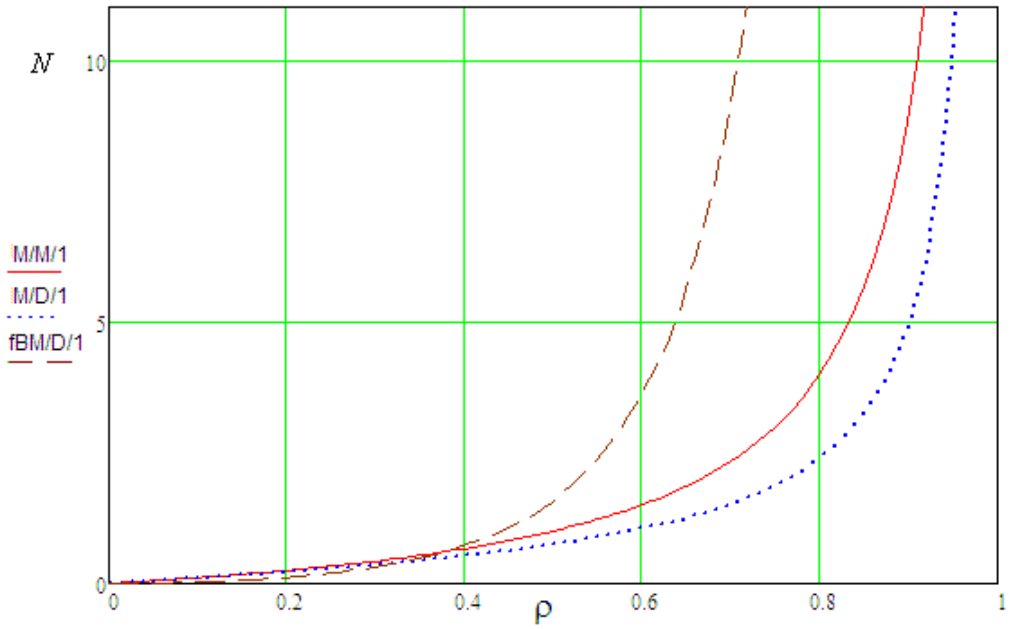


Рисунок 8.7 – Залежність середньої кількості вимог в системі N для моделей $M/M/1/\infty$, $M/D/1/\infty$ і $fBM/D/1/\infty$ при $H = 0{,}7$

До подібного висновку можна дійти й після аналізу графіка залежності середньої довжини черги Q від ρ у тих же системах, наведеному на рис. 8.8.

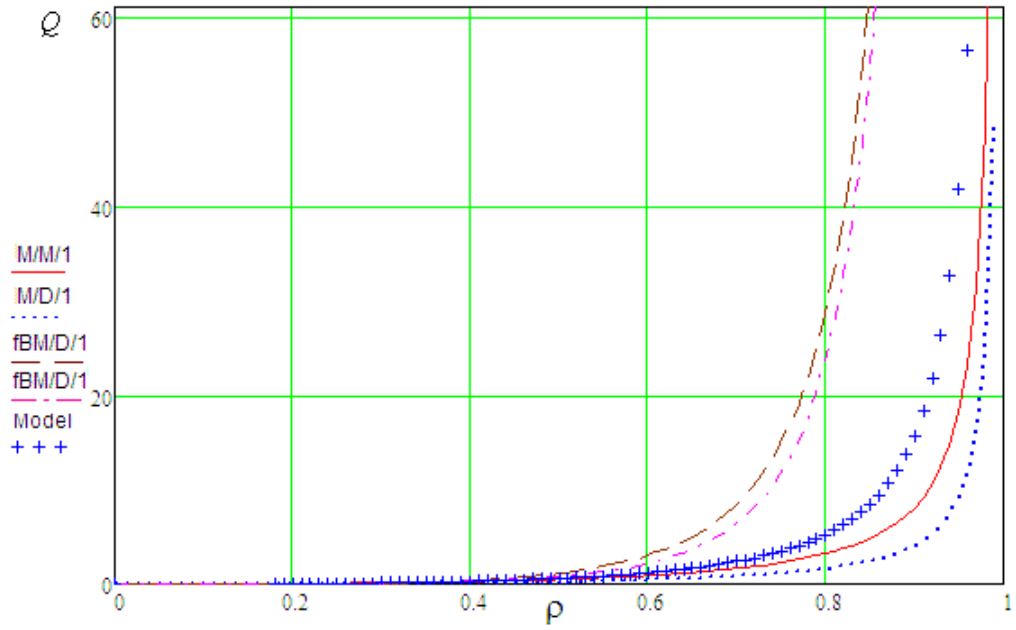


Рисунок 8.8 – Залежність середньої довжини черги Q для моделей M/M/1, $M/D/1/\infty$ і $fBM/D/1/\infty$ при $H = 0{,}7$

На всіх графіках можна бачити, що з ростом інтенсивності навантаження ρ погіршуються характеристики якості обслуговування T, N і Q, але ще більш істотно вони погіршуються за наявності властивостей самоподібності у вхідному потоці вимог. Результат у вигляді виразів (8.12), (8.13) і (8.14) показує ступінь цього впливу залежно від величини коефіцієнта H.

Однак, для даного результату (формули Нороса), отриманого в припущенні постійної тривалості обслуговування, при $H = 0{,}5$ вирази (8.12), (8.13) і (8.14) дають оцінки параметрів якості обслуговування, характерні для пуассонівського потоку з експонентним законом розподілу тривалості обслуговування, а не регулярного (модель $M/M/1/\infty$). Установити ступінь точності даного результату можна за допомогою імітаційного моделювання.

При імітаційному моделюванні достатньо оцінити тільки один з параметрів, наприклад N, оскільки параметри Q і T пов'язані з N відомими функціональними залежностями. Результати імітаційного моделювання СМО типу $fBM/D/1/\infty$ при $H = 0{,}7$ наведені на рис. 8.8 і показані лінією, у вигляді знаків „+”.

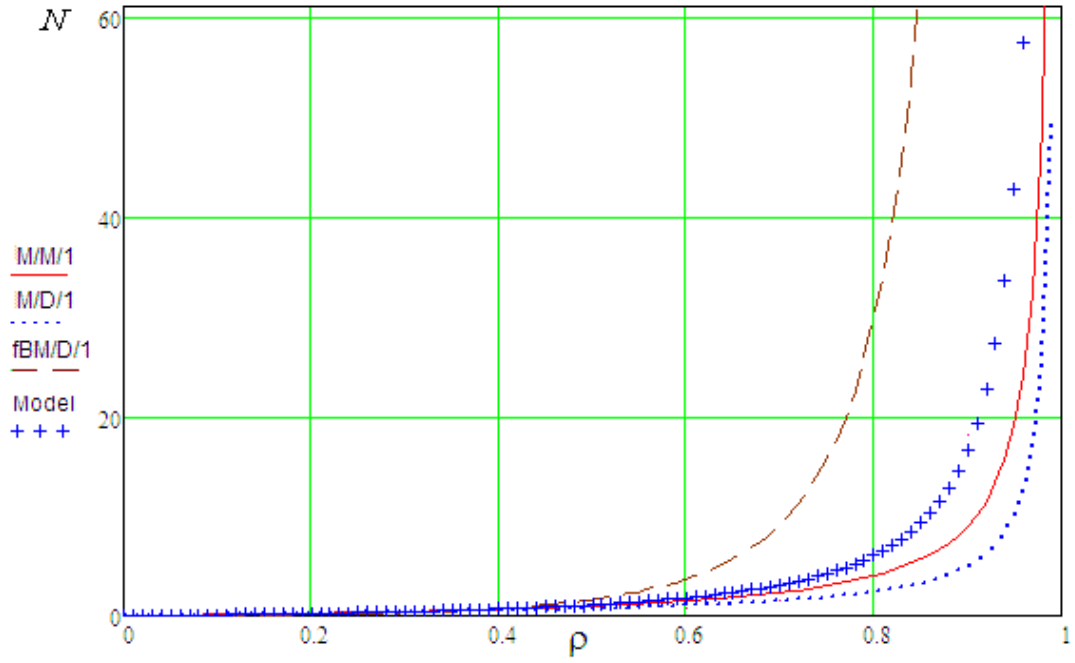


Рисунок 8.8 – Моделювання середньої кількості вимог у системі N для моделі $fBM/D/1/\infty$ при $H = 0{,}7$

Результати моделювання підтверджують висновок про те, що за наявності властивості самоподібності у вхідному потоці вимог з ростом інтенсивності навантаження ρ погіршуються характеристики якості обслуговування, але не настільки, як це передбачується за формулою Нороса. Розбіжність результатів моделювання й оцінок, отриманих за формулами (8.12), (8.13) і (8.14) може становити сотні відсотків. Очевидно, що оцінка Нороса значно завищена, а це вимагає знаходження більше точного розв’язання.

У зв'язку із цим найбільш перспективним є метод оцінки параметрів якості обслуговування самоподібного трафіка, у якому запропоновано використання методів розрахунку відомих класичних розподілів, ентропія яких найбільш близька до ентропії станів системи за умов обслуговування самоподібного трафіка [13]. При цьому стає можливим розрахунок характеристик *QoS* у моделях з самоподібним трафіком за будь-якого закону розподілу тривалості обслуговування за формулою Полачека-Хінчина.

Випадковий процес (ВП) надходження пакетів до СРІ на обслуговування, утворює потік пакетів (трафік), який характеризується певним законом розподілу, що встановлює зв'язок між значенням випадкової величини (кількістю пакетів) і імовірністю появи цього значення. У більшості випадків для розрахунку параметрів *QoS* досить знати про закон розподілу тільки деякі його числові характеристики – моменти розподілу різних порядків. Для розрахунку в умовах пуассонівського розподілу достатньо математичного сподівання Λ, а для нормального розподілу – необхідні математичне сподівання Λ і дисперсія D.

Основні характеристики випадкового процесу Λ і D, хоча й досить важливі, у той же час не є вичерпними, а іноді й марними для прогнозування значення випадкової величини. Іноді ВП характеризуються однаковими значеннями Λ і D, але внутрішня структура цих процесів різна. Одні можуть мати плавно мінливі реалізації, а інші – яскраво виражену коливальну структуру при стрибкоподібній зміні окремих значень випадкової величини (наприклад, різке зростання кількості пакетів в мережі, що призводить до „пачковості" трафіка). Для „плавних" процесів характерна велика передбачуваність реалізацій, а для „пачкових" – дуже мала імовірнісна залежність між двома випадковими величинами ВП. У таких випадках закон розподілу, що характеризує ВП, несе в собі деяку невизначеність і дозволяє з більшою або меншою надійністю прогнозувати значення випадкової величини.

Таким чином, використовувані імовірнісні закони розподілу, що описують трафік у пакетних мережах, не дають такої кількісної оцінки невизначеності стану системи масового обслуговування, як ентропія розподілу:

$$H(m) = -\sum_{j=1}^{m} P_j \log P_j .$$

Ентропія не залежить від значень, яких набуває випадкова величина, а тільки від їхніх ймовірностей.

Оцінки параметрів якості обслуговування самоподібного трафіка можлива ентропійним методом, який зводиться до використання методів розрахунку відомих розподілів, ентропія яких співпадає або найбільш близька до ентропії станів системи при обслуговуванні самоподібного трафіка [13].

Результатами моделювання встановлено, що в тих точках, де однакова ентропія розподілу станів системи, однакові й досліджувані параметри якості обслуговування, такі як середня довжина черги Q та середня тривалість очікування вимог W. Наприклад, для моделей $M/M/1/\infty$ і $fBM/D/1/\infty$ при коефіцієнті Херста $H = 0{,}8$ та $\rho = 0{,}6$ ентропії розподілів станів систему досить близькі і дорівнюють 1,683 та 1,719 відповідно. При цьому для моделі $fBM/D/1/\infty$ середня довжина черги $Q = 0{,}982$ і середня тривалість очікування всіх вимог $W = 1{,}611$, що перевищує відповідні значення для моделі $M/M/1/\infty$ усього на 3 % (на стільки ж відмінність і значень ентропії). Такий же збіг основних параметрів якості обслуговування СМО з чергою спостерігається в усіх інших точках, для яких однакові значення ентропії розподілу станів системи, незалежно від закону розподілу тривалості обслуговування. Тому для цих розрахунків можна застосовувати формули Поллачека-Хінчина з табл. 7.1 моделі $M/G/1/\infty$.

Алгоритм застосування ентропійного методу розрахунку *QoS* такий:

1. Для встановленого закону розподілу станів системи визначається ентропія розподілу H_{fBM} (за відомими формулами).

2. Зміною коефіцієнта варіації v_x для моделі, наприклад $M/HM/1/\infty$, досягається збіг значень ентропії $H_{M/HM/1} = H_{fBM}$.

3. За допомогою знайденого коефіцієнта варіації v_x визначається середня кількість пакетів у системі N за формулою Поллачека-Хінчина (табл. 7.1).

4. Через відомі співвідношення визначаються інші характеристики *QoS*.

Таким чином, розрахунок параметрів *QoS* в моделі з самоподібним трафіком за будь-якого закону розподілу тривалості обслуговування здійснимий. Необхідною умовою такого розрахунку є визначення ентропії розподілу станів системи.

Доказ властивості самоподібності реального трафіка виконується методом абсолютних моментів. В якості значень випадкового процесу розглядається кількість вимог, що надходить у СМО за одиницю часу. Вихідну послідовність кількості вимог довжиною N розділимо на блоки довжиною m (окремі агреговані процеси розміром m). На непересічних часових інтервалах, тобто на границях кожного блоку k-та послідовність має середнє значення:

$$X_k^{(m)} = \frac{1}{m}\sum_{j=1}^{m} X_{(k-1)m+1}, \qquad k = 1, 2, 3 \ldots, [N / m]$$

Після розрахунку середнього значення $\overline{X}$ для всієї послідовності, потім для кожного блоку k розрахуємо дисперсію D_k:

$$D_k^{(m)} = \frac{1}{N / m}\sum_{j=1}^{m} (X_k^{(m)} - \overline{X})^2 .$$

Для самоподібного процесу дисперсія агрегованих процесів повинна убувати повільніше, ніж величина, зворотна розмірові вибірки m [1]. Для виявлення цієї властивості побудуємо дисперсійно-часовий графік залежності дисперсій агрегованих процесів від ступеня агрегування m. Оскільки Херстом було показано, що:

$$\log\left(\frac{\max D - \min D}{D_k^{(m)}}\right) \approx H \log\left(\frac{N}{2}\right),$$

то графік цієї залежності будуємо теж у логарифмічному масштабі. Вираз в лівій частині цього рівняння $\dfrac{\max D - \min D}{D_k^{(m)}}$ називається *R*/*S* статистикою або нормованим розмахом. З отриманого графіка визначаємо коефіцієнт β, як тангенс кута нахилу апроксимуючої кривої до побудованої залежності. Дана апроксимація виконується методом мінімального середньоквадратичного відхилення від експериментальних даних. Коефіцієнт β ($0 < \beta < 1$), що задає асимптотичні властивості характеристик самоподібного випадкового процесу пов'язаний з параметром Херста наступним співвідношенням [1]:

$$H = 1 - \frac{\beta}{2}.$$

Для реальних процесів, що не мають властивості самоподоби, $H < 0{,}5$, а для самоподібних процесів з довгостроковою залежністю цей параметр змінюється в межах 0,65–0,8 (процес має тривалу пам'ять).

8.5 Система з необмеженою чергою *G*/*M*/1/∞

У мультисервісних пакетних мережах зв'язку вхідні інформаційні потоки можуть мати постійну (*CBR*), змінну (*VBR*) і змішану бітову швидкість, від чого математична модель потоку може бути від найпростішої пуассонівської до складної моделі фрактальних процесів (самоподібний трафік). Закон розподілу інтервалу часу між вимогами в цих потоках може бути довільний і тому в узагальненій моделі резонно досліджувати узагальнений (*G*) вид розподілу випадкової величини цього інтервалу. Довжина пакетів кожної зі служб загальної для них мультисервісної мережі (отже, і тривалість обслуговування) може бути різною – для одних служб постійною, а для інших – змінною. У такому випадку також бажано досліджувати загальний вид розподілу випадкової величини тривалості обслуговування.

Системи з чергою типу *G*/*G*/*m*/∞ є найважливішими моделями, що розглядаються у теорії телетрафіка. При їх дослідженні застосовувалися різні методи й отримані численні наближені результати. Окремий випадок – система з чергою й одним сервером (*m* = 1) розглядається, наприклад в [7], де отримано тільки приблизні результати. Однак дотепер у загальному випадку не існує простих і точних, безпосередньо застосовуваних на практиці, формул для розрахунку характеристик *QoS* у стаціонарному режимі.

У багатьох випадках односерверної системи хорошим наближенням є експонентна функція розподілу тривалості обслуговування вимог. Тому визначимо характеристик *QoS* для моделі *G*/*M*/1/∞ [14].

Коефіцієнт використання серверів системи ρ (*utilization factor*) визначається як відношення інтенсивності вхідного потоку вимог λ до інтенсивності обслуговування μ. В *m*-серверній системі всі сервери забезпечать інтенсивність обслуговування $m\mu = m\frac{1}{\bar{x}}$. Отже, в *m*-серверній системі $\rho = \frac{\lambda\bar{x}}{m}$. В односерверній системі ρ в *m* разів більше і співпадає з інтенсивністю навантаження Λ (5.3) або з інтенсивністю вхідного потоку вимог λ, якщо $\bar{x}$ є однією умовною одиницею часу обслуговування. За умови $0 \le \rho < 1$ процес в системі ергодичним і стаціонарний розподіл імовірностей станів системи існує.

Для будь-якої односерверної системи $\rho = 1 - p_0$, де p_0 – імовірність вільності системи (стан системи p_0 – зайнято 0 серверів). Отже, ρ – чисельно збігається з імовірністю зайнятості системи $P_{зн}$ (стан системи p_1 – зайнятий єдиний сервер, відповідає частці часу зайнятості серверу). З урахуванням вимог, що знаходяться в черзі, у стаціонарному режимі існує стаціонарний розподіл кількості вимог у системі p_k, де *k* – кількість вимог. Цей розподіл не залежить від моменту прибуття вимоги до системи.

За пуассонівського потоку імовірність очікування $P_{w>0}$ збігається з імовірністю зайнятості системи $P_{зн}$ (див. 7.47). Для односерверної моделі *M*/*G*/1/∞ з довільним розподілом тривалості обслуговування дані імовірності однакові і $P_{w>0} = \rho$. Однак для моделі *G*/*M*/1/∞ такої рівності нема, тобто за цим параметром моделі не інваріантні. В [5, с. 272] показано, що система *G*/*M*/1∞ призводить до геометричного розподілу кількості вимог у системі в моменти

надходження нових вимог r_k, де k – кількість вимог. Розподіл p_k відрізняється від розподілу r_k тим, що $p_0 = 1 - P_{зн}$ (або $p_0 = 1 - \rho$), у той час як $r_0 = 1 - P_{w>0}$. Для системи $M/G/1/\infty$ виконується рівняння $p_k = r_k$.

Вимога повинна очікувати обслуговування з імовірністю $P_{w>0} = 1 - r_0$. Тому при експонентному розподілі тривалості обслуговування безумовний розподіл тривалості очікування визначиться так:

$$W(t) = 1 - P_{w>0} e^{-\mu(1-P_{w>0})t}, \quad \text{при} \quad t \geq 0. \qquad (8.15)$$

З цього можна розрахувати середній час очікування у системі W і всі інші параметри якості обслуговування:

$$Đ_{çí} = \rho; \qquad Đ_{w>0} = 1 - \frac{\rho}{N};$$

$$W = \frac{Đ_{w>0}}{1 - Đ_{w>0}}; \qquad Q = \frac{\rho \cdot Đ_{w>0}}{1 - Đ_{w>0}};$$

$$t_q = T = \frac{1}{1 - P_{w>0}}; \qquad N = \frac{\rho}{1 - P_{w>0}}.$$

Оскільки $\mu = 1$, тобто середня тривалість обслуговування $\bar{x} = 1/\mu$ приймається за умовну одиницю часу, то W, t_q і T оцінюються в одиницях середньої тривалості обслуговування.

В табл. 8.1 наведено залежності між основними параметрами *QoS* для моделі $G/M/1/\infty$.

Таблиця 8.1 – Залежності між параметрами *QoS* у системі $G/M/1/\infty$

Хар-ка *QoS*	Характеристика *QoS*				
	Q	W	t_q	N	
$P_{зн}$	ρ	ρ	ρ	ρ	
$P_{w>0}$	$\frac{Q}{\rho + Q}$	$\frac{W}{1+W}$	$1 - \frac{1}{t_q}$	$1 - \frac{\rho}{N}$	$\boxed{\frac{Q}{N}}$
Q	–	$\rho \cdot W$	$\rho \cdot (t_q - 1)$	$N - \rho$	$N \cdot P_{w>0}$
W	$\frac{Q}{\rho}$	–	$\boxed{t_q - 1}$	$\frac{N}{\rho} - 1$	
t_q	$1 + \frac{Q}{\rho}$	$1 + W$	–	$\frac{N}{\rho}$	
N	$\rho + Q$	$\rho \cdot (1 + W)$	$t_q \rho$	–	$\frac{Q}{Đ_{w>0}}$
T	$1 + \frac{Q}{\rho}$	$1 + W$	t_q	$\frac{N}{\rho}$	

З таблиці випливає, що за наявності лише одного відомого параметра (наприклад, відомий параметр Q, W, t_q або N) всі інші параметри розраховуються через наведені в таблиці співвідношення.

Для моделі $G/M/1/\infty$ виявлено й перевірено за допомогою імітаційного моделювання важливі властивості односерверної системи, які виконуються тільки за експонентної тривалості обслуговування (виділені рамкою в табл. 8.1).

По-перше, середній час очікування в черзі t_q чисельно збігається із середнім часом перебування вимоги в системі T. Це означає, що середній час очікування в системі W менше середнього часу очікування в черзі t_q на величину середньої тривалості обслуговування (одиниця часу $\bar{x}$):

$$W = t_q - 1. \tag{8.16}$$

По-друге, імовірність очікування можна визначити як

$$P_{w>0} = \frac{Q}{N}. \tag{8.17}$$

За визначенням (6.2) імовірність очікування $P_{w>0}$ – це відношення кількості затриманих вимог до загальної кількості вимог, що надійшли, але як у даному випадку, частку вимог, що очікують, можна визначити і як відношення середньої кількості вимог у черзі Q до середньої кількості вимог у системі N. Звідси з урахуванням формули Літтла випливають кілька важливих співвідношень між параметрами QoS, справедливих вже для моделей $G/G/1/\infty$ та $G/G/m/\infty$ (у правому стовпчику табл. 8.1).

Для системи $GI/G/1/\infty$ Маршаллом запропонована наближена верхня (*High*) оцінка середнього стаціонарного часу очікування W_H [1]:

$$W_H \le \frac{\sigma_z^2 + \sigma_x^2}{2(\bar{z} - \bar{x})}.$$

Нижня (*Low*) оцінка середнього стаціонарного часу очікування W_L запропонована Штойяном в [7]:

$$W_L \ge \frac{\sigma_x^2}{2(\bar{z} - \bar{x})} - \frac{\bar{x}}{2}.$$

Аналіз системи типу $GI/G/1/\infty$, в якій враховується вид функції розподілу інтервалу часу між вимогами вхідного потоку та тривалості обслуговування вимог, за допомогою наведених наближених верхньої та нижньої оцінок іноді дає більш точний результат порівняно з максимальними оцінками, що отримуються при використанні марковських моделей.

8.6 Система з необмеженою чергою *G*/*D*/1/∞

У мультисервісних мережах зв'язку трафік має складну структуру, яка вимагає для свого опису суттєвого ускладнення математичної моделі потоку вимог. Закон розподілу інтервалу часу між вимогами в цих потоках може бути довільний і тому в моделі, що обслуговує мультисервісний трафік, резонно досліджувати узагальнений (*G*) вид розподілу випадкової величини цього інтервалу.

У мультисервісних пакетних мережах вимогою на обслуговування можна вважати кожний окремий пакет інформації. Довжина пакетів може бути незмінною за постійної тривалості їх обслуговування, наприклад, в технології *ATM*. В такому разі в дослідженнях можна обмежитися тільки детермінованим законом розподілу тривалості обслуговування (*D*). Однак і за змінної довжини пакетів при аналізі пакетних комутаторів слід враховувати наявність у кожного пакета заголовку фіксованої довжини, що вимагає обліку в тривалості обслуговування деякого постійного доданку, навіть якщо розподіл довжин пакетів є, наприклад, експонентним. Отже, розробка методу розрахунку основних характеристик якості обслуговування вимог у системах**,** представлених моделлю *G*/*D*/1/∞ також є актуальною.

У п. 8.7 показано, що для моделі *G*/*M*/1/∞ з геометричним розподілом r_k (розподіл кількості вимог у системі в моменти надходження нових вимог) за експонентного закону розподілу тривалості обслуговування середній час очікування в системі *W* визначається як

$$W = \frac{Đ_{w>0}}{1 - Đ_{w>0}}. \qquad (8.18)$$

З урахуванням цього результату та формули Літтла в табл. 8.2 наведено формули розрахунку характеристик якості обслуговування для моделі *G*/*M*/1/∞ і відомий результат для моделі *M*/*D*/1/∞, що випливає з формули Поллачека-Хінчина (7.66).

У формули розрахунку характеристик *QoS* моделі *G*/*G*/1/∞ можна ввести величину $t_q - W$ (табл. 8.2). Правильність цього кроку легко перевірити підстановкою в кожну з них справедливого за визначенням для будь-якої моделі рівняння:

$$P_{w>0} = \frac{W}{t_q}.$$

Для розрахунку характеристик *QoS* моделі *G*/*D*/1/∞ можна використати виявлену в п. 8.7 властивість моделі *G*/*M*/1/∞, а саме: $t_q = T$. Оскільки з визначення середнього часу перебування вимоги в будь-якій системі (в тому числі й в моделі *G*/*G*/1/∞) випливає, що

$$T = W + 1, \qquad (8.19)$$

то з урахуванням наведеної властивості моделі $G/M/1/\infty$

$$t_q - W = 1. \tag{8.20}$$

Видно, що при виконанні умови (8.20), кожна з формул розрахунку характеристик *QoS* моделі $G/G/1/\infty$ збігається з аналогічною формулою моделі $G/M/1/\infty$. Також відомо, що для моделі $M/D/1/\infty$:

$$t_q - W = 0{,}5. \tag{8.21}$$

Тому при виконанні умови (8.21) разом з умовою $P_{w>0} = \rho$, властивої за пуассонівського потоку, кожна з формул моделі $G/G/1/\infty$ перетвориться у відповідну формулу моделі $M/D/1/\infty$ (стовпчики 4 та 2 табл. 8.2).

Таблиця 8.2 – Основні параметри якості обслуговування

Хар-ка *QoS*	Модель			
	M/D/1/∞	G/M/1/∞	G/G/1/∞	$G^*/D/1/\infty$
			точно	*приблизно*
$P_{зн}$	ρ	ρ	ρ, $\dfrac{Q}{t_q P_{w>0}}$	
$P_{w>0}$	ρ	$1-\dfrac{\rho}{N}$	$\dfrac{W}{t_q}$, $\dfrac{Q}{t_q\rho}$	
Q	$\dfrac{\rho^2}{2(1-\rho)}$	$\dfrac{\rho P_{w>0}}{1-P_{w>0}}$	$\dfrac{\rho P_{w>0}}{1-P_{w>0}}(t_q-W)$	$\dfrac{\rho P_{w>0}}{2(1-P_{w>0})}$
W	$\dfrac{\rho}{2(1-\rho)}$	$\dfrac{P_{w>0}}{1-P_{w>0}}$	$\dfrac{P_{w>0}}{1-P_{w>0}}(t_q-W)$	$\dfrac{P_{w>0}}{2(1-P_{w>0})}$
t_q	$\dfrac{1}{2(1-\rho)}$	$\dfrac{1}{1-P_{w>0}}$	$\dfrac{1}{1-P_{w>0}}(t_q-W)$	$\dfrac{1}{2(1-P_{w>0})}$
N	$\rho+\dfrac{\rho^2}{2(1-\rho)}$	$\dfrac{\rho}{1-P_{w>0}}$	$\rho+\dfrac{\rho P_{w>0}}{1-P_{w>0}}(t_q-W)$	$\dfrac{\rho(2-P_{w>0})}{2(1-P_{w>0})}$
T	$1+\dfrac{\rho}{2(1-\rho)}$	$\dfrac{1}{1-P_{w>0}}$	$1+\dfrac{P_{w>0}}{1-P_{w>0}}(t_q-W)$	$\dfrac{2-P_{w>0}}{2(1-P_{w>0})}$

Якщо припустити, що умова (8.21) здійсненна і при $P_{w>0} \neq \rho$, то з тих же формул моделі $G/G/1/\infty$ можна отримати відповідні формули розрахунку характеристик *QoS* моделі $G/D/1/\infty$ (стовпчик 5 табл. 8.2). Але оскільки, як показує імітаційне моделювання, отримані у такий спосіб формули дають не зовсім точні і різні результати для різних вхідних потоків, то краще назвати таку модель з *умовно загальним* розподілом випадкової величини інтервалу часу між вимогами і позначити $G^*/D/1/\infty$.

Для моделі $G^*/D/1/\infty$ різниця (8.21) не завжди дорівнює 0,5. За результатами імітаційного моделювання, виконаного за допомогою алгоритму [9], встановлено, що для потоків, в яких коефіцієнт варіації тривалості інтервалу часу між вимогами $v_z < 1$ (більш вирівняний потік порівняно з пуассонівським) $t_q - W < 0{,}5$. Наприклад, за рівномірного розподілу інтервалу часу між вимогами (модель $U/D/1/\infty$) $v_z = 0{,}58$. Для пуассонівського потоку з експонентним розподілом зазначеного інтервалу при $v_z \equiv 1$ або з розподілом Вейбулла при підібраному коефіцієнті форми розподілу так, щоб $v_z = 1$, умова (8.20) виконується. Для гіперекспонентного, логарифмічно нормального або розподілів Парето та Вейбулла при параметрах розподілів, що забезпечують значення v_z від одиниць до декількох десятків (дуже нерівномірні потоки), різниця t_q і W завжди трохи більше 0,5.

У випадку $t_q - W < 0{,}5$ результати розрахунків трохи вище результатів моделювання, при $t_q - W = 0{,}5$ – вони цілком збігаються, а при $t_q - W > 0{,}5$ – результати розрахунків виявляються трохи заниженими. При цьому найбільша похибка обчислень за формулами моделі $G^*/D/1/\infty$ порівняно з результатами моделювання спостерігається у випадку самоподібного потоку вимог, згенерованому за розподілом Парето. Наприклад, при завданні параметра форми розподілу $a = 1{,}3$ (відповідає коефіцієнтові самоподібності трафіка $H = 0{,}85$ [13] і $v_z \geq 30$) похибка обчислень не перевищує 20 % для всіх параметрів *QoS*. При збільшенні a і пов'язаним з цим зменшенням H, наприклад, до значень 1,9 і 0,55 відповідно (ще самоподібний процес), дана похибка не перевищує вже 10 %.

Слід зазначити, що точність розрахунків декілька підвищується при зменшенні інтенсивності навантаження ρ і в багатьох випадках при $\rho < 0{,}5$ похибка розрахунків не перевищує 5 %.

Важливість даного методу мотивується істотним ускладненням моделей трафіка в сучасних мережах зв'язку і відсутністю адекватних цьому ускладненню методів розрахунку параметрів якості обслуговування. Особливо корисною даний метод може бути у випадку обслуговування самоподібного трафіка, оскільки його точність незрівнянно вище точності існуючих методів розрахунку.

9 ІМІТАЦІЙНЕ МОДЕЛЮВАННЯ СМО

9.1 Метод статистичних випробувань

Математичні моделі функціонування СМО, що розглянуто в розд. 7 та 8, орієнтовані на можливість отримання в тій або іншій формі аналітичних рішень для обумовлених характеристик СМО. Можливість отримання таких рішень істотно обмежується видом вхідних потоків, законом розподілу тривалості обслуговування та структурою СМО. Використання методів моделювання дозволяє помітно послабити ті обмеження, які відносяться до виду вхідних потоків вимог. Таким чином, основним інструментарієм дослідження задач, що не піддаються аналітичним і чисельним методам, є імітаційне моделювання.

Метод статистичних випробувань (метод Монте-Карло) заснований на моделюванні досліджуваного процесу. Основний його принцип – побудова такої штучної ймовірнісної моделі, параметри якої являють собою рішення поставленого завдання. Якщо така модель побудована, то користуючись методами математичної статистики, можна оцінити невідомі параметри, тобто знайти наближене рішення завдання.

Для моделювання процесу на ЕОМ необхідно перетворити його математичну модель СМО у спеціальний моделюючий алгоритм. При статистичному моделюванні реалізація моделюючого алгоритму є імітацією елементарних явищ, що складають досліджуваний процес, зі збереженням їхньої логічної структури, послідовності протікання в часі й особливо характеру й складу інформації про стани процесу. Можна вказати на наявну аналогію з дослідженням процесів в дійсності. В тому й у іншому випадках є можливість використати для рішення поставлених завдань будь-яку інформацію про стани процесу, якщо тільки вона доступна відповідній реєстрації.

Звідси витікає, що структура моделюючого алгоритму може слабко залежати від сукупності шуканих величин, а визначається головним чином побудовою математичної моделі. При дослідженні СМО метод дозволяє враховувати: вид всіх потоків подій; нестаціонарність потоків у системі; різного роду обмеження (наприклад, обмеження на час перебування в СМО); вплив стану СМО на інтенсивність потоків тощо.

Разом з тим результати моделювання, як і при будь-якому чисельному методі, завжди носять винятковий характер. Для отримання якісних висновків потрібно дослідити велику кількість СМО, причому й у цьому випадку якісний результат може носити винятковий характер. У ряді випадків може позначатися принципово обмежена точність одержуваних результатів, особливо, якщо використовується ЕОМ з невеликою довжиною розрядної сітки. Необхідність наявності досить потужної універсальної ЕОМ також може бути істотною перешкодою до використання методу. Тому, незважаючи на безсумнівні достоїнства, метод імітаційного моделювання не може замінити аналітичних методів дослідження СМО, а є їхнім доповненням.

9.2 Синтез моделюючих алгоритмів

Основними принципами побудови моделюючих алгоритмів є:

– принцип, що дозволяє визначати послідовні стани системи через деякі інтервали часу (принцип Δt);
– принцип послідовного проведення вимог.

Перший принцип полягає в тому, що стани СМО визначаються для моментів часу, розділених однаковими інтервалами часу Δt. При цьому виявляється, що для більшості таких моментів стан СМО не змінюється в порівнянні з попереднім моментом, тобто весь проміжок часу Δt є нецікавим для дослідження. Інтерес представляють особливі стани, що відповідають моментам надходження вимог у систему й моментам виходу вимог із СМО. При збільшенні Δt росте ймовірність влучення в інтервал особливого стану й знижується точність і вірогідність моделювання (за рахунок можливого влучення в один інтервал двох і більше особливих станів). Зменшення Δt приводить до різкого росту часу моделювання.

При моделюванні процесів обробки вимог у СМО зручніше будувати моделюючі алгоритми за принципом послідовного проведення вимогк. Ідея його полягає в послідовному відтворенні історії окремих вимог у порядку надходження їх у систему: алгоритм звертається до відомостей про інші вимоги лише в тому випадку, якщо це необхідно для рішення питання про подальший порядок обслуговування даної заявки. Такого роду моделюючі алгоритми досить економні, не вимагають спеціальних заходів для обліку особливих станів системи, однак вони мають досить складну логічну структуру й не завжди доступні для побудови.

При моделюванні доводиться вирішувати наступні завдання:

1. Отримання послідовності випадкових чисел, рівномірно розподілених в інтервалі [0, 1].

2. Перетворення чисел отриманої послідовності у величини, що мають деяку певну функцію розподілу.

3. Побудова логічної схеми алгоритму, що враховує особливості роботи СМО, що моделюється.

4. Побудова алгоритмів фіксації станів СМО й обробки результатів моделювання.

5. Вибір кількості реалізацій у відповідності до заданої точності обумовлених характеристик СМО.

6. Моделювання на ЕОМ.

7. Остаточна обробка й аналіз результатів.

Для моделювання потоків подій необхідно користуватися алгоритмами, за якими б вироблялися випадкові величини, що відповідають тривалостям інтервалів часу між сусідніми подіями необхідного виду потоку. Можлива нестаціонарність потоків може бути врахована уведенням в алгоритм формування випадкових величин залежності від поточного часу надходження (або початку обслуговування) певної вимоги. Ці випадкові величини повинні бути розподілені за певним законом, тобто має місце завдання отримання

послідовності випадкових величин з певною функцією розподілу. Звичайно спершу одержують послідовність чисел, рівномірно розподілених в інтервалі [0, 1], а потім цю послідовність перетворюють у послідовність випадкових величин з необхідною функцією розподілу. Існують наступні способи отримання випадкових чисел з рівномірним розподілом:

– алгоритмічне отримання послідовності випадкових (псевдовипадкових) чисел.
– початкове занесення в пам'ять ЕОМ таблиці випадкових чисел, що отримані з натурного експерименту випробовування СМО.

За першого способу послідовність псевдовипадкових чисел виробляється самою ЕОМ за допомогою спеціальних алгоритмів. При цьому пред'являються наступні вимоги.

1. Одержувана послідовність чисел повинна мати статистичну структуру, у достатньому ступені близьку до структури рівномірної сукупності.

2. Кількість операцій, яка необхідна для вироблення кожного числа послідовності, не повинна бути занадто великою.

Послідовність псевдовипадкових чисел характеризується періодичністю повторення. Бажано так підбирати „випадкові" числа x_i (i = 1, 2, ..., m), щоб довжина періоду повторення була найбільшою.

Для обґрунтування „випадковості" послідовності псевдовипадкових чисел необхідно використати систему наступних тестів:

1. Перевірка частот повторення чисел з різних інтервалів;
2. Перевірка пар;
3. Перевірка інтервалів;
4. Перевірка комбінацій.

Множина псевдовипадкових чисел, що задовольняє всім цим тестам, називається локально випадковою. Всі згадані вище тести характеризуються однією загальною властивістю: випробовувані псевдовипадкові числа (або розряди в них) класифікуються за деякими ознаками (різним для кожного тесту) і отриманий емпіричний розподіл порівнюється з теоретичним. Для порівняння використовуються звичайні статистичні критерії.

Маючи випадкові числа, що рівномірно розподілені в інтервалі [0, 1] можна отримати послідовність випадкових величин ξ із заданою густиною розподілу $f(x)$, якщо вирішити рівняння

$$\int_a^{\xi} f(x)dx = U,$$

де U – число з вихідної рівномірної послідовності; a – мінімально можливе число з послідовності випадкових величин із густиною розподілу $f(x)$

Найбільш поширеними є три підходи до статистичного моделювання систем масового обслуговування:

– моделювання марковського процесу;
– моделювання напівмарковського процесу;
– моделювання реального процесу обслуговування.

9.3 Моделювання марковського процесу

До СМО надходить потік вимог з функцією розподілу інтервалу часу між двома послідовними вимогами $A(z)$. Тривалість обслуговування є випадковою величиною з функцією розподілу $B(x)$. Необхідно визначити параметри *QoS*.

Моделювання марковського процесу можливе тільки за експонентних законів, де функція $A(z) = 1 - e^{-\lambda z}$ та $B(x) = 1 - e^{-\mu x}$. При цьому функціонування m-серверної системи з втратами описується марковським процесом $x(t)$, де його стани набувають значеннь 0, 1, 2, ..., m. Майже будь-яка траєкторія марковського процесу влаштована таким чином, що час перебування ξ_i у i-му стані має експонентний розподіл з параметром $\lambda + i$, тобто

$$P\{\xi_i > t\} = e^{-(\lambda+i)t}. \tag{9.1}$$

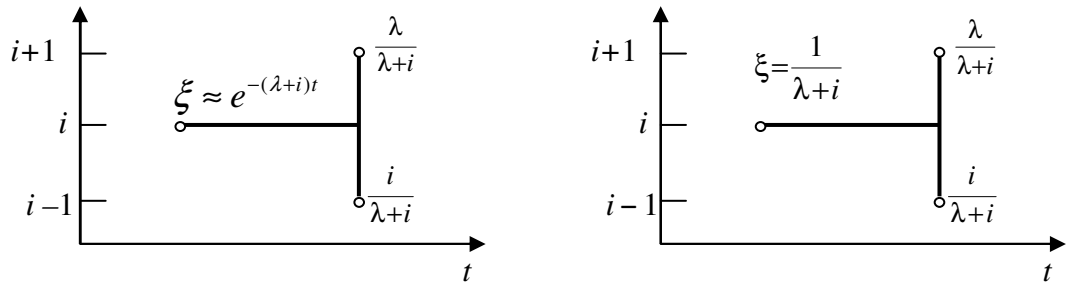
У момент виходу із цього стану здійснюється перехід у стан $i+1$ з імовірністю

$$p_i = \frac{\lambda}{\lambda + i} \tag{9.2}$$

або в стані $i - 1$ з ймовірністю

$$q_i = \frac{i}{\lambda + i} \tag{9.3}$$

При $i = m$ можливий перехід тільки в стан $m - 1$ з ймовірністю, що дорівнює 1 (рис. 3.23а). Звідси витікає, що при відтворенні роботи системи необхідно моделювати два випадкових механізми: час перебування у i-му стані згідно (9.1) і ймовірності переходів згідно (9.2) і (9.3), що показано на рис. 9.1.



а) б)

Рисунок 9.1 – Траєкторія марковского процесу з випадковим (а) і детермінованим (б) часом перебування в станах

Отже, для моделювання необхідні: програма реалізації випадкових величин, що розподілена за експонентним законом (9.1); програма генерування рівномірно розподілених випадкових чисел для вибору напрямку переходів по (9.2) і (9.3); комірки пам'яті для фіксації поточного часу системи, що міняється стрибками відповідно до випадкових періодів перебування в послідовних станах, і n розрядів (а не m комірок) для запису номерів зайнятих серверів або одна комірка для запису кількості зайнятих серверів. Такий підхід дає суттєву економію машинної пам'яті, що особливо важливо при моделюванні складних систем.

9.4 Моделювання вкладеного ланцюга Маркова

З аналізу властивостей траєкторії марковського процесу $x(t)$ виливає, що при обчисленні оцінки середньої кількості зайнятих серверів випадкові періоди перебування в окремих станах можна замінити їхніми середніми значеннями. Розглянемо траєкторію процесу $x(t)$ до моменту T_N, тобто до моменту N-го переходу. Необхідно обчислити

$$M_j(T_N) = \frac{1}{T_N} \int_0^{T_N} f[x(t)]dt . \qquad (9.4)$$

У нашому випадку $f[x(t)] = i$, якщо $x(t) = i$, $i = 0, 1, ..., m$. Функціонал (9.4) являє собою випадкову величину, що залежить від вибіркової траєкторії процесу $x(t)$. При заміні ξ_i випадкового часу перебування в стані i-го середнім значенням $1 / (\lambda + i)$ (рис. 9.1б) середнє значення функціонала (9.4) зберігається незмінним, а його дисперсія істотно зменшується, тому що усувається один із двох випадкових механізмів, що породжують траєкторію $x(t)$, а саме випадкові розподіли (9.1). Зменшення дисперсії є першою перевагою такого підходу до статистичного моделювання, а відсутність потреби в моделюванні випадкових інтервалів перебування – другою.

Опишемо останній алгоритм моделювання в загальному випадку. Нехай даний однорідний марковський процес $x(t)$ з дискретною множиною станів S, обумовлений матрицею інтенсивностей переходу $A = \|a_{xy}\|$, x, $y \in S$. Траєкторії процесу $x(t)$ є східчастими функціями і мають простий ймовірнісний зміст. Якщо в момент t процес $x(t)$ перебуває в стані x, то час, що залишився перебувати в цьому стані t_x є випадковою величиною, що розподілена за експонентним законом

$$P\{t_x > z\} = e^{-a_{xx}z} .$$

У момент виходу $t + t_x$ процес переходить у стан y з ймовірністю

$$p_{xy} = \frac{a_{xy}}{-a_{xx}}, \quad y \neq x, \quad y \in S .$$

Величини p_{xy} є перехідними ймовірностями вкладеного ланцюга Маркова. Отже, на процес $x(t)$ можна дивитися як на марковський ланцюг із приєднаними випадковими величинами. У стані x приєднаною випадковою величиною вважаємо t_x – випадковий час перебування в цьому стані.

Результатом моделювання звичайно є середнє значення інтеграла деякої функції $f(x)$, визначеної над процесом $x(t)$ згідно (9.4). У цьому випадку вигідно заміняти t_x на середнє значення $1 / (-a_{xx})$.

Цей алгоритм можна застосувати і у випадку довільних законів. Треба тільки їх попередньо апроксимувати такими розподілами, які є лінійними комбінаціями або згортками експонентних розподілів. Це приводить до деякого збільшення числа станів, проте, як показує практика моделювання, виходить економія машинного часу й обсягів пам'яті.

9.5 Моделювання реального процесу обслуговування

При імітаційному моделюванні алгоритм має відтворювати процес функціонування системи в часі, імітуючи складові процесу або елементарні явища зі збереженням їх часової та логічної структури. Алгоритм моделювання має бути таким, щоб за мінімальний час отримати статистичні оцінки максимальної точності. У зв'язку із цим існує кілька підходів до імітаційного моделювання систем масового обслуговування. Один з них полягає в моделюванні реального процесу обслуговування вхідного потоку вимог. При цьому використовуються підпрограми реалізації двох випадкових величин: відповідно до функції розподілу інтервалів часу між вимогами $A(z)$ і функції розподілу тривалості обслуговування $B(x)$ (див. розд. 3). Процес надходження вимог у систему моделюється як рекурентний – момент прибуття чергової вимоги дістаємо додаванням випадкового інтервалу $A(z)$ до попереднього, а моменти звільнення серверів – додаванням до поточного моменту випадкової тривалості обслуговування $B(x)$. Інтервали формуються датчиками псевдовипадкових чисел, налаштованими на необхідні закони розподілу.

Випадкові величини $A(z)$ і $B(t)$ з необхідним законом розподілу можуть бути отримані з послідовності випадкових величин з рівномірним розподілом на відрізку [0, 1], що в свою чергу, виходить із послідовності випадкових величин, розподілених за законом Бернуллі з параметром $p = 0{,}5$. Для одержання випадкових величин з необхідною функцією розподілу $F(t)$, випадкова величина t обчислюється як зворотна функція F^{-1} від аргументу, яким є величина x, рівномірно розподілена на відрізку [0, 1], тобто $t = F^{-1}(x)$ [2].

Оскільки потік вимог може бути заданий послідовністю інтервалів часу між вимогами z, то для моделювання будь-якого потоку достатньо на i-му кроці отримати значення величини z_i, що має відповідну модельованому потоку функцію розподілу $F(z)$. Для пуассонівського потоку з параметром λ експонентна функція розподілу інтервалу z між i-ю і $(i - 1)$-ю вимогами має вид:

$$F_i(z) = 1 - e^{-\int_0^z \lambda dt}, \qquad (9.5)$$

де λ – параметр потоку в інтервалі [0, z].

Для моделювання потоку з параметром λ необхідно знайти зворотну функцію до функції (9.5) і для експонентного розподілу отримуємо

$$z_i = \frac{-\ln(1 - x_i)}{\lambda}. \qquad (9.6)$$

Для гіперекспонентного розподілу густина розподілу визначається як

$$p(z) = p_1 \lambda_1 e^{-\lambda_1 \cdot z} + p_2 \lambda_2 e^{-\lambda_2 \cdot z}. \qquad (9.7)$$

Це значить, що з імовірністю p_1 інтервал часу між вимогами має експонентний розподіл з параметром λ_1, а з імовірністю p_2 – з параметром λ_2 (відповідно, $p_1 + p_2 = 1$). Для моделювання такого потоку вимог використовується допоміжний датчик випадкових рівномірно розподілених чисел на інтервалі [0, 1] y, що задає ймовірності p_1 і p_2 ($p_2 = 1 - p_1$). Випадкова

величина z_i визначається залежно від отриманого значення p_1 і відповідно до виразу (9.6), тому

$$z_i = \begin{cases} \dfrac{-\ln(1-x_i)}{\lambda_1} & \text{їðè} \quad y_i \le p_1 \\ \dfrac{-\ln(1-x_i)}{\lambda_2} & \text{їðè} \quad y_i > p_1 \end{cases}. \qquad (9.8)$$

Отже, при отриманні від допоміжного датчика випадкової величини y_i, що не перевищує заданої ймовірності p_1, інтервал z_i формується виходячи зі значення параметра λ_1. У протилежному випадку ($y_i > p_1$, „згенерована" імовірність p_2) – виходячи зі значення параметра λ_2. Описаний спосіб досить простий і не вимагає знаходження оберненої функції до щільності (9.7).

Формування самоподібного потоку може базуватись на суперпозиції декількох незалежних з однаковим розподілом ON/OFF джерел, інтервали між ON і OFF періодами якого мають ефект Ноа. Саме ефект Ноа в розподілі тривалостей ON/OFF періодів є базовим при моделюванні самоподібного трафіка і він є синонімом нескінченної дисперсії. Математично для досягнення цього ефекту можна використовувати розподіл Парето, який ще називають „розподілом з довгим хвостом". Густина розподілу Парето задається функцією:

$$f(x) = \frac{a}{b}\left(\frac{b}{x}\right)^{a+1},$$

де a – параметр форми, b – мода розподілу (мінімальне значення випадкової величини x). Причому, при $a \le 2$ дисперсія нескінченна (що й потрібно в якості однієї з умов самоподібності). Наявність у розподілі так званого „довгого хвоста" забезпечує властивість пачковості трафіка, оскільки в розподілі істотно зростають імовірності довгих інтервалів між вимогами (наприклад, відсутність пакетів на інтервалі) і для „підтримки" заданого середнього значення кількості вимог необхідна їхня концентрація (збільшення) на інших інтервалах часу.

Параметр форми a й параметр Херста H перебувають у такій залежності:

$$H = \frac{3-a}{2}.$$

У практичному моделюванні розподіл Парето утворюється шляхом переходу від рівномірного розподілу методом зворотної функції:

$$z_i = \frac{b}{\sqrt[a]{U_i}},$$

де z_i – i-й інтервал між вимогами, U – випадкове число, рівномірно розподілене на інтервалі [0, 1].

Алгоритм моделювання, схема якого показана на рис. 9.2, дозволяє досліджувати широкий спектр СМО типу *GI/G/m/r*. Це означає, що може бути будь-який тип вхідного потоку, будь-який розподіл тривалості обслуговування, будь-яка дисципліна обслуговування – з втратами при $r = 0$, з необмеженою чергою при $r = \infty$, та комбінована дисципліна при $0 < r < \infty$.

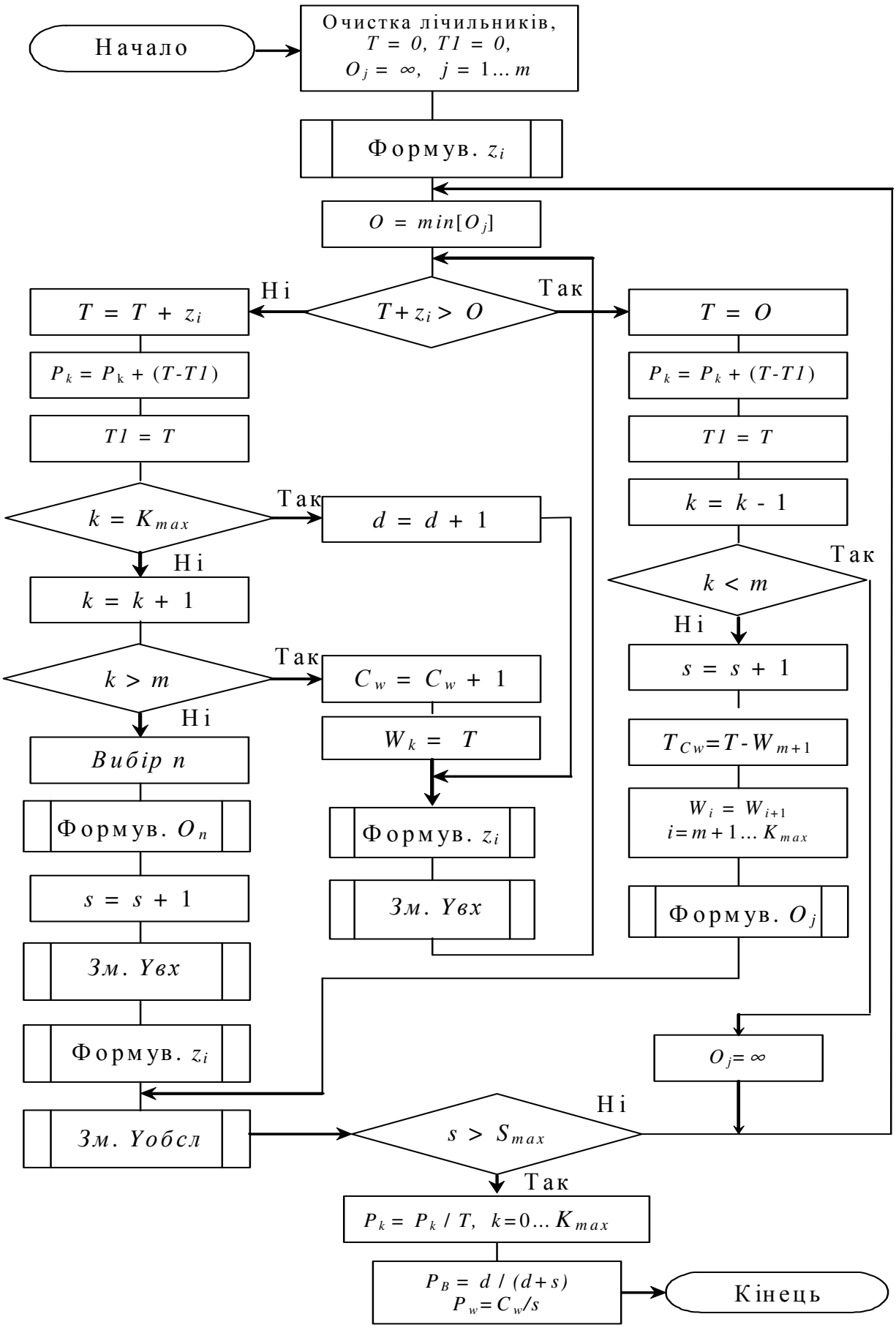


Рисунок 9.2 – Алгоритм імітаційної моделі

Для основних змінних імітаційної моделі прийняті наступні позначення (призначення інших змінних вказано далі за текстом):

- s – поточна кількість оброблених вимог;
- S_{max} – максимальна кількість обслужених вимог;
- d – кількість втрачених (не обслужених) вимог;
- k – поточна кількість вимог в системі (на обслуговуванні й у черзі);
- r – кількість комплектів очікування.
- m – кількість серверів в системі;
- n – номер займаного серверу ($n \le m$);
- K_{max} – максимальна кількість вимог, що є в системі ($K_{max} = m + r$);
- C_w – кількість вимог, що потрапили в чергу на очікування обслуговування;

Робота моделі починається з установки в нуль таймера (лічильника) поточного часу T, таймера моменту попередньої події $T1$ і всіх інших накопичувальних лічильників. Моментам виходу із системи вимог приписуються значення, що перевищують граничне значення таймера T (всі сервери вільні).

Момент виходу із системи вимоги відповідає моменту звільнення серверу й записується в індивідуальну комірку пам'яті, закріплену за кожним сервером. При зайнятті серверу в неї записується момент його майбутнього звільнення, визначений як сума значень таймера поточного часу та тривалості обслуговування $B(t)$. При звільненні серверу в цю комірку записується таке значення (наприклад, 10^{99} що символізує „нескінченність”), якого ніколи не досягне в результаті всього часу моделювання таймер T.

Після приведення схеми у вихідний стан формується випадковий момент z_i прибуття i-ї вимоги. З комірок пам'яті, закріплених за кожним сервером, вибирається мінімальний з моментів звільнення серверів O_j (j – номер серверу, що звільняється). Варіант обробки поточної події визначається співвідношенням між $T + z_i$ і O_j. Якщо $T + z_i \le \min[O_j]$, то поточною подією є надходження вимоги. Відповідно переміщується таймер поточного часу T. До лічильника P_k часу перебування системи в стані з k вимогами (у серверах і в черзі) додається різниця $T - T1$ і оновлюється значення $T1$. Якщо нова вимога застає в системі K_{max} вимог (зайняті всі m серверів і всі r комплектів очікування, тобто $k = K_{max}$), то до лічильника відмов d додається одиниця. Далі формується момент z_{i+1} надходження нової вимоги. Після підпрограми вимірювання надлишкового навантаження виконується повернення до оператора перевірки умови $T + z_i > O_j$ (оскільки z оновлене).

За наявності в системі вільних місць (сервери або комплекти очікування) $k < K_{max}$. Тому поточне значення кількості вимог в системі k збільшується на одиницю й перевіряється наявність вільних серверів. Якщо $k > m$, то їх немає, відповідно дана вимога направляється в комплект очікування. При цьому лічильник затриманих вимог C_w збільшується на одиницю, а в закріпленій за комплектом очікування комірці пам'яті W_k записується момент зайняття T.

Потім формується момент прибуття чергової вимоги z_{i+1} і запускається підпрограма виміру вхідного навантаження $Y_{вх}$.

Якщо є хоча б один вільний сервер ($k \le m$), то визначається його номер n по наявності признаку „∞” у закріпленій за ним комірці пам'яті. Далі формується момент майбутнього звільнення $O_n = T + B(t)$ n-го серверу та записується в його комірку пам'яті. Потім лічильник кількості обслугованих вимог s збільшується на одиницю. Після підпрограми виміру вхідного навантаження $Y_{вх}$ формується наступний момент z_{i+1} надходження нової вимоги. Після підпрограми виміру обслугованого $Y_{обсл}$ навантаження виконується перевірка досягнення заданої кількості обслужених вимог S_{max}. Якщо граничної кількості згенерованих вимог не досягнуто, то здійснюється повернення до оператора пошуку мінімального з моментів звільнення серверів O_j. В протилежному випадку провадиться розрахунок стаціонарних імовірностей P_k, які визначаються як частка загального часу T, протягом якого в системі було зайнято рівно k місць (серверів і комплектів очікування). Для цього накопичені значення лічильників P_k діляться на значення таймера загального часу моделювання T. Імовірність втрат за часом P_t відповідає долі часу, протягом якого в системі були зайняті всі m серверів, або відношенню сумарного часу заняття m серверів за інтервал часу спостереження до довжини інтервалу, тобто P_k / T при $k = m$. Імовірність втрат за вимогами P_B визначається як відношення кількості втрачених вимог d до загальної кількості вхідних вимог $d +$ s (втрачених і обслужених). Імовірність очікування $P_{w>0}$ визначається як відношення кількості затриманих вимог C_w до загальної кількості обслужених вимог s.

Якщо $T + z_i > \min[O_j]$, то черговою подією є завершення розпочатого обслуговування та звільнення місця в системі (серверу або комплекту очікування j). Тут також переміщується таймер поточного часу T, до лічильника P_k часу перебування системи в стані з k зайнятими місцями додається різниця $T - T1$ і оновлюється значення $T1$. Потім кількість зайнятих місць (кількість обслуговуваних системою вимог) k зменшується на одиницю. При $k \ge m$ є черга, а отже, після звільнення серверу з першого (головного) комплекту очікування вимога, що чекає, передається відразу ж у цей сервер на обслуговування. При цьому кількість обслужених вимог s збільшується на одиницю. Тривалість очікування для кожної вимоги із черги, що розраховують як різницю між поточним моментом часу T і моментом часу постановки вимоги в чергу на очікування (записано в комірку W_{m+1}), накопичується в спеціальному масиві T_{Cw} для визначення функції розподілу тривалості очікування й інших моментів. Слідом за тим вимоги (записи в комірках пам’яті про моменти постановки в чергу), що перебувають у комплектах очікування починаючи із другого, послідовно переміщаються в комплекти з номером на одиницю меншим (обслуговування вимог відбувається за правилом *FIFO*). Потім для даного серверу, що прийняв вимогу із черги на обслуговування формується момент майбутнього звільнення $O_n = T + B(t)$ і записується в його комірку пам'яті.

Якщо $k < m$, то черги немає а, отже, звільняється сервер системи. У комірку пам'яті, яка закріплена за звільненим сервером j, записується „нескінченність” – ознака того, що сервер вільний. Після цього виконується перехід на вибір наступного мінімального з моментів звільнення серверів O_j.

Підпрограми „вимірювання” обслуженого й вхідного (сума обслуженого і надлишкового) навантажень і їхньої дисперсії містять масиви для зберігання всіх інтервалів часу між вимогами й всіх тривалостей обслуговування вимог. Обслужене навантаження визначається як відношення сумарного часу зайняття всіх серверів до загального часу спостереження. Вхідне навантаження визначається як відношення середнього часу зайняття серверу до середнього часу інтервалу між двома вимогами (M_t / M_z). Воно ж визначається і як середнє число вимог за середню тривалість зайняття серверу M_t. При цьому дисперсія навантаження визначається як дисперсія кількості вимог за M_t для всіх інтервалів M_t, що містяться в періоді спостереження.

Для дослідження системи з дисципліною обслуговування з втратами необхідно вибрати кількість комплектів очікування $r = 0$, при цьому автоматично $K_{max} = m$.

Для перевірки коректності побудови імітаційної моделі необхідно виконати наступні тести:

– перевірка „випадковості” генератора рівномірно розподілених чисел на інтервалі [0, 1] за критерієм χ^2 Пірсона ;

– перевірка довільно розподілених випадкових чисел по центральних моментах розподілу, коефіцієнтам асиметрії й ексцесу.

Можна випробувати імітаційну модель на отримання відомих результатів. Для цього можна генерувати пуассонівський потік вимог (експонентний розподіл інтервалів часу між вимогами) і одержувані значення ймовірностей втрат порівнювати з розрахованими за формулою Ерланга значеннями.

Моделювання функцій $A(z)$ і $B(x)$ шляхом переходу від рівномірного розподілу методом зворотної функції виконується за формулами табл. 9.1.

Таблиця 9.1 – Метод моделювання випадкової величини

Вид функції	Параметр розподілу	Густина розподілу	Спосіб моделювання
Експонентна	λ – інтенсивність	$\lambda e^{-\lambda x}$	$x_i = -\frac{1}{\lambda}\ln U_i$
Релея	b – мода розподілу	$\frac{x}{b}e^{-\frac{x^2}{2b^2}}$	$x_i = b\sqrt{-\ln U_i}$
Вейбулла	a – параметр форми	$\lambda_0 a x^{a-1} e^{-\lambda_0 x^a}$	$x_i = \left(-\frac{1}{\lambda}\ln U_i\right)^{1/a}$
Парето	a – параметр форми b – мода розподілу (b –*мінімальне значення*)	$\frac{a}{b}\left(\frac{b}{x}\right)^{a+1}$	$x_i = \frac{b}{(U_i)^{1/a}}$

Контрольні запитання та задачі

1. У чому сутність задач аналізу телекомунікаційних систем?
2. Яка значущість задач синтезу телекомунікаційних систем?
3. У чому полягає мета задач оптимізації телекомунікаційних систем?
4. Назвіть основні методи розв'язання задач теорії телетрафіка.
5. Які складові входять до математичної моделі систем розподілу інформації?
6. Як описується дисципліна обслуговування вимог у СРІ?
7. Як описується вхідний потік вимог на обслуговування у СРІ?
8. Який вид розподілу має інтервал часу між вимогами в пуассонівському потоці?
9. Які види імовірнісних розподілів застосовуються для опису випадкових процесів, що відбуваються у СРІ (вхідний потік, тривалість обслуговування, стани системи)?
10. Назвіть види СРІ за способом обслуговування вимог.
11. Назвіть типи дисциплін обслуговування черги в СРІ.
12. Назвіть основні правила обслуговування вимог за пріоритетами.
13. Назвіть основні характеристики, що представляють структуру СРІ.
14. Поясніть структуру умовного позначення базової моделі СРІ за Кендаллом.
15. Що таке „інтенсивність потоку вимог"?
16. Що таке „інтенсивність обслуговування вимог"?
17. Що таке „навантаження" СРІ?
18. Що таке „інтенсивність навантаження" і якими способами її можна визначити?
19. У чому різниця між вхідним та обслуженим навантаженням СРІ?
20. Які властивості пуассонівському потоку вимог?
21. Які типи реального трафіка визначено для телекомунікаційних мереж і у чому різниця між ними?
22. Назвіть характеристики якості обслуговування для системи з втратами.
23. Назвіть характеристики якості обслуговування для системи з чергами.
24. Що таке „пропускна здатність" СРІ?
25. Якими методами досліджується функціонування СРІ за умови обслуговування пуассонівського потоку вимог?
26. Які ознаки пуассонівського потоку вимог?

27. Що визначає поняття „стан системи”?
28. Для розрахунку якої системи призначена *B*-формула Ерланга?
29. Який параметр *QoS* можна розрахувати за *C*-формулою Ерланга?
30. Що визначає формула Літтла?
31. Для розрахунку якої системи призначена Поллачека-Хінчина?
32. Як відрізняються характеристики якості обслуговування в моделях $M/M/1/\infty$ та $M/D/1/\infty$?
33. Що таке „гіперекспонентний розподіл”?
34. Як відрізняються між собою інтенсивність навантаження та її дисперсія для пуассонівського та реального (гіперекспонентного) потоку вимог?
35. Що таке „коефіцієнт скупченості” або „пікфактор” навантаження CPI?
36. Чим характерний трафік пакетних мереж зв’язку, і якою моделлю він може бути описаний?
37. У яких межах може бути коефіцієнт Херста для самоподібних та несамоподібних потоків трафіку.
38. Що визначає ентропія імовірнісного розподілу?
39. Визначити частку втрачених вимог для системи $M/M/1$, до якої надходить потік вимог з інтенсивністю $\lambda = 2$, а інтенсивність обслуговування $\mu = 4$.
40. Визначити середню кількість вимог в черзі в системі $M/M/1/\infty$, до якої надходить потік вимог з інтенсивністю $\lambda = 2$, а інтенсивність обслуговування $\mu = 2{,}5$.
41. Визначити імовірність очікування та імовірність втрати вимоги для системи $M/M/1/5$, до якої надходить потік вимог з інтенсивністю $\lambda = 2$, а інтенсивність обслуговування $\mu = 4$.
42. На скільки відрізняється середня кількість вимог у системах $M/M/1/\infty$ та $M/D/1/\infty$, до якої надходить потік вимог з інтенсивністю 2 вимоги на секунду за середньої тривалості обслуговування вимог $\bar{x} = 0{,}35$ с.
43. Скільки в середньому (в секундах) будуть очікувати вимоги в черзі системи $M/M/80/\infty$ за інтенсивності навантаження $\Lambda = 70$ Ерл та середньої тривалості обслуговування 35 с.
44. Визначити імовірність очікування в системи $M/M/80/\infty$ за інтенсивності навантаження $\Lambda = 70$ Ерл.
45. Довести, що імовірність очікування в системі $M/M/m/r$ завжди менша за цю ж імовірність в системі $M/M/m/\infty$.

Список літератури

1. Крылов В.В., Самохвалова С.С. Теория телетрафика и её приложения. – СПб.: БХВ-Петербург. – 2005. – 288 с.: ил.
2. Шнепс М.А. Системы распределения информации. Методы расчета: Справ. пособие. – М.: Связь, 1979. – 344 с., ил.
3. Корнышев Ю.Н., Фань Г.Л. Теория распределения информации: Учеб. пособие для вузов. – М.: Радио и Связь. – 1985. – 184 с., ил.
4. Ложковский А.Г., Захарченко Н.В., Горохов С.М. Экспериментальная оценка модели потока вызовов на современных телефонных сетях // Наукові праці ОНАЗ ім. О.С. Попова. – 2001. – №2. – С. 40–43.
5. Клейнрок Л. Теория массового обслуживания. Пер. с англ. – М.: Машиностроение. – 1979. – 432 с., ил.
6. Вентцель Е.С., Овчаров Л.А. Теория случайных процессов и её инженерные приложения. – М.: Наука. Ред. физ.-мат. лит. – 1991.– 384с.
7. Кениг Д., Штойян Д. Методы теории массового обслуживания. Пер. с нем. – М.: Радио и Связь. – 1981. – 128 с., ил.
8. Ложковський А.Г. Спрощений метод розрахунку багатоканальної системи з чергою в моделі M/D/m/∞ (Задача Кроммеліна) // Наукові праці ОНАЗ ім. О.С. Попова. – 2008. – № 2. – С. 69-76.
9. Ложковский А.Г., Салманов Н.С., Вербанов О.В. Моделирование многоканальной системы обслуживания с организацией очереди // Восточно-европейский журнал передовых технологий. – 2007. – №3/6(27). – С.72-76.
10. Ложковський А.Г. Нова методика оцінювання імовірності втрат викликів, наближена до реальних умов // К.: Зв'язок. – 2004. – №3. – С. 52–53.
11. Ложковский А.Г. Метод расчета систем обслуживания с ожиданием при произвольном потоке вызовов // К.: Зв'язок. – 2006. – № 1. – С. 57–60.
12. Ложковский А.Г. Сравнительный анализ методов расчета характеристик качества обслуживания при самоподобных потоках в сети // Моделювання та інформаційні технології. Зб. наук. пр. ІПМЕ ім. Г. Є. Пухова НАН України. – Вип. 47. – К.: 2008. – С. 187-193.
13. Ложковский А.Г., Ганифаев Р.А. Оценка параметров качества обслуживания самоподобного трафика энтропийным методом // Наукові праці ОНАЗ ім. О.С. Попова. – 2008. – № 1. – С. 57-62.
14. Ложковский А.Г. Расчет одноканальных систем с бесконечной очередью при экспоненциальной длительности обслуживания // Наукові праці ОНАЗ ім. О.С. Попова. – 2009. – № 2. – С. 10-13.
15. Ложковский А.Г. Рекуррентный метод расчета пропускной способности пакетной сети доступа / А.Г. Ложковский, Н.С. Салманов, Н.А. Чумак // Наукові праці ОНАЗ ім. О.С. Попова. – 2006. – № 2. – С. 44-48
16. Лившиц Б.С. Теория телетрафика: Учебник для вузов / Б.С. Лившиц, А.П. Пшеничников, А.Д. Харкевич // М.: Связь. – 1979. – 224 с., ил.
17. Степанов С.Н. Основы телетрафика мультисервисных сетей. – М.: Эко-Трендз. – 2010. – 392 с.: ил.

Додаткова література

18. Величко В.В., Субботин Е.А., Шувалов В.П., Ярославцев А.Ф. Телекоммуникационные системы и сети. Том 3. Мультисервисные сети. – М.: Горячая линия – Телеком, 2005. – 592 с.: ил.
19. Штермер Х., Белендорф Э. и др. Теория телетрафика. Основы расчета систем проводной связи. Перевод с нем. – М.: Связь, 1971. – 319 с.
20. Эллдин А., Линд Г. Основы теории телетрафика. – М.: Связь, 1972 г. – 200 с.
21. Бенеш В.Э. Математические основы теории телефонных сообщений.– Перевод с нем. – М.: Связь, 1968. – 291 с.
22. Башарин Г.П., Харкевич А.Д., Шнепс М.А. Массовое обслуживание в телефонии. – Изд-во «Наука», 1968.
23. Башарин Г.П., Кокотушкин В.А. О некоторых направлениях развития математической теории телетрафика. Обзор. // Теория телетрафика. Х. Штермер и др. – М.: Связь, 1971. – с. 292-304.
24. Башарин Г.П. О вычислении моментов обслуженной и избыточной нагрузок сложной системы – Изв. АН СССР. Техн. Кибернетика № 1, 1972. – с. 42-51.
25. Вентцель Е.С. Теория вероятностей. – М.: Гос. Издательство физ.-мат. литературы, 1962. – 564с.
26. Вентцель Е.С., Овчаров Л.А. Прикладные задачи теории вероятностей. – М.: Радио и связь. – 1983. – 416 с., ил.
27. Ионин Г.Л., Седол Я.Я. Статистическое моделирование систем телетрафика. - М.: Радио и связь. – 1982. – 184 с.
28. Кожанов Ю.Ф. Расчет и проектирование электронных АТС: Справочник. – М.: Радио и связь. 1991. – 144 с., ил.
29. Корнышев Ю.Н. Оптимизация проектных решений для сельских телефонных сетей. – М.: Связь, 1983. – 136 с.
30. Лившиц А.Л., Мальц Э.А. Статистическое моделирование систем массового обслуживания. – М.: Сов. Радио, 1978.
31. Нейман В.И. Структуры систем распределения информации. – 2-е изд, перераб. и доп. – М.: Радио и связь, 1983. – 216 с., ил.
32. Севастьянов Б.А. Эргодическая теорема для марковских процессов и ее приложение к телефонным системам с отказами. Сб. «Теория вероятностей и ее применения», 1957, т. 2, вып. 1. – с. 106-116.
33. Справочник по теории вероятностей и математической статистике / В.С. Королюк и др. – М.: Наука. Главная редакция физико-математической литературы, 1985. – 640 с.
34. Фигурин В.А., Оболонкин В.В. Теория вероятностей и математическая статистика: Учеб. Пособие. – Мн.: ООО «Новое знание», 2000. – 208 с.
35. Хинчин А.Я. Математические методы теории массового обслуживания. // Работы по математической теории массового обслуживания. Изд-во физ.-мат. Лит-ры. – М.: 1963. – с. 7-148.

Список позначень та скорочень

$C_m(\Lambda)$ – *C*-формула Ерланга

$E_m(\Lambda)$ – *B*-формула Ерланга

k – стан системи (кількість зайнятих серверів)

m – кількість серверів системи

N – середня кількість вимог у системі

P_k – імовірність стану системи в разі зайнятості k серверів

$P_{w>0}$ – імовірність очікування

P_B – імовірність втрати (блокування) вимоги

Q – середня довжина черги

r – кількість місць очікування в черзі

S – коефіцієнт скупченості навантаження (пікфактор трафіка)

T – середня тривалість перебування вимог у системі

t_q – середня тривалість очікування вимог в черзі

W – середня тривалість очікування вимог в системі

x – тривалість обслуговування вимоги

Y – інтенсивність обслуженого навантаження

z – тривалість інтервалу часу між вимогами

λ – інтенсивність потоку вимог

Λ – інтенсивність вхідного навантаження

μ – інтенсивність обслуговування вимог

ρ – інтенсивність питомого навантаження

FIFO – *First in first out* (перший обслуговується першим)

LIFO – *Last in first out* (останній обслуговується першим)

QoS – *Quality of Service* (якість обслуговування)

SIRO – *Service in random order* (випадкове обслуговування)

ГНН – година найбільшого навантаження

ЕОМ – електронна обчислювальна машина

СМО – Система масового обслуговування

СРІ – Система розподілу інформації

ТМО – Теорія масового обслуговування

Предметний покажчик

ДОДАТОК

Таблиця 1 – Інтенсивність навантаження Y, обслуженого m-серверним повнодоступним пучком при втратах 1, 3 і 5 ‰ та коефіцієнті S скупченості вхідного навантаження

m	***S* = 5**			***S* = 4**			***S* = 3**			***S* = 2**			***S* = 1**		
	Y в Ерл при втратах P_B			*Y* в Ерл при втратах P_B			*Y* в Ерл при втратах P_B			*Y* в Ерл при втратах P_B			*Y* в Ерл при втратах P_B		
	0,001	**0,003**	**0,005**	**0,001**	**0,003**	**0,005**	**0,001**	**0,003**	**0,005**	**0,001**	**0,003**	**0,005**	**0,001**	**0,003**	**0,005**
5	0,61	0,66	0,69	0,69	0,75	0,80	0,72	0,87	0,93	0,74	0,93	1,00	0,76	1,00	1,13
10	2,10	2,28	2,42	2,33	2,57	2,74	2,59	2,91	3,11	2,94	3,35	3,59	3,09	3,63	3,94
15	4,14	4,53	4,80	4,54	5,01	5,33	5,00	5,59	5,96	5,59	6,29	6,67	6,07	6,89	7,34
20	6,54	7,17	7,59	7,13	7,87	8,34	7,80	8,66	9,18	8,62	9,58	10,08	9,40	10,46	11,04
25	9,24	10,10	10,68	10,00	10,99	11,61	10,86	11,99	12,67	11,88	13,09	13,71	12,96	14,23	14,92
30	12,14	13,25	13,98	13,07	14,32	15,08	14,11	15,49	16,31	15,32	16,76	17,50	16,67	18,15	18,94
35	15,20	16,57	17,44	16,29	17,79	18,71	17,50	19.13	20,09	18,90	20,55	21,40	20,50	22,16	23,05
40	18,40	20,01	21,03	19,65	21,39	22,45	21,02	22.89	23,98	22,59	24,44	25,40	24,42	26,26	27,25
45	21,72	23,56	24,73	23,12	25,07	26,28	24,64	26.74	27,95	26,36	28,40	29,48	28,42	30,43	31,50
50	25,13	27,21	28,51	26,66	28,85	30,19	28,33	30.66	32,00	30,22	32,44	33,62	32,48	34,65	35,80
55	28,62	30,93	32,36	30,29	32,71	34,16	32,10	34.64	36,08	34,14	36,54	37,81	36,59	38,91	40,22
60	32,19	34,72	36,28	33,98	36,62	38,20	35,93	38.88	40,21	38,11	40,69	42,05	40,75	43,23	44,53
65	35,81	38,56	40,25	37,73	40,59	42,29	39,82	42.77	44,38	42,13	44,87	46,33	44,95	47,57	48,95
70	39,50	42,46	44,27	41,54	44,61	46,42	43,75	46.91	48,59	46,19	49,10	50,65	49,19	51,94	53,39
75	43,23	46,41	48,34	45,39	48,66	50,60	47,73	51.08	52,83	50,30	53,36	54,99	53,46	56,34	57,86
80	47,01	50,40	52,45	49,29	52,75	54,80	51,75	55.29	57,11	54,45	57,66	59,36	57,75	60,77	62,35
85	50,83	54,43	56,60	53,22	56,89	59,04	55,79	59.53	61,41	58,61	61,97	63,76	62,07	65,21	66,87
90	54,70	58,48	60,76	57,19	61,06	63,31	59,88	63.80	65,75	62,82	66,31	68,19	66,41	69,68	71,39
95	58,59	62,58	64,97	61,19	65,24	67,62	64,00	68.12	70,10	67,04	70,68	72,63	70,78	74,17	75,94
100	62,5	66,7	69,2	65.2	69,5	71,9	68,1	72,4	74,5	71,3	75,1	77,1	75,2	78,7	80,5
110	70,5	75,0	77,8	73,4	78,0	80,6	76,5	81,1	83,3	79,9	83,9	86,1	84,0	87,7	89,7
120	78,5	83,5	86,4	81,6	86,6	89,3	84,9	89,9	92,2	88,5	92,8	95,1	92,9	96,8	98,9
130	86,7	92,0	95,1	89,9	95,3	98,1	93,4	98,7	101,1	97,2	101,7	104,2	101,8	106,0	108,2
140	94,9	100,5	103,9	98,3	104,0	106,9	102,0	107,5	110,1	105,9	110,7	113,3	110,8	115,2	117,4
150	103,2	109,2	112,6	106,8	112,8	115,8	110,7	116,5	119,2	114,7	119,8	122,5	119,8	124,4	126,8
160	111,5	117,9	121,4	115,3	121,6	124,8	119,4	125,4	128,2	123,6	128,9	131,7	128,9	133,6	136,1
170	120,0	126,6	130,3	123,9	130,5	133,8	128.1	134,4	137,4	132,5	138,0	141,0	138,0	142,9	145,5
180	128,5	135,4	139,2	132,6	139,5	142,8	136,9	143,4	146,5	141,4	147,2	150,2	147,2	152,2	154,9
190	137,0	144,3	148,2	141,2	148,4	151,9	145,8	152,5	155,7	150,4	156,4	159,5	156,2	161,6	164,4
200	145,5	153,2	157,1	150,0	157,5	161,0	154,6	161,6	164,9	159,4	165,6	168,9	165,4	170,9	173,7
210	154,2	162,1	166,2	158,7	166,5	170,1	163,5	170,7	174,2	168,4	174,8	178,2	174,6	180,3	183,3
240	180,2	189,1	193,4	185,2	193,8	197,6	190,4	198,2	202,0	195,6	202,7	206,4	202,4	208,6	211,8

Навчальне видання

Анатолій Григорович Ложковський

ТЕОРІЯ МАСОВОГО ОБСЛУГОВУВАННЯ В ТЕЛЕКОМУНІКАЦІЯХ

підручник
для студентів ВНЗ, які навчаються за напрямом „Телекомунікації”

Издательство ОНАС им. А.С. Попова
(свидетельство ДК № 3633 от 27.11.2009)

Здано в набір 06.04.2010. Підписано до друку 10.04.2010
Формат 60х90 1/16. Зам. № 41
Тираж 300 прим. Обсяг 7,3 друк. арк.
Віддруковано на видавничому устаткуванні фірми RISO
у друкарні редакційно-видавничого центру ОНАЗ ім. О.С.Попова