

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Дніпровський національний університет імені Олеся Гончара

Механіко-математичний факультет

Кафедра статистики й теорії ймовірностей

100 років

*Дніпровському національному університету імені Олеся Гончара
(1918–2018)*

Гончаров С. В.

КОНСПЕКТ ЛЕКЦІЙ
З ТЕОРІЇ МАСОВОГО ОБСЛУГОВУВАННЯ

ЧАСТИНА I

Дніпро
Видавець Біла К. О.
2018

УДК 519.2
Г 65

*Рекомендовано до друку Вченою радою
механіко-математичного факультету
Дніпровського національного університету імені Олеся Гончара
(протокол № 8 від 27.03.2018)*

Рецензенти:

Послайко Н. І., канд. фіз.-мат. наук, доцент кафедри прикладної математики Дніпропетровського національного університету залізничного транспорту імені академіка В. Лазаряна;

Руденко О. О., канд. фіз.-мат. наук, доцент кафедри математичного аналізу і теорії функцій Дніпровського національного університету імені Олеся Гончара.

Гончаров С. В.

Г 65 Конспект лекцій з теорії масового обслуговування. Частина I / С. В. Гончаров. – Дніпро : Видавець Біла К. О., 2018. – 44 с.

Посібник містить лекції з курсу теорії масового обслуговування. Лекції охоплюють наступні теми: класифікація систем масового обслуговування (СМО), формула Літтла, пуассонів потік подій, процес народження й загибелі, марківські СМО, зокрема скінченноджерельні та з нетерпінням. Також у посібнику наведені лістинги симуляторів СМО на мовах програмування Python та R.

Для студентів спеціальності «Статистика» механіко-математичного факультету ДНУ, а також для усіх бажаючих ознайомитись з основами теорії масового обслуговування.

УДК 519.2

© Гончаров С. В., 2018

Зміст

Вступ. Класифікація СМО. Формула Літтла	4
1. Пуассонів потік подій	9
2. Марківські СМО	13
2.1. Марківські процеси	13
2.2. Процес народження й загибелі	17
2.3. Система $M/M/1$	18
2.4. Система $M/M/n$	22
2.5. Система $M/M/n/m$	27
2.6. Система $M/M/\infty$	31
2.7. Скінченноджерельні системи	31
2.8. Системи з нетерпінням	33
2.9. Перехідна поведінка	34
Симулятор $M/M/n$ на Python	36
Симулятор $M/M/n$ на R	39
Позначення	43
Список рекомендованої літератури	43

Вступ

Теорія масового обслуговування (ТМО) — це теорія систем масового обслуговування (СМО). Її ще називають теорією черг, Queueing Theory.

Початок розвитку ТМО відносять до першої половини ХХ ст., до робіт датського математика А. К. Ерланга, який працював у телефонній компанії Копенгагена. Він досліджував роботу телефонної мережі, кількості приладів та телефоністок (тогочасні телефонні станції потребували операторів, які вручну з'єднували відповідні номери, і абонент повинен був спочатку сказати оператору, з яким номером він прагне зв'язатись). Зокрема, уже в його роботах виділено два важливі типи таких систем: з чергами та без.

Типові приклади СМО:

- На березі річки чекає кілька човнів із човнярами. До річки з лісу виходять люди, що платять човнярам, і ті перевозять їх на інший берег.
- Холл на першому поверсі хмарочоса, де клієнти банку підходять до кас, аби зробити внесок, забрати гроші тощо.
- На шосе посеред пустелі машини під'їжджають до заправки.
- До серверу надходять запити від користувачів на авторизацію, завантаження контенту та його видалення тощо.

У будь-якій СМО ми виділяємо принаймні дві основні компоненти: події, які ще називають *запитами* або *заявками*, і те, що їх обробляє — *прилади*.

Розглянемо конкретну СМО та поставимо кілька питань:

- Який розподіл кількості запитів протягом певного проміжка часу?
- Як розподілена тривалість обробки одного запиту?
- Скільки приладів?
- Чи можуть запити ставати у чергу або в декілька черг?
- У якому порядку обслуговуються запити?
- Скільки в середньому запит чекає на обробку?
- Скільки в середньому проводить запит в системі?

тощо. Відповіді на ці питання відрізнятимуть дану систему від інших.

Чи не найважливішою особливістю СМО є стохастичність їхньої поведінки, відсутність розкладу, за яким запити надходили б до приладів та обслуговувались без будь-яких відхилень. Значення випадковості демонструє

ПР. Нехай система складається з одного приладу. Спостереження виявили, що в середньому за 1 годину надходить 6 запитів. Якою має бути тривалість обробки одного запиту, аби не виникало черг?

Розв'язок здається очевидним: $1 \text{ год} / 6 \text{ запитів} = 10 \text{ хвилин на запит}$, і щоб додати «запас» часу, відведемо 9 хвилин. В середньому система встигатиме обробляти всі запити й 6 хвилин на годину залишатиметься вільною.

Але реалізації такої системи із вказаними параметрами виявили проблему: часто перед приладом збиралась черга, довжина якої досягала 8 запитів.

Отже, будувати систему на основі таких міркувань, «розмиваючи» стохастичну поведінку усередненням, не годиться. Справжній потік запитів та його обробка системою відрізняються від усереднено-рівномірного.

Типові задачі, які виникають в ТМО:

а) для конкретної, вже існуючої СМО знайти її параметри, характеристики, оцінити ефективність її функціонування;

б) маючи певний набір умов та обмежень, запропонувати найоптимальнішу СМО, яка задовольняє цим умовам і забезпечує найбільшу ефективність.

Два загальні підходи до вирішення цих задач — аналітичний та симуляційний (часто вони комбінуються). При першому є аналітичні зв'язки між параметрами СМО та її ефективностями, які дозволяють вибрати параметри найкращим чином, а при другому СМО моделюється і обираються такі параметри, для яких модель показує більшу ефективність.

Класифікація СМО

Традиційно СМО розрізняють за такими основними характеристиками:

1) *Вхідний потік запитів*. Як розподілені тривалості проміжків між послідовними надходженнями запитів? Чи чекає запит у черзі, якщо вільних приладів немає, і як довго? Чи може декілька запитів надходити одночасно?

2) *Обробка запиту*. Який розподіл тривалості обробки? Чи обслуговує прилад декілька запитів одночасно? Чи залежить швидкість його роботи від кількості запитів у черзі?

3) *Кількість приладів*, що обробляють запити.

4) *Ємність системи*. Скільки запитів може одночасно перебувати в системі — тих, що обробляються, і тих, що чекають в черзі?

5) *Дисципліна обслуговування*. У якому порядку обробляються запити? Найрозповсюдженішою дисципліною є звичайна черга, або FCFS (first come, first served). Протилежний випадок — LCFS (last come, first served). Зустрічаються RSS (random selection service), коли наступний запит обирається з черги випадково, і SRPT (shortest remaining processing time), коли спочатку обробляється запит з найменшим залишковим часом обслуговування, а також різноманітні дисципліни з пріоритетами.

Нотація. Стандартним для ТМО є скорочене позначення типу СМО, започатковане Кендаллом (1953):

$$A / B / k / m / Z$$

• A та B — позначають розподіл вхідного потоку та розподіл часу обробки запиту. Типові варіанти: M для «марківського» (показникового) розподілу, D для «виродженого» (детермінованого, тривалість є константою), E_k для розподілу Ерланга порядку k , G для довільного (невизначеного) розподілу;

- k та m — відповідають п. 3 і 4 вище;
- Z — дисципліна обслуговування.

Зазвичай m і Z опускають, якщо $m = \infty$, $Z = FCFS$ (за умовчанням).

Отже, $M/G/3$ означатиме СМО з пуассоновим потоком запитів, тобто показниковим розподілом проміжків між моментами їх надходжень, довільним розподілом тривалості обробки запиту і 3-ма приладами, а також нескінченною ємністю і обробкою першим того запиту, який з'явився раніше.

Формула Літтла

Введемо деякі загальні параметри функціонування СМО та отримаємо одне з найважливіших співвідношень між ними, яке виконується майже без будь-яких додаткових припущень, тобто для системи $G/G/k$ (більшість розглядуваних далі СМО будуть її частинними випадками).

Основною вимогою до СМО, для яких ми встановимо це співвідношення, є *ергодичність* — існування стаціонарного режиму/розподілу, до якого система з часом приходить майже напевне, незалежно від початкового стану. Це дозволяє ототожнювати усереднення деяких показників за часом (time average) та за траєкторіями (ensemble average).

Також постулюємо, що кожний запит не залишає систему, поки не буде повністю оброблений, — говорять, що запити не «зникають». І вважатимемо, що система почала працювати у момент $t = 0$.

Нехай час $t > 0$ і $A(t)$ — кількість запитів, що надійшли у систему протягом $[0; t]$. Границя $\lambda = \lim_{t \rightarrow +\infty} \frac{A(t)}{t}$ існує майже напевне внаслідок ергодичності і називається *інтенсивністю вхідного потоку*, або *інтенсивністю надходження запитів до системи*. Інколи замість інтенсивності говорять про *середню швидкість* або *середній темп*. Цей параметр обернений до середнього проміжку між моментами надходжень послідовних запитів.

Другим важливим параметром є μ — *інтенсивність обробки*, або *середня швидкість обробки* одним приладом, — скільки в середньому запитів прилад обробляє за одиницю часу. Для СМО із k таких приладів загальна інтенсивність обробки становитиме $k\mu$.

Відношення $\rho = \frac{\lambda}{k\mu}$ для системи із k приладів називають *коефіцієнтом використання* (зокрема, у випадку єдиного приладу $\rho = \lambda/\mu$). Очевидно, при $\rho > 1$, — коли в одиницю часу запитів надходить більше, ніж система встигає обробити, — черга зростатиме необмежено, і система ніколи не досягне стаціонарного режиму. При $\rho = 1$ стаціонарний режим можливий тільки для детермінованих надходження й обслуговування за оптимального їх розкладу; будь-яке випадкове відхилення збільшує чергу і гальмує систему, причому з часом «відставання від графіку» тільки посилюється. Необхідною умовою існування стаціонарного режиму для стохастичної СМО є $\rho < 1$.

Поточна кількість запитів у системі $\mathcal{N} = \mathcal{N}(t)$ — випадкова величина (ВВ), розподіл якої у загальному випадку невідомий. «Середнє $\mathcal{N}$ » можна тлумачити як *усереднення за часом* на деякій траєкторії, $\overline{\mathcal{N}} = \lim_{t \rightarrow +\infty} \frac{1}{t} \int_0^t \mathcal{N}(u) du$, або як *усереднення за траєкторіями* в деякий момент часу, $M\mathcal{N} = M\mathcal{N}(t)$. Завдяки ергодичності для достатньо великих t вважатимемо $\overline{\mathcal{N}} = M\mathcal{N}$.

Час перебування запиту в системі $\mathcal{T}$ — ВВ з середніми $\overline{\mathcal{T}} = M\mathcal{T}$.

$\mathcal{N}_q$ та $\mathcal{T}_q$ — кількість запитів у черзі (queue) та тривалість перебування запиту в черзі відповідно;

$\mathcal{N}_s$ та $\mathcal{T}_s$ — кількість запитів, що обробляються (service), та тривалість обробки запиту.

$\mathcal{N} = \mathcal{N}_q + \mathcal{N}_s$ та $\mathcal{T} = \mathcal{T}_q + \mathcal{T}_s$, звідки $\overline{\mathcal{N}} = \overline{\mathcal{N}}_q + \overline{\mathcal{N}}_s$, $\overline{\mathcal{T}} = \overline{\mathcal{T}}_q + \overline{\mathcal{T}}_s$. Незалежно від розподілу $\mathcal{T}_s$ середнє $\overline{\mathcal{T}}_s = \frac{1}{\mu}$ — обернено пропорційне темпу обробки.

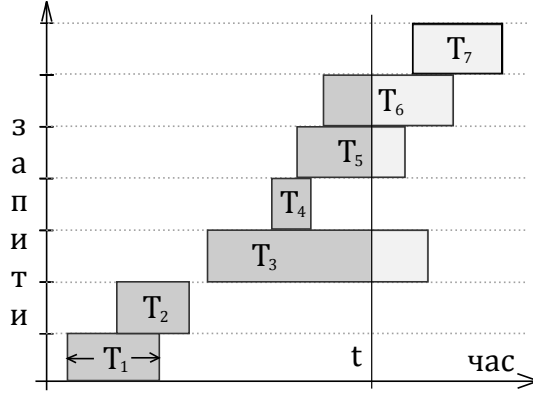
ТЕОР. (Формула Літтла). $\overline{\mathcal{N}} = \lambda \overline{\mathcal{T}}$: щоб визначити середню кількість запитів у системі, треба перемножити інтенсивність надходження запитів та середню тривалість перебування запиту в системі.

Цього висновку можна дійти, міркуючи так: за достатньо довгий час t до СМО надходить в середньому λt запитів, кожен з яких перебуває в ній у середньому $\overline{\mathcal{T}}$, отже, загальний час, проведений усіма ними в системі — $S = \lambda t \overline{\mathcal{T}}$. Тому середня кількість запитів у кожен момент, $\overline{\mathcal{N}} = S/t = \lambda \overline{\mathcal{T}}$.

Тепер наведемо більш строге доведення.

◀ Розглянемо довільну траєкторію роботи системи. Нехай $C(t)$ — кількість запитів, що були повністю оброблені (та залишили систему) за час $[0; t]$. Очевидно, $C(t) \leq A(t)$. Для ергодичної системи $\exists \lim_{t \rightarrow +\infty} \frac{C(t)}{t} \leq \lambda$, причому, оскільки запити не зникають, ця границя рівна λ .

На координатній площині уздовж осі абсцис відкладемо час, а уздовж осі ординат — запити, у порядку їх надходження, по одиничному відрізку на кожний. Відмітивши для k -го запиту прямокутник, відповідний інтервалу часу тривалістю T_k , коли даний запит перебував у системі, отримаємо графік:



Нехай $S(t)$ — спільна площа тих частин прямокутників, які (частини) лежать зліва від t . Тоді, з одного боку, при «горизонтальному» підрахунку

$$\sum_{k \in C(t)} T_k \leq S(t) \leq \sum_{k \in A(t)} T_k$$

(де $k \in C(t)$ означає індекси запитів, повністю оброблених до моменту t , і аналогічно для $A(t)$), а з іншого, при «вертикальному» $S(t) = \int_0^t \mathcal{N}(u) du$.

$$\begin{aligned} \frac{1}{t} \sum_{k \in C(t)} T_k &\leq \frac{1}{t} \int_0^t \mathcal{N}(u) du \leq \frac{1}{t} \sum_{k \in A(t)} T_k \Leftrightarrow \\ \Leftrightarrow \frac{C(t)}{t} \cdot \frac{1}{C(t)} \sum_{k \in C(t)} T_k &\leq \frac{1}{t} \int_0^t \mathcal{N}(u) du \leq \frac{A(t)}{t} \cdot \frac{1}{A(t)} \sum_{k \in A(t)} T_k \end{aligned}$$

Спрямуємо t до $+\infty$. При цьому $\frac{C(t)}{t}$ і $\frac{A(t)}{t} \rightarrow \lambda$, також $\frac{1}{t} \int_0^t \mathcal{N}(u) du \rightarrow \overline{\mathcal{N}}$.

$\frac{1}{C(t)} \sum_{k \in C(t)} T_k$ — це загальна тривалість запитів, оброблених повністю до моменту t , поділена на їх кількість — тобто середня тривалість вже оброблених запитів. При $t \rightarrow +\infty$ вона $\rightarrow \overline{\mathcal{T}}$; аналогічно для відношення з $A(t)$. Отже,

$$\lambda \overline{\mathcal{T}} \leq \overline{\mathcal{N}} \leq \lambda \overline{\mathcal{T}} \Leftrightarrow \overline{\mathcal{N}} = \lambda \overline{\mathcal{T}} \blacktriangleright$$

Аналогічно встановлюється $\overline{\mathcal{N}}_q = \lambda \overline{\mathcal{T}}_q$, звідки також $\overline{\mathcal{N}}_s = \lambda \overline{\mathcal{T}}_s$.

Н. $\overline{\mathcal{N}}_s = \frac{\lambda}{\mu}$; зокрема, для системи $G/G/1$ буде $\overline{\mathcal{N}}_s = \rho$.

ОЗН. Величина $r = \frac{\lambda}{\mu}$ називається *навантаженням* СМО.

Оскільки $\overline{\mathcal{N}}_s = r$ — середня кількість запитів, що обслуговуються у випадковий момент, то в системі з n приладами, внаслідок симетрії, на кожний прилад в середньому припадає $\frac{r}{n} = \rho$ запитів. Позначимо через p_b (busy) ймовірність того, що прилад зайнятий (обробкою запиту), тоді кількість запитів на даному приладі є ВВ $\xi \sim \begin{pmatrix} 0 & 1 \\ 1-p_b & p_b \end{pmatrix}$. Отже, $M\xi = 0 \cdot (1-p_b) + 1 \cdot p_b = p_b$. Звідси $p_b = \rho$, а ймовірність того, що прилад вільний — $1-p_b = 1-\rho$.

1. Пуассонів потік подій

Першим компонентом моделі СМО є модель вхідного потоку запитів. Називатимемо надходження запиту до системи подією. Потік подій описує траєкторія випадкового процесу (ВП) $N(t)$, де $N(t) \in \mathbb{Z}_+$ — кількість подій, що відбулись до моменту t . Вважаємо $t \geq 0$ (процес має початок у часі).

Ми розглянемо *найпростіший*, або *пуассонів*, потік подій.

ОЗН. ВП $N(t)$ має *незалежні прирости*, якщо $\forall \{t_i\}_{i=0}^n, 0 \leq t_0 < t_1 < \dots < t_n$: ВВ $\{N(t_k) - N(t_{k-1})\}_{k=1}^n$ Н/З в сукупності

ОЗН. ВП $N(t)$ має *стаціонарні прирости*, якщо для довільних s і $t, s < t$, розподіл приросту $N(t) - N(s)$ залежить тільки від тривалості $t - s$.

ОЗН. Потік подій $N(t)$ *пуассонівський з інтенсивністю λ* (ПП), якщо

1) $N(0) = 0$ м.н. (майже напевно): $P\{N(0) = 0\} = 1$;

2) $N(t)$ має Н/З прирости;

3) $\forall s, t, s < t$: $N(t) - N(s)$ має розподіл Пуассона з параметром $\lambda(t - s)$.

Із (3) відразу випливає, що $N(t)$ має стаціонарні прирости; із (1) і (3) — що $N(t) = N(t) - N(0)$ має розподіл Пуассона з параметром λt .

ТВ. Проміжки часу між послідовними подіями в ПП є Н/З ВВ, які мають показниковий розподіл з параметром λ .

◀ Нехай це проміжки $\{\tau_i\}_{i=1}^\infty$ (тоді k -та подія відбувається у момент $\sum_{i=1}^k \tau_i$).

Для $\forall n \in \mathbb{N}, \forall \{k_i\}_{i=1}^n$, де $k_i < k_{i+1}$, та $\forall \{t_i\}_{i=1}^n$, де $t_i > 0$:

$$\begin{aligned} & P\{\tau_{k_1} < t_1, \tau_{k_2} < t_2, \dots, \tau_{k_n} < t_n\} = \\ & = P\left\{N\left(\sum_{j=1}^{k_1-1} \tau_j + t_1\right) - N\left(\sum_{j=1}^{k_1-1} \tau_j\right) > 0, \dots, N\left(\sum_{j=1}^{k_n-1} \tau_j + t_n\right) - N\left(\sum_{j=1}^{k_n-1} \tau_j\right) > 0\right\} \stackrel{(2)}{=} \\ & = \prod_{i=1}^n \left[1 - P\left\{N\left(\sum_{j=1}^{k_i-1} \tau_j + t_i\right) - N\left(\sum_{j=1}^{k_i-1} \tau_j\right) = 0\right\}\right] \stackrel{(3)}{=} \\ & = \prod_{i=1}^n \left[1 - P\{N(t_i) - N(0) = 0\}\right] = \prod_{i=1}^n \left(1 - \frac{(\lambda t_i)^0}{0!} e^{-\lambda t_i}\right) = \prod_{i=1}^n (1 - e^{-\lambda t_i}) \blacktriangleright \end{aligned}$$

Можна показати, що ця властивість характеристична для ПП.

ТВ. (Розщеплюваність ПП). Нехай $N(t)$ — ПП з інтенсивністю λ , в якому кожна подія Н/З від інших має тип A чи B з ймовірністю p та $1 - p$ відповідно. Тоді потоки $N_A(t)$ і $N_B(t)$ подій типу A і B відповідно є Н/З ПП з інтенсивностями λp і $\lambda(1 - p)$.

$$\begin{aligned}
\blacktriangleleft \text{Згідно ФПЙ, } P\{N_A(t) = n, N_B(t) = m\} &= \\
&= \sum_{k=0}^{\infty} P\{N_A(t) = n, N_B(t) = m/N(t) = k\} \cdot P\{N(t) = k\} = \\
&= P\{N_A(t) = n, N_B(t) = m/N(t) = n+m\} \cdot P\{N(t) = n+m\}
\end{aligned}$$

Перший множник — ймовірність того, що у послідовності $n+m$ Н/З подій («випробувань») рівно n мають тип A («успіх»); другий визначений пуассоновістю вихідного потоку. Отже, $P\{N_A(t) = n, N_B(t) = m\} =$

$$= C_{n+m}^n p^n (1-p)^m \cdot \frac{(\lambda t)^{n+m}}{(n+m)!} e^{-\lambda t} = \frac{(\lambda p t)^n}{n!} \cdot \frac{(\lambda(1-p)t)^m}{m!} e^{-\lambda t}$$

$$\begin{aligned}
\text{Значить, } P\{N_A(t) = n\} &= \sum_{m=0}^{\infty} P\{N_A(t) = n, N_B(t) = m\} = \\
&= \frac{(\lambda p t)^n}{n!} e^{-\lambda t} \cdot \sum_{m=0}^{\infty} \frac{(\lambda(1-p)t)^m}{m!} = \frac{(\lambda p t)^n}{n!} e^{-\lambda t} \cdot e^{\lambda(1-p)t} = \frac{(\lambda p t)^n}{n!} e^{-\lambda p t}
\end{aligned}$$

і аналогічно $P\{N_B(t) = m\} = \frac{(\lambda(1-p)t)^m}{m!} e^{-\lambda(1-p)t}$. Звідси

$$\forall n, m: P\{N_A(t) = n, N_B(t) = m\} = P\{N_A(t) = n\} \cdot P\{N_B(t) = m\}$$

Поєднуючи це з Н/З приростів, маємо пуассоновість Н/З потоків $N_A(t)$ і $N_B(t)$ із вказаними інтенсивностями. $\blacktriangleright$

ТВ. (Рівномірність ПП). Якщо відомо, що протягом $[s; t]$ у ПП відбулось n подій, то вони розподілені Н/З і кожна — рівномірно на $[s; t]$.

$\blacktriangleleft$ Внаслідок стаціонарності можемо вважати $[s; t] = [0; h]$ (для спрощення позначень). Нехай B полягає у тому, що протягом $[0; h]$ відбулось рівно n подій. За пуассоновістю $P(B) = \frac{(\lambda h)^n}{n!} e^{-\lambda h}$. Узявши серед цих подій одну конкретну, «виділену», позначимо через A її потрапляння у проміжок $[a; b] \subseteq [0; h]$. Покажемо, що $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{b-a}{h}$.

Нехай $C_{s,r}$ — «у $[0; a]$ відбулось s подій, у $[a; b]$ відбулось r подій, серед яких виділена, у $[b; h]$ відбулось $n - s - r$ подій». Тоді $A \cap B = \bigcup_{s=0}^{n-1} \bigcup_{r=1}^{n-s} C_{s,r}$, і $C_{s,r}$ несумісні. Позначивши $N[x; y] = N(y) - N(x)$, внаслідок Н/З приростів:

$$\begin{aligned}
P(C_{s,r}) &= P\{N[0; a] = s\} \cdot P\{N[a; b] = r\} \times \\
&\times P\{N[b; h] = n - s - r\} \cdot P\{\text{виділена подія — серед } r \text{ з усіх } n\} = \\
&= \frac{(\lambda a)^s}{s!} e^{-\lambda a} \cdot \frac{(\lambda(b-a))^r}{r!} e^{-\lambda(b-a)} \cdot \frac{(\lambda(h-b))^{n-s-r}}{(n-s-r)!} e^{-\lambda(h-b)} \cdot \frac{r}{n} = \\
&= \frac{r}{n} \cdot \lambda^n e^{-\lambda h} \cdot \frac{a^s}{s!} \cdot \frac{(b-a)^r}{r!} \cdot \frac{(h-b)^{n-s-r}}{(n-s-r)!}
\end{aligned}$$

$$\begin{aligned}
\text{звідки } P(A \cap B) &= \frac{1}{n} \cdot \lambda^n e^{-\lambda h} \sum_{s=0}^{n-1} \frac{a^s}{s!} \sum_{r=1}^{n-s} \frac{(b-a)^r}{(r-1)!} \cdot \frac{(h-b)^{n-s-r}}{(n-s-r)!} \stackrel{k \leftarrow r-1}{=} \\
&= \frac{b-a}{n} \cdot \lambda^n e^{-\lambda h} \sum_{s=0}^{n-1} \frac{a^s}{s!(n-s-1)!} \sum_{k=0}^{n-s-1} C_{n-s-1}^k (b-a)^k (h-b)^{n-s-1-k} = \\
&= \frac{b-a}{n!} \cdot \lambda^n e^{-\lambda h} \sum_{s=0}^{n-1} C_{n-1}^s a^s (h-a)^{n-s-1} = \frac{b-a}{h} \cdot \frac{(\lambda h)^n}{n!} e^{-\lambda h}
\end{aligned}$$

поділивши на $P(B)$, отримаємо шукану рівність. $\blacktriangleright$

Тепер виведемо пуассоновість потоку із достатньо загальних вимог, які відображають природні властивості багатьох реальних потоків подій.

ОЗН. Потік називається *стаціонарним*, якщо для довільного скінченного набору неперетинних відрізків $\Delta_1, \dots, \Delta_n$ часу ймовірність появи протягом них відповідно $k_1, \dots, k_n$ подій залежить тільки від довжин $|\Delta_i|$ та чисел k_i , але не залежить від їх розташування на вісі часу.

ОЗН. Потік з *відсутністю післядії* — той, у якому ймовірність появи k подій протягом $[s; t]$ не залежить від того, скільки подій і як з'являлись до моменту s . За будь-яких припущень про події до цього інтервалу умовний розподіл кількості подій на ньому збігається з безумовним.

Завдяки стаціонарності та відсутності післядії можемо говорити про $P_k(h)$ — ймовірність появи k подій протягом проміжку часу тривалістю h .

ОЗН. *Ординарність* потоку виражає неможливість появи двох (чи більше) подій в один момент часу: позначивши $P_{>1}(h)$ ймовірність появи більше 1 події на проміжку тривалістю h , вимагаємо $P_{>1}(h) = o(h)$ при $h \rightarrow 0$.

ОЗН. Потік, що задовольняє ці 3 властивості — *найпростіший* (НП).

ТВ. Для НП $P_1(h) = \lambda h + o(h)$, $h \rightarrow 0$, де константа $\lambda > 0$. (*)

◀ Нехай $P_0(1) = \theta$. Розіб'ємо проміжок часу Δ довжини 1 на n проміжків тривалістю $\frac{1}{n}$. Жодної події на Δ не з'явилося **тоді** жодної події не з'явилося на кожному з них. Внаслідок стаціонарності та відсутності післядії $\theta = P_0^n(\frac{1}{n}) \Rightarrow P_0(\frac{1}{n}) = \theta^{\frac{1}{n}}$. Тому для $k \in \mathbb{N}$: $P_0(\frac{k}{n}) = \theta^{\frac{k}{n}}$.

При фіксованому n для довільного $t \geq 0$ число $k = [nt] + 1$ забезпечує $\frac{k-1}{n} \leq t < \frac{k}{n}$. Оскільки $P_0(t)$ незростаюча ($P_0(t + \delta) = P_0(t)P_0(\delta) \leq P_0(t)$), то $P_0(\frac{k-1}{n}) \geq P_0(t) \geq P_0(\frac{k}{n}) \Leftrightarrow \theta^{\frac{k-1}{n}} \geq P_0(t) \geq \theta^{\frac{k}{n}}$.

При $n \rightarrow \infty$ $\mathbb{Q}$ -наближення $\frac{k-1}{n} \rightarrow t$, $\frac{k}{n} \rightarrow t$, так що $P_0(t) = \theta^t$.

$P_0(t) \in [0; 1]$, тому $\theta \in [0; 1]$. Відкинемо «крайні» $\theta = 0$ і $\theta = 1$ (при $\theta = 0$ протягом довільного часу м.н. відбувається нескінченно багато подій, а при $\theta = 1$ м.н. не з'являється жодної), тоді для $\lambda = -\ln \theta$: $P_0(t) = e^{-\lambda t}$.

$P_0(h) + P_1(h) + P_{>1}(h) = 1$. Завдяки ординарності та визначній границі $\frac{e^x - 1}{x} \xrightarrow{x \rightarrow 0} 1 \Leftrightarrow e^x = 1 + x + o(x)$, $x \rightarrow 0$:

$$\begin{aligned} P_1(h) &= 1 - P_0(h) - P_{>1}(h) = 1 - e^{-\lambda h} - o(h) = \\ &= 1 - (1 - \lambda h + o(h)) + o(h) = \lambda h + o(h), \quad h \rightarrow 0 \blacktriangleright \end{aligned}$$

Визначимо $P_m(t+h)$ для $m \in \mathbb{N}$. Відповідна подія є сумою $m+1$ несумісних подій «на проміжку t відбулось k подій, на проміжку $h - (m - k)$ ». Враховуючи відсутність післядії та стаціонарність, $P_m(t+h) = \sum_{k=0}^m P_k(t)P_{m-k}(h)$.

Нехай $R_m = \sum_{k=0}^{m-2} P_k(t)P_{m-k}(h)$. Оскільки $0 \leq P_k(t) \leq 1$,

$$R_m \leq \sum_{k=0}^{m-2} P_{m-k}(h) = \sum_{s=2}^m P_s(h) \leq \sum_{s=2}^{\infty} P_s(h) = P_{>1}(h)$$

Згідно ординарності і $R_m = o(h)$, $h \rightarrow 0$, так що $P_m(t+h) = P_m(t)P_0(h) + P_{m-1}(t)P_1(h) + o(h)$. $P_0(h) = 1 - P_1(h) - \sum_{s=2}^{\infty} P_s(h)$, тож згідно ординарності та (*) маємо: $P_0(h) = 1 - (\lambda h + o(h)) - o(h) = 1 - \lambda h + o(h)$, $h \rightarrow 0$ (**) і ще раз застосувавши (*): $P_m(t+h) = P_m(t)(1 - \lambda h) + P_{m-1}(t)\lambda h + o(h)$, $h \rightarrow 0$, звідки $\frac{P_m(t+h)-P_m(t)}{h} = -\lambda P_m(t) + \lambda P_{m-1}(t) + o(1)$, $h \rightarrow 0$.

Так що, спрямувавши $h \rightarrow 0$, оскільки існує границя правої частини:

$$\frac{dP_m(t)}{dt} = -\lambda P_m(t) + \lambda P_{m-1}(t)$$

Прийшли до нескінченної послідовності диференціальних рівнянь (ДР). Додамо ще 1 ДР відносно $P_0(t)$. Вище відмічали, що $P_0(t+h) = P_0(t)P_0(h)$; внаслідок (**) $P_0(t+h) = P_0(t)(1 - \lambda h + o(h))$, $h \rightarrow 0$. Аналогічно приходимо до $\frac{dP_0(t)}{dt} = -\lambda P_0(t)$. Легко бачити, що це ДР має розв'язок $P_0(t) = Ce^{-\lambda t}$. Із (**) випливає $P_0(0) = 1$, отже $C = 1$ і $P_0(t) = e^{-\lambda t}$.

Послідовно підставлятимемо $P_m(t)$ у ДР $\frac{dP_{m+1}(t)}{dt} = -\lambda P_{m+1}(t) + \lambda P_m(t)$, аби визначити $P_{m+1}(t)$. Увівши заміну $P_m(t) = e^{-\lambda t}q_m(t)$: $-\lambda e^{-\lambda t}q_{m+1}(t) + e^{-\lambda t}q'_{m+1}(t) = -\lambda e^{-\lambda t}q_{m+1}(t) + \lambda e^{-\lambda t}q_m(t) \Leftrightarrow q'_{m+1}(t) = \lambda q_m(t)$ з урахуванням $q_0(t) \equiv 1$, $q_m(0) = 0$ при $m \geq 1$ отримаємо $q_m(t) = \frac{(\lambda t)^m}{m!}$, і нарешті $P_m(t) = \frac{(\lambda t)^m}{m!}e^{-\lambda t}$ для $m \in \mathbb{Z}_+$ — потік є пуассоновим.

І навпаки, ПП є стаціонарним, з відсутністю післядії та ординарним. Остання властивість випливає з того, що $P_{>1}(h) = P\{N[t; t+h] > 1\} = 1 - P\{N[t; t+h] = 0\} - P\{N[t; t+h] = 1\} = 1 - e^{-\lambda h} - \lambda h e^{-\lambda h} = 1 - (1 - \lambda h + o(h))(1 + \lambda h) = \lambda^2 h^2 + o(h) = o(h)$, $h \rightarrow 0$

У той же час спостереження різноманітних реальних потоків виявляють відхилення від найпростішого типу, навіть при детальнішому розгляді тих же явищ, де він є основною моделлю.

2. Марківські СМО

Це найпростіший і найбільш досліджений тип СМО, з аналізу якого почався розвиток ТМО.

2.1. Марківські процеси

ОЗН. ВП $\xi(t)$, $t \in T$ (де $T = \mathbb{Z}_+$ у дискретному або $T = [0; +\infty)$ у неперервному випадку) називається *марківським* процесом (МП), якщо

$$\forall \{t_i\}_{i=1}^n, \text{ де } t_i \in T, t_i < t_{i+1}, \text{ і } \forall \{x_i\}_{i=1}^n, \text{ де } x_i \in \mathbb{R}:$$

$$P\{\xi(t_n) < x_n / \xi(t_1) = x_1, \dots, \xi(t_{n-1}) = x_{n-1}\} = P\{\xi(t_n) < x_n / \xi(t_{n-1}) = x_{n-1}\}$$

— розподіл майбутнього залежить від теперішнього, але не від минулого.

У випадку **дискретного часу і дискретної множини станів** говорять про *ланцюг Маркова* (ЛМ), позначають також ξ_t , де $t = 0, 1, \dots$, а стани часто індексують, як і моменти часу, $\mathbb{Z}_+$ -числами. $P\{\xi_t = j / \xi_{t-1} = i\}$ називають *однокроковими ймовірностями переходу*, або просто *перехідними ймовірностями*. Якщо вони не залежать від часу t , $P\{\xi_t = j / \xi_{t-1} = i\} = p_{ij}$, то ланцюг є *однорідним*, при цьому p_{ij} утворюють *перехідну матрицю* P . В однорідного ланцюга n -крокові *перехідні ймовірності*, $P\{\xi_{t+n} = j / \xi_t = i\} = p_{ij}^{(n)}$, теж не залежать від t і утворюють матрицю $P^{(n)}$.

Безумовні ймовірності перебування в стані j у момент n , $P\{\xi_n = j\}$, позначатимемо вектор-рядком $\pi^{(n)} = \{\pi_j^{(n)}\}_j$. Початковий розподіл — $\pi^{(0)}$.

З адитивності ймовірності $p_{ij}^{(m)} = \sum_r p_{ir}^{(m-k)} p_{rj}^{(k)}$ для $0 < k < m$, або $P^{(m)} = P^{(m-k)} P^{(k)}$ — матрична форма рівнянь Колмогорова-Чепмена цього процесу.

Зокрема, $P^{(m)} = P P^{(m-1)} = P P P^{(m-2)} = \dots = P^m$.

Розподіл стану процесу на m -му кроці визначається розподілом на $(m-1)$ -му та перехідною матрицею: $\pi_j^{(m)} = \sum_i \pi_i^{(m-1)} p_{ij}$, або $\pi^{(m)} = \pi^{(m-1)} P$ (звідки $\pi^{(m)} = \pi^{(0)} P^m$). Позначивши одиничну матрицю E та увівши до розгляду матрицю $Q = P - E$, перепишемо цю рівність так: $\pi^{(m)} - \pi^{(m-1)} = \pi^{(m-1)} Q$.

P є *стохастичною* матрицею (сума коефіцієнтів у рядку рівна 1), а в Q така сума є 0. Відмітимо, що ми не вимагали скінченності множини станів.

Для МП з **неперервним часом та дискретною множиною станів**, якщо $p_{ij}(s; t) = P\{\xi(t) = j / \xi(s) = i\}$, рівняння Колмогорова-Чепмена

$$p_{ij}(s; t) = \sum_r p_{ir}(s; u) p_{rj}(u; t) \text{ для } u \in (s; t)$$

або, оскільки для кожної пари $(s; t)$ ймовірності $p_{ij}(s; t)$ теж утворюють матрицю $P(s; t)$, це можна записати матрично: $P(s; t) = P(s; u) P(u; t)$.

Аналогічно виникає поняття ймовірності перебування процесу в i -му стані у момент t , $p_i(t)$. Усі такі ймовірності утворюють розподіл $p(t)$ (вектор-рядок), $p(0)$ — розподіл початкового стану. Зв'язок між розподілами у послідовні моменти $s < t$: $p(t) = p(s)P(s; t)$, або через тривалість $\Delta t > 0$: $p(t + \Delta t) = p(t)P(t; t + \Delta t)$, тобто

$$p_i(t + \Delta t) = \sum_r p_r(t) p_{ri}(t; t + \Delta t)$$

Інфінітезімальний генератор. Якщо $\exists q_i(t), q_{ij}(t) \in C_{[0;+\infty)}$ такі, що

$$\begin{cases} P\{\xi(t + \Delta t) \neq \xi(t) / \xi(t) = i\} = 1 - p_{ii}(t; t + \Delta t) = q_i(t)\Delta t + o(\Delta t), \\ p_{ij}(t; t + \Delta t) = q_{ij}(t)\Delta t + o(\Delta t), \quad j \neq i \end{cases} \quad (*)$$

$$\text{тобто} \begin{cases} \frac{p_{ii}(t; t + \Delta t) - 1}{\Delta t} = -q_i(t) + o(1), \\ \frac{p_{ij}(t; t + \Delta t)}{\Delta t} = q_{ij}(t) + o(1), \quad i \neq j \end{cases} \quad (*)'$$

(де $o(\cdot)$ — відносно $\Delta t \rightarrow 0$), то за деяких додаткових припущень про регулярність з рівнянь Колмогорова-Чепмена випливає:

$$\begin{aligned} \frac{\partial}{\partial t} p_{ij}(s; t) &= -p_{ij}(s; t)q_j(t) + \sum_{r \neq j} p_{ir}(s; t)q_{rj}(t), \\ \frac{\partial}{\partial s} p_{ij}(s; t) &= q_i(s)p_{ij}(s; t) - \sum_{r \neq i} q_{ir}(s)p_{rj}(s; t) \end{aligned}$$

Це *пряме та зворотнє рівняння Колмогорова*. Розглянемо перше при $s = 0$ для *однорідного* процесу, коли $q_i(t) \equiv q_i$ та $q_{ij}(t) \equiv q_{ij}$ Н/З від часу:

$$\frac{d}{dt} p_{ij}(0; t) = -q_j p_{ij}(0; t) + \sum_{r \neq j} p_{ir}(0; t)q_{rj}$$

Домножимо на $p_i(0)$ та просумуємо по усіх i . Оскільки $p(t) = p(0)P(0; t)$,

$$\frac{d}{dt} p_j(t) = -q_j p_j(t) + \sum_{r \neq j} p_r(t)q_{rj} \text{ або матрично } p'(t) = p(t)Q, \text{ де}$$

$$p(t) = \|p_j(t)\|_j, \quad p'(t) = \left\| \frac{d}{dt} p_j(t) \right\|_j, \quad Q = \begin{pmatrix} -q_0 & q_{01} & q_{02} & \dots \\ q_{10} & -q_1 & q_{12} & \dots \\ q_{20} & q_{21} & -q_2 & \dots \\ \dots & \dots & \dots & \ddots \end{pmatrix}$$

Так як $\sum_j p_{ij}(t; t + \Delta t) = p_{ii}(t; t + \Delta t) + \sum_{j \neq i} p_{ij}(t; t + \Delta t) = 1$, то з (*)

$$[1 - q_i \Delta t + o(\Delta t)] + \sum_{j \neq i} (q_{ij} \Delta t + o(\Delta t)) = 1 \Leftrightarrow -q_i + o(1) + \sum_{j \neq i} (q_{ij} + o(1)) = 0$$

отже, $q_i = \sum_{j \neq i} q_{ij}$, тобто сума в кожному рядку матриці Q рівна 0.

Враховуючи (*'), $Q = \lim_{\Delta t \rightarrow 0} \frac{P(t; t + \Delta t) - E}{\Delta t}$. Вона відіграє, щодо процесів Маркова з неперервним часом, роль подібну до $P - E$ для процесів Маркова з дискретним часом. Q називається *матрицею інтенсивностей*, або *інфінітезімальним генератором* (ІГ) процесу Маркова з неперервним часом.

ТВ. ПП $N(t)$ є МП з неперервним часом і станами 0, 1, 2, ...

◀ Дійсно, узявши $t_1 < t_2 < \dots < t_n$ та $x_k \in \mathbb{Z}_+$, $x_1 \leq x_2 \leq \dots \leq x_n$:

$$P\{N(t_n) < x_n/N(t_1) = x_1, \dots, N(t_{n-1}) = x_{n-1}\} =$$

$$= P\{N[t_{n-1}; t_n] < x_n - x_{n-1}/N[t_{n-2}; t_{n-1}] = x_{n-1} - x_{n-2}, \dots, N[0; t_1] = x_1\}$$
що внаслідок Н/З приростів дорівнює $P\{N[t_{n-1}; t_n] < x_n - x_{n-1}\} =$

$$= P\{N[t_{n-1}; t_n] < x_n - x_{n-1}/N[0; t_{n-1}] = x_{n-1}\} =$$

$$= P\{N(t_n) < x_n/N(t_{n-1}) = x_{n-1}\} \blacktriangleright$$

Визначимо ІГ для ПП з інтенсивністю λ :

$$q_i = -\lim_{h \rightarrow 0} \frac{p_{ii}(t; t+h) - 1}{h} = -\lim_{h \rightarrow 0} \frac{P\{N[t; t+h]=0\} - 1}{h} = -\lim_{h \rightarrow 0} \frac{e^{-\lambda h} - 1}{h} \stackrel{\text{III Б.Г.}}{=} \lambda$$

$$q_{ij} = \lim_{h \rightarrow 0} \frac{p_{ij}(t; t+h)}{h} = \begin{cases} 0, & j < i, \\ \lim_{h \rightarrow 0} \frac{N[t; t+h]=1}{h} = \lim_{h \rightarrow 0} \frac{\lambda h e^{-\lambda h}}{h} = \lambda, & j = i + 1, \\ \lim_{h \rightarrow 0} \frac{N[t; t+h]=j-i}{h} = \lim_{h \rightarrow 0} \frac{\lambda^{j-i} e^{-\lambda h}}{(j-i)!} h^{j-i-1} = 0, & j > i + 1 \end{cases}$$

Отже, $Q = \begin{pmatrix} -\lambda & \lambda & 0 & 0 & \dots \\ 0 & -\lambda & \lambda & 0 & \dots \\ 0 & 0 & -\lambda & \lambda & \dots \\ \dots & \dots & \dots & \dots & \ddots \end{pmatrix}$, і $\frac{d}{dt} p_j(t) = -\lambda p_j(t) + \lambda p_{j-1}(t)$ — вже

зустрічали це ДР, коли доводили пуассоновість найпростішого потоку.

Розглянемо МП $N(t)$ з неперервним часом і $\mathbb{Z}_+$ -значеннями, що має *властивість народження та загибелі*:

$$\begin{cases} P\{|N(t + \Delta t) - N(t)| > 1\} = o(\Delta t), \\ P\{N(t + \Delta t) = n + 1 / N(t) = n\} = \lambda_n \Delta t + o(\Delta t), \quad n \geq 0, \\ P\{N(t + \Delta t) = n - 1 / N(t) = n\} = \mu_n \Delta t + o(\Delta t), \quad n \geq 1 \quad (\mu_0 = 0). \end{cases}$$

Перша вимога означає, що на нескінченно малому інтервалі часу приріст може бути лише -1 , $+1$ або 0 . У n -му стані розподіл часу до інкременту $(+1)$ та декременту (-1) — показниковий, відповідно з параметрами λ_n та μ_n .

У такого ВП $p_{nn}(t; t + \Delta t) = 1 - P\{N(t + \Delta t) = n \pm 1 / N(t) = n\} -$
 $- P\{|N[t; t + \Delta t]| > 1\} = 1 - (\lambda_n + \mu_n) \Delta t + o(\Delta t)$, тож

$q_0 = \lambda_0$, $q_n = \lambda_n + \mu_n$, $q_{n,n+1} = \lambda_n$, $q_{n,n-1} = \mu_n$, а $q_{n,n+k} = 0$ при $k \notin \{-1, 0, 1\}$

$$\text{ІГ } Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & \dots \\ \dots & \dots & \dots & \dots & \ddots \end{pmatrix}, \text{ і система ДР для такого}$$

процесу: $\begin{cases} \frac{d}{dt} p_0(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t), \\ \frac{d}{dt} p_n(t) = -(\lambda_n + \mu_n) p_n(t) + \lambda_{n-1} p_{n-1}(t) + \mu_{n+1} p_{n+1}(t), \quad n \geq 1. \end{cases}$

При $\mu_n \equiv 0$ маємо *процес чистого народження*; коли додатково $\lambda_n \equiv \lambda$ — пуассонів потік інтенсивності λ .

Вкладені ЛМ. Беручи стан МП з неперервним часом у певні моменти цього часу, матимемо *вкладений* ЛМ уже з дискретним часом. Як приклад, розглянемо описаний вище процес народження й загибелі у моменти зміни стану. Відповідний ЛМ має матрицю перехідних ймовірностей $P = \|p_{ij}\|$, де

$$p_{i,i+1} = \frac{\lambda_i}{\lambda_i + \mu_i} \text{ й } p_{i,i-1} = \frac{\mu_i}{\lambda_i + \mu_i} \text{ при } i \geq 1, p_{0,1} = 1, \text{ усі інші } p_{ij} = 0$$

(впливає з того, що для Н/З ВВ ξ та η , показниково розподілених з параметрами λ та μ , $P\{\xi < \eta\} = \frac{\lambda}{\lambda + \mu}$).

Граничні та стаціонарні розподіли. Ергодичність. Розглянемо ЛМ. Нехай $\exists \lim_{m \rightarrow \infty} p_{ij}^{(m)} = \pi_j$ при всіх i , тобто після достатньо тривалого часу ймовірність перебування процесу в стані j , якщо спочатку він був у стані i , не залежить від i . У матричній формі це означає $\exists \lim_{m \rightarrow \infty} P^{(m)}$ (а $P^{(m)} = P^m$) у тому сенсі, що усі рядки P^m стають однаковими з плином часу. $\{\pi_j\}_j$ називаються *граничними ймовірностями* даного ланцюга.

Безумовні ймовірності станів на m -му кроці, $\pi^{(m)} = \pi^{(0)} P^m$. Отже,

$$\lim_{m \rightarrow \infty} \pi_j^{(m)} = \lim_{m \rightarrow \infty} \sum_i \pi_i^{(0)} p_{ij}^{(m)} = \sum_i \pi_i^{(0)} \lim_{m \rightarrow \infty} p_{ij}^{(m)} = \sum_i \pi_i^{(0)} \pi_j = \pi_j \sum_i \pi_i^{(0)} = \pi_j$$
— ті самі, що й у $p_{ij}^{(m)}$, Н/З від початкового розподілу. Коли $\lim_{m \rightarrow \infty} \pi_j^{(m)}$ існують, їх можна знайти з рівності $\pi^{(m)} = \pi^{(m-1)} P \Rightarrow \lim_{m \rightarrow \infty} \pi^{(m)} = \lim_{m \rightarrow \infty} \pi^{(m-1)} P$; позначаючи $\pi = \{\lim_{m \rightarrow \infty} \pi_j^{(m)}\}_j$, оскільки $\lim_{m \rightarrow \infty} \pi^{(m)} = \lim_{m \rightarrow \infty} \pi^{(m-1)} = \pi$, маємо: $\pi = \pi P \Leftrightarrow \pi Q = \theta$. Додаючи до цих рівнянь ($\sum_j \pi_j q_{jk} = 0$ для усіх k) умову $\sum_j \pi_j = 1$, та розв'язавши отриману систему, знаходимо $\{\pi_j\}$.

Такі рівняння називаються *рівняннями стаціонарності*, а розв'язок — *стаціонарним розподілом*.

Коли граничний розподіл існує, він співпадає із стаціонарним. У загальному випадку рівняння $\pi Q = \theta$ може мати розв'язки навіть тоді, коли граничний розподіл $\{\lim_{m \rightarrow \infty} p_{ij}^{(m)}\}_j$ не існує (а може й не мати).

У МП з неперервним часом стаціонарний розподіл $p = \|p_j\| = \|\lim_{t \rightarrow \infty} p_j(t)\|$ визначається з рівнянь $pQ = \theta$, $\sum_j p_j = 1$, де Q — ІГ. Але для таких МП виникають складніші задачі, пов'язані з (нескінченними) системами ДР.

Під *ергодичністю* ВП розумітимемо можливість обчислення його статистичних характеристик за одною нескінченно тривалою реалізацією. Інакше кажучи, $\xi(t)$ ергодичний, коли середні за часом рівні середнім за траєкторіями. Оскільки середні — це моменти, котрих багато, уточнимо цю вимогу.

Нехай $\xi_0(t), \xi_1(t), \dots$ — траєкторії ВП $\xi(t)$.

k -й момент за часом $\in \overline{\xi^k(T)} = \frac{1}{T} \int_0^T \xi_0^k(t) dt$, а k -й момент за траєкторіями

в момент $t \in m_k(t) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_i^k(t)$.

ОЗН. $\xi(t)$ ергодичний по k -му моменту, якщо $\lim_{T \rightarrow \infty} \overline{\xi^k(T)} = \lim_{t \rightarrow \infty} m_k(t)$; (цілком) ергодичний, якщо він ергодичний по усіх моментах.

2.2. Процес народження й загибелі

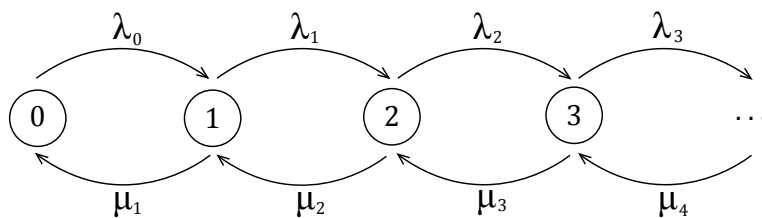
ОЗН. Процес народження й загибелі (ПНЗ) — ВП $N(t)$, $t \geq 0$, який:

- 1) приймає невід'ємні цілі значення;
- 2) перебуває в стані k протягом часу, показниково розподіленого з параметром α_k , незалежним від попередньої траєкторії процесу;
- 3) незалежно від попередньої траєкторії переходить зі стану k або у стан $k + 1$ з ймовірністю p_k , або у стан $k - 1$ з ймовірністю $q_k = 1 - p_k$.

ПНЗ є однорідним МП з неперервним часом і зліченною множиною станів.

Визначимо α_k і p_k через λ_k та μ_k : $\alpha_k = \lambda_k + \mu_k$, $p_k = \frac{\lambda_k}{\lambda_k + \mu_k}$, тож $q_k = \frac{\mu_k}{\lambda_k + \mu_k}$. λ_k та μ_k називаються інтенсивностями народження й загибелі: коли процес перебуває в стані k , розподіл часу до наступного народження є показниковим з параметром λ_k , а до наступної загибелі — з μ_k . Стан процесу ін/декрементується залежно від того, яка з цих подій відбудеться раніше.

Діаграма інтенсивностей переходів між станами виглядає так:



Нехай $p_k = \lim_{t \rightarrow \infty} p_k(t)$ є ймовірністю перебування процесу в k -му стані, Н/З від часу, тобто для великих t . Відомо, що за умов $\mu_i > \lambda_i$ для $i \geq i_0$ та $\frac{1}{\lambda_i + \mu_i} \leq K$ ПНЗ є ергодичним і має граничний розподіл (говорять, що він досягає статистичної рівноваги); за таких умов p_k існують. Визначити їх можна з рівняння $pQ = \theta$:

$$\begin{cases} -\lambda_0 p_0 + \mu_1 p_1 = 0, \\ -(\lambda_k + \mu_k) p_k + \lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1} = 0, \quad k \geq 1 \end{cases}$$

Ті ж рівняння можна отримати з міркувань *рівноваги/балансу потоку*: в стаціонарному режимі потік переходів у даний стан повинен бути таким же, як потік переходів з нього. При $k \geq 1$ процес може перейти в стан k із станів $(k-1)$ (з інтенсивністю λ_{k-1}) та $(k+1)$ (з інтенсивністю μ_{k+1}), а вийти з нього — у ті ж 2 стани, тільки з інтенсивностями λ_k і μ_k . Домножуючи інтенсивності на ймовірності перебування у відповідних станах, матимемо потоки:

$$\begin{cases} \mu_1 p_1 = \lambda_0 p_0, \\ (\lambda_k + \mu_k) p_k = \lambda_{k-1} p_{k-1} + \mu_{k+1} p_{k+1}, \quad k \geq 1 \end{cases} \Leftrightarrow \begin{cases} p_1 = \frac{\lambda_0}{\mu_1} p_0, \\ p_{k+1} = \frac{\lambda_k + \mu_k}{\mu_{k+1}} p_k - \frac{\lambda_{k-1}}{\mu_{k+1}} p_{k-1}, \quad k \geq 1 \end{cases}$$

При $k = 1$: $p_2 = \frac{\lambda_1 + \mu_1}{\mu_2} \cdot \frac{\lambda_0}{\mu_1} p_0 - \frac{\lambda_0}{\mu_2} p_0 = \left(\frac{\lambda_1 + \mu_1}{\mu_1} - 1 \right) \frac{\lambda_0}{\mu_2} p_0 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} p_0$.

ТВ. $\forall k \in \mathbb{N}$: $p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i}$ (*)

◀ Застосуємо індукцію по k . Базис, для $k = 1$ та $k = 2$, має місце. Якщо твердження вже доведене для $1, 2, \dots, k$, то $p_{k+1} =$

$$= \frac{\lambda_k + \mu_k}{\mu_{k+1}} p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} - \frac{\lambda_{k-1}}{\mu_{k+1}} p_0 \prod_{i=1}^{k-1} \frac{\lambda_{i-1}}{\mu_i} = p_0 \frac{\lambda_{k-1}}{\mu_{k+1}} \prod_{i=1}^{k-1} \frac{\lambda_{i-1}}{\mu_i} \cdot \left(\frac{\lambda_k + \mu_k}{\mu_k} - 1 \right) = p_0 \prod_{i=1}^{k+1} \frac{\lambda_{i-1}}{\mu_i} \blacktriangleright$$

Оскільки $1 = \sum_{k=0}^{\infty} p_k = p_0 \left(1 + \sum_{k=1}^{\infty} \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} \right)$, то $p_0 = \left(1 + \sum_{k=1}^{\infty} \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} \right)^{-1}$.

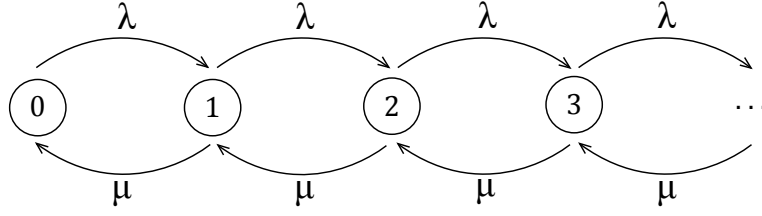
Також це означає, що необхідною та достатньою умовою існування стаціонарного режиму є збіжність ряду $\sum_{k=1}^{\infty} \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i}$.

Отримати p_k можна ще з *локальних* рівнянь балансу, які описують потік між сусідніми станами ($k-1$ та k), на відміну від розглядуваних тільки що *глобальних* рівнянь. У стаціонарному режимі крізь умовну межу між станами $k-1$ та k потік в один бік рівний потоку в інший. Потік в більший бік вичерпується потоком з $k-1$ до k , рівним $\lambda_{k-1} p_{k-1}$; потік у менший бік є $\mu_k p_k$. Отже, $\lambda_{k-1} p_{k-1} = \mu_k p_k \Rightarrow p_k = \frac{\lambda_{k-1}}{\mu_k} p_{k-1}$, звідки випливає (*).

2.3. Система М/М/1

Ми розглядаємо цю СМО у стаціонарному режимі. Проміжки між прибуттями запитів розподілені показниково з параметром λ , тривалості їх обробки — теж показниково, з параметром μ .

Моделлю буде ПНЗ $N(t)$, де величина $N(t) \in \mathbb{Z}_+$ позначає загальну кількість запитів у системі в момент t , «народження» відповідає прибуттю запиту до системи, а «загибель» — вибуттю.



$\lambda_k \equiv \lambda$, $\mu_k \equiv \mu$ — інтенсивності не залежать від кількості запитів у системі. Рівняння балансу потоку:

$$\begin{cases} \lambda p_0 = \mu p_1, \\ (\lambda + \mu)p_k = \lambda p_{k-1} + \mu p_{k+1}, \quad k \geq 1 \end{cases} \Leftrightarrow \begin{cases} p_1 = \frac{\lambda}{\mu} p_0, \\ p_{k+1} = (1 + \frac{\lambda}{\mu})p_k - \frac{\lambda}{\mu} p_{k-1}, \quad k \geq 1 \end{cases}$$

Розв'яжемо ці рівняння — знайдемо $\{p_k\}$ — 3-ма способами, аби проілюструвати їх використання. Щодо М/М/1 вони майже однаково ефективні; для інших систем їхня ефективність відносно один одного буває різною.

Ітераційний метод для ПНЗ описаний вище, отримали $p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i}$.

У даному випадку $p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k$, де $p_0 = \left[\sum_{k=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^k\right]^{-1} = 1 - \frac{\lambda}{\mu}$, коли $\lambda < \mu$; при $\lambda \geq \mu$ ряд розбігається. Це ще одне підтвердження того, що для існування стаціонарного режиму потрібно $\lambda < \mu$ (навіть якщо $\lambda = \mu$, кількість запитів у системі м.н. необмежено зростатиме).

Значить, кількість запитів у системі М/М/1 в стаціонарному режимі має геометричний розподіл з параметром $p_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$, де ρ — означений раніше коефіцієнт використання (для систем з 1 приладом він рівний $\frac{\lambda}{\mu}$):

$$p_k = (1 - \rho)\rho^k, \quad k \in \mathbb{Z}_+$$

$$\text{Середня кількість запитів, } \bar{N} = \frac{1-p_0}{p_0} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

Метод твірних функцій. Розглянемо $P(z) = \sum_{k=0}^{\infty} p_k z^k$, де $z \in \mathbb{C}$ і $|z| \leq 1$.

$$\text{Рівняння балансу: } \begin{cases} p_1 = \rho p_0, \\ p_{k+1} = (1 + \rho)p_k - \rho p_{k-1}, \quad k \geq 1. \end{cases}$$

Домножимо II рівняння на z^k і представимо $z^k = z^{-1} \cdot z^{k+1} = z \cdot z^{k-1}$:

$$z^{-1} p_{k+1} z^{k+1} = (1 + \rho) p_k z^k - \rho p_{k-1} z^{k-1}$$

Просумуємо по $k \in \mathbb{N}$:

$$\begin{aligned} z^{-1} \sum_{k=1}^{\infty} p_{k+1} z^{k+1} &= (1 + \rho) \sum_{k=1}^{\infty} p_k z^k - \rho z \sum_{k=1}^{\infty} p_{k-1} z^{k-1} \Leftrightarrow \\ \Leftrightarrow z^{-1} [P(z) - p_0 - p_1 z] &= (1 + \rho) [P(z) - p_0] - \rho z P(z) \stackrel{I}{\Leftrightarrow} \\ \Leftrightarrow P(z) - p_0(1 + \rho z) &= (1 + \rho) z [P(z) - p_0] - \rho z^2 P(z) \Leftrightarrow \\ \Leftrightarrow P(z) [1 - z - \rho z + \rho z^2] &= p_0(1 + \rho z) - p_0(z + \rho z) \Leftrightarrow \\ \Leftrightarrow P(z)(1 - z)(1 - \rho z) &= p_0(1 - z) \Leftrightarrow P(z) = \frac{p_0}{1 - \rho z} \end{aligned}$$

Оскільки $\frac{p_0}{1-\rho} = P(1) = \sum_{k=0}^{\infty} p_k = 1$, отримуємо $p_0 = 1 - \rho$ (звідки $\rho < 1$), і

$$P(z) = \frac{1-\rho}{1-\rho z} \text{ при } \rho < 1, |z| \leq 1$$

Зважаючи на $|\rho z| < 1$, розкладемо $\frac{1-\rho}{1-\rho z}$ у ряд як суму геометричної прогресії: $P(z) = \sum_{k=0}^{\infty} (1-\rho)\rho^k z^k$, так що $p_k = (1-\rho)\rho^k$.

ЗАУВ. Інколи виразити $P(z)$ у замкненій формі легше, ніж знайти p_k .

Цей підхід теж дозволяє знайти середню кількість запитів:

$$\bar{N} = \sum_{k=1}^{\infty} k p_k = \sum_{k=1}^{\infty} k p_k z^{k-1} \Big|_{z=1} = \frac{dP(z)}{dz} \Big|_{z=1} = \frac{\rho(1-\rho)}{(1-\rho z)^2} \Big|_{z=1} = \frac{\rho}{1-\rho}$$

Метод лінійних різницьових рівнянь. Рівняння $p_{k+1} = (1+\rho)p_k - \rho p_{k-1}$ є *лінійним різницьовим*, — значення p_{k+1} виражене через p_{k-1} і p_k . Це подібно лінійним ДР, лише замість неперервного x в $y(x)$ береться дискретне k в p_k . Загальна схема розв'язку подібна до розв'язку лінійних ДР; опишемо її.

Нехай $(a_0; a_1; \dots)$ — деяка послідовність. Визначимо лінійний оператор D на її елементах: $Da_k := a_{k+1}$. Тоді $D^m a_k = a_{k+m}$.

Розглянемо лінійне різницьове рівняння (РР) зі сталими коефіцієнтами:

$$\begin{aligned} C_0 a_n + C_1 a_{n+1} + \dots + C_k a_{n+k} &= 0 \Leftrightarrow \\ \Leftrightarrow \sum_{i=0}^k C_i (D^i a_n) &= 0 \Leftrightarrow \left(\sum_{i=0}^k C_i D^i \right) a_n = 0 \end{aligned}$$

Поліном у дужках визначає *характеристичне рівняння* (ХР) для даного РР: $\sum_{i=0}^k C_i D^i = 0$. Корені ХР, у свою чергу, визначають загальний розв'язок РР. Наприклад, якщо ХР квадратне з різними $\mathbb{R}$ -коренями r_1 і r_2 , то загальним розв'язком РР буде $a_n = d_1 r_1^n + d_2 r_2^n$, де d_i — довільні константи.

Для РР стаціонарного режиму $p_{k+1} - (1+\rho)p_k + \rho p_{k-1} = 0$ відповідне ХР $D^2 - (1+\rho)D + \rho = 0$ має корені $D = 1$ та $D = \rho$; загальний розв'язок — $p_n = d_1 + d_2 \rho^n$. Лишилося визначити d_1 та d_2 . Візьмемо «граничні» умови $p_1 = \rho p_0$ та $\sum_{k=0}^{\infty} p_k = 1$. Щоб ряд збігався, необхідно $d_1 = 0$. З першої умови $p_0 = d_2$, і $1 = \sum_{k=0}^{\infty} p_0 \rho^k = \frac{p_0}{1-\rho} \Rightarrow p_0 = 1 - \rho$. Отже, $p_k = (1 - \rho)\rho^k$.

Визначимо (статистичні) **характеристики ефективності** даної СМО.

Середня кількість запитів у системі знайдена вище: $\bar{N} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$ (МС ВВ, яка має геометричний розподіл з параметром $1 - \rho$).

Середня кількість запитів у черзі/довжина черги $\overline{N}_q$. Оскільки $N_q(t) = \max\{0; N(t) - 1\}$, то

$$\overline{N}_q = \sum_{k=1}^{\infty} (k-1)p_k = \sum_{k=1}^{\infty} kp_k - \left[\sum_{k=0}^{\infty} p_k - p_0 \right] = \overline{N} - (1 - p_0) = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$

ЗАУВ. Рівність $\overline{N}_q = \overline{N} - (1 - p_0)$ має місце для довільної СМО з єдиним приладом, котрий у кожний момент часу здатний обробляти рівно 1 запит.

Середня довжина непорожньої черги $\overline{N}'_q = M[N_q/N_q \neq 0] =$
 $\sum_{k=1}^{\infty} (k-1)p'_k$, де $\{p'_k\}$ — умовні ймовірності перебування системи у стані k ,
 коли черга непорожня $\Leftrightarrow N(t) \neq 0, 1$; при $k \geq 2$

$$p'_k = \frac{P\{N(t)=k\}}{P\{N(t) \geq 2\}} = \frac{p_k}{1 - P\{N(t)=0\} - P\{N(t)=1\}} = \frac{p_k}{1 - (1-\rho) - (1-\rho)\rho} = \frac{p_k}{\rho^2}$$

і $p'_0 = p'_1 = 0$. Звідси $\overline{N}'_q = \frac{1}{\rho^2} \sum_{k=2}^{\infty} (k-1)p_k = \frac{1}{\rho^2} \overline{N}_q = \frac{1}{1-\rho}$.

Середня тривалість перебування запиту в системі $\overline{T}$ отримується з формули Літтла: $\overline{N} = \lambda \overline{T} \Rightarrow \overline{T} = \frac{1}{\lambda} \overline{N} = \frac{1}{\mu - \lambda}$.

Середня тривалість перебування запиту в черзі $\overline{T}_q$ — аналогічно:

$$\overline{T}_q = \frac{1}{\lambda} \overline{N}_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

Видно, що при $\rho \rightarrow 1$ поведінка системи погіршується: намагання максимізувати продуктивність, тримаючи прилад завжди зайнятим, призводить до подовження затримки запитів у черзі.

Так як N має геометричний розподіл, $D[N] = \frac{\rho}{(1-\rho)^2}$. Тому й $D[N] \xrightarrow{\rho \rightarrow 1} \infty$ — при $\rho \rightarrow 1$ розкид кількості запитів у системі необмежено зростає, тобто передбачуваність її поведінки падає.

ЗАУВ. Наведені характеристики отримані для **стаціонарного** режиму.

Розподіл часу очікування у черзі T_q — ВВ, яка приймає певне значення для кожного запиту. Вже встановили $\overline{T}_q = \frac{\lambda}{\mu(\mu - \lambda)}$, тепер знайдемо її ФР $F_{T_q}(t) = P\{T_q < t\}$. Очевидно, $F_{T_q}(t) = 0$ для $t \leq 0$. T_q не є ані дискретною, ані АбсНеп: для строго додатніх значень її розподіл неперервний, але вона з ненульовою ймовірністю приймає значення 0, коли запит надходить до системи без інших запитів і відразу обробляється.

Нехай q_k , $k \in \mathbb{Z}_+$ — ймовірність події R_k : в момент прибуття запиту в системі вже є рівно k запитів. У загальному випадку $q_k \neq p_k$, але для ПП вони співпадають через рівномірність ПП: $q_k = p_k$.

$$\begin{aligned}
F_{T_q}(\{0\}) &= P\{T_q = 0\} = q_0 = p_0 = 1 - \rho. \text{ Для } t > 0 \\
F_{T_q}(t) &= P\{T_q < t\} = P\{T_q = 0\} + P\{0 < T_q < t\} = (1 - \rho) + P\{0 < T_q < t\} \\
&= P\{0 < T_q < t\} \stackrel{\text{ФПЙ}}{=} \sum_{k=1}^{\infty} p_k P\{0 < T_q < t/R_k\}
\end{aligned}$$

$\{0 < T_q < t/R_k\}$ означає, що протягом часу t всі k запитів, які були в системі у момент надходження даного (з них 1 оброблявся приладом, інші чекали у черзі), встигли обробитися, — сумарна тривалість їх обробки менше t . Оскільки тривалості обробки — Н/З ВВ, показниково розподілені з параметром μ , ця сумарна тривалість має розподіл Ерланга з параметрами $(k; \mu)$,

$$\text{щільність якого } p(k; \mu; u) = \begin{cases} 0, & u \leq 0, \\ \frac{\mu^k}{(k-1)!} u^{k-1} e^{-\mu u}, & u > 0. \end{cases}$$

$$\begin{aligned}
\text{Тому } F_{T_q}(t) &= 1 - \rho + (1 - \rho) \sum_{k=1}^{\infty} \rho^k \int_0^t \frac{\mu^k}{(k-1)!} u^{k-1} e^{-\mu u} du = \\
&= (1 - \rho) \left[1 + \int_0^t \mu \rho e^{-\mu u} \left[\sum_{k=1}^{\infty} \frac{(\mu \rho u)^{k-1}}{(k-1)!} \right] du \right] \stackrel{\mu \rho = \lambda}{=} (1 - \rho) \left[1 + \int_0^t \lambda e^{-(\mu-\lambda)u} du \right] = \\
&= (1 - \rho) \left[1 - \frac{\rho}{1-\rho} (e^{-(\mu-\lambda)t} - 1) \right] = 1 - \rho e^{-(\mu-\lambda)t}
\end{aligned}$$

Таким чином, T_q являє собою суму ВВ з показниковим розподілом із параметром $\mu(1-\rho) = \mu - \lambda$ (з коефіцієнтом ρ) та атомічного розподілу з масою $1 - \rho$ в т. $t = 0$. Звідси також випливає

$$\begin{aligned}
\overline{T_q} &= 0 \cdot (1 - \rho) + \rho \int_0^{\infty} \mu(1 - \rho) u e^{-\mu(1-\rho)u} du = \rho \int_0^{\infty} e^{-\mu(1-\rho)u} du = \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)} \\
&\text{— співпадає з середнім, отриманим за формулою Літтла.}
\end{aligned}$$

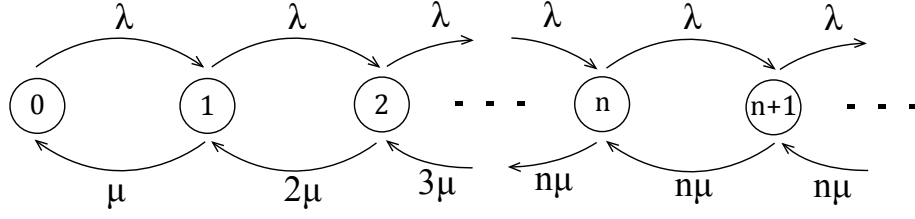
Аналогічно визначається **розподіл часу перебування в системі T** : міркуємо, як для T_q , тільки тепер $P\{0 < T < t/R_k\}$ слід знаходити з тої умови, що за час t оброблено $k + 1$ запитів (включаючи даний). Для $t > 0$

$$\begin{aligned}
F_T(t) &= (1 - \rho) \sum_{k=0}^{\infty} \rho^k \cdot \int_0^t \frac{\mu^{k+1}}{k!} u^k e^{-\mu u} du = \\
&= (1 - \rho) \mu \int_0^t e^{-\mu u} \left[\sum_{k=0}^{\infty} \frac{(\mu \rho u)^k}{k!} \right] du = \int_0^t (\mu - \lambda) e^{-(\mu-\lambda)u} du = 1 - e^{-(\mu-\lambda)t}
\end{aligned}$$

— T уже є АбсНеп ВВ, яка має показниковий розподіл з параметром $\mu - \lambda$.

2.4. Система М/М/п

Моделлю такої СМО (multiserver) теж є ПНЗ $N(t)$. Тривалість обробки запиту кожним приладом — показникова з параметром μ . Вхідний потік пуассонів з інтенсивністю λ , отже, $\lambda_k \equiv \lambda$. Однак інтенсивності вибуття різні: якщо в системі не менше n запитів, то n з них обробляються, тож сумарна інтенсивність вибуття $n\mu$, а якщо менше, k , — вони всі обробляються, і сумарна інтенсивність вибуття складає $k\mu$. Значить, $\mu_k = \mu \cdot \min\{k; n\}$.



$$p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} = \begin{cases} p_0 \prod_{i=1}^k \frac{\lambda}{\mu^i} = p_0 \frac{\lambda^k}{k! \cdot \mu^k}, & k < n, \\ p_0 \frac{\lambda^n}{n! \cdot \mu^n} \prod_{i=n+1}^k \frac{\lambda}{n\mu} = p_0 \frac{\lambda^k}{n! \cdot n^{k-n} \cdot \mu^k}, & k \geq n. \end{cases}$$

Нехай $r = \frac{\lambda}{\mu}$ і $\rho = \frac{\lambda}{n\mu} = \frac{r}{n}$. Розподіл $p_k = \begin{cases} p_0 \frac{r^k}{k!}, & k < n, \\ p_0 \frac{r^k}{n! \cdot n^{k-n}} = p_0 \frac{n^n}{n!} \rho^k, & k \geq n \end{cases}$

при $k \leq n$ схожий на пуассонів, а при $k \geq n$ — на геометричний.

Оскільки $\sum_{k=0}^{\infty} p_k = 1$, $p_0 = \left(\sum_{k=0}^{n-1} \frac{r^k}{k!} + \frac{n^n}{n!} \sum_{k=n}^{\infty} \rho^k \right)^{-1}$. Ряд збігається, коли $\rho < 1$

$$\Leftrightarrow \lambda < n\mu, \text{ при цьому } \sum_{k=n}^{\infty} \rho^k = \frac{\rho^n}{1-\rho} = \frac{r^n}{n^n(1-\rho)}, \text{ і } p_0 = \left(\sum_{k=0}^{n-1} \frac{r^k}{k!} + \frac{r^n}{(1-\rho)n!} \right)^{-1}.$$

Відмітимо, що звідси $\sum_{k=0}^{n-1} \frac{r^k}{k!} = \frac{1}{p_0} - \frac{r^n}{(1-\rho)n!}$.

При $n = 1$ ці результати перетворюються на отримані у випадку $M/M/1$.

Характеристики ефективності. Почнемо з **середньої довжини черги** $\overline{N}_q$; при її обчисленні можемо не враховувати стани $k \leq n$ (адже за такої кількості запитів в системі усі вони обробляються). Коли $N(t) > n$, у черзі перебуває $N_q(t) = N(t) - n$ запитів. Значить,

$$\begin{aligned} \overline{N}_q &= \sum_{k=n+1}^{\infty} (k-n)p_k = p_0 \frac{n^n}{n!} \sum_{k=n+1}^{\infty} (k-n) \rho^{k-n} \\ &= p_0 \frac{n^n}{n!} \sum_{s=1}^{\infty} s \rho^{s+n} = p_0 \rho \frac{(\rho n)^n}{n!} \sum_{s=1}^{\infty} s \rho^{s-1} = p_0 \rho \frac{r^n}{n!} \sum_{s=1}^{\infty} \frac{d}{d\rho} [\rho^s] = \\ &= p_0 \rho \frac{r^n}{n!} \frac{d}{d\rho} \left[\sum_{s=1}^{\infty} \rho^s \right] = p_0 \rho \frac{r^n}{n!} \frac{d}{d\rho} \left[\frac{\rho}{1-\rho} \right] = p_0 \frac{\rho r^n}{(1-\rho)^2 \cdot n!} \end{aligned}$$

За формулою Літтла **середній час перебування запиту в черзі** $\overline{T}_q = \frac{1}{\lambda} \overline{N}_q = p_0 \frac{r^n}{(n\mu)(1-\rho)^2 \cdot n!}$.

Згадуємо, що $\overline{T} = \overline{T}_q + \overline{T}_s = \overline{T}_q + \frac{1}{\mu}$, звідки **середній час перебування запиту в системі** $\overline{T} = \frac{1}{\mu} + p_0 \frac{r^n}{(n\mu)(1-\rho)^2 \cdot n!}$.

Знов-таки за формулою Літтла **середня кількість запитів у системі** $\overline{N} = \lambda \overline{T} = r + p_0 \frac{\rho r^n}{(1-\rho)^2 \cdot n!}$.

Тепер знайдемо $P\{T_q = 0\} = F_{T_q}(\{0\})$ — ймовірність того, що запит не чекає в черзі, а відразу після надходження до системи обробляється якимсь приладом. В момент надходження такого запиту у системі було менше n запитів (хоча б 1 прилад лишався вільним), значить,

$$P\{T_q = 0\} = \sum_{k=0}^{n-1} q_k \stackrel{\text{ПП}}{=} \sum_{k=0}^{n-1} p_k = p_0 \sum_{k=0}^{n-1} \frac{r^k}{k!} = p_0 \left[\frac{1}{p_0} - \frac{r^n}{(1-\rho)n!} \right] = 1 - p_0 \frac{r^n}{(1-\rho)n!} \quad (*)$$

Позначивши $C(n; r) := 1 - P\{T_q = 0\} = P\{T_q > 0\}$ (ймовірність того, що запит чекатиме в черзі):

$$C(n; r) = p_0 \frac{r^n}{(1-\rho)n!} = \frac{r^n}{(1-\rho)n!} \cdot \left[\sum_{k=0}^{n-1} \frac{r^k}{k!} + \frac{r^n}{(1-\rho)n!} \right]^{-1}$$

— *C-формула Ерланга*, виражає ймовірність чекання в черзі через r , ρ та n для системи М/М/п, — якщо нехтувати відмовами, повторними запитами, переходом у нестационарний режим, обмеженою довжиною черги.

Визначимо **розподіл тривалості чекання у черзі T_q** . При $t > 0$

$$F_{T_q}(t) \stackrel{\text{ФПЙ}}{=} P\{T_q = 0\} + \sum_{k=n}^{\infty} P\{0 < T_q < t/R_k\} P(R_k)$$

де, як і раніше, подія R_k полягає у тому, що в момент прибуття запиту до системи там вже було рівно k запитів (з них n обслуговувались приладами). $P\{0 < T_q < t/R_k\}$ є ймовірністю того, що за цієї умови протягом часу t буде оброблено $k - n + 1$ запитів: усі $k - n$ з черги та 1 з оброблюваних, аби звільнився 1 прилад. Оскільки в системі n приладів, проміжки між послідовними вибуттями запитів є Н/З ВВ, показниково розподілені з параметром $n\mu$, як $\min\{\tau_1; \dots; \tau_n\}$, де τ_i розподілені показниково з параметром μ .

Загальна тривалість обробки цих $k - n + 1$ запитів має тоді розподіл Ерланга з параметрами $(k - n + 1; n\mu)$. Також $P(R_k) = q_k = p_k$, і

$$\begin{aligned} F_{T_q}(t) &= P\{T_q = 0\} + p_0 \sum_{k=n}^{\infty} \frac{n^n}{n!} \rho^k \int_0^t \frac{(n\mu)^{k-n+1}}{(k-n)!} u^{k-n} e^{-n\mu u} du \stackrel{\rho^k = \rho^n \rho^{k-n}}{=} \\ &= P\{T_q = 0\} + p_0 n \mu \frac{r^n}{n!} \sum_{k=n}^{\infty} \int_0^t \frac{1}{(k-n)!} (n\mu \rho u)^{k-n} e^{-n\mu u} du = \\ &= P\{T_q = 0\} + p_0 n \mu \frac{r^n}{n!} \int_0^t e^{-n\mu u} \left[\sum_{s=0}^{\infty} \frac{(\lambda u)^s}{s!} \right] du = \\ &= P\{T_q = 0\} + p_0 n \mu \frac{r^n}{n!} \int_0^t e^{-(n\mu - \lambda)u} du = \\ &= P\{T_q = 0\} + p_0 n \mu \frac{r^n}{(n\mu - \lambda)n!} [1 - e^{-(n\mu - \lambda)t}] \stackrel{(*)}{=} \\ &= 1 - p_0 \frac{r^n}{(1-\rho)n!} + p_0 \frac{r^n}{(1-\rho)n!} [1 - e^{-(n\mu - \lambda)t}] = 1 - C(n; r) e^{-(n\mu - \lambda)t} \quad (**) \end{aligned}$$

(*) і (**) означають, що тривалість чекання в черзі T_q тут теж є сумою дискретної ВВ, зосередженої в 0, та АбсНеп показниково розподіленої ВВ.

При $t > 0$: $P\{T_q \geq t\} = 1 - F_{T_q}(t) = p_0 \frac{r^n}{(1-\rho)n!} \cdot e^{-(n\mu-\lambda)t}$, тому

$$P\{T_q \geq t/T_q > 0\} = \frac{P\{T_q \geq t\}}{P\{T_q > 0\}} = e^{-(n\mu-\lambda)t}$$

— умовний розподіл T_q при $T_q > 0$ показниковий з параметром $n\mu - \lambda$ (#)

Розподіл часу перебування в системі T знайдемо, розбивши усі запити на ті, які не чекали в черзі ($T_q = 0$) та ті, які чекали ($T_q > 0$). Запит належить до першого типу з ймовірністю $P\{T_q = 0\}$, при цьому тривалість його перебування в системі співпадає з тривалістю його обробки — показниковою ВВ з параметром μ , позначимо її τ ; до другого типу — з ймовірністю $P\{T_q > 0\}$, тоді час у системі є сумою тривалості чекання та тривалості обробки. Отже, ФР $F_T(t) = P\{T < t\} \stackrel{\text{ФПЙ}}{=}$

$$\begin{aligned} &= P\{T < t/T_q = 0\}P\{T_q = 0\} + P\{T < t/T_q > 0\}P\{T_q > 0\} = \\ &= P\{\tau < t/T_q = 0\}P\{T_q = 0\} + P\{T_q + \tau < t/T_q > 0\}P\{T_q > 0\} = \\ &= (1 - e^{-\mu t}) \left[1 - p_0 \frac{r^n}{(1-\rho)n!}\right] + P\{T_q + \tau < t/T_q > 0\} p_0 \frac{r^n}{(1-\rho)n!} \end{aligned}$$

Розподіл $T_q + \tau$ за умови $T_q > 0$ є згорткою двох показникових розподілів: одного (#) з параметром $n\mu - \lambda$, другого з параметром μ .

РЕС. Якщо ξ й η Н/З, показниково розподілені з параметрами a й b , то

щільність їх суми для $x > 0$ $p_{\xi+\eta}(x) = \int_{-\infty}^{+\infty} p_{\xi}(x-u)p_{\eta}(u)du =$

$$= \int_0^x a e^{-a(x-u)} b e^{-bu} du = a b e^{-ax} \frac{1}{a-b} e^{(a-b)u} \Big|_0^x = \frac{ab}{a-b} (e^{-bx} - e^{-ax}) \Rightarrow$$

$$\Rightarrow F_{\xi+\eta}(x) = \int_0^x p_{\xi+\eta}(u) du = \frac{1}{a-b} (b e^{-au} - a e^{-bu}) \Big|_0^x = 1 - \frac{1}{a-b} [a e^{-bx} - b e^{-ax}]$$

$$\begin{aligned} \text{Тож } P\{T_q + \tau < t/T_q > 0\} &= 1 - \frac{1}{n\mu - \lambda - \mu} [(n\mu - \lambda) e^{-\mu t} - \mu e^{-(n\mu - \lambda)t}], \text{ а } F_T(t) = \\ &= (1 - e^{-\mu t}) \left[1 - p_0 \frac{r^n}{(1-\rho)n!}\right] + \left\{1 - \frac{1}{n\mu - \lambda - \mu} [(n\mu - \lambda) e^{-\mu t} - \mu e^{-(n\mu - \lambda)t}]\right\} p_0 \frac{r^n}{(1-\rho)n!} = \\ &= 1 - e^{-\mu t} + p_0 \frac{r^n}{(1-\rho)n!} \left\{e^{-\mu t} - \frac{1}{n\mu - \lambda - \mu} [(n\mu - \lambda) e^{-\mu t} - \mu e^{-(n\mu - \lambda)t}]\right\} = \\ &= 1 - e^{-\mu t} + p_0 \frac{r^n}{(1-\rho)n!} \cdot \frac{1}{n\mu - \lambda - \mu} [\mu e^{-(n\mu - \lambda)t} - \mu e^{-\mu t}] = \\ &= 1 - e^{-\mu t} - p_0 \frac{r^n}{(1-\rho)(n-1-r)n!} [e^{-\mu t} - e^{-(n\mu - \lambda)t}] \end{aligned}$$

ЗАУВ. («Перевірка»). Узявши тут $n = 1$ ($\rho = r = \frac{\lambda}{\mu}$, $p_0 = 1 - \rho$):

$$F_T(t) = 1 - e^{-\mu t} - \frac{r}{(-r)} [e^{-\mu t} - e^{-(\mu - \lambda)t}] = 1 - e^{-(\mu - \lambda)t}$$

— результат, який отримали раніше для системи М/М/1.

Визначення кількості приладів n — виконується по-різному залежно від мети, яку ми хочемо досягти.

Подальші твердження стосуються СМО М/М/п з «великими» n .

У стаціонарному режимі внаслідок $\rho = \frac{r}{n} < 1$ буде $n > r \Leftrightarrow n = r + \Delta$, де $\Delta > 0$ (може бути нецілим) — кількість приладів «на додачу» до навантаження r . Тому вибір n можна тлумачити як вибір «додачі» Δ .

Розглянемо для прикладу М/М/п з $n = 12$, $r = \frac{\lambda}{\mu} = 9$ та $\rho = \frac{3}{4}$. Вважаємо, що її функціонування в стаціонарному режимі (який існує, $\rho < 1$) задовільне. Якщо збільшити навантаження в 4 рази, до $r = 36$, якою має стати кількість приладів, аби система також працювала задовільно? Відповідь залежить від того, яку характеристику прагнемо зберегти:

- *Темп*, коефіцієнт використання ρ . Оскільки $\rho = \frac{r}{n}$, матимемо $n = 48$.

Темп зберігається за рахунок збільшення сумарної ціни приладів. У той же час ймовірність чекання в черзі, $P\{T_q > 0\} \xrightarrow[r \rightarrow \infty]{} 0$, коли $\rho = \text{const}$. Іншими словами, при такому підході у СМО з великою кількістю приладів затримка в черзі буде короткою.

Отримати шукану кількість приладів просто: $n = \frac{r}{\rho}$.

Цей підхід називається *quality domain* (QD).

- *Дієвість*, кількість Δ приладів «на додачу». У початковій конфігурації $\Delta = n - r = 12 - 9 = 3$; для $r = 36$ буде $n = 36 + \Delta = 39$ приладів.

Тут, навпаки, зберігається ціна додаткових приладів за рахунок темпу обслуговування. При зростанні навантаження до ∞ , коли $\Delta = \text{const}$, матимемо $\rho \rightarrow 1$ та $P\{T_q > 0\} \rightarrow 1$. У великих СМО, побудованих з використанням цього підходу, майже напевне будуть затримки в черзі.

Кількість приладів теж знаходиться легко: $n = r + \Delta$.

Цей підхід називається *efficiency domain* (ED).

- *Темп та дієвість*, міру затримки в черзі; приміром, ймовірність чекання в черзі $P\{T_q > 0\} = C(n; r)$. У початковій системі, за С-формулою Ерланга, вона ≈ 0.266 . При $r = 36$, аби вона була не більшою, необхідно $n = 42$ прилади. У загальному випадку, якщо прагнемо обмежити ймовірність чекання пороговим $\alpha \in (0; 1)$, шукаємо $n_0 = \min\{n \in \mathbb{N}: P\{T_q > 0\} \leq \alpha\}$; тобто мова йде про обернення С-формули Ерланга.

Тут вже немає «простої» формули, яка б виражала n через r , ρ та Δ .

Цей підхід називається *quality and efficiency domain* (QED).

Конфігурація	r	n	ρ	Δ	$P\{T_q > 0\}$
Початкова	9	12	0.75	3	0.266
QD	36	48	0.75	12	0.037
ED	36	39	0.92	3	0.523
QED	36	42	0.86	6	0.246

QED є компромісом між QD та ED, — керуючись ним, ми підтримуємо рівновагу між ціною приладів та затримкою запитів у черзі. Залежність n від інших параметрів складніша, проте достатньо корисним є наближення за *правилом квадратного кореня*

$$n \approx r + \beta\sqrt{r} \Leftrightarrow \Delta \approx \beta\sqrt{r} \quad (\#\#)$$

(де β — описана нижче стала), що впливає з наступної

ТЕОР. (Halfin & Whitt, 1981). Нехай $\{M/M/n\}_{n \in \mathbb{N}}$ — послідовність СМО з навантаженнями $\frac{\lambda_n}{\mu_n} = r_n$.

$$\lim_{n \rightarrow \infty} C(n; r_n) = \alpha \in (0; 1) \quad \text{тоді} \quad \lim_{n \rightarrow \infty} \frac{n - r_n}{\sqrt{n}} = \beta > 0$$

де $\alpha = \frac{\varphi(\beta)}{\varphi(\beta) + \beta\Phi(\beta)}$ (тут $\varphi(\cdot)$ і $\Phi(\cdot)$ — щільність та ФР $N_{0,1}$).

($\#\#$) означає, що ймовірність затримки у черзі лишається приблизно сталою, коли додаткова кількість Δ приладів зростає пропорційно квадратному кореню з навантаження. Це правило дозволяє обирати параметри системи, навіть коли деякі характеристики невідомі; наприклад, якщо система працює задовільно, достатньо знати лише кількість Δ приладів на додачу до навантаження r , щоб при збільшенні навантаження в k разів узяти додаткову кількість, яка забезпечить таку ж задовільну затримку: $\Delta' \approx \sqrt{k} \cdot \Delta$.

Крім того, за допомогою ($\#\#$) визначається початкова кількість приладів: коли плановане навантаження r відоме, обираємо α , знаходимо відповідне β (Kolesar & Green показали (1998), що у якості β можна взяти $(1 - \alpha)$ -квантиль $N_{0,1}$ -розподілу), і покладаємо $n = r + \beta\sqrt{r}$.

2.5. Система $M/M/n/m$

відрізняється від попередньої обмеженою ємністю: у ній може перебувати загалом не більше m запитів. Очевидно, якщо $m < n$, прилади з номерами $> m$ ніколи не будуть задіяними (принаймні коли запит надходить до вільного приладу з найменшим номером), тобто така система працюватиме як $M/M/m/m$. Тому вважаємо $m \geq n$.

Відмінність моделюючого ПНЗ $N(t)$ полягає в тому, що $\lambda_k \equiv 0$ для $k \geq m$, і цей ВП має скінченну множину станів: від 0 до m . Тому $p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} =$

$\begin{cases} p_0 \frac{r^k}{k!}, & k < n, \\ p_0 \frac{n^n}{n!} \rho^k, & n \leq k \leq m \end{cases}$ — «пуассоноподібний» для $k < n$ та «геометричнопо-
дібний» у супротивному випадку. p_0 знаходимо з тої-таки рівності

$$\sum_{k=0}^m p_k = 1 \Leftrightarrow p_0 = \left[\sum_{k=0}^{n-1} \frac{r^k}{k!} + \frac{n^n}{n!} \sum_{k=n}^m \rho^k \right]^{-1}$$

Сума $\sum_{k=n}^m \rho^k$ скінченна, тому обмеження $\rho < 1$ можна послабити до $\rho \leq 1$.

Виникає 2 випадки: $p_0 = \left[\sum_{k=0}^{n-1} \frac{r^k}{k!} + \frac{r^n}{n!} \times \begin{cases} \rho^n \frac{1-\rho^{m-n+1}}{1-\rho}, & \rho < 1, \\ m-n+1, & \rho = 1 \end{cases} \right]^{-1} =$

$$= \left[\sum_{k=0}^{n-1} \frac{r^k}{k!} + \frac{r^n}{n!} \times \begin{cases} \frac{1-\rho^{m-n+1}}{1-\rho}, & \rho < 1, \\ m-n+1, & \rho = 1 \end{cases} \right]^{-1}$$

(при $\rho < 1$ буде $\rho^n n^n = r^n$, а при $\rho = \frac{r}{n} = 1$ буде $n^n = r^n$).

Характеристики ефективності. Почнемо з середньої довжини черги $\bar{N}_q = \sum_{k=n+1}^m (k-n)p_k = \sum_{k=n+1}^m (k-n)p_0 \frac{r^n}{n!} \rho^k \stackrel{s \leftarrow k-n}{=} p_0 \rho \frac{r^n}{n!} \sum_{s=1}^{m-n} s \rho^{s-1}$:

a) $\rho < 1$. $\bar{N}_q = p_0 \rho \frac{r^n}{n!} \sum_{s=0}^{m-n} \frac{d}{d\rho} [\rho^s] = p_0 \rho \frac{r^n}{n!} \frac{d}{d\rho} \left[\frac{1-\rho^{m-n+1}}{1-\rho} \right] =$

$$= p_0 \frac{\rho}{(1-\rho)^2} \cdot \frac{r^n}{n!} \left\{ -(m-n+1)\rho^{m-n}(1-\rho) - (-1)(1-\rho^{m-n+1}) \right\} =$$

$$= p_0 \frac{\rho}{(1-\rho)^2} \cdot \frac{r^n}{n!} \left\{ 1 - \rho^{m-n+1} - (m-n+1)(1-\rho)\rho^{m-n} \right\}$$

b) $\rho = 1$. $\bar{N}_q = p_0 \frac{r^n}{n!} \sum_{s=1}^{m-n} s = p_0 \frac{r^n}{n!} \cdot \frac{(m-n)(m-n+1)}{2}$.

Середня кількість запитів у системі $\bar{N}$. У моделі М/М/п/∞ рівність $\bar{N} = \bar{N}_q + r$ впливала з $\bar{T} = \bar{T}_q + \frac{1}{\mu}$ й формули Літтла; для обмежених черг формула Літтла змінюється, як і залежність $\bar{N}$ від $\bar{N}_q$, адже частка p_m запитів, що надходять до системи, відразу лишаться її, бо в стані $N(t) = m$ ємність вичерпана. Поєднаємо це з тим, що прибуваючі у складі ПП запити в середньому «спостерігають» середнє за часом (рівномірність ПП). Отже, темп надходження запитів, які не відразу вибувають, є $\lambda_{eff} = \lambda(1-p_m)$. Відповідно $r_{eff} = \frac{\lambda_{eff}}{\mu}$ і $\bar{N} = \bar{N}_q + \frac{\lambda_{eff}}{\mu} = \bar{N}_q + r(1-p_m)$.

Середня кількість оброблюваних запитів повинна бути менше за кількість приладів, $\bar{N}_s = \bar{N} - \bar{N}_q = r(1-p_m) < n$, звідки $\rho_{eff} = \frac{\lambda_{eff}}{n\mu} = \frac{\lambda(1-p_m)}{n\mu} < 1$.

За виправленою ФЛ для обмежених черг з іншим темпом появи «не відразу вибувають» запитів, **середній час чекання в черзі** $\bar{T}_q = \frac{\bar{N}_q}{\lambda_{eff}} = \frac{\bar{N}_q}{\lambda(1-p_m)}$, а **середній час перебування в системі** $\bar{T} = \frac{\bar{N}}{\lambda_{eff}} = \frac{\bar{N}}{\lambda(1-p_m)}$.

При $n = 1$, для СМО М/М/1/м, описані характеристики виражаються простіше: $\rho = r$ і

$$p_0 = \begin{cases} \left(1 + \rho \frac{1-\rho^m}{1-\rho}\right)^{-1} = \frac{1-\rho}{1-\rho^{m+1}}, & \rho < 1, \\ \frac{1}{m+1}, & \rho = 1, \end{cases} \text{ та } p_k = \begin{cases} \frac{1-\rho}{1-\rho^{m+1}} \rho^k, & \rho < 1, \\ \frac{1}{m+1}, & \rho = 1, \end{cases}$$

$$\begin{aligned}\overline{N}_q &= \begin{cases} p_0 \frac{\rho^2}{(1-\rho)^2} [1 - \rho^m - m(1-\rho)\rho^{m-1}], & \rho < 1, \\ p_0 \frac{m(m-1)}{2}, & \rho = 1 \end{cases} = \\ &= \begin{cases} \frac{\rho^2}{(1-\rho)(1-\rho^{m+1})} [1 - \rho^{m-1}(\rho + m - m\rho)], & \rho < 1, \\ \frac{m(m-1)}{2(m+1)}, & \rho = 1. \end{cases}\end{aligned}$$

$\overline{N} = \overline{N}_q + \rho$, звідки $\rho = 1 - p_0 = r_{eff} = \frac{\lambda(1-p_m)}{\mu} \Leftrightarrow \mu(1-p_0) = \lambda(1-p_m)$ — «вихідний» потік має співпадати з «вхідним» (виключаючи запити, які відразу вибувають, коли ємність вичерпана).

Розподіл тривалості чекання у черзі. На відміну від попередніх систем, $q_k \neq p_k$ (потік усіх запитів пуассонів, а тих, які не втрачаються — не пуассонів), адже нас цікавить стан черги в моменти прибуття запитів, які лишаються. Щоб знайти q_k для $k \leq m-1$, застосуємо формули Байєса, позначивши через A подію прибуття (arrival) запиту до системи:

$$\begin{aligned}q_k &= P(R_k) = P(N = k/A) = \frac{P(A/N=k)P\{N=k\}}{\sum_{s=0}^m P(A/N=s)P\{N=s\}} = \frac{P(A/N=k)p_k}{\sum_{s=0}^m P(A/N=s)p_s} = \\ &= \lim_{\Delta t \rightarrow 0} \frac{[\lambda \Delta t + o(\Delta t)] p_k}{\sum_{s=0}^{m-1} [\lambda \Delta t + o(\Delta t)] p_s} = \lim_{\Delta t \rightarrow 0} \frac{[\lambda + o(1)] p_k}{\sum_{s=0}^{m-1} [\lambda + o(1)] p_s} = \frac{\lambda p_k}{\lambda \sum_{s=0}^{m-1} p_s} = \frac{p_k}{1-p_m}\end{aligned}$$

Зазначимо, що $p_m \xrightarrow{m \rightarrow \infty} 0$, і для $M/M/n/\infty$ матимемо $q_k = p_k$.

Знов $F_{T_q}(t) = P\{T_q < t\} = P\{T_q = 0\} + \sum_{k=n}^{m-1} P\{0 < T_q < t/R_k\}P(R_k)$, а $P\{0 < T_q < t/R_k\}$ є ймовірністю того, що протягом часу t буде оброблено $k-n+1$ запитів, якщо спочатку їх було k . $q_m = 0$, бо такі запити не встають у чергу. T_q має тоді розподіл Ерланга з параметрами $(k-n+1; n\mu)$, і

$$\begin{aligned}F_{T_q}(t) &= P\{T_q = 0\} + \sum_{k=n}^{m-1} q_k \int_0^t \frac{n\mu}{(k-n)!} (n\mu u)^{k-n} e^{-n\mu u} du = \\ &= P\{T_q = 0\} + \sum_{k=n}^{m-1} q_k \left[1 - \int_t^\infty \frac{n\mu}{(k-n)!} (n\mu u)^{k-n} e^{-n\mu u} du \right]\end{aligned}$$

$\int_t^\infty \frac{\alpha}{s!} (\alpha u)^s e^{-\alpha u} du$ — ймовірність того, що ВВ, яка має розподіл Ерланга з параметрами $(s+1; \alpha)$, прийняла значення $\geq t$. Це рівносильно тому, що у ПП, інтенсивності α , $(s+1)$ -а подія відбулась не раніше t (момент $(s+1)$ -ї події має вказаний розподіл Ерланга як згортка показникових розподілів). Значить, протягом часу $[0; t]$ в цьому ПП відбулось не більше s подій. Маємо

$$\int_t^\infty \frac{\alpha}{s!} (\alpha u)^s e^{-\alpha u} du = \sum_{i=0}^s \frac{(\alpha t)^i}{i!} e^{-\alpha t}$$

Підставляючи тут $\alpha = n\mu$ та $s = k-n$, приходимо до

$$F_{T_q}(t) = P\{T_q = 0\} + \sum_{k=n}^{m-1} q_k - \sum_{k=n}^{m-1} q_k \sum_{i=0}^{k-n} \frac{(n\mu t)^i}{i!} e^{-n\mu t}$$

Оскільки $P\{T_q = 0\} = \sum_{k=0}^{n-1} q_k$ та $q_m = 0$, сума перших 2-х доданків є $\sum_{k=0}^{m-1} q_k = 1$, тому $F_{T_q}(t) = 1 - \frac{1}{1-p_m} \sum_{k=n}^{m-1} p_k \sum_{i=0}^{k-n} \frac{(n\mu t)^i}{i!} e^{-n\mu t}$.

Формула втрат Ерланга. Розглянемо випадок $m = n$ — СМО М/М/п/п, що не має черги: коли всі прилади зайняті, кожний прибуваючий запит відразу залишає систему (т.з. «нетерплячі» запити). Тоді

$$p_0 = \left[\sum_{k=0}^n \frac{r^k}{k!} \right]^{-1} \text{ та } p_k = p_0 \frac{r^k}{k!} \text{ для } k = \overline{0, n} \quad (*)$$

При $k = n$ величина $B(n; r) = p_n = \frac{r^n}{n!} / \left[\sum_{k=0}^n \frac{r^k}{k!} \right]$ — *В-формула Ерланга*, ймовірність того, що система повна, у стаціонарному режимі. Зважаючи на пуассоновість вхідного потоку та рівномірність ПП, це також частка запитів, які прибувають до повної системи і втрачаються (звідси назва).

ЗАУВ. (*) справедлива для довільної СМО М/Г/п/п, незалежно від розподілу тривалості обробки запиту, — стаціонарні ймовірності залежатимуть тільки від середньої тривалості. Доведення цього Ерлангом (1917) для сталої тривалості було виправлене та узагальнене подальшими роботами Vaultot (1927), Pollaczek (1932), Palm (1938), Kosten (1948) і Takács (1969).

$$\begin{aligned} \textbf{ТВ. 1. } \forall n \in \mathbb{N}: B(n; r) &= \frac{r^n/n!}{r^n/n! + \sum_{k=0}^{n-1} (r^k/k!)} = \frac{r^n/(n-1)!}{r^n/(n-1)! + n \sum_{k=0}^{n-1} (r^k/k!)} = \\ &= \frac{r \left[r^{n-1}/(n-1)! / \sum_{k=0}^{n-1} (r^k/k!) \right]}{n + r \left[r^{n-1}/(n-1)! / \sum_{k=0}^{n-1} (r^k/k!) \right]} = \frac{rB(n-1; r)}{n + rB(n-1; r)} \end{aligned}$$

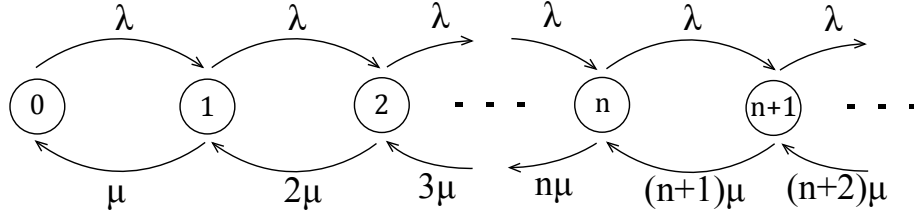
Вважаючи $B(0; r) = 1$, за допомогою цього рекурентного співвідношення можна знайти $B(n; r)$, уникаючи складнощів обчислення $\frac{r^k}{k!}$ для великих k .

$$\begin{aligned} \textbf{ТВ. 2. } \forall n \in \mathbb{N}: C(n; r) &= \frac{r^n / [(1-r/n)n!]}{\sum_{k=0}^{n-1} (r^k/k!) + r^n / [(1-r/n)n!]} = \frac{n \cdot r^n / n!}{(n-r) \sum_{k=0}^{n-1} (r^k/k!) + n \cdot r^n / n!} = \\ &= \frac{n \cdot r^n / n!}{n \sum_{k=0}^n (r^k/k!) - r \sum_{k=0}^{n-1} (r^k/k!)} = \frac{n \cdot r^n / n!}{(n-r) \sum_{k=0}^n (r^k/k!) + r \cdot r^n / n!} = \\ &= \frac{n \cdot [r^n/n! / \sum_{k=0}^n (r^k/k!)]}{n - r + r \cdot [r^n/n! / \sum_{k=0}^n (r^k/k!)]} = \frac{nB(n; r)}{n - r + rB(n; r)} \end{aligned}$$

Зауважимо, що $C(n; r)$ відноситься до СМО М/М/п/∞.

2.6. Система М/М/∞

Типовим прикладом такої СМО (ample service) є система з «самообслуговуванням» запитів, коли їх може бути багато, приміром, великий ліс, крізь який проходять люди. Зайшов у ліс — початок «обробки», вийшов — кінець. Або океан та кораблі, що його перепливають.

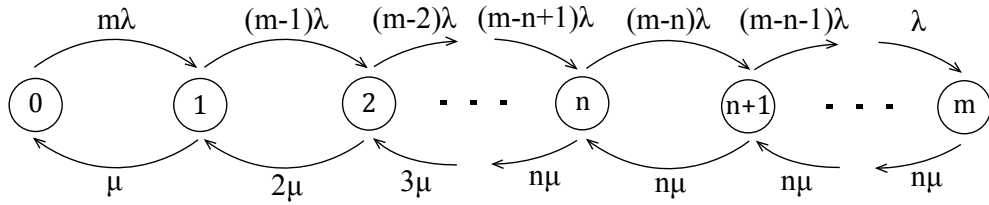


Моделюємо ПНЗ $N(t)$ з інтенсивностями $\lambda_k \equiv \lambda$, $\mu_k = k\mu$. Загальна формула $p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i}$ дає $p_k = p_0 \frac{r^k}{k!}$, де $p_0 = \left[\sum_{k=0}^{\infty} \frac{r^k}{k!} \right]^{-1} = e^{-r}$. Отже $p_k = \frac{r^k}{k!} e^{-r}$ — стаціонарний розподіл пуассонів з параметром $r = \frac{\lambda}{\mu}$. Навантаження r тут може бути довільним, як завгодно великим (коефіцієнт використання $\rho = \frac{r}{n} = 0$). Виявляється, такий розподіл в стаціонарному режимі притаманний будь-якій СМО М/Г/∞ і залежить тільки від середньої тривалості обробки запиту, а не від її розподілу.

Характеристики ефективності. Черги немає, $N_q \equiv 0$ і $T_q \equiv 0$. **Середня кількість запитів у системі** $\bar{N} \in \text{МС}$ пуассонової ВВ з параметром r , $\bar{N} = r$. Тривалість T перебування в системі співпадає з тривалістю обробки — показниково розподіленою з параметром μ ВВ, значить, **середня тривалість перебування в системі** $\bar{T} = \bar{T}_s = \frac{1}{\mu}$.

2.7. Скінченноджерельні системи

У попередніх СМО ми вважали джерело, з якого запити надходять до системи, нескінченним. Нехай тепер загальна кількість запитів у наборі «джерело + система» рівна m , — вони «циркулюють» між джерелом і системою, — та інтенсивність надходження їх з джерела залежить від стану системи, $N(t)$. Припустимо, що СМО має n приладів, кожний обслуговує запит протягом показниково розподіленого з параметром μ часу, а запит перебуває поза системою протягом часу, показниково розподіленого з параметром λ . Зокрема, це означає, що коли в системі перебуває k запитів, то у джерелі їх $(m-k)$, тож інтенсивність їх надходження становить $(m-k)\lambda$ (як $\min$ від $m-k$ Н/З показниково розподілених з параметром λ тривалостей).



Моделлю такої СМО (finite-source) є ПНЗ $N(t)$ з $\lambda_k = \max\{0; m-k\} \cdot \lambda$ та $\mu_k = \min\{k; n\} \cdot \mu$. Цей ВП має скінченну множину станів, $\{0; \dots; m\}$, і згідно загальної формули стаціонарного розподілу $p_k =$

$$= p_0 \begin{cases} \prod_{i=1}^k \frac{(m-i+1)\lambda}{i\mu} = \frac{m!}{(m-k)!} \cdot \frac{1}{k!} r^k = C_m^k r^k, & k < n, \\ \frac{m!}{n!(m-n)!} r^n \cdot \prod_{i=n+1}^k \frac{(m-i+1)\lambda}{n\mu} = \frac{m!}{n!(m-k)!} \cdot \frac{r^k}{n^{k-n}} = C_m^k \frac{k!}{n^{k-n} n!} r^k, & n \leq k \leq m \end{cases} \quad (**)$$

Покладемо $a_k = \begin{cases} C_m^k r^k, & k < n, \\ C_m^k \frac{k!}{n^{k-n} n!} r^k, & n \leq k \leq m, \end{cases}$ тоді $p_0 = \left[\sum_{k=0}^m a_k \right]^{-1}$; у такої

системи немає «замкненого», «простого» виразу для p_0 . З іншого боку, кількість $\{a_k\}$ скінченна, тому сума обчислюється за обмежений час.

У випадку $n = 1$: $p_k = C_m^k r^k k! p_0 = A_m^k r^k p_0$, де $p_0 = \left[\sum_{k=0}^m A_m^k r^k \right]^{-1}$.

Характеристики ефективності. За означенням МС ДВВ, середня кількість запитів у системі $\bar{N} = \sum_{k=1}^m k p_k = p_0 \sum_{k=1}^m k a_k$.

Аби застосувати формулу Літтла, виправимо її відповідно до «ефективної» інтенсивності надходження запитів у систему. У стані $N(t) = k$, згідно сказаному вище, інтенсивність прибуття є $(m-k)\lambda$. Загальна інтенсивність прибуття є сумою цих інтенсивностей, узятих з вагами — частками часу, коли система перебуває у такому-то стані. Тобто

$$\lambda_{eff} = \sum_{k=0}^m (m-k) \lambda p_k = \lambda \left[m \sum_{k=0}^m p_k - \sum_{k=1}^m k p_k \right] = \lambda(m - \bar{N})$$

Тоді $\bar{T} = \frac{1}{\lambda_{eff}} \bar{N} = \frac{\bar{N}}{\lambda(m-\bar{N})}$, $\bar{T}_q = \bar{T} - \bar{T}_s = \frac{1}{\lambda_{eff}} \bar{N} - \frac{1}{\mu}$, і $\bar{N}_q = \bar{N} - r_{eff} = \bar{N} - \frac{\lambda_{eff}}{\mu} = \bar{N} - r(m - \bar{N})$. $\bar{N}_q$ отримується як $\sum_{k=n+1}^m (k-n) p_k = p_0 \sum_{k=n+1}^m (k-n) a_k$.

ЗАУВ. (**) має місце для довільної скінченноджерельної СМО з показниковим розподілом тривалості обробки і середньою тривалістю перебування запиту в джерелі $\frac{1}{\lambda}$, яким би не був розподіл цієї тривалості.

Якщо кількість запитів рівна кількості приладів, $m = n$, то виконується твердження (*), отримане раніше щодо $M/G/n/n$, за умови показникового розподілу тривалості перебування запиту в джерелі.

2.8. Системи з нетерпінням

які ще називають «системами з нетерплячими запитами», тобто запити можуть залишати систему, не дочекавшись обробки. Загалом, це відбувається, коли час затримки запиту в черзі надто великий. Якщо у СМО значна кількість запитів втрачається саме через нетерпіння, і впливати на міру цього нетерпіння неможливо, треба змінювати параметри системи.

Можна виділити 3 види нетерпіння:

- *відмова вставати у чергу* (balking) відразу після прибуття до системи,
- *відмова продовжувати чекання у черзі* (reneging) внаслідок достатньо довгої затримки там,
- *перехід до іншої черги* (jockeying), коли черг декілька.

Розглянемо перші два для системи M/M/1 і дисципліни FCFS.

M/M/1 balking. Одну таку модель ми вже розглянули: M/M/n/m, з обмеженою ємністю. Але відносно багатьох реальних СМО вона спрощена; якщо m не однозначно визначене, поведінка запитів навряд чи буде настільки детермінованою: «вже є менше m запитів — чекати, більше — піти».

Інший підхід: змінити темп надходження запитів, зробивши його залежним від кількості запитів у системі, спадаючим при зростанні цієї кількості. Тобто ми розглядаємо «околицю» СМО, яку запити проходять і вирішують, чи йти далі у систему (і, можливо, чекати), чи не йти (ці втрачаються). Темп надходження до «околиці» вважаємо сталим, λ , але змінною є частка запитів, що вирішують йти у систему, і вона залежить від кількості запитів, які вже є всередині. Тоді темп прибуття запитів до системи — з околиці — теж залежить від цієї кількості k , він є λ_k .

Нехай $\{b_k\}_{k=0}^{\infty}$, $0 \leq b_{k+1} \leq b_k \leq 1$, $b_0 = 1$ — незростаюча послідовність, і $\lambda_k = \lambda b_k$. У ПНЗ, що моделює M/M/1 з інтенсивностями народження λ_k ,

$$p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} = p_0 \left(\frac{\lambda}{\mu}\right)^k \prod_{i=1}^k b_{i-1} = p_0 \rho^k \prod_{i=1}^k b_{i-1}, \text{ де } p_0 = \left(1 + \sum_{k=1}^{\infty} \rho^k \prod_{i=1}^k b_{i-1}\right)^{-1}$$

У якості $\{b_k\}$ природньо взяти послідовність типу $\frac{1}{1+k}$, $\frac{1}{1+k^\alpha}$, $e^{-\alpha k}$, ... — вона має відповідати поведінці запитів. Запит/клієнт при «підході» до СМО може оцінювати не довжину черги, а тривалість затримки; може встати у довгу чергу, якщо вона рухається швидко, і відмовитись вставати у коротку, якщо обслуговування повільне. За темпу обробки μ середня тривалість її $\bar{T}_s = \frac{1}{\mu}$, так що знаючи μ і спостерігаючи в системі $k-1$ клієнтів, черговий клієнт оцінює тривалість свого чекання як $\frac{k}{\mu}$ (включно з обслуговуванням).

Можна вважати $b_k = e^{-\alpha \frac{k}{\mu}}$ (де $\alpha > 0$ — «увірванюваність терпця»).

Систему M/M/1/m при такому підході задає набір $b_k = \begin{cases} 1, & 0 \leq k < m, \\ 0, & k \geq m. \end{cases}$

M/M/1 reneging + balking. Окрім відмови вставати у чергу, запит може піти з неї і з системи загалом, тобто втратитись, у будь-який момент, із тим вищою ймовірністю, чим більше оцінювана тривалість подальшого чекання. Вважатимемо, що така «нетерплячість» впливає на темп вибуття запитів з системи: $\mu_k = \mu + r(k)$, де «функція залишання» (reneging function)

$$r(k) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} Pr\{\text{запит йде з черги протягом часу } \Delta t \text{ / в системі } k \text{ запитів}\}$$

причому $r(0) = r(1) = 0$. Тоді $p_k = p_0 \prod_{i=1}^k \frac{\lambda_{i-1}}{\mu_i} = p_0 \lambda^k \prod_{i=1}^k \frac{b_{i-1}}{\mu + r(i)}$ для $k \geq 1$, і

$$p_0 = \left(1 + \sum_{k=1}^{\infty} \lambda^k \prod_{i=1}^k \frac{b_{i-1}}{\mu + r(i)}\right)^{-1}.$$

У якості $r(k)$ можна взяти аналогічну balking-випадку $e^{\beta \frac{k}{\mu}}$ з $\beta > 0$.

2.9. Перехідна поведінка

Розглянемо, як змінюється поведінка СМО у часі: визначимо $\{p_k(t)\}_k$, ймовірності того, що у момент $t > 0$ в системі перебуває k запитів. Обмежимося 3-ма випадками, бо за найменших послаблень умов пуассоновості та змін параметрів аналітичне відшукування цих ймовірностей різко ускладнюється.

Перехідна поведінка M/M/1/1. Очевидно, $p_k(t) \equiv 0$ для $k \geq 2$; лишається знайти $p_0(t)$ і $p_1(t)$. Скористаємось наслідком з прямого рівняння Колмогорова для однорідного процесу: $p'(t) = p(t)Q$, де $p(t) = \|p_k(t)\|_k$, а Q — ІГ. Раніше встановили, що для ПНЗ ця система ДР виглядає так:

$$\begin{cases} \frac{d}{dt} p_0(t) = -\lambda_0 p_0(t) + \mu_1 p_1(t), \\ \frac{d}{dt} p_k(t) = -(\lambda_k + \mu_k) p_k(t) + \lambda_{k-1} p_{k-1}(t) + \mu_{k+1} p_{k+1}(t), \quad k \geq 1 \end{cases}$$

У ПНЗ $N(t)$, моделі M/M/1/1: $\lambda_0 = \lambda$, $\lambda_k = 0$ при $k \geq 1$, $\mu_1 = \mu$,

$$\begin{cases} p'_0(t) = -\lambda p_0(t) + \mu p_1(t), \\ p'_1(t) = -\mu p_1(t) + \lambda p_0(t). \end{cases} \quad (*)$$

$\forall t \geq 0$: $p_0(t) + p_1(t) = 1 \Leftrightarrow p_0(t) = 1 - p_1(t)$. Підставимо це у 2-е рівняння і перепишемо його: $p'_1(t) + (\lambda + \mu) p_1(t) = \lambda$ (**)

— лінійне неоднорідне ДР I порядку зі сталими коефіцієнтами. Загальним розв'язком відповідного однорідного є $Ce^{-(\lambda+\mu)t}$, а одним з розв'язків неоднорідного є стала $\frac{\lambda}{\lambda+\mu}$. Значить, загальний розв'язок (**) має вигляд

$$p_1(t) = Ce^{-(\lambda+\mu)t} + \frac{\lambda}{\lambda+\mu}$$

При $t = 0$: $p_1(0) = C + \frac{\lambda}{\lambda+\mu}$, звідки $C = p_1(0) - \frac{\lambda}{\lambda+\mu}$. Таким чином,

$$\begin{cases} p_0(t) = \frac{\mu}{\lambda+\mu}(1 - e^{-(\lambda+\mu)t}) + p_0(0)e^{-(\lambda+\mu)t}, \\ p_1(t) = \frac{\lambda}{\lambda+\mu}(1 - e^{-(\lambda+\mu)t}) + p_1(0)e^{-(\lambda+\mu)t}. \end{cases}$$

За умови $N(0) = 0$ буде $p_0(0) = 1$, $p_1(0) = 0$.

Аби визначити стаціонарний розподіл, покладемо в (*) похідні рівними 0 і замінимо $p_i(t)$ на p_i : $\lambda p_0 = \mu p_1 \Leftrightarrow p_1 = \rho p_0$; додавши $p_0 + p_1 = 1$, отримаємо $p_0(1 + \rho) = 1 \Leftrightarrow p_0 = \frac{1}{1+\rho}$, $p_1 = \frac{\rho}{1+\rho}$. Граничний розподіл $\tilde{p}_i = \lim_{t \rightarrow \infty} p_i(t)$: $e^{-(\lambda+\mu)t} \xrightarrow{t \rightarrow \infty} 0$, тому $\tilde{p}_0 = \frac{\mu}{\lambda+\mu} = \frac{1}{1+\rho}$ і $\tilde{p}_1 = \frac{\rho}{1+\rho}$ — співпадає зі стаціонарним.

Перехідна поведінка М/М/∞. Вважатимемо $N(0) = 0$ — спочатку в системі немає запитів. Інтенсивності $\lambda_k \equiv \lambda$, $\mu_k = k\mu$, тож система ДР така:

$$\begin{cases} p'_0(t) = -\lambda p_0(t) + \mu p_1(t), \\ p'_k(t) = -(\lambda + k\mu)p_k(t) + \lambda p_{k-1}(t) + (k+1)\mu p_{k+1}(t), \quad k \geq 1. \end{cases}$$

Використовуючи ДР з частинними похідними та твірні функції, можна показати (наприклад, Gross & Harris (1985)), що має місце

ТЕОР. Твірна функція для $\{p_k(t)\}_{k=0}^{\infty}$, $P(z; t) = \sum_{k=0}^{\infty} p_k(t) z^k =$
 $= \exp\{(z-1)(1-e^{-\mu t})\frac{\lambda}{\mu}\} = \exp\{-r(1-e^{-\mu t})\} \cdot \exp\{zr(1-e^{-\mu t})\}$

Розкладемо у ряд: $p_k(t) = \frac{1}{k!} [r(1-e^{-\mu t})]^k \exp\{-r(1-e^{-\mu t})\}$, де $r = \frac{\lambda}{\mu}$.

Граничний розподіл при $t \rightarrow \infty$, $\tilde{p}_k = \frac{r^k}{k!} e^{-r}$ (і такий же стаціонарний) — розподіл Пуассона, отриманий раніше для цієї СМО.

Перехідна поведінка М/М/1. Аналітичне відшукування $\{p_k(t)\}$ досить складне. Перший розв'язок цієї задачі (приблизно через півстоліття після роботи Ерланга, де вперше розглянуто М/М/1) опублікували Ledermann & Reuter (1954), застосувавши спектральний аналіз ПНЗ загального виду. Інші способи запропонували Bailey (1954) і трохи згодом — Champagnowne (1956).

ТЕОР. Нехай $N(0) = s$ і $I_k(y) = \sum_{j=0}^{\infty} \frac{(y/2)^{k+2j}}{j!(k+j)!}$ — різновид функції Бесселя

І роду. Тоді $p_k(t) = e^{-(\lambda+\mu)t} \times$
 $\times \left[\rho^{\frac{k-s}{2}} I_{k-s}(2t\sqrt{\lambda\mu}) + \rho^{\frac{k-s-1}{2}} I_{k+s+1}(2t\sqrt{\lambda\mu}) + (1-\rho)\rho^k \sum_{j=k+s+2}^{\infty} \rho^{-\frac{j}{2}} I_j(2t\sqrt{\lambda\mu}) \right]$

Abatt & Whitt (1989) відмітили, що безпосередньо користуватись цією формулою незручно через обчислення суми ряду з видозмінених функцій Бесселя, і розглянули кілька способів ефективнішого обчислення $p_k(t)$.

Із властивостей функцій Бесселя випливає, що $p_k(t) \xrightarrow{t \rightarrow \infty} p_k = (1-\rho)\rho^k$ при $\rho = \frac{\lambda}{\mu} < 1$; коли $\rho \geq 1$, граничного розподілу не існує — $p_k(t) \xrightarrow{t \rightarrow \infty} 0$.

Симулятор М/М/п на Python

```
import random

def multvecbynum(vec, num):
    for i in range(len(vec)):
        vec[i] *= num
    return vec

class Runstat:
    def __init__(self, queue_len, busytimes, mean_queue_len,
                 mean_queue_wait):

        self.queue_len = queue_len
        self.busytimes = busytimes
        self.mean_queue_len = mean_queue_len
        self.mean_queue_wait = mean_queue_wait

class Systemul:
    def reset(self, servernum = -1, arrival_rate = -1,
              service_rate = -1):

        self.eventlist = []
        self.queuelist = []
        self.totalwaittime = 0
        self.totalwaitnum = 0
        self.totalqueue_len = 0
        self.totalqueue_num = 0
        if arrival_rate != -1:
            self.arrival_rate = arrival_rate
        if service_rate != -1:
            self.service_rate = service_rate
        if servernum == -1:
            servernum = len(self.serverstate)
        self.serverstate = []
        self.serverbusytime = []
        for i in range(servernum):
            self.serverstate.append(0)
            self.serverbusytime.append(0)

    def __init__(self, servernum = 1, arrival_rate = 1.0,
                 service_rate = 1.0):
        self.reset(servernum, arrival_rate, service_rate)

    def r_arrival_dur(self):
        return random.expovariate(self.arrival_rate)

    def r_service_dur(self):
        return random.expovariate(self.service_rate)
```

```

❏ def select_free_server(self):
    for i in range(len(self.serverstate)):
        if self.serverstate[i] == 0:
            return i
    return -1

def queue_size(self):
    return len(self.queueulist)

def queue_put(self, req):
    self.queueulist.append(req)

def queue_take(self):
    return self.queueulist.pop(0)

def update_totalwait(self, waitdur):
    self.totalwaittime += waitdur
    self.totalwaitnum += 1

def update_totalqueue(self, length):
    self.totalqueueulen += length
    self.totalqueueenum += 1

def add_event(self, etime, etype, edata):
    event = [etime, etype, edata]
    enum = len(self.eventlist)
    if (enum == 0) or (etime >= self.eventlist[enum - 1][0]):
        self.eventlist.append(event)
    elif etime < self.eventlist[0][0]:
        self.eventlist.insert(0, event)
    else:
        s = 0
        e = enum - 1
        while (e - s) > 1:
            m = (s + e) // 2
            if etime >= self.eventlist[m][0]:
                s = m
            else:
                e = m
        self.eventlist.insert(e, event)

def take_first_event(self):
    return self.eventlist.pop(0)

def load_server(self, ctime, s):
    self.serverstate[s] = 1
    dur = self.r_service_dur()
    self.serverbusytime[s] += dur
❏ ❏ self.add_event(ctime + dur, "server.released", [s])

```

```

❧ def dispatch_next_event(self):
    event = self.take_first_event()
    etime, etype, edata = (event[0], event[1], event[2])
    if etype == "request.arrived":
        self.add_event(etime + self.r_arrival_dur(), ¶
                        "request.arrived", [])
        self.update_totalqueue(self.queue_size())
        if self.queue_size() > 0:
            self.queue_put([etime])
        else:
            s = self.select_free_server()
            if s == -1:
                self.queue_put([etime])
            else:
                self.load_server(etime, s)
                self.update_totalwait(0)
    elif etype == "server.released":
        s = edata[0]
        self.serverstate[s] = 0

        if self.queue_size() > 0:
            ptime = self.queue_take()[0]
            self.update_totalwait(etime - ptime)
            self.load_server(etime, s)
    return etime

def run(self, duration):
    self.reset()
    self.add_event(0, "request.arrived", [])
    stime = 0
    while stime <= duration:
        stime = self.dispatch_next_event()
    return Runstat(self.queue_size(),
                  multvecbynum(self.serverbusytime, 1.0 / duration),
                  self.totalqueuelen / self.totalqueuenum,
                  self.totalwaittime / self.totalwaitnum)

def main():
    random.seed()
    systemul = Systemul(servernum = 5, arrival_rate = 0.9, ¶
                        service_rate = 0.2)

    rstat = systemul.run(1000)
    print("Final queue length: " + str(rstat.queue_len))
    print("Relative server busy times: " + str(rstat.busytimes))
    print("Mean queue length: " + str(rstat.mean_queue_len))
    print("Mean queue waiting time: " + str(rstat.mean_queue_wait))

main()

```

Симулятор М/М/п на R

```
eventlist <- NULL;
queuelist <- NULL;

serverstate <- rep(0, 5);
serverbusytime <- NULL;

totalqueueulen <- NULL;
totalqueueenum <- NULL;
totalwaittime <- NULL;
totalwaitnum <- NULL;

r.arrival.dur <- function() {
  return (rexp(1, rate=0.9));
}
r.service.dur <- function() {
  return (rexp(1, rate=0.2));
}

select.free.server <- function() {
  snum <- length(serverstate);
  for (i in 1:snum) {
    if (serverstate[i] == 0) {
      return (i);
    }
  }
  return (-1);
}

queue.size <- function() {
  return (length(queuelist));
}

queue.put <- function(curtime) {
  len <- length(queuelist);
  queuelist[[len + 1]] <- list(curtime);
}

queue.get <- function() {
  len <- length(queuelist);
  ptime <- queuelist[[1]][[1]];
  if (len == 1) {
    queuelist <- list();
  } else {
    queuelist <- queuelist[2:len];
  }
  return (ptime);
}
```

```

update.totalwait <- function(waitdur) {
  totalwaitnum <<- totalwaitnum + 1;
  totalwaittime <<- totalwaittime + waitdur;
}

update.totalqueue <- function(qlen) {
  totalqueueenum <<- totalqueueenum + 1;
  totalqueueelen <<- totalqueueelen + qlen;
}

add.event <- function(etime, etype, edata) {
  event <- list(etime, etype, edata);
  enum <- length(eventlist);

  new.eventlist <- list();

  if (enum == 0) {
    new.eventlist[[1]] <- event;
  } else if (etime < eventlist[[1]][[1]]) {
    new.eventlist[[1]] <- event;
    new.eventlist[2:(enum + 1)] <- eventlist;
  } else if (etime >= eventlist[[enum]][[1]]) {
    new.eventlist <- eventlist;
    new.eventlist[[enum + 1]] <- event;
  } else {
    S <- 1; E <- enum;
    while ((E - S) > 1) {
      M <- (E + S) %/% 2;
      if (etime >= eventlist[[M]][[1]]) {
        S <- M;
      } else {
        E <- M;
      }
    }

    new.eventlist[1:(E-1)] <- eventlist[1:(E-1)];
    new.eventlist[[E]] <- event;
    new.eventlist[(E + 1):(enum + 1)] <- eventlist[E:enum];
  }

  eventlist <<- new.eventlist;
}

```

```

take.first.event <- function() {
  event <- eventlist[[1]]; enum <- length(eventlist);
  if (enum == 1) {
    eventlist <- list();
  } else {
    eventlist <- eventlist[2:enum];
  }
  return (event);
}

load.server <- function(curtime, s) {
  serverstate[s] <- 1;
  dur <- r.service.dur();
  serverbusytime[s] <- serverbusytime[s] + dur;
  add.event(curtime + dur, "server.released", list(s));
}

dispatch.next.event <- function() {
  event <- take.first.event();
  etime <- event[[1]];
  etype <- event[[2]];
  edata <- event[[3]];

  if (etype == "request.arrived") {
    add.event(etime + r.arrival.dur(), "request.arrived", list());
    update.totalqueue(queue.size())

    if (queue.size() > 0) {
      queue.put(etime);
    } else {
      s <- select.free.server();
      if (s == -1) {
        queue.put(etime);
      } else {
        load.server(etime, s);
        update.totalwait(0);
      }
    }
  } else if (etype == "server.released") {
    s <- edata[[1]];
    serverstate[s] <- 0;
    if (queue.size() > 0) {
      ptime <- queue.get()
      update.totalwait(etime - ptime);
      load.server(etime, s);
    }
  }
  return (etime);
}

```

```

queueму.run <- function(duration) {
  eventlist <- list();
  add.event(0, "request.arrived", list());

  queuelist <- list();

  serverstate <- rep(0, length(serverstate));
  serverbusytime <- rep(0, length(serverstate));

  totalqueuelen <- 0;
  totalqueuenum <- 0;

  totalwaittime <- 0;
  totalwaitnum <- 0;

  stime <- 0;
  while (stime < duration) {
    stime <- dispatch.next.event();
  }

  return (list("Final queue length" = queue.size(),
    "Relative server busy times" = serverbusytime / duration,
    "Mean queue length" = totalqueuelen / totalqueuenum,
    "Mean waiting time" = totalwaittime / totalwaitnum));
}

>queueму.run(1000)

```

Позначення

ОЗН. — Означення.

РЕС. — Нагадування.

ТВ. — Твердження.

ЗАУВ. — Зауваження.

ПР. — Приклад.

В. — Властивість.

Н. — Наслідок.

ЛЕМА. — Лема.

ТЕОР. — Теорема.

◀ — Початок доведення.

▶ — Кінець доведення.

тоді́дот — тоді й тільки тоді.

⊗ — протиріччя.

Список рекомендованої літератури

- [1] Gross D. Fundamentals of Queueing Theory. 4th ed. [Text] / D. Gross, J. F. Shortle, J. M. Thompson, C. M. Harris. — Wiley. — 2008. — 512 p.
- [2] Harchol-Balter M. Performance modeling and design of computer systems: queueing theory in action. [Text] / M. Harchol-Balter. — Cambridge University Press. — 2013. — 573 p.
- [3] Гнеденко Б.В. Введение в теорию массового обслуживания. 2-е изд. [Текст] / Б. В. Гнеденко, И. Н. Коваленко. — М., Наука. — 1987. — 336 с.
- [4] Клейнрок Л. Теория массового обслуживания. [Текст] / Л. Клейнрок. — М., Машиностроение. — 1979. — 432 с.
- [5] Матвеев В.Ф. Системы массового обслуживания. [Текст] / В. Ф. Матвеев, В. Г. Ушаков. — М., Изд-во МГУ. — 1984. — 240 с.
- [6] Саати Т.Л. Элементы теории массового обслуживания и её приложения. [Текст] / Т. Л. Саати. — М., Сов. радио. — 1971. — 520 с.

Навчально-методичне видання

Гончаров Сергій Валерійович

КОНСПЕКТ ЛЕКЦІЙ
З ТЕОРІЇ МАСОВОГО ОБСЛУГОВУВАННЯ

ЧАСТИНА I

Видання друкується в авторській редакції

Відповідальний редактор *Біла К. О.*
Технічний редактор *Олексенко Н. С.*

Підп. до друку 24.04.18. Формат 60x84 1/16.
Ум. друк. арк. 1,6. Тираж 35 пр. Зам. № 0418-04/1.

Видавець та виготовлювач СПД Біла К. О.

Свідоцтво про внесення до Державного реєстру
суб'єктів видавничої справи ДК № 3618 від 06.11.09

Надруковано на поліграфічній базі видавничо-поліграфічного центру «Імпакт»
Україна, 49000, м. Дніпро, пр. Д. Яворницького, 111, оф. 2

+38 (099) 7805049; +38 (067) 2100256

www.impact.dp.ua
e-mail: impact.dnepr@gmail.com