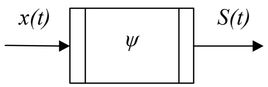
3. ЭЛЕМЕНТЫ ТЕОРИИ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ

3.1 Марковские цепи и потоки событий [11]

По определению, функция времени *x(t)*, которая принимает при каждом фиксированном значении аргумента случайное значение, называется ***случайным процессом***.

Пусть такой процесс воздействует на вход системы.



Наблюдаемое состояние системы при этом есть *S(t)*.

Определение. Если для любого момента времени t_0 вероятностные характеристики процесса *s(t)* в ***будущем*** зависят только от его состояния s_0 в данный момент и не зависят от ***предыстории процесса***, то процесс называется ***марковским***.

Примеры. 1) Поезд, идущий по маршруту, в график не укладывается. В какой-то момент опаздывает, а где то должен наверстывать. 2) Интенсивность обстрела цели определяется боезапасом и количеством стволов и не зависит напрямую от предыдущих стрельб.

Существует парадокс, согласно которому любой процесс может стать марковским, если все параметры из “прошлого” от которых зависит “будущее”, включить в настоящее. Вот часы: будут ли идти через месяц - два? Если их состояние “идут”, то процесс не является марковским. Если учитывать время завода или вставки батарейки, то он становится таковым.

В технике и при исследовании операций используется класс марковских процессов с ***дискретными состояниями*** и ***непрерывным временем***.

Основные черты таких процессов:

- состояния (значения) образуют конечное перечислимое множество;
- переход из состояния в состояние осуществляется скачком (мгновенно);
- одновременно процесс может находиться только в одном из состояний.

Математической моделью марковского процесса является ***матрица переходных вероятностей***

$$P(t) = \begin{pmatrix} p_{11}(t) & p_{12}(t) & \dots & p_{1k}(t) \\ p_{21}(t) & p_{22}(t) & \dots & p_{2k}(t) \\ \dots & \dots & p_{ij}(t) & \dots \\ p_{m1}(t) & p_{m2}(t) & \dots & p_{mk}(t) \end{pmatrix},$$

в которой элемент $P_{i,j}$ представляет вероятность перехода из *i*-ого состояния в *j*-ое.

Очевидно, что сумма по строке матрицы *P*(*t*) при фиксированном значении *t*, равна единице, поскольку состояния процесса образуют полный ансамбль.

Процессы такого рода удобно представлять в виде графа, где вершины соответствуют состояниям процесса, а дуги изображают возможные переходы между состояниями.

Ситуация, когда нет перехода из *i*-ого состояния в *j*-ое, в матрице переходных вероятностей соответствующая координата будет нулевой.

Частный случай марковского процесса – процесс с ***дискретными состояниями*** и с ***дискретным временем***, называемый ***цепью Маркова***.

Условимся считать переход из одного состояния в другое ***событием***. Пусть события происходят в заранее неизвестные моменты времени. Процесс, связанный с наступлением событий будем называть ***потоком событий*** или просто – ***потоком***.

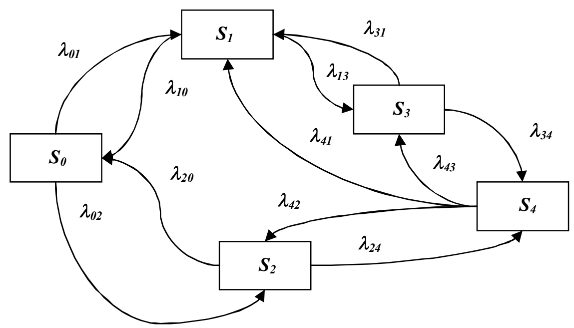
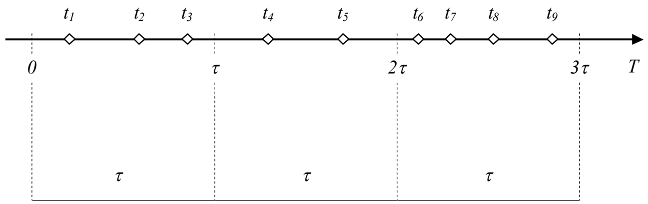


Рисунок 3.1 – Граф переходов между состояниями марковского процесса

Если события, происходящие в потоке, имеют одинаковую природу, поток называется ***однородным***. Для однородного потока матрица переходных вероятностей не зависит от времени

Однородные потоки событий являются одним из предметов изучения дисциплин исследования операций вообще и систем массового обслуживания, в частности.



Для оценки характеристик потоков оперируют понятиями теории вероятностей:

- среднее число событий, произошедших за единицу времени (интенсивность);
- среднее время между событиями.

3.2. Простейший поток событий [24, 25, 19, 50, 51]

В качестве ***простейшего потока*** в теории систем массового обслуживания (СМО) выбран поток, обладающий следующими свойствами:

- стационарен, по крайней мере, в широком смысле;
- ординарен (ни какие два события в потоке не происходят одновременно, всегда существует такой минимальный квант времени τ, в течение которого происходит только одно событие);
- не имеет последействия (события потока не связаны между собой).

Причинами выбора послужило:

- к простейшему потоку СМО приспособится наиболее трудно, поэтому система, рассчитанная на обработку простейшего потока, будут работать надежно при обработке других потоков, если ***их интенсивности одинаковы***;
- если на вход системы поступает одновременно несколько потоков разной структуры, то механическое суммирование этих потоков даёт поток, близкий к простейшему;
- относительная простота и возможность получения решения в аналитичекой форме для большинства практических приложений.

Можно считать, что простейший поток играет роль в теории СМО аналогичную нормальному закону распределения вероятностей.

3.3. Математические модели потоков

В литературе [6, 19, 25, 43, 51] простейший поток именуется пуассоновским, поскольку описывается моделью Пуассона:

$$P_k(\tau) = \frac{(\lambda\tau)^k}{k!} \cdot e^{-\lambda\tau},$$

которая характеризует вероятность наступления k событий за отрезок времени, равный τ, λ – интенсивность потока, равная математическому ожиданию числа событий, происходящих в единицу времени.

Вероятность отсутствия заявок за время определится выражением

$$P_k(\tau) = \mathrm{e}^{-\lambda\tau},$$

а отсутствие оных есть

$$P_k(\tau) = 1 - \mathrm{e}^{-\lambda\tau}.$$

Отсюда плотность распределения вероятностей наступления событий

$$f(t) = \lambda\,\mathrm{e}^{-\lambda t}, \qquad t > 0, \tag{3.1}$$

которое называется показательным или экспоненциальным распределением.

Для этого распределения характерно равенство

$$m_\tau = \sigma_\tau = \frac{1}{\lambda} -$$

– средний интервал времени между соседними заявками, совпадает со среднеквадратическим его отклонением.

В этом случае известный в статистике коэффициент вариации случайной величины

$$V_\tau = \frac{\sigma_\tau}{m_\tau} = 1.$$

Рассмотрим некую периодическую или равномерно дискретизированную последовательность. В этом случае $m_\tau = T_0$ – шагу дискретизации или периоду, а $\sigma_\tau = 0$, откуда имеем нулевой коэффициент вариации.

Отсюда становится ясным, почему к обработке простейшего потока труднее всего приспособится – он обладает максимальным коэффициентом вариации.

Если простейший поток прореживается, то есть сохраняется каждое *k*-ое событие, а остальные не учитываются (отбрасываются, отбраковываются), то возникает ***поток Эрланга*** *k*-ого порядка, *k =1, 2,…*

Плотность распределения вероятности такого потока описывается формулой:

$$f(t) = \frac{\lambda(\lambda \cdot t)^{k-1}}{(k-1)!} e^{-\lambda t}, \qquad t > 0.$$

Отметим, что простейший поток можно считать эрланговским при $k = 1$.

Для потока Эрланга характерно

- $m_k = \dfrac{k}{\lambda}$ – математическое ожидание, а
- $\sigma_k^2 = \dfrac{k}{\lambda^2}$ – дисперсия времени между наступлениями событий.

Иногда указанные формулы представляют в виде

$$m_k = \frac{1}{\lambda_k} \text{ и } m_k = \frac{1}{k\lambda_k^2}, \text{ где } \lambda_k = \frac{\lambda}{k},$$

но независимо от формы представления, коэффициент вариации определяется как

$$V_k = \frac{1}{\sqrt{k}}.$$

Поток Эрланга обладает следующими свойствами:

- стационарен;
- ординарен;
- обладает последействием;
- при возрастании *k* становится почти периодическим.

Когда модель потока априори не известна, её строят, по мере обретения информации о потоке в виде интегральной модели

$$f_k(\tilde{\tau}) = \tilde{\lambda} \cdot \int_0^{\tilde{\tau}} [\varphi_{k-1}(\tau) - \varphi_k(\tau)] d\tau \text{, где}$$

$\varphi_0(\tau)$- функцией Пальма.

$\varphi_k(\tau)$ - функция Пальма-Хинчина, k=1,2,…

Потоки, описываемые таким образом, называются потоками Пальма или ***рекуррентными*** в силу ядра интеграла. Интегральная модель достаточно достоверно отображает потоки, встречающиеся в практической деятельности.

Полагая функцию распределения вероятностей равной

$$f(t) = 1 - \varphi_0(t),$$

$\varphi_0(t)$ – определяемой формулой (3.1), а $\tilde{\lambda} = \lambda$, можно перейти к простейшему потоку.

3.4. Модель Колмогорова для описания систем с вероятностными состояниями [11, 51]

Выше было отмечено, что процессы, протекающие в системах, удобно представлять в виде графа состояний, как это показано на рисунке 3.1. На рисунке обозначено: S_i – состояния системы, λ_{ij} – потоки, переводящие процесс из *i*-го состояния, в *j*-ое состояние.

Колмогоров предложил описывать такие системы и процессы в виде системы дифференциальных уравнений, названных в его честь уравнениями Колмогорова.

Уравнения к конструируют по следующим правилам:

- число уравнений определяется количеством состояний системы (процесса);
- левая часть – производная по времени вероятности нахождение системы в *i*-ом состоянии;
- правая часть – сумма произведений вероятностей нахождения системы в *j*-ых состояниях, на интенсивности потоков, переводящих систему в *i*-е состояние из *j*-ых (так называемая, взвешенная сумма), за вычетом вероятности *i*-го состояния, умноженной на сумму интенсивностей потоков, выводящих систему из *i*-го состояния.

$$\frac{dP_i(\tau)}{d\tau} = \sum_{J+} \lambda_{ij} P_j(\tau) - P_i(\tau) \sum_{J-} \lambda_{ij}.$$

Для системы, изображённой на рисунке 3.1 можно построить следующую систему дифференциальных уравнений Колмогорова:

$$\left.\begin{aligned}
\frac{dP_0(\tau)}{d\tau} &= \lambda_{10}P_1(\tau) + \lambda_{20}P_2(\tau) - P_0(\tau)(\lambda_{01} + \lambda_{02}),\\
\frac{dP_1(\tau)}{d\tau} &= \lambda_{01}P_0(\tau) + \lambda_{31}P_3(\tau) + \lambda_{41}P_4(\tau) - P_1(\tau)(\lambda_{10} + \lambda_{13}),\\
\frac{dP_2(\tau)}{d\tau} &= \lambda_{02}P_0(\tau) + \lambda_{42}P_4(\tau) - P_2(\tau)(\lambda_{20} + \lambda_{24}),\\
\frac{dP_3(\tau)}{d\tau} &= \lambda_{13}P_1(\tau) + \lambda_{43}P_4(\tau) - P_3(\tau)(\lambda_{31} + \lambda_{43}),\\
\frac{dP_4(\tau)}{d\tau} &= \lambda_{24}P_2(\tau) + \lambda_{34}P_3(\tau) - P_4(\tau)(\lambda_{41} + \lambda_{42} + \lambda_{43}).
\end{aligned}\right\}$$

3.5. Схема “гибели – размножения” и её модель [33, 43]

Пусть на вход системы воздействует поток событий, которые переводят систему в ряд ***последовательных*** состояний. В свою очередь, система тем или иным способом реагирует на события и обладает свойством ***регенерируемости***, то есть обладает способностью “противиться” потоку и последовательно возвращаться к исходным состояниям.

Граф состояния такой системы, как это представлено на рисунке 3.2, будет вытянут, по аналогии с объектами биологии (где он впервые был использован), его называют схемой “гибели – размножения” (или “размножения – гибели”).

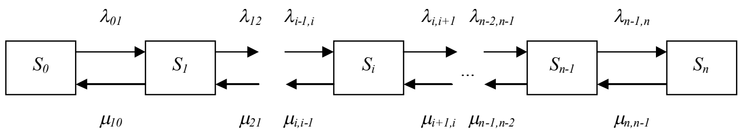


Рисунок 3.2 – Граф “гибели – размножения”

Для системы, описываемой с помощью такого графа можно составить уравнения Колмогорова в установившемся режиме. При этом предполагается, что:

существуют “финальные” вероятности состояний $\lim\limits_{t\to\infty} P_i(t) = P_i$;

состояния совокупно с их вероятностями образуют полный ансамбль $\sum\limits_{i=0}^{n} P_i = 1$.

Очевидно, что дифференциальные уравнения Колмогорова, при этом, преобразуются в систему линейных уравнений, благодаря чему представляется возможным найти финальные вероятности нахождения системы в соответствующих состояниях. Составим указную систему и отыщем финальные вероятности.

Для понимания дальнейшего изложения рассмотрим первые два уравнения системы и сравним их между собой.

$$\left.\begin{aligned}
0 &= \mu_{10}P_1 - \lambda_{01}P_0\\
0 &= \lambda_{01}P_0 + \mu_{21}P_2 - \mu_{10}P_1 - \lambda_{12}P_1
\end{aligned}\right\} \Rightarrow 0 = \mu_{21}P_2 - \lambda_{12}P_1.$$

Сама же система имеет вид

$$\begin{cases}
\mu_{10}P_1 = \lambda_{01}P_0,\\
\mu_{21}P_2 = \lambda_{12}P_1,\\
\ldots\\
\mu_{i,i-1}P_i = \lambda_{i-1,i}P_{i-1},\\
\ldots\\
\mu_{n-1,n}P_n = \lambda_{n-1,n}P_{i-1},\\
P_0 + P_1 + \ldots + P_i + \ldots + P_n = 1.
\end{cases}$$

Эта система путем последовательных подстановок легко решается:

$$\frac{+\begin{cases}
P_1 = \dfrac{\lambda_{01}}{\mu_{10}}P_0\\
P_2 = \dfrac{\lambda_{12}}{\mu_{21}}P_1 = \dfrac{\lambda_{01}\cdot\lambda_{12}}{\mu_{10}\cdot\mu_{21}}P_0\\
\ldots\\
P_i = \dfrac{\lambda_{01}\cdot\lambda_{12}\cdot\ldots\cdot\lambda_{i-1,i}}{\mu_{10}\cdot\mu_{21}\cdot\ldots\cdot\mu_{i,i-1}}P_0\\
\ldots
\end{cases}}{P_0 + \sum\limits_{i=1}^{n} P_i = 1} \tag{3.2}$$

Отсюда находится вероятность нахождения системы в начальном состоянии S_0:

$$P_0 = \left[1 + \frac{\lambda_{01}}{\mu_{10}} + \frac{\lambda_{01}\cdot\lambda_{12}}{\mu_{10}\cdot\mu_{21}} + \ldots + \frac{\lambda_{01}\cdot\lambda_{12}\cdot\ldots\cdot\lambda_{i-1,i}}{\mu_{10}\cdot\mu_{21}\cdot\ldots\cdot\mu_{i,i-1}} + \ldots + \frac{\lambda_{01}\cdot\lambda_{12}\cdot\ldots\cdot\lambda_{n-1,n}}{\mu_{10}\cdot\mu_{21}\cdot\ldots\cdot\mu_{n,n-1}}\right]^{-1}, \quad (3.3)$$

а через нее выражаются все остальные вероятности, согласно (3.2).

3.6. Понятие СМО. Формулы Литтла [7, 20, 53]

Система массового обслуживания (СМО) – специфическая техническая система, предназначенная для обработки потока.

В её состав входят обрабатывающие (обслуживающие) единицы, которые называются ***каналами обслуживания***. Физически это линии телефонной связи, рабочие точки, душевые кабинки, выездные бригады скорой медицинской помощи, продавцы и тому подобное.

Подразумевается, что поток представлен ***требованиями (заявками) на обслуживание***, при этом, для краткости употребляют термин ***поток заявок***.

Считается, что процесс работы СМО является процессом с дискретными состояниями и непрерывным временем, марковский. Заявки однородны по своей природе, система описывается схемой “разложение-гибель”.

- Пусть процессы, протекающие в СМО, связаны со следующими потоками событий: поток заявок, поступающих на вход СМО;
- поток заявок, покидающих СМО.

При наступлении (установке) в системе стационарного режима, эти потоки стационарны, сколько заявок входит, в среднем, в систему, столько и выходит. Система переходит из состояния в состояние скачкообразно (смотри рисунок 3.3).

Пусть

- $x(t)$ – число заявок, поступивших в СМО;
- $y(t)$ – число заявок, оставивших СМО;
- $S(t)$ – текущее состояние СМО.

Состояние системы определяется числом заявок, в ней находящихся (смотри граф “разложение-гибель” на рисунке 3.1), как обслуживаемых в настоящий момент, так и находящихся в очереди запросов:

$$S(t) = x(t) - y(t).$$

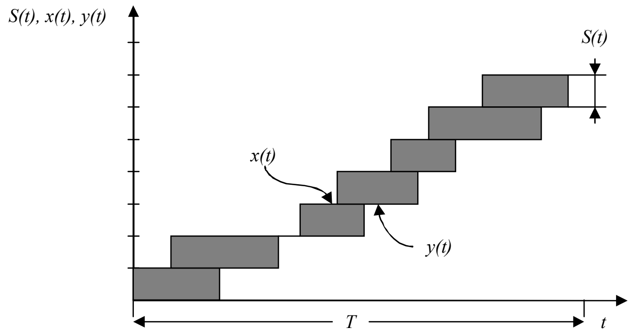


Рисунок 3.3 – Текущее состояние СМО в стационарном режиме

Среднее число заявок, находящихся в системе, есть

$$N_C = \lim_{T\to\infty}\frac{1}{T}\int_0^T S(t)dt,$$

где T – интервал наблюдения.

В силу дискретности $S(t)$, ординарности потоков (никакие два события несовместимы в пределах кванта времени) и установившегося режима (высота “ступеньки” единичная), как это показано на рисунке 3.3, интеграл можно заменить суммой

$$N_C = \lim_{T\to\infty}\frac{1}{T}\sum_i t_i,$$

где t_i – время наступления i-го события.

При этом, среднее число заявок, поступивших на вход системы за время наблюдения T, равно $m_з = \lambda T$. Учитывая это, избавимся от бесконечного предела, умножая и деля на λ.

$$N_C = \lim_{T\to\infty}\frac{1}{T}\cdot\frac{\lambda}{\lambda}\sum_i t_i = \lambda\cdot\frac{\sum_i t_i}{m_з}.$$

Обозначим среднее время пребывания заявки в системе как

$$T_C = \frac{\sum_i t_i}{m_3},$$

придём к выражению, называемому формулой Литтла [61]:

$$T_C = \frac{1}{\lambda} N_C. \qquad (3.4)$$

Справедливо утверждение, озвучивающее эту формулу.

При любом распределении времени обслуживания и любой дисциплине обслуживания, среднее время пребывания заявки в системе обратно пропорционально интенсивности входного потока и прямо пропорционально числу заявок в системе.

Подобные же рассуждения могут быть проделаны для получения связи среднего времени пребывания заявки в очереди и длины очереди.

$$T_{Оч} = \frac{1}{\lambda} N_{Оч}. \qquad (3.5)$$

При любом распределении времени обслуживания и любой дисциплине обслуживания, среднее время пребывания заявки в очереди обратно пропорционально интенсивности входного потока и прямо пропорционально числу заявок в очереди.

3.7. Примеры СМО. Одноканальная СМО с отказами [11]

К исходным данным для расчёта СМО данного типа следует отнести следующие параметры и допущения:

- система имеет один канал обслуживания;
- система имеет два состояния: «свободно» и «занято»;
- поток заявок, поступающих на вход системы, простейший, его интенсивность λ заявок в единицу времени задана;
- время обслуживания одной заявки определяется показательным законом $f(t) = \mu e^{-\mu t}, t > 0$, где μ – интенсивность обслуживания, заявок в единицу времени;
- заявка пришедшая, когда СМО свободна, принимается на обслуживание;

- заявка пришедшая, когда СМО занята, отвергается и аннулируется (пропадает).

Требуется определить:

- вероятности отказа и обслуживания;
- абсолютную и относительную пропускные способности.

Граф «размножения – гибели», в этом случае, простой

$$\begin{tikzcd} \boxed{S_0} \arrow[r, shift left, "\lambda"] & \boxed{S_1} \arrow[l, shift left, "\mu"] \end{tikzcd}$$

Его расчёт даёт значения финальных вероятностей

$$\begin{cases} P_0\lambda = P_1\mu, \\ P_0 + P_1 = 1 \end{cases} \Rightarrow \left| \begin{array}{l} P_1 = \rho P_0, \\ P_0 = (1+\rho)^{-1}, \end{array} \right.$$

где ρ – так называемая, ***приведённая интенсивность***. Последняя может трактоваться как

$$\rho = \frac{\lambda}{\mu} = \lambda \times \left(\frac{1}{T_{Обсуживания}} \right)^{-1} = \lambda \cdot T_{Обсуживания}$$

– среднее число заявок, ***поступающих на вход*** системы ***за время*** обслуживания ***одной*** заявки.

Окончательно имеем:

- вероятность отказа $P_1 = \dfrac{\lambda}{\mu + \lambda}$,
- вероятность обслуживания $P_0 = \dfrac{\mu}{\mu + \lambda}$.

  Последняя формула может трактоваться как предельное значение частоты

$$\lim_{N_{Поступающих} \to \infty} \frac{N_{Обслуженных}}{N_{Поступающих}} = P_0.$$

Таким образом, относительная пропускная способность (для данного типа СМО) совпадает с вероятностью нахождения системы в свободном состоянии и с вероятностью обслуживания.

Абсолютная же пропускная способность может быть определена двояко.

С одной стороны, уравнение $P_0\lambda = P_1\mu = \frac{\lambda\mu}{\lambda+\mu}$ описывает "баланс" обслуженных и поступающих заявок, а с другой стороны, можно воспользоваться понятийным аппаратом.

Абсолютная пропускная способность есть доля от числа заявок, поступающих на вход СМО в единицу времени, которые удаётся обслужить

$$A = q = \lambda P_0 = \frac{\lambda\mu}{\lambda+\mu}.$$

3.8. Примеры СМО. Многоканальная СМО с отказами [5, 25, 26, 35]

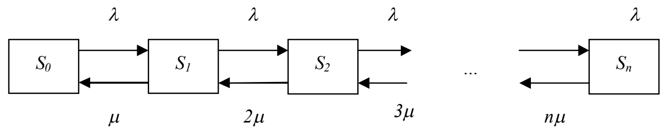
Расчёт многоканальной СМО с отказами получил наименования задача Эрланга. Исходными данными для проведения расчётов служат следующие:

- система имеет n однотипных каналов обслуживания:
- на вход поступает поток с интенсивностью λ заявок в единицу времени;
- интенсивность обслуживания одного канала составляет μ заявок в единицу времени;
- заявки принимаются на обслуживание, пока хотя бы один канал свободен;
- заявка, поступившая в момент времени, когда все каналы заняты, отвергается (пропускается, в терминах ПВО, когда эта задача впервые решалась).

Требуется рассчитать:

- вероятности обслуживания и отказа заявок;
- среднее число заявок, обслуживаемых в единицу времени (абсолютная пропускная способность);
- долю обслуживаемых заявок (относительная пропускная способность);
- среднее число занятых каналов.

Граф системы имеет вид (обратите внимание на интенсивности "возврата" в предыдущие состояния):



Вычисление финальных вероятностей для этого случая даёт следующее. Вероятность того, что система полностью свободна от заявок составляет

$$P_0 = \frac{1}{1+\rho+\frac{\rho^2}{2}+\ldots+\frac{\rho^n}{n!}},$$

где ρ – приведённая интенсивность.

По условию, пришедшая заявка будет отвергнута, когда все каналы заняты и система находится в состоянии S_n:

$$P_{OTK} = P_0\frac{\rho^n}{n!}.$$

Очевидно, что в противном случае заявка будет обслужена, и вероятность такого события есть:

$$P_{ОБСЛ} = 1 - P_{OTK}.$$

В данном случае, вероятность обслуживания, по понятным соображениям, совпадает по значению с относительной пропускной способностью. Отсюда находится абсолютная пропускная способность, как доля обслуженных заявок от заявок, поступающих в единицу времени:

$$A = q = \lambda P_{ОБСЛ}.$$

Среднее число занятых каналов

$$m_K = \frac{A}{\mu} = \rho P_{ОБСЛ}$$

есть отношение числа заявок, забираемых в единицу времени на обслуживание, к скорости обслуживания.

3.9 Примеры СМО. Одноканальная СМО с неограниченной очередью [5, 25, 26, 35]

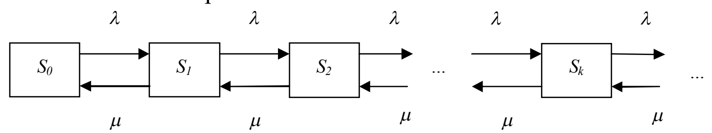
В модели СМО данного типа присутствуют следующие параметры и гипотезы:

- система имеет один канал обслуживания;
- поток заявок, поступающих на вход системы, простейший, его интенсивность λ заявок в единицу времени задана;
- время обслуживания одной заявки подчиняется показательному закону с интенсивностью обслуживания μ заявок в единицу времени;
- заявка пришедшая, когда СМО свободна, принимается на обслуживание;
- заявка пришедшая, когда СМО занята, помещается в очередь с дисциплиной обслуживания FIFO (первый пришедший обслуживается первым), длина которой теоретически ничем не ограничена.

Требуется определить:

- среднее число заявок в системе N_c;
- среднее время пребывания заявки в системе T_c;
- среднее число заявок в очереди $N_{оч}$.

Граф «размножения – гибели», в этом случае, имеет бесконечное число состояний и не ограничен



Используя ранее полученный результат решения модели Колмогорова (3.3), запишем

$$P_0 = \frac{1}{1+\rho+\rho^2+...+\rho^k+...}, \qquad (3.5)$$

где ρ – приведённая интенсивность.

В знаменателе дроби – геометрическая прогрессия с начальным членом 1 и знаменателем ρ. Сумма членов этой прогрессии будет существовать при $\rho < 1$ и равна $\frac{1}{1-\rho}$. В случае, когда $\rho \geq 1$, вероятность обслуживания (3.5) будет стремиться к нулю. Физически это означает, что очередь возрастает быстрее, нежели заявки попадают на обслуживание, и,

в перспективе, у «поздних» заявок неограниченно возрастает время ожидания в очереди и пребывания в системе.

Поэтому говорят, что если $\rho \geq 1$, то система с такими параметрами λ и μ нежизнеспособна. Совершенно очевидно, что для успешного функционирования одноканальной СМО с бесконечной очередью, необходимо обеспечить отношение $\lambda > \mu$.

Из (3.5), когда $\rho < 1$, следует, что вероятность отсутствия заявок в системе равна

$$P_0 = 1-\rho.$$

Вероятность же наличия k заявок в системе есть

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

Воспользуемся для нахождения среднего числа заявок в системе N_c известной формулой среднего дискретной случайной величины $m_x = \sum_i x_i P_i$. Имеем

$$N_C = \sum_{k=0}^{\infty} k(1-\rho)\rho^k,$$

применим элементарное алгебраическое преобразование:

$$N_C = (1-\rho)\rho\sum_{k=1}^{\infty} k\rho^{k-1}.$$

Под знаком суммы в последнем выражении стоит производная, поэтому

$$N_C = (1-\rho)\rho\sum_{k=1}^{\infty}\frac{d}{d\rho}\rho^k.$$

Так как операции дифференцирования и суммирования переставимы, а при перестановке под знаком суммы окажется геометрическая прогрессия, имеем ряд преобразований

$$N_C = (1-\rho)\rho\frac{d}{d\rho}\sum_{k=1}^{\infty}\rho^k = (1-\rho)\rho\frac{d}{d\rho}\left[\frac{\rho}{(1-\rho)}\right].$$

Окончательно среднее число заявок в СМО равно

$$N_c = \frac{\rho}{1-\rho}.$$

Воспользовавшись формулой Литтла (3.4), получим среднее время пребывания запроса в системе (время отклика):

$$T_c = \frac{N_c}{\lambda} = \frac{\rho}{\lambda(1-\rho)}.$$

Определим среднее число заявок, находящихся непосредственно на обслуживании

$$m = 0 \cdot P_0 + 1 \cdot (1 - P_0) = \rho .$$

Поэтому в очереди находится (средняя длина очереди), в среднем

$$N_{оч} = N_C - m = \frac{\rho^2}{1-\rho}.$$

заявок. Воспользовавшись формулой Литтла (3.5) получим среднее время ожидания в очереди:

$$T_{оч} = \frac{N_{оч}}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)}.$$

3.10. Пример решения практической задачи

Задача. Некая фирма работает 24 часа в сутки. В среднем, у 48 служащих возникает желание, чтобы размяться, по одному разу покрутить педали велотренажёра. Среднее время занятий на велотренажёре составляет 20 минут.

Служащие жалуются, что время ожидания у велотренажёра велико и просят у администрации установки ещё дополнительных велотренажёров.

Со своей стороны, дирекция считает, тренажёр занят только 2/3 всего времени, а установка дополнительного тренажёра – пустая трата времени.

Требуется вынести квалифицированное заключение по этому вопросу.

Воспользуемся для расчётов моделью СМО с бесконечной очередью.

Из условия задачи следует, что интенсивность обслуживания составит

$1/\mu = 20$ [мин]$\Rightarrow$ $\mu = 3$ [сотрудника/час].

Считая поток на входе тренажёрной комнаты пуассоновским, найдём его интенсивность:

$\lambda = 48/24 = 2$ [сотрудника/час].

Приведённая интенсивность, при этом, равна

$\rho = 2/3$.

Среднее время ожидания в очереди, составит:

$$T_{оч} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\left(\frac{2}{3}\right)^2}{2 \times \left(1 - \frac{2}{3}\right)} = \frac{1}{3}[\text{час}],$$

то есть сорок минут, а нахождение в спортивном зале, будет равно

$$T_c = \frac{\rho}{\lambda(1-\rho)} = \frac{\frac{2}{3}}{2 \times \left(1 - \frac{2}{3}\right)} = 1[\text{час}].$$

Всего же в системе будут, в среднем находится

$$N_c = \frac{\rho}{1-\rho} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$$

сотрудника фирмы, один из которых оздоровляется, а другой – бездельничает, ожидая своей очереди к велотренажёру.

Максимальное время в очереди для 90% времени составит

$$\pi(90\%) = T_{оч} \times \ln [10\rho] = 113{,}8 \text{ [мин]},$$

то есть, установка дополнительных велотренажёров необходима.

3.11. Сводные показатели эффективности СМО [7]

1. Вероятность потери требования (отказа) СМО.

2. Вероятность P_k занятости k приборов из n, для многоканальной СМО частными случаями являются:

- P_0 – все приборы свободны;
- $P_n = P_{отк}$ – все приборы заняты, если СМО без очереди.

3. Среднее число занятых приборов

$$N_{Зан} = \sum_{k=1}^{n} kP_k$$

характеризует степень занятости.

4. Среднее число свободных приборов

$$N_{Св} = \sum_{k=1}^{n} (n-k)P_k .$$

5. Коэффициент простоя приборов

$$K_{Пр} = \frac{N_{Св}}{n} .$$

6. Аналогично определяется коэффициент загрузки приборов или коэффициент занятости

$$K_{Зан} = \frac{N_{Зан}}{n} .$$

Если используется система с ожиданием в очереди, то дополнительно определяются:

7. Среднее время пребывания в очереди (ожидания)

$$T_{оч} = T_{ож} = \int_0^{\infty} t dP(t_{ож} > t) \text{ где } P(t_{ож} > t) = \sum_{k=0}^{n} \rho_k P_k (t_{ож} > t)$$

где $P_k(t_{ож} > t)$ – условная вероятность того, что время ожидания $t_{ож} > t$ при условии, что в момент поступления требования в систему в ней уже находилось k заявок.

8. Вероятность того, что время пребывания в очереди не превысит определённого, есть

$$P(t_{ож} < t) = \sum_{k=n}^{\infty} \rho_k P_k (t_{ож} < t) .$$

9. Средняя длина очереди, большей заданной длины n

$$N_{оч} = \sum_{k=n}^{\infty} (k-n)P_k .$$

10. Среднее число заявок в системе

$$N_C = N_{оч} + N_З = \sum_{k=1}^{\infty} kP_k .$$

11. Вероятность того, что число требований в очереди больше некоторого числа n

$$P_{k>n} = \sum_{k=n+1}^{\infty} P_k .$$

Существуют вполне очевидные связи показателей (характеристик)

- $n = N_З + N_{Св}$ по каналам обслуживания;
- $T_C = T_{обсл} + T_{оч}$ по времени.

В экономических расчётах применяется следующие показатели оценки СМО

- $q_{обсл}$ – стоимость обслуживания отдельного требования в СМО;
- $q_{ож}$ – стоимость потерь от ожидания в очереди;
- $q_{отк}$ – стоимость потерь от отказа заявке в обслуживании;
- $q_к$ – стоимость эксплуатации отдельного прибора в единицу времени;
- $q_{пк}$ – стоимость единицы времени простоя.

На базе этих показателей синтезируются следующие целевые функции в оптимизационных задачах:

- для систем с ожиданием $G = (q_{ож} \times T_{оч} + q_{пк} \times N_{Св} + q_к \times N_З) \times T;$
- для систем с отказами $G = (q_к \times N_З + q_{отк} \times P_{отк} \times \lambda) \times T;$
- для смешанных систем

$$G = (q_{пк} \times N_{Св} + q_{ож} \times T_{оч} + q_к \times N_З + + q_{отк} \times P_{отк} \times \lambda) \times T.$$

3.12. Вопросы для самоконтроля

1. Что такое поток?
2. Существуют ли реальные системы, в которых протекающие в них случайные процессы являются марковскими?
3. Что такое стационарность в широком и узком смысле?
4. Когда случайный процесс с непрерывным временем не обладает эргодическим свойством?
5. Обладает ли эргодическим свойством случайный процесс с непрерывным временем, имеющий бесконечное число состояний?
6. Как вы понимаете термин “ординарен”?
7. Как вы объясните, что означает отсутствие последействия?
8. Как связаны интенсивность потока и среднее время между событиями потока?
9. Когда оправдано использование предположения о простейшем характере потока заявок?
10. Каковы единицы измерения интенсивности потока?
11. Почему в моделях потоков положено $t \geq 0$?
12. Поясните, почему при больших значениях k поток Эрланга вырождается в периодическую последовательность?
13. По какой причине для проведения расчётов очень часто используется простейший (пуассоновский) поток?
14. Как влияет изменение параметра k на вид (форму) распределения?
15. Как влияет изменение параметра λ на вид (форму) распределения?
16. В чём заключается физический смысл понятия интенсивности входного потока?
17. Пояснить физический смысл понятия интенсивности обслуживания.
18. Каковы размерность интенсивности потока и величины обратной ей?
19. Какие допущения положены в основание модели одноканальной СМО с отказами?
20. Какие допущения положены в основание модели многоканальной СМО с бесконечной очередью?
21. Какие допущения положены в основание модели многоканальной СМО с ограниченной очередью?

22. Почему в СМО с бесконечной очередью даже если интенсивность обслуживания выше интенсивности поступления заявок, возникают очереди? В каких случаях они не возникают?
23. Как изменятся характеристики СМО при введении ограничений на длину очереди?
24. Как изменятся характеристики СМО при введении ограничений на время пребывания в очереди?
25. Что такое "установившийся режим" в СМО?
26. Какой вид имеют уравнения Колмогорова в "установившемся режиме"?
27. Каков физический смысл понятия "приведенная интенсивность"?
28. Пояснить физический смысл понятия интенсивности обслуживания.
29. Какие допущения положены в основание многоканальной модели СМО с отказами?
30. Как, по Вашему мнению, изменятся характеристики СМО при введении дополнительного числа каналов
31. Приведите примеры n-канальных СМО, используемых как в быту, так и в вычислительной технике.
32. Что включает в себя понятие n-канальная СМО??
33. Как вид распределения времени обслуживания заявки влияет на зависимости коэффициента загрузки и времени ожидания заявки от интенсивности входного потока заявок?