# Содержание

# НЕПРЕРЫВНЫЕ СОБЫТИЙНО-СТОХАСТИЧЕСКИЕ МОДЕЛИ ($Q$-СХЕМЫ)

Математические модели класса $Q$-схем предназначены для интерпретации выделенного в простом агрегате $\bar{\mathcal{A}}$ отношения $r\left(p_1, p_2, \cdots, p_{n_r}\right) \in R_{\bar{\mathcal{A}}}\left\{\vec{x}, \vec{v}, \vec{h}, \vec{y}, t\right\}$ как процесса обслуживания, удовлетворяющего следующим условиям:

- параметры отношения $r\left(p_1, p_2, \cdots, p_{n_r}\right)$ могут быть как детерминированными, так и случайными (*стохастическими*);
- параметр времени $t$ описывается моделью *непрерывного времени* - $M_{\mathcal{T}}(t_0, \delta t)$;
- входное воздействие $\vec{x}$ интерпретируется как появление в случайный момент непрерывного времени заявки на обслуживание процессом;
- процесс обслуживания $r\left(p_1, p_2, \cdots, p_{n_r}\right)$ представляется в непрерывном времени в дискретном конечном фазовом пространстве $Z_{r(p_1, p_2, \cdots, p_{n_r})}$ как последовательные скачки перехода ("одномоментно") в случайный момент времени $t_{i+1}$ из текущего $i$-ого состояния $z^{t_i}_{r(p_1, p_2, \cdots, p_{n_r})} \in Z_{r(p_1, p_2, \cdots, p_{n_r})}$ в случайное новое состояние $z^{t_{i+1}}_{r(p_1, p_2, \cdots, p_{n_r})} \in Z_{r(p_1, p_2, \cdots, p_{n_r})}$;
- выходная характеристика $\vec{y}$ интерпретируется как результат обслуживания процессом заявки $\vec{x}$, полученный в случайный момент непрерывного времени.

Такая модель поведения, например, характерна для информационных систем, функционирующих в режиме "клиент-сервер". В этом случае работу информационной системы можно интерпретировать как обслуживание сервером потока абстрактных заявок, поступающих от клиента через случайные интервалы времени. Например, сервер базы данных обслуживает случайные потоки заявок, поступающих от параллельных процессов, функционирующих в узлах компьютерной сети. Очевидно, что важной характеристикой качества работы таких систем является время обслуживания заявок. Функционирующие подобным образом системы называют *системами массового обслуживания* (СМО). Для работы таких систем характерным является случайное появление во времени заявки на обслуживание и/или случайное время обработки заявки. Так как обработка заявок не осуществляется мгновенно, то могут образовываться очереди заявок, ожидающих обслуживания. Поэтому СМО ещё называют *системами с очередями* (англ. $Q$ueueing system). Математические модели, используемые для моделирования систем массового обслуживания, относят к классу $Q$-схем. Использование моделей СМО для априорной оценки таких показателей информационной системы как среднее время обслуживания заявки, средняя длина очереди заявок, ожидающих обслуживания, вероятность отказа в обслуживании имеет важное значения на этапе проектирования.

## Модель обслуживания

### Простейший прибор обслуживания

При моделировании отношения $r(p_1, p_2, \cdots, p_{n_r})$ как сложной системы обслуживания полагают, что она может быть структурно представлена в виде взаимосвязанных по входу/выходу простейших приборов обслуживания (Рисунок 1).

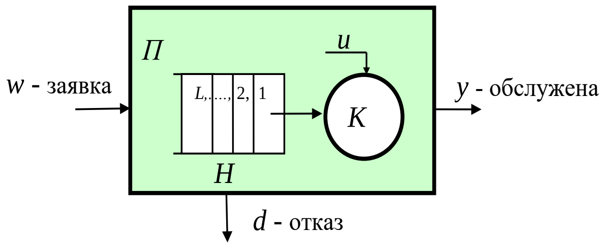



Рисунок 1 Простейший прибор обслуживания

Простейший прибор обслуживания $\Pi$ включает в себя ограниченный накопитель заявок $H$ и канал обработки заявки - $K$. Поступающая в $\Pi$ в некий момент непрерывного времени заявка $w$ может получить отказ в обслуживании, если накопитель $H$ полон, и тогда покидает прибор как не обслуженная заявка - $d$. Результаты обработки заявок $y$ на выходе одного прибора обслуживания могут поступать в качестве заявок на обслуживание $w$ на вход другого прибора обслуживания, связанного с ним по входу/выходу, образуя более сложную систему обслуживания.

Относительно поступившей на обслуживание заявки простейший прибор реализует элементарный акт обслуживания, в котором выделяют три основные составляющие:

- отказ заявке в обслуживании - $d$;
- приём заявки $w$ на обслуживание;
- ожидание заявкой $w$ обработки;
- обработка заявки и выдача результата обработки $y$.

Если отказ $d$ не последовал, то прибор *П* начинает элементарное обслуживание заявки, а именно:

- принимает заявку в накопитель заявок $H$ – это очередь к каналу обработки $K$, в которой могут одновременно находиться $l = 0,1, \dots, L$ заявок ($L$ - ёмкость накопителя $H$),
- при освобождении канала $K$ он ожидает поступления на обработку очередной заявки;
- из не пустого накопителя очередная заявка мгновенно поступает на обработку.

Текущее состояние $z \in Z$ прибора *П* определяется парой <"*состояние накопителя* $H$", "*состояние канала* $K$">, и изменяется скачком, когда в текущий момент непрерывного времени возникают случайные события, формирующие следующие потоки событий:

– $w$ – событие приёма заявки в прибор *П* на обслуживание – *поток принятых заявок на обслуживание*;

– $d$ –событие отказа заявке в обслуживании прибором *П* по причине отсутствия свободных мест в накопителе $H$ – *поток отказов в обслуживании*.

– $u$ - событие *завершения обработки заявки* каналом $K$ - *поток завершения обработки заявок*;

– $y$ – событие завершения обслуживания заявки прибором *П* - *поток завершения обслуживания заявок*;

Функционирование прибора обслуживания в непрерывном времени можно представить в виде сменяющихся во времени состояний прибора *П*, обусловленных событиями приёма заявки на обслуживание прибором *П* или завершения обработки заявки каналом $K$.

Определим конечное множество $Z$ дискретных состояний прибора *П*, и конечное дискретное множество событий $E$. Определим следующие состояния множества $Z$:

$z_0$ – канал *K* свободен, $l$=0 - накопитель заявок *H* пуст;

$z_1$ – канал *K* занят, $l$=0;

$z_2$ – канал *K* занят, $l$=1;

$z_3$ – канал *K* занят, $l$=2;

$z_4$ – канал *K* занят, $l$=3;

…

$z_{L+1}$ – канал *K* занят, $l$=$L$ - накопитель заявок *H* заполнен;

Множество событий $E$ так же будет дискретным и конечным и содержать следующие события:

$w$ – заявка принята прибором *П* на обслуживание;

$u$ – завершена обработка заявки каналом *K*;

Далее будем полагать, что события $w$ и $u$ не могут произойти одновременно. В каждый текущий момент времени $t \in (\dot{t}, \dot{t} + \Delta t)$, $\Delta t$ - текущий интервал времени, длящийся с момента $\dot{t}$ предыдущей смены состояния, в течение которого прибор обслуживания не меняет своё состояние. Интервал времени $\Delta t$ завершается случайно по событию либо $w$, либо $u$, и известны увеличивающиеся вероятности наступления в текущий момент времени $t = \dot{t} + \Delta t$ событий $w$ и $u$: $p_w(\dot{t}, \Delta t)$, $p_u(\dot{t}, \Delta t)$, с увеличением интервала $\Delta t$. Тогда вероятность того, что система сохранит своё текущее состояние в момент времени $t = \dot{t} + \Delta t$ будет равна: $1-\left(p_w(\dot{t},\Delta t)+p_u(\dot{t},\Delta t)\right)$. В результате прибор обслуживания *П* можно представить как конечный функционирующий в непрерывном времени событийно-стохастический автомат, с двумя вероятностными входными значениями – $X=\{w,u\}$, множеством вероятностных выходных значений $Y=\{\emptyset,y\}$, и множеством состояний - $Z=\{z_0,z_1,z_2,\ z_3,z_4,\dots,\ z_{L+1}\}$. Функционирование такого автомата можно изобразить в виде диаграммы состояний и переходов дискретного непрерывно-стохастического автомата (Рисунок 1):

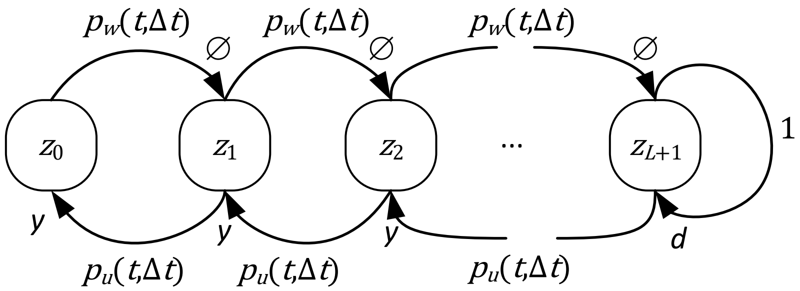



Рисунок 1 – Непрерывный событийно-стохастический автомат

Вероятность перехода прибора *П* за время $\Delta t$ из текущего состояния в следующее состояние равна вероятности возникновения в периоде времени $(\dot{t},\dot{t}+\Delta t)$ либо события $w$ – заявка принята на обслуживание, либо события $u$ – обработка заявки каналом *К* завершено. Вероятность остаться в текущем состоянии – это вероятность того, что события $w$ или $u$ в текущем периоде времени $(\dot{t},\dot{t}+\Delta t)$ не произошли. В общем случае с увеличением интервала времени $\Delta t$ вероятность перехода автомата в новое состояние увеличивается, а вероятность сохранить текущее состояние уменьшается. Из диаграммы видно, что переходы в очередное состояние под действием потока событий поступления заявок на обслуживание - $w$, ничего не порождают на выходе автомата. Переходы в новые состояния под действием потока событий завершения обработки заявки каналом – $u$, порождает на выходе обслуженную заявку. Обратим внимание на состояние $z_{L+1}$. В этом состоянии прибор *П* отказывает очередной возникшей заявке в приёме на обслуживание, а фиктивный переход оставляет прибор в том же состоянии, а на выходе формируется отказ $d$. То есть, некоторые заявки получают отказ в обслуживании.

Так как моделью прибора *П* является непрерывный событийно-стохастический автомат, то его основной характеристикой являются вероятности $\mathrm{P}_z(t)$ нахождения автомата в том или ином состоянии $z \in Z$ в момент времени $t$. Зная эти вероятности, можно, в зависимости от целей моделирования, оценивать различные производные от этих вероятностей характеристики эффективности функционирования прибора *П*, такие, например, как:

- вероятность того, что заявка не получит отказа в обслуживании;
- средняя доля (процент) заявок, получающих отказ в обслуживании;
- вероятность того, что поступившая в прибор *П* заявка будет немедленно (без ожидания в накопителе) принята на обработку каналом *К*;
- среднее количество заявок, *покидающих* прибор *П* обслуженными в единицу времени;
- среднее количество заявок, *обслуживаемых* прибором *П* в единицу времени (ожидающих в очереди и находящиеся в обработке каналом);
- среднее время обслуживания заявки прибором *П*;
- среднее время нахождения заявки в очереди - ожидания заявкой начала обработки;
- среднее количество заявок, ожидающих обработки;
- и т.п.

**Оценка эффективности процесса обслуживания**

В качестве фундаментальных количественных оценок эффективности функционирования прибора *П* используют две характеристики – *абсолютная* и *относительная пропускная способность* прибора обслуживания.

*Абсолютной пропускной способностью* – ***A***, называют *среднее количество заявок*, покидающих прибор *П* обслуженными в *единицу времени*. Эта характеристика позволяет оценить текущую продуктивность прибора обслуживания, а также оценивать, насколько эффективно в приборе *П* используется производительность канала обработки заявок $K$. Заметим, что если заявки на обслуживание вообще не поступают, то в текущий момент времени абсолютная пропускная способность прибора *П* будет равна нулю. По мере того, как интенсивность поступления заявок повышается, возрастает и абсолютная пропускная способность прибора обслуживания *П*, приближаясь к предельной интенсивности обработки заявок каналом $K$, но не может превысить её (производительность канала ограничена). Когда производительность канала полностью исчерпана, и накопитель полон, то поступающие в этот момент заявки получают отказ в обслуживании. При этом продуктивность прибора *П* является высокой.

Абсолютная пропускная способность позволяет оценить эффективность загрузки канала $K$, но не является достаточной характеристикой для оценки эффективности работы прибора *П* со стороны поступающих заявок. Чем выше интенсивность поступления заявок на обслуживание, тем хуже характеристики времени их обслуживания прибором *П* за счёт увеличения длительности ожидания заявками обслуживания в накопителе *Н* (время нахождения в очереди на обслуживание каналом *К*) или в значительном увеличении вероятности отказа в обслуживании. Поэтому для оценки эффективности функционирования прибора *П* со стороны заявок используют отношение абсолютной пропускной способности $A$, к среднему количеству заявок $\lambda$, поступающих на обслуживание в прибор *П* в единицу времени – *относительная пропускная способность*:

$$q = \frac{A}{\lambda}.$$

Относительная пропускная способность показывает долю, которую составляет среднее число заявок, обслуженных прибором *П* в единицу времени, от среднего количества заявок, поступающих на обслуживание в единицу времени. Коротко, это означает что текущая абсолютная пропускная способность прибора обслуживания может быть равна или быть меньше интенсивности поступления заявок. Если $q = 1$, то в общем случае это означает, что производительность канала обработки *К* полностью не задействована.

С увеличением интенсивности поступления заявок относительная пропускная способность может начать стремиться от 1 к 0. Если $q$=1, то интенсивность поступления заявок равна текущей интенсивности обработки заявок каналом $K$. При этом доля заявок, которые случайно получат отказ в обслуживании, будет существенно зависеть от ёмкости накопителя $H$. При $q$=1 существует минимальное значение ёмкости накопителя $L$, при котором доля заявок, случайно получающих отказ в обслуживании, практически сводится к нулю.

Если $q$<1, то это означает, что интенсивность поступления заявок на обслуживание превысила интенсивность обработки заявок каналом $K$. При $q \to 0$ вероятность отказа стремится к 1.

Таким образом, абсолютная и относительная пропускные способности являются основными взаимно дополняющими друг друга характеристиками оценки эффективности работы простейшего прибора обслуживания.

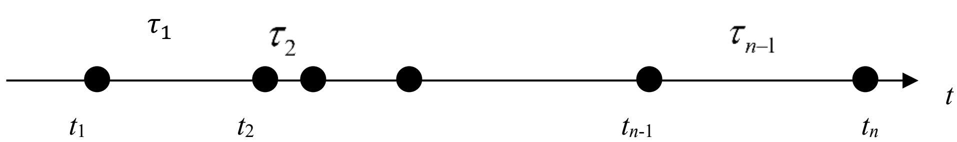
### *Моделирование и анализ процесса обслуживания*

При определённых допущениях относительно характеристик потока заявок и потока обслуживания их модель представляется так называемой *непрерывной марковской цепью*, при этом вероятности нахождения автомата в том или ином состоянии в текущий момент времени и могут быть выражены в форме системы связывающих их дифференциальных уравнений. Если эти вероятности удаётся получить в общем виде как решение полученной системы дифференциальных уравнений, то искомые характеристики СМО могут быть выражены в аналитическом виде через найденные вероятности состояний. Ниже рассмотрим теоретические основы математического моделирования СМО, позволяющие получать искомые характеристики в аналитическом виде, при условии, что потоки входных заявок и потоки обработанных заявок представляются в виде непрерывных марковских цепей.

## Модели потоков событий

Моменты поступления заявок на обслуживание или моменты завершения обслуживания заявок будем в общем случае рассматривать как случайно возникающие во времени *события*, образующие упорядоченные во времени потоки событий. Особенности появления событий во времени (условия формирования в потоке случайных интервалов между событиями) определяют специфические свойства потока, оказывающих существенное влияние на характеристики функционирования СМО. Более того, свойства потоков событий влияют и на возможность выражения вероятностей состояний СМО в аналитической форме. Ниже определим перечень свойств, характеризующих абстрактный поток событий, и на основании этих свойств введём классификацию потоков, а также определим характеристики свойств, которыми должен обладать поток событий, чтобы модель потока представлялась *непрерывной марковской*. Среди потоков, представляемых непрерывной марковской цепью, особую роль играют потоки событий, которые называют *простейшими*. Далее последовательно рассмотрим классификацию потоков событий и определим характеристики простейшего потока событий.

***Поток событий*** – это последовательность абстрактных событий, следующих одно за другим и возникающих в случайные моменты непрерывного времени - *вызывающие моменты*. Графически события можно обозначить точками на временно́й оси, помеченными вызывающими моментами времени их наступления - $t_i$:




Различают потоки однородных и неоднородных событий. Поток событий называется *однородным*, если события характеризуются только вызывающими моментами, и поток задаётся последовательностью $\{0 \le t_1 < t_2 \ldots < t_i < \ldots\}$ $t_i$ - неотрицательное вещественное число, являющееся вызывающим моментом $i$-ого события. Однородный поток событий может быть задан также в виде последовательности $\{\tau_1, \tau_2, \dots, \tau_i, \dots\}$, где $\tau_i$ - интервал времени между ($i$-1)-ым и $i$-ым вызывающим моментом. Поток событий называется *неоднородным*, если $i$-ое событие задаётся парой <$t_i$, $f_i$>, где $t_i$- вызывающий момент, $f_i$- отличительный признак события. В качестве признаков могут, например, быть заданы:

- идентификатор источника события,
- приоритет события,
- признак канала, предназначенного для обслуживания события и т.п.

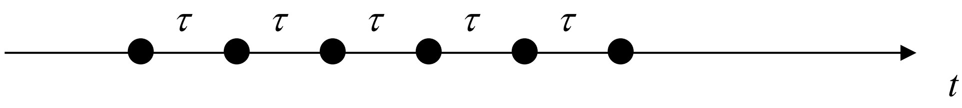
Если представить себе бесконечное упорядоченное множество параметров $\mathcal{T} = \{\mathrm{T}_1, \mathrm{T}_2, \dots, \mathrm{T}, \dots\}$, каждый из которых принимает случайные значения на мужестве положительных вещественных чисел, то поток однородных событий $\{\tau_1, \tau_2, \dots, \tau_i, \dots\}$, $(\tau_i > 0) \in R$ – множество положительных вещественных чисел, можно представить как случайную реализацию множества $\mathcal{T}$. Случайные параметры множества $\mathcal{T}$, в общем случае не являются независимыми и могут иметь разные законы распределения. В такой трактовке потоки однородных событий могут характеризоваться наличием различных корреляционных связей между упорядоченными случайными значениями $\tau_i, \tau_j \in \{\tau_1, \tau_2, \dots, \tau_i, \dots\}$. Иными словами, важно понимать, какое влияние оказывает на формирование очередного интервала $\tau_i$ предыстория потока событий.

По видам связей между упорядоченными случайными величинами $\tau_i$ однородные потоки событий делят на:

- регулярные потоки;
- потоки без последействия;
- потоки с ограниченным последействием;
- потоки Пальма;
- стационарные и нестационарные потоки;
- ординарные потоки;
- простейшие потоки;
- потоки Эрланга.

### Регулярные потоки

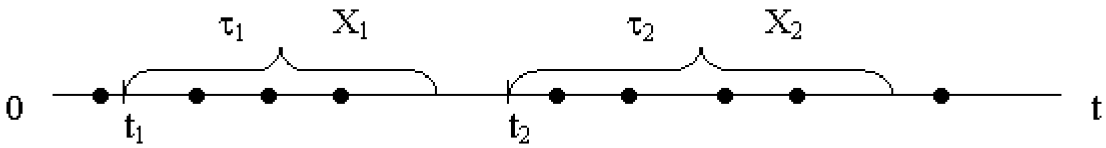
Поток событий называется *регулярным,* если события следуют одно за другим через детерминированные интервалы времени. Классическим примером регулярного потока является поток событий, возникающих через равные промежутки времени - $\tau_i=\tau_j=\tau$:




Очевидно, что события регулярного потока обладают однозначным последействием, так как вызывающий момент последующего события строго определён вызывающими моментами предыдущих событий. В вероятностной интерпретации промежутки времени между событиями являются достоверными значениями.

### Потоки без последействий

Поток событий называется потоком *без последействия*, если для любых непересекающихся периодов времени число вызывающих моментов $X_1$, попадающих на один из них, не зависит от того, сколько вызывающих моментов $X_2$ попадёт на другой период[1].




Образно говоря, непересекающиеся периоды времени, не связаны между собой как "сообщающиеся сосуды" событий. Если поток без последействия, то можно утверждать, что вызывающие моменты (события) наступают *независимо* друг от друга. Однако заметим, что если поток без последействия, то из этого не

[1] Условие отсутствия последействия означает, что заявки поступают в систему независимо друг от друга. Например, поток пассажиров, входящие на станцию метро, можно считать потоком без последействия потому, что причины, обусловившие приход отдельного пассажира именно в тот, а не другой момент, как правило, не связаны с аналогичными причинами для других пассажиров. Однако поток пассажиров, покидающих станцию метро, уже не может считаться потоком без последействия, так как моменты выхода пассажиров, прибывших одним и тем же поездом, зависимы между собой.

следует, что величины интервалов $\tau_i$ будут *независимыми* случайными значениями.

**Потоки с ограниченным последействием**

Теперь положим, что в потоке событий (вызывающих моментов) интервалы $\tau_i$ являются *независимыми* случайными величинами. Из этого в общем случае не следует, что этот поток событий будет потоком без последействия. Возможность «ограниченного» последействия потока событий при независимости $\tau_i$, как случайных величин, обуславливается тем, они в общем случае имеют разные законы распределения, а специфические свойства различных распределений вероятности каждой из независимых случайных величин $\tau_i$ могут оказывать влияние на формирование количества событий на одинаковых периодах времени на разных участках временной оси. Поэтому такие потоки событий с независимыми случайными величинами $\tau_i$ называются потоками с *ограниченным последействием*.

**Поток Пальма**

*Потоком Пальма* называется поток событий с *ограниченным последействием*, у которого случайные интервалы $\tau_i$ являются независимыми случайными величинами, а законы распределения вероятности интервалов $\tau_i$ являются *одинаковыми*.

**Стационарные и нестационарные потоки**

Поток событий называется *стационарным*, если вероятность появления того или иного количества вызывающих моментов (событий) на интервале времени $\Delta t$ зависит только от длины этого интервала и не зависит от того, на каком участке временно́й оси находится этот интервал. В противном случае поток будет нестационарным. Условию стационарности удовлетворяет поток событий, вероятностные характеристики которого не зависят от времени. В частности, стационарный поток характеризуется постоянным средним числом заявок, поступающих в единицу времени[2], называемым плотностью или интенсивностью потока.

**Ординарный поток**

Поток событий называется *ординарным*, если вероятность $p_{>1}(t,\Delta t)$ того, что на малый интервал времени $\Delta t$, непосредственно примыкающий к вызывающему моменту $t$, попадёт больше одного вызывающего момента пренебрежимо мала по сравнению с вероятностью того, что на этот же интервал времени $\Delta t$ попадает ровно одно событие - $p_1(t,\Delta t)$. То есть:

$$p_1(t,\Delta t) \gg p_{>1}(t,\Delta t) = 0(\Delta t),$$

где $0(\Delta t)$ - величина, порядок малости которой выше, чем $\Delta t$, т.е. $\lim\limits_{\Delta t\to 0} \frac{p_{>1}(t,\Delta t)}{\Delta t} = 0$. Качественно это означает, что ординарный поток событий характеризуется отсутствием по вероятности сгустков событий[3] на малом интервале времени $\Delta t$.

Так как для любого интервала $\Delta t$ любого потока событий будет выполняться условие:

$$p_0(t,\Delta t) + p_1(t,\Delta t) + p_{>1}(t,\Delta t)=1,$$

то для ординарного потока событий можно полагать, что:

$$p_0(t,\Delta t) + p_1(t,\Delta t) \approx 1.$$

С учётом этого для ординарного потока событий *среднее количество вызывающих моментов* событий, возникающих на интервале $\Delta t$, примыкающем к вызывающему моменту $t$, можно оценить как:

$$0\cdot p_0(t,\Delta t) + 1\cdot p_1(t,\Delta t) = p_1(t,\Delta t).$$

То есть, среднее количество вызывающих моментов событий на малом отрезке времени $\Delta t$ совпадает с вероятностью появления на отрезке времени $\Delta t$ одного

[2] В действительности потоки стационарны лишь на ограниченном участке времени, а распространение этого участка до бесконечности - лишь удобный приём, применяемый в целях упрощения анализа.

[3] Полагают, что заявки приходят поодиночке, а не парами, тройками и т. д. Поток клиентов, входящих в парикмахерскую, может считаться практически ординарным, чего нельзя сказать о потоке клиентов, направляющихся в ЗАГС для регистрации брака.

события. Тогда оценка *среднего количества* вызывающих моментов событий, возникающих на интервале $\Delta t$ в *единицу времени*, будет равна:

$$\frac{p_1(t,\Delta t)}{\Delta t}.$$

Рассмотрим предел этого выражения при $\Delta t \to 0$. Если этот предел существует, то обозначим его $\lambda(t)$:

$$\lim_{\Delta t \to 0} \frac{p_1(t,\Delta t)}{\Delta t} = \lambda(t).$$

Величина $\lambda(t)$ называется *интенсивностью* (или *плотностью*) ординарного потока событий в момент времени $t$. Качественно она характеризует скорость поступления событий в момент времени $t$. Чем выше интенсивность $\lambda(t)$, тем вероятнее возникновение события в момент времени $t$.

## Простейший поток

Для нестационарных потоков интенсивность потока зависит от времени $t$. Для стационарных потоков интенсивность постоянна, т.е. $\forall t\ \lambda(t)=\lambda=\text{const}$. Если *стационарный* поток является *ординарным* и *без последействий*, то такой поток называют *простейшим потоком*.

Простейший поток играет среди потоков событий особую роль, до некоторой степени аналогичную роли нормального закона среди других законов распределения. Известно, что при суммировании большого числа независимых случайных величин, подчинённых практически любым законам распределения, получается величина, приближённо распределённая по нормальному закону. Аналогично можно доказать, что при взаимном наложении (суммировании) большого числа ординарных, стационарных потоков с практически любым последействием получается поток, сколь угодно близкий к простейшему. Условия, которые должны для этого соблюдаться, аналогичны условиям центральной предельной теоремы, а именно - складываемые потоки должны оказывать на сумму приблизительно равномерно малое влияние.

Проиллюстрируем это положение элементарными рассуждениями[4]. Пусть имеется ряд независимых потоков $\Pi_1, \Pi_2, \ldots \Pi_n$. "Суммирование" потоков состоит в том, что все моменты появления событий сносятся на одну и ту же временну́ю ось $0t$ (Рисунок 2). Предположим, что потоки $\Pi_1, \Pi_2, \ldots \Pi_n$ сравнимы по своему влиянию на суммарный поток (т.е. имеют плотности одного порядка), а число их достаточно велико. Предположим, кроме того, что эти потоки стационарны и ординарны, но каждый из них может иметь последействие, и рассмотрим суммарный поток $\mathbf{\Pi}$ на оси $0t$ (Рисунок 2). Очевидно, что поток $\mathbf{\Pi}$ должен быть стационарным и ординарным, так как каждое слагаемое обладает этим свойством и они независимы. Кроме того, при увеличении числа слагаемых последействие в суммарном потоке, даже если оно значительно в отдельных потоках, должно постепенно слабеть. Действительно, рассмотрим на

[4] Вентцель Е.С. Теория вероятностей: Учеб. для вузов. — 6-е изд. стер. — М.: Высш. шк., 1999.— 576 с.

оси $0 \rightarrow t$ два не перекрывающихся отрезка $\tau_1$ и $\tau_2$ (Рисунок 2). Каждая из точек, попадающих в эти отрезки, случайным образом может оказаться принадлежащей тому или иному потоку, и по мере увеличения *n* удельный вес точек, принадлежащих одному и тому же потоку (и, значит, зависимых), должен уменьшаться, а остальные точки принадлежат разным потокам и появляются на отрезках $\tau_1$ и $\tau_2$ независимо друг от друга. Естественно ожидать, что при увеличении *n* суммарный поток будет терять последействие и приближаться к простейшему.

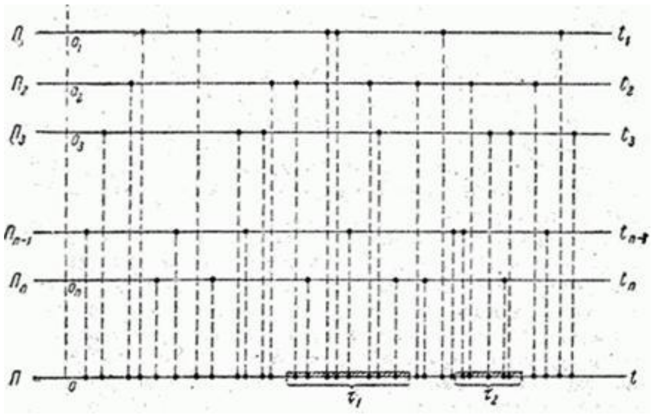


Рисунок 2

На практике оказывается, что обычно достаточно сложить 4-5 потоков, чтобы получить поток, с которым можно оперировать как с простейшим.

Теоретически доказано, что для *простейшего потока* случайный интервал времени $\tau$ между соседними вызывающими моментами событий подчиняется *экспоненциальному* закону распределения[5]:

$F(\tau) = 1 - e^{-\lambda\tau}$, $\tau>0$.

Соответственно плотность распределения вероятностей равна (Рисунок 3):

$f(\tau) = \lambda e^{-\lambda\tau}$.

5 Впервые это доказал в 1932 г. Александр Яковлевич Хи́нчин (1894 — 1959) — российский математик, один из наиболее значимых учёных в советской школе теории вероятностей.

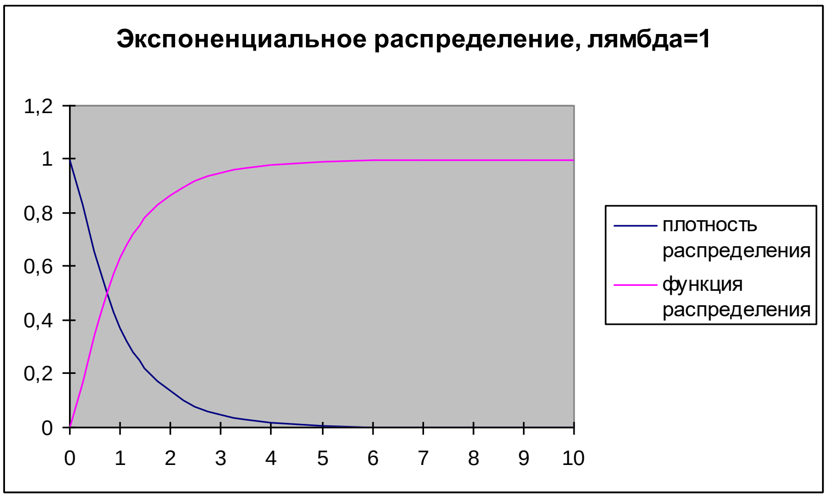



Рисунок 3

Математическое ожидание, дисперсия и среднеквадратическое отклонение интервала времени $\tau>0$ между соседними событиями простейшего потока имеют следующие значения:

$$\tau_{cp} = \int_0^{\infty} t\lambda e^{-\lambda t} dt = \frac{1}{\lambda},$$

$$D_{\tau} = \int_0^{\infty} t^2 \lambda e^{-\lambda t} dt - \frac{1}{\lambda^2} = \frac{1}{\lambda^2},$$

$$\sigma_{\tau} = \frac{1}{\lambda}.$$

Для простейшего потока вероятность того, что случайное количество вызывающих моментов - $\nu$, возникших в течение интервала времени $\Delta t$, окажется равным $m$, распределено в соответствии с законом Пуассона:

$$P\{\nu = m\} = \frac{a^m}{m!} e^{-a},$$

$a=a(\Delta t)=\lambda\Delta t$ – среднее количество вызывающих моментов, попадающих на отрезок времени $\Delta t$.

**Поток Эрланга**

*Поток Эрланга $k$-го порядка* получается из исходного простейшего потока с заданной интенсивностью $\lambda_0$ путём учёта только каждого $k$-ого вызывающего момента простейшего потока, а промежуточные вызывающие моменты отсеиваются. Исходный простейший поток является потоком Эрланга 1-го порядка. Иллюстрация получения потоков Эрланга 2-го и 3-го порядка представлена на рисунке 5.

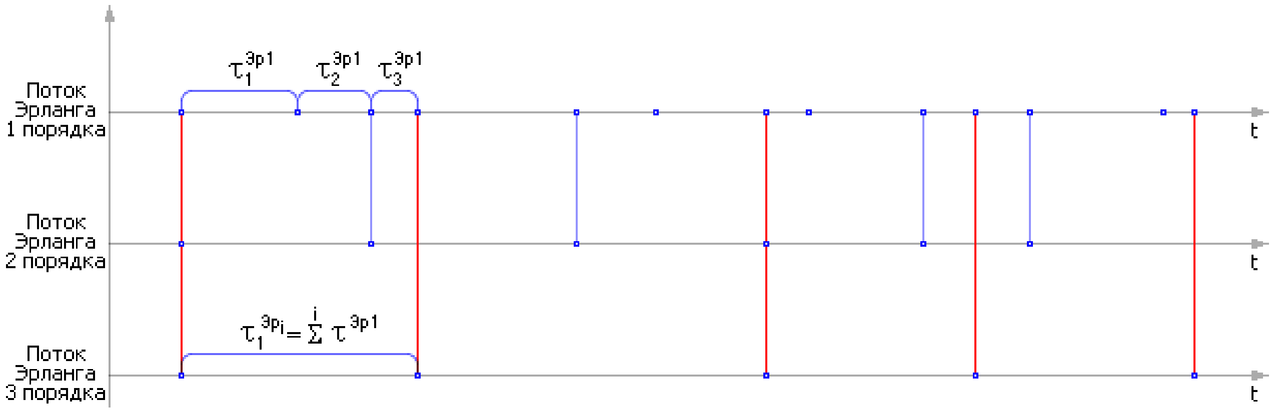



Рисунок 4

Для потока Эрланга интервалы времени между соседними вызывающими моментами событий равны по величине сумме $k$ независимых случайных интервалов времени между соседними вызывающими моментами исходного простейшего потока с интенсивностью $\lambda_0$, который подвергается прореживанию:

$$\tau_Э^{(k)} \equiv \tau = \tau_1 + \tau_2 + \cdots + \tau_k.$$

Показано, что плотность распределения вероятностей интервала времени между соседними вызывающими моментами событий потока Эрланга $k$-го порядка, полученного из прореживаемого простейшего потока с интенсивностью $\lambda_0$, равна:

$$f_Э^{(k)}(\tau) = \frac{(k\Lambda_k)^k}{(k-1)!}\tau^{k-1}e^{-k\Lambda_k\tau}, \tau > 0,$$

где $\Lambda_k = \frac{\lambda_0}{k}$ - интенсивность потока Эрланга $k$-го порядка.

В соответствии с теоремами сложения математических ожиданий и дисперсий формулы для вычисления среднего и дисперсии интервалов времени между событиями потока Эрланга $k$-го порядка будут иметь вид:

$$\tau_{ср}^{(k)} = \frac{k}{\lambda_0} = \frac{1}{\Lambda_k},$$

$$D_\tau^{(k)} = \frac{k}{\lambda_0^2} = \frac{1}{k\Lambda_k^2},$$

$$\sigma_\tau^{(k)} = \frac{\sqrt{k}}{\lambda_0} = \frac{1}{\sqrt{k}\Lambda_k}.$$

Из формул следует, что при $k$=1 получаются характеристики исходного простейшего потока. В то же время важно заметить, что если положить $\Lambda_k=\Lambda=const$, то при $k\to\infty$, сохраняя при этом значение $\Lambda$ неизменным (следовательно, увеличивая пропорционально увеличению $k$ значение интенсивности $\lambda_0$ исходного простейшего потока), дисперсия и среднеквадратическое отклонение потока Эрланга стремятся к нулю. Это означает, что *при* $k\to\infty$ *поток Эрланга с заданной интенсивностью* $\Lambda=const$ *неограниченно приближается к регулярному потоку с постоянным интервалом времени* $\tau = \dfrac{1}{\Lambda} = const$ *между вызывающими моментами событий*. Поэтому потоки Эрланга моделируют широкую гамму потоков Пальма - от простейшего потока (с отсутствием последействия), до практически регулярного потока.

### Непрерывные марковские цепи

Функционирование непрерывной событийно-стохастической системы в непрерывном времени выглядит как случайные переходы системы из одного дискретного состояния в другое дискретное состояние, обусловленные возникновением случайных событий через случайные промежутки времени. Так как с увеличением времени $\Delta t$, прошедшего с момента возникновения очередного случайного события, возрастает вероятность возникновения следующего случайного события, то вероятность сохранить текущее состояние уменьшается, а вероятность перейти в новое состояние увеличивается. В результате функционирование непрерывной событийно-стохастической системы во времени выглядит как *цепь последовательных событий* перехода системы из одного дискретного состояния в другое.

В теории непрерывных событийно-стохастических систем особо выделяют системы, функционирование которых выглядит так, что для каждого дискретного состояния системы, связанные с ним вероятности переходов в другие состояния не зависят от предыстории попадания в данное состояние, такой процесс функционирования системы называют *непрерывной марковской цепью*.

Характеристики системы, функционирование которой представляется непрерывной марковской цепью, полностью определяются вероятностями $\mathrm{P}_i(t)$ - вероятностями нахождения системы в момент времени $t$ в состоянии $z_i$ ($i$ =1,2,..,$K$). Поэтому нахождение этих вероятностей для анализа характеристик таких систем является основной задачей моделирования. Очевидно, что для любого текущего момента времени $t$ марковская цепь находится в одном из возможных состояний, следовательно:

$$\sum_{i=1}^{K} \mathrm{P}_i(t) = 1.$$

Для $t$=0 в качестве начальных условий задаётся вектор вероятностей начального состояния марковской цепи - $<\mathrm{P}_1(0), \mathrm{P}_2(0),\ldots, \mathrm{P}_k(0),\ldots, \mathrm{P}_K(0)>$, где $\mathrm{P}_k(0)$ – вероятность нахождения марковской цепи в начальный момент времени $t$=0 в состоянии $z_k$.

## Уравнения Колмогорова

Вероятности состояний системы, функционирующей в виде непрерывной марковской цепи, теоретически выводятся из её первоначального приближения в виде дискретного стохастического автомата, порождающего дискретную марковскую цепь, изменяющую своё состояние только по истечении достаточно малого тика $\Delta t$ реального времени. Искомые вероятности нахождения дискретного стохастического автомата в состоянии $z_i$ в произвольный момент времени $t$ и в течение тика $\Delta t$ обозначим в виде $\widetilde{\mathrm{P}}_i\,(t+\Delta t)$. Тогда вероятности состояний в произвольный момент времени $t$ для непрерывной марковской цепи определятся в результате предельного перехода - $\mathrm{P}_i\,(t) = \lim\limits_{\Delta t \to 0} \widetilde{\mathrm{P}}_i\,(t+\Delta t)$ к непрерывному времени.

Для нахождения вероятностей $\widetilde{\mathrm{P}}_i\ (t+\Delta t)$ положим, что в момент времени $t$ дискретная марковская цепь оказалась в состоянии $z_i$. Положим также, что на примыкающем к $t$ малом интервале $\Delta t$ состояние марковской цепи остаётся неизменным, а в момент $t+\Delta t$ система с вероятностью $p_{ij}(t,\Delta t)$ скачком меняет своё состояние на $z_j$ или с вероятностью $\bar{p}_{ii}(t,\Delta t)=1-p_{ij}(t,\Delta t)$ остаётся в прежнем состоянии $z_i$. Тогда матрица $P$ вероятностей перехода дискретной марковской цепи примет вид:

$$P=\left[p_{ij}(t,\Delta t)\right]_{i,j=\overline{1,K}}.$$

Введём в рассмотрение орграф, определяющий поведение дискретной марковской цепи в виде дискретного стохастического конечного автомата, функционирующего в реальном времени с тиком $\Delta t$, с заданной матрицей вероятностей перехода Вершины орграфа будут совпадать с состояниями аппроксимируемой непрерывной марковской цепи, а дуги, помеченные вероятностями перехода $p_{ij}(t,\Delta t)$ из состояния $z_i$, возникшего в момент $t$, в состояние $z_j$ спустя малый тик времени $\Delta t>0$. Вершины орграфа пометим вероятностями нахождения в них дискретной марковской цепи в начальный момент времени $\mathrm{P}_i\ (0)$ - вероятности начальных состояний. Пример орграфа дискретной марковской цепи, аппроксимирующей непрерывную марковскую цепь, с тремя возможными состояниями $\{z_1,\ z_2,\ z_3\}$, вероятностями начальных состояний $\{\mathrm{P}_1(0),\ \mathrm{P}_2(0),\ \mathrm{P}_3(0)\}$ и вероятностями переходов $\{p_{12}(t,\Delta t),\ p_{31}(t,\Delta t),\ p_{32}(t,\Delta t),\ p_{23}(t,\Delta t)\}$ показан ниже (рисунок 5):

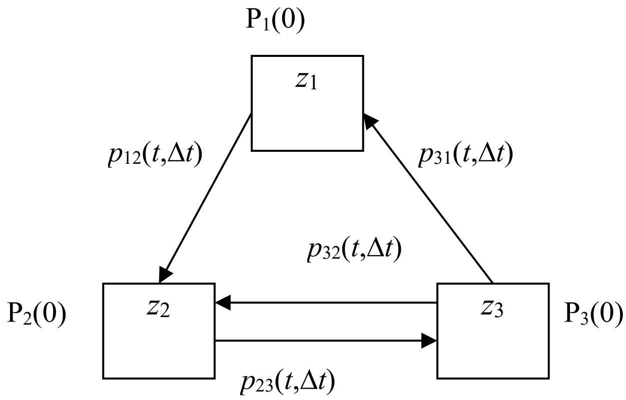



Рисунок 5

Положим, что порождаемая дискретная марковская цепь представляет собой *ординарный* поток вызывающих моментов перехода из состояния $z_i$ в состояние $z_j$, и в момент времени $t$ известна интенсивность ординарного потока $\lambda_{ij}(t) = \lim\limits_{\Delta t \to 0} \frac{p_{ij}(t,\Delta t)}{\Delta t}$ - среднее число событий перехода в единицу времени, возникающих в момент времени $t$. Пусть в момент времени $t$ дискретная марковская цепь оказалась в состоянии $z_i$, тогда для ординарного потока вероятность $p_{ij}(t, \Delta t)$ перехода автомата в состоянии $z_j$, спустя тик времени $\Delta t$, можно выразить через заданную интенсивность ординарного потока:

$$p_{ij}(t, \Delta t) = \lambda_{ij}(t)\Delta t.$$

Положим ещё, что ординарный поток событий перехода дискретной марковской цепи является однородным и стационарным, т.е. вероятности перехода не зависят от момента $t$: $p_{ij}(t, \Delta t) \equiv p_{ij}(\Delta t)$, и интенсивность $\lambda_{ij}(t) \equiv \lambda_{ij} = const$. Следовательно, $p_{ij}(\Delta t) = \lambda_{ij}\Delta t$. Тогда ординарный поток будет являться *простейшим потоком*. Для простейшего потока доказано, что вероятность перехода из текущего состояния $z_i(t_i)$ в новое состояние $z_j(t_i + \Delta t)$ за интервал времени $\Delta t$ подчиняется экспоненциальному закону $F_t(\Delta t) = 1 - e^{-\mu_{ij}\Delta t}$. То есть, вероятность перехода дискретной марковской цепи из состояния $z_i$ в состояние $z_j$ за время $\Delta t$ для экспоненциального закона будет равна:

$$p_{ij}(\Delta t) = \lambda_{ij}\Delta t = 1 - e^{-\mu_{ij}\Delta t}.$$

Тогда:

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{1 - e^{-\mu_{ij}\Delta t}}{\Delta t} = \left| \frac{0}{0} \right|.$$

Для разрешения неопределённости применим правило Лопиталя, взяв производные по $\Delta t$ от числителя и знаменателя, получим:

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{\mu_{ij} e^{-\mu_{ij}\Delta t}}{1} = \mu_{ij}.$$

Следовательно, параметр $\mu_{ij}$ соответствующего экспоненциального закона, определяющего вероятность перехода, будет просто совпадать с заданной для дискретной марковской цепи интенсивностью $\lambda_{ij}$ перехода из состояния $z_i$ в состояние $z_j$ за время $\Delta t$. А это значит, что зная заданную для дискретного марковского процесса интенсивность $\lambda_{ij}$, его вероятности перехода за время $\Delta t$ будут равны: $p_{ij}(\Delta t) = 1 - e^{-\lambda_{ij}\Delta t} = \lambda_{ij}\Delta t$. И теперь орграф дискретной марковской цепи можно для простейших потоков представить в следующем виде (рисунок 6):

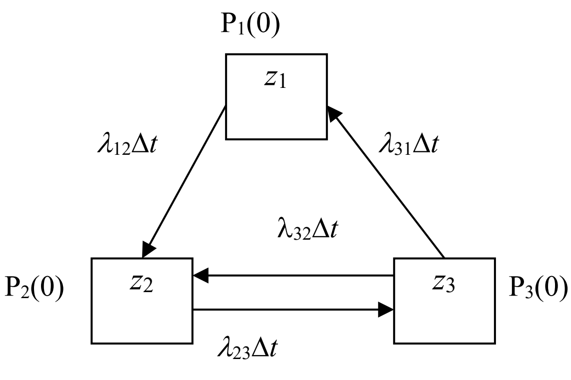



Рисунок 6

Используя полученный орграф, построим выражение для определения вероятности $P_1(t)$ - того, что однородная дискретная марковская цепь в некоторый момент времени $t$ окажется в состоянии $z_1$. Это может случиться в результате реализации в момент $t$ двух возможных предысторий:

- в момент времени $t$ с вероятностью $P_1(t)$ марковская цепь была в состоянии $z_1$ и спустя $\Delta t$ с вероятностью $\lambda_{11}\Delta t = 1 - p_{12}(\Delta t)$ *не вышла* из него,
- в момент времени $t$ система с вероятностью $P_3(t)$ находилась в состоянии $z_3$, а спустя $\Delta t$ с вероятностью $p_{31}(\Delta t) = \lambda_{31}\Delta t$ *перешла* в состояние $z_1$.

Тогда в первом случае вероятность будет равна:

$$P_1(t)(1 - p_{12}(\Delta t)) = P_1(t)(1 - \lambda_{12}\Delta t).$$

Вероятность второго случая равна:

$$P_3(t)p_{31}(\Delta t) = P_3(t)\lambda_{31}\Delta t.$$

Применяя правило сложения вероятностей, получим:

$$P_1(t + \Delta t) = P_1(t)(1 - \lambda_{12}\Delta t) + P_3(t)\lambda_{31}\Delta t,$$

или

$$\frac{P_1(t+\Delta t) - P_1(t)}{\Delta t} \approx -\lambda_{12}P_1(t) + \lambda_{31}P_3(t).$$

Чтобы для соответствующей непрерывной марковской цепи получить в момент времени $t$ аналогичное соотношение, произведём предельный переход к $t$ при $\Delta t \rightarrow 0$:

$$\lim_{\Delta t \to 0} \frac{P_1(t + \Delta t) - P_1(t)}{\Delta t} = -\lambda_{12}P_1(t) + \lambda_{31}P_3(t)$$

В результате для непрерывной марковской цепи, получим:

$$\frac{dP_1(t)}{dt} = -\lambda_{12}P_1(t) + \lambda_{31}P_3(t).$$

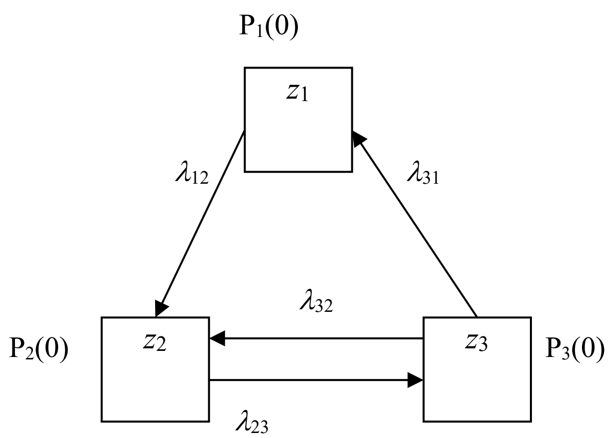
В итоге получили дифференциальное уравнение для определения вероятности нахождения уже непрерывной марковской цепи в заданный момент времени $t$ в состоянии $z_1$.

Действуя аналогично, построим дифференциальные уравнения для определения вероятностей всех других состояний однородной непрерывной марковской цепи. В итоге будет получена так называемая *система дифференциальных уравнений Колмогорова*[6].

$$\begin{cases} \dfrac{d\mathrm{P}_1(t)}{dt} = -\lambda_{12}\mathrm{P}_1(t) + \lambda_{31}\mathrm{P}_3(t), \\ \dfrac{d\mathrm{P}_2(t)}{dt} = -\lambda_{23}\mathrm{P}_2(t) + \lambda_{12}\mathrm{P}_1(t) + \lambda_{32}\mathrm{P}_3(t) \\ \dfrac{d\mathrm{P}_3(t)}{dt} = -\lambda_{31}\mathrm{P}_3(t) - \lambda_{32}\mathrm{P}_3(t) + \lambda_{23}\mathrm{P}_2(t) \end{cases}$$

Интегрирование этой системы уравнений при условии $\sum_{i=1}^{n} \mathrm{P}_i(t) = 1$ и заданных начальных условий для $t$=0: $\mathrm{P}_1(0)$, $\mathrm{P}_2(0)$,… $\mathrm{P}_k(0)$,…, $\mathrm{P}_K(0)$ позволяет получить искомые вероятности состояний непрерывной марковской цепи как функции времени.

С учётом полученных выше результатов соответствующий орграф однородной *непрерывной марковской цепи* с простейшими потоками перехода из состояния $z_i$ в состояние $z_j$ с интенсивностями $\lambda_{ij}$ будет представляться в виде (Рисунок 7).




[6]**Андрей Николаевич Колмогоров** (1903 — 1987) — советский математик, один из крупнейших математиков XX века.

Рисунок 7

На основе заданного для однородной непрерывной марковской цепи орграфа система дифференциальных уравнений Колмогорова формируется по вполне очевидному правилу, которое предписывает действовать следующим образом. При построении дифференциального уравнения относительно вероятности рассматриваемого состояния $z_i$ необходимо:

- В левую часть уравнения поставить производную вероятности состояния $z_i$ по времени - $\frac{dP_i(t)}{dt}$.
- Правая часть представляется суммой, содержащей столько членов, сколько дуг переходов орграфа связано с рассматриваемым состоянием $z_i$.
- Если дуга перехода направлена *из* рассматриваемого состояния $z_i$, соответствующий член суммы имеет знак минус - <->
- Если дуга перехода направлена *в* рассматриваемое состояние $z_i$, соответствующий член суммы имеет знак плюс - <+>.
- Каждый член суммы равен произведению интенсивности $\lambda_{ij}$ (помечающей дугу перехода) умноженной на вероятность того состояния, из которого эта дуга исходит: $\lambda_{ij} \cdot P_i(t)$.

Это правило составления системы дифференциальных уравнений для вероятностей состояний является общим и справедливо для любой однородной, стационарной и ординарной непрерывной марковской цепи. С его помощью можно сугубо алгоритмически выводить дифференциальные уравнения для вероятностей состояний по заданному размеченному орграфу состояний.

## Финальные вероятности

Если при $t \to \infty$ непрерывная марковская цепь переходит в стационарный режим, когда вероятности состояний перестают зависеть от времени, то такие вероятности называют *финальными вероятностями непрерывной марковской* *цепи*. На практике обычно именно нахождение финальных вероятностей является основной целью моделирования СМО.

## *Простейшие модели обслуживания*

Понятие "простейшие" не означает "примитивные". Основное допущение для простейшей СМО, заключается в формализации её поведения в виде непрерывной, стационарной и однородной марковской цепи с заданными интенсивностями перехода из одного состояния в другое по экспоненциальному закону, что позволяет построить систему дифференциальных уравнений Колмогорова для нахождения вероятностей состояний.

В дальнейшем при рассмотрении моделей простейших СМО интенсивности потоков входных заявок будем обозначать символом $\lambda$, а интенсивности потоков обслуживания заявок каналами - символом $\mu$.

### Одноканальное обслуживание с отказами

Пусть СМО состоит из одного прибора обслуживания с одним каналом обработки и без накопителя заявок. В прибор поступает простейший поток с интенсивностью $\lambda$ (Рисунок 8).

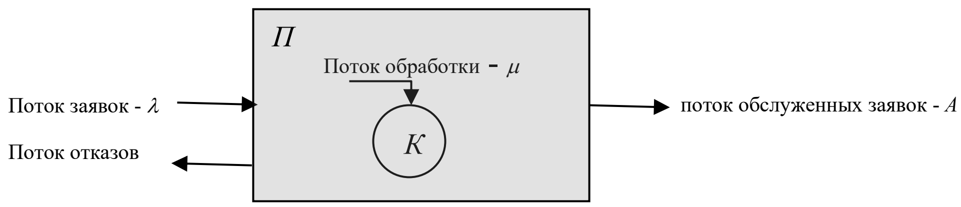



Рисунок 8

Обслуживание заявки осуществляется с интенсивностью $\mu$ - среднее количество заявок, обрабатываемых каналом $K$ в единицу времени (среднее время обслуживания $\bar{\tau}_{обсл} = 1/\mu$). В соответствии с описанием функционирования СМО определим два состояния марковской цепи:

$z_0$ - канал обработки $K$ свободен,

$z_1$ - канал обработки $K$ занят.

Определим вероятности состояний системы. Ниже показан орграф непрерывной марковской цепи, описывающий функционирование системы (Рисунок 9).

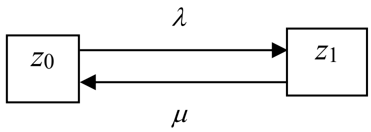



Рисунок 9

Обозначим вероятности состояний $z_0$ и $z_1$ в момент времени $t$ как $\mathrm{P}_0(t)$ и $\mathrm{P}_1(t)$. Естественно полагать, что в начальный момент времени $\mathrm{P}_0(0)=1$, $\mathrm{P}_1(0)=0$. Для полученного орграфа непрерывной марковской цепи составим систему дифференциальных уравнений Колмогорова в соответствии с правилами их построения:

$$\begin{cases} \dfrac{d\mathrm{P}_0(t)}{dt} = -\lambda \mathrm{P}_0(t) + \mu \mathrm{P}_1(t) \\ \dfrac{d\mathrm{P}_1(t)}{dt} = \lambda \mathrm{P}_0(t) - \mu \mathrm{P}_1(t) \end{cases}$$

с начальными вероятностями состояний: $\mathrm{P}_0(0)=1$, $\mathrm{P}_1(0)=0$. Добавим условие нормировки: $\mathrm{P}_0(t)+\mathrm{P}_1(t)=1$.

Известно, что результат решения полученной системы дифференциальных уравнений имеет вид:

$$\mathrm{P}_0(t) = \frac{\mu}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} e^{-(\lambda+\mu)t},$$

$$\mathrm{P}_1(t) = 1 - \mathrm{P}_0(t).$$

Теперь в аналитической форме можно выразить различные характеристики функционирования одноканальной СМО с отказами.

*Относительная пропускная способность* $q$.

Если в единицу времени поступает в среднем $\lambda$ заявок, а из них в среднем только $A(t)$ обслуживается в единицу времени, то вероятность $\mathrm{P}_0(t)$ для заявки быть обслуженной (т.е. застать прибор обслуживания свободным - в состоянии $z_0$)

равна отношению $A(t)$ к $\lambda$, т.е. совпадает с относительной пропускной способностью. Иными словами, относительная пропускная способность равна вероятности застать прибор обслуживания не занятым:

$$q(t) = \frac{A(t)}{\lambda} = \mathrm{P}_0(t).$$

Из этого следует, что абсолютная пропускная способность равна:

$$A(t) = \lambda\,\mathrm{P}_0(t).$$

Так как в формулах вероятностей состояний $\frac{\lambda}{\lambda+\mu}e^{-(\lambda+\mu)t}$ при $t\to\infty$ стремится к 0, то существуют финальные вероятности состояний:

$$\mathrm{P}_0 = \frac{\mu}{\lambda+\mu},$$

$$\mathrm{P}_1 = 1 - \mathrm{P}_0 = \frac{\lambda}{\lambda+\mu}.$$

Следовательно для финальных вероятностей состояний относительная и абсолютная пропускные способности примут вид:

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

Важной характеристикой является вероятность отказа (канал занят) - доля числа заявок, получивших отказ в обслуживании от числа поступивших заявок. Очевидно, что $\mathrm{P}_{\text{отк}} = \mathrm{P}_1$ – вероятность застать прибор занятым. Следовательно в стационарном режиме,

$$\mathrm{P}_{\text{отк}} = \frac{\lambda}{\lambda+\mu}.$$

**Многоканальное обслуживание с отказами**

Если количество каналов в приборе обслуживания больше одного, то СМО называют многоканальной. В начале рассмотрим двухканальную СМО без накопителя (Рисунок 10).

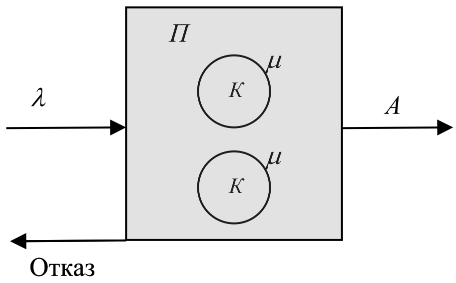



Рисунок 10

Состояния системы определим следующим образом:

- $z_0$ - все каналы свободны,
- $z_1$ - один канал занят,
- $z_2$ - заняты оба канала.

Тогда орграф состояний марковского процесса примет вид:

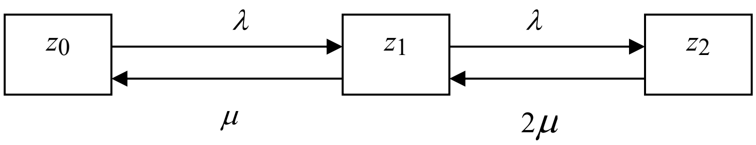



Рисунок 11

Составим систему дифференциальных уравнений для полученного орграфа состояний (Рисунок 11):

$$\begin{cases} \dfrac{d\mathrm{P}_0(t)}{dt} = -\lambda \mathrm{P}_0(t) + \mu \mathrm{P}_1(t), \\ \dfrac{d\mathrm{P}_1(t)}{dt} = \lambda \mathrm{P}_0(t) - (\lambda + \mu)\mathrm{P}_1(t) + 2\mu \mathrm{P}_2(t), \\ \dfrac{d\mathrm{P}_2(t)}{dt} = \lambda \mathrm{P}_1(t) - 2\mu \mathrm{P}_2(t). \end{cases}$$

Естественными начальными вероятностями состояний будут:

$$\mathrm{P}_0(0) = 1,\ \mathrm{P}_1(0) = 0, \mathrm{P}_2(0) = 0\,.$$

Добавим условие нормировки:

$$\mathrm{P}_0(t) + \mathrm{P}_1(t) + \mathrm{P}_2(t) = 1\,.$$

Решение полученной системы дифференциальных уравнений и получение вероятностей состояний сразу будем искать как финальные вероятности. Если финальные вероятности $P_0$ и $P_1$ существуют, то при $t\to\infty$ система дифференциальных уравнений превращается в систему линейных уравнений относительно финальных вероятностей $P_0$ и $P_1$:

$$\begin{cases} 0 = -\lambda P_0 + \mu P_1, \\ 0 = \lambda P_0 - (\lambda + \mu) P_1 + 2\mu P_2, \\ 0 = \lambda P_1 - 2\mu P_2. \end{cases}$$

или

$$\begin{cases} \lambda P_0 = \mu P_1, \\ (\lambda + \mu) P_1 = \lambda P_0 + 2\mu P_2, \\ \lambda P_1 = 2\mu P_2. \end{cases}$$

Введём обозначение:$\rho = \frac{\lambda}{\mu}$. Тогда система уравнений преобразуется к виду:

$$\begin{cases} \rho P_0 = P_1, \\ (\rho + 1) P_1 = \rho P_0 + 2P_2, \\ \rho P_1 = 2P_2. \end{cases}$$

Полученная система уравнений, не имеет свободного члена. Следовательно, система уравнений не замкнута и имеет линейную зависимость . Поэтому одно из уравнений (самое "неудобное") поменяем на условие нормировки $P_0+P_1+P_2=1$. Тогда получим:

$$\begin{cases} \rho P_0 = P_1, \\ P_0 + P_1 + P_2 = 1, \\ \rho P_1 = 2P_2. \end{cases}$$

Решение системы уравнений имеет вид:

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

$$P_1 = \rho P_0,$$

$$P_2 = \frac{\rho}{2} P_1 = \frac{\rho^2}{2} P_0 = \frac{\rho^2}{2(1+\rho+\frac{\rho^2}{2})}.$$

Зная вероятности состояний, можно найти следующие характеристики двухканальной СМО:

– вероятность отказа в обслуживании:

  $P_{отк} = P_2;$

– относительная пропускная способность (вероятность быть обслуженной):

  $q = (1\text{-}P_2);$

– абсолютная пропускная способность:

  $A=\lambda(1\text{-}P_2);$

– среднее число занятых каналов:

  $k_{ср} = 0 \cdot P_0 + 1 \cdot P_1 + 2 \cdot P_2 = \rho(1 - P_2).$

Ниже (рисунок 12) показан график изменения среднего числа занятых каналов от $\rho$.

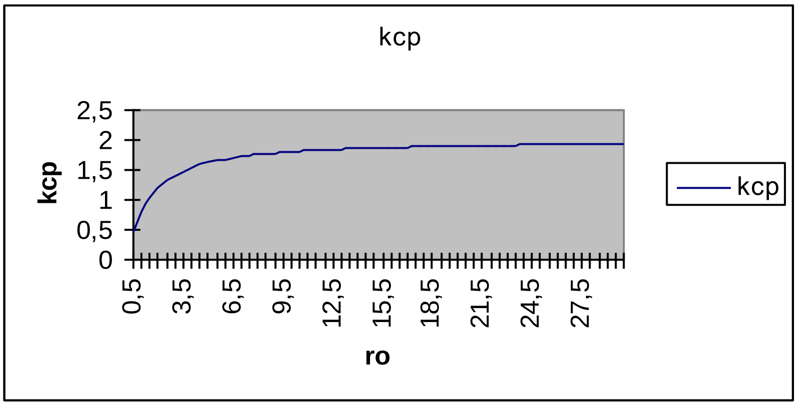



Рисунок 12

В общем случае для *n* каналов получены следующие формулы для финальных вероятностей системы (формулы Эрланга):

$$P_k = \frac{\lambda^k}{\mu \cdot 2\mu \ldots k\mu} P_0 = \frac{\rho^k}{k!} P_0, (k = 1,2, \ldots, n);$$

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

### Одноканальное обслуживание с ожиданием

Рассмотрим СМО с одним каналом обслуживания и накопителем для приёма заявок и ожидания обслуживания, если канал занят (

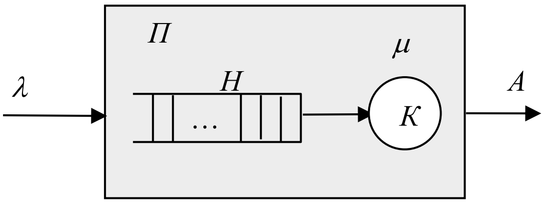



Рисунок 13). В СМО поступает простейший поток заявок с интенсивностью $\lambda$, а интенсивность обработки заявки в канале $K$ равна $\mu$.

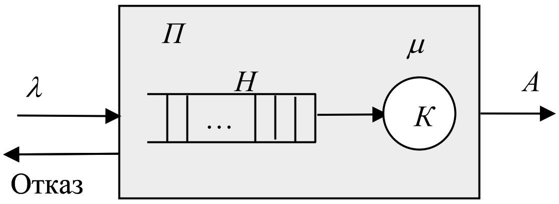



Рисунок 13

Заявка, поступившая в момент, когда канал занят, а накопитель не заполнен, становится в очередь и ожидает обслуживания. Если в накопителе мест нет, то заявка получает отказ и покидает СМО.

Определим следующие состояния для СМО с *двух*местным накопителем:

$z_0$ - канал свободен,

$z_1$ - канал занят, накопитель пуст,

$z_2$ - канал занят, в очереди 1-а заявка,

$z_3$ - канал занят, в очереди 2-е заявки.

Ниже представлен орграф состояний непрерывной марковской цепи (Рисунок 14).

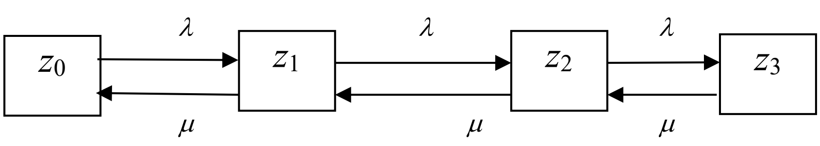



Рисунок 14

В соответствии с правилами построения составим систему дифференциальных уравнений Колмогорова:

$$\begin{cases} \dfrac{d\mathrm{P}_0(t)}{dt} = -\lambda \mathrm{P}_0(t) + \mu P_1(t), \\ \dfrac{d\mathrm{P}_1(t)}{dt} = \lambda \mathrm{P}_0(t) - (\lambda + \mu)\mathrm{P}_1(t) + \mu \mathrm{P}_2(t), \\ \dfrac{d\mathrm{P}_2(t)}{dt} = \lambda \mathrm{P}_1(t) - (\lambda + \mu)\mathrm{P}_2(t) + \mu \mathrm{P}_3(t), \\ \dfrac{d\mathrm{P}_3(t)}{dt} = \lambda \mathrm{P}_2(t) - \mu \mathrm{P}_3(t) \end{cases}$$

Начальные вероятности состояний очевидно равны:

$$\mathrm{P}_0(0) = 1,\ \mathrm{P}_1(0) = 0, \mathrm{P}_2(0) = 0, \mathrm{P}_3(0) = 0.$$

Кроме того, должно выполняться условие нормировки:

$$\mathrm{P}_0(t) + \mathrm{P}_1(t) + \mathrm{P}_2(t) + \mathrm{P}_3(t) = 1.$$

Для финальных вероятностей система уравнений преобразуется к виду:

$$\begin{cases} \lambda \mathrm{P}_0 = \mu \mathrm{P}_1, \\ (\lambda + \mu)\mathrm{P}_1 = \lambda \mathrm{P}_0 + \mu \mathrm{P}_2, \\ (\lambda + \mu)\mathrm{P}_2 = \lambda \mathrm{P}_1 + \mu \mathrm{P}_3, \\ \mu P_3 = \lambda \mathrm{P}_2. \end{cases}$$

Одно из уравнений (любое) заменим условием нормировки, после чего получим замкнутую систему уравнений.

Общее решение системы уравнений для $m$ мест в накопителе и $\rho = \frac{\lambda}{\mu}$ имеет вид:

$$\begin{cases} \mathrm{P}_k = \rho^k \mathrm{P}_0,\ (k = 1,2,\dots,m+1); \\ \mathrm{P}_0 = \frac{1}{1+\rho+\ldots+\rho^{m+1}}. \end{cases}$$

Заметим, что для $\mathrm{P}_0$ выражение в знаменателе является геометрической прогрессией, поэтому последнее уравнение можно заменить на:

$$\mathrm{P}_0 = \frac{1-\rho}{1-\rho^{m+2}}, \rho \neq 1.$$

В итоге получается:

$$\begin{cases} P_k = \rho^k P_0, \quad (k = 1,2,\ldots,m+1); \\ P_0 = \frac{1-\rho}{1-\rho^{m+2}},\ \rho \neq 1; \end{cases}.$$

Зная финальные вероятности состояний, получены аналитические выражения некоторых характеристик одноканальной СМО с очередью:

– Вероятность отказа в обслуживании:

$$P_{отк} = P_3 = \frac{\rho^3(1-\rho)}{1-\rho^4}.$$

– Относительная пропускная способность – эквивалентна вероятности, получить обслуживание:

$$q = 1 - P_{отк} = 1 - \frac{\rho^3(1-\rho)}{1-\rho^4}.$$

– Абсолютная пропускная способность:

$$A = \lambda q.$$

– Среднее число заявок в очереди:

$$\bar{r} = 1 \cdot P_2 + 2 \cdot P_3 + \ldots + mP_{m+1} = \frac{\rho^2[1-\rho^m(m+1-m\rho)]}{(1-\rho^{m+2})(1-\rho)}.$$

При $m$=2:

$$\bar{r} = 1 \cdot P_2 + 2 \cdot P_3 = \frac{\rho^2[1-\rho^2(3-2\rho)]}{(1-\rho^4)(1-\rho)}.$$

Так как среднее время обработки заявки равно $1/\mu$, то в общем случае:

– среднее время ожидания в очереди при $\lambda<\mu$ оценивается по формуле:

$$\bar{\tau}_{ож} = \frac{1}{\mu}P_1 + \frac{2}{\mu}P_2 + \ldots + \frac{m}{\mu}P_m = \frac{1}{\rho\mu}\bar{r} = \frac{\bar{r}}{\lambda}.$$

– среднее число заявок, находящихся в системе:

$$\bar{k} = 1\,P_1 + 2\,P_2 + \ldots + (m+1)\,P_{m+1} = \bar{r} + \frac{\rho-\rho^{m+2}}{1-\rho^{m+2}}.$$

## Замкнутые процессы обслуживания

До сих пор мы рассматривали СМО, в которых заявки приходили "извне". Интенсивность входного потока заявок не зависела от состояния системы. Рассмотрим СМО, в которых поток заявок на обслуживание зависит от текущего состояния прибора обслуживания. Такие СМО называют *замкнутыми*.

В качестве примера рассмотрим локальную сеть из двух компьютеров $C1$ и $C2$ (в общем случае $n$ компьютеров). Каждый компьютер с интенсивностью $\lambda$ может перестать работать и потребовать обслуживания системным администратором. Если администратор в этот момент свободен, то он принимает компьютер на обслуживание и обслуживает его в среднем а течение интервала времени $\tau_{обсл} = \frac{1}{\mu}$ ($\mu$- интенсивность обслуживания). Если компьютеру потребовалось обслуживание, но системный администратор занят, то компьютер ожидает обслуживания (Рисунок 15).

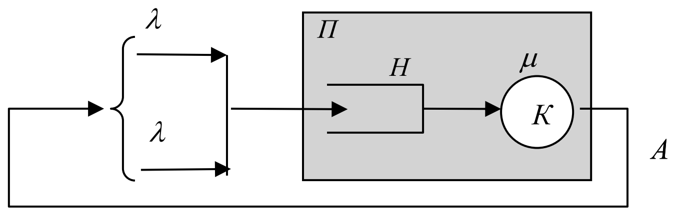



Рисунок 15

В качестве характеристик системы могут рассматриваться:

- вероятность того, что все компьютеры в работе,
- вероятность, что оба компьютера ждут обслуживания,
- среднее количество компьютеров, обслуживаемых администратором в единицу времени.
- среднее число компьютеров, требующих обслуживания

Для замкнутых СМО характерным является изменение интенсивности потока заявок. Действительно, в нашем случае интенсивность потока заявок может принимать значения: 0, $\lambda$, $2\lambda$. Однако в стационарном режиме их поведение может представлять собой непрерывную марковской цепь, что позволяет их исследовать как простейшую СМО.

Введём в рассмотрение следующее множество состояний:

– $z_0$ - все компьютеры исправны, администратор не занят;
– $z_1$ - один компьютер неисправен, администратор занят обслуживанием;
– $z_2$ - оба компьютера неисправны, один обслуживается, другой ожидает в накопителе.

Построим орграф непрерывного марковского процесса (Рисунок 16):

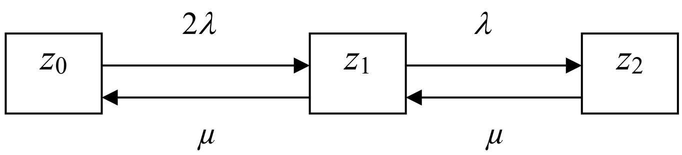



Рисунок 16

Система дифференциальных уравнений для данной марковской цепи примет вид:

$$\begin{cases} \dfrac{d\mathrm{P}_0}{dt} = -2\lambda \mathrm{P}_0(t) + \mu \mathrm{P}_1(t), \\ \dfrac{d\mathrm{P}_1}{dt} = 2\lambda \mathrm{P}_0(t) - (\lambda + \mu)\mathrm{P}_1(t) + \mu \mathrm{P}_2(t), \\ \dfrac{d\mathrm{P}_2}{dt} = \lambda P_1(t) - \mu \mathrm{P}_2(t). \end{cases}$$

В общем случае для $n$ компьютеров были получены следующие аналитические выражения для финальных вероятностей состояний, где $\rho = \frac{\lambda}{\mu}$:

$$\begin{cases} \mathrm{P}_0 = \dfrac{1}{1 + n\rho + n(n-1)\rho^2 + \ldots + \rho^n}, \\ \mathrm{P}_1 = n\rho P_0, \\ \ldots \\ \mathrm{P}_\mathrm{n} = n(n-1)\ldots 1 \cdot \rho^n \mathrm{P}_0. \end{cases}$$

На основании полученных вероятностей состояний можно найти интересующие характеристики функционирования системы.

Характеристикой занятости в процентах системного администратора обслуживанием компьютеров является величина:

$L = (1 - \mathrm{P}_0) \cdot 100\%.$

Среднее количество компьютеров, обслуживаемых администратором в единицу времени – это абсолютная пропускная способность администратора. Администратор занят обслуживает компьютеров в состояниях $z_1$ и $z_2$. Находясь в каждом из этих состояниях, он в среднем обслуживает в единицу времени $\mu$ компьютеров, тогда, с учётом вероятностей состояний, получим:

$A = 0\mathrm{P}_0 + \mu\mathrm{P}_1 + \mu\mathrm{P}_2 = \mu(\mathrm{P}_1 + \mathrm{P}_2) = (1 - \mathrm{P}_0)\mu.$

Вероятность того, что все компьютеры исправны:

$\mathrm{P}_{\text{исправны}} = \mathrm{P}_0.$

Среднее число неисправных компьютеров:

$C_{\text{неисправны}} = 1\mathrm{P}_1 + 2\mathrm{P}_2 + \dots + n\mathrm{P}_n.$

### *Обобщённые системы обслуживания*

Выше были рассмотрены математические модели простейших классических СМО. Их использование позволяет выявлять фундаментальные качественные связи между характеристиками СМО и параметрами простейших потоков заявок и потоков обслуживания, выраженные в аналитической форме. На практике же потоки заявок или потоки обслуживания зачастую не являются простейшими потоками. Кроме того, для произвольной СМО модель непрерывной марковской цепи может оказаться не адекватной из-за существенного влияния предыстории на характеристики функционирования СМО. Поэтому построить математическую модель СМО в виде системы дифференциальных уравнений Колмогорова для нахождения вероятностей состояний не представляется возможным.

Рассмотрим основные особенности функционирования обобщённой СМО (Рисунок 17).

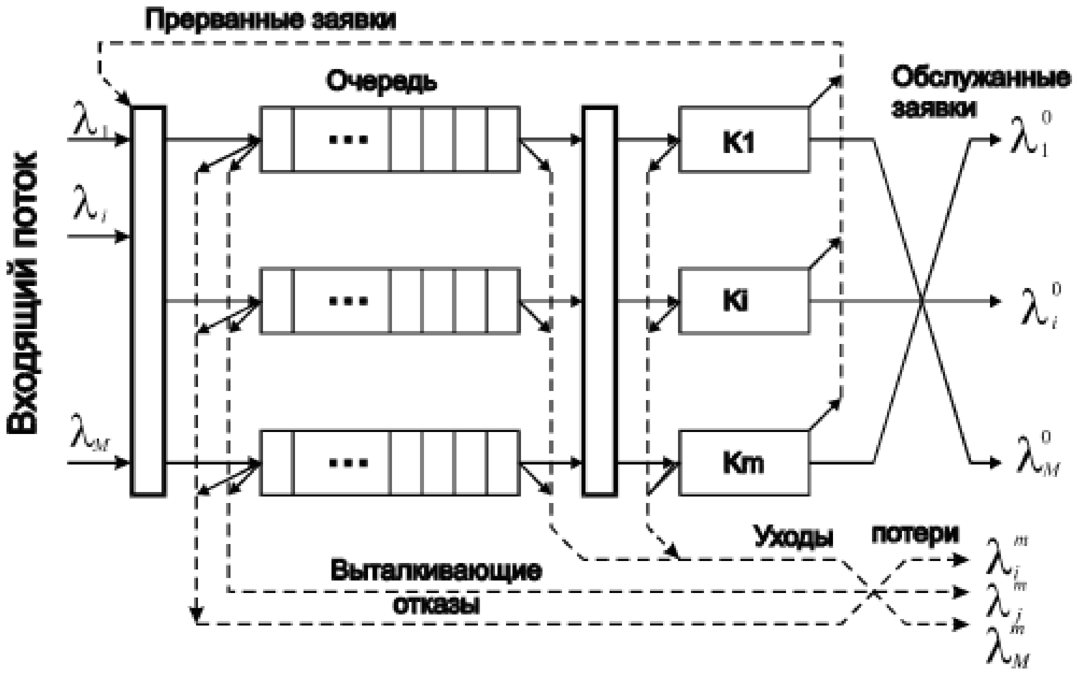



Рисунок 17 Обобщённая схема СМО

Принципиальной особенностью обобщённой СМО является то, что поступающие на вход потоки заявок в общем случае являются не однородными и отличаются от простейших. При этом заявки разного типа (например, с разным приоритетом) имеют различные условия обслуживания. Иными словами, вероятности состояний системы зависят от параметров заявки. Например, причиной отказа может служить отсутствие места в очереди на обслуживание для заявки данного типа. При этом заявка выталкивается из СМО, образуя поток *выталкивающих отказов* с интенсивностью $\lambda_j^m$.

Другой важной характеристикой заявки может быть ограниченное время ожидания начала обслуживания. По истечении этого времени заявка вынужденно покидает систему, образуя *поток уходов* с интенсивностью $\lambda_M^m$. Поток выталкивающих отказов вместе с потоком уходов образует обобщённый поток *потерь* системы. Заявки, поступившие на обслуживание в канал, и полностью обслужились покидают СМО, образуя на выходе поток обслуженных заявок с интенсивностью $\lambda_i^0$.

Различают разомкнутые и замкнутые СМО. В замкнутых СМО в зависимости от дисциплины обслуживания, реализуемой СМО, обслуживание заявки, поступившей в канал на обслуживание, может быть прервано, и недообслуженная заявка вновь отправляется в очередь на обслуживание (квантование времени обслуживания).

В разомкнутой СМО заявки, формирующие выходной поток, не могут снова поступить в какой-либо элемент СМО, т.к. обратная связь отсутствует.

Как правило, модели СМО, адекватно отражающие особенности реальных систем, являются композицией нескольких элементарных приборов обслуживания $П_i$. Если каналы $K_i$ различных приборов обслуживания соединены параллельно, то имеет место многоканальное обслуживание (*многоканальная* СМО), а если приборы $П_i$ и их параллельные композиции соединены последовательно, то имеет место многофазное обслуживание (*многофазная* СМО). При этом нужно заметить, что выходной поток (поток обслуженных заявок), покидающий элементарный прибор обслуживания, обычно имеет последействие, даже если входной поток его не имел. Чтобы убедиться в этом, рассмотрим одноканальную систему массового обслуживания, для которой время обслуживания одной заявки вполне определено и равно $\tau_{об}$. Тогда в потоке обслуженных заявок минимальный интервал времени между заявками, покидающими систему, будет равен $\tau_{об}$. Нетрудно убедиться, что наличие такого минимального интервала неизбежно приводит к последействию. Действительно, пусть стало известно, что в какой-то момент $t_1$ систему покинула обслуженная заявка. Тогда можно утверждать с достоверностью, что на любом участке времени $\tau$, лежащем в пределах $(t_1,t_1+\tau_{об})$, не появится обслуженной заявки; значит, будет иметь место зависимость между количествами событий на не перекрывающихся временных участках. Последействие, присущее выходному потоку, необходимо учитывать, если этот поток является входным для какой-либо другой системы обслуживания (например, при многофазном обслуживании).

Отметим, также, что самый простой на первый взгляд регулярный поток, в котором события отделены друг от друга равными интервалами, отнюдь не является "простейшим", так как в нем имеется ярко выраженное последействие: моменты появления следующих друг за другом событий связаны жёсткой, функциональной зависимостью. Именно из-за наличия последействия анализ процессов, протекающих в системе массового обслуживания при регулярном потоке заявок, гораздо сложнее, чем при простейшем.

Аналитические модели СМО применяются, прежде всего, для упрощённого качественного исследования характеристик СМО с классической структурой и "удобными" потоками входных заявок и законами обслуживания. Для исследования СМО сложной структуры они практически малопригодны. В этом случае применяют модели, основанные на компьютерной объектной имитации реальной работы СМО (так называемы имитационные модели) и получении характеристик системы методами статистической обработки результатов экспериментирования с моделью. В отличие от аналитического моделирования, имитационное моделирование практически всегда реализуемо и позволяет получать приемлемые результаты исследования СМО любой сложности.

### *Контрольные вопросы*

1. Понятие непрерывных событийно-стохастических систем.
2. Понятие систем массового обслуживания.
3. Элементарный прибор обслуживания и характеристики его функционирования.
4. Понятие абсолютной и относительной пропускной способности прибора обслуживания.
5. Понятие потока событий. Классификация потоков случайных событий.
6. Понятие регулярного потока.
7. Понятие потока без последействия.
8. Понятие потока с ограниченным последействием.
9. Поток Пальма.
10. Понятие стационарного потока событий.
11. Понятие ординарного потока событий.
12. Понятие простейшего потока случайных событий. Интенсивность потока.
13. Распределение вероятности интервалов простейшего потока.
14. Потоки Эрланга. Свойства потока Эрланга.
15. Понятие марковского случайного процесса. Непрерывные марковские цепи.
16. Орграф состояний непрерывного марковского процесса.
17. Математическое моделирования СМО с простейшими потоками. Уравнения Колмогорова. Финальные вероятности состояний.
18. Моделирование одноканальной СМО с отказами.
19. Моделирование многоканальной СМО с отказами.
20. Моделирование одноканальной СМО с ожиданием.

21. Моделирование замкнутых СМО.
22. Особенности моделирования СМО произвольной сложности.