НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ

ВМИ

БЛАГОТВОРИТЕЛЬНЫЙ
ФОНД В. ПОТАНИНА

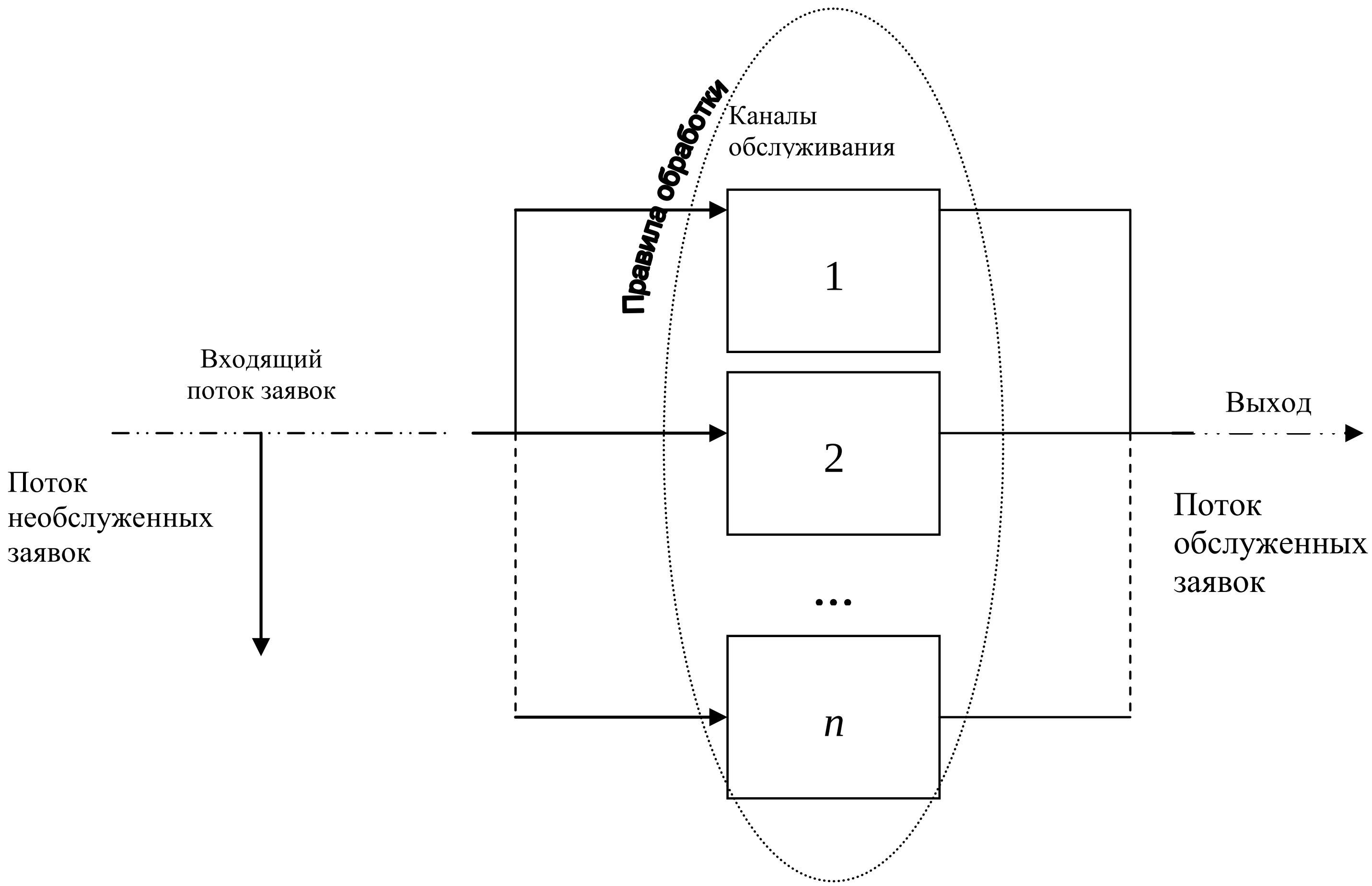
Дисциплина «Вероятностные модели»

Тема «Введение в теорию массового обслуживания**»**

Разработчик:
А.Т. Латипова, к.ф.м.н., доцент

Деятельность предприятий зачастую связана ***с многократной реализацией исполнения каких-то однотипных задач и операций*** (например, функционирование банков, страховых компаний, налоговых инспекций, сферы обслуживания – магазинов, больниц). Одним из видов моделей, с помощью которых можно описывать такими процессы, является ***системы массового обслуживания (СМО)***. Довольно эффективны ***подходы теории массового обслуживания (ТМО)*** при описании функционирования технических объектов – описании процессов в ЭВМ, их отдельных устройств и сетей составленных из ЭВМ, для решения задач для систем и сетей передачи информации.

Структура СМО



Каждая СМО включает в себя след. ***компоненты:***

1. *Каналы обслуживания (сервисы)* – лица, выполняющие те или иные операции – кассиры, менеджеры, операторы; или обслуживающие приборы, которые могут быть расположены <u>параллельно или последовательно</u>;
2. *Входящий поток заявок и требований* – который поступает на каналы обслуживания и ими обрабатывается, источник этого поток может иметь <u>конечную и бесконечную мощность</u>;
3. *Правила обслуживания на каждом канале* – инструкции, алгоритмы, привычки и т.д.

4. *Выходящий поток* – результат обслуживания заявки входного потока.
5. *Характеристики времени обслуживания каждого требования (заявки)*. Заявки или требования поступают на вход СМО в случайные моменты времени. Обслуживание заявок происходит за неизвестное обычно случайное время и зависит от множества разнообразных факторов. После обслуживания канал освобождается и готов к приему следующей заявки. Случайный характер потока заявок или требований и времени их обслуживания приводит к неравномерности загрузки СМО – перегрузке с образованием очередей заявок или недогрузке с простаиванием её каналов.
6. *Очередь*. У очереди может быть определенный порядок: FIFO, LIFO, случайный отбор. Длина очереди (число заявок в очереди) может быть ограниченной или неограниченной. Заявки могут выбираться СМО на основе их приоритетности. Клиенты могут перейти из одной очереди в другую для того, чтобы сократить время ожидания; или же покинуть очередь (отказаться от обслуживания).

Примеры систем массового обслуживания (СМО): телефонные станции, ремонтные мастерские, билетные кассы, справочные бюро, станочные и другие технологические системы, системы управления гибких производственных систем и т.д.

Каждая СМО состоит из какого – то количества обслуживающих единиц, которые называются **каналами обслуживания** (это станки, транспортные тележки, роботы, линии связи, кассиры, продавцы и т.д.). Всякая СМО предназначена для обслуживания какого – то **потока заявок** (требований), поступающих в какие – то случайные моменты времени.

Предмет теории массового обслуживания – построение математических моделей, связывающих заданные условия работы СМО (число каналов, их производительность, правила работы, характер потока заявок) с интересующими нас характеристиками – показателями эффективности СМО. Эти показатели описывают способность СМО справляться с потоком заявок. Ими могут быть: среднее число заявок, обслуживаемых СМО в единицу времени; среднее число занятых каналов; среднее число заявок в очереди; среднее время ожидания обслуживания и т.д.

Поступление заявок и обработка заявки СМО является **случайным процессом**. Случайный процесс, протекающий в системе, называется **марковским**, если для любого момента времени t_0 вероятностные характеристики процесса в будущем зависят только от его состояния в данный момент t_0 и не зависят от того, когда и как система пришла в это состояние.

Случайный процесс $x(t)$ называется **чисто случайным**, если случайные величины $x(t_1)$, $x(t_2)$, … взаимно независимы для любого конечного множества периодов времени t_1, t_2, ….

Случайный процесс $x(t)$ называется **марковским**, если для любого конечного множества $t_1 < t_2 < ... < t_{n-1} < t_n$

$$p(x(t_n), t_n \mid x(t_1), t_1; x(t_2), t_2; ...; x(t_{n-1}), t_{n-1}) = p(x(t_n), t_n \mid x(t_{n-1}), t_{n-1}).$$

Смысл формулы $p(x(t_n), t_n \mid x(t_1), t_1; x(t_2), t_2; ...; x(t_{n-1}), t_{n-1})$ – вероятность занчения случайной величины $x(t_n)$ для периода t_n при известных значениях $x(t_1), t_1; x(t_2), t_2; ...; x(t_{n-1}), t_{n-1}$ полностью определяется парой $x(t_{n-1}), t_{n-1}$, т.е. состоянием системы в предыдущий период t_{n-1}.

Марковский процесс задается **вероятностями перехода** $p(x(t_2),t_2 \mid x,t)$, $(t < t_2)$.

Если задано начальное состояние при $t = t_1$, то для марковских процессов можно использовать уравнения **Колмогорова-Смолуховского-Чепмена**

$p(x(t_2),t_2 \mid x(t_1),t_1) = \sum_{x(t)} p(x(t_2),t_2 \mid x,t) p(x,t \mid x(t_1),t_1)$, где $(t_1 \le t \le t_2)$.

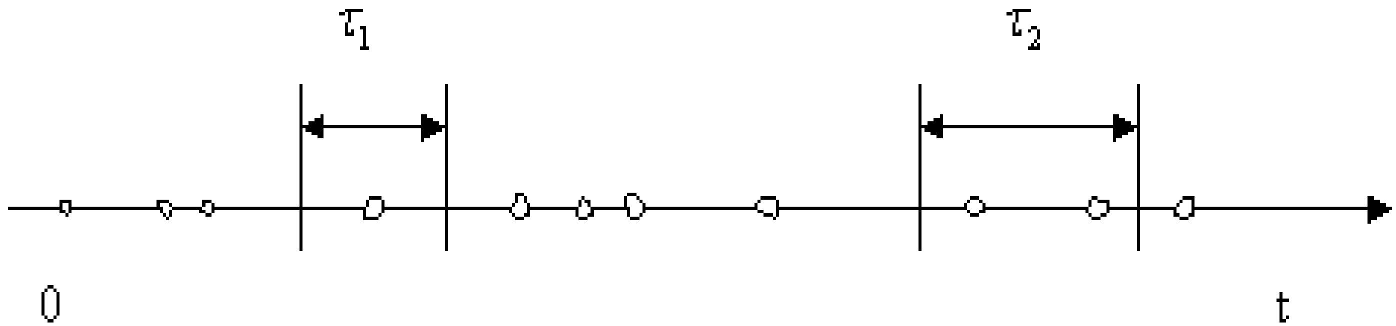
На практике марковские процессы в чистом виде обычно не встречаются. Но имеются процессы, для которых влиянием «предистории» можно пренебречь. И при изучении таких процессов можно применять марковские модели (в теории массового обслуживания рассматриваются и не марковские системы массового обслуживания, но математический аппарат, их описывающий, гораздо сложнее).

В исследовании операций большое значение имеют марковские случайные процессы с дискретными состояниями и непрерывным временем.

Процесс называется **процессом с дискретным состоянием**, если его возможные состояния S_1, S_2, … можно заранее определить, и переход системы из состояния в состояние происходит «скачком», практически мгновенно.

Процесс называется **процессом с непрерывным временем**, если моменты возможных переходов из состояния в состояние не фиксированы заранее, а неопределенны, случайны и могут произойти в любой момент. Далее рассматриваются только процессы с дискретным состоянием и непрерывным временем.

Поток событий – последовательность однородных событий, следующих одно за другим в какие-то случайные моменты времени. В предыдущем примере – это поток отказов и поток восстановлений. Другие примеры: поток вызовов на телефонной станции, поток покупателей в магазине и т.д. Поток событий можно наглядно изобразить рядом точек на оси времени *Ot*.



Изображение потока событий на оси времени

Положение каждой точки случайно, и здесь изображена лишь какая-то одна реализация потока.

Интенсивность потока событий (λ) – это среднее число событий, приходящееся на единицу времени. Интенсивность рассчитывается по формуле

$$\lambda = N / T,$$

где N – число заявок за период времени длительностью T.

Поток событий называется **стационарным**, если его вероятностные характеристики не зависят от времени.

В частности, интенсивность λ стационарного потока постоянна. Поток событий неизбежно имеет сгущения или разрежения, но они не носят закономерного характера, и среднее число событий, приходящееся на единицу времени, постоянно и от времени не зависит.

Поток событий называется **потоком без последствий**, если для любых двух непересекающихся участков времени τ_1 и τ_2 число событий, попадающих на один из них, не зависит от того, сколько событий попало на другой. Другими словами, это означает, что события, образующие поток, появляются в те или иные моменты времени **независимо друг от друга** и вызваны каждое своими собственными причинами.

Поток событий называется **ординарным**, если события в нем появляются поодиночке, а не группами по нескольку сразу.

Поток событий называется **простейшим (или стационарным пуассоновским),** если он обладает сразу тремя свойствами:

- стационарен,
- ординарен,
- не имеет последствий.

Простейший поток имеет наиболее простое математическое описание. Он играет среди потоков такую же особую роль, как и закон нормального распределения среди других законов распределения. А именно, при наложении достаточно большого числа независимых, стационарных и ординарных потоков (сравнимых между собой по интенсивности) получается поток, близкий к простейшему.

Для построения распределения используются **следующие аксиомы**:

<u>Аксиома 1</u>. Если $x(t)$ – число событий, происшедших на протяжении интервала времени $(0,t)$, то вероятностный процесс, описывающий $x(t)$ имеет независимые стационарные приращения; вероятность наступления события в интервале $(t,t+\tau)$ зависит лишь от его длины τ.

<u>Аксиома 2</u>. Вероятность того, что событие наступит на достаточно малом временном интервале $h>0$, положительна, но меньше 1.

<u>Аксиома 3</u>. На достаточно малом интервале $h>0$ может осуществиться не более одного события, т.е. $p(x(h)>1)=0$.

Тогда *переходные вероятности* обладают следующими свойствами при интенсивности λ:

$$p(x(t_2), t_2 \mid x, t) = \begin{cases} 0 & \text{при } x(t_2) < x, \\ 1 - \lambda\Delta t + o(\Delta t) & \text{при } x(t_2) = x, \\ \lambda\Delta t + o(\Delta t) & \text{при } x(t_2) = x+1, \\ o(\Delta t) & \text{при } x(t_2) > x+1, \end{cases}$$

где $\Delta t = t_2 - t$; $x, x(t_2) = 0, 1, 2, \ldots$, а $\lim\limits_{\Delta t \to 0} o(\Delta t) / \Delta t = 0$.

Обозначим $p(K,T) = p(x,t \mid x(t_1),t_1)$, где $(K = x - x(t_1) = 0,1,2,\ldots; T = t - t_1)$.
Воспользовавшись уравнениями Колмогорова, получим:

$$P(K,T+\Delta t) - P(K,T) = -\lambda P(K,T)\Delta t + \lambda P(K-1,T)\Delta t + o(\Delta t) \cdot (\quad K = 0,1,2,\ldots)$$

Поделим это уравнение на $\Delta t \to 0$. Тогда получаем дифференциальное уравнение:

$$\partial P(K,T) / \partial T = -\lambda P(K,T) + \lambda P(K-1,T).$$

Начальными условиями для этого дифференциального уравнения будут

$$P(-1,T) \equiv 0, \quad p(0,0) = p(x(t_1),t_1 \mid x(t_1),t_1) = 1,$$

$$p(K,0) = p(x(t_1)+K,t_1 \mid x(t_1),t_1) = 0 \quad (K > 0).$$

Тогда вероятность ненаступления события (неполучения заявки) получается след. образом.

$$\partial P(0,T)/\partial T = -\lambda P(0,T) + \lambda P(-1,T) \cdot$$

$$\Downarrow$$

$$\partial P(0,T)/\partial T = -\lambda P(0,T)$$

$$\Downarrow$$

$$\partial P(0,T)/P(0,T) = -\lambda \partial T$$

$$\Downarrow$$

$$\ln p(0,T) = -\lambda T + \ln C_1 \quad (C_1 = \mathrm{const})$$

$$\Downarrow$$

$$p(0,T) = e^{-\lambda T} C_1$$

$$\Downarrow$$

$$\left(p(0,0) = 1\right) \Rightarrow C_1 = 1$$

$$\Downarrow$$

$$p(0,T) = e^{-\lambda T}$$