

# Аналитико-статистическое моделирование информационных систем

Кафедра информационных управляющих систем

Лекции читает

канд.техн.наук, доцент

Литвинов Владислав Леонидович

## Список литературы:

1. О.И. Кутузов, Т.М. Татарникова

### **МОДЕЛИРОВАНИЕ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ**

<http://dvo.sut.ru/libr/ius/w101kutu/index.htm>

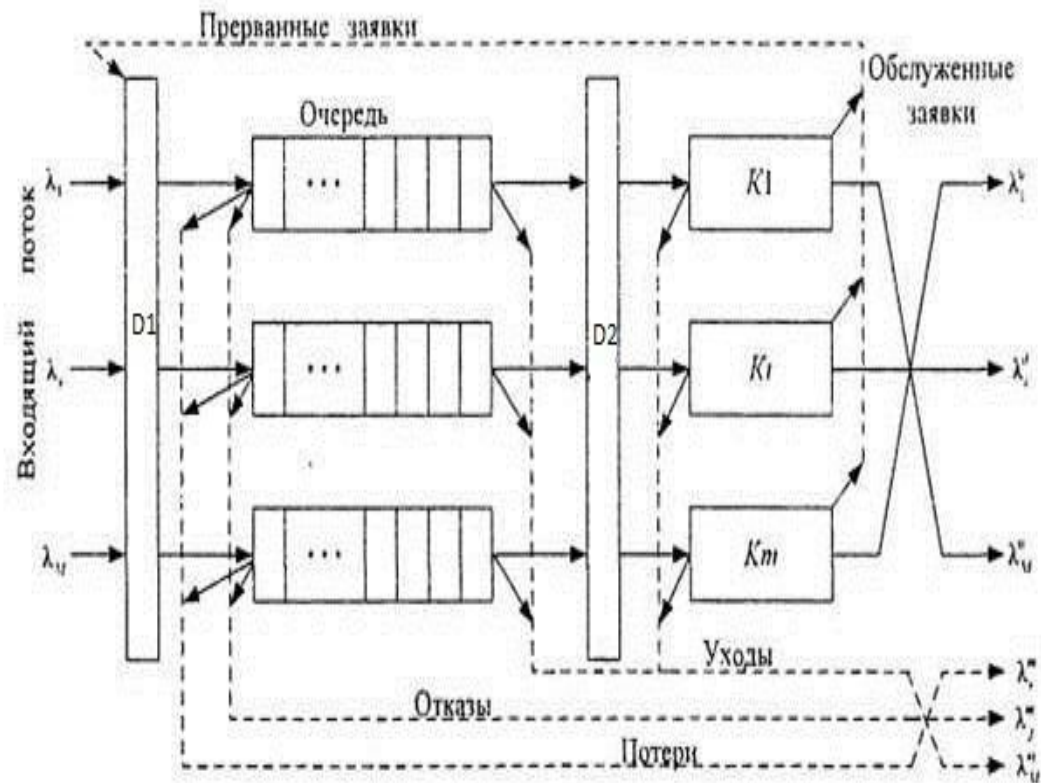
2. Боев В. Д, Моделирование систем. Инструментальные средства GPSS WORLD. Учеб. пособие — СПб.: БХВ-Петербург, 2004. — 368 с.
3. Боев В. Д, Сыпченко Р. П. Компьютерное моделирование. Элементы теории и практики. Учеб. пособие — СПб.: Военная академия связи, 2009. — 432 с.
4. Бражник А. Н, Имитационное моделирование: возможности GPSS WORLD — СПб.: Реноме, 2006. — 439 с.

# Тема лекции 3:

- **СИСТЕМЫ МАССОВОГО  
ОБСЛУЖИВАНИЯ И ИХ  
МОДЕЛИРОВАНИЕ**

## Общая характеристика систем массового обслуживания

- Одним из математических методов исследования сложных стохастических систем является теория массового обслуживания (ТМО), занимающаяся анализом эффективности функционирования так называемых систем массового обслуживания (СМО).
- *Под СМО понимают динамическую систему, предназначенную для эффективного обслуживания случайного потока заявок при ограниченных ресурсах системы.* Системы массового обслуживания называют также Q-схемами (англ. queuing system). Обобщенная схема СМО приведена на рис.1.



Работа любой такой системы заключается в обслуживании поступающего на нее потока требований или заявок (вызовы абонентов, приход покупателей в магазин, требования на выполнение работы в мастерской и т.д.). Заявки поступают в систему одна за другой в некоторые случайные моменты времени. Обслуживание поступившей заявки продолжается какое-то время, после чего система освобождается для обслуживания очередной заявки.

Каждая такая система может содержать конечное число элементов обслуживания, называемых *каналами обслуживания* (линия связи, приборы). Примерами таких систем могут быть телефонные станции, билетные кассы, аэродромы, радиолокационные станции, ЭВМ, вычислительные центры, информационные системы различных видов.

- Отметим случайный характер поступления заявок и случайный характер промежутков времени, необходимого для выполнения заявок. В целом, имеем случайный процесс, в котором возможны как перегрузки, так и простои. Если в момент появления заявки на входе СМО есть свободный канал, ее обслуживание начинается немедленно. Если СМО загружена (т.е. все каналы заняты), заявка занимает место в очереди. На число мест в очереди может быть наложено ограничение. При этом возможны конфликтные ситуации, решением которых может быть либо отказ системы принять заявку, либо принятие заявки в систему за счет выталкивания из очереди другой, менее ценной в данный момент времени заявки.
- Процесс продвижения заявок в СМО осуществляется по некоторому закону управления, который задается дисциплинами ожидания  $D1$  и обслуживания  $D2$ . Дисциплина ожидания определяет порядок приема заявок в систему и размещения их в очереди, а дисциплина обслуживания - порядок выбора заявок из очереди для назначения на обслуживание.



- Поступающие на вход СМО (рис. 1) однородные заявки в зависимости от порождающей причины делятся на типы. Совокупность заявок всех типов образует входящий поток СМО. Интенсивность потока заявок типа  $i$  ( $i = 1, \dots, M$ ) обозначим

Обслуживание заявок выполняется  $m$  каналами ( $K$ ). Различают универсальные и специализированные каналы обслуживания. Для универсального канала типа  $j$  считаются известными функции распределения  $F(t)$  длительности обслуживания заявок произвольного типа. Для специализированных каналов функции распределения длительности обслуживания заявок некоторых типов являются неопределенными.

- В качестве примера процесса обслуживания можно рассматривать различные по своей физической природе процессы функционирования экономических, производственных, технических и других систем, например, потоки поставок продукции некоторому предприятию, потоки деталей и комплектующих изделий на сборочном конвейере цеха, заявки на обработку информации ЭВМ от удаленных терминалов и т.д. При этом характерными для работы таких объектов являются случайное поведение заявок (требований) на обслуживание и завершение обслуживания в случайные моменты времени.

- Основными элементами сети связи, представляемой как СМО, выступают узлы и линии (каналы) связи. Эти элементы предназначены для обслуживания вызовов и, следовательно, являются системами массового обслуживания.
- Обслуживание вызова в узле характеризуется заданным алгоритмом и может быть описано как последовательность логических операций. Свойства узла СМО (производительность коммутатора, потери вызовов на групповых устройствах и т.д.) могут быть описаны моделями элементарных СМО, которые и будут рассмотрены в дальнейшем.
- Пропускная способность СМО зависит от числа каналов, их производительности и характера потока заявок.
- Теория массового обслуживания устанавливает зависимость между характеристиками потока заявок, числом каналов, их производительностью, правилами работы СМО, ее эффективностью и часто включает экономический аспект. Например, стремятся найти наименьшую полную стоимость единицы времени ожидания обслуживания требованиями в накопителе и простоя приборов. Кроме того, к задачам ТМО близки задачи повышения надежности технических устройств.

- Для характеристики СМО обычно применяют следующие показатели:
  - среднее число заявок, которые система может обслужить за единицу времени;
  - средний процент необслуженных заявок;
  - вероятность того, что поступившая заявка будет принята к выполнению;
  - среднее время ожидания в очереди;
  - закон распределения времени ожидания;
  - среднее число заявок в очереди;
  - закон распределения числа заявок в очереди;
  - средний доход системы в единицу времени.
- Наиболее удобны для анализа те системы, в которых случайный процесс является близким к **марковскому**, т.е. состояние системы в будущем зависит от состояния системы в настоящее время, но не зависит от того, каким образом эта система пришла к этому состоянию (не учитывает прошлое). В таком случае задачу ТМО можно описать обыкновенными дифференциальными уравнениями и выразить в явном виде основные характеристики обслуживания через параметры системы. В остальных случаях характеристики оцениваются приближенно.

В любом элементарном акте обслуживания можно выделить две основные составляющие: ожидание обслуживания заявки и собственно обслуживание заявки. Это можно отобразить в виде некоторого  $i$ -го прибора обслуживания  $\Pi$ , состоящего из накопителя заявок  $H$ , и канала обслуживания заявок  $K$ , (рис. 8.2). В накопителе может находиться одновременно  $l = 0, \dots, L$ , заявок, где  $L$  - емкость  $i$ -го накопителя. На каждый элемент прибора обслуживания  $\Pi$  поступают потоки событий: в накопитель  $H$  поток заявок  $w$  на канал  $K$  - поток обслуживания  $u$ .

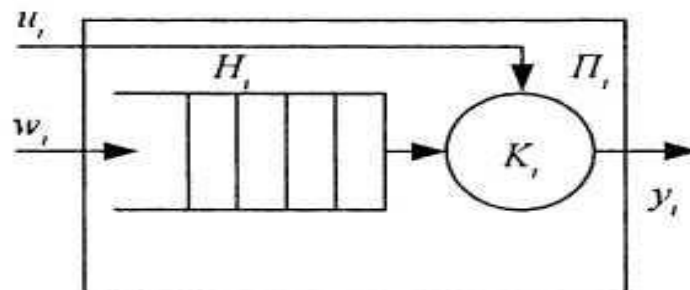


Рис. 8.2. Схема прибора СМО

- *Потоком событий (ПС) называется последовательность событий, происходящих одно за другим в какие-то случайные моменты времени.* Различают однородные и неоднородные потоки событий. *Однородный ПС (ОПС)* характеризуется только моментами поступления этих событий (вызывающими моментами) и задается последовательностью
- $\{t_n = \{0 \leq t_1 \leq t_2 \leq \dots \leq t_n\}$ , где  $t_n$  - момент поступления  $n$ -го события - неотрицательное вещественное число.
- Однородный ПС может быть также задан в виде последовательности промежутков времени между  $n$ - и  $(n-1)$ -ым событиями  $\{t_n\}$ .
- 
- *Неоднородным ПС* называется последовательность  $\{t_n; f_n\}$ , где  $t_n$  — моменты поступления заявки;  $f_n$  - набор признаков события.
- Например, может быть задана принадлежность к тому или иному источнику заявок, наличие приоритета, возможность обслуживания тем или иным типом канала и т.п.
- Рассмотрим ОПС, для которого  $t_i \in \{t_n\}$ . где  $t$  - случайные величины, независимые между собой. Тогда ПС называется *потоком с ограниченным последствием*.
- Поток событий называется *ординарным*, если пренебрежимо мала вероятность того, что на малый интервал времени  $\Delta t$ , примыкающий к моменту времени  $t$ , попадает больше одного события.

- *Стационарным ПС* называется поток, для которого вероятность появления того или иного числа событий на интервале времени  $\Delta t$  зависит от длины этого интервала и не зависит от того, где на оси времени взят этот интервал. Для ординарных ПС справедливо

$$0 \bullet P_0(t, \Delta t) + 1 \bullet P_1(t, \Delta t) = P_1(t, \Delta t)$$

- Среднее число событий, наступающих на интервале  $\Delta t$  в единицу времени, составляет  $P_1(t, \Delta t) / \Delta t$

- Рассмотрим предел этого выражения при  $\Delta t \rightarrow 0$ :

- $\lim_{\Delta t \rightarrow 0} P_1(t, \Delta t) / \Delta t = \lambda(t) \quad [1/\text{ед.врем}]$

- $\Delta t \rightarrow 0$

- Если этот предел существует, то он называется *интенсивностью (плотностью)* ОПС. Для стационарного ПС

- $\lambda(t) = \lambda = \text{const.}$

- Применительно к элементарному каналу обслуживания  $K_i$ , можно считать, что интервалы времени между моментами **появления заявок** на входе  $K_i$  образуют **подмножество неуправляемых переменных**, а **поток обслуживания**, т.е. интервалы времени между началом и окончанием обслуживания заявки образуют **подмножество управляемых переменных**. Заявки, обслуженные каналом  $K_i$  и заявки, по различным причинам не обслуженные и покинувшие прибор образуют выходной поток.
- Процесс функционирования прибора обслуживания  $\Pi$ , можно представить как процесс изменения состояний его элементов во времени  $Z_i(t)$ .
- Переход в новое состояние  $\Pi$  означает изменение количества заявок, которые в нем находятся (в канале  $K_i$  и накопителе  $H_i$ ).
- Таким образом, вектор состояний для  $\Pi_i$  имеет вид  $Z_i = (z_i^H; z_i^K)$ ,
- где  $z_i^H$  - состояния накопителя ( $z_i^H = 0$  - накопитель пуст,  $z_i^H = 1$  - в накопителе одна заявка,  $z_i^H = L_i^H$  - накопитель занят полностью);
- $z_i^K$  - состояние канала  $k$ , ( $z_i^K = 0$  - канал свободен,  $z_i^K = 1$  - канал занят).

- Следует отметить, что в ТМО, в зависимости от емкости накопителя, существуют:
- системы с потерями ( $L_t^H = 0$ , накопитель отсутствует);
- системы с ожиданием ( $L_t^H \rightarrow \infty$ , емкость накопителя бесконечно большая);
- системы с ограниченной емкостью накопителя (смешанные).



- В ТМО существуют следующие дисциплины обслуживания: беспriorитетная, приоритетная, со смешанным приоритетом.
- *Беспriorитетное обслуживание* осуществляется по правилу FIFO и LIFO или по случайному правилу. Дисциплины FIFO и LIFO по среднему значению или математическому ожиданию ничем не отличаются, однако, приоритетной считается дисциплина FIFO, которая *минимизирует дисперсию времени ожидания* (т.е. уменьшается разброс времени относительно среднего значения).
- *Приоритетное обслуживание* заявок, находящихся в очереди, может выполняться по правилу *относительного и абсолютного* обслуживания. Для приоритетного обслуживания заявкам необходимо указать уровень приоритета СМО ( $0 > 1 > 2 > \dots$ ):
- *относительная дисциплина обслуживания* заключается в следующем: заявка, пришедшая на обслуживание с более высоким приоритетом, не прерывает заявку, находящуюся на обслуживании, и только после окончания обслуживания в канале приоритетная очередная заявка поступает на обслуживание;
- *абсолютная дисциплина обслуживания* заключается в том, что пришедшая на канал заявка с более высоким приоритетом вытесняет с обслуживания заявку с меньшим приоритетом.
- При *смешанном обслуживании* наряду с различными дисциплинами приоритетного обслуживания используется беспriorитетное обслуживание.

# Классификация СМО

- В теории массового обслуживания приняты сокращенные обозначения, в основе которых лежит трехбуквенное обозначение вида
- $A/B/m$ ,
- где  $A$  и  $B$  описывают соответственно законы распределения промежутков времени между последовательно поступающими заявками и распределение времени их обслуживания, а величина  $m$  - число обслуживающих приборов;
- $A$  и  $B$  принимают значения из следующего набора символов:
- $M$ - показательное распределение;
- $E_r$  — распределение Эрланга порядка  $r$ ;
- $D$  - детерминированное распределение;
- $G$  - распределение произвольного вида.
- Так, система обслуживания  $M/G/1$  представляет собой систему с пуассоновским входным потоком, произвольным распределением времени обслуживания и одним обслуживающим прибором. Символ  $D$  означает, что время обслуживания - постоянная величина, поэтому система с пуассоновским входным потоком в этом случае обозначается как  $M/D/1$ .
-

- Иногда приходится указывать также емкость накопителя системы
- (которую обозначим через  $K$ ) или число источников нагрузки (которое обозначим через  $M$ ); в этом случае будет использоваться пятибуквенное обозначение:  $A / B / m / K / M$ . В случае отсутствия одного из последних индексов предполагается, что его значения сколь угодно велико.
- Запись вида  $D/M/2/20$  означает систему с двумя обслуживающими приборами, постоянным (детерминированным) временем между двумя последовательно поступающими заявками, показательным распределением длительности обслуживания и накопителем емкостью 20 заявок.
- Различают два вида СМО: *одноканальные и многоканальные*. Также различают *системы с отказами*, в которых заявка получает отказ, если все каналы заняты, и *системы без отказов (с ожиданием или очередью)*, которые делятся на упорядоченные и неупорядоченные.
- *Упорядоченные системы* отличаются тем, что освободившиеся каналы принимают заявки в порядке очереди. *Неупорядоченные системы* характеризуются тем, что при освобождении канала заявка выбирается случайным образом. Например, поступившее требование может занять место в самой короткой очереди; в этой очереди оно может расположиться последним (такая очередь будет упорядоченной), а может пойти на обслуживание вне очереди.
- Кроме того, системы с ожиданием разделяются на системы с *ограниченным и неограниченным временем ожидания*.
- *Смешанной системой* обслуживания называется система, в которой требование, заставшее все приборы занятыми, становится в очередь лишь в том случае, когда число требований, находящихся в системе, не превосходит определенного уровня (в противном случае происходит потеря требования).

# Показатели эффективности и основные характеристики СМО

- Показатели эффективности СМО зависят от вида систем.
- Для систем с отказами это абсолютная и относительная пропускная способность систем.
- *Абсолютная пропускная способность* - среднее число выполненных заявок в единицу времени.
- *Относительная пропускная способность* - средняя доля поступивших заявок, определяемая отношением среднего числа выполненных заявок к общему числу поступивших заявок в единицу времени.
- Кроме того, можно определить среднее число занятых каналов или среднее относительное время простоя одного канала или всей системы в целом.
- Для систем без отказов (с ожиданием) выбираются другие показатели эффективности. Если система с неограниченным ожиданием, то за показатели эффективности принимают:
  - *среднее число заявок в очереди,*
  - *среднее число заявок в системе,*
  - *время ожидания в очереди,*
  - *время выполнения заявок.*

- Для систем с ограниченным временем ожидания применимы обе группы показателей: абсолютная и относительная пропускная способность и характеристики ожидания. При этом нужно знать следующие параметры:
- $n$  - число каналов;  $\lambda$  — интенсивность потока заявок;  $\mu = 1/t_{\text{обсл}}$  - производительность (интенсивность обслуживания) каждого канала (среднее число заявок).
- Основные характеристики простейших СМО следующие:
  - Коэффициент загрузки устройства или канала —  $\rho = \lambda / \mu < 1$ . При этом коэффициент загрузки является вероятностью обслуживания заявки в канале.
  - Пусть СМО работает достаточно длительное время  $T$ , тогда число заявок в системе будет определяться как  $\lambda T$  и среднее время обслуживания заявок будет определяться как  $\lambda T t_{\text{обсл}}$ . В этом случае
    - $\lim_{T \rightarrow \infty} \lambda T t_{\text{обсл}} / T = \lambda t_{\text{обсл}} = \rho$
    - $T \rightarrow \infty$
  - Коэффициент загрузки  $\rho$  имеет смысл только лишь для установившихся режимов.
  - Коэффициент простоя канала  $\eta = 1 - \rho$ .
  - Время пребывания заявки в системе  $T_c = T_{\text{ож}} + M(T_{\text{обсл}})$  где  $T_{\text{ож}}$  - время ожидания заявки на обслуживание;  $M(T_{\text{обсл}})$  - среднее время обслуживания.
  - Время ожидания в СМО, которое в общем случае может состоять тоже из двух компонент:
    - $T_{\text{ож}} = T^H + T^П$ , где  $T^H$  - время ожидания обслуживания или время начала обслуживания;  $T^П$  - время ожидания в прерванном состоянии.
  - Длина очереди, которая определяется как  $l = \lambda T^H$
  - Среднее число заявок в системе  $n = \lambda(T^H + M(T_{\text{обсл}}))$  (формула Литтла).

# Структура системы массового обслуживания

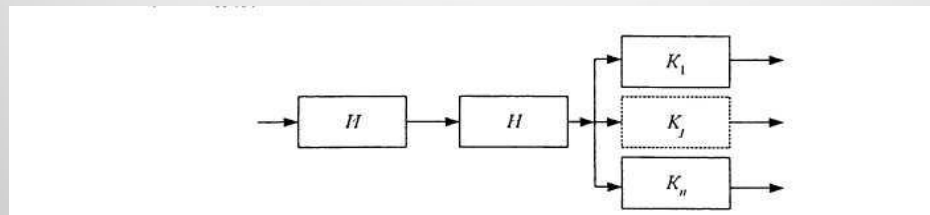
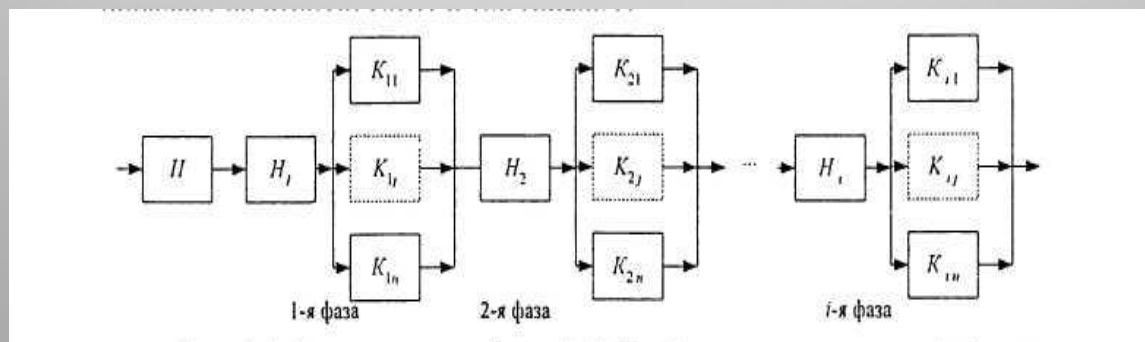


Рис. 3. Структура простейшей СМО:  $\Pi$ - источник заявок (при моделировании — генератор случайных событий - ГСС);  $H$  - накопитель;  $K_j$ - канал обслуживания.



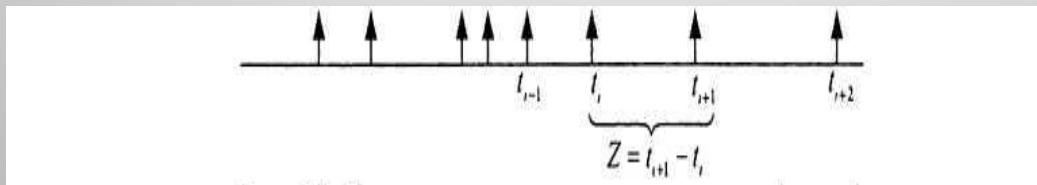
Для более сложных, многофазных систем выделяются фазы обслуживания, когда заявка, получив обслуживание в первой СМО, сразу попадает во вторую СМО и т.д. (рис.4).

- Система массового обслуживания считается заданной, если определены следующие характеристики:
- *Входящий поток требований* или, иначе говоря, моменты поступления требований в систему. (В дальнейшем будем считать, что источник располагает неограниченным числом требований и что требования однородны, т.е. различаются только моментами появления в системе.)
- *Структура системы обслуживания*, состоящая из накопителя и узла обслуживания, представляющего собой одно или несколько обслуживающих устройств, которые в дальнейшем будем называть *каналами*. Может оказаться, что требованиям придется ожидать, пока каналы освободятся. В этом случае требования находятся в накопителе, образуя одну или несколько очередей.
- *Время обслуживания* каждым каналом.
- *Дисциплина ожидания*, т.е. совокупность правил, регламентирующих хранение требований, находящихся в один и тот же момент времени в системе.
- *Дисциплина очереди*, т.е. совокупность правил, в соответствии с которыми требование отдает предпочтение той или иной очереди (если их несколько) и располагается в выбранной очереди.
- *Дисциплина обслуживания*, т.е. совокупность правил, в соответствии с которыми требование выбирает прибор, которым оно будет обслужено.

На практике обычно моменты поступления требований в систему случайны. Часто бывает, что поток требований тоже случайный, тогда случайна и длительность обслуживания.

По аналогии с вышеизложенным модели элементарных СМО, как правило, включают: модель обслуживания вызовов (заявок) группой устройств (пучком линий) при дисциплине обслуживания с потерями; модель обслуживания вызовов (заявок) устройством при дисциплине обслуживания с ожиданием.

Модель процесса поступления вызовов (заявок) показана на рис. 5.



Вероятность того, что промежуток времени  $Z$  между вызовами меньше некоторой величины  $z$ , называется функцией распределения вероятностей промежутков времени между вызовами  $P(Z < z) = F(z)$ . Каждый вызов характеризуется временем его обслуживания (временем занятия). Предположим, что продолжительность обслуживания вызовов потока является случайной величиной с экспоненциальным законом распределения, тогда

$$F(t) = 1 - e^{-\mu t} ; \quad \omega(t) = \mu e^{-\mu t} ; \quad M(T) = 1/\mu \quad (1)$$

где  $\mu$  - интенсивность обслуживания;  $M(T)$  - средняя продолжительность обслуживания.



Задача моделирования потока заявок и времени обслуживания заключается в получении последовательности случайных чисел с заданными законами распределения вероятностей. Модель процесса обслуживания вызовов показана на рис. 6.

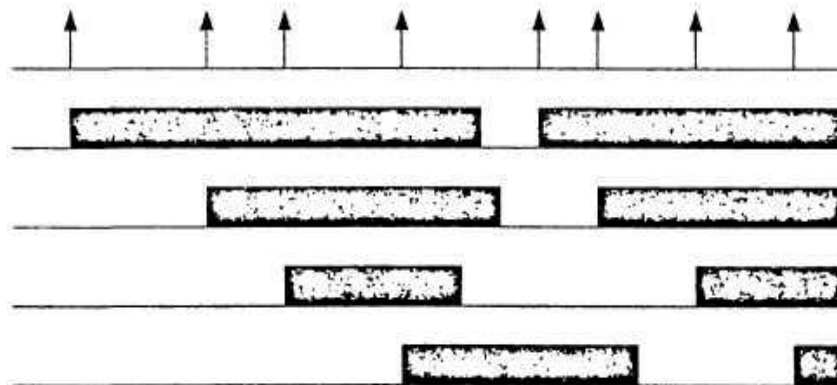


Рис. 8.6. Модель процесса обслуживания вызовов

# Системы массового обслуживания с ожиданием

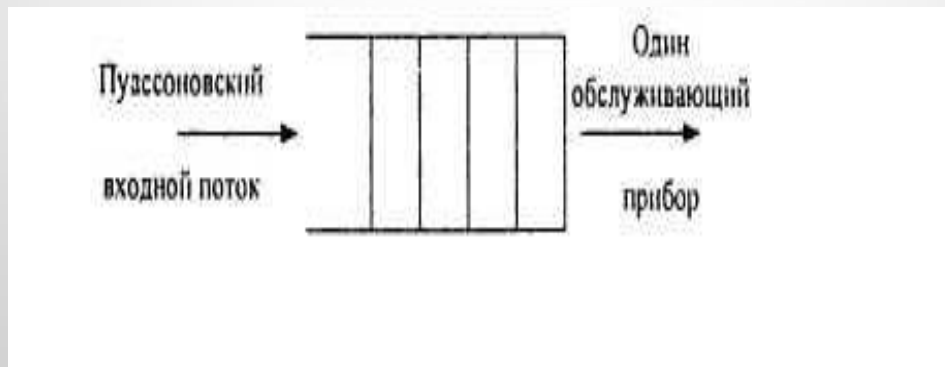
- Система обслуживания  $M/M/1$

- В качестве модели процесса поступления сообщений в такой СМО будем предполагать пуассоновский поток поступлений. В таком случае интервалы времени  $\tau$  между последовательными поступлениями сообщений являются непрерывными случайными величинами, распределенными по экспоненциальному закону:
- $$\omega(t) = \lambda e^{-\lambda t} \quad (2)$$
- Как было показано выше, среднее время между поступлениями сообщений в этом случае определяется из выражения

- $$M(\tau) = \int_0^{\infty} \tau \omega(\tau) d\tau = 1 / \lambda$$

- Параметр  $\lambda$ , введенный как коэффициент пропорциональности при определении пуассоновского процесса, определяет интенсивность входного потока. Будем также полагать, что длина сообщения  $r$  распределена по экспоненциальному закону со средним значением  $1/\mu$  (в теории очередей говорят об экспоненциальном распределении времени обслуживания), так что плотность распределения длины сообщения  $r$  определяется из выражений (1).
- Подобное предположение о распределении данных, как правило, не соответствует действительности. На практике длина сообщения является дискретной величиной (длина кратна используемой единице данных), в то время как величина  $r$  непрерывна. Вместе с тем, предположение об экспоненциальном распределении длины сообщения позволяет быстро получить искомый результат и облегчает изучение более сложных моделей обслуживания, поэтому в литературе его часто используют.
- Статистический анализ существующих сетей передачи данных показывает, что предположение об экспоненциальном распределении длины сообщения приводит к относительно хорошим результатам.

- Если пропускная способность выходного канала  $C$  единиц данных/сек, то очевидно, что на передачу или обслуживание сообщения ДЛИНОЙ  $r$  единиц потребуется  $r/C$  сек. Плотность распределения времени обслуживания можно записать в виде (8.1), т.е.
- $$\omega(t) = (C/L) e^{-(C/L)t}$$
- В этом случае  $M(T) = L/C$  -среднее время передачи сообщения.
- Заметим, что величина  $C/L$  используется для обозначения интенсивности обслуживания сообщений, а распределение времени обслуживания, как и распределение времени между последовательными поступлениями сообщений, является экспоненциальным.
- Рассмотрим модель системы обслуживания  $M/M/1$ , представленную на рис. 7. Процесс поступления сообщений - пуассоновский, объем буферной памяти не ограничен, распределение времени обслуживания - экспоненциальное. Сообщения обслуживаются по принципу FIFO («первым пришел - первым обслужен»). Такая модель обслуживания является одной из простейших в теории очередей.



Предположение о бесконечности числа мест для ожидания означает на практике выделение для хранения запросов, входной и выходной информации таких объемов памяти буферов и базы данных, которые при правильной эксплуатации гарантируют отсутствие информационных потерь в системе вследствие их возможного переполнения. В последние годы прослеживается устойчивая тенденция к существенному увеличению объемов оперативной и внешней памяти и их удешевлению в современных средствах электронно-вычислительной техники. Поэтому проблемы с недостатком памяти возникают все реже и в ближайшем будущем, по-видимому, перестанут вызывать практические затруднения. С учетом изложенного введенное предположение о бесконечности числа мест для ожидания в системе представляется вполне обоснованным.

- Для модели обслуживания  $M/M/1$  определим  $p_n(t)$  - вероятность того, что в буферной памяти в момент времени  $t$  находится  $n$  сообщений. Эта вероятность позволяет определить различные статистические характеристики рассматриваемой СМО (среднее число сообщений в буферной памяти, вероятность превышения заданного уровня заполнения буферной памяти и т. д.).
- Общее выражение для  $p_n$  имеет вид  $p_n = \rho^n p_0$ .
- Очевидно, что должно выполняться соотношение  $\rho = \lambda L / C = \lambda / \mu < 1$ , чтобы вероятности состояний уменьшались с ростом  $n$ .
- Справедливость этого соотношения согласуется с нашим интуитивным представлением о том, что среднее число сообщений, поступающих за единицу времени ( $\lambda$ ), должно быть меньше пропускной способности системы. Напомним, что объем буферной памяти предполагается неограниченным.

- В модели, учитывающей ограниченность объема буферной памяти при ее переполнении, дальнейшее поступление сообщения блокируется.

Следовательно,

- $$\sum_{n=0}^{\infty} p_n = 1 = p_0 \sum_{n=0}^{\infty} \rho^n = p_0 (1 - \rho) \quad (3)$$

- а также

- $$p_0 = 1 - \rho. \quad (4)$$

- Физическая интерпретация из уравнения (4):  $\rho = (1 - p_0)$  - вероятность того, что буферная память не пуста.
- Выражение для вероятностей различных состояний системы обслуживания  $M/M/1$  можно использовать для определения всех интересующих статистических характеристик системы. Например, средняя длина очереди

- $$M(n) = \sum_{n=0}^{\infty} n P_n = \rho / (1 - \rho) \quad (5)$$

- Средняя длина очереди превышает любые ограничения при  $\rho \rightarrow 1$ .
- График ее зависимости от  $\rho$  в  $M/M/1$  изображен на рис. 8. При  $\rho < 0,5$  среднее число сообщений в очереди меньше 1. При  $\rho > 0,5$  это число резко возрастет.
- Так, при  $M(n) = 3$  коэффициент загрузки  $\rho = 0,75$ ; при  $M(n) = 4$  значение  $\rho = 0,8$ ;
- при  $M(n) = 9$  значение  $\rho = 0,9$  и т. д.



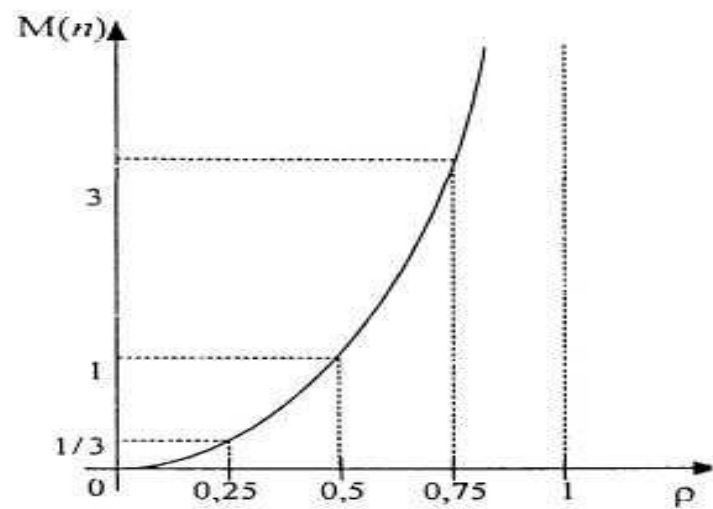


Рис. 8.8. Средняя длина очереди в системе  $M/M/1$

- Пусть в момент поступления очередного сообщения в буферной памяти уже находятся  $n$  сообщений. Тогда среднее время пребывания сообщения в концентраторе можно описать соотношением  $M(T) = T_{обсл} + T_{ож} = L/C + nL/C$ .
- Записанное выражение отражает зависимость между временем задержки и длиной очереди. Используя выражение (5), определяющее среднюю длину очереди, найдем среднее время задержки:
 
$$M(T) = L/C + L M(n)/C = L/(C (1-\rho)) = 1/(\mu-\lambda) \quad (6)$$
- Основное предположение, которое было сделано ранее, заключалось в том, что каждый узел сети может моделироваться как система обслуживания  $M/M/1$ , причем каждая система обслуживания статистически не зависит от других. Последнее предположение означает, что длины сообщений выбираются независимо из экспоненциального распределения в каждой узловой точке или точке концентрации. Очевидно, на практике это не так.
- Изучение реальных сетей показывает, что результаты, полученные при этом предположении, хорошо согласуются с практическими результатами.

- Другой, более общей формулой по сравнению с (6) является формула Литтла, соответствующая следующей теореме.
- **Теорема.** *Средняя длина очереди в системе равна среднему времени задержки, умноженному на интенсивность входного потока:*
- $M(T)\lambda = M(n) \quad (7)$
- Эта формула справедлива как для сети очередей, так и для отдельных систем обслуживания. Подставляя (5) в (7), получим формулу (6).