

Тема: Системы массового обслуживания (СМО)

1 Понятие СМО

2 Классификация СМО

3 Характеристики СМО

4 Структура обслуживания системы

5 Основные критерии эффективности функционирования СМО

6 Характеристики основных моделей СМО

1 Понятие СМО

Основоположник теории массового обслуживания датский ученый **А.К. Эрланг**

- 1909 г. «Теория вероятностей и телефонные переговоры»

Термин *теория массового обслуживания* предложен **А.Я. Хинчиным**

- 1932 г. «Математическая теория стационарной очереди»
- 1933 г. «О среднем времени простоя станков»
- 1963 г. «Работы по математической теории массового обслуживания»

В зарубежной литературе чаще используется термин *теория очередей*



Хинчин Александр
Яковлевич

Системы массового обслуживания – это такие системы, в которые в случайные моменты времени поступают заявки на обслуживание, при этом поступившие заявки обслуживаются с помощью имеющихся в распоряжении системы каналов обслуживания.

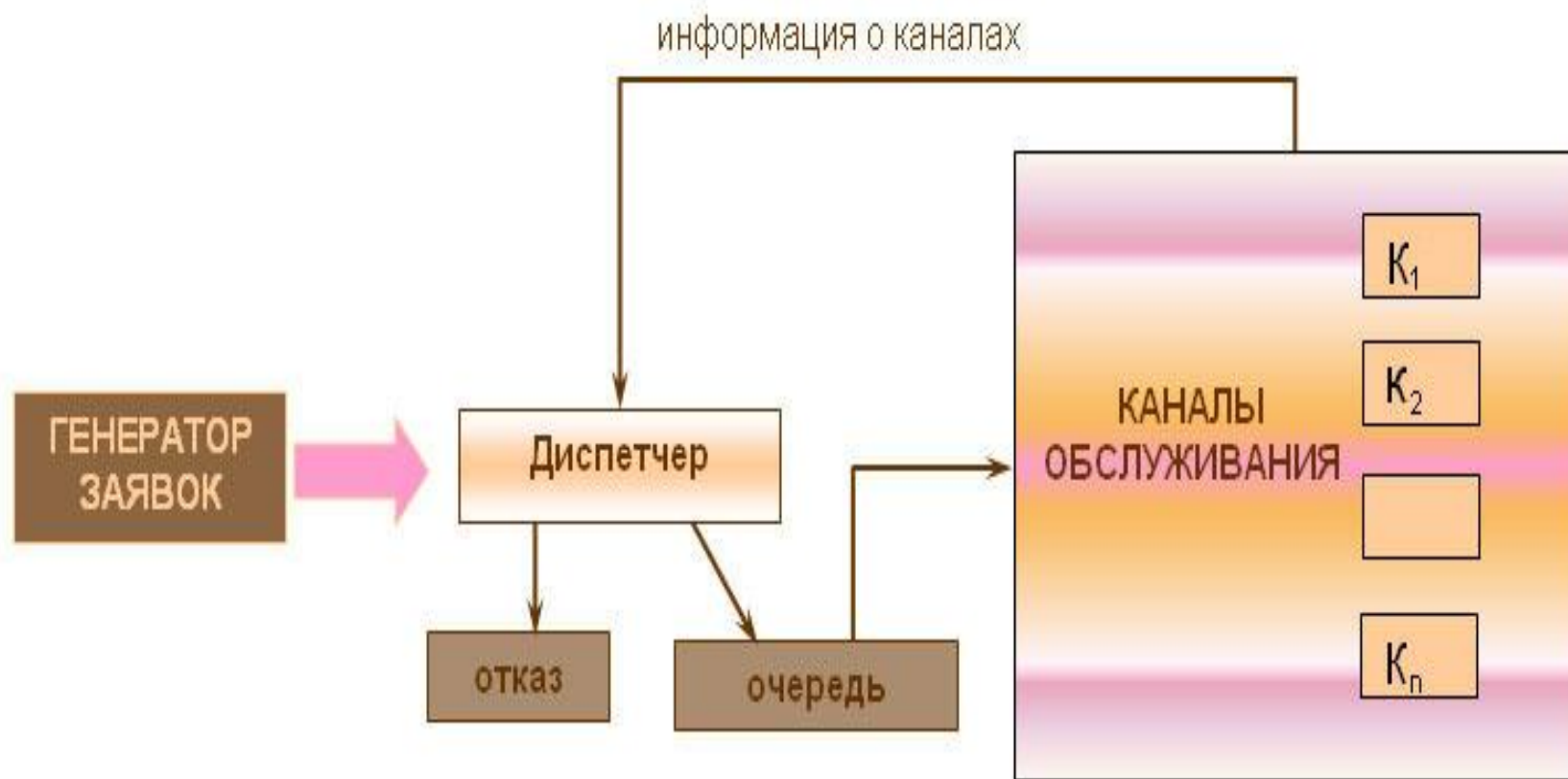
Цель теории массового обслуживания - выработка рекомендаций по рациональному построению СМО, рациональной организации их работы и регулированию потока заявок для обеспечения высокой эффективности функционирования СМО.

Примеры СМО

- Обслуживание покупателей в сфере розничной торговли
- Транспортное обслуживание
- Медицинское обслуживание населения
- Ремонт аппаратуры, машин, механизмов, находящихся в эксплуатации
- Обработка документов в системе управления
- Туристическое обслуживание

СМО	Узлы	Требования
Банк	Кассы	Клиенты
Больница	Врачи Санитары Койки	Пациенты
Производство	Станки Рабочие	Детали
Аэропорт	Выходы на взлетно-посадочные полосы Пункты регистрации	Пассажиры

Схема работы СМО



Генератор заявок – объект, порождающий заявки: улица, цех с установленными агрегатами. На вход поступает поток заявок.

Диспетчер – человек или устройство, которое знает, что делать с заявкой. Узел, регулирующий и направляющий заявки к каналам обслуживания. Диспетчер:

- принимает заявки;

- формирует очередь, если все каналы заняты;

- направляет их к каналам обслуживания, если есть свободные;

- дает заявкам отказ (по различным причинам);

- принимает информацию от узла обслуживания о свободных каналах;

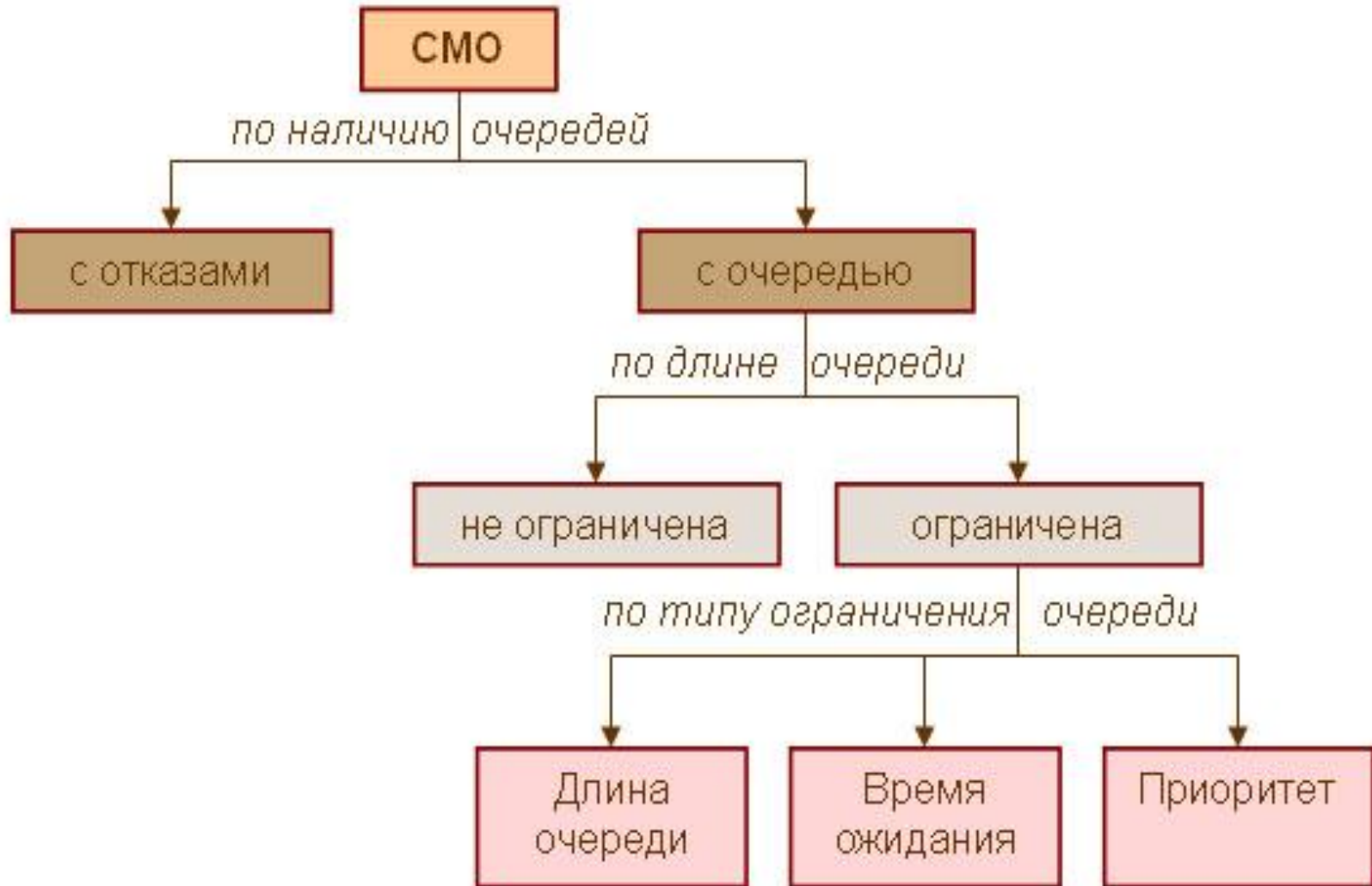
- следит за временем работы системы.

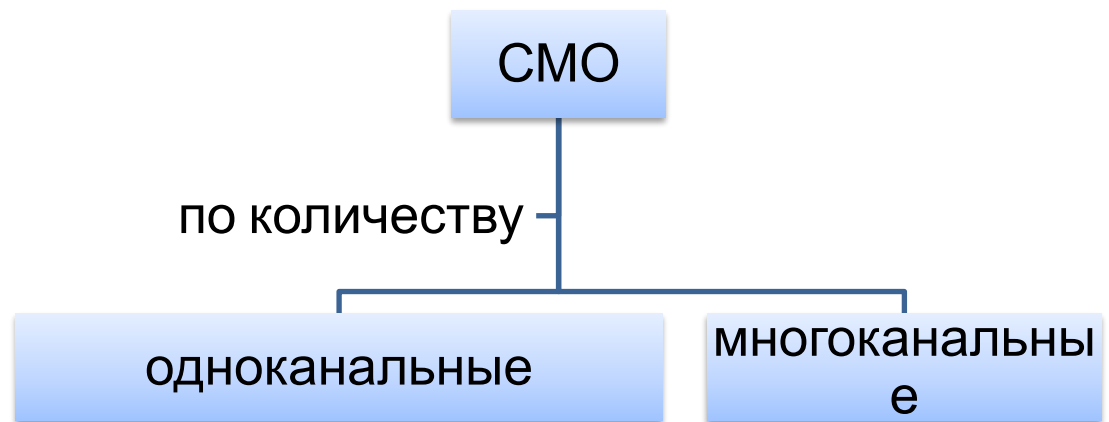
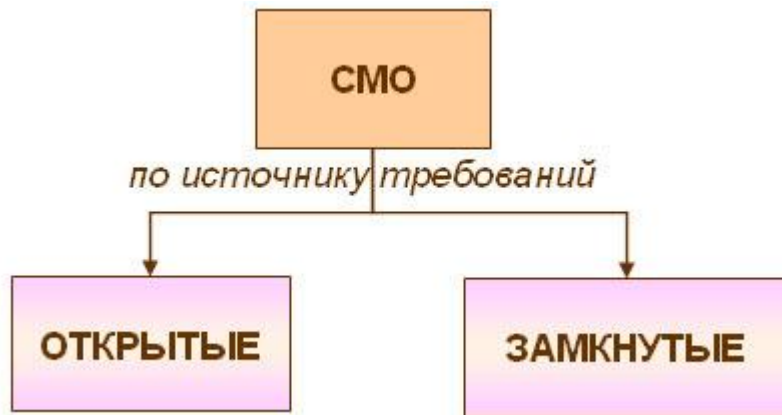
Очередь – накопитель заявок. Очередь может отсутствовать.

Узел обслуживания состоит из конечного числа каналов обслуживания. Каждый канал имеет 3 состояния: свободен, занят, не работает. Если все каналы заняты, то можно придумать стратегию, кому передавать заявку.

Отказ от обслуживания наступает, если все каналы заняты (некоторые в том числе могут не работать).

2 Классификация СМО





3 Характеристики СМО

Основными характеристиками системы массового обслуживания любого вида являются:

- входной поток поступающих требований или заявок на обслуживание;
- дисциплина очереди;
- механизм обслуживания.

Входной поток

Пусть:

A_i – время поступления между требованиями – независимые одинаково распределенные случайные величины;

$E(A)$ – среднее (МО) время поступления;

$\lambda = 1/E(A)$ – интенсивность поступления требований;

Характеристики входного потока:

Вероятностный закон, определяющий последовательность моментов поступления требований на обслуживание.

Количество требований в каждом очередном поступлении для групповых потоков.

Классическая теория массового обслуживания рассматривает так называемый пуассоновский (простейший) поток требований. Для этого потока число требований k для любого интервала времени распределено по закону Пуассона:

$$P_k(t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, k \geq 0, t \geq 0$$

где λ - интенсивность потока требований (число требований за единицу времени).

Дисциплина очереди

Очередь – совокупность требований, ожидающих обслуживания

Дисциплина очереди определяет принцип, в соответствии с которым поступающие на вход обслуживающей системы требования подключаются из очереди к процедуре обслуживания.

Чаще всего используются дисциплины очереди, определяемые следующими правилами:

- **первым пришел – первый обслуживаешься;**
(first in first out (FIFO))
- **пришел последним — обслуживаешься первым**
(LIFO)
- **случайный отбор заявок;**
- **приоритет:** некоторые находящиеся в очереди

Характеристики очереди

- ограничение времени ожидания момента наступления обслуживания (имеет место очередь с ограниченным временем ожидания обслуживания, что ассоциируется с понятием «допустимая длина очереди»);
- длина очереди.

Механизм обслуживания

- Механизм обслуживания** определяется характеристиками самой процедуры обслуживания и структурой обслуживающей системы. К характеристикам процедуры обслуживания относятся:
- ♦ количество каналов обслуживания (N);
 - ♦ продолжительность процедуры обслуживания (вероятностное распределение времени обслуживания требований);
 - ♦ количество требований, удовлетворяемых в результате выполнения каждой такой процедуры (для групповых заявок);
 - ♦ вероятность выхода из строя обслуживающего канала;
 - ♦ структура обслуживающей системы.

Пусть:

S_i – время обслуживания i -го требования;

$E(S)$ – среднее время обслуживания;

$\mu=1/E(S)$ – скорость обслуживания требований.

$N\mu$ – скорость обслуживания в системе, когда заняты все устройства обслуживания.

$\rho=\lambda/(N\mu)$ – называется **коэффициентом использования СМО**, показывает, насколько задействованы ресурсы системы.

T – среднее время пребывания требований в системе

$N=\lambda T$ – среднее количество требований в СМО (закон Литтла)

4 Структура обслуживающей системы

Структура обслуживающей системы определяется количеством и взаимным расположением каналов обслуживания

- ❖ С одним устройством обслуживания
- ❖ Параллельное обслуживание (многоканальные системы)
- ❖ Комбинированное обслуживание

Системы с одним устройством обслуживания

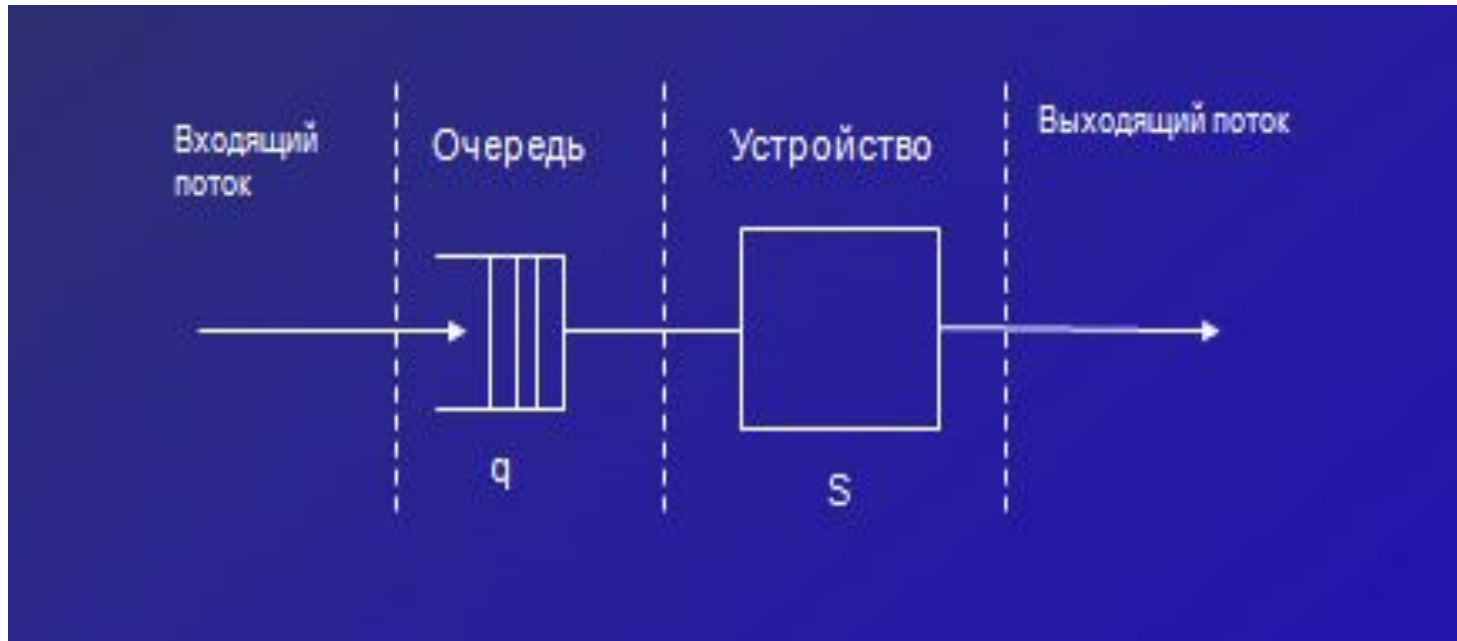


Рисунок 1 - Одноканальная СМО

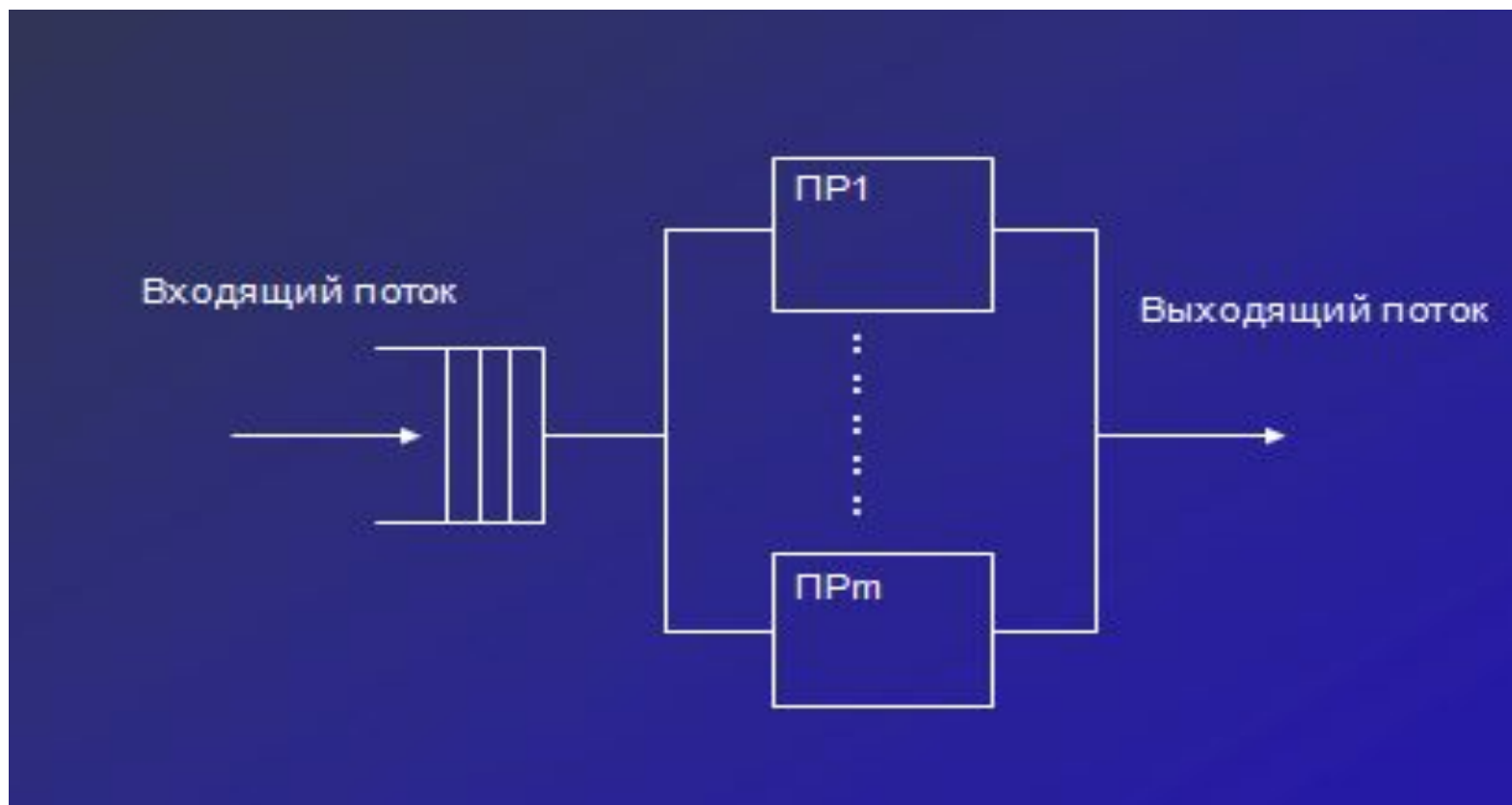


Рисунок 2 – Многоканальное обслуживание

**Функциональные возможности любой системы
массового обслуживания определяются
следующими основными факторами:**

- вероятностным распределением моментов поступлений заявок на обслуживание (единичных или групповых);
- мощностью источника требований;
- вероятностным распределением времени продолжительности обслуживания;
- конфигурацией обслуживающей системы (параллельное, последовательное или параллельно-последовательное обслуживание);
- количеством и производительностью обслуживающих каналов;
- дисциплиной очереди.

5 Основные критерии эффективности функционирования СМО

Судить о результатах работы СМО можно по **показателям**.

- ❖ вероятность обслуживания клиента системой;
- ❖ пропускная способность системы;
- ❖ вероятность отказа клиенту в обслуживании;
- ❖ вероятность занятости каждого из канала и всех вместе;
- ❖ среднее время занятости каждого канала;
- ❖ вероятность занятости всех каналов;
- ❖ среднее количество занятых каналов;
- ❖ вероятность простоя каждого канала;
- ❖ вероятность простоя всей системы;
- ❖ среднее количество заявок, стоящих в очереди;
- ❖ среднее время ожидания заявки в очереди;
- ❖ среднее время обслуживания заявки;
- ❖ среднее время нахождения заявки в системе.

Вероятность немедленного обслуживания
поступившей заявки ($P_{\text{обсл}} = K_{\text{обс}} / K_{\text{пост}}$);

Вероятность отказа в обслуживании
поступившей заявки ($P_{\text{отк}} = K_{\text{отк}} / K_{\text{пост}}$);

Очевидно, что $P_{\text{обсл}} + P_{\text{отк}} = 1$.

Задержка – один из критериев обслуживания СМО, время проведенное заявкой в ожидании обслуживания.

Пусть:

D_i – задержка в очереди требования i ;

$W_i = D_i + S_i$ – время нахождения в системе требования i .

Установившаяся средняя задержка очереди

$$d = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n D_i}{n} \quad \text{I В}$$

Установившееся среднее время на требования в СМО

$$w = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n W_i}{n}$$

Пусть:

$Q(t)$ – число требований в очереди в момент времени t ;

$L(t)$ – число требований в системе в момент времени t (плюс число требований, которые находятся на обслуживании в момент времени t).

Тогда рассчитывают показатели (если существуют)

Установившееся среднее по времени число требований в очереди;

$$Q = \lim_{T \rightarrow \infty} \frac{\int_0^T Q(t) dt}{T}$$

Установившееся среднее по времени число требований в системе.

$$L = \lim_{T \rightarrow \infty} \frac{\int_0^T L(t) dt}{T}$$

Заметим, что $\rho < 1$ – обязательное условие существования d , w , Q и L в системе массового обслуживания.

К наиболее общим и нужным результатам для систем массового обслуживания относятся уравнения сохранения

$$Q = \lambda d$$

$$L = \lambda w$$

Кроме этого можно также говорить о таких характеристиках, как:

абсолютная пропускная способность

системы – $A = P_{\text{обсл}} * \lambda$;

относительная пропускная способность

системы – $Q = \mu / (\mu + \lambda)$

Средняя задержка в очереди для системы массового обслуживания

$$d = \frac{\lambda \{ \sigma(S)^2 + E(S)^2 \}}{2[1 - \lambda E(S)]}$$

В России эта формула известна как формула Поллачека–Хинчина, за рубежом эта формула связывается с именем Росса (Ross).

- Для моделирования систем массового обслуживания важно знать характер потока заявок.
- Для многих потоков в справочниках можно найти формулы для расчёта характеристик СМО.
- Ошибка в определении характера потока заявок приводит к ошибке в оценке параметров СМО и, как следствие, либо к избыточным вложениям в их создание, либо к неработоспособности СМО.

Поток Пуассона

Стационарный ординарный поток, в котором длительность промежутков времени между возникновением транзактов является независимой случайной величиной



Поток Эрланга порядка k

Продолжительность промежутков между возникновением транзактов представляет собой сумму k независимых случайных величин, каждая из которых распределена по экспоненциальному закону



Поток Пуассона (простейший)

Характеризуется:

- экспоненциальным (показательным) распределением продолжительности промежутков между возникновением транзактов
- пуассоновским распределением вероятности возникновения n транзактов за период t

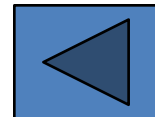
Плотность распределения интервала
времени между возникновением двух
транзактов в потоке Эрланга

$$f(\tau) = \frac{\lambda(\lambda\tau)^{k-1}}{(k-1)!} e^{-\lambda\tau}$$

λ среднее число заявок в единицу времени и

k порядок потока Эрланга

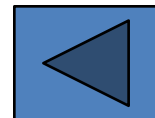
τ длительность промежутка времени



Дискретное распределение Пуассона

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

λ среднее число заявок в единицу времени и
 k число транзактов, возникших в течение
промежутка времени τ



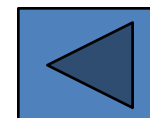
Экспоненциальное распределение

$$F(\tau) = \frac{1}{t} \cdot e^{-\frac{1}{t} \cdot \tau}$$

t средняя продолжительность времени между транзактами

τ заданный промежуток времени между транзактами

F вероятность того, что промежуток времени между транзактами не превысит величины τ



6 Характеристики основных моделей СМО

• 6.1 СМО с отказами

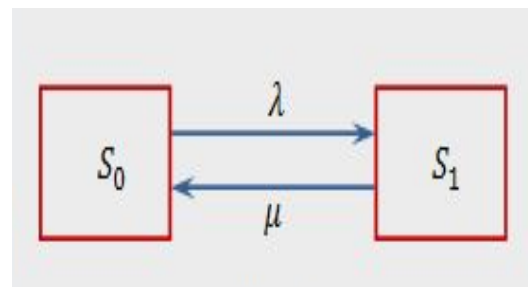
В качестве показателей эффективности СМО с отказами будем рассматривать:

- 1) A — **абсолютную пропускную способность СМО**, т.е. среднее число заявок, обслуживаемых в единицу времени;
- 2) Q — **относительную пропускную способность**, т.е. среднюю долю пришедших заявок, обслуживаемых системой;
- 3) $P_{отк}$ — **вероятность отказа**, т.е. того, что заявка покинет СМО необслуженной;
- 4) k — **среднее число занятых каналов** (для многоканальной системы).

Одноканальная система (СМО) с отказами

Рассмотрим задачу. Имеется один канал, на который поступает поток заявок с интенсивностью λ . Поток обслуживания имеет интенсивность μ . Найти предельные вероятности состояний системы и показатели ее эффективности.

Система (СМО) имеет два состояния: S_0 — канал свободен, S_1 — канал занят. Размеченный граф состояний представлен на рис.



В предельном, стационарном режиме система алгебраических уравнений для вероятностей состояний имеет вид

$$\begin{cases} \lambda \cdot p_0 = \mu \cdot p_1, \\ \mu \cdot p_1 = \lambda \cdot p_0, \end{cases}$$

- т.е. система вырождается в одно уравнение. Учитывая нормировочное условие $p_0 + p_1 = 1$, найдем из системы предельные вероятности состояний

$$p_0 = \frac{\mu}{\lambda + \mu}, \quad p_1 = \frac{\lambda}{\lambda + \mu},$$

Где

p_0 – вероятность того, что заявка будет обслужена.

Нетрудно убедиться, что для одноканальной СМО с отказами вероятность $P_0(t)$ есть не что иное, как **относительная пропускная способность** системы (Q).

Действительно, $P_0(t)$ – вероятность того, что в момент t канал свободен и заявка, пришедшая к моменту tt , будет обслужена, а следовательно, для данного момента времени t среднее отношение числа обслуженных заявок к числу поступивших также равно $P_0(t)$.

P_1 - Вероятность того, что в обслуживании будет отказано ($P_{отк}$).

Абсолютная пропускная способность

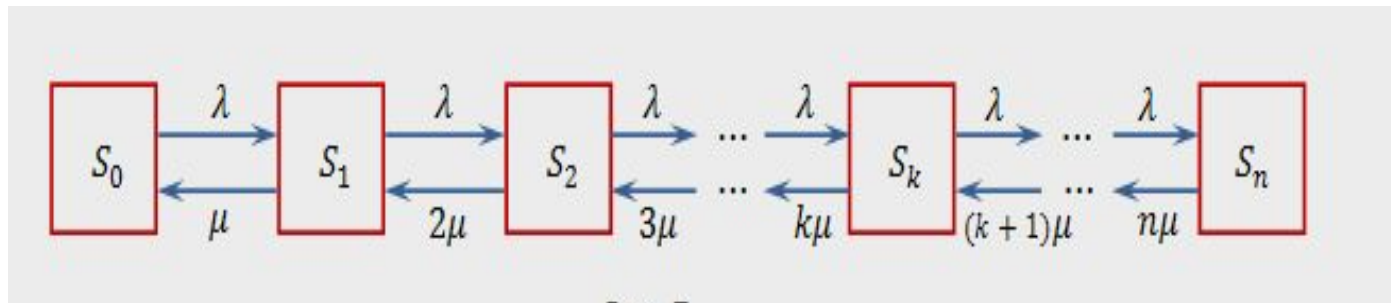
$$A = \lambda \cdot p_0 = \frac{\lambda \cdot \mu}{\lambda + \mu}$$

Многоканальная система (СМО) с отказами

Рассмотрим классическую **задачу Эрланга**. Имеется n каналов, на которые поступает поток заявок с интенсивностью λ . Поток обслуживания имеет интенсивность μ . Найти предельные вероятности состояний системы и показатели ее эффективности.

Система S (СМО) имеет следующие состояния $S_1, \dots, S_k, \dots, S_n$ (нумеруем их по числу заявок, находящихся в системе) где S_k — состояние системы, когда в ней находится k заявок, т.е. занято k каналов.

Граф состояний СМО соответствует процессу гибели и размножения и показан на рис.



Поток заявок последовательно переводит систему из любого левого состояния в соседнее правое с одной и той же интенсивностью λ . Интенсивность же потока обслуживания, переводящих систему из любого правого состояния в соседнее левое состояние, постоянно меняется в зависимости от состояния.

Действительно, если СМО находится в состоянии S_2 (два канала заняты), то она может перейти в состояние S_1 (один канал занят), когда закончит обслуживание либо первый, либо второй канал, т.е. суммарная интенсивность их потоков обслуживания будет 2μ . Аналогично суммарный поток обслуживания, переводящий СМО из состояния S_3 (три канала заняты) в S_2 , будет иметь интенсивность 3μ , т.е. может освободиться любой из трех каналов и т.д.

Предельная вероятность состояния

$$p_0 = \left(1 + \frac{\lambda}{\mu} + \frac{\lambda^2}{2!\mu^2} + \dots + \frac{\lambda^k}{k!\mu^k} + \dots + \frac{\lambda^n}{n!\mu^n} \right)^{-1},$$

$$p_1 = \frac{\lambda}{\mu} p_0$$

– вероятность, что занят один канал обслуживания;

$$p_2 = \frac{\lambda^2}{2\mu^2} p_0$$

– вероятность, что заняты два канала обслуживания;

и т. д.

Величина

$$\rho = \frac{\lambda}{\mu}$$

называется ***приведенной интенсивностью потока заявок*** или ***интенсивностью нагрузки канала***. Она выражает среднее число заявок, приходящее за среднее время обслуживания одной заявки.

Теперь

$$p_0 = \left(1 + \rho + \frac{\rho^2}{2!} + \dots + \frac{\rho^k}{k!} + \dots + \frac{\rho^n}{n!} \right)^{-1},$$

$$p_1 = \rho \cdot p_0, \quad p_2 = \frac{\rho^2}{2!} \cdot p_0, \quad \dots, \quad p_k = \frac{\rho^k}{k!} \cdot p_0, \quad \dots, \quad p_n = \frac{\rho^n}{n!} \cdot p_0.$$

Формулы для предельных вероятностей получили названия **формул Эрланга** в честь основателя теории массового обслуживания.

Вероятность отказа СМО есть предельная вероятность того, что все n каналов системы будут заняты, т.е.

$$P_{\text{отк}} = \frac{\rho^n}{n!} \cdot p_0.$$

Относительная пропускная способность — вероятность того, что заявка будет обслужена:

$$Q = 1 - P_{\text{отк}} = 1 - \frac{\rho^n}{n!} \cdot p_0.$$

Абсолютная пропускная способность:

$$A = \lambda \cdot Q = \lambda \cdot \left(1 - \frac{\rho^n}{n!} \cdot p_0 \right).$$

Среднее число занятых каналов $\bar{k}$ есть математическое ожидание числа занятых каналов:

$$\bar{k} = \sum_{k=0}^n (k \cdot p_k),$$

Однако среднее число занятых каналов можно найти проще, если учесть, что абсолютная пропускная способность системы есть не что иное, как интенсивность **потока обслуженных** системой заявок (в единицу времени). Так как каждый занятый канал обслуживает в среднем заявок (в единицу времени), то среднее число занятых каналов

$$\bar{k} = \frac{A}{\mu}$$

или:

$$\bar{k} = \rho \cdot \left(1 - \frac{\rho^n}{n!} \cdot p_0 \right).$$

6.2 СМО с ожиданием (очередью)

В качестве показателей эффективности СМО с ожиданием, кроме уже известных показателей — **A** абсолютной и **Q** относительной пропускной способности, **P_{отк}** вероятности отказа, среднего числа занятых каналов **k** (для многоканальной системы) будем рассматривать также следующие:

- 1) среднее число заявок в системе;
- 2) среднее время пребывания заявки в системе;
- 3) среднее число заявок в очереди (длина очереди);
- 4) среднее время пребывания заявки в очереди;
- 5) вероятность того, что канал занят (степень загрузки канала).

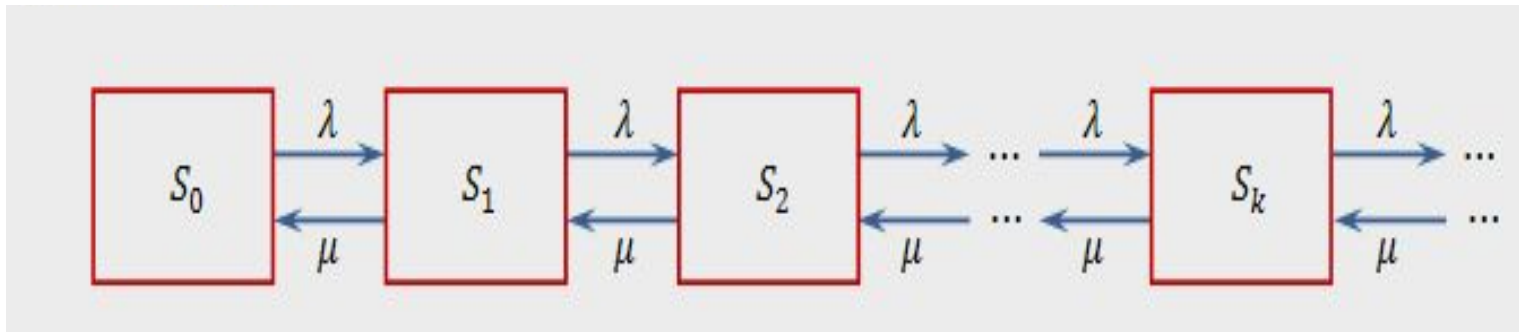
Одноканальная система с неограниченной очередью

Рассмотрим задачу.

Имеется одноканальная СМО с очередью, на которую не наложены никакие ограничения (ни по длине очереди, ни по времени ожидания). Поток заявок, поступающих в СМО, имеет интенсивность λ , а поток обслуживания — интенсивность μ . Необходимо найти предельные вероятности состояний и показатели эффективности СМО.

Система может находиться в одном из состояний $S_1, \dots, S_k$, по числу заявок, находящихся в СМО: S_0 — канал свободен; S_1 — канал занят (обслуживает заявку), очереди нет; S_2 — канал занят, одна заявка стоит в очереди; S_k — канал занят, $k-1$ заявок стоят в очереди и т.д.

Граф состояний СМО представлен на рис.



Предельные вероятности состояний:

$$p_0 = 1 - \rho,$$

$$p_1 = \rho(1 - \rho), \quad p_2 = \rho^2(1 - \rho), \quad \dots, \quad p_k = \rho^k(1 - \rho), \quad \dots$$

Предельные вероятности образуют убывающую геометрическую прогрессию со знаменателем $\rho < 1$, следовательно, вероятность p_0 — наибольшая. Это означает, что если СМО справляется с потоком заявок (при $\rho < 1$), то наиболее вероятным будет отсутствие заявок в системе.

Среднее число заявок в системе определим по формуле математического ожидания:

$$L_{\text{сист.}} = \sum_{k=1}^{\infty} k p_k = (1 - \rho) \sum_{k=1}^{\infty} k \rho^k$$

Или (при $\rho < 1$)

$$L_{\text{сист.}} = \frac{\rho}{1 - \rho}.$$

Среднее число заявок в очереди .

$$L_{\text{оч.}} = L_{\text{сист.}} - L_{\text{об.}}$$

где $L_{\text{об.}}$ — среднее число заявок, находящихся под обслуживанием

$$L_{\text{об.}} = P_{\text{зан.}} = \rho.$$

Тогда

$$L_{\text{och.}} = \frac{\rho^2}{1 - \rho}.$$

Доказано, что при любом характере потока заявок, при любом распределении времени обслуживания, при любой дисциплине обслуживания среднее время пребывания заявки в системе (очереди) равно среднему числу заявок в системе (в очереди), деленному на интенсивность потока заявок, т.е.

$$T_{\text{sist.}} = \frac{1}{\lambda} \cdot L_{\text{sist.}},$$

$$T_{\text{och.}} = \frac{1}{\lambda} \cdot L_{\text{och.}}.$$

Или

$$T_{\text{sist.}} = \frac{\rho}{\lambda(1 - \rho)},$$

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

Многоканальная СМО с неограниченной очередью

λ – среднее число транзактов, поступающих за единицу времени

t – среднее время обслуживания транзакта

$\mu = 1/t$ – среднее число транзактов, обслуживаемых за единицу времени

$\alpha = \lambda/\mu$ – среднее число занятых каналов

n – число каналов

- Вероятность того, что все n каналов свободны

$$P_0 = \frac{1}{\left(\sum_{k=0}^{n-1} \frac{\alpha^k}{k!} \right) + \frac{\alpha^n}{n!(1 - \alpha/n)}}$$

λ – среднее число транзактов, поступающих за единицу времени

t – среднее время обслуживания транзакта

$\mu = 1/t$ – среднее число транзактов, обслуживаемых за единицу времени

$\alpha = \lambda/\mu$ – среднее число занятых каналов

n – число каналов

- Вероятность того, что свободно $n-k$ каналов

$$P_k = \frac{\alpha^k}{k!} P_0, \quad 1 \leq k \leq n$$

- Вероятность наличия очереди из $k - n$ заявок

$$P_k = \frac{\alpha^k}{n! \cdot n^{k-n}} P_0, \quad k \geq n$$

λ – среднее число транзактов, поступающих за единицу времени

t – среднее время обслуживания транзакта

$\mu = 1/t$ – среднее число транзактов, обслуживаемых за единицу времени

$\alpha = \lambda/\mu$ – среднее число занятых каналов

n – число каналов

- Вероятность наличия очереди

$$P_Q = \frac{\alpha^{n+1}}{n!(n-\alpha)} P_0, \quad \alpha < n$$

■ Средняя длина очереди

$$L_Q = \frac{\alpha}{n-\alpha} P_k, \quad k = n, \alpha < n$$

λ – среднее число транзактов, поступающих за единицу времени

t – среднее время обслуживания транзакта

$\mu = 1/t$ – среднее число транзактов, обслуживаемых за единицу времени

$\alpha = \lambda/\mu$ – среднее число занятых каналов

n – число каналов

- Среднее время ожидания в очереди

$$t_Q = \frac{L_Q}{\lambda}$$

- Коэффициент простоя каналов

$$K_S = 1 - \alpha / n$$

- Необходимое условие работоспособности СМО

$$\alpha < n$$