Лекция 5
Математические модели и методы теории систем массового обслуживания, используемые в САПР КЭС

Вопросы лекции

1. Предметная область теории систем массового обслуживания.
2. Классификация систем массового обслуживания.
3. Методы анализа и синтеза систем массового обслуживания.

Вопрос 1.
Предметная область теории систем массового обслуживания

Система массового обслуживания (СМО) — система, которая производит обслуживание поступающих в неё требований.

Теория массового обслуживания (теория очередей) — раздел теории вероятностей, целью исследований которого является рациональный выбор структуры системы обслуживания и процесса обслуживания на основе изучения потоков требований на обслуживание, поступающих в систему и выходящие из неё, длительности ожидания и длины очередей.

Теория телетрафика - математическая теория, являющаяся одной из ветвей теории массового обслуживания. Применяется, прежде всего, для изучения и проектирования систем телекоммуникаций. Однако, разрабатываемые средства теории телетрафика являются независимыми от конкретной техники, и могут использоваться в области дорожного (авто) и воздушного (авиа) трафика, на производстве, при хранении и распределении готовых товаров, в общем, во всех системах

Основы теории телетрафика были заложены в работах **А. К. Эрланга** по исследованию пропускной способности полнодоступного пучка линий, обслуживающего простейший поток вызовов с потерями и с ожиданием. Труды А. К. Эрланга послужили толчком для других работ, которые были связаны с подтверждением, развитием или опровержением его результатов.

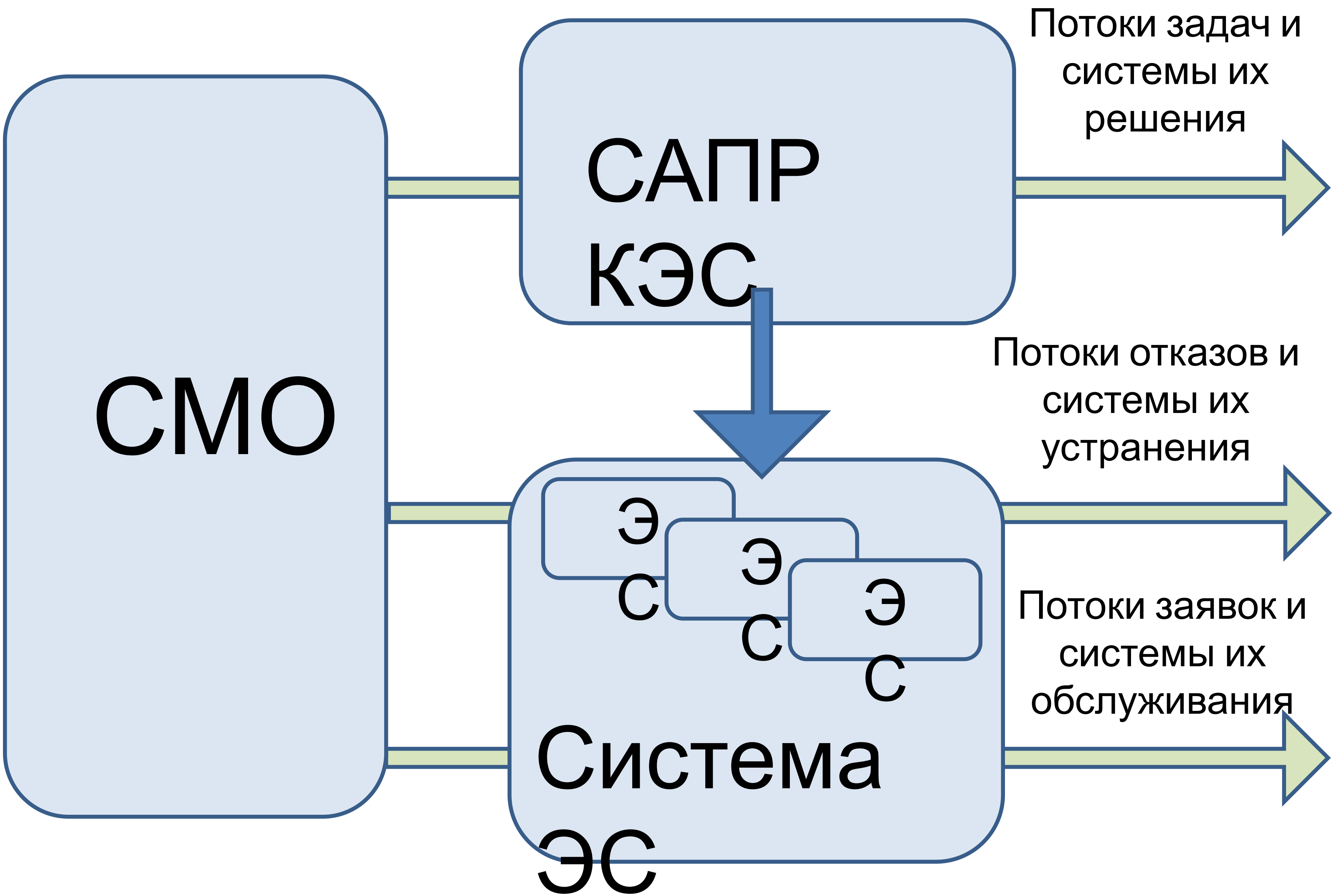
В 1918 году **Т.Энгсет** обобщил результаты А. К. Эрланга на случай обслуживания полнодоступным пучком потока вызовов от конечного числа источников нагрузки.

В 1933 году советский математик **А. Н. Колмогоров** выполнил свою классическую работу по аксиоматическому обоснованию теории вероятностей, в которой идеи А. К. Эрланга были увязаны с марковскими случайными процессами.

В этот же период появились первые работы **А. Я. Хинчина** по исследованию систем массового обслуживания с ожиданием.

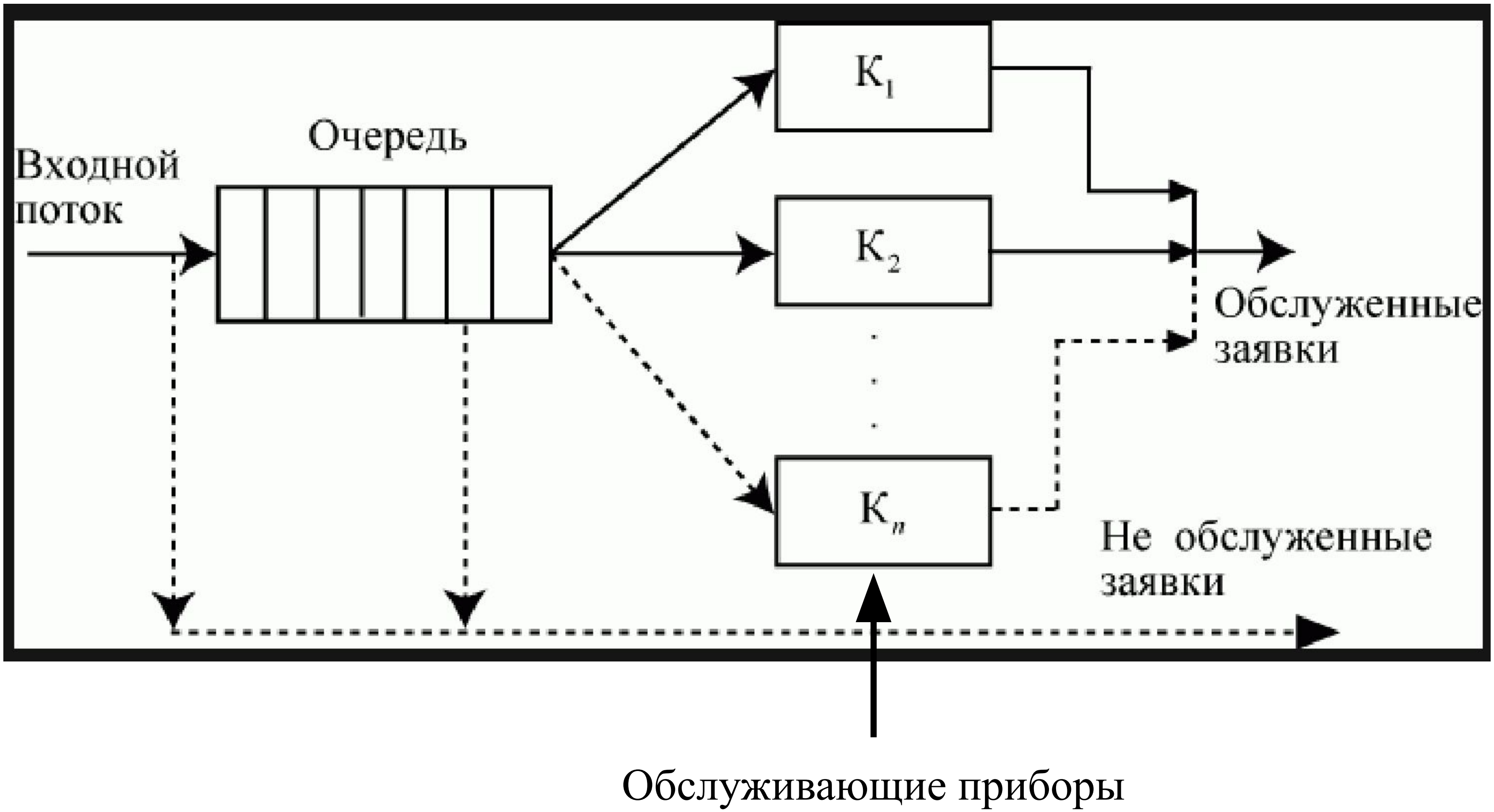
В 1943 году шведский ученый **К. Пальм** обобщил результаты А. К. Эрланга на случай обслуживания потока с ограниченным последействием, и получил важные результаты по изучению колебания телефонной нагрузки.

В 1964 году американский ученый **Л. Клейнрок** разработал основные принципы пакетной коммуникации, которые легли в основу современной


СМО
САПР КЭС
Потоки задач и системы их решения
Потоки отказов и системы их устранения
Потоки заявок и системы их обслуживания
ЭС
ЭС
ЭС
Система ЭС

Сеть связи – как система (сеть) массового обслуживания

Элементы типовой модели системы массового обслуживания



Модель потока дискретных событий в непрерывном времени

$\lambda = 1 / \tau_{\text{вх}}$

Интенсивность входного потока

$\mu = 1 / \tau_{\text{обс}}$

Интенсивность обслуживания

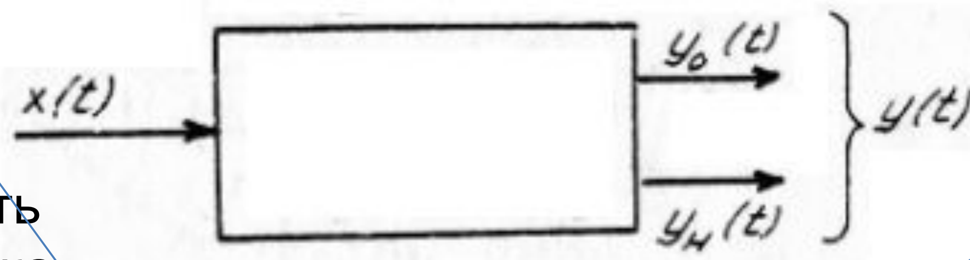


Рис.1

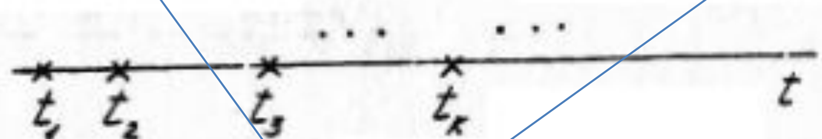


Рис.2

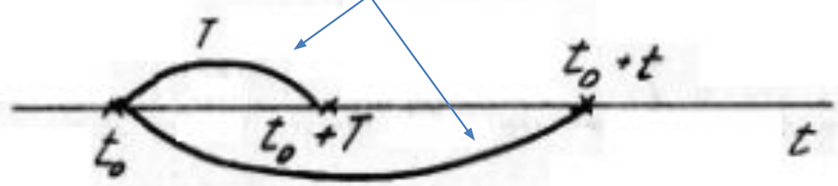


Рис.3

Вопрос 2

Классификация систем массового обслуживания

Классификация систем массового обслуживания

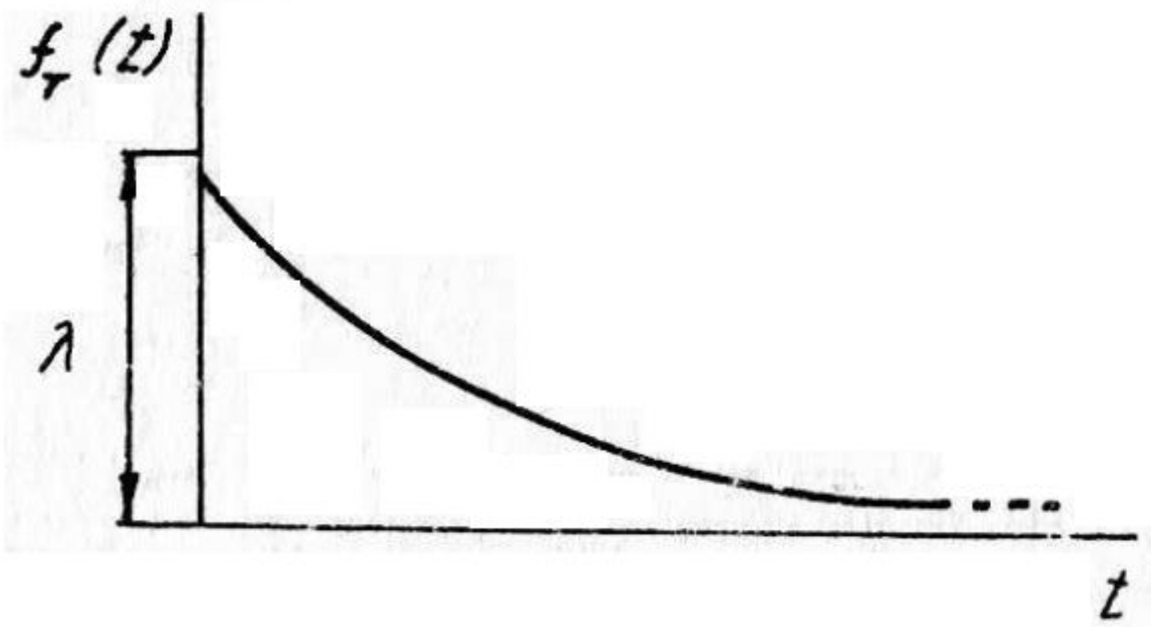
При моделировании СМО к наиболее важным свойствами потоков заявок относят следующие.

1. **<u>Стационарность</u>**. Поток событий считается стационарным, если вероятность попадания определенного числа событий на интервал времени длиной т зависит только от длины этого интервала и не зависит от того, в каком месте на временной оси расположен этот участок.
2. **<u>Отсутствие последействия</u>**. Поток событий называется потоком без последействия, если для любых неперекрывающихся интервалов времени число событий, попадающих на один временной отрезок, не зависит от числа событий, попадающих на другие отрезки времени.
3. **<u>Ординарность</u>**. Поток событий называется ординарным, если вероятность попадания двух или более событий на элементарный (малый) отрезок времени Δt пренебрежимо мала по сравнению с вероятностью попадания на этот временной интервал одного события.

Поток событий, обладающий тремя перечисленными свойствами, называется **<u>простейшим</u>**, или стационарным **<u>пуассоновским</u>**, поскольку для простейшего потока число событий, попадающих на любой фиксированный интервал времени, распределено по закону Пуассона.

$$p_k(\tau) = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau} \qquad F_T(t) = 1 - e^{-\lambda t} \quad (t > 0)$$

$$f_T(t) = \frac{\mathrm{d}F_T(t)}{\mathrm{d}t} = \lambda\, \mathrm{e}^{-\lambda t}$$



В теории массового обслуживания используют и другие модели потоков Пальма, Эрланга и т. д.

Ординарный поток однородных событий называется потоком с ограниченным последействием (потоком Пальма), если промежутки времени между последовательными событиями независимы. Простейший поток является частным случаем потока Пальма, когда независимые промежутки времени между соседними заявками распределены по показательному закону.

К потокам с ограниченным последействием относятся также потоки Эрланга, образуемые «просеиванием» простейшего потока. Если из простейшего потока исключить каждое второе требование, то оставшиеся требования образуют поток Эрланга первого порядка. Поток Эрланга k-го порядка получается при сохранении в исходном простейшем потоке каждого $(k+1)$-го требования и исключении остальных. Для потока Эрланга k-го порядка плотность распределения интервала между соседними заявками имеет вид

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

Поток Эрланга любого порядка представляет собой частный случай потока Пальма.

Реальные потоки, образуемые суммированием потоков от многих источников, оказываются близкими к простейшему.

Марковские модели систем массового обслуживания

Случайный процесс называется марковским, если его поведение в будущем определяется только настоящим состоянием и не зависит от того, когда и каким именно образом система перешла в текущее состояние.

Если динамика системы при выполнении ею своих задач может быть представлена в виде случайного процесса смены состояний и при этом вероятностные характеристики перехода из одного состояния в другое не будут зависеть от развития процесса смены состояний в прошлом, то для моделирования такой системы может быть применен математический аппарат марковских случайных процессов

В зависимости от характера изменения состояний и времени все многообразие марковских процессов подразделяется на следующие четыре типа:

- марковские процессы с дискретными состояниями и дискретным временем, называемые дискретными марковскими цепями или цепями Маркова;
- марковские процессы с дискретными состояниями и непрерывным временем, называемые непрерывными марковскими цепями;
- марковские процессы с непрерывными состояниями и дискретным временем;
- марковские процессы с непрерывными состояниями и непрерывным временем.

В практике моделирования сложных систем нашли широкое применение два типа из перечисленных выше марковских процессов, а именно дискретные и непрерывные марковские цепи.

Модели на основе дискретных марковских цепей

Задача разработки модели на основе дискретной марковской цепи формулируется следующим образом

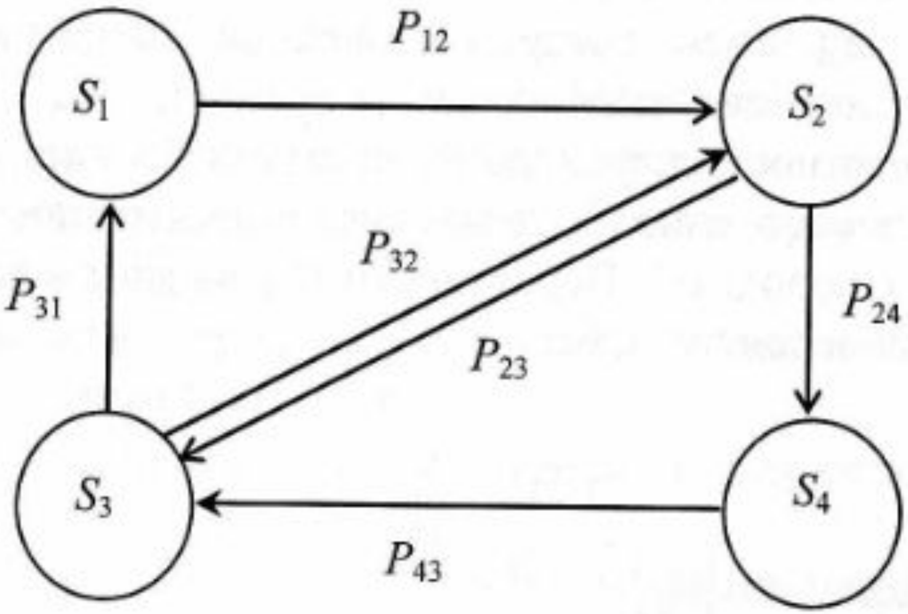
Пусть система S имеет n возможных состояний S_1, S_2, …, S_i, S_j, …, S_n, смена которых происходит мгновенно в определенные заранее известные моменты времени 1, 2,…, к,… K. Известны вероятности перехода P_{ij} системы S за один шаг из состояния S_i в состояние S_j, не изменяющиеся от шага к шагу (то есть цепь однородная). Вероятности P_{ij} заданы в виде матрицы переходных вероятностей:

$$\left[P_{ij}\right]=\begin{bmatrix} P_{11} & P_{12} & \dots & P_{1n} \\ P_{21} & P_{22} & \dots & P_{2n} \\ \dots & \dots & \dots & \dots \\ P_{n1} & P_{n2} & \dots & P_{nn} \end{bmatrix},$$

где для всех i, j справедливы утверждения:

$$P_{ij} = const;$$

$$\sum_{j=1}^{n} P_{ij} = 1.$$



Пример графа состояний и переходов системы

Математическая модель динамики рассматриваемой однородной марковской цепи получается на основе формулы полной вероятности в виде рекуррентного соотношения:

$$P_j(k) = \sum_{i=1}^{n} P_i(k-1) \cdot P_{ij} \ ,$$

Модели на основе непрерывных марковских цепей. Уравнения Колмогорова – Чепмена.

В непрерывных марковских цепях использование переходных вероятностей P_{ij} становится невозможным. Это объясняется тем, что вероятность перехода системы из одного состояния в другое в конкретный момент времени t для непрерывной случайной величины равна нулю. Поэтому вместо переходных вероятностей вводятся в рассмотрение плотности распределения вероятностей переходов λ_{ij}, задаваемые равенством :

$$\lambda_{ij} = \lim_{\Delta t \to 0} \frac{P_{ij}(\Delta t)}{\Delta t}, \qquad (1.10)$$

где $P_{ij}(\Delta t)$ – вероятность того, что система, пребывая в момент времени t в состоянии S_i, за время Δt перейдет в S_j.

С точностью до малых второго порядка из приведенной формулы (1.10) можно получить следующее выражение:

$$P_{ij}(\Delta t) \approx \lambda_{ij} \Delta t. \qquad (1.11)$$

Пример графа непрерывной марковской цепи с четырьмя состояниями S_1, S_2, S_3, S_4 приведен на рисунке 1.2.

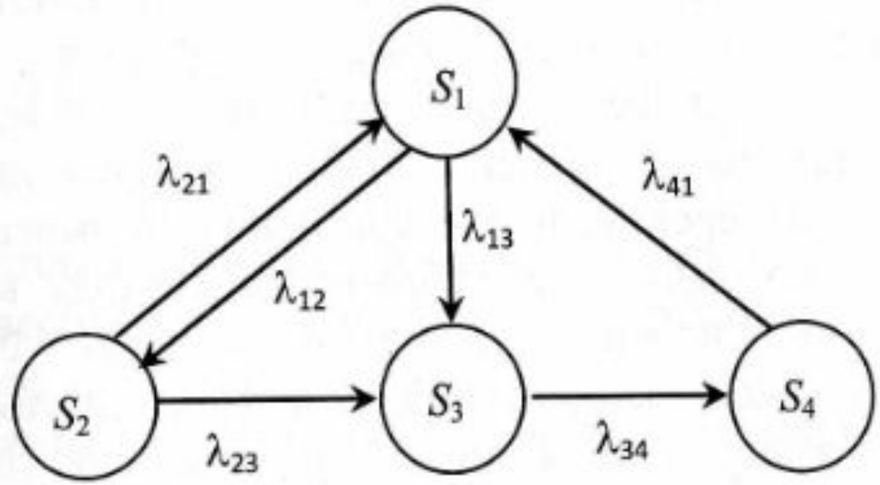


Рисунок 1.2 – Пример графа состояний непрерывной марковской цепи

Вероятности пребывания системы в любом j-м состоянии в момент t удовлетворяют уравнениям Колмогорова-Чепмена, которые являются обыкновенными дифференциальными уравнениями первого порядка с постоянными коэффициентами.

Составление уравнений Колмогорова-Чепмена производится с помощью следующего правила

Левая часть уравнения для каждого состояния представляет собой первую производную вероятности этого состояния по времени, а правая часть образуется сложением отрицательного произведения вероятности этого состояния на сумму интенсивностей всех возможных переходов из него в другие состояния и суммы произведений вероятностей всех предыдущих состояний на интенсивности перехода из них в рассматриваемое состояние. Общее число дифференциальных уравнений равно числу состояний системы.

Общий вид уравнений приведен ниже.

$$\frac{dP_j(t)}{dt} = -P_j(t)\sum_{k=1}^{n}\lambda_{jk} + \sum_{i=1}^{n}P_i(t)\lambda_{ij}, \quad j = 1,2,\ldots,n,$$

где $P_j(t)$ и $P_i(t)$ – вероятность j-го и предшествующего ему i-го состояния;

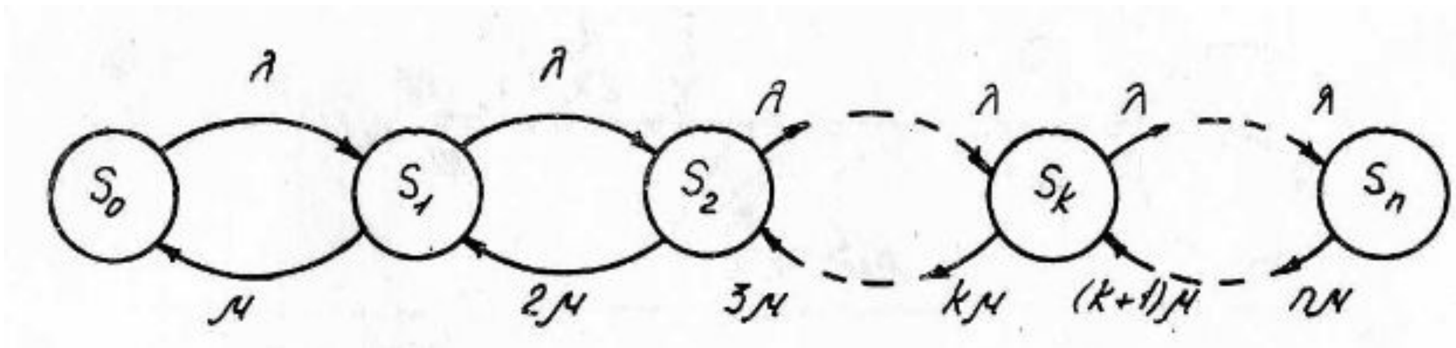
λ_{ij} и λ_{jk} – интенсивности потоков событий, переводящие систему из состояния i в состояние j и из состояния j в k соответственно.

Уравнения для определения предельных вероятностей нетрудно получить из системы уравнений Колмогорова-Чепмена, с учетом того, что в установившемся режиме вероятности P_j не изменяются. Тогда левые части уравнений Колмогорова - Чепмена обратятся в нуль. Модель однородной непрерывной марковской цепи будет представлена системой линейных алгебраических уравнений

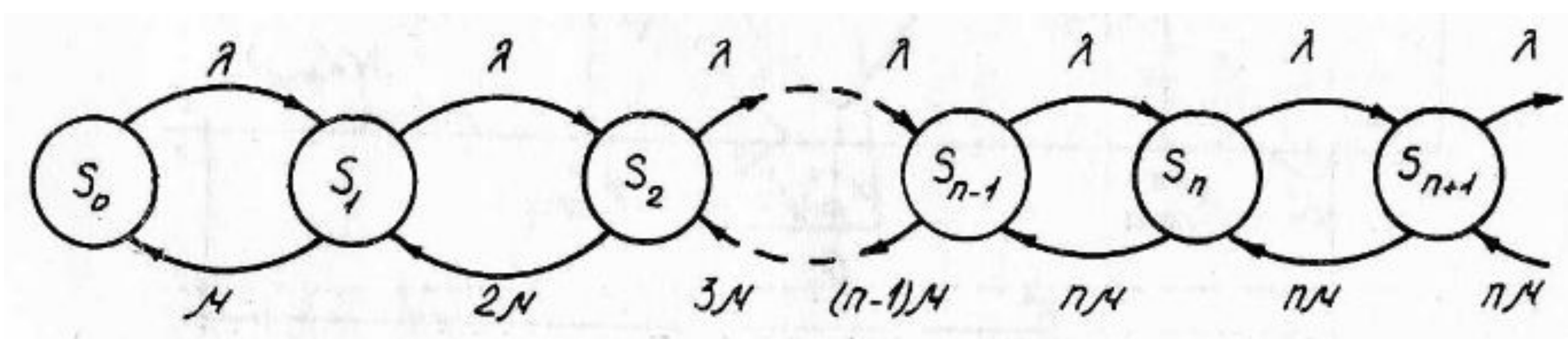
$$0 = -P_j \sum_{k=1}^{n} \lambda_{jk} + \sum_{i=1}^{n} P_i \lambda_{ij}, \quad j = 1,2,\ldots,n.$$

По известным значениям предельных вероятностей можно определить необходимые характеристики реальной системы в зависимости от интенсивностей λ_{ij}.

Пример марковской модели многоканальной СМО с отказами



Пример марковской модели многоканальной СМО с ожиданием



Одной из форм классификации систем массового обслуживания является кодовая (символьная) классификация **Д.Кендалла**.

При этой классификации характеристику системы записывают в виде трех, четырех или пяти символов, например A / B / S, где A — тип распределения входящего потока требований, B — тип распределения времени обслуживания, S — число каналов обслуживания.

Для экспоненциального распределения используют символ M, для любого (произвольного) распределения — символ G. Регулярный поток обозначают буквой D. Распределение Парето — символом P, самоподобный трафик — буквами fbm (фрактальное броуновское движение) и т.д.

Например, запись M / M / 3 означает, что входящий поток требований пуассоновский (простейший), время обслуживания распределено по экспоненциальному закону, в системе имеется три канала обслуживания.

Четвертый символ указывает допустимую длину очереди, а пятый — порядок отбора (приоритета) требований

<u>Вопрос 3</u>

Методы анализа и синтеза систем массового обслуживания

Основная цель теории телетрафика как одной из базовых ветвей теории массового обслуживания заключается **в разработке методов оценки качества** функционирования электронных систем распределения информации, т.е., построение математических моделей, более или менее адекватно отображающих реальные системы распределения и обработки информации, **что позволяет экономично проектировать системы и сети связи, а также их элементы (электронные средства)** при заданном качестве обслуживания.

Задачи теории телетрафика

анализ;
синтез;
оптимизация.

На первом месте стоят **задачи анализа**, т.е., отыскание зависимостей и значений величин, характеризующих качество обслуживания, от характеристик и параметров входящего потока вызовов, схемы и дисциплины обслуживания. Эти задачи в начальный период развития телефонной техники были более актуальными, чем задачи синтеза, и решались, как правило, с помощью теории вероятностей.

Развитие координатной, квазиэлектронной и электронной (цифровой) коммутационной техники поставило перед теорией телетрафика сложные вероятностно-комбинаторные **задачи синтеза**, в которых требуется определить структурные параметры коммутационных систем при заданных потоках, дисциплине и качестве обслуживания.

Близкими к задачам анализа и синтеза являются **задачи оптимизации**. Эти задачи при проектировании систем распределения информации формулируются следующим образом: определить такие значения структурных параметров коммутационной системы (алгоритмы функционирования), для которых:
при заданных потоках, качестве и дисциплине обслуживания стоимость или объем оборудования системы распределения информации минимальны и
при заданных потоках, дисциплине обслуживания и стоимости

$$p_k = \frac{\frac{a^k}{k!}}{\sum_{k=0}^{n} \frac{a^k}{k!}} \quad (0 \le k \le n) \qquad (1)$$

a = λ / μ

Формулы (1) называются формулами Эрланга. Они дают предельный закон распределения числа занятых каналов в зависимости от характеристик потока заявок и производительности системы обслуживания. Полагая в формуле (1') $k = n$, получим вероятность отказа (вероятность того, что поступившая заявка найдет все каналы занятыми):

$$P_{отк} = p_n = \frac{\frac{a^n}{n!}}{\sum_{k=0}^{n} \frac{a^k}{k!}}.$$

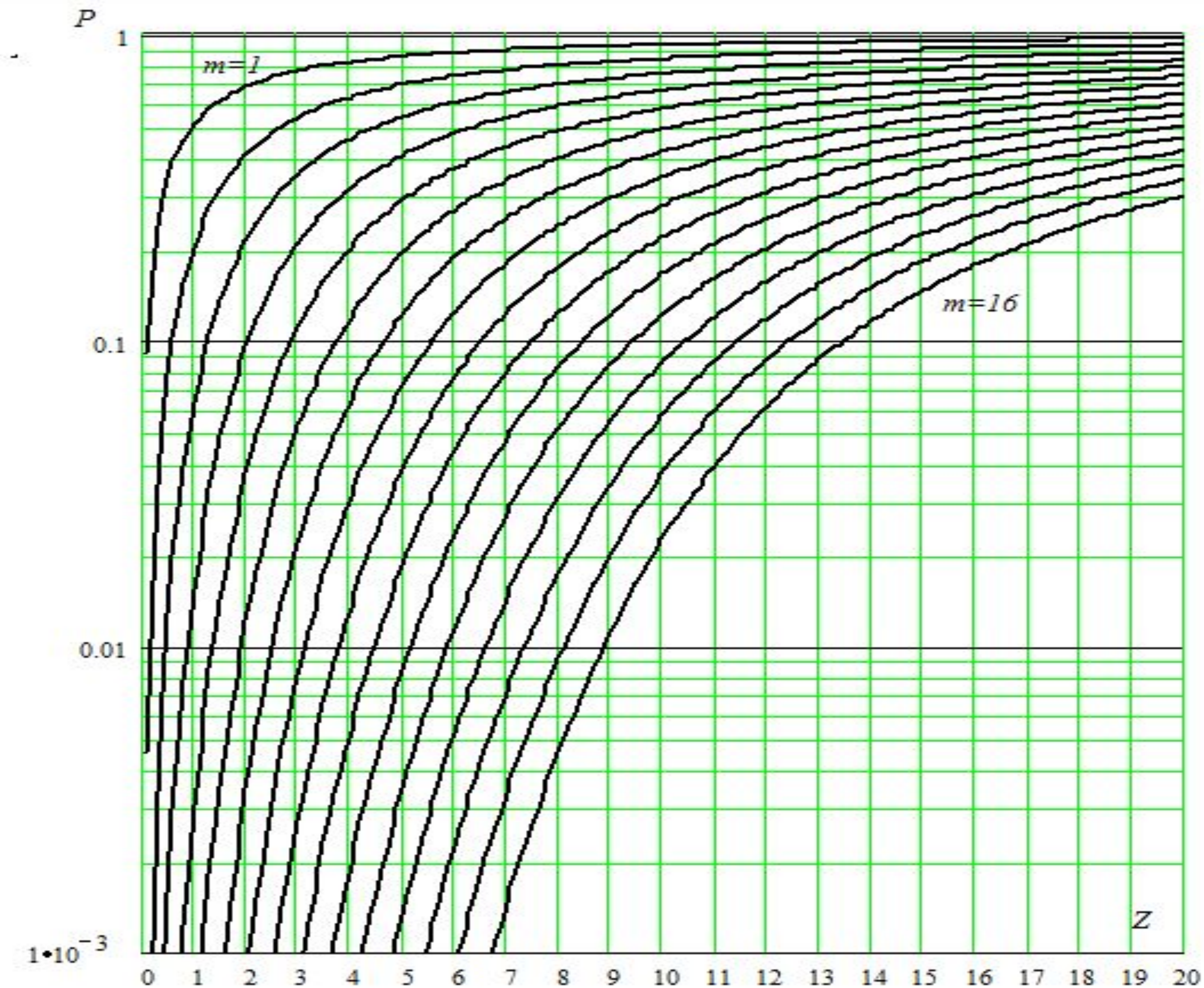
В частности, для одноканальной системы ($n = 1$)

$$P_{отк} = p_1 = \frac{a}{1+a},$$

а относительная пропускная способность

$$q = 1 - P_{отк} = \frac{1}{1+a}.$$

Пример графика, рассчитанного по формуле Эрланга



Задержка передачи отдельных пакетов τ_l по ЦК пути l определяется суммой четырех основных слагаемых:

$$\tau_l = \tau_{l.\text{расп}} + \tau_{l.\text{ком}} + \tau_{l.\text{кан}} + \tau_{l.\text{ож}} ,$$

где $\tau_{l.\text{расп}}$ – задержка распространения сигнала,

$\tau_{l.\text{ком}}$ – задержка коммутации,

$\tau_{l.\text{кан}}$ – задержка в канале передачи,

$\tau_{l.\text{ож}}$ – задержка из-за ожидания в очереди на передачу.

Задержка распространения сигнала $\tau_{l.\text{расп}}$ зависит в основном от физической длины пути $R_{l.\text{расп}}$ и от скорости $C_{l.\text{эмв}}$ распространения электромагнитных волн (ЭМВ) по используемой физической среде передачи: $\tau_{l.\text{расп}} = R_{l.\text{расп}} / C_{\text{эмв}}$.

Задержка коммутации $\tau_{l.\text{ком}}$ определяется производительностью устройства пакетной коммутации S_l, приведенной к входу ЦК рассматриваемого пути l. Если производительность S_l задана в пак/с с размером пакета, соответствующим размеру транспортного ПБД, то задержка коммутации будет равна $\tau_{l.\text{ком}} = 1/S_l$.

Задержка в канале передачи $\tau_{l.\text{кан}}$ определяется объемом (размером) V передаваемого ПБД (кадра, пакета) и скоростью передачи g_l в ЦК пути l.

$$\tau_{l.\text{кан}} = \frac{V}{g_l} .$$

Задержка из-за ожидания в очереди на передачу $\tau_{l.\text{ож}}$ существенно зависит от соотношения интенсивности протокольных блоков данных (ПБД) λ_l, поступающих на вход ЦК пути l, и интенсивности обслуживания ПБД в данном ЦК μ_l, равной обратной величине задержки в канале передачи $\tau_{l.\text{кан}}$ и прямо пропорциональной скорости передачи (канальному ресурсу) g_l в ЦК пути l:

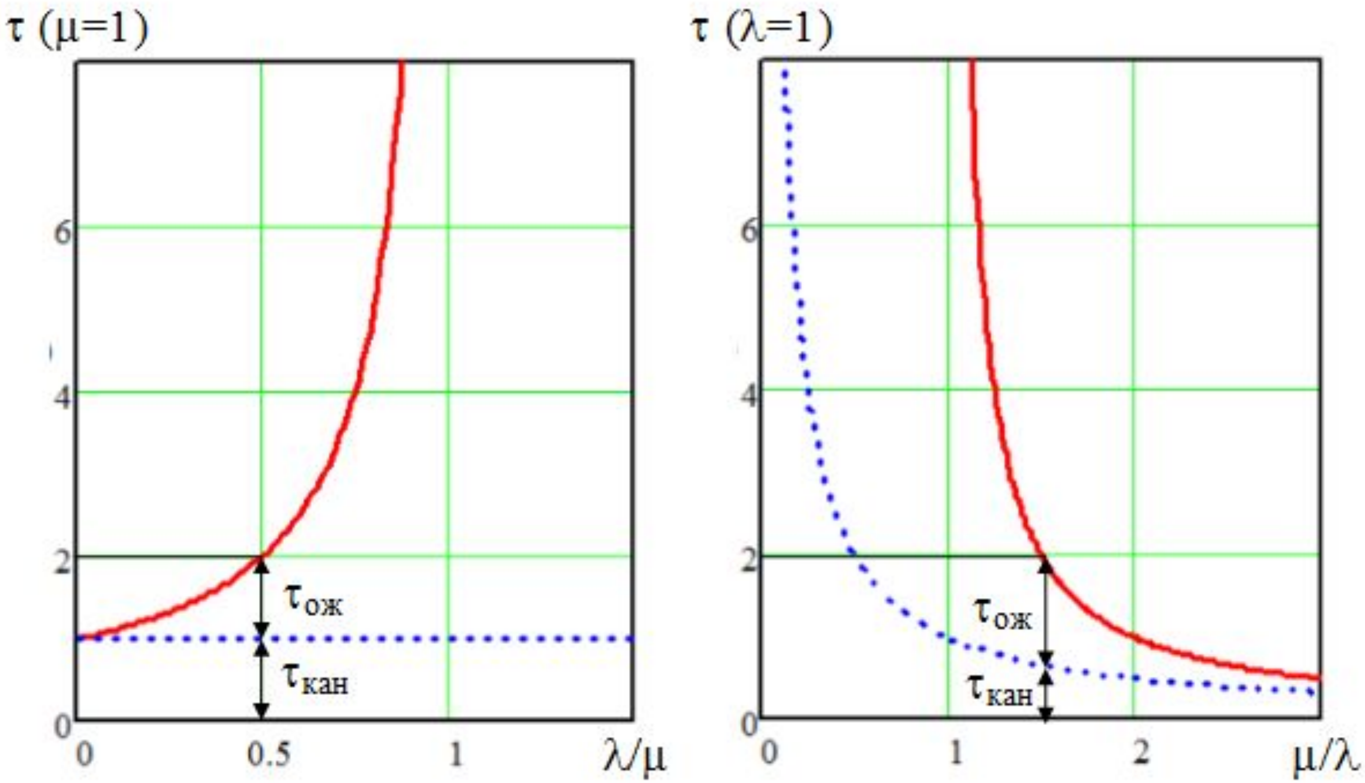
$$\mu_l = \frac{1}{\tau_{l.\text{кан}}} = \frac{g_l}{V}.$$

Отношение интенсивности ПБД λ_l, на входе ЦК к интенсивности обслуживания ПБД в ЦК μ_l принято называть нагрузкой на данный ЦК $\rho_l = \lambda_l / \mu_l$.

Из-за случайности моментов прихода очередных ПБД время ожидания в очереди на передачу является случайной величиной, из-за чего в расчетах обычно используют различные вероятностные и усредненные характеристики указанного времени. Методы расчета вероятностных характеристик подобных систем с ожиданием обслуживания являются предметом теории телетрафика и систем массового обслуживания.

$$\tau_{кан} = \frac{1}{\mu}, \quad \tau_{ож} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{1}{1-\rho},$$

$$\tau = \tau_{кан} + \tau_{ож} = \frac{1}{\mu - \lambda}$$



Зависимость относительной средней задержки τ в СМО М/М/1/∞ от соотношения входной интенсивности λ и интенсивности обслуживания μ

Расчет среднего времени задержки и вероятности потери пакетов

Модель СМО	Время задержки, τ	Вероятность потери пакетов, $p_{пот}$
M/M/1/*w*	$\tau = \frac{1}{\mu - \lambda}$	$p_{пот} = \frac{(1-\rho)\rho^{w}}{1-\rho^{w+1}}, \quad \rho=\lambda/\mu.$
G/G/n/*w*	$\tau = \frac{1}{\lambda} \cdot \frac{2 \cdot \rho(1-\rho) + C_a^2 + \left(\frac{\rho}{n}\right)^2 \cdot C_b^2}{2 \cdot (1-\rho)}$ $\rho < 1.$	$p_{пот} \cong p_{i>n-1} \cdot \frac{(1-\frac{\rho}{n})}{1-\left(\frac{\rho}{n}\right)^{\frac{2 \cdot w}{C_a^2 + C_b^2}+1}} \cdot \left(\frac{\rho}{n}\right)^{\frac{2 \cdot w}{C_a^2 + C_b^2}}.$
fbm/M(D)/1/*w*	$\tau = \frac{1}{\lambda}\left(\frac{\rho}{m} \cdot \frac{\rho^{\frac{1}{2(1-H)}}}{(1-\rho)^{\frac{-H}{H(1-H)}}} + \rho\right), \rho < 1.$	$p_{пот} = \bar{\Phi}\left(\frac{1}{\sqrt{a \cdot \rho}} \cdot \left(\frac{1-\rho}{H}\right)^{H} \cdot \left(\frac{m}{2} \cdot \frac{w+1}{1-H}\right)^{1-H}\right).$
P/P/1/*w*	$\tau = \frac{\rho}{\mu \cdot (1-\rho)} \cdot \frac{C_a^2 + C_b^2}{2}, \rho < 1.$	$p_{пот} = \frac{(1-\rho) \cdot \rho}{1-\rho^{\frac{2 \cdot w}{C_a^2 + C_b^2}+1}} \cdot \rho^{\frac{2 \cdot w}{C_a^2 + C_b^2}}.$

Для учета влияния случайных всплесков задержки на несвоевременность доставки ПБД $p_{\text{нсв}}$ необходимо знать функцию распределения $F(x)=\Pr(\Delta t<x)$ случайных задержек Δt. При этом $p_{\text{нсв}}= 1- \Pr(\Delta t<\tau_{\text{доп}}) = 1- F(\tau_{\text{доп}})$. Однако точный вид функции распределения, как правило, неизвестен. Наиболее популярной и простой является экспоненциальная функция распределения $F_{exp}(x)$:

$$F_{\text{exp}}(x) = 1 - \exp(-\frac{x}{\tau}) \qquad (1)$$

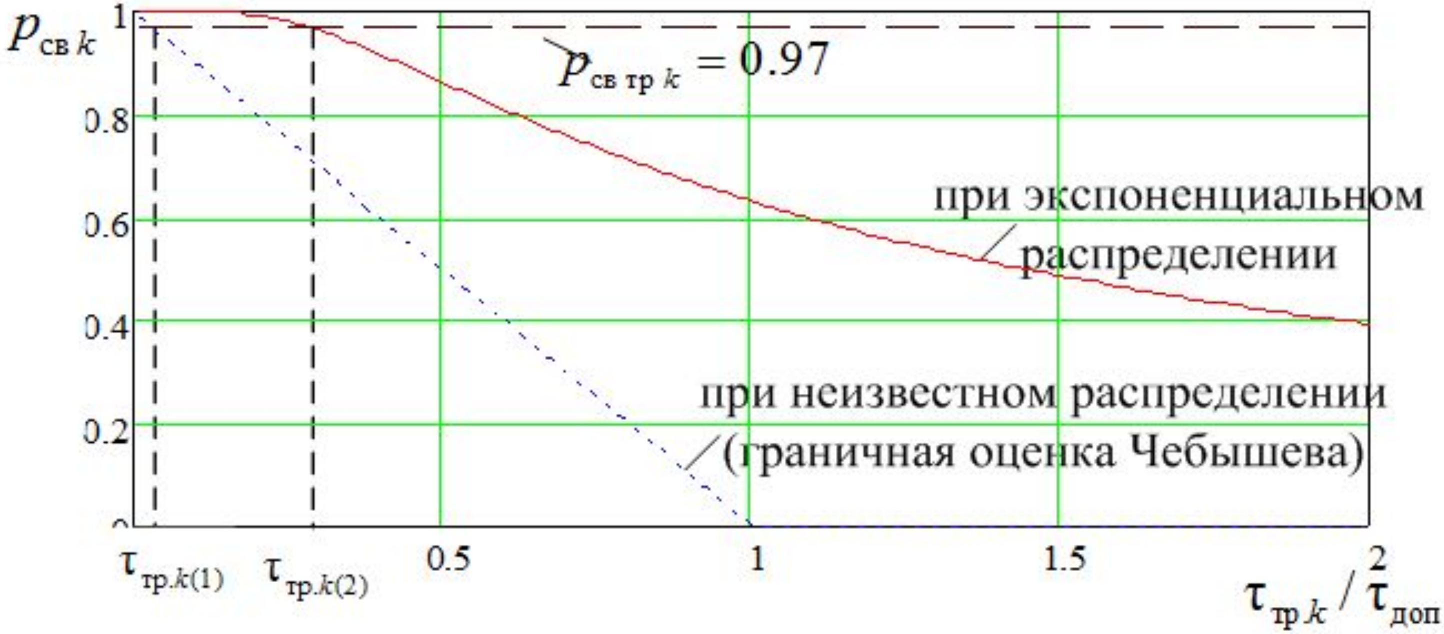
где τ – среднее значение (матожидание) случайной величины Δt.

Допуская справедливость экспоненциального распределения (1) можно рассчитать вероятность своевременной доставки пакетов $p_{\text{св}.k}$ при известной средней задержке τ_k по формуле

$$p_{\text{св}.k} = 1 - p_{\text{нсв}.k} = 1 - \exp(-\frac{\tau_{\text{доп}.k}}{\tau}) \qquad (2)$$

В случае, когда справедливость экспоненциального распределения (1) вызывает сомнение, можно оценить вероятность своевременной доставки пакетов $p_{\text{св}.k}$ по ее нижней границе Чебышева:

$$p_{\text{св}.k} = \begin{cases} 1 - \dfrac{\tau_k}{\tau_{\text{доп}.k}}, \tau_k < \tau_{\text{доп}.k} \\ 0, \qquad \tau_k \geq \tau_{\text{доп}.k} \end{cases} \qquad (3)$$



Пример пересчета требований к вероятности своевременной доставки за допустимое время в требования к средней задержке