Раздел 3. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ДИСКРЕТНЫХ СИСТЕМ

> «Соседняя очередь всегда движется быстрее. Как только вы перейдете в другую очередь, ваша бывшая начинает двигаться быстрее» (***Наблюдение Этторе***)

Исследование сложных систем предполагает построение абстрактных математических моделей, представленных на языке математических отношений в терминах определенной математической теории, позволяющей получить функциональные зависимости характеристик исследуемой системы от параметров. Изучение процессов, протекающих в дискретных системах со стохастическим характером функционирования, проводится в рамках *теории массового обслуживания (ТМО)* и *теории случайных процессов*. При этом многие модели реальных систем строятся на основе **моделей массового обслуживания (ММО)**, которые делятся на **базовые модели** в виде *систем массового обслуживания* и **сетевые модели** в виде *сетей массового обслуживания*, представляющие собой математические объекты, описываемые в терминах соответствующего математического аппарата.

3.1. Основные понятия

Для описания одного и того же понятия многочисленные литературные источники по моделям и методам теории массового обслуживания зачастую используют разные термины. Сама «теория массового обслуживания» часто называется «теорией очередей» (в англоязычной литературе Queue Theorie), наряду с термином «обслуживающий прибор» используются термины «устройство», «канал», «линия» и т.д. Обычно это связано с прикладной областью, в которой применяются модели массового обслуживания. Например, термины «вызов» и «линия» используются в телефонии (откуда собственно и пошла теория массового обслуживания), термин «клиент» – в моделях магазинов, банков, парикмахерских и т.д. В связи с этим, желательно иметь однозначные термины и понятия, которые будут использоваться при изложении материала в последующих разделах. Рассматривая модели массового обслуживания как абстрактные математические модели, ниже вводятся и используются термины безотносительно прикладной области применения этих моделей. Для каждого термина в круглых скобках перечислены термины-синонимы, которые могут встретиться в других источниках.

3.1.1. Система массового обслуживания

Система массового обслуживания (СМО) – математический (абстрактный) объект, содержащий один или несколько *приборов* **П** (*каналов*), обслуживающих *заявки* **З**, поступающие в систему, и

накопитель **Н**, в котором находятся заявки, образующие *очередь* **О** и ожидающие обслуживания (рис.3.1).

Заявка (требование, запрос, вызов, клиент) – объект, поступающий в СМО и требующий обслуживания в обслуживающем приборе.

Совокупность заявок, распределенных во времени, образуют ***поток заявок***.

Обслуживающий прибор или просто ***прибор (устройство, канал, линия)*** – элемент СМО, функцией которого является обслуживание заявок. В каждый момент времени в приборе на обслуживании может находиться только одна заявка.

Обслуживание – задержка заявки на некоторое время в обслуживающем приборе.

Длительность обслуживания – время задержки (обслуживания) заявки в приборе.

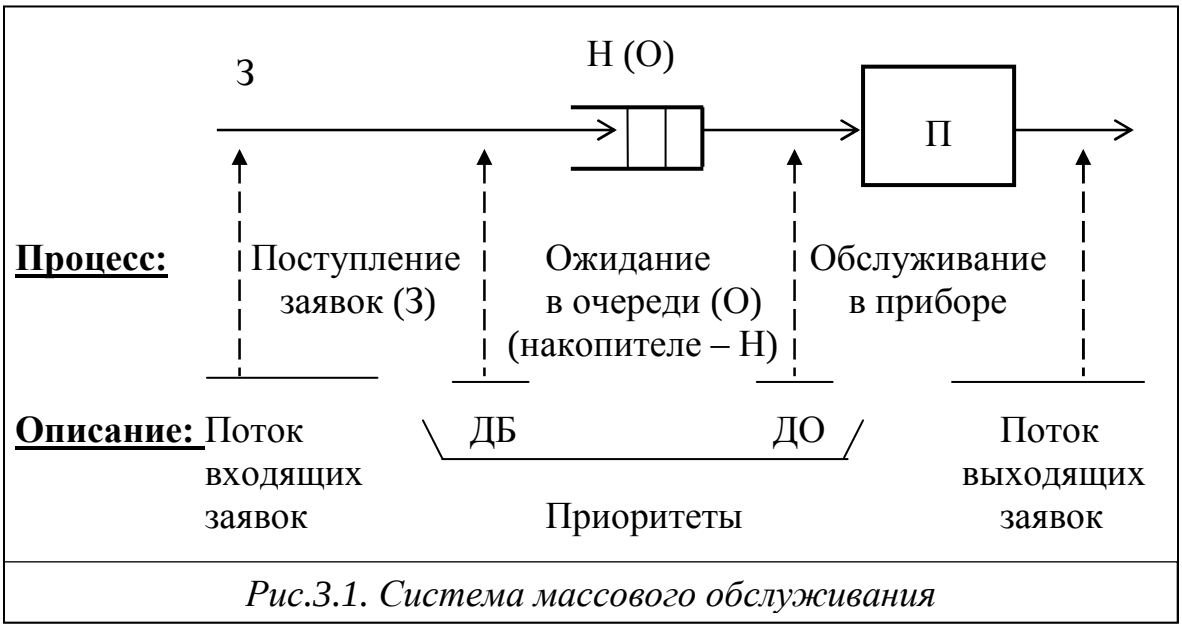


Рис.3.1. Система массового обслуживания

Накопитель (буфер) – совокупность мест для ожидания заявок перед обслуживающим прибором. Количество мест для ожидания определяет ***ёмкость накопителя***.

Заявка, поступившая на вход СМО, может находиться в двух состояниях:

- в состоянии *обслуживания* (в приборе);
- в состоянии *ожидания* (в накопителе), если все приборы заняты обслуживанием других заявок.

Заявки, находящиеся в накопителе и ожидающие обслуживания, образуют ***очередь*** заявок. Количество заявок, ожидающих обслуживания в накопителе, определяет ***длину очереди***.

Дисциплина буферизации – правило занесения поступающих заявок в накопитель (буфер).

Дисциплина обслуживания – правило выбора заявок из очереди для обслуживания в приборе.

Приоритет – преимущественное право на занесение (в накопитель) или выбор из очереди (для обслуживания в приборе) заявок одного класса по отношению к заявкам других классов.

Таким образом, СМО включает в себя:

- *заявки*, проходящие через систему и образующие *потоки заявок*;
- *очереди* заявок, образующиеся в накопителях;
- обслуживающие *приборы*.

Существует большое многообразие СМО, различающихся структурной и функциональной организацией. В то же время, разработка аналитических методов расчета характеристик функционирования СМО во многих случаях предполагает наличие ряда предположений, ограничивающих множество исследуемых СМО.

Ниже при рассмотрении СМО, если не оговорено другое, будем использовать следующие **предположения**:

- заявка, поступившая в систему, *мгновенно* попадает на обслуживание, если прибор свободен;
- в приборе на обслуживании в каждый момент времени может находиться только *одна* заявка;
- после завершения обслуживания какой-либо заявки в приборе очередная заявка выбирается на обслуживание из очереди мгновенно, то есть, другими словами, прибор *не простаивает*, если в очереди есть хотя бы одна заявка;
- поступление заявок в СМО и длительности их обслуживания не зависят от того, сколько заявок уже находится в системе, или от каких-либо других факторов;
- длительность обслуживания заявок не зависит от скорости (интенсивности) поступления заявок в систему.

3.1.2. Сеть массового обслуживания

Сеть массового обслуживания (СеМО) – совокупность взаимосвязанных СМО, в среде которых циркулируют заявки (рис.3.2,а).

Основными элементами СеМО являются узлы (У) и источники заявок (И).

Узел сети представляет собой систему массового обслуживания.

Источник – генератор заявок, поступающих в сеть и требующих определенных этапов обслуживания в узлах сети.

Для упрощенного изображения СеМО используется граф СеМО.

Граф СеМО – ориентированный граф, вершины которого соответствуют узлам СеМО, а дуги отображают переходы заявок между узлами (рис.3.2,б).

Переходы заявок между узлами СеМО, в общем случае, могут быть заданы в виде вероятностей передач.

Путь движения заявок в СеМО называется ***маршрутом***.

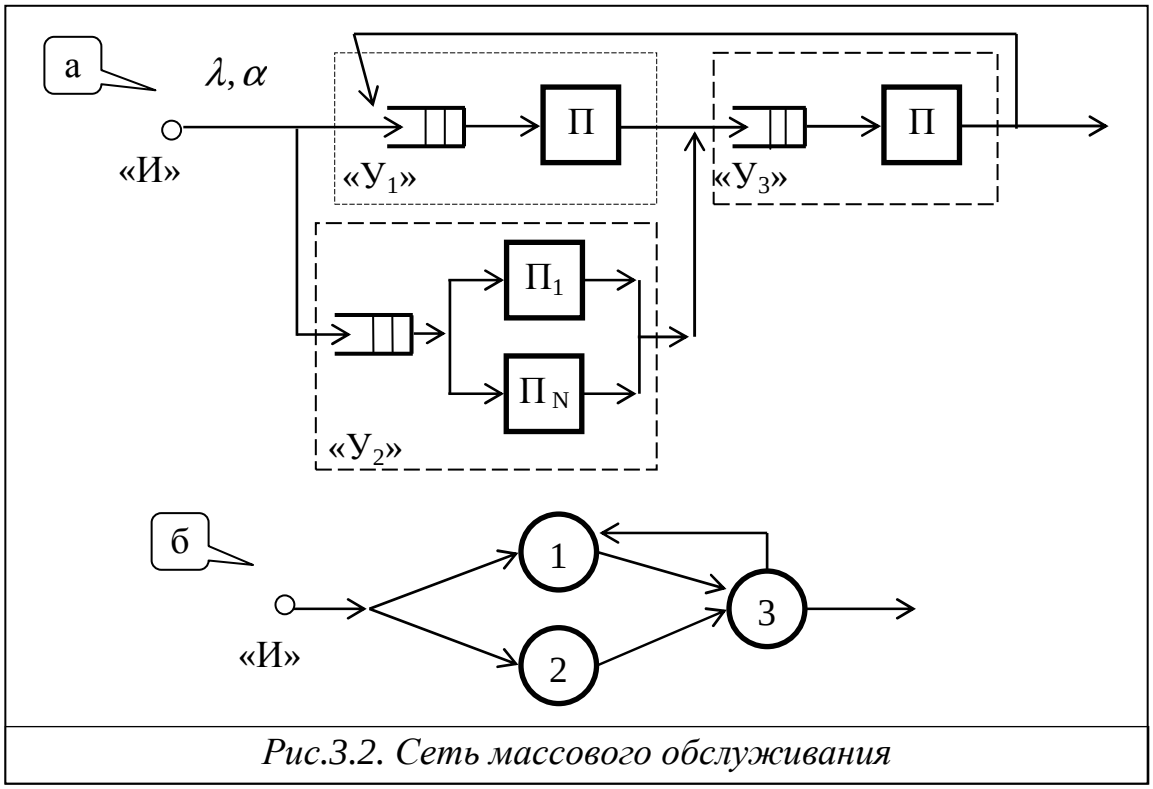


Рис.3.2. Сеть массового обслуживания

3.1.3. Поток заявок

Совокупность событий распределенных во времени называется ***потоком***. Если событие заключается в появлении заявок, имеем *поток заявок*.

Для описания потока заявок, в общем случае, необходимо задать интервалы времени $\tau_k = t_k - t_{k-1}$ между соседними моментами t_{k-1} и t_k поступления заявок с порядковыми номерами $(k-1)$ и k соответственно ($k = 1, 2, ...$; $t_0 = 0$ – начальный момент времени).

Основной характеристикой потока заявок является его ***интенсивность*** λ – среднее число заявок, проходящих через некоторую границу за единицу времени. Величина $a = 1/\lambda$ определяет *средний интервал времени между двумя последовательными заявками*.

Поток, в котором интервалы времени τ_k между соседними заявками принимают определенные заранее известные значения, называется ***детерминированным***. Если при этом интервалы одинаковы ($\tau_k = \tau$ для всех $k = 1, 2, ...$), то поток называется ***регулярным***. Для полного описания регулярного потока заявок достаточно задать интенсивность потока λ или значение интервала $\tau = 1/\lambda$.

Поток, в котором интервалы времени τ_k между соседними заявками представляют собой случайные величины, называется ***случайным***. Для полного описания случайного потока заявок, в общем случае, необходимо задать законы распределений $A_k(\tau_k)$ всех интервалов τ_k ($k = 1, 2, ...$).

Случайный поток, в котором все интервалы $\tau_1, \tau_2, \ldots$ между заявками независимы в совокупности и описываются функциями распределений $A_1(\tau_1)$, $A_2(\tau_2)$,…, называется потоком ***с ограниченным последействием***.

Случайный поток, в котором все интервалы $\tau_1, \tau_2, \ldots$ распределены *по одному и тому же закону* $A(\tau)$, называется ***рекуррентным***.

Поток заявок называется ***стационарным***, если интенсивность λ и закон распределения $A(\tau)$ интервалов между последовательными заявками не меняются со временем. В противном случае поток заявок является ***нестационарным***.

Поток заявок называется ***ординарным***, если в каждый момент времени t_k может появиться только одна заявка. Если в какой-либо момент времени может появиться более одной заявки, то имеем ***неординарный*** или ***групповой*** поток заявок.

Поток заявок называется потоком ***без последействия***, если заявки поступают *независимо* друг от друга, то есть момент поступления очередной заявки не зависит от того, когда и сколько заявок поступило до этого момента.

Стационарный ординарный поток *без последействия* называется ***простейшим***.

Интервалы времени τ между заявками в простейшем потоке распределены по *экспоненциальному закону* с функцией распределения

$$A(\tau) = 1 - e^{-\lambda\tau}, \qquad (3.1)$$

где $\lambda > 0$ – параметр распределения, представляющий собой интенсивность потока заявок.

Простейший поток часто называют ***пуассоновским***, поскольку число заявок k, поступающих за некоторый заданный промежуток времени t, распределено по *закону Пуассона*:

$$P(k,t) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}, \qquad (3.2)$$

где $P(k,t)$ – вероятность поступления ровно k заявок за некоторый фиксированный интервал времени t; λ – интенсивность потока заявок. Здесь k – дискретная случайная величина, принимающая целочисленные значения: $k = 0, 1, 2, \ldots$, а $t > 0$ и $\lambda > 0$ – параметры закона Пуассона.

Следует отметить, что пуассоновский поток, в отличие от простейшего, может быть:

- *стационарным*, если интенсивность λ не меняется со временем;
- *нестационарным*, если интенсивность потока зависит от времени: $\lambda = \lambda(t)$.

В то же время, простейший поток, по определению, всегда является стационарным.

Аналитические исследования моделей массового обслуживания часто проводятся в предположении о простейшем потоке заявок, что обусловлено рядом присущих ему замечательных особенностей.

1. **Суммирование (объединение) потоков**. Сумма H независимых стационарных ординарных потоков с интенсивностями $\lambda_1, \ldots, \lambda_H$ образует простейший поток с интенсивностью

$$\Lambda = \sum_{k=1}^{H} \lambda_k \qquad (3.3)$$

при условии, что складываемые потоки оказывают более или менее одинаково малое влияние на суммарный поток. На практике суммарный поток близок к простейшему при $H \geq 5$. Очевидно, что *при суммировании независимых простейших потоков суммарный поток будет простейшим* при любом значении H.

2. **Вероятностное разрежение потока**. *Вероятностное* (но *не детерминированное*) разрежение *простейшего потока* заявок, при котором любая заявка случайным образом с некоторой вероятностью p исключается из потока независимо от того, исключены другие заявки или нет, приводит к образованию *простейшего потока* с интенсивностью $\lambda' = p\lambda$, где λ – интенсивность исходного потока. Поток исключенных заявок – тоже *простейший* с интенсивностью $\lambda'' = (1-p)\lambda$.

3. **Простота**. Предположение о простейшем потоке заявок позволяет для многих математических моделей сравнительно легко получить в явном виде зависимости характеристик от параметров. Наибольшее число аналитических результатов получено для простейшего потока заявок. Анализ моделей с потоками заявок, отличными от простейших, обычно усложняет математические выкладки и не всегда позволяет получить аналитическое решение в явном виде. Свое название «*простейший*» поток получил именно благодаря этой особенности.

3.1.4. Длительность обслуживания заявок

Длительность обслуживания – время нахождения заявки в приборе – в общем случае величина случайная и описывается функцией $B(\tau)$ или плотностью $b(\tau) = B'(\tau)$ распределения. В случае неоднородной нагрузки длительности обслуживания заявок разных классов могут различаться законами распределений или только средними значениями. При этом обычно предполагается независимость длительностей обслуживания заявок каждого класса.

Часто длительность обслуживания заявок предполагается распределенной по *экспоненциальному закону*, что существенно упрощает аналитические выкладки. Это обусловлено тем, что процессы, протекающие в системах с экспоненциальным распределением интервалов времени, являются *марковскими* (см. раздел 5).

Величина, обратная средней длительности обслуживания b, характеризует среднее число заявок, которое может быть обслужено за единицу времени, и называется ***интенсивностью обслуживания*****:** $\mu = 1/b$.

Во многих случаях аналитические зависимости могут быть получены для произвольного закона распределения длительности обслуживания заявок. При этом для определения средних значений характеристик обслуживания, зачастую, как будет показано ниже, достаточно задать, кроме математического ожидания b, второй момент распределения (дисперсию) или коэффициент вариации ν_b длительности обслуживания.

Время T_0, оставшееся до завершения обслуживания заявки, находящейся в приборе, от момента поступления некоторой заявки в систему, и учитывающее, что на момент поступления в системе может и не оказаться заявок, то есть учитывающее простои системы, называется ***временем дообслуживания***. Математическое ожидание этого времени [9]:

$$M[T_0] = \lambda b^2 (1+\nu_b^2)/2, \qquad (3.4)$$

где λ – интенсивность *простейшего* потока заявок, поступающих в систему.

3.1.5. Стратегии управления потоками заявок

Стратегия управления потоками заявок в моделях массового обслуживания задается в виде:

- *дисциплины буферизации (ДБ)*;
- *дисциплины обслуживания (ДО)*.

ДБ и ДО могут быть классифицированы по следующим признакам:

- наличие приоритетов между заявками разных классов;
- способ (режим) вытеснения заявок из очереди (для ДБ) и назначения заявок на обслуживание (для ДО);
- правило вытеснения или выбора заявок на обслуживание;
- возможность изменения приоритетов.

Одна из возможных **классификаций дисциплин буферизации** в соответствии с перечисленными признаками представлена на рис.3.3.

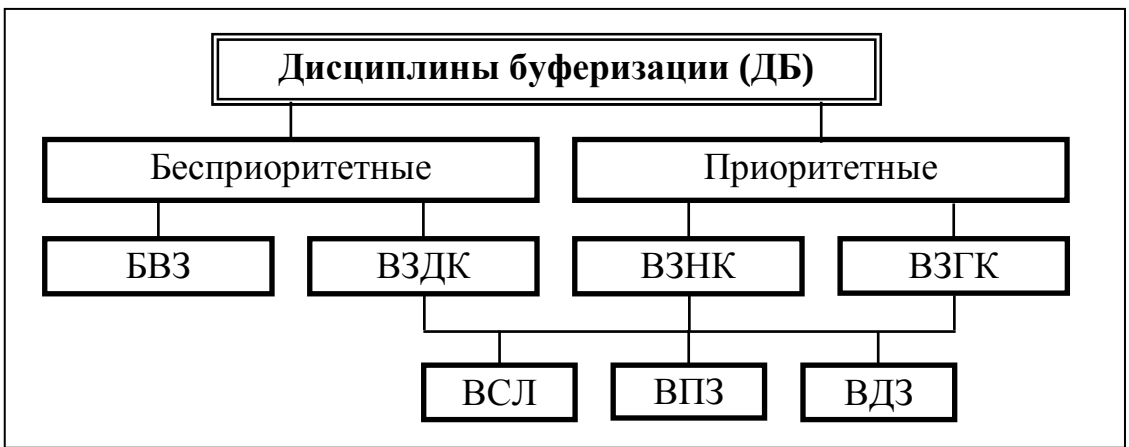


Рис.3.3. Классификация дисциплин буферизации

В зависимости от *наличия или отсутствия приоритетов* между заявками разных классов все ДБ могут быть разбиты на две группы:

- бесприоритетные;
- приоритетные.

По *способу вытеснения заявок из накопителя* можно выделить следующие классы ДБ:

- без вытеснения заявок (БВЗ) – заявки, поступившие в систему и заставшие накопитель заполненным до конца, теряются;
- с вытеснением заявки данного класса (ВЗДК), то есть такого же класса, что и поступившая;
- с вытеснением заявки самого низкоприоритетного класса (ВЗНК);
- с вытеснением заявки, принадлежащей группе низкоприоритетных классов (ВЗГК).

Два первых класса относятся к бесприоритетным ДБ, а остальные – к приоритетным.

ДБ могут использовать следующие *правила вытеснения заявок из накопителя*:

- вытеснение случайное (ВСЛ);
- вытеснение последней заявки (ВПЗ), то есть поступившей в систему позже всех;
- вытеснение «долгой» заявки (ВДЗ), то есть находящейся в накопителе дольше всех.

Часто ёмкость накопителя в моделях предполагается неограниченной, несмотря на то, что в реальной системе соответствующая ёмкость ограничена. Такое предположение оправдано в тех случаях, когда вероятность потери заявки в реальной системе из-за переполнения ограниченной ёмкости накопителя меньше 10^{-3}, поскольку в этом случае ДБ практически не влияет на характеристики обслуживания заявок.

На рис.3.4 представлена **классификация дисциплин обслуживания** заявок в соответствии с теми же признаками, что и для ДБ.

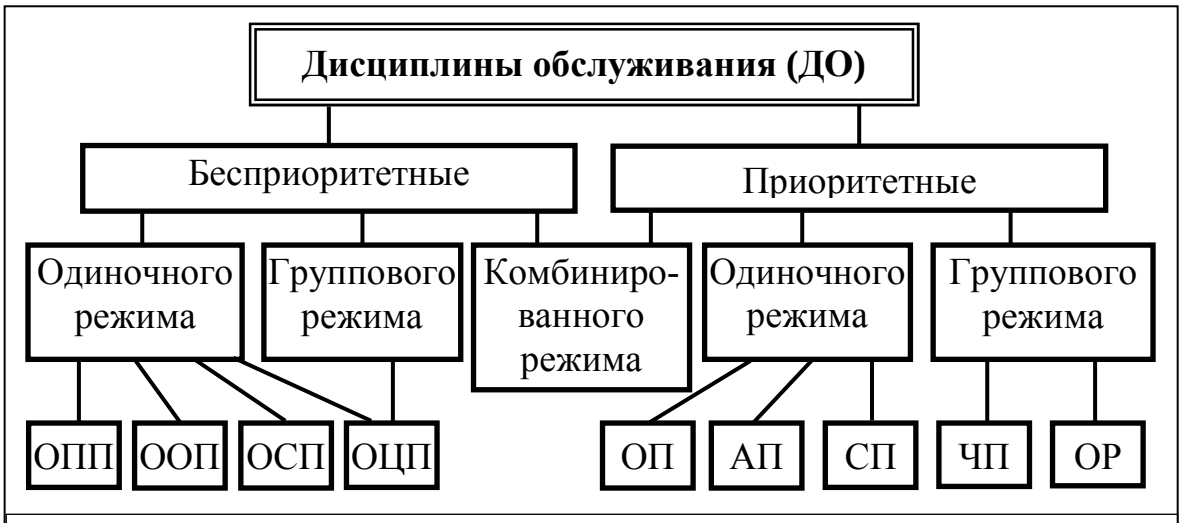


Рис.3.4. Классификация дисциплин обслуживания

В зависимости от *наличия или отсутствия приоритетов* между заявками разных классов все ДО, как и ДБ, могут быть разбиты на две группы:

- бесприоритетные;
- приоритетные.

По *способу назначения заявок на обслуживание* ДО могут быть разделены на дисциплины:

- одиночного режима;
- группового режима;
- комбинированного режима.

В ДО **одиночного режима** всякий раз на обслуживание *назначается только одна заявка* (просмотр очередей с целью назначения на обслуживание в приборе очередной заявки выполняется после обслуживания каждой заявки).

В ДО **группового режима** всякий раз на обслуживание *назначается группа заявок* одной очереди (просмотр очередей с целью очередного назначения на обслуживание выполняется только после обслуживания всех заявок ранее назначенной группы). В предельном случае назначаемая на обслуживание группа заявок может включать в себя все заявки данной очереди. Заявки назначенной на обслуживание группы *последовательно выбираются из очереди* и обслуживаются прибором, после чего на обслуживание назначается следующая группа заявок другой очереди в соответствии с заданной ДО.

Комбинированный режим – комбинация одиночного и группового режимов, когда часть очередей заявок обрабатывается в одиночном режиме, а другая часть – в групповом.

ДО могут использовать следующие *правила выбора заявок на обслуживание*:

- бесприоритетные:
  - **обслуживание в порядке поступления** (ОПП или FIFO – First In First Out), когда на обслуживание выбирается заявка, поступившая в систему раньше других;
  - **обслуживание в обратном порядке** (ООП или LIFO – Last In First Out) когда на обслуживание выбирается заявка, поступившая в систему позже других;
  - **обслуживание в случайном порядке** (ОСП), когда на обслуживание заявка выбирается случайным образом;
  - **обслуживание в циклическом порядке** (ОЦП), когда на обслуживание заявки выбираются в процессе циклического опроса накопителей в последовательности 1, 2, ..., H (H – количество накопителей), после чего указанная последовательность повторяется;
- приоритетные:

- с **относительными приоритетами** (ОП), означающими, что приоритеты учитываются только в моменты завершения обслуживания заявок при выборе новой заявки на обслуживание и не влияют на процесс обслуживания низкоприоритетной заявки в приборе; другими словами, поступление в систему заявки с более высоким приоритетом по сравнению с обслуживаемой в приборе не приводит к прерыванию обслуживаемой заявки;
- с **абсолютными приоритетами** (АП), означающими, что, в отличие от ОП, при поступлении высокоприоритетной заявки обслуживание заявки с низким приоритетом прерывается и на обслуживание принимается поступившая высокоприоритетная заявка; при этом прерванная заявка может быть возвращена в накопитель или удалена из системы; если заявка возвращена в накопитель, то её дальнейшее обслуживание может быть продолжено с прерванного места или начато заново, то есть с самого начала;
- со **смешанными приоритетами** (СП), представляющими собой любую комбинацию бесприоритетного обслуживания, ОП и АП;
- с **чередующимися приоритетами** (ЧП), являющимися аналогом ОП и проявляющимися только в моменты завершения обслуживания группы заявок одной очереди и назначения новой группы;
- **обслуживание по расписанию** (ОР), когда заявки разных классов (находящиеся в разных накопителях) выбираются на обслуживание в соответствии с некоторым расписанием (планом), задающим последовательность опроса очередей заявок, например, в случае трех классов заявок (накопителей) расписание может иметь вид: {1, 2, 1, 3, 1, 2}.

Дисциплины ОПП, ООП, ОП, АП и СП относятся к дисциплинам *одиночного режима*. Очевидно, что дисциплины *группового режима* ОЦП, ЧП и ОР, в частном случае могут быть реализованы как ДО одиночного режима, если размер назначаемой на обслуживание группы равен 1, при этом ДО ЧП вырождается в ДО ОП.

Среди представленных ДО особое место занимают дисциплины со смешанными приоритетами (СП), обладающие общностью по отношению к перечисленным ДО одиночного режима [3].

Для математического описания ДО СП используется **матрица приоритетов (МП)**, представляющая собой квадратную матрицу: $\mathbf{Q} = [q_{ij} \mid i, j = 1, \ldots, H]$, где H – число классов заявок, поступающих в систему.

Элемент q_{ij} матрицы задает приоритет заявок класса i по отношению к заявкам класса j и может принимать следующие значения:

- 0 – нет приоритета;
- 1 – приоритет относительный (ОП);
- 2 – приоритет абсолютный (АП).

Элементы МП должны удовлетворять следующим *требованиям*:

- $q_{ii} = 0$, так как между заявками одного и того же класса не могут быть установлены приоритеты;
- если $q_{ij} = 1$ или 2, то $q_{ji} = 0$, так как если заявки класса i имеют приоритет к заявкам класса j, то последние не могут иметь приоритет к заявкам класса i $(i, j = \overline{1, H})$.

В зависимости от *возможности изменения приоритетов* в процессе функционирования системы приоритетные дисциплины буферизации и обслуживания делятся на два класса:

- ***со статическим приоритетами***, которые не изменяются со временем;
- ***с динамическими приоритетами***, которые могут изменяться в процессе функционирования системы в зависимости от разных факторов, например, при достижении некоторого критического значения длины очереди заявок какого-либо класса, обладающего низким приоритетом, ему может быть предоставлен более высокий приоритет.

3.2. Классификация моделей массового обслуживания

3.2.1. Базовые модели

При моделировании реальных систем с дискретным характером функционирования широкое применение находят базовые модели в виде СМО, которые могут быть классифицированы (рис.3.5):

- по числу мест в накопителе;
- по числу обслуживающих приборов;
- по количеству классов заявок, поступающих в СМО.

1. По *числу мест в накопителе* СМО делятся на системы:

- **без накопителя**, в которых заявка, поступившая в систему и заставшая все обслуживающие приборы занятыми обслуживанием более высокоприоритетных заявок, получает отказ и теряется; такие системы называются ***СМО с отказами***;
- **с накопителем ограниченной ёмкости (*СМО с потерями*)**, в которых поступившая заявка теряется, если она застает накопитель заполненным до конца;
- **системы с накопителем неограниченной ёмкости (*СМО без потерь*)**, в которых для любой поступившей заявки всегда найдется место в накопителе для ожидания.

В дальнейшем, накопитель неограниченной ёмкости будем изображать так, как это показано на рис.3.5,а, и накопитель ограниченной ёмкости – как на рис.3.5,б.

Как уже было сказано выше, предположение о неограниченной ёмкости накопителя может использоваться для моделирования реальных систем, в которых вероятность потери заявки из-за переполнения накопителя ограниченной ёмкости меньше 10^{-3}.

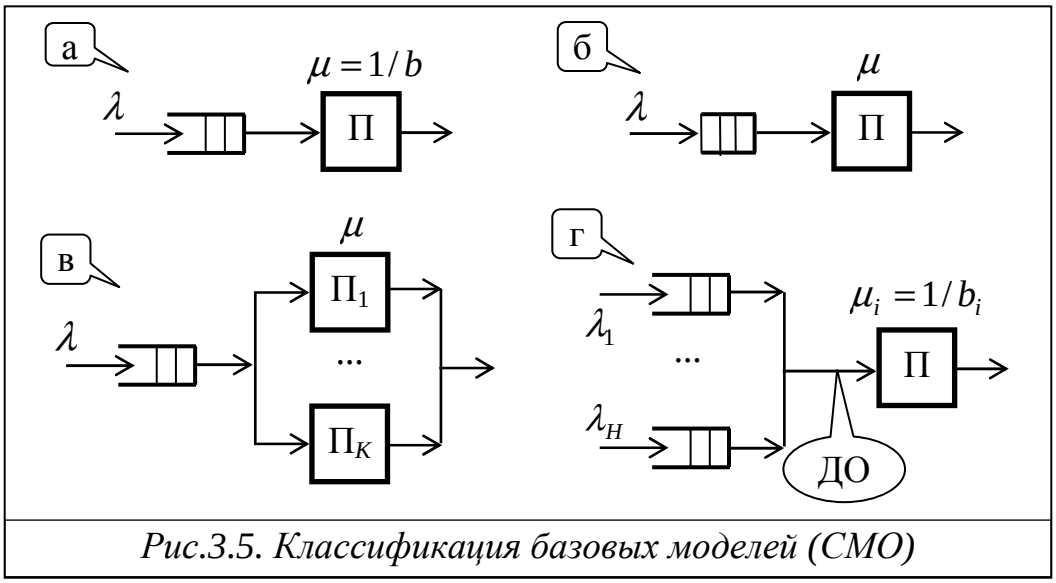


Рис.3.5. Классификация базовых моделей (СМО)

2. По *количеству обслуживающих приборов* СМО делятся на:

- **одноканальные** (рис.3.5,а, б, г), содержащие один прибор **П**;
- **многоканальные** (рис.3.5,в), содержащие K обслуживающих приборов $\mathbf{П}_1,...,\mathbf{П}_K$ ($K > 1$).

В многоканальных СМО обычно предполагается, что все приборы идентичны и равнодоступны для любой заявки, то есть при наличии нескольких свободных приборов поступившая заявка с равной вероятностью может попасть в любой из них на обслуживание.

3. По *количеству классов (типов) заявок*, поступающих в СМО, различают системы:

- **с однородным потоком** заявок (рис.3.5,а, б, в);
- **с неоднородным потоком** заявок (рис.3.5,г).

Однородный поток заявок образуют заявки одного класса, а неоднородный поток представляет собой поток заявок нескольких классов.

В СМО, представляющей собой абстрактную математическую модель, *заявки относятся к разным классам* в том случае, если они в моделируемой реальной системе различаются хотя бы одним из следующих факторов:

- *длительностью обслуживания*;
- *приоритетами*.

Если же заявки не различаются длительностью обслуживания и приоритетами, то в СМО они могут быть представлены как заявки одного класса, независимо от их физической сущности.

3.2.2. Сетевые модели

В зависимости от структуры и свойств исследуемых систем их моделями могут служить СеМО различных классов. Одна из возможных классификаций сетевых моделей приведена на рис.3.6.

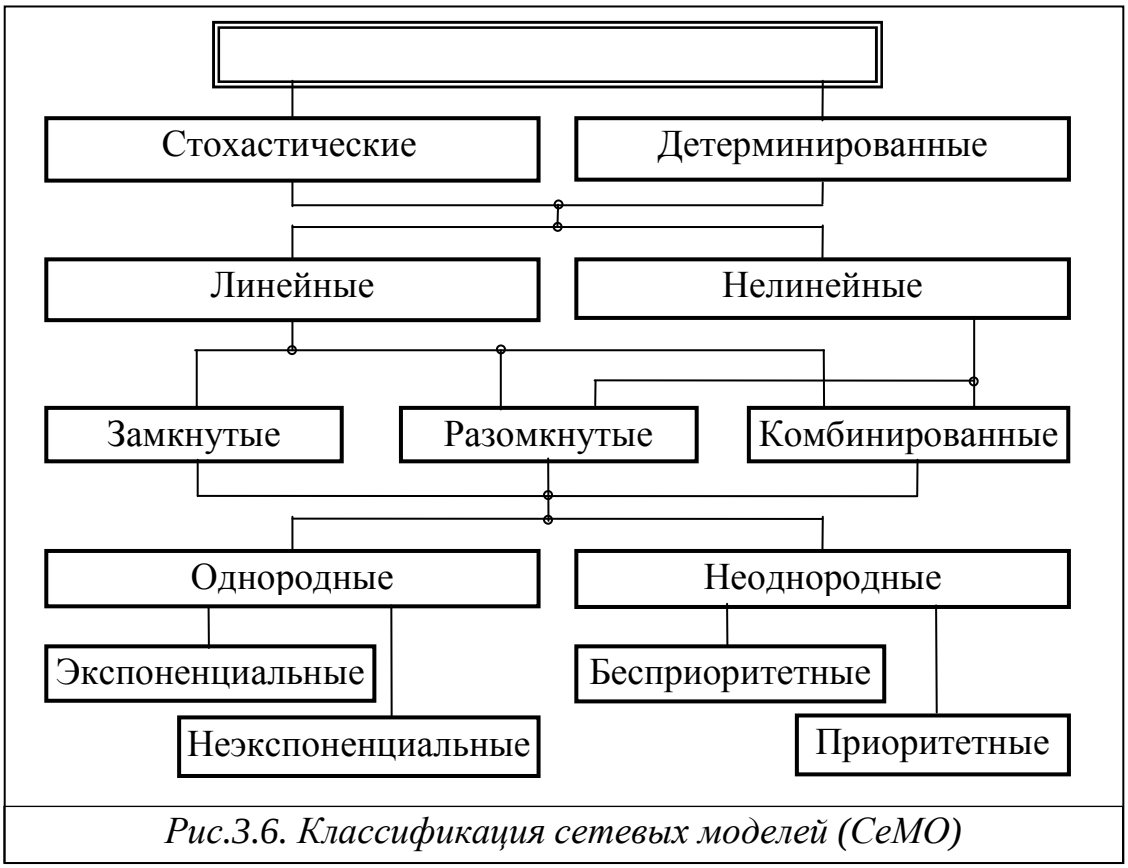


Рис.3.6. Классификация сетевых моделей (СеМО)

1. В зависимости от *характера процессов поступления и обслуживания заявок* в сети СеМО делятся на:

- **стохастические**, в которых процессы поступления и/или обслуживания заявок носят случайный характер, то есть интервалы времени между поступающими заявками и/или длительности их обслуживания в узлах представляют собой случайные величины, описываемые соответствующими законами распределений;
- **детерминированные**, в которых интервалы времени между поступающими заявками и длительности их обслуживания в узлах являются детерминированными величинами.

2. По *виду зависимостей, связывающих интенсивности потоков заявок в разных узлах*, СеМО делятся на:

- **линейные**, если эти зависимости линейные;
- **нелинейные**, если эти зависимости являются нелинейными.

В *линейных* СеМО, как это следует из определения, интенсивность потока заявок в узел j связана с интенсивностью потока заявок в узел i линейной зависимостью:

$$\lambda_j = \alpha_{ij}\lambda_i,$$

где α_{ij} – коэффициент пропорциональности, показывающий, во сколько раз отличаются интенсивности потоков заявок в узел j и в узел i $(i, j = \overline{1,n})$.

Поскольку указанная зависимость справедлива для любой пары узлов, это выражение можно записать в несколько ином виде и выразить интенсивность поступления заявок во все узлы $j = \overline{1,n}$ через одну и ту же интенсивность, например, через интенсивность λ_0 потока заявок, поступающих в СеМО из источника заявок:

$$\lambda_j = \alpha_j \lambda_0. \qquad (3.5)$$

В последнем выражении коэффициент пропорциональности $\alpha_j \geq 0$ показывает, во сколько раз интенсивность потока заявок в узел j $(i, j = \overline{1,n})$ отличается от интенсивности источника заявок, и называется ***коэффициентом передачи***. Коэффициент передачи может принимать любое положительное значение.

Коэффициент передачи играет важную роль при разработке математических зависимостей и расчете характеристик функционирования сетевых моделей. Это обусловлено тем физическим смыслом, который несет в себе коэффициент передачи.

Коэффициент передачи можно трактовать как *среднее число попаданий заявки в данный узел за время ее нахождения в сети*. Например, если коэффициент передачи узла СеМО равен 3, то это означает, что любая заявка за время нахождения в сети *в среднем* 3 раза побывает на обслуживании в данном узле. Значение коэффициента передачи, равное 0,25, будет означать, что *в среднем* только одна заявка из четырёх попадёт на обслуживание в данный узел, а три другие обойдут данный узел стороной.

В *нелинейных* СеМО интенсивности потоков заявок в узлах связаны более сложными нелинейными зависимостями, что значительно усложняет их исследование.

Нелинейность СеМО может быть обусловлена:

- *потерей заявок* в сети, например из-за ограниченной емкости накопителей в узлах;
- *размножением заявок* в сети, заключающимся, например, в формировании нескольких новых заявок после завершения обслуживания некоторой заявки в одном из узлов сети.

Таким образом, СеМО является линейной, если в ней заявки не размножаются и не теряются. Ниже рассматриваются, в основном, линейные СеМО.

3. По *числу циркулирующих в сети заявок* различают СеМО:

- разомкнутые;

- замкнутые;
- замкнуто-разомкнутые.

Разомкнутая (открытая) СеМО (РСеМО) содержит один или несколько *внешних независимых источников* заявок, которые генерируют заявки в сеть независимо от числа заявок, находящихся в сети (рис.3.7,а). В РСеМО одновременно может находиться *любое число заявок*, в том числе, и сколь угодно большое, то есть от 0 до бесконечности. С РСеМО связана внешняя среда, из которой поступают заявки в сеть и в которую они возвращаются после обслуживания в сети. Внешняя среда в РСеМО обозначается обычно как нулевой узел "0", и РСеМО, в этом случае, изображается в виде рис.3.7,б.

Замкнутая (закрытая) СеМО (ЗСеМО) не содержит *независимых внешних* источников заявок и характеризуется тем, что в ней циркулирует *постоянное число заявок M* (рис.3.7,в). На графе ЗСеМО из физических соображений, связанных с конкретным представлением процесса функционирования исследуемой реальной системы, обычно выделяется особая дуга, отображающая процесс завершения обслуживания заявок в сети и мгновенного формирования новой заявки с такими же параметрами обслуживания, что и завершившая обслуживание. Такая трактовка позволяет рассматривать завершившую обслуживание заявку как новую заявку, поступившую в сеть из *зависимого источника* заявок.

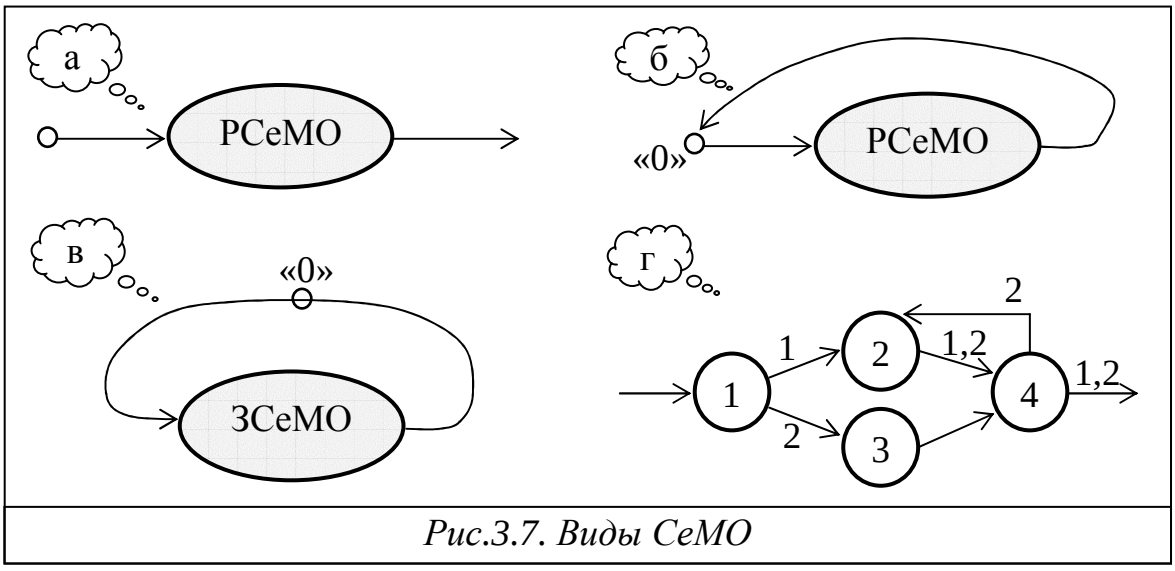


Рис.3.7. Виды СеМО

По аналогии с РСеМО на выделенной дуге ЗСеМО отмечается условная точка "0", рассматриваемая как нулевой узел и трактуемая иногда как фиктивная СМО с нулевой длительностью обслуживания или как зависимый источник заявок, генерирующий заявки только в момент поступления некоторой заявки на его вход. Выделение нулевого узла в ЗСеМО преследует двоякую цель: во-первых, достигается однозначность в представлении и математическом описании РСеМО и ЗСеМО; во-вторых, обеспечивается возможность определения временны*х* характеристик ЗСеМО относительно выделенного узла "0". В частности, ***время пребыва-***

ния заявок в ЗСеМО рассматривается как промежуток времени между двумя соседними моментами прохождения заявки через нулевой узел.

Замкнуто-разомкнутая СеМО (комбинированная) представляет собой комбинацию ЗСеМО и РСеМО, в которую, кроме постоянно циркулирующих в сети M^* заявок, из внешнего независимого источника поступают заявки такого же или другого класса, при этом суммарное число заявок в сети $M \geq M^*$.

4. По *типу циркулирующих заявок* различают СеМО:

- **однородные**, в которых циркулирует один класс заявок (однородный поток заявок);
- **неоднородные**, в которых циркулирует несколько классов заявок (неоднородный поток заявок), различающихся хотя бы одним из следующих факторов:
  - *длительностями обслуживания* в узлах;
  - *приоритетами*;
  - *маршрутами*.

Маршруты заявок разных классов задаются путем указания номеров классов заявок на соответствующих дугах сети (рис.3.7,г).

3.3. Параметры и характеристики СМО

«Чем больше ожидание, тем больше вероятность, что вы стоите не в той очереди» (***Принцип очереди***)

3.3.1. Параметры СМО

Для описания СМО используются три группы параметров:

- структурные;
- нагрузочные;
- функциональные параметры (параметры управления).

К *структурным параметрам* относятся:

- *количество обслуживающих приборов K*, равное 1 для одноканальной СМО и *K* >1 для многоканальной СМО;
- *количество k и ёмкости накопителей* E_j $(j = \overline{1,k})$;
- *способ взаимосвязи накопителей с приборами* (в случае многоканальных СМО), например в виде матрицы связей.

Нагрузочные параметры СМО включают в себя:

- количество поступающих в систему классов заявок *H*, которое равно 1 для СМО с однородным потоком заявок и *H* >1 для СМО с неоднородным потоком;
- закон распределения $A_i(\tau)$ интервалов времени между поступающими в систему заявками класса $i = \overline{1,H}$ или, по-крайней мере, первые два момента распределения, задаваемые, например, в виде интенсивности λ_i и коэффициента вариации ν_{a_i} интервалов;

- закон распределения $B_i(\tau)$ длительности обслуживания заявок класса $i = \overline{1, H}$ или, как минимум, первые два момента распределения, в качестве которых обычно используются средняя длительность b_i или интенсивность $\mu_i = 1/b_i$ обслуживания и коэффициент вариации ν_{b_i}.

Задание двух первых моментов нагрузочных параметров зачастую оказывается достаточным для оценки характеристик обслуживания заявок на уровне средних значений. Отметим, что *для описания простейшего потока достаточно задать только интенсивность поступления заявок* в систему.

Функциональные параметры задаются в виде конкретных стратегий управления потоками заявок в СМО, определяющих правило занесения заявок разных классов в накопители ограниченной ёмкости (дисциплина буферизации) и правило выбора их из очереди на обслуживание (дисциплина обслуживания).

3.3.2. Обозначения СМО (символика Кендалла)

Для компактного описания систем массового обслуживания часто используются обозначения, предложенные Д. Кендаллом [9], в виде:

A/B/N/L ,

где **A** и **B** – задают законы распределений соответственно интервалов времени между моментами поступления заявок в систему и длительности обслуживания заявок в приборе; **N** – число обслуживающих приборов в системе $(N = 1, 2, ..., \infty)$; **L** – число мест в накопителе, которое может принимать значения 0, 1, 2, … (отсутствие **L** означает, что накопитель имеет неограниченную ёмкость).

Для задания законов распределений **A** и **B** используются следующие обозначения:

G (General) – произвольное распределение общего вида;

M (Markovian) – экспоненциальное (показательное) распределение;

D (Deterministik) – детерминированное распределение;

U (Uniform) – равномерное распределение;

$\mathbf{E_k}$ (Erlangian) – распределение Эрланга *k*-го порядка (с *k* последовательными одинаковыми экспоненциальными фазами);

$\mathbf{h_k}$ (hipoexponential) – гипоэкспоненциальное распределение *k*-го порядка (с *k* последовательными разными экспоненциальными фазами);

$\mathbf{H_r}$ (Hiperexponential) – гиперэкпоненциальное распределение порядка *r* (с *r* параллельными экспоненциальными фазами);

g (gamma) – гамма-распределение;

P (Pareto) – распределение Парето и т.д.

Примеры:

M/M/1 – одноканальная СМО с накопителем неограниченной ёмкости, в которую поступает однородный поток заявок с экспоненциальным

распределением интервалов времени между последовательными заявками (простейший поток) и экспоненциальной длительностью обслуживания заявок в приборе.

M/G/3/10 – трёхканальная СМО с накопителем ограниченной ёмкости, равной 10, в которую поступает однородный поток заявок с экспоненциальным распределением интервалов времени между последовательными заявками (простейший поток) и длительностью обслуживания заявок, распределённой по закону общего вида.

D/E_2/7/0 – семиканальная СМО без накопителя (ёмкость накопителя равна 0), в которую поступает однородный поток заявок с детерминированными интервалами времени между последовательными заявками (детерминированный поток) и длительностью обслуживания заявок в приборе, распределённой по закону Эрланга 2-го порядка.

Для обозначения более сложных СМО дополнительно могут использоваться обозначения, описывающие неоднородный поток заявок и приоритеты между заявками разных классов.

3.3.3. Режимы функционирования СМО

СМО может работать в следующих режимах:

- **установившемся** или **стационарном**, когда вероятностные характеристики системы не изменяются со временем;
- **неустановившемся**, когда характеристики системы изменяются со временем, что может быть обусловлено:
  - *началом работы системы*, когда значения характеристик функционирования, меняясь со временем, стремятся в пределе к стационарным значениям (**переходной режим**);
  - *нестационарным характером* потока заявок и обслуживания в приборе (**нестационарный режим**).

Кроме этого, в некоторых системах, например в *СМО с накопителем неограниченной ёмкости*, неустановившийся режим функционирования может быть обусловлен *перегрузкой системы*, когда интенсивность поступления заявок превышает интенсивность обслуживания, и система не справляется с возлагаемой на нее нагрузкой (**режим перегрузки**). При этом характеристики функционирования СМО с течением времени растут неограниченно. В частности, длина очереди перед прибором с течением времени становится всё больше и в пределе стремится к бесконечности.

Обычно исследование СМО с накопителем неограниченной ёмкости проводится в предположении о существовании установившегося режима, непременным условием которого является требование отсутствия перегрузок, для чего необходимо, чтобы интенсивность поступления заявок была меньше, чем интенсивность обслуживания. Это требование записывается для одноканальных СМО в виде условия:

$\lambda < \mu$ или $\lambda b < 1$.

Для многоканальных СМО аналогичное условие имеет вид:

$$\lambda < K\mu \quad \text{или} \quad \frac{\lambda b}{K} < 1,$$

где K – число обслуживающих приборов, а значение $K\mu$ представляет собой суммарную интенсивность обслуживания заявок в K-канальной СМО

В СМО *с накопителем ограниченной ёмкости* превышение интенсивности поступления заявок над суммарной интенсивностью обслуживания не приводит к неограниченному росту длины очереди, что обусловлено потерей заявок. Следовательно, в СМО с накопителем ограниченной ёмкости перегрузки не приводят к работе системы в неустановившемся режиме, а приводят лишь к росту числа потерянных заявок. При этом потеря части поступающих в систему заявок при наличии накопителя ограниченной ёмкости может рассматриваться как один из механизмов борьбы с перегрузками.

3.3.4. Характеристики СМО с однородным потоком заявок

Характеристики систем со стохастическим характером функционирования являются *случайными величинами* и полностью описываются соответствующими законами распределений. На практике при моделировании часто ограничиваются определением только *средних значений* (математических ожиданий), реже – определением двух первых моментов этих характеристик.

В качестве основных характеристик СМО с однородным потоком заявок используются следующие величины:

- ***нагрузка*** системы:

$$\boxed{y = \lambda / \mu = \lambda b}; \qquad (3.6)$$

- ***коэффициент загрузки*** или просто ***загрузка*** системы, определяемая как доля времени, в течение которого система (в случае одноканальной СМО – прибор) работает, то есть выполняет обслуживание заявок; загрузка может быть рассчитана как отношение *среднего* времени T_p работы одного прибора многоканальной СМО, к общему времени наблюдения T:

$$\rho = \lim_{T \to \infty} \frac{T_p}{T}; \qquad (3.7)$$

время T_p для СМО с K обслуживающими приборами определяется путём усреднения времени работы по всем приборам:

$$T_p = \frac{1}{K} \sum_{i=1}^{K} T_i ,$$

где T_i - время работы прибора $i = \overline{1, K}$;

подставляя последнее выражение в (3.7) окончательно получим:

$$\rho = \lim_{T\to\infty} \frac{1}{KT} \sum_{i=1}^{K} T_i ;$$

очевидно, что $0 \le \rho \le 1$;

- ***коэффициент простоя*** системы:

$$\eta = 1 - \rho ; \qquad (3.8)$$

- ***вероятность потери*** заявок:

$$\pi_n = \lim_{T\to\infty} \frac{N_n(T)}{N(T)}, \qquad (3.9)$$

где T – время работы системы (наблюдения за системой); $N(T)$ – число заявок, поступивших в систему за время T; $N_n(T)$ – число потерянных заявок за время T;

- **вероятность обслуживания** заявки, то есть вероятность того, что поступившая в систему заявка будет обслужена:

$$\pi_0 = (1 - \pi_n) = \lim_{T\to\infty} \frac{N_0(T)}{N(T)}, \qquad (3.10)$$

где $N_0(T)$ – число обслуженных в системе заявок за время T, причем $N_n(T) + N_0(T) = N(T)$ и $\pi_0 + \pi_n = 1$;

- ***производительность*** системы, представляющая собой *интенсивность потока обслуженных заявок*, выходящих из системы:

$$\lambda' = \pi_0 \lambda = (1 - \pi_n)\lambda ; \qquad (3.11)$$

для *СМО с накопителем неограниченной ёмкости*, при условии отсутствия перегрузок, вероятность потери заявок $\pi_n = 0$ и, следовательно, производительность системы совпадает с интенсивностью поступления заявок в систему: $\lambda' = \lambda$;

- **интенсивность потока потерянных** (не обслуженных) **заявок** из-за ограниченной ёмкости накопителя:

$$\lambda'' = \pi_n \lambda = (1 - \pi_0)\lambda ; \qquad (3.12)$$

очевидно, что сумма интенсивностей потоков обслуженных и потерянных заявок должна быть равна интенсивности входящего в систему потока заявок: $\lambda' + \lambda'' = \lambda$;

- **среднее время ожидания** заявок в очереди: ***w***;
- **среднее время пребывания** заявок в системе, складывающееся из времени ожидания *w* и времени обслуживания *b*:

$$\boxed{u = w + b}; \qquad (3.13)$$

- **средняя длина очереди** заявок:

$$\boxed{l = \lambda' w}; \qquad (3.14)$$

- **среднее число заявок в системе** (в очереди и на обслуживании в *приборе*):

$$\boxed{m = \lambda' u}. \qquad (3.15)$$

Нагрузка и загрузка являются важнейшими характеристиками СМО, определяющими качество функционирования системы.

Нагрузка $y = \lambda b$ представляет собой интегральную оценку, объединяющую два нагрузочных параметра: частоту использования некоторого ресурса (прибора СМО), задаваемую в виде интенсивности λ поступления заявок в СМО, и время использования этого ресурса, задаваемое в виде средней длительности b обслуживания заявок в СМО. Нагрузка показывает количество работы, которую необходимо выполнить в системе. Если значение нагрузки $y < 1$, то заданная нагрузка может быть выполнена одним обслуживающим прибором, то есть одноканальная СМО будет работать без перегрузки. Если $y > 1$, то реализация заданной нагрузки в одноканальной СМО приведет к режиму перегрузки, означающему, что с течением времени всё большее число заявок будет оставаться не обслуженным, и в случае накопителя неограниченной емкости очередь заявок будет расти неограниченно. Для того чтобы система работала без перегрузок необходимо использовать многоканальную СМО, количество приборов которой должно быть больше, чем значение нагрузки: $K > y$.

В общем случае для любой СМО (с накопителем ограниченной и неограниченной ёмкости) загрузка системы может быть рассчитана через нагрузку следующим образом:

$$\boxed{\rho = \min\left(\frac{(1-\pi_n)y}{K}; 1\right)}, \qquad (3.16)$$

где K – число обслуживающих приборов в СМО; π_n – вероятность потери заявок.

Последнее выражение можно трактовать следующим образом: $\rho = \frac{(1-\pi_n)y}{K}$, если СМО работает без перегрузки, и $\rho = 1$, если СМО перегружена.

Покажем, что выражение (3.16) соответствует определению (3.7).

Рассмотрим достаточно большой промежуток времени $T \to \infty$, в течение которого работает СМО. За это время в систему поступит в среднем λT заявок, где λ – интенсивность поступления заявок в СМО, из которых будут обслужены системой $(1-\pi_n)\lambda T$ заявок ($\pi_n \lambda T$ заявок будут потеряны из-за ограниченной ёмкости накопителя). Обслуживание всех этих заявок будет длиться в течение времени $T_p = (1-\pi_n)\lambda T b$, если СМО – одноканальная, и в течение времени $T_p = \frac{(1-\pi_n)\lambda T b}{K}$, если СМО – многоканальная и содержит K обслуживающих приборов. Здесь b – средняя длительность обслуживания заявки в приборе.

Подставляя выражение для T_p в (3.7), получим:

$$\rho = \lim_{T\to\infty}\frac{T_p}{T} = \lim_{T\to\infty}\frac{(1-\pi_n)\lambda T b}{KT} = \frac{(1-\pi_n)\lambda b}{K} = \frac{\lambda' b}{K}, \qquad (3.17)$$

где $\lambda' = (1-\pi_n)\lambda$ – интенсивность обслуженных в СМО заявок.

Отметим, что загрузка системы, в отличие от нагрузки, определяется через интенсивность только *обслуженных* заявок, поскольку потерянные заявки не обслуживаются в приборах и, следовательно, не загружают систему.

Рассмотрим теперь СМО с накопителем неограниченной ёмкости и вспомним, что при возникновении перегрузок такая система не справляется с работой, что выражается в неограниченном росте очереди с течением времени.

Если $T_p < T$, то это означает, что система справляется с работой, то есть работает без перегрузок.

Если же время $T_p = \frac{\lambda T b}{K}$, которое требуется для обслуживания всех заявок, окажется больше, чем время наблюдения за системой $T_p > T$, то это означает, что система не справляется с нагрузкой, то есть работает в режиме перегрузки. В этом случае загрузка системы $\rho = 1$ (составляет 100%), а коэффициент простоя соответственно равен нулю.

Выражение (3.16) записано с учётом указанного обстоятельства.

Получим ещё одну полезную формулу для расчёта вероятности потери заявок по известному значению загрузки СМО.

Из (3.11) следует, что вероятность потери заявок в СМО с накопителем ограниченной ёмкости может быть рассчитана как

$$\pi_n = \frac{\lambda - \lambda'}{\lambda} = 1 - \frac{\lambda'}{\lambda}.$$

В то же время из (3.17) вытекает, что интенсивность обслуженных заявок

$$\lambda' = \frac{\rho K}{b}.$$

Подставляя последнее выражение в предыдущее, получим:

$$\boxed{\pi_n = 1 - \frac{\rho K}{\lambda b} = 1 - \frac{\rho}{y}K}, \qquad (3.18)$$

где $y = \lambda b$ – нагрузка системы.

Вероятность обслуживания поступившей в систему заявки:

$$\pi_0 = 1 - \pi_n = \frac{\rho}{y}K.$$

Выражение (3.18) оказывается полезным при расчёте характеристик обслуживания заявок в марковских моделях систем и сетей массового обслуживания (см. примеры в разделе 5).

Зависимости (3.14) и (3.15), связывающие средние значения временны*х* (w, u) и безразмерных (l, m) характеристик, известны как **формулы Литтла** и вместе с формулой (3.13) представляют собой *фундаментальные зависимости*, справедливые для широкого класса моделей массового обслуживания.

Из (3.15) можно получить зависимость, связывающую среднее число заявок в системе со средней длиной очереди заявок:

$$m = \lambda u = \lambda(w + b) = \lambda w + \lambda b = l + y,$$

откуда следует, что *нагрузка* $y = \lambda b$ *характеризует среднее число заявок, находящихся на обслуживании*.

При условии отсутствия перегрузок в одноканальной СМО загрузка совпадает с нагрузкой: $\rho = y = \lambda b$ и тогда $m = l + \rho$, то есть загрузку одноканальной СМО можно трактовать как среднее число заявок, находящихся на обслуживании в приборе. Отметим, что на обслуживании находится не одна заявка, как может показаться, а меньше единицы: $\rho < 1$. Это действительно так, если вспомнить, что речь идёт о *среднем* числе находящихся на обслуживании заявок, которое может быть рассчитано следующим образом. В приборе в каждый момент времени может находиться случайное число заявок, принимающее два значения: 1, если прибор работает, то есть обслуживает заявку, и 0, если прибор простаивает. Поскольку значение загрузки лежит в интервале от 0 до 1 ($0 \le \rho \le 1$) и показывает долю времени, в течение которого прибор работает, то *загрузку можно трактовать как вероятность того, что прибор работает*, а величину $\eta = (1 - \rho)$ – как вероятность простоя прибора. Тогда математическое ожидание случайной величины, принимающей значения 1 с вероятностью ρ и 0 с вероятностью $(1 - \rho)$, будет равно: $\rho \times 1 + (1 - \rho) \times 0 = \rho$, что и требовалось показать.

Обычно исследование систем проводится в предположении о стационарности входящего потока заявок и длительности обслуживания. В этом случае *условие существования установившегося режима* для СМО с накопителем неограниченной ёмкости совпадает с *условием отсутствия перегрузок* в СМО и записывается в виде: $\rho < 1$.

3.3.5. Характеристики СМО с неоднородным потоком заявок

Для СМО с неоднородным потоком заявок, в которую поступают H классов заявок с интенсивностями $\lambda_1, \ldots, \lambda_H$ и средними длительностями обслуживания $b_1, \ldots, b_H$, определяются две группы характеристик обслуживания заявок:

- характеристики по каждому классу (потоку) заявок;
- характеристики объединённого (суммарного) потока заявок.

Характеристики по каждому классу заявок $i = \overline{1, H}$ идентичны характеристикам СМО с однородным потоком:

- *нагрузка*, создаваемая заявками класса i: $y_i = \lambda_i / \mu_i = \lambda_i b_i$;
- *вероятность потери* заявок: π_{n_i};
- *вероятность обслуживания* заявки: $\pi_{0_i} = (1 - \pi_{n_i})$;
- *интенсивность* потока *обслуженных* заявок (производительность по i-му классу заявок): $\lambda_{0_i} = \pi_{0_i} \lambda_i = (1 - \pi_{n_i}) \lambda_i$;
- *интенсивность* потока *потерянных* заявок: $\lambda_{n_i} = \pi_{n_i} \lambda_i$.
- *загрузка* системы, создаваемая заявками класса i:

$\rho_i = \min\left(\dfrac{(1 - \pi_{n_i}) y_i}{K}; 1 \right)$, где π_{n_i} – вероятность потери заявок класса i из-за ограниченной ёмкости накопителя ($\pi_{n_i} = 0$, если ёмкость накопителя – неограниченная); K – число обслуживающих приборов в СМО;

- *время ожидания* заявок в очереди: w_i;
- *время пребывания* заявок в системе: $u_i = w_i + b_i$;
- *длина очереди* заявок: $l_i = \lambda_i w_i$;
- *число заявок в системе* (в очереди и на обслуживании): $m_i = \lambda_i u_i$.

Характеристики объединённого (суммарного) потока заявок позволяют определить усредненные по всем классам заявок показатели эффективности функционирования СМО:

- *суммарная интенсивность* поступления заявок в систему (интенсивность суммарного потока):

$$\Lambda = \sum_{i=1}^{H} \lambda_i ; \qquad (3.19)$$

- *суммарная нагрузка* Y и *суммарная загрузка* R системы:

$$Y = \sum_{i=1}^{H} y_i ; \qquad R = \min\left(\sum_{i=1}^{H} \rho_i ; 1 \right), \qquad (3.20)$$

причем условие отсутствия перегрузок в СМО с неоднородным потоком заявок и накопителем неограниченной ёмкости имеет вид:

$$R < 1; \qquad (3.21)$$

- *коэффициент простоя* системы: $\eta = 1 - R$;
- *среднее время ожидания* W и *среднее время пребывания* U заявок объединённого потока в системе:

$$W = \sum_{i=1}^{H} \xi_i w_i ; \qquad U = \sum_{i=1}^{H} \xi_i u_i , \qquad (3.22)$$

где $\xi_i = \lambda_i / \Lambda$ – коэффициент, учитывающий долю заявок класса i в суммарном потоке, который может трактоваться как *вероятность того, что поступившая в систему заявка принадлежит классу* i;

- *суммарная длина очереди* и *суммарное число заявок в системе*:

$$L = \sum_{i=1}^{H} l_i ; \qquad M = \sum_{i=1}^{H} m_i . \tag{3.23}$$

Можно доказать, что для характеристик объединённого (суммарного) потока справедливы те же фундаментальные соотношения (3.13) – (3.15) , что и для однородного потока:

$U = W + B$; $L = \Lambda W$; $M = \Lambda U$,

где B – среднее время обслуживания любой заявки суммарного потока:

$$B = \sum_{i=1}^{H} \xi_i b_i .$$

3.4. Параметры и характеристики СеМО

3.4.1. Параметры СеМО

Для описания *линейных разомкнутых и замкнутых однородных экспоненциальных* СеМО используется следующая совокупность параметров:

- *число узлов* в сети: n;
- *число обслуживающих приборов* в узлах сети: $K_1,...,K_n$;
- *матрица вероятностей передач*: $\mathbf{P} = [p_{ij} \mid i, j = 0,1,\ldots,n]$, где p_{ij} – вероятность передачи заявки из узла i в узел j;
- *интенсивность* λ_0 источника заявок, поступающих в **разомкнутую** СеМО (РСеМО), или *число заявок* M, циркулирующих в **замкнутой** СеМО (ЗСеМО);
- средние *длительности обслуживания* заявок в узлах сети: $b_1,\ldots,b_n$.

Заметим, что состав параметров разомкнутых и замкнутых СеМО различается только одним параметром, а именно: для ЗСеМО, в отличие от РСеМО, вместо интенсивности λ_0 поступления заявок в сеть необходимо задать число постоянно циркулирующих в сети заявок M.

Для *линейных* СеМО элементы матрицы вероятностей передач должны удовлетворять условию:

$$\sum_{j=0}^{n} p_{ij} = 1 \quad (i = \overline{0,n}). \tag{3.24}$$

Это условие отражает тот факт, что любая заявка, покинувшая некоторый узел, обязательно (с вероятностью 1) перейдёт в какой-то узел, включая тот же самый или нулевой. Переход заявки в нулевой узел означает, что заявка покинула сеть.

В случае *неэкспоненциальных* разомкнутых СеМО дополнительно необходимо задать законы распределения или, по крайней мере, вторые моменты интервалов времени между поступающими в разомкнутую сеть заявками и длительностей обслуживания заявок в узлах сети.

В случае *неоднородных* СеМО необходимо дополнительно задать количество классов заявок H в сети и для каждого класса – матрицы вероятностей передач **P(*h*)**, интенсивности $\lambda_0(h)$ или число заявок $M(h)$, а также средние длительности обслуживания $b_i(h)$ заявок класса $h=\overline{1,H}$ в узле $i=\overline{1,n}$. При необходимости могут быть заданы законы распределений интервалов между поступающими в РСеМО заявками и законы распределений длительностей обслуживания заявок разных классов в узлах сети.

3.4.2. Режимы функционирования СеМО

СеМО, как и СМО, может работать в установившемся и неустановившемся режимах. Последний может быть связан с началом работы системы (переходной режим), нестационарным характером потока заявок и обслуживания в приборе (нестационарный режим) и перегрузкой системы (режим перегрузки).

Очевидно, что для СеМО, как и для СМО, при использовании предположения о стационарности входящего потока заявок и длительностей обслуживания заявок в узлах *условие существования установившегося режима* совпадает с *условием отсутствия перегрузок*.

Рассмотрим это условие для разомкнутой и замкнутой СеМО.

Очевидно, что перегрузки в **разомкнутой СеМО** отсутствуют, если каждый узел сети работает без перегрузок. Если же хотя бы один из узлов сети не справляется с нагрузкой, то длина очереди в этом узле начнет увеличиваться до бесконечности и, следовательно, суммарное число заявок в РСеМО будет расти неограниченно.

Таким образом, для того чтобы в *разомкнутой* СеМО не было перегрузок, необходимо отсутствие перегрузок во всех узлах РСеМО, то есть загрузка ρ_j любого узла j ($j=\overline{1,n}$) должна быть строго меньше единицы:

$$\rho_j=\frac{\lambda_j b_j}{K_j}=\frac{\alpha_j\lambda_0 b_j}{K_j}<1 \qquad \text{для всех } j=\overline{1,n}.$$

Из последнего неравенства имеем:

$$\lambda_0<\frac{K_j}{\alpha_j b_j} \qquad \text{для всех } j=\overline{1,n}.$$

Это условие может быть записано также в следующем виде:

$$\boxed{\lambda_0<\min\left(\frac{K_1}{\alpha_1 b_1},\frac{K_2}{\alpha_2 b_2},...,\frac{K_n}{\alpha_n b_n}\right)}. \qquad (3.25)$$

Полученное условие налагает ограничение сверху на интенсивность поступления заявок в РСеМО из внешнего источника. Узлы, в которых указанное условие не выполняется, являются перегруженными. С течением времени это приводит к неограниченному росту числа заявок в сети, которые скапливаются в перегруженных узлах, имеющих накопители неограниченной ёмкости.

В дальнейшем при исследовании разомкнутых СеМО, если не оговорено другое, будем полагать, что в сети существует установившийся режим.

Несколько иначе дело обстоит для **замкнутых СеМО**. Поскольку в ЗСеМО циркулирует постоянное число заявок, то в узлах сети не могут образовываться очереди бесконечной длины, следовательно, в ЗСеМО всегда существует установившийся режим. Даже если в сети имеется очень «медленный» узел, в котором по сравнению с другими узлами слишком долго обрабатываются заявки, то это может привести только к тому, что все заявки будут постоянно скапливаться в очереди перед данным узлом, однако их количество будет всегда конечно и в пределе равно числу циркулирующих в сети заявок. Загрузка такого «медленного» узла будет близка к единице, поскольку постоянное наличие очереди перед этим узлом обусловливает непрерывную работу приборов узла. Такой узел обычно представляет собой так называемое «узкое место» сети.

3.4.3. Характеристики СеМО

Характеристики СеМО делятся на два класса:

- **узловые**, описывающие эффективность функционирования отдельных узлов СеМО;
- **сетевые**, описывающие функционирование СеМО в целом.

Состав *узловых характеристик* СеМО, работающей *в стационарном режиме*, такой же, как и для СМО, и для узла $j=\overline{1,n}$ включает в себя следующие характеристики:

- *нагрузка* узла: $y_j = \lambda_j b_j = \alpha_j \lambda_0 b_j$;
- *загрузка* узла: $\rho_j = \dfrac{y_j}{K_j} = \dfrac{\alpha_j \lambda_0 b_j}{K_j}$, причем $\rho_j < 1$;
- *коэффициент простоя* узла: $\eta_j = 1 - \rho_j$;
- *время ожидания* заявок в узле: w_j;
- *время пребывания* заявок в узле: $u_j = w_j + b_j$;
- *длина очереди* заявок узле: $l_j = \lambda_j w_j = \alpha_j \lambda_0 w_j$;
- *число заявок в* узле (в очереди и на обслуживании):

$m_j = \lambda_j u_j = \alpha_j \lambda_0 (w_j + b_j) = l_j + y_j$.

В приведенных выше формулах использован тот факт, что в линейных СеМО интенсивность поступления заявок в любой узел связана с интенсивностью источника соотношением (3.5).

На основе узловых характеристик рассчитываются *сетевые характеристики* СеМО:

- **суммарная нагрузка** во всех узлах, характеризующая *среднее число заявок, одновременно находящихся на обслуживании во всех узлах* сети:

$$Y = \sum_{j=1}^{n} y_j ,$$

где y_j – нагрузка узла j, причем $0 < Y \le \sum_{j=1}^{n} K_j$;

- **суммарная загрузка** всех узлов СеМО, характеризующая *среднее число параллельно работающих узлов* сети:

$$R = \sum_{j=1}^{n} \rho_j ,$$

где ρ_j – загрузка узла j, причем $0 < R \le n$;

- среднее число заявок, находящихся в очередях всех узлов сети и ожидающих обслуживания:

$$\boxed{L = \sum_{j=1}^{n} l_j}, \qquad (3.26)$$

где l_j – средняя длина очереди заявок в узле j;

- среднее число заявок, находящихся в сети:

$$\boxed{M = \sum_{j=1}^{n} m_j}, \qquad (3.27)$$

где m_j – среднее число заявок в узле j, причём для замкнутых сетей это выражение может быть использовано для проверки правильности проведенных расчетов, так как для них число заявок M в сети задано;

- среднее время ожидания заявок в сети:

$$\boxed{W = \sum_{j=1}^{n} \alpha_j w_j}, \qquad (3.28)$$

где w_j – среднее время ожидания заявок в узле j; α_j – коэффициент передачи для узла j, показывающий среднее число попаданий заявки в узел j за время её нахождения в сети; $W_j = \alpha_j w_j$ – представляет собой суммарное (полное) время ожидание заявки в узле j за время её нахождения в сети;

- среднее время пребывания заявок в сети:

$$\boxed{U = \sum_{j=1}^{n} \alpha_j u_j}, \qquad (3.29)$$

где u_j – среднее время пребывания заявок в узле j; $U_j = \alpha_j u_j$ – суммарное (полное) время пребывания заявки в узле j за время её нахождения в сети;

- **производительность замкнутой СеМО** λ_0, определяемая как интенсивность потока заявок, проходящих через выделенный нулевой узел замкнутой сети, и представляющая собой среднее число заявок, обслуженных в ЗСеМО за единицу времени; производительность ЗСеМО может быть рассчитана на основе выражения (3.5), из которого следует:

$$\boxed{\lambda_0 = \lambda_j / \alpha_j \quad (j = 1, \ldots, n)}; \tag{3.30}$$

Следует отметить, что для сетевых характеристик СеМО выполняются те же фундаментальные соотношения, что и для СМО, а именно:

$$L = \lambda_0 W; \tag{3.31}$$

$$M = \lambda_0 U; \tag{3.32}$$

$$M = L + Y; \tag{3.33}$$

$$U = W + B, \tag{3.34}$$

где $B = \sum_{j=1}^{n} \alpha_j b_j$ – суммарное время обслуживания заявки во всех узлах за время ее нахождения в сети.

Выражения (3.31) и (3.32) представляют собой формулы Литтла для расчёта сетевых характеристик СеМО.

Из (3.32) может быть получена ещё одна важная *формула для расчёта производительности ЗСеМО*:

$$\boxed{\lambda_0 = \frac{M}{U}}. \tag{3.35}$$

Для неоднородной СеМО перечисленные характеристики определяются как для каждого класса в отдельности, так и для объединенного (суммарного) потока заявок.

3.5. Резюме

1. В качестве математических моделей дискретных систем со стохастическим характером функционирования широко применяются модели массового обслуживания (ММО), которые делятся на *базовые модели* в виде одноканальных и многоканальных *систем массового обслуживания* (СМО) и *сетевые модели* в виде разомкнутых и замкнутых *сетей массового обслуживания* (СеМО).

Для описания СМО используются следующие понятия: *заявка* (требование, запрос, вызов, клиент), *поток заявок*, *обслуживающий прибор* (или просто прибор), *обслуживание*, *длительность обслуживания*, *накопитель*, *ёмкость накопителя*, *очередь*, *длина очереди*, *дисциплина буферизации*, *дисциплина обслуживания*, *приоритет*.

Для описания СеМО дополнительно используются такие понятия как *узел*, *источник*, *граф СеМО*, *маршрут*.

2. Описание потока заявок в простейшем случае предполагает задание его *интенсивности***.** Поток заявок может быть *детерминированным (регулярным)* или *случайным*, *стационарным* или нестационарным, *ординарным* или неординарным (групповым), с последействием или *без последействия*.

Стационарный ординарный поток без последействия называется простейшим (пуассоновским). Интервалы времени между заявками в простейшем потоке распределены по *экспоненциальному закону*. Аналитические исследования моделей массового обслуживания обычно проводятся в предположении о простейшем потоке заявок, что обусловлено рядом присущих ему особенностей (*суммирование потоков, вероятностное разрежение потока*), позволяющих во многих случаях получить сравнительно простые аналитические зависимости характеристик от параметров.

Длительность обслуживания заявок в приборе в простейшем случае может быть задана средним значением или величиной обратной – *интенсивностью обслуживания*, характеризующей среднее число заявок, которое может быть обслужено прибором за единицу времени.

Стратегия управления потоками заявок задается в виде *дисциплины буферизации (ДБ)* и *дисциплины обслуживания (ДО)*, которые могут быть классифицированы по следующим признакам: наличие приоритетов между заявками разных классов; способ (режим) вытеснения заявок из очереди или назначения заявок на обслуживание; правило вытеснения или выбора заявок на обслуживание; возможность изменения приоритетов.

Среди дисциплин обслуживания заявок в технических системах наибольшее распространение получили: бесприоритетная дисциплина *обслуживания в порядке поступления* (ОПП или FIFO) и приоритетные дисциплины: с *относительными* (ОП) и *абсолютными* (АП) *приоритетами*, которые могут быть *статическими* или *динамическими*.

3. Большинство СМО, используемых в качестве базовых моделей реальных систем, могут быть классифицированы: по числу мест в накопителе (*без накопителя* – СМО с отказами; *с накопителем ограниченной ёмкости* – СМО с потерями; *с накопителем неограниченной ёмкости* – СМО без потерь); по количеству обслуживающих приборов (*одноканальные* и *многоканальные*); по количеству классов заявок (с *однородным* и *неоднородным* потоком заявок).

Заявки относятся *к разным классам*, если они в моделируемой реальной системе различаются *длительностью обслуживания и/или приоритетами*.

4. Сетевые модели (СеМО) могут быть классифицированы: в зависимости от характера процессов поступления и обслуживания заявок (*стохастические*, *детерминированные*); по виду зависимостей, связывающих интенсивности потоков заявок в разных узлах СеМО (*линейные*, *нелиней-*

ные); по числу циркулирующих в сети заявок (*разомкнутые*, *замкнутые*, замкнуто-разомкнутые); по типу циркулирующих заявок (*однородные*, *неоднородные*).

В *линейных* СеМО интенсивность потока заявок в любом узле связана линейной зависимостью с интенсивностью источника через коэффициент передачи, который показывает *среднее количество попаданий заявки в данный узел за время ее нахождения в сети*.

В *нелинейных* СеМО интенсивности потоков заявок в узлах связаны нелинейными зависимостями. *Нелинейность СеМО* может быть обусловлена *потерей заявок* или *размножением заявок* в сети.

Разомкнутая СеМО содержит один или несколько *внешних независимых источников* заявок, причем в сети одновременно может находиться *любое число заявок*.

Замкнутая СеМО, в отличие от разомкнутой, не содержит *независимых внешних* источников заявок и характеризуется тем, что в ней циркулирует *постоянное число заявок M*.

5. Для компактного описания СМО используются обозначения в виде **A/B/N/L** , где **A** и **B** – задают законы распределений соответственно интервалов времени между моментами поступления заявок и длительностей обслуживания в приборе; **N** – число обслуживающих приборов в системе; **L** – число мест в накопителе.

6. Для описания СМО, в простейшем случае, используются следующие *параметры*:

- количество обслуживающих приборов *K*;
- количество *k* и емкости накопителей E_j ($j = \overline{1,k}$);
- количество поступающих в систему классов заявок *H*;
- интенсивность λ_i потока и коэффициент вариации ν_{a_i} интервалов времени между поступающими в систему заявками класса $i = \overline{1,H}$;
- среднее значение b_i и коэффициент вариации ν_{b_i} длительности обслуживания заявок класса $i = \overline{1,H}$;
- дисциплина буферизации и дисциплина обслуживания заявок.

СМО может работать в *установившемся* (*стационарном***)** режиме или в *неустановившемся* (переходном или нестационарном режиме). Кроме того, *СМО* может работать в *режиме перегрузки*, когда система не справляется с нагрузкой. При этом характеристики функционирования СМО *с накопителем неограниченной емкости* с течением времени растут неограниченно. Для того чтобы в такой СМО не было перегрузок, необходимо, чтобы нагрузка системы была меньше, чем число обслуживающих приборов, или, что то же самое, загрузка системы была

строго меньше единицы. В СМО *с накопителем ограниченной ёмкости* перегрузки не приводят к неустановившемуся режиму.

7. Характеристики систем со стохастическим характером функционирования являются случайными величинами и полностью описываются соответствующими законами распределений. На практике при моделировании часто ограничиваются определением только средних значений (математических ожиданий), реже – определением двух первых моментов этих характеристик.

В качестве основных *характеристик СМО* с однородным потоком заявок используются:

- *нагрузка* системы: $y = \lambda / \mu = \lambda b$;
- *загрузка* системы: $\rho = \min\left(\frac{(1-\pi_n)y}{K}; 1 \right)$;
- *коэффициент простоя* системы: $\eta = 1 - \rho$;
- *вероятность потери* заявок: $\pi_n = \lim\limits_{T \to \infty} \frac{N_n(T)}{N(T)}$;
- *вероятность обслуживания* заявки: $\pi_0 = (1-\pi_n);$;
- *производительность системы*: $\lambda' = \pi_0 \lambda = (1-\pi_n)\lambda$;
- *интенсивность потока потерянных заявок*: $\lambda'' = \pi_n \lambda = (1-\pi_0)\lambda$;
- *среднее время ожидания* заявок в очереди: $w = ?$ (подлежит определению для каждой конкретной СМО);
- *среднее время пребывания* заявок в системе: $u = w + b$;
- *средняя длина очереди* заявок: $l = \lambda' w$;
- *среднее число заявок* в системе: $m = \lambda' u$.

Для СМО с неоднородным потоком заявок определяются две группы характеристик обслуживания заявок: характеристики *по каждому классу заявок* и характеристики *суммарного (объединенного) потока* заявок.

8. Для описания линейных разомкнутых и замкнутых однородных экспоненциальных СеМО необходимо задать следующие *параметры*:

- число узлов в сети n;
- число обслуживающих приборов в узлах сети $K_1,...,K_n$;
- матрицу вероятностей передач $\mathbf{P} = [\, p_{ij} \mid i, j = 0, 1, \ldots, n]$;
- *интенсивность* λ_0 источника заявок, поступающих в РСеМО, или *число заявок* M , циркулирующих в ЗСеМО;
- средние длительности обслуживания заявок в узлах сети $b_1,...,b_n$.

СеМО, как и СМО, может работать в установившемся и неустановившемся режимах. Последний может быть связан с началом работы системы (переходной режим), нестационарным характером

процессов поступления и обслуживания заявок в приборе (нестационарный режим), а в разомкнутой СеМО, кроме того, перегрузкой системы (режим перегрузки). Условие отсутствия перегрузок в разомкнутой СеМО предполагает отсутствие перегрузок в каждом из узлов сети. *В замкнутой СеМО перегрузки не возникают*.

9. Характеристики СеМО делятся на узловые и сетевые. Состав узловых характеристик СеМО, работающей в стационарном режиме, такой же, как и для СМО. На основе узловых характеристик рассчитываются средние значения *сетевых характеристик* СеМО:

- суммарная нагрузка и загрузка: $Y = \sum_{j=1}^{n} y_j \quad R = \sum_{j=1}^{n} \rho_j$;
- среднее суммарное число заявок, находящихся во всех очередях сети: $L = \sum_{j=1}^{n} l_j$;
- среднее суммарное число заявок, находящихся в разомкнутой сети (во всех узлах): $M = \sum_{j=1}^{n} m_j$;
- среднее время ожидания и пребывания заявок в сети: $W = \sum_{j=1}^{n} \alpha_j w_j$; $W = \sum_{j=1}^{n} \alpha_j w_j$;
- производительность замкнутой СеМО: $\lambda_0 = \frac{M}{U}$.

Сетевые характеристики СеМО связаны между собой теми же фундаментальными соотношениями, что и характеристики СМО.

Для неоднородной СеМО перечисленные характеристики определяются как для каждого класса в отдельности, так и для объединенного (суммарного) потока заявок.

3.6. Практикум: обсуждение и решение задач

В разделе 3 рассмотрены модели массового обслуживания: СМО и СеМО, выполнена их классификация, перечислены параметры и рассчитываемые на их основе характеристики функционирования СМО и СеМО различных классов, приведены основные зависимости для расчета указанных характеристик.

Как и ранее, в процессе обсуждения представленного материала попытаемся ответить на некоторые конкретные вопросы практического характера.

Вопрос 1. Почему математическая модель называется абстрактной?

Обсуждение. Действительно, все математические модели являются абстрактными, собственно, как и сама математика. Абстрактность обусловлена переходом от параметров и характеристик реальной системы к её описанию в терминах определённого математического аппарата, например теории массового обслуживания. Затем выполняется анализ характеристик и исследование свойств этой математической модели, а полученные результаты интерпретируются применительно к реальной системе. Абстрактность математической модели состоит в том, что полученные с её помощью результаты могут быть применены к любой другой реальной системе, которая может быть представлена такой же моделью. Другими словами, одна и та же математическая модель может отображать функционирование совершенно разных по своей природе реальных систем, описываемых с помощью различных структурно-функциональных и нагрузочных параметров, состав и перечень которых определяются соответствующей прикладной областью.

Вопрос 2. Насколько предположение о простейшем характере потока заявок соответствует реальности?

Обсуждение. Простейший поток заявок является математическим представлением некоторого «идеального» потока, обладающего рядом замечательных свойств, благодаря которым для многих математических моделей удаётся получить достаточно простые аналитические зависимости, связывающие характеристики функционирования систем массового обслуживания с исходными параметрами. Одним из таких свойств является «отсутствие последействия», которое заключается в том, что поступление в систему очередной заявки не зависит от того, когда и сколько заявок поступило ранее. В реальной жизни наличие этого свойства означало бы следующее.

Представим, что вы, подходя к автобусной остановке, не успели на только что отправившийся автобус. Если поток автобусов, прибывающих на остановку, простейший, то в сложившейся ситуации это совсем не означает, что вам долго придётся ждать следующий автобус. Вполне возможно, что следующий автобус подойдет к остановке практически сразу. Точно так же, если вы пришли на автобусную остановку и застали большое число ожидающих пассажиров (что свидетельствует о том, что давно не было автобуса), то это совсем не означает, что скоро подойдет автобус. Кто-то скажет, что часто попадал в такие ситуации, и отсюда сделает вывод, что поток автобусов к остановке – простейший. В действительности же реальный поток автобусов может быть сколь угодно близок к простейшему, но не может быть простейшим по следующей причине. Если предположить, что поток автобусов к остановке – простейший, то существует (пусть и совсем ничтожная) вероятность того, что автобус вообще никогда не придёт, что, по всей видимости, невозможно (исключая случай, когда движение автобусов отменено, а все ожидающие

пассажиры не знали об этом). Наличие такой вероятности обусловлено тем, что интервалы времени между последовательными заявками (или автобусами) в простейшем потоке распределены по экспоненциальному закону, функция распределения которого ограничена слева (нулевым значением случайной величины), но не ограничена справа, то есть случайная величина, описывающая интервалы между последовательными заявками в простейшем потоке, может принимать сколь угодно большие значения, в том числе, равное бесконечности. Очевидно, что в реальных системах функция распределения обычно ограничена и справа.

Таким образом, отвечая на поставленный вопрос, можно сказать, что в реальной жизни вряд ли существует простейший поток. В то же время, многие реальные потоки могут быть достаточно близки к простейшему.

Вопрос 3. Когда оправдано использование предположения о простейшем характере потока заявок?

Обсуждение. Предположение о простейшем потоке широко используется не только из-за простоты получения математических зависимостей, но и по той причине, что многие реальные потоки близки к простейшим. Эта близость во многих случаях обусловлена следующим.

Во-первых, как сказано выше, суммирование (объединение) независимых *стационарных ординарных* потоков образует простейший поток при условии, что складываемые потоки оказывают более или менее одинаковое влияние на суммарный поток, причем на практике суммарный поток становится близким к простейшему уже при суммировании 5 потоков. Отметим, что к суммируемым потокам не предъявляется требование отсутствия последействия.

Во-вторых, можно показать, что стационарный ординарный поток заявок стремится к простейшему, если на него оказывает влияние множество случайных факторов. Именно этим можно объяснить близость потока автобусов, прибывающих на остановку, к простейшему. Действительно, если даже все автобусы отправляются с конечной остановки через одинаковые интервалы времени, то есть образуют детерминированный поток, то в процессе движения по улицам города интервалы между ними изменяются под влиянием многих, в основном случайных, факторов, таких как задержки перед светофорами, заторы и «транспортные пробки» на улицах, случайное время нахождения на остановках (зависящее от числа входящих и выходящих из автобуса пассажиров) и т.д. Всё это приводит к тому, что моменты прибытия к остановкам образуют случайный процесс, причем, чем ближе к конечной остановке, тем больше поток автобусов похож на простейший.

Предположение о простейшем характере входного потока заявок оправдано также в тех случаях, когда известно, что коэффициент вариации интервалов между последовательными заявками реального потока меньше единицы. В этом случае использование простейшего потока в модели

позволяет получить так называемые верхние оценки характеристик обслуживания заявок, гарантирующие, что в реальной системе значения характеристик будут не хуже, чем полученные на модели.

Вопрос 4. Почему в СМО с накопителем неограниченной емкости, работающей без перегрузок, возникают очереди? В каких случаях они не возникают?

Обсуждение. В СМО с накопителем неограниченной емкости перегрузки отсутствуют, если интенсивность поступления заявок меньше интенсивности обслуживания.

Рассмотрим случай, когда интенсивность поступления заявок равна 10 заявок в секунду, а интенсивность обслуживания – 1 заявка в секунду. За первую секунду в систему поступит 10 заявок, из которых будет обслужена одна заявка, а 9 – останутся в очереди. За вторую секунду в систему поступит ещё 10 заявок и одна заявка будет обслужена, в очереди окажется 18 заявок и т.д. Очевидно, что число заявок в очереди со временем будет возрастать до бесконечности, что свидетельствует о перегрузке системы, то есть система не справляется с нагрузкой.

Рассмотрим другой случай, когда интенсивность поступления заявок – 1 заявка в секунду, а интенсивность обслуживания – 10 заявок в секунду, или, что то же самое, средний интервал между последовательными заявками в потоке – 1 секунда, а средняя длительность обслуживания – 0,1 секунды. Таким образом, если заявки поступают с интервалом 1 секунда, а обслуживаются за 0,1 секунды, то возникает вопрос: откуда появляется очередь заявок?

Здесь следует обратить внимание на то, что речь идёт о *среднем* значении интервала между заявками и *среднем* значении длительности обслуживания. *Если процессы поступления и обслуживания заявок детерминированные, то очередь перед прибором не образуется*. Такие системы, естественно, не представляют интереса и не рассматриваются в теории массового обслуживания. Очередь появится только в том случае, если процесс поступления заявок в систему или процесс обслуживания их в приборе, или оба процесса – случайные. Тогда конкретное значение какого-то интервала между заявками может оказаться намного меньше среднего значения, например менее 0,1 секунды, а длительность обслуживания некоторой заявки – много больше среднего значения, например 2 секунды. Именно такие ситуации и приводят к появлению очереди перед прибором. Попутно отметим, что длина очереди – величина случайная, изменяющаяся случайным образом между нулём и некоторым максимальным значением.

Вопрос 5. Что в реальной системе может служить основанием для того, чтобы в соответствующей математической модели заявки были разделены на разные классы?

Обсуждение. Рассмотрим две модели обслуживания клиентов:

1) модель небольшого магазина, в котором только один продавец обслуживает покупателей, которыми являются и мужчины и женщины;

2) модель парикмахерской, в которой работает один мастер, делающий причёски мужчинам и женщинам.

Следует ли мужчин и женщин отнести к разным классам или же объединить их в модели в один класс?

Обе рассматриваемые модели представляют собой одноканальные СМО, в которых заявки соответствуют клиентам, а обслуживание заключается в затратах времени продавца или парикмахера на одного клиента.

В модели магазина мужчин и женщин при отсутствии у кого-нибудь из них преимущественного права (приоритета) на внеочередное обслуживание, скорее всего, можно объединить в один класс, поскольку время, затрачиваемее продавцом на одного покупателя примерно одинаково и не зависит от пола покупателя.

В парикмахерской, как известно, время, затрачиваемое на создание женской причёски много больше, чем на создание мужской причёски. В этом случае в модели парикмахерской заявки должны быть разбиты на два класса. Очевидно, что времена пребывания заявок разных классов в общем случае будут различаться, даже если их времена ожидания окажутся одинаковыми.

Вопрос 6. Когда в качестве модели реальной системы следует использовать разомкнутую, а когда замкнутую СеМО? Каким образом в замкнутой СеМО выбирается дуга, на которой отмечается точка «0»?

Обсуждение. Положим, что СеМО используется в качестве модели обслуживания покупателей в большом магазине с несколькими разными отделами, каждый из которых представляется в модели как узел сети. Покупатели в модели отображаются в виде заявок, перемещающихся между узлами СеМО.

Если количество покупателей, одновременно находящихся в магазине, может любым и принимать значения от 0 и, теоретически, до бесконечности, то в качестве модели такого магазина следует использовать разомкнутую СеМО.

Представим теперь, что мы хотим промоделировать работу этого магазина в час пик, когда в магазин стремится попасть большое число покупателей. Положим, что количество покупателей, которые могут одновременно находиться в магазине, определяется количеством корзинок или тележек, без которых вход в магазин запрещён. При отсутствии корзинок покупатели образуют очередь на входе и ожидают освобождения корзинок. Покупатель, покидающий магазин при выходе передает освободившуюся корзинку ожидающему на входе покупателю, который затем заходит в магазин. Таким образом, в магазине находится постоянное число покупателей, равное числу корзинок в магазине. Очевидно, что в

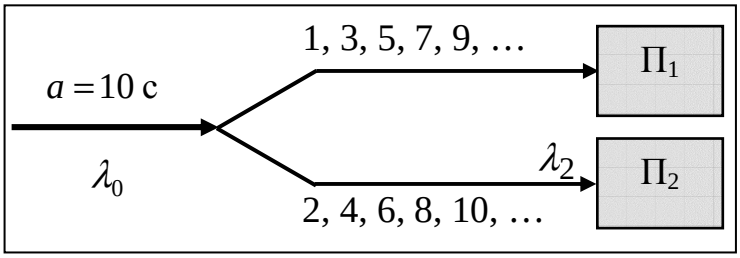
этом случае в качестве модели магазина должна использоваться замкнутая СеМО, а нулевая точка в модели должна быть выбрана на дуге, отображающей выход покупателя из магазина и вход нового покупателя.

Задача 1. В двухканальную СМО поступает простейший поток заявок со средним интервалом между соседними заявками 10 с, причем каждая вторая заявка направляется ко второму прибору. Чему равна интенсивность потока заявок ко второму прибору? Чему равен коэффициент вариации интервалов между заявками потока ко второму прибору?

Дано: СМО: $K = 2$; поток – простейший; $a = 10$ с.

Требуется:

- определить λ_2;
- определить ν_2.



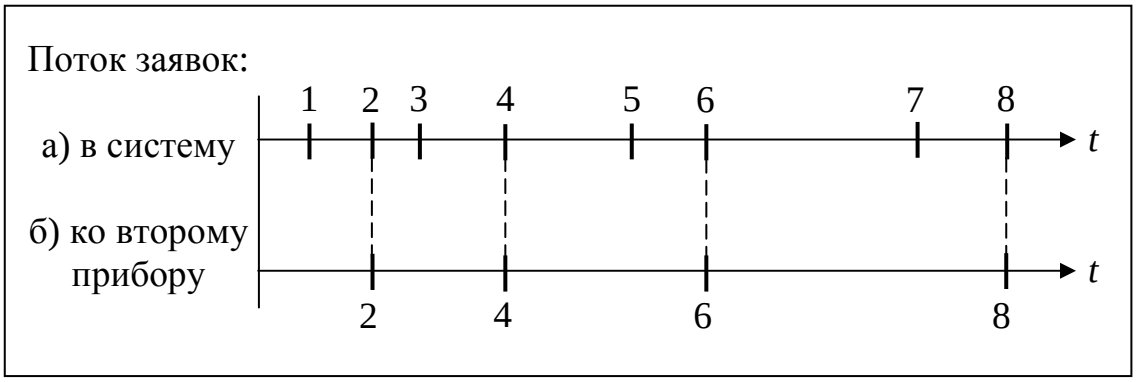
Решение.

1) Интенсивность потока заявок в СМО: $\lambda_0 = 1/a = 0{,}1\ \text{с}^{-1}$.

2) Поскольку каждая вторая заявка направляется ко второму прибору, то очевидно, что интенсивность поступления заявок ко второму прибору будет в два раза меньше, чем исходная интенсивность λ_0, то есть

$\lambda_2 = 0{,}5\lambda_0 = 0{,}05\ \text{с}^{-1}$.

2) Для определения коэффициента вариации ν_2 найдём вид закона распределения интервалов между заявками ко второму прибору, для чего построим временную диаграмму, отражающую процесс поступления заявок в систему (а) и ко второму прибору (б).



Как видно из диаграммы, интервалы между заявками ко второму прибору представляют собой сумму двух временных интервалов исходного простейшего потока заявок, поступающих в систему. Каждый такой временной интервал в случае простейшего потока представляет собой случайную величину, распределённую по экспоненциальному закону. Таким образом, интервалы между заявками ко второму прибору представляют собой случайную величину, равную сумме двух экспоненциально

распределённых величин, что соответствует распределению Эрланга 2-го порядка ($k = 2$).

Коэффициент вариации случайной величины, распределённой по закону Эрланга (см.п.2.5.5), зависит от порядка *k* и определяется по формуле:

$$\nu_2 = \nu_{Э_2} = \frac{1}{\sqrt{k}} = \frac{1}{\sqrt{2}} \approx 0{,}71.$$

Следует различать рассмотренное выше *детерминированное* разрежение потока от *вероятностного* разрежения. В случае вероятностного разрежения, когда заявки направляются ко второму прибору с вероятностью $p_2 = 0{,}5$, интенсивность поступления заявок ко второму прибору будет такой же, как и при детерминированном разрежении, то есть $\lambda_2 = p_2\lambda_0 = 0{,}5\lambda_0 = 0{,}05\ \text{с}^{-1}$. Однако коэффициент вариации в этом случае равен единице: $\nu_2 = 1$, поскольку, в соответствии с одним из сформулированных в п.3.1.3 замечательных особенностей простейшего потока, при вероятностном разрежении образуются простейшие потоки, в которых интервалы между последовательными заявками распределены по экспоненциальному закону, а не по закону Эрланга.

Задача 2. Проиллюстрировать на примере различие между дисциплинами группового и одиночного режима.

Решение. Рассмотрим следующие дисциплины обслуживания заявок:

1) одиночного режима:

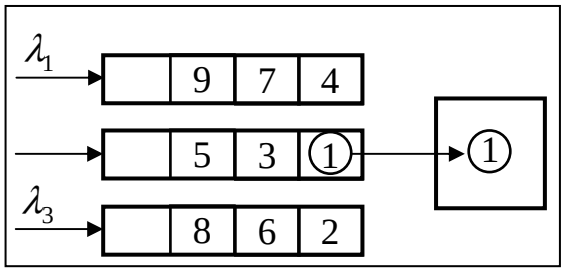
- обслуживание в порядке поступления (ОПП или FIFO);
- обслуживание в обратном порядке (ООП или LIFO);
- циклическое обслуживание в одиночном режиме (ЦО ОР), означающее, что всякий раз на обслуживание из очереди выбирается только одна заявка, после чего обслуживающий прибор переходит к следующей по порядку очереди, даже если в предыдущей очереди остались заявки;
- с относительными приоритетами (ОП), распределёнными по правилу: класс заявок с меньшим номером имеет более высокий приоритет;

2) группового режима:

- циклическое обслуживание в групповом режиме (ЦО ГР), отличающееся от одиночного режима тем, что обслуживание очереди заявок одного и того же класса осуществляется до тех пор, пока очередь не окажется пустой;
- чередующиеся приоритеты с размером группы, равным 2 (ЧП2), означающим, что из каждой очереди заявок последовательно выбирается на обслуживание не более двух заявок, после чего обслуживающий прибор переходит к непустой очереди с самым высоким приоритетом, даже если в предыдущей очереди остались заявки;

- чередующиеся приоритеты с неограниченным размером группы (ЧП), означающим, что обслуживание очереди заявок одного и того же класса осуществляется до тех пор, пока очередь не окажется пустой.

Положим, что в *некоторый фиксированный момент времени* в системе с тремя классами (очередями) заявок сложилась следующая ситуация (см. рисунок).



В системе находится 9 заявок. Номер заявки соответствует моменту поступления её в систему – чем меньше номер, тем раньше потупила заявка в систему, то есть заявка с номером 1 поступила раньше всех, а последней поступила заявка с номером 9. Все поступившие на рассматриваемый момент времени заявки распределены по классам (очередям) следующим образом: заявки самого высокоприоритетного первого класса поступили в систему в моменты 4, 7 и 9, заявки второго класса – в моменты 1, 3 и 5, заявки третьего низкоприоритетного класса – в моменты 2, 6 и 8. Положим, что в рассматриваемый момент времени на обслуживании в приборе находится заявка второго класса с номером 1. Полагая, что в систему более не поступят другие заявки, запишем последовательность обслуживания заявок при использовании перечисленных выше дисциплин обслуживания:

ОПП: 1, 2, 3, 4, 5, 6, 7, 8, 9
ООП: 1, 9, 8, 7, 6, 5, 4, 3, 2
ЦО ОР: 1, 2, 4, 7, 3, 6, 9, 5, 8
ЦО ГР: 1, 3, 5, 2, 6, 8, 4, 7, 9
ОП: 1, 4, 7, 9, 3, 5, 2, 6, 8
ЧП2: 1, 3, 4, 7, 9, 5, 2, 6, 8
ЧП: 1, 3, 5, 4, 7, 9, 2, 6, 8

Таким образом, изменение дисциплины обслуживания приводит к изменению последовательности выбора заявок на обслуживание из очередей и, следовательно, к изменению их времени ожидания. В частности, заявка с номером 9 будет иметь максимальное время ожидания при дисциплинах ОПП и ЦО ГР, а минимальное – при ООП.

Следует обратить внимание на то, что при групповом режиме заявки выбираются из очереди и обслуживаются в приборе так же по одной, как и при одиночном режиме, то есть последовательно друг за другом, а не группой. Понятие «групповой режим» лишь означает, что на обслуживание **назначается** (а не обслуживается) группа заявок (обычно одного класса), и прибор переходит к обслуживанию другой группы только после завершения обслуживания всех заявок назначенной группы.

3.7. Самоконтроль: перечень вопросов и задач

1. Выполнить классификацию СМО:
- по числу обслуживающих приборов;
- по емкости накопителя;
- по числу потоков заявок.

2. Какой поток заявок называется однородным? В каких случаях поток заявок в СМО является неоднородным?

3. В каких случаях заявки в СМО относятся к разным классам?

4. Нарисовать одноканальную СМО с неоднородным потоком заявок. Какие параметры необходимо задать для её описания? Какие характеристики функционирования СМО могут быть рассчитаны по этим параметрам?

5. Нарисовать многоканальную СМО с неоднородным потоком заявок. Какие параметры необходимо задать для её описания? Какие характеристики функционирования СМО могут быть рассчитаны по этим параметрам?

6. В чём различие между детерминированным и регулярным потоком заявок? Какой поток заявок является альтернативой детерминированного потока?

7. Как называется стационарный ординарный поток без последействия?

8. Когда поток заявок является стационарным? Привести примеры нестационарного потока заявок.

9. Какой поток заявок называется ординарным? Привести примеры неординарного потока заявок.

10. Каким является поток, в котором момент поступления очередной заявки не зависит от того, когда и сколько заявок поступило до этого момента?

11. В чём проявляется наличие последействия в потоке заявок? Привести примеры потоков заявок с последействием.

12. Понятие интенсивности потока и ее размерность. Что характеризует величина обратная интенсивности?

13. По какому закону распределены интервалы времени между заявками в простейшем потоке?

14. Какими замечательными особенностями обладает простейший поток заявок?

15. Чему равны математическое ожидание, коэффициент вариации и дисперсия интервалов времени между соседними заявками в простейшем потоке, интенсивность которого равна 2 заявки в секунду?

16. В систему поступают заявки с интервалом 80 секунд. Чему равно среднее число заявок, которые поступят в систему в течение 50-ти минут, в случае: а) детерминированного потока; б) простейшего потока; в) случайного потока?

17. В систему поступают заявки двух классов со средним интервалом между соседними заявками 0,2 с и 2 с соответственно. Определить суммарную интенсивность поступления заявок в систему. По какому закону распределены интервалы между заявками суммарного потока?

18. В систему поступают заявки трех классов со средним интервалом между соседними заявками 0,1 с; 0,2 с и 2 с соответственно. Определить суммарную интенсивность поступления заявок в систему. Чему равен коэффициент вариации интервалов между заявками суммарного потока?

19. В двухканальную СМО поступает простейший поток заявок со средним интервалом между соседними заявками 0,2 с, причем каждая третья заявка направляется ко второму прибору. Чему равна интенсивность потока заявок ко второму прибору? По какому закону распределены интервалы между заявками потока ко второму прибору?

20. В двухканальную СМО поступает простейший поток заявок с интенсивностью 15 заявок в секунду, причем с вероятностью 1/3 заявка направляется ко второму прибору. Чему равна интенсивность потока заявок к первому прибору? Чему равен коэффициент вариации интервалов между заявками потока к первому прибору?

21. Что понимается под обслуживанием заявок в СМО? Что такое интенсивность обслуживания заявок в СМО, и какова её размерность?

22. Чему равны математическое ожидание, коэффициент вариации и дисперсия длительности обслуживания заявок в СМО, распределенной по экспоненциальному закону, если известно, что интенсивность обслуживания равна 2 заявки в секунду?

23. В СМО поступают 2 класса заявок с интенсивностями 0,06 и 0,54 заявок в минуту, длительности обслуживания которых распределены по экспоненциальному закону со средними значениями 2 и 1 секунд соответственно. а) По какому закону распределена длительность обслуживания заявок суммарного (объединенного) потока? б) Чему равна средняя длительность обслуживания заявок суммарного потока?

24. Перечислить возможные дисциплины буферизации. В каких СМО не используются дисциплины буферизации?

25. Какие дисциплины обслуживания заявок относятся к бесприоритетным?

26. Краткая характеристика приоритетных дисциплин обслуживания заявок.

27. Проиллюстрировать на примере отличие дисциплин группового режима от дисциплин одиночного режима.

28. В чем отличие дисциплины с чередующимися приоритетами от дисциплины с относительными приоритетами. Проиллюстрировать на примере.

29. Что такое динамические приоритеты?

30. Что характеризуют нагрузка и загрузка? В чём отличие загрузки от нагрузки? В каких случаях нагрузка совпадает с загрузкой?

31. Перечислить факторы, обусловливающие нестационарный режим работы СМО.

32. Что такое и чем характеризуется перегрузка системы? При каких условиях возникают перегрузки системы? В каких СМО не возникают перегрузки?

33. При каком условии в одноканальной СМО отсутствуют перегрузки?

34. Раскрыть обозначение и дать краткое описание следующих СМО: а) D/M/2/3; б) M/H_2/3; в) E_3/D/2/5.

35. Привести обозначение СМО в символике Кендалла, имеющей следующее описание: двухканальная СМО с однородным простейшим потоком заявок, длительность обслуживания которых распределена по произвольному закону общего вида, с ограниченной емкостью накопителя, равной 5.

36. Перечислить характеристики одноканальной и многоканальной СМО с однородным потоком заявок и записать соотношения, устанавливающие их взаимосвязь.

37. Перечислить характеристики одноканальной СМО с неоднородным потоком заявок и записать соотношения, устанавливающие их взаимосвязь.

38. Почему в СМО, работающей в стационарном режиме, могут возникать очереди? В каких случаях они не возникают? Перечислите причины, обусловливающие возникновение очередей в СМО, работающей в стационарном режиме.

39. Какая СеМО называется линейной? Перечислить факторы, обусловливающие нелинейность СеМО.

40. Основные отличия замкнутых СеМО от разомкнутых.

41. Какая СеМО называется экспоненциальной? Перечислить факторы, обусловливающие неэкспоненциальность СеМО.

42. Какая СеМО называется неоднородной? Перечислить факторы, обусловливающие неоднородность СеМО.

43. Перечислить параметры разомкнутой и замкнутой однородной неэкспоненциальной СеМО.

44. Перечислить параметры разомкнутой и замкнутой неоднородной приоритетной СеМО.

45. Каким условиям должны удовлетворять элементы матрицы вероятно-стей передач в СеМО?

46. Узловые характеристики однородных СеМО и их взаимосвязь.

47. Сетевые характеристики разомкнутых и замкнутых однородных СеМО и их взаимосвязь.

48. Что такое "производительность замкнутой СеМО"? Какие соотношения используются для расчета производительности замкнутой СеМО?