ТЕМА 2. Основы построения имитационных моделей

1. Системы массового обслуживания.
2. Потоки событий.

1. Системы массового обслуживания.

Теория массового обслуживания (или **теория очередей**) имеет дело с процессами, для которых характерна следующая структура.

В **систему массового обслуживания** (СМО) (это могут быть линии связи, приемные пункты, подъездные пути, технологические агрегаты, ремонтные бригады и т. д.) в случайные моменты времени поступают заявки (или требования). Заявки на обслуживание образуют **входной поток**.

Если есть свободные каналы обслуживания, то требование выполняется. Если все каналы обслуживания заняты, то требование становится в очередь по определенным правилам или без обслуживания покидает систему. Выполненные требования образуют **выходной поток**.

Будем считать, что поток требований является простейшим с интенсивностью λ (среднее число требований, поступающих в единицу времени).

СМО состоит из определенного числа обслуживающих единиц – **каналов обслуживания**. Различают одноканальные СМО и многоканальные СМО.

Дисциплина очереди задает порядок прохождения заявки через очередь. Заявки из очереди могут выполняться в порядке поступления, с приоритетом, в случайном порядке и т. д. Очередь может быть конечной или бесконечной. СМО с очередями называют также СМО с ожиданием. Очереди могут ограничиваться по длине (по числу находящихся в ней заявок) или по времени ожидания обслуживания. В СМО с отказом очередь не предусмотрена, то есть заявка, пришедшая в момент, когда заняты все обслуживающие каналы, получает отказ.

Время обслуживания требований в системе является случайной величиной и обычно описывается экспоненциальным (показательным) законом распределения (то есть распределение длительности оставшейся части работ по обслуживанию не зависит от того, сколько оно уже продолжалось) с интенсивностью μ (среднее число требований, выполняемых в единицу времени). Это обусловлено рядом причин:

1) отсутствием последействия;
2) простотой и удобством аналитических выражений;
3) именно так устроены многие реальные системы.

Показательное распределение времени обслуживания имеет вид:

$$P_t = \mu e^{-\mu t} \quad (t \geq 0)$$

Тогда среднее время обслуживания одним каналом одного требования

$$t_{обсл} = \frac{1}{\mu}$$

Коэффициент загрузки СМО (среднее число каналов, которое должно быть для обслуживания в единицу времени всех поступающих требований)

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

Одноканальная СМО с отказами

СМО содержит один обслуживающий канал. На вход поступает простейший поток заявок с интенсивностью λ. Образование очереди не допускается. Если заявка застала обслуживающий канал занятым, то она покидает систему.

Время обслуживания заявки есть случайная величина, которая подчиняется экспоненциальному закону распределения с параметром μ. Среднее время обслуживания одной заявки $t_{обсл} = 1/\mu$.

Возможные состояния СМО $S0$ (канал свободен) и $S1$ (канал занят).

Размеченный граф состояний одноканальной СМО с отказами имеет следующий вид (рис.7.1):

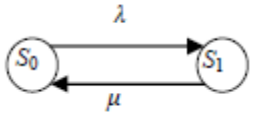

Рис. 7.1. Граф состояний одноканальной СМО с отказами

Показатели эффективности работы СМО:

1) вероятность отказа $p_{отк}$ (вероятность того, что заявка покинет СМО необслуженной, т.е. предельная вероятность состояния $S1$)

$$p_{отк} = p_1$$

2) относительная пропускная способность Q (отношение среднего числа обслуживаемых в единицу времени заявок к среднему числу поступивших за это время заявок)

$Q = 1 - p_{отк}$;

3) абсолютная пропускная способность A (среднее число заявок, которое СМО может обслужить в единицу времени)

$A = \lambda Q$.

Многоканальная СМО с отказами

СМО содержит n обслуживающих каналов. На вход поступает простейший поток заявок с интенсивностью λ. Образование очереди не допускается. Если заявка застала все обслуживающие каналы занятыми, то она покидает систему. Если в момент поступления требования имеется свободный канал, то он немедленно приступает к обслуживанию поступившего требования. Каждый канал может одновременно обслуживать только одно требование. Все каналы функционируют независимо. Время обслуживания заявки есть случайная величина, которая подчиняется экспоненциальному закону распределения с параметром μ. Среднее время обслуживания одной заявки $t_{обсл} = 1/\mu$.

Возможные состояния СМО $S0$ (все каналы свободны), $S1$ (один канал занят, остальные свободны), $S2$ (два канала заняты, остальные свободны), , Sn (все каналы заняты).

Размеченный граф состояний многоканальной СМО с отказами имеет следующий вид (рис.7.3):

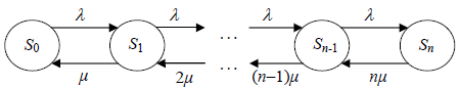

Рис. 2.3. Граф состояний системы

Приведенная интенсивность потока заявок (интенсивность нагрузки канала) $\rho = \lambda/\mu$.

Показатели эффективности работы СМО:

1) $p0$ (вероятность того, что все обслуживающие каналы свободны);

2) вероятность отказа *ротк* (вероятность того, что заявка покинет СМО необслуженной) $ротк=pn$;

3) pk (вероятность того, что в системе k требований)

$$p_k = \frac{\rho^k}{k!} p_0$$

4) относительная пропускная способность Q (отношение среднего числа обслуживаемых в единицу времени заявок к среднему числу поступивших за это время заявок) $Q=1–ротк$;

5) абсолютная пропускная способность A (среднее число заявок, которое СМО может обслужить в единицу времени) $A=\lambda Q$;

6) среднее число свободных от обслуживания каналов $N0$ есть математическое ожидание числа свободных каналов $N0=np0+(n–1)p1+…+1pn-1+0pn$;

7) коэффициент простоя каналов ;

8) среднее число занятых обслуживанием каналов;

9) коэффициент загрузки каналов .

Одноканальная СМО с неограниченной очередью

В этом случае клиенты формируют одну очередь к единственному пункту обслуживания. Пусть

λ – число заявок в единицу времени;

μ – число клиентов, обслуживаемых в единицу времени;

n – число заявок в системе.

Возможные состояния СМО $S0$ (канал свободен), $S1$ (канал занят, очереди нет), $S2$ (канал занят, в очереди одна заявка), $S3$ (канал занят, в очереди две заявки) и т.д.

Размеченный граф состояний одноканальной СМО с неограниченной очередью имеет следующий вид (рис. 7.5):

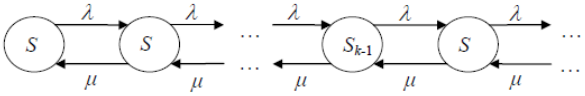

Рис. 7.5. Граф состояний одноканальной СМО с неограниченной очередью

Формулы для описания системы:

$L_{сист} = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$ – среднее число клиентов в системе;

$T_{сист} = \frac{1}{\mu-\lambda} = \frac{L_{сист}}{\lambda}$ – среднее время обслуживания одного клиента в системе (время ожидания в очереди плюс время обслуживания);

$L_{обсл} = \rho$ – среднее число заявок, находящихся под обслуживанием;

$L_{оч} = L_{сист} - L_{обсл} = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{1-\rho}$ – среднее число клиентов в очереди – средняя длина очереди;

$T_{оч} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{L_{оч}}{\lambda}$ – среднее время ожидания клиента в очереди;

$p_0 = 1 - \frac{\lambda}{\mu}$ – вероятность отсутствия заявок в системе;

$p_{зан} = 1 - p_0 = \rho$ – вероятность того, что канал занят;

$p_k = \rho^k(1-\rho)$ – вероятность того, что в системе ровно k клиентов;

$p_{n>k} = \rho^{k+1}$ – вероятность того, что в системе находится более чем k клиентов.

Многоканальная СМО с неограниченной очередью

В многоканальной системе для обслуживания открыты два канала или более. Предполагается, что клиенты ожидают в общей очереди и обращаются в первый освободившийся канал обслуживания. Пример такой многоканальной однофазовой системы можно увидеть во многих банках: из общей очереди клиенты обращаются в первое освободившееся окошко для обслуживания.

В многоканальной системе поток заявок подчиняется *пуассоновскому* закону с параметром λ, а время обслуживания – *экспоненциальному* с параметром μ. Приходящий первым обслуживается первым, и все каналы обслуживания работают в одинаковом темпе.

Возможные состояния СМО $S0$ (все каналы свободны), $S1$ (один канал занят, остальные свободны), $S2$ (два канала заняты, остальные свободны), , Sn (все каналы заняты), $Sn+1$ (все каналы заняты, в очереди одна заявка), $Sn+2$ (все каналы заняты, в очереди две заявки) и т.д.

Размеченный граф состояний многоканальной СМО с неограниченной очередью имеет следующий вид (рис. 7.6):

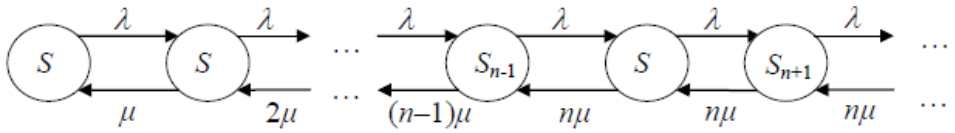

Рис. 7.6. Граф состояний системы

Формулы для описания системы:

$p_0 = \left(1+\frac{\rho}{1!}+\frac{\rho^2}{2!}+...+\frac{\rho^n}{n!}+\frac{\rho^{n+1}}{n!(n-\rho)}\right)^{-1}$ – вероятность того, что система свободна;

$p_n = \frac{\rho^n}{n!}p_0$ – вероятность того, что в системе находится n заявок;

$p_{n+1} = \frac{\rho}{n}\cdot\frac{\rho^n}{n!}p_0$; $p_{n+2} = \left(\frac{\rho}{n}\right)^2\cdot\frac{\rho^n}{n!}p_0$; $p_{n+3} = \left(\frac{\rho}{n}\right)^3\cdot\frac{\rho^n}{n!}p_0$ и т.д.

$p_q = \frac{\rho^{n+1}}{n!(n-\rho)}p_0$ – вероятность того, что заявка окажется в очереди;

$L_{оч} = \frac{\rho^{n+1}p_0}{n!n(1-\frac{\rho}{n})^2}$ – среднее число заявок в очереди;

$L_{сис} = L_{оч} + \rho$ – среднее число заявок в системе;

$T_{оч} = \frac{1}{\lambda}L_{оч}$ – среднее время нахождения заявки в очереди;

$T_{сис} = \frac{1}{\lambda}L_{сис}$ – среднее время нахождения заявки в системе.

Одноканальная СМО с ограниченной очередью

СМО содержит один обслуживающий канал. На вход поступает простейший поток заявок с интенсивностью λ. Если заявка застала обслуживающий канал занятым, то она встает в очередь и ожидает начала обслуживания. Число мест в очереди ограничено и равно m. Если заявка застала обслуживающий канал занятым и в очереди нет свободных мест, то она покидает систему необслуженной.

Время обслуживания заявки есть случайная величина, которая подчиняется экспоненциальному закону распределения с параметром μ. Среднее время обслуживания одной заявки $тобсл=1/\mu$.

Возможные состояния СМО $S0$ (канал свободен), $S1$ (канал занят, очереди нет), $S1+1$ (канал занят, в очереди одна заявка), $S1+2$ (канал занят, в очереди две заявки),…, $S1+m$ (канал занят, в очереди m заявок).

Размеченный граф состояний многоканальной СМО с неограниченной очередью имеет следующий вид (рис. 7.7):

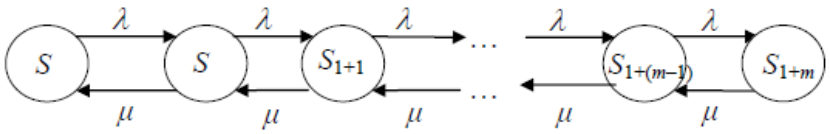

Рис. 7.7. Граф состояний системы

Формулы для описания системы:

1) $p_0 = \dfrac{1-\rho}{1-\rho^{m+2}}$ – вероятность того, что канал свободен;

2) $p_1=\rho p_0$;

3) $p_{1+k}=\rho^{1+k}p_0$ – вероятность того, что канал занят, в очереди k заявок;

4) $p_{отк}=p_{1+m}=\rho^{1+m}p_0$ – вероятность отказа (канал занят, в очереди нет свободных мест);

5) относительная пропускная способность Q (отношение среднего числа обслуживаемых в единицу времени заявок к среднему числу поступивших за это время заявок) $Q=1-p_{отк}$;

6) абсолютная пропускная способность A (среднее число заявок, которое СМО может обслужить в единицу времени) $A=\lambda Q$;

7) среднее число заявок в очереди $L_{оч} = \rho^2 \dfrac{1-\rho^m(m+1-m\rho)}{(1-\rho^{m+2})(1-\rho)}$;

8) среднее время нахождения заявки в очереди $T_{оч} = \dfrac{1}{\lambda} L_{оч}$;

9) среднее число заявок, находящихся под обслуживанием (среднее число занятых каналов)

$$L_{обсл}=1-p_0;$$

10) среднее число заявок в системе $L_{сис} = L_{оч} + L_{обсл}$;

11) среднее время нахождения заявки в системе $T_{сис} = \dfrac{1}{\lambda} L_{сис}$.

2. Потоки событий

Переходы СМО из одного состояния в другое происходят под воздействием вполне определённых событий – поступление заявок и их обслуживание.

Потоком событий называется последовательность однородных событий, следующих одно за другим в некоторые моменты времени. Например, это поток отказов или поток заявок на обслуживание вычислительного комплекса. События потока происходят в заранее неизвестные (обычно случайные) моменты времени t_1, t_2, ..., поэтому поток событий удобно изображать рядом точек на оси времени t (рис.1.5).

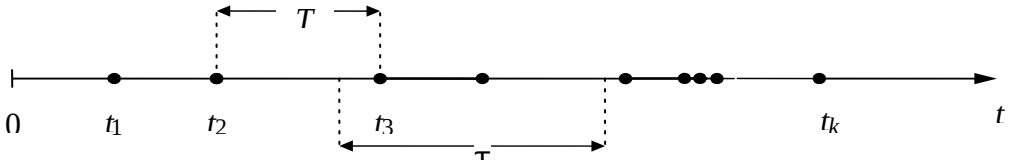

Рис. 1.5. Ось времени t

Регулярным (или детерминированным) потоком событий называется поток, в котором события следуют одно за другим через одинаковые промежутки времени.

Важной характеристикой любого потока является число m событий, происшедших за время τ. Если m случайная величина с возможными значениями $x \in \{0, 1, 2, ...\}$, то рассматривают или вероятности $p\,m(=x)$, или математические ожидания m_x и дисперсии σ_x^2, зависящие от τ.

Поток событий называется **стационарным,** если при любом x величина $p\,m(=x)$ зависит только от длины участка τ и не зависит от его расположения на оси времени.

Ординарным потоком событий называется такой поток, для которого вероятность $p\,m(\geq 2)$ попадания на отрезок τ пренебрежимо мала по сравнению с вероятностью $p\,m(=x)$. То есть практически невозможно попадание двух и более событий на достаточно малый интервал τ.

Потоком без последействия называется поток событий, в котором для любых двух непересекающихся интервалов τ_1 и τ_2 оси t число событий m_1, попадающих на τ_1, не зависит от того, сколько событий m_2 попало на τ_2. Это говорит о том, что события, образующие поток, появляются в последовательные моменты времени независимо друг от друга.

Поток событий называется **простейшим** (или стационарным пуассоновским), если он обладает одновременно тремя свойствами: стационарности, ординарности и отсутствием последействия. Интерес к этому потоку вызван простотой формального описания его статистических характеристик. Кроме того, при сложении нескольких независимых случайных потоков, стационарных и ординарных, образуется суммарный, который по своим характеристикам приближается к простейшему. Поэтому, при исследовании реальных потоков стремятся свести их к простейшим. Для этого нестационарные потоки представляются как стационарные на ограниченном отрезке времени. Неординарные потоки сводят к ординарным, рассматривая несколько одновременно наступающих событий как одно событие, и т.д.

Нестационарным пуассоновским потоком является поток событий, который не имеет последействия, ординарен, но не стационарен. В простейшем потоке (как стационарном, так и нестационарном) величина m подчиняется распределению ПуассонаВажной характеристикой потока событий является закон распределения длины промежутка между соседними событиями. Для простейшего потока с интенсивностью λ длина этого промежутка T (рис.1.5) распределена по показательному (экспоненциальному) закону с плотностью

Потоком Пальма (или потоком с ограниченным последействием) называется поток событий, для которого промежутки времени между последовательными событиями $T\,T_{1\,2}$, ,...,T_i,.... представляют собой независимые, одинаково распределенные случайные величины.

Очевидно, что простейший поток представляет собой частный случай потока Пальма, когда интервалы между событиями имеют показательное распределение. Так, пусть имеется система, состоящая из какого-то количества элементов, которые независимо друг от друга выходят из строя. Неисправный элемент тут же заменяется новым. Поток неисправностей образует поток Пальма.

Особый класс потоков Пальма образуют **потоки Эрланга,** которые получаются путем прореживания простейших потоков, т.е. отбрасыванием некоторых событий как несостоявшихся. Если в простейшем потоке сохраняется каждое $k-$е событие (считая от условного первого), а остальные просто не учитываются, то возникает поток Эрланга $k-$го порядка (k = 1, 2, ...), обозначаемый $Э_k$. На рис. 1.6 приведен поток Эрланга 3-го порядка (два события простейшего потока пропускаются, а третье учитывается).

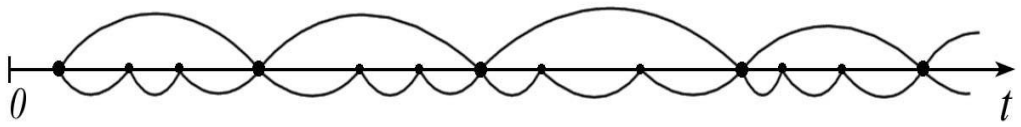

Рис. 1.6. Поток Эрланга 3-го порядка

Очевидно, что интервал времени T_k между любыми соседними событиями в потоке $Э_k$ – есть сумма k независимых случайных величин, а именно расстояний между соответствующими событиями простейшего потока.