Тема 8. МОДЕЛИРОВАНИЕ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ.

8.1. Понятие СМО

Системой массового обслуживания (СМО) называется любая система предназначенная для обслуживания каких-либо заявок (требований), поступающих на нее в случайные моменты времени.

В качестве процесса обслуживания могут быть представлены различные по своей физической природе процессы функционирования экономических, производственных, технических и других систем. Примеры систем массового обслуживания следующие: потоки поставок продукции некоторому предприятию, потоки деталей и комплектующих изделий на сборочном конвейере цеха, заявки на обработку информации ЭВМ от удаленных терминалов и т. д. При этом характерным для работы таких объектов является случайное появление заявок (требований) на обслуживание и завершение обслуживания в случайные моменты времени, т. е. стохастический характер процесса их функционирования. Остановимся на основных понятиях массового обслуживания, необходимых как при аналитическом, так и при имитационном подходе.

Работа любой системы массового обслуживания состоит в выполнении поступающего на ее вход потока ***заявок***. Заявки поступают в некоторые, в общем случае случайные, моменты времени. Обслуживание заявки продолжается какое-то время, также случайное, после чего канал освобождается для обслуживания следующей заявки. Предмет теории массового обслуживания – установление зависимостей между характером потока заявок, производительностью отдельного канала обслуживания, числом каналов и эффективностью обслуживания.

Различают СМО *с отказами и СМО с очередью.* В СМО с отказами заявка, пришедшая в момент, когда все каналы заняты, получает отказ, покидает СМО и в дальнейшем в процессе ее работы не участвует. В СМО с очередью заявка, пришедшая в момент занятости всех каналов, не покидает СМО, а становится в очередь и ждет, пока не освободится какой-нибудь канал. Число мест в очереди m может быть как ограниченным, так и неограниченным. При m = О СМО с очередью превращается в СМО с отказами. Очередь может иметь ограничения не только по количеству стоящих в ней заявок (длине очереди), но и по времени ожидания (такие СМО называются «системами с нетерпеливыми клиентами»).

СМО с очередью различаются не только по ограничениям очереди, но и по *дисциплине обслуживания:* обслуживаются ли заявки в порядке поступления, или в случайном порядке, или же некоторые заявки обслуживаются вне очереди (так называемые «СМО с приоритетом»). Приоритет может иметь несколько градаций или рангов.

Аналитическое исследование СМО является наиболее простым, если все потоки событий, переводящие ее из состояния в состояние, — простейшие (стационарные пуассоновские). Это значит, что интервалы времени между событиями в потоках имеют показательное распределение с параметром, равным интенсивности соответствующего потока. Для СМО это допущение означает, что как поток заявок, так и поток обслуживании — простейшие. Под *потоком обслуживании* понимается поток заявок, обслуживаемых одна за другой одним непрерывно занятым каналом. Этот поток оказывается простейшим, только если время обслуживания заявки $T_{обс}$ представляет собой случайную величину, имеющую показательное распределение. Параметр этого распределения μ есть величина, обратная среднему времени обслуживания. Вместо «поток обслуживании — простейший» часто говорят «время обслуживания — показательное». Условимся в дальнейшем для краткости всякую СМО, в которой все потоки простейшие, называть *простейшей* СМО.

Если всё потоки событий простейшие, то процесс, протекающий в СМО, представляет собой Марковский случайный процесс с дискретными состояниями и непрерывным временем. При выполнении некоторых условий для этого процесса существует финальный стационарный режим, при котором как вероятности состояний, так и другие характеристики процесса не зависят от времени.

Задачи теории массового обслуживания — нахождение вероятностей различных состояний СМО, а также установление зависимости между заданными параметрами (числом каналов n, интенсивностью потока заявок λ, распределением времени обслуживания и т, д.) и *характеристиками эффективности* работы СМО. В качестве таких характеристик могут рассматриваться, например, следующие:

- среднее число заявок *A,* обслуживаемое СМО в единицу времени, или *абсолютная пропускная способность* СМО;
- вероятность обслуживания- поступившей заявки Q или *относительная пропускная способность* СМО; $Q = A/\lambda$;
- вероятность отказа $P_{отк}$ т.е вероятность того, что поступившая заявка не будет обслужена, получит отказ; $P_{отк} = 1 - Q$;

Рассмотри процессы, протекающие в системе массового обслуживания.

8.2.Мнемоническое обозначение СМО.

В теории массового обслуживания приняты очень удобные сокращенные обозначения для различных СМО, позволяющие легко охарактеризовать систему. В основе этих обозначений лежит трехбуквенная комбинация вида A/B/N, где:

А — описывает распределение (или задает характер закона распределения) интервалов поступления заявок;

В — описывает распределение длительностей обслуживания заявок;

N — задает количество обслуживающих приборов в СМО.

Для СМО с очередью, приведенное обозначение расширяется до четырех букв A/B/N/К, где последняя буква (на самом деле число, как и N) К задает емкость накопителя (количество мест ожидания).

Приведенные трех или четырех буквенные обозначения называют обозначениями Кендалла. В этих обозначениях А и В могут принимать значения из следующего набора символов {М, D, Ek, Hk, G, U}. При этом:

а) А или В=М, если распределение интервалов поступления или длительностей обслуживания заявок является экспоненциальным (М — от слова Markovian — Марковский);

б) А или В=D, если интервалы поступления или длительности обслуживания являются детерминированными (D — Determinate);

в) А или В=Ek, если соответствующие распределения являются Эрланговскими порядка k (Е — Erlang);

г) А или В=Hk, в случае гиперэкспоненциальных распределений порядка k (Н — Hyperexponential);

д) А или В= G, в случае распределений общего (произвольного) вида (G — General — общий, общего вида);

е) A или B= U — при равномерных распределениях соответствующих случайных величин (U — Uniform distribution — равномерное распределение).

Так, например, обозначение вида:

M/M/1 означает СМО с простейшим потоком на входе и экспоненциально распределенной длительностью обслуживания заявок в приборе (один).

D/E2/3/5 — СМО с регулярным потоком на входе, длительностью обслуживания, распределенной по закону Эрланга 2-го порядка, тремя обслуживающими приборами и пятью местами ожидания;

M/G/2 — СМО с простейшим потоком на входе, длительностью обслуживания, распределенная по закону произвольного вида, и двумя обслуживающими приборами.

В случае СМО с неоднородной нагрузкой используются обозначения вида, где символ вектора над буквами A и B указывает на неоднородность нагрузки, а индекс H задает количество классов заявок. Например,

$M4/M/1$ — это обозначение СМО с одним обслуживающим прибором, четырьмя классами заявок, которые образуют на входе системы простейшие потоки и имеют общие законы распределения длительностей обслуживания.

8.3. СМО с отказами

Системы массового обслуживания делятся на системы с отказами и системы с ожиданием.

В системах с отказами заявка, поступившая в момент, когда все каналы обслуживания заняты, немедленно получает отказ, покидает систему и в дальнейшем в процессе обслуживания не участвует.

Пусть имеется n-канальная СМО с отказами. Рассмотрим конечное множество состояний этой системы:

z_0 – свободны все каналы;

z_1 – занят один канал;

. . . .

z_k – заняты k каналов;

. . . .

z_n – заняты все n каналов.

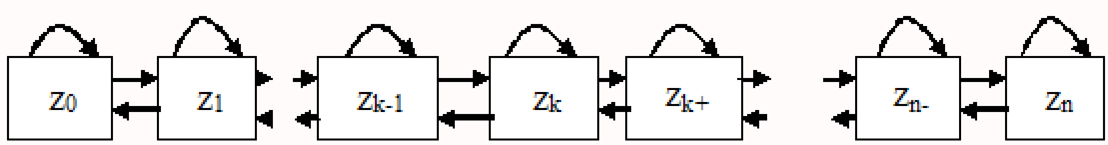

Рис. 3.1

Определим вероятности состояния системы pk(t) для любого момента времени в предположении, что поток заявок простейший, с интенсивностью λ, время обслуживания показательное, с параметром μ.

Поскольку оба потока заявок в системе (заявок и обслуживания) являются простейшими, то процесс, протекающий в системе будет марковским.

Очевидно, что для любого момента времени

$$\sum_{k=0}^{n} p_k(t) = 1$$

.

Составим дифференциальные уравнения для всех вероятностей состояний системы. Для этого, зафиксируем момент времени t и найдем вероятность pk(t+Δt) того, что в момент (t+Δt) система будет находиться в состоянии zk.

Для состояния z0 это может произойти двумя способами:

событие A – в момент времени t система находилась в состоянии z0 и осталась в этом состоянии. Вероятность этого события равна вероятности того, что за время Δt на вход системы не пришла ни одной заявки:

$$e^{-\lambda\cdot\Delta t} \approx 1 - \lambda\cdot\Delta t.$$

Следовательно, P(A)=p0(t)(1-λ˙Δt).

событие B – вероятность того, что система была в состоянии z1 и перешла в состояние z0. Вероятность этого события равна:

$$1 - e^{-\mu\cdot\Delta t} \approx \mu\cdot\Delta t.$$

Следовательно, P(B)=p_1(t)μ˙Δt.

Таким образом:

p_0(t+Δt)=p_0(t)(1-λ˙Δt)+ p_1(t)μ˙Δt.

$$\frac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t).$$

Аналогично составляются дифференциальные уравнения для других состояний системы. Для состояния zk вероятность p_k(t+Δt) определиться как сумма вероятностей трех событий:

событие A – в момент времени t система находилась в состоянии zk и осталась в этом состоянии. Вероятность этого события равна вероятности того, что за время Δt на вход системы не пришла ни одной заявки и ни одна из k заявок из системы не ушла (не обслужилась):

$$e^{-\lambda\cdot\Delta t}\left(e^{-\mu\Delta t}\right)^k = e^{-(\lambda+k\mu)\Delta t} \approx 1-(\lambda+k\mu)\cdot\Delta t.$$

Следовательно, P(A)=p_k(t)[1-(λ+kμ)˙Δt].

событие B – вероятность того, что система была в состоянии zk-1 и перешла в состояние zk. (пришла одна заявка). Вероятность этого события равна:

P(B)=p_{k-1}(t)λ˙Δt.

событие C – вероятность того, что система была в состоянии zk+1 и перешла в состояние zk. (обслужена одна заявка). Вероятность этого события равна:

$P(C)=p_{k+1}(t)(k+1)\mu\cdot\Delta t.$

Таким образом:

$p_k(t+\Delta t)= p_k(t)[1-(\lambda+k\mu)\cdot\Delta t]+ p_{k-1}(t)\lambda\cdot\Delta t+ p_{k+1}(t)(k+1)\mu\cdot\Delta t.$

$$\frac{dp_k(t)}{dt} = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + (k+1)\mu p_{k+1}(t).$$

Составим уравнение для последней вероятности pn:

$p_n(t+\Delta t)\approx p_n(t)(1-n\mu\cdot\Delta t)+ p_{n-1}(t)\lambda\cdot\Delta t.$

$$\frac{dp_n(t)}{dt} = \lambda p_n(t) - n\mu p_{n-1}(t).$$

Таким образом, получена система дифференциальных уравнений для вероятностей состояний системы:

$$\begin{cases} \dfrac{dp_0(t)}{dt} = -\lambda p_0(t) + \mu p_1(t), \\ \dots\dots\dots \\ \dfrac{dp_k(t)}{dt} = \lambda p_{k-1}(t) - (\lambda + k\mu) p_k(t) + (k+1)\mu p_{k+1}(t). \\ \dots\dots \qquad\qquad (0<k<n), \\ \dfrac{dp_n(t)}{dt} = \lambda p_n(t) - n\mu p_{n-1}(t). \end{cases}$$

Эти уравнения называются уравнениями Эрланга.

Вероятности $p_k(t)$ характеризуют среднюю загрузку системы и ее изменение с течением времени.

Вероятность $p_n(t)=P_{отк}$ есть вероятность того, что заявка, пришедшая в систему в момент времени t получит отказ.

Величина $q(t)=1-p_n(t)$ называется пропускной способностью системы.

Введем обозначение $\alpha=\lambda/\mu$ и назовем величину α *приведенной плотностью потока заявок*. Эта величина есть также среднее число заявок, приходящееся на среднее время обслуживания одной заявки: $\alpha=\lambda m_{обсл}$.

В новых обозначениях вероятности p_k принимает вид:

$$p_k = \frac{\alpha^k}{k!} p_0.$$

Приведенные выше формулы выражают вероятности p_k через p_0. Для того, чтобы выразить эти вероятности через характеристики системы α и n, воспользуемся условием нормировки:

$$\sum_{k=0}^{n} p_k = p_0 \sum_{k=0}^{n} \frac{\alpha^k}{k!} = 1,$$

откуда

$$p_0 = \frac{1}{\sum_{k=0}^{n} \frac{\alpha^k}{k!}}.$$

Окончательное выражение для вероятностей состояния системы принимают

$$p_k = \frac{\frac{\alpha^k}{k!}}{\sum_{k=0}^{n} \frac{\alpha^k}{k!}} \qquad (0 \le k \le n).$$

вид:

Вероятность отказа (все каналы заняты): $P_{отк} = p_n = \dfrac{\frac{\alpha^n}{n!}}{\sum_{k=0}^{n} \frac{\alpha^k}{k!}}.$

Для одноканальной системы (n=1): $P_{отк} = p_1 = \dfrac{\alpha}{1+\alpha}$

Относительная пропускная способность : $q = 1 - P_{отк} = \dfrac{1}{1+\alpha}.$

Формулы Эрланга и их следствия были получены в предположении о показательном распределении времени обслуживания заявок. Однако исследования показали, что эти формулы справедливы при любом законе распределения времени обслуживания, лишь бы входной поток был простейшим.

8.4. СМО с ожиданием

Система массового обслуживания называется системой с ожиданием, если заявка, заставшая все каналы занятыми, становится в очередь и ждет, пока не освободится какой-нибудь канал.

Если время ожидания заявки в очереди ничем не ограничено, то система называется *чистой системой с ожиданием*. Если оно ограничено некоторыми условиями, то система называется системой смешанного типа. Ограничения, наложенные на ожидание могут быть различного типа, например:

- ограничение на время пребывания заявки в очереди;
- ограничение на длину очереди;
- ограничение на время пребывания заявки в системе.

В системах с ожиданием существенную роль играет так называемая *дисциплина очереди*. Каждый тип системы с ожиданием имеет свои особенности и математическую теорию. Мы остановимся на простейшем случае смешанной системы, являющимся обобщением задачи Эрланга для системы с отказами.

Рассмотрим СМО с n каналами, на вход которой поступает простейший поток с параметром λ. Время обслуживания заявок также имеет показательное распределение с параметром μ. Заявка, заставшая все каналы занятыми, становится в очередь и ожидает обслуживания. Время ожидания заявки в очереди ограничено некоторым сроком $T_{ож}$. Если до истечения этого срока заявка не будет обслужена, то она покидает систему. Срок ожидания обслуживания будем полагать случайной величиной с показательным распределением и параметром ν. Очевидно, что при $\nu \to \infty$, система смешанного типа превращается в чистую систему с отказами, а при $\nu \to 0$, система смешанного типа превращается в чистую систему с ожиданиями.

Отметим, что в предположении о показательном распределении срока ожидания пропускная способность системы не зависит от того, обслуживаются ли заявки в порядке очереди ли в случайно порядке: для каждой заявки закон распределения оставшегося времени ожидания не зависит от того, сколько времени заявка стояла в очереди.

для любого k≤n $p_k = \dfrac{\lambda^k}{k!\mu^k} p_0$;

для любого s≥1: $$p_{n+s} = \frac{\lambda^{n+s} p_0}{n!\mu^n \prod_{m=1}^{s}(n\mu + m\nu)}.$$

В приведенных выше формулах в качестве сомножителя присутствует вероятность p_0. Определим эту вероятность из дополнительного условия:

$$p_0\left\{\sum_{k=0}^{n}\frac{\lambda^k}{k!\,\mu^k}+\sum_{s=1}^{\infty}\frac{\lambda^{n+s}}{n!\,\mu^n\prod_{m=1}^{s}(n\mu+m\nu)}\right\}=1$$

Введем обозначения:

$\lambda/\mu=\lambda m_{t обсл}=\alpha$; $\nu/\mu=\nu m_{t обсл}=\beta$.

Параметра α и β выражают соответственно среднее число заявок и среднее число необслуженных заявок приходящееся на среднее время обслуживания одной заявки.

В новых обозначениях приведенные выше выражения принимают вид:

$$p_k=\frac{\alpha^k}{k!}p_0; \qquad (0<k\leq n)$$

$$p_{n+s}=\frac{\frac{\alpha^{n+s}}{n!}p_0}{\prod_{m=1}^{s}(n+m\beta)}; \qquad (s\geq 1).$$

Зная вероятности состояния системы можно определить и другие интересующие нас характеристики, в частности вероятность того, что заявка покинет систему не обслуженной. Определим эту вероятность из следующих соображений: при установившемся режиме вероятность $P_н$ есть отношение среднего числа заявок, уходящих из очереди в единицу времени. Определим среднее число заявок, находящихся в очереди:

$$m_s=\sum_{s=1}^{\infty}sp_{n+s}.$$

Чтобы найти вероятность $P_н$, нужно среднее число заявок в очереди умножить на среднюю плотность уходов (определим среднее число заявок, покидающих систему) и умножим на интенсивность входного потока заявок:

$$P_н=m_s\cdot\frac{\nu}{\lambda}=\frac{\beta}{\alpha}m_s.$$

Относительная пропускная способность системы: $q=1-P_н$.

Очевидно, что пропускная система с ожиданиями выше, чем пропускная способность системы с отказами и пропускная способность увеличивается с увеличением среднего времени ожидания mtож=1/ν.

Рассмотрим, во что превратиться система с ожиданиями при изменении параметра β. Очевидно, что при β→∞ система с ожиданиями превращается в чистую систему с отказами, а при β→0 – в чистую систему с ожиданиями. В такой системе вероятность того, что заявка уйдет из системы не обслуженной, равна нулю. Однако, в такой системе не всегда имеется предельный стационарный режим при t→∞. Такой режим существует только при α<n, т.е., когда среднее число заявок, приходящееся на время обслуживания одной заявки не выходит за пределы возможностей n-канальной системы. В противном случае, число заявок в очереди будет неограниченно возрастать.

Полагая, что α<n, найдем предельные вероятности состояния системы (β→0):

$$p_0 = \frac{1}{\sum\limits_{k=0}^{n} \frac{\alpha^k}{k!} + \frac{\alpha^n}{n!} \sum\limits_{s=1}^{\infty} \frac{\alpha^s}{n^s}} = \frac{1}{\sum\limits_{k=0}^{n} \frac{\alpha^k}{k!} + \frac{\alpha^{n+1}}{n!(n-\alpha)}}.$$

Отсюда найдем (0?k?n): $p_k = \dfrac{\frac{\alpha^k}{k!}}{\sum\limits_{k=0}^{n} \frac{\alpha^k}{k!} + \frac{\alpha^{n+1}}{n!(n-\alpha)}}$

$$p_{n+s} = \frac{\frac{\alpha^{n+s}}{n!\,n^s}}{\sum\limits_{k=0}^{n} \frac{\alpha^k}{k!} + \frac{\alpha^{n+1}}{n!(n-\alpha)}} \qquad (s \geq 0).$$

Среднее число заявок в очереди:

$$m_s = \frac{\frac{\alpha^{n+1}}{n \cdot n!\left(1-\frac{\alpha}{n}\right)^2}}{\sum\limits_{k=0}^{n} \frac{\alpha^k}{k!} + \frac{\alpha^{n+1}}{n!(n-\alpha)}}.$$

8.5 Простейшая многофазовая СМО с очередью.

Анализ многофазовых СМО в общем случае затруднен, тем что входящий поток каждой последующей фазы является выходным потоком предыдущей и в общем случае имеет последействие. Однако *если на вход С МО с неограниченной очередью поступает простейший поток заявок, а время обслуживания показательное, то выходной поток, этой СМО — простейший,* с той же интенсивностью λ, что и входящий. Из этого следует, что многофазовую СМО с неограниченной очередью перед каждой фазой, простейшим входящим потоком заявок и показательным временем обслуживания на каждой фазе можно анализировать как простую последовательность простейших СМО.

Если очередь к фазе ограничена, то выходной поток этой фазы перестает быть простейшим и вышеуказанный прием может применяться только в качестве приближенного.