Системы массового обслуживания

Основные понятия

Теория массового обслуживания занимается разработкой и анализом математических моделей, описывающих системы массового обслуживания.

Системой массового обслуживания называется совокупность потока заявок (требований), поступающих в систему, и приборов (каналов), обслуживающих эти заявки. Поток заявок, как правило, носит случайный характер.

Примерами систем массового обслуживания являются: автоматические телефонные станции и поступающие на них вызовы, магазины и покупатели, предприятия бытового обслуживания и клиенты, ремонтные мастерские и техника, требующая ремонта, ЭВМ и задачи, поступающие на решение, аэропорты и самолеты, требующие посадки, преподаватели и сдающие экзамены студенты, и т.д.

Для перечисленных фрагментов коммерческой деятельности характерны массовость поступления товаров, денег, посетителей в случайные моменты времени, затем их последовательное обслуживание (удовлетворение требований, запросов, заявок) путем выполнения соответствующих операций, время выполнения которых носит также случайный характер. Все это создает неравномерность в работе, порождает недогрузки, простой и перегрузки в коммерческих операциях. Много неприятностей доставляют очереди, например, посетителей в кафе, столовых, ресторанах или водителей автомобилей на товарных базах, ожидающих разгрузки, погрузки или оформления документов. В связи с этим возникают задачи анализа существующих вариантов выполнения всей совокупности операций, например, торгового зала супермаркета, ресторана или в цехах производства собственной продукции для целей оценки их работы, выявления слабых звеньев и резервов для разработки в конечном итоге рекомендаций, направленных на увеличение эффективности коммерческой деятельности.

Неотъемлемой частью систем массового обслуживания является образование очереди на обслуживание, и поэтому теорию массового обслуживания принято называть также математической теорией очередей. Важно понимать, что в теории массового обслуживания речь идет о разработке математических моделей, обладающих достаточной степенью абстракции. Поэтому не важна природа обслуживаемых заявок и их физические свойства. Существенными являются лишь моменты появления этих заявок, так как от них зависит эволюция модели во времени. В абстрактной модели нет необходимости рассматривать физическую сторону процесса обслуживания. Обслужить заявку – это значит затратить на нее некоторое время в соответствии с принятой дисциплиной обслуживания.

Любая СМО предназначена для обслуживания поступающих в нее заявок. Заявки поступают на вход СМО, вообще говоря, в случайные моменты времени.

Заявка – любой запрос на удовлетворение какой-либо потребности (например, покупатели в магазине, заявки на телефонный разговор, заявка на получение товара, больные в поликлинике и др.).

Заявки в силу массовости поступления на обслуживание образуют ***потоки***, которые до выполнения операции обслуживания называются *входящими*, а после возможного ожидания начала обслуживания, т.е. простоя в очереди, образуют *потоки обслуживания* в каналах, а затем формируется *выходящий поток* заявок.

Под ***обслуживанием*** заявок понимается удовлетворение определенной потребности.

Обслуживание заявки осуществляется каналом обслуживания (в случайные моменты времени). В некоторых случаях обслуживание осуществляется одним человеком (один продавец, один врач), в некоторых – группой людей (врачебная комиссия в поликлинике, ГАК при защите диплома), в некоторых случаях техническим устройством (автоматы).

Обслуживающий канал – совокупность средств, которые непосредственно осуществляют обслуживание заявок.

В одном случае, например, повар в процессе приготовления блюд является каналом обслуживания, а в другом выступает в роли заявки на обслуживание, например, к заведующему производством за получением товара.

Если обслуживающие каналы способны удовлетворять одинаковые заявки, они называются ***однородными.***

Совокупность однородных каналов образует ***обслуживающую систему***.

Процедура обслуживания считается ***завершенной***, когда заявка на обслуживание покидает систему.

Случайный характер поступает заявок и их обслуживание приводят к тому, что в СМО может скапливаться большое количество заявок, т.е. образуют очередь, а в некоторые моменты времени система будет простаивать. В каждый момент времени 1 канал может обслуживать только 1 заявку.

Продолжительность интервала времени, требуемого для реализации процедуры обслуживания, зависит в основном от характера заявки на обслуживание, состояния самой обслуживаемой системы и канала обслуживания. Заявки, поступившие в систему обслуживания, могут покинуть ее и будучи не обслуженными. Например, покупатель не нашел в магазине товар или если товар имеется, но большая очередь, а покупатель не располагает временем.

Таким образом***, СМО включает в себя следующие элементы***:

- источник требований
- входящий поток заявок
- очередь
- обслуживающие устройства (каналы обслуживания)
- выходящий поток требований

Переход СМО из одного состояния в другое происходит под воздействием вполне определенных событий – поступление заявок и их обслуживание

Опишем один из возможных вариантов функционирования системы (рис.1). Предположим, что на обслуживание поступает поток заявок, который характеризуется параметром λ - интенсивность потока. В системе имеется r обслуживающих каналов (приборов). Если в системе есть свободные каналы, то вновь пришедшая заявка поступает на свободный канал и начинается ее обслуживание. Время обслуживания случайное и характеризуется параметром μ. По окончании обслуживания образуется поток обслуженных заявок. Если все каналы заняты обслуживанием, то вновь пришедшая заявка становится в очередь (поступает в бункер или накопитель) емкостью m. Это значит, что в очереди может находиться не более, чем m заявок. Если количество заявок в очереди превысит m, то такие заявки покидают систему не обслуженными, образуя поток необслуженных заявок.

Рис.1. Схематическое изображение системы массового обслуживания

Работа системы массового обслуживания сопровождается рядом случайных факторов. Поток поступающих заявок представляет собой случайный процесс – число заявок

является случайной функцией времени. Время, которое требуется для обслуживания одной заявки (время обслуживания) является случайной величиной.

Основными параметрами систем массового обслуживания являются следующие:

- входящий поток заявок и его интенсивность,
- число обслуживающих каналов (приборов),
- производительность каналов,
- время обслуживания заявки,
- время ожидания начала обслуживания,
- длина очереди на обслуживание,
- дисциплина и приоритет обслуживания заявки (порядок выбора заявок из очереди, возможность получения отказа в обслуживании, возможность простоя каналов),
- потоки очереди
- среднее время ожидания в очереди,

т.е. не то, насколько хорошо выполнено обслуживание, а то, насколько полно загружена система обслуживания, не простаивают ли каналы обслуживания, не образуются ли очереди.

Классификация СМО

Системы массового обслуживания классифицируются в зависимости от вида потока заявок и характера их обслуживания.

Различают системы с потерями (с отказами) и с очередью (с ожиданием). Если заявка поступает в ***систему с потерями*** в то время, когда все каналы заняты ($m = 0$), то она получает «отказ» и теряется. Примером такой системы может быть телефонная станция. В ***системах с очередью*** заявка, пришедшая в момент, когда каналы заняты, встает в очередь и ожидает, пока не освободится один из каналов. Существуют **системы с неограниченной очередью**, когда число мест в очереди не ограниченно ($m = \infty$) и **системы с ограниченной очередью**. Ограничения могут быть разными - по числу заявок, одновременно стоящих в очереди, по времени пребывания заявки в очереди, по времени работы системы и т.д.

По числу обслуживающих каналов различают ***одноканальные*** ($r = 1$) и ***многоканальные*** ($r > 1$) системы массового обслуживания. Для многоканальной системы будем предполагать, что каждая заявка может быть обслужена любым из каналов. Такая система каналов называется ***полнодоступным пучком.***

В системах с очередью учитывается также ***дисциплина обслуживания***. Обычно заявки обслуживаются в порядке их поступления в систему по принципу «первый пришел - первый обслужен» (прямой приоритет). Однако возможны и другие правила обслуживания заявок: «последний пришел - первый обслужен» (обратный приоритет), или «первой обслуживается заявка с заданным номером» (назначенный приоритет), или «первой обслуживается заявка со случайным номером» (случайный приоритет). Возможно также обслуживание заявки вне очереди. При этом заявка с более высоким приоритетом, поступив в систему, может оборвать уже начавшееся обслуживание заявки с меньшим приоритетом, а может дождаться окончания ее обслуживания. В первом случае говорят об *абсолютном*, а во втором - об *относительном приоритете*.

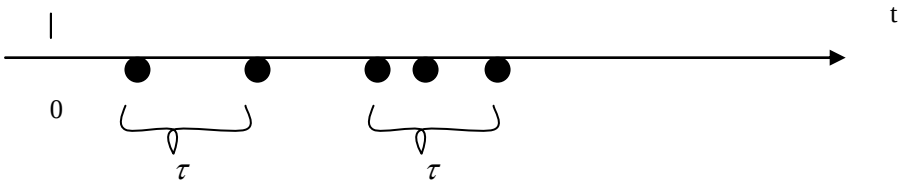
Основоположником теории массового обслуживания принято считать датского математика А.К.Эрланга, который в 1909 г. опубликовал важные результаты, полученные им при изучении математических моделей телефонных систем. В настоящее время модели и методы массового обслуживания находят приложения во многих областях науки и техники, начиная с контроля над приземлением самолетов и кончая теорией управления запасами, от исследований, связанных с ростом бактерий, – до составления больничных графиков.

Простейший поток

Переходы СМО из одного состояния в другое происходят под воздействием вполне определенных событий — поступления заявок и их обслуживания. Последовательность появления событий, следующих одно за другим в случайные моменты времени, формирует так называемый поток событий. Примерами таких потоков в коммерческой деятельности являются потоки различной природы — товаров, денег, документов, транспорта, клиентов, покупателей, телефонных звонков, переговоров. Поведение системы обычно определяется не одним, а сразу несколькими потоками событий. Например, обслуживание покупателей в магазине определяется потоком покупателей и потоком обслуживания; в этих потоках случайными являются моменты появления покупателей, время ожидания в очереди и время, затрачиваемое на обслуживание каждого покупателя. При этом основной ***характерной чертой потоков является вероятностное распределение времени между соседними событиями.*** Существуют различные потоки, которые отличаются своими характеристиками.

Свойства потоков:

1) ***Поток событий называется регулярным***, если в нем события следуют одно за другим через заранее заданные и строго определенные промежутки времени. Такой поток является идеальным и очень редко встречается на практике. Например, изделие на конвейере поступает к сборщику с постоянной скоростью. Чаще встречаются нерегулярные потоки, не обладающие свойством регулярности.
2) ***Поток событий называется стационарным***, если вероятность попадания любого числа событий на промежуток времени зависит только от длины этого промежутка и не зависит от того, как далеко расположен этот промежуток от начала отсчета времени. Стационарность потока ***означает*** независимость от времени его вероятностных характеристик, в частности, ***интенсивность стационарного потока есть среднее число событий в единицу времени и остается величиной постоянной***.



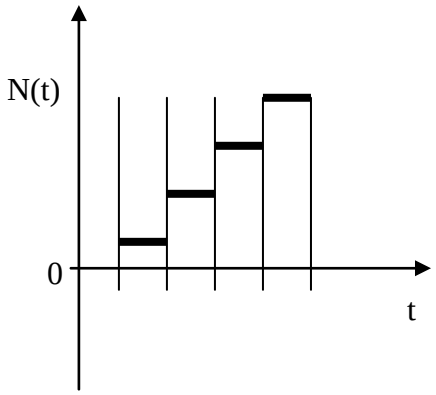
На практике обычно потоки могут считаться стационарными только на некотором ограниченном промежутке времени. Обычно поток покупателей, например, в магазине существенно меняется в течение рабочего дня. Однако можно выделить определенные временные интервалы, внутри которых этот поток допустимо рассматривать как стационарный, имеющий постоянную интенсивность.

3) ***Поток событий называется потоком без последствия,*** если число событий, попадающих на один из произвольно выбранных промежутков времени, не зависит от числа событий, попавших на другой, также произвольно выбранный промежуток, при условии, что эти промежутки не пересекаются между собой. В потоке без последствия события появляются в последовательные моменты времени независимо друг от друга. Например, поток покупателей, входящих в магазин, можно считать потоком без последствия потому, что причины, обусловившие приход каждого из них, не связаны с аналогичными причинами для других покупателей.
4) ***Поток событий называется ординарным,*** если вероятность попадания на очень малый отрезок времени сразу двух или более событий пренебрежимо мала по сравнению с вероятностью попадания только одного события. В ординарном потоке события происходят поодиночке, а не по два или более сразу.
5) Если поток одновременно обладает свойствами стационарности, ординарности и отсутствием последствия, то такой ***поток называется простейшим (или пуассоновским) потоком событий.*** Математическое описание воздействия такого

потока на системы оказывается наиболее простым. Поэтому, в частности, простейший поток играет среди других существующих потоков особую роль.

Если просуммировать (взаимно наложить) несколько стационарных и ординарных потоков практически с любым последействием (кроме регулярного), то в итоге получится поток, который можно рассматривать как простейший. При этом предполагается, что суммируемые потоки сравнимы по интенсивности. На практике достаточно сложить 4-6 потоков, чтобы поток был простейшим.

Рассмотрим функцию N(t), где N(t) – число событий на интервале (0, t). Таким образом, N(t) – неотрицательная целочисленная неубывающая функция, значения которой возрастают скачками в моменты осуществления событий.



Потоком однородных случайных событий называется случайный процесс N(t) с целочисленными неотрицательными значениями и непрерывным временем.

Задать случайный поток можно 2 способами:

1) определить $P_k(t) = P\{N(t) = k\}$

2) задать закон распределения интервала времени между соседними событиями, т.е. задать время пребывания процессов в некотором состоянии.

Используют следующие характеристики случайного потока:

1) $\lambda(t)$ - интенсивность потока – это среднее число событий, произошедших в единицу времени в момент t.

2) ведущая функция потока – математическое ожидание числа событий, произошедших на интервале (0, t).

Утверждение 1: сумма *n* простейших потоков с интенсивностями $\lambda_1, \lambda_2, \ldots, \lambda$ есть простейший поток с интенсивностью $\lambda = \sum_{i=1}^{n} \lambda_i$

Утверждение 2: Для простейшего потока число событий *k*, происшедших на интервале (0, t), распределено по закону Пуассона

$$P_k(t) = P\{N(t) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Утверждение 3: Длительность интервала времени T между соседними событиями распределена по показательному закону:

$$P(T < t) = F(t) = 1 - e^{-\lambda t}$$

- Вероятность попадания хотя бы одного события на малый промежуток времени Δt составляет

$P(T < \Delta t) = \lambda \cdot \Delta$t

• Плотность распределения промежутка времени между двумя последовательными событиями равна

$f(t) = \lambda \cdot e^{-\lambda t} \quad (t \geq 0)$

Числовые характеристики случайной величины *T* следующие:

- математическое ожидание $M[T] = \frac{1}{\lambda}$

- дисперсия $D(T) = \frac{1}{\lambda^2}$

- среднее квадратичное отклонение $\sigma = \frac{1}{\lambda}$

Таким образом, средний интервал времени *T* между любыми двумя соседними событиями в простейшем потоке в среднем равен $\frac{1}{\lambda}$, где λ - интенсивность потока, т.е. среднее число событий, происшедших в единицу времени.

• Среднее число событий, попадающих на интервал времени τ, равно $n = \lambda \cdot \tau$.

Рассмотрим другие ***характеристики СМО***:

• Случайное время ожидания в очереди начала обслуживания $t_{оч}$ распределено экспоненциально $f(t_{оч}) = v \cdot e^{-v \cdot t_{оч}}$,

где v – интенсивность потока прохода очереди, определяемая средним числом заявок, проходящих на обслуживание в единицу времени.

• $v = \frac{1}{T_{оч}}$, где $T_{оч}$ – среднее время ожидания обслуживания в очереди.

Выходной поток заявок связан с потоком обслуживания в канале, где $T_{обс}$ – длительность обслуживания также является случайной величиной и подчиняется во многих случаях показательному закону с плотностью распределения вероятности

$f(t_{обс}) = \mu \cdot e^{-\mu \cdot t_{обс}}$

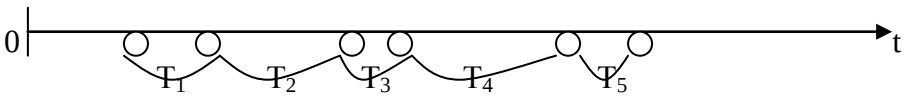
где μ - интенсивность потока обслуживания, т.е. среднее число заявок обслуживаемых в единицу времени $\mu = \frac{1}{t_{обс}}$.

• Важной характеристикой СМО является ***интенсивность нагрузки*** $\boldsymbol{\rho = \frac{\lambda}{\mu}}$, которая показывает степень согласования входных и выходных потоков заявок канала обслуживания и определяет устойчивость СМО. Для многоканальной СМО коэффициент загрузки равен $\rho = \frac{\lambda}{n \cdot \mu}$.

Условием стационарного режима работы СМО является условие, когда средняя длина очереди, среднее время ожидания заявки в очереди, среднее время пребывания заявки в системе являются конечными. Это происходит, когда для коэффициента загрузки выполняется ***условие*** $0 \leq \rho < 1$ ***(для многоканальной СМО*** $0 \leq \rho < n$ ***).*** Когда $\rho \geq 1$, в системе образуются бесконечные очереди

Кроме понятий простейшего потока событий, часто пользуются понятиями потоков других типов.

Поток событий называется ***потоком Пальма*** (или потоком с ограниченным последействием), если промежутки времени между соседними событиями являются независимыми, одинаково распределенными случайными, но в отличие от простейшего потока не обязательно распределенными по показательному закону.

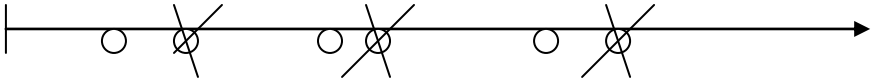


T_i – независимые случайные величины, одинаково распределенные.

Простейший поток является частным случаем потока Пальма.

Потоком Эрланга называется поток, полученный путем «прореживания» (просеивания) простейшего потока.

Поток Эрланга *k*-порядка ($E_к$) получается из простейшего, если оставить в нем только каждое k-ое событие.



E_2 – поток Эрланга 2-го порядка.

Очевидно, что E_1 – простейший поток.

Исследование СМО начинается с изучения того, что необходимо обслуживать, следовательно, с изучения входящего потока заявок и его характеристик.

Итак, наиболее общие характеристики СМО носят вероятностный характер. Это:

- интервал времени поступления заявок - τ
- продолжительность операций обслуживания - $t_{обс}$
- время ожидания в очереди - $t_{оч}$
- длина очереди – $l_{оч}$
- интенсивность поступления заявок - λ
- интенсивность потока обслуживания - μ
- интенсивность прохода очереди - v
- среднее время ожидания обслуживания в очереди - $T_{оч}$
- интенсивность нагрузки - ρ

Поток заявок называется простейшим, если вероятность поступления в систему ровно k заявок, $k = 0,1,2,\ldots$, в течение времени t определяется по формуле

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

где $\lambda > 0$ – постоянное число, называемое интенсивностью потока.