ВВЕДЕНИЕ В ТЕОРИЮ МАССОВОГО ОБСЛУЖИВАНИЯ ДЛЯ АНАЛИЗА СЛОЖНЫХ СИСТЕМ

Для оценки качества сложной системы с информационной точки зрения, в первую очередь, производительности системы, существуют следующие способы:

1. анализ производительности после реализации системы и внесении доработок в готовую систему;
2. выполнение оценки производительности на основании опыта разработки аналогичных систем;
3. аналитическая модель на основе теории массового обслуживания (теории очередей);
4. использование программ имитационного моделирования системного уровня.

Способ 1: используется в простых системах, которые можно легко откорректировать.

Способ 2: используется достаточно часто в относительно простых системах.

Способ 3: используется для оценочного расчета в системах с большим трафиком.

Способ 4: наиболее универсальный способ. Можно исследовать системы любой сложности.

Примеры программ имитационного моделирования компьютерных сетей:
ComNet, NetCracker.

Аналитический расчет выполняет оценочную функцию будущего проекта, с помощью теории массового обслуживания просто, быстро, без больших затрат вычислительных ресурсов, но результаты получаются менее точными, чем при моделировании.

Элементы модели СМО показаны на рис. 0.1. Элементы модели делятся на статические и динамические. К статическим элементам относятся обслуживающие устройства и очереди к ним, а к динамическим – транзакты или заявки на обслуживание.

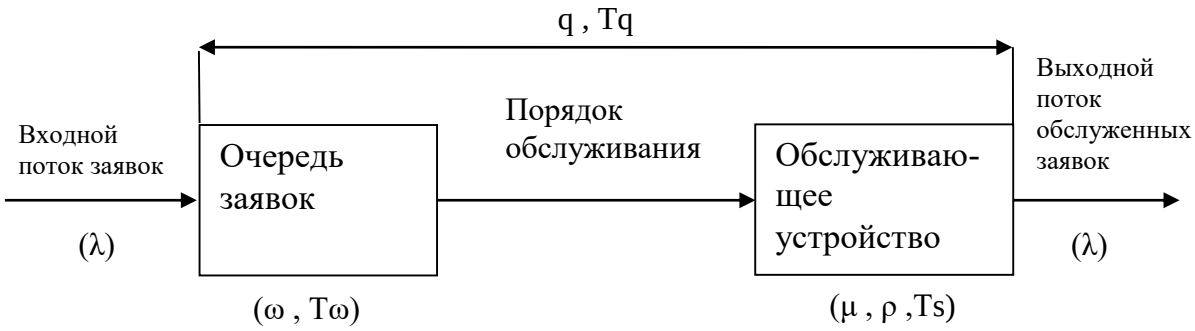


Рис.0.1. Элементы и параметры математической модели СМО

Модель СМО позволяет оценить такие параметры качества сложной системы:

λ – средняя скорость поступления заявок
ω – среднее количество заявок в очереди
Tω – среднее время ожидания заявки в очереди
Ts – среднее время обслуживания 1-й заявки
ρ – утилизация (загруженность) обслуживающего устройства
q – среднее количество элементов в системе

Tq – среднее время ожидания в системе

μ – средняя скорость обслуживания

Порядок обслуживания – это последовательность, в которой элементы выбираются из очереди устройством обслуживания

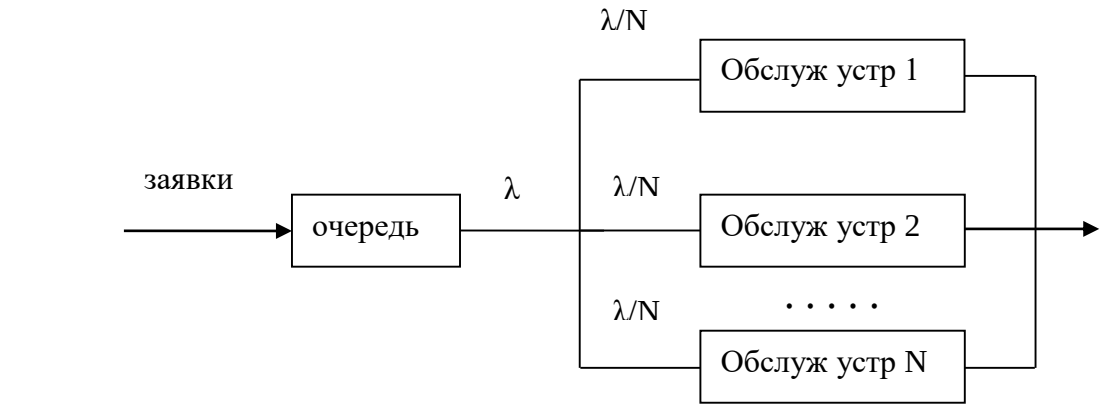
Возможны такие порядки как FIFO, приоритетный, взвешенное обслуживание и др.

Для использования математического аппарата теории очередей применяются определенные допущения.

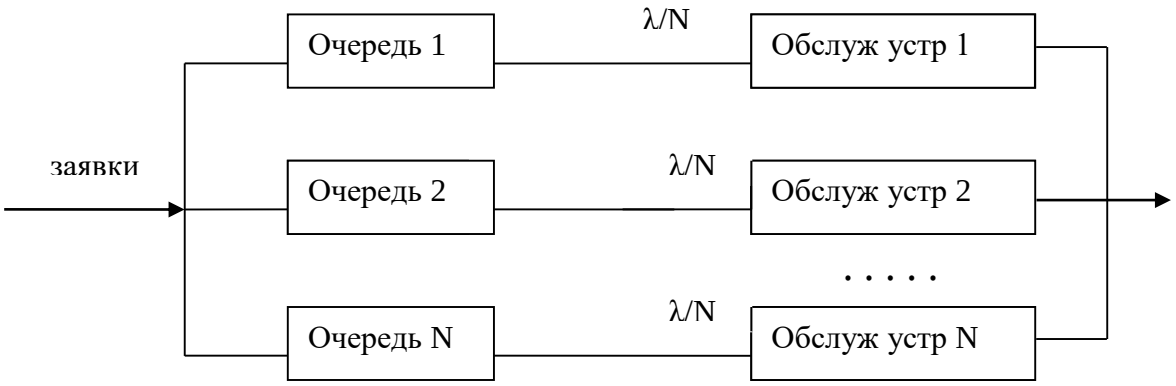
Допущения, принимаемые в СМО:

1. заявки в системе не теряются, емкость очереди бесконечна (очередь предназначена для сглаживания пульсаций поступления заявок);
2. если обслуживающее устройство свободно, оно немедленно выбирает из очереди заявку;
3. скорость поступления элементов в систему не зависит от числа элементов, которые в ней находятся;
4. после обработки заявки, она сразу покидает систему.

Существует ряд моделей СМО для описания систем с несколькими обслуживающими устройствами. Простейший случай модели с общей очередью к идентичным обслуживающим устройствам показан на рис.0.2.а. Такая модель, например, может соответствовать случаю, когда несколько процессоров (обслуживающих устройств) объединено в кластер. Можно сказать, что вероятность поступления заявок на обслуживающие устройства одинакова. Как только один из процессоров освобождается, он сразу выбирает следующую заявку из общей очереди. На рис.0.2.б. показан вариант модели с отдельными очередями к каждому обслуживающему устройству. Примером системы, описываемой такой моделью, может быть несколько одинаковых серверов, каждый из которых заранее закрепили за определенной группой пользователей в предположении, что активность групп примерно одинакова. В обоих случаях max скорость обслуживания системы будет N/Ts, т.е., пропускная способность системы возрастает во столько раз, сколько имеется обслуживающих устройств.



а



б

Рис.0.2. Модели СМО с несколькими обслуживающими устройствами: а – с общей очередью; б – с отдельными очередями.

Для каждой модели существуют готовые формулы расчета параметров. Для простоты будем рассматривать модель с общей очередью, тем более, что этот вариант соответствует многим системам, встречающимся на практике.

В рассматриваемой модели случайный характер носят:

- поток входящих заявок;
- время обслуживания заявок.

Случайный процесс поступления заявок описывается в модели функцией распределения интервалов времени между поступлениями заявок. Если его среднее значение равно T, то интенсивность поступления заявок в систему или скорость входного потока будет: $\lambda = 1/T$ заявок/сек.

Наиболее простые расчеты в СМО обеспечивает использование Пуассоновского закона распределения (рис.0.3). В аналитических расчетах практически всегда, когда последовательность событий разделена случайными интервалами времени и соблюдаются 3 приведенные ниже ограничения, используется Пуассоновский закон.

Ограничения эти следующие:

1) события происходят не зависимо друг от друга;
2) никогда не поступают 2 или более элементов данных одновременно;
3) среднее количество поступлений не изменяется со временем.

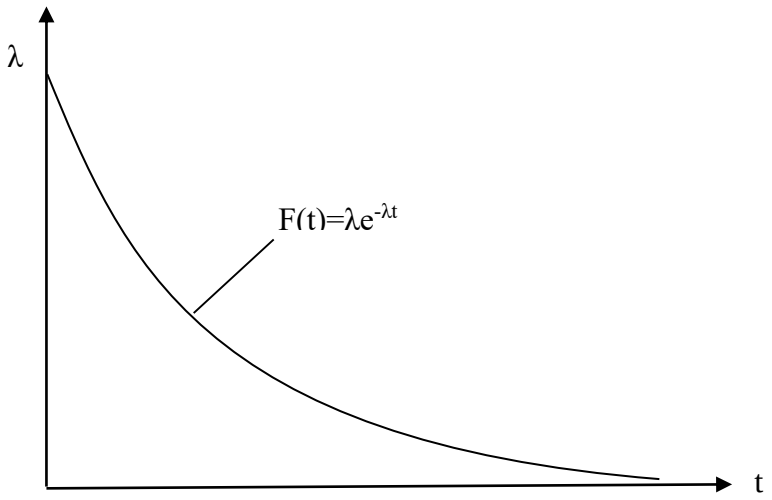


Рис. 0.3. Плотность распределения входного потока в соответствии с законом Пуассона

Из рисунка видно, что входной поток является существенно пульсирующим, так как есть не нулевая вероятность того, что интервал между заявками будет очень небольшим, близким к нулю, а также того, что он будет очень большим. Среднее отклонение интервалов также равно T, так, что стандартное отклонение равно T/T=1.

Время обслуживания чаще всего описывается экспоненциальным законом распределения, для которого справедливы те же самые 3 ограничения, что и для закона Пуассона. Разница состоит в том, что интервалы обслуживания могут чередоваться с

интервалом простоя, т.е. события случайного процесса не всегда плотно следуют друг за другом в отличие от Пуассоновского закона. (рис. 0.4.)

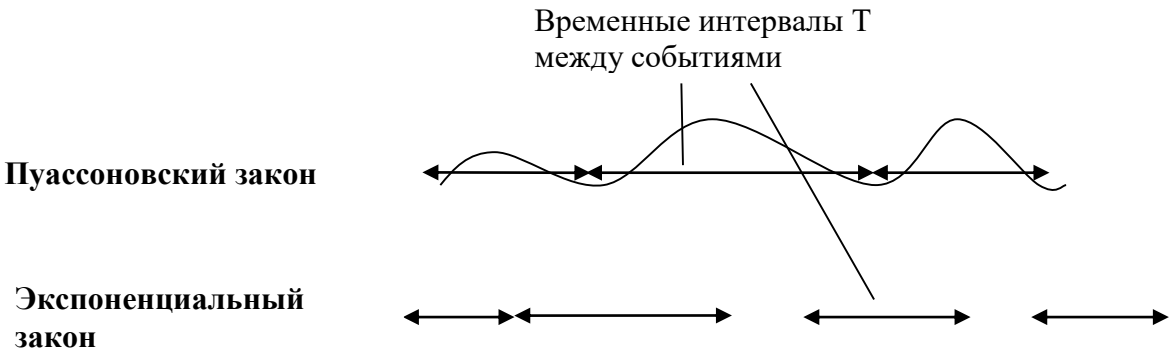


Рис.0.4. Последовательности событий, соответствующие Пуассоновскому и экспоненциальному закону распределения

Если считать, что среднее время обслуживания заявки Ts, то скорость обслуживания будет составлять $\mu=1/Ts$ заявок/сек. Коэффициент использования (загруженности) обслуживающего устройства будет составлять $\rho=: \lambda/\mu$.

Датский ученый Эрланг изучал теорию очередей применительно к телефонным сетям. Для классификации наиболее вероятных случаев организации систем с очередями Эрлангом была предложена классификация моделей и выведены формулы расчета их параметров.

Тип модели обозначаются: X /Y /N, где

X – закон распределения времени поступления элементов;

Y – закон распределения времени обслуживания элементов;

N – количество обслуживающих устройств.

G – нормальный закон для времени поступления или обслуживания элементов (заявок).

M – Пуассоновский закон для времени поступления заявок и Пуассоновский или экспоненциальный закон для времени обслуживания.

D – детерминированный закон для времени поступления или обслуживания.

Пример: M/M/1 - система с одним обслуживающим устройством, Пуассоновским законом распределения времени поступления заявок и экспоненциальным законом их обслуживания.

Время ответа системы, в которой используется разделяемая среда и случайный фактор доступа к ней, в зависимости от ее загруженности имеет экспоненциальный характер. Поскольку время ответа складывается из времени поступления в систему, времени доставки пользователю, времени обслуживания и времени ожидания заявки в очереди, то понятно, что изменение коэффициента загруженности системы влияет именно на последнюю составляющую (рис.0.5.).

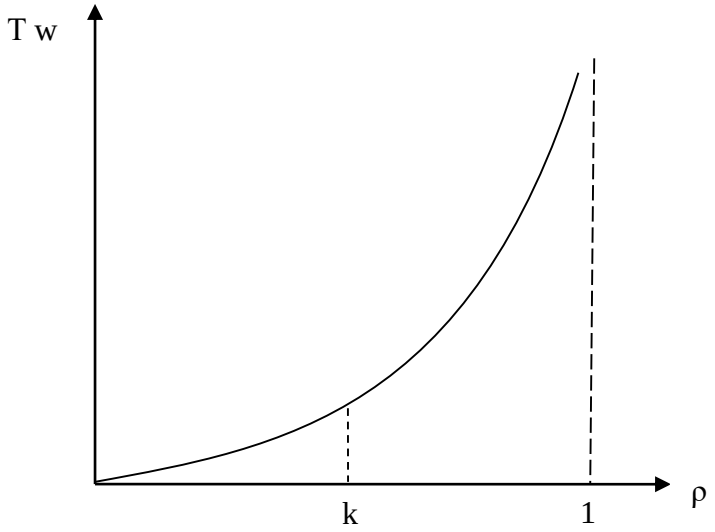


Рис.0.5. Зависимость времени ожидания заявки в очереди от коэффициента использования обслуживающего устройства

Теоретически, если :λ=1/Ts, тогда ρ=1 и средняя скорость поступления заявок в систему будет равна средней скорости обслуживания в ней. При этом размер очереди резко возрастает, и стремиться к ∞. Дело в том, что λ и μ – это средние значения соответствующих величин на длительном промежутке времени. На небольших временных интервалах эти значения могут существенно отклоняться от своих средних значений. В периоды, когда ρ>1 формируется очередь, а когда ρ<1 рассасывается. Увеличения времени ожидания не будет только в том случае, когда интервалы перегрузок будут чередоваться с достаточными по времени интервалами недогрузки, в течение которых очередь будет успевать рассасываться полностью. Если на интервале недогрузки очередь рассосалась не полностью, то на следующем интервале перегрузке ее длина увеличится. С ростом среднего значения ρ вероятность такого развития событий, когда накопление заявок в очереди неуклонно растет, будет увеличиваться.

Значение k, при котором время ожидания начинает нелинейно расти, зависит от особенностей метода доступа к разделяемой среде или дисциплины обслуживания.

В простейшем случае доступа – выборка на обработку из общей очереди в порядке FIFO, которым часто пользуются коммутаторы, маршрутизаторы, сервера и др. обслуживающие устройства, значение k в общем случае будет зависеть от коэффициента стандартного отклонения ρ от среднего. Отклонение ρ, в свою очередь, в основном зависит от степени пульсации входного потока (коэффициента стандартного отклонения скорости поступления λ от среднего), потому, что время обслуживания практически всегда варьируется в значительно меньших пределах.

Определение коэффициентов стандартного отклонения λ и μ полезно для оценки размеров очередей, а, следовательно, и буферов устройств в системе.

На практике выбирают значение ρ=0.6÷0.9. К сожалению модели СМО не дают простых аналитических зависимостей, позволяющих вычислить значение k в зависимости от конкретных значений коэффициентов отклонения λ и μ. Поэтому для получения таких данных прибегают к средствам имитационного моделирования или проведению измерений в реальных системах. Общая закономерность проявляется в том, что с увеличением степени пульсации входного потока значение k уменьшается. Следовательно, для снижения задержек обслуживания в системе необходимо снижать значение ρ и сглаживать пульсации входного потока.

Наилучшие характеристики будет иметь система, которая описывается детерминированными (определенными) законами распределения для μ и λ. Наихудшие

значения покажет расчет с использованием экспоненциального закона для μ и Пуассоновского - для λ. Эту модель можно рассматривать как модель «наихудшего случая». Если коэффициент стандартного отклонения от среднего для λ в реальных наблюдениях будет < 1, то скорость поступления нужно брать постоянной, так как Пуассоновский закон даст завышенную оценку длины очередей и задержек. Для его применения нужно, чтобы коэффициент стандартного отклонения был равен 1. А, если коэффициент будет >1 (сверх пульсирующий входной поток), то вероятность перегрузки системы будет больше рассчитанной. В этом случае необходимо выбирать режимы работы реальной системы с меньшими значениями ρ.

Самый простой способ обеспечения минимальных задержек в системе – это ее работа в «недогруженном режиме» при ρ= 0.2-0.3. Но такой режим делает использование системы неэффективным. Это особенно актуально в глобальных сетях, оборудование которых стоит очень дорого. Поэтому в коммутаторах и маршрутизаторах для разных типов трафика создают разные очереди, а также могут выделять для их обслуживания разное количество ресурсов. Например, для чувствительного к задержкам трафика с λ1 в выходном порту маршрутизатора выделяют отдельную очередь и обеспечивают ее первоочередную обработку из такого расчета, чтобы ρ1=λ1/μ=0.2-0.3. Для нечувствительного к задержкам трафика с λ2 выделяют другую очередь, которую обслуживают только тогда, когда первая очередь будет пуста. Для второй очереди должно обеспечиваться ρ2=λ2/μ=0.5-0.6. В этом случае для нечувствительного трафика реальный коэффициент загрузки порта будет составлять ρ=λ1+λ2/μ≈0.9. Для того, чтобы очередь нечувствительного трафика не переполнялась, и пакеты не терялись, необходимо предусмотреть для нее большую буферную память, оценить размеры которой можно с помощью модели СМО.