Аналитико-статистическое моделирование информационных систем

Кафедра информационных управляющих систем

Лекции читает

канд.техн.наук, доцент

Литвинов Владислав Леонидович

Список литературы:

1. О.И. Кутузов, Т.М. Татарникова
   МОДЕЛИРОВАНИЕ ТЕЛЕКОММУНИКАЦИОННЫХ СЕТЕЙ
   http://dvo.sut.ru/libr/ius/w101kutu/index.htm
2. *Боев В. Д,* Моделирование систем. Инструментальные средства GPSS WORLD. Учеб. пособие — СПб..: БХВ-Петербург, 2004. — 368 с.
3. *Боев В. Д, Сыпченко Р. П.* Компьютерное моделирование. Элементы теории и практики. Учеб. пособие — СПб..: Военная академия связи, 2009. — 432 с.
4. *Бражник А. Н,* Имитационное моделирование: возможности GPSS WORLD — СПб..: Реноме, 2006. — 439 с.

Тема лекции 4:

- **Система обслуживания M/G/1**

Для исследования системы массового обслуживания (однолинейная система с пуассоновским входным потоком и произвольным распределением времени обслуживания) используется подход, отличный от описанного выше. Этот подход состоит в изучении случайного процесса изменения состояний системы в моменты окончания обслуживания сообщений (передача сообщения заканчивается, и оно покидает концентратор).
В буфер поступает последовательность сообщений, каждое из которых обладает случайной переменной длиной (или временем обслуживания). Обслуживающее устройство обрабатывает сообщения поочередно (т. е. по каналу сообщения передаются одно за другим). Закончив обработку одного сообщения, оно приступает к обработке следующего сообщения по принципу «первым пришел-первым обслужен».

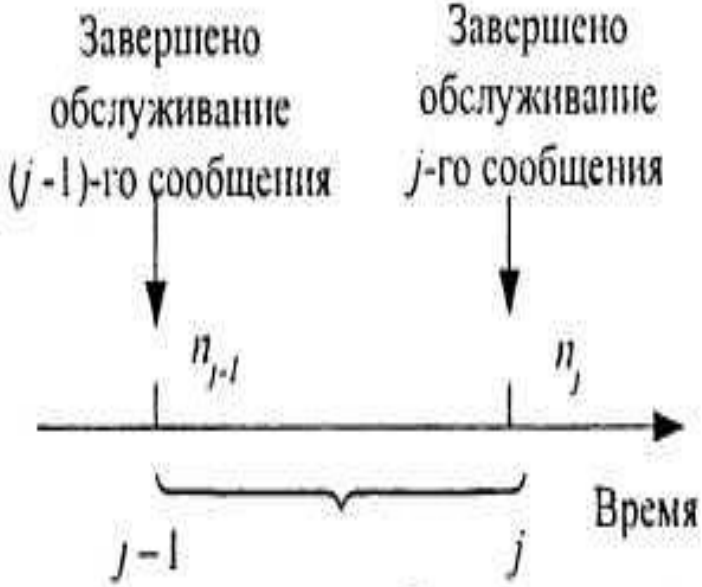

Рис. 8.9. Временная диаграмма обслуживания *j*-го сообщения

- Пусть n_j - длина очереди после окончания обслуживания *j-го* сообщения (j - переменный индекс). Можно записать простое соотношение, связывающее n_j и n_{j-1} — длину очереди после завершения обслуживания (j-1)-го сообщения. Получим
- $$n_j = \begin{cases} (n_{j-1} - 1) + v_j, & n_{j-1} >= 1 \\ v_j, & n_{j-1} = 0 \end{cases} \quad (1)$$
- Здесь v_j - число сообщений, поступивших в течение времени обслуживания *j-го* сообщения (или *j-го* требования). (Это, естественно, также случайная величина.)
- После преобразований получаем выражение для средней длины очереди:
- $$M(n) = \frac{1}{1-\rho}\left[\rho - \frac{1}{2}\rho^2(1-\mu^2\sigma^2)\right] \quad (2)$$
- Здесь σ^2 — дисперсия длины сообщения.

- Таким образом, средняя длина очереди сообщений в буфере зависит непосредственно от ρ, средней длительности сообщения $1/\mu$ и дисперсии длины сообщения σ^2.
- ***Пример 1***. Пусть длительность сообщения распределена по экспоненциальному закону. Тогда
- $\sigma^2 = 1/\mu^2$ *и* $M(n) = \rho/(1-\rho)$.
- Этот результат был получен ранее для системы *M/ M/1.*
- ***Пример 2***. Пусть сообщения имеют фиксированную длительность
- $\tau_0 = 1/\mu$.
- Тогда
  $\sigma^2 = 0$ и $M(n) = (1 - \rho/2)(\rho/[1 - \rho])$.
- Таким образом, при сообщениях фиксированной длительности средняя занятость буферной памяти, меньше, чем для модели *M/M/1.*

- Теперь, используя формулу Литтла, можно найти среднее время задержки для сообщений, поступающих в буферную память. Учитывая, что
- ρ=λ/μ для модели *MIGI1* получим
- M(T)= (1/λ)M(n) = $\frac{1}{2\mu(1-\rho)}\left[2-\rho(1-\mu^2\sigma^2)\right]$ (3)
- Для экспоненциального распределения длины сообщения, когда
- $\sigma^2\mu^2 = 1$, находим
- M(T)=1/[μ(1-ρ)].
- При фиксированной длине сообщения
- M(T)=(1-ρ/2)/[μ(1-ρ)]
- Другие случаи можно исследовать аналогично.

- Выражение для среднего времени задержки в системе *MIGI1* аналогично выражению, полученному для системы *Ml Ml1,* и отличается лишь наличием в формуле (3) второго члена в скобках, зависящего от величины разности $(1 - \sigma^2\mu^2)$. Для законов распределений длин сообщений, у которых $\sigma^2 < 1/\mu^2$, среднее время задержки меньше, чем в случае экспоненциального распределения длины сообщения. Для законов распределений длин сообщений, у которых $\sigma^2 > 1/\mu^2$, среднее время задержки выше.
- Интуитивно ясно, что с увеличением дисперсии вероятность появления более длинных сообщений увеличивается и, следовательно, время задержки растет.

- Формула (3), называемая также формулой Поллачека - Хинчина, определяет среднее время задержки для модели *MIGI1.* Эквивалентная формула в более компактной форме может быть записана для среднего времени, затрачиваемого на ожидание в очереди на обслуживание. Это время равно разности времен задержки и обслуживания (передачи) сообщения.
- Формула для времени ожидания будет использоваться позже при
- описании систем массового обслуживания с приоритетами. Среднее время *M(T)* равно сумме среднего времени ожидания $M(T_{0ж})$ и среднего времени обслуживания (передачи) сообщения $1/\mu$ (рис.).
- $M(T) = M(T_{0ж}) + 1/\mu.$ (4)
- Подставляя выражение (4) в (3), решая его относительно $M(T_{0ж})$ и упрощая, получим
- $M(T_{0ж})=(\lambda/2)[M(\tau^2)/(1-\rho)]$ (5)
-

  Таким образом, среднее время ожидания $M(T_{0ж})$ зависит от второго момента $M(\tau^2)$ распределения длины сообщения.

Рис. 8.10. Время ожидания в системе обслуживания

$$M(T_{ож}) = \frac{\lambda}{2} \frac{M(\tau^2)}{1-\rho}.$$

Сети с большим числом узлов, соединенных каналами связи

- Рассмотрим сеть, содержащую большое число узлов, соединенных каналами связи. Пусть для буферной памяти, связанной с i-й линией, интенсивность входного потока λ_i, среднее число ожидающих и обслуживающихся сообщений $M(n_i)$ и среднее время задержки $M(T_i)$. Тогда, согласно теореме Литтла,
- $M(T_i) = M(n_i)$, $i = 1...m$, где m - число узлов в сети.
- Рассмотрим теперь модель полной сети, заключенную в традиционный «черный ящик». Пусть Υ - среднее число сообщений, поступающих в «черный ящик» за единицу времени; $M(T)$ - среднее значение времени задержки сообщений в сети и $M(n)$ - среднее число сообщений, находящихся в сети. «Черный ящик» сам ведет себя как система обслуживания, для которой по теореме Литтла можно записать:
- $M(T)\Upsilon = M(n)$.

- Тогда, суммируя по всем системам обслуживания в сети и применяя теорему Литтла к каждой из них, получим

$$M(n) = \sum_{i=1}^{m} M(n_i) = \sum_{i=1}^{m} \lambda_i M(T_i)$$

- Часто представляет интерес нахождение вероятности того, что размер очереди превышает определенную величину. Выражение для этой вероятности обычно используют при решении задачи выбора объема буферной памяти при наличии ограничения на вероятность переполнения этой памяти. Может быть найдена вероятность того, что число сообщений, ожидающих в очереди в системе *M/M/1,* превысит заданное число *N:*

$$P(n < N) = \sum_{n=N+1}^{\infty} p_n = (1-\rho) \sum_{n=N+1}^{\infty} \rho^n = \rho^{N+1}$$

- Как видно, вероятность уменьшается экспоненциально с ростом *N,* что следует из выражения .
- При ρ=0,6
- N=1 P(n<N)=0,36
- N=3 0,13
- N=9 $6,1*10^{-3}$
- N=19 $3,7*10^{-5}$

- Рассмотрим модель обслуживания, в которой предполагаются ограниченный объем буферной памяти, пуассоновский входной поток и экспоненциальное распределение длин сообщений. При выводе формулы (3) для стационарных вероятностей длины очереди не использовалось предложение о неограниченном объеме буферной памяти. Следовательно, выражение для $p_п$ является справедливым и в случае ограниченного объема буферной памяти, с той лишь разницей, что сумма вероятностей конечного числа состояний должна быть равна единице.

- Используя выражения для суммы членов геометрической прогрессии, имеем

$$\sum_{n=0}^{N} p_n = 1 = p_0 \sum_{n=0}^{N} \rho^n = p_0 \frac{1-\rho^{N+1}}{1-\rho}$$

- Следовательно,

- $$p_0 = \frac{1-\rho}{1-\rho^{N+1}} \quad \text{и} \quad p_n = \frac{(1-\rho)\rho^n}{1-\rho^{N+1}} \qquad (11)$$

- Вероятность того, что буферная память заполнена и очередное поступающее сообщение получит отказ (блокируется), равна вероятности того, что N сообщений находятся в буферной памяти:

- $$p_N = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}} \qquad (12)$$

Рассмотрим систему обслуживания *M/M/1* с ограниченным объемом буферной памяти (рис. 8.11). Предположим, что интенсивность входного потока - *λ* (сообщений/с) и объем буферной памяти рассчитан на прием и хранение только *N* сообщений. Если вероятность блокировки обозначить через $P_Б$, то интенсивность потока сообщений, поступающего на вход системы, должна составлять $\lambda(1- P_Б)$ сообщений/с

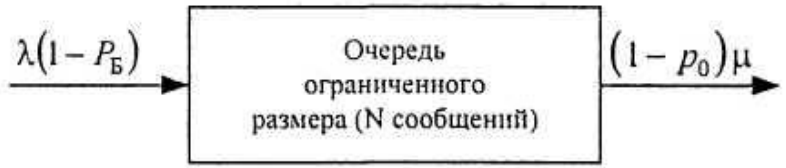

Рис. 8.11. Система обслуживания *M/M/1* с ограниченным объемом буферной памяти

- В стационарном режиме эта интенсивность должна совпадать с интенсивностью выходящего потока. Вероятность того, что сообщение покинет систему за интервал Δt, равна μΔt, если хотя бы одно сообщение находится в системе.
- Вероятность того, что очередь не пуста, определяется как $(1 - p_Q)$. Следовательно, интенсивность выходного потока равна $(1 - p_0)\mu$
- и тогда
- $\lambda (1-P_В)=(1- p_0)\mu$ (20)

- Для системы *Ml Ml1* с ограниченным объемом буферной памяти вероятность p_o определяется формулой (11). Подставляя ее в (20), и учитывая, что $\rho = \lambda/\mu$, получим выражение для вероятности блокировки $P_Б$, которое совпадает с (12).

Приоритетное обслуживание

- В сетях связи для ЭВМ характерной является передача сообщений с различными приоритетами. Коротким сообщением, содержащим подтверждения, часто назначают более высокий приоритет, чем информационным сообщениям. По сети могут передаваться сообщения двух и более категорий срочности. Например, некоторые пользователи, передающие в среднем сообщения более короткие, чем у других абонентов, получают приоритет для ускорения общей доставки сообщений. В связи с этим представляет интерес исследование системы *M/G/1* с несколькими классами сообщений, обладающих разными приоритетами.

- Для упрощения этой задачи основное внимание будет уделено определению среднего времени ожидания, а не времени задержки. Как видно из выражения (4), среднее время задержки всегда можно получить, добавив среднее время передачи сообщения к среднему времени ожидания. Будем предполагать, что сообщения разных классов обладают относительными приоритетами. При этом сообщение с более высоким приоритетом располагается в очереди перед сообщениями с более низким приоритетом, но уже начавшееся обслуживание сообщений с более низким приоритетом не прерывается.
- В рассматриваемой системе обслуживания предполагается, что классы сообщений, обозначаемые индексом $p = 1, 2, 3, ..., r,$ пронумерованы в порядке уменьшения приоритета. Рассмотрим обслуживание (начало передачи) с момента времени t_1 с целью получения общего соотношения для среднего времени $M(T_{ож})$ ожидания сообщения с приоритетом p.

- Для этого разберем, из каких компонентов складывается $T_{ож}$. Очевидно, что сюда входят: время T_0 необходимое для завершения текущего обслуживания; времена $T_к$ необходимые для обслуживания $m_к$ сообщений с приоритетами $к$ = 1, 2, ...,p-1, уже ожидающих обслуживания в очереди к моменту поступления рассматриваемого сообщения, и времена $T_к^{1}$ $к$=1,2, ..., p – 1 необходимые для обслуживания сообщений с более высоким приоритетом, которые могут поступить за интервал ожидания и будут обслужены раньше данного сообщения. Суммируя средние значения всех этих случайных величин, получим

$$M(T_{ожс})_p = M(T_0) + \sum_{k=1}^{p} M(T_k) + \sum_{k=1}^{p-1} M(T_k^1)$$

- Для оценки $M(T_к)$ допустим, что среднее число ожидающих сообщений с приоритетом *к* составляет $M(m_к)$. Если каждое из них требует для обслуживания в среднем $1/\mu_к$ единиц времени, то
- $M(T_к) = M(m_к)/\mu_к$. (22)
- Но $M(m_к)$ представляет собой разность двух величин - среднего числа сообщений, ожидающих и обслуживаемых в системе $M(n_к)$, и среднего числа обслуживаемых сообщений. Число последних составляет
- $\rho_к = \lambda_к/\mu_к$, где $\lambda_к$ - интенсивность потока сообщений *к-й* категории. Из теоремы Литтла следует, что

- $M(n_к) = M(m_к) + \rho_k$
- Следовательно
- $M(T_к) = \rho_k M(T_{ож})_k$

- По аналогии
- $M(T_к^{1}) = \rho_k M(T_{ож})_p$

Можно показать, что время ожидания для сообщений с приоритетом p можно найти по формуле

$$M(T_{ож})_p = \frac{M(T_0)}{(1-\sigma_p)(1-\sigma_{p-1})}$$

Где

$$\sigma_p = \sum_{k=1}^{p} \rho^k \qquad (23)$$

- Определим теперь величину времени $M(T_0)$, необходимого для завершения текущего обслуживания. Рассмотрим сначала систему обслуживания *MIGI1* с одним классом требований. Сравнивая выражения (5) и (23), получим в этом случае
- $M(T_0) = \lambda M(\tau^2)/2.$
- С целью проверки предположим, что распределение длин сообщений экспоненциальное. Тогда легко показать, что $M(T_0) = \rho/\mu.$
- Указанная величина может рассматриваться как произведение вероятности занятости системы обслуживания ρ на среднюю длину сообщения 1/μ.
- В более общем случае, для системы обслуживания с несколькими классами требований, получим

$$M(T_0) = \frac{1}{2}\sum_{k=1}^{r}\lambda_k M(\tau_k^2), \lambda = \sum_{k=1}^{r}\lambda_k$$

Система обслуживания *MIMINIm*

- Пусть на СМО *MIMINIm* с числом обслуживающих приборов N и числом мест для ожидания *m* поступает поток заявок с интенсивностью λ, которые обслуживаются каждым прибором с интенсивностью μ.
- Пусть также время ожидания в очереди распределено по экспоненциальному закону с параметром (интенсивностью) v.
- Определим вероятность обслуживания требований, вероятность ожидания требованием начала обслуживания и среднее время ожидания. (Задача Бухмана)
- В рассматриваемой СМО существуют следующие рабочие состояния:

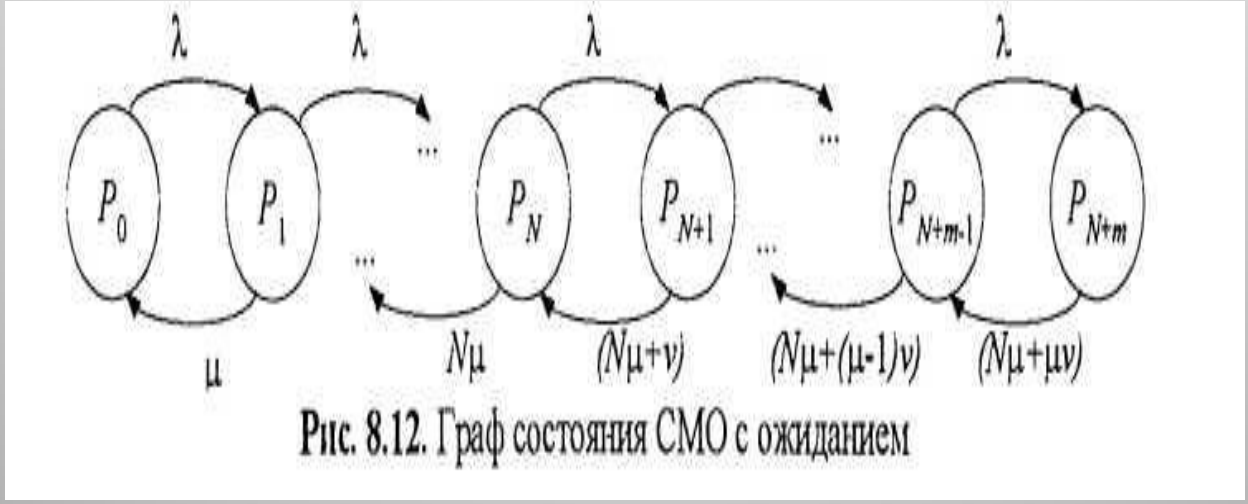

Рис. 8.12. Граф состояния СМО с ожиданием

- система обслуживает *s* требований с интенсивностью *sμ,* если *0<=s<N;*
- система ставит требование в очередь, если число требований больше числа обслуживающих приборов, но меньше числа мест ожидания *N<= s <m,* при этом интенсивность поступления требований из очереди равна *(s - N)v*
- система отказывает требованиям в обслуживании, если *s > (N+ m).*
- Под состоянием сети будем понимать значение числа требований, находящихся на обслуживании (в системе распределения ресурса и в очереди) в момент времени *t.* Обозначим через s = O, ..., *S* номер состояния СМО (число требований в ней), где *S= N+m.*
- Для аппроксимации вероятностно-временного механизма перехода СМО из одного состояния в другое используем аппарат марковских цепей.
- Решение задачи было найдено Эрлангом

- Среднее время ожидания начала обслуживания

- $T_{ож} = P(t_{ож}>0)/(\mu N-\lambda)$

- Средняя длина очереди вычисляется по формуле Литтла.

- Математический аппарат ТМО охватывает широкий класс СМО с простейшими, примитивными и рекуррентными потоками и может быть использован для анализа и синтеза СМО с отказами, с ожиданием и ненадежными единицами ресурса. Трудность аналитического разрешения уравнений состояния для СМО большой размерности делает целесообразным применение для их исследования методов имитационного моделирования и численных методов расчета на ЭВМ. Особо следует отметить важность постановки и решения оптимизационных задач для СМО. В качестве целевых функций критериев при этом целесообразно использовать полученные вероятностно-временные характеристики (ВВХ), а оптимизируемыми переменными могут стать интенсивности входящего потока требований, число мест для ожидания, число обслуживающих приборов, дисциплина обслуживания, алгоритм предоставления ресурса.