**Министерство образования Республики Беларусь**

**Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»**

**Кафедра Сетей и устройств телекоммуникаций**

**РЕФЕРАТ**

**На тему:**

«**Дисциплины обслуживания. Модель с приоритетами. Дисциплины обслуживания с приоритетами, зависящими от времени**»

**МИНСК, 2008**

**Дисциплины обслуживания. Модель с приоритетами.**

Дисциплина обслуживания – это способ определения того, какое требование в очереди должно обслуживаться следующим. Решение может основываться на одной из приведенных ниже характеристик или на их совокупности:

1. мера, определяемая относительным временем поступления рассматриваемого требования в очередь;
2. мера требуемого или полученного до сих пор времени обслуживания;
3. функция, определяющая принадлежность требования к той или иной группе.

Примерами дисциплин обслуживания являются постоянно используемая модель «первый пришел - первый обслужен» (FCFS-first came-first served), называемая в русскоязычной литературе «дисциплина обслуживания в порядке поступления»-ОПП. Приведем здесь список некоторых типичных дисциплин обслуживания.

ОПП-обслуживание в порядке поступления (FCFS);

ООП – обслуживание в обратном порядке, т.е. последнее поступившее требование обслуживается первым (LCFS);

ПК – первоочередное обслуживание требований с кратчайшей длительностью обслуживания (SPT/SJE);

ПКД – первоочередное обслуживание требований с кратчайшей длительностью дообслуживания (SRPT);

ПКС – первоочередное обслуживание требований с кратчайшей средней длительностью обслуживания (SEPT);

ПКСД – первоочередное обслуживание требований с кратчайшей средней длительностью дообслуживания (SERPT);

ПКОВ – первоочередное обслуживание требований с кратчайшим обязательным временем (SIPT).

Если сравнивать эти дисциплины по среднему времени ожидания попарно, и обозначать тот факт, что среднее время ожидания для дисциплины D1 ,больше или равно среднему времени ожидания для дисциплины D2 следующим образом: D1→D2, то можно построить следующую диаграмму

ПО → ОПП → ПК → ПКД

↓ ↑

↓ → ОП →→→→ ↑

Итак, основным предметом анализа различных дисциплин обслуживания будем считать расчет среднего времени ожидания требования в очереди или среднего времени пребывания в системе.

Предположим, что требования принадлежат одному из *P* различных приоритетных классов, обозначаемых индексом *p*=1,2,3…*P*. Каждому требованию, находящемуся в системе в момент времени *t* ставится в соответствие значение некоторой приоритетной функции *qp(t).* Чем больше значение этой функции, тем выше приоритет требования. Всякий раз, когда принимается решение для выбора требования на обслуживание, выбор делается в пользу требования с наибольшим значением приоритетной функции. В простейшем случае в качестве приоритетной функции выбирается просто значение *p*. В этом случае приоритет требования тем больше, чем больший номер класса принадлежности оно имеет. Рассмотрим достаточно общую модель, основанную на системе M/G/1. Предположим, что требования из приоритетного класса *p* образуют пуассоновский поток с интенсивностью *λp* требований в секунду. Время обслуживания каждого требования из этого класса выбирается независимо в соответствие с распределением с плотностью вероятности *bp(x)* со средним значением

.

Введем следующие определения

Здесь *ρ* интерпретируется как доля времени, в течение которого сервер занят (*ρ*<1), а каждый из парциальных коэффициентов *ρp* – как доля времени, в течение которого сервер занят обслуживанием заявок из приоритетного класса с номером *p*.

Если требование в процессе обслуживания может быть удалено из сервера и возвращено в очередь при поступлении требования с более высоким приоритетом, то говорят, что система работает с абсолютным приоритетом, если обслуживание любого требования, находящегося в сервере не может быть прервано, то говорят что СМО работает с относительным приоритетом.

##

## Основная модель расчета среднего времени ожидания

Будем использовать далее следующие обозначения для среднего значения времени ожидания в очереди требований из приоритетного класса *p* - *Wp*, и среднего времени пребывания в системе для требований этого класса - *Tp*:

.

Основное внимание будем уделять системам с относительным приоритетом. Рассмотрим процесс с момента поступления некоторого требования из приоритетного класса *p*. Будем далее называть это требование меченым. Первая составляющая времени ожидания для меченого требования связана с требованием, которое оно застает в сервере. Эта составляющая равна остаточному времени обслуживания другого требования. Обозначим теперь и будем использовать это обозначение и далее, среднюю задержку меченого требования, связанную с наличием другого требования на обслуживании *W0*. Зная распределение времени между соседними поступлениями входных требований для каждого приоритетного класса, можно всегда вычислить эту величину. В нашем предположении пуассоновского закона для потока заявок каждого класса можно записать

.

Вторая составляющая времени ожидания для меченого требования определяется тем, что перед меченым требованием обслуживаются другие требования, которые меченое требование застало в очереди. Обозначим далее число требований из класса *i*, которое застало в очереди меченое требование (из класса *p*) и которые обслуживаются перед ним *Nip*. Среднее значение этого числа будет определять величину среднего значения этой составляющей задержки

.

Третья составляющая задержки связана с требованиями, поступившими после того как пришло меченое требование, однако получившими обслуживание раньше его. Число таких требований обозначим *Mip*. Среднее значение этой составляющей задержки находится аналогично и составляет

.

Складывая все три составляющие, получаем, что среднее время ожидания в очереди для меченого требования определяется формулой

.

Очевидно, что независимо от дисциплины обслуживания число требований, *Nip* и *Mip* в системе не может быть произвольным, поэтому существует некоторый набор соотношений, связывающий между собой задержки для каждого из приоритетного класса. Важность этих соотношений для СМО позволяет называть их ЗАКОНАМИ СОХРАНЕНИЯ. Основой законов сохранения для задержек является тот факт, что незаконченная работа в любой СМО в течение любого интервала времени занятости не зависит от порядка обслуживания, если система является консервативной (требования не исчезают внутри системы и сервер не простаивает при непустой очереди).

Распределение времени ожидания существенно зависит от порядка обслуживания, но если дисциплина обслуживания выбирает требования независимо от времени их обслуживания (или любой меры, зависящей от времени обслуживания), то распределение числа требований и времени ожидания в системе инвариантно относительно порядка обслуживания.

Для СМО типа M/G/1 можно показать, что для любой дисциплины обслуживания должно выполняться следующее важное равенство

Это равенство означает, что взвешенная сумма времен ожидания никогда не изменяется, независимо от того, насколько сложна или искусно подобрана дисциплина обслуживания. Если удается сократить задержку для одних требований, то она немедленно возрастет для других.

Для более общей системы с произвольным распределением времени поступления требований G/G/1 закон сохранения может быть записан в виде

.

Общий смысл этого соотношения таков: взвешенная сумма времен задержки остается постоянной. Просто в правой части стоит разность средней незавершенной работы и остаточного времени обслуживания. Если предположить пуассоновский характер входного потока, то выражение для незавершенной работы можно записать в виде

.

Подставляя его в предыдущее выражение, сразу получается приведенный ранее закон сохранения для СМО типа M/G/1.

Рассмотрим теперь расчет среднего времени ожидания для СМО с обслуживанием в порядке приоритета, задаваемого приоритетной функцией

.

На рис. 1 приведена схема функционирования СМО с такой дисциплиной обслуживания: поступающее требование ставится в очередь слева от требования с равным или большим приоритетом.

Рис. 1 СМО с обслуживанием в порядке приоритета.

Воспользуемся формулой для *Wp*. Исходя из механизма функционирования, можно сразу выписать

Все требования более высокого, чем у меченого приоритета будут обслужены раньше. Из формулы Литтла число требований класса *i* находящихся в очереди, будет равно:

Требования более высокоприоритетных классов, поступившие в систему после меченого требования, пока оно находится в очереди, также будут обслужены перед ним.

Так как меченое требование будет находиться в очереди в среднем *Wp* секунд, то число таких требований будет равно

.

Непосредственно из формулы (\*) получаем:

Эта система уравнений может быть решена рекуррентно, начиная с *W1,W2* и т.д.

Полученная формула позволяет рассчитывать характеристики качества обслуживания для всех приоритетных классов. На рисунке 2 показано, как изменяется нормированная величина времени ожидания в очереди для СМО с пятью приоритетными классами с равной интенсивностью потока требований каждого приоритетного класса и равным средним временем обслуживания требований каждого класса (нижний рисунок детализирует кривые при значениях малой нагрузки).

Рис. 2 Обслуживание в порядке приоритетов в случае относительных приоритетов (Р=5, λР= λ/5, ).

Особую задачу представляет определение законов распределения времени ожидания.

Рассмотрим теперь систему с абсолютными приоритетами и обслуживанием в порядке приоритета с дообслуживанием. Применим подход полностью аналогичный рассмотренному ранее. Средняя задержка в системе меченого требования также состоит из трех составляющих: первая составляющая- это среднее время обслуживания, вторая – это задержка из-за обслуживания тех требований равного или более высокого приоритета, которые меченое требование застало в системе. Третья составляющая средней задержки меченого требования представляет собой задержку за счет любых требований, поступающих в систему до ухода меченого требования и имеющих строго больший приоритет. Расписывая все эти три составляющие общего времени нахождения в системе, получим

.

Весьма интересной задачей является выбор приоритетов для заявок различных классов. Поскольку имеет место закон сохранения, оптимизация имеет смысл только при рассмотрении некоторых дополнительных атрибутов каждого класса требований. Предположим, что можно оценить каждую секунду задержки заявки приоритетного класса p некоторой стоимостью *Cp*. Тогда средняя стоимость секунды задержки для системы может быть выражена через среднее число требований каждого класса, находящихся в системе

Решим задачу нахождения дисциплины обслуживания с относительными приоритетами для системы M/G/1, которая минимизирует среднюю стоимость задержки *C*. Пусть имеется *P* приоритетных классов заявок с заданной интенсивностью поступления и средним временем обслуживания. Перенесем в левую часть постоянную сумму и выразим правую часть через известные параметры

Задача состоит в минимизации суммы в правой части этого равенства путем выбора соответствующей дисциплины обслуживания, т.е. выбора последовательности индексов *p*.

Обозначим

В этих обозначениях задача выглядит так: нужно минимизировать сумму произведений при условии

Условие независимости суммы функций *gp* от выбора дисциплины обслуживания определяется законом сохранения. Иначе говоря задача состоит в минимизации площади под кривой произведения двух функций , при условии , что площадь под кривой одной из них постоянна.

Решение состоит в том, что сначала упорядочим последовательность значений *fp*: .

А затем выберем для каждого *fp* свое значение *gp*, так, чтобы минимизировать сумму их произведений. Интуитивно ясно, что оптимальная стратегия выбора состоит в подборе наименьшего значения *gp* для наибольшего *fp* , далее для оставшихся значений следует поступать тем же образом. Поскольку *gp*=*Wpρp*, то минимизация сводится к минимизации значений средней задержки. Таким образом, решение рассматриваемой задачи оптимизации состоит в том, что из всех возможных дисциплин обслуживания с относительным приоритетом минимум средней стоимости обеспечивает дисциплина с упорядоченными приоритетами в соответствие с неравенствами

.

**Дисциплины обслуживания с приоритетами, зависящими от времени**

На практике часто встречается задача назначения приоритетов в зависимости от времени поступления заявки. Например, для того, чтобы никакие требования не задерживались в системе очень долго, несмотря на общую нагрузку, организуют дисциплину обслуживания, при которой чем дольше заявка находится в системе, тем ее приоритет становится выше.

Рассмотрим приоритетные функции, линейно зависящие от времени с крутизной нарастания, зависящей от номера класса, к которому принадлежит требование.

Предположим, что некоторое меченое требование поступает в момент *τ* и получает в момент *t* приоритет, определяемый значением приоритетной функции

Всякий раз, когда сервер готов к обслуживанию нового требования он выбирает из очереди требование с наивысшим мгновенным приоритетом- наибольшим значением приоритетной функции. Требования из класса с большим значением *p* наращивают приоритет с большей скоростью, чем требования из низшего приоритетного класса. На рисунке 3 показан пример того, как поступившее позже требование, но из высшего приоритетного класса, может получить обслуживание раньше, чем поступившее ранее требование из менее приоритетного класса. Это произойдет, если сервер освободится позже момента *t0* . При освобождении сервера до этого момента, обслуживание получит первое из поступивших требований.

Рис. 3 Взаимодействие между приоритетными функциями для СМО с приоритетами, зависящими от времени.

Исследуем эту систему при экспоненциальном распределении времени обслуживания.

Найдем среднее число требований , поступивших позже меченого , из классов с *p* ≥ *i* , которые будут обслужены раньше меченого. Очевидно, что для таких требований скорость нарастания приоритетной функции меньше скорости нарастания приоритетной функции меченого требования и , следовательно число таких требований равно нулю. Теперь определим число таких требований для классов с большей, чем у меченого скоростью нарастания приоритетной функции *p*< *i*. Из рассмотрения рисунка 3 можно видеть, что задержка меченого требования в системе для этого случая *Wp=t0-τ* связана с интервалом времени на котором поступают заявки, опережающие меченое требование *VI = τi - τ* соотношением

Отсюда получаем, что этот интервал равен

Следовательно, при интенсивности *λi*входящего потока для требований *i*-го класса находим среднее число «обгоняющих» требований

Рассмотрим теперь меченое требование из класса *p*, поступающее в момент *τ=0* и находящееся в очереди в течение времени *tp*.

Рис. 4 График приоритета qp(t), используемый для получения .

На рисунке 4 показано, что значение функции приоритета этого требования к моменту поступления на сервер будет равно *bptp*. При поступлении меченого требования оно застает в очереди *ni*требований из класса *i* . Одно из таких требований показанное на рисунке 4 поступило в момент *t=-t1*. Определим теперь среднее число требований из класса *i* , которые поступают до нулевого значения момента времени, находятся в нулевой момент еще в очереди и получают обслуживание раньше меченого требования. Из рисунка 4 можно видеть, что этому условию удовлетворяет требование из класса *i* , которое поступает в момент *-t1*и ожидает в очереди в течение времени

Из рассмотрения соотношений на рисунке видно, что

При *i > p*

Подставив вычисленные средние значения для «обгоняющих» требований получим систему линейных уравнений для величин задержки меченого требования

Производя преобразования, можно свести решение этой системы уравнений к рекурсивной форме

Полученная формула представляет собой главный результат анализа дисциплины обслуживания с приоритетами, зависящими от времени. Типичная характеристика СМО с проанализированной дисциплиной обслуживания приведена на рисунке 5 Штриховая линия показывает характеристику для системы без приоритетов. Видно, что действие закона сохранения проявляется здесь в том, что хотя одна часть заявок, получившая высший приоритет будет иметь меньшее время ожидания, чем в системе без приоритетов с обслуживанием в порядке поступления, другая часть заявок при этом обязательно будет задержана на большее, чем в бесприоритетной системе время.

Рис. 5 для СМО с относительными приоритетами, зависящими от времени (Р=5, λр=λ/5,).


##

## Оптимизация назначения приоритетов

Рассмотрим систему M/G/1 с интенсивностью поступающего пуассоновского потока *λ* требований в секунду и произвольной функцией плотности вероятности для времени обслуживания с заданным средним временем обслуживания. Пусть плата за требование *Y* является случайной величиной с произвольной функцией распределения .

Система функционирует следующим образом: новое требование, поступившее в систему «предлагает» неотрицательную плату *Y* «организатору очереди». После этого требованию предоставляется место в очереди такое, что все требования внесшие меньшую плату оказываются позади, большую впереди данного требования. В каждый момент времени сервер, завершив обслуживание очередного требования, принимает на обслуживание требование, оказавшееся впереди всей очереди. До полного завершения обслуживания требование не покидает сервер. Требования, внесшие одинаковую плату, обслуживаются в порядке поступления.

Найдем среднее время ожидания в очереди для требования, внесшего плату *Y=y*. Это время складывается из трех составляющих. Во-первых, это время на дообслуживание требования, находящегося в данный момент в сервере. Во-вторых – время обслуживания требований, которые поступили раньше и внесли большую или равную плату. Наконец меченому требованию придется ждать обслуживания всех требований поступивших позже его, но внесли большую плату. Среднее число требований, плата которых лежит в интервале (*u,u+du*) определяется по формуле Литтла :, где .

Используя обозначения для нижнего и верхнего предела функции β(u) можно записать суммарное выражение для времени ожидания в очереди для меченого требования в виде:

Используя ряд соотношений и обозначений можно найти, что при разрывной функции распределения вероятности это соотношение может быть приведено к виду

При абсолютно непрерывной функции плотности вероятности получим

.

Таким образом, мы получили конечное среднее время ожидания для всех требований, которые вносят плату выше, чем некоторое критическое значение

В пределе, когда размер платы стремится к бесконечности, среднее время ожидания стремится к *W0*. Нетрудно убедиться, что когда размер платы для всех требований одинаков

Это известный результат для СМО типа M/G/1 при обслуживании в порядке поступления, как и следовало ожидать, поскольку равная плата равносильна ее отсутствию с точки зрения распределения приоритетов.

При распределении приоритетов можно рассмотреть и другие стоимостные задачи. Определим оптимальное распределение платы за приоритеты в следующих предположениях. Пусть имеется зависимость стоимости от времени задержки в очереди для каждого требования, т.е. есть возможность определить, сколько стоит каждая секунда ожидания в очереди. Опишем эту зависимость с помощью случайного коэффициента нетерпения *α*.

Очевидно, что общие затраты клиента при обслуживании будут состоять из платы за место в очереди и потерь от времени ожидания.

Для требования с фиксированным коэффициентом нетерпения эти затраты равны

Пусть для всей совокупности клиентов можно определить функцию распределения вероятностей коэффициентов нетерпения

Сформулируем следующую задачу оптимизации: найти функцию *yα* , которая минимизирует среднюю стоимость *С* при условии ограничения всей средней платы некоторой заданной величиной *B*.

Определим

Преобразуя минимизируемый интеграл, получим

Из закона сохранения в непрерывной форме

следует, что решение задачи минимизации стоимости сводится к нахождению такой функции, при которой минимальна площадь под кривой произведения :

.

В то время как площадь под кривой, определяемой первым сомножителем должна оставаться постоянной.

Путем рассуждений о согласованности возрастания и убывания функций, входящих в произведение, можно сделать вывод, что решением задачи являются все функции, удовлетворяющие условию

Множество S такое, что.

Выберем самую простую строго возрастающую функцию – линейную. Таким образом, будем считать, что плата пропорциональна коэффициенту нетерпения.

.

Применяя ограничение средней платы

получим, что, если считать средний коэффициент нетерпения равный *A*

Это и есть функция оптимальной платы.

В качестве примера рассмотрим систему с показательным распределением платы

Время ожидания можно непосредственно вычислить:

Используя рассмотренное правило оптимальной платы можно найти распределение коэффициента нетерпения

Следовательно, средняя стоимость получается:

Описанная оптимизация является глобальной и позволяет найти функцию платы, которые минимизируют общую среднюю стоимость.