АНАЛИТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
Одноканальные СМО с неоднородным потоком заявок

Рассмотрим одноканальную СМО с неоднородным потоком заявок, в которую поступают H классов заявок, образующие простейшие потоки с интенсивностями λ_1,.., λ_H. Длительность τ_{b_k} обслуживания заявок класса k распределена по произвольному закону со средним значением b_k и коэффициентом вариации v_{b_k}. Выбор заявок из очереди на обслуживание осуществляется в соответствии с заданной дисциплиной обслуживания, в качестве которой будем рассматривать:

- дисциплину обслуживания бесприоритетную (ДО БП), при которой заявки выбираются на обслуживание в порядке поступления;
- дисциплину обслуживания заявок с относительными приоритетами (ДО ОП);
- дисциплину обслуживания заявок с абсолютными приоритетами (ДО АП).

В качестве основной характеристики, описывающей эффективность функционирования системы, будем рассматривать средние времена ожидания заявок разных классов, на основе которых легко могут быть

рассчитаны все остальные характеристики с использованием известных фундаментальных зависимостей (формул Литтла).

Основные предположения:
1) СМО содержит один обслуживающий прибор, который в каждый момент времени может обслуживать только одну заявку;
2) СМО имеет накопитель заявок неограниченной ёмкости, что означает отсутствие отказов поступающим заявкам при их постановке в очередь, то есть любая поступающая заявка всегда найдёт в накопителе место для ожидания независимо от того, сколько заявок уже находится в очереди;
3) заявки разных классов, поступающие в СМО независимо друг от друга, образуют простейшие потоки;
4) длительности обслуживания заявок каждого класса в приборе распределены по произвольному закону и не зависят друг от друга;
5) обслуживающий прибор не простаивает, если в системе (накопителе) имеется хотя бы одна заявка любого класса, причем после завершения обслуживания очередной заявки мгновенно из накопителя выбирается следующая заявка в соответствии с заданной дисциплиной обслуживания;
6) при использовании ДО БП заявки разных классов выбираются на обслуживание только в зависимости от времени поступления в систему по правилу «раньше пришел – раньше обслужен», независимо от номера класса, к которому принадлежит заявка;
7) при использовании приоритетных дисциплин (ДО ОП и ДО АП) приоритеты классам заявок назначены по принципу «класс с меньшим номером имеет более высокий приоритет», то есть наивысшим приоритетом обладают заявки класса 1;
8) в случае ДО АП заявка, обслуживание которой прервано более высокоприоритетной заявкой, возвращается в накопитель, где ожидает дальнейшего обслуживания, причем ее обслуживание продолжается с прерванного места.

Характеристики и свойства ДО БП

При бесприоритетной ДО средние времена ожидания одинаковы для всех классов заявок и определяются по следующей формуле:

$$w_k^{\text{БП}} = w^{\text{БП}} = \frac{\sum_{i=1}^{H} \lambda_i b_i^2 (1+\nu_{b_i}^2)}{2(1-R)} \qquad (k=1,\ldots,H),$$

где $R = \sum_{i=1}^{H} \rho_i = \sum_{i=1}^{H} \lambda_i b_i$ - суммарная загрузка системы.

Формула получена в предположении, что в системе существует стационарный режим и отсутствует перегрузка: R < 1.

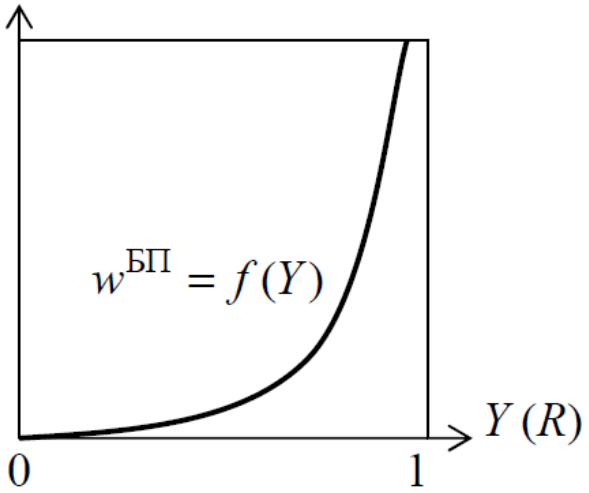
Основные выводы:

1. *Среднее время ожидания заявок разных классов* при использовании ДО БП *одинаково* при любых интенсивностях поступления $\lambda_1,..,\lambda_H$ и законах распределений $B_1(\tau),...,B_H(\tau)$ длительностей обслуживания заявок. Средние времена пребывания в системе заявок разных классов, в общем случае, различны, так как различны длительности обслуживания: $u_k^{\text{БП}} = w^{\text{БП}} + b_k \quad (k=1,\ldots,H)$.
2. *Среднее время ожидания заявок* в очереди минимально при постоянной (детерминированной) длительности обслуживания заявок каждого класса, когда коэффициент вариации длительности обслуживания = 0, и *увеличивается с ростом коэффициента вариации (дисперсии) длительности обслуживания*. Зависимость среднего времени ожидания от коэффициента вариации ν_{b_k} носит нелинейный характер. Так, например, при экспоненциально распределенной длительности обслуживания, когда $\nu_{b_k} = 1$, среднее время ожидания заявок увеличивается в 2 раза, а при $\nu_{b_k} = 2$ – в 5 раз, по сравнению с детерминированным обслуживанием.

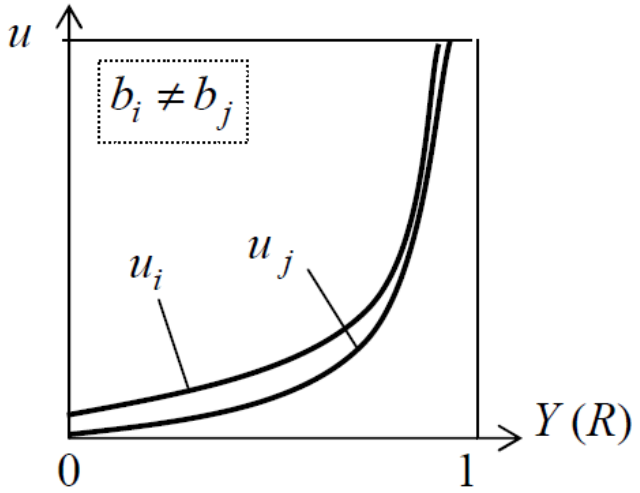
3. Среднее время ожидания заявок существенно зависит от суммарной нагрузки Y (загрузки R) системы. При $Y \geq 1$ ($R \to 1$) *время ожидания* заявок всех классов *возрастает неограниченно*: $w^{БП} \to \infty$, то есть заявки могут ожидать обслуживания сколь угодно долго. Увеличение суммарной нагрузки может быть обусловлено двумя факторами: увеличением интенсивностей поступления в систему заявок разных классов или увеличением длительности обслуживания заявок (например, за счет уменьшения скорости работы обслуживающего прибора).

   Зависимость среднего времени пребывания в системе заявок разных классов от суммарной нагрузки аналогична зависимости времени ожидания. Единственное отличие состоит в том, что *средние времена пребывания в системе заявок разных классов, в общем случае, различны*, что обусловлено различием длительностей обслуживания заявок разных классов.

а)



б)

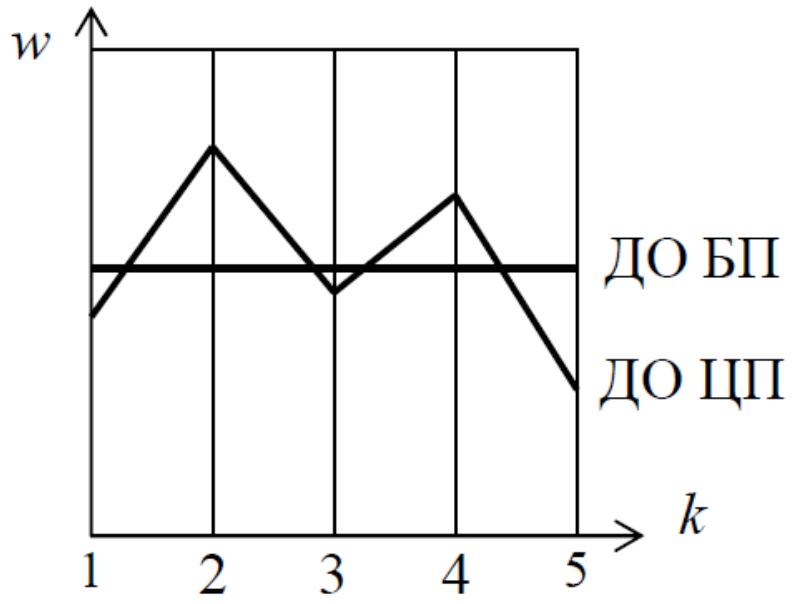


Аналогично, на графиках, отображающих зависимости средних длин очередей и числа заявок в системе от суммарной нагрузки, в общем случае, будут изображаться несколько кривых, соответствующих разным классам заявок. *Средние длины очередей заявок разных классов*, несмотря на одинаковое время ожидания, *в общем случае, различны* и, в соответствии с формулой Литтла, совпадают только в случае равенства интенсивностей поступления заявок разных классов в систему.

4. Можно показать, что *для бесприоритетной дисциплины обслуживания в обратном порядке (ООП)*, когда заявки на обслуживание выбираются по правилу «последний пришёл – первый обслужен», *средние времена ожидания заявок будут такими же*, как и при обслуживании в порядке поступления (ОПП), *но дисперсия* времени ожидания *будет больше*. Это обусловлено тем, что заявки, поступившие последними, будут ожидать незначительное время, в то время как заявки, попавшие в начало очереди, могут ожидать обслуживания достаточно долго, что обусловливает большой разброс значений времени ожидания.

5. Аналитическое исследование дисциплины обслуживания в циклическом порядке (ДО ЦП) достаточно сложно и связано с громоздкими математическими выкладками. Поэтому, не выписывая громоздких формул, отметим лишь наиболее характерные особенности, присущие этой ДО.

Для дисциплины обслуживания в циклическом порядке среднее время ожидания заявок разных классов в общем случае не одинаково. Это различие зависит от соотношения параметров потоков ($\lambda_1,..,\lambda_H$) и обслуживания ($B_1(\tau),...,B_H(\tau)$) заявок разных классов. В некоторых случаях ДО ЦП позволяет обеспечить меньшую суммарную длину очереди заявок, чем ДО БП. Зависимость среднего времени ожидания заявок каждого класса от суммарной нагрузки Y имеет такой же вид, как и для ДО БП.



Характеристики и свойства ДО ОП

Приоритеты называются ***относительными***, если они учитываются только в момент выбора заявки на обслуживание и не сказываются на работе системы в период обслуживания заявки любого класса (приоритета). Относительность приоритета связана со следующим. После завершения обслуживания какой-либо заявки из очереди на обслуживание выбирается заявка класса с наиболее высоким приоритетом, поступившая ранее других заявок этого класса (такого же приоритета). Если в процессе её обслуживания в систему поступят заявки с более высоким приоритетом, то обслуживание рассматриваемой заявки не будет прекращено, то есть эта заявка, захватив прибор, оказывается как бы более приоритетной. Таким образом, приоритет *относителен* в том смысле, что он имеет место лишь в момент выбора заявок на обслуживание и отсутствует, если прибор занят обслуживанием какой-либо заявки.

Введение относительных приоритетов (ОП) позволяет уменьшить по сравнению с ДО БП время ожидания высокоприоритетных заявок. При описании свойств для определённости будем полагать, что относительные приоритеты назначены по правилу: «более высокий приоритет – классу заявок с меньшим номером».

Для ДО ОП среднее время ожидания заявок класса k определяется по следующей формуле:

$$w_k^{\text{ОП}} = \frac{\sum_{i=1}^{H} \lambda_i b_i^{\,2} (1 + \nu_{b_i}^{\,2})}{2(1 - R_{k-1})(1 - R_k)} \qquad (k = 1, \ldots, H),$$

где R_{k-1} и R_k – суммарные загрузки, создаваемые заявками, которые имеют приоритет не ниже $(k-1)$ и k соответственно:

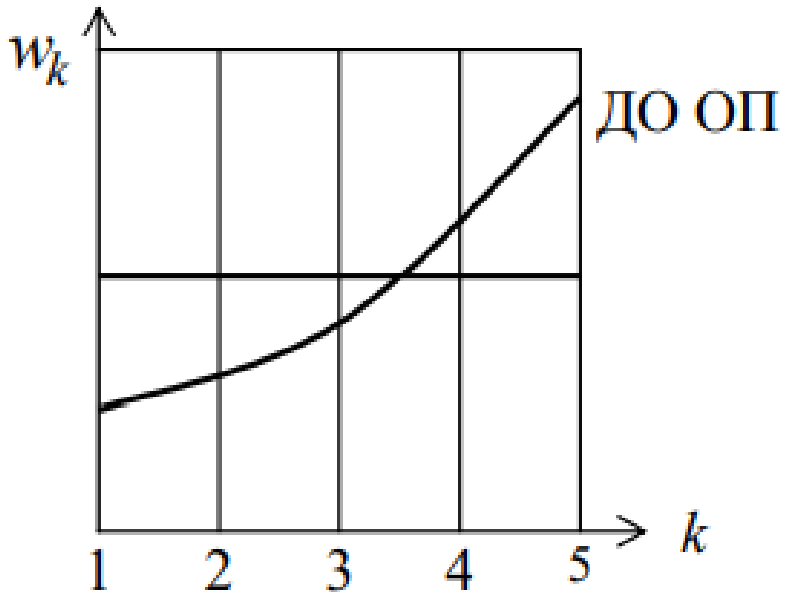
$$R_{k-1} = \sum_{i=1}^{k-1} \rho_i; \qquad R_k = \sum_{i=1}^{k} \rho_i .$$

Основные выводы:

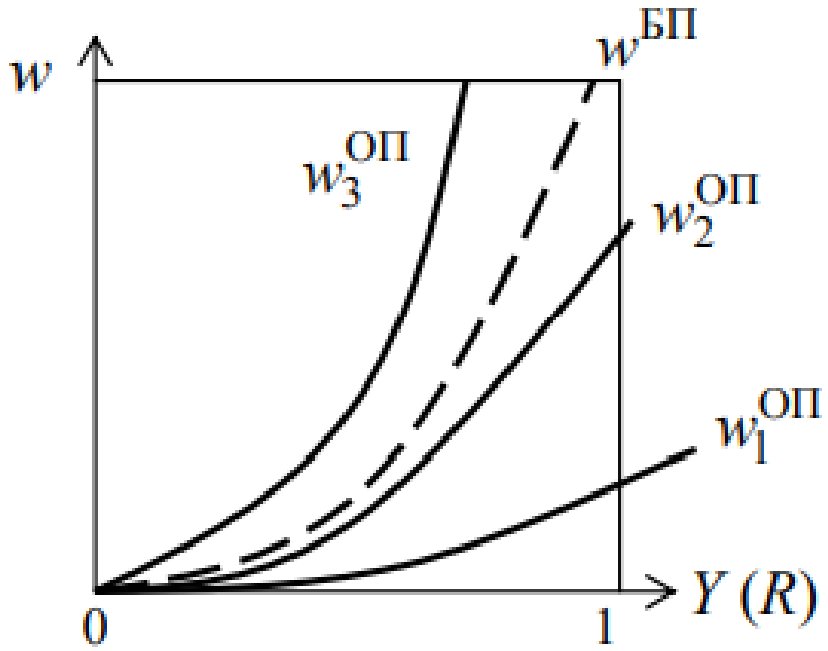
1. Введение относительных приоритетов по сравнению с ДО БП приводит к *уменьшению времени ожидания высокоприоритетных заявок* первого класса и к *увеличению времени ожидания низкоприоритетных заявок* класса H: $w_1^{\text{ОП}} < w_1^{\text{БП}}$ и $w_H^{\text{ОП}} > w_H^{\text{БП}}$.
2. При использовании ДО ОП *средние времена ожидания заявок монотонно увеличиваются с уменьшением приоритета* при любых интенсивностях поступления $\lambda_1,\ldots, \lambda_H$ и законах распределения $B_1(\tau),\ldots, B_H(\tau)$ длительностей обслуживания: $w_1^{\text{ОП}} < w_2^{\text{ОП}} < \ldots < w_H^{\text{ОП}}$.

(для средних времён пребывания заявок разных классов последнее соотношение, в общем случае, может и не выполняться).

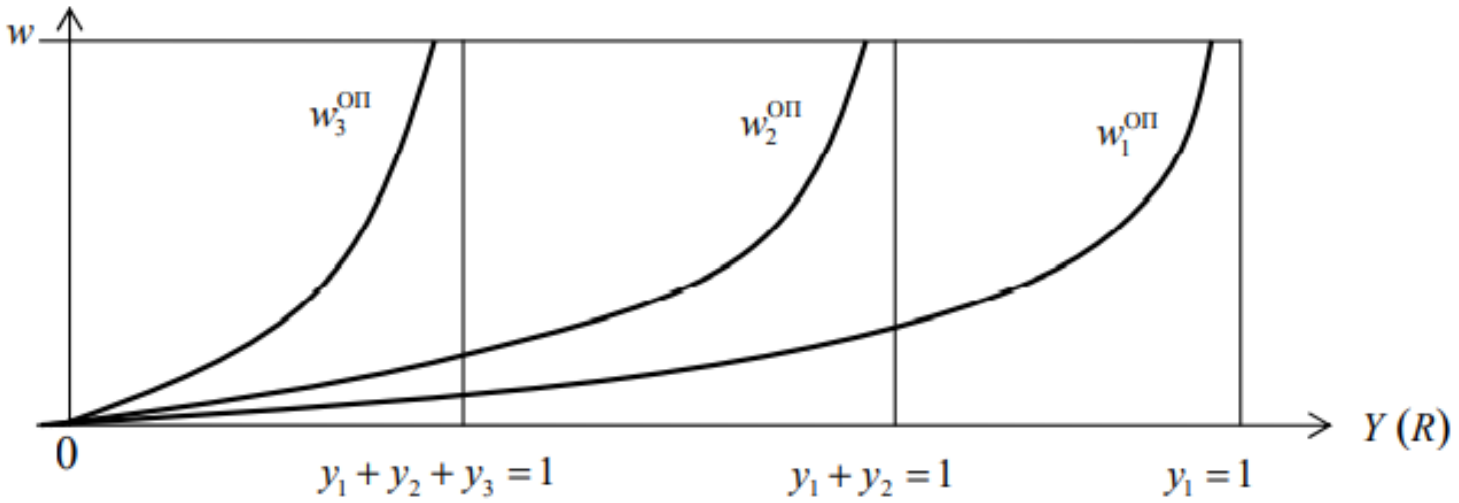
Свойства, сформулированные выше, иллюстрируются рисунке а, показывающим характер зависимости среднего времени ожидания заявок w_k от номера класса k при использовании ДО БП и ДО ОП.



3. На приведенном рисунке показаны зависимости среднего времени ожидания заявок разных классов от суммарной нагрузки Y системы при использовании ДО ОП. Здесь же для сравнения приведена аналогичная зависимость для ДО БП (штриховая линия). Характер зависимостей свидетельствует о том, что для ДО ОП при $Y \to 1$ резко увеличивается время ожидания заявок низкоприоритетных классов, в то время как для высокоприоритетных заявок это увеличение незначительно. Более того, для высокоприоритетных заявок обеспечивается достаточно хорошее качество обслуживания, то есть небольшое время ожидания даже при возникновении перегрузок, когда суммарная нагрузка становится больше единицы: $Y \geq 1$. Это свойство, называемое ***защитой от перегрузок***, обеспечивается за счет отказа в обслуживании низкоприоритетным
заявкам, время ожидания которых при этом резко возрастает. При ДО БП защита от перегрузок *отсутствует* для всех классов заявок



4. Рассмотрим свойство защиты от перегрузок при ДО ОП. На рисунке ниже построены зависимости среднего времени ожидания заявок трех классов при значительном росте нагрузки Y.



При достижении суммарной нагрузки, создаваемой заявками всех трех классов, значения (*y1 + y2 + y3 =1*) время ожидания заявок 3-го класса устремляется в бесконечность, что означает отказ в обслуживании, при этом заявки классов 1 и 2 продолжают обслуживаться и имеют конечное время ожидания. Дальнейшее увеличение нагрузки приводит к отказу в обслуживании заявок второго класса при *y1 + y2 =1* , то есть когда создаваемая заявками 1-го и 2-го классов нагрузка достигнет значения 1. Заявки первого класса получат отказ в обслуживании при *y1 =1* . Таким образом, в отличие от ДО БП при ДО ОП система полностью перестаёт обслуживать заявки, то есть функционировать, только в том случае, если нагрузка, создаваемая заявками самого высокоприоритетного (первого) класса, достигнет значения 1.

Характеристики и свойства ДО АП

Приоритет, прерывающий обслуживание низкоприоритетной заявки, называется абсолютным, а соответствующая дисциплина – дисциплиной обслуживания с абсолютными приоритетами (ДО АП). Прерванная заявка может быть потеряна или возвращена в накопитель, где она будет ожидать дальнейшего обслуживания. В последнем случае возможны два варианта продолжения обслуживания прерванной заявки:

- обслуживание с начала, то есть прерванная заявка будет обслуживаться заново с самого начала;
- дообслуживание, когда обслуживание прерванной заявки в приборе будет выполняться с прерванного места.

В дальнейшем, если не оговорено иное, будем предполагать дообслуживание прерванной заявки.

Для ДО АП среднее время ожидания заявок класса k определяется по следующей формуле:

$$w_k^{AП} = \frac{\sum_{i=1}^{K} \lambda_i b_i (1+\nu_{b_i}^2)}{2(1-R_{k-1})(1-R_k)} + \frac{R_{k-1} b_k}{1-R_{k-1}} \qquad (k=1,\ldots,H) \qquad (*)$$

где R_{k-1} и R_k – суммарные загрузки, создаваемые заявками, которые имеют приоритет не ниже $(k-1)$ и k соответственно:

$$R_{k-1} = \sum_{i=1}^{k-1} \rho_i; \qquad R_k = \sum_{i=1}^{k} \rho_i .$$

Основные выводы.

1. Выражение (*) содержит два слагаемых: $w_k^{АП} = s_k + z_k$, отображающих *среднее время ожидания начала обслуживания* s_k и *среднее время ожидания в прерванном состоянии* z_k соответственно:

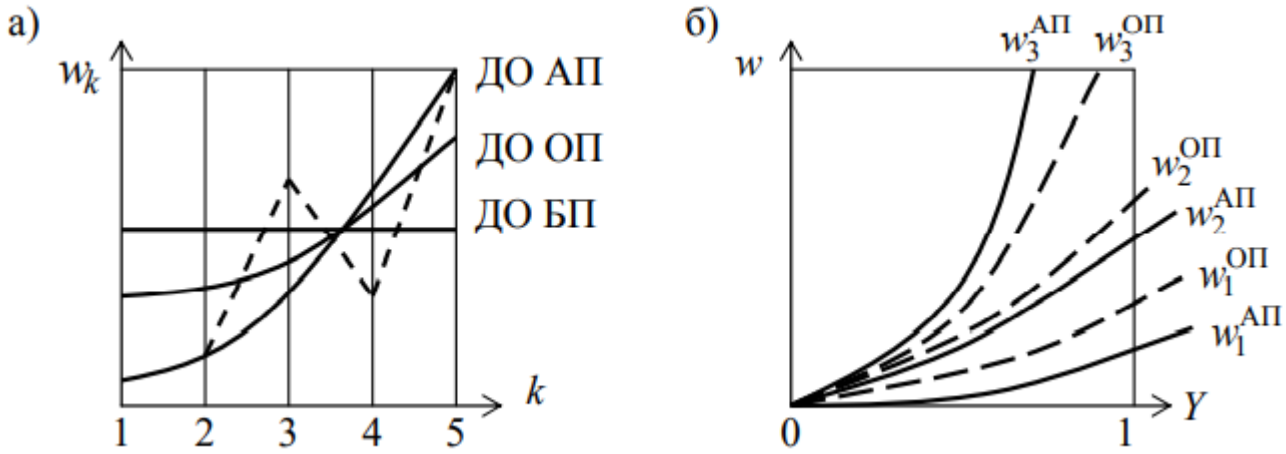
$$s_k = \frac{\sum_{i=1}^{k} \lambda_i b_i (1+\nu_{b_i}^2)}{2(1-R_{k-1})(1-R_k)}, \qquad z_k = \frac{R_{k-1} b_k}{1-R_{k-1}} \qquad (k = 1,\ldots,H).$$

2. Время ожидания заявок класса *k* зависит только от значений параметров классов *1*,…,*k* заявок, имеющих более высокий или такой же приоритет, и не зависит от параметров классов заявок *k +1,…, H* , имеющих более низкий приоритет.

3. Для заявок класса 1, имеющих самый высокий абсолютный приоритет, обеспечивается минимально возможное время ожидания по сравнению со всеми другими ДО, то есть при любой другой ДО среднее время ожидания заявок первого класса не может быть меньше, чем при ДО АП. Это объясняется тем, что в случае ДО АП заявки первого класса обслуживаются как бы в изоляции, независимо от заявок других классов.

4. Времена ожидания начала обслуживания s_k монотонно увеличиваются с уменьшением приоритета, однако время ожидания высокоприоритетной заявки в прерванном состоянии z_k может оказаться больше времени ожидания z_{k+1} заявки с более низким приоритетом, если длительности обслуживания связаны соотношением $b_k >> b_{k+1}$, так как количество прерываний заявками более высокого приоритета и, следовательно, время ожидания в прерванном состоянии прямо пропорционально зависит от длительности обслуживания заявок данного класса. Вследствие этого, полное время ожидания заявок высокоприоритетного класса, складывающееся из времени ожидания начала обслуживания и времени ожидания в прерванном состоянии, может оказаться больше, чем у заявок класса с низким приоритетом: $w_k^{АП} >> w_{k+1}^{АП}$.
5. Введение АП по сравнению с ОП приводит к уменьшению среднего времени ожидания самых высокоприоритетных заявок первого класса и к его увеличению для заявок класса H:. $w_1^{АП} < w_1^{ОП}$ и $w_H^{АП} > w_H^{ОП}$.

На рис. а построена зависимость среднего времени ожидания от номера класса. Для ДО АП пунктиром показан случай, когда $w_3^{АП} >> w_4^{АП}$, из чего следует, что $b_3 >> b_4$. Зависимость полного времени ожидания от суммарной нагрузки Y системы при использовании ДО АП аналогична зависимости для ДО ОП (рис. б) с тем лишь отличием, что при ДО АП высокоприоритетные заявки лучше защищены от перегрузок



Законы сохранения

Изменение ДО позволяет уменьшить время ожидания высокоприоритетных заявок за счет увеличения времени ожидания низкоприоритетных заявок. Очевидно, что за счет изменения ДО нельзя добиться того,

чтобы уменьшилось или увеличилось время ожидания заявок всех классов. Этот факт сформулирован в виде *закона сохранения времени ожидания.*

Для любой дисциплины обслуживания (ДО) $\sum_{i=1}^{H} \rho_i w_i = Const_{\text{ДО}}$, *или* $\sum_{i=1}^{H} \rho_i w_i = \frac{R\sum_{i=1}^{H} \lambda_i b_i^2 (1+\nu_{bi}^2)}{2(1-R)}$

то есть сумма произведений загрузок ρ_i *на среднее время ожидания* w_i $(i = 1 \ldots H)$ *заявок всех классов инвариантна относительно ДО.*

Закон сохранения времени ожидания выполняется при следующих условиях:

• система без потерь – все заявки на обслуживание удовлетворяются;

• система простаивает лишь в том случае, когда в ней нет заявок;

• при наличии прерываний длительность обслуживания прерванных заявок распределена по экспоненциальному закону;

• все поступающие потоки заявок – простейшие, и длительности обслуживания не зависят от интенсивностей потоков заявок.

Закон сохранения времени ожидания универсален и справедлив для всех ДО, удовлетворяющих указанным условиям. Его можно использовать для оценки достоверности приближенных результатов, полученных при исследовании сложных ДО и проведении имитационного моделирования, а также при решении задач синтеза.

Закон сохранения времени пребывания:

$$\sum_{i=1}^{H} \rho_i u_i = \frac{R \sum_{i=1}^{H} \lambda_i b_i^{\,2} (1 + \nu_{bi}^{\,2})}{2(1-R)} + \sum_{i=1}^{H} \rho_i b_i \,.$$

Когда средние длительности обслуживания заявок разных классов одинаковы, можно получить новую формулировку закона сохранения *в виде закона сохранения суммарной длины очереди заявок:*

$$\sum_{i=1}^{H} \lambda_i w_i = L = \underset{\text{ДО}}{Const} \,.$$

Пример. Дано: СМО с отказами и с двумя входящими потоками заявок: обычный с интенсивностью λ_1 и приоритетный с интенсивностью λ_2. Интенсивности обслуживания соответствующих заявок - μ_1 и μ_2. Система может находиться в трех состояниях: S_0 – свободна, S_1 - обработка обычной заявки, S_2 - обработка приоритетной заявки. Первоначально система находится в состоянии S_0. В случае поступления обычной заявки система переходит в состояние S_1. Если до завершения обслуживания обычной заявки поступила приоритетная, система прерывает обслуживание текущей заявки и приступает к обслуживанию приоритетной, т. е. переходит в состояние S_2 . После завершения обслуживания любой заявки система возвращается в исходное состояние S_0.