Динамические приоритеты в системах массового обслуживания

Научный доклад по статье В.В. Рыкова и Э.Е. Лемберг (1967)

Легиньких Галина Андреевна

Содержание

1 ВВЕДЕНИЕ И ОСНОВНЫЕ ПОНЯТИЯ

1.1 Что такое динамические приоритеты?

Динамические приоритеты — это система управления очередью, где правило выбора следующего требования на обслуживание зависит от текущего состояния всей системы.

Почему это важно изучать?

- Интуитивно кажется, что гибкие, адаптивные правила должны быть лучше жестких
- На практике часто пытаются внедрить “умные” алгоритмы
- Но математика показывает неожиданный результат!

Простая аналогия:

Представьте врача в приемном покое, который решает:

- **Не по приоритету:** “Всегда сначала реанимация, потом переломы, потом царапины”
- **По динамическому приоритету:** “Смотрю на всех пациентов. Если много реанимационных — беру их. Если много царапин, но они ждут уже 3 часа — беру их. Решаю в каждый момент заново”

Математически: Разбиваем пространство состояний (сколько кто ждет) на зоны и для каждой зоны свое правило.

1.2 Отличие от обычных (статических) приоритетов

Обычные (статические) приоритеты	Динамические приоритеты
Фиксированное правило	**Правило зависит от ситуации**
Пример: “Всегда сначала тип A”	Пример: “Если A > 3, то A; иначе если B ждет >1 час, то B”
Не смотрит на очередь	**Смотрит на всю картину**
Просто: “1 -> 2 -> 3”	Сложно: “Анализирую и решаю”
Легко реализовать	**Сложно реализовать**

1.3 Основная задача исследования

Вопрос: Дают ли динамические приоритеты преимущество по сравнению с обычными (статическими) приоритетами в смысле минимизации средних потерь?

Ответ в статье: НЕТ! Оптимальными оказываются обычные приоритеты с правильным упорядочением.

2 МАТЕМАТИЧЕСКАЯ МОДЕЛЬ

2.1 Параметры системы и их смысл

Рассматривается однолинейная система массового обслуживания с ожиданием, имеющая следующие характеристики:

1. **Один обслуживающий прибор** (однолинейная система)
2. N **входящих потоков** разнотипных требований
3. **Интенсивности потоков:** $\lambda_i > 0, i = 1, 2, \ldots, N$
4. **Нормировка:** $\sum_{i=1}^{N} \lambda_i = 1$ (условная вероятность принадлежности вновь поступившего требования к i-му типу)
5. **Время обслуживания** требования i-го типа имеет произвольную функцию распределения $B_i(x)$
6. **Штрафы:** $a_i > 0$ — потери за единицу времени ожидания требования i-го типа в очереди. Это **экономические потери** за один час ожидания требования типа i. Отражает важность/срочность требований

2.1.1 Формула 1: Интенсивности потоков λ_i

$$\lambda_i > 0, \quad \sum_{i=1}^{N} \lambda_i = 1$$

λ_i — это **среднее число требований типа i**, поступающих в систему за единицу времени. Пример: $\lambda_1 = 0.3$ означает, что в среднем каждую минуту поступает 0.3 требования первого типа

Нормировка $\sum \lambda_i = 1$ — это технический приём для упрощения вычислений. По сути, мы выбираем такую единицу времени, чтобы суммарная интенсивность равнялась 1.

Зачем нужна нормировка? Без нормировки формулы были бы более громоздкими. С нормировкой λ_i можно интерпретировать как **вероятность того, что вновь поступившее требование принадлежит к i-му типу**.

2.1.2 Формула 2: Функция распределения времени обслуживания $B_i(x)$

$$B_i(x) = P\{\text{время обслуживания требования типа } i \leq x\}$$

Модель работает для любых распределений: нормальное, равномерное, Эрланга и т.д.

Почему это важно? В реальных системах время обслуживания редко бывает экспоненциальным. Эта модель более реалистична!

2.2 Целевая функция: что мы минимизируем?

2.2.1 Формула 3: Основная целевая функция L

$$L = \sum_{i=1}^{N} a_i \lambda_i v_i$$

1. v_i — среднее время ожидания требования типа i в очереди.

   Пример: $v_1 = 0.5$ часа означает, что требования первого типа ждут в среднем 30 минут

2. $\lambda_i v_i$ — по формуле Литтла для систем массового обслуживания:

   $$\lambda_i v_i = \text{среднее число требований типа i в очереди}$$

   Доказательство: Если в среднем поступает λ_i требований в час и каждое ждет v_i часов, то в системе в среднем находится $\lambda_i v_i$ требований.

3. $a_i \lambda_i v_i$ — средние потери от ожидания требований типа i:

$$a_i \times (\lambda_i v_i) = \frac{\text{рубль}}{\text{час}} \times \text{требование} = \frac{\text{рубль}}{\text{час}}$$

4. **Суммирование** по всем типам даёт общие средние потери системы

2.2.2 Формула 4: Альтернативное представление через длины очередей

$$L = \sum_{i=1}^{N} a_i M q_i$$

q_i — случайная величина: число требований типа i в очереди в случайный момент времени

Mq_i — математическое ожидание длины очереди типа i

По формуле Литтла: $Mq_i = \lambda_i v_i$, поэтому формулы 3 и 4 эквивалентны

3 ФОРМАЛЬНОЕ ОПРЕДЕЛЕНИЕ ДИНАМИЧЕСКИХ ПРИОРИТЕТОВ

3.1 Пространство состояний

3.1.1 Формула 5: Вектор состояния x

$$x = (x_1, x_2, \dots, x_N), \quad x_i \geq 0 \text{ — целые числа}$$

x_i — **количество требований типа i**, находящихся в системе в данный момент

“В системе” = в очереди + на обслуживании (если обслуживается тип i)

3.2 Динамический приоритет как разбиение пространства

3.2.1 Формула 6: Разбиение на области $\mathcal{E}$

$$\mathcal{E} = \{E_1, E_2, \dots, E_N\}, \quad \bigcup_{i=1}^{N} E_i = X, \quad E_i \cap E_j = \emptyset \, (i \neq j)$$

X — всё пространство состояний

$\{E_i\}$ — разбиение на N непересекающихся областей

В каждой области E_i действует одно и то же правило: “выбирать требование типа i”

3.2.2 Формула 7: Функция управления $d(x)$

$$d(x) = i, \quad \text{если } x \in E_i$$

Это формальная запись правила: “посмотрели на состояние x, определили, к какой области оно принадлежит, применили соответствующее управление”

$d(x) = i$ означает: “в состоянии x выбрать для обслуживания требование типа i”

4 АНАЛИЗ ДЛЯ ДВУХ ПОТОКОВ ($N = 2$)

4.1 Вложенная цепь Маркова

Идея: Рассматривать систему не в произвольные моменты времени, а в специально выбранные — моменты окончания обслуживания.

Почему это удобно?

- В эти моменты принимаются решения о выборе следующего требования
- Процесс становится дискретным во времени
- Можно применить теорию марковских цепей

4.2 Рекуррентные соотношения для очередей

4.2.1 Формула 8: Рекуррентные уравнения

$$\begin{cases} q_i' = \xi_{ji} & \text{с вероятностью } \lambda_j, \text{ если } t \text{ — 0-момент} \\ q_i' = q_i + \xi_{ji} - \delta_{ji} & \text{если } t \text{ — } j\text{-момент} \end{cases}$$

t **и** t' — два последовательных момента окончания обслуживания

q_i **и** q_i' — длина очереди типа i в моменты t и t'

0-момент — момент, когда прибор был свободен - Начинаем обслуживать новое требование - Какое требование? С вероятностью λ_j — типа j - За время обслуживания приходят новые требования: ξ_{ji} типа i - Поэтому новая очередь: $q_i' = \xi_{ji}$

j-**момент** — момент, когда закончили обслуживать требование типа j - Была очередь q_i - За время обслуживания пришли ξ_{ji} требований типа i - Обслужили одно требование: если $j = i$, то убрали одно требование типа i (вычитаем δ_{ji}) - Поэтому: $q_i' = q_i + \xi_{ji} - \delta_{ji}$

ξ_{ji} — случайная величина: число требований типа i, пришедших за время обслуживания требования типа j - Если время обслуживания = t, то $\xi_{ji} \sim Poisson(\lambda_i t)$ - Условное распределение: $P(\xi_{ji} = k \mid \text{время} = t) = \frac{(\lambda_i t)^k}{k!} e^{-\lambda_i t}$

δ_{ji} — символ Кронекера:

$$\delta_{ji} = \begin{cases} 1, & \text{если } j = i \\ 0, & \text{если } j \neq i \end{cases}$$

- Математический способ сказать: "если обслужили требование типа i, то убрать одно требование этого типа из очереди"

4.3 Производящие функции

4.3.1 Формула 9: Условные производящие функции $G_i(z_1, z_2)$

$$G_i(z_1, z_2) = M(z_1^{q_1} z_2^{q_2} \mid E_i)$$

В нашем случае: $G_i(z_1, z_2)$ — это производящая функция СОВМЕСТНОГО распределения длин очередей ПРИ УСЛОВИИ, что система находится в области E_i.

4.4 Ключевое стационарное уравнение

4.4.1 Формула 10: Основное уравнение баланса

$$\pi_0 \left[1 - \lambda_1 b_1(u) - \lambda_2 b_2(u)\right] + \pi_1 G_1 \left(1 - \frac{b_1(u)}{z_1}\right) + \pi_2 G_2 \left(1 - \frac{b_2(u)}{z_2}\right) = 0$$

Левая часть состоит из трех слагаемых:

1. **Первое слагаемое:** $\pi_0[1 - \lambda_1 b_1(u) - \lambda_2 b_2(u)]$
   - π_0 — вероятность 0-момента (прибор свободен)
   - $1 - \lambda_1 b_1(u) - \lambda_2 b_2(u)$ — вклад от этого случая
2. **Второе слагаемое:** $\pi_1 G_1 \left(1 - \frac{b_1(u)}{z_1}\right)$
   - π_1 — вероятность 1-момента (выбрали тип 1)
   - G_1 — производящая функция при условии E_1
   - $1 - \frac{b_1(u)}{z_1}$ — оператор, соответствующий рекуррентному соотношению для случая "обслужили тип 1"
3. **Третье слагаемое:** аналогично для типа 2

Параметр u:

$$u = \lambda_1(1 - z_1) + \lambda_2(1 - z_2)$$

Функция $b_i(u)$:

$$b_i(u) = \int_0^{\infty} e^{-ux} dB_i(x)$$

Что такое $b_i(u)$? Это **преобразование Лапласа-Стилтьеса** функции распределения $B_i(x)$.

Физический смысл:

Если ξ_{ji} — число требований типа i, пришедших за время обслуживания требования типа j, то $M(z_i^{\xi_{ji}}) = b_j(\lambda_i(1 - z_i))$

В нашем случае два типа, поэтому: $M(z_1^{\xi_{j1}} z_2^{\xi_{j2}}) = b_j(\lambda_1(1-z_1)+\lambda_2(1-z_2)) = b_j(u)$

4.5 Вероятности состояний системы

4.5.1 Формула 11: Стационарные вероятности π_0, π_1, π_2

$$\pi_0 = 1 - R, \quad \pi_1 = \lambda_1 R, \quad \pi_2 = \lambda_2 R$$

где:

$$\rho_i = \lambda_i b_i, \quad R = \rho_1 + \rho_2$$

Вывод этих формул (почему они такие):

1. $\rho_i = \lambda_i b_i$ — это **коэффициент загрузки** прибора потоком i

λ_i требований типа i поступает в час.Каждое обслуживается b_i часов. Следовательно, прибор занят обслуживанием требований типа i в среднем $\lambda_i b_i$ часов в час. Это и есть доля времени, когда прибор занят обслуживанием типа i

2. $R = \rho_1 + \rho_2$ — **общая загрузка системы**

   По теореме Литтла: Для системы M/G/1:

   - Среднее число занятых приборов = R
   - Для однолинейной системы: либо 0, либо 1 прибор занят
   - Вероятность, что прибор занят = R
   - Вероятность, что прибор свободен = $1 - R$

3. $\pi_0 = 1 - R$ — вероятность 0-момента (прибор свободен)

4. $\pi_i = \lambda_i R$ — вероятность i-момента

4.6 Условные математические ожидания длин очередей

4.6.1 Формула 12: Обозначение g_{ij}

$$g_{ij} = M(q_j \mid E_i)$$

g_{ij} — **среднее число требований типа j в очереди**, при условии что система находится в области E_i

4.6.2 Формула 13: Система уравнений для g_{ij}

$$\begin{cases} (1-\rho_1)g_{11} - \rho_2 g_{21} = \lambda_1 C + (1-\rho_1) \\ -\rho_1 g_{12} + (1-\rho_2)g_{22} = \lambda_2 C + (1-\rho_2) \end{cases}$$

Как получается эта система?

1. **Берем производные** от Формулы 10 по z_1 и z_2
2. **Подставляем** $z_1 = z_2 = 1$ (это стандартный прием для получения моментов из производящих функций)
3. **Используем свойства:**
   - $\frac{\partial G_i}{\partial z_j}\Big|_{z_1=z_2=1} = g_{ij}$
   - $\frac{\partial b_k(u)}{\partial z_j}\Big|_{z_1=z_2=1} = -\lambda_j b_k$
   - $\frac{\partial^2 b_k(u)}{\partial z_j^2}\Big|_{z_1=z_2=1} = \lambda_j^2 b_k^{(2)}$
4. **После преобразований** получаем систему двух уравнений с четырьмя неизвестными: $g_{11}, g_{12}, g_{21}, g_{22}$

4.6.3 Формула 14: Константа C

$$C = \frac{b^{(2)}}{2R}, \quad b^{(2)} = \lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}$$

$b^{(2)}$ — средневзвешенный второй момент времени обслуживания:

$$b^{(2)} = \lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)}$$

где $b_i^{(2)} = \int_0^\infty x^2 dB_i(x)$ — второй момент для типа i

C — нормированная мера дисперсии времени обслуживания:

- Деление на $2R$ — результат математических преобразований
- C характеризует, насколько “размазано” время обслуживания

- Если все $b_i^{(2)} = b_i^2$ (дисперсия 0), то $C = \frac{R}{2}$
- Если есть дисперсия, C больше

4.7 Среднее время ожидания: ключевая формула

4.7.1 Формула 15: v_i через g_{ii}

$$v_i = R\frac{g_{ii} - 1}{\lambda_i}$$

Подробный вывод этой формулы:

1. **По определению:** v_i — среднее время ожидания требования типа i
2. **По формуле Литтла:** $\lambda_i v_i = Mq_i$ — среднее число требований типа i в очереди
3. **Разложим Mq_i по полной вероятности:**

$$Mq_i = \pi_0 M(q_i \mid E_0) + \pi_1 M(q_i \mid E_1) + \pi_2 M(q_i \mid E_2)$$

   Но $M(q_i \mid E_0) = 0$ (если прибор свободен, очереди нет)

4. **Уточнение:** $M(q_i \mid E_i)$ включает обслуживаемое требование! Поэтому:
   - Среднее число требований типа i В ОЧЕРЕДИ при условии E_i: $g_{ii} - 1$
   - Среднее число требований типа i В ОЧЕРЕДИ при условии E_j ($j \neq i$): g_{ji}
5. **Получаем:**

$$Mq_i = \pi_i(g_{ii} - 1) + \pi_j g_{ji}, \quad j \neq i$$

6. **Из системы уравнений** (Формула 13) можно выразить g_{ji} через g_{ii}
7. **После алгебраических преобразований** получаем удивительно простую формулу:

$$v_i = R\frac{g_{ii} - 1}{\lambda_i}$$

4.8 Фундаментальное ограничение для любых разбиений

4.8.1 Формула 16: Важное равенство

$$\rho_1 v_1 + \rho_2 v_2 = \frac{CR^2}{1-R}$$

Левая часть: $\rho_1 v_1 + \rho_2 v_2$

- $\rho_i = \lambda_i b_i$ — загрузка
- v_i — время ожидания
- Произведение $\rho_i v_i$ имеет смысл: “загрузка, умноженная на время ожидания”

Правая часть: $\frac{CR^2}{1-R}$

- $C = \frac{b^{(2)}}{2R}$ — нормированная дисперсия
- R — общая загрузка
- $(1-R)$ в знаменателе — характерно для систем с ожиданием

Это равенство выполняется для **ЛЮБОГО РАЗБИЕНИЯ** $\mathcal{E}$!

- Неважно, как мы разбиваем пространство состояний
- Неважно, насколько сложное правило придумали
- Сумма $\rho_1 v_1 + \rho_2 v_2$ всегда одна и та же!

Следствие: Мы не можем одновременно уменьшить v_1 и v_2. Уменьшение одного ведёт к увеличению другого.

4.9 Случай обычных приоритетов

4.9.1 Формула 17: Условие обычного приоритета

При приоритете типу 1: $g_{12} = 0$

Почему?

Если тип 1 имеет высший приоритет, и мы выбираем требование типа 1 (система в E_1), то в очереди не должно быть требований типа 2, которые ждут дольше. В пределе это даёт $g_{12} = 0$ (среднее число требований типа 2 равно 0)

Аналогично: при приоритете типу 2: $g_{21} = 0$

4.9.2 Формула 18: Решение при $g_{12} = 0$

$$g_{11} = \frac{\rho_1 C}{1 - \rho_1}$$

Вывод: Подставляем $g_{12} = 0$ в систему уравнений (Формула 13) и решаем.

4.9.3 Формула 19: Время ожидания при приоритете типу 1

$$v_1^{(1)} = \frac{CR}{1 - \rho_1}, \quad v_2^{(1)} = \frac{CR}{(1 - \rho_1)(1 - R)}$$

Вывод:

1. Из Формулы 18: $g_{11} = \frac{\rho_1 C}{1 - \rho_1}$
2. Подставляем в Формулу 15: $v_1^{(1)} = R\frac{g_{11} - 1}{\lambda_1} = \frac{CR}{1 - \rho_1}$
3. Из Формулы 16 находим $v_2^{(1)}$: $\rho_1 v_1^{(1)} + \rho_2 v_2^{(1)} = \frac{CR^2}{1 - R}$
4. Решаем относительно $v_2^{(1)}$

4.9.4 Формула 20: Время ожидания при приоритете типу 2

$$v_1^{(2)} = \frac{CR}{(1 - \rho_2)(1 - R)}, \quad v_2^{(2)} = \frac{CR}{1 - \rho_2}$$

(Вывод аналогичен)

4.10 Область достижимых значений

4.10.1 Формула 21: Границы для v_1 и v_2

$$v_1^{(1)} \leq v_1 \leq v_1^{(2)}, \quad v_2^{(2)} \leq v_2 \leq v_2^{(1)}$$

$v_1^{(1)}$ — минимально возможное v_1 (достигается при высшем приоритете типу 1)

$v_1^{(2)}$ — максимально возможное v_1 (достигается при низшем приоритете типу 1)

Аналогично для v_2

Геометрически: Точка (v_1, v_2) лежит в прямоугольнике на плоскости.

4.11 Минимизация потерь

4.11.1 Формула 22: Выражаем L через v_1

Из Формулы 16 выражаем v_2 через v_1:

$$v_2 = \frac{1}{\rho_2}\left(\frac{CR^2}{1-R} - \rho_1 v_1\right)$$

Подставляем в $L = a_1\lambda_1 v_1 + a_2\lambda_2 v_2$:

$$L = a_1\lambda_1 v_1 + a_2\lambda_2 \cdot \frac{1}{\rho_2}\left(\frac{CR^2}{1-R} - \rho_1 v_1\right)$$

Упрощаем:

$$L = \left(a_1\lambda_1 - a_2\lambda_2\frac{\rho_1}{\rho_2}\right) v_1 + \frac{a_2\lambda_2 CR^2}{\rho_2(1-R)}$$

4.11.2 Формула 23: Преобразование коэффициента

Учитываем, что $\rho_i = \lambda_i b_i$:

$$a_1\lambda_1 - a_2\lambda_2\frac{\rho_1}{\rho_2} = a_1\lambda_1 - a_2\lambda_2\frac{\lambda_1 b_1}{\lambda_2 b_2} = \lambda_1\left(a_1 - a_2\frac{b_1}{b_2}\right)$$

Домножим и разделим на b_1/b_2:

$$= \lambda_1\frac{b_1}{b_2}\left(\frac{a_1}{b_1} - \frac{a_2}{b_2}\right)$$

4.11.3 Формула 24: Оптимальное правило

Так как $\lambda_1\frac{b_1}{b_2} > 0$, то знак коэффициента при v_1 определяется знаком:

$$\frac{a_1}{b_1} - \frac{a_2}{b_2}$$

Оптимальное правило:

1. Если $\frac{a_1}{b_1} \geq \frac{a_2}{b_2}$, то коэффициент при $v_1 \geq 0$
   - Чтобы минимизировать L, нужно взять минимальное v_1
   - $v_1^{\text{min}} = v_1^{(1)}$ (приоритет типу 1)
2. Если $\frac{a_1}{b_1} \leq \frac{a_2}{b_2}$, то коэффициент при $v_1 \leq 0$
   - Чтобы минимизировать L, нужно взять максимальное v_1
   - $v_1^{\text{max}} = v_1^{(2)}$ (приоритет типу 2)

Итог: Сравниваем отношения $\frac{a_i}{b_i}$ и даем приоритет типу с большим отношением.

5 ОБОБЩЕНИЕ НА СЛУЧАЙ N > 2 ПОТОКОВ

5.1 Система уравнений

5.1.1 Формула 25: Общая система

Для любых k, l ($k < l$):

$$\lambda_l \sum_{i=1}^{N} \rho_i g_{ik} + \lambda_k \sum_{i=1}^{N} \rho_i g_{il} - \lambda_k g_{lk} - \lambda_l g_{kl} = \lambda_l \rho_k + \lambda_k \rho_l - \lambda_k \lambda_l \frac{b^{(2)}}{R}$$

Для каждого l:

$$\sum_{i=1}^{N} \rho_i g_{il} - g_{ll} = -(1 - \rho_l) - \lambda_l \frac{b^{(2)}}{2R}$$

Это обобщение Формулы 13 на случай N потоков

Первое уравнение — для пар (k, l)

Второе уравнение — для каждого l отдельно

Всего уравнений: $\frac{N(N+1)}{2}$

Неизвестных: N^2 (g_{ij} для всех i, j)

5.2 Множество достижимых значений

5.2.1 Формула 26: Определение V

$$V = \{v(\mathcal{E}) = (v_1(\mathcal{E}), \dots, v_N(\mathcal{E})) \mid \mathcal{E} \text{ — разбиение пространства состояний}\}$$

где

$$v_i(\mathcal{E}) = R\frac{g_{ii} - 1}{\lambda_i}$$

Свойства множества V:

1. $V \subset \mathbb{R}^N$
2. V — выпуклое множество (доказывается)
3. Крайние точки V соответствуют обычным приоритетам

5.3 Теоремы и леммы

5.3.1 Лемма (выпуклость): V — выпуклое множество

Доказательство: Пусть $v', v'' \in V$. Рассмотрим $v = \gamma v' + (1-\gamma)v''$, $0 \leq \gamma \leq 1$. Так как система уравнений линейна относительно g_{ii}, а v_i линейно зависит от g_{ii}, то $v \in V$.

5.3.2 Теорема (главный результат): В классе систем с динамическими приоритетами оптимальной является система с обыкновенными приоритетами

5.3.3 Алгоритм определения оптимальных приоритетов

1. Для каждого типа i вычислить:
$$r_i = \frac{a_i}{b_i}$$
2. Упорядочить типы по убыванию r_i:
$$r_{i_1} \geq r_{i_2} \geq \cdots \geq r_{i_N}$$
3. Назначить приоритеты: тип i_1 — высший, ..., тип i_N — низший

6 ПРАКТИЧЕСКИЙ ПРИМЕР: ОПТИМИЗАЦИЯ CALL-ЦЕНТРА

6.1 Постановка задачи

Компания имеет call-центр с одним оператором и тремя типами звонков:

Тип звонка	Описание	λ_i (зв./час)	b_i (мин)	a_i (руб/мин)
1	VIP-клиенты	2	5	100
2	Постоянные клиенты	4	8	40
3	Новые клиенты	6	10	10

Задача: Определить оптимальную стратегию ответов на звонки для минимизации потерь.

6.2 Подготовка данных

Приводим к одним единицам измерения:

- $\lambda_1 = 2$ зв/час = 2/60 зв/мин = 0.0333 зв/мин
- $\lambda_2 = 4$ зв/час = 0.0667 зв/мин

- $\lambda_3 = 6$ зв/час = 0.1000 зв/мин
- $b_1 = 5$ мин
- $b_2 = 8$ мин
- $b_3 = 10$ мин
- $a_1 = 100$ руб/мин
- $a_2 = 40$ руб/мин
- $a_3 = 10$ руб/мин

Проверяем нормировку:

$$\sum \lambda_i = 0.0333 + 0.0667 + 0.1000 = 0.2000 \text{ зв/мин}$$

Нужно нормировать, чтобы сумма была 1. Умножаем все λ_i на 5:

- $\lambda_1 = 0.1667$
- $\lambda_2 = 0.3333$
- $\lambda_3 = 0.5000$
- $\sum \lambda_i = 1.0000$

6.3 Проверка устойчивости системы

Загрузки:

$$\rho_1 = \lambda_1 b_1 = 0.1667 \times 5 = 0.8335$$

$$\rho_2 = \lambda_2 b_2 = 0.3333 \times 8 = 2.6664$$

$$\rho_3 = \lambda_3 b_3 = 0.5000 \times 10 = 5.0000$$

Общая загрузка:

$$R = \rho_1 + \rho_2 + \rho_3 = 0.8335 + 2.6664 + 5.0000 = 8.4999 > 1$$

Проблема: Система перегружена! $R = 8.5 > 1$ означает, что оператор не успевает обслуживать звонки.

Решение: Увеличим время работы или наймем больше операторов. Для примера предположим, что мы увеличили количество операторов или уменьшили интенсивность.

Скорректированные данные (уменьшаем интенсивность в 10 раз):

- $\lambda_1 = 0.01667$
- $\lambda_2 = 0.03333$
- $\lambda_3 = 0.05000$
- $\sum \lambda_i = 0.1$ (нормировка не соблюдается, но для расчета устойчивости это не важно)

Тогда:

$$\rho_1 = 0.01667 \times 5 = 0.08335$$

$$\rho_2 = 0.03333 \times 8 = 0.26664$$

$$\rho_3 = 0.05000 \times 10 = 0.50000$$

$$R = 0.08335 + 0.26664 + 0.50000 = 0.84999 < 1$$

6.4 Определение оптимальных приоритетов

Вычисляем отношения a_i/b_i:

1. VIP-клиенты: $\frac{100}{5} = 20$
2. Постоянные клиенты: $\frac{40}{8} = 5$
3. Новые клиенты: $\frac{10}{10} = 1$

Упорядочиваем:

$$20 > 5 > 1$$

Оптимальный порядок приоритетов:

1. **Высший:** VIP-клиенты

2. **Средний:** Постоянные клиенты
3. **Низший:** Новые клиенты

6.5 Расчет времени ожидания

Предположения: Время обслуживания имеет экспоненциальное распределение, тогда:

$$b_i^{(2)} = 2b_i^2$$

Вычисляем: - $b_1^{(2)} = 2 \times 5^2 = 50$

- $b_2^{(2)} = 2 \times 8^2 = 128$
- $b_3^{(2)} = 2 \times 10^2 = 200$

Средневзвешенный второй момент:

$$b^{(2)} = \lambda_1 b_1^{(2)} + \lambda_2 b_2^{(2)} + \lambda_3 b_3^{(2)}$$

$$= 0.01667 \times 50 + 0.03333 \times 128 + 0.05000 \times 200$$

$$= 0.8335 + 4.2662 + 10.0000 = 15.0997$$

Константа C:

$$C = \frac{b^{(2)}}{2R} = \frac{15.0997}{2 \times 0.85} = \frac{15.0997}{1.7} = 8.8822$$

6.5.1 Формула 27: Время ожидания при обычных приоритетах (общая формула)

Для типа с приоритетом k:

$$v_{i_k} = \frac{CR}{(1 - \sum_{j=1}^{k-1} \rho_{i_j})(1 - \sum_{j=1}^{k} \rho_{i_j})}$$

Применяем для нашего случая:

1. **VIP-клиенты (k=1):**

$$v_1 = \frac{8.8822 \times 0.85}{(1-0)(1-0.08335)} = \frac{7.5499}{0.91665} = 8.234 \text{ мин}$$

2. **Постоянные клиенты (k=2):**

$$v_2 = \frac{8.8822 \times 0.85}{(1-0.08335)(1-0.08335-0.26664)} = \frac{7.5499}{0.91665 \times 0.65001}$$

$$= \frac{7.5499}{0.5958} = 12.672 \text{ мин}$$

3. **Новые клиенты (k=3):**

$$v_3 = \frac{8.8822 \times 0.85}{(1-0.08335-0.26664)(1-0.85)} = \frac{7.5499}{0.65001 \times 0.15}$$

$$= \frac{7.5499}{0.0975} = 77.435 \text{ мин} \approx 1 \text{ час } 17 \text{ мин}$$

6.6 Экономические потери

Используем исходные интенсивности (до нормировки):

- $\lambda_1 = 2$ зв/час = 0.0333 зв/мин
- $\lambda_2 = 4$ зв/час = 0.0667 зв/мин
- $\lambda_3 = 6$ зв/час = 0.1000 зв/мин

Потери в минуту:

$$L = \sum a_i \lambda_i v_i$$

$$= 100 \times 0.0333 \times 8.234 + 40 \times 0.0667 \times 12.672 + 10 \times 0.1000 \times 77.435$$

$$= 27.42 + 33.80 + 77.44 = 138.66 \text{ руб/мин}$$

Потери в час: $138.66 \times 60 = 8319.6$ руб/час

6.7 Сравнение с другими стратегиями

Стратегия 1: FIFO (первый пришел — первый обслужен)

Приближенная формула для M/G/1:

$$v_{\text{FIFO}} = \frac{CR}{1-R} = \frac{8.8822 \times 0.85}{1-0.85} = \frac{7.5499}{0.15} = 50.333 \text{ мин}$$

Все типы ждут одинаково: 50.333 мин

Потери:

$$L_{\text{FIFO}} = (100 \times 0.0333 + 40 \times 0.0667 + 10 \times 0.1000) \times 50.333$$

$$= (3.33 + 2.668 + 1.000) \times 50.333 = 6.998 \times 50.333 = 352.2 \text{ руб/мин}$$

В час: $352.2 \times 60 = 21132$ руб/час

Экономия от оптимальной стратегии: $21132 - 8319.6 = 12812.4$ руб/час (60.6% экономии!)

Стратегия 2: Обратные приоритеты (новые -> постоянные -> VIP)

Пересчитываем аналогично, получаем: $v_3 = 8.234$ мин, $v_2 = 12.672$ мин, $v_1 = 77.435$ мин

Потери:

$$L = 10 \times 0.1000 \times 8.234 + 40 \times 0.0667 \times 12.672 + 100 \times 0.0333 \times 77.435$$

$$= 8.234 + 33.80 + 257.86 = 299.89 \text{ руб/мин}$$

В час: $299.89 \times 60 = 17993.4$ руб/час

Экономия от оптимальной стратегии: $17993.4 - 8319.6 = 9673.8$ руб/час (53.8% экономии)

6.8 Выводы для call-центра

1. **Оптимальная стратегия:** Сначала VIP, затем постоянные, потом новые клиенты
2. **Ожидаемое время ответа:**
    - VIP: 8.2 минуты
    - Постоянные: 12.7 минуты
    - Новые: 77.4 минуты (нужно предупреждать!)
3. **Экономический эффект:** 60.6% экономии по сравнению с FIFO
4. **Рекомендации:**
    - Внедрить систему автоматического определения типа клиента
    - Для новых клиентов предложить callback-систему
    - VIP-клиентам гарантировать ответ в течение 10 минут

7 Рассмотрим банк с двумя типами клиентов:

Данные: - Тип 1 (VIP): $\lambda_1 = 0.2, b_1 = 0.3$ часа, $a_1 = 5000$ руб/час

- Тип 2 (обычные): $\lambda_2 = 0.3, b_2 = 0.4$ часа, $a_2 = 1000$ руб/час

Вычисления:

1. $\rho_1 = 0.2 \times 0.3 = 0.06$
2. $\rho_2 = 0.3 \times 0.4 = 0.12$
3. $R = 0.06 + 0.12 = 0.18$
4. Предположим $C = 0.05$ (вычисляется из вторых моментов)

Применяем Формулу 23 при $v_1 = 0.5$ часа:

$$L = 5000 \times 0.2 \times 0.5 + 1000 \times 0.3 \times \frac{1}{0.12} \left(\frac{0.05 \times 0.18^2}{1 - 0.18} - 0.06 \times 0.5 \right)$$

Вычисляем по частям:

1. Первое слагаемое: $5000 \times 0.2 \times 0.5 = 500$
2. В скобках: $\frac{0.05 \times 0.0324}{0.82} - 0.03 = \frac{0.00162}{0.82} - 0.03 = 0.001976 - 0.03 = -0.028024$
3. Второе слагаемое: $1000 \times 0.3 \times 8.3333 \times (-0.028024) = 300 \times 8.3333 \times (-0.028024) = 2500 \times (-0.028024) = -70.06$

Итого: $L = 500 - 70.06 = 429.94$ руб/час

Анализ: Формула 23 показывает, что при данных параметрах потери уменьшаются при увеличении v_1 (второе слагаемое отрицательное). Следовательно, для минимизации L нужно взять максимальное v_1, то есть $v_1^{(2)}$ — приоритет типу 2.