Теория массового обслуживания

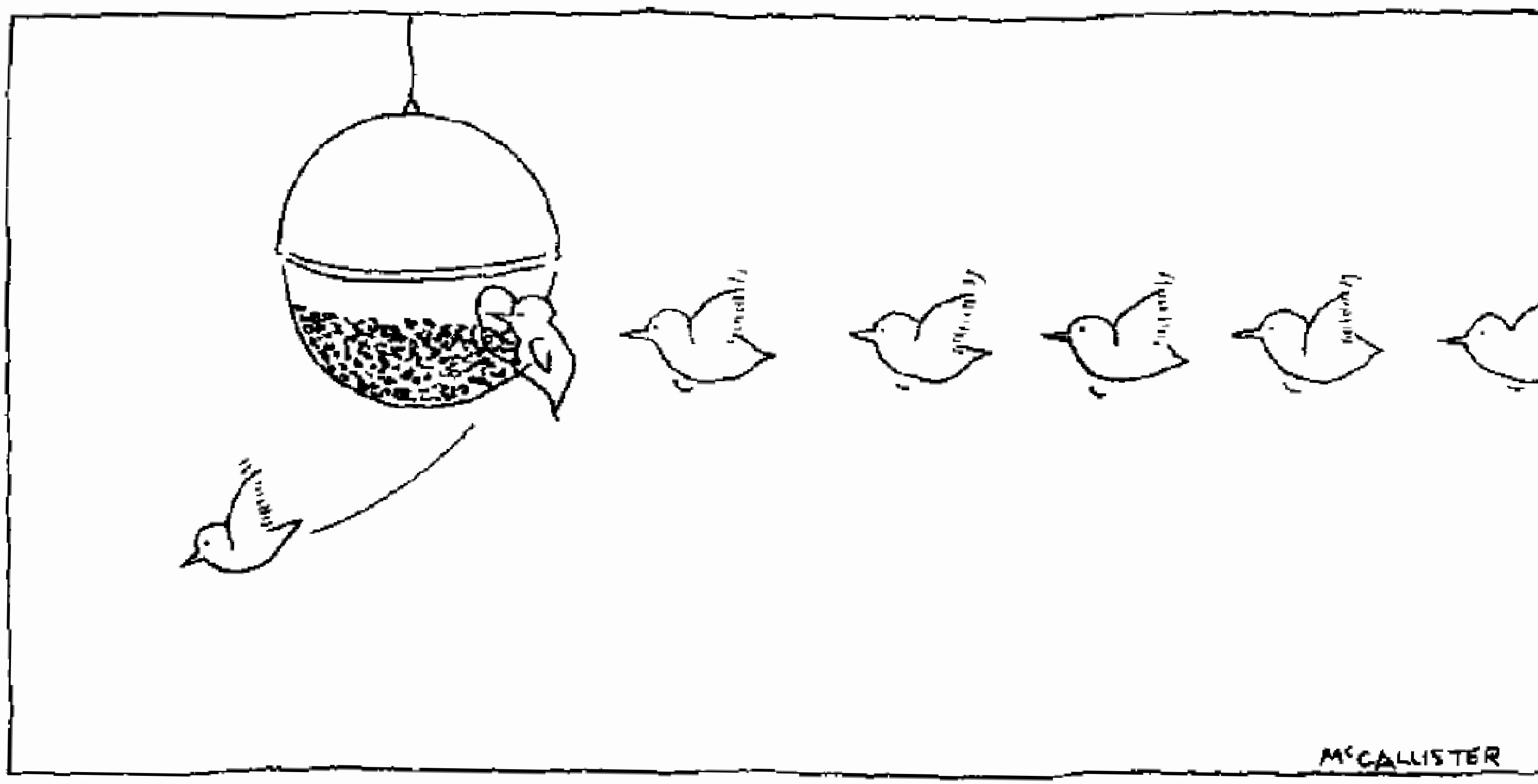
Лекция 1

Системы массового обслуживания
и их характеристики
+
Геометрическая интерпретация
математического ожидания

лектор: Фурманов Кирилл Константинович

О чём будет речь?

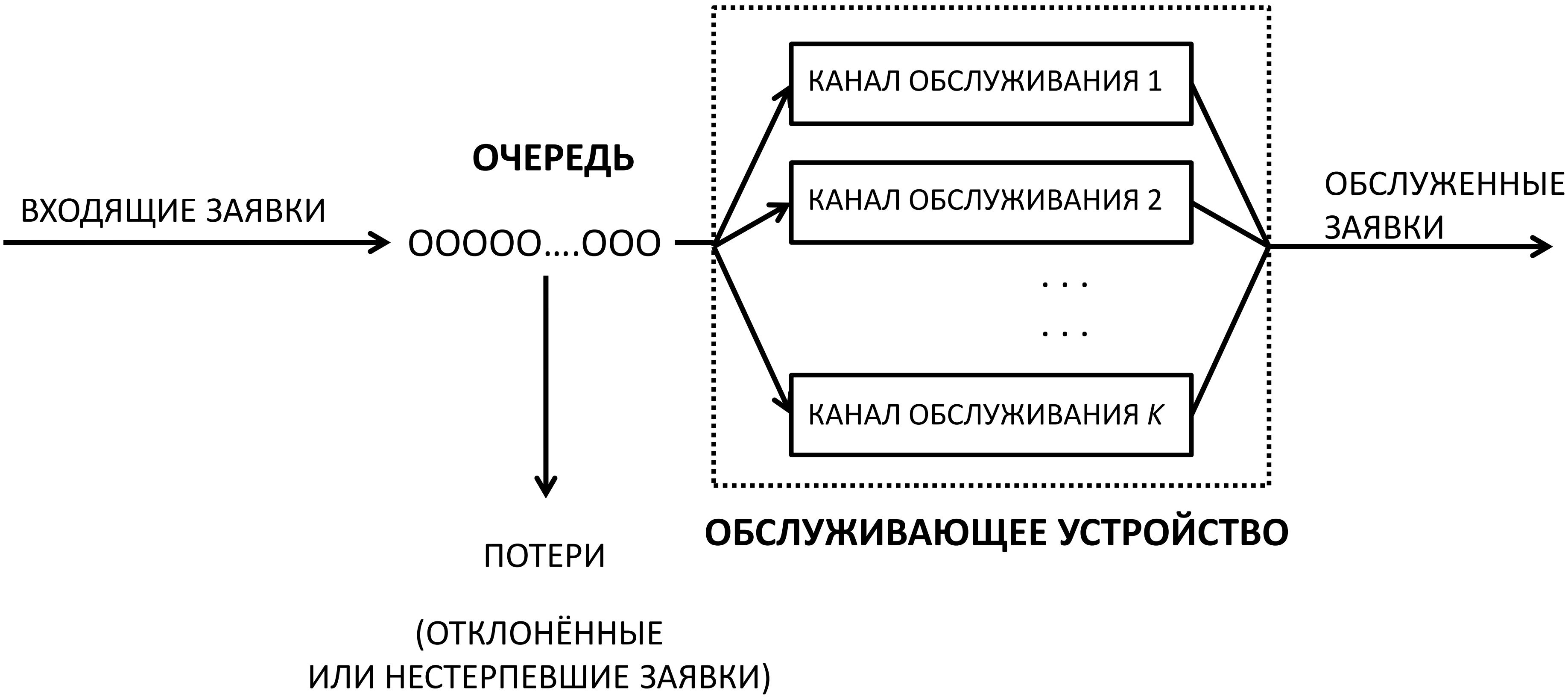
О системах массового обслуживания (СМО) и об очередях.
СМО выглядят как-то так:



Drawing by McCallister; © 1977 *The New Yorker* Magazine, Inc.

Картинка из книги R.B. Cooper “Introduction to Queueing Theory”

Система массового обслуживания на схеме



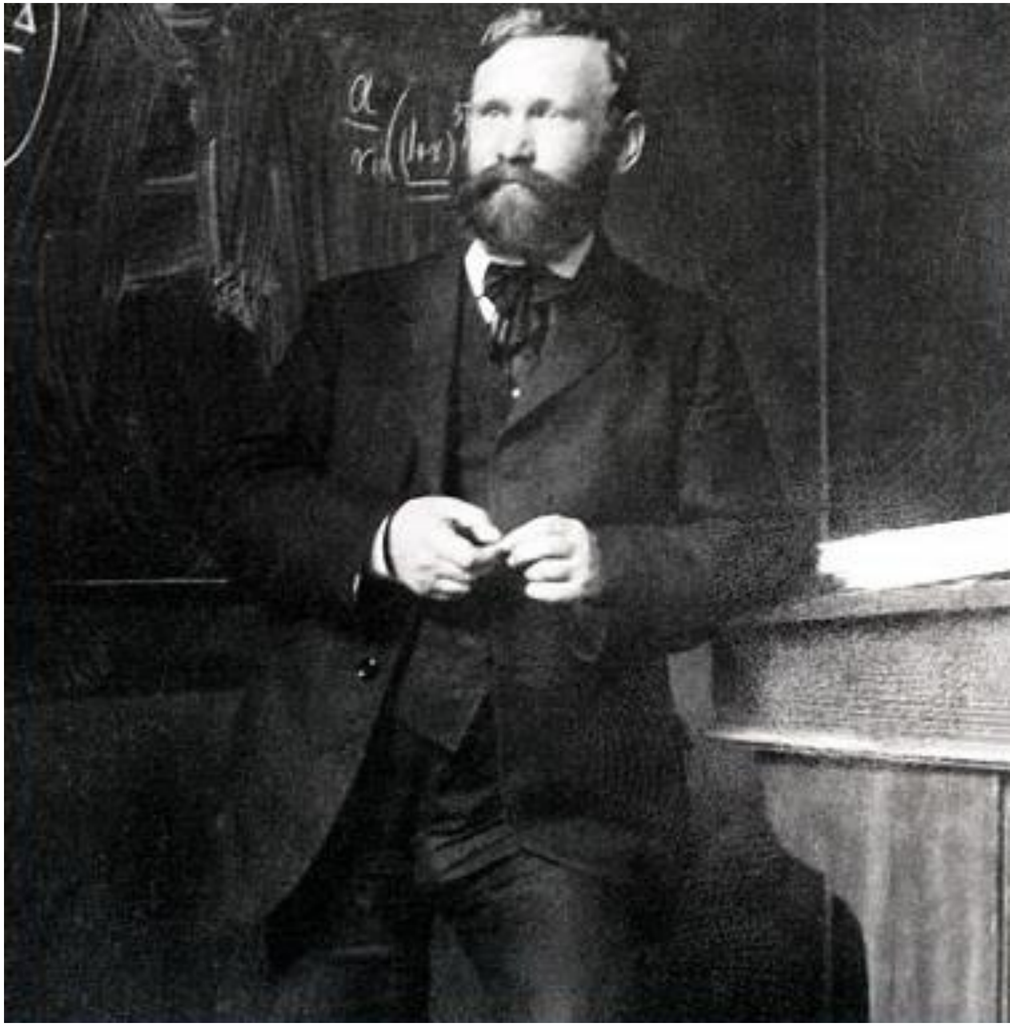
“Заявки” – очень условный термин. За ним может скрываться что угодно.

Примеры СМО

- телекоммуникационные системы (оригинальное приложение);
- колл-центр;
- сетевой сервер;
- больница;
- гостиница;
- аэропорт;
- многие другие …

Агнер Краруп Эрланг →

Что это за импозантный мужчина?



Основные характеристики процессов обслуживания

(i) Характеристики входящего потока заявок (интенсивность и не только).
(ii) Распределение времени обслуживания.
(iii) Число каналов обслуживания.
(iv) Число стадий обслуживания.
(v) Ёмкость системы.
(vi) Дисциплина обслуживания.

Случайные величины, представляющие особый интерес

Число заявок в системе (в момент t) = число в очереди + число на обслуживании;
Время пребывания заявки i в системе = время ожидания в очереди + время обслуживания;
Длительность цикла занятости = занятое время + время простоя;
(ещё …)

Характеристики эффективности СМО

- Средняя длина очереди;
- Среднее время ожидания;
- Вероятность потерь;
- Коэффициент загрузки мощностей;
- Пропускная способность;
- ещё …

Что эти слова означают?

Пример

После изготовления изделия обследуются службой контроля качества.
В пункт контроля изделия поступают с интервалом в десять минут: $t = 0, 10, 20, \ldots$
Для первых пяти изделий известно время завершения обследования:

12	18	32	40	46

Перед поступлением изделия в момент $t = 0$ пункт контроля свободен.
По данным о первых пятидесяти минутах работы рассчитайте:
(а) долю времени простоя; (б) среднее число изделий на пункте контроля.

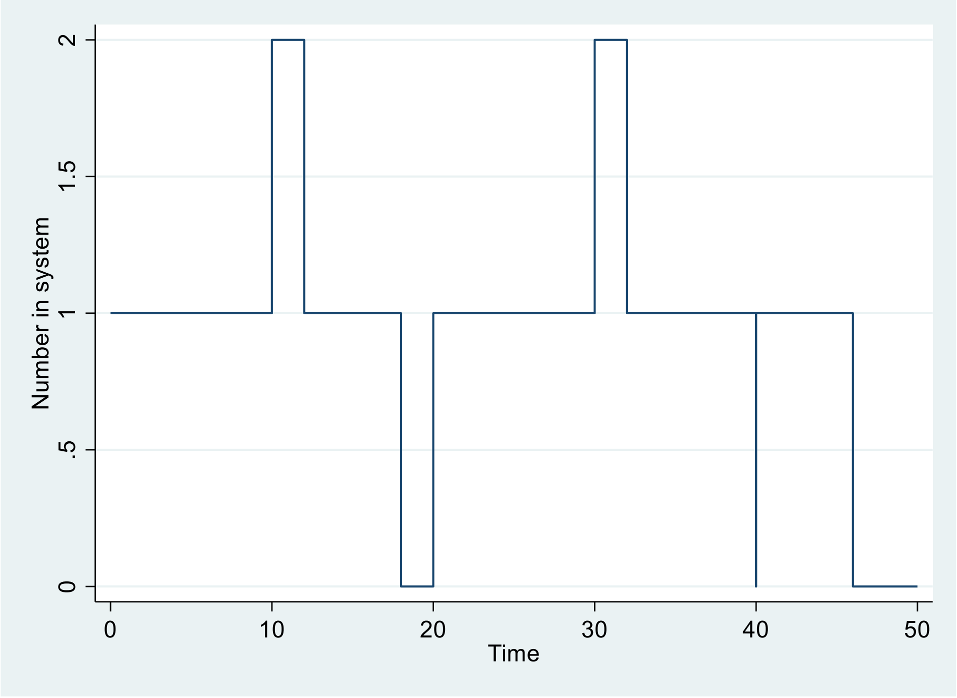
Пример

После изготовления изделия обследуются службой контроля качества.
В пункт контроля изделия поступают с интервалом в десять минут: $t = 0, 10, 20, \ldots$
Для первых пяти изделий известно время завершения обследования:

12 18 32 40 46

Перед поступлением изделия в момент $t = 0$ пункт контроля свободен.
По данным о первых пятидесяти минутах работы рассчитайте:
(а) долю времени простоя; (б) среднее число изделий на пункте контроля.

Решение. (а)



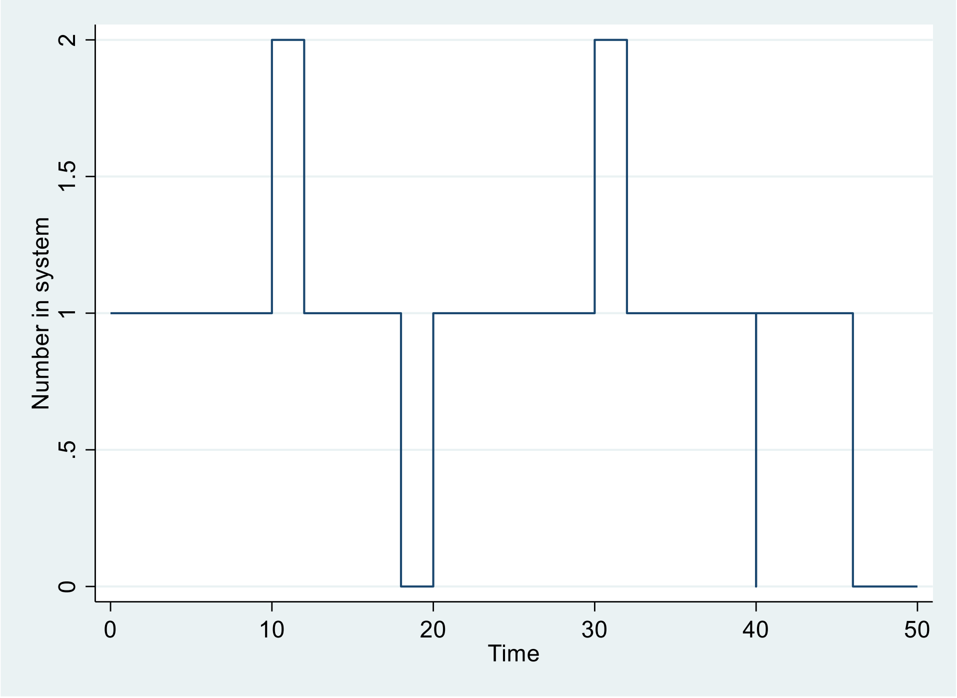
Пример

После изготовления изделия обследуются службой контроля качества.
В пункт контроля изделия поступают с интервалом в десять минут: $t = 0, 10, 20, \ldots$
Для первых пяти изделий известно время завершения обследования:

12 18 32 40 46

Перед поступлением изделия в момент $t = 0$ пункт контроля свободен.
По данным о первых пятидесяти минутах работы рассчитайте:
(а) долю времени простоя; (б) среднее число изделий на пункте контроля.

Решение. (а) пункт контроля простаивает с 18 по 20 минуту и с 46 по 50 – всего 6 минут.



Доля времени простоя: $\frac{6}{50} = 12\%$.

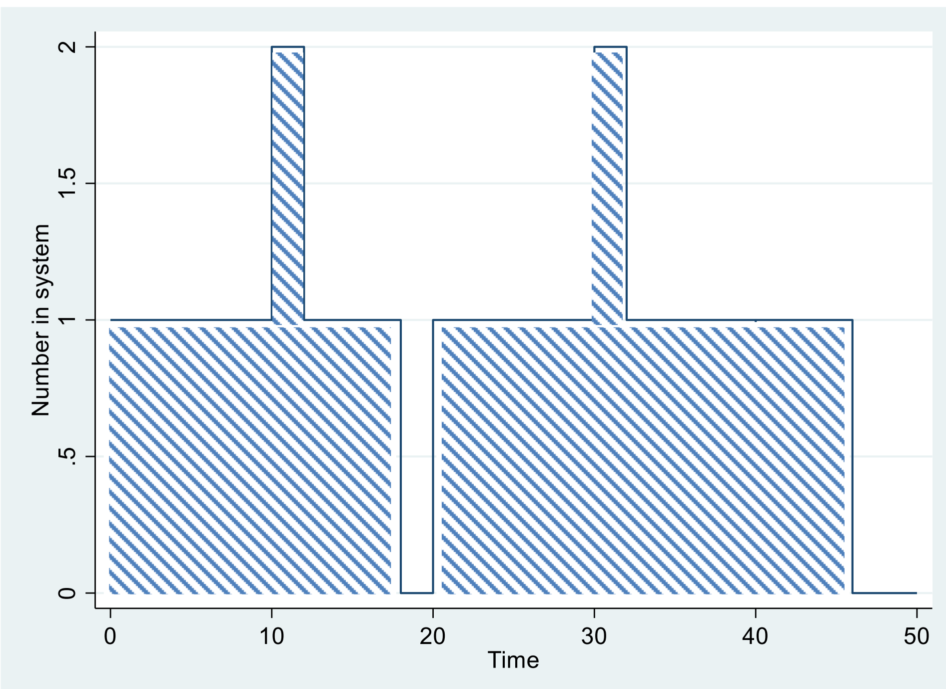
Пример

После изготовления изделия обследуются службой контроля качества.
В пункт контроля изделия поступают с интервалом в десять минут: $t = 0, 10, 20, \dots$
Для первых пяти изделий известно время завершения обследования:

12	18	32	40	46

Перед поступлением изделия в момент $t = 0$ пункт контроля свободен.
По данным о первых пятидесяти минутах работы рассчитайте:
(а) долю времени простоя; (б) среднее число изделий на пункте контроля.

Решение. (б) Площадь заштрихованной фигуры – суммарное время, которое все изделия пробыли на пункте контроля.



6 мин из 50 – 0 изделий на пункте;
40 мин из 50 – 1 изделие;
4 мин из 50 – 2 изделия.

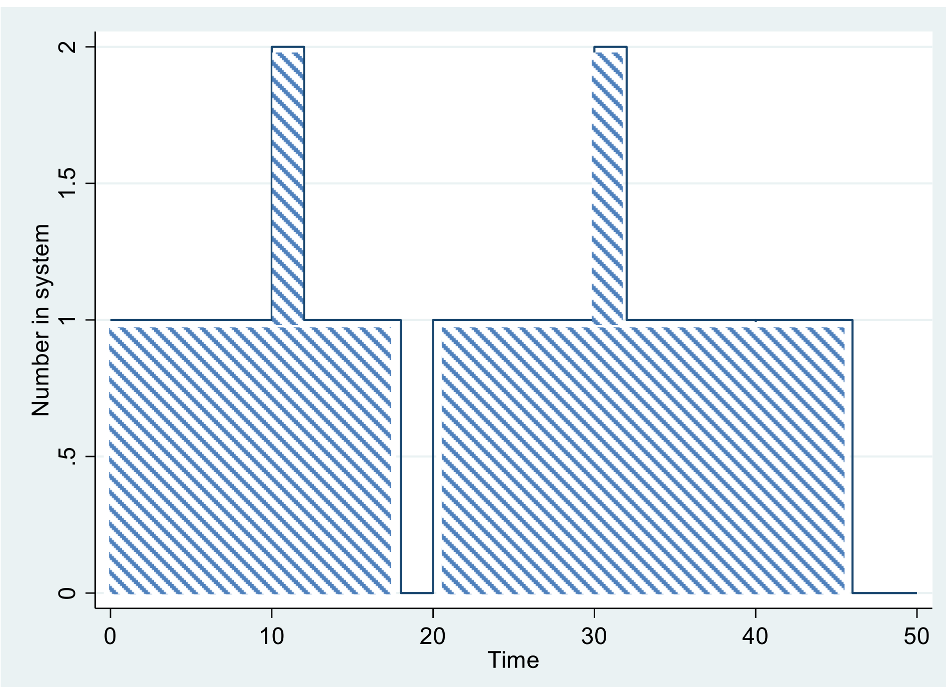
Пример

После изготовления изделия обследуются службой контроля качества.
В пункт контроля изделия поступают с интервалом в десять минут: $t = 0, 10, 20, \ldots$
Для первых пяти изделий известно время завершения обследования:

12	18	32	40	46

Перед поступлением изделия в момент $t = 0$ пункт контроля свободен.
По данным о первых пятидесяти минутах работы рассчитайте:
(а) долю времени простоя; (б) среднее число изделий на пункте контроля.

Решение. (б) Площадь заштрихованной фигуры – суммарное время, которое все изделия пробыли на пункте контроля.



6 мин из 50 – 0 изделий на пункте;
40 мин из 50 – 1 изделие;
4 мин из 50 – 2 изделия.

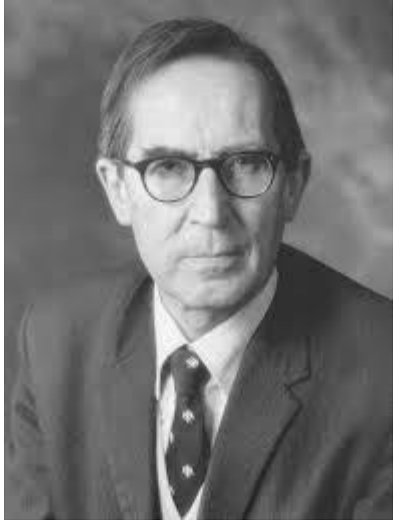
Суммарное время:

$$6 \cdot 0 + 40 \cdot 1 + 4 \cdot 2 = 48 \text{ мин}$$

Среднее число изделий в системе:

$$L = \frac{48}{50} = 0.96.$$

Нотация Кендалла



Дэвид Джордж Кендалл

СМО часто описывается последовательностью символов вида:

$$A \,/\, B \,/\, X \,/\, Y \,/\, Z$$

Здесь *A* означает распределение времени между заявками;
B – распределение времени обслуживания;
X – число каналов обслуживания;
Y – ограничение ёмкости системы (макс. число заявок в системе);
Z – дисциплина обслуживания.

Примеры использования:

"Consider an $M/M/\infty$ queue …"
"This result holds for $M/G/1$ queue"
"Erlang loss formula describes the probability of loss for $M/G/c/c$ queue"
"We derive it for $M/M/1/\infty/FCFS$ model but the results are valid for $M/M/1/\infty/GD$"

ЧТО ЭТО ЗНАЧИТ?

Нотация Кендалла

Распределения (времени между заявками и времени обслуживания):

M	– экспоненциальное (*с какой стати?!*);
E_k (или Er_k)	– распределение Эрланга порядка *k*;
D	– детерминированное;
G (или *GI*)	– произвольное (general или general independent);

Дисциплина обслуживания:

FCFS (или *FIFO*)	– первый пришёл – первый обслуживается;
LCFS (или *LIFO*)	– последний пришёл – первый обслуживается;
RSS (или *SIRO*)	– случайный выбор из очереди;
PR (или *PRI*)	– приоритетное обслуживание;
PS	– совместное обслуживание;
GD	– произвольная дисциплина.

По умолчанию:

не указано макс. число заявок -> неограниченная ёмкость;
не указана дисциплина -> FCFS или произвольная.

Нотация Кендалла

Примеры:

$M/M/1$	– время между заявками и время обслуживания экспоненциально распределено; – 1 канал; – неограниченная ёмкость системы (может вместить неограниченное число заявок).
$D/G/3/3$	– заявки приходят в строго определённое время; – время обслуживания может иметь любое распределение; – 3 канала обслуживания; – не более трёх заявок в системе (=> не может быть очереди).
$M/E_2/1/\infty/PR$	– экспоненциально распределённые промежутки между заявками; – время обслуживания распределено по закону Эрланга второго порядка; – 1 канал; – неограниченная ёмкость; – порядок обслуживания определяется приоритетом заявки.

Нотация Кендалла

Иногда используется расширенная нотация:

A / B / X / Y /Population size/ Z

Здесь *A* означает распределение времени между заявками;
B – распределение времени обслуживания;
X – число каналов обслуживания;
Y – ограничение ёмкости системы (макс. число заявок);
Population size – общее число клиентов во вселенной (∞ по умолчанию);
Z – дисциплина обслуживания.

Пример:

M/*G*/1/∞/30/*FCFS*

- экспоненциальное время между заявками,
- произвольное распределение времени обслуживания,
- один канал обслуживания,
- неограниченная ёмкость,
- всего существует 30 клиентов,
- «первый пришёл – первый обслужен».

Примеры?

А что когда все клиенты будут обслужены?

Идентификация системы: пример 1

Малое предприятие располагает шестью автомобилями, которые сдаёт в пользование. Заявки на аренду автомобилями поступают в среднем по пять в день, время между заявками экспоненциально распределено. Клиент арендует машину на время, распределённое по закону Эрланга третьего порядка со средним 1.5 дня. Плата за пользование автомобилем – 110 гульденов в день. Если все автомобили сданы, поступающие заявки получают отказ.

Опишите процесс аренды нотацией Кендалла.

Идентификация системы: пример 1

Малое предприятие располагает шестью автомобилями, которые сдаёт в пользование. Заявки на аренду автомобилями поступают в среднем по пять в день, время между заявками экспоненциально распределено. Клиент арендует машину на время, распределённое по закону Эрланга третьего порядка со средним 1.5 дня. Плата за пользование автомобилем – 110 гульденов в день. Если все автомобили сданы, поступающие заявки получают отказ.

Опишите процесс аренды нотацией Кендалла.

Ответ.

$$M/E_3/6/6$$

M	– экспоненциальное время между заявками;
E_3	– распределение времени обслуживания – Эрланг-3;
6	– число каналов обслуживания (автомобилей);
6	– максимальное число заявок в системе;

Что с дисциплиной?

Идентификация системы: пример 2

Луноход отсылает сигналы на станцию с промежутками ровно в 1 сек. Станция обрабатывает сигналы в порядке поступления за экспоненциально распределённое время со средним 0.4 сек. Приёмник станции имеет буфер, вмещающий три сообщения (помимо находящегося в обработке). Сигналы, поступающие при заполненном буфере, теряются.

Опишите систему в нотации Кендалла.

Идентификация системы: пример 2

Луноход отсылает сигналы на станцию с промежутками ровно в 1 сек. Станция обрабатывает сигналы в порядке поступления за экспоненциально распределённое время со средним 0.4 сек. Приёмник станции имеет буфер, вмещающий три сообщения (помимо находящегося в обработке). Сигналы, поступающие при заполненном буфере, теряются.

Опишите систему в нотации Кендалла.

Ответ.

$$D/M/1/4/FCFS$$

Пример задачи

Заявки поступают через интервалы ровно в 20 минут.
Время обслуживания распределено от 12 до 22 минут равномерно.
В системе один канал обслуживания.
Заявки, поступающая в занятую систему, получают отказ (очереди нет).

Найдите вероятность, что заявке с порядковым номером *k* будет отказано.

Обозначим эту вероятность $p_l(k)$.

Между прочим, это система D/U/1/1, где U – равномерное распределение (uniform distribution)

Пример задачи

Заявки поступают через интервалы ровно в 20 минут.
Время обслуживания распределено от 12 до 22 минут равномерно.
В системе один канал обслуживания.
Заявки, поступающая в занятую систему, получают отказ (очереди нет).

Найдите вероятность, что заявке с порядковым номером *k* будет отказано.

Обозначим эту вероятность $p_l(k)$

Решение:
а) *k* = 1: первая заявка точно обслуживается, так что $p_l(1) = 0$.

Пример задачи

Заявки поступают через интервалы ровно в 20 минут.
Время обслуживания распределено от 12 до 22 минут равномерно.
В системе один канал обслуживания.
Заявки, поступающая в занятую систему, получают отказ (очереди нет).

Найдите вероятность, что заявке с порядковым номером *k* будет отказано.

Обозначим эту вероятность $p_l(k)$

Решение:
а) *k* = 1: первая заявка точно обслуживается, так что $p_l(1) = 0$.

k = 2: вторая заявка получит отказ, если обслуживание первой займёт > 20 мин. Значит,

$$p_l(2) = \frac{22 - 20}{22 - 12} = 0.2$$

Пример задачи

Заявки поступают через интервалы ровно в 20 минут.
Время обслуживания распределено от 12 до 22 минут равномерно.
В системе один канал обслуживания.
Заявки, поступающая в занятую систему, получают отказ (очереди нет).

Найдите вероятность, что заявке с порядковым номером *k* будет отказано.

Обозначим эту вероятность $p_l(k)$

Решение:
а) *k* = 1: первая заявка точно обслуживается, так что $p_l(1) = 0$.

k = 2: вторая заявка получит отказ, если обслуживание первой займёт > 20 мин. Значит,

$$p_l(2) = \frac{22 - 20}{22 - 12} = 0.2$$

k = 3: третья заявка получит отказ, если второй не будет отказано, а её обслуживание займёт более 20 мин.

$$p_l(3) = (1 - 0.2) \cdot 0.2 = 0.16$$

Получаем разностное уравнение:

$$p_l(k) = (1 - p_l(k - 1)) \cdot 0.2$$

Пример задачи

$$p_l(k) = (1 - p_l(k-1)) \cdot 0.2 = 0.2 - 0.2p_l(k-1)$$

Все решения сходятся к стационарному:

$$p_l = 0.2 - 0.2p_l \Rightarrow p_l = \frac{1}{6}.$$

После короткого *переходного периода* система сходится к *стационарному режиму*.

Примерный план курса

1 модуль:

Введение в дифференциальные и разностные уравнения;
Моделирование входящего потока заявок;
Марковские цепи в дискретном времени.

2 модуль:

Марковские цепи в непрерывном времени;
Процесс размножения и гибели;
Одноканальные экспоненциальные системы массового обслуживания.

3 модуль:

Многоканальные экспоненциальные СМО;
Система M/G/1.

Вариативная часть:

- СМО с неординарным потоком заявок (заявки пакетами);
- обслуживание с приоритетами;
- нетерпеливые заявки;
- эрланговские модели.

Формы контроля

- две контрольные работы;
- два домашних задания;
- экзамен.

Оценка за 1 модуль: 0.6*[к/р 1] + 0.4*[д/з 1].
Оценка за 3 модуль: 0.2*[к/р 2] + 0.2*[д/з 2] + 0.6*[экзамен].
Итоговая оценка: 0.3*[оценка за 1 модуль] + 0.7*[оценка за 3 модуль].

Книги

При подготовке курса я опирался на эту книгу:
▶ Donald Gross, Carl M. Harris (+ соавторы). *Fundamentals of Queueing Theory*.

Дополнительная литература:

▶ S.M. Ross. Introduction to Probability Models.
▶ A.A. Allen. Probability, Statistics and Queueing Theory with Computer Science Applications.
▶ Е.С. Вентцель. Исследование операций.
▶ Б.В. Гнеденко, И.Н. Коваленко. Введение в теорию массового обслуживания.
▶ Б.В. Гнеденко. Беседы о теории массового обслуживания.

Классика может оказаться полезной:
▶ А.Я. Хинчин. *Математические методы в теории массового обслуживания*.

Геометрическая интерпретация математического ожидания

Пусть T – неотрицательная случайная величина с функцией распределения

$$F(t) = P(T \leq t).$$

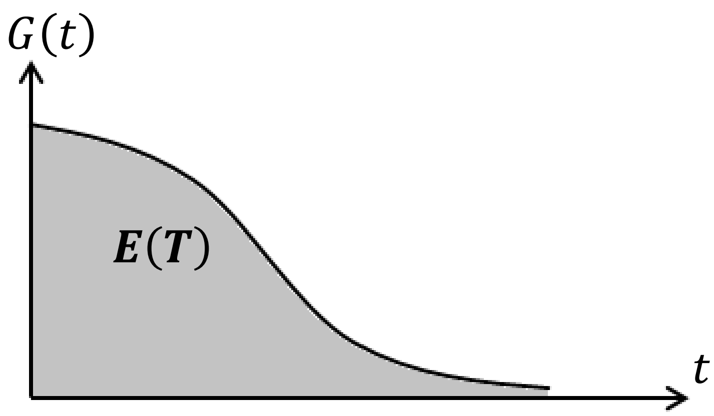
Дополнительная функция распределения (она же функция надёжности, функция дожития):

$$G(t) = P(T > t) = 1 - F(t).$$

Полезная формула:

$$E(T) = \int_0^\infty G(t)dt$$

Матожидание случайной величины равно площади под графиком дополнительной функции распределения.

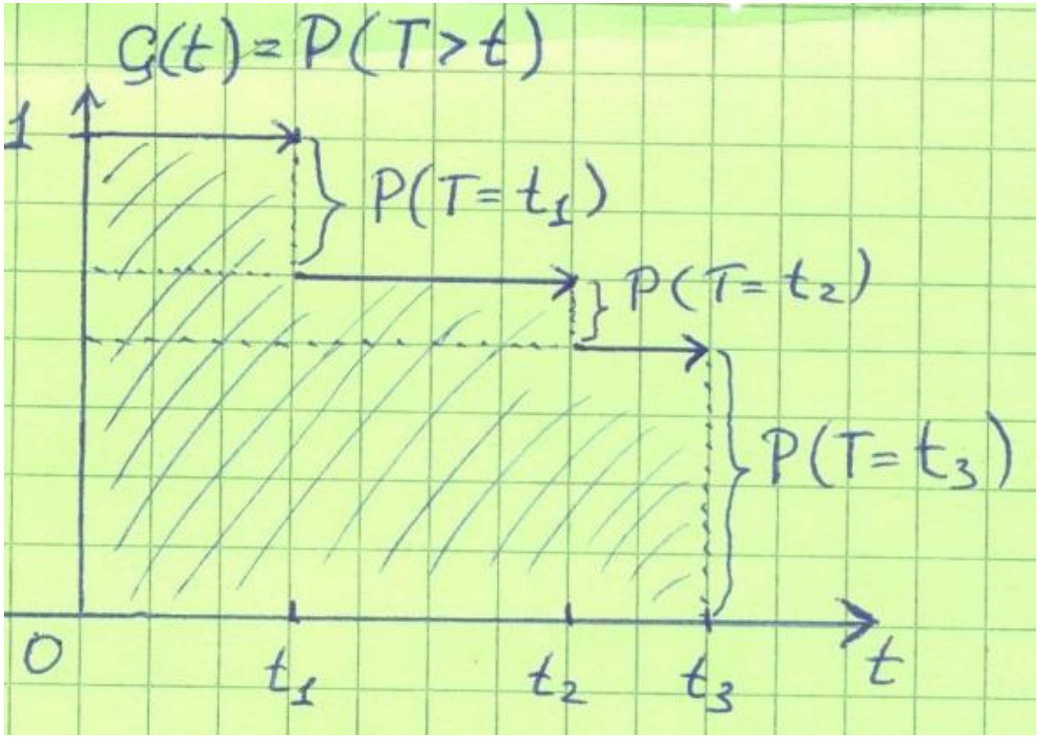


Это верно для **всех** неотрицательных с.в. : дискретных, непрерывных, смешанных.

Есть вариант и для с.в., принимающих отрицательные значения.

Почему так?

Разберёмся с дискретным случаем.

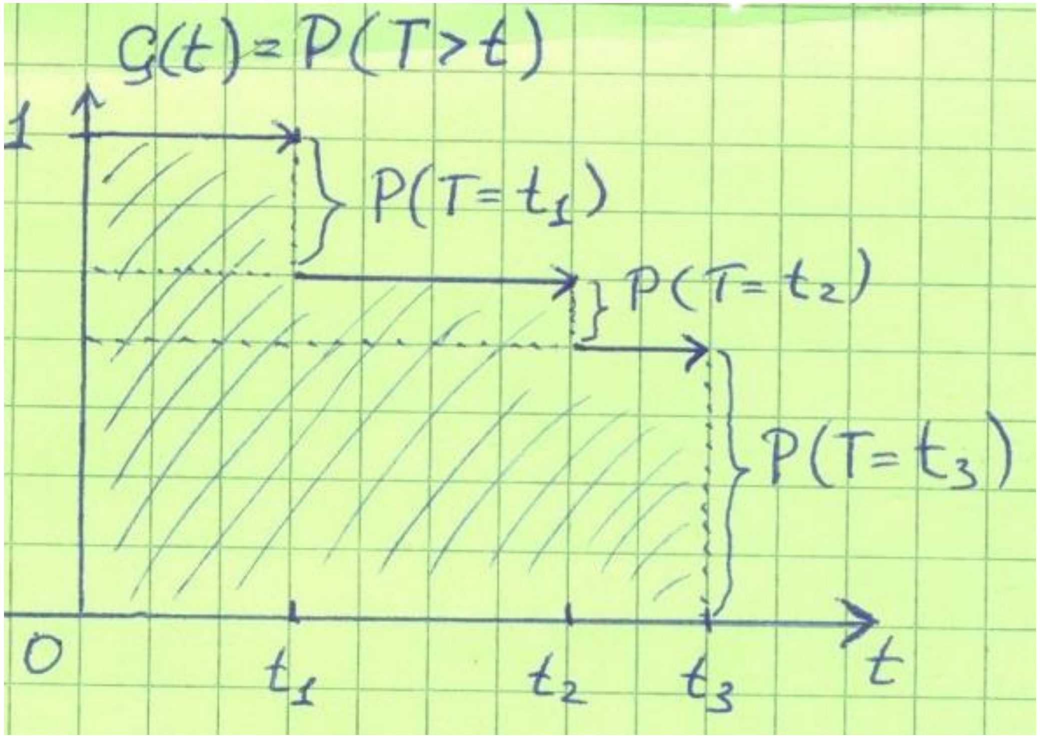


$$\int_0^\infty G(t)dt = \sum_{i=1}^{k} t_i P(T = t_i) = E(T).$$

А для непрерывной или смешанной случайной величины?

Почему так?

Разберёмся с дискретным случаем.



$$\int_0^\infty G(t)dt = \sum_{i=1}^{k} t_i P(T = t_i) = E(T).$$

А для непрерывной или смешанной случайной величины?

Берётся последовательность $T_1, T_2, \dots$ с функциями распределения, стремящимися к $F(t)$.

$$\int_0^\infty G_j(t)dt = \sum_{i=1}^{k_i} t_i P(T_j = t_i) \underset{j\to\infty}{\rightarrow} \int_0^\infty t dF(t) = E(T)$$

Пример: найдём м.о. экспоненциальной случайной величины

Пусть $T \sim \mathrm{Exp}(\lambda)$, так что $F(t) = 1 - e^{-\lambda t}, t \geq 0$.

Дополнительная функция распределения: $G(t) = 1 - F(t) = e^{-\lambda t}$.

Пример: найдём м.о. экспоненциальной случайной величины

Пусть $T \sim \mathrm{Exp}(\lambda)$, так что $F(t) = 1 - e^{-\lambda t}, t \geq 0$.

Дополнительная функция распределения: $G(t) = 1 - F(t) = e^{-\lambda t}$.

Математическое ожидание:

$$E(T) = \int_0^\infty e^{-\lambda t} dt = -\frac{1}{\lambda} e^{-\lambda t} \Big|_0^\infty = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}\ .$$

Другие моменты тоже можно выразить через дополнительную функцию распределения.
Но мы обойдёмся.

Ещё пример: матожидание смешанной случайной величины

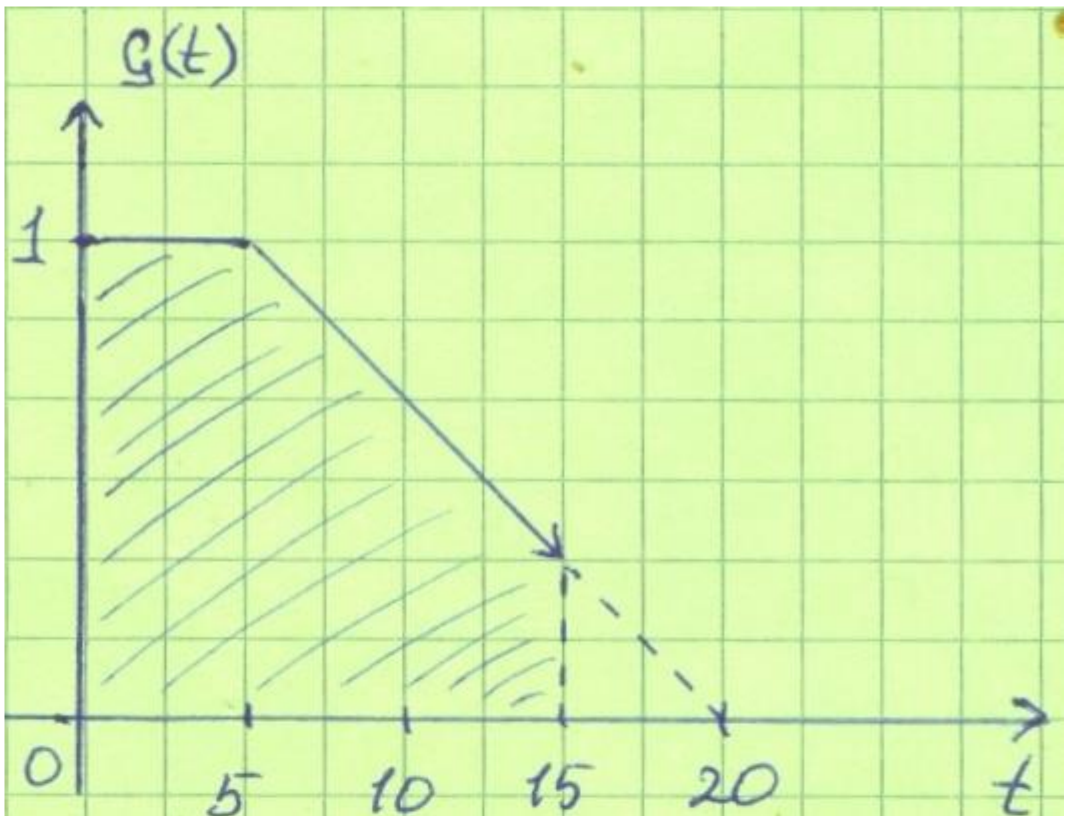
Обслуживание клиента требует времени, равномерно распределённого от 5 до 20 мин, но клиенты нетерпеливы: если их обслуживание длится уже 15 минут, они убегают прочь из системы.

Каково среднее время, которое клиент проводит на обслуживании?

Ещё пример: матожидание смешанной случайной величины

Обслуживание клиента требует времени, равномерно распределённого от 5 до 20 мин, но клиенты нетерпеливы: если их обслуживание длится уже 15 минут, они убегают прочь из системы.
Каково среднее время, которое клиент проводит на обслуживании?

Решение. $E(T) = 5 + \frac{1+\frac{1}{3}}{2} \cdot (15-5) = 11\frac{2}{3}$ мин.



Вопрос напоследок

Есть ли смысл изучать теорию массового обслуживания?

▶ Если интересно, что ещё нужно?

▶ Даже если вы не будете изучать системы массового обслуживания, некоторые темы курса могут оказаться полезными.

▶ А вдруг и правда будете заниматься ТМО?

We did it!