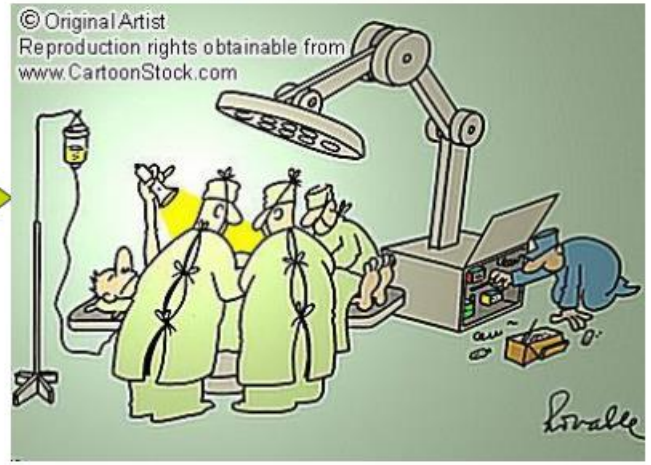
Теория очередей и основы моделирования систем массового обслуживания

Время ожидания в пункте экстренной помощи

Время ожидания

Время обслуживания

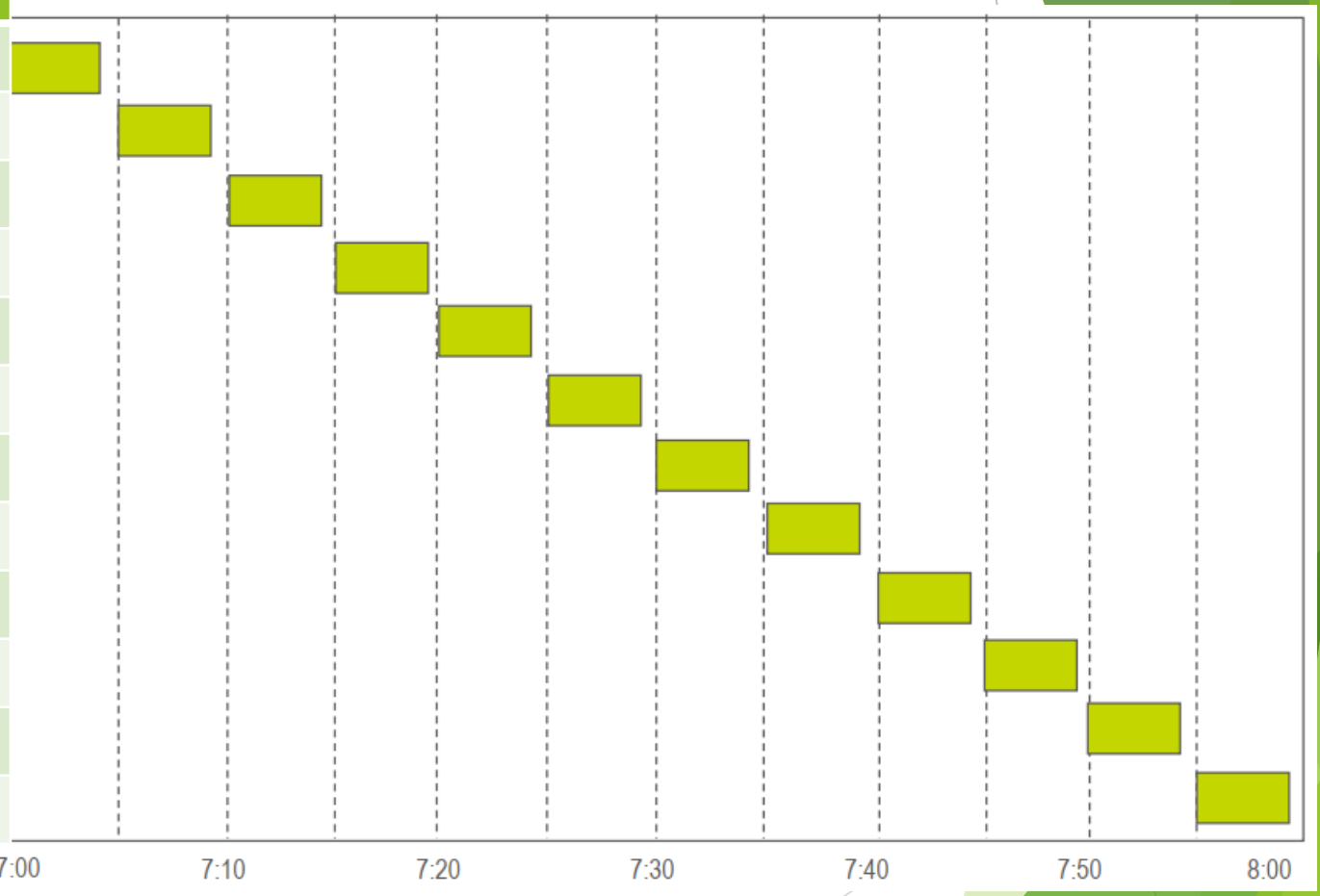


Общее время в системе=ожидание + обслуживание

Странный процесс

Пациент	Время прибытия	Время между прибытиями	Время обслуживания
1	7:00	0	4
2	7:05	5	4
3	7:10	5	4
4	7:15	5	4
5	7:20	5	4
6	7:25	5	4
7	7:30	5	4
8	7:35	5	4
9	7:40	5	4
10	7:45	5	4
11	7:50	5	4
12	7:55	5	4

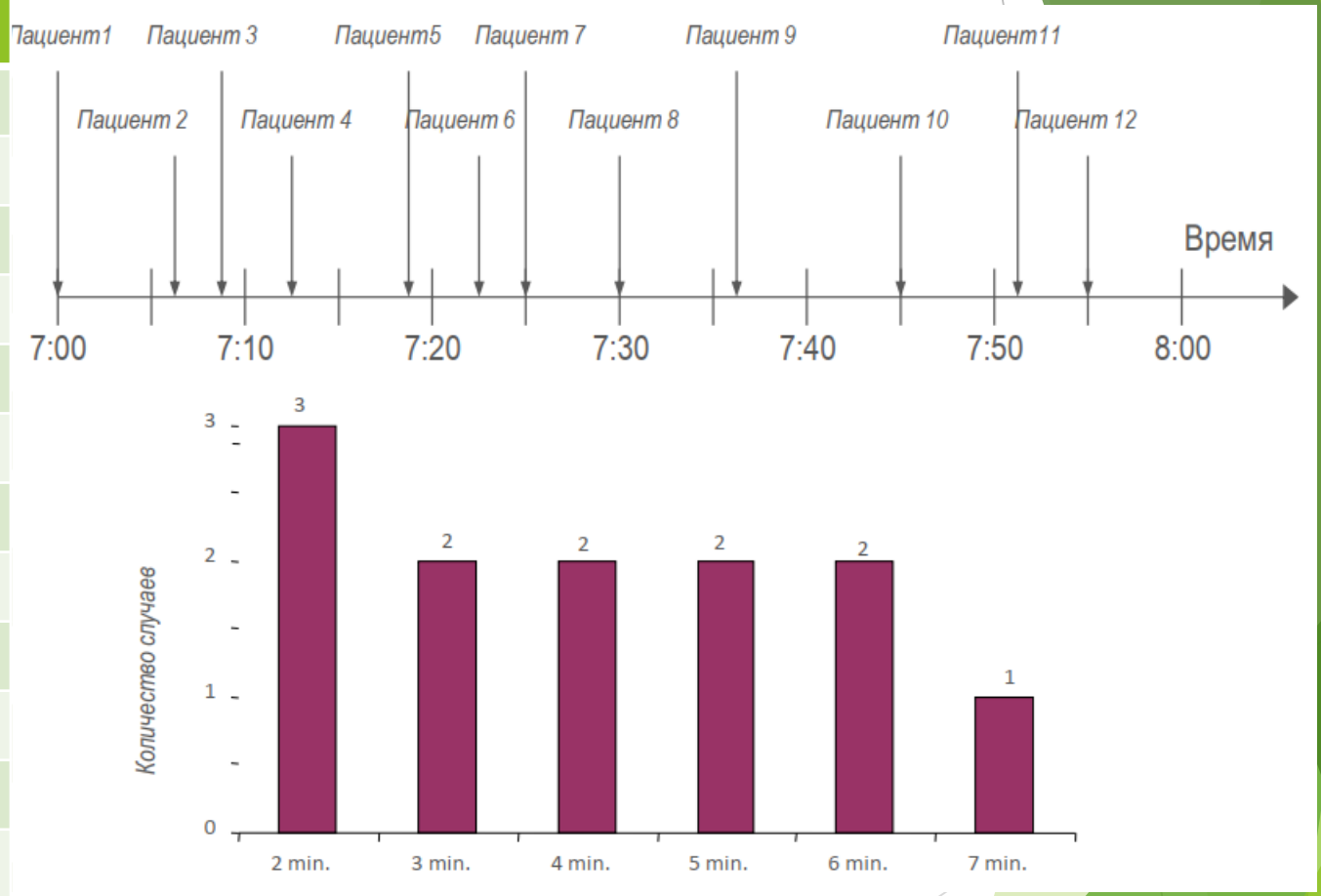
Диаграмма Ганта, иллюстрирующая услугу



- Среднее время ожидания оказания услуги 5 минут, среднее время обслуживания – 4 минуты. Что странного в этом процессе оказания услуги?

Более реалистичный процесс

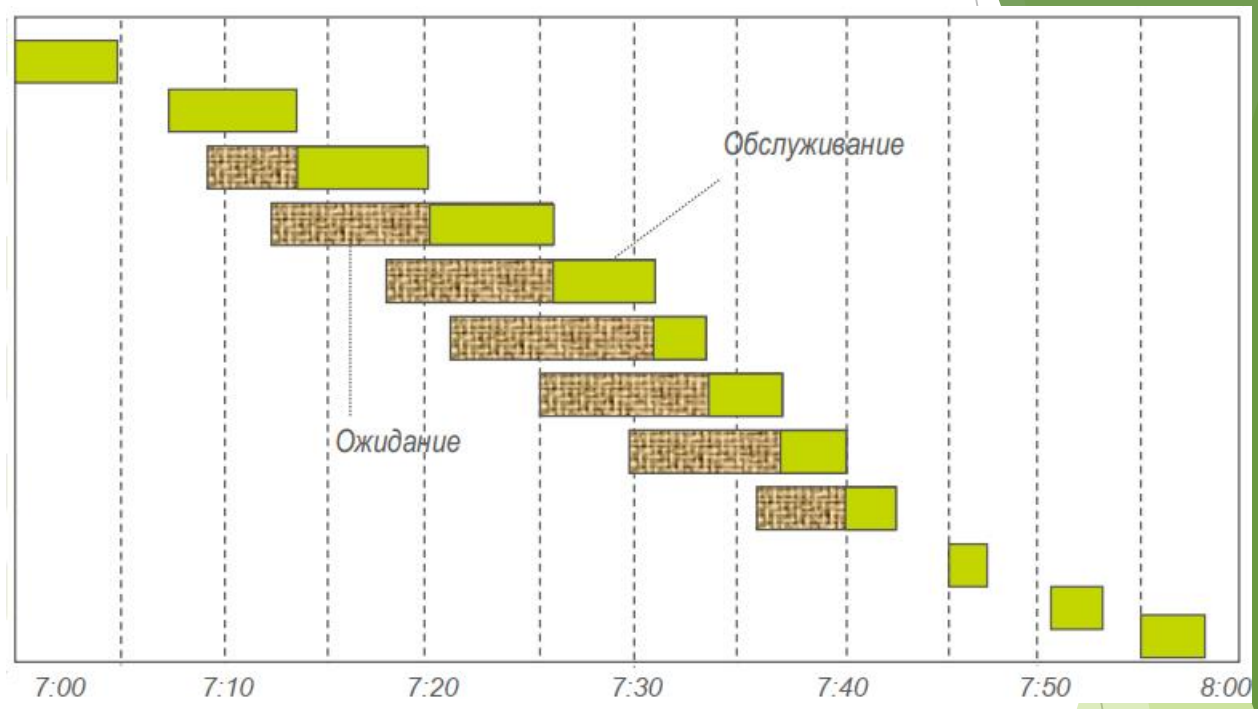
Пациент	Время прибытия	Время между прибытиями	Время обслуживания
1	7:00	0	5
2	7:07	7	6
3	7:09	2	7
4	7:12	3	6
5	7:18	6	5
6	7:22	4	3
7	7:25	3	4
8	7:30	5	3
9	7:36	6	4
10	7:45	9	2
11	7:51	6	2
12	7:55	4	2



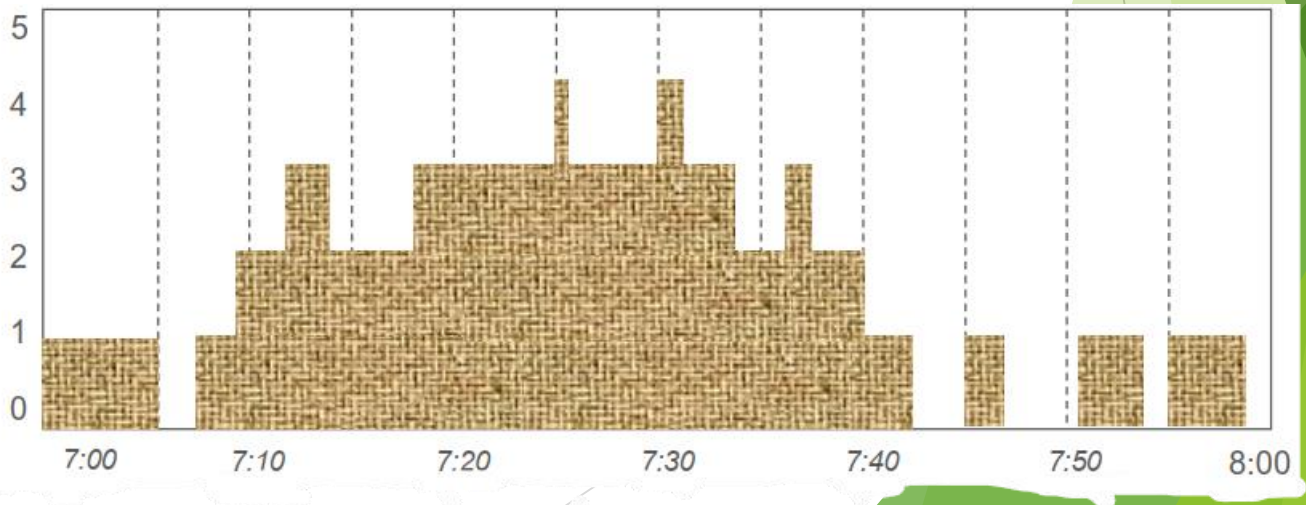
- Среднее время ожидания услуги 5 минут, среднее время оказания услуги 4 минуты. Будет ли производительность процесса такой же, как и в предыдущем случае?

Вариации ведут к ожиданию и промежуточным запасам

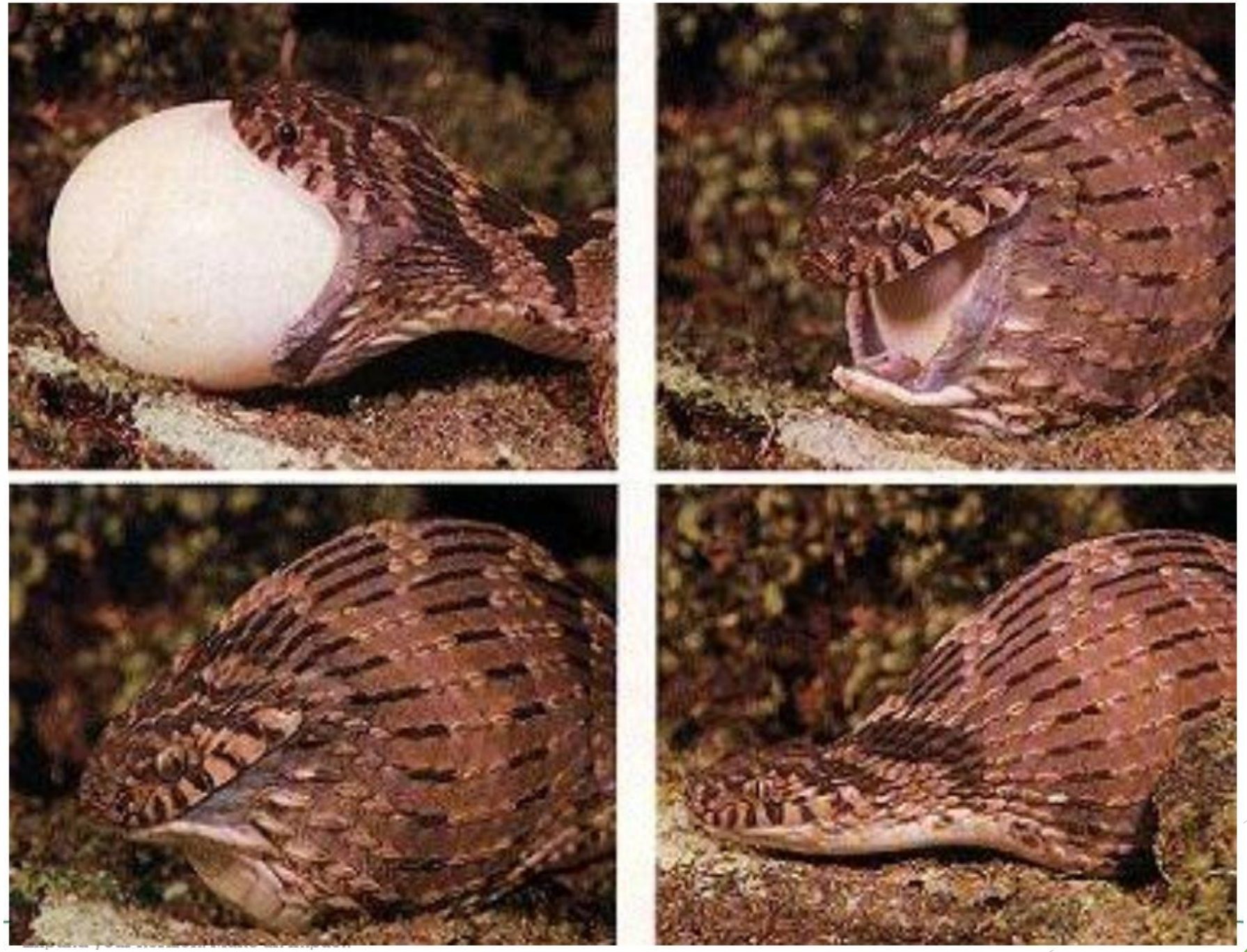
Пациент	Время прибытия	Время между прибытиями	Время обслуживания
1	7:00	0	5
2	7:07	7	6
3	7:09	2	7
4	7:12	3	6
5	7:18	6	5
6	7:22	4	3
7	7:25	3	4
8	7:30	5	3
9	7:36	6	4
10	7:45	9	2
11	7:51	6	2
12	7:55	4	2



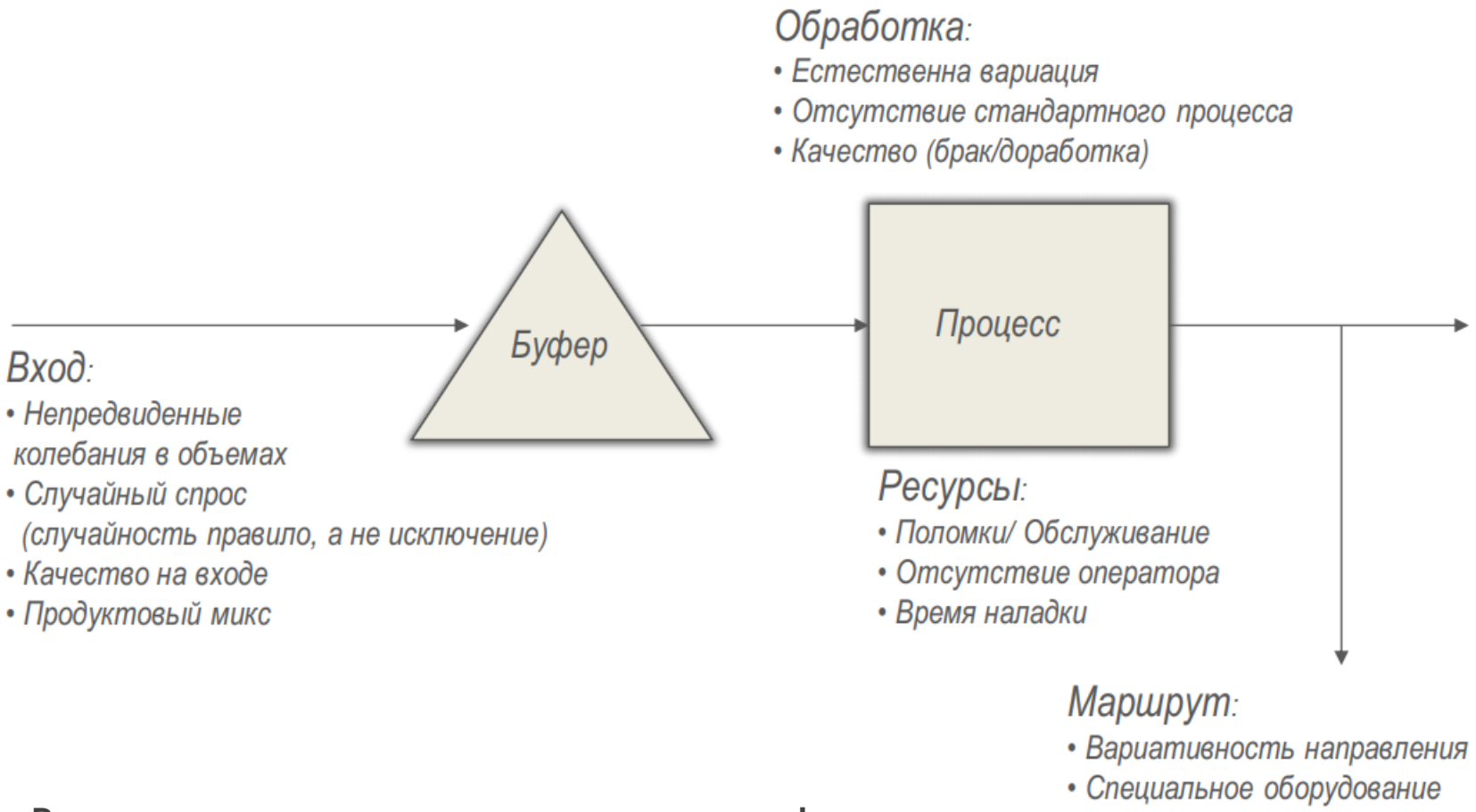
запасы (очередь)



- Вариации это зло!



Вариации – откуда они берутся?



- Вариации это норма, а не исключение!

От процесса к параметрам

Параметры:

- Количество ресурсов: s
- Частота (средняя) прибытия клиентов: λ
- Среднее время сервиса: τ (скорость сервиса $\mu = 1/\tau$)
- Загруженность $\rho = \lambda * \tau / s$
- Коэффициент вариации: КВ = стандартное отклонение / математическое ожидание (либо для периодов между прибытием клиентов либо для времени сервиса):

$$\text{КВ Прибытия} = \text{КВ}_\text{П} = \frac{\delta_\text{П}}{\lambda_\text{П}}$$

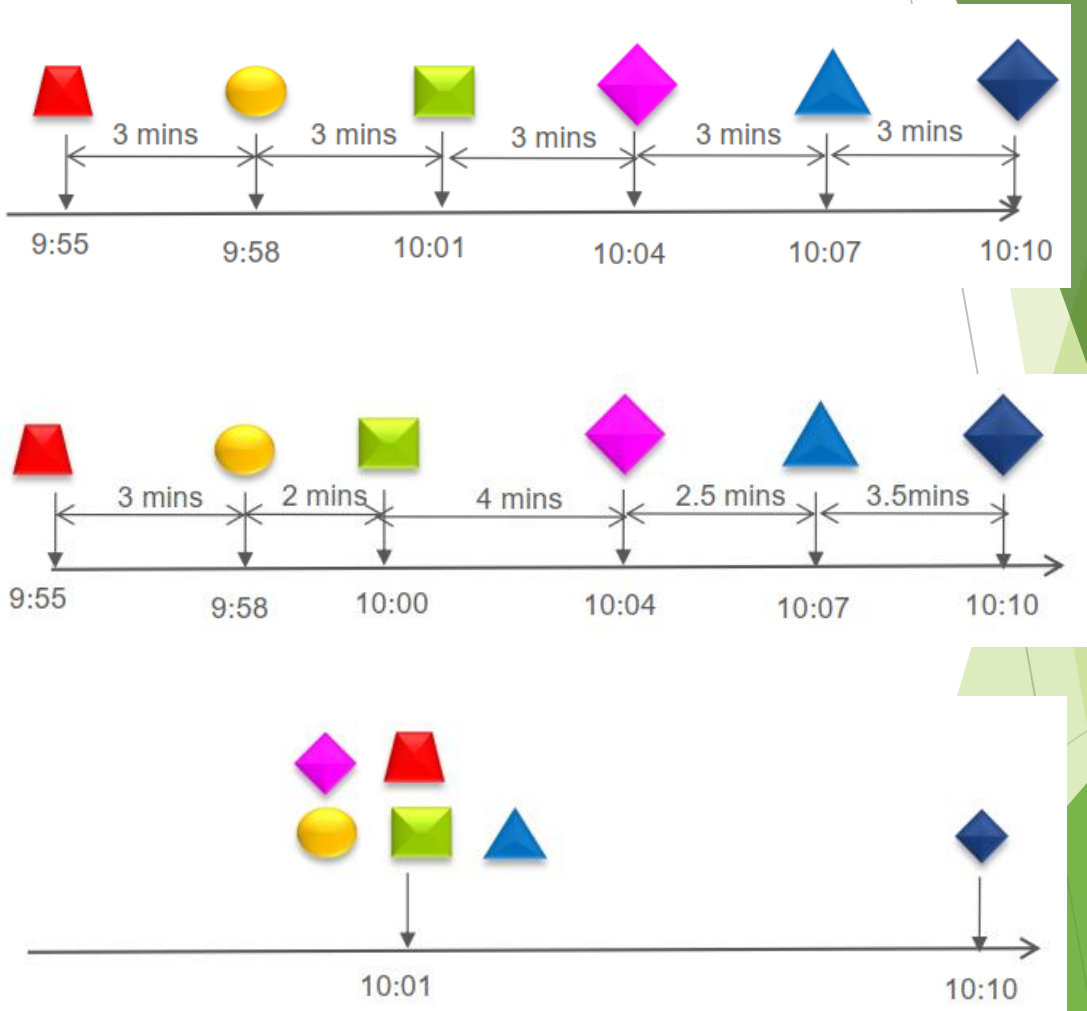
$$\text{КВ Сервиса} = \text{КВ}_\text{С} = \frac{\delta_\text{С}}{\lambda_\text{С}}$$

Усреднённые метрики эффективности:

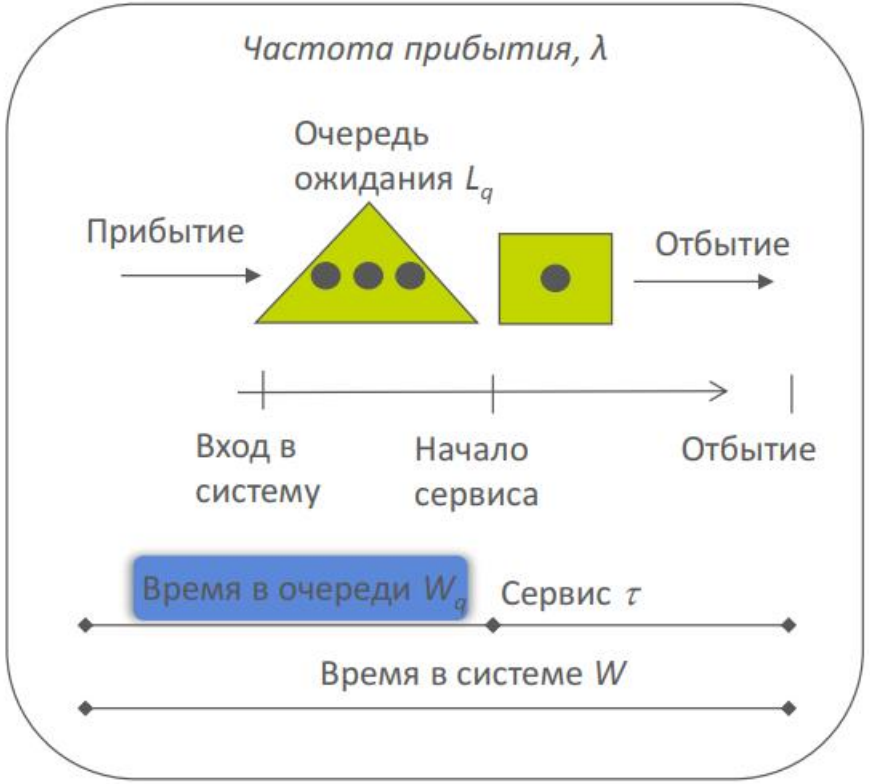
- Время ожидания: W_q
- Общее время в системе: $W = t + W_q$
- Число потребителей в очереди: L_q
- Число потребителей в системе L

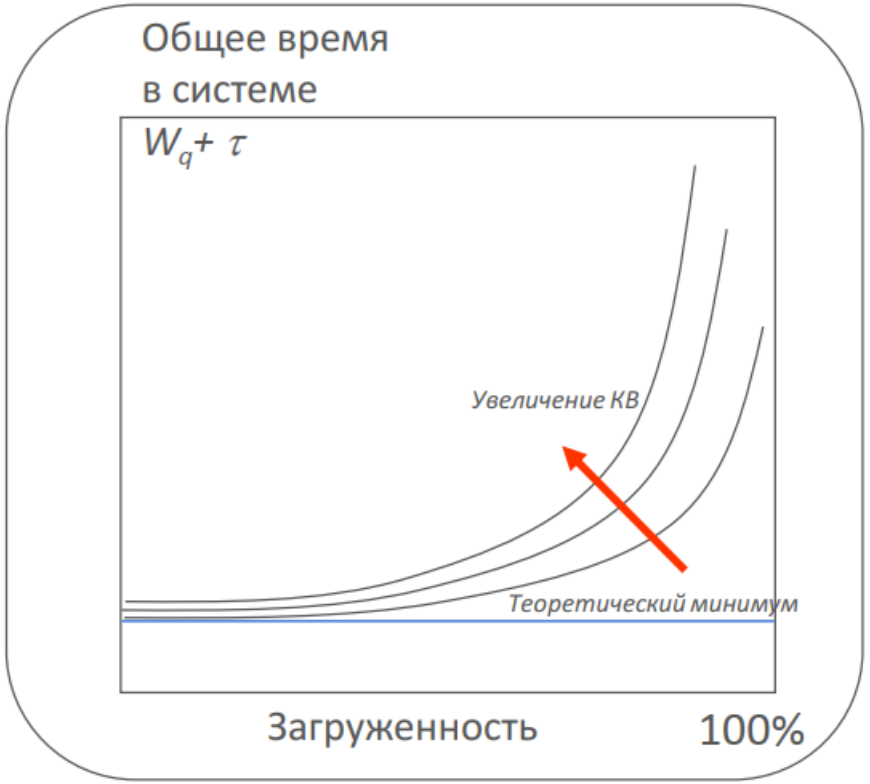
Каков смысл коэффициента вариации?

- Процесс с КВ = 0: Прибытия чётко по графику, например выход продуктов с механической производственной линии
- Процесс с КВ = 1: Прибытия клиентов абсолютно независимы. Например, звонки в телефонный центр. Время между звонками имеет экспоненциальное распределение. Другими словами, прибытия происходят в соответствии с распределением Пуассона
- Процесс с КВ >> 1: Групповые прибытия клиентов: например в обеденный перерыв



Формула ожидания (приближение для 1-го ресурса)





$$W_q = \tau * \left(\frac{\rho}{1-\rho}\right) * \left(\frac{КВ_П^2 + КВ_С^2}{2}\right)$$

Эффект вариативности

Эффект загруженности

Эффект шкалы

Механика вычислений с одним ресурсом

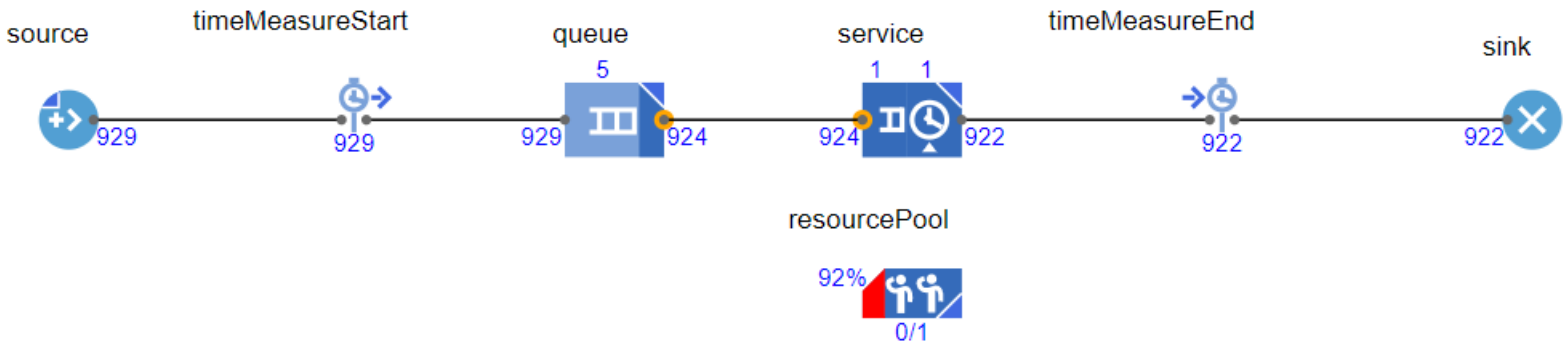
Общий подход

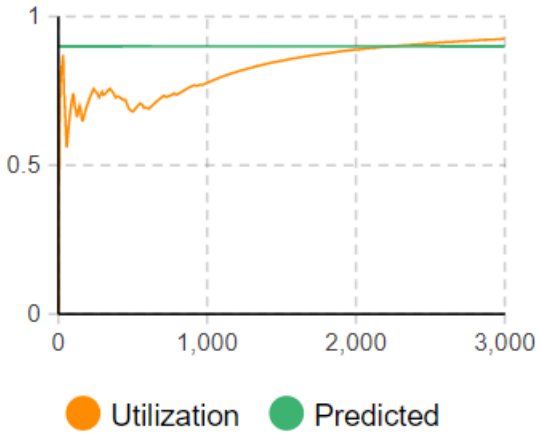
- Каковы входные параметры? $\lambda, \tau, \text{КВ}_\text{П}, \text{КВ}_\text{С}$
- Найти загруженность $\rho = \lambda * \tau$
- Найти время ожидания W_q по формуле $W_q = \tau(\frac{\rho}{1-\rho})\frac{1}{2}(\text{КВ}_\text{П}^2 + \text{КВ}_\text{С}^2)$
- $L_q = \lambda W_q$
- $W = W_q + \tau$
- $L = \lambda W$

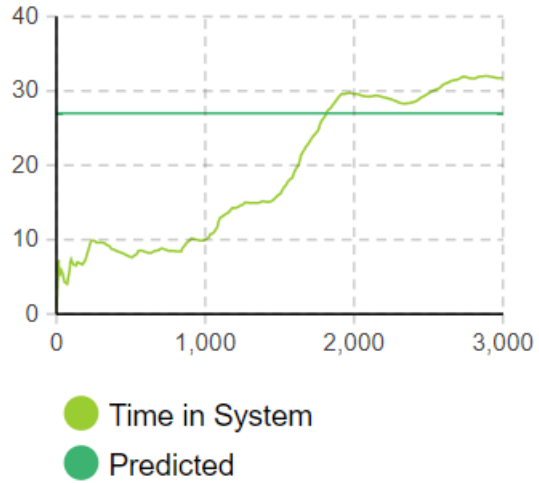
Пример (обслуживание в банке)

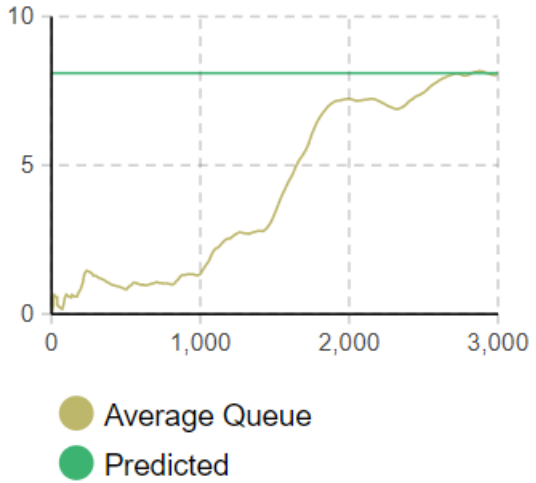
- Клиенты прибывают со скоростью $\lambda = 0{,}3$ в минуту, скорость сервиса в среднем $\mu = 0{,}33$ клиента в минуту, среднее время сервиса $\tau = \frac{1}{0{,}33} = 3$ минуты, $\text{КВ}_\text{П} = \text{КВ}_\text{С} = 1$
- $W_q = 3 * \left(\frac{0{,}9}{0{,}1}\right) * 0{,}5 * (1 + 1) = 27$ мин.
- Клиентов в очереди $L_q = 0{,}3 * 27 = 8{,}1$
- Время в системе $W = 27 + 3 = 30$
- Клиентов в системе $L = 0{,}3 * 30 = 9$

Имитационная модель



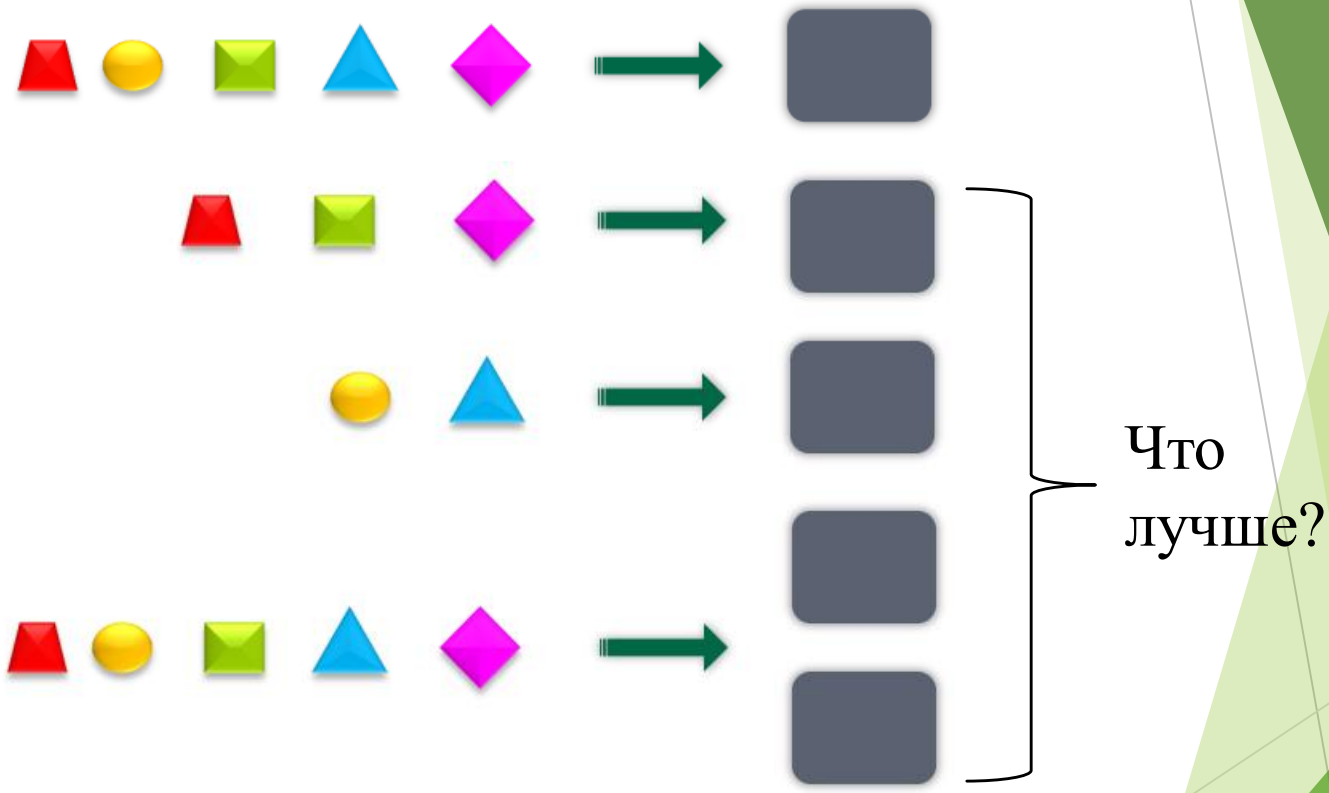






Системы с одним или несколькими ресурсами

- Один ресурс – одна очередь
- Два ресурса – две очереди
- Два ресурса – одна очередь

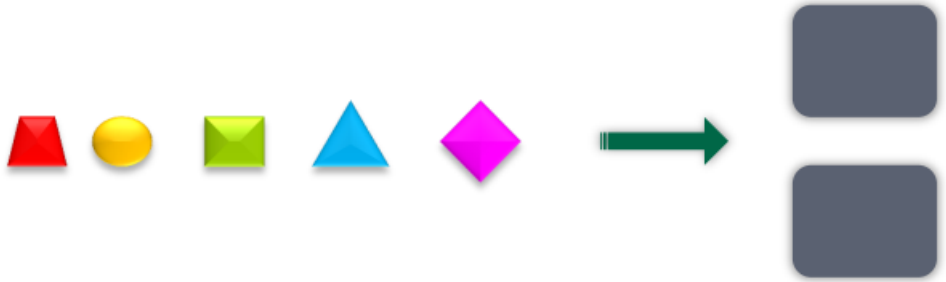


Более общая формула (приближение)

$$W_q = \underbrace{\left(\frac{\tau}{S}\right)}_{\text{Эффект шкалы}} * \underbrace{\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{1-\rho}\right)}_{\text{Эффект загруженности}} * \underbrace{\left(\frac{\text{КВ}_\text{П}^2 + \text{КВ}_\text{С}^2}{2}\right)}_{\text{Эффект вариативности}}$$

- Помните, что $\rho = \lambda * \frac{\tau}{s}$
- Разберём пример:

Механика вычислений с несколькими ресурсами



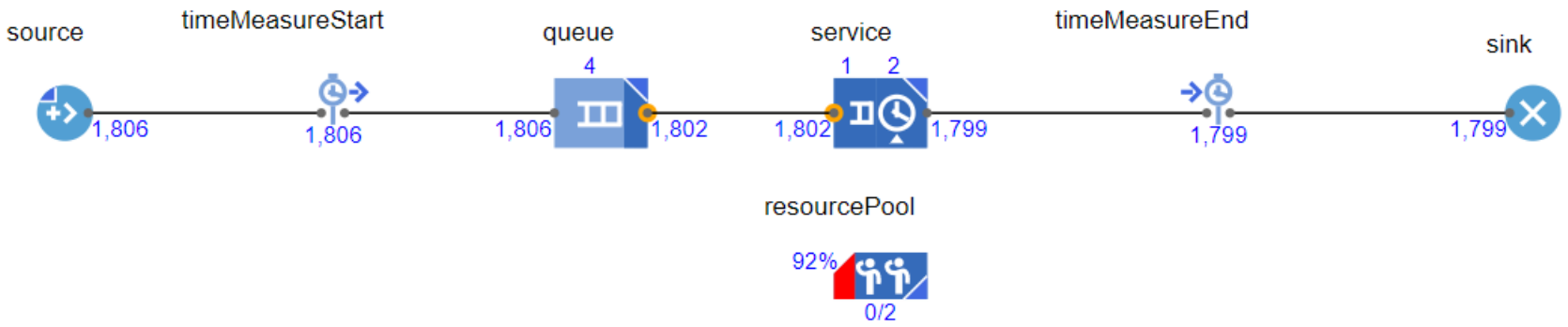
Общий подход

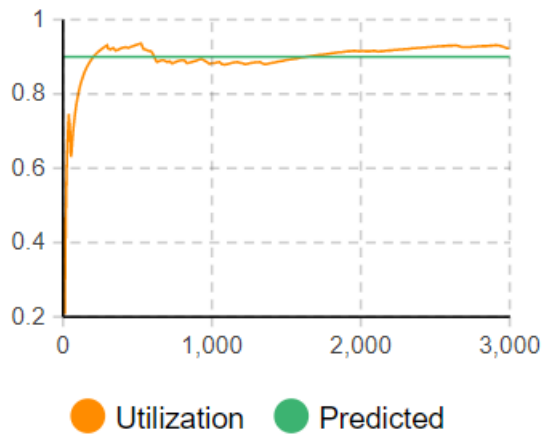
- Каковы входные параметры? $\lambda, \tau, s, \text{КВ}_\text{П}, \text{КВ}_\text{С}$
- Найти загруженность $\rho = \frac{\lambda * \tau}{s}$
- Найти время ожидания W_q по формуле $W_q = (\frac{\tau}{s})(\frac{\rho^{\sqrt{2(s+1)}-1}}{1-\rho})\frac{1}{2}(\text{КВ}_\text{П}^2 + \text{КВ}_\text{С}^2)$
- $L_q = \lambda W_q$
- $W = W_q + \tau$
- $L = \lambda W$

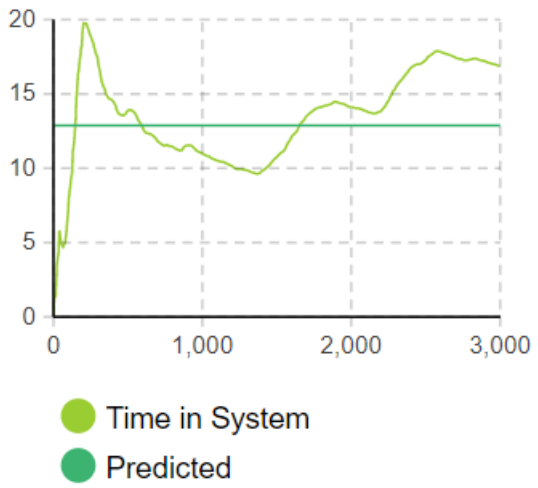
Пример (обслуживание в банке)

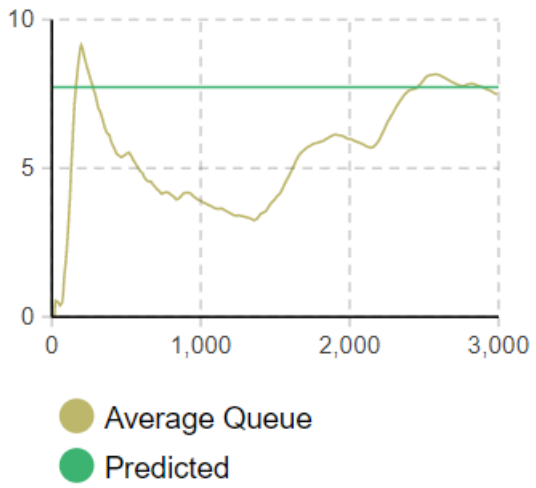
- Клиенты прибывают со скоростью $\lambda = 0{,}6$ в минуту, скорость сервиса в среднем $\mu = 0{,}33$ клиента в минуту, среднее время сервиса $\tau = \frac{1}{0{,}33} = 3$ минуты, s=2 ресурса, $\text{КВ}_\text{П} = \text{КВ}_\text{С} = 1$
- $W_q = \frac{3}{2} * \frac{0{,}9^{\sqrt{2(2+1)}-1}}{1-0{,}9} = 12{,}88$ мин.
- Клиентов в очереди $L_q = 12{,}88 * 0{,}6 = 7{,}73$
- Время в системе $W = 12{,}88 + 3 = 15{,}88$
- Клиентов в системе $L = 0{,}6 * 15{,}88 = 9{,}52$

Имитационная модель



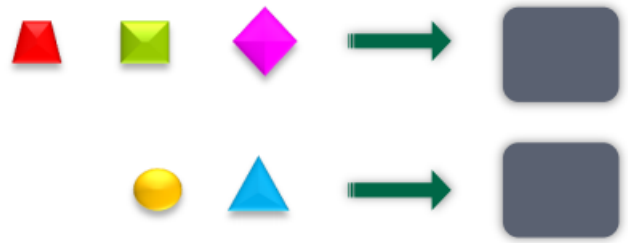




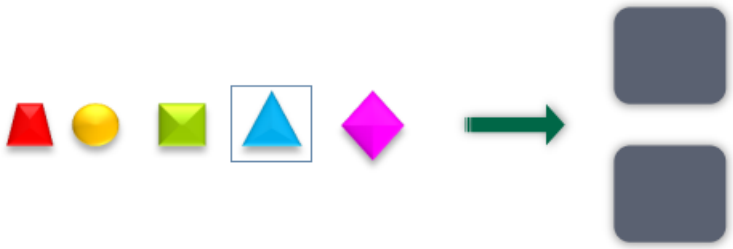


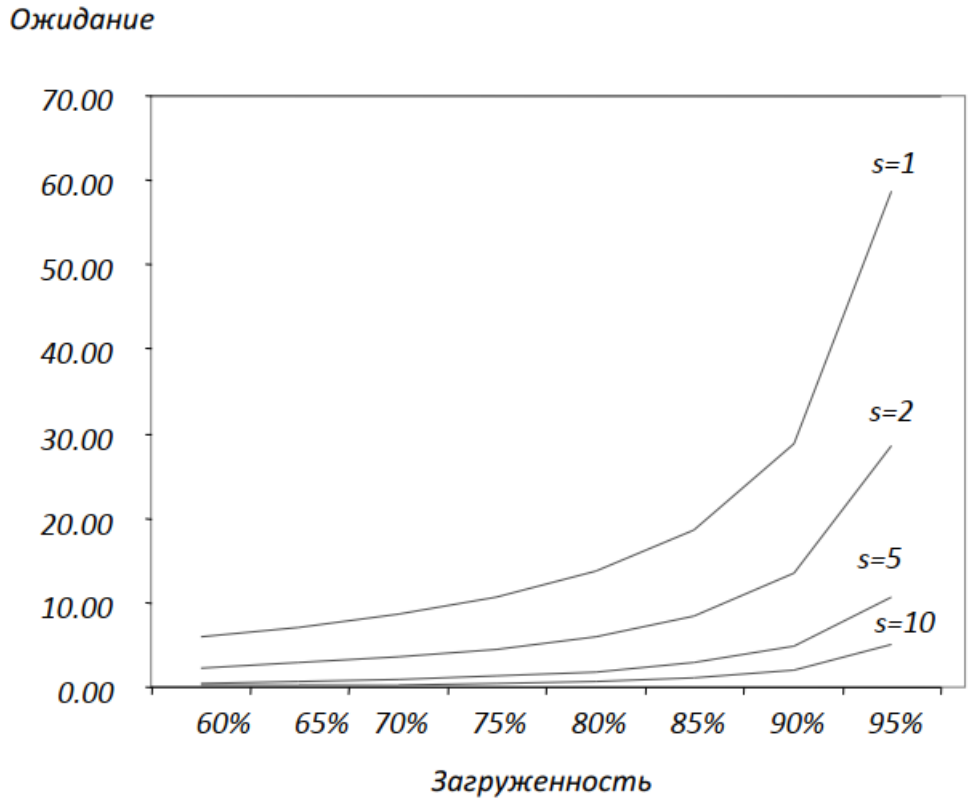
Сила объединения ресурсов

- Два ресурса – две очереди
- Ожидание для клиента 27 минут

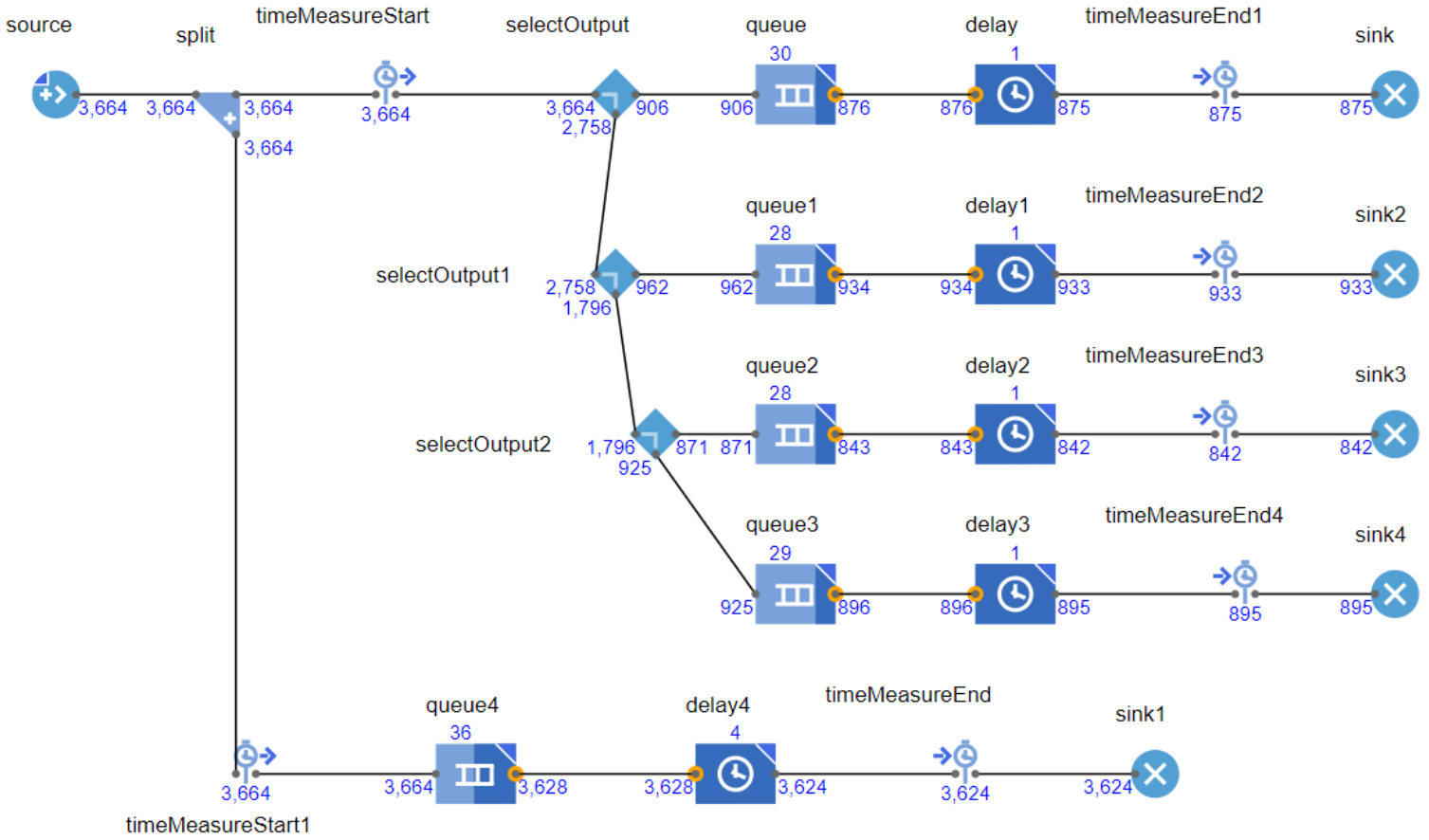


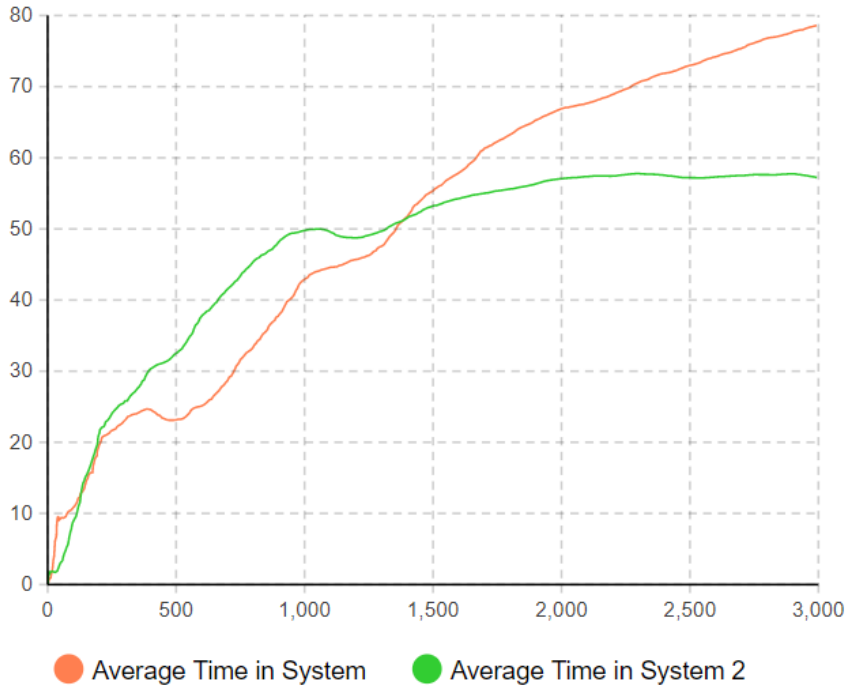
- Два ресурса – одна очередь
- Ожидание для клиента 12,88 минут





Обеспечиваем чистоту эксперимента





Управление системами с ожиданием

- Закон Литтла фундаментален – запасы, производительность и время цикла связаны. Этот закон помогает связать операционные решения с оборачиваемостью запасов, циклом конверсии наличности и прибылью.
- Вариативность ведёт к ожиданию и плохому сервису даже если загруженность ресурсов <100% !
- Вариативность – это норма, а не исключение! Необходимо понять откуда она берётся
- и минимизировать её источники
  - Расписание прибытия клиентов
  - Стимулы прибыть в незагруженные времена
  - Обучение и технологии
  - Чёткие процессы (неправильно, но единообразно)
  - Тренировка клиента
- Остаточной вариативностью нужно эффективно управлять, используя объединение ресурсов и добавочные ресурсы
- **Используйте имитационное моделирование**, чтобы:
  - получить качественное описание системы
  - проанализировать рекомендации/сценарии
- Помните: 100% загруженность ресурсов ведёт к бесконечной очереди, если присутствует вариативность в системе!

Теория ограничений Голдратта (синопсис книги «Цель»)

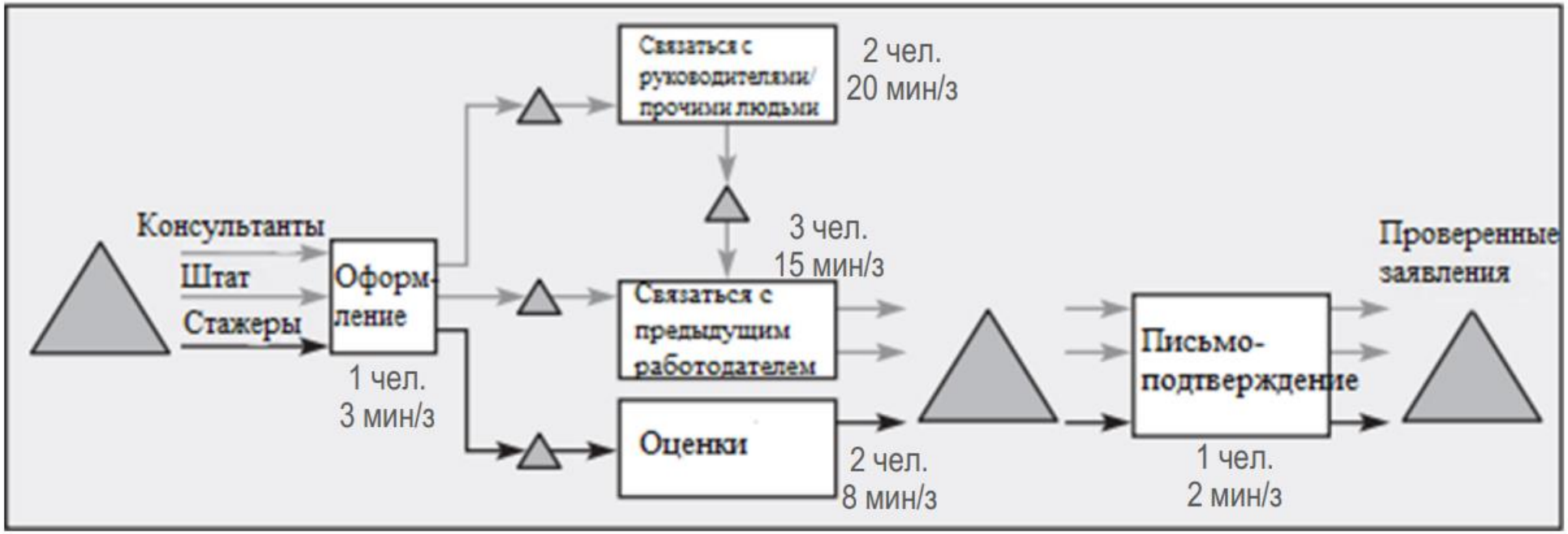
- Практический подход к оптимизации процессов:
  - Определить ограничивающие факторы (узкие места)
  - Подчинить всё остальное этой цели
  - Использовать ограничивающие факторы наилучшим образом.
  - Ликвидировать / уменьшить влияние ограничивающих факторов.
  - Вернуться к шагу 1 – Не допускать инертности!
- Теория ограничений помогает увеличить производительность и наладить плавное протекание процессов

От простого процесса к сложному

- А что если:
  - Единицы процесса разделяются на несколько потоков. Например, в банке в зависимости от сложности кредитной ситуации клиента, возможны разные пути обработки запросов на кредит с исключением разных стадий
  - Имеется несколько видов единиц процесса, которые представляют, например, разные типы клиентов. Например, жалобы от клиентов могут требовать технической, экономической или юридической экспертизы
- Наличие узкого места может зависеть от разнообразия клиентов/потоков: недостаточно знать, что операция занимает много времени, нужно также знать, насколько часто требуется эта операция.
- В этом случае критическое значение имеет правильный выбор *единицы процесса*! (начиная со стадии построения диаграммы).
- Напоминаю что:

$$\text{Предполагаемая загрузка} = \frac{\text{Мощность, необходимая для удовлетворения спроса}}{\text{Имеющаяся мощность}}$$

Пример: как работать с более сложными процессами?



- Спрос: 180 заявлений/день (10 часов в день). Из них:
  - 30 заявлений/день – консультанты
  - 110 заявлений/день – штатные должности
  - 40 заявлений/день - стажёры

Подход №1: единица труда – это заявление

- В этом случае полагаем, что разные типы заявлений приходят случайным образом:
  - С вероятностью 3/18, заявление на должность консультанта
  - С вероятностью 11/18, заявление на штатную должность
  - С вероятностью 4/18, заявление на должность стажера
- По сути, мы отталкиваемся от спроса и определяем требуемую мощность каждого участка исходя из продуктового микса, получаемого от случайного прибытия разных продуктов
- Затем сравниваем требуемую мощность с ресурсами

Принимаем за единицу труда заявления

	Длительность операции [мин/заявка]	Число работников	Имеющаяся мощность [заявок/час]	Требуемая мощность [заявлений/час]				Предпо-лагаемая загрузка
				Консультанты	Штат	Стажеры	Всего	
Оформление	3	1	60/3 = 20	3	11	4	18	18/20 = 90%
Связаться с людьми	20	2	2*60/20 = 6	3	0	0	3	3/6 = 50%
Связаться с работодателями	15	3	3*60/15 = 12	3	11	0	14	14/12 = 117%
Анализ оценок / школы	8	2	3*60/15 = 12	0	0	4	4	4/15 = 27%
Письмо-подтверждение	2	1	60/2 = 30	3	11	4	18	18/30 = 60%

Подход №1: единица труда – это минута работы

- В этом случае мы сначала рассчитываем имеющуюся мощность на этапе как (число работников) × 60 [минут/час]
- Далее находим требуемую мощность: (сколько заявлений различного типа нужно обработать за час) × (сколько минут работы требует обработка заявления каждого из видов на данном участке процесса)
- Сравниваем с имеющейся мощностью чтобы найти узкое место

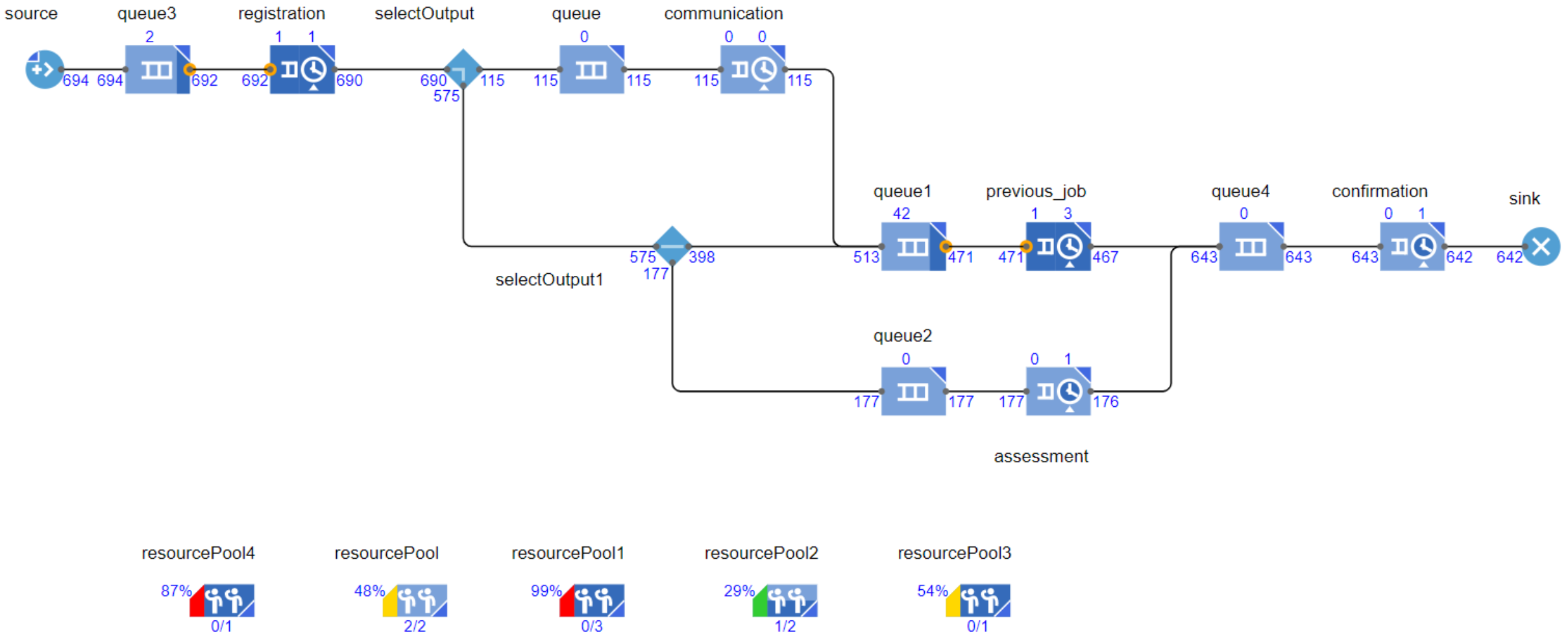
Принимаем за единицу труда минуту работы

	Длительность операции [мин/заявка]	Число работников	Имеющаяся мощность [минут/час]	Требуемая мощность [минут/час]				Предпо-лагаемая загрузка
				Консультанты	Штат	Стажеры	Всего	
Оформление	3	1	60	3 * 3 = 9	11 * 3 = 33	4 * 3 = 12	54	54/60 = 90%
Связаться с людьми	20	2	120	3 * 20 = 60	0	0	60	60/120 = 50%
Связаться с работодателями	15	3	180	3 * 15 = 45	11 * 15 = 165	0	210	210/180 = 117%
Анализ оценок / школы	8	2	120	0	0	4 * 8 = 32	32	32/120 = 27%
Письмо-подтверждение	2	1	60	3 * 2 = 6	11 * 2 = 22	4 * 2 = 8	36	36/60 = 60%

Замечания по поводу двух подходов

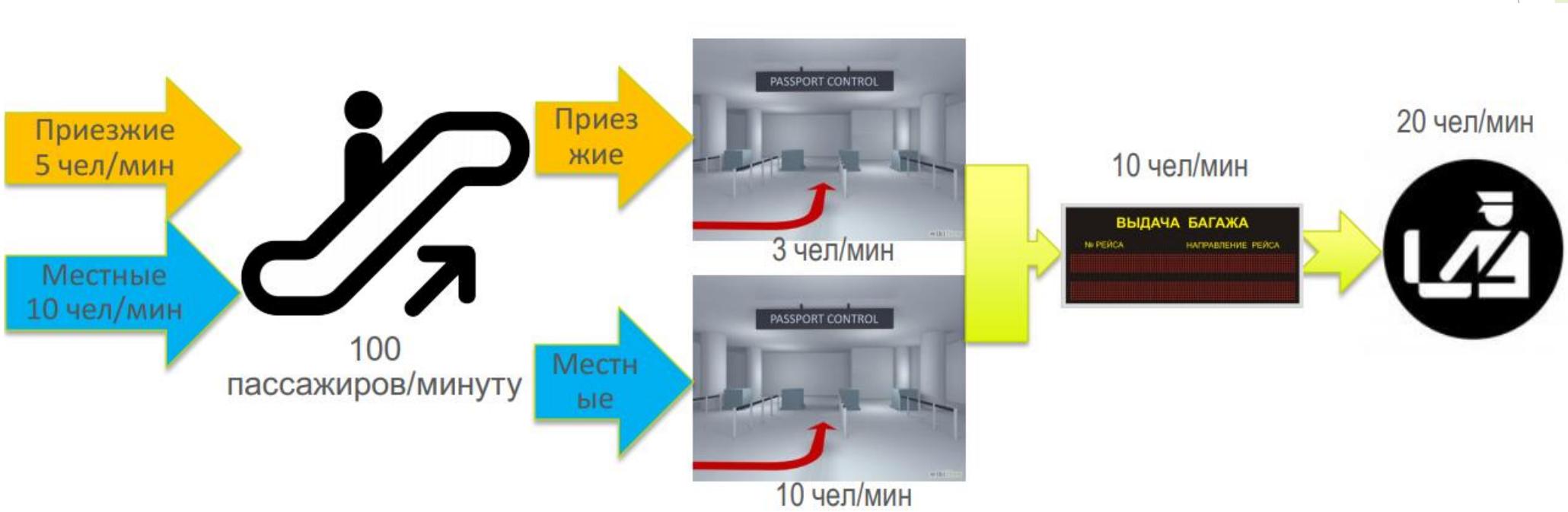
- Обе процедуры нахождения узкого места в случае ассортимента продукции эквивалентны. Ни один из двух подходов не превосходит другой
- Следует помнить, что:
  - Мощность каждого участка можно выразить в терминах этой единицы процесса
  - Каждый вид спроса можно выразить в терминах требуемого числа единиц процесса
- Например, если за единицу процесса взять «одно заявление», то мы можем оценить мощность каждого участка в терминах числа обрабатываемых заявлений в единицу времени
- Если единицей процесса является «одна минута работы», то мы выражаем мощность каждого участка в количестве «минут работы» в единицу времени, и аналогично каждый вид спроса может быть выражен в количестве «минут работы», которое требуется на данном участке

Имитационная модель



Что дальше?

- Мы нашли узкое место или определили обладает ли процесс достаточной мощностью
- Значит ли это, что мы знаем, какова будет реальная скорость протекания процесса и где добавить ресурсы (если такая возможность имеется)? Не всегда! Разберём другой пример: международный аэропорт



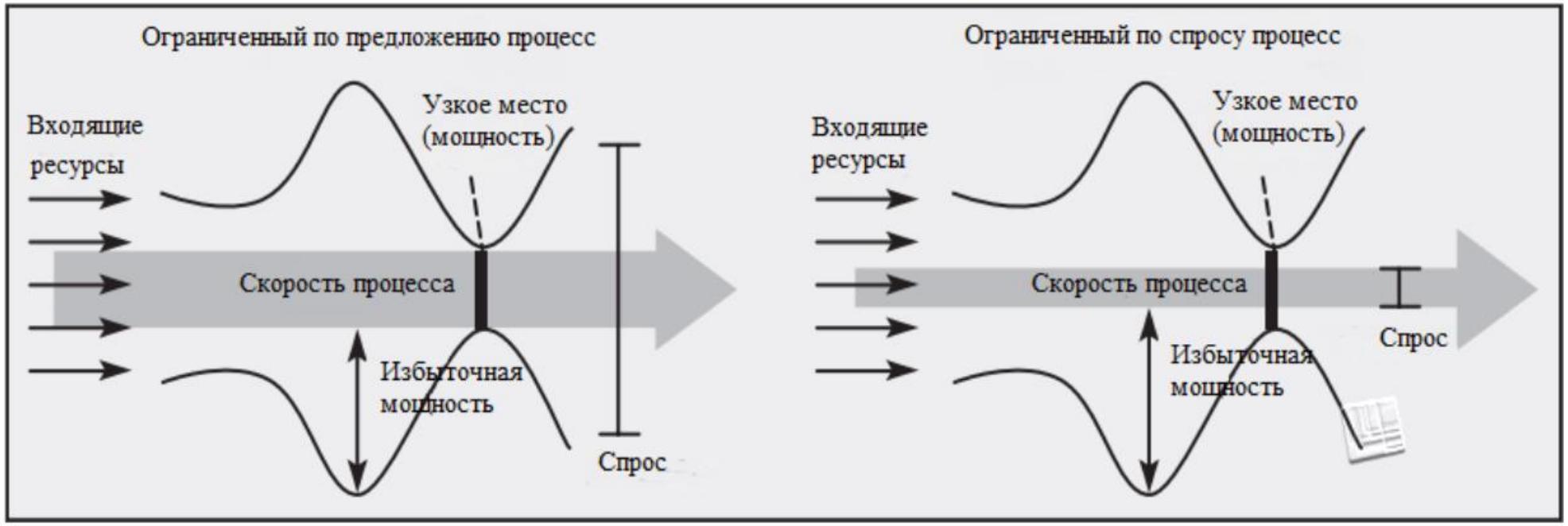
Вычисления

Этап	Спрос со стороны местных граждан [пассажиров / минуту]	Спрос со стороны Приезжих [пассажиров / минуту]	Мощность [пассажиров / минуту]	Предполагаемая загрузка [пассажиров / минуту]
Эскалатор	10	5	100	15/100 = 15%
Паспортный контроль – местные	10	-	10	10/10 = 100%
Паспортный контроль - приезжие	-	5	3	5/3 = 167%
Выдача багажа	10	5	10	15/10 = 150%
Таможенный контроль	10	5	20	15/20 = 75%

Что мы получаем

- Узким местом является участок паспортного контроля приезжих.
- Три приезжих в минуту покидают зону паспортного контроля и направляются в зону выдачи багажа.
- Вместе с 10 местными пассажирами в минуту, это создает поток из 13 пассажиров в минуту на этапе выдачи багажа, который имеет мощность всего 10 пассажиров в минуту
- Таким образом, очередь создается на этапах паспортного контроля приезжих и на этапе выдачи багажа. Предположим, нашей задачей является максимизация количества обслуженных пассажиров:
  - Максимум {Местных + Приезжих}
  - Местные =< 10, Приезжие =<5 (ограничения по спросу)
  - Местные =< 10, Приезжие =<3 (паспортный контроль)
  - Местные + Приезжие =<10 (выдача багажа)
- Можно обслуживать 7 местных и 3 приезжих в минуту, либо 10 местных и 0 приезжих, либо любую комбинацию из 10 человек в минуту. Следует определить, какие задачи стоят перед системой! (после этого задача решается используя AnyLogic)

Заключение по узким местам



- Главный шаг при определении узких мест сложных процессов – это определение единицы процесса и потом произведение подсчетов с этой единицей
- Определение узкого места – не самая сложная задача в процессах с несколькими типами продуктов: зачастую, чтобы понять, как себя поведет система, требуется поставить перед ней четкие задачи.

А что ещё дальше? Добавляем вариативность

- Подсчет пропускной способности системы, который мы производили сегодня, игнорирует вариативность
- Теория очередей моделирует вариативность, но очень быстро углубляется в математические дебри – трудно проанализировать что-то по-настоящему сложное
- **Если есть нужда смоделировать сложную систему, то предпочтительный подход - это имитационное моделирование!**