# Прелести чужой очереди

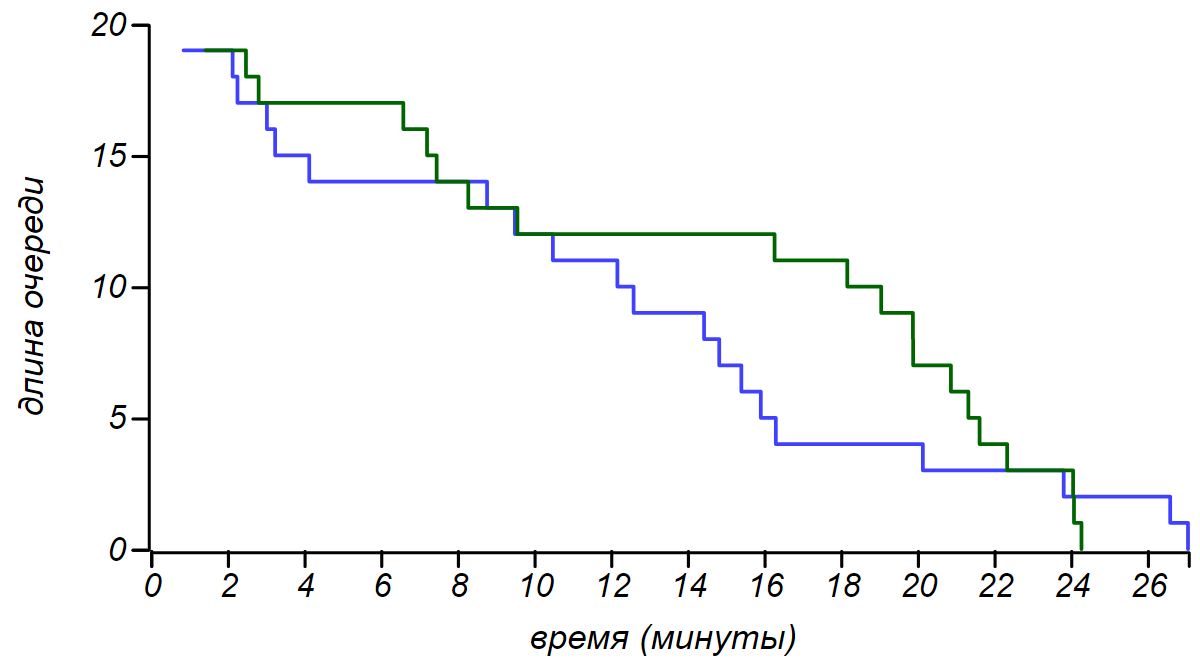
Мы порассуждаем о том как не заскучать, стоя в длинном хвосте у стойки регистрации, о том почему застревают бумаги в кабинете у чиновника и как можно испортить дорожную сеть, расширив её.

Я размышляю о законах подлости, стоя в аэропорту в очереди на регистрацию пассажиров и оформление багажа. Очередь длинная, люди разные и заметные со всеми своими сумками, детьми или клетками.  Сзади слышу ворчание: «Как обычно, наша очередь тормозит. Вон, гляди, тот усатый в кепке наравне с нами стоял а теперь вон где... Вот ведь закон подлости!» Этот закон зовётся наблюдением Этторе:

Соседняя очередь всегда движется быстрее.

Что же это, психологический эффект или причуды математики?

Мы уже достаточно вооружены знаниями, чтобы проанализировать очередь, в которой стоим. За неимением других данных разумно предположить, что выход из очереди происходит по-пуассоновски, то есть через экспоненциально распределённые промежутки времени. Перемещения наблюдателя, стоящего в очереди, будет иметь вид монотонно изменяющейся ступенчатой линии, с одинаковыми шагами, случающимися через случайные промежутки времени. Получаемая таким образом зависимость от времени называется пуассоновскимпроцессом. Его примеры приведены на рисунке.



Перемещения двух очередей, как пуассоновских процессов с равной интенсивностью.

В свою очередь, *разница* двух одинаковых пуассоновских процессов, а именно её наблюдает человек скучающий в хвосте и исследующий соседнюю очередь, представляет собой своеобразное случайное блуждание. А раз так, то мы готовы сделать некоторые качественные выводы. Первый: расстояние между одновременно вставшими в одинаковые очереди людьми будет то увеличиваться, то уменьшаться, при этом будут образовываться характерные меандры с постоянно меняющейся длительностью. Второй вывод: из-за самоподобия случайного блуждания и для коротких очередей и для длинных, меандры будут иметь длительность, соизмеримую со временем стояния в очереди, а значит, они будут заметны, а меандры — это уже повод для недовольства. Третий вывод: заранее неизвестно какая очередь пройдёт быстрее, ведь случайное блуждание равновероятно уходит как вверх, так и вниз. И, наконец, четвёртое заключение: очереди движутся независимо, то и дело опережая и нагоняя друг друга, в среднем, они движутся одинаково, и ожидаемая разница между ними стремится к нулю, но разброс вокруг среднего со временем растёт (в описанном нами случае, величина отставания одной очереди от другой подчиняется распределениюСкеллама). Выходит, или угадал с быстрой очередью или нет — никаких подлых штучек со стороны злодейки судьбы!

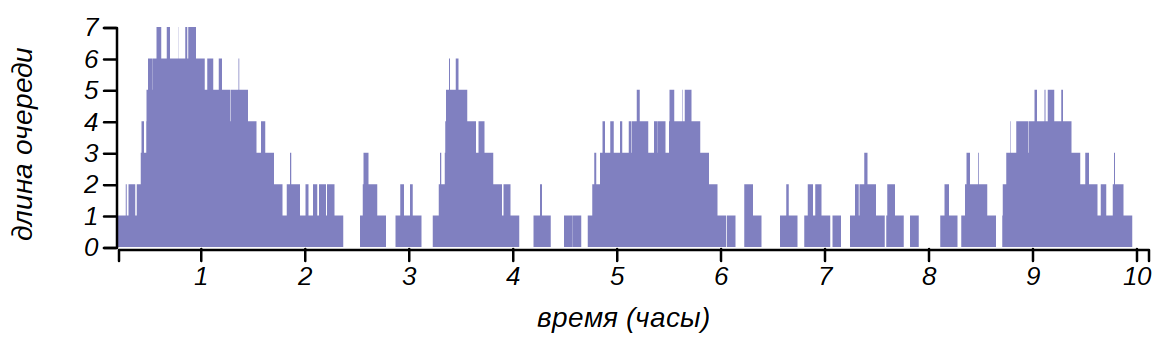
Правда, если нам не повезло оказаться во временно отстающей очереди, то мы в ней проведём больше времени, и, согласно закону велосипедиста, у нас будет больше возможностей посетовать на судьбу! А теперь, внимание, хорошие новости: в любой выбранный интервал времени тех, кому повезёт попасть в быструю очередь, будет больше чем невезунчиков, ведь быстрая очередь может пропустить больше людей! Но, увы, это вовсе не утешит того, кто надолго застрял в хвосте.

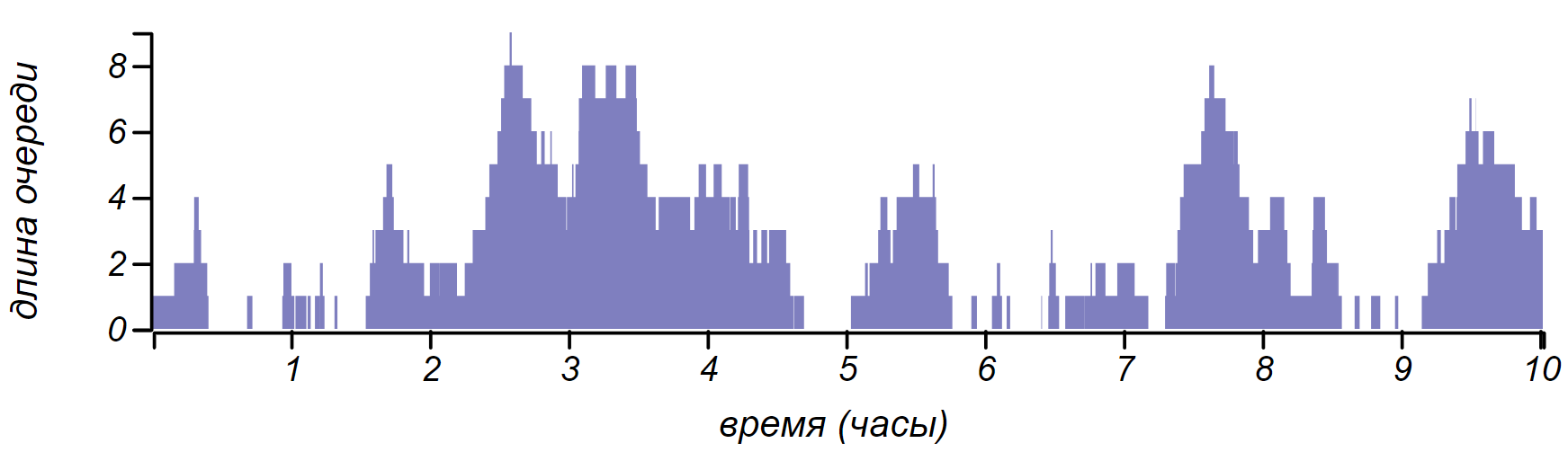
Но тем и хороша математика, что она способна сделать даже стояние в очереди увлекательным процессом. Например, можно прикинуть, сколько ещё предстоит ждать своей очереди, но для этого, как ни странно, надо посмотреть не вперёд, а назад, на растущий хвост. Если подождать какое-то время, скажем, 10 минут и посчитать, сколько человек выстроилось за вами, то разделив количество людей, стоящих перед вами на полученное число вы получите время ожидания в десятках минут. Например, за десять минут хвост за вами вырос на 5 человек, если в момент подсчёта перед вами стоит семь человек, то ожидаемое время ожидания составит . Понятно, что эта оценка будет весьма грубой, но любопытно, что она действительно соответствует среднему времени ожидания, согласно теореме Литтла – одного из общих результатов теории очередей.

## Теория для заскучавших в коридоре

Теория очередей берёт своё начало в самом начале XX века, с первых работ Агнера Эрланга, занимавшегося только зарождающейся областью телекоммуникаций. За сотню лет результаты исследований Эрланга прочно вошли в нашу жизнь, настолько, что возникает ощущение того, что это мы вошли в мир телекоммуникаций. Результаты этой теории важны для проектирования магазинов и залов ожидания, оптимального управления операционной системой в компьютере и операционным залом в банке, для грамотной разработки бюрократической машины, для управления дорожной сетью и в оценке рисков страховой компании.

Отправной точкой для моделирования очереди служит всё тот же пуассоновский поток, поскольку для него требуется минимум дополнительных допущений. Представьте себе очередь, в которую люди встают, согласно некоторому распределению временных интервалов , со средним значением . Время, которое оператор тратит на работу с клиентами подчинено распределению со средним значением . Очередь является стабильной, если , в противном случае, хвост будет расти неограниченно, как пробка на дороге, в которую въезжает больше автомобилей, чем может выехать. От характера распределений и зависит динамика очереди и такие её характеристики как распределения для длины очереди, времени ожидания клиентом и времени занятости оператора. Для очередей создана целая номенклатура, называемая формулой Кэндэлла. Например, простая очередь, в которую люди подходят равномерно и равномерно уходят, как, например, в аэропорту при посадке на рейс, обозначается D/D/1 (D здесь обозначает дельта-распределение, а единица – одного оператора). Въезд и выезд автомашин на территорию аэропорта через три автоматических шлагбаума можно описать очередью M/D/3 – буквой M обозначается пуассоновский (марковский) процесс, то есть случайный процесс без памяти. В очередь на регистрацию билетов и оформление багажа новые люди приходят по-пуассоновски, и багаж у всех разный, так что клиенты будут выходить из очереди тоже по-пуассоновски, для пяти стоек такая очередь обозначается M/M/5. На рисунке показан пример того, как могут «жить» M/D/1 и M/M/1-очередь с и .





Динамика очередей тоже склонна образовывать полосы. Очередь «дышит», то удлинняясь, то сокращаясь.

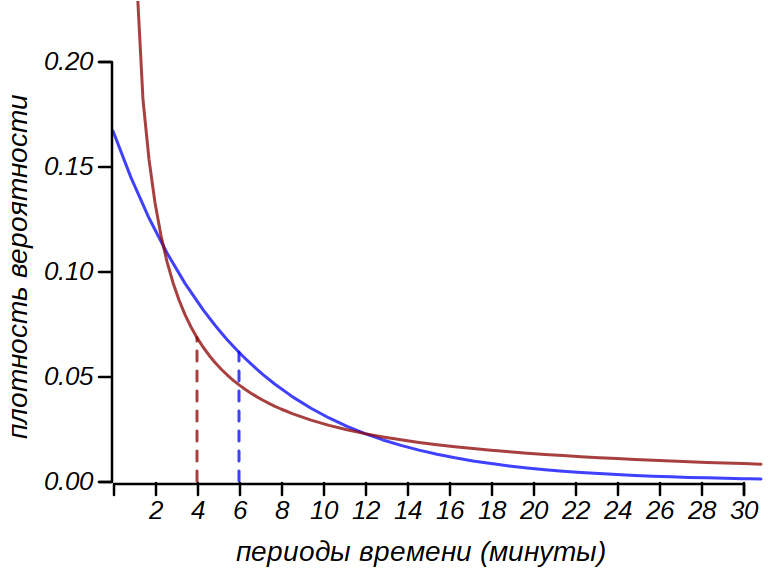
Длина M/M/1-очереди, описывается геометрическим распределением:

мы встречали его в этой главе, рассматривая простейшую несимметричную марковскую цепь. Зная это распределение, можно вычислить ожидаемую длину . Обратите внимание на то, что M/D/1-очередь склонна к образованию несимметричных «горбов» с крутым подьёмом и пологим спуском, тогда как рост и уменьшение M/M/1-очередей вполне симметричны. Если средняя длина очереди зависит от соотношения средних значений для интервалов между приходом новых клиентов и длительности обслуживания, то за симметрию роста и убывания очереди отвечает соотношение их дисперсий. Время обслуживания клиента (то есть сумма времени ожидания своей очереди и, собственно, времени работы с оператором), описывается экспоненциальным распределением с параметром , а значит, среднее время ожидания . Как видно, для стационарной M/M/1-очереди выполняется равенство:

Это и есть формула Литтла, которой мы воспользовались, стоя в очереди и от нечего делать, занявшись подсчётами. Будучи очень простой, эта формула на удивление сильна, она выполняется для очень широкого класса очередей и в самых разных задачах.

Важной характеристикой очереди является время занятости оператора , то есть длительность непрерывных периодов времени, в которые оператор обслуживает клиентов. Такие периоды перемежаются периодами простоя, когда по какой-то причине клиентов в очереди не оказывается. Клиенты приходят, ждут и уходят, а оператор остаётся работать, так что разумно предположить, что . В действительности, как ни странно, ожидаемое, то есть, среднее время занятости для M/M/1-очередей равно среднему времени ожидания, то есть, . Это уже кажется не вполне интуитивным результатом, однако, и это ещё не всё: при той же интенсивности работы оператора, среднее время обслуживания клиента может стать существенно больше среднего времени работы оператора!

Тут всё дело в разбросе данных, то есть, в дисперсии распределения . Для очереди, в которой распределение времени обслуживания клиента оператором уже не экспоненциальное (они обозначаются M/G/1, где G обозначает обобщённое распределение), среднее время ожидания начинает зависеть от дисперсии этого распределения . Ещё в 1930-е годы австрийскому математику Феликсу Поллачеку удалось в общем виде вычислить отношение для произвольной M/G/1-очереди:

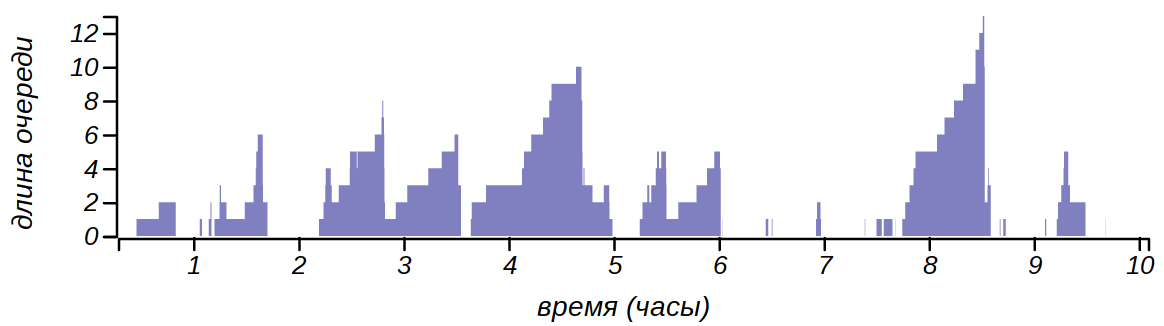


Распределения для периодов времени между появлением новых клиентов (синяя линия, экспоненциальное распределение) и времени обслуживания одного клиента (красная линия, гамма-распределение).

В случае M/M/1-очереди и это отношение равно единице. Но может так случиться, что при том же значении среднего распределение будет иметь большую дисперсию. И тогда может оказаться сколь угодно больше . На рисунке показан пример, в котором экспоненциально с , а описывается гамма-распределением, соответствующим интенсивности , с дисперсией .

Очередь остаётся стабильной: клиенты в среднем обслуживаются быстрее, чем приходят новые. Оператор работает хорошо: большая часть клиентов обслуживается очень быстро, но обратите внимание на то, что велика доля «трудных» клиентов, которые формируют достаточно толстый хвост распределения. Их немного, но каждый из них отнимает много времени и все в очереди вынуждены их ждать. Для примера, приведённого на рисунке, среднее время ожидания оказалось равно 24 минутам, тогда как среднее время занятости оператора составило только 12 минут.

Динамика такой очереди отличается от динамики M/M/1. Для неё характерен несимметричный пилообразный рисунок с плавной восходящей линией и резким сбросом. Пока оператор занят «трудным» клиентом постепенно вырастает длинный хвост, а потом, освободившись, оператор очень быстро с ним справляется.



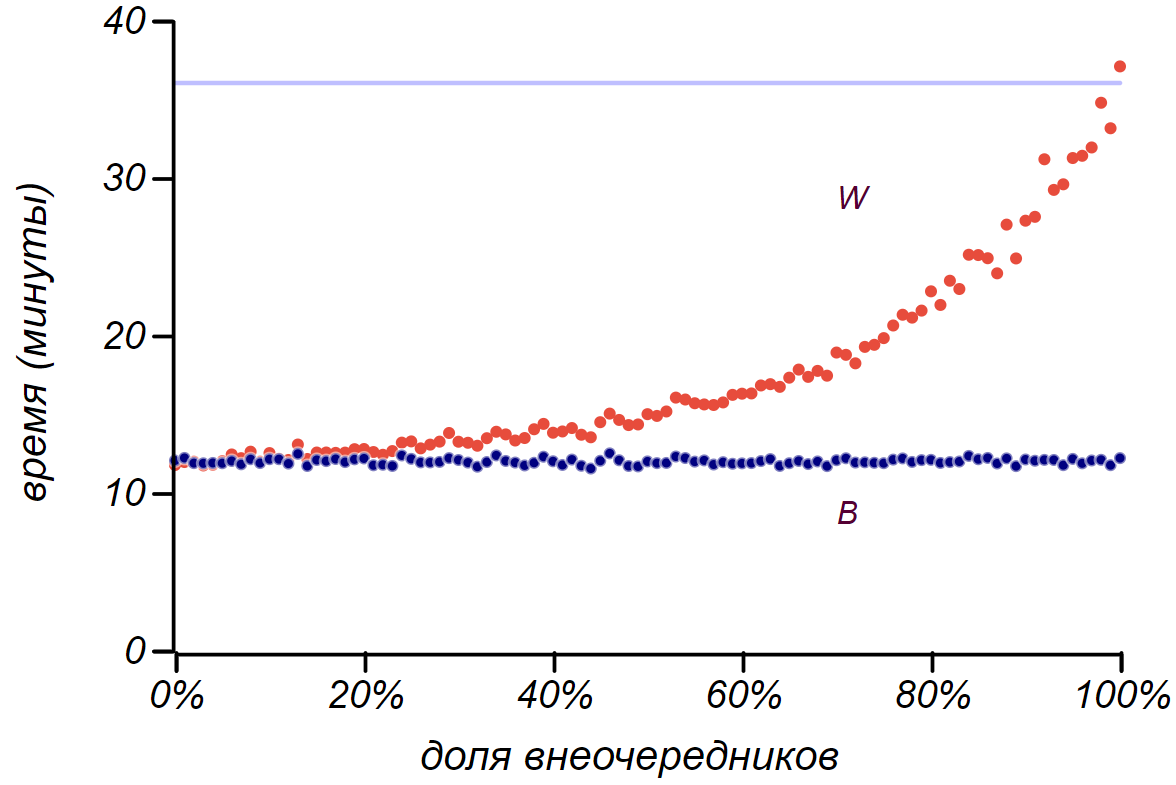
Динамика M/G/1-очереди, в которой время ожидания клиентов вдвое превосходит время занятости оператора.

Получается, что не переставая работать, оператор, в среднем, филонит, пока мы страдаем в очереди от безделья!

## Мне только спросить!

Есть в нашей жизни такое досадное явление – «обочечники» – ушлые водители, объезжающие пробку по обочине и потом встревающие в поток. Есть настырные посетители поликлиник и касс, норовящие просочиться к окошку или двери с заветной формулой «Мне только спросить..». В любую отлаженную бюрократическую систему то и дело врываются неотложные дела, не терпящие промедления. Понятно, что иногда без таких случаев не обойтись: в больницах бывают неотложные пациенты, в операционной системе есть задачи с очень высоким приоритетом, наконец, на дороге мы обязаны пропускать спецтранспорт, едущий по экстренному случаю. Но насколько такие внеочередники опасны всей для очереди? Подобные случаи моделируются очередями с приоритетом и для них тоже есть развитая теория, поскольку в жизни они встречаются чуть ли не чаще простых очередей.

Мы не будем здесь влезать в дебри теории, здесь она уже становится достаточно сложной, а рассмотрим один простой случай. Пусть в нашей M/M/1-очереди с вероятностью могут появляться нетерпеливые клиенты, которые встают не в конец очереди, а в её начало, пододвигая всех стоящих назад. При этом, они, всё же, дают оператору завершить работу с текущим клиентом, не прерывая его. Для очередей с приоритетом тоже выполняется формула Литтла, связывающая длину очереди, время ожидания и интенсивность заполнения очереди. Более того, время занятости оператора оказывается независимым от значения , в отличие от времени ожидания, которое может существенно измениться. К сожалению, простым выражением оно уже не описывается. На рисунке показан результат численного эксперимента, который показывает, как изменяется время ожидания с ростом доли внеочередников для нашего случая с и :



Внеочередники увеличивают время ожидания для всех клиентов,  
однако, не мешают оператору

Если доля внеочередников не превышает трети, очередь этого почти не заметит, но когда их становится больше половины, время ожидания уже вырастает существенно. В предельном случае наша очередь превращается в то, что программисты называют стеком – последовательность элементов, которая подчиняется правилу «*первым вошёл, последним вышел*» в противовес очереди, для которой выполняется правило «*первым вошёл, первым вышел*». Время ожидания в этом случае вырастает втрое по сравнению с очередью стоящей «на общих основаниях». Для произвольных и , время ожидания при становится равно и если лишь немного превышает , очередь остаётся устойчивой, но время ожидания в ней вырастает катастрофически!

Очередь из одних только внеочередников выглядит неестественно, но давайте вместо людей рассмотрим пачку документов, тогда мы можем узнать знакомую картину на рабочем столе. Если входящие документы не сортируются по времени, а просто складываются в стопку по мере поступления, а потом обрабатываются, начиная сверху, то получается та самая перевёрнутая очередь, то есть, стек. При этом оператор, то есть, бюрократ, обрабатывающий бумаги может ничего не заметить, его время занятости не изменится. Начальник бюрократа тоже увидит, что из кабинета подчинённого бумаги выходят с нормальной интенсивностью. И только сами документы, или их авторы почувствуют, как время их обработки нещадно вырастает.

Эта же картина наблюдается и в шкафу, в который мы складываем вещи с мыслью разобрать потом. Но потом, мы задвигаем то, что уже лежит в шкафу поглубже и добавляем в него новые вещи. Так что даже если мы, всё же станем их постепенно разбирать, до «ископаемых» у самой стенки руки дойдут нескоро.

Через любую стационарную очередь клиенты проходят с интенсивностью , то есть, не накапливаются в ней, собственно, это и есть условие стационарности. Таким образом, если, стремясь оптимизировать работу учреждения, начальство снизит число кассиров или операторов до минимально необходимого, ориентируясь на входящий или выходящий поток клиентов, то рискует получить , а мы знаем, что . Таким образом, это разумное, как кажется, решение приведёт к неограниченному и плохо контролируемому росту времени ожидания. Метастабильная очередь может долгое время вести себя «хорошо», а потом внезапно превратиться в коллапс.

## Хотели как лучше

Наконец, говоря об очередях, нельзя не упомянуть и совершенно возмутительный парадокс Браэса, приводящий к тому, что в коммуникационной сети, содержащей очереди добавление новых простых связей (не стохастических) может привести к уменьшению пропускной способности всей сети. Этот парадокс получил достаточно широкую известность тогда, когда… позже…