# Прелести чужой очереди

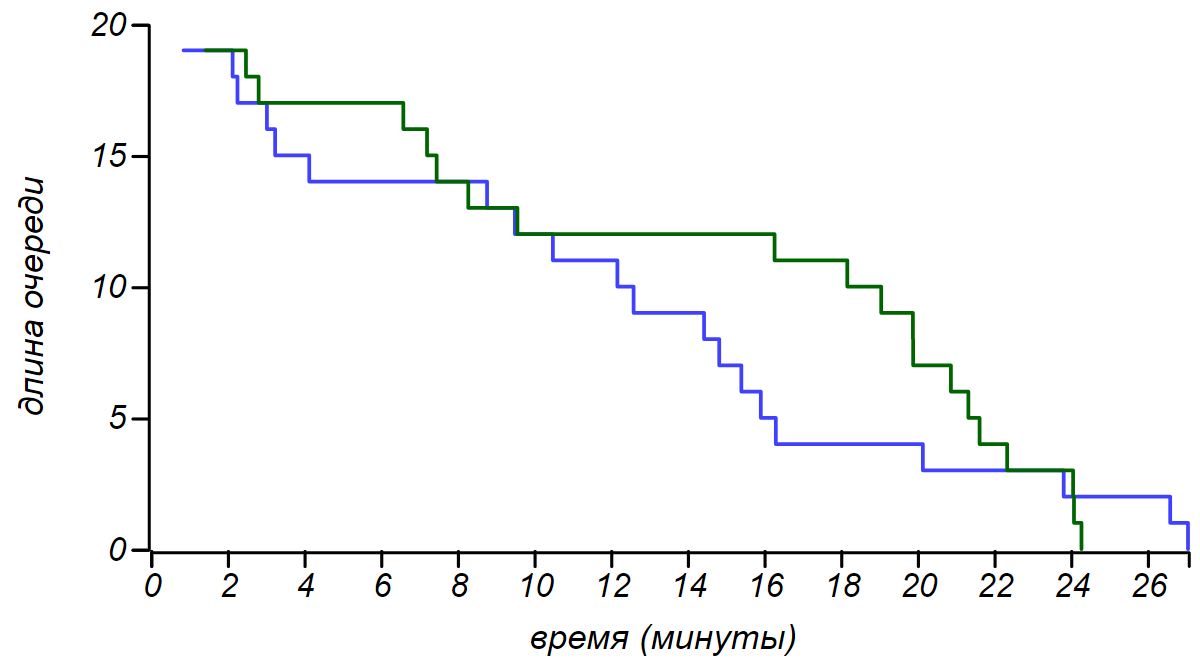
В этой главе мы нацчимся не заскучать, стоя в длинном хвосте у стойки регистрации, узнаем о том почему застревают бумаги в кабинете у чиновника и как можно испортить дорожную сеть, расширив её.

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 немного 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. Такая зависимость от времени называется пуассоновскимпроцессом. Пара примеров пуассоновских процессов приведена 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. Ну, а раз так, мы уже 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.

## Теория для заскучавших в коридоре

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 среднее время ожидания в десятках минут. Например, пусть за десять минут хвост 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, об этом говорит теорема Литтла – один из самых ранних и самых общих результатов теории очередей.

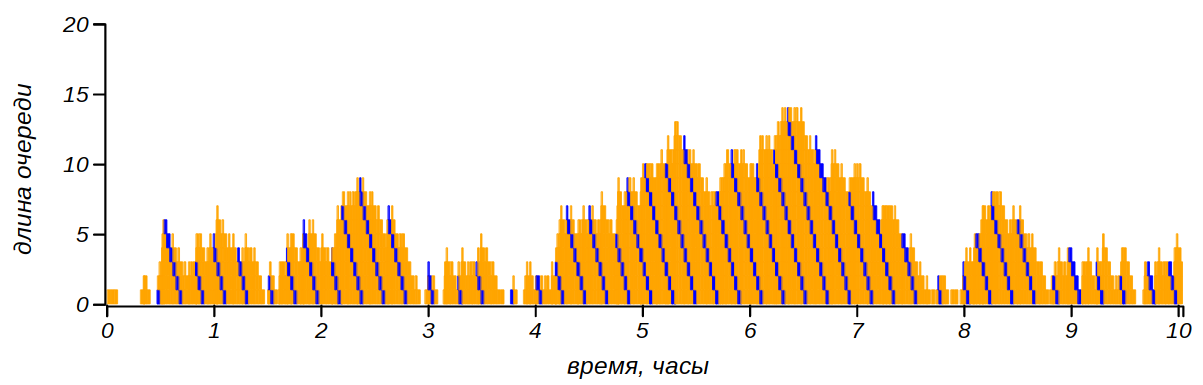
Теория очередей берёт своё начало в самом начале XX века, с первых работ Агнера Эрланга, который занимался только-только зарождающейся областью телекоммуникаций. За сотню лет результаты исследований Эрланга прочно вошли в нашу жизнь, настолько, что возникает ощущение того, что это мы вошли в мир телекоммуникаций. 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. В очередях могут стоять люди (покупатели, клиенты, пассажиры), автотранспорт и грузы, задачи и документы; а обрабатывать их могут кассиры, операторы, регистраторы, серверы и бюрократы. Чтобы не путаться и не утопать в деталях, мы будем называть стоящих в очереди клиентами, а того, кто их обслуживает – оператором.

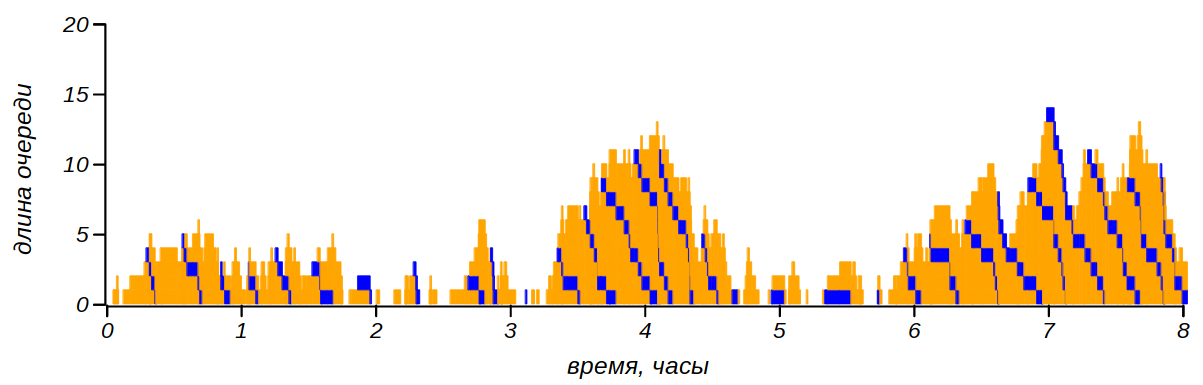
Представьте себе очередь, в которую люди встают, согласно некоторому распределению временных интервалов , со средним значением . Время, которое оператор тратит на работу с клиентами подчинено распределению со средним значением . Такую очередь описывают как марковский процесс с *бесконечным числом состояний*. Каждое состояние соответствует некоторой длине очереди: состояние 0 — в очереди никого, состояние 1 — один клиент, состояние 2 — два клиента и так далее… и, в принципе, ничто не запрещает очереди вырасти сколь угодно длинной. Таким образом, для анализа очереди придётся иметь дело с матрицей переходов, содержащей бесконечное число строк и столбцов. В предыдущей главе мы уже имели дело с марковскими процессами, и для анализа стационарного состояния цепи нам понадобилось возводить матрицу переходов в бесконечную степень. Так что же надо будет вычислить бесконечную матрицу, возведённую в бесконечную степень? Математиков эта задача не испугала и уже в тридцатые годы XX века были придуманы методы для таких вычислений. Результатом этого анализа будут свойства стационарного состояния очереди. Это состояние не меняется со временем, но все параметры очереди, такие как её длина или время ожидания в ней представляют собой случайные величины. То есть, они сами по себе постоянно меняются, но при этом описываются какими-то распределениями, независимыми от времени. И чего только не придумаешь, скучая в зале ожидания!

Свойства очереди сильно зависят от соотношения и . Если , то хвост будет расти неограниченно, как пробка на дороге, в которую въезжает больше автомобилей, чем может выехать. Такая очередь перекрывает поток клиентов. Для очередь является стабильной. Она может расти или уменьшаться, но клиенты в ней не исчезают и не появляются, а значит, сколько их вошло в зону ожидания, столько и выйдет. То есть, стабильная очередь может затормозить тех, кто в ней стоит, но она не может изменить поток людей, проходящих сквозь неё. И если на входе мы имеем человек в единицу времени, то и на выходе мы должны получить такой же поток, независимо от скорости работы оператора. Случай рассматривается отдельно, такая метастабильная очередь ведёт себя неустойчиво и моделируется даже не марковским процессом, а процессом случайного блуждания или диффузии.

Далее мы будем рассматривать только стабильные очереди. От характера распределений и зависит динамика очереди и такие её характеристики как распределения для её длины, 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.

В этой главе мы будем исследовать неприятности и неожиданности, наблюдаемые в очредях на примере очереди с и . В среднем, новые клиенты будут поступать в неё с интервалом в мин, а обрабатываться оператором за минуту секунд. На рисунке показан пример того, как могут «жить» M/D/1 и M/M/1-очередь с такими параметрами.





Динамика MD1 и MM1 очередей. Синим цветом выделены траектории каждого седьмого клиента в очереди. Длина очереди склонна к своеобразным колебаниям, она «дышит», то удлинняясь, то сокращаясь, оставаясь при этом стационарной.

В стационарном, состоянии 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 . Для нашего примера средняя длина очереди составит человек. 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 . Не смотря на то, что среднее время работы с каждым клиентом не превышает двух минут, среднее время ожидания равно 15 минутам. Как видно, для стационарной M/M/1-очереди выполняется равенство:

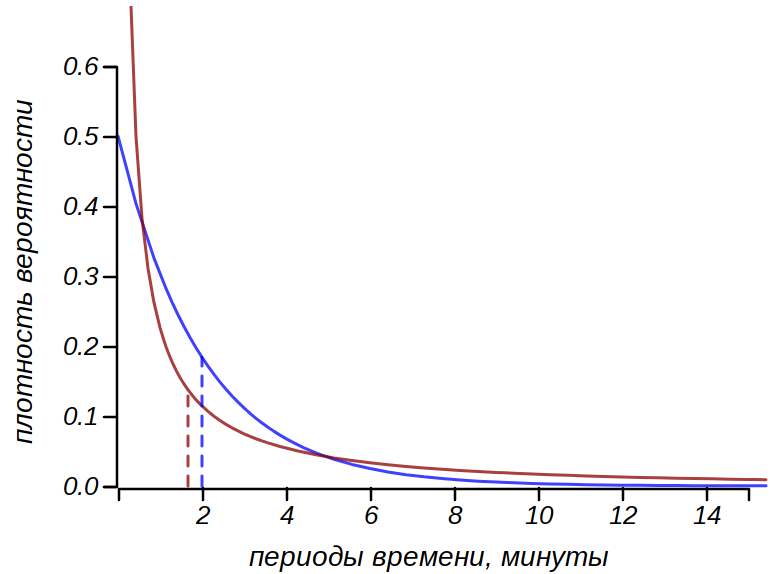
Это и есть формула Литтла, которой мы воспользовались, стоя в очереди и от нечего делать, занявшись подсчётами. Будучи очень простой, эта формула на удивление сильна, она выполняется для очень широкого класса очередей и в самых разных задачах. То что в формулу Литтла входит только , а не отражает основное свойство стабильной очереди, она может задерживать клиентов, но не меняет их поток, который остаётся равным . И даже если скорость работы оператора будет очень велика, среднее время ожидания будет всё равно определяться входным потоком. А так как , то мы получаем ещё один закон подлости:

Даже с идеальным кассиром, время стояния в очереди на кассу определяется неспешными покупателями.

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:

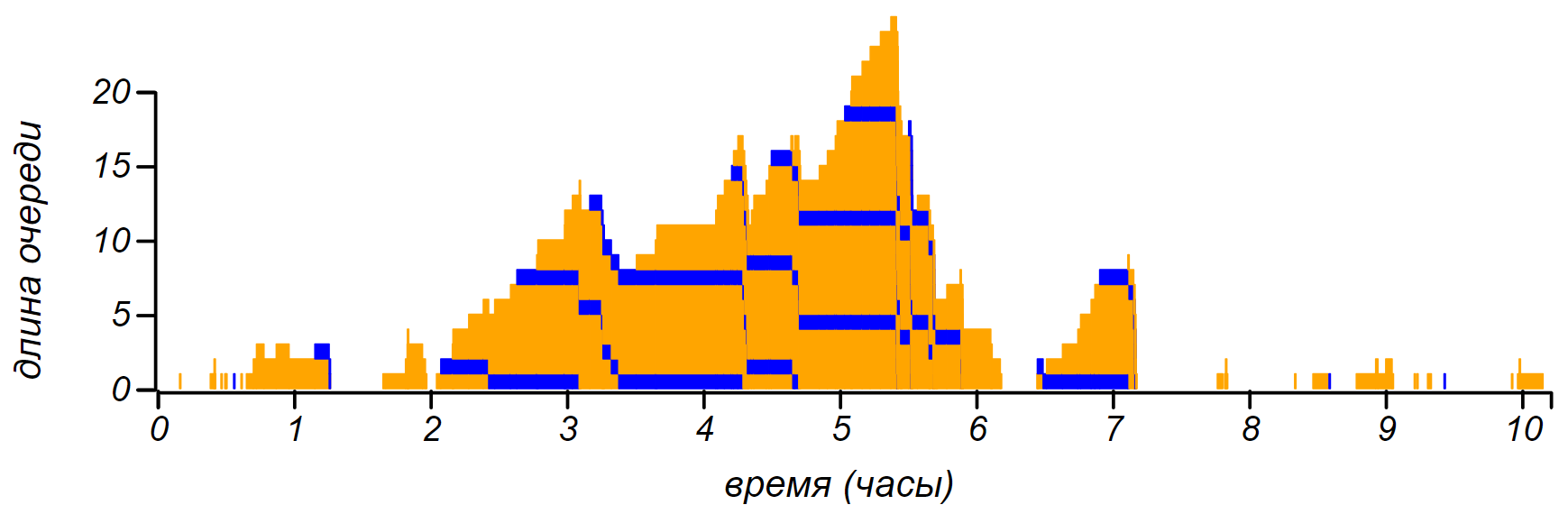
В случае M/M/1-очереди и это отношение равно единице. Но может так случиться, что при том же значении среднего распределение будет иметь большую дисперсию, и тогда может оказаться сколь угодно больше . На рисунке показан пример, в котором экспоненциально с , а описывается гамма-распределением, соответствующим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



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

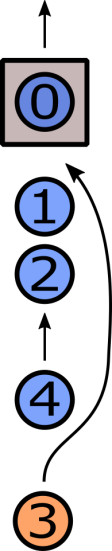
Очередь остаётся стабильной: клиенты в среднем обслуживаются быстрее, чем приходят новые. Оператор работает хорошо: большая часть клиентов обслуживается очень быстро, но обратите внимание на то, что велика доля «трудных» клиентов, которые формируют достаточно толстый хвост распределения. Их немного, но каждый из них отнимает много времени и все в очереди вынуждены их ждать. Для примера, приведённого на рисунке, среднее время ожидания оказалось равно 35 минутам, тогда как среднее время занятости оператора осталось прежним (15 минут). Получается, что не переставая работать, оператор, в среднем, филонит, пока мы страдаем в очереди от безделья!

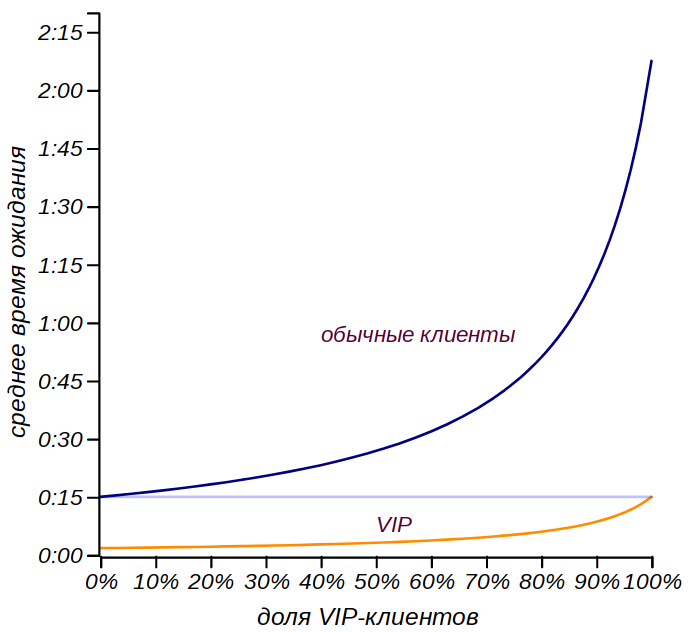
Динамика такой очереди отличается от динамики M/M/1. Для неё характерен несимметричный пилообразный рисунок с плавной восходящей линией и резким сбросом. Пока оператор занят «трудным» клиентом постепенно вырастает длинный хвост, а потом, освободившись, оператор очень быстро с ним справляется.



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

## 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 компьютера есть задачи с очень высоким приоритетом, наконец, на дороге мы обязаны пропускать спецтранспорт, едущий по экстренному случаю. Но как такие внеочередники влияют на всю очередь? Подобные случаи моделируются очередями с приоритетом и для них тоже есть развитая теория, поскольку в жизни они встречаются чуть ли не чаще простых очередей.

Пусть в нашей M/M/1-очереди с вероятностью могут появляться особые клиенты, назовём их VIP (very impatient person – очень нетерпеливые персоны), которые встают не в конец очереди, а вклиниваются в её начало, заставляя ждать всех стоящих позади. При этом, они, всё же, дают оператору завершить работу с текущим клиентом, не прерывая его. Если внеочередников наберётся несколько, они могут образовать свою VIP-очередь. Если вспомнить, что пуассоновский поток можно представить как случайное «расбрасывание» по временному интервалу какого-то известного количества событий, то по нашему условию, мы получим поток нетерпеливых клиентов и поток обычных клиентов , при этом, общий поток останется неизменным. Среднее время ожидания для VIP будет равно , как в любой M/M/1-очереди, поскольку они в своей VIP-очереди «не замечают» присутствия обычных клиентов. Для тех, что ждёт на общих основаниях, время ожидания вырастет, и составит уже:

Соотношение средних времён ожидания   
для очереди

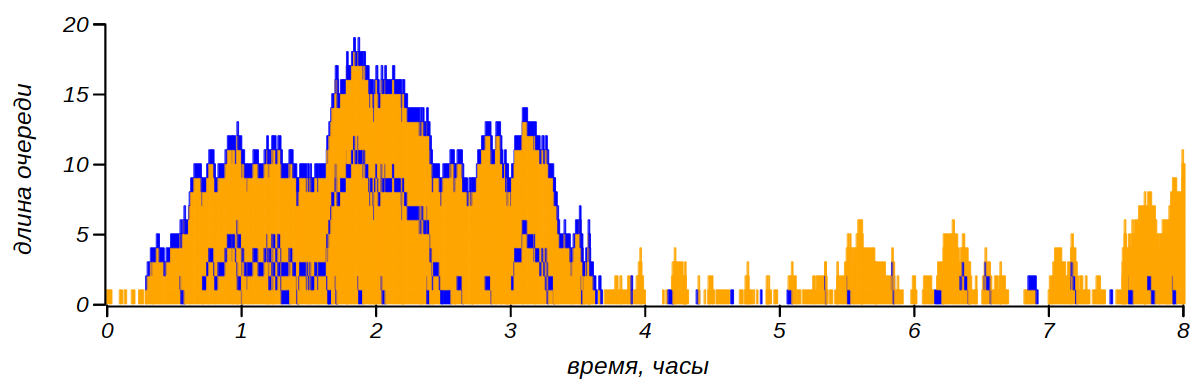
Пока VIP-ов немного, очередь их не замечает, он если доля внеочередников оказывается близкой к единице, то никакого преимущества они уже не имеют, зато немногочисленным скромным очередникам приходится ждать существенно дольше. Для произвольных и , время ожидания при становится равно (больше двух часов в нашем случае!) и, вообще, если лишь немного превышает , очередь остаётся устойчивой, но время ожидания в ней вырастает катастрофически!

Но вот, что любопытно. Можно найти среднее время ожидания для всех клиентов, как взвешенную сумму и она окажется равной , то есть, такой же, как для обыкновенной M/M/1-очереди без всяких VIP-ов. Выходит, системе в целом внеочередники не мешают. На время занятости оператора они тоже не влияют и распределения времён ожидания остаётся экспоненциальным. Мы уже говорили в предыдущей главе, что для экспоненциального распределения кривая Лоренца и, соответственно, коэффциент Джини не зависят от параметра распределения. А значит, все M/M/1-очереди имеют одинаковую степень несправедливости – . Отсюда следует, что наш обобщённый критерий несправедливости для всех ожидающих в очереди также останется равным .

## C:\tmp\podlost\ToH\work\figures\queue\g8140.pngСтационарный бардак

А теперь немного изменим политику очерёдности. Пусть внеочередники будут сверхнаглыми, и если так случится, что один такой клиент придёт вслед за другим, то вместо формирования нормальной очереди, второй вклинится перед первым. Эта задача уже отличается от классического подхода к очередям с приоритетом. Давайте сразу рассмотрим предельный случай, когда доля наглых клиентов равна единице. Тогда наша очередь превращается в то, что программисты называют стеком – в последовательность элементов, подчиняющуюся правилу «*первым вошёл, последним вышел*» (FIFO – first in, first out) в противовес очереди, для которой выполняется правило «*первым вошёл, первым вышел*» (FILO – first in, last out).

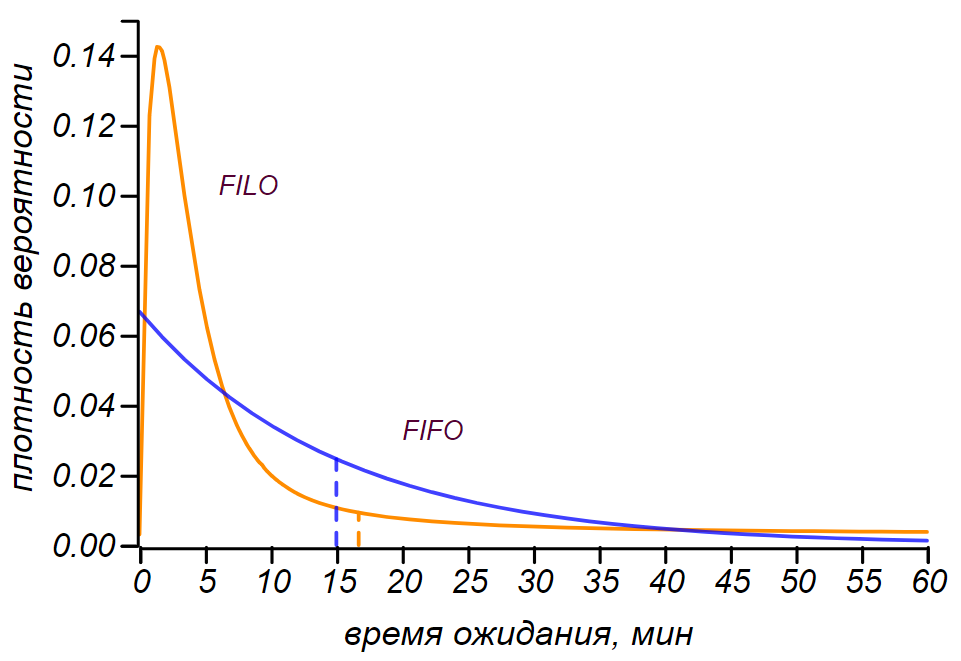
Такая «очередь наоборот» выглядит неестественно, но давайте вместо людей рассмотрим пачку документов, тогда мы можем узнать знакомую картину на рабочем столе. Если входящие документы не сортируются по времени, а просто складываются в стопку по мере поступления, а потом обрабатываются, начиная сверху, то получается та самая очередь FILO, то есть, стек. Удивительно, но в стационарном состоянии и среднее время ожидания и среднее время занятости оператора будут точно такими же, как и в FIFO-очереди. Что же поменяется? Давайте посмотрим на динамику такой «очереди»:



Динамика «очереди наоборот», то есть стека, или стопки документов, которые поступая, кладутся наверх и обрабытываются, начиная сверху.

Мы видим, что вместо целенаправленного движения к оператору, клиенты могут двигаться, то к нему, то от него. Время ожидания для самого последнего клиента существенно удлиняется, однако пока он ждёт, через оператора проходит большое число вновь поступающих клиентов, которые обрабатываются почти мгновенно. В среднем же, мы получаем время ожидания, примерно такое же, как для «нормальной» очереди. Но мы уже много раз убеждались в том, что среднее значение не может характеризовать случайную величину в полной мере.

Глядя на динамику очереди легко понять, что время ожидания клиента W должно быть близким к времени занятости оператора B. Время занятости определяется как период от момента прихода первого клиента до момента выхода последнего, но в стеке первый клиент и является последним. Нужно ещё учесть, что очередной клиент не прерывает оператора, и поэтому к времени его ожидания добавится время обслуживания клиента, с который уже работает оператор. Если это время распределено экспоненциально, то, как уже обсуждалось со временем ожидания автобуса, и добавочное время будет распределено точно также. В конечном итоге, время ожидания будет распределено как сумма времени занятости оператора и периода работы с одним клиентом. На рисунке показаны распределения времени ожидания для M/M/1-очередей, обыкновенной -- с политикой FIFO и придерживающейся правила FILO. В обоих случаях и человека в час.



Распределения времени ожидания для M/M/1-очередей с различной политикой.

Распределения сильно отличаются, но средние значения у них практически одинаковые! Несмотря на то, что распределение времени ожидания для FILO-очереди кажется сконцентрированным около моды (близкой к , то есть, времени работы с одим клиентом) у него очень длинный тяжёлый хвост, который сильно увеличивает дисперсию и увеличивает среднее значение. Медиана этого распределения равна , это значит, что в половине случаев клиент будет ждать не более минут, но весьма велики и шансы застрять очень надолго.

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 внезапно 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.

Поведение очереди на пределе её пропускной способности моделируется случайным блужданием, подобно тому, как мы моделировали перепады настроения. Это блуждание выражается в возможности практически неограниченного роста длины очереди, притом, что рано или поздно, она сократится и даже исчезнет. Но при этом отклонение времени оджидания и времени работы оператора от седнего будут столь велики, что динамика очереди станет практически непредсказуемой.