Информационные процессы, Том 7, № 3, 2007, стр. 248–322.

МАТЕМАТИЧЕСКИЕ МОДЕЛИ

РАЗДЕЛЕНИЕ ПРОЦЕССОРА: ОБЗОР МАТЕМАТИЧЕСКОЙ ТЕОРИИ [1]

С.Ф.Яшков*, А.С.Яшкова**

**Институт проблем передачи информации им. А.А.Харкевича, Российская академия наук*
19, Большой Каретный переулок, 127994 Москва ГСП–4, Россия.
E-mail: yashkov@iitp.ru
***Муниципальный институт г.Жуковского,*
кафедра прикладной информатики и программирования,
15, улица Маяковского, 140180 г.Жуковский, Московской обл., Россия.
E-mail: annpost1@mail.ru

Поступила в редколлегию 20.04.2007

Аннотация—За последнее десятилетие эгалитарное разделение процессора (EPS) стало играть заметную роль в прикладной теории вероятностей, в особенности в теории очередей и ее компьютерных приложениях. В то время как парадигма EPS возникла в 1967 как идеализация циклического (round–robin: RR) алгоритма диспетчеризации в вычислительных системах с разделением времени, недавно возобновился интерес к ней как к мощной концепции для моделирования WEB серверов. Статья суммирует важнейшие результаты по точным решениям системы обслуживания M/GI/1 с эгалитарным разделением процессора. Материал почерпнут, в основном, из статей авторов, которые дополнены в небольшой степени другими смежными результатами. Многие дальнейшие результаты установлены под непосредственным влиянием наших более ранних статей. Главной целью является дать обзор состояния и достижений современной теории системы обслуживания M/GI/1—EPS. Упор сделан на методах точного и асимптотического анализа этой системы обслуживания. В отличие от стандартных обзоров, в статью включены сокращенные доказательства (или идеи доказательств) ключевых теорем и следствий. Мы очертим недавние достижения в точных решениях системы M/GI/1—EPS по стационарным и нестационарным (транзиентным) распределениям основных вероятностно–временных характеристик. В частности, данная статья включает результаты по совместному нестационарному распределению времени пребывания в системе требования, поступающего в момент t с требуемой длительностью обслуживания (длиной) u, и числа требований в системе M/GI/1—EPS в момент $t-$, полученные в терминах многомерных преобразований. Демонстрируется также, как могут быть использованы нестационарные решения для получения известных и новых результатов, позволяющих предсказать поведение системы обслуживания и обнаружить ее новые неожиданные свойства. Мы также обсудим ряд предельных теорем, возникающих при исследовании очередей с разделением процессора.

ОГЛАВЛЕНИЕ

[1] Работа частично поддержана Программой фундаментальных исследований ОИТВС РАН, проект 4.6.

1. ВВЕДЕНИЕ

1.1. Предварительные замечания

За последнее десятилетие эгалитарное разделение процессора (Egalitarian Processor Sharing: EPS) стало играть заметную роль в прикладной теории вероятностей, в особенности в теории очередей и ее компьютерных приложениях. В то время как парадигма EPS возникла в 1967 [107] как идеализация циклического (Round–Robin: RR) алгоритма диспетчеризации в вычислительных системах с разделением времени [10,109,120] , недавно возобновился интерес к ней как к мощной концепции для моделирования WEB серверов [131,180,183]. В отличие от стандартных моделей очередей крайне важной проблемой в теории массового обслуживания состоит в нахождении точного решения для стационарного времени пребывания требования в системе обслуживания M/GI/1 с дисциплиной эгалитарного (справедливого) разделения процессора[2] Долгое время эта проблема оставалась открытой, впервые полное ее решение было получено Яшковым (Yashkov) с помощью предложенного им нового аналитического метода (первоначально для системы M/M/1 с дисциплиной кругового циклического доступа RR [197] (1977)), адаптированного позднее для случая системы M/GI/1 с эгалитарным разделением процессора[3] (см., например, [104] (1978), [199] (1981), [201] (1983), а также [206], [212] для необходимых ссылок). Теперь многие авторы считают это одним из наиболее значительных достижений теории очередей. Метод решения, улучшенный немного позже, оказался весьма плодотворным для получения дальнейших результатов, например, для нестационарного распределения числа требований в рассматриваемой и родственной моделях [207] (1988), [208] (1989), [211] (1991), [212] (в основном потоке западной литературы по теории очередей до конца девяностых годов эти задачи считались неразрешимыми с аналитической точки зрения). Новый подход к анализу системы M/GI/1—EPS был назван “методом декомпозиции на элементы задержек” в [201, p. 134].

Отметим, что предположения, которые требуются для использования стационарных решений любой системы массового обслуживания, редко удовлетворяются в реальной жизни. Для реального применения теории очередей при проектировании и анализе технических систем использование только стационарных результатов теории во многих случаях явно недостаточно. Например, часто бывает необходимо исследовать поведение системы массового обслуживания,

[2] Когда система содержит $n > 0$ требований, то каждое из них обслуживается со скоростью $1/n$, т.е. все они обслуживаются одновременно. Первоначальная идея EPS принадлежит Клейнроку (Kleinrock) [107] (1967).

[3] Система обслуживания EPS может рассматриваться как непрерывный предел системы round–robin (подробности см. в третьем абзаце раздела 2.1).

приближающееся к стационарному с ростом времени (если и когда стационарное состояние существует). При этом даже средняя длина очереди в момент t дает намного больше информации по сравнению со стационарным средним числом требований. Однако известно мало стохастических систем, для которых получены нестационарные (транзиентные) решения по вероятностным распределениям процессов обслуживания. Как правило, к таким системам относятся очереди M/GI/1 с классическим дисциплинами обслуживания (например, FCFS (первым пришел — первым обслужен), а также с относительными или абсолютными приоритетами, см. Джейсуол (Jaiswal) [86], Прабху (Prabhu) [144], Такач (Takács) [175,176], а также Асмуссен (Asmussen) [15], Бочаров и Печинкин [32], Гнеденко и Коваленко [73], Коэн (Cohen) [46], Купер (Cooper) [50], Кокс и Смит (Cox and Smith) [52]). (Книга Джейсуола [86], например, является одной из редких великолепных монографий, которая почти полностью посвящена проблематике нестационарных приоритетных систем обслуживания.) Кроме того, все нестационарные решения для очередей типа M/GI/1 получены в терминах двойных преобразований (по пространству и времени), из которых весьма затруднительно извлечь необходимую информацию относительно поведения системы обслуживания. (Подчеркнем дополнительно, что для получения нестационарных решений необходима более продвинутая математическая техника по сравнению с анализом системы в стационарном режиме.) Впрочем, некоторые исключения составляют варианты системы M/M/1, для которых известны транзиентные решения в явном виде. Как правило, точные результаты анализа в переходном режиме включают бесконечные суммы бесселевых функций (см., например, книги Асмуссена (Asmussen) [15], Коэна (Cohen) [46], Кокса и Смита (Cox and Smith) [52], Прабху (Prabhu) [144], Саати (Saaty) [154] или обзор Трипати (Tripathi) и Дуда (Duda) [179], а также Бхат (Bhat) [26] и [21,111].). Впервые такое решение было получено Бейли (Baily) [21], который использовал производящие функции в дифференциальных уравнениях, описывающих процесс числа требований. В более общих ситуациях получение точных решений в явном виде маловероятно. Поэтому были предложены различные численные и асимптотические методы, а также аппроксимации для вычисления вероятностей состояний и некоторых показателей производительности простых систем обслуживания. Однако в данной статье мы не предполагаем обсуждать роль аппроксимаций при анализе очередей.

Прежде всего, эта статья посвящена точным решениям для вероятностных распределений основных характеристик системы M/GI/1—EPS. Полученные сначала для стационарного режима, эти решения в последнем десятилетии удалось распространить и на нестационарный случай. Иными словами, были найдены точные транзиентные решения в терминах многомерных преобразований. (В общем, такие решения могут рассматриваться как явные, по крайней мере в принципе, если мы считаем решение в терминах преобразований явным.) Эти решения развиты для получения совместного нестационарного распределения виртуального времени пребывания в системе требования, поступающего в момент t и имеющего в этот момент требуемую длительность обслуживания (длину) u, и числа требований в момент $t-$ в системе обслуживания M/GI/1—EPS (маргинальное нестационарное распределение времени пребывания требования является условным распределением по u).

Основные цели статьи состоит в более детальном по сравнению с [212] объяснении метода получения указанных результатов и обзоре точных аналитических решений, в особенности, полученных в последнем десятилетии. Асимптотическим и, в особенности, алгоритмическим решениям уделяется меньше внимания. Тем не менее, мы обсуждаем некоторые асимптотические решения, хотя это не первоочередная цель. В наши задачи не входит рассмотрение алгоритмических решений. Оставлены в стороне и точные решения для системы M/GI/1 с

другой популярной дисциплиной разделения процессора FBPS [4], несмотря на то, что эта модель исследована почти столь же глубоко как M/GI/1—EPS[5] в отличие от других вариантов дисциплин разделения процессора. (Относительно достижений в анализе модели FBPS см., например, [202] (1984), [206], [158] (1988), [208] (1989), [212] (1990), [32] (1995), [225] (2004) и [14] (2005)[6].) Мы продемонстрируем также как большинство (если не все) известных или новых аналитических решений системы M/GI/1—EPS могут быть получены в качестве частных случаев, используя стандартные или нестандартные аргументы (например, с помощью абелевых/тауберовых (Abelian/Tauberian) теорем). В сущности, наши конструкции восходят к аналитическому методу из статей [197], [199], [201], детально отраженному и в монографическом виде [208] (1989). Основные результаты [208] доступны в виде обширного обзора [212] (1992) и для англоязычных читателей[7]. Статья продолжает, частично дополняет и углубляет обзоры [206] (отразивший достижения по 1987 г.), [212] (отразивший достижения по 1990 г.), но она отнюдь не претендует на полноту по сравнению, например, с обзорной статьей [212][8]. Еще одна сверхзадача обзора состоит в том, чтобы обеспечить руководство по математической теории системы M/GI/1 с эгалитарным разделением процессора (и ее методах), развитой за последние десятилетия. Стоит отметить, что этот обзор довольно нестандартен в том, что упор сделан не на сборе результатов. а на прояснении тех аспектов теории системы M/GI/1—EPS, которые делают удобным и естественным использование новых аналитических методов, разработанных в наших статьях.

За последнее десятилетие объем литературы по математической теории систем обслуживания с разделением процессора процессора резко увеличился. Общее число статей составляет около тысячи (если учитывать и работы прикладного характера). Поэтому практически невозможно отразить в одной статье некоторые результаты. Кроме того, среди них имеется ряд тривиальных или заимствованных результатов. Мы пытались сократить до минимума число ссылок на такие работы, но не смогли полностью избежать ссылок на работы такого сорта. Однако надеемся, что большинство интересных результатов отражено в этом обзоре. Впрочем, недавно появился ряд неполных обзорных статей. Некоторые из них будут упомянуты далее. Это позволит сконцентрировать усилия только на сформулированных целях данной обзорной

[4] FBPS — аббревиатура термина Foreground Background Processor Sharing. Эта дисциплина осуществляет следующее приоритетное правило, скомбинированное с EPS: высший абсолютный приоритет всегда предоставляется требованию (требованиям), получившему минимальную величину времени обслуживания. Если имеется $n > 0$ таких требований, то каждое из них обслуживается со скоростью $1/n$, т.е. эти требования делят ресурс сервера согласно EPS. Однако, ресурс прибора разделяется только самыми молодыми (недавно вошедшими в систему) требованиями. В отсутствии информации относительно длин (времени обслуживания) требований, дисциплина FBPS пытается выполнить короткие требования с максимально возможной быстротой посредством смещения в сторону требований с минимальным возрастом. О различных вариантах дисциплин разделения процессора см. [87, 206, 208, 212], здесь мы не останавливаемся на этом.

[5] Основной вклад в глубокое изучение обоих моделей EPS и FBPS сделан российской вероятностной школой к началу восьмидесятых годов. В монографии [208] и озоре [212] отражены главные достижения.

[6] Кроме того, в дополнение к [206, 212], в недавних статьях [1, 2] обнаружены новые факты относительно классов распределений, при которых дисциплина FBPS является неоптимальной.

[7] Однако представляется, что перевод на английский язык указанной обзорной статьи не вполне удовлетворителен с терминологической точки зрения, поскольку неизвестный переводчик, по–видимому, не является экспертом в теории очередей.

[8] Отметим также более ранние обзоры по этой проблематике и смежной тематике, принадлежащие Мак Кинни (McKinney) [122] (1969), Клейнроку (Kleinrock) (1970), (1972) (включенные позднее в его монографию [109] (1976)), Кобаяши и Конхейму (Kobayashi and Konheim) [112] (1977), Джейсуолу (Jaiswal) [87] (1982), Митрани (Mitrani) [126] (1986), Нейману [134] (1978) и Соловьеву [167] (1981). К этим работам примыкают, в частности, статья Варда и Витта (Ward and Whitt) [187] (2000) и диссертация ван ден Берга, в которой даны дополнительные комментарии к аналитическому методу декомпозиции на элементы задержек, введенному в [104, 201, 203, 208]. Некоторые частные вопросы теории систем с разделением процессора рассматриваются также в работах [23, 53, 60], [98, 113], [37, 111, 172], [101, 115, 182, 195], посвященных, в целом, другим проблемам.

статьи. В первую очередь мы стремимся указывать основные статьи, в которых были доказаны теоремы (представленные в последующих разделах).

Данная статья является расширенным и переработанным вариантом русско–язычной работы [230] (2006). В представленном варианте сделаны небольшие изменения в подразделах 1.1, 1.2, 2.4, 3.3, подразделы 2.3, 2.11, 3.1, 3.2 переработаны и значительно улучшены в нескольких пунктах, подразделы 2.5, 2.7 являются новыми.

1.2. Некоторые смежные направления исследований

Как уже отмечалось выше, в первую очередь мы обсуждаем точные аналитические решения системы M/GI/1 с эгалитарным разделением процессора и некоторые специальные случаи этой системы обслуживания, а также некоторые асимптотические решения. В этом подразделе дается лишь небольшое представление о некоторых смежных направлениях и их истоках.

Для обширной библиографии по различным типам аппроксимаций в теории очередей см., например, Бхат, Шелеби и Фишер (Bhat, Shalaby and Fisher) [27] (1979) (см. также [26]). Наболее распространенным подходом к аппроксимации классических систем обслуживания является их анализ в условиях высокой загрузки (явно или неявно опирающийся на центральную предельную теорему), предложенный Кингманом (Kingman) [103](1961). Цель такого анализа состоит в выводе более простых выражений в явном виде для предельных распределений характеристик системы. Можно также отметить тесно связанную с этим подходом диффузионную аппроксимацию классических систем обслуживания. Она детально исследована для однолинейных и многолинейных систем обслуживания систем обслуживания Виттом (Whitt) и Иглехартом и Виттом (Iglehart and Whitt), см. обзор [190] (1974) и [82] (1970). Обзор родственных работ в этом направлении исследований сделан, в частности, Глинном (Glynn) [72] (1990). Весьма популярна стала также флюидная, т.е. жидкостная аппроксимация (опирающуюся на усиленный закон больших чисел), предложенная Ньювеллом (Newell) в 1968 г. (см. [135, Ch. 6] (1971), [109, Ch. 2] (1976)).

С теоретической точки зрения случай большой загрузки интересен тем, что для классических систем обслуживания удается получить довольно простые асимптотические результаты в форме некоторых предельных теорем, которые зачастую можно распространить на более общие случаи (в частности, на случай рекуррентного входящего потока (типа GI) или даже типа G). В то же время точные результаты, если они достижимы в принципе, имеют очень сложный вид. Недавние достижения в этих направлениях исследований для классических систем обслуживания хорошо изложены в монографиях [96, 192][9].

В последнее десятилетие весьма популярны стали аналогичные задачи, но уже для систем обслуживания с дисциплинами разделения процессора. В качестве примеров укажем на некоторые работы из многочисленных публикаций в этом направлении, см., например, [74, 160, 214, 233].

Литература по асимптотическим решениям для стационарных характеристик систем с разделением процессора весьма обширна. Много статей посвящено так называемой “редуцированной эквивалентности по загрузке ” (reduced load equivalence), т.е. недавним результатам в духе теоремы 2.12 (см. также текст близ (2.75) и замечание 2.22). Эти асиммптотические результаты

[9] В сети Internet доступны даже дополнительные материалы к книге Витта [192] (суммарный объем этой книги с дополнением превышает 1000 стр.). Следует отметить также обзоры Стидхэма (Stidham) [169], Сайски (Syski) [174] и Джалалова (Dhalalow) [60] (1995), которые дают общее представление об основных достижениях классической теории очередей (и ее методах) за последние 50 лет. Тем не менее в статье [169], отражающей, впрочем, персональные взгляды автора, системы с разделением процессора упомянуты лишь мимоходом на стр. 199.

отражены также в обзоре [35][10]. Мы предпочитаем ссылаться на немногие оригинальные работы, которые, основаны на некоторых теоремах из статей [105,125,199,157,145], [47,203,204], [104, 207, 211, 74], [160, 214, 222]. Прежде всего, среди них имеются относящиеся к делу статьи [104] (1978), [199] (1981), [204] (1986), [74] (1991). В [199] выведены асимтотические оценки дисперсии стационарного времени пребывания в системе M/GI/1—EPS при малой и большой длине требования (важные дополнения и уточнения содержатся в [204]). В последнем случае эта оценка совпадает с оценкой Крамера (Cramér) [65,176] для риска разорения в обобщенном пуассоновском процессе. Позднее в этом направлении сделано несколько последующих шагов в [74, 234] и [91, 92], в которых из формул работ [104, 105, 199] удалось извлечь дальнейшую информацию об асимптотическом поведении стационарного времени пребывания в системе M/GI/1—EPS для некоторых подклассов распределений длин требований с так называемым "тяжелым" хвостом "heavy tailed"[11].Однако отметим, что основной результат статей [92, 234] был предсказан (но не доказан) Клейнроком [109, p. 175] (1976).

Еще один более или менее интересный результат относится к явному виду асимптотического решения для числа требований в момент t (с нормировкой t) в системе M/GI/1—EPS, работающей в условиях перегрузки. Первое довольно сложное и тяжеловесное доказательство принадлежит Гришечкину [74] (1991). Позднее этот результат распространен на случай рекуррентного входящего потока Жаном–Мари и Робером (Jean–Marie and Robert) [90](1994) (см.также [89]). Более простые версии доказательств для случая пуассоновского входа содержатся, например, в [12] (в контексте дискриминаторного разделения процессора[12]), [39,231] и т.п.

Относительно новым направлением исследований является изучение свойств ветвящихся процессов, которые возникают при изучениии систем с разделением процессора. Классическим примером, восходящим к статьям Бореля (Borel) [33] (1942) и Кендалла (Kendall) [99] (1951), является получение распределения периода занятости в системе M/GI/1—FCFS (см. также [65]). В теории ветвящихся процессов (см. [18], [114, гл. 13], [159]) такие процессы изучены очень глубоко. Однако было замечено (см. [207], [208, стр. 95], [p. 117][212]), что исследование нестационарного распределения числа требований в системе M/GI/1—EPS может привести к исследованию более сложных и слабее изученных ветвящихся процессов Крампа–

[10] Возможно, это первая из обзорных статей из серии однотипных и повторяющих друг друга обзоров с небольшими вариациями содержания. Они написаны более или менее одними и теми же авторами в течение 2002–2007. См. *Perform. Evaluation*, 2003, vol. 54, pp. 175–206, *Queueing Syst.*, 2006, vol. 53, pp. 31–51, etc. Авторы этих обзоров освещают только одну сторону анализа систем с разделением процессора и родственными дисциплинами, введенными в [48], а именно: асимптотику поведения хвостов распределений характеристик. Получение асимптотики хвоста распределения представляется более скромной целью по сравнению с нахождением самого распределения. Ясно, что первая цель должна быть подчинена второй из указанных, несмотря на тот факт, что в настоящее время преобразованияи Лапласа несколько пренебрегают. Авторы этих обзоров преувеличивают роль асимптотики и даже свой личный вклад в эту проблематику, который был сделан под непосредственным влиянием работ [104, 105, 199, 204, 208, 74]. В этих статьях впервые разработана теория системы обслуживания M/GI/1—EPS, но в указанных обзорах вообще отсутствуют ссылки на работы [104, 105, 199, 204, 208, 211]. Ряд результатов этих обзоров основан на статье [141], которая, в свою очередь, содержит только небольшие дополнения к статьям [104, 105, 199, 200, 201] (см. замечание 2.13 и сноску 31 для деталей). Кроме того, в этих обзорах не отражены другие оригинальные статьи, в особенности, принадлежащие российской школе. Обзоры являются менее полными по сравнению, например, с более ранней обзорной статьей Коваленко [113] по теории редких событий.

[11] Теория таких вероятностных распределений впервые появилась в финансовой математике. Ее детальное изложение можно найти, в частности, в монографиях [62] (1997), [165] (1998). Некоторые результаты частично отражены в §2.10.

[12] Дискриминаторное разделение процессора, введенное в [64] есть обобщение EPS на несколько классов требований, при котором каждое требование имеет собственный параметр относительного приоритета и получает обслуживание в режиме разделения процессора, пропорциональное своему параметру. См. [11, 206, 212] для обзоров.

Мода–Ягерса. Позднее такие процессы были изучены в контексте стационарного распределения времени пребывания в системе EPS в [74] (1991). (Ветвящихся процессы Крампа– Мода–Ягерса были введены для решения некоторых биологических проблем, см. [85] (1975) и недавнюю монографию [79] (2005) или обзор [182] (1993).) Переформулировка методов и результатов анализа системы M/GI/1—EPS из работ [105,199,201,203], где они излагались с позиций теории очередей, в терминах этих процессов была осуществлена в [74, 75]. Это позволило получить ряд интересных предельных теорем.

Теория ветвящихся процессов и теория очередей, в общем, имеют различные цели, хотя и имеют ряд общих аспектов, см. замечание 3.4. Тем не менее, системы обслуживания с разделением процессора — объект теории очередей, и в данном обзоре они рассматриваются с позиций именно этой теории с дополнительным привлечением некоторых нетривиальных вероятностных конструкций, в частности, случайной замены времени.

Статья организована следующим образом. Раздел 2 содержит основные результаты теории системы обслуживания M/GI/1—EPS в стационарном режиме. В нем введены некоторые концепции и детально описаны основные результаты. В частности, один из ключевых результатов представлен теоремой 2.5 в §2.3 вместе с ее доказательством. Детально обсуждаются некоторые асимптотические результаты и предельные теоремы. Раздел 3 посвящен теории системы обслуживания M/GI/1 с эгалитарным разделением процессора в нестационарном режиме. Внимание концентрируется на транзиентном распределении числа требований (§3.1). В §3.2 выводится совместное нестационарное распределение времени пребывания требования длины u, которое поступает в момент t, и числа требований в момент $t-$. Мы представим также некоторые интересные следствия главного результата (теорема 3.6), многие из которых доказывались ранее как автономные (самостоятельные) теоремы, но теперь они выводятся как частные случаи. Последний раздел содержит несколько заключительных замечаний.

Принятые обозначения незначительно отличаются от обозначений, использованных в обзоре [212]. Мы используем стандартные двойные номера для теорем, следствий и замечаний (первое число соответствует номеру раздела, второе число указывает на номер утверждения такого же типа в разделе). Литература (независимо от языка) дается единым списком в порядке, соответствующем латинскому алфавиту. Случайные величины и их функции распределения обозначаются прописными латинскими или греческими буквами, а для преобразований Лапласа–Стилтьеса распределений и их моментов используются соответствующие строчные буквы. Порядок момента указывается нижним индексом. Одинаковые обозначения используются для плотности распределения и ее преобразования Лапласа, поэтому они часто различаются только аргументами. Хотя тильда над символом часто указывает на двойное преобразование Лапласа, этот же знак может иногда использоваться для обычных преобразований Лапласа. Прописные латинские буквы применяются для обозначения случайных процессов, аргумент ω по традиции опускается. Символами $\mathsf{E}[\cdot]$ и $\mathsf{Var}[\cdot]$ обозначаются математическое ожидание и дисперсия, соответственно. Некоторые дополнительные комментарии и вспомогательные результаты приведены в сносках. Это помогает подойти ближе к автономной работе.

2. ТЕОРИЯ СИСТЕМЫ ОБСЛУЖИВАНИЯ M/GI/1—EPS. СТАЦИОНАРНЫЙ РЕЖИМ

2.1. Вводные замечания

В этом разделе мы представим наиболее существенные достижения (и ключевые особенности их вывода) в точном анализе системы M/GI/1 с эгалитарным разделением процессора. Понимание метода получения стационарных распределений основных вероятностно–временных характеристик позволит сократить объяснения для нестационарных решений. Основное внимание уделяется проблеме нахождения (условного) распределения времени пребывания требования в системе M/GI/1—EPS в терминах преобразований Лапласа–Стилтьеса (ПЛС), полу-

ченным решениям, методам решений и некоторым смежным результатам. По ходу изложения дается представление о наиболее важных достижениях, существовавших до наших работ и появившихся после них, о других подходах к исследованию, а также некоторая вспомогательная информация.

Рассмотрим систему M/GI/1 с дисциплиной эгалитарного разделения процессора (EPS)[13]. Эгалитарное разделение процессора обычно рассматривалось как предельная форма дисциплины RR при при бесконечно малой величине кванта обслуживания. Однако дисциплина RR, в общем, аналически неразрешима в системе обслуживания M/GI/1 (см. [208, §1.4], [195] или [216, Предложение 2.1]), это утверждение частично основано на [194]. Отсюда вытекает, что более удобным с математической точки зрения является изучение системы M/GI/1—EPS как системы с переменной скоростью обслуживания[14].

Дисциплина EPS введена Клейнроком [107] как предельный аналог дисциплины RR при устремлении величины кванта к нулю. Мы рассматриваем систему с дисциплиной EPS иначе — как систему, одновременно обслуживающую каждое из присутствующих требований с одинаковой скоростью, которая зависит от общего числа требований и равна $1/n$, если число требований в текущий момент равно n, $n = 1, 2, \ldots$.(Как следствие, дисциплина EPS позволяет коротким требованиям обгонять более длинные.) В моменты поступлений новых требований или уходов обслуженных происходят скачки скорости обслуживания. Если называть остаточной длиной требования количество работы по его обслуживанию с единичной скоростью, измеряемое в единицах времени, то под скоростью обслуживания понимается предел отношения изменения остаточной длины требования за промежуток времени Δ к Δ при $\Delta \to 0$

[13] Во всех классических дисциплинах обслуживания в однолинейных системах в любой момент времени обрабатывается не более одного требования, причем скорость обслуживания любого требования равна либо нулю при ожидании, либо единице при обслуживании. Другим классом дисциплин в таких системах обслуживания, включающим предельные формы дисциплин разделения времени, являются дисциплины разделения процессора. В этих моделях предполагается, что все требования или некоторое множество требований обслуживаются одновременно единственным прибором (процессором) с переменной скоростью, зависящей от состояния системы. Наиболее важной из дисциплин разделения процессора является EPS в силу следующих фактов. Долгое время такая система оставалась изученной только на поверхностном уровне (удавалось вывести формулы только для средних характеристик), поэтому проблема нахождения (стационарного) распределения времени пребывания стала считаться аналитически неразрешимой с математической точки зрения. Эскиз ее решения появился в [104] (1978), но полное доказательство дано впервые в [199] (1981) и усовершенствовано в [201] (1983). Таким образом, система M/GI/1—EPS превратилась в пригодную для анализа (tracktable) модель в последние десятилетия. Кроме того, дисциплина EPS более широко используется в практических приложениях по сравнению с другими дисциплинами.

[14] Дополнительное обоснование этого утверждения следующее.В любой момент времени ресурс процессора поровну разделяется между всеми требованиями, присутствующими в системе, и такое обслуживание эволюционирует как обслуживание с переменной скоростью. Этот взгляд на EPS основан на том, что модель EPS является непрерывным пределом циклической дисциплины RR [109]. Более конкретно, в системе round–robin [194, ?, 156], [120, 197, 54], все требования, присутствующие в системе, перемежают получаемое обслуживание некоторым квантом Θ. Система EPS соответствует случаю, когда Θ стремится к нулю. Способ интерпретации этой модели состоит в предположении, что требование имеет свои часы, которые работают только когда оно обслуживается. В частности, x–требование (с длиной x) выходит из системы, когда его часы показывают x. По сравнению с обычными часами, эти часы идут с переменной скоростью: когда n дополнительных требований присутствуют в системе, часы каждого из них идут в $(n + 1)$ раз медленнее. К сожалению, точное нахождение распределения времени пребывания (даже в стационарной ситуации) в системе M/GI/1 с дисциплиной round–robin является аналитически неразрешимой проблемой [216]; иизвестны только точные решения для случая M/M/1 посредством различных аналитических методов (см.Adiri и Avi–Itzhak [10] (1969), Nakamura, Murao и Tsukamoto [133] (1972), Schassberger [156] (1973), Yashkov [197] (1977), [208, Ch. 1] (1989), [216, Th. 2.1] (2002)). Это объясняет причины изучения системы M/GI/1—EPS как системы с переменной скоростью обслуживания. Рассмотрение системы M/GI/1—EPS как предела аналитически неразрешимой модели round–robin model не ведет к успеху. Только в первой постановке можно составить и решить систему дифференциальных уравнений (нетривиальных, но решаемых), описывающих процесс времени пребывания в модели EPS (см. §2.3).

(термин "длина" требования будет заменять неоднозначно трактуемый термин "длительность обслуживания". Здесь концепция времени ожидания не имеет смысла, поскольку обслуживание начинается немедленно. Фактически, время пребывания требования является реальным временем обслуживания.).

Дисциплину EPS можно описать следующим образом. Поступившее в систему требование сразу начинает обслуживаться (очередь в традиционном ее понимании отсутствует) и обслуживается с переменной скоростью (понимаемой в кинематическом смысле) до тех пор, пока его остаточная длина не станет равной нулю. В моменты изменения общего числа требований в системе происходят скачки скорости обслуживания. Скорость обслуживания флюктуирует во времени, а длительность пребывания отдельного требования в системе зависит не только от ранее поступивших, но и от последующих поступлений. Это на порядок усложняет анализ системы EPS по сравнению, например, с системой FCFS. Если в момент t имеется n требований, то в процессе обслуживания за бесконечно малый промежуток времени Δ происходит уменьшение остаточной длины каждого из n требований на $\Delta/n + o(\Delta)$. При этом ресурс процессора делится поровну (равномерно) между всеми обслуживаемыми требованиями. Такой способ разделения ресурса с точки зрения требований будет справедливым, что и подчеркивается в названии.

Практическое значение такой математической идеализации состоит в том, что она отражает наиболее существенную особенность систем с разделением времени (если под требованием понимать отдельное задание пользователя) или мультиплексных узлов коммутации пакетов (требованием является сообщение или его сегмент) — замедление обслуживания каждого требования, пропорциональное их общему числу в текущий момент. Пусть λ — интенсивность входящего пуассоновского потока, а длины требований — независимые и одинаково распределенные случайные величины (сл.в.) с произвольной функцией распределения (ф.р.) $B(x) = \mathrm{P}(B \le x),$, $B(0+) = 0$. Пусть первый момент $B(\cdot)$ есть

$$\beta_1 = \int_0^\infty (1 - B(x))dx < \infty$$

и ПЛС

$$\beta(s) = \int_{0-}^\infty \mathrm{e}^{-sx}\, dB(x) = s \int_{0-}^\infty \mathrm{e}^{-sx}\, B(x)\, dx.$$

Необходимым и достаточным условием существования стационарного режима является

$$\rho = \lambda\beta_1 < 1. \tag{2.1}$$

Пусть $V(u)$ — стационарное время пребывания в системе виртуального требования (называемого также "помеченным"), которое в момент поступления имеет длину u, т. е. реализацию сл.в. B, равную u (информация о величине u по каждому требованию недоступна для системы). Понятно, что $V(u)$ представляет собой также длительность фактического обслуживания требования длины u (для краткости такое требование будет иногда называться u–требованием). Положим $v(s,u) \doteq \mathsf{E}[\exp(-sV(u))]$. Здесь u показывает, что $v(s,u)$ определяет ПЛС *условного* распределения $V(x|u) = \mathsf{P}(\mathsf{V}(\mathsf{u}) \le \mathsf{x}|\mathsf{B} = \mathsf{u})$.

Коффман и др. [44] (1970) использовали марковское свойство системы M/M/1—EPS, чтобы найти (условное) распределение сл.в. $V(u)$; для системы M/G/1—EPS долгое время был известен только первый момент ф.р. сл.в. $V(u)$, полученный впервые Сакатой и др. [155] (1969) (см. (2.30)). Тем не менее, до недавнего времени появлялось много работ по различным способам вычисления только $\mathsf{E}[\mathsf{V}(\mathsf{u})]$ в этой системе (см., например, [17,68,138], [109,88,67]). Стационарное распределение времени пребывания требования длины u в системе M/GI/1—EPS впервые найдено в [104, 199, 200] в терминах ПЛС с помощью обобщения и развития предложенного

в [197] подхода к получению точного решения для $v(s,u)$ в системе M/M/1—RR (дополнительные подробности можно найти в [208, гл. 1]). В [201] усовершенствован метод нахождения распределения времени пребывания и получен ряд новых вероятностных характеристик системы M/GI/1—EPS. Исключительно интересные свойства этой системы были впервые обнаружены в [199, 201] на основе этих характеристик, существенно дополняющие свойства, отмеченные в [109, гл.4]. Одна из основных тем данного раздела — описание метода получения $v(s,u)$ в системе M/GI/1—EPS. Следует отметить, что здесь предполагается существование стационарного эргодического распределения $V(x|u)$, поэтому во введенных обозначениях не указана зависимость от времени.

Зная $v(s,u)$, легко найти ПЛС распределения безусловного времени пребывания V требования в системе

$$v(s) \doteq \mathsf{E}[\mathrm{e}^{-sV}] = \int_0^\infty v(s,u)\,dB(u). \tag{2.2}$$

Условное время ожидания требования длины u (точнее, потерянное время сверх необходимого для обслуживания со скоростью единица) определяется равенством $W(u) = V(u) - u$. Отсюда, зная $v(s,u)$, легко получить $w(s,u) \doteq \mathsf{E}[\mathrm{e}^{-sW(u)}]$.

Моменты распределения $V(x|u)$ вычисляются (с формальной точки зрения) как

$$v_j(u) = \lim_{s\downarrow 0}(-1)^j \partial v(s,u)/\partial s, \quad \mathsf{Var}[V(u)] = v_2(u) - v_1^2(u). \tag{2.3}$$

Кратко остановимся на распределении (стандартного) периода занятости Π. Положим $\pi(s) \doteq \mathsf{E}[\mathrm{e}^{-s\Pi}]$. Отсчет времени — с момента начала периода занятости. Все дисциплины разделения процессора принадлежат классу *консервативных* дисциплин.

Система обслуживания называется консервативной, если:

(i) на время обслуживания каждого требования не влияют задержки или прерывания, которые могли бы предписываться дисциплинами обслуживания,

(ii) прибор не простаивает, если имеются требования, ожидающие обслуживания.

Иными словами, не порождается и не уничтожаются работа внутри системы. В частности, класс консервативных дисциплин включает дисциплины с абсолютным приоритетом (прерываниями) и дообслуживанием.

Свойство консервативности означает, что длины требований не зависят от дисциплины и отсутствуют искусственные простои прибора. Нет также никаких потерь требований. Возможна следующая интерпретация в однолинейных системах обслуживания: в каждый момент времени на периоде занятости общая сумма мгновенных скоростей обслуживания всех имеющихся требований равна единице.

Теорема 2.1. *[104] В системе M/GI/1 с любой дисциплиной разделения процессора ПЛС распределения (стандартного) периода занятости $\pi(s)$ удовлетворяет известному для дисциплин LCFS и FCFS функциональному уравнению Такача*

$$\pi(s) = \beta(s + \lambda - \lambda\pi(s)). \tag{2.4}$$

Уравнение (2.4) определяет единственную функцию $\pi(s)$, аналитическую в полуплоскости $\mathrm{Re}\ s > 0$, *в которой* $|\pi(s)| \le 1$. *Функция $\pi(s)$ вполне монотонна. Если $\rho < 1$, то $\pi(0+) = \Pi(+\infty) = p^* = 1$, в противном случае $\pi(0+) < 1$, $\Pi(+\infty) = p^*$, где p^* — единственный корень уравнения $p^* = \beta(\lambda - \lambda p^*)$, лежащий в (0,1).*

Комментарии к доказательству. Два варианта вывода (2.4) даны в [208, стр. 57, 62-63]. Один из них появился ранее в [104]. Доказательство в [208, §2.3] основано на редукции к ветвящимся процессам. Классическим примером служит вывод (2.4) для системы M/GI/1—FCFS

(см., например, [65, гл. 14, §4], [48, §8.3], [52, §5.6]). Для вывода этого результата рассматриваются хорошо известные ветвящиеся процессы, возникающие в системе M/GI/1—LCFS и доказывается справедливость (2.4) для системы с такой дисциплиной. Финальным шагом служит высказывание, что распределения периодов занятости в обеих системах M/GI/1—LCFS и M/GI/1—FCFS совпадают. Стоит отметить, что, насколько нам известно, любое докзательство 2.4) сводится к анализу периода занятости при дисциплине LCFS. Отсутствуют прямые автономные (self–contained) доказательства только для дисциплины FCFS без привлечения LCFS.

Чтобы доказать (2.4) для системы M/GI/1—EPS, были введены в [104,199,201] более сложно устроенные ветвящиеся процессы с новым правилом порождения потомков (в отличие от [65], это правило основано на равновероятном случайном выборе). Обозначив через $\Pi(x)$ время, затрачиваемое на обслуживание предка длины x и всех его потомков, можно вывести и решить дифференциальное уравнение, которому удовлетворяет ПЛС распределения периода занятости. В результате

$$\pi(s,x) \doteq \mathsf{E}[\mathrm{e}^{-s\Pi(x)}] = \mathrm{e}^{-x(s+\lambda-\lambda\pi(s))}. \tag{2.5}$$

Снятие условия по x приводит к (2.4. Таким образом, (2.4) справедливо для системы M/GI/1—EPS.

Далее рассматривается процесс незаконченной работы $U(t)$ (сумма остаточных длин всех имеющихся требований в момент t). Поскольку сл.в. Π есть интервал времени, на котором $U(t) > 0$, то отсюда следует, что ф.р. сл.в. Π инвариантна относительно любой консервативной дисциплины в системе M/GI/1. Подклассом таких дисциплин являются все дисциплины разделения процессора, не только EPS. Они, в частности, удовлетворяют закону сохранения Клейнрока [109, §4.9], и не используют точной информации о требуемом времени обслуживания. □

Замечание 2.1. Равенство (2.5) совпадает с ПЛС ф.р. стационарного времени пребывания $V(x)$ требования длины x (x–требование)в системе M/GI/1 с дисциплиной LCFS–P (обслуживание в обратном порядке с прерываниями). При дисциплине LCFS–P каждому новому поступлению дается высший абсолютный приоритет, а в очереди, организованной по принципу стека, размещаются прерванные требования для дальнейшего дообслуживания.

2.2. Стационарное распределение числа требований

Ряд обозначений для модели M/GI/1—EPS введен выше. Долгое время для нее был известен только первый момент стационарного распределения времени пребывания требования длины u (см. формулу (2.30)) и стационарное распределение числа требований (см. формулу (2.8)[15]). Введем в рассмотрение кусочно–линейный марковский процесс $\{X_0(t) = \{L(t); x_i(t), i = 1, \ldots, L(t); t \in \mathbf{R}_+\}$, характеризующий состояние системы в момент t. Здесь $L(t)$ — число требований в системе EPS в момент t, а дополнительные переменные $x_i(t)$ указывают на остаточную длину i–го требования в момент t (порядок нумерации требований не имеет значения в силу специфики дисциплины EPS). Пространство состояний процесса $X_0(t)$ есть $\{\mathbf{Z}_+ \times \{\emptyset \cup \mathbf{R}_+{}^1 \cup \mathbf{R}_+{}^2 \ldots\}\}$. Здесь $\emptyset$ обозначает состояние, когда $L(t) = 0$, $\mathbf{Z}_+ = \{0, 1, 2, \ldots\}$. Пусть также процесс $\{V_t(u), t \in \mathbf{R}_+\}$ описывает время пребывания требования, которое приходит в момент t и имеет в этот

[15] Эскиз доказательства приведен впервые в [155]. Применён метод введения дополнительных переменных, в качестве которых использовались прошедшие длительности обслуживания (т.е. достигнутые величины полученного обслуживания) i-го требования, $i = 1, 2, \ldots, L(t)$, обслуживаемого в момент t. Поведение исследуемого процесса описывалось системой интегро–дифференциальных уравнений с некоторыми граничными условиями. Отсутствовал вывод этой системы уравнений, поэтому доказательство было неполным.

момент длину u. В силу следствия из теоремы Смита для регенерирующих процессов [168] (точками регенерации служат моменты окончания периодов занятости) справедлива

Теорема 2.2. *При выполнении условия (2.1) процессы $\{V_t(u)\}$ и $X_0(t)$ обладают единственными стационарными распределениями.*

Следствие 2.1. *Процесс $X_0(t)$ ограничен по вероятностям состояний.*

Теорема 2.2 обеспечивает существование и единственность следующих пределов:

$$v(s,u) = \lim_{t\to\infty} \mathsf{E}[\mathrm{e}^{-sV_t(u)}];$$

$$\mathsf{P}_n(x_1,\dots,x_n) = \lim_{t\to\infty} \mathsf{P}_n(t;x_1,\dots,x_n) = \lim_{t\to\infty} \mathsf{P}(L(t)=n; x_i(t)\in[x_i,x_i+dx_i), i=1,2,\dots,n); \tag{2.6}$$

$$\mathsf{P}_n = \lim_{t\to\infty} \mathsf{P}(L(t)=n).$$

В (2.6) $\mathsf{P}_n(x_1,\dots,x_n)$ означает стационарную плотность вероятности пребывания процесса $X_0(t)$ в состоянии $(n;x_1,\dots,x_n)$. Здесь неявно предполагается, что финальное распределение процесса $X_0(t)$ при фиксированном числе требований $n\geq 1$ обладает плотностью. Это условие выполняется, если предположить для простоты, что ф.р. $B(x)$ имеет плотность $\beta(x)$.

Теорема 2.3. *Если выполняется условие (2.1) в системе M/GI/1− EPS, то*

$$\mathsf{P}_n(x_1,\dots,x_n) = \mathsf{P}_0\lambda^n \prod_{i=1}^{n}[1-B(x_i)],\ n\geq 1, \tag{2.7}$$

where $\mathsf{P}_0 = \mathsf{P}\{\emptyset\} = 1-\rho$.

Эти результаты находятся после составления и решения дифференциальных уравнений Колмогорова, которым удовлетворяют плотности одномерных распределений компонент процесса $X_0(t)$. Те же самые результаты в общем случае можно получить при замене плотностей $\mathsf{P}_n(x_1,\dots,x_n)$ дифференциалами соответствующих ф.р. с привлечением более утонченных обобщенных функций из теории распределений Шварца.

Плотность $\mathsf{P}_n(x_1,\dots,x_n)$ симметрична относительно любой перестановки переменных x_i, $i=1,2,\dots,n$ в силу того, что предположение об упорядоченности дополнительных координат процесса $X_0(t)$ не делалось.

Отметим, что (2.7) было получено в [105] при рассмотрении процесса $A_0(t)=\{L(t),a_i(t),i=1,\dots,L(t);t\in\mathbf{R}_+\}$ как вспомогательный результат, необходимый при доказательстве пуассоновского характера выходящего потока. Отличие процесса $A_0(t)$ от $X_0(t)$ заключается в том, что $a_i(t)$ имеет смысл не остаточной длины, а достигнутого времени обслуживания, уже полученным i–м требованием. Плотности вероятностей состояний процесса $A_0(t)$ описываются равенством, аналогичным (2.7) при соответствующей замене переменных, т. е.

$$\mathsf{P}_n(a_1,\dots,a_n) = \mathsf{P}_0\lambda^n \prod_{i=1}^{n}[1-B(a_i)],\ n\geq 1,\ \text{где } \mathsf{P}_0 = \mathsf{P}\{\emptyset\}. \tag{2.8}$$

Попутно отметим, что от (2.8) нетрудно перейти к (2.7), однако найти простую связь (2.7) с (2.8) в обратную сторону не удается.

Следствие 2.2. *Число требований в системе $M/G/1$ с дисциплиной EPS имеет геометрическое распределение*

$$\mathsf{P}_n = (1-\rho)\rho^n,\ n = 0, 1, 2, \ldots, \tag{2.9}$$

инвариантное относительно вида $B(x)$ при фиксированном среднем β_1.

Доказательство.

$$\mathsf{P}_n = \int_0^\infty \ldots \int_0^\infty \mathsf{P}_n(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = \mathsf{P}_0\rho^n,\ n \geq 1.$$

Из условия нормировки $\sum_{n=0}^\infty \mathsf{P}_n = 1$ находим $\mathsf{P}_0 = 1-\rho$. □

Еще одно доказательство следствия 2.2 дано после следствия 3.3.

Свойства инвариантности такого же типа характерны и для многолинейных систем. Например, равновесная вероятность нахождения n требований в классической системе $\mathrm{M/GI/\infty}$ с пуассоновским входящим потоком интенсивности λ имеет вид

$$\mathsf{P}_n = \mathrm{e}^{-\rho}\,\rho^n/n!, \quad n = 0, 1, 2, \ldots, \quad \rho = \lambda\beta_1.$$

Замечание 2.2. Отмеченная аналогия становится более понятной, если рассмотреть процесс обслуживания в системе $M/GI/1$—EPS в новом масштабе времени: интервалы времени, в течение которых в системе обслуживается $n > 1$ требований, сжимаются в n раз [105] (эта идея предложена А.Д.Соловьевым в 1975). При таком изменении масштаба времени мгновенная скорость обслуживания уже не является случайной переменной, зависящей от числа требований, а постоянна и равна единице. Однако при этом усложняется структура входящего потока. Свойство отсутствия последействия сохраняется, но интенсивность становится переменной, в зависимости от числа обслуживаемых требований. Иными словами, на том интервале времени, когда в системе находится n требований, $n \geq 1$, интенсивность новых поступлений равна $n\lambda$ (и λ при $n = 0$). Входящий поток на этом интервале времени можно рассматривать как суперпозицию n независимых пуассоновских потоков интенсивности λ. Таким образом, в новом масштабе времени система $M/GI/1$—EPS превращается в систему $\tilde{M}/GI/\infty$ с марковским входящим потоком переменной интенсивности $n\lambda$ (фактически интенсивность входа является случайным процессом, так как n есть значение $L(t)$, числа требований в момент t). Теперь новое поступление может занять с равной вероятностью любой свободный прибор в бесконечно–линейной системе обслуживания $\tilde{M}/GI/\infty$. Такое преобразование случайных процессов позволило дать некоторое теоретическое обоснование новому методу анализа системы $\mathrm{M/GI/1}$ с дисциплиной EPS и с некоторыми другими дисциплинами разделения процессора [105,201,203,208,211]. Это преобразование является одним из примеров случайной замены времени в теории случайных процессов. Случайная замена времени впервые введена Волконским [186] (1958) в контексте абстрактной теории случайных процессов и позднее была развита многими авторами (см., например, обзоры Сайски [173,174] или монографии [31, Ch. 5], [124, Ch. 7, §2], [165], [114,188]). В частности, случайная замена времени использовалась в теории стохастического интегрирования [165] (обобщенная формула Ито для выпуклых функций от броуновского движения, преобразование любого непрерывного локального мартингала с квадратической вариацией, стремящейся к бесконечности, в броуновское движение, интегрирование по возрастающему процессу [124,31] и т.п.). По–видимому, подобное преобразование было впервые применено в теории очередей в [105, 202, 203] для другого класса случайных процессов, которые возникли при изучении систем обслуживания с разделением процессора. Ключевой идеей является использование инверсии некоторых аддитивных функционалов в качестве случайной замены времени в процессе $\{L(t),\ t \geq 0\}$, в котором $L(t)$ есть число требований в момент t. Такая идея

крайне необыкновенна в теории очередей. Это позволило теоретически обосновать аналитические методы, введенные в [197, 199, 201].

Мы теперь готовы доказать еще одно следствие теоремы 2.3.

Следствие 2.3. *Стационарные вероятности состояний* $\mathsf{P}^\star_n(x_1,\dots,x_n)$ *системы* $\tilde{M}/G/\infty$ *имеют вид*

$$\mathsf{P}^\star_n(x_1,\dots,x_n) = \mathsf{P}^\star_0 \frac{\lambda^n}{n} \prod_{i=1}^{n} [1 - B(x_i)],\ n \geq 1, \tag{2.10}$$

где

$$\mathsf{P}^\star_n = \int_0^\infty \dots \int_0^\infty \mathsf{P}^\star_n(x_1,\dots,x_n)\, dx_1 \dots dx_n = \mathsf{P}^\star_0 \rho^n / n, \tag{2.11}$$

$$\mathsf{P}^\star_0 = \mathsf{P}\{\emptyset\} = (1 - \ln(1-\rho))^{-1},\ \rho < 1.$$

Доказательство. Когда система M/GI/1—EPS содержит $n > 1$ требований, то "новое" время течет в ней в n раз быстрее реального времени (см. замечание 2.2). Поэтому отношение стационарных вероятностей пребывания в состояниях $(n; x_1, \dots, x_n),\ n \geq 1$ для систем M/GI/1—EPS и $\tilde{M}/GI/\infty$ равно n. В силу этого факта из формулы (2.7) вытекает (2.10). Равенство (2.10) приводит к (2.11), что позволяет получить с учетом условия нормировки и ряда Тейлора для натурального логарифма формулу для стационарной вероятности $\mathsf{P}^\star_0$ отсутствия требований в системе $\tilde{M}/GI/\infty$. □

2.3. Метод декомпозиции на элементы задержки

В дальнейшем обсуждении мы следуем [201, 208, 216]. Опишем метод нахождения $v(s,u) = \mathsf{E}[\mathrm{e}^{-sV(u)}]$ в системе M/GI/1—EPS. Одновременно будет восстановлено доказательство теоремы 2.4, которое вновь дается в расширенном виде в этом подразделе. Под сл.в. $V(u)$ будем понимать промежуток времени, за который остаточная длина некоторого требования уменьшится на u. Эквивалентно, это промежуток времени, за который требование получит приращение достигнутого времени обслуживания, равное u. Суть метода состоит в изучении динамики прохождения через систему некоторого виртуального требования длины u. Такое требование будем считать помеченным и поступающим в систему в момент $t = 0$ при начальном условии, что в момент своего поступления оно встречает n других требований в системе с остаточными длинами $x_1, \dots, x_n$. Метод основан на декомпозиции сл.в. $V(u)$ на сумму некоторых "*элементов задержки*". В действительности будет исследоваться процесс накопления обслуживания, полученного этим виртуальным требованием, которое имеет начальную длину u.

Другими словами, состояние процесса $X_0(t)$ есть $(n; x_1, x_2, \dots, x_n)$ в момент непосредственно до поступления помеченного требования (напомним, что система находится в стационарном режиме (в равновесии), т.е. она начала работать в момент $t = -\infty$). Хорошо известно, что пуассоновские поступления наблюдают систему M/GI/1 в таком состоянии с вероятностью (2.7) благодаря свойству PASTA (Poisson Arrivals See Time Averages: пуассоновские поступления видят временны́е средние). Главный (и наиболее трудный) этап анализа заключается в следующем. Нам необходимо найти условное ПЛС: $\mathsf{E}[\mathrm{e}^{-sV(u)} | (n; x_1, \dots, x_n)]$, которое будет представлено теоремой 2.4. Чтобы сделать это, с каждым из n требований, которых встречает помеченное при своем поступлении, свяжем некоторые ветвящиеся процессы. Кроме того, с самим помеченным требованием свяжем некоторый дополнительный ветвящийся процесс. Каждое из этих требований будем считать *предком*, в то время как новые поступления после момента $t = 0$ будут считаться *потомками* предков. Случайный механизм, относящий новые поступления к какому-либо потомству, основан на следующем правиле: когда в системе EPS

присутствует $(n+1)$ предков, то каждое новое поступление объявляется *прямым потомком* любого (но только одного) из этих предков с вероятностью $1/(n+1)$. Будут учитываться только те потомки, которые поступят до ухода из системы помеченного виртуального требования (напомним, что помеченное требование также является предком. Таким образом, изучаются *обрывающиеся* ветвящиеся процессы. Каждый из этих $(n+1)$ ветвящихся процессов формируется одним предком и его прямыми потомками.

Пусть $\Phi(x,u)$ есть сумма приращений достигнутого времени обслуживания некоторого предка длины x и его прямых потомков за интервал времени, в течение которого остаточная длина другого предка (скажем, помеченного требования) уменьшается на u. Эта сл.в. может также рассматриваться как некоторый (марковский) функционал от соответствующего ветвящегося процесса, описывающий его полное время жизни. Проще интерпретировать сл.в. $\Phi(x,u)$ как длительность некоторого *обрывающегося* (суб)периода занятости, который открывается предком длины x. Он обрывается в момент, когда приращение достигнутого обслуживания помеченного требования достигает уровня u. Вероятностная структура компонент этого субпериода занятости напоминает структуру компонент обычного периода занятости с тем существенным отличием, что каждая последующая компонента зависит от момента обрыва ветвящегося процесса и длины потомка. Поэтому последующая компонента "стохастически меньше" предыдущей (в смысле введенного Штойяном (Stoyan) [170, 171] отношения порядка $\overset{1}{\leq}$ для функций распределения) [16] Имеются также другие типы стохастического порядка.

Отметим, что при $u \to \infty$ сл.в. $\Phi(x,u)$ превращается в стандартный период занятости, у которого длина открывающего его требования фиксирована и равна x. (ПЛС распределения этой сл.в. при $u \to \infty$ дается правой частью (2.5). Иными словами, это время первого достижения уровня $x=0$ процессом незаконченной работы $\{U(t),\ t \geq 0\}$ при условии $U(0)=x$.) При $x \geq u$ сл.в. $\Phi(x,u)$ не зависит от x. Удобно ввести специальное обозначение в этом случае

$$D(u) \stackrel{d}{=} \Phi(x,u) \qquad x \geq u. \tag{2.12}$$

Напомним, что система EPS находится в состоянии $(n; x_1, \ldots, x_n)$ в момент $t=0$. Тогда время пребывания помеченного требования можно представить как

$$V_n(u, x_1, \ldots, x_n) \stackrel{d}{=} \sum_{i=1}^{n} \Phi(x_i, u) + D(u), \tag{2.13}$$

при указанном выше начальном условии. Компоненты (2.13) (которые были названы "элементами задержки") — независимы друг от друга. Независимость этих сл.в. — нетривиальный факт, который элегантно доказан с помощью случайной замены времени [203] (см. замечание 3.1). Предыдущее аргументы из [199, 200, 201] для доказательства этой независимости представлены в замечании 2.3.

Таким образом, время пребывания $V(u)$ декомпозируется в сумму некоторых функционалов от *независимых* ветвящихся процессов, распределения которых могут быть найдены. Это

[16] Напомним определения только для двух типов стохастического порядка $\overset{(1)}{\leq}$ и $\overset{(2)}{\leq}$. Первый из них, distributional order, обозначается также как $\leq_{st}$ или $\leq_d$ или $\leq_p$; последний из них, выпуклый (convex) или variational order, обозначается также как $\leq_c$ (мы предпочитаем первоначальное обозначения):

$B_i(x) \overset{(1)}{\leq} B_j(x)$, если $1 - B_i(x) \leq 1 - B_j(x)$ для $\forall x \geq 0$;

$B_i(x) \overset{(2)}{\leq} B_j(x)$, если $\int_x^\infty (1 - B_i(y))\, dy \leq \int_x^\infty (1 - B_j(y))\, dy$ для $\forall x \geq 0$.

краеугольный камень теории системы M/GI/1—EPS[17]. Отметим, что что компоненты правой части (2.13) — некоторые аналоги элементов задержки в системе M/M/1 с дисциплиной round–robinRR (см. [197,208]). Но рассматриваемый случай является значительно более сложным из–за произвольного распределения $B(\cdot)$ и одновременного обслуживания с переменной скоростью. Поэтому для нахождения распределений компонент (2.13) необходимо составить и решить некоторую систему дифференциальных уравнений. Положим $\varphi(s,x,u) \doteq \mathsf{E}[\mathrm{e}^{-\mathsf{s}\Phi(\mathsf{x},\mathsf{u})}]$ и $\delta(s,u) \doteq \mathsf{E}[\mathrm{e}^{-\mathsf{s}\mathsf{D}(\mathsf{u})}]$. Рассматривая инфинитезимальные изменения сл. в. $\Phi(x,u)$ за малый интервал времени Δ, можно вывести систему уравнений с начальными и граничными условиями условиями, которым удовлетворяет $\varphi(s,x,u)$ и $\delta(s,u)$

$$\frac{\partial\varphi(s,x,u)}{\partial x} + \frac{\partial\varphi(s,x,u)}{\partial u} + \left[s+\lambda-\lambda\int_0^\infty \varphi(s,y,u)\,dB(y)\right]\varphi(s,x,u) = 0, \tag{2.14}$$

$$\frac{\partial\delta(s,u)}{\partial u} + \left[s+\lambda-\lambda\int_0^\infty \varphi(s,y,u)\,dB(y)\right]\delta(s,u) = 0, \tag{2.15}$$

$$\delta(s,0) = \varphi(s,0,u) = \varphi(s,x,0) = 1. \tag{2.16}$$

Для вывода этих уравнений удобно использовать стохастические равенства (2.17) и (2.18), описывающие изменение элементов задержки виртуального помеченного требования с точностью до членов порядка $o(\Delta)$. Пустьt η будет интервалом времени, за который $(x+\Delta)$–требование превращается в x–требование.

$$\Phi(x+\Delta,u+\Delta) = \begin{cases} \Delta+\Phi(x,u), & \text{если нет потомков } (x+\Delta)\text{–требования за интервал } \eta,\\ \Delta+\Phi(x,u)+\Phi(y,u), & \text{если появился прямой } y\text{–потомок за интервал } \eta. \end{cases} \tag{2.17}$$

(Равенство (2.17) выполняется в области $x<u$. Аналогично,

$$D(u+\Delta) = \begin{cases} \Delta+D(u), & \text{если нет потомков } (u+\Delta)\text{–требования за интервал } \eta,\\ \Delta+D(u)+\Phi(y,u), & \text{если появился прямой } y\text{–потомок за интервал in time interval } \eta. \end{cases} \tag{2.18}$$

Замечание 2.3. Члены правых частей равенств (2.17) и (2.18) независимы. Это можно доказать следующим образом. Если Δ инфинитезимален, то η тоже инфинитезимален в силу следствия 2.1. А именно, когда имеется n требований, то $\eta = n\Delta$. Но каково бы ни было n, вероятности отсутствия или появления прямого потомка за in η независимы от n, потому что вероятности первого и последнего из указанных событий соответственно, равны:

$$(1-\lambda\eta)+\lambda\eta\frac{n-1}{n}+o(\eta) = 1-\lambda\Delta+o(\Delta), \quad \frac{1}{n}(\lambda\eta+o(\eta)) = \lambda\Delta+o(\Delta)$$

в силу правила порождения потомков, описанного во втором абзаце этого подраздела. Теперь можно убедиться, что компоненты равенства (2.13) независимы в силу:
(i) число прямых потомков каждого предка зависит только от длины предка,
(ii) независимость длин требований,
(iii) независимость приращений пуассоновского процесса.

[17] По–видимому, такая стохастическая декомпозиция являетсЯ новой. К примеру, хорошо известная декомпозиция времени ожидания Фурмана и Купера [49, p.222], [50, pp.503–504] и Такаги [177] в моделях с прогулками прибора и их приложениям к системам поллинга не попадает в рамки (2.13).

Если переписать (2.17) и (2.18) в терминах ПЛС и применить затем формулу полного математического ожидания, то результате предельного перехода при $\Delta \to 0$ мы получим систему дифференциальных уравнений (2.14) и (2.15). В свою очередь, эти уравнения можно свести к дифференциальному уравнению с частными производными первого порядка, решение которого находится с помощью классического метода характеристик. Все детали этого решения были даны в [199,201], позднее они были повторены в [141,136,137,187]. Для удобства мы напомним основные шаги решения из [199,201].

Уравнение (2.15) можно переписать как

$$\frac{\partial}{\partial u}\ln\delta(s,u) = -\left[s+\lambda-\lambda\int_0^\infty \varphi(s,y,u)dB(y)\right].$$

Подстановка этого последнего уравнения в (2.14) приводит к

$$\frac{\partial\varphi(s,x,u)}{\partial x}+\frac{\partial\varphi(s,x,u)}{\partial u}-\varphi(s,x,u)\frac{\partial}{\partial u}\ln\delta(s,u)=0. \tag{2.19}$$

Дифференциальное уравнение с частными производными первого порядка (2.19) решается с помощью классического метода характеристик. Для этого следует решить вспомогательную (сопряженную) систему обыкновенных дифференциальных уравнений (система уравнений записана в симметрической форме)

$$\frac{dx}{1}+\frac{du}{1}=\frac{d\varphi}{\varphi(\ln\delta)'},$$

где $\varphi=\varphi(s,x,u)$ и $\delta=\delta(s,u)$. Находим первые интегралы этой системы уравнений $C_1 = u-x$ и $C_2=\varphi/\delta$. Поскольку φ входит только в C_2, то общим решением уравнения (2.19) является $C_2=f(C_1)$. Здесь $f=f(s,x,u)$ есть произвольная дифференцируемая функция, которая в нашем случае находится с помощью условий (2.16). Виду того, что $\varphi(s,0,u)=1$, это дает $f=1/\delta$ при $x=0$, что в конечном счете позволяет получить соотношение, которому удовлетворяют функции φ и δ в области $x<u$. Оно дается второй строкой правой части равенства (2.25). Первая строка очевидна в силу равенства (2.12). Подставляя (2.25) в (2.15), приходим к следующему уравнению, которому удовлетворяет неизвестная функция δ

$$\frac{\delta(s,u)}{\partial u}+\left[s+\lambda-\lambda\int_0^u\frac{\delta(s,u)}{\delta(s,u-y)}\,dB(y)-\int_u^\infty dB(y)\right]\delta(s,u)=0. \tag{2.20}$$

Так как каждое решение уравнения (2.20) должно удовлетворять оценке (ср. с (2.5))

$$\delta(s,u)>\mathrm{e}^{-u(s+\lambda-\lambda\pi(s))}\quad \text{for } \mathsf{Re}\ s>0, \tag{2.21}$$

то функция δ может быть представлена в виде равенства (2.26), в котором неизвестная функция $\psi(s,u)<\mathrm{e}^{-\lambda\pi(s)u}$ при $\mathsf{Re}\ s>0$. Подстановка (2.26) в (2.20) приводит к уравнению, которому удовлетворяет неизвестная функция $\psi(s,u)$

$$\frac{\psi(s,u)}{\partial u}+\lambda\int_0^u\mathrm{e}^{-y(s+\lambda)}\,\psi(s,u-y)dB(y)+\lambda(1-B(u))\,\mathrm{e}^{-u(s+\lambda)}=0 \tag{2.22}$$

с дополнительными условиями $\psi(s,0)=1$ и $\psi(0,u)=\mathrm{e}^{-\lambda u}$. Эти условия находятся из равенств (2.26) и (2.16).

Уравнение (2.22) решается с помощью применения преобразования Лапласа (ПЛ). Определим

$$\tilde{\psi}(s,q)=\int_0^\infty\mathrm{e}^{-qu}\psi(s,u)\,du. \tag{2.23}$$

Имееет смысл отметить, что $\tilde{\psi}(s,q)$ является двумерным преобразованием Лапласа функции $\Psi(x,u)$ от двух переменных. Эта функция имеет плотность вероятностей по переменной x, т.е.

$$\tilde{\psi}(s,q)=\int_0^\infty\int_0^\infty \mathrm{e}^{-sx-qu}d_x\Psi(x,u)\,du,\quad s>0,\ q>-\lambda\pi(s),$$

в то время как $\psi(s,u)$ является обычным ПЛ по x плотности функции $\Psi(x,u)$. (Ясно, что функции $\Psi(x,u)$ и даже $\psi(s,u)$ не могут быть представлены в явном виде.)

Применяя преобразование (2.23) к каждому члену уравнения (2.22), получаем

$$q\tilde{\psi}(s,q)-\psi(s,0)+\lambda\tilde{\psi}(s,q)\beta(q+s+\lambda)+\frac{\lambda[1-\beta(q+s+\lambda)]}{q+s+\lambda}=0.$$

Это простое уравнение приводит к утверждению (2.27). Одновременно мы получили решение уравнения (2.22) в терминах ПЛ по u. Конечно, необходимо использовать обращение преобразований Лапласа для получения более прозрачных формул, однако это возможно только для некоторых специальных случаев $B(x)$ (см. §2.6). Применение ПЛС к обеим сторонам равенства (2.13) приводит к (2.24). Это завершает доказательство ключевой теоремы 2.4. □

Теорема 2.4. *(1978) [104, 199, 200, 201] For the M/GI/1—EPS queue, we have*

$$\mathsf{E}[\mathrm{e}^{-sV(u)}|(n;x_1,\dots,x_n]=\mathsf{E}[\mathrm{e}^{-sV_n(u,x_1,\dots,x_n)}]=\delta(s,u)\prod_{i=1}^n\varphi(s,x_i,u),\quad Re\ s>0, \tag{2.24}$$

where

$$\varphi(s,x,u)=\begin{cases}\delta(s,u) & for\ x\ge u,\\ \delta(s,u)/\delta(s,u-x) & for\ x<u,\end{cases} \tag{2.25}$$

$$\delta(s,u)=\mathrm{e}^{-(s+\lambda)u}/\psi(s,u),\quad u\ge 0, \tag{2.26}$$

$$\tilde{\psi}(s,q)\doteq\int_0^\infty e^{-qu}\psi(s,u)\,du=\frac{s+q+\lambda\beta(s+q+\lambda)}{(s+q+\lambda)[q+\lambda\beta(s+q+\lambda)]},\quad Re\ s>0,\ q>-\lambda\pi(s). \tag{2.27}$$

Here $\pi(s)$ is the solution of the equation (2.4).

Комментарии к доказательству.[18] Сокращенное доказательство впервые дано в [104] (1978), два варианта полных доказательств представлены в [199, 200] (1981). Усовершенствованный вариант доказательства содержится в [201] (1983). Некоторые дополнительные аргументы доказательств обсуждались также в [203,206,212]. monograph [208, Ch. 2]. □

[23,33,35,43–45,47]. Основные результаты вместе с дополнительным обсуждением введенных конструкций включены в монографию [208, гл.2] (1989). □

Замечание 2.4. Стоит упомянуть, что случайная величина $D(u)$ в равенстве (2.13) представляет собой "главную" компоненту времени пребывания: она имеет распределение времени пребывания требования с длиной u, которое приходит в пустую (свободную от требований) систему. Когда система занята, то i–ое требование (среди требований, разделяющих ресурс процессора), имеющее остаточную длину x_i, "добавляет" некоторую задержку $\Phi(x_i,u)=\Phi(x_i\wedge u,u)$ к длительностям пребывания новых требований.

[18] Выше дано объяснение, в основном, к аналитической части доказательств. Мы упомянули про одну из трудностей: доказательство независимости компонент декомпозиции (2.13). Существует другая утонченная трудность, но ее обсуждение выходит за рамки данной статьи. По этим причинам доказательство неоднократно модифицировалось. Внешняя простота метода обманчива.

Замечание 2.5. При доказательстве теоремы 2.4 не использовалось равенство $\pi(0+) = 1$, которое имеет место при выполнении условия (2.1), поэтому теорема 2.4 справедлива и при $\rho \geq 1$.

Замечание 2.6. В некоторых случаях могут быть полезны эквивалентные формы записи равенства (2.25). В качестве примера приведем одну из них

$$\varphi(s,x,u) = \mathrm{e}^{-(x\wedge u)(s+\lambda)+\lambda\int_0^{x\wedge u}\varphi_B(s,u-y)\,dy},\ x\in[0,\infty),$$

где

$$\varphi_B(s,t) \doteq \int_0^\infty \varphi(s,x,t)\,dB(x) = \int_0^t \mathrm{e}^{-\int_{t-x}^t(s+\lambda-\lambda\varphi_B(s,y))\,dy}\,dB(x)+(1-B(t))\,\mathrm{e}^{-\int_0^t(s+\lambda-\lambda\varphi_B(s,y))\,dy}.$$

Последняя формула представляет собой некоторое функциональное уравнение, которому удовлетворяет функция $\varphi_B(s,\cdot)$. $\varphi_B(s,t)$ есть ПЛС распределения некоторого обрывающегося периода занятости (ТВР). ТВР оканчивается (обрывается) в момент ухода помеченного t–требования из системы M/GI/1—EPS, когда мы рассматриваем эту систему в старом (реальном) масштабе времени. Если мы рассматриваем систему EPS в новом масштабе времени (см. замечание 2.2), то ТВР оканчивается в момент t нового времени. Это нетривиальное понятие для теории очередей (среди других понятий, введенных в [201, 208]), потому что распределение такого обрывающегося периода занятости не инвариантно относительно консервативных дисциплин. Например, система M/GI/1—FBPS имеет другое распределение ТВР [208, 212, 225]. Решение приведенного выше уравнения было получено в терминах функции ψ $(\psi(s,t) \doteq \exp(-\lambda\int_0^t \varphi_B(s,y)\,dy))$ (точнее, в терминах изображения по Лапласу этой функции, см. (2.27)). Это также показывает, что изучение процесса времени пребывания в системе M/GI/1 (даже в стационарном режиме) требует более глубокого анализа по сравнению с анализом, который ожидается на поверхностном уровне. К сожалению, до сих пор многие работы в этой области выполняются именно на таком привычном уровне.

Отметим, что дополнительные объяснительные комментарии к нашему методу анализа содержатся также непосредственно после замечания 3.7. Комментарии даются с различных точек зрения и могут оказаться полезными для читателей.

Замечание 2.7. Более полная информация о различных формах интерпретации уравнений (2.14), (2.15) и способах их составления содержится в [208]. Отметим, что [200] описывает другой способ для более легкого вывода таких уравнений. Этот способ основан на применении так называемого метода введения дополнительного события (окраски требований) (the method of collective marks), который введен ван Данцигом [56], Раненбургом [153] and Климовым (см. [108, Ch. 7], [110]). Метод был развит, в частности, в [123] и [128] для приоритетных систем обслуживания. Он использовался также при изучении системы M/GI/1 с более простой *gated processor sharing* дисциплиной разделения процессора [110][19]. Однако, как верно заметил Клейнрок [108, p. 269],

> "...этот метод не дал пока никаких результатов, которые не были бы ранее известны благодаря применению других методов".

[19] Такая дисциплина (точнее, класс дисциплин) введена Клейнроком в 1970 в контексте более широкого класса *эгоистических (selfish)* дисциплин разделения процессора (см. ссылку [49] в [109, Ch. 4], а также [212, pp. 13–14]), ее средние характеристики изучены в [139], стационарные распределения основных характеристик выведены в [110].

В качестве последнего шага следует подчеркнуть, что пуассоновские поступления создают моменты случайного наблюдения за системой M/GI/1 в произвольный момент времени. Следовательно, вероятности состояний системы EPS в моменты поступлений совпадают с вероятностями состояний в произвольный момент времени в силу закона PASTA. Поэтому при $\rho < 1$, можно снять в равенстве (2.24) условие по $(n; x_1, \ldots, x_n)$ с помощью теоремы 2.3, которая гарантирует мультипликативную форму плотности совместного распределения стационарного числа требований $L = n$ и их остаточных длин $x_i \in [x_i, x_i + dx_i),\ i = 1, \ldots, n$. Это позволяет получить

$$v(s,u) \doteq \mathsf{E}[\mathrm{e}^{-sV(u)}] = \frac{(1-\rho)\delta(s,u)}{1 - \lambda \int_0^\infty \varphi(s,x,u)(1-B(x))\,dx}. \tag{2.28}$$

Это является представлением сл.в. $V(u)$ в виде некоторой геометрической случайной суммы (см., например, монографию Калашникова [94] относительно свойств геометрических сумм). Учитывая (2.25) — (2.27), можно переписать равенство (2.28) в более детальном виде. В действительности это быстро приводит к тому же виду результата, в котором данное решение появилось к концу семидесятых[20].

Теорема 2.5. *(1978) [104, 199, 200, 201, 157] В системе M/GI/1—EPS при выполнении условия (2.1)*

$$v(s,u) = \frac{(1-\rho)\,\mathrm{e}^{-u(s+\lambda)}}{\psi(s,u) - \lambda \int_0^u \mathrm{e}^{-x(s+\lambda)}\,\psi(s,u-x)(1-B(x))\,dx - \lambda\,\mathrm{e}^{-u(s+\lambda)} \int_u^\infty (1-B(x))\,dx}, \tag{2.29}$$

$$Re\ s > 0.$$

Можно показать, что (2.29) есть преобразование безгранично делимого распределения[21] на $[0, \infty)$. Иными словами, положительный корень m-й степени $v^{1/m}(s,u),\ \ m = 1, 2, \ldots$, есть ПЛС некоторого вероятностного распределения.

Замечание 2.8. Теорему 2.4 можно расширить на случай групповых поступлений при помощи тех же самых аргументов, которые представлены выше. Но в отличие от системы M/GI/1—FCFS, это приводит к более громоздким уравнениям и не позволяет продвинуть исследование достаточно далеко. В действительности, вид результата делает невозможным его использование в приложениях без снятия условия по $(n; x_1, \ldots, x_n)$. К сожалению, теорема 2.5 не распространяется на случай групповых поступлений. Даже $\mathsf{E}[V(u)]$ может быть представлена только в виде решения некоторого интегрального уравнения [109, p.184]. Но это решение нельзя получить в замкнутом виде при произвольном $B(x)$. Обычно упомянутое интегральное уравнение решается при накладываемом на $B(x)$ ограничении, гарантирующем, что $B(x)$ имеет экспоненциальный хвост. Впрочем, это ограничение может быть ослаблено [22].

Также крайне затруднительно распространить эту теорему на многие другие ситуации. Мы подразумеваем нетривиальное расширение, которое должно содержать новые идеи, а не стандартные технические манипуляции с формулами.

[20] Отметим, что первое издание превосходной монографии [15, Notes on p. 101] (1987) приписывает этот результат только двум другим авторам, из которых только один независимо продвинул ту же самую проблему в [157] (1984). Это неверная точка зрения, см. текст после замечания 2.13.

[21] Вероятностное распределение $V(x|u)$ на положительной полуоси является безгранично делимым тогда и только тогда, когда его ПЛС допускает представление $\int_0^\infty \exp(-sx)V(dx|u) = \exp[-\int_0^\infty (1-\exp(-sx))\nu(dx|u)]$, где $\nu(dx|u)$, the "мера Леви" вероятностного закона $V(x|u)$, есть σ–конечная мера на $[0,\infty)$, удовлетворяющая условию интегрируемости $\int_0^\infty (x \wedge 1)\nu(dx|u) < \infty$ [28, 165].

Приведем только пример тривиального расширения теоремы 2.5 на случай нескольких классов требований в системе M/GI/1—EPS (см. [208, p. 69–70] или [233, p. 43–44]). Чтобы дать представление об этом, достаточно рассмотреть только два входящих потока (индексированных посредством $m = 1, 2$) с интенсивностями λ_m, второй поток может быть объединением нескольких входящих потоков (это стандартный прием в теории приоритетных очередей). Положим $\lambda = \lambda_1 + \lambda_2$ (нижние индексы соответствуют классу требований $m = 1, 2$), $\beta_m(s)$ будет обозначать ПЛС распределения длин требований $B_m(x)$ класса m, имеющего первый момент $\beta_{1m} < \infty$. Пусть также $\rho_m = \lambda_m \beta_{1m}$ и $\rho = \rho_1 + \rho_2 < 1$. Тогда при $\beta(s) = \lambda^{-1}(\beta_1(s) + \beta_2(s))$ (это ПЛС ф.р. $B(x) = \lambda^{-1}(B_1(x) + B_2(x))$), теорема 2.5 дает ПЛС ф.р. стационарного времени пребывания требования длины u, принадлежащего классу m (нет необходимости снабжать индексом m левую часть равенства (2.29) в силу того, что случайные величины $V_m(u)$ одинаково распределены для каждого m, поскольку скорости обслуживания требований каждого класса зависят только от общего числа $n = i + j$ всех требований независимо от числа требований класса 1 (i) и класса 2 (j)). При таком обобщении формула (2.2) для ПЛС распределения безусловного времени пребывания требования класса m, очевидно, принимает вид:

$$v_m(s) = \int_0^\infty v(s, u)\, dB_m(u).$$

Приведем также аналог следствия 2.2 для системы $\mathsf{M}_2/\mathsf{GI}_2/1$—EPS:

$$\mathsf{P}_{ij} = \left(1 - \sum_{m=1}^{2} \rho_m\right) \binom{i+j}{i} \rho_1^i \rho_2^j, \quad i, j = 0, 1, 2, \ldots,$$

где $\rho_1 = \lambda_1 \beta_{11}, \quad \rho_2 = \lambda_2 \beta_{12}$.

Имеются другие факты, характеризующие метод декомпозиции на элементы задержек, но приведенные здесь результаты включают наиболее полезные утверждения и отражают особенности предмета рассмотрения. Некоторые дополнительные факты будут обсуждаться далее.

2.4. Некоторые следствия. Свойства системы M/GI/1—EPS.

В качестве следствий теоремы 2.5 получены, в частности, математическое ожидание и дисперсия распределения сл. в. $V(u)$.

Следствие 2.4. *(1969) [155]*

$$\mathsf{E}[V(u)] = u/(1 - \rho). \tag{2.30}$$

Следствие 2.5. *(1979) [105]*

$$\mathsf{Var}[V(u)] = u^2(1 - \rho)^{-2} - 2(1 - \rho)^{-1} \int_0^u \delta_1(y)\, dy, \tag{2.31}$$

где

$$\delta_1(u) = \mathsf{E}[D(u)] = u + \int_0^u (u - x) \sum_{n=1}^{\infty} \rho^n f^{n*}(x)\, dx. \tag{2.32}$$

Здесь $f^{n*}(x)$ — n-кратная свертка плотности остаточных длин $f(x) = (1 - B(x))/\beta_1$, $f^{0*}(x) \equiv \delta(x)$, где δ — δ-функция Дирака.

Равенство (2.30) хорошо известно с 1969 г. [155][22], равенство (2.31) впервые анонсировано в [104] (с небольшой ошибкой) и корректно выведено несколько раз в [105, 208, 219] с помощью решения двух систем дифференциальных уравнений, которым удовлетворяют $\mathsf{E}[\Phi(x, u)^j]$

[22] Существует не менее двух десятков различных способов вывода формулы (2.30).

и $\mathsf{E}[D(u)^j]$, $j = 1, 2$. Эти уравнения являются более простыми вариантами уравнений (2.14) и (2.15), потому что они выводятся не для ПЛС $v(s, u)$, а только для второго и первого моментов. Отправными пунктами для их вывода служат равенства (2.17) и (2.18)[23]. Решения этих систем уравнений были получены в терминах преобразований Лапласа, для которых удалось найти обратные преобразования. В частности, (2.32) есть результат инверсии следующего преобразования Лапласа: $\delta_1(s) = s^{-2}[1 - \lambda(1 - \beta(s))/s]^{-1}$ (см., например, формулу (2.63) в [208] или (A.14) в [219]). Аналогичным способом был выведен затем третий момент.

Эквивалентная формула для дисперсии была выведена из равенств (2.31) и (2.32) [105, 199, 201, 208] с помощью некоторых манипуляций.

Следствие 2.6. *(1979) [105]*

$$\mathsf{Var}[V(u)] = 2(1 - \rho)^{-2} \int_0^u (u - x)(1 - W(x))\, dx, \tag{2.33}$$

где

$$W(x) = (1 - \rho) \sum_{n=0}^{\infty} \rho^n F^{n*}(x), \tag{2.34}$$

$$F^{n*}(x) = \int_0^x F^{(n-1)*}(x - y)\, dF(y), \quad F(x) = F^{1*}(x) = \int_0^x f(y)\, dy = \beta_1^{-1} \int_0^x (1 - B(y))\, dy^{24}. \tag{2.35}$$

Здесь $F^{0*}(x) = \mathbf{1}(x)$ есть функция Хевисайда.

В (2.34) $W(x) = \lim_{t\to\infty} \mathsf{P}\{\mathsf{U(t)} \le \mathsf{x}\}$ — предельное распределение процесса незаконченной работы, т.е. сумма остаточных длин всех имеющихся в системе требований в текущий момент. Это распределение совпадает со стационарным распределением времени ожидания в системе M/GI/1—FCFS[25].

Двадцать лет спустя рекуррентные формулы для вычисления любого момента сл.в. $V(u)$ выведены в [222, 223] (1999) и [234] (2000) с помощью разложения в ряд $v(s, u)^{-1}$, где $v(s, u)$ найдено в теореме 2.5. Для получения такого представления (2.28)(см. следствие 2.8) использован прием, предложенный Хевисайдом [81] более чем сотню лет назад.

Замечание 2.9. Тем не менее, предыдущий способ вычисления моментов [208, §2.7] до сих пор представляет самостоятельный интерес, поскольку при этом проливается дополнительный

[23] Стоит отметить, что ПЛС $v(s, u)$ имеет довольно сложный вид (см. формулу (2.29)), следовательно, стандартный прямой способ вычисления моментов с помощью (2.3) срабатывает только при нахождении математического ожидания (формула (2.30)). Для вычисления $\mathsf{Var}[V(u)]$ необходимо дважды продифференцировать функцию $v(s, u)$ по s и затем сделать предельный переход при $s \to 0$. Однако это ПЛС очень трудно дифференцировать по s более одного раза. По этой причине формула (2.31) была выведена в 1979 г. с помощью уравнений, которым удовлетворяют $\mathsf{E}[\Phi(x, u)^j]$ и $\mathsf{E}[D(u)^j]$, $j = 1, 2$. Удобной ссылкой является [219], которая содержит расширения и дополнения.

[24] Ф.р. $F(x)$ имеет много различных названий. Например, она называется случайной модификацией распределения $B(x)$ или эксцессом или интегрированным хвостом времени обслуживания или прямым временем возвращения, и т.д.

[25] Побочным результатом вычисления $\mathsf{Var}[V(u)]$ в системе M/GI/1—EPS является знаменитая формула Поллачека–Хинчина. Она дает стационарное распределение времени ожидания в системе M/GI/1-FCFS. Известно несколько десятков различных способов ее получения. В дополнение к основным результатам, проведенный анализ дисциплины EPS привел также к новому выводу формулы Поллачека–Хинчина (2.34) (это же равенство в терминах ПЛС дается формулой (2.38), см. также сноску 38). Отметим, что (2.34) является одним из краеугольных камней многих областей теории очередей. Например, такое направление исследований как теория приоритетных очередей возникло в результате (нетривиального) обобщения системы M/GI/1—FCFS на случай двух классов требований с различными фиксированными приоритетами.

свет на структуру полученных формул. В частности (см. [208, формула (2.66)]),

$$\mathsf{E}[\mathsf{V}(\mathsf{u})] = \mathsf{u}/(1-\rho) = \delta_1(\mathsf{u}) + \lambda(1-\rho)^{-1}\int_0^\infty \varphi_1(\mathsf{x},\mathsf{u})(1-\mathsf{B}(\mathsf{x}))\,\mathsf{dx},$$

где $\delta_1(u)$ дается (2.32), а $\varphi_1(x,u) \doteq \mathsf{E}[\Phi(\mathsf{x},\mathsf{u})] = \delta_1(\mathsf{u}) - \delta_1(\mathsf{u}-\mathsf{x})$ при $x < u$ и $\varphi_1(x,u) = \delta_1(u)$ при $x \geq u$. Это способствовало успешному доказательству нетривиальной предельной теоремы 2.8, а также прояснило причины возникновения так называемого "парадокса", связанного с крайне простым видом формулы (2.30). В сущности, приведенная выше декомпозиция (частный случай декомпозиции (2.13) после снятия условий) заранее дает ответ на поставленный в статье [67, p. 674] (1989) вопрос о "парадоксальности" вида формулы (2.30). В [67] дано объяснение этому "парадоксу", которое можно рассматривать как дополнительное к нашему более раннему объяснению.

Замечание 2.10. В формуле (2.24) можно частично снять условия только по $(x_1,\ldots,x_n)$, получив в итоге некоторый аналог теоремы из [197]. Это дает $\mathsf{E}[\mathrm{e}^{-sV(u)}\,|n]$ для системы M/GI/1—EPS. Приведем лишь формулу для (условного) среднего времени пребывания

$$(a) \qquad \mathsf{E}[V|(u,n)] \equiv \mathsf{E}[V(u)|n] = \delta_1(u) + n\beta_1^{-1}\int_0^\infty \varphi_1(x,u)(1-B(x))\,dx.$$

Из (2.32) и (2.34) вытекает, что

$$(b) \qquad \delta_1(u) = (1-\rho)^{-1}\int_0^u W(x)\,dx.$$

Тогда с учетом формул из замечания 2.9, равенство (a) сводится к

$$(c) \qquad \mathsf{E}[V(u)|n] = u/(1-\rho) + (n-\rho/(1-\rho))\rho^{-1}\int_0^u (1-W(x))\,dx.$$

Приведем теперь эквивалентную форму равенства (2.34), которая может оказаться полезной для более глубокого изучения вероятностных законов, управляющих случайными флюктуациями времени ожидания в системе M/GI/1—FCFS при некоторых специальных классах распределений $B(x)$ с тяжелым хвостом.

Замечание 2.11. В эквивалентной форме равенство (2.34) можно представить как [26]

$$W(x) = \mathrm{e}^{-\lambda}\sum_{n=0}^\infty \lambda^n H^{n*}(x)/n!,$$

где $H(x) = H^{1*}(x) = -(\ln(1-\rho))^{-1}\sum_{n=1}^\infty \rho^n F^{n*}(x)/n$. Попутно отметим взаимосвязь констант перед знаком суммы в этой формуле для $H(x)$, $\mathsf{P}_0^\star$ в следствии 2.3 и константы в правой части равенства, приведенного в [225, p. 183] между формулой (3.6) и следствием 3.4.

Выражение (2.33) впервые позволило установить ряд новых свойств системы M/GI/1—EPS [199, 201, 208]. Здесь отметим только немногие из них. Формула (2.34) совпадает с обратным преобразованием ПЛС распределения времени ожидания в системе M/GI/1—FCFS, выполненным впервые Бенешем (Beneš) [24]. Важно подчеркнуть, что геометрические весовые коэффициенты в (2.34) равны стационарным вероятностям системы M/M/1—FCFS, которые

[26] Мы не утверждаем, что такое представление в виде сложного распределения Пуассона хорошо известной ф.р. $W(x)$ является новым, но, по крайней мере, оно нетипично в теории очередей.

не совпадают с вероятностями состояний системы M/GI/1—FCFS (но совпадают с вероятностями состояний системы M/GI/1—EPS в силу следствия 2.2). Попытки дать объяснение этого простого, но удивительного результата, предпринятые Клейнроком [108, §5.7] и Купером [49, p.219], оказались безуспешными. Однако это объяснил Прабху [144, §1.7] посредством интерпретации формулы (2.34) в терминах лестничных процессов, а затем Купер и Чен [51] успешно объяснили эту формулу с помощью результатов анализа системы M/GI/1—LCFS–P. Сравнение обеих систем FCFS и EPS позволило дать в [208, стр. 86] еще одну дополнительную интерпретацию формулы (2.34). Ключевая идея здесь состоит в совпадении распределений процессов незаконченной работы в системе M/GI/1 с дисциплинами FCFS, LCFS–P и EPS[27].

Формула (2.31) позволила установить еще одно новое свойство системы M/GI/1—EPS. Это свойство восходит к [199] (1981).

Следствие 2.7. *(1981). Если $\beta_2 < \infty$, то ф.р. сл.в. $V(u)$ всегда имеет коэффициент вариации меньше 1. Таким образом, это некоторый новый представитель класса гипоэкспоненциальных распределений*[28].

Новое доказательство теоремы 2.5 дано Шассбергером (Schassberger) [157] (1984). Его доказательство имеет принципиально новый характер и оно совершенно отлично от доказательств в [199, 201]. Главные результаты Шассбергера (в частности, вывод формулы (2.29) в другом виде) получены с помощью анализа системы M/GI/1 в дискретном времени при небольшой модификации стандартной round–robin дисциплины: вновь поступающее требование немедленно получает свой квант времени обслуживания и только после этого становится в конец очереди, если не закончится обслуживание за отведенный квант времени обслуживания (такая модификация системы M/GI/1–RR с дискретным временем предложена в [54]). Тем самым удается обойти основную трудность, связанную с тем, что порядок, в котором незавершенные требования ожидают своих следующих квантов обслуживания в классической модели RR, не сохраняется в последующих раундах (циклах) ожидания квантов. Именно это обстоятельство объясняет причины всех неудачных попыток получить точное аналитическое решение для распределения времени пребывания в системе M/GI/1–RR с непрерывным временем (см. [208,

[27] Кроме того, в то время как аналитические формулы являются точными для системы M/GI/1, они дают ориентир в более общих ситуациях, в частности, при рекуррентном входящем потоке. Например, известно [15, 163, 118, 119], что вероятностная структура стационарного времени ожидания в системе GI/GI/1—LCFS–P примерно такая же как (2.34). Весовые коэффициенты в (2.34) все еще геометрические, но с другим параметром, который не может быть вычислен в явном виде в общем случае. Кроме того, распределение времени ожидания включает в себя сумму "восходящих лестничных высот" вместо остаточных длин. Аналогичный ориентир к догадкам относительно форм асимптотики для системы GI/GI/1—EPS обеспечивается асимптотическими формулами для системы M/GI/1—EPS, которые, в свою очередь, выводятся в §2.8 и 2.10 из наших точных результатов. Следствие 3.4 является типичным примером. Статьи [90, 39] и последующие несколько десятков их "производных" относительно "жидкостных пределов сильно загруженной или перегруженной системы EPS появились благодаря этому следствию (см., например, статью [76] (2004) и ссылки в ней). Этот результат обычно приписывается статье [90] (1994), хотя Гришечкин первым доказал его для системы M/GI/1—EPS в [74] (1991) и позднее распространил его на систему GI/GI/1—EPS case (1994).

[28] Коэффициент вариации сл.в. Q, c_v, есть отношение $\sqrt{\mathsf{Var}[\cdot]}$ к среднему $\mathsf{E}[\cdot]$. Он характеризует изменчивость ф.р. $Q(x)$. Ф.р. $Q(x)$, ПЛС которой удовлетворяет $-q'(s) < (>)[1-q(s)]/[sq(s)]$, $\forall s > 0$ называется гипо(гипер)экспоненциальной. Гамма распределение при $a > 1$ (см. сноску 39), детерминированное, равномерное, усеченное нормальное являются гипоэкспоненциальными. Можно показать из свойства полной монотонности ПЛС [65, Ch.13], что $c_v < (>)1$ для гипо(гипер)экспоненциальных распределений. Однако это не является достаточным условием. Можно также любую абсолютно непрерывную ф.р. $Q(x)$ представить в виде $Q(x) = 1 - \mathrm{e}^{-\int_0^x \mu(y)dy}$, где $\mu(x) = dQ(x)/[1-Q(x)]$ называется функцией интенсивности отказов. Хорошо известное утверждение [171, p.16–19] связывет $\mu(x)$ и c_v: когда $\mu(x)$ возрастает (убывает), то $c_v < (>)1$. Можно показать с помощью аргументов, использующих свойство выпуклости, что любое рациональное распределение, полюса которого — действительные числа, является гиперэкспоненциальным. В частности, этот класс включает любую выпуклую комбинацию экспоненциальных распределений.

§1.5] или [216, Предложение 2.1]). В качестве дискретной единицы времени используется постоянный размер кванта обслуживания. Пуассоновский входящий поток аппроксимируется потоком Бернулли, $B(x)$ аппроксимируется произвольным арифметическим распределением с тем же временны́м дискретом. Одним из основных результатов [157] является производящая функция для дискретного распределения времени пребывания (условного по длине требования) в модифицированной системе RR. Также Шассбергером доказано, что последовательность стационарных длительностей пребывания в модифицированной системе RR сходится по распределению к сл.в. $V(u)$ в системе M/GI/1—EPS, когда размер кванта стремится к нулю. Вид этого предела (в терминах ПЛС) определяется формулой, которую можно редуцировать к (2.29). Позднее было также дано в [148] некоторое дополнительное обоснование к методу анализа Шассбергера и его важным достижениям.

Замечание 2.12. Фактически, до сих пор возможно различать только два разных аналитических метода для точного нахождения распределения времени пребывания в модели M/G/1—EPS: прямой метод декомпозиции на элементы задержек [104, 199, 201] (см. §2.3) и не прямой метод аппроксимации модифицированной системы RR в дискретном времени [54, 157], обсуждавшийся выше. Множество последующих статей по точным решениям моделей EPS (или их вариантов) эксплуатируют один из этих методов (или их комбинацию). Они существенно основаны на соответствующих результатах (непосредственно или в скрытой форме). Реально эти статьи имеют дело только с построением простых дополнительных приспособлений к известным нетривиальным решениям. Некоторые примеры можно найти в замечаниях 2.13 и 2.19. К сожалению, почти любое обобщение не позволяет, как правило, получить точные решения по распределениям времени пребывания в обобщенных моделях на таком глубоком уровне, который соразмерим с глубиной исследования первоначальной модели M/GI/1—EPS. Безуспешные попытки распространить теорему 2.5 на случай системы M/GI/1—EPS с групповыми поступлениями можно упомянуть в качестве примера. Кроме того, было бы более плодотворным различать фундаментальные обобщения от второстепенных. Литература по очередям с разделением процессора обильна по незначительным обобщениям. Другими словами, представляется, что как правило много статей, написанных компьютерными специалистами являются малоинтересными с математической точки зрения из–за отсутствия собственных новых идей[29].

Замечание 2.13. Статьи Отта [141] (1984) и Гришечкина [74] (1991) содержат обобщения непринципиального характера: получение $\mathsf{E}[\mathrm{e}^{-sV(u)}\, z^L]$ и $\mathsf{E}[\mathrm{e}^{-sV(u}\,|(n; x_1, \ldots, x_n)]$ для той же самой системы EPS и системы EPS с групповыми поступлениями, соответственно. Относительно последнего случая см. также замечание 2.8. В сущности, доказательства повторяют основную аргументацию [104, 199, 201] в переформулированном виде, например, в терминах теории ветвящихся процессов плюс балласт непривычной терминологии, не приспособленной к теории очередей в последнем случае [74, 75]. (Статья [75] является расширенной версией [74])[30]. Тем не менее, статья [74] содержит ряд интересных предельных теорем, которые ранее не были доказаны. Это резко контрастирует со статьей [141], в которой переформулировка и повторение аргументов [201] выполнено на более поверхностном уровне по сравнению с [74]. Первоначальные более ранние статьи [104, 199] вообще не цитируются Оттом. Его основной вклад состоит в изменении обозначений, переформулировке уравнений (2.14), (2.15) и их решению

[29] В дополнение, можно отметить следующую тенденцию. Число очень значительных достижений в теории очередей в течение промежутка между второй мировой войной и примерно до 1960—1965 г.г. было огромным. Их поток сохранился примерно до 1970—1975 г.г., и затем заметно ослабел — почти в обратной пропорции к количеству новых статей и книг, публикуемых по теории очередей. Мы знаем не так уж много столь разительных примеров диалектического перехода количества в качество.

[30] Статья Риджа и Сенгупты [146] дает более простое решение той же задачи, которое также основано на методе и результатах [199, 201] (см. §2.3 выше).

в терминах функций h и f, которые совпадают с нашими δ и $1/\delta$ при $z = 1$. В действительности, статья [141] является только некоторой "производной" от статей [104, 199, 201], а также от неопубликованной главы Яшкова про разделение процессора в монографии, цитированной Оттом во второй ссылке на p. 378 его статьи [141][31].

Отметим другую статью Риджа и Сенгупты [147] (1994), в которой раскрыты дополнительные возможности изложенного в §2.3 метода, примененного к системе M/GI/1 с дисциплиной дискриминаторного разделения процессора.

Если бы Эйсер (Asare) и Фостер (Foster) [17] (1983) ввели декомпозицию (2.13), то они смогли бы близко подойти к теореме 2.5. Asare и Foster использовали подобные нашим функционалы от ветвящихся процессов, но они смогли вычислить для них только математические ожидания. В итоге главным результатом [17] оказалось лишь равенство (c) из замечания 2.10. Тем не менее, Фостер — один из первых, кто применил понятия ветвящихся процессов (на поверхностном уровне) для анализа системы M/M/1—EPS в статье [68] (1973). Выводу (c) из замечания 2.10 в частном случае $B(x)$ типа M посвящен §6 этой статьи. В том же §6 Фостер указал на "важную нерешенную проблему": обобщение на случай произвольного распределения $B(x)$ (типа GI), сделанное позднее в [17]. Напомним, что (c) является элементарным следствием теоремы 2.4.

Ван ден Берг [25] (1990) дал еще одно доказательство теоремы 2.5. Он опирался на некоторое обобщение системы M/M/1—FCFS с одним из вариантов бернуллиевской обратной связи. Некоторые комментарии к [25] (и [148], дополняющей работу ван ден Берга) содержатся в [212, p.113–114]. Использованный метод анализа восходит к методу Шассбергера [157], искусственно модифицированному с помощью некоторых довольно громоздких конструкций из теории очередей с обратной связью и скомбинированному с нашими ключевыми результатами. Можно упомянуть книгу [58], содержащую исследование очередей с обратной связью.

Метод декомпозиции на элементы задержек, описанный в §2.3, был также применен для анализа системы M/M/1—EPS с ненадежным прибором [136] и системы M/M/1—EPS в случайной среде, управляемой процессом квазирождения и гибели [137]. Таким образом, усложненные варианты системы M/M/1—EPS были изучены в рамках метода, введенного в [104, 199, 201]. Однако, в соответствии с замечанием 2.8, теорему 2.5 не удалось распространить на эти случаи при произвольном распределении $B(x)$.

Дальнейшее развитие нашего метода позволило впервые исследовать нестационарные процессы в системе M/GI/1—EPS и найти их транзиентные распределения в терминах многомерных преобразований [208, 212, 221], [232, 225]. Эти результаты до сих пор остаются, по–видимому, недостижимыми с помощью других методов, упомянутых выше.

2.5. Дополнительно о моментах

Решение, представленное теоремой 2.5, содержит контурные интегралы Бромвича. Это затрудняет использование результата (2.29) в практических приложениях. Более простой контурный интеграл $\delta_1(s)$ также содержался в выражении для $\mathsf{Var}[V(u)]$, но для него было аналитически получено обратное преобразование. В противном случае, вычисление $\mathsf{Var}[V(u)]$ (2.31) могло бы рассматриваться в качестве примера. Следующее утверждение дает удобный вид (2.29) без контурных интегралов.

[31] См. также [206, p. 8–9] (где была объяснена суть небольшого обобщения в [141]) относительно дополнительной мотивировки утверждения про заимствованный характер результатов Отта. Учитывая, что Отт заимствовал идеи, метод и основные уравнения из [104, 105, 199, 201], включение его в авторы одного из наиболее значительных продвижений в теории очередей с разделением процессора есть только миф, который искусственно поддерживается в некоторых западных публикациях по теории очередей (см., например, сноску 20), несмотря на очевидные факты.

Следствие 2.8. *(1999) Эквивалентной формой (2.29) (без контурных интегралов Бромвича) является*

$$\frac{1}{v(s,u)} = \sum_{n=0}^{\infty} \frac{s^n}{n!} \xi_n(u), \tag{2.36}$$

где

$$\xi_0(u) = 1, \quad \xi_n(u) = \frac{n}{(1-\rho)^n} u^{n-1} * W^{(n-1)*}(u), \quad n = 1, 2, \ldots. \tag{2.37}$$

Здесь $W^{(n-1)*}(u)$ есть $(n-1)$-кратная свертка стационарного распределения времени ожидания $W(u)$ (см. (2.34) в знакомой системе M/GI/1-FCFS ($W^{0*}(u) = \mathbf{1}(u)$, $W^{1*}(u) = W(u)$), ПЛС ф.р.$W(u)$ дается хорошо известной формулой Поллачека–Хинчина

$$w(q) = \frac{1-\rho}{1-\rho f(q)}, \tag{2.38}$$

где $f(q) = (1-\beta(q))/(q\beta_1)$ есть ПЛС ф.р.$F(x)$ по x (аргумент q), введенное в (2.35) (плотность $F(x)$ обозначается через $f(x)$, см. следствие 2.5).

Доказательство. Перепишем (2.29) в виде, напоминающем [157, Th. 3.2] (см. также (5.5) в [206]), а именно

$$v(s,u) = \frac{(1-\rho)\delta(s,u)}{1-\rho\delta(s,u)\left[\int_0^u \frac{dF(x)}{\delta(s,u-x)} + (1-F(u))\right]}, \quad \mathrm{Re}\ s \geq 0. \tag{2.39}$$

Здесь $F(x)$ введено в (2.35). напомним, что $\mathsf{E}\left[\mathrm{e}^{-sF}\right] = (1-\beta(s))/(s\beta_1)$. Для достижения нашей цели применим ПЛ по u от $1/\delta(s,u)$ u (аргумент q), которое находится из (2.27) как $\tilde{\psi}(s, q-s-\lambda)$, $\quad s \geq 0,\ q > s + \lambda - \lambda\pi(s)$ (ср. также с третьей строкой на p.8 в [206]). Теперь после простых преобразований получаем следующее разложение в степенной ряд ПЛ от функции $1/v(s,u)$, $\quad s \geq 0,\ u \geq 0$

$$\begin{aligned}\int_0^\infty \mathrm{e}^{-qu} \frac{1}{v(s,u)}\, du &= \frac{1}{q}\left[1 + \frac{1}{1-\rho}\frac{s}{q}\frac{1}{1-\frac{1}{1-\rho}\frac{s}{q}w(q)}\right] \\ &= \frac{1}{q}\left[1 + \sum_{n=1}^{\infty}\left(\frac{1}{1-\rho}\frac{s}{q}\right)^n w(q)^{n-1}\right]\end{aligned} \tag{2.40}$$

где $w(q)$ is given by (2.38). Отметим, что $\left|\frac{sw(q)}{(1-\rho)q}\right| < 1$ когда $q > s + \lambda - \lambda\pi(s),\ \rho < 1$. Легко получить обратное преобразование (по аргументу q) каждого члена степенного ряда по s в (2.40). Результат дается (2.37), что приводит к (2.36), правая часть которого есть степенной ряд по s с коэффициентами $\xi_n(u)/n!$. □

Идея такого подхода восходит к Хевисайду [81, Chapter 5]. Этот результат получен в [222, 223] (1999) и также в [233, 234] (2000). Однако отправной пункт доказательства в [233, 234] является другой формой равенства (2.29) в теореме 2.5, повторенной в [141] в других терминах, а именно, равенство (5.6) при $z = 1$ в [206]; кроме того, наше доказательство не использует информации о $\mathsf{Var}[V(u)]$. В действительности, вид $\mathsf{Var}[V(u)]$ (см. (2.33) в следствии 2.6) наводит на догадку о возможности такого разложения в ряд.

Замечание 2.14. Формула для $W^{n*}(x)$ в (2.37) модет быть представлена в следующем виде

$$W^{n*}(x) = (1-\rho)^n \sum_{k=0}^{\infty} \binom{k+n-1}{n-1} \rho^k F^{k*}(x).$$

Это делается, к примеру, посредством обращения $w(q)^n$, где $w(q)$ дается равенством (2.38).

Замечание 2.15. Мы снова заметим, что побочным результатом анализа является формула Поллачека–Хинчина (2.34) для ф.р. $W(x)$, ПЛС которой дается (2.38). Однако анализ системы EPS дает другую величину (соответствующую не вероятнстной мере) $W^\circ(x) = W(x)/(1-\rho)$. Вид ПЛС $W^\circ(x)$ хорошо известен: $w^\circ(q) = \sum_{n=0}^{\infty} \rho^n f^n(q)$. В отличие от $W(x)$, $W^\circ(x)$ корректно определено для всех $\rho > 0$ и $x > 0$. Можно показать, что $W^\circ(x) < \infty$ для всех $\rho > 0$, $x > 0$ и для любой $B(\cdot)$ (несмотря на то, что при $\rho \geq 1$, $W^\circ(x) \to \infty$ as $x \to \infty$).

Следствие 2.9. *(1999) Если $v_n(u) = \mathsf{E}[V(u)^n]$, $n = 1, 2, \ldots$, то справедлива следующая рекурсивная формула:*

$$v_n(u) = \sum_{i=1}^{n} \binom{n}{i} v_{n-i}(u)\xi_i(u)(-1)^{i+1} \tag{2.41}$$

Доказательство. Поскольку $v(s,u)$ есть аналитическая функция по s (в частности, близ $s = 0$), можно использовать разложение в ряд Тейлора we can use the Tailor series expansion of $v(s,u)$ при малых $s > 0$

$$v(s,u) = 1 - \frac{s}{1!}v_1(u) + \frac{s^2}{2!}v_2(u) - \frac{s^3}{3!}v_3(u) + \ldots \tag{2.42}$$

Произведение рядов (2.42) и (2.36) дает

$$\begin{aligned} -\frac{s}{1!}[v_1(u) - \xi_1(u)] + \frac{s^2}{2!}[v_2(u) - 2v_1(u)\xi_1(u) + \xi_2(u)] \\ - \frac{s^3}{3!}[v_3(u) - 3v_2(u)\xi_1(u) - 3v_1(u)\xi_2(u) - \xi_3(u)] + \ldots = 0 \end{aligned}$$

и это приводит к (2.41) после n–кратного дифференцирования по s и перехода к пределу при $s \to 0$. □

Формула (2.37) показывает, что that $\xi_1(u) = \mathsf{E}[V(u)]$ в (2.30). Легко проверить, что выражения for первых двух моментов $V(u)$ совпадают с (2.30) и (2.33).

Формула (2.41) в следствии 2.9 выведена в [222, 223] и [233, 234]. По—видимому, наш вывод (2.41) короче.

Замечание 2.16. Следствие 2.9 важно для приложений, так как оно приводит к решению проблемы моментов.

2.6. Некоторые частные случаи.

Ниже мы обсудим два частных случая распределения времени пребывания, когда $B(x)$ специфицировано. В частности, сконцентрируем внимание на случаях экспоненциального и детерминированного распределений $B(x)$.

Случай 1. Сначала рассмотрим простейший случай этого типа, когда $B(x) = 1 - \mathrm{e}^{-\mu x}$. Тогда ПЛС ф.р.$B(x)$ имеет вид $\beta(s) = \mu/(s+\mu)$ и $\beta_1 = 1/\mu$. Кроме того, мы можем также анализировать родственный случай, когда система M/M/1 управляется дисциплиной со случайным порядком обслуживания (RO). Прямым следствием теоремы 2.5 является

Следствие 2.10. *Для системы M/M/1—EPS при $\rho < 1$ справедливо следующее выражение для ПЛС (условного) распределения времени пребывания*

$$v(s,u) = \frac{(1-\rho)(1-\rho r^2)\exp(-u(s+\lambda-\lambda r))}{(1-\rho r)^2 - \rho(1-r)^2\exp(-\mu u(1-\rho r^2)/r)}, \quad Re\ s > 0, \tag{2.43}$$

где $r = \pi(s)$ есть решение уравнения (2.4), в котором $\beta(s) = \mu/(s+\mu)$.

Доказательство. См. [208, с.73-75], [220] для деталей. Основные этапы: функциональное уравнение (2.4) сводится к квадратому уравнению, которому удовлетворяет r; уравнение (2.15) упрощается, позволяя в итоге преобразовать равенство (2.29) к виду

$$v(s,u) = \frac{(1-\rho)\,\mathrm{e}^{-u(s+\lambda)}}{\psi(s,u) + \frac{1}{\mu}\frac{\partial\psi(s,u)}{\partial u}};$$

формула (2.27) сводится к

$$\tilde{\psi}(s,q) = \frac{q+s+\mu}{(q-q_1)(q-q_2)},$$

где $q_1 = -\lambda r$, $q_2 = -\mu/r$ есть два простых полюса функции $\tilde{\psi}(s,q)$.

Остается выполнить обращение последней формулы по комплексному аргументу q с помощью классической теоремы Коши о вычетах. Предварительно следует вычислить два вычета функции $f(s,q) = \tilde{\psi}(s,q)\,\mathrm{e}^{qu}$ (относительно q), которые расположены в $q_1 = -\lambda r$ и $q_2 = -\mu/r$ комплексной q–плоскости. В итоге, это приводит к утверждению (2.43) после некоторых выкладок. □

Формула (2.43) впервые получена в [44] (1970) (см. также [109, Ch. 4]). Метод доказательства в [44] основан на использовании процессов рождения и гибели для описания процесса обслуживания; технически он довольно громоздок (приходится, например, решать дифференциальные уравнения с частными производными первого порядка) и непригоден для анализа более общих случаев. Некоторые простые варианты такого подхода описаны также в недавних книгах Асмуссена [15, Ch. 3, §9], Джагермана [84] и Рольски [150, pp. 64–65] (естественно, только для случая M/M/1—EPS). Отметим, что (2.43) можно также вывести из более общих результатов статьи [197, Th. 1, p. 47] после выполнения предельного перехода, отражающего устремление к нулю размера кванта. При этом приходится решать только два обыкновенных дифференциальных уравнения. Однако эта техника не переносится на систему M/GI/1—EPS.

Замечание 2.17. Сенгупта и Джагерман [161] нашли альтернативное выражение для ПЛС распределения времени пребывания, условному только по числу требований, которые видимы при поступлении нового требования.

Формула (2.33) для дисперсии времени пребывания $V(u)$ сводится к виду

$$\mathrm{Var}[V(u)] = \frac{2\rho u}{\mu(1-\rho)^3} - \frac{2\rho}{\mu^2(1-\rho)^4}\left[1 - \mathrm{e}^{-\mu u(1-\rho)}\right]. \tag{2.44}$$

Используя (2.43), можно получить интегральное представление преобразования $v(s)$ (см. (2.2)) и затем выполнить обращение этой формулы. Такое обращение было впервые сделано Моррисоном [129] (1985) для случая $\mu = 1$ (см. также [208, §2.6, pp. 75–78] (1989)). Обращение выполняется посредством деформирования контура интегрирования в комплексной плоскости для ПЛ хвоста ф.р.

$$\mathsf{P}(V > x) = \frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} \frac{\mathrm{e}^{sx}}{s}[1 - v(s)]ds$$

в контур вокруг разреза между двумя точками ветвления. Это приводит к единственному интегральному представлению для $\mathsf{P}(V > x)$, а именно,

$$\mathsf{P}(V > x) = \frac{2}{1-\rho}\int_0^\pi \frac{\mathrm{e}^{-\gamma\left[2\sqrt{\rho}-(1+\rho)\cos\gamma\right]/[(1-\rho)\sin\gamma] - \mu x(1-\rho)^2/\left(1+\rho-2\sqrt{\rho}\cos\gamma\right)}}{1 + \mathrm{e}^{-\pi\left[2\sqrt{\rho}-(1+\rho)\cos\gamma\right]/[(1-\rho)\sin\gamma]}}\sin\gamma\,\mathrm{d}\gamma. \tag{2.45}$$

Отметим, что выражение (2.45) переформулировано здесь для общего случая, когда $\beta_1 = 1/\mu$, а не $\mu = 1$.

Случай обслуживания в случайном порядке. Позднее результат подобный (2.45) получен Флатто [66] для системы M/M/1 с дисциплиной обслуживания в случайном порядке (RO: Random Order). При дисциплине RO, требования обслуживаются в случайном порядке: всякий раз, когда прибор становится свободным, следующее требование выбирается случайным образом (равновероятно) из присутствующих в очереди требований. Пусть W_{RO} есть стационарное время ожидания в системе M/M/1—RO. Тогда

$$\mathsf{P}(W_{RO} > x) = \frac{2(1-\rho)}{\rho} \int_0^\pi \frac{\mathrm{e}^{[2\phi(\theta)-\theta]\cot\theta}}{\mathrm{e}^{\pi\cot\theta}+1} \frac{\mathrm{e}^{-\left[1-2\rho^{-1/2}\cos\theta+\rho^{-1}\right]\lambda \mathrm{x}}}{\left(1-2\rho^{-1/2}\cos\theta+\rho^{-1}\right)^2} \sin\theta \mathrm{d}\theta, \tag{2.46}$$

где

$$\phi(\theta) = \arctan\frac{\sin\theta}{\cos\theta - \rho^{1/2}}, \quad 0 \le \phi(\theta) \le \pi. \tag{2.47}$$

В [66] время нормализовано так, что требования поступают с единичной интенсивностью, но здесь это переформулировано для общего случая.

Формулы (2.45) и (2.46) эквивалентны в следующем смысле

$$\rho\mathsf{P}(V_{EPS} > x) = \mathsf{P}(W_{RO} > x) = \rho\mathsf{P}(W_{RO} > x | W_{RO} > 0). \tag{2.48}$$

Соотношение эквивалентности (2.48) установлено с помощью различных методов в ряде работ (см., например, [47,208,218]). Здесь мы отметим два способа доказательств. Доказательство опирается на тот факт, что время пребывания в системе GI/M/1—EPS равно по распределению времени ожидания требования, поступающего в не пустую систему RO. Это важное утверждение сделано Коэном [47], который, стартуя от результатов Рамасвами [145], показал, что преобразование распределения задержки удовлетворяет в обоих случаях одному и тому же дифференциальному уравнению (с единственным решением). См. для деталей [208, Th. 2.8]. Это приводит к (2.48). Другими словами, время пребывания V_{EPS} тесно связано с условным временем ожидания $W_{RO}|W_{RO} > 0$ в системе GI/M/1 queue через простую мультипликативную константу, а именно,

$$\mathsf{P}(\mathsf{V_{EPS}} > \mathsf{x}) = \frac{1}{\eta}\mathsf{P}(\mathsf{W_{RO}} > \mathsf{x}), \quad \mathsf{x} \ge 0,$$

где $\eta = \mathsf{P}(W_{RO} > 0)$. Отметим, что η совпадает с вероятностью $\mathsf{P}(L > 0)$, где L есть число требований в системе GI/M/1 в моменты поступлений, и η сводится к ρ в случае пуасссоновского входа [145, 47, 208].

Опишем теперь суть второго способа. Выражение (2.48) вытекает после переформулировки (2.45) к виду (2.46) в духе преобразований, напоминающих преобразования из [129] или [208, pp. 75–77]. Чтобы это сделать, можно использовать некоторые простые факты из комплексного анализа и тригонометрии. Конечно, можно стартовать от (2.46) и переписать эту формулу к виду (2.45).

Более глубокая связь между разделением процессора и случайным порядком обслуживания наблюдалась в [201, Remark 2].

Насколько нам известно, интегральные формулы наподобие (2.45) или (2.46) существуют только для системы M/M/1 при дисциплинах EPS и RO. Конечно, эти точные интегральные выражения для $\mathsf{P}(V > x)$ слишком сложны для прямых приложений. Это хороший пример, когда необходимы некоторые аппроксимации (см. конец §2.10).

Случай 2. Рассмотрим теперь второй частный случай, когда длины требований имеют детерминированное (вырожденное) распределение

$$B(x) = \begin{cases} 0, & 0 \le x < u, \\ 1, & x \ge u. \end{cases}$$

Следовательно, ПЛС этой ф.р. имеет вид $\beta(s) = \mathrm{e}^{-su}$ с моментами $\beta_i = u^i$, $i = 1, 2, \ldots$. Загрузка равна $\rho = \lambda u < 1$. В этом частном случае ф.р. условного и безусловного времени пребывания совпадают, значит можно использовать $V = V(u)$ для обозначения стационарного времени пребывания в этой системе M/D/1—EPS. Ясно, что $v(s,u) = v(s)$. Справедливо

Следствие 2.11. *В системе M/D/1—EPS при $\rho = \lambda u < 1$ ПЛС распределения времени пребывания имеет явный вид*

$$v(s) = v(s,u) = \frac{(1-\rho)(s+\lambda)^2 \mathrm{e}^{-u(s+\lambda)}}{s^2 + \lambda[s + (s+\lambda)(1-\rho)]\mathrm{e}^{-u(s+\lambda)}}, \quad Re\ s > 0. \tag{2.49}$$

Доказательство. См. [208, с.73], [220] для деталей. В доказательстве удобно использовать уравнение (2.28) (ср. с равенством (2.33) из [208]). В нашем случае упомянутое уравнение сводится к виду

$$v(s) = v(s,u) = \frac{(1-\rho)\delta(s,u)}{1 - \lambda\delta(s,u)\int_0^u \frac{dx}{\delta(s,u-x)}}, \tag{2.50}$$

где

$$\delta(s,t) = \frac{s+\lambda}{\lambda + s\mathrm{e}^{t(s+\lambda)}}, \quad t \le u. \tag{2.51}$$

Чтобы получить (2.51), легче использовать уравнение (2.20), которому удовлетворяет неизвестная функция $\delta(s,u)$ (ср. с уравнением (2.29) в [208]), чем находить обратное преобразование функции $\tilde{\psi}(s,q)$, которая дается (2.27). В нашем случае уравнение (2.20) сводится к виду

$$\frac{\partial\delta(s,t)}{\partial t} + (s+\lambda)\delta(s,t) - \lambda\delta(s,t)^2 = 0 \tag{2.52}$$

с дополнительным условием $\delta(s,0) = 1$. Это уравнение Бернулли. Оно сводится к линейному уравнению после деления каждого члена на $\delta(s,t)^2$ и замены переменной $1/\delta(s,t) = u$. Решение уравнения (2.52) дается (2.51). Подстановка (2.51) в (2.50) приводит к окончательному утверждению (2.49). □

Решение для $v(s)$ (см. следствие 2.11) было также получено в [141] (1984) и [25] (1990) с помощью более громоздкой (точнее: довольно искусственной) и менее удобной аргументации. Напомним, что более общий результат для системы M/GI/1—EPS (теорема 2.5) был известен к тому времени, когда статья [141] (1984) была опубликована, см., например, [104] (1978) или [199] (1981). Выше только сущность некоторых наших выкладок конца семидесятых годов была описана.

Замечание 2.18. Выражение для дисперсии времени пребывания $V(u) = V$ (см. (2.31) или (2.33)) сводится к виду [208, (2.70) на p. 82], [212, p. 115]

$$\mathrm{Var}[V(u)] = \frac{u^2}{(1-\rho)^2} - \frac{2u^2(\mathrm{e}^\rho - 1 - \rho)}{\rho^2(1-\rho)}.$$

Недавно, стартуя от наших более ранних результатов, Шалмон [164] вывел формулу замечания 2.18 с помощью некоторого (нестандартного) представления стационарного времени пребывания как случайного блуждания с ветвлением. Его подход интересен и, по–видимому, имеет дополнительные применения помимо частного случая, рассмотренного в цитированной статье.

2.7. Некоторые границы

Следствие 2.7 служит ориентиром для некоторых уточнений и, в частности, для получения верхней и нижней границ распределения времени пребывания (и его моментов). Эти границы легко вывести с помощью использования следствий 2.5 и 2.8. Из (2.37) вытекает, что $\xi_0(u) = 1$ и

$$\frac{u^n}{1-\rho} \le \xi_n(u) \le \left(\frac{u}{1-\rho}\right)^n, \quad n = 1, 2, \ldots$$

Эти неравенства и (2.36) позволяют установить

Следствие 2.12. *(2006) [42, 229]*

$$\mathrm{e}^{-su/(1-\rho)} \le v(s,u) \le \frac{1-\rho}{\mathrm{e}^{su}-\rho} \le \frac{1}{1+su/(1-\rho)}. \tag{2.53}$$

Следствие 2.12 означает, что мы нашли некоторые границы для ф.р. случайной величины $V(u)$ такие, что эта сл.в. лежит в некотором смысле между детерминированной и экспоненциально распределенной случайными величинами с одинаковыми средними $\mathsf{E}[\mathsf{V}(\mathsf{u})]$. Неравенства (2.53) вводят стохастический порядок (по преобразованиям Лапласа) для ф.р. сл.в. of $V(u)$ и, в частности, приводят к следующему уточнению следствия 2.7

Следствие 2.13. *(2006) [42, 229] Если ф.р. $B(x)$ имеет конечный второй момент, то ф.р. сл.в. $V(u)$ принадлежит $\mathcal{L}$–классу распределений*[32].

Другими словами, сл.в. $V(u)$ имеет ф.р. $\mathsf{P}(V(u) \le x | B = u)$, для которой $\int_0^\infty \mathrm{e}^{-sx}(1 - \mathsf{P}(V(u) \le x | B = u))dx \ge \mathsf{E}[V(u)]/(1 + s\mathsf{E}[V(u)])$ (в сущности, это еще одно определение $\mathcal{L}$–класса распределений). Следствие 2.13 позволяет найти точные (tight) границы распределения хвоста сл.в. $V(u)$, различные оценки моментов и т.д., см. [229, 42]. В частности, имеет место следующее утверждение:

Следствие 2.14. *(2006)[229,42] Следующие инвариантные верхняя и нижняя границы имеют место для системы M/GI/1—EPS при фиксированном β_1 :*
(a) $\mathsf{P}(V(u) > x) \le C_2^{-1}(u)$ *при* $x > u/(1-\rho)$,
(b) $\mathsf{P}(V(u) > x) \ge 1 - C_2^{-1}(u)$ *при* $x \le u/(1-\rho)$,
где $C_2(u) = (x(1-\rho)/u)^2 - 2x(1-\rho)/u + 2$. *Верхняя (a) и нижняя (b) оценки являются острыми (sharp), когда* $x(1-\rho)/u \ge 2 + \sqrt{2}$ *и* $2 - \sqrt{2} \le x(1-\rho)/u \le 1$, *соответственно.* □

Некоторые другие свойства сл.в.$V(u)$.

Сл.в. $V(u)$ имеет коэффициент вариации $c_v \le \sqrt{\rho}$.

Также $v_j(u) \Leftrightarrow \rho < 1$. Кроме того, $\mathsf{E}[V^j] < \infty \Leftrightarrow \beta_j < \infty$. Здесь $v_j(u)$ и V определяются посредством (2.3) и (2.2).

[32] Различные классы распределений в теории надежности и в моделях технической эксплуатации и ремонта детально обсуждаются, например, в [171]. Эта книга содержит также определение $\mathcal{L}$–класса распределений в терминах стохастического порядка по ПЛ. Дадим адаптированное определение. Скажем, что $V(u) \in \mathcal{L}$–классу распределений, если для некоторой экспоненциально распределенной сл.в. $V^e(u)$, такой что $\mathsf{E}[V(u)^e] = \mathsf{E}[V(u)]$, выполняется стохастический порядок по ПЛ $V(u) \ge_L V^e(u)$, т.е. $v(s,u) \le v^e(s,u)$ для всех $\mathsf{Re}\ s \ge 0$. См. также [106] и цитированные там статьи относительно $\mathcal{L}$–класса распределений. Например, $\mathcal{L}$–класс шире класса HNBUE.

2.8. Свойства монотонности.

Следующее свойство монотонности было найдено для $\mathsf{Var}[V(u)]$ [105,199,208]. Пусть индексы M, E_m и H_m, соответственно, обозначают экспоненциальное, эрланговское (порядка m) и гирерэкспоненциальное (порядка m) распределения $B(\cdot)$ с одним и тем же (фиксированым) первым моментом β_1. Справедливо

Следствие 2.15. *(1979) [105,199,208] Когда $\rho < 1$, то следующие неравенства имеют место для системы M/GI/1—EPS*

$$\mathsf{Var}_{E_m}[V(u)] \le \mathsf{Var}_M[V(u)] \le \mathsf{Var}_{H_m}[V(u)]. \tag{2.54}$$

Идея доказательства. (См. [199, 208] для деталей.) Вывод неравенства (2.54) основан на установлении связи между распределениями случайных величин $V(u)$ в системах EPS (с распределениями длин из классов E_m, M, H_m) в терминах отношений стохастического порядка типа $\overset{(2)}{\le}$ (см. сноску 16). При этом решающую роль играет формула (2.33). □

Приведем еще одно свойство монотонности ф.р. сл.в. $V(u)$ в системе EPS.

Следствие 2.16. *(1979) Вероятность $\mathsf{P}(V(u) > t)$ не убывает по $u \ge 0$. Отсюда вытекает, что каждый момент $v_j(u) = \mathsf{E}[\mathsf{V(u)^j}]$, $j \in \mathsf{N}$ не убывает по u.*

Доказательство. На всех реализациях управляющих последовательностей в системе M/G/1—EPS точка выхода требования длины $u_2 > u_1$ находится правее точки выхода требования длины u_1 $(u_1, u_2 \in \mathbf{R}_+)$. □

Недавно формула (2.33) позволила заметить еще один факт монотонности в системе EPS:

Следствие 2.17. *(2006) [189, Th. 3.1] $\mathsf{Var}[V(u)]/u$ строго монотонно возрастает по x при любом $B(\cdot)$.*

Доказательство. Продифференцируем $\mathsf{Var}[V(u)]/u$. В результате: $\frac{d}{du}(\mathsf{Var}[V(u)]/u) > 0$ for all $u > 0$. □

Напомним, что $\mathsf{E}[V(u)]/u = 1/(1-\rho)$ постоянна при любом $u > 0$.

Замечание 2.19. Кроме сказанного в замечании 2.12, следствие 2.17 дает пример тривиального дополнительного приспособления к (2.33). Конечно, это следствие имеет более элементарный характер, чем результаты, отмеченные в замечании 2.13 или даже сама формула (2.33). Было намного труднее получить (2.33), чем эксплуатировать эту формулу. Тем не менее, подобные "достижения" типичны для многих недавних статей западных компьютерных специалистов. Можно привести несколько десятков других примеров, но мы предпочитаем вообще не отмечать тривиальные результаты.

2.9. Асимптотика

Стоит отметить, что точное выражение для дисперсии времени пребывания содержит интегральный член, затрудняющий точные вычисления. Поэтому мы рассмотрим способ упрощения вычисления $\mathsf{Var}[V(u)]$ в системе M/GI/1—EPS. Его суть состоит в получении асимптотических оценок дисперсии при малой и большой длине требования. Используя разложение правой части (2.33) в ряд Тейлора в точке $u = 0$ и ограничиваясь первыми четырьмя членами разложения, приходим к утверждению.

Следствие 2.18. *[105, 199, 204, 25] Если выполняется условие (2.1), то*

$$\mathsf{Var}[V(u)] \sim \rho(1-\rho)^{-2}u^2 - \lambda(3(1-\rho))^{-1}u^3, \quad u \to 0. \tag{2.55}$$

Предположим, что:

(a) существуют первые три момента распределения $B(x)$ ($\beta_i,\ i = 1,2,3$). Интерпретируем $W(x)$ в (2.34) как распределение времени жизни обрывающегося процесса восстановления с распределением (несобственным) промежутков времени между последовательными моментами восстановления $G(x) = \rho F(x)$ ($G(0) = 0,\ G(+\infty) = \rho < 1$). Здесь $F(x)$ приведено в (2.35) и дефект $(1-\rho)$ представляет собой вероятность обрыва $(1-\rho)$. В дополнение к предположению (a): $\beta_i < \infty,\ i = 1,2,3$, будем считать, что:

(b) распределение $B(x)$ обладает следующим свойством: существует некоторая константа $\Theta > 0$ такая, что

$$\int_0^\infty \mathrm{e}^{\Theta x}\, dG(x) = \lambda \int_0^\infty \mathrm{e}^{\Theta x}(1 - B(x))\, dx = 1$$

и

$$C = \lambda \int_0^\infty \mathrm{e}^{\Theta x}\, x(1 - B(x))\, dx < \infty.$$

Введем обозначения $r_1 = \int_0^\infty (1-W(y))\, dy = \rho(1-\rho)^{-1} f_1$, где $f_j = \beta_{j+1}((j+1)\beta_1)^{-1},\ j = 1,2,\ldots$ и $r_2 = \int_0^\infty y(1-W(y))\, dy = 2(1-\rho)^{-1}\left(f_2 + 2\rho(1-\rho)^{-1} f_1^2\right)$.

Следствие 2.19. *(1981) [199, 204] Если выполняются условия (a), (b) выше и $\rho < 1$, то для системы M/GI/1—EPS*

$$\mathsf{Var}[V(u)] \sim 2(1-\rho)^{-2}\left(r_1 u - r_2 + (1-\rho)\,\mathrm{e}^{-\Theta u}/[C\Theta^3]\right), \quad u \to \infty. \tag{2.56}$$

Доказательство. (Первое доказательство было дано в [199], см. также [204] для уточнений или [208, p. 82–85].) В предположении (a) перепишем (2.33) в терминах r_1 и r_2

$$\mathsf{Var}[V(u)] = 2(1-\rho)^{-2}\left[r_1 u - r_2 + \int_u^\infty (x-u)(1-W(x))\, dx\right], \tag{2.57}$$

где r_1 и r_2 приведены выше. Равенство (2.57) есть третья эквивалентная форма $\mathsf{Var}[V(u)]$[33]; первые две формы даются равенствами (2.31) и (2.33).) Дополнительно к (a) введем теперь предположение (b). В силу следствия из теоремы 2 из [65, Ch. 11, §6,] имеет место следующая оценка для $1 - W(x)$, когда $x \to \infty$

$$1 - W(x) \sim (1-\rho)\,\mathrm{e}^{-\Theta x}/[C\Theta],$$

которая совпадает со знаменитой асимптотической оценкой Крамéра для риска разорения в обобщенном пуассоновском процессе. Тогда последний член в правой части равенства (2.57) при $u \to \infty$ асимптотически равен $(1-\rho)\,\mathrm{e}^{-\Theta u}/[C\Theta^3]$ при $u \to \infty$. Это приводит к (2.56). □

Отметим, что при $B(x) = 1 - \mathrm{e}^{-\mu x}$ формула (2.56) становится точной для всех $u \geq 0$ и может быть преобразована к виду

$$\mathsf{Var}[V(u)] = 2\rho u/[\mu(1-\rho)^3] - 2\rho[1 - \mathrm{e}^{-\mu(1-\rho)u}]/[\mu^2(1-\rho)^4].$$

[33] Восемь лет спустя та же самая формула (2.57) была повторно получена Ави–Ицхаком и Халфиным в [19, (44) on p. 998], стартуя от (2.33). Теперь (2.57) иногда приписывается им или кому–либо еще.

Более грубая по сравнению с (2.56) асимптотическая оценка дисперсии

$$\mathsf{Var}[V(u)] \sim 2(r_1 u - r_2)/(1-\rho)^2 \qquad \text{при } u \to \infty$$

получена ранее в [105,200]. Десять лет спустя тот же самый результат получил повторно ван ден Берг [25]. Отметим, что точность этой формулы, а также асимптотики (2.55) достаточно высока (см. [199,208] и [25] для численных примеров).

Замечание 2.20. Во многих приложениях вполне можно пользоваться даже еще более грубой по сравнению с приведенными выше асимптотической оценкой дисперсии:

$$\mathsf{Var}[V(u)] \sim 2r_1 u/(1-\rho)^2 \qquad \text{при } u \to \infty$$

при условии, что $\beta_2 < \infty$. Эта формула эквивалентна следующему асимптотическому разложению второго момента ф.р. сл.в. $V(u)$ в системе M/GI/1—EPS

$$v_2(u) = v_1^2(u) + \lambda\beta_2(1-\rho)^{-3}u + o(u), \quad u \to \infty.$$

Главный член асимптотики совпадает с точной формулой для дисперсии длительности (стандартного) периода занятости $\Pi(u)$, открываемого требованием длины u в системе M/GI/1 (ПЛС ф.р. сл.в. $\Pi(u)$ приведено в равенстве(2.5)). Напомним, что формула для $\mathsf{Var}[V(u)]$ в системе M/GI/1—LCFS–P имеет тот же самый явный вид (ср. с замечанием 2.1), т.е.

$S\,\mathsf{Var}[V(u)]_{LCFS-P} = \lambda\beta_2(1-\rho)^{-3}u.(2.3),(2.5).$

Анализ остаточного члена асимптотического разложения в замечании 2.20 проведен в следствии 2.19, которое дает, в частности, скорость сходимости $\int_u^\infty (x-u)(1-W(x))\,dx$ к нулю при $u \to \infty$ в (2.57).

Замечание 2.21. Из (2.57) и выражения для $\mathsf{Var}[V(u)]_{LCFS-P}$ в замечании 2.20 вытекает, что

$$\mathsf{Var}[V(u)]_{EPS} \le \mathsf{Var}[V(u)]_{LCFS-P}$$

при любом $B(\cdot)$, см. также [201].

Позднее некоторые родственные аппроксимации $\mathsf{Var}[V(u)]_{EPS}$ были рассмотрены в [8,185].

2.10. Связь с теорией страхового риска

Условия (a), (b), введенные между следствиями 2.18 и 2.19, известны как условия Крамера–Лундберга в теории страхового риска[34]. В более общем случае условие (a) формулируется как

$$f(-\Theta) = \int_0^\infty \mathrm{e}^{\Theta x}\,dF(x) = \rho^{-1} < \infty, \quad \Theta > 0. \tag{2.58}$$

Это эквивалентно следующему предположению: существует производящая функция моментов (экспоненциальный момент) распределения $F(x)$. Другими словами, ПЛС

$$f(s) = (1-\beta(s))/(s\beta_1) < \infty$$

[34] Другое название этой теории — теория разорения. Теория разорения как математическая дисциплина началась с работы Лундберга (1903). Эта теория изучает вероятностные законы, которым подчиняются случайные флюктуации резервного капитала страховой компании. Основной вклад в теорию разорения был сделан шведской вероятностной школой.

существует в непустой окрестности нуля (у $f(s)$ абсцисса абсолютной сходимости $s_b < 0$), откуда вытекает, что хвост $1-F(x)$ (а следовательно, и $1-B(x)$) экспоненциально ограничен: существуют $K < \infty$, $\epsilon > 0$ и $x_0 \geq 0$, такие, что

$$1 - B(x) \leq K\,\mathrm{e}^{-\epsilon x}, \quad x \geq x_0. \tag{2.59}$$

Это означает, что большие длины требований очень маловероятны (с экспоненциально малыми вероятностями).

Таким образом, условие (2.58) выделяет во множестве всех функций распределения $B(x)$ класс распределений с так называемым легким хвостом. Ключевые примеры: гамма–распределения, ф.р. Вейбулла $B(x) = 1 - \exp(-cx^b)$ при $c > 0$ и $b \geq 1$ и любая ф.р. с ограниченным носителем). Условие (2.58) не выполняется, когда $f(s)$ имеет существенную особенность в нуле ($f(-\epsilon) = \infty$ при $\forall \epsilon > 0$). Это выделяет класс $\mathcal{K}$ распределений на $(0, \infty)$ с так называемым тяжелым хвостом (heavy-tailed). Ключевые примеры: ф.р. Вейбулла при $c > 0$ и $b \in (0, 1)$, логнормальное распределение, семейство распределений Парето ($B(x) = 1-(b/(x+a))^\nu$, $b, \nu > 0$), и т.д. Распределения из класса $\mathcal{K}$ имеют ряд патологических свойств, основные из которых: отсутствие некоторых (или всех) старших моментов, отсутствие характеризации в терминах ПЛС, более медленное, чем экспоненциальное, убывание хвоста. Известны различные подклассы таких распределений и о некоторых из них мы напомним. Самым узким является класс ф.р. $B(x)$ ($B(x) < 1$ при $\forall x \geq 0$) с правильно меняющимся (regularly varying) хвостом. Класс всех правильно меняющихся функций обозначается как $RV(-a)$:

$$B(x) \in RV(-a),\ a \geq 0,\ \text{если}\ \lim_{x\to\infty} \frac{1 - B(tx)}{1 - B(x)} = t^{-a},\ \forall t > 0. \tag{2.60}$$

Из (2.60) вытекает, что $1 - B(x) \sim x^{-a}\ell(x)$, где $\ell(x)$ — медленно меняющаяся (на бесконечности) функция, т.е. положительная измеримая функция, для которой $\ell(tx)/\ell(x) \to 1$ при $x \to \infty$ и $\forall t > 0$. Примеры $\ell(x)$: функции, сходящиеся к положительной константе, константы и (повторные) логарифмы (например, $\ln(\ln(\mathrm{e}+x))$), и т.д. Отметим, что для $\forall \epsilon > 0 \quad \exists x_0$, такое, что

$$x^{-\epsilon} \leq \ell(x) \leq x^{\epsilon}, \quad x > x_0.$$

Стоит упомянуть, что распределения Вейбулла и логнормальное не являются правильно меняющимися.

$\mathcal{K}$ оказался слишком широким классом для разработки теории, полезной для применений. Наиболее успешно применяется подкласс $\mathcal{S} \subset \mathcal{K}$, называемый классом субэкспоненциальных распределений. Класс $\mathcal{S}$ ($1 - B(x) > 0$ при $x \geq 0$), введен впервые В.П.Чистяковым в [Ch64?] (1964) в контексте ветвящихся процессов и изучен в [18] (1972) (см. также [28, 34]). (В частности, идея характеризации поведения хвоста распределений в (2.61) через асимптотическое свойство сверток принадлежит Чистякову.)

Класс $\mathcal{S}$ определяется как:

$$B(x) \in \mathcal{S},\ \text{если}\ \lim_{x\to\infty} \frac{1 - B^{n*}(x)}{1 - B(x)} = n, \quad n \geq 2. \tag{2.61}$$

Если условие в (2.61) выполняется для некоторого $n \geq 2$, то оно справедливо для всех $n \geq 2$.

В эквивалентной форме (2.61) можно переписать как

$$\mathsf{P}(B_1 + \ldots + B_n > x) \sim \mathsf{P}(\max\{B_1, \ldots, B_n\} > x) \sim n(1 - B(x)),$$

где $B_1, \ldots, B_n$ — независимые копии сл.в. B с ф.р. $B(x)$. Мы видим, что хвост частного (partial) максимума случайных величин существенно определяет хвост частной суммы тех же самых сл.в.

Если $B(x) \in \mathcal{S}$, то $\lim_{x\to\infty} \mathrm{e}^{\epsilon x}(1-B(x)) = \infty$, $\epsilon > 0$, т.е. неравенство (2.59) не выполняется. Полезным свойством $\mathcal{S}$ является: для $\forall \epsilon > 0$ существует положительная константа $K(\epsilon) < \infty$, такая, что

$$\frac{1-B^{n*}(x)}{1-B(x)} \le K(\epsilon)(1+\epsilon)^n$$

справедливо при $n \ge 1$ и $x \ge 0$.

Некоторые другие свойства $\mathcal{S}$:

если $B(x) \in \mathcal{S}$ и $1-B(x) \sim 1-G(x)$, то $G(x) \in \mathcal{S}$;

если $B(x) \in \mathcal{S}$ и $1-G(x) = o(1-B(x))$ при $x \to \infty$, то $B(x) * G(x) \in \mathcal{S}$, причем $1-B(x)*G(x) \sim 1-B(x)$;

если $B(x) \in \mathcal{S}$ и $1-G(x) = c(1-B(x))$ при $x \to \infty$ и $c > 0$, то $G(x) \in \mathcal{S}$, $B(x) * G(x) \in \mathcal{S}$, причем $1-B(x)*G(x) \sim (1+c)(1-B(x))$.

Отметим еще, что класс $\mathcal{S}$ не замкнут относительно операции свертки, т.е. если B_1 и B_2 являются независимыми членами класса $\mathcal{S}$, то $B_1 + B_2$ может не принадлежать $\mathcal{S}$.

Обозначим через $\mathcal{L}_\circ$ класс ф.р. $B(x)$ с длинным хвостом (long-tailed), который определяется асимптотическим соотношением

$$1-B(x-y) \sim 1-B(x) \text{ при } x \to \infty \text{ и всех } y > 0.$$

Из (2.60) следует, что $B(x) \in \mathcal{L}_\circ$ тогда и только тогда, когда $1-B(\log x)) \in RV(0)$, где $RV(0)$ — класс медленно меняющихся функций. Основное свойство $\mathcal{L}_\circ$: если сл.в. $B \in \mathcal{L}_\circ$ и Y — любая неотрицательная сл.в., не зависящая от B, то $\mathsf{P}(B-Y>x) \sim \mathsf{P}(B>x)$.

Между введенными классами вероятностных распределений выполняются следующие отношения включения $RV(-a) \subset \mathcal{S} \subset \mathcal{L}_\circ \subset \mathcal{K}$.

Более подробные сведения по классам ф.р. с тяжелым хвостом и их свойствам можно найти в [28], [62], [165] и частично в [15, 65, 94, 142, 166]. Важным является следующий факт. Неизвестно, как распознать некоторую ф.р. $B(x) \in \mathcal{S}$ по виду ее ПЛС $\beta(s)$, так как не существует характеризации класса $\mathcal{S}$ в терминах ПЛС. Класс $\mathcal{S}$ характеризуется в других терминах, см., например, [28, p.430][35].

Однако исходная информация и результаты решения задач теории очередей, как правило, формулируются в терминах ПЛС. Поэтому нам необходимо дополнительно применять нетрадиционные инструменты при изучении очередей с распределениями из классов $RV(-a)$[36] or $\mathcal{S}$:

[35] В частности, асимптотическая эквивалентность $1-W(x)$ и $1-F(x)$ в (2.64) — одна из характеризаций класса $\mathcal{S}$: $W(x) \in \mathcal{S} \Leftrightarrow F(x) \in \mathcal{S}$.

[36] Когда $B(x) \in RV(-a)$, известна, в частности, характеризация асимптотического поведения $1-B(x)$ в терминах остаточного члена разложения в ряд Тейлора ПЛС $\beta(s)$ в точке $s \downarrow 0$, см. теорему ниже или Theorem 8.1.6 в [28, pp. 333–334]. Такая характеризация называется иногда тауберовой теоремой Бингхема и Дони. Она впервые доказана в [29, Th. A] (1974) в контексте суперкритических процессов Гальтона–Ватсона.

Теорема. *Предположим, что сл.в. Q имеет n моментов $q_1, \dots, q_n$ и $q_0 = 1$. Пусть $q(s)$ есть ПЛС функции распределения сл.в Q. Определим*

$$d_n(s) = (-1)^{n+1}\left[q(s) - \sum_{i=0}^{n}(-s)^i/i!\right].$$

При $n < a < n+1$, $n = 1, 2, \dots$ и константе $C > 0$, следующие утверждения эквивалентны:

$$d_n(s) = (C+o(1))s^a \ell(1/s), \quad s \downarrow 0,$$

различные варианты теоремы Карамата при $B(x) \in RV(-a)$ [95] (эта теорема тесно связана с тауберовой теоремой Харди—Литтлвуда), другие неклассические тауберовы и абелевы теоремы, нетривиальные асимптотические разложения, разработанные в асимптотической теории экстремальных порядковых статистик [70], теорию больших уклонений, а также ряд новых результатов из продвинутой в последние десятилетия теории правильно меняющихся функций, называемой теорией де Хаана [37], и т.д.

Это — область очень активных исследований, потому что теория очередей очень часто имеет дело только с вероятностными распределениями, имеющими легкий хвост. Но пространственно-временна́я динамика потоков трафика в современных компьютерных сетях описывается распределениями со степенными хвостами и, как правило, бесконечной дисперсией (так называемый *the Noah effect* в информатике). Кроме того, убывание автокорреляционных функций процессов трафика имеет гиперболический, а не экспоненциальный характер (феномен "самоподобности" (*self-similar*) или "фрактальности" (*fractal*), другой термин — *the Joseph effect*). Такие неожиданные эффекты впервые обнаружены только в 1993 г. в сети ETHERNET.

Ранее подобные случаи изучались в теории страхового риска, однако переформулировка результатов этой теории на системы обслуживания может оказаться трудной. Относительно просто получаются результаты в той ситуации, когда $B(x) \in RV(-a)$ и $a \notin \mathbf{N}$. Но случаи, когда $a = 1, 2, \ldots$, сложнее. Как правило, они изучаются с помощью теории де Хаана. Те же результаты часто (но не всегда) оказываются справедливыми и в более общих ситуациях, когда, например, $B(x) \in \mathcal{S}$.

Замечание 2.22. Приведем пример применения теоремы из сноски 36. Де Мейер и Тейгелс [125] изучили хвост распределения периода занятости в системе M/GI/1, когда $B(x) \in RV(-a), a \geq 1$. Их стартовой точкой было уравнение (2.4). Они применили тауберову теорему Бингхема и Дони, чтобы доказать

Теорема 2.6. *(1980) [125] Справедлива следующая эквивалентность: при $\rho < 1$, $a \geq 1$ и $x \to \infty$,*

$$\mathsf{P}(B > x) \sim x^{-a}\ell(x) \Leftrightarrow \mathsf{P}(\Pi > x) \sim (1-\rho)^{-a-1}x^{-a}\ell(x).$$

Теорема покрывает случай, когда $a = 1, 2, \ldots$, при доказательстве которого дополнительно используется теория де Хаана. Это — наиболее тяжелые этапы доказательства и поэтому в приложениях теоремы обычно исключаются степенные показатели $a = 2, 3, \ldots$ и $a = 1$. Мы видим, что поведение хвостов случайных величин B и Π практически одинаково, если ограничить свое внимание правильно меняющимися хвостами. В виду (2.1), следует предполагать в приложениях теории очередей, что, как минимум, $a > 1$, поэтому $\beta_1 < \infty$, и, когда $a < 2$, $\beta_2 = \infty$. В эквивалентной форме эта теорема записывается как

$$\mathsf{P}(\Pi > x) \sim (1-\rho)^{-1}\mathsf{P}(B > (1-\rho)x), \quad x \to \infty. \tag{2.62}$$

$$1 - Q(x) = (C + o(1))(-1)^n x^{-a}\ell(x)/\Gamma(1-a), \quad x \to \infty,$$

где $\Gamma: (0,\infty) \to \mathbf{R}_+$ *есть Гамма функция, см. сноску 39 для ее определения.*

Случай, когда a есть целое, более сложен. Важный пример применения этой теоремы содержится в замечании 2.22. По существу, эта теорема является одним из вариантов формального обоснования для класса $RV(-a)$ первого принципа операционного исчисления, установленного Хевисайдом [81] в более общем случае, см., например, [3]. См. также [59, p.254] относительно условий, накладываеых на $B(x)$, при которых принципы Хевисайда оправдываются.

[37] Теория правильно меняющихся функций часто называется теорией Карамата (J.Karamata) в честь J.Karamata [95], который первым дал характеризацию таких функций. Эта теория имеет дело с изучением асимптотических соотношенияй вида $h(tx)/h(x) \to g(t) = O(t^{-a}), \quad x \to \infty$ при $\forall t > 0$. Теория де Хаана (L.de Haan) изучает более общие соотношения вида $(h(tx) - h(x))/h(x) \to k(t) \in \mathbf{R}, \quad x \to \infty$ при $\forall t > 0$.

Теорема 2.6 обобщена в [233] (2001) на случай системы GI/GI/1—FCFS при $B(x) \in RV(-a)$. Заметим, что в этом и последующем случаях константа $(1-\rho)^{-1}$ в (2.62) заменяется $\mathsf{E}[N]$, т.е. средним числом требований, обслуженных в течение периода занятости. Позднее асимптотика (2.62) распространена на случай $B(x) \in \mathcal{S}$ при некоторых дополнительных условиях, гарантирующих гладкость ф.р. $B(x)$, в частности, $\mathsf{P}(B > x) \sim \mathsf{P}(B > x - \sqrt{x}), x \to \infty$ [93]. Этот класс включает распределения Вейбулла с параметром $b < 1/2$. Хвосты функций распределения, рассмотренных в [93], тяжелее, чем $\mathrm{e}^{-\sqrt{x}}$. Как показано в [16], это условие является критическим для того, чтобы асимптотика (2.62) имело место. Обсуждение подробностей этого вопроса не входит в наши цели.

Следует подчеркнуть, что важность распределений с тяжелыми хвостами была обнаружена в теории страхового риска на несколько десятилетий раньше, чем в информатике и теории очередей. Другой механизм разорения возникает при таком классе распределений, и это порождает необходимость более точной оценки вероятности разорения по сравнению с возможностями классических схем. Отметим, что в контексте теории страхового риска вероятность разорения за бесконечное время (модель Крамéра–Лундберга с бесконечным горизонтом) в случае положительных страховых сумм и премий есть $1 - W(x)$, где ф.р. $W(x)$ дается (2.34) [38] (это справедливо для случая положительных размеров страховых сумм и положительной премиальной ставке, см. [62,176]). До сих пор даже простая формула (2.34) (или ее копия в терминах ПЛС (2.38)) позволяет находить новые асимптотические свойства системы M/G/1—FCFS, когда ф.р. $B(x) \in \mathcal{S}$. Некоторым аналогом формулы для вероятности разорения $1 - W(x,t)$ за конечный промежуток времени для модели риска с конечным горизонтом (в терминах двойных преобразований) служит решение уравнения Такача в теории очередей (см. равенство (3.47)).

2.11. Другие асимптотические формулы. Предельные теоремы.

Приведем результат, который анонсирован в [233,234] и доказан в [216]. Он позволяет изучать случай, когда условие Крамера–Лундберга не выполняется. Другими словами, равенство (2.58) не справедливо и $\beta_2 = \infty$.

Следствие 2.20. *(2000) [233, 234, 216] Пусть $B(x) \in RV(-a)$ (см. определение (2.60)) при $a \in (1,2)$. Тогда в системе M/GI/1—EPS при $\rho < 1$*

$$\mathsf{Var}[V(u)] \sim \frac{2\Gamma(2-a)}{\Gamma(4-a)} \frac{\lambda}{(1-\rho)^3(a-1)} u^{3-a}\ell(u), \quad u \to \infty, \tag{2.63}$$

где $\ell(u)$ — медленно меняющаяся (на бесконечности) функция и Γ есть обычная эйлеровская Гамма функция[39].

[38] Х.Крамéр в 1930 г. нашел преобразование Фурье ф.р. $W(x)$, а ПЛС этой ф.р. было независимо получено Ф.Поллачеком (1930) и А.Я.Хинчиным (1932) в контексте теории очередей. Формула (2.34) может быть также интерпретирована как распределение максимума процесса случайного блуждания с отрицательным сносом (см. также замечание 2.11). Попутно заметим, что ф.р. $W(x)$ в равестве (2.34) имеет еще одно представление в виде классического интегрального уравнения Вольтерра второго рода

$$W(x) = 1 - \rho + \lambda \int_0^x [1 - F(x-y)]W(y)\, dy \quad \text{for } x \geq 0.$$

[39] Действительнозначная функция $\Gamma: (0,\infty) \to \mathbf{R}_+$, определяемая как $\Gamma(a) = \int_0^\infty x^{a-1}\mathrm{e}^{-x}\, dx, \quad a > 0$, называется Гамма функцией. Интегрирование по частям показывает, что $\Gamma(a+1) = a\Gamma(a)$ при $a > 0$ и $\Gamma(n+1) = n!$ при $n = 0, 1, \ldots$. Также $\Gamma(1/2) = \sqrt{\pi}$.

Доказательство. При $1-B(x)\sim x^{-a}\ell(x),\ \ a\in(1,2)$ из утверждения (d) предложения A3.8 в [62] вытекает, что $\beta_2=\infty$, следовательно, теорема 2.19 и замечание 2.20 не применимы в этом случае. В силу леммы 1 из [65, Ch. 5, §6] (или леммы 8.1.5 из [28]) существует момент $\beta_{1+\varepsilon}<\infty$ порядка $1+\varepsilon,\ \varepsilon\in(0,1)$, который позволяет получить искомое асимптотическое приближение из формулы (2.33) с помощью двух теорем. По теореме Коэна–Пэйкса [28, Th.8.10.3], эквивалентной варианту II теоремы Крамéра–Лундберга в теории страхового риска, [62, Th.1.3.8],

$$1-W(x)\sim\rho(1-\rho)^{-1}(1-F(x)),\quad x\to\infty, \tag{2.64}$$

причем обе функции распределения $W(x)$ и $F(x)$ принадлежат классу $\mathcal{S}$[40]. Так как $RV(-a)\subset\mathcal{S}$, то подставляя в (2.64) формулу (**??**) и применяя один из вариантов теоремы Карамата ([95], [28, Prop.1.5.8], [65, Ch. VIII, §9])[41], правая часть асимптотического равенства (2.64) редуцируется к виду $\lambda(1-\rho)^{-1}(a-1)^{-1}x^{1-a}\ell(x)$. Теперь предел при $u\to\infty$ интеграла в (2.33) после подстановки $x=ut$ ($\ell(ut)\sim\ell(t)$ при $u\to\infty$ и $\forall t$) сводится к бета–интегралу (интегралу Эйлера первого рода) $B(p,q)=\int_0^1(1-t)^{p-1}t^{q-1}\,dt$ при $p=2$ и $q=2-a$. Множитель перед бета–интегралом стал асимптотически равен $2\lambda(1-\rho)^{-3}(a-1)^{-1}u^{3-a}\ell(u)$. Осталось вспомнить, что бета–интеграл выражается через Гамма функцию как $B(p,q)=\Gamma(p)\Gamma(q)/\Gamma(p+q)$,. Это приводит к (2.63). □

Замечание 2.23. Следствие 2.20 дополняет следствие 2.19 в случае с тяжелыми хвостами.

Аналогичный подход позволяет легко получить асимптотические оценки и старших моментов ф.р. сл.в. $V(u)$, когда $B(x)\in RV(-a)$ в соответствующем диапазоне a.

Это делает возможным изучать асимптотическое поведение систем обслуживания при некоторых дополнительных условиях в форме предельных теорем. Многие из них получены в условиях высокой загрузки (когда $\rho\ to 1$ в случае однолинейных систем). Имеется два подхода к анализу систем обслуживания в таких условиях. Первый подход связан с асимптотическим анализом явных формул или уравнений, описывающих (стационарные) случайные процессы, такие как число требований в моментt или время пребывания в системе. Первые предельные теоремы в теории очередей о нормированных надлежащим образом случайных величинах L (стационарное число требований) или W (стационарное время ожидания) в условиях высокой

Ф.р. $Q(x)=\frac{\mu^a}{\Gamma(a)}\int_0^x y^{a-1}\mathrm{e}^{-\mu y}dy$ is называется гамма распределением с параметрами $a>0$ и $\mu>0$. Также оно при $a=2,3,\dots$ называется эрланговским распределением и при $a<1$ — гирерэкспоненциальным распределением. В этих случаях оно символически обозначается через E_a и H_a, соответственно. Если $a=1$, то $Q(x)=1-\mathrm{e}^{-\mu x}$.

Отметим, что a может быть комплексной переменной. Тогда комплекснозначная гамма функция может быть определена как преобразование Меллина функции e^{-x} посредством той же самой формулы как выше при $\mathsf{Re}\ a>0$. Однако такая гамма функция не используется в статье.

Интеграл с параметром $\Gamma(a,z)=\int_z^\infty x^{a-1}\mathrm{e}^{-x}\,dx,\quad a>0$ есть неполная Гамма функция. Здесь $z\notin(-\infty,0)$ может быть комплексной переменной. По определению, $\Gamma(a,0)\equiv\Gamma(a)$.

[40] Отметим, что (2.64) принципиально отличается от асимптотической формулы Крамéра (см. формулу для $1-W(x)$ в доказательстве следствия 2.19) в том смысле, что оценку Крамéра нельзя получить из (2.64) с помощью формального распространения (2.64) на класс ф.р. с легким хвостом.

[41] В сущности, теорема Карамата утверждает, что интегралы от правильно меняющихся функций снова правильно меняются и тем же свойством обладают урезанные моменты ф.р. из класса RV. В частности, при $x_0>0$ таком, что $\ell(x)$ локально ограничена на $[x_0,\infty)$), имеют место:

$$(a)\qquad \int_{x_0}^x t^a\ell(t)\,dt\sim\ell(x)\int_{x_0}^x t^a\,dt\sim(a+1)^{-1}x^{a+1}\ell(x)\quad\text{при } x\to\infty\quad\text{и } a>-1,$$

$$(b)\qquad \int_x^\infty t^a\ell(t)\,dt\sim-(a+1)^{-1}x^{a+1}\ell(x)\quad\text{при } x\to\infty\quad\text{и } a<-1.$$

загрузки были получены именно таким образом. Подмеченные закономерности оказались справедливыми и в более общих условиях, когда явные формулы были уже недоступны. Поэтому соответствующие результаты получали с помощью второго подхода, связанного с изучением на основе теории слабой сходимости предельного поведения случайных процессов, возникших в рассматриваемых системах. Ю. В. Прохоровым было установлено, что в основе явлений, возникающих в системах обслуживания в условиях высокой загрузки, лежит принцип инвариантности (ныне называемый принципом Донскера–Прохорова). Этот принцип инвариантности дает возможность использовать винеровский процесс (броуновское движение) для аппроксимации довольно сложных процессов, порожденных суммами случайных величин [34,114,165,188,192]. Во многих исследованиях в качестве предельного процесса фигурирует винеровский процесс или его многомерные аналоги.

Однако применение общих методов для исследования поведения систем обслуживания с разделением процессора в условиях высокой загрузки сталкивается с серьезными трудностями, так как появляющиеся процессы имеют значительно более сложную структуру по сравнению с процессами, возникающими в классических системах обслуживания возникающих процессов по сравнению с классическими системами обслуживания. По–видимому, отсутствие процессов простой структуры в очередях с разделением процессора не обещает значительного дальнейшнго прогресса в применении второго подхода к таким системам обслуживания. Появление в некоторых случаях новых предельных распределений (см., например, [96], [212, Th. 6.7], [226, Th. 2.2], [192]) также не способствует использованию общих методов для исследования предельного поведения очередей с разделением процессора.

Используя результаты §2.3 и §2.4, мы можем получать предельные теоремы для системы EPS значительно проще с помощью первого подхода. Далее приводятся некоторые из них.

Теорема 2.7. *В системе M/GI/1—EPS*

$$\lim_{\rho\uparrow 1}\mathsf{P}(\mathsf{L}(1-\rho)/\rho<\mathsf{x})=1-\exp(-\mathsf{x}),\ \mathsf{x}>0. \tag{2.65}$$

Доказательство. См. [210,74]. □

То же самое утверждение справедливо и для системы M/GI/1—FCFS, но при дополнительном предположении, что $\beta_2<\infty$. Это хорошо известный результат, восходящий к классическим статьям Кингмана, см., например, [103] или [46, Ch.7, §2]. Случай $\beta_2=\infty$ (например, когда $B(x)\in RV(-a),\ 1<a<2$) значительно сложнее и только с конца последнего века наметился прогресс в доказательствах соответствующих предельных теорем для системы FCFS. Приведенный результат для дисциплины EPS верен и в случае $\beta_2=\infty$.

Следующее утверждение анонсировано в [210, 212] (1990) и доказано в [74, 160, 213, 214] с помощью различных приемов. Доказательство в [213, 214], по–видимому, наиболее простое. Оно основано на разложении в ряд Тейлора в малой окрестности точки $s=0$ комплексной плоскости формулы (2.29) для $\mathsf{E}[\mathrm{e}^{-\mathsf{sV(u)}}]$. Эта техника довольно стандартна, если не принимать во внимание силу самой теоремы 2.5. Однако успех доказательства зависел также от нетривиальной формулы в замечании 2.9, представляющей собой неожиданное декомпозиционное свойство хорошо известного первого момента $\mathsf{E}[V(u)]=u/(1-\rho)$.

Теорема 2.8. *(1990) [210, 212, 213, 74, 160] Если $u\in[0,\infty)$ фиксировано, то справедливо*

$$\lim_{\rho\uparrow 1}\mathsf{P}(\mathsf{V(u)}(1-\rho)/\mathsf{u}<\mathsf{x})=1-\exp(-\mathsf{x}),\ \mathsf{x}\ge 0 \tag{2.66}$$

для системы M/GI/1—EPS.

Доказательство. (См. [213, 214] для дополнительных деталей и [227] для дальнейших расширений.) Далее будут использоваться обозначения

$$\delta_j(u) = \lim_{s\downarrow 0}(-1)^j \frac{\partial^j \delta(s,u)}{\partial s^j}, \quad \varphi_j(x,u) = \lim_{s\downarrow 0}(-1)^j \frac{\partial^j \varphi(s,x,u)}{\partial s^j}, \quad j = 1,2,\ldots$$

в дополнение к обозначениям теорем Theorems 2.4 и 2.5. Положим $\varepsilon = 1-\rho$, $\varepsilon << 1$. Заменяя s в (2.28) на εs и используя разложение в ряд Тейлора в точке εs при малой $\varepsilon > 0$, получаем

$$v(\varepsilon s, u) = \frac{\varepsilon\left[1 - \varepsilon s\delta_1(u) + \frac{\varepsilon^2 s^2}{2!}\delta_2(u) - \ldots\right]}{1-(1-\varepsilon)\left[1 - \varepsilon s\overline{\varphi}_1(u) + \frac{\varepsilon^2 s^2}{2!}\overline{\varphi}_2(u) - \frac{\varepsilon^3 s^3}{3!}\overline{\varphi}_3(u) + \ldots\right]}, \tag{2.67}$$

где

$$\overline{\varphi}_j(u) = \beta_1^{-1}\int_0^\infty \varphi_j(x,u)(1-B(x))\,dx, \quad j = 1,2,\ldots. \tag{2.68}$$

Учитывая следствие 2.9, равенство (2.67) сводится к

$$v(s,u) = \frac{1 - \varepsilon s\delta_1(u) + O(\varepsilon^2)}{1 + su - \varepsilon\left[s\delta_1(u) + \frac{1-\varepsilon}{2!}s^2\overline{\varphi}_2(u)\right] + O(\varepsilon^2)} \tag{2.69}$$

в точке εs равномерно по $s \in (0,\infty)$.

Из (2.69) вытекает, что

$$\lim_{\varepsilon\downarrow 0} v(\varepsilon s, u) = 1/(1+su).$$

Мы узнаем ПЛС функции распределения $1 - \mathrm{e}^{-x/u}$ в правой части последнего равенства. Это завершает доказательство. □

Теперь мы в состоянии изучить предельное поведение безусловного распределения времени пребывания. Общий вид ПЛС ф.р. (безусловного) времени пребывания V (см. (2.2) ничего не говорит об этом. Но из теоремы 2.8 вытекает

Следствие 2.21. *(1992) [160, 213]*

$$\lim_{\rho\uparrow 1} v(s(1-\rho)) = \int_0^\infty \frac{1}{1+su}\,dB(u). \tag{2.70}$$

Доказательство. [160, 213]. Результат непосредственно следует из (2.2), теоремы 2.8 и теоремы Лебега о мажорируемой сходимости (для обоснования изменения порядка операций интегрирования и перехода к пределу). См. также [214]. □

Спустя семь лет доказательства теоремы 2.8 и следствия 2.21 были повторены в [233, 234] с незначительными различиями.

Замечание 2.24. Формула (2.70) есть так называемая свертка Меллина–Стилтьеса (см. [28]) двух независимых функций распределения: экспоненциальной со средним 1 и $B(\cdot)$. Другими словами, предельное распределение нормированной должным образом случайной величины V является произведением двух независимых случайных величин, функции распределения которых указаны выше.

Замечание 2.25. Отметим, что свертка Меллина–Стилтьеса (обозначаемая через $\overset{S}{*}$) некоторых функций распределения $A(x)$ и $B(x)$ $(x \in \mathbf{R}_+)$ совпадает с обычной стилтьесовской сверткой (обозначаемой через $*$) функций распределения $A(\mathrm{e}^x)$ и $B(\mathrm{e}^x)$. Другими словами,

$$C(x) \doteq A \overset{S}{*} B(x) = \int_0^\infty A(x/y)dB(y) = A * B(\mathrm{e}^x).$$

Комбинирование этих результатов с (2.70) позволяет использовать хорошо известные свойства произведения двух случайных величин (см, например, Феллер [65]) с целью нахождения явных выражений для предельных распределеий нормированного времени пребывания. Это имеет практический интерес, в особенности, для случая распределений $B(x)$ с тяжелыми хвостами (точнее, в случае субэкспоненциальных $B(x)$, например, с правильно меняющимися хвостами на бесконечности). В частности, предложение 3 в Брейман (Breiman) [38] (1965), которое спрятано в [28, Theorem 8.15.3], говорит, что если сл.в. B имеет ф.р., хвост которой удовлетворяет $1 - B(x) \sim x^{-\alpha}\ell(x),\ \alpha > 0, x > 0$, где $\ell(x)$ есть медленно меняющаяся функция (на бесконечности), и если A есть другая неотрицательная сл.в., независимая от B, и удовлетворяющая $\mathsf{E}[A^\gamma] < \infty$ при некотором $\gamma > \alpha$, то

$$\mathsf{P}(AB > x) \sim \mathsf{E}[A^\alpha]\mathsf{P}(B > x), \quad x \to \infty. \tag{2.71}$$

Хотя первоначальное утверждение предложения Бреймана предполагает случай $0 < \alpha < 1$, его доказательство может быть распространено на случай $\alpha > 1$.

Благодаря (2.71), выражение (2.70) хорошо подходит для практических вычислений в случае высокой загрузки. Например, если $B(x)$ имеет распределение Парето, то предельное распределения времени пребывания в системе M/GI/1—EPS при $\rho \uparrow 1$ принадлежит классу "смесь по Парето экспоненциальных распределений" (PME: Pareto Mixtures of Exponentials), первоначально введенному в [3] для другого случая. Если скомбинировать формулы (2.70) и (2.71) с, результатами, скажем, из [3] или [165, Ch. 4], то можно легко получить много новых выражений в явном виде.

Рассмотрим континуум систем M/GI/1—EPS и для каждой из них будем наблюдать стационарное время пребывания требования длины u. Эти наблюдения дают процесс $\{V(u) : u \geq 0\}$. Из доказательств теорем 2.4 и 2.5 вытекает, что этот процесс имеет стационарные и независимые приращения (но они не одинаково распределены). Этот факт отмечен также в [67]. Пусть $\mathcal{N}(x)$ — ф.р. нормального закона со средним 0 и дисперсией 1.

Теорема 2.9. *(1991) [74] Если $\rho < 1$ и $\mathsf{Var}[V(u)] < \infty$, то в системе M/GI/1—EPS*

$$\lim_{u\to\infty} \mathsf{P}\left(\frac{V(u) - u(1-\rho)^{-1}}{(\lambda\beta_2 u)^{1/2}(1-\rho)^{-3/2}} \leq x\right) = \mathcal{N}(x). \tag{2.72}$$

Теорема 2.10. *(1991) [74] Если $\rho = 1$ и $\mathsf{Var}[V(u)] < \infty$, то в системе M/GI/1—EPS*

$$\lim_{u\to\infty} \mathsf{P}\left(\lambda\beta_2 V(u)/u^2 \leq x\right) = G_{1/2}(x), \tag{2.73}$$

где $G_{1/2}(x) = 2[1-\mathcal{N}(x^{-1/2})],\quad x > 0$ есть ф.р. устойчивого закона с индексом (показателем) 1/2 и ПЛС $\mathrm{e}^{-\sqrt{2s}}$.

Эти две теоремы восходят к Прабху (Prabhu) [144, Ch. 3, §4], который доказал их аналоги в контексте стохастических процессов теории запасов[42], но для другой системы, интерпретируемой как система M/GI/1—LCFS–P в теории очередей. Однако, те же самые результаты справедливы для системы M/GI/1—EPS [74,213,216].

Существует подобный результат для предельного поведения случайной величины $V_n(u, x_1, \ldots, x_n)$, которая представлена равенством (2.13). Этот аналог теоремы 2.9 был доказан в [74] с помощью вероятностной техники и повторно доказан в [187], используя как вероятностную технику, так и аналитически с помощью преобразований (посредством преобразований Лапласа и Фурье в последнем случае).

[42] См. также книгу Рубальского [152], которая посвящена проблемам управления стохастическими процессами хранения.

Теорема 2.11. *(1991) [74, 187] Если $\rho < 1$ и $\beta_2 < \infty$, то*

$$\lim_{u\to\infty} \mathsf{P}\left(u^{-1/2}[V_n(u, x_1, \ldots, x_n) - u(1-\rho)^{-1}] \le x\right) = \mathcal{N}^\circ(x), \tag{2.74}$$

где $\mathcal{N}^\circ(x)$ есть ф.р. нормального закона со средним ноль и дисперсией $\rho\beta^2/(1-\rho)^3$.

Напомним, что точное решение для ПЛС (безусловного) распределения времени пребывания в системе M/GI/1—EPS дается равенством (2.2), где $v(s,u)$ представлено формулой (2.29).

Теорема 2.12. *(2000) [233, 234] Следующие утверждения эквивалентны для системы M/GI/1—EPS as $B(x) \in RV(-a)$, $a > 1$, $a \notin \mathbf{N}$ and $\rho < 1$:*

(i) $1 - B(x) \sim x^{-a}\ell(x)$ for $x \to \infty$;

(ii) $\mathsf{P}(V > x) \sim (1-\rho)^{-a}x^{-a}\ell(x)$ for $x \to \infty$.

В простых терминах, теорема 2.12 означает, что хвост функции распределения сл.в. V так же тяжел как хвост функции распределения сл.в. B, т.е.

$$\mathsf{P}(V > x) \sim \mathsf{P}(B > (1-\rho)x), \qquad x \to \infty. \tag{2.75}$$

Равенство (2.75) называетсяis called *редуцированной эквивалентностью по загрузке* (RLE: Reduced Load Equivalence). По–видимому, первая догадка про возможность подобной эквивалентности (в ослабленной форме) принадлежит Клейнроку (см. [109, p. 175]). В действительности, до сих пор объяснение Клейнрока используется для комментариев к RLE в ряде недавних статей. Вторым примером является теорема 2.6 в замечании 2.22, которую можно интерпретировать как RLE для времени пребывания в системе M/GI/1—LCFS–P в силу замечания 2.1. Мы отсылаем читателя к [35] за списком других ссылок.

Стартовой точкой для аналитического доказательства теоремы 2.12 является следствие 2.8 теоремы 2.5. Доказательство впервые выполнено в [234](2000). В основу доказательства положены результаты §2.3 и §2.4 (в частности, формула для дисперсии $\mathsf{Var}[V(u)]$) и некоторые результаты §2.5, а также тауберова теорема Бингхема и Дони, упомянутая в сноске 36. Теперь известны различные обощения теоремы 2.12, в особенности, на случаи, когда $B(x)$ принадлежит некоторым подклассам $\mathcal{S}$, а именно, субэкспоненциальным вогнутым распределениям. Точнее, если $\mathsf{P}(B > x - \sqrt{x}) \sim \mathsf{P}(B > x)$ при $x \to \infty$, то (2.75) справедливо [91, 92]. Обычно используются вероятностный метод доказательств в обобщениях теоремы 2.12, который, как правило, основывается на теории больших уклонений и границе Чернова [40].

Асимптотика для высших моментов сл.в. $V(u)$ находится из следствия 2.9 тем же самым способом, которым были выведены формулы в замечании 2.20 из (2.33). Например,

$$v_n(u) = v_1^n(u) + n(n-1)(1-\rho)^{-n}r_1u^{n-1} + o(u^{n-1}),\ u \to \infty,\ n = 1, 2, \ldots, \tag{2.76}$$

где r_1 введено непосредственно перед следствием 2.19. Это равенство справедливо для $B(x)$ с легкими хвостами.

В случае $B(x) = 1 - \mathrm{e}^{-x}$, полный анализ поведения хвоста ф.р. $\mathsf{P}(V \le x)$ выполнен Моррисоном [129] (1985) при $x \to \infty$ и $\rho \to 1$. Точнее, используя (2.45) как отправной пункт, он доказал, что при $x = T/\varepsilon >> 1, T = O(1)$ и $\rho = 1 - \varepsilon, 0 < \varepsilon << 1$

$$\mathsf{P}(V > T/\varepsilon) \sim 2\sqrt{T}K_1(2\sqrt{T}) + \varepsilon\left[2\sqrt{T}K_1(2\sqrt{T}) - \sqrt[3]{T^2}K_3(2\sqrt{T})/3\right],$$

где $K_m(\cdot)$ — функция Макдональда [140]. ($K_m(\cdot)$ есть одно из стандартных решений уравнения Бесселя. Оно выражается через функции Ганкеля (Hankel) $H_m^{(1)}(x)$ порядка m как

$$K_m(x) = \pi i\mathrm{e}^{m\pi i/2}H_m^{(1)}(x\mathrm{e}^{\pi i/2})/2,$$

которые, в свою очередь, часто представляются через другое стандартное решение уравнения Бесселя в терминах модифицированных бесселевых функций первого рода.) Он показал также, что в $x = 0$ происходит сингулярное возмущение. Кроме того, Моррисон доказал, что при $x = T/\varepsilon = O(1)$

$$\mathsf{P}(V > x) \sim 1 + \varepsilon x[2\gamma - 1 + \log(\varepsilon x)] + \varepsilon[2/3 - Q_1(x)],$$

где γ — константа Эйлера, а $Q_1(x)$ — сумма двух определенных интегралов, которые не приводятся здесь. Кроме того, найден главный член в асимптотической аппроксимации $\mathsf{P}(V > x)$ при $\varepsilon^2 x = O(1)$ и $\varepsilon x >> 1$

$$\mathsf{P}(V > x) \sim \sqrt{\pi}(\varepsilon x)^{1/4} \mathrm{e}^{-2(\varepsilon x)^{1/2} - \varepsilon^2 x/6}, \tag{2.77}$$

когда $\varepsilon = 1 - \rho \to 0$.

Анализ опирается на изящное использование асимптотических свойств модифицированных бесселевых функций.

Доказательство (2.77) довольно изощренное. Позднее в [78] (2001) получена более грубая эвристическая оценка вероятности $\mathsf{P}(V > x)$ для "достаточно больших" x, ее вывод проще доказательства (2.77 в [129]. Хотя эта недавняя асимптотическая оценка может иметь силу только при $\rho > 1/2$, такая аппроксимация является весьма точной при высоких значениях ρ. Даже при ρ в диапазоне $[0.55, 07]$, асимптотика все еще обеспечивает удовлетворительную точность для больших x.

Флатто [66] использует (2.46) как отправной пункт для нахождения формулы для асимптотики хвоста ф.р. времени ожидания в системе M/M/1—RO. Доказательство чисто аналитическое, оно основано на применении метода Лапласа. Основной результат становится подобным формулам выше, если принять во внимание (2.48). Приведем формулу Флатто, которая приспособлена здесь к системе M/M/1—EPS

$$\mathsf{P}(V > x) \sim \bar{c}_1 x^{-5/6} \mathrm{e}^{-\bar{c}_2 x - \bar{c}_3 x^{1/3}}, \quad x \to \infty, \tag{2.78}$$

где константы $\bar{c}_1, \bar{c}_2, \bar{c}_3$ известны. Эти константы приведены в [66] (см. также [218]).

Стоит отметить, что шестьдесят лет назад результат (2.78) (в контексте системы M/M/1—RO) уже был анонсирован Поллачеком в [143]. Мы остаемся в изумлении от осознания того факта, что многое из рассказанного про систему M/M/1—RO было известно ему в 1946 г.

Представляет интерес выяснить асимптотику $\mathsf{P}(V(u) > x)$ для системы M/GI/1—EPS в случае легкого хвоста $B(x)$. Главный инструментарий для этого введен в [199, 201] и подробно обсуждался в §2.3. Осталось сделать только небольшой шаг к этому побочному вопросу посредством перевода основных результатов §2.3 на язык производящих функций моментов (в действительности, это является дальнейшим развитием следствия 2.19, которое было установлено для $\mathsf{Var}[V(u)]$).

Снимем условие в декомпозиции (2.13) only on $(x_1, \ldots, x_n)$, что приводит к

$$V(u) \stackrel{d}{=} D(u) + \sum_{i=1}^{L} \Phi_i(u),$$

где L есть число требований в системе EPS (распределенное геометрически, см. следствие 2.2) и $\Phi_i(u)$ совпадает с $\Phi(x_i, u)$ после усреднения по $dF(x) = \beta_1^{-1}(1 - B(x))dx$ (см. начало §3.1). Сл.в. $\Phi_i(u) \stackrel{d}{=} \Phi(u)$ имеет ф.р. $\Phi(x|u) = \mathsf{P}(\Phi(u) \le x | B = u)$. Принимая во внимание (2.12) и замечания 2.4 и 2.3 (или 3.1), получаем

$$\mathsf{P}\left(\sum_{i=1}^{L} \Phi_i(u) > x\right) = (1 - \rho) \sum_{n=1}^{\infty} \rho^n \left[1 - \Phi^{n*}(x|u)\right],$$

где $\Phi^{n*}(x|u)$ есть n–кратная свертка распределения $\Phi(x|u)$.

Теперь уточним условие $\beta(-s) < \infty$ при некотором $s > 0$ следующим образом (ср. с условием (b) перед следствием 2.19):

(i) существует некоторая константа $\theta(u) > 0$ (показатель Лундбергathe), такая, что

$$\mathsf{E}\left[\mathrm{e}^{\theta(u)\Phi(u)}\right] = \rho^{-1}, \tag{2.79}$$

(ii) кроме того, существует некоторая константа $C(u)$, такая чтоДоказана

$$C(u) = \rho \int_0^\infty x\mathrm{e}^{\theta(u)x} d\Phi(x|u) < \infty. \tag{2.80}$$

Доказана

Теорема 2.13. *(2006)[229][43] Если выполняются условия (2.79) и (2.80), то при любом конечном и справедливо*

$$\mathsf{P}(V(u) > x) \sim C_1(u)\mathrm{e}^{-\theta(u)x}, \qquad x \to \infty, \tag{2.81}$$

где

$$C_1(u) = (1-\rho)\mathsf{E}\left[\mathrm{e}^{\theta(u)D(u)}\right] / [C(u)\theta(u)]. \tag{2.82}$$

Здесь $\theta(u)$ есть решение уравнения (2.79). Достаточное условие для существования этого решения вытекает из неравенства

$$\rho^{-1} < \mathsf{E}\left[\mathrm{e}^{s_1\Phi(u)}\right] < \infty \quad \textit{при некотором } s_1 > 0 \tag{2.83}$$

(другими словами, ρ должно быть достаточно высоким).

Теорема 2.13 имеет более простые формы в частных случаях. Условия (2.80) и (2.83) можно также уточнять. Например, если $\beta_1 = 1$, то (2.83) может быть представлено как

$$\mathsf{E}\left[\mathrm{e}^{s_2(1,u)(u\wedge F)}\right]^{-1} < \rho \quad \text{for some } s_2 > 0.$$

Отметим, что (2.80) эквивалентно $\rho\frac{d}{ds}\mathsf{E}\left[\mathrm{e}^{s\Phi(u)}\right]\big|_{s=\theta(u)} < \infty$. Константа $C_1(u)$ в (2.82) сводится к виду

$$C_1(1) = [(1-\lambda)(\lambda-\theta(1)]/[2\lambda(1-\lambda) - \theta(1)\lambda(2-\lambda)]$$

для случая системы M/D/1—EPS, когда $\beta(s) = \mathrm{e}^{-su}$ и $u = 1$. Здесь $\theta(1)$ есть единственное положительное действительное решение уравнения, в которое превращается (2.79):

$$\left[\lambda(\lambda-s) + s - s\mathrm{e}^{\lambda-s}\right] / \left[(\lambda-s)(\lambda - s\mathrm{e}^{\lambda-s})\right] = \lambda^{-1}.$$

Этот частный случай изучается также в [61] с дополнительными деталями.

Теорема 2.13 позволяет изучать экспоненциальную асимптотику хвоста распределения времени пребывания требования с длиной u, в частности, *скорость убывания* (decay rate) (это обратная величина времени релаксации, которое обсуждается в конце §3.1 близ следствия 3.5) в системе M/GI/1—EPS при достаточно большом ρ.

[43] Англо–язычный вариант теоремы 2.13 опубликован также в e–журнале *Information Processes*, 2006, vol. 6, no. 3, pp. 256–257 (http://jip.ru/).

3. ТЕОРИЯ СИСТЕМЫ ОБСЛУЖИВАНИЯ M/GI/1—EPS. НЕСТАЦИОНАРНЫЙ РЕЖИМ.

3.1. Нестационарное распределение числа требований

Обсудим ряд результатов исследования системы EPS, обобщающих и усиливающих теоремы 2.4 и 2.5, дающих обоснование метода анализа и приводящих, кроме того, к получению нестационарного распределения процесса числа требований $\{L(t),\ t \in \mathbf{R}_+\}$. Пусть в момент $t = 0+$ начинается первый период занятости системы и в этот момент она находится в состоянии $(n; x_1, \ldots, x_n)$, т.е. содержит $L(0+) = n$ требований, $n > 0$, с остаточными длинами $x_i \in [x_i,\ x_i + dx_i)$, $i = 1, 2, \ldots$. Таким образом, мы изучаем нестационарный процесс $\mathrm{X}_0(t), \S 2.22.2. \zeta = \inf\{t > 0 :\ L(t) = 0\}$. Рассмотрим процесс $\{L(t),\ t \in [0, \zeta)\}$ на первом периоде занятости (считаем $L(t) = 0$ при $t > \zeta$). Введем следующий непрерывный аддитивный функционал, построенный на траекториях процесса $\{L(t)\}$ (точнее, на траекториях процесса $X_0(t)$) [203,205,207]:

$$X(t) = \int_0^t \frac{du}{L(u) \vee 1}. \tag{3.1}$$

(Относительно различных функционалов аддитивного типа см., например, обзор [173].) Будем предполагать, что $X(0) = 0$, $X(t) = X(\zeta^-)$ при $t > \zeta$ и $\lim_{t\to\infty} X(t) = \infty$. Заметим, что $X(t) < \infty$ при $t < \infty$.

Случайная замена времени, связанная с процессом $\{X(t)\}$, определяется как

$$\tau(t) = \inf\{u > 0 :\ X(u) \geq t\}, \quad t \in [0, \infty). \tag{3.2}$$

Это возрастающая кусочно гладкая функция. Семейство таких функций образует процесс $\{\tau(t)\}$ со стационарными приращениями. Пусть $\tau(0) = 0$ и $\tau(t) = \infty$, если $\{u > 0 :\ X(u) \geq t\} = \emptyset$ (это доопределяет $\{\tau(t)\}$ для всех ω). Отметим, что при каждом фиксированнном t сл.в. $\tau(t)$ есть марковский момент (a stopping time) процесса $X_0(t)$ относительно возрастающего семейства σ–алгебр. Положим

$$M(t) = L(\tau(t)). \tag{3.3}$$

Будем считать $M(t) = 0$ при $\tau(t) = \infty$. Равенство (3.3) означает, что процесс $\{M(t)\}$, определенный почти наверное, получен из процесса $\{L(t)\}$ случайной заменой времени (3.2). (См. замечание 2.2 и дополнительно [31,114,124], [165,173,202,203] относительно случайной замены времени.) Процесс $\{M(t)\}$ непрерывен справа; рассматриваемый вместе со своими дополнительными координатами $x_1(\tau(t)), \ldots, x_{L(\tau(t))}(\tau(t))$, этот процесс обладает строго марковским свойством. Другие подробности о строении преобразованных процессов в этой и родственных системах можно найти в [202,203,207,208].

Пусть $\zeta_m = \inf\{t :\ M(t) = 0\}$. Отметим, что $X(\zeta) = \zeta_m$. Кроме того,

$$\tau(X(t)) = t, \quad X(\tau(t)) = t \quad \text{для каждого } t. \tag{3.4}$$

Справедлива

Теорема 3.1. *(1985)[203,205,207,208,216]. Между траекториями процессов $\{L(t)\}$ и $\{M(t)\}$ с соответствующими дополнительными координатами существует взаимно-однозначное соответствие, при котором для каждого фиксированного t*

$$\tau(t) = \int_0^t M(u)\, du. \tag{3.5}$$

Доказательство. (См. [208, p. 93–94], [216, Предложение 3.3] для деталей.) Равенство (3.5) доказывается с помощью формулы Лебега, играющей центральную роль в случайной замене времени [31, Ch. 5, §2], [124, Ch. 7, §2]: если $\{X(t),\ t \geq 0\}$ — неотрицательный процесс с монотонно возрастающими траекториями и $\{\tau(t),\ t \geq 0\}$ — процесс, определяемый (3.2) (т.е. с помощью обратной к $X(t)$ функции), то для любого процесса $\{Y(t),\ t \geq 0\}$ с измеримыми по Борелю траекториями выполняется следующее равенство

$$\int_0^\infty Y(t)\, X(dt) = \int_0^{X(\infty)} Y(\tau(t))\, dt.$$

Теперь применим эту формулу при фиксированном t к процессу $\{X(t), t \geq 0\}$, определяемому равенством (3.1). Положим $Y(t) = \mathbf{1}_{(0,\tau(b)]}(t)L(t)$. Учитывая (3.4), это дает $\mathbf{1}_{(0,\tau(b)]}(\tau(t)) = \mathbf{1}_{(0,\tau(b)]}$. Тогда правая часть формулы Лебега сводится к $\int_0^b M(t)\, dt$ в силу (3.3). Учитывая (3.1), левая часть формулы Лебега принимает вид

$$\int_0^\infty \mathbf{1}_{(0,\tau(b)]}(t)(L(t) \vee 1)\frac{1}{L(t) \vee 1}\, dt = \tau(b).$$

Это приводит к (3.5). □

Из теоремы следует, что (строго возрастающие) траектории процесса $\{\tau(t)\}$ абсолютно непрерывны (этот факт был отмечен в [205, 207]). Для каждого фиксированного t сл.в. $\tau(t)$ в (3.5) интерпретируется так: $\tau(t)$ на множестве $\{\zeta_m > t\}$ представляет собой общую сумму достигнутых длительностей обслуживания требований, которые обслуживаются в момент t (а также требований, обслуженных до момента t).

Процесс $\{M(t)\}$ интерпретируется как процесс числа требований в системе $\tilde{M}/G/\infty$ с марковским входящим потоком переменной интенсивности $\lambda L(t)$ (см. также замечание 2.2). Еще одна интерпретация процесса $\{M(t)\}$ состоит в рассмотрении каждого требования как частицы со временем жизни, распределенным согласно $B(x)$ (ср. с §2.3). Каждая частица за время своей жизни порождает другие частицы (потомков) с интенсивностью λ. Тогда $\{M(t)\}$ представляет собой сумму $L(0) = n$ независимых ветвящихся процессов, которые эволюционируют параллельно, чтобы порождать новые частицы (потомков). Каждый из ветвящихся процессов образуется единственной частицей (предком) из $L(0) = n$ первичных частиц с оставшимся временем жизни x_i (т.е. с длиной x_i на момент $t = 0$) и ее потомками. Это делает возможным представить $\{M(t)\}$ при начальном условии $(L(0) = n; x_1, \ldots, x_n)$ в виде

$$M_n(t, x_1 \ldots, x_n) = \sum_{i=1}^n N_{x_i}(t), \tag{3.6}$$

где $N_x(t)$ — число частиц в момент t в отдельном ветвящемся процессе, порожденном единственным предком со временем жизни x. Траектории каждого процесса $\{N_x(t)\}$ непрерывны справа.

Время жизни такого ветвящегося процесса, обрывающегося в момент t, определяется как

$$\Phi(x, t) = \int_0^t N_x(u)\, du. \tag{3.7}$$

(В иной терминологии $\Phi(x, t)$ было введено ранее в §2.3 перед формулой (2.12) как “элемент задержки”.) В (3.6) и (3.7) время измеряется по новой шкале.

Замечание 3.1. Независимость составляющих декомпозиции (2.13) становится теперь очевидной. Это дополняет аргументы замечания 2.3.

Введем обозначение $\varphi(z,s,x,t) \doteq \mathsf{E}[\mathrm{e}^{-s\Phi(x,t)}\, z^{N_x(t)}]$ для преобразования совместного распределения времени жизни $\Phi(x,t)$ и числа частиц $N_x(t)$ в ветвящемся процессе с единственным x–предком. Справедлива

Теорема 3.2. *(1986) [204, 207, 208] На первом периоде занятости системы, в которую преобразована система M/GI/1—EPS посредством случайной замены времени (3.2), выполняется равенство*

$$\mathsf{E}[\mathrm{e}^{-\mathsf{s}\tau(\mathsf{t})}\, \mathsf{z}^{\mathsf{M}(\mathsf{t})}\mathbf{1}_{(\zeta_{\mathsf{m}}>\mathsf{t})}|(\mathsf{n};\mathsf{x}_1,\ldots,\mathsf{x}_{\mathsf{n}})] = \prod_{\mathsf{i}=1}^{\mathsf{n}} \varphi(\mathsf{z},\mathsf{s},\mathsf{x}_{\mathsf{i}},\mathsf{t}), \tag{3.8}$$

где

$$\varphi(z,s,x,t) = \begin{cases} \delta(z,s,t) & \text{при } x \ge t, \\ \delta(z,s,t)/\delta(z,s,t-x) & \text{при } x < t, \end{cases} \tag{3.9}$$

$$\delta(z,s,t) = z\,\mathrm{e}^{-(s+\lambda)t}/\psi(z,s,t), \quad t \ge 0. \tag{3.10}$$

Здесь функция ψ задается своим ПЛ по t (аргумент q)

$$\tilde{\psi}(z,s,q) = \int_0^\infty \mathrm{e}^{-qt}\,\psi(z,s,t)\,dt = \frac{q+s+\lambda-\lambda z+\lambda z\beta(q+s+\lambda)}{(q+s+\lambda)(q+\lambda\beta(q+s+\lambda))}, \tag{3.11}$$

$Re\ s>0,\ q>-\lambda\pi(s),\ |z|\le 1,\ \pi(s)=\mathsf{E}[\mathrm{e}^{-\mathsf{s}\zeta}]$ – решение уравнения (2.4).

Комментарии к доказательству. (Теорема 3.2 была впервые анонсирована в [204]. См. [207, 208, 209] для деталей доказательства.) Формулы (3.9) – (3.11) являются решением системы уравнений, которые аналогичны по типу уравнениям (2.14), (2.15) и отличаются от них только наличием еще одного переменного z в неизвестных функциях $\delta(z,s,t)$, $\varphi(z,s,x,t)$ и начальными условиями $\varphi(z,s,0,t)=1$, $\varphi(z,s,x,0)=\delta(z,s,0)=z$. Эти условия отражают тот факт, что $\Phi(0,t)=\Phi(x,0)=0$, $N_0(t)=0$ и $N_x(0)=1$. Уравнения выводятся при рассмотрении инфинитезимальных изменений функций φ и $\delta=\varphi$ при $x\ge t$. Составление уравнений аналогично выводу уравнений (2.15) и (2.14) для $\delta(s,u)=\delta(1,s,u)$ и $\varphi(s,x,u)=\varphi(1,s,x,u)$. Техническая часть доказательства теоремы 3.2 заключается в формальном распространении доказательства теоремы 2.4 на случай учета текущего числа потомков. Отметим лишь, что при этом функция ψ определяется равенством (ср. с замечанием 2.6)

$$\psi(z,s,t) \doteq \exp\left[-\lambda\int_0^t \varphi_B(z,s,u)\,du\right],$$

где

$$\varphi_B(z,s,t) = \int_0^\infty \varphi(z,s,x,t)\,dB(x),$$

откуда следует

$$\int_0^t \varphi_B(z,s,y)\,dy = -\lambda^{-1}\ln\psi(z,s,t), \tag{3.12}$$

а также $\varphi_B(z,s,t) = -\lambda^{-1}\frac{\partial}{\partial t}\psi(z,s,t)/\psi(z,s,t)$ и $\pi(s) = -\lambda^{-1}\lim_{t\to\infty}\ln\psi(1,s,t)$. □

Замечание 2.5 также справедливо для теоремы 3.2.

Пусть в момент $t=0$ система M/GI/1—EPS находится в состоянии $(n;x_1,\ldots,x_n)$. Введем обозначения

$$p_{n,\mathbf{x}}(z,s) = \int_0^\infty \mathrm{e}^{-st}\,\mathsf{E}[\mathsf{z}^{\mathsf{L}(\mathsf{t})}\mathbf{1}_{(\zeta>\mathsf{t})}|(\mathsf{n};\mathsf{x}_1,\ldots,\mathsf{x}_{\mathsf{n}})]\,\mathsf{dt},\ \mathsf{n}\ge 1,\ \mathsf{Re\ s}>0,\ |\mathsf{z}|\le 1, \tag{3.13}$$

$$p_n(z,s) = \int_0^\infty \mathrm{e}^{-st}\,\mathsf{E}[\mathsf{z}^{\mathsf{L(t)}}\mathbf{1}_{(\zeta>\mathsf{t})}|\mathsf{L}(0)=\mathsf{n}]\,\mathsf{dt}.$$

Используя теорему Фубини, можно переписать (3.13) в следующей эквивалентной форме

$$p_{n,\mathbf{x}}(z,s) = \mathsf{E}\left[\int_0^\zeta \mathrm{e}^{-\mathsf{st}}\,\mathsf{z}^{\mathsf{L(t)}}|(\mathsf{n};\mathsf{x}_1,\ldots,\mathsf{x}_\mathsf{n})\right]\,\mathsf{dt}. \tag{3.14}$$

Заменяя в (3.14) переменную интегрирования на $\tau(t)$ (мы интегрируем по возрастающему процессу) и снова применяя теорему Фубини, можно получить с учетом (3.3) и (3.5)

$$p_{n,\mathbf{x}}(z,s) = z\int_0^\infty \frac{\partial}{\partial z}\left(\mathsf{E}[\mathrm{e}^{-\mathsf{s}\tau(\mathsf{t})}\,\mathsf{z}^{\mathsf{M(t)}}\mathbf{1}_{(\zeta_\mathsf{m}>\mathsf{t})}|(\mathsf{n};\mathsf{x}_1,\ldots,\mathsf{x}_\mathsf{n})]\right)\,dt. \tag{3.15}$$

Снимем условия по остаточным длинам требований, находящихся в системе в момент $t=0$. Принимая во внимание (3.12), это приводит к

Теорема 3.3. *(1988) [207,208,209] При любом конечном $\rho>0$ справедливо следующее интегральное представление для преобразования Лапласа производящей функции числа требований в момент t на первом периоде занятости системы M/G/1—EPS, который начинается при наличии $n>0$ требований в момент 0*

$$p_n(z,s) = \frac{z}{\lambda^n}\int_0^\infty \frac{\partial}{\partial z}\left[-\frac{\partial}{\partial t}\ln\psi(z,s,t)\right]^n\,dt,\ Re\ s>0,\ |z|\le 1, \tag{3.16}$$

где ПЛ функции ψ задается равенством (3.11).

Ограничиваясь случаем $n=1$ (первый период занятости открывается единственным требованием), можно получить $p_1(z,s)$ в явном виде.

Теорема 3.4. *(1989) [208,211,226] В системе M/G/1—EPS при любом конечном $\rho>0$*

$$p_1(z,s) = \frac{z(1-\pi(s))}{s+\lambda(1-z)(1-\pi(s))},\quad Re\ s>0,\ |z|\le 1. \tag{3.17}$$

Схема доказательства. (Полное доказательство дано впервые в [208, с.97-98], см. также [210,211,226] для уточнений.) Простые алгебраические преобразования позволяют представить $\tilde\psi(z,s,q)$ в (3.11)) в виде

$$\tilde\psi(z,s,q) = \tilde\kappa_1(s,q) - z\tilde\kappa_2(s,q).$$

В этом равенстве

$$\tilde\kappa_1(s,q) = [q+\lambda\beta(q+s+\lambda)]^{-1},\ \tilde\kappa_2(s,q) = (\lambda-\lambda\beta(q+s+\lambda))\tilde\kappa_1(s,q)(q+s+\lambda)^{-1}. \tag{3.18}$$

Из (3.18) следует, что неизвестные функции $\kappa_i(s,t) = \mathcal{L}^{-1}(\tilde\kappa_i(s,q))(s,t),\ i=1,2$ являются плотностями функций восстановления обрывающихся (транзиентных) процессов восстановления. Здесь $\mathcal{L}^{-1}$ — оператор обращения (двумерного) преобразования Лапласа (контурный интеграл Бромвича); аргументу q соответствует переменная t. (Нет необходимости применять $\mathcal{L}^{-1}$ к одномерному преобразованию Лапласа $\kappa_i(s,t)$.)

Методом преобразования несобственного (defective) уравнения восстановления к собственному [65, Ch. 11, §6], опирающимся на предельную теорему для обрывающихся процессов восстановления, находятся асимптотики $\kappa_i(s,t)\,\mathrm{e}^{\lambda\pi(s)t}$ при $t\to\infty$. Подчеркнем, что для этого следует сначала проверить некоторые условия, гарантирующие возможность преобразований

несобственных уравнений восстановления, и далее ввести новую (экспоненциально смещенную) вероятностную меру для интервалов между моментами восстановления [208, 226]. Это позволяет "вогнать" $\kappa_i(s,t)$ в стандартные рамки теории восстановления. (Такое преобразование вероятностной меры называется преобразованием Эшера (Esscher transformation)[63][44] в теории страхового риска [62, 165]. Вследствие этого преобразования появляется в конечном счете дополнительный экспоненциальный множитель в левой части приводимых далее формул.)

Асимптотические решения двух таких собственных уравнений восстановления находятся с помощью хорошо известной узловой теоремы восстановления [168]. Эти решения имеют вид

$$\lim_{t\to\infty} \kappa_1(s,t)\,\mathrm{e}^{\lambda\pi(s)t} = [C_1(s)(\lambda - \lambda\pi(s))]^{-1},$$

$$\lim_{t\to\infty} \kappa_2(s,t)\,\mathrm{e}^{\lambda\pi(s)t} = [C_2(s)(s + \lambda - \lambda\pi(s))]^{-1},$$

где $C_1(s) < \infty$ и $C_2(s) < \infty$ — некоторые зависящие от s константы. Учитывая упомянутые выше условия, нетрудно установить, что $C_1(s) = C_2(s)$ при $s > 0$. Тогда из двух последних приведенных равенств вытекает, что

$$\lim_{t\to\infty} \kappa_2(s,t)/\kappa_1(s,t) = \lambda(1 - \pi(s))/(s + \lambda - \pi(s)). \tag{3.19}$$

Стоит подчеркнуть, что, когда $n = 1$, можно поменять порядок дифференцирования по z и по t в подынтегральном выражении в правой части равенства (3.16). Это позволяет свести (3.16) после интегрирования к виду

$$p_1(z,s) = \frac{z}{\lambda} \lim_{t\to\infty} \frac{\kappa_2(s,t)}{\kappa_1(s,t) - z\kappa_2(s,t)},$$

используя тот факт, что $\psi(z,s,0) = 1$. Принимая во внимание (3.19), можно привести последнее равенство к виду 3.17). □

Распределение процесса $\{L(t)\}$ на всей положительной полуоси $[0, \infty)$ можно теперь найти, используя более стандартные результаты теории восстановления тем же самым способом, как это делается в[86, Ch. 1], где изучается процесс длины очереди в системе M/GI/1—FCFS. Однако невнимательно опущено из §1.3 в [86], что нам необходимо рассмотреть два процесса восстановления, которые образуются моментами начала и окончания периодов занятости, следующих друг за другом. Первый процесс восстановления состоит из последовательности моментов, в которые система EPS становится свободной. Считаем, что начальный момент 0 тоже является моментом восстановления. Второй процесс восстановления состоит из последовательности моментов поступления требований в свободную систему. Он отличается от первого процесса восстановления только тем, что его первый момент восстановления сдвинут на случайную величину (свободный период), которая имеет экспоненциальную функцию распределения с параметром λ вследствие пуассоновского входящего потока. Для простоты мы предполагаем, что вся система EPS свободна от требований в момент $t = 0$, т.е. $\mathsf{P}(L(0) = 0) = 1$[45].

Введем обозначение $g_0(z,s) \doteq \int_0^\infty \mathrm{e}^{-st}\,\mathsf{E}[z^{L(t)}|L(0) = 0]\,dt$. Нетрудно вывести [208, p. 99] следующее уравнение, которому удовлетворяет функция $g_0(z,s)$

$$g_0(z,s) = \eta(s)/\lambda + \eta(s)p_1(z,s),$$

[44] Преобразование Эшера родственно хорошо известному преобразованию Гирсанова–Маруямы вероятностных мер в семимартингалах [165, 188].

[45] В принципе, нетрудно перейти к призвольному начальному условию от этого нулевого начального условия, но это приведет к более громоздким результатам, см., например, замечание 3.9 и цитированную в нем статью.

где $\eta(s) = \lambda/(s + \lambda - \lambda\pi(s))$ есть преобразование Лапласа по t плотности восстановления во втором процессе восстановления, упомянутом выше. Следовательно,

$$g_0(z,s) = [1 + \lambda p_1(z,s)](s + \lambda - \lambda\pi(s))^{-1},\ Re\ s > 0,\ |z| \le 1. \tag{3.20}$$

Замечание 3.2. Равенство (3.20) связывает характеристики процесса числа требований $\{L(t)\}$ на $[0,\infty)$ при нулевом начальном условии с подобными характеристиками этого же самого процесса на $[0,\zeta)$ (на первом периоде занятости). Оно выполняется для любой консервативной дисциплины в системе M/GI/1. Дисциплиной определяется только вид функции $p_1(z,s)$. Однако, нахождение вида функции $p_1(z,s)$ является наиболее трудным этапом анализа в случае дисциплины EPS. Для сравнения, напомним вид $p_1(z,s)$ для системы the M/GI/1—FCFS [86, Ch. 1, §2]

$$p_1(z,s)_{FCFS} = \frac{1 - \beta(s + \lambda - \lambda z)}{s + \lambda - \lambda z}\,\frac{z - \pi(s)}{1 - z^{-1}\beta(s + \lambda - \lambda z)}\,. \tag{3.21}$$

Равенства (3.17) и (3.20) приводят к утверждению

Теорема 3.5. *(1989) [208, 210, 211] Функция*

$$g_0(z,s) = [s + \lambda(1 - z)(1 - \pi(s))]^{-1} \tag{3.22}$$

определяет вид преобразования Лапласа вероятностной производящей функции числа требований в момент t в системе M/GI/1—EPS при начальном условии $\mathsf{P}(L(0) = 0) = 1$.

Одним из следствий теоремы 3.5 является вид преобразования Лапласа вероятности отсутствия требований в системе EPS в момент t

Следствие 3.1. *(1989) [208, 210, 211]*

$$g_0(0,s) \equiv \tilde{p}_{00}(s) = \int_0^\infty \mathrm{e}^{-st}\,\mathsf{P}_{00}(t)\,dt = [s + \lambda - \lambda\pi(s)]^{-1}. \tag{3.23}$$

Здесь и далее дополнительный нижний индекс 0 будет указывать на начальное состояние системы в момент $t = 0$.

Формула (3.23) показывает, что $\mathsf{P}_{00}(t)$ принадлежит классу стандартных p-функций, введенному Кингманом при изучении феномена регенерации[46]. Следовательно, многие свойства $\mathsf{P}_{00}(t)$ вытекают из свойств p–функций. В частности, $|\mathsf{P}_{00}(t_2) - \mathsf{P}_{00}(t_1)| < \rho|t_2 - t_1|$ для всех положительных t_1 и t_2, поскольку $\mathsf{P}'_{00}(0) = -\rho$. Следовательно, $\mathsf{P}_{00}(t)$ абсолютно непрерывна относительно меры Лебега. Это означает, что если ввести функцию $A(t) = (1 - \mathsf{P}_{00}(t))/\rho$ $\mathsf{P}_{00}(0) = 1$, $\mathsf{P}_{00}(\infty) = 1 - \rho$, то она имеет плотность $\alpha(t)$ и $0 \le \alpha(t) \le 1$. (Однако $A(t)$ может не быть дифференцируемой при всех t, примером служит система M/D/1.) См., например, Abate и Whitt [6] про свойства вероятности отсутствия требований. Кроме того, замечание 3.5 содержит некоторые факты относительно аппроксимаций $\mathsf{P}_{00}(t)$.

Следующее утверждение содержит другой вид следствия 3.1 (а также теоремы 2.1).

[46] В общем, $\mathsf{P}_{00}(t)$ — хорошо изученный объект в контексте системы M/GI/1—FCFS; следствие 3.1 утверждает формально, что система M/GI/1—EPS имеет ту же самую характеристику. Следствие 3.1 гарантирует, что все известные утверждения, связанные с этой характеристикой системы M/GI/1—FCFS, автоматически распространяются и на систему M/GI/1–EPS, например, аналоги следствия 3.5. Тем не менее, существуют некоторые пробелы в изучении вероятности отсутствия требований $\mathsf{P}_{00}(t)$.

Следствие 3.2. *(1994) [6] Для системы M/GI/1, ПЛ $\tilde{p}_{00}(s)$ есть решение функционального уравнения*

$$\tilde{p}_{00}(s) = [s + \lambda - \lambda\beta(1/\tilde{p}_{00}(s))]^{-1}.$$

Доказательство. Положим $\theta = \theta(s) = s+\lambda-\lambda\pi(s)$, Re $s > 0$ в (2.4). Утверждение выводится из (3.23) и (2.4) после некоторых алгебраических преобразований. □

Функциональные уравнения (2.4) теоремы 2.1 и следствие 3.2, очевидно, эквивалентны, т.е. из решения одного уравнения вытекает решение другого.

Еще одна эквивалентная форма уравнения (2.4) представляется как

$$\pi(s) = \beta(l^{-1}(s)), \tag{3.24}$$

где $l^{-1}(s) = \theta(s) = s + \lambda - \lambda\pi(s)$ есть обратная функция к функции $l(s) = \theta(s) = s - \lambda + \lambda\beta(s)$. Функциональное уравнение (3.24) выводится в [151] прямым мартингальным методом.

Второй способ установления результата аналогичного (3.22) состоит в применении метода допонительных переменных (ср. с теоремой 2.3 и равенством (2.8), которое справедливо в стационарном режиме). Тем не менее, в отличие от классических очередей, только успешное доказательство нетривиальных теорем 3.2—3.4 позволяет опираться на этот метод в нестационарном случае. Пусть E_i при $1 \le i \le L(t)$ являются прошедшими (достигнутыми) длительностями обслуживания каждого из $L(t) \ge 1$ требований, находящихся в системе в момент t. Тогда $A_0(t) = \{L(t), E_i(t),\ i = 1, \ldots, L(t);\ t \ge 0\}$ будет марковским процессом с пространством состояний

$$\{0\} \cup \{n, x_1, x_2, \ldots, x_n : n \ge 1,\ x_i \ge 0 \quad \text{for } 1 \le i \le L(t)\},$$

где x_i обозначает величину достигнутого времени обслуживания у i–го требования, находящегося в процессоре, при $i \ge 1$.

Введем обозначения

$$\mathsf{P}_n(t; x_1, \ldots, x_n)dx_1dx_2\ldots, dx_n = \mathsf{P}\{L(t) = n;\ E_i(t) \in [x_i, x_i + dx_i), i = 1, 2, \ldots, n\}.$$

Напомним, что $\mathsf{P}_{00}(t) = \mathsf{P}\{L(t) = 0|L(0) = 0\}$.

С помощью хорошо известного метода можно вывести дифференциальные уравнения, которым удовлетворяют совместные нестационарные функции плотностей $\mathsf{P}_{00}(t)$, $\mathsf{P}_n(t; x_1, \ldots, x_n)$, $n \ge 1$

$$\left(\frac{\partial}{\partial t} + \lambda\right)\mathsf{P}_{00}(t) = \int_0^\infty \mu(x_1)\mathsf{P}(t; x_1)\, dx_1, \tag{3.25}$$

$$\frac{1}{n}\sum_{i=1}^{\infty}\left[\frac{\partial}{\partial t} + \frac{\partial}{\partial x_i} + \mu(x_i) + \lambda\right] P_n(t; x_1, \ldots, x_n) = \int_0^\infty P_{n+1}(t; x_1, \ldots, x_n, x_{n+1})\mu(x_{n+1})\, dx_{n+1}, n \ge 1, \tag{3.26}$$

где $\mu(x)$ есть функция интенсивности отказов распределения $B(x)$ (см. конец сноски 28). Граничные условия имеют вид

$$\mathsf{P}_1(t, 0) = \lambda\mathsf{P}_{00}(t), \tag{3.27}$$

$$P_{n+1}(t; x_1, \ldots, x_n, 0) = \lambda P_n(t; x_1, \ldots, x_n), \quad n \ge 1. \tag{3.28}$$

Условием нормировки для системы уравнений (3.25)—(3.28) является

$$\mathsf{P}_{00}(t) + \sum_{n=1}^{\infty}\int_0^\infty \ldots \int_0^\infty \mathsf{P}(t; x_1, \ldots, x_n)dx_1 \ldots dx_n = 1. \tag{3.29}$$

В действительности, уравнения (3.25)—(3.28) являются простым распространением уравнений (4.4) из [105] на нестационарный случай. Их решение в стационарном режиме дается равенством (2.8), в котором использован символ a_i вместо x_i.

После некоторых математических манипуляций с использованием ПЛ и ПЛС, можно убедиться, что решение системы уравнений (3.25)—(3.28) принимает вид

$$\int_0^\infty \mathrm{e}^{-st}\mathsf{P}_n(t;x_1,\ldots,x_n)\,dt = \tilde{p}_{00}(s)[1-s\tilde{p}_{00}(s)]^n \prod_{i=1}^{n}(1-B(x_i)), \tag{3.30}$$

где $\tilde{p}_{00}(s)$ дается равенством (3.23). Отметим, что $[1-s\tilde{p}_{00}(s)]$ в (3.30) совпадает с правой частью равенства (3.19).

Равенство(3.30) в конечном счете приводит к (3.22).

Замечание 3.3. Используя (3.20), можно формально установить, что транзиентное поведение процесса $\{L(t)\}$ в сисиеме the M/M/1 такое же как и в системе FCFS при нулевом начальном условии. Уравнение(2.4) при $\beta(s) = \mu/(s+\mu)$ сводится к квадратному уравнению с решением

$$\pi(s) = \left[(s+\lambda+\mu) - ((s+\lambda+\mu)^2 - 4\lambda\mu)^{1/2}\right]/(2\lambda).$$

(Следует выбрать знак "-" перед корнем, поскольку $|\pi(s)| \le 1$.) Теперь можно показать после простых алгебраических преобразований, что $g_0(z,s)$, где $p_1(z,s)$ дается равенством (3.21), совпадает с (3.22). Это верно, когда $\mathsf{P}(L(0)=0)=1$. Еще один способ доказательства этого утверждения заключается в том, чтобы показать, что процессы длины очереди в обеих системах EPS и FCFS имеют одинаковый инфинитезимальный оператор (generator) в случае, когда $B(x) = 1-\mathrm{e}^{-\mu x}$.

Следующее утверждение содержит информацию относительно среднего числа требований в в системе M/GI/1—EPS в момент t.

Следствие 3.3. *(1989) [208, 210, 211]*[47]

$$\int_0^\infty \mathrm{e}^{-st}\mathsf{E}[L(t)]\,dt = \lambda(1-\pi(s))/s^2 \ . \tag{3.31}$$

Доказательство. Простое вычисление предела производной функциии $g_0(z,s)$ из теоремы 3.5 по z, а именно,

$$\int_0^\infty \mathrm{e}^{-st}\mathsf{E}[L(t)]\,dt = \lim_{z\to 1}\frac{\partial g_0(z,s)}{\partial z}.$$

□

Следствие 3.4. *(1990) [74, 210, 212, 213, 214, 231] когда $\rho > 1$ в системе M/GI/1—EPS, то*

$$\lim_{t\to\infty}\frac{L(t)}{t} = \gamma \quad \textit{с вероятностью } 1, \tag{3.32}$$

[47] Для классической системы M/GI/1—FCFS ПЛ по t функциии $\mathsf{E}[L(t)]$ имеет другой вид по сравнению с (3.31), который зависит также от $\beta(s)$. Именно,

$$\int_0^\infty \mathrm{e}^{-st}\mathsf{E}[L(t)_{FCFS}]\,dt = \frac{\lambda}{s^2} - \frac{\lambda}{s}\frac{\beta(s)}{1-\beta(s)}\frac{1-\pi(s)}{s+\lambda-\lambda\pi(s)}.$$

Конечно, эта формула становится эквивалентной (3.31) в силу замечания 3.3, когда $B(x) = 1-\mathrm{e}^{-\mu x}$.

где $\gamma = \lambda(1 - \pi(0)) > 0$ *есть решение уравнения*

$$\lambda \int_0^\infty \mathrm{e}^{-\gamma x}(1 - B(x))dx = 1. \tag{3.33}$$

Comments to the proof. Первое (трудное) доказательство дано в [74], где эксплуатировались довольно экзотические теоремы из суперкритических ветвящихся процессов. Очень простое и короткое доказательство дано в [231]. Оно основано на следствии 3.3 и некоторых транзиентных свойств процесса $\{L(t)\}$ в системе M/GI/1—EPS при $\rho > 1$. Промежуточные доказательства и обобщения упомянуты в середине §1.2 и в сноске 27. Все доказательства, за исключением [231], довольно сложны, поскольку транзиентное распределение процессаs $\{L(t)\}$ считалось неизвестным. □

Хорошо известно, что при любой консервативной дисциплины оставшаяся работа (workload) в однолинейной системе растет со скоростью $\rho - 1$, когда $\rho > 1$, см. [144] или [46, p.633]. То же самое справедливо для числа требований в момент t в перегруженной системе M/GI/1—FCFS. Другими словами, хорошо известный эквивалент следствия 3.4 для дисциплины FCFS (когда $\rho > 1$) выглядит как

$$\lim_{t\to\infty} \frac{L(t)}{t} = \rho - 1 \quad \text{with probability 1,}$$

см., например, [144]. Следствие 3.4 демонстрирует, что транзиентное поведение числа требований в системе M/GI/1–EPS не такое простое, какого можно было бы ожидать на первый взгляд.

Замечание 3.4. Константа γ в следствии 3.4 называется мальтусовским параметром в теории ветвящихся процессов, в которой, как правило, не используется концепция периода занятости. В противоположность этому, теория очередей базируется, как правило, на концепции периода занятости, но не использует концепции мальтусовского параметра. Это оодин из примеров трудностей, которые возникают при попытках использовать некоторые полезные результаты из, скажем, теории очередей в ветвящихся процессах (и наоборот). Решение уравнения (2.4) и γ родственны, их связь дается следствием 3.4 в случае системы M/GI/1–EPS. Мальтусовский параметр равен 0, когда $\rho < 1$ (стационарный режим). Это наименее интересный случай для теории ветвящихся процессов и наиболее популярный случай для теории очередей, потому что $\pi(0) = 1$ в такой ситуации. Таким образом, наиболее интересная цель теории очередей (точный или асимптотический анализ транзиентного поведения) молчаливо игнорируется или заменяется своим суррогатом (изучение стационарного поведения). Возможно, сказанное может дополнить конец замечания 2.12.

Стоит упомянуть про относительно мало известный аналог следствия 3.4 для системы M/GI/1—FBPS при $\rho > 1$

$$\lim_{t\to\infty} \frac{L(t)}{t} = \lambda(1 - B(x^*)) \quad \text{с вероятностью 1,} \tag{3.34}$$

где x^* есть решение уравнения $\lambda \int_0^x (1 - B(y))dy = 1$. Скорость роста процесса $\{L(t)\}$ в перегруженной системе M/GI/1—FBPS, представленная равенством (3.34), выводится из результатов [202, 208, 212]. Простейший способ доказательства — использовать точное решение для нестационарного распределения длины очереди [208, 212, 225].

Приведем теперь

Другое доказательство следствия 2.2. [208] Если $\rho < 1$, то равенство $\mathsf{P}_n = (1-\rho)\rho^n$, $n = 0, 1, 2, \ldots$, вытекает из (3.22) после применения классической тауберовой теоремы $\lim_{s\downarrow 0} g_0(z,s) = \lim_{t\to\infty} \mathsf{E}[z^{L(t)}] = (1-\rho)/(1-\rho z)$ и последующего обращения этой вероятностной производящей функции. □

Из теоремы 3.5 нетрудно извлечь ряд других новых результатов по характеризации переходного поведения системы M/GI/1—EPS [210,211], но далее мы ограничимся кратким обсуждением времени релаксации и скорости убывания (decay rate).

Можно выписать в явном виде нестационарные вероятности состояний системы M/M/1—EPS (см., например, [210,211]) в терминах модифицированных функций Бесселя первого рода. Эти вероятности состояний совпадают с вероятностями состояний системы M/M/1—FCFS [46, 52, 128, 154] в силу замечания 3.3. Асимптотика для среднего числа требований в момент t известна для экспоненциальной ф.р. $B(x)$ при $t \to \infty$ ($\rho < 1$) (см., например, формулу (2.55) для $\mathsf{E}[L(t)|L(0) = 0]$ в [212]), из которой следует, что $\mathsf{E}[\mathsf{L(t)}|\mathsf{L(0)} = \mathsf{0}]$ стремится к $\rho/(1-\rho)$ экспоненциально быстро. Константа $T_c = [\mu(1-\sqrt{\rho})^2]^{-1}$ при $\rho = \lambda/\mu < 1$ характеризует скорость сходимости $\mathsf{E}[\mathsf{L(t)}]$ к стационарному значению при $t \to \infty$; она называется *временем релаксации* (ср. с [46,30]). Обратная величина константы T_c часто называется "асимптотической скоростью убывания".

Более удобное определение времени релаксации[48] дается равенством

$$T_c = T_c(\mathsf{P}_{00}(t)) = \inf\{T \geq 0 : \ \mathsf{P}_{00}(t) - \mathsf{P}_{00}(\infty) = O(\mathrm{e}^{-t/T})\}.$$

Это определение пригодно для любой однолинейной системы обслуживания при соблюдении условий: (i) $\rho < 1$, (ii) обе ф.р. интервалов между поступлениями и длин требований имеют легкий хвост[49]. Поскольку нас интересует система M/GI/1, то достаточно предположить, что $\beta(s) \uparrow \infty$ при $s \downarrow s_b$, где $s_b < 0$ — абсцисса абсолютной сходимости для ПЛС $\beta(s)$. В этом случае $\beta(s)$ имеет аналитическое продолжение на часть левой полуплоскости, следовательно, $B(x) \notin \mathcal{K}$. (Варьируя определение T_c и используя вместо $\mathsf{P_{00}(t)}$ другие нестационарные характеристики системы, можно находить для них оценки скорости сходимости, отличающиеся от приведенной далее). Оценка в следствии 3.5 получена для простейшей характеристики $\mathsf{P_{00}(t)} - (1-\rho)$.

Следствие 3.5. *(1986) [30, 210, 211, 212]. При сделанном предположении относительно $\beta(s)$ время релаксации T_c для системы M/G/1—EPS можно найти из равенства $T_c = -s_0^{-1}$, где s_0 — сингулярность (особая точка) с наибольшей действительной частью (исключая полюс в точке $s = 0$) функции $\tilde{p}_{00}(s)$, определяемой равенством (3.23).*

Доказательство. Из (3.23) вытекает,что $\theta = \theta(s) = s + \lambda - \lambda\pi(s)$, $\mathrm{Re}\, s > 0$ есть единственное решение уравнения $\beta(\theta) = (s + \lambda - \theta)/\lambda, \mathrm{Re}\, \theta > 0$. В силу предположения о виде ПЛС $\beta(s)$, единственное решение уравнения $\beta'(\theta) = -1/\lambda$ есть θ_0 на $(s_b, 0)$. Поэтому функция $\theta(s)$ имеет точку ветвления в $s_0 = \theta_0 + \lambda(\beta(\theta_0) - 1)$. Аналитическое продолжение функции $\theta(s)$ на область $\mathrm{Re}\, s > s_0$ имеет единственный ноль при $\mathrm{Re}\, s > s_0$, который является полюсом функции $\tilde{p}_{00}(s)$. Следовательно, s_0 есть особая точка с наибольшей действительной частью функциии $\tilde{p}_{00}(s)$, за исключением полюса в точке $s = 0$. □

[48] Понятие времени релаксации ввел впервые Морз (P.Morse) в 1955 г., но его определение несколько отличается от приведенного. См. также [109, §2.8].

[49] Отметим, что ф.р. времени ожидания в системе GI/M/1—FCFS экспоненциальна, даже если распределение интервалов между поступлениями имеет тяжелый хвост. Однако, система G/M/1—FCFS с последовательностью зависимых интервалов между поступлениями может привести к распределению времени ожидания с тяжелым хвостом, see [149].

Это утверждение доказано в [30] (1986) в контексте системы M/GI/1—FCFS. Однако небольшая адаптация результатов [52] (1961) позволяет получить то же самое и даже больше. Мы обсудим этот вопрос после замечания 3.5.

Время релаксации T_c легко вычисляется с помощью следствия 3.5 для некоторых систем M/GI/1, когда $B(x)$ имеет дробно–рациональное ПЛС $\beta(s)$ со знаменателем степени 1 или 2. Такое ПЛС полностью определяется средним β_1, коэффициентом вариации ф.р. $B(x)$ и наибольшим полюсом s_0 этой ф.р. Отметим, что T_c есть возрастающая функция загрузки ρ, коэффициента вариации и s_0 при распределениях $B(x)$ такого типа.

Замечание 3.5. Литература по аппроксимациям характеристик производительности системы M/GI/1 с дисциплинами FCFS и LCFS достаточно обильна. Но некоторые характеристики совпадают с характеристиками системы M/GI/1—EPS, например, $\mathsf{P}_{00}(t)$ и $\mathsf{E}[\mathrm{e}^{-s\Pi}]$. Разнообразные аппроксимации вероятности $\mathsf{P}_{00}(t)$ могут быть извлечены из статей [71, 178]. Точнее, эти статьи содержат аппроксимации среднего виртуального времени ожидания $\mathsf{E}[W(t)]$ в системе M/GI/1—FCFS, стартующей из свободного состояния. Хорошо известна простая формула (3.48) для ПЛ функции $\mathsf{E}[W(t)]$ как следствие результата Такача (3.47). Но она выражается в терминах $\tilde{p}_{00}(s)$ (см. (3.23)), следовательно, нетрудно получить аппроксимации вероятности $\mathsf{P}_{00}(t)$. В [71,178] рассмотрены оба условия недогрузки ($\rho < 1$) и перегрузки ($\rho > 1$) в случае распределений $B(x)$ с легкими хвостами. Случай $B(x)$ с тяжелым хвостом резко отличается (см. (3.39)).

Рассмотрим теперь асимптотику периода занятости при легких хвостах. Отметим, что в силу следствия 3.2, асимптотики вероятности $\mathsf{P}_{00}(t)$ и хвоста периода занятости тесно связаны. Пусть $\pi(x)$ будет плотностью распределения периода занятости при нерешетчатых ф.р. времени обслуживания, тогда ПЛ $\pi(s)$ этой плотности по x дается равенством (2.4). В общем, отсутствует представление для ПЛС $\pi(s)$ распределения периода занятости в явном виде, но возможно вычислить плотность $\pi(x)$ при любом x посредством прямого обращения ПЛС $\pi(s)$. Это основано на следующем разложении в ряд

$$\pi(x) = \sum_{n=0}^{\infty} \mathrm{e}^{-\lambda x}(\lambda x)^n beta^{(n+1)*}(x)/(n+1)!,$$

где
$beta^{(n+1)*}(x)$ есть плотность суммы из $n+1$ независимых и одинаково распределенных длительностей обслуживания, см. [52,175,108].

Кокс и Смит [52] исследовали асимптотическое поведение плотности $\pi(x)$, используя метод седловой точки (т.е. метод перевала или наискорейшего спуска)

$$\pi(x) \sim C_b(\pi x^3)^{-1/2}\mathrm{e}^{-x/T_c} \quad x \to \infty. \tag{3.35}$$

Здесь

$$T_c^{-1} = \lambda + \theta - \lambda\beta(-\theta) \tag{3.36}$$

называется скоростью убывания (decay rate), θ есть единственное действительное число y слева от всех сингулярностей производящей функции моментов $\beta(-s)$, такой, что

$$\beta'(-y) = -\lambda^{-1} \tag{3.37}$$

и $C_b = [2\lambda^3\beta''(-\theta)]^{-1/2}$.

Формула (3.35) установлена как равенство (49) в [52, p.156], за исключением обозначений. Главный член асимптотического разложения хвоста распределения периода занятости

$\int_x^\infty \pi(t)\,dt$ приведен формулой (50) на той же странице книги [52]. (Формула (50) содержит неверный множитель $x^{-1/2}$ вместо корректного множителя $x^{-3/2}$, ср. с [7].)

Стоит отметить, что $\mathsf{P}_{00}(t)$ имеет ту же самую скорость убывания T_c^{-1} как в (3.35) и

$$\mathsf{P}_{00}(t) - (1-\rho) = C_e(\pi x^3)^{-1/2}\mathrm{e}^{-x/T_c} \quad x \to \infty, \tag{3.38}$$

где $C_e = \lambda C_b/\theta^2$, θ есть решение уравнения (3.37), а C_b дано выше.

Когда $B(x) \in RV(-a)$, то (3.38) не справедлива. Поведение хвоста в этом случае установлено Асмуссеном и Тейгельсом[50]

$$\mathsf{P}_{00}(t) - (1-\rho) \sim \lambda(1-\rho)^{1-a}(a-1)^{-1}x^{-(a-1)}\ell(x), \quad a > 1, \qquad x \to \infty. \tag{3.39}$$

3.2. Нестационарное совместное распределение числа требований и времени пребывания

Метод разложения на элементы задержки оказался настолько мощным, что его дальнейшее усовершенствование позволило получить точное решение задачи, указанной в заголовке этого подраздела. Приведем результаты точного решения проблемы вычисления совместного распределения сл.в. $V(t,u)$ (времени пребывания некоторого требования длины u, поступившего в момент t) и $L(t)$ (числа требований в момент $t-$) в системе M/GI/1—EPS [221,222], [223,215], [216,217].

Пусть $\zeta = \inf(t > 0 : \ L(t) = 0)$ and $\pi(s) = \mathsf{E}[\mathrm{e}^{-s\zeta}]$ — ПЛС распределения стандартного периода занятости (см. теорему 2.1), т.е. единственное положительное решение (с наименьшим абсолютным значением) функционального уравнения (2.4). Введем определение

$$\tilde{v}_0(z,r,s,u) \doteq \int_0^\infty \mathrm{e}^{-st}\mathsf{E}\left[\mathrm{e}^{-rV(t,u)}z^{L(t)} \mid L(0) = 0\right] dt \quad (\mathrm{Re}\ s, r > 0,\ |z| \le 1). \tag{3.40}$$

Эта функция определяет двумерное совместное (нестационарное) распределение процессов $\{V(t,u)\}$ и $\{L(t)\}$ при нулевом начальном условии. Для системы M/GI/1—EPS невозможно получить распределения процессов $\{V(t,u)\}$ или $\{L(t)\}$ в явном виде. Наилучшее, что можно сделать, это найти (многомерные) преобразования, ассоциированные с этими распределениями. Напомним, что даже в стационарном режиме известен только явный вид распределения числа требований. Наш обобщенный результат представлен утверждением Справедлива

Теорема 3.6. *(1997) [221,222,223,215] Для любой положительной загрузки $\rho = \lambda\beta_1$, точное решение для функции $\tilde{v}_0$ имеет вид:*

$$\tilde{v}_0(z,r,s,u) = \tilde{p}_{00}(s)\frac{\delta(r,u)}{1 - \tilde{a}(z,r,s,u)/\psi(r,u)} \tag{3.41}$$

где $\tilde{p}_{00}(s) = [s+\lambda-\lambda\pi(s)]^{-1}$, $\delta(r,u) = \mathrm{e}^{-u(r+\lambda)}/\psi(r,u)$, $\psi(r,u)$ дается своим преобразованием Лапласа по u (аргумент q):

$$\tilde{\psi}(r,q) = \frac{q+r+\lambda\beta(q+r+\lambda)}{(q+r+\lambda)(q+\lambda\beta(q+r+\lambda))} \quad (r \ge 0,\ q > -\lambda\pi(r)), \tag{3.42}$$

$$\begin{aligned}\tilde{a}(z,r,s,u) = z\lambda\Bigg\{&\psi(r,u) * \left[\mathrm{e}^{-u(r+\lambda\pi(s)-s)}\int_u^\infty \mathrm{e}^{-y(s+\lambda-\lambda\pi(s))}\,dB(y)\right] \\ &+ \mathrm{e}^{-u(r+\lambda\pi(s)-s)}\int_u^\infty \mathrm{e}^{-y(s+\lambda-\lambda\pi(s))}(1-B(y))\,dy\Bigg\}.\end{aligned} \tag{3.43}$$

[50] Возможно, их статья опубликована в *Adv. Appl. Probab.* примерно в 1996–1998 г.г. Эти номера журнала были недоступны для нас, поэтому оказалось легче вывести (3.39) снова.

Здесь $*$ — символ стилтьесовской свертки and $\pi(s)$ — минимальное решение функционального уравнения (2.4).

Комментарии к доказательству. Доказательство основывается на распространении наших аргументов из [105], [199], [201], [208] (см. §2.3 и теоремы 2.4, 2.5 и 3.2) на рассматриваемый случай. Оно использует также объединение приемов работы с обрывающимся периодом занятости, случайной замены времени, а также некоторые конструкции из теории восстановления, регенерирующих процессов и теории ветвящихся процессов. Частные случаи $z = 1 \quad (t \to \infty,\ \rho < 1)$ и $r = 0 \quad (0 < \rho < \infty)$ обсуждались выше (see also [207,211,212]). Эти статьи содержат также ряд важных понятий, которые являются (до некоторой степени) краеугольными камнями метода анализа. Чтобы восстановить опущенные шаги доказательства, полезно использовать указанные выше доказательства частных случаев в качестве промежуточных этапов. В сущности, мы вывели выражение для $\tilde{v}_0$ с помощью того же метода декомпозиции на элементы задержки, т.е. посредством представления условного времени пребыывания $V(t,u)$ и числа требований $L(t)$ в момент t в виде некоторого обобщенного функционала от (нетривиального) ветвящегося процесса на первом периоде занятости[51]. Используя структуру ветвящегося процесса, мы вывели и решили в терминах многомерных преобразований систему дифференциальных уравнений с частными производными первого порядка, определяющую $\tilde{v}_0$ на первом периоде занятости. Окончательный вид $\tilde{v}_0$ на всей положительной полуоси времени был найден с помощью стандартных приемов теории восстановления. Решение содержит контурные интегралы Бромвича, т.е. операторы обращения преобразований Лапласа $\mathcal{L}^{-1}$ (например, $\psi(r,u) = \mathcal{L}^{-1}(\tilde{\psi}(r,q))$ в (3.43)).

Замечание 3.6. Теорема 3.6 может быть представлена в нескольких эквивалентных формах. Здесь выбран вид представления результата, напоминающий формулу (5.2) из [206] (см. также [221, Th. 3], где этот результат появляется в другом виде). Результат кажется довольно неудобным для приложений. Однако мы покажем в §3.3, что такое мнение может оказаться несколько ошибочным.).

Замечание 3.7. Этот результат (среди других) обсуждается также в [216], которая включает более детальное доказательство теоремы 3.6, чем представлено здесь. Однако, (сокращенная) аргументация из [216] включена ниже как дополнительный комментарий к доказательству.

Дополнительный комментарий к доказательству. Одна из основных идей доказательства состоит в изучении аддитивного функционала, описывающего накопление во времени достигнутого обслуживания у некоторого помеченного виртуального требования

$$X(t) = \int_0^t \frac{1}{1 + L(y)}\, dy$$

где $L(y)$ есть число требований в момент y (ср. с уравнением (2.34) в [212]). (Можно показать, что процесс $\{L(\cdot)\}$ является положительно рекуррентным и эргодическим по Харрису при $\rho < 1$. Это не ставит проблем при $\rho > 1$, этот и родственные процессы будут транзиентными в таком случае.) Если определить для $u \geq 0$

$$V(0,u) = \inf\,(t \geq 0 :\ X(t) \geq u),$$

[51] Связь с ветвящимися процессами не является классической, как в работах Бореля [33] и Кендалла [99]. Расматриваемый ветвящийся процесс был описан в §2.3 и §3.1. Он напоминает ветвящийся процесс Крампа—Мода–Ягерса (см., например, Jagers [85], Kovalenko, Kuznetsov и Shurenkov [114, §13.7], Vatutin и Zubkov [182]).

то $V(0,u)$ будет временем пребывания помеченного виртуального требования, поступающего в момент 0. С другой стороны, время пребывания $V(t,u)$ удовлетворяет уравнению

$$\int_t^{t+V(t,u)} \frac{1}{1+L(y)}\,dy = u.$$

Изучение таких процессов на первом периоде занятости позволяет декомпозировать $V(t,u)$ на некоторые обрывающиеся (суб)периоды занятости (введенные в [199, 201] и упомянутые в замечании 2.6) и вывести дифференциальные уравнения с частными производными, описывающими эволюцию этих обрывающихся периодов занятости. Дело облегчается, если ввести новый масштаб времени посредством случайной замены времени (такой прием из теории вероятностей использован в для анализа моделей разделения процессора в [105] (1979), см. также [202] (1984), [203, 208, 212, 225]). В новом масштабе времени система M/GI/1—EPS преобразуется в систему $\tilde{M}/G/\infty$, в которой интенсивность входа равна $\lambda L(t)$ (и равна λ, если $L(t)=0$), см. замечание 2.2. После решения соответствующих уравнений (обобщенные варианты уравнений (2.14) и (2.15)[52]), можно использовать аргументацию из теории восстановления, чтобы связать $\mathsf{E}[\mathrm{e}^{-sV(t,u)}\mathbf{1}_{(\zeta>t)}]$ на первом периоде занятости с $\mathsf{E}[\mathrm{e}^{-sV(t,u)}]$ на всей положительной полуоси, см. текст после схемы доказательства теоремы 3.4. □

Замечание 3.8. Теорема 3.6 справедлива не только для условия устойчивости $\rho<1$, но и для $\rho\geq 1$. Все транзиентное и равновесное поведение процессов в системе M/G/1—EPS содержится в этой теореме, и мы можем вывести из нее большинство (если не все) известных и новых результатов. Кроме того, теорема 3.6 содержит все утверждения предыдущих подразделов как частные случаи. Некоторые из них будут также приведены в §3.3.

Замечание 3.9. Решение для других начальных условий получается более громозким, оно частично отражено (только для случайной величины $V(t,u)$) в [232] (мы не обсуждаем этот результат в статье).

Замечание 3.10. Прорыв в проблему нахождения распределения времени пребывания в системе M/GI/1—EPS произошел относительно недавно в in [104, 199] (конец семидесятых годов). В то время, за исключением авторов [104], только немногие другие исследователи (например, А.Д.Соловьев, Р.Шассбергер и, возможно, Ф.Фостер и Р.Вулф) осознавали внутреннюю тяжесть этой задачи и трудности, возникающие при попытках получить ее решение(я). В течение пары десятилетий, теоремы 2.4 и 2.5 были повторно переоткрыты, но в целом развитие теории разделения процессора затормозилось, за исключением, например, работ [157,158,207,208,209,211], Y92,FK,Gr91,RS93,Se92, [195,221,222,223]. (Асимптотика и предельные теоремы являются, в основном, побочными результатами.) Клейнрок написал [108, p. 226]:

> "... наше изучение периода занятости в действительности является изучением неустановившихся явлений и это одна из причин того, что развитие теории очередей затормозилось (увязло в болоте)".

Это замечание Клейнрока дополнительно объясняется следствием 3.2, но его слова имеют более глубокий смысл. Аналогичные, но более сильные причины (ср. с замечанием 2.6) характеризуют состояние теории разделения процессора. Теперь теоремы 3.3—3.5 и 3.6 дают возможности выбраться из "болота" на правильную дорогу несмотря на "лапласовский занавес" Кендалла[53]. Возможно, правильная дорога приведет к другому "болоту", например, к

[52] См. также комментарии к доказательству теоремы 3.2 и другие объяснения к теоремам 3.3 и 3.4 в §3.1.

[53] Многие результаты в теории очередей выводятся часто в терминах преобразований. Сорок лет назад это побудило Кендалла [100] сделать свое знаменитое замечание про "лапласовский занавес, который затемнил много деталей теоретической сцены в теории очередей".

численному обращению многомерных преобразований Лапласа или к аппроксимациям в духе замечания 3.5, но это уже будет намного более знакомым и неглубоким "болотом". Один подход к взгляду за лапласовский занавес состоит в использовании асимптотического анализа для развития аппроксимаций, часто с помощью преобразований, см. [45] и §2.10, 2.11. В основном, асимптотический анализ применяется для стационарных характеристик производительности. Метод численного обращения ПЛ и производящих функций, алгоритмы которого основаны на рядах Фурье, описывается в [5]. Даже двумерные преобразования могут быть эффективно обращены численно по крайней мере для более простых дисциплин обслуживания, таких как FCFS или LCFS–P. Это делает возможным численно находить распределение зависимой от времени оставшейся работы (т.е. виртуальное время ожидания) в системе M/GI/1—FCFS с достаточно высокой точностью и небольшим временем счета. Читатель может также получить представление о других работах в этой области из обзоров [57,4], а также недавней статье ден Исегера [83]. Статья [83] содержит ряд интересных результатов про алгоритмы численного обращения преобразований с помощью гауссовских квадратур.

3.3. Некоторые важные следствия.

Этот подраздел дает ряд примеров для демонстрации возможностей приведенных выше теорем. Приступим к некоторым следствиям теоремы 3.6.

Следствие 3.6. *(1989) [208, 210] Для любого положительного ρ, преобразование Лапласа (по t, аргумент s) вероятностной производящей функции (аргумент z) числа требований в момент t, имеет вид:*

$$\tilde{v}_0(z,0,s,u) = g_0(z,s) \doteq \int_0^\infty \mathrm{e}^{-st}\mathsf{E}\left[z^{L(t)} \mid L(0)=0\right] dt = \frac{1}{s+\lambda(1-z)(1-\pi(s))}. \tag{3.44}$$

Proof. Полагая $r=0$, (3.43) сводится к $\tilde{a}(z,0,s,u) = z\tilde{p}_{00}(s)(\lambda-\lambda\pi(s))\mathrm{e}^{-\lambda u}$. Теперь результат вытекает из (3.41). □

Следствие 3.6 совпадает с теоремой 3.5 (см. также [208, p.100], [211, Th. 3.2], [212, Th. 2.18]).

Замечание 3.11. Тот же самый результат справедлив для системы M/GI/1—LCFS–P (см. [228] (2005)). Появление этого результата в случае системы M/GI/1—LCFS–P было предсказано Соловьевым в [167, p.174] (1981). Кроме того, такая формула (но только для системы M/GI/1—LCFS–P) может быть извлечена из результатов [119] после некоторых (не простых) математических манипуляций. Автор [119] могла бы подойти ближе к (3.44), если бы она рассматривала систему M/GI/1—LCFS–P под несколько другим углом.

Следствие 3.7. *(2005) [228] Пусть T будет экспоненциально распределенной случайной величиной с параметром $s>0$. Тогда $L(T)$ есть геометрически распределенная случайная величина с вероятностью успеха $s\tilde{p}_{00}(s)$, где $\tilde{p}_{00}(s)$ дается равенством (3.23). В противоположность стационарному случаю (см. следствие 2.2), $\mathsf{E}[z^{L(T)}]$ зависит от всего распределения $B(x)$.*

Proof. Правая часть равенства (3.44) может быть переписана как $\mathsf{E}[z^{L(T)}]/s$ (здесь числитель есть обычная производящая функция числа требований, выбранных в экспоненциально распределенный момент времени T с параметром s). Отметим, что выборка в экспоненциально распределенный момент времени T с параметром $s>0$ эквивалентна взятию ПЛС по времени [110,153]. Теперь из (3.44) вытекает, что

$$\mathsf{E}[z^{L(T)}] = \sum_{n=0}^{\infty} \frac{s}{s+\lambda-\lambda\pi(s)} \left(\frac{\lambda-\lambda\pi(s)}{s+\lambda-\lambda\pi(s)}\right)^n z^n, \tag{3.45}$$

где $\frac{\lambda-\lambda\pi(s)}{s+\lambda-\lambda\pi(s)}$ совпадает с $1-s\tilde{p}_{00}(s)$. Важно, что $\frac{\lambda-\lambda\pi(s)}{s+\lambda-\lambda\pi(s)}$ найдена также в (3.19) нетривиальным способом: как некоторая асимптотика, ассоциированная с обрывающимся процессом восстановления в системе M/GI/1—EPS (см. также [208, p.98], [211, p.202] или [226, eq.(2.14)]), и она совпадает с вероятностью обрыва соответствующего обрывающегося процесса восстановления. □

Замечание 3.12. Следствие 3.7 может быть использовано как отправной пункт для исследования свойств выходящего процесса из системы M/GI/1—EPS в нестационарном режиме. В настоящее время известно только распределение момента первого выхода при начальном условии $\mathsf{P}(L(0)=1)$, которое выводится из следствия 3.6 (точнее, из теоремы 3.6). Мы не обсуждаем этот результат. В отличие от случая стационрного режима, проблема изучения транзиентного выходящего процесса далека от своего решения. Кстати, изучение стационарного выходящего процесса только относительно недавно привело к неожиданному успеху. Теперь хорошо известно, что система M/GI/1—EPS имеет пуассоновский выходящий процесс в установившемся состоянии [97, 105, 198, 208]. Последнее даже имеет место для более широкого класса дисциплин [198], который включает в себя класс *симметричных* дисцилин Келли как частный случай (см. также [208, 212]). Кроме того, метод доказательства пуасссоновского характера стационарного выходящего процесса из системыan M/GI/1 [198, 208] полностью отличен от метода Келли. Конечно, следует исключить дисциплину FCFS. Хорошо известно, что дисциплина FCFS не превращает пуассоновский вход в пуассоновский выход, когда $B(x)$ произвольно. Стационарный выходящий процесс в этом случае не является даже рекуррентным процессом. Другие более ранние результаты по стационарным выходящим процессам из классических систем обслуживания содержатся в [55].

Следствие 3.8. *(1997) [221, 223] При любом положительном ρ, ПЛ (по t, аргумент s) преобразования Лапласа –Стилтьеса (аргумент r) виртуального времени пребывания $V(t,u)$ требования, поступающего в момент t с длиной u, дается равенством (3.41), где $\tilde{a}$ находится из (3.43) как $\tilde{a}(1,r,s,u)$. Мы опускаем формулу для $\tilde{a}$, так как она имеет почти тот же самый вид, как (3.43).*

Proof. Положим $z=1$ in (3.41). □

Замечание 3.13. При $z=1$, теорема 2.1 в [232] совпадает с этим следствием. Однако, основной результат [232, §4, Th. 4.1] в действительности был полнее, чем представленный здесь вариант теоремы 3.6, покрывая случай, когда имеется $K\geq 0$ дополнительных постоянно существующих требований бесконечной длины. Заинтересованный читатель может обратиться к этому результату, который доступен через Интернет. Отметим, что другие исследования очередей с разделением процессора при $K>0$ перманентных требованиях рассматривают только стационарный режим, см., например, [25, 36, 41, 191].

Если снять условие по u в уравнении (3.41) посредством усреднения по $B(\cdot)$, то это даст также безусловное распределение нестационарного виртуального времени пребывания $V(t)$:

$$\int_0^\infty \mathrm{e}^{-st}\mathsf{E}\left[\mathrm{e}^{-rV(t)}|L(0)=0\right]\,dt=\tilde{p}_{00}(s)\int_0^\infty \frac{\delta(r,u)}{1-\tilde{a}(r,s,u)/\psi(r,u)}\,dB(u), \tag{3.46}$$

где $\psi(r,u)=\mathcal{L}^{-1}(\tilde{\psi}(r,q))(r,u)$ и $\tilde{\psi}(r,q)$ дается равенством (3.42).

Формула (3.46) намного сложнее хорошо известного результата Такача для нестационарного распределения виртуального времени ожидания в системе M/GI/1—FCFS при нулевых начальных условиях (см. [175]):

$$\int_0^\infty \mathrm{e}^{-st}\mathsf{E}\left[\mathrm{e}^{-rW(t)}|L(0)=0\right]\,dt=\tilde{p}_{00}(s)\frac{s+\lambda-\lambda\pi(s)-r}{s+\lambda-\lambda\beta(r)-r}. \tag{3.47}$$

Мы снова обнаруживаем из (3.46) и (3.47) как модель EPS, как это ей свойственно, отображает свою скрытую нетривиальную натуру.

Такач получил (3.47)[54] посредством составления и решения своего интегро–дифференциального уравнения для процесса оставшеся (незаконченной) работы в системе M/G/1—FCFS [175]. Однако его метод не срабатывает для получения (3.46). Поэтому мы выводим (3.46) посредством другого аналитического метода декомпозиции на элементы задержек, который обсуждался выше. Отметим, что метод Такача имеет еще один недостаток: его нельзя использовать для вычисления характеристик длины очереди. Вероятностная интерпретация уравнения Такача дана Ранебургом [153].

Из (3.47) вытекает простая формула для ПЛ по t функции $\mathsf{E}[W(t)]$ [175, p.55]

$$w_1(s)\doteq\int_0^\infty \mathrm{e}^{-st}\mathsf{E}[W(t)]dt=\frac{\tilde{p}_{00}(s)}{s}+\frac{\rho-1}{s^2}. \tag{3.48}$$

Другими словами,

$$\mathsf{E}[W(t)|W0=0]=\int_0^t \mathsf{P}_{00}(y)dy+(\rho-1)t. \tag{3.49}$$

Следствие 3.9. *(1999) [222, 223, 215]* $\mathsf{E}[V(t,u)]$ *определяется своим LT по to* t *(argument* s*)*

$$v_1(s,u)=\frac{\delta_1(u)}{s^2\tilde{p}_{00}(s)}-\frac{\lambda\delta_1(u)}{s^2\tilde{p}_{00}(s)}*\left[\mathrm{e}^{u/\tilde{p}_{00}(s)}\int_u^\infty \mathrm{e}^{-y/\tilde{p}_{00}(s)}dB(y)\right], \tag{3.50}$$

где $\delta_1(u)=u+\int_0^u\sum_{n=1}^\infty\rho^nF^{n*}(x)dx$ *for* $\rho<\infty$*. Здесь* $\tilde{p}_{00}(s)$ *дано выше в следствии 3.1 (это ПЛ вероятности опустошения системы в момент* t*:* $\mathsf{P}\{L(t)=0\}$*), а* $F^{n*}(x)$ *есть* n*–кратная свертка функции распределения* $F(x)=F^{1*}(x)=\beta_1^{-1}\int_0^x(1-B(y))dy$*, см. (2.35).*

Proof. Из (3.41) при $z=1$ через $v_1(s,u)=-\lim_{r\downarrow0}\partial\tilde{v}_0(1,r,s,u)/\partial r$. □

Корректный вид формулы для $\delta_1(u)$ был впервые получен в [105] (см. текст после следствия 2.5 для комметариев к выводу). Напомним еще раз две другие эквивалентные формы $\delta_1(u)$: выражение (2.32)

$$\delta_1(u)=u+\int_0^u(u-x)\sum_{n=1}^\infty\rho^nf^{n*}(x)dx$$

и выражение (b) в замечании 2.10

$$\delta_1(u)=(1-\rho)^{-1}\int_0^u W(x)dx,$$

где $W(x)$ есть стационарное распределение времени ожидания в системе M/GI/1—FCFS, которое представлено равенством (2.34) (см. также замечание 2.15).

Замечание 3.14. Когда $\rho<1$, из следствия 3.9 вытекает известный результат для (условного) стационарного среднего времени пребывания $\mathsf{E}[V(u)]=u/(1-\rho)$ при $t\to\infty$ с помощью применения классической тауберовой теоремы.

[54] Мы не рассматриваем здесь случай $W(0)\neq0$ в системе FCFS, см. Киприэноу [116] (1971) для интересных дополнительных результатов (относительно квази–стационарного распределения виртуального времени ожидания в системе M/GI/1—FCFS) и дальнейших подробностей.

Замечание 3.15. Формула (3.50) намного сложнее, чем ПЛ функции $\mathsf{E}[W(t)]_{FCFS}$ (см. (3.48)). В противоположность известным аппроксимациям функции $\mathsf{E}[W(t)]_{FCFS}$ (см. замечание 3.5), до сих пор ничего не известно про любую аппроксимацию нестационарного среднего времени пребывания $\mathsf{E}[V(t,u)]_{EPS}$ несмотря на существование аппроксимаций для $\mathsf{P}_{00}(t)$.

Следствие 3.10. *При $\rho < 1$ ПЛС распределения стационарного времени пребывания $(V(u))$ в системе M/GI/1—EPS требования с длиной u, выражается как*

$$v(r,u) \doteq \mathsf{E}[\mathrm{e}^{-rV(u)}] = \frac{(1-\rho)\delta(r,u)}{1-\tilde{a}(1,r,0,u)/\psi(r,u)}\,, \tag{3.51}$$

где

$$\tilde{a}(1,r,0,u) = \lambda\psi(r,u) * \left[\mathrm{e}^{-u(r+\lambda)}(1-B(u))\right] + \lambda \mathrm{e}^{-u(r+\lambda)} \int_u^\infty (1-B(x))dx. \tag{3.52}$$

Proof. При $\rho < 1$ имеем $\pi(0) = 1$. Следовательно, при $z = 1$ результат вытекает из (3.41) с помощью применения классической тауберовой теоремы

$$\lim_{s\downarrow 0} s\tilde{v}_0(1,r,s,u) = \lim_{t\to\infty} \frac{1}{t}\int_0^t \mathsf{E}[\mathrm{e}^{-rV(t,u)}]dt = v(r,u),$$

принимая также во внимание правило Лопиталя. □

Нетрудно проверить, что это следствие совпадает с теоремой 2.5.

4. ЗАКЛЮЧИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

Можно было бы сказать больше про систему M/GI/1—EPS, но для целей этой обзорной статьи сказано вполне достаточно. Мы дали обзор наиболее важных достижений по тточным (и асимптотическим) решениям для системы M/GI/1 с эгалитарным разделением процессора (EPS), включая распределение длительностей пребывания. Ключевые стационарные результаты, их сокращенные доказательства (или схемы доказательств) были суммированы в §2.2. Было чрезвычайно трудно получить распределение времени пребывания при EPS даже в стационарном режиме. Тем не менее, мы обсудили даже нестационарные (транзиентные) решения и ключевые идеи доказательств в §2.3. Упор был сделан на новые аналитические методы и приемы. Статья обеспечила также унифицированное руководство по современной математической теории системы обслуживания M/GI/1 с эгалитарным разделением процессора. Мы также пытались восстановить историческую правду относительно вклада в теорию системы M/GI/1—EPS с помощью указаний на статьи, в которых были впервые решены ключевые проблемы.

Для обсуждения других аспектов анализа систем обслуживания с эгалитарным разделением процессора For a discussion of other aspects of analysis of egalitarian processor sharing queues (или близко родственных систем), не обсуждавшихся в данной статье, мы отсылаем интересующегося читателя к Баччелли и Бремо [20] (2003), Брандт и Брандт [37] (2006), Гросс и Харрис [77] (1998), Хэмпшир и др. [80] (2006), Ким и Ким [102] (2007), Ли и др. [117] (2005), Масуяма и Такине [121] (2003), Моулин [130] (2007), Серикола и др. [162] (2005) и Ву и Такаги [196] (2005).

СПИСОК ЛИТЕРАТУРЫ

1. Aalto, S., M/G/1/MLPS Compared with M/G/1/PS within Service Time Distribution Class IMRL. *Math. Methods Oper. Research*, 2006, vol. 64, no. 2, pp. 309–325.

2. Aalto, S. and Ayesta, U., On the Nonoptimality of the Foreground–Background Discipline for IMRL Service Times, *J. Appl. Probab.*, 2006, vol. 43, no. 2, pp. 523–534.

3. Abate, J., Choudhuri, G.L., and Whitt, W., Waiting-Time Probabilities in Queues with Long-Tail Service-Time Distributions, *Queueing Syst.*, 1994, vol. 16, no. 3–4, pp. 311–333.

4. Abate, J., Choudhuri, G.L., and Whitt, W., An Introduction to Numerical Transform Inversion and its Application to Probability Models, in *Computational Probability*, Grassman, W., Ed., Boston: Kluwer, 1999, pp. 257–323.

5. Abate, J., and Whitt, W., The Fourier–Series Method for Inverting Transforms of Probability Distributions, *Queueing Syst.*, 1992, vol. 10, no. 1, pp. 5–98.

6. Abate, J. and Whitt, W., Transient Behavior of the M/G/1 Workload Process, *Oper. Research*, 1994, vol. 42, no. 4, pp. 751–764.

7. Abate, J. and Whitt, W., Limits and Approximations for the M/G/1 LIFO Waiting–Time Distribution, *Oper. Res. Lett.*, 1992, vol. 20, pp. 199–206.

8. Abrahao, B., Almeida, V., Zhang, A. et al., Self–Adaptive SLA–Driven Capacity Management for Internet Services, in *Proc. IEEE/IFIP Networks Operations and Manag. Symp.*, Vancouver, 2006.

9. Abramowitz, M. and Stegun, L.A., *Handbook of Mathematical Functions*, New York: Dover, 1972.

10. Adiri, I. and Avi–Itzhak, B., A time–sharing queue, *Management Sci.*, 1969, vol. 15, No. 11, pp. 639–657.

11. Altman, E., Avrachenkov, K., and Ayesta, U., A Survey on Discriminatory Processor Sharing, *Queueing Syst.*, 2006, vol. 53, no. 1–2, pp. 53–63.

12. Altman, E., Jiménez, T., and Kofman, D., DPS Queues with Stationary Ergodic Service Times and the Performance of TCP in Overload, *Proc. IEEE INFOCOM'04*, 2004, paper 0-783-8356-7/04.

13. Artamonov, G.T. and Brekhov, O.M., *Analiticheskie veroyatnostnye modeli funktsionirovaniya EVM* (Analytical Probability Models of Computer Functioning), Moscow: Energiya, 1978 (in Russian).

14. D'Apice, C. and Pechinkin, A.V., The $BMAP_k/G_k/1$ Finite Queue with the Foreground-Background Processor-Sharing Discipline, Autom. and Remote Control, 2006, vol 67, no. 3.

15. Asmussen, S., *Applied Probability and Queues*, Heidelberg: Springer, 2003.

16. Asmussen, S., Klüppelberg, C., and Sigman, K., Sampling at Subexponential Times, with Queueing Applications, *Stochastic Processes and their Appl.*, 1999, vol. 79, pp. 265–286.

17. Asare, B.K. and Foster, F.G., Conditional Response Times in the M/G/1 Processor-Sharing System, *J. Appl. Probab.*, 1983, vol. 20, no. 4, pp. 910–915.

18. Athreya, K.B. and Ney, P.E., *Branching Processes*. New York: Springer, 1972.

19. Avi–Itzhak, B. and Halfin, S., Server Sharing with a Limited Number of Service Positions and Symmetric Queues, *J. of Appl. Probab.*, 1987, vol. 24, No. 4, pp. 990–1000.

20. Baccelli, F. and Brémaud, P., *Elements of Queueing Theory*, Heidelberg: Springer, 2003.

21. Bailey, N.T.J., A Continuous Time Treatment of a Simple Queue Using Generating Functions, *J. Royal Statist. Soc., ser.B*, 1954, vol. 16, pp. 288-291.

22. Bansal, N., Analysis of the M/G/1 Processor Sharing Queue with Bulk Arrivals, *Oper. Res. Lett.*, 2003, vol. 31, pp. 401–405.

23. Basharin, G.P. and Tolmachev, A.L., Theory of Queueing Networks and its Application to the Analysis of Information-Computing Systems, *J. Soviet Math.*, 1985, vol. 99, no. 1, pp. 951–1050.

24. Beneš, V.E., On Queues with Poisson Arrrivals. *Annals Math. Statist.*, 1957, vol. 28, pp. 670-677.

25. van den Berg, J.L., Sojourn Times in Feedback and Processor-Sharing Queues. *PhD Dissertation*, Utrecht: Rijksuniversiteit, 1990.

26. Bhat, U.N., *Introduction to Queueing Theory (Draft of the Lecture Course)*, Dallas, 2005. Available at http://www.faculty.smu.edu/nbhat.

27. Bhat, U.N., Shalaby, M., and Fisher, M.J., Approximation Techniques in the Solution of Queueing Problems, *Naval Research Logist. Quart.*, 1979, vol. 26 , pp. 311–326.

28. Bingham, N.H., Goldie, C.M., and Teugels, J.L., *Regular Variation,* vol. 27 of *Encycl. of Math. and its Appl.*, Cambridge: Cambridge Univ. Press, 1987.

29. Bingham, N.H. and Doney, R.A., Asymptotic Properties of Super–Critical Branching Processes. I, *Adv. Appl. Probab.*, 1974, vol. 6, no. 4, pp. 711–731.

30. Blanc, J.P.C. and van Doorn, E.A., Relaxation times for queueing systems, in *Mathematics and Computer Science*, Hazewinkel, M. and Lenstra, J.K., Eds, Amsterdam: North Holland, 1986, pp. 139–162.

31. Blumenthal, R. and Getoor, R., *Markov Processes and Potential Theory,* New York: Academic Press, 1968.

32. Bocharov, P.P. and Pechinkin, A.V., *Teoriya massovogo obsluzhivaniya* (Queueing Theory), Moscow: Univ. Druzhby Narodov, 1995 (in Russian).

33. Borel, E., Sur l'Emploi du Théorème de Bernoulli Pour Faciliter le Calcul d'Une Infinité de Coefficients. Application au Problème de l'Attente à un Guichet, *Comptes Rendus Hebd. des Séanc. de l'Académie des Sciences*, 1942, vol. 214, pp. 452–456.

34. Borovkov, A.A., *Asymptotic Methods in Queueing Theory*, New York: Wiley, 1984 (Translation from the Russian edition of 1980).

35. Borst, S.C, Boxma, O.J., and Nunez–Queija, R., Heavy tails: the effect of the service discipline, in *Computer Performance Evaluation — Modelling Techniques and Tools. Proc. 12th Int. Conf. (London, Apr. 2002)*, Field, T. et al., Eds., Berlin: Springer, 2002, pp. 1–30.

36. Brandt, A. and Brandt, M., On the Sojourn Times for Many-Queue Head-of-the-Line Processor-Sharing Systems with Permanent Customers. *Math. Methods Oper. Research*, 1998, vol. 47, pp. 181–220.

37. Brandt, A. and Brandt, M., A Sample Path Relation for the Sojourn Times in G/G/1–PS Systems and its Applications, *Queueing Syst.*, 2006, vol. 52, no. 4, pp. 281–286.

38. Breiman, L., On Some Limit Theorems Similar to Arc–Sin Law, *Theor. Prob. Appl.*, 1965, vol. 10, pp. 323–331..

39. Chen, H., Kella, O., and Weiss, G., Fluid Approximation for a Processor–Sharing Queue, *Queueing Syst.*, 1997, vol. 27, no. 1–2, pp. 99–125..

40. Chernoff, H., A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations, *Ann. Math. Statist.*, 1952, vol. 23, no. 4, pp. 493–507.

41. Cheung, S.-K., van den Berg, J.L., and Boucherie, R.J., Decomposing the Queue Length Distribution of Processor-Sharing Models into Queue Lengths of Permanent Customer Queues, *Proc. of IFIP Performance'2005*, Juan-les-Pins, France, Oct. 3–7, 2005.

42. Cheung, S.-K., van den Berg, J.L., and Boucherie, R.J., Insensitive Bounds for the Moments of the Sojourn Time Distribution for the M/G/1 Processor-Sharing Queue, *Queueing Syst.*, 2006, vol. 53, no. 1–2, pp. 7–18.

43. Chistyakov, V.P., A Theorem on Sums of Independent, Positive Random Variables and its Applications to Branching Processes, *Theor. Prob. Appl.*, 1964, vol. 9, pp. 640–648.

44. Coffman, E.G., Muntz, R., and Trotter, H., Waiting Time Distributions for Processor-Sharing Systems. *J. Assoc. Comput. Mach.*, 1970, vol. 17, no. 1, pp. 123–130.

45. Cohen, J.W., Asymptotic Relations in Queueing Theory, *Stoch. Proc. Appl.*, 1973, vol. 1, pp. 107–124.

46. Cohen, J.W., *The Single Server Queues*, Amsterdam: North-Holland, 1982.

47. Cohen, J.W., On Processor Sharing and Random Service (Letter to the editor), *J. Appl. Probab.*, 1984, vol. 21, no. 4, pp. 937–937.

48. Conway, R.W., Maxwell, V.L. and Miller, L.W., *Theory of Scheduling.* Reading MA: Addison–Vesley, 1967. Russian edition: Conway R.W., Maxwell V.L., Miller L.W. *Theory of Scheduling.* Ed. Basharin G.P. Moscow: Nauka, 1975.

49. Cooper, R.B., *Introduction to Queueing Theory,* New York: North-Holland, 1981, 2nd edition.

50. Cooper, R.B., Queueing theory, in *Handbooks in Oper. Res. and Manag. Sci.*, Heyman, D.P. and Sobel, M.J., Eds., New York: Elsevier, 1990, vol. 2, pp. 469–518.

51. Cooper, R.B. and Shun-Chen Niu, Beneš's Formula for M/G/1—FIFO "Explained" by Preemptive Resume LIFO, *J. Appl. Probab.*, 1986, vol. 23, no. 2, pp. 550–554.

52. Cox, D.R. and Smith, W.L., *Queues*, London: Methuen, 1961.

53. Daduna, H., *Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Queueing Networks* (Lect. Notes on Comput. Sci., vol. 2046), Heidelberg: Springer, 2001.

54. Daduna, H. and Schassberger, R., A Discrete–Time Round–Robin Queue with Bernoulli Input and General Arithmetic Service Time Distributions, *Acta Informatica*, 1981, vol. 15, no. 3, pp. 251–263.

55. Daley, D.J., Queueing Output Processes, *Adv. Appl. Probab.*, 1976, vol. 8, pp. 395–415.

56. van Dantzig, D., Sur la Méthode des Fonctions Génèratrices, *Colloques Internationaux du CNRS*, 1948, vol. 13, pp. 29–45.

57. Davies, B., and Martin, B.L., Numerical Inversion of the Laplace Transform: A Survey and Comparison of Methods, *J. Comput. Phys.*, 1979, vol. 33, pp. 1–32.

58. Disney, R.L. and Kiessler, P., *Traffic Processes in Queueing Networks. A Markov Renewal Approach,* Baltimore: John Hopkins Univ. Press, 1987.

59. Doetsch, G., *Introduction to the Theory and Applications of the Laplace Transformation,* New York: Springer, 1974.

60. Dshalalow, J.H., An Antology of Classical Queueing Methods, in *Advances in Queueing: Theory, Methods, and Open Problems,* Dhalalow, J.H., Ed., Boca Raton: CRC Press, 1995, pp. 1–42.

61. Egorova, R., Zwart, B., and Boxma, O., Sojourn Time Tails in the M/D/1 Processor Sharing Queue, *Probab. Engrg. Inf. Sci.*, 2006, vol. 20, no. 3, pp. 429–446.

62. Embrechts, P., Klüppelberg, C., and Mikosch, T., *Modelling Extremal Events for Insurance and Finance.* Heidelberg: Springer, 1997.

63. Esscher, F., On the Probability Function in the Collective Theory of Risk, *Skandinavisk Aktuarietidskrift.*, 1932, vol. 15, pp. 175–195.

64. Fayolle, G., Mitrani, I. and Jasnogorodski, R., Sharing a Processor among Many Job Classes, *J. Assoc. Comput. Mach.*, 1980, vol. 27, no. 3, pp. 519–532.

65. Feller, W., *Introduction to Probability Theory and its Applications,* New York: Wiley, 1966, vol. 2.

66. Flatto, L., The Waiting Time Distribution for the Random Order Service M/M/1 Queue, *Ann. Appl. Probab.*, 1997, vol. 7, no. 2, pp. 382–409.

67. Foley, R.D. and Klutke, G.-A., Stationary Increments in the Accumulated Work Process in Processor-Sharing Queues, *J. Appl. Probab.*, 1989, vol. 26, no. 3, pp. 671–677.

68. Foster, F.G., Stochastic processes, *Proc. 6th IFORS Int. Conf. Operational Research'72*, Ross, M., Ed., Amsterdam: North-Holland, 1973, pp. 223–239.

69. Franken, P., König, D., Arndt U. and Schmidt, V., *Queues and Point Processes*, Berlin: Akademie, 1981.

70. Galambosh, J., *The Asymptotic Theory of Extreme Order Statistics*, New York,: Wiley, 1978.

71. Gaver, D.P. and Jacobs, P.A., On Inference Concerning Time–Dependent Queue Performance (Invited paper), *Queueing Syst.*, 1990, vol. 6, no. 3, 261–276.

72. Glynn, P.W., Diffusion approximation, in *Handbooks in Oper. Res. and Manag. Sci.*, Heyman, D.P. and Sobel, M.J., Eds., New York: Elsevier, 1990, vol. 2, pp. 145–198.

73. Gnedenko, B.V. and Kovalenko, I.N., *Introduction to Queueing Theory*, Boston: Birkhäuser, 1991 (Translation from the Russian second edition of 1987).

74. Grishechkin, S.A., The Crump–Mode–Jagers Branching Processes as a Method for Studying the Processor-Sharing System M/G/1, *Theor. Prob. Appl.*, 1991, vol. 36, no. 1, pp. 19–35.

75. Grishechkin, S.A., On a Relationship Between Processor Sharing Queues and Crump–Mode–Jagers Branching Processes, *Adv. Appl. Probabb.*, 1992, vol. 24, no. 3, pp. 653–698.

76. Gromoll, H.C., Diffusion Approximation for a Processor Sharing Queue in Heavy Traffic, *Annals Appl. Probab.*, 2004, vol. 14, no. 2, pp. 555–611.

77. Gross, D. and Harris, C.M., *Fundamentals of Queueing Theory*, New York: Wiley, 1998.

78. Guillemin, F. and Boyer, J., Analysis of the M/M/1 Queue with Processor Sharing via Spectral Theory, *Queueing Syst.*, 2001, vol. 39, no. 4, pp. 377–397.

79. Haccou, P., Jagers, P., and Vatutin, V.A., *Branching Processes: Variation, Growth, and Extinction of Populations,* Cambridge: Cambridge Univ. Press, 2005.

80. Hampshire, R.C., Harchol–Balter, M., and Massey, W.A., Fluid and Diffusion Limits for Transient Sojourn Times of Processor Sharing Queues with Time Varying Rates, *Queueing Syst.*, 2006, vol. 53, no. 1–2, pp. 19–30.

81. Heaviside, O., *Electromagnetic Theory,* London: Electrician Co., 1893, vol. 1; 1899, vol. 2; 1912, vol. 3.

82. Iglehart, D.L. and Whitt, W., Multiple Channel Queues in Heavy Traffic. I, II, *Adv. Appl. Probab.*, 1970, vol. 2, pp. 150–177, pp. 355–369.

83. den Iseger, P., Numerical Transform Inversion Using Gaussian Quadrature, *Probab. Eng. Inf. Sci.*, 2006, vol. 20, no. 1, pp. 1–44.

84. Jagerman, D.L., *Difference Equations with Applications to Queues*, Marcel Dekker, 2000.

85. Jagers, P., *Branching Processes with Biological Applications*, London: Wiley, 1975.

86. Jaiswal, N.K., *Priority Queues*, New York: Academic Press, 1968.

87. Jaiswal, N.K., Performance Evaluation Studies for Time-Sharing Computer Systems, *Performance Evaluation*, 1982, vol. 2, no. 4, pp. 223–236.

88. Jansen, U., Conditional Expected Sojourn Times in Insensitive Queueing Systems and Networks, *Adv. Appl. Probab.*, 1984, vol. 16, no. 4, pp. 906–919.

89. Jean-Marie, A., On Overloaded Queue, *30th Int. Conf. on the Mathematics of Operations Research*, Lunteren, Netherlands, 2005.

90. Jean–Marie, A. and Robert, Ph., On Transient Behavior of the Processor-Sharing Queue, *Queueing Syst.,* 1994, vol. 17, pp. 129–136.

91. Jelenković, P. and Momčilović, P., Resource Sharing with Subexponential Distributions, in *Proc. IEEE INFOCOM'02*, New York, 2002, vol. 3, pp. 1316–1325.

92. Jelenković, P. and Momčilović, P., Large Deviation Analysis of Subexponential Waiting Times in a Processor Sharing Queue, *Math. Oper. Research*, 2003, vol. 28, no. 3, pp. 587–608.

93. Jelenković, P. and Momčilović, P., Large Deviation of Square–Root Insensitive Random Sums, *Math. Oper. Research*, 2004, vol. 29, no. 2, pp. 398–406.

94. Kalashnikov, V., *Geometric Sums: Bounds for Rare Events with Applications*, Doldrecht: Kluwer, 1997.

95. Karamata, J., Sur un Mode de Croissance Réguliére des Fonctions, *Mathematica Cluj*, 1930, vol. 4, pp. 38–53.

96. Karpelevich, F.I. and Kreinin, A.Ya., *Heavy Traffic Limits for Multiphase Queues,* Providence: AMS, 1994.

97. Kelly, F.P., *Reversibility and Stochastic Networks,* New York: Wiley, 1979.

98. Kel'bert, M.Ya. and Sukhov, Yu.M., Mathematical Theory of Queueing Networks, *J. Soviet Math.*, 1990, vol. 50, no. 3, pp. 1527–1600.

99. Kendall, D.G., Some Problems in the Theory of Queues, *J. Roy. Statist. Soc., Ser.B*, 1951, vol. 13, pp. 151–185.

100. Kendall, D.G., Some Recent Work and Further Problems in the Queueing Theory, *Theor. Prob. Appl.*, 1964, vol. 9, no. 1, pp. 1–13.

101. Kim, J. and Kim, B., Sojourn Time Distribution in the M/M/1 Queue with Discriminatory Processor Sharing, *Performance Evaluation*, 2004, vol. 58, no. 4, pp. 341–365.

102. Kim, J. and Kim, B., The Processor–Sharing Queue with Bulk Arrivals and Phase–Type Service, *Performance Evaluation*, 2007, vol. 64, no. 4, pp. 277–287.

103. Kingman, J.F.C., The Single Server Queue in Heavy Traffic, *Proc. Cambridge Philos. Soc.*, 1961, vol. 57, pp. 902–904.

104. Kitayev, M.Yu. and Yashkov, S.F., Distribution of the Conditional Sojourn Time in a System with Division of Time of Servicing, *Eng. Cybernetics*, 1978, vol. 16, no. 4, pp. 162–167.

105. Kitayev, M.Yu. and Yashkov, S.F., Analysis of a Single–Channel Queueing Systems with the Discipline of Uniform Sharing of a Device, *Eng. Cybernetics*, 1979, vol. 17, no. 6, pp. 42–49.

106. Klar, B., A Note on the $\mathcal{L}$-class of Life Distributions, *J. Appl. Probab.*, 2002, vol. 39, no. 1, pp. 11–19.

107. Kleinrock, L., Time–Shared Systems: A Theoretical Treatment, *J. Assoc. Comput. Mach.*, 1967, vol. 14, no. 2, pp. 242–251.

108. Kleinrock, L., *Queueing Systems*, vol. 1: *Theory*. New York: Wiley, 1975.

109. Kleinrock, L., *Queueing Systems,* vol. 2: *Computer Applications*, New York: Wiley, 1976.

110. Klimov, G.P., Lyakhu, A.K., and Matveev, V.F., *Matemeticheskie modeli sistem s razdeleniem vremeni* (Mathematical Models of Time Sharing Systems), Kishinev: Shtiintsa, 1983 (in Russian).

111. Knessl, Ch., A Diffusion Model for Two Parallel Queues with Processor Sharing: Transient Behavior and Asymptotics, *J. Appl. Math. Stoch. Analysis*, 1999, vol. 12, no. 4, 311–338.

112. Kobayashi, H. and Konheim, A., Queueing Models of Computer Communications System Analysis, *IEEE Trans. Commun.*, 1977, vol. 25, no. 1, pp. 2–28.

113. Kovalenko, I.N., Rare Events in Queueing Systems—A Survey, *Queueing Syst.*, 1994, vol. 16, pp. 1–49.

114. Kovalenko, I.N., Kuznetsov, N.Yu., and Shurenkov, V.M., *Models of Random Processes. A Handbook*, Skorokhod, A.V., Ed., Boca Raton: CRC Press, 1996 (Translation from the Russian edition of 1983).

115. König, D., Rykov, V.V., and Schmidt, V., Stationary Queueing Systems with Dependencies, *J. Soviet Math.*, 1983, vol. 21, no. 6, pp. 938–994.

116. Kyprianou, E.K., On the quasi–stationary distribution of the virtual waiting time in queues with Poisson arivals, *J. Appl. Probab.*, 1971, vol. 8, no. 3, pp. 494–507.

117. Li, Q.-L., Lian, Z., and Liu, L., A RG-factorization Approach for a BMAP/M/1 Generalized Processor-Sharing Queue, *Stochastic Models*, 2005, vol. 21, no. 2–3, pp. 507–530.

118. Limic, V., On the Behavior of LIFO Preemptive Resume Queues in Heavy Traffic, *Elec. Comm. Probab.*, 1999, vol. 4, pp. 13–27.

119. Limic, V., A LIFO Queue in Heavy Traffic, *Annals Appl. Probab.*, 2001, vol. 11, pp. 301–331.

120. Lipayev, V.V. and Yashkov, S.F., *Effektivnost' metodov organizazii vychislitel'nogo processa v ASU* (Efficiency of Methods of Organizing Computation process in Automatic Control Systems), Moscow: Statistika, 1975 (in Russian).

121. Masuyama, H. and Takine, T., Sojourn Time Distribution in a MAP/M/1 Processor-Sharing Queue, *Oper. Res. Lett.*, 2003, vol. 31, pp. 406–412.

122. McKinney, J., A Survey of Analytical Time-Sharing Models, *Computing Surveys*, 1969, vol. 1, no. 2, pp. 105–116.

123. Matveev, V.F. and Ushakov, V.G., *Sistemy massovogo obsluzhivaniya* (Queueing Systems), Moscow: Moscow State Univ., 1984 (in Russian).

124. Meyer, P., *Probability and Potentials*, Waltham, Mass.: Blaisdell Publ. Co., 1966.

125. de Meyer, A. and Teugels, J.L., On the Asymptotic Behaviour of the Distributions of the Busy Period and Service–Time in M/G/1, *J. Appl. Probab.*, 1980, vol. 17, pp. 802–813.

126. Mitrani, I., Response Time Problems in Communication Networks, *J. Royal Statist. Soc., ser. B*, 1986, vol. 47, no. 3, pp. 396–406.

127. Mitrani, I., *Probabilistic Modeling*, Cambridge: Cambridge Univ. Press, 1997.

128. Mishkoi, G.K., *Veroyatnosti sostoyaniya prioritetnykh sistem v nestatsionarnom regime* (Probabilities of a State of Priority Systems in a Non-Stationary Mode), Kishinev: Shtiintsa, 1979 (in Russian).

129. Morrison, J., Response Time for a Processor-Sharing System, *SIAM J. Appl. Math.*, 1985, vol. 45, no. 1, pp. 152–167.

130. Moulin, H., Minimizing the Worst Slowdown: Off-Line and On-Line, *Oper. Research*, 2007 (to appear).

131. Nabe, M., Murata, M., and Miyahara, H., Analysis and Modelling of World Wide Web traffic for Capacity dimensioning of Internet Access Lines, *Performance Evaluation*, 1998, vol. 34, pp. 249–271.

132. Nadarajah, S. and Kotz, S., On the Laplace Transform of the Pareto Distribution, *Queueig Syst.*, 2006, vol. 54, pp. 243-244.

133. Nakamura, G., Murao, Y., and Tsukamoto, K., Analysis of waiting time distribution for the TSS round–robin scheduling scheme, *Review of Electr. Commun. Lab.*, 1972, vol. 20, No. 3–4, pp. 210–219.

134. Neiman, V.I., Seti svyazi elektronnykh vychislitel'nykh mashin (Computer Communication Networks), in *Progress in Sci. and Techn., ser. Electrosvyaz'*, Moscow: Vses. Inst. Sci.–Techn. Info., 1978, vol. 9, pp. 5–119 (in Russian).

135. Newell, G.F., *Applications of Queueing Theory,* London: Chapman and Hall, 1971.

136. Núñez–Queija, R., Sojourn Times in a Processor Sharing Queue with Service interruptions, *Queueing Syst.*, 2000, vol. 34, no. 1–4, pp. 351–386.

137. Núñez–Queija, R., Sojourn Times in Non–Homogeneous QBD Processes with Processor Sharing, *Stochastic Models*, 2001, vol. 17, no. 1, pp. 61–92.

138. O'Donovan, T.M., Direct Solutions of M/G/1 Processor Sharing Models, *Oper. Research*, 1974, vol. 22, no. 6, pp. 1232–1235.

139. O'Donovan, T.M., The Queue M/G/1 When Jobs Are Scheduled within Generations, *Oper. Research*, 1975, vol. 23, no. 4, pp. 821–824.

140. Olver, F.W.J., *Introduction to Asymptotics and Special Functions*, New York: Academic Press, 1974.

141. Ott, T., The Sojourn–Time Distribution in the M/G/1 Queue with Processor Sharing, *J. Appl. Probab.*, 1984, vol. 21, no. 2, pp. 360–378.

142. Pakes, A.G., On the Tails of Waiting–Time Distributions, *J. Appl. Probab.*, 1975, vol. 12, no. 3, pp. 555–564.

143. Pollaczek, F., La Loi de l'Attente des Appels Téléphoniques, *Comptes Rendus Hebd. des Séanc. de l'Académie des Sciences*, 1946, vol. 222, pp. 352–355.

144. Prabhu, N.U., *Stochastic Storage Processes,* New York: Springer, 1980.

145. Ramaswami, V., The Sojourn Time in the GI/M/1 Queue with Processor Sharing, *J. Appl. Probab.*, 1984, vol. 21, no. 2, pp. 437–442.

146. Rege, K.M. and Sengupta, B., The M/G/1 Processor-Sharing Queue with Bulk Arrivals, in *Modelling and Performance Evaluation of ATM Tecnology*, Amsterdam: Elsevier, 1993, pp. 417–432.

147. Rege, K.M. and Sengupta, B., A Decomposition Theorem and Related Results for the Discriminatory Processor Sharing Queue, *Queueing Syst.*, 1994, vol. 18, no. 3–4, pp. 333–351.

148. Resing, J.A.C., Hooghiemstra, G., and Keane, M.S., The M/G/1 Processor Sharing Queue as the Almost Sure Limit of Feedback Queues, *J. Appl. Probab.*, 1990, vol. 27, no. 4, pp. 913–918.

149. Resnick, S. and Samorodnitsky, G., Performance Decay in a Single Server Exponential Queueing Model with Long Range Dependence, *Oper. Research*, 1997, vol. 45, pp. 235–243.

150. Rolski, T., *Twisting in Applied Probability (Draft of the Lecture Course)*, Edinburgh: Heriot–Watt Univ., Dept. of Acturial Math. and Statistics, 2004.

151. Rosenkrantz, W.A., Calculation of the Laplace Transform of the Length of the Busy Period for the M/G/1 Queue via Martingales, *Annals Probab.*, 1983, vol. 11, no. 3, pp. 817–818.

152. Rubalskii, G.B., *Upravlenie zapasami pri sluchainom sprose* (Control of Inventories under Random Demand), Moscow: Sov. Radio, 1977 (in Russian).

153. Runnenburg, J. Th., On the Use of the Method of Collective Marks in Queueing Theory, in *Proc. Symp. on Congestion Theory*, Smith, W.A. and Wilkinson, W.E., Eds., Chapel Hill: Univ. of North Carolina Press, 1965, pp. 399–438.

154. Saaty, T.L., *Elements of Queueing Theory with Applications,* New York: McGraw-Hill, 1961.

155. Sakata, M., Noguchi, S., and Oizumi, J., Analysis of a Processor Shared Model for Time Sharing Systems, *Proc. 2nd Hawaii Int. Conf. on System Sci.*, Honolulu: Univ. of Hawaii, 1969, pp. 625–628.

156. Schassberger, R., *Warteschlangen,* Wien: Springer, 1973.

157. Schassberger, R., A New Approach to the M/G/1 Processor-Sharing Queue, *Adv. Appl. Probab.*, 1984, vol. 16, no. 1, pp. 202–213.

158. Schassberger, R., The Steady State Distribution of Spent Service Present in the M/G/1 Foreground-Background Processor–Sharing Queue, *J. Appl. Probab.*, 1988, vol. 25, no. 1, pp. 194–203.

159. Sevast'yanov, B.A., *Vetvyashchiesya protsessy* (Branching Processes), Moscow: Nauka, 1971 (in Russian).

160. Sengupta, B., An Approximation for the Sojourn–Time Distribution for the GI/G/1 Processor-Sharing Queue, *Stochastic Models*, 1992, vol. 8, no. 1, pp. 35–57.

161. Sengupta, B. and Jagerman, D.L., A Conditional Response Time of the M/M/1 Processor–Sharing Queue, *AT & T Techn. J.*, 1985, vol. 64, no. 2, pp. 409–421.

162. Sericola, B., Guillemin, F., and Boyer, J., Sojourn Times in the M/PH/1 Processor Sharing Queue, *Queueing Syst.*, 2005, vol. 50, no. 1, pp. 109–130.

163. Shalmon, M., Analysis of the GI/G/1 Queue and its Variations via the LCFS preeemptive Resume Discipline and its Random Walk Interpretation, *Probab. Engrg. Inf. Sci.*, 1988, vol. 2, pp. 215–230.

164. Shalmon, M., Explicit Formulas for the Variance of Conditioned Sojourn Times in M/D/1–PS, *Oper. Res. Lett.*, 2007, vol. 35, to appear.

165. Shiryaev, A.N., *Essentials of Stochastic Finance: Facts, Models, Theory*, Singapore: World Scientific, 1999.

166. Sigman, K., A Primer on Heavy Tailed Distributions, *Queueing Syst.*, 1999, vol. 33, pp. 261–275.

167. Solov'yev, A.D., Analysis of an M/G/1 Queue under Various Queueing Disciplines, in *Teoriya massovogo obsluzhivaniya* Queueing Theory, Gnedenko, B.V. and Kalashnikov, V.V., Eds., Moscow: VNIISI, 1981, pp. 172–178 (in Russian).

168. Smith, W., Regenerative Stochastic Processes, *Proc. of the Royal Soc., Ser. A*, 1955, vol. 232, pp. 6–31.

169. Stidham, S., Analysis, Design, and Control of Queueing Systems, *Oper. Research*, 2002, vol. 50, no. 1, pp. 197–216.

170. Stoyan, D., *Qualitative eigenschaften und abschätzungen Stochastischer Modelle,* Berlin: Akademie, 1977.

171. Stoyan, D., *Comparison Methods for Queues and Other Stochastic Models*, New York: Wiley, 1983.

172. Stuck, B.W. and Arthurs, E., *A Computer & Communications Network Performance Analysis Primer,* Englewood Cliffs: Prentice-Hall, 1985. Available at http://www.signallake.com/publications/Primer.

173. Syski, R., Markov Functionals in Teletraffic Theory, in *Teletraffic Analysis and Computer Performance Evaluation*, Boxma, O.J. and Tijms, H.C., Eds., Amsterdam: North-Holland, 1986, pp. 303–317.

174. Syski, R., A Personal View of Queueing Theory, in *Frontiers in Queueing: Models and Applications in Science and Engineering*, Dshalalow, J., Ed., Boca Raton: CRC Press, 1997, pp. 3–18.

175. Takács, L., *Introduction to the Theory of Queues*, Oxford: Oxford Univ. Press, 1962.

176. Takács, L., *Combinatorial Methods in the Theory of Stochastic Processes*, New York: Wiley, 1967.

177. Takagi, H., Time–Dependent Analysis of M/G/1 Vacation Models with Exhaustive Service, *Queueing Syst.*, 1990, vol. 6, pp. 369–390.

178. Teugels, J.L., The Average Virtual Waiting Time as a Measure of Performance, *Queueing Syst.*, 1990, vol. 6, pp. 327–334.

179. Tripathi, S.K. and Duda, A., Time-Dependent Analysis of Queueing Systems, *Info. Syst. Oper. Researh (INFOR),* 1986, vol. 24, no. 3, pp. 199–219.

180. Trub, I.I., Answers to Queries to Internet: An Optimal Generation Strategy, *Autom. and Remote Control*, 2003, vol. 64, no. 6, pp. 935–942.

181. Tsitsiashvili, G.Sh., An Ergodicity Condition for a Cyclic Queueing System, *Cybern. and Syst. Anal.*, 2001, vol. 37, no. 1, pp. 149–150.

182. Vatutin, V.A. and Zubkov, A.M., Branching Processes. II, *J. Soviet Math.,* 1993, vol. 67, no. 6, pp. 3407–3485.

183. Vishnevsky, V.M., *Teoreticheskie osnovy proektirovaniya kompyuternykh setei* (The Theoretical Fundamentals for Design of Computer Networks), Moscow: Tekhnosphera, 2003.

184. Vishnevsky, V.M. and Semenova, O.V., Mathematical Methods to Study the Polling Systems, *Autom. and Remote Control*, 2006, vol. 67, no. 2, pp. 173–220.

185. Villela, D., Pradhan, P., and Rubenstein, D., Provisioning Servers in the Application Tier for E-commerce Systems, *Proc. 12th IEEE Int. Workshop on Quality of Service*, Montreal, 2004, pp. 57–66.

186. Volkonskii, V.A., Random Substitution of Time in Strong Markov Processes, *Theor. Prob. Appl.*, 1958, vol. 3, no. 3, pp. 310–325.

187. Ward, A. and Whitt, W., Predicting Response Times in Processor-Sharing Queues, in *Proc. Fields Institute Conf. on Commun. Networks*, Glynn, P., MacDonald, D., and Turner, S., Eds., Providence: AMS, 2000, pp. 1–29.

188. von Weizsäcker, H. and Winkler, G., *Stochastic Integrals,* Braunschweig: Fridr. Vieweg, 1990.

189. Wierman, A. and Harchol-Balter, M., Classifying Scheduling Policies with Respect to Higher Moments of Conditional Response Time, *Proc. ACM 2005 Sigmetrics Int. Conf. on Measurement and Modeling of Computer Systems*, Bannf, Canada, 2005, pp. 229–240.

190. Whitt, W., Heavy Traffic Limit Theorems for Queues: A Survey, *Mathematical Methods in Queueing Theory. Proc. of a Conf. at Western Michigan Univ., May 10–12, 1973* (Lect. Notes in Economics and Math. Syst., vol. 98), Heidelberg, Springer, 1974, pp. 307–350.

191. Whitt, W., The M/G/1 Processor–Sharing Queue with Long and Short Jobs, *Unpublished manuscript*, 1998 (available at http://www.research.att.com/resourses/trs/TRs/ 98/98.37/98.37.1/body.ps).

192. Whitt, W., *Stochastic–Process Limits: An Introduction to Stochastic–Process Limits and Their Applications to Queues*, New York: Springer, 2002.

193. Widder, D.V., *The Laplace Transform*, Princeton: Univ. Press, 1946.

194. Wolff, R.W., Time–Sharing with Priorities, *SIAM J. Appl. Math.*, 1970, vol. 19, no. 3, pp. 566–574.

195. Wolff, R.W., *Stochastic Modeling and the Theory of Queues,* Englewood Cliffs: Prentice-Hall, 1989.

196. Wu, D.-A. and Takagi, H., Processor–Sharing and Random-Service Queues with Semi-Markovian Arrivals, *J. Appl. Probab.*, 2005, vol. 42, no. 2, pp. 478–490.

197. Yashkov, S.F., Distribution of the Conditional Waiting Time in a System with Division of Time, *Eng. Cybernetics*, 1977, vol. 15, no. 5, pp. 44–52.

198. Yashkov S.F., Properties of Invariance of Probabilistic Models of Adaptive Scheduling in Shared–Use Systems. *Autom. Contr. and Comput. Sci.*, 1980, vol. 14, no. 6, pp. 46–51.

199. Yashkov, S.F., Some Results of Analyzing a Probabilistic Model of Remote Processing Systems, *Autom. Contr. and Comput. Sci.*, 1981, vol. 15, no. 4, pp.1–8.

200. Yashkov, S.F., System with Processor-Sharing and Its Applications in the Models of Computer Networks, *VIII All-Union Conf. on Coding and Information Transmission*, Kuibyshev, 1981, pt. 3, pp. 67–72 (in Russian).

201. Yashkov, S.F., A Derivation of Response Time Distribution for an M/G/1 Processor-Sharing Queue, *Problems of Control and Info. Theory*, 1983, vol. 12, no. 2, pp. 133–148.

202. Yashkov, S.F., Analysis of a System with Priority–Based Processor–Sharing, *Autom. Contr. and Comput. Sci.*, 1984, vol. 18, no. 3, pp. 27–36.

203. Yashkov, S.F., New Application of Random Time Change to Analysis of Processor-Sharing Queues, *4th Int. Vilnius Conf. on Prob. Theory and Math. Statistics,* Vilnius: Inst. of Math. and Cybern., 1985, vol. 4, pp. 343–345.

204. Yashkov, S.F., A Note on Asymptotic Estimates of the Sojourn Time Variance in the M/G/1 Queue with Processor-Sharing, *Syst. Analysis, Modelling, Simulation*, 1986, vol. 3, no. 3, pp. 267–269.

205. Yashkov, S.F., New Solutions for Processor–Sharing Queues, in *1-st World Congress of Bernoulli Soc. ISI.* (Tashkent, 1986), Prokhorov, Yu.V., Ed., Moscow: Nauka, 1986, vol. 2, pp. 558–558.

206. Yashkov, S.F., Processor–Sharing Queues: Some Progress in Analysis (Invited paper), *Queueing Syst.*, 1987, vol. 2, no. 1, pp. 1–17.

207. Yashkov, S.F., The Non–Stationary Distribution of Numbers of Calls in the M/G/1 Processor-Sharing Queue, *Proc. 3rd Int. Symp. on Systems Analysis and Simulation*, Berlin: Akademie, 1988, vol. 2, reprinted in *Advances in Simulation,* Lukar, P.A. and Schmidt, B., Eds., Berlin: Springer, 1988, vol. 2, pp. 158–162.

208. Yashkov, S.F., *Analiz ocheredei v EVM* (Analysis of Queues in Computers), Moscow: Radio i Svyaz', 1989.

209. Yashkov, S.F., On Method of Analysis of Processor Scharing Queue, in *Teoriya teletrafika v sistemakh informatiki* (Teletraffic Theory in Systems of Informatics), Moscow: Nauka, 1989, pp. 38–45.

210. Yashkov, S.F., On Some Characteristics of Non–Classical Models of Processor Sharing Queue, in *Modeli i metody informatsionnykh setei* (Models and Methods of Information Networks), Moscow: Nauka, 1990, pp. 14–19.

211. Yashkov, S.F., Time–Dependent Analysis of Processor-Sharing Queue, *Proc. ITC–13. Queueing, Performance and Control in ATM,* Cohen, J.W. and Pack, C.D., Eds., Amsterdam: Elsevier, 1991, pp. 199–204.

212. Yashkov, S.F., Mathematical Problems in the Theory of Shared–Processor Systems, *J. Soviet Math.*, 1992, vol. 58, no. 2, pp. 101–147 (English translation of the original Russian paper of 1990).

213. Yashkov, S.F., Some Limit Theorems for Processor Sharing Queueing Systems, in *Long–Range Tools of Telecommunication and Integrated Communication Systems*, Kuznetsov, N.A., Ed., Moscow: Inst. Problem Peredachi Inf., 1992, Pt. 1. pp. 214–220 (in Russian).

214. Yashkov, S.F., On a Heavy Traffic Limit Theorem for the M/G/1 Processor-Sharing Queue, *Stochastic Models*, 1993, vol. 9, no. 3, pp. 467–471.

215. Yashkov, S.F., Modelling Processor Sharing Queue, *Proc. of the 16th IMACS World Congress 2000 on Sci. Computation, Appl. Math. and Simulation*, Lausanne, Switzerland, 2000.

216. Yashkov, S.F., *Matematicheskie modeli sistem s razdeleniem vremeni* (Mathematical Models of Time–Sharing Systems), Moscow: Inst. Problem Peredachi Inf., 2002 (in Russian).

217. Yashkov, S.F., On Sojourn Time Problem in Processor Sharing Queue, *Int. Conf. "Kolmogorov and Contemporary Mathematics,"* Moscow: Moscow State Univ., 2003, pp. 594–595.

218. Yashkov, S.F., On Random Order of Service and Processor Sharing, *Information Processes*, 2006, vol. 6, no. 2, pp. 160–163 (available at http://www.jip.ru/).

219. Yashkov, S.F., The Moments of the Sojourn Time in the M/G/1 Processor Sharing System, *Information Processes*, 2006, vol. 6, no. 3, pp. 237–249 (available at http://www.jip.ru/).

220. Yashkov, S.F., Two Special Cases of the M/G/1—EPS Queue. *Information Processes*, 2006, vol. 6, no. 3, pp. 250–255 (available at http://www.jip.ru/).

221. Yashkov, S.F. and Yashkova, A.S., The M/G/1 Processor-Sharing System: The Transient Solutions, *Proc. 2nd Int. Conf. on Distributed Computer Commun. Networks (DCCN'97),* Tel-Aviv, 1997, Moscow: Inst. Problem Peredachi Inf., 1997, pp. 261–272.

222. Yashkov, S.F. and Yashkova, A.S., Processor Sharing Queue: Additional Results, *Proc. 3rd Int. Conf. on Distributed Comput. Commun. Networks (DCCN'99),* Tel-Aviv, 1999, Moscow: Inst. Problem Peredachi Inf., 1999, pp. 216–221.

223. Yashkov, S.F. and Yashkova, A.S., Processor Sharing Queue: Transient Solutions, *Proc. 2nd Int. Conf. on Computer Sci. and Info. Technologies,* Yerevan: Nat. Acad. of Sci. of Armeniya, 1999, pp. 99–103.

224. Yashkov, S.F. and Yashkova, A.S., On Conditions of Asymptotic Equivalence of the Tail's Distributions for Sojourn and Service Times in the M/G/1 System with Processor Sharing, *Obozrenie Prikl. Promyshl. Matem.*, 2003, vol. 10, no. 3, pp. 781–782 (in Russian).

225. Yashkov, S.F. and Yashkova, A.S., The Time–Dependent Solution of the M/G/1—FBPS Queue, *Information Processes*, 2004, vol. 4, no. 2, pp. 175–187 (available at http://www.jip.ru/).

226. Yashkov, S.F. and Yashkova, A.S., A Note on Asymptotics Associated with Limit Theorem for Terminating Renewal Process in Processor Sharing Queue, *Information Processes*, 2004, vol. 4, no. 3, pp. 256–260 (available at http://www.jip.ru/).

227. Yashkov, S.F. and Yashkova, A.S., Some Extension of the Heavy Traffic Limit Theorem for the M/G/1—EPS Queue, *Information Processes*, 2004, vol. 4, no. 3, pp. 269–274 (available at http://www.jip.ru/).

228. Yashkov, S.F. and Yashkova, A.S., Some Insight to the Time-Dependent Properties of the Queue Length Process in the M/G/1—EPS and LCFS–P Queues, *Information Processes*, 2005, vol. 5, no. 2, pp. 102–105 (available at http://www.jip.ru/).

229. Yashkov, S.F. and Yashkova, A.S., Asymptotics of the Sojourn Time in the M/G/1 System with Processor Sharing in the Case of Light Tails, *Obozrenie Prikl. Promyshl. Matem.*, 2006, vol. 13, no. 4 (in Russian).

230. Yashkov, S.F. and Yashkova, A.S., Egalitarian Processor Scharing, *Informatsionnye Processy*, 2006, vol. 6, no. 4, pp. 396–444 (in Russian) (available at http://www.jip.ru/).

231. Yashkova, A.S., A Note on Limit Theorem for Overloaded Processor Sharing Queue, *Information Processes*, 2003, vol. 3, no. 2, pp. 151–153 (available at http://www.jip.ru/).

232. Yashkova, A.S. and Yashkov, S.F., Distribution of the Virtual Sojourn Time in the M/G/1 Processor Sharing Queue, *Information Processes*, 2003, vol. 3, no. 2, pp. 128–137 (available at http://www.jip.ru/).

233. Zwart, A.P., Queueing Systems with Heavy Tails, *PhD Dissertation*, Eindhoven: Univ. of Technology, 2001.

234. Zwart, A.P. and Boxma, O.J., Sojourn Time Asymptotics in the M/G/1 Processor Sharing Queue, *Queueing Syst.*, 2000, vol. 35, pp. 141–166.

Статью представил к публикации член редколлегии В.И.Венец