ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 60 **2024** **Вып. 2**

УДК 621.391 : 519.872

 А.Ю. Карамышев, Е.Д. Порай, Е.М. Хоров

ОЦЕНКА ЕМКОСТИ СИСТЕМЫ СВЕРХНАДЕЖНОЙ СВЯЗИ С НИЗКИМИ ЗАДЕРЖКАМИ С ПОМОЩЬЮ АППРОКСИМАЦИЙ ДЛЯ МНОГОСЕРВЕРНЫХ СИСТЕМ МАССОВОГО ОБСЛУЖИВАНИЯ $G/G/s$ [1]

Для анализа производительности беспроводных локальных и сотовых сетей, обеспечивающих сверхнадежную связь с низкими задержками, требуются методы быстрой и точной оценки эффективной емкости системы, т.е. того объема трафика, для которого удается выполнить заданные требования к надежности и времени доставки. Эти методы могут использовать теорию массового обслуживания, например, моделируя систему связи как многосерверную систему массового обслуживания $G/G/s$. Однако существующие методы оценки показателей производительности системы $G/G/s$ обладают либо высокой вычислительной сложностью, либо высокой погрешностью в области малых значений задержки обслуживания, а также малых вероятностей испытать эту задержку. В статье исследуются приближенные методы оценки показателей производительности многосерверных систем $G/G/s$, потенциально применимых для оценки эффективной емкости системы сверхнадежной связи с низкими задержками. Предлагается метод оценки вероятности превысить ограничение на время пребывания в очереди, и численно показана его низкая ошибка. Также приведен асимптотический анализ метода, результаты которого могут быть полезны при реализации алгоритмов планирования радиоресурсов в беспроводных локальных и сотовых сетях.

Ключевые слова: аппроксимация, эффективная емкость, теория массового обслуживания, $G/G/s$, чувствительный к задержкам трафик, сверхнадежная связь с низкими задержками.

DOI: `10.31857/S0555292324020049`, **EDN:** GLDKHC

§ 1. Введение

В последние годы активно развиваются приложения, требующие гарантированно высоких скоростей передачи данных, низких задержек и высокой надежности при доставке данных. Например, при использовании облачных приложений виртуальной реальности необходимо быстро и надежно доставлять большие объемы данных, чтобы избежать ухудшения качества восприятия передаваемого видеоизображения [1]. Другим примером являются сценарии индустриальной автоматизации, где для эффективной координации мобильных роботов требуется доставлять команды управления с крайне низкой задержкой почти без потерь [2].

Для обслуживания такого трафика со строгими требованиями к качеству обслуживания (Quality of Service, QoS) в системах сотовой связи пятого поколения (5G) вводится сервис сверхнадежной связи с низкими задержками (URLLC). Аналогично, в дополнениях IEEE 802.11be и IEEE 802.11bn к стандарту Wi-Fi предлагаются

[1] Исследование выполнено в ИППИ РАН за счет гранта Российского научного фонда № 23-19-00756, https://rscf.ru/project/23-19-00756/

различные методы поддержки приложений реального времени и обеспечения сверхвысокой надежности доставки данных.

Особенностью указанных систем беспроводной связи является то, что за счет использования технологий множественного доступа с ортогональным частотным разделением (OFDMA) и многопользовательских многоантенных передач (MU-MIMO) одновременно в сети могут передаваться данные различных пользователей. Этот эффект еще сильнее увеличивается, когда базовые станции (точки доступа) координируют свои одновременные передачи для снижения взаимной интерференции, что является, например, одним из ключевых направлений развития сетей Wi-Fi 8.

Для эффективного планирования частотно-временных ресурсов, контроля доступа новых потоков в сеть (admission control), планирования сетевой инфраструктуры, разделения радиоресурсов между сервисами и виртуальными операторами требуется с высокой точностью определять объем ресурсов, который окажется достаточным для обслуживания трафика, требующего высокую надежность и низкие задержки (далее – трафика URLLC). Требования к качеству обслуживания такого трафика могут задаваться следующим образом: 99,999% пакетов должны быть доставлены не более чем за 1–10 мс [3, 4].

Для решения описанных задач могут использоваться математические модели, в том числе использующие методы теории массового обслуживания (ТМО) [5–10]. Отметим, что многим современным системам беспроводной связи, использующим частотно-временное разделение радиоресурсов, можно сопоставить многосерверную систему массового обслуживания, как показано в [5].

При этом доля заявок, обслуженная с не превышающей заданное ограничение задержкой, может быть определена в явном виде лишь для некоторых систем массового обслуживания, в то время как для остальных необходимы достаточно объемные вычисления [5]. Как результат, большую ценность имеет разработка аппроксимаций для оценки показателей производительности систем массового обслуживания. Известные из литературы аппроксимации имеют ряд серьезных недостатков при анализе сценариев обеспечения сверхнадежной связи с низкими задержками. Многие аппроксимации для систем массового обслуживания построены для оценки средней задержки и/или высоких значений вероятностей ($\gtrsim 0{,}1$) превысить ограничение на задержку в очереди, что делает их неприменимыми к анализу систем связи, в которых требуется обеспечить крайне низкие вероятности не выполнить ограничение на задержку.

Ряд авторов [8, 11, 12] строят аппроксимации для марковских систем, не соответствующих реальным системам связи. Наконец, некоторые методы, использующие аппарат ТМО, адаптированы к очень узким сценариям использования и обладают высокой вычислительной сложностью, например, [5], что делает их едва ли применимыми для расчетов в режиме реального времени.

В данной статье разрабатывается и исследуется приближенная формула для оценки эффективной емкости многосерверной системы массового обслуживания в области малых значений задержки обслуживания, а также малых вероятностей испытать эту задержку. В принятых обозначениях ТМО объектом исследования является многосерверная система $G/G/s$, где нотация G обозначает общий (general) вид распределения времени поступления и обслуживания заявок, а параметр $s \gg 1$ задает число обслуживающих устройств (серверов). Под эффективной емкостью здесь и далее понимается максимально достижимый объем трафика, который может быть обслужен в единицу времени при условии выполнения требований к качеству обслуживания.

Основной вклад статьи заключается, во-первых, в том, что в ней для системы $M/M/s$ исследуется точность приближенных методов оценки вероятности испытать задержку большую, чем некоторое малое ограничение по уровню низких вероятностей ($\ll 0{,}1$), характерных для сценариев обеспечения сверхнадежной связи с низкими задержками. Во-вторых, строится аппроксимация эффективной емкости системы $G/G/s$, а также оценивается и анализируется точность оценки эффективной емкости при помощи этих аппроксимаций.

Дальнейшее изложение построено следующим образом. В § 2 содержится обзор литературы по теме как с технической, так и с математической стороны. В § 3 формулируется объект исследования и ставятся решаемые задачи. Затем в § 4 описываются идеи и процесс построения приближенных формул, а также предлагается их асимптотический анализ. После чего в § 5 осуществляется оценка и анализ точности разработанных аппроксимаций. Наконец, в § 6 представлены основные выводы статьи.

§ 2. Обзор литературы

2.1. ТМО как инструмент оценки производительности современных систем связи. На протяжении более чем ста лет ТМО является одним из ключевых инструментов оценки производительности различных систем связи. Не являются исключением и современные системы сотовой связи 5G и беспроводные локальные сети Wi-Fi, несмотря на высокую сложность технологий заложенной в них логики работы и соответствующую громоздкость ее описания в виде формул.

Из-за применения частотно-временного разделения радиоресурсов обслуживание трафика в таких системах связи можно описать с помощью многосерверных систем массового обслуживания. При этом сервером является фундаментальная единица частотно-временных ресурсов – ресурсный блок, число которых может достигать сотен. Так, в работе [11] предлагается использовать системы массового обслуживания $M/M/s$ с конечной очередью для оценки пропускной способности системы 5G при обеспечении сверхнадежной связи с малой задержкой. Работа [7] предлагает оценивать среднюю задержку в системе 5G при помощи приближенной формулы для системы $G/G/s$. Для этого авторы адаптируют выражения для средней задержки в системе $M/M/s$, внося поправки, предложенные в [13]. Кроме того, работы [8, 12] предлагают при анализе систем сверхнадежной связи с низкими задержками пользоваться “правилом квадратного корня”, описанным подробно в п. 4.1 и основанным на асимптотическом поведении системы $M/M/s$.

2.2. Аппроксимации для анализа систем ТМО. Точное описание показателей производительности систем $G/G/s$ при помощи явных математических формул возможно только для некоторых частных случаев распределений, задающих поток поступления и процессы обслуживания заявок, например, для системы $M/M/s$. Как результат, достаточно распространенной является задача построения аппроксимаций для предсказывания характеристик произвольных систем массового обслуживания. Эта область активно развивалась на протяжении XX века и продолжает развиваться в настоящее время.

Фокусируясь на построении аппроксимаций для систем $G/G/s$, выделим фундаментальные работы [14–16], благодаря которым рассматриваемый раздел теории массового обслуживания сформировался в его современном виде. В частности, в работе [14] для многосерверной системы $M/M/s$ сформулирована и доказана теорема о существовании предела вероятности испытать задержку пребывания в очереди, отличную от нуля, что позже в своей упрощенной формулировке превратилось в “правило квадратного корня” и вошло во множество учебников по стохастическим процессам и компьютерным наукам [17, гл. 19; 18, § 5.3]. Несмотря на то, что результат работы [14] был выведен для системы $M/M/s$, которая имеет решение в явном виде, эта работа стала основой для построения аппроксимаций более сложных систем.

Нельзя обойти вниманием и классические работы [19–22], многие выкладки и идеи которых стали ключевыми для построения аппроксимаций для систем $G/G/s$.

Наконец, выделим ряд современных работ [23–26], предлагающих доработку и переосмысление фундаментальных результатов. Так, например, работы [23, 24] расширяют область применения выкладок [14] на более широкий класс систем, например, $M/D/s$.

2.3. Открытые вопросы. Сверхнадежная связь с низкими задержками накладывает жесткие требования к качеству обслуживания, а именно: задержка может превышать малое пороговое значение крайне редко, например, с вероятностью 10^{-5}. Анализ событий, которые происходят с такой низкой вероятностью, оказывается чувствительным к изменению параметров системы, в том числе к подмене распределений поступления и обслуживания заявок. Поэтому применение выкладок для системы $M/M/s$ к произвольной системе $G/G/s$, например, как это сделано в работах [8, 11, 12], приводит к высокой ошибке при оценке показателей производительности: задержки, вероятности превысить ограничение на задержку, пропускной способности сети и т.д.

Вопрос построения аппроксимаций непосредственно для систем $G/G/s$ уже рассматривался в работах [16, 27–29], как и вопрос их применения к реальным технологиям связи [7, 13]. Однако все эти работы предлагают аппроксимации для оценки среднего времени пребывания в системе или для оценки вероятности испытать ненулевую задержку пребывания в очереди, а не для оценки вероятности того, что задержка пребывания заявки в очереди превысит заданный порог. Более того, точность оценки предлагаемых формул для низких вероятностей ($\ll 0{,}1$) испытать задержку также остается неисследованной.

Как результат, предложенные в литературе аппроксимации не позволяют получить достаточно точную оценку показателей производительности в условиях, характерных для сверхнадежной связи с низкими задержками. В частности, открытым является вопрос построения вычислительно простых аппроксимаций для систем $G/G/s$, подходящих для анализа производительности указанных выше систем связи.

§ 3. Объект исследования и постановка задачи

Объектом исследования является система массового обслуживания $G/G/s$, в которой поток поступления заявок задается функцией распределения времени между поступлениями заявок в систему, обозначаемой через $\mathcal{A}$ (arrival), а процесс обслуживания заявок определяется функцией распределения времени обслуживания заявки на сервере, обозначаемой через $\mathcal{S}$ (service). Эти распределения являются распределениями общего вида. Условимся, что времена обслуживания на каждом сервере одинаково распределены и независимы. Для описания системы используются классические для теории массового обслуживания обозначения: интенсивность поступления заявок $\lambda = 1/\,\mathbf{E}\,\mathcal{A}$, интенсивность обслуживания каждой заявки $\mu = 1/\,\mathbf{E}\,\mathcal{S}$ и коэффициент использования $\rho = \lambda/(\mu s)$. Кроме того, для упрощения формул воспользуемся популярным приемом и положим $\mu = 1$, перенормировав все остальные показатели системы на размерную часть μ.

Хотя параметры рассматриваемой системы и требования к качеству обслуживания трафика мотивированы системами сверхнадежной связи с низкими задержками, оговоримся, что точное описание реальной системы связи, например, системы сотовой связи 5G, требует построения более детальной модели массового обслуживания нежели $G/G/s$: со слотированной политикой обслуживания, конечной очередью и возможностью группового поступления заявок. Учет этих факторов, безусловно, влияет на показатели производительности, однако выходит за рамки настоящей статьи и является предметом будущих исследований.

Рассмотрим время пребывания заявки в очереди (queueing time) $t_q = t_q(\lambda, s, \mathcal{A}, \mathcal{S})$, т.е. время пребывания заявки в системе за вычетом непосредственного времени обслуживания этой заявки. Нашей целью является построить и исследовать аппроксимации вероятности $\mathbf{P}(t_q > D)$ превысить ограничение D на время t_q. Вычисление данной вероятности является ключевым шагом на пути оценки введенной ранее эффективной емкости системы λ^*, которая математически определяется как

$$\lambda^* = \max\bigl\{\lambda \mid \mathbf{P}\left(t_q(\lambda, s, \mathcal{A}, \mathcal{S}) > D\right) \leqslant \varepsilon\bigr\}, \tag{1}$$

где D и ε – определяемые требованиями к качеству обслуживания ограничения на задержку и надежность, характерные значения которых – несколько средних времен обслуживания $1/\mu$ (несколько единиц, если $\mu = 1$) и 10^{-5} соответственно.

§ 4. Построение приближенных формул

4.1. Аппроксимации для системы $M/M/s$. Несмотря на то, что система $M/M/s$ является одной из немногих, для подсчета показателей производительности которой существуют точные математические формулы, логичным является начать рассмотрение аппроксимаций именно с этой системы. Такой порядок изложения во многом обусловлен исторической хронологией. Кроме того, экспоненциальные функции распределения времени, задающие процессы $\mathcal{A}$ и $\mathcal{S}$, являются одним из ключевых предельных предположений при построении более сложных аппроксимаций.

Предполагая процесс математического описания системы $M/M/s$ классическим, подробности которого можно найти, например, в [18, гл. 5], сошлемся лишь на один примечательный результат:

$$\mathbf{P}(t_q > D) = \mathbf{P}(t_q > 0)e^{-(s-\lambda)D}, \tag{2}$$

т.е. вероятность испытать задержку в очереди, большей D, при условии того, что задержка уже случилась, убывает экспоненциально с увеличением D.

Пусть $\Phi(\cdot)$ и $\varphi(\cdot)$ – функция и плотность стандартного нормального распределения соответственно. Тогда ключевой результат работы [14] формулируется в виде следующей теоремы.

Теорема 1 [14]. *Пусть $\lambda \to \infty$ и $s \to \infty$, однако $\rho = \lambda/s < 1$, т.е. система стабильна. Для системы $M/M/s$ невырожденный предел (его значение заключено строго между 0 и 1) вероятности*

$$\lim_{\lambda\to\infty} \mathbf{P}\left(t_q(\lambda, s) > 0\right) = \left(1 + \frac{\beta\Phi(\beta)}{\varphi(\beta)}\right)^{-1} \triangleq \mathcal{HW}(\beta) \in (0, 1) \tag{3}$$

существует тогда и только тогда, когда

$$\lim_{\lambda\to\infty} \sqrt{s}(1-\rho) \triangleq \beta > 0. \tag{4}$$

Данную теорему, а точнее условие (4), можно интерпретировать как руководство для оценки $\mathbf{P}(t_q > 0) \to \mathcal{HW}(\beta)$, где $\beta = \beta(\lambda, s) > 0$ – неявная функция, задающая соотношение между λ и s. При этом условие (4) допускает свободу в построении и выборе функции $\beta(\lambda, s)$, хотя и требует выполнения предельного соотношения.

Замечание 1. Если выбирать λ и s согласно формуле $s = \lambda + \beta_1\sqrt{s}$, $\beta_1 > 0$, то

$$\beta_1 = \frac{s-\lambda}{\sqrt{s}} = \sqrt{s}\,(1 - \lambda/s) = \sqrt{s}\,(1-\rho)\,,$$

т.е. условие (4) выполняется по построению. Следовательно, в пределе большого числа серверов и большой нагрузки $\mathbf{P}\left(t_q > 0\right) \to \mathcal{HW}(\beta_1)$.

Замечание 2. Поскольку для определения эффективной емкости (1) интерес представляет $\rho \to 1$, то условие (4) можно заменить на эквивалентное, домножив выражение $\sqrt{s}(1-\rho)$ на любую функцию от ρ, такую что ее предел равен 1 при $\lambda \to \infty$, например, $1/\sqrt{\rho}$. В частности, выбор параметров λ и s согласно $s = \lambda + \beta_2\sqrt{\lambda}$ дает

$$\beta_2 = \frac{s-\lambda}{\sqrt{\lambda}} = \frac{\sqrt{s}(1-\lambda/s)}{\sqrt{\lambda/s}} = \frac{\sqrt{s}(1-\rho)}{\sqrt{\rho}}.$$

Таким образом, обеспечивается предел $\mathbf{P}\left(t_q > 0\right) \to \mathcal{HW}(\beta_2)$.

Соотношение $s = \lambda + \beta_2\sqrt{\lambda}$ называется “правилом квадратного корня” для выбора числа серверов [17, гл. 19]. В некоторых работах (см., например, [12]) вместо формулы (3) используют более простую:

$$\mathbf{P}\left(t_q > 0\right) \to \mathcal{HW}(\beta_2) \approx 1 - \Phi(\beta_2). \tag{5}$$

Отметим, что эта упрощенная аппроксимация хороша лишь для малой нагрузки. Однако для упрощенной формулы при $\rho \to 1$ оказывается, что

$$\mathbf{P}(t_q > 0) \to 1 - \Phi(0) = 1/2,$$

хотя система перестает быть стабильной и должно выполняться $\mathbf{P}(t_q > 0) \to 1$, что явно демонстрирует ошибку оценки вероятности испытать задержку.

Авторы [24] предлагают альтернативную формулу для $\beta(\lambda, s)$, обеспечивающую совпадение более высоких порядков разложения для $s \to \infty$, чем β_1 и β_2:

$$\beta_{JLZ} \triangleq \sqrt{-2s\left(1 - \frac{\lambda}{s} + \log\left(\frac{\lambda}{s}\right)\right)} = \sqrt{-2s\left(1 - \rho + \log\rho\right)}.$$

Заметим, что для β_1, β_2, β_{JLZ} справедливы следующие соотношения:

$$\beta_{JLZ}^2 - \beta_1^2 = 2s\sum_{k=3}^{+\infty}\frac{(1-\rho)^k}{k} \geqslant 0,$$

$$\beta_2^2 - \beta_{JLZ}^2 = s\sum_{k=3}^{+\infty}\left(1 - \frac{2}{k}\right)(1-\rho)^k \geqslant 0,$$

откуда следует, что

$$\beta_1 \leqslant \beta_{JLZ} \leqslant \beta_2, \quad \forall\rho \in [0, 1),$$

т.е. β_{JLZ} в некотором смысле является промежуточным вариантом аппроксимации между β_1 и β_2.

С учетом экспоненциальной зависимости от допустимой задержки D согласно (2) оценку искомой вероятности превысить ограничение на время пребывания в очереди можно подсчитать как

$$\mathbf{P}(t_q(\lambda, s) > D) \approx \mathcal{HW}(\beta(\lambda, s))e^{-(s-\lambda)D}, \quad \beta(\lambda, s) \in \{\beta_1, \beta_2, \beta_{JLZ}\}. \tag{6}$$

В заключение этого пункта отметим, что хотя эти формулы так или иначе встречались в литературе, исследование их сходимости в области малых значений задержки обслуживания, а также малых вероятностей испытать эту задержку, не проводилось. Это делается в п. 5.1 данной статьи.

4.2. Построение аппроксимация для системы $G/G/s$. Следуя наблюдению, отраженному в (2), вероятность $\mathbf{P}(t_q > D)$ можно описать в виде произведения двух множителей: первый из которых – оценка вероятности $\mathbf{P}(t_q > 0)$ испытать ненулевую задержку в очереди в принципе, а второй – множитель, отвечающий за скорость уменьшения размера непустой очереди. Так, разделяя построение аппроксимации для системы $G/G/s$ на две части, проанализируем каждую их них в следующих подразделах, после чего приведем итоговую формулу.

Оценка вероятности испытать ненулевую задержку в очереди. Рассмотрим систему $G/G/\infty$ (с бесконечным числом серверов) в предположении, что все серверы функционируют независимо и процесс обслуживания заявок не зависит от потока поступления пакетов.

В работах [15, 22, 30] показано, что если процесс поступления заявок в систему может быть приближен броуновским движением, то в пределе высокой нагрузки стационарное число заявок в системе Q может быть описано нормальным распределением $\mathcal{N}$ со средним λ и дисперсией λz:

$$Q \approx \mathcal{N}(\lambda, \lambda z), \tag{7}$$

где λ – среднее число активных серверов ($\mu = 1$),

$$z \triangleq 1 + (c_{\mathcal{A}}^2 - 1)\eta_{\mathcal{S}},$$

$$\eta_{\mathcal{S}} \triangleq (\mathbf{E}\,\mathcal{S})^{-1} \int\limits_0^\infty (1 - F_{\mathcal{S}}(t))^2\, dt,$$

$c_{\mathcal{A}}^2 \triangleq \mathbf{Var}\,\mathcal{A}/(\mathbf{E}\,\mathcal{A})^2$ – квадрат коэффициента вариации распределения $\mathcal{A}$,

$F_{\mathcal{S}}(t)$ – функция распределения $\mathcal{S}$.

Данное утверждение означает, что асимптотическая дисперсия числа заявок в системе равна λz, т.е. отклонение числа заявок в системе Q от среднего значения λ составляет величину порядка $\mathcal{O}(\sqrt{\lambda z})$ при больших значениях λ. Множитель изменения дисперсии z называется асимптотической "остроконечностью" (peakedness) распределения [15].

Применим предельное соотношение (7) для системы $G/G/s$ ($s \gg 1$) в предположении, что характер зависимости (7) сохранится для большого числа серверов. Обратим внимание, что ненулевая задержка в очереди означает, что число заявок в системе оказалось больше числа доступных серверов, т.е. $\mathbf{P}\,(t_q > 0) = \mathbf{P}\,(Q > s)$. Тогда с учетом того, что число заявок в системе $G/G/s$ можно приблизить нормальным законом с дисперсией, измененной на величину асимптотической "остроконечности" z, имеем:

$$\mathbf{P}\,(t_q > 0) = \mathbf{P}\,(Q > s) \approx \mathbf{P}\,(\mathcal{N}(\lambda, \lambda z) > s) = 1 - \Phi((s - \lambda)/\sqrt{\lambda z}) = 1 - \Phi(\beta_2/\sqrt{z}).$$

Сравнивая полученный результат с правой частью соотношения (5) для системы $M/M/s$ и осуществляя обратный переход к исходной формуле в (5), получаем для системы $G/G/s$, что

$$\mathbf{P}(t_q(\lambda, s) > 0) \approx \mathcal{HW}(\beta(\lambda, s)/\sqrt{z}),$$

где $\beta(\lambda, s)$ – функция, удовлетворяющая условию (4), однако теперь применительно к системе $G/G/s$.

Следуя [15], обратим внимание, что ключевой величиной для вычисления введенной поправки z в (7) является величина $\eta_{\mathcal{S}}$, которая принимает значения от 0 до 1. При этом 0 соответствует бесконечной дисперсии процесса $\mathcal{S}$, а 1 – случаю, когда весь вес функции распределения времени обслуживания сосредоточен в одной точке, т.е. в случае вырожденного распределения. Если дисперсия времени обслуживания заявки возрастает, а среднее сохраняется, то $\eta_{\mathcal{S}}$ уменьшается.

Заметим, что вклад распределения времени обслуживания $\mathcal{S}$ в (7) зависит от знака $(c_{\mathcal{A}}^2 - 1)$. Более того, система формально невосприимчива к виду распределения времени обслуживания $\mathcal{S}$, если число поступающих в систему пакетов в единицу времени подчиняется пуассоновскому распределению ($c_{\mathcal{A}}^2 = 1$), т.е. в случае системы $M/G/s$.

Оценка условной вероятности превысить допустимую задержку. Для описания нормированного числа заявок в системе $\widehat{Q} = (Q-s)/\sqrt{s}$ известен предел высокой нагрузки для системы $G/G/s$ в условиях, аналогичных указанным в теореме 1 [14, 15]. В частности, в этих работах утверждается, что при условии $\widehat{Q} > 0$ (т.е. при $Q > s$) дальнейшее изменение $\widehat{Q}$ приближается броуновским движением с дискретным временем со средним (параметром сноса) $-\mu\beta$ и дисперсией (параметром диффузии) $\mu(c_{\mathcal{A}}^2 + c_{\mathcal{S}}^2)$. Для такого процесса справедливо, что

$$\mathbf{P}\big(\widehat{Q} > x \mid \widehat{Q} > 0\big) \approx e^{-\mu\beta x/\left(\mu(c_{\mathcal{A}}^2 + c_{\mathcal{S}}^2)\right)} = e^{-x\beta/v}, \tag{8}$$

где $v \triangleq (c_{\mathcal{A}}^2 + c_{\mathcal{S}}^2)/2$ – коэффициент масштабирования броуновского движения.

В этом случае превышение числа заявок в системе над числом серверов $Q - s = x\sqrt{s}$, причем в среднем потребуется время D ($\mu = 1$), чтобы принять в обработку все накопившиеся на текущий момент заявки, т.е. $x(D) \approx D\sqrt{s}$.

Тогда из (8) получаем:

$$\mathbf{P}\left(t_q > D \mid t_q > 0\right) \triangleq \mathbf{P}\big(\widehat{Q} > x(D) \mid \widehat{Q} > 0\big) \approx e^{-x(D)\beta/v} \approx e^{-D\sqrt{s}\beta/v}. \tag{9}$$

Оговоримся, что область применимости предложенной аппроксимации ограничена [15, 16]. В частности, ожидается, что в случае чрезмерно больших значений дисперсий и/или большого дисбаланса между $c_{\mathcal{A}}^2$ и $c_{\mathcal{S}}^2$, что усложняет переход к стационарному распределению, потребуются дополнительные улучшения аппроксимации, например, введение функции балансировки влияния между процессами $\mathcal{A}$ и $\mathcal{S}$. Поправка v мотивирована лишь первым порядком разложения характеристик системы по степеням $(1 - \rho)$, в то время как учет коэффициентов более высокого порядка потенциально мог бы существенно улучшить точность аппроксимации, однако вычисление этих коэффициентов для произвольной системы $G/G/s$ является достаточно сложной и все еще открытой задачей.

Замечание 3. Для системы $M/M/s$ имеем $c_{\mathcal{A}} = c_{\mathcal{S}} = 1 \Rightarrow v = 1$. В результате для $\beta = \beta_1(\lambda, s)$ формула (9) упрощается до уже известного соотношения (2).

Замечание 4. Поправка приближения для систем $G/G/s$ (9) согласуется с известными аппроксимациями для среднего времени ожидания в очереди для системы $G/G/1$, например, в [19].

Замечание 5. Соотношение (9) имеет ряд распространенных аналогов, например, экспоненциальную функцию затухания (exponential decay rate) в теории стохастических сетевых исчислений (stochastic network calculus) [31].

Итоговая аппроксимация для системы $G/G/s$. Объединяя проделанные выше результаты, получаем следующий метод оценки вероятности того, что задержка в очереди превысит пороговую $\mathbf{P}(t_q(\lambda, s, \mathcal{A}, \mathcal{S}) > D)$:

$$\mathbf{P}(t_q(\lambda, s, \mathcal{A}, \mathcal{S}) > D) \approx \mathcal{HW}\left(\beta(\lambda, s)/\sqrt{z}\right) e^{-D\sqrt{s}\beta(\lambda,s)/v}, \tag{10}$$

$$\mathcal{HW}(\xi) = \left(1 + \frac{\xi\Phi(\xi)}{\varphi(\xi)}\right)^{-1},$$

$$\beta(\lambda,s)\in\{\beta_1,\beta_2,\beta_{JLZ}\}=\left\{\frac{s-\lambda}{\sqrt{s}},\frac{s-\lambda}{\sqrt{\lambda}},\sqrt{-2s\left(1-\frac{\lambda}{s}+\log\left(\frac{\lambda}{s}\right)\right)}\right\},$$
$$z=1+(c_{\mathcal{A}}^2-1)\eta_{\mathcal{S}},$$
$$\eta_{\mathcal{S}}=(\mathbf{E}\,\mathcal{S})^{-1}\int\limits_0^\infty(1-F_{\mathcal{S}}(t))^2\,dt,$$
$$v=\frac{1}{2}\left(c_{\mathcal{A}}^2+c_{\mathcal{S}}^2\right),$$
$$c_{\mathcal{A}}^2=\frac{\mathbf{Var}\,\mathcal{A}}{(\mathbf{E}\,\mathcal{A})^2},\quad c_{\mathcal{S}}^2=\frac{\mathbf{Var}\,\mathcal{S}}{(\mathbf{E}\,\mathcal{S})^2}.$$

Чтобы очертить область применимости полученной аппроксимации, явно выделим потенциальные проблемы данной формулы, которые вытекают из нарушений предположений, так или иначе использовавшихся при построении:

- нарушение предположения $s\to\infty$ $(s\gg 1)$;
- нарушение предположения $\lambda\to\infty$ $(\rho\to 1)$;
- сильное различие между процессами $\mathcal{A}$ и $\mathcal{S}$ в терминах $c_{\mathcal{A}}^2$ и $c_{\mathcal{S}}^2$, т.е. $|c_{\mathcal{A}}^2-c_{\mathcal{S}}^2|\gg 0$;
- чрезмерно большие значения коэффициентов вариации $c_{\mathcal{A}}^2$ и/или $c_{\mathcal{S}}^2$ [16];
- невозможность выполнения условий вывода z: малое число серверов и зависимые функции распределения времени обслуживания заявки на этих серверах.

Валидация приведенной аппроксимации в сценариях, характерных для сверхнадежной связи с малой задержкой, приведена в п. 5.2.

Заметим, что предложенная аппроксимация допускает свободу в выборе $\beta(\lambda,s)$; в § 5 даются рекомендации по выбору $\beta(\lambda,s)$ из числа сформулированных с целью повышения точности аппроксимации оценки эффективной емкости системы.

4.3. Асимптотический анализ построенной аппроксимации. Исследуем полученную приближенную формулу (10) применительно к вычислению эффективной емкости системы (1). Рассмотрим "резерв емкости" $\Delta\triangleq s-\lambda^*$, т.е. разницу между числом серверов и эффективной емкостью системы, при которой обеспечивается выполнение требований к качеству обслуживания, заданные параметрами D и ε.

Исследуем формулу (10) в следующих предельных предположениях:

- число серверов велико, т.е. $s\gg 1$;
- имеет место предел высокой нагрузки, т.е. $\Delta/s\to 0$, что эквивалентно $\rho\to 1$;
- величина допустимой задержки в очереди D мала, но $D>0$.

Для предельного случая (4), т.е. $\mathbf{P}\left(t_q(\lambda^*,s)>D\right)=\varepsilon$, из (10) получим:

$$\log\varepsilon=\log\mathcal{HW}\left(\beta/\sqrt{z}\right)-\frac{D}{v}\sqrt{s}\beta. \tag{11}$$

Разложим каждую из функций составной формулы (11), пользуясь предположениями о большом числе серверов и высокой нагрузке, а также применяя разложения для функции $\Phi(\xi)$ и плотности $\varphi(\xi)$ стандартного нормального распределения:

$$\log\mathcal{HW}(\xi)=-\log\left(1+\frac{\xi\Phi(\xi)}{\varphi(\xi)}\right)=$$
$$=-\log\left(1+\frac{\xi\left(\frac{1}{2}+\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}\left(\xi+\sum\limits_{k=1}^{+\infty}\frac{\xi^{2k+1}}{(2k+1)!!}\right)\right)}{\frac{1}{\sqrt{2\pi}}e^{-\xi^2/2}}\right)=$$

$$
= -\log\left(1+\sqrt{\frac{\pi}{2}}\xi e^{+\xi^2/2}+\xi^2+\mathcal{O}\left(\xi^4\right)\right) =
$$
$$
= -\sqrt{\frac{\pi}{2}}\xi-\left(1-\frac{\pi}{4}\right)\xi^2+\mathcal{O}\left(\xi^3\right), \quad \xi\to 0;
$$
$$
\xi=\beta/\sqrt{z}=\frac{\Delta}{\sqrt{sz}}\left(1+\mathcal{O}\left(\Delta/s\right)\right), \quad \Delta/\sqrt{s}\to 0, \quad s\to\infty,
$$

причем разложение для ξ справедливо для любой из функций $\beta\in\{\beta_1,\beta_2,\beta_{JLZ}\}$.

Возвращаясь к исходному соотношению (11) и удерживая члены разложения вплоть до $\mathcal{O}\left(\Delta/\sqrt{s}\right)$, имеем асимптотическую формулу:

$$
-\log\varepsilon=\sqrt{\frac{\pi}{2sz}}\Delta+\frac{D}{v}\Delta+\mathcal{O}\left(\Delta/s\right).
$$

Из этой формулы можно выразить

$$
\Delta\approx\frac{-\log\varepsilon}{\sqrt{\dfrac{\pi}{2sz}}+\dfrac{D}{v}}\xrightarrow{s\to\infty}-\log\varepsilon\cdot\frac{v}{D}.
$$

Заметим, что по построению $\Delta=s-\lambda^*\leqslant s$ (так как $\lambda^*\geqslant 0$), что может нарушаться в асимптотической формуле в случае экстремально малых D и/или малой нагрузки ρ. В качестве грубой поправки на это ограничение, перепишем искомое соотношение для Δ в виде

$$
\Delta\approx\min\left\{s,-\log\varepsilon\left(\frac{D}{v}+\sqrt{\frac{\pi}{2sz}}\right)^{-1}\right\}\xrightarrow{s\to\infty}-\log\varepsilon\cdot\frac{v}{D}, \quad D>0. \tag{12}
$$

Альтернативным способом улучшения формулы в области малых s является разложение до более высоких порядков $\Delta/\sqrt{s}$, что может быть выполнено аналогично представленной процедуре, но уже потребует явного выбора вида функции β.

Полученное асимптотическое соотношение $\Delta=-\log\varepsilon\cdot v/D$ не зависит ни от s, ни от λ^*, что фактически означает, что любые зависимости эффективной емкости λ^* от s для $D>0$ имеют линейную асимптоту при $s\gg 1$, параллельную прямой $\lambda^*=s$ и смещенную относительно этой прямой на величину Δ, зависящую только от требований к качеству обслуживания (параметров ε и D), а также от коэффициентов $c_{\mathcal{A}}^2$ и $c_{\mathcal{S}}^2$, которые определяются лишь первыми двумя моментами соответствующих распределений. В свою очередь это открывает возможность оценки требуемого числа обслуживающих устройств в зависимости от запрашиваемых значений $\{\lambda^*,D,\varepsilon\}$ при помощи линейной функции, простота вычисления которой является ключом к построению алгоритмов управления передачей данных, упомянутых в §1 и работающих в режиме реального времени.

Отдельно отметим, что существование линейной асимптоты эффективной емкости наблюдается и для других систем [5], более сложных, чем $G/G/s$, факт чего обсуждается при анализе численных результатов в §5.

§ 5. Численные результаты

Построенные в §4 аппроксимации валидируются с использованием дискретно-событийной платформы имитационного моделирования ns-3 [32]. Для этого в этой платформе реализован сетевой сценарий точка-точка со множеством каналов, полностью соответствующий математическому описанию системы $G/G/s$.

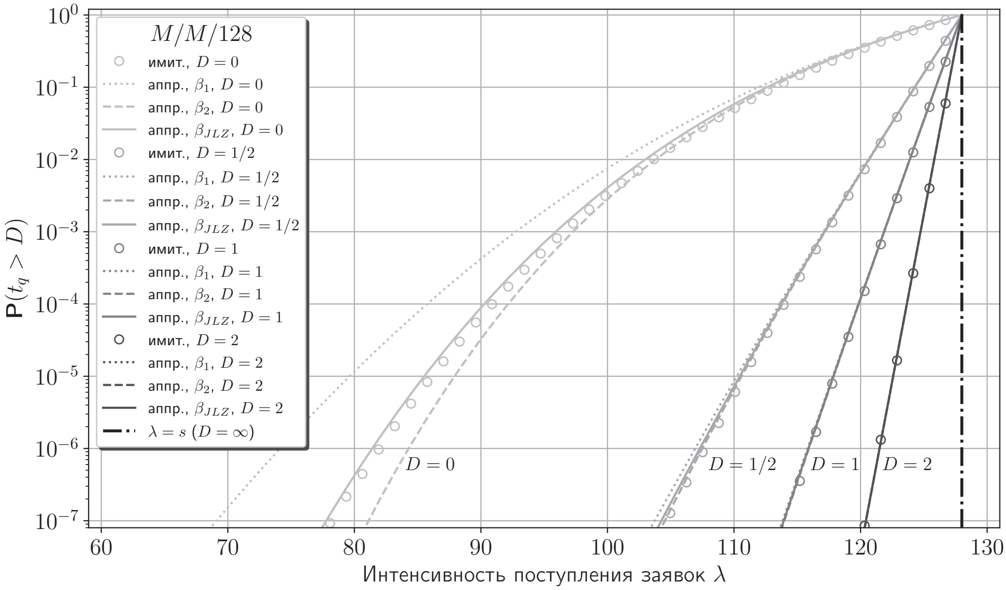

Рис. 1. Валидация приближенных формул (6) для системы $M/M/128$

5.1. Аппроксимации для системы $M/M/s$. На рис. 1 представлено сравнение рассмотренных в статье аппроксимаций (6) для системы $M/M/s$ с $s = 128$ серверами для различных ограничений на задержку в очереди D.

Данные, полученные при помощи имитационной модели, обозначаются в легенде как “имит.”, в то время как аппроксимации, отвечающие различным функциям $\beta(\lambda, s)$, имеют обозначение “аппр., β”.

Результаты показывают сходимость всех трех вариантов выбора функции $\beta(\lambda, s)$ в области $\rho \to 1$, что согласуется с предельными соотношениями, в которых эти аппроксимации были построены. При этом в области умеренной нагрузки вариант β_{JLZ} показывает наилучшие результаты, с высокой точностью совпадая с имитационными результатами, что происходит благодаря более точному асимптотическому представлению пуассоновского распределения [24]. С ростом допустимой задержки D разница между аппроксимациями фактически исчезает за счет того, что система оказывается в области $\rho \to 1$.

В итоге, если система $M/M/s$ требует построения аппроксимации, рекомендуется использовать формулу (6) с вариантом

$$\beta_{JLZ} = \sqrt{-2s\left(1 - \frac{\lambda}{s} + \log\left(\frac{\lambda}{s}\right)\right)},$$

обеспечивающим малую ошибку приближения.

5.2. Аппроксимации для системы $G/G/s$. Чтобы исследовать точность аппроксимаций (10), рассмотрим системы $G/G/s$ с различными функциями распределения, задающими процессы $\mathcal{A}$ и $\mathcal{S}$ и представленными в табл. 1. Распределения имеют одинаковые средние значения $1/\lambda$ или $1/\mu = 1$ соответственно.

Поскольку система $D/D/s$ не является стохастической, она исключена из рассмотрения.

Таблица 1

Параметры тестовых распределений

Обозначение	Распределение	c^2	η
M	Экспоненциальное	1	1/2
U	Равномерное, $\max = 5 \cdot \min$	4/27	7/9
W	Вейбулл, модуль=1/2	5	1/4
D	Вырожденное	0	1

На рис. 2, 3 представлены примеры сравнения аппроксимаций для систем $U/M/s$ и $W/D/s$ при $s = 128$. Обратим внимание на то, что теперь β_2 показывает результаты, наиболее близкие к результатам имитационной модели, в особенности для $D > 0$. Это происходит из-за того, что неточности при аппроксимациях $\mathbf{P}(t_q > 0)$ и $\mathbf{P}(t_q > D \mid t_q > 0)$ для β_2 компенсируют друг друга, в то время как для других функций β ошибка апокcимации обоих частей оценки вносит вклад одного знака, что увеличивает общую ошибку оценки искомой вероятности. Как результат, рекомендуется использовать аппроксимацию (10) в совокупности с $\beta_2 = (s - \lambda)/\sqrt{\lambda}$ для систем $G/G/s$, обособляя случай системы $M/M/s$, для которого предпочтительным остается вариант формулы (6) с β_{JLZ}.

Поскольку одной из целей построения аппроксимаций (10) является оценка эффективной емкости системы, рассмотрим ошибку оценки λ^* для

$$\mathbf{P}(t_q > D) \leqslant \varepsilon = 10^{-5}.$$

Значения эффективной емкости λ^*, полученные при помощи имитационного моделирования и предложенной аппроксимации для различных значений D, представлены в табл. 2. Кроме того, в табл. 2 подсчитана относительная ошибка такого предсказания.

Отметим, что для рассмотренных распределений и параметров системы погрешность оценки не превышает 5,5%. Ошибка оценки эффективной емкости для $D \geqslant 1$

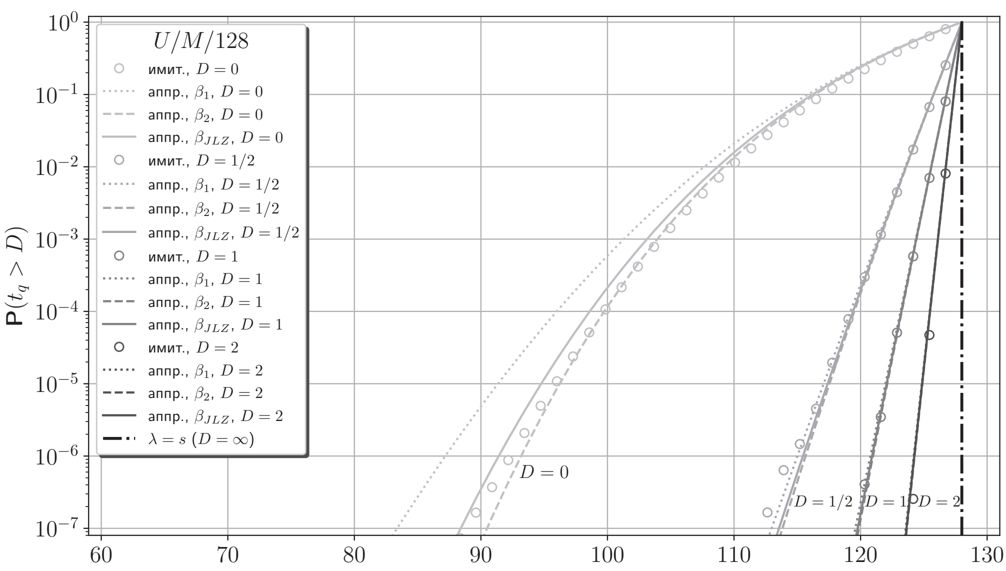

Рис. 2. Валидация приближенных формул (10) для системы $U/M/128$

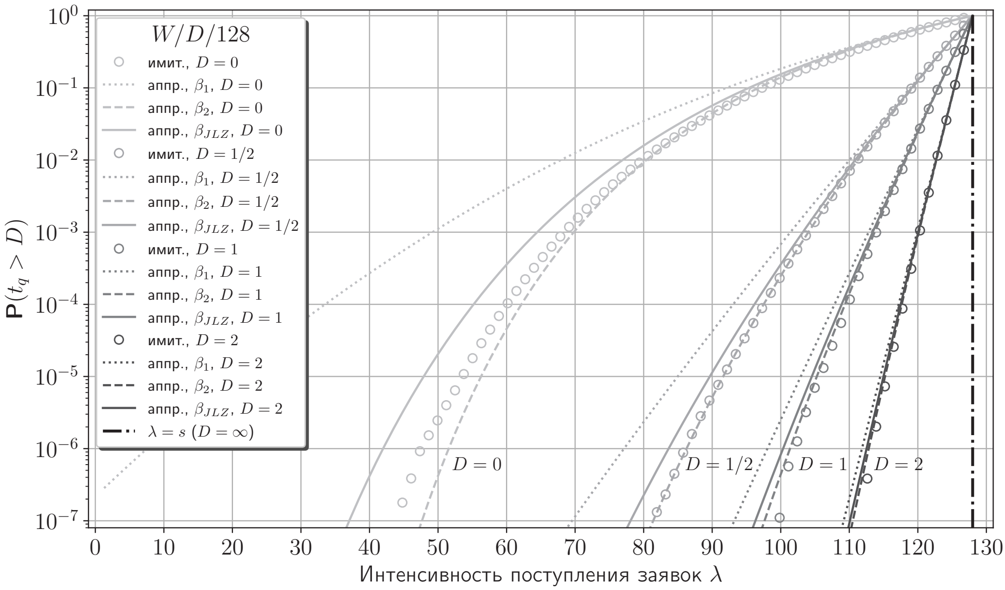

Рис. 3. Валидация приближенных формул (10) для системы $W/D/128$

вовсе не превышает 2%, что является остаточным эффектом конечной гранулярности сетки имитационного моделирования.

5.3. Асимптотическое поведение зависимостей эффективной емкости системы. Наконец, убедившись в том, что построенные аппроксимации (10) могут быть использованы для оценки эффективной емкости λ^* с приемлемой ошибкой, обратимся к проверке асимптотических выкладок, сделанных в п. 4.3.

Рассмотрим систему $W/D/s$ и исследуем поведение кривых $\lambda^* = f(s \mid D)$ для $\varepsilon = 10^{-5}$. Обратим внимание, что аппроксимации для системы $W/D/s$ обладают одной из наибольших ошибок оценки эффективной емкости согласно табл. 2. Это происходит из-за крайне большой дисперсии входного потока: рассматриваемое распределение Вейбулла с модулем 1/2 является классическим распределением “с тяжелыми хвостами” и высокой плотностью вероятности в нуле; а также из-за сильного дисбаланса между процессами поступления и обслуживания заявок, в частности, $|c^2_{\mathcal{A}} - c^2_{\mathcal{S}}| = 5$, что является максимальным значением для рассматриваемых комбинаций распределений из табл. 1. Все это дает основание считать выводы, валидные для системы $W/D/s$, скорее пессимистичными относительно других систем $G/G/s$. Иными словами, закономерности, прошедшие проверку для системы $W/D/s$, можно считать справедливыми для крайне широкого класса систем $G/G/s$.

Кроме того, система $W/D/s$ идейно близка к опорному реальному объекту – системе сверхнадежной связи с малой задержкой, работающей по технологии 5G или Wi-Fi 7 с расписанием. Во-первых, для такого сценария характерно входное распределение времени между поступлениями заявок с высокой плотностью вероятности в нуле и “с тяжелыми хвостами” из-за преобладания низкоскоростных и надежных сигнально-кодовых конструкций, доставка пакетов на которых требует большого числа ресурсных блоков. Данные свойства также присущи распределению Вейбулла с модулем 1/2. Во-вторых, в этом сценарии имеет слотированный характер работы системы, из-за чего возникает вырожденное распределение времени обслуживания заявки ресурсными блоками.

Таблица 2

Валидация аппроксимации для оценки эффективной емкости системы $G/G/s$
$\lambda^* = \max\{\lambda \mid \mathbf{P}\,(t_q(\lambda) > D) \leqslant \varepsilon = 10^{-5}\}$

Система	D	имит.	аппр., β_2	ошибка	Система	D	имит.	аппр., β_2	ошибка
$M/U/128$	0	86,3	87,9	1,9%	$W/M/128$	0	63,8	67,2	5,3%
	1/2	116,7	117,0	0,3%		1/2	90,8	91,0	0,2%
	1	122,1	121,9	−0,2%		1	103,0	103,1	0,1%
	2	124,9	124,8	−0,1%		2	113,1	113,3	0,2%
$M/W/128$	0	86,0	87,9	2,2%	$W/U/128$	0	57,4	60,5	5,4%
	1/2	101,0	99,3	−1,7%		1/2	92,2	91,9	−0,3%
	1	108,2	106,7	−1,4%		1	105,3	105,0	−0,3%
	2	115,0	114,4	−0,5%		2	115,6	114,9	−0,6%
$M/D/128$	0	86,2	87,9	2,0%	$W/W/128$	0	72,0	75,4	4,7%
	1/2	118,1	118,2	0,1%		1/2	89,8	87,7	−2,3%
	1	122,7	122,6	−0,1%		1	98,3	96,6	−1,7%
	2	125,2	125,2	0,0%		2	107,7	107,0	−0,6%
$U/M/128$	0	95,8	96,2	0,4%	$W/D/128$	0	53,5	56,2	5,0%
	1/2	117,1	117,5	0,3%		1/2	91,6	91,6	0,0%
	1	122,1	122,1	0,0%		1	105,6	105,2	−0,4%
	2	125,0	124,8	−0,2%		2	115,5	115,2	−0,3%
$U/U/128$	0	103,1	102,8	−0,3%	$D/M/128$	0	98,0	98,0	0,0%
	1/2	124,8	124,8	0,0%		1/2	118,6	118,7	0,1%
	1	126,4	126,3	−0,1%		1	122,9	122,8	−0,1%
	2	127,2	127,1	−0,1%		2	125,3	125,2	−0,1%
$U/W/128$	0	90,6	91,6	1,1%	$D/U/128$	0	108,3	107,1	−1,1%
	1/2	104,1	102,5	−1,5%		1/2	126,3	126,3	0,0%
	1	110,3	109,3	−0,9%		1	127,2	127,1	−0,1%
	2	116,2	116,2	0,0%		2	127,6	127,5	−0,1%
$U/D/128$	0	110,6	110,6	0,0%	$D/W/128$	0	91,5	92,4	1,0%
	1/2	126,4	126,3	−0,1%		1/2	104,9	103,1	−1,7%
	1	127,2	127,1	−0,1%		1	110,9	109,7	−1,1%
	2	127,5	127,5	0,0%		2	116,5	116,5	0,0%

На рис. 4 представлены зависимости эффективной емкости системы λ^* ($\varepsilon = 10^{-5}$) от числа обслуживавших устройств s для различных значений D. В частности, сравниваются кривые, полученные при помощи имитационного моделирования, с аппроксимациями, посчитанными по формуле (10) с β_2. Кроме того, для демонстрации асимптотического характера построенных зависимостей рис. 4 содержит асимптоты (12) для $D > 0$ и $\rho = \lambda^*/s > 1/2$, обозначенные "асимпт." в легенде.

Рис. 4 показывает, что предложенные аппроксимации обладают умеренной ошибкой для различных значений числа серверов s, в особенности для $D \geqslant 1$. Также результаты демонстрируют, что зависимости $\lambda^* = f(s \mid D)$ для $D > 0$ действительно имеют линейную асимптоту, параллельную прямой пропорциональности, как и предсказывалось в п. 4.3. Смещение данных асимптот неплохо предсказывается формулой (12) в случае $D > 1$ и большого числа серверов s. При этом, как правило, ограничение на допустимое время задержки в очереди составляет от единиц до нескольких десятков времен обслуживания [33], т.е. практические запросы быстрой оценки эффективной емкости системы оперируют с $D > 1$.

Наконец, отметим, что выведенный линейный эффект справедлив для гораздо более широкого класса систем, нежели $G/G/s$. Так, например, в работе [5] сделаны аналогичные наблюдения для детализированной модели 5G системы. Это дает надежду на то, что разработанные аппроксимации и их асимптотики окажутся крайне полезны для реализации различных легковесных алгоритмов планирования в широком классе реально существующих систем связи.

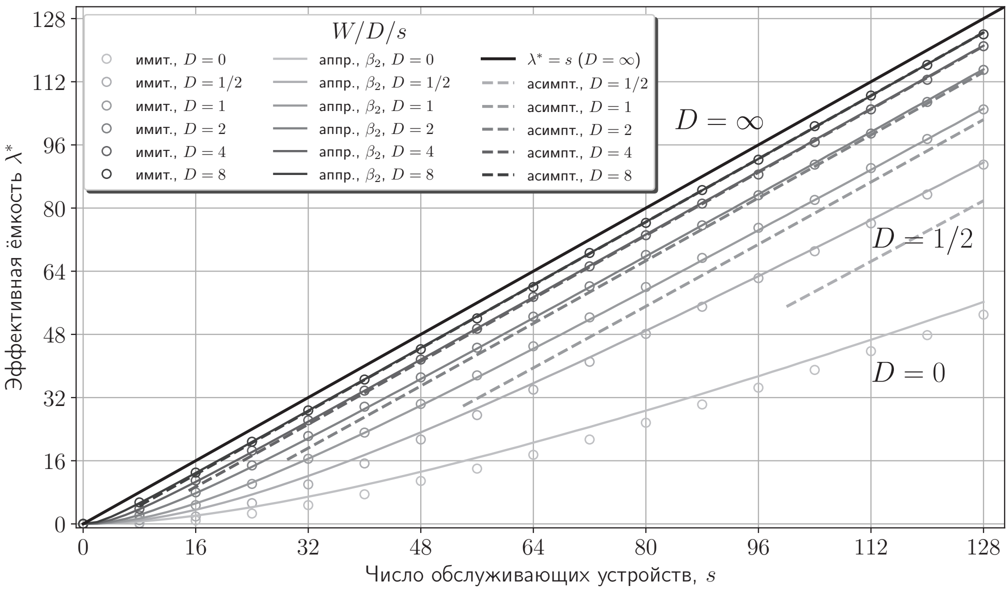

Рис. 4. Асимптотический характер зависимости эффективной емкости λ^* от числа обслуживающих устройств s на примере системы $W/D/s$, $\varepsilon = 10^{-5}$

§ 6. Заключение

В данной статье производится разработка и исследование аппроксимаций для оценки эффективной емкости многосерверных систем $G/G/s$ с акцентом на применимость этих аппроксимаций в условиях сценариев со строгими требованиями к качеству обслуживания трафика.

В частности, описан процесс построения универсальной приближенной формулы для оценки вероятности превысить ограничения на время пребывания в очереди для произвольной системы $G/G/s$. Аппроксимации этой вероятности являются ключевым шагом при оценке эффективной емкости системы, т.е. объема трафика, который можно обслужить, удовлетворив требования к качеству обслуживания, заданные в виде ограничений на задержку и надежность доставки данных. Приводится оценка точности построенной аппроксимации для широкого класса систем $G/G/s$ в области малых значений задержки обслуживания, а также малых вероятностей испытать эту задержку. На основании полученных численных результатов сделан вывод, что предложенные приближенные формулы показывают высокую точность при оценке вероятности превысить заданное ограничение на задержку, а также при оценке эффективной емкости системы.

В качестве результата, готового к использованию, рекомендуется формула (6) с β_{JLZ} для системы $M/M/s$ и формула (10) с β_2 для систем $G/G/s$, за исключением $M/M/s$.

Наконец, читателю предлагается асимптотический анализ разработанной приближенной формулы оценки емкости системы, причем отмечается, что сделанные наблюдения справедливы для систем гораздо шире классических $G/G/s$. Последнее позволяет надеяться, что асимптотические выкладки могут быть применены к крайне широкому классу реальных объектов и сценариев, а значит, полученные результаты будут полезны для реализаций алгоритмов управления передачей данных, работающих в режиме реального времени.

Отметим, что для адаптации предложенных результатов к более широкому классу сценариев может понадобиться усовершенствование разработанных формул для систем с конечной очередью, расширение на возможность группового поступления заявок в очередь, слотированного обслуживания, добавление учета взаимного влияния серверов (из-за взаимной интерференции между пространственными потоками MU-MIMO и одновременной передачей несколькими точками доступа) и т.д., что является предметом будущих исследований.

СПИСОК ЛИТЕРАТУРЫ

1. *Korneev E., Liubogoshchev M., Bankov D., Khorov E.* How to Model Cloud VR: An Empirical Study of Features That Matter // IEEE Open J. Commun. Soc. 2024. V. 5. P. 4155–4170. `http://doi.org/10.1109/OJCOMS.2024.3409472`
2. *Karamyshev A., Liubogoshchev M., Lyakhov A., Khorov E.* Enabling Industrial Internet of Things with Wi-Fi 6: An Automated Factory Case Study // IEEE Trans. Ind. Inform. Early access paper, August 2024. P. 1–11. `http://doi.org/10.1109/TII.2024.3431086`
3. Study on Scenarios and Requirements for Next Generation Access Technologies (3GPP Tech. Rep. TR 38.913; version 18.0.0. Release 18). May 2024.
4. *Shashin A., Belogaev A., Krasilov A., Khorov E.* Adaptive Parameters Selection for Uplink Grant-Free URLLC Transmission in 5G Systems // Comput. Netw. 2023. V. 222. P. 109527. `http://doi.org/https://doi.org/10.1016/j.comnet.2022.109527`
5. *Karamyshev A., Khorov E., Krasilov A., Akyildiz I.F.* Fast and Accurate Analytical Tools to Estimate Network Capacity for URLLC in 5G Systems // Comput. Netw. 2020. V. 178. P. 107331. `http://doi.org/https://doi.org/10.1016/j.comnet.2020.107331`
6. *Adamuz-Hinojosa O., Sciancalepore V., Ameigeiras P., Lopez-Soler J.M., Costa-Pérez X.* A Stochastic Network Calculus (SNC)-Based Model for Planning B5G uRLLC RAN Slices // IEEE Trans. Wireless Commun. 2023. V. 22. № 2. P. 1250–1265. `https://doi.org/10.1109/TWC.2022.3203937`
7. *Chinchilla-Romero L., Prados-Garzon J., Ameigeiras P., Muños P., Lopez-Soler J.M.* 5G Infrastructure Network Slicing: E2E Mean Delay Model and Effectiveness Assessment to Reduce Downtimes in Industry 4.0 // Sensors. 2022. V. 22. № 1. P. 229 (29 pp.). `http://doi.org/10.3390/s22010229`
8. *Yang P., Xi X., Quek T.Q.S., Chen J., Xianbin C., Dapeng W.* Network Slicing for URLLC // Ultra-Reliable and Low-Latency Communications (URLLC) Theory and Practice: Advances in 5G and Beyond. Hoboken, NJ, USA: Wiley, 2023. Ch. 7. P. 215–239. `https://doi.org/10.1002/9781119818366.ch7`
9. *Zhbankova E., Khakimov A., Markova E., Gaidamaka Yu.* The Age of Information in Wireless Cellular Systems: Gaps, Open Problems, and Research Challenges // Sensors. 2023. V. 23. № 19. P. 8238 (28 pp.). `http://doi.org/10.3390/s23198238`
10. *Markova E., Manaeva V.E., Zhbankova E., Moltchanov D., Balabanov P., Koucheryavy Ye., Gaidamaka Yu.* Performance-Utilization Trade-Offs for State Update Services in 5G NR Systems // IEEE Access. 2024. V. 12. P. 129789–129803. `http://doi.org/10.1109/ACCESS.2024.3442825`
11. *Chinchilla-Romero L., Prados-Garzon J., Muñoz P., Ameigeiras P., Lopez-Soler J.M.* URLLC Achieved Data Rate through Exploiting Multi-Connectivity in Industrial Private 5G Networks with Multi-WAT RANs // Proc. 2023 IEEE Wireless Communications and Networking Conf. (WCNC). Glasgow, United Kingdom. Mar. 26–29, 2023. P. 1–6. `http://doi.org/10.1109/WCNC55385.2023.10119085`
12. *Anand A., de Veciana G.* Resource Allocation and HARQ Optimization for URLLC Traffic in 5G Wireless Networks // IEEE J. Select. Areas Commun. 2018. V. 36. № 11. P. 2411–2421. `http://doi.org/10.1109/JSAC.2018.2874122`
13. *Whitt W.* The Queueing Network Analyzer // Bell Syst. Tech. J. 1983. V. 62. № 9. P. 2779–2815. `https://doi.org/10.1002/j.1538-7305.1983.tb03204.x`
14. *Halfin S., Whitt W.* Heavy-Traffic Limits for Queues with Many Exponential Servers // Oper. Res. 1981. V. 29. № 3. P. 567–588. `https://doi.org/10.1287/opre.29.3.567`

15. *Whitt W.* Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues. New York: Springer, 2002. https://doi.org/10.1007/b97479
16. *Whitt W.* A Diffusion Approximation for the $G/GI/n/m$ Queue // Oper. Res. 2004. V. 52. № 6. P. 922–941. https://doi.org/10.1287/opre.1040.0136
17. *Harchol-Balter M.* Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge: Cambridge Univ. Press, 2013.
18. *Tijms H.C.* A First Course in Stochastic Models. New York: Wiley, 2003.
19. *Kingman J.F.C.* The Single Server Queue in Heavy Traffic // Math. Proc. Cambridge Philos. Soc. 1961. V. 57. № 4. P. 902–904. http://doi.org/10.1017/S0305004100036094
20. *Köllerström J.* Heavy Traffic Theory for Queues with Several Servers. II // J. Appl. Probab. 1979. V. 16. № 2. P. 393–401. https://doi.org/10.2307/3212906
21. *Боровков А.А.* Вероятностные процессы в теории массового обслуживания. М.: Наука, 1972.
22. *Боровков А.А.* Асимптотические методы в теории массового обслуживания. М.: Наука, 1980.
23. *Janssen A.J.E.M., van Leeuwaarden J.S.H., Zwart B.* Corrected Asymptotics for a Multi-Server Queue in the Halfin–Whitt Regime // Queueing Syst. 2008. V. 58. P. 261–301. http://doi.org/10.1007/s11134-008-9070-0
24. *Janssen A.J.E.M., van Leeuwaarden J.S.H., Zwart B.* Gaussian Expansions and Bounds for the Poisson Distribution Applied to the Erlang B Formula // Adv. in Appl. Probab. 2008. V. 40. № 1. P. 122–143. http://doi.org/10.1239/aap/1208358889
25. *Puhalskii A.A., Reed J.E.* On Many-Server Queues in Heavy Traffic // Ann. Appl. Probab. 2010. V. 20. № 1. P. 129–195. https://doi.org/10.1214/09-AAP604
26. *van Leeuwaarden J.S.H., Mathijsen B.W.J., Zwart B.* Economies-of-Scale in Many-Server Queueing Systems: Tutorial and Partial Review of the QED Halfin–Whitt Heavy-Traffic Regime // SIAM Rev. 2019. V. 61. № 3. P. 403–440. http://doi.org/10.1137/17M1133944
27. *Seelen L.P., Tijms H.C.* Approximations for the Conditional Waiting Times in the $GI/G/c$ Queue // Oper. Res. Lett. 1984. V. 3. № 4. P. 183–190. https://doi.org/10.1016/0167-6377(84)90024-5
28. *Kimura T.* A Two-Moment Approximation for the Mean Waiting Time in the $GI/G/s$ Queue // Manag. Sci. 1986. V. 32. № 6. P. 751–763. https://doi.org/10.1287/mnsc.32.6.751
29. *Whitt W.* Approximations for the $GI/G/m$ Queue // Prod. Oper. Manag. 1993. V. 2. № 2. P. 114–161. https://doi.org/10.1111/j.1937-5956.1993.tb00094.x
30. *Боровков А.А.* О предельных законах для процессов обслуживания в многоканальных системах // Сиб. матем. журн. 1967. Т. 8. № 5. С. 983–1004. https://www.mathnet.ru/rus/smj5444
31. *Bouillard A.* Stochastic Network Calculus with Localized Application of Martingales. http://arxiv.org/abs/2211.05657 [cs.PF], 2024.
32. Network Simulator 3 (ns-3). https://www.nsnam.org/. Доступ: 01.08.2024.
33. *Jiang X., Luvisotto M., Pang Z., Fischione C.* Reliable Minimum Cycle Time of 5G NR Based on Data-Driven Channel Characterization // IEEE Trans. Ind. Inform. 2021. V. 17. № 11. P. 7401–7411. http://doi.org/10.1109/TII.2021.3052922

Карамышев Антон Юрьевич
Порай Екатерина Дмитриевна
Хоров Евгений Михайлович
Институт проблем передачи информации
им. А.А. Харкевича Российской академии наук, Москва
Московский физико-технический институт
(государственный университет), Москва
karamyshev@wireless.iitp.ru
porai@wireless.iitp.ru
khorov@wireless.iitp.ru

Поступила в редакцию
15.08.2024
После доработки
25.09.2024
Принята к публикации
27.09.2024