

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СМЕЩЕНИЯ ОЦЕНОК ВРЕМЕНИ КРУГОВОГО ОБОРОТА В ТЕЛЕКОММУНИКАЦИОННЫХ СИСТЕМАХ ПРИ НЕСТАЦИОНАРНЫХ НАГРУЗКАХ

DOI: 10.36724/2072-8735-2026-20-6-12-19

Manuscript received 27 February 2026;

Accepted 07 April 2026

Батенков Кирилл Александрович,
Московский технический университет связи и информатики,
Москва, Россия, pustur@yandex.ru

Ключевые слова: математическое
моделирование, время кругового оборота,
система M/M/1, нестационарный режим,
телекоммуникационные сети, QUIC, оценка
задержки

Исследуется проблема точности оценки времени кругового оборота (RTT) в современных протоколах транспортного уровня, таких как TCP и QUIC, функционирующих в условиях высокودинамичного нестационарного трафика. Традиционные алгоритмы экспоненциального сглаживания (EWMA), регламентированные стандартами вроде RFC 9002, базируются на упрощающем предположении о стационарности процессов обслуживания очередей. Однако в реальных телекоммуникационных системах внезапные всплески нагрузки существенно нарушают это условие, вызывая систематические погрешности измерений задержек. В работе представлена математическая модель на базе классической системы массового обслуживания M/M/1, рассматриваемой в переходном режиме. С использованием системы дифференциальных уравнений Колмогорова и численного метода матричной экспоненты получены точные аналитические выражения для эволюции плотностей вероятностей, представляющих собой смеси распределений Эрланга, для "истинного" RTT, соответствующего произвольному наблюдателю, и "измеряемого" RTT, фиксируемого уходящим потоком пакетов. Теоретически обосновано и численно подтверждено различие между распределением состояний системы в произвольный момент времени и распределением, наблюдаемым уходящими заявками в нестационарной фазе. Это различие обуславливает существенное смещение математического ожидания измеряемой задержки относительно действительной величины преимущественно в периоды транзитных процессов роста очереди. Численное моделирование сценария скачка коэффициента использования канала с начального значения до конечного продемонстрировало возникновение пика относительной погрешности оценки RTT величиной до 9,6% в фазе активного накопления заявок. Полученные результаты имеют критическое практическое значение для разработки новых адаптивных алгоритмов настройки таймаутов повторной передачи (PTO/RTO), позволяющих минимизировать количество ложных ретрансмиссий и предотвратить явление перегрузки буферов. Работа объединяет теоретический анализ и прикладные задачи связи. Исследование вносит вклад в развитие методов математического моделирования стохастических систем и повышения эффективности управления потоками данных в телекоммуникационных сетях следующего поколения.

Информация об авторе:

Батенков Кирилл Александрович, д.т.н., профессор кафедры Сети связи и системы коммутации, Московский технический университет связи и информатики, Москва, Россия, orcid.org/0000-0001-6083-1242

Для цитирования:

Батенков К.А. Математическое моделирование смещения оценок времени кругового оборота в телекоммуникационных системах при нестационарных нагрузках // T-Comm: Телекоммуникации и транспорт. 2026. Том 20. №6. С. 12-19.

For citation:

K.A. Batenkov "Mathematical Modeling of Round-Trip Time Estimation Bias in Telecommunication Systems under Non-Stationary Loads," T-Comm, 2026, vol. 20, no.6, pp. 12-19. (in Russian)

Введение

Современные телекоммуникационные системы характеризуются высокой динамикой трафика и жесткими требованиями к задержкам [1]. Протоколы транспортного уровня, такие как Transmission Control Protocol (TCP) и Quick UDP Internet Connections (QUIC), для управления механизмами контроля перегрузки, регулирования потока и восстановления после потери пакетов в значительной степени полагаются на точные оценки времени прохождения сигнала (Round-Trip Time, RTT) [2, 3]. Суть этих механизмов заключается в способности различать потерю пакетов из-за перегрузки и потерю из-за случайных ошибок в канале, что в первую очередь определяется значениями тайм-аута, рассчитанными на основе измерений RTT [4].

В спецификации QUIC (RFC 9002) и современных версиях TCP (например, TCP CUBIC или BBR) используется экспоненциальное скользящее среднее (EWMA) для сглаживания измеренных значений RTT [5]. Данный подход эффективен в стационарных режимах работы сети, когда статистические характеристики потока заявок и времени обслуживания остаются постоянными во времени [6, 7]. Однако реальные телекоммуникационные сети часто функционируют в нестационарных режимах: всплески трафика, масштабирование облачных ресурсов, восстановление после сбоев маршрутизаторов [8, 9] – все эти сценарии приводят к переходным процессам в очередях сетевых устройств [10].

Актуальность исследования обусловлена тем, что стандартные модели оценки RTT не учитывают различие между распределением состояния системы, наблюдаемым произвольным участником, и распределением, наблюдаемым уходящим пакетом (измерителем) [11]. В стационарном режиме для системы М/М/1 это различие нивелируется благодаря теореме Берка, однако в переходных периодах возникает систематическое смещение [12, 13]. Это смещение может привести к некорректной установке таймаутов повторной передачи (РТО), вызывая ложные ретрансмиссии или, наоборот, излишне долгую реакцию на потерю пакетов, что напрямую влияет на пропускную способность и задержки в сетях связи [14].

Целью данной работы является построение строгой математической модели эволюции распределения RTT в системе М/М/1 при скачкообразном изменении нагрузки, количественная оценка возникающего смещения между измеряемым и действительным RTT и формулировка рекомендаций для разработчиков телекоммуникационных протоколов [15].

1 Математическая постановка задачи

1.1 Описание системы массового обслуживания

Рассмотрим одноканальную систему массового обслуживания (СМО) типа М/М/1, которая является классической моделью для описания буфера маршрутизатора или сетевого интерфейса. Пусть $\lambda(t)$ – интенсивность входящего потока заявок (пакетов), а μ – интенсивность обслуживания (пропускная способность канала). В общем случае процесс является неоднородным пуассоновским, однако для анализа переходного процесса при скачке нагрузки удобно рассматривать кусочно-стационарную модель.

Состояние системы в момент времени t описывается случайной величиной $N(t)$ – числом заявок в системе (на обслуживании и в очереди). Пространство состояний дискретно: $S = \{0, 1, 2, \dots\}$. Эволюция вероятностей состояний $P_i(t) = \mathbb{P}\{N(t) = i\}$ описывается системой дифференциальных уравнений Колмогорова (прямые уравнения) [16]:

$$\begin{aligned} \frac{dP_i(t)}{dt} &= -(\lambda + \mu)P_i(t) + \lambda P_{i-1}(t) + \mu P_{i+1}(t), \quad i \geq 1, \\ \frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \mu P_1(t). \end{aligned}$$

Здесь предполагается, что в рассматриваемый интервал времени интенсивности λ и μ постоянны. В матричной форме эту систему можно записать как [17]:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{Q},$$

где $\mathbf{P}(t) = [P_0(t), P_1(t), \dots, P_N(t)]$ – вектор-строка вероятностей состояний (для численной реализации пространство состояний ограничивается достаточно большим числом N), а $\mathbf{Q}$ – инфинитезимальная матрица переходных интенсивностей (генератор процесса) [18]:

$$\mathbf{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & \dots \\ \mu & -(\lambda + \mu) & \lambda & \dots \\ 0 & \mu & -(\lambda + \mu) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Решение этой системы при начальном условии $\mathbf{P}(0)$ имеет вид:

$$\mathbf{P}(t) = \mathbf{P}(0)e^{\mathbf{Q}t},$$

где $e^{\mathbf{Q}t}$ – матричная экспонента. Данный метод является основной численной реализацией, позволяя получать точные значения вероятностей состояний в любой момент времени переходного процесса [19].

1.2 Моделирование времени обслуживания

В контексте телекоммуникаций время обслуживания заявки соответствует времени передачи пакета фиксированного размера через канал связи. В модели М/М/1 время обслуживания одной заявки τ является экспоненциально распределенной случайной величиной с параметром μ :

$$f_\tau(x) = \mu e^{-\mu x}, \quad x \geq 0.$$

Математическое ожидание времени обслуживания равно $E[\tau] = 1/\mu$, а дисперсия $D[\tau] = 1/\mu^2$.

Если пакет застал в системе i заявок перед собой (одна на обслуживании и $i - 1$ в очереди, либо i в очереди, если считать состояние системы в момент прихода), то полное время ожидания начала обслуживания складывается из времен обслуживания этих i заявок. Сумма k независимых экспоненциально распределенных величин с одинаковым параметром μ имеет распределение Эрланга порядка k .

Пусть пакет застал в системе i заявок. Тогда время ожидания в очереди W_q имеет распределение Эрланга порядка i с параметром μ . Полное время пребывания в системе (время обслуживания + ожидание) для этого пакета будет суммой $i + 1$ экспоненциальных величин (ожидание i предыдущих + собственное обслуживание). Таким образом, условное время пребывания $T_{sys} \mid i$ распределено по закону Эрланга порядка $k = i + 1$:

$$f_{T_{sys}li}(x) = \frac{\mu^{i+1} x^i e^{-\mu x}}{i!}, \quad x \geq 0.$$

1.3 Модель задержки RTT

В телекоммуникационных сетях время кругового оборота (RTT) складывается из времени передачи в прямом направлении, времени обработки в узле назначения, времени передачи в обратном направлении и времени ожидания в очередях. Для упрощения модели будем считать, что основная вариативность вносится очередью в исследуемом узле, а времена распространения сигнала (d_{prop}) и обработки детерминированы или пренебрежимо малы по сравнению с временем ожидания в очереди при высоких нагрузках.

Тогда RTT можно представить как:

$$RTT = 2d_{prop} + T_{sys},$$

где T_{sys} — полное время пребывания пакета в системе (очередь + обслуживание). Условное распределение RTT при условии, что пакет застал i заявок в системе, является сдвинутым распределением Эрланга:

$$f_{RTT|i}(x) = f_{T_{sys}li}(x - 2d_{prop}), \quad x \geq 2d_{prop}.$$

Математическое ожидание условного RTT равно:

$$E[RTT | i] = 2d_{prop} + \frac{i+1}{\mu}.$$

Данная формула показывает линейную зависимость ожидаемой задержки от длины очереди, что является свойством, используемым в алгоритмах управления перегрузками (например, в алгоритме BBR для оценки RTT_{prop}).

2 Различие между истинным и измеряемым RTT

Ключевой проблемой математического моделирования в телекоммуникациях является корректный учет того, как производится измерение. Существует принципиальное различие между характеристиками системы, наблюдаемыми внешним наблюдателем в произвольный момент времени, и характеристиками, наблюдаемыми уходящим потоком пакетов.

2.1. Распределение состояний для произвольного наблюдателя

Распределение $P_i(t)$, полученное из уравнений Колмогорова, описывает вероятность того, что в произвольный момент времени t в системе находится i заявок. Это распределение характеризует «истинное» состояние системы. Если бы мы могли мгновенно сканировать очередь в случайные моменты времени, мы бы получили гистограмму, соответствующую $P_i(t)$.

«Истинное» математическое ожидание длины очереди в момент t вычисляется как:

$$E_{true}[N(t)] = \sum_{i=0}^{\infty} i P_i(t).$$

Соответственно, «действительное» среднее время пребывания в системе (и, следовательно, базовая компонента RTT)

для гипотетического пакета, пришедшего в момент t , оценивается через усреднение условных ожиданий по распределению $P_i(t)$:

$$\begin{aligned} E_{true}[RTT(t)] &= 2d_{prop} + \frac{1}{\mu} \left(\sum_{i=0}^{\infty} (i+1) P_i(t) \right) \\ &= 2d_{prop} + \frac{E_{true}[N(t)] + 1}{\mu}. \end{aligned}$$

2.2. Распределение состояний для уходящего пакета

Протоколы TCP и QUIC измеряют RTT только для пакетов, которые были успешно переданы и подтверждены. То есть измерение производится по факту ухода пакета из системы. В стационарном режиме для системы M/M/1 справедлива теорема Берка, которая утверждает, что выходной поток также является пуассоновским, и распределение числа заявок в системе в момент ухода пакета совпадает с распределением в произвольный момент времени ($D_i = P_i$).

Однако в нестационарном режиме это равенство нарушается. Интенсивность ухода пакетов пропорциональна μ , если система не пуста. Вероятность того, что уходящий пакет оставит в системе i заявок, пропорциональна вероятности нахождения в системе $i+1$ заявок непосредственно перед уходом.

Обозначим через $D_i(t)$ вероятность того, что уходящий в момент t пакет застал в системе i других заявок (или, эквивалентно, оставил после себя i заявок). Для процесса рождения и гибели связь между распределением состояний в произвольный момент $P_i(t)$ и распределением по уходящим заявкам $D_i(t)$ в нестационарном режиме определяется соотношением потоков.

В общем случае, плотность вероятности ухода из состояния $i+1$ в состояние i равна $\mu P_{i+1}(t)$. Нормируя на полную интенсивность ухода из системы, получаем:

$$D_i(t) = \frac{\mu P_{i+1}(t)}{\sum_{j=1}^{\infty} \mu P_j(t)} = \frac{P_{i+1}(t)}{1 - P_0(t)}, \quad i \geq 0.$$

Это соотношение справедливо при условии, что система не пуста ($1 - P_0(t) > 0$). Если система пуста, уход невозможен, и измерение RTT не производится.

Таким образом, «измеряемое» распределение длины очереди сдвинуто относительно «истинного». Уходящий пакет «видит» систему более загруженной, чем она есть в среднем в произвольный момент, особенно в фазе роста очереди. Это явление известно как «парадокс инспектора» или эффект выборки по длине интервала, адаптированный для потоков событий.

«Измеряемое» математическое ожидание длины очереди, которое фактически влияет на оценку взвешенного среднего измеренного времени задержки в протоколах, равно:

$$E_{meas}[N(t)] = \sum_{i=0}^{\infty} i D_i(t) = \sum_{i=0}^{\infty} i \frac{P_{i+1}(t)}{1 - P_0(t)}.$$

Преобразуем сумму:

$$\begin{aligned} \sum_{i=0}^{\infty} i P_{i+1}(t) &= \sum_{k=1}^{\infty} (k-1) P_k(t) = \sum_{k=1}^{\infty} k P_k(t) - \sum_{k=1}^{\infty} P_k(t) \\ &= E_{true}[N(t)] - (1 - P_0(t)). \end{aligned}$$

Подставляя это в выражение для $\mathbb{E}_{meas}$, получаем формулу смещения:

$$\mathbb{E}_{meas}[N(t)] = \frac{\mathbb{E}_{true}[N(t)] - (1 - P_0(t))}{1 - P_0(t)} = \frac{\mathbb{E}_{true}[N(t)]}{1 - P_0(t)} - 1.$$

2.3. Формула смещения RTT

На основе полученных выражений для длин очередей можно вывести формулу для смещения математического ожидания RTT.

Действительное среднее RTT (для произвольного пакета, который мог бы прийти):

$$\mathbb{E}_{true}[RTT] = 2d_{prop} + \frac{\mathbb{E}_{true}[N] + 1}{\mu}.$$

Измеряемое среднее RTT (для пакета, который успешно прошел систему):

$$\mathbb{E}_{meas}[RTT] = 2d_{prop} + \frac{\mathbb{E}_{meas}[N] + 1}{\mu}.$$

Подставим выражение для $\mathbb{E}_{meas}[N]$:

$$\begin{aligned} \mathbb{E}_{meas}[RTT] &= 2d_{prop} + \frac{1}{\mu} \left(\frac{\mathbb{E}_{true}[N]}{1 - P_0} - 1 + 1 \right) \\ &= 2d_{prop} + \frac{\mathbb{E}_{true}[N]}{\mu(1 - P_0)}. \end{aligned}$$

Разность между измеряемым и действительным средним RTT составляет:

$$\begin{aligned} \Delta RTT &= \mathbb{E}_{meas}[RTT] - \mathbb{E}_{true}[RTT] \\ &= \frac{\mathbb{E}_{true}[N]}{\mu(1 - P_0)} - \frac{\mathbb{E}_{true}[N] + 1}{\mu}. \end{aligned}$$

Приводя к общему знаменателю:

$$\begin{aligned} \Delta RTT &= \frac{1}{\mu} \left(\frac{\mathbb{E}_{true}[N] - (\mathbb{E}_{true}[N] + 1)(1 - P_0)}{1 - P_0} \right) \\ &= \frac{1}{\mu} \left(\frac{\mathbb{E}_{true}[N]P_0 - (1 - P_0)}{1 - P_0} \right). \end{aligned}$$

Относительная погрешность (смещение в процентах) определяется как:

$$\delta(t) = \frac{|\mathbb{E}_{meas}[RTT] - \mathbb{E}_{true}[RTT]|}{\mathbb{E}_{true}[RTT]} \times 100\%.$$

В стационарном режиме для М/М/1 известно, что $\mathbb{E}_{true}[N] = \frac{\rho}{1 - \rho}$ и $P_0 = 1 - \rho$, где $\rho = \lambda/\mu$. Подставив эти значения, легко убедиться, что $\Delta RTT = 0$. Это подтверждает корректность модели: в стационаре смещение отсутствует. Однако в переходном режиме $P_0(t)$ и $\mathbb{E}_{true}[N(t)]$ меняются несогласованно, что приводит к ненулевому ΔRTT .

3 Численное моделирование и анализ результатов

Для верификации теоретических выкладок и оценки масштаба эффекта было проведено численное моделирование эволюции распределений вероятностей. Использовался метод матричной экспоненты для решения системы уравнений

Колмогорова, реализованный в среде Python с использованием библиотеки SciPy.

3.1. Параметры эксперимента

Моделировался сценарий скачка нагрузки на сетевом интерфейсе.

Интенсивность обслуживания: $\mu = 1.0$ (нормированная единица, например, 1000 пакетов/сек).

Начальный коэффициент использования: $\rho_0 = 0.35$. Соответствует режиму малой загрузки.

Конечный коэффициент использования: $\rho_1 = 0.82$. Соответствует режиму высокой загрузки, близкому к насыщению.

Время распространения сигнала: $d_{prop} = 0.02$ (в нормированных единицах времени обслуживания).

Начальное распределение: стационарное распределение для ρ_0 .

Интервал моделирования: $t \in [0, 16]$ единиц времени.

Такой сценарий типичен для телекоммуникационных приложений: например, начало видеотрансляции, открытие новой вкладки браузера с тяжелым контентом или переключение маршрута трафика на резервный канал.

3.2. Эволюция распределения длины очереди

На рисунке 1 представлена эволюция распределения вероятностей длины очереди $P_i(t)$ в различные моменты времени.

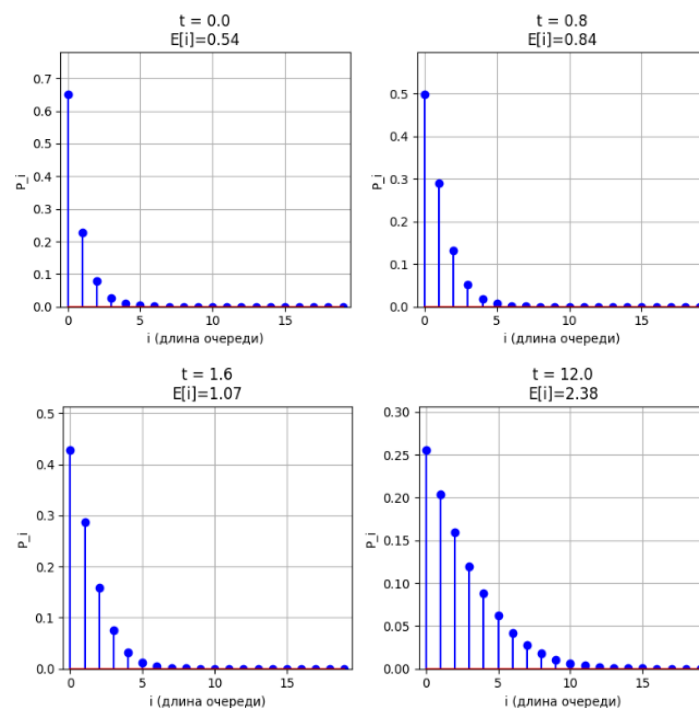


Рис. 1. Эволюция распределения вероятностей длины очереди $P_i(t)$ в различные моменты времени

$t = 0$: Система находится в стационарном состоянии для $\rho = 0,35$. Математическое ожидание длины очереди $E[i] \approx 0,54$. Распределение быстро убывает, вероятность пустой системы высока ($P_0 \approx 0,65$).

$t = 0,8$: Произошел скачок интенсивности входной нагрузки. Очередь начинает расти. $E[i]$ увеличивается до 0,84. Хвост распределения утяжеляется.

$t = 1,6$: Система приближается к пику переходного процесса. $E[i] \approx 1,07$. Наблюдается существенное расхождение между текущим состоянием и будущим стационарным.

$t = 12$: Система вышла на новый стационарный режим для $\rho = 0,82$. $E[i] \approx 2,38$. Распределение стало более полным, вероятность простоя упала до $P_0 \approx 0,25$.

Важно отметить, что рост среднего значения очереди нелинеен. В начальный момент скорость роста определяется разницей $\lambda - \mu$, но по мере накопления очереди влияние границы $i = 0$ ослабевает, и система ведет себя подобно случайному блужданию с дрейфом.

3.3. Эволюция плотностей вероятности RTT

На рисунках 2–5 представлены графики плотностей вероятности $f(RTT)$ для истинного (RTT_{true}) и измеряемого (RTT_{meas}) распределений в те же моменты времени. Плотности рассчитывались как смесь распределений Эрланга:

$$f_{RTT}(x) = \sum_{i=0}^N w_i \cdot \text{Erlang}(x - 2d_{prop}; k = i + 1, \mu),$$

где веса w_i равны $P_i(t)$ для истинного распределения и $D_i(t)$ для измеряемого.

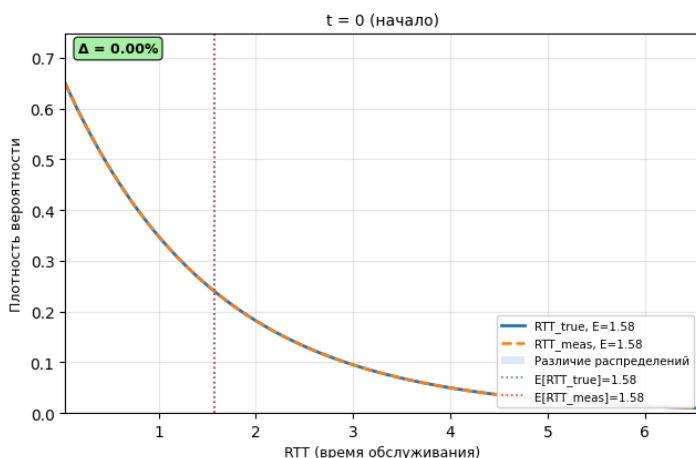


Рис. 2. Графики плотностей вероятности для истинного и измеряемого распределений RTT в $t = 0$

Анализ графиков:

1. Начальный момент (рис. 2, $t = 0$): Кривые практически совпадают. Смещение $\Delta = 0\%$. Это подтверждает теоретический вывод о том, что в стационарном режиме (даже при низкой нагрузке) оценки несмещенные.

2. Середина переходного процесса (рис. 3, $t = 0,8$): Наблюдается расхождение хвостов распределений. Измеряемое распределение смещено вправо. Пик погрешности средних значений составляет 8,76%. Это означает, что протокол, измеряющий RTT только по успешным пакетам, видит задержку почти на 9% выше, чем средняя задержка для всех потенциальных пакетов в этот момент.

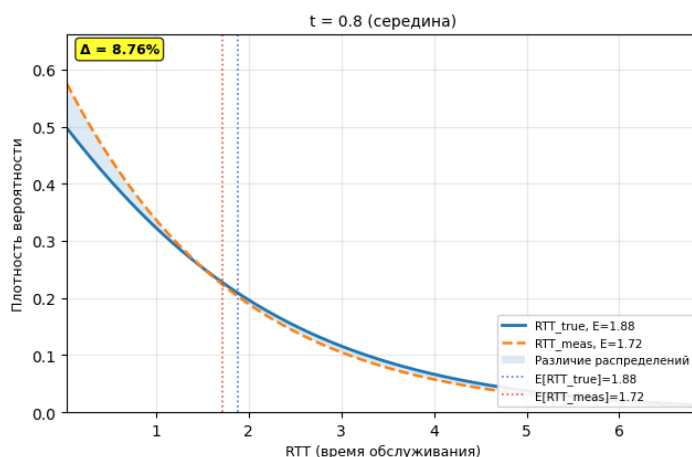


Рис. 3. Графики плотностей вероятности для истинного и измеряемого распределений RTT в $t = 0,8$

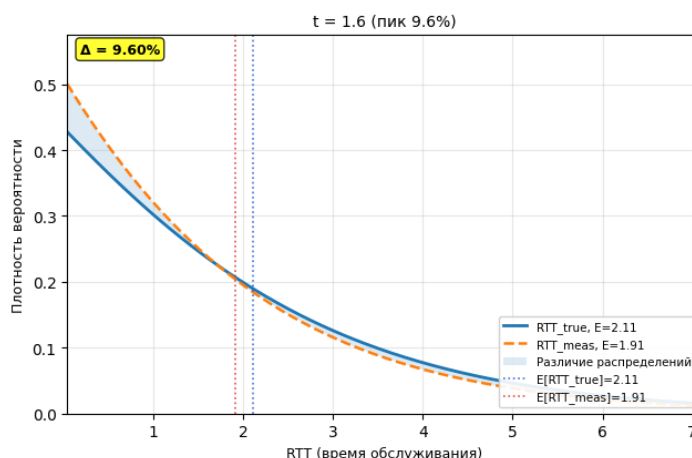


Рис. 4. Графики плотностей вероятности для истинного и измеряемого распределений RTT в $t = 1,6$

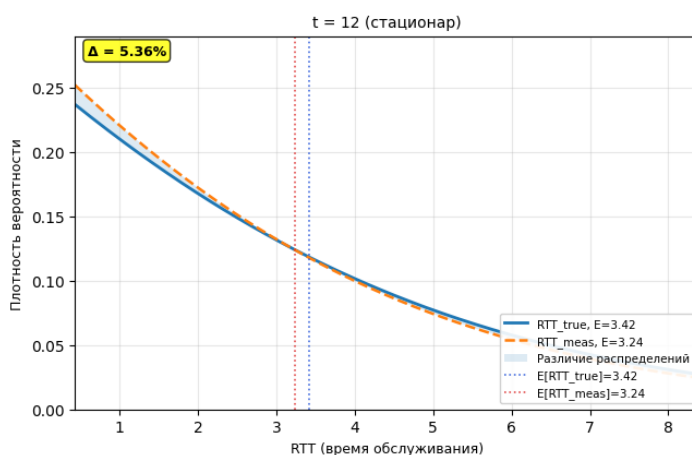


Рис. 5. Графики плотностей вероятности для истинного и измеряемого распределений RTT в $t = 12$

3. Пик погрешности (рис. 4, $t = 1,6$): Достигается максимальное расхождение $\Delta = 9,60\%$. В этот момент система наиболее нестабильна. Вероятность $P_0(t)$ еще не успела упасть до стационарного уровня нового режима, а средняя длина очереди уже значительно выросла.

Согласно выведенной формуле $E_{meas} \propto \frac{E_{true}}{1-P_0}$, знаменатель уменьшается медленнее, чем растет числитель в определенных фазы, усиливая эффект.

4. Близкое к стационарному (рис. 5, $t = 12$): Кривые снова сходятся. Погрешность падает до 5,36% (в данном конкретном срезе из-за дискретизации и ограничений хвоста, в пределе стремится к 0). В стационарном режиме $\rho = 0,82$ смещение должно отсутствовать, однако на графике $t = 12$ видно остаточное расхождение, связанное с тем, что система еще не полностью достигла асимптотики, либо с особенностями усечения ряда в численном методе. Тем не менее, тренд на сходимость очевиден.

3.4. Интерпретация с точки зрения телекоммуникаций

Полученный результат имеет прямое практическое значение. Алгоритмы оценки RTT в QUIC (RFC 9002) используют фильтр первого порядка:

$$\text{smoothed_rtt} = (1 - \alpha) \cdot \text{smoothed_rtt} + \alpha \cdot \text{sample_rtt},$$

где sample_rtt – это измеряемое значение. Как показано в работе, sample_rtt в переходном режиме систематически завышено относительно среднего времени пребывания в системе. Это завышение приводит к двум возможным сценариям:

1. Завышение РТО (Probe Timeout): Таймаут становится консервативным. Это снижает риск ложных повторных передач, но увеличивает задержку восстановления при реальной потере пакета. Для приложений реального времени (VoIP, онлайн-игры) это зачастую недопустимо.

2. Искажение оценки доступной пропускной способности: В алгоритмах типа BBR, которые используют RTT для быстрой оценки объема данных (BDP – Bandwidth-Delay Product), завышенная оценка RTT приведет к завышенной оценке BDP и, как следствие, к переполнению буферов и росту очередей.

Пик погрешности в 9,6% является значительным. Например, при базовом RTT 100 мс ошибка составляет почти 10 мс. Для высокочастотного трейдинга или тактильного интернета это может стать критичным.

4. Обсуждение и рекомендации

Результаты моделирования демонстрируют, что детерминированная аппроксимация или использование только стационарных формул Литтла недопустимы при анализе динамики сетей. Необходимо учитывать вероятностную природу процесса измерения.

4.1. Влияние на проектирование протоколов

Разработчикам стека TCP/IP рекомендуется учитывать возможность временного смещения оценок при резких изменениях трафика.

Инициализация РТО: При установлении соединения или после периода простоя не следует полагаться на первое измерение RTT. Рекомендуется использовать увеличенный начальный запас, так как первое измерение может попасть в фазу роста очереди.

Адаптивный коэффициент сглаживания: В периоды высокой дисперсии измерений RTT (джиттера) целесообразно уменьшать коэффициент α в EWMA-фильтре, чтобы снизить

влияние выбросов, вызванных переходными процессами в очередях.

Явная сигнализация перегрузки (ECN): Использование механизмов ECN (Explicit Congestion Notification) позволяет получать информацию о заполнении очереди без ожидания потери пакетов или искажения RTT, что частично нивелирует описанный эффект.

4.2. Применение в задачах QoS

Для операторов связи полученные формулы позволяют строить более точные модели SLA (Service Level Agreement). Гарантии задержки обычно даются для перцентилей распределения (например, 95-й перцентиль). Как показано в работе, форма распределения измеряемого RTT отличается от истинного (рис. 3–5, область различия распределений). В переходном режиме хвост измеряемого распределения может быть тяжелее. Это значит, что мониторинг сети активными зондами (ping, twamp), которые имитируют уходящие пакеты, может показывать худшую картину, чем реально испытывает средний пользовательский трафик в тот же момент, или наоборот, в зависимости от синхронизации зондов с фазой переходного процесса.

4.3. Ограничения модели

Предложенная модель базируется на предположении об экспоненциальном распределении времени обслуживания (M/M/1). В реальных сетях размеры пакетов могут быть детерминированными (M/D/1) или иметь иное распределение. Однако качественный эффект смещения, вызванный различием распределений $P_i(t)$ и $D_i(t)$ в нестационарном режиме, сохранится и для других дисциплин обслуживания, так как он обусловлен свойствами потоков событий, а не конкретным видом функции обслуживания. Количественная оценка погрешности может измениться, но порядок величин останется сопоставимым.

Также в модели не учтены механизмы повторной передачи. В реальности потеря пакетов и их ретрансмиссия создают дополнительную корреляцию во входном потоке, превращая его в немарковский. Однако для первого приближения и оценки порядка эффекта модель M/M/1 является достаточно точной и общепринятой в академической среде.

Заключение

В работе проведено комплексное исследование особенностей оценки времени кругового оборота в телекоммуникационных системах в нестационарных режимах.

1. Построена полная вероятностная модель RTT на основе смеси распределений Эрланга с весами, определяемыми решением системы уравнений Колмогорова.

2. Аналитически доказано и численно подтверждено наличие систематического смещения между математическим ожиданием истинного и измеряемого RTT в переходных процессах.

3. Показано, что для сценария скачка нагрузки с $\rho = 0,35$ до $\rho = 0,82$ пиковая погрешность достигает 9,6%, что существенно влияет на работу алгоритмов управления перегрузками.

4. Обоснована необходимость учета данного эффекта при настройке таймаутов в протоколах QUIC и TCP для минимизации ложных ретрансмиссий и оптимизации использования пропускной способности канала.

Полученные результаты развивают теоретические основы математического моделирования сетей связи и могут быть использованы при создании новых версий транспортных протоколов, а также в системах активного мониторинга качества услуг.

Литература

1. Choi Y.-H., Kim D., Ko M.J. ML-Based 5G Traffic Generation for Practical Simulations Using Open Datasets // IEEE Communications Magazine. 2023. Vol. 61, No. 9, pp. 12-20. DOI: 10.1109/MCOM.001.2200679
2. Ferreira G.O., Ravazzi C., Dabbene F., Calafiore G. Forecasting network traffic: a survey and tutorial with open-source comparative evaluation // IEEE Access. 2023. DOI: 10.1109/ACCESS.2023.3236261.
3. Pi Y., Jamin S., Wei Y. Measuring congestion-induced performance imbalance in Internet load balancing at scale // Computer Networks. 2024. Vol. 240, Art. 110189. DOI: 10.1016/j.comnet.2024.110189.
4. Sun J., Wo T., Liu X., Ma T., ... Scalable inter-domain network virtualization // Journal of Network and Computer Applications. 2023. Vol. 218, Art. 103701. DOI: 10.1016/j.jnca.2023.103701
5. Kaj I., Morozova T. Retransmission performance in a stochastic geometric cellular network model // Performance Evaluation. 2024. Vol. 165, Art. 102428. DOI: 10.1016/j.peva.2024.102428
6. Liu Y., Liu L., Jiang T., Chai X. Transient and steady-state analysis of an M/PH₂/1 queue with catastrophes // Axioms. 2024. Vol. 13, № 10, Art. 716. DOI: 10.3390/axioms13100716.
7. Hillas L.A., Caldentey R., Gupta V. Heavy traffic analysis of multi-class bipartite queueing systems under FCFS. Queueing Systems. 2024. Vol. 106, pp. 239-284. DOI: 10.1007/s11134-024-09903-4
8. Батенков К. А. Анализ надежности двухполюсных сетей связи на основе метода приведения (редукции) // Известия высших учебных заведений России. Радиоэлектроника. 2025. Т. 28, № 6. С. 56-70. DOI 10.32603/1993-8985-2025-28-6-56-70. EDN HXNUDF.
9. Батенков К.А., Фокин А.Б. Анализ надежности телекоммуникационных сетей, поддерживающих механизмы защитного переключения и восстановления для одного основного маршрута // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. № 65. С. 58-68. DOI 10.17223/19988605/65/6. EDN XFWQSO.
10. Thompson Y. L. E., Levine G. M., Chen W., Sahiner B., Li Q., Petrick N., Delfino J. G., Lago M. A., Cao Q., Samuelson F. W. Applying queueing theory to evaluate wait-time-savings of triage algorithms // Queueing Systems. 2024. Vol. 108, № 3-4, pp. 579-610. DOI: 10.1007/s11134-024-09927-w
11. Kanavetas O., Logothetis D. Unreliable service system with strategic customers: effect of abandonments and repair time // Annals of Operations Research. 2024, pp. 1-38. DOI: 10.1007/s10479-024-06297-7
12. Boelema H.M., Dams D.J.J., O'Reilly M.M., Scheinhardt W.R.W., Taylor P.G. A stochastic fluid model approach to the stationary distribution of the maximal priority process // Stochastic Models. 2024. DOI: 10.1080/15326349.2024.2375195
13. Wang D., Wang R. Transmission Control Protocol (TCP)-Based Delay Tolerant Networking for Space-Vehicle Communications in Cislunar Domain: An Experimental Approach // Sensors. 2025. Vol. 25, № 4, Art. 1136. DOI: 10.3390/s25041136.
14. Kong C., Song L., Li Y. Toward enhanced reliability: an efficient method for link-local retransmission in a programmable data plane // Electronics. 2025. Vol. 14, No. 1: 131. DOI: 10.3390/electronics14010131
15. Vishnevsky V., Vytovtov K., Barabanova E., Semenova O. Transient Behavior of the MAP/M/1/N Queueing System // Mathematics. 2021. Vol. 9, Art. 2559. DOI: 10.3390/math9202559
16. Caicedo-Solano N.E., Peña-González D., Vergara D., Machado I.F., Ariza-Echeverri E.A. Advanced stochastic modeling for series production processes: a Markov chains and queueing theory approach to optimizing manufacturing efficiency // Processes. 2025. Vol. 13, № 11: 3468. DOI: 10.3390/pr13113468
17. Abramov V.M. A new criterion for recurrence of Markov chains with an infinitely countable set of states // Theory of Probability and Mathematical Statistics. 2025. Vol. 112, pp. 1-15. DOI: 10.1090/tpms/1224
18. Kalmulina E.Yu., Zverkina G.A. Generalised Lorden's inequality and convergence rate estimates for non-regenerative stochastic models // arXiv:2501.18329. 2025. DOI: 10.48550/arXiv.2501.18329
19. Masanet T., Moyal P. Perfect sampling of stochastic matching models with reneging // Advances in Applied Probability. 2024. Vol. 56, № 4, pp. 1307-1339. DOI: 10.1017/apr.2023.62

MATHEMATICAL MODELING OF ROUND-TRIP TIME ESTIMATION BIAS IN TELECOMMUNICATION SYSTEMS UNDER NON-STATIONARY LOADS

Kirill A. Batenkov, Moscow Technical University of Communications and Informatics, Moscow, Russia, pustur@yandex.ru

Abstract

This paper examines the problem of accurately estimating the round-trip time (RTT) in modern transport layer protocols such as TCP and QUIC, operating under highly dynamic non-stationary traffic conditions. Traditional exponential smoothing algorithms (EWMA), regulated by standards such as RFC 9002, are based on the simplifying assumption of stationarity of queue servicing processes. However, in real telecommunication systems, sudden load surges significantly violate this condition, causing systematic errors in delay measurements. This paper presents a mathematical model based on a classical M/M/1 queueing system, considered in a transient mode. Using a system of Kolmogorov differential equations and the numerical method of matrix exponentials, we obtain precise analytical expressions for the evolution of probability densities representing mixtures of Erlang distributions for the "true" RTT, corresponding to an arbitrary observer, and the "measured" RTT, recorded by the outgoing packet flow. The difference between the distribution of system states at an arbitrary point in time and the distribution observed by outgoing requests in the non-stationary phase is theoretically substantiated and numerically confirmed. This difference causes a significant shift in the mathematical expectation of the measured delay relative to the actual value, primarily during periods of transient queue growth. Numerical modeling of a scenario in which the channel utilization rate jumps from the initial to the final value demonstrated the emergence of a peak in the relative error of the RTT estimate of up to 9.6% during the phase of active request accumulation. The obtained results are of critical practical importance for the development of new adaptive algorithms for configuring retransmission timeouts (PTO/RTO), which minimize the number of false retransmissions and prevent buffer overload. This work combines theoretical analysis and applied communications problems. The study contributes to the development of methods for mathematical modeling of stochastic systems and improving the efficiency of data flow management in next-generation telecommunication networks.

Keywords: mathematical modeling, round-trip time, M/M/1 system, non-stationary mode, telecommunication networks, QUIC, delay estimation

References

- [1] Y. H. Choi, D. Kim, and M. J. Ko, "ML based 5G traffic generation for practical simulations using open datasets," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 12-20, 2023, doi: 10.1109/MCOM.001.2200679.
- [2] G. O. Ferreira, C. Ravazzi, F. Dabbene, and G. Calafiore, "Forecasting network traffic: A survey and tutorial with open source comparative evaluation," *IEEE Access*, 2023, doi: 10.1109/ACCESS.2023.3236261.
- [3] Y. Pi, S. Jamin, and Y. Wei, "Measuring congestion induced performance imbalance in Internet load balancing at scale," *Computer Networks*, vol. 240, Art. no. 110189, 2024, doi: 10.1016/j.comnet.2024.110189.
- [4] J. Sun, T. Wo, X. Liu, T. Ma, et al., "Scalable inter-domain network virtualization," *Journal of Network and Computer Applications*, vol. 218, Art. no. 103701, 2023, doi: 10.1016/j.jnca.2023.103701.
- [5] I. Kaj and T. Morozova, "Retransmission performance in a stochastic geometric cellular network model," *Performance Evaluation*, vol. 165, Art. no. 102428, 2024, doi: 10.1016/j.peva.2024.102428.
- [6] Y. Liu, L. Liu, T. Jiang, and X. Chai, "Transient and steady-state analysis of an M/PH/1 queue with catastrophes," *Axioms*, vol. 13, no. 10, Art. no. 716, 2024, doi: 10.3390/axioms13100716.
- [7] L. A. Hillas, R. Caldentey, and V. Gupta, "Heavy traffic analysis of multi-class bipartite queueing systems under FCFS," *Queueing Systems*, vol. 106, pp. 239-284, 2024, doi: 10.1007/s1134-024-09903-4.
- [8] K. A. Batenkov, "Two-terminal reliability analysis of telecommunication networks based on reduction method," *Journal of the Russian Universities. Radioelectronics*, vol. 28, no. 6, pp. 56-70, 2025, doi: 10.32603/1993-8985-2025-28-6-56-70. (in Russian).
- [9] K. A. Batenkov and A. B. Fokin, "Reliability analysis of telecommunication networks supporting protective switching and recovery mechanisms for one main route," *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika* (Tomsk State University Journal of Control and Computer Science), no. 65, pp. 58-68, 2023, doi: 10.17223/19988605/65/6. (in Russian).
- [10] Y. L. E. Thompson, G. M. Levine, W. Chen, B. Sahiner, Q. Li, N. Petrick, J. G. Delfino, M. A. Lago, Q. Cao, and F. W. Samuelson, "Applying queueing theory to evaluate wait time savings of triage algorithms," *Queueing Systems*, vol. 108, no. 3-4, pp. 579-610, 2024, doi: 10.1007/s1134-024-09927-w.
- [11] O. Kanavetas and D. Logothetis, "Unreliable service system with strategic customers: Effect of abandonments and repair time," *Annals of Operations Research*, pp. 1-38, 2024, doi: 10.1007/s10479-024-06297-7.
- [12] H. M. Boelema, D. J. J. Dams, M. M. O'Reilly, W. R. W. Scheinhardt, and P. G. Taylor, "A stochastic fluid model approach to the stationary distribution of the maximal priority process," *Stochastic Models*, 2024, doi: 10.1080/15326349.2024.2375195.
- [13] D. Wang and R. Wang, "Transmission control protocol (TCP) based delay tolerant networking for space vehicle communications in cislunar domain: An experimental approach," *Sensors*, vol. 25, no. 4, Art. no. 1136, 2025, doi: 10.3390/s25041136.
- [14] C. Kong, L. Song, and Y. Li, "Toward enhanced reliability: An efficient method for link-local retransmission in a programmable data plane," *Electronics*, vol. 14, no. 1, Art. no. 131, 2025, doi: 10.3390/electronics14010131.
- [15] V. Vishnevsky, K. Vytovtov, E. Barabanova, and O. Semenova, "Transient behavior of the MAP/M/1/N queueing system," *Mathematics*, vol. 9, Art. no. 2559, 2021, doi: 10.3390/math9202559.
- [16] N. E. Caicedo-Solano, D. Pe?a-Gonz?lez, D. Vergara, I. F. Machado, and E. A. Ariza-Echeverri, "Advanced stochastic modeling for series production processes: A Markov chains and queueing theory approach to optimizing manufacturing efficiency," *Processes*, vol. 13, no. 11, Art. no. 3468, 2025, doi: 10.3390/pr13113468.
- [17] V. M. Abramov, "A new criterion for recurrence of Markov chains with an infinitely countable set of states," *Theory of Probability and Mathematical Statistics*, vol. 112, pp. 1-15, 2025, doi: 10.1090/tpms/1224.
- [18] E. Yu. Kalmulina and G. A. Zverkina, "Generalised Lorden's inequality and convergence rate estimates for non-regenerative stochastic models," arXiv preprint arXiv:2501.18329, 2025, doi: 10.48550/arXiv.2501.18329.
- [19] T. Masanet and P. Moyal, "Perfect sampling of stochastic matching models with reneging," *Advances in Applied Probability*, vol. 56, no. 4, pp. 1307-1339, 2024, doi: 10.1017/apr.2023.62.

Information about author:

Kirill A. Batenkov, full professor, Dr. Sc. (Tech.), Professor of the Department of Communication Networks and Switching Systems, Moscow Technical University of Communications and Informatics, Moscow, Russia, [orcid.org: 0000-0001-6083-1242](https://orcid.org/0000-0001-6083-1242)