Введение

Понятие трафика (телетрафика) изначально было связано с проектированием телефонных сетей, направленным на минимизацию количества линий, соединяющих различные АТС, с учетом вероятности потери вызова и длительности ожидания абонента. Первые работы в этой области проводились в начале XX века А.К. Эрлангом. В этих работах в качестве моделей потока запросов, поступающих в узел сети, рассматривался ординарный поток без последействия (пуассоновский поток). Подобный подход надолго лег в основу проектирования нагрузки на телефонные станции, поскольку потоки запросов, поступавших на станцию, являлись суперпозицией большого числа потоков малой интенсивности, поступавших от конечного числа абонентов сети. Более того, все основные положения теории массового обслуживания, являющейся основой математического аппарата для теории телетрафика, были получены в предположении независимости (а следовательно, и некоррелированности) интервалов времени между поступающими заявками и, соответственно, потока интервалов времени обслуживания поступающих заявок. Предполагалось также, что процессы поступления и обслуживания независимы.

Явление зависимости в процессах поступления в очереди активно начало изучаться в 70–80-х годах прошлого века [1–7].

Дальнейшее развитие технологий, появление компьютерных сетей, сетей ISDN (Integrated Service Digital Networks), сетей мобильной связи и многих других новых технологий показали, что модель использования входного стационарного пуассоновского потока, имеющего коэффициент автокорреляции Пирсона ρ для интервалов времени между заявками, равный 0, для описания трафика в таких сетях не подходит, поскольку трафик в таких сетях, как минимум, имеет коэффициент вариации C_v, превосходящий 1 (для пуассоновского потока $C_v = 1$). Именно поэтому появилась необходимость в математических моделях узлов сетей,

представленных как система массового обслуживания (СМО), позволяющая учесть корреляцию во входном потоке [8, 9].

Зависимость между временем поступления и временем обслуживания может оказывать серьезное влияние на характеристики СМО [10, 11]. Оценивание такой зависимости было впервые определено как важная проблема в сетях связи и впервые была серьезно исследована Л. Клейнроком [12]. Л. Клейнрок рассматривал модель сети массового обслуживания (как сеть связи), в которой заявки проходили через ряд очередей на пути от источника к приемнику. Поскольку одна и та же заявка посещает каждую очередь, время обслуживания каждой заявки в последовательных очередях обычно будет иметь положительную корреляцию и даже может быть идентичным. Кроме того, если есть зависимость между временем обслуживания в двух последовательных очередях, то время поступления и время обслуживания во второй очереди будут являться зависимыми, а следовательно, и коррелированными. В такой ситуации, когда время поступления и время обслуживания сильно коррелированы, задержки во второй очереди СМО имеют тенденцию быть меньше, чем в случае отсутствия зависимости [13–16]. Зависимость интервалов времени обслуживания и ее влияние на основные характеристики СМО рассматривается, например, в [17–20].

Следует отметить основные гипотезы, объясняющие причины появления автокорреляции трафика. Большинство причин были описаны в различных работах примерно 20 лет назад, и с тех пор не появилось принципиально новых публикаций, объясняющих возникновение автокорреляции. Первая гипотеза заключается в том, что источник автокорреляции находится на прикладном уровне. Вторая гипотеза рассматривает происхождение коррелированного трафика на транспортном уровне, в частности из-за влияния алгоритмов управления TCP. Также в качестве причин возникновения самоподобного (а следовательно, и коррелированного) трафика указываются механизмы генерации данных. Следует заметить, что в [21] было показано, что выборочные автокорреляционные функции реального трафика, обладающего самоподобными свойствами, сходятся, но становятся более плавными, когда размер выборки увеличивается, что допускает отождествление самоподобного и коррелированного трафика. Проведенные измерения показывают [22, 23], что самоподобность в сетевом трафике является двумерным свойством,

которое относится частично к распределениям времен между поступлениями элементов трафиковых последовательностей (пакеты, кадры и т. д.) и частично к распределениям размеров этих элементов. Различные источники могут генерировать на самом верхнем уровне трафик со статистически различными характеристиками, но его поведение, характеризуемое корреляционной структурой, как правило, остается инвариантно при прохождении различных элементов сети [24, 25].

Наиболее широко известная, эмпирически подтвержденная теория происхождения коррелированного трафика на прикладном уровне утверждает, что коррелированный трафик является причиной поведения пользователей. Впервые она была разработана в [26, 27], где трафик в Интернете представляется как большое количество ON/OFF-источников с одинаковыми распределениями длительности периодов активности, что хорошо характеризует, например, распределение времени обдумывания пользователя [23].

В [28–30] была обнаружена связь автокорреляции трафика и распределения длин передаваемых по сети файлов. Размеры файлов, передаваемых по сети, были исследованы в [31–34], где было продемонстрировано, что они имеют ярко выраженные «тяжелохвостные» распределения, что и обозначалось как причина возникновения автокорреляции. В целом эта теория ON/OFF-источников в основном присуща клиент-серверной сетевой архитектуре [22, 35, 36], и ее варианты стали доминирующим объяснением появления коррелированного сетевого трафика и являются основой наиболее часто используемой модели для симуляции трафика данных.

Также одной из основных причин являются, предположительно, алгоритмы управления TCP. Протокол TCP является доминирующим протоколом в Интернете, определяющим динамику трафика. Он является протоколом, основанным на обратной связи, работа которого может быть оценена только частично с использованием аналитических или стохастических моделей.

Существование невырожденных конечных корреляционных структур, присущих именно коррелированному, а не самоподобному трафику, наблюдалось на различных уровнях, таких как трафик Ethernet [37], трафик глобальных сетей [38] и WWW-трафик [28]. Существует множество исследований, посвященных изучению различных аспектов появления автокорреляции,

например рассматриваются сопутствующие модели трафика [39-=42], определяется влияние автокорреляции на производительность сети [43, 44]. В [45–47] на основе анализа контроля перегрузки TCP было показано, что контроль перегрузки (вследствие его хаотичности) может привести к появлению не только автокорреляции в трафике, но и самоподобия. Вместе с тем исследования, проведенные в [48–50], указывают на то, что сетевой трафик не является строго самоподобным и что автокорреляция присутствует только в определенных границах временных масштабов, значения которых напрямую зависят от механизмов управления перегрузкой и предотвращения перегрузки в TCP. Косвенно это может подтверждаться исследованиями, показывающими, что наличие автокорреляции во входном потоке при загрузках, превышающих 1, может снижать средний размер очереди [25].

Также стоит заметить, что агрегирование (статистическое мультиплексирование) трафика предполагает наложение множества ON/OFF-источников. Если в какой-то момент агрегированный трафик станет проявлять корреляционные свойства, то он сохранит их до самого приемника [23, 24]. Присутствие корреляции (медленно убывающей зависимости в трафике передачи данных в пакетных сетях), причины ее возникновения, способы описания и моделирования были показаны достаточно давно [22].

Таким образом, при анализе работы какого-либо устройства, представленного системой массового обслуживания, можно ожидать наличие следующих типов проявления зависимостей:

- корреляция между последовательными интервалами времени между поступлением заявок;
- корреляция между последовательными интервалами времени обслуживания заявок;
- взаимная корреляция указанных последовательностей интервалов времени.

Важность учета корреляции обусловлена ее существенным влиянием (как негативным, так и позитивным) на основные характеристики системы массового обслуживания. Вычисления и результаты расчетов на реальных сетях показывают, что наличие положительной корреляции во входном потоке существенно ухудшает основные показатели работы СМО (увеличивает среднее время обслуживания, вероятность отказа и т. д.). Например,

при одной и той же средней скорости поступления запросов, одинаковом распределении времени обслуживания и одинаковой емкости буфера вероятность потери запроса может отличаться на несколько порядков [8, 51]. Даже простое упорядочивание интервалов времени между заявками в простейшем стационарном потоке заявок с экспоненциальным распределением интервалов времени между заявками таким образом, чтобы «короткие» интервалы оказались в начале интервала времени анализа, а «длинные» — в конце, существенно увеличивает значение коэффициента корреляции Пирсона и средний размер очереди [52].

Существует большое количество способов описания потоков входных заявок, использующих пуассоновский поток, самоподобные модели, аппроксимационные модели входного потока, групповые марковские входные потоки (BMAP) [53] и др. Однако приведенные методы обладают определенными недостатками, такими как сложность проведения расчетов для определения конченых показателей качества обслуживания, слишком высокая сложность аналитических моделей, неполное задание элементов системы массового обслуживания и др.

Для преодоления сложностей использования указанных методов анализа и моделирования коррелированного трафика и для получения относительно простых методов количественного анализа показателей качества обслуживания трафика в моографии рассматривается аппроксимационный подход к решению поставленной задачи, основанный на использовании гиперэкспоненциальных распределений и корреляционных функций, полученных путем измерений на реальных трассах. В случае взаимной коррелированности последовательностей интервалов времени поступления и обслуживания предлагается анализ на основе синтеза двумерной плотности вероятностей, который можно осуществить, вводя в рассмотрение понятие функции копулы от одномерных распределений и коэффициентов ранговой корреляции Кендалла и Спирмена.

Также приводится методика представления коррелированного трафика эквивалентной последовательностью некоррелированных случайных величин (а именно, некоррелированной последовательностью интервалов времени между поступлениями заявок и некоррелированными значениями интервалов времени обслуживания заявок) для упрощения анализа (и прогнозирования) вероятностно-временных свойств сетевых узлов обработки

трафика. Показано, что на практике с точки зрения экономии вычислительного ресурса реализовать декорреляцию последовательности интервалов времени между заявками целесообразно с использованием вейвлет-преобразования Хаара.