

Направления развития теории очередей и её применение при проектировании телекоммуникационных сетей

В.М. Вишнеvский¹

¹Федеральное государственное бюджетное учреждение науки Институт проблем управления им. В.А. Трапезникова Российской академии наук

Математические модели сетей и систем массового обслуживания (СМО) широко применяются для исследования различных технических, экономических, производственных, медицинских, военных и других систем. Особенно эффективно эти модели используются при проектировании современных телекоммуникационных сетей, включая анализ существующих и перспективных сетевых протоколов, оптимизацию алгоритмов маршрутизации и топологической структуры сети и т.д.

В настоящей лекции будет дано описание основных направлений и этапов развития теории очередей и их тесной связи с прогрессом в области телекоммуникационных сетей. Следует отметить, что выделение этапов развития теории очередей является достаточно условным, но оно отражает основные тенденции развития теории очередей от момента её становления до настоящего времени. На конкретных примерах проектирования телекоммуникационных систем и сетей будет обоснована необходимость и эффективность применения моделей теории очередей.

Первый этап связан с появлением в конце 19-го века телефонных сетей и, соответственно, проектирования автоматических телефонных станций (АТС). Впервые возникла проблема доступа к ограниченному коммуникационному ресурсу. Её разрешил русский инженер М.Ф. Фрейденберг, представивший инженерный расчёт по оценке производительности АТС в зависимости от числа абонентов и вероятности отказа в их обслуживании. Строгая математическая модель была разработана основоположником теории очередей датским математиком и инженером А.К. Эрлангом в рамках его работ по исследованию телефонных сетей в период 1909-1922гг. Он предположил, что случайный поток телефонных запросов является стационарным, ординарным пуассоновским потоком без последствия, а время обслуживания имеет показательное (экспоненциальное) распределение. При выполнении этих предположений число запросов в рассматриваемой А.К. Эрлангом системы описывалось однородным марковским процессом, что позволило найти его стационарное распределение. Результаты рассмотренной модели хорошо согласовывались с

результатами измерений на реальных телефонных сетях. Позднее этот факт был объяснён благодаря работам А.Я. Хинчина, Г.Г. Ососкова и Б.И. Григелиониса, которые доказали, что суперпозиция большого числа рекуррентных потоков малой интенсивности сходится к стационарному пуассоновскому потоку. В дальнейшем развитие теории очередей шло в направлении усложнения моделей СМО. Появились многочисленные статьи и книги по исследованию многолинейных и приоритетных СМО, систем стохастического поллинга и т.д., а также СМО с пуассоновским входным потоком и произвольной функцией распределения времени обслуживания.

Была доказана знаменитая формула Полячека-Хинчина по оценке стационарных характеристик СМО типа M/G/1. В рамках доказательства было найдено преобразование Лапласа-Стилтьеса $\psi(s)$ функции распределения времени ожидания заявок в очереди такой СМО. Соответственно, путём дифференцирования $\psi(s)$ при $s \rightarrow 0$ определялось среднее время ожидания (первый момент $T^{(1)}$).

$$T^{(1)} = \lim_{s \rightarrow 0} \frac{d\psi(s)}{ds} = \frac{\lambda b^{(2)}}{2(1-\rho)}, \text{ где}$$

λ – интенсивность входного пуассоновского потока, $b^{(2)}$ – второй момент функции распределения времени обслуживания и ρ – загрузка систем. Аналогично определялись и моменты более высокого порядка. Например, второй момент функции распределения времени ожидания

$$T^{(2)} = \lim_{s \rightarrow 0} \frac{d^2\psi(s)}{d^{(2)}s}.$$

В дальнейшем были разработаны различные методы анализа стационарных характеристик СМО с пуассоновским входным потоком и произвольной функцией распределения времени обслуживания, включая метод вложенных цепей Маркова, метод введения дополнительного события и т.д.

Второй этап развития теории очередей связан с появлением и широким внедрением компьютерных сетей, в которых использовался эффективный метод коммутации пакетов в отличие от метода коммутации каналов, применявшегося в телефонных сетях. Возникла проблема разработки нового математического аппарата – исследование моделей сетей массового обслуживания (СeМО) для оптимального проектирования компьютерных сетей с коммутацией пакетов. Решению этой проблемы посвящены работы американского учёного Леонарда Клейнрока, который в 1976г. опубликовал монографию «Вычислительные системы с очередями», Его исследования по открытым СеМО эффективно использовались для синтеза топологической структуры, управления маршрутизацией, выбора оптимальных параметров

сетевых протоколов и т.д. В дальнейшем результаты Д. Клейнрока были расширены и обобщены на класс замкнутых, открытых и смешанных СеМО с мультипликативным представлением вероятности состояний, удовлетворяющих ограничениям теоремы ВСМР. Простейшим примером такой сети является замкнутая экспоненциальная сеть, состоящая из M узлов, каждый из которых представляет собой однолинейную СМО с неограниченным накопителем. Количество заявок в сети постоянно и равно N , а переходы из одного узла в другой осуществляются в соответствии с маршрутной матрицей $Q = \|Q_{ij}\|$, где Q_{ij} – вероятность перехода из узла i в узел j , $i, j = \overline{1, M}$.

Функционирование замкнутой СеМО описывается многомерным марковским процессом

$$N(t) = \{n_1(t), n_2(t), \dots, n_M(t)\},$$

где n_k – число заявок, находящихся в k -й СМО в момент t ($k = \overline{1, M}$). Стационарные вероятности состояний этой сети имеют мультипликативный вид

$$P(n) = G^{-1}(N, M) \prod_{i=1}^M d_i^{n_i}, n \in S(N, M),$$

где $S(N, M)$ – пространство состояний марковского процесса, а $G(N, M)$ – нормализующая константа, определяемая из условия нормировки. Для вычисления стационарных характеристик открытых, замкнутых и смешанных СеМО были разработаны эффективные вычислительные алгоритмы, описание которых было приведено в книге Bruell S.C., Balbo G. *Computational Algorithms for Queuing Network*. 1980. В 1988 г. была опубликована книга Вишнеvский В.М., Жожикашвили В.А. «Сети массового обслуживания. Теория и применение в сетях ЭВМ», в которой было приведено систематизированное изложение теории СеМО и, главное её применение при проектировании системы «Сирена». Разработанная нами первая в стране компьютерная сеть «Сирена» охватила всю территорию бывшего СССР. В сети были реализованы новейшие для того времени методы пакетной коммутации, адаптивная маршрутизация, управление потоками и т.д. Сеть была реализована на полностью отечественных аппаратно-программных комплексах. Теоретические методы проектирования и внедрения сети «Сирена», включая новые методы и алгоритмы теории СеМО, были опубликованы в монографии В.М, Вишнеvский «Теоретические основы проектирования компьютерных сетей», опубликованной в 2003 г.

Новый импульс развития теоретических исследований (3 этап) связан с появлением цифровых сетей интегрального обслуживания (ЦИО)

информационные потоки в которых являются не стационарными и главное коррелированными. Было доказано, что применение моделей СМО с пуассоновским входящим потоком, в котором коэффициент корреляции между соседними интервалами равен 0, при проектировании вычислительных систем и сетей нового поколения приводит к значительным погрешностям. В связи с этим в 80-е годы возникла проблема разработки математической модели коррелированных потоков, адекватно описывающих информационные потоки в технических и социальных сетях нового поколения. Эта проблема была решена двумя группами научных коллективов – американским под руководством М. Ньютса и российским под руководством Г.П. Башарина, предложивших в качестве такой модели МАР-потоки и их обобщение ВМАР, ММАР и т.д. До настоящего времени ведутся активные исследования СМО с коррелированными входящими потоками. Большой вклад в становление и развитие теории СМО с коррелированными потоками внесен школой Белорусского государственного университета под руководством профессора Дудина А.Н. В 2000 г. появилась одна из первых книг этого направления: А.Н. Дудин и В.И. Клименок «Теория очередей с коррелированными потоками». За последние два десятилетия опубликовано огромное количество статей в этом направлении. В 2020г. опубликована книга В.М. Вишнеvский, А.Н. Дудин, В.И. Клименок «Стохастические системы с корректированными потоками. Теория и применение в телекоммуникационных сетях», в которой систематизированы методы и подходы анализа СМО с коррелированными входными потоками.

Однако до настоящего времени осталось множество задач по исследованию СМО и СеМО большой размерности с входящим МАР-потоком, для решения которых применение традиционных методов и алгоритмов отыскания численных решений либо затруднено, либо вовсе невозможно. Примерами таких задач являются: исследование многолинейных приоритетных СМО с входными ММАР потоками; анализ характеристик сетей и многофазных СМО большой размерности; исследования систем адаптивного, динамического поллинга и СМО типа fork-join с коррелированным входным потоком и PH-распределением времени обслуживания и др. Отсутствие методов и подходов к аналитическому и численному решению таких задач сдерживает практическое применение моделей теории очередей при оценке производительности и проектировании сложных технических и социальных сетей.

Один из подходов для решения этой проблемы заключается в объединении аналитических, численных и имитационных методов

моделирования таких СМО с методами машинного обучения. В результате такого объединения создаются обученные модели машинного обучения, пригодные для качественного и, что главное, быстрого вычисления характеристик таких СМО. Для иллюстрации этого подхода далее мы подробно рассмотрим пример проектирования современной широкополосной беспроводной сети с использованием модели многофазной СМО большой размерности и входящим МАР-потокком, для исследования которой эффективно применяются методы машинного обучения. Такая беспроводная сеть вдоль окружной дороги города Казань (М7 Волга) была разработана нами по заказу ГИБДД республики Татарстан несколько лет назад и эффективно функционирует до настоящего времени. Сеть предназначена для оперативной передачи (со скоростью до 300 мбит/с) информации с камер видеонаблюдения и автомобилей ГИБДД в центр управления. На рисунках 1-3 приведены схемы, поясняющие назначения и структуру разработанной широкополосной беспилотной сети.

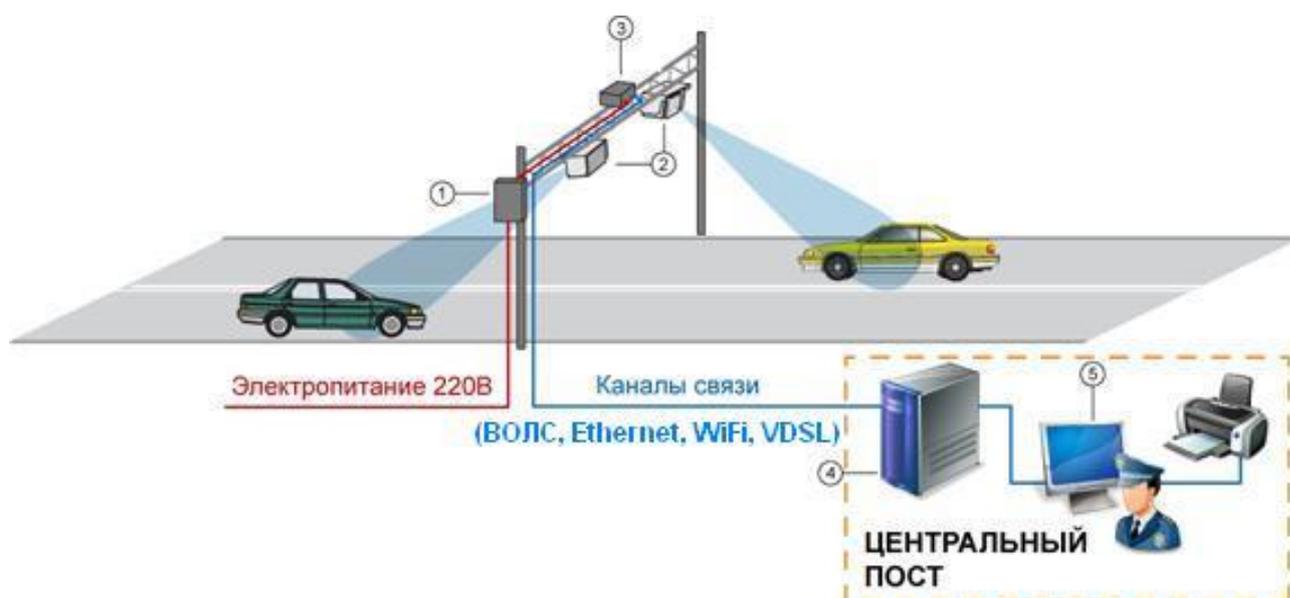


Рис. 1. Схема видеонаблюдения транспортных средств и возможные варианты передачи информации в центр управления



Рис. 2. Фрагмент беспроводной сети, функционирующий в сантиметровом диапазоне радиоволн

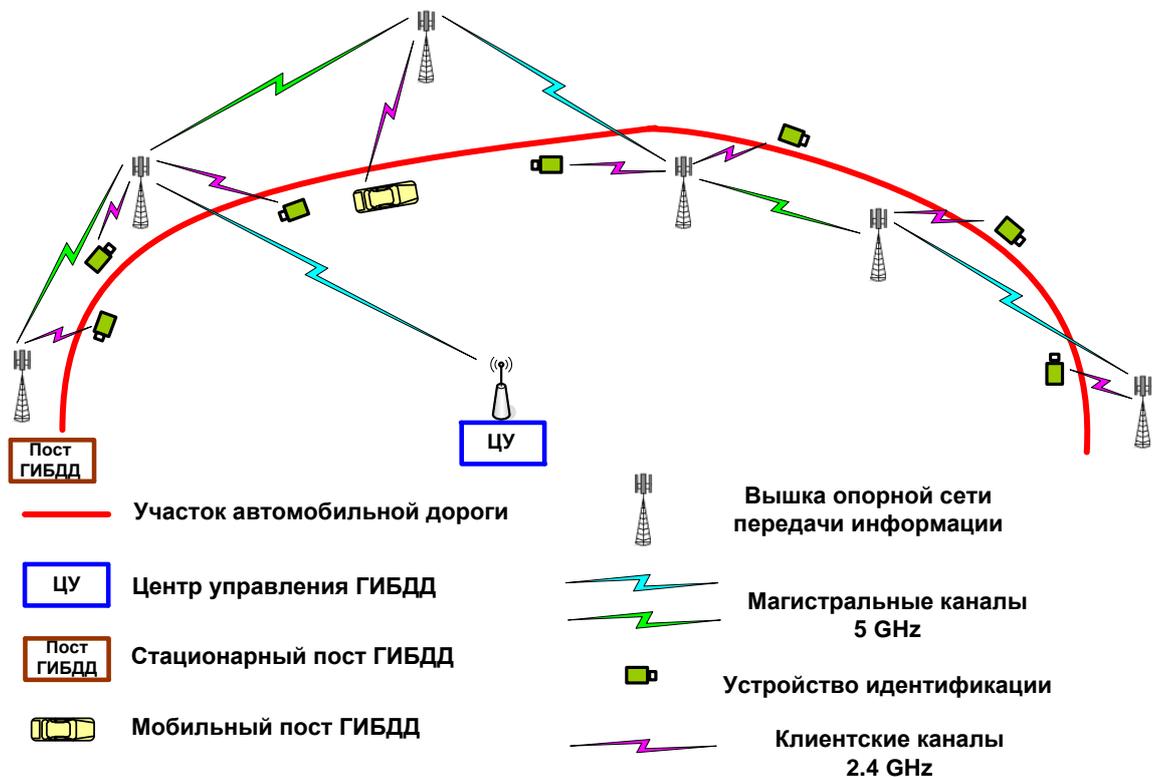


Рис. 3. Магистральная сеть (100 км) с беспроводными каналами IEEE802.11n (300 Мбит/с)

Математической моделью, адекватно описывающей функционирование беспроводной сети является многофазная СМО с коррелированными входными потоками, РН-распределением времени обслуживания и ограниченными буферными накопителями на фазах (рис. 4).

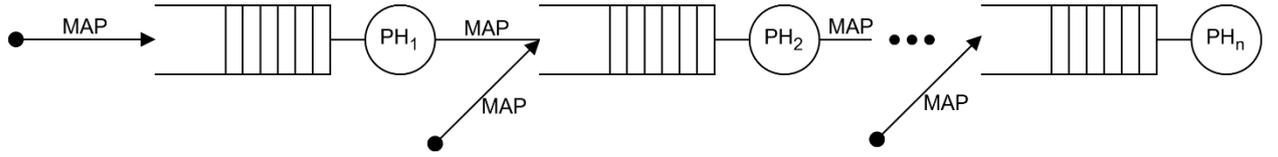


Рис. 4. Математическая модель многофазной системы

Исследование двухфазных СМО посвящено большое количество публикаций отечественных и зарубежных авторов. Однако отсутствуют точные аналитические решения и алгоритмы для многофазных СМО с входящим MAP-поток и числом фаз > 2 . Для решения этой задачи при проектировании беспроводной сети мы использовали замечательное свойство СМО MAP/PH/1/M – замкнутость на множестве MAP-потоков согласно следующим двум теоремам:

Теорема 1. Поток выходных (обслуженных) пакетов в системе MAP/PH/1/M, где входной MAP-поток задается $X: \text{MAP}(D_0, D_1)$, а время обслуживания имеет фазовое распределение $Y: \text{PH}(S, \bar{\tau})$, является MAP-поток $Z \sim \text{MAP}(D'_0, D'_1)$, матрицы которого имеют вид

$$D'_0 = \begin{bmatrix} D_0 \otimes I_V & D_1 \otimes (\bar{\tau} \otimes \bar{I}_V) & 0 & \dots & 0 & 0 \\ 0 & D_0 \otimes S & D_1 \otimes I_V & \dots & 0 & 0 \\ 0 & 0 & D_0 \otimes S & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & D_0 \otimes S & D_1 \otimes I_V \\ 0 & 0 & 0 & \dots & 0 & (D_0 + D_1) \otimes S \end{bmatrix},$$

$$D'_1 = \begin{bmatrix} 0 & \dots & 0 & 0 \\ I_W \otimes C_t & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & I_W \otimes C_t & 0 \end{bmatrix},$$

где $C_t = (-S\bar{I}_V) \otimes \bar{\tau}$, а I_V, I_W – единичные матрицы порядков V и W соответственно. Отметим, что эта теорема является обобщением известной теоремы Бурке.

Теорема 2 Суперпозиция MAP-потоков X_1 и X_2 , $X_i \sim \text{MAP}(D_0^{(i)}, D_1^{(i)})$, $i=1,2$ – MAP-поток

$$X = X_1 \oplus X_2 \sim \text{MAP}(D_0^{(1)} \oplus D_0^{(2)}, D_1^{(1)} \oplus D_1^{(2)}),$$

где \oplus – сумма Кронекера. Если потоки X_1 и X_2 имеют размерности W_1 и W_2 , то размерность суммарного потока X равна $W = W_1 W_2$.

Обозначим Z_i выходящий поток с i -го узла тандемной системы, а \hat{X}_i – общий входящий поток на i -й узел (см. рис. 1). Тогда согласно теоремам 1 и 2 потоки \hat{X}_i и Z_i являются MAP-потоками. Таким образом, i -й узел представляет собой систему массового обслуживания $\text{MAP}_i/\text{PH}_i/1/M_i$, интенсивность поступления пакетов в которой λ_i . Для данной системы

хорошо известны формулы расчета основных характеристик производительности: средняя длина очереди m_i ; вероятность потери пакета $P_L^{(i)}$; среднее время пребывания (задержка) пакета на i -м узле T_i и др. Вычисление этих характеристик позволяет определить значения искомых параметров вероятности потери пакетов в тандемной системе

$$P_L = 1 - \prod_{i=1}^N (1 - P_L^{(i)})$$

и время межконцевой задержки

$$T = \sum_{i=1}^N T_i = \sum_{i=1}^N \frac{m_i^{(i)}}{(1 - P_L^{(i)})\lambda_i}.$$

В следующем подразделе приведен формальный алгоритм нахождения стационарных характеристик производительности тандемной сети.

Алгоритм точного расчета стационарных характеристик производительности тандемной сети

Формально алгоритм точного расчета характеристик тандемной системы имеет следующий вид.

Шаг 1. Положим $i := 1$.

Шаг 2. Если $i = 1$, то положим $\hat{X}_i = X_1$. Если же $i > 1$, то вычисляем \hat{X}_i : $\hat{X}_i = Z_{i-1}$, если в сети нет кросс-трафика, иначе $\hat{X}_i = Z_{i-1} \oplus X_i$. Обозначим матрицы потока \hat{X}_i как $\hat{\mathbf{D}}_{i,0}$ и $\hat{\mathbf{D}}_{i,1}$, т. е. $\hat{X}_i = \text{MAP}(\hat{\mathbf{D}}_{i,0}, \hat{\mathbf{D}}_{i,1})$.

Шаг 3. С помощью теоремы 1 вычисляем матрицы $\mathbf{D}'_{i,0}, \mathbf{D}'_{i,1}$ MAP-потока $Z_i = \mathcal{D}(\hat{X}_i, Y_i, M_i)$.

Шаг 4. Для выходящего MAP-потока Z_i вычисляем его стационарное распределение $\bar{\theta}^{(i)}$ с помощью системы линейных алгебраических уравнений

$$\begin{cases} \bar{\theta}^{(i)} (\mathbf{D}'_{i,0} + \mathbf{D}'_{i,1}) = 0, \\ \bar{\theta}^{(i)} \mathbf{1} = 1. \end{cases}$$

Шаг 5. Рассчитываем среднее число заявок в очереди на i -й фазе

$$m_i^{(i)} = \sum_{k=0}^{M_i+1} k \sum_{j=1}^{V_i \hat{W}_i} \theta_{kV_i \hat{W}_i + j}^{(i)},$$

где $V_i = |Y_i|$ – порядок РН-распределения Y_i , а $\hat{W}_i = |\hat{X}_i|$ – порядок входящего MAP-потока \hat{X}_i .

Шаг 6. Определяем стационарное распределение вероятностей $\bar{\pi}^{(i)}$ входящего потока \hat{X}_i . Если в сети нет кросс-трафика и $i > 1$, то полагаем $\bar{\pi}^{(i)} \equiv \bar{\theta}^{(i-1)}$. В противном случае находим $\bar{\pi}^{(i)}$ как решение системы линейных алгебраических уравнений

$$\begin{cases} \bar{\pi}^{(i)} (\hat{\mathbf{D}}_{i,0} + \hat{\mathbf{D}}_{i,1}) = 0, \\ \bar{\pi}^{(i)} \mathbf{1} = 1. \end{cases}$$

Шаг 7. С помощью найденного на предыдущем шаге стационарного распределения $\bar{\pi}^{(i)}$ входящего потока \hat{X}_i вычисляем интенсивность поступления заявок на i -й узел

$$\lambda_i = \bar{\pi}^{(i)} \hat{\mathbf{D}}_{i,1} \bar{\mathbf{1}}.$$

Шаг 8. Рассчитываем распределение состояний входящего МАР-потока при наличии в системе $M_i + 1$ заявки (т. е. при заполненной системе)

$$\bar{\Psi}^{(i)} = \left(\sum_{j=1}^{V_i} \{\bar{\theta}_{M_i+1}^{(i)}\}_j, \dots, \sum_{j=1}^{V_i} \{\bar{\theta}_{M_i+1}^{(i)}\}_{(\hat{W}_{i-1})V_i+j} \right).$$

Здесь вектор $\bar{\theta}_{M_i+1}^{(i)}$ – часть вектора $\bar{\theta}^{(i)}$, относящаяся к состояниям системы, когда в ней находится $M_i + 1$ заявка.

Шаг 9. Вычисляем вероятность потери заявки из-за переполнения i -й очереди

$$P_L^{(i)} = \bar{\Psi}^{(i)} \frac{\hat{\mathbf{D}}_{i,0} \bar{\mathbf{1}}}{\lambda_i}.$$

Шаг 10. Вычисляем среднюю задержку на i -м узле

$$T_i = \frac{m_1^{(i)}}{(1 - P_L^{(i)})\lambda_i}.$$

Шаг 11. Если $i < N$, то увеличиваем $i := i + 1$ и переходим на шаг 2. В противном случае переходим далее, на шаг 12.

Шаг 12. Вычисляем вероятность потери заявки $P_L = 1 - \prod_{i=1}^N (1 - P_L^{(i)})$.

Шаг 13. Вычисляем общую задержку $T = \sum_{i=1}^N T_i$.

Оценка сложности алгоритма нахождения точных характеристик производительности многофазной системы

Предложенная схема проста в вычислении. По сути, на каждом шаге с помощью нескольких операций произведения Кронекера строятся блочные матрицы для выходящего МАР-потока. Далее решается система линейных алгебраических уравнений для определения стационарных вероятностей входящего и исходящего потока. Наконец, с помощью нескольких операций умножения найденных распределений на матрицы потоков вычисляются искомые характеристики – вероятность потери пакета, средний размер системы и межконцевая задержка. Главный недостаток этой схемы расчета – чрезвычайно высокая вычислительная сложность.

Утверждение 1. Пусть входящие МАР-потоки имеют порядок W , PH-распределения – порядок V , емкость буфера на каждой фазе равна m и сеть содержит N станций. Тогда итерационная схема расчета характеристик тандемной сети имеет сложность:

– $O((MVW)^{3N})$, если в сети есть кросс-трафик;

– $O(W^3(MV)^{3N})$, если кросс-трафика в сети нет.

Д о к а з а т е л ь с т в о. Рассмотрим i -ю итерацию алгоритма, $i \leq N$, т. е. расчет характеристик на i -м узле сети. Отметим сперва, что при $i > 1$ порядок выходящего потока с предыдущего $i-1$ -го узла есть $(M+2)V\hat{W}_i$, где \hat{W}_i – порядок входящего потока на i -й узел. При наличии в сети кросс-трафика $U_i = ((M+2)VW)^i$, а если кросс-трафика нет, то $U_i = W((M+2)V)^i$.

Сложность итерации определяется шагами 4 и 6 алгоритма, на которых необходимо решать системы линейных алгебраических уравнений, причем порядок матрицы системы на шаге 4 (генератор выходящего потока) заведомо выше, чем системы на шаге 6 (генератор входящего потока). Полагая, что для решения системы используется алгоритм наподобие метода Гаусса, на шаге 4 потребуется $O(U_i^3)$ операций. Остальные шаги имеют более низкую сложность: шаги 1, 10 и 11 – $O(1)$, шаг 2 – $O(U_{i-1}^2 W^2)$, шаг 3 – $O(U_i^2)$, шаг 5 – $O(VW+M)$, шаги 7 и 9 – $O(U_i^2)$, шаг 8 – $O(VM)$. Сложность шагов 12 и 13 есть $O(N)$. ♦

Таким образом, если в сети есть кросс-трафик, сложность алгоритма составит

$$O((VWM)^3) + O((VWM)^6) + \dots + O((VWM)^{3N}) + \\ + O(N) = O(VWM)^{3N},$$

а если кросс-трафика в сети нет, то

$$O(W^3(VM)^3) + O(W^3(VM)^6) + \dots + O(W^3(VM)^{3N}) + \\ + O(N) = O(W^3(VM)^{3N}).$$

Таким образом, искать решение с помощью описанного алгоритма становится сложно даже при относительно небольших значениях N , V и W . В табл. 1 приведены примеры значений порядков МАР-потоков в зависимости от значений параметров системы. Из таблицы следует, что получение точного решения возможно лишь при числе узлов $N < 5$. Для практического применения тандемных систем большой размерности с узлами МАР/РН/1/М необходимы более эффективные методы расчета.

Таблица 1 – Порядки выходящих МАР-потоков в зависимости от порядков РН-распределения (V), входящих МАР-потоков (W) и емкости буфера (M)

Параметры системы			Номер узла				
W	V	M	1	2	3	4	5
Сети без кросс-трафика							
1	1	1	3	9	27	81	243
1	1	3	5	25	125	625	3 125
2	2	2	16	128	1 024	8 192	65 536
3	1	3	15	75	375	1 875	9 375
1	3	3	15	225	3 375	50 625	759 375
3	3	3	45	675	10 125	151 875	2 278 125
Сети с кросс-трафиком							
1	1	1	3	9	27	81	243
1	1	3	5	25	125	625	3 125
2	2	2	16	256	4 096	65 536	1 048 576
3	1	3	15	225	3 375	50 625	759 375
1	3	3	15	225	3 375	50 625	759 375
3	3	3	45	2 025	91 125	4 100 625	184 528 125

Для решения данной задачи предложен подход, базирующийся на комбинации методов имитационного моделирования и машинного обучения (рис. 5). В рамках данного метода на разных наборах входных параметрах с помощью имитационного моделирования генерируется набор данных, в котором рассчитаны характеристики производительности тандемной сети. Далее с помощью сгенерированного набора данных применяются алгоритмы машинного обучения для получения быстрых оценок характеристик производительности. Метод эффективно применялся при решении различных задач теории очередей с коррелированными входными потоками.

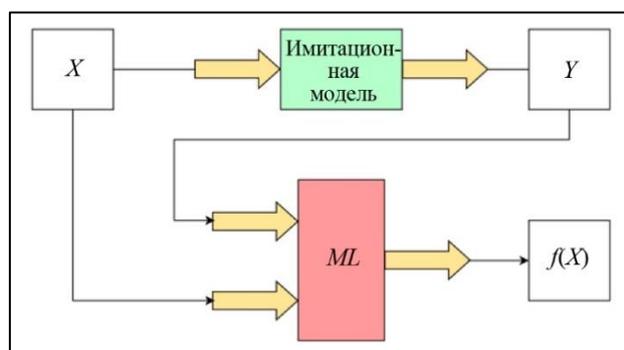


Рис. 5. Методология получения быстрых оценок стационарных характеристик производительности сети

Основными этапами являются:

1. Точный расчёт характеристик многофазной системы небольшой размерности.

2. Имитационная модель и формирование датасета.
3. Применение алгоритмов машинного обучения и формирование результатов.

В недавно опубликованной книге В. М. Вишнеvский, Д. В. Ефросинин «Теория очередей и машинное обучение» дано описание многочисленных алгоритмов машинного обучения и систематизированное описание подхода к исследованию сложных СМО с использованием машинного обучения.

Заканчивая выступление, хочу отметить, что в нашей стране ведутся активные исследования в области теории очередей и её применения при проектировании компьютерных сетей. Такие исследования ведутся в школе РУДН под руководством профессора Самойлова К.Е., в Корельской научной школе под руководством профессора Морозова Е. В., Вологодской школе под руководством профессора Зейфмана А.И., Владивостокской школе под руководством профессора Цициашвили Г.Ш., Школе по теории надежности под руководством профессора Рыкова В.В. и др. Отдельно отмечу научную школу Омского государственного университета под руководством профессора Назарова А.А. и профессора Моисеевой С.П., в которой возникла и успешно реализована замечательная идея организации этой Международной конференции.

Спасибо за внимание!