ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ И ТЕОРИЯ ОЧЕРЕДЕЙ

Ю. И. Рыжиков (Санкт-Петербург)

На первой и в особенности второй конференциях ИММОД ряд авторов – как молодых, так и маститых – ставил вопрос о *теории имитационного моделирования*. В докладах, заявляющих «концепции», «методологии» и даже «парадигмы», недостатка не было. Однако в задаваемых вопросах и в дискуссии ощущалась явная тоска по стройной, прозрачной и *считающей* математической теории. Если признать объектами имитации системы и сети с очередями, что будет лишь незначительным сужением предметной области, то *такая теория есть*. Она так прямо и называется – queuing theory (в русскоязычном варианте – теория массового обслуживания – ТМО). Эта теория насчитывает почти сто лет и накопила огромный, но мало востребованный вычислительный потенциал. Сведения из нее остро необходимы нынешней компьютерной молодежи, натасканной на технологии, но не осознавшей *базовых идей.* К примеру, не было необходимости моделировать, чтобы убедиться, что при регулярном потоке среднее время задержки меньше, чем при экспоненциальном: элементарный здравый смысл подсказывает, что при равных средних любое увеличение степени неопределенности ухудшает качество обслуживания. Столь же очевидно, что внесение в цикл диспетчирования временных задержек при передаче заявок на обслуживание не приводит к повышению уровня обслуживания и прибыльности.

В докладе дается очерк основных понятий и возможностей ТМО с одновременным разбором типичных, к сожалению, девиаций – по материалам ранее состоявшихся конференций ИММОД.

ТМО – прикладная ветвь теории вероятностей

Теория очередей (она же – теория массового обслуживания) в ее современном состоянии позволяет исследовать процессы накопления и рассасывания очередей в стационарных (установившихся) режимах работы систем и сетей. Соответственно исходные данные и результаты расчета задаются вероятностными распределениями и/или их числовыми характеристиками. В качестве последних обычно выбираются моменты распределений. При не слишком стеснительных условиях моменты однозначно определяют преобразование Лапласа от распределения и, следовательно, его плотность и функцию распределения.

На результирующие показатели систем обслуживания влияют не только первые моменты (средние значения), но и высшие – см. рис. 1, 2 – графики дополнительной функции распределения (ДФР) времени пребывания заявки в системах *M/G*/1 и *GI/M*/1 с немарковским распределением Вейбулла с проставленными на них коэффициентами вариации *v* при единичных средних и коэффициенте загрузки 0,8.

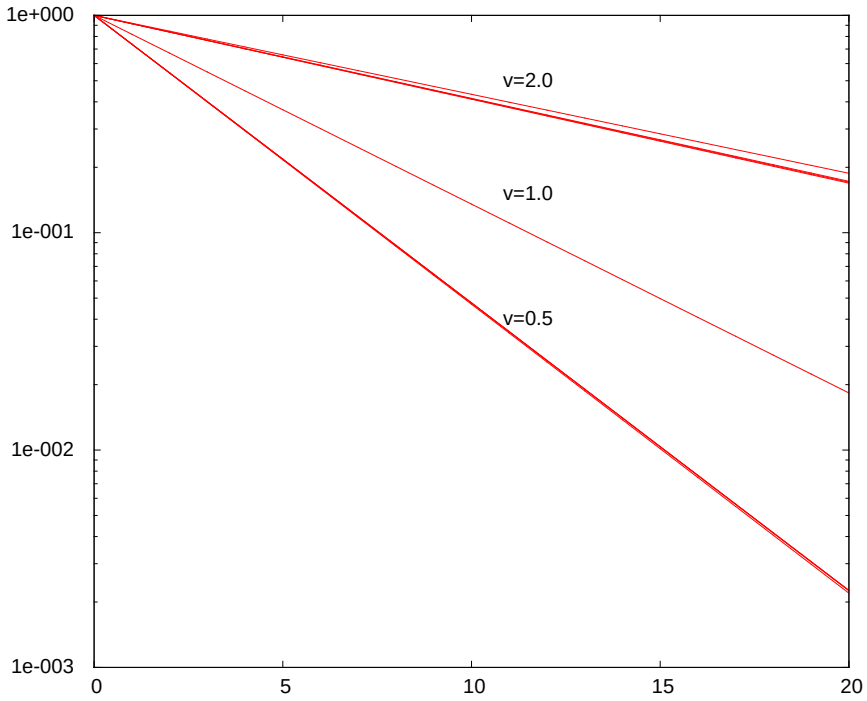

Рис. 1. Распределение числа заявок в системе *GI*/*M*/1

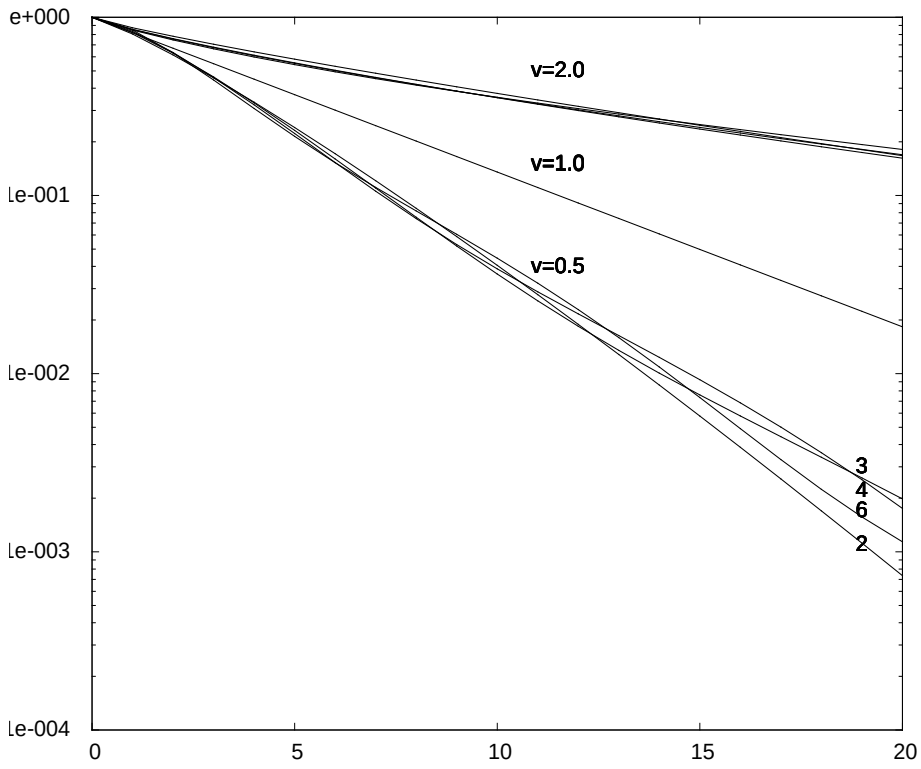

Рис. 2. Распределение числа заявок в системе *M*/*G*/1

Кривые строились с учетом выравнивания от одного до шести моментов распределения Вейбулла гамма-плотностью с поправочным многочленом. Выравнивание одного момента соответствовало подстановке экспоненты (v=1). Очевидны следующие выводы:

1. Разница в значениях ДФР на правой границе графиков достигает двух порядков. Это говорит о категорическом императиве учета для немарковских распределений, заметно отличающихся от показательного, минимум двух моментов.

2. Выравнивание двух моментов обеспечивает приемлемую точность в области значений ДФР, превышающих 10^{-2}. Большие относительные погрешности на «хвостах» распределений теоретически неизбежны при любом числе моментов.

3. Влияние моментов на искомую ДФР быстро убывает с увеличением их порядка. К тому же реальные моменты высокого порядка определяются с большими относительными погрешностями. Поэтому распределения уже с небольшим числом совпадающих моментов могут считаться взаимными аппроксимациями. Поскольку погрешность исходных (статистических) моментов с увеличением их порядка быстро возрастает, практически достаточно выравнивать 2–4 момента. Кроме того, реальные моменты высокого порядка определяются со значительной относительной погрешностью. В качестве аппроксимаций следует выбирать распределения, доставляющие какие-либо расчетные преимущества (в разных целях – фазовые, гамма-, Вейбулла).

Примерно так же выглядят графики распределения числа заявок в системе.

Одно из распределений – показательное с ДФР вида $\overline{B}(t) = e^{-\mu t}$ – обладает уникальным *марковским* свойством: распределение оставшейся длительности обслуживания *не зависит* от того, сколько времени оно уже продолжалось. Как установил основатель ТМО А. К. Эрланг, оно хорошо описывает длительность телефонных контактов (В. Феллер ехидно объясняет это особенностью дамских разговоров, создающих основную загрузку сети). Указанное свойство резко упрощает расчеты. По этой причине «исследователи» правдами и неправдами стараются аргументировать экспоненциальное распределение длительностей обслуживания, например: «поскольку разброс времени обслуживания относительно среднего невелик». Но ведь это распределение имеет *единичный*, т. е. весьма значительный, коэффициент вариации, так что на самом деле из вышеприведенного тезиса вытекает гипотеза о *регулярном* обслуживании!

Поскольку показательное распределение – однопараметрическое, оно позволяет выравнивать только первый момент (среднее значение случайной величины). Его выбор автоматически определяет и дисперсию, равную квадрату среднего.

Элементы модели ТМО

Модель теории очередей образуют**:**

1) поток заявок;

2) процесс обслуживания (число каналов, их однородность и специализация, распределения длительности обслуживания, и т. д.);

3) дисциплина обслуживания (FCFS, LCFS, случайный выбор из очереди, квантованное обслуживание, различные варианты приоритетов и др.);

4) показатели эффективности (они включаются в модель, поскольку выбор показателя может существенно повлиять на метод решения задачи).

Поток заявок обычно считается рекуррентным (с произвольным, но одинаковым распределением между смежными заявками) и описывается распределением интервалов между смежными заявками. Если оно подчинено показательному закону, мы имеем дело с простейшим потоком (Пуассона). Это предположение весьма часто выполняется на практике, чему имеются теоретические основания (теоремы о суммировании и прореживании потоков). Однако бывают и исключения.

Поток может состоять из *пачек заявок.* В ряде докладов (лечебно-эвакуационная работа медслужбы дивизии при нанесении удара по аэродрому, сборка персональных компьютеров) это обстоятельство игнорировалось; однако его учет *радикально* меняет результаты вычислений.

Типична также ошибка, когда с помощью аппарата для стационарных ситуаций обрабатываются *нестационарные* (та же медслужба; обращения граждан в учреждения – как правило, перед началом обслуживания уже имеется очередь, которая затем слабо пополняется со временем). В таких случаях полезной является диффузионная аппроксимация нестационарных процессов. Добавим к этому, что в обоих обсуждаемых случаях надо было моделировать не систему, а *сеть* обслуживания.

Каналы обслуживания, как правило, считаются тождественными, полнодоступными (неполнодоступными схемами давно и обстоятельно занимается телефония) и работающими независимо от других. Длительность обслуживания заявки задается функцией распределения *B(t).* Она в большинстве случаев заметно отличается от показательной, что делает применение моделей *M/M*/n практически бессмысленным. К сожалению, изложение ТМО в большинстве вузов этими моделями исчерпывается, что и объясняет (но не оправдывает!) их «принудительное» использование.

Под *дисциплиной обслуживания* главным образом понимается порядок выбора заявок из очереди. Типичной считается FCFS («первый пришел – первый обслужен»). Весьма характерен и случайный выбор из очереди (RANDOM), расчет которого сопряжен со значительными, но преодолимыми трудностями. Имеется возможность рассчитывать одноканальные системы со статическими (постоянными) приоритетами – без прерывания и с разными вариантами такового, а также с динамическим относительным приоритетом, возрастающим по времени ожидания.

Показатели эффективности систем массового обслуживания (СМО) классифицируются в двух смыслах: по точке зрения (хозяин или пользователь) и по типу значений (счетные и временные). *Пользователя* интересуют в основном реактивность системы (распределение времени ожидания и его моменты, средняя длина очереди), распределение полного времени пребывания, а также вероятность отказа в приеме на обслуживание (при ограниченной очереди). Для *хозяина* по экономическим соображениям желателен высокий *коэффициент загрузки системы* (калька с английского utilization для ρ в русском языке перегружена нежелательными ассоциациями). Он считается по формуле $\rho = \lambda b / n,$ где λ – интенсивность потока заявок, b – среднее время обслуживания, n – число каналов. Очевидна противоположность интересов сторон: чем выше коэффициент загрузки, тем меньше будут удовлетворены заказчики. С общесистемных позиций в показателе качества работы СМО должны учитываться интересы обеих сторон (см. задачу о восстанавливаемом ЗИПе [1]).

Очевидно, что для *систем с отказами* увеличение числа каналов монотонно снижает вероятность отказа. Однако приходилось слышать утверждение, что *при увеличении* числа модемов в пуле за 51 уровень потерь *стабилизируется* на 4%.

С обеих упомянутых точек зрения интересна *производительность* системы, естественно измеряемая средним числом обслуженных в единицу времени заявок. В стационарном (установившемся) режиме при неограниченном ожидании она равна среднему числу заявок, прибывающих в систему за то же время, т. е. *известной входной величине*. При заданном входящем потоке производительность системы приходится рассчитывать только для систем с потерями (ограниченном ожидании) и замкнутых систем с примитивным входящим потоком. Высказывалась, однако, точка зрения что «ресурс производительности многопроцессорной системы при бесконечном объеме памяти можно считать неисчерпаемым».

К сожалению, докладчики наших конференций порой путают производительность системы с ее *реактивностью* (скоростью реакции), измеряемой, например, средним временем ожидания начала обслуживания. Например, в связи с задачей управления подвижными объектами утверждалось, что «рост интенсивности потока приводит к

снижению производительности» – ухудшается реактивность, но не производительность! На самом деле последняя лишь *ограничивается* суммарным быстродействием каналов *n/b*, а среднее время ожидания стремится к бесконечности (реактивность – к нулю).

Простейшие методы (Эрланга) позволяют рассчитывать только марковские системы – с чисто показательными временными распределениями. Более сложные системы подвергаются *марковизации* методами линейчатых процессов, вложенных цепей Маркова, фазовых аппроксимаций или их комбинациями. Теоретическими основами численных методов расчета являются аппроксимация распределений по методу моментов (выравнивание некоторого числа последних) и *законы сохранения* заявок, стационарной очереди, вероятностей состояний и объема работы.

Законы сохранения СМО

Законы сохранения, в комплексе сформулированные для стационарных СМО в середине 1960-х гг. (М. Краковски), позволили резко упростить анализ сложных систем обслуживания. Из их кажущихся банальными словесных формулировок элементарно выводятся нетривиальные количественные следствия. Рассмотрим, к примеру, закон сохранения объема работы: *в классе дисциплин, реализация которых не создает системе дополнительной работы, объем находящейся в системе работы постоянен и не зависит от конкретной дисциплины.* Для одноканальной системы среднее время w ожидания начала обслуживания вновь прибывшей заявкой как раз и равно среднему объему (длительности выполнения) накопленной работы. Этот объем состоит из среднего остатка текущего обслуживания $b_2/2b_1$, умноженного на вероятность занятости системы $\rho = \lambda b_1$, и средней длительности обслуживания скопившейся очереди. Средняя длина последней $q = \lambda w$ (*формула Литтла*, вытекающая из закона сохранения очереди), причем обслуживание каждой заявки в среднем требует b_1 единиц времени. Итак,

$$w = \lambda b_2/2 + \lambda w b_1,$$

откуда следует формула (Полячека–Хинчина)

$$w = \frac{\lambda b_2}{2(1-\lambda b_1)} = \frac{\lambda(b_1^2 + D)}{2(1-\lambda b_1)}.$$

Уже из этой формулы и формулы Литтла можно вывести ряд весьма поучительных следствий:

- *Ожидаемая* длина очереди отлична от нуля при потоке сколь угодно малой интенсивности. Между тем, в «военно-медицинском» докладе утверждалось, что «очередь образуется при интенсивности потока 13 раненых в час и более».
- Коэффициент загрузки $\rho = \lambda b_1/n$ должен быть *строго меньше* единицы, поскольку при $\rho \to \infty$ среднее время ожидания стремится к бесконечности (это верно для любых систем обслуживания). Проверка этого условия должна *предшествовать* любым попыткам моделирования систем и сетей (наиболее очевидный пример сочетания аналитического подхода и имитации). Однако все еще встречаются попытки планировать максимальную производительность системы обслуживания *равной* интенсивности входящего потока или доказывать на имитационных моделях, что сильно загруженные системы очень чувствительны к увеличению нагрузки. Попутно отметим, что для таких систем быстро возрастает *дисперсия результатов* имитационного моделирования.

- Для показательно распределенного обслуживания дисперсия $D = b_1^2$, и среднее время ожидания при прочих равных условиях будет *вдвое* больше, чем при регулярном обслуживании. Это говорит о недопустимости огульного применения марковских моделей. Приходилось слышать мнение, что «по формуле Полячека–Хинчина мы можем иметь расхождение системных характеристик по сравнению с традиционными результатами (подразумевается система *M*/*M*/1) ровно в два раза, но не более». Фактически же для распределений с коэффициентом вариации, *большим единицы,* разница *может быть сколь угодно велика*.

Два последних тезиса количественно подтверждают «философски» ожидаемый результат: увеличение неопределенности в любом из элементов модели ухудшает качество обслуживания. Поэтому установление на модели преимущества в этом смысле регулярного потока перед пуассоновым трудно считать вкладом в науку.

Численные методы теории очередей

Если понимать под аналитическими методами ТМО совокупность формул, допускающих ручной счет, то они исчерпываются приведенными выше формулами Эрланга и Полячека–Хинчина.

Есть мнение (одного из докладчиков ИММОД-2005), что «возможности аналитических методов ТМО ограничены обязательно экспоненциальным распределением наработки на отказ и восстановления работоспособности…Она учитывает (и дает) только средние значения показателей». Было заявлено, что аналитическое решение задачи Эрланга (система с отказами) непомерно трудно и требует «программного обеспечения, обрабатывающего большие числа». Неужели 10^{4932} (расширенный формат для чисел при работе ПК с двойной точностью) будет недостаточно? К тому же известны простые *асимптотические* зависимости.

Продвинутые разделы классической ТМО позволяют рассчитывать модели с *одной* немарковской составляющей (типа $GI/E_k/1$) методами вложенных цепей Маркова, но требуют обязательного применения ЭВМ и нетривиального программирования. Наиболее общая схема *численного* расчета системы ориентирована на фазовые аппроксимации немарковских распределений и представляется в следующем виде:

- выбор фазовых аппроксимаций и определение пространства состояний;
- расчет параметров аппроксимаций и разметка графа состояний;
- построение (желательно, автоматическое) матриц интенсивностей переходов между состояниями смежных ярусов диаграммы;
- расчет вероятностей микросостояний итерационным решением уравнений баланса;
- переход к укрупненным вероятностям;
- расчет моментов распределения времени ожидания;
- расчет моментов распределения времени *пребывания* заявки в системе и построение по ней ДФР;
- расчет распределения интервалов между заявками выходящего потока (для звеньев сети).

Сети обслуживания (разомкнутые, замкнутые и смешанные) рассчитываются методом их потокоэквивалентной декомпозиции:

- из уравнений необходимого баланса потоков между узлами находятся средние интенсивности потоков, входящих в каждый узел;
- каждый узел рассчитывается как изолированная СМО для суммарного потока на ее входе; вычисляются моменты распределения времени пребывания заявок в них при однократном посещении;

- вычисляется преобразование Лапласа–Стилтьеса (ПЛС) распределения времени пребывания заявки в сети и численным дифференцированием ПЛС – моменты распределения;
- строится аппроксимация ДФР времени пребывания заявки в сети.

При наличии циклических маршрутов требуется итерационный пересчет потоков. Все эти методы отчасти описаны в работе [1] и полностью – в пока не изданной монографии автора «Компьютерный расчет систем и сетей с очередями».

К сожалению, редко кто осведомлен о возможностях *современных* численных методов теории очередей. Молодые специалисты по компьютерным сетям (**Кульгин М.** Теория очередей и расчет параметров сети//BYTE – Россия, ноябрь 1999. С. 26–33) убеждены в том, что информационные технологии вполне заменяют знание основ ТМО. Автор вышеупомянутой статьи полагает, что «аналитик может провести анализ очередей в заданной сетевой структуре, используя уже готовые таблицы очередей или простые компьютерные программы, которые занимают несколько строк кода», и далее демонстрирует математическую безграмотность и полное непонимание проблемы.

Для длительности обслуживания им предлагается «закон интервалов, или экспоненциальный закон». Здесь родовое понятие отождествляется с его частным случаем. Формула $\Pr(\Theta \ge \theta) = 100\exp(-\mu\theta)$ из-за невесть откуда появившейся сотни может давать значения вероятностей до 100 включительно. Далее утверждается, что μ – уровень обслуживания, в данном случае равный коэффициенту загрузки ρ. Поскольку ρ – величина безразмерная, показатель степени приобретает размерность времени – совершенно вопиющий абсурд. На самом деле μ есть интенсивность обслуживания (величина, обратная его средней длительности).

Утверждается далее, что «если существует среда, в которой есть разделяемые каналы связи, то производительность такой системы обычно изменяется по экспоненциальному закону при увеличении нагрузки… При дальнейшей загрузке системы ее производительность будет резко снижаться». На самом деле производительность системы измеряется не временем ответа, а числом заявок, обслуженных в единицу времени. Она равна интенсивности входящего потока, следовательно, с ростом коэффициента загрузки ρ будет *возрастать*, пока он не достигнет единицы. После этого производительность системы будет постоянна и равна ее максимальной производительности. Экспоненциальному же закону при докритическом режиме подчиняется не производительность системы, а *распределение времени пребывания заявки* в ней. Среднее время ответа при простейших допущениях меняется по *гиперболическому* закону с вертикальной асимптотой в точке ρ=1. Наконец, в статье, вопреки ее названию, дело до заявленного расчета *сетей обслуживания* так и не дошло.

Продвинутая теория очередей (см. для начала университетский курс [2]) накопила богатейший арсенал методов, реализация которых требует, однако, хорошей математической и программистской подготовки (отметим, что в [2] нет ни примеров программ, ни результатов счета). Для примера сошлемся на одну из самых простых задач этого класса – расчет системы *GI/M*/1. Доказано, что распределение времени пребывания заявки в ней – показательное с параметром $\mu(1-\omega)$, где μ – интенсивность обслуживания, а ω – корень уравнения

$$\omega = \int_0^\infty e^{-\mu(1-\omega)t} dA(t).$$

Это уравнение приходится решать методом итераций. Практически необходимо выбрать такое представление плотности распределения интервалов между смежными

заявками, чтобы интеграл в правой части последнего равенства выражался явно (плотность должна быть показательно-степенной функцией). Кроме того, нужно задать хорошее начальное приближение и, возможно, позаботиться об ускорении сходимости итераций. Очевидно, что для решения таких задач требуется владение основами математического анализа, вычислительной математики и теории вероятностей. Утверждалось, однако, что для аналитического моделирования требуется лишь знание Excel. Еще более крайнюю позицию занимают некоторые «икономисты» (так в тексте их доклада): «определение соответствия схем и характеристик теории СМО разновидностям производственно-экономических систем не требует особого ума и труда не составляет».

Из структуры уравнения ясно, что ω определяется *всеми* существующими моментами распределения *A(t)*. Не было никакой необходимости доказывать *на модели,* что среднее время пребывания заявки в системе с приоритетами зависит от высших моментов распределения *A(t)* интервалов между заявками рекуррентного потока, в том числе от третьего.

В некоторых докладах проводилось исследование простых систем массового обслуживания «для новых видов распределений» – например, Вейбулла и Парето – с особыми свойствами («толстыми хвостами»). Общие схемы и примеры таких расчетов известны, а «толстый хвост» – просто признак распределения с коэффициентом вариации, превышающим единицу. Уж если заниматься такими исследованиями, то следует определять влияние разницы в высших моментах *при равных первых двух*. Можно наперед сказать (см. рисунки, приведенные в начале доклада), что этот эффект будет весьма мал.

В связи с изложенным нет никакой необходимости внедрения GPSS/W в курс теории телетрафика «для изучения характеристик потоков вызовов с такими распределениями, как Вейбулла, Пирсона, логнормальное». Перечисленные задачи легко и намного точнее решаются *численными* методами теории очередей.

Ряд докладчиков считал проблемой хорошо изученный расчет приоритетных режимов для одноканальных систем. Между тем для многих случаев известны [2] алгоритмы расчета ПЛС распределения времени ожидания и в некоторых случаях – высших моментов этого распределения.

Некоторые полагают, что введение приоритетов улучшает все показатели обслуживания. На самом деле введение приоритетов лишь *перераспределяет* имеющийся ресурс – в пользу более приоритетных заявок и *за счет* менее приоритетных. Для консервативных дисциплин это перераспределение управляется инвариантом $\sum_i \rho_i w_i = \text{const}$ (Л. Клейнрок). Недавние расчеты автора показали, что для приоритетов с прерыванием в случае распределений времени обслуживания, отличных от показательного, этот инвариант нарушается, но лишь на десятые доли процента.

Пакеты программ для расчета систем с очередями

Сложность решения достаточно содержательных задач ТМО сравнительно рано сделала их объектом усилий программистов. Одно из первых упоминаний (1966г) о результатах таковых описывает программу, позволяющую рассчитывать итерационным методом марковские СМО с числом состояний до 5000.

Бурный рост числа публикаций по *пакетам* программ для задач ТМО отмечается с конца 1970-х гг. Впечатляет география появления уже первых пакетов: США, Пуэрто-Рико, Европа, Израиль, Япония и ЮАР. Очерк истории создания зарубежных пакетов дается в обзорной статье Сойера и Мак-Нейра [3]. В сборнике [4] приводится описание ряда более поздних разработок. К ним следует добавить пакеты, выполнен-

ные под руководством Ю. И. Митрофанова и В. М. Вишневского, а также комплекс программ, создание которых координировалось В. Ф. Матвеевым (МГУ).

Сопоставление названных пакетов затруднительно, поскольку опубликованные сведения о них неполны и часто имеют рекламный характер, а техническая документация широкой общественности не доступна. Ниже дано описание возможностей и технологии применения разработанного автором пакета МОСТ по состоянию на май 2007 г.

Пакет МОСТ (Массовое Обслуживание – СТационарные задачи) в качестве теоретических основ использует:

- аппроксимацию распределений по методу моментов;
- законы сохранения теории очередей;
- потокоэквивалентную декомпозицию сетей обслуживания.

В настоящее время пакет записан на Фортране для персональных ЭВМ (Windows 98, Фортран версии MS PowerStation 1.0), насчитывает около 140 процедур и эксплуатируется в трех вариантах:

- «чайный» – для убежденных непрофессионалов с ограниченными возможностями. Здесь система в процессе диалога с пользователем автоматически формирует и запускает ведущую Фортран-программу;
- профессиональный – с вышеописанной логикой использования и полным спектром возможностей;
- учебный – технологический аналог профессионального, но с усеченными возможностями.

Все версии пакета будут обеспечены описанием его теоретических основ (уже упоминавшаяся монография автора «Компьютерный расчет систем и сетей с очередями» – 400 с.), усеченной версией последнего для «чайников» и Руководствами к лабораторным работам (учебная версия – 4, профессиональная – 9 работ).

Пакет МОСТ *в умелых руках* позволяет рассчитывать одно- и многоканальные системы и сети обслуживания с произвольными (заданными своими моментами) распределениями интервалов между входящими заявками и длительностей обслуживания.

Чему учит теория очередей

Перечислим на *качественном* уровне выводы, которые следуют из современной теории очередей и должны быть приняты на вооружение всеми «имитаторами».

1. Современное состояние теории очередей позволяет решать широкий круг задач по расчету систем обслуживания – в том числе многоканальных, с немарковскими входящими потоками и распределениями обслуживания, с отличными от FCFS дисциплинами выборки из очереди, а также по расчету *сетей* обслуживания.
2. Теория очередей «не требует большого ума и знания математики» лишь в самых элементарных случаях, имеющих чисто учебную ценность, да и то для «ограниченного контингента». *Реальной замены знанию не существует.*
3. При сохранении средних любое увеличение неопределенности (дисперсии интервалов между заявками, длительности обслуживания, объема пачки) *ухудшает* характеристики обслуживания, особенно для сильно загруженных систем. Наиболее наглядно это следует из формулы Полячека–Хинчина.
4. Для существования стационарного режима коэффициент загрузки системы (и каждого узла сети) должен быть *строго меньше единицы.* Производительность (она же – пропускная способность системы или узла сети) ограничена суммарным быстродействием каналов n/b.

5. Влияние высших моментов на показатели обслуживания увеличивается с ростом коэффициента загрузки и быстро убывает по числу учтенных моментов. Среднее время пребывания в заявки в системе *M/G*/1 определяется двумя моментами распределения обслуживания, а в *GI/M*/1 – *всеми* моментами распределения интервалов между смежными заявками (переоткрывать этот факт имитацией не было необходимости). Учет только средних (чисто показательные аппроксимации, т. е. работа с формулами Эрланга) явно недостаточен; необходимо выравнивать 2–3 момента.
6. Дробление суммарной производительности многоканальных систем приводит к уменьшению среднего времени ожидания заявки, но увеличивает среднее время пребывания. При пропорциональном увеличении интенсивностей входящего потока и обслуживания средние времена ожидания и пребывания в системе уменьшаются во столько же раз (этот неожиданный факт для одноканальной марковской системы доказывается элементарно).
7. Введение приоритетов лишь *перераспределяет* ресурс системы в пользу более приоритетных заявок за счет менее приоритетных, причем *относительный* выигрыш во времени ожидания для первых существенно превосходит проигрыш для вторых.
8. «Классическая» теория сетей обслуживания из-за слишком стеснительных условий ее применимости («теорема BCMP») практически бесполезна. Имитационное моделирование подтвердило приемлемую точность методов потокоэквивалентной декомпозиции сети. Любой форме расчета или имитации сети должно предшествовать выявление ее *узких мест* – наиболее загруженных узлов – и их «расшивка».
9. Расчет достаточно адекватных реальности систем и сетей обслуживания практически должен (и может) опираться на разработанные специалистами пакеты прикладных программ (ППП). Сознательное их применение требует если не знания, то по крайней мере понимания сути вышеупомянутых методов. Такие расчеты свободны от известных недостатков ИМ и могут служить как для непосредственного решения практических задач, так и для предварительного обсчета моделируемых систем и сетей и тестирования моделей.
10. Методы ИМ:
    а) имеют ограниченную точность (хотя бы из-за неидеальности равномерных ДСЧ) – см. пример с расчетом числа π в пленарном докладе Ю. И. Рыжикова, Б. В. Соколова и Р. М. Юсупова;
    б) требуют большого числа испытаний для определения вероятностей редких событий и расчета сильно загруженных систем (дисперсия времени ожидания в M/M/1 для нарастающих ρ: 0.7 – 10.11; 0.8 – 24; 0.9 – 99; 0.95 – 399);
    в) практически непригодны для оптимизации систем и сетей (малые изменения целевой функции маскируются статистическими флюктуациями).
11. Методами ТМО неудобно решать задачи многоресурсные, с расщеплением и слиянием заявок, с циклическими временными режимами и т. п. Здесь ИМ незаменимо.

Вытекающий отсюда *аргументированный* вывод – «каждому (применению) свое» – не нов, но предполагает значительное смещение интересов «имитаторов» и их подготовки в сторону численных методов ТМО. Овладение основами этих методов вполне по силам студенту-второкурснику и инженеру со стандартной для втуза математической подготовкой; их практическое использование требует соответствующих пакетов программ.

Литература

1. **Рыжиков Ю. И.** Теория очередей и управление запасами. СПб.: Питер, 2001. 376 с.
2. **Климов Г. П.** Стохастические системы обслуживания. М.: Наука, 1966. 243 с.
3. **Sauer C. H., McNair E. A.** The Evolution of the Research Queuing Package//Modelling Technique and Tools for Performance Analysis'85. Proc. of the Internat. Conf. Amsterdam: North-Holland Publ. Co., 1986. 365 pp.
4. Modelling Technique and Performance Evaluations//Proc. of the International Workshop. Eds. S. Fdida, G. Pujolle. Amsterdam: North-Holland Publ. Co., 1987. 340 pp.