Аналитическая модель приоритетной системы массового обслуживания релейного типа

С.Ш. Кутбитдинов (ГУП «UNICON.UZ»), В.В. Лохмотко (СПбГУТ)

В данной статье предлагаются формулы оценки среднего времени пребывания пакета в однолинейных системах массового обслуживания с абсолютным или относительным приоритетом в режиме «включен–выключен».

Ушбу мақолада пакетнинг «уланган – узиб қўйилган» режимида абсолют ёки нисбий устуворликка эга бир линияли оммавий хизмат кўрсатиш тизимларида бўлиш ўртача вақтини баҳолаш учун формулалар таклиф этилади.

IPTD formulas for one-linear systems with exponential priority-service discipline in on/off mode are offered in this article.

Введение. Для подготовки бизнес–планов и инвестиционных проектов создания и развития современных инфокоммуникационных сетей на платформе IP необходимы компактные аналитические модели оценки значений нормируемых показателей качества обслуживания QoS [1], адекватно представляющие реальные процессы передачи и обработки сетевого трафика и учитывающие влияние на его свойства возможных воздействий ряда факторов (всплески сетевой нагрузки, программные прерывания и сбои, технологические перерывы, отказы сетевого оборудования и т.п.), снижающих эффективно используемую пропускную способность линий связи и QoS [2–4]. Попытки как-то приблизить результаты расчета значений QoS-показателей, полученные в условиях идеальной среды передачи, к реальной действительности, переориентировали инструментарий моделирования систем массового обслуживания (СМО) с очередями с традиционных пуассоновских и марковских моделей в сторону «тяжелохвостных» распределений [5]. Однако авторы данной статьи считают, что возможности экспоненциальных моделей еще далеко не исчерпаны и демонстрируют это на примере сервера, как типового элемента инфокоммуникационной сети.

Построение экспоненциальной модели. Рассматривается сервер, как система массового обслуживания пакетов *P* пользовательских приложений в режиме ON/OFF («включен-выключен»). При этом работа источников нагрузки предполагается непрерывной, периоды нормальной работы на ON–интервалах чередуются с периодическими блокировками на OFF–интервалах (рис.1), сопровождающимися простоями со средней продолжительностью T_{off} и вероятностью ρ_{off} появления OFF–интервала (ρ_{off} + ρ_{on} = 1). Формулируется задача построения модели вида $T_p\left(F, \lambda_p, \overline{h}_{pn}\right)$, функционально связывающей среднее время пребывания пакета *p*–го типа (в частном случае *p*–го приоритета) с *F*–параметром (раскрывается ниже) ON/OFF–режима, интенсивностью посту-

пления λ_p пакетов *p*–го типа и моментами до *n*–го порядка включительно времени обслуживания $\overline{h}_{pn}$, $p = \overline{1,P}$.

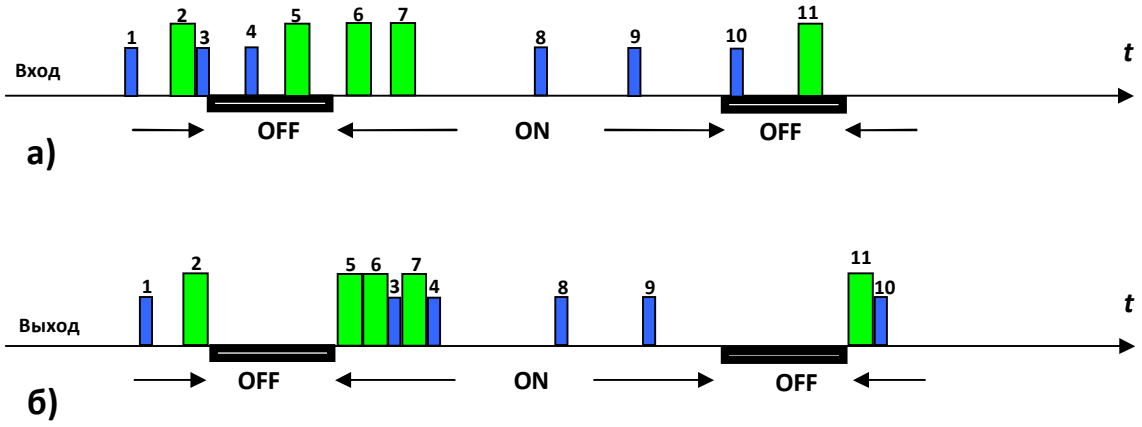


Рис.1. Временная диаграмма режима «включен-выключен»

Диаграмма отображает процесс поступления (а) и завершения (б) обслуживания пакетов с учетом «торможения» со стороны OFF–интервала (пакетам 2, 5, 6, 7 и 11 – предоставлен приоритет).

Предполагается стационарность, статистическая независимость и структурированность сетевых процессов (сессий, транзакций), позволяющих описывать пользовательские приложения тремя параметрами: интенсивностью поступления λ_p, средним объемом пакета V_p (или средней продолжительностью передачи) и нормами t_{zp} на среднюю задержку пакета *p*–го пользовательского приложения, $p = \overline{1,P}$.

Исследование проводится при следующих модельных допущениях и ограничениях:

– отсутствие помех и потерь в каналах, время распространения сигнала не учитывается;

– структура буферной памяти обслуживающего прибора СМО не детализируется, а ее емкость – неограниченна;

– многомерные пуассоновские входящие потоки;

– экспоненциальный закон распределения всех случайных величин, в том числе и длительности ON и OFF–интервалов;

– приоритетное обслуживание пакетов с предоставлением OFF–интервалам старшего (фиктивного, *p* = 0) приоритета;

– обслуживание пакетов внутри *p*–го приоритетного класса осуществляется в соответствии с дисциплиной FIFO;

– механизм подтверждения правильности приема пакета не учитывается;

– стохастические процессы рассматриваются на временном отрезке до наступления тайм-аута и последующей блокировки источников нагрузки;

– возможность возникновения «пачечности» пакетов типа 5,6,3,7,4 (рис.1) и «наслоения» OFF–интервалов (например, сбоев на технологические перерывы).

В качестве отправной точки выбрана классическая *P*–приоритетная однолинейная СМО типа $M_p/\mathrm{G}_p/1/\infty$ [2] с многомерным пуассоновским входящим потоком, для которой среднее время пребывания пакета *p*–го приоритетного класса T_p определяется как:

$$T_p = \frac{1/\mu_p(1-\sigma_p)\chi + S}{(1-\sigma_p)(1-\sigma_{p-1})}, \qquad p = \overline{1,P}, \tag{1}$$

где: $\sigma_p = \sum_{i=1}^{p}\rho_i$ – суммарная загрузка исследуемой СМО потоками пакетов различных классов до *p*–го класса включительно;

$\rho_i = \lambda_i/\mu_i$ – загрузка СМО потоком пакетов i–го класса для случая индексации потоков по переменной i, $i = \overline{1,p}$; причем $\rho_{i-1} = 0$ при i = 1;

$1/\mu_i$ – среднее время передачи (обработки) пакета *i*–го класса;

$S = \sum_{i=1}^{\nu}\lambda_i h_{i2}/2$ – среднее остаточное время обслуживания пакетов, h_{i2} – второй начальный момент времени передачи (обработки) пакета i–го класса.

Переменные χ (неполная сумма) и ν (верхний предел суммирования) соответствуют следующим вариантам приоритетного обслуживания: при $\chi = 1-\sigma_{p-1}$ и $\nu = P$ – дисциплине без прерывания обслуживания (относительный приоритет); при $\chi = 1$ и $\nu = p$ – дисциплине с прерыванием и дообслуживанием (абсолютный приоритет).

При выборе законов распределения случайных величин в формуле (1) предпочтение отдается экспоненциальному распределению, отличающемуся компактностью записи, гибкостью, воспроизводимостью, простотой машинной обработки и многовариантностью в применении. Если к экспоненциально распределенному времени обслуживания добавить какое-то постоянное время, то результат будет следовать гамма–распределению [4]. Гиперэкспоненциальные и гипоэкспоненциальные аппроксимации реальных распределений времени обслуживания также базируются на экспоненте.

Для экспоненциального закона распределения времени обработки ($h_{i2} = 2/\mu_i^2$) информационных запросов и длительности ON и OFF–интервалов при условии предоставления им старшего (фиктивного) приоритета ($\rho_0 = \rho_{off}$, $1/\mu_0 = T_{off}$) модель (1) превращается в модель однолинейной СМО типа $M_p/M_p/\bar{1}/\infty$, для которой IPTD с обслуживанием по абсолютному приоритету представляется выражением

$$T_p^{абс} = \frac{1/\mu_p(\rho_{on}-\rho_1-\ldots-\rho_p) + \rho_{off}T_{off} + \sum_{i=1}^{p}\frac{\rho_i}{\mu_i}}{(\rho_{on}-\rho_1-\ldots-\rho_{p-1})(\rho_{on}-\rho_1-\ldots-\rho_p)} \tag{2}$$

и, соответственно, по относительному приоритету

$$T_p^{отн} = \frac{1}{\mu_p} + \frac{\rho_{off}T_{off} + \sum_{i=1}^{P}\frac{\rho_i}{\mu_i}}{(\rho_{on}-\rho_1-\ldots-\rho_{p-1})(\rho_{on}-\rho_1-\ldots-\rho_p)}, \quad p = \overline{1,P}\ . \tag{3}$$

Среднее время ожидания пакета *p*–го приоритета в очереди, в терминах, приведенных [3], определяется как $W_p = T_p - \frac{1}{\rho_{on}\mu_p}$, а математическое ожидание числа пакетов *p*–го приоритета в очереди [2], соответственно как $N_p^Q = \lambda_p W_p$, $p = \overline{1,P}$.

В частном случае (для $P = 1$) средняя задержка пакета в однолинейной СМО типа $M/M/1/\infty$ с прерыванием обслуживания будет определяться как

$$T^{абс} = \frac{1+F\mu_э}{\mu_э - \lambda} = T^{отн} + \rho_{off}/\mu_э \quad , \qquad \mu_э > \lambda , \tag{4}$$

где: $F = \rho_{off} T_{off} (1-\rho_{off})^{-2}$ – параметр, определяемый скважностью ON/OFF–процесса (соотношением между средней длительностью периода «включен» и средней длительностью «выключен») и приблизительно равный остаточному времени [2] завершения текущего OFF–интервала. Численные значения средней продолжительности T_{off} и частости ρ_{off} могут быть получены путем обработки [6] статистики по прерываниям или анализа данных по прототипам (каналам, серверам или маршрутизаторам). В ряде случаев может оказаться достаточной приближенная оценка параметра *F* в следующем виде $F \approx \rho_{off} T_{off}$;

$\mu_э$ – эффективная интенсивность обслуживания [3] обслуживающего прибора СМО, в зависимости от контекста представляемая долей либо интенсивности обслуживания μ (пакетов/с), либо битовой скорости $C = \mu V$;

$$\mu_э = \mu\rho_{on} = \mu - \mu\rho_{off} = \rho_{on} C/V = (1-\rho_{off})C/V$$

Независимо от типа приоритета релейным формулам (2–4) свойственны:

– непроизводительные потери $\mu\rho_{off}$ пропускной способности обслуживающего прибора, обуславливаемые присутствием OFF–периодов неготовности;

– увеличенный по сравнению с идеальными условиями временной масштаб $\overline{\alpha} = T_p / T_{p(F=0)}$, $p = \overline{1,P}$, определяемый отношением задержки пакета в режиме ON/OFF к аналогичной задержке при $F = 0$, в частности, при $P = 1$

$$\alpha = T^{абс} / T^{абс}_{F=0} \approx 1 + \mu\rho_{off} T_{off}, \tag{5}$$

– бо′льшие, по сравнению с идеальными условиями, размеры очередей из ожидающих пакетов, оцениваемые отношением $\overline{\gamma} = N_p^Q / N_{p(F=0)}^Q$, $p = \overline{1,P}$ средней длины очереди в ON/OFF–режиме к аналогичному показателю при $F = 0$;

– момент насыщения и перегрузки $\rho^* = \rho_{on}$, который наступает раньше, чем в идеализированных системах $\rho^* < \rho = \lambda / \mu$. Порог ρ^* отделяет область, в

которой для усеченных распределений математическое ожидание и дисперсия не существуют (нестационарный режим) [6], а расчетные значения IPTD, полученные вблизи точки ($\rho^* - \delta$, $\delta \to 0$) не устойчивы к перегрузкам и практического интереса не представляют;

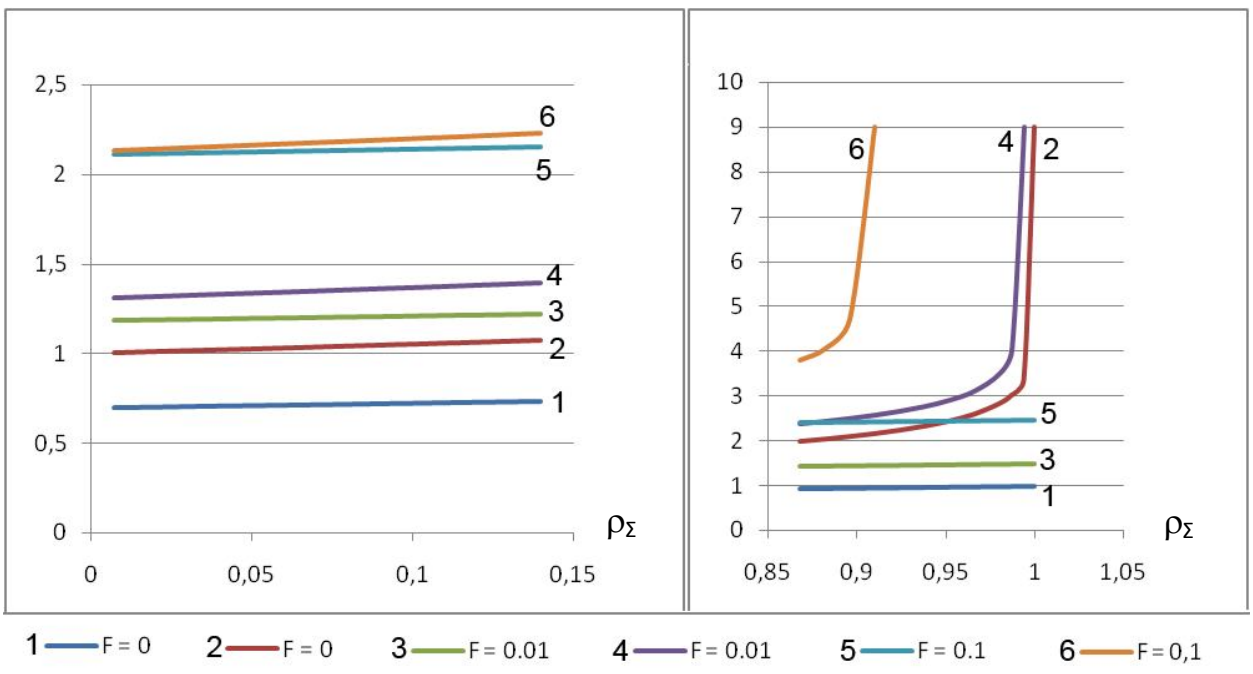
– определяемые скважностью ON/OFF–процесса асимптоты:

$$\lim_{\lambda \to 0} T^{абс} = \lim_{\lambda \to 0} T^{отн} = F + \mu_э^{-1}, \qquad \lim_{\lambda_\Sigma \to 0, \mu \to \infty} T_p = F, \quad p = \overline{1,P}, \tag{6}$$

прямо указывающие на причину снижения качества обслуживания, производительности и быстродействия СМО $M_p/M_p/\bar{1}/\infty$ по сравнению с идеализированной системой $M_p/M_p/1/\infty$ (при $F = 0$);

– вырождение в известные модели СМО типа $M_p/M_p/1/\infty$ и $M/M/1/\infty$ [2] при $F = 0$.

Численные оценки задержки пакета в двухприоритетной системе «клиент–сервер», рассчитанные по формуле (2) для трех значений параметра F (F = 0,01; 0,1 и 0), приведены на рисунках 2а и 2б, где по оси абсцисс отложена суммарная загрузка $\rho_\Sigma = \rho_1 + \rho_2 + \rho_{off}$ сервера с учетом интервалов его неготовности, а по оси ординат в логарифмическом масштабе – задержки T_p пакетов для заданных скорости канала 128 кбит/с и соотношения объемов пакета V_1:V_2 = 1: 2.



а) б)

Рис. 2. Средняя задержка T_p пакета в двухприоритетной системе в зависимости от суммарной загрузки ρ_Σ и параметра F (а – диапазон малых нагрузок; б – область насыщения)

Полученные графики показывают, что задержки T_p пакетов 1–го и 2–го приоритетов (кривые 1, 3, 5 соответствуют старшему приоритету; кривые 2, 4, 6 соответствуют младшему приоритету), рассчитанные для ON/OFF–режима, могут в несколько раз превышать аналогичные для идеального случая (кривые 1 и 2 соответствуют параметру F = 0), а выполнение жестких норм на IPTD, например, $t_z \le$ 100мс осуществимо не всегда (верхние кривые расположены выше отметки $\log_{10}$(100мс) = 2).

Практические рекомендации. Результаты проведенных исследований показывают, что программные прерывания и сбои, а еще хуже продолжительные технологические перерывы и отказы сетевого оборудования, принципиально изменяют стохастическую структуру сетевых процессов, протекающих в системе «клиент-сервер» в режиме «включен–выключен», и ухудшают показатели эффективности:

– осредненное по множеству реализаций среднее время доставки пакета увеличивается в α ($\alpha > 1$) раз;

– среднее время пребывания пакета в очереди увеличивается в $\gamma(\gamma > 1)$ раз, что предрасполагает к появлению длинных серий пакетов;

– эффективно используемая пропускная способность каналов и производительность серверов снижается на $100\rho_{off}$ %.

Параметры α и γ определяются расчетным путем с помощью (2)–(4), ρ_{off} задается исходными данными. Для идеализированного случая ρ_{off} = 0; α = 1; γ = 1.

Из (5) видно, что негативное влияние OFF–интервалов неготовности на быстродействие системы «клиент–сервер» можно ослабить резервированием и реконфигурированием системы, а также выбором процессора большей производительности μ. Например, переход к кластерной структуре, образованной двумя серверами с общим дисковым массивом, позволит приложению в случае отказа (сбоя) одного из серверов мигрировать на другой и, тем самым, сократить среднее время восстановления и, соответственно, коэффициент простоя ρ_{off} .

При резервировании пропускной способности IP–сети, расчете настроечных параметров алгоритмов–планировщиков доступа к сетевым ресурсам и нормировании QoS–параметров дифференцированного по классам обслуживания трафика должны учитываться следующие особенности ON/OFF–режима:

– QoS–нормы, выполняемые в идеальных условиях, могут оказаться нарушенными в реальных условиях эксплуатации по причине реальных сбоев, прерываний и отказов аппаратного и программного обеспечения;

– для предотвращения перегрузок суммарная загрузка ρ_Σ системы «клиент–сервер» приоритетными потоками не должна превышать показателя готовности ρ_{on};

– чрезмерно жесткие нормы t_{zp} на величину задержки IPTD T_p пакета останутся не выполненными, если они заданы за пределами области работоспособности модели, т.е. $t_{zp} < F$;

– устойчивость системы «клиент–сервер» к перегрузкам определяется отношением норматива t_{zp} к нижнему порогу F минимально достижимой задержки и увеличивается при $(F - t_{zp}) \to \max$;

– чувствительность системы «клиент–сервер» к нарушению нормативов на IPTD определяется k–м пользовательским приложением, для которого запас

времени между нормативом и задержкой пакета T_k в рабочей точке $\min_k (t_{zk} - T_k),\ k = \overline{1, P}$ минимален.

По аналогии с [2] на базе (2) и (3) в дальнейшем может быть синтезирована модель сети массового обслуживания, предназначенная для расчета и оптимизации QoS–параметров сетей пакетной коммутации в архитектуре IMS (IP Multimedia Subsystem) с ненадежными обслуживающими приборами (каналами, серверами, маршрутизаторами и т.п.).

Литература

1. ITU-T Recommendation Y.1541. Global information infrastructure, internet protocol aspects and next – generation networks. Internet protocol aspects – Quality of service and network performance.

2. Клейнрок Л. Вычислительные системы с очередями.– М.: Мир.1979.– 600с.

3. Захаров Г.П. Методы исследования сетей передачи данных.– М.:Радио и связь, 1982. – 208с.

4. Дж. Мартин. Системный анализ передачи данных. Том II.– М.:Мир.1975. – 431с.

5. Городецкий А.Я., Заборовский В.С. Информатика. Фрактальные процессы в компьютерных сетях.: Учеб. пособие СПб.: Изд-во СПбГТУ, 2000. – 102 с.

6. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.:Финансы и статистика, 1983. – 471 с.