МЕТОДЫ УСКОРЕННОЙ ИМИТАЦИИ ПРОЦЕССОВ С ИНТЕНСИВНЫМИ ПРЕРЫВАНИЯМИ

В. Н. Задорожный (Омск)

1. Проблема моделирования интенсивных прерываний

При непосредственном имитационном моделировании таких систем массового обслуживания (СМО), в которых интенсивность λ потока приоритетных заявок на несколько порядков превосходит интенсивность λ' потока неприоритетных (рядовых) заявок, затраты компьютерного времени могут возрастать также на несколько порядков. Это связано с тем, что для получения представительной статистической выборки по обслуживанию N рядовых заявок приходится воспроизводить прохождение приблизительно $(\lambda/\lambda')N$ приоритетных заявок, т. е. общее число имитируемых событий оказывается на несколько порядков выше, чем число событий наблюдаемых.

Подобная проблема возникает, например, при имитации вычислительных систем, в которых управляющие программы ненадолго, но часто прерывают выполнение прикладных программ, создавая некоторую дополнительную загрузку ρ центрального процессора. Обычно при имитации ВС проблему обходят путем исключения из модели потока приоритетных заявок и учета их влияния «в среднем». Например, в [1] предлагается исключаемый поток прерываний компенсировать занижением быстродействия моделируемого процессора на долю, равную ρ. Однако при таком учете времени прерываний теряется его дисперсия, вклад которой в среднее время ожидания рядовых заявок может быть существенным, когда показатель $\alpha = \lambda/\lambda'$ составляет $10^1 \div 10^3$.

Исследование интенсивно прерываемых процессов обслуживания в СМО типа $G_2|G_2|1$ с абсолютными приоритетами позволило установить ряд общих аналитических результатов. В частности, найден достаточно простой и эффективный метод такого исключения из СМО потока приоритетных заявок, при котором время прерываний учитывается с точностью до распределения. Точная аналитическая форма полученных результатов достигается с помощью асимптотического анализа свойств системы $G_2|G_2|1$ при $\alpha \to \infty$. Увеличение α интерпретируется здесь как масштабное преобразование кумулятивных распределений $A(t)$ и $B(t)$, задающих время поступления и обслуживания приоритетных заявок в исходной системе. Эти распределения рассматриваются как заданные параметрически через базовые распределения $A^*(t)$ и $B^*(t)$:

$$A(t) = A^*(\alpha t), \quad B(t) = B^*(\alpha t). \qquad (1)$$

Характеристики СМО, которые определяются только функциями $A(t)$ и $B(t)$ и имеют размерность времени, изменяются при увеличении α пропорционально α^{-1}.

2. Основные соотношения

Обозначим τ время между приходами смежных приоритетных заявок, x – время обслуживания приоритетной заявки, τ' – время между приходами рядовых заявок и x' – время обслуживания рядовой заявки. Средние значения этих случайных величин (сл. в.) будут обозначаться, соответственно, в виде $\overline{\tau}, \overline{x}, \overline{\tau}', \overline{x}'$. В дальнейшем в качестве символа среднего любой сл. в. будем использовать ее надчеркнутое обозначение.

Интенсивности λ, λ' приоритетного и неприоритетного потоков выражаются через средние интервалы поступления заявок: $\lambda = 1/\overline{\tau}$, $\lambda' = 1/\overline{\tau}'$. Коэффициент загрузки системы ρ_Σ положим меньшим единицы:

$$\rho_\Sigma = \rho + \rho' < 1, \qquad (2)$$

где $\rho = \lambda \bar{x}$ – коэффициент загрузки СМО приоритетными заявками, $\rho' = \lambda' \bar{x}'$ – коэффициент ее загрузки рядовыми заявками.

Приоритетные заявки в системе «не ощущают» рядовых заявок, поэтому с их точки зрения рассматриваемая СМО является системой G|G|1 с одним входным потоком заявок. Систему, получаемую из исходной СМО удалением потока рядовых заявок, назовем системой S. Обозначим через π длину периода занятости в этой системе, через ψ – длину периода незанятости. Период занятости и следующий за ним период незанятости образуют период регенерации [2]. Процессы, которые по определению принадлежат разным периодам регенерации, статистически независимы.

При больших α чистое время обслуживания x' рядовой заявки с высокой вероятностью многократно превышает среднюю длину $\overline{\psi}$ периода незанятости СМО приоритетными заявками. Поскольку рядовая заявка обслуживается только во время этих периодов незанятости, то ее обслуживание завершается, когда их сумма перекрывает заданное значение величины $x' = T$. Независимость и одинаковое распределение всех периодов ψ_i, покрывающих в сумме заданное время обслуживания T, позволяют рассматривать их как *поток восстановлений* и применять к ним соответствующую хорошо разработанную теорию [3].

Последовательные периоды π_i занятости СМО приоритетными заявками представляют собой приращения суммарного времени Z_T прерываний рядовой заявки в процессе ее обслуживания. Между собой приращения π_i независимы, как и периоды незанятости ψ_i. Однако любые два периода π_i и ψ_i , составляющие вместе *один* период регенерации (в системе S), в общем случае зависимы. В [3] с точностью до обозначений рассматриваются асимптотические свойства именно такой последовательности интервалов восстановления (у нас это ψ_i) с независимыми приращениями (π_i), в котором допускается зависимость внутри соответствующих пар сл. в. (π_i и ψ_i). Известны характеристики суммы приращений Z_T, которая накапливается в процессе покрытия интервалами восстановления большого (относительно них) отрезка времени T. Поскольку в нашей системе $G_2|G_2|1$ сл. в. Z_T представляет собой суммарное время прерываний рядовой заявки при *фиксированном* времени ее обслуживания $x' = T$, то эти известные характеристики являются ее *условными* вероятностными характеристиками.

Общий подход к имитации системы $G_2|G_2|1$ при больших α будет состоять в том, чтобы перейти от нее к моделированию системы S' класса G|G|1, которая получается удалением (так или иначе скомпенсированным) из исходной системы потока приоритетных заявок. В системе S' присутствует только поток рядовых заявок, и их скорректированное время обслуживания определяется как сумма чистого времени обслуживания и времени прерываний, вычисляемого по его условным характеристикам. Задержкой начала обслуживания рядовой заявки, которая может возникать из-за ее прихода во время незавершенного периода занятости π, пренебрежем, т. к. при $\alpha \to \infty$ она в среднем сводится к нулю.

3. Метод усреднения времени прерываний

Из свойств восстановлений с приращениями [3] вытекает, что время Z_T прерываний обслуживания рядовой заявки при больших значениях α в среднем пропорционально ее чистому времени обслуживания T:

$$\overline{Z}_T \sim \frac{\overline{\pi}}{\overline{\psi}} \cdot T = \frac{\rho}{1-\rho} \cdot T, \qquad (3)$$

где $\overline{\pi}$ и $\overline{\psi}$ – средние значения сл. в. π и ψ соответственно, а символ ~ обозначает сходимость с нулевой относительной погрешностью (при $\alpha \to \infty$).

Суммируя чистое время обслуживания рядовой заявки T со средним временем прерываний $\overline{Z}_{\mathrm{T}}$, получаем общее время обслуживания $h_{\mathrm{T}} = \mathrm{T} + \overline{Z}_{\mathrm{T}} = \mathrm{T}/(1-\rho)$. Таким образом, скомпенсировать удаление приоритетного потока при моделировании системы S' можно увеличением времени обслуживания рядовой заявки в $(1-\rho)^{-1}$ раз. Именно это и реализуется известным практическим приемом, упомянутым в [1].

Однако такой метод компенсации, пренебрегающий дисперсией времени прерываний Z_{T}, занижает и дисперсию скорректированного времени обслуживания, и вместе с нею занижает среднее время $\overline{w}'$ ожидания рядовых заявок.

4. Метод суммы периодов занятости

Используя результаты теории восстановлений [3], второй момент времени Z_{T} прерываний можно выразить в форме коэффициента вариации следующим образом:

$$C_{Z\mathrm{T}}^2 \quad \sim \quad \frac{\overline{\psi}}{\mathrm{T}}\left(C_{\pi}^2 - 2rC_{\pi}C_{\psi} + C_{\psi}^2\right), \tag{4}$$

где $C_{Z\mathrm{T}}$ – коэффициент вариации времени прерываний Z_{T},

$C\pi$, $C\psi$ – коэффициенты вариации сл.в. π и ψ соответственно,

r – коэффициент корреляции периодов π и ψ в одном периоде регенерации.

Время прерываний Z_{T} сходится по распределению к нормальной сл. в., имеющей коэффициент вариации (4) и среднее значение (3).

В формуле (4) величина $\overline{\psi}$ пропорциональна α^{-1}. Следовательно, коэффициент вариации $C_{Z\mathrm{T}}$, а вместе с ним и погрешности известного метода усреднения, пропорциональны $\alpha^{-1/2}$, т. е. уменьшаются достаточно медленно.

Для того, чтобы при компенсации времени прерываний Z_{T} учитывать его второй момент (4), необходимо определить соответствующие характеристики системы S, в которой присутствует только приоритетный поток заявок. Это можно сделать с помощью моделирования системы S, которое имеет «обычную» трудоемкость. Поэтому в целом моделирование системы $G_2|G_2|1$ получается двухэтапным. На втором этапе моделируется система S' (только с рядовыми заявками). В системе S' перед обслуживанием рядовой заявки сначала определяется ее чистое время обслуживания $x' = \mathrm{T}$, а затем к нему добавляется суммарное время прерываний, которое имеет среднее $\overline{Z}_{\mathrm{T}}$ и коэффициент вариации $C_{Z\mathrm{T}}$, вычисляемые через T по формулам (3) и (4).

Асимптотические свойства рассмотренных процессов восстановления позволяют также определить два момента числа γ_{T} прерываний обслуживания

$$\overline{\gamma}_{\mathrm{T}} \quad \sim \quad \mathrm{T}/\overline{\psi}, \tag{5}$$

$$C_{\gamma\mathrm{T}}^2 \quad \sim \quad (\overline{\psi}/\mathrm{T})\cdot C_{\psi}^2, \tag{6}$$

а также коэффициент корреляции числа прерываний и времени прерываний:

$$corr(\gamma_{\mathrm{T}}, Z_{\mathrm{T}}) \quad \sim \quad \frac{C_{\psi} - rC_{\pi}}{\sqrt{C_{\pi}^2 - 2rC_{\pi}C_{\psi} + C_{\psi}^2}}, \tag{7}$$

где r – коэффициент корреляции сл. в. π, ψ внутри периода регенерации.

Сл. в. γ_T и Z_T при больших α распределены по двухмерному нормальному закону, и коэффициент корреляции между ними не зависит от α.

Метод суммы периодов занятости исследован аналитическими средствами и экспериментально. В качестве критерия точности метода использовалась ошибка $\Delta\overline{w}'$, вносимая компенсацией в среднее время $\overline{w}'$ ожидания рядовых заявок.

Аналитическая проверка метода на системах класса $M_2|G_2|1$, выполненная с помощью имеющихся для этого класса СМО точных решений [4], показала, что при их имитации ошибка метода $\Delta\overline{w}' = 0$ при любом α.

Экспериментальная проверка, выполненная с моделями на языке GPSS, подтвердила высокую точность метода в классе систем $G_2|G_2|1$. Коэффициенты вариации интервала поступления и времени обслуживания заявок варьировались в пределах от 0 до 3. На практике, в широком диапазоне параметров СМО, уже при $\alpha = 10 \div 20$ условное время прерываний Z_T с высокой точностью отвечает нормальному распределению с параметрами (3), (4).

Недостатком метода, усложняющим планирование имитационного эксперимента, является необходимость предварительной имитации системы S для оценки величин r, $\overline{\psi}$, C_π и C_ψ, используемых на основном этапе моделирования.

Заметим, что в случае, если система S относится к классу M|G|1, эти величины известны: $r = 0$, $\overline{\psi} = \overline{\tau}$, $C_\psi = 1$ и $C_\pi^2 = (C_x^2 + \rho)/(1-\rho)$.

5. Метод суммы периодов обслуживания

Формирование времени прерываний Z_T можно представить и в виде другого механизма восстановлений и накоплений, определяемого прямо через сл. в. τ и x, которые имеют заданные распределения вероятностей $A(t)$ и $B(t)$. Сл.в. Z_T может быть представлена как сумма длительностей обслуживания всех прерывающих заявок:

$$Z_T = x_1 + x_2 + \ldots + x_{\nu_T}, \tag{8}$$

где ν_T – число прерывающих заявок. При этом ν_T определяется как такое число независимых слагаемых вида $(\tau_i - x_i)$ в сумме $(\tau_1 - x_1) + (\tau_2 - x_2) + \ldots + (\tau_{\nu_T} - x_{\nu_T})$, при котором добавление еще одного слагаемого $(\tau_{\nu_T+1} - x_{\nu_T+1})$ приводит к превышению заданного чистого времени T обслуживания рядовой заявки. Система сл. в. $\{x_i, (\tau_i - x_i)\}$, $(i = 0, 1, 2, \ldots, \nu_T)$ имеет, с точностью до обозначений, те же свойства, которые установлены выше для системы сл.в. $\{\pi_i, \psi_i\}$, $(i = 0, 1, 2, \ldots, \gamma_T)$, за тем несущественным при больших α исключением, что слагаемые $(\tau_i - x_i)$ могут быть отрицательными. Следовательно, в соотношениях (3) – (7) все параметры сл.в. π и ψ можно просто заменить соответствующими им параметрами сл.в. x и $(\tau - x)$. Этот способ применения теории восстановлений менее очевиден, но приводит к более простым и хорошо интерпретируемым формулам. В роли интервала восстановления здесь выступает сл. в. $(\tau - x)$.

Выполняя в формулах (3) – (7) после оговоренной замены переменных алгебраические преобразования, учитывающие простые связи между разными представлениями моментов сл. в., получаем следующие результаты.

Во-первых, из (3) и (4) после соответствующих замен и упрощений находим:

$$\overline{Z}_T \sim \frac{\overline{x}}{\overline{\tau} - \overline{x}} \cdot T = \frac{\rho}{1-\rho} \cdot T, \tag{9}$$

$$C_{ZT}^2 \sim \frac{\overline{\tau}}{T} \cdot \left(\frac{C_\tau^2 + C_x^2}{1-\rho} \right), \tag{10}$$

где Cτ и Cx – коэффициенты вариации сл. в. τ и x соответственно,

ρ – коэффициент загрузки СМО приоритетными заявками.

Формулы (9) и (10) выражают моменты сл. в. Z_T непосредственно через характеристики интервала поступления τ и времени обслуживания x приоритетных заявок, вследствие чего отпадает необходимость этапа предварительной имитации системы S перед моделированием системы S'.

Во-вторых, преобразуя таким же способом соотношения (5) и (6), находим два первых момента числа прерывающих заявок ν_T:

$$\overline{\nu}_T \sim \frac{T}{\overline{\tau}} \cdot \frac{1}{(1-\rho)}, \tag{11}$$

$$C^2_{\nu T} \sim \frac{\overline{\tau}}{T} \cdot \left(\frac{C_\tau^2 + \rho^2 C_x^2}{1-\rho} \right), \tag{12}$$

а также коэффициент корреляции между ν_T и Z_T:

$$corr(\nu_T, Z_T) \sim \frac{1+\rho Q^2}{\sqrt{1+Q^2} \cdot \sqrt{1+\rho^2 Q^2}}, \tag{13}$$

где $Q = C_x / C_\tau$.

В-третьих, приравнивая выражения (4) и (10) одного и того же параметра и умножая полученное равенство на T, приходим к соотношению

$$\overline{\psi}(C_\pi^2 - 2rC_\pi C_\psi + C_\psi^2) = \overline{\tau} \cdot \left(\frac{C_\tau^2 + C_x^2}{1-\rho} \right), \tag{14}$$

которое обязано быть точным при $\alpha \to \infty$ и, следовательно, является точным при любом значении α, т. к. не зависит от него. Соотношение (14), таким образом, представляет собой инвариант, выполняющийся точно *в любой системе* G|G|1.

Метод суммы периодов обслуживания эквивалентен по точности методу суммы периодов занятости, но значительно проще в практическом применении. Однако он несколько уступает в плане универсальности; он не позволяет, например, имитировать число γ_T прерывающих периодов занятости.

С учетом (9), (10) и формулы Кингмана [5] для $\rho_\Sigma \to 1$, теперь можно оценить погрешность известного метода усреднений. При больших ρ_Σ оценка $\overline{w}'_o$ величины $\overline{w}'$ методом усреднений удовлетворяет приближенному равенству:

$$(\overline{w}' - \overline{w}'_o)/\overline{w}'_o \approx \alpha^{-1} \cdot [\rho^2/(1-\rho)] \cdot [(C_\tau^2 + C_x^2)/C_{x'}^2], \tag{15}$$

где $C_{x'}$ – коэффициент вариации времени обслуживания рядовой заявки. При фиксированных α и ρ относительное занижение времени ожидания (15) сверху не ограничено.

Разработанные методы суммирования, напротив, практически не вносят погрешности, т. к. уже при $\alpha > 10 \div 20$ учитывают время прерываний с точностью до распределения. При значениях α порядка 10^n, когда обычное усреднение прерываний из-за существенных ошибок может оказаться неприемлемым, эти методы позволяют без потери точности ускорить имитацию, соответственно, примерно в 10^n раз.

Литература

1. **Максимей И. В.** Функционирование вычислительных систем (Измерения и анализ). – М.:Советское радио, 1979. – 272 с.
2. **Iglehart D. L.** The regenerative method for simulation analysis//In Current Trends in Programming Methodology. Vol. III: Software Modelling, K.M. Chandy and R.T. Yen, Eds., Prentice-Hall, Englewood, Cliffs N.J., 1978. P. 52–71.
3. **Кокс Д. Р.** Теория восстановления /Кокс Д.Р, Смит В. Л.: Пер. с англ./Под ред. Ю. К Беляева. – М.: Сов. радио, 1967 г. – 312 с.
4. **Гнеденко Б. В.** Приоритетные системы обслуживания/ Б. В. Гнеденко , Э. А. Даниэлян , Б. Н. Димитров и др. – М.: Изд-во Московского университета, 1973. – 447 с.
5. **Клейнрок Л.** Вычислительные системы с очередями: Пер. с англ./Под ред. Б. С. Цыбакова. – М.: Мир, 1979. – 600 с.