АСИМПТОТИЧЕСКИЕ ПРИБЛИЖЕНИЯ В ИМИТАЦИОННОМ МОДЕЛИРОВАНИИ ПРИОРИТЕТНЫХ СИСТЕМ С ОЧЕРЕДЯМИ

В. Н. Задорожный (Омск)

Наличие статистической погрешности в численных результатах имитационного моделирования (ИМ) затрудняет использование градиентных методов [1] и численных алгоритмов решения уравнений или неравенств. Одним из возможных способов решения проблемы является использование аналитической информации о свойствах исследуемых зависимостей, в частности, использование асимптотических приближений, основанных на точном предельном анализе математических соотношений, которыми определяются свойства модели. В данной статье предлагаются асимптотические приближения, разработанные для моделирования приоритетных систем с очередями.

Приоритетные системы с относительно интенсивными прерываниями

Методами анализа вложенных процессов восстановления и накопления [2, 3] для приоритетных систем массового обслуживания (СМО) $GI_2 \mid GI_2 \mid n$ с абсолютными приоритетами и дообслуживанием заявок установлен ряд асимптотических соотношений, полезных для решения проблемы разномасштабных интенсивностей. Она состоит в том, что при $\alpha = \lambda_1 / \lambda_2 \to \infty$ (где λ_1 и λ_2 – интенсивности прихода приоритетных и неприоритетных заявок соответственно) затраты компьютерного времени на имитационный эксперимент (ИЭ) также стремятся к бесконечности. Методы [2, 3] можно рассматривать как основанное на теории восстановления [4] распространение (асимптотически точное) метода декомпозиции очередей [5, 6] (разработанного Гейвером для систем $M_2 \mid GI_2 \mid 1$) на более широкий класс приоритетных систем. Найденные для систем $GI_2 \mid GI_2 \mid n$ асимптотические соотношения позволяют исключать из модели поток приоритетных заявок, освобождаясь от разномасштабности интенсивностей. Исключая поток приоритетных заявок, мы заменяем время обслуживания x_2 остающихся неприоритетных заявок на длительность h их цикла обслуживания в исходной системе и получаем систему $G \mid G \mid n$, время ожидания заявок в которой сходится по распределению (при $\alpha = \lambda_1 / \lambda_2 \to \infty$) ко времени ожидания неприоритетных заявок в исходной системе.

Рассмотрим, в частности, систему $GI_2 \mid GI_2 \mid n$ с разделением неприоритетной работы, в которой каждая неприоритетная заявка обслуживается всеми каналами одновременно. Как только в процессе ее обслуживания системой какой-либо канал освобождается от обслуживания приоритетных заявок, на него тут же переносится часть незавершенного обслуживания обрабатываемой неприоритетной заявки. Скорость ее обслуживания k каналами в k раз больше скорости обслуживания одним каналом. Новая неприоритетная заявка обслуживается только после того, как завершится обслуживание предыдущей неприоритетной заявки.

Для такой системы $GI_2 \mid GI_2 \mid n$ при $\alpha \to \infty$ установлены соотношения [3]:

$$h_{\mathrm{T}} = T + Z_{\mathrm{T}}, \quad \overline{Z}_{\mathrm{T}} \sim \frac{\rho_1}{1-\rho_1} \cdot T, \quad C^2_{Z_{\mathrm{T}}} \sim \frac{n\overline{\tau}_1}{T} \cdot \left(\frac{C^2_{\theta} + C^2_{x_1}}{1-\rho_1} \right), \qquad (1)$$

где h_{T} – условная длительность цикла обслуживания неприоритетной заявки, рассматриваемая при условии, что чистое время x_2 ее обслуживания равно T; Z_{T} – условное время прерываний обслуживания неприоритетной заявки; символом надчеркивания случайной величины (сл.в.) в виде $\overline{\theta}$ обозначается переход к математическому ожида-

нию θ; символ $\sim$ обозначает сходимость с нулевой относительной погрешностью; $\rho_1 = \bar{x}_1/(n\bar{\tau}_1) = \lambda_1\bar{x}_1/n$ – приоритетный коэффициент загрузки системы; x_1 и τ_1 – интервал обслуживания и интервал поступления приоритетных заявок; в виде C_θ^2 обозначается квадратичный коэффициент вариации (к.в.) сл.в. θ. Распределение сл.в. Z_T сходится к нормальному.

Безусловное календарное время y обслуживания неприоритетной заявки (время от начала до завершения обслуживания) имеет характеристики:

$$\bar{y} \sim \frac{\bar{x}_2}{n(1-\rho_1)}; \quad C_y^2 \sim C_{x_2}^2 + \frac{n\rho_1(C_{\tau_1}^2 + C_{x_1}^2)}{(1-\rho_1)} \cdot \frac{\bar{x}_1}{\bar{x}_2}; \quad f_y(t) \sim b_2\big(n(1-\rho_1)\cdot t\big). \qquad (2)$$

где $f_y(t)$ и $b_2(t)$ – плотности распределения вероятностей сл.в. y и x_2.

Эти формулы справедливы при любых «рациональных» правилах выбора доступных каналов приоритетными заявками и могут использоваться для приближенных вычислений уже при $\alpha = 1...10$.

Пакетирование поступающих заявок

В существующих СМО пакетирование наиболее часто осуществляется по отношению к поступающим неприоритетным заявкам. Накопленный пакет неприоритетных заявок обслуживается системой в перерывах между обслуживанием приоритетных заявок, т.е. в фоновом режиме. Способ формирования пакетов влияет на интервал Θ накопления (прихода) пакетов, на число N заявок в нем, на его суммарную трудоемкость ξ и на длительность H периода его фонового обслуживания [7]. В частности, при Т-пакетировании заявок, когда интервал Θ накопления пакета поступающих неприоритетных заявок фиксирован ($\Theta = T_0$), нетрудно установить следующие асимптотические соотношения, справедливые при $T_0 \to \infty$:

$$\bar{N} \sim \frac{T_0}{\bar{\tau}_2}; \quad D(N) \sim \frac{D(\tau_2)}{\bar{\tau}_2^3} \cdot T_0; \quad C_N^2 \sim \frac{D(\tau_2)}{\bar{\tau}_2 \cdot T_0}; \qquad (3)$$

$$\bar{\xi} \sim \frac{T_0}{\bar{\tau}_2} \cdot \bar{x}_2; \quad D(\xi) \sim T_0 \cdot \left(\frac{\sigma_{x_2}^2}{\bar{\tau}_2} + \frac{\sigma_{\tau_2}^2 \cdot \bar{x}_2^2}{\bar{\tau}_2^3} \right); \quad C_\xi^2 \sim \frac{\bar{\tau}_2}{T_0} \cdot (C_{x_2}^2 + C_{\tau_2}^2), \qquad (4)$$

где τ_2 – интервал поступления неприоритетных заявок, а в виде $D(\theta)$ или σ_θ^2 обозначается дисперсия сл.в. θ. Распределения сл.в. N и ξ асимптотически нормальны.

Пример применения асимптотических приближений в ИМ

Серверный центр (СЦ) состоит из коммуникационной ЭВМ (КЭВМ) и многопроцессорного сервера, процессоры которого $\Pi_1,\ldots,\Pi_n$ имеют общую полнодоступную оперативную память (рис. 1), позволяющую организовать общую очередь задач к процессорам. КЭВМ принимает из внешней сети приоритетные оперативные задания, поступающие с интервалом $\tau_{1,1}$, и без задержки передает их серверу. Кроме того, КЭВМ пакетирует поступающие из внешней сети неприоритетные задания и периодически (дважды в сутки) передает очередной накопленный пакет серверу. Наряду с внешними заданиями, сервер обрабатывает внутренние задания, поступающие из локальной сети СЦ и обладающие таким же высшим приоритетом, как и оперативные задания из внешней сети. Оба приоритетных потока – внутренний и внешний – пуассо-

новские. Неприоритетный внешний поток тоже пуассоновский, но имеет разную интенсивность в разное время суток: в течение первых $t_0 = 8$ ч средний интервал поступления заявок $\overline{\tau}_2 = \overline{\tau}_{2,1} = 15$ мин, а в остальные 16 ч $\overline{\tau}_2 = \overline{\tau}_{2,2} = 5$ мин.

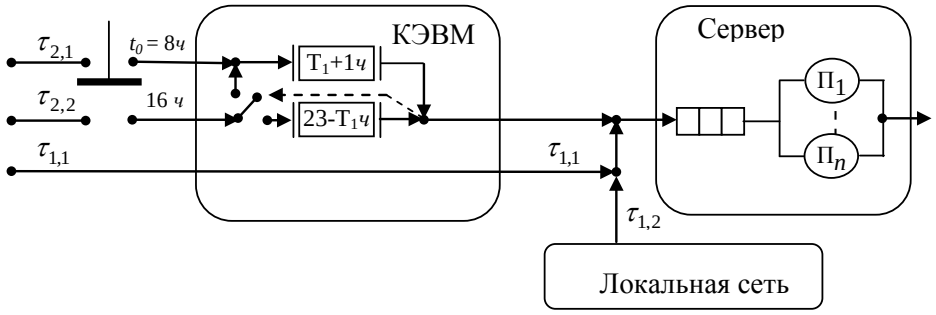

Рис. 1. Схема процесса обработки заданий в серверном центре

Средняя трудоемкость неприоритетных заданий $\overline{x}_2 = 5$ мин постоянна. Распределение трудоемкости x_2 – симметричное треугольное, на интервале от 0 до 10 мин. Выполнение неприоритетных заданий идеально распараллеливается между *n* процессорами сервера. Его начальная конфигурация – двухпроцессорная. Приоритетные задания выполняются в однопроцессорном режиме: одно задание – одним процессором. Приоритетные внешние задания поступают со средним интервалом $\overline{\tau}_{1,1} = 1$ мин и требуют в среднем $\overline{x}_{1,1} = 0{,}5$ мин процессорного времени, внутренние поступают в среднем через $\overline{\tau}_{1,2} = 0{,}5$ мин и выполняются в среднем за $\overline{x}_{1,2} = 0{,}1$ мин. Время $x_{1,1}$ выполнения приоритетных внешних заданий распределено по закону Эрланга 2-го порядка. Время $x_{1,2}$ выполнения приоритетных внутренних заданий имеет гиперэкспоненциальное распределение вероятностей 2-го порядка и характеризуется квадратичным к.в., равным 2.

В течение первых $T_1 + 1 > 8$ (ч), где T_1 – момент начала часового профилактического перерыва, $T_1 + 1$ – момент его завершения, накапливается ПЕРВЫЙ пакет неприоритетных заявок. В момент $T_1 + 1$ он передается на исполнение серверу, и начинается накопление ВТОРОГО пакета, которое продолжается в течение всего оставшегося времени суток $(23 - T_1)$ ч. В начале следующих суток ВТОРОЙ пакет передается серверу для выполнения. В тот же момент КЭВМ вновь начинает накопление первого в этих новых сутках пакета. Далее процесс повторяется циклически, изменяются только значения составляющих его сл.в. При этом ПЕРВЫЙ пакет выполняется между моментом $T_1 + 1$ и концом суток, а ВТОРОЙ – между началом суток и моментом T_1. Во время профилактического часа $T_1 \le t \le T_1 + 1$ пакеты не обрабатываются.

Задача моделирования заключается в оптимальном выборе момента T_1: необходимо обеспечить минимум суммы вероятностей p_1 и p_2 невыполнения в срок пакетов (ПЕРВОГО и ВТОРОГО) соответственно. Кроме того, необходимо определить, как повлияет на функционирование СЦ прогнозируемый на ближайшие полгода (180 дней) приблизительно линейный рост спроса на услуги СЦ, из-за которого утренняя интенсивность $\lambda_{2,1} = 1/\overline{\tau}_{2,1}$ и дневная интенсивность $\lambda_{2,2} = 1/\overline{\tau}_{2,2}$ ежедневно будут расти на 0,5% каждая. Если этот дрейф интенсивностей будет угрожать качеству обслуживания внешних клиентов, то следует дать рекомендации по предотвращению этой угрозы.

Применяя к СЦ асимптотические приближения (1)–(4), выражаем его показатели через искомое T_1, регулируемое n и дрейфующие $\lambda_{2,1} = 1/\bar{\tau}_{2,1}$ и $\lambda_{2,2} = 1/\bar{\tau}_{2,2}$:

$$\bar{\xi}_1 \sim \left[\frac{480}{\bar{\tau}_{2,1}} + \frac{60\cdot(\mathrm{T}_1 - 7)}{\bar{\tau}_{2,2}}\right]\bar{x}_2, \quad D(\xi_1) \sim 480\left(\frac{\sigma^2_{x_2}}{\bar{\tau}_{2,1}} + \frac{\sigma^2_{\tau_{2,1}}\cdot\bar{x}_2^2}{\bar{\tau}^3_{2,1}}\right) + 60(T_1 - 7)\left(\frac{\sigma^2_{x_2}}{\bar{\tau}_{2,2}} + \frac{\sigma^2_{\tau_{2,2}}\cdot\bar{x}_2^2}{\bar{\tau}^3_{2,2}}\right);$$

$$\bar{\xi}_2 \sim \frac{60\cdot(23 - \mathrm{T}_1)}{\bar{\tau}_{2,2}}\cdot\bar{x}_2; \quad D(\xi_2) \sim 60\cdot(23 - T_1)\cdot\left(\frac{\sigma^2_{x_2}}{\bar{\tau}_{2,2}} + \frac{\sigma^2_{\tau_{2,2}}\cdot\bar{x}_2^2}{\bar{\tau}^3_{2,2}}\right);$$

$$\bar{y}_1 \sim \frac{\bar{\xi}_1}{n(1-\rho_1)}; \qquad C^2_{y_1} \sim C^2_{\xi_1} + \frac{n\rho_1}{(1-\rho_1)}\cdot(C^2_{\tau_1} + C^2_{x_1})\cdot\frac{\bar{x}_1}{\bar{\xi}_1};$$

$$\bar{y}_2 \sim \frac{\bar{\xi}_2}{n(1-\rho_1)}; \qquad C^2_{y_2} \sim C^2_{\xi_2} + \frac{n\rho_1}{(1-\rho_1)}\cdot(C^2_{\tau_1} + C^2_{x_1})\cdot\frac{\bar{x}_1}{\bar{\xi}_2}, \qquad (5)$$

где ξ_1 и ξ_2 – трудоемкости первого и второго пакетов соответственно (в минутах);

y_1 и y_2 – время выполнения первого и второго пакетов соответственно;

$\rho_1 = \rho_{1,1} + \rho_{1,2} = \bar{x}_{1,1}/(\bar{\tau}_{1,1}n) + \bar{x}_{1,2}/(\bar{\tau}_{1,2}n)$; $\bar{x}_1 = \bar{x}_{1,1}/3 + 2\bar{x}_{1,2}/3 = 7/30$;

$\sigma^2_{x_1} = (\sigma^2_{x_{1,1}} + \bar{x}^2_{1,1})/3 + 2(\sigma^2_{x_{1,2}} + \bar{x}^2_{1,2})/3 - \bar{x}_1^2 = 0{,}09055$; $C^2_{x_1} = \sigma^2_{x_1}/\bar{x}_1^2 = 1{,}6633$.

Используя последнее из соотношений (2) и учитывая асимптотическую нормальность сл.в. ξ_1 и ξ_2, теперь можно записать для вероятностей $p_1 = \mathbf{P}(y_1 > T_{01})$ и $p_2 = \mathbf{P}(y_2 > T_{02})$ следующие выражения:

$$p_1 \approx 1 - \Phi\left(\frac{T_{01} - \bar{y}_1}{\sqrt{D(y_1)}}\right); \qquad p_2 \approx 1 - \Phi\left(\frac{T_{02} - \bar{y}_2}{\sqrt{D(y_2)}}\right), \qquad (6)$$

где $\Phi_{01} = 60(23 - T_1)$ – интервал в минутах, выделенный для выполнения ПЕРВОГО пакета; $\Phi_{02} = 60T_1$ – интервал для ВТОРОГО пакета; Φ – стандартная нормальная функция распределения вероятностей. Графики зависимостей $p_1(T_1)$ и $p_2(T_1)$, построенные по формулам (6), (5), приведены на рис. 2, *а*. На рис. 2, *б* к ним добавлены результаты полуторачасового аттестующего проверочного ИЭ (отмечены маркерами + и ×).

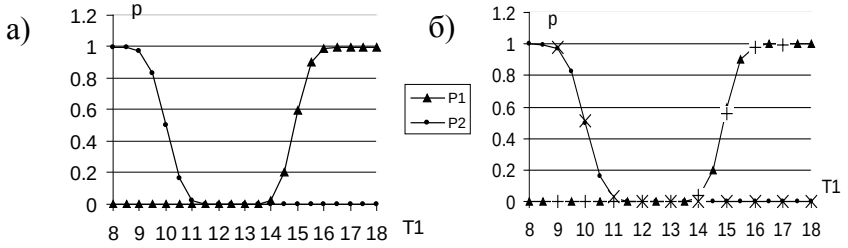

Рис. 2. Влияние момента T1 начала профилактики на вероятности превышения сроков обработки:

а – расчет по асимптотическим приближениям;
б – сравнение с результатами моделирования

Поскольку при оптимальном выборе T_1 имеет место $p_1(T_1) = p_2(T_1)$, то оптимальное T_1 определяется из уравнения, получаемого приравниванием выражений в (6). Его численное решение элементарно и для первого дня работы СЦ дает оптимальное значение $T_1 = 12{,}583$ (*ч*) = 12 ч 35 мин, при котором $p_1 = p_2 \approx 1{,}1 \cdot 10^{-8}$.

На рис. 3 изображена часть графиков для характеристик $p_1(T_1)$ и $p_2(T_1)$, найденных для различных сценариев эволюции СЦ с учетом дрейфа интенсивностей. Для расчета по асимптотическим формулам десятков таких графиков потребовалось не более секунды компьютерного времени. При непосредственном ИМ на их вычисление ушло бы несколько суток с существенной потерей точности. Поскольку эти расчеты проводятся при растущей интенсивности потока приоритетных заявок, то параметры предельного перехода здесь также возрастают, а следовательно, используемые асимптотические приближения становятся более точными. Поэтому необходимость в их аттестации с помощью непосредственного ИМ далее отпадает.

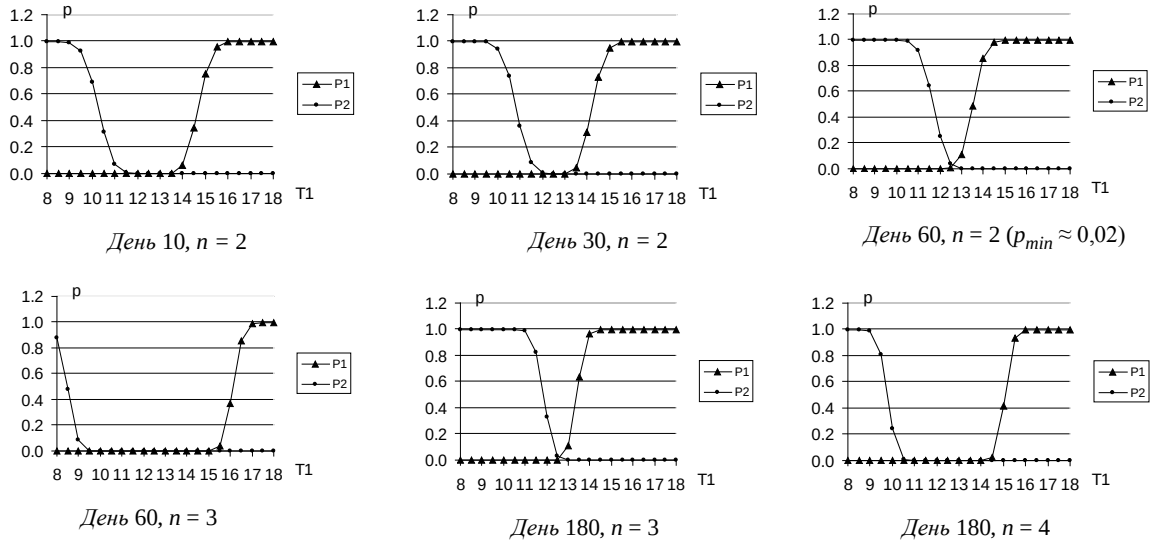

Рис. 3. Характеристики $p_1(T_1)$ и $p_2(T_1)$ для различных сценариев развития СЦ

Результаты исследования точности предлагаемых асимптотических приближений приводятся в [7], рекомендации по их практическому использованию при ИМ приоритетных циклических систем с пакетированием заявок излагаются в [8].

Выводы

Использование в ИМ предлагаемых в статье асимптотических приближений позволяет в ряде случаев существенно сократить затраты машинного времени на решение оптимизационных задач и одновременно повысить точность результатов. В общем случае для применения этих аппроксимаций требуется предварительная аттестация их точности с помощью непосредственного ИМ. Если решение задачи продолжается в такой области факторного пространства, где аттестованные аппроксимации могут только уточняться, то дальнейшее их использование может проводиться без применения ИМ.

Литература

1. **Рыжиков Ю. И.** Имитационное моделирование. Теория и технологии. СПб.: КОРОНА принт. М.: Альтекс-А, 2004. 384 с.
2. **Задорожный В. Н.** Методы ускоренной имитации процессов с интенсивными прерываниями//Материалы II Всероссийской конференции (ИММОД-2005). СПб.: ФГУП ЦНИИ ТС, 2005. Т. 1, С. 101–106.
3. **Задорожный В. Н.** Анализ систем с приоритетами методом декомпозиции//Омский научный вестник, 2005. № 3(32). С. 126–132.
4. **Кокс Д. Р., Смит В. Л.** Теория восстановления/ Пер. с англ.; под ред. Ю. К Беляева. М.: Сов. радио, 1967. 312 с.
5. **Gaver D. P.** A waiting line with interrupted service, including priorities/J. Roy. Stat. Soc., Ser. B 25 (1962). P. 73–90.
6. **Jaiswal N. K.** Priority Queues (Academic Press, New York, 1968)/Пер. на рус. яз.: Джейсуол Н. К. Очереди с приоритетами. М.: Мир, 1973. 280 с.
7. **Задорожный В. Н.** Асимптотический анализ периодов повышенной нагрузки в приоритетных системах//Омский научный вестник. 2006. № 3(36). С. 117–124.
8. **Задорожный В. Н.** Комбинированный метод моделирования циклических систем обслуживания//Омский научный вестник. 2006. № 9(46). С. 156–163.