ИМИТАЦИОННОЕ МОДЕЛИРОВАНИЕ В ОБОСНОВАНИИ МЕТОДИК РАСЧЕТА МНОГОКАНАЛЬНЫХ ПРИОРИТЕТНЫХ СИСТЕМ

Ю. И. Рыжиков (Санкт-Петербург)

Задачи управления сложными техническими комплексами в реальном времени обычно требуют использования многомашинных или многопроцессорных систем – как из соображений производительности, так и с учетом требований надежности и организации технического обслуживания. Различия в важности задач, их трудоемкости и требованиях оперативности решения приводят к необходимости введения *приоритетных* дисциплин обслуживания.

Методы анализа одноканальных систем с приоритетами разработаны достаточно хорошо [1, 2], хотя численная их реализация и ставит ряд непростых проблем. Однако сложность этих задач резко возрастает при переходе к многоканальным системам. Такие задачи обычно решаются лишь в простейшем (экспоненциальном) варианте с одинаковыми по типам заявок средними – случай нетипичный и малоинтересный [3, 5]. Все попытки создания реально применимых методик, учитывающих находящиеся в каналах и в очередях количества заявок каждого вида [4], заведомо обречены на неудачу в связи с непомерным разрастанием пространства состояний. Не всегда помогают и имитационные системы: GPSS World не позволяет моделировать многоканальные устройства с приоритетными прерываниями – более того, даже одноканальные с кратными прерываниями.

В подобных случаях могут быть эффективны комбинации аналитических подходов с имитационными моделями, причем последние используются трояко:

- для получения первичных исходных данных;
- для генерации и проверки гипотез (допущений), закладываемых в аналитические фрагменты алгоритма;
- для окончательной комплексной проверки.

Задача анализа многоканальной приоритетной системы сводится к расчету среднего времени недоступности системы для обслуживания «меченой» заявки типа j. В случае приоритета с прерываниями к нему добавляется средняя длительность прерывания, умноженная на число прерываний. Решения этих задач для одноканальной системы элементарны, а для многоканальных отсутствуют. Для качественного осознания возникающих эффектов пришлось прибегнуть к имитационному моделированию.

При моделировании системы с абсолютным приоритетом длительность обслуживания заявки определялась в момент ее постановки в очередь: при входе в систему – с помощью соответственно настроенного датчика случайных чисел, после прерывания – с остатком ранее сформированной длительности. Прерванные заявки помещались в голову соответствующей очереди.

Одной из новых проблем оказалось определение необходимости прерывания и выбор прерываемого канала. Оказалось целесообразным иметь список занятых каналов, упорядоченный по убыванию приоритетов обслуживаемых заявок (при равных приоритетах – по возрастанию моментов прибытия в систему), в этом случае приоритет вновь прибывшей заявки было достаточно сопоставить с обслуживаемой в последнем канале. Разумеется, это потребовало переупорядочения списка при прерывании и при выборе заявки из очереди после завершении обслуживания

Изучение *конечных* результатов имитационного моделирования многоканальных систем могло привести в лучшем случае к эмпирическим аппроксимациям с областью применения, ограниченной исследованным диапазоном параметров. Поэтому имитаци-

онная модель была дополнена сбором статистики по искомым «внутренним» показателям.

Для расчета среднего времени недоступности удалось найти *аналитический* подход. Распределение времени ожидания начала обслуживания меченой заявки в одноканальной системе было аппроксимировано по двум моментам, из которых первый дается общеизвестной формулой, а правильное выражение для второго имеется в [1] (его варианты с разными ошибками приводятся по крайней мере в 5 источниках). Для аппроксимации применялась дополнительная функция распределения Вейбулла $\overline{F}(t) = \exp(-t^k / T)$, с которой связаны теоретические моменты $f_m = T^{m/k}\Gamma(1+m/k), \quad m = 1,2,...$ В n-канальном случае ДФР времени ожидания является n-й степенью вышеприведенной и сводится к тому же распределению с делением параметра T на n. Соответственно, среднее время ожидания вычисляется согласно $w(n) = (T/n)^{1/k}\Gamma(1+1/k)$, так что $w(n)/w(1) = (1/n)^{1/k}$. Моделирование подтвердило хорошую точность этого метода.

Теперь рассмотрим методы решения второй и третьей задач, найденные с помощью имитационного моделирования. Типы заявок считаются упорядоченными по убыванию приоритетов.

Период непрерывной занятости прерываниями

Моделирование периодов непрерывной занятости (ПНЗ) многоканальной системы однородными заявками при поддержании постоянной удельной загрузки на канал показало (табл. 1), что при марковском (М), т. е. показательном распределении обслуживания, средняя длина ПНЗ

$$\pi(n) = \frac{b_1}{n(1-\lambda b_1 / n)},$$

где b_1 – средняя длительность обработки головной заявки периода занятости. В остальных случаях эта зависимость могла рассматриваться лишь как грубое приближение. В связи с этим было принято решение искать общую формулу в виде

$$\pi(n) = \frac{b_1(1+\Delta(\rho, v, n))}{n(1-\rho)},$$

где ρ – коэффициент загрузки канала и v – коэффициент вариации обслуживания. Необходимая поправка Δ вычислялась через наблюденное в эксперименте среднее значение периода непрерывной занятости $\pi(n)$ по формуле

$$\Delta = (n-1)\pi(n)b_1 - 1.$$

Отмеченное выше нулевое значение поправки для показательного распределения обслуживания определило ее мультипликативное строение и необходимость обращения в нуль при $v = 1$. Кроме того, в широком диапазоне коэффициентов вариации (от нуля до двух) зависимость от v при прочих равных условиях оказалась близка к линейной. Далее, модуль поправки был приблизительно пропорционален коэффициенту загрузки ρ. Наконец, поправка нелинейно росла по числу каналов n, обнаруживая тенденцию к насыщению, и по определению равнялась нулю при $n = 1$ (последнее требование – еще один аргумент за мультипликативную форму поправки). В итоге оказалось, что следует принять $\Delta = \rho(v-1)(n-1)/(4n)$ (числовой множитель подобран экспериментально).

В табл. 1 представлены результаты вычисления средней длины периода непрерывной занятости системы однородными заявками на имитационной модели (И) и по расчету (Р). Распределением H_2 заменялось гамма-распределение с коэффициентом вариации 2. Во всех случаях предполагалось $b_1=1$.

Таблица 1

Средние длительности непрерывной занятости

n	ρ	Распределение обслуживания							
		D		E_3		M		H_2	
		И	Р	И	Р	И	Р	И	Р
1	0.5	1.997	2.000	1.998	2.000	2.003	2.000	2.008	2.000
	0.7	3.332	3.333	3.339	3.333	3.349	3.333	3.325	3.333
	0.9	9.971	10.000	10.121	10.000	9.975	10.000	9.510	10.000
2	0.5	0.917	0.938	0.974	0.974	0.998	1.000	1.053	1.062
	0.7	1.496	1.521	1.599	1.605	1.673	1.667	1.781	1.812
	0.9	4.421	4.438	4.732	4.762	4.972	5.000	5.110	5.562
3	0.5	0.593	0.611	0.633	0.643	0.667	0.667	0.723	0.722
	0.7	0.959	0.981	1.044	1.056	1.114	1.111	1.243	1.241
	0.9	2.809	2.833	3.113	3.122	3.312	3.333	3.609	3.813
4	0.5	0.441	0.453	0.473	0.480	0.502	0.500	0.553	0.547
	0.7	0.708	0.724	0.774	0.787	0.836	0.833	0.950	0.943
	0.9	2.048	2.078	2.275	2.321	2.517	2.500	2.747	2.922
5	0.5	0.351	0.360	0.377	0.383	0.401	0.400	0.443	0.440
	0.7	0.560	0.573	0.614	0.627	0.667	0.667	0.773	0.760
	0.9	1.619	1.640	1.810	1.848	1.990	2.000	2.262	2.360

Ожидаемое число прерываний

Вновь поровну разделим входящий поток между каналами обслуживания. Для одноканальной системы ожидаемое число прерываний $\overline{k_j} = \Lambda_{j-1} b_{j,1}$, где $\Lambda_{j-1} = \sum_{i=1}^{j-1} \lambda_j$ есть интенсивность потока заявок с правом прерывания j-й. В n-канальном случае заявка, прибывшая во время обслуживания j-й, может вообще ее не прерывать (если хотя бы один канал свободен или занят обслуживанием менее приоритетной заявки). С другой стороны, меченую заявку могут прервать и те, которые «изначально» пришлись на другие каналы. Ясно, что первый эффект будет преобладать для заявок относительно высокого приоритета, а второй – для низкого. Приближенно можно считать, что прерывание j-заявки происходит при выполнении следующих условий:

- она имеется хотя бы в одном из каналов (вероятность равна $1-(1-\rho_j / R_k)^n$);
- ни в одном из остальных каналов нет заявок более низкого приоритета (вероятность $\left(R_j / R_k\right)^{n-1}$).

В этих формулах $\{R_j\}$ суть кумулянтные коэффициенты загрузки системы заявками до j-го типа включительно, а k – индекс последнего типа заявок. Приходящие прерывающие заявки приходятся в среднем на $\rho_j = \lambda_j b_{j,1}$ прерываемых. Таким образом, среднее число прерываний j-заявки можно оценить по формуле

$$\overline{k}_j = \Lambda_{j-1} b_{j,1} \left[1 - \left(\frac{\rho_j}{R_k} \right)^n \right] \left(\frac{R_j}{R_k} \right)^{n-1} \Big/ \left(\lambda_j b_{j,1} \right) = \frac{\Lambda_{j-1}}{\lambda_j} \left[1 - \left(\frac{\rho_j}{R_k} \right)^n \right] \left(\frac{R_j}{R_k} \right)^{n-1}.$$

В табл. 2 результаты такого расчета сопоставляются с полученными на имитационной модели. Рассматривалась система с тремя типами заявок при средних длительностях обслуживания $b_{1,1}$=0.45, $b_{2,1}$ = 0.90, $b_{3,1}$ =1.35 и интенсивностях потоков на канал $\lambda_1 = 0.2$, $\lambda_2 = 0.3$, $\lambda_3 = 0.4$ (коэффициент загрузки 0.9 поддерживался умножением этих интенсивностей на число каналов). Поскольку обнаружилось, что типы распределений длительности обслуживания на кратность прерываний практически не влияют, объем таблицы соответственно сокращен.

Таблица 2

Средние кратности прерываний

n	Типа 2		Типа 3	
	И	Р	И	Р
1	0.180	0.180	0.674	0.674
2	0.116	0.122	0.827	0.944
3	0.071	0.063	0.882	1.051
4	0.043	0.029	0.892	1.094
5	0.026	0.013	0.897	1.111

Таблица иллюстрирует качественное соответствие результатов ожидаемым и неплохое согласие полученных кратностей для низкоприоритетных заявок типа 3, т. е. как раз там, где прерывания могут дать заметную дополнительную задержку. Для заявок высокого приоритета согласие заметно хуже, но сами кратности весьма малы, да и длительности прерываний тоже будут малыми. Здесь итоговая погрешность заведомо несущественна.

Сопоставление конечных результатов

Предложенные подходы были запрограммированы на Фортране 77. В качестве эталонных использовались результаты имитационного моделирования многоканальных систем при 200 тыс. наблюдений по заявкам высшего приоритета.

Для верификации моделей решалась одноканальная задача при вышеуказанных исходных данных, приведенных в разделе о кратности прерываний. Достаточно большой суммарный коэффициент загрузки (R = 0.9) обеспечивал существенную роль длительностей ожидания и прерываний низкоприоритетных заявок. Результаты расчета сведены в табл. 3.

Таким образом, полученные из анализа имитационных экспериментов допущения относительно средней длины ПНЗ и кратности прерываний позволили получить несложную и достаточно точную методику приближенного расчета многоканальных приоритетных систем.

Таблица 3

Средние времена пребывания в системе с абсолютным приоритетом

n	Тип	Распределение обслуживания							
		D		E_3		M		H_2	
		И	Р	И	Р	И	Р	И	Р
1	1	0.472	0.472	0.480	0.480	0.493	0.495	0.559	0.561
	2	1.233	1.242	1.311	1.323	1.478	1.486	2.192	2.216
	3	10.038	10.104	12.875	12.741	18.364	18.014	40.557	41.745
2	1	0.452	0.453	0.452	0.454	0.454	0.455	0.459	0.462
	2	1.000	1.006	1.020	1.023	1.064	1.060	1.223	1.229
	3	5.718	5.640	6.840	6.851	9.536	9.264	20.065	20.108
3	1	0.450	0.451	0.450	0.451	0.450	0.452	0.449	0.453
	2	0.942	0.947	0.948	0.954	0.959	0.969	1.019	1.040
	3	4.113	4.161	4.858	4.928	6.398	6.456	13.168	13.314
4	1	0.450	0.450	0.451	0.451	0.452	0.451	0.446	0.451
	2	0.919	0.925	0.923	0.929	0.926	0.937	0.951	0.976
	3	3.350	3.424	3.911	3.979	5.053	5.083	9.597	10.036
5	1	0.450	0.450	0.450	0.455	0.450	0.450	0.447	0.450
	2	0.910	0.925	0.911	0.918	0.916	0.923	0.920	0.947
	3	2.956	2.985	3.399	3.416	4.311	4.274	7.920	8.122

Литература

1. **Джейсуол Н.К.** Очереди с приоритетами/пер. с англ. – М.: Мир, 1973. – 279 с.
2. Приоритетные системы обслуживания. – М.: МГУ, 1973. – 447 с.
3. **Сотский Н.М., Чуркин Е.А**. Стационарное распределение длины очереди в многоканальной системе массового обслуживания с приоритетами//АиТ. – 1985. – №1. – С. 69–76.
4. **Хомоненко А.Д.** Вероятностный анализ приоритетного обслуживания в многопроцессорных системах//АВТ. – 1990. – №2. – С. 55–61.
5. **Gail H.R., Hantler S.L., Taylor B.A.** Analysis of a non-preemptive priority multiserver queue//Advances in applied prob. – 1988. – v. 20. – P. 852.