УДК 519.6

МОДЕЛИРОВАНИЕ ИНФОРМАЦИОННЫХ ПРОЦЕССОВ В ВЫЧИСЛИТЕЛЬНОЙ ПОДСИСТЕМЕ С ПРИМЕНЕНИЕМ АГРЕГАТИВНЫХ МОДЕЛЕЙ

А. Н. ПРИВАЛОВ
В. Л. КУЛЕШОВ

Тульский артиллерийский инженерный институт

e-mail: alexandr_prv@rambler.ru

Рассматривается применение агрегативных моделей для моделированияя информационных процессов в вычислительной подсистеме, спроектированной на основе систем с распределенной обработкой данных. Приводятся результаты эксперимента.

Ключевые слова: агрегативная модель, сеть массового обслуживания, информационные процессы.

Функционирование любой тренажёрной системы на основе вычислительных систем с распределенной обработкой данных (СРОД) может быть представлено в виде совокупности взаимодействий пользователей (операторов, руководителей обучения) с системой [1]. С формальной точки зрения любое такое взаимодействие можно отобразить в виде последовательности этапов передачи и обработки информации.

Для оценки взаимодействия однородных информационных процессов (ИП) в вычислительной подсистеме тренажёрных систем предлагается подход, основанный на агрегативном описании систем с использованием особенностей их структуры.

Существо подхода заключается в выделении из рассматриваемой системы некоторой подсистемы (агрегата) и последующем ее детальном исследовании; при этом учитывается влияние остальной части системы, которая представляется в виде обобщенной сети массового обслуживания (СМО) с интенсивностью обслуживания, зависящей от числа заявок в ней.

Полагается, что эта СМО ведет себя по отношению к оставшейся подсистеме аналогично той части системы, которую она заменяет.

В общем случае СРОД достаточно сложной структуры может быть разбита на несколько подсетей, которые исследуются отдельно. При этом объединение отдельных СМО в подсети осуществляется таким образом, чтобы:

– взаимодействия элементов внутри подсети (внутренние взаимодействия) могли быть исследованы без учета;

– взаимодействий между подсетями; взаимодействия между подсетями (внешние взаимодействия) можно было анализировать без учета внутренних взаимодействий.

Указанные условия будут выполняться, если частота внутренних взаимодействий много больше частоты внешних взаимодействий. В этом случае говорят о почти разложимой системе.

Агрегативные модели дают точное решение для сетей, допускающих решение в виде произведения. Так можно показать, что точное решение может быть получено для сетей МО, описываемых моделью Гордона – Ньюэлла [2]. Агрегативный подход часто связывают с замкнутыми сетями, однако он может быть использован и в отношении открытых сетей МО.

Рассмотрим функционирование вычислительного комплекса (ВК) в режиме реального масштаба времени, ограничиваясь анализом ИП на уровне ВК. Пусть на ВК поступает пуассоновский поток заявок интенсивностью λ, а реализуемые информационно-вычислительные работы (ИВР) связаны с выполнением достаточно длинной цепочки переходов: процессор–внешнее устройство- процессор.... Время обслуживания одного запроса на устройство $j, j = \overline{1, J}$ распределено по экспоненциально-

му закону со средним $1/?_j$. Здесь индекс «*1*» относится к узлу процессора. Вероятность выхода на узел (устройство) $j, j = \overline{2,J}$, после обслуживания в узле процессора есть θ_{1j}. Наибольший допустимый уровень мультипрограммирования в вычислительном комплексе – *N*.

Для исследования ИП в подобной ситуации может быть использован как аппарат открытых, так и замкнутых сетей МО. Однако в первом случае не будет учитываться ограничение на уровень мультипрограммирования, что может привести к существенному завышению оценок пропускной способности вычислительного узла, особенно при большой загрузке системы. Во втором случае получаемые оценки будут занижены, поскольку при расчетах будет приниматься максимальная пропускная способность вычислительного комплекса, соответствующая уровню мультипрограммирования *N*, в то время как при числе активных ИП в вычислительном узле, меньшем *N*, эта величина может заметно отличаться от ее максимального значения.

В данной ситуации ошибка в расчетах увеличивается с уменьшением загрузки системы. Для преодоления указанных трудностей необходима модель, отражающая реальную зависимость пропускной способности рабочей станции от числа активных ИП. Такая модель получается при использовании агрегативного подхода. В самом деле, интенсивность переходов между состояниями, связанными с занятием отдельных ресурсов (процессоров, ВУ) при реализации активного ИП, значительно больше интенсивности последовательных активизаций ИП. Поэтому можно выделить и рассмотреть отдельно замкнутую подсеть, отображающую процесс реализации в мультипрограммном режиме *n* активных ИП, $n = \overline{1,N}$.

На рис. 1. а схематически представлена рассматриваемая открытая сеть МО, а на рис. 1,б – образованная из нее подсеть, которую условно можно считать замкнутой.

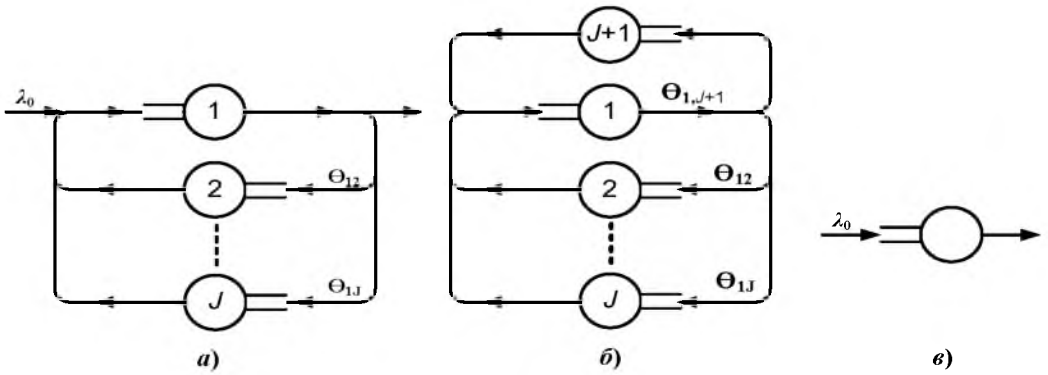

Рис. 1. Агрегирование открытой сети МО

В последнем случае для удобства вычислений введен фиктивный узел *J+1*, для которого полагаем $1/?_{J+1}=0$.

Пропускная способность вычислительного комплекса определяется зависимостью (1)

$$\lambda(n) = e_{J+1} G(n-1)/G(n), n = \overline{1,N}, \quad (1)$$

т.е. является функцией от *n*. При этом e_{J+1} определяется из решением системы:

$e_j = \sum_{k=1}^{J} e_k \theta_{kj}$, а *G(K)* рассчитывается по рекуррентной формуле

$$G_j(k) = G_{j-1}(k) + p_j G_j(k-1), j = \overline{2,J}, k = \overline{1,K}, \quad (2)$$

где *K* –число заявок, циркулирующих в сети.

Рассмотренную подсеть заменим эквивалентным обслуживающим прибором, интенсивность обслуживания которым выражается следующим образом:

$$\gamma_i = \lambda(i), i = \overline{1,N}, \gamma_i = \lambda(N), \gamma > N. \quad (3)$$

Допуская теперь, что время обслуживания эквивалентным прибором можно считать распределенным по экспоненциальному закону, рассматриваемую систему представим в виде СМО типа M/M/1 с переменной интенсивностью обслуживания, задаваемой соотношением (3). Решение такой СМО может быть записано в виде:

$$P_i = \begin{cases} \lambda^i P_0 \Big/ \prod_{k=1}^{i} \lambda(k), 1 \le i \le N; \\ \lambda^i P_0 \Big/ \left\{ [\lambda(N)]^{i-N} \prod_{k=1}^{N} \lambda(k) \right\}, i > N; \end{cases} \quad (4)$$

$$P_0 = \left\{ 1 + \sum_{i=1}^{N} \left[\lambda^i \Big/ \prod_{k=1}^{i} \lambda(k) \right] + \lambda^{N+1} \Big/ \left[\prod_{k=1}^{N} \lambda(k)(\lambda(N) - \lambda) \right] \right\}^{-1}.$$

Рассчитав стационарные вероятности состояний (4), можно получить требуемые характеристики реализации ИП в системе. Например, среднее число заявок в системе выразится как

$$n_1 = P_0 \left\{ \left[\sum_{i=1}^{N} i\lambda^i \Big/ \prod_{k=1}^{i} \lambda(k) \right] + \frac{1}{\prod_{k=1}^{N} \lambda(k)} \frac{\lambda^{(N+1)}}{\lambda(N) - \lambda} \left[N + 1 + \frac{\lambda}{\lambda(N) - \lambda} \right] \right\}.$$

Отсюда среднее время реакции рабочей станции определяется на основании формулы Литтла: $\upsilon_1 = n_1 / \lambda$.

Рассмотрим пример, демонстрирующий основные особенности расчетных схем рассматриваемой модели.

Пусть необходимо оценить пропускную способность ВК, схематически изображенного на рис. 2, где показана его структура: процессор (узел 1) и три внешних запоминающих устройства (ВЗУ) (узлы 2, 3, 4). Каналы обмена данными в ВК не являются узким местом, так что очереди образуются лишь к процессору и ВЗУ. Объем оперативной памяти ВК такой, что возможна одновременная реализация лишь двух ИП (работа в двухпрограммном режиме). Параметры соответствующих ИП заданы; время одного обслуживания любым из устройств можно считать распределенным по экспоненциальному закону с интенсивностями $\mu_1 = 5, \mu_j = 1, j = \overline{2,4}$; после завершения обслуживания процессором с равной вероятностью 0,3 выдаются запросы на устройства 2, 3,4 и с вероятностью 0,1 ИП завершается.

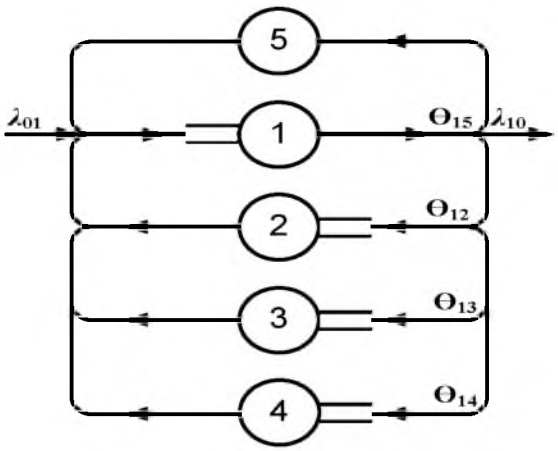

Рис. 2. Пример структуры ВК

При сравнении оценок основных характеристик реализации, полученных на основе агрегативного подхода, будем полагать, что пропускная способность рабочей станции определяется как $\lambda(N)$. Для простоты будем полагать, что предельный уровень мультипрограммирования N=2. Результаты расчетов представлены в табл. 1, где приведены значения среднего числа ИП в системе ($N1$) и среднего времени реакции рабочей станции(υ_1).

Таблица 1

Результаты вычислительного эксперимента с различными типами моделей

λ	Замкнутая модель		Агрегативная модель		Открытая модель	
	N_1	υ_1	N_1	υ_1	N_1	υ_1
0,05	0,5	10	0,68	13,4	0,65	13
0,1	2,0	20	2,3	23	0,69	6,9
0,14	13	93	14,3	102	2,6	18
0,2	-	-	-	-	4,8	24
0,33	-	-	-	-	299	906

Анализируя эти данные, можно обратить внимание, что при малой загрузке рабочей станции (малых значениях входящего потока заявок на ИВР) результаты расчетов, полученные на основе агрегативного подхода и в результате анализа открытой сети МО, весьма близки. Это вполне естественно, поскольку в данном случае система работает с недогрузкой и ограничение на уровень мультипрограммирования почти не проявляется. При большой загрузке системы $(\lambda \approx 0{,}14)$ оценки, полученные на основе агрегативного подхода, близки к соответствующим оценкам, полученным путем анализа замкнутой сети МО $\left[\gamma_i = \lambda(N), i = \overline{1,N}\right]$. Этот факт также легко объясняется: в условиях большой загрузки система почти постоянно работает при максимальном уровне мультипрограммирования, что соответствует наивысшей пропускной способности. Таким образом, использование агрегативного подхода в данном случае позволяет наиболее полно отобразить реальные условия функционирования ВК.

Аналогичным образом агрегатирование применимо и при исследовании замкнутых систем. Рассмотрим в качестве примера ВК, обслуживающий некоторое число M абонентов, каждый из которых обращается к рабочей станции с заявками на ИВР.

Время между последовательными обращениями абонента к рабочей станции распределено по экспоненциальному закону с параметром λ. Причем новая заявка может быть сформирована абонентом лишь после получения ответа на предыдущую. Следовательно, при наличии в рабочей станции(на обслуживании и в очереди) i заявок поток заявок от всех абонентов будет иметь пуассоновский характер с интенсивностью

$$\lambda_i = (M - i)\lambda, i = \overline{0, M}. \quad (5)$$

Пусть наибольший уровень мультипрограммирования для рабочей станции есть N.

Теперь, рассматривая рабочую станцию как замкнутую сеть МО, можем, как и ранее, определить величины $\lambda(n)$ в соответствии с (1). Предположим, что при работе рабочей станции с уровнем мультипрограммирования i поток завершения решений задач является пуассоновским с параметром γ_i, причем γ_i определяется в соответствии с (3).

В этом случае из анализа замкнутой сети с двумя приборами с переменной интенсивностью обслуживания, определяемой выражениями (3) и (5), получаем соотношения для расчета вероятностей P_i (числа заявок в ВК):

$$P_i = \lambda^i M! P_0 \Big/ \left[(M - i)! \prod_{k=1}^{i} \gamma_k \right], i = \overline{1, M}; \quad (6)$$

$$P_0 = \left[1 + \sum_{r=1}^{M} \frac{\lambda^r M!}{\prod_{k=1}^{r} \gamma_k (M-r)!} \right]^{-1}.$$

Отсюда среднее время реакции системы выразится как

$$\upsilon_1 = M \left[\sum_{i=1}^{N} P_i \lambda(i) + \lambda(N) \sum_{i=N+1}^{M} P_i \right]^{-1} - 1/\lambda. \qquad (7)$$

Расчеты с использованием (7) показывают, что учет ограничений на уровень мультипрограммирования при применении агрегативного подхода позволяет более точно оценивать основные характеристики функционирования мультипрограммных ВК при конечном числе источников заявок. Особенно это относится к условиям большой загрузки ВК, когда при игнорировании указанных ограничений пропускная способность рабочей станции может быть завышена в несколько раз [1].

В [2] приведена одна из наиболее общих методик реализации агрегативного подхода, основанная на итеративном вычислении характеристик замкнутых сетей МО с общим распределением времени обслуживания заявок в порядке их поступления.

Методика дает точные результаты для сетей, удовлетворяющих условиям локального баланса.

Рассмотрим замкнутую сеть МО, в которой циркулирует I заявок и имеется J узлов обслуживания. Время обслуживания в узле $j, j = \overline{1,J}$ распределено по некоторому закону $B_i(t)$ со средним b_{j1}. Заданы вероятности θ_{jk} перехода в узел k после обслуживания в узле j. Это позволяет рассчитать величины e_j путем решения системы уравнений вида:

$$e_j = \sum_{k=1}^{J} e_k \theta_{kj}.$$

Такая сеть именуется сетью A. В соответствии с методикой [2] строится некоторая последовательность сетей $A_0, A_1, \ldots$ аппроксимирующих сеть A. При этом сеть A_0 получается из A путем перехода к экспоненциальному распределению времени обслуживания в узлах с тем же значением среднего времени обслуживания. Сети A_r, r=1,2,... отличаются от A_0 лишь значением среднего времени обслуживания в узлах $1/?_j(r)$.

Расчетная процедура методики, определяющая переход от A_k к A_{k+1} и условия завершения вычислений, заключается в реализации некоторой последовательности шагов, на первом шаге которой средние значения времени обслуживания в узлах сети A_0 полагаются равными соответствующим величинам в сети A, а распределение времени обслуживания – экспоненциальным. Далее для каждого узла j строится подсистема, содержащая все узлы сети, кроме j. При этом исследуется сеть с двумя узлами, так что выделенная подсистема отображается некоторой СМО, эквивалентной ей в смысле воздействия на узел j (для простоты анализа производится перенумерация: узел j нумеруется как 1, а эквивалентный прибор – как 2).

Для сети с двумя узлами известными методами рассчитываются значения среднего числа заявок в узле $j(n_{j1})$ и среднего числа заявок, обслуживаемых в единицу времени (λ_j), а также приведенное среднее число заявок, обслуживаемых в узле j в единицу времени, $\gamma_j=\lambda_j/e_j$. После этого осуществляется проверка сходимости величин n_{j1} и γ_j на основе соотношений

$$\left| I - \sum_{j=1}^{J} n_{j1} \right| \le \varepsilon I; \left| \gamma_j - \sum_{k=1}^{J} \gamma_k / J \right| \le \varepsilon \gamma_j. \qquad (8)$$

Здесь ε – малая величина, задающая порядок допустимой ошибки (например, ε=0,01). Выполнение условий (8) на некоторой итерации указывает на приемлемость полученного отображения A с помощьюA_0. В противном случае осуществляется корректировка средних значений времени обслуживания в узлах с помощью выражений (в зависимости от выполнения того или иного из условий (8) и производится следующая (r+1)-я итерация.

Описанная методика использовалась для расчета характеристик некоторых замкнутых сетей, а результаты расчетов сравнивались с результатами имитационного моделирования на ряде экспериментов [2]. В одном из этих экспериментов рассматривалась сеть, в которой число узлов варьировалось от 2 до 5, число заявок – также от 2 до 5, а времена обслуживания полагались распределенными по законам: экспоненциальному, Эрланга 2-го порядка и гиперэкспоненциальному. Результаты эксперимента показали, что расхождение в вычислении значения загрузки узлов не превышает 0,05, а в вычислении среднего числа заявок в узле в основном находится в пределах 0,05. В другом эксперименте исследовались сети из 6-7 узлов, число заявок в которых варьировалось от 2 от 12, а времена обслуживания полагались распределенными по законам гиперэкспоненциальному, Эрланга порядка 2, 4 и 6, экспоненциальному, а также считались постоянными. В результате наблюдалось хорошее согласование вычисляемых характеристик (в пределах 0,05 для загрузки узлов и 0,05J для среднего числа заявок в узле). Расчет тех же сетей точными методами в предположении экспоненциального времени обслуживания давал более значительные расхождения в оценке указанных характеристик.

Литература

1. Привалов А.Н., Моделирование информационных процессов в вычислительной подсистеме тренажёрных систем специального назначения [Текст] / А.Н. Привалов. – Тула: Изд-во ТулГУ, 2009. – 215 с.

2. Балыбердин, В.А., Оптимизация информационных процессов в распределённых системах обработки данных [Текст] / В.А. Балыбердин, А.М. Белевцев, О.А. Степанов. – М.: Технология, 2002. –280 с.

PERFORMANCE EVALUATION OF VIRTUAL TRAINING SYSTEM WITH DIDTRIBUTED DATA PROCESSING

A. N. PRIVALOV
V. L. KULESHOV

Tula artillery engineering institute

e-mail: alexandr_prv@rambler.ru

Application is considered aggregate models for modeling of information processes in computing subsystem, designed on basis of systems with distributed data processing. Experimental results are resulted.

Key words: Aggregate model, queuing network, information processes.