квадратичного программирования используется метод множителей Лагранжа [1, 2], а для поиска граничных решений - методы линейного программирования (метод Вольфа) [3]. При этом приходится многократно решать систему из n+m линейных уравнений с помощью симплекс-алгоритма, соблюдая условия обращения в ноль либо функции ограничения, либо соответствующего множителя Лагранжа, где m - общее количество функций ограничений типа неравенства. В результате получим одно граничное решение x_i^*, которое для задачи (1) может оказаться неоптимальным.

Предлагаемый в работе подход, даже в предусматриваемых случаях перебора вариантов решений на границах, значительно упрощает решение класса задач (1) - (3) и легко поддается программной реализации. Множество решений, получаемое в результате перебора некоторых вариантов, близких к оптимальному, может использоваться в задачах управления с принятием решения.

ЛИТЕРАТУРА

1. **Химмельблау Д.** Прикладное нелинейное программирование. - М.: Мир, 1975. - 534 с.
2. **Аоки М.** Введение в методы оптимизации. - М.: Наука, 1977. - 344 с.
3. **Дегтярев Ю.И.** Исследование операций. - М.: Высшая школа, 1986. - 320 с.

ГИУА

04.12.1997

Изв. НАН и ГИУ Армении (сер. ТН), т. LI, № 3, 1998, с. 345 - 351

УДК 519.95:681.3

АВТОМАТИЗАЦИЯ И СИСТЕМЫ УПРАВЛЕНИЯ

В.Г. СААКЯН, Д.А. МОВСЕСЯН, Г.С. КЮРЕГЯН

ИССЛЕДОВАНИЕ ВРЕМЕННЫХ ХАРАКТЕРИСТИК СТОХАСТИЧЕСКОЙ МОДЕЛИ ИНФОРМАЦИОННОЙ СИСТЕМЫ

Բաշխված ինֆորմացիոն համակարգերում սերվերների ինֆորմացիայի տեղադրման և ինդեքսավորման արդյունավետության բարձրացման համար առաջարկվում են ստատիստիկական փորձարկումների մեթոդներ: Ստացված տվյալների հիման վրա հետազոտվում են մոդելավորվող համակարգերի ժամանակային բնութագրերը և հերթերի երկարությունները: Առաջարկվում է համակարգի սերվերների ինֆորմացիայի բաշխման բարելավման ալգորիթմ:

Предлагаются методы статистических испытаний для оценки и повышения эффективности размещения информации на серверах в распределенных информационных системах. На основе полученных результатов исследуются временные характеристики и длины очередей в моделируемой системе. Предлагается алгоритм улучшения распределения информации по серверам системы.

Ил. 1. Библиогр: 5 назв.

The methods of statistic tests for evaluating and increasing the effective information location on servers in distributed networks are suggested. Temporary characteristics and lengths of queues of the modeling system are studied on the basis of data received. The algorithm for improving the information distribution on servers of the system is proposed.

Il. 1. Ref. 5.

В настоящее время во многих организациях реализуются локальные информационные системы на основе технологии Intranet (внутреннего фрагмента Internet), используя протоколы и программное обеспечение Internet для своих распределенных информационных систем (РИС). Данная технология предполагает использование протокола с квитированием TCP/IP для передачи информации между клиентами и серверами системы. Поскольку, зачастую, информация в подобных распределенных системах накапливается без предварительной оценки ожидаемых временных характеристик ее каналов, то желательно оценивать и улучшать временные характеристики уже действующих систем. Известно, что временные характеристики в распределенных информационных системах существенно зависят от способа размещения информации по ее серверам. При этом необходимо иметь механизм предварительной оценки способов его изменения с тем, чтобы новое размещение улучшало характеристики всей системы в целом. Для систем с квитированием, в которых загрузка каналов в определенные промежутки времени может превышать их пропускные способности, использовать методы оценки временных параметров на уровне средних нецелесообразно из-за больших разбросов. Кроме того, функции, характеризующие эффективность подобных систем, как правило, имеют нелинейный характер, что еще больше усложняет исследование подобных моделей аналитическими методами.

В настоящей работе предлагается использовать метод статистических испытаний для оценки и повышения эффективности размещения информации на серверах в распределенных системах. При этом предполагается, что известны топология системы, пропускные способности ее каналов, а также интенсивности обращений клиентов к блокам информации в системе. Рассматриваются процессы передачи информации в распределенной информационной системе, использующей протокол TCP/IP. Оцениваются временные характеристики эффективности размещения информации по серверам РИС, предлагается алгоритм их улучшения.

Для формализации моделируемой информационной системы целесообразно ввести следующие обозначения [1].

Имеется РИС, состоящая из M узлов-клиентов информационной системы - U_i, $i=1,...,M$. Под узлами будем понимать элементы РИС, которые хранят, используют и маршрутизируют информацию с временным сохранением транзитных сообщений (пакетов) в буферах очередей (рис.).

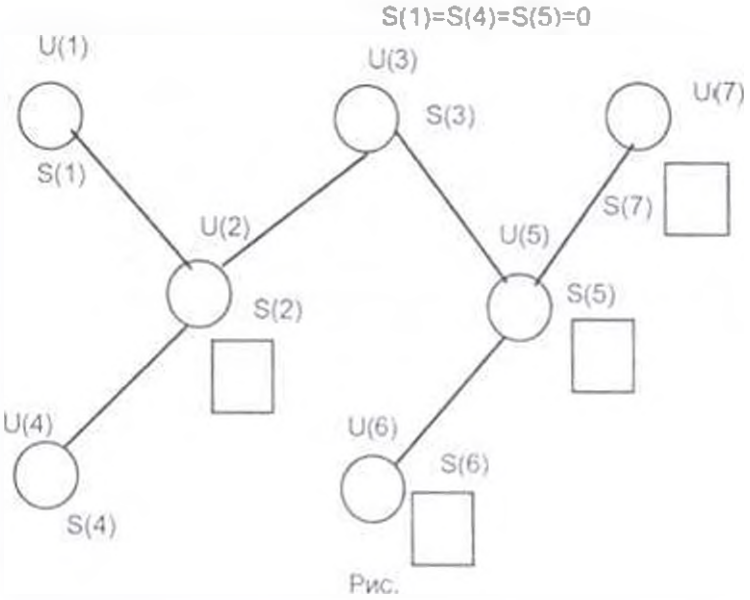

Рис.

В узлах системы имеются однотипные серверы с объемами дискового пространства $S_1, S_2, \ldots, S_M$, $S_j \geq 0$, $j=1,\ldots,M$. Задаются пропускные способности $C=\{c_{ij}\}$ каналов, связывающих узлы распределенной системы, где $c_{ij}>0$- пропускная способность канала, связывающего узлы U_i и U_j, $i,j=1,\ldots,M$ ($c_{ij}=0$, если узлы не связаны между собой).

Информационная система содержит K блоков информации с объемами $V_1, V_2, \ldots, V_K$, $V_i>0$, $i=1,\ldots,K$ ($\sum S > \sum V$). Интенсивности обращения пользователей узлов к информационным блокам определяются экспериментально или задаются априори. Пусть $a_{ij} \geq 0$ ($i=1,\ldots,K$, $j=1,\ldots,M$) - интенсивность пересылки i-го блока пользователям узла j-го узла. Обозначим через $X=\|x_{ij}\|$ способ размещения блоков информации по серверам, где $x_{ij}=1$, если i-й блок находится на сервере j-го узла, и $x_{ij}=0$, если i-й блок не находится на сервере j-го узла. Вопрос выбора начального распределения блоков информации по серверам РИС обсуждается, например, в [2] и выходит за рамки данной работы.

Введем следующее ограничение на матрицу X: каждый блок должен быть размещен на одном из серверов, $\sum_{j=1}^{M} x_{ij} = 1$. В общем случае суммарный объем размещенных на j-ом сервере блоков не может превышать его объема $\sum_{i=1}^{K} x_{ij} V_i \leq S_j$. На основе результатов моделирования может приниматься решение о перераспределении дискового пространства серверов РИС.

В качестве целевой функции рассмотрим среднее время передачи информации от серверов к клиентам:

$$\tau = \sum n_i \tau_i / n \rightarrow \min \quad (i = 1,\ldots,K),$$

где τ_i - среднее время доставки блока i-го блока ко всем затребовавшим его клиентам; n_i - количество доставок блока информации ко всем клиентам, n - общее количество доставок.

Опишем подробнее процесс моделирования РИС, использующих протоколы TCP/IP и UDP для передачи информации.

Как известно, протокол TCP/IP использует на транспортном уровне передачу информации с коммутацией каналов, а на сетевом уровне по проложенному виртуальному каналу передаются информационные блоки, разбитые на пакеты [3]. Протокол же UDP, используемый для передачи управляющей информации, построен по принципу коммутации сообщений.

Пусть P_{ij} - это последовательность номеров узлов, по которым проходит сообщение из i-го узла в j-й узел. На основе экспериментальных данных определяются функции распределения промежутков времени между запросами к блокам информации i-го типа со стороны i-го узла- $F_{ij}(\tau)$ и интенсивности $a_{ij} \geq 0$ обращения пользователей i-го узла к j-му блоку информации, используемые в дальнейшем для генерации обращений-событий в стохастической модели. Обычно функция $F_{ij}(\tau)$ имеет экспоненциальный закон распределения. В зависимости от способа распределения информационных блоков по серверам определяются функции распределения промежутков времени между поступлениями запросов к серверам системы $S_{ij}(\tau) = \sum_{k=1}^{K} F_{ij}(\tau | x_k)$. На основе значений $S_{ij}(\tau)$ вырабатываются моменты запросов t_{ij}^1, t_{ij}^2, $t_{ij}^3, t_{ij}^4, \ldots$ со стороны узлов-клиентов к соответствующим информационным блокам на узлах-серверах системы. Используя характеристики блока, для него порождается процесс передачи от источника (сервера) к адресату (клиенту РИС) - $P_{ij}(V_K)$. Если по истечении некоторого промежутка времени подтверждение не получено, источник передает пакет заново. Виртуальный канал может предоставляться процессам передачи двумя способами:

1. Система без отказов с выделением некоторой доли пропускной способности канала. В этом случае процессы передачи анализируют все каналы на маршруте от источника к адресату и запрашивают некоторую фиксированную долю от свободной пропускной способности каждого канала. По окончании каждого процесса передачи посылаются сигналы всем остальным процессам с тем, чтобы они пересчитали пропускные способности своих каналов с учетом освободившихся каналов.

2. Система с отказами при отсутствии каналов с достаточной пропускной способностью. В этом случае каждый процесс анализирует все каналы на маршруте от источника к адресату, и при наличии достаточной пропускной способности процессу предоставляется данный канал. В противном случае процесс получает отказ на предоставление ему канала и посылает следующий запрос на предоставление ему канала с необходимой пропускной способностью через некоторый промежуток времени. Данная модель

применима в современных сетях передачи информации, где для поддержки некоторых услуг (аудио- и видеоконференции) необходимо наличие каналов с высокой пропускной способностью.

Суммарное время передачи пакета от сервера к клиенту определяется как сумма времен передачи по каждому каналу и времен ожидания в очереди на передачу.

Каждое звено маршрута $i \rightarrow j$ в графе, описывающем топологию системы (рис.), может рассматриваться как система массового обслуживания. Процессы передачи $P_u(V_k)$ последовательно посылают заявки на его использование. При предоставлении виртуального канала по первому способу поступают заявки от активных процессов передачи от серверов к клиентам $P_u(V_k)$ на выделение определенной доли своей пропускной способности. Для каждого нового процесса передачи анализируются все звенья на маршруте от сервера к клиенту и на каждом звене выделяется определенный процент (66%) от свободной доли данного канала. В случае системы с отказами при поступлении заявки $r(P_u(V_k))$ на предоставление максимальной пропускной способности канала анализируются все звенья на маршруте от сервера к клиенту и выбирается звено с наименьшей пропускной способностью. Если его пропускная способность c_{min} не меньше требуемой минимальной пропускной способности канала, то начинается передача информации по заданному каналу. В противном случае заявка на передачу $r(P_u(V_k))$ проверяет звенья маршрута через случайный промежуток времени τ.

Итак, процессу $P_u(V_k)$ предоставляется канал с пропускной способностью $cP_u(V_k)$, равной минимуму из выделенных пропускных способностей. На время передачи $t = V_k / c(P_u(V_k))$ пропускная способность всех звеньев на маршруте $i \rightarrow j$ уменьшается на $c(P_u(V_k))$. По истечении времени соответствующие звенья маршрутов от серверов к клиентам освобождаются.

Если в процессе передачи заявки на предоставление канала по некоторому звену маршрута оно занято, то заявка размещается в очереди соответствующего узла. Выбор очередной заявки на передачу из очереди производится в соответствии с принятой дисциплиной FIFO, LIFO или RS. Если суммарное время передачи пакета (включая время доставки квитанции) превышает заданный предел, то процесс его передачи начинается заново.

В процессе моделирования определяются следующие параметры эффективности системы при заданном размещении информационных блоков по ее серверам: P_{ij} - коэффициенты загрузки каналов между ее серверами и клиентами; $Z = W + T$ - средние длины очередей в буферах узлов и среднее время задержки при передаче информации по каналу. По известным моментам начала и окончания процесса передач на каждом звене маршрута коэффициент загрузки для каждого канала можно определить по формуле

$$\rho_{lm} = \sum_{(i \to j) \in P_{lm}} \frac{t^{k}_{ij} - t^{n}_{ij}}{T},$$

где $l,m = 1,\ldots,M$; T - период исследования. Система будет функционировать, если $\rho_{lm} \le 1$ для всех каналов.

Рассмотрим алгоритм улучшения распределения информации по серверам РИС, перемещающий один блок информации за один шаг. Выбор алгоритма улучшения существенно зависит от начального размещения информационных блоков по серверам РИС [1-4].

Известно, что объем вычислений в подобных задачах дискретной оптимизации возрастает экспоненциально с увеличением размерности задачи [5]. Кроме того, даже при ограничении множества допустимых решений модель весьма чувствительна к изменениям системы. Поэтому целесообразно методом статистических испытаний предварительно оценить те временные характеристики РИС, которые используются при улучшении распределения информации по серверам РИС. В результате моделирования определяются: τ_{ij} - среднее время доставки i-го блока в j-й узел; n_{ij} - количество передач i-го блока в j-й узел; φ_{ij} - среднее время доставки пакета информации по каналу $i \to j$; n_{il} - количество передач пакетов i-го блока по каналу l, (по всем узлам, путь к которым содержит канал L); n_{i0} - количество передач пакетов блока i_0 в самом узле.

Правило улучшения размещения информации по серверам РИС выбираем в соответствии с целевой функцией:

$$\tau = \sum n_i \tau_i / n, \text{ где } n = \sum n_i; \; n_i = \sum n_{il}, \; \tau_i = \sum \tau_{ij}.$$

Рассмотрим последовательность $(n_i \tau_i)$, i=1,...k, показывающую долю вклада каждого блока в суммарный объем информации, передаваемой в системе за период моделирования.

Расположим величины этой последовательности в убывающем порядке

$$n_{i(1)}\tau_{i(1)} \ge n_{i(2)}\tau_{i(2)} \ge \ldots \ge n_{i(K)}\tau_{i(K)}. \qquad (1)$$

Обозначим

$$\Phi_1 = \varphi_{j(0) \to j(1)} \qquad K_0 = n_{i(1)},$$
$$\ldots \qquad K_1 = n_{j(0) \to j(1)},$$
$$\Phi_s = \varphi_{j(0) \to j(s)}, \qquad \ldots$$
$$K_s = n_{i(1) \to j(s)}.$$

Если $K_0 > K_0 + K_1 + \ldots + K_s$, тогда блок i_1 переносится в узел j_s, в противном случае производится проверка значений K_s для соседних узлов.

Если условие ни для одного из соседних узлов не выполняется, то берется блок l_2, соответствующий следующему элементу последовательности (1), и процесс переноса блока повторяется для соответствующего узла.

Из элементов последовательности (1) целесообразно рассматривать не все, а лишь ее первые элементы, которые оказывают существенное влияние на временные характеристики системы, т.е. алгоритм целесообразно завершить после перебора некоторой части p_t элементов последовательности. Правило выбора перемещаемого блока информации допускает небольшие отклонения в худшую сторону (из-за выбора не первого элемента последовательности (1)) с дальнейшим улучшением ситуации. Поэтому величина p_t должна быть близка к 1: $0,8 < p_t < 1$.

Предложенный алгоритм достаточно эффективен, поскольку информационные системы, построенные на основе технологии Intranet, как правило, имеют в своей основе сети с достаточно простой топологией и сравнительно малым количеством серверов.

В настоящее время продолжаются работы по реализации системы моделирования РИС, построенных на основе технологии Intranet, и улучшению предложенного агоритма распределения информации по ее серверам.

ЛИТЕРАТУРА

1. **Movsesyan D., Sahakyan V., Shoukourian A.** Optimum Allocation of Information in Distributed Information Systems of Networks with Low Loaded Transfer Channels // Proceedings of Conference of Computer Science and Information Technologies.- Yerevan, 1997. - P. 296-298.
2. **Гаврилов А.Л., Постельник Д.Я.** Задача синтеза информационной архитектуры сети // Информационные технологии -М: Машиностроение, 1997 - № 2. - С. 26-28.
3. **Tanenbaum A.** Computer Networks. - Engwood Cliffs. NJ. Prentice Hall, 1995. - 374 с.
4. **Зайченко Ю.П., Гонта Ю.В.** Структурная оптимизация сетей ЭВМ. - Киев: Техника. - 1986. - 167 с.
5. **Пападимитроу Х.** Комбинаторная оптимизация. - М.: Мир, 1985. -510 с.

ГИУА,
Ин-т проб.инф.и авт. НАН РА

25.12.1997