ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА

УДК 681.3

В. А. Богатырев, И. Ю. Голубев, В. Ф. Беззубов

ОРГАНИЗАЦИЯ МЕЖМАШИННОГО ОБМЕНА В ДУБЛИРОВАННЫХ ВЫЧИСЛИТЕЛЬНЫХ КОМПЛЕКСАХ

Проводится анализ надежности двухмашинных вычислительных комплексов при различных подходах к организации взаимосвязи между полукомплексами. Показано преимущество организации межмашинного обмена на основе двойного прямого доступа к памяти.

Ключевые слова: *дублированный вычислительный комплекс, отказоустойчивость, надежность, межмашинный обмен.*

Введение. Высокая надежность и отказоустойчивость [1, 2] управляющих компьютерных систем достигается при их построении на основе дублированных (двухмашинных) вычислительных комплексов (ДВК), зачастую объединяемых в кластеры [3—5].

В системах компьютерного управления двухмашинные комплексы функционируют либо в режиме дублированных вычислений (параллельной работы, при которой каждый запрос направляется на обслуживание в два полукомплекса, а результаты вычислений сравниваются), что повышает достоверность работы, либо в режиме разделения нагрузки, что позволяет повысить производительность системы, но снижает достоверность результатов вычислений и может привести к их потере.

Эффективность дублированных комплексов и кластеров на их основе во многом определяется организацией межмашинного обмена [6, 7], что обусловливает важность анализа при проектировании ДВК результативности использования известных вариантов организации межмашинного обмена и возможностей их модификации с учетом особенностей построения систем.

Проанализируем потенциальные возможности повышения эффективности дублированных комплексов в результате организации межмашинного обмена с двойным прямым доступом к памяти (ПДП) [8—10], суть которого заключается в конвейерном совмещении передачи данных с использованием ПДП одновременно в обоих полукомплексах [11]. Двойной ПДП потенциально позволяет ускорить межмашинный обмен при повышении отказоустойчивости дублированных комплексов [11,12] .

Организация дублированного комплекса. В качестве типовой рассмотрим реализацию дублированного комплекса (рис.1), каждый из полукомплексов которого содержит процессор (*P*) и модуль памяти (*M*). Реконфигурация системы и обмен данными между полукомплексами осуществляются с использованием переключателя (*S*) [12, 13].

При работе дублированного комплекса в режиме разделения нагрузки по мере накопления отказов при реконфигурации возможен переход (деградация) от обслуживания запросов двумя полукомплексами к их обслуживанию одним полукомплексом, формируемым, в частности, из исправного оборудования разных полукомплексов.

Если время выполнения запросов в системе является критичным и при отказе оборудования выполняемый запрос не может быть возобновлен без риска срыва процесса управления, прерванную обработку запросов следует восстанавливать, используя контрольные точки. В контрольных точках полукомплексы обмениваются данными, необходимыми для взаимоконтроля и восстановления вычислительного процесса.

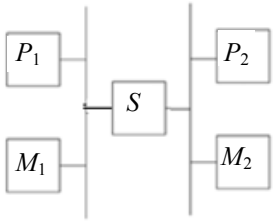

Рис. 1

В режиме дублированных вычислений организация межмашинного обмена в целях контроля осуществляется путем сравнения окончательных или промежуточных (в контрольных точках) результатов вычислений.

Время, затрачиваемое на межмашинный обмен, и возможности восстановления работоспособности комплекса после сбоев и отказов зависят от варианта реализации межмашинного обмена.

Оценка готовности дублированного комплекса. Рассмотрим варианты построения дублированного комплекса с реализацией переключателей, позволяющих организовать программно управляемый обмен и обмен на основе ПДП и двойного ПДП.

При построении марковской модели надежности восстанавливаемого комплекса с разделением нагрузки будем считать, что известны интенсивности отказов λ_p , λ_m , λ_s и восстановлений μ_p , μ_m , μ_s процессора P, модуля памяти M и переключателя S, причем восстановление производится одним ремонтником после любого отказа. Ниже представлена матрица интенсивностей переходов для марковской модели надежности исследуемой системы. Состояния системы отображаются пятью двоичными разрядами. Два старших и два младших разряда отображают состояния („0“ — исправное, „1“ — отказавшее) процессоров P и модулей памяти M соответственно первого и второго полукомплексов. Третий разряд отображает состояние переключателя S . Коды состояний записаны в шестнадцатеричном виде.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**A**	**B**	**D**	**E**	**F**	**12**	**13**	**16**	**17**	**1B**	**1F**
0	0	$2\lambda_m$	$2\lambda_p$	0	λ_s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	μ_m	0	0	λ_p	0	λ_s	0	0	λ_m	λ_p	0	0	0	0	0	0	0	0	0	0
2	μ_p	0	0	λ_m	0	0	λ_s	0	0	λ_m	0	0	0	0	λ_p	0	0	0	0	0
3	0	μ_p	μ_m	0	0	0	0	λ_s	0	0	λ_m	0	0	0	0	λ_p	0	0	0	0
4	μ_s	0	0	0	0	$2\lambda_m$	$2\lambda_p$	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	μ_s	0	0	μ_m	0	0	λ_p	0	0	0	λ_m	λ_p	0	0	0	0	0	0	0
6	0	0	μ_s	0	μ_p	0	0	λ_m	0	0	0	0	λ_m	0	0	0	λ_p	0	0	0
7	0	0	0	μ_s	0	μ_p	μ_m	0	0	0	0	0	0	λ_m	0	0	0	λ_p	0	0
9	0	μ_m	0	0	0	0	0	0	0	0	$2\lambda_p$	λ_s	0	0	0	0	0	0	0	0
A	0	μ_p	μ_m	0	0	0	0	0	0	0	λ_m	0	λ_s	0	0	λ_p	0	0	0	0
B	0	0	0	μ_m	0	0	0	0	μ_p	μ_m	0	0	0	λ_s	0	0	0	0	λ_p	0
D	0	0	0	0	0	μ_m	0	0	μ_s	0	0	0	0	$2\lambda_p$	0	0	0	0	0	0
E	0	0	0	0	0	μ_p	μ_m	0	0	μ_s	0	0	0	λ_m	0	0	0	λ_p	0	0
F	0	0	0	0	0	0	0	μ_m	0	0	μ_s	μ_p	μ_m	0	0	0	0	0	0	λ_p
12	0	0	μ_p	0	0	0	0	0	0	0	0	0	0	0	0	$2\lambda_m$	λ_s	0	0	0
13	0	0	0	μ_p	0	0	0	0	0	μ_p	0	0	0	0	μ_m	0	0	λ_s	λ_m	0
16	0	0	0	0	0	0	μ_p	0	0	0	0	0	0	0	μ_s	0	0	$2\lambda_m$	0	0
17	0	0	0	0	0	0	0	μ_p	0	0	0	0	μ_p	0	0	μ_s	μ_m	0	0	λ_m
1B	0	0	0	0	0	0	0	0	0	0	μ_p	0	0	0	0	μ_m	0	0	0	λ_s
1F	0	0	0	0	0	0	0	0	0	0	0	0	0	μ_p	0	0	0	μ_m	μ_s	0

Решение дифференциальных уравнений, составленных по матрице интенсивностей переходов, позволяет определить вероятности всех состояний комплекса и, в результате суммирования вероятностей работоспособных состояний, вычислить нестационарный коэффициент готовности (функцию готовности) $K(\tau)$ комплекса [1].

В режиме межмашинного обмена с двойным ПДП состояние комплекса относится к работоспособным в случае исправности хотя бы одного процессора и хотя бы одного модуля памяти в любом полукомплексе. При программно управляемом обмене состояние комплекса относится к работоспособным, если исправны модуль памяти и процессор одновременно хотя бы в одном полукомплексе.

Результат расчета нестационарного коэффициента готовности $K(\tau)$ ДВК представлен на рис. 2: кривые *1* и *2* соответствуют комплексу на основе межмашинного обмена без ПДП и с использованием двойного ПДП. Расчет проведен при λ_p =0,00005 ч$^{-1}$, λ_m =0,00015 ч$^{-1}$, λ_s=0,0001 ч$^{-1}$; μ_p=μ_m=μ_s=0,5 ч$^{-1}$. При тех же исходных данных в результате решения системы алгебраических уравнений найдены значения стационарного коэффициента готовности $K_г$ комплекса без ПДП и с использованием двойного ПДП, они равны соответственно 0,9999997 и 0,9999998.

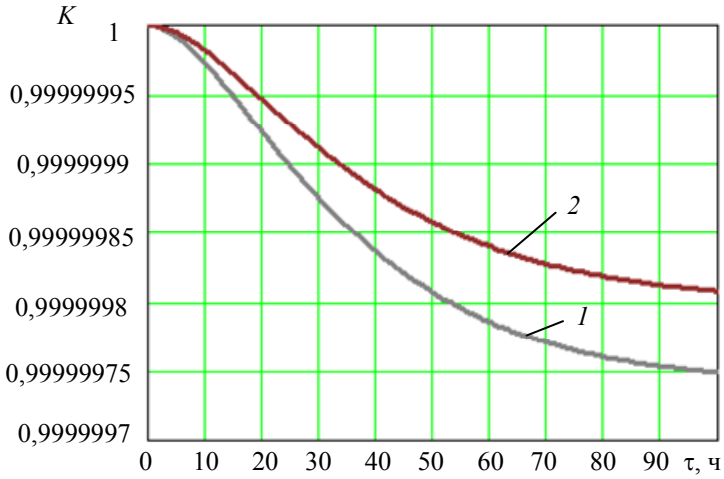

Рис. 2

Оценка эффективности межмашинного обмена в дублированном комплексе. Сравним эффективность ДВК при следующих вариантах межмашинного обмена:

— вариант B_1: обмен в режиме ПДП с конвейерным совмещением передачи данных из модуля памяти M_1 первого полукомплекса в буфер переключателя S и из него в модуль памяти M_2 второго полукомплекса по магистралям обоих полукомплексов (обмен с двойным ПДП);

— вариант B_2: обмен под управлением процессора P с конвейерным совмещением передачи данных из модуля памяти M_1 в буфер переключателя S и из него в модуль памяти M_2 по магистралям обоих полукомплексов (программно управляемый обмен с конвейеризацией);

— вариант B_3: обмен в режиме ПДП с занесением кадра из модуля памяти M_1 в буферную память переключателя S с дальнейшей передачей этого кадра (после его полного приема) в модуль памяти M_2 в режиме ПДП;

— вариант B_4: обмен под управлением процессора P с занесением кадра из модуля памяти M_1 в буфер переключателя с дальнейшей передачей этого кадра (после его полного приема) в модуль памяти M_2 под управлением процессора P .

Время межмашинного обмена при передаче кадра из L слов для вариантов B_1—B_4 вычисляется соответственно как

$$T_1 = (L+1)t + d\,,\ T_2 = (L+1)2t + D\,,\ T_3 = 2(Lt + d)\,,\ T_4 = 4tL + D\,,$$

где t — время передачи одного слова, d и D — время инициализации и установления режима ПДП и режима прерывания.

Среднее время обмена для вариантов B_1—B_4 с учетом повторных передач кадров в случае сбоев определяется соответственно как

$$T_1 = \big((L+1)t + d\big)\sum_{i=1}^{\infty} i b_1 (1-b_1)^{i-1}, \quad b_1 = e^{-\big((L+1)t+d\big)(\lambda_2+\lambda_3)};$$

$$T_2 = \big((L+1)2t + D\big)\sum_{i=1}^{\infty} i b_2 (1-b_2)^{i-1}, \quad b_2 = e^{-\big((L+1)2t+D\big)(\lambda_1+\lambda_2+\lambda_3)};$$

$$T_3 = 2\big(Lt + d\big)\sum_{i=1}^{\infty} i b_3 (1-b_3)^{i-1}, \quad b_3 = e^{-\big(2(Lt+d)\big)(\lambda_2+\lambda_3)};$$

$$T_4 = \big(4tL + D\big)\sum_{i=1}^{\infty} i b_4 (1-b_4)^{i-1}, \quad b_4 = e^{-\big(4tL+D\big)(\lambda_1+\lambda_2+\lambda_3)},$$

где λ_1, λ_2, λ_3 — интенсивности сбоев процессора P, модуля памяти M и переключателя S.

Результаты расчета среднего времени T межмашинного обмена в зависимости от длины L массива передаваемых данных (количества слов) без учета повторных передач из-за сбоев для вариантов межмашинного обмена B_1—B_4 представлены на рис. 3 соответствующими кривыми. Расчеты выполнены в предположении, что t=10^{-7} ч, d=5t ч, D=10t ч. Представленные зависимости показывают эффективность межмашинного обмена на основе двойного ПДП, причем эта эффективность растет с увеличением объемов передаваемых данных.

При функционировании ДВК в режиме дублированных вычислений, когда в полукомплексах решаются одни и те же задачи, программно управляемый обмен может быть организован без прерываний. Для этого режима результаты расчета среднего времени межмашинного обмена при различных вариантах его организации приведены на рис. 4. Анализ рисунка показывает, что существует граница целесообразности обмена с двойным ПДП.

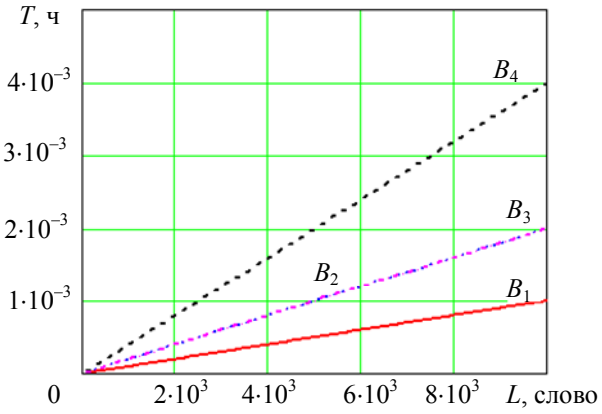

Рис. 3

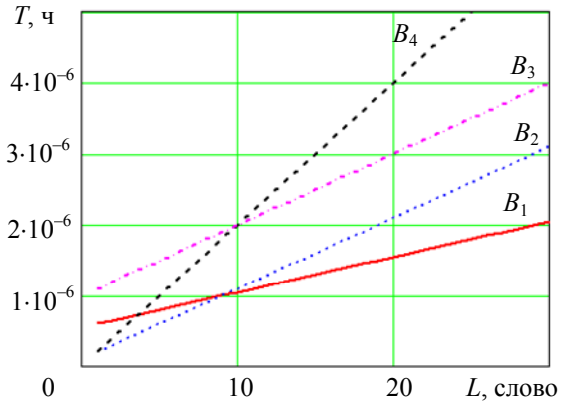

Рис. 4

При передаче больших массивов данных с использованием варианта B_1 возможно их разбиение на части (кадры) с организацией канала двойного ПДП между полукомплексами для каждого кадра. Очевидно, что в отсутствие сбоев (их пренебрежимо малой вероятности) весь массив данных наиболее быстро удается передать без его разбиения на кадры, так как передача каждого кадра связана с временными потерями на установление канала ПДП. В реальных условиях разбиение передаваемого массива данных на кадры (и соответственно уменьшение их длин) приводит, с одной стороны, к снижению вероятностей повторных передач из-за ошибок (сбоев), а с другой — к возрастанию издержек времени на организацию каналов прямого доступа. Таким образом, возникает задача оптимизации числа кадров, формируемых при передаче массива данных в режиме двойного ПДП.

Среднее время межмашинного обмена (T_1) с установлением канала двойного ПДП при разбиении передаваемого массива данных из L слов на k кадров вычисляется как

$$T_1 = \left(\left(\frac{L}{k}+1\right)t + d\right)k\sum_{i=1}^{\infty} ib(1-b)^{i-1}, \qquad b_1 = e^{-\left(\left(\frac{L}{k}+1\right)t+d\right)(\lambda_2+\lambda_3)}.$$

Зависимость величины T от числа k кадров, формируемых при передаче массива данных длиной L слов, представлена на рис. 5 для интенсивности сбоев $\lambda_1 = \lambda_2 = \lambda_3 = \lambda$, когда $\lambda = 10^{-3}$ ч$^{-1}$ и $\lambda = 10^{-4}$ ч$^{-1}$. Из графиков видно, что существует оптимальное значение k, при котором в условиях сбоев (ошибок передачи) достигается минимальное время межмашинного обмена в режиме двойного ПДП.

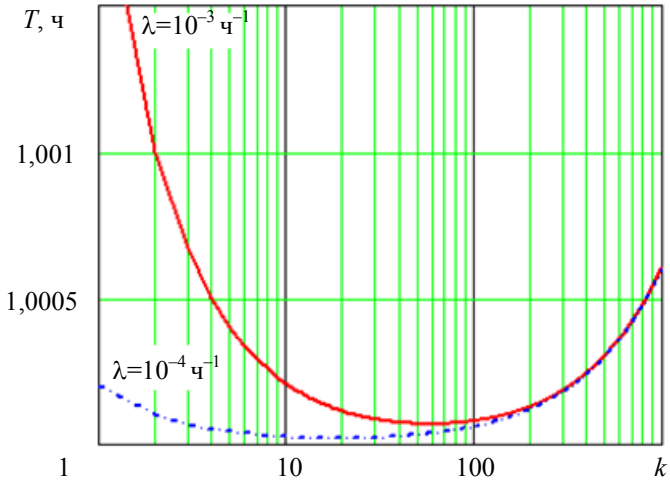

Рис. 5

Заключение. Представленные в настоящей статье результаты показывают:

— существенность влияния организации межмашинного обмена на эффективность отказоустойчивого дублированного вычислительного комплекса;

— преимущество межмашинного обмена на основе использования двойного ПДП при условии, что длина передаваемых кадров превышает некоторое граничное значение, зависящее от времени установления ПДП;

— наличие в режиме двойного ПДП оптимального числа кадров, формируемых при передаче массива данных, при котором в условиях сбоев время межмашинного обмена минимально.

СПИСОК ЛИТЕРАТУРЫ

1. *Половко А. М., Гуров С. В* Основы теории надежности: Учеб. пособие. СПб: БВХ–Петербург, 2008. 704 с.

2. Активная защита от отказов управляющих модульных вычислительных систем / *И. Б. Шубинский, В. И. Николаев, С. К. Колганов, А. М. Заяц.* СПб: Наука, 1993. 285 с.

3. *Богатырев В. А.* Отказоустойчивые многомашинные вычислительные системы динамического распределения запросов при дублировании функциональных ресурсов // Изв. вузов. Приборостроение. 1996. Т. 39, № 4. С. 81—84.

4. *Богатырев В. А.* Оценка надежности и оптимальное резервирование кластерных компьютерных систем // Приборы и системы. Управление, контроль, диагностика. 2006. № 10. С. 18—21.

5. *Богатырев В. А.* Мультипроцессорные системы с динамическим перераспределением запросов через общую магистраль // Изв. вузов СССР. Приборостроение. 1985. Т. 28, № 3. С. 33—38.

6. *Богатырев В. А.* Оптимальное резервирование системы разнородных серверов // Приборы и системы. Управление, контроль, диагностика. 2007. № 12. С. 30—36.

7. *Bogatyrev V. A.* Exchange of duplicated computing complexes in fault tolerant systems // Automatic Control and Computer Sciences. 2011. Vol. 46, N 5. P. 268—276.

8. Пат. 1679493 СССР, G 06 F 13/00. Устройство для сопряжения ведущей и ведомой ЭВМ / *В. Ф. Беззубов* и др. Б.И. 1993. № 8.

9. А.с. 1462341 СССР, G 06 F 15/16. Устройство для сопряжения ЭВМ / *В. Ф. Беззубов*. Б.И. 1989. № 8.

10. А.с. 1798946 СССР, H 05 K 10/00, G 06 F11/20. Резервированная вычислительная система / *В. Ф. Беззубов* и др. Б.И. 1991. № 35.

11. *Беззубов В. Ф.* Сравнительный анализ методов обмена в многопроцессорных системах // Вестник компьютерных и информационных технологий. 2006. № 4. С. 51—56.

12. *Голубев И. Ю, Богатырев В. А., Беззубов В. Ф.* Сравнительный анализ структур отказоустойчивых дублированных вычислительных комплексов // Информационно-измерительные и управляющие системы. 2011. Т. 9, № 2. С. 8—12.

13. *Богатырев В. А., Башкова С. А., Беззубов В. Ф.* Надежность дублированных вычислительных комплексов // Науч.-техн. вестн. СПбНИУ ИТМО. 2011. Вып. 6. С. 74—78.

Сведения об авторах

Владимир Анатольевич Богатырев — д-р техн. наук, профессор; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; E-mail: Vladimir.bogatyrev@gmail.com

Иван Юрьевич Голубев — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники; E-mail: www.golubev@mail.ru

Владимир Федорович Беззубов — аспирант; Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кафедра вычислительной техники

Рекомендована кафедрой вычислительной техники

Поступила в редакцию 23.11.11 г.