УДК 519.724.2

НАРАЩИВАЕМЫЕ МНОГОКОЛЬЦЕВЫЕ НЕКОММУТИРУЕМЫЕ СЕТИ СВЯЗИ ДЛЯ МНОГОПРОЦЕССОРНЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

В. С. Подлазов

Институт проблем управления им. В. А. Трапезникова, г. Москва

Предложена концепция построения мультикольцевой сети связи, пропускная способность которой может быть прямо пропорциональной числу узлов сети при малом числе колец.

Памяти И. В. Прангишвили посвящается

ВВЕДЕНИЕ

В настоящее время среди многопроцессорных вычислительных систем (МВС) высокой производительности наиболее перспективными по критерию производительность/стоимость считаются симметричные МВС с общей разделяемой памятью (SMP-системы) и неоднородным доступом к ней (SMP NUMA-системы) [1—3]. Они могут содержать несколько сотен процессоров. Для них необходимы пакетные коммутаторы с малой задержкой и хорошей наращиваемостью [2, 3].

К ним относятся такие серверы, как NUMA-Q 2000 фирмы "Sequent", AV25000 фирмы "Data General", Starfire Ultra Enterprise 10000 фирмы SUN, Cray Origin 2000 фирмы "Silicon Graphics", Octa SCALE фирмы NCR, Exemplar X-Class фирмы "Hewlett Packard" [4, 5], содержащие десятки и сотни процессоров.

Одно из важнейших требований с таким серверам заключается в наличии свойств RAS (Reliability, Availability, Scalability) — надежности, непрерывной готовности и наращиваемости. Наращиваемость понимается как линейный рост производительности системы при увеличении числа процессоров и объема памяти.

Компромиссное разрешение этих противоречивых требования нашлось в применении архитектуры ccNUMA (cache-coherent Non-Uniform Memory Access). Для согласования кэшей третьего уровня используется интерфейс SCI (Scalable Coherent Interface — IEEE P1596), который требует сети связи с пропускной способностью не менее 1 Гбайт/с. Этот интерфейс используется также в системах только с кэшевой памятью (COMA) и с рефлексивной памятью (RM) [2, 3].

Простейшей сетью, в которой применяется интерфейс SCI, является однонаправленное цепочечное кольцо. Оно обеспечивает значительно более высокую скорость передачи, чем шина, так как в нем используются только двухточечные физические соединения. Подобное кольцо со скоростью передачи 1 Гбайт/с применяется в сервере NUMA-Q 2000. Оно функционирует как синхронное сегментированное кольцо. Целое число сегментов циркулирует по кольцу как контейнеры для переноса кадров данных. В нем любой приемник удаляет адресованные им кадры из сегментов, обеспечивая возможность одновременной передачи кадров по кольцу многими источниками. Такое кольцо можно называть *кратным*. Кратность выражается в пространственном распараллеливании процесса передачи кадров между узлами сети. Поэтому кратное кольцо имеет много большую пропускную способность, чем применяемые в локальных сетях кольца с передачей жезла (FDDI, Token Ring). В них кадры удаляются из кольца источниками, и передача кадров осуществляется последовательно по узлам. Пропускная способность кратного кольца больше скорости передачи благодаря пространственному распараллеливанию передачи.

Кольцо легче всего удовлетворяет требованию надежности, так как в нем кадры могут передаваться без промежуточной буферизации в узлах, и можно иметь узлы с очень простой структурой. В этом случае кольцо обеспечивает также минимальное время доставки кадров по сети.

В кольце легко обеспечить устойчивость к отказам узлов путем введения избыточных узлов, используемых в режиме скользящего резервирования. Целостность кольца, нарушаемая при отказе любого узла, восстанавливается путем обвода этого узла в кольце.

Кольцо позволяет применять централизованную синхронизацию, благодаря чему еще более упрощать узлы. Они могут вообще не иметь в тракте кольца элементов памяти (сдвиговых регистров), а только коммутирующие элементы [6, 7]. Оптоэлектронные коммутаторы

сигналов позволяют вообще не производить в узлах преобразований "свет — электричество — свет" при генерации и ретрансляции сигналов и иметь светоизлучающие элементы только в центре синхронизации и генерации световых сигналов. При этом надежность кольца приближается к надежности шины [7] при сохранении пространственного распараллеливания передачи данных.

Два встречных кратных кольца со скоростью передачи в каждом 0,5 Гбайт/с, применяются в сервере AV25000 (рис. 1). Здесь любой источник передает кадры данных в кольцо с кратчайшим маршрутом до приемника, что обеспечивает в 4 раза большую пропускную способность, чем в однотипном одиночном кольце. Фирма "Data General" планирует использование кольцевых сетей с числом колец больше двух.

Для сетей связи в SMP-системах основной моделью передачи пакетов данных является *сетевая модель*, при которой имеет место произвольное отображение индивидуальных адресов от источников к приемникам. Кольца идеально подходят и для передачи групповых и широковещательных кадров.

Наращиваемость в этой модели передачи означает рост пропускной способности пропорционально числу узлов. Одно- или двухкольцевые сети связи обеспечивают наращиваемость только за счет запаса пропускной способности, т. е. обладают ограниченной наращиваемостью. При этом увеличение пропускной способности пропорционально числу колец не снимает указанного ограничения, а только ослабляет его. Аналогичное ограничение имеет место и для сетей связи других конфигураций. Например, в новейшем сервере Origin Altix 3000

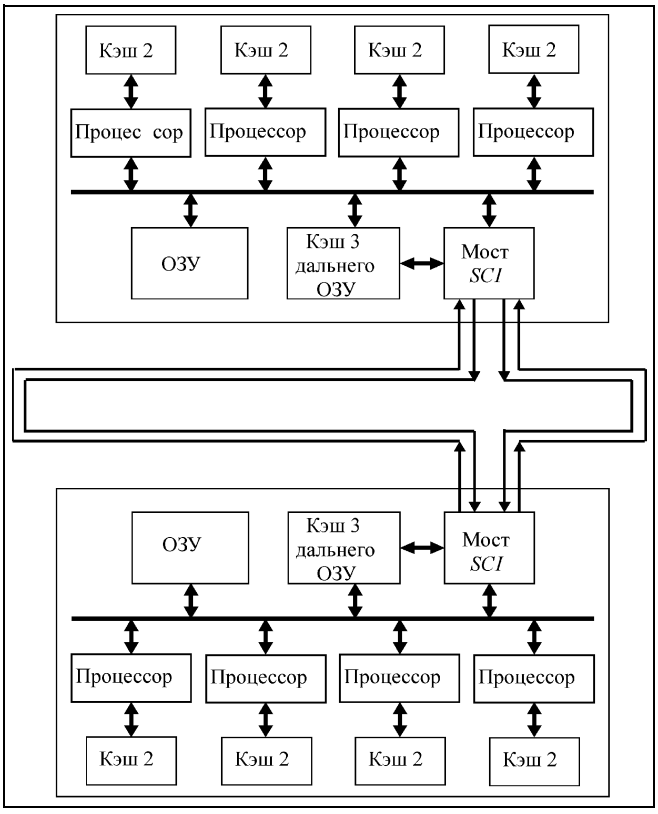

Рис. 1. Архитектура сервера AV25000

Таблица 1

Пропускная способность сети связи сервера Altix 3000

Число процессоров	Пропускная пособность, (Гбайт/с)	Пропускная способность на один процессор (Гбайт/с)
8	12,8	1,6
16	12,8	0,8
32	25,6	0,8
64	25,6	0,4
128	51,2	0,4

[8] используется сеть связи переменной структуры, которая в зависимости от числа процессоров N имеет вид кольца ($N = 8$), звезды ($N = 16$), $2D$-гиперкуба ($N = 32; 64$), "толстого" дерева ($N = 128$). Пропускная способность этой сети представлена в табл. 1.

Возникает задача построения мультикольцевых сетей (*мультиколец*), пропускная способность которых растет с ростом *числа колец* быстрее, чем линейно, и может быть пропорциональной числу узлов при *малом числе* колец. Для сохранения высоких надежностных и временных характеристик такое мультикольцо должно состоять из *некоммутируемых* колец. Предполагается его строить, применяя кольца разной топологи, т. е. кольца с разными последовательностями соединений узлов.

1. ТОПОЛОГИЯ МУЛЬТИКОЛЕЦ И СТАТИЧЕСКИЕ РАСПИСАНИЯ

Предполагается, что мультикольцо состоит из кратных колец, имеющих постоянный шаг приращения номеров соседних узлов. Формально это можно выразить следующим образом.

Предположим, что узлы перенумерованы целыми числами из $[0, N-1]$, где N — число узлов. Пусть мультикольцо состоит из $m \geqslant 1$ колец. В j-м кольце номера узлов образуют последовательность ${}^{j}X_{i+1} = ({}^{j}X_i + {}^{j}S) \bmod N$, где $X_i \in [0, N-1]$, $i = 0, 1, ,$ и ${}^{j}S > 0$ — шаг j-го кольца, $1 \leqslant j \leqslant m$. Мультикольцо задается набором шагов $S_m = ({}^{1}S = 1, {}^{2}S, ..., {}^{m}S)$, где ${}^{1}S < {}^{2}S < ... < {}^{m}S$.

Кольцо с положительным шагом ${}^{i}S \geqslant N/2$ будем также называть встречным кольцом с отрицательным шагом ${}^{j}S = -(N - S)$. Пример мультикольца на 16 узлов из 4-х колец с набором шагов $S_4 = (1, 3, 13, 15) = (\pm 1, \pm 3)$ приведен на рис. 2. В дальнейшем кольцо с шагом ${}^{j}S$ будем называть кольцом ${}^{j}S$, а его дугу — дугой ${}^{j}S$.

Будем различать два вида колец — обычные и расщепленные. Обычное кольцо имеет взаимно простые числа N и ${}^{j}S$. В расщепленном кольце числа N и ${}^{j}S$ имеют наибольший общий делитель d, и последовательность X_i разделяется на d непересекающихся последовательностей ${}^{j}X_i$ с периодом N/d, $0 \leqslant j \leqslant d-1$, ${}^{j}X_0 = j$, которые в совокупности содержит все номера из $[0, N-1]$. Физически расщепленное кольцо состоит из d миниколец по N/d узлов в каждом.

В некоммутируемом мультикольце любой пакет от узла отправления до узла назначения доставляется толь-

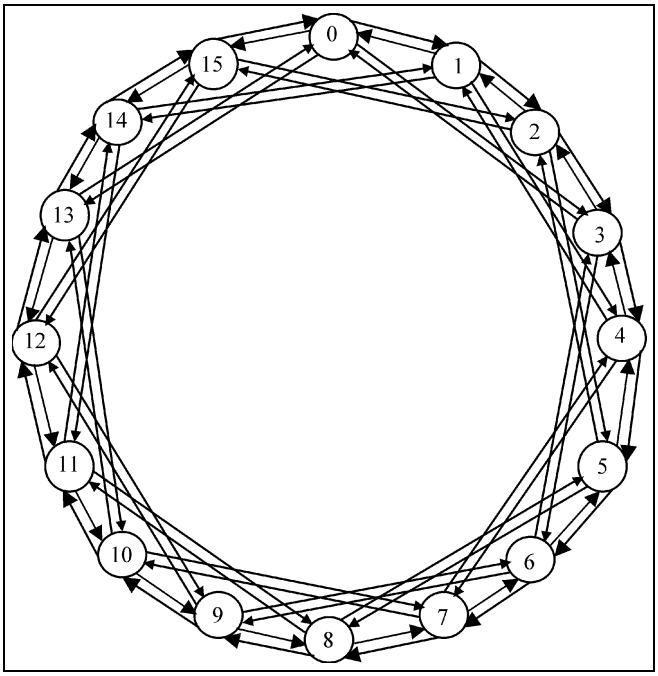

Рис. 2. Мультикольцо S_4 = (1, 3, 13, 15) = (±1, ±3)

ко по одному кольцу. Выбор кольца для передачи зависит от длины пути (числа дуг) по этому кольцу. Правило выбора задается *статическим расписанием*, которое определяет вероятность назначения маршрута заданной длины в каждое кольцо. В сетевой задаче узлы сети генерируют потоки пакетов одинаковой интенсивности с заданными распределением маршрутов по их длинам. В случае однородных узлов все узлы имеют одинаковое распределение *g*.

Пусть в кольце ${}^jS$ существует кратчайший путь от узла отправления ${}^jX_i$ в узел назначения ${}^jX_{(i+1)\bmod N}$. Длина такого пути равна *l*. Длина пути по кольцу 1 называется длиной маршрута между узлами отправления и назначения. Пусть ${}^jl_r$ задает длину пути в ${}^jS$ маршрута длины *r*, а $s_{g,r}$ — вероятность появления маршрута длины *r* в распределении *g*. Теперь расписание $R_g(S_m)$ для мультикольца S_m — это набор *m* строк ${}^jp_g = ({}^jp_{g,1}, ..., {}^jp_g, r, ..., {}^jp_{g,N-1})$, в котором *r*-й элемент *j*-й строки задает вероятность назначения маршрута длины *r* в кольцо ${}^jS$.

Используя введенные обозначения, можно определить:

- ${}^ju_g = \sum_{r=1}^{N-1} s_{g,r}\,{}^jp_{g,r}$ — вероятность передачи маршрута длины *r* по кольцу ${}^jS$;
- ${}^jL_g = \sum_{r=1}^{N-1} {}^jl_r\,{}^ju_g$ — условную среднюю длину маршрутов по кольцу ${}^jS$;
- ${}^j\bar{L}_g = {}^jL_g/{}^ju_g$ — среднюю длина маршрутов в кольце ${}^jS$.

2. ПРОПУСКНАЯ СПОСОБНОСТЬ И ЕМКОСТЬ МУЛЬТИКОЛЕЦ С ОДНОРОДНЫМИ УЗЛАМИ

Пропускная способность *W* одиночного кратного кольца задается выражением $W = cv$, где v — скорость передачи по кольцу (бит/с), а *c* — емкость кольца, задаваемая средним числом пакетов, переданных в одном сегменте за один оборот по кольцу в условиях максимальной нагрузки. При однородных узлах емкость одного кольца задается выражением

$$c_g = N/\bar{L}_g, \tag{1}$$

и $c_g = 2$ для любых распределений *g* длин маршрутов [9—12].

Мультикольцо, состоящее из двух встречных колец $S_2 = (1, N-1) = (\pm 1)$, исследовалось при различных симметричных невозрастающих распределениях длин маршрутов — равномерном [9], треугольном, гиперболическом, обратных квадратов, показательном с основанием $0 < q < 1$ и одноступенчатом с высотой ступени *A* и ее полушириной *D* [11, 12]. Эти распределения определены в табл. 2 для четных *N* и симметричных длин маршрутов, когда маршрут длины *r* имеет ту же вероятность появления, как и маршрут длины $N - r$.

Емкость мультикольца $C_g(N)$ рассчитывается как сумма емкостей обоих колец. Асимптотические значения этой емкости представлены в табл. 2 для мультикольца с однородными узлами. Табл. 2 показывает две зависимости величины емкости мультикольца от числа узлов — слабую и сильную. Слабая зависимость — это не более чем логарифмический рост с увеличением числа узлов. Сильная зависимость — это рост не медленнее, чем корень квадратный от числа узлов.

Слабая зависимость имеет место для однородного, линейного, гиперболического и одноступенчатого (при $B \leqslant O(\ln N)$) распределений. Сильная зависимость имеет место для квадратичного, показательного и одноступенчатого при ($B = O(N/\ln N)$) распределений. Наличие сильной зависимости практически полностью решает проблему наращиваемости мультикольца уже в случае двух колец. Наличие слабой зависимости вынуждает исследовать характер поведения емкости мультикольца при числе колец $m > 2$.

В общем случае при $m > 2$ появляется возможность выбора состава колец и оптимизации расписания для увеличения пропускной способности мультикольца. Она задается как $W = C_g(m, N)v$, где $C_g(m, N)$ — емкость мультикольца. Но как определить эту емкость? Ее можно определить как сумму емкостей ${}^jc_g$ колец ${}^jS$, т. е. как

$$C_g(m, N) = \sum_{j=1}^{m} {}^jc_g. \tag{2}$$

При заданном расписании емкости ${}^jc_g$ рассчитываются по формуле (1). Однако имитационное моделирование показало [9—12], что реальная емкость мультикольца совпадает с формулой (2) только при бесконечном выходном буфере в каждом узле или при симметричном наборе колец и симметричном по кольцам расписании.

Причиной расхождения является недогрузка отдельных колец вследствие нехватки в выходных буферах узлов кадров с некоторыми длинами маршрутов для передачи по некоторым кольцам. Была доказана теорема [9, 11, 12], что реальная емкость правильно задается так называемой эффективной емкостью $\tilde{C}_g(m, N)$. Она по-прежнему определяется как сумма по кольцам среднего числа кадров, доставленных за один оборот сегмента в каждом кольце, и измеряется как реальная емкость при имитационном моделировании, В случае однородных узлов эффективная емкость задается следующей формулой:

$$\tilde{C}_g(m, N) = N / \max_{1 \le j \le m} {}^{j}L_g. \qquad (3)$$

Заметим, что в формулу (3) входит условная средняя длина ${}^{j}L_g$, а не средняя длина ${}^{j}\overline{L}_g$, как в формулу (1).

Входящие в состав узла процессоры могут генерировать независимые потоки кадров с разными распределениями длин маршрутов g с удельными весами f_g, которые составляют некоторое распределение d. Вследствие последовательного прохождения кадров через общую шину узла вероятность маршрута длины r задается как $p_{d,r} = \sum_g f_g p_{g,r}$, а эффективная емкость мультикольца задается выражением:

$$\tilde{C}_d(m, N) = N / \max_{1 \le j \le m} \sum_g f_g {}^{j}L_g. \qquad (3')$$

Имитационное моделирование показало, что формулы (3) и (3') выполняются с высокой точностью для любых наборов колец S_m, для всех исследованных распределений g и для любых построенных в экспериментах расписаний R.

Возможность рассчитать емкость мультикольца открывает возможность решения для него оптимизационной задачи — поиска набора колец S_m и расписа-

Таблица 2

Распределения *g* длин маршрутов

Распределение g	Вероятность маршрута длины r	Нормировочная функция	Асимптотическая емкость мультикольца $S_2 = (\pm 1)$
Равномерное u	$s_{u,r} = 1/(N-1)$	—	$C_u(N) \to 8$ $N \to \infty$
Треугольное t	$s_{t,r} = t(N)(N/2 + 1 - r)$	$t(N) = (2\sum_{r-1}^{N/2}(N - r + 1) - 1)^{-1}$	$C_t(N) \to 12$ $N \to \infty$
Гиперболическое h	$s_{\mathrm{h},\mathrm{r}} = h(N)/r$	$h(N) = (2(\sum_{e-1}^{N/2} 1/r - 1) + 2/N)^{-1}$	$C_h(N) \to 4\ln N$ $N \to \infty$
Обратных квадратов s	$s_{\mathrm{s},\mathrm{r}} = s(N)/r^2$	$s(N) = (2(\sum_{r-1}^{N/2} 1/r^2 - 1) + 4/N^2)^{-1}$	$C_s(N) \underset{N \to \infty}{\to} \frac{\pi^2 N}{3\ln N}$
Показательное e (q)	$s_{t,r} = e(N)\, q^r$	$e(N) = (2(\sum_{r-1}^{N/2} q^r - 1) + q^{N/2})^{-1}$	$C_e(N) \to 2(1-q)N$ $N \to \infty$
Одноступенчатое o (A, D)	$s_{o,r} = \begin{cases} o(N)A, & 1 \leqslant r \leqslant D < N \\ o(N), & r > R \end{cases}$	$o(N) = (2(A-1)\, D + N)^{-1}$	$C_o(N) \to 8(B+1)$, $N \to \infty$, если $B = (N/2 - D)/AD$

Таблица 3

Кратчайшее *r* и выровненное *R* расписания для набора колец (±1, ±3) с реальными емкостями $\tilde{C}_u = 19{,}9$ и $\tilde{C}_u = 21{,}6$, соответственно

r	1	2	3	4	5	6	7	$8 = -8^*$
Кольцо 1	1,0	1,0	0,0	0,5	1,0	0,0	0,0	0,25
Кольцо 3	0,0	0,0	1,0	0,0	0,0	1,0	0,0	0,25
Кольцо −3	0	0	0	0,5	0	0,0	1,0	0,25
Кольцо −1	0	0	0	0	0	0,0	0,0	0,25
R	$15 = -1^*$	$14 = -2^*$	$13 = -3^*$	$12 = -4^*$	$11 = -5^*$	$10 = -6^*$	$9 = -7^*$	$\mathbf{8 = -8^*}$
Кольцо 1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	**0,125**
Кольцо 3	0,0	0,0	0,0	0,5	0,0	0,0	1,0	**0,375**
Кольцо −3	0,0	0,0	1,0	0,0	0,0	1,0	0,0	**0,375**
Кольцо −1	1,0	1,0	0,0	0,5	1,0	0,0	0,0	**0,125**

*Равенство по модулю 16.

ния $R_g(S_m)$, которые обеспечивают наибольшую емкость $\tilde{C}_g(m, N)$ для любых распределений длин маршрутов g. Эту задачу можно сравнить с задачей составления расписания для выполнения некоторой работы на N процессорах m типов за минимальное время [13]. При этом доля работы, выделяемая для процессора типа j, это аналог величины $\sum_r^j p_{g,r}$. Такое сравнение показывает, что данная оптимизационная задача является NP-полной. Она решалась путем перебора состава колец S_m для мультиколец с числом узлов $N \leqslant 64$ и стохастической оптимизацией при построении расписаний. Применялась следующая методика.

При малых $N \leqslant 32$ использовался полный перебор всевозможных наборов колец. В две стадии отбирались 16 наборов с максимальными значениями эффективной емкости. На первой стадии рассматривалось только "кратчайшее" расписание с передачей по кратчайшим путям. На второй стадии из него строилось "выровненное" расписание путем перераспределения назначений маршрутов по кольцам с увеличением минимальной кольцевой нагрузки. Для этого применялась эвристическая процедура сложности $O(m)$. В ней многократно выравнивались кольцевые нагрузки с минимальным и максимальным значениями при условии, чтобы новое значение было меньше предыдущего максимального. Это выравнивание проводилось путем расщепления некоторых маршрутов на несколько долей для передачи по разным кольцам. В табл. 3 приведены "кратчайшее" и "выровненное" расписания для мультикольца, показанного на рис. 2.

Для проверки оптимальности полученных таким образом расписаний и для дальнейшей их оптимизации к некоторым из оптимальных наборов колец применялась процедура отжига. Метод отжига — это эвристический метод стохастической оптимизации [14]. Процедура оптимизации расписания по этому методу осуществляется следующим образом. Текущее расписание подвергается случайному изменению и оценивается по формуле (3). Если оно оказывается лучше текущего расписания, то заменяет его в качестве текущего и в качестве искомого расписаний. Если оно оказывается хуже текущего расписания, то заменяет его с некоторой вероятностью. Последняя замена необходима, чтобы не застрять в локальном экстремуме. Вероятность такой замены падает, а время ее использования растет, по мере осуществления процедуры отжига. Исходное расписание может быть любым.

С помощью процедуры отжига, как правило, удавалось повысить эффективную емкость на десятки процентов по сравнению с выровненным расписанием в области $m \approx N/2 \pm N/3$. В области малых m процедура отжига давала небольшой прирост эффективной емкости.

При числе узлов в диапазоне $32 < N \leqslant 64$ полный перебор применялся только к наборам с четным числом колец, которые состоят из различных пар встречных колец с одинаковым шагом. Оптимизация расписаний для каждого набора проводилась также, как и при малых N.

Далее рассматриваются зависимости $\tilde{C}_g(m, N)$ для оптимальных наборов колец с кратчайшими и выровненными расписаниями. Для них наборы и расписания

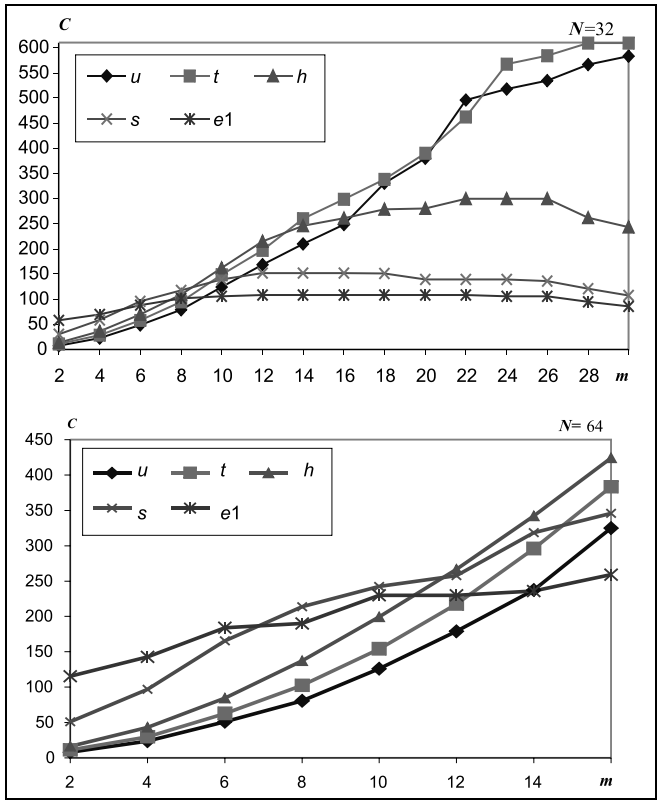

Рис. 3. Зависимость $C = \tilde{C}_g(m, N)$ для "выровненного" расписания ($e1 = e(0,1)$) при $N = 32$ и $N = 64$

зависят не только от N и m, но и от g, т. е. они могут быть разными для различных распределений длин маршрутов. Поэтому приведенные ниже зависимости являются лишь теоретически достижимыми.

При $m < N/4$ для "кратчайшего" расписания $\tilde{C}_u(m, N) \approx m^2$ при $m < N/4$ для равномерного распределения и $\tilde{C}_t(m, N) \approx 1{,}5m^2$ для треугольного распределения. Для "выровненного" расписания на рис. 3 приведена зависимость $\tilde{C}_g(m, N)$ для $N = 32$ и $N = 64$. По ним можно сделать вывод, что для любого числа узлов существует такое число колец, при котором емкость для всех распределений не хуже, чем у показательного. Поэтому она растет линейно с увеличением числа узлов.

Приведенные ранее наблюдения позволяют предположить, что путем увеличения числа колец оптимальной топологии можно обеспечить полную наращиваемость системы. Однако остается открытым вопрос, а существуют ли наборы колец, которые по своим характеристикам близки к оптимальным наборам во всех распределениях. Ответ на него положительный, если допускать отклонения от оптимальных значений в среднем на 10...20 %. В этих условиях удается найти даже семейство наращиваемых мультиколец. В них любой набор колец содержит наборы с меньшим число колец. В табл. 4 приведены наборы S_m этого семейства с указанием (в процентах) среднего (по рассматриваемому множест-

ву распределений) отклонения эффективной емкости от ее значений для оптимальных наборов.

Для мультиколец с числом узлов $N = 128$ и $N = 256$ это семейство было по аналогии расширено благодаря мультикольцам с наборами $S_{14} = (\pm1, \pm2, \pm3, \pm5, \pm7, \pm9, \pm11, \pm13)$ и $S_{16} = (\pm1, \pm2, \pm3, \pm5, \pm7, \pm9, \pm11, \pm13)$. Для последнего мультикольца проведен решающий эксперимент (имитационное моделирование с "выровненным" расписанием), который показал, что реальная емкость для всех рассмотренных распределений g находится в коридоре $0{,}8N \leqslant \tilde{C}_g(16{,}256) \leqslant 2{,}6N$.

Теперь основной научный результат для некоммутируемого мультикольца с однородными узлами можно сформулировать так: *полную наращиваемость некоммутируемого мультикольца можно обеспечить путем использования семейства наращиваемых мультиколец. Для обеспечения роста эффективной емкости мультикольца прямо пропорционально числу N число колец m должно расти как* $m \leqslant \sqrt{N}$.

Здесь необходимо отметить, что реальное число узлов колец может быть существенно меньше $\sqrt{N}$, если в исходной сети имеется некоторый запас пропускной способности. Так, сервер NUMA-Q 2000 при 8-ми узлах (32-х процессорах) на широком классе задач имеет загрузку кольца (используемую долю пропускной способности) $\rho \leqslant 0{,}35$. Фирма "Sequent" позиционирует данный сервер на число процессоров до 128 (до 32-х узлов). Поэтому возникает следующая задача. Пусть одно кольцо при числе узлов M имеет загрузку $\rho_0 < 1$. При каком числе колец m мультикольцо будет иметь загрузку ρ, если число процессоров увеличится до N?

Полная загрузка кольца достигается, если используется вся его пропускная способность, характеризуемая его емкостью $c = 2$. Для достижения такого состояния число узлов необходимо повысить до величины M/ρ_0. Для того, чтобы мультикольцо с $Z \geqslant M$ кольцами было загружено до максимума, его эффективная емкость должна быть повышена до величины $\tilde{C}(m, Z) = 2Z\rho_0/M$. Это достигается при числе колец $m \leqslant \sqrt{2Z\rho_0/M}$. Тогда загрузка ρ_0 будет достигаться при числе узлов $N = Z\rho_0$. Поэтому $m \leqslant \sqrt{2N/M}$. Аналогично, загрузка ρ будет достигаться при $N = Z\rho$, и тогда $m \leqslant \sqrt{2N\rho_0/M\rho}$.

Использование в мультикольце кольца сервера NUMA-Q 2000 позволяет иметь загрузку $\rho \leqslant 0{,}35$ для числа узлов $N = 64$ (256 процессоров) при числе колец $m = \sqrt{128/8} = 4$, а для $N = 256$ (1024 процессоров) — $m = \sqrt{512/8} = 8$. Это означает, что можно использовать мультикольца с наборами колец $(\pm1, \pm3)$ и $(\pm1, \pm2, \pm3, \pm7)$ соответственно. Аналогично, загрузка $\rho \leqslant 0{,}7$ при $N = 256$ достигается при $m = 6$ и можно использовать набор колец $(\pm1, \pm2, \pm3)$.

3. УСЛОВИЯ ДОСТИЖИМОСТИ ТЕОРЕТИЧЕСКИХ ХАРАКТЕРИСТИК

В § 2 были выявлены максимально возможные значения пропускной способности некоммутируемого наращиваемого мультикольца, выраженные через его реальную или эффективную емкости. В дальнейшем не будем их различать и будем именовать емкостью мультикольца. Однако даже теоретическая достижимость полученных значений ограничивается рядом факторов.

Один из них состоит в том, что "выровненные" или оптимальные расписания различны для разных распределений g, а одинаковые "кратчайшие" расписания обеспечивают на десятки процентов меньшую емкость. Имеется экспериментальный факт, что оптимизация расписания для равномерного распределения u заметно повышает емкость для всех распределений по сравнению с "кратчайшим" расписанием.

Другой фактор — недостаточный объем выходного буфера кадров в каждом узле и возможность параллельной или последовательной выдачи кадров из него в каждое кольцо. При имитационном моделировании этот буфер поддерживается постоянно заполненным потоком кадров с заданным распределением g. Оказалось, что емкость мультикольца сильно зависит от размера буфера V. На рис. 4 приведен пример такой зависимости для мультикольца с однородными узлами при $N = 64$ и $m = 8$. В нем каждый узел генерирует смесь одинаковых потоков пакетов с равномерным u и показательным e распределениями. Горизонтальная асимптота задает емкость мультикольца, рассчитанную по формуле (3'). Это

Таблица 4

Наращиваемые мультикольца с однородными узлами

N	m				
	4	6	8	10	12
16	±1, ±3 2 %	±1, ±2, ±3 4 %	Мультикольца не рассматривались		
32	±1, ±3 2 %	±1, ±2, ±3 7 %	±1, ±2, ±3, ±7 8 % ±1, ±2, ±3, ±5 11 %		
64	±1, ±3 4 %	±1, ±2, ±3 7 %	±1, ±2, ±3, ±7 11 % ±1, ±2, ±3, ±5 10 %	±1, ±2, ±3, ±5, ±7 11 %	±1, ±2, ±3, ±5, ±7, ± 9 16 %

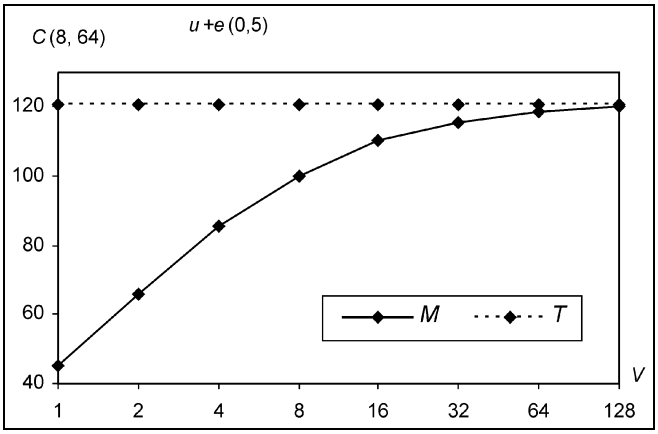

Рис. 4. Зависимость емкости от размера выходного буфера узла: T — теоретическое, M — модельные значения

означает, что мало обеспечить высокую емкость мультикольца, надо еще уметь создать поток кадров, который ее "наполнит". Все имитационное моделирование проводилось с буфером в 100 кадров и с параллельной выдачей кадров в кольца.

Следующий фактор заключается в способе выдачи кадров из буфера в кольца — параллельном или последовательном. Все имитационное моделирование проводилось для параллельного способа, который обеспечивает достижение максимальной емкости.

Наконец, еще один фактор — допущение об однородности узлов. Для преодоления этого ограничения в работах [12, 15] был исследован простейший случай неоднородных узлов. Неоднородность задавалась долей f_g числа узлов, которые имеют распределение g, и предполагалось их равномерное размещение на мультикольце. Эти допущения трактовались так, что произвольный узел имеет распределение g с вероятностью f_g, что справедливо при быстрой смене вида распределений в узлах. Доказана теорема, что емкость мультикольца с неоднородными узлами задается формулой:

$$\tilde{C} = N \max_{1 \le j \le m} G(i), \qquad (4)$$

где $G(j) = \sum_g f_g / {}^j u_g / \sum_g f_g {}^j L_g / {}^j u_g = \sum_g f_g / {}^j u_g / \sum_{r=1}^{N-1} {}^j l_r$.

Для случая неоднородных узлов имитационное моделирование показало, что раздельная оптимизация расписаний для каждого распределения g зачастую не приводит к повышению эффективной емкости мультикольца или даже приводит к ее снижению. Этот результат объясняется тем, что при раздельной оптимизации расписаний для сравнения вариантов используются формулы (3) или (3'), а не формула (4). Для совместной оптимизации расписаний требовалось полное перепрограммирование метода отжига. Поэтому он был заменен на более универсальный и более быстродействующий метод стохастической оптимизации — метод генетических алгоритмов [16, 17].

Однако совместная оптимизация расписаний для различных составов типов узлов может проводиться по формуле (4), только если она имеет приемлемую точность. Для ее проверки было выполнено имитационное моделирование с постоянным размещением различных узлов в отельном сеансе моделирования и с усреднением результатов по небольшому ряду сеансов. Точность формулы (4) оказалось не очень высокой — 10...15 %. Однако расхождения обычно не превосходили дисперсии [12, 15].

Моделирование показало, что для любого числа узлов $N \leqslant 256$ и любого размещения f_g разнотипных узлов существует наращиваемое мультикольцо с числом узлов $m \leqslant \sqrt{N}$, которое имеет емкость $C = O(1)N$. Более того, это свойство обеспечивается даже при использовании "кратчайшего" расписания, что исключает зависимость расписания от f_g.

В табл. 5 дан пример зависимости емкости мультикольца с неоднородными узлами от числа колец и числа узлов в случае, когда равномерное распределение имеет заданную долю f_u, а треугольное, гиперболическое, обратных квадратов и показательное с основанием $q = 0{,}5$ распределения имеют равные доли $(1 - f_u)/4$. Для сравнения необходимо помнить, что одиночное кольцо имеет емкость, равную 2.

4. НЕКОТОРЫЕ ВОПРОСЫ ПРАКТИЧЕСКОЙ РЕАЛИЗАЦИИ

В смысле свойств мультикольца очень удобно выбирать число резервных узлов таким, чтобы общее число узлов было простым. Например, 17 при 16-ти рабочих узлах, 37 при 32-х рабочих узлах, 67(71) при 64-х рабочих узлах, 131(137) при 128-ми рабочих узлах и т. д. В этом случае все кольца сети не имеют расщепления на миникольца, что дает возможность просто осуществлять центральную синхронизацию мультикольца. Наличие

Таблица 5

Емкость наращиваемых мультиколец при "кратчайшем" расписании

N	m	S_m	$f_u = 1$	$f_u = 0{,}75$	$f_u = 0{,}5$	$f_u = 0{,}25$	$f_u = 0$
64	2	(±1)	8	9	11	14	19
64	4	(±1, ±3)	19	20	22	26	34
64	6	(±1, ±2, ±3)	37	39	42	49	61
256	6	(±1, ±2, ±3)	38	42	49	59	77
256	8	(±1, ±2, ±3, ±7)	56	81	109	108	114
256	10	(±1, ±2, ±3, ±5, ±7)	89	117	162	197	172

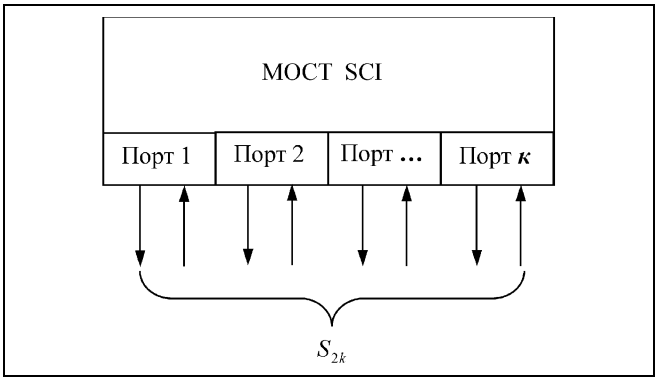

Рис. 5. Многопортовый мост SCI

резервных узлов обеспечивает повышение отказоустойчивости системы благодаря применению скользящего резервирования.

Исследования показали, что удельная (на узел) емкость для простого числа узлов $N > 2^n$ выше, чем для числа узлов $N = 2^n$ [12, 15]. Кроме того, отказ избыточных узлов не ведет к уменьшению удельной емкости по сравнению с нерезервированными конфигурациями. Это означает, что отказ небольшого числа узлов не ведет к снижению пропускной способности сети для множества рабочих узлов, что обеспечивает $(N - 2^n)$-узловую отказоустойчивость мультикольца.

Применение некоммутируемых мультиколец в МВС типа SMP NUMA требует только увеличения числа коммуникационных портов на внешней стороне SCI-мостов (рис. 5). При этом сам протокол SCI совершенно не затрагивается, а весь эффект радикального повышения пропускной способности сети связи достигается только за счет выбора последовательности соединения узлов в кольцах.

ЗАКЛЮЧЕНИЕ

Для многопроцессорных вычислительных систем (МВС) с общей разделяемой памятью и сотнями узлов предложен новый оригинальный класс коммуникационных сетей — некоммутируемые мультикольца — в виде набора простых кратных кольцевых каналов с разными топологиями. Сети этого класса позволяют сочетать низкие задержки доставки пакетов и высокую пропускную способность с хорошей наращиваемостью и отказоустойчивостью.

Построены наращиваемые наборы мультиколец и статические расписания для них, которые обеспечивают для сетевой модели передачи пропускную способность, пропорциональную числу узлов N при числе колец не больше, чем $\sqrt{N}$. Показано, что пропускная способность современных колец и структура узлов в них позволяют иметь 4—8 колец при числе узлов в диапазоне 64—256 (числе процессоров 256—1024).

Область применимости некоммутируемых мультиколец не ограничивается МВС с общей памятью. Они эффективно могут использоваться и в кластерных МВС для межкластерной связи [18].

ЛИТЕРАТУРА

1. *Андреев А., Воеводин В., Жуматий С.* Кластеры и суперкомпьютеры — близнецы или братья? // Открытые системы. — 2000. — № 5-6. — С. 9—14.
2. *Корнеев В.* Архитектуры с распределенной разделяемой памятью // Там же. — 2001. — № 3.— С. 9—17.
3. *Корнеев В.* Будущее высокопроизводительных вычислительных систем // Там же. — 2003. — № 5. — С. 10—17.
4. *Data* General's NUMALiiNE Technology: The Foundation for the AV25000 Server. — www.dg.com/about/html/av25000_foundation.html
5. *AViiON* AV25000 ccNUMA Server. — www.dg.com/aviion/html/av_25000_enterprise_server.html
6. *Локальные* управляющие сети для тесно связанных распределенных систем жестокого реального времени / Г. Г. Стецюра, В. С. Подлазов, А. Н. Смирнов и др. // Вычислительная техника. Системы. Управление. — 1991. — Вып. 6. — С. 82—96.
7. *Надежность* и пропускная способность кратного последовательного канала ЛВС на ВОЛС / В. С. Подлазов, В. В. Слободчук, А. Н. Смирнов, Г. Г. Стецюра // Вопросы кибернетики. Отказоустойчивые вычислительные системы реального времени. — М., 1990. — С. 100—136.
8. *Кузьминский М.* Реинкарнация Origin. Серверы SGI Altix: ccNUMA на базе процессоров Itanuim 2 // Открытые системы. — 2003. — № 7-8. — С. 12—15.
9. *Алленов А. В., Подлазов В. С., Стецюра Г. Г.* Пропускная способность набора кольцевых каналов. I. Класс наборов колец. Наборы с простыми узлами // Автоматика и телемеханика. — 1996. — № 3. — С. 135—144.
10. *Подлазов В. С.* Возможности кольцевых каналов в масштабируемых многопроцессорных вычислительных системах с общей разделяемой памятью // Труды Ин-та пробл. управл. РАН. — 1999. — Т. VI. — С. 91—99.
11. *Подлазов В. С., Подлазова А. В.* Обеспечение наращиваемости многопроцессорных систем с общей памятью с использованием многокольцевых некоммутируемых сетей связи (однородные узлы) // Там же. — 2002. — Т. XVI. — С. 103—116.
12. *Подлазов В. С.* Распределенные коммутаторы со статическими расписаниями для многопроцессорных вычислительных систем: Дисс. д-ра техн. наук. — М.: Ин-т пробл. управл. РАН. — Гл. 2.1.
13. *Теория* расписаний и вычислительные машины / Под ред. Э. Г. Коффмана. — М.: Наука, 1984. — 335 с.
14. *Лоскутов А. Ю., Михайлов А. С.* Введение в синергетику. — М.: Наука, 1990. — С. 217—225 (гл. 25. Сложные задачи комбинаторной оптимизации).
15. *Подлазов В. С., Подлазова А. В.* Обеспечение наращиваемости отказоустойчивых многопроцессорных систем с общей памятью с использованием многокольцевых некоммутируемых сетей связи с неоднородными узлами // Труды Ин-та пробл. управл. РАН. — 2002. — Т. XVIII. — С. 164—181.
16. *Стецюра Г. Г.* Эволюционные методы в задачах управления, выбора и оптимизации // Приборы и системы управления. — 1998. — № 3. — С. 54—62.
17. *Емельянов В. В., Курейчик В. М., Курейчик В. В.* Теория и практика эволюционного моделирования. — М.: Наука, 2003. — 432 с.
18. *Frank S., Bukhart H. and Rothnie J.* The KSR1: Bridging the gap between shared memory and MPPs // Proc. COMPCON'93. — CS Press, 1993. — P. 285—294.

☎ *(495) 334-78-31*

E-mail: podlazov@ipu.ru ❑