УДК 004.272.43

В. Н. ЕВГРАФОВ

СВОЙСТВА БЕЗБУФЕРНЫХ МНОГОСТУПЕНЧАТЫХ СЕТЕЙ ДЛЯ ПРОИЗВОЛЬНОГО ЧИСЛА «ГОРЯЧИХ» МОДУЛЕЙ ПАМЯТИ

Приведено характеристики багатоступеневих мереж в умовах передачі «гарячого» трафіку при наявності будь-якої кількості «гарячих» модулів пам'яті. Приведені характеристики використовуються для побудови аналітичної моделі мережі та розрахунку її швидкодії.

There is characteristics of multistage interconnection network presented for "hot spot" kind of traffic with any quantity of "hot" memory modules. Following assumptions are used for developing of analytics model of multistage interconnection network and performance evaluation.

Постановка проблемы. При производстве суперкомпьютерных систем необходимо решить задачу соединения большого количества процессоров в рамках одной вычислительной системы. Так, например, в максимальной конфигурации суперсистемы «Терра» насчитывается 256 процессоров, 512 модулей памяти, 256 контроллеров ввода/вывода, 256 модулей кэша ввода/вывода. Для соединения функциональных модулей суперсистемы применяются многоступенчатые сети (МС).

Необходимо иметь адекватную модель для оценки производительности МС при произвольном количестве "горячих" модулей памяти.

Общие сведения. Непрямой сетью (НС) называется сеть, у которой переключающие элементы отдалены от каналов ввода и вывода. «Источники» посылают сообщения на каналы ввода сети. Сообщения проходят через все ступени сети и достигают «приемников». В мультипроцессорных суперсистемах каналы ввода могут быть соединены с процессорными модулями, а каналы вывода с модулями памяти и/или с другими процессорными модулями.

Рассматриваются многоступенчатые, синхронные, переключающие непрямые многоступенчатые сети (НМС). Ступень, состоящую из всех «источников», обозначим, как нулевую ступень. Ступень, состоящую из переключающих элементов (ПЭ), подсоединенных непосредственно к «источникам», обозначим, как первую ступень. Ступень, состоящую из ПЭ, подсоединенных к ПЭ первой ступени, обозначим, как вторую ступень. ПЭ, который имеет два входных канала и два выходных, называется ПЭ первого порядка и обозначается 2×2.

В самонаправляющих сетях каждое сообщение (пакет) содержит заголовок и данные. В заголовок включена информация, необходимая для доставки пакета из «источника», откуда пришел пакет к «приемнику»,

который является конечным пунктом назначения пакета. Для выбора пути доставки пакета используется только эта информация в заголовке. Никакая другая глобальная информация о состоянии соседних ПЭ не используется.

Адрес, содержащийся в заголовке пакета, имеет длину]log28[бита. ПЭ, находящийся на нулевой ступени, направит пакет на верхний выходной канал, если младший нулевой бит сброшен в ноль, или на нижний выходной канал, если нулевой младший бит установлен в единицу. Затем пакет попадает на ПЭ, находящийся на второй ступени, который направит пакет в свой верхний выходной канал, если первый младший бит очищен в ноль, или на нижний канал, если первый младший бит установлен в единицу. Так, например, пакет, адресованный «приемнику» с адресом 6 = 1102, покидает нулевую ступень через верхний канал, первую ступень через нижний канал и вторую ступень через нижний канал.

Блокировка возникает в случае, когда на оба входа ПЭ пришли пакеты, которые должны быть направлены на один выходной канал. Оба запроса не могут быть обработаны, поэтому один пакет направляется на выходной канал, а другой блокируется. В нашей модели все пакеты имеют одинаковые предпочтения, поэтому оба пакета имеют одинаковую вероятность быть заблокированными в ситуации коллизии.

Весь процесс подчиняется следующим предположениям: источники генерируют пакеты независимо друг от друга. С заданной вероятностью p_i i-й источник генерирует пакет в начале каждого цикла. Каждый сгенерированный пакет направляется на ПЭ нулевой ступени; в результате работы внутреннего синхронизатора процесс передачи пакетов носит синхронный характер. В каждом цикле пакеты передаются с i-й на $(i+1)$-ю ступень; случайная величина d_i представляет адрес «приемника» пакета, сгенерированного i-м источником в начале цикла. Случайные величины d_i независимы и одинаково распределены. Распределение может быть задано как параметр модели.

Пусть НМС имеет N источников и N приемников. Обозначим НМС как Δ_n, где $n = \log_2 N$; S_n – перестановочные ступени, а через ES_n обозначим переключающие ступени. Конкатенацию сетей обозначим символом «;». Пусть двоичный адрес входного канала – $[b_{n-1}, b_{n-2},...,b_0]$. Тогда $S_n([b_{n-1}, b_{n-2},...,b_0]) = [b_{n-2}, b_{n-3},...,b_0, b_{n-1}]$. Введем оператор $\Im(\Delta_n) = \Delta_{n+1}$, который увеличивает размерность сети на единицу, тем самым, увеличивая количество входных каналов вдвое. Первые 2^n входных каналов связываются с ПЭ независимо от вторых 2^n каналов. $ES_n(1)$ соответствует ступени ПЭ первого порядка (ES_1), число которых – 2^{n-1}. К каждому ПЭ

присоединена соответствующая пара входных каналов. Переключающий элемент первого порядка ES_1 – ПЭ размерности 2×2.

Для оператора выполняются следующие условия:

1. $(\Im P_n)([b_n, b_{n-1}, \ldots, b_0]) = [b_n, P_n([b_{n-1}, b_{n-2}, \ldots, b_0])]$.
2. $\Im^k(\Delta_n) = \Im^{k-1}(\Im\Delta_n)$.
3. $ES_n(1) = \Im^{n-1} ES_1$.

В классическом, плоском представлении сети оператор $\Im$ соответствует наложению одной идентичной сети на другую.

Безбуферные ННС. Безбуферная архитектура подразумевает отсутствие буфера у ПЭ. При возникшей коллизии заблокированные пакеты безвозвратно теряются. Процесс передачи пакетов в условиях безбуферной архитектуры подчиняется следующим условиям: в рамках вычислительной системы существует $N = 2^n$ процессорных элемента, $n \in N$. Процессорный элемент под номером i обозначается, как PE_i, модуль памяти под номером j обозначается MM_j, где $0 \le i, j \le N-1$; в результате работы внутреннего синхронизатора процесс передачи пакетов носит синхронный характер. В каждом цикле пакеты передаются с *i*-й на (*i*+1)-ю ступень. Все процессорные элементы вырабатывают запрос к модулям памяти в начале каждого цикла. Модули памяти обрабатывают запросы процессоров за равное время; пакет содержит непосредственно данные и адрес назначения.

Для разрешения ситуации коллизии применяется логика случайного выбора. ПЭ случайным образом выбирает один входной канал и блокирует находящийся там пакет; при возникшей коллизии пакеты, которые не могут быть переданы, теряются.

Процессор генерирует пакеты независимо от того, был ли заблокирован пакет на предыдущем цикле; процессор генерирует пакеты независимо от остальных процессоров; процессор генерирует пакет в начале каждого цикла с вероятностью p_0; подразумевается, что пакет, сгенерированный процессорным элементом, равновероятно направляется на любой модуль памяти, за исключением «горячего» модуля памяти. На него пакеты направляются с большей вероятностью, чем на другие модули памяти.

Анализ литературы. Безбуферная архитектура предполагает повторную передачу заблокированных пакетов. В сетях, построенных по буферной архитектуре, заблокированные пакеты сохраняются в промежуточных буферах и передаются при последующих циклах. Производительность как буферных, так и безбуферных сетей хорошо описана в работах [1 – 3]. При расчете производительности МС большинство авторов предполагают равномерный доступ к памяти. Это означает, что пакеты направляются ко

всем модулям памяти с равной вероятностью [4 – 6]. Однако такое предположение является неприемлемым для реальных систем, где трафик имеет неоднородный характер. В работе [7] описываются условия возникновения неоднородного трафика. Из результатов этой работы следует, что неоднородный трафик возникает в большинстве приложений. Анализ производительности безбуферной сети был проведен в работе [8] для единственного "горячего" модуля памяти.

В данной статье впервые рассматривается случай, когда в рамках вычислительной системы имеется более одного «горячего» модуля памяти. Такая модель является более реалистической и соответствует реальной картине, которая возникает при передаче трафика в MIMD машинах; согласно предыдущему предположению, пакеты, прибывшие по входным каналам ПЭ, распределяются по выходным каналам *неравномерно*.

В реальных вычислительных системах заблокированные пакеты посылаются повторно на последующих циклах. Отказ от учета повторно посылаемых пакетов в рамках данной статьи, позволил упростить аналитическую и имитационную модель. Анализ более сложных систем для подобной проблемы [4 – 5] показал, что предположение о генерации пакетов независимо от блокировки на предыдущем цикле вносит очень *незначительные* ошибки допущения.

Цель статьи. Для построения аналитической модели оценки производительности МС при произвольном числе "горячих" модулей памяти, необходимо зафиксировать свойства трафика и расширить классификацию каналов данных. Данная статья определяет новые классы каналов передачи данных и описывает свойства потока данных. Формулируются определения *"горячего" множества каналов*, описываются свойства каналов, имеющих *конечный*, *общий*, *граничный* статусы.

Свойства сети. На рисунке изображена Ω-сеть, размерности $N = 8$, у которой MM_4, MM_7 – «горячие» модули памяти.

Пакеты, предназначенные для "горячего" модуля памяти, назовем «горячими» пакетами, а каналы, по которым могут быть доставлены «горячие» пакеты, назовем «горячими» каналами. "Горячие" каналы обозначены жирными линиями.

Все ПЭ, присоединенные к "горячим" каналам, называются "горячими". "Горячие" ПЭ обозначены жирными прямоугольниками.

Справедливо следующее утверждение: все "горячие" выходные каналы на любой ступени сети являются одновременно либо нижними, либо верхними.

Назовем множество модулей памяти, которые достижимы из данного ПЭ, областью видимости ПЭ. Множество модулей памяти, которые достижимы из канала, – областью видимости канала. Тогда область

видимости ПЭ разбивается на две непересекающиеся равные области видимости верхнего и нижнего каналов.

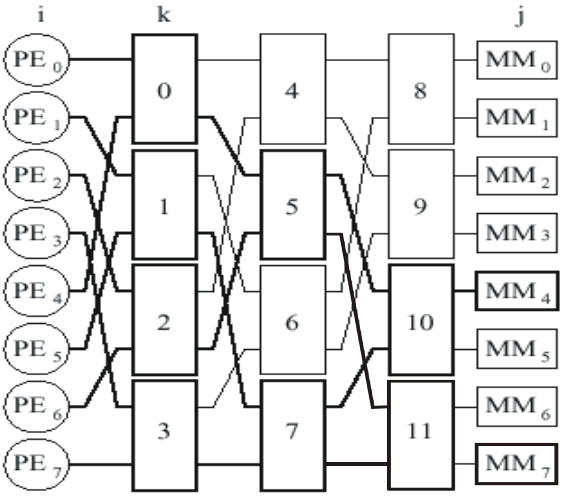

Рис.

После прохождения ступени любое множество входных каналов разбивается на два множества выходных каналов одинаковой мощности, причем, объединение областей видимости первого множества выходных каналов и второго множества выходных каналов в результате дает область видимости множества входных каналов.

Покажем, что уровень потока данных на выходных каналах горячего ПЭ различен. Пусть два выходных канала ПЭ передают запросы для двух непересекающихся множеств модулей памяти. Выходной канал, который передает запросы для множества модулей памяти, который включает в себя "горячий" модуль памяти, будет иметь больший уровень потока данных, чем канал, передающий запросы к множеству, не содержащему "горячего" модуля памяти. Пусть, к примеру, верхний и нижний выходной каналы нулевого ПЭ передает трафик для множеств модулей памяти $\{MM_0, MM_1, MM_2, MM_3\}$ и $\{MM_4, MM_5, MM_6, MM_7\}$ соответственно. Уровень потока данных для нижнего выходного канала будет выше, чем для верхнего канала, из-за наличия "горячих" модулей в области видимости нижнего канала.

Два выходных канала передают трафик для двух непересекающихся множеств модулей памяти, которые имеют равную вероятность быть

запрошенными. Поэтому уровень данных на обоих выходных каналах одинаков.

Оба выходных канала "горячего" ПЭ, имеющего конечный статус, никогда не могут быть "горячими" одновременно. "Горячие" пакеты, прибывающие на вход "горячего" ПЭ, имеющего крайний статус, направляются на "горячий" выходной канал, который может передавать также и обыкновенные пакеты. Обыкновенные пакеты направляются только на обыкновенный выходной канал.

Множество каналов, по каждому из которых вероятность прохождения "горячего" пакета не равна нулю, назовем – "горячим" множеством каналов.

Каждое "горячее" множество входных каналов может иметь один из 3-х статусов:

– если область видимости одного "горячего" множества выходных каналов включает в себя *один* "горячий" модуль памяти, а область видимости другого множества выходных каналов не содержит *ни одного* "горячего" модуля памяти, то такое "горячее" множество входных каналов имеет **конечный статус**;

– если область видимости одного горячего множества выходных каналов включает в себя *более одного* "горячего" модуля памяти, а область видимости другого множества выходных каналов не включает *ни одного* "горячего" модуля памяти, то такое "горячее" множество входных каналов имеет **общий статус**;

– если область видимости одного и другого "горячих" множеств выходных каналов одновременно включает *более одного* "горячего" модуля памяти, то такое "горячее" множество входных каналов имеет **граничный статус**.

Вершина (множество входных каналов), которая входит одновременно в два или более пути, но находится не на разветвлении путей, имеет общий статус. Вершина, которая входит одновременно в два или более пути, и находится на разветвлении путей, имеет граничный статус. Вершина, которая входит только в один путь, имеет конечный статус.

Горячее множество каналов, имеющее крайний статус, порождает вдвое меньше горячих ПЭ на последующей ступени сети. Количество горячих выходных каналов на последующей ступени также уменьшается вдвое. Число "горячих" выходных каналов на ступени S_k равно $2^n / 2^{k+1}$. Количество горячих ПЭ на ступени S_k равно 2^{n-1-k}.

Оба входных канала любого ПЭ одновременно являются либо "горячими" каналами, либо обыкновенными каналами. На одном из входов ПЭ x_u, $0 \le u \le 1$ имеется пакет, который должен быть направлен на один из выходов ПЭ y_v, $0 \le v \le 1$. Вероятность того, что он будет направлен на

выход y_v, равно $r_v / \sum_{w=0}^{1} r_w$, где r_w – это сумма вероятностей, с которыми процессор запрашивает множество модулей памяти, достижимых из выходного канала y_v.

Выводы. Рассмотрены особенности безбуферной архитектуры. Определены *конечный*, *общий*, *граничный* статусы множеств входных каналов и изучена эволюция потока данных. Приведенные выше свойства и определения позволяют, в перспективе, получить более общую модель передачи "горячего" трафика в многоступенчатой сети. Аналитическая модель может быть использована для оценки производительности МС.

Список литературы: 1. *Wilkinson B.* Overlapping connectivity interconnection networks for shared memory multiprocessors systems // J. Parall Distrib Comput. – 1992. – 15 (1). – P. 49 – 61. **2.** *Liu Y.C., Wang C.* Analysis of prioritized crossbar multiprocessor systems // J. Parall Distrib Comput. – 1991. – 7. – P. 321 – 334. **3.** *Valero M., Llaberia J.M., Labarta J., Sanvicente E., Lang T.* A performance evaluation of the multiple bus network for multiprocessors // Sigmetrics Conference on Measurment and Modelling of Computer Systems August. – 1983. – P. 200 – 206. **4.** *Chang D.Y., Kuck D.J. and Lawrie D.H.* On the effective bandwith of parallel memories // IEEE Transactions on Computers. – 1977. – Vol l. – P. 480 – 489. **5.** *Basket F., Smith A.J.* Interference in multiprocessor computer systems with interleaved memory // Comunications of ACM. – 1976. – Vol. 19. – № 6. – P. 327 – 334. **6.** *Yang Q., Bhuyan L.N.* Analysis of packet-switched multiple-bus multiprocessors // IEEE Trans. Comput. – 1991. – 40 (3). – P. 352 – 357. **7.** *Kim H.S., Leon-Garcia A.* Performance of buffered Banyan networks under non-uniform traffic patterns // IEEE Trans. Commun. – 1990. – Vol. 38 (5). – 648 – 658. **8.** *Atiquzzaman M., Akhtar M.S.* Effect of hot spots on the performance of multistage interconnection networks // FRONTIERS 92: The Forth Symposium on the Frontiers of Massively Parallel Computations. – Virginia. – October 19-21, 1992. – P. 504 – 505.

Поступила в редакцию 31.09.2004