УДК 004.274.43

ОЦЕНКА МИНИМАЛЬНЫХ ТРЕБОВАНИЙ К АППАРАТУРЕ И ТОПОЛОГИИ ПРИ ПОСТРОЕНИИ ВЫСОКОСКОРОСТНЫХ КОММУНИКАЦИОННЫХ СЕТЕЙ ДЛЯ СУПЕРКОМПЬЮТЕРОВ С ОБЩЕЙ ПАМЯТЬЮ

А. А. Корж[1], **Д. В. Макагон**[1]

Рассматриваются проблемы построения эффективных высокоскоростных коммуникационных сетей для суперкомпьютеров с распределенной общей памятью. Приводятся оценки агрегатной пропускной способности сети в зависимости от топологических характеристик и шаблонов доступа к данным при решении задач с интенсивной нерегулярной работой с памятью. На основе полученных соотношений проводится оценка минимальных аппаратно-технических требований для наиболее распространенных топологий. Статья подготовлена по материалам доклада авторов на международной научной конференции "Параллельные вычислительные технологии" (ПаВТ-2008; http://agora.guru.ru/pavt).

1. Введение. Существующие высокопроизводительные системы используют принцип параллельной обработки данных на многих вычислительных узлах. Каждый такой узел содержит несколько процессоров с локальной памятью. Для обмена информацией и синхронизации работы узлы соединяются между собой коммуникационной сетью.

В настоящее время большое внимание уделяется разработке высокопроизводительных систем, использующих распределенную общую память [1 – 3]. В таких системах коммуникационная сеть становится ключевой в подсистеме памяти; следовательно, от ее производительности зависит и производительность всей системы. Для обеспечения эффективной работы с распределенной общей памятью необходимо, чтобы пропускная способность сети была сопоставима с пропускной способностью локальной памяти узла [3].

С другой стороны, возможности доступных технологий являются ограниченными, в частности, ограниченной является агрегатная пропускная способность маршрутизатора и скорость каналов сети. Отсюда, казалось бы, следует логичный вывод, что число каналов, подключаемых к каждому маршрутизатору (степень маршрутизатора), должно быть достаточно большим, чтобы позволить использовать максимально возможную для современной технологии агрегатную пропускную способность кристалла. Однако теоретически возможен и альтернативный вариант — в кристалле вместо одного маршрутизатора высокой степени можно использовать несколько маршрутизаторов меньшей степени.

В [4, 5] для доказательства оптимальности использования сетей с высокой степенью связности (high radix networks) приводятся выкладки, в которых минимизируется задержка передачи сообщения при ограниченной агрегатной пропускной способности маршрутизаторов. Такой подход нам представляется малоэффективным, поскольку для современных систем более важной является не задержка доставки одного сообщения в пустой сети, что редко отвечает коммуникационным шаблонам использующихся задач, а пропускная способность сети в условиях нагрузок, близких к максимальным.

Кроме того, крайне важно обеспечить баланс пропускной способности памяти и сети [1, 2]. Это труднодостижимо, поскольку соответствующего высокоскоростного канала (порядка 25 – 50 Гб/с) пока не существует и в ближайшее время не ожидается.

Одним из возможных решений в данной ситуации является использование нескольких обычных каналов (1 – 4 Гб/с) для подключения процессора к сети. Кроме того, для увеличения пропускной способности, приходящейся на узел, возможно подключение каждого узла сразу к нескольким маршрутизаторам [3, 4], при этом часто увеличивают и количество маршрутизаторов относительно количества узлов для повышения общей пропускной способности сети.

Нашей целью является теоретическая оценка производительности рассмотренных вариантов реализации коммуникационной сети с учетом коммуникационного шаблона задачи. В отличие от [4, 5], проблема

[1] Научно-исследовательский центр электронной вычислительной техники (НИЦЭВТ), Варшавское шоссе, 125, 117587, Москва; e-mail: anton@korzh.ru; makagond@mail.ru

видится не только в необходимости использования сетей высокой связности, но и в эффективности применения (в том числе с точки зрения стоимости сети) различных топологий, в частности таких широко используемых, как kD-тор и сети Клоса.

2. Терминология. Вычислительная система с распределенной общей памятью состоит из узлов, каждый из которых содержит процессорный элемент (processor element, PE) с локальной памятью. Обмен данными между узлами вычислительной системы обеспечивается коммуникационной сетью.

Сеть состоит из маршрутизаторов, которые соединены между собой двусторонними каналами связи (линками). То, каким образом PE подсоединяются к маршрутизаторам и как маршрутизаторы соединяются между собой, задается топологией сети. Обычно каждый PE за единицу времени способен принять и отправить гораздо больший объем данных, чем может передать сетевой линк, поэтому на каждый PE выделяется сразу несколько линков.

Каждый маршрутизатор имеет определенное число портов, к каждому из которых подсоединяется линк, идущий либо в другой маршрутизатор, либо к PE. Порты, подключенные к PE, называются PE-портами. Число портов у маршрутизатора, подключенных к другим маршрутизаторам, называется его степенью.

В рассматриваемой нами модели в целях упрощения предполагается, что если некоторые два маршрутизатора соединены, то лишь одним линком; пропускные способности всех линков одинаковы.

Понятие PE-порта используется для обобщения нескольких возможных типов соединения маршрутизаторов с PE, а именно: подключение одного PE сразу к нескольким маршрутизаторам, подключение PE к маршрутизатору несколькими линками и подключение к одному маршрутизатору нескольких PE. Далее мы будем рассматривать лишь общее число PE-портов у маршрутизаторов (а не общее число PE), не затрагивая вопроса, как именно распределены эти PE-порты между PE.

В наших дальнейших рассуждениях мы будем ориентироваться на регулярные топологии, т.е. такие, в которых число портов и PE-портов у всех маршрутизаторов одинаковое. Это, несомненно, является существенным допущением, однако на практике почти всегда используются именно такие топологии, поскольку для них возможны наиболее простые и эффективные алгоритмы маршрутизации.

Основная используемая нами характеристика компонентов сети (PE, порт, линк) — это пропускная способность, показывающая, сколько данных передает компонент сети за единицу времени. Пиковая пропускная способность — это теоретически максимальная пропускная способность.

Для сравнения топологий мы будем использовать две важнейшие характеристики: диаметр и средний диаметр. Диаметр сети — это максимальная длина кратчайшего пути в сети. Средний диаметр — это средняя длина кратчайшего пути в сети.

3. Основные соотношения. В общем случае вычислительная система состоит из N_{PE} узлов (процессорных элементов), пиковая пропускная способность каждого узла — $\mathrm{PBW}_{\mathrm{PE}}$. Коммуникационная сеть, соединяющая узлы системы, состоит из N_{Rt} маршрутизаторов степени deg, каждый маршрутизатор соединен с вычислительными узлами ν линками (т.е. имеет ν PE-портов). Пиковые пропускные способности линка, порта и PE-порта равны $\mathrm{PBW}_{\mathrm{Link}}$.

Общее число PE-портов в сети (N_{PEports}) определяется числом процессоров и соотношением пропускных способностей процессора и линка:

$$N_{\mathrm{PEports}} = N_{\mathrm{PE}} \frac{\mathrm{PBW}_{\mathrm{PE}}}{\mathrm{PBW}_{\mathrm{Link}}}. \tag{1}$$

С другой стороны, общее число PE-портов в сети равно произведению количества маршрутизаторов (N_{Rt}) на число PE-портов у каждого маршрутизатора (ν): $N_{\mathrm{PEports}} = N_{\mathrm{Rt}}\,\nu$. Пиковая пропускная способность маршрутизатора ($\mathrm{PBW}_{\mathrm{Rt}}$) равна произведению его степени (deg) на пиковую пропускную способность линка:

$$\mathrm{PBW}_{\mathrm{Rt}} = \deg \times \mathrm{PBW}_{\mathrm{Link}}. \tag{2}$$

Рассмотрим общий случай взаимодействия процессоров через произвольную сеть. Пусть N_{PE} процессоров инжектируют пакеты в сеть со скоростью $\mathrm{PBW}_{\mathrm{PE}}$, т.е. агрегированный входной трафик сети $\mathrm{PBW}_{\mathrm{Inj}} = \mathrm{PBW}_{\mathrm{PE}}\,N_{\mathrm{PE}}$, а инжекция в каждый маршрутизатор ($\mathrm{PBW}_{\mathrm{Inj\,Rt}}$) выражается формулой

$$\mathrm{PBW}_{\mathrm{Inj\,Rt}} = \mathrm{PBW}_{\mathrm{PE}} \frac{N_{\mathrm{PE}}}{N_{\mathrm{Rt}}}. \tag{3}$$

В зависимости от топологии сети и специфики задачи можно вычислить среднюю длину пути пакета в сети, т.е. среднее число маршрутизаторов на пути пакета; пусть оно равно $H + 1$ (H — среднее число линков на пути пакета). Значение H зависит не только от топологии сети, но и от вида нагрузки на сеть.

В частности, задачи, использующие сеточные методы, хорошо ложатся на многомерный тор, поскольку активно обмениваются данными только соседние узлы; поэтому средняя длина пути пакета в сети в этом случае близка к 1. Однако для задач, в которых каждый узел может равновероятно обмениваться данными с любым другим узлом, значение H близко к среднему диаметру сети. В наихудшем случае, когда узлы обмениваются пакетами с наиболее далекими узлами, значение H близко к диаметру сети.

Общий транзитный трафик (не считая эжекции в PE) на выходе маршрутизаторов сети будет близок к $\mathrm{PBW}_{\mathrm{Inj}} H$, поэтому имеет место следующее соотношение:

$$N_{\mathrm{Rt}}\,\mathrm{PBW}_{\mathrm{Rt}} \geqslant N_{\mathrm{PE}}\,\mathrm{PBW}_{\mathrm{PE}}\,H. \tag{4}$$

Учитывая (2) и (4), получим оценку для пиковой пропускной способности линка:

$$\mathrm{PBW}_{\mathrm{Link}} \geqslant \mathrm{PBW}_{\mathrm{PE}}\,\frac{N_{\mathrm{PE}}}{N_{\mathrm{Rt}}}\,\frac{H}{\deg}\,. \tag{5}$$

Учитывая (1) и (5), получим оценку максимального числа PE-портов у маршрутизатора:

$$\nu \leqslant \frac{\deg}{H}\,. \tag{6}$$

Формулы (4) – (6) позволяют связать основные характеристики топологии и пиковые пропускные способности двух главных компонентов сети — линка и PE. Благодаря этому можно вывести оценку для любого из этих параметров через остальные.

Ниже мы воспользуемся формулами (4) – (6) применительно к конкретным топологиям с целью ответить на часто возникающий вопрос: сколько необходимо маршрутизаторов (с каким числом PE-портов), чтобы соединить заданное число PE (с заданной $\mathrm{PBW}_{\mathrm{PE}}$) линками (с заданной $\mathrm{PBW}_{\mathrm{Link}}$). Другими словами, сколько необходимо маршрутизаторов (с каким числом PE-портов), чтобы сеть имела заданное число PE-портов.

При построении коммуникационных сетей используется весьма ограниченное число различных классов топологий; мы подробно рассмотрим наиболее широко используемые — многомерные торы и сети Клоса (fat tree).

Для определенности будем в дальнейших рассуждениях ориентироваться на задачи, создающие нагрузку в сети, близкую к случайному равномерно распределенному трафику. Это позволит нам в формулах (4) – (6) заменить H на средний диаметр сети D_{Av}.

4. Многомерный тор. Многомерный тор [7] задается размерами сторон и числом измерений $K = \dfrac{\deg}{2}$; для простоты мы будем рассматривать самый оптимальный вид тора — равносторонний. Для тора с K измерениями и N_{Rt} узлами средний диаметр находится по формуле $D_{\mathrm{Av}} = \dfrac{K}{4}\sqrt[K]{N_{\mathrm{Rt}}}$. Из соотношения (6) и исходя из необходимого числа PE-портов, получим следующие оценки числа маршрутизаторов и числа PE-портов у каждого маршрутизатора:

$$N_{\mathrm{Rt}} \geqslant \left(\frac{N_{\mathrm{PEports}}}{8}\right)^{\deg/(\deg-2)}, \tag{7}$$

$$\nu \leqslant \left(\frac{8^{\deg}}{N_{\mathrm{PEports}}^2}\right)^{1/(\deg-2)}. \tag{8}$$

Рассмотрим K-мерный равносторонний тор со стороной n. Общее число узлов в нем имеет вид $N_{\mathrm{Rt}} = n^K$, средний диаметр $D_{\mathrm{Av}} = \dfrac{nK}{4}$. Из соотношения (6) получим

$$\nu \leqslant \frac{8}{n}\,. \tag{9}$$

Анализируя полученное соотношение и учитывая, что n — целое, большее или равное двум, получим, что для тора максимальное ν всегда меньше 4, а равно 4 только для гиперкуба с удвоенными ребрами (в частном случае тора при длине стороны 2 все циклы превращаются в удвоенные ребра гиперкуба).

Если же рассматривать ситуацию, когда к каждому маршрутизатору идет только один линк от PE (т.е. у каждого маршрутизатора всего один PE-порт, $\nu = 1$), получим, что $n \leqslant 8$, а это значит, что в принципе нецелесообразно рассматривать торы, у которых стороны больше 8.

Отсюда же очевидно, что независимо от степени тора неэффективно выбирать сторону тора больше, чем $\dfrac{8}{\nu}$.

Учитывая (9) и (1), получим оценку максимального общего числа PE-портов для равносторонних торов разной размерности:

$$N_{\mathrm{PEports}} \leqslant \nu \left(\frac{8}{\nu}\right)^{K}. \tag{10}$$

Это соотношение указывает максимальное число PE-портов, которое может быть удовлетворительно связано сетью типа равностороннего K-мерного тора с $\dfrac{N_{\mathrm{PEports}}}{\nu}$ маршрутизаторами (т.е. с ν PE-портами у каждого маршрутизатора).

5. Сеть Клоса. Сеть Клоса [4, 6] задается числом портов у маршрутизаторов разных стадий (deg) и общим числом стадий $(2D-1)$, где D — диаметр сети. Для простоты мы будем рассматривать частный случай сетей Клоса, в которых все маршрутизаторы имеют одинаковое число портов; такие сети имеют максимальную бисекцию. Для сети Клоса с $(2D-1)$ стадиями и N_{Rt} узлов средний диаметр очень близок к D ($D \approx D_{\mathrm{Av}}$), а степень маршрутизатора находится по формуле

$$\deg = 2\left(\frac{N_{\mathrm{Rt}}}{D+1}\right)^{2/D}.$$

В сети Клоса PE-порты есть лишь у маршрутизаторов первого слоя, у остальных маршрутизаторов PE-портов нет; число PE-портов у маршрутизаторов первого слоя выражается формулой

$$n = \nu\,\frac{D+1}{2}.$$

Поскольку сеть Клоса нерегулярна, формулы (4) – (6) необходимо изменить, а именно, следует учесть, что маршрутизаторы первого слоя имеют меньшую пропускную способность:

$$\mathrm{PBW}^{I}_{\mathrm{Rt}} = (\deg - n)\mathrm{PBW}_{\mathrm{Link}}.$$

Общее количество маршрутизаторов первого слоя задается равенством

$$N^{I}_{\mathrm{Rt}} = \frac{2N_{\mathrm{Rt}}}{D+1} = \frac{N_{\mathrm{PEports}}}{n}.$$

Формулы (4) – (6) принимают следующий вид:

$$(N_{\mathrm{Rt}} - N^{I}_{\mathrm{Rt}})\mathrm{PBW}_{\mathrm{Rt}} + N^{I}_{\mathrm{Rt}}\,\mathrm{PBW}^{I}_{\mathrm{Rt}} \geqslant N_{\mathrm{PE}}\,\mathrm{PBW}_{\mathrm{PE}}\,H, \tag{11}$$

$$\mathrm{PBW}_{\mathrm{Link}} \geqslant \mathrm{PBW}_{\mathrm{PE}}\,\frac{N_{\mathrm{PE}}}{N_{\mathrm{Rt}}}\,\frac{H}{\deg - \dfrac{n}{D+1}}, \tag{12}$$

$$\nu \leqslant \frac{\deg}{H+1}, \tag{13}$$

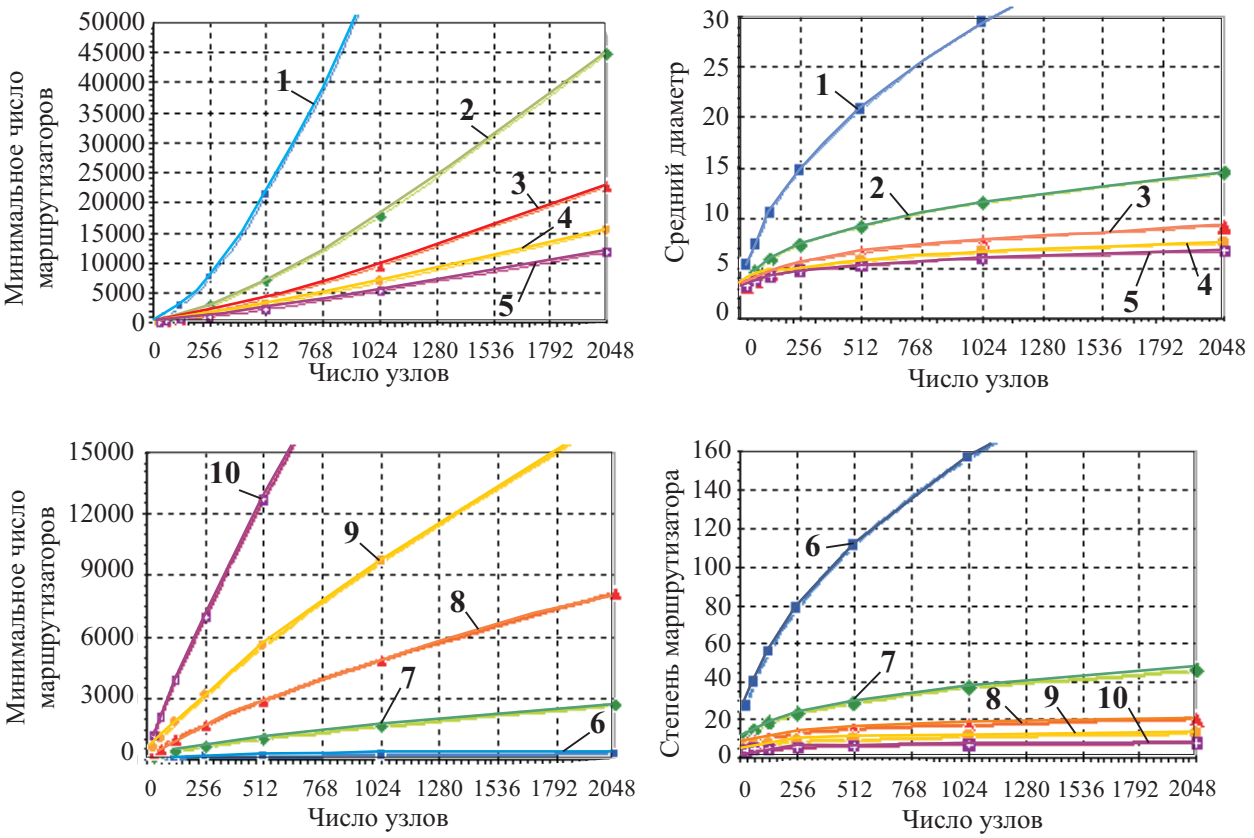
$$n \leqslant \frac{\deg}{2}\,\frac{D+1}{H+1}. \tag{14}$$

Формулы (13) и (14) эквивалентны. При $H \approx D$ получим $n \leqslant \dfrac{\deg}{2}$; это означает, что для эффективной работы сети необходимо, чтобы у каждого маршрутизатора первого слоя сети Клоса было не более чем $\dfrac{\deg}{2}$ PE-портов.

Из соотношения (14) и условия $H \approx D$ и исходя из необходимого числа PE-портов, получим следующие оценки числа маршрутизаторов и числа PE-портов у маршрутизаторов первого слоя:

$$N_{\mathrm{Rt}} \geqslant (D+1)\left(\frac{N_{\mathrm{PEports}}}{2}\right)^{D/(D+2)}, \tag{15}$$

$$n \leqslant \frac{1}{2}\left(N^{2}_{\mathrm{PEports}}\,2^{D}\right)^{1/(D+2)}. \tag{16}$$



Минимальное число маршрутизаторов, требующееся для сетей типа kD-тор различной размерности и сетей Клоса с различным числом стадий, чтобы обеспечить указанное число узлов с пропускной способностью в 12 раз большей, чем пропускная способность канала. Обозначение линий, представленных на рисунке: 1) 3-мерный тор, 2) 4-мерный тор, 3) 5-мерный тор, 4) 6-мерный тор, 5) 7-мерный тор, 6) 3-стад. сеть Клоса, 7) 5-стад. сеть Клоса, 8) 7-стад. сеть Клоса, 9) 9-стад. сеть Клоса, 10) 11-стад. сеть Клоса

На рисунке слева представлены графики, построенные на основе соотношений (7), (8), (15), (16) и показывающие, какое минимальное число маршрутизаторов требуется для сетей типа kD-тор различной размерности и сетей Клоса с различным числом стадий, чтобы обеспечить указанное число узлов с пропускной способностью в 12 раз большей, чем пропускная способность канала. Последнее требование, по мнению авторов, актуально при построении высокопроизводительной системы с распределенной общей памятью, для соблюдения баланса между пропускной способностью сети и подсистемы памяти [4].

Следует отметить, что при фиксировании размерности сети kD-тор диаметр сети зависит от количества маршрутизаторов, в то время как при фиксировании количества стадий сети Клоса диаметр остается неизменным, но приходится увеличивать степень маршрутизаторов, что отражено на графиках справа.

6. Заключение. Основным результатом статьи можно считать формулы (4) – (6), которые ограничивают отношение пропускной способности инжекционно-эжекционных каналов каждого маршрутизатора к суммарной пропускной способности межузловых каналов. Исходя из этого, фактическая инжекционная пропускная способность на каждый маршрутизатор также будет ограничена, причем это ограничение зависит от коммуникационного шаблона.

Таким образом, для каждого типа трафика можно оценить, сколько узлов и с какой пропускной способностью сможет поддержать сеть. Кроме того, в случае применения маршрутизаторов с низкой степенью их количество должно быть больше, чем в случае применения маршрутизаторов и топологии сети с высокой связностью, хотя в последнем случае значительно увеличивается количество каналов связи.

На основе полученных формул в будущем планируется провести сравнение сетей с различными топологиями по стоимости при одинаковой агрегатной пропускной способности. При этом стоимость будет рассмотрена как с точки зрения сложности кристаллов маршрутизаторов, так и с точки зрения количе-

ства высокоскоростных каналов между маршрутизаторами.

С целью оценки точности полученных выводов в будущих работах планируется провести моделирование различных вариантов подключения узла к нескольким маршрутизаторам, в том числе, рассмотреть вопрос использования сетей типа kD-тор с числом маршрутизаторов, превышающим число узлов, с целью повышения пропускной способности, приходящейся на узел.

Авторы статьи выражают благодарность руководителю работы по разработке суперкомпьютера с мультитредово-потоковой архитектурой Л. К. Эйсымонту.

СПИСОК ЛИТЕРАТУРЫ

1. *Фролов А., Семенов А., Корж А., Эйсымонт Л.* Программа создания перспективных суперкомпьютеров // Открытые системы. 2007. № 9. 42–51.
2. *Слуцкин А., Эйсымонт Л.* Российский суперкомпьютер с глобально адресуемой памятью // Открытые системы. 2007. № 9. 20–21.
3. *Sterling T.* Critical Factors and Directions for Petaflops-scale Supercomputers // Presentation on IFIP WG10.3 e-Seminar Series (www.ifipwg103.org/seminar).
4. *Scott S., Abts D., Kim J., Dally W.J.* The BlackWidow high-radix Clos network // Proc. Int. Symposium on Computer Architecture (ISCA). Boston, MA (US). 2006. 16–28.
5. *Kim J., Dally W.J., Towles B., Gupta A.K.* Microarchitecture of a high-radix router // Proc. Int. Symposium on Computer Architecture (ISCA). Madison, WI (US). 2005. 420–431.
6. *Clos C.* A study of non-blocking switching networks // The Bell System Technical J. 1953. **42**, N 2. 406–424.
7. *Dally W.J.* Performance analysis of k-ary n-cube interconnection networks // IEEE Transactions on Computers. 1990. **39**, N 6. 775–785.

Поступила в редакцию
11.03.2008