О РЕАЛИЗАЦИИ В ПЛИС МАРШРУТИЗАТОРА ВЫСОКОПРОИЗВОДИТЕЛЬНОЙ СЕТИ

А.Ю. Орлов, А.Б. Шворин

Введение

Всем понятна необходимость изобретения новых архитектур для высокопроизводительных вычислений. При этом, в большинстве случаев, разработчики не могут позволить себе делать машину, построенную на новых принципах, "с нуля": слишком много времени пройдет, прежде чем будет создан или портирован весь необходимый инструментарий, и какие-то реальные задачи будут запущены на ней -- к этому времени машина устареет, ее не удастся окупить. Отсюда возникает идея *гибридных* машин, которые содержат как традиционную кластерную часть, так и экспериментальную "добавку". Для реализации этой экспериментальной "добавки" идеально подходят программируемые микросхемы (ПЛИС) -- это дешевый с точки зрения разработки способ опробовать те или иные аппаратные решения.

Встает вопрос о том, какого рода архитектурные дополнения хотелось бы иметь вдобавок к традиционной кластерной организации суперкомпьютера. Термин "гибридность" ассоциируется с понятием "сопроцессора", когда нетрадиционная аппаратура строго локализована внутри одного узла и служит ускорению вычислений в нем. Однако ясно, что узкое место кластерной архитектуры -- это слабая сеть, и именно здесь хотелось бы добиться принципиальных улучшений. Мы понимаем под "гибридной машиной" архитектуру, сочетающую в себе традиционные и новые решения для организации не только вычислений, но и коммуникаций. Оказывается, современные ПЛИС позволяют организовывать коммуникации на предельных скоростях, доступных сегодня для общеиспользуемых соединений, и это дает возможность строить сети на ПЛИС, успешно конкурирующие с такими решениями как Infiniband, Myrinet и т.д. Как правило, выигрыш достигается за счет специализации под конкретный класс задач. Кроме того, технология ПЛИС позволяет весьма дешево ставить эксперименты для обкатки новых идей в области коммуникаций.

На рынке появляются ПЛИС с достаточно большой вентильной емкостью и встроенными высокоскоростными трансиверами. Например, Altera Stratix IV GX [1] имеет встроенные ядра для работы с PCI Express Gen. 2 и соответствующие трансиверы, обеспечивающие пропускную способность связи с центральным процессором на уровне 8 GB/s на один линк 8x. Такие же трансиверы можно использовать и для объединения ПЛИС, стоящих в разных узлах, в единую сеть.

Остановимся на преимуществах и недостатках использования ПЛИС как платформы для построения сети суперкомпьютера, и далее приведем пример конкретного решения: маршрутизатора сети типа 3D-тор.

Преимущества использования ПЛИС

– У разработчиков суперкомпьютера появляется возможность создания решений для огромного класса задач, где сеть является узким горлом. Использование ПЛИС позволяет вывести машины кластерного типа на совершенно новый уровень:

- реализовав в ПЛИС маршрутизатор сети типа 3D-тор, получаем недорогую сеть, эквивалентную используемой в IBM Blue Gene [2];
- поддержав в ПЛИС эффективную одностороннюю передачу коротких сообщений, получаем возможности, доступные до сих пор лишь в машинах Cray и SGI;
- и т.д.

1. Простота и скорость разработки, удобство программирования, отладки и экспериментирования по сравнению с разработкой специализированной интегральной схемы для коммуникационного оборудования.
2. Помимо обеспечения коммуникаций, ПЛИС может выполнять и традиционную работу вычислительного сопроцессора. В отличие от традиционных гибридных схем, когда результат вычислений в сопроцессоре передается центральному процессору, ПЛИС могут обмениваться данными для вычислений в сопроцессорах напрямую, не мешая работе центрального процессора. Кроме того, при исполнении вычислений и коммуникаций физически в одном устройстве, снижается задержка на коммуникационных операциях, что дает значительное преимущество на широком классе задач.
3. Удобство перепрограммирования ПЛИС позволяет реализовывать в ней алгоритмы маршрутизации, специализированные под задачу. Простейшим примером может служить проект QPACE [3], где в ПЛИС реализуется высокореактивная сеть, обеспечивающая связь только соседних узлов в трехмерном пространстве. Такая сеть очень удобна для решения задач квантовой хромодинамики.

Недостатки ПЛИС

1. Более низкая эффективность по сравнению со специализированными интегральными схемами.

2. Малый объем встроенной статической памяти (SRAM).

Специфика использования ПЛИС в качестве маршрутизатора позволяет нивелировать указанные недостатки. А именно: алгоритмы маршрутизации, как правило, не являются вычислительно сложными, поэтому даже сравнительно низкая частота ПЛИС является достаточной для решения поставленных задач. Что касается объема встроенной памяти, то она требуется в основном лишь для буферизации данных. В тех случаях, когда буферной памяти оказывается недостаточно, имеется возможность подключить внешнюю память DRAM или воспользоваться системной памятью узла. Отметим, что использование внешней памяти увеличит задержку, однако это не является существенным недостатком, так как задержка в буферах может быть скрыта путем конвейерезации.

Маршрутизатор сети 3D-тор

Идея использовать тороидальную топологию, в том числе трехмерную, не нова. В качестве примера можно отметить разработки Cray T3E и IBM BlueGene/L (см. [6,2]).

Новизна предлагаемого решения заключается в использовании ПЛИС для организации сети суперкомпьютера. Ниже приведены характеристики этого конкретного решения -- маршрутизатора для сети типа 3D-тор.

- Топология сети -- 3D-тор, сеть рассчитана на тысячи узлов.
- Пропускная способность одного линка -- 10 Гбит/с в одном направлении; линки полнодуплексные. Таким образом, полная ПС маршрутизатора 10*6*2 = 120 Гбит/с.
- Интерфейс с центральным процессором (процессорами) -- два порта PCI-E Gen.2 8x. Суммарная пропускная способность - 16 ГБ/с.
- Особенности маршрутизации (некоторые упоминаемые ниже понятия определены в [4,5]).
  25. Пузырьковая передача данных (bubble routing) по принципу VCT (virtual-cut-through).
  26. Виртуальная подсеть с детерминированной маршрутизацией гарантирует отсутствие дедлоков.
  27. Опциональная виртуальная подсеть с адаптивной маршрутизацией обеспечивает более равномерную загрузку, позволяет обходить "заторы".
  28. Гарантируется отсутствие лайвлоков (бесконечного блуждания пакета по сети).

Более подробно про маршрутизацию. Единица передачи данных -- пакет -- в процессе передачи может "размазываться" по нескольким узлам, то есть голова пакета может начать передаваться в следующий узел до того, как подошел хвост. Это обеспечивает низкую задержку (не нужно дожидаться, пока весь пакет соберется в узле), однако требует дополнительных механизмов, гарантирующих отсутствие дедлоков. Одним из таких механизмов является правило пузырька -- дисциплина передачи данных, требующая наличия свободного места в буфере узла-получателя для помещения туда всего передаваемого пакета. Проблема дедлоков и пути ее решения исследованы в работах Х.Дуато (см., например, [4]).

Реализация такого маршрутизатора в ПЛИС обладает следующими характеристиками (для варианта с одним виртуальным каналом):

- Сложность: 20 тыс. эквивалентных логических элементов (LE)
- Суммарный объем занимаемой внутренней памяти ПЛИС: 100кбит
- Задержка: < 20 нс
- Частота работы ПЛИС: 250МГц

Заключение

Прогресс в развитии технологии программируемых микросхем позволяет использовать их для создания относительно дешевых конкурентоспособных высокопроизводительных сетей. У разработчиков суперкомпьютеров появилась возможность экспериментировать с важнейшим компонентом машины, где раньше все зависело лишь от нескольких крупных производителей, не отказываясь при этом и от традиционных решений.

Авторами выполняется работа по реализации в ПЛИС вышеописанного маршрутизатора сети типа 3D-тор. Работы ведутся в рамках суперкомпьютерной программы "СКИФ-ГРИД" Союзного государства [7]. Также авторы выражают благодарность за поддержку РФФИ, грант №08-07-00280-а.

ЛИТЕРАТУРА:

1. *Altera Stratix IV FPGA: transceiver overview* // Веб-ресурс: http://www.altera.com/products/devices/stratix-fpgas/stratix-iv/transceivers/stxiv-transceivers.html.
2. N. R. Adiga, et.al, *"Blue Gene/L Torus Interconnection Network*" // IBM J. Research and Development, March/May 2005
3. H. Baier, et.al, *"Status of the QPACE Project"* // ArXiv e-prints 0810.1559, Oct 2008 http://adsabs.harvard.edu/abs/2008arXiv0810.1559B
4. Duato J., *"A necessary and sufficient condition for deadlock-free routing in wormhole networks"* // IEEE Transactions on Parallel and Distributed. Systems, Vol. 6, No 10, pp. 1055-1067, 1995

5. V. Puente, C. Izuy, R. Beivide, J.A. Gregorio, F. Vallejo, J.M. Prellezo, *"The Adaptive Bubble Router"*, 2000
6. Steven L. Scott, Gregory M. Thorson., *"The Cray T3E Network: Adaptive Routing in a High Performance 3D Torus"* // HOT Interconnects IV, Stanford University, August 15-16, 1996, 10 pp.
7. *Суперкомпьютерная программа "СКИФ-ГРИД" (2007-2010 гг.)* // Веб-ресурс: http://skif.pereslavl.ru/psi-info/rcms-skif/index.ru.html#программа_СКИФ-ГРИД