Организация параллельных вычислительных систем на основе протокола PCI Express

А.В. Сухих

Поволжский государственный технологический университет

Потребности в высокопроизводительной системе коммутации с высокой надежностью транзакций и низкими показателями задержек возникают не только при организации систем хранения. Они находят свое применение при решении специфических задач, например, при организации кластерных вычислений.

Для параллельных вычислительных систем (ПВС) очень важна правильная организация системы взаимосвязи компонентов, следовательно, коммутационная система непосредственно влияет на производительность всего вычислительного кластера. С развитием специализированных высокопроизводительных коммуникационных сред, таких как Fibre Channel, InfiniBand и PCI Express, разница между производительностью суперкомпьютеров, построенных, например, на MPP-системах, и кластеров стала резко сокращаться. Существуют различные подходы для организации межкомпонентного взаимодействия внутри кластерной системы в зависимости от её масштаба.

Один из подходов к созданию кластерной системы, основанный на прямой коммутации каналов, был разработан в НИИ «Квант» [1]. Этот подход подразумевает наличие коммутатора PCI Express, к которому подключаются вычислительные узлы, один из которых является ведущим. Этот узел формирует для остальных узлов «окна перекрестного доступа», через которые всем ведомым узлам предоставляется доступ к общему адресному пространству. В этом случае ведущий узел представляет единую точку отказа, что снижает надежность всей системы, кроме того для него требуется большой объем оперативной памяти.

Другой способ создания топологии вычислительного кластера на базе шины PCI Express предложен американской компанией IDT [2]. Этот подход подразумевает модификацию протокола PCI Express для возможности трансляции адресов в непрозрачных узлах, что приводит к возрастанию нагрузок на приемо-передатчики, установленные на вычислительных узлах.

Предлагаемая автором система представляет собой кластер, состоящий из вычислительных модулей, поддерживающих использование внутрисистемной шины PCI Express. Все вычислительные модули коммутируются в единую вычислительную сеть посредством применения специализированной кабельной системы, разработанной в соответствии с кабельной спецификацией PCI Express. Для возможности подключения устройства в сеть оно должно иметь специализированный адаптер шины, к которому подключается кабель. Ядром системы является специализированный коммутатор, который строит таблицу маршрутизации на основе проходящего через него трафика. Все порты коммутатора выполнены по технологии непрозрачного моста. Данное решение выгодно отличается от системы с прямой коммутацией каналов тем, что упрощается процесс построения вычислительной сети. Все узлы равнозначны между собой, все используют одинаковый набор оборудования и вся нагрузка по коммутации трафика (построение адресных таблиц и т.д.) лежит только на коммутаторе

Дальнейших исследований требуют вопросы механизма трансляции адресов, разработки оптимальной структуры таблиц маршрутизации и алгоритмов анализа трафика.

Литература

1. Заявка на пат. № 2011128993/08 РФ МПК G06F15/163. Кластерная система с прямой коммутацией каналов / Четверушкин Б. Н., Смольянов Ю.П., Лацис А.О. и др. – Опубл. 13.07.2011
2. Patent US8429325 B1. PCI Express switch and method for multi-port non-transparent switching/ Peter Z. Onufryk, Cesar A. Talledo; заявитель и патентообладатель Integrated Device Technology Inc. – Опубл. Apr. 23, 2013