***В.С. Горбунов [2], А.О. Лацис [1], А.Н. Иванов [2]***

# О ПОСТРОЕНИИ СУПЕРКОМПЬЮТЕРОВ НА ОСНОВЕ ИНТЕРФЕЙСА PCE-EXPRESS

*[1] Институт прикладной математики им. М.В. Келдыша РАН, г. Москва, Россия,*
*[2] ФГУП «НИИ «Квант», г. Москва, Россия,*
*vitech@rdi-kvant.ru*

## Введение

Для решения задач высокой вычислительной и емкостной (размер обрабатываемых данных) сложности широко используются параллельные вычислительные системы (вычислительные кластеры), создаваемые с использованием универсальных процессоров в рамках SMP-, NUMA- и MPP- архитектур.

При этом хорошо известны достоинства и недостатки каждого из подходов. Однако наиболее важным принципиальным отличием этих подходов является устройство коммуникационной системы. Традиционно при построении вычислительных кластеров используются коммуникации на уровне набора системной логики микропроцессоров (внутриузловые коммуникации), оригинальные коммуникационные системы межузловых обменов (например сеть Sea Star ф. Cray) или широко распространенные сетевые решения (например Infiniband).

В настоящем докладе рассмотрены возможности использования коммуникационной системы, созданной на основе периферийного интерфейса PCI Express. Разработанная коммуникационная система позволяет реализовать ряд функциональных преимуществ суперкомпьютеров верхней ценовой категории, создаваемых на основе дорогостоящих оригинальных коммуникационных систем. При этом сохраняется основное преимущество вычислительных кластеров, высокая доступность и низкая стоимость.

Эксперименты, проведенные на образце разработанной коммуникационной системы, показывают низкую латентность (немногим более 1мкс) передачи данных и исключительно малую задержку выдачи небольших порций данных (70-80нс).

## Базовая модель параллельного программирования

Для реализации параллельных алгоритмов на SMP и ccNUMA-архитектурах чаще всего используют так называемую парадигму программирования с общей памятью (shared memory paradigm).

Основной проблемой программирования в парадигме с общей памятью является сложность реализации операций синхронизации с памятью и высокая длительность операций чтения данных. Так, например, в современных системах на базе PCI Express минимальное время записи – менее 0.1 мкс, а минимальное время чтения может достигать 2 мкс и более. Такой медленный доступ, даже эпизодический, оправдывает себя только при наличии возможности кэширования. Но это означает, что для полного использования парадигмы общей памяти, дающей действительно заметные преимущества программисту по сравнению с системами без общей памяти, эта общая память должна быть кэш-когерентной. Наличие кэш-когерентной общей памяти помимо преимуществ таит в себе и существенный недостаток – необходимость обеспечивать корректную работу параллельных процессов с ней, реализуя механизмы синхронизации ячеек путем выполнения дорогостоящих операций блокировки и разблокировки. Преимущество в программировании входит в противоречие с масштабируемостью приложений.

Эти совсем не сложные рассуждения приводят нас к весьма нетривиальному выводу большой практической значимости. Невзирая на то, что формально новые технологии позволяют строить системы с общим полем памяти, доступным непосредственно по машинному указателю, непосредственное использование таких систем прикладным программистом невозможно. Без кэш-когерентности однородность доступа к общему полю памяти обесценивается, построенные на его базе технологии программирования могут дезориентировать пользователя.

Для промежуточного (между SMP-системами и традиционными кластерами) класса коммуникационного оборудования специалистам ИПМ им. М. В. Келдыша РАН удалось сформулировать промежуточную (между Open MP и MPI) по простоте, удобству и эффективности модель параллельного программирования. Эта модель оказалась хорошо известной в теории параллельного программирования моделью односторонних обменов, глубоко оптимизированной по задержкам и латентности. Она является простейшим, базовым вариантом также хорошо известной в теории модели PGAS (Разделённое глобальное адресное пространство). Именно эта модель оказалась наилучшей абстракцией именно той «степени общности» общей памяти, которую нам реально дают упомянутые выше новые коммуникационные технологии (PCI Express) с их низкой латентностью доступа в «чужую» память без участия процессора – «хозяина» этой памяти, но без кэш-когерентности.

Базовым программистским интерфейсом здесь является не доступ в чужую память по машинному указателю, а библиотека shmem, обладающая всеми свойствами рассмотренной в докладе модели программи-

рования. Это библиотека односторонних обменов, рассчитанная на сеть с очень низкими уровнями латентности и задержки на запуск обмена.

**Макет минисуперкомпьютера «МВС-Экспресс»**

ИПМ им. М. В. Келдыша совместно с ФГУП «НИИ «Квант» был разработан макет суперкомпьютера «МВС-Экспресс».

В поиске наилучшей реализации описанной схемы разработчики МВС-Экспресс рассматривали два варианта. В обоих пространственно разнесенные части системы коммутации соединялись медными кабелями длиной 1-2 метра.

В первом случае логика коммутатора и непрозрачных мостов была реализована в FPGA. Построенный работоспособный макет был признан дорогим и нетехнологичным.

Во втором случае решено было использовать микросхемы готовых коммутаторов PCI Express фирмы PLX. Непрозрачные мосты конструктивно выполнены в виде «сетевых плат», коммутатор на 8 линков – в отдельном корпусе, соединение – стандартными медными кабелями Infiniband длиной 2м. При необходимости объединить более 8 узлов возможно построение решетки коммутаторов, каждый из которых обслуживает 4 узла. Коммутаторы соединяются между собой оставшимися каналами.

Этот вариант параллельной системы производится ФГУП «НИИ «Квант» и в настоящее время используется в ИПМ им. М. В. Келдыша РАН в качестве минисуперкомпьютеров гибридной архитектуры.

***В.С. Горбунов, В.К. Левин, С.В. Яблонский***

## СУПЕРКОМПЬЮТЕРЫ СЕГОДНЯ И ЗАВТРА

*ФГУП НИИ «Квант», г. Москва, Россия,*
*info@rdi-kvant.ru*

1. Создание и применение наиболее мощных вычислительных систем – суперкомпьютеров – является существенным фактором научно-технического прогресса и показателем стратегического потенциала, входит в число важнейших приоритетов развитых стран. Выполнение больших расчетов и компьютерное моделирование сложных конструкций и процессов становится необходимым условием проведения современных исследований и разработок в различных областях.

Специфика суперкомпьютеров как некоторого класса средств вычислительной техники определяется их предназначением для решения задач, которые не могут быть решены на других средствах из-за вычислительной сложности и большого объёма обрабатываемых