



Архитектура глобальной памяти с общим доступом SGI NUMAflex и интерконнекта SGI NUMAlink в серверах и суперкомпьютерах Silicon Graphics

Национальный дистрибьютор Silicon Graphics в России, группа компаний Arbyte, поставляет полный спектр серверных и суперкомпьютерных решений от SGI — серверы и суперкомпьютеры семейства Altix на 64-разрядных процессорах Intel Itanium 2: Altix 330 (1-16 процессоров), Altix 350 (1-32 процессора), Altix 3700 (до 512 процессоров) и уникальный Altix 4700 — блейд-сервер (до 512 процессоров). Все они основаны на архитектуре глобальной памяти с общим доступом SGI NUMAflex и интерконнекте SGI NUMAlink. Об этих технологиях мы и расскажем в данной публикации.

Архитектура глобальной памяти с общим доступом SGI NUMAflex

Архитектура глобальной памяти с общим доступом SGI NUMAflex обеспечивает работу с оперативной памятью до 24 Тбайт, при этом вся память в системе может быть адресована любым процессором.

Большинство пользователей персональных компьютеров знают, что они могут получить большие преимущества от добавления в систему оперативной памяти, чем от наращивания процессорной мощности, — и это в еще большей степени верно для суперкомпьютеров. Серверы и суперкомпьютеры SGI Altix обладают наиболее масштабируемой архитектурой в отрасли, предоставляя до 24 Тбайт адресуемой памяти на одну систему, что сильно контрастирует с кластерными архитектурами, в которых адресуемая память обычно ограничена объемом в 32 Гбайт.

Системы с глобально адресуемой памятью обеспечивают эффективный прямой доступ ко всем данным в памяти, исключая необходимость перемещения данных через

средства ввода-вывода или через перегруженные сетевые каналы. При этом достигается потрясающая производительность, так как латентность, или время ожидания при связи с доменом памяти, в 1000-10 000 раз меньше, чем при работе через системные средства ввода-вывода или сетевые каналы. Для преодоления этого недостатка кластеры, не имеющие глобальной памяти с общим доступом, вынуждены перемещать копии данных (часто в форме сообщений), что крайне осложняет программирование и снижает производительность, увеличивая время, которое процессоры вынуждены проводить в ожидании данных.

Почему глобальная память важна? Пользователи могут сделать больше и получить результаты раньше. Целые базы данных могут находиться непосредственно в памяти. В комплексе аппаратное и программное обеспечение зачастую становится дешевле. Системные администраторы и пользователи тратят меньше времени на загрузку и настройку системы, нежели в случае кластера. Разработчики имеют больше пространства для маневра при выборе программной модели и обеспечении наращиваемости приложения.

Рассмотрим свойства и преимущества интерконнекта SGI NUMAlink, обеспечивающего функционирование архитектуры глобальной памяти с общим доступом SGI NUMAflex и фактически неотделимого от нее.

Свойства и преимущества интерконнекта SGI NUMAlink

С микропроцессорами, способными совершать более миллиона операций в секунду (меньше 1 нс на операцию), быстрый доступ к памяти важен для достижения сбалансированной, устойчивой производительности при обработке технических задач. Данные, передаваемые через коммутатор SGI NUMAlink, совершают полный оборот всего за 50 нс (это меньше, чем требуется лучу света для прохождения 160 м) по сравнению с 10 000 нс и более, характерных для многих стандартных кластерных интерконнектов. Более того, технология SGI NUMAlink является единственным интерконнектом, обеспечивающим режим глобальной памяти с общим доступом между узлами.

Наивысшая в отрасли производительность технологии интерконнекта NUMAlink становится очевидной при сравнении ее показателей с параметрами полосы пропускания и латентности других технологий (табл. 1). Это ведет к большей производительности всей системы на приложениях MPI, а также на стандартных отраслевых тестах, таких как Linpack (табл. 2).

Для лучшего понимания преимуществ глобальной памяти с общим доступом приведем примеры ее практического применения в различных отраслях.

Первое из этих преимуществ — ускорение работы приложений:

- **лучшие в отрасли результаты MPI** — с параметром в 1 микросекунду технология интерконнекта SGI NUMAlink является отраслевым лидером по латентности и ширине полосы пропускания MPI (6,4 Гбайт/с, дуплекс на линк), обеспечивая исключительную про-



NUMAflex против традиционной кластерной архитектуры



изводительность на кодах MPI. Это подтверждает беспрецедентная производительность SGI Altix на широко применяемых в мире кодах для предсказания погоды mesoscale: WRF и MM5 (см. <http://www.mmm.ucar.edu/mm5/mpp/helpdesk/20040304a.html>);

- **наращиваемость до 512 процессоров и более** — в NASA при минимуме усилий была достигнута экстраординарная масштабируемость наиболее широко применяемых приложений в значительной степени благодаря тому, что в агентстве описывают как «тривиальную» природу балансирования нагрузок на 512-процессорных системах Altix с глобальной памятью с общим доступом (см. <http://www.nas.nasa.gov/About/Projects/projects.html>). Благодаря этой модели время программирования при масштабировании и оптимизации крайне больших комплексных приложений сократилось с многих человеко-лет до нескольких недель. Результаты работы 512-процессорной системы достигли и превзошли результаты системы Earth Simulator стоимостью 500 млн. долл.;

- **биоинформатика** — в расположенном в Великобритании фармацевтическом исследовательском институте был разработан алгоритм поиска геномных совпадений, который выдает результат в тысячу раз быстрее, чем обычно используемое приложение BLAST. Алгоритм полагается на память в 192 Гбайт, имеющуюся на каждом из двух серверов SGI Altix с 4 и 16 процессорами соответственно.

Второе преимущество глобальной памяти с общим доступом — возможность наращивания производительности при ограниченном количестве процессоров:

- **NWChem** — у одного из клиентов SGI приложение NWChem всегда обрабатывается на системе SGI Altix, включающей всего четыре процессора, но использует при этом 80% установленной в ней глобальной памяти с общим доступом размером в 2 Тбайт. В то время как вычисление не может быть распараллелено более чем на несколько процессоров, существенное преимущество достигается благодаря простому увеличению объема памяти;
- **MSC.Nastran** — определенный класс задач MSC.Nastran может требовать больших объемов памяти, но не может быть распараллелен. На этом коде достигается существенное ускорение, когда обрабатываемая задача имеет доступ к большему объему памяти с общим доступом.

Третье преимущество — возможность взаимодействия в реальном времени с массивными пакетами данных. Например, **взаимодействие с большими пакетами сейсмических данных** — перед компанией Marathon Oil стоит задача поставлять геофизикам для интерпретации всё большие и большие объ-

Таблица 1. Латентность и полоса пропускания различных интерконнектов

Технология	Поставщик	Латентность MPI усец, короткие сообщения	Полоса пропускания на линк (все направления, Мбайт/с)
NUMalink 4 (Altix)	SGI	1	3200
RapidArray (XD1)	Cray	1,8	2000 ¹
QsNet II	Quadrics	2	900 ²
Infiniband	Voltaire	3,5	830 ³
High Performance Switch	IBM	5	1000 ⁴
Myrinet XP2	Myricom	5,7	495 ⁵
SP Switch 2	IBM	18	500 ⁶
Ethernet	Various	30	100

Примечание. Данные по полосам пропускания взяты из следующих источников:

- ¹ <http://www.cray.com/products/xd1/index.html#RapidArrayInterconnect>
- ² [http://doc.quadrics.com/Quadrics/QuadricsHome.nsf/DisplayPages/81DD13F71CFD762580256EAD0010AA75/\\$File/Performance.pdf](http://doc.quadrics.com/Quadrics/QuadricsHome.nsf/DisplayPages/81DD13F71CFD762580256EAD0010AA75/$File/Performance.pdf)
- ³ <http://nowlab.cis.ohio-state.edu/projects/mpi-iba/>
- ⁴ <http://publib-b.boulder.ibm.com/Redbooks.nsf/f338d71ccde39f08852568dd006f956d/55258945787efc2e85256db00051980a?OpenDocument>
- ⁵ <http://www.myricom.com/myrinet/performance/>
- ⁶ http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/sp_switch_perf.pdf

Таблица 2. Сравнение производительности систем Linpack из списка Top-500 за ноябрь 2004 года

Система/интерконнект	Средняя эффективность Linpack для 256-процессорных систем, %*	Размер выборки, количество систем в списке *
SGI Altix/NUMalink	84	14
HP Superdome	79	18
Various/Quadrics	75	4
Various/Infiniband	75	3 (одна система с 288 процессорами)
Various/Myrinet	63	19
Various/Gigabit Ethernet	59	14

* Величины Linpack Rmax/Rpeak для 256-процессорных систем на ноябрь 2004 года даны в списке Top-500 — см. www.top500.org.

емы сейсмических данных. Впервые геофизики могут визуализировать и взаимодействовать в реальном времени с более чем 4000 Гбайт данных. Этот показатель более чем в четыре раза превосходит предыдущий рекорд (см. <http://www.lgc.com/>).

Четвертое преимущество — широкие вычислительные возможности:

- **крупномасштабный структурный анализ** — недавно ANSYS объявила, что из компаний, занимающихся инженерным моделированием, она первой решила модель структурного анализа более чем со 100 млн. степеней свободы (DOF), обеспечив своим клиентам возможность обрабатывать при полном разрешении модели авиационных двигателей, автомобилей, строительного оборудования и других цельных систем. Совместными усилиями ANSYS и Silicon Graphics, Inc. (SGI) задача структурного анализа в 111 млн. DOF была решена всего за несколько часов на небольшом количестве процессоров, но при большом объеме памяти, на компьютере SGI Altix. По результатам этой работы ANSYS вступила в трехгодичное партнер-

ство с SGI, нацеленное на развитие возможностей ANSYS в параллельных вычислениях и на больших объемах памяти;

- **анализ данных с быстрым откликом** — правительственное агентство США использует 32-процессорные системы SGI Altix с 4 Тбайт памяти в качестве единственного решения для масштабного сбора данных и работы поисковых алгоритмов, крайне важных для их деятельности;
- **базы данных, находящиеся в памяти**, — Xcelerix IMDB и ускоритель баз данных продемонстрировали на системах SGI Altix рост производительности на порядки по сравнению с традиционными базами данных, размещенными на дисках, для баз, содержащих до 500 млн. записей (120 Гбайт).

С конкретными решениями SGI в области серверов и в других областях (в системах хранения SGI InfiniteStorage, системах визуализации, FPGA-системах реконфигурируемых вычислений SGI RASC, GRID-решениях от SGI) и с обширной областью специализированного программного обеспечения можно ознакомиться на сайте <http://www.silicongraphics.ru>. ▀