Параллельные кластеры: перспективы и применение

В середине февраля с.г. официально объявлено о запуске в эксплуатацию суперкомпьютера "СКИФ Cyberia" на базе 566 двухъядерных процессоров Intel® Xeon® серии 5150 (SN № 4/29, 2006), ставшим самым мощным вычислительным комплексом на территории России, СНГ и Восточной Европы, а также одним из 100 мощнейших компьютеров мира. SN обратился с просьбой высказаться по перспективам применения параллельных кластеров в России к ведущим специалистам, активно использующим и разрабатывающим суперкомпьютерные системы.

Введение

Суперкомпьютеры были и остаются одним из основных инструментов (или технологической платформой) решения стратегических задач крупнейших государств мира, прежде всего, в таких отраслях, как: оборонная, аэрокосмическая, геология (управление запасами недр Земли). Однако в последнее время сфера их применения стала значительно расширяться. Они стали широко использоваться в фармакологии, машиностроении и даже в таких областях, как легкая промышленность, являясь уже сами по себе локомотивом внедрения новых достижений в промышленности и научных исследованиях.

Крупнейшие разработчики аппаратного и программного обеспечения, до недавнего времени не специализирующиеся на разработках для параллельных вычислений, объявили их одними из приоритетных и активно их развивают. Так, например, компания Microsoft – крупнейший разработчик ПО для массового рынка – в октябре 2006 г. объявила о выходе в России комплексного решения для параллельных вычислительных систем: Windows® Compute Cluster Server 2003 (WCCS) – http://www.microsoft.com/rus/hpc. Среди преимуществ WCCS следующие: упрощенная система развертки и управления кластером; улучшенный процесс интеграции с существующей IT-инфраструктурой; поддержка ключевых технологий (64-процессорные кластерные вычислительные узлы; интерфейс MPI2); 1 Gbit Ethernet, InfiniBand и др.).

С прошлого года на рынке стали доступны первые настольные параллельные кластерные системы для домашнего использования.

11 февраля с.г. компания Intel объявила о разработке исследовательского 80-ядерного чипа с пиковой производительностью 1 Тфлоп/с и потребляемой мощностью всего 62 Вт. Появление систем на базе таких чипов, которые Intel пока не планирует делать доступными для рынка, позволяет на новом уровне подходить к решению многих задач глобальной геополитики, в сжатые сроки решать сложнейшие отраслевые задачи и совершенно по-новому смотреть на мир.

Использование сверхмощных вычислительных систем в отдельных отраслях, как правило, ведущих, обеспечивает совершенно новый градиент (который недостижим простой консолидацией менее мощных систем, хотя, например, в ядерной физике GRID-вычисления активно используются) не только их развития, но и целого ряда смежных отраслей. Экономический эффект подобных внедрений суперкомпьютерных систем может на порядки превосходить первоначальные затраты на их имплементацию.

Основную долю суперкомпьютерных систем в мире занимают параллельные кластерные системы (72% списка Top500). Распределение кластерных систем по отраслям в России&СНГ представлено в табл. 1 (http://www.supercomputers.ru/?page=rating). Абсолютные значения рей-

Табл. 1. Использование параллельных кластеров по отраслям в мире и в России&СНГ

Области использования	В мире (Top-500)	Top50 (4-я ред.)	Top50 (5-я ред.)
Промышленность	41,8%	28%	30%
– электронная	14,2%	0%	0%
– тяжелая (автомобильная, авиационная, аэрокосмическая, металлургия и др.)	10,2%	12%	12%
– добывающая (геологоразведка, нефте- и газодобыча)	10.4%	16%	18%
– медиа (цифровые технологии)	7%		
Исследования стратегические – ядерные, космические, в области безопасности, окружающей среды; прикладные в различных областях, в т.ч. вычислительных технологий	27,2%	16%	18%
Наука и образование в т.ч. биоинформатика и молекулярная динамика для разработки новых лекарств, нанотехнологии, астрофизика, квантовая химия, математика и др.	16,6%	20%	22%
Финансы банки, финансовые компании, страхование, финансовые прогнозы и консалтинг	9,6%	32%	28%
Управление, в т.ч. корпоративные бизнес-приложения, государственное управление	2,2%	–	–
Телекоммуникации различные сервисы для больших баз данных абонентов	2,6%	–	–

Табл. 1. Сравнение по производительности мирового суперкомпьютерного рейтинга Top500 и рейтинга самых мощных компьютеров СНГ Top50

Отрасли экономики	Верхний уровень (GFlops)			Средний уровень (GFlops)		
	в мире	Top50-4	Top50-5	в мире	Top50-4	Top50-5
Исследования	280600 (Пиверморская Национальная Лаборатория, США)	6680	6680 (меньше в 42 раза)	~3500	260	260 (меньше в 13 раз)
Финансы	4924 (Финансовые услуги, Великобритания)	642	642 (меньше в 7,6 раз)	~2500	250	250 (меньше в 10 раз)
Промышленность	12312 (Геофизика, США)	768	768 (меньше в 16 раз)	~2800	250	250 (меньше в 11 раз)

тинга Top-50 (табл. 2) ниже общемирового, однако и они не дают возможности понять многих перспектив применения суперкомпьютерных систем в России.

Прокомментировать эту проблематику SN любезно согласились ***Георгий Гогоненков*** *– д.т.н., первый зам. ген. директора Центральной Геофизической Экспедиции (ЦГЭ),* ***Юрий Зеленков*** *– зам. директора по IT НПО "Сатурн", а также* ***Андрей Слепухин*** *– руководитель "Центра кластерных технологий "Т-Платформы".*

Мощные кластерные системы – это геологоразведка новых сложных месторождений нефти, острая потребность в которой возникнет уже через 5 лет

SN. *Чтобы лучше понимать проблематику геологоразведки и сейсморазведки, в частности, а также что связано с обработкой геофизической информации, вкратце расскажите, пожалуйста, об их истории развития.*

Георгий Гогоненков – д.т.н., первый заместитель генерального директора "Центральной Геофизической Экспедиции".

Г.Г. Самые первые шаги, которые мы делали в 70-х годах – это уточнение с помощью сейсморазведки структурного строения геологического разреза: расположения по глубине сейсмических границ, их повышение, понижение или структурные нарушения. Однако сейсмический сигнал обычно регистрировался до глубин 5 км, а дальше возникали проблемы, т.к. мы не могли выделить сигнал на фоне помех.

Первый этап, который сделала сейсморазведка в России в 80-х годах с переходом на цифровую обработку – это возможность выделять принципиально новые структурные горизонты. Например, в районе Поволжья стали картироваться девонские горизонты, содержащие многомиллиардные запасы нефти, которые были вовлечены в разработку в последние 30 лет. В Западной Сибири стала картироваться двухкилометровая толща юрского комплекса. В Закавказье это была толща триаса. Этот этап закончился в начале 90-х годов.

Второй этап длился примерно 10 лет – до 2000 г. В это время геофизика перешла на промышленное изучение свойств продуктивных пластов: их пористости, толщины, а в благоприятных случаях удавалось определять, чем насыщен пласт – нефтью, газом или водой.

Третий этап – с 2000 г. по настоящее время – обеспечил повышение разрешающей способности сейсморазведки. Если раньше мы могли прогнозировать пласты толщиной 25–20 м, а в лучшем случае – 15 м, то теперь стоит задача прогноза пластов 5- и даже 3-метровой толщины.

Соответственно, каждый технологический шаг имел в качестве своей предпосылки увеличение вычислительной мощности. И поскольку ЦГЭ первоначально была создана для разработки соответствующего ПО и технологии компьютерной обработки результатов геологоразведки, то с первых шагов мы четко отслеживали зависимость качества геологического результата от числа машинных операций на единицу геофизической информации.

Сейчас на одно 4-байтное слово геофизической информации в среднем приходится несколько сот тысяч операций. И каждый год эта цифра увеличивается в 3–5 раз. При этом возможности однопроцессорных машин мы исчерпали 7–8 лет назад, и перешли на кластеры, первые из которых были 8–16-процессорные. Сейчас мы работаем с 64-процессорным кластером, а в ближайший год планируем перейти на 256-процессорный. Для сравнения: крупнейшие геофизические западные компании имеют вычислительные системы с числом 10–20 тысяч и более процессоров. Эта разница обусловлена тем, что геологоразведочные работы в России пока в основном выполняются на суше, а на Западе – на море. Сам процесс морской сейсморазведки намного производительнее. К судну крепят "косу", причем не одну, а несколько, и такой "караван" одновременно регистрирует 8–10 тысяч сейсмических трасс по ходу движения судна. Т.е. за 30 с достигается сбор такого объема информации, который на суше мы получаем за полдня. Поэтому объем морской геоинформации больше, чем получаемый на суше, примерно, на 2 порядка. Я думаю, что к потребности, измеряемой тысячами процессоров, мы подойдем примерно через 5 лет.

А если говорить о геофизике в целом, то по объему применения кластерных систем в мире она стоит на 2-3 местах после военных и космических отраслей.

SN. *Хотелось бы уточнить, есть ли ограничения в распараллеливании геофизических задач?*

Г.Г. Практически нет. Геофизические задачи хорошо распараллеливаются, потому что мы работаем с большим количеством единиц информации, которой является сейсмическая трасса. Она представляет собой цифровую последовательность колебаний, которые приходят к приемнику или группе приемников, ее объем порядка 2–3 тыс. 4-байтных слов. Это наша единица. Таких слов от одного возбуждения мы получаем несколько сот или несколько тысяч. А потом, когда мы делаем обработку, для получения одного выходного слова мы суммируем 100–200 тыс. таких единичных трасс. Для получения каждой выборки (или одного выходного слова) нам необходимо одновременно манипулировать с 10 млн. выборок. А общий объем выходной информации, которую мы получаем для одной "единичной площади", составляет порядка 100 Мбайт.

На Западе используют кластеры с двумя и большим числом процессоров в чипе, загружая один процессор на 100%, другой – на 35% и меньше. Мы стараемся работать с однопроцессорными системами, но по мере совершенствования алгоритмов переходим и к многоядерным процессорам.

Следует различать два комплекса задач, с которыми мы работаем. Один комплекс – это обработка данных, смысл которого заключается в том, чтобы выделить сигнал на фоне помех. Эти задачи забирают львиную долю машинного времени, но все делается в поточном режиме. Однако есть и второй класс задач, который мы называем интерпретационной обработкой. Смысл их заключается в обработке выделенного сигнала с целью извлечения из этих данных геофизической информации – получения параметров геологического разреза – слоев, их характеристик. Для этого созданы специализированные системы, и до последнего времени они были достаточно простыми – в основном однопроцессорными. Но сейчас и в этой области происходит коренной перелом. У нас появились стохастические решения, которые проверяются на генерации большого количества вариантов реализаций, на которых мы и оцениваем стохастические параметры. Это не один параметр плюс его ошибка. Мы создаем "облако" распределений возможных параметров и при этом более точно рассчитываем сам параметр и зону неопределенности. И такие задачи тоже становятся процессороемкими. Сейчас мы работаем с 8-процессорными системами, в ближайшее время переходим на 16-процессорные. Т.е. здесь примерно с отставанием на порядок (в сравнении с системами для выделения сигнала) начинает активизироваться применение кластерных систем. Кстати, для справки, число сотрудников, которые у нас занимаются подобными задачами, раз в 5-6 больше числа персонала, связанного с поточной обработкой на высокоразмерных кластерах.

SN. *Сейчас ваш "вычислительный" потенциал ограничен средствами, которые вы можете выделить на эти цели. Если абстрагироваться от этого ограничения, то что, по Вашему мнению, это дало бы и нужно ли вообще увеличивать этот вычислительный потенциал сегодня? Например, есть ли потребность в этом с точки зрения повышения разрешенности геологических разрезов?*

Г.Г. При решении геофизических задач практически нет предела потребности в вычислительных мощностях. Но если говорить обобщенно, то, если бы мы смогли на порядок увеличить вычислительную мощность нашего 64-процессорного кластера, то при том же количестве персонала и прочих равных параметрах мы смогли бы в 1,5 раза повысить геологическую эффективность своих работ. Здесь нужно учитывать несколько факторов.

Во-первых, повышение вычислительной мощности это, прежде всего, повышение точности геофизических работ. А это приведет к уменьшению числа разведочных скважин, каждая из которых стоит $1–3 млн и даст возможность проектирования более эффективных – горизонтальных скважин. Например, если вертикальная скважина может дать 20 т нефти в сутки, то горизонтальная – уже 300–500 т. Она, конечно, дороже, например, в 3 раза, но зато нефти может дать уже в 10 раз больше. Чтобы спроектировать такую скважину, нужно очень точно знать расположение и характеристики пласта, т.к. на глубине 1,5–2 км необходимо попасть в пласт толщиной 5–3 м.

Месторождения с толщиной нефтяного слоя несколько сот метров в мире можно пересчитать по пальцам. В большинстве случаев толщина нефтяных пластов составляет несколько метров, но их в месторождении может быть несколько, и каждый нужно отдельно изучать, и для каждого нужно строить отдельную систему разработки. Именно такой подход обеспечивает цивилизованную добычу, сводящую к минимуму неоправданные потери углеводородного сырья. Поэтому есть прямая связь между развитием геофизики и увеличением эффективности добычи углеводородов.

Другой аспект этого же направления, связанного с точностью обработки сейсмоданных, заключается в стремительно развивающейся алгоритмической базе обработки, что в среднем через 4 года делает экономически оправданной переобработку старых сейсмических материалов. И в нашем объеме работ переобработка данных уже занимает 30–40%.

Во-вторых, это возможность освоения новых территорий. Задачи геофизики очень многомерны. В одних случаях, на существующем оборудовании мы можем рассчитывать пласты толщиной в 3 м, а в отдельных – нельзя говорить и о 15 м потому, что там очень сложные условия в верхней части разреза. Например, условия геологоразведки и, соответственно, добычи в районах Западной и Восточной Сибири существенно отличаются. Западная Сибирь это гигантский бассейн с нефтедобычей, находящейся сегодня в затухающей фазе, близкой к спаду. Если не в ближайшие годы, то через 10–15 лет там начнут круто уменьшаться объемы добычи, потому что основные месторождения уже вовлечены в разработку.

Есть гигантская территория в Восточной Сибири. Здесь открыты отдельные месторождения, но проблема в том, что на этой огромной территории разлиты неоднородные лавы базальта. И базальт как экран или подушка (толщиной от 15 м до 1,5 км) не дает возможности сейсморазведчикам “пробиться” к осадочным породам, которые содержат нефть. Поэтому качество геофизических работ и информативность получаемых данных в этом регионе намного ниже, чем в Западной Сибири. И здесь, опять-таки за счет использования более мощных компьютеров, появляется возможность вовлечь эти перспективные территории в георазведку. В Восточной Сибири еще будут сделаны очень серьезные открытия.

SN. *Поскольку SN активно продвигает технологии хранения и управления данными, хотелось бы узнать о ваших потребностях в объемах хранения при решении задач геофизики.*

Г.Г. В среднем, каждый наш проект требует 1–2 Тбайт дисковой емкости. Прежде всего, требования к объемам памяти возрастают из-за необходимости хранить несколько различных вариантов обсчитываемых данных. После завершения проекта все сбрасывается на долговременные носители. Всего у нас в центре сейчас около 15 Тбайт, и нам этого мало. А всего в своих архивах мы храним порядка 60 Тбайт информации.

SN. *Когда мы говорим о будущем геологоразведки, в качестве основного инструмента имеется в виду только сейсморазведка или будут использоваться другие методы – в частности, авиа- и космические исследования?*

Г.Г. Сегодня и в просматриваемой перспективе сейсморазведка является основным методом геологоразведки как на суше, так и на море, позволяющим очень точно изучать строение геологических пластов. Авиа- и космическая разведка – это, в основном, региональные методы изучения. Если взять за 100% весь объем геологоразведки, то на сейсморазведку будет приходиться порядка 75–80% из десятка других методов. А потребности в вычислительных мощностях сейсморазведки раз в 100 больше, чем всех остальных методов, вместе взятых.

SN. *Можно сопоставить результаты геологоразведки и геофизики при одних и тех же затратах?*

Г.Г. Как мы уже говорили, стоимость одной скважины в Западной Сибири составляет $2–3 млн. При этом, согласно современной системе категоризации запасов, считается, что в радиусе 1 км от скважины свойства пласта изучены. А в геофизике за $3 млн можно произвести “съемку” 100 км² площади, что по информативности эквивалентно порядка 20 скважинам, Но при этом необходимо учитывать, что обработка данных это наиболее дешевая составляющая геофизической разведки. Самая дорогая и тяжелая – это сбор данных: сложная работа в тайге в зимних условиях, необходимость доставки большого количества тяжелой техники (практически минизавода на колесах – 2-х десятков тракторов, гигантских вибраторов, 10 тыс. сейсмоприемников, буровых станков и др.).

SN. *Как Вы уже говорили, на Западе достаточно активно проводятся морские геоисследования. Почему же в России актуальность этого направления не столь высока?*

Г.Г. В геологоразведке всегда идут от простого к сложному. И пока у нас были большие запасы на суше, а разработка на суше дешевле, чем на море, мы и концентрировались на суше. Все морские работы в Советском Союзе велись, в основном, в Азербайджане. Но сейчас ситуация меняется. Учитывая то, что нефтяной бассейн в Западной Сибири, в основном, разведан, интерес к разведке в шельфовой зоне значительно возрастает. Всем известно Штокмановское месторождение в Баренцевом море, и мы уверены, что там могут быть найдены еще 5–6 такого же масштаба месторождений. Баренцево и Карское моря это объекты номер один плюс Охотское море и Сахалин (сейчас там уже активно ведут работы российские и западные компании). А дальше геологоразвдка будет двигаться на восток и северо-восток – это Восточно-Сибирское море.

SN. *Осознают ли нефтяные компании, что долго “продержаться” им на старых запасах не удастся и начали ли они вкладывать деньги в георазведку?*

Г.Г. Начали, но не очень активно. Пока мы все же находимся только у истоков этого процесса, он еще впереди. Если в Западной Сибири процесс добычи находится на затухающей, то в Волго-Уральском районе – уже почти завершился и значение этого района намного меньше, чем Западной Сибири. Сейчас на первом плане геологоразведки – Восточная Сибирь, а далее – шельф.

SN. *Сейчас много разговоров о том, что разведанных запасов углеводородов в России хватит на 10–15 лет. Соответствует ли это действительности?*

Г.Г. Да, это действительно так. Если мы говорим о разведанных запасах, то через 15–20 лет они подойдут к концу. Вывод из этого только один: надо вкладывать деньги в будущие разработки. Прогнозы, что нефть кончится, давали и 20 и 30 лет назад, но все это время открывали новые месторождения, и запасы росли. Я думаю, что и сейчас ресурсов нефти и газа в земле достаточно много. И лет 50 человечество еще может жить спокойно. А вот по истечении этого срока вопрос о замене энергоресурсов может стать очень острым, а лет через 80 может резко встать вопрос уже о замене химического сырья.

Высокопроизводительные кластеры позволят сократить затраты на разработку авиадвигателя до 10 раз, а время – до двух лет

SN. *Пожалуйста – вкратце о проблемах при разработках новых авиадвигателей.*

Ю.З. Для современной авиадвигателестроительной компании инженерные расчеты являются одним из ключевых элементов бизнеса, поскольку только виртуальное моделирование конструкции позволяет сократить затраты на ее доводку “в металле”. Цикл проектирования нового изделия можно разбить на 2 этапа: первый – разработка конструкции, второй – изготовление опытных образцов, их испытание и доводка до необхо-

димых параметров. Соотношение затрат времени и ресурсов между этими этапами приблизительно оценивается 1:10. Таким образом, основной эффект от внедрения информационных технологий (прежде всего CAD/CAM/CAE) в процесс проектирования можно получить за счет уменьшения количества опытных

Юрий Зеленков – заместитель директора по IT НПО "Сатурн".

экземпляров, изготавливаемых на втором этапе, времени их испытания и т.д. Это позволяет сократить общий цикл разработки нового авиадвигателя до 3-4 лет и затраты на его создание – в 5-6 раз. Необходимо стремиться к тому, чтобы вообще избежать испытаний для доводки конструкции, выполняя только испытания, обязательные для сертификации авиационного двигателя.

SN. *Несколько слов о истории развития вычислительных систем на вашем предприятии.*

Ю.З. Создание специализированных систем для высокопроизводительных инженерных расчетов (CAE) на НПО "Сатурн" началось в 2001 г. с приобретения нескольких SMP-серверов на RISC-процессорах. Данные системы были предназначены для аэродинамических и прочностных расчетов узлов и деталей проектируемых двигателей. Опыт их эксплуатации показал, что тщательный просчет конструкции позволяет сократить количество изготавливаемых опытных экземпляров в 3-4 раза. Однако SMP-системы имеют сравнительно высокую стоимость выполнения одной операции с плавающей точкой в секунду, поэтому в 2003 г. был реализован проект по созданию первого кластера на процессорах Intel Xeon, объединенных по технологии SCI.

В 2005 г. на НПО "Сатурн" был введен в эксплуатацию второй вычислительный кластер на основе технологии Infiniband и процессорах Intel Xeon с пиковой производительностью 0,92 Тфлоп/с, который занял 4-е место в списке 50 самых производительных суперкомпьютеров и стал самым производительным в промышленности России и СНГ *(www.supercomputers.ru)*. Появление данной системы позволило стандартизировать процедуры 3D-аэродинамических расчетов на установившихся режимах и 3D-анализ динамической прочности и передать выполнение данных функций из специализированного подразделения инженерного анализа в конструкторские отделы, отвечающие за проектирование отдельных узлов. Это дало возможность сократить затраты времени на анализ конструкции, увеличить число специалистов, выполняющих расчеты. В результате в 2006 г. средняя загрузка кластера превысила 95%.

В 2006 г. была также решена задача предоставления удаленного доступа к кластеру по относительно медленным каналам связи. Теперь специалисты из инженерных центров в Москве и Перми могут работать с системами расчетов и визуализации на оборудовании, физически находящемся в Рыбинске. Для обеспечения этой возможности специалистами НПО "Сатурн" было реализовано уникальное решение на базе компрессии протокола X11.

SN. *Если можно, расскажите о перспективах авиадвигателестроения с использованием высокопроизводительных вычислительных систем.*

Ю.З. Следующим этапом в развитии инженерных вычислений является освоение методов многоцелевой оптимизации. Данные методы позволяют найти такие варианты конструкции двигателя, которые имеют оптимальные характеристики одновременно по прочностным, весовым и аэродинамическим параметрам. Однако в процессе поиска решения задачи многоцелевой оптимизации необходимо выполнить значительное количество (500 и более) расчетов вариантов конструкции с различными входными данными. Опыт показывает, что для аэродинамического расчета только одного варианта вентилятора и бустера требуется выполнить более 150 млн операций с плавающей точкой – на кластере производительностью около 1 Тфлоп/с – это занимает более 40 часов.

Вторая перспективная задача – анализ узлов на нестационарных режимах работы двигателя при изменении частоты вращения ротора. Предварительные оценки показывают, что для расчета нестационарного режима двигателя целиком необходимы вычислительные мощности порядка 10 Тфлоп/с.

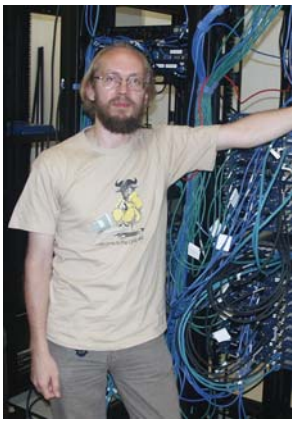
Для решения данных задач на НПО "Сатурн" разработана программа наращивания вычислительных мощностей на 2007–2009 гг.

Создание кластерных систем общего назначения и специализированных с производительностью до 600 Тфлоп/с возможно в России уже сегодня

SN. *Каковы основные мировые направления развития суперкомпьютерной отрасли и возможные параметры кластерных систем в России?*

А.С. Можно выделить 3 направления развития суперкомпьютерной отрасли. Первое – создание стандартных кластерных систем на основе широко применяемых компонентов, использование которых становится все более массовым. Второе – создание единичных систем очень большого масштаба, например, таких как: BlueGene/L (eServer Blue Gene Solution, IBM, US), Red Storm (Sandia/Cray Red Storm, Opteron 2.4 GHz dual core, Cray Inc., US). Они ориентированы, прежде всего, на стратегические исследования. И хотя архитектура этих систем во многом похожа на кластер, и для их построения также широко используются стандартные компоненты, называть их кластером было бы неправильно. Дизайн таких систем разрабатывается индивидуально, и собрать такую систему "самостоятельно" нельзя. Третье направление связано с развитием альтернативных технологий в области высокопроизводительных систем. Это – применение процессоров с нетрадиционной архитектурой (например, процессор Cell, IBM); применение специальных аппаратных ускорителей либо на базе специализированных микросхем, либо микросхем с программируемой логикой; применение графических спецпроцессоров для решения специфических задач.

Если говорить более конкретно о кластерных системах, то сегодня возможно создание систем, включающих до 10 тыс. вычислительных узлов. Для примера: система с 10 тыс. вычислительными узлами (единица конструктива, которая может иметь до нескольких процессоров) к концу этого года может иметь теоретическую пиковую производительность 600–700 Тфлоп/с.



Андрей Слепухин – руководитель Центра кластерных технологий "Т-Платформы".

Создание таких систем возможно не только на Западе, но и в России. И технологически, например, наша компания к этому готова. Такие системы архитектурно будут похожи на существующие кластеры, изменения в архитектуре в первую очередь будут направлены на повышение их надежности, без чего эффективность их эксплуатации может снижаться. Для примера: система Blue Gene, надежность которой на порядки превышает надежность обычных кластеров, в среднем дает сбой 1 раз в 6 дней (по информации одного из открытых источников).

Другая проблема это электропитание. Если взять нашу последнюю систему в Томске, в которой было всего 282 вычислительных узла, то общая потребляемая мощность всей установки, включая затраты на охлаждение, дополнительную мощность, потребляемую самой системой электропитания, составила 150 КВт. Для 1000 узлов это будет уже 600 КВт, для 10 тыс. – 6 МВт, т.е. мощность небольшой электростанции.

SN. *В данном контексте – какова эффективность перехода на блэйд-компоненты?*

А.С. Использование блэйдов не очень эффективно в плане затрат на электропитание. Если говорить о процессорах с пониженным электропитанием, то они

могут быть установлены и в стандартные узлы. Если сейчас 2-процессорный стандартный вычислительный узел потребляет в среднем 400 Вт (процессоры, память, системная плата, блок питания), то с переходом на блэйд-архитектуру (при двух процессорах на узел) можно сэкономить примерно до 90 Вт, или до 23% мощности.

SN. *Как Вы прокомментируете анонсирование Intel 80-ядерного чипа?*

A.C. Во-первых, подобный чип нельзя рассматривать как эквивалент кластера. Прежде всего – это набор специализированных процессоров, не являющихся процессорами общего назначения. Чип представляет собой множества процессоров (в разных конфигурациях) для графики, операций с плавающей точкой, преобразований сигналов и др. Например, 80-ядерный чип может состоять из 50 ядер только для арифметических операций, из 20 – для ввода/вывода, из 10 – для специальных функций. Соответственно, за счет специализации каждого ядра достигается бóльшая эффективность всего чипа с точки зрения пересчета на 1 процессор (меньше транзисторов, меньше потребляемая мощность). Во-вторых, появление систем на базе таких чипов может вызвать настоящую революцию в программировании. Для того чтобы писать программы для подобных систем, потребуются совершенно новые технологии, новые алгоритмы, и адаптация существующего ПО на подобные системы будет очень непростой задачей. Но, в любом случае, до выхода подобных чипов на рынок пройдет не менее трех лет.

SN. *Несколько слов – о наиболее "правильных", с Вашей точки зрения, файловых системах и системах хранения, используемых в составе HPC-кластеров.*

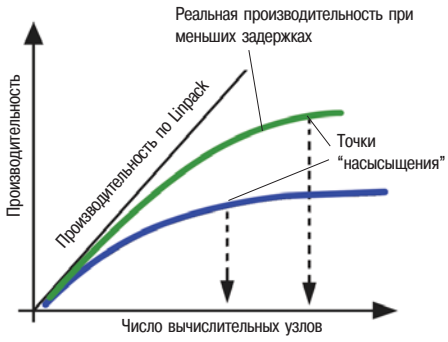
A.C. Для кластеров с производительностью более 1 TFlops, особенно если планируется расширение, наиболее эффективным является использование сетевой системы хранения с параллельной файловой системой. Для кластеров среднего и большого размера мы предлагаем законченное программно-аппаратное решение с интегрированной файловой системой T-Platforms ReadyStorage ActiveScale Cluster, разработанное компанией Panasas Inc. Файловая система ActiveScale File System позволяет увеличить объем хранилища путем простого добавления модулей хранения, обеспечивая при этом линейный рост производительности. Кроме того, с ростом объема хранилища сохраняется единое глобальное пространство имен, что позволяет существенно упростить администрирование. Можно также использовать другие аппаратные решения совместно со свободно распространяемыми файловыми системами, такими как Lustre. Для небольших кластерных систем можно использовать также сетевые системы хранения с интерфейсом NAS или SAN, работающие по протоколу NFS. Например, в небольших инсталляциях мы активно используем систему хранения T-Paltforms ReadyStorage SAN, что позволяет, в частности, организовать единую инфраструктуру для обмена сообщениями и доступа к данным на базе интерконнекта InfiniBand и обеспечить таким образом дополнительную экономичность, надежность и легкость расширения системы.

SN. *Какие новые технологии и тенденции для создания кластерных систем Вы можете отметить?*

A.C. Одна из технологий, которая в ближайшее время появится на рынке, это технология интеграции оптических соединений прямо на кристалл микросхемы. Она сильно удешевит стоимость интерконнекта и позволит существенно повысить пропускную способность. Так, если сейчас это максимум 20 Гбит/с на порт, то в течение уже этого года следует ожидать появления интерконнекта 40–100 Гбит/с на порт.

Другая тенденция в области интерконнекта это все большее внедрение на рынок массовых открытых технологий типа Infiniband и Ethernet 10 Gb, а такие интерфейсы как QsNET, Quadrics, Myrinet и др. снижают свое присутствие на рынке и потихоньку исчезают.

В начале прошлого года мы приобрели у компании GDA Technologies лицензию на реализацию интерфейса высокоскоростной шины HyperTransport(tm) для микросхем программируемой логики (FPGA) и приступили к разработке собственных компонентов интерконнекта с пониженным временем задержки для высокопроизводительных вычислительных комплексов. И сейчас мы достигли задержек в интерконнекте порядка 1,5 мкс. Для сравнения: наилучшие показатели сегодня это – 1,3 мкс, а для Infiniband – порядка 3-4 мкс. Необходимо понимать, что производительность кластеров, измеряемая на стандартных пакетах типа Linpack, растет практически линейно до нескольких сот и тысяч узлов. Реальная производительность в зависимости от класса задач может достигать максимума при значительно меньшем числе узлов. Это связано с тем, что, помимо счета, узлы должны обмениваться сообщениями между собой. И, например, при прочностных расчетах этот максимум может быть достигнут уже на 10 узлах, при расчетах газодинамики – на 60–120 узлах. Поэтому, чем меньше задержка, тем больше производительность и, соответственно, насыщение возникает при большем числе узлов, если огрубленно, то это порядка 10-25% прироста реальной производительности.



SN. *Насколько эффективно применение специализированных чипов и чипов с программируемой логикой?*

A.C. Построение систем на базе подобных чипов для отдельных классов задач позволяет получить производительность системы в десятки раз большую, чем на стандартных вычислительных узлах. Проблема здесь в том, что программирование подобных систем несоизмеримо сложнее, чем для стандартных кластеров. Например, аппаратные ускорители компании PSP позволяют гораздо эффективнее решать задачи линейной алгебры. В эпоху бурного "расцвета" микропроцессоров они были забыты, но сейчас начинают возвращаться, вследствие того, что дальнейший рост частоты становится все более трудным, а увеличение числа ядер не ведет к линейному росту производительности.

Стоимость разработки подобных чипов может составлять сотни тысяч долларов. Что касается решений, то стоимость приобретения платы аппаратного ускорителя составляет несколько тысяч долларов, но при этом в сравнении со стандартным модулем (стоимостью порядка $1 тыс.) он обеспечивает повышение производительности в 10-100 раз, а на некоторых задачах – до 1000 раз. Т.е. это реальная альтернатива построения высокопроизводительных систем в сравнении с аналогичными на стандартных компонентах без слишком больших затрат на энергопотребление. Проблема в том, что, как правило, для рынка разрабатывается комплекс взаимоувязанных программ, и каждая такая система требует перепрограммирования. Простая перекомпиляция здесь невозможна, т.е. требуется отдельный проект.

Заключение

Ряд последних анонсов по внедрению суперкомпьютерных систем в учебные заведения и в отраслях промышленности в России – свидетельство развития суперкомпьютерной отрасли. Но, все же, как уже было сказано, Россия находится только в начале пути. Возрастающее понимание того, что без использования высокопроизводительных вычислительных систем практически становится невозможным решение не только стратегических задач, но и достижение необходимого уровня эффективности производства, вселяет определенный оптимизм. Однако во многих ведущих отраслях и компаниях (где государство является единственным собственником или мажоритарным акционером) – нефтегазовая, автомобиле-, авиастроение, космос, микроэлектроника и др. – внедрение высоких технологий происходит медленнее, чем хотелось бы. Интерес со стороны Государственной Думы к этому направлению, безусловно, даст ему мощный импульс. И уже до конца года, SN надеется дать информацию об анонсах нескольких крупных проектов.