Кластерные HPC-системы: особенности, перспективы

Данная публикация является первой из серии, посвященной кластерным системам для параллельных вычислений – одному из наиболее перспективных направлений развития IT- отрасли на ближайшие годы. По мнению ряда экспертов, кластерный принцип развития IT-инфраструктуры, может стать основным при ее масштабировании уже в ближайшие несколько лет.

При подготовке публикации были использованы материалы, предоставленные С. Жуматием (НИВЦ МГУ).

Введение

История параллельных вычислений уходит в 70-е годы, когда для отдельных областей применения компьютеров (геология, распознавание образов, ядерные исследования, научное моделирование, прогноз погоды и др.) была необходима производительность на операциях с плавающей точкой, на порядки превышающая производительность универсальных ЭВМ. Одной из родоначальниц суперкомпьютеров является фирма Cray, выпустившая в эти годы модель – Cray-1. Началом интенсивного развития рынка суперкомпьютеров можно считать середину 80-х, когда за появлением 4-конвейерной модели Cyber-205 компании CDC и выпуском 48-узловой Cray X-MP, были представлены разработки HAP-1 и VP100/200 компаний Hitachi и Fujitsu соответственно.

Как видно, история HPC (High Performance Computing) имеет более чем 20-летний опыт. В чем специфика настоящего момента и чем обусловлен резко возросший интерес к HPC-системам (речь идет прежде всего о кластерных HPC-системах на стандартных серверах)?

<u>Первое</u> – в ценовой доступности колоссальной вычислительной мощности гораздо более широкому слою потребителей. Если в недавнем прошлом стоимость суперкомпьютера определялась миллионами долларов (US), то в настоящее время она измеряется от нескольких десятков тысяч до нескольких сот тысяч долларов, т.е. упала более чем на 2 порядка! Это вызвано технологическими революциями на рынке компонентов HPC-систем и практически полным переходом на стандартные компоненты (производимые в массовом количестве промышленностью, а, соответственно, и дешево). Прежде всего это относится к микропроцессорам, когда 64-разрядные чипы стоимостью $5-6 уже стали стандартом при производстве коммерческих серверов. Также это значительный прогресс на рынке интерконнекторов, когда с началом поставок (2004 г.) Infiniband-продуктов в несколько раз упала их стоимость (например, в сравнении с 10GE) как на адаптеры, так и на порты коммутаторов. Уже в конце с.г. ожидается появление оптоволоконного (сейчас производится только медный кабель) Infiniband-интерфейса, что позволит значительно расширить оптимальную его длину (по стоимости) с настоящих 15 м. Заметим, что в текущий момент в стандартной 19" стойке размещается до 80 одноядерных узлов, что более чем достаточно для построения многосотузловых систем существующего уровня технологии Infiniband.

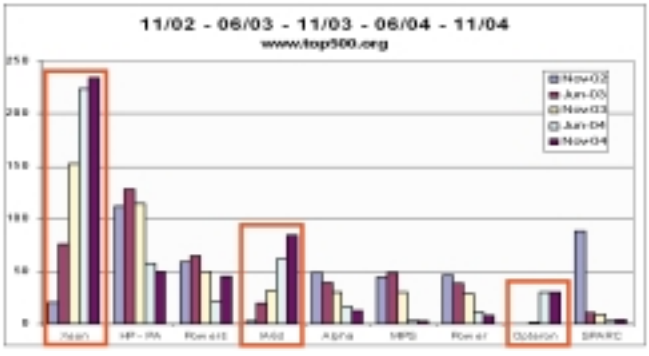

Рис. 1. Пятилетний анализ проектов списка top500 показывает неуклонный рост реализаций HPC систем на стандартных архитектурах.

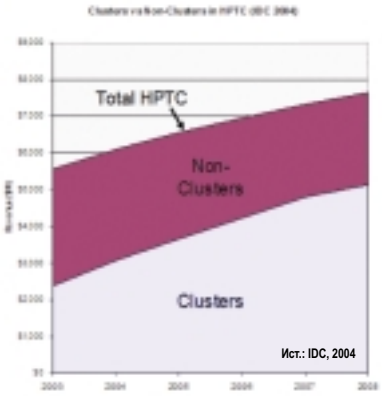

Рис. 2. По оценкам IDC, рынок HPC-кластерных систем к 2008 г. по сравнению с 2004 г. почти удвоится и в 2 раза будет превышать рынок некластерных HPC-систем.

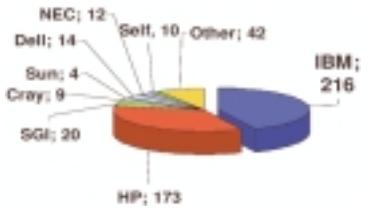

Рис. 3. Количество инсталляций суперкомпьютеров каждого из вендоров по списка top500.

Пятилетний анализ списка "top500" самых мощных суперкомпьютеров (www.top500.org) мира показывает (рис. 1) неуклонный рост числа проектов, построенных на стандартных микропроцессорах. Эта тенденция подтверждается прогнозами IDC, в соответствии с которыми рынок HPC-кластерных систем к 2008 г. по сравнению с 2004 г. почти удвоится и в 2 раза будет превышать рынок некластерных HPC-систем (рис. 2).

Второе (вследствие снижения уровня ценовой доступности суперсистем) – в значительном расширении сферы применения HPC-систем: от решения в основном научных и стратегических задач до большого круга коммерческих приложений. Так по заявлениям представителей IBM, кластерные HPC-системы могут с успехом использоваться как front-end системы в банковской IT-инфраструктуре. А вследствие их значительно мéньшей стоимости могут уже в ближайшей перспективе существенно "подвинуть" рынок RISC-серверов в соответствующих применениях.

Обзор рынка

Согласно 24-й редакции (ноябрь 2004 г.) списка "top500" самых мощных суперкомпьютеров мира *(www.top500.org)* лидирующее положение занимает IBM (рис.3). Если рассматривать первую сотню проектов списка, то соотношение между вендорами следующее: IBM (50%), HP (12%), Dell (5%), Cray (4%), Fujitsu (4%), Hitachi (3%) и др. (табл. 1). 98 место в этом списке занимает система СКИФ К-1000 *(SN № 3/21, 2004)* – совместная разработка России и Беларуси в рамках программы "СКИФ".

Данные, представленные IDC в конце 2004 г., в целом по разделению мирового рынка кластерных систем (общий объем – $3,38 млрд) представлены на рис. 4.

Статистические данные по региону СНГ по официально установленным и представленным суперкомпьютерам собираются в рамках совместного проекта НИВЦ МГУ и Межведомственного Суперкомпьютерного Центра РАН – "top50" *(www.parallel.ru)*. В соответствии с последними данными (конец марта 2005 г.) первое место занимает

Рис. 4. Разделение мирового рынка кластерных систем, по данным IDC, 2004.

Табл. 1.

Rank	Site Country / Year	Computer / Processors Manufacturer	Computer Family Model	Inst. type Installation Area	R_{max} R_{peak}	N_{max} N_{half}
1	IBM/DOE United States/2004	BlueGene/L beta-System BlueGene/L DD2 beta-System (0.7 GHz PowerPC 440) / 32768 IBM	IBM BlueGene/L BlueGene/L	Research	70720 91750	933697
2	NASA/Ames Research Center/NAS United States/2004	Columbia SGI Altix 1.5 GHz, Voltaire Infiniband / 10160 SGI	SGI Altix SGI Altix 1.5 GHz, Infiniband	Research	51870 60960	1.29024e+06
3	The Earth Simulator Center Japan/2002	Earth-Simulator / 5120 NEC	NEC Vector SX6	Research	35860 40960	1.0752e+06 266240
4	Barcelona Supercomputer Center Spain/2004	MareNostrum eServer BladeCenter JS20 (PowerPC 970 2.2 GHz), Myrinet / 3564 IBM	IBM Cluster JS20 Cluster, Myrinet	Academic	20530 31363	812592
5	Lawrence Livermore National Laboratory United States/2004	Thunder Intel Itanium2 Tiger4 1.4GHz - Quadrics / 4096 California Digital Corporation	NOW - Intel Itanium Itanium2 Tiger4 Cluster - Quadrics	Research	19940 22938	975000 116000
6	Los Alamos National Laboratory United States/2002	ASCI Q AlphaServer SC45, 1.25 GHz / 8192 HP	HP AlphaServer SC Alpha-Server-Cluster	Research	13880 20480	633000 225000
96	Energy Company United States/2004	Integrity rx4640-4s (16 Itanium2 1.3 GHz, Gig) / 900 HP	HP Cluster Integrity rx4640 Itanium2 Cluster, GigEthernet	Industry Geophysics	2059 4160	
97	Oracle Corporation United States/2004	DL380G3, Pentium4 Xeon 2.8 GHz, Myrinet / 702 HP	HP Cluster HP DL380 Cluster	Industry Database	2044 3921.2	
98	United Institute of Informatics Problems Belarus/2004	SKIF K-2000 Opteron 2.2 GHz, Infiniband / 576 Self-made	NOW - AMD NOW Cluster - AMD - Infiniband	Academic	2032 2534.4	274000 28000
99	Geoscience (C) United Kingdom/2004	xSeries Xeon 2.8 GHz, Gig-Ethernet / 2296 IBM	IBM Cluster xSeries Cluster Xeon - GigE	Industry Geophysics	2026 12633	
100	Credit Suisse Switzerland/2004	BladeCenter HS20 Xeon 2.8 GHz, Gig-Ethernet / 1500 IBM	IBM Cluster xSeries Cluster Xeon - GigE	Industry	2026 9408	

Табл. 2.

N	Место	Тип	Кол-во CPU	Архитектура (тип процессора / сеть)	Область применения	Производительность (расч. / тест.)	Разработчик
1	Москва МСЦ РАН 2004 г.	кластер	552	узлов: 276 (2xPowerPC 970 2.2 GHz 4 GB RAM) сеть: Myrinet/2xGigabit Ethernet	Наука и образование	3052	4657.6	ФГУП "Квант", ИПМ РАН, МСЦ
2	Минск ОИПИ НАНБ 2004 г.	кластер	576	узлов: 288 (2xOpteron 248 2.2 GHz 4 GB RAM) сеть: Infiniband/Gigabit Ethernet/ICSMB-Servfed	Исследования	2832	2534.4	СКИФ
3	Москва HEWLETT 2004 г.	кластер	360	узлов: 88 (2xOpteron 248 2.2 GHz 4 GB RAM) сеть: Infiniband/Gigabit Ethernet/Fast Ethernet	Наука и образование	532	704	Hewlett-Packard
4	Москва МСЦ РАН 2005 г.	кластер	512	узлов: 256 (2xAlpha 21264A 766 MHz 1 GB RAM) сеть: Myrinet 2000/Fast Ethernet	Наука и образование	1013.7	794.4	ФГУП "Квант", ИПМ РАН, МСЦ
5	Ереван ИПИИ НАН РА 2004 г.	кластер	128	узлов: 64 (2xXeon 3 GHz 1 GB RAM) сеть: Myrinet/Gigabit Ethernet	Наука и образование	1024	763.36	ИПИИ НАН РА, ИСП РАН, C.I.Technology
6	Минск ОИПИ НАНБ 2003 г.	кластер	128	узлов: 64 (2xXeon 2.8 GHz 2 GB RAM) сеть: SCI/Gigabit Ethernet/ICSMB-Servfed	Наука и образование	1075.2	736.8	СКИФ
7	Москва ChipBase 2003 г.	SMP	256	HP SuperDome PA-8550 750 MHz	Финансы	1024	768	Hewlett-Packard
8	Киев ИК НАН Украины 2004 г.	кластер	64	узлов: 32 (2xItanium 2 1.4 GHz 2 GB RAM) сеть: SCI/Gigabit Ethernet/Fast Ethernet	Исследования	179.2	358.4	ИК НАН Украины / ГНЦП "ЮСТАР" г. Киев
9	Москва отдел высокопроизводительных вычислительных систем ОАО "НИЦЭВТ" 2003 г.	кластер	32	узлов: 16 (2xXeon 2.8 GHz 2 GB RAM) сеть: SCI/Gigabit Ethernet	Исследования	358	403.2	СКИФ, ОАО "НИЦЭВТ"
10	Москва Paradigm 2005 г.	кластер	40	узлов: 24 (2xXeon EM64T 3.4 GHz 4 GB RAM) сеть: Gigabit Ethernet/Gigabit Ethernet	Промышленность	219.5	312.0	T-Платформы

совместная разработка ФГУП "Квант", ИПМ РАН и МСЦ – кластер с 276 узлами (2xPowerPC 970 2.2 GHz 4 GB RAM) на базе платформы IBM BladeCenter JS20 (табл. 2).

Обзор архитектур

Практически все мощные вычислительные установки на сегодняшний день можно разделить на два класса: с общей памятью и массивно-параллельные. Системы с общей памятью, в свою очередь, можно разделить на SMP- (Symmetric MultiProcessing – симметричная многопроцессорная архитектура) и NUMA-системы (Non-Uniform Memory Access – архитектура с неравномерным доступом к памяти). Массивно-параллельные установки могут различаться внутренним устройством и называться либо кластером, либо нет, но последнее зависит только от того, собрана ли установка из серийных компонентов.

Архитектура SMP

Симметричная многопроцессорная обработка — расширение базовой однопроцессорной архитектуры (рис.4). SMP это единый компьютер, в котором множество центральных процессоров разделяют доступ к системным ресурсам типа памяти или внешних устройств по общедоступной системной шине (как правило, серверы стандартной архитектуры) или через системный коммутатор (например, серверы Sun Microsystems).

В SMP для каждого из процессоров любой участок оперативной памяти доступен на равных правах с остальными. Программирование на подобных системах наиболее удобно, т.к. любое изменение, сделанное одним процессором, сразу доступно всем остальным.

Однако в реализации такое решение весьма сложно, т.к. если все процессоры могут обращаться к одному участку памяти одновременно, то необходимо обеспечить и высо-

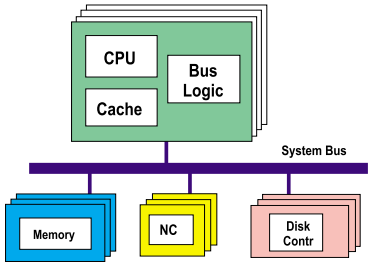

Рис. 4. Архитектура SMP

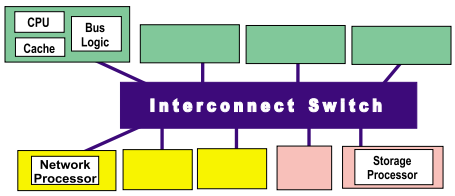

Рис. 5. Архитектура NUMA

кую скорость работы памяти, и непротиворечивость данных в кэшах процессоров. Если "посадить" все процессоры на одну шину доступа к памяти, то ее производительности может просто не хватить на всех. В случае 2-х процессоров это не сильно заметно, но уже на 4-процессорных системах это уже создает проблемы.

Архитектура NUMA

NUMA-машины основываются на парадигме разделяемой памяти SMP-компьютеров, но отличаются тем, что физически не имеют централизованной памяти и SMP-шина заменяется общим коммутатором (рис. 5). Существует несколько вариантов NUMA-серверов в зависимости от уровня связности (когерентности), поддерживаемой среди блоков памяти.

В NUMA-варианте у каждого процессора свой банк памяти, но он также имеет возможность обращаться и к банкам памяти других процессоров. В случае обращения к "чужим" банкам скорость выполнения запроса ниже. Это наиболее популярная в данный момент архитектура для вычислительных установок с общей памятью.

Даже в архитектуре NUMA приходится решать проблему с адекватностью данных в кэшах процессоров. Существует даже специальное обозначение для NUMA-архитектур, поддерживающих корректность данных в кэшах: cc-NUMA (cache coherent NUMA). Так как поддерживать корректность данных в кэшах непросто, особенно без существенного снижения скорости работы, то такие системы весьма дороги. Стоимость их растет непропорционально числу процессоров и максимальное число процессоров в них сравнительно невелико.

Кластерная архитектура

Кластеры, или массивно-параллельные компьютеры, основаны на понятии автономного разделения, т.е. коллекции узлов, которые обычно работают независимо друг от друга

Рис. 6. Классификация типов кластеров.

и каждый из которых ответствен за долю рабочей нагрузки, но имеет возможность доступа к памяти одного или более других узлов. Или, в другом определении, кластер это многосерверная система состоящая из взаимосвязанных компьютеров одной или несколькими сетями, а также системы хранения, которые унифицируются посредством общего управления и сетевого программного обеспечения для выполнения определенных задач. При этом каждый узел имеет свою оперативную память и напрямую обращаться в память других узлов (в общем случае) не может. На каждом узле работает свой экземпляр ОС. В общем случае можно выделить 5 типов кластеров (рис. 6), из которых к HPC-кластерам относятся только два.

Степень, в которой многосерверные системы показывают следующие характеристики, определяет может ли кластер называться высокопроизводительным:

- выделенная частная VLAN;
- на всех узлах работает одно и то же приложение/пакет приложений;
- единая точка управления ПО /распространения приложений и управления;
- единая точка управления аппаратной частью;
- межузловая связь;
- взаимозависимость узлов.

Типовая организация высокопроизводительного кластера представлена на рис. 7.

Если встал выбор, поставить кластер или большую SMP-систему, то лучше протестировать свою задачу на одном из доступных кластеров – возможно задаче настолько часто надо обмениваться информацией внутри процессов, что даже самая лучшая сеть не

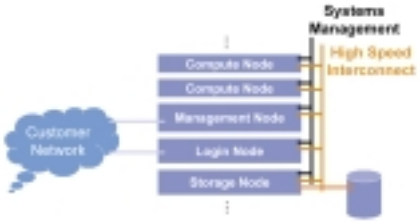

Рис. 7. Типовая организация HPC-кластера.

поможет. В этом случае другого выхода не будет: надо считать на SMP. Если же прирост производительности есть, то стоит подумать о более дешевом кластере, но не промахнуться с выбором сети. Думая о размере SMP или кластерной системы, не забывайте о законе Амдала: если доля последовательных (нераспараллеленных) операций в программе равна f, то ускорение задачи на p процессорах не может быть болше, чем 1/(f+(1-f)/p). Как следствие, программа не может быть ускорена более чем в 1/f раз ни на каком числе процессоров. Если в программе 1/10 часть операций исполняется не параллельно, то ускорить ее более чем в 10 раз невозможно.

Особенности построения HPC-кластеров

Архитектура HPC-кластеров во многом определяется заранее самим алгоритмом предполагаемых для решения на нем задач. От того, как строится выполнение задачи и насколько архитектура кластера соответствует логике решения, во многом зависит и эффективность использования самого кластера в дальнейшем.

Рис. 8. Классификация задач, решаемых на HPC-системах, с точки зрения требований к архитектуре HPC-кластера.

В качестве параметров задачи, существенным образом влияющих на архитектуру кластера, можно назвать:

- размер "порции" программы при распараллеливании;
- количество параллельных процессов;
- используемый вид синхронизации процессов;
- способ управления доступом к разделяемым данным.

Если рассматривать архитектурные компоненты, определяющие скорость решения задачи, то среди множества можно выделить три основных (рис. 8):

- мощность вычислителя (или производительность CPU кластера);
- пропускная способность и задержка межузлового соединения (тип и характеристики интерконнектора);
- пропускная способность и задержка памяти (организация доступа и хранения обрабатываемых данных).

Вследствие достаточного множества платформ и несколько вариантов коммуникационных сетей, а также вариантов организации доступа к данным, HPC-кластеры очень разнообразны.

Из аппаратных платформ при построении узлов HPC-кластера наиболее популярны Intel Xeon, Intel Itanium^2, AMD Opteron. Сейчас появился Intel Xeon EMT64, на базе которого тоже, по всей видимости, будет построено немало кластеров. Вариантов коммуникационных сетей на данный момент несколько: Fast Ethernet, 1Gbit Ethernet, InfiniBand, SCI, Myrinet, QsNET.

Что выбрать для построения кластера – задача нелегкая. Ориентироваться в ее решении следует на два ключевых момента – целевую задача (какие задачи будут решаться на кластере) и стоимость. Первый критерий самый важный, т.к. если попытаться слишком много сэкономить, то вместо ускорения счёта можно получить даже замедление.

Проведем сравнение процессоров:

1. Intel Xeon – наиболее популярное на данный момент решение. Распространенный процессор, большое количество готового программного обеспечения. Однако произ-

водительность наиболее низкая из всех перечисленных выше платформ.

2. Intel Itanium[2] – наиболее производительный процессор. Минусы – высокая стоимость, а также необходимость перекомпиляции любых программ. Если используются скомпилированные программы или библиотеки и не имеется исходных текстов, то применение этого процессора может быть проблематичным.

3. AMD Opteron – процессор, сочетающий скорость, относительно невысокую цену и возможность работы в 2-х режимах: исполнения кода процессоров семейства x86; исполнения своей 64-битной системы команд. Даже в режиме "эмуляции" он нередко показывает производительность значительно более высокую, чем аналогичный по частоте Xeon.

Не стоит забывать о таком важном параметре, как размер кэша. Чем больше кэш, тем больше информации может быть доставлено из памяти с бóльшей скоростью при повторном обращении к ней. Это не будет играть существенной роли, если программа обрабатывает "за раз" большие порции данных, которые заведомо в кэш не помещаются. Но если обработку сделать небольшими порциями, подобрав их под размер кэша, то программа будет работать значительно быстрее.

Краткое резюме: 1) если программе требуется производить большое количество вычислений с плавающей точкой, то лучший выбор – Itanium [2] (если, конечно, на него хватит средств); 2) если хочется сэкономить, то самый просто выбор – Xeon; 3) золотая середина – Opteron. Желательно проверить, как поведет себя программа на каждом из процессоров. Возможно Xeon'а окажется вполне достаточно, и на Opteron'е или Itanium[2] значительного ускорения не получится. Это зависит только от конкретной задачи и реализованного в ней алгоритма. Если негде протестировать задачу, можно обратиться в Internet, где возможно есть данные о задачах, использующих аналогичные алгоритмы.

Выбор коммуникационной сети важен не менее, чем выбор процессора. На что нужно обратить внимание? На скорость передачи, на величину латентности передачи, на цену и на возможность расширения. Чем выше скорость передачи, тем быстрее процессы параллельной задачи смогут обменяться данными друг с другом. Куда менее очевидно значение латентности (время на подготовку к пересылке и приему сообщения). Однако пренебрегать им не следует: если обмены происходят часто, то большая латентность выльется в серьезные задержки. Если обменов мало, то латентность будет играть небольшую роль.

Возможность расширения, тоже немаловажный фактор. Если возникает необходимость добавить вычислительные узлы, то коммутатор должен поддерживать возможность добавления портов (все виды скоростных сетей, кроме SCI на данный момент построены на базе коммутаторов). Наращивание сети путем стыкования коммутаторов обычно очень неэффективно.

Кратко рассмотрим плюсы и минусы перечисленных сетей. Ниже будет упомянута библиотека MPI, которая является стандартом de-facto при программировании на кластерах.

1. Fast Ethernet. Наиболее старая и распространенная технология. Скорость передачи невысока, при этом очень высокая латентность (150 мкс). В рамках MPI скорость – около 6–7 Мбайт/с.

2. 1GB Ethernet. Развитие Fast Ethernet – скорость передачи увеличена до 100 Мбайт/с. Латентность по-прежнему высокая (150 мкс). В рамках MPI скорость около 45 Мбайт/с. При покупке оборудования важно учесть, что не все коммутаторы способны поддерживать скорость 1GB/s на всех портах одновременно.

3. Myrinet 2000. Позволяет передавать данные со скоростью 2 Gbit/sec. Латентность – порядка 5 мкс. На приложениях MPI латентность составляет около 10 мкс, скорость передачи данных – до 400 Мбайт/сек (в дуплексном режиме).

4. SCI. Скорость передачи данных – 400 Мбайт/с. В рамках MPI – более 300 Мбайт/с. Аппаратная латентность – 1,2 мкс, для MPI-приложений – 4 мкс. В отличие от других технологий, здесь не используется коммутатор – узлы соединяются напрямую в кольцо, либо 2- или 3-мерный тор.

5. QsNET. Наиболее дорогая на сегодняшний день технология. Пропускная способность – 1064 Мбайт/с (на реальных приложениях – до 900 Мбайт/с). Латентность в рамках MPI – 3 мкс.

6. InfiniBand. Самая новая, а поэтому активно развивающаяся технология. На сегодняшний день существуют решения, дающие скорость до 1000 Мбайт/с с латентностью 6–7 мкс. На реальных приложениях получены скорости до 900 Мбайт/с и латентность – 7–8 мкс.

Выбирая сеть, важно хорошо представлять, как будут организованы обмены сообщениями в запускаемых задачах – если обменов немного (например, данные задачи разбиваются на большие порции и долго обрабатываются и только потом процессы обмениваются результатами), то и требования к сети невысокие. Если объемы пересылаемых данных велики, но количество их небольшое, то важна будет только скорость, но не латентность. Если же обменов много, то в любом случае латентность будет играть большую роль: чем она меньше, тем лучше.

Например, большинство задач квантовой химии на Ethernet запускать просто не имеет смысла: почти все ускорение съедается за счет высокой латентности при пересылках по сети. В то же время, задачи кодирования видео или оцифровки 3-мерных сцен требуют небольшого количества обменов, поэтому хорошо работают даже на Ethernet.

Не стоит недооценивать и значение периферии кластера: дисковой подсистемы и системы охлаждения. Последняя особенно важна для мощных процессоров (например, для Itanium [2]), т.к. пренебрегая ею – можно просто спалить дорогостоящее оборудование. В отличие он "настольных" систем, в случае кластеров внутреннего охлаждения вентиляторами недостаточно: необходимо заранее продумать и просчитать систему кондиционирования.

Дисковая подсистема может стать причиной серьезного замедления работы программ, если они используют сетевой диск. А использование сетевого диска – наиболее удобная схема для кластеров (т.к. все узлы "видят" одну и ту же информацию). Поэтому крайне желательно организовать отдельное сетевое хранилище и отдельную сеть для доступа к нему (чтобы не смешивать вычислительный трафик со служебным). Для организации служебной сети, как правило, не требуется высокоскоростной сети, и использование Ethernet бывает достаточно. Однако требования к сетевому хранилищу при этом весьма серьезны, ведь клиентов, одновременно работающих с ним, будет много – все узлы кластера. Следует помнить, что производительность наиболее распространенной сетевой файловой системы – NFS – довольно невелика, а нагрузка на сервер и его диски при этом значительна.

Для оценки требований к сетевому хранилищу снова нужно обратиться к задачам, которые будут считаться. Если они активно пишут на общий диск, хранилище должно быть очень производительным. Если есть возможность реорганизовать программы так, чтобы временные файлы писались на локальные диски, то это нужно сделать обязательно, т.к. запись и чтение с сетевого диска все равно медленнее, чем с локального.

Из промышленно предлагаемых параллельных файловых систем для обеспечения разделяемого высокоскоростного доступа к данным при работе с многоузловым вычислительным кластером можно выделить две: систему IBM General Parallel File System (GPFS) для Linux (поставляемую в составе кластера e1350 и ряда RISC-систем) и систему на основе Lustre™ технологии – свободно распространяемую и поставляемую, например, в составе продукта HP StorageWorks Scalable File Share.

GPFS – это высокоэффективная, масштабируемая файловая система с совместным доступом к дисковой памяти, разработанная для обеспечения высокоскоростного доступа к данным всех узлов Linux-кластера. GPFS обеспечивает свободный совместный доступ к общим файлам с использованием стандартных интерфейсов файловых систем UNIX® как для приложений, работающих одновременно на нескольких узлах кластера, так и для приложений, работающих на отдельно взятом узле. Кроме того, возможна настройка GPFS с автоматическим переходом на резервные ресурсы при отказе как дисков, так и серверов. Главное, что дает применение GPFS для Linux – это высочайшие производительность, масштабируемость и готовность. GPFS способна масштабироваться по мере роста Linux-кластера и предоставляет возможность экспорта в NFS за пределами кластера.

Основные преимущества GPFS:

- увеличение общей производительности и общей пропускной способности; производительность SAN-систем по значительно более низкой цене;
- проверенные технологии для AIX (RS/6000® SP™ heritage);

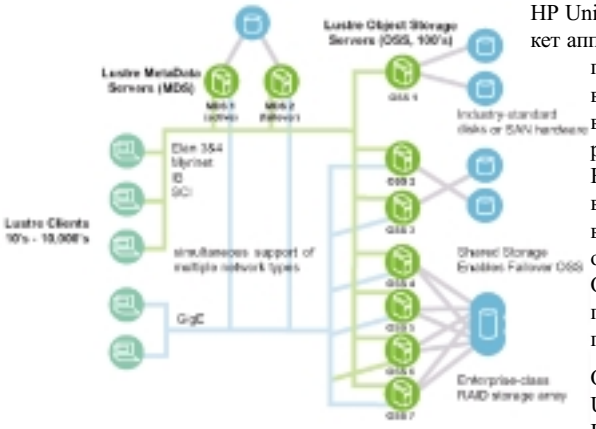

Рис. 9. Организация параллельной системы хранения на основе технологии Lustre для HPC-кластера.

- функция разделяемого виртуального диска для IBM Linux кластеров;
- технология IBM Reliable Scalable Cluster Technology (RSCT) обеспечивает автоматическое восстановление доступа к системам хранения данных; функции журналирования (быстрое восстановление данных);
- независимость от приложений –доступ к единому образу файла;
- файловая система до 75 Тбайт.

В основе технологии Lustre лежит идея разбиения файла на несколько частей или потоков, так же, как это делается в обычных дисковых RAID. В настоящее время поддерживается уровень RAID-0, планируется поддержка RAID-1 и RAID-5. Управление файловой системой осуществляется на уровне Object Storage Target (OST, размер ограничен размером 2 Тбайт на Linux 2.4 или 4 Тбайт на Linux 2.6). Однако сами OST могут объединяться в одну агрегированную файловую систему, соответственно, с максимальной емкостью – 800 Тбайт. Среди других показателей Lustre: масштабируемая пропускная способность от 100 Мбайт/с до 12 Гбайт/с; масштабируемое число узлов (клиентов) – от 10 до 10 000; масштабируемая надежность.

Технологиия Lustre обеспечивает построение очень гибких параллельных систем хранения (рис. 9) с использованием самых разных по емкости и производительности единичных Lustre Object Storage Servers (OSSs) – от low-end до high-end систем хранения.

Обзор предложений на рынке

В настоящее время, по имеющимся данным, в России промышленно изготавливаемые кластеры предлагаются тремя вендорами: HP, IBM и Sun Microsystems.

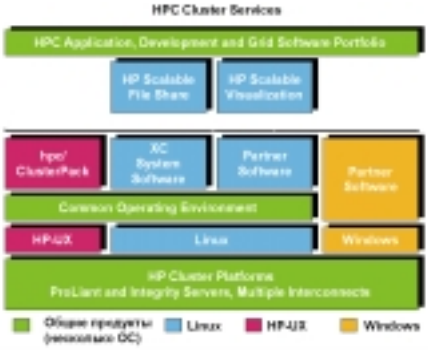

Рис. 10. HP Unified Cluster Portfolio – модульный пакет аппаратных средств, ПО и услуг для построения HPC-кластеров.

HP Unified Cluster Portfolio – модульный пакет аппаратных средств, программного обеспечения и услуг для масштабируемых вычислений, управления данными и визуализации (рис. 10). Все кластеры строятся на базе 3 платформ: HP Cluster Platform 3000/4000/6000, в составе которых могут использоваться 4 типа интерконнекта и 3 операционные системы (табл. 3). Общее число узлов для каждой платформы – 512 (общее число процессоров на кластер – до 1024).

Одна из основных особенностей HP Unified Cluster Portfolio – доступность HP-MPI библиотек и драйверов, которые обеспечивают поддержку продуктов, разрабатываемых независимыми поставщиками ПО (ISV). Обеспечивая общий дизайн (на основе (HP-MPI с Red Hat 3.0), отдельный кластер способен поддержать смесь различных ISV-приложений.

HP Unified Cluster Portfolio обеспечивает выбор управляющего ПО как для Linux, так и HP-UX операционного окружения, включая HP XC System Software, HP-UX hpc/ clusterpack, а также разработки третьих фирм, например: Scali Manage и Scali MPI Connect. Также пользователи имеют возможность инсталляции открытого ПО, такого как OSCAR и ROCKs.

Табл. 3.

	Compute node and processor	Interconnects	Operating Systems
HP Cluster Platform 3000 (до 512 узлов)	HP ProLiant DL360 or DL380 servers with Intel EM64T Xeon™ processors	Myrinet, InfiniBand, or Ethernet	Linux or Microsoft Windows
HP Cluster Platform 4000 (до 512 узлов)	HP ProLiant DL145 or DL 585 servers with AMD Opteron processors	Quadrics, InfiniBand, Myrinet, or Ethernet	Linux or Microsoft Windows
HP Cluster Platform 6000 (до 512 узлов)	HP Integrity rx1620, rx2600, rx2620, and rx4640 servers with Intel Itanium ² processors	Quadrics, InfiniBand, or Ethernet	Linux or HP-UX

Для ряда задач, требующих высокопроизводительного параллельного доступа к данным и высокопроизводительных ресурсов для визуализации данных в составе вычислительного кластера могут поставляться, соответственно, еще 2 специализированные кластерные системы: HP StorageWorks Scalable File Share (построенной на Lustre технологии) и HP Scalable Visualization Cluster (SVC, на основе технологии SEPIA).

HP SVC – система визуализации на Linux кластерах, которая распределяет данные изображения среди множества рабочих станций, параллельно обрабатывающих часть изображения.

Промышленные кластерные системы IBM (до 512 узлов) строятся на основе платформы e1350. Основным элементом e1350 является модуль IBM Blade Center, интегрирующий в своем конструктиве все основные многочисленные компоненты кластера: сетевую инфраструктуру клиентов, вычислительные узлы, сетевую инфраструктуру для подключения SAN (рис. 11). Один Blade Center может содержать до 14 серверов, в свою очередь, каждый сервер может иметь до 4 процессоров. Благодаря высокой плотности компоновки, в стандартном 42U 19" шкафу может размещаться до 6 Blade Center с общим числом 84 блэйд-сервера или общей производительно-

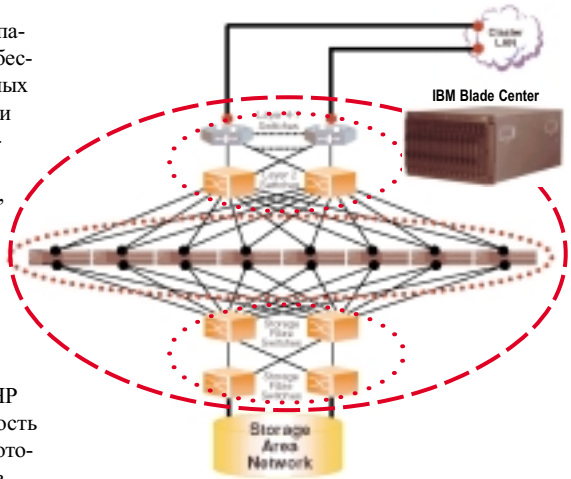

Рис. 11. IBM Blade Center интегрирует в одном конструктиве: сетевую инфраструктуру клиентов; вычислительные кластеры; сетевую инфраструктуру для доступа к SAN.

стью свыше 1 TFlops. При этом не предъявляется никаких особых требований по охлаждению системы, кроме выделенного помещения с кондиционированной средой.

Предусматриваются различные варианты исполнения блэйд-серверов как на базе процессоров Intel, так и AMD. Соединение узлов (блэйд-серверов) – на основе технологий Gigabit Ethernet, Myrinet или Infiniband. Одна из реализаций с использованием IBM BladeCenter – кластер МСЦ (см. табл. 2).

Компания Sun Microsystems предлагает промышленные системы на базе процессоров AMD Opteron и только для 2 базовых инсталляций: решения задач геофизики и Oracle-приложений.

Основной недостаток промышленных кластерных систем – более высокая стоимость в сравнении с аналогичными системами, собираемыми из отдельных компонент. Поэтому основная доля проектов разрабатывалась и разрабатывается по отдельным заказам для конкретных заказчиков.

Общее число российских компаний-интеграторов, предлагающих услуги по построению HPC-кластеров, весьма ограниченно и пока не превышает 10.

Примеры реализаций в России HPC кластерных систем

Необходимо отметить, что многие из реализованных проектов публично не представляются и недоступны для прессы. Об одном наиболее значительном проекте 2004 г., реализованном в рамках программы "СКИФ", мы уже писали *(SN № 3/21, 4004)*. Реализация следующего проекта была завершена в марте с.г.

24-узловая HPC кластерная система для "НК "Роснефть"

В апреле 2004 г. "РН-Телепорт" – дочерняя компания ОАО "НК "Роснефть" – завершила проект по созданию для "НК "Роснефть" информационно-расчетной системы, предназначенной для обработки информации о состоянии нефтяных запасов и сейсмических процессов в районах нефтедобычи. Новое уникальное решение на основе специализированного ПО Omega компании WesternGeco позволяет реализовать сейсмическое моделирование нефтяных месторождений в "НК

"Роснефть". Проект стал уникальным не только для российского рынка, но и для всей мировой IT-индустрии: впервые в мире для развертывания функционала Omega была выбрана HPC кластерная система на базе серверов HP стандартной архитектуры.

Традиционно для решения задач сейсмического моделирования, которые связаны с огромным объемом вычислений, использовались многопроцессорные RISC-серверы, работающие под управлением одного из коммерческих вариантов Unix. Стоимость подобных компьютеров измеряется многими сотнями тысяч долларов, а в некоторых случаях может превышать и миллион долларов. Кроме того, их покупатели оказываются заложниками фирменной архитектуры RISC/Unix одного производителя и из-за значительных инвестиций в приобретение RISC-системы им крайне невыгодно переводить свои приложения на серверы другого производителя.

В последние годы ситуация изменилась в связи с широкой популярностью вычислительных кластеров на базе операционной системы Linux, которые построены из небольших серверов стандартной архитектуры с процессорами Intel. Прежде всего, такие кластеры очень привлекательны по стоимостным показателям – например, заказчик может приобрести кластер минимальной конфигурации, состоящий из 4-8 двухпроцессорных серверов, и затем постепенно наращивать его производительность, подключая к нему дополнительные серверы. В то же время использование в этих кластерах стандартной архитектуры Intel и открытой операционной системы Linux гарантирует, что их покупатель не будет зависеть от одного поставщика и сможет воспользоваться всем богатейшим выбором ПО, разработанного сообществом Open Source.

Выбор программно-аппаратной платформы для информационно-расчетного центра "Роснефти" производился на основе тендера, в котором кроме HP участвовали компании Sun и Racksaver. При подготовке к тендеру HP провела тестирование оборудования на его совместимость с программным решением Omega – в российском офисе HP был собран тестовый кластер, функционирующий как испытательная модель решения, а для тщательного тестирования и установки Omega компания HP передала кластер из восьми HP ProLiant в лабораторию WesternGECO в г. Хьюстон, штат Техас. Результаты исследований, проводившихся в Москве и Хьюстоне, доказали полную совместимость аппаратной платформы HP-Intel, функционирующей под управлением ОС Linux, с Omega компании WesternGeco.

Развернутая в НК "Роснефть" информационно-расчетная система, представляет собой кластер Linux, который состоит из 24 серверов HP ProLiant DL360 G3. Каждый сервер оснащен двумя процессорами Intel® Xeon®, двухгигабайтной оперативной памятью и двумя жесткими дисками. Двухпроцессорный HP ProLiant модели DL380 G3 обеспечивает управление кластером, а еще один обслуживает систему хранения кластера. Основу программной платформы внедренного решения представляет сейсмическая вычислительная система (Seismic Processing System – SPS) Omega, предназначенная для широкоформатных и узконаправленных сейсмических вычислений, визуализирующая их результаты в реальном времени, что позволяет значительно улучшить производительность добычи ископаемых.

Инсталляцию решения HP-WesternGeco в "Роснефти" проводил "РН-Телепорт", отвечавший за системную интеграцию и координацию проекта, в партнерстве с компаниями ЛАНИТ, которая выполняла поставку готового кластерного решения HP, и Roy International , отвечавшая за внедрение программного обеспечения. Техническую поддержку заказчику оказывает Roy International, а компания HP через своего партнера ЛАНИТ обеспечивает трехлетнее гарантийное обслуживание.

В общей сложности с момента начала проекта и до поставки оборудования в "Роснефть" прошло около 2 месяцев, а инсталляция, пуск и наладка решения заняли 3 недели.

Внедрение решения HP-WesternGeco должно стать первым шагом на пути реализации масштабной инициативы компании "Роснефть" по созданию центра геолого-геофизической информации, где будут накапливаться и храниться все данные о месторождениях, на основе которых будет проводиться планирование разработки месторождений. Компания планирует повысить мощность кластера, на котором выполняется Omega, добавив в него несколько серверов, построить новый кластер для расчетов, связанных с гидродинамическим моделированием, а также систему хранения геологических данных.

В место заключения

Кластерные системы в IT – одно из самых перспективных быстро развивающихся направлений, касающееся всех уровней IT – средств вычисления, хранения данных и их визуализации – связанное, прежде всего, с ценовой доступностью решений и с тем, что в ряде случаев дальнейшее "последовательное" развитие IT-систем/подсистем становится невозможным без использования параллелизма в их архитектуре. Развитие кластеризации в IT, безусловно, явится мощным стимулятором развития промышленных технологий и научных исследований.