Гибридные кластеры как средства организации бюджетных суперкомпьютеров и вычислительных облаков

Н.П. Васильев, М.М. Ровнягин (НИЯУ МИФИ)

Рассмотрено современное направление вычислительной техники - вычислительные кластеры с гибридной CPU/GPU-архитектурой, которые широко применяются в различных областях науки и техники. Особое внимание уделено принципам построения недорогих, но достаточно мощных кластеров для промышленных расчетов, технологического моделирования и построения собственных вычислительных облаков.

Ключевые слова: вычислительные кластеры, гибридная архитектура, облачные вычисления, компьютерное проектирование и моделирование.

Введение

В настоящее время средства вычислительной техники используются во всех областях науки и промышленности. При помощи методов численного моделирования решаются сложные технологические проблемы, возникающие на этапе проектирования различной техники. Так в авиации применение компьютерного моделирования позволяет существенно удешевить и ускорить процесс выпуска новой авиатехники: многие испытания, которые раньше проводились только в аэродинамической трубе, теперь выполняются виртуально на моделях будущих самолетов. Несмотря на доступность компьютеров, вычисления такого уровня сложности требует совсем иных, нежели персональные ЭВМ, вычислительных систем.

В последние годы все чаще используются технологии облачных вычислений. С точки зрения пользователей — это получение доступа к вычислительным услугам как к обычному сервису: пользователь подготовил свою задачу, запустил ее "где-то там" в облаке провайдера, затем получил результаты расчетов. С точки зрения провайдеров, для предоставления такого рода услуг необходима гибкая вычислительная инфраструктура, которая позволяет быстро организовать наиболее подходящую вычислительную среду для пользовательской задачи: операционную систему, перечень необходимого прикладного ПО и т. п. В основе облачных технологий лежит виртуализация ресурсов и вычислительных процессов, расплатой же за гибкость являются высокие требования к вычислительным системам облачной инфраструктуры.

Для решения сложных вычислительных задач с середины XX века создавались специальные ЭВМ, превосходящие по производительности существующие компьютеры на порядок. Подобные вычислительные системы принято называть суперкомпьютерами. Целью настоящей статьи является обзор современных технологий организации высокопроизводительных вычислений; особое внимание уделено гибридным (или "гетерогенным", как пишут зарубежные авторы) вычислительным системам, в которых вычисления производятся как с помощью обычных, так и посредством графических процессоров. Построение многомашинных вычислительных комплексов (кластеров) с гибридной архитектурой позволит получить в свое распоряжение мощный вычислительный центр, который в дальнейшем можно использовать как при решении вычислительно сложных задач, так и в качестве облачной инфраструктуры. В то же время затраты на построение такой системы могут быть весьма скромны: такой кластер можно создать даже на базе уже имеющихся офисных компьютеров, объединенных в обычную локальную сеть.

Суперкомпьютеры и кластерные системы

К началу 90-х годов XX века сформировались следующие практически востребованные виды архитектур суперЭВМ:

1) Single Processor — суперкомпьютер, построенный на базе одного вычислителя;

2) MPP — система с массовым параллелизмом, которая проектируется на базе стандартных процессоров. Средства коммутации (interconnect) и размещения (placement), как правило, являются проприетарными разработками производителей суперЭВМ;

3) SIMD — суперкомпьютер, построенный на базе векторных процессоров;

4) SMP — система, включающая несколько процессоров, работающих с общей памятью.

Летом 1994 г. в научно-космическом центре NASA был введен в эксплуатацию многомашинный вычислительный комплекс Beowulf. Система была построена на базе 16 процессоров 486DX4/100МГц, использующих стандартные комплектующие, установленные в обыкновенные корпуса и соединенные при помощи Ethernet-сети с пропускной способностью 10Мбит/с. Вычислительный комплекс обладал высокой производительностью, но стоил в несколько раз дешевле MPP-аналогов, в которых преобладали проприетарные технологии. Идея «вычислительного кластера» получила свое дальнейшее развитие с появлением новых технологий коммутации узлов. Начиная с 1993 г., два раза в год публикуется рейтинг Top500 supercomputers (http://top500.org), составляемый на основе результатов тестирования производительности суперЭВМ программой LINPACK. Производительность суперкомпьютеров измеряется в FLOPS (FLoating-point Operations Per Second).

Применение кластерных систем в первую очередь обосновано в расчетных задачах с высокой степенью независимости параллельных вычислений; системах массового обслуживания и виртуализации (где требу-

ется выделение обособленных ресурсов для решения какой-либо конкретной задачи или предоставления услуги индивидуальному заказчику); в системах высокой надежности, в которых присутствует необходимость дублирования отдельных узлов и подсистем, поддержание их в "горячем резерве". Соответственно существует несколько видов кластерных систем.

Кластеры высокой доступности (High Availability clusters, HA clusters, failover clusters) — группа компьютеров, имеющая в своем составе дополнительные узлы, обеспечивающие функционирование всего комплекса в целом в случае отказа одного или нескольких компонентов системы. Кластеры HA имеют в своем составе специальные программные средства, наблюдающие за процессами в системе.

Кластеры распределения нагрузки используют метод распределения нагрузки по нескольким вычислительным узлам посредством сетевых и программных решений. Благодаря этому можно достичь оптимального использования всех ресурсов кластера, увеличить пропускную способность, минимизировать время отклика, избежать перегрузок и сбоев.

Вычислительные кластеры. Со времен появления первого кластера Beowulf задачи, решаемые с помощью высокопроизводительных кластеров (HPC Cluster — High-performance computing cluster), не изменились. По-прежнему остается актуальным решение сложных вычислительных задач из различных областей физики, химии, биоинформатики и других направлений науки. Архитектура вычислительных кластеров сильно отличается от систем высокой доступности и балансировки нагрузки, в которых превалирует параллелизм задач. Рассматривая систему HPC, принято говорить о параллелизме процессов, их выполнении на различных узлах и передаче сообщений между ними в рамках одного большого параллельного приложения. В большинстве своем, узлы вычислительного кластера однотипны и обладают ограниченной конфигурацией с минимальным числом подсистем ввода/вывода и долговременного хранения информации. Последнее выражается в том, что многие узлы являются бездисковыми с отсутствующим графическим адаптером, дисководом и пр.

Гибридные вычислительные технологии и их практическое применение

Развитие гибридных суперкомпьютерных систем (в современном понимании) во многом связано с историей эволюционирования графических акселераторов. Созданные для увеличения скорости отображения графической информации первые ускорители графики способны были лишь на выполнение растеризации (преобразование треугольников в массив пикселей). По мере своего развития, графические процессоры приобретали и новую функциональность, например, возможность вершинного процессинга и обработки микропрограмм обсчета динамического освещения от нескольких источников.

Большинство операций с графикой хорошо поддается распараллеливанию. Именно этот факт обуславливает сложившуюся архитектуру графического процессора, в котором присутствует большое число функционально независимых процессорных ядер с маленьким объемом кеш-памяти и примитивной структурой конвейера.

В настоящее время значительная часть суперкомпьютеров, занимающих высокие позиции в TOP500, обладает гибридной архитектурой. В качестве ускорителей применяются процессоры NVIDIA 2050, NVIDIA 2070, AMD FirePro S10000 и др. Первое место в юбилейном 40-м рейтинге TOP500 (ноябрь 2012 г.) занял гибридный суперкомпьютер Titan Cray XK7, расположенный в Национальной лаборатории Окриджа. Titan имеет гибридную MPP архитектуру (спроектирован на базе процессоров AMD Opteron 6274 и ускорителей Nvidia Tesla K20X), общее число ядер — 560640 ед., в том числе 299008 ядер CPU и 261632 ядер (SMX) GPU. Обладая производительностью 17,59 PFLOPS, суперкомпьютер имеет отличный показатель производительность/потребляемая мощность, составляющий 2,14 TFLOPS/КВт. У японского конкурента, суперЭВМ классической негибридной архитектуры K-computer фирмы Fujitsu, этот показатель составляет лишь 0,83 TFLOPS/КВт при соизмеримой производительности.

Таким образом, технология гибридных суперкомпьютерных вычислений получила большое распространение. Наиболее производительные решения имеют MPP-архитектуру и используют в качестве интерконнекта проприетарные интерфейсы. СуперЭВМ из среднего и высокого сегмента по производительности, как правило, имеют кластерную архитектуру и используют стандартные технологии межмашинного обмена, такие как 100-гигабитный Ethernet и InfiniBand.

Особого внимания заслуживает применение гибридных вычислительных кластеров в промышленности. На электронном ресурсе http://www.nvidia.ru/object/cuda_app_tesla_ru.html приведен перечень как различных научно-технических и промышленных направлений, так и конкретных программных продуктов, ориентированных на использование гибридные вычисления. Это такие продукты, как AutoCAD, Matlab, Simulink, Mathematica, SEMCAD, Gauda и т.д.— перечень такого рода приложений, широко применяемых в промышленности, очень велик.

Немаловажно также применение гибридных кластеров для создания облачных инфраструктур. Существенно, что гибридные технологии обеспечивают значительно более высокую удельную производительность в расчете на киловатт мощности, чем классические (негибридные) системы [1].

Аппаратное и программное обеспечение гибридных кластерных решений

Современные кластерные системы имеют широкий спектр наименований устройств и технологий, используемый для их построения. Некоторые из них получили широкое распространение.

К наиболее популярным технологиям организации межузловых соединений можно отнести решения, построенные на базе Ethernet и InfiniBand. Производительность последних очень высока. InfiniBand хорошо масштабируется — возможно объединение базовых двунаправленных шин в группы (1x, 4x, 12x). Существует несколько режимов работы: Single Data Rate (SDR), Double Data Rate (DDR), Quad Data Rate (QDR), fourteen data rate (FDR) и enhanced data rate (EDR) с базовыми скоростями двунаправленных шин 2,5, 5, 10, 14, 25 Гбит/с соответственно. Одним из главных достоинств архитектуры InfiniBand является применение простой дифференциальной схемы обмена сигналами на физическом уровне. Длина InfiniBand-линий ограничена 17 м. Благодаря низкому энергопотреблению возможно создание однокристальных интегральных InfiniBand-коммутаторов с большим числом (>32 ед.) интегрированных приемопередатчиков. InfiniBand имеет стандартизированные подходы к организации RDMA-обменов, IPoIB (IP over InfiniBand), SRP (SCSI RDMA Protocol) и других протоколов.

В 2010 г. появился новый стандарт IEEE Std 802.3ba-2010, описывающий продвинутую технологию передачи данных 100-гигабитный Ethernet. При сохранении общей преемственности в рамках линейки стандартов 802.3 (сохранен формат кадров Ethernet 802.3, сохранены минимальные и максимальные размеры кадра FrameSize) был существенным образом переработан физический уровень (PHY). Крупнейшие вендоры (Mellanox, Reflex Photonics, Sumitomo Electric Industries, OpNext) уже представили свои решения для создания коммутационных суперкомпьютерных подсистем, используя в качестве физического уровня CFP-модули для многомодовых и одномодовых оптических волокон. Еще более перспективными с точки зрения стоимости, надежности и мощностных характеристик выглядят решения с оптической матрицей коммутации для создания объединительных плат. На рынке доступных средств коммутации по-прежнему лидирует аппаратура стандарта IEEE 802.3ab (Gigabit Ethernet). Устройства, построенные по этой технологии, обладают достаточной пропускной способностью и низкой стоимостью. Однако высокая латентность (> 100 мкс) затрудняет их использование в высокопроизводительных решениях.

В качестве ускорителей вычислений часто используются устройства компаний NVIDIA, ATI, а также матрицы ПЛИС фирм Xilinx и Altera.

Схемы ПЛИС (FPGA) могут выполнять функции любого устройства, если в них загрузить его логическое описание. Непосредственно процесс загрузки занимает несколько секунд и может выполняться неограниченное число раз. В последнее время среди исследователей и инженеров в области суперкомпьютерных вычислений большую популярность получила концепция «процессора одной задачи» (http://school-2010.hpc-russia.ru), которая заключается в оснащении узлов вычислительного кластера платами с микросхемами FPGA. У подобного подхода к организации вычислений есть несколько недостатков: ПЛИС на порядок уступают «жесткой» логике по скорости вычислений, соответственно выигрыш от распараллеливания задач на FPGA-укорителях должен как минимум компенсировать это отставание; отсутствие приемлемых средств преобразования кода «программистской» модели представления вычислений в схемотехническое логическое описание схемы; высокая стоимость FPGA.

Компания ATI предлагает свою технологию организации параллельных вычислений AMD FireStream. Первыми устройствами, поддерживающими эту технологию, являлись видеокарты Radeon X1900. В дальнейшем технология совершенствовалась. В качестве основного языка разработки параллельных модулей используется межплатформенный язык OpenCL [2]. Межплатформенная переносимость OpenCL-решения достигается путем предоставления программисту абстрактных моделей платформы, памяти, исполнения и программирования. Написание и отладка параллельных программ обеспечивается посредством среды ATI APP Software Development Kit. Основным недостатком технологии ATI является сложность написания программных модулей на OpenCL из-за низкоуровневого характера последнего.

Наибольшее распространение среди разработчиков высокопроизводительных кластерных систем получила технология NVIDIA CUDA. Программная модель CUDA (Compute Unified Device Architecture) позволяет внедрить описание параллелизма и иерархии памяти в язык программирования. Технология включает аппаратную поддержку параллельных вычислений, кроссплатформенный компилятор и систему исполнения.

Для упрощения «навигации» по всему множеству CUDA-устройств NVIDIA ввела понятие compute capability, выраженное парой чисел, разделенных точкой: major.minor. Здесь major обозначает глобальную архитектурную версию, а minor — небольшие изменения. Первые CUDA вычислители серии 8800 обладали compute capability 1.0. На момент написания статьи ускорители с архитектурой Kepler (Nvidia Tesla K20X) имели compute capability 3.5. Второе поколение GPU с архитектурой Kepler (K20 и K20X) имеет пиковую производительность для вычислений одинарной точности с плавающей точкой 3,52 и 3,95 TFLOPS (http://www.nvidia.com).

Помимо компиляторов высокоуровневых языков и средств построения параллельных вычислений на каждом вычислительном узле кластера должны быть установлены библиотеки организации межмашинного обмена. Наиболее популярной технологией программирования межузловых передач является MPI (Message Passing Interface). Интерфейс передачи сообщений MPI [3] предназначен для организации информационного взаимодействия в вычислительных системах с распределенной памятью. В соответствии с концепцией MPI для решения задачи разрабатывается одна параллельная программа, которая запускается одновремен-

но на всех процессорах суперкомпьютерной системы. По стандарту (последний датирован 21 сентября 2012 г.) MPI представляет собой программные средства, содержащие целое множество операций передачи данных, организованные в виде библиотек программных модулей для языков C и Fortran. Основу библиотеки составляют функции передачи сообщений между процессами. Операции передачи могут быть парные и коллективные. Стандарт реализован во множестве бесплатных и коммерческих библиотек: MPICH, Open MPI, LAM/MPI, Intel MPI, HP-MPI и др.

Выбор системного ПО (СПО) может оказаться непростым занятием по причине присутствия на рынке большого числа разнообразных дистрибутивов, компиляторов и библиотек от различных производителей. Существуют также различные дистрибутивы, специально собранные для использования в кластерных системах.

Возможные варианты СПО:

- ОС Windows, набор библиотек для вычислений, компиляторов и драйверов NVIDIA;
- UNIX-подобная ОС, например, один из дистрибутивов Linux и соответствующий набор для компиляции и запуска программ;
- специализированный дистрибутив для создания высокопроизводительного кластерного решения, например, CLIC или Rocks. В данном случае процесс настройки будет частично автоматизирован;
- LiveCD дистрибутив для наладки динамического кластера (PelicanHPC GNU Linux). В данном случае ОС на консоли запускается с диска и настраивается в автоматизированном режиме, а пользовательские программы и файлы хранить в виртуальном каталоге в ОЗУ. Загрузка узлов осуществляется по сети.

Построение бюджетного гибридного кластера

Для построения малобюджетного гибридного вычислительного кластера можно предложить следующую конфигурацию. В качестве средств коммутации использовать коммутатор и адаптеры стандарта IEEE 802.3ab (Gigabit Ethernet). Вычислительные узлы построить на базе платформы Intel (для возможности использования оптимизированных библиотек, например Intel Math Library). Центральный процессор не обязательно должен быть многоядерным (⩾6 ядер), так как в гибридном вычислительном кластере CPU отведена роль планировщика задач. Наилучшими показателями по соотношению производительность/цена сейчас обладают четырехядерные процессоры архитектур Ivy Bridge и Sandy Bridge (Socket 1155). Материнская плата, подбираемая под процессор, должна быть построенной на качественной элементной базе и обеспечивать стабильное питание процессору. Наличие дополнительных возможностей по «разгону» процессора, подключению дополнительных устройств в данном случае не обязательно и приведет лишь к удорожанию конечной системы. Объем оперативной памяти должен быть наибольшим из доступных по бюджету (⩾4 Гб). В качестве ускорителя вычислений можно использовать видеоадаптеры из пользовательского сегмента рынка, они гораздо дешевле профессиональных решений типа Tesla. Зачастую имеет смысл выбирать графические адаптеры из предыдущей линейки моделей, так как цены на них значительно ниже из-за смещения спроса в сторону новых моделей. Компания NVIDIA маркирует высокопроизводительные адаптеры префиксом GTX, именно на них стоит обратить внимание. Управляющий узел (консоль) кластера не обязательно должен обладать графическим CUDA-вычислителем и процессором мощнее, чем у узлов. Скорее всего, консоль придется оснастить дополнительным сетевым адаптером для доступа из внешней сети.

Если каждый узел оснащается двумя сетевыми адаптерами, можно организовать физическое разделение сетей передачи сообщений MPI и сообщений сетевой файловой системы NFS. Для этого также понадобится еще один обычный либо управляемый коммутатор с удвоенным числом портов и возможностью создания VLAN. Сконфигурировав вычислительную сеть подобным образом, можно перейти к концепции бездисковых узлов кластера, организовав общее хранилище на консоли либо на отдельном узле.

После установки ОС, драйверов видеокарт и сетевого оборудования, необходимых библиотек, можно приступать к установке прикладного ПО. Также имеется немало и свободно распространяемого ПО. В итоге можно получить собственный малобюджетный суперкомпьютер.

Заключение

Гибридные вычислительные кластеры заслуживают внимания как относительно недорогой способ создания своих собственных суперкомпьютеров. Такой кластер может быть организован даже на базе уже приобретенных персональных компьютеров, оснащенных современными дискретными видеокартами и объединенных в локальную сеть. Все современные ОС, в том числе и свободно распространяемые, имеют поддержку гибридных кластеров. Поддержка виртуализации на аппаратном (встроено в современные микропроцессоры) и программном (поддерживается современными ОС) уровне позволяет использовать эти кластеры в качестве инфраструктуры для облачных вычислений. Широкое многообразие прикладного ПО (как проприетарного, так и свободно распространяемого) позволяет решать сложные задачи технологического проектирования и моделирования.

Список литературы

1. *Черняк Л.* Облако // СуперКомпьютеры. №2(2). 2010.
2. *Линев А.В., Боголепов Д.К., Бастраков С.И.* Технологии параллельного программирования для процессоров новых архитектур. М.: МГУ. 2010.
3. *Гергель В.П.* Современные языки и технологии параллельного программирования. М.: МГУ. 2012.

Васильев Николай Петрович *– канд. техн. наук, доцент,* ***Ровнягин Михаил Михайлович*** *– аспирант кафедры «Компьютерные системы и технологии» НИЯУ МИФИ.*
Контактный телефон (499) 323-92-83.
E-mail: NPVasilyev@mephi.ru