РАЗРАБОТКА СИСТЕМЫ ДИНАМИЧЕСКОГО РАЗДЕЛЕНИЯ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ СУПЕРКОМПЬЮТЕРА НА ИЗОЛИРОВАННЫЕ ЧАСТИ

К.В. Бородулин, П.С Костенецкий, Ф.М. Мелёхин

Лаборатория суперкомпьютерного моделирования ФГБОУ ВПО «ЮУрГУ» (НИУ)

В настоящее время на суперкомпьютерах Лаборатории суперкомпьютерного моделирования ЮУрГУ [8] производятся расчеты большого количества задач, часть из которых требует выделения изолированных вычислительных ресурсов, например, для организации вычислений под ОС Microsoft Windows коммерческим заказчикам. При этом использование виртуализации не всегда является решением, так для работы виртуализации необходимо часть ресурсов выделять на поддержку гипервизора, что уменьшает количество ресурсов, доступных виртуальным машинам [1]. Проведенные в Лаборатории суперкомпьютерного моделирования эксперименты показали, что гипервизор VMWare [3] использует меньшее количество ресурсов для организации работы, чем открытые аналоги (KVM, Xen), но имеет высокую стоимость (порядка 45 млн. руб.), что не подходит для организации системы виртуализации. Также, на суперкомпьютере «Торнадо ЮУрГУ» проводились эксперименты по внедрению виртуализации адаптеров Infniniband, которые имели отрицательный результат. Другим подходом для управления изолированными вычислительными ресурсами является менеджер загрузки суперкомпьютера MOAB HPC Workload Manager [4], но данный продукт имеет большую стоимость и не имеет гибкой конфигурации, например, изоляция сети Infiniband через технологию Partitioning [2]. В связи с этим актуальна разработка системы динамического разделения вычислительных ресурсов суперкомпьютера на изолированные части.

Система динамического разделения вычислительных ресурсов суперкомпьютера на изолированные части, разработанная в Лаборатории суперкомпьютерного моделирования, использует очередь задач SLURM [5] для выделения узлов, систему применения конфигурации Puppet [6] для восстановления узлов для работы в очереди задач, комплекс утилит для администрирования распределенных компьютерных систем xCAT [7]. При изолировании узлов система выполняет автоматическое развертывание нужной операционной системы и ее настройку под задачу. При возврате узла, система выполняет реконфигурацию узла, используя Puppet для проверки и настройки конфигурации операционной системы узла для работы в очереди задач суперкомпьютера. Система представляет собой 3 главных программы для реализации алгоритма работы по выделению/возврату узлов в очередь:

1. главный сервис для управления узлами, недоступен пользователям;
2. скрипт для возврата узлов в нормальный режим работы после завершения расчета, выполняется пользователем;
3. скрипт для выделения необходимого числа узлов для создания изолированной части вычислительных ресурсов, выполняется пользователем.

Для выделения узлов из очереди задач используются команды очереди задач SLURM. При выделении узла производится замена Vlan порта коммутатора Ethernet путем выполнения скрипта vlan.sh, работающего следующим образом:

1. Поиск номера порта в базе данных коммутаторов xCAT;
2. Выполнение скрипта, выполняющего вход на нужный коммутатор;
3. проверка, существует ли в базе Vlan коммутатора Vlan с нужным номером, если его нет, то производятся его создание;
4. изменение конфигурации необходимого порта на коммутаторе.

Для всех узлов, изолированных под одну задачу (заказчика), применяется один Vlan. Другие узлы не могут просмотреть пакеты, проходящие в данном Vlan.

Также, в данной системе предусмотрена возможность двойной загрузки узлов для быстрого выделения узлов без развертывание нужной операционной системы и ее настройку под задачу (операционная система настраивается 1 раз на узле и после используется уже готовая конфигурация). Данная технология позволяет значительно сократить время выделения узлов, с 20 мин для 80Гб образа ОС до 3 мин.

Дальнейшим направлением работ будет динамическое выделение ресурсов под часть OpenStack и реализация изолирования сети Infiniband между группами узлов.

ЛИТЕРАТУРА:

1. Lance Albertson. Comparsing Ganeti to other Private Cloud Platforms: [Электронный ресурс] URI: http://linuxfestnorthwest.org/sites/default/files/slides/Comparing Ganeti to other private cloud platforms.pdf (дата обращения: 29.05.14).

2. Rajkumar Buyya, Toni Cortes, Hai Jin. High Performance Mass Storage and Parallel I/O: Technologies and Applications, Wiley, 2011, pp: 688.
3. Marshal D., Beaver S., McCarty J. VMware ESX Essentials in the Virtual Data Center. CRC Press, 2008. 256p.
4. MOAB HPC Workload Manager. URL: http://www.adaptivecomputing.com/products/hpc-products/comparison-grid/ (дата обращения: 29.05.14).
5. SLURM: A Highly Scalable Resource Manager [Электронный ресурс] (URL: https://computing.llnl.gov/linux/slurm, дата обращения 29.05.14)
6. Sean Walberg. Automate system administration tasks with puppet, Linux Journal archive, 2008, Article No. 5
7. Egan Ford, Brad Elkin, Scott Denham, Benjamin Khoom, Matt Bohnsack, Building a Linux HPC Cluster with xCAT. IBM Redbooks, 2002, pp: 276.
8. Суперкомпьютеры СКЦ ЮУрГУ URL: http://supercomputer.susu.ru/computers/ (дата обращения 29.05.14)