Создание комплекса высокопроизводительных вычислений Сибирского федерального университета

А.П. Бугай, С.В. Маколов

Сибирский федеральный университет, Красноярск

В настоящее время кластерные системы уже не роскошь доступная крупным компания, а необходимость для любого учреждения, занимающегося исследованиями в той или иной области. При создании Сибирского федерального университета (СФУ) встал вопрос о необходимости вычислительного комплекса способного решать актуальные задачи фундаментальной и прикладной науки. Для создания необходимого вычислительного ресурса было закуплено оборудование компании IBM, а именно:

16 шасси IBM BladeCenter H.

224 узла кластера IBM Blade HS21.

4 сервера IBM x3650.

2 сервера IBM x3950.

Система хранения IBM DS3400

Оборудование объединено в единый комплекс с использованием сети Ethernet 1 Gb и InfiniBand 4X. Также были приобретены еще несколько малых кластерных систем, в состав которых входят:

1 шасси IBM BladeCenter H.

14 узлов кластера IBM Blade HS21.

1 сервер IBM x3650.

Система хранения IBM DS3400

Ethernet 1Gb.

Все оборудование функционирует под управлением операционной системы SLES 10, также установлены система управления кластером IBM Cluster System Management и параллельная файловая система IBM General Parallel File System. Но данного программного обеспечения не достаточно для эффективного управления и использования всего комплекса.

Для использования любой кластерной системы необходимо решить следующие проблемы:

- установка и обновление необходимо программного обеспечения;
- мониторинг состояния вычислительных узлов и вычислительных ресурсов;
- предоставления пользователю удобного интерфейса для запуска задач и проверки их состояния;
- динамическое распределение нагрузки на вычислительные узлы кластера.

Рассмотрим один из возможных способов решения данных проблем.

Установка и обновление необходимо программного обеспечения.

Решить данную задачу позволяет уже установленный программный пакет IBM CSM (Cluster System Management). CSM предоставляет в распоряжение системных администраторов единую точку контроля для установки, настройки, эксплуатации и обновления узлов кластера, работающих под управлением Linux. Для этого необходимо написать исполняемые скрипты, которые по некоторому условию запустят установку или обновления операционных систем на узлах и настройку программного обеспечения. CSM позволяет устанавливать драйверы для оборудования и программные пакеты во время установки ОС. Данная возможность позволяет быстро и в автоматическом режиме устранять неисправности при функционировании узлов кластера, если они вызваны не физическим повреждением оборудования.

В настоящий момент решено перевести управление кластером на другую систему, а именно xCAT (Extreme Cloud Administration Tool). Связано это с тем, что в последней

версии IBM CSM отсутствует поддержка ОС SLES 11. Кроме того xCAT является продуктом Open Source, и предоставляет более удобный интерфейс для управления кластерными комплексами и xCAT совместим со всеми программами которые мы используем.

Мониторинг состояния вычислительных узлов и вычислительных ресурсов.

В условиях роста вычислительных центров потребность в эффективных инструментах мониторинга вычислительных ресурсов становится важной как никогда. Реализовать мониторинг кластеров СФУ, было принято, используя зарекомендовавшие себя системы мониторинга кластеров Ganglia и Nagios.

Ganglia – масштабируемая распределенная система мониторинга для высокопроизводительных кластеров с иерархической архитектурой. Система распространяется бесплатно и доступна в открытых исходных кодах. Данная система мониторинга проста в установке и конфигурировании. Не смотря на свою простоту, данный сервис позволяет собирать разнообразную статистику по работе узлов кластера, начиная временем загрузки вычислительного узла, и заканчивая количеством принятых и отправленных пакетов.

Nagios в большей степени применяется как средство уведомления, в отличие от Ganglia, больше сфокусированного на сборе и отслеживании изменений значений различных параметров – метрик. Nagios следит за указанными узлами и службами, и оповещает администратора в том случае, если какие-то из служб прекращают (или возобновляют) свою работу. Данная система имеет ряд особенностей, которых не имеет Ganglia:

- отправка оповещений в случае возникновения проблем со службой или хостом (с помощью почты, пейджера, смс, или любым другим способом, определенным пользователем через модуль системы);
- возможность определять обработчики событий произошедших со службами или хостами для проактивного разрешения проблем.

Эти системы (Ganglia и Nagios) местами имеют схожую функциональность, но все-таки они довольно сильно различаются, и их совместное использование может компенсировать недостатки каждого продукта.

Предоставления пользователю удобного интерфейса для запуска задач и проверки их состояния.

В настоящий момент разрабатывается интернет портал Комплекса Высокопроизводительных Вычислений Сибирского Федерального университета. Данный портал должен объединить в себе:

- мониторинг состояния узлов;
- инструментарий по постановки задачи на счет и получение результатов;
- систему управления кластером для администраторования.

Динамическое распределение нагрузки на вычислительные узлы кластера.

Для динамического распределения нагрузки было решено использовать менеджер ресурсов Torque в связке с локальным планировщиком задач Maui. Именно эти программные продукты были выбраны потому, что дальнейшее развитие создаваемого комплекса высокопроизводительных вычислений связано с внедрением его в грид-консорциум RDIG, а для корректной работы в его структуре рекомендуется данное программное обеспечение. Еще одним большим плюсом к выбору именно этого ПО послужило то, что это продукты Open Source.

Менеджер ресурсов Torque позволяет автоматически распределять вычислительные ресурсы между задачами, управлять порядком их запуска, временем работы, получать информацию о состоянии очередей. При невозможности запуска задач немедленно, они ставятся в очередь и ожидают, пока не освободятся нужные ресурсы.