ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 2(027), 2015 с.??—??

УДК (004.45)

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

Система динамического разделения суперкомпьютера на изолированные логические кластеры

Аннотация. Описываемая в данной работе программная система позволяет пользователям Лаборатории суперкомпьютерного моделирования Южно-Уральского государственного университета самостоятельно управлять вычислительными ресурсами суперкомпьютера Торнадо ЮУрГУ, обеспечивая создание для их расчетов логически изолированных вычислительных кластеров. Получаемые пользователями кластеры содержат требуемое количество физических вычислительных узлов с требуемой пользователем операционной системой. Система позволяет клонировать образы операционных систем с одного узла на другой. Вычислительные узлы различных клиентов изолированы между собой на аппаратном уровне Ethernet и InfiniBand сетей. Клиенты получают прямой доступ (RDP, SSH) к вычислительным узлам путем подключения к ним публичного адреса.

Ключевые слова и фразы: Управление ресурсами суперкомпьютера, Оптимизация загрузки суперкомпьютера, Система развертывания сервисов.

Введение

В настоящее время в связи с развитием информационных технологий возрастает потребность в вычислительных ресурсах. Примером такого роста может служить рейтинг Топ-500 мощнейших суперкомпьютеров мира [3]. С увеличением количества вычислительных узлов в суперкомпьютерах подобного рода возрастает потребность в системах, способных упростить задачи администрирования и мониторинга.

Зачастую для решения различных вычислительных задач необходимы различные операционные системы, имеющие разные наборы программного обеспечения. К примеру, для решения задач 3D моделирования чаще всего используется ОС Windows, а для решения задач виртуализации чаще всего используется ОС Linux [5].

Предложенная в данной работе система позволяет пользователям суперкомпьютера самостоятельно управлять вычислительными ресурсами суперкомпьютера, а также ресурсами коммуникационных сетей. Используемые пользователем вычислительные и коммуникационные ресурсы объединяются в проект, доступ к которому имеет только пользователь. Благодаря подсистеме двойной загрузки, нет необходимости каждый раз заново производить установку операционной системы, т.к.

ОС пользователя уже установлена в качестве дополнительной на каждом вычислительном узле суперкомпьютера.

1. Структура программной системы

Разработанная система состоит из четырех подсистем (см. **рис. 1**):

- управляющая подсистема (controller);
- сетевая подсистема (network-agent);
- подсистема развертывания (deployment-server и deployment-client);
- подсистема обеспечения двойной загрузки.

Все подсистемы (кроме подсистемы обеспечения двойной загрузки) используют единую базу данных.

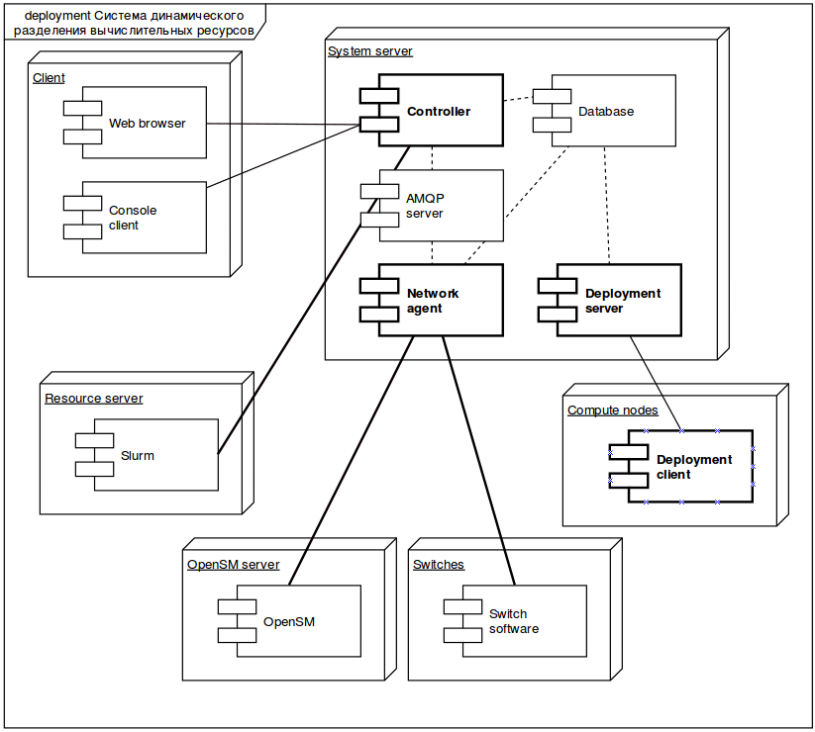

Рис. 1. Диаграмма размещения

1.1. Управляющая подсистема

Управляющая подсистема предоставляет пользователю возможность взаимодействовать с системой посредством веб-интерфейса или RESTful API [2]. Взаимодействие между пользователем и системой осуществляется с использованием защищенного протокола HTTPS. Благодаря данной подсистеме пользователь имеет следующие возможности:

- контролировать процесс установки операционных систем на выделенных ему вычислительных узлах;
- управлять публичными интернет адресами, необходимыми для осуществления доступа к узлам из сети интернет;
- управлять созданием образов операционных систем вычислительных узлов, необходимых для последующего развертывания;
- перезагружать любой выделенный ему вычислительный узел;
- возвращать выделенные ранее узлы в вычислительное поле суперкомпьютера.

Данная подсистема взаимодействует с менеджером ресурсов SLURM, который обеспечивает выделение свободных вычислительных узлов из очереди суперкомпьютера для дальнейшего перевода в монопольное пользование.

Также управляющая подсистема взаимодействует с сетевой подсистемой путем отправки ей уведомлений по протоколу AMQP [1] что обеспечивает оперативное уведомление подсистемы об изменении топологии сетей Ethernet и InfiniBand, изменении IP-адресов узлов, добавлении и удалении узлов из сети, а также о подключении публичных адресов к узлам суперкомпьютера.

1.2. Сетевая подсистема

Сетевая подсистема (network-agent) предназначена для управления Ethernet и InfiniBand сетями. Сетевая подсистема необходима для обеспечения процессов развертывания, создания образов и конфигурирования узлов.

Данная подсистема обеспечивает следующие возможности:

- производит динамическое конфигурирование DHCP сервера, что позволяет вычислительным узлам получать актуальную информацию об ip адресах и другой информации о сетях;
- производит конфигурирование Ethernet свитчей, что позволяет динамически изменять логическую топологию Ethernet сети, благодаря чему Ethernet сети различных пользователей изолированы между собой;
- производит конфигурирование InfiniBand свитчей посредством изменения конфигурации менеджера сетей OpenSM, что позволяет динамически изменять логическую топологию InfiniBand сети, благо-

даря чему InfiniBand сети различных пользователей изолированы между собой;
- производит подключение/отключение публичных интернет адресов к внутренним адресам узлов, что позволяет пользователям получать доступ к узлам из сети Интернет;
- производит создание и конфигурирование сетевых интерфейсов сервера, на котором установлена данная подсистема, что необходимо для функционирования системы.

В InfiniBand сети каждый порт вычислительного узла имеет уникальный guid, а каждая сеть имеет свой уникальный pkey. Guid – это 64-битный идентификатор порта InfiniBand устройства. Pkey (partition key) – это 16 битный идентификатор раздела.

Раздел определяет набор InfiniBand узлов, которые могут взаимодействовать друг с другом. Разделы используются для логического разделения сетей проектов, что обеспечивает изоляцию различных проектов пользователей друг от друга [6].

Сетевая подсистема в случае получения уведомления по протоколу AMQP от управляющей подсистемы производит обновление конфигурационного файла менеджера OpenSM. Обновление производится на основе информации об узлах, хранящейся в базе данных.

1.3. Подсистема развертывания

Подсистема развертывания является основной подсистемой, участвующей в процессе развертывания и создания шаблонов на основе образа локального диска узла [4].

Данная подсистема состоит из двух частей, взаимодействующих по протоколу HTTP:

(1) сервер развертывания (deployment-server);
(2) клиент развертывания (deployment-client).

В подсистеме развертывания была реализована возможность «отложенной установки», данная возможность позволяет пользователю выделить узлы на основе создаваемого шаблона еще до его полного завершения. Данная возможность позволяет уменьшить время развертывания ОС за счет того, что процесс инициализации клиентов развертывания происходит параллельно с процессом создания шаблона.

1.4. Подсистема обеспечения двойной загрузки

Для обеспечения возможности работы двух операционных систем на вычислительных узлах суперкомпьютера была реализована подсистема обеспечения двойной загрузки. Данная подсистема позволяет в случае изменения образа диска виртуальной машины одного из клиентов динамически синхронизировать образ операционной системы с содержимым

локальных дисков вычислительных узлов суперкомпьютера без необходимости в остановке задач пользователей и переустановке операционной системы.

Поддержка двойной загрузки позволяет использовать вторую операционную систему сразу после выделения вычислительного узла из очереди и перезагрузки, что позволяет уменьшить время подготовки узла к использованию исключив этап переустановки операционной системы. Подсистема обеспечения двойной загрузки состоит из двух частей, взаимодействующих друг с другом по HTTP протоколу (см. **рис. 2**):

- сервер двойной загрузки;
- клиент двойной загрузки.

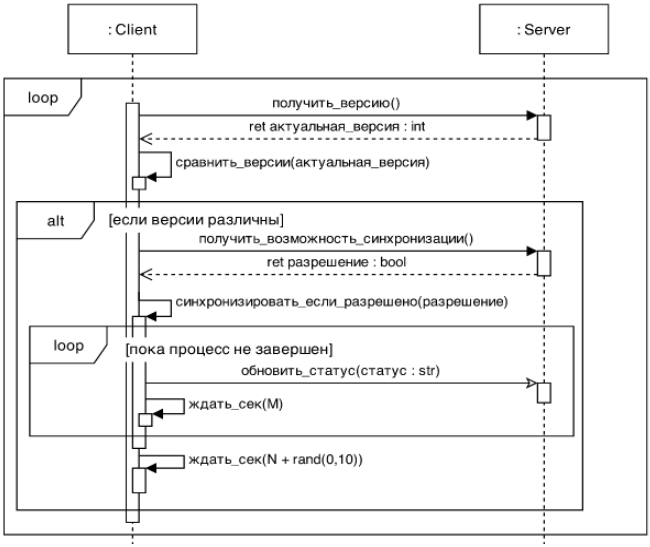

Рис. 2. Диаграмма взаимодействия клиента и сервера двойной загрузки

Заключение

Система позволяет клиентам суперкомпьютера разворачивать любые поддерживаемые аппаратным обеспечением операционные системы на вычислительные узлы суперкомпьютера. После завершения использования узла клиентом он может сохранить образ диска узла, с настроенной необходимым ему образом операционной системой, для последующего развертывания (клонирования) на другие узлы, либо может сразу вернуть узел в вычислительное поле суперкомпьютера, где SSD накопи-

тель узла будет отформатирован и на него будет установлена базовая ОС суперкомпьютера.

Благодаря подсистеме двойной загрузки есть возможность иметь две предустановленные операционные системы (Linux+Windows) на каждом твердотельном накопителе вычислительного узла, что в конечном итоге позволяет продлить срок службы накопителей, т.к. нет необходимости перезаписывать ОС при каждом выделении узла клиентом. К тому же это приводит к сокращению времени выделения узла для клиента с 20 минут (время перезагрузки и установки) до 3-4 минут (только время перезагрузки и выбора другой ОС).

После выделения узлов из вычислительного поля суперкомпьютера происходит смена vlan на портах Ethernet коммутаторов и перестроение логической топологии InfiniBand сети (partitions). Таким образом, все вычислительные узлы различных клиентов (проектов) изолированы между собой на аппаратном уровне.

Клиенты получают прямой доступ (RDP, SSH) к вычислительным узлам путем подключения к ним публичного адреса. Также, система предоставляет работающий через HTTPS протокол публичный RESTful API, что позволяет упростить интеграцию с автоматизированными системами клиента.

Список литературы

[1] Keig A. Instant Rabbitmq Messaging Application Development How-To. – UK: Packt Publishing Ltd, 2013. – 54 p.

[2] Richardson K., Ruby S. RESTful Web Services. – USA: O'Reilly Media, Inc., 2008. – 454 p.

[3] Strohmaier E., Dongarra J., Simon H., Meuer M. TOP500 Supercomputer Sites. URL: http://top500.org/.

[4] Мелёхин Ф.М., Бородулин К.В., Костенецкий П.С. Разработка системы динамического разделения вычислительных ресурсов суперкомпьютера на изолированные части // Научный сервис в сети Интернет: многообразие суперкомпьютерных миров: Труды Международной суперкомпьютерной конференции. М.: Издательство МГУ, 2014. С. 321–322.

[5] Костенецкий П.С., Семенов А.И., Соколинский Л.Б. Создание образовательной платформы ”Персональный виртуальный компьютер” на базе облачных вычислений // Научный сервис в сети Интернет: экзафлопсное будущее: Труды Международной суперкомпьютерной конференции. М.: Издательство МГУ, 2011. С. 374–377.

[6] Бородулин К.В., Мелёхин Ф.М. Логическая изоляция сегментов сети InfiniBand при разделении ресурсов вычислительного кластера между пользователями // Параллельные вычислительные технологии

(ПаВТ'2015): Труды международной научной конференции, 2015. С. 498.

Об авторе:

Мелёхин Федор Михайлович
ФГБОУ ВПО «ЮУрГУ» (НИУ), программист Суперкомпьютерного центра лаборатории «Суперкомпьютерное моделирование»,
e-mail: melekhinfm@susu.ru

Бородулин Кирилл Владимирович
ФГБОУ ВПО «ЮУрГУ» (НИУ), директор Суперкомпьютерного центра ЛСМ, старший преподаватель кафедры Системного программирования
e-mail: borodulinkv@susu.ru

Костенецкий Павел Сергеевич
ФГБОУ ВПО «ЮУрГУ» (НИУ), руководитель Лаборатории суперкомпьютерного моделирования, к.ф.-м.н., доцент кафедры Системного программирования
e-mail: kostenetskiy@susu.ru

F.M. Meljohin, K.V. Borodulin, P.S. Kostenetsky. *A System for Dynamic Dividing of Supercomputer's Computational Resources into Isolated Logical Clusters*

ABSTRACT. Described system allows users to manage computing resources of the supercomputer, provides formation of logically isolated cluster for user's computing projects. Users gets clusters with the required number of nodes and the required operation system are installed on node.

System allows cloning images of operating system from one node to another. Computing nodes of various clients are isolated among themselves at the Ethernet and InfiniBand network. Clients get direct access (RDP, SSH) to computing nodes by attaching public IP address to them.

Key Words and Phrases: Supercomputer resource management, Service deployment system.