ОРГАНИЗАЦИЯ ИНФРАСТРУКТУРЫ ОБЛАЧНЫХ ВЫЧИСЛЕНИЙ НА ОСНОВЕ SDN СЕТИ

УДК 681.3

Алексей Анатольевич Ефименко,
к.т.н., старший научный сотрудник, Институт точной механики и вычислительной техники имени С. А. Лебедева РАН
Тел.: 8 (910) 421-49-50
Эл. почта: alex192@mail.ru

Сергей Витальевич Федосеев,
к.т.н., доц., зав. каф. Математического обеспечения информационных систем и инноватики, Московский государственный университет экономики, статистики и информатики (МЭСИ)
Тел.: 8 (495) 442-80-98
Эл. почта: SFedoseev@mesi.ru

В статье представлены основные подходы к организации инфраструктуры облачных вычислений на основе SDN сети в современных Центрах обработки данных (ЦОД). Определены основные показатели эффективности управления сетевой инфраструктурой ЦОД. Приведены примеры решений по созданию виртуальных сетевых устройств.

Ключевые слова: облачные вычисления, SDN сеть, Центры обработки данных, виртуальные сетевые устройства.

Alexey A. Efimenko,
PhD in Technological Sciences, Senior Researcher, Lebedev Institute of Precision Mechanics and Computer Engineering, Russian Academy of Sciences
Tel.: 8 (910) 421-49-50
E-mail: alex192@mail.ru

Sergey V. Fedoseev,
PhD in Technological Sciences, Associate Professor, Head of the Chair of Software of Information Systems and Innovations, Moscow State University of Economics, Statistics and Informatics (MESI)
Tel.: 8 (495) 442-80-98
E-mail: SFedoseev@mesi.ru

ORGANIZATION OF CLOUD COMPUTING INFRASTRUCTURE BASED ON SDN NETWORK

The article presents the main approaches to cloud computing infrastructure based on the SDN network in present data processing centers (DPC). The main indexes of management effectiveness of network infrastructure of DPC are determined. The examples of solutions for the creation of virtual network devices are provided.

Keywords: cloud computing, SDN network, data processing centers, virtual network devices.

1. Принципы создания виртуальной сетевой инфраструктуры современных центров обработки данных

Виртуализация серверов и облачные вычисления в публичных, частных и гибридных сетях послужили толчком к изменениям в вычислительных сетях центров обработки данных (ЦОД). Вычислительные сети ЦОД вступают в период инноваций и преобразований. Виртуализированное прикладное программное обеспечение заменяет сетевые устройства и сетевая инфраструктура становится все более программируемой.

SDN (программно-конфигурируемая сеть) обычно рассматривается как описательный термин для широкого диапазона развивающихся решений, которые вносят развитые логико-информационные возможности к взаимодействию между сетевыми службами и сетевой инфраструктурой, чтобы динамически приспособить сеть к потребностям ЦОД.

Основные принципы SDN:

- разделение процессов передачи и управления данными;
- единый, унифицированный, независящий от поставщика интерфейс между уровнем управления и уровнем передачи данных;
- логически централизованное управление сетью, осуществляемое с помощью контроллера с установленной сетевой операционной системой и реализованными поверх сетевыми приложениями;
- виртуализация физических ресурсов сети.

В архитектуре SDN обычно выделяют три уровня:

- *инфраструктурный уровень*, предоставляющий набор сетевых устройств (коммутаторов и каналов передачи данных);
- *уровень управления*, включающий в себя сетевую операционную систему, которая обеспечивает приложениям сетевые сервисы и программный интерфейс для управления сетевыми устройствами и сетью;
- *уровень сетевых приложений* для гибкого и эффективного управления сетью.

Наиболее перспективным и активно развивающимся стандартом для SDN является OpenFlow – открытый стандарт, в котором описываются требования, предъявляемые к коммутатору, поддерживающему протокол OpenFlow для удаленного управления.

Согласно спецификации 1.3 стандарта OpenFlow [1], взаимодействие контроллера с коммутатором осуществляется посредством протокола OpenFlow – каждый коммутатор должен содержать одну или более таблиц потоков (flow tables), групповую таблицу (group table) и поддерживать канал (OpenFlow channel) для связи с удаленным контроллером – сервером. Спецификация не регламентирует архитектуру контроллера и API для его приложений. Каждая таблица потоков в коммутаторе содержит набор записей (flow entries) о потоках или правила. Каждая такая запись состоит из полей-признаков (match fields), счетчиков (counters) и набора инструкций (instructions).

Логически-централизованное управление данными в SDN сети предполагает вынесение всех функций управления сетью на отдельный физический сервер, называемый контроллером, который находится в ведении администратора сети. Контроллер может управлять как одним, так и несколькими OpenFlow-коммутаторами и содержит сетевую операционную систему, предоставляющую сетевые сервисы по низкоуровневому управлению сетью, сегментами сети и состоянием сетевых элементов, а также приложения, осуществляющие высокоуровневое управление сетью и потоками данных.

SDN технология OpenFlow применяется для захвата и анализа сетевого трафика в сервисах обеспечения безопасности и мониторинга, используются в обычных и облачных центрах обработки данных.

2. Управление сетевой инфраструктурой ЦОД

Управление сетевой инфраструктурой в OpenFlow представляет собой процесс максимизации надежности и производительности сетевых ресурсов в целях оптимизации сетевой доступности (готовности) и времени реакции в сети .

Эффективность управления сетью определяют шесть факторов:

– объекты управления (физические элементы сети и программные элементы);

– персонал, который участвует в управлении (уровень знаний персонала);

– уровни доступа к управлению элементами сети;

– средства управления (специальные программные средства и инструменты);

– степень интеграции других процессов в управление сетевой инфраструктурой;

– ожидаемые услуги и качество предоставляемых услуг (QoS).

Оценка эффективности управления сетевыми инфраструктурами представляет собой анализ производительности и надежности (Performance Management) и определяется по критериям, путем выставления весовых или метрических коэффициентов. Порядок оценки эффективности в соответствии с Моделью управления сети ISO:

1) Сбор информации об эффективности по тем переменным, которые представляют интерес для администраторов сети: сформулировать критерии эффективности работы сети. Чаще всего такими критериями служат производительность и надежность, для которых в свою очередь требуется выбрать конкретные показатели оценки, например, время реакции и коэффициент готовности, соответственно;

2) Анализ информации для определения базовых уровней: определить множество варьируемых параметров сети, прямо или косвенно влияющих на критерии эффективности. Эти параметры действительно должны быть варьируемыми, то есть нужно убедиться в том, что их можно изменять в некоторых пределах по вашему желанию. Так, если размер пакета какого-либо протокола в конкретной операционной системе устанавливается автоматически и не может быть изменен путем настройки, то этот параметр в данном случае не является варьируемым, хотя в другой операционной системе он может относиться к изменяемым по желанию администратора, а значит и варьируемым. Другим примером может служить пропускная способность внутренней шины маршрутизатора – она может рассматриваться как параметр оптимизации только в том случае, если вы допускаете возможность замены маршрутизаторов в сети;

3) Определение соответствующих порогов эффективности: определить порог чувствительности для значений критерия эффективности. Так, производительность сети можно оценивать логическими значениями «Работает» / «Не работает», и тогда оптимизация сводится к диагностике неисправностей и приведению сети в любое работоспособное состояние. Другим крайним случаем является тонкая настройка сети, при которой параметры работающей сети (например, размер кадра или величина окна неподтвержденных пакетов) могут варьироваться с целью повышения производительности (например, среднего значения времени реакции) хотя бы на несколько процентов. Как правило, под оптимизацией сети понимают некоторый промежуточный вариант, при котором требуется выбрать такие значения параметров сети, чтобы показатели ее эффективности существенно улучшились, например, пользователи получали ответы на свои запросы к серверу баз данных не за 10 секунд, а за 3 секунды, а передача файла на удаленный компьютер выполнялась не за 2 минуты, а за 30 секунд.

3. Перспективные решения виртуализации сетевой инфраструктуры

Вопрос единого управления группой коммутаторов чрезвычайно актуален в традиционных сетях передачи данных, и именно потребность в его решении во многом подстегивает интерес к SDN. Такое управление уже достаточно давно реализовано для разработок, которые можно объединить общим понятием «виртуальное шасси». Речь идет об объединении нескольких физических коммутаторов в логически единое устройство с общей системой управления.

Многие решения класса «виртуальное шасси» ведут свое происхождение от обычных стековых коммутаторов. В процессе развития этих продуктов короткие шины, ограничивающие возможность пространственного разнесения коммутаторов стека, заменялись высокоскоростными линиями связи (медными или оптическими), которые позволяли размещать устройства в разных стойках ЦОД, – такие решения иногда называют горизонтальными стеками.

«Виртуальные шасси» предлагают многие производители. Одни из самых известных решений – это Virtual Switching System (VSS) компании Cisco [3] и Intelligent Resilient Framework (IRF) компании HP (получено ею в результате покупки 3Com). Согласно заявлениям этих производителей, при использовании 10-гигабитных линий связи коммутаторы в рамках одного «виртуального шасси» могут быть разнесены не только по разным стойкам одного ЦОД, но и на расстояния в десятки километров, что позволяет построить единую сеть для территориально распределенного ЦОД.

В большинстве решений класса «виртуальное шасси» предусматривается, что один из коммутаторов группы получает статус главного (мастер). Это ставит вопрос о том, как «виртуальное шасси» поведет себя в случае нарушения связи между его компонентами и что будет с теми устройствами, которые окажутся отрезанными от мастера. Результат может быть различным – вплоть до потери их работоспособности. Обычно в «отрезанном» сегменте быстро выбирается свой мастер, но это может привести к другой проблеме – конфликту с основным мастером при восстановлении связи. Каждый производитель предлагает подробные рекомендации по грамотному проектированию, настройке и обслуживанию таких систем, но использование нестандартных (фирменных) алгоритмов практически исключает возможность дать какие-либо общие советы на этот счет.

Следует заметить, что большинство виртуальных шасси предоставляют все преимущества поддержки множественных путей передачи для внешних устройств.

Одна из задач SDN-обеспечение высокого уровня масштабирования (от сотен до тысяч 10-гигабитных портов) при неизменности ключевых характеристик. Другими словами, при увеличении числа портов в SDN ни задержка, ни уровень переподписки не должны возрастать. При традиционных методах расширения сети эти задачи решить чрезвычайно сложно, поскольку подключение каждого нового коммутатора увеличивает число транзитных узлов, а потому – и задержку. Кроме того, вся подсистема коммутации с точки зрения управления должна была выглядеть как одно устройство и обеспечивать передачу любого трафика (L2, L3, FCoE, iSCSI, NAS и пр.).

Отправной точкой для разработки виртуальных коммутаторов SDN послужила архитектура обычного модуль-

ного коммутатора, которая отвечает большинству перечисленных требований за исключением одного, но очень важного: возможности масштабирования ограничены размерами шасси (числом слотов) такого коммутатора. Традиционные способы наращивания емкости предполагают установку еще одного устройства, но при этом очевидно повышается сложность и ухудшается управляемость системы. В SDN кардинально изменяется модель масштабирования.

По сути в SDN предлагается применять распределенный коммутатор, в котором вместо пассивной шины, обычно соединяющей линейные карты и платы матрицы коммутации, используются оптические каналы, связывающие устройства Node (выполняют функции линейных карт) и Interconnect (применяются вместо традиционных матриц коммутации). Таким образом устранено ограничение, накладываемое размером физического шасси, при сохранении большинства преимуществ единого устройства. Управление реализовано с помощью внешнего контроллера Director. Служебный трафик передается по выделенной сети управления (out-of-band)

Развертывание ЦОД можно начинать с установки в качестве устройств Interconnect двух небольших продуктов от Juniper QFX3600-I, к которым каналами 40G подключается до 16 узлов Node (всего 384 порта 10G) [2]. По мере расширения ЦОД можно добавить еще два устройства QFX3600-I, тогда число поддерживаемых портов увеличивается в два раза – до 768. ЦОД, в котором функционал QF/Interconnect реализуется с помощью устройств QFX3600-I, имеет суффикс M (Micro).

Следующая стадия масштабирования – замена QFX3600-I на более мощное оборудование межсоединения: модульные устройства QFX3008 с восемью слотами вмещают до 128 портов 40G (QSFP+). При использовании четырех таких устройств общая емкость фабрики превышает 6000 портов 10G. Заметим, что при модернизации устройства QFX3600-I можно использовать в качестве Node, для чего каждый порт 40G потребуется конвертировать в четыре порта 10G с помощью специального разветвителя. Пока максимальное расстояние между устройствами Node и Interconnect составляет 100 и 150 м при использовании оптики OM3 и OM4 соответственно.

Интересно отметить, что внутри ЦОД используются протоколы IS-IS и BGP, которые относятся к уровню L3. Например, когда новый узел Node подключается к сети, именно протокол IS-IS служит для его автоматического обнаружения, после чего Director автоматически генерирует конфигурационные параметры (IP-адрес и пр.) и направляет их новому узлу. Протокол BGP нужен для распределения динамической информации, например новых MAC-адресов.

Другой пример масштабирования коммутатора – решение Cisco Fabric Extender (FEX). Его основу составляют «материнские» коммутаторы (серий Cisco Nexus 5000, Nexus 7000 или UCS Fabric Interconnect), к которым подключаются выносы FEX, выполняющие функции удаленной линейной платы. В качестве выносов могут использоваться коммутаторы Nexus 2000, сетевые модули UCS 2100 Fabric Extender для блейд-серверов Cisco UCS, а также решение Cisco Nexus B22 Fabric Extender для блейдсерверов HP. Особенность таких ЦОД состоит в том, что она может охватывать интерфейсы сервера (с помощью Cisco Adapter FEX) и даже виртуальные машины (технология VM-FEX). Это означает проникновение сети «внутрь» сервера, при этом в единой плоскости коммутации могут находиться не только порты коммутаторов, но и серверные адаптеры и виртуальные машины.

Технология TRILL определена в серии документов организации IETF (RFC 5556, 6325, 6327, 6349), но некоторые механизмы находятся только в стадии рассмотрения. Часто ее называют маршрутизацией на уровне L2. Как известно, классическая маршрутизация выполняется на основании информации уровня L3, при этом решение о выборе маршрута осуществляется по результатам вычисления кратчайшего пути. TRILL реализует похожую логику, но только не для IP-, а для MAC-адресов. Не удивительно, что «на языке» TRILL поддерживающие эту технологию коммутаторы называются маршрутизирующими мостами, или RBridge.

Для вычисления наилучшего пути до пункта назначения коммутаторы RBridge используют протокол IS-IS, основанный на известном алгоритме Shortest Path First (SPF). Коммутатор, находящийся на входе в облако TRILL, с помощью IS-IS сразу определяет 16-разрядный идентификатор коммутатора на выходе. Каждый последующий коммутатор (транзитный узел) в облаке пересылает трафик на основе этого идентификатора, благодаря чему внутри облака не требуется поддерживать таблицу внешних MAC-адресов. Узлы оперируют очень небольшим объемом адресной информации, что упрощает их задачу, в частности, по распределению трафика по множеству путей. В технологии TRILL вводится такой важный параметр, как «время жизни» – Time To Live (TTL): при прохождении кадром каждого узла в сети TRILL значение этого параметра уменьшается. Этот механизм отсутствует в классической технологии Ethernet, что во многом и является причиной зацикливания трафика – без поля TTL кадр Ethernet может бесконечно долго «путешествовать» по сети, если не достигнет адресата.

В настоящее время несколько производителей при описании своих решений упоминают о технологии TRILL. В частности, Cisco называет свою технологию FabricPath, поддерживаемую устройствами серий Nexus 5000 и 7000, совместимой с TRILL. Однако независимые эксперты отмечают ряд отступлений от стандарта – в частности, другой формат кадра, который используется для передачи трафика между коммутаторами. Но поскольку Cisco активно участвует в продолжающейся стандартизации TRILL, высока вероятность, что фирменные функции со временем станут частью стандартов. Собственно, такое уже многократно происходило при формировании стандартов на другие сетевые технологии.

Литература

1. Open Flow, http://www.openflow.org/, май 2013 г.

2. Juniper, http://newsroom.juniper.net/press releases/juniper-networks-delivers-openflow-application-to--nyse-jnpr-0813497, февраль 2013 г.

3. http://cisco.com, «Cisco Easy Virtual Network At-A-Glance», 2012 Cisco Systems, октябрь 2012 г.

References

1. Open Flow, http://www.openflow.org/, May 2013.

2. Juniper, http://newsroom.juniper.net/press releases/juniper-networks-delivers-openflow-application-to--nyse-jnpr-0813497, February 2013.

3. http://cisco.com, «Cisco Easy Virtual Network At-A-Glance», 2012 Cisco Systems, October 2012.