

Технологии ЦОД: общая информация по коммутаторам QFX

В статье рассмотрены основные принципы построения сетей в центрах обработки данных, приведены примеры аппаратных решений от компании Juniper Networks. Глубокого технического анализа она не несет, но возможно будет полезна тем кто собирается организовать свой первый небольшой центр обработки данных, или присматривается к альтернативным решениям и новым возможностям в уже существующих ЦОД.

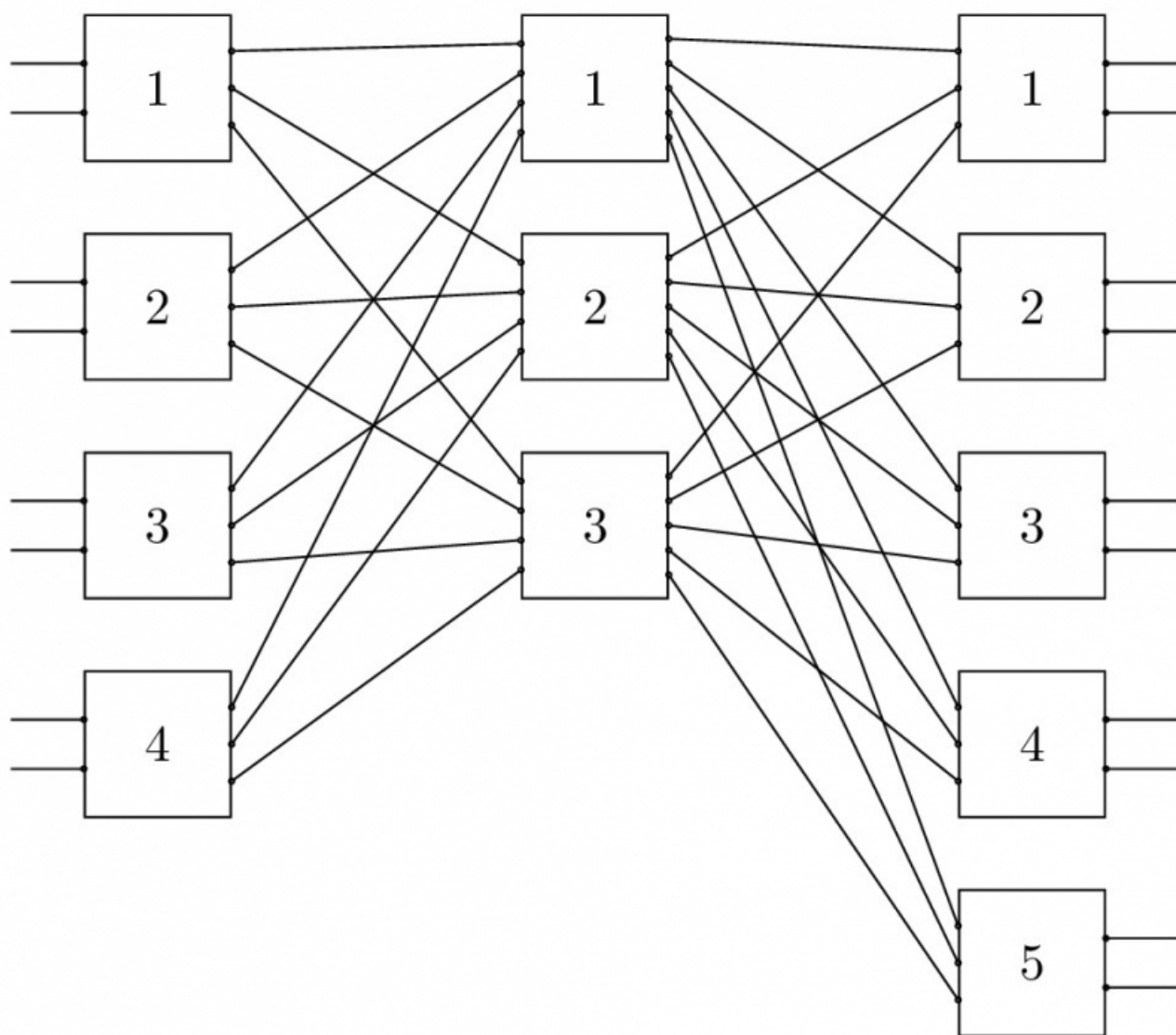
Ни одного коммутатора или ЦОДа в процессе написания статьи не пострадало!

Тема построения сетей связи в центрах обработки данных уже далеко не нова на сегодняшний день. Это направление сейчас находится в стадии бурного развития, что обосновано все возрастающими потребностями по хранению информации, а также массовой миграцией сервисов и технологий в "облака".

В связи с этим, требования к сетям передачи данных в ЦОД повышаются каждый день. Это заставляет человечество разрабатывать новые технологии управления сетью, обработки трафика и работы оборудования. Оптимизировать существующие решения, и конечно снижать нагрузку на человеческие ресурсы, дабы исключить ошибки связанные с человеческим фактором.

Немаловажным стало изменение направления движения трафика в ЦОД - если раньше большие потоки трафика проходили снизу вверх - вертикально, от серверов во внешний мир, то сейчас ситуация совсем иная, и теперь сервера общаются между собой, передавая большие потоки информации друг другу, что изменило пропорции объемов трафика в пользу "горизонтального" обмена.

В 1953 году математик Чарльз Клоз, решал задачу по оптимизации сетевых подключений для применения в телеграфных сетях. Метод подключения "каждый с каждым" был избыточен и требовал больших затрат ресурсов на реализацию. Идея Клоза заключалась в построении сети по схеме трех уровней. Вход - Промежуточный - Выход. В такой схеме количество физических соединений резко уменьшалось, но общая производительность сети была достаточной для работы.



Этот же принцип был взят за основу построения сетей передачи данных в ЦОД. С единственным изменением - уровни "вход" и "выход" были совмещены в один, так как сервера являются как источником информации, так и ее потребителем. В конечном итоге, получили схему подключения сети по принципу Spine - Leaf.

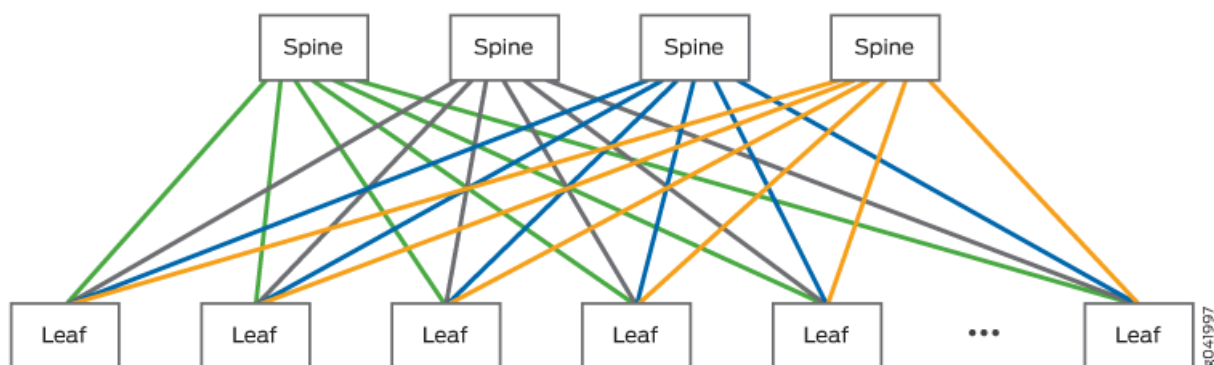


Схема построена по аналогии с растениями в природе. Leaf - Лист. Листья крепятся к стволу. Spine - ствол (или хребет). В итоге, мы получаем одну общую структуру.

Leaf. Оборудование уровня доступа, куда подключаются сервера в центре обработки данных. Это коммутаторы Top of Rack, с большой плотностью портов. Как можно увидеть из рисунка, все оборудование Leaf включается отдельными линиями в каждый коммутатор уровня Spine. Это позволяет достигнуть высокого уровня резервирования и отказоустойчивости.

Spine. Оборудование уровня "агрегации" - обычно производственное оборудование с большой плотностью портов 25/40/50/100 Гбит/с. Отличительной чертой данного оборудования является

широкий спектр поддерживаемого функционала, и большая производительность по сравнению с уровнем Leaf.

В результате, мы получаем схему, в которой потоки трафика проходят через оборудование уровня Spine. Такое распределение потоков трафика может показаться не оптимальным в случае, если оба сервера, источник и получатель подключены к одному коммутатору уровня Leaf. И это правильно, в таком случае трафик проходит напрямую через коммутатор Leaf не поднимаясь на уровень выше.

QFX - серия коммутаторов для построения сетей центров обработки данных.



Аббревиатура QFX не несет никакого логического смысла, поэтому не будем вдаваться в смысл этих трех букв. Данное оборудование разработано и представлено на рынке компанией Juniper Networks, занимающей передовые позиции по разработке и производству телекоммуникационного оборудования.

Вопросами коммутации и маршрутизации в центрах обработки данных в наше время занимается любой уважающий себя производитель телекоммуникационного оборудования, будь то Cisco, Arista, Extreme и некоторые другие.

Интересную позицию занимают производители сетевого программного обеспечения, которое возможно использовать совместно с оборудованием без предустановленного ПО, либо напрямую на серверных платформах. Развитие этого направления происходит в тесном сотрудничестве с производителями телекоммуникационного оборудования, создающих новые продукты и технологии.

Вернемся к оборудованию QFX от компании Juniper Networks. Вся серия была разработана с учетом потребностей центров обработки данных. Программное обеспечение написано с учетом необходимого функционала, архитектура оборудования учитывает требования по отказоустойчивости и резервированию.



Линейка коммутаторов для центров обработки данных представлена следующими сериями:

- Серия QFX51xx - оборудование уровня Leaf.
- Серия QFX52xx - оборудование которое можно использовать как на уровне Leaf, так и на уровне Spine.
- Серия QFX10к - оборудование уровня Spine.

В зависимости от серии, оборудование построено на чипах компании Broadcom, либо на собственных разработках компании Juniper. И тот, и другой вариант имеет свои плюсы и минусы.

Рассмотрим QFX51xx - более старая модель QFX5100, а также обновленный вариант QFX5110. Оба коммутатора построены на основе Broadcom Trident 2+, но разных версий. В итоге, мы получили не одинаковое поведение коммутатора в работе, и, конечно же, различающуюся поддержку программных функций, и портовой емкости. Как было отмечено ранее, данное оборудование позиционируется, в основном, как Leaf-устройство. Возможен вариант использования серии 51xx как Spine-устройства в небольших центрах обработки данных.

QFX51xx поддерживает резервирование блоков питания по схеме AC/AC, DC/DC.

Программно, оборудование строится с применением механизмов виртуализации. В отличие от других серий коммутаторов компании, операционная система Junos OS запускается как виртуальная машина под гипервизором KVM, который работает на устройстве. Благодаря этому, решение поддерживает технологию обновления ПО - [TISSU](#). В процессе обновления создаются дополнительные виртуальные машины Junos OS, и происходит переключение между ними с минимальными перерывами сервиса.

QFX5100 построен на более старой версии чипа Trident 2+ и на мой скромный взгляд уже достоин отправиться на пенсию, уступая дорогу своему старшему брату QFX5110.

QFX5100 выполнен в нескольких модификациях:

- QFX5100-48S - коммутатор 48 портов SFP+ 6 портов QSFP+.
- QFX5100-96S - коммутатор 96 портов SFP+ 8 портов QSFP+.
- QFX5100-24Q - коммутатор 24 порта QSFP+.
- QFX5100-48T - коммутатор 48 портов 10G BASE-T 6 портов QSFP+.

QFX5110 – более новая модель в серии 51xx, построенная на обновленной версии чипа Trident 2+. На мой взгляд, сейчас это лучшее из доступных решений у компании Juniper для работы в качестве Leaf-устройств.



QFX5110 выполнен в нескольких модификациях

- QFX5110-48S - коммутатор 48 портов SFP+ 4 порта QSFP28.
- QFX5110-32Q - коммутатор 32 порта QSFP+ или 20 портов QSFP28.

В целом, можно отметить большую плотность портов 10 Гбит/с для включения оконечных устройств, и довольно "широкие" аплинк-порты 100 Гбит/с, что позволяет получить уровень переподписки на аплинк до "1:1.2" при полной загрузке коммутатора модели 5110. Я считаю, что это хороший показатель.

Небольшая сравнительная таблица:

	QFX5100	QFX5110
Inter VXLAN routing	нет	да
100 G ports	нет	4
Ingress filters	4к	16к
VRF	1к	4к
MPLS labels	16к	32к
Port buffer	12Mб	16Mб
PtP connectors	нет	да

Рассмотрим серию QFX52xx. Оборудование данной серии представлено двумя моделями QFX5200, а также QFX5210. Вся серия построена на чипе Broadcom Tomahawk первого поколения. Это накладывает свои ограничения на функциональность и производительность решения. Серия позиционируется как Spine-устройства небольших центров обработки данных, либо как Leaf-устройства, если необходимо подключать оконечное оборудование 25 Гбит/с. Это обусловлено высокой плотностью портов QSFP28 и тем, что оборудование построено на коммерческом чипе от

компании Broadcom, с ограниченным функционалом.

QFX52xx поддерживает резервирование блоков питания по схеме AC/AC, DC/DC.

Все сказанное выше о работе ПО на оборудовании QFX51xx справедливо и для серии QFX52xx. Операционная система Junos OS работает так же, как виртуальная машина, с возможностью проводить TISSU.

QFX5200 представлен двумя моделями, более старой QFX5200-32C, построенной на чипе Broadcom Tomahawk, и более новой моделью QFX5200-48Y, построенной на чипе Broadcom Tomahawk+:

- QFX5200-32C - коммутатор 32 порта QSFP28.
- QFX5200-48Y - коммутатор 48 портов SFP28 6 портов QSFP28.

QFX5210 представлен одной моделью, являющейся, по сути, двумя коммутаторами QFX5200-32C, что дает нам те же ограничения.

- QFX5210-64C - коммутатор 64 порта QSFP28.

В конечном итоге, мы получаем большую плотность портов 100 Гбит/с за относительно небольшие деньги, что делает данное оборудование хорошим решением уровня Spine в небольших и средних центрах обработки данных. Также, хочется отметить тот момент, что оборудование позволяет осуществить миграцию уровня Leaf на скорости 25 Гбит/с

Перейдем к серии QFX10k. Оборудование данной серии построено на собственной разработке компании Juniper, чипе Q5, с производительностью 500 Гбит/с. Данный чипсет имеет возможность подключения внешней памяти для хранения маршрутной информации RIB/FIB. Это позволяет строить оборудование большой производительности и внедрять новый функционал по мере его появления на рынке через программные обновления. Коммутаторы серии QFX10k разработаны с учетом всех современных потребностей рынка, для использования в больших центрах обработки данных на уровне Spine, а также для построения связей между двумя разнесенными географическими центрами обработки данных - DCI - Data Center Interconnect.

Коммутаторы QFX10k выпускаются в фиксированном и в модульном исполнении.

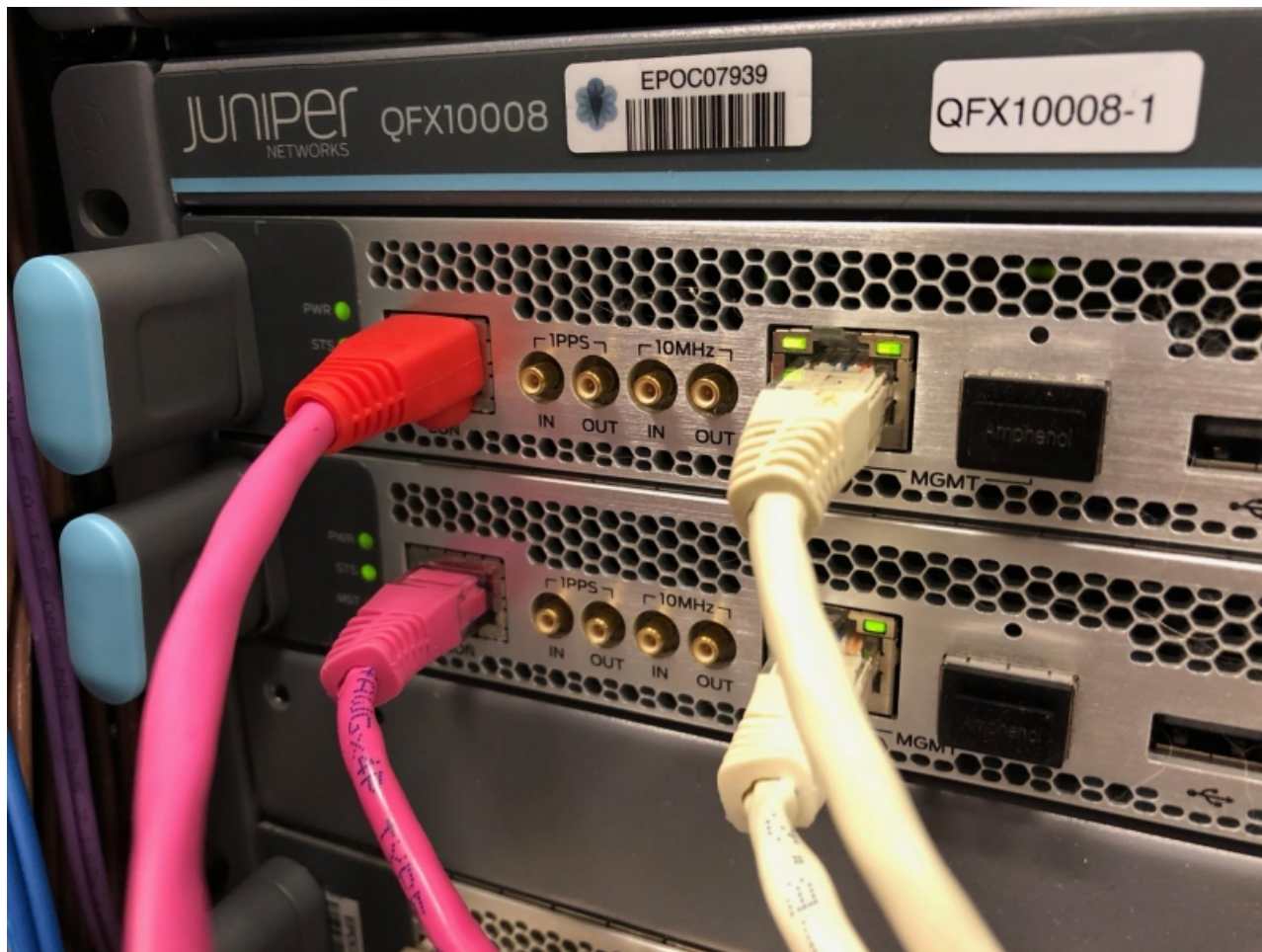


QFX10002 является коммутатором фиксированного исполнения размерами 2RU. Серия включает в себя несколько моделей:

- QFX10002-36/72Q коммутатор с поддержкой 36/72 порта QSFP+ либо 12/24 портов QSFP28.
- QFX10002-60C коммутатор с поддержкой 60 портов QSFP28.

QFX100xx является коммутатором модульного исполнения. Это большие шасси с возможностью подключения различных типов линейных карт, обеспечивающими поддержку разнообразных видов интерфейсов (SFP, SFP+, SFP28, QSFP+, QSFP28). Оборудование позволяет использовать резервирование компонент внутри одного шасси, таких как модули управления, линейные карты.

Серия включает в себя шасси с возможностью включения 8 либо 16 линейных карт.



На сегодняшний день оборудование QFX10k занимает лидирующие позиции в сегменте "Spine-устройства" для сетей центров обработки данных. Коммутаторы составляют сильную конкуренцию оборудованию других производителей, в данном сегменте.

Все оборудование поддерживает механизмы удаленной конфигурации NetConf, сбор статистики и аналитики с применением Junos Telemetry.

В целом, мы провели обзор физической топологии построения сети - в варианте best practice - возможного оборудования от компании Juniper. Хочется отметить, что аналоги таких коммутаторов имеются и у других производителей оборудования. Используя наш обзор, всегда можно выбрать то, что вам больше подходит.

В заключении, предлагаю рассмотреть логические варианты построения сетей связи в центрах обработки данных.

На сегодняшний день я могу выделить два варианта логического построения сети:

1. Основанный на закрытых алгоритмах и протоколах работы.
2. Основанный на открытых алгоритмах и протоколах работы.

Мы по-прежнему рассматриваем производителя Juniper Networks.

В первом случае можно выделить два основных решения для построения сетей в центрах обработки данных :

1. Virtual chassis fabric
2. Junos Fusion Datacenter

Virtual Chassis Fabric. Решение является развитием технологии Virtual Chassis от компании Juniper. Используется тот же принцип объединения коммутаторов в одно общее шасси с помощью внутреннего протокола VCCP. VCCP основан на открытом протоколе IS-IS, и является его переработанной версией. В результате, механизм работы решения скрыт от конечного пользователя, и отдается на откуп разработчикам, что затрудняет решение любых проблем, возникающих во время эксплуатации, но позволяет пользоваться завершенным продуктом.

- Единая точка управления через главный свич.
- Поддержка архитектуры Spine - Leaf.
- Максимальное число устройств - 20 штук (4 Spine и 16 Leaf).
- Максимальное масштабирование до 1536 портов доступа с поддержкой 10 Гбит/с - FCoE в одной фабрике.
- Система управления Junos Space.

Junos Fusion DataCenter. Решение является новым продуктом компании, разработано специально для применения в центрах обработки данных. Поддерживаются только коммутаторы серии QFX - 51xx/10k. Решение основано на технологиях 802.1BR, netconf, json-rpc. Единая логическая структура с возможностью управления из одной консоли. Решение активно развивается, и вносится дополнительный функционал.

- Основано на IEEE802.1BR.
- Единая точка управления.
- Поддержка технологии Spine - Leaf.
- Максимальное число устройств 65 штук.
- Максимальное масштабирование до 6144 портов 10 Гбит/с -FCoE в одной фабрике.
- Возможность использования 1 либо 2 устройств на уровне Spine.
- Уровень Spine представлен коммутаторами QFX10k.
- Уровень Leaf представлен коммутаторами QFX51xx.
- Поддержка локальной коммутации на уровне Leaf.
- Возможность автонастройки vlan на портах уровня Leaf.
- Система управления Junos Space.
- Поддержка MC-LAG.

В случае, когда сеть строится на открытых алгоритмах и протоколах работы - решение только одно - **IP Fabric.** Решение позволяет использовать оборудование любых производителей. IP fabric работает исключительно на третьем уровне OSI. Не имеет каких-то общих точек управления и единой консоли, но возможности масштабирования, а также траблшутинга, довольно обширны.

Разделение на уровни Underlay и Overlay. В качестве технологии Underlay используется известные нам протоколы IGP - такие, как OSPF/IS-IS , для небольших центров обработки данных. Если центр обработки данных имеет множество стоек и используется большое количество подключенных оконечных устройств - то всеми любимый iBGP/eBGP. На мой взгляд, тут при выборе решение underlay - это дело вкуса, и здравого смысла.

В качестве решения overlay выбирается что нибудь вроде VXLAN, который поддерживается большинством производителей сетевых решений, совместно с технологиями MPLS или, допустим, EVPN. Оборудование компании Juniper может быть использовано для построения подобных решений наряду с другими производителями. Никаких особенностей тут выделить, я думаю, нельзя.

- Широкий выбор вариантов развертывания.
- Поддержка технологии Spine - Leaf.
- Открытый выбор технологий и протоколов.
- Возможность использования решений от различных производителей.
- Самая высокая масштабируемость по сравнению с другими решениями.
- Работа только на третьем уровне модели OSI.

В данной статье я провел общий обзор принципов построения сетей для центров обработки данных, а также привел возможности использования оборудования и решений компании Juniper Networks для их реализации.

Все фотоматериалы любезно предоставлены компанией Juniper Networks в лице инженера Бугакова Евгения. За что ему отдельное спасибо.