Сетевая инфраструктура ЦОД для самых маленьких (с)

GetNet Конференция для IT-специалистов

Обо мне

@LIKE_A_BUS_CHANNEL

- Инженер, ~~CCIE #65101~~
- Последние 6 лет по большей части занимаюсь инфраструктурой центров обработки данных
- За период с 2018 по 2025 год, совместно с коллегами, построил и сдал в эксплуатацию более 20 фабрик на основе VXLAN-EVPN разных масштабов
- Преподаватель на онлайн платформах и наставник во внутренней стажёрской программе

Целевая аудитория

- Сетевые инженеры, архитекторы и их руководители, которые хотят освежить или получить базу по сетевой инфраструктуре ЦОД
- Смежные профессии, которые хотят поглубже понять структуру ЦОД с точки зрения сетевого инженера
- Если вы уже управляете ЦОДами на сотни или тысячи стоек, это доклад вряд ли для вас, уверен вы и сами можете многое мне рассказать

О чем доклад

- Сформируем задачу
- Определим топологию
- Выберем оборудование
- Проанализируем стек технологий
- Рассмотрим первичную автоматизацию
- Бонусом будет всё остальное

Задача

- 10 стоек по 7 КВт и в каждой максимум 12 серверов (2 порта по 10 Gbps)
- Обязательно наличие резервирования на уровне сетевой связности
- Должна быть возможность разделять потоки трафика на зоны безопасности
- На текущий момент инфраструктура ограничена одним сайтом, но есть потенциал расширения

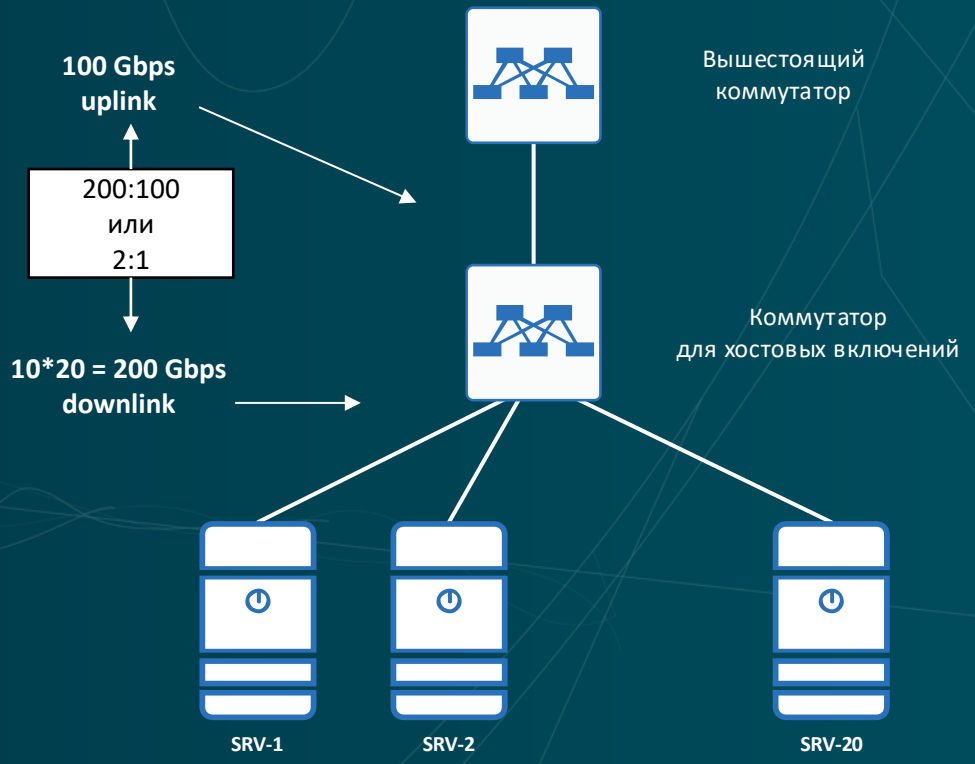
Задача

- Итого выходит:
  - Максимально возможное количество серверов – 10 стоек * 12 штук = 120 штук, или суммарно 240 портов по 10 Gbps.
  - Сети управления нет, её тоже нужно учесть. В среднем 1 хост – 1 порт. Всего 120 серверов + N сетевых устройств.
  - Необходимы коммутаторы:
    - для «основного трафика» с портами 10G как downlink и 100G как uplink
    - для «трафика управления» с портами 1G как downlink и 10G как uplink
  - Потенциально мы можем вырасти, стоит брать с запасом.

Задача

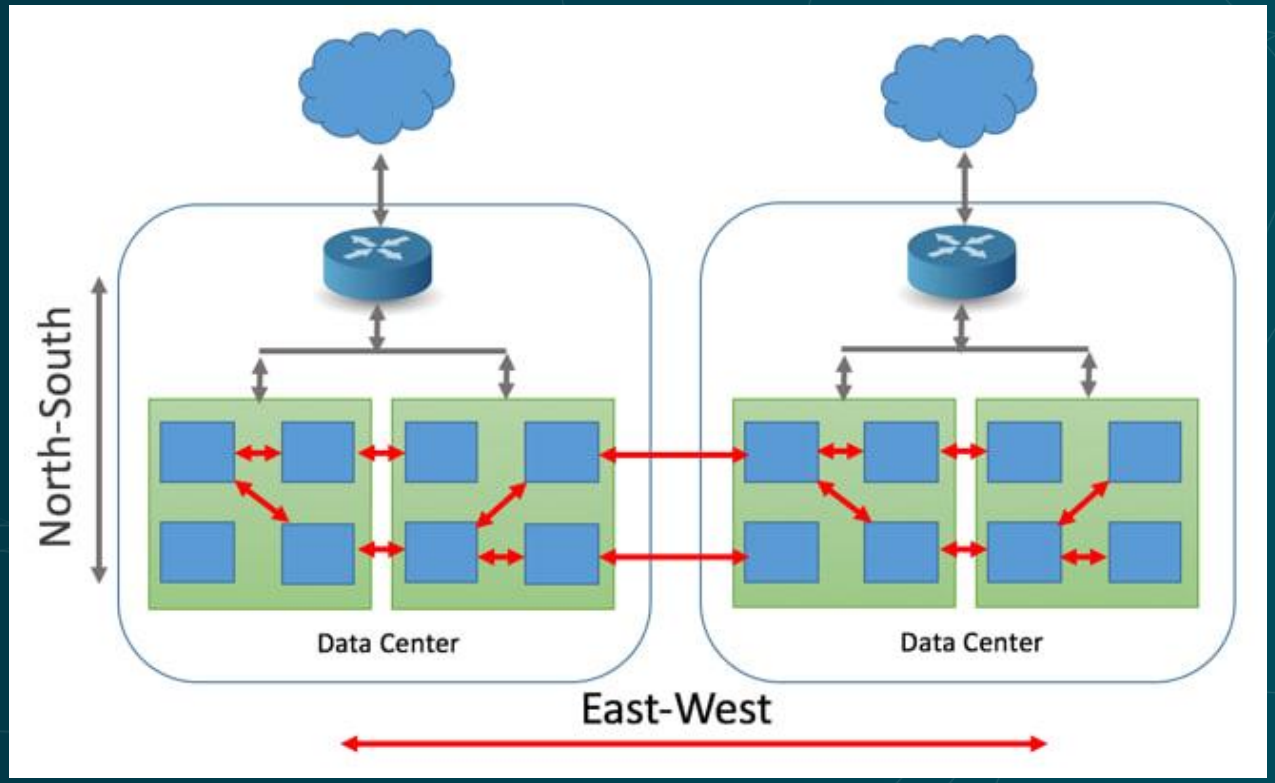
- OverSubscription или переподписка - соотношение суммарной емкости «хостовых» портов к суммарной емкости «сетевых» портов.

- На что опираться при расчете?
  - Чем больше east-west, тем больше требования
  - Для «стандартного корп ЦОДа» 4:1
  - Для управления не считаем



Направление трафика

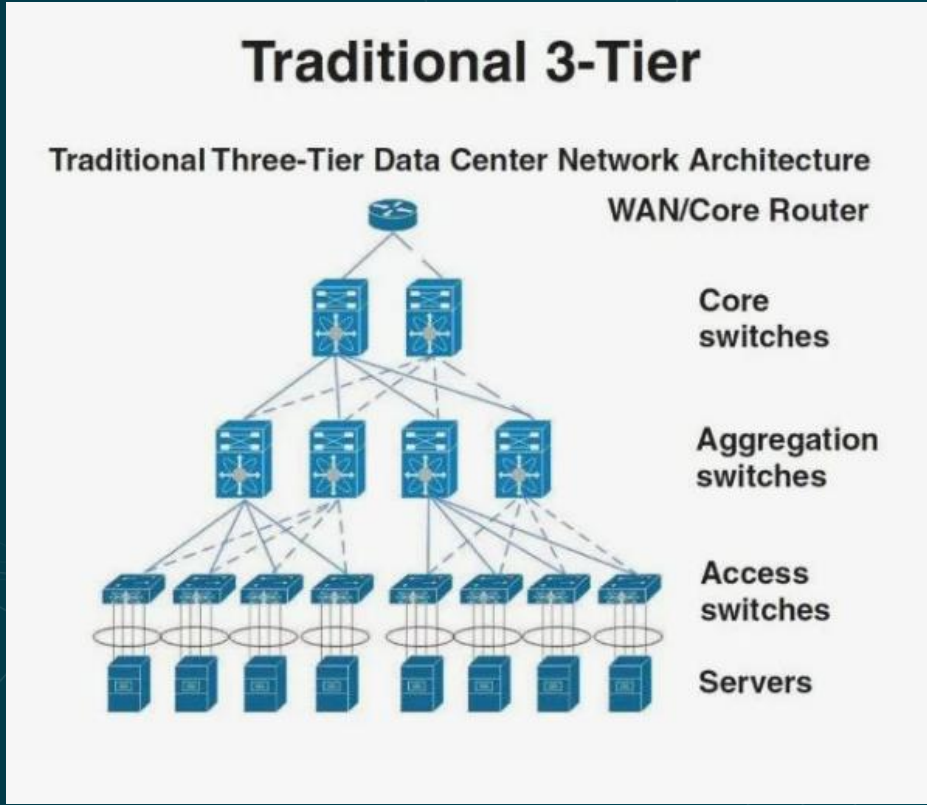
- North-South
- East-West



Топология

- Трёхуровневая, классическая топология

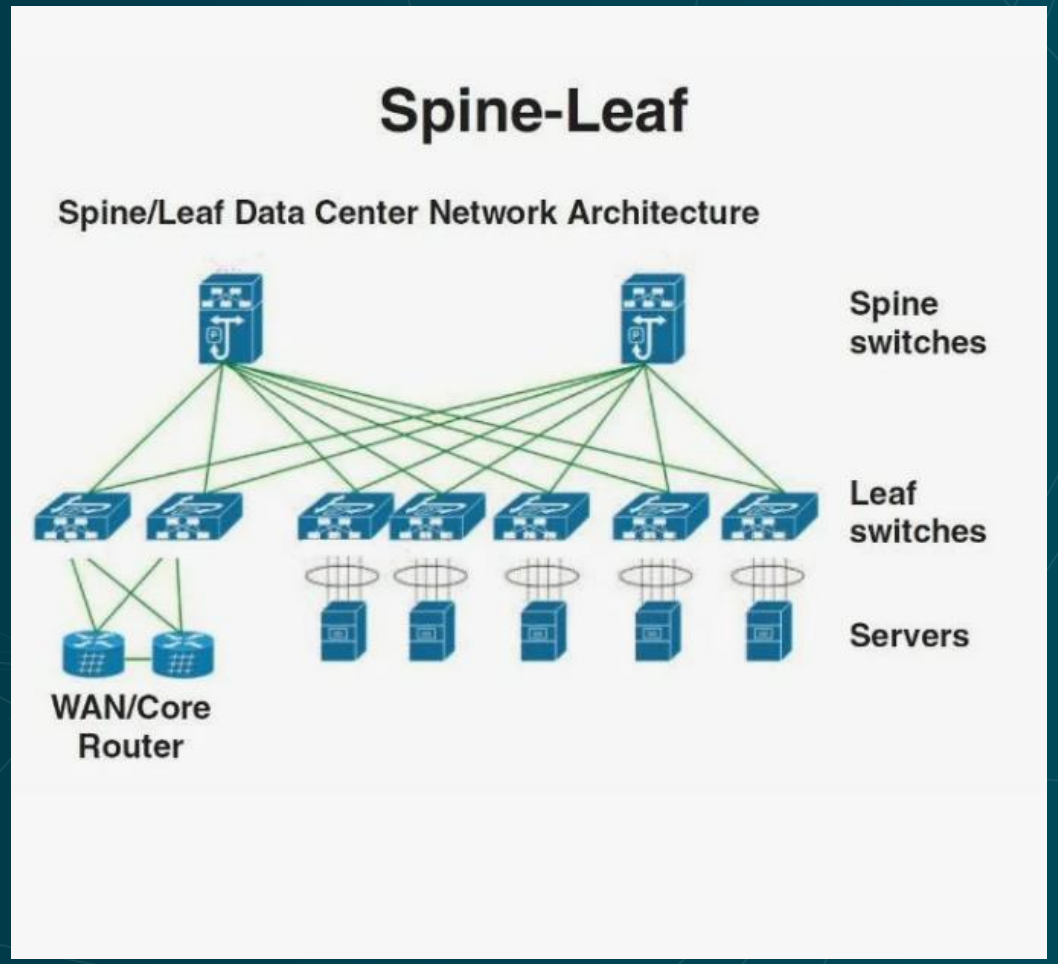
- Особенности:
  - Наличие Core уровня
  - Адаптирована для North-South
  - Отлично подходит для OOB



Топология

- CLOS / Spine-Leaf / Fat-tree

- Особенности:
  - Все линки зарезервированы
  - Адаптирована для East-West
  - Отлично подходит для Data трафика
  - Дорого



Топология

- Что еще сейчас можно встретить:
  - Dragonfly+
  - Jellyfish
  - DRing
  - Что-то безумное

Оборудование

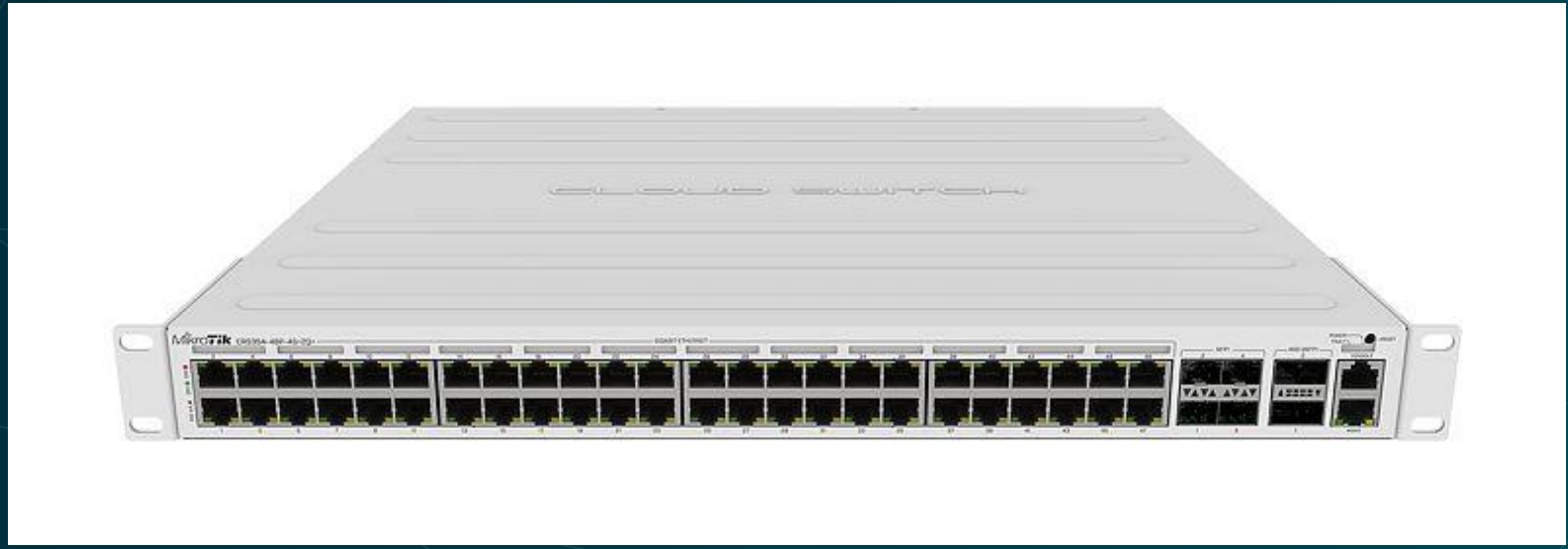
- Модульный коммутатор
  - Не ваш друг
  - Неприятный в эксплуатации

- «Пиццабокс»
  - Ваш друг
  - Отлично встраивается в CLOS

Оборудование

@LIKE_A_BUS_CHANNEL

- Коммутаторы уровня Aggregation и Access
  - Нужно было 120 серверов + сетевое оборудование = 48 * 4 = 192 порта
  - 4 шт, порты 48x1G и 4x10G SFP+



Оборудование OOB network

- Коммутаторы уровня Aggregation и Access
  - 4 шт, порты 48x1G и 4x10G SFP+
- Маршрутизатор умеющий в IPSec
  - 1 шт, порты 4x1G и 2x10 SFP+

Оборудование OOB network

- Коммутаторы уровня Aggregation и Access
  - 4 шт, порты 48x1G и 4x10G SFP+
- Маршрутизатор умеющий в IPSec
  - 1 шт, порты 4x1G и 2x10 SFP+
- Терминальный сервер для консольного доступа
  - 1 шт, порты 24x1G

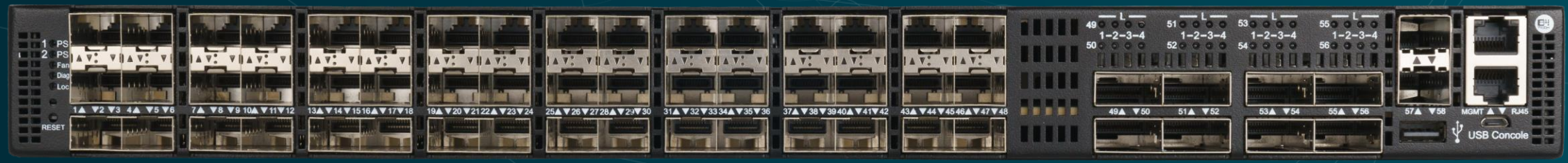
Оборудование data network

- Коммутаторы уровня Leaf
  - Служит точкой включения хостов. Порты должны быть и 100G и 10G, функционал богаче по сравнению со Spine
  - Нужно было 240 портов 10G в сторону хостов => 240 портов / 10 коммутаторов = 24 порта на каждом устройстве и они уже заняты
  - 2 коммутатора будем использовать как border leaf, для всех внешних подключений

Оборудование data network

@LIKE_A_BUS_CHANNEL

- Коммутаторы уровня Leaf
  - Служит точкой включения хостов. Порты должны быть и 100 и 10, функционал богаче по сравнению со Spine
  - Нужно было 240 портов 10G в сторону хостов => 240 портов / 10 коммутаторов = 24 порта на каждом устройстве и они уже заняты
  - 12 шт, порты 48x10G SFP28 и 8x100G QSFP28 (+2 берём из-за пары border leaf)
  - Можно вырасти x2, и переподписка в максимальной набивке составит 48*10/2*100 = 2.4:1



Оборудование data network

@LIKE_A_BUS_CHANNEL

- Коммутаторы уровня Spine
  - Простая «молотилка» трафика. Порты пожирнее, функционал победнее
  - Нужно было 10 коммутаторов, с каждого по 1x100G подключению, т.е. 10 портов заняты сразу
  - 2 шт, порты 32x100G QSFP28
  - Можно вырасти ещё на 20 лифов

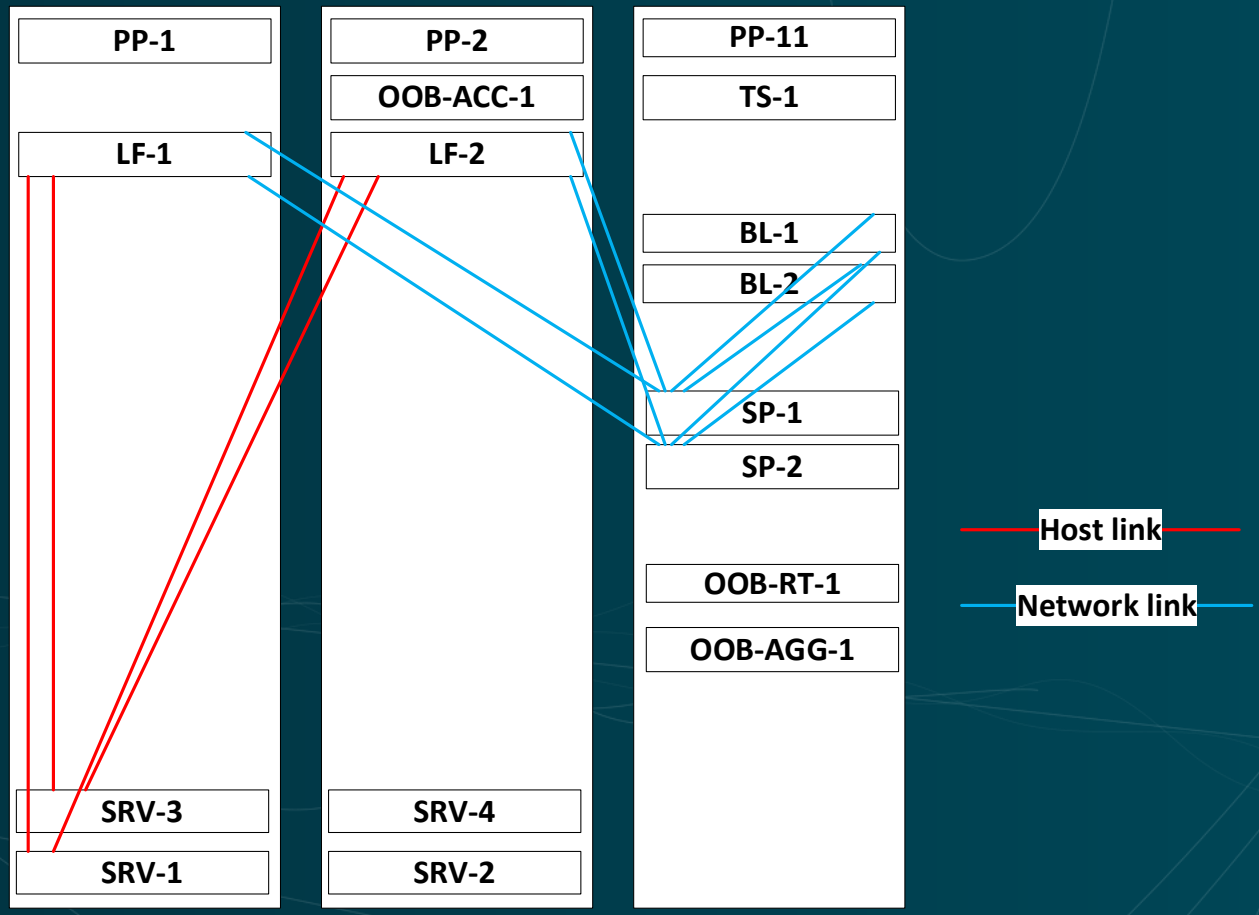
Оборудование

- На что ещё обратить внимание?
  - Схему вентиляции (Front-to-Back или Back-to-Front)
  - Резервирование блоков питания
  - Наличие устройства в ТОРП...

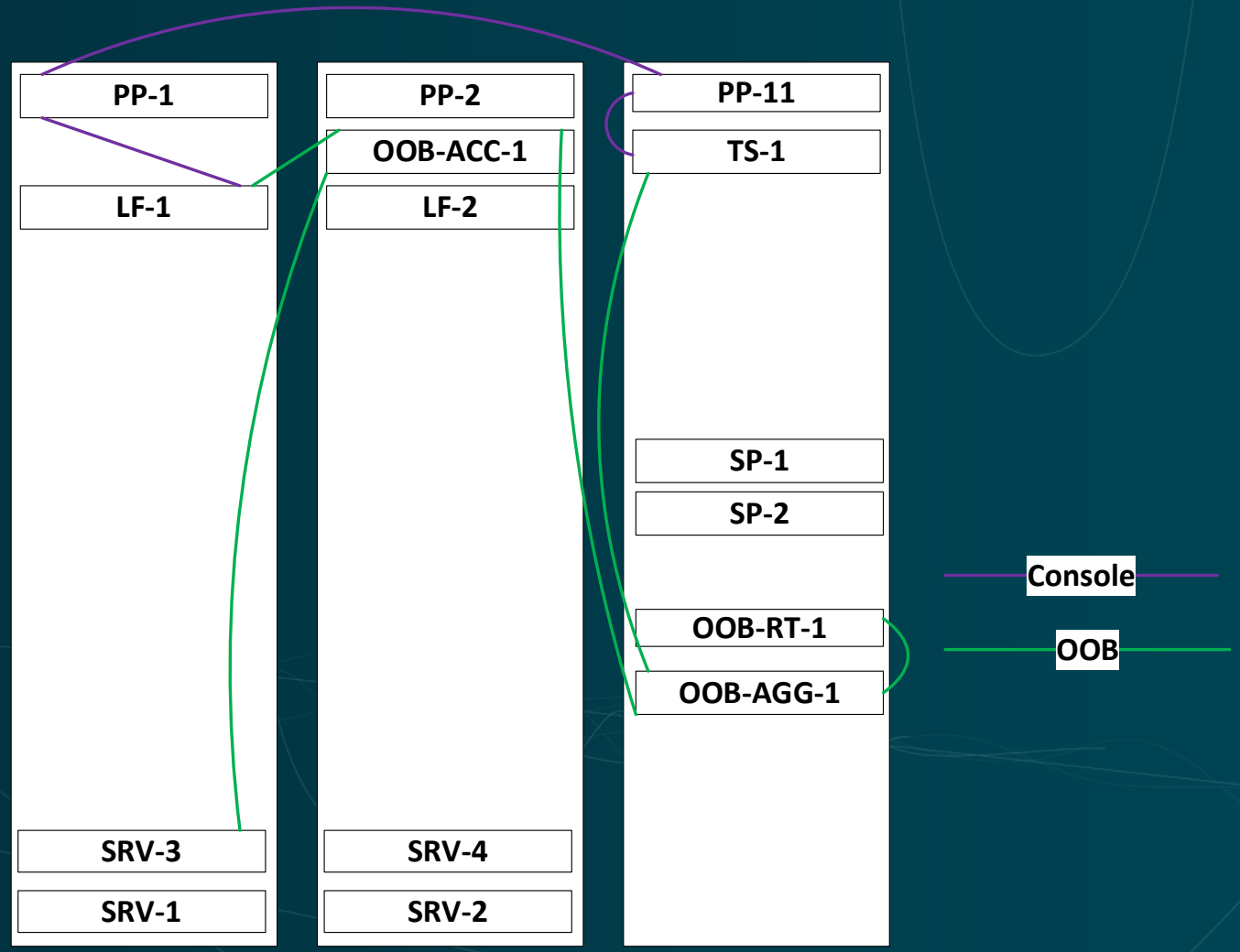
Оборудование

- Как сэкономить?
  - Проанализировать лицензии, потенциально могут быть платными те, которые использовать не будут
  - Уровни поддержки. Обычно вендор имеет несколько, выберите удобный вам, необязательно максимальный
  - Есть примеры вендоров, которые заранее стоят дороже, но готовы реактивно исправлять баги, откликаться на feature request и т.д

Промежуточный итог



Промежуточный итог



Промежуточный итог

- Управление:
  - Коммутаторы уровня Access - 4 шт, порты 48x1G и 4x10G SFP+
  - Маршрутизатор - 1 шт, порты 4x1G и 2x10 SFP+
  - Терминальный сервер - 1 шт, порты 24x1G

- Основное направление:
  - Коммутаторы уровня Leaf - 12 шт, порты 48x10G SFP28 и 8x100G QSFP28
  - Коммутаторы уровня Spine - 2 шт, порты 32x100G QSFP28

Технологический стек

- STP + MLAG
  - Заблокирована половина линков
  - Нет возможности разделить потоки трафика по зонам безопасности
  - Ограничение в 4000 подсетей (вланов)

  - Работает, но...

Технологический стек

- EVPN-VXLAN
  - Разделение на плоскости (underlay + overlay)
  - Доступны все линки для трафика
  - Можно сделать много сетей и все разделить на зоны безопасности
  - Сложно в управлении, BGP + L2/L3VPN = must have

  - Нормально, рабочая схема.

Технологический стек

- EVPN-VXLAN
  - Разделение на плоскости (underlay + overlay)
  - Доступны все линки для трафика
  - Можно сделать много сетей и все разделить на зоны безопасности
  - Сложно в управлении, BGP + L2/L3VPN = must have

  - Нормально, рабочая схема.

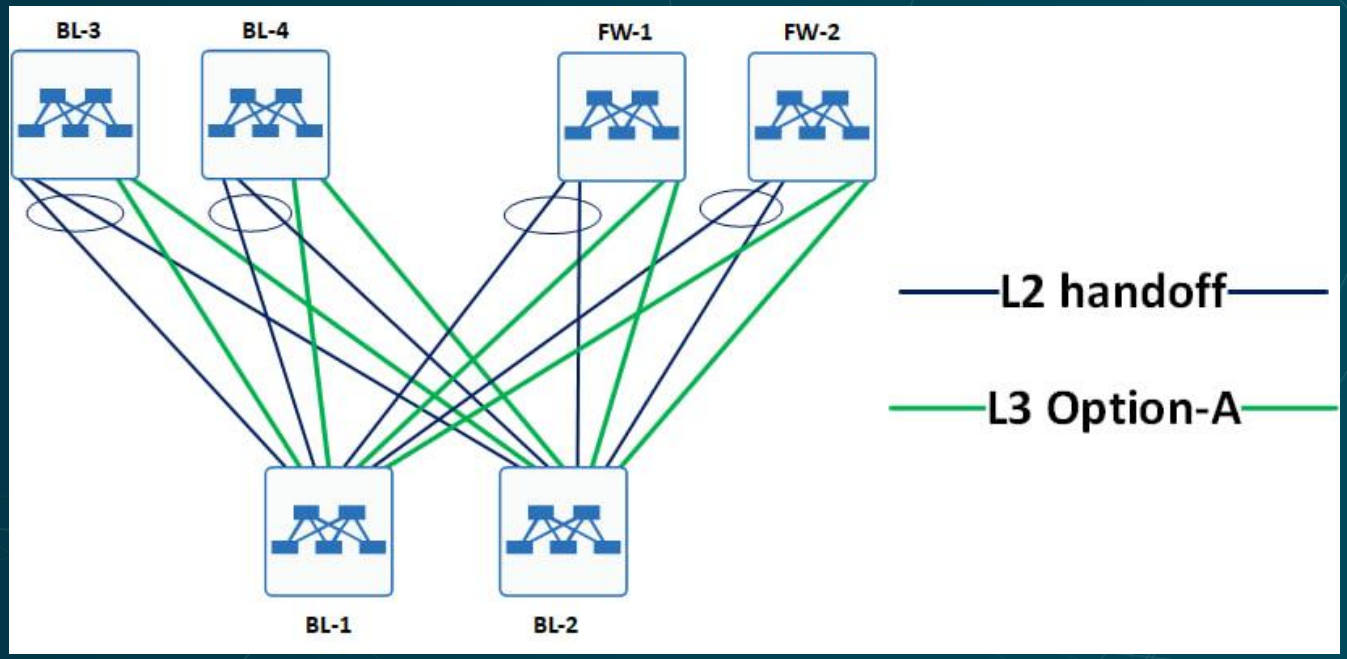
Технологический стек

- SRv6
  - Большинство плюсов EVPN-VXLAN
  - По-настоящему раскрывается при наличии контроллера
  - Требуется поддержка коммутаторами

  - Пока что рано, потенциал огромен

Организация внешних подключений в EVPN-VXLAN

- Через border-leaf
- Проприетарное решение:
  - Multisite
  - Segment VxLAN
- Независимое от вендора:
  - L2 hand-off
  - L3 Option-a (VRF-Lite)



Обязательные службы

- SNMP
- NTP
- Syslog
- AAA (TACACS+ / RADIUS)

Без любого из них ваша инфраструктура в очень уязвимом состоянии

Первичная автоматизация

@LIKE_A_BUS_CHANNEL

- Не бойтесь писать скрипты - они на вашей стороне
- ChatGPT - пишет хорошо, но Python учить стоит
- Минимум то из автоматизации, что мы регулярно используем:
  - Проверка коммутации
  - Создание конфигураций
  - Плагины в Netbox
  - Сбор специфичных данных с узлов и их парсинг

Важные инструменты

- Netbox
  - Менеджмент IP адресов, инфраструктуры ЦОДа (стойки, юниты и т.д.) и описание логической структуры сети (VRF, VLAN и т.д.)
- Oxidized
  - Резервирование конфигураций
- Git
  - Скрипты
  - Примеры шаблонов конфигураций

только не храните внутри этого же ДЦ, пожалуйста...

Различные нюансы

- Все работы проводите с умом и тщательно планируя
- Старайтесь проводить работы днём
- Запланируйте регулярные проверки отказоустойчивости
- Сформируйте план восстановления, он же DR

...мимолётный ветер шепчет «СДЭЭЭЭК»...

В заключении мы получили

- Инфраструктуру на 10 стоек, для 120 серверов, где:
  - Data network - CLOS + EVPN-VXLAN
  - OOB network - 3-tier + VLAN
  - Различные службы (NTP, SNMP, Syslog и т.д.)
  - Набор инструментов для эксплуатации
  - Скрипты для автоматизации рутины
  - План работ, план DR, регулярные проверки отказоустойчивости

Полезные материалы

- Топологии

  https://linkmeup.ru/blog/1262/

  https://habr.com/ru/companies/yandex/articles/859794/

  https://link.springer.com/book/10.1007/978-981-16-9368-7

- Стек технологий

  https://habr.com/ru/articles/352564/

  https://linkmeup.gitbook.io/sdsm/12.1.-mpls-evpn

  https://www.cisco.com/c/dam/en/us/td/docs/switches/datacenter/nexus9000/sw/vxlan_evpn/VXLAN_EVPN.pdf

  https://www.youtube.com/watch?v=KV_He_Xnaa8

- Автоматизация

  https://pyneng.readthedocs.io/ru/latest/

  https://pynet.twb-tech.com/

  https://github.com/jeremyschulman

  https://github.com/manwithwine

- Полезное

  https://blog.cortel.cloud/2022/11/30/avarijnoe-vosstanovlenie-vsyo-o-disaster-recovery-za-15-minut/

  https://habr.com/ru/companies/first/articles/711004/

Полезные инструменты

- **IPAM + DCNM + CMDB**

  Netbox - https://netboxlabs.com/

- **SNMP**

  Zabbix - https://www.zabbix.com/

- **Syslog**

  Graylog - https://graylog.org/

- **Backup**

  Oxidized - https://github.com/ytti/oxidized

- **Git**

  GitLab - https://about.gitlab.com/

- **AAA**

  FreeRADIUS - https://www.freeradius.org/

- **NTP**

  chrony - https://chrony-project.org/

Спасибо за внимание!

Буду рад ответить на все ваши вопросы сейчас
или свяжитесь со мной в будущем:

Сергей Бочарников

sergey.bocharnikov.v@gmail.com

https://t.me/like_a_bus_channel