Ucloud

SD-WAN 技术方案 (初稿)

2018年5月4日

版本 v0.0.1

陈煌栋

1. SD-WAN 技术方案

1.0 概述

随着企业 IT 向云架构转型的不断推进、AWS 等公有云的崛起和流行， 引导着企业数据中心 （DC）等基础设施云化，越来越多的企业开始在公有云安家，从而打破企业 IT 的传统封闭架构， 引领企业网络架构走向开放之路；与此同时，企业的关键应用也逐渐云化，依赖于应用服务商提 供的 SaaS 服务（如 office 365、生产 ERP 系统、销售 Saleforce 等），企业通过互联网从云端 访问日常办公所需关键应用的趋势日渐明显。

云化不仅仅是一场技术的变革， 也是商业模式的变革。 但是企业分支的业务向云端迁移， 面临的挑战依然不少：

- **传统专线费用贵**：WAN 流量激增，成本合理的全场景连接成为基本需求

  由于业务云化，企业需频繁与云中心互动，这就需要更高的网络带宽。为了保 证服务质量，企业 WAN 网络互联通常采用运营商 MPLS 专线。虽然专线网络质 量有保障，但价格过于昂贵，平均占企业 OPEX 达 50% 以上。

- **应用体验难保障**：海量应用带宽共享，业务冲突导致体验不佳

  随着 Internet 的普及，其网络的覆盖范围和网络质量有了很大的提高，Internet 成为许多企业除了传统专线之外新的重要选择，但是 Internet 网络本身并不保 障服务质量。此外传统网络对业务不感知，无法获知应用的状态，当遭遇突发 流量链路拥塞或质量劣化的时候，往往会造成关键业务体验无法保障。

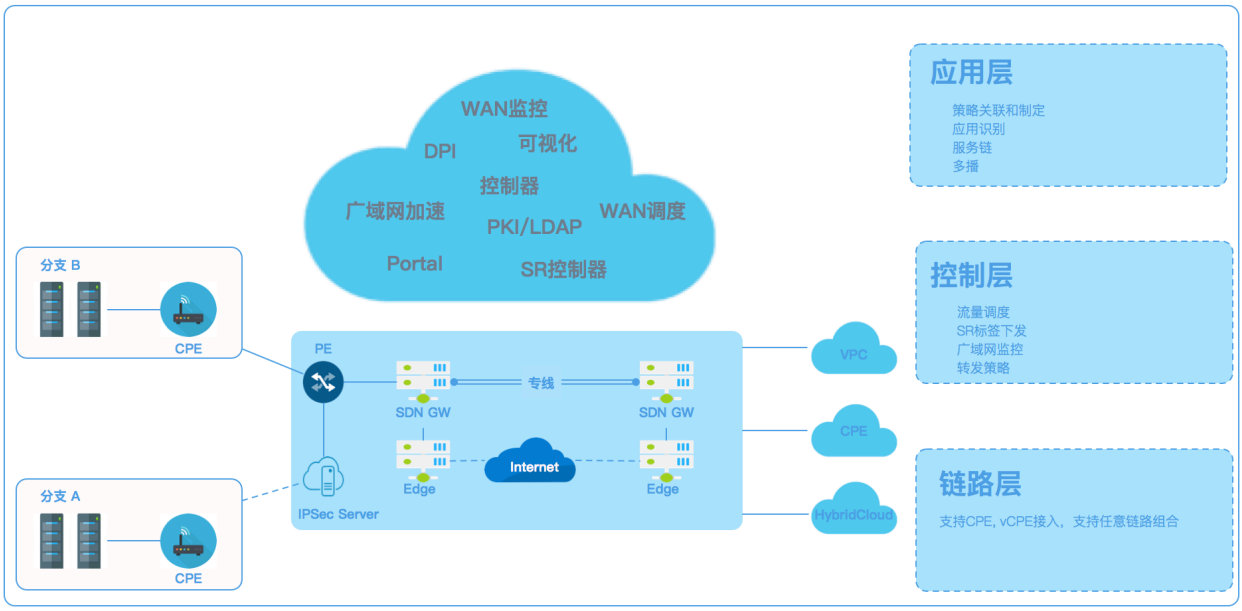
- **业务上线周期长**：传统方式难以满足业务灵活部署的诉求

  传统的专线新业务发放速度慢，需要经历营业厅申请、业务调试、现场配置等 多个环节。从业务申请到开通往往需要长达 1 ～ 3 个月的时间。云化趋势下， 企业业务更新发展迅速，当前网络难以满足快速上线要求。

- **网络运维难度大**：业务流量不可视，运维效率低下

  传统专线需要专人到现场对设备进行维护，但随着企业分支跨地域分布越来越 广泛，数量激增，导致维护难度大、成本高。此外随着业务不断增多和业务云化， WAN 网络中分支到分支、公有云、私有云的流向更加复杂，传统的网络运维方 式已经难以适应业务的发展。

1.1 整体架构



用户可在创建主机时或VPC页面下创建虚拟网卡。虚拟网卡具有如下属性：object_id，VPC，子网，关联ip和mac，名称，备注，是否是默认网卡等属性。

按照功能划分，主要分为链路层、控制层和应用层，其中：

- 链路层：负责CPE/vCPE 的接入，IPSec VPN隧道的建立，MTU的探测与植入，CPE作为站点边缘的Internet接入等功能；
- 控制层：负责CPE的ZTP接入、配置下发、CPE认证，负责秘钥和路由推送，负责应用识别和广域网链路质量监控，负责SR标签计算和流量调度，负责广域网优化和加速等；
- 应用层：负责策略、标签的制定和关联，负责App-based的转发策略制定，负责Service Chain的制定等；

1.2 CPE

1.2.1 CPE 选型

CPE 支持硬件、软件和SDK等：

- 对于硬件，普通的、轻量级的分支-总部互联场景，可以选择基于 ARM 的CPE 平台；
- 对于流量、性能要求较高的 IDC 接入，可以选择 x86 + DPDK 的CPE 平台；
- 对于友商公有云等场景，可以选择软件形式的 VNF CPE，提供 docker、vm 镜像等方式；

1.2.1.1 x86 平台

对于 IDC 等流量较大的场景，基于x86 + DPDK 提供相对灵活+高性能的CPE平台。

端口上，提供千兆或万兆互联RJ-45和SFP光纤接口，WAN口提供1-2个，LAN口提供2-8个。WAN口侧可选提供T1/E1等传输接口。

对于提升IPSec性能，可集成 Crypto Accelerator 等ASIC芯片用于IPSec加速；对于CPE的身份认证可集成 TPM 芯片。

CPE 可选提供 Qugga 等路由软件，用于向站点内部交换机发布远端路由等，可选运行PIM协议，通告控制器下发的组播路由等。

1.2.1.2 ARM 平台

对于通常情况下的分支机构，可提供基于 ARM 平台的CPE方案。ARM-based CPE可基于 Openwrt 开发，Openwrt是一个支持 ARM、MIPS、x86 等多种硬件架构的linux发行版，提供完善的 router 开发框架。

CPE在分支机构作为外网出口，需提供出口路由器的功能，如 PPPoE、NAT、Firewall 等功能，这部分 Openwrt 里已经有成熟的实现。

1.2.1.3 VNF

对于公有云等场景，需要提供基于 VM镜像、Docker 等封装的 vCPE。

1.2.1.4 SDK

对于移动应用，提供SD-WAN接入的SDK。

1.2.2 CPE 接入

CPE Wan接入可以选择基于物理专线连接到 PE，或者基于 IPSec VPN Over Internet 到最近 PE的接入。

其中，基于专线可提供稳定、可靠的接入能力，提供SLA保证，基于Internet的接入，会受到运营商线路质量抖动的干扰，无法提供SLA保证。

CPE 到 SP 的接入，支持 Dual-Wan 等 Multi-Home 方式。在 CPE 端需要监控每条线路的链路质量，当某条线路 QoS 出现下降时，可以在多条 WAN 线路之间切换。

对于有 BGP 接入的 IDC 场景等，可以将 IPSec VPN 的隧道地址配置在 Loopback 口，并向 SP 发布路由，在多条 BGP 线路间做到高可用。

其中，接入多条 SP 线路时，每个 SP 会通告一条默认路由，转发时可选支持按权重转发、互相热备等。

部分 CPE 接入，可能无公网IP、固定IP等。GRE、VxLan 等都无法穿越 NAT，IPSec 通过借助 NAT-T（(NAT Traversal，将ESP协议封装进 UDP 中）可支持穿越 NAT。

1.2.3 CPE ZTP 部署

理想情况下 CPE 是 ZTP (零接触) 部署的。CPE 可通过唯一的序列号（如SN）等确定唯一身份，客户收到 CPE 并家电后，CPE 主动寻找 Staging Server，完成自身的注册。客户通过输入关键信息，完成 CPE 和 Staging Server 的双向认证，认证成功后从控制器获取 IP、密钥、路由等配置信息，并建立 IPSec VPN 隧道。

ARM based CPE 提供 U-BOOT 和 Web Console 提供 Firmware Update。对于配置更新通过 Staging Server 动态下发，对于程序、Patch等可提供在线更新方式。

1.3 流量调度

1.3.1 Underlay 调度

Underlay 对于 QoS 的支持 可选 RSVP/TE 或者 Segment Routing，优选 Segment Routing。Segment Routing对SDN的支持性较好，支持源路由转发，以标签交换为基础，通过控制面SR控制器决定标签转发路径，SR交换机根据标签转发。提供良好的 TE 流量工程支持和 FRR 快速重路由。

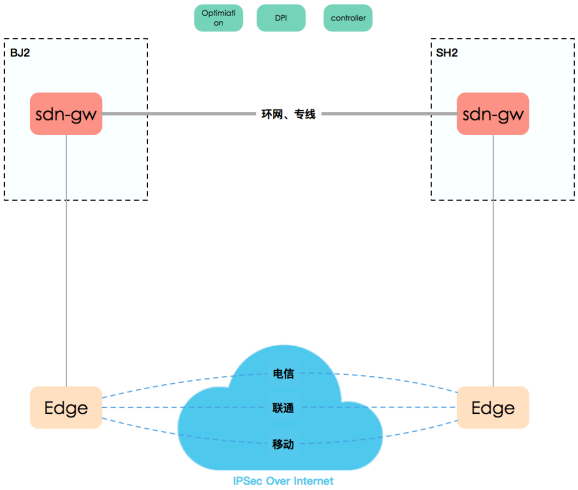
SR 的调度流程：

- BGP-LS 收集网络节点、拓扑、链路状态和带宽、时延、丢包等信息传递给 SDN Controller；
- SDN Controller 基于应用要求的带宽、时延、保护等策略，完成路径的计算；
- Controller 完成选路后，将选路信息下发给 SR 路由器；
- SR 路由器按照标签栈进行转发；

通过 Underlay 的 SR 提供底层的策略转发支持。

1.3.2 WAN 调度

WAN 调度的关键技术在于广域网质量监测和应用识别。SD-WAN 提供的关键能力中，是 Application Based 的转发能力。



SDN-GW 收到数据包后，对一条流的前几个数据包上送 Controller ， Controller 通过送给 DPI，完成应用识别。当 DPI 根据流识别出应用后，下发明细 Flow （最粗基于 5-Tuple，并应指定 idle timeout)，后续该流的流量不再经过 Controller 和 DPI。

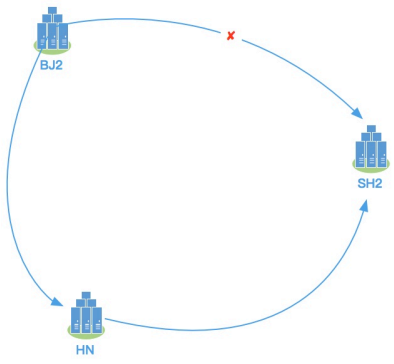
控制器根据流量类型，和当前的骨干网和 SP Internet 线路情况（利用率、时延、丢包、抖动等信息），选择转发线路：专线（UDPN）或者是 Internet。

若选路 UDPN，则根据应用关联的策略信息，给数据包插上指定的Tag信息（DSCP等），发送给 Wan Optimation Server。若基于 Internet 转发，则封装 IPSec VPN 后发送给 Wan Optimation Server。经过 Wan Optimation Server，流量交由底层转发。

WAN Edge 应提供多条运营商 WAN 线路用于备份和流量调度。IPSec VPN 应基于主流 SP 如电信、联通和移动建立 LoopBack 隧道。通过广域网流量监控，监控运营商线路的丢包、时延、抖动、利用率等，动态调整转发线路，当某条线路 SLA 严重下降或不可用时，完全切换到其余线路。探测手段如 ICMP/OWAMP/TWAMP等，对于 SaaS 服务，可基于 HTTP Ping 做监控。

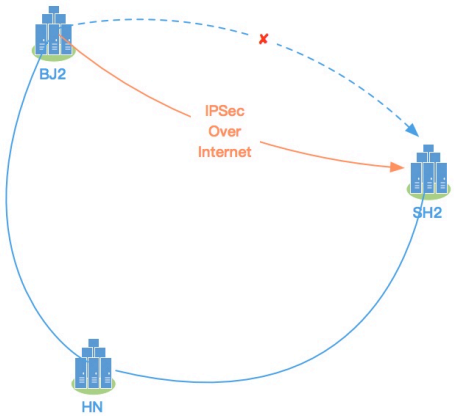
1.3.3 SDN GW流量调度

都某条环网不可用时，通过评估延迟、利用率等信息，可绕行环网，牺牲延迟提供内网互通能力。改切换方式主要依靠 Underlay 的路由切换能力。



环网互切

也可切换备份 Internet 线路。主要依靠 SDN GW 上的切换能力，并在多个 SP 之间选择最优 WAN 线路。



Internet备份链路互切

1.4 应用识别

应用识别基于DPI（深度包检测）完成，DPI 有开源实现如 Netifyd，nDPI 等。对于 HTTPS 等无法解密的流量，可以通过识别 IP 地址，并维护 URL-IP 的 DNS映射关系，或周知的IP库来判断应用类型。

对于 DPI 的实现和效果，需要再做详细调研。

1.5 广域网优化

对于 VoIP 类型的电话会议，可以通过 `ovs group table`，来向多条wan线路同时传输。因为 `VOIP client` 会以收到的第一个包为准，并忽略后续的重复数据包，而不同wan线路丢掉同一个包的概率基本为 0，以此优化 VoIP 类型的传输。

对于视频会议，可以通过插入FEC纠错码等纠错算法来完成优化，如 [基于内容关键性的高效 FEC 抗网络丢包算法](https://cloud.tencent.com/developer/article/1020364)。

此外，优化手段包括 TCP 传输算法优化（ BBR/ZetaTCP）；数据压缩，对 header、payload进行数据压缩；内存缓存，如 http cache 将热点数据保存在本地直接返回；以及一些 HTTP 层的优化等。

针对广域网优化手段，需要再做详细调研并研究实现方式。

1.6 系统架构