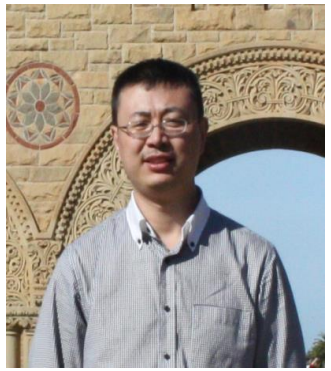
IP网络协议

李振斌 (Robin)
华为首席IP协议专家
IETF互联网架构委员会（IAB）委员

HUAWEI



李振斌

华为首席IP协议专家
IETF互联网架构委员会（IAB）委员
https://www.iab.org/about/iab-members/

- 负责华为IP协议创新研究和标准化工作。
- 2000年加入华为，曾负责华为IP操作系统（VRP）和MPLS子系统的架构设计和开发工作。
- 2015 - 2017年担任SDN架构师，负责控制器的研究、架构设计与开发等工作。
- 自2009年起积极参与IETF标准创新工作，持续推动了SDN的BGP、PCEP、Netconf/YANG等的协议创新和标准化。当前研究的重点包括SRv6、5G承载、Telemetry、网络智能等。
- 主导和参与的IETF RFC/草案累计100余篇(www.ipv6plus.net/ZhenbinLi)，申请专利110多项。
- 2019年当选IETF互联网架构委员会（IAB）委员，承担2019 - 2021年的互联网架构管理工作。

HUAWEI

IETF的组织架构

有一部分面向研究的，research group
比如对去中心化的研究，利用区块链技术，因为现在根域名服务器没有在中国的，如果美国关闭服务，中国是不是就从互联网中消失了。这是一个互联网的主权问题
有一部分面向实际的，工程组

IESG: Internet Engineering Steering Group

- Areas and Working Groups here!

IESG分为七个域

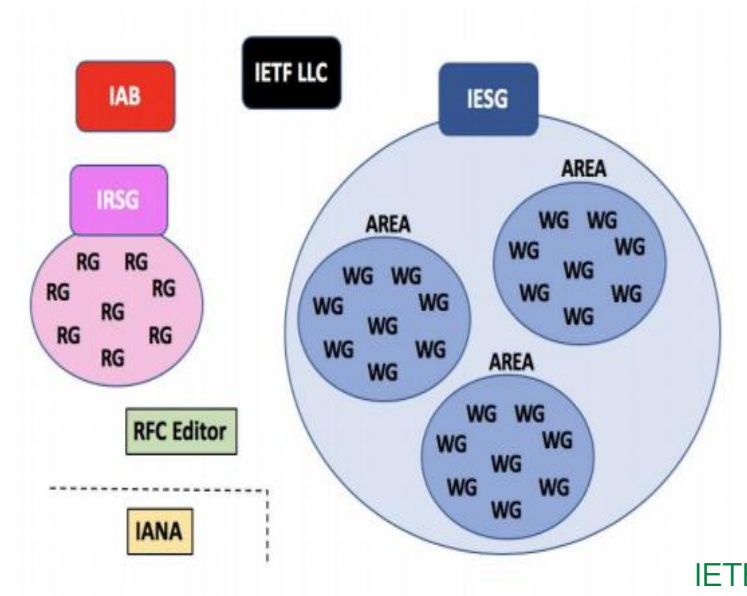
IRTF: Internet Research Task Force

- Research Groups here!

IAB: Internet Architecture Board

IETF LLC: IETF Administration LLC

Applications and Real-Time (ART)	• Application protocols and architectures • Real-time (communication) and non-real-time
Transport (TSV)	• Mechanisms related to data transport on the Internet - Includes congestion control
Routing (RTG)	• Routing and signaling protocols
Internet (INT)	• IPv4/IPv6, DNS, DHCP, mobility
Operations and Management (OPS)	• Network management • Operations: IPv6, DNS, security, routing
Security (SEC)	• Security protocols and mechanisms
General (GEN)	• Activities focused on supporting and updating IETF processes



- ✓ 工作组和个人参与者是IETF的基础
- ✓ AD作为IESG成员负责指导各领域的工作

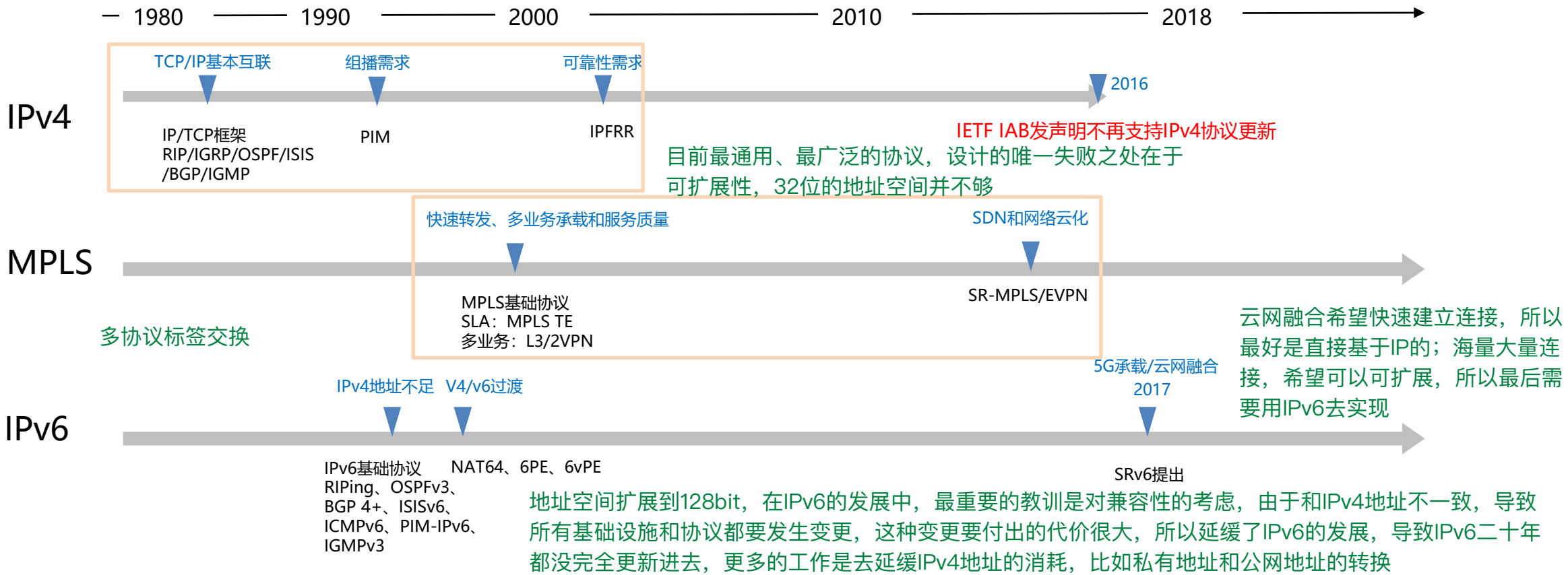
IETF可以认为是一个民间组织，但也正是因为这种特性，才能保证不受政府政治影响，保证互联网的自由自主性，它定的标准都是open的，不涉及未公开技术的讨论
所以在中国不允许访问谷歌时，IETF会发声明说中国不对，美国不用微信，IETF同样也发声明

HUAWEI

IETF协议工作流程

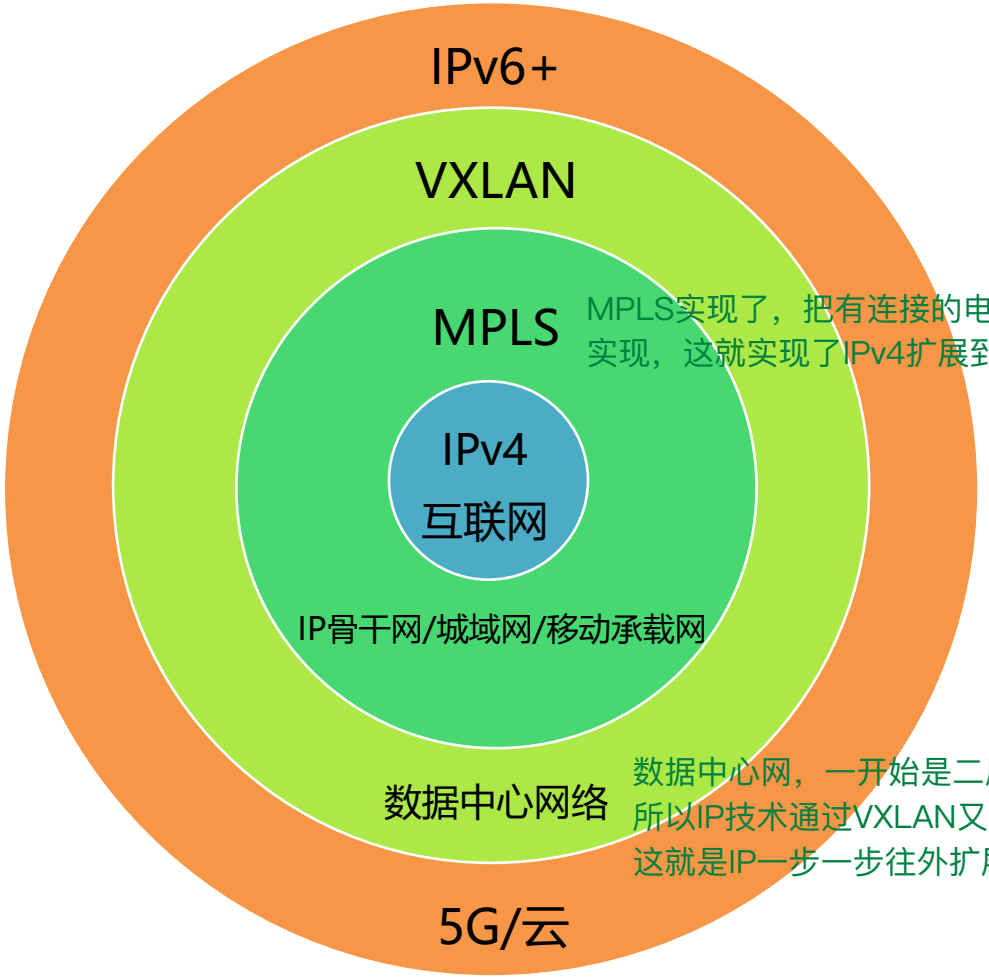
- ✓ 选择一个工作组
- ✓ 通过技术优势获得社区支持
- ✓ 工作组接纳
- ✓ 标准开发
- ✓ 工作组Last Call
- ✓ 撰写Shepherd
- ✓ IETF Last Call
- ✓ IESG Evaluation
- ✓ 编辑
- ✓ 出版
- ✓ 检索

IP网络变革历史



HUAWEI

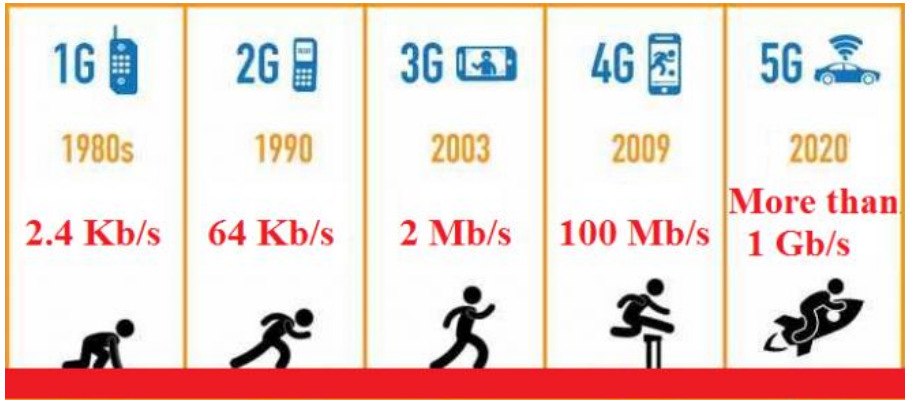
IP历史规律：业务应用带动了IP技术的发展



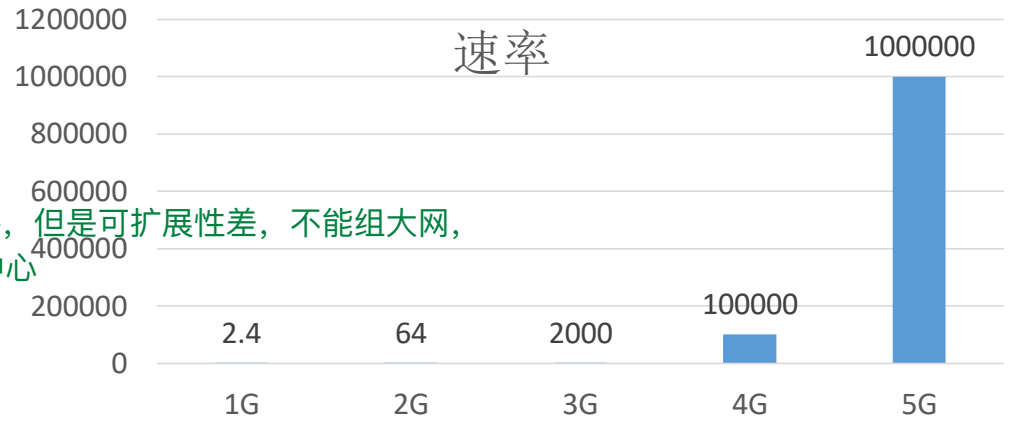
MPLS实现了，把有连接的电信网，用IP网来实现，这就实现了IPv4扩展到电信网

数据中心网，一开始是二层的以太网连接，但是可扩展性差，不能组大网，所以IP技术通过VXLAN又扩展到了数据中心
这就是IP一步一步往外扩展的过程

IP：兼容、连接，围城圈地



图片来自网络，侵删

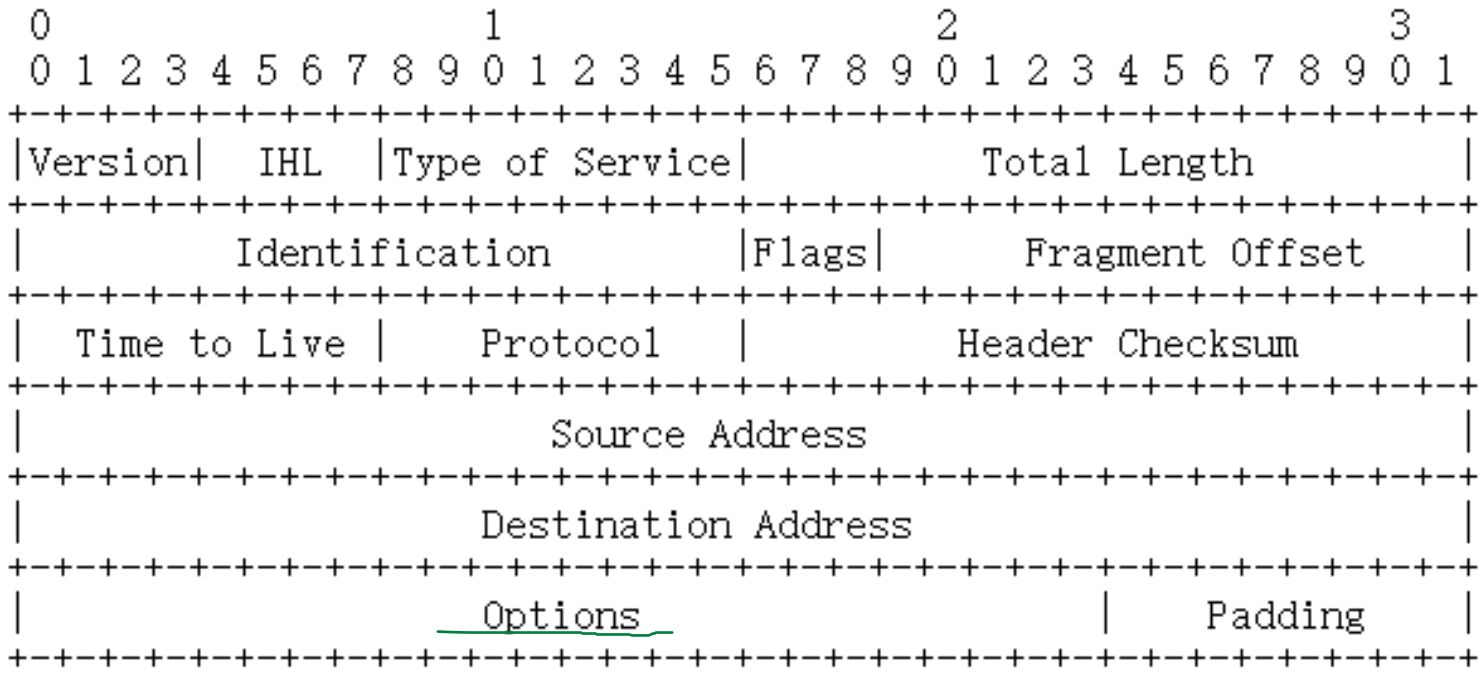


无线、光：更高更快更强

诺基亚一开始是做木材的，然后转型到通信，不仅仅是手机，包括通信基站，诺基亚手机是怎么接入无线的，这些都是诺基亚要做的；华为也一样，并不是只卖手机，更多的是通信设备

HUAWEI

IPv4

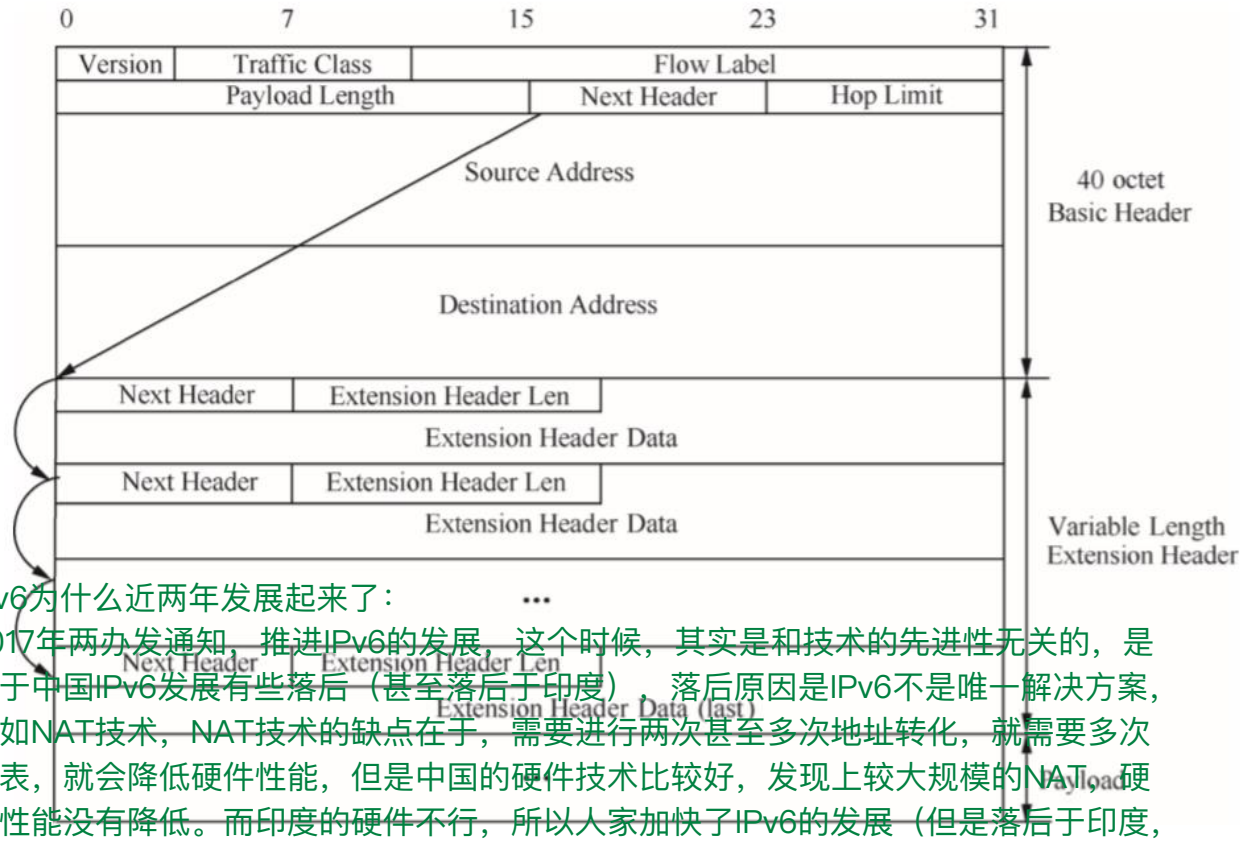


IPv4报文头格式

options字段，本身是想可以自定义的设置一些东西，但是主流的厂商没有一家是支持该字段的
比如源路由选项，在options里就有
为什么不支持，因为硬件的能力有限，硬件在处理定长的东西时表现很好，如果是变长，就很麻烦，所以不支持options字段

IPv4是互联网最基础的协议，也是目前应用最广的协议

IPv6



IPv6 Header

IPv4的升级版本IPv6。IPv6报文头考虑到可扩展性问题，增加了扩展头的设计，为支持新业务带来了更多可能。

在IPV6中，原来的option字段都设计为了扩展头

IPv6基本报文头（IPv6 Header）；

- 逐跳选项扩展报文头（Hop-by-Hop Options Header）
- 目的选项扩展报文头（Destination Options Header）
- 路由扩展报文头（Routing Header）
- 分片扩展报文头（Fragment Header）
- 认证扩展报文头（Authentication Header）
- 封装安全有效载荷扩展报文头（Encapsulating Security Payload Header）
- 目的选项扩展报文头（Destination Options Header，指那些将由IPv6 报文的最终目的地处理的选项）
- 上层协议报文（Upper-Layer Header）

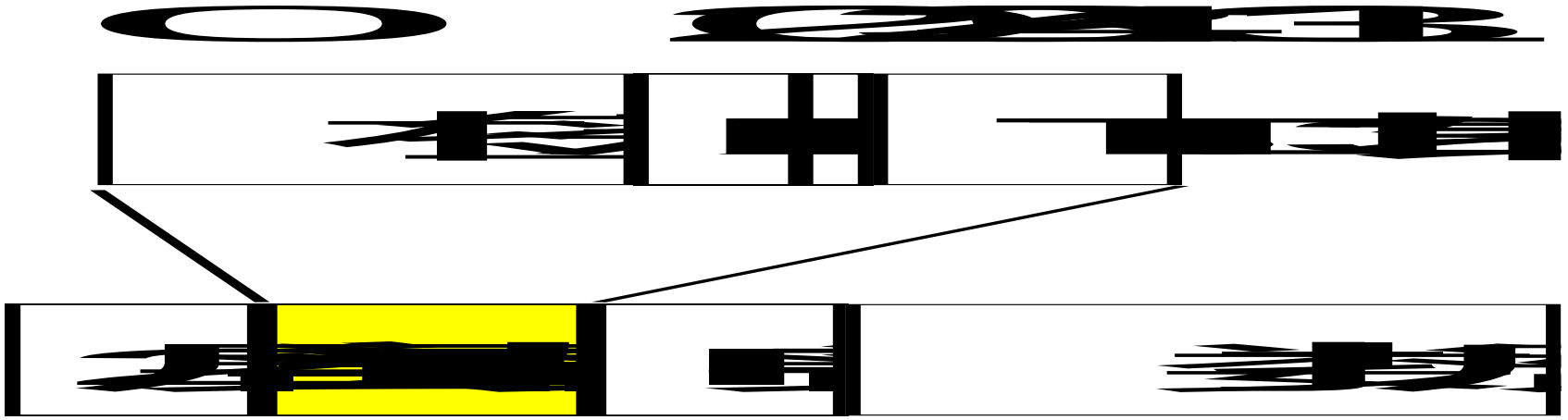
但实际扩展头的功能都没开，因为发现在实际上，扩展头的弊大于利
但是后来发现扩展头的用处很大，就出现了IPv6+时代，所以近两年IPv6得到了较大发展

ipv6为什么近两年发展起来了：
2017年两办发通知，推进IPv6的发展，这个时候，其实是和技术的先进性无关的，是由于中国IPv6发展有些落后（甚至落后于印度），落后原因是IPv6不是唯一解决方案，比如NAT技术，NAT技术的缺点在于，需要进行两次甚至多次地址转化，就需要多次查表，就会降低硬件性能，但是中国的硬件技术比较好，发现上较大规模的NAT，硬件性能没有降低。而印度的硬件不行，所以人家加快了IPv6的发展（但是落后于印度，舆论不允许）
NAT的另一个问题是安全溯源困难
这些原因推动了IPv6的建设，行政力量的推动
同时正好17年3月，SRv6技术的提出

IPv6的出现是因为地址，但是实验证明地址的驱动是失败的，最终的发展驱动力是“服务”

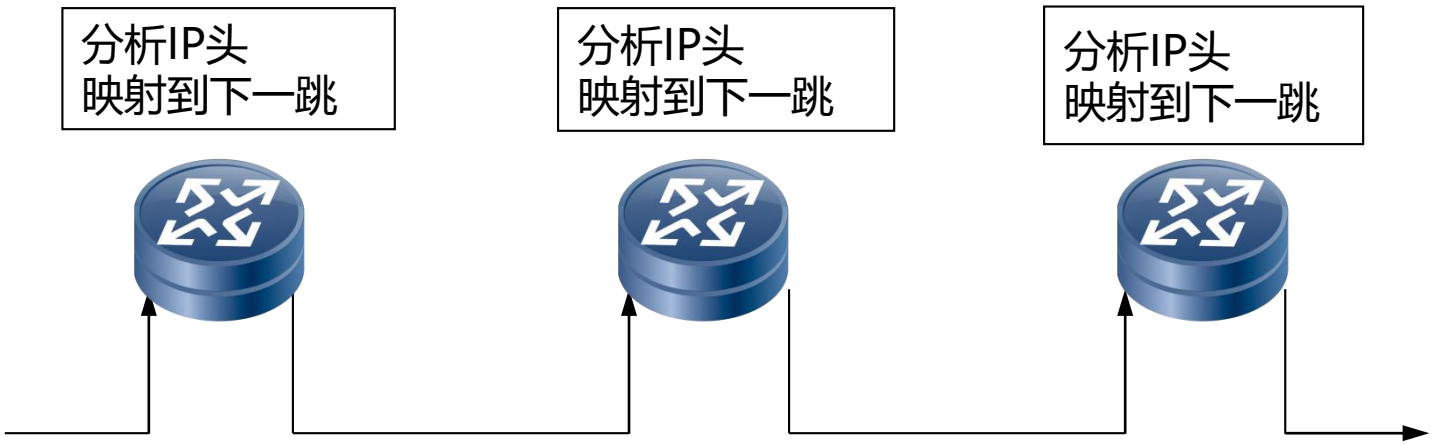
HUAWEI

MPLS封装格式与标签



- Label: 20bit，标签字段，用于存储标签值。
- TC：3bit，实验位，通常用作CoS（Class of Service）。
- S：栈底位，用于指示该标签是否为最后一层标签。如果值为1,表示当前标签头部为栈底；如果值为0，则表示当前标签头部之后依然还有其他标签头部。
- TTL：8bit，与IP报文头中的TTL作用相同，防环。

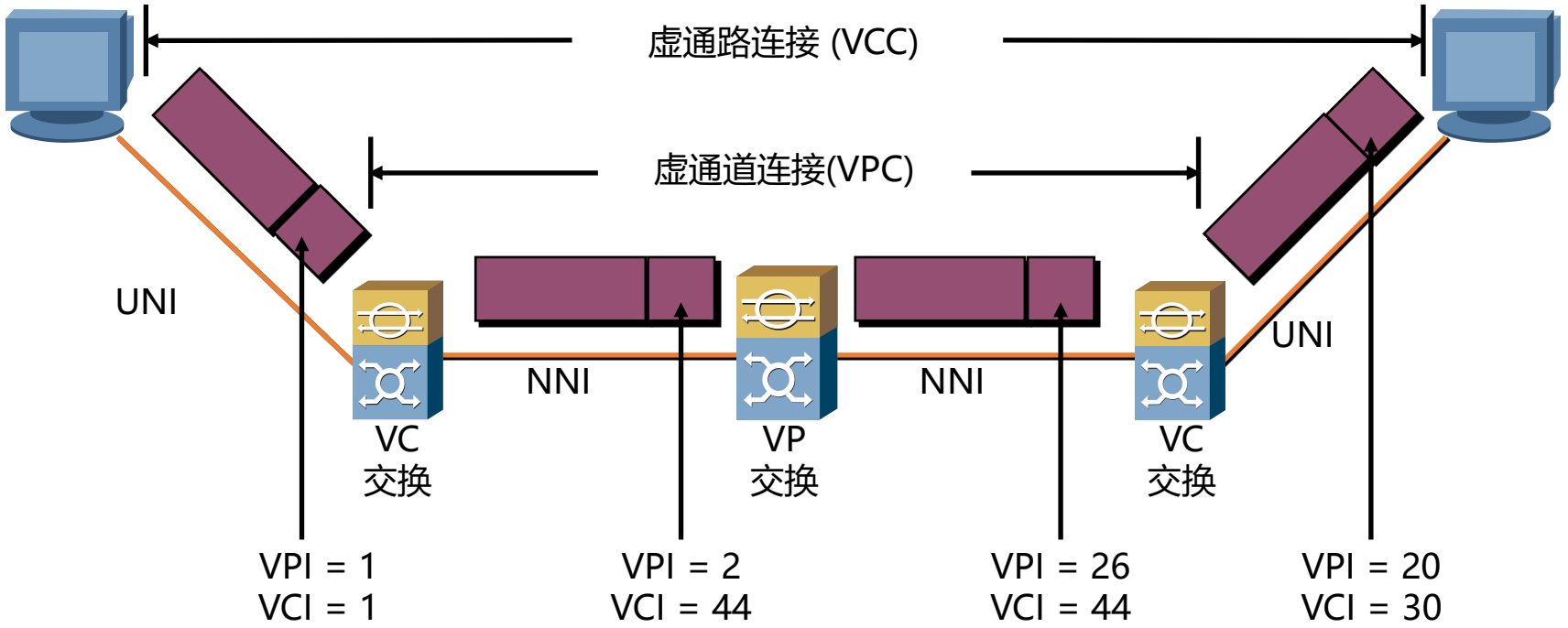
传统IP转发



- 每一跳分析IP头，效率低
- QoS难于部署，而且效率低
- 所有路由器都要知道整个网络的所有路由

ATM交换的过程

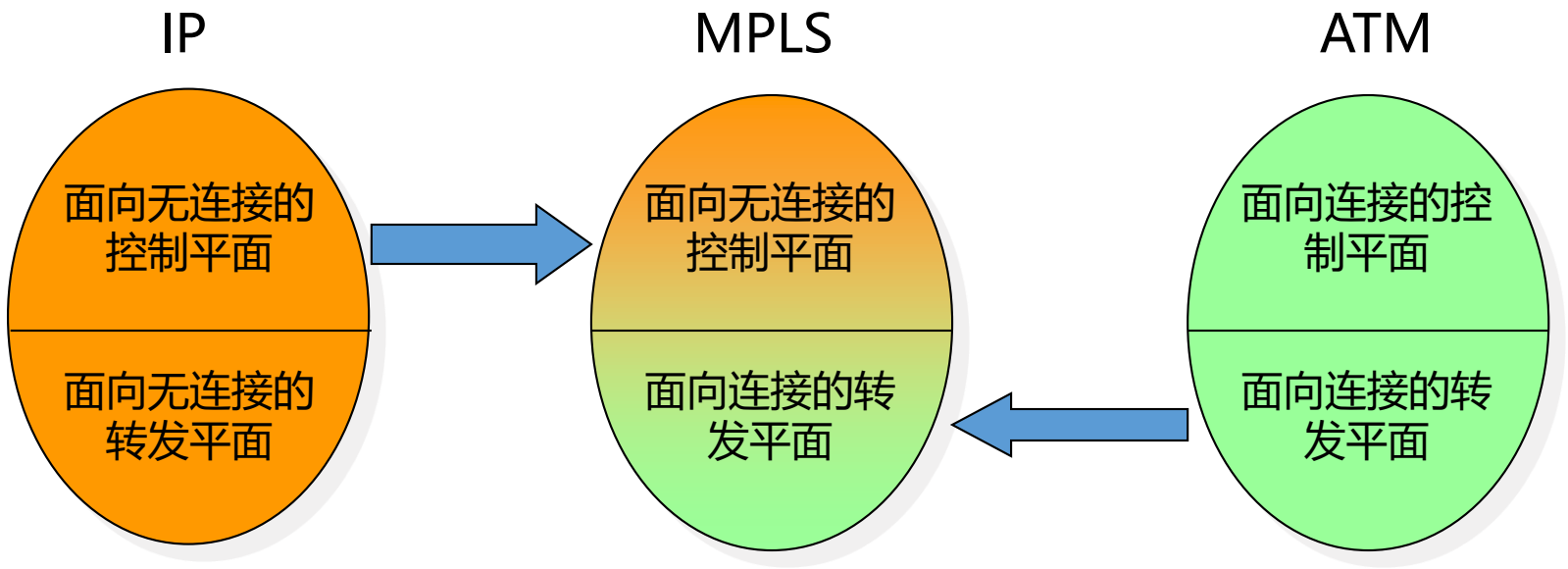
异步传输模式



- 面向连接，有N^2 问题
- 靠链路层选路，基于VPI/VCI或标签
- 业务质量有保证，可保证实时业务

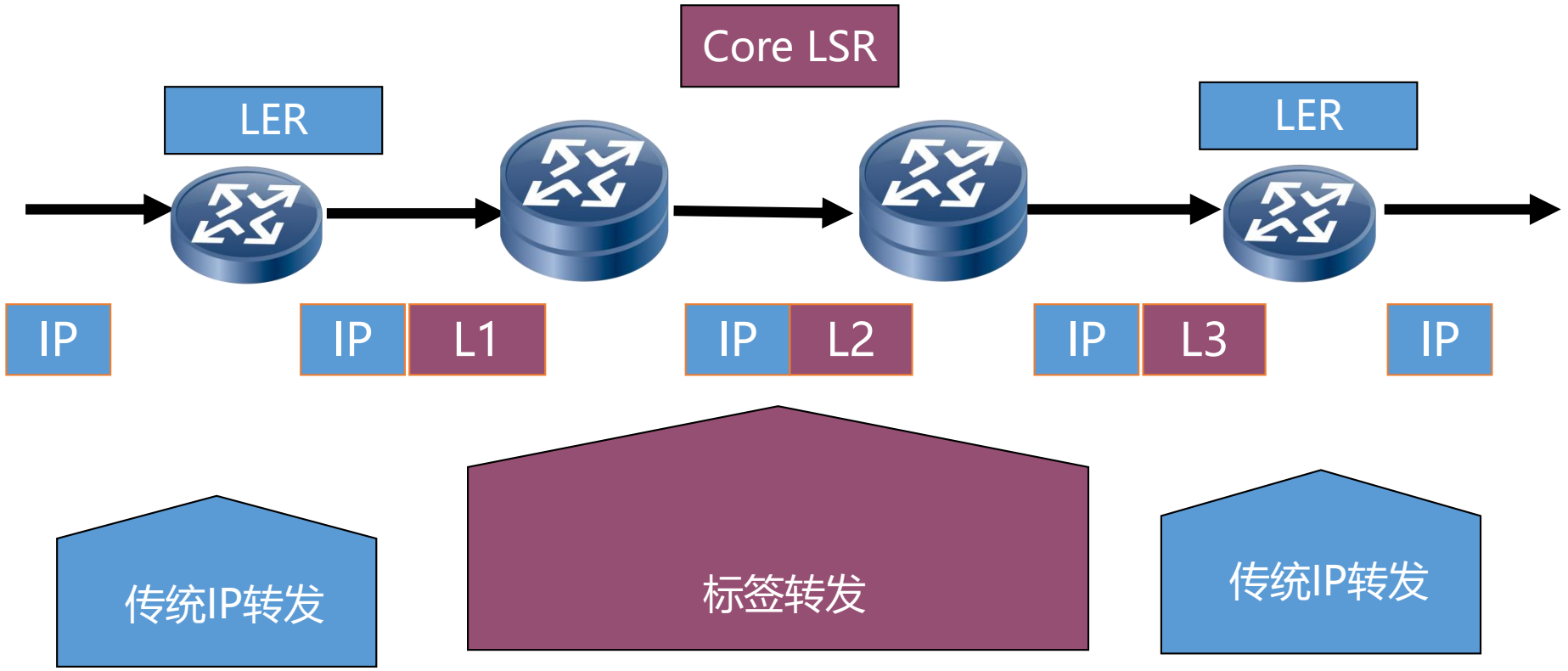
HUAWEI

MPLS：为了将IP与ATM结合



HUAWEI

MPLS 基本工作过程



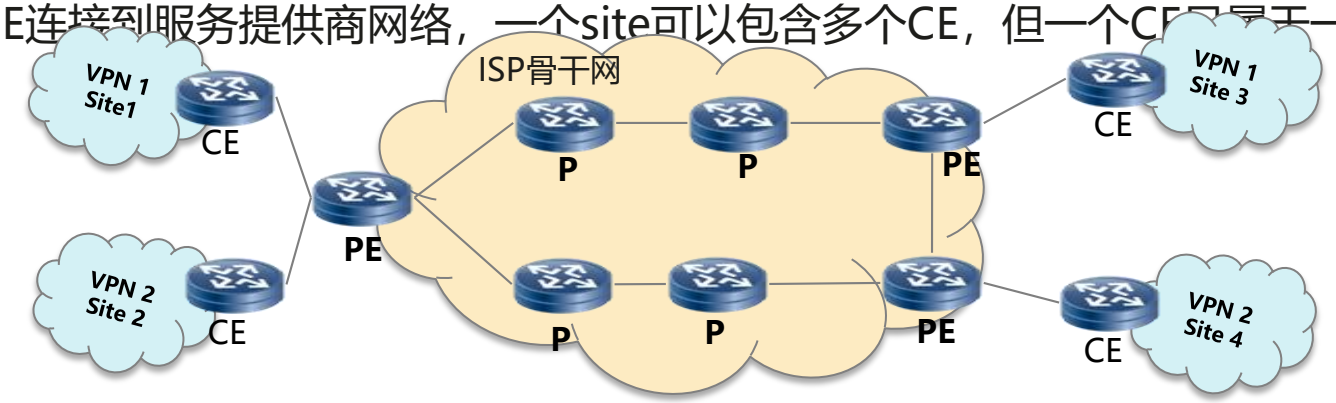
- 以短的、固定长度的标签代替IP头作为转发依据，提高转发速度
- IP与ATM更好地结合，提供有链接的服务
- 提供增值业务，同时不损害效率：
  - VPN
  - 流量工程
  - QOS

MPLS VPN

一开始用MPLS是为了速度快，IP一跳一跳的太慢了，但是后来发现查表也很快
但是MPLS在一些应用场景里还是有用的，用于扩展IPv4的功能，IPv4的头不够用

❑VPN的基本模型

- VPN由以下3部分组成：
  1) CE：CE是用户网络的边缘设备，与SP（service provider，服务提供商）相连。CE可以是路由器或交换机，也可以是一台主机。CE感知不到VPN的存在，也不需要支持VPN的承载协议，如MPLS或SRv6。
  2) PE：PE是服务提供商网络的边缘设备，与用户的CE直接相连。在VPN中，对VPN的所有处理都发生在PE上。
  3) P（provider）：P是服务提供商网络中的骨干设备，不与CE直接相连。P设备不感知VPN，只需要具备基本的网络转发能力（MPLS转发或IPv6转发能力）即可。
- Site（站点）：
  - site是指相互之间具备IP连通性的一组IP系统，并且这组IP系统的IP连通性不需通过服务提供商网络实现。
  - Site的划分是根据设备的拓扑关系，而不是地理位置，尽管在大多数情况下一个site中的设备地理位置相邻。地理位置隔离的两组IP系统，如果它们使用专线互联，不需要通过服务提供商网络就可以互通，那么这两组IP系统也组成一个site。一个site中的设备可以属于多个VPN，换言之，一个site可以属于多个VPN。
  - Site通过CE连接到服务提供商网络，一个site可以包含多个CE，但一个CE只属于一个site。

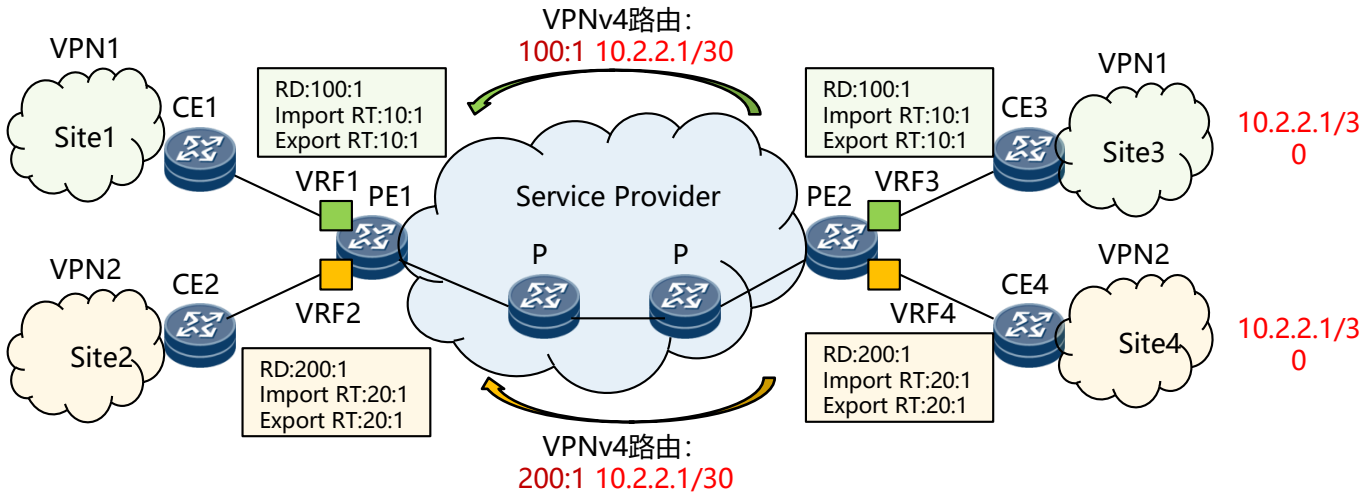


HUAWEI

MPLS VPN

❑MPLS L3VPN

BGP/MPLS IP VPN是一种L3VPN（Layer 3 Virtual Private Network）。它使用BGP在服务提供商骨干网上发布VPN路由，使用MPLS在服务提供商骨干网上转发VPN报文。这里的IP是指VPN承载的是IP报文。



VPNV4地址结构：

Route Distinguisher(8个字节)	IPv4 地址

- VRF（VPN路由转发表，也称为VPN Instance）
- Route Distinguisher (路由区别符）：用于区分使用相同地址空间的IPv4前缀
- VPN Target（也称为Route Target）来控制VPN路由信息的发布

- MP-BGP (Multiprotocol BGP)

传统的BGP-4（RFC1771）只能管理IPv4的路由信息，无法正确处理地址空间重叠的VPN的路由。 MP-BGP采用地址族（Address Family）来区分不同的网络层协议，既可以支持传统的IPv4地址族，又可以支持其它地址族。在MPLS L3VPN组网中PE和PE之间建立通iBGP邻居，通过MP-BGP的VPNv4地址族传递VPN路由。

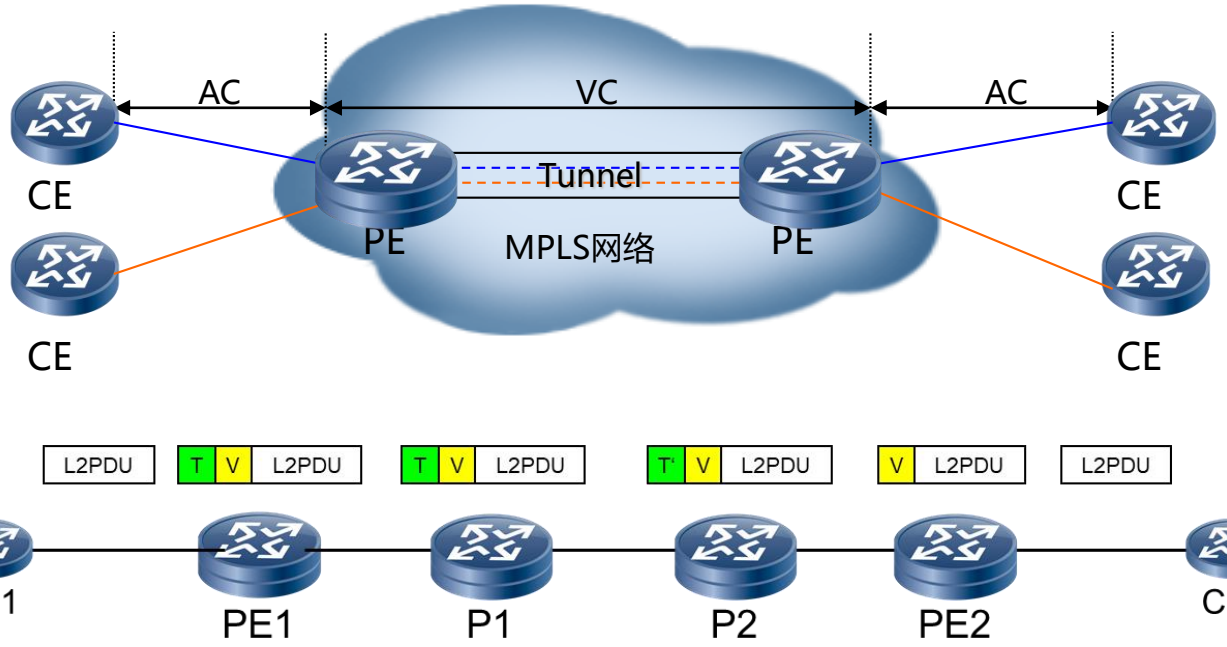
- MPLS (Multi Protocol Label Switching)

服务提供商网络部署MPLS，PE之间创建LSP，使用MPLS LSP为公网隧道对私网数据报文进行封装传输。

MPLS VPN

❑MPLS L2VPN

MPLS L2VPN就是在MPLS网络上透明传递用户的二层数据。从用户的角度来看，这个MPLS网络就是一个二层的交换网络，通过这个网络，可以在不同站点之间建立二层的连接。



- 接入电路AC（Attachment Circuit）：一条连接CE和PE的独立的链路或电路。AC接口可以是物理接口或逻辑接口。
- 虚电路VC（Virtual Circuit）：两个PE节点之间的一种逻辑连接。
- 隧道Tunnel（MPLS Tunnel）：用于在PE之间透明地传输用户数据。

- MPLS L2VPN通过标签栈实现用户报文在MPLS网络中的透明传送。
- 外层标签（称为Tunnel标签）用于将报文从一个PE传递到另一个PE。
- 内层标签（在MPLS L2VPN中称为VC标签）用于区分不同VPN中的不同连接，接收方PE根据VC标签决定将报文转发给哪个CE（哪个AC接口）。

VC提供用户二层数据穿越运营商骨干网络的通道，可以将其简单地理解为连接两个AC接口虚拟线路（点到点连接），将两条用户侧的AC“短接”起来。因此在MPLS L2VPN的实现中，VC又被称为PW（Pseudo Wire，伪线）。

MPLS VPN

❑跨域VPN

国内运营商的不同城域网之间，或相互协作的运营商的骨干网之间都存在着跨越不同自治域的情况。

目前通用的三种跨域VPN解决方案是：

- 跨域VPN-OptionA（Inter-Provider Backbones Option A）方式：

  基本BGP/MPLS IP VPN在跨域环境下的应用，不需要专门配置。这种方式下，两个AS的边界路由器ASBR直接相连，ASBR同时也是各自所在自治系统的PE（这里被称为ASBR PE）。两个ASBR PE都把对端ASBR PE看作自己的CE设备，使用EBGP方式向对端发布IPv4路由。

- 跨域VPN-OptionB（Inter-Provider Backbones Option B）方式：

  两个ASBR通过MP-EBGP交换它们从各自AS的PE路由器接收的标签VPN-IPv4路由。

- 跨域VPN-OptionC（Inter-Provider Backbones Option C）方式：

  ASBR不维护或发布VPN-IPv4路由，PE之间直接交换VPN-IPv4路由。

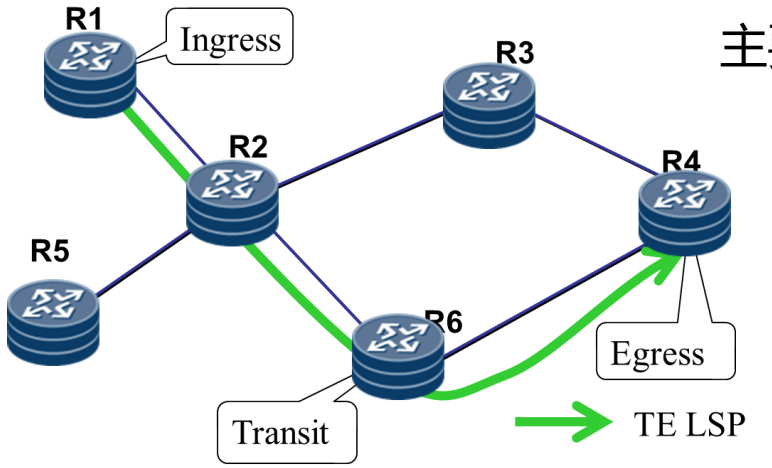
MPLS VPN

❑跨域VPN三种方式的比较

跨域VPN	特点
OptionA	• 优点是配置简单：由于ASBR之间不需要运行MPLS，也不需要为跨域进行特殊配置。 • 缺点是可扩展性差：由于ASBR需要管理所有VPN路由，为每个VPN创建VPN实例。这将导致PE上的VPN-IPv4路由数量过大。并且，由于ASBR间是普通的IP转发，要求为每个跨域的VPN使用不同的接口（可以是子接口、物理接口、捆绑的逻辑接口），从而提高了对PE设备的要求。如果跨越多个自治域，中间域必须支持VPN业务，不仅配置量大，而且对中间域影响大。在需要跨域的VPN数量比较少的情况，可以优先考虑使用。
OptionB	• 不同于OptionA，OptionB方案不受ASBR之间互连链路数目的限制。 • 局限性：VPN的路由信息是通过AS之间的ASBR路由器来保存和扩散的，当VPN路由较多时，ASBR负担重，容易成为故障点。因此在MP-EBGP方案中，需要维护VPN路由信息的ASBR一般不再负责公网IP转发。
OptionC	• VPN路由在入口PE和出口PE之间直接交换，不需要中间设备的保存和转发。 • VPN的路由信息只出现在PE设备上，而P和ASBR路由器只负责报文的转发，使得中间域的设备可以不支持MPLS VPN业务，只需支持MPLS转发，ASBR设备不再成为性能瓶颈。因此跨域VPN-OptionC更适合在跨越多个AS时使用。 • 更适合支持MPLS VPN的负载分担。 • 缺点最主要是安全问题，以及需要维护一条端到端的PE连接管理代价较大。

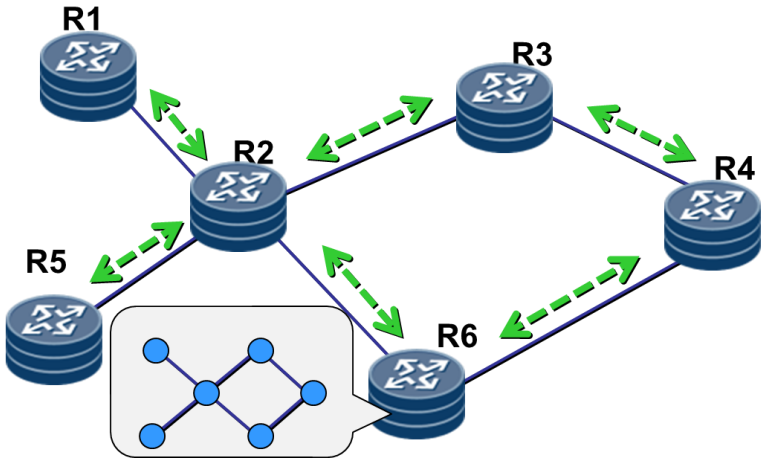
MPLS TE 流量工程

不走最短路径，最短路径可能很堵
用于建立非最短连接



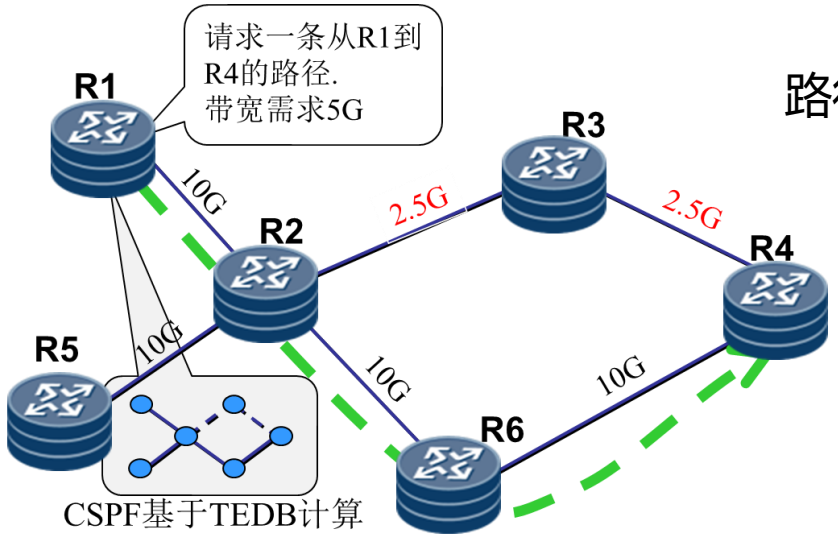
主要组件:

- 链路状态发布组件
  - ISIS-TE
  - OSPF-TE
- 路径计算组件
  - CSPF
- 路径建立组件
  - RSVP-TE
- 报文转发组件
  - 标签转发
  - QOS处理



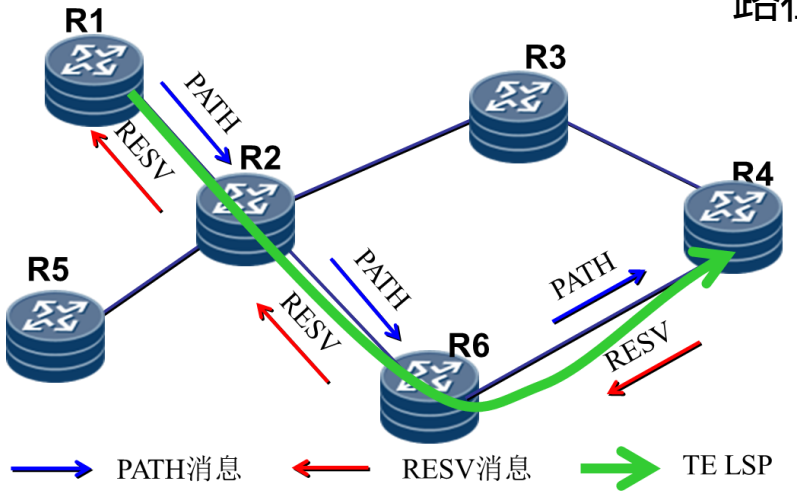
链路状态发布:

- 发布的链路状态参数
  - 本端地址
  - 远端地址
  - 物理带宽
  - 最大可预留带宽
  - 未预留带宽
  - TE metric
  - 管理组属性
- 通过ISIS或者OSPF在对应域内泛洪
- 在ISIS或者OSPF域内的每个节点都会建立TEDB



路径计算:

- 一般由Ingress节点完成路径计算。
- 路径计算以TEDB为基础。
- 使用SPF算法，计算过程中考虑算路请求中的约束。
- 路径计算成功后，就开始路径建立。



路径建立协议RSVP-TE:

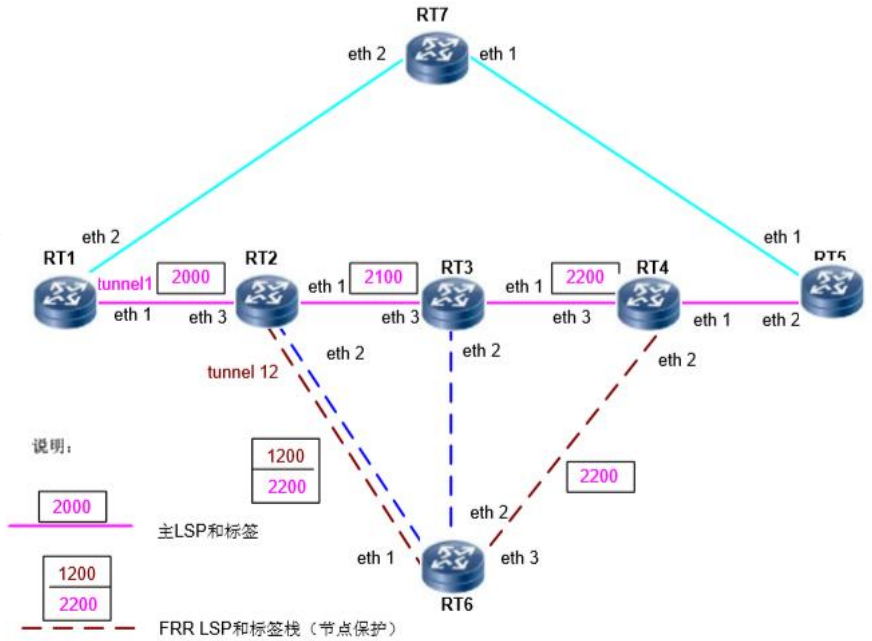
- 通过PATH消息发起路径建立请求，逐跳转发。
- 通过RESV消息回复路径建立确认消息，并携带标签。
- 通过PATH和RESV的软状态消息刷新，维持LSP状态。
- 相对于RSVP的关键扩展:
  - LABEL_REQUEST (PATH)
  - LABEL (RESV)
  - EXPLICIT_ROUTE
  - RECORD_ROUTE (PATH/RESV)
  - SESSION_ATTRIBUTE (PATH)

HUAWEI

MPLS TE FRR

快速路由切换

- MPLS TE 快速重路由是 MPLS TE 中一套用于链路保护和节点保护的机制。当 LSP 链路或者节点故障时，在发现故障的节点进行保护，这样可以允许流量继续从保护链路或者节点的隧道中通过，以使得数据传输不至于发生中断，同时头节点就可以在数据传输不受影响的同时继续发起主路径的重建。
- MPLS TE 快速重路由的基本原理是用一条预先建立的 LSP 来保护一条或多条 LSP。预先建立的 LSP 称为快速重路由 LSP，被保护的 LSP 称为主 LSP。
- 主LSP的建立过程与普通LSP相同，RSVP从头节点（0中的RT1）逐跳向下游发送PATH消息（经过RT1-RT2-RT3-RT4-RT5），从尾节点（0中的RT5）逐跳向上游发送RESV消息。在处理RESV消息时分配标签，预留资源，建立LSP。
- Bypass LSP 的建立可以有两种方式，一种是手工方式，一种是自动方式；手工 Bypass LSP 是当一个没有快速重路由属性的隧道被指定保护一个物理接口以后，它所对应的 LSP 就成为 Bypass LSP。自动 Bypass LSP 是对手工方式的配置简化，当主 LSP 需要被 FRR 保护时，PLR 可以选择或自动建立一条 Bypass LSP，用来保护这个主 LSP，这种方式就叫自动 Bypass



HUAWEI

VXLAN原理与技术

VXLAN基本概念

基于NVo3的二层Fabric组网

NVO3(Network Virtualization Over Layer 3)，基于三层IP overlay网络构建虚拟网络技术统称为NVO3，目前比较有代表性的有：VXLAN、NVGRE、STT。

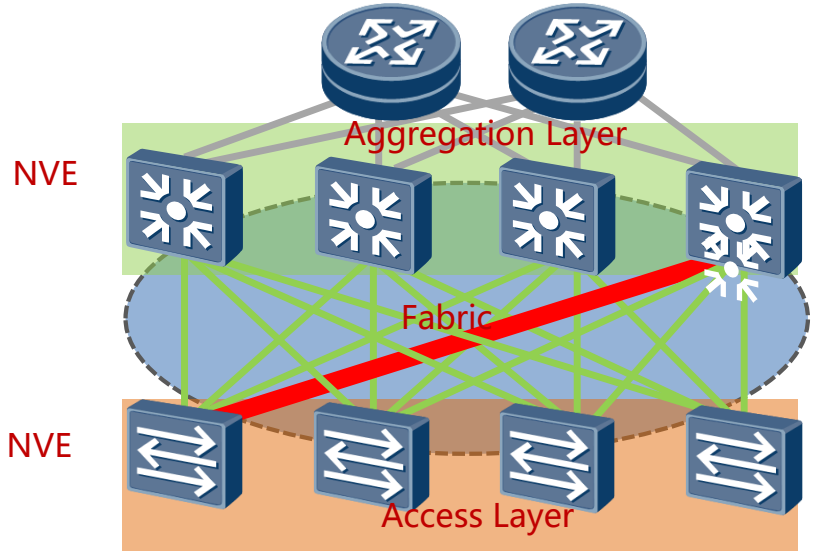
运行NVO3的设备叫做NVE（Network Virtualization Edge），它位于overlay网络的边界，实现二、三层的虚拟化功能；

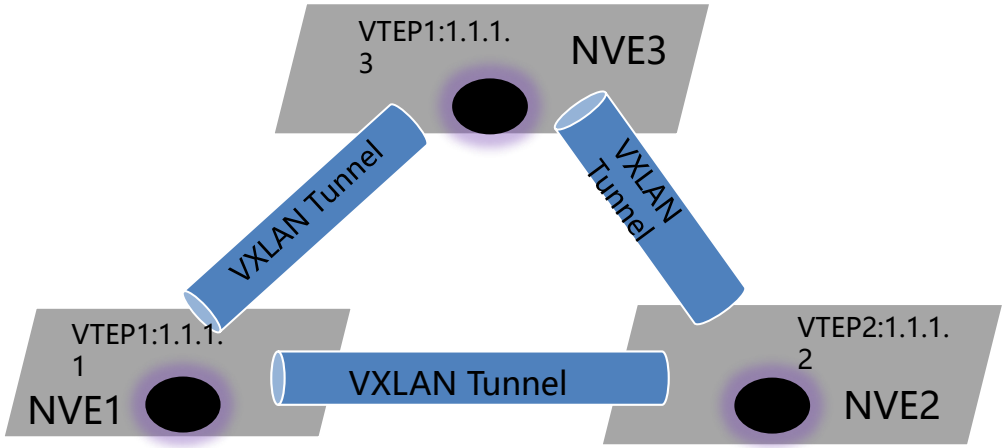
VXLAN(Virtual Extensible LAN，虚拟可扩展局域网)是目前NVO3中影响力最为广泛的一种。它通过LMAC in UDP的报文封装方式，实现基于IP overlay的虚拟局域网。

VXLAN网络中的NVE以VTEP进行标识，VTEP（VXLAN Tunnel EndPoint，VXLAN隧道端点）；

每一个NVE至少有一个VTEP，VTEP使用NVE的IP地址表示；

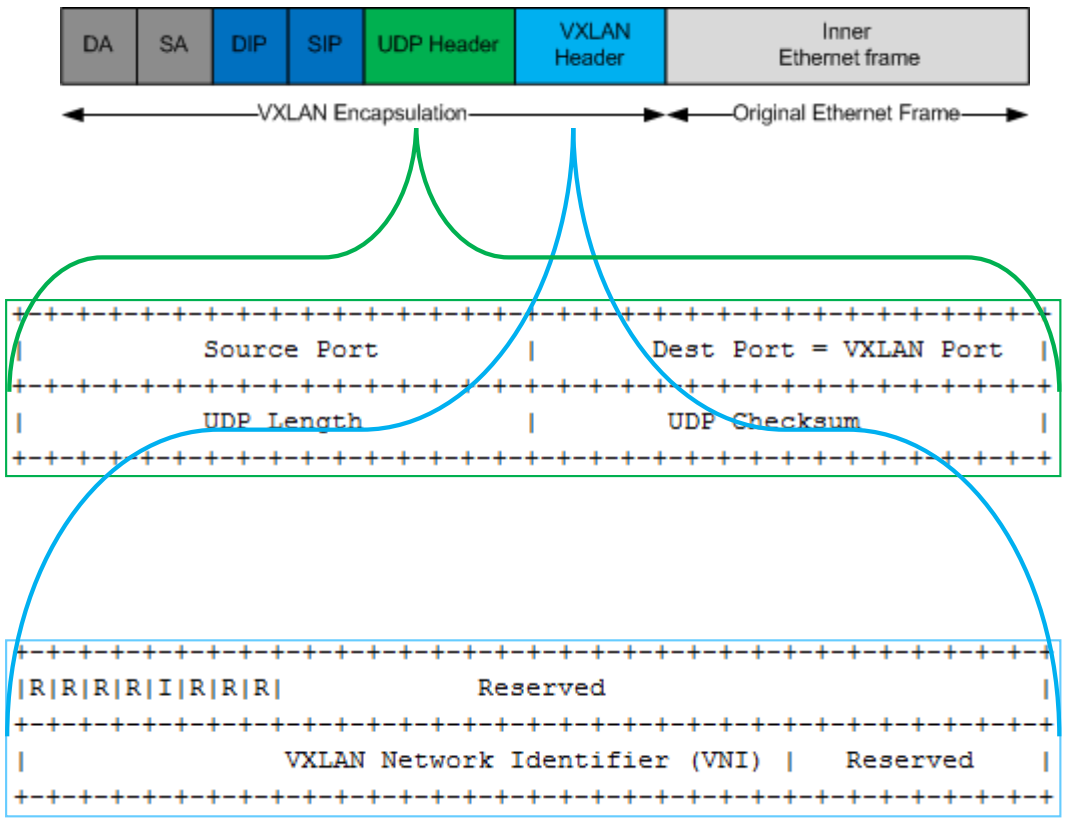
两个VTEP可以确定一条VXLAN隧道





VXLAN原理与技术

VXLAN报文封装



VXLAN报文格式　RFC7348

DA：外层目的MAC，单播为下一跳路由器MAC，组播复制为组播MAC。

SA：外层源MAC，为每一跳路由设备自身MAC。

DIP: 目的NVE的IP地址。

SIP：源NVE的IP地址。

UDP Dest Port：VXLAN保留UDP目的端口号，默认为4789。

UDP Source Port：根据数据流HASH动态生成。

VXLAN I flag：必须置为1，标识VNI字段有效。

VXLAN VNI：VXLAN Network Identifier，24比特，用于标识虚拟网络，最大支持16M。

Original Ethernet Frame：按照标准建议，报文进行VXLAN封装后，需要剥掉原始报文的VLAN TAG，即使不剥掉，在egress NVE也仅仅基于VNI转发（忽略原始报文的VLAN）。

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI