这是什么?

路由器是连接两个或多个网络或子网的设备 [Cloudflare]

路由器使用包含各种数据的分组来引导和指导网络数据, 例如文件、通信和简单的传输 (如 Web 交互) [Cisco]。

路由器是一个网关, 在一个或多个局域网之间传递数据 [Juniper]。

什么是路由器?

(图片幻灯片)

互联网路由

互联网服务提供商的规模各不相同

大规模: 数千个自治网络, 称为自治系统 (AS)

竞争性合作:

– 必须合作以实现全球连通性

– 自利: 独立的经济实体

概述

互联网是一个全球性的网络之网络。

– 连通性通过路由器和路由协议实现

互联指的是网络附加和交换流量的各种方式。

– 一套商业实践和技术机制的集合, 允许各自独立管理的网络为此目的而连接在一起

松散协调

没有中央权威机构管理互联网互联:

– 系统由互联的各个参与者做出的许多双边和多边决策组成

高带宽应用和内容源数量的变化正在改变互联网上的流量增长率和流量传递方法

1995年之前的互联

互联网在美国的最初形式:

– 由美国政府运营的单一骨干网络。

– 较小的区域网络连接到该网络, 形成简单的层次结构。

– 从互联网一部分到另一部分的流量被交接到此骨干网络, 由其携带到目的网络。

多年来, 提供此功能的路由协议的技术要求很简单, 无需处理商业问题。

约1992年的互联

由 Al Gore (时任副总统) 主导的倡议

信息高速公路

1992年克林顿-戈尔竞选期间, 互联变得政治化

1990年: 互联网上只有 313,000 台计算机

1996年: 1000万台

约1995–2005年的互联

骨干网络最终从单一的政府运营骨干网络过渡到由多个商业网络运营商组成的联邦骨干网络模型。

创建了一个新的路由协议, 称为边界网关协议 (BGP), 允许多方商业提供骨干连接。

BGP 允许网络运营商管理这个更复杂和更具竞争性的空间, 并在路由上表达至少一组有限的商业约束。使用 BGP, 网络运营商可以指定其偏好的流量流入和流出其网络的方式。

今天的互联：由于内容提供商网络的整合和扁平化

- 更扁平、高度互联
- 没有单一的大型骨干网络
- 内容分发网络的崛起

流量与互联

超大规模企业的主导地位体现在流量流动方式上

互联网流量模式的变化与互联网连接模型的巨大变化相伴随行

2009年, 互联网流量的一半来自约 150 家公司

2014年, 只有 30 家公司占据了所有流量的一半

2022年, 所有互联网流量的 65% 来自六家公司 (Facebook、Amazon、Google、Apple、Netflix 和 Microsoft)

什么是路由前缀?

互联网上的目的地被聚合到路由前缀中

18.31.0.82 实际上是 32 位的比特串: 00010010 00111110 00000000 01010010

路由器的转发表条目对应于地址前缀 (具有公共前缀比特串的地址范围)

18.0.0.0/8 代表所有范围从 00010010 00...0 到 00010010 11...1 的 IP 地址 (即 2^24 个地址)

什么是路由前缀?

18.31.0.0/17 代表范围为 2^15 个连续 IP 地址的范围 (前 17 位相同)

子网络的大小可以是 1, 2, 4, 8, ... (通常最大为 2^24), 并可以进一步递归划分

转发使用最长前缀匹配

– 在每个路由器上, 路由的形式为 “对于这个地址范围, 使用这条路由”

– 选择与目的地址具有最长匹配前缀的路由

互联网是如何随时间增长的?

互联网上的目的地被聚合到路由前缀中

路由表包含关于前缀的路由信息

路由表增长 (IPv4) — 见图表

https://www.cidr-report.org/

自治系统（ASes）

- 具有相同策略的网络集合
- 单一路由协议
- 通常在单一行政控制下
- 拥有唯一的 ASN (以前是 16 位, 现在是 32 位)

AS 随时间的增长

16 位 -> 32 位 ASN

https://www.cidr-report.org/

自治系统（ASes）

示例:

MIT: 3, CMU: 9

AT&T;: 7018, 6341, 5074, ...

UUNET: 701, 702, 284, 12199, ...

Sprint: 1239, 1240, 6211, 6242, ...

自治系统如何互联以提供全球连通性? 路由信息如何交换?

https://ipinfo.io/AS3

今天剩余内容

理解互联网商业模式

– 互联网路由只是偶然关于转发分组

– 它主要关于金钱

– BGP 中几乎所有的东西都可以用激励和经济学来解释

– 思考: 金钱向哪个方向流动?

路由表条目的详细信息

互联网商业模式（简化）

提供商 (Provider): 付费使用

对等方 (Peer): 免费使用

客户 (Customer): 收费使用

金钱从客户流向提供商

客户得到什么回报？ 客户获得路由的访问权。提供商将其路由表的可见性提供给客户以换取金钱。提供商有动力将其整个路由表提供给客户, 因为提供商可以通过客户发送/接收更多流量。

互联网商业模式（简化）

问: 那为什么我们要有对等互联 (peering)?

互联网商业模式（简化）

问: 那为什么我们要有对等互联 (peering)?

因为在树的顶端没有单一的超级提供商。至少在层次结构的顶部,
我们需要几个互相对等互联且不互相付费的提供商。

无结算对等互联 (Settlement-free peering): 两个 AS 互相交换路由器。

转接（Transit）

只是一个反映客户/提供商关系的技术术语

如果一个 AS 为客户 AS 提供“转接”服务, 它可以在该客户的网络和所有其他互联网端点之间携带流量。

转接关系可以是:

- “完整的” — 客户从其转接提供商获得所有互联网目的地的路由
- “部分的” — 客户获得某些互联网端点子集的路由

转接通常被视为一种收费服务。

实现转接

过滤

来自客户的路由: 向所有人广播

因为这将增加使用你的客户 ($$) 的可能性

实现转接

过滤

来自客户的路由: 向所有人广播

因为这将增加使用你的客户 ($$) 的可能性

来自提供商的路由: 只向客户广播

否则 N 不会从中赚钱, 没有动力向客户之外的任何人广播

N 应该向谁广播路由?

实现转接

过滤

来自客户的路由: 向所有人广播

来自提供商的路由: 只向客户广播

对等互联的激励

MIT 和 Harvard 都向提供商付费。

大学之间的研究合作产生大量流量。

更好的方法是什么?

对等互联（Peering）

对等互联通常发生在两个 AS 之间有大量流量交换且流量有一定对称性的时候

MIT 与 YouTube 对等互联有意义吗?

对等互联（Peering）

如果一个 AS 与另一个 AS “对等互联”，两个 AS 同意仅在它们自己的端点和其客户网络的端点之间交换流量。

– 该协议可以是正式的或非正式的。

– 当对等互联协议被正式化时，通常包含保密和不披露条款

对等互联协议历史上一直是非正式的“握手”协议，但似乎越来越多地涉及合同关系。

在转接关系中，转接提供商通常会向客户广播其所有路由，而在对等互联关系中，对等方只会向另一个对等方广播其客户路由和自己的路由。

实现对等互联

过滤

向对等方广播你自己网络中的前缀和来自客户的路由

来自对等方的路由: 只向客户广播

BGP 的策略

BGP 提供执行各种策略的能力

策略不是 BGP 的一部分: 它们作为配置信息提供给 BGP

BGP 通过从多个备选方案中选择路径并控制向其他 AS 的广播来执行策略

导入和导出策略

导入策略（Import policy）

- 如何处理从邻居学到的路由?
- 选择最佳路径

导出策略（Export policy）

- 向邻居广播哪些路由?
- 取决于与邻居的关系

导出策略

向客户

广播从对等方、提供商和客户学到的所有路由, 以及自有路由

向提供商

广播从客户学到的路由和自有路由

向对等方

广播从客户学到的路由和自有路由

导入路由（从邻居学到的路由）

提供商路由 | 对等方路由 | 客户路由 | 自有路由

(表格图示)

导入路由（从邻居学到的路由）

(表格图示变体)

导出路由：向邻居广播的路由

提供商路由 | 对等方路由 | 客户路由 | 自有路由

(表格图示)

导出路由：向邻居广播的路由

(带过滤器的表格图示)

过滤器阻止某些广播

互联的物理设施

网络要互联, 就必须将它们的网络设备物理连接在一起。

– 这需要网络在共同位置会合, 在能够支持互联所需设备的设施中

– 这些共置设施为客户租赁安全的空间来放置和运行设备

存在点（PoP）: 通信提供商网络的接入点。

互联：公共和私有

互联两个网络需要:

(1) 物理连接性, 以及

(2) 网络连接性。

常见的互联选项:

直接互联: 两个网络之间的私人双边安排, 使用专用物理连接

公共连接: 多边安排, 所有网络连接到公共互联网交换机

公共和私有互联

左图: 具有直接互联的简单共置设施

右图: 还提供通过公共交换机 (或“交换结构”) 的 IX 的共置设施

BGP 如何实现这些路由策略?

互联网路由协议：BGP

自治系统（ASes）— 路由广播

路由表示例: 目的地 | 下一跳 | AS 路径

BGP 会话在直接连接的路由器之间建立。

BGP: 路径向量路由

距离向量路由的扩展

– 支持灵活的路由策略

核心思想: 广播完整路径

– 距离向量: 发送每个目的地 d 的距离度量

– 路径向量: 发送每个目的地 d 的完整路径

用于 BGP

路径向量：灵活的策略

每个节点可以应用本地策略

– 选择: 使用哪条路径?

– 导出: 广播哪些路径?

节点 2 偏好 "2, 3, 1" 而不是 "2, 1"

节点 1 不让 3 听到路径 "1, 2"

BGP 的两种形式

eBGP（外部 BGP）

在 AS 之间交换路由

iBGP（内部 BGP）

在 AS 内部的路由器之间分发到外部目的地的路由

IGP（内部网关协议）

在 AS 内部的路由器之间分发到内部目的地的路由 (不是 BGP)

iBGP

全网状: 每个 eBGP 路由器与 AS 中的每个其他路由器都有一个 iBGP 会话

路由反射: 每个 eBGP 路由器与一个 (逻辑上的中心) 路由反射器有一个 iBGP 会话

BGP 路由表示例

完整路由表: show ip bgp

网络 | 下一跳 | 度量 | LocPrf | 权重 | 路径

特定条目示例, 可以进行最长前缀查找:

BGP 路由表条目: 130.207.0.0/16

AS 路径: 10578 11537 10490 2637

下一跳: 192.5.89.89

Origin IGP, metric 0, localpref 150, valid, internal, best

路由属性和路由选择

BGP 路由具有以下属性, 路由选择过程基于这些属性:

Local preference (LOCALPREF): 由路由策略分配的数值。值越高越优先。

AS 路径长度: 路径中 AS 级别跳数

Multiple exit discriminator (MED): 允许一个 AS 指定某个出口点比另一个更优先。值越低越优先。

eBGP 优先于 iBGP

到下一跳的最短 IGP 路径成本: 离开 AS 前往目的地的最短路径 (“热土豆”路由)

Router ID 平局: 任意平局, 因为只能选择单个“最佳”路由

Local Preference（本地偏好）

控制出站流量

不跨 AS 传递

实现路由偏好的粗糙工具

对于偏好来自某个 AS 的路由而不是另一个很有用 (例如主备用语义)

AS 路径长度

在具有最高本地偏好的路由中, 选择 AS 路径长度最短的路由

最短 AS 路径 != 最短路径, 无论如何解释 “最短路径”

热土豆路由（Hot-Potato Routing）

偏好 IGP 路径成本到下一跳更短的路由

思想: 流量尽快离开 AS

常见实践: 根据传播延迟设置 IGP 权重 (例如英里等)

热土豆路由的问题

IGP 权重的小变化可能导致大规模的流量转移

Multi-Exit Discriminator（多出口判别器）

60

BGP 的复杂性

BGP 是一个非常复杂的协议

– 太多的参数

– 需要适应 (次优的) ISP 策略

– 需要复杂的人工配置

头疼!

BGP 的陷阱和问题

- 没有认证
- 配置错误
- 收敛问题
- 性能问题
- 可靠性问题
- 稳定性问题
- 安全问题
- 问题列表还在继续……

BGP 缺乏认证

设计缺陷: 没有认证或验证机制。

(网络拓扑图示)

幻灯片来源: Cecilia Testart

BGP 缺乏认证

设计缺陷: 没有认证或验证机制。

(带问号的网络拓扑图示)

幻灯片来源: Cecilia Testart

BGP 缺乏认证

危害:

1. 可用性丧失
2. IP 地址滥用
3. 虚假端点

幻灯片来源: Cecilia Testart

使用 BGP 劫持窃取加密货币

https://www.theverge.com/2018/4/24/17275982/myetherwallet-hack-bgp-dns-hijacking-stolen-ethereum

使用 BGP 劫持窃取加密货币

Amazon BGP 劫持: eNET 宣布了 Amazon 地址空间的一部分, 持续了数小时。

虚假 DNS 服务器 -> 虚假 myetherwallet.com

窃取了 1700万美元的以太坊

幻灯片来源: Cecilia Testart

最喜欢的替罪羊!

网络社区 -> BGP