大数据的信息基础设施

传统网络结构

陈一帅

yschen@bjtu.edu.cn

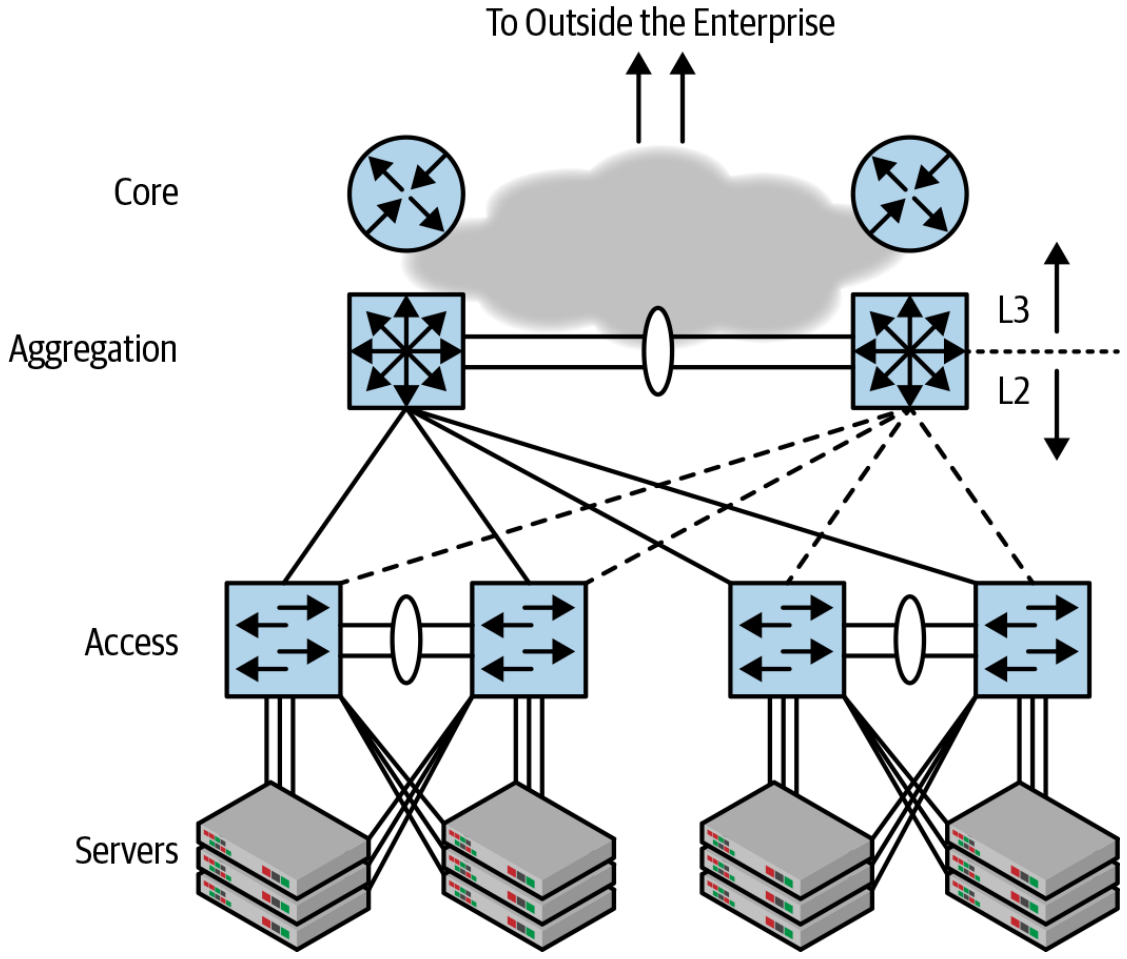
北京交通大学电子信息工程学院

背景

- 传统网络结构
- 数据中心以太网
- STP 桥接机制及其难题

传统网络设计

- Access-Aggregation-Core



背景

- 传统网络结构
- 数据中心以太网
- STP 桥接机制及其难题

数据中心以太网

- 1973 年，Xerox Palo Alto 研究中心（PARC）
- 局域网
  - 可变大小的帧格式，尽力而为传递，错误-检测机制
- CSMA/CD 链路控制
  - 10Mbps，100 Mbps，1 Gbps 版本同时具有 CSMA /CD 和全双工版本
  - 10G 及以上更高速版本中没有 CSMA/CD
  - 现在甚至在家里或 10 Mbps 时都没有 CSMA/CD

以太网速度

- 常用配置
  - 住宅，企业办公室中：1 Gbps
  - 数据中心：10 Gbps，过渡到 40 Gbps 和 100 Gbps
  - 某些运营商核心网络：100G
- 100G 仍比 10×10G 贵
- 进展
  - IEEE 802.3cu 正在使用 400G 以太网标准
  - 以太网联盟正在讨论 800G/1.6T 标准

以太网的分层结构

- 分层架构，是以太网经久不衰的最大原因
- 系统组织为数据链路层和物理层，将与介质相关的方面与与框架相关的操作区分开
- 可以自由采用新的布线和传输速度，同时使用完全相同的第 2 层特性

命名

- 三个值
  - 代表 Mbps 传输速度的数字
  - BASE 表示基带传输
  - 一两个字母指定使用的媒体
- 10BASE-T
  - 10 Mbps
  - 基带传输
  - 双绞线电缆（Twisted-pair）

媒体

- 同轴
- 双绞线
- 光纤
  - 多模、单模
- 直连双轴电缆
  - 在数据中心非常流行

Twinax cables

- 在数据中心非常流行
  - IEEE 要求传输 10E12 位 1 个错误
  - 双芯电缆的误码率（传输 10E18 位为 1 个错误）低得多
- 不同媒介，10 Gigabit Ethernet 的性能比较

Technology	Maximum Distance (m)	Power[1] (W)	Latency (Microseconds)
Twinax Passive	5	0.1	0.1
Twinax Active	10	0.5	0.1
10GBASE-T	100	2.5[2] to 6.5	1.5[2] to 2.5
10GBASE-SR	400	1	0.1

[1]Each side

[2]Short-reach mode (up to 30 meters)

以太网段

- Segment（段）
  - 与使用同轴电缆作为共享通信介质的原始以太网总线有关
  - 在一个段中，所有连接的设备都接收所有传输的帧
- 以太网集线器（Hub）也属于一段
  - 被引入以替代易于出错的同轴总线
- 段定义了冲突域
  - 一个帧在该段上传时，传输另一个帧会出现错误

以太网桥

- 桥
  - 允许多个网段之间通信，不会形成一个大的冲突域

广播域

- 桥定义了广播域
  - 以太网中，广播是必须的
  - 如果桥未在该网段上检测到其目的 MAC 地址，它将帧从一个网段传输到另一网段
  - 它必须将广播帧转发到所有桥接段
  - ARP 学到的 MAC 只保留 300s，DHCP 也依靠广播
- 桥的广播机制
  - flooding，来者不回
  - Spanning Tree 避免 loop

交换机

- 交换机
  - 桥概念的演变
  - 它的转发过程基于硬件
  - 通常具有比网桥更多的端口
  - 一个交换机也定义了一个广播域

背景

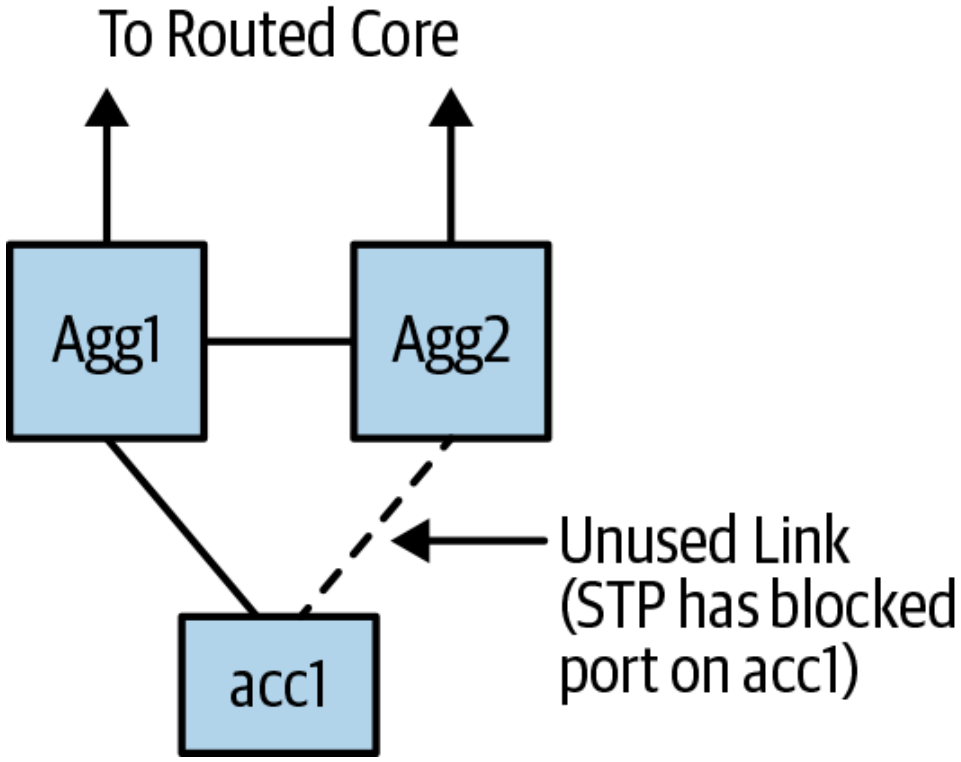
- 传统网络结构
- 数据中心以太网
- STP 桥接机制及其难题

STP 环破除

- Spanning Tree Protocol
- 为了容忍链路失败，会有冗余链路
- 冗余链路带来 Loop
- 与 IP 数据包不同，以太网帧没有生存时间（TTL）标头字段
- 在 Loop 中，ARP，DHCP 广播包，带未知 MAC 地址的包会无限循环
- 几毫秒内，把整个链路撑满，网络崩溃

STP 环破除

- 破除环


To Routed Core
Agg1
Agg2
Unused Link
(STP has blocked
port on acc1)
acc1

STP 环破除

- Radia Pearlman（也称为“互联网之母”）1980 年代创建
- 检测桥接网络中的环路，阻止所选端口中的流量
- 默认路径成本

Bandwidth	Short-Path Cost Method	Long-Path Cost Method
10 Mbps	100	2,000,000
100 Mbps	19	200,000
1 Gbps	4	20,000
10 Gbps	2	2000
40 Gbps	1	500
100 Gbps	1	200

STP 环破除算法

- 类似距离矢量算法
- 交换机根据以下决策选择是否阻止所选端口中的流量（“死亡”）
  - 最小的根网桥 ID
  - 到根网桥的最低路径成本
  - 最小发送方网桥 ID
  - 最小端口号
- 也叫死亡匹配

STP 环破除收敛过程

- Radia Pearlman 可爱的 Algorhyme 诗

```
I think that I shall never see
A graph more lovely than a tree.
A tree whose crucial property
Is loop-free connectivity.
A tree that must be sure to span
So packets can reach every LAN.
First, the root must be selected.
By ID, it is elected.
Least-cost paths from root are traced.
In the tree, these paths are placed.
A mesh is made by folks like me,
Then bridges find a spanning tree.
```

Flooding 的可扩展性问题

- 无论如何分割，自学习桥的""flood and learn""模型都无法扩展
- MAC 地址不是分层的。因此，MAC 转发表是对 VLAN 和数据包的目标 MAC 地址的简单 60 位查找
- 通过泛洪和学习来学习一百万个 MAC 地址，并由于超时而定期重新学习它们，几乎每个网络架构师都认为这是不可行的
- 终端或者虚拟终端将被迫处理百万个数据包的周期性洪泛

STP 在实际中的难题：不稳定，不可预测

- 一个普通的故障就可能导致环，因此带来严重故障，如
  - 对等 STP 出于某种原因而无法及时发送 hello 数据包（例如，因为它正在处理 ARP 风暴），则其他对等 STP 假定远端没有运行 STP，并开始将数据包转发出链路到不堪重负的开关。这将立即导致环路，并引发广播风暴，从而完全破坏网络。
  - 一个案例：交换硅片有一个错误，该错误导致数据包泄漏出阻塞的交换机端口，无意间形成了环路，从而引发了广播风暴

STP 在实际中的难题：不稳定，不可预测

- STP 根选举程序可能会被取消，导致错误的设备被选举为根
  - 案例：向网络中添加新设备时，遇到了太多的网络故障，以至于要求交换机端口在配置之前被禁用。默认情况下，在禁用端口的情况下，新添加的交换机上的 STP 不会自动加入网络并选择根
- 许多活动部件（通常是专有部件）的存在也导致网络变得不可预测且难以进行故障排除

小结

- 传统网络结构
- 数据中心以太网
- STP 桥接机制及其难题

练习

- STP 算法