Lecture 2 Internet and Datacenter

Internet

What's "Internet"

The **Internet** (or **internet**) is the global system of interconnected computer networks that uses the Internet protocol suite (TCP/IP) to communicate between networks and devices.

Internet is a "Network of network"

- Campus Networks
- Enterprise Networks
- Internet Service Providers (ISP)

Each network is owned and managed by some organization, and connected together

Autonomous Systems (AS)

Internet is divided into Autonomous Systems

- Node: Autonomous System (AS)
- Edge: Two ASes connect to each other

Autonomous System Numbers

AS Numbers are 16bit values.

Internet Service Provider (ISP)

- An Internet service provider(ISP) is an organization that provides services for connecting to the Internet
- An ISP can have multiple ASes
- In fact, there are grades for ISPs in Internet

Ties of ISPs

The Relation Between Neighboring Nodes

- Neighboring ASes have business contracts
  - How much traffic to carry
  - Which destination to reach
  - How much
- Common business relationships
  - Customer-Provider (上下关系)
  - Peer-Peer (同层级关系)

Customer-Provider Relationship

Customer needs to be reachable from everyone

- Provider tells *all neighbors how to reach the customer*
  Customer doesn't want to provide transit service
- Customer doesn't let its providers route through it

Multi-Homing: Two or More Providers

- Motivations for multi-homing
  - Higher reliability, survive single ISP failure
  - Better performance by selecting better path
  - Financial leverage through competition
  - Gaming the 95th-percentile billing

Peer-Peer Relationship

Some networks have an incentive to connect directly, to reduce their bill with their own provider

Peers exchange traffic between customers

- AS exports *only customer routes* to a peer
- AS exports a peer's routes *only to its customers*
- Often the relationship is settlement-free (免费结算)

How to Connect ASes?

Internet eXchange Points (IXPs)

- In fact, the connection between ISPs is not C_n^2
- Interconnecting each network to its neighbors one-by-one is not cost effective
- Many networks **connect in one location** ⇒ IXP

Tier-1 Providers

Tier-1 Provider

- Has *no upstream provider* of its own
- Typically has a national or international backbone
- Usually *no fee between each Tier-1 Providers*
  Top of the Internet Hierarchy of 12-20 ASes
- *Full peer-peer connections* between *tier-1* providers

Tier-2 Providers

- Provide transit service to downstream customers

- But, need at least one provider of their own
- Typically have national or regional scope

Characteristics of AS Paths

- AS Path may be longer than shortest path
- Router Path may be longer than shortest path

Routing

- inter-domain routing: Find paths *between* networks ⇒ BGP
- intra-domain routing: Find paths *within* a network ⇒ OSPF / RIP / ISIS

Intra Domain Routing

.....

Inter Domain Routing - BGP

⇒ BGP ⇒ very complex !

BGP is Path Vector Protocol

- *BGP announcement* carry **complete path information** instead of distances.
- Every "node" inherits the "Prefix" and put its own Node_Num into it, so that it can realize the "complete path information"
- Complete path enables
- Each AS is free to select and use any path preferably, maybe the cheapest one.
- BGP is Policy Based (very flexible)

The Common Used BGP Policy

- Prefer the path with the highest WEIGHT.
- Prefer the path with the highest LOCAL_PREF.
- Prefer the path that was locally originated via a **network** or **aggregate** BGP subcommand or through redistribution from an IGP. Local paths that are sourced by the **network** or **redistribute** commands are preferred over local aggregates that are sourced by the **aggregate-address** command.
- Prefer the path with the shortest AS_PATH. (judged by length of AS_PATH)
- Prefer the path with the lowest origin type. (IGP / EGP)
- Prefer the path with the lowest multi-exit discriminator (MED).
  ……

More information is listed in CISCO_BGAlggorithm

It's so complex that many issues are caused by BGP

- Globally issues
- The configuration is of great significance and is hard to config for people.
- SDN is excellent but we still use old-system mostly now.
- Therefore, if SDN is commonly used in social life, the BGP can be discarded !

Data Center Networks

Traditional Architecture

Traits

- Three Layers: core / distribution (aggregation, 聚合/汇总) / access
- Layer 2 ⇒ MAC ⇒ STP solves the Loop problem
- Layer 3 ⇒ IP

Limitations

- Waste of Bandwidth
  - STP block ports except one
- Latency
  - Server to server traffic: need many hops
- No for the cloud
  - rise of east-west traffic
  - VM migration（虚拟机迁移）

Solution: Fat-Tree

基础特点：上面的链路带宽大，下层的链路带宽小

Clos Network

⇒ multi-stage switching

Theorem 1

> If m ≥ 2n-1, the Clos Network is strict-sense nonblocking

Theorem 2

> If m ≥ n, the Clos Network is Rearrangeably nonblocking

Folded 3-Stage Clos Network

Traits

- Spine - Leaf Structure
- **Equal-cost multi-path (ECMP)**: IP and Routing ⇒ All traffic is evenly distributed!
- Server to server traffic: 3 hops
- Large bandwidth for east-west traffic
- VM migration

Limitation

Problem: each switch still needs a lot of ports

Benes Network

Solution: Adding more stages

⇒ reducing an N x N cross-bar switch to two N/2 x N/2 crossbar switches and two N-input exchange switches

N x N Benes Network

2 log2N−1 stages, each containing N/2 2×2 crossbar switches, and use a total of N log2N − N/2 2×2 crossbar switches.

Top-of-Rack (机柜顶部) Architecture

- Rack of servers
  - top-of-rack switch
  - commodity servers
- Modular Design
  - Preconfigured racks
  - Power, Network, and storage cabling

Fat Tree

Structure Traits

- Each switch has k *ports*
- There are k *pods*
- Each pod has k/2 *aggr.Switches* and k/2 *edge.Switches*
- Each edge.Switch is connected to k/2 hosts and k/2 aggr.Switches

- Each *core.Switch* is connected to **an aggr.Switch for each pod**

Some Details

- How many switches are needed?
  - k * (k/2 + k/2) + $(1/4)k^2$
- How many servers that it can hold?
  - (k/2) * (k/2) *k = $(1/4)k^3$
- How many hops between each pair of hosts at different pods?
  - ≥ 6
- How many 6-hop paths between hosts at different pods?
  - (k/2)*(k/2) = (1/4)k^3

Extensions

Method: Add more cheap switches!

Facebook DCN Topology

Google Jupiter Topology

LinkedIn DCN Topology