# How Computer Network Route Your Packet

## Overview of BGP, AS, and Anycast Networks

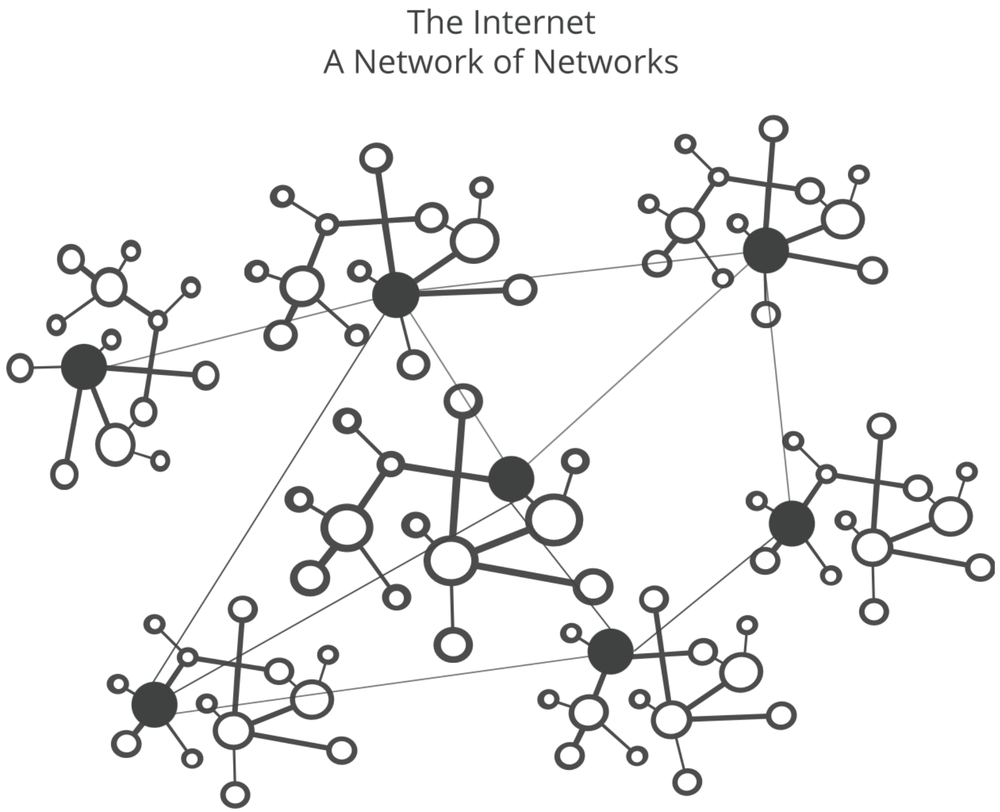
TheUnknownThing

“Border Gateway Protocol is the postal service of the Internet. When someone drops a letter into a mailbox, the Postal Service processes that piece of mail and chooses a fast, efficient route to deliver that letter to its recipient. ”

**https://www.cloudflare.com/learning/security/glossary/what-is-bgp/**

# How Does Our Network Function?

**What you think v.s. What it actually is**

# A Network of Networks

## Autonomous System

ASes are like individual post office branches.

We may have tens of millions of mailboxes worldwide, but the mail in those boxes must go through the local postal branch before being routed to another destination.

Examples? 中国电信，中国移动，中国联通——AS4134, AS9808, AS4837

Routers in this map would only have to remember their “local information.”

# A Network of Networks

## Autonomous System

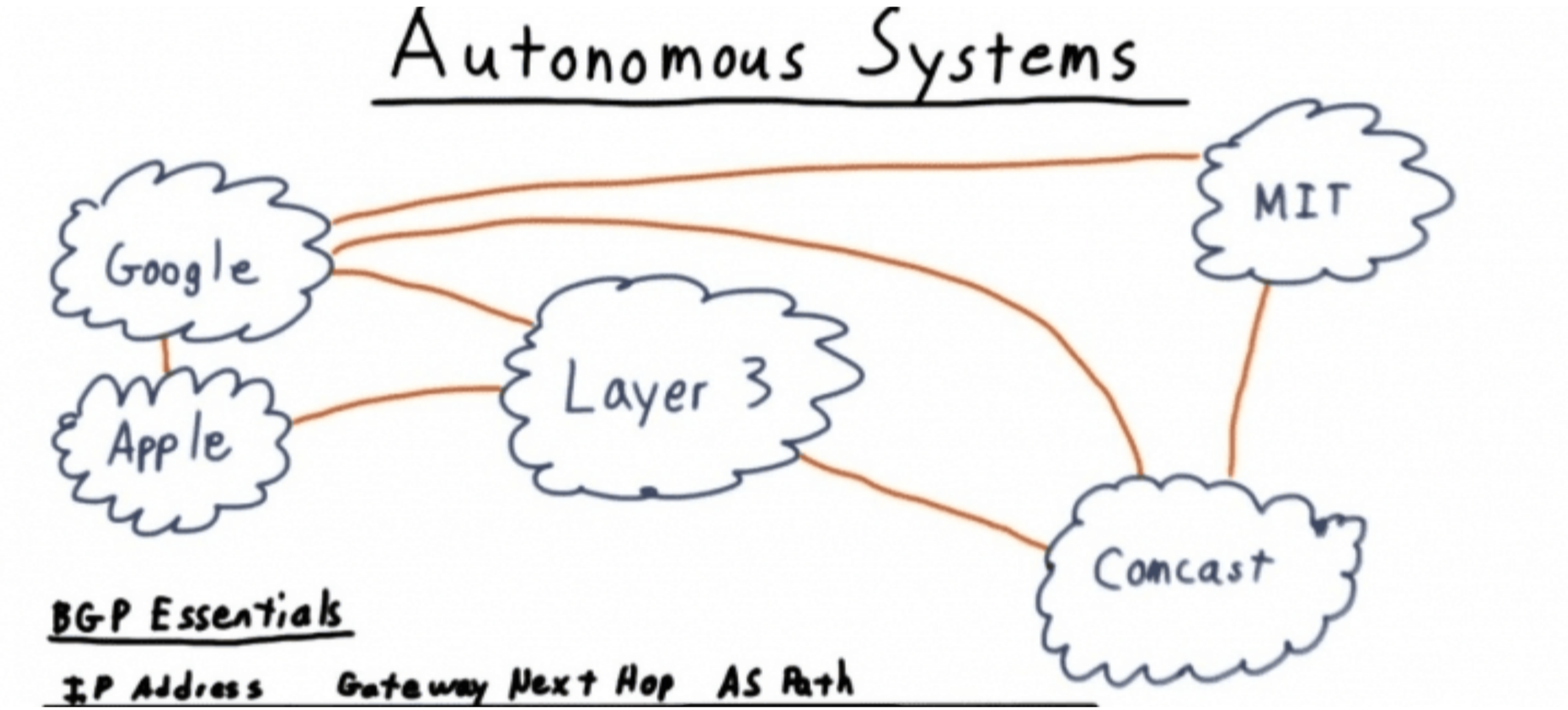
Formally, an Autonomous System (AS) is a collection of connected IP routing prefixes under the control of a single administrative entity that presents a common, clearly defined routing policy to the Internet.

An AS is typically operated by a single large organization, such as an Internet Service Provider (ISP), a major technology company, a university, or a government agency.

Internet Assigned Numbers Authority (IANA) assigns each AS a globally unique number known as an Autonomous System Number (ASN).

# A Network of Networks

**How do ASes Connect?**



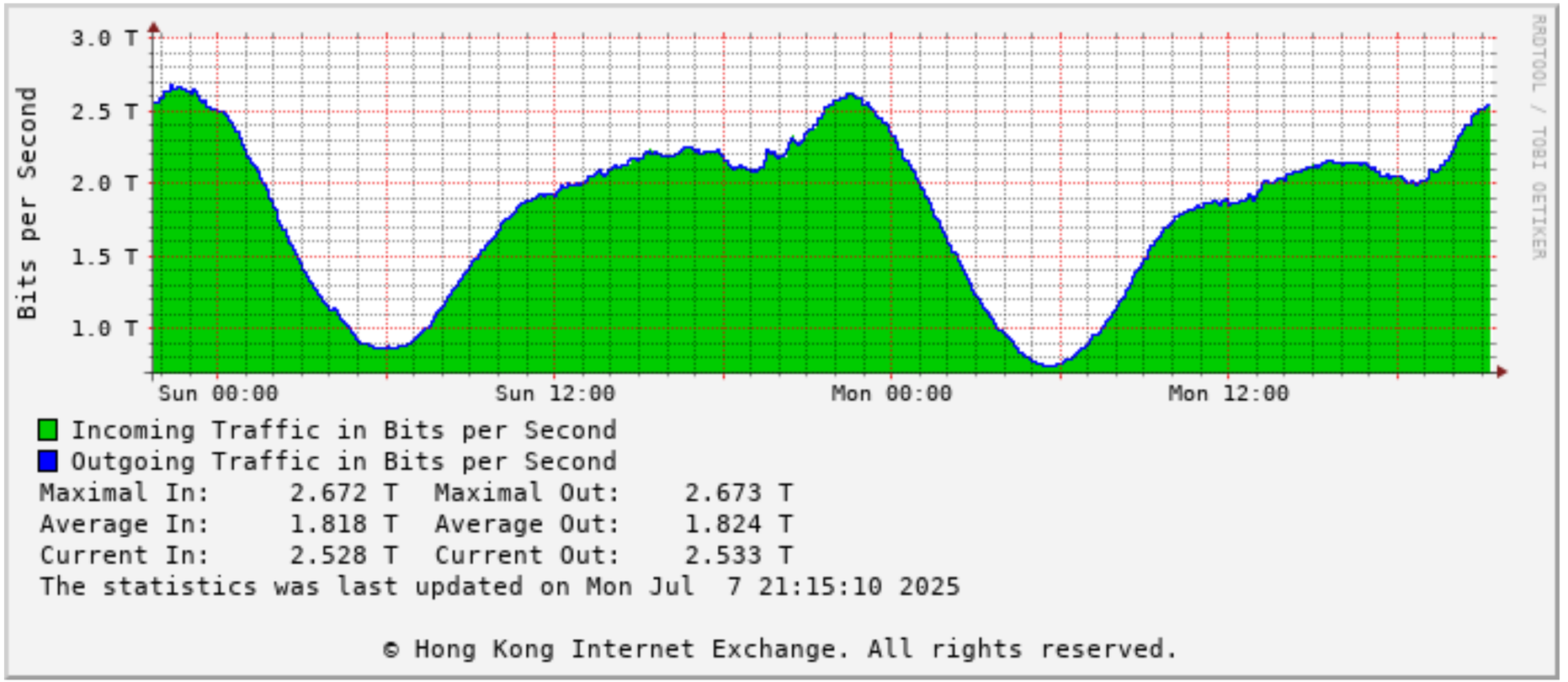
# A Network of Networks

## How do ASes Connect?

- **Transit:** Paid Service. A smaller network (the "customer") pays a larger network (the "provider" or "upstream") to carry its traffic to and from the rest of the Internet. This creates a hierarchical structure where smaller ISPs connect to larger ones, forming a path to the global Internet.
- **Peering:** 2 ASes of roughly equal size and traffic volume agree to exchange traffic directly between their respective customers. Peering often occurs at physical locations called Internet Exchange Points (IXPs), which are large data centers containing switches that allow dozens or even hundreds of ASes to interconnect efficiently.

# A Network of Networks

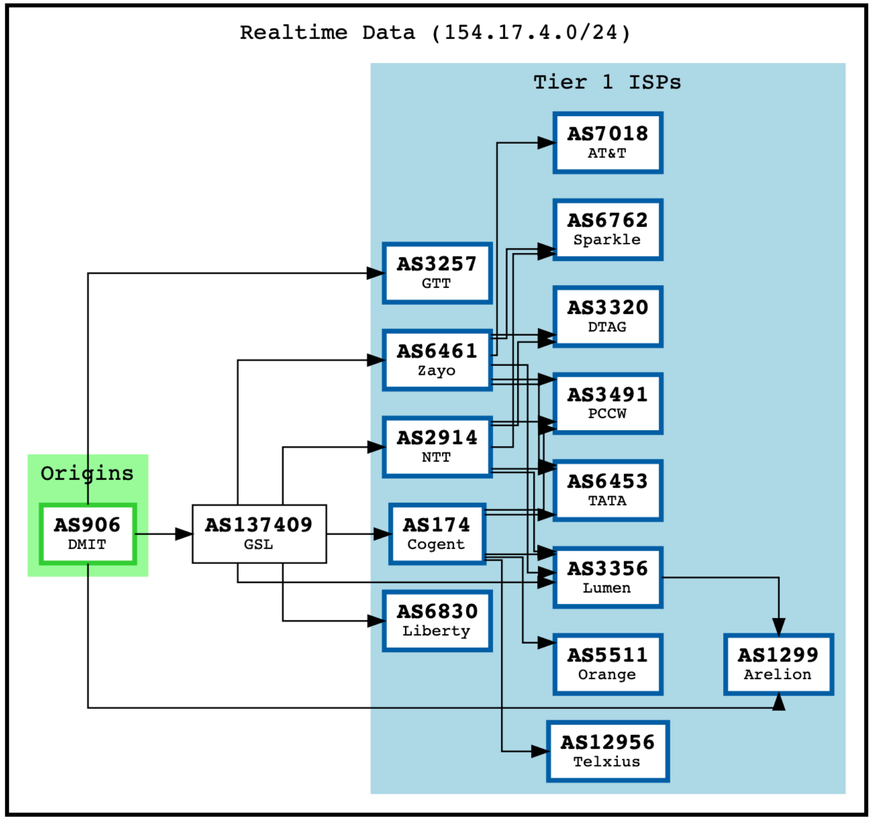
## How do ASes Connect?

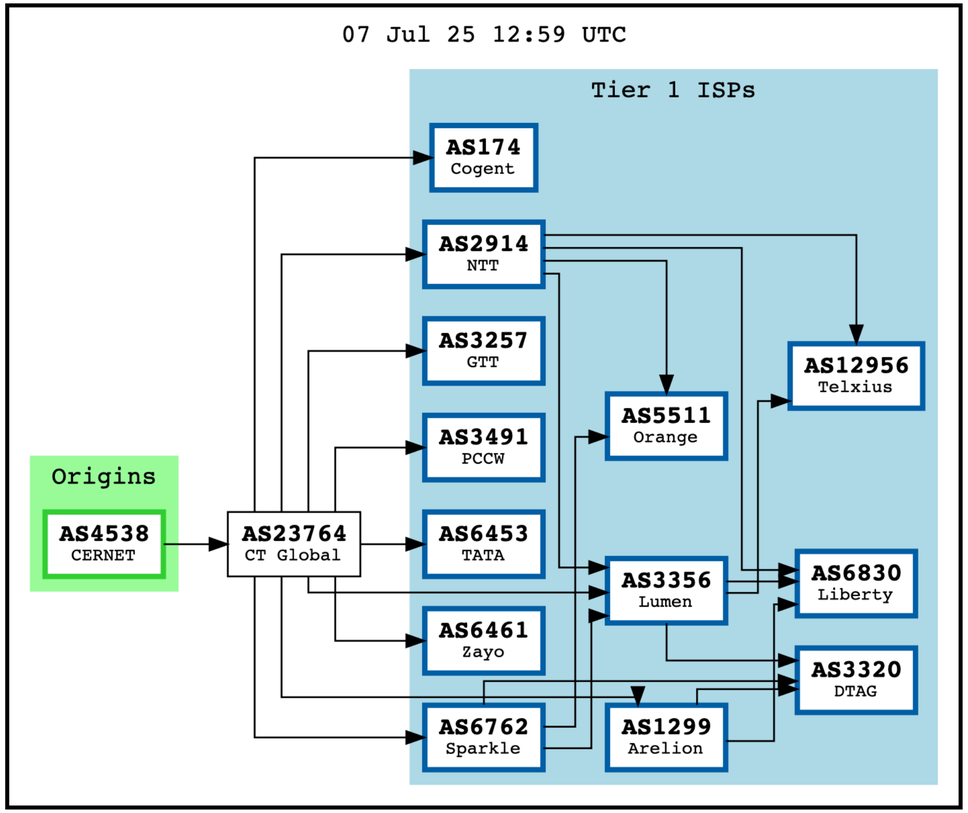



https://www.hkix.net/hkix/stat/aggt/hkix-aggregate.html

# A Network of Networks

## How do ASes Connect?




https://bgp.tools

# A Network of Networks

## How do ASes Connect?

- **T1 ISPs:** Massive global networks that do not purchase transit from any other network. They have a complete view of the Internet's routes and can reach every other network through settlement-free peering with all other Tier 1 providers. This core of the Internet is often referred to as the “Default-Free Zone” (DFZ).
- Do we always choose the shortest path? Emm… That’s complicated.

# Let’s Visualize This!

# From AS to Broader Networks

# Which Layer?

Q: ***What's the responsibility of BGP?***

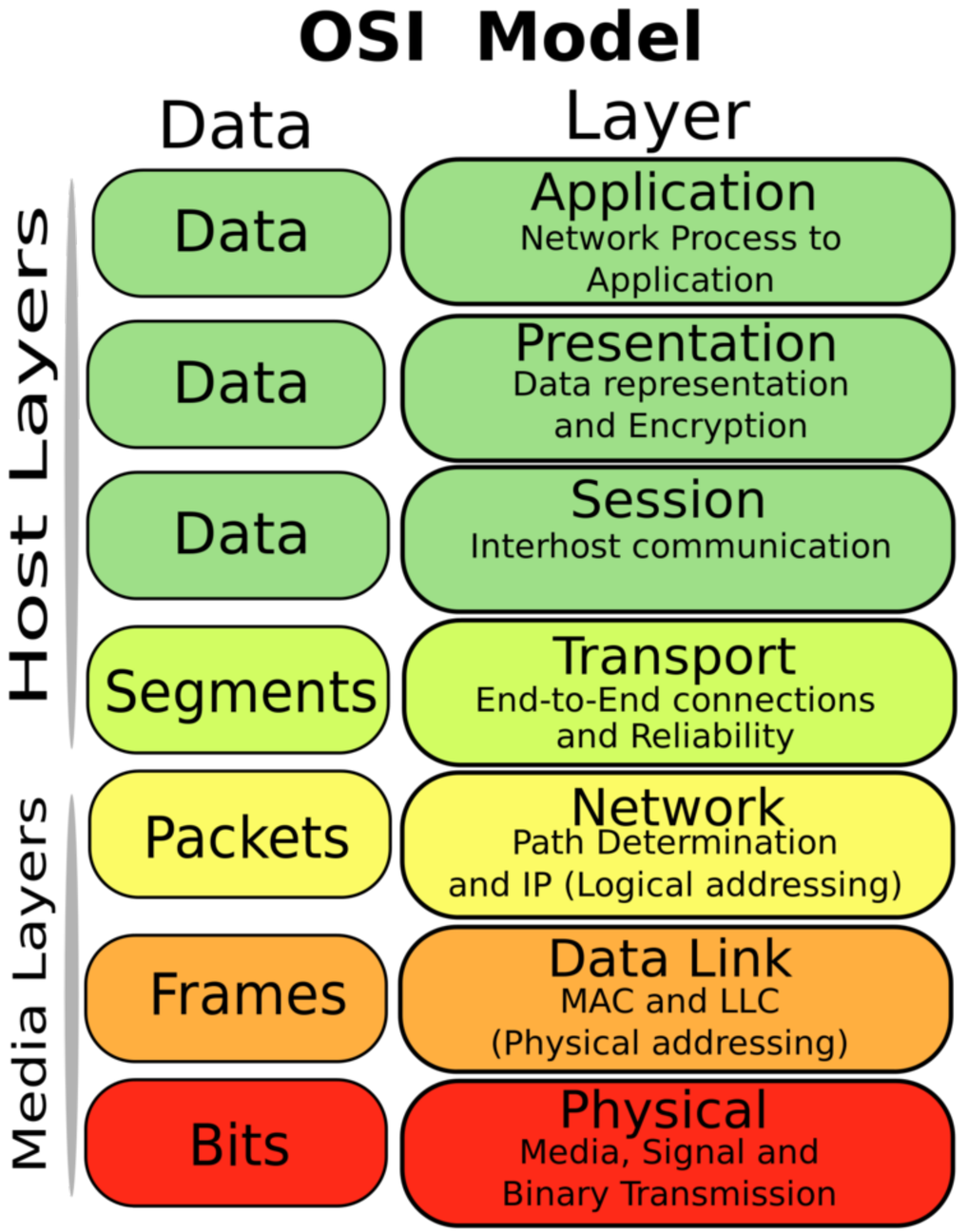
A: Exchange network routing and reachability information.

Q: ***So Layer 3 then?***

A: Not exactly…

BGP does not handle its own session reliability. Instead, it establishes a connection using the **TCP** on port 179.

By definition, any protocol that uses a Layer 4 service for its transport is an Application Layer protocol, i.e., layer 7.

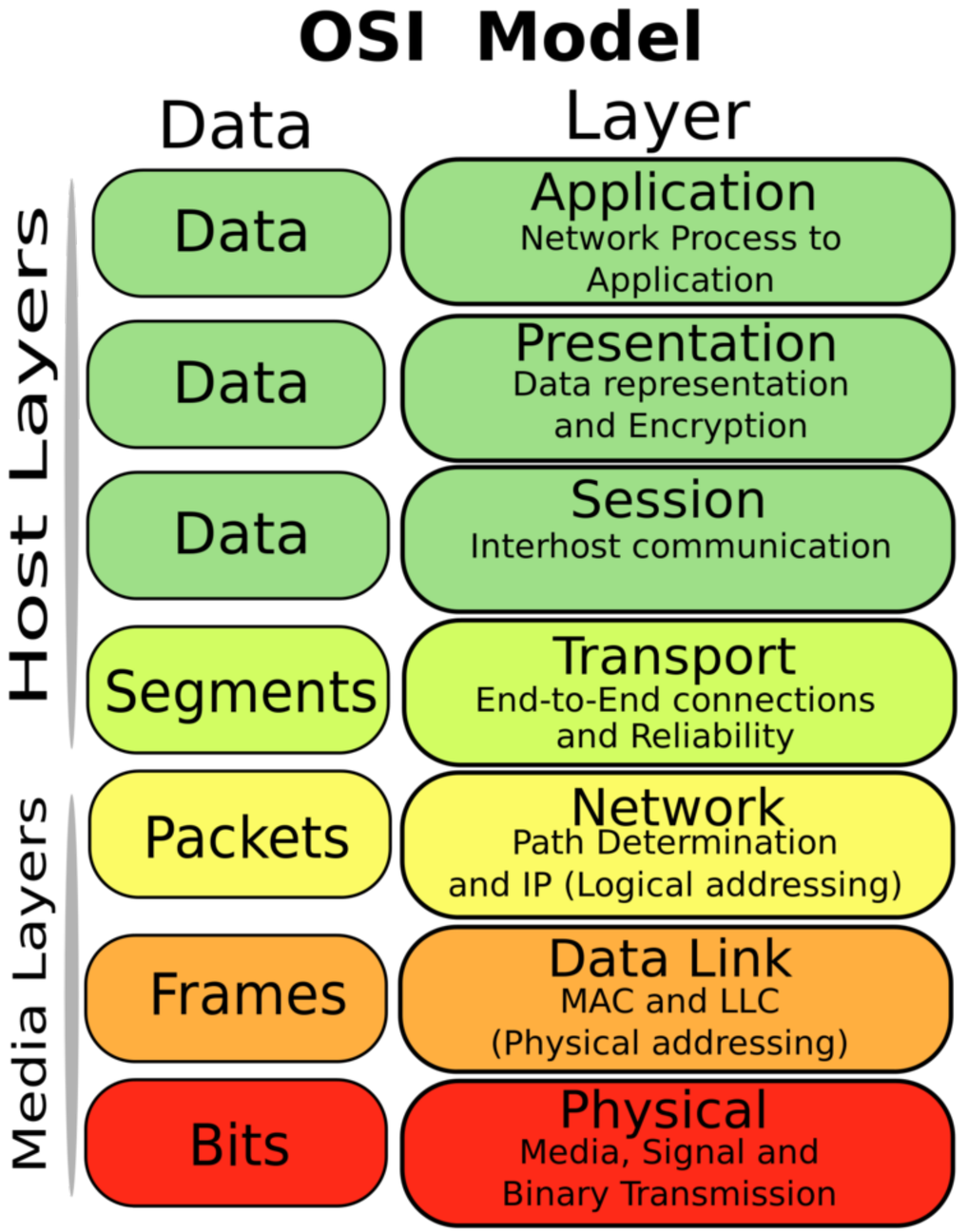


# Which Layer?

Simple notion:

Postal Offices do not deliver mails themselves!

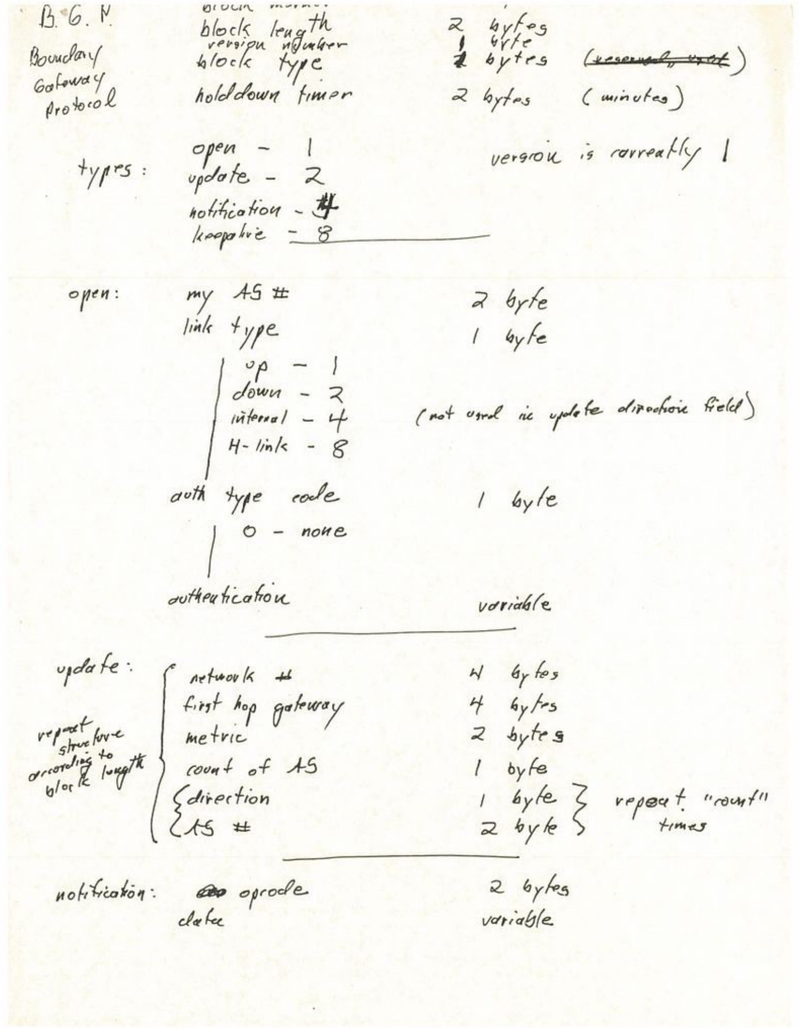
The company itself is an application, even though its product is used for navigation (routing).



# Introduction to BGP

## A “two-napkin” Protocol

The genesis of BGP was in 1989 when Kirk Lougheed, Len Bosack and Yakov Rekhter were sharing a meal at an IETF conference. They famously sketched the outline of their new routing protocol on the back of some napkins, hence often referenced to as the “Two Napkin Protocol”



# Introduction to BGP

## A “two-napkin” Protocol

Historical Background: 1989-BGP, 1995-SSL

草台班子：If your AS announced a prefix, other networks simply **believed you.** There was no central authority or automated system to check if your AS was actually the legitimate owner of those IP addresses. 😈😈😈

# Introduction to BGP

## A “two-napkin” Protocol

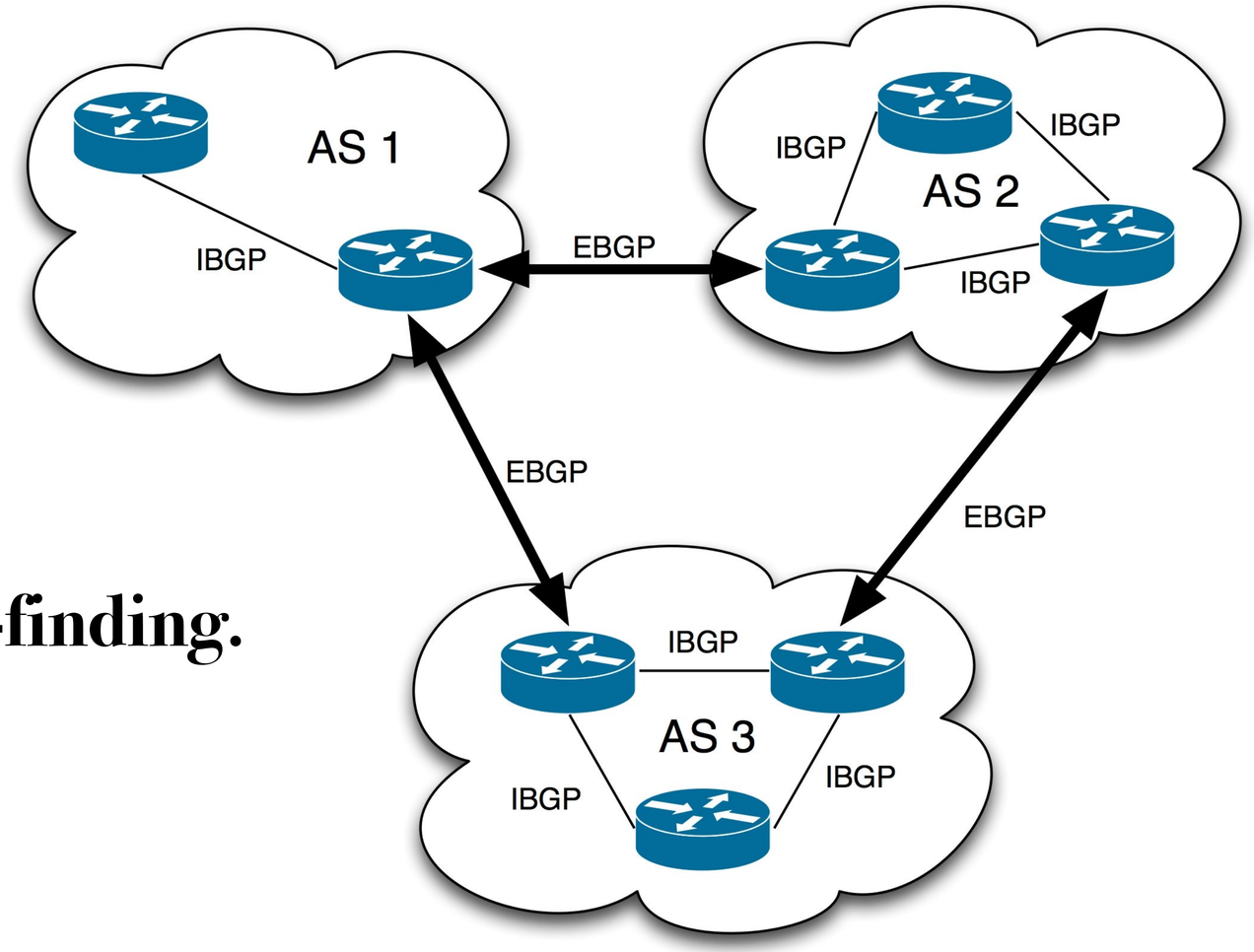
BGP is unique among major routing protocols because it runs over the TCP on the well-known (well, maybe not that well-known) port 179.

The routers that run the BGP protocol are called BGP “speakers”. These speakers establish connections, known as “sessions,” with other BGP speakers. These neighbors are referred to as “peers”.

Unlike IGPs that can often automatically discover their neighbors, BGP peerings must be manually configured by network administrators. This manual configuration is a security feature, ensuring that an AS only exchanges routing information with explicitly trusted partners.

# IGP, BGP, iBGP, eBGP

WTF? What's the difference?

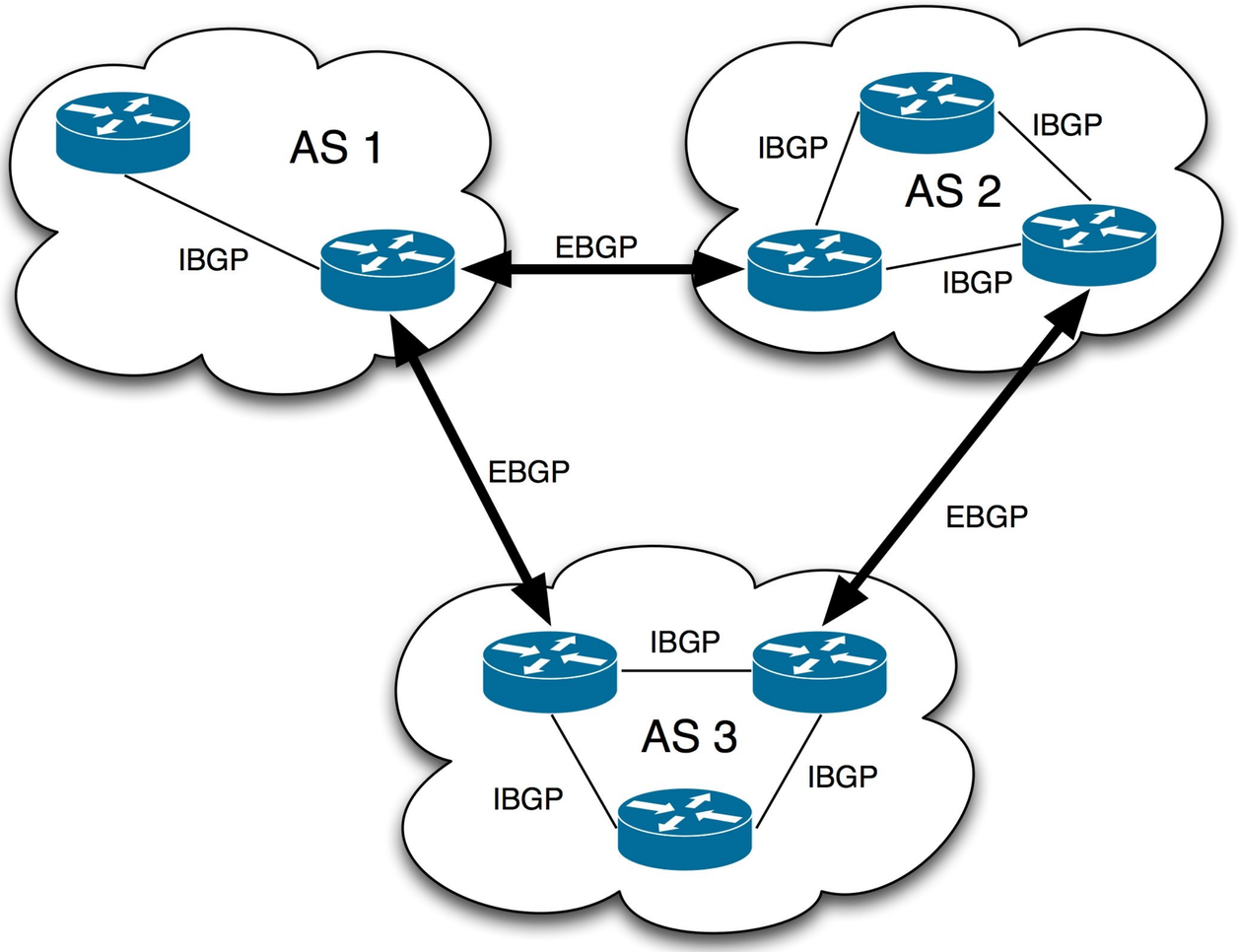



**Billions of hosts,**
**but only ≈ 900 k routers do the path-finding.**

# IGP, BGP, iBGP, eBGP

WTF? What's the difference?

IGP – Interior Gateway Protocol
• "Interior" = used only inside one AS.
• Goal: pick the shortest or cheapest path for traffic that *stays inside* the AS.
• Metric-based (cost, bandwidth, delay), converges very fast.




BGP – Border Gateway Protocol
• Policy-based: lets each AS accept, reject, prefer, or de-prefer routes according to business or traffic-engineering rules.
• Runs over TCP port 179 and easily carries the 900 k+ public Internet routes.

# IGP, BGP, iBGP, eBGP

## Design Choices in IGP

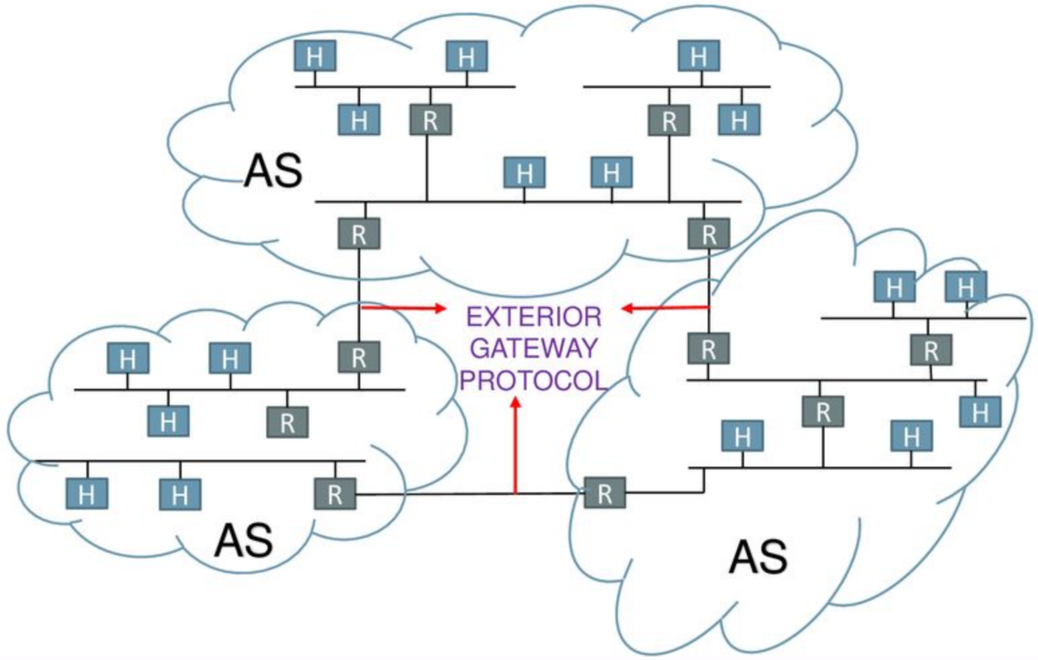
- OSPF, IS-IS: LSDB (Link-State DataBase) → all the links and costs in the area. Dijkstra.
- RIP/EIGRP: distance-vector tables. Bellman-Ford.

Why some dijkstra, some bellman-ford?

Dijkstra (Link-State) - “Global View”

Router A’s Database (Same on All Routers)

```
Link B-C: cost 5
Link A-C: cost 20
Link C-D: cost 15
```




Bellman-Ford (Distance-Vector) - “Partial View”

Router B’s Database (Ask my neighbours )

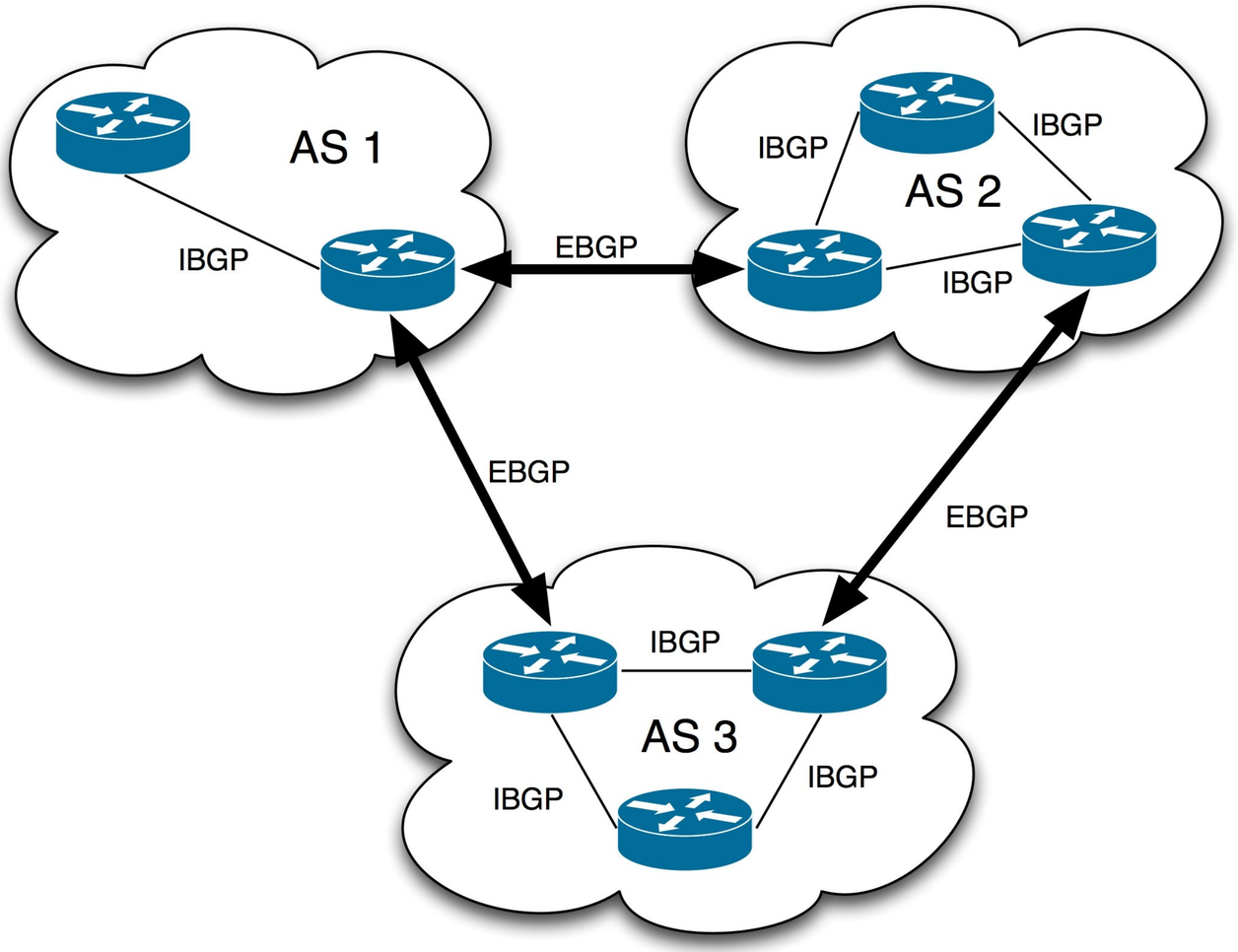
```
To reach Net X: via B, cost 15
To reach Net Y: via C, cost 8
To reach Net Z: via B, cost 22
```

# IGP, BGP, iBGP, eBGP

WTF? What's the difference?

IGP – Interior Gateway Protocol
• "Interior" = used only inside one AS.
• Goal: pick the shortest or cheapest path for traffic that *stays inside* the AS.
• Metric-based (cost, bandwidth, delay), converges very fast.

BGP – Border Gateway Protocol
• Policy-based: lets each AS accept, reject, prefer, or de-prefer routes according to business or traffic-engineering rules.
• Runs over TCP port 179 and easily carries the 900 k+ public Internet routes.



# BGP Message Types

1. **OPEN:** Used to initiate a BGP session.
2. **UPDATE:** Advertise new routes, withdraw previously advertised routes, and communicate the attributes associated with those routes (such as the AS path). After the initial exchange of the full routing table, BGP sends **only incremental updates as the network changes**, making it more efficient than protocols that require periodic full refreshes.
3. **KEEPALIVE:** Small, 19-byte messages sent periodically (every 30 or 60 seconds) to confirm that the peer is still active and the TCP session is healthy. If a router doesn't receive a KEEPALIVE message from its peer, it assumes the peer has failed and tears down the BGP session.
4. **NOTIFICATION:** Indicate an error condition. After sending a NOTIFICATION message, the BGP speaker immediately closes the TCP connection.

# Decision Tree

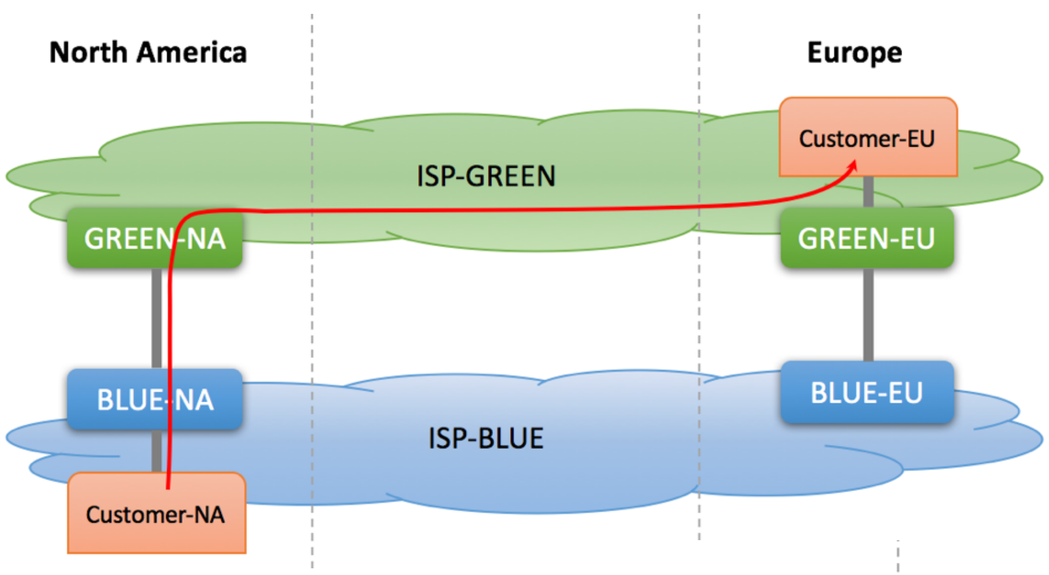
| Step | Attribute | Preference | Scope | Description & Analogy |
|---|---|---|---|---|
| 1 | **Weight** | Highest | Local to Router | A "sticky note" the local admin puts on a route. "I like this one best, regardless of what anyone else thinks." It's a purely local instruction, not shared with other routers. A higher value wins. |
| 2 | **LOCAL_PREF** | Highest | AS-wide (iBGP) | The company-wide policy for outbound traffic. "All our offices should use ISP A to get to this destination." It's communicated to all iBGP peers within the AS. Higher is better; default is 100. |
| 3 | **Locally Originated** | Yes > No | Local to Router | "I trust routes I created myself (via network or redistribute commands) more than routes I learned from others." A router will always prefer to use a path it injected into BGP itself. |
| 4 | **AS_PATH Length** | Shortest | Global (eBGP) | The default "shortest trip" metric. A path traversing two ASes is better than one traversing four. This is the most intuitive step and a primary factor in global routing. |
| 5 | **Origin Code** | IGP < EGP < Incomplete | Global (eBGP) | A measure of how the route first entered BGP. 'i' (IGP, from a network command) is most trusted. 'e' (EGP, an obsolete protocol) is next. '?' (Incomplete, from redistribution) is least trusted. |
| 6 | **MED** (Multi-Exit Discriminator) | Lowest | Inter-AS (eBGP) | A suggestion to your neighbor. "If you have multiple paths to reach me, please use this one; it's cheaper/faster on my end." The neighbor can choose to ignore this hint. Lower is better. |
| 7 | **Neighbor Type** | eBGP > iBGP | Local to Router | Prefer routes learned from external partners (eBGP) over routes learned from internal colleagues (iBGP). External routes are considered closer to the source. |
| 8 | **IGP Metric to Next Hop** | Lowest | AS-wide (iBGP) | If there are multiple internal paths to the same exit router, choose the one with the lowest internal cost (e.g., OSPF cost). |
| 9 | **Oldest Path** | Oldest | eBGP only | For stability, if all else is equal between two external paths, stick with the path you've known the longest to minimize route flapping. |
| 10 | **Router ID** | Lowest | Tie-breaker | An arbitrary tie-breaker. Prefer the path from the neighbor with the numerically lowest BGP Router ID. |
| 11 | **Neighbor IP Address** | Lowest | Tie-breaker | The final, arbitrary tie-breaker. If router IDs are the same, prefer the path from the neighbor with the numerically lowest source IP address used in the TCP session. |

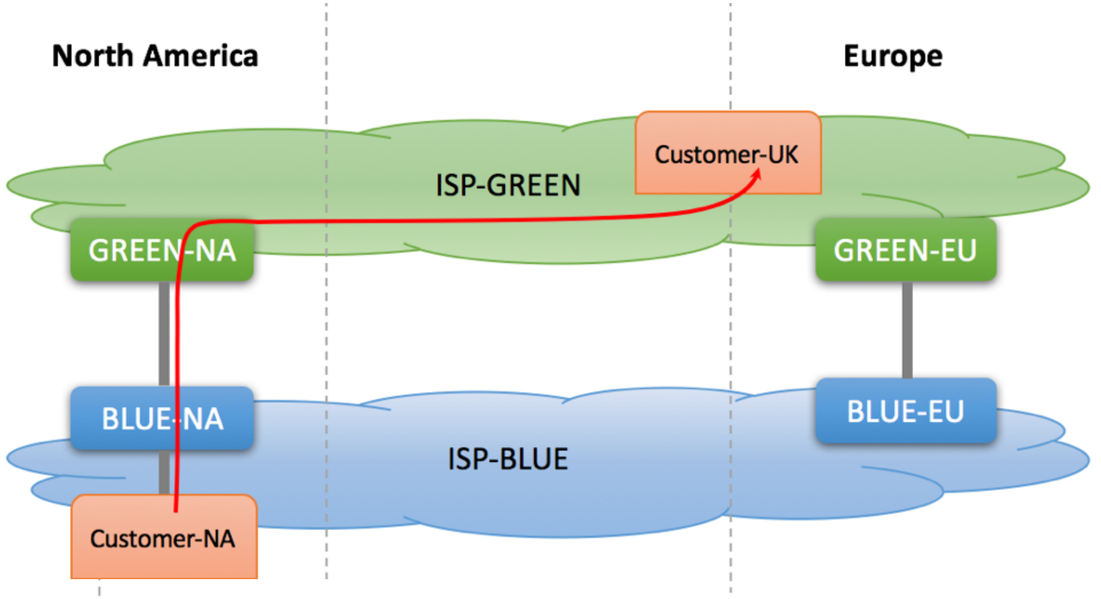
# Decision Tree

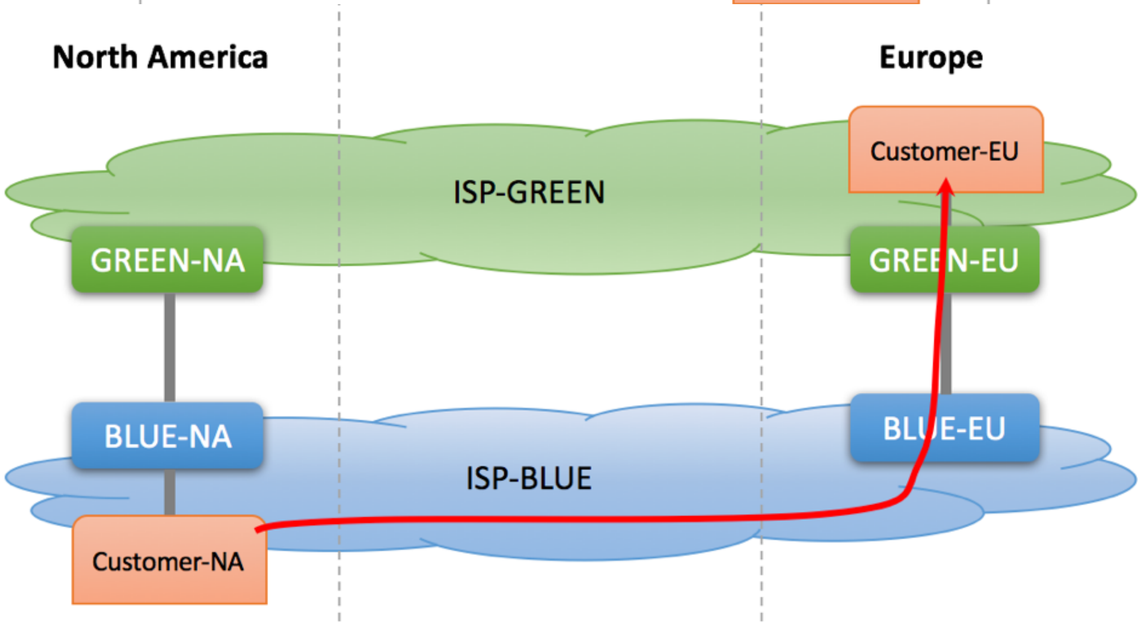
**LOCAL_PREF**

**MED & AS-PATH**

# Hot v.s. Cold


North America
Europe
ISP-GREEN
Customer-EU
GREEN-NA
GREEN-EU
BLUE-NA
BLUE-EU
ISP-BLUE
Customer-NA

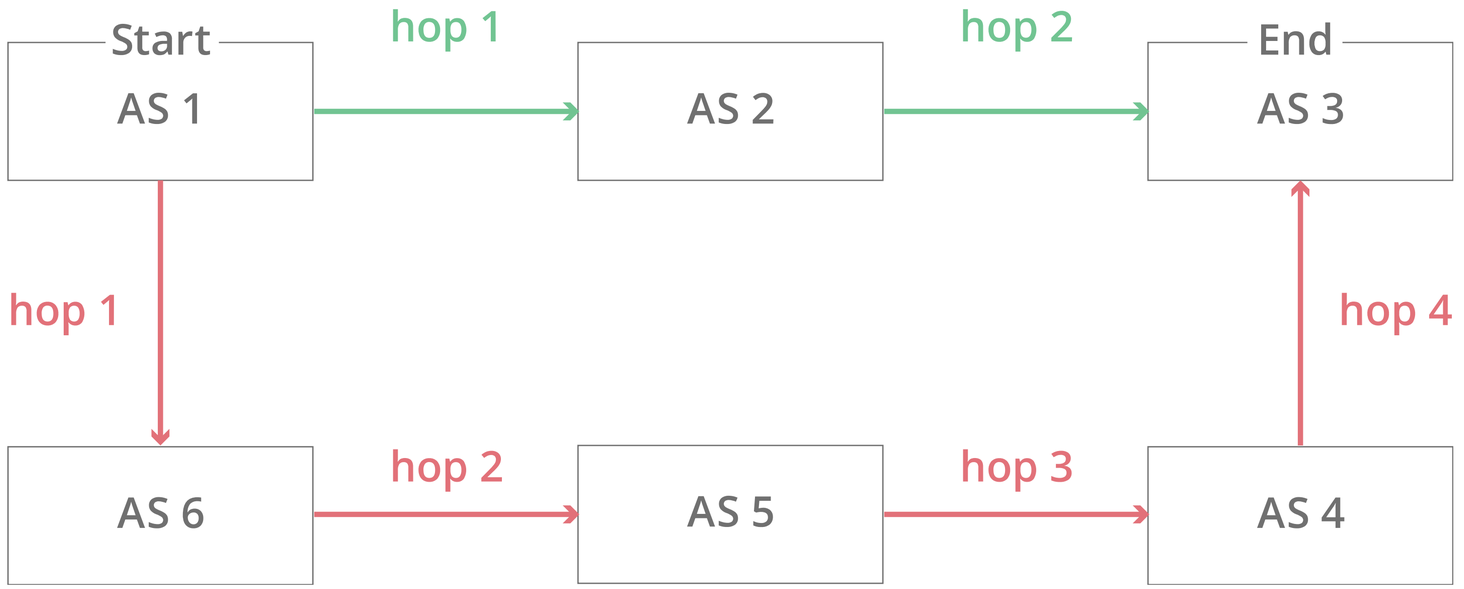

North America
Europe
ISP-GREEN
Customer-UK
GREEN-NA
GREEN-EU
BLUE-NA
BLUE-EU
ISP-BLUE
Customer-NA


North America
Europe
ISP-GREEN
Customer-EU
GREEN-NA
GREEN-EU
BLUE-NA
BLUE-EU
ISP-BLUE
Customer-NA

# Decision Tree

Option 1:

Option 2:




https://www.cloudflare.com/img/learning/security/glossary/what-is-bgp/bgp-simplified.svg

“Border Gateway Protocol is the postal service of the Internet. When someone drops a letter into a mailbox, the Postal Service processes that piece of mail and chooses a **fast**, **efficient** route to deliver that letter to its recipient.”

https://www.cloudflare.com/learning/security/glossary/what-is-bgp/

“BGP’s goal is not necessarily to find the *fastest* path, but to find the best path that adheres to the business and operational policies of every AS along the way.”

**Google Gemini 2.5 Pro Deep Research**

# Yet Another Visualization.

# From Unicast to Anycast

**Unicast:** one-to-one mapping between an IP address and a single network interface on a specific device. When you send a packet to a unicast IP, there is only one correct destination for it on the entire Internet.

**Anycast:** single IP address is assigned to and advertised by multiple servers. “one-to-one-of-many” or “one-to-nearest” relationship.

# Magic Behind Anycast

**It’s just BGP!**

1. **BGP Announcements:** The BGP router at each of these locations announces a route to this shared IP address to its peers, propagating the advertisement into the global Internet.
2. **Standard Path Selection:** A user's local router, upon receiving a request for 8.8.8.8, will now see multiple paths to this same prefix from its various BGP peers.
3. **Topological Proximity:** The router then executes its standard BGP Best Path Selection Algorithm.
4. **One-to-Nearest Routing:** The result is that the user’s traffic is automatically forwarded along the path with the fewest AS hops.

# Real-World Applications

1. CDN
2. DNS
3. DDoS Mitigation
4. 😴 😴 😴
5. How to choose your own VPS? 🤓 🤓 🤓
6. Can I have my own ASN? 😋 😋 😋

# Real-World Applications

1. How to choose your own VPS?
   

China unicom中国联通
让一切自由连通
世界可及
只要主干网活着

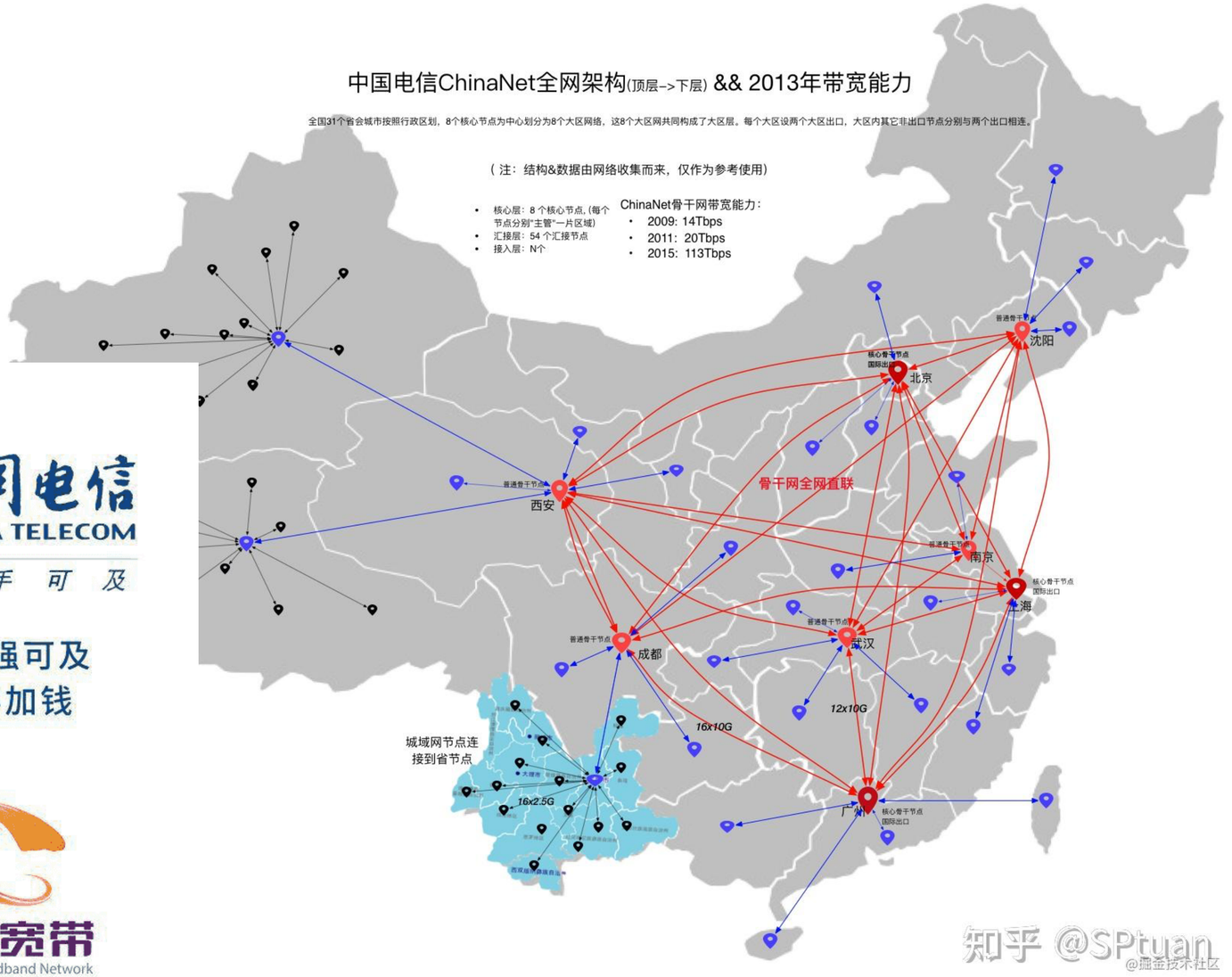
中国移动
China Mobile
世界不可及
加钱也不行，除非借道

中国电信
CHINA TELECOM
世界触手可及
世界勉强可及
想快得加钱

CERNET
世界不一定可及
借道也不一定行

中国科技网
China Science & Technology Network
世界可及
就是接不到

长城宽带
Great Wall Broadband Network
UNKNOW
ERROR



# Real-World Applications


1. How to choose your own VPS? 🤓 🤓 🤓

在校园内： CERNET接入10Gbps的HKIX

交大WIFI：中国联通IPv4出口/CERNET IPv6

交大有线网：CERNET IPv4 & IPv6

选购建议：一般来说只要接入了HKIX，且有IPv6的香港机器延迟都不错。(~60ms，出口在清华)

Budget：Free ~ 10USD/y

# Real-World Applications


1. How to choose your own VPS? 🤓 🤓 🤓

三大运营商：精品网络

中国电信：AS4809 CN2 (CN2 GT, CN2 GIA), AS23764 (CTG GIA, CTGNET)

中国联通：AS9929 CNC Backbone, AS10099 CUG Backbone

中国移动：AS58807 CMIN2

对于香港而言，只要能直连的机器一般都不错（贵，~25USD/mo per 1Mbps 优化带宽）

Budget: 美国优化线路 25~50 USD/y，香港优化 40 ~ 100USD/y

# Real-World Applications


1. How to choose your own VPS? 🤓 🤓 🤓

歪门邪道：利用你服务商的T1 ISP接入

E.g. 中国移动国际 AS58453 在香港接入 Level 3 (Lumen) 的 Transit。

所以香港接入 Lumen 的机器到中国移动的网络表现还可以。坏处：只有移动网络很快乐。

Budget: ~10USD/y

更加有钱的富哥——跨境专线（>100USD/y）+ 实名认证

# Real-World Applications

1. How to choose your own VPS? 🤓 🤓 🤓


**Let’s Look at Traceroute of Each Network!**

# Real-World Applications

Can I have my own ASN? 😋 😋 😋

In short, yes! And it’s not that (well, merely affordable) expensive actually.

* https://blog.lyc8503.net/post/asn-1-asn-registration/

目前地球上直接负责管理 IP 地址和 ASN 分配的机构是区域互联网注册管理机构 - Regional Internet registry, RIR. RIR 由 IANA 管理, IANA 又由 ICANN 管理.(好乱)

RIR 目前有 ARIN, RIPE NCC, APNIC, LACNIC, AfriNIC 这五家.

# Real-World Applications

Can I have my own ASN? 😋 😋 😋

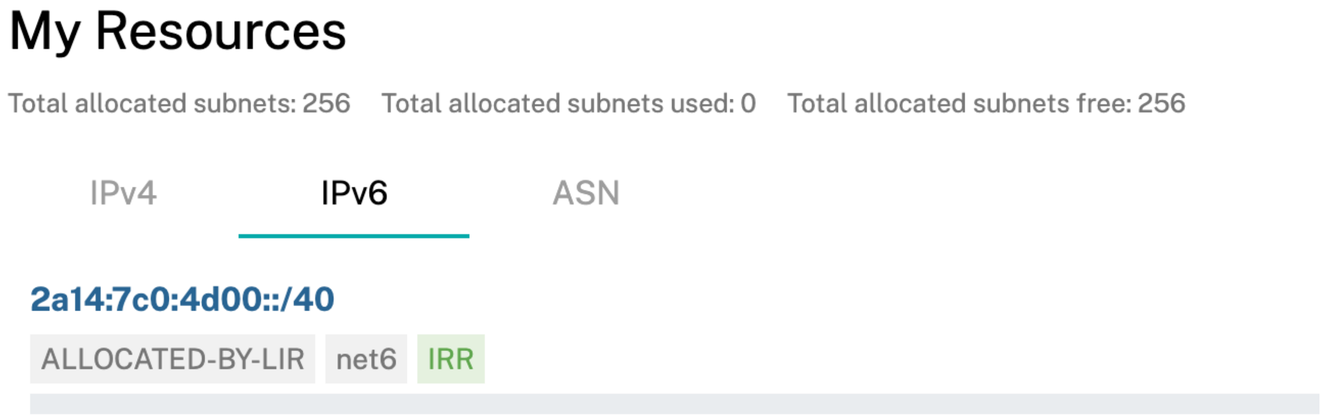
如果你需要直接向 RIPE NCC 申请资源的话, 需要先成为 RIPE NCC 的 member, 每年缴纳 €1,550 的会员费. 这对大多数人来说显然太贵了, 但 RIPE NCC 允许自己的 member 代他人申请, 此时 member 又被称为 *Local Internet registry, LIR*. 我们可以直接找一个交了会员费的 LIR, 请他帮我们申请需要的 ASN 和 IP 地址段. 所以你需要为你的 ASN 给 LIR 至少支付 **€50**/年的年费 + **LIR** 的服务费用, 具体价格咨询你选定的 LIR.

# Real-World Applications

Can I have my own ASN? 😋 😋 😋

1. 个人身份证或公司注册文件 (身份证后续需要在 iDenfy 在线验证)
2. 近期欧洲范围内支持 BGP 的服务商的发票一份 (BuyVM / Vultr 买台虚拟机即可)
3. 联系邮箱及通信地址 (联系邮箱还挺重要的, 之后很多验证会用到, 地址应该无所谓)
4. 两个你的上游 ASN (可以随便写欧洲范围内支持 BGP 的服务商, 其中之一可能要和你提供的发票一致)

# Real-World Applications




## Can I have my own ASN? 😋 😋 😋

得到 ASN 和 IP 地址段之后还有一个步骤: 设置 RPKI 和 ROA.

你需要通过 RPKI 上的 ROA (路由源授权) 来**将你的 IP 段授权给你的 AS 使用**.
目前, 部分上游已经开始强制要求基于 RPKI 的验证, 所以我们现在就来完成这一步, 避免之后的麻烦.

你得到的 IP 资源分为 PA (Provider Aggregatable) 和 PI (Provider Independent) 两种. 对于 LIR 赠送的或者你主动购买的 IP 地址资源, 它一般属于 PA 资源, PA 资源要修改 RPKI 的话, **直接联系 LIR 即可**.

所以直接联系 LIR 说明: 为 xxx IP 段添加一个 ASxxxx 的 ROA, LIR 应该就会帮你设置好了.

之后如果有多个 ASN / IP 段的话, 需要修改 ROA 也要联系 LIR.

接下来还需要创建route6对象。

和 RPKI 类似的, route6 对象也是 AS 和 IP 段之间的对应关系, 有些上游会要求. 这个需要我们自己在 RIPE Database 中创建.

登陆 RIPE Database, 选择 Create an Object, 类型选择 route6, 直接填入自己的 AS 和 IP 段即可.

Thanks!

Questions?