Kết nối nhiều backbone (AS) với BGP

Phạm Huy Hoàng
SoICT/HUST
hoangph@soict.hust.edu.vn

ONE LOVE. ONE FUTURE.

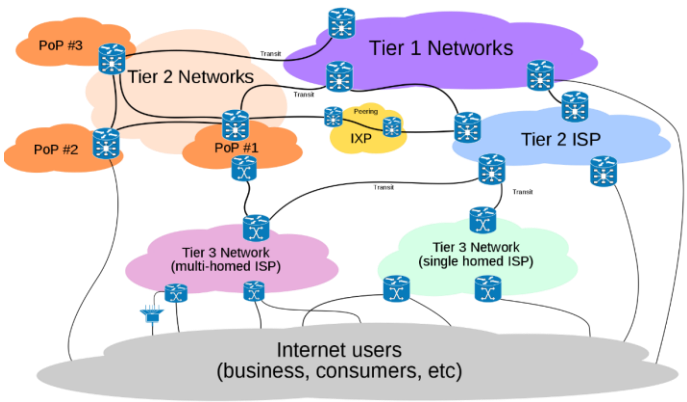
1

Nội dung

- BGP Introduction
- Hoạt động chung của BGP
- Hiện trạng Internet backbone BGP
- BGP packet format
- BGP routing policy
- Hoạt động của router BGP
- eBGP và iBGP
- Thực hành

Kiến trúc Internet với các AS

- Internet backbone
- Các mạng tiers (AS)
- Kết nối các AS
  - Peering
  - Downlink /Uplink
- Vai trò của router
- Bên trong AS: IGP
- Kết nối các AS: EGP (BGP)



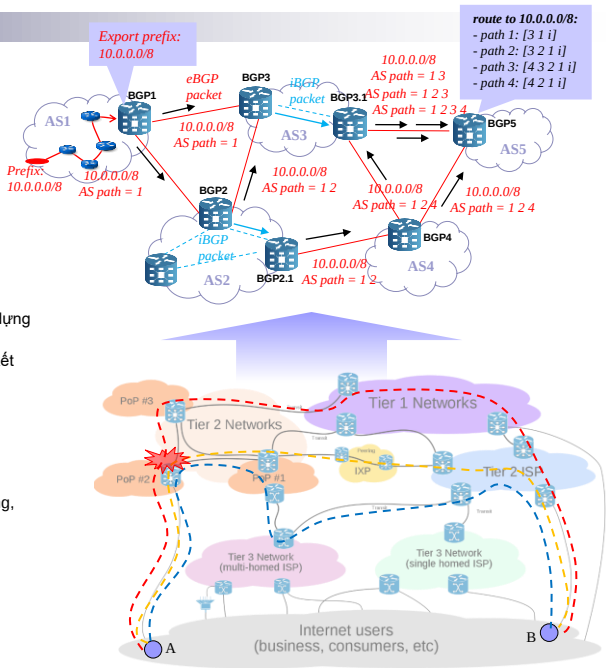
BGP Introduction

- Border Gateway Protocol: routing protocol giữa các AS
- Là "chất keo" kết dính toàn bộ hệ thống Internet hiện nay
- Ra đời năm 1994, phiên bản hiện nay là version 4 chuẩn hóa năm 2006 (RFC4271[1])
- Hỗ trợ CIDR
- Định tuyến theo policy, hơn là theo đường đi ngắn nhất
- "Đơn vị" routing là Autonomous System - tìm đường đi theo các kết nối giữa các AS
- Kết hợp với các IGP (RIP, OSPF, v.v..) trong AS để tạo nên giải pháp dynamic routing hoàn chỉnh trên toàn bộ hệ thống Internet

[1] https://tools.ietf.org/html/rfc4271

Hoạt động chung của BGP

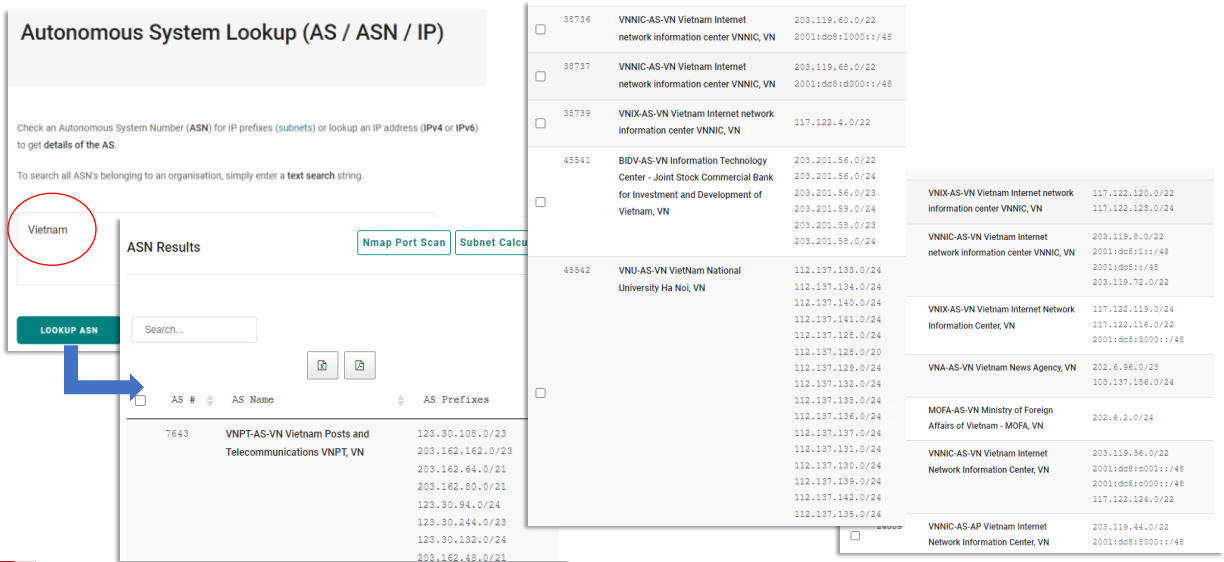
- Dựa trên hoạt động láng giềng: 2 loại láng giềng BGP
  - BGP láng giềng được khai báo (cấu hình)
  - Các BGP bên trong một AS được khai báo là láng giềng của nhau.
  - Giữa 2 AS: BGP router gửi message trực tiếp cho nhau ➔ eBGP
  - Bên trong một AS: BGP router gửi message dựa trên các IGP ➔ iBGP
- BGP speaker & propagation process:
  - Sử dụng kênh TCP (cổng 179) để kết nối láng giềng
  - BGP hoạt động giống RIP ở khâu quảng bá các network prefix và xây dựng đường routing về gốc
  - Loan báo (speak) các BGP packet đến láng giềng cung cấp khả năng kết nối (reachability) đến một network được "export"
  - Cập nhật AS path để đi về "exported network" mỗi khi đi qua eBGP
- Quyết định lựa chọn AS-Path theo Policy
  - Không nhất thiết là đường đi ngắn nhất
  - BGP cho phép xác định nhiều AS-Path để route từ A đến B
  - Chọn AS-Path nào là do các mạng Tier áp dụng policy riêng (băng thông, kinh tế, chính trị, v.v..)
- AS number (được gán cho AS theo thủ tục đăng ký)
  - 16 bit nhị phân. Dải number 64512-65535 được quy hoạch cho private
  - Check AS number with Prefixes: https://hackertarget.com/as-ip-lookup/
  - FPT: AS18403 (FPT-AS-AP FPT Telecom Company, VN)
  - Vietel: AS7552 (VIETEL-AS-AP Viettel Group, VN)
  - Route = AS path (AS1, AS2, AS3, v.v…)



5

Check AS number with Prefixes

https://hackertarget.com/as-ip-lookup



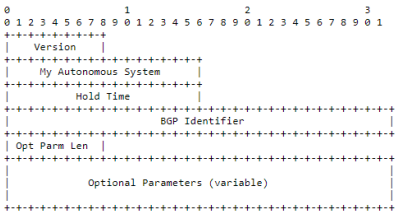
SOICT TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
School of Information and Communication Technology

6

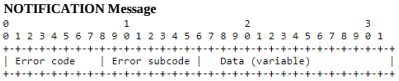
BGP packet format

- Cấu trúc: *<Marker(16-octet), Length(2-octet), Type(1-octet)><Data>*
  - Type: 1 - OPEN, 2 - UPDATE, 3 - NOTIFICATION, 4 - KEEPALIVE
- OPEN: được gửi từ mỗi BGP router ngay khi kết nối TCP được thiết lập. KEEPALIVE được trả về nếu router chấp nhận kết nối.
  - *My Autonomous System:* AS number
  - *Hold Time:* Timeout, tính bằng giây. Duy trì liên kết khi chưa nhận được KEEPALIVE hoặc UPDATE hoặc NOTIFICATION.
  - *BGP Identifier:* tương tự Router ID trong OSPF (admin cấu hình hoặc lấy theo IP)
- KEEPALIVE: không có data, chỉ có header với Type=4. Được gửi trước khi Hold time hết hạn, để thông báo duy trì kết nối
- NOTIFICATION: dùng để thông báo lỗi
- UPDATE: quảng bá các đường route có khả năng được thiết lập giữa các router BGP
  - Có thể được sử dụng để thông báo hủy bỏ một số route:
    - *Withdrawn Router Length:* độ dài trường Withdrawn Routers ngay sau
    - *Withdrawn Routers:* danh sách các route cần hủy - *List<length, network prefix>*
  - Cùng một UPDATE message có thể thông báo hủy một số route và thêm một số route có thể được bổ sung hoặc cập nhật:
    - *Total Path Attribute Length:* độ dài trường Path Attributes ngay sau
    - *Path Attributes:* danh sách dạng *List<attribute type, attribute length, attribute value>*
    - *Network Layer Reachability Informartion:* danh sách địa chỉ IP mạng (cũng dưới dạng bộ đôi *List<length, network prefix>*) mà BGP router có thể route đến (độ dài trường này được tính bằng độ dài gói tin trừ đi độ dài các trường trên)

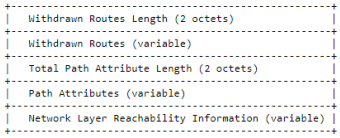
OPEN Message



NOTIFICATION Message



UPDATE Message



BGP Policy: Decision Process[1]

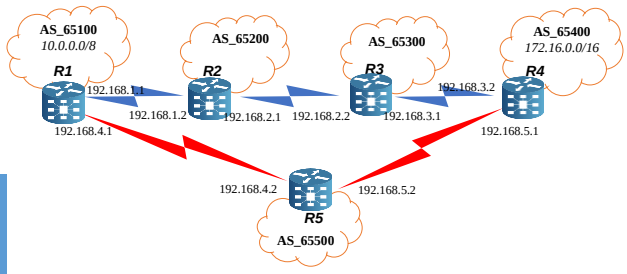
- 1. Weight check: prefer higher local weight routes to lower routes.
- 2. Local preference check: prefer higher local preference routes to lower.
- 3. Local route check: Prefer local routes (statics, aggregates, redistributed) to received routes.
- 4. AS path length check: Prefer shortest hop-count AS_PATHs.
- 5. Origin check: Prefer the lowest origin type route. That is, prefer IGP origin routes to EGP, to Incomplete routes.
- 6. MED check: Where routes with a MED were received from the same AS, prefer the route with the lowest MED. See BGP MED.
- 7. External check: Prefer the route received from an external, eBGP peer over routes received from other types of peers.
- 8. IGP cost check: Prefer the route with the lower IGP cost.
- 9. Multi-path check: If multi-pathing is enabled, then check whether the routes not yet distinguished in preference may be considered equal. If bgp bestpath as-path multipath-relax is set, all such routes are considered equal, otherwise routes received via iBGP with identical AS_PATHs or routes received from eBGP neighbours in the same AS are considered equal.
- 10 Already-selected external check: Where both routes were received from eBGP peers, then prefer the route which is already selected. Note that this check is not applied if bgp bestpath compare-routerid is configured. This check can prevent some cases of oscillation.
- 11. Router-ID check: Prefer the route with the lowest router-ID. If the route has an ORIGINATOR_ID attribute, through iBGP reflection, then that router ID is used, otherwise the router-ID of the peer the route was received from is used.
- 12. Cluster-List length check: The route with the shortest cluster-list length is used. The cluster-list reflects the iBGP reflection path the route has taken.
- 13. Peer address: Prefer the route received from the peer with the higher transport layer address, as a last-resort tie-breaker.

[1] BGP Decision Process: https://tools.ietf.org/html/rfc4271#section-9.1

thực hành: BGP cơ bản

- Các BGP router R1, R2, R3, R4 nối serial với nhau để kết nối các AS 65100, 65200, 65300, 65400
- Tại AS 65100 có mạng 10.0.0.0/8 và tại AS 65400 có mạng 172.16.0.0/16
- Sau khi chạy BGP, các mạng trên đã xuất hiện trong các routing table của tất cả các router

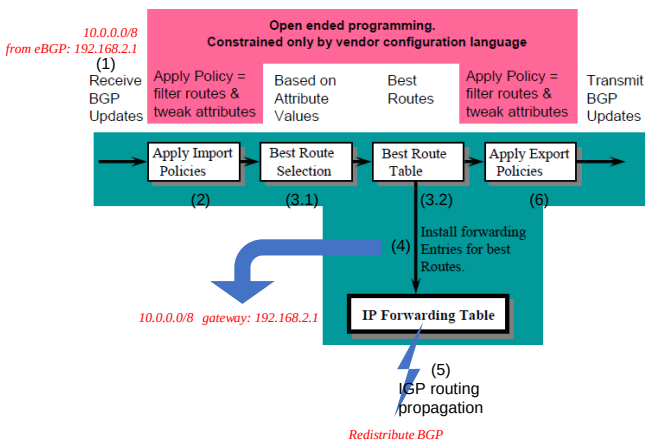
```
R2> route -n
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
10.0.0.0        192.168.1.1     255.0.0.0       UG    20     0        0 enp0s8
172.16.0.0      192.168.2.2     255.255.0.0     UG    20     0        0 enp0s9
192.168.1.0     0.0.0.0         255.255.255.0   U     0      0        0 enp0s8
192.168.2.0     0.0.0.0         255.255.255.0   U     0      0        0 enp0s9
```



- Bổ sung AS 65500 với BGP R5 nối R1 & R4. Đường đi từ R1 đến 172.16.0.0/16 được đổi sang R5: do policy chọn AS path bé nhất.
- Thay đổi weight của kết nối R1-R2: 101, kết nối R1-R5: 55. Đường đi được chuyển sang qua R2: do policy chọn weight lớn nhất có mức ưu tiên cao hơn policy AS path nhỏ nhất

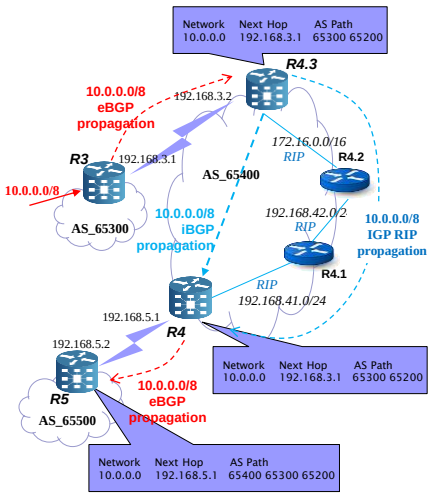
Xử lý bên trong router BGP

1. Kết nối TCP láng giềng và nhận gói tin BGP (chứa thông tin export một mạng nội bộ nào đó) ➔ incoming Routing Information Base (RIB-in)
2. Kiểm tra import policy để quyết định có xử lý update RIB-in hay không
3. Xử lý BGP routing process:
   1. Trích xuất thông tin từ RIB-in để cập nhật AS path từ BGP router hiện tại đến mạng nội bộ mà đang được lan tỏa trong RIB-in.
   2. Lựa chọn AS path "phù hợp nhất theo policy", tính toán next hop và các thông số khác để chuẩn bị cập nhật bảng routing. Lưu ý là "next hop" của AS path khác với next hop đường định tuyến (trường hợp iBGP)
4. Cập nhật tuyến đường đã chọn trong bước 3 cùng với next hop và các thông số liên quan (ví dụ RIP metric hay OSPF cost) vào bảng định tuyến để áp dụng cho giao thức IP – routed protocol
5. Kích hoạt tiến trình lan truyền IGP để quảng bá đến các router nội bộ AS cập nhật thông tin tuyến đường mới vừa được BGP đưa vào bảng định tuyến
6. Áp dụng export policy để xác định RIB-out, phục vụ update cho BGP láng giềng kế tiếp



eBGP và iBGP

- IGP (RIP)
  - Loan báo thông tin network lần lượt qua các router đôi một láng giềng
  - Kết nối routing giữa các network
  - Routing next hop = router
  - RIP dùng đường loan báo xác định "khoảng cách" và đường routing (bằng cách đi ngược lại đường loan báo) ➔ router RIP láng giềng trực tiếp
- BGP
  - Loan báo thông tin network lần lượt qua các router đôi một láng giềng
  - Kết nối routing giữa các AS
  - Routing next hop = AS
  - BGP loan báo trên kênh kết nối TCP cổng 179 để thiết lập AS path ➔ không cần láng giềng kết nối trực tiếp
  - BGP giữa 2 AS: kết nối trực tiếp ➔ eBGP
  - BGP bên trong AS: kết nối gián tiếp (sử dụng IGP nội vùng AS để chuyển gói tin giữa 2 router BGP ➔ iBGP

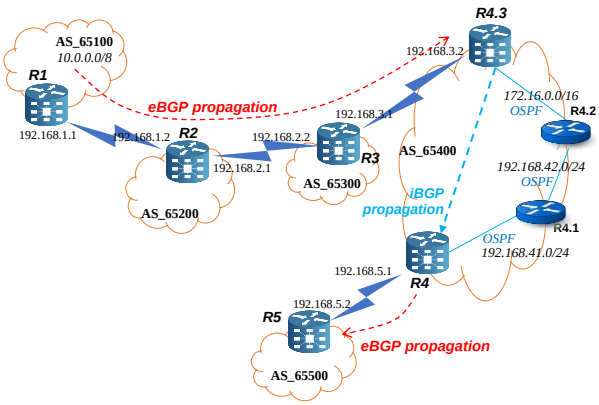


- ❑ eBGP và iBGP đều là giao thức BGP
- ❑ Xử lý phù hợp với môi trường inter-AS và intra-AS
- ❑ Routing inter-AS: next hop (BGP) = next hop (routing)
- ❑ Routing intra-AS: next hop (BGP) != next hop (routing)
- ❑ BGP routing qua một AS (AS_65400 trong hình vẽ) có khả năng không thành công nếu lan tỏa BGP đi trước IGP: thông tin routing cho mạng 10.0.0.0/8 được cập nhật trong BGP R5 trong khi các IGP trong AS 65400 chưa kịp lan tỏa thông tin về mạng này ➔ vấn đề synchronization!

11

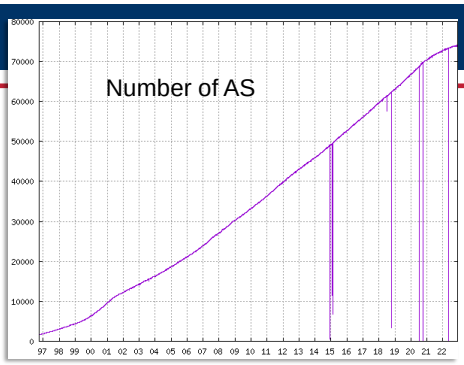
thực hành: eBGP & iBGP

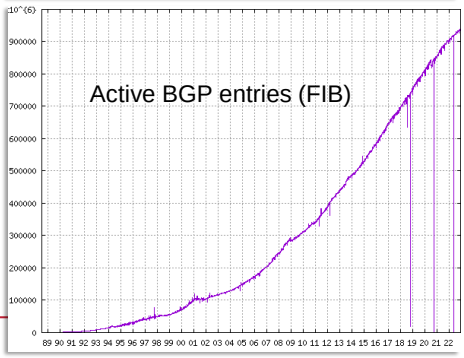
- Tiếp tục bài thực hành BGP, bổ sung thêm trong AS 65400 có 2 BGP router: R4.3 và R4
- Cấu hình kết nối láng giềng giữa các BGP inter-AS như bài trước (là eBGP)
- Cấu hình kết nối láng giềng giữa các BGP R4.3 với R4 bên trong AS 65400 (là iBGP)
- Kiểm tra AS Path đến mạng 10.0.0.0/8 được export từ AS_65100 trên R4.3 và R4: thấy giống nhau (cùng nhận R3 là next hop
- Cấu hình IGP cho AS 65400 bằng OSPF như bài trước, kiểm tra lan tỏa 10.0.0.0/8 từ bên ngoài AS vào bên trong AS
- Kiểm tra lan tỏa 10.0.0.0/8 từ R4 đến R5
- Khai báo tích hợp (redistribute) giữa BGP và IGP (OSPF) của AS 65400. Phân tích các bảng routing của các router trong AS65400 về đường đi đến mạng 10.0.0.0



Hiện trạng Internet backbone BGP

- Nhắc lại BGP: khi một nhà cung cấp dịch vụ internet (hoặc AS - Autonomous System) mua một khối địa chỉ IP toàn cầu (gọi là tiền tố - prefix), họ quảng bá nó đến bảng định tuyến toàn cầu trên Internet. Điều này là để tất cả các AS khác trên thế giới biết rằng tiền tố mới hiện đã hoạt động và cũng biết cách và nơi để gửi bất kỳ gói tin IP nào được định đến nó.
- Số lượng các địa chỉ mạng (prefix) trong bảng định tuyến toàn cầu Internet đang tăng nhanh chóng.
- Hầu hết các tổ chức đã giải quyết vấn đề này bằng cách chi tiêu tiền vào đó. Họ đến các nhà cung cấp như Cisco Systems hoặc Juniper Networks và mua các router của họ. Những router này được xây dựng đặc biệt để có thể thực hiện chuyển tiếp định tuyến ở tốc độ cao ngay cả khi bảng định tuyến phát triển vượt qua một triệu tuyến đường.
- Nguồn số liệu: CIDR report[1]





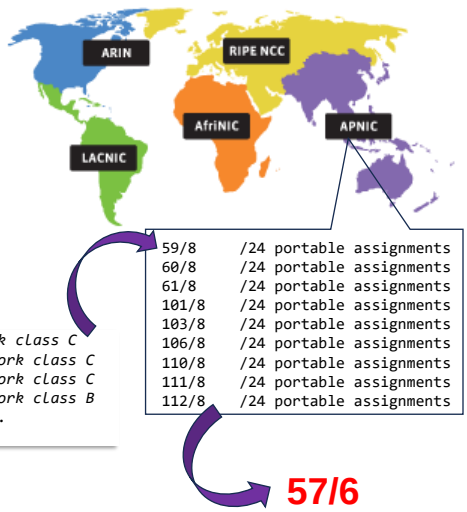
SOICT TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG School of Information and Communication Technology

[1] https://www.cidr-report.org/as2.0

13

Tối ưu BGP routing bằng quản lý địa chỉ IP phân vùng

- 1 triệu bản ghi (prefix) trên BGP backbone!!!
- IANA (Internet Assigned Numbers Authority) & RIR (Regional Internet Registries)
  - APNIC: Asia Pacific
  - AFRINIC: Africa
  - ARIN: Canada, many Caribbean and North Atlantic islands, and the United States
  - LACNIC: Latin America and the Caribbean
  - RIPE NCC: Europe, the Middle East, and parts of Central Asia
- Tối ưu BGP:
  - Không export từng prefix trong AS
  - Khai báo gộp: 59/8, 60/8, 101/8, v.v..
  - Các block địa chỉ IP phải được cấp phát phù hợp



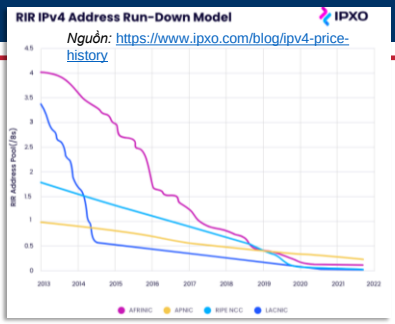
SOICT TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG School of Information and Communication Technology

14

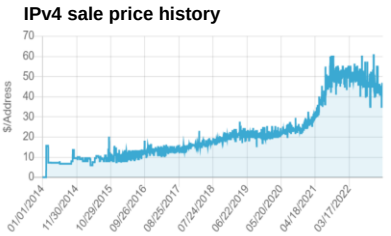
14

IPv4: Điều gì đã xảy ra từ 2010*?

- 2011
  - Tháng 2/2011, IANA hết địa chỉ IPv4, sau khi 5 block cuối cùng được cấp cho RIRs (Regional Internet Registry).
  - Tháng 4/2011, APNIC là RIR đầu tiên cạn kiệt địa chỉ IP để cấp phát cho các tổ chức.
  - Hình thành dự án thương mại đầu tiên giữa Nortel và Microsoft thực hiện chuyển đổi IPv4 sang IPv6.
  - Tháng 5/2011, IPv4 Market Group, LLC bắt đầu vận hành.
  - Tháng 9/2011, APNIC áp dụng chính sách Inter-RIR Transfer Policy để sử dụng các block địa chỉ IP từ RIR khác.
- 2012
  - Tháng 7/2012, ARIN theo chân APNIC, áp dụng chính sách Inter-RIR Transfer Policy vì cạn kiệt địa chỉ IP.
  - Tháng 12/2012, RIPE là RIR tiếp theo hết địa chỉ IP để cấp phát.
- 2014
  - Tính đến cuối năm 2014, đơn giá cấp phát địa chỉ IPv4 (Price/IP) cho các dải lớp B (/16) là khoảng $5/IP.
  - Đơn giá cấp phát địa chỉ IPv4 đã tăng rất nhiều từ 2014 đến nay (xem biểu đồ)
- 2015
  - Tháng 1/2015, IPv6 đạt 5% theo số liệu thống kê người dung từ Google.
- 2018
  - Tháng 1/2028, người dung sử dụng IPv6 đạt 20%.



Nguồn: https://www.ipxo.com/blog/ipv4-price-history



Nguồn: https://www.ipxo.com/blog/ipv4-price-history

[*] https://ipv4marketgroup.com/a-brief-history-of-ipv4/