TW DNS ANYCAST

TWNIC 許乃文

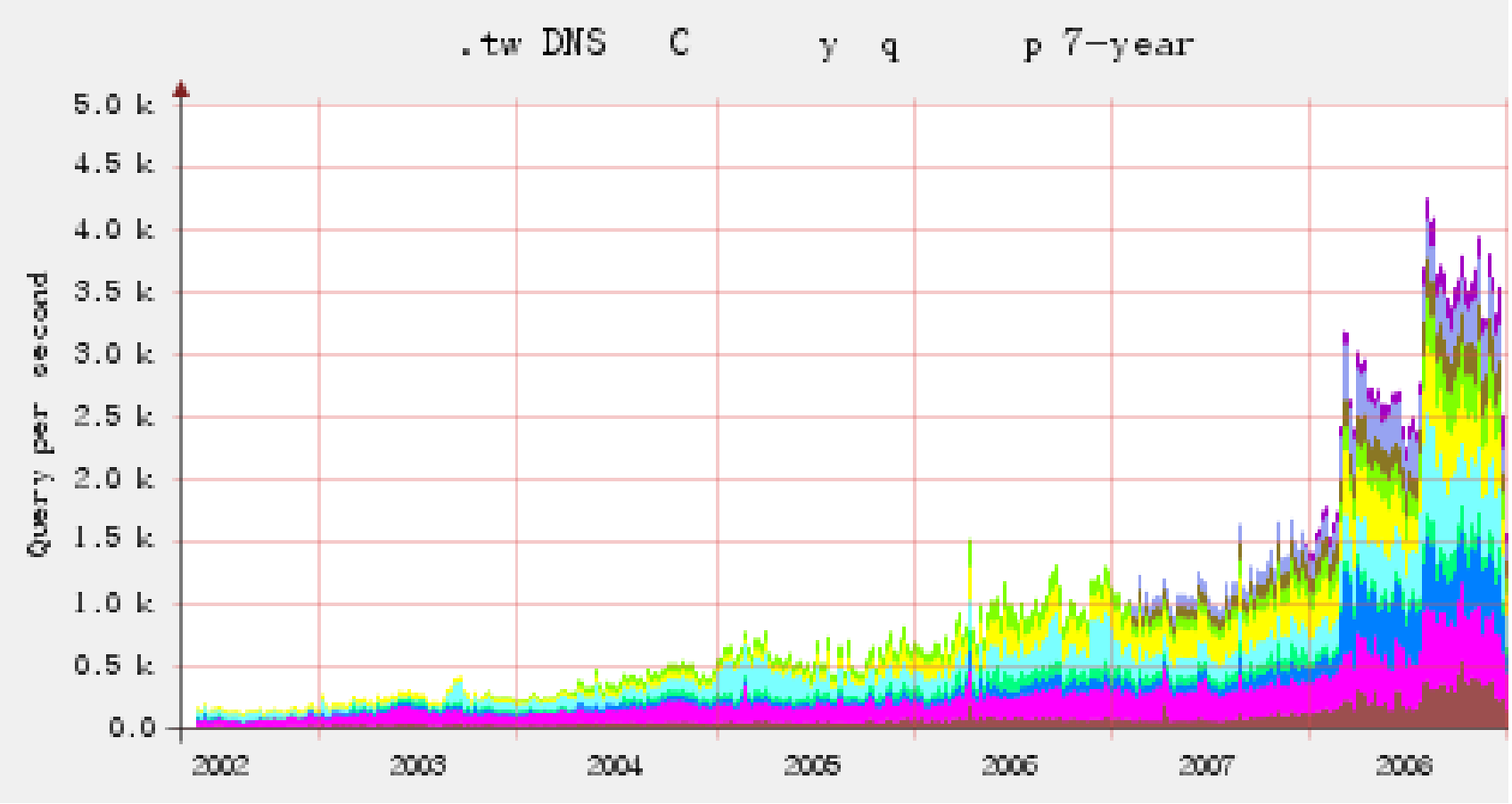
內容

- DNS的重要性及特性
- DNS的效能
- Anycast
- BGP anycast之優缺點
- 使用BGP anycast之實例

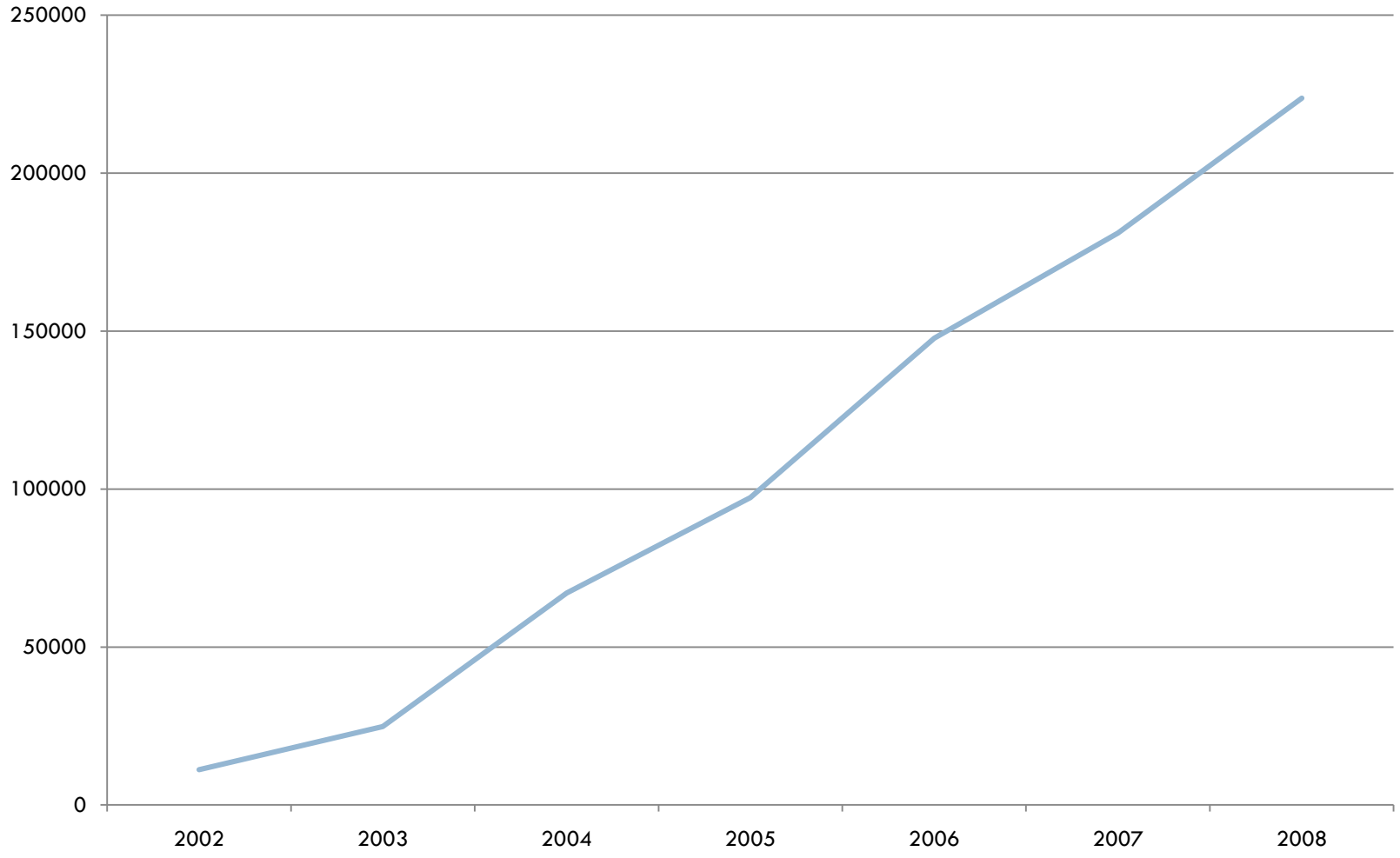
DNS的重要性(1)

- Internet上的服務大都仰賴DNS作解析服務
  - 一個ISP的DNS故障可導致骨幹流量掉到十分之一以下
  - 新的社交網站導致大量的DNS查詢
    - 一個單一的MySpace頁面就可能產生200到300次DNS查詢
    - 一個帶有廣告的新聞網站可能產生10到15次DNS查詢
- 依VeriSign的統計: .COM DNS的流量每十個月成長一倍

.TW DNS流量成長


.tw DNS C y q p 7-year
Query per second
5.0 k
4.5 k
4.0 k
3.5 k
3.0 k
2.5 k
2.0 k
1.5 k
1.0 k
0.5 k
0.0
2002
2003
2004
2005
2006
2007
2008

台灣對外國際頻寬成長


250000
200000
150000
100000
50000
0
2002
2003
2004
2005
2006
2007
2008

DNS的重要性(2)

- 當DNS受到攻擊時
  - 大部份的情況只能用更佳的效能來迎接攻擊
  - 無法使用Firewall來防護DNS—Firewall的效能可能比DNS還差
  - 當某些DNS封包被Firewall擋下時
    - 對方收不到回應→re-transmit→造成更多的封包
    - tw.com.tw的實例

DNS的特性

- 效能上
  - 使用UDP協定
  - 封包大小在512 bytes以內(傳統DNS協定)
- 負載平衡
  - Round-robin, load balance

將來DNS可能的改變

- 封包大小可超過512 bytes
  - RFC 2671中定義的edns0允許超過512 byes
- DNSSEC
  - 解決現有DNS協定上的弱點
  - 可能會大量使用TCP協定來取代UDP協定

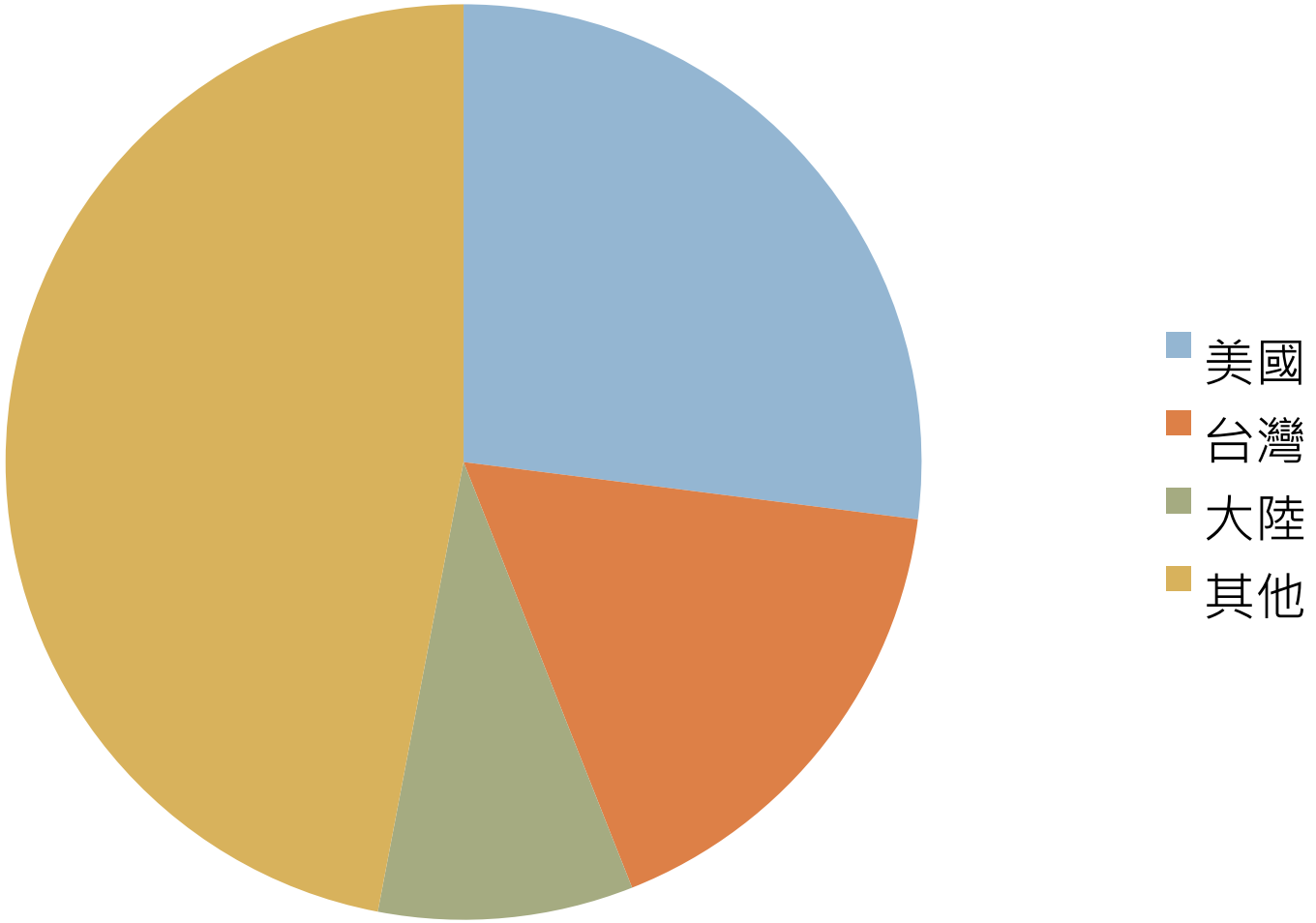
DNSSEC查詢例子

- $ dig @a.ns.se. se. ns +dnssec
- ; <<>> DiG 9.6.0a1 <<>> @a.ns.se. se. ns +dnssec
- ; (2 servers found)
- ;; global options: printcmd
- ;; Got answer:
- ;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 34920
- ;; flags: qr aa rd; QUERY: 1, ANSWER: 11, AUTHORITY: 0, ADDITIONAL: 27
- ;; WARNING: recursion requested but not available
- .
- ;; OPT PSEUDOSECTION:
- ; EDNS: version: 0, flags: do; udp: 4096
- .
- .
- .
- ;; Query time: 336 msec
- ;; SERVER: 192.36.144.107#53(192.36.144.107)
- ;; MSG SIZE rcvd: 2706

DNS效能

- BIND: 7000次查詢/秒
  - 最高紀錄: 93000次查詢/秒
  - 目前使用最廣泛的DNS server軟體，root servers超過一半以上使用BIND軟體
- NSD: 約為BIND的2-3倍
- CNS或ANS: 約為BIND的4-8倍

TW DNS查詢來源統計

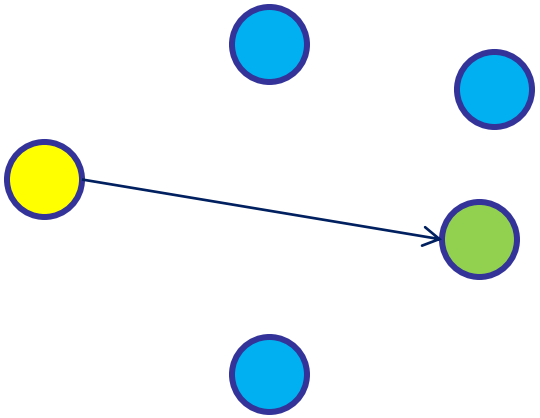


幾個可增加DNS效能的方法

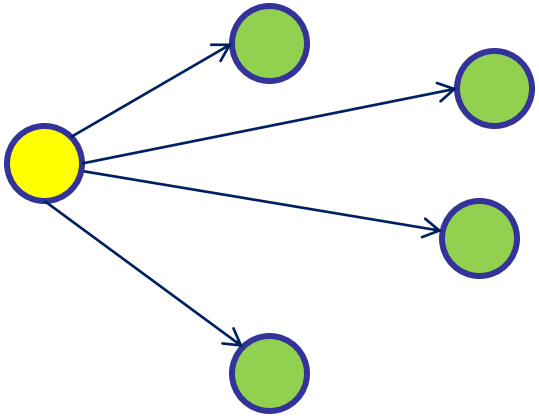
- 較好的設備及軟體
- 負載平衡設備
  - Layer 4 switch
  - Application proxy
- BGP anycast

什麼是anycast

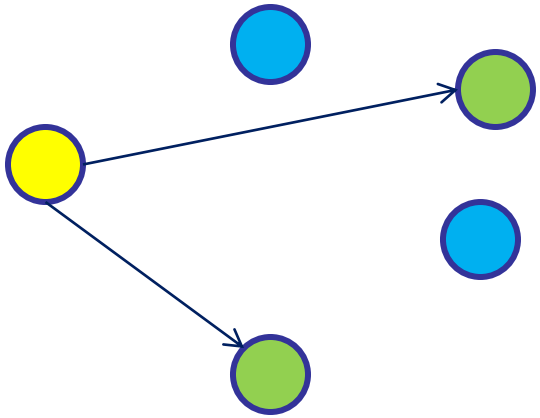
- Unicast
  - 一對一通訊
- Broadcast
  - 廣播
- Multicast
  - 選擇性廣播
- Anycast
  - 就近連線通訊



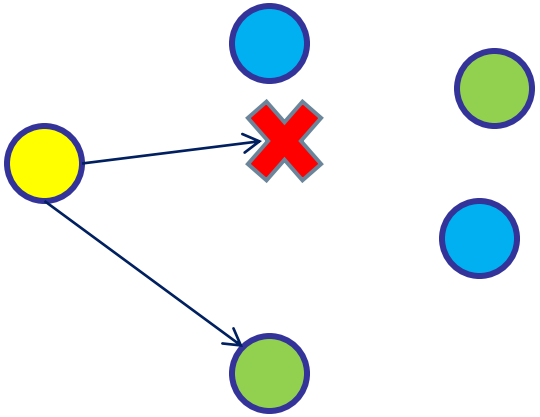
Unicast



Broadcast



Multicast



Anycast

IPv6 anycast

- 網路上所有的裝置都必須有一個獨一無二的IP Address以資識別
- 在RFC 2373中定義了IPv6 anycast的位址
  - 讓一個IPv6 anycast address可以使用在多個設備及網卡上
  - Client端連線到IPv6 anycast address時會自到連線到最近的點

IPv6 anycast的問題(1)

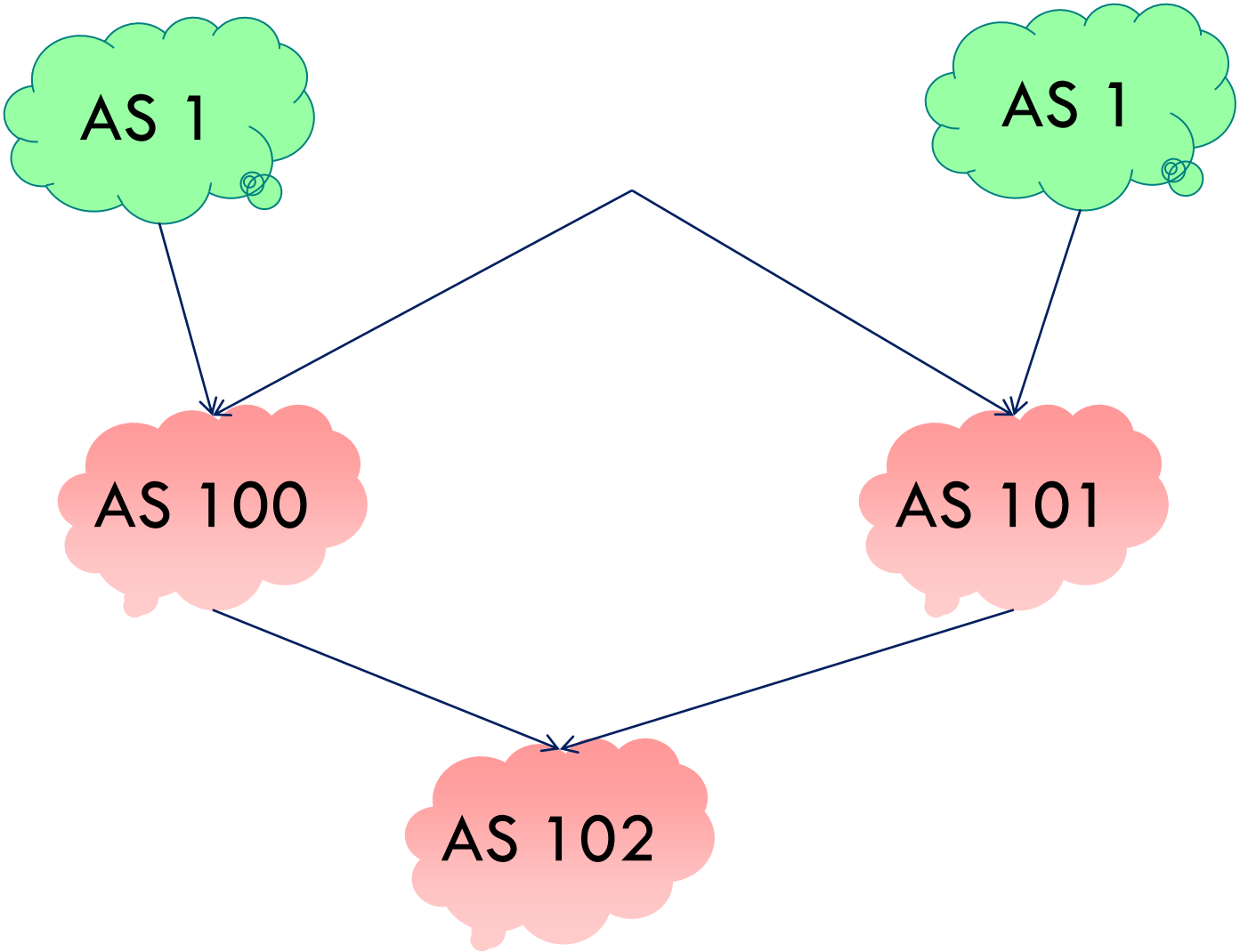
- Anycast address不能用在source address，只能用原unicast address作為source address
- 無法/不易建立TCP連線
- 看來使用UDP協定的DNS較可能利用anycast，但…

IPv6 anycast的問題(2)

- 如果DNS server使用IPv6 anycast address
  - 查詢封包
    - Client→DNS server之IPv6 anycast address
  - 回應封包
    - DNS server之unicast address→Client
  - Client會拒收DNS server送過來之封包，因為source address不等於anycast address

BGP anycast

- 多部設備設定同一個IP來提供相同的服務
- 使用BGP routing protocol在不同的地點和不同或相同的ISP進行路由交換
- 每個地點的AS number及交換的路由都相同
- 由路由器依路由表來決定連到那一部設備
- 對使用者而言是查覺不到的


AS 1
AS 1
AS 100
AS 101
AS 102

使用BGP anycast的優點

- AS-Path prepend可用來調整DNS間之負載
- 就近查詢，減少查詢回應時間
- 減低被攻擊時的影響
- IPv4/IPv6皆可使用
- Clients, Servers, Routers都不需使用特殊軟體

使用BGP anycast的注意事項

- 每個地點都必需一部路由器及相關網路設備
- 每個地點的設備都必須設定一致，包括IP及服務等
- 每個地點的每個設備都必須有兩個IP，除了anycast的IP外還要另一個ISP提供的unicast IP以作為維護使用
- 在BGP AS PATH相同的地方有可能是輪流的方式傳送封包→TCP的連線可能無法建立

Root servers

Server	Locations	IP Addresses	AS Number
A	Sites: 2	IPv4: 198.41.0.4 IPv6: 2001:503:BA3E::2:30	19836
B	Sites: 1	IPv4: 192.228.79.201 IPv6: 2001:478:65::53	
C	Sites: 6	IPv4: 192.33.4.12	2149
D	Sites: 1	IPv4: 128.8.10.90	27
E	Sites: 1	IPv4: 192.203.230.10	297
F	Sites: 48	IPv4: 192.5.5.241 IPv6: 2001:500:2f::f	3557
G	Sites: 1	IPv4: 192.112.36.4	568
H	Sites: 1	IPv4: 128.63.2.53 IPv6: 2001:500:1::803f:235	13
I	Sites: 33	IPv4: 192.36.148.17	29216
J	Sites: 52	IPv4: 192.58.128.30 IPv6: 2001:503:C27::2:30	26415
K	Sites: 17	IPv4: 193.0.14.129 IPv6: 2001:7fd::1	25152
L	Sites: 2	IPv4: 199.7.83.42 IPv6: 2001:500:3::42	20144
M	Sites: 6	IPv4: 202.12.27.33 IPv6: 2001:dc3::35	7500

Root servers

國內使用BGP anycast的實例

- 中華電信168.95.1.1及168.95.192.1等的流量已大於Layer 4 Switch一個port的上限(1Gbps)，已經不是增加DNS設備的問題
- 在中、南部分別安裝server farm並使用BGP anycast技術以提供168.95.1.1等DNS服務

國外使用BGP anycast的實例

- CNNIC使用路由器取代Layer 4 switch
- 伺服器使用gated軟體與路由器交換BGP
- 伺服器設定相同IP及環境以提供DNS服務

參考文件

- RFC 3258: Distributing Authoritative Name Servers via Shared Unicast Addresses
- RFC 4786: Operation of Anycast Services

- AS112: http://www.as112.net

問題與討論