

# Internet Operation

Akira Kato



The Univ. of Tokyo/WIDE Project  
kato@wide.ad.jp

## Overview of the course

### ☆ Internet Operation

- "Real" Internet may not be described in textbooks
  - Understanding of the textbooks is not enough
  - Some books covers partially, however
- Operational environments/requirements varies in time
  - varies in location, in community, in ...

### ☆ Speakers are selected from/around WIDE Project

- <http://www.soi.wide.ad.jp/class/20040013>
- With operational experiences

### ☆ Various topics in the Internet operations

- Not all of the topics are covered
- Profitable to those who operate something
  - ISP, Campus Network, Laboratory, ...

## Overview of the course

### ☆ Prerequisites

- Overall knowledge of the Internet
  - Items covered by the textbooks
- Brief understanding of each protocol
  - BGP, OSPF, Spanning Tree, SNMP, DNS, ...

### ☆ Lecture focuses on how they are operated

- Basic overview is skipped
  - Sometimes review of the protocols is given
- Practical understanding of the protocol
- Hints to the trouble shooting

## Is the prerequisites appropriate?

### ☆ Can you explain the following briefly?

- Dijkstra's Algorithm
- How traceroute works?
- What is the role of Router ID?

### ☆ If you can explain them clearly enough

- You may need not to participate the course :-)
- But brief understanding is important

### ☆ Please tell us if such assumption makes sense

## At the end of the course

### ☆ We are planning to hold a workshop

- For select persons
  - mainly due to fiscal reason
- Score of assignments are subject for selection
- Possible styles
  - Hands-on workshops, Site visit,
- Keiko/Shoko is going to announce
  - Stay tuned
  - Feed forward is welcome

## Feedback

### ☆ Distance education environment

- Speakers hardly feel if you understand something
- Overdone actions are welcome
  - Positive/negative feedback
  - Good to prevent from falling sleep :-)
- To raise questions are good idea
  - Most of the lectures are "interrupt enabled"
  - Raise questions at anytime

### ☆ Use a BBS

[http://www.soi.wide.ad.jp/soi-asia/class\\_bbs/20040013/](http://www.soi.wide.ad.jp/soi-asia/class_bbs/20040013/)

### ☆ Email to 20040013\_faculty@soi.wide.ad.jp

- Speakers can see your messages
- Other course attendees can't
- Don't hesitate to post your comments/questions

## For further study

### ☆ **APRICOT**

- AP region operators' get together
  - NANOG/RIPE counterpart
  - APNIC Member meeting is colocated
- APRICOT2005 Feb 16-25, 2005 in Kyoto, JP
  - <http://www.2005.apricot.net/>

### ☆ **SANOG: <http://www.sanog.org/>**

- Southasian Network Operators' group

### ☆ **APAN: <http://www.apan.net/>**

- Not a operators' get together
- Focus on academic advanced networking
- APAN2005 Jan 24-28 in Bangkok, TH

## Brief review of Routing

## Review of Routing

### ☆ Routing: determine direction of the packet

- Which nexthop is appropriate to get the destination?
- Routing is one-way
  - Returning packets get routed separately
  - Asymmetric routing in many cases

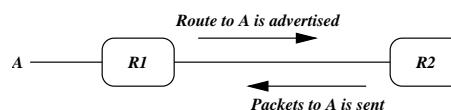
### ☆ Packet forwarding

- Best-match table lookup in IP (v4/v6)
  - If two entries matches, use one with longest match
- Route is specified with prefix length
  - Example: 202.12.27.0/24
- Routing table example
  - A: 0.0.0.0/0
  - B: 203.178.128.0/18
  - C: 203.178.128.0/24
- Which RT entry matches the destinations below?
  - 203.178.127.5, 203.178.128.5, 203.178.129.5

## BGP: implementing routing policy

### ☆ Routing in general

- If you advertise a route
  - route: reachability info for a prefix
- then your neighbor will send traffic
  - whose destinations match with advertised prefixes



### ☆ RT propagation is opposite direction to packet flow

- Especially true for distance vector routing
  - In BGP, it is in AS-level granularity
- It may not be true in link-state routing

## Implementation of routing policy

### ☆ If we need to restrict the use of the link

- e.g. customers can use the link
- then we need to control "advertising"
- It can be implemented easily in distance vector
  - advertise/not-advertise can control the usage
- It may not be easily implemented in link-state
  - permission database need to be established

### ☆ How to control the advertisement

- per prefix
- more convenient way?
  - per AS

## BGP's policy

### ☆ A set of attributes are associated with a route

- AS-Path : sequence of ASes
- MED : preference of multiple interconnects
- NextHop : nexthop router's address

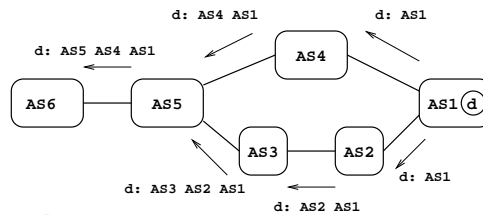
### ☆ AS is a (currently) 16bit globally unique number

- 64512 -- 65534 : private AS Number
- About half of entire space have been assigned
  - May need to expand to 32bit in near future
  - draft-ietf-idr-as4bytes-08.txt

### ☆ AS-Path can be used for control advertisement

- which routes are to accept
- which routes are to advertise

## Propagation of BGP AS-Path



- ☆ **AS5 has a choice**
  - d: AS4 AS1, or d: AS3 AS2 AS1
- ☆ **AS5 makes a decision**
  - shorter AS-Path
  - other configured policy
    - ex. avoid AS4, prefer AS3, ...
- ☆ **AS6 have no freedom to select the route to "d"**
  - It depends on AS5's decision

## BGP Configuration Example

```
1: router bgp 2500
2: neighbor 203.178.133.139 remote-as 9407
3: neighbor 203.178.133.139 description DragonTap
4: neighbor 203.178.133.139 send-community
5: neighbor 203.178.133.139 soft-reconfiguration inbound
6: neighbor 203.178.133.139 password <password>
7: neighbor 203.178.133.139 filter-list 10 in
8: neighbor 203.178.133.139 route-map bgp-dragontap-out out
```

## BGP Configuration Example

```
9: ip as-path access-list 10 permit ^(9407_)+$
10: ip as-path access-list 10 permit ^(9407_)+(4538_)+$

11: ip as-path access-list 20 permit ^4717_

12: ip as-path access-list 90 permit ^$

13: ip prefix-list LOCAL seq 100 permit 131.113.0.0/16
14: ip prefix-list LOCAL seq 100 permit 133.12.0.0/16
15: ip prefix-list LOCAL seq 100 permit 133.27.0.0/16

16: route-map bgp-dragontap-out permit 10
17: match ip address prefix-list LOCAL
18: match as-path 90
19: route-map bgp-dragontap-out permit 20
20: match as-path 20
```

## BGP Sessions

```
% show ip bgp summary
```

```
Neighbor      V  AS MsgRcvd MsgSent TblVer InQ OutQ Up/Down
State/PfxRcd
203.178.133.139 4 9407 132947 135936 84795 0 0 1w5d 50
```

```
% show ip bgp regexp ^9407
```

```
Network      Next Hop      Metric LocPrf Weight Path
*> 59.64.0.0/14 203.178.133.139          0 9407 4538 i
*> 59.64.0.0/13 203.178.133.139          0 9407 4538 i
*> 59.72.0.0/15 203.178.133.139          0 9407 4538 i
*> 162.105.0.0 203.178.133.139          0 9407 4538 i
```

## BGP Path Preference

### ☆ If there are multiple paths available for a destination

- (by Cisco, may be vary by vendors)
- Highest LOCAL\_PREF
- Shortest AS-PATH
- Smallest ORIGIN value
- Lowest MED value
- Smallest metric to NEXTHOP
- Lowest IP value of BGP Router ID
- Tie breaking rule

## BGP Operational issues

### ☆ Correct policy configuration is not that easy

- Accepting prefixes is straightforward
- It is difficult to check the announced prefixes

### ☆ Solution:

- "show ip bgp neighbor <address> advertised-routes"
- Use a dumb router to check the advertisement
- Check the prefixes with RouteViews, etc
- Use Looking glass sites

## BGP route viewer

- ☆ **Oregon Route Views: <http://www.route-views.net/>**
  - Telnet-capable cisco/juniper/zebra boxes
  - Feed routes from participating ISPs
    - Good coverage,
  - try "telnet route-views.oregon-ix.net"
- ☆ **RIPE RIS**
  - <http://www.ripe.net/ripencc/pub-services/np/ris/index.html>
  - No a direct login account provided
  - Accessible via various tools/web interfaces
- ☆ **Data archives available on both of above**
  - It helps to trace a past routing incident

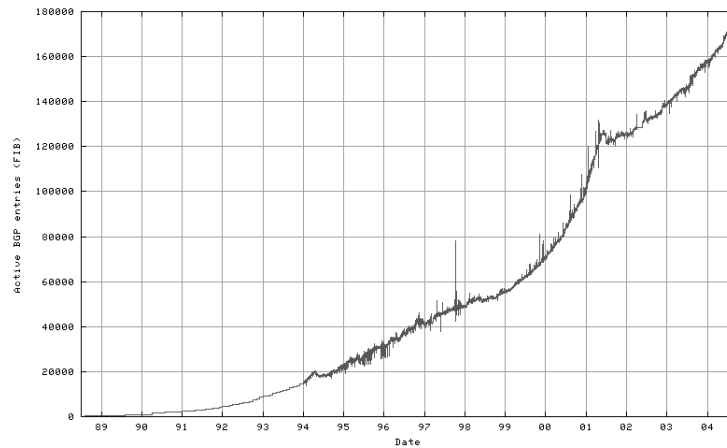
## Looking glass

- ☆ **Router viewer with limited capability**
  - subnet of "show" commands
  - ping/traceroute
- ☆ **BGP4 Wiki: <http://www.bgp4.net/>**
  - lists number of looking glass sites
  - LG service is provided by ISP's kindness
- ☆ **Abilene CoreNode Proxy**
  - <http://ratt.uits.iu.edu/routerproxy/abilene/>
  - More commands than a regular LG service

## How routing system scales?

### ☆ Number of prefixes in the BGP table grows

- <http://bgp.potaroo.net/>



## How routing system scales?

### ☆ If you add ONE prefix

- It will consume some bandwidth to propagate
- It will consume CPU cycles of the backbone routers
- It will consume all backbone routers' memory

### ☆ You need a router with a large amount of memory

- e.g. An 1U router equipped with a 1GB memory

## Route Aggregation

### ☆ You can "aggregate" adjacent prefixes into one

- 202.249.0.0/18 from AS2500
- 202.249.64.0/18 from AS2500
- then a single prefix works fine
  - 202.249.0.0/17 from AS2500

### ☆ If everybody does this

- 147,000 prefixes will be 87,000 prefixes

router bgp 2500

```
aggregate-address 202.249.0.0 255.255.192.0 summary-only
```

- "summary-only" suppress specifics
  - unaggregated routes are not advertised

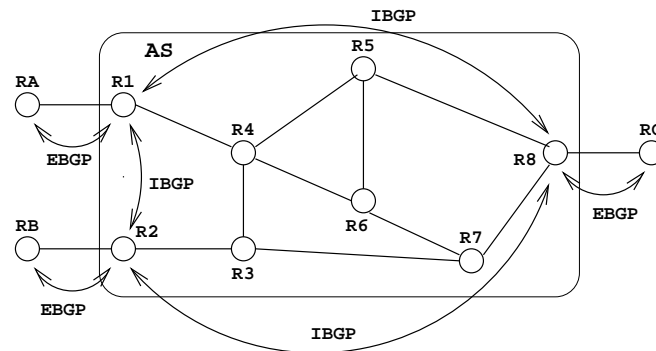
## Minimum prefix length

### ☆ Most of the ISP rejects routes with /25 or longer

- No host route is acceptable
- /24 for 192.0.0.0/8 and a few other blocks
- /21 for other spaces
  - RIR's (e.g. APNIC) minimum allocation size

### ☆ You are encouraged to aggregate as much as possible

## Interaction with IGP



### ☆ All routers must know routes from other AS

- In this figure. Stub routers are exempt.
- Otherwise AS can not provide transit service
- Traditionally, BGP routes were injected to IGP
- BGP/OSPF interaction (RFC1403, historic)
  - Use OSPF Tag to match BGP/OSPF routes

## Interaction with IGP

### ☆ OSPF can not handle 100,000 routes

### ☆ IBGP is used to feed routes to the intermediate routers

- IBGP as (partial) IGP
- Still OSPF/RIP need to be run
  - To IBGP TCP session get established
  - To resolve nextthop

### ☆ But IBGP should be maintained as Full Mesh

- If you have 5 BGP speakers, you are happy
- If you have more than 10 BGP speaks...
  - $(n-1)^2$  configurations.... a nightmare
- Some tools are necessary to overcome  $O(n^2)$

## Interaction with IGP

### ☆ Use BGP Route Reflector (RFC2796)

- Everbody peers with Route Reflector
- RR forwards the update to others
- $O(n^2) \rightarrow O(n)$
- You may need multiple RR for redundancy

### ☆ Introduction of AS hierarchy (RFC3065)

- IDRPs Confederation is imported to BGP partially
- $O(n^2) \rightarrow O(m^2)$ ,  $m < n$

## Today's Assignment

### ☆ 1) Check your IP address with RouteView or RIS

### ☆ 2) Choose one looking glass site

- perform traceroute to you
- perform traceroute to the LG site
- compare both

### ☆ 3) Give 2-3 lines of feedback

### ☆ Submission dues on Sept. 22 00:00 UTC

### ☆ Submit your assignment through the web

- at the bottom of the following URL
- <http://www.soi.wide.ad.jp/class/20040013/>