


Slide 1

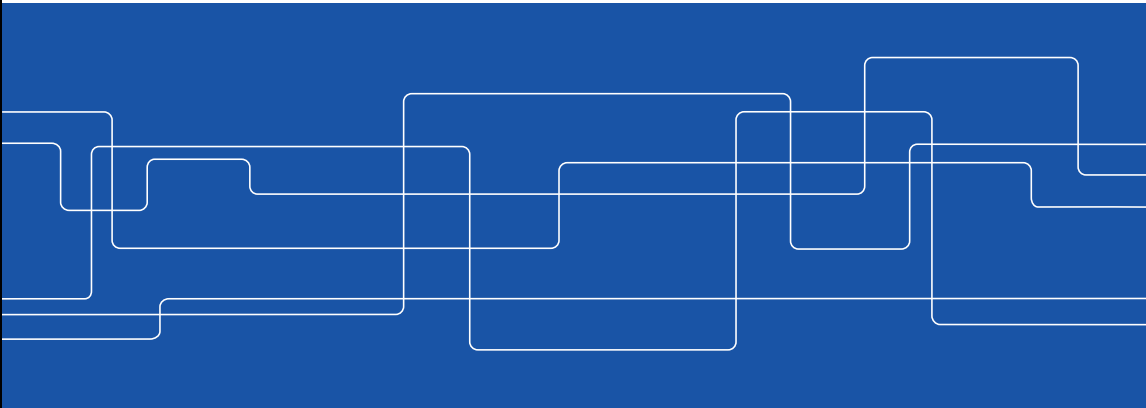


# **IK1552**

# **Internetworking/Internetteknik**

prof. Gerald Q. Maguire Jr. <http://people.kth.se/~maguire/>

School of Electrical Engineering and Computer Science (EECS), KTH Royal Institute of Technology  
IK1552 Spring 2019, Period 4 2019.03.13 © 2019 G. Q. Maguire Jr. All rights reserved.



## Slide 2



## Module 7: Dynamic Routing

Lecture notes of G. Q. Maguire Jr.

For use in conjunction with James F. Kurose and Keith W. Ross, *Computer Networking: A Top-Down Approach*.

IK1552

SPRING 2019

SLIDE 2

Slide 3



## Outline

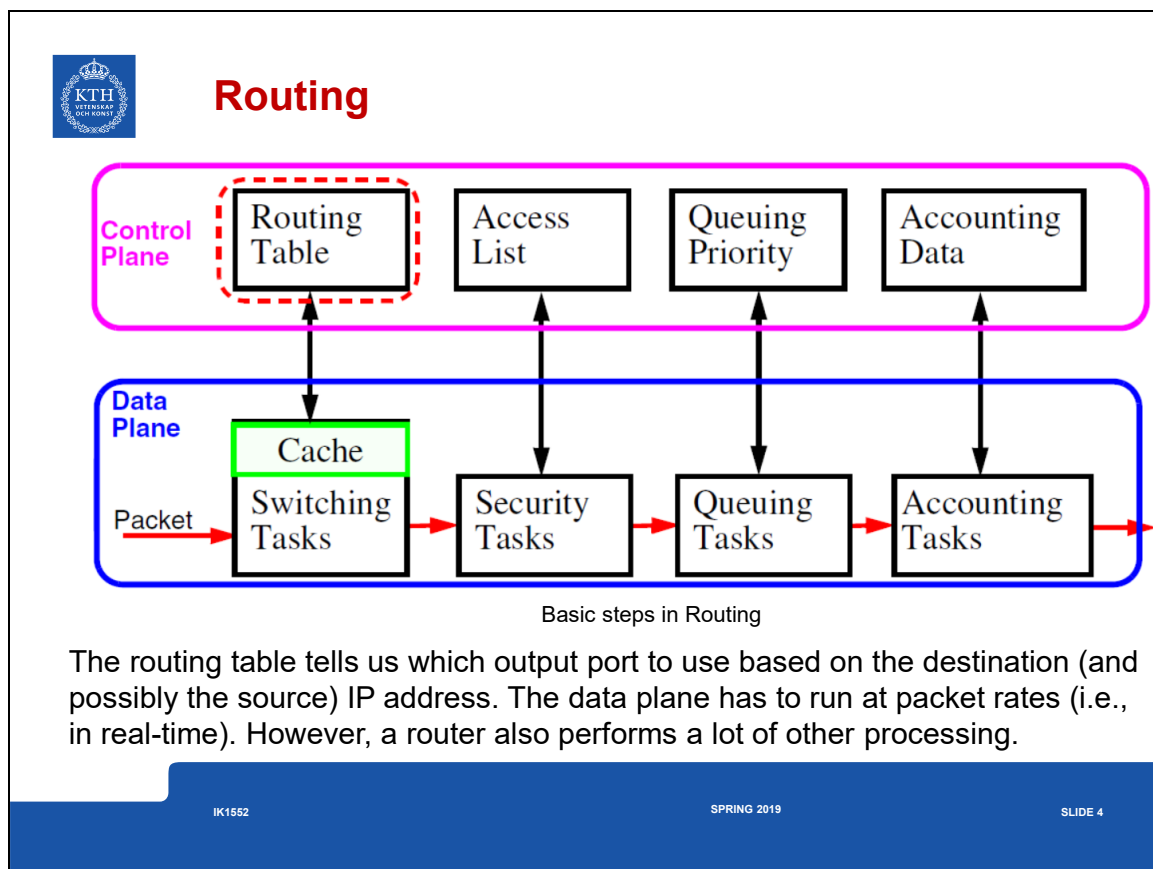
Dynamic Routing Protocols

IK1552

SPRING 2019

SLIDE 3

## Slide 4





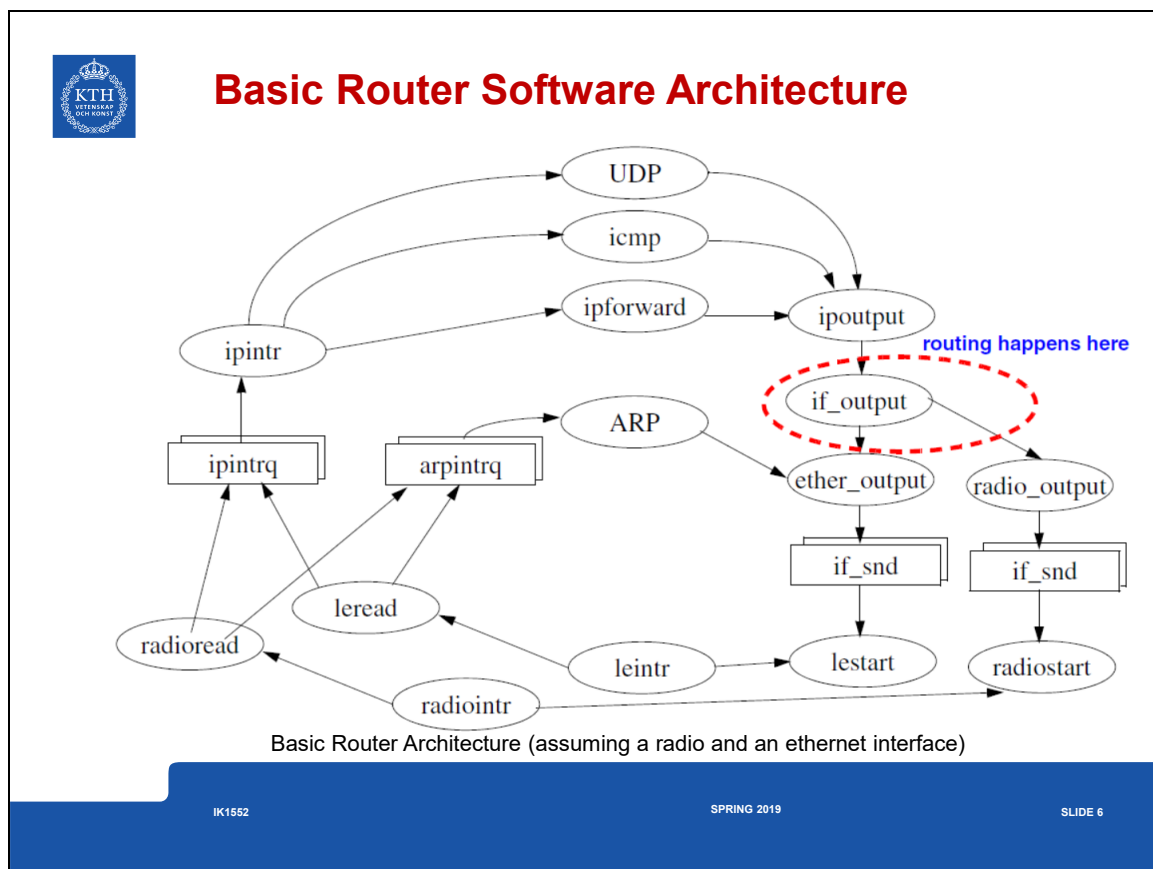
## Routing Principles

- **Routing Mechanism:** Use the *most specific* route  
IP provides the mechanism to route packets
- **Routing Policy:** What routes should be put in the routing table? <<< **today's topic!**  
Use a routing daemon to provide the *routing policy*

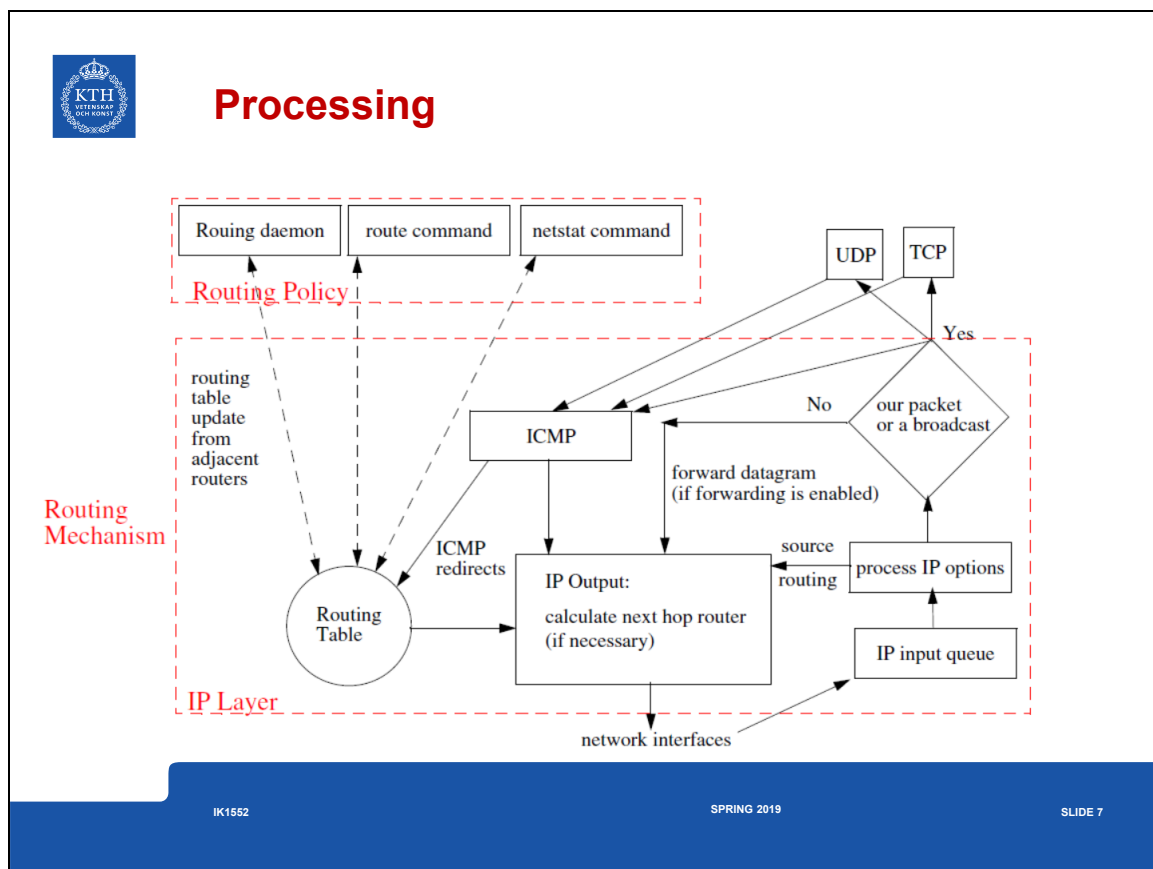
For further information see:

Bassam Halabi, *Internet Routing Architectures*, Cisco Press, 1997, ISBN 1-56205-652-2. -- especially useful for IGRP.

## Slide 6



Slide 7





## Routing packets in the Internet

Router needs to know where to route packets, to do this they need routing information. Such information can be provided by **manually entered routes** or ICMP Redirect or learning of routes via **a routing protocol**.

Dynamic routing protocols are based on routers talking to each other.

Intradomain - within an AS (aka Interior Gateway protocols)

- RIP-1 - Routing Information Protocol (version 1)
- RIP-2 - Routing Information Protocol (version 2)
- OSPF - Open Shortest Path First

Interdomain - between ASs (aka Exterior Gateway protocols)

- BGP - Border Gateway Protocol





## Autonomous systems (ASs) – RFC 1930

Each of which is generally administered by a **single entity**.

Each autonomous system selects the routing protocol to be used **within** the AS.

Network	AS number
Swedish University Network (SUNET)	AS1653
SUNET-KI	AS2837
Stockholm University - SU	AS2838
KTH	AS2839

For statistics about the number of AS, etc.:

<http://www.cidr-report.org/as2.0/>

<http://www.cidr-report.org/v6/as2.0/>

For a list of AS number to name mappings: <http://www.cidr-report.org/as2.0/>

To find out who is responsible for a given autonomous system see

<https://apps.db.ripe.net/search/query.html> search for AS2839

J. Hawkinson and T. Bates, 'Guidelines for creation, selection, and registration of an Autonomous System (AS)', *Internet Request for Comments*, vol. RFC 1930 (Best Current Practice), Mar. 1996 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1930.txt>



## Routing Metrics

A measure of which route is better than another:

- Number of hops
- Bandwidth
- Delay
- Cost
- Load
- ...

It is possible that the metric uses some **weighted combination** of the above.



## Routing Algorithms

- Static vs. Dynamic
- Single path vs. Multi-path
- Flat vs. Hierarchical
- Host-intelligent vs. Router-intelligent
- Intradomain (interior) vs. Interdomain (exterior)
- Link state vs. Distance vector

### Issues:

- Initialization (how to get started)
- Sharing
- Updating
- When to share & Who to share with



## Intradomain routing protocols

also called “Interior Gateway protocols”

Examples:

- HELLO - an old IGP protocol
- RIP - widely used
- OSPF - increasingly used

Routers that speak any dynamic routing protocol **must** speak RIP and OSPF.



## Routing Information Protocol (RIP) version 1

RFC 1058 {Note it was written years after the protocol was in wide use!}

RIP is a distance-vector protocol - RIP messages contain a vector of **hop counts**. RIP messages are carried via UDP datagrams which are **broadcast**.

0	8	16	31
Command	Version = 1	Reserved	
Family		All 0s	
Network Address			
All 0s			
All 0s			
Distance			

- a command: request or reply
- a version number (in this case 1)
- up to 25 instances (entries) containing:
  - address family (2 = IP addresses)
  - Network Address (allocated 14 bytes, for an IP address we only need 4 bytes - and they are aligned to a 4 byte boundary - hence the leading and trailing zeros)
  - metric [hop count]

RIP message format (see Forouzan figure 14.9 pg. 394)

C. L. Hedrick, 'Routing Information Protocol', *Internet Request for Comments*, vol. RFC 1058 (Historic), Jun. 1988 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1058.txt>

J. Halpern and S. Bradner, 'RIPv1 Applicability Statement for Historic Status', *Internet Request for Comments*, vol. RFC 1923 (Informational), Mar. 1996 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1923.txt>

## Slide 14



## RIP v1 operation

As carried out by UNIX daemon “routed” using UDP port 520

Initialization:

for all interface which are up

- { send a request packet out each interface asking for the other router's complete routing table

- [command=1, address family=0 {== unspecified}<sup>†</sup>, metric=16}

Request received:

- if whole table requested, then send it all 25 at a time

- else if a specific set of routes

  - then fill in the metric

  - else set metric to 16

    - [16 == “infinity” == we don't know a route to this address]

Response received:

- if valid (i.e., not 16), then update/add/delete/modify routing table

<sup>†</sup> Page 24 of RFC 1058 says “If there is exactly one entry in the request, with an address family identifier of 0 (meaning unspecified), and a metric of infinity (i.e., 16 for current implementations), this is a request to send the entire routing table.”

IK1552

SPRING 2019

SLIDE 14

C. L. Hedrick, ‘Routing Information Protocol’, *Internet Request for Comments*, vol. RFC 1058 (Historic), Jun. 1988 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1058.txt>



## When are routes sent?

**Solicited response:** Send a response when a request is received

**Unsolicited response:**

- If a metric for a route changes, then (trigger) send update, else send all or part of the table every 30 seconds.
- If a route has not been updated for 180 seconds (3 minutes = 6 update cycles), then set metric to 16 and then **after** 60 seconds (1 minute) delete route.

Metrics are in units of hops, thus this protocol leads to selection between routes based on the minimum number of hops.

Summary of RIPv1 Timers:

- **Periodic** timer - regular updates random value [25..35] mean 30 s.
- **Expiration** timer - routes not updated within (180 s) expire
- **Garbage collection** timer - 120 s after expiring entries are GC'd




## Problems with RIP v1

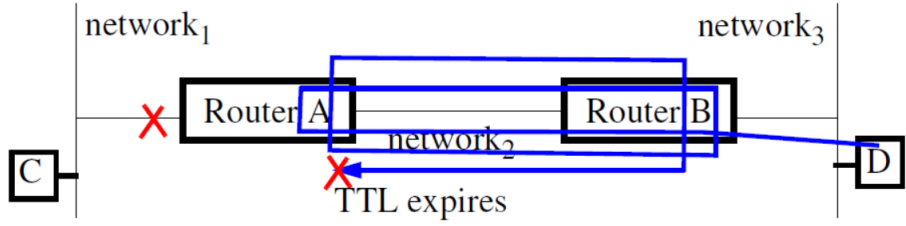
- RIPv1 does **not** know about subnets (or assumes all interfaces on the network have the same netmask)
- after a router or link failure RIP takes **minutes** to stabilize (since each neighbor only speaks ~every 30 seconds; so the time for the information to propagate several hops is minutes)  $\Rightarrow$  while it is unstable it is possible to have routing loops, etc.
- Hops count may **not** be the best indication for which is the best route
- Since the maximum useful metric value is 15, the network diameter must be less than or equal to 15.
- RIP will accept updates from anyone - so one misconfigured device can disrupt the entire network.
- RIP uses more bandwidth than other protocols, since it sends the whole routing table in updates.



## Slide 17



## Count to Infinity



- Router A advertises it knows about routes to networks 1 and 2
- Router B advertises it knows about routes to networks 2 and 3
- After one update cycles A and B know about all 3 routes.

If A's interface to Network<sub>1</sub> goes down, then A learns from B - that B knows a route to Network<sub>1</sub>; so A now thinks it can reach Network<sub>1</sub> via B. So if D sends a packet for C, it will simply loop back and forth between routers A and B, until the TTL counts down to 0.

IK1552

SPRING 2019

SLIDE 17



## Split Horizon

To counter the count to infinity, the split horizon algorithm - never sends information on an interface that it learned from this interface.

RIPv1 implements: **Split Horizon with Poison Reverse Update** - rather than not advertise routes to the source, advertise them with a metric of 16 (i.e., unreachable) - hence the source simply ignores them.

Unfortunately split horizon only prevents loops between **adjacent** routers (so if there are three or more routers involved the previous problem re-appears)



## Triggered updates and Hold-Downs

To decrease convergence time - when the topology changes send out an update immediately.

However, if a node can learn about connectivity from more than one source, then if a delete happens before the add, then the existence of the route is asserted; therefore the hold-down rules says:

When a route is removed, no update of this route is accepted for some period of time - to give everyone a chance to remove the route.

This period of time is the Hold-down time.

The result is to decrease the rate of convergence; but the combined effect of triggered updates + hold-downs leads to **faster** convergence than not using triggered updates.



## RIP extensions (aka RIP-2)

Defined in RFC 1388 and revised in RFC 2453. Version number is 2.

- for each of up to 25 entries we add the fields:
- Route tag - carries the AS number
- Subnetmask - to be used with this address (to support classless addressing)
- Next-hop IP address, either the IP address of where packets to this destination should be sent or zero [which means send them to the system which sent the RIP message]
- One entry can be replaced by Authentication data

RIP-2 supports multicast to address 224.0.0.9, to reduce load on hosts **un**interested in RIP-2 messages

0 8 16 31

Command	Version = 2	Reserved	
0xFFFF		Authentication type	
Authentication data (16 bytes)			
if Authentication type = 2, this is a clear text password to be used to authenticate this message			
Family		Route tag	
Network Address			
Subnet mask			
Next-hop address			
Distance			

RIPv2 message format (see Forouzan figures 14.13 pg. 397 and 14.14 pg. 398)

G. Malkin, 'RIP Version 2 Carrying Additional Information', *Internet Request for Comments*, vol. RFC 1388 (Proposed Standard), Jan. 1993 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1388.txt>

G. Malkin, 'RIP Version 2 - Carrying Additional Information', *Internet Request for Comments*, vol. RFC 1723 (INTERNET STANDARD), Nov. 1994 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1723.txt>

G. Malkin, 'RIP Version 2', *Internet Request for Comments*, vol. RFC 2453 (INTERNET STANDARD), Nov. 1998 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc2453.txt>



## Why would anyone use RIP?

After all these problems you might ask this question.

### Answer

- Because RIP is generally the only routing protocol which **all** UNIX machines understand!
- Relatively easy to configure
- It is widely available, since it **must** exist if the device is capable of routing!



## Interior Gateway Routing Protocol (IGRP)

Cisco's IGRP [Hedrick 1991] - a proprietary protocol with the following goals:

- stable, optimal routing for large networks - with no routing loops
- fast response to changes in net topology
- low overhead in both bandwidth and processor utilization
- ability to **split traffic across several parallel routes** if they are (or nearly are) equal.

It is a distance-vector protocol based on many of the ideas from RIP.

IK1552

SPRING 2019

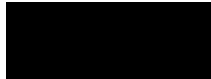
SLIDE 22

Charles L. Hedrick, "An Introduction to IGRP", Cisco, Document ID: 26825, Technology White Paper, August 1991 <http://www.cisco.com/warp/public/103/5.html>



## IGRP Metrics

- a vector of metrics each with a 24 bit value
- composite metric is



where K1 and K2 are constants, B the unloaded path bandwidth, D a topological delay, and R is reliability

- also we pass the **hop count** and **Maximum Transmission Unit** values

K1 is the weight assigned to bandwidth (by default 10,000,000)

K2 is the weight assigned to delay (by default 100,000)

If up to 4 paths are within a defined **variance** of each other, Cisco's IOS (Internetwork Operating System) will split the traffic across them in inverse proportion to their metric.



## IGRP Route Poisoning

IGRP poisons routes which increase by a factor of 10% or more after an update  
[they are thought to be: “too good to be true”].

While this rule may temporarily delete valid routes (which will get reinstated after the next regular update) - it allows use of a zero hold-down time, which leads to faster convergence.





## IGRP Default Gateway

Rather than using the fake network 0.0.0.0 to indicate the default network, IGRP allows a real network to be flagged as the default network.

Periodically, IGRP scans the routes offering a path to this flagged network and selects the path with the lowest metric.

Note: Default gateways help to keep the size of local routing tables smaller.



## Enhanced IGRP (EGRP)

Uses the distance-vector technology of IGRP, but changes the way routes are advertised and the calculation of entries for the routing table.

EGRP uses:

- a neighbor discover/recovery process of hello packets to learn its neighbors
- a Reliable Transport Protocol to ensure guaranteed, ordered delivery of routing updates
- a Diffusing Update Algorithm (DUAL) - which selects both the best route for insertion into the table **and** a feasible **successor** (for if the primary route fails)
- Variable length subnet masks (VLSM)

EGRP is a Cisco proprietary technology.

IK1552

SPRING 2019

SLIDE 26

“Enhanced Interior Gateway Routing Protocol”, Cisco, Technology White Paper, Document ID: 16406, April 19, 2005 <http://www.cisco.com/warp/public/103/eigrp-toc.html>



## Open Shortest Path First (OSPF)

OSPF defined in RFC 2328

OSPF is a link-state protocol. OSPF messages (Link State Advertisements (LSAs)) tell the **status of links** of each of its neighbors and propagates this info to its neighbors. Each router uses this link-state information to build a **complete** routing table. Uses IP directly (protocol field = OSPF (89)) ⇒ does **not** use UDP or TCP.

### Advantages

- link-state protocols converge faster than distance-vector protocols
- can calculate a route per IP service type (i.e., TOS)
- each interface can have a per TOS cost
- if there are several equally good routes ⇒ can do **load balancing**
- supports variable length subnet masks
- enable point to point links to be unnumbered (i.e., don't need an **IP address**)
- uses clear text passwords
- uses multicasting

IK1552

SPRING 2019

SLIDE 27

J. Moy, 'OSPF Version 2', *Internet Request for Comments*, vol. RFC 2328 (INTERNET STANDARD), Apr. 1998 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc2328.txt>



## OSPF (continued)

OSPF uses the Shortest Path First algorithm (also known as **Dijkstra's algorithm**). OSPF networks are generally divided into areas such that cross-area communication is minimal.

Some routers with multiple interfaces become **border area routers** (with one interface in one area and another interface in another area).

The only way to get from one area to another area is via the **backbone** - which is area 0. Note: The backbone need **not** be continuous.

Link state advertisements are sent to all routers in a given area (via **flooding**), rather than just neighbors (as in the distance-vector approach) - thus periodic updates are infrequent (every 1 to 2 hours).

A key feature of OSPF is **route aggregation** - which minimizes the size of routing tables and the size of the topological database; in addition, it keeps protocol traffic to a minimum.



## OSPF building blocks

### 1. Hello protocol

Check for & with neighbors and learn designated routers

### 2. Synchronization of Databases

Exchange of Link State Database between neighbors

- Get LSA **headers**
- Request the transfer of necessary LSAs

### 3. Flooding protocol

When links change or when your knowledge is old

- Send Link State updates to neighbors and flood recursively
- If not seen before, propagate updates to all adjacent routers, except the router you received it from

## Slide 30



## OSPF Packets

### Common header

0	8	16	31
Version	Type	Message length	
Source Router IP Address			
Area Identification			
Checksum		Authentication type	
Authentication (64 bits) <sup>†</sup>			

OSPF Common Header (see Forouzan figure 14.26 pg. 408)

<sup>†</sup>Note that the Authentication field is 64 bits, Forouzan figure 14.26 pg. 408 incorrectly shows it as 32 bits.

5 types of OSPF packets:

Type	Description
1	Hello
2	Database Description
3	Link State Request
4	Link State Update
5	Link State Acknowledgment

## Slide 31



## Hello packet

0	8	16	31
Version	Type = 1	Message length	
Source Router IP Address			
Area Identification			
Checksum		Authentication type	
Authentication (64 bits)			
Network Mask			
Hello interval (seconds)		All zeros	E T Priority
Dead interval (seconds)			
Designated router IP address			
Backup designated router IP address			
Neighbor IP address			
...			
Neighbor IP address			

- E = 1 indicates a stub area OSPF Link state update packet (see Forouzan figure 14.44 pg. 419)
- T = 1 indicates router supports multiple metrics
- priority = 0 indicates that this router should not be considered as a designated or backup designated router
- Dead interval is the time before a silent neighbor is assumed to be dead
- list of neighboring routers (of the router which sent the hello packet)

## Slide 32



## Database Description packet

0	8	16	31
Version	Type = 2	Message length	
Source Router IP Address			
Area Identification			
Checksum		Authentication type	
Authentication (64 bits)			
Interface MTU			
Hello interval (seconds)		All zeros	E B All zeros I M M S
Database Description sequence number			
LSA header (20 bytes)			
...			
LSA header			

OSPF Database Description packet (see Forouzan figure 14.45 pg. 420)

Rather than send the entire database - send an **outline** of it:

E = 1 indicates the advertising router is an autonomous boundary router (i.e., E  $\equiv$  external)

B = 1 indicates the advertising router is an autonomous border router

I = 1 initialization flag; M = 1  $\equiv$  More flag; M/S flag: 0=slave, 1=Master

Database Description sequence number

LSA header(s) - gives information about the link - but **without** details; if details are desired they can be requested



## Slide 33



## Link State Announcement (LSA) header

0	8	16	24	31
Link state age (seconds)		Reserved	E	T
Link state ID				
Advertising router				
Link state sequence number				
Link state checksum		Length		

OSPF Link State Advertisement general header format (see Forouzan figure 14.28 pg. 410)

E = 1 indicates a stub area

T = 1 indicates router supports multiple metrics

Link state type	Link state ID	Description
1	IP address of the router	Router-LSAs
2	IP address of the designated router	Network-LSAs
3	IP address of the network	Summary-LSAs (IP network)
4	IP address of the border router	Summary to AS border router (BR)
5	IP address of the external network	AS-external-LSAs



## Link State Announcement (LSA) header

(continued)

**Advertising router** - IP address of the router advertising this message

**Link state checksum** - a Fletcher's checksum of all but age field

**Length** of the whole packet in bytes

## Slide 35



## Link state update packet

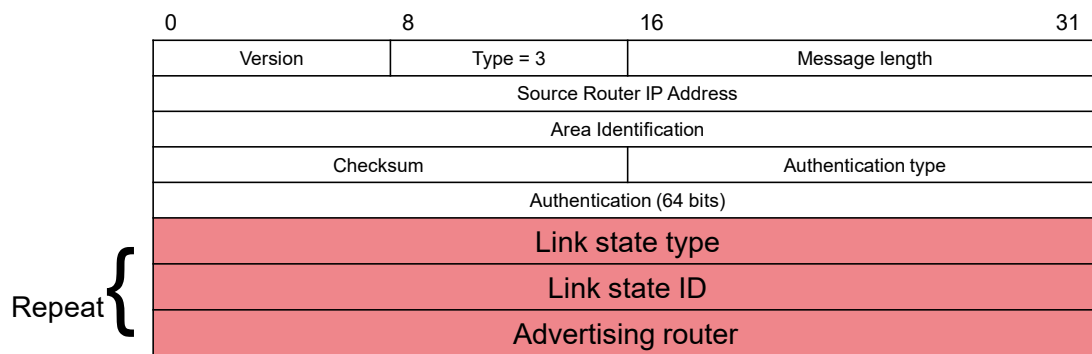
0	8	16	31
Version	Type = 4	Message length	
Source Router IP Address			
Area Identification			
Checksum		Authentication type	
Authentication (64 bits)			
Number of link state advertisements			
Link state advertisement (LSA)			
...			
Link state advertisement (LSA)			

OSPF Link state update packet (see Forouzan figure 14.27 pg. 409)

## Slide 36



## Link state request packet



OSPF Link state request packet (see Forouzan figure 14.46 pg. 420)

To ask for information about a specific route (or routes); reply is an update packet.

## Slide 37



## Link state acknowledgement packet

0	8	16	31
Version	Type = 5	Message length	
Source Router IP Address			
Area Identification			
Checksum		Authentication type	
Authentication (64 bits)			
LSA header			

OSPF Link state acknowledgement packet (see Forouzan figure 14.47 pg. 421)

To acknowledge receipt of an update packet.



## Extensions to OSPF

OSPF with traffic engineering extensions (OSPF-TE)

OSPF-TE extended with information about node resource information – see:

- Amir Roozbeh, Resource monitoring in a Network Embedded Cloud: An extension to OSPF-TE, Master's thesis, KTH, ICT/CoS, TRITA-ICT-EX-2013:85, June 2013  
<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-124367>
- A. Roozbeh, A. Sefidcon, and G. Q. Maguire Jr., 'Resource monitoring in a Network Embedded Cloud : An extension to OSPF-TE', in *Proceeding UCC '13 Proceedings of the 2013 IEEE/ACM 6th International Conference on Utility and Cloud Computing*, Dresden, Germany, Dresden, Germany: IEEE, 2013, pp. 139–146. DOI: 10.1109/UCC.2013.36



## Interdomain routing protocols

also called “Exterior Gateway Protocols (EGPs)” - used between ASs

Examples:

- EGP - an old EGP protocol
- BGP - Border Gateway Protocol

**Intradomain** routing protocols:

- Do not scale up to the large numbers of routers involved in interdomain routing (due to the huge computational resources required)
- Distance vector routing becomes unstable beyond a few hops



## Exterior Gateway Protocol (EGP)

an exterior gateway protocol with three components:

- neighbor acquisition
- neighbor reach ability, and
- routing information

EGP was designed to provide more automation in configuring routers.

EGP is similar to the distance-vector protocols, but **omits the metrics**, since EGP was designed for the internet where typically routers are connected to a backbone (with its own routing domain) via a single router.

But since there are **no** metrics, if there is more than one path, then there can be a loop!





## BGP - Border Gateway Protocol

An exterior gateway protocol to exchange routing information between routers in different ASs.

BGP version 3 defined in RFC 1267, while version 4 defined in RFC1654, RFC 1771, and RFC 4271.

BGP routers exchange routing information with other BGP routers.

For further information see:

John W. Stewart III, *BGP4: Inter-Domain Routing in the Internet*, Addison-Wesley, 1999, ISBN: 0-201-37951-1

See also: <http://www.bgpexpert.com/>

Ilijtsch van Beijnum, *BGP*, O'Reilly, 1st Edition September 2002, ISBN 0-596-00254-8

K. Lougheed and Y. Rekhter, 'Border Gateway Protocol 3 (BGP-3)', *Internet Request for Comments*, vol. RFC 1267 (Historic), Oct. 1991 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1267.txt>

P. Traina, 'BGP-4 Protocol Document Roadmap and Implementation Experience', *Internet Request for Comments*, vol. RFC 1656 (Informational), Jul. 1994 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1656.txt>

Y. Rekhter and T. Li, 'A Border Gateway Protocol 4 (BGP-4)', *Internet Request for Comments*, vol. RFC 1654 (Proposed Standard), Jul. 1994 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1654.txt>

Y. Rekhter and T. Li, 'A Border Gateway Protocol 4 (BGP-4)', *Internet Request for Comments*, vol. RFC 1771 (Draft Standard), Mar. 1995 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1771.txt>

Y. Rekhter, T. Li, and S. Hares, 'A Border Gateway Protocol 4 (BGP-4)', *Internet Request for Comments*, vol. RFC 4271 (Draft Standard), Jan. 2006 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc4271.txt>

Ilijtsch van Beijnum, web site <http://www.bgpexpert.com/>, last modified Monday, March 3, 2014 11:42:51

## Slide 42



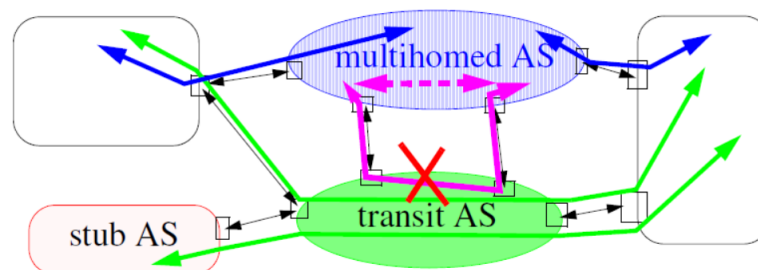
## Local vs. Transit traffic

**Local traffic:** originates or terminates in an AS

**Transit traffic:** all other traffic

⇒ 3 types of ASs

- |               |   |
|---------------|---|
| stub AS       | connected to only one other AS, carries only local traffic        |
| multihomed AS | connected to multiple ASs, but refuses transit traffic            |
| transit AS    | connected to multiple ASs, carries both local and transit traffic |





## BGP operation

BGP routers exchange information based on traffic which transits the AS, derives a graph of AS connectivity; with loop pruning.

Routing policy decisions can be enforced as to what is allowed to transit whom⇒

### **policy-based routing**

- based on economic/security/political/... considerations.
- BGP does **not** implement the policy decisions, but allows the information on which such decisions can be made to propagate as necessary

Uses **TCP** (port 179) to create a session between BGP routers:

- initially two systems exchange their entire BGP routing table,
- then they simply send updates as necessary.

BGP is a **path-vector** protocol - which **enumerates** the route to each destination (i.e., the sequence of AS numbers which a packet would have to pass through from a source to its destination) = a path vector



## BGP operation

(continued)

BGP does **not** transmit metrics.

However, each path is a list of attributes:

- **well-known attributes** (which every router must understand)
  - well-known **mandatory** attribute - must appear in the description of a route
  - well-known **discretionary** attribute - may appear, but must be recognized, in the description of a route
- **optional attributes**
  - optional **transitive** attribute - must be passed to the next router
  - optional **nontransitive** attribute - the receiving router must discard it if it does not recognize it

For examples of the use of an attribute see [RFC 1997 ] and [RFC 1998].

BGP detects failures (either links or hosts) by sending **keepalive** messages to its neighbors. Generally sent every 30 seconds and as they are only 19 bytes each  $\Rightarrow$  only **~5 bits/second** of bandwidth, but with very long lived TCP connections (semi-permanent connections)

A major feature of BGP version 4 is its ability to do **aggregation** - to handle CIDR and supernetting. For more information on aggregation see chapter 5 of [Halabi 1997].

IK1552

SPRING 2019

SLIDE 44

R. Chandra, P. Traina, and T. Li, 'BGP Communities Attribute', *Internet Request for Comments*, vol. RFC 1997 (Proposed Standard), Aug. 1996 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1997.txt>

E. Chen and T. Bates, 'An Application of the BGP Community Attribute in Multi-home Routing', *Internet Request for Comments*, vol. RFC 1998 (Informational), Aug. 1996 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1998.txt>

Bassam Halabi, *Internet routing architectures*. Indianapolis, IN: Cisco Press : New Riders Pub., 1997.



## Classless Inter-Domain Routing (CIDR)

A standard for both classless addressing and classless interdomain routing scheme (RFCs 1517 .. 1520).

- Basic concept: to allocate/collapse a block of contiguous IP addresses into a single routing table entry: (network address, count). e.g., 192.5.48.0, 192.5.49.0, 192.5.50.0 = (192.5.48.0, 3)
- **Hierarchical Routing Aggregation** minimizes routing table entries; enables "route aggregation" in which a single high-level route entry can represent many lower-level routes in the global routing tables.  
Reduces the growth of routing table.
- Allows the addresses assigned to a single organization to span multiple classed prefixes.
- Envisioned a hierarchical Internet.

CIDR addressing scheme and route aggregation has two major user impacts:

- you have to **justifying** IP Address Assignments
- get address from your ISP, i.e., **renting** them vs. being **assigned** them

Internet Engineering Steering Group and R. Hinden (Editor), 'Applicability Statement for the Implementation of Classless Inter-Domain Routing (CIDR)', *Internet Request for Comments*, vol. RFC 1517 (Historic), Sep. 1993 [Online]. Available:

<http://www.rfc-editor.org/rfc/rfc1517.txt>

Y. Rekhter and T. Li, 'An Architecture for IP Address Allocation with CIDR', *Internet Request for Comments*, vol. RFC 1518 (Historic), Sep. 1993 [Online]. Available:

<http://www.rfc-editor.org/rfc/rfc1518.txt>

V. Fuller, T. Li, J. Yu, and K. Varadhan, 'Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy', *Internet Request for Comments*, vol. RFC 1519 (Proposed Standard), Sep. 1993 [Online]. Available: [http://www.rfc-](http://www.rfc-editor.org/rfc/rfc1519.txt)

[editor.org/rfc/rfc1519.txt](http://www.rfc-editor.org/rfc/rfc1519.txt)

Y. Rekhter and C. Topolcic, 'Exchanging Routing Information Across Provider Boundaries in the CIDR Environment', *Internet Request for Comments*, vol. RFC 1520 (Historic), Sep. 1993 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1520.txt>

V. Fuller and T. Li, 'Classless Inter-domain Routing (CIDR): The Internet Address Assignment and Aggregation Plan', *Internet Request for Comments*, vol. RFC 4632 (Best Current Practice), Aug. 2006 [Online]. Available: [http://www.rfc-](http://www.rfc-editor.org/rfc/rfc4632.txt)  
[editor.org/rfc/rfc4632.txt](http://www.rfc-editor.org/rfc/rfc4632.txt)



## Redistribution of Route Information between protocols

Redistribution: allows a router running more than one routing protocol to distribute information from one protocol to another.

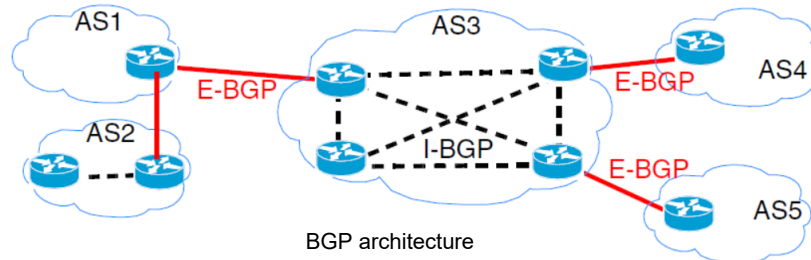
Thus at the border, a router will translate the information obtained from one routing domain and pass it to the other routing domain in the appropriate manner.

- **Advertise** (aggregated) *interior routes* to the Internet
- **Inject** (some) *exterior routes* into the interior network

Usually the redistributed routes are filtered (as not all the information needs to cross the border).



## Resulting BGP Architecture




Two kinds of BGP sessions:

- External (E-BGP) coordinates between border routers **between** ASs
- Internal BGP (I-BGP) coordinates between BGP peers **within** an AS

Note it must be a full mesh, but this does not scale  $\Rightarrow$  organization into subASs, ... .

## Slide 48



## BGP Messages

- Open
- Update
- Keepalive
- Notification

IK1552

SPRING 2019

SLIDE 48



## Slide 49



## BGP Open Message

0	7	8	15	16	23	24	31
Marker (16 bytes)							
Length				Type = 1		Version	
My autonomous system				Hold time			
BGP identifier							
Option length		Option (variable length)					

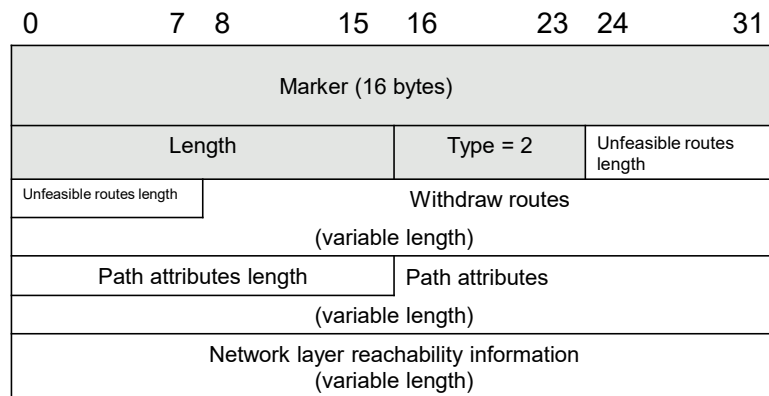
Marker (for authentication), Length, and Type are common to all BGP messages.

- Version = 4
- My autonomous system - the AS number
- Hold time - maximum time to wait for a keepalive or update, otherwise the other party is considered to be dead
- BGP identifier - identifies the router sending this message (typically its IP address)
- Option length - zero if none
- Option - options in the form (length of parameter, parameter value)

## Slide 50



## BGP Update Message



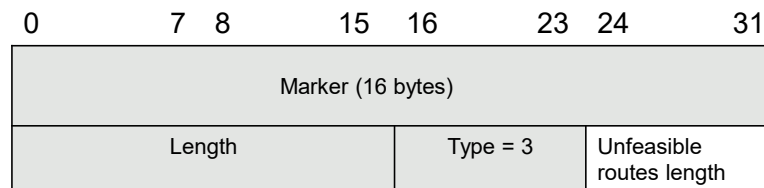
BGP Update message (see Forouzan figures 14.54 pg. 428 and 14.52 pg. 426)

- Unfeasible routes length (2 bytes) - length of next field
- Withdraw routes - list of all routes that must be deleted
- Path attributes length (2 bytes) - length of next field
- Path attributes - specifies the attributes of the path being announced
- Network layer reachability information - prefix length and IP address (to support CIDR)

## Slide 51



## BGP Keepalive Message



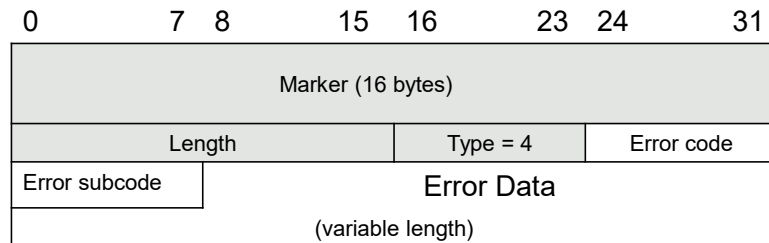
BGP Keepalive message (see Forouzan figures 14.55 pg. 429 and 14.52 pg. 426)

Sent to reassure your peer that you are still alive

## Slide 52



## BGP Notification Message



BGP Notification message (see Forouzan figures 14.56 pg. 429 and 14.52 pg. 426)

Error code (1 bytes) - category of error

- 1 = Message header error
- 2 = Open Message error
- 3 = Update Message error
- 4 = Hold time expired
- 5 = Finite state machine error
- 6 = Cease

- Error subcode (1 bytes) - the particular error in this category
- Error Data - information about this error



## Interconnections of networks

Since different networks have different users, policies, cost structure, etc.

But, the value of a network is proportional to the (number users)<sup>2</sup> {Metcalf's Law} Therefore, network operators want to connect their networks to other networks.

⇒ Internet eXchange Points (IXs or IXPs)

For a discussion of why IXPs are important see [McLaughlin 2002] and

<http://www.internetsociety.org/blog/development/2014/02/we-need-your-feedback-ixp-toolkit-and-best-practices-guide>

- No internet exchange points ⇒ no internetworking!
- Cost advantages in peering
- QoS advantages

Andrew McLaughlin, "Internet Exchange Points: Their Importance to Development of the Internet and Strategies for their Deployment: The African Example", Global Internet Policy Initiative (GIPI), 6 June 2002 (revised 3 May 2004),  
<http://www.internetpolicy.net/practices/ixp.pdf>

Jane Coffin, We Need Your Feedback: IXP Toolkit And Best Practices Guide, Internet Society, web page,  
Date published 24 February 2014,  
<http://www.internetsociety.org/blog/development/2014/02/we-need-your-feedback-ixp-toolkit-and-best-practices-guide>



## Federal Internet eXchange (FIX)

A top-level routing domain - i.e., it does not use default routes.

Each was initially built around an FDDI ring which interconnected routers from each of the operators.

Each of these routers was in turn connected to the rest of the operator's network via a high speed link (often at speeds up to 45Mbps).



Note that it need not be a physical ring, but was often an FDDI switch (such as the DEC Gigaswitch/FDDI).



## Commercial Internet eXchange (CIX)

A nonprofit trade association of Public Data Internetwork Service Providers:

- a neutral forum - for forming consensus on legislation and policies
- fundamental agree for all CIX members to interconnect with on another
- no restriction on traffic between member networks
- no “settlements” or traffic charges



## Global Internet eXchange (GIX)

Global Internet eXchange (GIX), Guy Almes, Peter Ford, Peter Lothberg. proposed in June 1992 - Stockholm D-GIX became Netnod-IX





## Further information about Sweden's & other Internet exchange points (IXPs)

- NorrNod <http://www.norrnod.se/>
- NETNOD Internet eXchange <http://www.netnod.se/>
- SOL-IX - Stockholm <http://www.sol-ix.net/>
- STHIX - Stockholm Internet eXchange <http://www.sthix.net/>
- ...

### Other useful contacts:

Swedish Network Users Society (*SNUS*): <http://www.snus.se/>

“... its goal, from the users perspective, to force the evolution and development of the networks and interconnections between networks, to arrange seminars, to exchange information between the members, and to write agreements with companies.”

North American Network Operators' Group (NANOG) <http://www.nanog.org/>

Euro-IX <https://www.euro-ix.net/> an association of (IXPs)



## Network Access Points (NAPs)

At the NAP a high-speed network or switch is used to interconnect a number of routers for the purpose of exchanging traffic.

Started with several NSF sponsored NAPs:

- Sprint NAP - in Pennsauken NJ
- PacBell NAP - in San Francisco
- Ameritech Advanced Data Services (AADS) NAP - in Chicago
- MAE-East - in Washington DC
- MAE-West - in San Jose CA

In addition to handling IP packets, NAPs were required to support InterDomain Routing Protocol (IDRP) {the ISO OSI Exterior Gateway Protocol} and route CLNP (ConnectionLess Networking Protocol) packets.



## NAPs today

Using GigE, switch fabrics, resilient packet ring (Spatial Reuse Protocol (SRP)) technology, e.g., Cisco's Dynamic Packet Transport (DPT), ... with dedicated fiber connections to/from members.

NAP managers are increasingly concerned about security, reliability, and accounting & statistics.

Various NAP have different policies, methods of dividing costs, fees, co-location of operators equipment at the NAP, etc.

List of Internet Exchange Points (IXP) at:

<http://www.datacentermap.com/ixps.html>



## Router Arbiter Project

Router Arbiter (RA) - <http://www.ra.net> (July 1994 through March 1998)

- provide a common database of route information (network topology and policies) [Routing Arbiter Database (RADb) became Merit Routing Assets Database (RADb) <http://www.radb.net/> ]
- promote stability and manageability of networks

Instead of a full mesh connection between providers, all the providers peer with a central [router server](#). A Router server (RS):

- maintains a database of all information operators need to set their routing policy (written in RIPE 181, see RFC 1786  $\Rightarrow$  Routing Policy Specification Language (RPSL) - see RFCs 2622 & 4012))
- does not forward packets or perform any switching function
- a distributed rover runs at each RS and collects information which the central network management system later queries.

T. Bates, E. Gerich, L. Joncheray, J.-M. Jouanigot, D. Karrenberg, M. Terpstra, and J. Yu, 'Representation of IP Routing Policies in a Routing Registry (ripe-81++)', Internet Request for Comments, vol. RFC 1786 (Informational), Mar. 1995 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc1786.txt>

C. Alaettinoglu, C. Villamizar, E. Gerich, D. Kessens, D. Meyer, T. Bates, D. Karrenberg, and M. Terpstra, 'Routing Policy Specification Language (RPSL)', Internet Request for Comments, vol. RFC 2622 (Proposed Standard), Jun. 1999 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc2622.txt>

L. Blunk, J. Damas, F. Parent, and A. Robachevsky, 'Routing Policy Specification Language next generation (RPSLNg)', Internet Request for Comments, vol. RFC 4012 (Proposed Standard), Mar. 2005 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc4012.txt>

C. Villamizar, C. Alaettinoglu, D. Meyer, and S. Murphy, 'Routing Policy System Security', Internet Request for Comments, vol. RFC 2725 (Proposed Standard), Dec. 1999 [Online]. Available: <http://www.rfc-editor.org/rfc/rfc2725.txt>



## Internet Routing Registry (IRR)

<http://www.irr.net/>

- a neutral Routing Registry using the Routing Policy Specification Language (RPSL).
- set of routing DBs which includes
  - RIPE Routing Registry (European ISPs) - <http://www.ripe.net/data-tools/db/the-ripe-routing-registry> & <http://www.ripe.net/data-tools/db/the-internet-routing-registry-history-and-purpose>
  - MCI Routing Registry (MCI customers)
  - CA\*net Routing Registry (CA\*net customers)
  - ANS Routing Registry (ANS customers)
  - JPRR Routing Registry (Japanese ISPs)
  - Routing Arbiter Database (RADB) (all others)
- entries are maintained by each service provider

Internet Performance and Analysis Project (IPMA) IRR Java Interface



## Euro6IX

The European IPv6 Internet Exchanges (Euro6IX) project [Morales et al. 2002] examined how to build an IPv6 exchange. It contains good examples of how to combine both switching and routing mechanisms to build a high performance internet exchange.

They also describe additional services which such an exchange might provide. These range from DNS to content distribution.

[http://www.euro6ix.org/main/e\\_objectives.php](http://www.euro6ix.org/main/e_objectives.php)

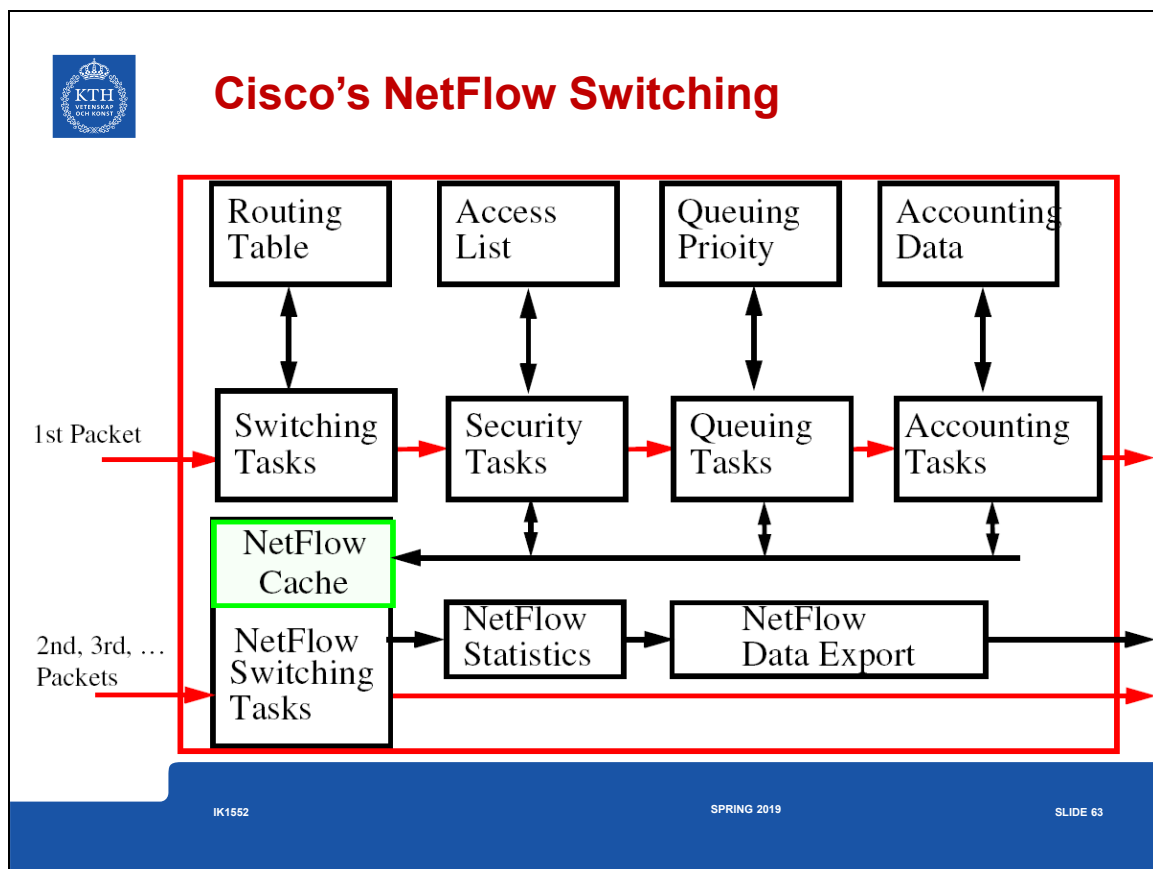
European Internet Exchange Association (**Euro-IX**)

<https://www.euro-ix.net/>

Cesar Olvera Morales, Jordi Palet Martinez, Alvaro Vives, Alain Baudot, Carlos Parada, Raffaele D'Albenzio, Mario Morelli, David Fernandez, and Tomás de Miguel, Specification of the Internal Network Architecture of each IX Point, Deliverable D2.1, Euro6IX: European IPv6 Internet Exchanges Backbone Project, IST-2001-32161, version 4.4, 30 July 2002

[http://www.euro6ix.org/Reports/public/euro6ix\\_public\\_d2\\_1\\_v4\\_4.pdf](http://www.euro6ix.org/Reports/public/euro6ix_public_d2_1_v4_4.pdf)

## Slide 63





## Flows

A flow is defined as a “uni-directional stream of packets between a given source network-layer address and port number and a specific destination network-layer address and port number”.

Since many applications use well known transport-layer port numbers, it is possible to identify **flows per user per application basis**.

There is a well defined Netflow Switching Developer's Interface which allows you to get the statistics concerning the NetFlow cache and the per flow data (the latter gives you essentially billing records). Flow information can be used in many ways, see [Visockas 2012].

A general introduction to NetFlow Switching is available at

[http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod\\_white\\_paper0900aecd80406232.html](http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html)

**Introduction to Cisco IOS NetFlow - A Technical Overview**, Last updated: May 2012

[http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod\\_white\\_paper0900aecd80406232.html](http://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html)

Configuring NetFlow Switching, a chapter in Cisco IOS Switching Services Configuration Guide,

<http://www.net130.com/book/cisco/typical/Configuring%20NetFlow%20Switching.pdf>

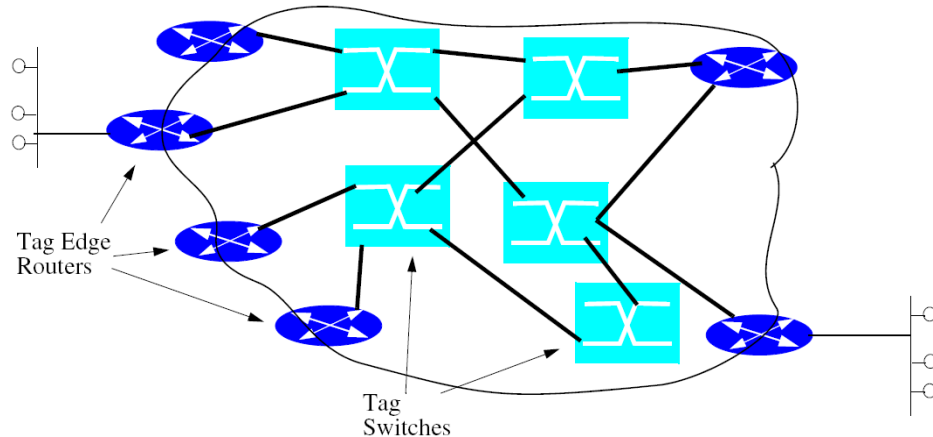
[Visockas 2012] Vilius Visockas, Comparing Expected and Real-Time Spotify Service Topology, Master's thesis, KTH, ICT/CoS, TRITA-ICT-EX-2012:63, May 2012 <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-96352>





## Cisco's Tag Switching

Combine routing with the performance of switching, based on the concept of “label swapping”, in which units of data (e.g., a packet or a cell) carry a short, fixed length label that tells switching nodes how to process the data.



## Slide 66



A Tag Edge router labels a packet based on its destination, then the Tag Switches make their switching decision based on this tag, **without** having to look at the contents of the packet.

The Tag Edge routers and Tag Switch exchange tag data using Tag Distribution Protocol (TDP).

**Basics of Tag switching:**

1. Tag edge routers and tag switches use standard routing protocols to identify routes through the internetwork.
2. Using the tables generated by the routing protocols the tag edge routers and switches assign and distribute tag information via the tag distribution protocol (TDP). When the Tag routers receive this TDP information they build a forwarding database.
3. When a tag edge router receives a packet it analyzes the network layer header, performs applicable network layer services, selects a route for the packet from its routing tables, applies a tag, and forwards the packet to the next hop tag switch.
4. The tag switch receives the tagged packet and switches the packet based **solely** on the tag.
5. The packet reaches the tag edge router at the egress point of the network, the tag is stripped off and the packet delivered as usual.



## Tag Locations

- in the Layer 2 header (e.g., in the VCI field for ATM cells)
- in the Layer 3 header (e.g., in the flow label field in IPv6) or
- in between the Layer 2 and Layer 3 headers



## Creating tags

Since tag switching decouples the tag distribution mechanisms from the data flows - the tags can be created:

- when the first traffic is seen to a destination or
- in advance - so that even the first packet can immediately be labelled.

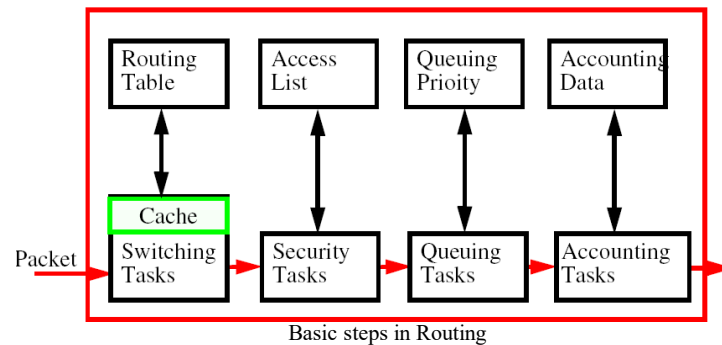
See also: Multiprotocol Label Switching (MPLS) and Generalized Multi-Protocol Label Switching (GMPLS)

Pontus Sköldström, Multi-region GMPLS control and data plane integration, Master's thesis, KTH Royal Institute of Technology, School of Information and Communication Technology, Stockholm, Sweden, COS/CCS 2008-16, August 2008

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-91851>



## Routers do more than routing



Earlier we have looked at the routing step, but today much of the excitement is in the details of the other functions. For example, in order to support various QoS features you might want to use more sophisticated queue management: Weighted Round Robin, Fair Queuing, Weighted Fair Queuing, Random Early Detection (RED), Weighted RED, ... .




## Summary

This module has discussed:

- IP routing
  - Dynamic routing protocols: RIP, OSPF, BGP, CIDR
  - Cisco's NetFlow Switching and Tag Switching
- NAPs and other interconnect points

Slide 71



# ¿Questions?

IK1552

SPRING 2019

SLIDE 71