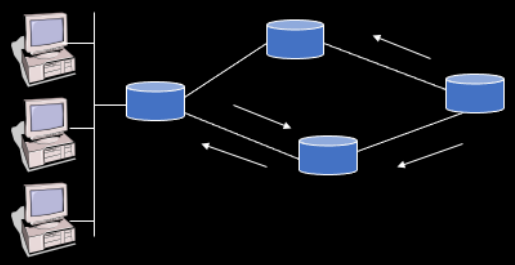
Introduction to Routing and the Dijkstra Algorithm

Miguel Rio

UCL

What is routing ?

- The propagation of connectivity information in order to build the routing tables. Routing tables are used for packet forwarding
- Final hosts usually are configured to talk with one router



UCL

Routing Tables

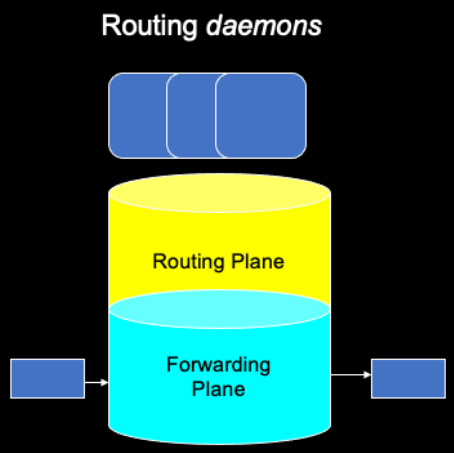
Routing

- Routing protocols act before any data packets go on the network.
- They are not involved directly in data transmission
- They are equivalent to the people who put traffic signs on roads

UCL

Routing/Forwarding Plane

- Routers are just computers with more than one network interface
- The forwarding plane forwards every packet. It needs to be fast
- On the Routing Plane, one or more Routing daemons (normal programs) talks with daemons on other routers, to exchange routing information and updates the routing tables in the forwarding plane



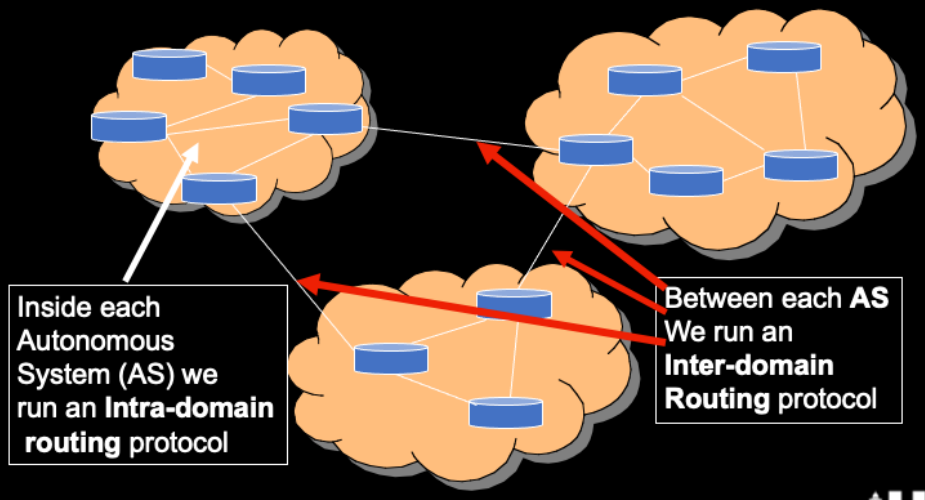
UCL

Building the Routing Tables

- Routing tables can be built by hand but…
  - In some cases there are many entries
  - They may have to change quickly when things change in the network
- Today we need Routing Protocols to automatically build and update the Routing Tables

UCL

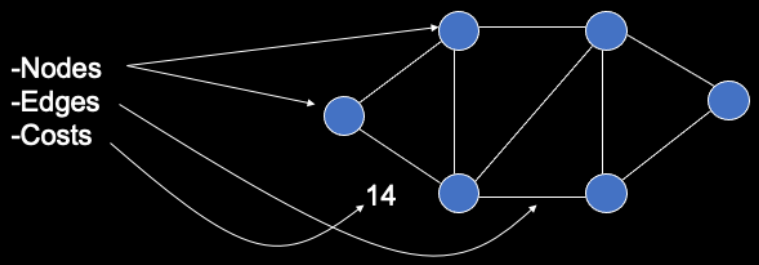
The Big Picture: Intra-domain vs inter-domain



UCL

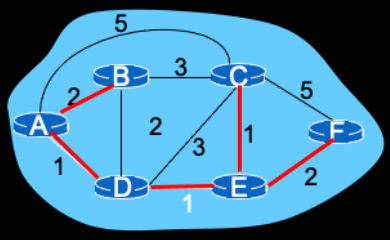
Graphs - A useful mathematical abstraction

- Several real-life problems can be approached by using graphs
- These are used in lots of fields like road traffic optimization
- Graphs have:



Abstraction

- Mathematically we model networks with graphs G=(V,E) with costs f(A,B)=Cab
- Nodes are routers, edges are links
- It becomes easier to think about algorithms
- The information is stored in a matrix
- It just represents a snapshot in time



$$\begin{bmatrix} 0 & 2 & 5 & \infty & 1 & \infty \\ 2 & 0 & 3 & 2 & \infty & \infty \\ 5 & 3 & 0 & 3 & 1 & 5 \\ 1 & 2 & 3 & 0 & 1 & \infty \\ \infty & \infty & 1 & 1 & 0 & 2 \\ \infty & \infty & 5 & \infty & 2 & 0 \end{bmatrix}$$

UCL

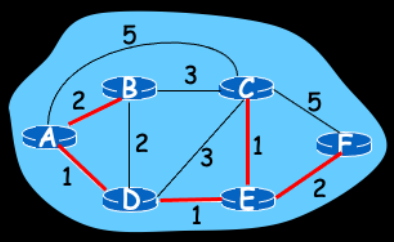
Dijkstra's Best Path algorithm: example

Step	start N	D(B),p(B)	D(C),p(C)	D(D),p(D)	D(E),p(E)	D(F),p(F)
0	A	2,A	5,A	1,A	infinity	infinity
1	AD	2,A	4,D		2,D	infinity
2	ADE	2,A	3,E			4,E
3	ADEB		3,E			4,E
4	ADEBC					4,E
5	ADEBCF					

```
Initialization:
N = {A}
for all nodes v
  if v adjacent to A
    then D(v) = c(A,v)
    else D(v) = infinity

Loop
  find w not in N such that D(w) is a minimum
  add w to N
  update D(v) for all v adjacent to w and not in N:
      D(v) = min( D(v), D(w) + c(w,v) )
  /* new cost to v is either old cost to v or known
   shortest path cost to w plus cost from w to v */

until all nodes in N
```



Summary

- Routing: Intra-domain vs Inter-domain
- Graphs
- The Dijkstra algorithm

RIP

Miguel Rio

UCL

Distance Vector vs. Link-State

Bellman-Ford	Dijkstra
Distribution of aggregate information	Distribution of information on local links
Distribution only to neighbours	Distribution to every node
↓	↓
Distance Vector Algorithms e.g.: RIP Protocol	Link-state Algorithms e.g.: OSPF Protocol

UCL

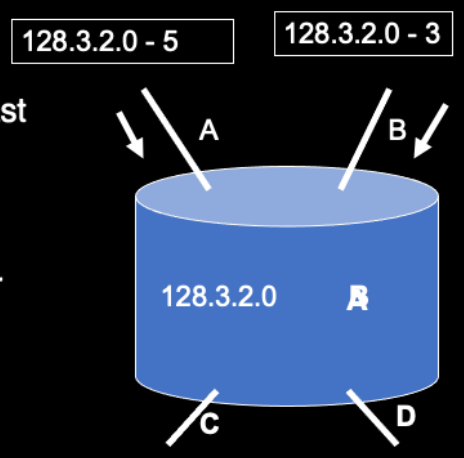
RIP - Routing Information Protocol

- Original Routing protocol. Developed in 1969
- Defined in RFC 2453
- Gained popularity with inclusion in UNIX BSD in 1982
- RIP itself is an Application level protocol running over UDP but it affects the network layer

UCL

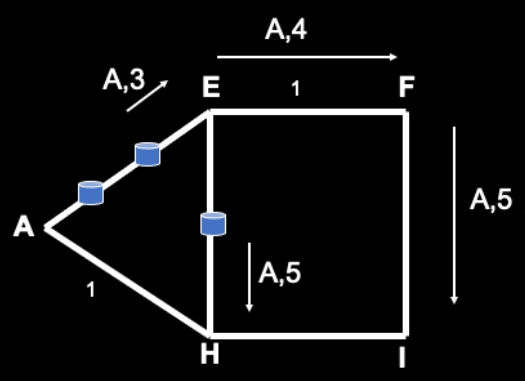
RIP Operation

- A routing running RIP broadcast a routing message every 30 seconds to all it neighbours
- Each update contains a pair <network,distance>.Where distance is usually the number of hops (in certain cases managers may increase the weight of a hop
- RIP participants update their routing tables based on these broadcasts.



UCL

RIP Example



Important:Routers never Know the topology of The network

Don't forget: router 1 Will also send a message To node **3** which will also Propagate it. Routers will Choose the best metric To reach **1**

All routers send equivalent messages

UCL

RIP Operation (contd.)

- Hysteresis - A RIP router does not change the route if the distance received is the same
- What if the first router that advertised the network fails ?
  - If a router does not receive an update after 180 seconds the route is deleted
- The maximum distance for RIP is 16. If an administrator wants to have bigger value it has to partition the network.

UCL

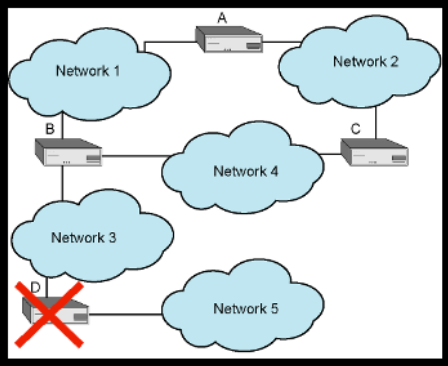
Split horizon

1 - B maintains info that it reaches network 5 through D with distance of 2

2 - Router D goes down

3 - After 180 seconds B deletes that entry

4 - … But receives announcements by C and D - slow convergence

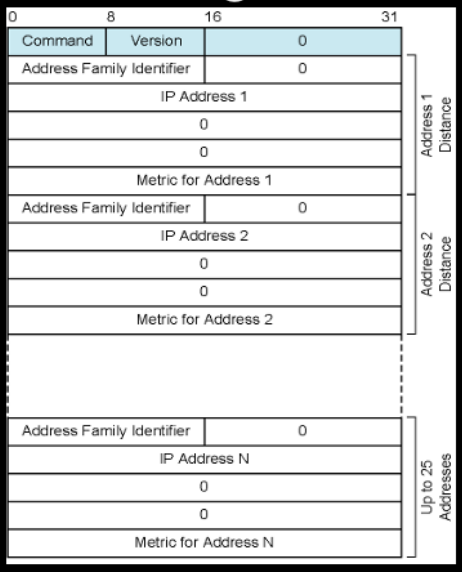


CURE: In RIP routers do not announce routes on the links these were received - Split horizon

UCL

B maintains info that it reaches network 5 through D with distance of 2
Router D goes down
After 180 seconds B deletes that entry
… But receives announcements by C and D
In RIP routers do not announce routes on the links these were received

RIP Message



- Command: 1=request 2=reply
  - Updates are replies whether asked for or not
  - Initializing node broadcasts request
  - Requests are replied to immediately
- Version: 1 or 2
- Address family: 2 for IP
- IP address: non-zero network portion, zero host portion
  - Identifies particular network
- Metric
  - Path distance from this router to network
  - Typically 1, so metric is hop count

Summary

- RIP – Routing Internet Protocol
- Split Horizon

OSPF

Miguel Rio

OSPF - Open Shortest Path First

- To encourage the use of link state routing protocols a working group of the IETF has designed OSPF
- Defined in RFCs 1245, 1246, 1247, 2328 (v2), 5340 (v3 for IPv6)

UCL

OSPF Operation

- Differences from RIP:
  - It is link-state based. Every router knows about every link in the network
  - Includes type of service routing
  - Authentication of routing messages
  - Definition of areas
  - Allows load-balancing of traffic
  - … several minor improvements…

UCL

OSPF Overview (OSPF=Flooding + Dijkstra)

Router maintains descriptions of state of local links as a directed graph

1 - Transmits updated state information to all routers it knows about using flooding

- It sends a message about each link to all the neighbours. These replicate that message to all the neighbours except the one that sent the message
- If routers receive a message they already broadcasted they just drop it

2 - Router receiving update must acknowledge

- Lots of traffic generated

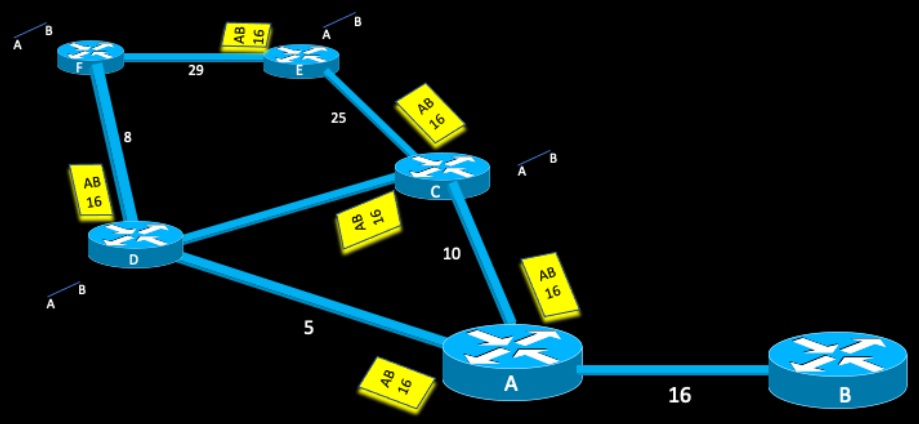
UCL

More OSPF…

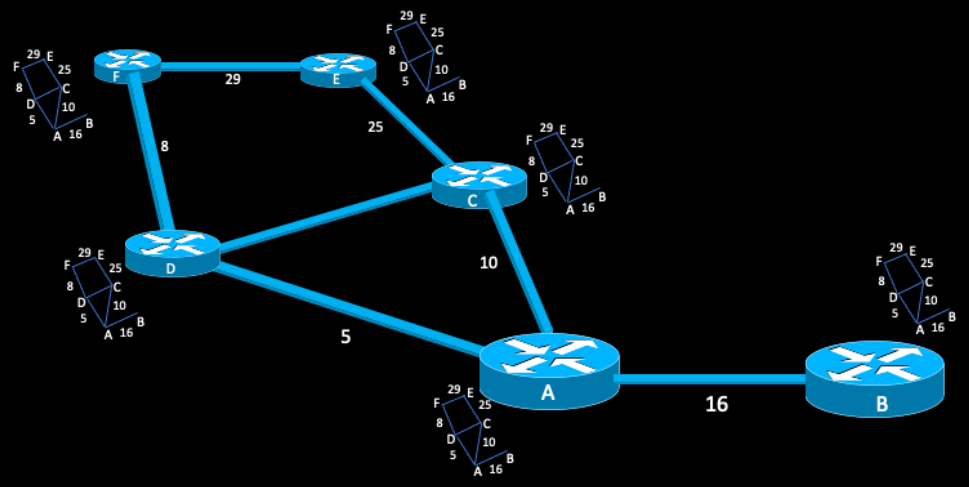
3. After receiving all the message from the network OSPF routers calculate what is the shortest path to reach all the destinations
4. With that information they calculate what is the link to be used for every network
5. Because all the routers have the same information and calculate the same paths using the same algorithm all the forwarding decisions are consistent
6. If for some reason information is not consistent packets may get looped. Example: link goes down and that information has not arrived to all the routers. Remember TTL !!!

UCL

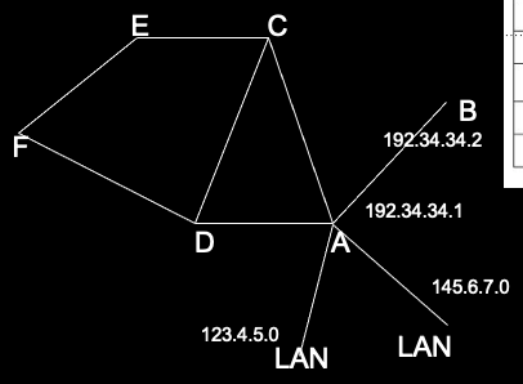
OSPF flooding



After all the messages



In more detail

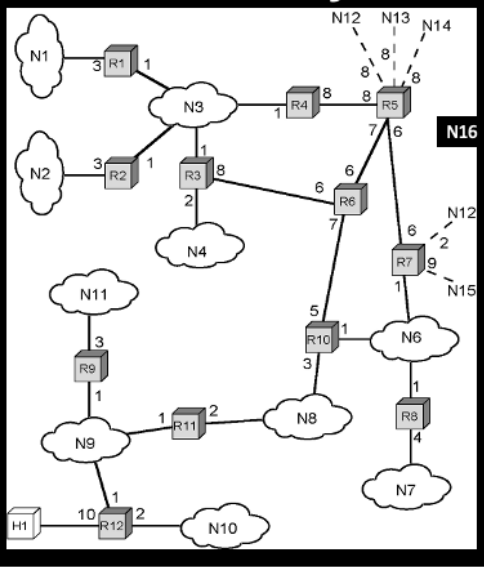


0	8	16	24	31
OSPF With Header Type=2				
Must Be Zero				
Database Sequence Number				
Link Type				
Link ID				
Advertising Router				
Link Sequence Number				
Line Checksum		Link Age		
.....				

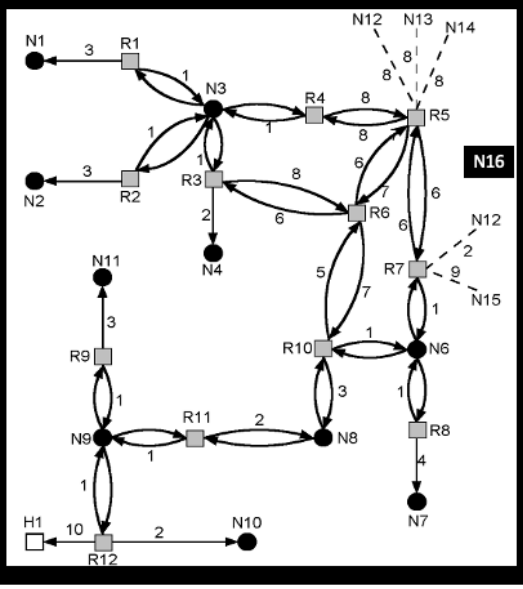
In reality OSPF announces links or networks
The type is indicated in the "Link type" of the Database Description message (above)

UCL

Sample Autonomous System



Directed Graph of Autonomous System

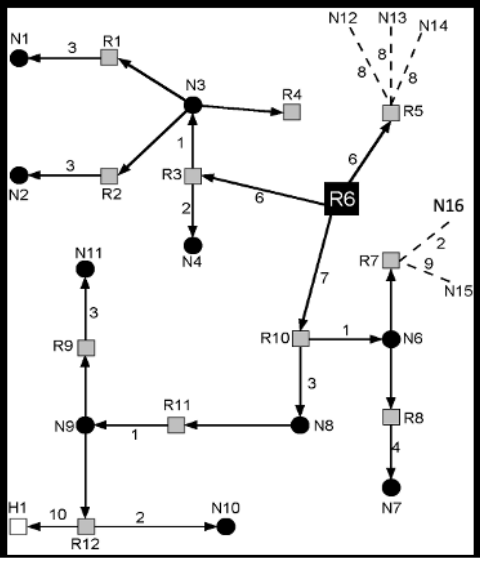


Link Costs

- Cost of each hop in each direction is called routing metric
- OSPF provides flexible metric scheme based on type of service (TOS)
  - Normal (TOS) 0
  - Minimize monetary cost (TOS 2)
  - Maximize reliability (TOS 4)
  - Maximize throughput (TOS 8)
  - Minimize delay (TOS 16)
- Each router generates 5 spanning trees (and 5 routing tables)

UCL

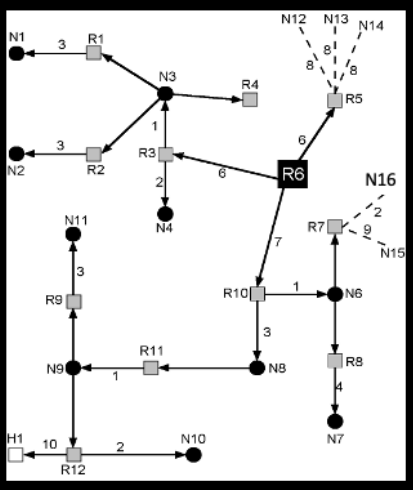
The Shortest Path Tree for Router R6



Routing Table:
N1: R3
N2:R3
N3:R3
N4:R3
N6:R10
N7:R10
N8:R10
N9:R10
N10:R10
N11:R10
N12:R5
N13:R5
N14:R5
N15:R10
N16:R10

UCL

Default Routing for R6



Routing Table:
N1: R3
N2:R3
N3:R3
N4:R3
N6:R10
N7:R10
N8:R10
N9:R10
N10:R10
N11:R10
N12:R5
N13:R5
N14:R5
N15:R10
N16:R10

Routing Table:
N1: R3
N2:R3
N3:R3
N4:R3
N12:R5
N13:R5
N14:R5
Default:R10

UCL

Summary

OSPF:

- Flooding
- Dijkstra for shortest path
- Routing table
- Default Routing

OSPF continued

Miguel Rio

UCL

Areas

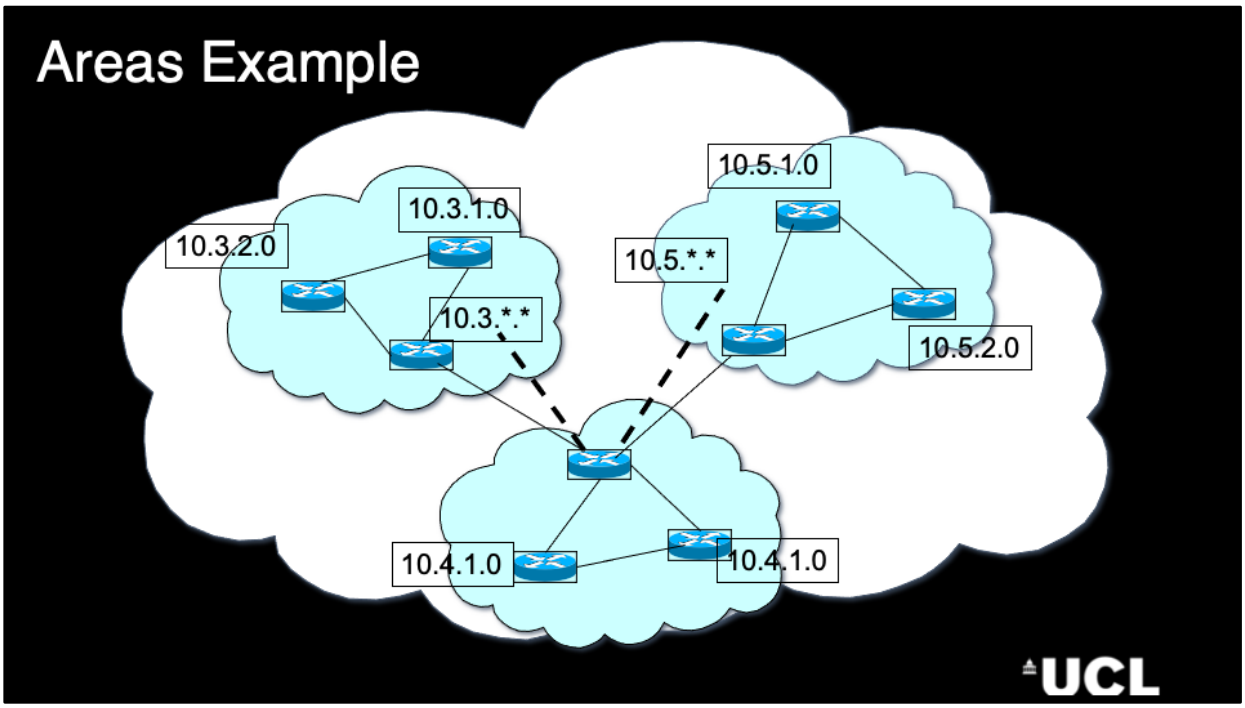
- Make large internets more manageable
- Configure as backbone and multiple areas
- Area – Collection of contiguous networks and hosts plus routers connected to any included network
- Backbone – contiguous collection of networks not contained in any area, their attached routers and routers belonging to multiple areas

UCL

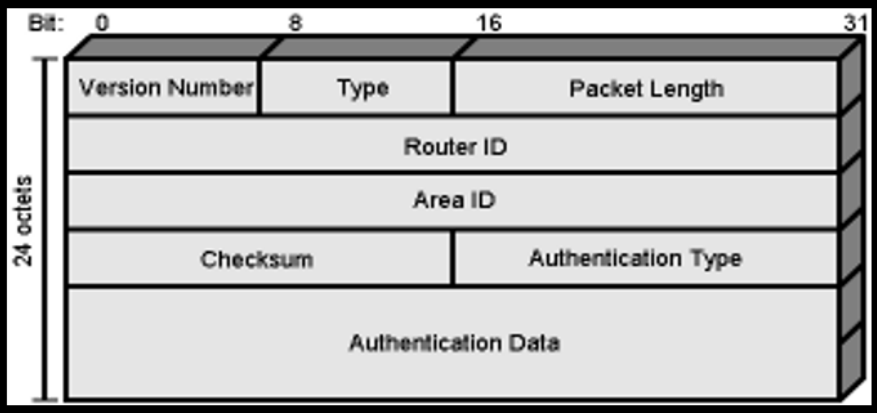
Operation of Areas

- Each area runs a separate copy of the link state algorithm
  - Topological database and graph of just that area
  - Link state information broadcast to other routers in area
  - Reduces traffic
  - Intra-area routing relies solely on local link state information

UCL

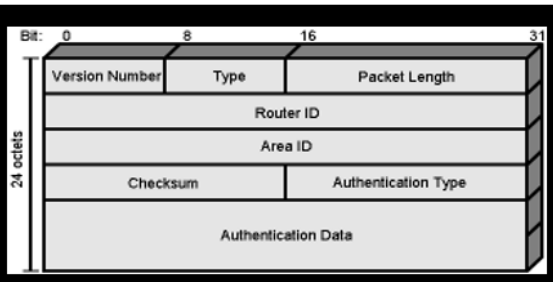

Areas Example
10.3.1.0
10.3.2.0
10.3.*.*
10.5.1.0
10.5.*.*
10.5.2.0
10.4.1.0
10.4.1.0
UCL

OSPF Packet Header



UCL

Packet Format Notes



- Version number: 2 is current
- Type: one of 5, see next slide
- Packet length: in octets including header
- Router id: this packet's source, 32 bit
- Area id: Area to which source router belongs
- Authentication type: null, simple password or encryption
- Authentication data: used by authentication procedure

UCL

OSPF Packet Types

- Hello: used in neighbour discovery
- Database description: Defines set of link state information present in each router's database
- Link state request
- Link state update
- Link state acknowledgement
- Runs directly over IP

UCL

IS-IS

- Has its roots in OSI
- Similar to OSPF: also link state. Also uses Dijkstra
- Runs directly over layer 2 (parallel to IP)
- Several minor differences:
  - OSPF supports point to multipoint links
  - OSPF supports virtual link
  - An OSPF router can belong to multiple areas whereas an IS-IS router can belong to only one area

UCL

OSPF weight calculation

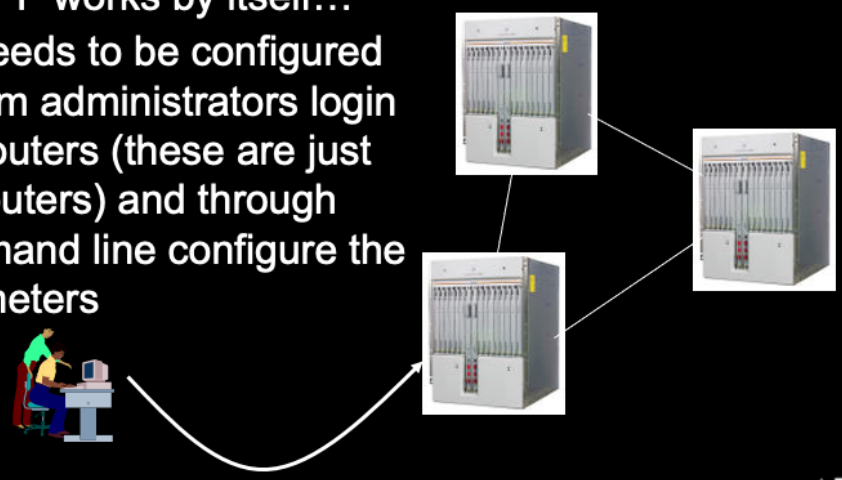
- There is no standard
- 10^8/(interface bandwidth)
- More elaborate methods:
  - Take into account traffic matrix
  - Spread the traffic over all links: minimizing the most used link

UCL

OSPF Configuration

OSPF works by itself…

But needs to be configured
System administrators login
Into routers (these are just
Computers) and through
Command line configure the
parameters



UCL

OSPF Parameters (CISCO example)

Command	Parameter
Router(config-if)#/\ip **ospf cost** *cost*	Explicitly specifies the cost of sending a packet on an OSPF interface.
Router(config-if)#/\ip **ospf retransmit-interval** *seconds*	Specifies the number of seconds between link-state advertisement (LSA) retransmissions for adjacencies belonging to an OSPF interface.
Router(config-if)#/\ip **ospf transmit-delay** *seconds*	Sets the estimated number of seconds required to send a link-state update packet on an OSPF interface.
Router(config-if)#/\ip **ospf priority** *number-value*	Sets priority to help determine the OSPF designated router for a network.
Router(config-if)#/\ip **ospf hello-interval** *seconds*	Specifies the length of time between the hello packets that the Cisco/\ IOS software sends on an OSPF interface.
Router(config-if)#/\ip **ospf dead-interval** *seconds*	Sets the number of seconds that a device must wait before it declares a neighbor OSPF router down because it has not received a hello packet.
Router(config-if)#/\ip **ospf authentication-key** *key*	Assigns a password to be used by neighboring OSPF routers on a network segment that is using the OSPF simple password authentication.
Router(config-if)#/\ip **ospf message-digest-key** *key-id* **md5** *key*	Enables OSPF MD5 authentication. The values for the *key-id* and *key* arguments must match values specified for other neighbors on a network segment.
Router(config-if)#/\ip **ospf authentication** [**message-digest** | **null**]	Specifies the authentication type for an interface.

UCL

Summary

- Routing tables are built by applications running on the routers that communicate using routing protocols
- Graphs are a mathematical abstraction used to calculate routing tables
- Each Autonomous System (AS) runs the same protocol across all its routers.
- There are two types of routing protocols: link-state and distance-vector
- OSPF is the most used interior routing protocol
- OSPF uses a flooding algorithm to propagate link-state information and the Dijkstra algorithm to calculate the best path to every other node in the network
- Given the best path, calculating the routing table is straightforward

UCL

BGP

Miguel Rio

UCL

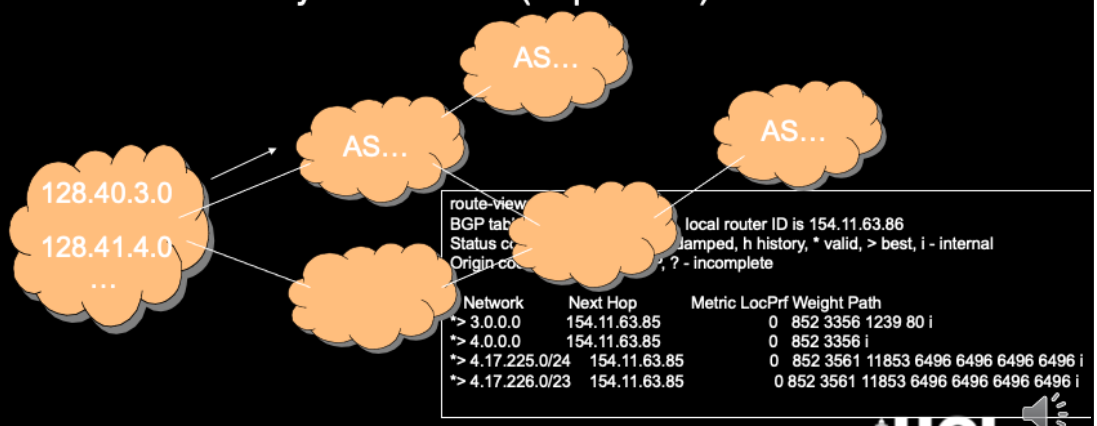
Routing Continued

- At this point you should know how IP packets traverse an Autonomous system to reach a destination inside that AS. Here all the routers are under the same administration
- Now we will look at how are packets routed over several Autonomous systems ?
  - Many more networks
  - Owned by different organizations

UCL

BGP - Border Gateway Protocol

- Connects Autonomous Systems
- Transmits reachability to networks (or prefixes)



What is an Autonomous System ?

- The Internet is not controlled by any central authority
- If an organization is big enough it can form an Autonomous System.
- These Autonomous Systems make bilateral agreements between themselves. There isn't any other form of organization !!!
- An AS can be a big company (e.g IBM), it can be a Research network (e.g. JANET) and it can be an ISP
- Each AS, runs one instance of an IGP (OSPF, RIP, etc)

UCL

Autonomous System Numbers

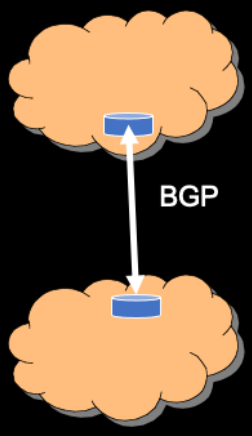
Name	AS Number	Degree
Sprint	1239	1,710
UUNET Technologies, Inc.	701	2,346
AT&T WorldNet Services	7018	1,917
Level 3 Communications, LLC	3356	1,228
Qwest	209	1,105
Cogent Communications	174	840
Global Crossing	3549	634
Savvis	3561	568
MCI EMEA	702	548
SBC Internet Services	7132	545
Swisscom Enterprise Solutions Ltd	3303	545

Allocated by IANA www.iana.org

UCL

Border Gateway Protocol (BGP)

- Allows routers (gateways) in different ASs to exchange routing information
- Messages sent over TCP
- Three functional procedures
  - Neighbour acquisition
  - Neighbour reachability
  - Network reachability

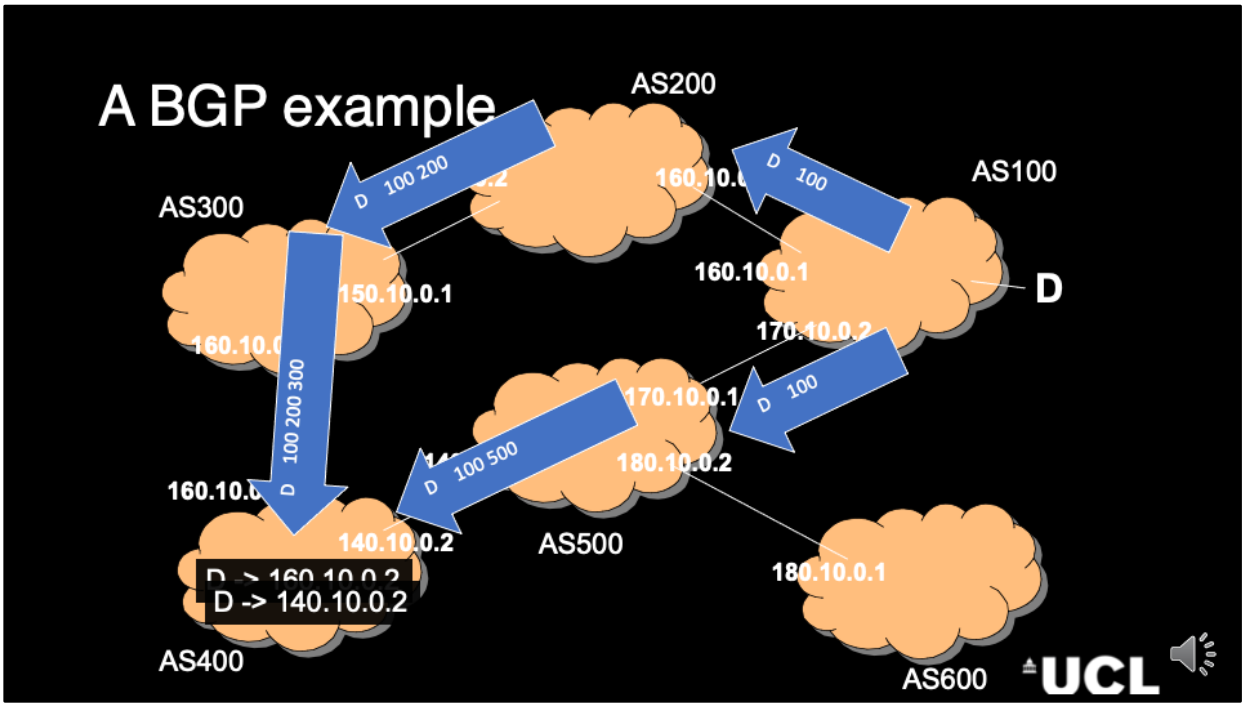


UCL

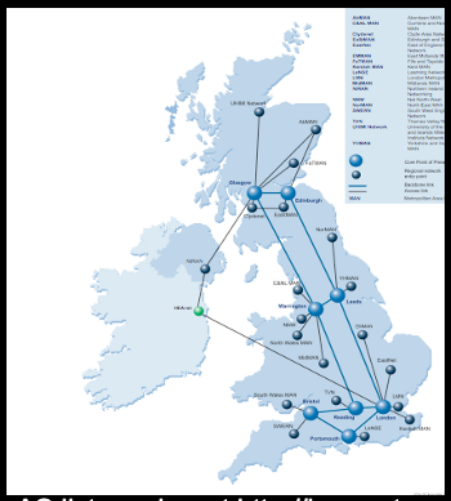
BGP Messages

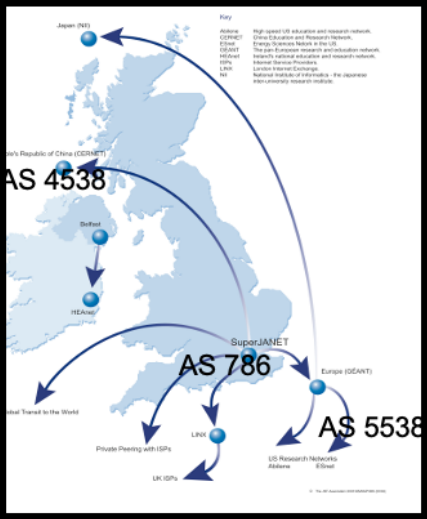
- Open
  - Start neighbour relationship with another router
- Update
  - Transmit information about single route
  - List multiple routes to be withdrawn
- Keepalive
  - Acknowledge open message
  - Periodically confirm neighbour relationship
- Notification
  - Send when error condition detected

UCL


A BGP example
AS200
AS100
AS300
AS400
AS500
AS600
D
D 100 200
D 100
D 100 200 300
D 100 500
D 100
150.10.0.1
160.10.0.1
170.10.0.2
170.10.0.1
180.10.0.2
140.10.0.2
180.10.0.1
D -> 140.10.0.2
UCL

UCL -> UK -> World



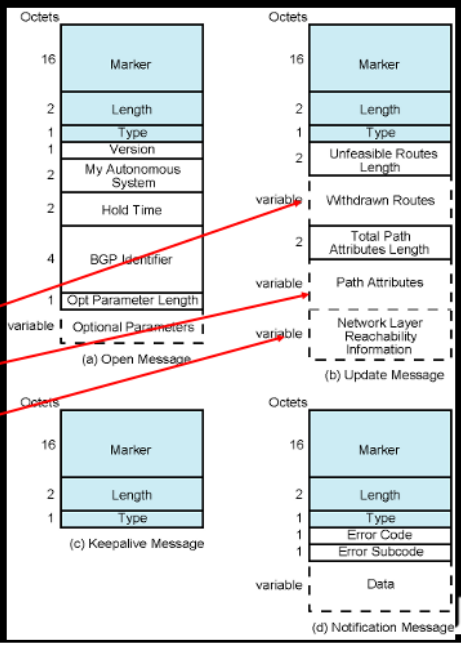


AS list number at http://bgp.potaroo.net/cidr/autnums.html

UCL

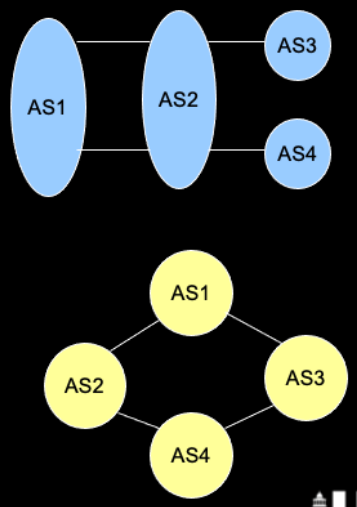
BGP Message Formats

- Four types of messages:
  - OPEN - Starts a BGP peering
  - KEEPALIVE - Keeps alive the connection
  - UPDATE - updates any information about network reachability using
    - Withdrawn Routes
    - Attributes
    - Network Layer Reachability (networks/prefixes)
  - NOTIFICATION - notifies of any protocol errors. Must close TCP connection after it is sent



Most used Attributes

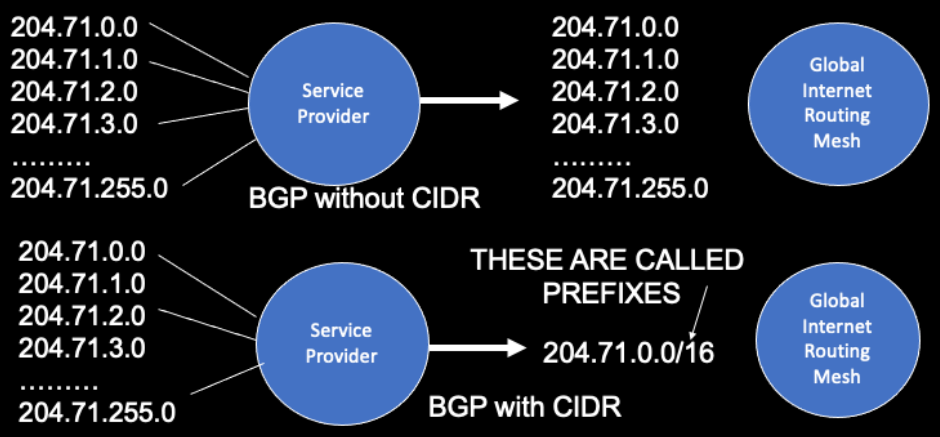
- ORIGIN: Tells how a route was learned (1 - IGP, 2 iBGP, 3 - Other
- ASPATH: List of AS numbers that packets will traverse for this announcement
- NEXT-HOP: IP address of the router that packets need to be sent
- MULTI-EXIT DESCRIMINATOR (MED): When a AS has several options to a given destination, it may include information for preferring one of them
- LOCAL-PREF: Allows a specific AS to signal internal preference for a route to a destination (when there is more than one



UCL

BGP and CIDR

CIDR allows BGP to transmit much less information due to Aggregation of routes

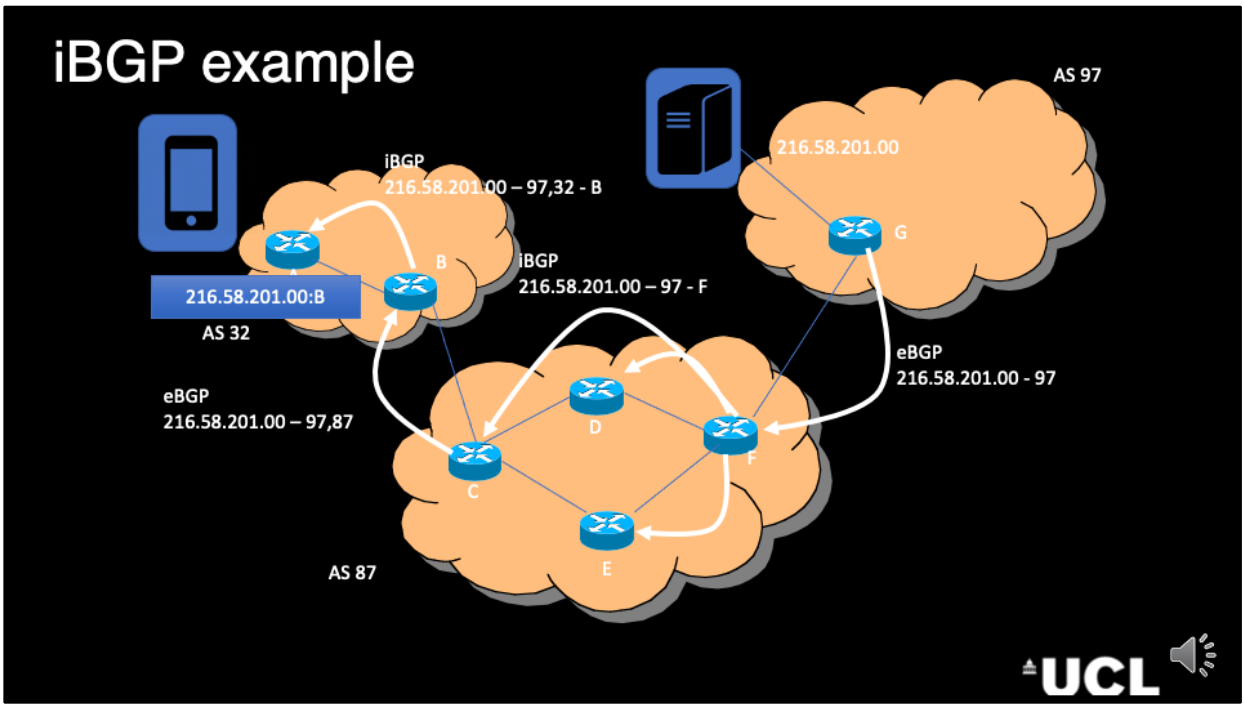


UCL

BGP= eBGP+ iBGP

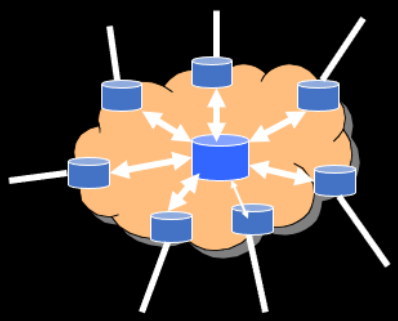
- When we talk about BGP we usually mean eBGP
- iBGP it is the part of iBGP concerned with transmitting external connectivity information inside the Autonomous System
- IT IS NOT an Interior Routing Protocol. Routers will not learn how to reach a router inside their network through iBGP

UCL


iBGP example
AS 97
216.58.201.00
G
iBGP
216.58.201.00 – 97,32 - B
B
iBGP
216.58.201.00 – 97 - F
216.58.201.00:B
AS 32
eBGP
216.58.201.00 - 97
eBGP
216.58.201.00 – 97,87
D
F
C
E
AS 87
UCL

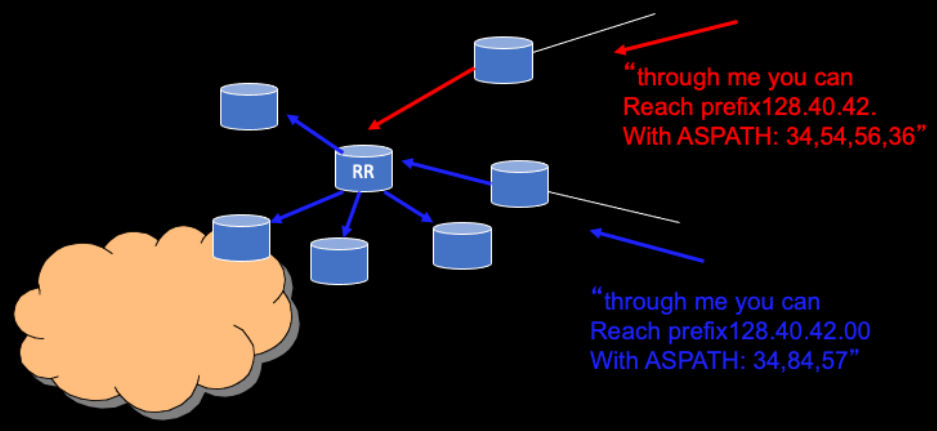
BGP Route Reflectors

- Some Autonomous Systems have a big number of external connections (e.g. Tier1 ISPs)
- If every router transmits every message received to every router in the AS there will be lots of traffic generated
- iBGP uses sometimes **Route Reflectors**. These receive the messages and send summaries to all the routers



UCL

Route Reflector Example



For ASes with more than 100 nodes
It is recommended to use a route reflector instead of a full mesh

UCL

Which Route to Choose ?

Every router in an AS needs now to decide which route to take to all prefixes based on the following rules (by this order):

1. Looks at LOCAL-PREF
2. The route with less ASes in the AS list
3. Looks at the MED (Multi-Exit Discriminator)
4. IGP distance to the NEXT-HOP
5. If all routes were learned through iBGP, choose the the neighbour with the lowest BGP identifier

UCL

Summary

- BGP implements Inter-domain routing
- eBGP and iBGP
- Attributes
- Route selection

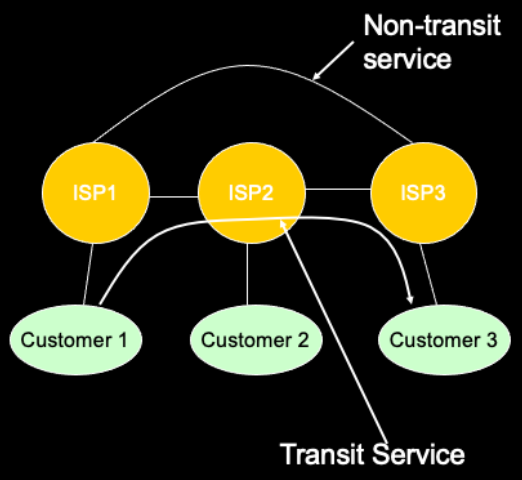
UCL

BGP Continued

Miguel Rio

UCL

Transit vs non Transit



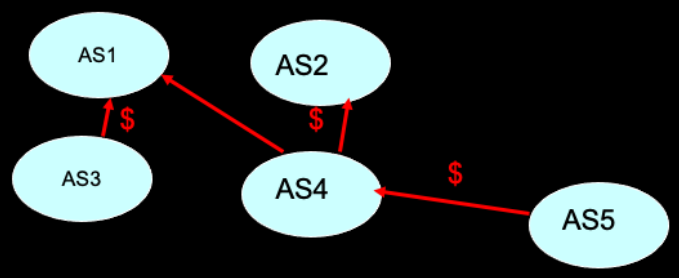
- When a ISP is providing connectivity to its clients, it is also providing connectivity to clients of other ISPs !!!
- Imagine that traffic from Customer 1 was going to Customer 3 via ISP2
- Neither C1 or C3 are paying ISP2 …ISP2 would be carrying traffic that was not paid
- Two options: Transit Service and Non-transit service

BGP Policies

- In theory, BGP allows each domain to define its own routing policy
- In practice, there are two common policies:
  - customer-provider peering
    - Customer C buys Internet connectivity from provider P
  - shared-cost peering
    - Domains x and y agree to exchange packets by using a direct link or through an interconnection point

UCL

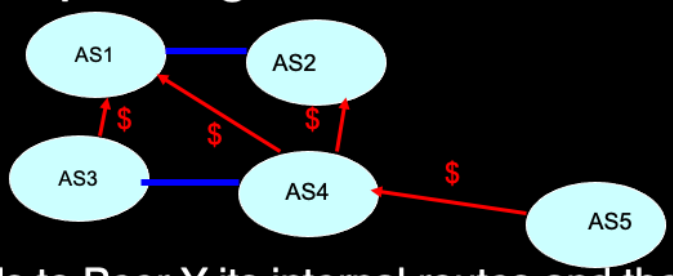
Customer-provider peering



- Customer sends to its provider its internal routes and the routes learned from its own customers => Provider will advertise those routes to the entire Internet to allow anyone to reach the Customer
- Provider sends to its customers all known routes => Customer will be able to reach anyone on the Internet

UCL

Shared-cost peering



- Peer X sends to Peer Y its internal routes and the routes learned from its own customers
  - Peer Y will use shared link to reach Peer X and Peer X's customers
  - Peer X's providers are not reachable via the shared link

UCL

Traffic cost models

- Fixed Cost: One fixed payment per month
- Linear Cost: Payment proportional to traffic
- 95th percentile:
  1. Bandwidth is measured (or sampled) from the switch or router and recorded in a log file. In most cases, this is done every 5 minutes.
  2. At the end of the month, the samples are sorted from highest to lowest, and the top 5% (which equal to approximately 36 hours of a 30-day billing cycle) of data is thrown away.
  3. The next highest measurement becomes the billable use for the entire month.

UCL

Routing Policies

- Routing policies implement business relationships between domains
- The routing policy of a domain is implemented via the route filtering mechanism on BGP routers:
  - Inbound filtering: Upon reception of a route from a peer, a BGP router decides whether the route is acceptable, and if so whether to change some of its attributes.
  - Outbound filtering: Before sending its best route towards a destination, a BGP router decides which peers should receive this route and whether to change some of its attributes before sending it.

UCL

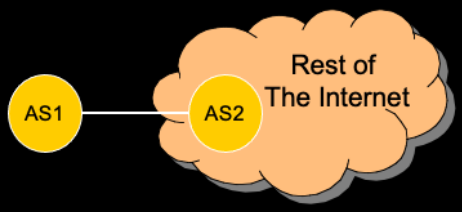
Route Instabilities

Several types of Instabilities affect the behaviour of BGP:

- Instabilities of the IGP
- Hardware Failures
- Software Problems
- Insufficient Horsepower
- Insufficient memory
- Network Upgrades
- Human Error
- Backup Link Overloads

UCL

Route Flap Dampening

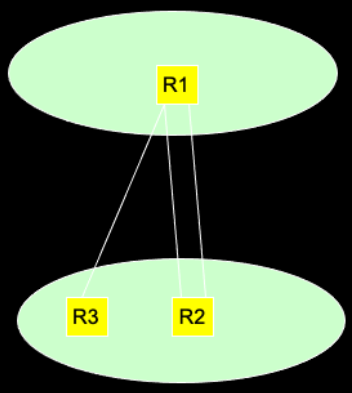


- Imagine that the link AS1-AS2 goes up and down frequently
- This generates disturbances in the entire Internet
- This could be aggravated if this is done intentionally
- In this case AS2 does Route Flap Dampening. That is it stops advertising routes to AS1 to the rest of the Internet.

UCL

Redundancy with Multi-homing

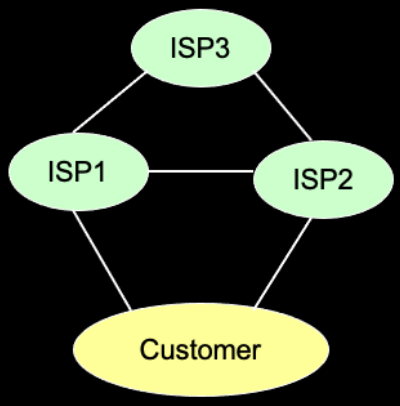
- Multi-homing refers to a single network having more than one connection to the Internet
- This can be done mainly for two reasons:
  - Load-balancing
  - Fault-tolerance
- Two types
  - Single provider
  - Multiple providers



UCL

Multi-homing with Multiple Providers

- More interesting and useful is when a customer has two providers
- Multi-homing brings some problems:
  - Destroys route symmetry
  - May cause packet reordering
  - Makes firewalls more difficult
  - Impact on the BGP tables: CIDR becomes difficult



UCL

Load Balancing

- Inbound Load Balancing can be achieved by announcing some of our routes in one link and some in the other(s)
- An easy way of achieving Outbound Load Balancing is to use a Route Reflector. The Route Reflector can be configured to announce some routes to some of our routers and other routes to the others

UCL

Interaction with the IGP

- The interaction of the IGP with BGP needs to be careful thought by the network administrator
- You can have the default that all your networks are announced on all the BGP peerings or you can decide which areas are announced to
- You can even decide not to announce some networks (for example networks used on the administration of the AS)
- In certain circumstances networks obtained from BGP can be inserted into the IGP: Risky !!

UCL

Anycast

- Anycast is another message paradigm: A packet is sent to the closest member of a group
- Today in the IPv4 internet this is achieved by announcing the same address/prefix from several locations in the Internet. BGP will redirect packets from different sources to the closest destinations
  - e.g. DNS root servers

https://arstechnica.com/information-technology/2012/10/one-big-cluster-how-cloudflare-launched-10-data-centers-in-30-days/

UCL

Real Life BGP Information

```
route-server>sh ip bgp
BGP table version is 2735630693, local router ID is 193.196.190.135
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*>i3.0.0.0          213.248.79.41          0    100      0 1299 701 703 80 i
*>i4.0.0.0          213.248.79.41          0    100      0 1299 3356 i
*>i4.23.114.0/24    213.248.79.41          0    100      0 1299 174 21889 ?
*>i4.23.180.0/24    213.248.79.41          0    100      0 1299 6128 30576 i
*>i4.36.200.0/21    213.248.79.41          0    100      0 1299 3549 14135 i
* i4.79.181.0/24    188.1.38.53            0     50      0 680 3549 3549 3549 3549 3549 3549 3549 10310 14780 i
*>i                 188.1.38.205           0     50      0 680 3549 3549 3549 3549 3549 3549 10310 14780 i
*>i6.1.0.0/16       213.248.79.41          0    100      0 1299 701 668 i
*>i6.2.0.0/22       213.248.79.41          0    100      0 1299 701 668 i
*>i6.3.0.0/18       213.248.79.41          0    100      0 1299 701 668 i
*>i6.4.0.0/16       213.248.79.41          0    100      0 1299 701 668 i
*>i6.5.0.0/19       213.248.79.41          0    100      0 1299 701 668 i
*>i6.8.0.0/20       213.248.79.41          0    100      0 1299 701 668 i
*>i6.9.0.0/20       213.248.79.41          0    100      0 1299 701 668 i
*>i6.10.0.0/15      213.248.79.41          0    100      0 1299 701 668 i
```

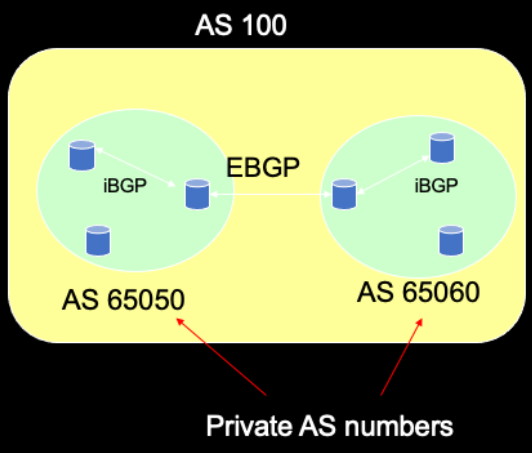
What Network is packet Going ?

AS Path

UCL

Confederations

- Another way of dealing with big ASes is to break them into confederations.
- A confederation is an AS broken into sub-ASes
- These can run different IGPs
- Used in company acquisitions

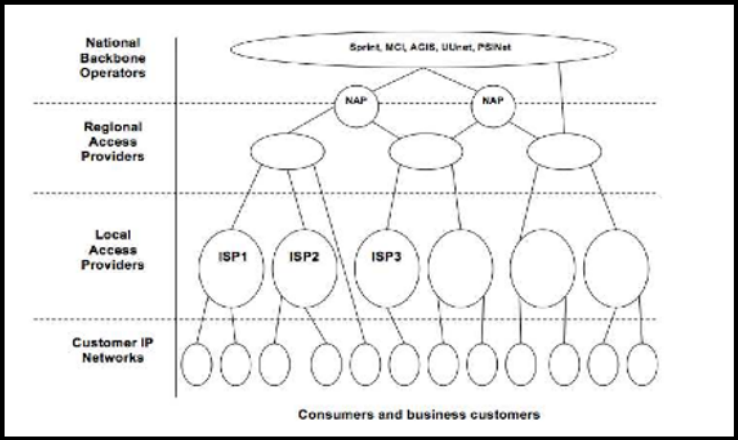


Forwarding Summary: Longest Prefix Match

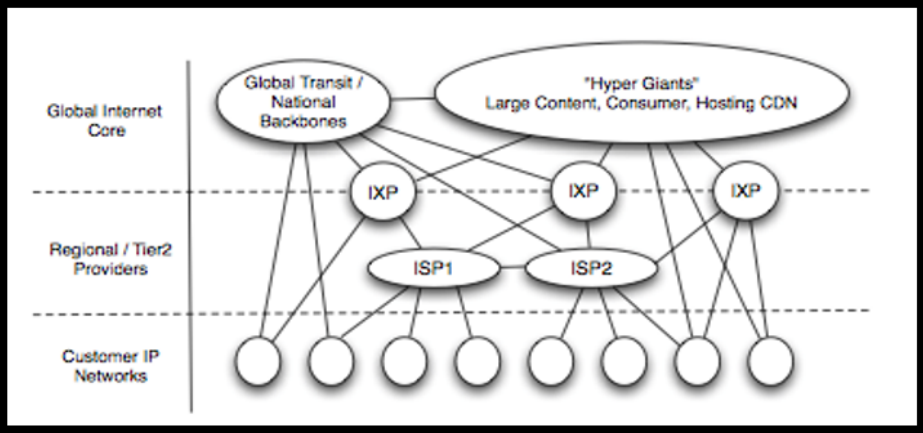
- Since with prefixes/CIDR there is not a fixed distinction between “network” and “host” routes select the next hop using Longest Prefix Match
- Imagine that the routing table has two entries
  - 138.39.0.0/16 -> Link 1
  - 138.39.16.0/24 -> Link 2
- If a packet arrives with destination address 138.39.16.32 the router uses link2 because 24 bits match against 16

UCL

Old Internet



New Internet



http://ccr.sigcomm.org/online/files/p75_0.pdf

UCL

Summary

- Relationships between ASes
- Traffic cost models
- Routing policies
- Route Flat Dumpening
- Multi-homing
- Anycasting
- Longest Prefix Match

UCL