# Inter-Domain Routing

IETF 70
Geoff Huston
APNIC

December 2007

# Agenda

- Scope
- Background to Internet Routing
- BGP
- IDR
- Views, Opinions and Comments

# Agenda

- Scope
- Background to Internet Routing
- BGP
- Current IETF Activities
- Views, Opinions and Comments

# Today, lets talk about ...

- How self-learning routing systems work
- The Internet's routing architecture
- The design of BGP as our current IDR of choice
- BGP features
- Inter-Domain Routing
- Possible futures, research topics and similar

# We won’t be talking about ...

- How to write a BGP implementation
- How to configure the control knobs on your favourite vendor’s BGP
- Operating your network
- Peering and Transit
- Debugging your favourite routing problem!

# Agenda

- Scope
- Background to Internet Routing
- BGP
- Current IETF Activities
- Views, Opinions and Comments

# Background to Internet Routing

- The routing architecture of the Internet is based on a decoupled approach to:
  - Addresses
  - Forwarding
  - Routing
  - Routing Protocols
- There is no single routing protocol, no single routing configuration, no single routing state and no single routing management regime for the entire Internet
- The routing system is the result of the interaction of a collection of many components, hopefully operating in a mutually consistent fashion!

# IP Addresses

- IP Addresses are not locationally significant
  - An address does not say “where” a device may be within the network
  - An address does not determine how a packet is passed across the network
  - Any address could be located at any point within the network
- It’s the role of the *forwarding system* to direct packets to this location

# Forwarding

- Every IP routing element is equipped with one (or more!) forwarding tables.
  - The forwarding table contains mappings between address prefixes and an outgoing interface
  - Switching a packet involves a lookup into the forwarding table using the packet's destination address, and queuing the packet against the associated output interface
  - End-to-end packet forwarding relies on mutually consistent populated forwarding tables held in every routing element
- The role of the *routing system* is to maintain these forwarding tables

# Routing

- The routing system is a collection of switching devices that participate in a self-learning information exchange (through the operation of a routing protocol)

- There have been many routing protocols, there are many routing protocols in use today, and probably many more yet to come!
- Routing protocols differ in terms of applicability, scale, dynamic behaviour, complexity, style, flavour and colour

# Routing Approaches

- All self-learning routing systems have a similar approach:
  You tell me what you know and I'll tell you what I know!

- All routing systems want to avoid:
  - Loops
  - Dead ends
  - Selection of sub-optimal paths

- The objective is to support a distributed computation that produces consistent "best path" outcomes in the forwarding tables at every switching point, at all times
  - where "best" is a flexible term requiring consistent interpretation

# Distance Vector Routing

- I’ll tell you my “best” route for all known destinations
- You tell me yours
- If any of yours are better than mine I’ll use you for those destinations
- And I’ll let all my other neighbours know

# Relative properties - DV

- Distance Vector routing
  - Is simple
  - Can be very verbose (and slow) as the routing system attempts to converge to a stable state
  - Finds it hard to detect the formation of routing loops
  - Ensures consistent forwarding states are maintained (even loops are consistent!)
  - Can't scale

# Link State Routing

- I’ll tell everyone about all my connections (links), with link up/link down announcements
- I’ll tell everyone about all the addresses I originate on each link

- I’ll listen to everyone else’s link announcements
- I’ll build a topology of every link (map)
- Then I’ll compute the shortest path to every address

- And trust that everyone else has assembled the same map and performed the same relative path selection

# Relative properties - LS

- Link State Routing
  - Is more complex
  - Converges extremely quickly
  - Should be loop-free at all times
  - Does not guarantee consistency of outcomes
  - Relies on a “full disclosure” model and policy consistency across the routing domain
  - Still can’t scale

# Routing Structure

- The Internet's routing architecture uses a 2-level hierarchy, based on the concept of a *routing domain* ("Autonomous System")
- A "domain" is an interconnected network with a single exposed topology, a coherent routing policy and a consistent metric framework

- *Interior Gateway Protocols* are used *within* a domain
- *Exterior Gateway Protocols* are used to *interconnect* domains

# IGPs and EGPs

- IGPs
  - Distance Vector: RIPv1, RIPv2, IGRP, EIGRP
  - Link State: OSPF, IS-IS
- EGPs
  - Distance Vector: EGP, BGPv3 BGPv4

# Agenda

- Scope
- Background to Internet Routing
- BGP
- Current IETF Activities
- Views, Opinions and Comments

# Why BGP?

- Simple protocol to implement and operate
- Very simple distance metric
- Occludes local policies from external inspection
- Limited inter-SP coordination required
- Mature deployment

# Border Gateway Protocol - BGP

- Developed as a successor to EGP
  - Version 1
    - RFC1105, Experimental,  June 1989
  - Version 2
    - RFC1163, RFC 1164, Proposed Standard, June 1990
  - Version 3
    - RFC1267, Proposed Standard, October 1991
  - Version 4
    - RFC1654, Proposed Standard, July 1994
    - RFC1771, Draft Standard, March 1995
    - RFC4271, Draft Standard, January 2006

# BGPv4

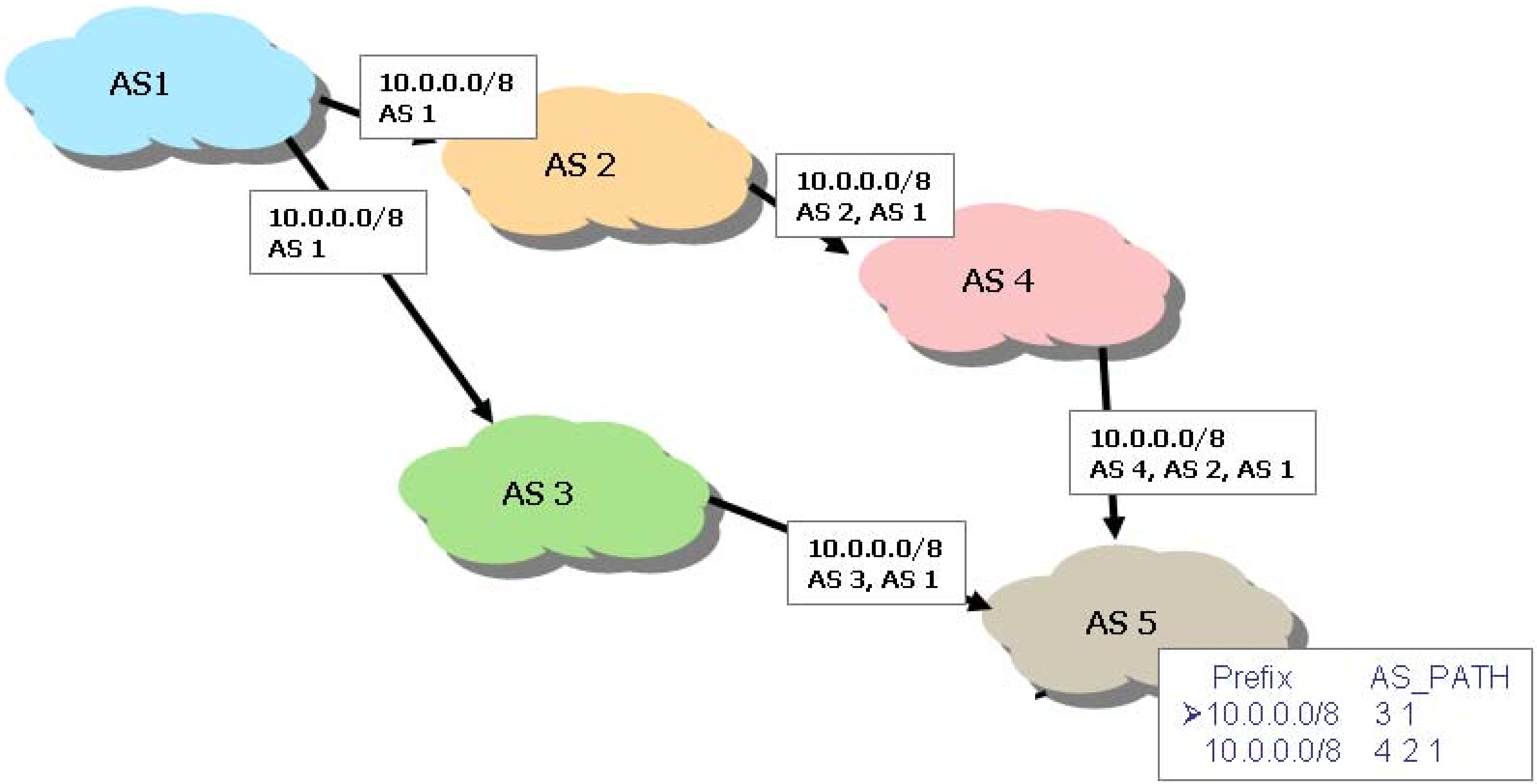
- BGP is a Path Vector Distance Vector exterior routing protocol
- Each routing object is an address and an attribute collection
  - Attributes: AS Path vector, Origination, Next Hop, Multi-Exit-Discriminator, Local Pref, ...
- The Path Vector is a vector of AS identifiers that form a viable path of AS transits from this AS to the originating AS
  - Although the Path Vector is only used to perform loop detection and route comparison for best path selection

# BGP is an inter-AS protocol

- Not hop-by-hop
- Addresses are bound to an “origin AS”
- BGP is an “edge to edge” protocol
  - BGP speakers are positioned at the inter-AS boundaries of the AS
  - The “internal” transit path is directed to the BGP-selected edge drop-off point
  - The precise path used to transit an AS is up to the IGP, not BGP

- BGP maintains a local forwarding state that associates an address with a next hop based on the “best” AS path
  - Destination Address -> [*BGP Loc-RIB*] -> Next Hop address
  - Next_Hop address -> [*IP Forwarding Table*] -> Output Interface

# BGP Example


AS1
10.0.0.0/8
AS 1
AS 2
10.0.0.0/8
AS 1
10.0.0.0/8
AS 2, AS 1
AS 4
AS 3
10.0.0.0/8
AS 4, AS 2, AS 1
10.0.0.0/8
AS 3, AS 1
AS 5
Prefix AS_PATH
➢10.0.0.0/8 3 1
10.0.0.0/8 4 2 1

# BGP Example

```
bgpd# show ip bgp
BGP table version is 0, local router ID is 203.119.0.116
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale, R Removed
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*> 0.0.0.0          193.0.4.28                             0 12654 34225 1299 i
*  3.0.0.0          193.0.4.28                             0 12654 7018 701 703 80 i
*>                  202.12.29.79                           0 4608 1221 4637 703 80 i
*> 4.0.0.0          193.0.4.28                             0 12654 7018 3356 i
*                   202.12.29.79                           0 4608 1221 4637 3356 i
*> 4.0.0.0/9        193.0.4.28                             0 12654 7018 3356 i
*                   202.12.29.79                           0 4608 1221 4637 3356 i
*> 4.23.112.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.23.113.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.23.114.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.36.116.0/23    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.36.116.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.36.117.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.36.118.0/24    193.0.4.28                             0 12654 7018 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
```

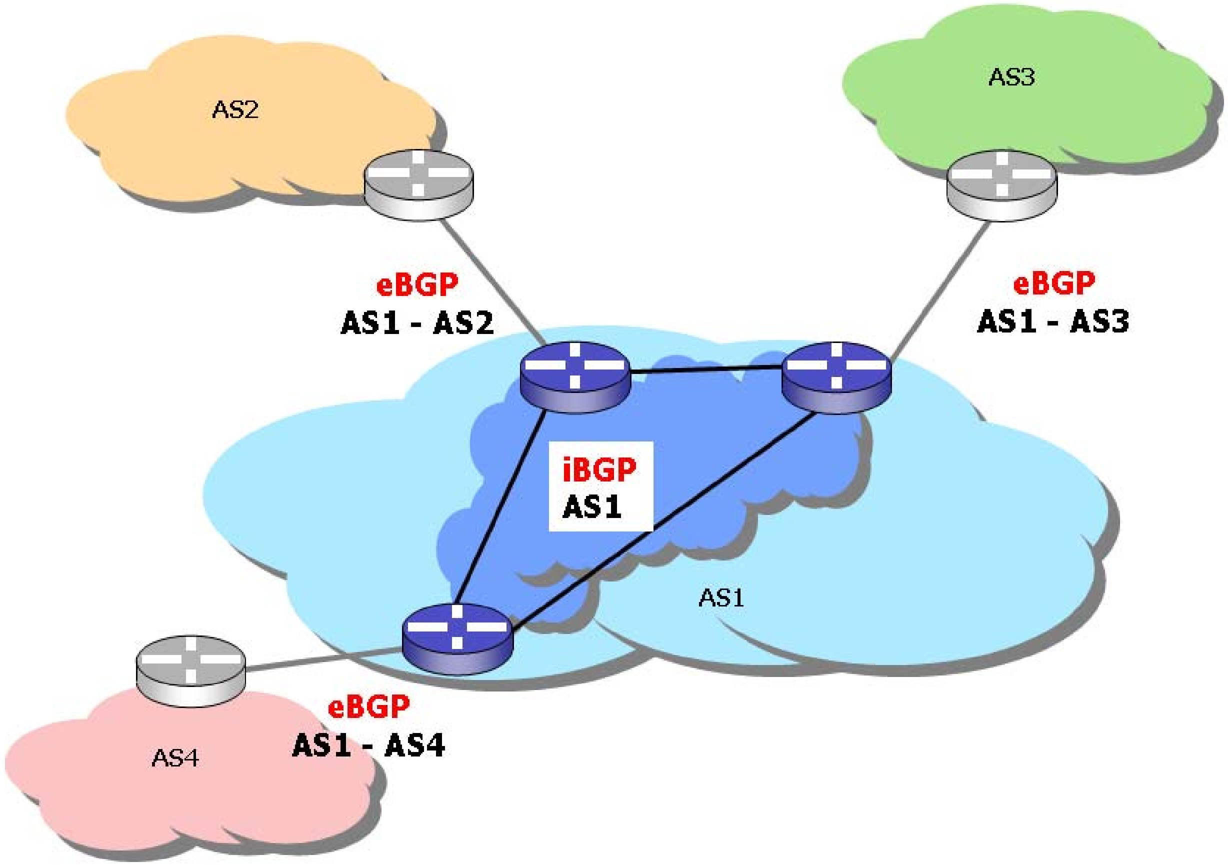
# BGP is a Distance Vector Protocol

- Maintains a collection of local “best paths” for all advertised prefixes
- Passes incremental changes to all neighbours rather than periodic full dumps
- A BGP update message reflects changes in the local database:
  - A new reachability path to a prefix that has been installed locally as the local best path (update)
  - All local reachability information has been lost for this prefix (withdrawal)

# iBGP and eBGP

- eBGP is used across AS boundaries
- iBGP is used within an AS to synchronise the decisions of all eBGP speakers
  - iBGP is auto configured (vie a match of MyAS in the OPEN message)
  - iBGP peering is manually configured
  - iBGP needs to emulate the actions of a full mesh
  - Typically configured as a flooding hierarchy using Route Reflectors
  - iBGP does not loop detect
  - iBGP does not AS prepend

# iBGP and eBGP



# BGP Transport

- TCP is the BGP transport
  - Port 179
  - Reliable transmission of BGP Messages
    - Messages are never repeated!
  - Capability to perform throttling of the transmission data rate through TCP window setting control
- May operate across point-to-point physical connections or across entire IP networks

# Messaging protocol

- BGP is not a data stream protocol
- The TCP stream is divided into messages using BGP-defined “markers”
- Each message is a standalone protocol element
- Each message has a maximum size of 4096 octets

# BGP Messages

```
UPDATE: 2007/07/15 01:46
    ATTRS: nexthop 202.12.29.79,
           origin i,
           aggregated by 64642 10.19.29.192,
           path 4608 1221 4637 3491 3561 2914 3130
    U_PFX: 198.180.153.0/24

UPDATE: 2007/07/15 01:46
    W_PFX: 64.31.0.0/19,
           64.79.64.0/19
           64.79.86.0/24

UPDATE: 2007/07/15 01:46
    ATTRS: nexthop 202.12.29.79,
           origin i,
           aggregated by 65174 10.17.204.65,
           path 4608 1221 4637 16150 3549 1239 12779 12654
    U_PFX: 84.205.74.0/24

UPDATE: 2007/07/15 01:47
    ATTRS: nexthop 202.12.29.79,
           origin i,
           aggregated by 64592 10.17.204.65,
           path 4608 1221 4637 4635 34763 16034 12654
    U_PFX: 84.205.65.0/24
```

# BGP Message Format – Marker

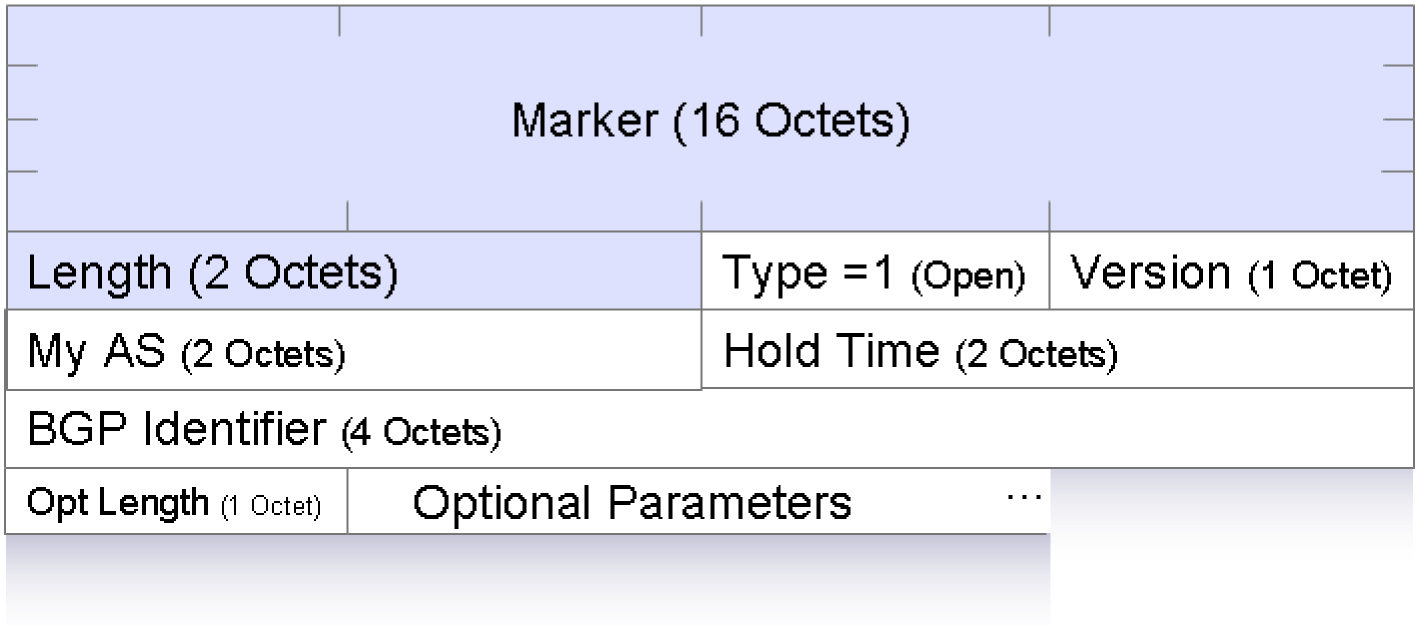
| | |
|---|---|
| Marker (16 Octets) | |
| Length (2 Octets) | Type (1 Octet) |

1 – OPEN
2 – UPDATE
3 – NOTIFICATION
4 – KEEPALIVE
5 – ROUTE-REFRESH

# Mark

- Mark is a record delimiter
  - Value all 1's (or a security encoded field)
- Length is message size in octets
  - Value from 19 to 4096
- Type is the BGP message type

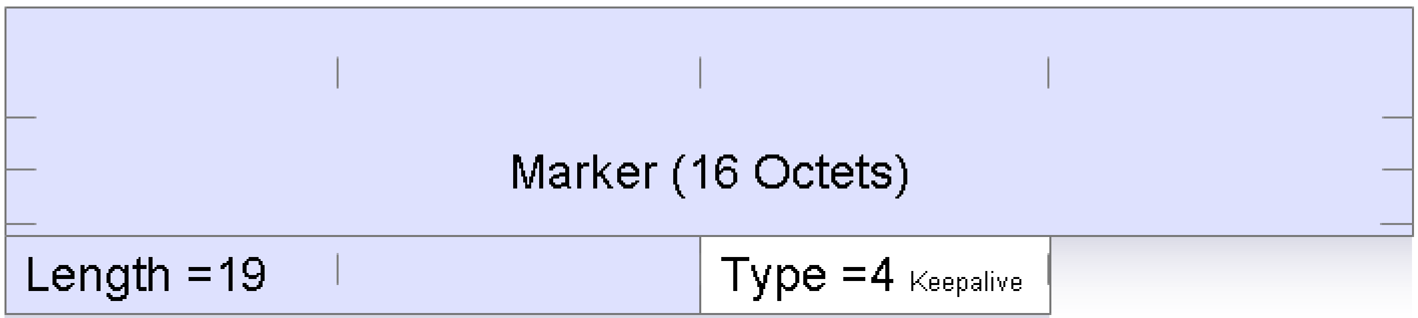
# BGP OPEN Message



# Open

- Session setup requires mutual exchange of OPEN messages
- Version is 4
- MyAS field is the local AS number
- Hold time is inactivity timer
- BGP identifier code is a local identification value (loopback IPv4 address)
- Options allow extended capability negotiation
  - E.g. Route Refresh, 4-Byte AS, Multi-Protocol

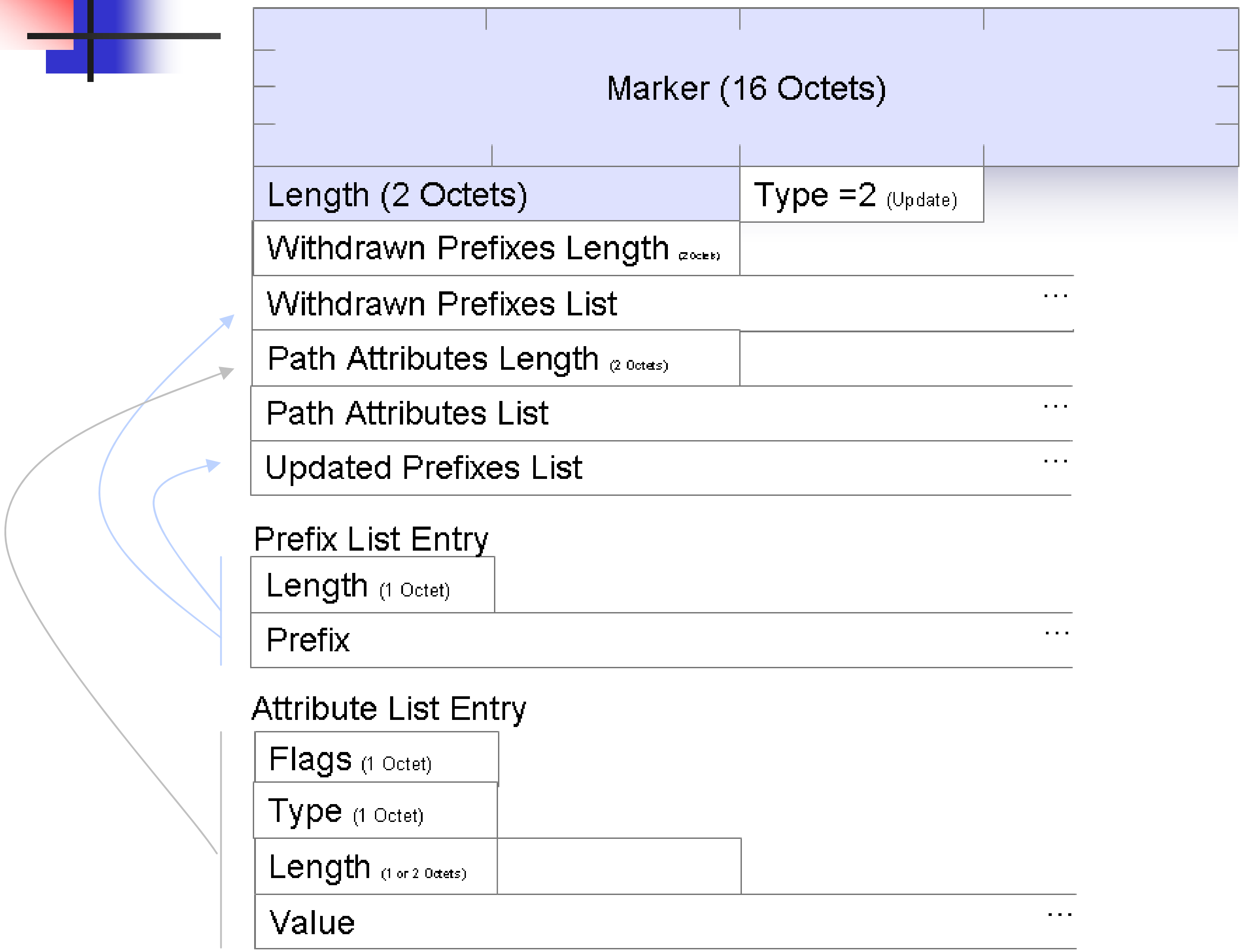
# BGP KEEPALIVE Message



# Keepalive

- “null” message
- Sent at 1/3 hold timer interval
- Prevent the remote end triggering an inactivity session reset

# BGP UPDATE Message



# UPDATE

- Used for announcements, updates and withdrawals
- Can piggyback withdrawals onto announcements
- List of withdrawn prefixes
- List of updated prefixes
- Set of “Path Attributes” common to the updated prefix list

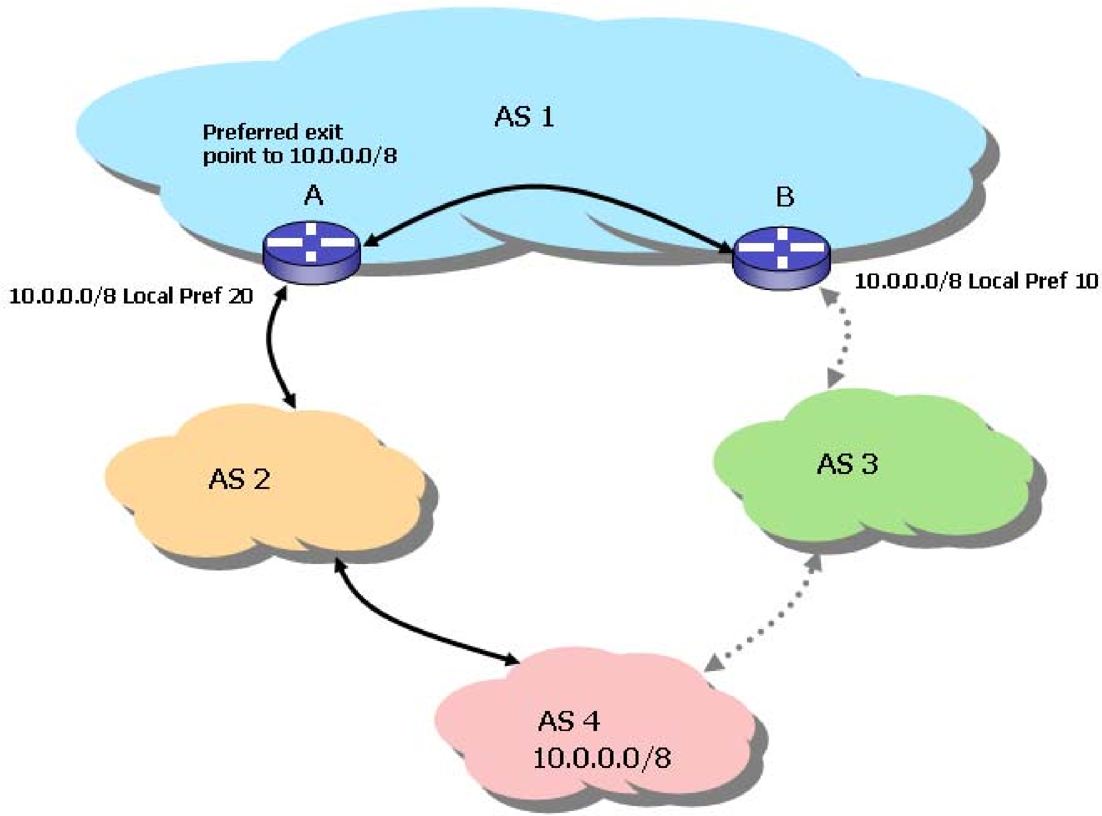
# Update Path Attributes

- Additional information that is associated with an address
- Attributes can be:
  - Optional or Well-Known
  - Transitive or Point-to-point
  - Partial or Complete
  - Extended Length or not

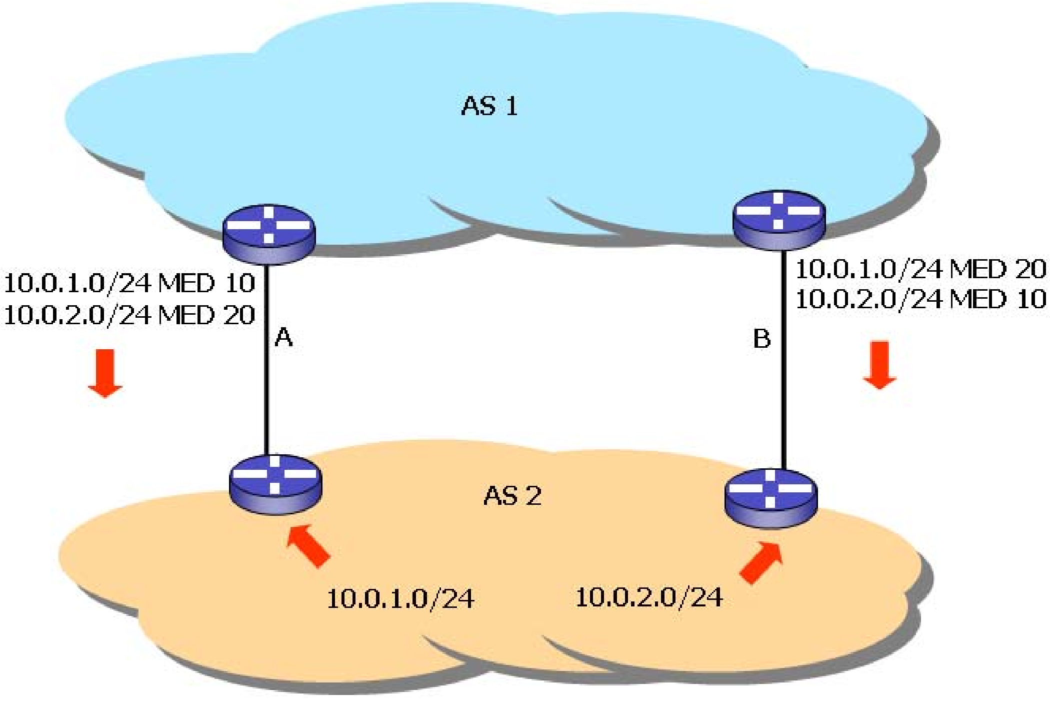
# Update Path Attributes

- **Origin** : how this route was injected into BGP in the first place
- **Next_Hop** : exit border router
- **Multi-Exit-Discriminator** : relative preference between 2 or more sessions between the same AS pair
- **Local Pref** : local preference setting
- **Atomic Aggregate** : Local selection of aggregate in preference to more specific
- **Aggregator** : identification of proxy aggregator
- **Community** : locally defined information fields
- **Destination Pref** : preference setting for remote AS

# Local Pref Example



# MED Example



AS 1
10.0.1.0/24 MED 10
10.0.2.0/24 MED 20
A
B
10.0.1.0/24 MED 20
10.0.2.0/24 MED 10
AS 2
10.0.1.0/24
10.0.2.0/24

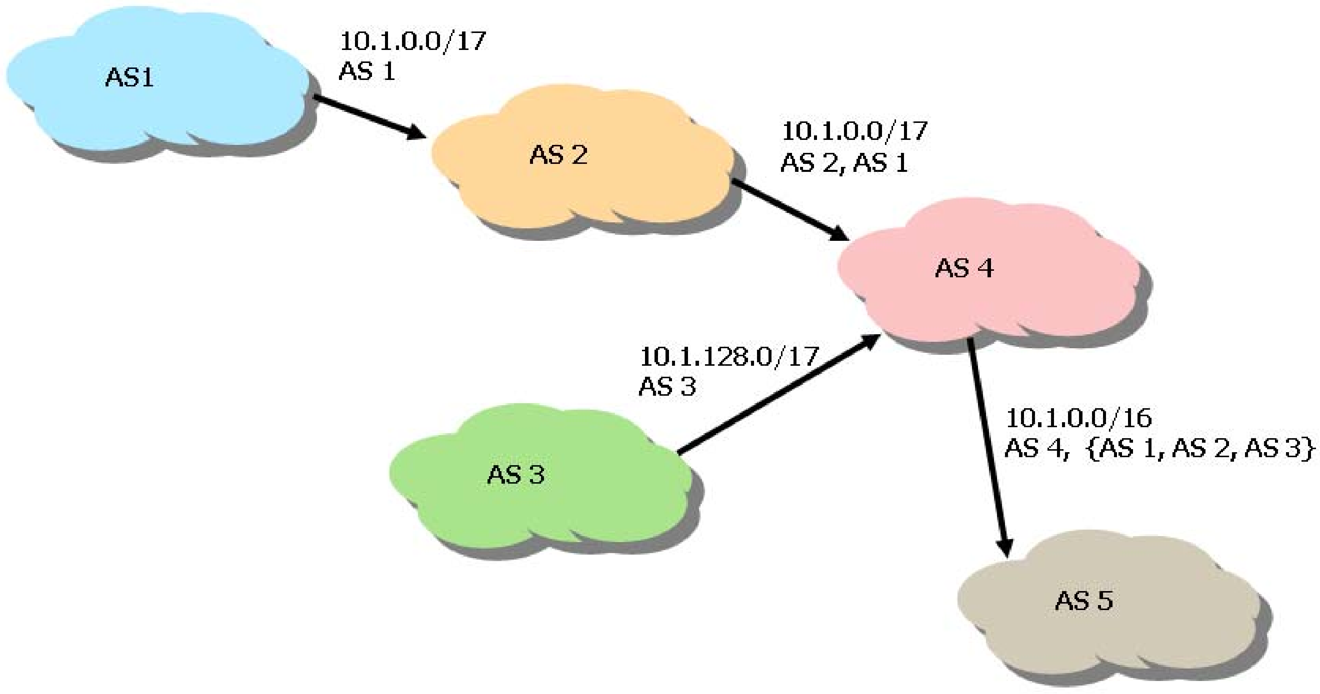
# AS Path

- **AS_PATH** : the vector of AS transits forming a path to the origin AS
  - In theory the BGP Update message has transited the reverse of this AS path
  - In practice it doesn’t matter
    - The AS Path is just a loop detector and a path metric

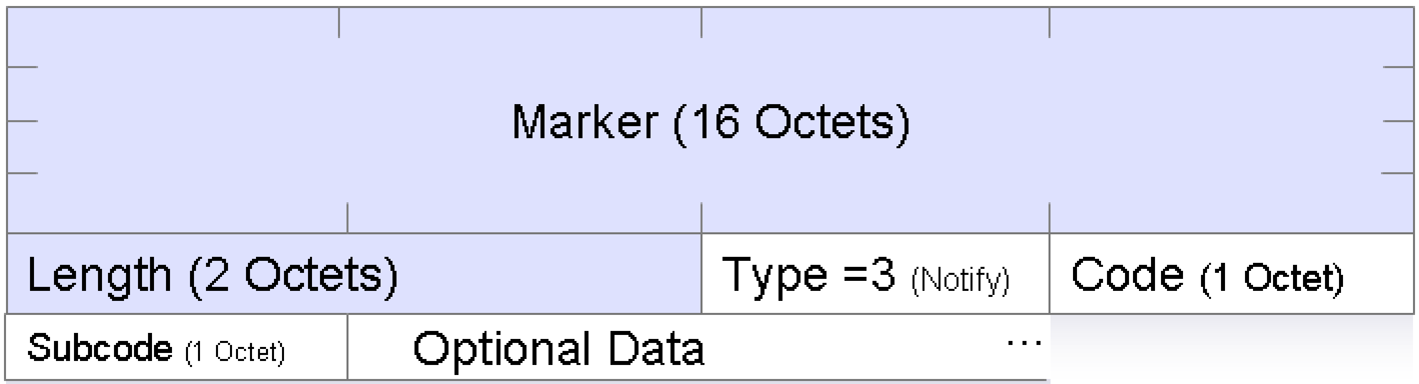
# AS Path

- AS Path is a vector of AS values, optionally followed by an AS Set
- AS Set : If a BGP speaker aggregates a set of BGP route objects  into a single object, the set of AS's in the component updates are placed into an unordered AS_Set as the final AS Path element

# AS Path Example



# BGP NOTIFICATION Message



# BGP ROUTE REFRESH Message

| | |
|---|---|
| Marker (16 Octets) | |
| Length =19 | Type =4 Refresh |

# Route Selection Algorithm

- For a set of received advertisements of the same address prefix then the local "best" selection is based on:
  - Highest value for Local-Pref
    - Local setting
  - Shortest AS Path length
    - External preference
  - Lowest Multi_Exit_Discriminator value
    - Egress tie break for multi-connected ASes
  - Minimum IGP cost to Next_Hop address
    - iBGP tie break
  - eBGP learned routes preferred to iBGP-learned routes
  - Lowest BGP Identifier value
    - Last point tie break

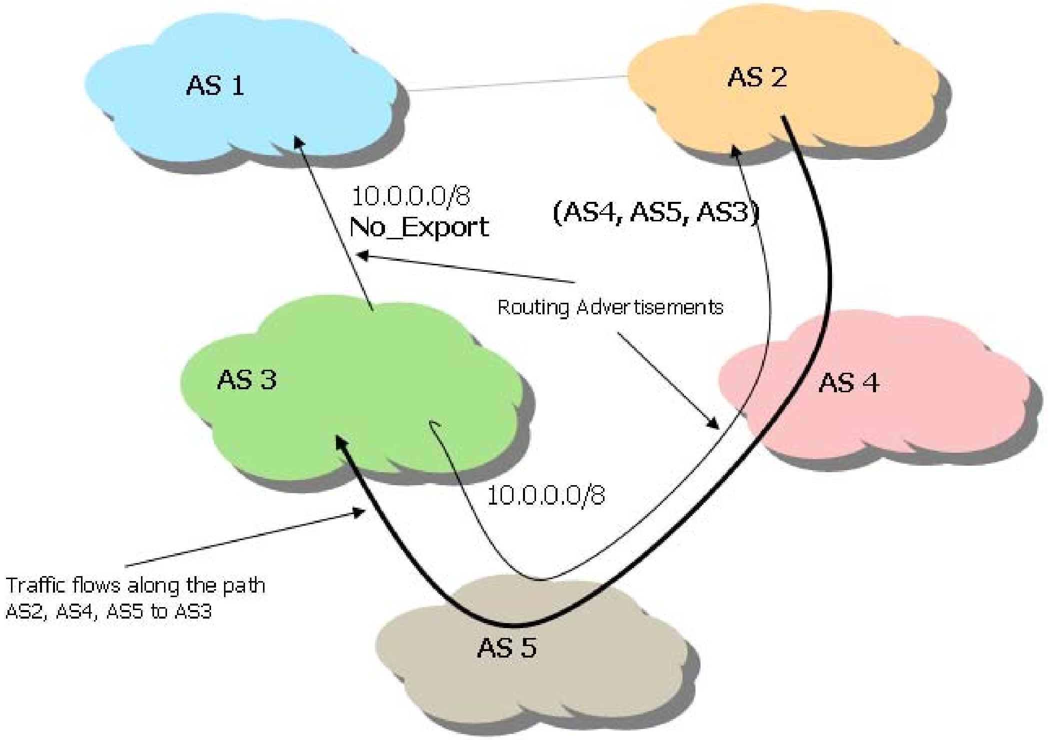
# Communities

- Communities are an optional transitive path attribute of an Update message, with variable length
  - Well-Known Communities
  - AS-Defined communities
- A way of attaching additional information to a routing update

# Well-Known Communities

- Registered in an IANA Registry
- Created by IETF Standards Action
  - NO_EXPORT
    - Do not export this route outside of this AS, or outside of this BGP Confederation
  - NO_ADVERTISE
    - Do not export this route to any BGP peer (iBGP or eBGP)
  - NO_EXPORT_SUBCONFED
    - Do not export this route to any eBGP peer
  - NOPEER
    - No do export this route to eBGP peers that are bilateral peers

# Community Example: NO_EXPORT



# AS-Defined Communities

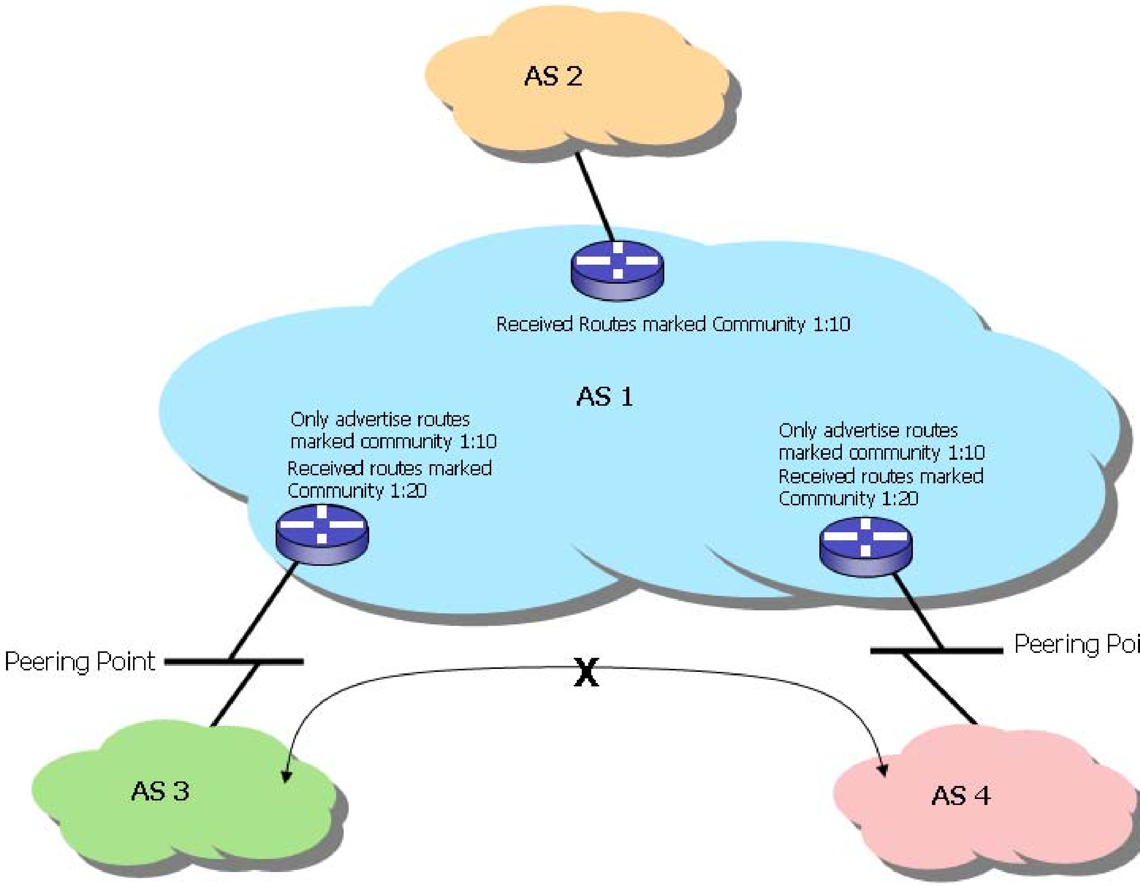
- Optional Transitive Attribute
  - AS value
  - AS-specific value
- Used to signal to a specific AS information relating to the prefix and its handling
  - Local pref treatment
  - Prepending treatment
- Use to signal to other ASs information about the local handling of the prefix within this AS

# Extended Communities

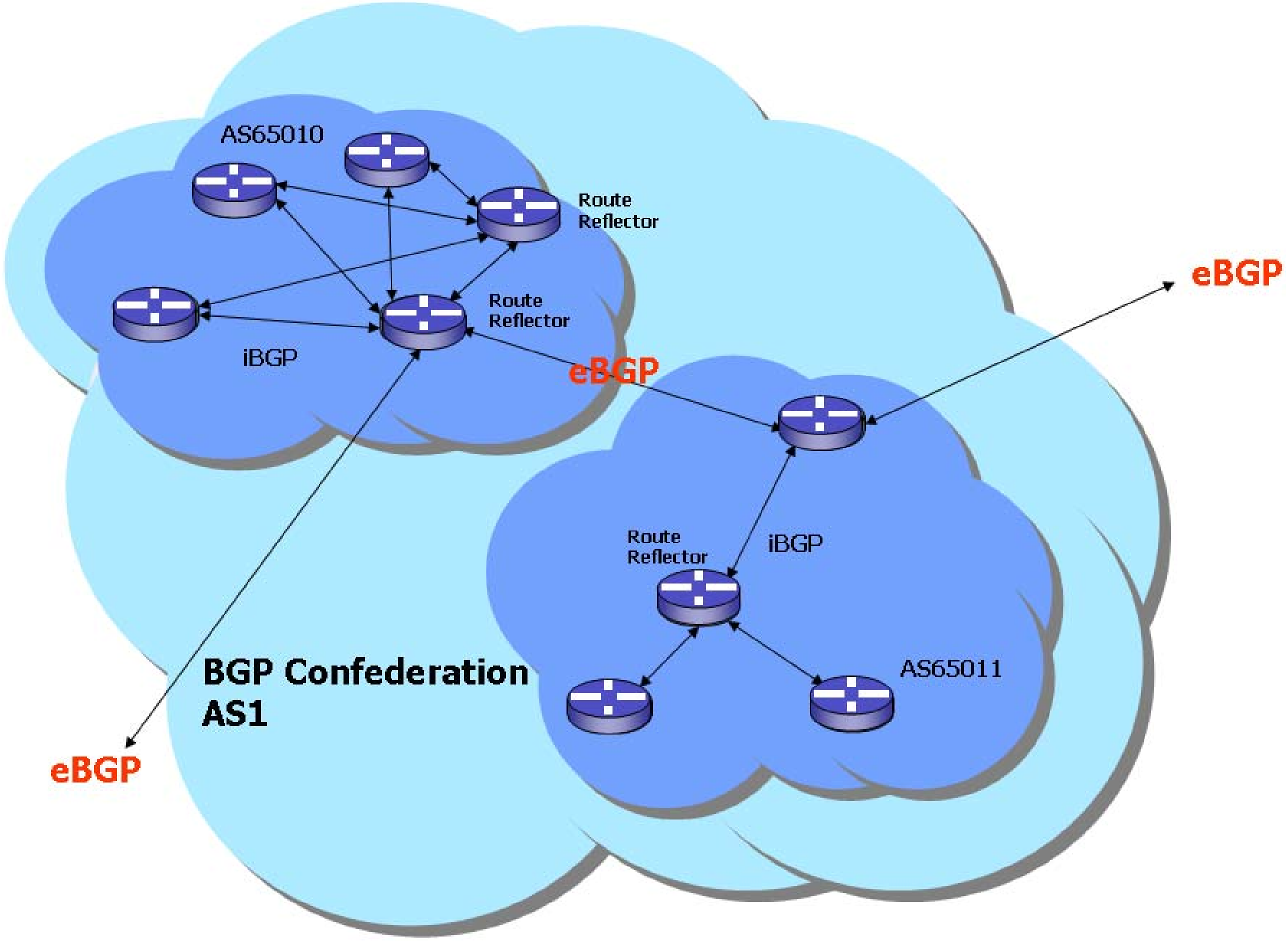
- Negotiated capability
- Adds a Type field to the community
- 8 octet field
  - 2 octets for type
    - 1 bit for IANA registry
    - 1 bit for transitive
  - 6 octets for value
    - 2 octets for AS
    - 4 octets for value

    or
    - 4 octets for AS
    - 2 octets for value

# Community Example: Policy Signalling in iBGP



# Route Reflectors and Confederations



# BGP and IPv6

- IPv6 support in BGP is part of a generalized multi-protocol support in BGP
- Capability negotiated at session start
- New non-transitive optional attributes
  - MP_REACH_NLRI
    - Carries reachable destinations and associated next hop information, plus AFI/Sub-AFI
    - V6 -> AFI = 2, SAFI = 1 (unicast)
  - MP_UNREACH_NRLI
    - Unreachable destinations, AFI/Sub-AFI
- Like tunnelling, the MP-BGP approach places IPv6 BGP update information inside the MP attributes of the outer BGP update message

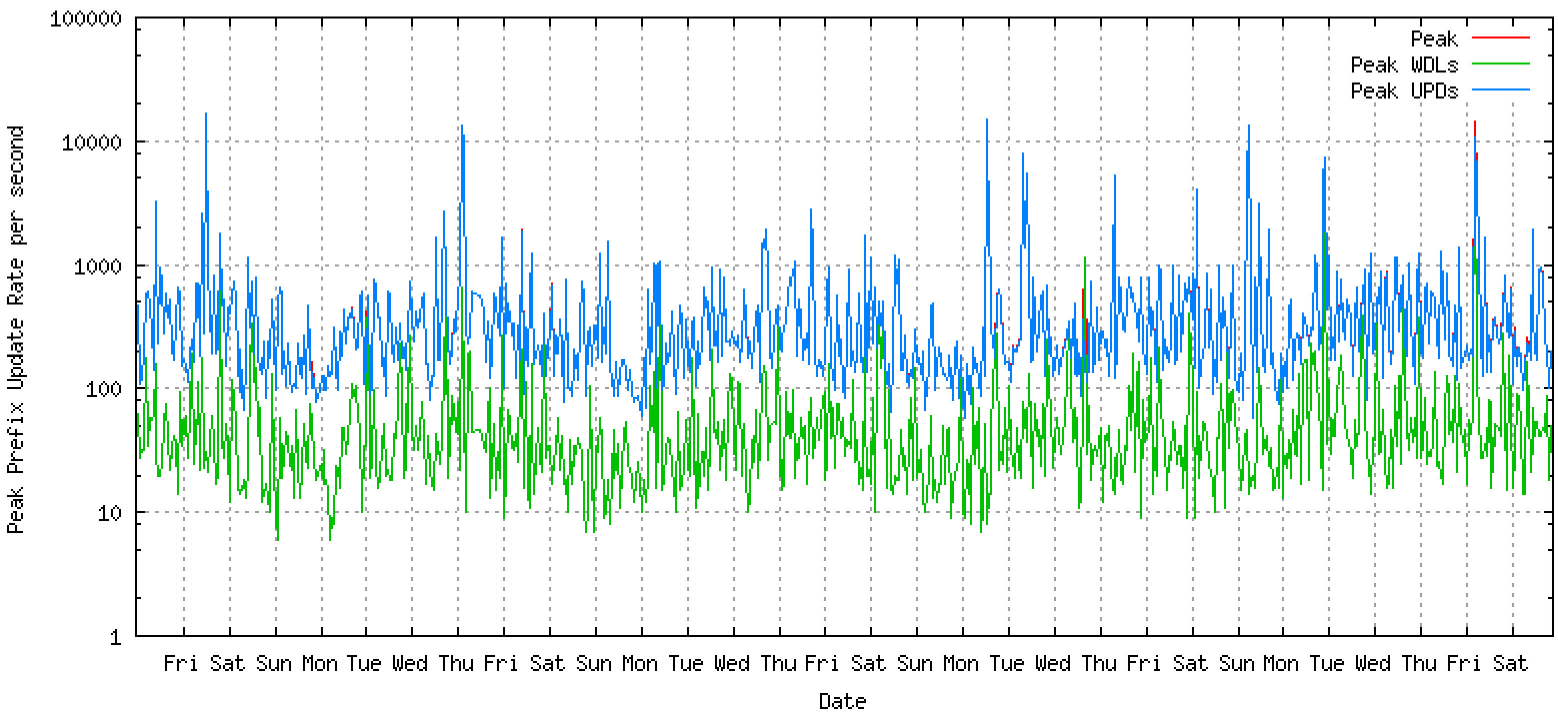
# BGP Session Security

- The third party TCP reset problem
  - TTL Hack
  - TCP hack
  - MD5 Signature Option
  - IPSEC for BGP

# BGP Update Loads

- BGP does not implicitly suppress information
  - Anything passed into BGP is passed to all BGP speakers
  - Local announcements and withdrawals into eBGP are propagated to all BGP speakers in the entire network
- BGP can be a "chatty" protocol
  - Particularly in response to a withdrawal at origin
- The instanteous peak "update loads" in BGP can be a significant factor in terms of processor capability for BGP speakers  and overall convergence times

# Peak Update loads – IPv4 Network




Hourly peak per second BGP update loads – measured at AS2.0 in July 2007

# Load Shedding - RFD

- Route Flap Damping
  - “Two flaps are you are out!”
  - For each prefix / eBGP peer pair have a “penalty” score
  - Each Update and Withdrawal adds to the penalty
  - The penalty score decays over time
  - If the penalty exceeds the suppression threshold then the route is damped
  - The route is damped until the panelty score decays to the re-advertisement threshold
  - Fallen into disfavour these days
    - Single withdrawal at origin can trigger multi-hour outages

# Load Shedding – MRAI and WMRAI

- Applied to the ADJ-RIB-OUT queue
- Wait for the MRAI timer interval (30 seconds) before advertising successive updates for the same prefix to the same peer
- Coarser: only advertise updates to a peer at 30 second intervals
- Coarser: Only advertise updates at 30 second intervals
- WMRAI : Include Withdrawal in the same timer

- A very coarse granularity filter
- Some implementations have MRAI enabled by default, others do not
- The mixed deployment has been simulated to be worse than noone or everyone using MRAI!

# Load Shedding – SSLD

- Relative simple hack to BGP
- Use the sender side to perform loop detection looking for the eBGP peer's AS in the AS Path, suppress sending the update is found

# Influencing Route Selection

- Local selection (outbound path selection) can be adjusted through setting the Local_Pref values applied to incoming routing objects
- But what about inbound path selection?
  - How can a AS “bias” the route selection of other ASs?
    - BGP Communities
    - Advertise more specific prefixes along the preferred path
    - Use own-AS prepending to advertise longer AS paths on less preferred paths
    - Use poison-AS set prepending to selectively eliminate path visibility

# Operational Practices

- Maximize business outcomes
  - Prefer customers over peers over upstream
  - Maximize choices and avoid upstream lock-in
- Follow the topology
  - Prefer to minimize delay and maximize bandwidth
  - Maximize network utilization efficiency
  - Leverage topology diversity
- Reduce complexity
- Reduce risk

# Agenda

- Scope
- Background to Internet Routing
- BGP
- Current IETF Activities
- Views, Opinions and Comments

# Current (and Recent) IETF Activities

- Working Groups that directly relate to BGP work in the IETF:
  - Inter-Domain Routing (IDR)
  - Routing Protocol Security Requirements (RPSEC)
  - Secure Inter-Domain Routing (SIDR)
  - Global Routing Operations (GROW)

# 4-Byte AS Numbers

- RFC4893
  - Extends the Autonomous System identifier from 16 bits to 32 bits
    - Due to run-out concerns of the 16 bit number space first identified in 1999
  - An excellent example of a clearly through out backward-compatible transition arrangement
  - IDR activity undertaken from 2000 - 2007

# Current IDR topics

- Outbound Route Filter
  - Extension BGP signalling that requests the peer to apply a specified filter set to the updates prior to passing them to this BGP speaker
- AS Path Limit
  - A new BGP Path Attribute that functions as a form of TTL for BGP Route Updates

# RPSEC Topics

- BGP Security Requirements
  - What are the security requirements for BGP?
  - This work is largely complete – the major outstanding topic at present is the extent to which the AS Path attribute of BGP updates could or should be secured

# SIDR

- Currently Working on basic tools for passing security credentials
  - Digital signatures with associated X.509 certification and a PKI for signature validation
- Then will work on approaches to fitting this into BGP in a modular fashion
  - Based on the RPSEC requirements this is a study of what and how various components of the BGP information could be digitally signed and validated

# GROW

- Operational perspectives on BGP deployment
  - Recent activity:
    - MED Considerations
    - CIDR revisited
    - BGP Wedgies
- Currently setting a new work agenda

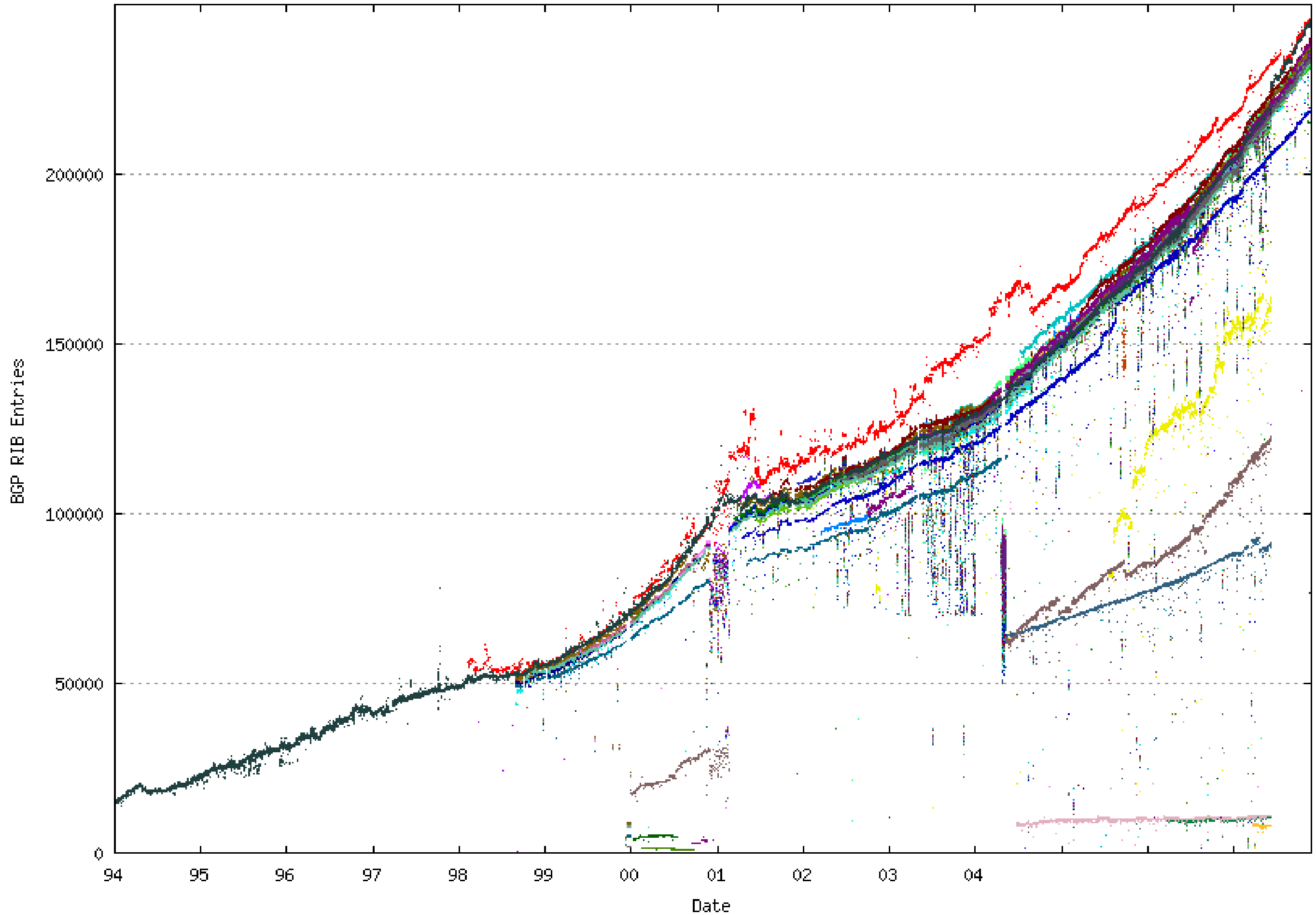
# Agenda

- Scope
- Background to Internet Routing
- BGP
- Current IETF Activities
- Views, Opinions and Comments

# Some open questions

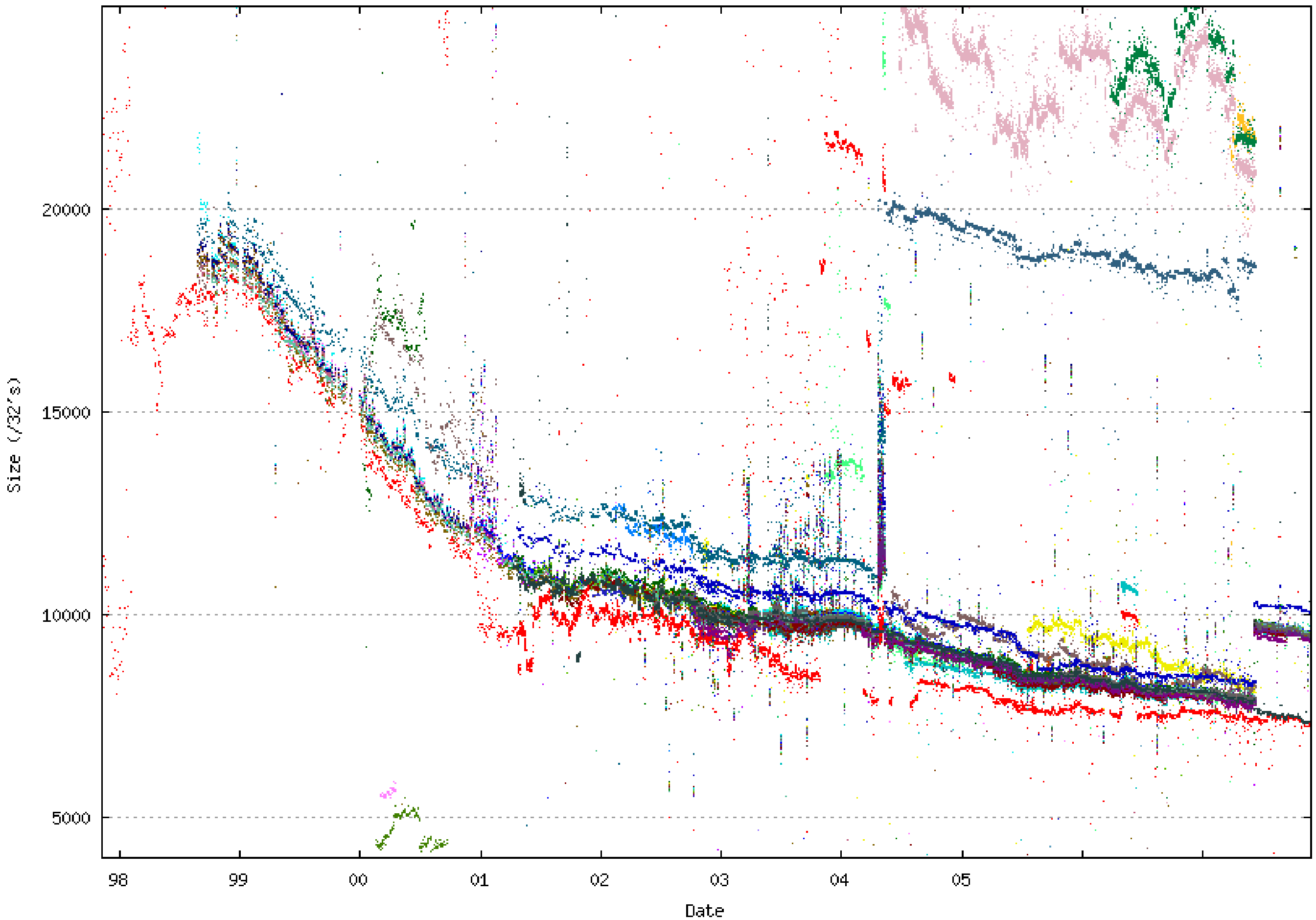
- How big does the routing world get?
- How important are routing behaviours to mobility, ad hoc networking, sensor nets, ... ?
- While IP addresses continue to use overloaded semantics of forwarding and identity then there is continual pressure for persistent identity properties of addresses
  - Which places pressure on the routing system

# Scaling – How big can it get?



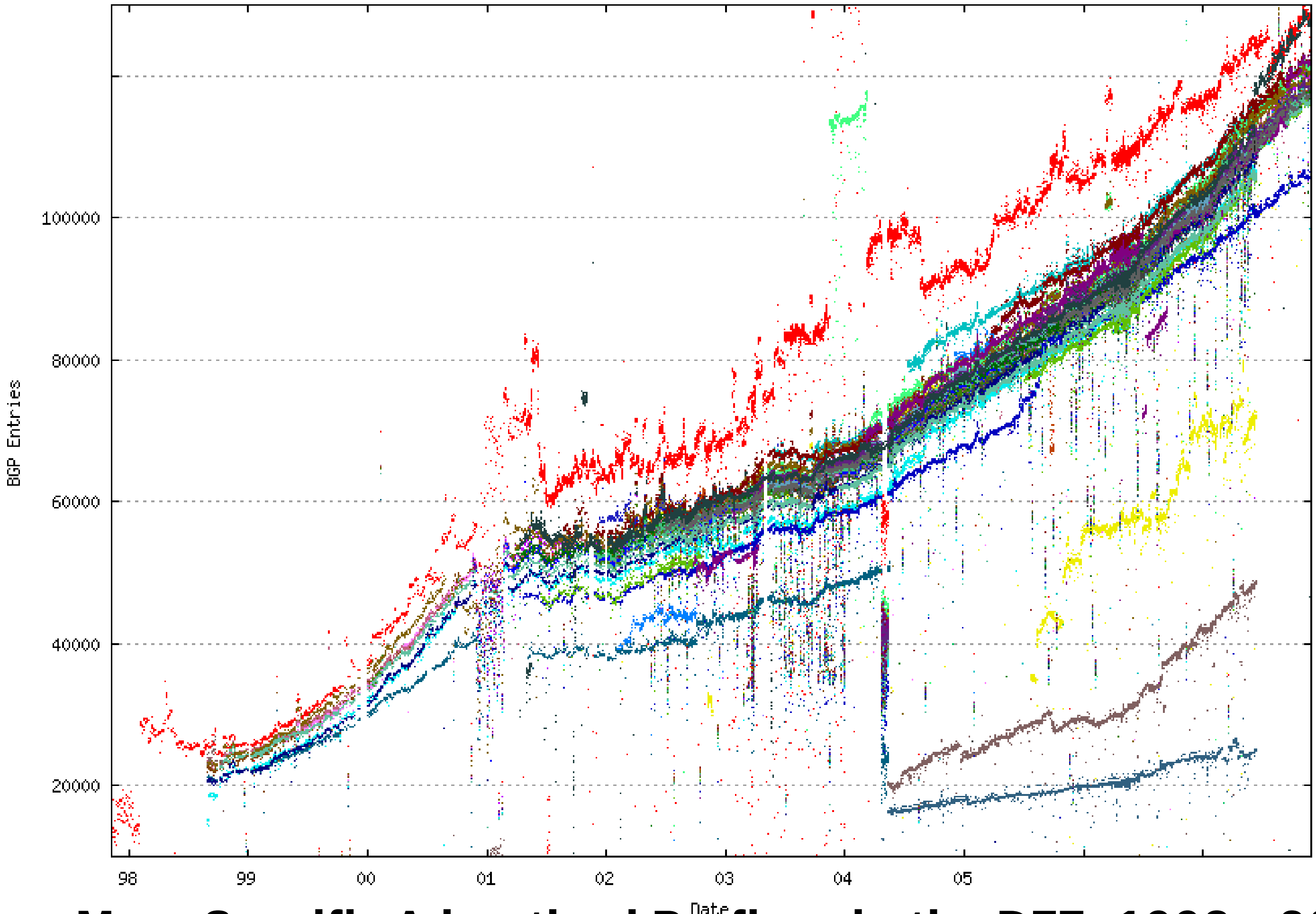

**Size of the DFZ IPv4 Routing Table: 1994 - 2007**

# Scaling – Micro and Macro




**Average Advertised Prefix Size DFZ: 1998 - 2007**

# Scaling –More Specifics




**More Specific Advertised Prefixes in the DFZ: 1998 - 2007**

# Research Perspectives

- How well does BGP scale?
  - Various views ranging from perspectives of short term scaling issues through to no need for immediate concern
    - Is it the number of route objects or their dynamic behaviour that's the pressing problem here?
  - Recent interest in
    - examining BGP to improve some aspects of its dynamic behaviour
    - looking at alternative approaches to addressing and routing, generally based on forms of tunnelling and landmark routing to reduce the route object population

# Looking Forward

- A number of studies over the years to enumerate the requirements and desired properties of an evolved routing system in the Routing Research Group
- It is unclear whether there is an immediate need to move the entire Internet to a different inter-domain routing protocol
- The decoupled routing architecture of the network does not prevent different routing protocols and different approaches to routing being deployed in distinct routing realms within the Internet

# Questions and Comments?