

Copyright © William C. Cheng

AS Categories

- ↳ **Stub:** an AS that has only a single connection to one other AS - carries only local traffic
- ↳ **Multi-homed:** an AS that has connections to more than one AS, but does not carry transit traffic
- ↳ **Transit:** an AS that has connections to more than one AS, and carries both transit and local traffic (under certain policy restrictions)

Computer Communications - CSC1 551

Copyright © William C. Cheng

Autonomous Systems

- ↳ **What is an AS?**
 - = a set of routers under a single technical administration
 - = uses an *interior gateway protocol (IGP)* and common metrics to route packets within the AS
 - = uses an *exterior gateway protocol (EGP)* to route packets to other AS's
- ↳ AS may use multiple IGPs and metrics, but appears as single AS to other AS's
- ↳ Why have both EGP and IGP?
 - = know different levels of detail
 - = different levels of trust
 - = policy issues are much more important in EGP

Computer Communications - CSC1 551

Copyright © William C. Cheng

Where And Why BGP?

- ↳ **Where?**
 - = multi-homed hosts
 - = E-BGP for inter-domain routing (between AS's)
 - = I-BGP for intra-domain routing (within an AS)
- ↳ **Why?**
 - = to deal with dynamics (link failure/recovery)
 - = configurable policies on routes

Computer Communications - CSC1 551

Copyright © William C. Cheng

Example

Computer Communications - CSC1 551

Copyright © William C. Cheng

BGP Terminology

- ↳ **AS:** autonomous system
- ↳ **Peer:** an adjacent router (Note: not the same meaning as ISP peering)
- ↳ **Exchange point:** place where many ISPs have routers and connections
- ↳ **RIB:** routing information base
- ↳ **Adj-RIB-In:** incoming routing information
- ↳ **Loc-RIB:** local routing information
- ↳ **Adj-RIB-Out:** outgoing routing information

Computer Communications - CSC1 551


Copyright © William C. Cheng

Complicate: Multihoming

- = links to multiple ISPs
- = static routes do not work well

Computer Communications - CSC1 551

Copyright © William C. Cheng




Solution: DV with Path Vectors

- Each routing update carries the entire path (AS's appear in the path)
- Loops are detected as follows:
 - when AS gets route check if AS already in path
 - if yes, reject route
 - if no, add itself and (possibly) advertise route further
- Advantage:
 - metrics are local - AS chooses path, protocol ensures no loops

Computer Communications - CSC1 551

Copyright © William C. Cheng




Hop-by-hop Model

- BGP advertises to neighbors only those routes that it uses
 - consistent with the hop-by-hop Internet paradigm
 - e.g., AS1 cannot tell AS2 to route to other AS's in a manner different than what AS2 has chosen (need source routing for that)

Computer Communications - CSC1 551

Copyright © William C. Cheng




BGP Messages

- **OPEN**: sets up timeout, AS, id, etc.
- **UPDATE**: update (inject, withdraw) routes with attributes
- **NOTIFICATION**: error reporting
- **KEEPALIVE**: no change, but link is up
 - TCP has keepalive option, why BGP keepalive also?
 - end-to-end argument: TCP connection can be alive but routing mechanism can be hung, must have BGP keepalive

Computer Communications - CSC1 551

Copyright © William C. Cheng




EGP Protocol Choices

- Link state or distance vector?
 - no universal metric - policy decisions
 - Problems with distance-vector:
 - Bellman-Ford algorithm slow to converge (counting to infinity problem)
 - Problems with link state:
 - metric used by routers in different AS's is not the same - may create loops
 - link state database too large - entire Internet
 - may expose policies to other AS's

Computer Communications - CSC1 551

Copyright © William C. Cheng




Interconnecting BGP Peers

- BGP uses TCP to connect peers (port 179)
- Advantages:
 - simplicity from reliability and ordering
 - I-BGP communicate over multiple hops
 - no need for periodic refresh - routes are valid until withdrawn, or the connection is lost (hard state or soft state?)
 - incremental updates
- Disadvantages
 - congestion control on a routing protocol?
- TCP has keepalive option, why BGP keepalive also?
 - **end-to-end argument**: TCP connection can be alive but routing mechanism can be hung, must have BGP keepalive

Computer Communications - CSC1 551

Copyright © William C. Cheng



Protocol Observations

- How does BGP know when a link is down/out?
 - timeout (hold time)
 - see [Shaihkh00a]
- How does BGP avoid looping paths?
 - path vector via AS_PATHS
 - loop detection on route receipt (or transmission [Labovitz00a])

Computer Communications - CSC1 551

Copyright © William C. Cheng

20

Policy With BGP

- BGP provides capability for enforcing various policies
- Policies are not part of BGP (no policy messages)
- they are provided to BGP as configuration information
- BGP enforces policies by *choosing paths from multiple alternatives and controlling advertisement to other AS's*

Computer Communications - CSC1 551

Copyright © William C. Cheng

22

BGP Is NOT Needed If:

- Single homed network (stub)
- AS does not provide downstream routing
- AS uses a default route

AS100
Upstream Provider
204.10.0.23
Static Route
Default Route

Computer Communications - CSC1 551

Copyright © William C. Cheng

24

Routing Information Bases (RIB)

- Routes are stored in RIBs
- Adj-RIBs-in: routing info that has been learned from other routers (unprocessed routing info)
- Loc-RIB: local routing information selected from Adj-RIBs-in (routes selected locally)
- Adj-RIBs-Out: info to be advertised to peers (routes to be advertised)

Computer Communications - CSC1 551

Copyright © William C. Cheng

19

BGP Attributes

- ORIGIN: where prefix originates
- AS_PATH: path for routing
- NEXT_HOP: where to send data
- MULTI_EXIT_DISCRIMINATOR: used to influence multi-homing
- Why BGP Attributes?
 - want to do policy routing
 - want some way to prevent looping in DV routing (AS_PATH)
 - flexibilities (allows extensibility)

Computer Communications - CSC1 551

Copyright © William C. Cheng

21

Examples of BGP Policies

- A multi-homed AS refuses to act as transit
- limit path advertisement
- A multi-homed AS can become transit for some AS's
- only advertise paths to some AS's
- An AS can favor or disfavor certain AS's for traffic transit from itself

Computer Communications - CSC1 551

Copyright © William C. Cheng

23

BGP-4

- Latest version of BGP
- BGP-4 supports CIDR

Computer Communications - CSC1 551

Copyright © William C. Cheng

25

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

Length (2 bytes) | Type (1 byte)

16 bytes
Marker (security and message delineation)

0 1 2 3

BGP Common Header

Computer Communications - CSC 551

Copyright © William C. Cheng

26

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

My autonomous system: ID assigned to that AS

Hold timer: max interval between KEEPALIVE or UPDATE messages

BGP ID: address of one (typically virtual) interface and is same for all messages

Optional parameters <type, length, value>

Parameter length

BGP Identifier

My autonomous system | Hold time

Length | Type: open | version

Marker (security and message delineation)

0 1 2 3

BGP OPEN Message

Computer Communications - CSC 551

Copyright © William C. Cheng

29

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

Used for error notification (update error, expired timer, FSM, cease)

TCP connection is closed immediately after notification.

Date

Error sub-code

Length | Type: notification | Error code

Marker (security and message delineation)

0 1 2 3

BGP UPDATE Message

Computer Communications - CSC 551

Copyright © William C. Cheng

27

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

Many prefixes may be included in UPDATE, but must share same attributes.

UPDATE message may report multiple withdrawn routes.

Path attribute len | Path attributes (variable) - origin, path, metrics, etc.

... | -routes len | Withdrawn routes (variable)

Length | Type: update | Withdrawn..

Marker (security and message delineation)

0 1 2 3

BGP UPDATE Message

Computer Communications - CSC 551

Copyright © William C. Cheng

28

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

Flags: optional v.s. well-known

transitive v.s. non-transitive (passed on)

partial (someone in path did not understand this attribute)

extended length (2 bytes instead of 1)

Attribute types: *Origin, AS_PATH, Next_Hop (more later..)*

Attribute type field

Attribute flags (1 byte) | Attribute type code (1 byte)

Type-Length-Value encoding

Attribute type (2 bytes) | Attribute length (1-2 bytes) | Attribute Value (variable)

0 1 2 3

Path Attributes

Computer Communications - CSC 551

Copyright © William C. Cheng

30

Types: OPEN, UPDATE, NOTIFICATION, KEEPALIVE

Used for error notification (update error, expired timer, FSM, cease)

TCP connection is closed immediately after notification.

Length | Type: keepalive

Marker (security and message delineation)

0 1 2 3

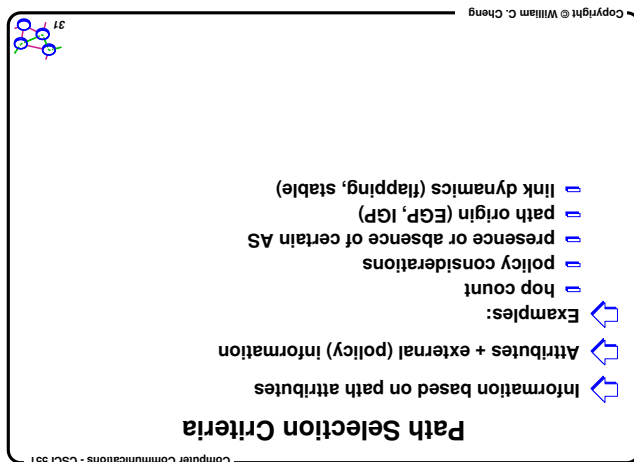
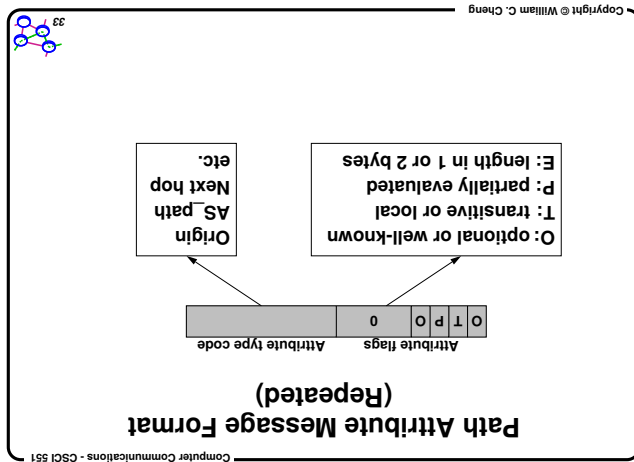
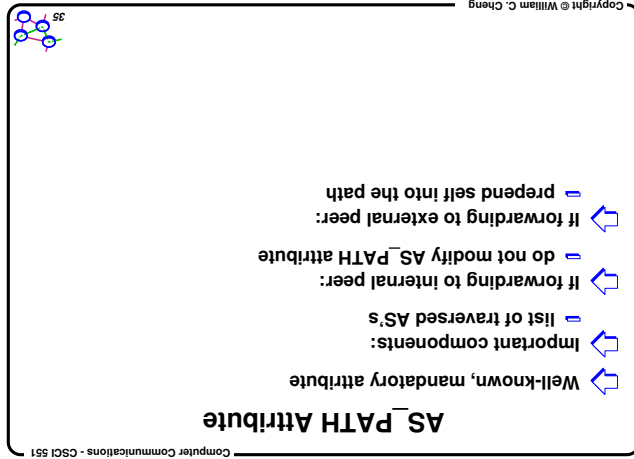
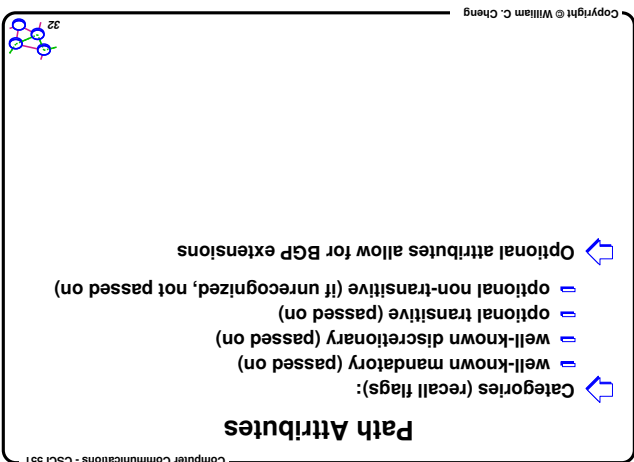
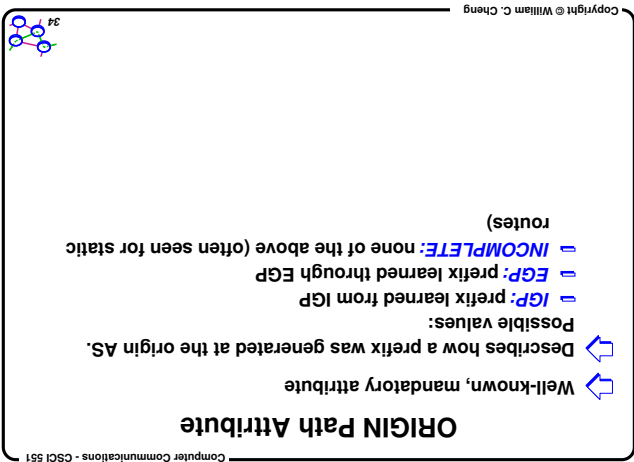
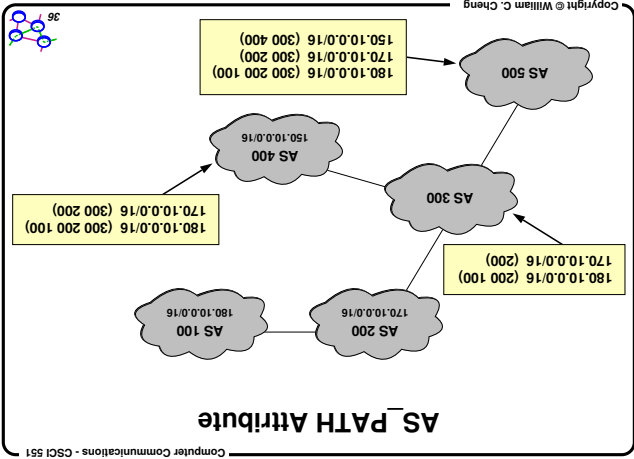
BGP KEEPALIVE Message

Computer Communications - CSC 551

Note: hold_time = zero means no keepalives will be sent

Sent periodically (but before hold timer expires) to peers to ensure connectivity.

Sent in place of an UPDATE message.



Copyright © William C. Cheng

CIDR and BGP

AS1 → 3: 10.1/16:a(1) 10.3/16:x [Internal]
 AS1's state: 10.1/16:b(1) 10.2/16:d(2)
 AS2 → 3: 10.2/16:c(2)
 AS2's state: 10.2/16
 AS3 (provider) 10/8
 AS3 → 4: 10.3/16:e(3) 10.2/16:e(3 2) 10.1/16:e(3 1) 10.2/16:e(3 2)
 AS3's state: 10.3/16:e(3) 10.2/16:e(3 2) 10.1/16:e(3 1)
 AS4 → 4: 10.8:e(3 [1 2])

↑ aggregation with set notation

42

Copyright © William C. Cheng

CIDR and BGP

AS1 → 3: 10.1/16:a(1) 10.3/16:x [Internal]
 AS1's state: 10.1/16:b(1) 10.2/16:d(2)
 AS2 → 3: 10.2/16:c(2)
 AS2's state: 10.2/16
 AS3 (provider) 10/8
 AS3 → 4: ?
 AS4 → 4: ?

Notation:
 = export: destination network:exit router(AS_PATH)
 = state: destination network:entry router(AS_PATH)

40

Copyright © William C. Cheng

CIDR and BGP

AS1 → 3: 10.1/16:a(1)
 AS1's state: 10.1/16:b(1)
 AS2 → 3: ?
 AS2's state: 10.2/16
 AS3 (provider) 10/8
 AS3 → 4: ?
 AS4 → 4: ?

Notation:
 = export: destination network:exit router(AS_PATH)
 = state: destination network:entry router(AS_PATH)

38

Copyright © William C. Cheng

CIDR and BGP

AS1 → 3: 10.1/16:a(1) 10.3/16:x [Internal]
 AS1's state: 10.1/16:b(1) 10.2/16:d(2)
 AS2 → 3: 10.2/16:c(2)
 AS2's state: 10.2/16
 AS3 (provider) 10/8
 AS3 → 4: 10.3/16:e(3) 10.2/16:e(3 2) 10.1/16:e(3 1)
 AS3's state: 10.3/16:e(3) 10.2/16:e(3 2) 10.1/16:e(3 1)

- too many entries in routing tables
- need aggregation for scalability
- also need enough information in AS_PATH for loop detection
- how?

41

Copyright © William C. Cheng

CIDR and BGP

AS1 → 3: 10.1/16:a(1) 10.3/16:x [Internal]
 AS1's state: 10.1/16:b(1) 10.2/16:d(2)
 AS2 → 3: 10.2/16:c(2)
 AS2's state: 10.2/16
 AS3 (provider) 10/8
 AS3 → 4: ?
 AS4 → 4: ?

Notation:
 = export: destination network:exit router(AS_PATH)
 = state: destination network:entry router(AS_PATH)

39

Copyright © William C. Cheng

CIDR and BGP

AS1 → 3: ?
 AS1's state: ?
 AS2 → 3: ?
 AS2's state: 10.2/16
 AS3 (provider) 10/8
 AS3 → 4: ?
 AS4 → 4: ?

37

Copyright © William C. Cheng

NEXT-HOP Path Attribute

- Well-known, mandatory attribute
- NEXT-HOP: IP address of border router to be used as next hop
- Usually, next hop is the router sending the UPDATE message
- Useful when some routers do not speak BGP

44

Copyright © William C. Cheng

BGP Attributes For Policy Control

- So, why policy routing?
- business relationships
- control (optimize) routes
- multi-homing: control traffic over multiple links

45

Copyright © William C. Cheng

LOCAL-PREF

Higher preference wins

160.10.0.0/16 500 > 160.10.0.0/16 800

AS 400
AS 200
AS 300
AS 100 (160.10.0.0/16)

46

Copyright © William C. Cheng

Sets and Sequences

- Solution: restructure AS-PATH attribute as:
 - path: (Sequence (3), Set (1, 2))
 - if AS4 wants to advertise path:
 - Path: (Sequence (4, 3), Set (1, 2))
- In practice used only if paths in set have same attributes

47

Copyright © William C. Cheng

Example of NEXT-HOP

UPDATE MSG through BGP

Traffic 138.39.0.0/16

138.39.0.0/16

C (no BGP)

A (BGP)

B (BGP)

48

Copyright © William C. Cheng

Policy 1: LOCAL-PREF Path Attribute

- Well-known, discretionary
- Provided by a BGP router to all other internal BGP routers
- denotes degree of preference for each destination
- From local configuration
- affects *your* AS only
- (does not propagate to others)
- can influence any prefixes
- Pick with path to prefer for a prefix
- Rule: BGP prefers paths with higher LOCAL-PREF

49

Copyright © William C. Cheng

54

AS-PATH Inflation Example

You are AS1 with two links A & B to AS2. How to make link A primary and B backup for incoming traffic?

AS1 exports: ?

AS2's state: ?

Computer Communications - CSC1 551

Copyright © William C. Cheng

52

LOCAL-PREF Example 2

You are AS1 with two links A & B to AS2. How to load-share AS2-bound traffic between links A & B?

AS1's routing table:

- 12.0/9:a(2) w/LP 10
- 12.0/9:c(2) w/LP 5
- 12.128/9:a(2) w/LP 5
- 12.128/9:c(2) w/LP 10

Computer Communications - CSC1 551

Copyright © William C. Cheng

50

LOCAL-PREF Example 1

You are AS1 with two links A & B to AS2. How to force all traffic to AS2's prefix 128 through link A?

AS1's routing table:

- 128:a(2) w/LP=10
- 128:c(2) w/LP=5

Computer Communications - CSC1 551

Copyright © William C. Cheng

53

Policy 2: AS-PATH Inflation

- From local configuration
- affects *all* AS's in the Internet
- affects only your prefixes
- Make a path look worse than its
- Rule: *BGP prefers shorter AS-PATHS*

Computer Communications - CSC1 551

Copyright © William C. Cheng

51

LOCAL-PREF Example 2

You are AS1 with two links A & B to AS2. How to load-share AS2-bound traffic between links A & B?

AS1's routing table: ?

Computer Communications - CSC1 551

Copyright © William C. Cheng

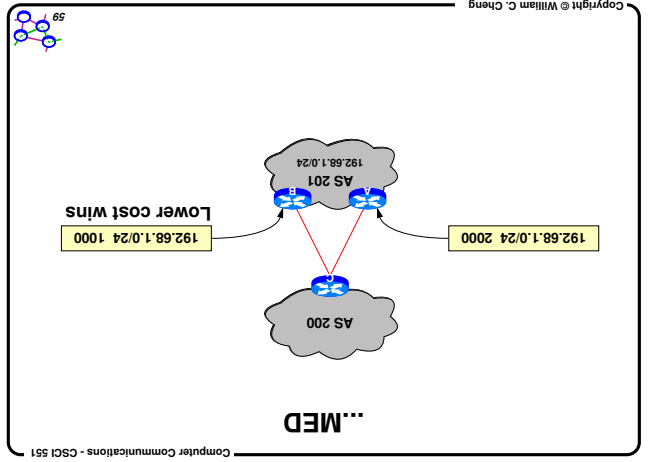
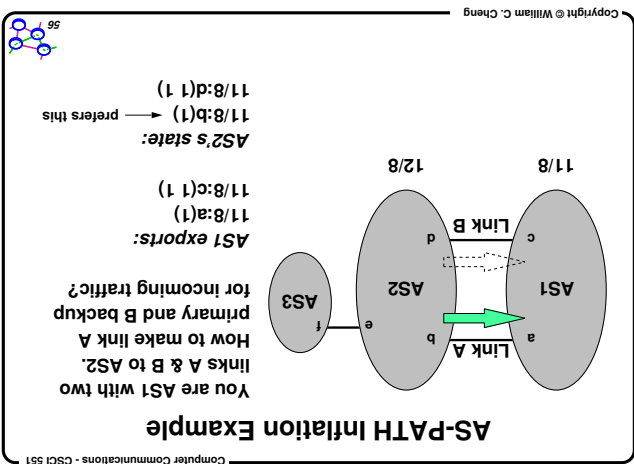
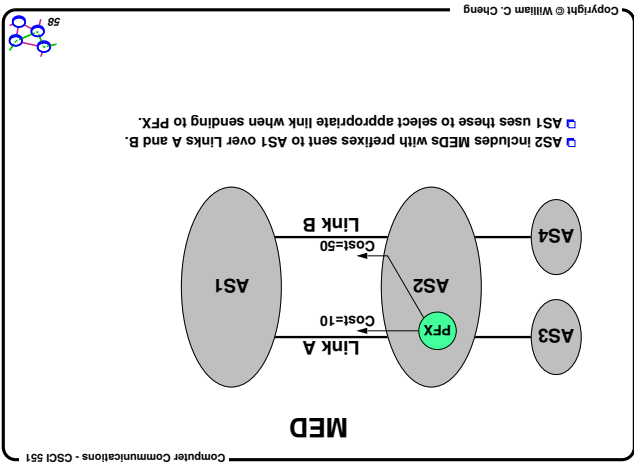
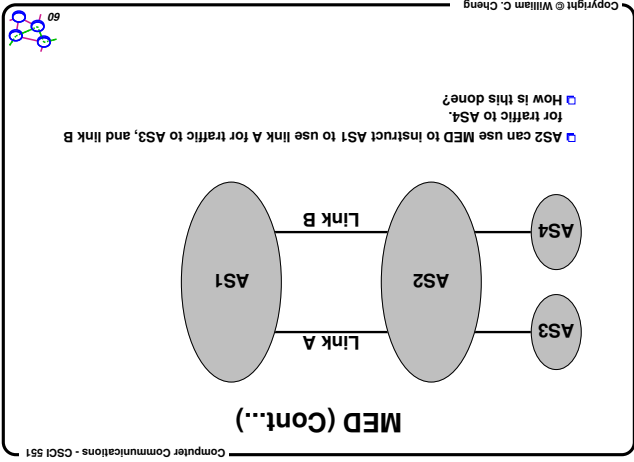
49

LOCAL-PREF Example 1

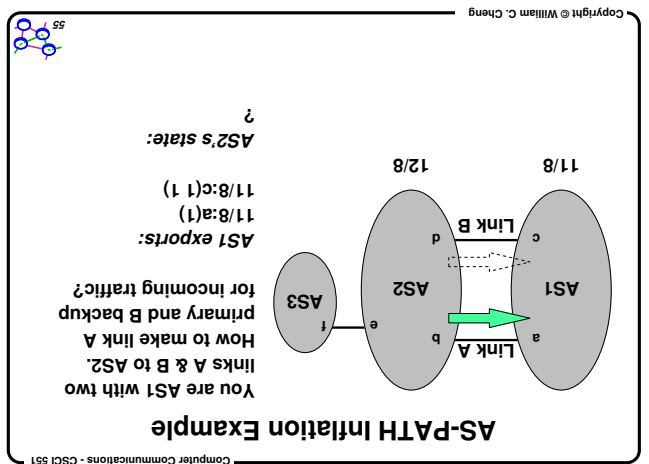
You are AS1 with two links A & B to AS2. How to force all traffic to AS2's prefix 128 through link A?

AS1's routing table: ?

Computer Communications - CSC1 551



- Copyright © William C. Cheng
- Policy 3: Multi-exit Discriminator (MED)**
- Optional, non-transitive attribute
 - Used when two AS's connect to each other in more than one place
 - Carries a metric expressing degree of preference
 - From local configuration
 - affects prefixes you propagate
 - affects adjacent AS's
 - Used to help others pick the right exit point
 - therefore they probably trust you (e.g., client/provider relationship)
- Rule: BGP prefers the lowest MED
- Computer Communications - CSC1 551



Copyright © William C. Cheng

UPDATE Message Handling

- Unrecognized, optional, non-transitive attributes are ignored. Unrecognized, optional, transitive attributes are the Partial bit to be set.
- WITHDRAWN routes are processed first.
- Feasible routes are placed in Adj-RIB-in, replacing old ones, if any.

Computer Communications - CSC1 551

Copyright © William C. Cheng

Route Selection

Question: which routes should be installed in the forwarding table?

- Input: All routes that have been learned and accepted by a router
- If only one route, then select it
- If multiple routes (with same length prefix) then we have a decision to make

Computer Communications - CSC1 551

Copyright © William C. Cheng

MED is Non-transitive

- AS1 sends MEDs to AS2, AS2 will not pass these MEDs to AS3 and AS4
- MEDs are relative to links A and B only
- Cannot combine or compare MEDs from different AS's
- AS1 learns two ways to reach AS3, one from AS2 and one from AS4, cannot compare MEDs

Computer Communications - CSC1 551

Copyright © William C. Cheng

MED (Cont...)

- MED is typically used in provider/subscriber scenarios.
- It can lead to unfairness if used between ISPs because it may force one ISP to carry more traffic.
- ISP1 ignores MED from ISP2
- ISP2 obeys MED from ISP1
- ISP2 ends up carrying traffic most of the way
- "hot potato routing"
- Results: MED ignored by ISP's that don't trust each other

Computer Communications - CSC1 551

Copyright © William C. Cheng

MED Example

You are AS1 with two links A & B to AS2. How can you make AS2 send north traffic to link A and south traffic to link B?

AS1 exports:
 11.0/16:a(1) w/ MED=10
 11.0/16:c(1) w/ MED=20
 11.1/16:a(1) w/ MED=20
 11.1/16:c(1) w/ MED=10

Computer Communications - CSC1 551

Copyright © William C. Cheng

MED Example


You are AS1 with two links A & B to AS2. How can you make AS2 send north traffic to link A and south traffic to link B?

AS1 exports: ?

Computer Communications - CSC1 551

Computer Communications - CSC1 551

Copyright © William C. Cheng




BGP's Importance

- BGP is a very powerful protocol
- support for *policy* is unique among deployed routing protocols
- The key to global connectivity of the Internet
- Yet, it is so complex that many pathologies are being discovered even now, nearly a decade after initial deployment
- delayed convergence [Labovitz00]
- persistent oscillation (Varadhan 1996 and Griffin 2000)
- router-reflector pathologies (Basu 2002)

Computer Communications - CSC1 551

Copyright © William C. Cheng



Decision Process

- as follows (apply following steps until one route is left):
- 1) Select route with *highest LOCAL-PREF*
- 2) Select route with *shortest AS-PATH*
- 3) Apply MED (if routes learned from same neighbor), choose *lowest MED*
- 4) Select route with smallest NEXT-HOP cost (from IBGP, cost to edge router)
- 5) Select route learned from E-BGP peer with lowest BGP ID
- 6) Select route from I-BGP neighbor with lowest BGP ID
- Install selected route in Loc-RIB
- Disseminate routes to peers, update Adj-RIB-Out
- Done