BGP Routing

CSci551: Computer Networks
SP2006 Thursday Section
John Heidemann

Prerequesites

- Addressing
  - subnetting and CIDR
  - aggregation
- Routing algorithms

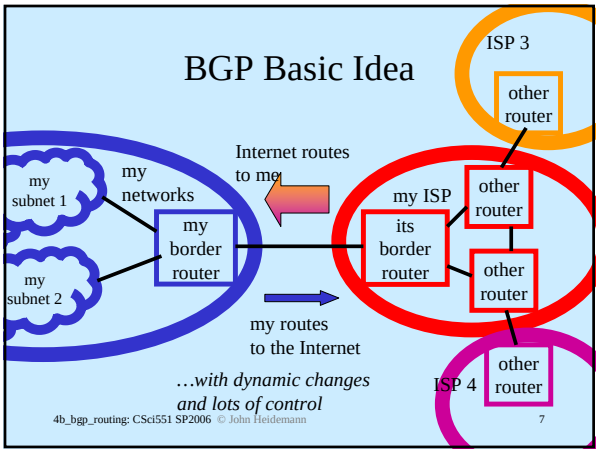
References

- [Gao00b] sections 2.1 and 3.1
  - excellent terse overview of BGP
  - see also [Gao02a] section II
- *BGP4: Inter-Domain Routing in the Internet*, by John Stewart
  - highly recommend for details about this material
- RFC-1654 [Rekhter95a]

Why BGP?

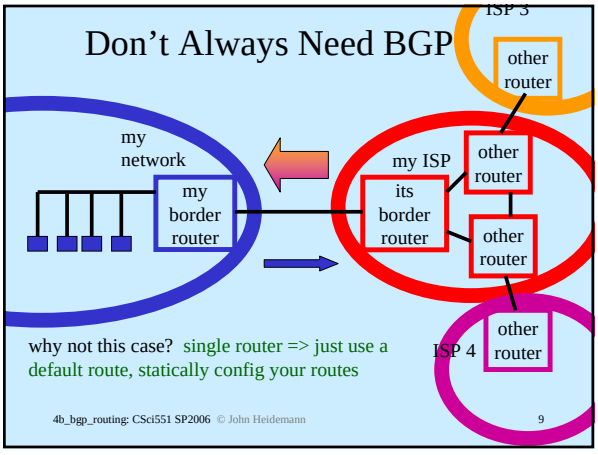
- widely used in the Internet
- primary EGP today
- real protocol
- very flexible
  - allows implementation of policies

BGP Basic Idea

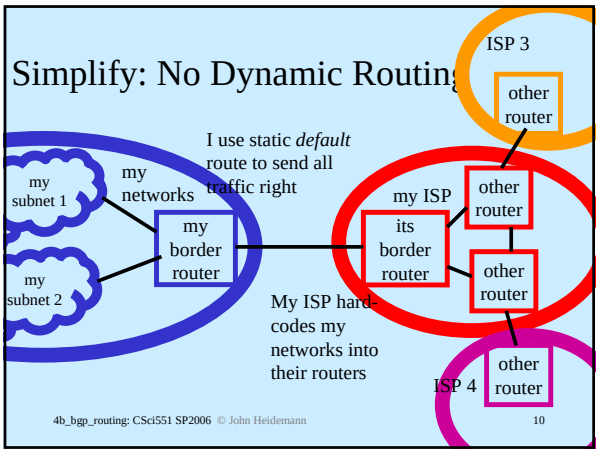


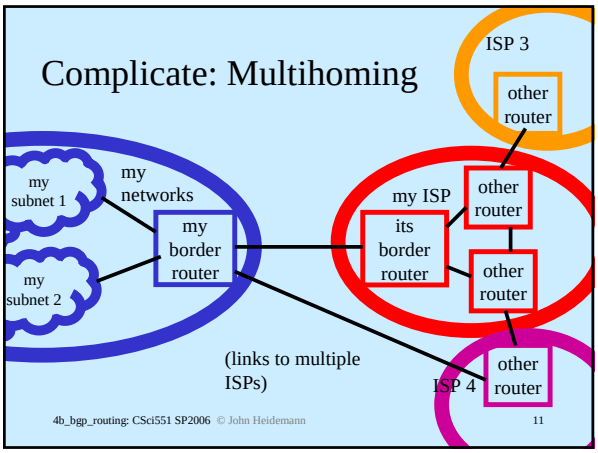
…with dynamic changes and lots of control

Don't Always Need BGP



why not this case? single router => just use a default route, statically config your routes





Where and Why is BGP Used?

- where
  - Internet routers
  - betweens ISPs (primarily EGP)
- why
  - to control routing to improve efficiency
  - control policy for business reasons

Aspects of BGP

- **BGP as a protocol**
- mechanisms to control policy
- E-BGP vs. I-BGP
- Multihoming
- Other issues

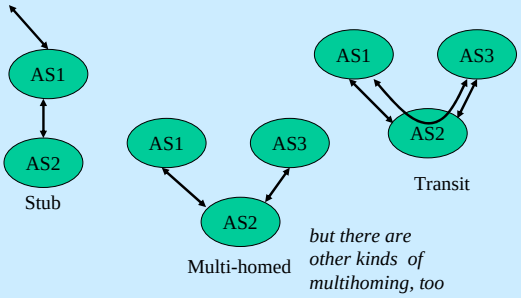
BGP Terminology

- AS: autonomous system (a area of the network under a single adminstrative control for routing)
- peer: an adjacent router (a router you swap routes with)
  - peering: two ASes that agree to swap traffic
- exchange point: location where many ISPs come together
  - NAP, POP
- RIB: routing information base

AS Types

- stub: AS with one connection to the internet
- multi-homed: multiple connections to other ISPs
- transit traffic: send traffic from you to a third party
- default route: xxx

AS categories

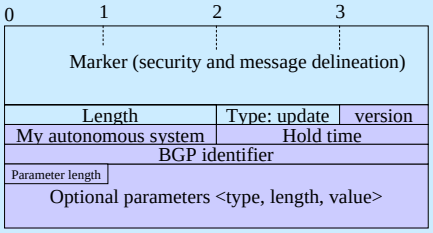


but there are other kinds of multihoming, too

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

BGP OPEN message



- My AS: 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

Messages go over TCP

- Why TCP?
  - reliablity
  - routing tables can be large (need ordering, bytestream, etc.)
  - want to send routing info over multiple hops
  - compare OSPF vs. BGP in [Shaikh00a]
- What and why are BGP keep-alives?
  - [see Shaikh00a]

Protocol Observations

- how does BGP know when the link is out?
  - KEEPALIVES or periodic messages
  - …we'll talk about this in [Shaikh00a]
- how does BGP avoid looping paths?
  - AS-PATH with each route
    - list of all the ASes were in the route, as far as you know
  - …we'll talk about this in [Labovitz00a]

BGP Attributes

- ORIGIN: where prefix originates
- AS-PATH: path for routing
- NEXT-HOP: where to send data
- MULTI-EXIT-DISCRIMINATOR: used to influence multihoming

will cover policy stuff shortly

Why BGP attributes?

- for correctness
  - ex: loop elimination
- to implement policy
  - can do a lot more

AS-PATH

- Why?
  - want to do loop detection to avoid counting to infinity
- What?
  - accumulate list of each AS that forwards route
  - but what about aggregation?
    - support sets and sequences in AS path (ex.)

Aspects of BGP

- BGP as a protocol
- **mechanisms to control policy**
- E-BGP vs. I-BGP
- Multihoming
- Other issues

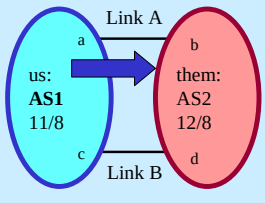
Why policy?

- to implement commercial/business constraints
- could be fore security reasons
  - ex: when stopping a DoS attack
- control packet flow for efficiency
  - different links with diff bandwidth
  - load balancing

Policy 1: LOCAL-PREF

- 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*

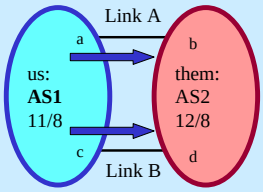
LOCAL-PREF Example 1



You are AS1 with two links A & B to AS2.
How to force all traffic to AS2's prefix 12/24 through link A?

12/8: b (LP=12)
12/8: d (LP=8)

LOCAL-PREF Example 2



but... xxx

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

12.0/9: b (LP=18)
12.0/9: d (LP=8)
12.128/9: b (LP=8)
12.128/9: d (LP=18)

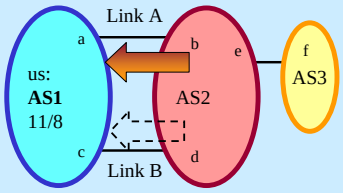
How Does BGP Pick Routes?

- overall goal:
  - determine connectivity
    - need current info (it's changing)
    - need to converge (mostly getting the same answer; avoiding loops)
  - allow policy to control routing
- rules
  1. highest local pref
  2. shortest AS-PATH
  3. lowest MED multi-exit discriminator
  4. minimum IGP distance to NEXT-HOP
  5. E-BGP route from speaker with lowest ID
  6. I-BGP router from speaker with lowest ID
- rules 5-6 are there to break ties consistently

Policy 2: AS-PATH Inflation

- From local configuration
  - affects all ASs in the Internet
  - affects only your prefixes
- Make a path look worse than it is
- Rule: BGP prefers shorter AS-PATHs
- AS path:
  - 128.9/16:a (AS152 AS3465 AS388)
  - 128.9/16:b (AS152 AS44)
  - idea: fewer AS hops is correlated with shorter paths
- path inflation: repeat AS numbers
  - 128.9/16:b (AS152 AS44 AS44 AS44)

AS-PATH Inflation Example



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

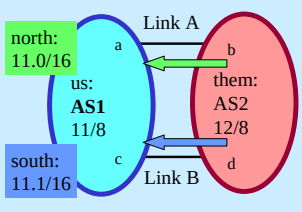
export from AS1:
11/8(a) (AS1)
11/8(c) (AS1 AS1)

comes from AS2
11/8(a) (AS2 AS1)
11/8(c) (AS2 AS1 AS1)

Policy 3: MULTI-EXIT-DISCRIMINATOR

- From local configuration
  - affects prefixes you propagate
  - affects *adjacent* ASs
- Used to help others pick the right exit point
  - therefore they probably trust you (ex. client/provider relationship)
- Rule: *BGP prefers the lowest MED*

MED Example 1

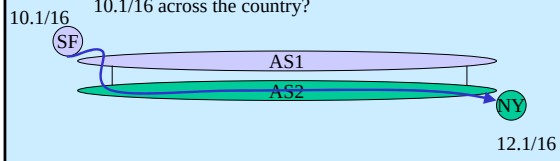


You are AS1
with two links A & B
to AS2.
How can you make
B send north traffic to
link A and south traffic
to link B?

11.0/16(a) MED=5
11.0/16(c) MED=10
11.1/16(a) MED=10
11.1/16(c) MED=5

MED Ex 2: hot potato routing

MED requires trust: suppose AS1 and AS2 both have long-haul nets. Who routes traffic from 10.1/16 across the country?



Arguably, AS1 *should* route the traffic (10.1/16 is its customer). But if AS1 ignores 2's MEDs and sends MEDs to AS2 it can make 2 carry the traffic most of the way.

How Does BGP Prefer Paths?

1. (ordering from before)

Aspects of BGP

- BGP as a protocol
- mechanisms to control policy
- **E-BGP vs. I-BGP**
- Multihoming
- Other issues

EGP vs. IGP

- exterior vs. interior
- world vs. me
- little control vs. complete admin control
- BGP (and GGP,Hello,EGP) vs. (RIP, OSPF, IS-IS,IGRP,EIGRP)

Why BGP as an IGP?

- I-BGP has mechanisms to forward BGP policy directives across AS
- often use I-BPG *with* some other IGP that does internal routing
- but…
  - because IGP is all under your control, it's relatively easy to do other protocols
    - static routing
    - propriatory protcolos

E-BGP vs. I-BGP

- E-BGP connects ASes (*external* GP)
- I-BPG is *intra*-AS (*internal* GP)

⇒ differences in operation

- direct vs. indirect connections
- different failure modes
- special attributes for internal use
- details are in [Stewart99a]

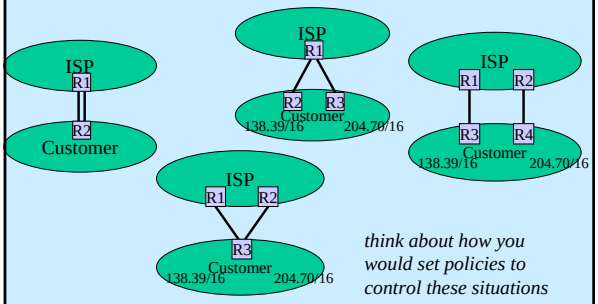
Aspects of BGP

- BGP as a protocol
- mechanisms to control policy
- E-BGP vs. I-BGP
- **Multihoming**
- Other issues

Multihoming: What and Why?

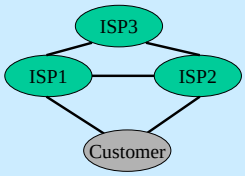
- what?
  - an ISP with multiple connections to the rest of the Internet
- why?
  - load balancing
  - routing policies (ex. education only)
  - robustness

Multihoming to Single Provider



think about how you would set policies to control these situations

Multihoming to Multiple Providers



- very robust
- but who gives you address space?

⇒one ISP, but then there are aggregation problems

⇒both ISPs, but still problems

⇒own address space, but then no aggregation

Aspects of BGP

- BGP as a protocol
- mechanisms to control policy
- E-BGP vs. I-BGP
- Multihoming
- **Other issues**

Other BGP-related Issues

- Router synchronization [Floyd94b]
- Convergence time [Labovitz00a]
- Congestion [Shaikh00a]
- Policy and convergence [Gao00a, Tangmunarunkit01a]
- Misconfiguration [Mahajan et al, 2002]
- other other issues
  - route flap dampening
  - routing arbiter—central DB of policies
  - robustness in the face of router resource exhaustion [Chang, Govindan, Heidemann]

Some BGP Stats (as of 25-Jan-06)

data from Japan (bgp-stats@lists.apnic.net)

- BGP routing table entries: 179,268
  - prefixes after max. aggregation: 100,443
- addresses announced: 1,482,719,520
  - 40% of available address space announced
  - 59% of the allocated address space announced
  - 68% of available address space allocated
- ASes in Internet rtg table: 21,369
  - origin-only ASes: 18,588
  - origin-only ASes w/only one prefix: 8,804
  - transit ASes: 2,781
- AS path length
  - mean: 4.5
    - down from 5.0 in 2004 and 5.3 in 2003, same as 2005
  - maximum seen: 22 (up from 19 in 2004 and 17 in 2003, same as 2005)

Other questions?

- what happens if someone doesn't use BPG?
  - the internet is made of peerings