Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the de facto* standard

- BGP provides each AS a means to:
  1. Obtain subnet reachability information from neighboring ASs.
  2. Propagate the reachability information to all routers internal to the AS.
  3. Determine “good” routes to subnets based on reachability information and policy.

- Allows a subnet to advertise its existence to rest of the Internet: “I am here”
BGP-4

- BGP = Border Gateway Protocol
- Is an exterior routing protocol (EGP)
- Is a Policy-Based routing protocol
- Is the de facto EGP of today’s global Internet
- Has a reputation for being complex
- Supports hierarchical routing
- Is a distance vector protocol
BGP history

- 1989: BGP-1 [RFC 1105]
  - Replacement for EGP (1984, RFC 904)
- 1990: BGP-2 [RFC 1163]
- 1991: BGP-3 [RFC 1267]
- 1995: BGP-4 [RFC 1771] (only 57 pages!)
  - Support for CIDR

Changes primarily driven by scalability issues.
Development dominated by Cisco.
Routing tasks: BGP

- Neighbor?
  - Discovery
  - Maintenance

- Database?
  - Granularity
  - Maintenance – updates
  - Synchronization

- Routing table?
  - Metric
  - Calculation
  - Update
BGP Basics

- Pairs of routers (BGP peers) exchange routing info over semi-permanent TCP connections: **BGP sessions**
- Note that BGP sessions do not correspond to physical links.
- When AS2 advertises a prefix to AS1, AS2 is *promising* it will forward any datagrams destined to that prefix towards the prefix.
  - AS2 can aggregate prefixes in its advertisement
Distributing reachability info

- With eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
- 1c can then use iBGP to distribute this new prefix reachability info to all routers in AS1.
- 1b can then re-advertise the new reachability info to AS2 over the 1b-to-2a eBGP session.
- When a router learns about a new prefix, it creates an entry for the prefix in its forwarding table.
Current Internet architecture

Arbitrary Internetwork of Autonomous Systems

An Autonomous System is a unified administrative domain with a consistent routing policy.

A few years ago about 7000 AS numbers are assigned, about 4200 in use.
Routing policy

- Reflects goals of network provider
  - Which routes to accept from other ASes
  - How to manipulate the accepted routes
  - How to propagate routes through network
  - How to manipulate routes before they leave the AS
  - Which routes to send to another AS
Policies with BGP

- BGP provides capabilities for enforcing various policies
- Policies are **not** part of BGP!
- Policies are used to configure BGP
- BGP enforces policies by choosing paths from multiple alternatives and controlling advertisements to other AS’s
Routing policy examples

- Honor business relationships
  (e.g., customers get full-table; peers only customer prefixes)
  (e.g., prefer customer routes over peer routes over upstream routes)

- Allow customers a choice of route
  (e.g., on customer request do not export prefix to AS x, etc.)

- Enable customer traffic engineering
  (e.g., prepend x times to all peers or to specified AS)

- Enable DDoS defense for customers
  (e.g., blackholing by rewriting the next hop)

- ...

Why policy should win over distance metrics

Even if it is the shortest path!
Stub vs. multihomed networks

Multihomed Networks

AS23

AS2006

AS300

AS1717

AS400

Stub Networks

Multihomed networks are “required” to run BGP
Routing at Stub ASs

Upstream Provider

AS100

Static Route

204.10.0/23

Default Route
Policy: Transit vs. Nontransit

A transit AS allows traffic with neither source nor destination within AS to flow across the network.

A nontransit AS allows only traffic originating from AS or traffic with destination within AS.
BGP operations simplified

Establish Peering on TCP port 179

Peers Exchange All Routes

Exchange Incremental Updates

BGP Route = network prefix + attributes

While connection is ALIVE exchange route UPDATE messages
Path attributes & BGP routes

- When advertising a prefix, advertisement/update includes BGP attributes.
  - prefix + attributes = “route”

- Two important attributes:
  - **AS-PATH**: Contains the ASs through which the advertisement for the prefix passed: AS 67 AS 17
    - Used for loop detection / policies
  - **NEXT-HOP**: Indicates the specific internal-AS router to next-hop AS. (There may be multiple links from current AS to next-hop AS.)

- When gateway router receives route advertisement, uses **import policy** to accept/decline.
AS Path attribute

135.104.0.0/16
AS Path = 701 144

AS702
Alternet (Uunet)

135.104.0.0/16
AS Path = 702 701 144

AS1849
Uunet UK

135.104.0.0/16
AS Path = 5459 5413 1 144

AS5459
LINX

135.104.0.0/16
AS Path = 5413 1 144

AS5413
GXN

135.104.0.0/16
AS Path = 1 144

AS144
Bell Labs

135.104.0.0/16
Route Originated

AS1
BBN

135.104.0.0/16
AS Path = 1 144
Next Hop attribute

AS 200
150.10.0.0/16

AS 100
160.10.0.0/16

AS 300

150.10.1.1
150.10.1.2

150.10.0.0/16
150.10.1.1
160.10.0.0/16
150.10.1.1
BGP attributes

- AS path (well-known, mandatory)
- Next Hop (well-known, mandatory)
- Origin (well-known, mandatory)
- Multiple Exit Discriminator (MED) (Optional, nontrans, eBGP)
- Local Preference (LocPref) (well-known, discretionary, iBGP)
- Community (Optional, transitive)
- Atomic Aggregate (well-known, discretionary)
- Aggregator (Optional, transitive)
- Originator ID (Optional, nontransitive, Cisco)
- Other vendor-specific optional attributes ...
BGP route processing

Receive BGP Updates
Apply Policy = filter routes & tweak attributes
Based on Attribute Values
Best and Alternate Routes
Apply policies only to Best Routes!
Transmit BGP Updates

Apply Import Policies
Best Route Selection
BGP Route Table
Apply Export Policies

Install Best Routes
IP Forwarding Table

Only this is Detailed in RFC 1771
BGP route selection

- Router may learn about more than one route to some prefix.
- Router must select route.
- Elimination rules:
  1. Local preference value attribute: policy decision
  2. Shortest AS-PATH
  3. Route with lowest MED
  4. Closest NEXT-HOP router: hot potato routing
  5. Additional criteria
  6. Pick route from router with lowest IP address (break tie)
Peers exchange BGP messages using TCP

BGP messages:

- **OPEN:**
  - Opens TCP conn. to peer
  - Authenticates sender

- **UPDATE:**
  - Advertises new path (or withdraws old)

- **KEEPALIVE:**
  - Keeps conn alive in absence of UPDATES
  - Serves as ACK to an OPEN request

- **NOTIFICATION:**
  - Reports errors in previous msg;
  - Closes a connection
BGP routing policy

- A, B, C are provider networks
- X, W, Y are customer (of provider networks)
- X is dual-homed: attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C
BGP routing policy (2)

- A advertises to B the path AW
- B advertises to X the path BAW
- Should B advertise to C the path BAW?
  - No way! B gets no “revenue” for routing CBAW since neither W nor C are B’s customers
  - B wants to force C to route to w via A
  - B wants to route only to/from its customers!
Why different Intra- and Inter-AS routing?

**Policy:**
- Inter-AS: Admin wants control over how its traffic routed, who routes through its net.
- Intra-AS: Single admin, so no policy decisions needed

**Scale:**
- Hierarchical routing saves table size, reduced update traffic

**Performance:**
- Intra-AS: Can focus on performance
- Inter-AS: Policy may dominate over performance

We need BOTH!
Local Preference attribute

Path with highest local preference wins
Local Preference – common uses

- Handle traffic directed to multi-homed transit customers
  - Allows providers to prefer a route

- Peering vs. transit
  - Prefer to use peering connection
  - Customer > peer > provider
Multi-Exit Discriminator (MED)

- Non-transitive
- Used to convey the relative preference of entry points
- Influences best path selection
- Comparable if paths are from same AS
- IGP metric can be conveyed as MED
MED attribute

- Used to convey the relative preference of entry points
- Comparable if paths are from same AS
- IGP metric can be conveyed as MED
Communities

- Used to group prefixes and influence routing decisions (accept, prefer, redistribute, etc.), e.g., via route-maps to realize routing policies
- Represented as an integer
  Range: 0 to 4,294,901,760
- Each destination could be member of multiple communities
- Community attribute carried across AS’s
BGP communities

Community 10:200
Community 10:300
Community 10:200
Community 10:300
Load balancing

- BGP does not load-balance traffic; it chooses & installs a “best” route.

“Since BGP picks a ‘best’ route based upon most specific prefix and shortest AS_PATH, it becomes non-trivial to figure out how to manually direct specific portions of internal traffic (prefixes) in a distributed fashion across multiple external gateways.”
Difficulties in load balancing

192.10.0/16
AS100
204.10.14.0/23

AS200

AS300
Multi-homing

- Multi-homing:
  - Network has several connections to the Internet.
- Improves reliability and performance:
  - Can accommodate link failure
  - Bandwidth is sum of links to Internet
- Challenges
  - Getting policy right (MED, etc..)
  - Addressing
Multi-homing to multiple providers

- **Major issues:**
  - Addressing
  - Aggregation

- **Customer address space:**
  - Delegated by ISP1
  - Delegated by ISP2
  - Delegated by ISP1 and ISP2
  - Obtained independently
Address space from one ISP

- Customer uses address space from ISP1
- ISP1 advertises /16 aggregate
- Customer advertises /24 route to ISP2
- ISP2 relays route to ISP1 and ISP3
- ISP2-3 use /24 route
- ISP1 routes directly
- Problems with traffic load?
Pitfalls

- ISP1 aggregates to a /19 at border router to reduce internal tables.
- ISP1 still announces /16.
- ISP1 hears /24 from ISP2.
- ISP1 routes packets for customer to ISP2!
- Workaround: ISP1 must inject /24 in I-BGP.
Address space from both ISPs

- ISP1 and ISP2 continue to announce aggregates
- Load sharing depends on traffic to two prefixes
- Lack of reliability: If ISP1 link goes down, part of customer becomes inaccessible.
- Customer may announce prefixes to both ISPs, but still problems with longest match as in case 1.
Independent address space

- Offers the most control, but at the cost of aggregation.
- Still need to control paths
- Many ISP’s ignore advertisements of less than /19
Internal BGP (iBGP)

- Same routing protocol as BGP, different application
- iBGP should be used when AS_PATH information must remain intact between multiple eBGP peers
- All iBGP peers must be fully meshed, logically; An iBGP peer will not advertise a route learned by one iBGP peer to another iBGP peer (readvertisement restriction to prevent looping)
iBGP peers must be fully meshed

- $N$ border routers means $N(N-1)/2$ peering sessions
  - this does not scale
- Currently three solutions:
  - Break an AS up into smaller Autonomous Systems
  - Route Reflectors
  - Confederations
Route reflectors

Route Reflectors must be fully meshed

Route Reflectors pass along updates to client routers
Confederations

To the global internet, this looks just like AS100
Link failures

- Two types of link failures:
  - Failure on an E-BGP link
  - Failure on an I-BGP Link

- These failures are treated completely different in BGP

- Why?
Failure of an E-BGP link

- If the link R1-R2 goes down
  - The TCP connection breaks
  - BGP routes are removed
- This is the desired behavior

Diagram:
- AS1
- R1
- E-BGP session
- R2
- AS2
- Physical link
- 138.39.1.1/30
- 138.39.1.2/30
Failure on an I-BGP link

- Link R1-R2 down ⇒ R1 and R2 can still exchange traffic
- The indirect path through R3 must be used
- E-BGP and I-BGP use different conventions with respect to TCP endpoints
  - E-BGP: no multihop – I-BGP: multihop OK
BGP summary

- Neighbors
  - discovery configured
  - maintenance keep-alives

- Database
  - granularity prefix
  - maintenance incremental updates & filter
  - synchronization full exchange

- Routing table
  - metric policies
  - calculation route selection
Routing protocols summary

- IS-IS
- OSPF
- RIP
- BGP

*Network layer*
- UDP
- Transport
- TCP

*Link layer*

*Physical layer*
A few problems

- BGP used to realize routing policy
- BGP dynamics
- Internet topology?
- Source routing?
- Naming?
- Security?
- How can ISPs make a profit?
- Simplicity vs. complexity?
Routing policy

Current state of the art:

- **Ill-specified** (e.g., policy database is the network itself)
- Undergoes constant adjustments
- Customer specific
- Conglomerate of BGP statements
- Realized by manual configuration of routers which routes to send to another AS
BGP dynamics

- Number of routes
  - 400K and growing
    - Traffic engineering
    - Protection
    - Alternative routes

- Route propagation
  - Better route: < 5 minutes
  - Route no longer reachable: < 20 minutes

- Dynamics
  - Small number prefix responsible for most churn

- Hard to pinpoint origin or route instability
BGP is not guaranteed to converge!

- BGP is not guaranteed to converge to a stable routing. Policy inconsistencies can lead to "livelock" protocol oscillations.

- Goal:
  - Design a simple, tractable, and complete model of BGP modeling
  - Example application: sufficient condition to guarantee convergence.
BGP may have multiple solutions

First solution

Second solution

DISAGREE
BGP routing policies for DISAGREE

1 2 0
1 0

1

import : from AS2 action pref = 0; accept ANY;
from AS0 action pref = 10; accept ANY;
export : to AS2 announce ANY;

0

export : to AS1, AS2 announce AS0;

2

import : from AS1 action pref = 0; accept ANY;
from AS0 action pref = 10; accept ANY;
export : to AS1 announce ANY;

2 1 0
2 0
BGP routing policies for DISAGREE (2)

Assume AS1 and AS2 use “neighbor send-community” command ....
Multiple solutions => “Route Triggering”

Remove primary link

Restore primary link
BAD GADGET: always diverges

The routing policies of this system have no solution—the protocol always diverges

See “Persistent Route Oscillations in Inter-domain Routing” by K. Varadhan, R. Govindan, and D. Estrin. ISI report, 1996
Bad Gadget: No solution

Stage 1:
1: [10]
2: [210]
3: [30]

Stage 2:
1: [130]
2: [20]
3: [320]

Back to stage 1
Bad Gadget: No solution

Stage 1:
1: [10]
2: [20]
3: [320]

Stage 2:
1:[130]
2:[210]
3:[30]

Back to stage 1
How to ensure no policy conflicts

Strawman Proposal: Perform Global Policy Check

- Require each AS to publish its policies
- Detect and resolve conflicts

Problems:

- ASes typically unwilling to reveal policies
- Checking for convergence is NP-complete
- Failures may still cause oscillations
Think globally, act locally

- Key features of a good solution
  - **Safety**: Guaranteed convergence
  - **Expressiveness**: Allow diverse policies for each AS
  - **Autonomy**: Do not require revelation/coordination
  - **Backwards-compatibility**: No changes to BGP

- *Local* restrictions on configuration semantics
  - Ranking
  - Filtering
Gao and Rexford Scheme

Gao & Rexford, “Stable Internet Routing without Global Coordination”, *IEEE/ACM ToN*, 2001

- Permit only two business arrangements
  - Customer-provider
  - Peering
- Constrain both *filtering* and *ranking* based on these arrangements to guarantee safety
- *Surprising result*: These arrangements correspond to today’s common behavior
Signs of routing instability

- Monitored BGP messages at major exchanges
- Orders of magnitude more updates than expected
  - Bulk: Duplicate withdrawals
    - Stateless implementation of BGP – did not keep track of information passed to peers
    - Impact of few implementations
  - Strong frequency (30/60 sec) components
    - Interaction with other local routing/links etc.
Route flap storm

- Overloaded routers fail to send Keep_Alive message and marked as down
- I-BGP peers find alternate paths
- Overloaded router re-establishes peering session
- Must send large updates
- Increased load causes more routers to fail!
Route flap dampening

- Route flap
  - Going up and down of path
  - Change in attribute
- Ripples through the entire Internet
- Consumes CPU
- Dampening
  - Reduce scope of route flap propagation
  - History predicts future behavior
  - Suppress oscillating routes
  - Fast convergence for normal route changes
Flap dampening: Operation

- Add penalty for each flap
- Exponentially decay penalty
- Penalty above suppress-limit—Do not advertise up route
- Penalty decayed below reuse-limit—Advertise route
- History path
Route flap dampening

Penalty

Time

Suppress-Limit

Reuse-Limit
Flap dampening: Operation (cont.)

- Done only for external path
- Alternate paths still usable
- Suppress-limit, reuse-limit and half-life time give control
- Less overhead
BGP soft reconfiguration

- Soft reconfiguration allows BGP policies to be configured & activated without clearing the BGP session
- Does not invalidate forwarding cache, hence no short-term interruptions
- Outbound preferable over inbound reconfiguration