Internet Routing Protocols
Lecture 03
Inter-domain Routing

Advanced Systems Topics

Lent Term, 2008

Timothy G. Griffin
Computer Lab
Cambridge UK

Autonomous Routing Domains

A collection of physical networks glued together using IP, that have a unified administrative routing policy.

- Campus networks
- Corporate networks
- ISP Internal networks
- ...

Autonomous Systems (ASes)

An autonomous system is an autonomous routing domain that has been assigned an Autonomous System Number (ASN).

> … the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it.
>
> RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System

AS Numbers (ASNs)

ASNs are 16 bit values (soon to be 32 bits)

64512 through 65535 are "private"

Currently nearly 30,000 in use.

- **JANET: 786**
- **MIT: 3**
- **Harvard: 11**
- **UC San Diego: 7377**
- **AT&T: 7018, 6341, 5074, ...**
- **UUNET: 701, 702, 284, 12199, ...**
- **Sprint: 1239, 1240, 6211, 6242, ...**
- **...**

ASNs represent units of routing policy

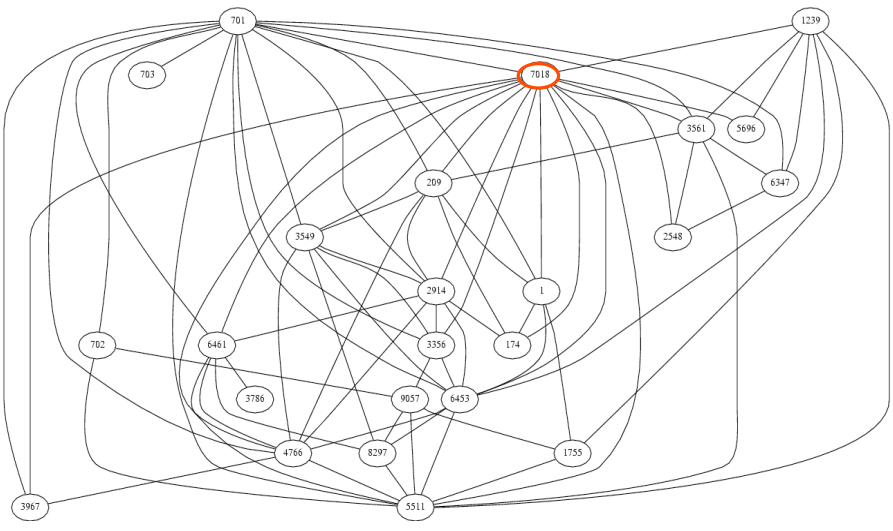
BGP Routing Tables

```
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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 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
*>                  203.50.0.33                            0 65056 4637 174 21889 i
*                   202.12.29.79                           0 4608 1221 4637 174 21889 i
*> 4.78.22.0/23     193.0.4.28                             0 12654 3257 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909 i
*                   203.50.0.33                            0 65056 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909
*                   202.12.29.79                           0 4608 1221 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13
*> 4.78.56.0/23     193.0.4.28                             0 12654 3257 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909 13909 i
*                   203.50.0.33                            0 65056 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13909
*                   202.12.29.79                           0 4608 1221 4637 1299 1239 19151 13909 13909 13909 13909 13909 13909 13909 13909 13
*  4.79.181.0/24    193.0.4.28                             0 12654 3741 10310 14780 i
*>                  203.50.0.33                            0 65056 4637 10310 14780 i
*                   202.12.29.79                           0 4608 1221 4637 10310 14780 i
```

Thanks to Geoff Huston.
http://bgp.potaroo.net on Feb 1, 2008

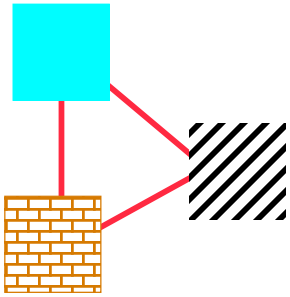
AS Graphs Can Be Fun



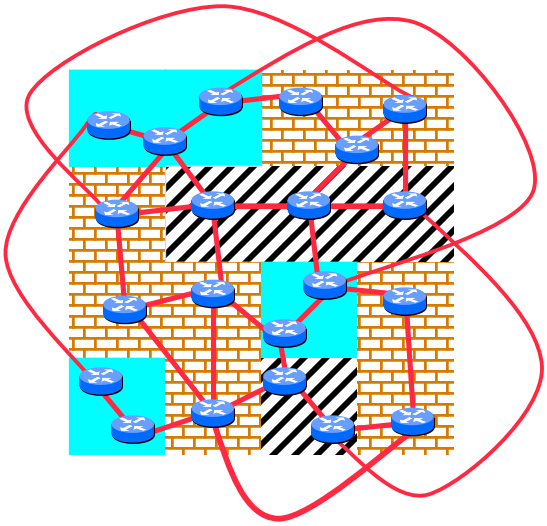
The subgraph showing all ASes that have more than 100 neighbors in full graph of 11,158 nodes. July 6, 2001. **Point of view: AT&T route-server**

AS Graphs Do Not Show "Topology"!

BGP was designed to throw away information!

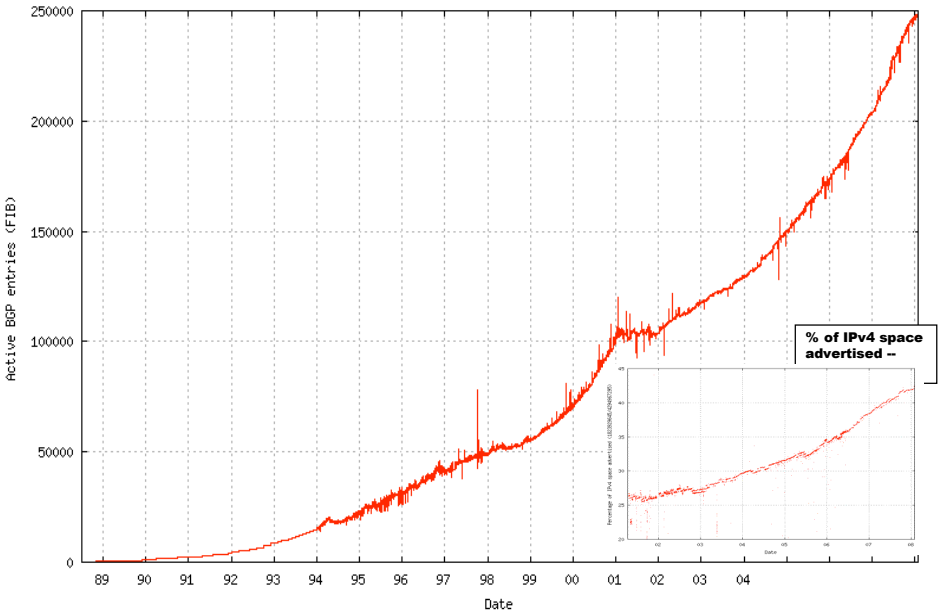


The AS graph may look like this.



Reality may be closer to this...

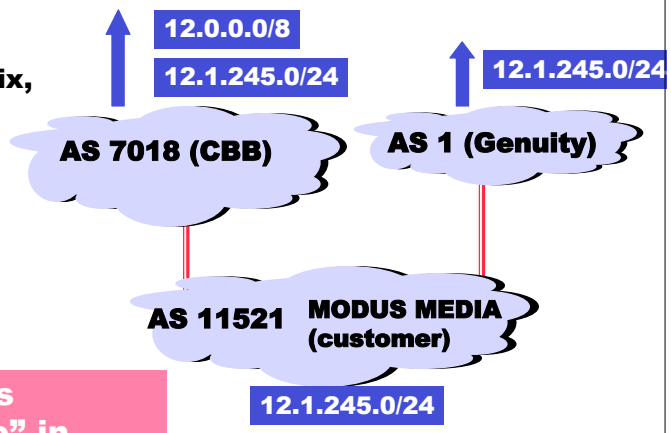
Routing Table Growth



Thanks to Geoff Huston. http://bgp.potaroo.net on Feb 1, 2008

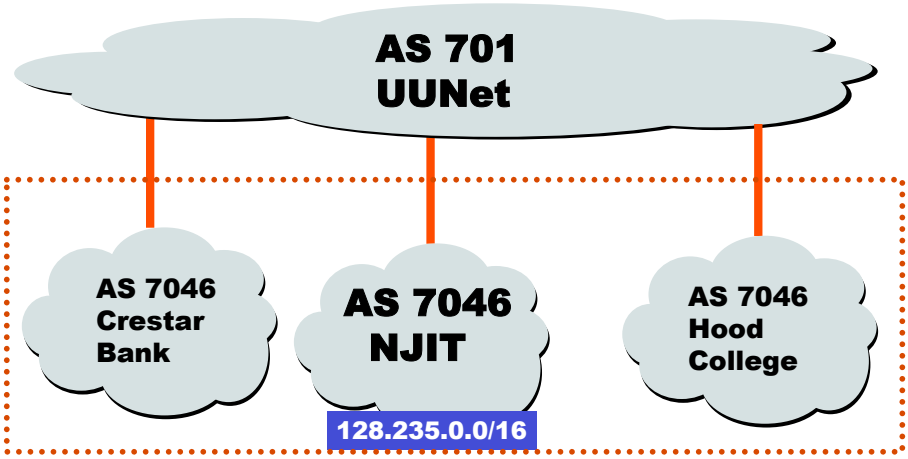
Deaggregation Due to Multihoming May Contribute to Table Growth

If AT&T does not announce the more specific prefix, then traffic to MODUS MEDIA will go through Genuity because it has a longer match....



MODUS MEDIA is "punching a hole" in the 12.0.0.0/8 CIDR block

ASNs Can Be "Shared" (RFC 2270)



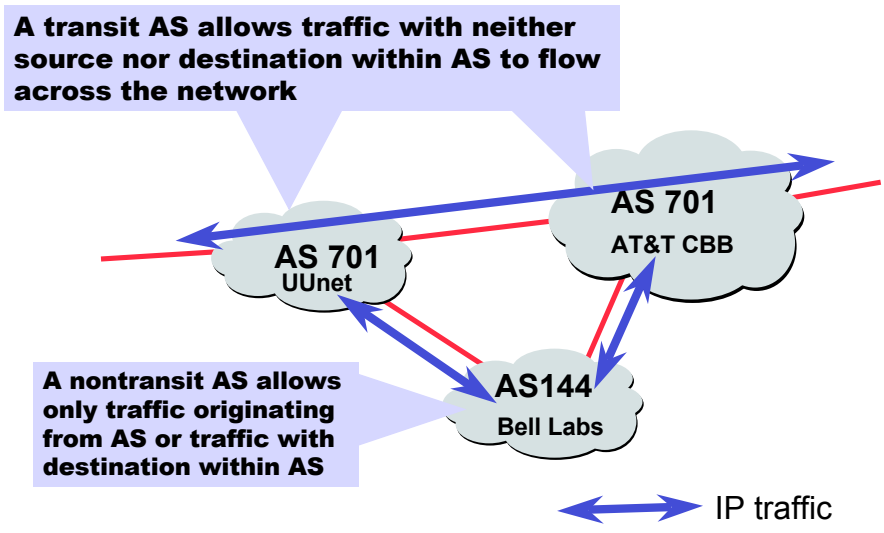
ASN 7046 is assigned to UUNet. It is used by Customers single homed to UUNet, but needing BGP for some reason (load balancing, etc..) [RFC 2270]

ARD != AS

- Most ARDs have no ASN (statically routed at Internet edge)
- Some unrelated ARDs share the same ASN (RFC 2270)
- Some ARDs are implemented with multiple ASNs (example: Worldcom)

ASes are an implementation detail of Interdomain routing

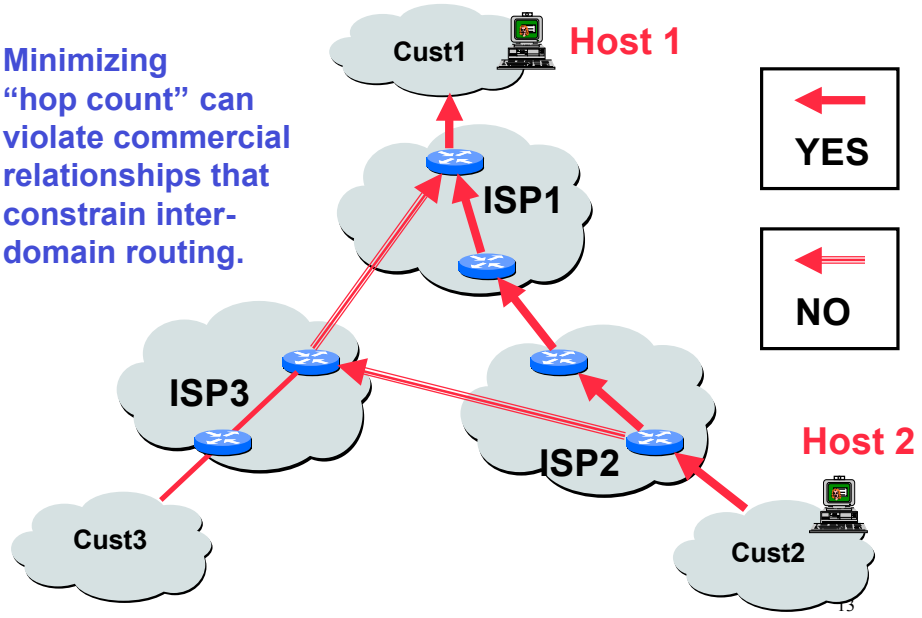
Policy : Transit vs. Nontransit



12

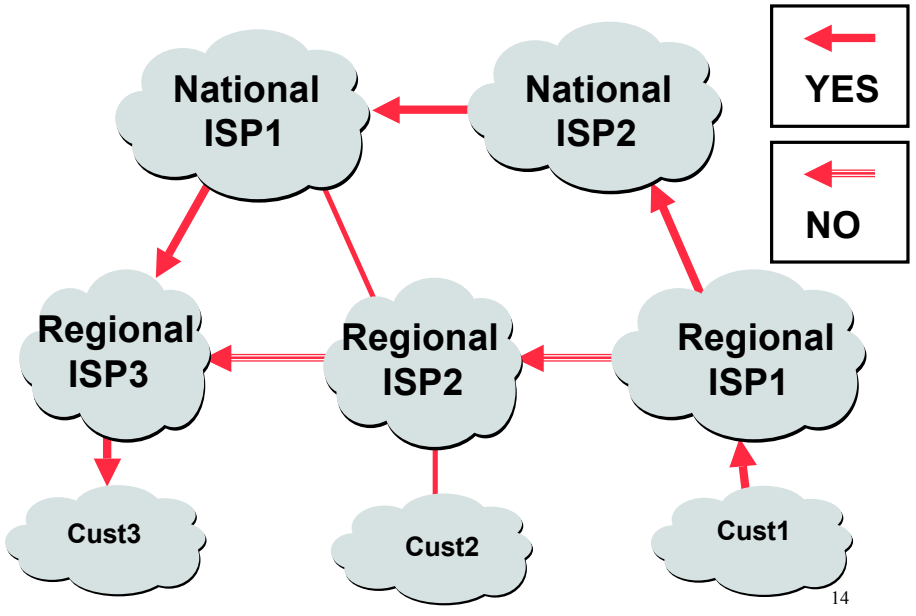
Policy-Based vs. Distance-Based Routing?

Minimizing "hop count" can violate commercial relationships that constrain inter-domain routing.



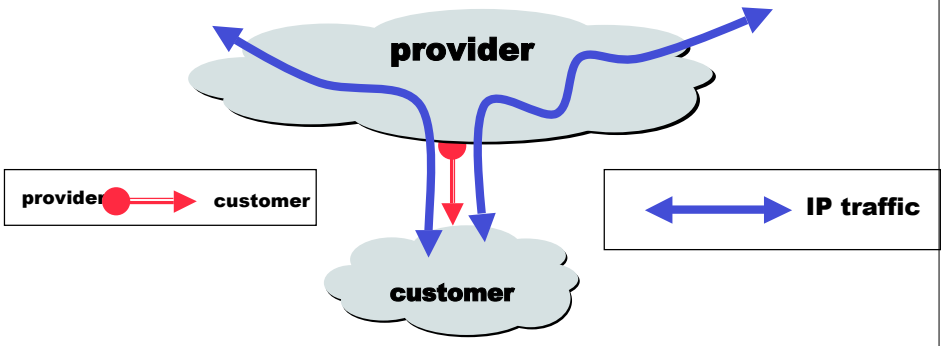
13

Why not minimize "AS hop count"?



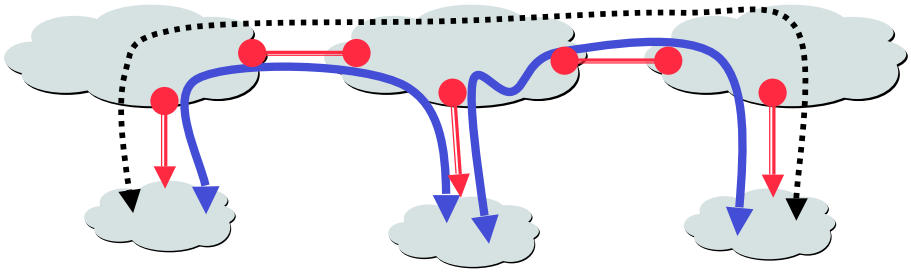
14

Customers and Providers



Customer pays provider for access to the Internet

The "Peering" Relationship



peer — peer

provider → customer

traffic allowed

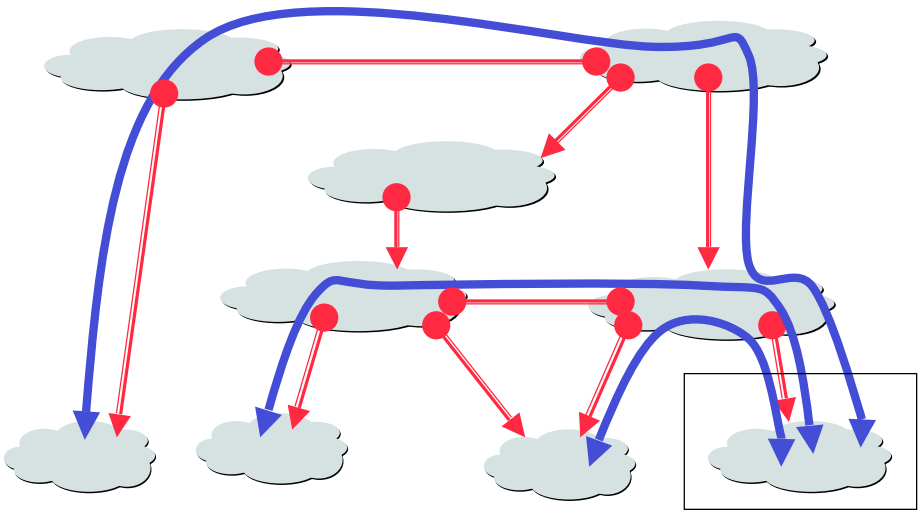
traffic NOT allowed

Peers provide transit between their respective customers

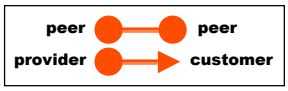
Peers do not provide transit between peers

Peers (often) do not exchange $$$

Peering Provides Shortcuts



Peering also allows connectivity between the customers of "Tier 1" providers.



Peering Wars

Peer

- Reduces upstream transit costs
- Can increase end-to-end performance
- May be the only way to connect your customers to some part of the Internet ("Tier 1")

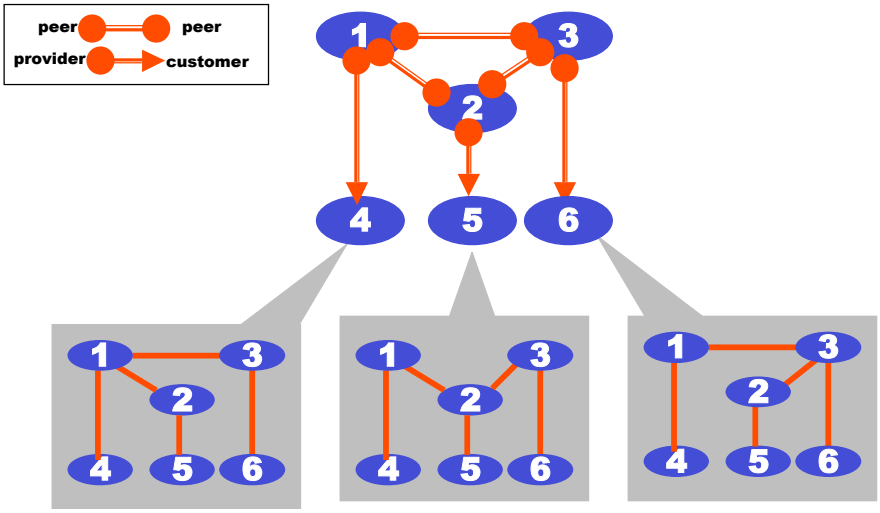
Don't Peer

- You would rather have customers
- Peers are usually your competition
- Peering relationships may require periodic renegotiation

Peering struggles are by far the most contentious issues in the ISP world!

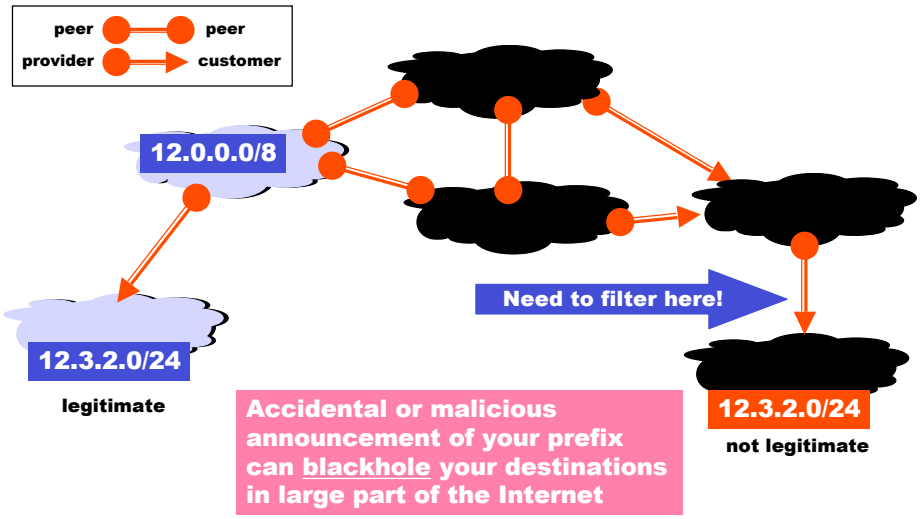
Peering agreements are often confidential.

AS Graphs Depend on Point of View



This explains why there is no UUNET (701) Sprint (1239) link on previous slide!

Blackholes



Accidental or malicious announcement of your prefix can blackhole your destinations in large part of the Internet

Commandments of Interdomain Routing

- Thou shall prefer customer routes over all others
- Thou shall use provider routes only as a last resort
- Thou shall not provide transit between peers or providers
- Thou shall verify customer address space, or burn in hell

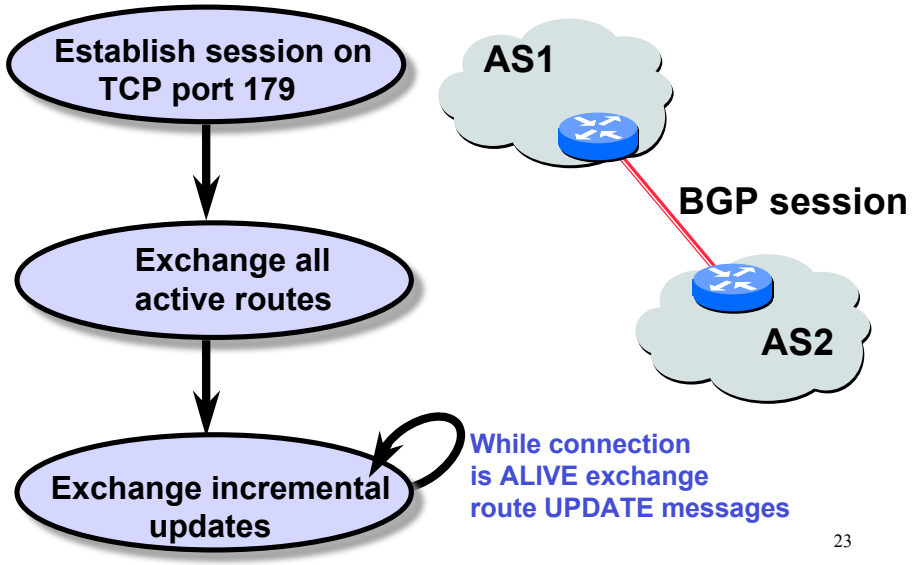
BGP-4

- **BGP** = **B**order **G**ateway **P**rotocol
- Is a **Policy-Based** routing protocol
- Is the **de facto EGP** of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.

- **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]**
  - **Support for Classless Interdomain Routing (CIDR)**
- **2006 : BGP-4 [RFC 4271]**

22

BGP Operations (Simplified)



23

Four Types of BGP Messages

- **Open** : Establish a peering session.
- **Keep Alive** : Handshake at regular intervals.
- **Notification** : Shuts down a peering session.
- **Update** : Announcing new routes or withdrawing previously announced routes.

announcement
=
prefix + attributes values

24

BGP Attributes

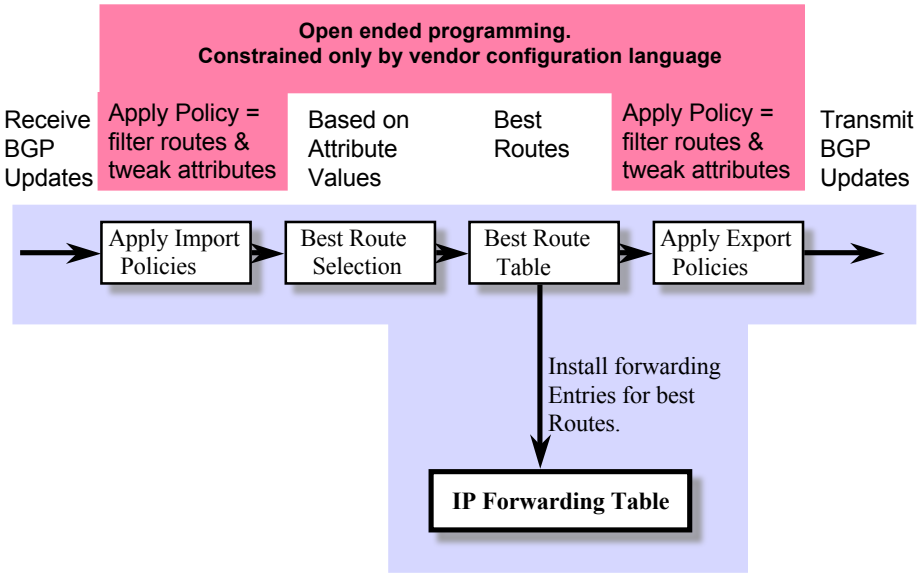
Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	[Rosen]
...		
255	reserved for development	

From IANA: http://www.iana.org/assignments/bgp-parameters

Most important attributes

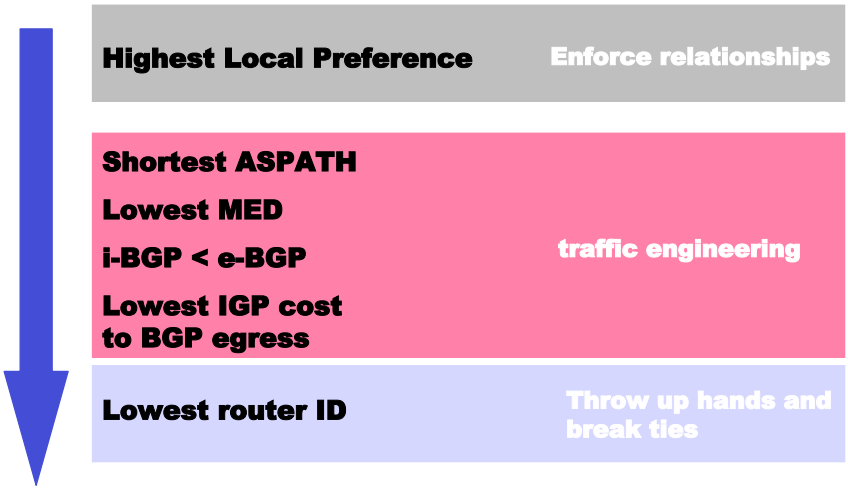
Not all attributes need to be present in every announcement

BGP Route Processing

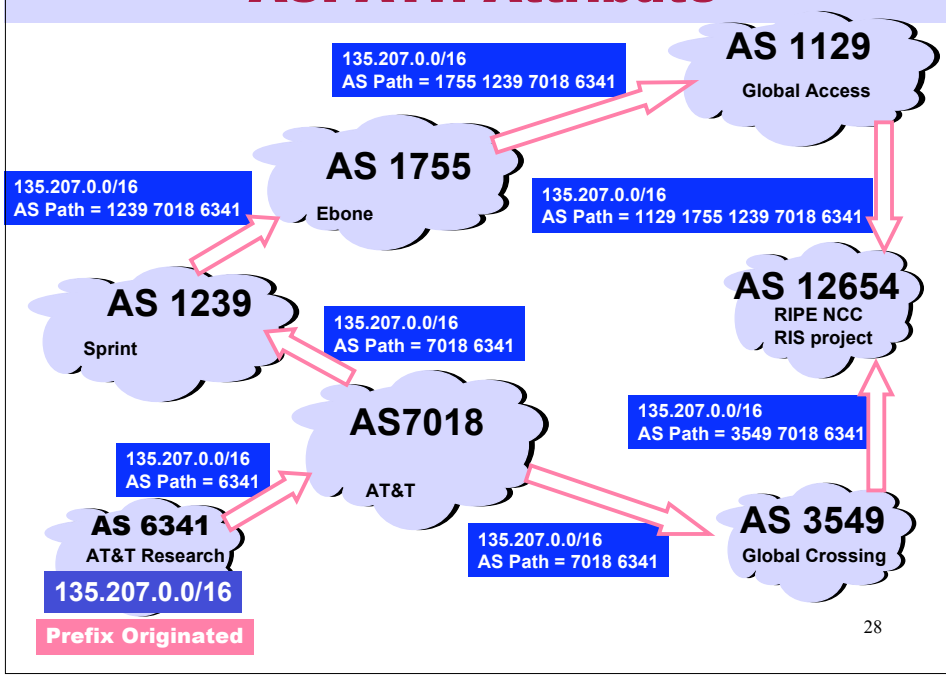


26

Route Selection Summary

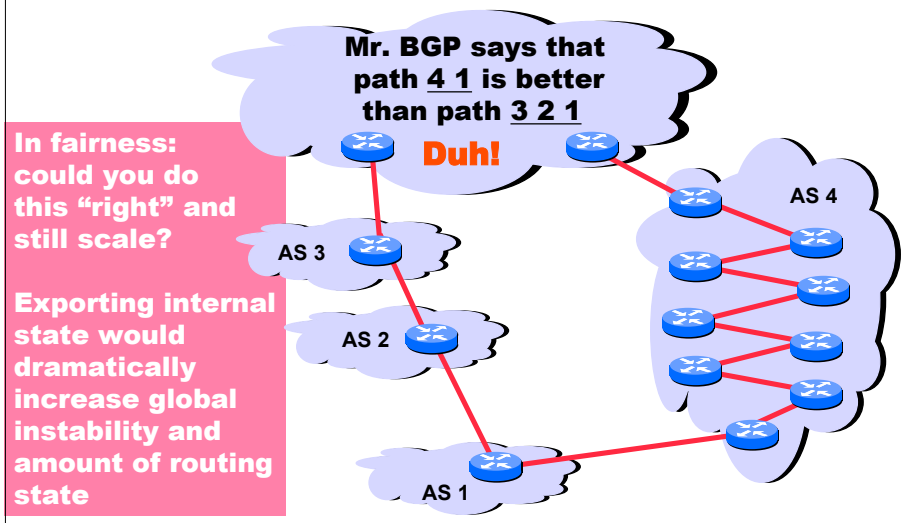


ASPATH Attribute



28

Shorter Doesn't Always Mean Shorter

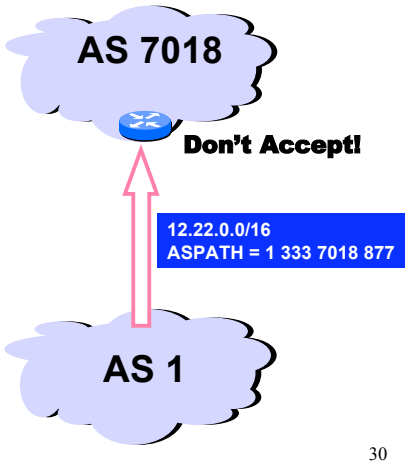


In fairness: could you do this "right" and still scale?

Exporting internal state would dramatically increase global instability and amount of routing state

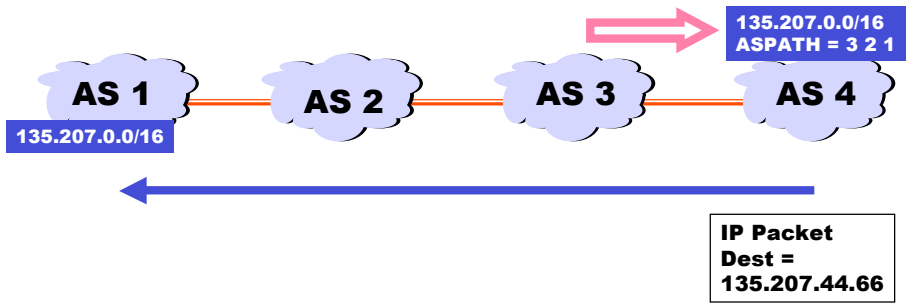
Interdomain Loop Prevention

BGP at AS YYY will never accept a route with ASPATH containing YYY.

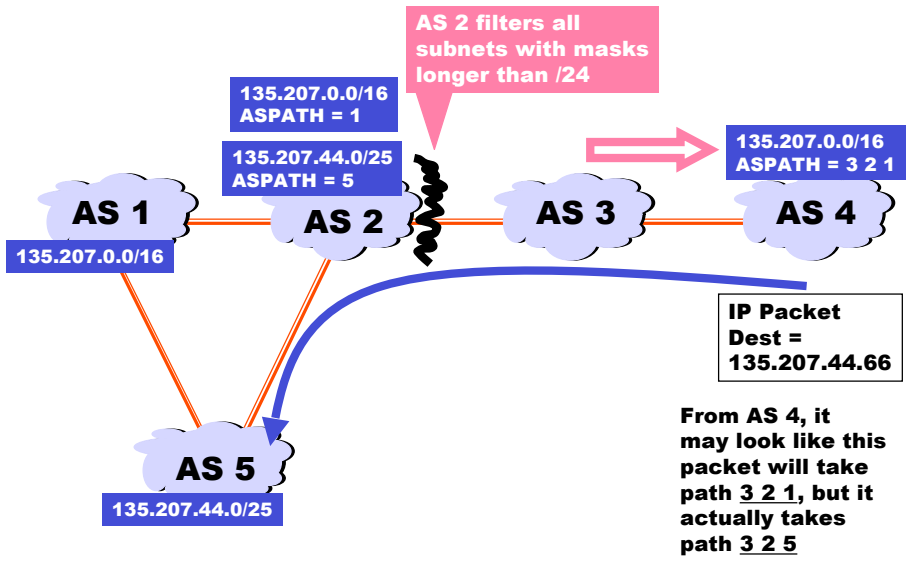


30

Traffic can follow ASPATH



... but It might not

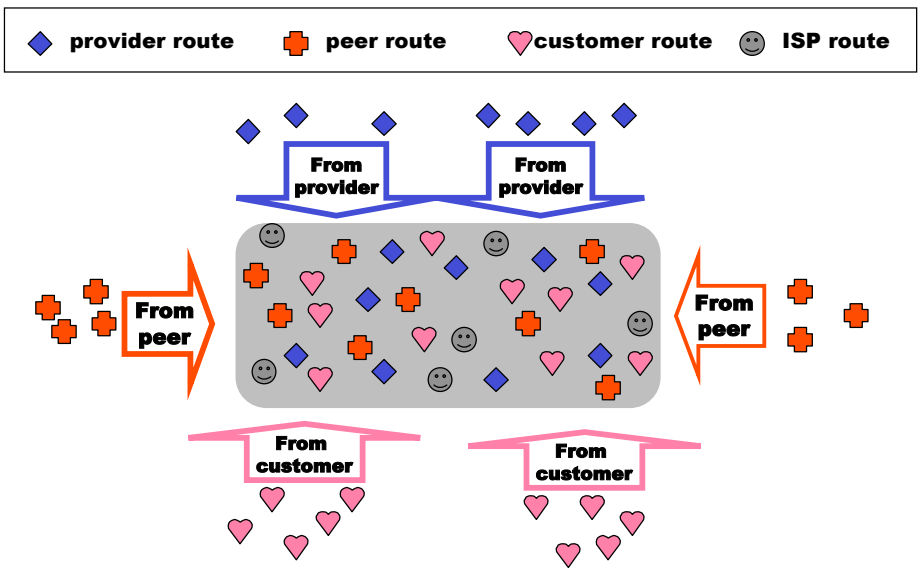


Implementing Customer/Provider and Peer/Peer relationships

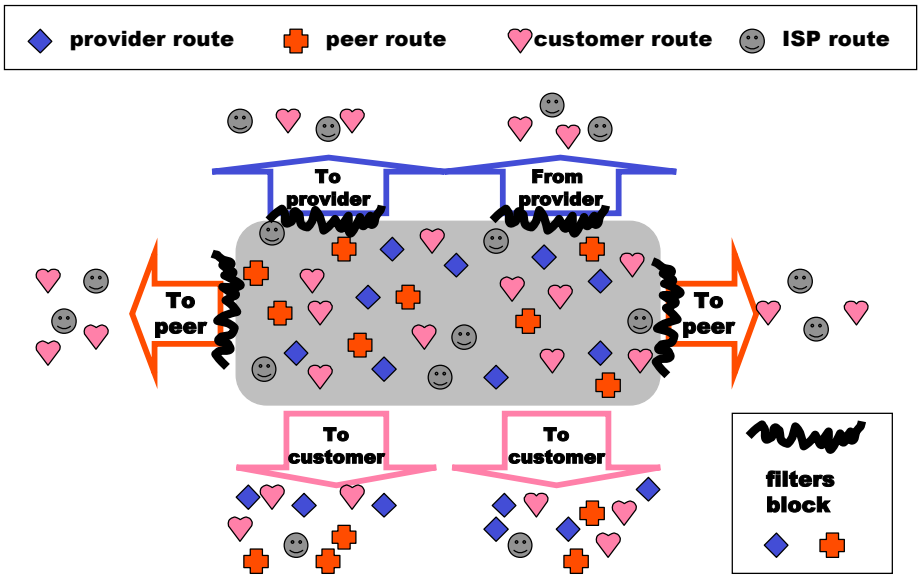
Two parts:

- Enforce transit relationships
  - Outbound route filtering
- Enforce order of route preference
  - provider < peer < customer

Import Routes

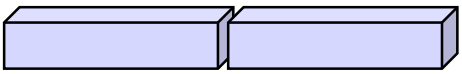


Export Routes



How Can Routes be Colored? BGP Communities!

A community value is 32 bits



By convention, first 16 bits is ASN indicating who is giving it an interpretation

community number

Used for signally within and between ASes

Very powerful BECAUSE it has no (predefined) meaning

Community Attribute = a list of community values. (So one route can belong to multiple communities)

RFC 1997 (August 1996)

Reserved communities

no_export = 0xFFFFFF01: don't export out of AS

no_advertise 0xFFFFFF02: don't pass to BGP neighbors

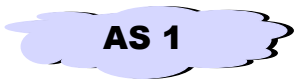
Communities Example

- 1:100
  - Customer routes
- 1:200
  - Peer routes
- 1:300
  - Provider Routes

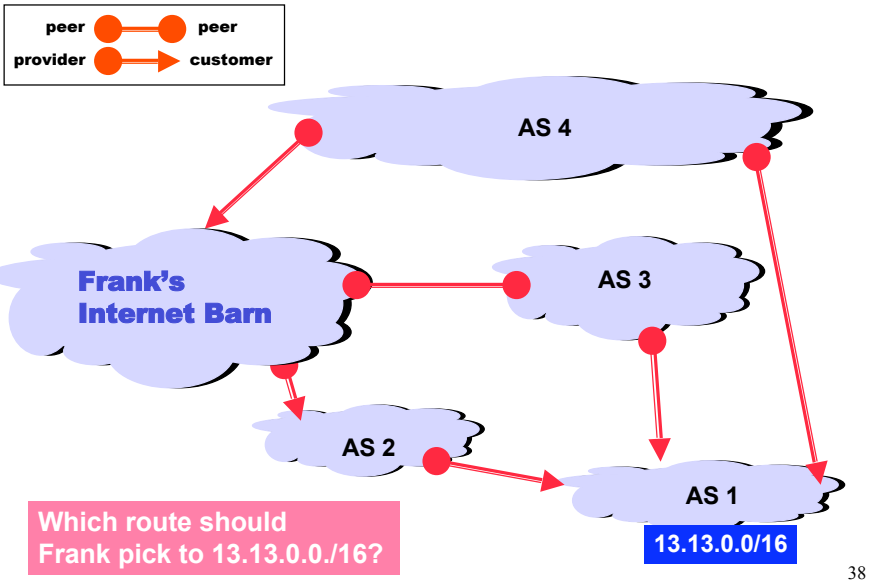
Import

- To Customers
  - 1:100, 1:200, 1:300
- To Peers
  - 1:100
- To Providers
  - 1:100

Export



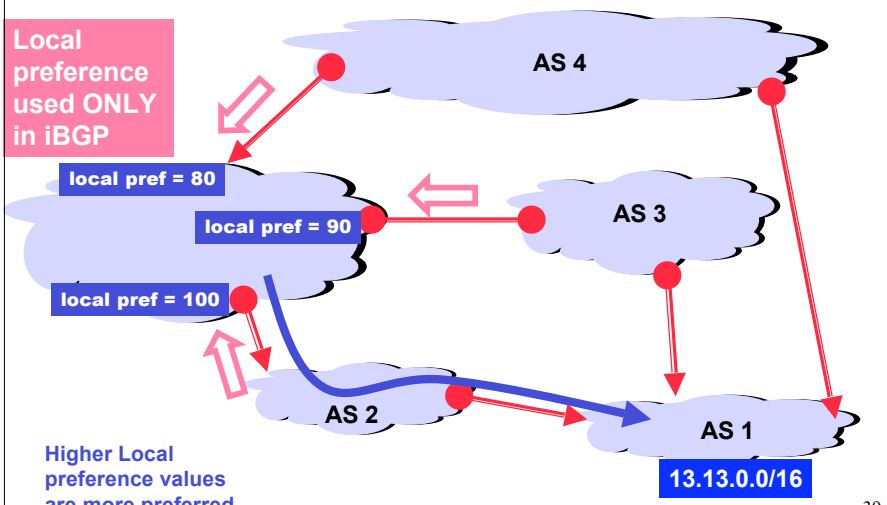
So Many Choices



Which route should Frank pick to 13.13.0.0./16?

38

LOCAL PREFERENCE



39