Interdomain Interconnections and Routing

Slides credit:

Hari Balakrishnan, Nick Feamster, Vyas Sekar, Cecilia Testart

What's this?

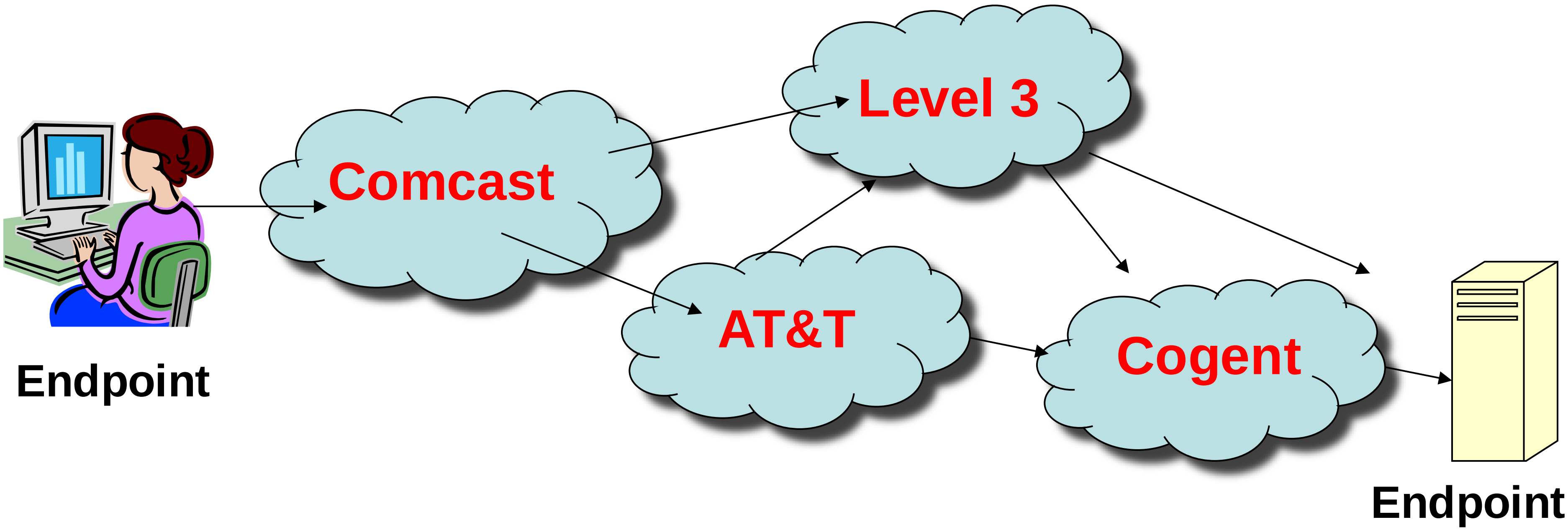
A router is a device that connects two or more networks or subnetworks [Cloudfare]

Routers guide and direct network data, using packets that contain various kinds of data—such as files, communications, and simple transmissions like web interactions [Cisco].

A router is a gateway that passes data between one or more local area [Juniper].

What’s a router?

Internet Routing



- Internet service providers with varying different sizes
- Large-scale: Thousands of autonomous networks called Autonomous Systems (AS)
- Competitive cooperation:
  - Must cooperate for global connectivity
  - Self-interest: Independent economic entities

Overview

- **The Internet is a global network of networks.**
  - Connectivity is achieved by routers and routing protocols
- ***Interconnection* refers to the various ways networks attach and exchange traffic.**
  - A collection of business practices and technical mechanisms that allow individually managed networks to connect together for this purpose

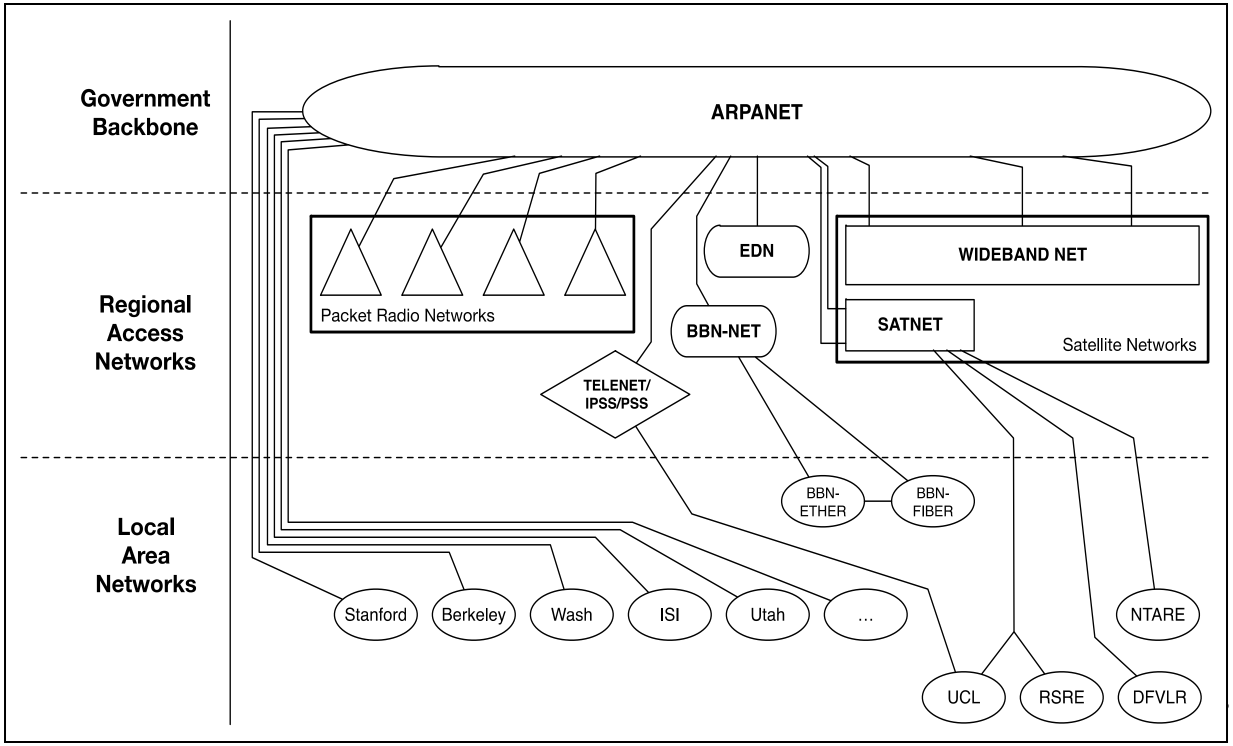
Loose Coordination

- **There is no central authority that manages Internet interconnection:**
  – System made up of the many bilateral and multilateral decisions made by various actors that interconnect
- **High bandwidth applications and changes in the number of sources of content is altering traffic growth rates on the Internet and the methods to deliver traffic**

Interconnection Pre-1995

- Initial form of the Internet in the US:
  - A single backbone network that was operated by the U.S. government.
  - Smaller, regional networks connected to this network forming a simple hierarchical structure.
  - Traffic from one part of the Internet to another was handed off to this backbone network, which carried it to the destination network.

For many years, the technical requirements on the routing protocol providing this function were simple, and there was no need to deal with business issues.



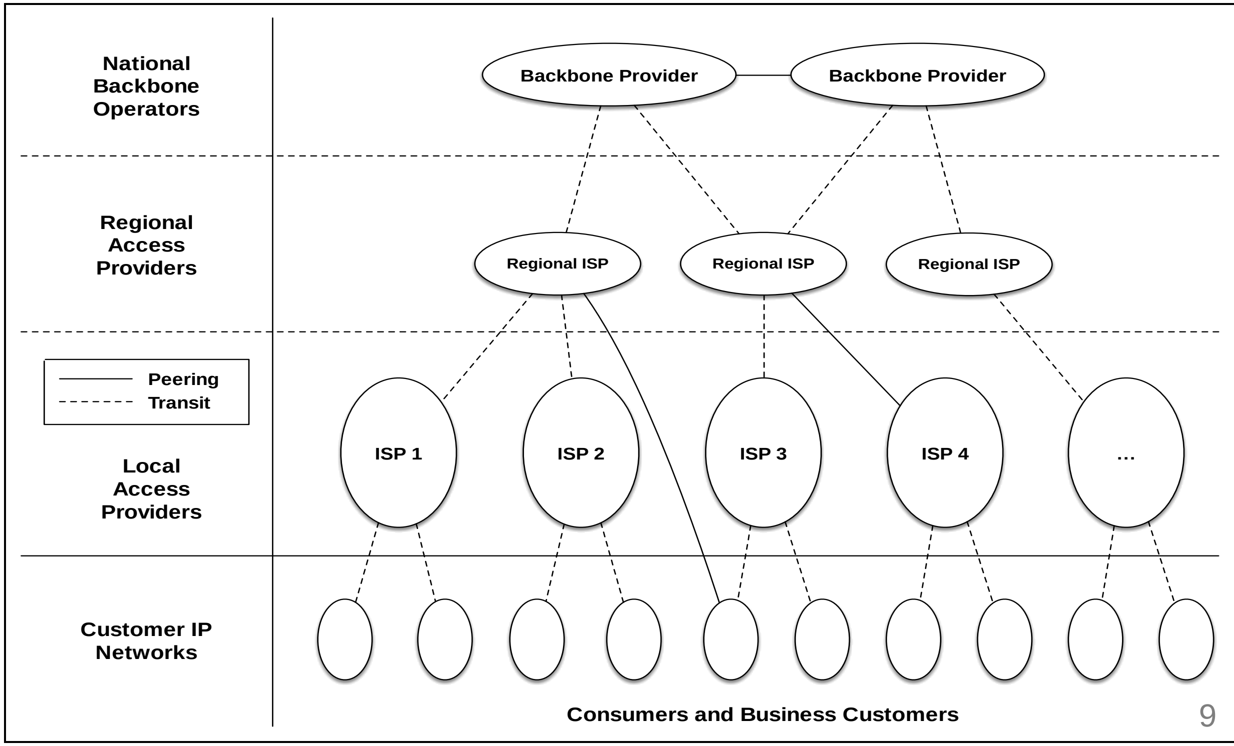
Interconnection Circa 1992

- Initiative led by Al Gore (then VP)
- Information Superhighway
- Interconnection became politicized during the 1992 Clinton–Gore election campaign
- 1990: just 313,000 computers on the Internet
- 1996: 10 million

Interconnection Circa 1995-2005

- The backbone eventually transitioned from a single government-operated backbone to a federated backbone model comprised of multiple commercial network operators.
- A new routing protocol, called Border Gateway Protocol (BGP), was created, allowing for commercial provision of backbone connectivity by multiple parties.
- BGP allowed network operators to manage this more complex and competitive space, and to express at least a limited set of business constraints on routing.

Using BGP a network operator can specify how it prefers traffic flow into and out of its network.



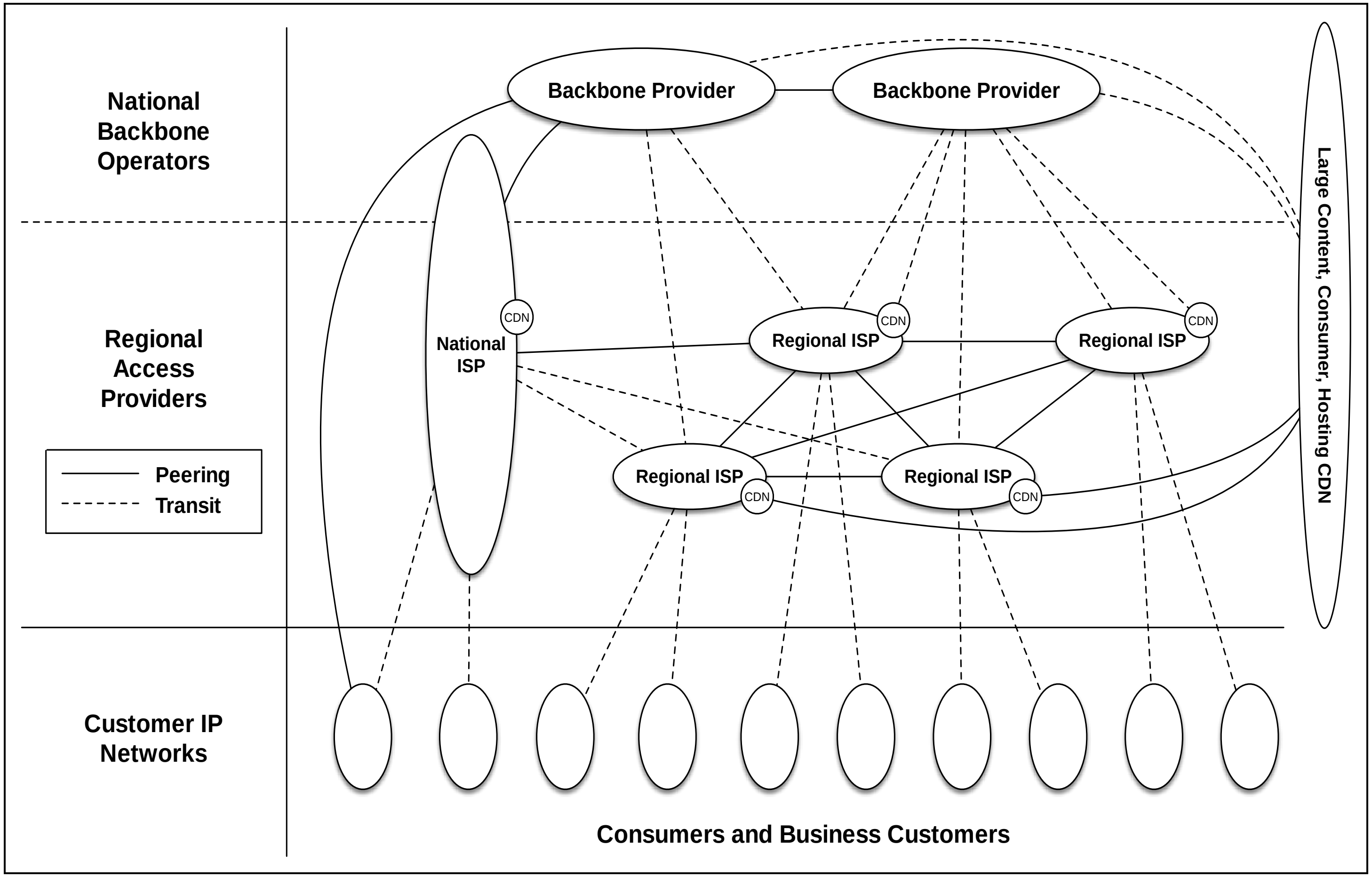
Interconnection Today:
Consolidation and flattening due to content provider networks

Flatter, highly interconnected

No single large backbone network

Rise of content distribution networks

https://www.bitag.org/documents/Interconnection-and-Traffic-Exchange-on-the-Internet.pdf



Content distribution networks

Content providers (Google, Amazon, Meta, ...)

Traffic and Interconnection

- Hyperscaler dominance manifested in how traffic flows
- Changes in Internet traffic patterns have coincided with a dramatic change in the Internet connectivity model
- In 2009 half of all Internet traffic from approx. 150 companies
- In 2014, only 30 companies account for half of all traffic
- In 2022, 65% of all internet traffic came from six companies (Facebook, Amazon, Google, Apple, Netflix, and Microsoft)

What is a routing prefix?

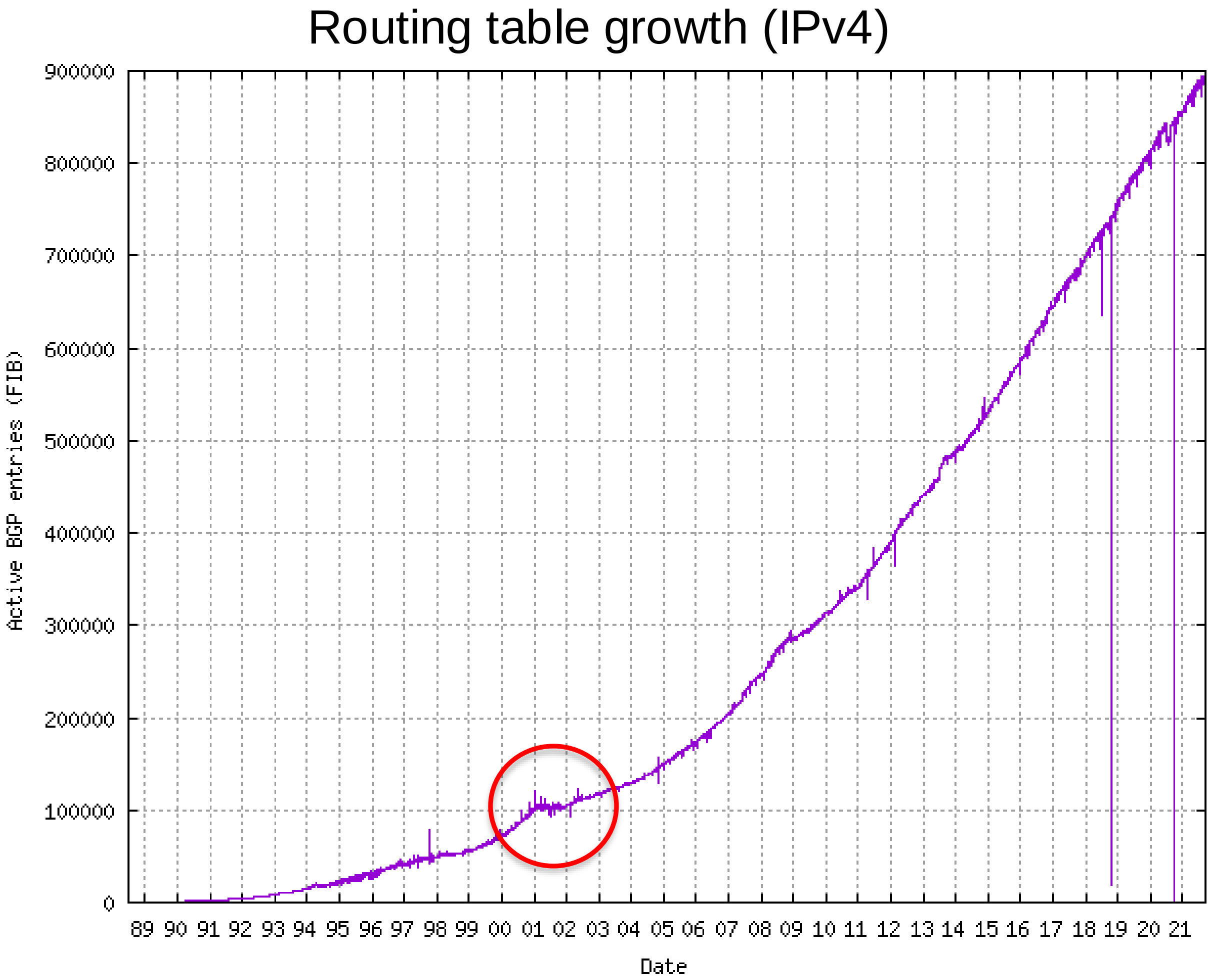
- Destinations on the Internet are aggregated together into routing prefixes
- 18.31.0.82 is actually the 32 bit string
  00010010 00111110 00000000 01010010
- Routers have forwarding table entries corresponding to an address *prefix* (a range of addresses with common prefix bitstring)
- 18.0.0.0/8 stands for all IP addresses in the range 00010010 00…0 to 00010010 11…1 (i.e., 2^{24} addresses of the form 00010010*)

What is a routing prefix?

- 18.31.0.0/17 stands for a range of 2^{15} consecutive IP addresses of the form 00010010001111100* (1^{st} 17 bits are the same for each address in that range)
- *subnetworks* may be of size 1, 2, 4, 8, … (maxing out at 2^{24} usually), and may be recursively divided further
- Forwarding uses *longest prefix match*
- At each router, routes are of the form “For this range of addresses, use this route”
- Pick the route that has the longest matching prefix with destination address

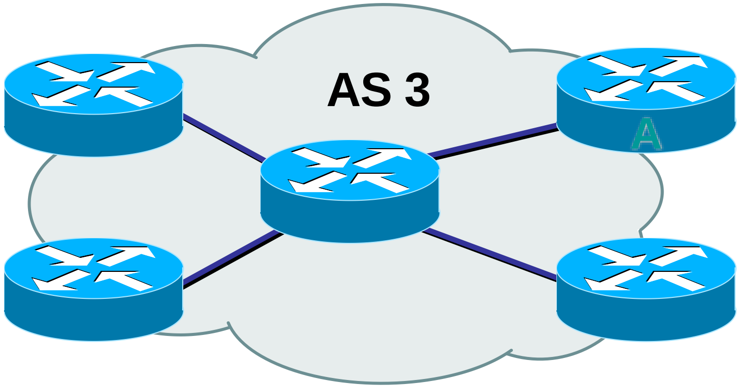
How has the Internet grown over time?

- Destinations on the Internet are aggregated together into routing prefixes
- A routing table contains routing information about the prefixes



https://www.cidr-report.org/

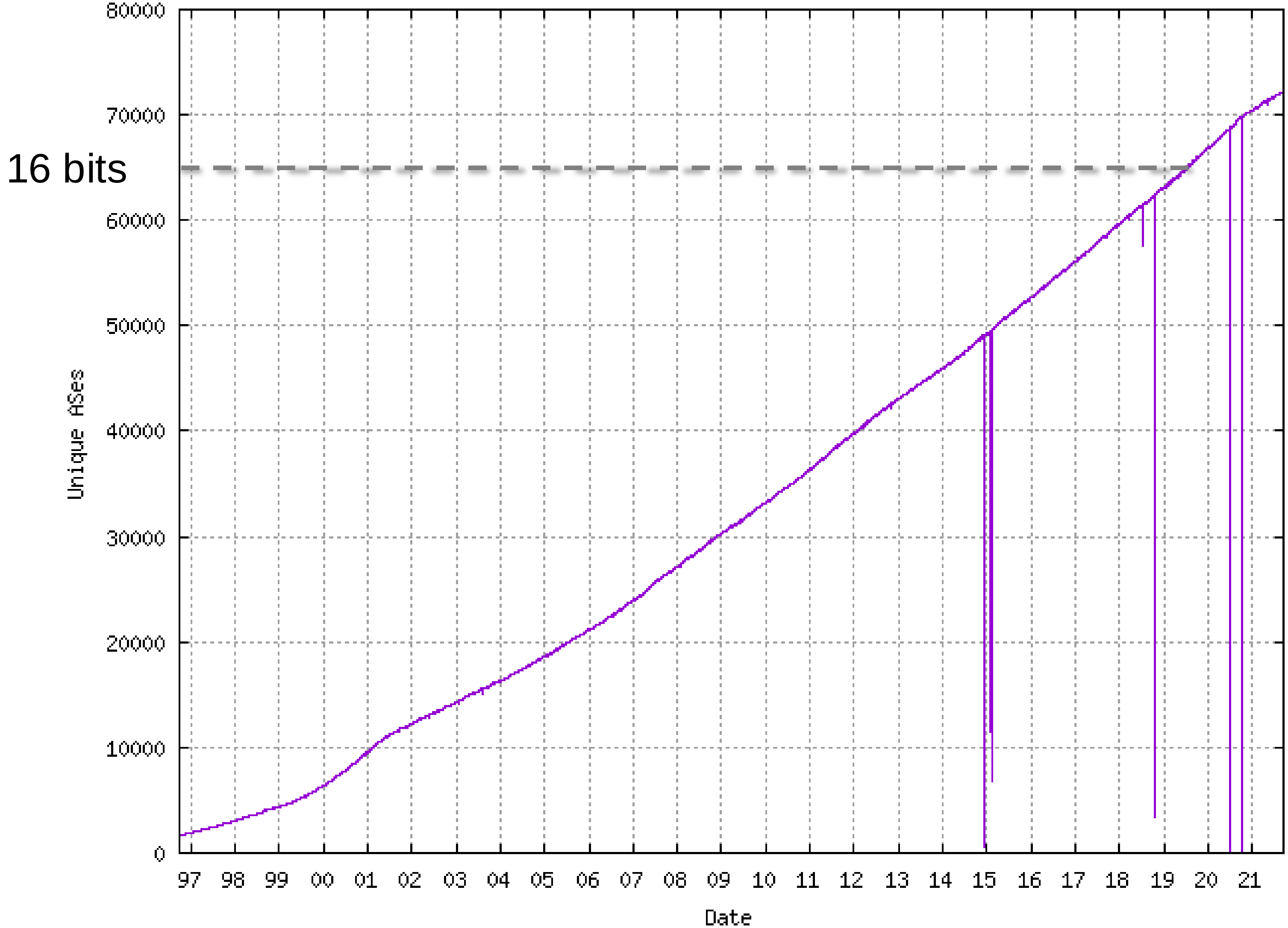
Autonomous Systems (ASes)



- Collection of networks with same policy
- Single routing protocol
- Usually under single administrative control
- Have a unique ASN (used to be 16 bits but it’s now 32 bits)

AS Growth with Time

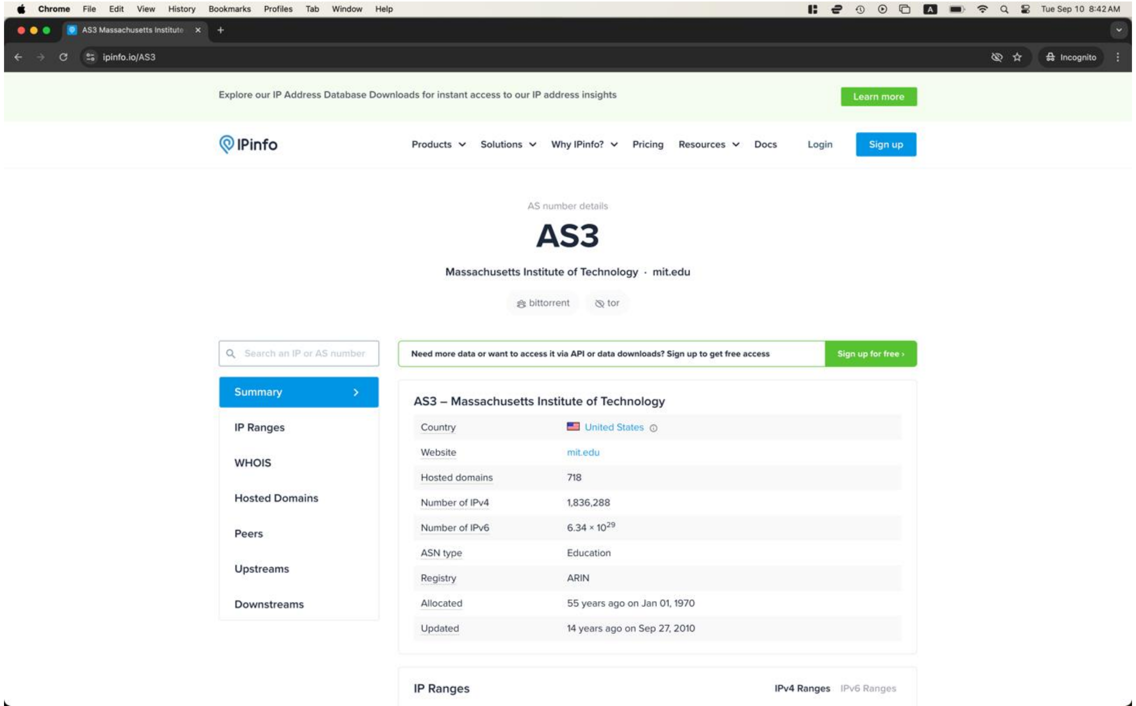
https://www.cidr-report.org/



Autonomous Systems (ASes)

- Examples:
  - MIT: 3, CMU: 9
  - AT&T: 7018, 6341, 5074, …
  - UUNET: 701, 702, 284, 12199, …
  - Sprint: 1239, 1240, 6211, 6242, …
- How do ASes interconnect to provide global connectivity?
- How does routing information get exchanged?

https://ipinfo.io/AS3

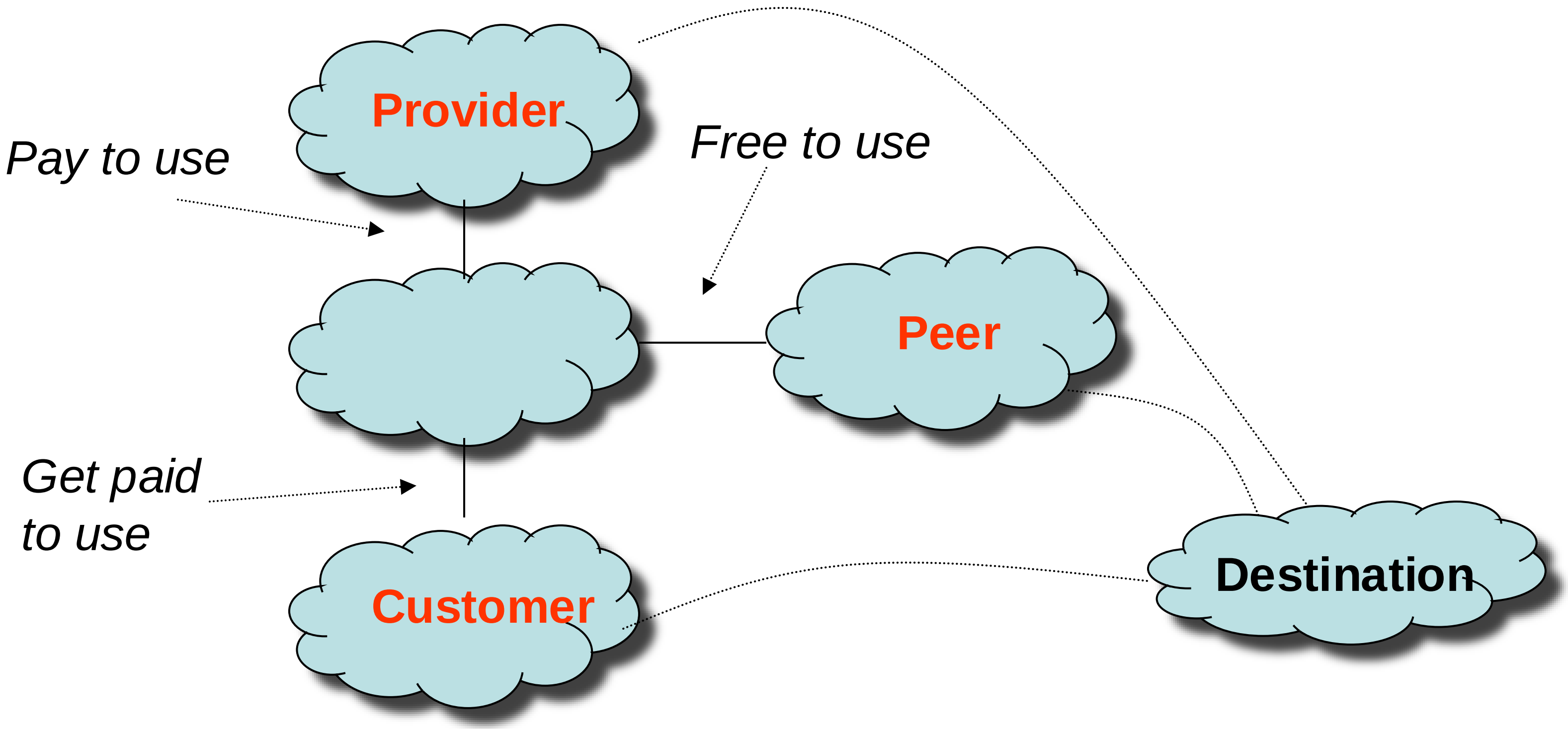


Rest of today

- Understanding the Internet business model
  - Internet routing is only *incidentally* about forwarding packets
  - It's mostly about money
  - Almost everything in BGP can be justified with *incentives* and *economics*
  - Think: Which way is the money flowing?

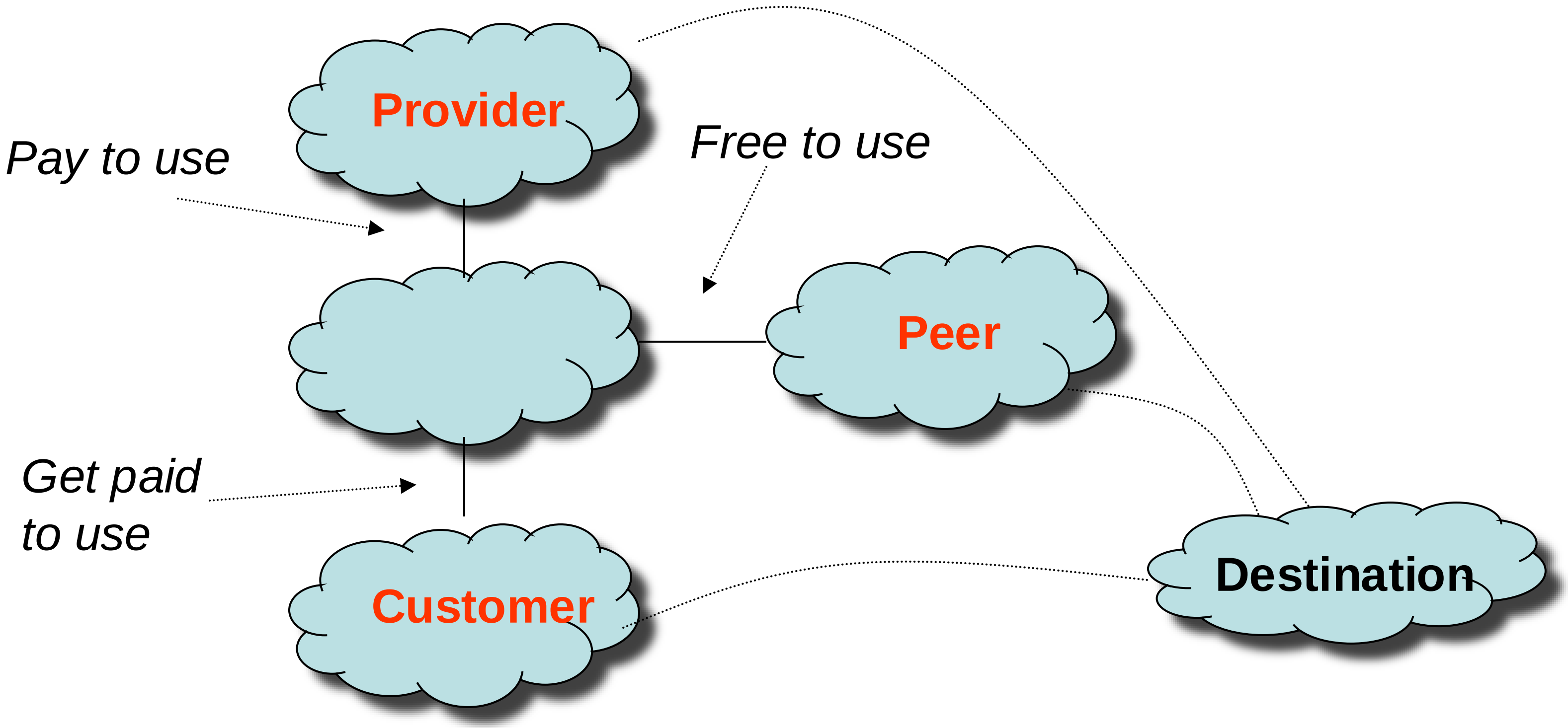
- Details of routing table entries

Internet Business Model (Simplified)



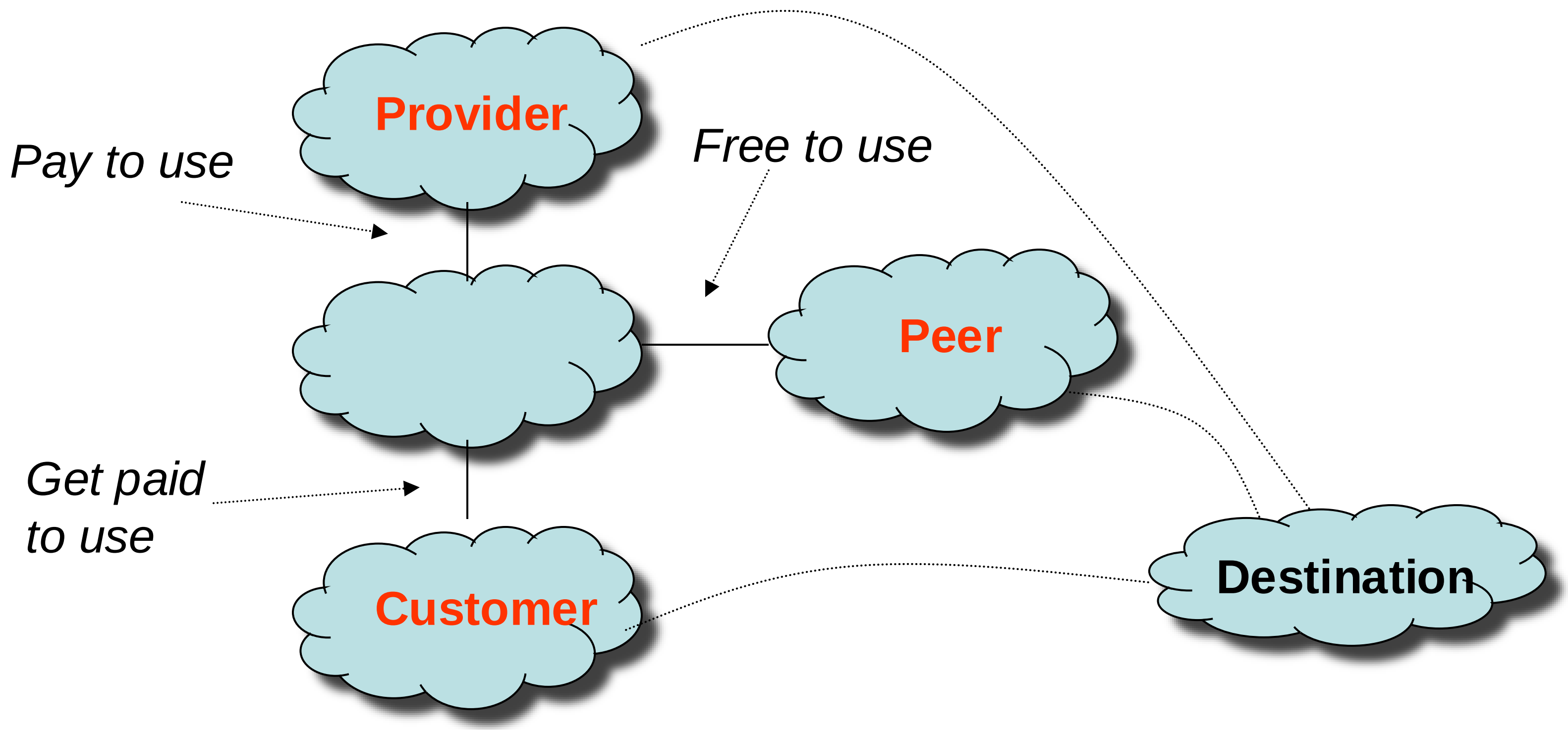
- Money flows from customer to provider
- What does the customer get in return?
- The customer gets access to routes.
- The provider gives the customer the visibility into their routing tables in exchange for money.
- It's in the provider's interest to give the customer their entire routing table. Because the provider can send/receive more traffic through customer

Internet Business Model (Simplified)



- Why do we have peering then?

Internet Business Model (Simplified)



- Why do we have peering then?
- Because there is no single uber provider at the top of the tree.
- At least at the top of the hierarchy, we need a few providers who peer with each other and they don’t pay money to each other.
- Settlement-free peering: Two ASes exchange routers with one another.

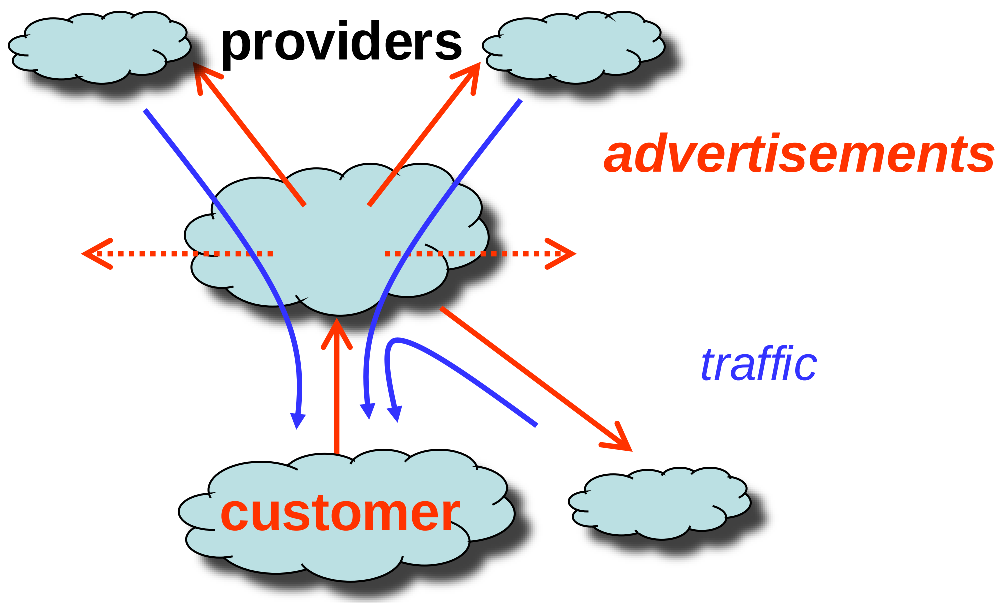
Transit

- Just a technical word to reflect the customer/provider relationship
- If an AS provides “transit” service for a customer AS, it can carry traffic between that customer’s network and all other Internet endpoints.
- Transit relationships may be:
  - “***full***” – the customer receives routes for all Internet destinations from its transit provider), or
  - “***partial***” – the customer receives routes for some subset of all Internet endpoints.
- Transit is usually thought of as a service offered for a fee.

Implementing Transit

Filtering

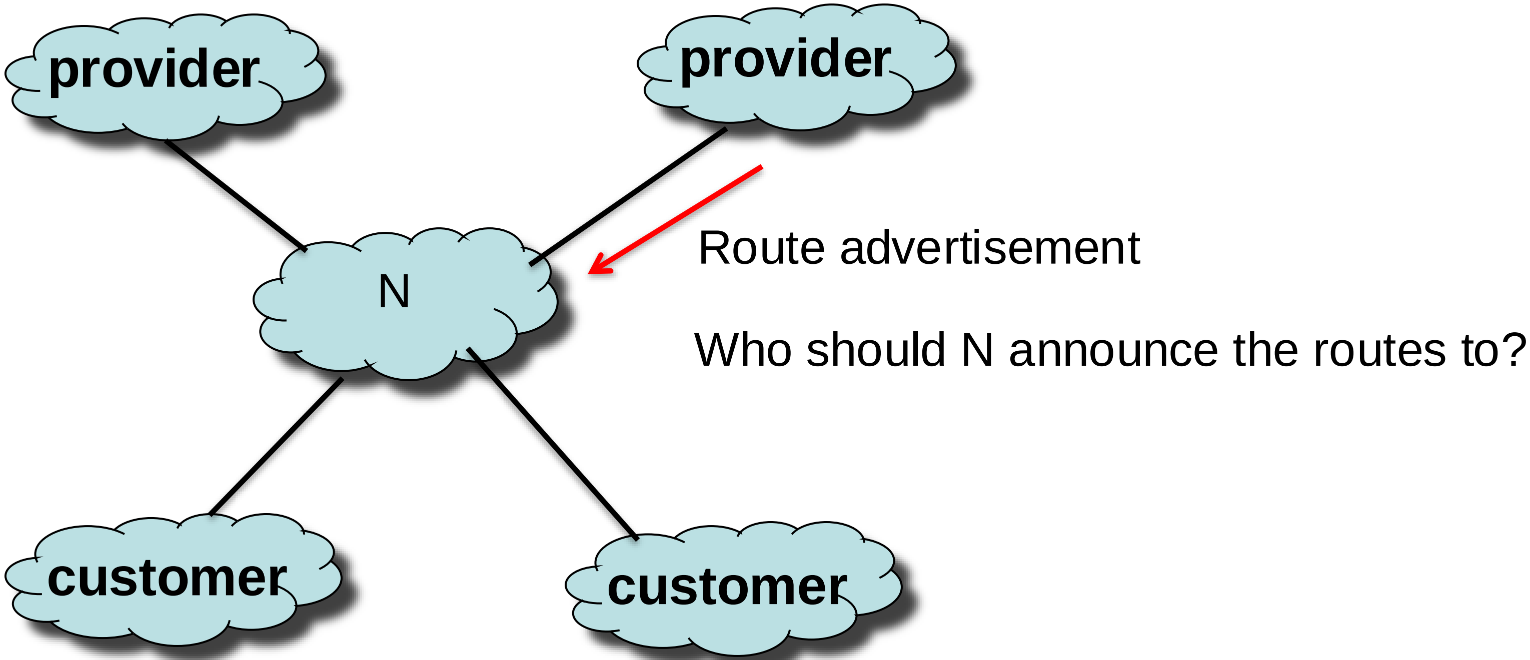
- Routes from customer:
  - Announce to *everyone*
  - Because it will increase the likelihood of using your customer ($$)



Implementing Transit

Filtering

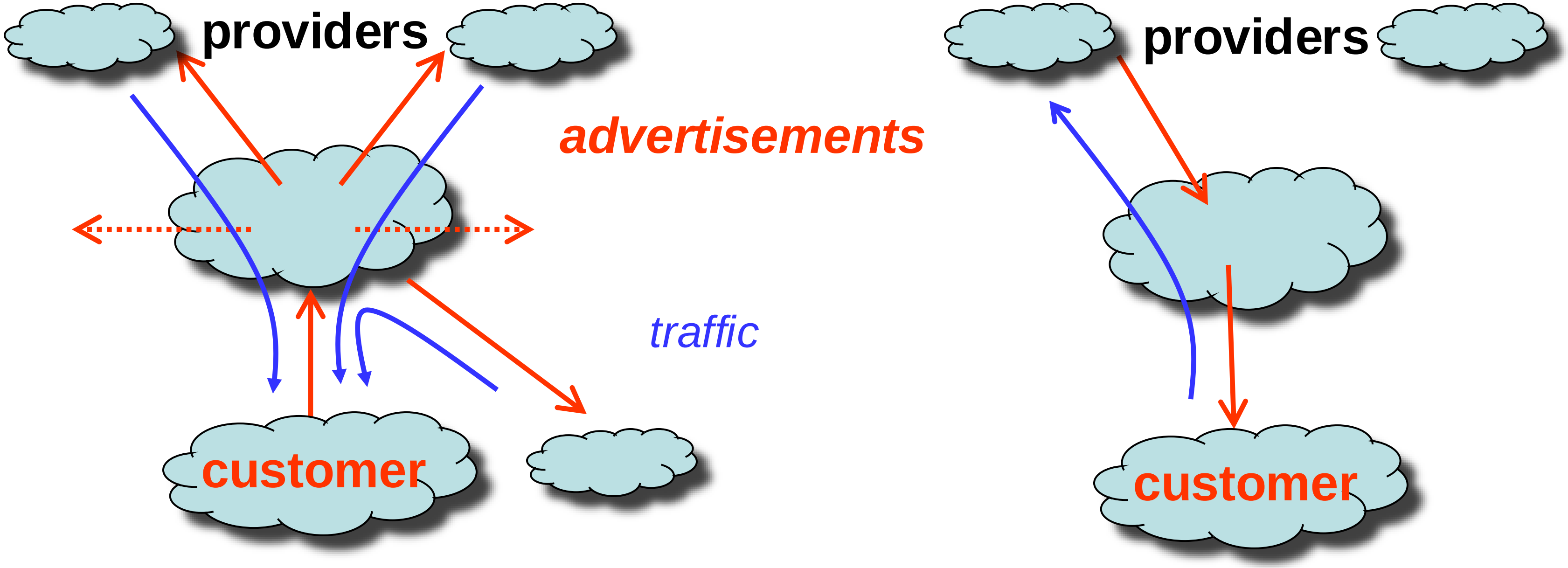
- Routes from customer:
  - Announce to *everyone*
  - Because it will increase the likelihood of using your customer ($$)
- Routes from provider:
  - only to *customers*
  - Because otherwise N will not make money from it, no incentive to announce it to anyone beside its customers



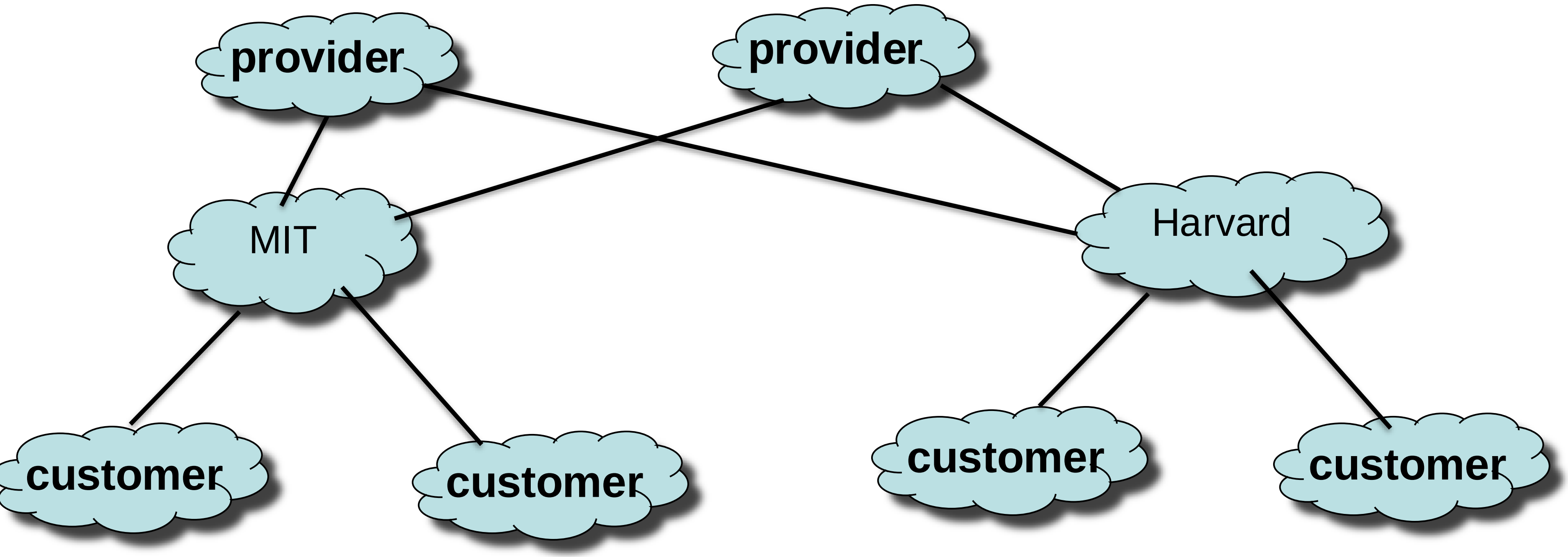
Implementing Transit

Filtering

- Routes from customer: announce to *everyone*
- Routes from provider: only to *customers*



Peering incentives



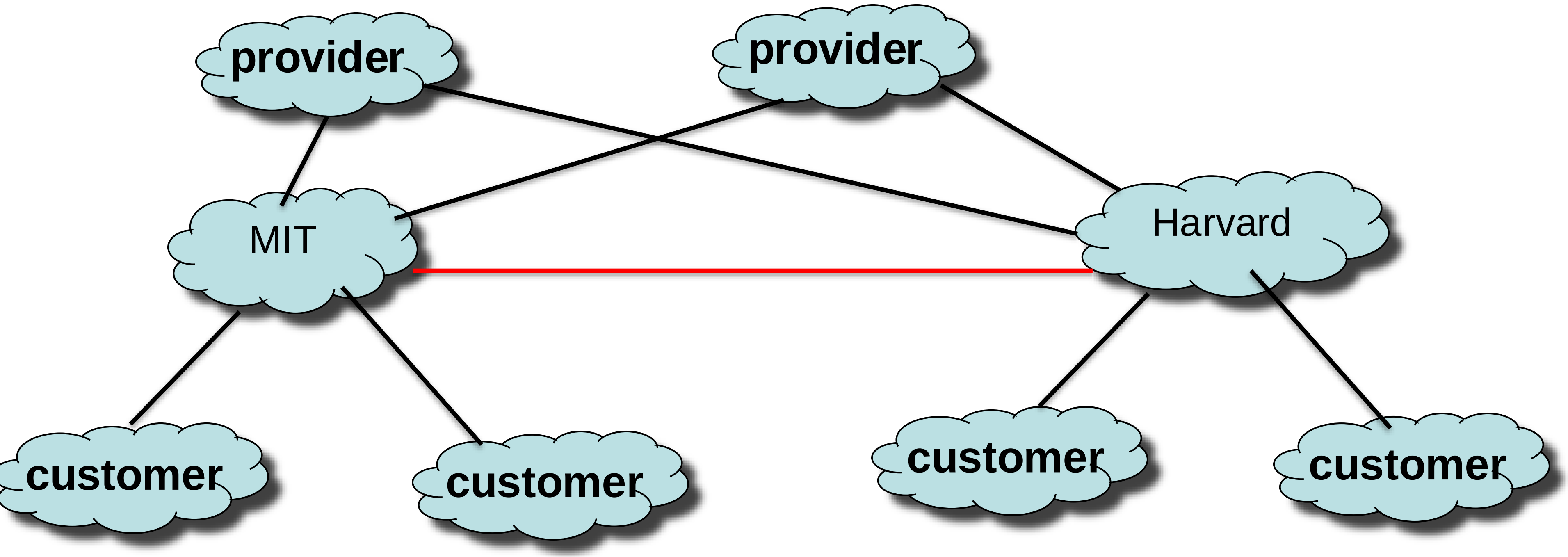
MIT is paying the providers

Harvard is also paying the providers

- Research collaboration between the university with massive amount of traffic

What’s a better approach?

Peering



- Peering tends to happen when there is a lot of traffic exchange between two ASes, and the traffic is somewhat symmetric
- Does it make sense for MIT to peer with YouTube?

Peering

- If an AS "peers" with another AS, the two ASes agree to exchange traffic only between their own endpoints and the endpoints in their customers' networks.
  - This agreement can be formal or informal.
  - Where a peering agreement is formalized, it will usually include confidentiality and non-disclosure terms

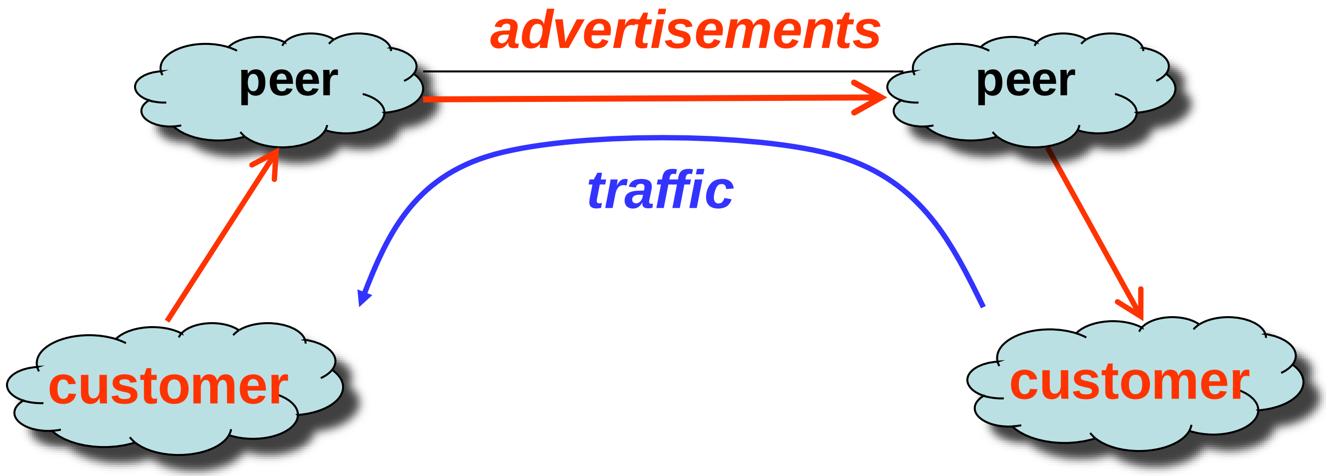
- Peering agreements have historically been informal "hand shake" agreements but appear to increasingly involve contractual relationships.

- In a transit relationship, a transit provider will typically advertise all of its routes to a customer, whereas in a peering relationship, a peer will only advertise its customer routes and its own routes to another peer.

Implementing Peering

Filtering

– Announce your own prefixes in your network and routes from customers to your peers
– Routes from peer: only to customers



Policy with BGP

- BGP provides capability for enforcing various policies
- Policies are **<u>not</u>** part of BGP: they are provided to BGP as configuration information
- BGP enforces policies by choosing paths from multiple alternatives and controlling advertisement to other ASes

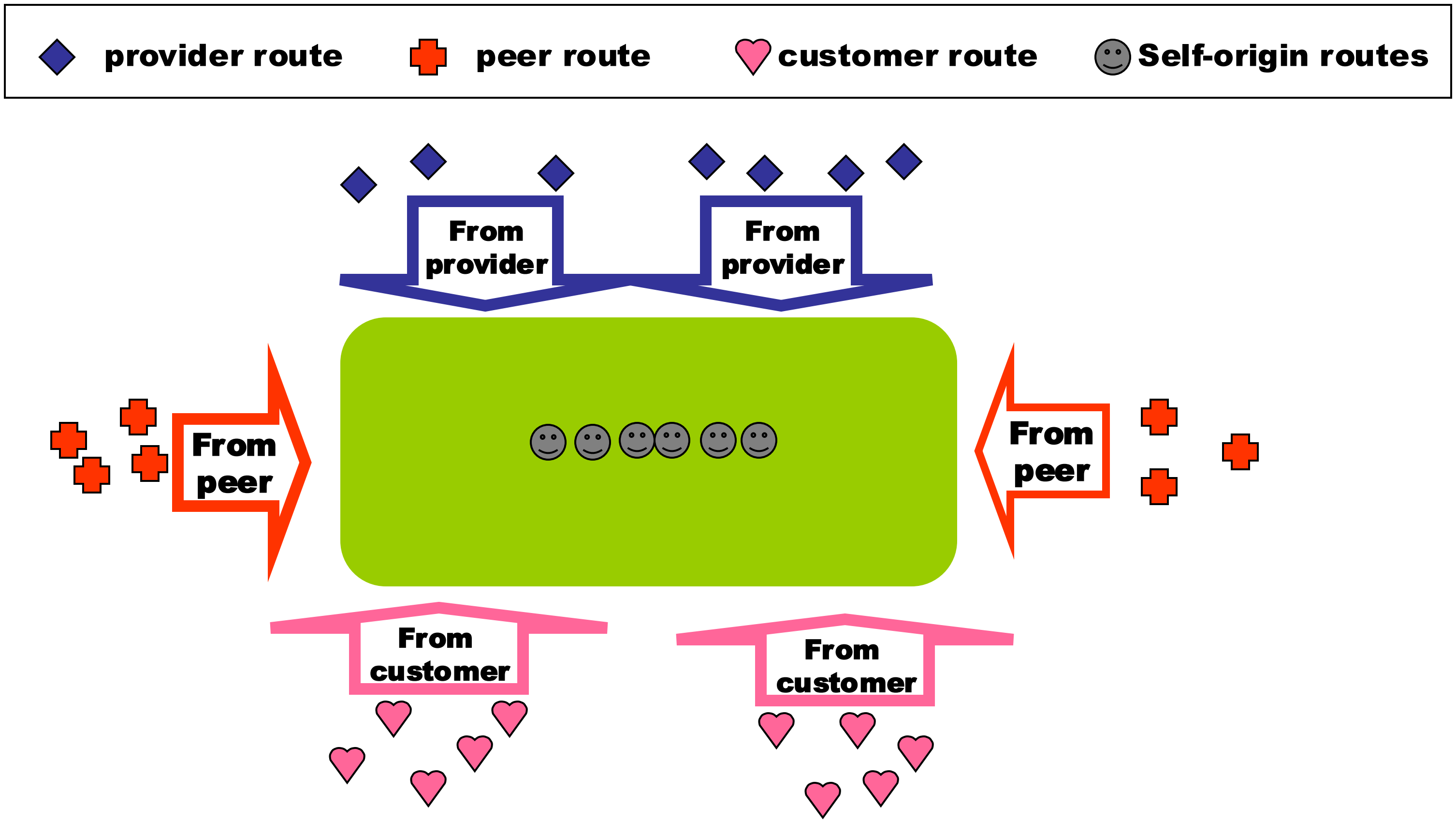
Import and Export policies

- Import policy
  - What to do with routes learned from neighbors?
  - Select best path
- Export policy
  - What routes to announce to neighbors?
  - Depends on relationship with neighbor

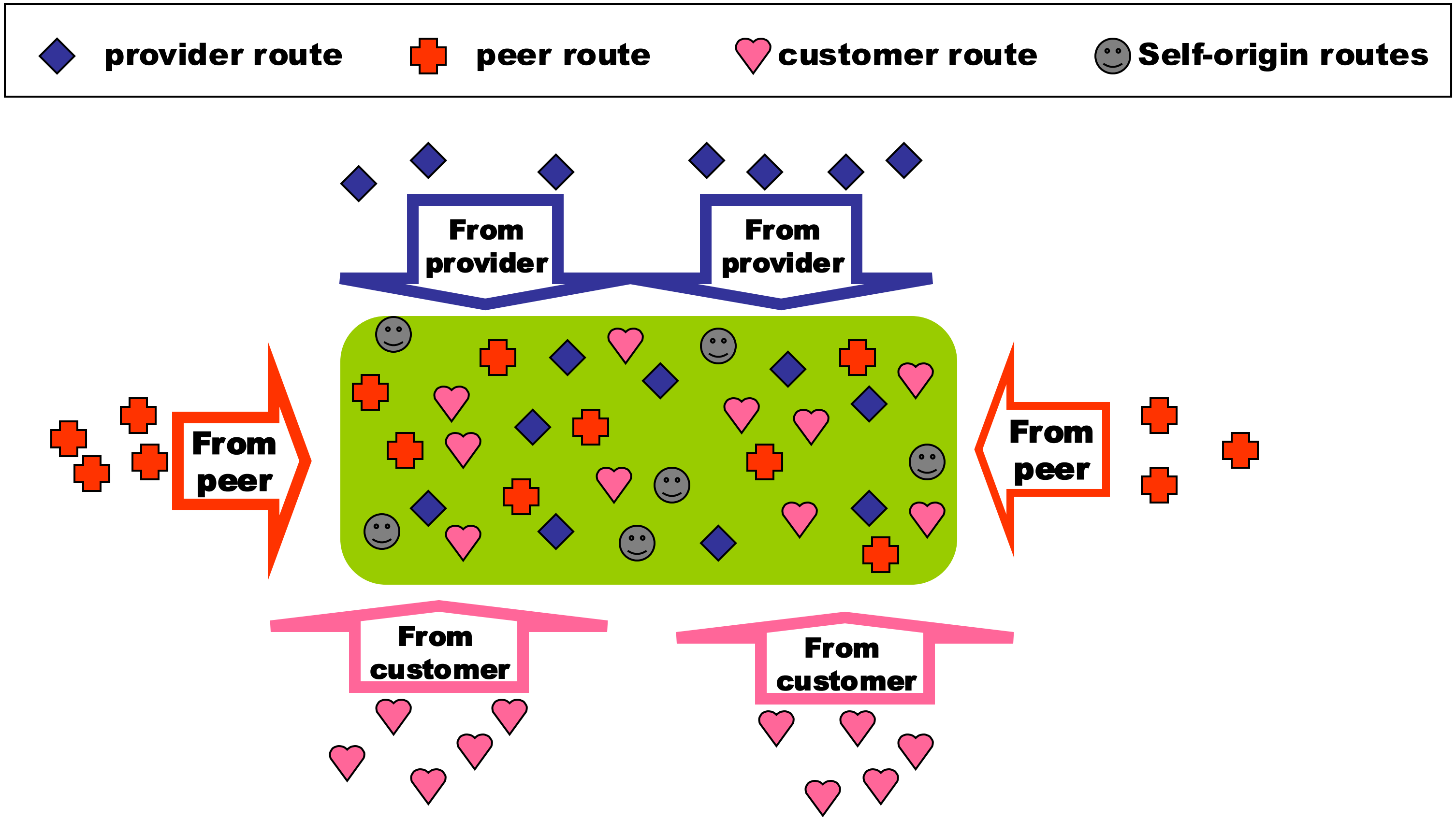
Export Policy

- To Customers
  – Announce all routes learned from peers, providers and customers, and self-origin routes
- To Providers
  – Announce routes learned from customers and self-origin routes
- To Peers
  – Announce routes learned from customers and self-origin routes

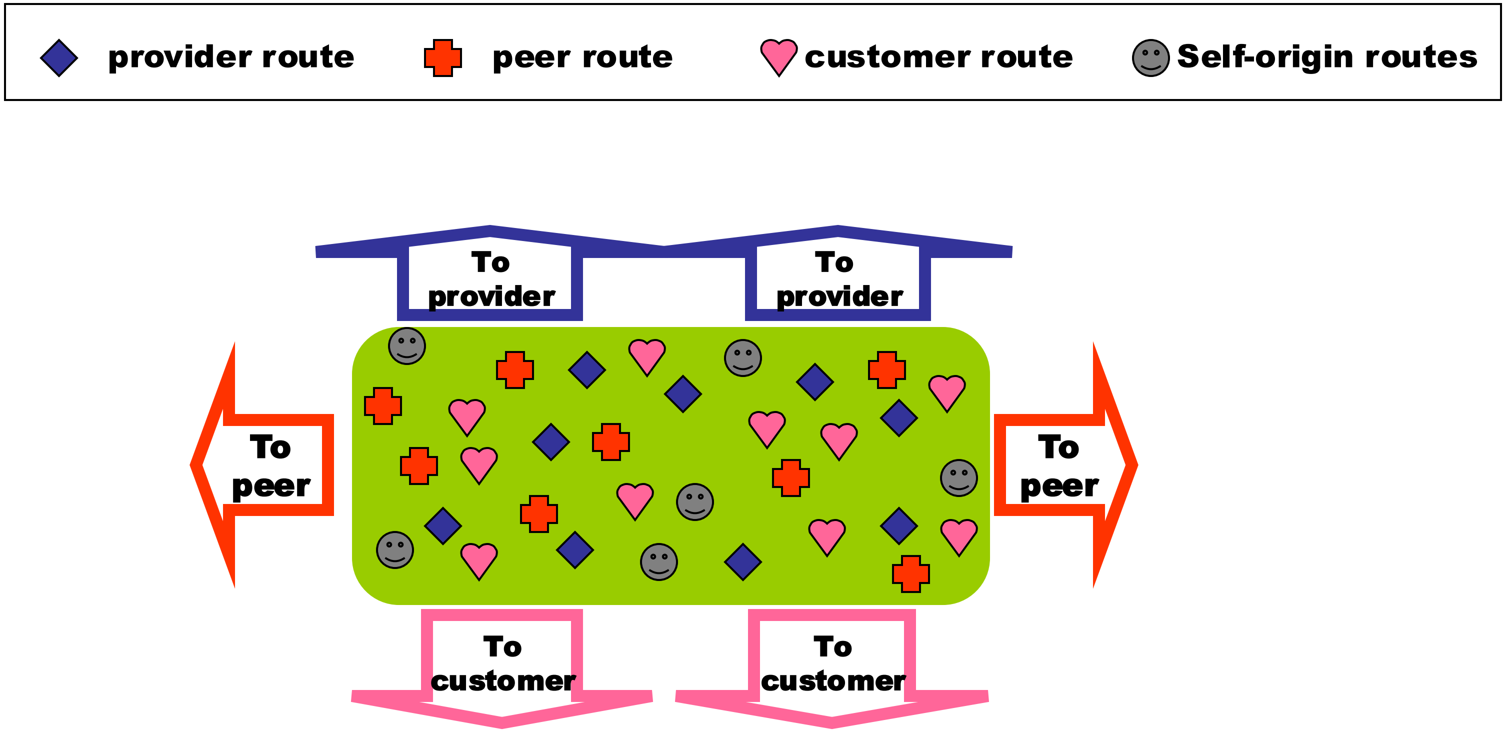
Import Routes (routes learned from neighbors)



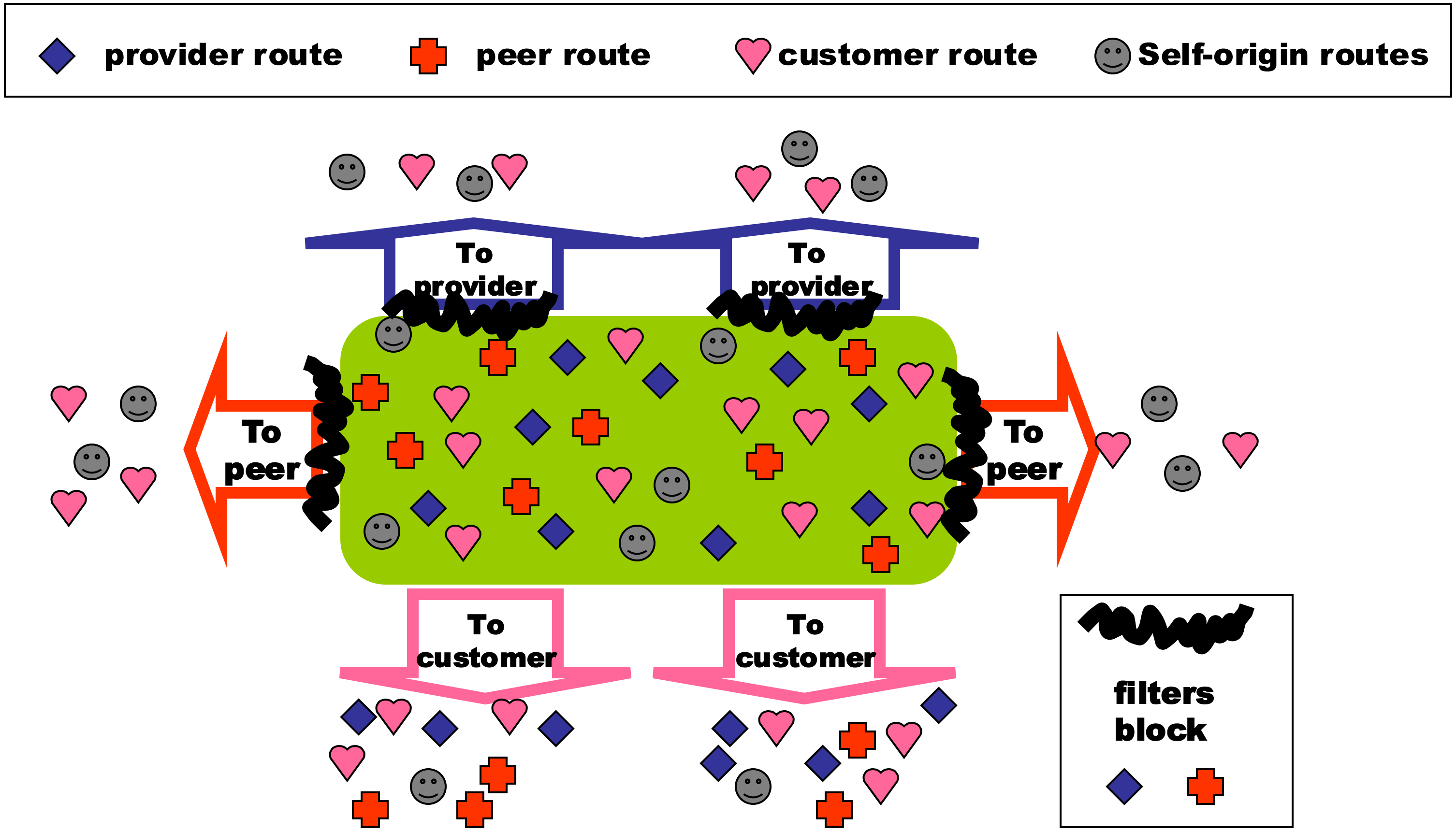
Import Routes (routes learned from neighbors)



Export Routes: routes to announce to neighbors



Export Routes: routes to announce to neighbors



Physical Facilities for Interconnection

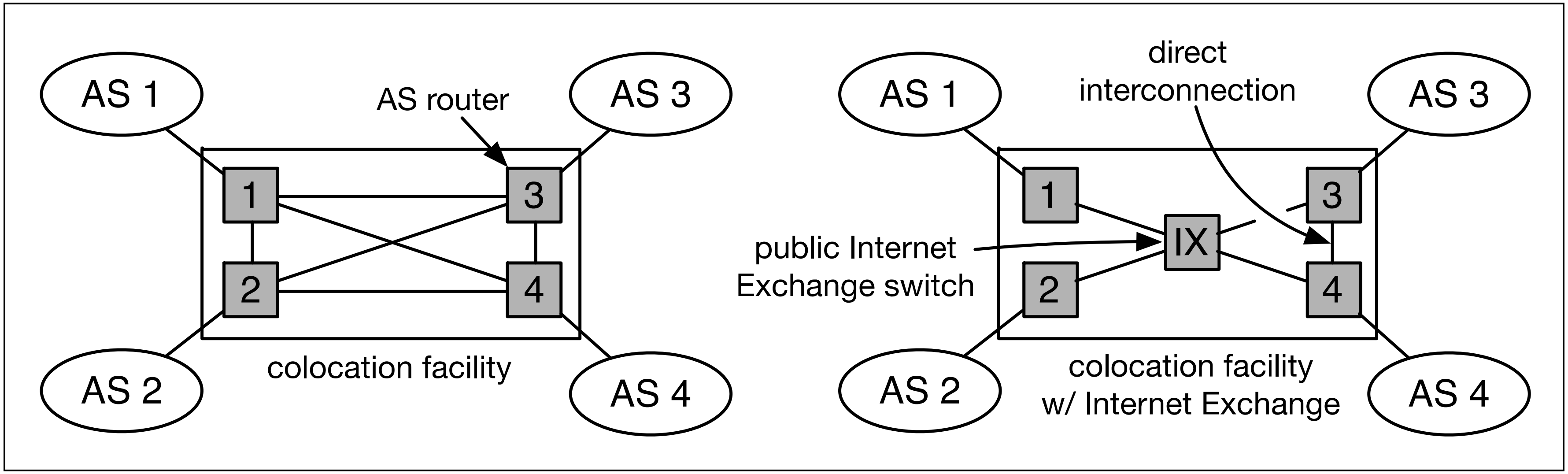
- **For networks to interconnect, they have to *physically* connect their networking equipment with each other.**
  - This requires the networks to meet in a common location, in facilities capable of supporting the equipment required for interconnection.
  - These colocation facilities lease their customers secure space to locate and operate equipment
- **Point of Presence (PoP)**
  - An access point to a communication provider's network.

Interconnection: Public & Private

- **Interconnecting two networks requires both:**
  - (1) physical connectivity, *<u>and</u>*
  - (2) network connectivity.

- **Common options for interconnection are either:**
  - *Direct interconnection:*
    - Private bilateral arrangement between two networks using a dedicated physical connection
  - *Public connection:*
    - A multilateral arrangement where all networks connect into a public Internet Exchange switch.

Public and Private Interconnection



- **At left**: Simple colocation facility with direct interconnects
- **At right**: colocation facility that also offers IX through a public switch (or “switching fabric”)

HYPERSCALERS OPERATIONS & MANAGEMENT

PAIX: A Key Hub, From Alta Vista To Facebook

The Palo Alto Internet Exchange facility was the first major carrier-neutral Internet exchange point, and continues to be a key part of Switch and Data's PAIX interconnection business.

Rich Miller
February 11, 2009

3 Min Read

paix-rows

A look at the battery room in the Switch and Data PAIX data center in Palo Alto, Calif.

Inside Switch Data's Palo Alto data center.

paix-gear

Equinix Palo Alto

Platinum Sponsor

Peers	Connections	Open Peers	Total Speed	% with IPv6
79	89	38	3.1T	87

EXPORT

Organization	Equinix, Inc.
Also Known As	
Long Name	Equinix Internet Exchange Palo Alto
City	Palo Alto
Country	US
Continental Region	North America
Service Level	24/7 Support
Terms	Recurring Fees
Last Updated	2024-07-26T15:38:10Z
Notes ?	Formerly PAIX
	EQUINIX

Contact Information

Company Website	https://ix.equinix.com
Traffic Stats Website	https://ix.equinix.com/home/locations-and-traffic/#traffic
Technical Email	support@equinix.com
Technical Phone ?	
Policy Email	support@equinix.com
Policy Phone ?	
Sales Email	
Sales Phone ?	
Health Check	

LAN

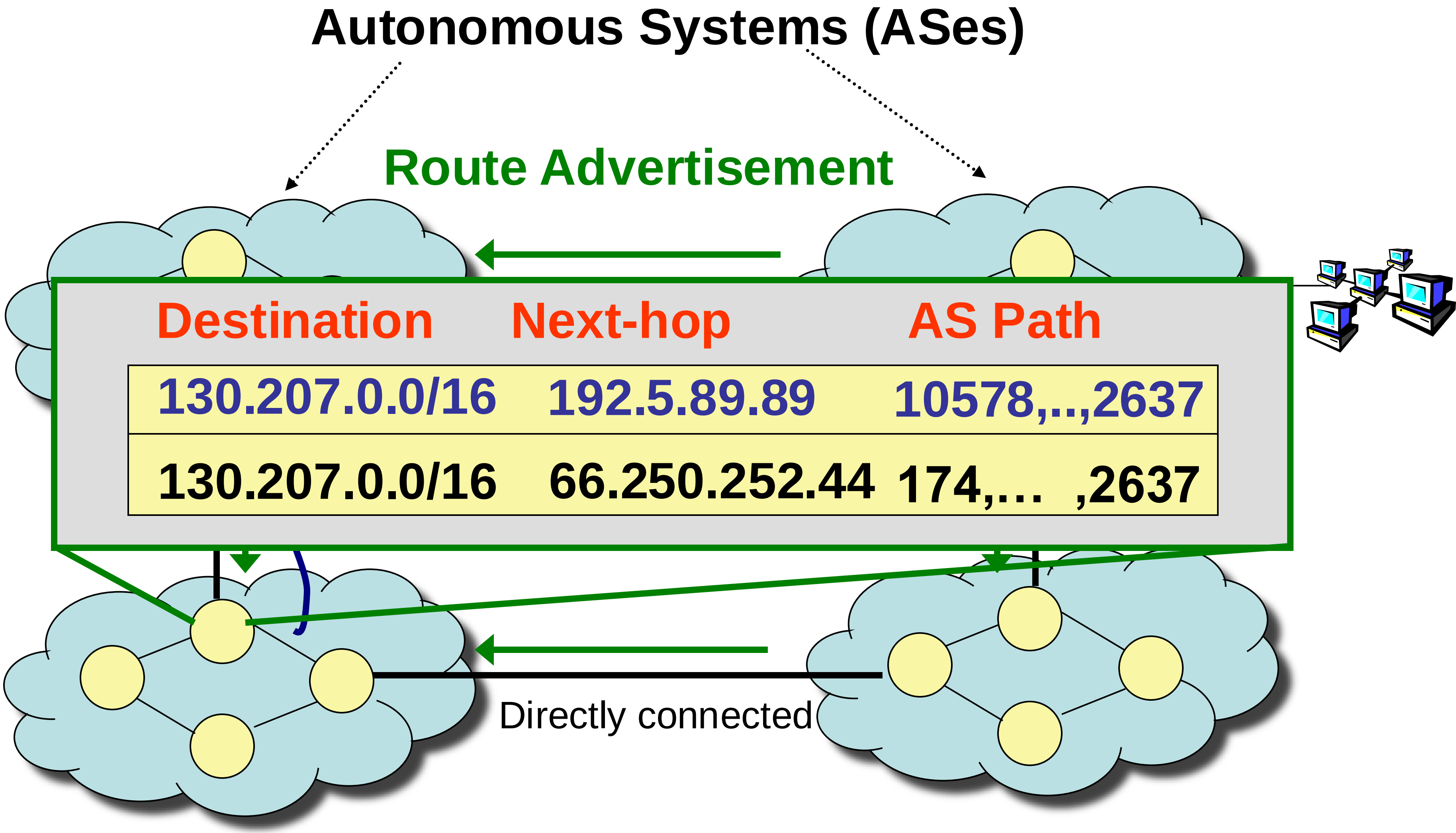
Peers at this Exchange Point

Filter

Peer Name / IPv4	ASN / IPv6	Speed / Port Location	Policy ?
365 Data Centers (BroadbandONE) 198.32.176.164	19151 2001:504:d::9151:1	10G	Selective
6connect 198.32.176.51	8038 2001:504:d::33	1G	Open
AARNet 198.32.176.177	7575 2001:504:d::b1	10G	Selective
Academia Sinica Network(ASNet) 198.32.176.174	9264 2001:504:d::ae	10G	Selective
Advanced Wireless Network Co. Ltd.(IIG) 198.32.176.129	45430 2001:504:d::4:5430:1	1G	Selective
Akamai Prolexic DDoS Mitigation 198.32.176.228	32787 2001:504:d::3:2787:1		Selective
Akamai Technologies 198.32.176.127	20940 2001:504:d::2:940:1	100G	Open
Alibaba 198.32.176.180	45102	20G	Open
Amazon IVS 198.32.176.232	46489 2001:504:d::4:6489:1	30G	Selective
Amazon IVS 198.32.176.32	46489 2001:504:d::4:6489:2	30G	Selective
Amazon.com	16509	60G	Selective

How does BGP implement these routing policies?

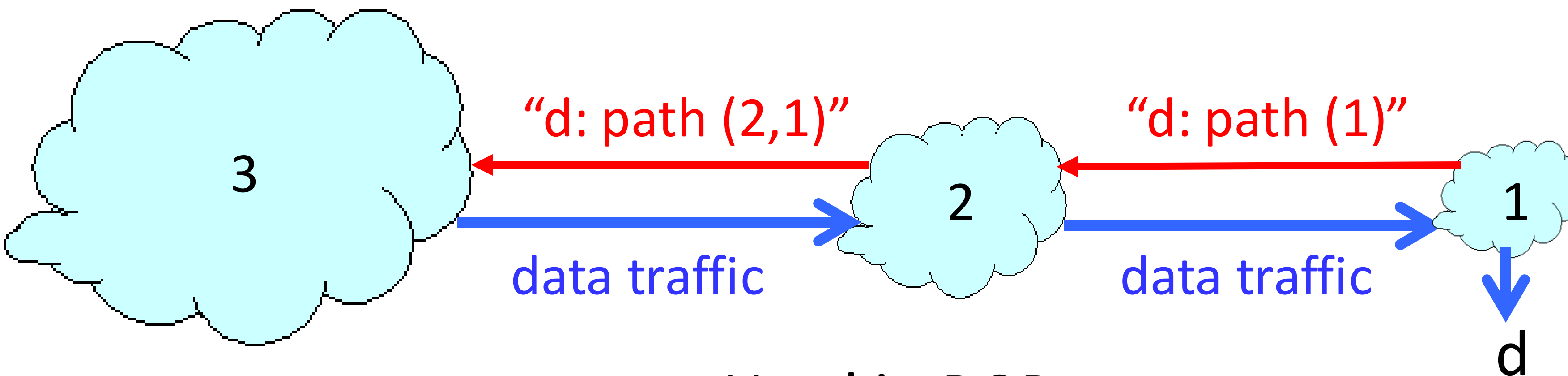
Internet Routing Protocol: BGP



Destination	Next-hop	AS Path
130.207.0.0/16	192.5.89.89	10578,..,2637
130.207.0.0/16	66.250.252.44	174,… ,2637

BGP: Path-Vector Routing

- Extension of distance-vector routing
  - Support flexible routing policies
- Key idea: advertise the entire path
  - Distance vector: send *distance metric* per dest d
  - Path vector: send the *entire path* for each dest d

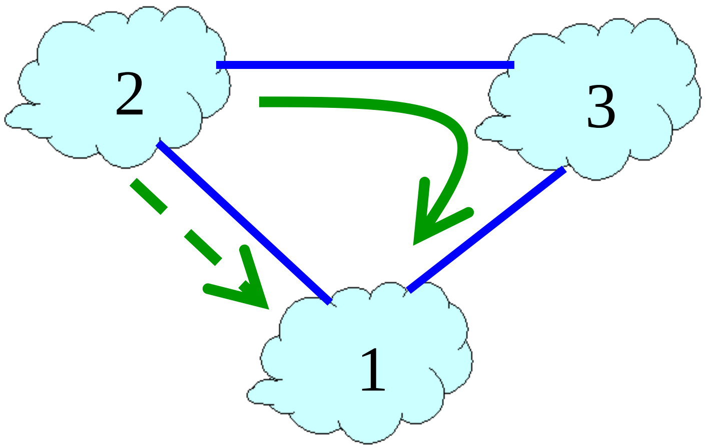


Used in BGP

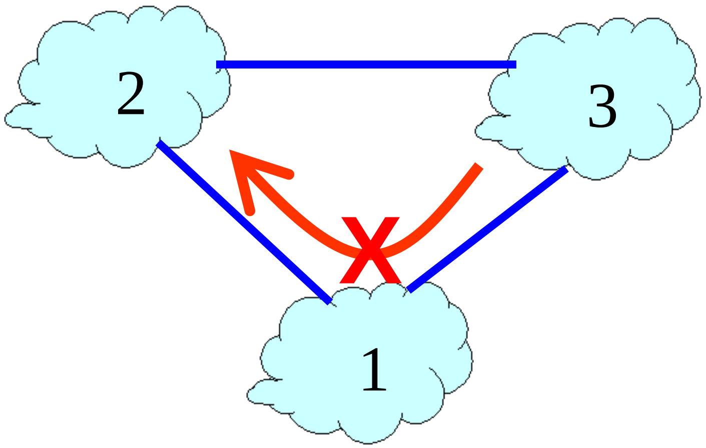
Path-Vector: Flexible Policies

- Each node can apply local policies
  - Selection: Which path to use?
  - Export: Which paths to advertise?

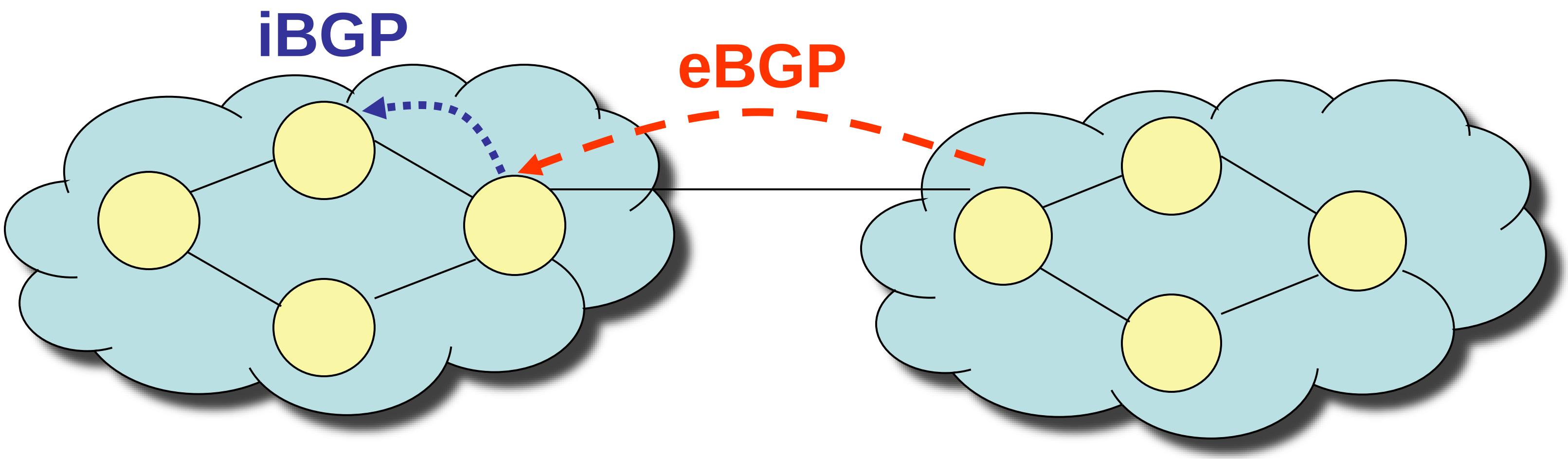
Node 2 prefers
"2, 3, 1" over "2, 1"



Node 1 doesn't let 3
hear the path "1, 2"



Two Flavors of BGP



- **External BGP (eBGP):** exchanging routes *between* ASes
- **Internal BGP (iBGP):** distributing routes to external destinations among routers *within an AS*
- **Interior Gateway Protocol (IGP):** distributing routes to interior destinations among routers within an AS (not BGP)

iBGP

- Full mesh: each eBGP router has an iBGP session with every other router in the AS

- Route reflection: each eBGP router has an iBGP session with a (logically central) route reflector, and each router has an iBGP session with the route reflector

Example BGP Routing Table

The full routing table

```
> show ip bgp

 Network             Next Hop          Metric LocPrf Weight Path
*>i3.0.0.0           4.79.2.1               0    110      0 3356 701 703 80 i
*>i4.0.0.0           4.79.2.1               0    110      0 3356 i
*>i4.21.254.0/23     208.30.223.5          49    110      0 1239 1299 10355 10355 i
* i4.23.84.0/22      208.30.223.5         112    110      0 1239 6461 20171 i
```

Specific entry. Can do longest prefix lookup:

```
> show ip bgp 130.207.7.237
BGP routing table entry for 130.207.0.0/16
Paths: (1 available, best #1, table Default-IP-Routing-Table)
  Not advertised to any peer
  10578 11537 10490 2637
    192.5.89.89 from 18.168.0.27 (66.250.252.45)
      Origin IGP, metric 0, localpref 150, valid, internal, best
      Community: 10578:700 11537:950
      Last update: Sat Jan 14 04:45:09 2006
```

Prefix

AS path

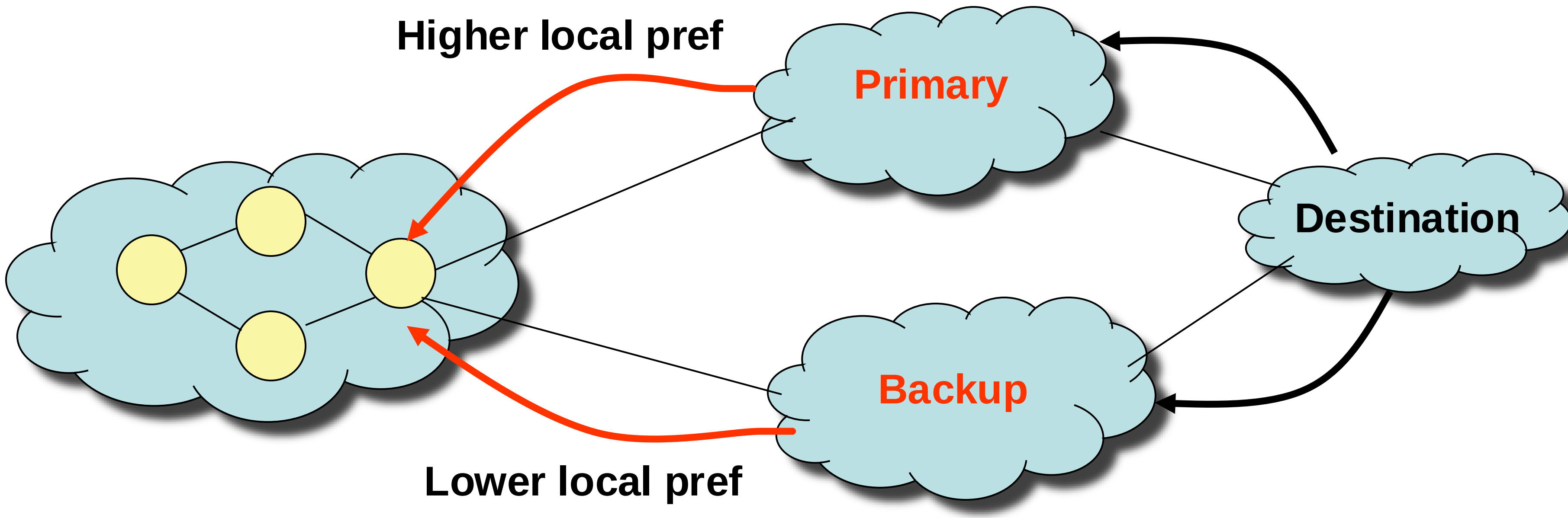
Next-hop

Routing Attributes and Route Selection

BGP routes have the following attributes, on which the route selection process is based:

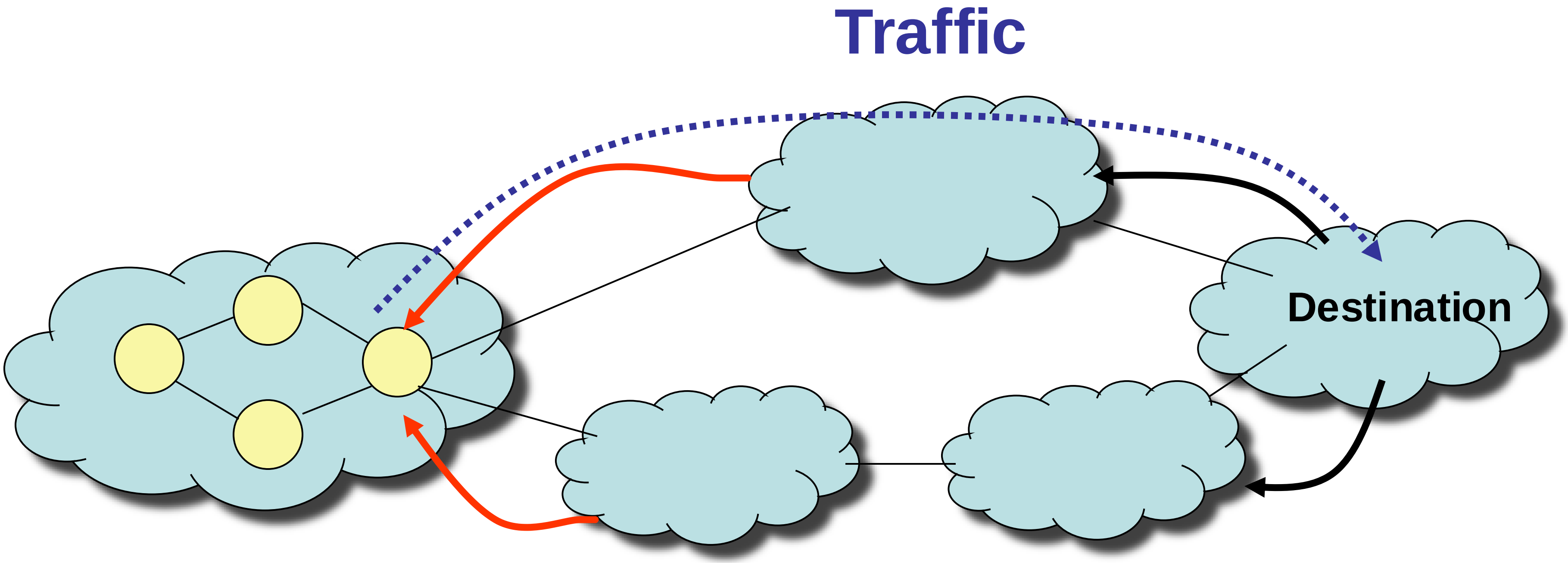
- **Local preference ("LOCALPREF"):** numerical value assigned by routing policy. Higher values are more preferred.
- **AS path length:** number of AS-level hops in the path
- **Multiple exit discriminator ("MED"):** allows one AS to specify that one exit point is more preferred than another. Lower values are more preferred.
- **eBGP over iBGP**
- **Shortest IGP path cost to next hop:** shortest way to get out of the AS en route to the destination ("hot-potato" routing)
- **Router ID tiebreak:** arbitrary tiebreak, since only a single "best" route can be selected

Local Preference



- **Control over *outbound* traffic**
- *Not* transitive across ASes
- Coarse hammer to implement route preference
- Useful for preferring routes from one AS over another (*e.g.*, primary-backup semantics)

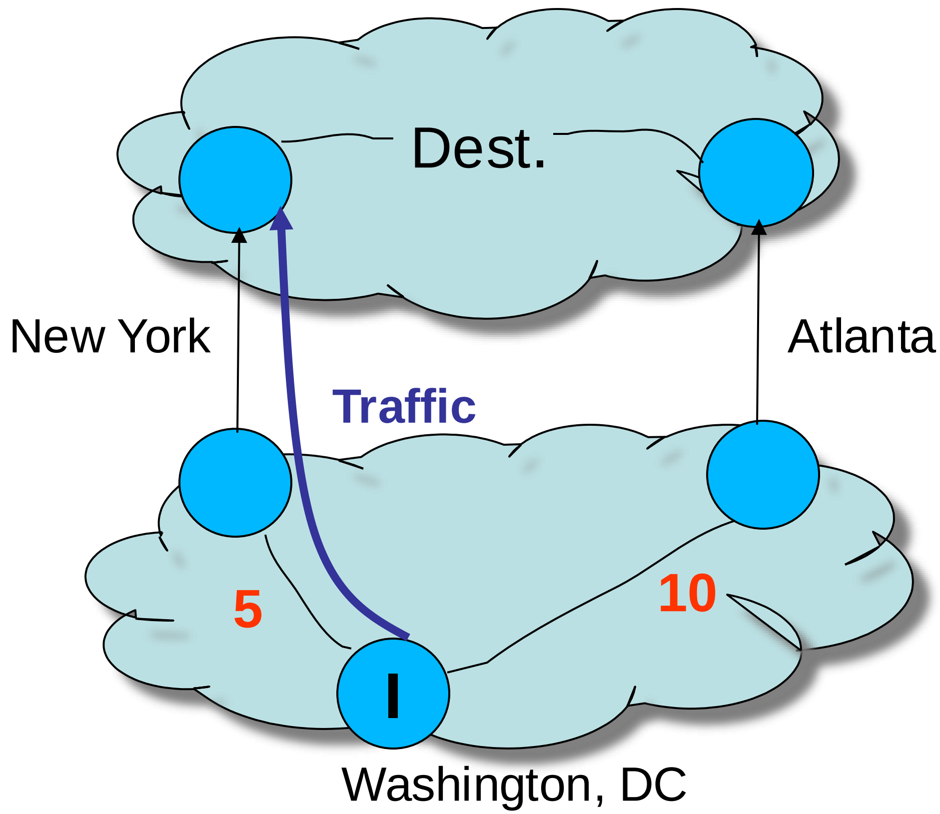
AS Path Length



- Among routes with highest local preference, select route with shortest AS path length
- Shortest AS path != shortest path, for *any* interpretation of "shortest path"

Hot-Potato Routing

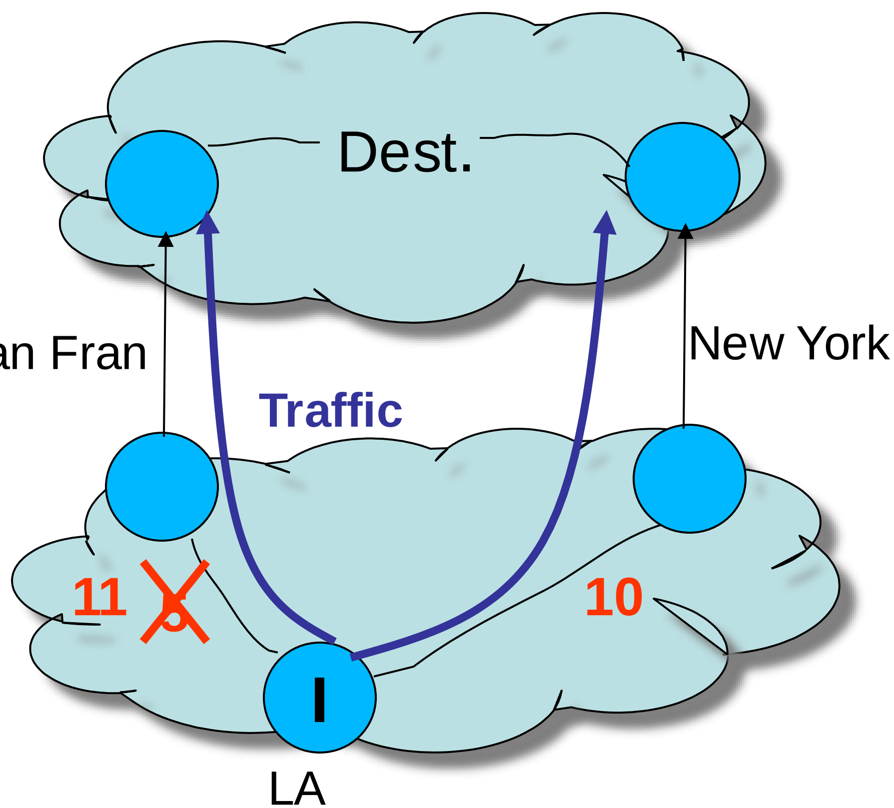
- Prefer route with shorter IGP path cost to next-hop
- *Idea:* traffic leaves AS as quickly as possible



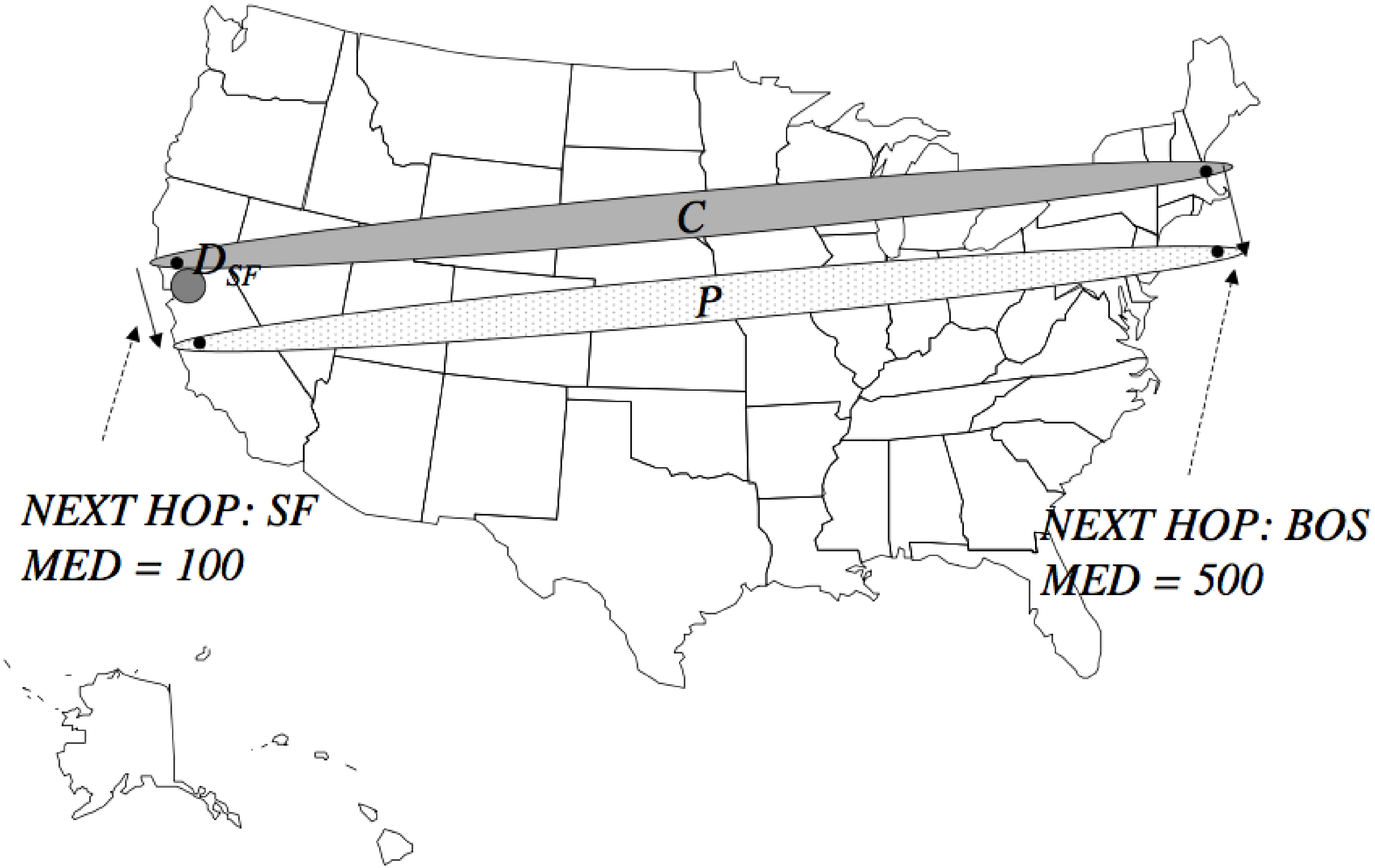
Common practice:
Set IGP weights in accordance with propagation delay (*e.g.*, miles, etc.)

Problems with Hot-Potato Routing

- Small changes in IGP weights can cause large traffic shifts



Multi-Exit Discriminator



BGP Complexity

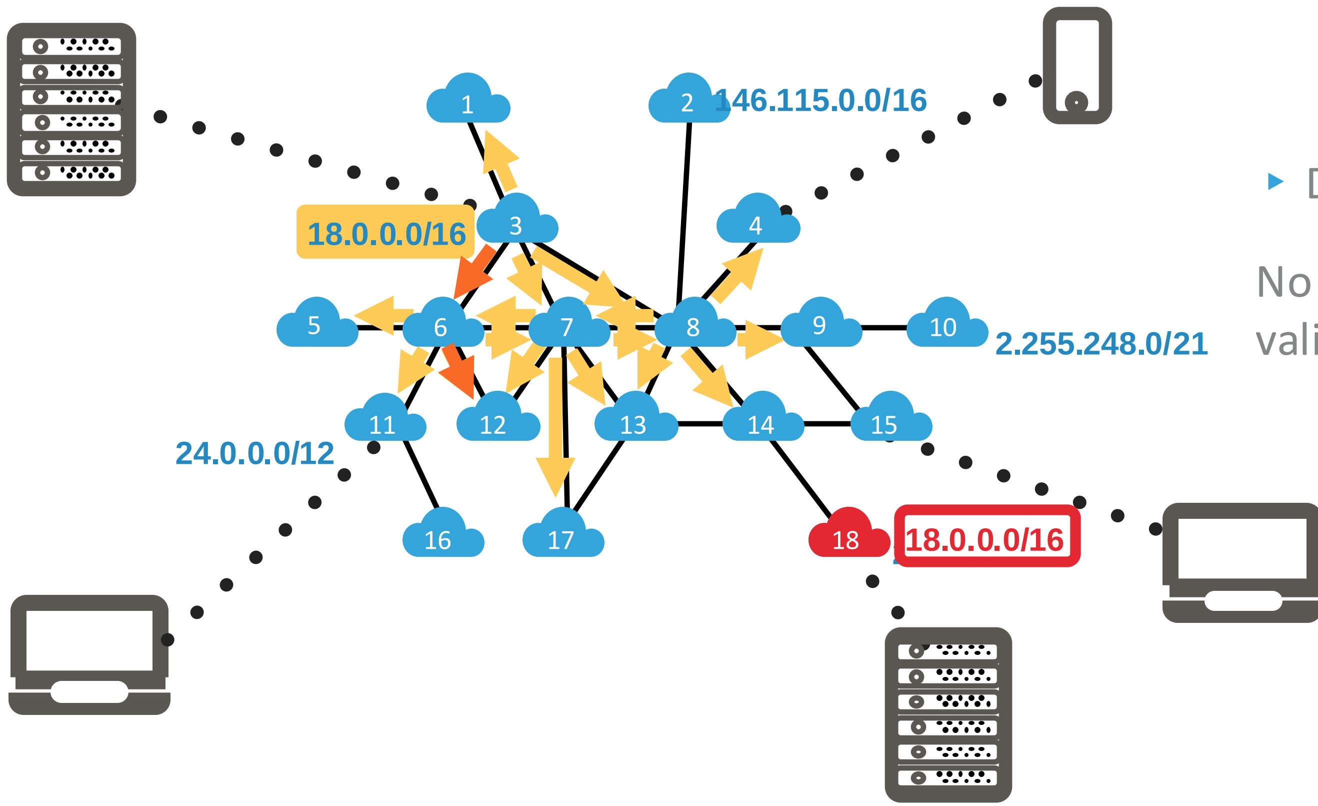
- BGP is a very complicated protocol
  - Too many knobs
  - Need to accommodate (sub-optimal) ISP policies
  - Requires complex, human configuration

Headache!

BGP Pitfalls and Problems

- Pitfalls and problems
  - No authentication
  - Misconfiguration
  - Convergence
  - Performance
  - Reliability
  - Stability
  - Security
  - And the list goes on…

Lack of authentication in BGP

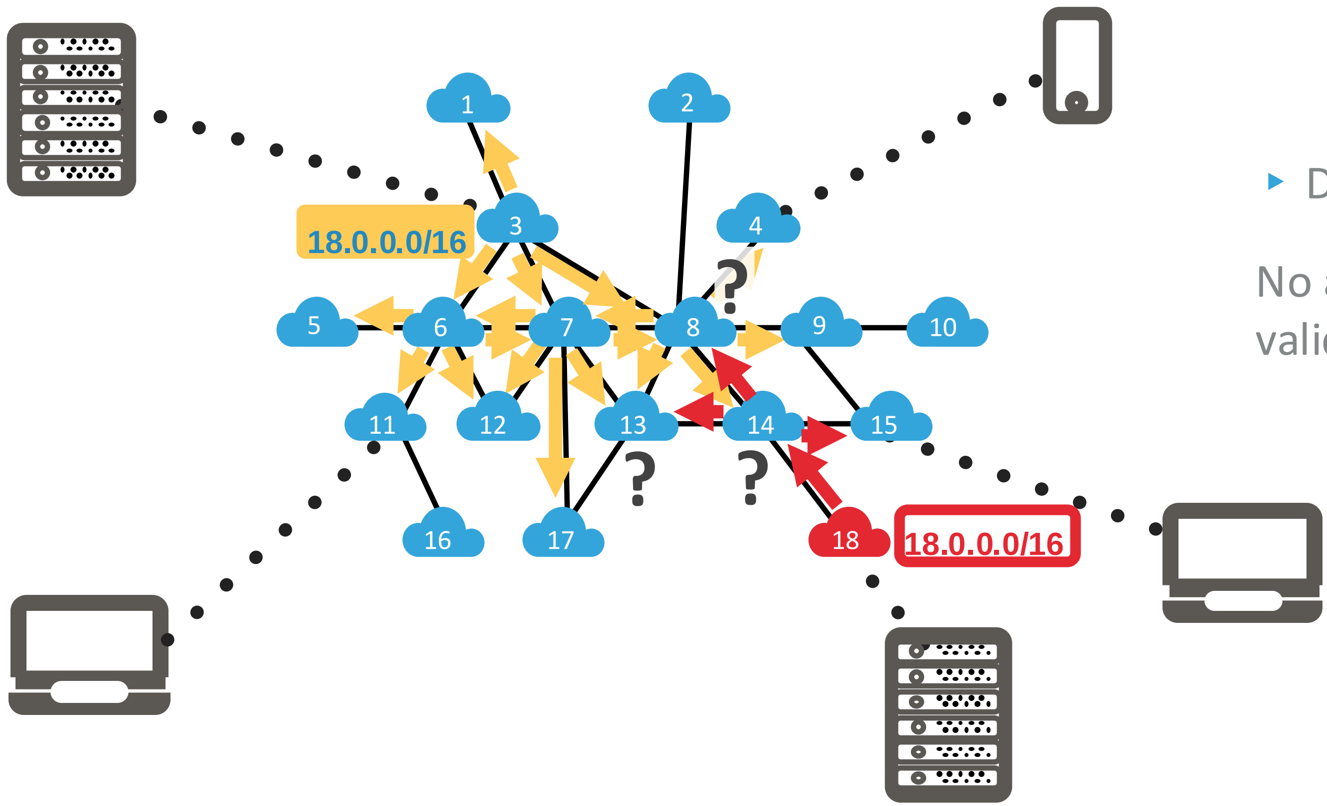


- Design flaw:

No authentication nor validation mechanism.

Slides credit: Cecilia Testart

Lack of authentication in BGP

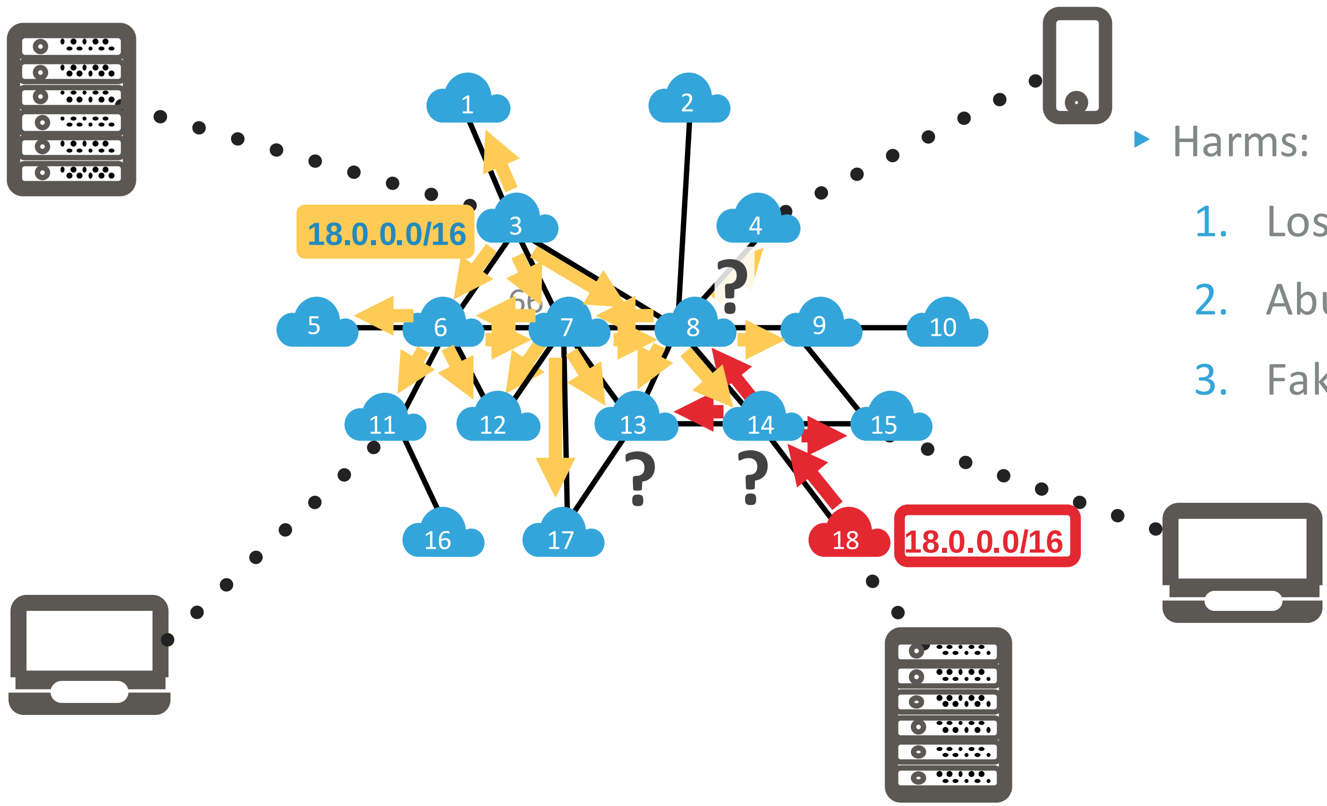


- Design flaw:

No authentication nor validation mechanism.

Slides credit: Cecilia Testart

Lack of authentication in BGP



- Harms:
  1. Loss of availability
  2. Abuse of IP addresses
  3. Fake end point

Slides credit: Cecilia Testart

Using BGP hijacks to steal cryptocurrencies

TECH / SECURITY / CRYPTO

Hackers emptied Ethereum wallets by breaking the basic infrastructure of the internet

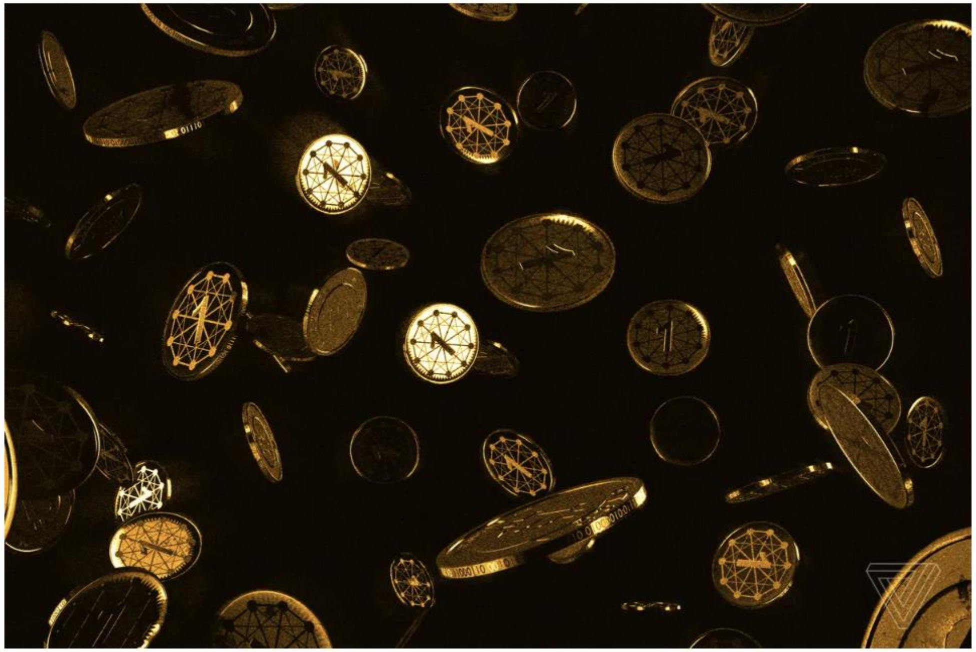


Illustration by Alex Castro / The Verge

By Russell Brandom

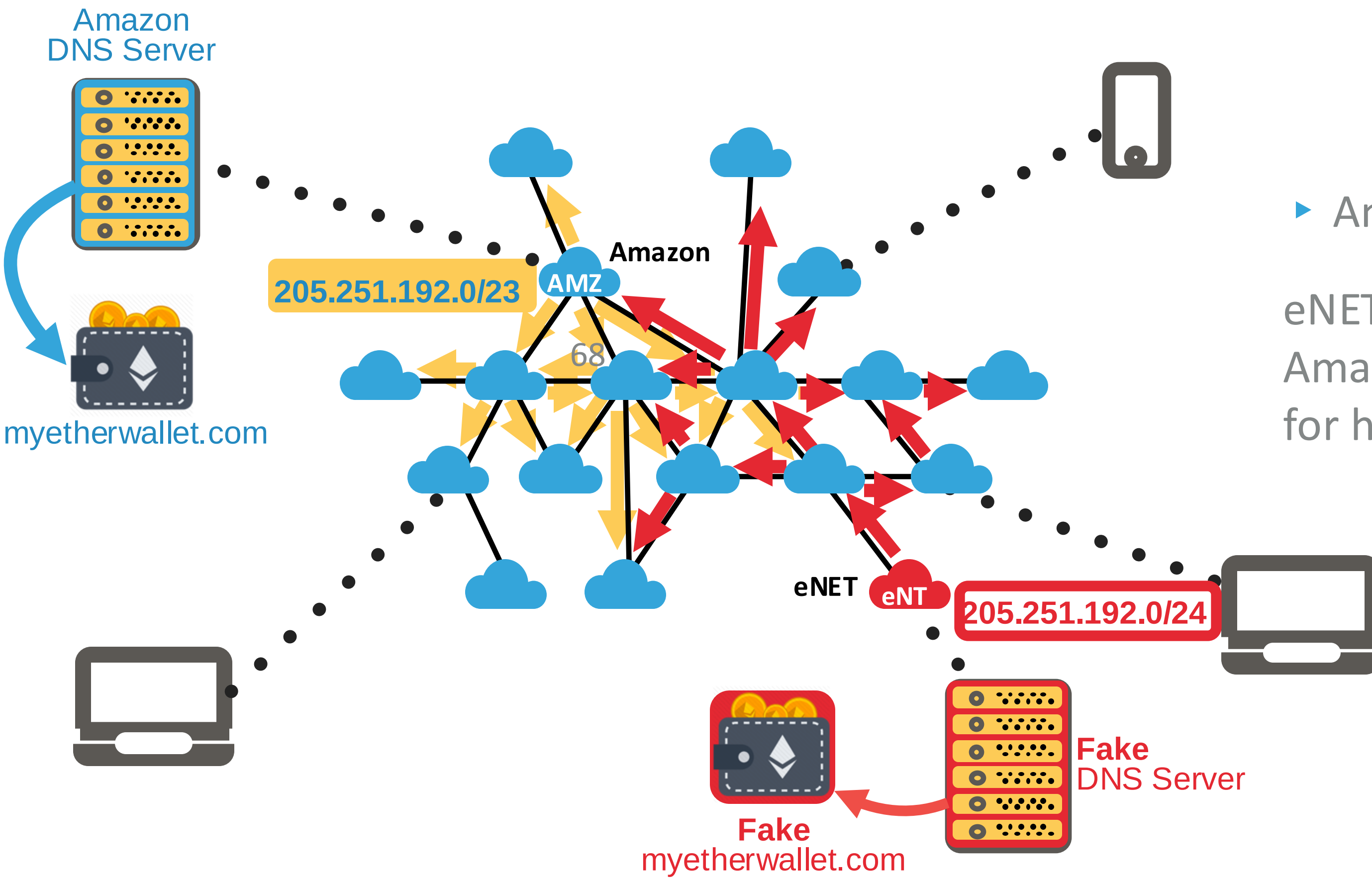
Apr 24, 2018, 1:40 PM EDT |

0

Comments

https://www.theverge.com/2018/4/24/17275982/myetherwallet-hack-bgp-dns-hijacking-stolen-ethereum

Using BGP hijacks to steal cryptocurrencies



- Amazon BGP hijack:

eNET announced part of Amazon address space for hours.

Stole $17M in Etherium

Slides credit: Cecilia Testart

Favorite Scapegoat!